The Thesis Committee for Michael David Wittig
Certifies that this is the approved version of the following thesis:

Search for Selection Pressures Associated with Aggregation Propensity Following Whole Genome Duplication in *S.cerevisiae.*

Committee:

William Press, Supervisor

Edward Marcotte

Search for Selection Pressures Associated with Aggregation Propensity Following Whole Genome Duplication in *S.cerevisiae.*

By

Michael David Wittig, B.S.Bio.Sci.;B.S.Comp.Sci.

Thesis

Presented to the Faculty of the Graduate School

of the University of the Texas at Austin

in partial fulfillment

of the requirements

for the Degree of

Master of Arts

The University of Texas at Austin
December 2011

Search for Selection Pressures Associated with Aggregation Propensity Following Whole Genome Duplication in *S.cerevisiae.*

Michael David Wittig, MA

The University of Texas at Austin, 2011

Supervisor: William Press

It has been theorized that most proteins are under selection pressure to be soluble in crowded cellular spaces. To maintain solubility a proteins' aggregation propensity should be inversely proportional to their maximum likely concentration. This theory was examined by comparing the proteome of the model organism *S. cerevisiae*, which has previously undergone a Whole Genome Duplication (WGD) event to the proteome of the closely related yeast *K. waltii*, which has not undergone WGD. This comparison revealed the following: 1) Predicted aggregation propensities are higher in *S. cerevisiae* than *K. waltii*. 2) Aggregation propensity does not predict which genes reverted to a single copy after WGD. 3) In genes which were retained as duplicates in *S. cerevisiae* after WGD, aggregation propensities rose from the inferred common ancestral protein. 4) Genes retained as duplicates showed less of an increase relative to their homologues in *K. waltii* than genes which were not retained as duplicates. 5) The relationship between the log predicted aggregation propensity and log mRNA expression level or log protein abundance was not linear as previously predicted. These results suggest that while there is broad selection pressure for reduced aggregation pressure for genes which have been duplicated, the precise relationship between aggregation propensity and gene expression is more complicated than previously predicted. These results also allow speculation that the whole genome duplication in *S.cerevisiae* may have been made possible by a general relaxation of aggregation-related selection pressure.

Table of Contents

List of Figures

Introduction

Protein aggregation is a large and growing topic with great medical relevance due to various human neurodegenerative diseases which are linked with aggregation and amyloid formation [1]. One important open question in the field is whether aggregation effects are limited to a small number of proteins which then cause disease, or whether all proteins are aggregation-prone and thus possibly disease-associated [2]. If a particular gene is problematic, then focusing on that gene should lead to useful therapies. If all proteins are aggregation prone, then a more general solution will be needed. It is also unclear whether aggregation is really one process or many, with research suggesting that there aggregation in general should be distinguished from the fibril formation seen in neurodegenerative diseases [3]. There is even some evidence for a viral component to Alzheimer's [4], in which case aggregation is merely the proximate cause of symptoms and not the ultimate cause, leading toward an entirely different set of therapies aimed at the virus.

In the quest to better understand protein aggregation, multiple groups have developed algorithms for predicting protein aggregation. The Vendruscolo lab has produced an algorithm called Zyggregator which predicts protein aggregation propensity [5]. Zyggregator uses the hydrophobicity, the secondary structure propensity, and the charge of the amino acid (hereafter, AA) to predict the effects of mutations upon the aggregation propensity of a protein. To predict the absolute aggregation rate, the same factors are used plus the protein is rated on whether or not it has alternating hydrophobic and hydrophilic residues. Their model also accounted for extrinsic factors such as pH and ionic strength. They have argued that aggregation is potentially an issue for any

and all proteins, based on their finding of a strong inverse relationship between mRNA expression and protein abundance for a small set of human genes [2]. They conclude that all proteins are poised on the edge of aggregation-related disorders with no safety margin. If true, this would have tremendous ramifications, both for our understanding of aggregation and our approaches to treating aggregation related diseases. It is a bold conclusion based on a very small dataset, and therefore warrants additional testing. Unfortunately, the Zyggregator algorithm is only available via a web form that accepts only a single AA sequence, making it difficult to use for large data sets.

The Serrano lab has developed an algorithm called TANGO which predicts aggregation propensity by examining the tendency of the protein to assume beta-sheet conformations rather than competing secondary structures [67]. Specifically, it calculates the partition function of the phase space specified by the Boltzmann distribution, which states that the frequency of each structural state depends only on the energy of that state. The distribution is calculated for each amino acid in isolation, since the problem becomes intractable for entire proteins, and so the distribution is only an approximation. They have used this algorithm to suggest that aggregation is mostly caused by hydrophobic aggregation prone regions which are surrounded by hydrophilic gatekeeper amino acid residues. They have further concluded that while aggregation is a fairly universal protein trait, the formation of amyloid fibrils is not necessarily associated with aggregation and appears to be a fundamentally different process [3]. The authors have made the TANGO algorithm publicly available in the form of a downloadable executable, allowing other researchers to use it on large data sets. Therefore, this prediction algorithm will be used for all aggregation propensity prediction in this paper.

Additional detail on these three topics follows.

Protein Aggregation Background

The following information is primarily taken from the review paper *Cellular Strategies for Controlling Protein Aggregation*, by Tiedmers, et al. Proteins must bury their hydrophobic residues by folding properly in order to function [8] [9]. If they fail to do so, those hydrophobic surfaces can instead cause aggregation by recruiting other proteins and trapping them in a misfolded state, which is eventually toxic in some cases if left unchecked [10] [11] [12]. Cells use a number of systems to deal with misfolding [13] [14] [15] [16] [17]. Chaperone proteins, especially the Heat Shock Proteins (HSPs), help proteins fold properly [14] [18] [19]. Misfolded proteins that are not refolded are degraded by cytosolic ATP-dependent AAA+ proteases [13] or acidic hydrolases after they are moved to the lysosomal compartment [17] [20] [21] [22]. Protein aggregation seems to result form the exhaustion of these functions, either due to a single severe defect or from a combination of moderate conditions, with the defects falling into four broad categories: (i) Mutations that result in proteins very prone to misfolding or that disrupt the protein quality-control systems [23] [24] [25] [26] [27] [28], (ii) defects in protein biogenesis due to translation errors or defects in the assembly of protein complexes [29] [30], (iii) environmental stress conditions such as reversible heat-induced unfolding [31] and irreversible oxidative damage in the form of peptide backbone fragmentation or carbonylation [32] [33], and (iv) the slower effects of aging, due to both an accumulation of aggregates the cell is unable to deal with and to the progressive exhaustion of the quality-control systems [34] [35] [36] [37] [38]. The primary structural feature of aggregates appears to be intramolecular beta-sheet, with variation in the degree of organization of those sheets, with the most organization found in amyloid fibrils [39] [40] [41].

It appears that in some cases, the toxic elements are the soluble oligomers, rather than the aggregates themselves [24], and that cells respond to the potentially toxic misfolded proteins by collecting them in aggregates [42 43 44 45 46 47 48] and then sequestering the aggregates at particular sites to be dealt with [49 50 51]. In yeast [52], multiple patterns of localization occur. Heat stress causes aggregation that is not specific to particular parts of the cell [53], with most types of aggregates able to be reactivated by chaperones during a recovery period. Aggregates composed of unrecoverable proteins, such as oxidatively damaged proteins, ubiquitinated proteins, and others, may be localized to one of two sites [51]. The first is the juxtanuclear quality-control compartment (JUNC), which is the localization site for ubiquitinated proteins. The other is found near the vacuole, called the insoluble protein deposit (IPOD)[54]. In mammals, a specialized form of inclusion bodies is termed the aggresome [55 56]. These structures are not normally present but appear in various disease states [57 58], localizing near the microtubule-organizing center near the nuclear envelope [57]. It appears that smaller aggregates are dragged to this site along the microbutules from elsewhere in the cell [55].

Once a protein has accumulated in an aggregate, it may be dealt with by the cell using a number of mechanisms. A bi-chaperone system involving the Hsp70 and Hsp104/ClpB systems deals with heat-aggregated proteins by helping them refold into their proper state and by protecting the damaged proteins from the protease systems[53]. The bi-chaperone system appears to work by a threading activity which leads to a one-by-one extraction of misfolded proteins from the aggregate[59]. Small heat-shock proteins (sHSPs) respond to high temperatures by binding tightly to misfolded species[60 61]. This provides cells with a handle on aggregates, increasing their solubility, allowing transport,

creating a reservoir of misfolded proteins during heat shock, and allowing for more efficient disaggregation by chaperone systems. In bacteria, several additional AAA+ chaperones also possess a disaggregation activity[62]. In eukaryotes, those chaperones are found only in the mitochondria and chloroplasts[63], and while cytosolic disaggregation still occurs, the chaperones responsible are not known[64]. One candidate is valosin-containing protein (VCP), which is an ATP and ubiquitin dependent AAA+ chaperone[65]. Eukaryotes are thought to rely more heavily on protein degradation relative to protein refolding, possibly via the ubiquitination system[66]. They also make use of macroautophagy[21], where a specialized, cytosolic, double-membrane structure engulfs substrates to form autophagic vesicles that ultimately fuse with the lysosome for degradation of their contents.

Cells may also deal with aggregates through asymmetrical partitioning during cell division. By moving all of the aggregates to one cell, the other cell benefits from a reduced aggregate load. In *E.coli*, protein aggregates are retained along with the old cell pole, producing aggregate free daughter cells which reproduce faster[67]. In yeast, the mother cell retains aggregates and produces aggregate-free daughter cells by budding[68], with the aggregates transported out of the bud via actin[69]. In mammalian cells, asymmetrical partitioning has been observed in some cases, with the shorter-lived cell receiving the aggregates[70], with the mechanism suspected to involve the centrosome[71].

Edge Theory Papers:

Tartaglia and Vendruscolo have proposed that human proteins have evolved precisely enough aggregation resistance to avoid aggregation at their current expression

levels but with no margin of error such that any increase in expression level or aggregation rate will trigger aggregation[2]. They collected data from the literature on mRNA expression levels and measured in vivo aggregation propensities of human proteins. They were able to obtain this data for 12 genes. They excluded one gene from analysis due to it being a functional amyloid. When they graphed expression levels vs. aggregation propensities on a log-log scale, they discovered a strong inverse linear relationship with r=0.97. They then concluded that human proteins are produced at expression levels at the limits of aggregation, an idea which they dubbed the Edge Theory. This analysis has a number of potential flaws, however. It considers only 11 genes. Of those genes nine are known to be involved in disease and six are known to be involved specifically in aggregation-related diseases; this suggests that their results may not hold true for all genes. It excludes a gene on the basis of it being a functional amyloid without giving a clear definition of the criterion for such, and including that gene would dramatically reduce the correlation coefficient. This analysis uses mRNA levels, while one would expect that protein abundances would be limited by aggregation propensities. Finally and most critically, the linear relationship they have demonstrated shows only that the safety factor for expression levels vs. aggregation is *constant*, not that it is zero. Biological safety factors commonly range from 1.3 to 6 for load-bearing structural components, and the most common safety factor across systems where it has been measured is 2[72]. It would be surprising to find that proteins are produced with no safety factor at all, and the authors present no evidence supporting this conclusion.

