Evolution of Protein Domain Repeats in Metazoa

Andreas Schüler and Erich Bornberg-Bauer*
Institute for Evolution and Biodiversity, Westfalian Wilhelms University, Huefferstrasse 1, Muenster, Germany
***Corresponding author:** E-mail: ebb@wwu.de.

Associate editor: Jeffery Thorne

Abstract

Repeats are ubiquitous elements of proteins and they play important roles for cellular function and during evolution. Repeats are, however, also notoriously difficult to capture computationally and large scale studies so far had difficulties in linking genetic causes, structural properties and evolutionary trajectories of protein repeats. Here we apply recently developed methods for repeat detection and analysis to a large dataset comprising over hundred metazoan genomes. We find that repeats in larger protein families experience generally very few insertions or deletions (indels) of repeat units but there is also a significant fraction of noteworthy volatile outliers with very high indel rates. Analysis of structural data indicates that repeats with an open structure and independently folding units are more volatile and more likely to be intrinsically disordered. Such disordered repeats are also significantly enriched in sites with a high functional potential such as linear motifs. Furthermore, the most volatile repeats have a high sequence similarity between their units. Since many volatile repeats also show signs of recombination, we conclude they are often shaped by concerted evolution. Intriguingly, many of these conserved yet volatile repeats are involved in host-pathogen interactions where they might foster fast but subtle adaptation in biological arms races.

***Key Words*: protein evolution, domain rearrangements, protein repeats, concerted evolution.**

Introduction

Most proteins are composed of one or more domains which are smaller units with an evolutionary conserved structure. Quite often, these domains are reused across many proteins in a different context, i.e., they combine with other domains in changing order and quantity (Apic et al. 2001; Björklund et al. 2005; Moore et al. 2008; Kummerfeld and Teichmann 2009; Bornberg-Bauer and Alb 2013). One striking property of multi-domain proteins is the frequent occurrence of domain repeats in proteomes across the tree of life and in particular in eukaryotic proteomes (Marcotte et al. 1999; Björklund et al. 2006; Schaper et al. 2014; Jernigan and Bordenstein 2015). However, estimates on how frequent repeat-containing proteins are depend on the dataset and the method being used for repeat detection. In an early study, 13% of proteins in Swiss-Prot were found to be repeat-containing (Marcotte et al. 1999), but more recent estimates (Pellegrini et al. 2012) put this fraction at 25%. Protein repeats may be a very important source of gene novelties from which new proteins or fragments thereof arise (Lupas et al. 2001; Bornberg-Bauer et al. 2010; Bornberg-Bauer and Albà 2013; Toll-Riera and Albà 2013; Laurino et al. 2016). Some protein repeats are hyper-variable, i.e., the number of repeat units within the domain repeat changes within evolutionary short time by insertion (expansion) or deletion (contraction) of units. In some proteins, these changes occur so rapidly that they can be observed not only between species but also within a population under both adaptive (Verstrepen et al. 2005) and neutral conditions (Chevanne et al. 2010).

The most common known biological function of domain repeats is binding to other proteins, to RNA and DNA and to smaller ligands such as ATP (Pawson and Nash 2003). Due to their flexibility in binding targets, domain repeats occur in proteins which are responsible for many different biological functions such as gene regulation, protein transport, and protein cell cycle control (Andrade et al. 2001).

The length of the *repeated unit* in a *protein repeat* varies significantly, from just a single amino acid to more than 100 amino acids. Repeats with very short unit lengths, such as mono- and dipeptide repeats, are rare and often associated with disease (Mirkin 2006, 2007; Gemayel et al. 2010). Other repeats with slightly longer unit lengths form extended fibrous structures like coiled coils and collagen (Kajava 2001). For longer repeat units, each repeat unit often corresponds to one entire globular protein domain (Kajava 2012). We refer to repeats composed of units of at least 12 amino acids as *domain repeats*. A widely used classification of such domain repeats is based on whether the individual units fold independently into separate globular domains or cooperatively into one globular domain. For cooperative folding, the classification further distinguishes long elongated structures and circular barrel- or propeller-like folds (Kajava 2012). However, such a classification is complicated by proteins or parts thereof that do not fold into a regular structure composed of, e.g., α helices and/or β sheets but, instead, are "disordered" (van der Lee et al. 2014). Protein repeats are often partially or completely disordered (Tompa 2003) and the "purity" of repeats is positively correlated with the degree of disorder (Jorda et al. 2010). Repeat purity here is defined as the average pairwise sequence identity between all individual repeat units within one protein. For some domain repeats, artificial protein constructs with varying degrees of repeat purity have been designed and it has been demonstrated that above 40% sequence identity, protein aggregation and misfolding become very likely (Wright et al. 2005). This is a likely reason for why repeats above that degree of similarity are exceedingly rare in naturally occurring proteins (Wright et al. 2005; Street et al. 2006; Reshef et al. 2010). The low degree of similarity between the individual repeat units of many repeat-containing proteins is also the main reason for why the prediction of such repeats with bioinformatics methods is a very challenging problem (Biegert and Söding 2008).

The number of domain repeat units in protein families is variable and, on average, higher in eukaryotes compared with prokaryotes (Björklund et al. 2006). Several studies have described changes in domain arrangements (i.e., the N- to C-terminal order of domains in a protein) and concluded that the dominating forces of rearrangement are duplication of protein coding genes, fusion and terminal domain losses (Apic et al. 2003; Ye and Godzik 2004; Kummerfeld and Teichmann 2005; Weiner and Bornberg-Bauer 2006; Weiner et al. 2006; Wang and Caetano-Anollés 2009; Zmasek and Godzik 2011, 2012; Forslund and Sonnhammer 2012; Moore et al. 2013) whereas fissions and emergence of novel domains are rather rare (Kummerfeld and Teichmann 2005; Moore and Bornberg-Bauer 2012; Kersting et al. 2012). In all of these studies, domain repeats have been treated as a single domain, i.e., repeat units have been computationally "collapsed" into a single unit for better handling. Therefore, possible changes in repeat unit number have not been considered and relatively little is currently known about the expansion and contraction of domain repeats through insertions and deletions (indels) of repeat units. However, recent methodological advances have made it possible to reliably infer protein repeat indels (Szalkowski and Anisimova 2013; Schaper and Anisimova 2014). First results indicate that, within the mammalian clade, the number of repeat units in human repeat-containing proteins is well conserved (Schaper and Anisimova 2014). Some human protein families on the other hand appear to experience frequent repeat indels, particularly zinc-finger proteins (Schaper and Anisimova 2014). The underlying reasons for the high variability of some protein families with respect to repeat indels are currently not well understood.

Here we present a study on the evolution of protein domain repeats in a comprehensive dataset of proteins from 109 completely sequenced metazoan genomes. We use a recently developed approach (Szalkowski and Anisimova 2013) to infer domain repeat indels in protein families and correlate the frequency of indels with protein structure, disorder, and function. Based on this, we provide several explanations for the question as to why the numbers of units in some domain repeats are highly variable whereas many others are well or even perfectly conserved in this respect.

Results and Discussion

Frequency of Domain Repeat Insertions and Deletions

We analyzed protein domain repeats in the sequences of the TreeFam database (version 9) (Ruan et al. 2008). TreeFam classifies proteins from 109 completely sequenced genomes, 104 animal species and five outgroups. Proteins are grouped into 15,322 *TreeFam* families of proteins that each most likely evolved from a common ancestral gene and multiple sequence alignments (MSAs) are provided for each family. We scanned 15,322 TreeFam families for the presence of repeats with a minimum unit length of 12 amino acids using HHRepID (Biegert and Söding 2008). Whereas most approaches for *de novo* identification of repeats rely on comparing a protein sequence to itself, HHRepID further utilizes evolutionary information by aligning homologous sequences to the query sequence, leading to a significantly improved sensitivity (Biegert and Söding 2008). A *de novo* approach was preferred over annotating domain repeats with profile hidden Markov models (HMMs) for known domain repeats, such as those provided by Pfam (Punta et al. 2012) or SUPERFAMILY (Wilson et al. 2009) for two reasons. First, the set of known domain repeats in these databases is not exhaustive and, accordingly, *de novo* approaches find many repeats that do not match any already known domain (Ponting et al. 2001; Jorda et al. 2012). Second, many HMMs provided by Pfam and SUPERFAMILY do not specifically match an individual repeat unit, but rather stretches of several repeat units in tandem. While this feature may be useful for many applications, it prevents determining the exact number and boundaries of individual repeat units in a protein and hence complicates analysis of repeat indels.

