Published online 14 May 2021

NAR Genomics and Bioinformatics, 2021, Vol. 3, No. 2 **1**
doi: 10.1093/nargab/lqab032

PolyQ length co-evolution in neural proteins

Serena Vaglietti[1] and Ferdinando Fiumara [1,2,*]

[1] *Rita Levi Montalcini* Department of Neuroscience, University of Torino, Torino 10125, Italy and [2]National Institute of Neuroscience (INN), University of Torino, Torino 10125, Italy

Received December 30, 2020; Revised February 10, 2021; Editorial Decision March 30, 2021; Accepted March 31, 2021

ABSTRACT

Intermolecular co-evolution optimizes physiological performance in functionally related proteins, ultimately increasing molecular co-adaptation and evolutionary fitness. Polyglutamine (polyQ) repeats, which are over-represented in nervous system-related proteins, are increasingly recognized as length-dependent regulators of protein function and interactions, and their length variation contributes to intraspecific phenotypic variability and interspecific divergence. However, it is unclear whether polyQ repeat lengths evolve independently in each protein or rather co-evolve across functionally related protein pairs and networks, as in an integrated regulatory system. To address this issue, we investigated here the length evolution and co-evolution of polyQ repeats in clusters of functionally related and physically interacting neural proteins in Primates. We observed function-/disease-related polyQ repeat enrichment and evolutionary hypervariability in specific neural protein clusters, particularly in the neurocognitive and neuropsychiatric domains. Notably, these analyses detected extensive patterns of intermolecular polyQ length co-evolution in pairs and clusters of functionally related, physically interacting proteins. Moreover, they revealed both direct and inverse polyQ length co-variation in protein pairs, together with complex patterns of coordinated repeat variation in entire polyQ protein sets. These findings uncover a whole system of co-evolving polyQ repeats in neural proteins with direct implications for understanding polyQ-dependent phenotypic variability, neurocognitive evolution and neuropsychiatric disease pathogenesis.

INTRODUCTION

Homopolymeric amino acid repeats (AARs), such as polyglutamine (polyQ), are increasingly recognized as mediators and regulators of protein function, interactions and evolution (e.g. 1,2,3,4,5,6,7). AAR length variation in proteins contributes to intraspecific phenotypic variability and interspecific divergence (3,5,7,8,9,10,11).

In particular, polyQ repeats have been more intensively studied in relation to the nervous system given their frequent occurrence in neural proteins, and because of their involvement in polyQ expansion-related neurodegenerative disorders such as Huntington disease (HD; 12,13,14). PolyQ repeats can mediate and regulate the interactions and activity of their parent proteins, which are often transcription factors and translational regulators (1,4,10,15,16).

While the preferential involvement of polyQ proteins in basic biological processes, such as transcription and translation, is well defined at the cellular and molecular levels (e.g. 4,7), their functional roles in higher-order biological processes at the organismal level, and their involvement in human diseases besides expansion-related disorders (13), are still not clearly understood. PolyQ repeats occur frequently in proteins involved in the development of the nervous system (12) with some degree of regional specificity (5), suggesting that polyQ variation may have a considerable impact on neural development and evolution. However, the overall functional meaning and quantitative dynamics of the complex, proteome-wide evolutionary variation patterns of polyQ stretches remain largely obscure.

Molecular co-evolution is a process thought to optimize physiological performance in pairs and networks of functionally related proteins, representing a general determinant of molecular co-adaptation and evolutionary fitness (17,18,19). In fact, coordinated evolutionary changes in different proteins facilitate the establishment or refinement of molecular interactions (18,19,20), to the point that the detection of molecular co-evolution between proteins predicts their functional and interaction dependencies (21,22,23,24).

While several examples are known of molecular co-evolution at the level of single amino acid substitutions arising from single nucleotide polymorphisms (SNPs; 19), it is unclear whether also AARs, by means of insertion/deletion ('indel') mutations, undergo similar co-evolutionary processes. Thus, it would be important to determine whether the observed evolutionary length variation of AARs, such as polyQ, occurs independently in each protein or, rather, in a more coordinated manner across functionally related proteins, as in an integrated regulatory system. Indeed, polyQ length co-evolution, through coordinated variation

*To whom correspondence should be addressed. Tel: +39 0116708486; Fax: +39 0116708174; E-mail: ferdinando.fiumara@unito.it

in the length and stability of CC structures mediating protein interactions, may have considerable functional impact on polyQ protein networks (1,4,6,7,25,26,27).

To address these issues, we focus here on polyQ repeats in Primates in search of possible signatures of length co-evolution in functionally related neural proteins. Toward this aim, we first develop a cross-database mapping of the higher-order functional roles and disease associations of polyQ proteins in the nervous system and allied structures, tracing at the same time the evolutionary history of polyQ length variation in Primates. We find evidence of function-related polyQ enrichment and hypervariability in protein clusters related especially to cognitive, behavioral and neuropsychiatric phenotypes, uncovering generalized patterns of polyQ length co-evolution in functionally related, physically interacting proteins pairs and networks.

MATERIALS AND METHODS

Data sources and *a priori* clustering of gene annotations and disease associations

The complete proteome of *Homo sapiens* was downloaded from Uniprot (uniprot.org) and polyQ repeats of four or more glutamine residues were detected using a previously developed script (5). In proteins with multiple repeats, the total polyQ length was calculated as the sum of all the individual repeat lengths. Lists of human genes associated with gene ontology (GO) terms, human phenotype ontology (HPO) terms and specific diseases were derived, respectively, from the Gene Ontology AmiGO database (amigo.geneontology.org; 28), the HPO database (hpo.jax.org; 29) and the genome-wide association study (GWAS) Catalog database (ebi.ac.uk; 30). The GWAS Catalog entries were filtered to include only intragenic SNP and insertion/deletion (indel) mutations, either coding or non-coding, thus excluding intergenic mutations or non-univocal gene-phenotype/-disease associations. These lists of terms and associated genes were analyzed using *ad hoc* Perl scripts (5) to select terms/associations of interest based on string/keyword matching. A list of text strings and keywords of interest covering the major aspects of neural pathophysiology (e.g. 'synap' and 'Parkinson'; Supplementary Table S1) was manually compiled using textbooks and manuals in the fields of neurobiology (31), neurology (32) and psychiatry (DSM-V; 33). The resulting list of terms/disease associations was then manually revised, removing obviously spurious terms, duplicated terms found using different strings and terms only tangentially related to the strings of interest. Thus, we obtained a final list of 1937 terms (Supplementary Table S2; 944 GO-BP terms, 739 HPO terms and 254 GWAS-derived disease associations). These terms were then manually clustered semantically into 49 clusters associated with major categorical/nosological subdomains (each comprising on average ∼40 GO/HPO/GWAS terms) grouped into 21 superclusters (Supplementary Figure S1A and Table S3). These, in turn, belonged to three major domains (Supplementary Figure S1A and Tables S3–S4), the first one related to neuroanatomy/neurobiology/neurological disorders (*NEU*), the second to neuropsychology/psychiatric disorders (*PSY*) and the third to structures/tissues allied to the nervous system in terms of ectodermal embryological derivation or close anatomical association, which are frequently involved in genetic syndromes with a neuropsychiatric component (e.g. skin/adnexa and cranio-facial features; '*other*', *OTH,* domain; (34,35). The clusters and superclusters are: *Supercluster* NEA: General neuroanatomy, neurobiology and related disorders. *Clusters*: *cel*, cellular elements of the nervous system (neurons, synapses, neurotransmitters, glia); *cns*, central nervous system; *pns*, peripheral nervous system; *aut*, automomic nervous system; *ent*, enteric nervous system; *atr*, atrophy of nervous system elements; *Supercluster* SEN: Sensory systems and related disorders. *Clusters*: *sen*, sensory system in general; *tou*, touch; *mec*, mechanosensation; *pro*, proprioception; *ter*, thermosensation; *pai*, pain; *vis*, vision; *eye*, eye; *hea*, hearing; *equ*, equilibrium; *che*, chemosensation; *sme*, smell; *tas*, taste; *Supercluster* MOV: Motor system and related disorders. *Clusters*: *mov*, motor system, movement; *ref*, nervous reflexes; *Supercluster* EPI: Epilepsy and seizures. *Clusters*: *epi*, epilepsy; *Supercluster* MYE: Myelination and related disorders. *Clusters*: *mye*, *myelin*; *msc*, multiple sclerosis; *Supercluster* CIR: Brain circulation, blood-brain barrier, cerebrospinal fluid and related disorders. *Clusters*: *bbb*, blood brain barrier; *csf*, cerebrospinal fluid; *str*, stroke; *ane*, brain aneurysm; *syc*, syncope; *Supercluster* BEH: Behavior and related disorders. *Clusters*: *beh*, behavior; *Supercluster* LCO: Language, cognition and releated disorders. *Clusters*: *lan*, language; *cog*, cognition; *dem*, dementia; *Supercluster* NDE: Autism and other neurodevelopental disorders. *Clusters*: *asd*, autism spectrum disorders; *adh*, ADHD; *Supercluster* PSY: Schizophrenia and psychosis. *Clusters*: *sch*, schizophrenia; *psy*, psychosis; *Supercluster* MOO, Mood disorders. *Clusters*: *dep*, depression; *bip*, bipolar disorder; *Supercluster* ANX, Anxiety-related disorders. *Clusters*: *anx*, anxiety; *Supercluster* OCD: Obsessive-compulsive disorder. *Clusters*: *ocd*, obsessive-compulsive disorder; *Supercluster* SOM: Somatic symptom and related disorders. *Clusters*: *ast*, astasia; *Supercluster* EAT: Eating disorders. *Clusters*: *edi*, eating disorders; *Supercluster* SLE: Sleep and related disorders. *Clusters*: *sle*, sleep; *Supercluster* CND: Disruptive, impulse-control and conduct disorders. *Clusters*: *cnd*, disruptive, impulse-control and conduct disorders; *Supercluster* ADD: Addiction. *Clusters*: *add*, addiction; *Supercluster* PRS: Personality disorders. *Clusters*: *prs*, personality disorders; *Supercluster* CRA: Skull and related disorders. *Clusters*: *cra*, cranium; *Supercluster* AND: Skin, adnexa and related disorders. *Clusters*: *hai*, hair and keratinization. The superclusters NEA, SEN, MOV, EPI, MYE, SLE and CIR belong the *NEU domain,* BEH, LCO, NDE, PSY, MOO, ANX, OCD, SOM, EAT, CND, ADD and PRS to the *PSY domain*, whereas CRA and ADN to the *OTH domain.* A detailed description of the clusters/superclusters is reported in Supplementary Table S2. Each cluster, if not obviously related to specific diseases, also included the disorders related to the associated semantic area. Thus, for example, the cluster 'hearing' included terms/disease associations related, for instance, to 'hearing loss'. In each cluster, the gene/protein IDs related to all the terms/disease associations were pooled, and a list of unique gene/protein IDs was obtained for subsequent polyQ protein enrichment analyses. The relative contribution of GO-BP terms, HPO

terms and GWAS-derived disease associations to each cluster reflected the nature of both the annotation databases and the clusters, some of which are mostly biology-centered, others more clinically focused, and some others covering both domains (Supplementary Figure S1A and Table S4). Thus, the 'cellular' (*cel*) and 'dementia' (*dem*) clusters, comprised mostly GO-BP- or GWAS-derived terms, respectively, whereas many clusters such as the 'language' (*lan*) or 'cognition' (*cog*) were more heterogeneous. The clusters also varied in size in relation to their semantic extension, so that, for example, the 'motor system' (*mov*) cluster comprised 90 terms whereas the 'proprioception' (*pro*) cluster only two. The number of genes in each cluster also varied as a function of the number of genes associated with each individual term of a given cluster.

PolyQ protein enrichment analysis

We performed an analysis of the relative enrichment of proteins bearing polyQ repeats in the 49 protein clusters that were defined previously using χ^2 tests on 2×2 contingency tables reporting the number of proteins with or without polyQ repeats in either a cluster or in the whole proteome, followed by a Benjamini–Hochberg correction for multiple testing (36), with a false discovery rate (FDR) set to 0.05. These tests compared, for each cluster, the observed proportion of polyQ proteins within the cluster with the proportion expected based on the proteome-wide occurrence of polyQ proteins.