The authors' next paper examines the relationship between mRNA expression levels and protein solubility in *E.coli* [73]. Since protein solubility data is not widely available, the solubility is instead predicted via analysis of the amino acid sequences.

By combining predictions based on factors like hydropathy, secondary structure propensity, and translation factors using a support vector machine, the final method is able to predict 83% of expression levels to within one order of magnitude and 92% to within two orders of magnitude. For context, mRNA expression levels in E.coli vary over about six orders of magnitude. The authors then use this method to predict the soluble fraction of 746 human proteins expressed in E.coli, and 86% were assigned to the correct soluble fraction of four. One problem with this analysis is that the methods are not compared to any baseline or null hypothesis, such as blindly assigning all proteins to the same order of magnitude or soluble fraction as the median or mode. I was unable to find the mRNA expression set used in the study, and could not determine the performance of a baseline algorithm directly from the dataset. However, it seems reasonable to guess that the distribution is approximately log-normal (this is the case in *S.cerevisiae*, results not shown), in which case their algorithm might well perform only slightly better than the baseline described above. It is not clear why the authors switched from their previous linear correlation metric to the classifying to within an order of magnitude of the correct value

The authors continue their search for anticorrelations between protein aggregation propensities and mRNA expression levels by examining sub-cellular localizations of human proteins[74]. They show that when the average predicted aggregation propensities for each sub-cellular compartment are plotted against the average mRNA expression levels for those compartments, there is a strong anticorrelation (r=-0.93 on the log-log scale). They then show that both of these properties strongly correlate or anti-correlate with the sub-cellular compartment volume (r=0.88 for aggregation propensity and r=-0.87 for mRNA level). It would be interesting

to see the results of a straightforward correlation analysis of predicted aggregation propensity vs. mRNA levels across all human genes. They did calculate the correlations within each sub-cellular compartment and found that they varied between -0.34 and 0.26. This suggests that the edge theory does NOT apply across all human proteins, although the scatter plots are not provided for the set of all proteins or for any of the compartments, making it difficult to tell precisely how the edge theory is failing in those cases.

Tango

The Serrano lab has developed an algorithm called TANGO which predicts aggregation propensities of proteins[6]. It uses the amino acid sequence, the pH, concentration, and ionic strength. Each segment of a protein can take up different conformations. The likelihood of the conformations is determined by the energy of the conformation according to the Boltzmann distribution. TANGO predicts the beta-aggregating portions of a peptide by calculating the partition function of the conformational phase-space. In short, TANGO predicts how readily each segment of the protein adopts a beta-aggregation conformation. Other factors included include hydrophobicity, beta-sheet propensity, electrostatics, hydrogen-bonding, and competition from alpha-helix and beta-turn conformation. Many of these factors are actually calculate by separate program called AGADIR. Five or more consecutive residues in the beta-aggregation conformation state were considered a strong predictor of aggregation (92% success rate on the test set). The authors note that the algorithm is less accurate at low levels of aggregation propensity and that TANGO cannot be used to compare proteins which differ widely in sequence.

Question of Interest #1 - Predicted aggregation propensity of Orthologs of paralogs vs. orthologs of non-paralogs

Is there a difference in TANGO scores in *K. Waltii* genes whose orthologs remained duplicated in *S.cerevisiae* after WGD versus those genes whose orthologs reverted to a single copy?

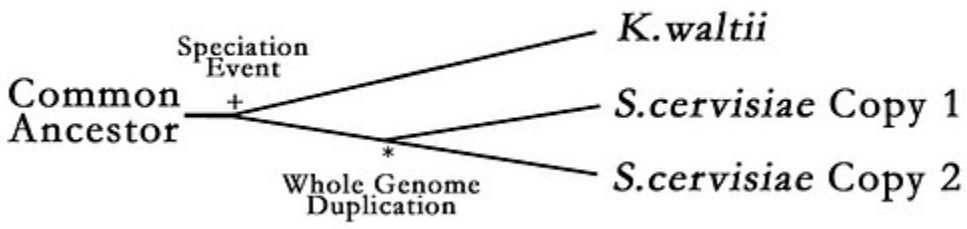

Figure 1. S.cerevisiae and K.waltii phylogenetic tree. Not to scale.

Introduction:

It has been theorized that most proteins in cells are under selection pressure to have a low enough aggregation propensity to remain soluble at the expression levels required for the cell to function, also known as the Edge Theory [2]. If true, any effect that produced changes in copy count of the gene would trigger strong selection pressure on the aggregation propensity of the protein. Whole genome duplication (WGD) results in two copies of a gene, which would be expected to increase expression level. It has been shown that the duplication of chromosomes produces an approximate doubling of gene expression across the entire duplicated chromosome in aneuploid yeast [75], and the

effect of duplicating all the chromosomes can be expected to be similar. If yeast proteins were already at the edge, the increase in expression level from WGD would push proteins over the edge and favor a return to a single copy, increased solubility, or reduced expression levels. The selection should be strongest in those proteins which are closest to the edge, namely those with high aggregation tendency, high protein abundance, or both. We tested this hypothesis by looking at the popular model organism *S.cerevisiae*, which has been previously shown to have undergone a WGD event,[76] possibly as an adaptation for rapid sugar metabolism[77]. *K.waltii* is a closely related yeast species which split off from *S.cerevisiae* ~150 MYA [78], prior to the WGD event, and serves as a control. I used the TANGO algorithm to predict the aggregation propensity for all *K.waltii* genes based on their amino acid sequence and compared genes which reverted to single copy in *S.cerevisiae* with those that were maintained as two copies. This allows me to use *K.waltii* as a baseline to control for changes in aggregation propensity which occurred after the WGD event in *S.cerevisiae*. The same experiment was repeated using *S.cerevisiae*. Finally, overall TANGO scores for the *K.waltii* and *S.cerevisiae* proteomes were compared.

Methods:

K.waltii protein sequences were obtained from the supplementary data of the Kellis WGD paper, as was the list of genes that were retained as duplicates in *S.cerevisiae*. The protein sequences were fed into TANGO.exe using the standard settings and the total TANGO score for each was recorded.

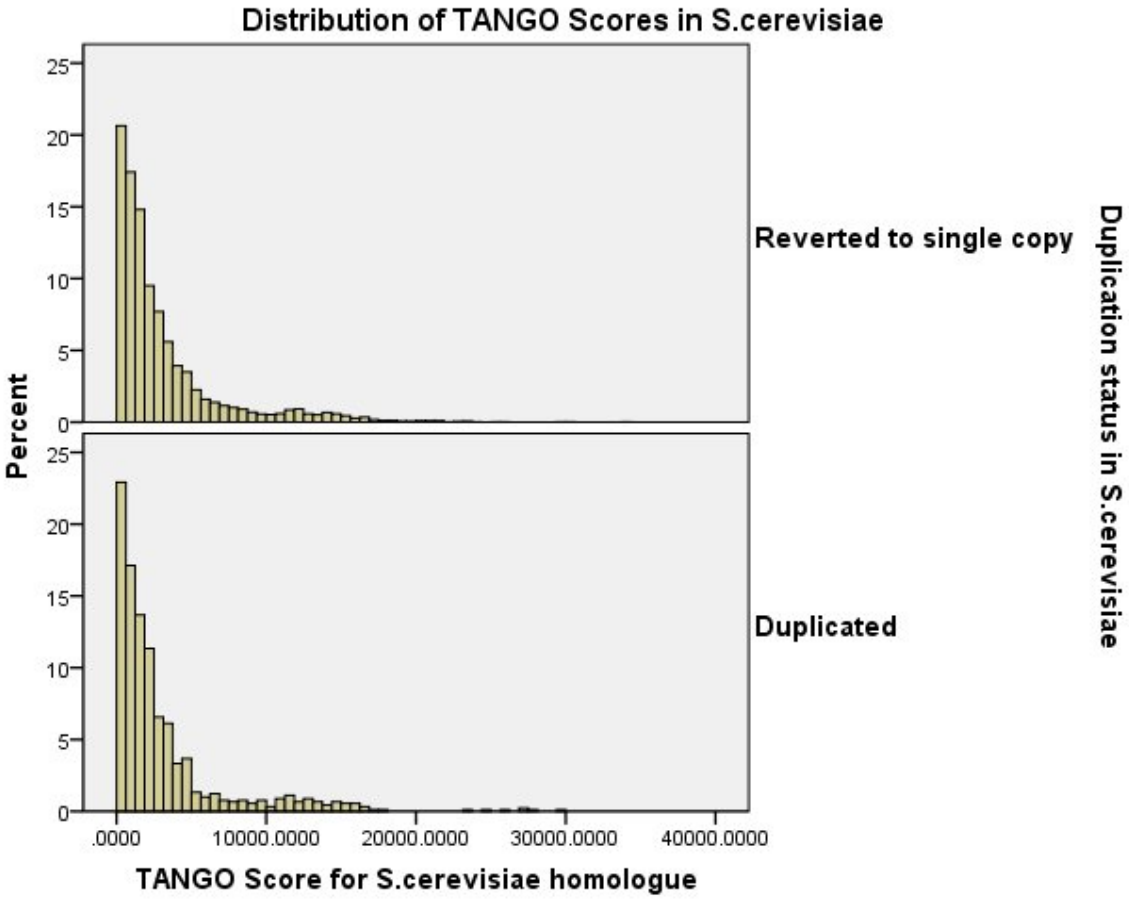

Figure 2: The distributions of TANGO score in S.cerevisiae for genes which were retained as duplicates vs. genes which reverted to singletons. The median TANGO score for duplicated genes was 1712, while the median TANGO score for singletons genes was 1752. The distributions were not statistically significant (p=0.16 WMW, n1=899, n2=4342).