We discarded repeats that do not occur in at least 10 different proteins within one *TreeFam* protein family. This results in a dataset of 3,838 domain repeat families that we refer to as *TreeFamRep*. Note that one TreeFam protein family can include more than one domain repeat family because some proteins include more than one type of repeat. For a family of proteins with, for example, Ankyrin repeats and BTB repeats, which are not evolutionary related with each other and would be detected separately by HHRepID, we would create two separate domain repeat families. For each domain repeat family, we built custom HMMs and used HHsearch (Söding 2005) to compare these HMMs against the Pfam (Finn et al. 2010) and SUPERFAMILY (Wilson et al. 2009) databases. For 3,095 (80.6%) domain repeat families, we find significant similarity (e-value $< 1e-5$) to known Pfam or SUPERFAMILY domains. In other words, roughly 1 out of 5 families in the TreeFamRep dataset are "novel" in the sense that they are not homologous to any known protein domain. These newly detected repeat domains have interesting properties such as a significantly higher proportion of protein disorder and higher indel rates on average, which we discuss further below.

For each family, we input the full protein sequences and repeat locations as determined by HHRepID into the phylogeny aware ProGraphMSA + TR multiple sequence alignment tool (Szalkowski and Anisimova 2013). We extracted the predicted number of repeat indels for each family from the ProGraphMSA + TR output. Figure 1 shows the frequency distribution of observed repeat indels per repeat family. Many repeat families are well-conserved, and for 923 (24%) of all families we observe no indels at all (fig. 1A). For 2024 repeat families (52.6%), we observe 0.1 or fewer indels per protein (fig. 1B). Note that it is not possible to reliably distinguish between perfectly conserved families where all proteins

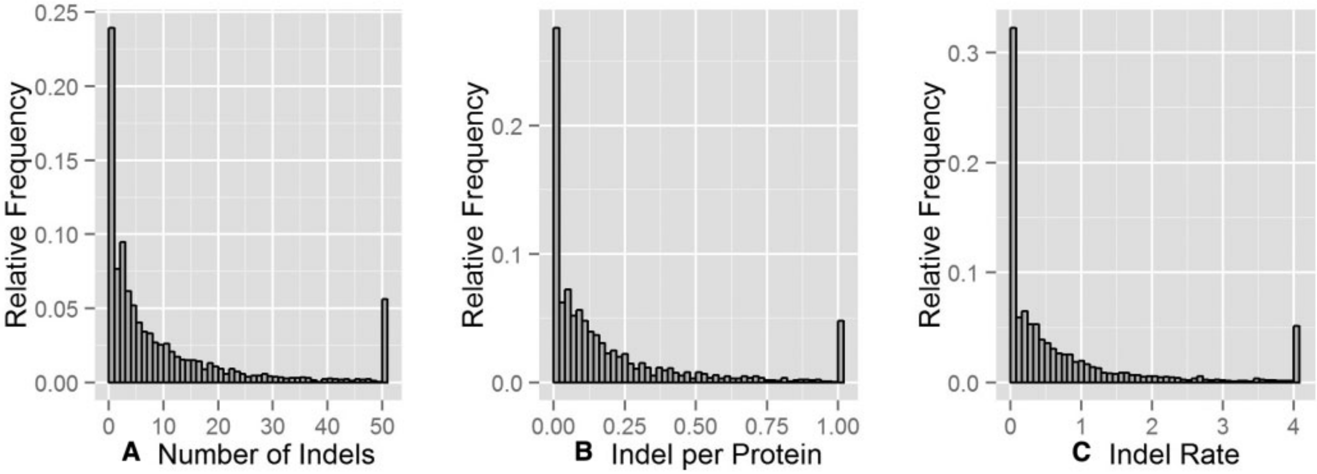

Fig. 1. Distributions of repeat domain indel frequency: (*A*) frequencies of the total number of repeat domain indels within one domain repeat family. (*B*) Frequencies of the number of indels divided by the number of proteins within the respective. (*C*) Frequencies of indel rates (see main text for details). For each graph, the rightmost bin cumulates all values that were higher than the limit of the *x*-axis.

have the same number of repeat units that can be aligned to each other in an MSA and those that are only very well but not perfectly conserved (see “Materials and Methods” for details). Results so far suggest that the length of repeats within a protein family remain frequently conserved over evolutionary long time scales. However, the distribution of indels per protein is long-tailed with many outliers. For 216 repeat families (5.6%), we observe 50 or more indels.

The ratio of indels per protein can in some cases produce misleading results. For a repeat-containing protein family which includes many recently duplicated paralogs, the time-frame in which repeat indels could have occurred would be much lower compared with a protein family that does not contain many recently emerged paralogs. The ratio of indels per protein would then not be comparable between the two cases. Therefore, we also calculated a different measure of domain repeat indels. We divided the number of indels by the total length of the maximum likelihood phylogenetic tree as provided by TreeFam for those proteins (fig. 1C). We will refer to the latter as *indel rate* from now on. The indel rate only depends on the rate of evolutionary changes being approximately equal over evolutionary long time frames between the families that are compared. Strong deviations from such a clock-like process would render rates not comparable between families. For the following analyses, we decided to use the indel rate instead of the indel per protein ratio. The justification for this is that retention rates after gene duplication, which affects the number of paralogs in a protein family, are highly variable between different protein families (Shiu et al. 2006; Morel et al. 2015). Significant and long deviations from the molecular clock hypothesis, however, are rare at the sequence level (Kumar 2005) and for domain rearrangements (Kersting et al. 2012; Moore et al. 2013).

The strong conservation of unit numbers for the majority of repeat families is in agreement with a recent study which found conservation for 61% of human protein repeats since the origin on the mammalian clade, and for 17% since the origin of the vertebrate clade (Schaper et al. 2014). The low average indel rate in our data is, however, accompanied by a very high variance because there is a significant fraction of outliers with extremely high indel rates. About 364 (9.4%) families have indel rates higher than $(Q3 + 1.5 \times IQR)$, with $Q3$ being the third quartile of the indel rate distribution and IQR being the interquartile range, i.e., the range between third quartile and first quartile. We refer to these 364 families as *TreeFamRepHind* and the remaining families as *TreeFamRepLoind*. The underlying reasons for this discrepancy between many well conserved and few highly volatile repeat families will be addressed next.

Comparison of Domain Repeat Indel Rates for Different Protein Structure Classes

We first turn our attention to the possible effect which protein structure might have on the volatility of protein repeats. We classified domain repeats according to the schema established by Kajava (2001, 2012). In this classification, domain repeats are composed of units that fold cooperatively into either long “elongated” structures, circular “closed” structures, or a collection of units that can fold independently (see fig. 2 top). The RepeatsDB project (Di Domenico et al. 2014) provides annotations of structures from PDB entries (Berman et al. 2000) according to this schema. We used hmmbuild and hmmscan from the HMMER3 package (Eddy 2011) to first construct HMMs based on MSAs of each Treefam repeat family, and then scanned these HMMs against all protein sequences with PDB structures annotated by RepeatsDB. For 1293 (33.7%) repeat families we detected significant hits in PDB. Of these, 538 were annotated as “elongated”, 309 as “closed”, and 446 as “independent” folds.

In figure 2, we show comparisons of the indel rates for elongated, closed and independent folds. While the median indel rates for closed and elongated repeat families are quite similar, we observe a significantly and much higher median indel rate of 0.95 for independently folding repeats. The higher indel rate for independently folding repeats was to be expected, because independently folding repeats are

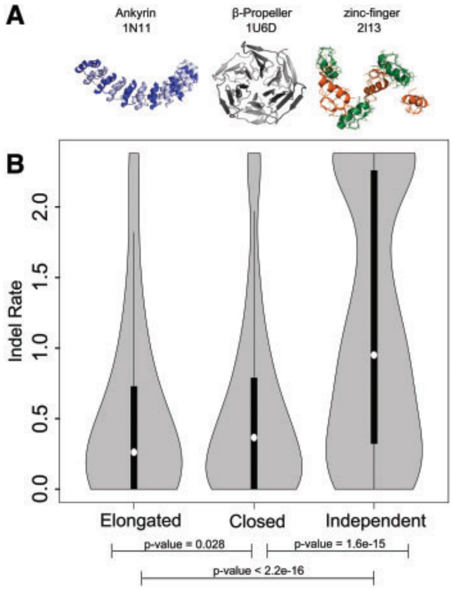

Fig. 2. Comparison of indel rates for elongated, closed and independent domain repeat families. (*A*) a representative structure with PDB ID for each class. The individual repeat units are highlighted with alternating colors. (*B*) Violin plots for the repeat indel rates. All outliers have been collapsed into the highest value on the *y*-axis for readability (a visualization without collapsed outliers is shown in supplementary fig. S1, Supplementary Material online).

more likely to tolerate indels which would not adversely affect the fold (and thus the function) of adjacent units. It is however surprising that elongated repeats do, on average, not have significantly higher indel rates than closed repeats but, on the contrary, do have even smaller rates. The closed category includes several barrel-shaped structures that bind their substrates in the interior of the structure. Since indels would change the diameter of such a barrel-shaped structure it seems unlikely that the original function would be preserved. For elongated structures with binding functions on the other hand, duplications seem intuitively less likely to destroy the original function. However, the individual units in elongated repeats are often not functional units. Protein binding interfaces in elongated repeats for example frequently consist of a specific number of individual units (Grove et al. 2008), and this could represent additional constraints because only insertions that correspond to this number are likely to be functional. This would also fit with the observation that the size of indels is not randomly distributed in many repeat families but rather corresponds to indels that match a specific number of units, so that some families have a typical indel size corresponding to a single repeat unit whereas others have indels typically matching, two, three or even more repeat units (Björklund et al. 2006).