Analysis of polyQ repeats in primate protein orthologs

To analyze the length variation of the polyQ repeats found in human proteins, we downloaded from Ensembl (*ensembl.org*) the list of coding transcripts of *H. sapiens* (Hom sap) and selected, for each gene, the longest transcript and the corresponding protein. We also downloaded the gene/protein IDs of the available orthologs of human polyQ proteins in 22 primate species (*Pan troglodytes* (Pan tro), *Pan paniscus* (Pan pan), *Gorilla gorilla* (Gor gor), *Pongo abelii* (Pon abe), *Nomascus leucogenys* (Nom leu), *Mandrillus leucophaeus* (Man leu), *Cercocebus atys* (Cer aty), *Papio anubis* (Pap anu), *Macaca nemestrina* (Mac nem), *Macaca mulatta* (Mac mul), *Macaca fascicularis* (Mac fas), *Chlorocebus sabaeus* (Chl sab), *Rhinopithechus roxellana* (Rhi rox), *Rhinopithechus bieti* (Rhi bie), *Colobus* angolensis (Col ang), *Saimiri boliviensis* (Sai bol), *Cebus capucinus* (Ceb cap), *Callithrix jacchus* (Cal jac), *Aotus nancymae* (Aot nan), *Tarsius/Carlito syrichta* (Tar syr), *Propithechus coquereli* (Pro coq), *Otolemur garnettii* (Oto gar)) belonging to the major taxonomic groups of Primates, including only orthologs coded as '*one-to-one*' in the database, that we combined in a single orthology table. The primate proteomes were then analyzed to identify polyQ repeats and quantify their total length in each ortholog protein, as described above for the human proteome (5). For each polyQ protein, we also calculated the coefficient of variation of the total repeat length (CV_{rl}) across orthologs, counting proteins with either phylogenetically constant (i.e. $CV_{rl} = 0$) or variable (i.e. $CV_{rl} > 0$) polyQ lengths. For each cluster, we quantified the proportion of proteins with either constant or variable total polyQ length and the ratio between these two proportions. Furthermore, for the subset of proteins in each cluster with $CV_{rl} > 0$, we calculated the mean CV_{rl}.

Interactome analyses

The total human protein–protein interaction (PPI) network was downloaded from BioGrid (*thebiogrid.org*). We extracted from this database all the physical interactions involving two polyQ proteins (including self-interactions) using an *ad hoc* Perl script. Using this dataset, we performed two types of analyses. First, we extracted subnetworks of polyQ proteins either belonging to a given cluster (containing at least 20 polyQ proteins) or chosen at random in equal numbers among the polyQ proteins. Some of the proteins found in the clusters or in the random control groups were not present in the total polyQ interaction network because their interactions are not known or listed in the Biogrid database. However, the number of these missing proteins was similar in the cluster-related and in the random control polyQ protein sets, so that the related subnetworks that were analyzed were on average equinumerous (see Supplementary Figure S4E). For each cluster-related subnetwork, we extracted five control subnetworks. In statistical analyses of the subnetwork properties, the values of the parameters of interest (*number of nodes*, *average number of neighbors*, *network density*, *clustering coefficient*, *isolated nodes* and *connected components*) associated with the five control subnetworks were averaged. These averaged values of the random subnetworks were then compared with those of the cluster-related subnetworks. Second, we extracted subnetworks of polyQ proteins based on their known interactions or, alternatively, at random using a Perl script, and compared their *Z*-score distributions and mean values in subsequent polyQ length co-evolution analyses. A subnetwork of the first type was obtained by initially selecting an arbitrary set of 10 polyQ 'seed' proteins at random and then searching systematically for all their known interactors in the total polyQ protein interactome. Thus, this subnetwork would be formed by the 10 initial seed proteins plus their *n* known interactors and all their mutual pairwise interactions. A corresponding random control subnetwork was generated by selecting the same 10 seed proteins and *n* other polyQ proteins selected at random irrespective of their known interactions. As for the first type of analysis, some of the proteins found in the interactome-based or in the random control sets were not present in the total polyQ interaction network, but again the number of these missing proteins was not different in the two types of protein sets, so that the related subnetworks that were analyzed were on average equinumerous (see Supplementary Figure S4E). We generated 100 interactome-based subnetworks and 100 control subnetworks in pairwise combinations and used them for polyQ length co-evolution analyses (see below).

Analysis of polyQ length co-evolution

To identify pattern of polyQ length co-evolution in protein pairs and networks, we performed three types of analyses. First, we performed a cluster analysis of polyQ length variation in primates using a *protein* x *species* matrix containing, when available, the total polyQ length in each protein

ortholog. This table was analyzed using Cluster 3.0 (37) with data normalization in both rows and columns, selecting 'Spearmann rank correlation' as similarity metric and 'average linkage' as clustering method (5). TreeView (38) was used to visualize the heatmap. Second, using an *ad hoc* R script, we generated a matrix reporting the *r* correlation coefficients between the total polyQ lengths of two proteins, for all the pairwise combinations of polyQ proteins. This table included only proteins ($n = 225$) with at least 10 available orthologs in primates and a $CV_{rl} > 0$. We used this table to identify pairs of proteins with high levels of correlation (positive or negative) between their total polyQ lengths in primate phylogeny. Third, to compare the degree of correlation (i.e. length co-evolution) of entire sets of proteins in primates, we transformed the calculated r coefficients into *Z*-scores (using the formula $Z = [\ln(1+r) - \ln(1-r)]/2$) which have a normal distribution (39). From this *Z*-score table, we extracted sets of *Z*-values related to pairs of polyQ proteins belonging to functional clusters or to equinumerous sets of randomly selected polyQ proteins. For each cluster-related protein set, we generated 10 random protein sets and averaged their mean *Z*-values. These averaged control *Z*-values were then used for comparisons with the cluster-derived values. The *Z*-values used when comparing two 'control' protein sets ('random 1' (*r1*) *versus* 'random 2' (*r2*), see 'Results' section) with each other were similarly generated by averaging the mean *Z*-values of 10 random protein sets. Of the total number of pairwise combinations between proteins belonging to a functional cluster, or to a random set, only those between proteins that were both present in the table (>10 available orthologs, $CV_{rl} > 0$) had a defined *Z*-value. For each protein set, we measured the distribution and mean value of absolute *Z*-scores as indexes of the overall level of correlation in the set. A similar analysis was performed when comparing sets of polyQ proteins selected because of their known interactions *versus* randomly selected polyQ proteins (see above 'Interactome analyses' and the 'Results' section). Overall, in these analyses there was no difference in the proportion of protein combinations with no defined *Z*-value in the cluster-/interactome-related *versus* randomly selected protein sets (see 'Results' section).

Phylogenetic trees

The standard primate phylogenetic tree was derived from TimeTree (timetree.org; 40) in Newick format and then elaborated using MEGA7 (41) and iTOL (42) to generate scaled and unscaled trees. The primate silhouettes were downloaded from PhyloPic (phylopic.org) and modified. Credits: *H. sapiens*, *P. paniscus*, *G. gorilla* (after Colin M. L. Burnett) and *Pongo pygmaeus* by T. Michael Keesey; *P. troglodytes* by Jonathan Lawley; *C. jacchus* by Yan Wong from a drawing by T. F. Zimmermann; *P. anubis* by Owen Jones; *Rhinopithecus* by Yan Wong from a drawing by Joseph Smit; *C. angolensis* by Yan Wong from a drawing in The Century Dictionary (1911); *C. capucinus* by Sarah Werning (https://creativecommons.org/licenses/by/3.0/), *A. nancymae* by E. Lear, 1819 (vectorization by Yan Wong), *C. syrichta* by Yan Wong, *P. coquereli* by Terpsichores (https://creativecommons.org/licenses/by/3.0/); *O. garnettii* by Joseph Wolf (1863) vectorization by Dinah Challen; Cercopithecinae (for *C. sabaeus*) by I.A., *Macaca* spp., *N. leucogenys*, *M. leucophaeus*, *C. atys* and *S. boliviensis* uncredited).

Software

Ad hoc software was written in Perl language (www.perl.org). Alignments of protein sequences were obtained using ClustalW (43), and protein schemes using Prosite MyDomains (44). Cytoscape (45) was used to generate and analyze network graphs. Data analysis and statistics were performed using Excel (Microsoft), SPSS-25 (IBM) and Statistica (Tibco Data Science) software, which were also used to generate graphs. Photoshop Elements 11 (Adobe) was used for image editing.

Data analysis and statistics

Data are displayed as mean ± standard error of mean (SEM). Statistical tests (Student's *t*-test, one-way ANOVA, Fisher's exact test and χ^2 test) were performed where appropriate, as indicated in 'Results' section. A value of $P < 0.05$ was considered as statistically significant in all instances.

RESULTS

A priori cross-database clustering of gene annotations related to the nervous system and allied structures

As a first step in our analysis of the evolution and co-evolution of polyQ repeats in neural proteins, we developed a higher-order classification of the functional roles and disease associations of human polyQ proteins. Toward this aim, by using a cross-database integration of protein functional annotations and disease associations, we obtained clusters of functionally related proteins linked to major aspects of the pathophysiology of the nervous system and allied structures, and then we analyzed the occurrence and enrichment of polyQ proteins in each one of these clusters (Figure 1 and Supplementary Figure S1).

Specifically, we developed an *a priori* semantic clustering of gene functional annotations (GO, biological process (GO-BP) terms), phenotypic annotations (HPO terms) and GWAS-derived disease associations, and we pooled the associated genes/proteins into clusters (Figure 1A; Supplementary Figure S1A and Tables S1–S3). We semantically clustered 1937 terms/disease associations (Supplementary Table S1; 944 GO-BP terms, 739 HPO terms and 254 GWAS-derived disease associations) into 49 clusters related to major functional/nosological subdomains (e.g. *language*, *schizophrenia*), each comprising on average ∼40 GO/HPO/GWAS terms (Supplementary Figure S1A, Table S2 and 'Materials and Methods' section). These clusters were in turn grouped into 21 superclusters belonging to three major domains (Supplementary Figure S1A and Tables S2–S3), the first one related to neuroanatomy/neurobiology/neurological disorders (*NEU domain*), the second one to neuropsychology/psychiatric disorders (*PSY domain*), and the third one to structures/tissues allied to the nervous system, in terms of anatomical association or ectodermal embryological derivation, which are frequently involved in genetic

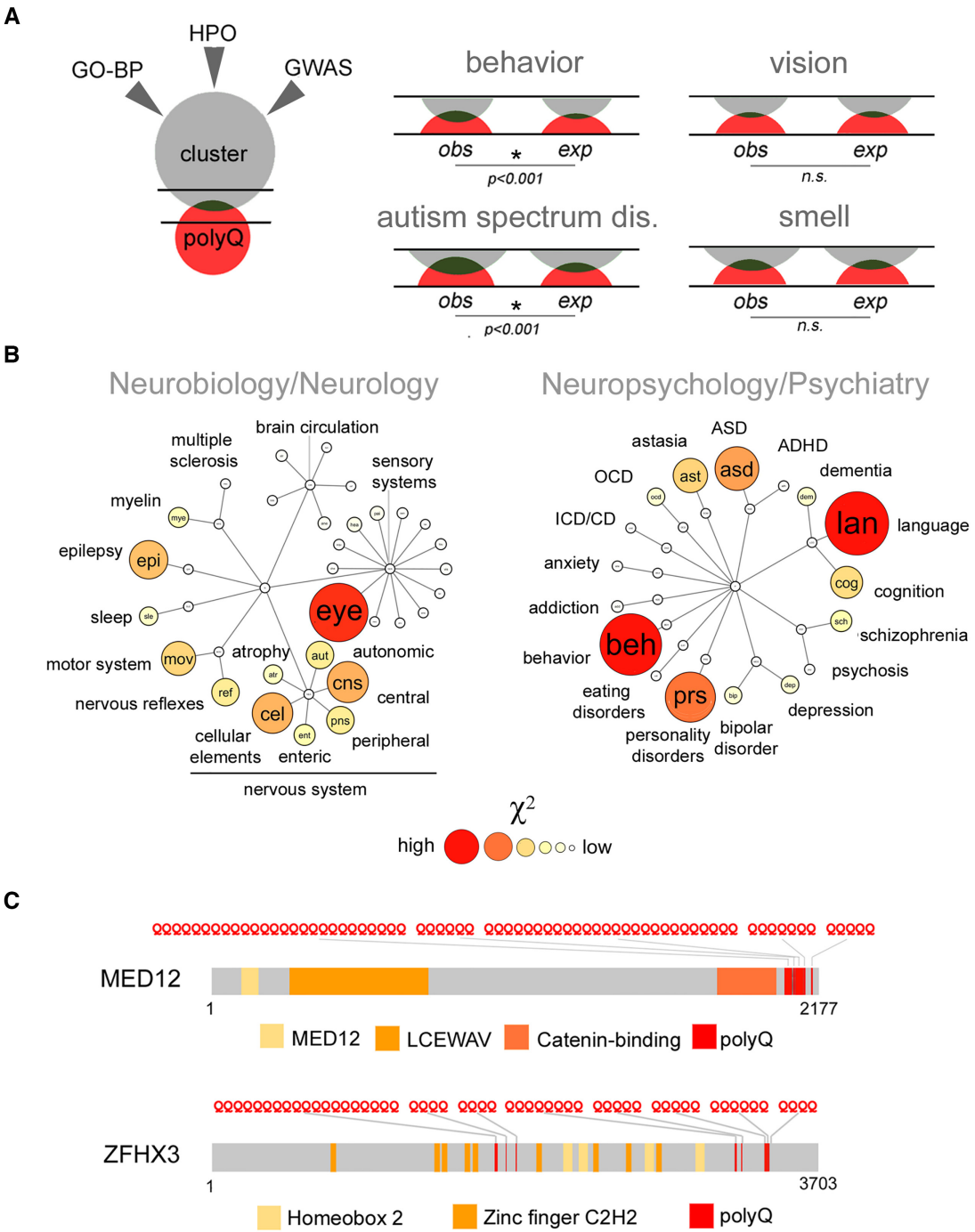

Figure 1. PolyQ proteins are over-represented in specific functional clusters of neural proteins. (**A**) Schematic representation of the enrichment analysis of polyQ proteins in protein sets associated with clusters of functional terms (GO-BP), phenotypic terms (HPO) and disease associations (GWAS). The *left panel* shown a Venn diagram representing, as a *red circle,* the polyQ protein set and, as a *gray circle*, the set of proteins associated with one of the 49 clusters of interest. The *darker overlap area* between the two circles represents polyQ proteins which are also part of the cluster. The *panels on the right side* represent details of the observed and expected (assuming a random distribution of polyQ proteins) overlap areas of two circles such as those shown in the *left panel* (between the two *horizontal lines*) for the indicated clusters. Note how for the *behavior* and *autism spectrum disorders* clusters, but not for the *vision* and *smell* clusters, the overlap area exceeds random expectations. (**B**) Network representations of the clusters and superclusters of the NEU (*left*) and PSY (*right*) domains. The *central nodes* in the two networks represent domains, the more *peripheral nodes* represent clusters and the intermediate nodes represent *superclusters*. The size of the peripheral cluster nodes (*small* to *large*) and color hue (from *light yellow* to *red*) indicate increasing χ^2 values proportional to the relative enrichment of polyQ proteins. The smallest peripheral nodes (clusters) in *white* are those in which there was no statistically significant enrichment of polyQ proteins. (**C**) Representative examples of proteins physiopatologically linked to language, i.e. MED12 (88) and epilepsy, i.e. ZFHX3 (89), in the PSY and NEU domains, respectively. Each protein is represented by a *gray bar* with the polyQ repeats highlighted in *red*. The primary sequence of the repeats is *above each bar*. Other protein domains are highlighted in other colors as indicated *below each bar*. The length of the primary sequence of each protein is indicated *below each bar*.

syndromes with a neuropsychiatric component (e.g. neuro-/splancno-cranium and skin/adnexa; '*other*', *OTH domain*; 34,35).