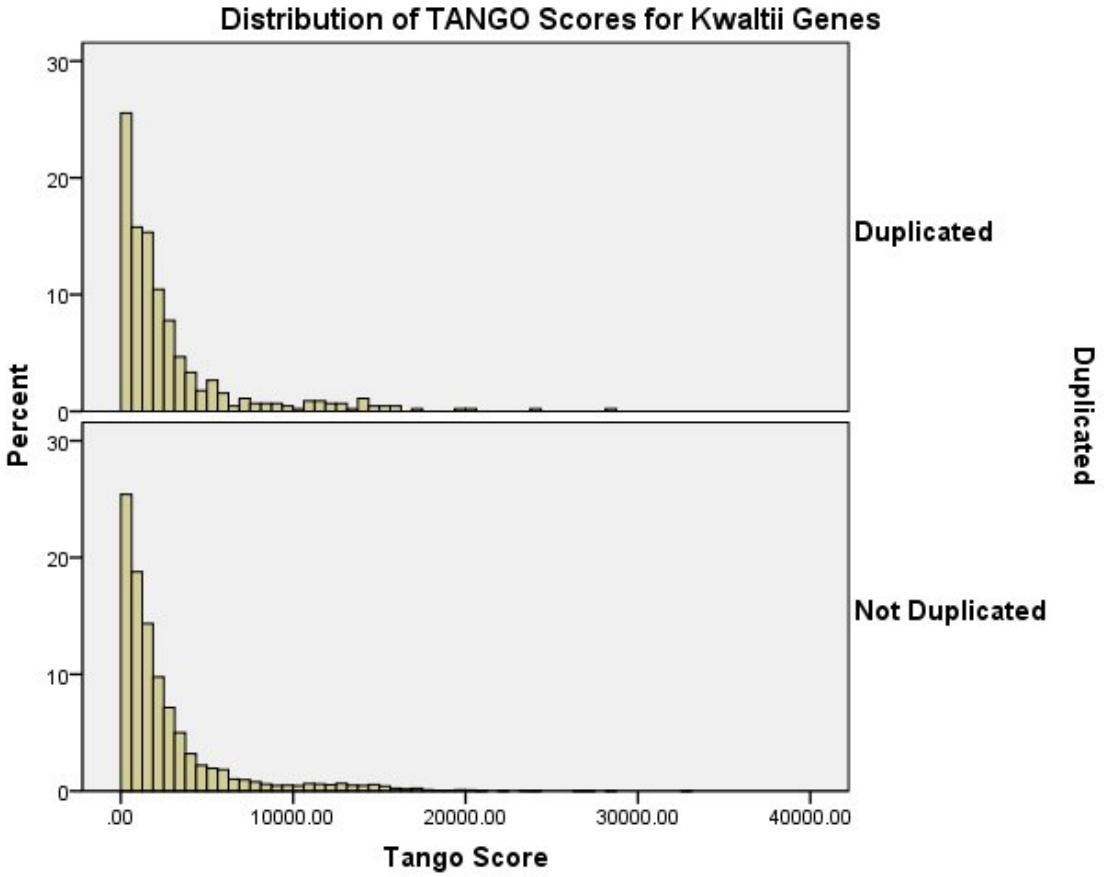

Figure 3 Distributions of TANGO scores for orthologs of genes where both duplicates were retained as duplicates vs. orthologs of genes which reverted to singletons. Distributions were not significantly different. (p=0.550 WMW test, p=0.705 KS test)

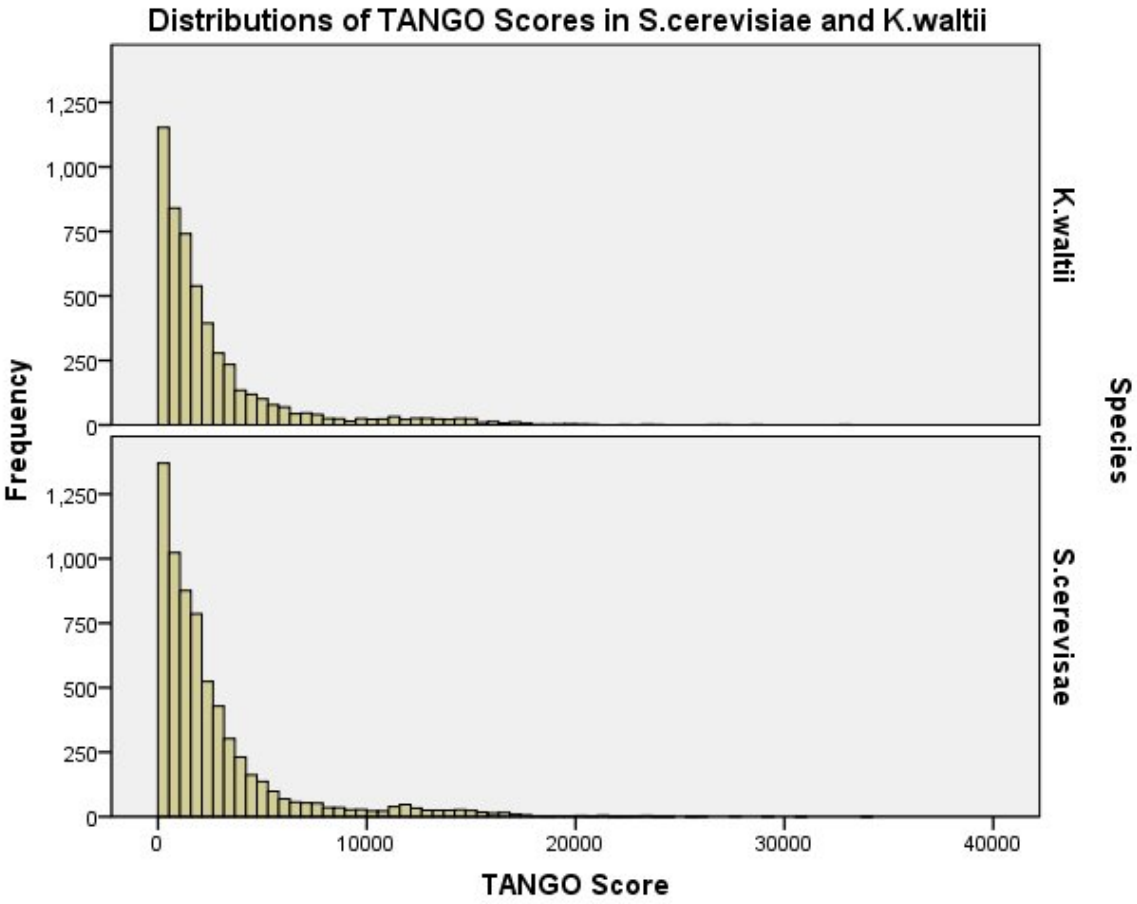

Figure 4 Distributions of log TANGO scores for all genes in *K.waltii* and all genes in *S.cerevisiae*. *S.cerevisiae* has a statistically significantly higher median score (1477 vs. 1626, p=0.003 WMW)

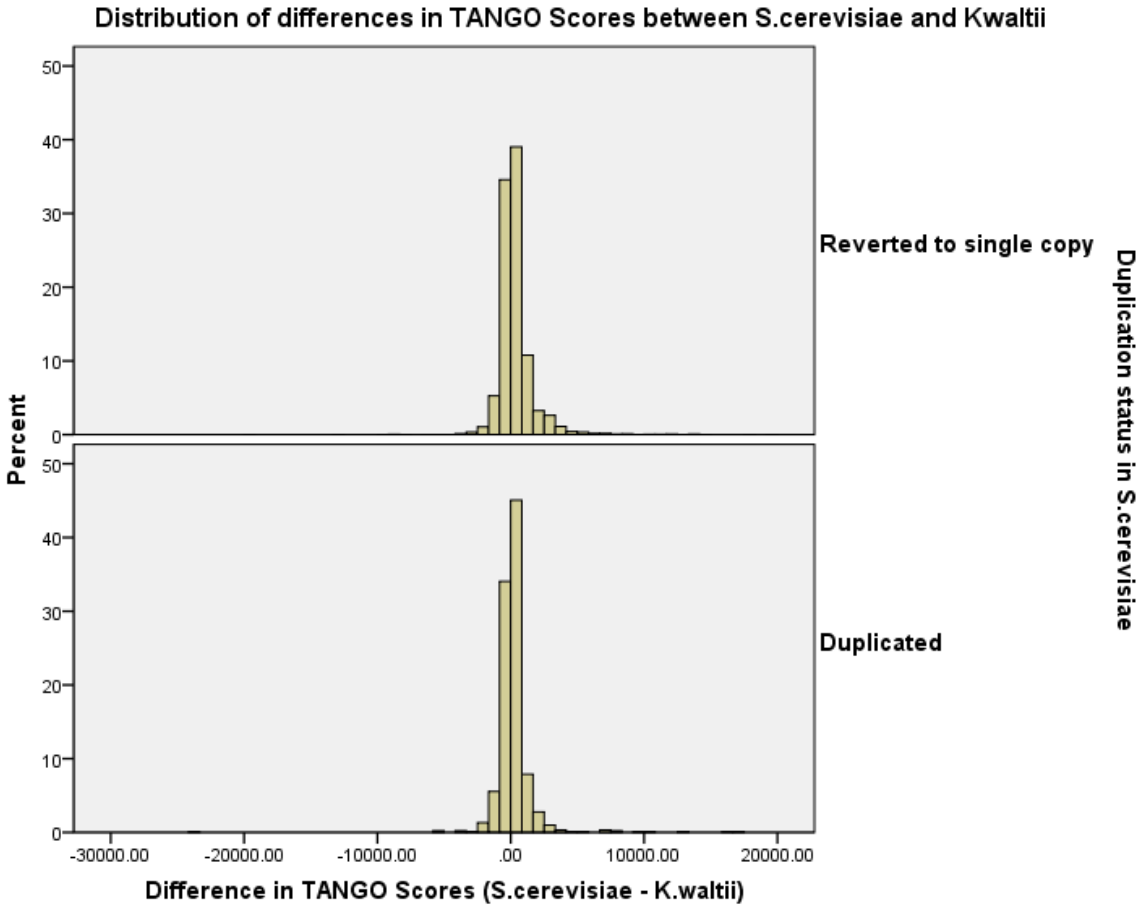

Figure 5: The distribution of differences in TANGO score between orthologous genes in *S.cerevisiae* and *K.waltii*. Median difference was smaller for duplicated genes (42 vs. 74). Distributions were significantly different (KS test, p=0.005)

