Cell

Intrinsic Protein Disorder and Interaction Promiscuity Are Widely Associated with Dosage Sensitivity

Tanya Vavouri,[1] Jennifer I. Semple,[1] Rosa Garcia-Verdugo,[1] and Ben Lehner[1,2,*]
[1]EMBL-CRG Systems Biology Unit
[2]ICREA
Centre for Genomic Regulation, UPF, Dr. Aiguader 88, Barcelona 08003, Spain
*Correspondence: ben.lehner@crg.es
DOI 10.1016/j.cell.2009.04.029

SUMMARY

Why are genes harmful when they are overexpressed? By testing possible causes of overexpression phenotypes in yeast, we identify intrinsic protein disorder as an important determinant of dosage sensitivity. Disordered regions are prone to make promiscuous molecular interactions when their concentration is increased, and we demonstrate that this is the likely cause of pathology when genes are overexpressed. We validate our findings in two animals, *Drosophila melanogaster* and *Caenorhabditis elegans*. In mice and humans the same properties are strongly associated with dosage-sensitive oncogenes, such that mass-action-driven molecular interactions may be a frequent cause of cancer. Dosage-sensitive genes are tightly regulated at the transcriptional, RNA, and protein levels, which may serve to prevent harmful increases in protein concentration under physiological conditions. Mass-action-driven interaction promiscuity is a single theoretical framework that can be used to understand, predict, and possibly treat the effects of increased gene expression in evolution and disease.

INTRODUCTION

Most of the genetic variation between any two individuals or species consists of regulatory or copy number variants that alter gene expression rather than coding sequence (Stranger et al., 2007). Despite the importance of altered gene expression to disease and evolution, it is not understood why only certain genes are pathological when their expression is increased (are dosage sensitive), and what the molecular mechanisms are that drive these phenotypic changes (Semple et al., 2008). Indeed there are no known molecular mechanisms that are predictive of dosage sensitivity across the genome of an organism (Gelperin et al., 2005; Semple et al., 2008; Sopko et al., 2006). As a result, it is currently very difficult to understand the consequences of increased gene expression in either disease or evolution.

In yeast, ~80% of genes can be constitutively overexpressed without any severe detrimental effect on growth (Gelperin et al., 2005; Sopko et al., 2006). In contrast a subset of genes are harmful when overexpressed. These dosage-sensitive genes are enriched for diverse and multiple functions (Gelperin et al., 2005; Sopko et al., 2006), and they do not significantly overlap the set of genes that are harmful when their expression is decreased (Deutschbauer et al., 2005; Semple et al., 2008). Unlike essential genes, dosage-sensitive genes are not enriched among the subunits of protein complexes (Sopko et al., 2006). Moreover, whereas the loss-of-function phenotype of one subunit of a protein complex is highly predictive of the loss-of-function phenotype of the other subunits (Fraser and Plotkin, 2007; Hart et al., 2007), this is not true for overexpression phenotypes (Semple et al., 2008). Indeed, in the majority of cases examined overexpression causes phenotypic effects that are different from underexpression (Niu et al., 2008; Sopko et al., 2006). It has also been shown that dosage-sensitive genes are only very weakly enriched for cell-cycle-regulated genes (Sopko et al., 2006), so forced expression of periodically expressed genes cannot be a major cause of phenotypic change. In short, it is not understood why cells function robustly following the overexpression of most genes but are very sensitive to increases in the levels of a subset of genes. It is also not clear what the most important molecular mechanisms are that cause gain-of-function phenotypes following gene overexpression.

To resolve this, we systematically tested possible causes of dosage sensitivity in yeast. We find that the intrinsic disorder content of a protein is an important determinant of dosage sensitivity. These disordered regions are prone to make promiscuous molecular interactions when their concentration is increased, and we present evidence that this is a frequent cause of dosage sensitivity. We confirm our findings in two animals, *Drosophila melanogaster* and *Caenorhabditis elegans*, and we show that the properties of dosage-sensitive genes detected in model organisms are also strongly associated with dosage-sensitive oncogenes in mice and humans. Finally, we show that dosage-sensitive genes are tightly regulated at the transcriptional, RNA, and protein levels, and we argue that this control acts to prevent potentially harmful increases in protein concentration under physiological conditions. The interaction promiscuity theory yields predictions for future experimental studies and

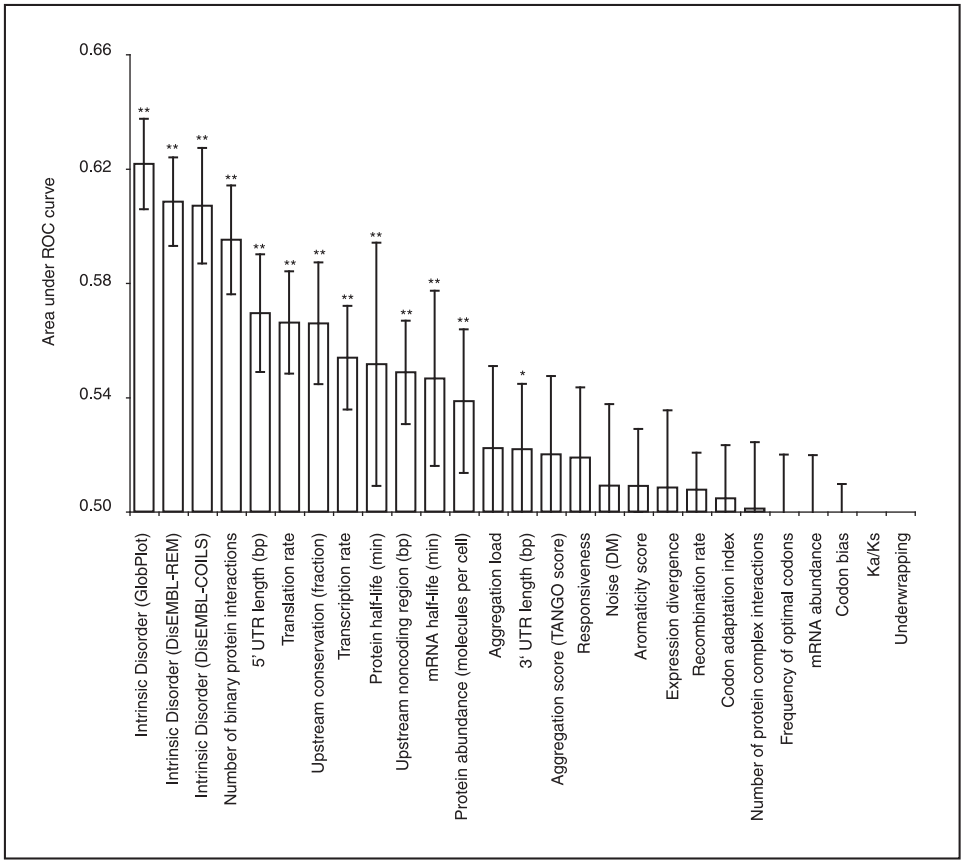

Figure 1. Features that Predict Dosage-Sensitive Genes in Yeast

Twenty-seven genomic and experimental features tested for their ability to predict dosage-sensitive genes in yeast by measuring the average area under a receiver operating characteristic (ROC) curve in a tenfold cross-validation experiment. The features and their correlations with dosage sensitivity are described in Table S1. Features that are significantly predictive are indicated, **p < 0.01, *p < 0.05 (one-tailed t test). Error bars show the 95% confidence interval for each predictor.

provides a single theoretical framework for understanding, predicting, and potentially treating dosage sensitivity in disease and evolution.

RESULTS

Testing Possible Determinants of Dosage Sensitivity in Yeast

Loss-of-function phenotypes resulting from decreased gene expression can be globally predicted in both unicellular and multicellular animals (Lee et al., 2008; Pena-Castillo et al., 2008). In contrast, the genes that are harmful when their expression levels are increased cannot be predicted, primarily because the mechanisms that drive overexpression phenotypes are unknown (Gelperin et al., 2005; Semple et al., 2008; Sopko et al., 2006).