Overall, most of the clusters (77.5%) contained between 100 and ∼4000 genes/proteins, with a minority of the clusters with <100. These proportions were essentially similar in all three domains in comparison with the overall distribution ($P > 0.72$ in all instances, Fisher's exact test; Supplementary Figure S1B and C).

These results show that the *a priori* cross-database clustering of terms, and pooling of the associated genes, is effective in generating large clusters of functionally related genes/proteins suitable for polyQ enrichment analyses in the domains of interest.

Function-related enrichment of polyQ proteins in neural protein clusters

We then analyzed the relative distribution and enrichment of polyQ proteins in each cluster. Toward this aim, we first identified the entire repertoire of human polyQ proteins, i.e. proteins containing at least one stretch of four consecutive glutamine residues (as in 5; 472 proteins; Supplementary Table S4). This length threshold allows one to include proteins bearing polyQ repeats at different stages of their 'life cycle' and within regions of cryptic simplicity with more fragmented polyQ repeats (5,10,46). Then, in each one of the 49 clusters, we identified the polyQ proteins (Figure 1A and B; Supplementary Table S5) and calculated their relative enrichment in comparison with the whole proteome using χ^2 tests with the Benjamini–Hochberg correction (FDR = 0.05). We found that 27 of the 49 clusters displayed a significant over-representation of polyQ proteins (Figure 1B and Supplementary Table S5), thus indicating a non-generalized, but rather function-related enrichment of polyQ repeats in neural proteins. Thus, for instance, we found a significant over-representation of polyQ proteins in the clusters related to *behavior* (beh) and *autism spectrum disorders* (asd) but not in those related to *vision* (vis) or to the sense of *smell* (smell; Figure 1A and B).

Notably, four of the five clusters with significant polyQ enrichment and the highest χ^2 values belonged to the PSY domain (*language*, *behavior*, *personality disorders*, *autism spectrum disorders,* Figure 1B and Supplementary Table S5). In the NEU domain, the highest ranking clusters related to neuroanatomy/neurobiology were the *eye* (eye), *central nervous system* (cns) and *cellular elements* (cel) clusters, whereas *epilepsy* (epi) and *motor system* (mov) were the highest-ranking neurology-related clusters (Figure 1B and Supplementary Table S5).

Interestingly, we also observed a significant enrichment of polyQ repeats in both cluster of the OTH domain related to the pathophysiology of neuro-/splancno-cranium (cluster *cra*; e.g. cranial ossification, nose/palate development) and neuroectodermal epidermal/adnexal structures (cluster *hai*, e.g. skin keratinization, hair features; Supplementary Figure S1D and Table S5). This indicates that polyQ-dependent regulation may also impact structures that are anatomically or embryologically allied to the nervous system (see 'Discussion' section).

Figure 1C and Supplementary Figure S1E show examples of proteins in clusters with the highest degrees of polyQ enrichment in the three domains.

Taken together, these findings indicate that polyQ repeats are significantly enriched only in specific subsets of functionally related neural proteins, and especially in certain clusters of the PSY domain including *language*, *cognition* and *behavior*.

Evolutionary history of polyQ length variation in Primates

As a second step in our analysis, we reconstructed the proteome-wide evolutionary history of polyQ repeats, in terms of both occurrence and length variation, in 23 species belonging to the major taxonomic groups of Primates, ranging from *O. garnettii* to *H. sapiens* (Figure 2A and Supplementary Figure S2A).

Toward this aim, we first identified the polyQ proteins in the 23 primate proteomes and we calculated their percent occurrence, which was substantially stable (∽2%) throughout primate phylogeny, displaying no significant correlation with evolutionary distances (Supplementary Figure S2B).

Then, we systematically analyzed the length variation of polyQ repeats across human proteins and their available orthologs in the other 22 primate species (Figure 2B and C; Supplementary Figure S2C). We included in this analysis polyQ proteins with at least 10 available orthologs (i.e. 406 proteins, 86% of the total). For each protein, we calculated the total polyQ length as the sum of the lengths of all the individual glutamine repeats of at least four residues. The mean total polyQ length in the human proteins was 9.29 ± 0.6 residues (Supplementary Table S6). We found that the total polyQ length in each protein can vary broadly throughout primate phylogeny or it can be, at the opposite, surprisingly stable (Figure 2B and Supplementary Figure S2C). Overall (Figure 2C, *left panel*), 47% of the polyQ proteins displayed a constant total polyQ length (coefficient of variation of total repeat length, $CV_{rl} = 0$; $n = 191$), whereas 53% displayed variable degrees of length variation ($CV_{rl} > 0$, mean $CV_{rl} = 0.10 \pm 0.005$, $n = 215$).

Evolutionary polyQ length hypervariability in neuropsychology-/psychiatry-related protein clusters

Our previous findings prompted us to explore whether the length of polyQ repeats may also have varied throughout primate phylogeny in a function-related manner. Thus, for each cluster (containing at least 20 polyQ proteins), we measured two parameters defining the overall level of polyQ variation (Figure 2D and E; Supplementary Figure S2D).

The first parameter is the ratio, in each cluster, between the number of proteins with variable polyQ length (i.e. with $CV_{rl} > 0$) and the number of proteins with constant polyQ length (i.e. with $CV_{rl} = 0$) across primates. This analysis revealed different proportions of variable *versus* non-variable polyQ lengths across clusters (Supplementary Figure S2D). Interestingly, the clusters with the highest ratios (i.e. greater variation) belonged to the PSY domain (Supplementary Figure S2D). Indeed, the average ratio was significantly higher in the PSY domain clusters in comparison with non-PSY domain clusters (1.56 ± 0.30 *versus* 0.93 ± 0.05, $n = 8$ *versus* 13, respectively, $P < 0.02$ *t*-test; Figure 2D).

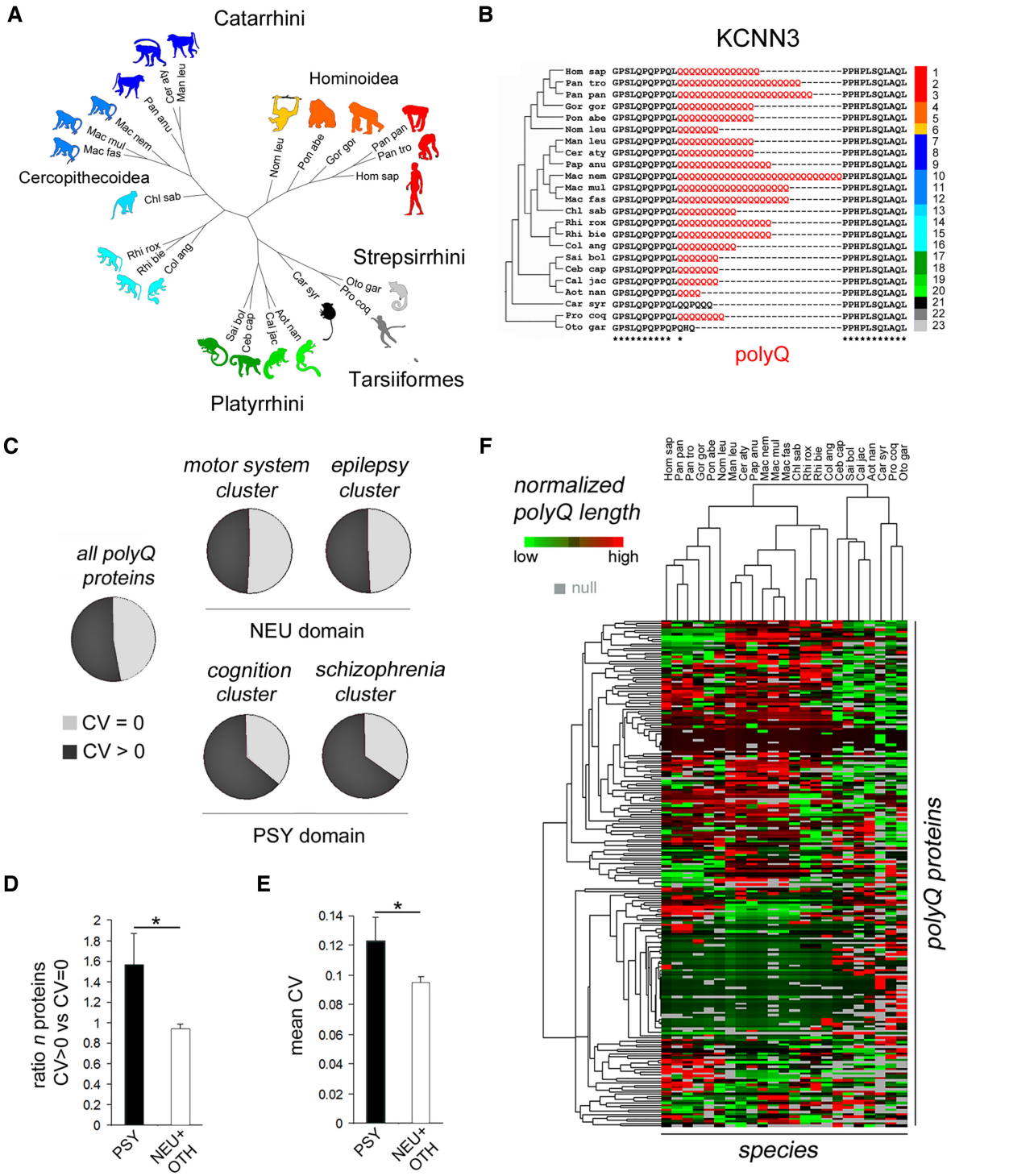

Figure 2. Evolutionary history of polyQ repeats in primates and their hypervariability in clusters of the PSY domain. (**A**) Unrooted, unscaled phylogenetic tree of primate species whose polyQ protein lengths were analyzed in this study. Species name abbreviations are reported in the 'Materials and Methods' section. The tree covers all the major taxonomic groups of Primates, i.e. Catarrhini, Platyrrhini, Tarsiiformes and Strepsirrhini. Lower level taxa are color-coded as indicated in Supplementary Figure S2A. (**B**) Partial alignment of a region of the primary sequence of the primate KCNN3 orthologs bearing a variable polyQ stretch. Phylogenetic relationships are reproduced in the tree *on the left side*, taxa are color-coded *on the right side* as indicated in Supplementary Figure S2A. (**C**) Pie charts indicating the relative proportion of human polyQ proteins whose primate orthologs display either constant total polyQ length ($CV_{rl} = 0$) or variable total polyQ length ($CV_{rl} > 0$), either in the entire set of polyQ proteins (*left*) or in protein sets associated with the four indicated clusters. Note the higher proportion of proteins with variable polyQ lengths associated with the two clusters belonging to the PSY domain. (**D**) Bar graph displaying the mean ratio between the number of proteins with variable ($CV_{rl} > 0$) and the number of proteins non-variable ($CV_{rl} = 0$) total polyQ lengths in clusters of the PSY domain *versus* clusters of the other two domains (NEU + OTH). (**E**) Bar graph displaying the mean CV_{rl} value in protein sets related to clusters of the PSY domain *versus* clusters of the other two domains (NEU + OTH). (**F**) Heat map of the total repeat length variation in polyQ proteins (*rows*) in the indicated 23 primate species (*columns*).

