Human capacitance to dosage imbalance: Coping with inefficient selection

Ariel Fernández and Jianping Chen

Genome Res. 2009 19: 2185-2192 originally published online October 9, 2009
Access the most recent version at doi:10.1101/gr.094441.109

Letter

Human capacitance to dosage imbalance: Coping with inefficient selection

Ariel Fernández[1] and Jianping Chen

Department of Bioengineering, Rice University, Houston, Texas 77005, USA

Proteins rely on associations to improve packing quality and thus maintain structural integrity. This makes packing deficiency a likely determinant of dosage sensitivity, that is, of the fitness impact of concentration imbalances relative to the stoichiometry of the protein complexes. This hypothesis was validated by examining evolution-related dosage imbalances: Duplicates of genes encoding for deficiently packed proteins are less likely to be retained than genes coding for well-packed proteins. This selection pressure is apparent in unicellular organisms, but is mitigated in higher eukaryotes. In human, this effect reveals a capacitance toward dosage imbalance. This capacitance is not expected in organisms with larger population size, where evolutionary forces are more efficient at promoting adaptive functional innovation and purifying selection, thus curbing the concentration imbalance arising from gene duplication. By examining miRNA target dissimilarities within human gene families, we show that the capacitance is operative at a post-transcriptional regulatory level: The higher the packing deficiency of a protein, the more likely that its paralogs will be dissimilarly targeted by miRNA to mitigate dosage imbalance. For families with low capacitance, paralog sequence divergence and family size correlate tightly with packing deficiency, just like in unicellular eukaryotes. Thus, a major component of human tolerance toward dosage imbalances is rooted in the paralog-discriminating capacity of miRNA regulation. The results may clarify the evolutionary etiology of aggregation-related diseases, since aggregation is often promoted by overexpression (a dosage imbalance) and aggregation propensity is associated with extreme packing deficiency.

[Supplemental material is available online at http://www.genome.org.]

Dosage imbalances occur when protein concentration levels at specific locations in tissues or metabolic/developmental phases do not fit the stoichiometry of the complexes in which the proteins are involved (Veitia 2002, 2004; Papp et al. 2003; Kondrashov and Koonin 2004; Liang et al. 2008). The complexes may be transient, adventitious, or obligatory with regard to maintaining the structural integrity of the protein (Veitia 2002, 2004; Fernández and Scheraga 2003), and hence the effects of the imbalances may vary widely. Therefore, dosage sensitivity, that is, the impact of dosage imbalances on fitness, must be influenced not only by whether the protein is part of a complex, as previously observed (Papp et al. 2003), but also by the extent of reliance of the protein on its binding partners to maintain structural integrity and functional competence (Fernández and Scheraga 2003; Fernández et al. 2003; Fernández 2004; Pietrosemoli et al. 2007).

While overexpression, gene duplication, misfolding, and self-aggregation may all cause dosage imbalance, the structural or molecular properties determining the magnitude of the resulting effects remain largely unknown. For example, as we focus on gene duplication, we notice that paralog proteins, identical when they initially diverge, are subject to higher or lower selection pressure depending on their dosage sensitivity (Liang et al. 2008).

Recently, cross-examination of genetic and structural information revealed that a molecular quantifier of dosage sensitivity is the packing deficiency of the protein (Liang et al. 2008), a measure of its reliance on binding partners to maintain the integrity of the native fold (Fernández and Scheraga 2003; Fernández et al. 2003; Fernández 2004). Thus, a deficiently packed protein is more likely to be engaged in an obligatory complex (Fernández and Scheraga 2003; Pietrosemoli et al. 2007) and its concentration imbalances, relative to the complex stoichiometry, are likely to impact fitness more than those of a well-packed protein (Liang et al. 2008).

The packing quality of a protein refers to the capacity of the soluble fold to shield backbone hydrogen bonds (BHBs) from hydration by "wrapping" them with side-chain nonpolar groups (Pietrosemoli et al. 2007). Poorly wrapped intramolecular hydrogen bonds represent structural vulnerabilities (Fernández et al. 2003), since their exposure to solvent promotes hydration of backbone amides and carbonyls, a process that eventually triggers the dismantling of structure (Fernández et al. 2003; Fernández 2004). These exposed bonds are also sticky (Fernández and Scott 2003), since further removal of surrounding water upon protein association enhances the underlying electrostatic interaction and stabilizes the bond (Fernández 2004). The stabilization arises because the nonbonded state, with amide and carbonyl hindered from hydration, becomes unstable. The stabilization has been experimentally shown to contribute, on average, ~3.9 kJ/mol per packing improvement of a protein hydrogen bond (Fernández and Scott 2003). Thus, besides representing local weaknesses in the protein structure, solvent-exposed backbone hydrogen bonds (SEBHs) are determinants of protein associations (Fernández and Scheraga 2003). This is also evidenced by their higher abundance in all protein–protein interfaces from PDB complexes, with SEBH abundance being between 1.5 and 10 times higher than surface average in PDB complexes (Pietrosemoli et al. 2007).

Thus, we operationally define packing deficiency as the percentage of BHBs that are SEBHs. Experimental and structural bioinformatics evidence reveals that a defectively packed protein is not only reliant on binding partnerships to maintain its structural integrity, but actually promotes such interactions (Fernández and Scott 2003). Hence, a deficiently packed protein, rich in SEBHs, is likely to possess higher dosage sensitivity than a well-packed protein (Liang et al. 2008), a hypothesis tested in this work.

[1]Corresponding author.
E-mail arifer@rice.edu; fax (713) 348-3699.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.094441.109.

In unicellular organisms, the packing quality of soluble gene products correlates with the number of paralogs or family size (Liang et al. 2008). However, this correlation becomes less significant in higher eukaryotes. Thus, paralog survival is dependent on the packing quality of protein structure with $P < 10^{-16}$ in *Escherichia coli* or yeast (*Saccharomyces cerevisiae*), and $P < 6.7 \times 10^{-3}$ in human (Wilcoxon rank-sum test) (Liang et al. 2008). This contrast between simple and complex organisms is hard to interpret due to wide differences at the proteome level. However, alternative measures point to a similar trend (Liang et al. 2008). For example, the average difference in packing deficiency between singletons and duplicate genes is 21% in *E. coli*, 13% in *S. cerevisiae*, 8% in worm (*Caenorhabditis elegans*), and ~6% in human (*Homo sapiens*), fly (*Drosophila melanogaster*), and thale cress (*Arabidopsis thaliana*).