In the budding yeast *Saccharomyces cerevisiae* ~18% of genes have a detrimental effect on growth when their expression is increased (Gelperin et al., 2005; Sopko et al., 2006). In most cases the overexpression phenotypes differ from loss-of-function phenotypes, suggesting that they normally represent gain-of-function effects (Gelperin et al., 2005; Semple et al., 2008; Sopko et al., 2006). Some of these phenotypic changes may result from the "misexpression" of a regulatory gene in a condition in which that gene is not normally expressed (Sopko et al., 2006). However, this cannot explain the vast majority of overexpression phenotypes. First, nearly all of the genes that are harmful when overexpressed are constitutively expressed during normal growth (93%; Holstege et al., 1998). Second, 87% of them have expression patterns that do not alter in level during the cell division cycle during normal growth (Gauthier et al., 2008). Third, 85% of them do not encode proteins that are considered to have regulatory functions (Segal et al., 2003). Therefore, misexpression is likely to explain only a few cases of overexpression phenotypes.

To identify alternative causes of dosage sensitivity, we tested a total of 27 genomic and experimental features for their relationship with overexpression phenotypes and used cross-validation to assess the use of each feature as a predictor of dosage sensitivity (see Experimental Procedures and Table S1). Among the features for which we find no relationship with dosage sensitivity are the abundance of an mRNA, the number of protein complex interactions, the aromaticity of a protein, the underwrapping of a protein (a measure of backbone exposure), and the aggregation propensity of a protein (Figure 1). Thus, for most genes,

sensitivity to a dosage increase must be caused by a mechanism distinct from misassembly of protein complexes, forced misexpression, or protein aggregation.

Intrinsic Protein Disorder and Linear Motif Content Are Predictive of Dosage Sensitivity

Many proteins contain both structured regions and intrinsically unstructured, or disordered, regions (Russell and Gibson, 2008). We find that the content of these intrinsically disordered regions is a good predictor of dosage sensitivity in yeast ($\rho = 0.94$, $p = 2.9 \times 10^{-5}$, Figure 2). This is seen using three alternative measures of intrinsic disorder (Figure 1). Moreover the strong relationship with disorder is seen when only considering genes with low (Figure 2B), medium (Figure 2C), or high (Figure 2D) levels of endogenous expression. It is also very strong when excluding all genes with expression levels that change during the cell cycle (Figure 2E) and when excluding all regulatory genes (Figure 2F). Disorder is therefore predictive of dosage sensitivity for many different types of genes in yeast.

To understand the mechanism that connects disordered regions to dosage sensitivity it is necessary to consider the functions of these regions. Unstructured protein regions are important because they contain short, linear functional sites within proteins (Russell and Gibson, 2008). Many recognition events within a cell—for example protein associations and posttranslational modifications—are mediated by the binding of globular protein domains to linear peptide sequence motifs contained within unstructured regions (Castagnoli et al., 2004; Collins et al., 2008; Russell and Gibson, 2008). It is possible therefore that dosage sensitivity is related to the ability of proteins to make molecular interactions via linear sequence motifs. Indeed, predicting instances of known linear motifs (Obenauer et al., 2003) across the yeast proteome shows that the number of known linear motifs within a protein is also highly correlated with dosage sensitivity ($\rho = 0.91$, $p < 2.3 \times 10^{-4}$). That is, both the intrinsic disorder content (Figures 2A–2F) and the linear motif content (Figure 2G) of a protein are predictive of dosage sensitivity in yeast.

The Interaction Promiscuity Hypothesis

The binding of two molecules depends not just on their affinity but also on their concentration. That is, as a simple consequence of mass action, any two molecules will associate if their concentration is high enough. Within a cell, many proteins have both physiological targets to which they bind with high affinity as well as additional targets to which they will bind if their concentration is increased. For high-affinity molecular interactions mediated via the large and complex interaction interfaces of two globular domains there are few potential "off target" interactions within a cell. In contrast, for interactions mediated by short linear motifs—for example the recognition of linear peptide motifs or the binding of transcription factors to DNA—there are many potential off target interactions with only marginally reduced affinities. This is because linear motifs are short and degenerate and so occur at high frequencies by chance in biological sequences (Diella et al., 2008; Russell and Gibson, 2008). Moreover, motif-binding protein domains are present in families of proteins with very similar binding site preferences and so will bind to each others' physiological targets if their concentration is increased (Diella et al., 2008; Russell and Gibson, 2008). For example, proteins containing SH3 domains bind to sequence motifs based on the consensus sequence PxxP, and the binding site preferences of individual proteins are both highly promiscuous and overlapping (Tong et al., 2002). Thus the profile and promiscuity of the interactions of proteins that contain linear motifs, or that are able to bind to linear motifs, are inherently sensitive to increases in protein concentration (Jones et al., 2006).

We propose that it is this potential for concentration-dependent interaction promiscuity, mediated via linear motif interactions, that is a major cause of dosage sensitivity in yeast.

Dosage Sensitivity Correlates with Binary Protein Interaction Degree

As a test of the interaction promiscuity theory, we asked whether there is any relationship between the number of protein interactions known for a protein and its likelihood of being dosage sensitive. As predicted by the hypothesis, there is (Figure 3A). Proteins that have more known binary interaction partners are much more likely to be dosage sensitive ($\rho = 0.92$, $p = 1.6 \times 10^{-4}$). This is not true when considering stable (high-affinity) interactions that can be identified using purification techniques (Figure 1, Table S1 available online) but only when considering binary interactions detected by sensitive interaction assays. This is exactly what is expected from the interaction promiscuity hypothesis, which predicts that it is the number of potential low-affinity interactions that is the important determinant of dosage sensitivity.

Linear Motif-Binding Proteins Are Dosage Sensitive

A further prediction of the promiscuity theory is that if linear motif interactions are an important cause of dosage sensitivity, then not just linear motif-containing proteins but also linear motif-binding proteins should be dosage sensitive. Increasing the concentration of a protein that can bind to short linear motifs should cause mass-action-driven promiscuous interactions just as increasing the concentration of a motif-containing protein does. To test this, we compiled a set of yeast proteins that contain domains that recognize linear sequence motifs and asked whether these proteins are also more likely to be dosage sensitive: they are (Figure 3B). Proteins that can bind to linear motifs are highly dosage sensitive ($p = 6.7 \times 10^{-15}$). Thus both linear motif-containing and linear motif-binding proteins are dosage sensitive, in agreement with the promiscuity hypothesis.

Dosage Sensitivity in *Drosophila*

In yeast we find that four measures of the potential of a protein to make promiscuous molecular interactions when overexpressed—the disorder content, the linear motif content, binary protein interaction degree, and the ability to bind linear motifs—are all predictive of dosage sensitivity. To test the generality of this result, we asked whether the same four measures are also predictive of dosage sensitivity in a second species, the fly *Drosophila melanogaster*. We used systematic data from screens in which ~1000 genes have been overexpressed in specific tissues and their phenotypic consequences assayed (Rorth, 1996; Toba et al., 1999). Just as in yeast, we find that

