Evolutionary Systems Biology in Yeast

By Guang-Zhong Wang

A thesis submitted to the Heinrich-Heine-Universität Düsseldorf
for the degree of *Doctor rerum naturalium* (Dr. rer. nat.)

Supervised by Professor Martin J. Lercher

Institute of Computer Science
Heinrich-Heine-University Düsseldorf
June 2010

Aus dem Institut für Informatik
der Heinrich-Heine-Universität Düsseldorf

Gedruckt mit der Genehmigung der
Mathematisch-Naturwissenschaftlichen Fakultät der
Heinrich-Heine-Universität Düsseldorf

Referent: Prof. Dr. Martin Lercher
Korrferent: Prof. Dr. William Martin

Tag der mündlichen Prüfung: August 25, 2010

Declaration

This thesis is submitted for the degree of Doctor rerum naturalium at the Heinrich-Heine-University Düsseldorf. It has not been submitted to any other university for a degree. I agree that the University library may lend out or copy this thesis freely.

Guang-Zhong Wang.
June, 2010.

Acknowledgements

Many thanks to everybody in the Martin Lercher lab, past and present, for the creative scientific environment they provided: Gabriel Gelius-Dietrich, Wei-Hua Chen, Sabine Thuß, Thomas Laubach, Na Gao, Christian Eßer, Wolfgang Kaisers, Jan Wolfertz, Milan Majtanik and Janina Maß.

A special thank you to the system administrators in the lab during my PhD, Jochen Kohl and Lutz Voigt, and to the secretary Anja Walge. Without their daily help, it would have been difficult to keep going.

Many thanks to Professor Laurence Hurst and his group members: Claudia Weber, Tobias Warnecke, Catherine Pink, and Lu Chen, for their helpful discussions during my visit to the University of Bath. I had a great time in Bath and learned so much in 4 months.

Many thanks to all of the project students, visitors and thesis students in the Martin Lercher lab, and to the people in Bill Martin's group for the joint seminars, as well as anyone else that I may have forgotten.

Many thanks to Martin Lercher for his kind and supportive supervision during the development of each project.

Finally, I owe my gratitude to my parents, to my sister and to my girlfriend for their love and constant support.

Abstract

Since the release of the yeast whole genome sequence in 1996 and the advances in high-throughput technologies, scientists have studied the evolution of the yeast *Saccharomyces cerevis iae* at the genomics level in two major ways: comparative genomics and systems biology. Both of these approaches have provided important insights into how the yeast genome is organized and how networks evolve to achieve phenotypic features. In this thesis, I study these issues by analyzing both genomics and networks data.

The major questions I asked and the findings raised by this thesis are as follows: First, I studied why there are so many yeast non-coding transcripts (>60%) that are transcribed in bi-directional orientation with a coding gene. For some genes, notably essential genes, expression when expression is needed is vital, hence low noise in expression is favorable. For other genes, noise is necessary for coping with environmental stochasticity or for providing dice-like mechanisms to control cell fate. But how is noise in gene expression modulated? We hypothesize that gene orientation may be crucial, as for divergently organized gene pairs, expression of one gene could affect the chromatin conformation of a neighbouring gene, thereby reducing noise for that gene. Transcription of antisense non-coding RNA from a shared promoter is similarly argued to be a noise-reduction mechanism. Our stochastic simulation models confirm the expectation. The model correctly predicts: that protein coding genes with bi-promoter architecture, including those with a ncRNA partner, have lower noise than other genes; divergent gene pairs uniquely have correlated expression noise; distance between promoters predicts noise; ncRNA divergent transcripts are associated with genes that *a priori* would be under selection for low noise; essential genes reside in divergent orientation more than expected; bi-promoter pairs are rare subtelomerically, cluster together and are enriched in essential gene clusters. We conclude that gene orientation and transcription of ncRNAs, even if unstable, are candidate modulators of noise levels.

Second, I studied whether ancestrally neighbouring genes still remain

co-expressed even after they have been separated by chromosomal rearrangements. Although there is clear evidence that closely spaced gene pairs tend to be highly co-expressed, it is not clear if this co-expression is solely due to a mechanistic neighbourhood effect, or if the co-expression is selectively favorable. Thanks to the multiple fungi genome sequencing projects, it is now possible to answer this question. Using a reconstruction of gene order in an ancestral yeast based on parsimony, we found a significant co-expression signal for many separated gene pairs. Moreover, even genes that are neighbouring in other fungi but were never genomic neighbours in the evolutionary history of *Saccharomyces cerevisiae* show higher co-expression than expected. We conclude that co-expression of neighbouring genes is indeed often favoured by natural selection.

Third, I studied how network neighbours influence the evolutionary rate of a protein and why. Recently it was shown that the level of protein expression is the main predictor of the protein evolution rate. Thus, if two genes have similar expression levels, they should also have a similar rate of evolution. We found that this can explain the fact that neighbouring gene evolve similarly in most biological networks, regardless of the different network topologies. Namely, controlling for expression level, neighbouring genes no longer show correlated evolution in almost all networks studied. But in co-expression network, even controlling for expression abundance as well as for gene essentiality and gene length, neighbouring (*i.e.*, co-expressed) genes still co-evolve. This finding suggests that both expression level and co-expression influence the rate of protein evolution in networks.

Finally, I focused on the phenotypic effect of genetic hubs. Robustness is a basic feature of biological networks, and we expect different proteins to make different contributions to the overall robustness of the network. In genetic interaction networks, when the genetic hubs function abnormally, the offspring is expected to exhibit more phenotypic variation (both genetically caused and non-genetically caused). We observed that the number of strong negative genetic interactions (synthetic lethality) is indeed positively correlated with phenotypic variation of the respective single-gene knockouts in yeast.

Furthermore, there is a high correlation between haploid fitness of the knockouts and phenotypic variation. Thus, haploid fitness and genetic interactions are two predictors of phenotypic variation in mutants. This further suggests that the release of phenotypic variation in mutants is mostly not due to a specific buffering function of the mutated gene (as in the case of chaperones, *e.g.,* Hsp90), but that compromised function of one part of the network reduces the cell's ability to compensate for sub-optimal pathways elsewhere.

Zusammenfassung

Die Evolution der Hefe *Saccharomyces cerevisiae* wurde seit der Veröffentlichung der kompletten Genomsequenz 1996 und den Fortschritten in Hochdurchsatz-Technologien auf Genomebene hauptsächlich auf zwei Arten untersucht: durch vergleichende Genomik und durch Systembiologie. Beide Ansätze haben wichtige Einsichten in die Genomorganisation und in die Evolution von Netzwerken geliefert. Diese Forschungsthemen bilden auch das Thema der vorliegenden Doktorarbeit. Hierzu untersuche ich sowohl Netzwerk- als auch genomische Daten.

Die wichtigsten behandelten Fragen sind die folgenden: Zunächst untersuche ich die Frage, warum das Hefegenom so viele nicht-kodierende Transkripte enthält, die bidirektional mit einem protein-kodierenden Gen abgelesen werden (>60%). Für einige Gene, insbesondere für essentielle Gene, ist es lebenswichtig, dass Genexpression genau dann stattfindet, wenn sie benötigt wird; für diese Gene bietet ein niedriges ‚Expressionsrauschen' (d.h. Variabilität in Genexpression) einen Selektionsvorteil. Andere Gene benötigen Expressionsrauschen, um mit stochastischen Umgebungen umzugehen oder um Mechanismen für die zelluläre Regulation zu liefern, die nach dem Muster von Würfelspielen funktionieren. Aber wie wird Expressionsrauschen vom Genom gesteuert? Wir stellen die Hypothese auf, dass Genorientierung hier eine entscheidende Rolle spielen könnte: bei divergent transkribierten Genpaaren könnte die Expression eines Gens die Chromatin-Konformation des benachbarten Gens beeinflussen, wodurch das Expressionsrauschen des zweiten Gens vermindert würde. Transkription von nicht-kodierender Antisense-DNS ausgehend von einem gemeinsamen Promoter könnte analog zu einer Reduzierung des Expressionsrauschens führen. Unsere stochastischen Simulationen bestätigen diese Erwartungen. Unser Modell sagt folgende Beobachtungen richtig voraus: Protein-kodierende Gene mit bi-promoter Architektur, einschließlich solcher mit einem nicht-kodierenden Partner, haben ein niedrigeres Expressionsrauschen als andere Gene; divergente Genpaare zeichnen sich durch korreliertes Expressionsrauschen aus; der Abstand

zwischen Promotoren sagt das Ausmaß an Expressionsrauschen voraus; nicht-kodierende RNS involviert in divergente Transkription haben Partnergene, von denen *a priori* erwartet wird, dass niedriges Expressionsrauschen einen Selektionsvorteil bietet; essentielle Gene befinden sich öfter in divergenter Orientierung als zufällig erwartet; bi-promoter Genpaare treten selten in sub-telomeren Regionen auf, bilden Gruppen und treten häufig in essentiellen Gengruppen auf. Wir schließen daraus, dass Genorientierung und Transkription von nicht-kodierenden RNS, selbst wenn diese instabil sind, mögliche Regelungssysteme für das Ausmaß von Expressionsrauschen sind.

Zweitens habe ich untersucht, ob Gene, die in einem Vorläufergenom Nachbarn waren, immer noch co-exprimiert sind selbst wenn sie durch genomische Umordnungen auseinandergerissen wurden. Es steht außer Frage, dass eng benachbarte Genpaare häufig stark co-exprimiert sind; es ist allerdings unklar, ob dies lediglich durch rein mechanische Nachbarschafts-Effekte verursacht wird, oder ob co-Exprimierung einen Selektionsvorteil bietet. Diese Frage kann nun dank mehrerer Pilz-Genomprojekte beantwortet werden. Mit Hilfe einer Rekonstruktion der Genanordnung in einer ancestralen Hefespezies konnten wir zeigen, dass viele Genpaare auch nach einer genomischen Umordnung noch signifikant co-exprimiert sind. Wir fanden weiterhin, dass sogar benachbarte Genpaare in anderen Pilzarten – die in der Evolutionsgeschichte der Hefe nie benachbart waren – stärker co-exprimiert sind als erwartet. Wir schließen daraus, dass co-Expression benachbarter Gene tatsächlich häufig im Einklang mit natürlicher Selektion ist.

Drittens habe ich untersucht wie und warum Nachbarn in biologischen Netzwerken die Evolutionsrate von Proteinen beeinflussen. Es wurde kürzlich gezeigt, dass die Expressionshöhe diejenige Variable ist, die die beste Vorhersage der Evolutionsrate ermöglicht. Zwei Gene mit ähnlicher Expressionsrate sollten daher auch ähnliche Evolutionsraten zeigen. Wir zeigen hier, dass dieser Zusammenhang erklären kann, warum benachbarte Gene in den meisten biologischen Netzwerken ähnliche Evolutionsraten haben: wenn wir Genexpression herausrechnen, dann verschwindet die Ähnlichkeit der Evolutionsraten in fast allen Netzwerken. Eine Ausnahme bildet das

co-Expressions-Netzwerk, in dem benachbarte Gene immer noch ähnliche Evolutionsraten haben nachdem wir die Effekte von Genexpression, Gen-Essentialität und Genlänge herausrechnen. Diese Beobachtung legt nahe, dass sowohl Expressionshöhe als auch co-Expression die Evolutionsrate in Netzwerken beeinflussen.

Schließlich habe ich mich auf den phenotypischen Einfluß von genetischen Netzwerk-Achsen (d.h. Genen mit vielen genetischen Interaktionen) konzentriert. Robustheit ist eine grundlegende Eigenschaft biologischer Netzwerke, und wir erwarten, dass verschiedene Proteine unterschiedlich zur Robustheit des Netzwerkes beitragen. Wenn Achsen in genetischen Interaktionsnetzwerken ihre normale Funktion verlieren, dann erwartet man höhere phänotypische Variabilität in den Nachkommen. Wir beobachten dass die Anzahl stark negativer genetischer Interaktionen (synthetische Letalität) tatsächlich mit der phänotypischen Variabilität der entsprechenden Gen-Knockouts korreliert. Weiterhin finden wir eine starke Korrelation zwischen der haploiden Fitness der Knockouts und phänotypischer Variabilität. Einerseits tragen also sowohl genetische Interaktionen als auch haploide Knockout-Fitness zur Vorhersage phänotypischer Variabilität bei. Daraus folgt weiterhin, dass die Freisetzung phänotypischer Variabilität häufig nicht die Konsequenz einer spezifischen Pufferfunktion des mutierten Gens ist (so wie bei Chaperonen, z.B. Hsp90), sondern dass eine beeinträchtigte Funktion in einem Teil des Netzwerkes die Fähigkeit der Zelle reduziert, sub-obtimale Reaktionspfade in anderen Teilen des Netzwerkes auszugleichen.

Abbreviations

CUTs	cryptic unstable transcripts
FOP	optimal codon usage
ncRNA	non-coding RNA
NFR	nucleosome free region
SUTs	stable annotated transcripts
TF	transcription factor
TSS	transcription start site
PPI	protein-protein interaction
YGOB	the Yeast Gene Order Browser
WGD	whole-genome duplication
ORF	open reading frame
S.cerevisiae	*Saccharomyces cerevisiae*
TR	transcription regulation
SLI	Synthetic lethal interaction
MetI	metabolic interaction
CoEx	co-expression interaction
dS	synonymous substitutions per synonymous site
dN	non-synonymous substitutions per non-synonymous site

Table of contents

Chapter 1 Introduction

In this chapter, I will discuss the genetics and life cycle of budding yeast, *Saccharomyces cerevisiae*, which is the main subject of the thesis. After that, I will discuss the general genome evolution of *S. cerevisiae*. Finally, I will discuss the evolution of yeast on the network and phenotype levels.

1.1 The genetics of budding yeast

1.1.1 *Saccharomyces cerevisiae*: background

Yeast is one of the oldest domesticated organism in human history: it has been used, *e.g.,* in beer brewing since around 6000 B.C.. In the 1960s, budding yeast was introduced as a model organism for molecular biology. Today it is widely used in food technologies, fermentation industries, environmental technologies and healthcare industries. Beyond this, it plays very important roles in biological and biochemical research for the following reasons: it can easily be cultured and grows rapidly; it can be transformed with genes from other genomes (Bonnefoy, Remacle, Fox 2007); and its gene content is known, as the complete genome was sequenced in 1996 (Goffeau *et al*. 1996).

Figure 1-1 represents the phylogenetic position of *S. cerevisiae* among yeast species. A major division in yeast evolutionary history occurred when a common ancestor experienced a whole genome duplication, which happened 100–150 million years ago (Sugino, Innan 2005), shortly after the divergence of the *S. cerevisiae* lineage from *Kluyveromyces lactis*. This duplication event apparently helped the species in the *Saccharomyces* complex branch (tagged sensu stricto in Figure 1-1) to gain the ability for rapid growth in an anaerobic environment and to preferentially ferment glucose by generating alcohol in oxygen-rich environments.

1.1.2 The life cycle and mating types of *S. cerevisiae*

There are two forms in which yeast cells can survive and grow: haploid and diploid (Figure 1-2). The diploid cell is the preferential form of yeast; in this

form yeast cells can undergo meiosis. There are two mating types in the haploid cell: a and α. The mating of a and α results in an a/α diploid cell, which can live in harsher environments. The diploid cells are unable to mate but in stress conditions will divide by meiosis into four haploid yeast cells: two a and two α.

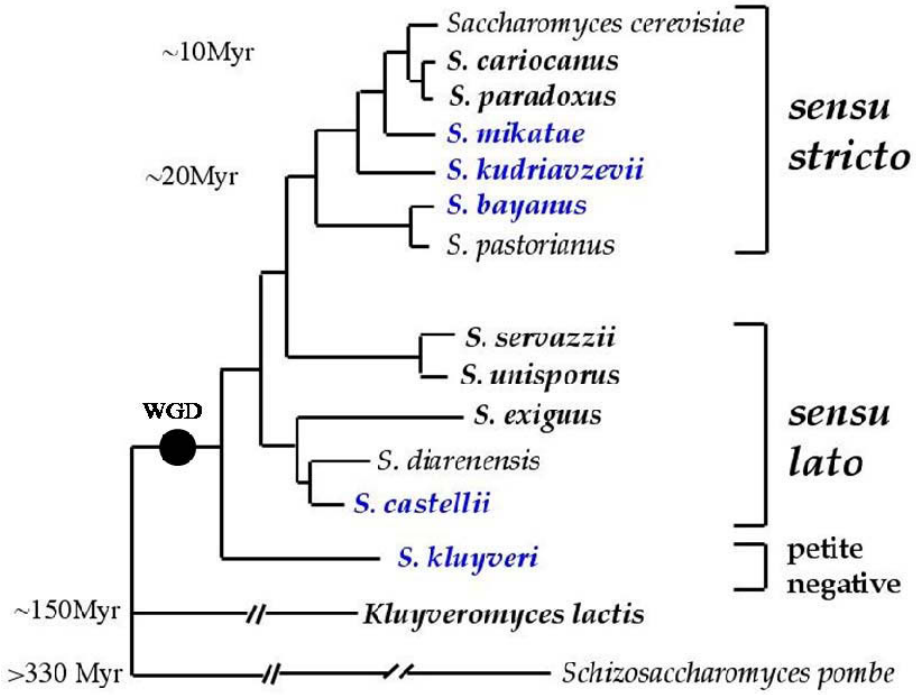


Figure 1-1 Phylogeny of yeast species with whole genome duplication event (WGD) which happened 100~150 Myr ago noted by the black cycle[1].

The mating type of *S. cerevisiae*, which is determined by a single locus MAT, only has consequences in haploid cells. In haploid yeast cells, the a cells produce an "a-factor", which prompts the a-cell to approach an α-cell. Conversely, α-cells attract a-cells by producing an αfactor.

The locus MAT in a and α cells encodes different proteins. In the a cell it is called MATa, and in the α cell it is called MATα. The MATa allele encodes a1, which activates the a-specific transcriptional program, and the MATα allele encodes the α1 and α2 genes, which activate the α-specific transcriptional program and make the a-cell into a α-cell.

The mating type of the yeast cell can be switched, which is an advantage to

[1] http://www.genetics.wustl.edu/saccharomycesgenomes/yeast_phylogeny.html

yeast population that could lead of both a-cell and α-cell represent in the population (Houston, Simon, Broach 2004).

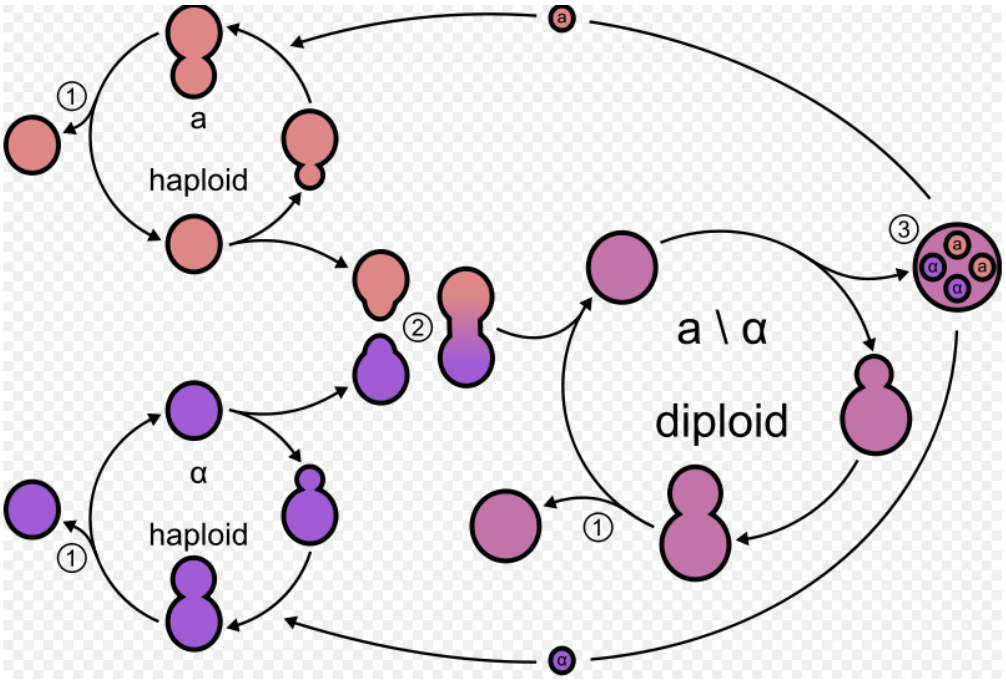


Figure 1-2 Life cycle of budding yeast which comes from Wikipedia[2]. Major events in the life cycle: 1. Budding; 2. Conjugation; 3. Spore.