The second parameter is the mean CV_{rl} in each cluster. A one-way ANOVA revealed an overall significant difference between the clusters ($F_{(20,1178)} = 1.60$, $P < 0.05$, $n = 21$). Moreover, 6 of the 10 clusters with the highest mean CV_{rl} belonged to the PSY domain (i.e. *dementia* (dem), *schizophrenia* (sch), *cognition* (cog), *personality disorder*s (prs), *language* (lan), *behavior* (beh); Supplementary Figure S2D). Again, the mean CV_{rl} in the PSY domain clusters was significantly higher than in the non-PSY domain clusters (0.12 ± 0.01 *versus* 0.09 ± 0.003, $n = 8$ *versus* 13 clusters, respectively, $P < 0.05$ *t*-test; Figure 2E).

Taken together, these findings indicate that polyQ repeat lengths varied differentially across functional clusters during primate evolution, and that clusters of the PSY domain display polyQ length hypervariability in comparison with those of the NEU and OTH domains.

Extensive patterns of polyQ length co-evolution in primates

Finally, we integrated the results of our functional and evolutionary analyses of polyQ proteins with available interactomics datasets, to identify possible co-evolution patterns of polyQ repeats in pairs and networks of primate proteins.

To obtain an overall picture of the polyQ length co-variation in primates, we performed a cluster analysis of polyQ repeat lengths in protein orthologs across 23 primate proteomes (Figure 2F). Interestingly, this analysis uncovered extensive patterns of polyQ length co-variation in multiple protein sets. Notably, these co-variation patterns appeared to be complex and not simply related to clock-like trends of repeats expansion or contraction across taxa. Moreover, the same analysis clustered the 23 species according to their known phylogenetic relationships, with a minor deviation from the standard phylogeny, as shown by the species clustergram represented as an unrooted tree (Supplementary Figure S2E). This indicates that not only the evolutionary occurrence of polyQ repeats (5), but also their length variation carries considerable phylogenetic signal.

PolyQ length co-evolution in pairs of functionally related and physically interacting proteins

Based on the previous findings, we further explored the possible significance of the observed co-evolution patterns of polyQ repeats in relation to the function and interactions of their parent proteins.

Toward this aim, we first asked whether the observed patterns of polyQ length co-variation may be associated with the coordinated evolution of functionally related proteins. To address this question quantitatively, we sought to determine whether polyQ repeats in clusters of functionally related proteins, such as those we identified, display higher degrees of co-variation than those in randomly selected polyQ protein sets.

For this analysis, we selected polyQ proteins with at least 10 available orthologs and polyQ $CV_{rl} > 0$ (n = 225; Supplementary Table S7; see 'Materials and Methods' section). Then, we calculated the Pearson's *r* correlation coefficients between the total polyQ lengths of each pairwise combination of these proteins across species, thus obtaining an overall correlation matrix.

An initial analysis of this matrix highlighted numerous protein pairs in which polyQ length varied with considerable degrees of correlation (Figures 3–4). For instance, we found that the polyQ lengths of KMT2D and PAXIP1 display a significant level of positive correlation ($r = 0.73$, $n = 21$; $P < 0.001$; Figure 3A and B, *left panel*), which was even higher when averaging the values of related species belonging to the same subtaxa ($r = 0.98$, $n = 9$; $P < 0.001$; Figure 3B, *middle* and *right panels*). Notably, KMT2D and PAXIP1 are closely related functionally, as they are part of the same functional complex, interacting physically (47), and mutations in either one of the two proteins can cause one same disease, i.e. Kabuki syndrome (48). Interestingly, heptad-spaced hydrophobic residues (mostly leucine) were interspersed in the polyQ repeats of both proteins, consistent with the formation of stabilized polyQ CC interfaces (1,10).

We also observed cases of negative correlation of polyQ lengths, as for TBP and NCOA6 ($r = -0.61$, $n = 22$, $P < 0.01$; Figure 4A and B, *left panel*), by which polyQ elongation in one protein is accompanied by polyQ contraction in the other. The correlation was even stronger when averaging values of closely related species ($r = -0.71$, $n = 9$, $P < 0.04$; Figure 4B, *middle* and *right* panels). As for KMT2D and PAXIP1, the two proteins are known to interact physically (49), indicating functional relatedness.

These findings strongly suggest that the length of polyQ repeats in protein pairs, as CC-based protein interaction surfaces, may co-evolve to vary the strength and stability of the interaction between the two proteins.

Co-evolution of polyQ repeats in clusters of functionally related proteins

To compare systematically the overall co-variation of polyQ lengths between protein sets, we performed a Fisher's Z transformation of the *r* coefficients (39; see 'Materials and Methods' section). We visualized this *Z*-score matrix as an overall 'correlation network' (Figure 5A, *middle panel*) in which *nodes* represent polyQ proteins, *edge thickness* and *color*, respectively, the degree (*Z*-score) and sign (*red*, positive; *blue*, negative) of the polyQ length correlation.

Using this overall network, we compared the distribution and mean value of *Z*-scores in subnetworks of polyQ proteins belonging to one same functional cluster (e.g. *language*) *versus* control subnetworks of equinumerous, randomly selected polyQ proteins (Figure 5A).

In analyzing the distribution of *Z*-scores, we found a significant difference between the percentage of protein pairs with higher degrees of polyQ length correlation (absolute $Z \geq 0.5$), in cluster-related *versus* random subnetworks (21.80 ± 0.60 *versus* $16.22 \pm 0.31\%$, respectively, $n = 15$ in both groups, $P < 0.0001$, *t*-test; Figure 5B and C; Supplementary Figure S3E). A further control analysis revealed instead no significant difference when we compared two sets of randomly selected polyQ proteins (16.25 ± 0.26 *versus* $16.22 \pm 0.31\%$, respectively, $n = 15$ in both groups, $P = 0.95$, *t*-test; Figure 5D and E; Supplementary Figure S3F).

Similar results were obtained when we compared the mean *Z*-score value in cluster-related *versus* random control subnetworks (33.22 ± 0.63 *versus* 28.17 ± 0.21, respectively,

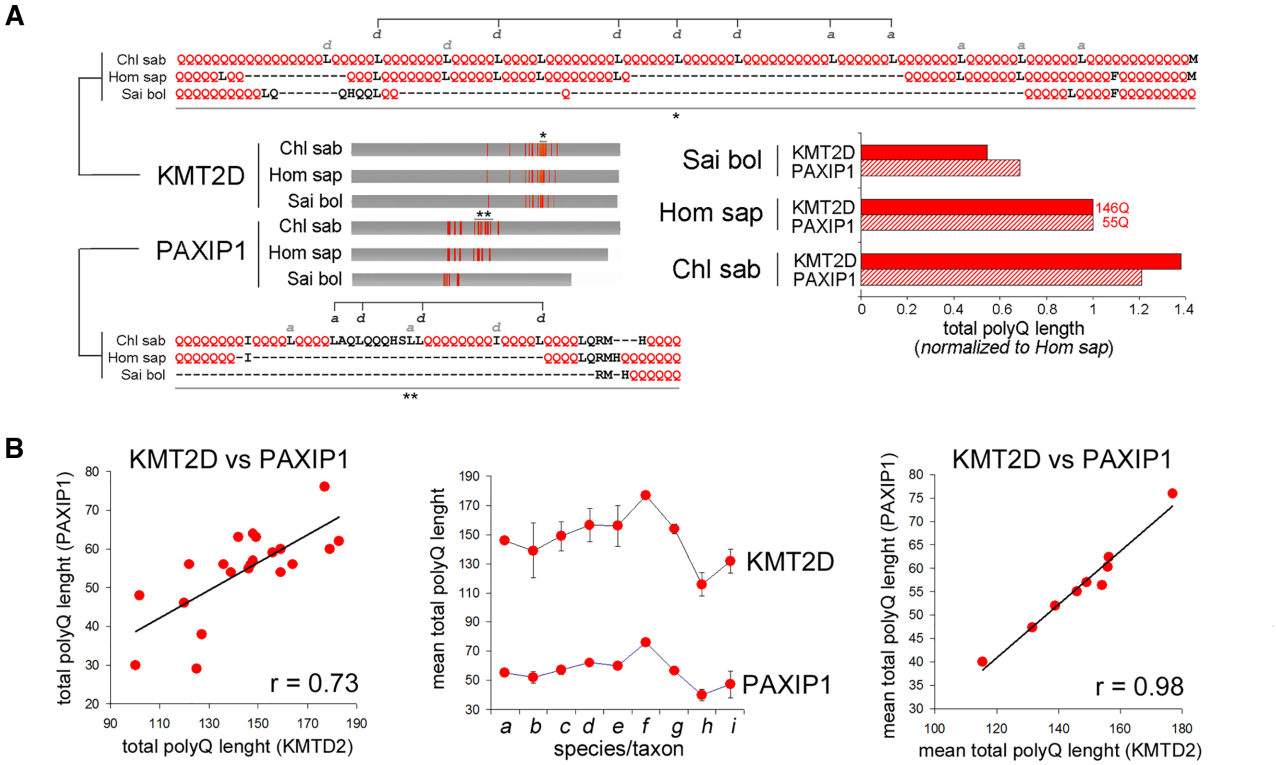

Figure 3. Co-evolution of polyQ repeats in pairs of functionally related, physically interacting proteins: positive correlation of polyQ lengths. (**A**) *Gray bars* schematize the orthologs of KMT2D and PAXIP1 in *Saimiri*, *Homo* and *Chlorocebus* with their polyQ repeats represented as *red vertical bars*. The length of the orthologs of each protein is normalized to the length of the *Saimiri* ortholog. Alignments of the main polyQ regions of the two proteins are shown *above* and *below* the gray bars, respectively. PolyQ stretches (≥4Q) are highlighted in *red*. *Black letters* (*a,d*) above the alignments highlight leucine/isoleucine residues, interspersed among polyQ repeats, and spaced consistent with a coiled coil heptad register in positions *a*/*d*. Other leucine/isoleucine residues follow a similar periodicity but shifted with respect to the main register, suggesting the presence of two hydrophobic interfaces (1). The histograms *on the right* show the relative total polyQ lengths in the three species, with values normalized to those in *Homo* (146Q in KMT2D and 55Q in PAXIP1). Note the parallel variation of the total polyQ lengths in the two proteins across species. (**B**) *Left graph:* scatterplot displaying the positive correlation of the total polyQ lengths in KMT2D and PAXIP1 in 21 ortholog pairs. *Middle graph:* mean values of the total polyQ lengths of KMT2D and PAXIP1 in representative species/taxa: *a*, Hominina (Hom sap); *b*, other Homininae (Pan tro, Pan pan, Gor gor); *c*, Ponginae/Hylobatidae (Pon abe, Nom leu); *d*, Papionini (Man leu, Cer aty, Pap anu); *e*, Macaca (Mac nem, Mac mul, Mac fas); *f*, Cercopithecini (Chl sab); *g*, Colobinae (Rhi rox, Rhi bie, Col ang); *h*, Ceboidea (Sai bol, Ceb cap, Cal jac, Aot nan); *i*, Tarsiiformes/Strepsirrhini (Car syr, Pro coq, Oto gar). Note the parallel trends of the mean polyQ repeat lengths in the two proteins across taxa. *Right graph:* scatterplot displaying the positive correlation of the mean total polyQ lengths in KMT2D and PAXIP1 in the nine species/taxa as defined in the *middle panel*.

$n = 15$ in both groups, $P < 0.0001$, *t*-test; Supplementary Figure S3A, B and G). Again, this difference was not found when comparing two random sets of polyQ proteins (28.51 ± 2.25 *versus* 28.14 ± 2.15, $n = 15$ in both groups, $P = 0.78$, *t*-test; Supplementary Figure S3C, D and H).

These findings show that the lengths of polyQ repeats within clusters of functionally related proteins co-varied during primate evolution significantly exceeding random expectations, supporting the notion that such co-variation is function-related and not simply the result of the random accumulation of indel mutations in unrelated genes.

Annotation-based clustering of polyQ proteins captures interaction networks

The observed patterns of polyQ co-variation in functionally related proteins may ultimately regulate their interactomes by finely tuning the strength of their mutual interactions. If this is the case, one may hypothesize, first, that the proteins within the functional polyQ protein clusters should display higher interactivity than randomly selected polyQ proteins, and, second, that significant polyQ length co-variation should also be present in clusters of proteins selected based on their known interactions rather than their function.

To test the first hypothesis, we derived from Biogrid (50) all the known physical interactions between polyQ proteins. From this overall polyQ interaction network (Figure 6A, *middle panel*), we extracted subnetworks of proteins belonging to the same functional clusters as well as equinumerous control subnetworks of randomly selected polyQ proteins. We quantified the properties of these subnetworks and found that the cluster-derived ones displayed significantly higher internal connectivity in comparison with the control subnetworks, as shown by multiple parameters (Figure 6B; $P < 0.03$ in all instances, *t*-test, $n = 17$ in all groups). Specifically, we detected higher *average number of neighbors*, *network density*, and *clustering coefficient* and lower numbers of *isolated nodes* and *connected components* in the cluster-related subnetworks. The number of nodes in the cluster-related and control subnetworks were not different, as expected ($P = 0.74$ *t*-test, $n = 17$ in both groups; Figure 6B).