In human, this insensitivity to dosage imbalance may be attributed in part to selection inefficiency arising from smaller population size (Lynch and Conery 2003), implying that the selection pressure exerted on paralogs of deficiently packed proteins has simply not become operative. Other "escape routes" to dosage imbalance in higher eukaryotes, such as alternative splicing, promoter polymorphism, or allosteric protein oligomerization may also be important (Liang et al. 2008). Alternatively, the higher complexity of expression regulation in higher eukaryotes may introduce a tolerance to dosage imbalance not found in unicellular organisms. This work explores and validates this latter possibility focusing on evolution-related dosage imbalances.

We first assess the selection pressure on gene duplicates exerted, as paralogs are coexpressed at the mRNA (messenger RNA) level and hence are likely to compete for their interactive partners. Then we relate packing deficiency with differences in post-transcriptional regulation patterns within families. Thus, we investigate how differences in miRNA-target patterns (Bartel 2009), telling apart paralogs through different patterns of translational repression, impinge on the selection pressure on duplicate genes by mitigating dosage imbalances. In human these patterns will be shown to be significantly dissimilar across paralogs of poorly packed proteins, while nearly coincident across paralogs of well-packed proteins, thus underscoring a means to buffer dosage imbalance effects arising from gene duplication.

This miRNA-based capacitance is not expected to be nearly as significant in species with larger effective population size due to the higher efficiency of evolutionary forces in such organisms when compared with human (Lynch and Conery 2003). Thus, the selection pressure affecting the retention of gene duplicates is likely to be more efficient in these organisms promoting adaptation through functional innovation or purifying selection. In this regard, the capacitance of *C. elegans*, *D. melanogaster*, and *A. thaliana* would provide adequate controls. However, the scarcity of combined structural, genetic, expression, and regulatory information (Supplemental material), even for these well-studied species, precludes statistically significant conclusions at this juncture. The biggest limitation in the combined data collection arises from the dearth of structural information on paralogs inferred from PDB representation for these species when compared with human.

To study the effects of unmitigated selection pressure, we shall focus on human families with significant paralog coexpression and miRNA-target coincidence (low capacitance). Under these presumed "no escape route" conditions, we shall show that the selection pressure imposed by dosage imbalance correlates well with packing deficiency, just as was previously observed in unicellular eukaryotes (Liang et al. 2008).

Taken together, these results support the claim that protein packing quality is a determinant of dosage sensitivity and uphold the view that resilience to dosage imbalance is achieved in human by diversifying miRNA-regulatory patterns across paralogs.

Results

Human tolerance to dosage imbalance

The *packing deficiency* (ν) of a soluble protein is defined by the percentage of BHBs that are solvent-exposed (SEBHs) (Fernández and Scheraga 2003; Fernández 2004; Pietrosemoli et al. 2007). SEBHs are vulnerable hydrogen bonds, protected by an insufficient number of side-chain nonpolar groups (Fernández and Scheraga 2003; Fernández 2004) (see Methods) and hence, prone to be disrupted by further hydration of the amide and carbonyl.

Packing deficiency and SEBH patterns (Fig. 1) are computationally determined from structure coordinates by assessing the number of nonpolar groups within the microenvironment of the intramolecular hydrogen bonds (see Methods). Through binding partnerships, soluble proteins may further protect their BHBs and improve packing quality by increasing the number of nonpolar groups in their hydrogen-bond microenvironments (Fernández and Scheraga 2003; Fernández 2004). Hence, the extent of intermolecular protection determines whether the complex is obligatory, ephemeral, or adventitious (Fernández and Scheraga 2003), and thus packing quality has been recognized as an important factor in determining dosage sensitivity (Liang et al. 2008).

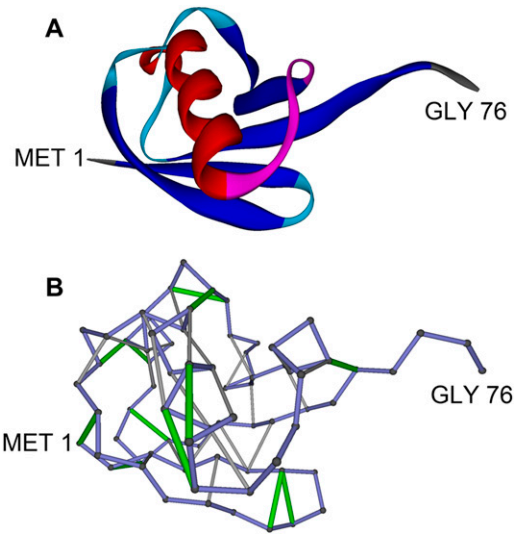

Figure 1. Packing deficiency of human ubiquitin. (*A*) Ribbon representation of ubiquitin structure (PDB. 1UBI) (Ramage et al. 1994). (*B*) Ubiquitin SEBH-pattern (solvent-exposed backbone hydrogen bond). The protein backbone is shown as virtual bonds (blue) joining α-carbons. Light-gray segments joining α-carbons represent well-wrapped (protected) backbone hydrogen bonds (BHBs) pairing residues defined by their α-carbon positions, and green segments represent SEBHs. The solvent-exposure extent of a hydrogen bond is determined from atomic coordinates (Fernández 2004; Pietrosemoli et al. 2007) by calculating the number of nonpolar side-chain groups within its microenvironment (see Methods). SEBHs are those BHBs protected by an insufficient number of nonpolar groups as statistically defined in Methods. The packing deficiency ν, defined as the ratio of SEBHs to the overall number of BHBs, is 31.4% ($\nu = 11/35$).

Gene duplication introduces dosage imbalance and the resulting selection pressure on paralogs (Veitia 2002) appears to depend on the packing deficiency of the parental gene (Liang et al. 2008). This trend is clear in *E. coli* and *S. cerevisiae*, but not so apparent in higher eukaryotes (Liang et al. 2008). This observation suggests that expression dissimilarities at the mRNA level and at post-transcriptional levels may be exploited to separate paralogs and avoid competition for the binding partners of the parental gene. Thus, to study human capacitance to dosage imbalance arising from gene duplication, we examined families with paralog coexpression at the mRNA level (Su et al. 2004), and assessed post-transcriptional microRNA (miRNA) regulation patterns in relation to the packing quality of the proteins in the family.