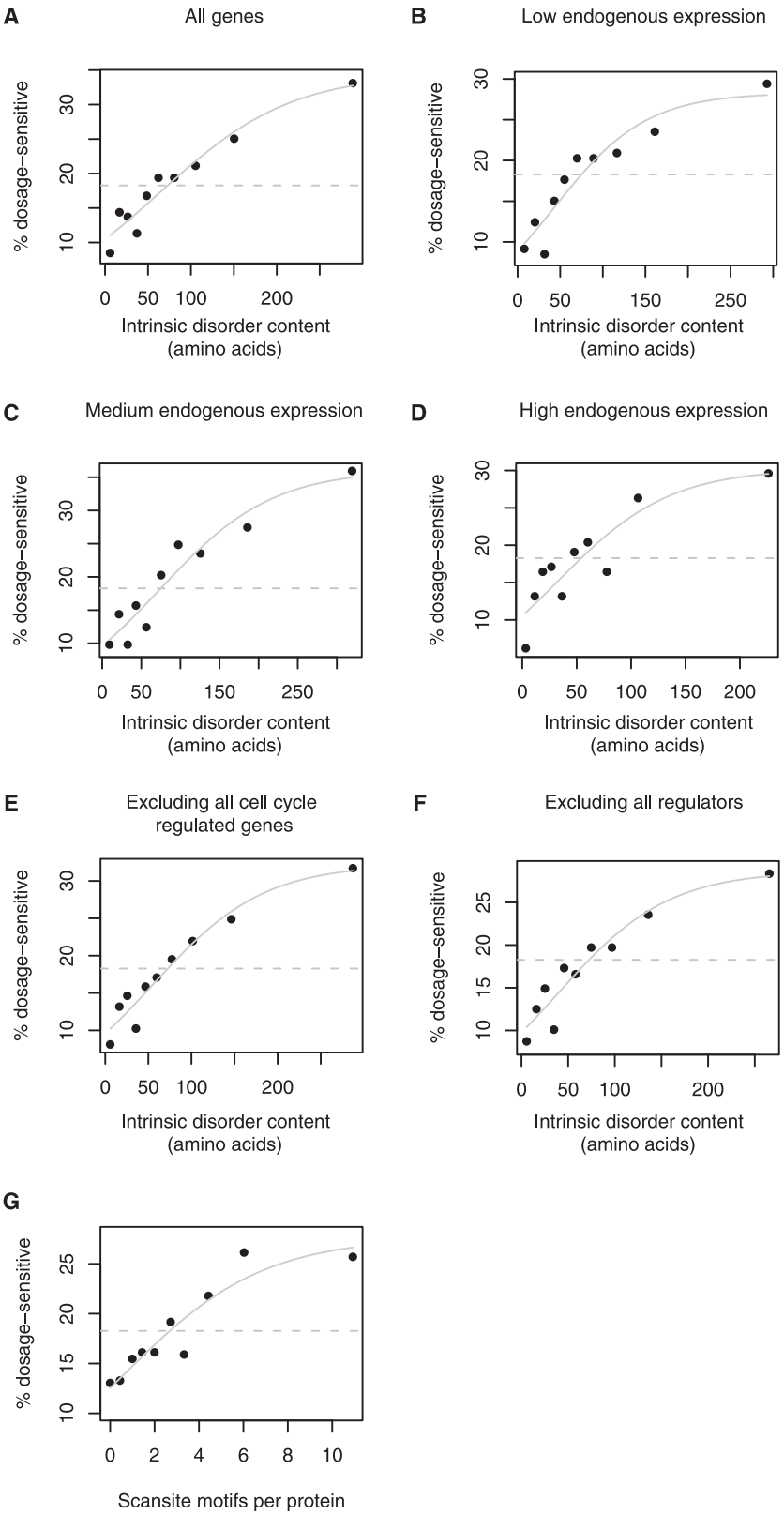

Figure 2. Intrinsic Protein Disorder and Linear Motif Content Are Associated with Dosage Sensitivity in Yeast

(A) There is a very good correlation between the total length of intrinsically disordered regions within a protein and dosage sensitivity in yeast (Spearman's rank correlation coefficient $\rho = 0.94$, $p = 2.9 \times 10^{-5}$). This correlation is still strong after normalizing by protein length (Figure S1A). The strong relationship between dosage sensitivity and intrinsic disorder is also seen when only considering genes with low (B), medium (C), or high (D) levels of endogenous expression (Beyer et al., 2004). It is also seen when excluding all genes with cell-cycle-regulated (Gauthier et al., 2008) expression patterns (E) and when excluding all regulatory genes (Segal et al., 2003) (F). There is also a strong correlation between the number of predicted linear motifs (Obenauer et al., 2003) a protein contains and its dosage sensitivity ($\rho = 0.91$, $p = 2.4 \times 10^{-4}$) (G). The effect is still strong after normalizing by protein length (Figure S1B). Also, the trend is strong when only considering either enzymatic motif-binding sites (Chi squared test for trend, $p = 5.1 \times 10^{-11}$) or nonenzymatic motif-binding sites (Chi squared test for trend, $p = 2.8 \times 10^{-3}$). The dashed lines indicate the frequency of dosage-sensitive genes for the whole yeast genome.

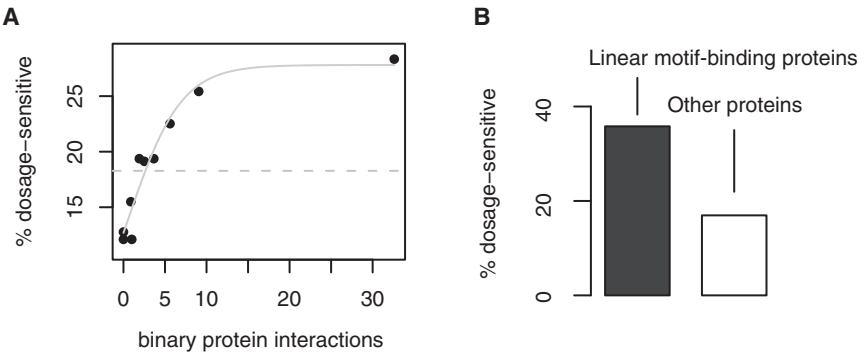

Figure 3. Dosage-Sensitive Genes in Yeast Have Many Binary Protein Interactions and Bind to Short Linear Motifs

(A) There is a good correlation between the number of binary protein interactions known for a protein and its dosage sensitivity ($\rho = 0.92$, $p = 1.6 \times 10^{-4}$). Moreover, just as linear motif content is predictive of dosage sensitivity (Figure 2), so is the ability to bind to linear motifs (B) ($p = 6.7 \times 10^{-15}$, Fisher's exact test). The same result is seen when only considering either enzymatic motif-binding domains ($p = 2.94 \times 10^{-5}$) or nonenzymatic motif-binding domains ($p = 5.35 \times 10^{-11}$).

all four predictions of the interaction promiscuity theory are validated in an animal. The intrinsic disorder content ($\rho = 0.83$, $p = 3.2 \times 10^{-5}$, Figure 4A), the linear motif content ($\rho = 0.78$, $p = 8.0 \times 10^{-3}$, Figure 4B), the number of binary interactions ($\rho = 0.73$, $p = 0.01$, Figure 4C), and the ability to bind to linear motifs (Fisher's exact test, $p = 2.2 \times 10^{-3}$, Figure 4D) are all predictive of dosage sensitivity.

We conclude that the potential for concentration-dependent interaction promiscuity is predictive of dosage sensitivity in both yeast and flies.