There are also notable variations of indel rates within each structural group and sometimes also between different families which contain the same domain type as repeat unit. One reason for such differences is probably that very similar protein structures can be used for very different biological functions, some of which would tolerate variance in the number of repeat units whereas others would not. Ankyrin repeat domains, for example, often mediate protein–protein interactions, with each repeat unit corresponding to one binding interface (Mosavi et al. 2004). Accordingly, it is possible in many proteins to add more Ankyrin repeat units by insertions without affecting the original binding functionalities. For the TRP-N family of mechanoreceptor proteins however, Ankyrin repeats fulfill a different functional role and serve as an elastic mechanical spring (Liang et al. 2013). For this specific function, a repeat number of approximately 29 is optimal because it allows a transmission of mechanical force from one end of the Ankyrin repeat structure to the other end without creating a torque (Howard and Bechstedt 2004). This structural role of Ankyrins in the TRP-N family coincides with a very low rate of indels across the whole metazoan tree, which is rather atypical for Ank repeats (Schüler et al. 2015).

Protein Disorder Is Linked to High Domain Repeat Indel Rates

We next study the intricate relationship between disorder and protein repeats. In an early analysis of 126 largely disordered proteins, it was shown that tandem repeats occur frequently within disordered regions and that interspecies variations in repeat unit number are frequent (Tompa 2003). It has been hypothesized that disordered regions more readily tolerate indels due to the low degree of long-range interactions between amino acids in those regions (Tompa 2003). A recent study however reported only a weak association between repeats and length variation of disordered protein regions (Light et al. 2013b).

For all proteins within the TreeFamRep set, we predicted disorder contents based on the IUPred scheme (Dosztányi et al. 2005) as implemented in the ANCHOR program (Mészáros et al. 2009). This method estimates, for each amino acid within a protein, how likely it can form enough energetically favorable interactions to overcome the entropy loss of folding, given its context in the protein sequence. For each protein sequence in the TreeFamRep set, we calculated the IUPred scores and classified protein regions with an IUPred score $\geq$0.5 as disordered. We then determined for each domain repeat family the average fraction of amino acids that correspond to disordered regions within the repeat sequences. About 738 (19.2%) of the domain repeat families in TreeFamRep are largely disordered, i.e., the average fraction of disordered amino acids within the repeat sequences is $\geq$50%. We refer to this dataset as *TreeFamRepDis* from now on and to all other domain repeat families as *TreeFamRepStruc*.

We find a median value for the repeat indel rate of 0.55 for the TreeFamRepDis families and 0.28 for the TreeFamRepStruc families (*P*-value = 1.4e−14, Kolmogorov–Smirnov test) (fig. 3). This difference thus explains some of the variance of indel rates for domain repeat families reported above. Since disordered regions form few or negligible interactions with the rest of the protein chain, structure and function of the remaining protein will be less affected by repeat indels. This seems to lead to indels being more readily tolerated and preserved in disordered regions, as has been hypothesized earlier (Tompa 2003).

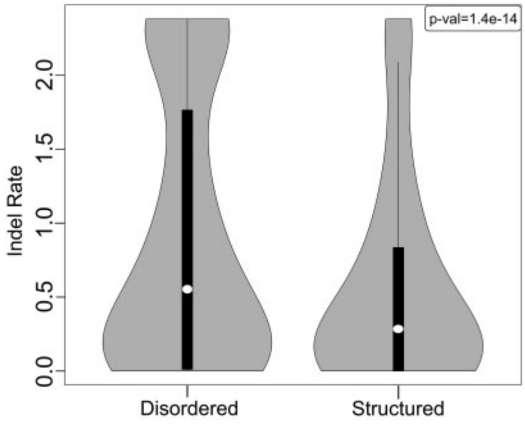

Fig. 3. Comparison of domain repeat indel rates between repeats that are largely disordered and repeats that are largely structured. The violin plots correspond to box plots with added probability density distribution of the data (mirrored to the left and right of the box plot). For better visualization, all outliers have been collapsed into the highest point on the *y*-axis.

Previously, it had been assumed that disordered regions have no role other than being flexible linkers between structured and functional folded domains (van der Lee et al. 2014). However, it is now well established that disordered regions very often play more active roles for protein function. This is mainly due to short linear motifs being present in disordered regions that represent sites for post-translational modifications (Iakoucheva et al. 2004) or recognition sites that mediate transient interactions with other proteins (Davey et al. 2012). Furthermore, disordered regions are often capable of coupled folding and binding, i.e., they undergo a disorder to order transition in physical proximity of a suitable binding partner (Dyson and Wright 2002).

To distinguish between these possible functionalities we annotated repeats with known linear binding motifs from the ELM database (Dinkel et al. 2012) and identified disordered regions with a predicted potential for coupled folding and binding based on the ANCHOR program (Dosztányi et al. 2009). For simplicity, we refer to ELM linear binding motifs as "linear" and predicted potential for coupled folding and binding as "folding" from now on. Note that folding residues are also by definition disordered. Linear motifs are also only predicted for residues that are also disordered according to the IUPred scheme because linear motifs need to be exposed in order to be functional (Dinkel et al. 2012). Figure 4 shows a comparison of how many residues within repeat containing proteins were identified as linear or folding. We distinguish between linear and folding residues within repeat regions on the one hand, and non-repeat regions in the same proteins on the other. For this comparison, we considered all TreeFamRep proteins with at least one disordered region of at least 20 amino acids, not only the largely disordered proteins in the TreeFamRepDis set. Note that the TreeFamRepDis set includes only families that are *largely* disordered, whereas short disordered stretches can and do also occur in families outside the TreeFamRepDis set. The median proportion of folding residues is 0.42 which is significantly lower than the median proportion of 0.58 for folding residues in repeats (*P*-value $< 2.2\mathrm{E}{-16}$, Kolmogorov–Smirnov test). The median frequencies/proportions of linear residues in repeat containing proteins are very similarly significant (*P*-value $< 2.2\mathrm{E}{-16}$, Kolmogorov–Smirnov test), with a median of 0.35 in non-repeat residues and 0.54 in repeat residues.

Disordered regions have increasingly been recognized as an indispensable elements in the molecular toolkit of cells (van der Lee et al. 2014) and our results indicate that disordered regions in domain repeats are most likely also highly functional elements. Besides the rich potential of possible binding, recognition, and modification sites, disordered repeats further provide the potential for evolutionary novelty due to their high indel rates. Out of the 738 families in the TreeFamRepDis set, 428 (58%) do not match any known domain from Pfam or SUPERFAMILY, highlighting the relevance of *de novo* approaches for repeat detection. This need not be a methodological error of the orthology based prediction methods, such as Pfam, but can also be attributed to genomic novelties which give rise to novel domains which are, per definition, not represented by predefined HMMs. Interestingly, the fraction of repeat families that do not match any known domain in Pfam or SUPERFAMILY is only 19.4% for the entire TreeFamRep dataset. With 58%, the fraction of "novel" repeat domains is thus almost three times as high in the TreeFamRepDis set.

Disordered regions often evolve faster than structured regions (Dosztányi et al. 2010; Szalkowski and Anisimova 2011; Light et al. 2013a). The consequential strong divergence at the sequence level could explain why most largely disordered domain repeat families cannot be matched to any domain HMM in Pfam or SUPERFAMILY. To test whether disordered regions indeed show more divergence within the TreeFamRepDis set compared with the TreeFamRepStruc set, we compared amino acid conservation by calculating normalized conservation scores using the al2co program (Pei and Grishin 2001). The distribution of conservation scores for disordered and structured repeats is shown in the supplementary figure S2, Supplementary Material online. Indeed, we find that the residues in structured domain repeats are on average better conserved than the rest of the protein, whereas residues in largely disordered domain repeats are equally conserved as the residues in the rest of the protein (see supplementary fig. S2, Supplementary Material online).