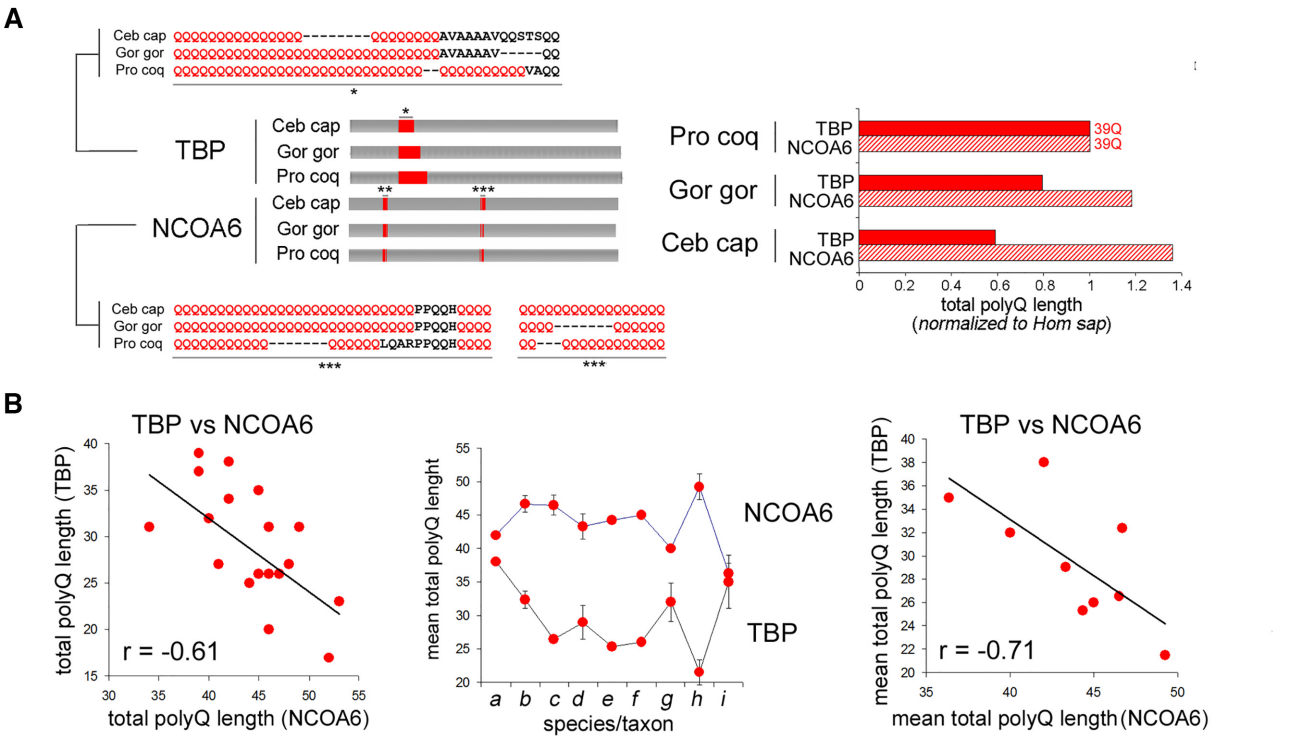

Figure 4. Co-evolution of polyQ repeats in pairs of functionally related, physically interacting proteins: negative correlation of polyQ lengths. (**A**) Graphs and data as in Figure 3A but for the polyQ protein pair TBP-NCOA6 in *Gorilla*, *Cebus* and *Propithechus*. Values in the bar graph are normalized to those in *Propithechus*. (**B**) Scatterplots and graph as in Figure 3B but for the polyQ protein pair TBP-NCOA6. Note how the total polyQ lengths in the two proteins are inversely correlated, so that the increase in the total polyQ length in one protein is accompanied by polyQ length contraction in the other.

These results confirm the first hypothesis and support the notion that the functional meaning of polyQ co-variation may be related to the modulation of interactome properties.

Co-evolution of polyQ repeat length in clusters of physically interacting proteins

Finally, we tested the second hypothesis, i.e. whether sets of polyQ proteins selected only based on their known physical interactions show significant polyQ length co-evolution. Thus, we extracted 100 subnetworks of interacting polyQ proteins from the total polyQ correlation network (Figure 7A, *middle panel*), each formed by a different core of 10 arbitrarily selected polyQ proteins and their known direct interactors found in Biogrid (Figure 7A, *left panel*; see 'Materials and Methods' section). For each one of these interactome-derived subnetworks, we extracted an equinumerous control subnetwork formed by the same core of 10 polyQ proteins and other randomly selected polyQ proteins (Figure 7A, *right panel*).

For each interactome-related subnetwork (i) and its paired control subnetwork (r), we calculated the same two indexes of polyQ co-variation as in the previous analysis, i.e. the proportion (%) of protein pairs with higher correlation (absolute $Z \geq 0.5$), and the mean absolute Z. For each i-r subnetwork pair we calculated the difference (Δ) for each parameter between the two paired subnetworks (Figure 7B–D). As a further control, we performed the same analysis by comparing pairs of random control subnetworks ($r1$ and $r2$; Figure 7C–E).

These analyses revealed that interactome-derived subnetworks had a higher proportion of protein pairs with absolute $Z \geq 0.5$ than their control subnetworks ($i > r$) in ~2/3 of the instances (Figure 7B), which significantly exceeded random expectations (i.e. 1/2 of the instances; 65/35 observed *versus* 50/50 expected, $P < 0.05$, Fisher's exact (FE) test). The two subsets of random subnetworks instead did not differ in this respect (Figure 7C, 52/48 observed *versus* 50/50 expected; $P = 0.88$, FE test). Overall, the proportion of protein pairs with absolute $Z \geq 0.5$ in the interactome-derived subnetworks was significantly higher than in random controls (17.93 ± 0.36 *versus* 16.31 ± 0.36, $n = 100$, $P < 0.01$, paired t-test; Supplementary Figure S4A), while it did not differ when comparing two sets of random subnetworks (15.72 ± 0.35 *versus* 16.31 ± 0.36, $n = 100$, $P = 0.28$, paired t-test; Supplementary Figure S4B).

Similarly, interactome-derived subnetworks had a higher mean absolute Z than their control subnetworks ($i > r$) in ~2/3 of the instances (Figure 7D), which significantly exceeded random expectations (66/34 observed *versus* 50/50 expected, $P < 0.05$, FE test), while the two subsets of random subnetworks did not differ in this respect (Figure 7E, 50/50 observed *versus* 50/50 expected, $P = 1$, FE test). Also the mean absolute Z across all the interactome-derived subnetworks was significantly higher than in the random controls (29.75 ± 0.26 *versus* 28.29 ± 0.30, $n = 100$, $P < 0.001$, paired t-test; Supplementary Figure S4C) while it did not differ again when comparing two sets of random subnetworks (28.09 ± 0.28 *versus* 28.29 ± 0.30, $n = 100$, $P = 0.64$, paired t-test; Supplementary Figure S4D).

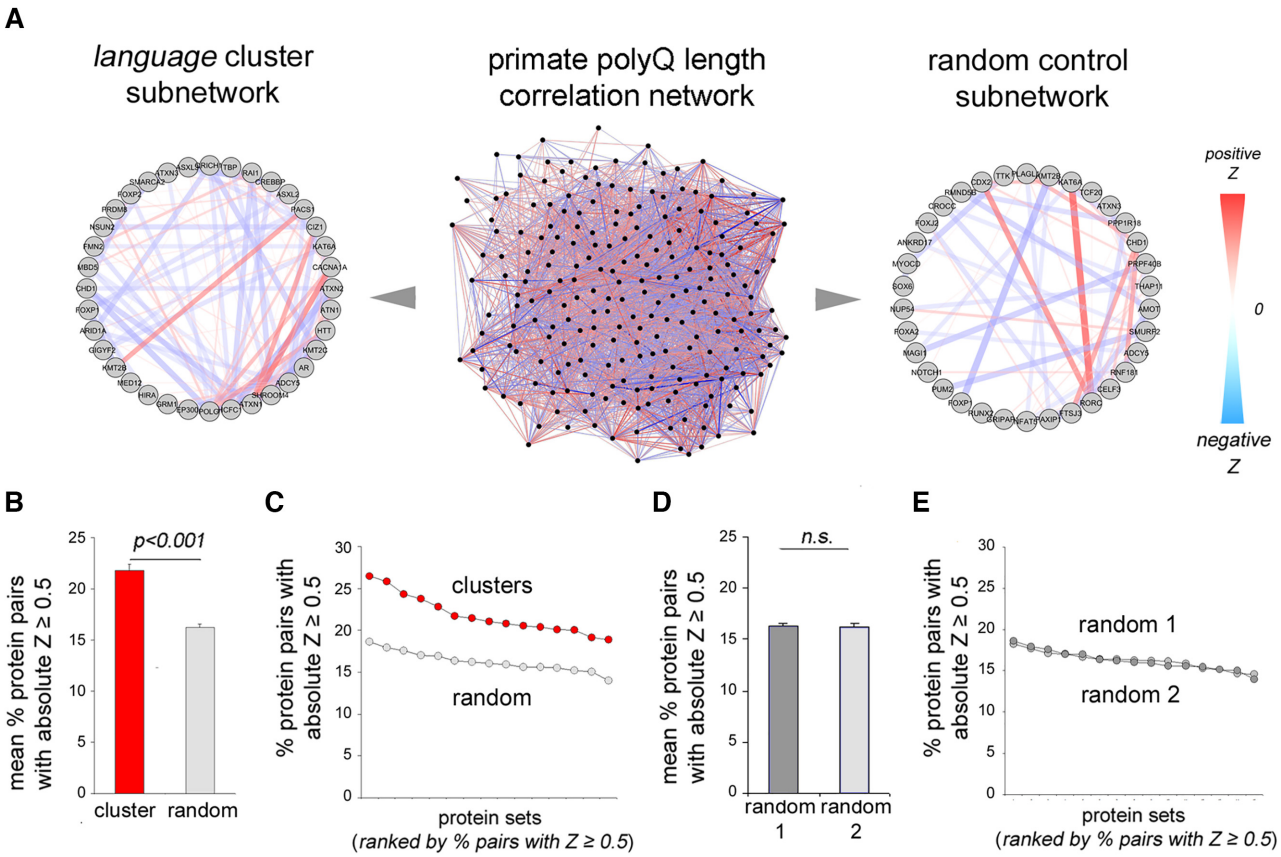

Figure 5. Co-evolution of polyQ repeat lengths in clusters of functionally related proteins. (**A**) The *central graph* represents, in the form of a network, the overall evolutionary co-variation patterns of polyQ repeat lengths in all the pairwise combinations of polyQ proteins. *Nodes* represent polyQ proteins and *edges* represent correlations whose sign is coded by color (positive in *red*, and negative in *blue*) and whose magnitude is coded by edge thickness (legend on the *right side*). From this overall primate polyQ length 'correlation network' we extracted subnetworks of polyQ proteins belonging to functional clusters (e.g. *language*, *left* subnetwork) or equinumerous control subnetworks of randomly selected polyQ proteins (e.g. *right* subnetwork). (**B**) Bar graph plotting the mean percentage of polyQ protein pairs with an absolute Z-score ≥ 0.5 (i.e. with relatively higher levels of positive or negative correlation) in subnetworks of functionally related polyQ proteins ('cluster') or in equinumerous control subnetworks of randomly selected polyQ proteins ('random'). (**C**) Percentage of polyQ protein pairs with an absolute Z-score ≥ 0.5 in individual subnetworks of functionally related polyQ proteins ('clusters') or in equinumerous random control subnetworks ('random'), as in (B). In each series, values are ranked from the highest to the lowest. Only clusters with more than 20 polyQ proteins (i.e. 15/49) were included in this analysis. **(D** and **E)** As in (B and C), but comparing two sets of random subnetworks ('random 1' and 'random 2').

In all these analyses, as expected, the subnetwork sets were equinumerous (Supplementary Figure S4E) and the mean number of protein pairs with defined Z-values was not different (Supplementary Figure S4F), ($P > 0.05$ in all instances, t-test).

These results show that polyQ co-variation in sets of interacting proteins significantly exceeds random expectations, further supporting the notion that polyQ co-evolution is related to the functional regulation of interactomes.

DISCUSSION

The results of our functional and evolutionary analyses reveal polyQ length co-evolution in pairs and clusters of functionally related, physically interacting neural proteins in Primates. Overall, this novel evidence identifies polyQ repeats as a system of co-evolving sequences in protein networks. These findings have direct implications for understanding the regulatory impact of polyQ repeats on neural function and dysfunction.

A higher-order functional and nosological mapping of polyQ proteins in the nervous system and allied structures

We developed a novel functional and nosological classification of the >450 human polyQ proteins through a cross-database integration of gene functional/phenotypic annotations and disease associations, with an *a priori* semantic clustering of terms and pooling of the associated genes (e.g. 5,51). Such analytical approach can overcome inherent constraints of conventional single-database enrichment analyses (i.e. intrinsic domain limitation, annotation biases, and small number of genes with annotations of interest), often requiring *a posteriori* clustering of complex lists of heterogeneous terms (51,52,53,54).

Previous functional classifications of polyQ proteins were focused mostly on the cellular and molecular levels (e.g.