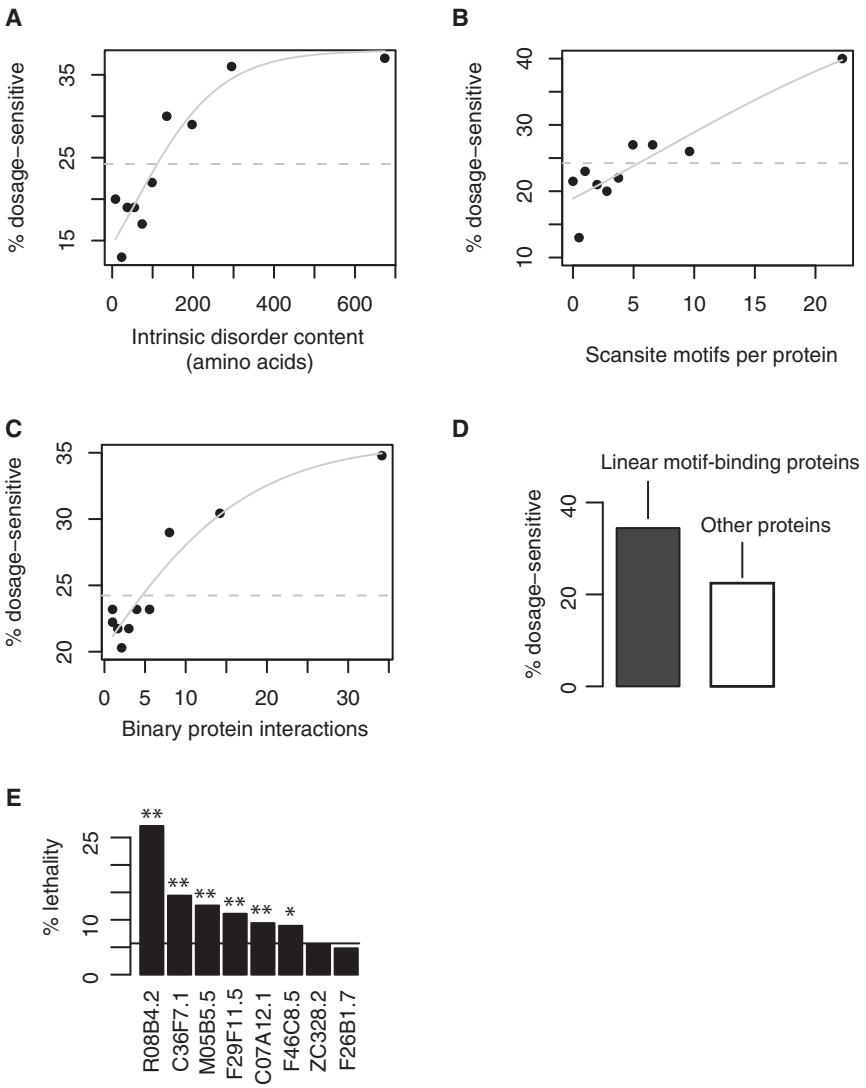

Figure 4. Dosage Sensitivity in *Drosophila melanogaster* and *Caenorhabditis elegans*

Just as in *S. cerevisiae*, in *D. melanogaster* the intrinsic disorder content (A) ($\rho = 0.83$, $p = 3.2 \times 10^{-5}$), the linear-motif content (B) ($\rho = 0.78$, $p = 8.0 \times 10^{-3}$), the number of binary protein interactions (C) ($\rho = 0.73$, $p = 0.01$), and the ability to bind to linear motifs (D) ($p = 2.2 \times 10^{-3}$) are predictive of dosage sensitivity.

(E) Integrating information on intrinsic disorder, binary interaction degree, and linear motif binding, we predicted dosage-sensitive genes in *C. elegans* (see Experimental Procedures). We overexpressed 8 of these genes and found that 6 (75%) induced embryonic lethality. The horizontal line indicates the background rate of lethality following heat shock. $^{**}p < 0.01$, $^{*}p < 0.05$ (Fisher's exact test).

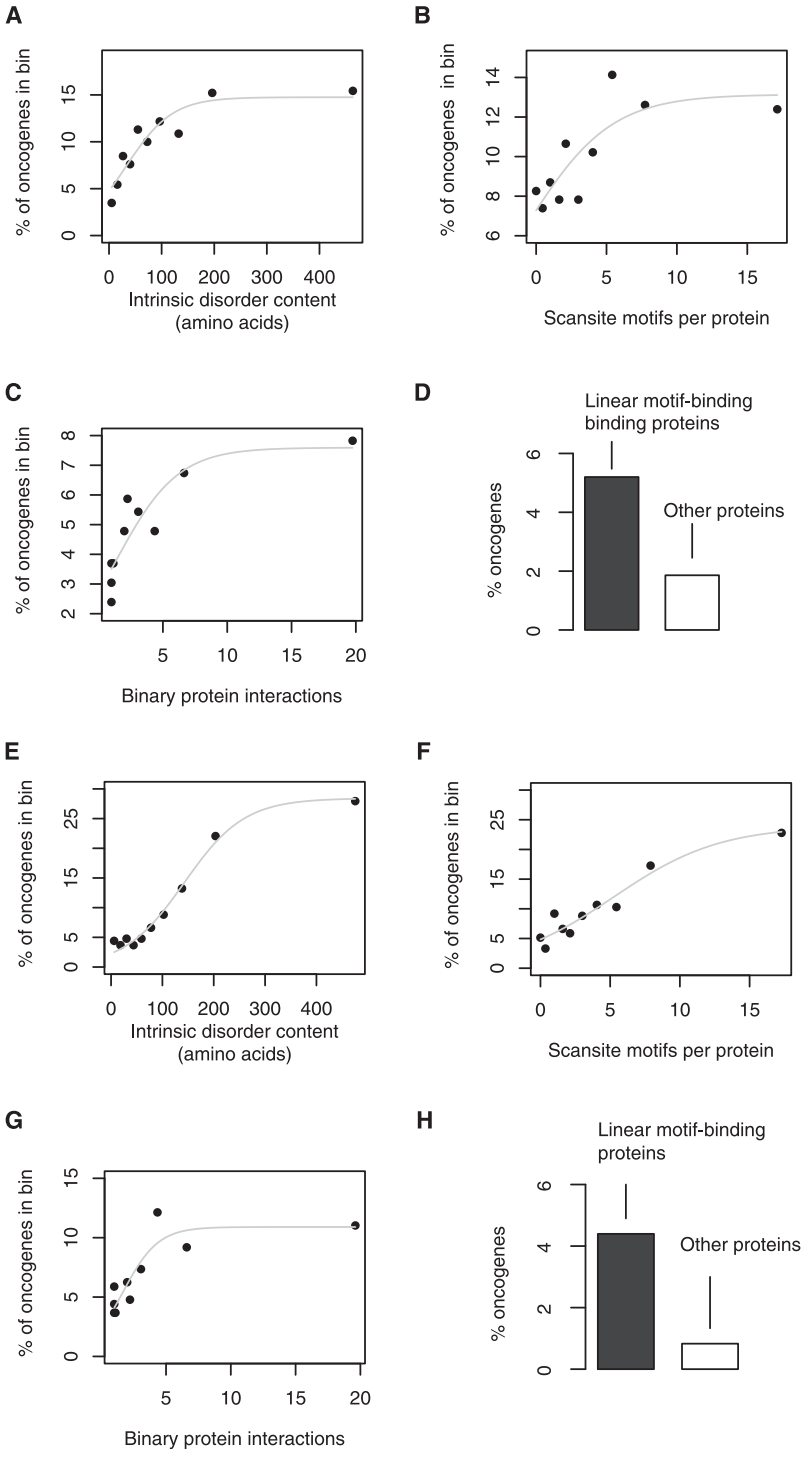

Figure 5. Dosage Sensitivity and Cancer in Mice and Humans

Intrinsic disorder (A) ($\rho = 0.93$, $p < 2.2 \times 10^{-16}$), linear motif content (B) ($\rho = 0.74$, $p = 0.013$), binary protein interaction degree (C) ($\rho = 0.91$, $p = 2.2 \times 10^{-4}$), and linear motif binding (D) ($p < 2.2 \times 10^{-16}$) are all also highly associated with dosage-sensitive genes that cause cancer when activated by retroviral insertion in mice (Akagi et al., 2004). In (A) to (C) the recall of oncogenes is shown for each equally sized bin of genes. As in mice, intrinsic disorder (E) ($\rho = 0.92$, $p = 1.6 \times 10^{-4}$), linear motif content (F) ($\rho = 0.89$, $p = 1.1 \times 10^{-3}$), binary protein interaction degree (G) ($\rho = 0.83$, $p = 2.9 \times 10^{-3}$), and linear motif binding (H) ($p = 2.2 \times 10^{-16}$) are highly associated with dosage-sensitive genes that cause cancer when activated in humans. The relationship between protein interaction degree and dosage sensitivity is also seen when only using data from high-throughput assays and so is not an artifact of ascertainment bias (data not shown). In (E) to (G) the recall of oncogenes is shown for each equally sized bin of genes.

Predicting Dosage Sensitivity in *C. elegans*

To further confirm our findings, we integrated information on protein disorder, linear motif binding, and protein interaction degree to predict dosage-sensitive genes in a third species, the nematode *Caenorhabditis elegans* (see Experimental Procedures). We tested 8 of the most highly ranked genes and verified 6 (75%) as causing lethality when overexpressed (Figure 4E). Thus using criteria and parameters derived from yeast, we are able to successfully predict dosage-sensitive genes in *C. elegans*.