To assess the role of miRNA regulation in the human capacitance to dosage imbalance, we selected human genes from an exhaustive set of 583 nonsingleton families for which genetic (Birney et al. 2006), evolutionary (Yang and Nielsen 2000), structural (Liang et al. 2008), expression (Su et al. 2004), and post-transcriptional (Lewis et al. 2005; Liang and Li 2007; Bartel 2009; Friedman et al. 2009) data are available for at least two paralogs (see Methods). We collected their mRNA-expression profiles (Su et al. 2004), and their putative miRNA-target patterns (Friedman et al. 2009). Putative conserved target sites in the 3′ UTR (untranslated region) of each gene for 156 conserved microRNA families were identified using TargetScanS (version 5.1). Thus, to determine coexpression and coregulation patterns across paralogs, each gene i was represented by two vectors:

1. A normalized mRNA-expression vector $\Phi_i/||\Phi_i||$, where the vector Φ_i has 73 entries indicating mRNA expression levels in 73 normal tissues (Su et al. 2004) and $||\Phi_i||$ is the norm of the vector.
2. A normalized miRNA vector $\Psi_i/||\Psi_i||$ of 156 entries representing the pattern of miRNA-related repression efficacy on gene i, with $||\Psi_i||$ = vector norm. The nth entry in ψ_i is $\Psi_i(n) = 1 - 2^{s(i,n)}$, where $s(i,n) \leq 0$ is the context score of conserved miRNA-binding site n in the 3′ UTR of gene i (Grimson et al. 2007). Thus, $\Psi_i(n) = 1$ indicates full repressive efficacy of the nth miRNA conserved site on gene i ($s(i, n) = -\infty$), while $\Psi_i(n) = 0$ (or $s(i, n) = 0$) indicates absolute lack of regulatory power (Methods).

Only paralogs that are significantly coexpressed are likely to produce dosage imbalances if the genes have not diverged significantly. Thus, similarities between mRNA expression profiles of two genes i, j will be assessed by the Pearson correlation coefficient $\eta(i, j)$ of their expression vectors Φ_i and Φ_j (Methods).

For paralogs with significant coexpression, a tolerance to dosage imbalance may still arise through differences in translational repression patterns. Thus, orthogonal miRNA-repression patterns for paralogs with high dosage sensitivity may introduce an escape route to the selection pressure introduced by the dosage imbalance. To test this hypothesis, we introduce the extent of miRNA-target coincidence $\tau(i,j)$, defined as the scalar (dot) product of the two miRNA-target vectors: $\tau(i,j) = \Psi_i/||\Psi_i|| \cdot \Psi_j/||\Psi_j||$.

To determine the dosage sensitivity we calculated the packing deficiency of each gene-encoded protein based on its PDB coordinates, if available. Otherwise, packing deficiency was determined based on homology-threaded structure coordinates adopting as templates PDB-reported paralogs with at least 40% sequence identity, i.e., within the reliability range (Aloy et al. 2003) (see Methods).

To assess the selection pressure imposed by dosage imbalance we first consider an exhaustive set of 457 nonsingleton human gene families with paralog coexpression at the mRNA level: $\langle\eta\rangle > 0$, $< >$ = family average. This condition is essential since paralogs expressed in different cell types cannot introduce dosage imbalance, regardless of their extent of identity. The families with paralog coexpression are selected to discern the factors that buffer dosage imbalance caused by gene duplication. Thus, the selection pressure may be assessed at the post-transcriptional level in terms of dissimilarities in miRNA-targeting patterns across paralogs.

The families with significant mRNA coexpression were deemed likely to generate dosage imbalance. To assess how these imbalances impinge on the degree of divergence in post-transcriptional repression patterns across paralogs, we must compare families with similar divergence time of gene duplicates. This is so since significant regulatory dissimilarities across paralogs may simply result from long divergence times. Thus, we adopt K_s, the synonymous nucleotide divergence (Yang and Nielsen 2000), as a proxy for divergence time (Gu et al. 2002) and bin human families with $\langle\eta\rangle > 0$ according to their respective maximum K_s over paralog pairs. Each class contains families whose duplicate divergence is located in time vis-à-vis particular speciation events. Thus, we construct four classes of human families with significantly coexpressed paralogs (Supplemental Table 1): class I: $K_s > 2.98$ (378 families); class II: $2.98 > K_s > 1.48$ (68 families); class III: $1.48 > K_s > 1.11$ (six families) and class IV: $K_s < 1.11$ (five families), in accord with the K_s values between human and orangutan (*Pongo pygmaeus*) ($K_s = 2.98$), human and gorilla (*Gorilla gorilla*) ($K_s = 1.48$), and human and chimpanzee (*Pan troglodytes*) ($K_s = 1.11$) (Chen and Li 2001). All K_s values are given as percentages.

The conservation-based reliability of miRNA site prediction (Bartel 2009) is the highest in class I and decreases with lower divergence times for duplicate genes. This is so since the condition: K_s (duplicate genes) > K_s (speciation) implies that orthologs of the paralog human genes are likely to be found in the diverging species (Gao and Innan 2004). Thus, paralogs for families in class I are likely to have orthologs in orangutan, gorilla, and chimpanzee, those in class II, only in gorilla and chimpanzee, etc.

Human families with paralog coexpression and the most reliable miRNA site inference (class I) exhibit a tight anti-correlation ($R^2 = 0.697$) between packing deficiency and miRNA-target coincidence (Fig. 2A): Paralogs with deficient packing are more likely to be localized separated from each other as dictated by their dissimilar miRNA-target patterns of post-transcriptional regulation: $\langle\tau\rangle \rightarrow 0$ as $\langle v\rangle \rightarrow$ maximum $\approx 58\%$. These disjoint localization patterns reduce paralog competition for binding partners, thereby buffering the evolution-related dosage imbalance. This result highlights the role of miRNA regulation as a capacitor for dosage imbalance.

An even tighter anti-correlation between packing deficiency and miRNA target coincidence is found for family class II ($R^2 = 0.792$; Fig. 2B). The slope of the linear fit obtained by the least-squares linear regression is now significantly larger in magnitude (-69.34 versus -57.54 for class I). This implies that for a fixed level of packing deficiency, a more effective buffer (lower miRNA target coincidence) is needed for the newer families (K_s-class II) than for the older ones (K_s-class I). This result is expected since a longer exposure of surviving paralogs to the selection pressure promoted by dosage imbalance is likely to promote a higher level of adaptation through functional divergence, and hence, as older paralogs become more differentiated, a capacitance to dosage imbalance becomes less necessary. The same trend is apparent as we examine class III (slope -72.53, $R^2 = 0.786$; Fig. 2C), although the scarcity of the data precludes a reliable statistical analysis. Class IV consists of