Statistical Evaluation of Predictors for Domain Repeat Indel Rates

So far, we have already derived two quantities that are significantly different between domain repeat families with high indel rates and families with low indel rates, namely, the average number of repeat units per protein and the average proportion of disordered residues in the repeat region. Earlier results analyzing the human proteome also showed that repeat proteins with high indel rates have a higher *repeat purity*, i.e., a higher average sequence identity between repeat units than protein repeats with few indels (Schaper et al. 2014). For very short repetitive sequences (i.e., Microsatellites), the average length of the repeated units within a repeat is also

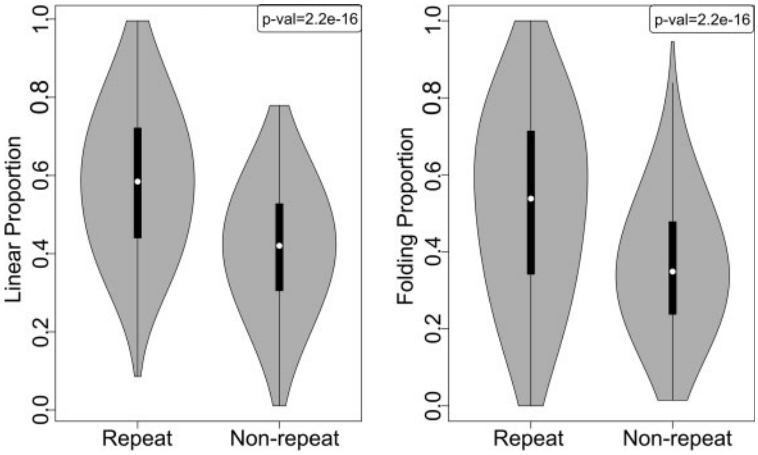

Fig. 4. Comparison of the proportions of linear and folding amino acids in disordered regions of repeats (Repeat) and non-repeat regions (Non-repeat). Left: distribution of amino acid frequencies which match known ELM linear motifs, Right: distribution of amino acids frequencies which were predicted by ANCHOR to have coupled folding and binding potential.

strongly correlated with indel rates (Legendre et al. 2007; Gemayel et al. 2010) but it is unknown whether this correlation still holds for longer repeat units.

It seems unlikely that those four variables (number, length, purity and disorder of repeats) are all statistically independent of each other, and indeed, significant correlations between some have already been reported. Specifically, the amount of disorder in protein repeats seems to correlate with short repeat lengths and high repeat purity (Jorda et al. 2010). We can reproduce this correlation and find lower but still highly significant correlations between all other combinations of repeat unit number, unit length, purity, and disorder, except for the combination of number of repeat units and disorder content within repeats (see supplementary fig. S3, Supplementary Material online).

To test whether those four variables have an independent correlation with indel rate as opposed to merely being associated with it through covariance with one or more of the other variables, we calculated partial correlations (Kim and Yi 2007) (see “Materials and Methods” for details). Normal and partial Spearman correlations are shown in table 1.

Even after controlling for the effects of the other variables, the amount of intrinsic disorder retains a very significant correlation with the indel rate. The previously discussed properties of some disordered regions (linear and folding) have no independent correlation with the indel rate. When the amount of intrinsic disorder is controlled for, both have negligible partial correlations with P-values > 0.05.

Table 1. Correlation of Repeat Indel Rate with Other Properties.

Spearman correlations with repeat indel rate		
	Normal	Partial
#Repeat units	0.57**	0.60**
Repeat length	0.04	0.08*
Repeat purity	0.16**	0.21**
Disorder	0.21**	0.20**

*P-value $<$ 1e−5 and
**P-value $<$2.2e−16.

The correlation between repeat purity and indel rate actually becomes stronger when the other three variables are controlled for. This indicates that for purity, one or more of the remaining three variables act as suppressor variables (Tzelgov and Henik 1991). These suppressor variables covary with purity, but have a covariance with it that is independent of the association between purity and indel rate. Controlling for this covariance thus has the effect of suppressing variance in the purity variable, but only variance that is independent of the association with the indel rate, such that the observed association with the indel rate becomes stronger.

This effect is also observed for repeat length, which has no significant normal correlation with indel rate, but a significant partial correlation. The repeat number retains its extremely strong correlation with indel rate when covariance with the other three variables is controlled for. All correlations except for the correlations of repeat length with indel rate are highly significant (table 1).

The weak correlation of repeat length and indel rate is in contrast to microsatellite repeats in coding regions, where repeat length is a very good predictor for indel rates (Legendre et al. 2007). This difference between microsatellite repeats and domain repeats is probably due to the different genetic mechanisms that underlie repeat indels. Microsatellites change their length mainly through replication slippage, and the likelihood of replication slippage sharply decreases as the length of the repeat units increases (Brinkmann et al. 1998). Protein domain repeats however are too long for replication slippage to occur, and indels are thus caused by other mechanisms such as unequal crossing-over, which are less length-dependent.

The extremely good correlation between repeat unit number and indel rate can be easily explained. An individual domain repeat unit in the TreeFamRep set is at least 12 amino acids long, 48 amino acids on average, and the emergence of many repeat units of this length from a non-coding region is unlikely. The only alternative to an emergence of the entire repeat region from a non-coding sequence is that an ancestral protein had much fewer repeat units or even just a single one, which have expanded through repeat insertions. Therefore, high indel rates are to be expected for domain repeat families with high average repeat number. The correlation between high average repeat unit numbers and high repeat indel rates entails a further conclusion. Protein families with a high number of repeat units that has been fixed in a distant ancestor, and did not experience any indels in all or most extant species, are rare.

Different Repeat Families Can Belong to the Same Structure Class

It is possible that evolutionary related repeat families can be assigned to two (or more) different TreeFam families. There are two possible reasons for this. First, some domains are promiscuous in the sense that they co-occur with many different other domains in many different arrangements of domains (Weiner et al. 2008), a property that is also called “domain versatility”. If a set of proteins have zinc-finger domain repeats in common, but one half also has a BTB domain whereas the other half has a KRAB domain instead, automated clustering methods based on sequence similarity would likely assign them to two different clusters. A second possible reason is that the number of repeat units is extremely variable so that, for example, some proteins have only three repeated units whereas others have 30. This could lead to long and short proteins ending in separate clusters despite bearing the same domain repeats. To test this, we used the assignments to SUPERFAMILY and analyzed whether different repeat families in the TreeFamRep set sometimes correspond to the same SUPERFAMILY domain. We observe 241 different SUPERFAMILY domains in the TreeFamRep set, with 108(45%) corresponding to a single TreeFamRep family and 191(80%) corresponding to five or less families in the TreeFamRep set (supplementary fig. S4*A*, Supplementary Material online).

Most SUPERFAMILY domains thus correspond to few or just a single TreeFamRep families, but we again have significant outliers. Two SUPERFAMILY domains, the WD40 repeat-like and beta-beta-alpha zinc finger domains, occur in more than 100 different TreeFamRep families each (130 and 289, respectively). In a comprehensive assessment of domain versatility, repeat domains have been shown to typically have a low versatility (Weiner et al. 2008). However, there some repeat domains are an exception to this trend, including the WD40 and beta-beta-alpha zinc finger domains (Weiner et al. 2008; Stirnimann et al. 2010). This is consistent with the first explanation for why different TreeFamRep families have the same repeat domains, mentioned above.

We calculated the coefficient of variation (CV), defined as the standard deviation divided by the mean, for the average numbers of repeat units within TreeFamRep families that belong to the same SUPERFAMILY domain (supplementary fig. S4*C*, Supplementary Material online). The distribution of CV values for the average numbers of repeat units within families that belong to the same SUPERFAMILY domain has a median of 0.27 (supplementary fig. S4*C*, Supplementary Material online). Given that the CV for the entire TreeFamRep set is 0.7, this indicates that the variation of repeat unit numbers is typically comparatively small between families that map to the same SUPERFAMILY domain. This suggests that strong variations in length are not the reason for why proteins with the same repeat domain are sometimes assigned to different TreeFam families.

To test if repeat indel rates are conserved among TreeFamRep families that map to the same SUPERFAMILY domain, we also used the CV. The distribution of CV values between families that map to the same SUPERFAMILY domain has a median of 1.19 (supplementary fig. S4*B*, Supplementary Material online) compared with a CV of 1. 82 for the entire TreeFamRep dataset. This means that indel rates are highly variable even between different TreeFamRep families that belong to the same SUPERFAMILY domain, but they are typically much more similar than they would be for a random sample from the TreeFamRep set.

Some Domain Repeat Families Are Shaped by Concerted Evolution

Now that we have provided some explanations for the variance in domain repeat indel rates, we will focus on the TreeFamRepHind set, i.e., the 364 TreeFam domain repeat families with an indel rate $> Q3 + 1.5 \times IQR$.