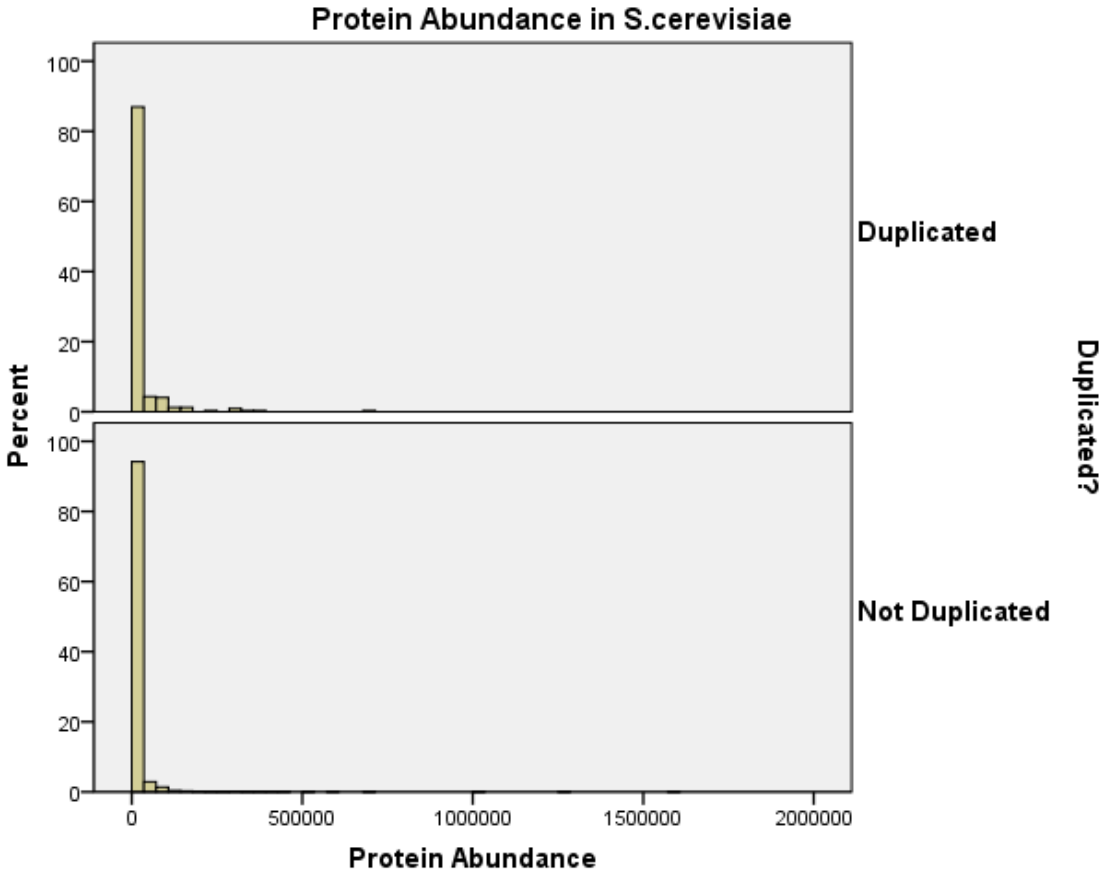

Figure 6: Distribution of protein abundances in *S.cerevisiae*. The median abundance for duplicated genes was 2490, while the mean abundance for singleton genes was 2750. This difference was statistically significant (p=0.042, KS test, n1 = 321, n2 = 3159).

Results:

Duplication status does not affect TANGO score in *S.cerevisiae*. The median TANGO score for duplicated genes was 1712, while the median TANGO score for singletons genes was 1752. This difference was not statistically significant (p =0.16 WMW, n1=899, n2=4342).

The aggregation propensity of genes in *K.waltii* does not predict whether or not their paralogs in *S.cerevisiae* were retained as duplicates or reverted to singletons. In

K.waltii, paralogs of duplicated genes had a median TANGO score of 1575. Paralogs of singleton genes had a median TANGO score of 1477. The two samples were not significantly different (p=0.550, WMW test, n1=441, n2=5150).

TANGO Scores are higher in *S.cerevisiae*. The median TANGO score in *S.cerevisiae* was 1626, vs. 1427 for *K.waltii*, and this difference was statistically significant (p=0.003, WMW, n1=5150, n2=6610).

TANGO scores of duplicated genes are less different from their orthologs than genes which reverted to singletons. This is a paired analysis, using the differences between orthologous genes in *S.cerevisiae* and *K.waltii*. TANGO scores are very similar for orthologous genes (scatter plot not shown), so a paired analysis is reasonable. The median difference was smaller for duplicated genes (42 vs. 74). The distributions of differences were significantly different for duplicated genes vs. genes which returned to single copy (KS test, p=0.005).

Duplicated genes have higher protein abundance per gene than genes which reverted to a single copy. In *S.cerevisiae*, the median abundance for duplicated genes was 2490, while the mean abundance for singleton genes was 2750. This difference was statistically significant (p=0.042, KS test, n1 = 321, n2 = 3159).

Conclusion:

The clearest result is that *S.cerevisiae* genes have higher aggregation propensities than *K.waltii*. This could be due to WGD or selection pressures due to environmental differences, particularly since *S.cerevisiae* has been domesticated for ~6000 years[79]. TANGO scores are not significantly different between duplicates and singletons in *S.cerevisiae*, or between their orthologs in *K.waltii*. This suggests that

aggregation propensity did not influence which genes were retained as duplicates. While TANGO scores are higher overall in *S.cerevisiae* than *K.waltii*, the differences are larger in genes which reverted to single copy relative to those which remained as duplicates in *S.cerevisiae*. This suggests that after WGD, there was selection pressure on the genes which were retained as duplicates to reduce their aggregation propensity, or at least pressure which prevented their aggregation propensity from rising as fast as other genes.

Assuming that TANGO score is an accurate measure of aggregation propensity, then aggregation propensity is not a major selective force driving the loss or elimination of duplicate copies of genes after WGD. TANGO scores are higher in *S.cerevisiae* than in *K.waltii*, but it is unclear whether this rise is due to the WGD event. Since the ancestral state is unknown, it is unclear whether TANGO scores increased in *S.cerevisiae* or fell in *K.waltii*. Furthermore, the *K.waltii* proteome was generated by predicting ORFs in the *K.waltii* genome after sequencing, and may not be accurate. In particular, ORF prediction must guess at intron locations and in many cases will incorrectly identify start codon locations. Finally, TANGO was not intended to be used to measure absolute aggregation propensity, only relative propensities of mutants. Therefore, the comparison of duplicated vs. nonduplicated genes may not be a valid application of the algorithm. These two factors may have contributed to a lack of power in the comparison of duplicated to non-duplicated genes. The analyses which compared orthologs using TANGO should be a valid application, and these analyses found significant differences. Highly abundant proteins are more likely to be retained as duplicates, which is the opposite of what the edge theory would predict. However, abundance is strongly correlated with essentiality, which may overwhelm the effects of

any aggregation-related selection pressure. Taken all together, this evidence suggests that genes for highly abundant proteins tend to be retained after WGD, and that genes which remain duplicated reduce their aggregation propensity.

Question of Interest #2: AA sequence pressure

Does WGD produce selection pressures on aggregation propensity at the level of amino acid sites? First, do I see more mutations after WGD at sites that have high or low aggregation propensity? Second, if mutations do occur do those mutations raise or lower the aggregation propensity?

Introduction

The Edge Theory predicts that duplication of genes should produce an increase in gene copy count which would raise expression levels and thus cause aggregation. Therefore, there should be selection pressure on genes which remained as duplicates to reduce their aggregation propensity. This was tested by looking at point mutations which occurred after whole genome duplication in genes where both copies were retained in the yeast *S.cerevisiae*. In order to tell which point mutations occurred after the WGD event, the closely related yeast *K.waltii* was used as a control. I used the 450 gene alignments of *S.cerevisiae* paralogs with their *K.waltii* orthologs from Kellis, et al. [76]. A section of one such alignment is shown.

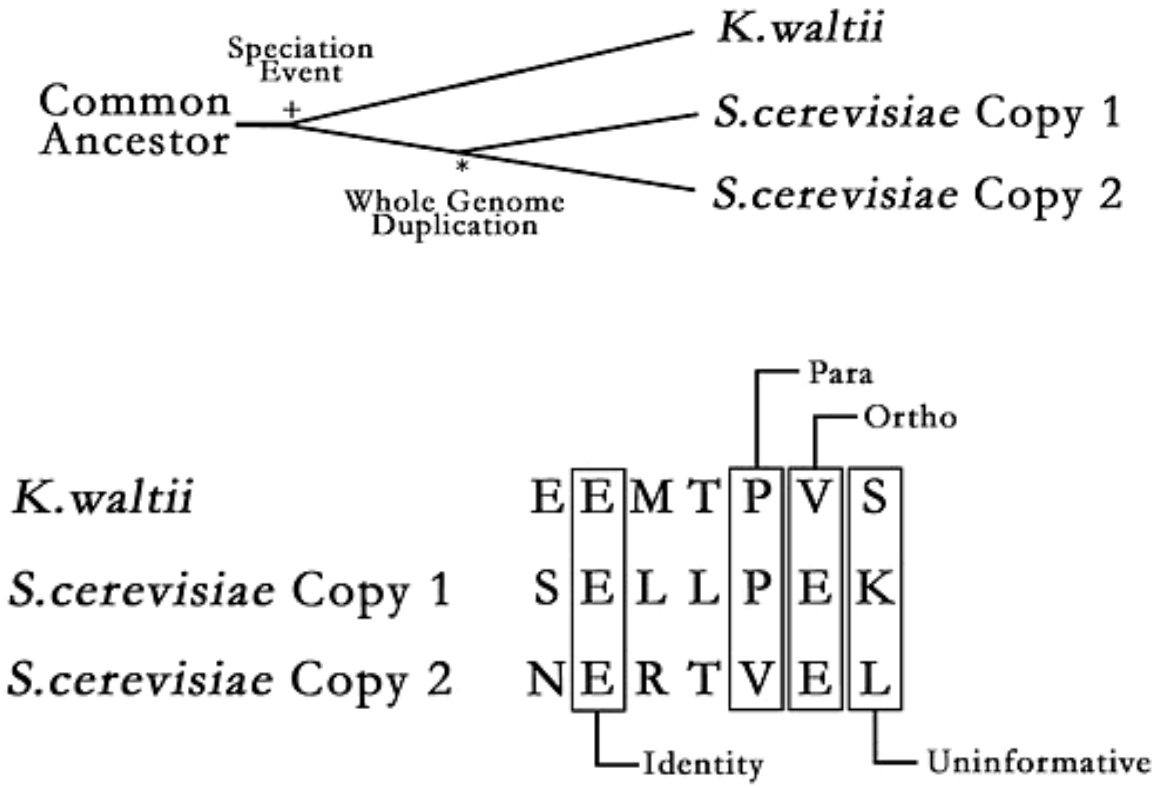

Figure 7. Phylogeny of K.waltii and S.cerevisiae, as well as a selected subsection of a three-gene alignment between K.waltii and the two homologues in S.cerevisiae resulting from whole-genome duplication (WGD). Para sites are those where one S.cerevisiae copy matches the AA in K.waltii, but the other S.cerevisiae copy has a different AA. These indicate a mutation occurring in S.cerevisiae after WGD. Ortho sites are those were both S.cerevisiae copies match each other but differ from K.waltii. These indicate that a mutation occurred either in S.cerevisiae before WGD or in K.waltii. Sites where all three genes have different AA are uninformative because they do not allow us to infer the ancestral state.