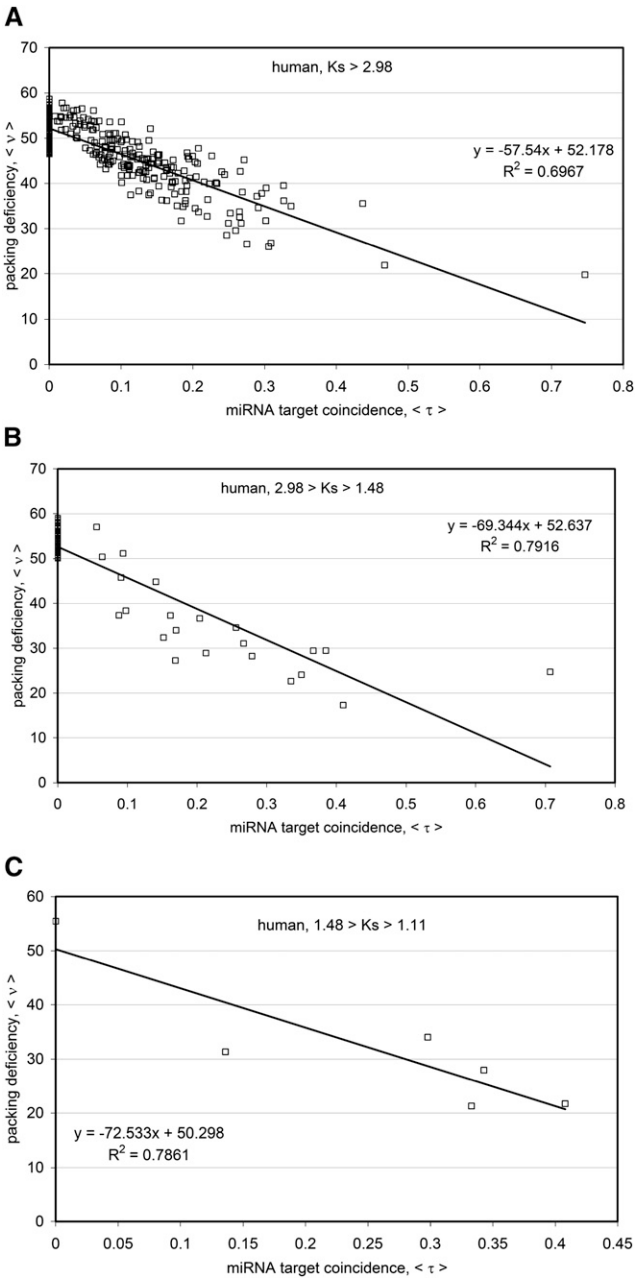

Figure 2. Anti-correlation between packing deficiency (ν) and miRNA target coincidence (τ) for human families in K_s-classes I (*A*), II (*B*), and III (*C*). The linear fits were obtained by least-squares linear regression.

only five families and hence no trend can be established, except that all families have zero miRNA target coincidence irrespective of their packing deficiency (Supplemental Table 2). This fact is clearly indicative of a pressing need to buffer dosage imbalances arising from duplicates that have not yet undergone sufficient functional differentiation.

The trends in terms of tighter $\nu-\tau$ anti-correlation and steeper slope as classes with lower K_s are considered (Fig. 2A–C) implies that a miRNA-based capacitance to dosage imbalance is more operative for younger families (classes II–IV versus class I). This result is compatible with the fact that selection pressure on more recent paralogs has had comparably less time to promote adaptation through functional divergence and hence duplication-related dosage imbalances are more significant than those in older families.

These results reveal that the human capacitance to dosage imbalance is, in part, required due to the inefficiency of the selection pressure (Lynch and Conery 2003) on duplicate genes, precluding sufficient differentiation over the evolutionary times of the latest speciations, thereby maintaining an evolutionarily related dosage imbalance.

The homo-oligomerization of protein domains can, in principle, become a confounding factor in the assessment of the role of packing deficiency as a quantifier of dosage sensitivity. This is so since oligomerization or allostery generally improves the packing quality of the domain (Fernández and Scheraga 2003). However, the Gene Ontology (GO) Database (http://www.geneontology.org) singles out only two of the 583 families under examination as confirmed to contain homo-oligomers: ENSF00000001444 (Diacylglycerol kinase delta protein) and ENSF00000001818 (Hypoxanthine-guanine phosphoribosyltransferase). Hence, these cases do not compromise the statistical inferences made in this work.

The results of Figure 2 imply that miRNA target dissimilarity across paralogs may be assimilated to a capacitance to dosage imbalance effects arising from gene duplication. The severity of such effects is in turn quantified by packing deficiency (ν): Dosage imbalances are less tolerated for deficiently packed proteins forcing paralogs to be localized separately from each other.

Selection pressure on gene duplicates at low capacitance

The picture presented above is further supported by examination of dosage imbalance effects for a subset of families with significant paralog mRNA-coexpression and high miRNA target coincidence. If this case indeed represents low capacitance, packing deficiency should be now tightly anti-correlated with paralog survival, precisely as found in unicellular organisms (Liang et al. 2008) and also tightly correlated with paralog sequence divergence as tolerance to dosage imbalance decreases with packing deficiency. The results of Figure 3 confirm this scenario. Hence, the selection pressure exerted on paralogs is not mitigated in these families, much like in unicellular eukaryotes (Liang et al. 2008). Furthermore, colocalized paralogs of a protein with extremely deficient packing should tend to diverge as much as possible to avoid competing with the parent protein for its obligatory binding partners, while paralogs of a well-packed protein are not subject to this constraint and hence their lesser divergence should reflect a lower selection pressure. This scenario is corroborated by the results shown in Figure 3A.

Paralog sequence divergence δ of a family is defined as the Hamming distance between paralog sequence pairs averaged over all pairs in a family. At low capacitance (signaled by $\langle\tau\rangle >> 0$ according to Fig. 2), selection pressure arising from duplication of deficiently packed genes is relatively high, as evidenced by the correlation between packing deficiency and paralog sequence divergence ($R^2 = 0.433$; Fig. 3A). By contrast, no significant correlation ($R^2 = 0.009$) is found for families endowed with maximum capacitance ($\langle\tau\rangle = 0$), pointing to an efficient curbing of dosage imbalance effects. The deviation in sequence divergence from the $\delta-\nu$ fitting trend found for families with low capacitance (Fig. 3A) can be attributed to family size (gene copy number) effects. The number of paralogs is a determinant of dosage-related purifying selection. Whatever the divergence topology within the family, dosage imbalance is dependent on family size since the number of surviving paralogs determines the extent of competition for the binding partners of the parental gene. This contribution is