Some protein repeats have indel rates so high that variants occur even within a population. Experimentally well-characterized examples are the *FLO1* gene in *Saccharomyces* (Verstrepen et al. 2005), *het-e* and *het-d* in fungi (Paoletti et al. 2007), and *dumpy* in *Drosophila* (Carmon et al. 2010). The underlying mechanisms causing such a high indel rate are recombinatorial processes such as unequal crossing over and DNA-repair after double-strand breakage. These processes not only lead to high indel rates, but they also cause regular homogenization of repeats through recombination of repeat units. As a consequence, repeats stay highly similar to each other within a protein and do not evolve independently from each other but rather in concert (Liao 1999). This phenomenon is called “concerted evolution”. In figure 5, we show a case study of concerted evolution for *Drosophila* CtripleX proteins and contrast it to the typical divergent mode of sequence evolution for domain repeats. We next test if repeat purity is particularly high in the TreeFamRepHind set. To calculate the repeat purity for each protein, we first extracted the sequences corresponding to the individual repeat units. An MSA for those sequences was then inferred and the all-vs.-all pairwise sequence identity (in %) between the repeat units was calculated. We further determined the average purity $\bar{P}$ and maximum purity $\hat{P}$ values for each repeat family and compared $\bar{P}$ and $\hat{P}$ in the TreeFamRepHind dataset to the corresponding values observed for the TreeFamRepLoind set. The results are shown in figure 6. $\bar{P}$ tends to be much higher in the TreeFamRepHind set, with a median = 35.9%, compared with a median of 27.1% in the TreeFamRepLoind set. When we look at $\hat{P}$, we find an even more significant increase in the *TreeFamRepHind* set, with a median purity = 59.5% compared with a median purity of 43.4% in all other domain repeat families. This difference between the TreeFamRepHind and TreeFamRepLoind sets is consistent with many proteins in the TreeFamRepHind set being shaped

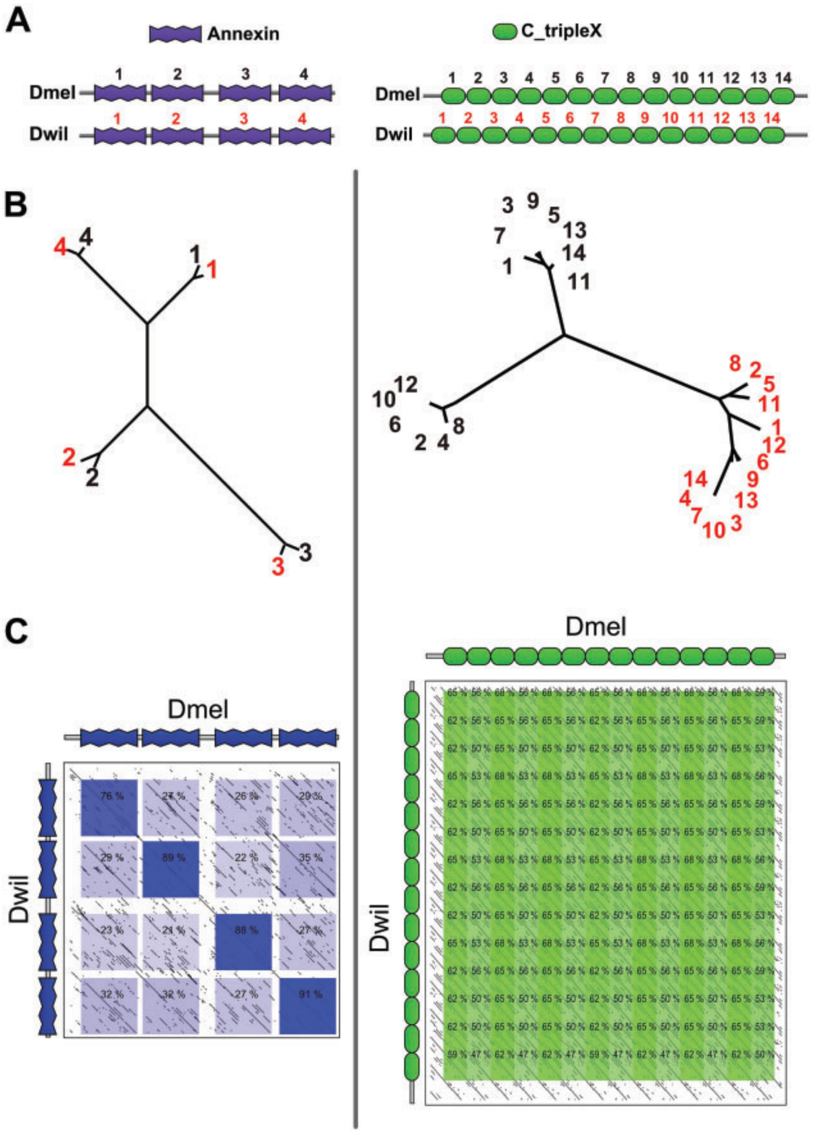

Fig. 5. Case studies illustrating divergent and concerted evolution of protein domain repeats. (*A*) Domain arrangements of two Annexin domain repeat proteins (left) and two CtripleX domain repeat proteins (right). In both cases, the two proteins correspond to one *Drosophila melanogaster (Dmel)* and one *Drosophila willistoni (Dwil)* protein. The individual domains repeat units are numbered. (*B*) Phylogenies for the domain sequences of the proteins shown in A. In the left example, each *Dmel* domain clusters together with its orthologous domain in the *Dwil* protein. In the right case however, all domains from one protein cluster together and are extremely similar or even completely identical to each other. (*C*) Domain–dot-plots for the two case studies. The rectangles within the dot-plots correspond to domain boundaries and color intensity scales with the sequences similarity in the region encompassed by the rectangle.

by concerted evolution, leading to high purity values through frequent recombination.

Since high purity values and high indel rates are consistent with concerted evolution but not sufficient to reliably conclude that concerted evolution indeed has shaped these proteins, we applied further tests. Specifically, we applied statistical tests for recombination. The pairwise homoplasy index (PHI) and neighborhood similarity score (NSS) tests as implemented in the Phi program (Bruen et al. 2006) were used to test all protein sequences in the TreeFamRepHind and TreeFamRepLoind datasets for recombination. For each protein, we extracted the sequences of the individual repeat units and used Muscle to align them. This protein alignment was translated into a cDNA alignment with PAL2NAL (Suyama et al. 2006). For each of those alignments, we then applied the PHI and NSS tests. We classify proteins with a domain repeat purity >50% and a *P*-value < 0.05 according to the PHI or NSS test as likely to have been shaped by concerted evolution. In total, we find 1027 (5%) repeats with significant evidence for recombination according to the PHI test in the TreeFamRepHind set compared with only 0.6% in the TreeFamRepLoind set. This is a highly significant enrichment (*P*-value < 2.2E−16, Fisher test). Note however that, even with this enrichment, concerted evolution is still rare and not characteristic of repeat families with high indel rates. Most likely, finding 1,027 proteins is an underestimate

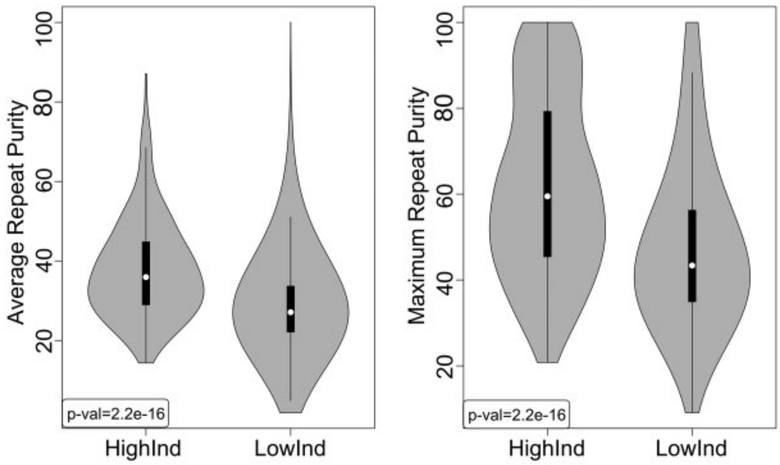

Fig. 6. Average repeat purity (left) and maximum repeat purity (right) in the HighInd repeat families (High Indrate) compared with all other repeat families (Low Indrate). *P*-values for both comparisons are $< 2.2e-16$ (Kolmogorov–Smirnov test).

because we are testing very short sequences for recombination. The resulting short cDNA alignments (compared to cDNA alignments of entire genes or even longer sequences) cannot contain much information indicative of recombination, even if recombination happens at high rates. Therefore, a high rate of false negatives is almost inevitable. Concerted evolution is the exception, not the norm. For the case study of concerted evolution we present in figure 5, concerted evolution is only detectable for the two *Drosophila* proteins shown, whereas no homolog in any of the closely related *Drosophila* species shows a comparable pattern.

Apparently only some proteins within a protein family have been shaped by concerted evolution whereas most have not and this raises the question what biological reasons underlie this apparent lineage-specificity of concerted evolution. To address this question, we annotated Gene Ontology (GO) terms to all proteins in the TreeFamRep set based on their Pfam and SUPERFAMILY domains (see "Materials and Methods" for details). Significant GO clusters for the Biological Process (BP) and Cellular Component (CC) GO categories are shown in supplementary figures S5 and S6, Supplementary Material online.