1.2 Yeast genome evolution

1.2.1 The mechanisms of gene duplication

Gene duplication refers to the duplication of a large DNA fragment containing a gene. The duplicated copy is often regarded as being initially free from selection. However, there are several different fates for newly duplicated gene pairs: the additional copy may become lost (which is most often the case), maybe via first becoming a pseudogene; one copy may gain a new function by neo-functionalization; or both copies could share the original function by subfunctionalization (Zhang 2003).

There are at least four ways in which a duplicated copy of a gene may arise: homologous recombination; retrotransposition; segmental duplication; and

[2] http://en.wikipedia.org/wiki/Yeast

whole genome duplication.

Homologous recombination (often synonymous with recombination) refers to the exchange of highly similar DNA sections with a sister chromosome. This process is not only useful for repairing potentially deleterious double-strand breaks in DNA, but is also essential in producing new genetic combinations during meiosis. Moreover, homologous recombination is a way to facilitate horizontal gene transfer between bacteria and viruses (Fall *et al*. 2007).

A retrotransposon is a type of transposon that is particularly widespread in plants. Retrotransposons first transcribe one gene into an mRNA, then the mRNA is retrotranscribed to complementary DNA (cDNA) by transposase and then inserted into the chromosome somewhere else. The two mechanisms generate very different duplicated genes: homologous recombination usually produces tandem duplicates with the introns and regulatory sequences of the genes retained, while retrotransposition usually produces gene copies that are located far away from the original copy and do not contain introns or regulatory sequences (Zhang 2003). Therefore retrotransposition often generates pseudogenes.

Segmental duplication gives rise to low copy repeats. Although the mechanism is not very clear, segmental duplication is reported to be an important evolutionary force in primate evolution (Cheung *et al*. 2003; Bailey, Eichler 2006). However, its role in yeast evolution is likely to be more limited.

Finally, whole genome duplication is a very different form of generating duplicated genes. In whole genome duplications, the genes on each chromosome are duplicated. Many of the redundant copies are subsequently lost. Although initially whole genome duplication was regarded as very rare in nature, whole genome duplications have now been discovered in plants (Maere *et al*. 2005; Tuskan *et al*. 2006; Jaillon *et al*. 2007; Paterson *et al*. 2010), fishes (Brunet *et al*. 2006), yeasts (Wolfe, Shields 1997) and basal vertebrates (Blomme *et al*. 2006).

1.2.2 Yeast gene order browser (YGOB)

With the rapid development in sequencing technology, hundreds of genomes have been sequenced to date, including more than 20 fungal species. This data has facilitated the comparison of genome order, which is of great importance for studying the evolutionary history of yeasts. Based on the current gene arrangement in the genomes, it is also possible to reconstruct the ancestral gene orders. The yeast gene order browser (YOGB)[3], which currently contains information on the genome arrangement of 11 yeasts, is an extremely useful tool in this respect (Byrne, Wolfe 2005). The information in this database is manually curated. More than 550 gene pairs that were retained after the ancestral whole genome duplication have been identified using this browser (Byrne, Wolfe 2005).

A comparison of the gene order of the three yeast species *Saccharomyces cerevisiae*, *Saccharomyces castellii* and *Candida glabrata* revealed that the gene loss process differs at different loci. There is about a 20% loci difference in these three species (Scannell *et al*. 2006). It was also found that the three lineages diverged shortly after the whole genome duplication. This lead the authors to propose a simple model based on passive complementary gene losses to explain the rapid speciation after whole genome duplication (WGD). Additionally, YGOB can also be used to assign the content of the ancestral genome before speciation. Aligning the gene order of *Kluyveromyces polysporus* to the other post-WGD yeasts showed that the common ancestor contained >9,000 genes (Scannell *et al*. 2007).

The ancestral yeast gene order before the whole genome duplication event was constructed based on the integration of the gene order of 11 yeasts. This ancestor contained 8 chromosomes and 4703 loci in total. Comparison of the ancestral genome information with *S. cerevisiae* indicates that 124 genes were gained during whole genome duplication and 88 loci were lost in *S. cerevisiae*. The function of gained genes is biased towards ethanol production, growth in hypoxic environments, or making use of alternative nutrient sources (Gordon,

[3] http://wolfe.gen.tcd.ie/ygob/

Byrne, Wolfe 2009).

1.2.3 Whole genome duplication (WGD) in yeast

Gene duplication is an important evolutionary force that was recognized by geneticists in the 1970s, in particular including Susumu Ohno. In his famous book "Evolution by Gene Duplication", he also proposed that polyploidy (whole genome duplication) played a particularly important role in evolutionary history (Ohno 1970). Not until the completion of sequencing of the first eukaryote genome (Goffeau *et al*. 1996) was it possible to test his hypothesis. In 1997, Wolfe and his colleagues aligned the chromosomes of yeast by dot matrices (Wolfe, Shields 1997). They found that there are 55 large-scale duplicated regions, which in sum contain 376 homologous gene pairs. This was consistent with Ohno's hypothesis. Thus they proposed that *Saccharomyces* experienced a whole genome duplication during its recent evolution (Wolfe, Shields 1997).

Although the large duplicated chromosomal segments provided support for whole genome duplication, there was a debate over whether these segments were created by WGD or by many independent segment duplications (Friedman, Hughes 2001; Piskur 2001). The controversy was not resolved until the publication of a genome-wide sequence of a pre-WGD species -- *K. waltii* (Kellis, Birren, Lander 2004). *K. waltii* contains 8 chromosomes, which is exactly half the number of chromosomes in *Saccharomyces cerevisiae*. Further analysis of the genome of *K. waltii* shows that *K. waltii* has a 1:2 mapping of the genome with *Saccharomyces cerevisiae*, namely, each chromosome region of *K. waltii* corresponds to two regions of *Saccharomyces cerevisiae* (Kellis, Birren, Lander 2004).

1.2.4 Models for the retention of duplicated genes

Two main models have been proposed to explain the process by which a newly created duplicated copy is preserved, namely subfunctionalization and neofunctionalization. Both are supported by a wide range of literature. Subfunctionalization proposes that after duplication, the two genes suffer complementary mutations in their regulatory sequences, resulting in each of the

copies being able to perform a subset of the original function of the ancestral gene. Thus there is no adaptive evolution in the duplicated genes (Force *et al*. 1999), although subfunctionalization may typically be followed by an increased specialisation of the proteins. Tissue-specific expression of duplicated genes and reciprocal degenerative coding-region changes are usually cited as examples of separate function (Force *et al*. 1999).

Conversely, neofunctionalization proposes that one of the two duplicated loci retains the ancestral function, while the other copy gains a new function; this is usually accompagnied by an increased rate of evolution. After neofunctionalisation, it is presumed that the second locus confers a selective advantage.

In addition, other models have been suggested to explain the preservation of duplicated genes. One model assumes that selection for increased dosage may be responsible for the preservation of the duplicated gene (Seoighe, Wolfe 1999). In another model, both of the copies could perform the required function of the ancestral protein but a decrease in gene expression level or coding region impairment could lead to the retention of the two copies. This situation is likely to exist in particularly small populations (Lynch, Force 2000). Finally, deleterious effects of dosage imbalance, *e.g.*, when only some subunits of a duplicated protein complex get lost, may hinder the loss of sets of interacting proteins (Papp, Pal, Hurst 2003).

1.2.5 The consequences of gene and genome duplication

The rate of duplication events in the genome is estimated at 0.002 to 0.020 per gene per million years, which appears relatively high and underlines the importance of duplication events in the evolution of eukaryotes (Lynch, Conery 2000). Although both of the copies are perhaps created equally, most of the duplicated copies are lost fairly rapidly: on average they have a half-life of 1~10 million years (Lynch, Conery 2000). In addition, there is a rapid divergence of expression and regulatory network evolution after gene duplication (Gu, Zhang, Huang 2005). The preserved genes become stable in the genome, either by

gaining a new function (neofunctionalization) or by inheriting part of the function of the ancestral gene (subfunctionalization).

Since the duplication of genes is the main way of gaining novel functions in eukaryote evolution, it is critically important to study the consequences of gene duplications. Duplicated genes are enriched in several functional categories. In both *E. coli* and *S. cerevisiae*, duplicated genes tend to be transcription factors and about half of all transcription interactions were gained after duplication (Teichmann, Babu 2004). In protein interaction networks, duplication not only increases the number of proteins, but also of interactions (Bergman, Siegal 2003). What is more, duplicated genes may compensate the function of each other, hence increasing the genetic robustness of the system (Gu *et al*. 2003).

The consequences of small-scale duplications are very different to the consequences of whole genome duplications, as whole genome duplication provides each gene with an equal chance of evolving new functions. Stress-related genes tend to be duplicated in small-scale duplications, while growth-related genes do not (Wapinski *et al*. 2007); however, in whole genome duplications this constraint has been circumvented. Moreover, new modules rarely arise by gene duplication (Wapinski *et al*. 2007). A further important consequence of whole genome duplication in yeast (perhaps the most important consequence) is the increase in glycolytic flux, which provided the post-wgd yeast with the ability for rapid glucose fermentation, which in turn improved the adaptation to glucose-rich environments (Conant, Wolfe 2007).

1.2.6 Non-random gene order in yeasts

Although it is generally accepted that genes that form the same operon are usually from similar functional categories and thus not randomly distributed in prokaryotes (Suyama, Bork 2001), even a decade ago there was little evidence that gene order is not random in the eukaryote genome. But the situation has changed dramatically in the past ten years due to the effort of geneticists. First in yeast, (Cohen *et al*. 2000), then in human (Lercher, Urrutia, Hurst 2002; Semon, Duret 2006), *Caenorhabditis elegans* (Roy *et al*. 2002), *Mus musculus*

(Reymond *et al*. 2002; Semon, Duret 2006), *Rattus n orvegicus* (Fukuoka, Inaoka, Kohane 2004), *Arabidopsis thaliana* (Birnbaum *et al*. 2003; Williams, Bowles 2004; Chen, de Meaux, Lercher 2010), and *Plasmodium falciparum* (Florens *et al*. 2002), and now in almost every examined eukaryote genome, people have found evidence for non-random clustering of genes. Understanding the origin and the underlying selective reason for these clusters is of considerable importance to both genomics and gene therapy.

The first genome-wide evidence showing that gene order is not random in yeast was the paper by Cohen *et al*. (Cohen *et al*. 2000). Using whole genome-wide cell cycle mRNA expression data, they found that genes on the same chromosome show correlated expression patterns. Moreover, they demonstrated that adjacent genes also show correlated expression independent of gene orientation, while genes located close together on the same chromosome also tend to have similar functions. Shortly afterwards, Hurst and colleagues found evidence that the co-expression of neighbouring genes cannot fully be accounted for by their physical proximity to the promoter region, and thus a clear signal of selection on these clusters is seen (Hurst, Williams, Pal 2002). Later, through the analysis of a "cleaner" expression dataset that was obtained from a large Northern blot study, it was shown that the co-expression domains in yeast chromosomes could be as large as 100 kb (Lercher, Hurst 2006). Given the fact that gene order is not random in yeast, which factors are more important than others in determining the conserved gene order? A systematic examination based on a logistic regression model suggested that intergene distance is the strongest predictor, then the local density of essential genes and co-regulation, followed by co-expression and the recombination rate (Poyatos, Hurst 2007).

1.2.7 Different mechanisms are responsible for the co-expressed gene clusters

Different mechanisms have been proposed to be responsible for the non-randomness of gene order. The first one is the null hypothesis, holding that only intergenic distance plays a major role. As for recombination, gene pairs that are separated by a large distance are more likely to be recombined while

gene pairs that are close to each other should be more conserved. Closely linked genes are expected to be more conserved than gene pairs that are far away from each other. Actually many researchers have observed that intergenic distance is a strong predictor of neighbouring gene conservation (Poyatos, Hurst 2007; Tsai *et a l.* 2009). In addition, clusters of essential genes usually have low recombination rates and larger clusters are usually found in low recombination rate regions, indicating that these clusters are not only physical clusters but also genetic clusters (Pal, Hurst 2003).

Another mechanism that has been proposed is the sharing of transcription regulatory systems (Kruglyak, Tang 2000). Regardless of orientation, many gene pairs could potentially be regulated by just one regulatory motif, and thus the transcription of one gene will automatically activate another. Under this situation, the adjacent gene pairs show incredibly similar expression profiles. Kruglyak and Tang reported several such gene pairs, most of them being related to the cell cycle (Kruglyak, Tang 2000).

Furthermore, by comparing the expression profile of genes in more than 80 experiments, Spellman and Rubin reported that more than 200 groups of adjacent genes showed similar expression profiles, accounting for 20% of the total number of genes (Spellman, Rubin 2002). This phenomenon led the authors to propose that the regulation of chromatin structure plays an important role because it cannot be explained by either gene function or homology (Spellman, Rubin 2002). As chromatin is not static, transitions between active chromatin and inactive chromatin could be an efficient mechanism leading neighbouring genes to be highly co-expressed (Poyatos, Hurst 2007; Chen, de Meaux, Lercher 2010). In yeast, chromatin remodeling is thought to have a major role in the co-expression of neighbouring genes, especially for those for which the co-expression level is relatively low (Pearson correlation coefficient between 0.2 and 0.4) (Batada, Urrutia, Hurst 2007).

1.2.8 Noise measurement and single molecular technology

The transcription and translation processes of genes are built from fundamental

biochemical reactions, and thus they are stochastic. It is not long since people first noticed this point, although even in 1957, Novick and Weiner found that the expression of beta-galactosidase varies between cells (Novick, Weiner 1957). Recently, interest in this topic has increased, one important reason being the advances in using fluorescent proteins, *e.g.,* GFP. Combining fluorescent proteins and microscopy, flow cytometry and time-lapse imaging technology, different methods have been developed to detect both single mRNA and protein molecules. Table 1-1 is a summary of the advantages and disadvantages of different methods in this field, modified from a previously published review (Raj, van Oudenaarden 2009).

The technology described above has helped to advance our understanding of both the benefits and the harmful effects of noise. An excellent example where the cell utilizes noisy gene expression to achieve a specific function involves odorant receptors. There are over a thousand different odorant receptors in different sensory neurons, and it is believed that the mouse developed a simple "Monte Carlo" strategy to randomly express a special odorant receptor for each cell (Vassar, Ngai, Axel 1993). Another example is the stochastic expression of ComK, which is responsible for the competence transitions in *Bacillus subtilis*: noise reduction of this protein could also reduce the number of competent cells (Maamar, Raj, Dubnau 2007).

Despite these examples of advantageous noise, is likely that noisy gene expression is mostly harmful to organisms. This may be especially true in developmental systems, as the accurate implementation of a special developmental program at a certain time is crucial. However, the potentially harmful effects of noisy gene expression have not yet been fully studied experimentally. One well known example is the positive correlation between the noise of gene expression and the age of animals (Bahar *et al*. 2006; Somel *et al*. 2006).

1.2.9 Gene order evolved to reduce gene expression noise

Although several specific examples have been discovered regarding the benefit

of noisy gene expression, it was recently shown that the cell uses a genome-wide strategy to minimize overall expression noise in both prokaryotes and eukaryotes. In *E. coli*, the gene order is not random because of operon structures: genes encoded in the same operon often belong to the same metabolic pathway or the same protein complex (Suyama, Bork 2001). Recently, Kovacs *et al.* discovered that even within the same operon, the gene order is still not random: enzymes acting first in their metabolic pathway are often located close to the 5' terminus (Kovacs, Hurst, Papp 2009). These authors further constructed a theoretical model to show that this arrangement can be explained by the minimization of stochastic stalling of metabolism, especially for proteins that are expressed at low levels.

Moreover, in yeast, essential genes tend to cluster together (Pal, Hurst 2003; Hurst, Pal, Lercher 2004), one indication of the non-randomness of gene order. The origin of these clusters is obscure. A model proposed by Batada and Hurst shows that these clusters could effectively avoid transcriptional bursting because they are located in relatively open chromatin, and thus reduce expression noise (Batada, Hurst 2007). Thus, noise control appears to be an important force in genomic evolution.

1.3 Yeast networks and phenotypic evolution

Since none of the proteins or enzymes in the cellular machinery work fully independently, pure reductionism, which focuses on individual components and protein function, is incapable of providing a meaningful picture of natural processes. Since the release of the human genome sequence (Lander *et al.* 2001; Venter *et al.* 2001), integrated methods based on high-throughput technologies have been providing a great opportunity for the systematic investigation of cellular function. Among these technologies are microarrays, large-scale tandem affinity purification coupled to mass spectrometry (TAP-MS), and two-hybrid or protein chips. This increased trend towards the study of interactions is known as network biology or systems biology. So far, large amounts of data have been accumulated, *e.g.*, on protein–protein and genetic interactions (Breitkreutz *et al.*

2008) as well as on metabolic networks. Based on the knowledge thus acquired, we are also able to predict interactions in newly sequenced species (Jensen *et al.* 2009).

To date the best-annotated networks in yeast are the co-expression network, protein–protein interaction network, genetic interaction network, metabolic network and transcriptional regulatory network. While some signaling pathways are available, they are far from combining into a near-complete signaling network.

1.3.1 Protein-protein interaction network

The earliest data on the yeast protein–protein interaction (PPI) network came from two large-scale yeast two-hybrid screens. In the first study in 2000, a total of 957 interactions were detected involving 1004 proteins (Uetz *et al.* 2000), while the second study identified 4546 interactions among 3278 proteins (Ito *et al.* 2001). Note that the detected interactions do not overlap very much, indicating that the complete interaction network is indeed much larger. More recently, another two independent studies have greatly increased the number of interactions: a protein-fragment complementation assay (PCA) method applied to 1124 endogenously expressed proteins revealed as many as 2770 interactions (Tarassov *et al.* 2008), and a "second-generation" high-quality, high-throughput yeast two-hybrid screen found another 1809 interactions (of which more than 1500 are new compared with the previous two investigations) (Yu *et al*. 2008). Protein–protein interaction networks are essentially dynamic during the cell cycle, with many new interactions being formed and lost (de Lichtenberg *et al.* 2005).

The evolution of the protein–protein interaction network has received much interest. Protein complexes, which are aggregates of several proteins working together, play a particularly important role in this network evolution. A comparative investigation of yeast PPI with PPIs in humans revealed that 90% of co-complex membership is conserved during evolution (van Dam, Snel 2008), indicating that the evolution of complexes in the protein–protein interaction

Table 1-1 Different technologies of detecting single molecular (mRNA and protein) which is modified from (Raj, van Oudenaarden 2009)

	Method	Endogenous protein detection?	Real-time measurements?	Multiple proteins species?	Other advantages	Other disadvantages
Detection of single mRNA molecule	MS2	No	Yes	No	No need for external interventions (e.g., microinjection), yields spatial information	mRNA tend to form clumps, requires long UTR sequence elements
	FISH	Yes	No	<= 3	Yields spatial information	Imaging can be difficult in small organisms
	Single-cell RT-PCR	Yes	No	<= 5	Simple to perform with a large dynamic range	Requires careful standardization, questions about efficiency and sensitivity at low numbers, no spatial information
	Digital single-cell RT-PCR	Yes	No	<= 2	Easily interpretable signals, sensitive at low numbers of molecules	Requires microfluidics, questions about RT efficiency, no spatial information
	Molecular beacons	No	Yes	Yes	No clumping of transcripts, yields spatial information	Requires microinjection or other invasive delivery methods, requires long UTR sequence elements
Detection of single protein molecule	β-galactosidase microfluidics	No	Yes	No	Works in a variety of cell types	Requires microfluidics, cells must be permeabilized, not as effective with large molecule numbers, no spatial information
	Single fluorescent protein imaging	No	Yes	Potentially	Yields spatial information	Unlikely to work in present form in organisms larger than bacteria
	Single-protein antibody labeling	Yes	No	Potentially	Simple to perform with a large dynamic range, works in many organisms	Requires microfluidics and complex optics, questions about antibody efficiency and sensitivity at low numbers, no spatial information

network is mainly due to gain and loss of complex members rather than network rewiring. Most of the protein complexes originated very early during evolution, and gene duplication played an important role in the growth of these complexes (Yosef *et al.* 2009).