Dosage Sensitivity in Mice

To test whether our findings also apply to mammals, we considered dosage-sensitive genes that are oncogenic when overexpressed. In mice these genes have been systematically identified in genetic screens using the integration of retroviruses to activate gene expression (Akagi et al., 2004). As in yeast, flies, and worms, these dosage-sensitive genes are strongly associated with protein disorder (Figure 5A) and have the presence of many linear motifs (Figure 5B), a high binary protein interaction degree (Figure 5C), and the ability to bind to linear motifs (Figure 5D). Thus the same properties associated with dosage-sensitive genes in yeast are able to predict dosage-sensitive cancer genes in mice.

Dosage Sensitivity in Human Cancer

In humans, the small set of genes that are known to be causally amplified in cancer are also strongly enriched for disorder ($p = 5.8 \times 10^{-4}$, Wilcoxon rank sum test), linear motifs ($p = 4.9 \times 10^{-4}$), binary protein interactions ($p = 0.035$), and the ability to bind to linear motifs ($p = 5.0 \times 10^{-9}$). This is also true of a larger set of oncogenes activated by either amplification or translocation (Figures 5E–5G). Thus, as in mice, dosage-sensitive oncogenes share the same properties as dosage-sensitive genes in model organisms.

We conclude that the properties of dosage-sensitive genes that we identify in yeast are also conserved for dosage-sensitive genes in mice and in human disease. Thus the principle of mass-action-driven interaction promiscuity can be used to

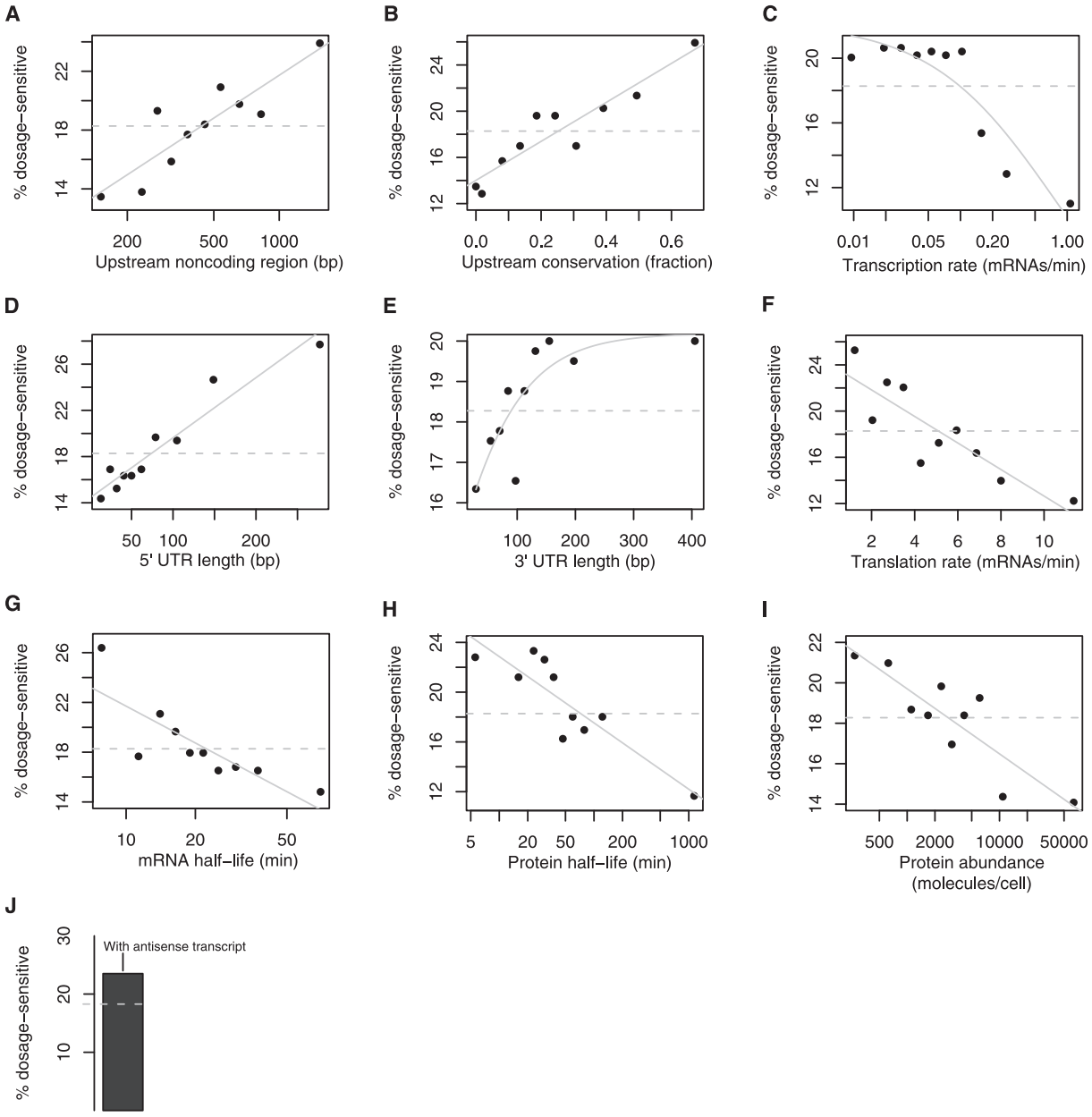

Figure 6. Dosage-Sensitive Genes Are Tightly Regulated and Rapidly Cleared to Prevent Harmful Increases in Protein Concentration

Dosage-sensitive genes in yeast have more extensive (A) and more conserved (B) upstream regulatory regions, slower transcription rates (C), larger 5′ (D) and 3′ (E) untranslated regions in their mRNAs, and lower translation rates (F). Dosage-sensitive genes also have faster rates of mRNA decay (G) and protein degradation (H), ensuring that they are rapidly cleared from the cell after use and resulting in lower overall protein abundances (I). They are also more likely to have overlapping antisense transcripts (J). All plots for quantitative variables are shown for ten evenly sized bins of genes, ranked according to the variable under consideration. Spearman's rank correlation coefficients (and p values): (A) 0.81 (7.5×10^{-3}), (B) 0.93 (8.2×10^{-5}), (C) −0.66 (3.0×10^{-2}), (D) 0.89 (5.5×10^{-4}), (E) 0.87 (1.0×10^{-3}), (F) −0.88 (2.0×10^{-5}) (G) −0.66 (3.8×10^{-2}), (H) −0.83 (3.0×10^{-3}), (I) −0.79 (6.5×10^{-3}). For (J), p = 0.014 by Fisher's exact test.

successfully predict dosage sensitivity across many different species.

Dosage-Sensitive Gene Products Are Tightly Regulated and Rapidly Degraded in Yeast

Genes that are harmful when overexpressed should be tightly regulated to prevent such harmful increases under physiological conditions. To test this prediction we used global datasets on gene regulation in yeast. In short, we find that this is the case, and that dosage-sensitive genes are tightly regulated at many levels.