There is a speciation event between *K.waltii* and *S.cerevisiae*, and then a WGD in *S.cerevisiae*. To simplify the analysis, only amino acid sites compatible with the assumption of one mutation from the common ancestor of the two organisms will be

considered (assumption of parsimony). If a mutation occurs in *K.waltii* after speciation, then both copies in *S.cerevisiae* will have the same amino acid, and the *K.waltii* copy will have a different amino acid at that site. If there was a mutation in *S.cerevisiae* after speciation but before the WGD event, the same signature appears. Those mutations will be referred to as ortho mutations because the orthologs have different amino acids at that site. In both of these cases, the mutation occurred when there was only one copy in the cell. These mutations will serve as a control. The other informative case is when there is a mutation in *S.cerevisiae* after WGD, and this produces a signature where the two copies in *S.cerevisiae* are different, but one of them will match the *K.waltii*. These are mutations which must have occurred after the WGD event and in the presence of two copies of the gene. These mutations will be referred to as para mutations because the amino acids at that site are different between two paralogs. If the edge theory is correct, I would expect to see selection pressure associated with aggregation effects at these sites. If there were no mutations, then the amino acid will be the same in all three copies and this provides no information about selection pressure. If all three copies have a different amino acid, then there were multiple mutations and I don't know where they occurred, so they provide no information about selection pressure.

Strong selection pressures to reduce aggregation are predicted to have two effects: The mutations should be preferentially located at sites which have a high local aggregation propensity (as changes at these sites would have the most dramatic effects) and the mutations should result in a lower aggregation propensity.

Methods:

450 alignments of three genes were taken from Kellis, et al. [76]. Alignments that had fewer than 3 para sites (n=40) or fewer than 3 ortho sites (n=6) were discarded as uninformative. Alignments in which more than 50% of sites were insertion/deletion mutations were discarded as well (n=63); aligning sequences with high indel rates is difficult to do reliably in any case and automated methods in particular tend not to do well on these sets. This is exacerbated by the fact that open reading frames (ORFs) identification was automated by Kellis, et al.; these alignments do not correspond to sequenced proteins but merely ORFs which were predicted to correspond to actual genes. As such, the start codons and location of introns were predicted based on genome sequence alignment with *S.cerevisiae*. Many of the rejected genes have large deletions in *K.waltii* at the N-terminus, which is probably due to incorrectly identified start codons rather than true deletion mutations. Since the following analysis is entirely dependent on the assumption that the alignments given are correct, the 63 genes with a high percentage of insertion/deletion sites in the alignment were rejected. This left 375 alignments.

TANGO.exe was used to generate TANGO scores for each site of all three genes of the alignments. The mean score was calculated over all sites, all para sites, and all ortho sites for each alignment. The experimental unit is a three-gene alignment rather than a particular AA site. It would be unreasonable to assume that the mutations in the same gene are independent of each other, so analysis was carried out on the mean scores for each alignment.

In addition, at para sites, the ancestral state could be inferred and so the difference in TANGO score from the new mutation to the ancestral state was calculated.

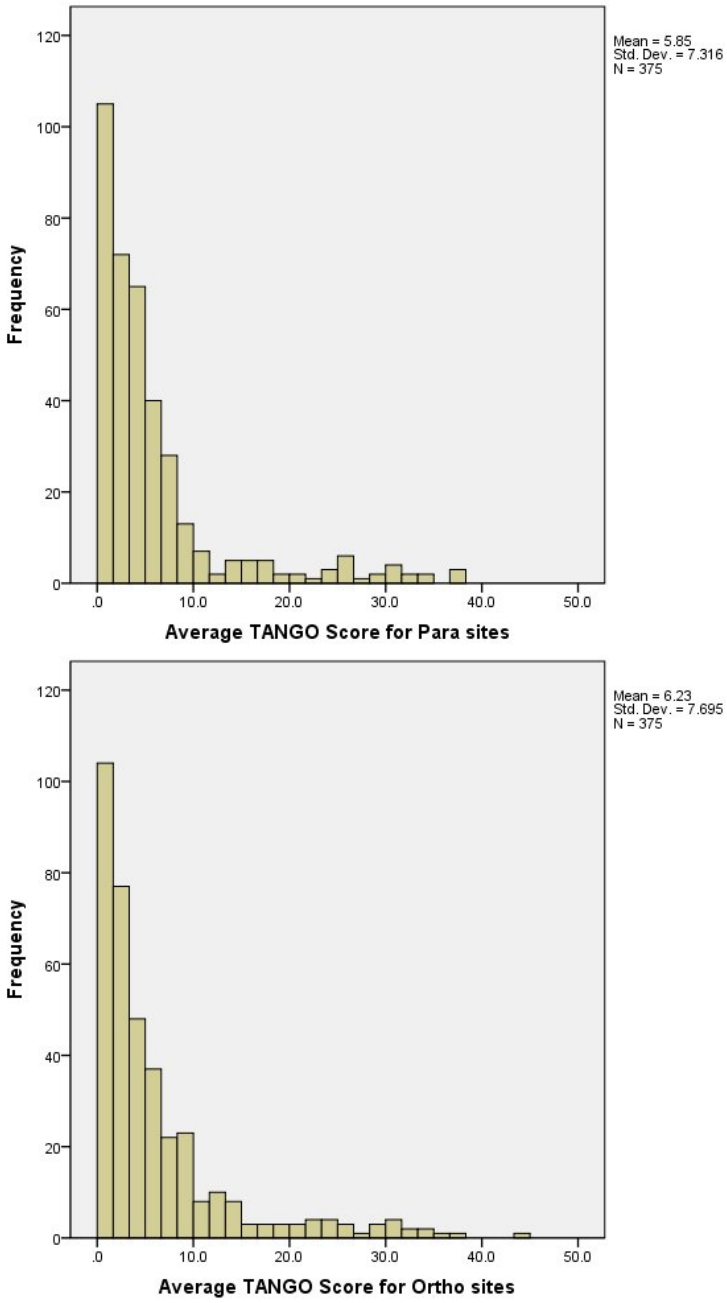

Figure 8. Distribution of mean TANGO scores for para and ortho sites for three-gene alignments. TANGO scores were not significantly different (p=0.09 for one-sample t-test of the differences).

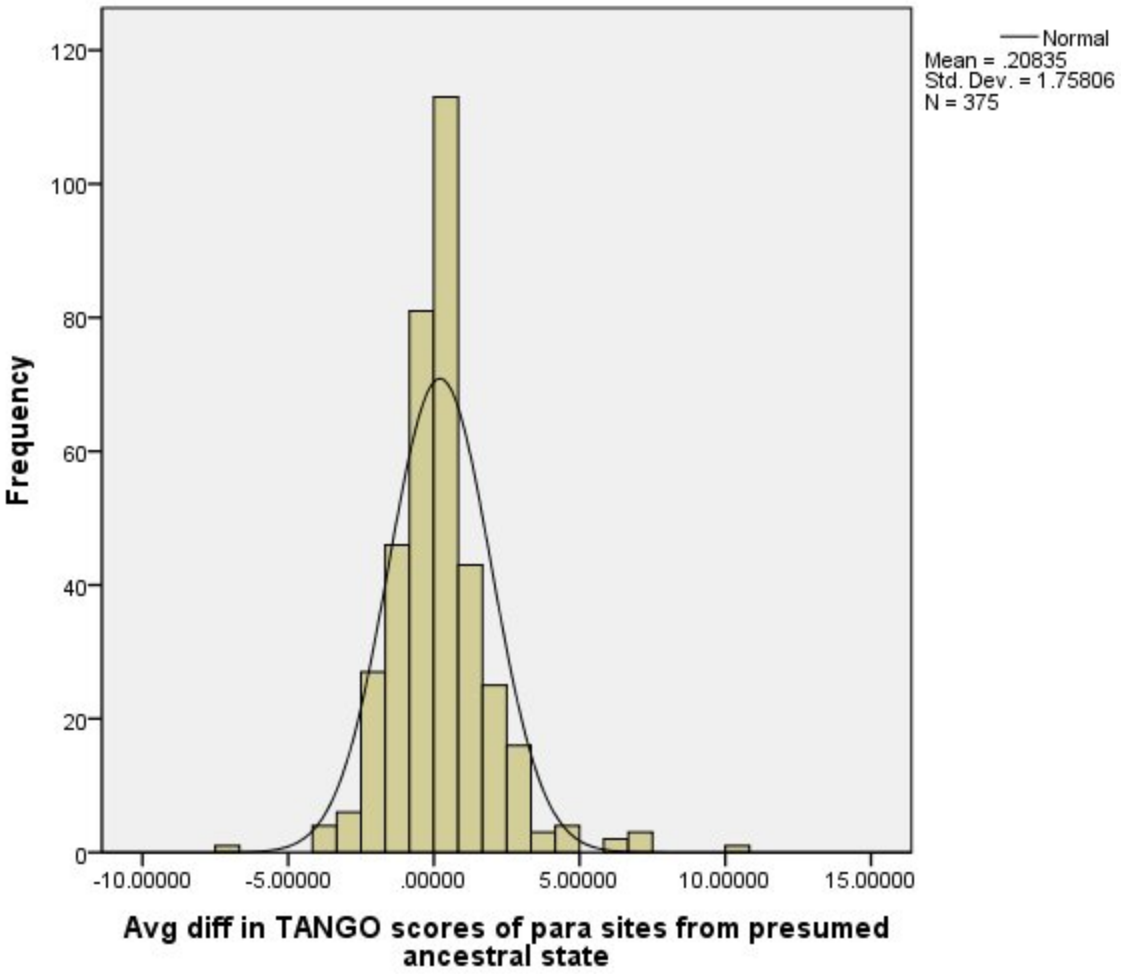

Figure 9. The distribution of the average difference in TANGO scores of the new mutation from the inferred ancestor for 375 genes retained as duplicates. Mean TANGO scores increased on average by 0.208, which was statistically significant (p=0.022, one-sample t-test).

Results:

Higher mean TANGO scores were observed at ortho sites than at para sites: 6.35 vs. 5.98. This difference was not significant (p=0.09 for one-sample t-test of differences).

Para mutations resulted in a mean increase in TANGO score of 0.208. This increase was statistically significant (one sample t-test, p=0.022, 95% CI of 0.0298, 0.386).