1.3.2 Transcriptional regulatory network

In the cell, the spatial and temporal expression of each gene differs. How does the cell control this process? How are the genes that are involved in development transcribed at a specific time of the lifecycle? To answer these questions, one has to know the exact DNA sequences that determine the transcription of a gene and the exact transcription factors that bind to the DNA, resulting in transcriptional interactions. Many technologies have been developed to detect such interactions; the most used method is the ChIP-chip (or more recently ChIP-seq) method. ChIP-chip was developed first in yeast and then applied to other organisms. The technology includes several steps, including preparation of cross-linked chromatin; chromatin immunoprecipitation; amplification and fluorescence labeling of chromatin immunoprecipitation DNA; microarray hybridization (or DNA sequencing) and data analysis; and finally, validation of the identified binding sites (Kim, Ren 2006).

Today, the known transcriptional network of yeast consists of more than 10,000 interactions, involving about 150 transcription factors. Different transcription factors (TFs) bind to different numbers of target genes: those TFs binding to many target genes usually regulate multiple cellular processes, while TFs that bind to few target genes specifically regulate one or a few cellular processes (Balaji *et al.* 2006).

1.3.3 Co-expression network

Gene co-expression networks are often constructed using microarray data. They are usually used to predict the function of unknown genes, because genes with high co-expression levels are often also functionally related (Luo *et al.* 2007; Ruan, Dean, Zhang 2010). Two methods are frequently used, employing either weighted or unweighted co-expression matrices. In unweighted networks, two

genes are linked if their co-expression values are larger than a threshold (hence links have no weights and are just either present or absent), whereas in weighted networks, the co-expression level is used to weight the link (Zhang, Horvath 2005). As a consequence, weighted co-expression networks are much larger than unweighted networks.

Co-expression networks typically exhibit a scale-free, small-world structure, which means that most of the nodes have few connections while a small number of nodes has high numbers of interactions. It has been reported that this structure can be explained by a purely neutral model under simple assumptions about the duplication and loss of genes and of transcription factor binding sites (van Noort, Snel, Huynen 2004).

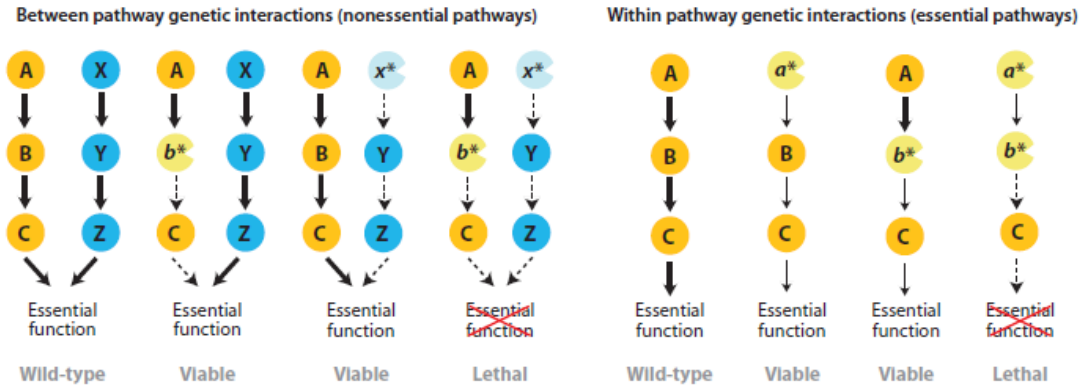
1.3.4 Metabolic network

Since the advent of molecular biology in the 1950s, biochemists have been working on various enzyme activities, accumulating an enormous amount of metabolic reaction data. But not until the very late 20th century (1999) were genome-scale metabolic reaction maps constructed in a prokaryote, *Haemophilus influenza* (Edwards, Palsson 1999). Then in 2003, the first genome-scale metabolic network for a eukaryote (yeast) was released (Forster *et al.* 2003). The network accounted for 708 proteins, corresponding to 1035 reactions. In total about 16% of yeast genes were included in this metabolic network. To date, more than 50 genome-scale metabolic networks of various species have been published, including networks of specific human cell types.

What is the use of the constructed metabolic model? One common use is for metabolic engineering (Feist, Palsson 2008), which aims to discover new metabolic pathways and improve the performance of old pathways. For example, *S. cerevisiae* could be engineered to produce 59 g/l of malate, which is five times higher than the amount previously reported (Zelle *et al*. 2008). The other common use is for the study of differences in species-specific networks, in order to gain information on network evolution. For instance, a metabolic network model of four halophilic archaea has been developed and used to compare

phenotypic differences (Falb *et al.* 2008); a yeast model was used to study the adaptation of metabolic networks, indicating that the topological properties of a metabolic network may be formed as a byproduct of selection for other phenotypes; thus, topological properties *per se* do not necessarily give selective advantages (Papp, Teusink, Notebaart 2009).

a:



b:

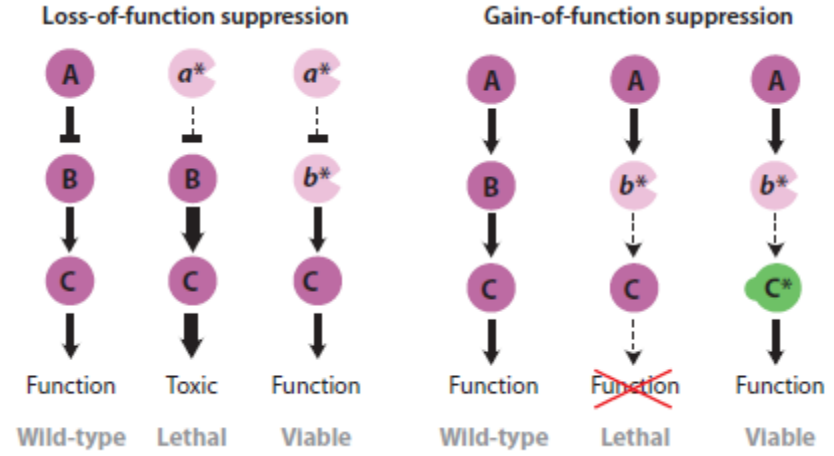


Figure 1-3 Molecular mechanisms for negative genetic interactions (a) and positive genetic interactions (b), edited from (Dixon *et al*. 2009).

1.3.5 Genetic interaction network

Genetic interaction refers to an interaction between phenotypic effects of mutations (mostly complete knockouts) in two different genes. Three effects are possible: no interaction (*i.e.,* the phenotype of the double mutant is as expected - a multiplicative combination of the single mutant phenotypes), positive or

alleviating interactions (*i.e.,* the double mutant is less impaired than expected), and negative or aggravating interactions (*i.e.,* the double mutant is more severly impaired than expected). The strongest negative interaction results in synthetic lethality.

Most gene pairs show no genetic interaction (Tong *et al.* 2004). Negative genetic interactions arise, *e.g.,* from the disruptive consequences of two compensatory pathways, or a situation where each gene contributes to the decreased flux of one essential pathway (Dixon *et al.* 2009) (Figure 1-3a). Positive interactions often occur when two genes share the same protein complex, when two genes act in a single non-essential pathway, or when the inactivation of gene B decreases the accumulation of the toxic protein product caused by the inactivation of gene A (Figure 1-3b).

Researchers have also studied the evolution of genetic interaction networks. Comparisons based on both manually curated and high-throughput data of *S. cerevisiae* and *S. pombe* revealed that synthetic lethal interactions are strongly conserved between distantly related species (Dixon *et al.* 2008).

1.3.6 The concept of phenotypic capacitors

Natural environments are subject to unpredictable events, such as rain, wind, or extreme temperatures. All organisms living on Earth were capable of adapting to these events in their environments, with few phenotypic changes. This indicates that living organisms can buffer variation in their environments. Furthermore, there is also enormous intrinsic (non-environmental) variation, such as genetic differences within populations, and stochastic variations in expression arising via gene transcription and translation systems. The buffering effect of the system is often believed to stem from the robustness of the underlying networks, and the genes that contribute to this kind of robustness are called phenotypic capacitors (Masel, Siegal 2009). When a phenotypic capacitor is functionally impaired, the organism will become less robust and thus exhibit more phenotypic variation (Masel, Siegal 2009). The most famous example of this process is the chaperone Hsp90, which targets many signaling transduction

proteins. The fact that changing Hsp90's function leads to more morphological variability has been demonstrated in both *Drosophila* and *Arabidopsis* (Rutherford, Lindquist 1998; Queitsch, Sangster, Lindquist 2002).

Are there also phenotypic capacitors in yeast? Using numerical simulations in gene networks, it was predicted that there would be a large set of genes for which abnormal function could lead to increased morphological variability (Bergman, Siegal 2003). Based on high-throughput morphological phenotyping of individual yeast cells following single-gene deletion, more than 300 phenotypic capacitors were subsequently identified in yeast (Levy, Siegal 2008).

1.3.7 Network and protein evolution

Protein evolution is of particular interest to biologists, as proteins form the fundamental building blocks of organisms. In the past ten years, much information has been gathered on the factors that influence protein evolution rates, such as expression level (Pal, Papp, Hurst 2001), recombination rate (Rattray, Strathern 2003), mutation rate (Datta, Jinks-Robertson 1995), and protein connectivity in networks (von Mering *et al.* 2002). Among these factors, the level of protein expression is thought to be the main contributor to the evolutionary rate of the sequence. In yeast, the expression level can explain more than 50% of the variation in the synonymous substitution rate and nearly half of the variation in the non-synonymous substitution rate (Drummond, Raval, Wilke 2006). The underlying reason for this correlation is probably that for highly expressed proteins, selection against protein misfolding is increased, as for such proteins, the toxicity caused by the aggregation of misfolded proteins would be extremely harmful. Thus, selection for increasing translational robustness causes proteins with high expression levels to evolve slowly (Drummond *et al*. 2005; Drummond, Wilke 2008).

Different network properties have also been found to influence the rate of protein evolution; the most highly studied property is protein connectivity (Fraser *et al.* 2002). A negative correlation was observed between the number of

connections in the protein–protein interaction network and evolutionary rate; this was assumed to be the result of co-evolution of important interacting sites (Fraser *et al* . 2002). What is more, proteins that are more "central" in the network evolve more slowly (Hahn, Kern 2005).

1.3.8 Network and phenotype

One of the ultimate objectives of network biology is to understand the phenotypic consequences of the networks under different environmental stimuli; here phenotypic outcome includes cell behavior, nutrition, and the response to toxic material. Understanding the network behavior of individual cells would also greatly facilitate our understanding of the process of cell–cell signaling and multi-cell response.

1.3.9 Understanding phenotypic effects on the network level – modules

Modular structure has been shown to be a common feature of complex networks, such as biological networks, the internet and social networks. In the protein–protein network, a group of proteins that function together, such as protein complexes, can be treated as a module (Hartwell *et al.* 1999). A module might be the basic functional unit of many critical cell events, such as chromosome segregation.

Some research groups have highlighted the potential connections between modules and the phenotypic features of the cell. In one particular study, 10 modules that are potentially connected with breast cancer were identified (Niida *et al.* 2009). A further investigation based on the Iterative Clique Enumeration algorithm revealed 19 modules that showed expression correlations with tumor stage (Shi, Derow, Zhang 2010). Additionally, a robust module network including microRNAs was introduced as a potential modulator in cancer cells (Bonnet *et al* . 2010). While these resulst highlight the connection between functional modules and phenotypic features of a cell, it is evident that further investigations are needed to fully understand the underlying mechanisms.

Chapter 2 Gene Orientation and Non-coding Transcripts Modulate Noise Levels

2.1 Project summary

For some genes, such as essential genes, expression when expression is needed is vital, hence low noise is selectively favourable. For others, such as stress-response genes, some level of noise is necessary for coping with stochasticity. Noise also provides a dice-like mechanism to enable cell fate choice. But how is noise controlled? Recent evidence suggests that chromatin opening and shutting can be an important mechanism modulating noise. Assuming this, we hypothesise that gene orientation may be a key determinant of noise. We suggest that genes in divergent orientation may benefit from the close proximity, on the opposite strand, of the promoter of the neighbour, this being especially so when the promoters of each gene overlap (bi-promoter genes). A stochastic simulation model confirms the expectation that if expression of one gene ensures some resilience to stochastic chromatin shutting of a neighbour (and vice versa) noise can be reduced. The hypothesis has the potential to explain why some genes have an antisense non-functional non-coding RNA transcribed from a shared promoter. The model correctly predicts that 1) protein coding genes with bi-promoter architecture, including those with a ncRNA partner, have lower noise than other genes; 2) ncRNA antisense transcripts are associated with genes that a priori would be under selection for low noise (essential genes); 3) divergent gene pairs have correlated expression noise, while convergent or co-oriented gene pairs do not; 4) mean expression noise level of divergent transcripts correlates with the distance between the two promoters, 5) essential genes reside in divergent orientation more than expected by chance, while 6) stress response genes tend not to be divergent. We conclude that gene orientation is an important modulator of noise levels with an impact on gene

order evolution.

2.2 Introduction

Between genetically identical cells we see variation in abundance of any given transcript or protein. This variation is noise in gene expression (Elowitz *et al*. 2002; Ozbudak *et a l*. 2002; Blake *et al* . 2003; Kaern *et al* . 2005; Raj, van Oudenaarden 2008). There is also considerable variation between genes in the level of noise (Bar-Even *et al*. 2006; Newman *et al*. 2006; Sigal *et al*. 2006). In part the between-gene variation in noise, assayed as the coefficient of variation (standard deviation/mean across individuals), is accounted for by expression level, there being lower noise for more highly expressed genes (Raser, O'Shea 2004; Bar-Even *et al* . 2006; Newman *et al* . 2006; Yin *et al* . 2009). Even controlling for this, using an abundance corrected noise measure, there remains, however, striking variation (Bar-Even *et al.* 2006; Newman *et al.* 2006). What are the underlying determinants of this abundance-independent variation in noise levels between genes and might the variation between genes in their noise levels reflect the activity of selection?

For some genes high noise is likely to be significantly deleterious. In particular, essential genes are, by definition, genes for which reductions (but not necessarily increases) in dosage are highly deleterious. Stochastic fluctuation in abundance of such proteins is thus likely to be highly deleterious as dose can, by chance, sink to fitness-reducing low levels (Fraser *et al*. 2004). We should then expect such proteins to be under selection to have low noise. That they do have low noise is consistent with such a model (Fraser *et al* . 2004; Batada, Hurst 2007). Haplo-insufficient genes have yet lower noise, as might be expected (Batada, Hurst 2007; Lehner 2008). Conversely noise can be advantageous to some degree. Noise, for example, can provide the underlying basis of dice-like behaviour necessary for alternative cell fate specification in a genetically uniform population of cells (e.g. the developing embryo) (Chang *et al*. 2008; Choi *et al*. 2008). Further, if the environment is stochastic, noisy gene expression can be an effective mechanism to cope with uncertainty (Kussell, Leibler 2005; Blake *et al.* 2006). Noise in the expression of metabolic import channels is, for example,

potentially advantageous when nutrient availability is fluctuating. It is striking that of all metabolic genes, import channels are the most noisy (Zhang, Qian, Zhang 2009). Stress response genes are also expected to be high noise genes, these also being responsive to an uncertain environment (Lopez-Maury, Marguerat, Bahler 2008).

While noise may then be an important target of selection, this leaves the issue of how mechanistically noise is modulated. At the transcript level slow translation rates and low mRNA half lives are likely to reduce noise (Blake *et al*. 2003; Kaern *et al*. 2005; Ramsey *et al*. 2006; Kar *et al*. 2009). Much noise modulation is probably achieved at the transcriptional control level. TATA controlled genes, in particular, tend to be especially noisy (Raser, O'Shea 2004; Blake *et al*. 2006; Field *et al*. 2008) and expression noise of genes is increased when the binding site of GAL1 promoter is moved closer to a TATA-box (Murphy, Balazsi, Collins 2007). The underlying cause of an association with TATA is unresolved. The high expression variation of TATA-box containing gene may be owing to the binding stability of transcription-mediating factor TBP (Blake *et al*. 2006) or related to the high nucleosome occupancy (Field *et al*. 2008; Tirosh, Barkai 2008), suggesting a link to chromatin dynamics. A recent report of the lack of activating histone modifications in this region (Choi, Kim 2008; Choi, Kim 2009) supports the latter.

The above results go someway to unifying TATA control with chromatin level control, also thought to be important in noise modulation (Newman *et al*. 2006). In one striking example (Raj *et al*. 2006), a pair of genes inserted in tandem showed co-ordinated spiking in their gene expression, while the same pair when unlinked showed little co-ordination. This result suggests a model whereby opening of chromatin permits accessibility to transcription factors. Regular opening and closing of chromatin then leads to co-ordinated expression, and correlated noise levels, of neighbours. Such a model correctly predicts that across a genome, controlling for similarity of transcription factor control, linked genes show much higher levels of co-expression than do unlinked genes (Batada, Urrutia, Hurst 2007). This in turn is related to nucleosome occupancy (Batada, Urrutia, Hurst 2007). The magnitude of this effect is noteworthy: two random

unlinked genes regulated by the same set of transcription factors show no higher co-expression than a pair of linked genes with no similarity in their transcription factors (Batada, Urrutia, Hurst 2007).

This class of model has led to the suggestion that the genomic distribution of essential genes and chromatin control should co-evolve such that essential genes end up clustered into domains with largely open chromatin, thereby ensuring low noise and expression when expression is needed (Batada, Hurst 2007). The model has some predictive power. It correctly predicts, for example, that essential genes should be rare subtelomerically in yeast, these being domains inconsistent with permanently open chromatin. It also correctly predicts nucleosome occupancy in domains rich in essential genes and that noise levels of non-essential genes should be predicted by the local density of essential genes (Batada, Hurst 2007). Here we extend the logic of chromatin mediated noise modulation to propose that modulation of noise by DNA dynamics might affect gene pairs differentially dependent on their orientation.

Gene pairs can come in one of three orientations: convergent (→←), co-oriented (→→or←←) or divergent (←→). These three classes are not equally conserved. In human, mouse, and rat bidirectional gene organization tends to be both ancient and more conserved than alternative orientations (Trinklein *et al*. 2004; Sigal *et al*. 2006). Similarly, through the fungi, divergent gene pairs are more conserved in orientation than convergent or co-oriented gene pairs (Kensche *et al*. 2008). In some cases of divergent genes the promoter domains overlap. Here we define such bidirectional-promoter genes as those where the nucleosome free region (NFR) of the two genes overlap. In *Saccharomyces cer evisiae* we find, in agreement with prior results (Xu *et al* . 2009), that more than 60% of non-overlapping divergent protein coding transcripts share the same promoter region. For convenience we refer to genes with a bidirectional promoter as bi-promoter genes.

Bipromoter gene pairs are especially well conserved as a pair. This can be seen when comparing the current gene order in *Saccharomyces cerevisiae* with that seen in the ancestor, prior to the whole genome duplication (Gordon, Byrne,

Wolfe 2009). Comparing bipromoter pairs to divergent but non-bipromoter pairs using logistic regression, we find that bipromoter pairs are much better conserved as a pair ($p = 3 \times 10^{-7}$), even when controlling for co-expression level ($p = 0.03$) and intergene distance ($p = 0.00136$), known predictors of pair conservation (Hurst, Williams, Pal 2002; Poyatos, Hurst 2007). This conservation may reflect nothing more than the fact that inversions that break up bidirectional gene pairs are more likely to disrupt promoter architecture.

Here we note that divergent orientation, bipromoter architecture in particular, is peculiar in that it puts in proximity the promoters of the two genes. This we argue may well have consequences for noise levels as for divergent genes the transcription, or priming for transcription by PolII loading, makes the transcription of the neighbour more likely, either because it might decrease the probability that the relevant chromatin stochastically closes or increases the probability of it being opened. That neighbouring genes show co-ordinated expression (Batada, Urrutia, Hurst 2007; Raj, van Oudenaarden 2008), that such co-ordination is not simply owing to similarity in transcription factors and is related to local nucleosome occupancy (Batada, Urrutia, Hurst 2007), while noise of a transgene is dependent on the insertion site (Becskei, Kaufmann, van Oudenaarden 2005) all point to a coupling between chromatin neighbourhood and noise. That transcription affects chromatin status (Li, Carey, Workman 2007) suggests in turn that bipromoter genes are unlikely to have uncoupled expression. Indeed, in humans, it has been shown that intensive transcription at one locus frequently spills over into its physical neighbouring loci (both upstream and downstream) resulting in a time lagged burst of expression subsequent to the upregulation of the focal gene (Ebisuya *et al*. 2008). This spill over is thought to be at least in part owing to local relaxation of chromatin associated with the expression of the focal gene, as evidenced by changes in histone modifications (Ebisuya *et al*. 2008). The same effect is seen in yeast, only here the effect is much more highly localized, the spillover extending no further than 3kb (Ebisuya *et al*. 2008) as opposed to 100kb in humans.