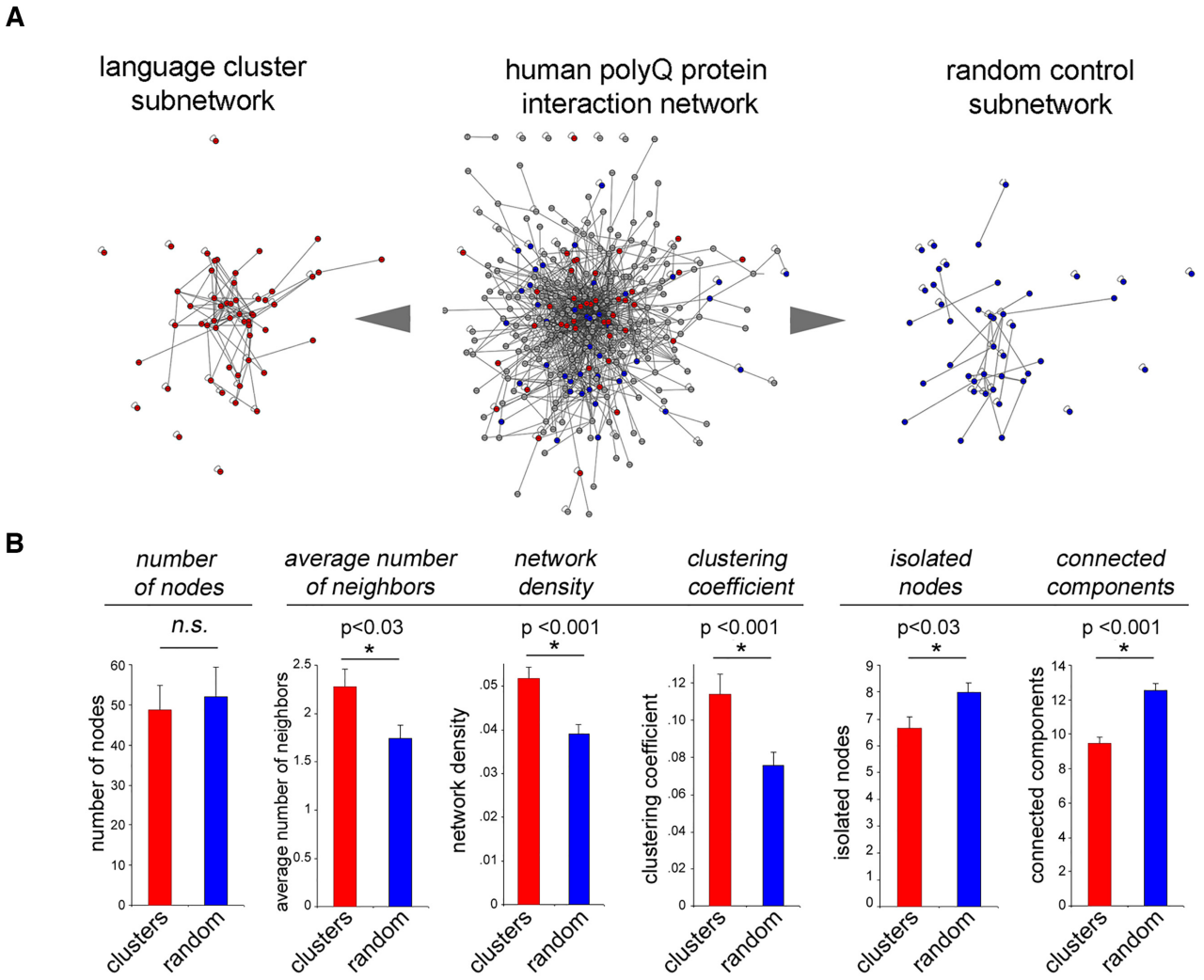

Figure 6. Interactomes of functionally related polyQ proteins (**A**) Schematic representation of the interactome analyses. *Middle graph*: PPI network of polyQ proteins derived from BioGrid. *Nodes* represent proteins and *edges* represent known protein interactions. *Left graph*: subnetwork of polyQ proteins belonging to the *language* cluster. *Right graph*: equinumerous control subnetwork formed by randomly selected polyQ proteins. (**B**) Mean values of the indicated descriptive parameters in the subnetworks of functionally related proteins ('clusters', *red* bars; $n = 15$) and in the set of equinumerous control subnetworks ('random', *blue* bars; $n = 15$). Only clusters with more than 20 polyQ proteins (i.e. 15/49) were included in this analysis.

4,55,56,57), typically through GO-term enrichment approaches, while the roles of these proteins in higher-order pathophysiological processes of the nervous system (e.g. cognition, behavior and neuropsychiatric disorders) were scarcely defined. The classification that we developed, by means of cross-database integration and *a priori* clustering of functional terms/disease associations, offers new perspectives on the higher-level functional and (dys)functional roles of polyQ proteins in the nervous system.

Function-related polyQ enrichment in neural protein clusters

Our analyses revealed that polyQ repeats are specifically overrepresented only in certain protein clusters, in both the neurobiology/neurology (NEU) and neuropsychology/psychiatry (PSY) domains. This specificity is in agreement with previous results obtained in the analysis of AAR enrichments in developmental proteins (5), supporting the notion of a relative functional/regional specialization of AAR-containing protein networks (5,25).

Notably, some of the most polyQ-enriched clusters belong to the PSY domain, such as those linked to high-level cognitive and behavioral functions (e.g. *language* and *behavior*) and related disorders (e.g. *personality disorders* and *autism spectrum disorders*). At the level of individual genes, a few examples exist of how the length variation of coding or non-coding repeats, such as those in the androgen and vasopressin 1a receptors, can regulate psychological, social and complex behavioral traits (58,59,60,61,62,63). Thus, our findings indicate that polyQ repeats may have overall a complex pathophysiological impact on cognitive and behavioral traits, as a whole system of functional regulators. This conclusion is consistent with the hypothesis that different kinds of genetic dynamics, including repeat length variation, may have behavioral consequences (64).

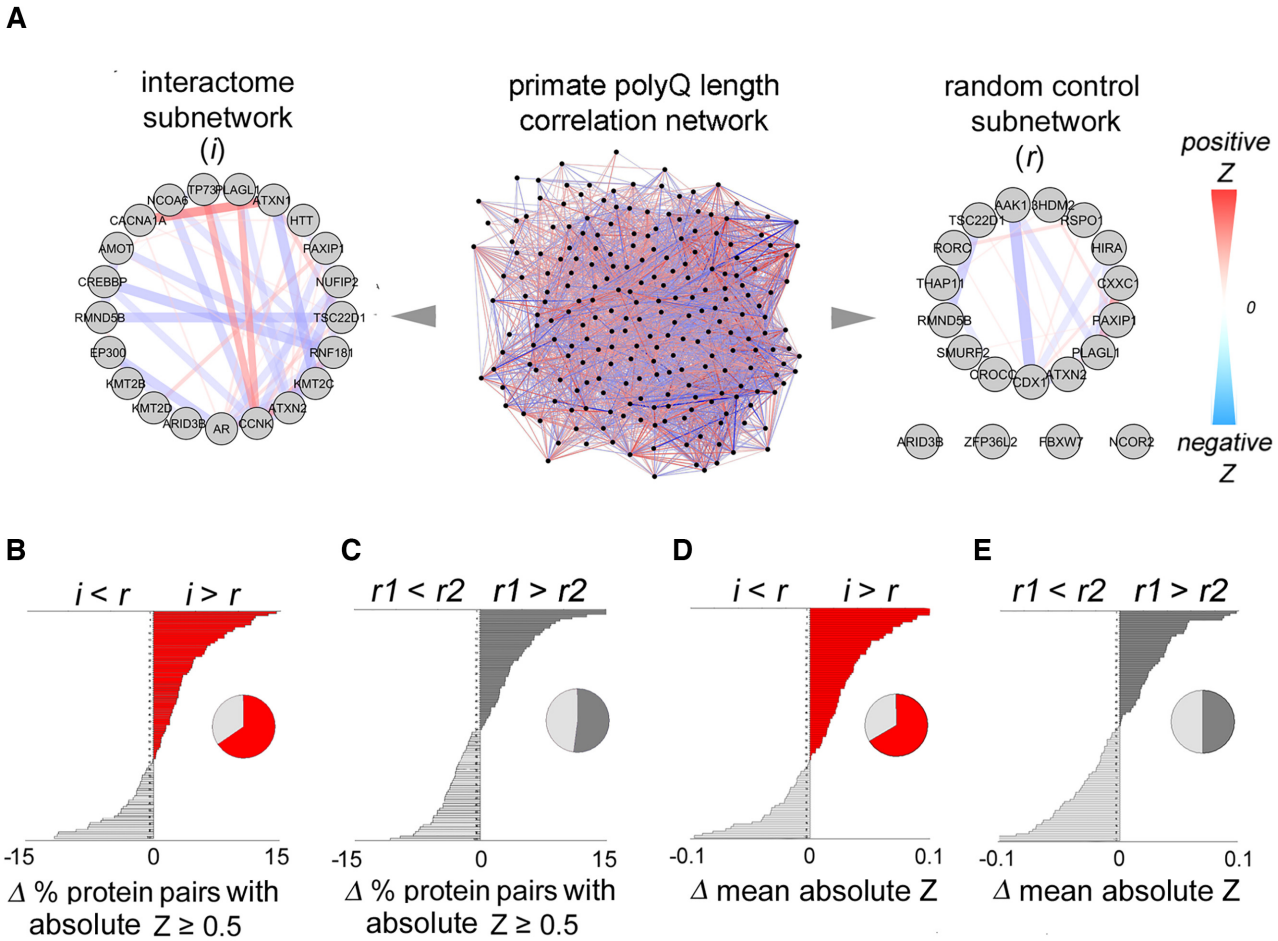

Figure 7. Co-evolution of polyQ repeats in sets of physically interacting proteins. (**A**) Schematic representation of polyQ length co-variation analyses in sets of interacting proteins. The *central network* represents the degree of evolutionary co-variation of polyQ lengths in all the pairwise combinations of polyQ proteins, as in Figure 5A. From this network, we extracted subnetworks formed by a set of 10 arbitrarily selected polyQ proteins and their *n* known interactors (*left,* 'interactome' subnetwork) or equinumerous control subnetworks formed by the same set of 10 polyQ proteins and other *n* proteins selected randomly (see 'Materials and Methods' section). (**B**) A total of 100 pairs of *interactome versus control* subnetworks were generated, as in panel (A). In all these 200 subnetworks, we calculated the percentage of protein pairs with an absolute *Z*-score ≥ 0.5, as an index of the overall degree of correlation in the subnetworks. For each one of the 100 *interactome versus control* subnetwork pairs, we calculated the difference (Δ) between these percentages in the two subnetworks (*i* and *r*). The bar graph plots these 100 Δ values ranked from positive to negative ones. Note that in about two thirds of the pairs (as also indicated in the *pie chart* inset) the *interactome* subnetwork had a higher percentage of protein pairs with an absolute *Z*-score ≥ 0.5 than the matched control subnetwork (*i*>*r*), indicating overall a higher level of polyQ length co-variation. (**C**) As in panel (B), for two sets of 100 random subnetworks each (*r1 versus r2*). In this case, *r1* subnetworks had a higher percentage of protein pairs with a *Z*-score ≥ 0.5 (*r1*>*r2*) only in about half of the *r1-r2* subnetwork pairs (48%), as expected in a random distribution. (**D** and **E**) As in panels (B and C), for the difference (Δ) in mean absolute *Z*-score in the indicated subnetwork pairs, another index of the overall degree of correlation in the subnetworks. Again, the mean absolute *Z*-value is greater in the *interactome* (*i*) subnetworks in two thirds of the *i-r* subnetwork pairs (*panel D*), while the mean *Z*-values in pairs of random networks (*r1* versus *r2*) were as expected in a random distribution (i.e. *r1*>*r2* only in 50% of the *r1-r2* subnetwork pairs; *panel E*).

In the NEU domain, the clusters with the highest polyQ enrichment were *eye,* in the neurobiological/neuroanatomical subdomain, and *epilepsy* and *motor system* in the neurological subdomain. This latter observation is in good correlation with the fact that polyQ expansions cause Huntington disease and related movement disorders (13) and that epilepsy is often associated with these diseases (65). Importantly, our findings can rationalize clinical observations showing that the length of certain polyQ repeats modulates the risk to develop non-polyQ-related motor disorders, such ALS and Parkinson disease (66,67), thus supporting the notion of a generalized involvement of polyQ repeats in motor regulation. Interestingly, polyQ repeats were also particularly enriched in eye-related proteins. Thus, the known roles of ATXN3 and ATXN7 in the differentiation and degeneration of photoreceptors (68,69,70) may represent an instance of a more generalized involvement of polyQ proteins in eye development and physiology.

We found that polyQ repeats display a parallel enrichment in protein clusters related both to the nervous system and to anatomically (skull) or embryologically (skin/adnexa) allied structures. Recent observations show that proteins involved in brain development and physiology also have important regulatory roles in craniofacial and epidermal/adnexal phenotypes (71,72,73), and many

developmental neuropsychiatric syndromes display cranial or cutaneous/adnexal abnormalities that may have a common genetic basis (e.g. 35). Moreover, our findings provide a more general framework for putting into perspective punctual observations on individual genes, such as RUNX2, in which polyQ and polyalanine (polyA) variation is known to regulate molecular function and thereby cranial morphology (10,11,74). Our results, therefore, indicate that polyQ protein networks may have relevant regulatory roles in the coordinated development/function of cephalic and neuroectodermal structures. Strikingly, a similar enrichment of polyQ repeats has been observed in proteins regulating head morphology in the insect *Teleopsis dalmanni* (75), suggesting that the involvement of polyQ protein networks in the cephalic compartment may be phylogenetically very ancient. This also suggests that polyQ length co-evolution may occur in other taxa besides Primates, consistent with the pervasive presence of complex AAR dynamics in eukaryotic proteomes, from yeast to humans (5,10).