The average magnitude of the change in TANGO score per gene was higher for para sites than for ortho sites (median of 2.99 for para sites, 2.84 for ortho sites, n=366, p=0.025 for a paired samples t-test of the log-transformed data, mean log para was 0.3870 vs. mean log ortho 0.3290.).

Para mutations are roughly twice as common as ortho mutations (44232 para mutations vs. 21345 ortho mutations, significantly different, paired t-test of the percent para per alignment vs. the percent ortho per alignment, p<0.0005).

Conclusion:

The edge theory predicts that any increase in copy count or aggregation propensity will produce damaging levels of aggregation. If this is correct, whole genome duplication should produce an increase in gene copy count and a corresponding increase in expression level for those genes retained as duplicates [80] [81], based on the fact that aneuploidy in yeast results in a doubling of gene expression along the entire length of a doubled chromosome [75]. This increase can be compensated for by mutations. The most common compensatory mutation is simple deletion or inactivation. However, for those genes where both copies are retained, I would expect mutations to preferentially reduce the aggregation load.

First, I would expect the mutations to occur preferentially at sites strongly contributing to the overall aggregation propensity of the protein. However, the average aggregation propensity of sites where mutations occurred under WGD conditions is not

statistically significantly different from the average aggregation propensity of sites where mutations occurred under non-WGD conditions. This says that there is no pressure for mutations to occur at highly aggregation-prone sites.

Second, I would expect mutations to occur which lower the aggregation propensity of the protein. For sites where I can identify a mutation as having occurred under WGD conditions, I can also identify the ancestral state and determine the direction of the change in aggregation propensity. Contrary to what the Edge Theory would predict, I actually see mutations which cause an increase in aggregation propensity under WGD conditions. However, aggregation propensity in *S.cerevisiae* is higher overall. Since *S.cerevisiae* has been domesticated for thousands of years [79], it is possible that the increase in aggregation propensity is an adaptation to domestication rather than a response to WGD, or to another difference in environments. The ideal test would be to examine the genomes of many organisms which have undergone WGD and compare them to close relative which have not.

We would expect aggregation selection pressures to produce larger changes in aggregation propensity than other selection pressures. Therefore, para sites should show larger changes in TANGO score than ortho sites, regardless of the direction. This was confirmed (paired samples t-test of log-transformed scores $p=0.025$) although the evidence is not very strong.

Cells could also respond to increased aggregation loads through mutations which reduce expression levels or protein abundance, but those kinds of mutations are more difficult to identify and were not analyzed.

Question of Interest #3: Testing the Edge Theory using the *S.cerevisiae* data

Is there a strong inverse linear relationship between the log aggregation propensity of a protein and the log protein abundance or log mRNA expression for proteins in *S.cerevisiae*? Is the same trend present for any subgroups of genes?

Introduction:

I am going to directly reproduce the original Edge Theory experiment [2], which was to compare aggregation propensity to protein abundance, except that I am going to do it for the entire *S.cerevisiae* proteome rather than a selection of human proteins and I will use the TANGO algorithm to predict the aggregation propensity rather than measuring it in vitro. Then, the predicted aggregation propensity will be compared to the measured protein abundance as well as to the measured mRNA expression levels.

Methods:

Predicted protein aggregation propensities were calculated using the TANGO algorithm. Protein abundance data for *S.cerevisiae* genes was obtained from the Marcotte lab, with collection details described in Laurent et al [82]. The data is available at http://www.marcottelab.org/MSdata. mRNA expression levels were taken from the same paper.

Results:

I compared TANGO scores vs. protein abundance and three different mRNA expression level datasets. Since TANGO scores, protein abundance, and mRNA levels are approximately log-normally distributed, and because the linear regression observed in the Edge Theory paper was done on the log-transformed aggregation and expression levels, all analysis were performed on the log-transformed data. Linear regressions

were performed for TANGO score vs. abundance and TANGO score vs. mRNA expression levels.

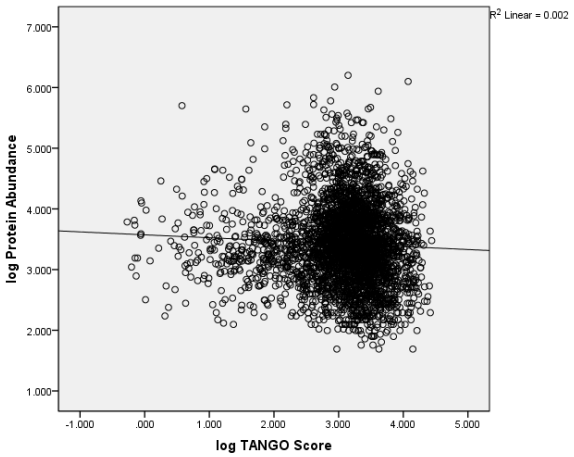

Figure 10. Log-transformed protein abundance vs log TANGO score. r = 0.049, p < 0.005.

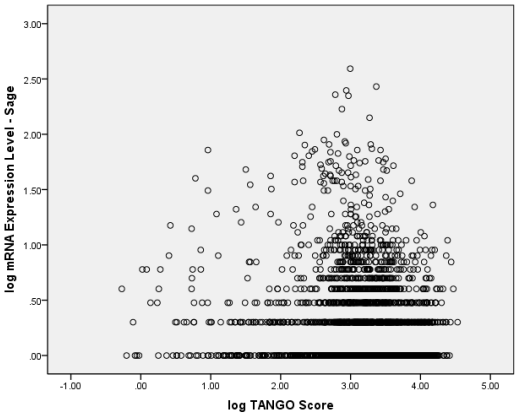

Figure 11. Log transformed mRNA expression level (Sage data set) vs. log TANGO score. r = 0.095, p < 0.005

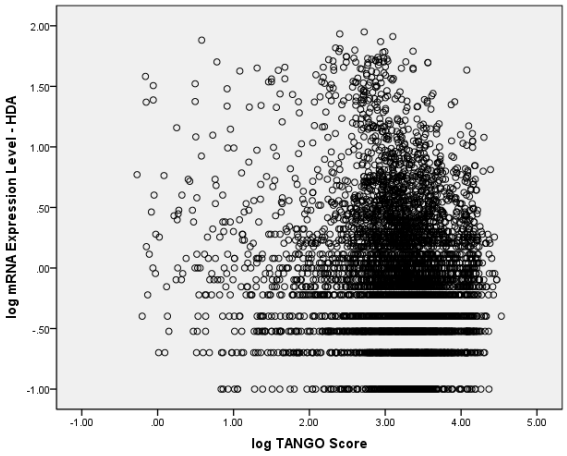

Figure 12. Log transformed mRNA expression level (HDA data set) vs. log TANGO score. r=0.115, p<0.005.

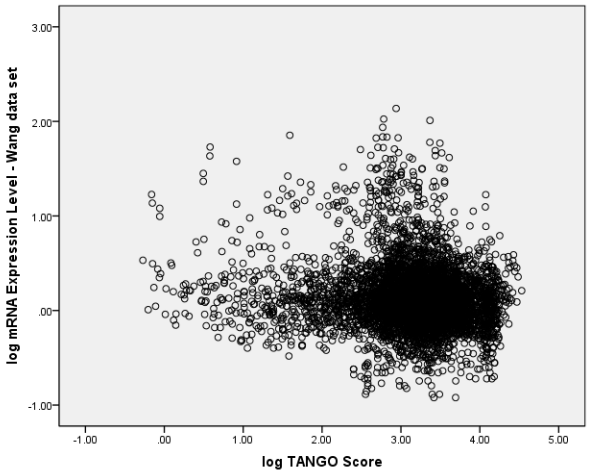

Figure 5. Log transformed mRNA expression level (Wang data set) vs. log TANGO score. r=0.128, p<0.005.

	N	r	r^2	P
Protein Abundance	2517	0.049	0.0024	< 0.005
SAGE mRNA expression level	2467	0.095	0.009	< 0.005
HDA mRNA expression level	5301	0.115	0.013	< 0.005
Wang mRNA expression level	5546	0.128	0.016	< 0.005

Figure 6. Linear regressions of protein abundance and mRNA expression level datasets against TANGO score. All datasets log-transformed.

All of the regressions were significant with negative slope but all had dramatically smaller r values than the 0.95 reported by Tartaglia, et al[2]. The edge theory predicts that all proteins evolve to the highest aggregation propensity allowed by their ideal expression level, and that the relationship between these is linear on a log-log scale. If correct, all graphs should show a strong linear relationship. Instead, none of the graphs show a linear relationship, let alone a strong one. It looks like there might be two different populations of genes, one with similar aggregation propensities and varied abundance and another with differing aggregation propensities but similar abundances.

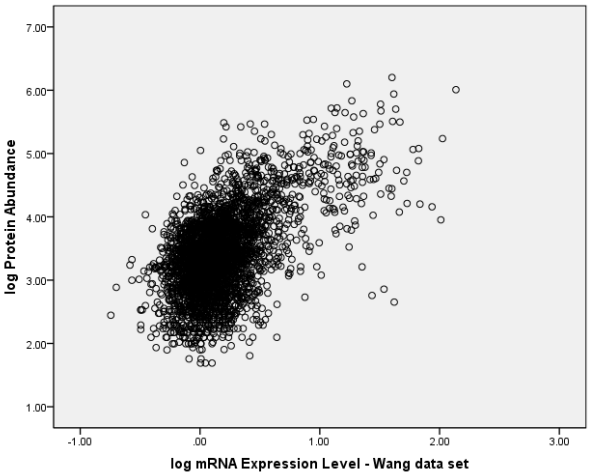

Figure 7. Log protein abundance vs. log mRNA expression level (Wang data set). $r = 0.556$, $p<0.0005$.

Protein abundance and mRNA expression levels are loosely correlated. For the SAGE mRNA expression level data set, 25% of the variation in log protein abundance is explainable by the variation in log mRNA expression level ($p < 0.0005$). For the HDA data set and Wang data set, the numbers are 37.9% and 31%. For the SAGE dataset, the abundance varies over 2-3 orders of magnitude for a particular mRNA expression level. For the HDA data set, the abundance varies over about 2 orders of magnitude for any given mRNA expression level. For the Wang dataset, it varies over up to 4 orders of magnitude.

Similar plots were constructed for subgroups of protein based on gene ontology classifications as well as on sub-cellular compartment localization. It was hoped that the Edge Theory might hold for some of these subgroups, despite not appearing to hold for proteins in general. Scatter plots were generated for the following compartments: actin, ambiguous, bud, bud neck, cell periphery, cytoplasm, early Golgi, endosome, ER, ER to Golgi, Golgi, late Golgi, lipid particles, microtubules, mitochondria, nuclear periphery,

nucleolus, nucleus, peroxisome, punctate composite, spindle pole, vacuolar membrane, and vacuole. None of the subcompartments showed a strong linear correlation by visual inspection. Scatter plots were also generated for 67 groups of genes by gene-ontology keyword. Visual inspection again failed to find any groups which showed strong linear correlations. The cytosolic compartment data looks very similar to the overall data, which eliminates any concern that transmembrane proteins affected the overall analysis.