Based on these observations, we propose that for bipromoter genes, the gene pair acts as it were as a partially self re-inforcing domain of open chromatin. Such

bi-promoter domains should, we hypothesise, increase the net likelihood that chromatin is open and should thus be conducive to low noise, enabling expression when expression is needed. This could explain why some genes have non-coding unstable RNAs produced off a bidirectional promoter. Below we start by examining the hypothesis by reference to stochastic simulations.

2.3 Materials and methods

2.3.1 Dataset

All yeast (*Saccharomyces cer evisiae*) transcripts as observed by tiling arrays under three conditions (YPE, YPD and YPGal) and their genomic coordinates were obtained from (Xu *et al* . 2009). Two transcripts were considered as bi-promoter transcripts if they share the same 5' nucleosome free region (NFR), where NFR was defined as a nucleosome deplete region ≥80bp, according to (Xu *et al*. 2009). These transcripts were defined as divergent (←→), convergent t (→←) or co-oriented (→→ or←←) by their coordinates in the genome. Essential genes in rich media were downloaded from the web site of the *Saccharomyces* Genome Deletion Project[4]. Both the yeast gene order (Version 2) and genome annotation information were taken from YFOB[5]. For more than 2,000 proteins, expression noise data in rich media were obtained from (Newman *et al* . 2006). We used the distance to median noise level (DM_YEPD) in our analysis to get rid of the confounding influence of protein abundance. Genes whose promoter contains a TATA-box were derived from a large TATA-box gene enquiry experiment (Basehoar, Zanton, Pugh 2004). Codon usage bias (FOP) was obtained from (Drummond, Raval, Wilke 2006). The relationships between transcription factors (TF) and their target genes were derived from the yeast transcriptional regulatory network (Balaji *et al*. 2006). In total, 12873 regulatory interactions were indentified in this network. Stress-related genes and growth-related genes were obtained from (Wapinski *et al*. 2007) and co-expression level of adjacent gene pairs as previously reported (Batada, Urrutia, Hurst 2007). Haploinsufficent genes were taken from

[4] http://www-sequence.stanford.edu/group/yeast_deletion_project/deletions3.html

[5] http://wolfe.gen.tcd.ie/ygob/

(Deutschbauer *et al*. 2005) and Genes with type I and type II promoters were obtained from (Field *et al*. 2008). 431 type I genes and 565 type II genes were included in our analysis. Protein complexes were gained from (Yin *et al*. 2009).

2.3.2 Data analysis

Transcripts that share the same 5' NFR were described in Xu *et al*. (Xu *et al*. 2009). The noise of each protein measured by Newman *et al*. (Newman *et al*. 2006) was used to represent the noise of the transcript. In the comparison of the noise of proteins derived from divergent transcripts to the noise of proteins without divergent transcripts, transcripts with complex annotations were excluded (e.g. the annotation "other", which means the transcript contains multiple open reading frames or is a mixture of non-coding and coding parts). In the calculation of the correlation between noise levels of protein pairs, transcripts that contain multiple annotation features (e.g. the annotation "other", which means the transcript contains multiple open reading frames or is a mixture of non-coding and coding parts) were excluded. In the calculation of the correlation between noise level and the distance between transcription start sites, we used the mean noise level of the two proteins if the noise of both proteins had been measured. If one gene transcript shares its promoter with a non-coding transcript, the noise of this gene was chosen to represent the noise of the two transcripts in the calculation. We used the lawstat package in R to perform the Brunner-Munzel test (Brunner, Munzel 2000; Hui, Gel, Gastwirth 2008).

2.3.3 Randomization test of the correlation between noise levels of gene pairs.

Our model predicts that the expression noises of two divergent genes should be positively correlated due to the shared chromatin regulation, as chromatin regulation processes are responsible for much of the expression noise in yeast (Choi, Kim 2008; Choi, Kim 2009). To check if there is a positive correlation between expression noise in divergent, convergent and co-oriented gene pairs, and to obtain the significance level of any such correlation, we employed a randomization procedure. In this we extract the noise level for each protein,

orient the gene pairs by their strand location for divergent and convergent gene pairs, by their transcription order for co-oriented gene pairs, calculate the spearman correlation level for this data, randomize one column of genes 10,000 times and determine the correlation for each. The significance level of the observed correlation is (m+1)/10001 where m is the rank of the true correlation compared against the randomizations.

2.3.4 Randomization test to determine whether essential-essential gene pairs are more likely to be divergent gene pairs.

The *S. cerevisiae* gene order was taken from the Yeast Gene Order Browser[6], Version 2. The procedure is as follows: 1: count the number of divergent essential gene pairs in the *S.cerevisiae* genome; 2. randomize the position of essential genes in each chromosome 1,000 times and calculate the number of divergent essential gene pairs for each; 3. The significance level of this number is (m+1)/1001, where m is the rank of the true number compared with the randomizations.

2.3.5 Method to test to the density of essential genes in different gene types.

To calculate the density of essential genes surrounding essential bi-promoter genes and essential non-bi-promoter genes, a +/- 5 gene window was used to scan the yeast chromosomes (the *S. cer evisiae* gene order we used is from YGOB, as described above). To avoid biases caused by the fact that essential genes tend to be in divergent gene pairs, the direct (+1 and -1) gene neighbours were excluded from the scan.

2.3.6 Randomization test of bi-promoter genes show significantly lower noise level after control for transcription factor number.

In order to determine the significant level that bi-promoter genes have lower noise than other genes after control for the number of transcription factors (TFs), a randomization test was utilized. The procedure as follows: 1). Assume that all

[6] http://wolfe.gen.tcd.ie/ygob/

the genes have more than 10 transcription factors have 10 transcription factors. 2). for each TF categories (1 to 10, there is no bi-promoter genes that have 0 TF in our dataset), we calculated the absolute value of difference between the mean noise level of bi-promoter genes and other genes (abs (mean noise of bi-promoter genes – mean noise of other genes)). 3). sum up the difference of all the categories to get the overall difference value. 4). Random assign the noise data in each categories and repeat the steps 2) and 3). After randomize the data 10000 times, we got the significant level of the test (m+1)/10001, where m is the rank of the true number compared with the randomizations.

2.3.7 Calculation of the evolutionary rate of non-coding transcripts

The alignment of non-coding RNAs in 38 *S. cere visiae* strains and 4 related yeast species were obtained from recently published genomic sequencing data (Liti *et al* . 2009). For each transcript alignment, we also obtained the downstream 100 bp. Only ncRNAs with clearly mapped end, and only downstream sequences without overlap with any annotated features were included, resulting in 234 alignments.

For each non-coding transcript alignment and downstream 100 bp alignment, we calculated Nucleotide diversity Pi (p) within the *S. cerevisiae* population and divergence K (the average number of substitutions per site, using the Jukes-Cantor model to correct for multiple hits) among 5 yeast species (Rozas, Rozas 1999).

2.4 Results

2.4.1 The stochastic simulation model

Consider a pair of neighbouring genes. The promoter of each we presume can exist in one of two states, either in open chromatin or closed. Transcription is only possible, we assume, when chromatin is open. Here, note, we ignore the possibility that transcription factors might also act to open chromatin. Assuming independent behaviour of the two genes, the probability that open chromatin closes within a fixed time interval is p_c, while the probability closed chromatin

opens in p_o. If chromatin is open, then transcription is possible, occurring with a probability p_t. A transcriptional event results in N proteins before the mRNA is lost and protein decays with probability p_d.

The novel component of the simulation is to suppose that transcription of one gene might alter chromatin dynamics of the other and in turn affect transcription. There are two ways (not mutually exclusive) by which transcription of one gene might mediate such effects: either by reducing the probability that chromatin of the other promoter will shut, if open, or by increasing the probability of the chromatin opening if shut. We model both independently and consider a third model combining both.

We start by considering the case where the probability of shutting alone is modified (model 1). We can then define a parameter, i, for the level of independence between the genes, such that if one gene is being transcribed the probability that chromatin associated with the other gene's promoter will shut will be $i.p_c$. For i=1, the two genes are perfectly independent (e.g. not bidirectional). For i=0, transcription of one gene holds open the chromatin of the other gene, if the chromatin was already open. In this model, if the chromatin of the other gene is closed, it isn't forced to open by the activity of the neighbour. This coupling is hence in the form of resilience to chromatin closure.

In the second model, we consider that transcription of one gene increases the chances that the promoter of the other is opened if closed, but doesn't affect the probability of closure if open. If one gene is been transcribed, the probability that the chromatin of the other gene will open, if closed, is $(2- i).p_O$. In the final model (model 3), we incorporate both effects. For further details see supplementary experimental procedures 1.

For each simulation we follow the chromatin state, the transcriptional state and the protein level over 10,000 time units, updating status each time unit. Noise for the protein is defined as the standard deviation in protein level over the time course / mean level (note that variation over the time course is equivalent to

variation between unsynchronized replicates at any given time). An analogous definition is used for the transcript-level noise. Co-expression between the two genes is the Pearson product moment correlation through the time course of the pair. Chromatin fluctuation is the probability of observing a change in chromatin state in a randomly chosen iteration.

In the model in which transcription exclusively increases the chances of closed chromatin opening (model 2), in nearly all parameter space increasing interdependence (i -> 0) promotes low noise. Given this, we present in detail the less permissive model (model 1). The results from models 2 and 3 are presented in Figures 2-1b and 2-1c.

A typical result for model 1 is presented in Figure 2-1a. Here $p_c = p_o = 0.5$. Note that as the likelihood of coupling decreases so noise goes up and co-expression is reduced. More generally, for a variety of parameter values we need to consider the correlation between noise level and i. If, as in figure 1, this is positive, then increased coupling, (i->0), ensures reduced noise. We consider simulations in which for the two genes all parameters are the same, but we vary independently both p_c and p_o over the range 0.05 to 1 under increments of 0.05 with 10 replicates for each set of parameter values. We find that as regards transcriptional noise a positive correlation is always seen (Figure 2-2, blue points). However, owing to stochasticity in protein degradation this does not necessarily translate to protein level noise always decreasing with decreasing independence.

We find that protein noise can be increased when the stochastic probability of chromatin closure is high and thus most of the time no transcription is happening (Figure 2-2, red points). This is largely dependent on p_c not being too high (Figure 2-3a), as opposed to variation in p_o (Figure 2-3b). The causality of the negative correlation when closure probability is high is intimately related to effects on protein abundance. When closure probabilities are high (and transcription rates low), there is a positive correlation between protein abundance and protein noise (Figure 2-4), while, when closure rates are lower this correlation switches to a negative correlation. This most likely reflects the

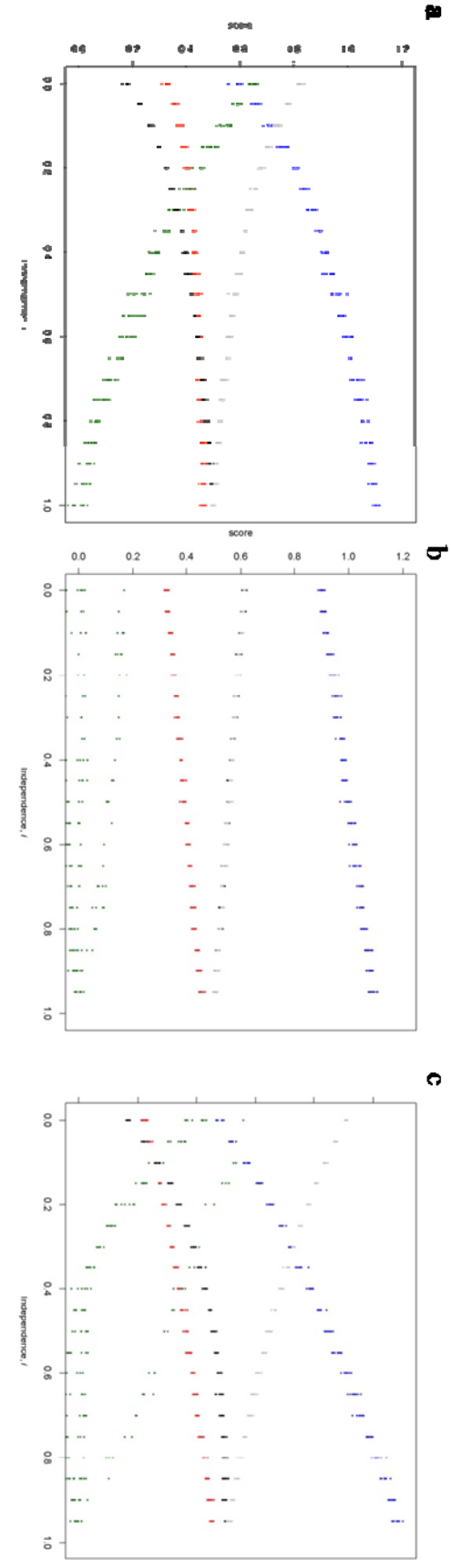


Figure 2-1 The relationship between the independence between two neighbouring genes and various noise and co-expression parameters. For this plot $p_c=p_o=0.5$. Data: transcriptional noise, blue; protein noise, red; co-expression, green; chromatin fluctuation, black; proportion of time chromatin open, grey. Other parameter values: N=100, p_f= 0.9, d=0.7. a) the model that transcription of one gene decrease the chance of the close open chromatin of another gene; b) the model that transcription of one gene increase the chance of the open closed chromatin of another gene; c) in the model that both of the effects exist.

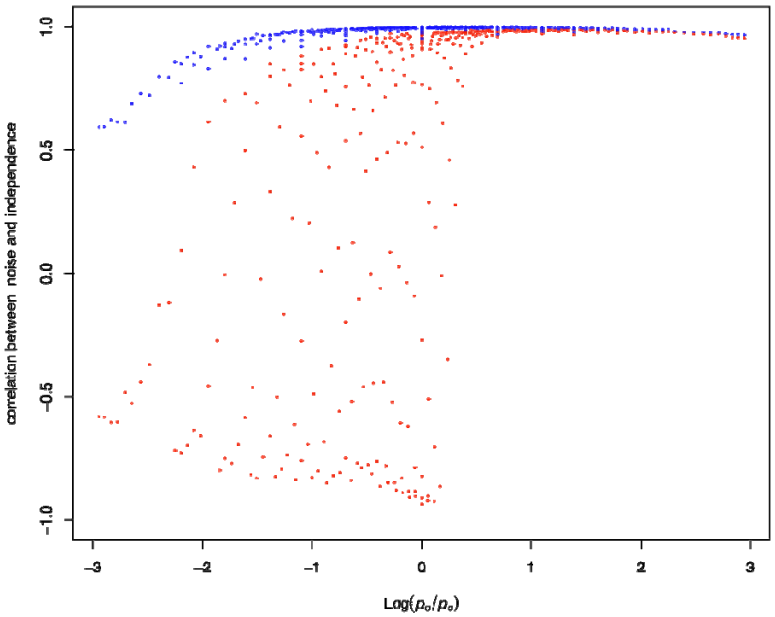


Figure 2-2 The correlation between noise and independence as a function of the ratio of the probabilities of chromatin opening and shutting. A positive correlation indicates decreased noise with increasing inter-dependence. Protein noise, red; transcriptional noise, blue. Other parameter values, N=100, p_f= 0.9, d=0.7.

fact that when closure rates are high, little transcription is seen and protein levels can descend to zero, thereby reducing the variance in levels until the next transcriptional event. With some degree of coupling between the genes the protein abundance level is raised and so noise is raised. The transcripts are, however, rare and lost almost immediately. As in yeast we see a negative correlation between protein noise and protein abundance (Newman *et al*. 2006), we surmise that true closure rates are relatively low, predicting a decrease in protein noise with increasing coupling.

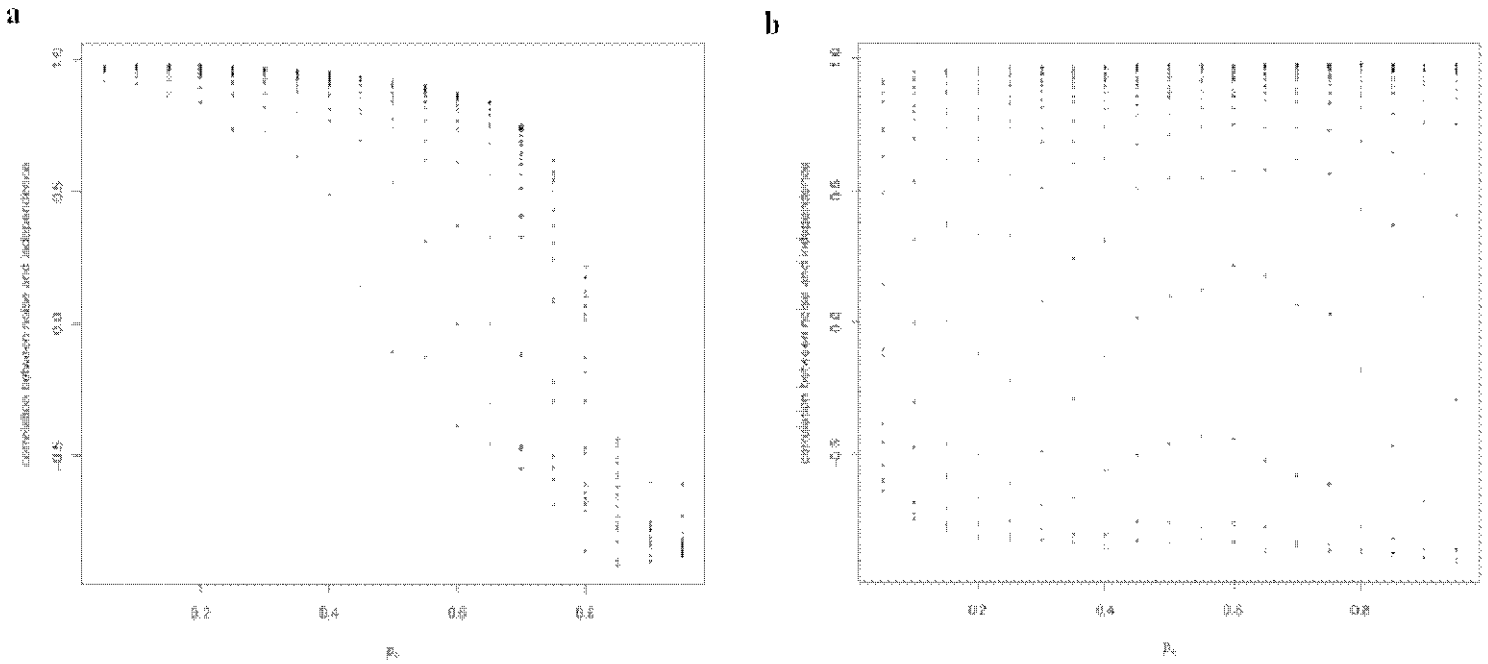


Figure 2-3 The correlation between noise and independence as a function of the probabilities of chromatin shutting (3a) and opening (3b). A positive correlation indicates decreased noise with increasing inter-dependence. Protein noise, red; transcriptional noise, blue. Other parameter values, N=100, p_f= 0.9, d=0.7.

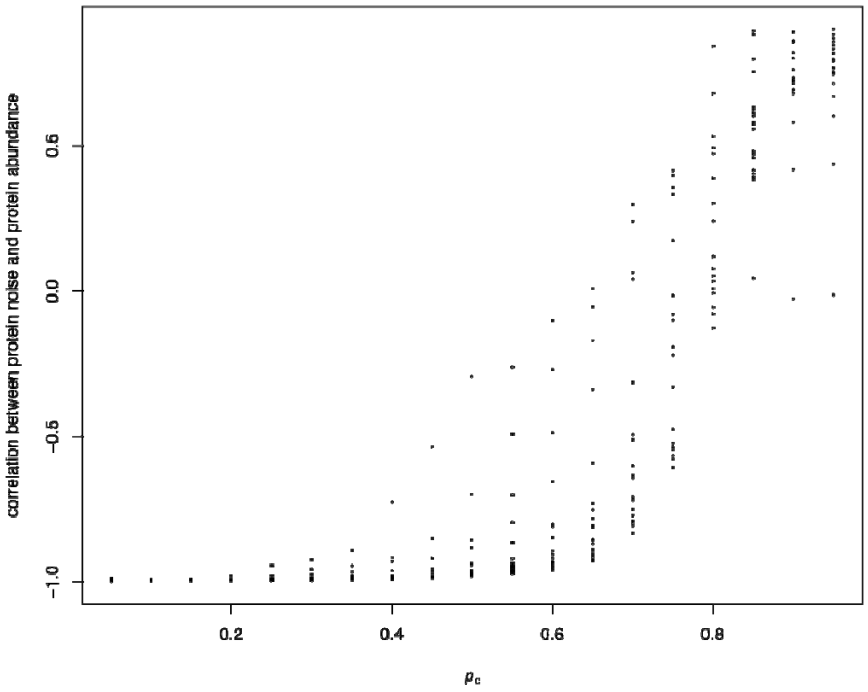


Figure 2-4 The relationship between the correlation between protein noise and protein abundance as a function of the probability of chromatin closure.