Protein function-related polyQ length variation in primate evolution

The *de novo* occurrence of AARs and their subsequent length variation are thought to contribute to phenotypic evolution (e.g. 5,8,9). While on large evolutionary scales the occurrence rate of polyQ repeats in proteomes varies considerably (5), we found that, within Primates, the proteome-wide occurrence of polyQ proteins is substantially stable. However, we found instead considerable levels of repeat length variation in the orthologs of more than half of the human polyQ proteins. Such widespread variability indicates that polyQ length-dependent regulatory effects (e.g. 11,76) may have generalized roles in primate evolution. Notably, the overall polyQ length variation observed in Primates carries a strong phylogenetic signal, sufficient to reconstruct their phylogeny in fine detail. This observation is consistent the notion that polyQ variation is subject to some degree of evolutionary pressure (e.g. 77), as entirely random intra-/inter-specific patterns of length variation, that one could expect from unstable CAG microsatellite repeats as such, would arguably occlude the phylogenetic signal, especially in closely related species.

Strikingly, polyQ length varies differentially in functionally distinct protein clusters. Notably, polyQ length variation is overall significantly higher in protein clusters of the PSY domain, which are also among those most enriched in polyQ repeats. Such parallel enrichment and hypervariability of polyQ repeats in these same protein clusters concurrently indicate that the evolution of neuropsychological phenotypes may have been particularly subject to polyQ length-dependent regulation in primates.

Together with observations that behavioral traits evolve faster than morphological ones (78,79,80,81), and that hypervariable polyQ repeats are found in rapidly evolving proteins (77,82), our findings indicate that polyQ length variation in proteins of the PSY domain may have contributed to the rapid cognitive and behavioral evolution of primates (83,84).

Co-evolution of polyQ repeats in functionally related proteins

Our analyses uncovered significant levels of polyQ length co-variation during primate evolution. To our knowledge, these findings represent the first report of intermolecular co-evolution related to the length of AARs. These observations expand the repertoire of molecular co-evolution beyond SNP-based mechanisms, which lead to single amino acid substitutions, to a different class of mutations (indels) and amino acid changes (homopolymer length variation). More broadly, as suggested by the observation that AARs in general display highly complex and interrelated evolutionary dynamics (5,10), these findings may also apply to AARs other than polyQ (S. Vaglietti and F. Fiumara, unpublished observations). The results of our analyses fit well with the view of molecular co-evolution as a driver of molecular co-adaptation, interactivity, and, ultimately, of organismal fitness (17,18,19,20,21,22,23,24). Indeed, we found, first, that functionally related polyQ proteins, which display polyQ-length co-variation, also have higher degrees of physical interactivity. Second, we observed that polyQ protein sets selected only on the basis on their known physical interactions also exhibit significant levels of polyQ length co-evolution. These findings together strongly suggest that the observed evolutionary polyQ co-variation is functional in regulating protein interactivity in complex networks.

In mechanistic terms, polyQ co-evolution translates directly into the coordinated variation of the length and stability of CC structures, i.e. protein interaction modules, in functionally related protein pairs/complexes (1,4,25,26,27). In this respect, it is interesting to note that CC structures can undergo molecular co-evolution (e.g. 85,86). Moreover, CCs are also physiologically involved in the formation of supramolecular assemblies, through liquid–liquid phase separation and other structural dynamics, underlying prion-like functional switches (e.g. 1,16,87). Thus, the evolutionary length variation of CC-embedded polyQ repeats may have profound functional consequences by modulating the formation of homo-/hetero-typic molecular assemblies.

In conclusion, our findings reveal polyQ length co-evolution in functionally related, physically interacting protein pairs and networks, supporting the notion that polyQ repeats are functional sequences that may have played—through coordinated length variation—a key role in primate neural evolution.

DATA AVAILABILITY

The published article includes as supplemental tables all the datasets generated or analyzed during this study. The Perl scripts that were generated are available at GitHub (https://github.com/ffiumara/SV-FF-NARgab-2021).

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

We thank Davide Corà, Mirella Ghirardi and Maurizio Giustetto for helpful discussions and critical reading of the manuscript.

FUNDING

Ministero dell'Istruzione, dell'Università e della Ricerca (MIUR) [FFABR-2017 to F.F.].
Conflict of interest statement. None declared.

REFERENCES

1. Fiumara,F., Fioriti,L., Kandel,E.R. and Hendrickson,W.A. (2010) Essential role of coiled coils for aggregation and activity of Q/N-rich prions and PolyQ proteins. *Cell*, **143**, 1121–1135.
2. Gemayel,R., Vinces,M.D., Legendre,M. and Verstrepen,K.J. (2010) Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annu. Rev. Genet.*, **44**, 445–477.
3. Gemayel,R., Chavali,S., Pougach,K., Legendre,M., Zhu,B., Boeynaems,S., van der Zande,E., Gevaert,K., Rousseau,F., Schymkowitz,J. *et al.* (2015) Variable glutamine-rich repeats modulate transcription factor activity. *Mol. Cell.*, **59**, 615–627.
4. Schaefer,M.H., Wanker,E.E. and Andrade-Navarro,M.A. (2012) Evolution and function of CAG/polyglutamine repeats in protein–protein interaction networks. *Nucleic Acids Res.*, **40**, 4273–4287.
5. Pelassa,I., Cibelli,M., Villeri,V., Lilliu,E., Vaglietti,S., Olocco,F., Ghirardi,M., Montarolo,P.G., Corà,D. and Fiumara,F. (2019) Compound dynamics and combinatorial patterns of amino acid repeats encode a system of evolutionary and developmental markers. *Genome Biol. Evol.*, **11**, 3159–3178.
6. Chavali,S., Chavali,P.L, Chalancon,G., de Groot,N.S., Gemayel,R., Latysheva,N.S., Ing-Simmons,E., Verstrepen,K.J., Balaji,S. and Babu,M.M. (2017) Constraints and consequences of the emergence of amino acid repeats in eukaryotic proteins. *Nat. Struct. Mol. Biol.*, **24**, 765–777.
7. Chavali,S., Singh,A.K., Santhanam,B. and Babu,M.M. (2020) Amino acid homorepeats in proteins. *Nat.Rev. Chem.*, **4**, 420–434.
8. Dover,G.A. (1989) Slips strings and species. *Trends Genet.*, **5**, 100–102.
9. Fondon,J.W. 3rd and Garner,H.R. (2004) Molecular origins of rapid and continuous morphological evolution. *PNAS*, **101**, 18058–18063.
10. Pelassa,I., Cora,D., Cesano,F., Monje,F.J., Montarolo,P.G. and Fiumara,F. (2014) Association of polyalanine and polyglutamine coiled coils mediates expansion disease-related protein aggregation and dysfunction. *Hum. Mol. Genet.*, **23**, 3402–3420.
11. Ritzman,T.B., Banovich,N., Buss,K.P., Guida,J., Rubel,M.A., Pinney,J., Khang,B., Ravosa,M.J. and Stone,A.C. (2017) Facing the facts: the Runx2 gene is associated with variation in facial morphology in primates. *J. Hum. Evol.*, **111**, 139–151.
12. Karlin,S. and Burge,C. (1996) Trinucleotide repeats and long homopeptides in genes and proteins associated with nervous system disease and development. *Proc. Natl Acad. Sci. U.S.A.*, **93**, 1560–1565.
13. Almeida,B., Fernandes,S., Abreu,I.A. and Macedo-Ribeiro,S. (2013) Trinucleotide repeats: a structural perspective. *Front. Neurol.*, **4**, 76.
14. Silva,A., de Almeida,A.V. and Macedo-Ribeiro,S. (2018) Polyglutamine expansion diseases: more than simple repeats. *J. Struct. Biol.*, **201**, 139–154.
15. Gerber,H.P., Seipel,K., Georgiev,O., Höfferer,M., Hug,M., Rusconi,S. and Schaffner,W. (1994) Transcriptional activation modulated by homopolymeric glutamine and proline stretches. *Science*, **263**, 808–811.
16. Fiumara,F., Rajasethupathy,P., Antonov,I., Kosmidis,S., Sossin,W.S. and Kandel,E.R. (2015) MicroRNA-22 gates long-term heterosynaptic plasticity in aplysia through presynaptic regulation of CPEB and downstream targets. *Cell. Rep.*, **11**, 1866–1875.
17. Dover,G.A. and Flavell,R.B. (1984) Molecular coevolution: DNA divergence and the maintenance of function. *Cell*, **383**, 622–623.
18. Pazos,F. and Valencia,A. (2008) Protein co-evolution. co-adaptation and interactions. *EMBO J.*, **27**, 2648–2655.
19. De Juan,D., Pazos,F. and Valencia,A. (2013) Emerging methods in protein co-evolution. *Nat. Rev. Genet.*, **14**, 249–261.
20. Dover,G. (2000) How genomic and developmental dynamics affect evolutionary processes. *Bioessays*, **22**, 1153–1159.
21. Peixoto,A.A., Hennessy,J.M., Townson,I., Hasan,G., Rosbash,M., Costa,R. and Kyriacou,C.P. (1998) Molecular coevolution within a *Drosophila* clock gene. *Proc. Natl Acad. Sci. U.S.A.*, **95**, 4475–4480.
22. Travers,S.A. and Fares,M.A. (2007) Functional coevolution.ary networks of the Hsp70–Hop–Hsp90 system revealed through computational analyses. *Mol. Biol. Evol.*, **24**, 1032–1044.
23. Tillier,E.R. and Charlebois,R.L. (2009) The human protein coevolution. network. *Genome Res.*, **19**, 1861–1871.
24. Wang,Y., Correa Marrero,M., Medema,M.H. and van Dijk,A.D.J. (2020) Coevolution-based prediction of protein–protein interactions in polyketide biosynthetic assembly lines. *Bioinformatics*, **36**, 4846–4853.
25. Pelassa,I. and Fiumara,F. (2015) Differential occurrence of interactions and interaction domains in proteins containing homopolymeric amino acid repeats. *Front.Genet.*, **6**, 345.
26. Hosp,F., Gutiérrez-Ángel,S., Schaefer,M.H., Cox,J., Meissner,F., Hipp,M.S., Hartl,F.U., Klein,R., Dudanova,I. and Mann,M. (2017) Spatiotemporal proteomic profiling of Huntington's disease inclusions reveals widespread loss of protein function. *Cell Rep.*, **21**, 2291–2303.
27. Lilliu,E., Villeri,V., Pelassa,I., Cesano,F., Scarano,D. and Fiumara,F. (2018) Polyserine repeats promote coiled coil-mediated fibril formation and length-dependent protein aggregation. *J. Struct. Biol.*, **204**, 572–584.
28. Carbon,S., Ireland,A., Mungall,C.J., Shu,S., Marshall,B., Lewis,S., Lomax,J., Mungall,C., Hitz,B., Balakrishnan,R. and Web Presence Working Group. (2009) AmiGO: online access to ontology and annotation data. *Bioinformatics*, **25**, 288–289.
29. Köhler,S., Gargano,M., Matentzoglu,N., Carmody,L.C., Lewis-Smith,D., Vasileversusky,N.A., Danis,D., Balagura,G., Baynam,G., Brower,A.M. *et al.* (2021) The human phenotype ontology in 2021. *Nucleic Acids Res.*, **49**, D1207–D1217.
30. Buniello,A., Macarthur,J.A.L., Cerezo,M., Harris,L.W., Hayhurst,J., Malangone,C., McMahon,A., Morales,J., Mountjoy,E., Sollis,E. *et al.* (2019) The NHGRI-EBI GWAS catalog of published genome-wide association studies targeted arrays and summary statistics 2019. *Nucleic Acids Res.*, **47**, D1005–D1012.
31. Kandel,E.R., Schwartz,J.H., Jessell,T.M., Siegelbaum,S. and Hudspeth,A.J. (2012) In: *Principles of Neural Science*. McGraw-Hill Education. NY.
32. Pinessi,L., Gentile,S. and Rainero,I. (2015) In: *Neurology. Neurological anatomy physiology and semiology*. Edi-Ermes, Milan, IT.
33. American Psychiatric Association (2013) In: *Diagnostic and statistical manual of mental disorders*. (5th edn.), Arlington, VA.
34. Gilbert,S.F. and Barresi,M.J.F. (2016) In: *Developmental Biology*. Sinauer Associates, Sunderland, MA.
35. Richtsmeier,J.T. and Flaherty,K. (2013) Hand in glove: brain and skull in development and dysmorphogenesis. *Acta Neuropathol. (Berl)*, **125**, 469–489.
36. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B Methodol.*, **57**, 289–300.
37. De Hoon,M.J., Imoto,S., Nolan,J. and Miyano,S. (2004) Open source clustering software. *Bioinformatics*, **20**, 1453–1454.
38. Saldanha,A.J. (2004) Java Treeview—extensible visualization of microarray data. *Bioinformatics*, **20**, 3246–3248.
39. Fisher,R.A. (1915) Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, **10**, 507–521.
40. Hedges,S.B., Dudley,J. and Kumar,S. (2006) TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics*, **22**, 2971–2972.
41. Kumar,S., Stecher,G. and Tamura,K. (2016) MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.*, **33**, 1870–1874.
42. Letunic,I. and Bork,P. (2019) Interactive Tree Of Life iTOL v4: recent updates and new developments. *Nucleic Acids Res.*, **47**, W256–W259.
43. Sievers,F., Wilm,A., Dineen,D., Gibson,T.J., Karplus,K., Li,W., Lopez,R., McWilliam,H., Remmert,M., Söding,J. *et al.* (2011) Fast scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, **7**, 539.