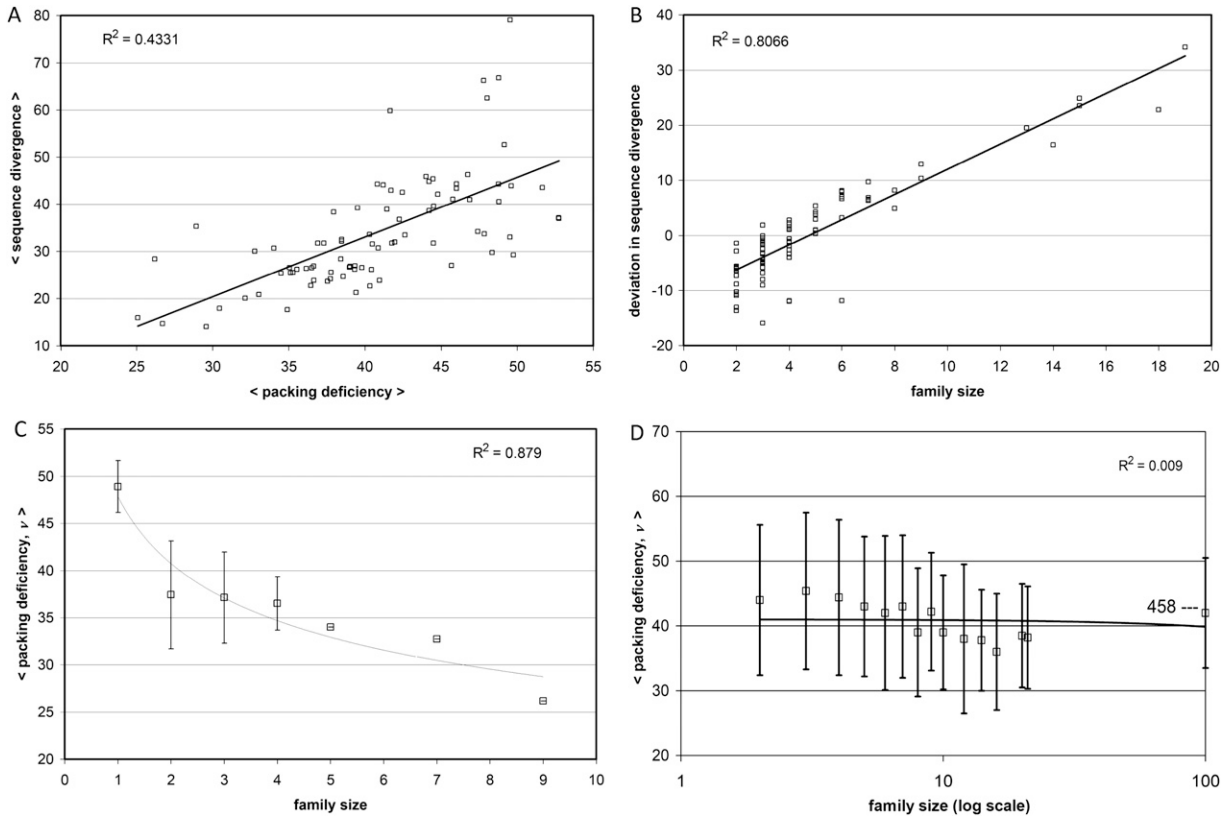

Figure 3. Selection pressure associated with evolution-related dosage imbalance for human families with low and high capacitance. (*A*) Correlation between paralog sequence divergence (δ) and packing deficiency (ν) for human families with low miRNA capacitance. Paralogs in families with high packing deficiency are under strong selection pressure, reflected in significant sequence divergence. (*B*) Correlation between family size (number of paralogs) and deviation in sequence divergence from the $\delta-\nu$ fitting trend shown *A*. (C) Anti-correlation between packing deficiency (error bars denote dispersion) and family size for low-capacitance human families. A tight anti-correlation is observed between packing deficiency and number of surviving paralogs in 690 human families (including singletons) with low capacitance ($\langle\tau\rangle > 0$ and significant mRNA coexpression $\langle\eta\rangle \geq 0.162$). Paralog survival is dependent on the packing quality of the protein with $P < 10^{-16}$ (Wilcoxon rank-sum test) (Supplemental Table 3). (*D*) Weak anti-correlation between packing deficiency (error bars denote dispersion) and family size for the 402 human families with high capacitance ($\langle\tau\rangle = 0$). High capacitance clearly mitigates selection pressure on paralogs, suppressing deleterious effects.

reflected in the significant correlation ($R^2 = 0.807$; Fig. 3B) between the number of paralogs in each family and the deviation from the low-capacitance $\delta-\nu$ trend shown in Figure 3A.

If indeed packing deficiency is a good measure of dosage sensitivity as the evidence from unicellular organisms suggests, then human families with low capacitance should be subject to a deleterious selection pressure comparable to that found for unicellular organisms. In other words, we should expect a family size well anti-correlated with packing deficiency, as previously found for unicellular organisms (Liang et al. 2008). Indeed, when the regulatory and evolutionary routes of paralog differentiation reducing evolution-related dosage imbalances are absent, the selection pressure determined by gene-duplication imbalances correlates well with the packing deficiency of the proteins, as in unicellular organisms. This is shown in Figure 3C, where the sizes of 690 human families (including singletons) with low miRNA capacitance are shown to be sharply anti-correlated with packing deficiency. At low capacitance, paralog survival is dependent on the packing quality of the protein with $P < 10^{-16}$ (Wilcoxon rank-sum test), a value comparable to that found at low organismal complexity (Liang et al. 2008).

On the other hand, a very weak anti-correlation between packing deficiency and family size is found (Fig. 3D) for human families with high capacitance ($\langle\tau\rangle = 0$), revealing the escape route of selection pressure introduced by localization dissimilarity.

Discussion

The results from this work underscore the importance of a molecular attribute, the packing deficiency of a protein, as a quantifier of dosage sensitivity. The latter property indicates the fitness impact of an imbalance in protein concentration relative to the stoichiometry of the protein complexes (Veitia 2002, 2004; Papp et al. 2003; Kondrashov and Koonin 2004; Liang et al. 2008). In this work we examine dosage imbalances that have an evolutionary origin. Thus, gene duplication events generate dosage imbalances that impose selection pressure on paralogs (Papp et al. 2003), and the magnitude of the effects of this pressure depend on the packing quality of the gene product (Liang et al. 2008). However, this dependence varies widely from unicellular to higher eukaryotes, with human being particularly insensitive to dosage imbalances (Liang et al. 2008). In human, there is a significant amount of genes with packing deficiency, which are nevertheless extensively duplicated. This suggests that humans are resilient to evolution-related dosage imbalances, a capacitance that may be rationalized in terms of escape routes available to human, but not to unicellular organisms,

where dosage imbalances have clear deleterious effects (Papp et al. 2003; Liang et al. 2008). This work focuses on the molecular/mechanistic basis for this major difference.