2.4.2 Noise reduction has abundance-dependent and abundance-independent components

While in the above models we see a robust relationship between coupling and noise, much of this effect is likely to be owing to there commonly being lower noise for highly abundant proteins. In the simulations, increased coupling increases the abundance of the protein product by permitting a higher opportunity for transcription. This agrees with the prior suggestion that, for essential genes, an increase in dose may be beneficial as it both reduces noise and moves the mean expression level away from the danger zone, where low dose equates to large fitness effects (Choi *et al*. 2007; Yin *et al*. 2009). Note too that dose sensitive genes, such as essential genes are asymmetrically dose sensitive. While reduction of dosage is very costly (hence they are deemed essential) increases in dose do not have any similar effect. Indeed, it is notable that the set of genes for which gross over-expression has a phenotype shows

little overlap with the set showing fitness on reduction in dosage (Sopko *et al*. 2006). We conclude that it is likely to be advantageous for some dose sensitive genes to be configured in bipromoter architecture as it increases net abundance.

Given the above logic, we might also ask whether bipromoter genes are expected to have lower noise, even allowing for the increased abundance. To approach this we consider simulations in which we alter abundance by modifying factors that affect protein abundance independent of the effects of chromatin opening and shutting, and transcriptional bursting. We can then ask whether an independent gene pair (i=1) and a coupled pair (i=0) show different noise levels when steady state protein abundance levels are equal owing to differences in decay rates (higher for coupled genes). We find for all three models that bipromoter genes still show lower noise levels at any given abundance level. We also consider the possibility that transcripts that result from bipromoter activity produce fewer translated proteins than do those from independent genes, keeping the decay rates constant. Again we find that controlling for net protein abundance that coupled genes (i=0) have much lower noise than do independent ones. We conclude that noise modulation by modification of transcriptional bursting, owing to coupled gene activity, can have both abundance-dependent and abundance-independent causality. These results are in many regards comparable to those of Cook *et al*.(Cook, Gerber, Tapscott 1998), who, in examining a role for ploidy in noise modulation, identify both an abundance-dependent and abundance-independent component to noise modulation.

2.4.3 Bi-promoter transcribed genes have low expression noise

We tested the hypothesis that bi-promoter protein coding genes have low protein noise with the help of recently published yeast whole genome transcription data (Xu *et al*. 2009) to define gene orientation and presence of ncRNA, coupled with high resolution noise data on rich media provided for over 2000 protein coding genes specified by Newman *et al*. (Newman *et al*. 2006). In all, we analysed 7,272 well identified transcripts, of which 1,772 are non-coding

transcripts (stable unannotated transcripts and cryptic unstable transcripts, SUTs and CUTs) which is approximately 25% of all transcripts (Xu *et al* . 2009). Among transcripts with a mapped 5' nucleosome free region (NFR), 61% of the unannotated transcripts and 48% of the protein-coding transcripts initiated bidirectionally from shared 5' NFRs rather than initiating from their own promoters (Xu *et al*. 2009).

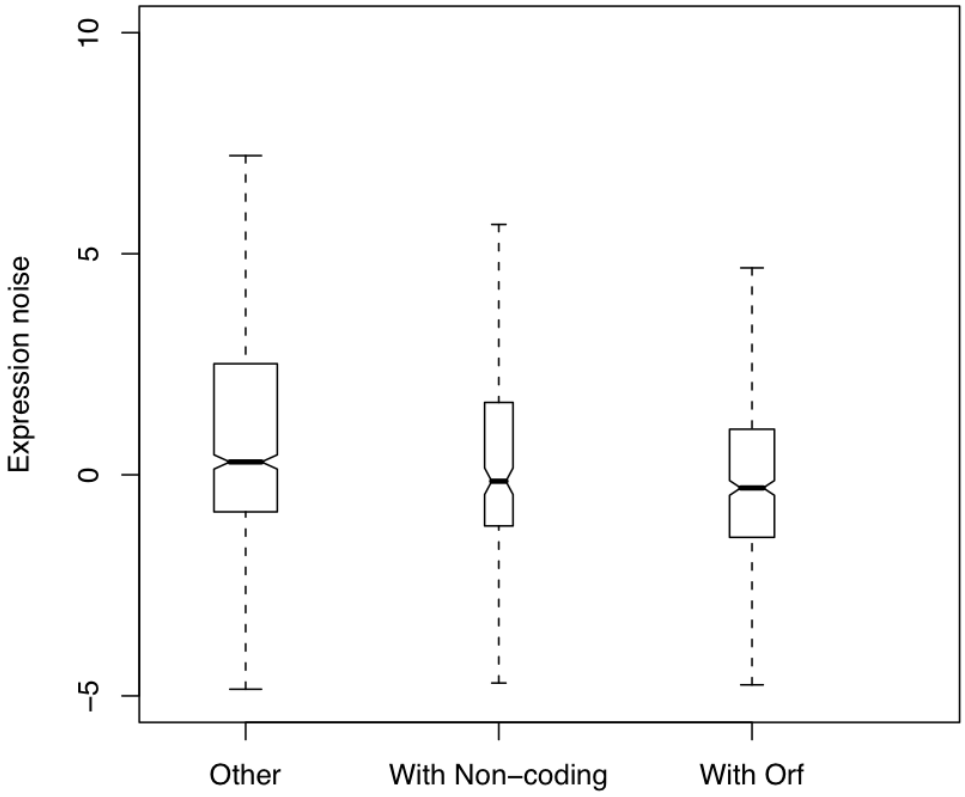


Figure 2-5 Genes which share a promoter (5' NFR) with either a non-coding transcript or coding transcript (ORF) show lower expression noise than genes without any bi-promoter transcript.

If our hypothesis is correct, protein-coding genes with a bi-promoter architecture (shared with either a protein coding gene or a ncRNA) should show lower expression noise. As we are not so interested in the hypothesis that bipromoter architecture might modify noise through modification of abundance, we restrict analysis to abundance corrected noise measures, as defined by Newman *et al* . (Newman *et al* . 2006). We also repeated analysis using

residuals from a loess regression of noise against abundance and find no important differences (data not shown). After removing the confounding transcript types (5'NFR tandem transcript, 3'NFR antisense transcript and 3'NFR tandem transcript) annotated by (Xu *et al*. 2009), we find that protein-coding genes with a bi-promoter structure, sharing their 5' NFR either with a coding gene or with a non-coding gene, show significantly lower expression noise than the genes that do not have a bi-promoter transcript structure (mean noise of bi-promoter genes = 0.33 +/- 0.11; of all non-bi-promoter genes: 1.76 +/- 0.15; Brunner-Munzel test $p = 4.1\text{x } 10^{-13}$, Figure 2-5). More generally, divergent genes (regardless of their NFR) have lower noise than those in alternative configurations (noise of non-divergent genes = 1.50 +/- 0.18, mean noise of divergent genes = 0.88 +/- 0.12, Brunner-Munzel test $p = 0.0077$). By contrast, convergent genes don't show significant differences in noise level compared with co-oriented genes ($p = 0.68$, Brunner-Munzel test).

2.4.4 Noise reduction and divergent ncRNA

This model not only has applicability in the case where both genes in the pair are protein coding. It also has the potential to explain why some genes have antisense non-coding RNA specified from a bi-directional domain. Such transcripts are now widely reported. In yeast, of the unannotated transcripts (ncRNA) which have mapped 5' NFR, 61% are bidirectional initiated from a shared promoter region (Xu *et al*. 2009). Similarly, mapping millions of short RNA reads generated from murine embryonic stem cells and other differentiated cell types has revealed abundant short transcription start site–associated RNAs, many of which are antisense transcripts (Seila *et al*. 2008). Likewise in humans, depletion of the exonucleolytic RNA exosome reveals lots of highly unstable RNA of promoter upstream transcripts (Preker *et al*. 2008). Similar RNAs are reported in chicken and *Drosophila* (Taft *et al*. 2009).

One model sees these as spurious transcripts, a consequence of illegitimate transcription factor activity (Brosius 2005). Our model suggests a functional explanation. For the chromatin to remain open and for noise to be reduced, permitting expression when expression is needed, polII priming or transcription

of a ncRNA through a promoter on the opposite strand to that of the focal protein coding gene would be an efficient mechanism to enable accessibility of the promoter domain of the focal gene. As expected, we find that bi-promoter protein coding genes have low noise both when they are partnered with a protein coding gene (p = 5.0x 10^{-14} compared with all other genes; Brunner-Munzel test), and when the partner is not protein coding (p = 0.0030, Figure 2-5).

2.4.5 Is noise more important than co-expression?

In simulations we find that co-expression is higher when genes are coupled (r=-0.86). While then the above results support the noise model, can we be confident that the function of bi-promoter architecture is ever to reduce noise rather than to increase co-expression levels? For the most highly co-expressed 2% gene pairs it is known that they tend to belong to the same functional class, are preserved as a pair over evolutionary time and are enriched in divergent orientation (Batada, Urrutia, Hurst 2007). For these there is little doubt that co-expression is functionally relevant. However, several findings support the proposition that noise modification is relevant. First, we see no significant correlation between co-expression level and mean noise, neither for divergent gene pairs (r = -0.064, p = 0.424), convergent gene pairs (r = -0.1038, p = 0.152), nor co-oriented gene pairs (r = -0.0672, p = 0.257). We do, nonetheless and as expected, find higher co-expression rates for divergent gene pairs (divergent gene pairs: mean co-expression =0.140 +/- 0.012; convergent gene pairs, mean co-expression: 0.107 +/- 0.010; co-oriented gene pairs: mean co-expression: 0.101 +/- 0.009; p = 0.0467 between divergent and convergent; p = 0.0019 between divergent and co-oriented and p = 0.333 between convergent and co-oriented, Brunner-Munzel test).

Second, co-presence of the product of transcription is unlikely to be the case for one class of ncRNA, cyptic unstable transcripts (CUTS), as these tend to be rapidly targeted for degradation (Neil *et al*. 2009; Xu *et al*. 2009). Importantly then, we find that when we consider protein coding genes partnered with CUTs through bi-promoters, they too have lower noise than other genes (p=0.012), but

no different from that of protein coding genes partnered with protein coding genes in a bi-promoter architecture (p>0.05).

A third line of evidence derives from examination of a class of genes where *a priori* we might know the fellow genes with which they might benefit from being co-expressed. The best candidates in this regard are proteins that belong to the same protein complex, that do indeed have high co-expression scores with fellow members (mean co-expression of genes from same complex: 0.1877 +/- 0.0026 and mean co-expression of genes from different complexes: 0.0253 +/- 0.0001, $p < 2.2 \times 10^{-16}$ in Wilcoxon rank sum test) (data from Wang *et al*. 2009). Given the need for transcription when transcription is needed, as expected complex-associated genes do indeed have low noise ($p = 7.3 \times 10^{-7}$ Brunner-Munzel Test). Further, as we would expect, genes specifying proteins in a complex tend to have bipromoter architecture more than expected by chance ($p < 2.2 \times 10^{-16}$, Fisher's Exact Test), this being true after control for essentiality ($p < 2.2 \times 10^{-16}$, Fisher's Exact Test). While, however, complex related genes both have low noise and are found more commonly in bipromoter architecture than expected by chance, we find no cases where two genes specifying proteins in the same complex are located in the same bi-promoter pair. These results strongly suggest that noise modulation above co-expression is key to selection on bi-promoter genes. A very few bi-promoter genes may well also benefit from their mutual co-expression, but the more relevant force may well be selection for noise modulation.

2.4.6 For noise, orientation of the ncRNA matters

While above we show that nCRNA in divergent orientation is associated with low noise of the protein coding gene, this does not demonstrate that oritentation *per se* is important. Is then low noise a general property of genes associated with ncRNAs, regardless of orientation, or is the divergent orientation important? We find that noise levels of proteins with an ncRNA from the same strand as the protein coding gene have higher expression noise than proteins with a ncRNA derived from a bidirectional promoter (bi-promoter with ncRNA noise=0.65, co-oriented with ncRNA noise=2.07, p=0.036; Brunner-Munzel test). This both

supports the hypothesis that the function of bi-promoter ncRNA is to reduce noise of the paired protein-coding gene and suggests that noise, rather than co-expression, can be the focus of selection. Moreover, genes with ncRNA from the same strand as the protein coding gene have higher expression noise than the protein coding genes which have a same strand protein coding gene neighbour (p = 0.026). This suggests that co-oriented ncRNAs may be a means to increase expression noise, a possibility we will not examine further.

Table 2-1 Binding of particular transcription factors cannot explain the low noise of bi-promoter genes. The noise level of bi-promoter genes is significantly lower than that of other genes both in the case of genes regulated by the same common transcription factor, and for those regulated by other transcription factors. p-values from Brunner-Munzel tests.

	Regulated by particular TFs	Regulated by other TFs
Bi-promoter genes	0.09 +/- 0.24 (322)	0.43 +/- 0.12 (1789)
Non-bi-promoter genes	1.15 +/- 0.26 (568)	1.57 +/- 0.15 (3921)
p	0.0013	4.0e-08

2.4.7 Results are robust to covariate controls

The above results are all consistent with our hypothesis but may have alternative explanations. Previous analysis of divergent promoters in mammals suggests that several particular binding motifs are enriched in bi-promoter structures (Lin *et al*. 2007a) and a particular binding protein, GABP, binds to more than 80% percent of divergent promoters (Collins *et al*. 2007). This raises the possibility that differential utilization of transcription factors might explain the low noise of bi-promoter genes.

To test this, we take three transcription factors (Balaji *et al*. 2006) that each regulate more than 100 genes and ask whether the mean expression noise of bi-promoter genes bound by these three TFs is lower than the noise of other genes that are bound by the same TFs. Second, we ask whether the expression noise of bi-promoter genes bound exclusively by other TFs (i.e. not the main three) is lower than that of non-bi-promoter genes bound exclusively by other

TFs. The results show that TF binding cannot explain the low noise in bi-promoter genes (Table 2-1). Further, when we control for the number of transcription factors regulating a gene, bi-promoter genes still show lower expression noise than other genes (p<0.0001 from randomization; Figure 2-6).

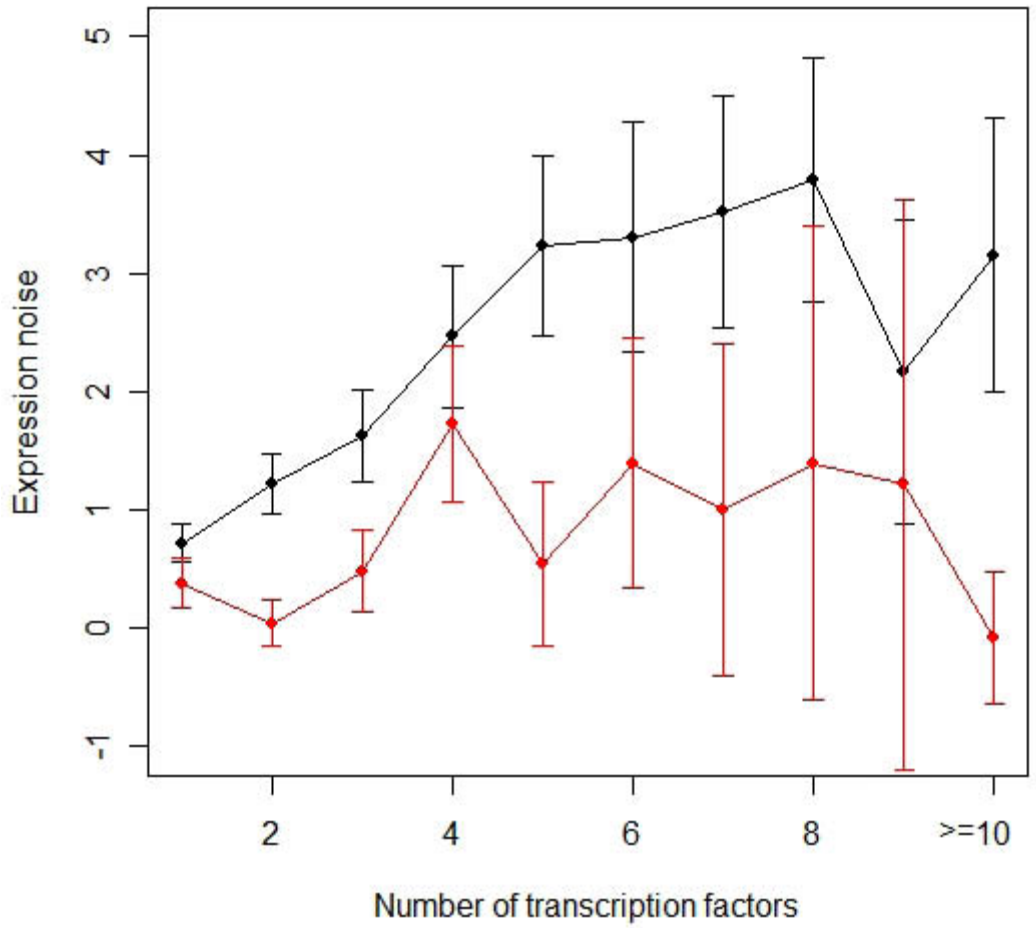


Figure 2-6 Controlling for the number of TFs, bi-promoter genes show lower expression noise than other genes. Bi-promoter genes, red; other genes, black. Significance determined by randomization with the sum difference (non-modular) between the means for each increment on the x-axis being the reporting statistic (p<0.0001). Randomizations preserved for each gene the number of transcription factors and randomized noise levels between genes with the same number of regulating transcription factors.

The existence of a TATA-box appears to be linked to increased noise levels (Raser, O'Shea 2004; Blake *et al*. 2006; Murphy, Balazsi, Collins 2007; Choi,

Kim 2008; Field *et a l*. 2008). As bi-directional genes in both human and *Drosophila melanogaster* (Trinklein *et al* . 2004; Yang, Yu 2009) often lack TATA control, the result could reflect TATA presence/absence rather than bidirectionality *per se* . In yeast, we find the same bias: of the 2111 protein coding genes involved in bi-promoter pairs, only 509 are annotated as containing a TATA-box, which is significantly lower compared to other genes (p < 2.2e-16, Fisher's Exact Test).

Table 2-2 The low noise of bi-promoter genes cannot be explained by TATA boxes. Noise levels of bi-promoter genes are significantly lower than those of other genes, both in genes with TATA box containing promoters in TATA-less genes. Mean noise+/-standard error (number of genes*)*. *p*- values from Brunner-Munzel tests.

	TATA box-containing genes	TATA-less genes
Bi-promoter genes	1.01+/-0.25 (509)	0.13+/-0.11 (1602)
Non-bi-promoter genes	2.71+/-0.27 (1587)	0.70+/-0.12 (2902)
p-value	2.1e-06	0.0013

We thus compared the noise of bi-promoter TATA-containing genes with that of non-bi-promoter TATA-containing genes, and the noise of bi-promoter TATA-less genes with that of non-bi-promoter TATA-less genes. As expected, TATA is a predictor of noise (e.g. in bi-promoter genes, genes with a TATA-box show higher noise levels than genes without a TATA-box, p = 0.0064, Brunner-Munzel test). However, this fails to explain the low noise of bi-promoter genes: bi-promoter genes have lower noise than non-bi-promoter genes even when only considering those genes without a TATA-box; the same holds when considering only genes with a TATA-box (Table 2-2).

2.4.8 Type II promoters already are nucleosome free and so don't benefit from bidirectional architecture.

There are two types of promoter regions: those that favour nucleosomes, and those that don't (Field *et al* . 2008; Tirosh, Barkai 2008). Genes with nucleosome-favoring promoters usually have high expression noise, while genes with nucleosome disfavoring promoters usually have low expression noise

(Field *et al*. 2008). How does this relate to gene orientation?

We utilized a prior definition of type I and type II promoters (Field *et al*. 2008). Here a type I promoter is defined as a promoter containing a TATA-box with at least 80% of the length of its binding sites covered by nucleosomes. A type II promoter is TATA-less with at most 20% of the total length of its binding sites covered by nucleosomes. We find that non-bi-promoter genes have higher noise than bi-promoter genes when restricting our analysis to nucleosome-favouring promoters (>80% occupancy; mean noise =1.81 in bi-promoter genes, noise=5.46 in other genes, p = 0.00020, Brunner-Munzel test; Table 2-3). This remains true after controlling for gene essentiality (Table 2-3).