Conclusion:

Protein aggregation does not have a simple correspondence to mRNA expression levels or to protein abundance. Protein abundance correlates less well with TANGO scores than the various measurements of mRNA expression levels, although since the relationship is not linear this may not be a useful metric. The original Edge paper also used mRNA expression levels rather than abundance[2]. The fact that mRNA is more closely associated with aggregation than abundance is fairly surprising, since the TANGO score measures the aggregation propensity of the protein rather than the mRNA coding for it. Aggregation effects should act directly on the protein and only indirectly on the mRNA expression levels. This may be explainable by the fact that mRNA expression levels are more closely related to turnover rates, while protein abundance measures how much of the protein is present. Properly folded proteins should be less vulnerable to aggregation since their hydrophobic residues are buried, so aggregation is a concern only before the protein has folded or if the protein is later unfolded. Proteins with a long half life might spend the same amount of time folding but much more time in their folding state and thus present a lower aggregation risk. It might be very interesting to compare aggregation propensities to turnover rates.

Final Conclusions

Aggregation propensity does not affect which genes are retained as duplicates after WGD, regardless of whether the aggregation propensity is measured in *S.cerevisiae* or for the homologues in *K.waltii*. Aggregation propensity was higher in the species with WGD, but the difference was smaller for genes which were retained as duplicates. Abundant genes are more likely to be retained as duplicates, but this may be related to essentiality rather than aggregation selection pressures.

At the amino acid level, we did not see a preference for mutations to occur at sites with high aggregation propensities. Mutations which occurred under conditions of duplication tended to increase TANGO scores, but this may be due to the overall rise in TANGO scores in *S.cerevisiae*. It does suggest that the difference in overall aggregation propensity for *S.cerevisiae* proteins vs. *K.waltii* proteins is due to an increase in S.cerevisiae rather than a decrease in *K.waltii* from the ancestral state. Mutations which occurred under conditions of duplication also produced larger changes in aggregation propensity than mutations which did not occur under conditions of duplication. Finally, mutations which occurred under conditions of duplication occurred twice as often as mutations not under conditions of duplication. Combined with the result that proteins retained as duplicates had less of an increase in aggregation than other proteins, these results suggest that there was fairly noticeable selection pressures on duplicates after WGD. There appears to have been a general relaxation of aggregation-related selection pressure in *S.cerevisiae* after WGD, but the relaxation was less dramatic for duplicated genes, suggesting that duplication really did increase aggregation.

A potential confounding factor is the possibility that proteins preferentially aggregate with themselves; with two different copies of a gene, each copy may mutate to the point where it forms two different pools of protein, each of which aggregates separately. If so, then having two versions of the gene actually decreases the selection pressures associated with aggregation since the total amount of protein which can be safely produced is effectively doubled. Indeed, studies of Alzheimer's disease genotypes suggest that having two different alleles of the protein which forms amyloid fibrils in Alzheimer's can produce less severe symptoms than being homozygous for either allele[83].

The Edge theory is based on an observation that mRNA expression levels and protein aggregation propensity were strongly inversely correlated. This trend was not observed in *S.cerevisiae*. It is possible that aggregation is more about protein turnover rate than protein abundance in general, and this might be an interesting area for future research.

Based on this research, it seems that there is no simple rule which relates a protein's abundance or expression level to the aggregation propensity in *S.cerevisiae*. There is some evidence that duplication of genes triggers increased aggregation pressures for those genes, but abundance rather than aggregation propensity predicts which genes were retained as duplicates. Since *S.cerevisiae* has higher overall aggregation propensities than its close relative, it may be that one of the prerequisites for WGD is a relaxation of aggregation pressures, so that the rise in aggregation issues from duplication can be tolerated. An analysis which constructed phylogenetic trees that included many more related yeast species would allow the inference of the ancestral state for all genes, which would help answer that question. None of my results can

comment directly on the claim of the Edge Theory that proteins have no safety factor in their expression level vs. their aggregation propensity, but given the wide variations in both mRNA expression levels and protein abundance levels at particular aggregation propensities, as well as the apparently small selection pressures associated with aggregation effects, it seems likely that most proteins in yeast are not right at the limit. However, the fact that duplicated genes had smaller aggregation propensity differences suggests that the safety factor between the actual aggregation propensity and the ideal propensity is less than two for at least some genes.

References

1. Come, J.H., Fraser, P.E. & Lansbury, P.T. A kinetic model for amyloid formation in the prion diseases: importance of seeding. *Proc Natl Acad Sci U S A* **90**, 5959-5963 (1993).
2. Tartaglia, G.G., Pechmann, S., Dobson, C.M. & Vendruscolo, M. Life on the edge: a link between gene expression levels and aggregation rates of human proteins. *Trends Biochem. Sci.* **32**, 204-206 (2007).
3. Rousseau, F., Schymkowitz, J. & Serrano, L. Protein aggregation and amyloidosis: confusion of the kinds? *Curr. Opin. Struct. Biol.* **16**, 118-126 (2006).
4. Wozniak, M.A., Mee, A.P. & Itzhaki, R.F. Herpes simplex virus type 1 DNA is located within Alzheimer's disease amyloid plaques. *J. Pathol.* **217**, 131-138 (2009).
5. Tartaglia, G.G. & Vendruscolo, M. The Zyggregator method for predicting protein aggregation propensities. *Chemical Society Reviews* **37**, 1395 (2008).
6. Fernandez-Escamilla, A.-M., Rousseau, F., Schymkowitz, J. & Serrano, L. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat Biotech* **22**, 1302-1306 (2004).
7. Linding, R., Schymkowitz, J., Rousseau, F., Diella, F. & Serrano, L. A comparative study of the relationship between protein structure and beta-aggregation in globular and intrinsically disordered proteins. *J. Mol. Biol.* **342**, 345-353 (2004).
8. Tyedmers, J., Mogk, A. & Bukau, B. Cellular strategies for controlling protein aggregation. *Nat Rev Mol Cell Biol* **11**, 777-788 (2010).
9. Voet, D. & Voet, J.G. *Biochemistry (BIOCHEMISTRY*. (Wiley: 2010).
10. Chai, Y., Shao, J., Miller, V.M., Williams, A. & Paulson, H.L. Live-cell imaging reveals divergent intracellular dynamics of polyglutamine disease proteins and supports a sequestration model of pathogenesis. *Proceedings of the National Academy of Sciences* **99**, 9310 -9315 (2002).
11. Nucifora, F.C. *et al.* Interference by Huntingtin and Atrophin-1 with CBP-Mediated Transcription Leading to Cellular Toxicity. *Science* **291**, 2423 -2428 (2001).
12. Steffan, J.S. *et al.* The Huntington's disease protein interacts with p53 and CREB-binding protein and represses transcription. *Proc. Natl. Acad. Sci. U.S.A.* **97**, 6763-6768 (2000).
13. Goldberg, A.L. Protein degradation and protection against misfolded or damaged proteins. *Nature* **426**, 895-899 (2003).
14. Hartl, F.U. & Hayer-Hartl, M. Converging concepts of protein folding in vitro and in vivo. *Nat Struct Mol Biol* **16**, 574-581 (2009).
15. Ellgaard, L. & Helenius, A. Quality control in the endoplasmic reticulum. *Nat. Rev. Mol. Cell Biol.* **4**, 181-191 (2003).
16. Meusser, B., Hirsch, C., Jarosch, E. & Sommer, T. ERAD: the long road to destruction. *Nat Cell Biol* **7**, 766-772 (2005).
17. Rubinsztein, D.C. The roles of intracellular protein-degradation pathways in neurodegeneration. *Nature* **443**, 780-786 (2006).
18. Bukau, B., Weissman, J. & Horwich, A. Molecular chaperones and protein quality control. *Cell* **125**, 443-451 (2006).

19. Horwich, A.L., Fenton, W.A., Chapman, E. & Farr, G.W. Two Families of Chaperonin: Physiology and Mechanism. *Annual Review of Cell and Developmental Biology* **23**, 115-145 (2007).
20. Kirkin, V., McEwan, D.G., Novak, I. & Dikic, I. A Role for Ubiquitin in Selective Autophagy. *Molecular Cell* **34**, 259-269 (2009).
21. Nakatogawa, H., Suzuki, K., Kamada, Y. & Ohsumi, Y. Dynamics and diversity in autophagy mechanisms: lessons from yeast. *Nat Rev Mol Cell Biol* **10**, 458-467 (2009).
22. He, C. & Klionsky, D.J. Regulation Mechanisms and Signaling Pathways of Autophagy. *Annu Rev Genet* **43**, 67-93 (2009).
23. Powers, E.T., Morimoto, R.I., Dillin, A., Kelly, J.W. & Balch, W.E. Biological and Chemical Approaches to Diseases of Proteostasis Deficiency. *Annual Review of Biochemistry* **78**, 959-991 (2009).
24. Chiti, F. & Dobson, C.M. Protein Misfolding, Functional Amyloid, and Human Disease. *Annual Review of Biochemistry* **75**, 333-366 (2006).
25. Ross, C.A. & Poirier, M.A. Protein aggregation and neurodegenerative disease. *Nature Medicine* **10**, S10-S17 (2004).
26. Abbas, N. *et al.* A Wide Variety of Mutations in the Parkin Gene Are Responsible for Autosomal Recessive Parkinsonism in Europe. *Human Molecular Genetics* **8**, 567 - 574 (1999).
27. Olzmann, J.A. *et al.* Parkin-mediated K63-linked polyubiquitination targets misfolded DJ-1 to aggresomes via binding to HDAC6. *The Journal of Cell Biology* **178**, 1025 - 1038 (2007).
28. Chin, L.-S., Olzmann, J.A. & Li, L. Parkin-mediated ubiquitin signalling in aggresome formation and autophagy. *Biochem Soc Trans* **38**, 144-149 (2010).
29. Drummond, D.A. & Wilke, C.O. Mistranslation-Induced Protein Misfolding as a Dominant Constraint on Coding-Sequence Evolution. *Cell* **134**, 341-352 (2008).
30. Pierre, P. Dendritic cells, DRiPs, and DALIS in the control of antigen processing. *Immunological Reviews* **207**, 184-190 (2005).
31. Ben-Zvi, A.P. & Goloubinoff, P. Review: mechanisms of disaggregation and refolding of stable protein aggregates by molecular chaperones. *J. Struct. Biol.* **135**, 84-93 (2001).
32. Stadtman, E.R. & Levine, R.L. Protein Oxidation. *Annals of the New York Academy of Sciences* **899**, 191-208 (2000).
33. Nystrom, T. Role of oxidative carbonylation in protein quality control and senescence. *EMBO J* **24**, 1311-1317 (2005).
34. Morley, J.F., Brignull, H.R., Weyers, J.J. & Morimoto, R.I. The threshold for polyglutamine-expansion protein aggregation and cellular toxicity is dynamic and influenced by aging in Caenorhabditis elegans. *Proceedings of the National Academy of Sciences* **99**, 10417 -10422 (2002).
35. Wang, J. *et al.* Progressive aggregation despite chaperone associations of a mutant SOD1-YFP in transgenic mice that develop ALS. *Proceedings of the National Academy of Sciences* **106**, 1392 -1397 (2009).
36. Münch, C. & Bertolotti, A. Exposure of Hydrophobic Surfaces Initiates Aggregation of Diverse ALS-Causing Superoxide Dismutase-1 Mutants. *Journal of Molecular Biology* **399**, 512-525 (2010).