Interestingly, we observe an enrichment of GO processes and GO localizations which are relevant for various stress responses, neural development, and both innate and adaptive immunity (see supplementary fig. S5, Supplementary Material online). In plants, repeats occur in many proteins related to stress response (Sharma and Pandey 2015). It has been hypothesized that the presence of repeats in stress-related proteins confers an advantage because repeat expansions and contradictions create genetic variability that helps a species adapt to adverse environmental conditions (Sharma and Pandey 2015). Our results support this hypothesis. Repeats, particularly leucine-rich repeats (LRRs) and fibronectin repeats, have also been implicated in neural development in earlier studies (Dolan et al. 2007; de Wit et al. 2011; de Wit and Ghosh 2014). It has been observed that LRR families involved in neural development experienced many gene duplications in higher animals compared with worms (Dolan et al. 2007) and that the number of repeat units in those families is highly variable (Dolan et al. 2007).

Proteins related to host cell invasion and immune evasion of parasitic protists often have repeats, and a repeat purity significantly higher than that of free-living protists (Mendes et al. 2013). There is also evidence for recombination based repeat expansions in proteins which are relevant for virulence of pathogenic fungi (Butler et al. 2009). We therefore hypothesize that concerted evolution in metazoan domain repeats is sometimes triggered by host–parasite co-evolution. This would also explain why concerted evolution only seems to happen for a subset of the proteins in those families. In this case, the high repeat indel rates mediated by concerted evolution would only be expected for species where the protein repeat is currently participating in an evolutionary arms race with a pathogen.

Conclusions

Taken together, our results reveal some intricate relationships between structure, folding, function and genetic mechanisms affecting protein repeats. It appears that proteins with domain repeats with independently folding units have fewer constraints to vary their repeat numbers compared with proteins where repeat units fold cooperatively. This indicates that insertions and deletions into a stretch of cooperatively folding repeat units would usually have detrimental effects on a protein's native function. Consequently, such insertions are selected against whereas the duplication of an independently folding repeat unit is more likely to be tolerated. However, note that if selection does not weed out variants with repeat indels such indels need not be adaptive, they could well be largely neutral changes.

The higher indel rate in largely disordered domain repeats compared with largely structured ones may also be a consequence of the fact that an inserted disordered unit may not adversely affect the fold and function of the remaining protein sequence. It is surprising that the majority of the disordered repeat families which we could identify are not covered by domain databases such as Pfam and SUPERFAMILY although these repeats are well conserved across family members. This lower coverage by established domains is despite those repeats occurring in many different proteins with an average number of 36.1 proteins for the families in the TreeFamRepDis set. We hypothesize that the low coverage is at least partly caused by the sequence divergence of disordered regions, which is faster compared with ordered ones. Furthermore, many HMM models depend on X-ray structures in PDB in which disordered regions are strongly underrepresented (Peng et al. 2004). The sequence constraints to maintain disorder are not very specific because it is sufficient to keep a large fraction of charged and polar amino acids (van der Lee et al. 2014). This capability to maintain disorder against the backdrop of strong sequence variation is also apparent in linear motifs such as binding sites. Linear motifs are frequent in disordered regions and have only very few constraints which enforce high conservation on some positions in the sequence.

For the cases of concerted evolution, we found that they are strongly over-represented in the subset of families with exceptionally high indel rates. However, even with this over-representation, the phenomenon is still overall very rare. Since concerted evolution was most frequently observed in protein families characterized by high repeat indel rates, which are enriched in functions related to host–parasite co-evolution, stress response and neural development, we speculate that concerted evolution is sometimes a beneficial evolutionary strategy. Adaptation might speed up the variation of repeat unit number whereas attenuating sequence variations between repeat units. This would explain the positive correlation between repeat purity and repeat indel rates. For example, if a species starts to be involved in an evolutionary arms race with a parasite, concerted evolution could facilitate high repeat indel rates to increase genetic variability and in turn more efficiently evade the parasite.

It has been reported that repeat proteins which are related to binding in yeast and which vary their number of repeat units quite rapidly, such as flocculin, are located in hypervariable regions of the genome (Verstrepen et al. 2005). Based on our results, we speculate that binding proteins relevant for immunity and stress response in Metazoa are sometimes similarly located in genomic regions that are prone to recombination. Further studies are certainly required to investigate in more depth this intertwined relationship between genetic processes and selection pressure on protein repeats.

Materials and Methods

Annotation of Protein Repeats and Identification of Indels

We downloaded release 9 of the TreeFam dataset (Ruan et al. 2008), which corresponds to 15,316 multiple sequence alignments and phylogenetic trees for over a million protein sequences from 109 metazoan species. As outgroups, it contains the choanoflagellates *Monosiga brevicollis* and *Proterospongia sp.*, the Baker's yeast *Saccharomyces cerevisiae*, the fission yeast *Schizosaccharomyces pombe* and the cress *Arabidposis thaliana*). The family classification and alignment procedure is analogous to the Ensembl Compara pipeline [see Vilella et al. (2009) for a detailed explanation of the methodology].

We used the HHRepID program (Biegert and Söding 2008) to identify protein repeats. For each of those alignments, we scanned all proteins included in the alignment with HHRepID and stored information about the start and stop positions of all predicted repeats. For each repeat type predicted by HHRepID within one protein, we checked whether the predictions overlap with predicted repeats in other proteins in the alignment and joined them into one repeat family if they do. We then extracted the sequences of each individual repeat unit within a family and inferred a new multiple sequence alignment with Muscle (Edgar 2004) and used those alignments to construct HMMs with HHmake (Söding 2005). If there was more than one repeat family for the respective TreeFam alignment after this procedure, we used HHsearch (Söding 2005) for an all-vs.-all comparison of the corresponding HMMs and joined two families into one if they are significantly similar (HHsearch e-value $< 10E-5$).

For each repeat family, we then reformatted the repeat annotations into the T-Reks format readable by ProGraphMSA + TR and extracted all protein sequences that correspond to it from the original alignment. We provided those sequences and the repeat annotations for them to build a new MSA with ProGraphMSA + TR (using default parameters), and extracted the number of inferred indels and the estimated maximum likelihood distances between the sequences from the output files. Note that it is not possible to reliably distinguish between perfectly conserved repeat families where all proteins have the same number of repeat units that can be aligned to each other in a MSA and those that are only very well but not perfectly conserved. Repeat deletions could be falsely inferred because the protein sequence is incomplete. Particularly for proteins encoded by genes with a complex exon–intron structure and long introns, it can happen that gene prediction pipelines miss some exons or falsely predict the start or stop codons, which would lead to falsely predicted repeat deletions (Schüler et al. 2015).

Calculation of Repeat Purity

For each repeat containing protein, we extracted the sequences corresponding to the individual repeat units and aligned them with Muscle (Edgar 2004). We then used T-Coffee (Notredame et al. 2000) to infer the all-vs.-all pairwise sequence identity matrix of this alignment and calculated the average sequence identity between the sequences.

Comparison to Known Protein Domains

To evaluate whether the repeat array corresponds to known repetitive protein domains, we used the HHmake and HHsearch programs from the HHsuite (Söding 2005). We transformed the alignment of repeat sequences for each repeat families into a HMM with hh and scanned it against the HMM libraries of the SUPERFAMILY (Wilson et al. 2009) and Pfam (Punta et al. 2012) databases. For each repeat family, we extracted the hit with the most significant e-value, and if no hit could be detected (e-value < 10E−5), we considered the family to be unknown.

Annotation of Disorder and ELM Motifs

We used the ANCHOR program (Dosztányi et al. 2009) to calculate the IUPred score for each amino acid within a protein and to detect regions prone to undergo disorder to order transitions. Regions with an IUPred score >0.5 were annotated as "disordered". To identify instances of ELM motifs, we downloaded the most recent version of the ELM database (as of November 15th 2015) and translated them into Perl regular expressions. Protein regions matching those regular expressions were annotated as belonging to the corresponding ELM motif.

Testing for Recombination in Protein Repeats

We used the pairwise homoplasy index (PHI) and neighborhood similarity score (NSS) tests (Bruen et al. 2006) to test for recombination in alignments of domain repeats. The alignments necessary for this test were produced by running the Muscle alignment program (Edgar 2004) on the repeat sequences corresponding to each set of repeat sequences contained within one protein. These alignments were translated to DNA alignments based on their respective cDNA sequences with the PAL2NAL program (Suyama et al. 2006). These alignments were used as input for the PHI program to calculate the p-values based on the PHI and NSS tests.

Partial Correlations

A normal Pearson or Spearman correlation test cannot distinguish between two variables being directly correlated to each other and two variables being indirectly correlated to each other through covariation with a third variable. To account for that, we also calculated partial correlations with an R script developed by Kim and Yi (2007). We use those partial correlations to measure the degree of independent association between a variable and the indel rate whereas controlling for the effects of the other variables we consider in this study.