Table 2-3 Nucleosome favouring bi-promoter genes have lower noise than nucleosome favouring non-bi-promoter genes. In the control for essentiality we just examine the non-essentials. p value determined by the Brunner-Munzel test.

	nucleosome favouring	nucleosome disfavouring
bi-promoter	1.81 +/- 0.66 (103)	0.09 +/- 0.30 (233)
non bi-promoter	5.463+/- 0.698 (328)	0.10+/- 0.17 (331)
p	0.00020	0.16

control for essentiality	nucleosome favoured	nucleosome disfavoured
bi-promoter	2.491 +/- 0.829 (87)	-0.2801 +/- 0.2018 (162)
non bi-promoter	6.1434 +/- 0.756 (301)	0.188 +/- 0.203 (241)
p	0.0028	0.065

By contrast, for genes with nucleosome-disfavouring promoters (occupancy <20%), we see no evidence for a noise reduction through bi-promoter architecture (Table 3). If Seila *et al* (Seila *et al*. 2009) are correct this result is to be expected. They conjecture that RNAPII complexes are simultaneously engaged at the boundaries of the nucleosome-depleted region surrounding TSSs and that these divergently engaged polymerases could directly reinforce the -1 and +1 nucleosome positions, effectively enhancing the boundaries of the nucleosome-free region, allowing transcription factors access to the promoter

(Seila *et al*. 2009), and maybe further maintaining the "loose" chromatin during transcription (Preker *et al* . 2009) Such genes are, in effect, primed for transcription, regardless of orientation: an 'interrupted form' of bi-directional transcription occurs even if there is no bi-promoter. For those bi-promoter pairs that do not exclude nucleosomes in this manner from the bi-promoter region during transcription (type I pairs), dependence between the two genes is re-inforced and noise reduced, much as we modelled. If the above picture is true, we would expect that the class II (nucleosome-free) genes in non-bidirectional orientation should have lower noise than class I genes in the same orientation, which indeed we observe (mean noise level is 0.10 +/- 0.17 and 5.46 +/- 0.70, respectively. p < 2.2e-16, Brunner-Munzel test; this remains true when controlling for gene essentiality). In short, nucleosome depletion and bidirectional orientation we suggest to be two alternative mechanisms to ensure low noise by resisting stochastic chromatin closure.

2.4.9 Only bi-promoter genes show correlated noise of neighbours

For any gene we can assay its noise level under a variety of parameter values. The simulation suggests that when two genes are coupled (i->0) the noise levels of the two proteins across these multiple conditions are correlated. More generally, across all simulations we consider the correlation in protein noise between the neighbours for a given value of independence i. We find this to be strongest when coupling is strongest (r=-0.96). Our simulations thus predict that the correlation in noise levels between neighbours should be strongest when coupling is strongest and hence when genes are divergent. If independence of divergent genes is in turn modulated by intergene distance, by the same logic we expect for divergent genes the correlation in noise levels to be higher when intergene distance is lower.

Confirming these predictions, we find a significant correlation of the noise of two divergent transcripts. Conversely, neither convergent nor co-oriented gene pairs show correlated noise levels (Spearman rank correlation for divergent pairs r = 0.148, p = 0.031 (r = 0.151, p = 0.047 after removing type II genes); for convergent pairs r = 0.0089, p = 0.45; for co-oriented pairs r = -0.0008, p =

0.51; *p*-values determined by randomization). Also as predicted the mean noise level of the transcripts in divergent gene pairs is correlated with the distance between transcription start sites, a correlation not seen for convergent and co-oriented pairs (Spearman rank correlation for divergent pairs $r = 0.0936$, $p = 0.0055$; for convergent pairs $r = -0.0194$, $p = 0.49$; for co-oriented pairs $r = -0.0282$, $p = 0.29$).

2.4.10 Essential genes tend to be low noise with bi-promoter architecture, while the opposite is seen for stress response genes.

Of all genes, those that are lethal on knockout (i.e., essential) are most likely to be under selection for reduced noise levels (Fraser *et al*. 2004). Conversely, stress related genes are thought to be under selection for high noise (Lopez-Maury, Marguerat, Bahler 2008). Many features of essential genes are consistent with low noise. They tend to be highly expressed, but even controlling for this they have low noise (Newman *et al*. 2006; Batada, Hurst 2007). Counter-intuitively for highly expressed genes the mRNAs have short half lives (Pal, Papp, Hurst 2001), a feature consistent with low noise (Fraser *et al*. 2004). They tend not to be TATA controlled and reside clustered in genomic low noise/open chromatin domains (Batada, Hurst 2007).

If bi-promoter architecture is a mechanism to enable low noise and expression when needed, we might also expect such genes to be in divergent or bi-promoter orientation more than expected by chance. This is indeed the case in yeast. Of 6600 protein coding genes in yeast, 2627 are divergent with a partner protein coding gene. Of these, 537 (20.4%) are essential, while only 577 (14.5%) of the 3973 non-divergent genes are essential. There is thus enrichment of essential genes in the divergent class ($p = 4.9 \times 10^{-10}$, Fisher's exact test). There is a corresponding enrichment of essential genes in gene pairs with bi-promoter architecture. Of 2111 genes in bi-promoter organization, 22% are essential, while only 649 of 4489 (14.4%) non-bi promoter genes are essential ($p=5.9 \times 10^{-13}$, Fisher's exact test). An analogous excess in divergent orientation has recently been reported in Drosophila (Yang, Yu 2009). Moreover, we see more bidirectional pairs of two essentials genes than expected

by chance: there are 79 bidirectional essential gene pairs in yeast, this being more than ever found in 1000 gene order randomizations, $p<0.001$). Also as expected, haploinsufficent genes tend to be in bipromoter architecture more than expected (41% versus 31% of all others; $p=0.005$).

For stress-related genes, where we expect selection for high noise, we see the opposite pattern. While those that are bi-promoter have lower noise than stress related genes in different configurations (mean noise for bi-promoter stress genes 1.59 +/- 0.30, for non-bi-promoter stress genes 3.63 +/- 0.27, $p=1.6 \times 10^{-8}$, Brunner-Munzel test), stress related genes tend to avoid having a bi-promoter architecture. Only 509 (24.1%) bi-promoter genes are stress related, while 1525 (34.0%) of non-bi-promoter genes are stress related (Fisher's exact test, $p = 2.7 \times 10^{-16}$). Similarly, stress genes tend not to be in divergent orientation (28% divergent, 32.5% non-divergent; $p = 0.00024$, Fisher's exact test).

What of the essential genes that are not bi-promoter with another protein coding gene? We predict to see more cases of antisense ncRNA than expected by chance associated with such genes, if ncRNA is a mechanism of noise reduction. This we observe. Of 309 genes with an antisense CUT, 65 (21%) are essential genes, while only 624 (14.1%) of 4441 genes without an antisense CUT are essential ($p = 0.0014$, Fisher's exact test).

If there are peculiar features of essential genes (e.g. short half life, low usage of optimal codons), can we exclude the possibility that bi-promoter genes have low noise just because of this enrichment for essential genes? Mean noise level of the 1646 non-essential bi-promoter genes is significant lower than other non-essential genes (0.43+/- 0.12 versus 1.83+/- 0.15, $p= 5.5 \times 10^{-12}$ in Brunner-Munzel test). That non-essential genes with bi-promoter control have lower expression noise than essential genes (in all orientations) ($p = 0.035$) further suggests that dispensability cannot alone account for the low noise of bi-promoter genes.

There must, however, be alternative methods to modulate noise. Notably, we find that the mean noise of bi-promoter essential genes (with either an ncRNA

or a protein coding gene partner) is not significantly lower than the noise of non-bi-promoter essential genes (0.18+/-0.22 versus 0.22+/- 0.26, p=0.82 in Brunner-Munzel test; 0.03+/-0.20 versus 0.29+/-0.30, p = 0.76 after removing type II genes). These results are then consistent with bi-promoter architecture being a means to reduce noise, but, unsurprisingly, not the only mechanism.

What the other mechanisms might be is not immediately transparent. For example, while essential genes have a shorter mRNA half life than non-essential genes (p = 2.8x 10^{-16}, Brunner-Munzel test), the mean mRNA half life for bi-promoter essential genes is no different to that of non-bi-promoter essential gene (16.65 versus 16.91 respectively: p = 0.25, Brunner-Munzel test). Increased usage of codons that specify abundant tRNAs is expected to enable fast translation and be associated with high noise. As expected, there is a positive correlation between the frequency of optimal codon usage (FOP) and expression noise in yeast (r = 0.107, p = 4.6 x 10^{-07}, Spearman's rank correlation). However, FOP of bi-promoter essential genes does not differ from that of either essential non-bi-promoter genes or essential non-divergent genes (p = 0.16 and 0.63, respectively, Brunner-Munzel tests).

2.4.11 Bi-promoter gene pairs and CUTs are rare in noisy subtelomeric domains

Does the fact that bi-promoter gene pairs have low noise affect not only which sort of genes are found in this architecture but also where on chromosomes they are found? Previously it was reported that essential genes and non-essential genes flanked by a high density of essential genes tend to have low noise (Batada, Hurst 2007). Could it be that non-bi-promoter essential genes tend to reside in essential gene clusters, thus giving them low noise? Alternatively might genes requiring low noise not only adopt bi-promoter architecture but also aggregate into low noise chromosomal domains? Ignoring genes +1 and -1 from a focal essential gene (direct neighbours) and then asking about the number of essential genes in the flanking 5 genes on either side, we find that both bi-promoter essential genes (p=0.022) and bi-promoter non-essential genes (p=0.018) have more essential genes in their vicinity than expected by chance

(Table 2-4). Thus bi-promoter genes tend to be enriched in the vicinity of essential gene clusters, these having unusually low noise levels (Batada, Hurst 2007). Clustering of bipromoter genes doesn't however fully account for the low noise of genes in such domains. Examining non-bipromoter genes, those in essential gene clusters have lower noise than those not in clusters (P=0.0007; controlling for essentiality, P=0.01).

Table 2-4 The density of essential genes among the 10 genes flanking focal genes. Here we ignore genes +1 and -1 of a focal gene (direct neighbours).

Bi-promoter		**Not bi-promoter**	***p*-value**
Essential	0.212 +/- 0.006	0.195 +/- 0.005	0.022
Not essential	0.188 +/- 0.003	0.180 +/- 0.003	0.018
p-value	0.00089	0.010	

Yeast subtelomeric domains are high-noise domains and are depauperate in essential genes (Batada, Hurst 2007). From the logic that bi-promoter architecture is a genomic device to minimize noise, we might expect that genes found in subtelomeric domains should be favoured to be high noise genes and hence not in a bi-promoter architecture. Considering all genes, 28 of 324 gene pairs (8.6%) are bi-promoter in subtelomeric domains (20kb from chromosome ends), while 2083 of 6276 (33%) non-subtelomerics are bi-promoter ($p<2.2 \times 10^{-16}$, Fisher's exact test). However, as essential genes tend to be bi-promoter and avoid subtelomeric domains, we may be seeing nothing more than the biased distribution of essential genes. Considering only non-essential genes, we see the same bias (8% subtelomeric non-essential genes in bi-promoter architecture versus 31% non-subtelomeric, $p < 2.2 \times 10^{-16}$, Fisher's exact test). We similarly find that bi-promoter CUT associated genes are rare subtelomerically (1.2% subtelomeric genes have a bi-promoter CUT compared with 4.8% otherwise, p=0.001 Fisher's exact test; this remains when controlling for essentiality of the neighbour, p=0.006). The high noise of subtelomeric genes and the avoidance of subtelomeric domains by bipromoter genes cannot explain the low noise of bipromoter genes, as they have low noise even compared with genes that are not subtelomeric ($P<10^{-11}$).

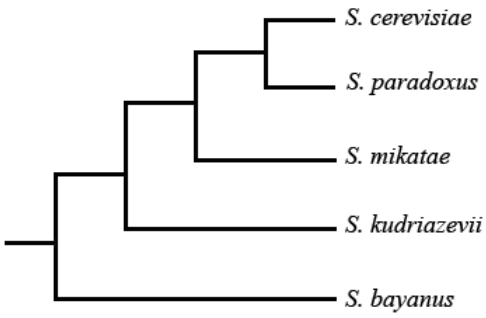


Figure 2-7 The phylogenetic relationships of yeasts used in the comparison[7].

2.4.12 Non-coding RNAs evolve fast in yeast, suggesting the sequence *per se* is not functional

Do non-coding transcripts show obvious conservation signals? If the transcribed sequence is functional, we expect that it is conserved across different evolutionary lineages. The recent literature however reports that many non-coding transcripts do not show detectable signs of evolutionary conservation. For example, Liang and Li (Liang, Li 2009) studied 383 human microRNAs and found that about 30% of them lack sequence conservation signals and appear almost free of selection pressures. Using recently released yeast population genetics data, we calculated the variation of non-coding sequences both within *S. cerevisiae* and between yeast species (see Figure 2-7). We then compared these measures to those of the (presumably non-functional) downstream 100bp regions. If the sequence divergence of non-coding transcripts is significantly lower than the downstream regions, then this indicates evolutionary conservation (Liang, Li 2009). As seen from Figure 2-8, the difference between ncRNA and downstream region is not significant in any of the 5 comparisons (p>0.05 in each case). Thus, non-coding RNAs and downstream untranscribed sequences show very similar diversity and divergence. What is more, this appears to be equally true for various subsets of non-coding RNAs (for example, in stable annotated transcripts and unstable annotated transcripts; data not shown).

[7] http://www.broadinstitute.org/science/projects/fungal-genome-initiative/fungal-genome-initiative

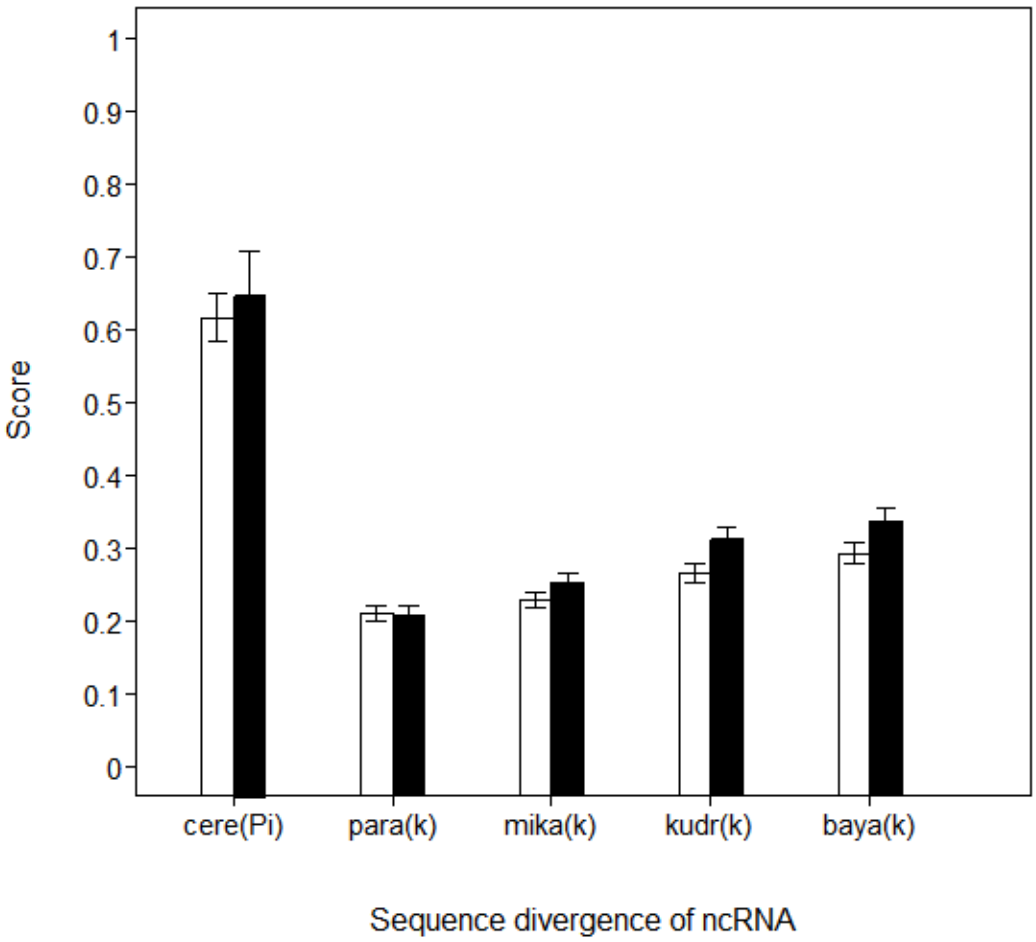


Figure 2-8 Sequence diversity (Pi) and divergence (K) of non-coding RNAs (white bars), compared with non-transcribed un-annotated downstream 100 bp seqences (black bars). x-axis labels: cere = diversity within *S. cerevisiae;* para, mika, kudr, baya = divergence between *S. cerevisiae* and *S. paradoxus, S. miyake, S. kurowski,* and *S. bayanus,* respectively.

2.5 Discussion

We have found, via simulation, that if transcription of one gene increases the probability of transcription of a neighbour and vice versa, then low noise of both is expected across broad and realistic parameter space. We propose that divergent gene pairs, bi-promoter gene pairs in particular, are thus expected to be low noise genes, even allowing for any effect on protein abundance. This model has striking predictive ability. Bi-promoter genes are indeed low noise

and, as predicted, the noise is modulated by intergene distance. Similarly, bipromoter pairs have correlated noise. The model can predict biases both in which genes are or are not in bipromoter architecture (essential/complex genes and stress response genes respectively) and which classes of gene should be more likely to have ncRNA in bi-promoter architecture. Indeed, that our model can predict noise levels and skew in gene type associated with CUTs, strengthens the view that noise control, independent of co-expression modulation, is a focus of selection. The model also predicts that bipromoter pairs should be rare subtelomerically as observed, such domains being high noise domains.

These results suggest that gene orientation may well be an important feature in the control of noise, they also suggest that, as with transcription at SER3 (Martens, Laprade, Winston 2004), it is the act of transcription, rather than the product of transcription, that can be important. While the CUT associated with SER3 (a sense transcript) is associated with control of the expression of the downstream gene, we argue that transcription from the opposing strand is an effective mechanism for priming a focal sense strand gene for expression and hence for reduction in noise. The transcript may well be unwanted, but it doesn't follow that the making of the transcript is without functional relevance. This is also supported by the observation that upstream RNA PolII transcripts usually cannot be elongated effectively (Core, Waterfall, Lis 2008; Seila *et al*. 2009).

We might then also wonder how much expression in protein coding genes from bidirectional promoters is to enable noise control rather than produce the protein product itself. Such a hypothesis could explain why many relatively highly co-expressed neighbours ($0.4>r>0.2$) in yeast have no functional (GO class) similarity (Batada, Urrutia, Hurst 2007).

These findings add to recent evidence that a substantial component of selection on gene arrangement within genomes is to modulate noise levels. In yeast the clustering of essential genes may be owing to such selection (see also (Keller, Knop 2009)). In bacteria co-linearity, the tendency for genes to appear in the

same order in the operon as the proteins are needed in a temporal fashion, appears also best explained by the consequences of selection on noise (Kovacs, Hurst, Papp 2009; Lovdok *et al* . 2009). What remains to be resolved is whether noise modulation mediated by changes in gene order/orientation is relevant in less compact genomes, such as those of mammals.

Chapter 3 Co-expression of Linked Gene Pairs Persists Long After Their Separation

3.1 Project summary

In many organisms, physically linked gene pairs tend to be co-expressed. While co-expressed gene clusters appear to be opposed by natural selection in mammals, they seem to be under stabilizing selection in yeast. Here, we analysed expression patterns of gene pairs that have lost their linkage in the evolution of *S. cerevisiae* since its last common ancestor with *K. waltii*. We demonstrate that co-expression of linked genes is retained long after their separation, and is thus likely to be functionally important. Contrary to previous suggestions, functional co-expression is not restricted to bi-directional promoters, and cannot be explained by gene essentiality alone.