In the absence of expression dissimilarity, the initially identical paralogs of deficiently packed proteins are subject to high selection pressure because they compete for binding partners needed to maintain structural integrity. Conversely, tight protein packing reduces dosage sensitivity, thereby curbing selection pressure. Cross-examination of genetic and structural data reveals that humans have a built-in resilience or capacitance to dosage imbalances. The determinant of this human capacitance is traced in this contribution to the paralog-discriminatory power of miRNA regulatory patterns. In this way, dissimilar paralog localization governed by post-transcriptional regulation of protein levels mitigates the competition of paralogs for common binding partners that become obligatory for proteins of low packing quality. In other words, dissimilarity in paralog localization operative through miRNA control offers an escape route to dosage imbalances created by gene duplication, and this escape route becomes more necessary as protein packing deficiency makes these dosage imbalances less tolerable.

Alternative mechanisms different from those explored in this work may be invoked as responsible to curb dosage imbalance effects in human: Alternative splicing with splicing variants offering the escape routes to dosage imbalances (Romero et al. 2006), regulatory adjustors of gene expression levels (chaperones, proteases, noncoding RNAs, etc.) (Rockman and Wray 2002), allosteric oligomerization of duplicate proteins (Kuriyan and Eisenberg 2007), and selection inefficiency due to the smaller population of humans enabling a duplicate of a deficiently packed protein to be fixed in the population (Lynch and Conery 2003). While no doubt all these possibilities deserve independent study, the results of Figures 2 and 3 reveal that the miRNA-based capacitance is a dominant and quantifiable factor responsible for the human resilience to dosage imbalance.

If selection is indeed inefficient in human, one may wonder how miRNA-based capacitance could be achieved through random genetic drift. The removal of a miRNA binding site is readily achievable through a single deleterious mutation in one paralog and is unlikely to occur at the same binding site in another paralog. For instance, if M nonoverlapping miRNA-binding sites are present in the 3′UTR of two paralogs ($1 << M < 156$), the probability that a pair of mutations (one in each paralog) will discriminate paralogs is $(M-1)/M$, while the probability that they both occur at the same site in the two paralogs and hence do not contribute to increase the capacitance is $1/M$. Thus, dissimilarity and even orthogonality in post-transcriptional repressive patterns is readily achievable through widespread random mutation and hence may be the result of the nonadaptive forces prevalent in human.

In the context of human disease particular attention should be paid to overexpression of genes with high packing deficiency and low capacitance, since the somatic consequences arising from their dosage imbalance are likely to be important and unmitigated. Thus, pathogenic self-templating aggregation of proteins is promoted by overexpression (a dosage imbalance), while the aggregation propensity relates to extreme packing deficiencies in soluble proteins (Fernández et al. 2003). Hence, by removing the need for binding partners, protein self-aggregation is likely to curb dosage imbalance effects, which would be otherwise extremely severe. In other words, aberrant aggregation may prevail as an alternative to miRNA-based capacitance because it offers an escape route to dosage imbalances for human proteins with very high dosage sensitivity (high packing deficiency). Thus, this study is likely to impact our understanding of aggregation-induced clinical phenotypes.

Methods

Gene family databases

Combined structural, genetic, expression, and regulatory information were collected for *H. sapiens, D. melanogaster, C. elegans*, and *A. thaliana*. However, at this time, the scarcity of combined data for the latter three species (17, 4, and 1 fully described gene families, respectively) precludes statistically significant conclusions except for the human species.

We obtained gene information from the following sources: *A. thaliana*, The *Arabidopsis* Information Resource (TAIR) (Swarbreck et al. 2008); *C. elegans*, WormBase (http://www.wormbase.org/) (WB170); *D. melanogaster*, Berkeley *Drosophila* Genome Project (http://www.fruitfly.org/) (BDGP 4.3); and *H. sapiens* Ensembl Genome Database (NCBI36). Using the Ensembl gene family annotation (Birney et al. 2006), 20,173 *C. elegans* genes were grouped into 11,503 families (a singleton gene is counted as one family in our analysis), 14,116 *D. melanogaster* genes were grouped into 9477 families, and 22,357 human genes were grouped into 12,394 families.

Gene expression data for different species were obtained from different sources: Novartis Gene Expression Atlas (Su et al. 2004) for human, FlyAtlas for *D. melanogaster* (Chintapalli et al. 2007), PUMAdb for *C. elegans* (Capra et al. 2008), and ArrayExpress (Schmid et al. 2005) for *A. thaliana*. For human, the gene expression data set contains expression levels across a panel of 79 human tissues. We discarded six cancer tissues, hence samples of cancer tissues were not included: Colorectal Adenocarcinoma, leukemia lymphoblastic (molt4), lymphoma burkitts Raji, leukemia promyelocytic, lymphoma burkitts Daudi, leukemia chronic myelogenous (k562). The PUMAdb data set contains gene expression levels for *C. elegans* at six different developmental time points (egg, L1, L2, L3, L4, and young adult) in two different strains (N2 and CB4856).

Synonymous nucleotide divergence, K_s, across paralog pairs was determined using the PAML package (Yang and Nielsen 2000). MicroRNA-target predictions for human, *D. melanogaster*, and *C. elegans* were carried out with TargetScan (Bartel 2009; Friedman et al. 2009). MicroRNA target results for *A. thaliana* were downloaded from *Arabidopsis* Small RNA Project (ASRP) (Gustafson et al. 2005).

Thus, we identified exhaustive sets of nonsingleton families for which genetic, evolutionary, expression, structural (Liang et al. 2008), and post-transcriptional (Lewis et al. 2005; Liang and Li 2007; Bartel 2009; Friedman et al. 2009) information is available for at least two paralogs. Under the condition of positive paralog coexpression at the mRNA level, these exhaustive sets are made up of 457, 17, 4, and 1 families for human, *D. melanogaster*, *C. elegans*, and *A. thaliana*, respectively (Supplemental material).