Gene Ontology Analysis

Go terms that are significantly over-represented in the HighInd set compared with all repeat proteins have been determined with TopGO (Alexa et al. 2006). Out of the available algorithms implemented in the TopGO package, the parent–child algorithm (Grossmann et al., 2007) has been chosen. The list of significantly enriched GO terms (*P*-value <0.01) has been filtered with REVIGO (Supek et al. 2011) using the SimRel semantic similarity measure and a similarity cutoff of 0.5.

Supplementary Material

Supplementary Figures S1–S5 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Author Contributions

A.S. performed all computational analyses and drafted the manuscript. Both authors conceived the study, interpreted results, wrote the manuscript and approved of the final version.

Acknowledgments

The authors thank E. Dohmen for assembling a protein sequence dataset corresponding to the RepeatsDB annotations. A.S. was supported by the Volkswagen Foundation Grant I/84830 and a DFG grant BO2544/4-1 to E.B-B. E.B-B's. ORCID is 0000-0002-1826-3576, ResID is A-1563-2013.

References

Alexa A, Rahnenführer J, Lengauer T. 2006. Improved scoring of functional groups from gene expression data by decorrelating go graph structure. *Bioinformatics* 22:1600–1607.

Andrade MA, Perez-Iratxeta C, Ponting CP. 2001. Protein repeats: structures, functions, and evolution. *J Struct Biol*. 134:117–131.

Apic G, Gough J, Teichmann SA. 2001. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J Mol Biol*. 310:311–325.

Apic G, Huber W, Teichmann SA. 2003. Multi-domain protein families and domain pairs: comparison with known structures and a random model of domain recombination. *J Struct Funct Genomics* 4:67–78.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The protein data bank. *Nucleic Acids Res*. 28:235–242.

Biegert A, Söding J. 2008. De novo identification of highly diverged protein repeats by probabilistic consistency. *Bioinformatics* 24:807–814.

Björklund AK, Ekman D, Elofsson A. 2006. Expansion of protein domain repeats. *PLoS Comput Biol*. 2:e114.

Björklund AK, Ekman D, Light S, Frey-Skött J, Elofsson A. 2005. Domain rearrangements in protein evolution. *J Mol Biol*. 353:911–923.

Bornberg-Bauer E, Albà MM. 2013. Dynamics and adaptive benefits of modular protein evolution. *Curr Opin Struct Biol*. 23:459–466.

Bornberg-Bauer E, Huylmans AK, Sikosek T. 2010. How do new proteins arise? *Curr Opin Struct Biol*. 20:390–396.

Brinkmann B, Klintschar M, Neuhuber F, Hühne J, Rolf B. 1998. Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. *Am J Hum Genet*. 62:1408–1415.

Bruen TC, Philippe H, Bryant D. 2006. A simple and robust statistical test for detecting the presence of recombination. *Genetics* 172:2665–2681.

Butler G, Rasmussen MD, Lin MF, Santos MA, Sakthikumar S, Munro CA, Rheinbay E, Grabherr M, Forche A, Reedy JL, et al. 2009. Evolution of pathogenicity and sexual reproduction in eight candida genomes. *Nature* 459:657–662.

Carmon A, Larson M, Wayne M, MacIntyre R. 2010. The rate of unequal crossing over in the dumpy gene from *Drosophila melanogaster*. *J Mol Evol*. 70:260–265.

Chevanne D, Saupe SJ, Clavé C, Paoletti M. 2010. WD-repeat instability and diversification of the *Podospora anserina* hnwd non-self recognition gene family. *BMC Evol Biol*. 10:134.

Davey NE, Van Roey K, Weatheritt RJ, Toedt G, Uyar B, Altenberg B, Budd A, Diella F, Dinkel H, Gibson TJ. 2012. Attributes of short linear motifs. *Mol Biosyst*. 8:268–281.

de Wit J, Ghosh A. 2014. Control of neural circuit formation by leucine-rich repeat proteins. *Trends Neurosci*. 37:539–550.

de Wit J, Hong W, Luo L, Ghosh A. 2011. Role of leucine-rich repeat proteins in the development and function of neural circuits. *Annu Rev Cell Dev Biol*. 27:697–729.

Di Domenico T, Potenza E, Walsh I, Parra RG, Giollo M, Minervini G, Piovesan D, Ihsan A, Ferrari C, Kajava AV, Tosatto SC. 2014. RepeatsDB: a database of tandem repeat protein structures. *Nucleic Acids Res*. 42:D352–D357.

Dinkel H, Michael S, Weatheritt RJ, Davey NE, Van Roey K, Altenberg B, Toedt G, Uyar B, Seiler M, Budd A, et al. 2012. ELM–the database of eukaryotic linear motifs. *Nucleic Acids Res*. 40:D242–D251.

Dolan J, Walshe K, Alsbury S, Hokamp K, O'Keeffe S, Okafuji T, Miller SFC, Tear G, Mitchell KJ. 2007. The extracellular leucine-rich repeat SUPERFAMILY; a comparative survey and analysis of evolutionary relationships and expression patterns. *BMC Genomics* 8:320.

Dosztányi Z, Csizmók V, Tompa P, Simon I. 2005. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J Mol Biol*. 347:827–839.

Dosztányi Z, Mészáros B, Simon I. 2009. ANCHOR: web server for predicting protein binding regions in disordered proteins. *Bioinformatics* 25:2745–2746.

Dosztányi Z, Mszros B, Simon I. 2010. Bioinformatical approaches to characterize intrinsically disordered/unstructured proteins. *Brief Bioinform*. 11:225–243.

Dyson HJ, Wright PE. 2002. Coupling of folding and binding for unstructured proteins. *Curr Opin Struct Biol*. 12:54–60.

Eddy SR. 2011. Accelerated profile HMM searches. *PLoS Comput Biol*. 7:e1002195.

Edgar RC. 2004. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 32:1792–1797.

Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, et al. 2010. The Pfam protein families database. *Nucleic Acids Res*. 38:D211–D222.

Forslund K, Sonnhammer ELL. 2012. Evolution of protein domain architectures. *Methods Mol Biol*. 856:187–216.

Gemayel R, Vinces MD, Legendre M, Verstrepen KJ. 2010. Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annu Rev Genet*. 44:445–477.

Grossmann S, Bauer S, Robinson PN, Vingron M. 2007. Improved detection of overrepresentation of gene-ontology annotations with parent child analysis. *Bioinformatics* 23:3024–3031.

Grove TZ, Cortajarena AL, Regan L. 2008. Ligand binding by repeat proteins: natural and designed. *Curr Opin Struct Biol*. 18:507–515.

Howard J, Bechstedt S. 2004. Hypothesis: a helix of ankyrin repeats of the NOMPC-TRP ion channel is the gating spring of mechanoreceptors. *Curr Biol*. 14:R224–R226.

Iakoucheva LM, Radivojac P, Brown CJ, O'Connor TR, Sikes JG, Obradovic Z, Dunker AK. 2004. The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res*. 32:1037–1049.

Jernigan KK, Bordenstein SR. 2015. Tandem-repeat protein domains across the tree of life. *Peer J*. 3:e732.

Jorda J, Baudrand T, Kajava AV. 2012. Prdb: Protein repeat database. *Proteomics* 12:1333–1336.

Jorda J, Xue B, Uversky VN, Kajava AV. 2010. Protein tandem repeats—the more perfect, the less structured. *FEBS J* 277:2673–2682.

Kajava AV. 2001. Review: proteins with repeated sequence–structural prediction and modeling. *J Struct Biol*. 134:132–144.

Kajava AV. 2012. Tandem repeats in proteins: from sequence to structure. *J Struct Biol*. 179:279–288.

Kersting AR, Bornberg-Bauer E, Moore AD, Grath S. 2012. Dynamics and adaptive benefits of protein domain emergence and arrangements during plant genome evolution. *Genome Biol Evol*. 4:316–329.

Kim SH, Yi SV. 2007. Understanding relationship between sequence and functional evolution in yeast proteins. *Genetica* 131:151–156.

Kumar S. 2005. Molecular clocks: four decades of evolution. *Nat Rev Genet*. 6:654–662.

Kummerfeld SK, Teichmann SA. 2005. Relative rates of gene fusion and fission in multi-domain proteins. *Trends Genet*. 21:25–30.

Kummerfeld SK, Teichmann SA. 2009. Protein domain organisation: adding order. *BMC Bioinformatics* 10:39.

Laurino P, Tóth-Petróczy G, Meana-Pañeda R, Lin W, Truhlar DG, Tawfik DS. 2016. An ancient fingerprint indicates the common ancestry of Rossmann-fold enzymes utilizing different ribose-based cofactors. *PLoS Biol*. 14:e1002396.

Legendre M, Pochet N, Pak T, Verstrepen KJ. 2007. Sequence-based estimation of minisatellite and microsatellite repeat variability. *Genome Res*. 17:1787–1796.