3.2 Introduction

A gene's expression pattern is influenced by its genomic location, both in prokaryotes and eukaryotes. In prokaryotes, neighbouring genes often form operons, resulting in tight co-expression of neighbouring genes. In eukaryotes, physically linked gene pairs also show higher co-expression than randomly chosen gene pairs (Cohen *et al*. 2000; Kruglyak, Tang 2000; Lercher, Urrutia, Hurst 2002; Williams, Hurst 2002; Lercher, Blumenthal, Hurst 2003; Hurst, Pal, Lercher 2004; Singer *et al*. 2005; Lercher, Hurst 2006; Semon, Duret 2006; Batada, Urrutia, Hurst 2007; Kensche *et al*. 2008). For example, in yeast, adjacent gene pairs show correlated expression regardless of their relative orientation (Cohen *et al*. 2000; Kruglyak, Tang 2000), and this co-expression relationship spans up to 30 neighbouring genes (Lercher, Hurst 2006). In the worm *Caenorhabditis elegans,* many genes are organized into operons (Lercher,

Blumenthal, Hurst 2003). In the mouse genome, both immune system genes and tissue specific genes are found to be expressed in clusters (Williams, Hurst 2002). In the human genome, housekeeping genes also show strong clustering (Lercher, Urrutia, Hurst 2002). Such co-expression clusters may be selectively favourable in mammals, as they are usually maintained during evolution (Singer *et al*. 2005).

The co-expression of neighbouring genes in prokaryotic operons is conceptually simple. In eukaryotes, many mechanisms have been proposed to be responsible for the co-expression of closely spaced genes. Neighbouring genes with similar functions have lead to the proposal that the co-expression of linked genes may be related to gene function (Cohen *et al*. 2000; Michalak 2008). Neighbouring genes that have divergent orientation suggest that bi-directionally active promoters are often responsible for the co-expression of divergently transcribed neighbouring genes (Cohen *et al*. 2000; Kruglyak, Tang 2000; Kensche *et al*. 2008). Chromatin structure also likely has an impact on the co-expression of closely located genes (Hurst, Pal, Lercher 2004; Batada, Urrutia, Hurst 2007); indeed, chromatin remodeling is a major source for the co-expression of linked genes (Batada, Urrutia, Hurst 2007). Finally, gene pairs that share the same transcription factors, and gene pairs that may be prone to a failure of transcription termination ('transcriptional read-through') were also reported to be responsible for the co-expression of neighbouring genes (Semon, Duret 2006; Batada, Urrutia, Hurst 2007; Michalak 2008).

Thus, it appears clear that neighbouring genes tend to be co-expressed. But is this co-expression really selectively favourable, or is it solely a mechanistic by-product of genomic neighbourhood? If neighbour co-expression is indeed functional, then co-expression should be maintained even if the neighbourhood is broken up by genomic rearrangements. In this chapter, we compare the effects of current and ancestral gene order on the current gene expression patterns. We show that gene pairs which were neighbours in the evolutionary past, but are separated now, also show higher co-expression than randomly chosen gene pairs. This indicates a significant role of natural selection in the co-expression of linked yeast genes.

3.3 Materials and methods

3.3.1 Data source

The *Saccharomyces cerevisiae* gene order as well as the ancestral gene order were taken from the yeast gene order database (Byrne, Wolfe 2005). We only retained genes with known positions in both data sets for further analysis. For the genes that retained duplicated copies in the ancestor genome, both of the two copies are taken into account for the calculation of co-expression level of ancestor neighbouring gene pair. *S. cerevisiae* genome sequence data was downloaded via ftp from the SGD database[8]. Ancestral gene order from 8 reconstructed chromosomes were taken from(Gordon, Byrne, Wolfe 2009) and the information of gain and lose of neighbourhood of gene pairs in yeast phylogeny were taken from (Kensche *et al*. 2008).

3.3.2 Expression data

Combined data from 40 time-series microarray experiments was used to constructed the 5883 gene pair co-expression relationships (Kafri, Bar-Even, Pilpel 2005). For each two genes, we calculate the Pearson correlation coefficient of their expression levels across experiments and time points; missing values were omitted. For the ancestor genes that have two copies in the current genome, the mean correlation values of the two paralogs were used.

3.3.3 Dollo parsimony method to calculate ancestor state of each gene pair

Phylogenetic relationship of 19 yeasts and the neighbouring gene pairs of orthologs in these fungi were downloaded from the supplementary file of (Kensche *et al*. 2008). Then we used the Dollo parsimony method from PAUP* to calculate the ancestor state of each gene pair (Wilgenbusch, Swofford 2003). That method provides us the gain and loss information of each node.

[8] ftp://genome-ftp.stanford.edu/pub/yeast/

3.4 Results

3.4.1 Gene order comparison between the reconstructed ancestor and the current *S. cerevisiae* genome

We used the recently reconstructed gene order of the pre-whole genome duplication (pre-WGD) yeast ancestor (Gordon, Byrne, Wolfe 2009), which is believed to be about 100~150 million years old (Sugino, Innan 2005). We compared the co-expression of the gene pairs which are conserved between the two genomes to the co-expression of gene pairs newly formed in *S. cerevisiae*.

Three possible scenarios exist: (i) If the conserved gene pairs are less likely to be co-expressed compared to newly formed gene pairs, then highly co-expressed neighbouring gene pairs may be generally disadvantageous, as was observed recently in mammals (Liao, Zhang 2008). (ii) If conserved gene pairs share similar co-expression profiles with newly formed gene pairs, then neighbour co-expression is likely to be largely selectively neutral. (iii) If the conserved gene pairs generally show higher co-expression levels compared to newly formed gene pairs, then this suggests that neighbour co-expression is generally advantageous, as previously suggested (Singer *et al*. 2005; Semon, Duret 2006).

Table 3-1 shows the results of this comparison. For divergently oriented *S. cerevisiae* gene pairs (<- ->), those that were already in this orientation in the ancestral genome show higher co-expression compared to newly formed divergent gene pairs. No such difference between conserved and new pairs was found for convergent or co-oriented gene pairs. This indicates that in yeast, only divergent gene pairs are under selection for high co-expression. Surprisingly, there is no difference between the co-expression of newly formed divergent gene pairs and convergent gene pairs (p = 0.59 comparing new divergent gene pairs with conserved convergent gene pairs, and p = 0.59 comparing new divergent gene pairs with newly formed convergent gene pairs, Brunner-Munzel tests). Thus, divergent gene pairs do not always show higher co-expression compared to other types of gene pairs in yeast.

Table 3-1 Only divergent gene pairs show higher co-expression in ancient compared to new neighbours (*p*-values from Brunner-Munzel tests)

	Neigbhours in ancestor & *S. cerevisiae*		Neighbours in *S. cerevisiae only*		
	Mean	*N*	mean	*N*	*p*
divergent	0.14	738	0.11	502	0.00064
convergent	0.11	708	0.11	561	0.95
co-oriented	0.084	1140	0.084	1090	0.85

The results further show that highly co-expressed linked gene pairs tend to be evolutionarily conserved only for divergent gene pairs. This is likely related to the activity of bi-directionally active promoters ('bipromoters') (Kruglyak, Tang 2000). Consistent with this hypothesis, we found that 454 out of 638 bipromoter gene pairs (71%) were already present in the ancestral genome, while the same is true for only 47% of the non-bipromoter divergent gene pairs. More importantly, there is no difference between the co-expression level of conserved bipromoter gene pairs and new bipromoter gene pairs. Conversely, conserved non-bipromoter gene pairs show higher co-expression compared to newly formed non-bipromoter gene pairs.

These results have two important implications. On one hand, they suggest that co-expression *per se* cannot explain the conservation of bipromoter structures; this result is consistent with our hypothesis of a role of bipromoters in regulating expression noise (see Chapter 2). On the other hand, the results indicate that there is a selective advantage for the retention of bi-promoter structure.

For the 2765 ancestrally neighbouring gene pairs that are separated in the current genome, there is also a significant difference between their co-expression level and 10000 randomly chosen gene pairs (p=5.3x10^{-6}, Wilcoxon rank sum test). That result suggests that the high co-expression profile of linked gene in the ancestor genome cannot solely be explained by the gene pairs neighbouring in the current genome. but we do not find a significant

differences between divergent, convergent and cooriented gene pairs ($p = 0.084$ between divergent and convergent pairs, $p = 0.13$ between divergent and cooriented gene pairs and $p = 0.70$ between cooriented and convergent pairs, *p*-values from Brunner-Munzel tests), that could potentially be explained by the hypothesis that chromatin remodeling account for most of the low level gene co-expression of linked gene pairs(Batada, Urrutia, Hurst 2007) (after separation, the chromatin effect disappears).

3.4.2 Dollo parsimony method results

The results presented above suggest that at least part of the high co-expression level of neighboring yeast gene pairs is due to natural selection. Thus, genes still need to be co-expressed when pairs are separated through a genomic rearrangement. To further verify this prediction, we used recent data based on gene pair conservation across 19 fungi (Kensche *et al*. 2008). We reconstructed the gene order in the common ancestor of these species, using Dollo parsimony as implemented in PAUP* (Wilgenbusch, Swofford 2003).

We only analysed genes that were direct neighbours in the ancestral genome, but that are now located on different chromosomes, because genes located nearby on a yeast chromosome still show similar expression profiles even when separated by tens of genes (Lercher, Hurst 2006).

As predicted, separated gene pairs still show slightly higher co-expression compared to random gene pairs (Figure 3-1; $p = 3.1 \times 10^{-6}$, Brunner-Munzel Test). Again, there is no difference between the co-expression of divergent, convergent and co-oriented ancestral gene pairs after their separation ($p = 0.18$ between divergent and convergent pairs, $p = 0.32$ between divergent and co-oriented gene pairs, and $p = 0.76$ between co-oriented and convergent pairs; Brunner-Munzel tests).

If gene neighbourhood is under positive selection for genes that need to be co-expressed, then we would further expect that orthologs of co-expressed *S. cerevisiae* genes are more likely to be genomic neighbours in other yeast

species; consequently, genes neighbouring in at least one other yeast species but neither in *S. cerevis iae* nor in the common ancestor should show higher co-expression than random gene pairs. This is indeed the case (p = 1.2x10^{-5}, Brunner-Munzel test).

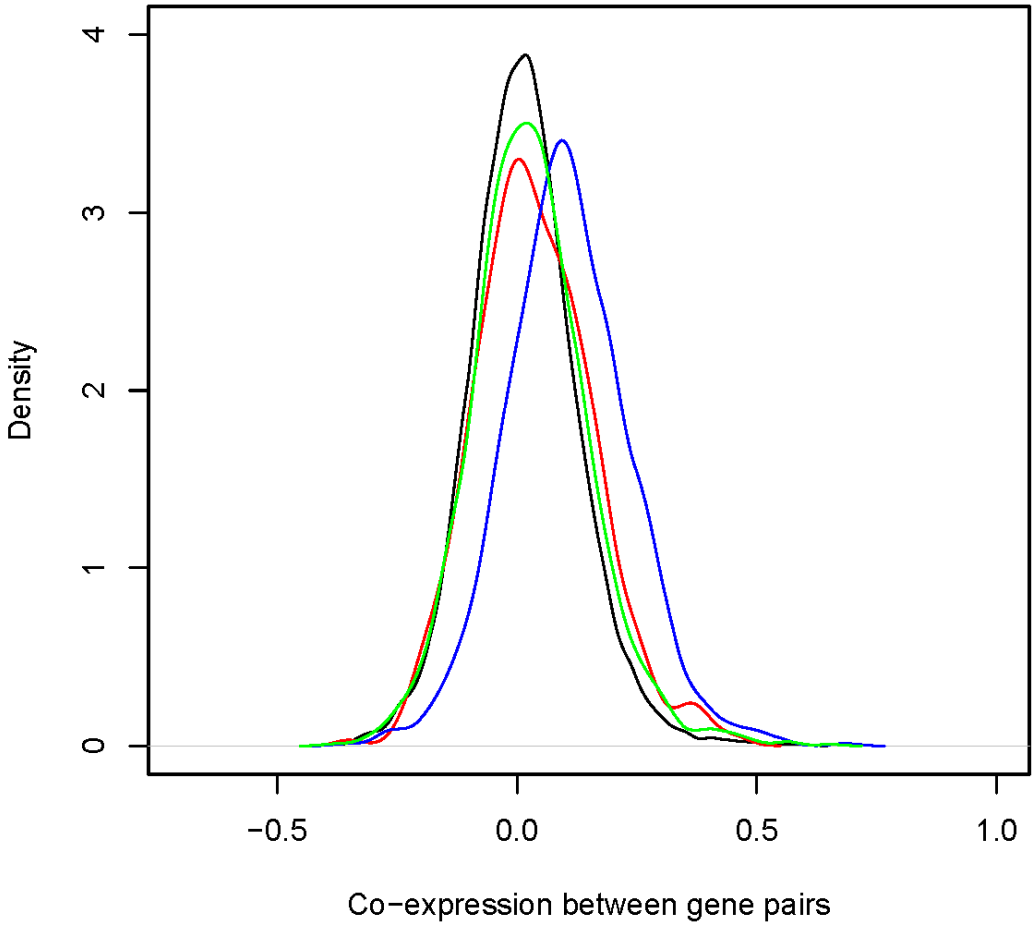


Figure 3-1 Co-expression of gene pairs neighbouring in the ancestral or current genome. Black, 10000 random gene pairs; red, pairs neighbouring in the ancestor but separated in *S. cer evisiae*; blue, pairs only neighbouring in *S.cerevisiae*; green, gene pairs only neighbouring in other yeast species.

3.5 Discussion

Using ancestral gene order information gained from the yeast gene order browser, we confirmed that among neighbouring gene pairs, divergently oriented pairs are the ones that were most likely to be conserved during genome

evolution (Kensche *et al* . 2008). This conservation implicates stabilising selection on the relative positioning of a subset of the divergently arranged gene pairs, most likely because of bi-directional promoters. However, after separation of neighbouring gene pairs through genomic rearrangements, we no longer found any difference between divergent and convergent or co-oriented gene pairs; all three types of ancestrally neighbouring gene pairs show higher than expected co-expression in *S. cerevisiae* after their separation through genomic rearrangements. It is possible that the two genes in these co-expressed separated pairs had part of their *cis*-regulatory apparatus in common even before their separation, so that co-expression could be partially maintained after the rearrangement; conversely, it may be that co-expression was initially lost in the rearrangement, and was re-instated through *cis*-regulatory changes afterwards.

The co-expression of ancestrally neighbouring gene pairs that are now located on different chromosomes is sharply reduced compared to the pairs that are neighbours in the current yeast genome. This observation is expected, as factors such as chromatin remodelling are known to strongly influence the co-expression of linked genes in yeast (Batada, Urrutia, Hurst 2007). Thus, while part of the neighbour co-expression is likely maintained by natural selection, it is likely that a substantial component of neighbour co-expression is purely mechanistic.

When discussing the properties of neighbouring gene pairs, these are usually classified by their relative orientation into three categories – divergent gene pairs (head to head), convergent gene pairs (tail to tail), and co-oriented gene pairs. Those three types of gene pairs appear to have different properties – divergent gene pairs are the most conserved gene pairs and show stronger co-expression than the other two orientations (Kensche *et al*. 2008). Here we show that as far as co-expression is concerned, there are essentially only two types of neighbouring gene pairs in the genome – bipromoter gene pairs and non-bipromoter gene pairs. Bipromoter gene pairs show strong signals of conservation and co-expression, while non-bipromoter gene pairs don't. After separation through genomic rearrangements, ancestral divergent gene pairs no longer exhibit higher co-expression compared to other gene pairs, supporting

the view that chromatin remodeling dominated the co-expression of most neighbouring gene pairs (Batada, Urrutia, Hurst 2007).

In conclusion, we have shown that not only gene neighbourhood in the current yeast genome, but also gene order in the ancestral genome is predictive of co-expression. This conservation of co-expression is evidence in favour of a role for natural selection in the establishment and maintenance of neighbour co-expression in yeast, and argues against a purely mechanistic view that considers neighbour co-expression as a purely neutral (or even slightly deleterious) phenomenon.

Chapter 4 The effects of Network Neighbours on Protein Evolution

4.1 Project summary

Interacting proteins may experience similar selection pressures. Thus, we may expect that neighbouring proteins in biological interaction networks evolve at similar rates. This is indeed the case for co-expression network, protein-protein interaction network, metabolic network and genetic interaction networks. However, the strongest known predictor of the rate of protein evolution remains expression level. Hence, when testing for network effects, expression has to be incorporated into the Null model against which we compare our results. We found that similar expression levels of neighbours indeed explain their similar evolution rates in protein-protein and metabolic networks. In co-expression network, on the other hand, neighbouring genes still show similar evolutionary rates even after controlling for expression level, gene essentiality and gene length. This suggests that both expression level and co-expression shape the rate of protein evolution in networks.

4.2 Introduction

Recently, there has been increased interest in the influence of biological networks on protein evolution. Network connectivity, i.e., the number of connections that an individual protein has, was the first parameter reported to influence protein evolution (Fraser *et al*. 2002; Fraser, Wall, Hirsh 2003; Jordan, Wolf, Koonin 2003; Saeed, Deane 2006; Vitkup, Kharchenko, Wagner 2006). A negative correlation between connectivity and evolutionary rate is observed not only in protein-protein interaction networks (Fraser *et al*. 2002; Carlson *et al*. 2006), but also in metabolic networks (Vitkup, Kharchenko, Wagner 2006) and genetic interaction networks (Costanzo *et al*. 2010). Another network parameter, betweenness, which is an measure of network centrality, was found to correlated

with evolution rate in protein-protein interaction networks: proteins with high betweenness are more likely to be essential genes and thus evolve more slowly (Joy *et al*. 2005). In the metabolic network of yeast, centrality is also a predictor of evolution rate of enzymes (Vitkup, Kharchenko, Wagner 2006). In contrast, transcription factors that are more central in the regulatory network were shown to evolve faster than other genes (Jovelin, Phillips 2009), indicating that the transcription network has dramatically different properties compared to other biological networks.

Are there other features in the network that influences protein evolutionary rate? Here we studied the relationship between the evolutionary rate of a protein and the evolutionary rate of its network neighbours. It has been reported that in the protein-protein interaction network, interacting proteins tend to have similar evolution rates (Fraser *et al*. 2002). But we do not know if this conclusion holds generally true for all types of biological networks.

If the protein and its network partners co-evolve or co-adapt (Juan, Pazos, Valencia 2008), we indeed expect that the proteins show similar rates of evolution. For example, in the protein-protein interaction network, interacting binding sites usually show co-evolution (Kann *et al.* 2009). Physically interacting human proteins (protein protein interaction network) show stronger signs of co-evolution than proteins in the same biochemical pathway (metabolic network) (Tillier, Charlebois 2009). In co-expression networks, neighbouring genes are often involved in the same biological function, and in genetic interaction networks, the mutation of one protein changes the fitness effects of mutations in its partners. By comparing the number of substitutions per site between interacting proteins, we tested this hypothesis in the yeast protein-protein interaction network, co-expression network, metabolic network, genetic interaction network, and transcriptional regulatory network.

There is an ongoing debate if the co-evolution of interacting proteins is caused by compensatory mutations between binding partners, or if it is simply due to similar selective constraints, like those resulting from similar expression levels. An investigation of the three-dimensional structures of about 100 yeast proteins

indicated that buried residues – which are located on a stable interaction surface between protein units – are under stronger evolutionary constraints than solvent exposed sites (Lin *et al*. 2007b), even after excluding the effect of expression level. Moreover, residues close to the binding sites responsible for protein-protein interactions show higher co-evolution signals than residues outside the binding region (Kann *et al* . 2009). However, another analysis observed that correlations purely based on the co-evolution of proteins surfaces and binding interfaces are not higher than the correlation when considering the whole interacting proteins (Hakes *et al*. 2007).

From an analysis of the evolution rate of each focal protein in the network and the mean rate of its neighbours, we show that there is indeed a positive correlation in most biological networks. Further, we find that the correlation can be explained by shared evolutionary constraints, in particular related to similar expression levels. These results support the view that the co-evolution of binding sites or functional similarity plays only a minor role in determining network effects on protein evolution. Finally, our results suggest that co-expression relationships are another factor that influences evolutionary rate on the level of biological networks.

4.3 Materials and methods

4.3.1 Evolutionary rates

The evolutionary rates of yeast genes (*dN* and *dN*/dS) were obtained from the comparison of 4 closely related species (Hirsh, Fraser, Wall 2005).

4.3.2 Network data

All network and other data is for *Saccharomyces cerevisiae*. For all networks, only genes which have evolutionary rate values were taken into account. The co-expression network was obtained from a combination of 40 time-series microarray experiments (Kafri, Bar-Even, Pilpel 2005). Pearson's correlation coefficient *r* across all experiments was used as a measure of the co-expression level of two genes. Two genes are linked in the resulting co-expression network

if their expression profiles are correlated with $r >= 0.5$.