At the DNA level, genes with overexpression phenotypes have both larger (Figure 6A) and more conserved (Figure 6B) upstream regions, reflecting tighter transcriptional control (Chin et al.,

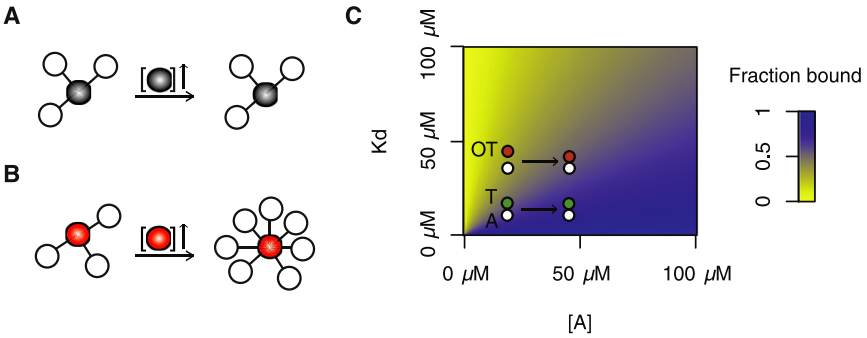

Figure 7. Mass-Action-Driven Interaction Promiscuity

As a result of mass action, increasing the concentration of a protein can dramatically alter its profile of cellular interactions. High-affinity domain-domain interactions have complex binding interfaces and very few potential "off-target" interactions. Their interaction profiles therefore change little in response to alterations in protein concentration (A). In contrast interactions with short, degenerate linear motifs are low affinity and have many potential off-target interactions within a cell due to the large families of motif-binding proteins and the high frequency of motifs and disordered regions in proteins (B) (Castagnoli et al., 2004; Neduva and Russell, 2005). The profiles of linear motif interactions in a network can therefore become much more promiscuous following increases in protein concentration. This is further illustrated in the phase-plane diagram shown in (C), which shows the sensitivity of linear motif interactions to changes in free protein concentration over realistic ranges of dissociation constants (K_D) (Neduva and Russell, 2005) and cellular concentrations (Wu and Pollard, 2005). As the free concentration of a protein (A) is increased it will interact both with physiological targets (A + T ↔ AT) and also with off-target molecules (A + OT ↔ AOT) to which it binds with lower affinity. Fraction bound is the proportion of target (T) or off-target (OT) proteins bound to protein A.

2005). They also have a lower rate of transcription (Figure 6C). At the mRNA level they have both larger 5′ (Figure 6D) and 3′ (Figure 6E) untranslated regions and faster rates of mRNA turnover (Figure 6G). They also have a lower translation rate (Figure 6F) and are subject to rapid degradation at the protein level (Figure 6H), and as a result, they have lower overall protein levels (Figure 6I). Dosage-sensitive genes are also more likely to have overlapping antisense transcripts (Figure 6J), suggesting a negative (or positive; Faghihi et al., 2008) role for antisense transcription in the regulation of dosage-sensitive genes. The tight regulation of dosage-sensitive genes is also seen when only considering disordered proteins (Figure S2, Gsponer et al., 2008), proteins with a high linear motif content (Figure S3), proteins with a high protein interaction degree (Figure S4), or proteins that bind to linear motifs (Figure S5).

We conclude that dosage-sensitive genes, and gene products that have the potential to make promiscuous interactions, are tightly regulated in yeast and, in particular, that they are slowly produced and rapidly cleared at both the mRNA and protein levels. These regulatory "safety mechanisms" may act to limit harmful accumulations of protein concentration during normal growth.

DISCUSSION

Mass-Action-Driven Interaction Promiscuity and the Mechanisms of Dosage Sensitivity

Within any cell there are tens of thousands of physical interactions that can occur between macromolecules. Although they are often represented as static structures, these networks of molecular interactions actually have topologies that alter in response to changes in protein concentration (Figure 7). As a consequence of mass action, an increase in the concentration of a protein can result in that protein making more promiscuous molecular interactions. What we term the interaction promiscuity hypothesis states that it is these promiscuous molecular interactions, primarily involving linear sequence motifs, that drive pathological changes in response to increased gene dosage (Figure 7).

In yeast there is good evidence that interaction promiscuity is a major cause of dosage sensitivity. First, the intrinsic disorder content of a protein is a good predictor of dosage sensitivity in this organism. Second, the more linear motifs a protein contains, and the more binary protein interactions that are known for a protein, the more likely it is to be dosage sensitive. Third, proteins that can bind to linear motifs are also highly dosage sensitive.

The same four measures of the potential for interaction promiscuity—intrinsic disorder, linear motif content, protein interaction degree, and motif binding—are all also predictive of dosage sensitivity in an animal, *Drosophila melanogaster*. Further, using the same four measures, it is possible to successfully predict de novo dosage-sensitive genes in a third species, *C. elegans*. We conclude that the potential for interaction promiscuity, mediated via linear motifs, is widely associated with dosage-sensitive genes across eukaryotes.

Interaction Promiscuity in Human Disease

The properties of dosage-sensitive genes in yeast, flies, and worms are also strongly associated with dosage-sensitive oncogenes in both mice and humans. It seems therefore that interaction promiscuity may provide a general method for predicting the changes in gene expression that are most likely to be pathological in humans. In any disease there are often many genes overexpressed or upregulated, and a central challenge for human genetics is to identify which of these are etiologically important. Interaction promiscuity provides one framework to do this.

Two previous observations also support our findings. First, protein kinases that are activated in cancer have more promiscuous substrate specificities than other kinases (Miller et al., 2008). Second, a quantitative study of the interactions of members of the ErbB family of cell-surface receptors showed that oncogenic family members become more promiscuous in their interactions when overexpressed (Jones et al., 2006). Again this is consistent with ectopic interactions being a widespread cause of gain-of-function phenotypes.

In animals, mass-action-driven interaction promiscuity also predicts an additional class of genes that should be particularly dosage sensitive—miRNAs. The interactions of miRNAs also depend on short, degenerate sequence motifs that are found

in very many cellular mRNAs. These interactions should also therefore be sensitive to overexpression. Consistent with this, there are many examples of miRNAs that are known to be pathological when overexpressed (Table S2). It is likely that many of these effects result from the concentration-dependent binding of miRNAs to nonphysiological target sequences.

Finally, our findings suggest that it may be possible to alleviate dosage-sensitive phenotypes in humans by using competitive inhibitors of linear motif interactions. The interactions of proteins with linear sequence motifs, which normally bind in deep surface clefts, are intrinsically more "druggable" than other protein interactions (Russell and Gibson, 2008) and so represent good candidates for therapeutic intervention.

Concluding Remarks

Most importantly, we demonstrate here a molecular mechanism that is widely predictive of dosage sensitivity, and one that is predictive across many different species. This makes it possible to consider predicting the effects of both decreased (Lee et al., 2008; Pena-Castillo et al., 2008) and increased gene expression in disease and evolution. Our findings highlight the importance of considering global interaction networks as having dynamic, not static, structures, and it is likely that further work in this area will illuminate many other areas of biology.

EXPERIMENTAL PROCEDURES

Testing Features for Their Ability to Predict Dosage Sensitivity in Yeast

Yeast genes with overexpression phenotypes were identified in two genome-wide screens (Gelperin et al., 2005; Sopko et al., 2006). There are a total of 839 genes with overexpression phenotypes out of 4591 genes tested. The complete set of sequence and experimental features tested for their ability to predict dosage sensitivity are described in Table S1. For each feature we first tested for a correlation between the feature and dosage sensitivity (Table S1). We then used a tenfold cross-validation experiment to test the ability of each feature to predict dosage-sensitive genes. We use the mean area under a receiver operating characteristic (ROC) curve for each of the cross-validation experiments as a measure of the performance of each feature as a predictor (Figure 1, Table S1).

Intrinsically Disordered Regions

Intrinsically disordered regions were identified using Globplot (Linding et al., 2003) using the default settings and using DisEMBL as described (Beltrao and Serrano, 2005). Genes were classified as low, medium, and highly expressed using three equally sized bins (Beyer et al., 2004). Cell-cycle-modulated genes were identified using the data from Cyclebase.org (Gauthier et al., 2008). Regulatory genes were taken from the classification of Segal et al. (2003).

Linear Motif Content

Predicted instances of known linear motifs were identified using Scansite 2.0 in the most stringent setting and using the following motif families relevant to yeast: pST_bind, SH3, acid_ST_kin, baso_ST_kin, DNA_dam_kin, Pro_ST_kin, kin_bind, PDZ (Obenauer et al., 2003). For the distinction between enzymatic and nonenzymatic motifs, we used the following grouping: pST_bind, SH3, and PDZ as nonenzymatic and the remainder as enzymatic.