Liang X, Madrid J, Gärtner R, Verbavatz JM, Schiklenk C, Wilsch-Bräuninger M, Bogdanova A, Stenger F, Voigt A, Howard J. 2013. A NOMPC-dependent membrane-microtubule connector is a candidate for the gating spring in fly mechanoreceptors. *Curr Biol*. 23:755–763.

Liao D. 1999. Concerted evolution: molecular mechanism and biological implications. *Am J Hum Genet*. 64:24–30.

Light S, Sagit R, Ekman D, Elofsson A. 2013a. Long indels are disordered: a study of disorder and indels in homologous eukaryotic proteins. *Biochim Biophys Acta*. 1834:890–897.

Light S, Sagit R, Sachenkova O, Ekman D, Elofsson A. 2013b. Protein expansion is primarily due to indels in intrinsically disordered regions. *Mol Biol Evol*. 30:2645–2653.

Lupas AN, Ponting CP, Russell RB. 2001. On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? *J Struct Biol*. 134:191–203.

Marcotte EM, Pellegrini M, Yeates TO, Eisenberg D. 1999. A census of protein repeats. *J Mol Biol*. 293:151–160.

Mendes TAO, Lobo FP, Rodrigues TS, Rodrigues-Luiz GF, daRocha WD, Fujiwara RT, Teixeira SMR, Bartholomeu DC. 2013. Repeat-enriched proteins are related to host cell invasion and immune evasion in parasitic protozoa. *Mol Biol Evol*. 30:951–963.

Mirkin SM. 2006. DNA structures, repeat expansions and human hereditary disorders. *Curr Opin Struct Biol*. 16:351–358.

Mirkin SM. 2007. Expandable DNA repeats and human disease. *Nature* 447:932–940.

Moore AD, Björklund AK, Ekman D, Bornberg-Bauer E, Elofsson A. 2008. Arrangements in the modular evolution of proteins. *Trends Biochem Sci*. 33:444–451.

Moore AD, Bornberg-Bauer E. 2012. The dynamics and evolutionary potential of domain loss and emergence. *Mol Biol Evol*. 29:787–796.

Moore AD, Grath S, Schüler A, Huylmans AK, Bornberg-Bauer E. 2013. Quantification and functional analysis of modular protein evolution in a dense phylogenetic tree. *Biochim Biophys Acta*. 1834:898–907.

Morel G, Sterck L, Swennen D, Marcet-Houben M, Onesime D, Levasseur A, Jacques N, Mallet S, Couloux A, Labadie K, et al. 2015. Differential gene retention as an evolutionary mechanism to generate biodiversity and adaptation in yeasts. *Sci Rep*. 5:11571.

Mosavi LK, Cammett TJ, Desrosiers DC, Peng ZY. 2004. The ankyrin repeat as molecular architecture for protein recognition. *Protein Sci*. 13:1435–1448.

Mészáros B, Simon I, Dosztányi Z. 2009. Prediction of protein binding regions in disordered proteins. *PLoS Comput Biol*. 5:e1000376.

Notredame C, Higgins DG, Heringa J. 2000. T-coffee: a novel method for fast and accurate multiple sequence alignment. *J Mol Biol*. 302:205–217.

Paoletti M, Saupe SJ, Clavé C. 2007. Genesis of a fungal non-self recognition repertoire. *PLoS One* 2:e283.

Pawson T, Nash P. 2003. Assembly of cell regulatory systems through protein interaction domains. *Science* 300:445–452.

Pei J, Grishin NV. 2001. Al2co: calculation of positional conservation in a protein sequence alignment. *Bioinformatics* 17:700–712.

Pellegrini M, Renda ME, Vecchio A. 2012. Ab initio detection of fuzzy amino acid tandem repeats in protein sequences. *BMC Bioinformatics* 13 Suppl 3:S8.

Peng K, Obradovic Z, Vucetic S. 2004. Exploring bias in the protein data bank using contrast classifiers. *Pac Symp Biocomput*. 435–446.

Ponting CP, Mott R, Bork P, Copley RR. 2001. Novel protein domains and repeats in *Drosophila melanogaster*: insights into structure, function, and evolution. *Genome Res*. 11:1996–2008.

Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, et al. 2012. The Pfam protein families database. *Nucleic Acids Res*. 40:D290–D301.

Reshef D, Itzhaki Z, Schueler-Furman O. 2010. Increased sequence conservation of domain repeats in prokaryotic proteins. *Trends Genet*. 26:383–387.

Ruan J, Li H, Chen Z, Coghlan A, Coin LJ, Guo Y, Hériché JK, Hu Y, Kristiansen K, Li R, et al. 2008. Treefam: 2008 update. *Nucleic Acids Res*. 36:D735–D740.

Schaper E, Anisimova M. 2014. The evolution and function of protein tandem repeats in plants. *New Phytol*. 206(1):397–410.

Schaper E, Gascuel O, Anisimova M. 2014. Deep conservation of human protein tandem repeats within the eukaryotes. *Mol Biol Evol*. 31:1132–1148.

Schüler A, Schmitz G, Reft A, Özbek S, Thurm U, Bornberg-Bauer E. 2015. The rise and fall of TRP-N, an ancient family of mechanogated ion channels, in metazoa. *Genome Biol Evol*. 7:1713–1727.

Söding J. 2005. Protein homology detection by HMM-HMM comparison. *Bioinformatics* 21:951–960.

Sharma M, Pandey GK. 2015. Expansion and function of repeat domain proteins during stress and development in plants. *Front Plant Sci*. 6:1218.

Shiu SH, Byrnes JK, Pan R, Zhang P, Li WH. 2006. Role of positive selection in the retention of duplicate genes in mammalian genomes. *Proc Natl Acad Sci U S A*. 103:2232–2236.

Stirnimann CU, Petsalaki E, Russell RB, Müller CW. 2010. WD40 proteins propel cellular networks. *Trends Biochem Sci*. 35:565–574.

Street TO, Rose GD, Barrick D. 2006. The role of introns in repeat protein gene formation. *J Mol Biol*. 360:258–266.

Supek F, Bošnjak M, Škunca N, Šmuc T. 2011. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One* 6:e21800.

Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res*. 34:W609–W612.

Szalkowski AM, Anisimova M. 2011. Markov models of amino acid substitution to study proteins with intrinsically disordered regions. *PLoS One* 6:e20488.

Szalkowski AM, Anisimova M. 2013. Graph-based modeling of tandem repeats improves global multiple sequence alignment. *Nucleic Acids Res*. 41:e162.

Toll-Riera M, Albà MM. 2013. Emergence of novel domains in proteins. *BMC Evol Biol*. 13:47.

Tompa P. 2003. Intrinsically unstructured proteins evolve by repeat expansion. *Bioessays* 25:847–855.

Tzelgov J, Henik A. 1991. Suppression situations in psychological research: definitions, implications, and applications. *Psychol Bull*. 109(3):524–536.

van der Lee R, Buljan M, Lang B, Weatheritt RJ, Daughdrill GW, Dunker AK, Fuxreiter M, Gough J, Gsponer J, Jones DT, et al. 2014. Classification of intrinsically disordered regions and proteins. *Chem Rev*. 114:6589–6631.

Verstrepen KJ, Jansen A, Lewitter F, Fink GR. 2005. Intragenic tandem repeats generate functional variability. *Nat Genet*. 37:986–990.

Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. 2009. Ensemblcompara genetrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res*. 19:327–335.

Wang M, Caetano-Anollés G. 2009. The evolutionary mechanics of domain organization in proteomes and the rise of modularity in the protein world. *Structure* 17:66–78.

Weiner J, 3rd, Beaussart F, Bornberg-Bauer E. 2006. Domain deletions and substitutions in the modular protein evolution. *FEBS J*. 273:2037–2047.

Weiner J, 3rd, Bornberg-Bauer E. 2006. Evolution of circular permutations in multidomain proteins. *Mol Biol Evol*. 23:734–743.

Weiner J, 3rd, Moore AD, Bornberg-Bauer E. 2008. Just how versatile are domains? *BMC Evol Biol*. 8:285.

Wilson D, Pethica R, Zhou Y, Talbot C, Vogel C, Madera M, Chothia C, Gough J. 2009. SUPERFAMILY-sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Res*. 37:D380–D386.

Wright CF, Teichmann SA, Clarke J, Dobson CM. 2005. The importance of sequence diversity in the aggregation and evolution of proteins. *Nature* 438:878–881.

Ye Y, Godzik A. 2004. Comparative analysis of protein domain organization. *Genome Res*. 14:343–353.

Zmasek CM, Godzik A. 2011. Strong functional patterns in the evolution of eukaryotic genomes revealed by the reconstruction of ancestral protein domain repertoires. *Genome Biol*. 12:R4.

Zmasek CM, Godzik A. 2012. This déjà vu feeling—analysis of multidomain protein evolution in eukaryotic genomes. *PLoS Comput Biol*. 8:e1002701.