Gene expression correlation

Gene expression data at the mRNA level were collected for human, *C. elegans*, *D. melanogaster*, and *A. thaliana* as specified above. Similarities between gene expression profiles were determined by Pearson correlation coefficients of expression vectors. In general, for two expression vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$, the Pearson coefficient is given by

$$\eta(\boldsymbol{X},\boldsymbol{Y}) = \mathrm{Corr}(X,Y) = \frac{<(X - <X>)(Y - <Y>)>}{\sqrt{<X^2> - <X>^2}\sqrt{<Y^2> - <Y>^2}},$$

where X, Y are generic coordinates in the vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively, and $\langle\ \rangle$ indicates mean over cell types, developmental phases, or metabolic adaptation phases as the case may be.

miRNA target coincidence

We identified target sites for the 156 conserved miRNA families (broadly conserved, intermediately conserved, and mammalian specific) in 17,444 human genes (Bartel 2009; Friedman et al. 2009). The miRNA target profile for a human gene was defined as a vector with 156 entries. The value of component n was set to be nonzero, if the gene 3′ UTR was predicted to be a target for conserved miRNA family n, and 0, otherwise. The normalized miRNA vector $\Psi_i/||\Psi_i||$ represents the pattern of miRNA-related repression efficacy on gene i, with $||\Psi_i||$ = vector norm. The nth entry in Ψ_i is $\Psi_i(n) = 1 - 2^{s(i,n)}$, where $s(i,n) \leq 0$ is the context score of the conserved miRNA-binding site n for gene i. The context score is known to correlate tightly with the post-transcriptional down-regulation efficacy, $2^{s(i,n)}$, of the predicted binding site for the nth miRNA family within the 3′ UTR of gene i (Grimson et al. 2007). Thus, $2^{s(i,n)} \approx g(i)/g(i,n)$, where $g(i)$ is the translation level for gene i and $g(i,n)$ is the i-translation level with knockout of the nth miRNA family. If the n-site is not predicted in the 3′ UTR of gene i, $g(i) = g(i,n)$ and therefore $s(i,n) = 0$. The miRNA-target coincidence τ between two genes is defined by the dot product of their respective normalized target vectors.

Sequence divergences within human gene families

The attribution of paralog sequence divergence for each human family from Ensembl 2006 was carried out in three steps: (1) prioritization of encoded domains; (2) computation of sequence distances for the domain with highest priority aligned across the gene family; and (3) average of such distances over all paralog pairs in the family. Because of the need to contrast sequence-based with structure-based data, domain prioritization is performed according to quality of sequence alignment (number of in-frame insertions or deletions) followed by PDB-representativity (number of aligned domains in the family that have PDB representation).

Calculation of packing deficiency and identification of SEBHs for soluble proteins

The protein packing deficiency is measured by the ratio of the SEBHs to the total number of BHBs in the structure. To determine the extent of solvent exposure of a BHB in a soluble protein structure, we determine the extent of bond protection from atomic coordinates. This parameter, denoted ρ, is given by the number of side-chain nonpolar groups contained within a desolvation domain (hydrogen-bond microenvironment) defined as two intersecting balls of fixed radius (~thickness of three water layers) centered at the alpha-carbons of the residues paired by the hydrogen bond (Pietrosemoli et al. 2007). In structures of PDB-reported soluble proteins, BHBs are protected, on average, by $\rho = 26.6 \pm 7.5$ side-chain nonpolar groups for a desolvation ball radius 6 Å. Thus, SEBHs lie in the tail of the distribution, i.e., their microenvironment contains 19 or fewer nonpolar groups, so their ρ-value is below the mean ($\rho = 26.6$) minus one standard deviation ($\sigma = 7.5$) (Fernández 2004; Pietrosemoli et al. 2007). An across-PDB analysis reveals that the ρ-distribution of hydrogen bonds is slightly bimodal, peaking at $\rho = 17$ and $\rho = 28$, with the bulk of the distribution weight, 86.2% of the 6,211,302 PDB-reported hydrogen bonds (PDB latest update as of June 2, 2009), in the range $22 \leq \rho \leq 30$ and 12.7% in the under-protection range $7 \leq \rho \leq 19$. No PDB-reported hydrogen bond has less than seven nonpolar protectors for desolvation radius 6 Å.

Homology threading

In cases where the protein structures were unavailable from the PDB, we generated atomic coordinates through homology threading adopting the PDB-reported homolog as template and using the program Modeller (Sali and Blundell 1993; Martí-Renom et al. 2000; Eswar et al. 2003). Modeller is a computer program that models three-dimensional (3D) structures of proteins subject to spatial constraints (Sali and Blundell 1993), and was adopted for homology and comparative protein structure modeling. We thus generate the alignment of the target sequence to be modeled with the PDB-reported homologs and the program computes a model with all nonhydrogen atoms. The input for the computation consists of the set of constraints applied to the spatial structure of the amino acid sequence to be modeled and the output is the 3D structure that best satisfies these constraints. The 3D model is obtained by optimization of a molecular probability density function with a variable target function procedure in Cartesian space that employs methods of conjugate gradients and molecular dynamics with simulated annealing.

In cases where protein structures were not reported in PDB, structural coordinates were generated through homology threading adopting the PDB-reported homolog as template. Reliable models require sequence identity (s.i.) greater than 40% (Aloy et al. 2003; Hillisch et al. 2004). Thus, data on homology-inferred packing deficiency for 30% < s.i. < 40%, with 30% = lower threshold to infer homology, are considered to be unreliable and are not reported in Figure 2. These data yield $\nu-\tau$ anti-correlations with far larger dispersions ($R^2 = 0.113$ and 0.178, respectively, for low and high expression correlations).

The resulting homology model was validated by comparing its inferred SEBH pattern with the SEBH pattern predicted from a sequence-based computation of disorder score (Obradovic et al. 2005) for a sliding window along the gene (Pietrosemoli et al. 2007). The inability of an isolated protein fold to protect specific intramolecular hydrogen bonds from water attack may lead to a structure-competing backbone hydration with concurrent local or global dismantling of the structure (Pietrosemoli et al. 2007). This view of under-wrapping implies a strong correlation between the degree of solvent exposure of intramolecular hydrogen bonds and the local propensity for structural disorder. Hence, the latter parameter was used to validate the former.

The disorder propensity was determined by a sequence-based score f_d ($0 \leq f_d \leq 1$; $f_d = 1$, certainty of disorder; $f_d = 0$, certainty of order) assigned to each residue. This parameter was generated by the highly accurate predictor of native disorder PONDR-VSL2 (Obradovic et al. 2005). The extent of intrinsic disorder of a domain is defined as the percentage of residues predicted to be disordered relative to a predetermined f_d threshold ($f_d = 0.5$). PONDR-VSL2 takes into account residue attributes, such as hydrophilicity, aromaticity, and their distribution within the window interrogated. The disorder score was assigned to each residue within a sliding window, representing the predicted propensity of the residue to be in a disordered region. Only 6% of 1100 nonhomologous PDB proteins gave false-positive predictions of disorder in sequence windows of 40 amino acids. A strong correlation (over 2806 nonredundant nonhomologous PDB domains) between disorder score of a residue and the ρ-parameter for the hydrogen bond engaging the residue (if any) implies that SEBHs correspond to an order–disorder intermediate region with $0.35 \leq f_d \leq 1$ (Pietrosemoli et al. 2007). Hence, the SEBHs inferred for paralogs of PDB-reported domains using threading homology were

confirmed through sequence-based disorder predictions. Only confirmed SEBHs were included in the ν computation.