Protein Interactions

Yeast protein interactions were downloaded from BioGRID 2.0.33 (Stark et al., 2006). We divided the interaction data into two sets—those detected by affinity purification methods (protein complex interactions), and those only detected by other methods (binary interactions). There are a total of 27,517 and 13,142 interactions in each dataset for the proteins tested for overexpression phenotypes.

Linear Motif-Binding Proteins

Linear motif-binding proteins are proteins containing protein domains that are known to bind to linear peptide motifs, as listed (Diella et al., 2008) (Interpro identifiers [Mulder et al., 2007] and the number of tested proteins are indicated, total n = 324 proteins in yeast): SH3 (IPR001452, 21), 14-3-3 (IPR000308, 2), PDZ (IPR001478, 2), EVH1 (IPR000697, 1), VHS (IPR002014, 4), FHA (IPR000253, 15), EH (IPR000261, 4), BRCT (IPR001357, 9), Bromo (IPR001487, 9), Chromo (IPR000953, 4), GYF (IPR003169, 3), ER retention receptor (IPR000133, 1), kinase (IPR011009, 122), Ser/Thr phosphatase (IPR006186, 12), dual-specificity phosphatase (IPR000340, 6), plus sequence-specific DNA-binding proteins (118; MacIsaac et al., 2006). For the distinction between enzymatic and nonenzymatic motif-binding proteins, we used the following grouping: kinase, Ser/Thr phosphatase, and dual-specificity phosphatases as enzymatic and the remainder as nonenzymatic.

Gene Regulation

Upstream intergenic distances were calculated from the SGD database (ftp://ftp.yeastgenome.org/yeast/). The following additional genomic datasets were used: transcription rate (Garcia-Martinez et al., 2004); mRNA half-life (Wang et al., 2002); protein half-life (Belle et al., 2006); translation rate (Beyer et al., 2004); protein abundance (Beyer et al., 2004); antisense transcripts (David et al., 2006); upstream conservation (the fraction of the upstream region overlapping with sequences conserved in 7 yeast species (Gustafson et al., 2006)); and 5′ and 3′UTR lengths (Nagalakshmi et al., 2008).

Drosophila Datasets

Genes tested for overexpression phenotypes in flies using the Gal4-driven overexpression system (Rorth, 1996; Toba et al., 1999) were downloaded from Flybase on December 2, 2008 (Crosby et al., 2007). A total of 1068 genes have been tested in overexpression screens, of which 279 have a reported morphological overexpression phenotype in Flybase. These data are available as Table S3. Protein interactions (a total of 4821 binary protein-protein interactions), motif-binding domains, linear motifs, and disorder predictions were defined as for yeast, with the addition of tyrosine kinase (Y_kin), SH2 motifs, and the SH2 domain (IPR000980).

Predicting and Testing Dosage-Sensitive Genes in *C. elegans*

We predicted dosage-sensitive genes in *C. elegans* using a generalized linear model fitted on the yeast data to rank *C. elegans* genes according to their likelihood of being dosage sensitive (0.004D + 0.009PI + 0.587L + 1.914, measuring the number of disordered residues ["D"] and the ability to bind to linear motifs ["L"] as defined for *Drosophila*, as well as binary protein interaction degree ("PI") using the dataset of Simonis et al., 2009). We focused on genes between 1 and 1.2 kb to facilitate cloning (a total of 2801 genes) and tested the first 8 of the top 20 ranked genes for which we obtained transgenic animals. Open reading frames were cloned into the heat-shock-inducible promoter vectors pMB1 and pMB7 (kindly provided by Mike Boxem) and microinjected into *C. elegans* with a *myo2::mCherry* cotransformation marker. Overexpression was induced using a 30 min heat shock at 35°C. Following heat shock worms were allowed to lay eggs for 24 hr at 20°C and then removed from the wells. Embryonic lethality was scored 24 hr later. A strain with heat-shock-inducible expression of green fluorescent protein (TJ375; Rea et al., 2005) was used as an internal control in all experiments.

Human and Mouse Datasets

Mouse oncogenes activated by retrovirus insertions are from the RTCGD database (http://rtcgd.abcc.ncifcrf.gov/), a total of 460 genes (Akagi et al., 2004), excluding all insertions that disrupt open reading frames. Human oncogenes activated by amplification (n = 9) or translocation (n = 263) are from the Sanger Cancer Gene Census (Futreal et al., 2004) (http://www.sanger.ac.uk/genetics/CGP/Census/). Motif-binding domains, linear motifs, and disorder were defined as for *Drosophila*. Human protein interaction data were taken from an integration of 21 different databases (Bossi and Lehner, 2009) and filtered to only include binary interactions detected by two-hybrid assays, removing

interactions also detected in complex purification methods (a total of 13,352 interactions). The same interaction dataset was used for mouse, using 1:1 orthology relationships identified by Ensembl.

SUPPLEMENTAL DATA

Supplemental Data include five figures and five tables and can be found with this article online at http://www.cell.com/supplemental/S0092-8674(09)00454-1.

ACKNOWLEDGMENTS

We thank Mark Isalan and Madan Babu for helpful discussions, Pedro Beltrao for providing DisEMBL scores, and Mike Boxem for *C. elegans* vectors. This work was funded by a European Research Council (ERC) Starting Grant, ICREA, the Spanish Ministry of Science and Innovation (MICINN), the CRG-EMBL Systems Biology Program, and a Marie Currie Intra-European Training Fellowship to T.V.

Received: October 30, 2008
Revised: February 3, 2009
Accepted: April 6, 2009
Published: July 9, 2009

REFERENCES

Akagi, K., Suzuki, T., Stephens, R.M., Jenkins, N.A., and Copeland, N.G. (2004). RTCGD: retroviral tagged cancer gene database. Nucleic Acids Res. *32*, D523–D527.

Belle, A., Tanay, A., Bitincka, L., Shamir, R., and O'Shea, E.K. (2006). Quantification of protein half-lives in the budding yeast proteome. Proc. Natl. Acad. Sci. USA *103*, 13004–13009.

Beltrao, P., and Serrano, L. (2005). Comparative genomics and disorder prediction identify biologically relevant SH3 protein interactions. PLoS Comput. Biol. *1*, e26. 10.1371/journal.pcbi.0010026.

Beyer, A., Hollunder, J., Nasheuer, H.P., and Wilhelm, T. (2004). Post-transcriptional expression regulation in the yeast Saccharomyces cerevisiae on a genomic scale. Mol. Cell. Proteomics *3*, 1083–1092.

Bossi, A., and Lehner, B. (2009). Tissue specificity and the human protein interaction network. Mol. Syst. Biol. *5*, 260.

Castagnoli, L., Costantini, A., Dall'Armi, C., Gonfloni, S., Montecchi-Palazzi, L., Panni, S., Paoluzi, S., Santonico, E., and Cesareni, G. (2004). Selectivity and promiscuity in the interaction network mediated by protein recognition modules. FEBS Lett. *567*, 74–79.

Chin, C.S., Chuang, J.H., and Li, H. (2005). Genome-wide regulatory complexity in yeast promoters: separation of functionally conserved and neutral sequence. Genome Res. *15*, 205–213.

Collins, M.O., Yu, L., Campuzano, I., Grant, S.G., and Choudhary, J.S. (2008). Phosphoproteomic analysis of the mouse brain cytosol reveals a predominance of protein phosphorylation in regions of intrinsic sequence disorder. Mol. Cell. Proteomics *7*, 1331–1348.

Crosby, M.A., Goodman, J.L., Strelets, V.B., Zhang, P., and Gelbart, W.M. (2007). FlyBase: genomes by the dozen. Nucleic Acids Res. *35*, D486–D491.

David, L., Huber, W., Granovskaia, M., Toedling, J., Palm, C.J., Bofkin, L., Jones, T., Davis, R.W., and Steinmetz, L.M. (2006). A high-resolution map of transcription in the yeast genome. Proc. Natl. Acad. Sci. USA *103*, 5320–5325.