44. Sigrist,C.J.A., De Castro,E., Cerutti,L., Cuche,B.A., Hulo,N., Bridge,A., Bougueleret,L. and Xenarios,I. (2012) New and continuing developments at PROSITE. *Nucleic Acids Res.*, **41**, D344–D347.
45. Shannon,P., Markiel,A., Ozier,O., Baliga,N.S., Wang,J.T., Ramage,D., Amin,N., Schwikowski,B. and Ideker,T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.
46. Buschiazzo,E. and Gemmell,N.J. (2006) The rise, fall and renaissance of microsatellites in eukaryotic genomes. *Bioessays*, **28**, 1040–1050.
47. Patel,S.R., Kim,D., Levitan,I. and Dressler,G.R. (2007) The BRCT-domain containing protein PTIP links PAX2 to a histone H3 lysine 4 methyltransferase complex. *Dev. Cell*, **13**, 580–592.
48. McVeigh,T.P., Banka,S. and Reardon,W. (2015) Kabuki syndrome: expanding the phenotype to include microphthalmia and anophthalmia. *Clin. Dysmorphol.*, **24**, 135–139.
49. Lee,S.K., Anzick,S.L., Choi,J.E., Bubendorf,L., Guan,X.Y., Jung,Y.K., Kallioniemi,O.P., Kononen,J., Trent,J.M., Azorsa,D. *et al.* (1999) A nuclear factor ASC-2 as a cancer-amplified transcriptional coactivator essential for ligand-dependent transactivation by nuclear receptors in vivo. *J. Biol. Chem.*, **274**, 34283–34293.
50. Chatr-Aryamontri,A., Breitkreutz,B.J., Oughtred,R., Boucher,L., Heinicke,S., Chen,D., Stark,C., Breitkreutz,A., Kolas,N., O'Donnell,L. *et al.* (2015) The BioGRID interaction database: 2015; update. *Nucleic Acids Res.*, **43**, D470–D478.
51. Ayllón-Benítez,A., Mougin,F., Allali,J., Thiébaut,R. and Thébault,P. (2018) A new method for evaluating the impacts of semantic similarity measures on the annotation of gene sets. *PLoS One*, **13**, e0208037.
52. Haynes,W.A., Tomczak,A. and Khatri,P. (2018) Gene annotation bias impedes biomedical research. *Sci. Rep.*, **8**, 1362.
53. Griesemer,M., Kimbrel,J.A., Zhou,C.E., Navid,A. and D'haeseleer,P. (2018) Combining multiple functional annotation tools increases coverage of metabolic annotation. *BMC Genomics*, **19**, 948.
54. Townes,F.W. and Miller,J.W. (2020) Identifying longevity associated genes by integrating gene expression and curated annotations. *PLoS Comput. Biol.*, **16**, e1008429.
55. Alba,M.M. and Guigo,R. (2004) Comparative analysis of amino acid repeats in rodents and humans. *Genome Res.*, **14**, 549–554.
56. Faux,N.G., Bottomley,S.P., Lesk,A.M., Irving,J.A., Morrison,J.R., De La Banda,M.G. and Whisstock,J.C. (2005) Functional insights from the distribution and role of homopeptide repeat-containing proteins. *Genome Res.*, **15**, 537–551.
57. Harrison,P.M. (2006) Exhaustive assignment of compositional bias reveals universally prevalent biased regions: analysis of functional associations in human and Drosophila. *BMC Bioinformatics*, **7**, 441.
58. Gettler,L.T., Ryan,C.P., Eisenberg,D.T.A., Rzhetskaya,M., Hayes,M.G., Feranil,A.B., Bechayda,S.A. and Kuzawa,C.W. (2017) The role of testosterone in coordinating male life history strategies: the moderating effects of the androgen receptor CAG repeat polymorphism. *Horm. Behav.*, **87**, 164–175.
59. Manuck,S.B., Marsland,A.L., Flory,J.D., Gorka,A., Ferrell,R.E. and Hariri,A.R. (2010) Salivary testosterone and a trinucleotide CAG length polymorphism in the androgen receptor gene predict amygdala reactivity in men. *Psychoneuroendocrinology*, **35**, 94–104.
60. Lenz,B., Jacob,C., Frieling,H., Jacobi,A., Hillemacher,T., Muschler,M., Watson,C., Kornhuber,T. and Bleich,S. (2009) Polymorphism of the long polyglutamine tract in the human androgen receptor influences craving of men in alcohol withdrawal. *Psychoneuroendocrinolog*, **34**, 968–971.
61. Vermeersch,H., T'sjoen,G., Kaufman,J.M., Vincke,J. and Van Houtte,M. (2010) Testosterone androgen receptor gene CAG repeat length mood and behaviour in adolescent males. *Eur. J. Endocrinol.*, **163**, 319.
62. Garai,C., Furuichi,T., Kawamoto,Y., Ryu,H. and Inoue-Murayama,M. (2014) Androgen receptor and monoamine oxidase polymorphism in wild bonobos. *Meta Gene*, **2**, 831–843.
63. Hammock,E.A., Lim,M.M., Nair,H.P. and Young,L.J. (2005) Association of vasopressin 1a receptor levels with a regulatory microsatellite and behavior. *Genes Brain Behav.*, **4**, 289–301.
64. Rubenstein,D.R., Ågren,J.A., Carbone,L., Elde,N.C., Hoekstra,H.E., Kapheim,K.M., Keller,L., Moreau,C.S., Toth,A.L., Yeaman,S. *et al.* (2019) Coevolution of genome architecture and social behavior. *Trends Ecol. Evol.*, **34**, 844–855.
65. Estrada-Sánchez,A.M., Levine,M.S. and Cepeda,C. (2017) Epilepsy in other neurodegenerative disorders: Huntington's and Parkinson's diseases. In: *Models of Seizures and Epilepsy*, Academic Press, Cambridge, MA, pp. 1043–1058.
66. Elden,A.C., Kim,H.J., Hart,M.P., Chen-Plotkin,A.S., Johnson,B.S., Fang,X., Armakola,M., Geser,F., Greene,R., Lu,M.M. *et al.* (2010) Ataxin-2 intermediate-length polyglutamine expansions are associated with increased risk for ALS. *Nature*, **466**, 1069–1075.
67. Luoma,P.T., Eerola,J., Ahola,S., Hakonen,A.H., Hellström,O., Kivistö,K.T., Tienari,P.J. and Suomalainen,A. (2007) Mitochondrial DNA polymerase gamma variants in idiopathic sporadic Parkinson disease. *Neurology*, **69**, 1152–1159.
68. Yanicostas,C., Barbieri,E., Hibi,M., Brice,A., Stevanin,G. and Soussi-Yanicostas,N. (2012) Requirement for zebrafish ataxin-7 in differentiation of photoreceptors and cerebellar neurons. *PLoS One*, **71**, e50705.
69. Lebon,C., Behar-Cohen,F. and Torriglia,A. (2019) Cell death mechanisms in a mouse model of retinal degeneration in Spinocerebellar Ataxia 7. *Neuroscience*, **400**, 72–84.
70. Toulis,V., García-Monclús,S., de la Peña-Ramírez,C., Arenas-Galnares,R., Abril,J.F., Todi,S.V., Khan,N., Garanto,A., do Carmo Costa,M. and Marfany,G. (2020) The deubiquitinating enzyme ataxin-3 regulates ciliogenesis and phagocytosis in the retina. *Cell Rep.*, **33**, 108360.
71. Benitez-Burraco,A. and Boeckx,C. (2015) Possible functional links among brain-and skull-related genes selected in modern humans. *Front. Psychol.*, **6**, 794.
72. Sawaya,M.E. and Shalita,A.R. (1998) Androgen receptor polymorphisms CAG repeat lengths in androgenetic alopecia hirsutism and acne. *J. Cutan. Med. Surg.*, **3**, 9–15.
73. Yang,Z., Yu,H., Cheng,B., Tang,W., Dong,Y., Xiao,C. and He,L. (2009) Relationship between the CAG repeat polymorphism in the androgen receptor gene and acne in the Han ethnic group. *Dermatology.*, **218**, 302–306.
74. Morrison,N.A., Stephens,A.A., Osato,M., Polly,P., Tan,T.C., Yamashita,N., Doecke,J.D., Pasco,J., Fozzard,N., Jones,G. *et al.* (2012) Glutamine repeat variants in human RUNX2 associated with decreased femoral neck BMD broadband ultrasound attenuation and target gene transactivation. *PLoS One*, **7**, e42617.
75. Birge,L.M., Pitts,M.L., Baker,R.H. and Wilkinson,G.S. (2010) Length polymorphism and head shape association among genes with polyglutamine repeats in the stalk-eyed fly Teleopsis dalmanni. *BMC Evol. Biol.*, **10**, 227.
76. Staes,N., Sherwood,C.C., Wright,K., De Manuel,M., Guevara,E.E., Marques-Bonet,T., Krützen,M., Massiah,M., Hopkins,W.D., Ely,J.J. *et al.* (2017) FOXP2 variation in great ape populations offers insight into the evolution. of communication skills. *Sci. Rep.*, **7**, 16866.
77. Mularoni,L., Toll-Riera,M. and Alba,M.M. (2008) Comparative genetics of trinucleotide repeats in the human and ape genomes. doi:10.1002/9780470015902.a0020844.
78. Adams,D.C. (2014) Quantifying and comparing phylogenetic evolutionary rates for shape and other high-dimensional phenotypic data. *Syst. Biol.*, **63**, 166–177.
79. Huey,R.B. and Bennett,A.F. (1987) Phylogenetic studies of coadaptation: preferred temperatures versus optimal performance temperatures of lizards. *Evolution.*, **41**, 1098–1115.
80. Gittleman,J.L., Anderson,C.G., Kot,M. and Luh,H.K. (1996) Phylogenetic lability and rates of evolution: a comparison of behavioral morphological and life history traits. In: Martins,E.P. (ed). *Phylogenies and the comparative method in animal behavior*. Oxford University Press, Oxford, UK, pp. 166–205.
81. Blomberg,S.P. and Garland,T. Jr (2002) Tempo and mode in evolution: phylogenetic inertia adaptation and comparative methods. *J. Evol. Biol.*, **15**, 899–910.
82. Legendre,M., Pochet,N., Pak,T. and Verstrepen,K.J. (2007) Sequence-based estimation of minisatellite and microsatellite repeat variability. *Genome Res.*, **17**, 1787–1796.
83. Whiten,A. and Erdal,D. (2012) The human socio-cognitive niche and its evolutionary origins. *Philos.Trans. R. Soc. B Biol. Sci.*, **367**, 2119–2129.
84. Roth,G. and Dicke,U. (2019) Origin and evolutionof human cognition. In: *Progress in Brain Research*. Elsevier, Amsterdam, NL, Vol. **250** pp. 285–316.

85. Kanzaki,H., Yoshida,K., Saitoh,H., Fujisaki,K., Hirabuchi,A., Alaux,L., Fournier,E., Tharreau,D. and Terauchi,R. (2012) Arms race co-evolution of Magnaporthe oryzae AVR-Pik and rice Pik genes driven by their physical interactions. *Plant J.*, **72**, 894–907.
86. Mier,P., Alanis-Lobato,G. and Andrade-Navarro,M.A. (2017) Protein–protein interactions can be predicted using coiled coil co-evolution. patterns. *J. Theor. Biol.*, **412**, 198–203.
87. Fang,X., Wang,L., Ishikawa,R., Li,Y., Fiedler,M., Liu,F., Calder,G., Rowan,B., Weigel,D., Li,P. *et al.* (2019) Arabidopsis FLL2 promotes liquid–liquid phase separation of polyadenylation complexes. *Nature*, **569**, 265–269.
88. Prontera,P., Ottaviani,V., Rogaia,D., Isidori,I., Mencarelli,A., Malerba,N., Cocciadiferro,D., Rolph,P., Stangoni,G., Vulto-van Silfhout,A. *et al.* (2016) A novel MED12 mutation: evidence for a fourth phenotype. *Am. J. Med. Genet. Part A*, **170**, 2377–2382.
89. Fuller,T.D., Westfall,T.A., Das,T., Dawson,D.V. and Slusarski,D.C. (2018) High-throughput behavioral assay to investigate seizure sensitivity in zebrafish implicates ZFHX3 in epilepsy. *J. Neurogenet.*, **32**, 92–105.