Acknowledgment

The research of A.F. was supported by NIH/NIGMS grant R01 GM072614.

References

Aloy P, Ceulemans H, Stark A, Russell RB. 2003. The relationship between sequence and interaction divergence in proteins. *J Mol Biol* **332:** 989–998.

Bartel D. 2009. MicroRNAs: Target recognition and regulatory functions. *Cell* **136:** 215–233.

Birney E, Andrews D, Caccamo M, Chen Y, Clarke L, Coates G, Cox T, Cunningham F, Curwen V, Cutts T, et al. 2006. Ensembl 2006. *Nucleic Acids Res* **34:** D556–D561.

Capra EJ, Skrovanek SM, Kruglyak L. 2008. Comparative developmental expression profiling of two *C. elegans* isolates. *PLoS One* **3:** e4055. doi: 10.1371/journal.pone.0004055.

Chen F, Li W-H. 2001. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am J Hum Genet* **68:** 444–456.

Chintapalli VR, Wang J, Dow JA. 2007. Using FlyAtlas to identify better *Drosophila* models of human disease. *Nat Genet* **39:** 715–720.

Eswar N, John B, Mirkovic N, Fiser A, Ilyin V, Pieper U, Stuart AC, Marti-Renom MA, Madhusudhan MS, Yerkovich B, et al. 2003. Tools for comparative protein structure modeling and analysis. *Nucleic Acids Res* **31:** 3375–3380.

Fernández A. 2004. Keeping dry and crossing membranes. *Nat Biotechnol* **22:** 1081–1084.

Fernández A, Scheraga H. 2003. Insufficiently dehydrated hydrogen bonds as determinants for protein interactions. *Proc Natl Acad Sci* **100:** 113–118.

Fernández A, Scott R. 2003. Adherence of packing defects in soluble proteins. *Phys Rev Lett* **91:** 018102. doi: 10.1103/PhysRevLett.91.018102.

Fernández A, Kardos J, Scott R, Goto Y, Berry RS. 2003. Structural defects and the diagnosis of amyloidogenic propensity. *Proc Natl Acad Sci* **100:** 6446–6451.

Friedman RC, Farth KK, Burge CB, Bartel DP. 2009. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res* **19:** 92–105.

Gao L, Innan H. 2004. Very low gene duplication rate in the yeast genome. *Science* **306:** 1367–1370.

Grimson A, Farth KK, Johnston WK, Garret-Engele P, Lim LP, Bartel DP. 2007. MicroRNA targeting specificity in mammals: Determinants beyond seed pairing. *Mol Cell* **27:** 91–105.

Gu Z, Nicolae D, Lu HH, Li W-H. 2002. Rapid divergence in expression between duplicate genes inferred from microarray data. *Trends Genet* **18:** 609–613.

Gustafson AM, Allen E, Givan S, Smith D, Carrington JC, Kasschau KD. 2005. ASRP: The Arabidopsis Small RNA Project Database. *Nucleic Acids Res* **33:** D637–D640.

Hillisch A, Pineda LF, Hilgenfeld R. 2004. Utility of homology models in the drug discovery process. *Drug Discov Today* **9:** 659–669.

Kondrashov FA, Koonin EV. 2004. A common framework for understanding the origin of genetic dominance and evolutionary fates of gene duplications. *Trends Genet* **20:** 287–290.

Kuriyan J, Eisenberg D. 2007. The origin of protein interactions and allostery in colocalization. *Nature* **450:** 983–990.

Lewis B, Burge C, Bartel D. 2005. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120:** 15–20.

Liang H, Li WH. 2007. MicroRNA regulation of human protein–protein interaction network. *RNA* **13:** 1402–1408.

Liang H, Rogale-Plazonic K, Chen J, Li WH, Fernández A. 2008. Protein under-wrapping causes dosage sensitivity and decreases gene duplicability. *PLoS Genet* **4:** e11. doi: 10.1371/journal.pgen.0040011.

Lynch M, Conery JS. 2003. The origins of genome complexity. *Science* **302:** 1401–1404.

Martí-Renom M, Stuart A, Fiser A, Sánchez R, Melo F, Sali A. 2000. Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct* **29:** 291–325.

Obradovic Z, Peng K, Vucetic S, Radivojac P, Dunker K. 2005. Exploiting heterogeneous sequence properties improves prediction of protein disorder. *Proteins* **61:** 176–182.

Papp B, Pal C, Hurst LD. 2003. Dosage sensitivity and the evolution of gene families in yeast. *Nature* **424:** 194–197.

Pietrosemoli N, Crespo A, Fernández A. 2007. Dehydration propensity of order-disorder intermediate regions in soluble proteins. *J Proteome Res* **6:** 3519–3526.

Ramage R, Green J, Muir T, Ogunjobi O, Love S, Shaw K. 1994. Synthetic, structural and biological studies of the ubiquitin system: The total chemical synthesis of ubiquitin. *Biochem J* **299:** 151–158.

Rockman MV, Wray GA. 2002. Abundant raw material for *cis*-regulatory evolution in humans. *Mol Biol Evol* **19:** 1991–2004.

Romero P, Zaidi S, Fang Y, Uversky VN, Radivojac P, Oldfield CJ, Cortese MS, Sickmeier M, LeGall T, Obradovic Z, et al. 2006. Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms. *Proc Natl Acad Sci* **103:** 8390–8395.

Sali A, Blundell T. 1993. Comparative protein modeling by satisfaction of spatial restraints. *J Mol Biol* **234:** 779–815.

Schmid M, Davison TS, Henz SR, Pape UJ, Demar M, Vingron M, Schölkopf B, Weigel D, Lohmann JU. 2005. A gene expression map of *Arabidopsis thaliana* development. *Nat Genet* **37:** 501–506.

Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, et al. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci* **101:** 6062–6067.

Swarbreck D, Wilks C, Lamesch P, Berardini TZ, Garcia-Hernandez M, Foerster H, Li D, Meyer T, Muller R, Ploetz L, et al. 2008. The *Arabidopsis* Information Resource (TAIR): Gene structure and function annotation. *Nucleic Acids Res* **36:** D1009–D1014.

Veitia RA. 2002. Exploring the etiology of haploinsufficiency. *Bioessays* **24:** 175–184.

Veitia RA. 2004. Gene dosage balance: Deletions, duplications and dominance. *Trends Genet* **21:** 33–35.

Yang Z, Nielsen R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol Biol Evol* **17:** 32–43.

Received March 29, 2009; accepted in revised form September 30, 2009.