Deutschbauer, A.M., Jaramillo, D.F., Proctor, M., Kumm, J., Hillenmeyer, M.E., Davis, R.W., Nislow, C., and Giaever, G. (2005). Mechanisms of haploinsufficiency revealed by genome-wide profiling in yeast. Genetics *169*, 1915–1925.

Diella, F., Haslam, N., Chica, C., Budd, A., Michael, S., Brown, N.P., Trave, G., and Gibson, T.J. (2008). Understanding eukaryotic linear motifs and their role in cell signaling and regulation. Front. Biosci. *13*, 6580–6603.

Faghihi, M.A., Modarresi, F., Khalil, A.M., Wood, D.E., Sahagan, B.G., Morgan, T.E., Finch, C.E., St Laurent, G., 3rd, Kenny, P.J., and Wahlestedt, C. (2008). Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of beta-secretase. Nat. Med. *14*, 723–730.

Fraser, H.B., and Plotkin, J.B. (2007). Using protein complexes to predict phenotypic effects of gene mutation. Genome Biol. *8*, R252.

Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., and Stratton, M.R. (2004). A census of human cancer genes. Nat. Rev. Cancer *4*, 177–183.

Garcia-Martinez, J., Aranda, A., and Perez-Ortin, J.E. (2004). Genomic run-on evaluates transcription rates for all yeast genes and identifies gene regulatory mechanisms. Mol. Cell *15*, 303–313.

Gauthier, N.P., Larsen, M.E., Wernersson, R., de Lichtenberg, U., Jensen, L.J., Brunak, S., and Jensen, T.S. (2008). Cyclebase.org–a comprehensive multi-organism online database of cell-cycle experiments. Nucleic Acids Res. *36*, D854–D859.

Gelperin, D.M., White, M.A., Wilkinson, M.L., Kon, Y., Kung, L.A., Wise, K.J., Lopez-Hoyo, N., Jiang, L., Piccirillo, S., Yu, H., et al. (2005). Biochemical and genetic analysis of the yeast proteome with a movable ORF collection. Genes Dev. *19*, 2816–2826.

Gsponer, J., Futschik, M.E., Teichmann, S.A., and Babu, M.M. (2008). Tight regulation of unstructured proteins: from transcript synthesis to protein degradation. Science *322*, 1365–1368.

Gustafson, A.M., Snitkin, E.S., Parker, S.C., DeLisi, C., and Kasif, S. (2006). Towards the identification of essential genes using targeted genome sequencing and comparative analysis. BMC Genomics *7*, 265.

Hart, G.T., Lee, I., and Marcotte, E.R. (2007). A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality. BMC Bioinformatics *8*, 236.

Holstege, F.C., Jennings, E.G., Wyrick, J.J., Lee, T.I., Hengartner, C.J., Green, M.R., Golub, T.R., Lander, E.S., and Young, R.A. (1998). Dissecting the regulatory circuitry of a eukaryotic genome. Cell *95*, 717–728.

Jones, R.B., Gordus, A., Krall, J.A., and MacBeath, G. (2006). A quantitative protein interaction network for the ErbB receptors using protein microarrays. Nature *439*, 168–174.

Lee, I., Lehner, B., Crombie, C., Wong, W., Fraser, A.G., and Marcotte, E.M. (2008). A single gene network accurately predicts phenotypic effects of gene perturbation in Caenorhabditis elegans. Nat. Genet. *40*, 181–188.

Linding, R., Russell, R.B., Neduva, V., and Gibson, T.J. (2003). GlobPlot: Exploring protein sequences for globularity and disorder. Nucleic Acids Res. *31*, 3701–3708.

MacIsaac, K.D., Wang, T., Gordon, D.B., Gifford, D.K., Stormo, G.D., and Fraenkel, E. (2006). An improved map of conserved regulatory sites for Saccharomyces cerevisiae. BMC Bioinformatics *7*, 113.

Miller, M.L., Jensen, L.J., Diella, F., Jorgensen, C., Tinti, M., Li, L., Hsiung, M., Parker, S.A., Bordeaux, J., Sicheritz-Ponten, T., et al. (2008). Linear motif atlas for phosphorylation-dependent signaling. Sci. Signal. *1*, ra2.

Mulder, N.J., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Buillard, V., Cerutti, L., Copley, R., et al. (2007). New developments in the InterPro database. Nucleic Acids Res. *35*, D224–D228.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. Science *320*, 1344–1349.

Neduva, V., and Russell, R.B. (2005). Linear motifs: evolutionary interaction switches. FEBS Lett. *579*, 3342–3345.

Niu, W., Li, Z., Zhan, W., Iyer, V.R., and Marcotte, E.M. (2008). Mechanisms of cell cycle control revealed by a systematic and quantitative overexpression screen in S. cerevisiae. PLoS Genet. *4*, e1000120. 10.1371/journal.pgen.1000120.

Obenauer, J.C., Cantley, L.C., and Yaffe, M.B. (2003). Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. Nucleic Acids Res. *31*, 3635–3641.

Pena-Castillo, L., Tasan, M., Myers, C.L., Lee, H., Joshi, T., Zhang, C., Guan, Y., Leone, M., Pagnani, A., Kim, W.K., et al. (2008). A critical assessment of

Mus musculus gene function prediction using integrated genomic evidence. Genome Biol. *9* (*Suppl 1*), S2.

Rea, S.L., Wu, D., Cypser, J.R., Vaupel, J.W., and Johnson, T.E. (2005). A stress-sensitive reporter predicts longevity in isogenic populations of Caenorhabditis elegans. Nat. Genet. *37*, 894–898.

Rorth, P. (1996). A modular misexpression screen in Drosophila detecting tissue-specific phenotypes. Proc. Natl. Acad. Sci. USA *93*, 12418–12422.

Russell, R.B., and Gibson, T.J. (2008). A careful disorderliness in the proteome: sites for interaction and targets for future therapies. FEBS Lett. *582*, 1271–1275.

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., and Friedman, N. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat. Genet. *34*, 166–176.

Semple, J.I., Vavouri, T., and Lehner, B. (2008). A simple principle concerning the robustness of protein complex activity to changes in gene expression. BMC Syst. Biol. *2*, 1.

Simonis, N., Rual, J.F., Carvunis, A.R., Tasan, M., Lemmens, I., Hirozane-Kishikawa, T., Hao, T., Sahalie, J.M., Venkatesan, K., Gebreab, F., et al. (2009). Empirically controlled mapping of the Caenorhabditis elegans protein-protein interactome network. Nat. Methods *6*, 47–54.

Sopko, R., Huang, D., Preston, N., Chua, G., Papp, B., Kafadar, K., Snyder, M., Oliver, S.G., Cyert, M., Hughes, T.R., et al. (2006). Mapping pathways and phenotypes by systematic gene overexpression. Mol. Cell *21*, 319–330.

Stark, C., Breitkreutz, B.J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006). BioGRID: a general repository for interaction datasets. Nucleic Acids Res. *34*, D535–D539.

Stranger, B.E., Forrest, M.S., Dunning, M., Ingle, C.E., Beazley, C., Thorne, N., Redon, R., Bird, C.P., de Grassi, A., Lee, C., et al. (2007). Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science *315*, 848–853.

Toba, G., Ohsako, T., Miyata, N., Ohtsuka, T., Seong, K.H., and Aigaki, T. (1999). The gene search system. A method for efficient detection and rapid molecular identification of genes in Drosophila melanogaster. Genetics *151*, 725–737.

Tong, A.H., Drees, B., Nardelli, G., Bader, G.D., Brannetti, B., Castagnoli, L., Evangelista, M., Ferracuti, S., Nelson, B., Paoluzi, S., et al. (2002). A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. Science *295*, 321–324.

Wang, Y., Liu, C.L., Storey, J.D., Tibshirani, R.J., Herschlag, D., and Brown, P.O. (2002). Precision and functional specificity in mRNA decay. Proc. Natl. Acad. Sci. USA *99*, 5860–5865.

Wu, J.Q., and Pollard, T.D. (2005). Counting cytokinesis proteins globally and locally in fission yeast. Science *310*, 310–314.