37. Erjavec, N., Larsson, L., Grantham, J. & Nyström, T. Accelerated aging and failure to segregate damaged proteins in Sir2 mutants can be suppressed by overproducing the protein aggregation-remodeling factor Hsp104p. *Genes & Development* **21**, 2410 -2421 (2007).
38. Ben-Zvi, A., Miller, E.A. & Morimoto, R.I. Collapse of proteostasis represents an early molecular event in Caenorhabditis elegans aging. *Proceedings of the National Academy of Sciences* **106**, 14914 -14919 (2009).
39. Fändrich, M. On the structural definition of amyloid fibrils and other polypeptide aggregates. *Cellular and Molecular Life Sciences* **64**, 2066-2078 (2007).
40. Maji, S.K., Wang, L., Greenwald, J. & Riek, R. Structure–activity relationship of amyloid fibrils. *FEBS Letters* **583**, 2610-2617 (2009).
41. Nelson, R. & Eisenberg, D. Recent atomic models of amyloid fibril structure. *Current Opinion in Structural Biology* **16**, 260-265 (2006).
42. Kitada, T. *et al.* Mutations in the parkin gene cause autosomal recessive juvenile parkinsonism. *Nature* **392**, 605-608 (1998).
43. Arrasate, M., Mitra, S., Schweitzer, E.S., Segal, M.R. & Finkbeiner, S. Inclusion body formation reduces levels of mutant huntingtin and the risk of neuronal death. *Nature* **431**, 805-810 (2004).
44. Tanaka, M. *et al.* Aggresomes Formed by α-Synuclein and Synphilin-1 Are Cytoprotective. *Journal of Biological Chemistry* **279**, 4625 -4631 (2004).
45. Douglas, P.M. *et al.* Chaperone-dependent amyloid assembly protects cells from prion toxicity. *Proceedings of the National Academy of Sciences* **105**, 7206 -7211 (2008).
46. Cohen, E., Bieschke, J., Perciavalle, R.M., Kelly, J.W. & Dillin, A. Opposing Activities Protect Against Age-Onset Proteotoxicity. *Science* **313**, 1604 -1610 (2006).
47. Chesebro, B. *et al.* Anchorless Prion Protein Results in Infectious Amyloid Disease Without Clinical Scrapie. *Science* **308**, 1435 -1439 (2005).
48. Saudou, F., Finkbeiner, S., Devys, D. & Greenberg, M.E. Huntingtin Acts in the Nucleus to Induce Apoptosis but Death Does Not Correlate with the Formation of Intranuclear Inclusions. *Cell* **95**, 55-66 (1998).
49. Taylor, J.P. *et al.* Aggresomes protect cells by enhancing the degradation of toxic polyglutamine-containing protein. *Human Molecular Genetics* **12**, 749 -757 (2003).
50. Christian Wigley, W. *et al.* Dynamic Association of Proteasomal Machinery with the Centrosome. *The Journal of Cell Biology* **145**, 481 -490 (1999).
51. Kaganovich, D., Kopito, R. & Frydman, J. Misfolded proteins partition between two distinct quality control compartments. *Nature* **454**, 1088-1095 (2008).
52. Coughlan, C.M. & Brodsky, J.L. Use of Yeast as a Model System to Investigate Protein Conformational Diseases. *Molecular Biotechnology* **30**, 171-180 (2005).
53. Parsell, D.A., Kowal, A.S., Singer, M.A. & Lindquist, S. Protein disaggregation mediated by heat-shock protein Hspl04. *Nature* **372**, 475-478 (1994).
54. Tyedmers, J. *et al.* Prion induction involves an ancient system for the sequestration of aggregated proteins and heritable changes in prion fragmentation. *Proceedings of the National Academy of Sciences* **107**, 8633 -8638 (2010).
55. García-Mata, R., Zsuzsa Bebök, Sorscher, E.J. & Sztul, E.S. Characterization and Dynamics of Aggresome Formation by a Cytosolic Gfp-Chimera□. *The Journal of Cell Biology* **146**, 1239 -1254 (1999).

56. Johnston, J.A., Ward, C.L. & Kopito, R.R. Aggresomes: A Cellular Response to Misfolded Proteins. *The Journal of Cell Biology* **143**, 1883 -1898 (1998).
57. Kopito, R.R. Aggresomes, inclusion bodies and protein aggregation. *Trends Cell Biol.* **10**, 524-530 (2000).
58. Garcia□Mata, R., Gao, Y. & Sztul, E. Hassles with Taking Out the Garbage: Aggravating Aggresomes. *Traffic* **3**, 388-396 (2002).
59. Weibezahn, J. *et al.* Thermotolerance Requires Refolding of Aggregated Proteins by Substrate Translocation through the Central Pore of ClpB. *Cell* **119**, 653-665 (2004).
60. Franzmann, T.M., Menhorn, P., Walter, S. & Buchner, J. Activation of the Chaperone Hsp26 Is Controlled by the Rearrangement of Its Thermosensor Domain. *Molecular Cell* **29**, 207-216 (2008).
61. Haslbeck, M., Franzmann, T., Weinfurtner, D. & Buchner, J. Some like it hot: the structure and function of small heat-shock proteins. *Nat Struct Mol Biol* **12**, 842-846 (2005).
62. Dougan, D.A., Reid, B.G., Horwich, A.L. & Bukau, B. ClpS, a Substrate Modulator of the ClpAP Machine. *Molecular Cell* **9**, 673-683 (2002).
63. Mosser, D.D., Ho, S. & Glover, J.R. Saccharomyces cerevisiae Hsp104 Enhances the Chaperone Capacity of Human Cells and Inhibits Heat Stress-Induced Proapoptotic Signaling†. *Biochemistry* **43**, 8107-8115 (2004).
64. Yamamoto, A., Lucas, J.J. & Hen, R. Reversal of Neuropathology and Motor Dysfunction in a Conditional Model of Huntington's Disease. *Cell* **101**, 57-66 (2000).
65. Yeung, H.O. *et al.* Insights into adaptor binding to the AAA protein p97. *Biochem. Soc. Trans.* **36**, 62-67 (2008).
66. Bedford, L. *et al.* Depletion of 26S Proteasomes in Mouse Brain Neurons Causes Neurodegeneration and Lewy-Like Inclusions Resembling Human Pale Bodies. *The Journal of Neuroscience* **28**, 8189 -8198 (2008).
67. Lindner, A.B., Madden, R., Demarez, A., Stewart, E.J. & Taddei, F. Asymmetric segregation of protein aggregates is associated with cellular aging and rejuvenation. *Proceedings of the National Academy of Sciences* **105**, 3076 -3081 (2008).
68. Henderson, K.A. & Gottschling, D.E. A mother's sacrifice: what is she keeping for herself? *Current Opinion in Cell Biology* **20**, 723-728 (2008).
69. Aguilaniu, H., Gustafsson, L., Rigoulet, M. & Nyström, T. Asymmetric Inheritance of Oxidatively Damaged Proteins During Cytokinesis. *Science* **299**, 1751 -1753 (2003).
70. Rujano, M.A. *et al.* Polarised Asymmetric Inheritance of Accumulated Protein Damage in Higher Eukaryotes. *PLoS Biol* **4**, e417 (2006).
71. Singhvi, A. & Garriga, G. Asymmetric divisions, aggresomes and apoptosis. *Trends in Cell Biology* **19**, 1-7 (2009).
72. Weibel, E.R., Taylor, C.R. & Bolis, L. *Principles of Animal Design: The Optimization and Symmorphosis Debate*. (Cambridge University Press: 1998).
73. Tartaglia, G.G., Pechmann, S., Dobson, C.M. & Vendruscolo, M. A relationship between mRNA expression levels and protein solubility in E. coli. *J. Mol. Biol.* **388**, 381-389 (2009).
74. Tartaglia, G.G. & Vendruscolo, M. Correlation between mRNA expression levels and protein aggregation propensities in subcellular localisations. *Mol Biosyst* **5**, 1873-1876 (2009).
75. Torres, E.M. *et al.* Effects of Aneuploidy on Cellular Physiology and Cell Division in Haploid Yeast. *Science* **317**, 916 -924 (2007).

76. Kellis, M., Birren, B.W. & Lander, E.S. Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. *Nature* **428**, 617-624 (2004).
77. Conant, G.C. & Wolfe, K.H. Increased glycolytic flux as an outcome of whole-genome duplication in yeast. *Molecular Systems Biology* **3**, (2007).
78. Qian, W., He, X., Chan, E., Xu, H. & Zhang, J. Measuring the evolutionary rate of protein–protein interaction. *Proceedings of the National Academy of Sciences* **108**, 8725 -8730 (2011).
79. Fay, J.C. & Benavides, J.A. Evidence for Domesticated and Wild Populations of Saccharomyces cerevisiae. *PLoS Genet* **1**, e5 (2005).
80. Lipson, D., Ben-dor, A., Dehan, E. & Yakhini, Z. Joint Analysis of DNA Copy Numbers and Gene Expression Levels. *PROCEEDINGS OF ALGORITHMS IN BIOINFORMATICS: 4TH INTERNATIONAL WORKSHOP, WABI 2004* **3240**,
81. Pollack, J.R. *et al.* Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proceedings of the National Academy of Sciences* **99**, 12963 -12968 (2002).
82. Laurent, J.M. *et al.* Protein abundances are more conserved than mRNA abundances across diverse taxa. *Proteomics* **10**, 4209-4212 (2010).
83. Lehmann, D.J. *et al.* Replication of the association of HLA-B7 with Alzheimer's disease: a role for homozygosity? *J Neuroinflammation* **3**, 33-33