Protein protein interaction data was obtained from the CCSB interactome database [9]. To ensure high data quality, literature-based interactions (LC-multiple) and co-complex associations for which we are not sure if the two proteins are in direct contact with each other (Combined-AP/MS) were not included. In total, we got four dataset (CCSB-YI1, Ito-Core, Uetz-Screen and Y2H-Union), containing a total of 6273 protein-protein interactions. We built the union of these four sets, removing duplicated interactions. This led to 4349 interactions in the final data set.

A synthetic lethality (strong negative genetic interaction) network was extracted from BIOGRID, version 2.0.60 (Breitkreutz *et al*. 2008). Only interactions tagged with "Synthetic Lethality" were used, resulting in a total of 15196 interactions. After removing duplicate interactions, we obtained a final data set of 13030 interactions. Another genetic interaction data set was published by (Costanzo *et al*.). Only interactions below a stringent cutoff were used, resulting in a set of 74984 interactions. The yeast metabolic network was obtained from (Forster *et al*. 2003) and compiled according to the procedure previously reported (Vitkup, Kharchenko, Wagner 2006). After removing duplicate interactions, we got 11179 interactions in our dataset (14283 in the raw data).

4.3.3 Other datasets

Protein abundance in log phase growth were taken from (Ghaemmaghami *et al*. 2003), yeast mRNA expression level was from (Holstege *et al*. 1998), and CAI and *dN* were obtained from (Hirsh, Fraser, Wall 2005). We used *dN* to represent the evolutionary rate of yeast. Alternatively using *dN*/dS does not change the results. Protein length was calculated based on the protein sequence of *Saccharomyces cerevisiae* (Cherry *et al*. 1998) via a Perl script. The identity of more than 1100 essential genes was download from the *Saccharomyces* Genome Deletion Project web page[10].

[9] http://interactome.dfci.harvard.edu/index.php?page=home

[10] http://yeastdeletion.stanford.edu/

4.4 Results

4.4.1 Proteins evolve at similar rates as their network neighbours

By comparing the differences of evolutionary rate of interacting proteins and randomly chosen protein pairs in *S. cerevisiae*, Fraser *et al* (Fraser *et al*. 2002) found that interacting proteins have similar evolutionary rates. Since the sample size they used was quite small, their observation might not hold true in larger networks. Thus, it is important to double check their results using updated protein interaction data. In order to ensure that all protein-protein interactions in the dataset refer to direct contact between the proteins, protein interactions within the same complex but without direct contact were excluded.

We considered each protein in turn as the 'focal' protein, and calculated the average evolutionary rate across its network neighbours. If adjacent proteins show similar evolutionary rates, we would expect a positive correlation between the evolutionary rate of the focal protein this neighbour average. We indeed found the expected correlation in the protein-protein interaction data (Figure 4-1; for dN, $r = 0.15$, $p = 3.7\text{x}10^{-6}$; for dN /dS, $r = 0.14$, $p = 2.1\text{x}10^{-5}$).

We thus confirmed that neighbouring proteins in the yeast protein-protein interaction network evolve at similar rates. Is this correlation a general feature of all biological networks? If all types of interactions impose constraints on sequence evolution, this correlation would be expected. To test this hypothesis, we used recently published yeast network data, including co-expression data (Kafri, Bar-Even, Pilpel 2005), genetic interaction data (Breitkreutz *et al*. 2008), transcription regulation data (Balaji *et al*. 2006), and metabolic data (Forster *et al*. 2003). After removal of duplicated links, we obtained final datasets with 14283 interactions in the metabolic network, 12873 interactions in the transcription network, 13030 interactions in the synthetic lethal interaction network, and 689100 interactions in the co-expression network. Note that for our first analysis we only chose synthetic lethal interactions for genetic interactions. After that we checked the effects in a much larger data.

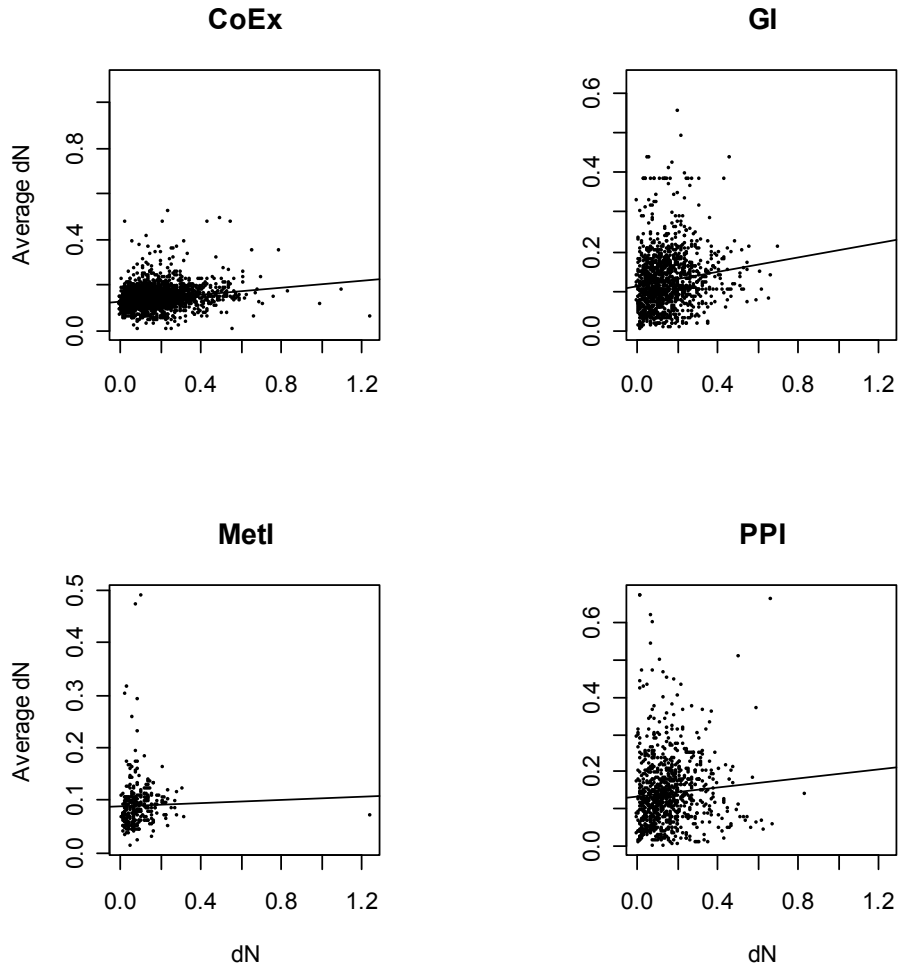


Figure 4-1 Correlations between the evolutionary rate *dN* of focal proteins and the average rate of their network neighbours. CoEx: co-expression network; GI: genetic interaction network; MetI: metabolic network; PPI: protein protein interaction network.

As seen in Table 4-1, except for the transcription regulation network, each of biological networks has a significant correlation between the evolutionary rates of the focal proteins and the average evolutionary rates of their neighbours (p<0.01 from comparison to random pairs in each case). These correlations are indeed stronger than those seen for the protein-protein interaction network. Thus, interacting neighbours evolve at similarly rates for all available biological networks, with the sole exception of transcription regulation networks.

For the transcription regulation network, there is no signal of a correlation between evolutionary rates of network neighbours (Table 4-1). Thus, there appears to be no connection between the sequence evolution of transcription

factors and their target genes. Since network rewriting is the main evolutionary force of transcription regulation (Ihmels *et al*. 2005), sequence co-evolution between transcription factors and target genes is apparently not a major force of transcription network evolution.

Table 4-1 Significant correlations between the evolutionary rate of proteins and the average rate of their network neighbours, except for the transcription network.

	Pearson's *r* for *dN*	*P*	Pearson's *r* for *dN*/dS	*p*
PPI	0.15	$3.7x10^{-6}$	0.14	$2.1x10^{-5}$
SLI	0.18	$6.2x10^{-11}$	0.16	$8.5x10^{-9}$
MetI	0.21	$1.6x10^{-4}$	0.18	0.0017
CoEx	0.27	$3.0x10^{-51}$	0.23	$1.5x10^{-37}$
TR	-0.02	0.34	-0.02	0.50

4.4.2 The influence of network neighbourhoods on protein evolution can largely be explained by expression level

Although our preliminary analysis shows that in most of the networks, neighbouring genes have similar evolution rates, we have to control for confounding variables. The first parameter one might think of in this context is network connectivity, because in almost all of the networks previous analysis found that connectivity (the number of direct neighbours) influences evolutionary rates. It was first reported that there is a negative correlation between evolution rate and connectivity in the protein-protein interaction network (Fraser *et al*. 2002), although other researchers later noticed that only hub protein (those with exceptionally high connectivity) evolve slowly (Jordan, Wolf, Koonin 2003). In the metabolic network, based on the analysis of 671 enzymes, Vitkup *et al* showed that both highly connected enzymes and enzymes with large metabolic fluxes evolve slowly (Vitkup, Kharchenko, Wagner 2006). People also found that hubs tend to evolve slowly both in the yeast co-expression network (Carlson *et al*. 2006) and in recently released genetic interaction data (Costanzo *et al*.). For all of the networks analysed here, except for the transcription regulation network, we confirmed a negative correlation

between connectivity and evolutionary rate *dN* (Table 4-1).

Table 4-2 Correlation between *dN* and avergage *dN* of the neighbours after controlling for protein abundance, CAI, mRNA expression and nod number.

dN	control for protein abundance		control for CAI		control for mRNA expression		control for connectivity	
	Cor	*r p*	Cor	*r p*	Cor	*r p*	Cor	*r p*
PPI	0.068	0.08	0.031	0.4	0.059	0.08	0.08	0.025
SLI	0.13	$5x10^{-5}$	0.10	0.0003	0.14	$6x10^{-7}$	0.1	$4.5x10^{-7}$
MetI	0.014	0.8	-0.040	0.5	0.0034	1	0.03	0.62
TR	-0.013	0.7	-0.023	0.4	-0.017	0.5	-0.005	0.84
CoEx	0.20	0	0.143	$3x10^{-15}$	0.17	0	0.2	0

However, these correlations with connectivity are not sufficient to explain the observed correlations among network neighbours. After controlling for connectivity using partial regression analysis, only the correlation between neighbours in the metabolic network became non-significant (Table 4-2). Thus, connectivity cannot fully explain why neighbouring proteins have similar evolution rate.

The most important factors that determine protein evolutionary rated revealed in recent years is expression abundance. Drummond *et al* (Drummond, Raval, Wilke 2006) observed that, by using principal component analysis, protein expression level and abundance nearly explain half of the variation in protein evolutionary rate in yeast. So if two proteins co-evolve, it is very likely that because they have similar expression level. What is more, An analysis based on 4,708 protein protein interactions shows that co-evolution of interacting protein in protein interaction network are largely not due to the compensatory mutations between interface, similar constrains like expression level could account for most of this (Hakes *et al*. 2007). So it is very likely that in most of the networks level, expression level play an important role for similar evolutionary rate of neighbouring genes.

It is widely accepted that there are three variables that measure aspects of

protein expression in yeast: mRNA expression level, codon usage bias (measured, e.g., as codon adaptation index, CAI), and protein abundance. After controlling for expression level using any one of these three factors, both the protein-protein interaction network and the metabolic network do not show any significant correlations among neighbours anymore. Similarly, the significance of the correlation in the genetic interaction network is greatly reduced (Table 4-2). But in the co-expression network, the correlation between focal protein *dN* and average *dN* of the neighbours is still highly significant. Moreover, even after we control for the other two potential confounding factors, protein length and gene essentiality, the correlation in the co-expression network still remains significant ($p = 0.017$ after controlling for protein length and $p < 10^{-15}$ after controlling for protein dispensability). This result highlights the importance of co-expression relationships in protein evolution. Recently, it was shown that in humans, co-expression also influences protein evolution rate (Vinogradov 2010).

4.4.3 The correlation between the evolutionary rate of proteins and their neighbours is not observed in a large-scale genetic interaction network

In the case of the genetic interaction network, it was not quite clear how much of the observed correlation is in fact due to expression level. Thus, we decided to corroborate our findings by analysis of a recent large set of genetic interaction data released (Costanzo *et al*. 2010). Only interactions fulfilling a stringent cutoff criterion were used in order to ensure high data quality. If the correlation observed above is a general feature of genetic interacting networks, we expect to see a corresponding signal in the larger dataset, even if the latter includes weaker as well as positive interactions (which were not present in the synthetic lethal network analyzed above). But we did not observe any significant correlations among evolutionary rates of network neighbours, neither for the total network, nor for only the negative interactions (total network: $p = 0.3$, Spearman's $r = 0.024$; negative network: $p = 0.31$, $r = 0.024$). Thus the evidence that neighbouring proteins show similar evolutionary rates in genetic interaction networks appears not convincing; possibly only synthetic lethal

interactions have an influence on protein evolution, while weaker or positive interactions do not.

4.5 Discussion

Neighbouring proteins in different biological networks evolve at similar rates. By controlling for other factors that may constrain protein evolution, we have shown that network connectivity and expression level are capable of explaining this effect in the metabolic network. We also did not find solid evidence of correlated evolution among neighbours in the genetic interaction network. In the remaining networks, similar expression levels of neighbours alone is sufficient to explain the correlated evolutionary rates. Thus, it appears that neighbouring genes evolve at similar rates largely because they have similar expression levels. These results raise an interesting question, namely, if many factors affecting protein evolution actually act through constraints on expression level. From the analysis of the co-expression network, we find that co-expression (beyond expression level) is also an important factor for protein evolution in yeast.

Chapter 5 The Predictors of Phenotypic Capacitors

5.1 Project summary

Many single-gene knockouts result in increased phenotypic (*e.g.*, morphological) variability among their offspring. This has been interpreted as an intrinsic ability of genes to buffer genetic and environmental variation, as exemplified by the chaperone function of Hsp90. Testing five different genomic and network variables, we show that only the fitness effect of the haploid knockout and the number of genetic interactions are strongly correlated with 'phenotypic potential'. This leads us to suggest that it is not failure of a specific buffering function that causes the release of phenotypic variation in mutants; instead, cells that are functionally compromised by a mutation are no longer capable of compensating for a sub-optimal pathway by either flux re-routing or by increased activity of other network components. Furthermore, we demonstrate a stronger phenotypic potential of genes involved in essential complexes, which again can be traced to stronger fitness effects and more genetic interactions of the mutants.

5.2 Introduction

One of the most fundamental and challenging problems in biology is the relationship between genotype and phenotype. Little is known about this relationship in most biological systems. The genotype-phenotype relationship is complicated by the fact that loss-of-function mutations of some individual genes are capable of increasing phenotypic variability in a wide range of traits; such genes are termed 'phenotypic capacitors' (Rutherford, Lindquist 1998; Hartman, Garvik, Hartwell 2001; Mitchell-Olds, Knight 2002; Queitsch, Sangster, Lindquist 2002; Bergman, Siegal 2003; Wagner 2003; Suzuki, Nijhout 2006). The induced variability may either be due to cryptic genetic variation that is released by the mutation (Hermisson, Wagner 2004), or may be

non-genetic(Levy, Siegal 2008). Phenotypic capacitors have been discovered in many species, including *Drosophila*, *Arabidopsis*, *Manduca*, *E. coli* and yeast (Rutherford, Lindquist 1998; Hartman, Garvik, Hartwell 2001; Queitsch, Sangster, Lindquist 2002; Kitano 2004; Cooper *et a l*. 2006; Suzuki, Nijhout 2006; Sangster *et al*. 2008).

Theoretical simulations of complex cellular networks have suggested that most genes reveal cryptic genetic variation when functionally compromised (Bergman, Siegal 2003). More generally, phenotypic release of hidden genetic variation appears to be a generic property of models with epistasis or genotype-environment interactions (Hermisson, Wagner 2004) This process may even be selectively favourable: an allele for the revelation of cryptic genetic variation can invade a population if revelation is sometimes selectively favourable (Masel 2005). However, it is unclear if such alleles do in fact exist, as phenotypic capacitance can arise as a direct consequence of network structure (Bergman, Siegal 2003; Hermisson, Wagner 2004).

The standard model of a phenotypic capacitor is the heat shock protein Hsp90, a chaperone that helps many proteins to achieve their correct 3-D structure. This assistance in folding likely removes some selective constraints on amino acid sequence evolution, as reported for another chaperone, GroEL (Warnecke, Hurst 2010). This allows the accumulation of polymorphisms that would impede correct protein folding in the absence of the chaperone. Knockout of Hsp90 then releases these hidden polymorphisms, resulting in variation in protein folding efficiency between genetically different individuals (Milton *et al*. 2003). However, the same study reported that – at least in flies – purely non-genetic phenotypic variation was not released by Hsp90 knockout. Thus, it is currently unclear if phenotypic capacitors differ in their effects on genetic and non-genetic variability.

In a recent study, Levy and Siegal examined non-genetic variation among yeast cells (Levy, Siegal 2008). They found more than 300 phenotypic capacitors, and reported that many of these had a large number of synthetic lethal interactions; at the same time, the authors did not observe strong effects of the

knockouts on growth rates. These findings were interpreted as evidence for incomplete functional redundancy at multiple levels in the genetic architecture(Levy, Siegal 2008).

Thus, theory predicts that (i) phenotypic capacitors should have similar effects on genetically and non-genetically caused phenotypic variation (Milton *et al.* 2003); (ii) strong functional impairment of a majority of genes leads to the revelation of cryptic genetic variation (Bergman, Siegal 2003; Hermisson, Wagner 2004); and (iii) the release of cryptic genetic variation by knockout mutations is related to genetic interactions of the mutated gene (Hermisson, Wagner 2004). This suggests that phenotypic capacitors that reveal genetically caused phenotypic variation should often be genes with severe knockout-effects and with many synthetic lethal interactions. Here, we test if these two factors can predict 'phenotypic potential', *i.e.,* the ability of genes to act as phenotypic capacitors for non-genetically caused phenotypic variability.

Many genes perform their biological functions as parts of protein complexes. Thus, functional impairment is a consequence of compromised complex function rather than an effect of the individual gene. Consistent with this view, it was shown that protein complexes contain subunits of similar essentiality (Dezso, Oltvai, Barabasi 2003; Fraser, Plotkin 2007; Hart, Lee, Marcotte 2007). Similarly, we expect phenotypic potential to be a feature of the protein complex rather than of individual subunits. Thus, we also test if essential complexes (those containing essential genes) contain more phenotypic capacitors than non-essential complexes.

5.3 Methods

5.3.1 Phenotypic potential

The knockout of individual genes can cause increased phenotypic variability. Here, we used previously published data measuring non-genetically caused morphological variation among the offspring of single-gene deletion *S. cerevisiae* strains (Levy, Siegal 2008). This dataset contains measurements of

'phenotypic potential' (i.e., the amount of variation observed across many independent morphological traits in the offspring) for 4683 non-essential genes, resulting in the identification of 502 phenotypic capacitors.

5.3.2 Other *S. cerevisiae* data

Essential genes were taken from the Saccharomyces genome deletion project website[11]. Membership of 408 protein complexes was obtained from CYC2008 (Pu *et al*. 2009), which is based on manual curation. Complexes that contained at least one essential gene were termed 'essential complexes', while other complexes were considered non-essential. Connectivity of proteins in the protein-protein interaction network was obtained from (Levy, Siegal 2008).

5.3.4 Statistics

Distribution shifts were examined using the Brunner-Munzel test, which is in many situations more robust than, *e.g.*, the Wilcoxon rank sum test. Linear models were examined using the glm function implemented in R (Ihaka, Gentleman 1996). The relative importance of individual predictor variables was estimated using the *lmg* statistic implemented in the *relaimpo* package for R (Grömping 2006); this statistic can be viewed as an average over the contribution of the variables in models of different sizes. 95% confidence intervals were calculated using bootstrapping.

5.4 Results

5.4.1 Phenotypic potential is influenced by haploid fitness, genetic interactions, and protein length

Based on the arguments outlined above, we expect phenotypic potential to be correlated with the severity of fitness reductions in gene knockouts, and with the number of synthetic lethal interactions. Both predictions are indeed confirmed by experimental data on non-genetically caused morphological variation(Levy, Siegal 2008). We find a significant negative correlation between phenotypic potential and haploid fitness (Figure 5-1; Pearson's r=-0.29, $p<10^{-15}$), and a

[11] http://www-sequence.stanford.edu/group/yeast_deletion_project/deletions3.html

significant positive correlation between phenotypic potential and the number of synthetic lethal interactions (Figure 5-2; r=0.30, p<10^{-15}; interaction number on log-scale). Thus, both haploid fitness and connectivity in the synthetic lethal network explain about 9% of the variation in phenotypic potential.

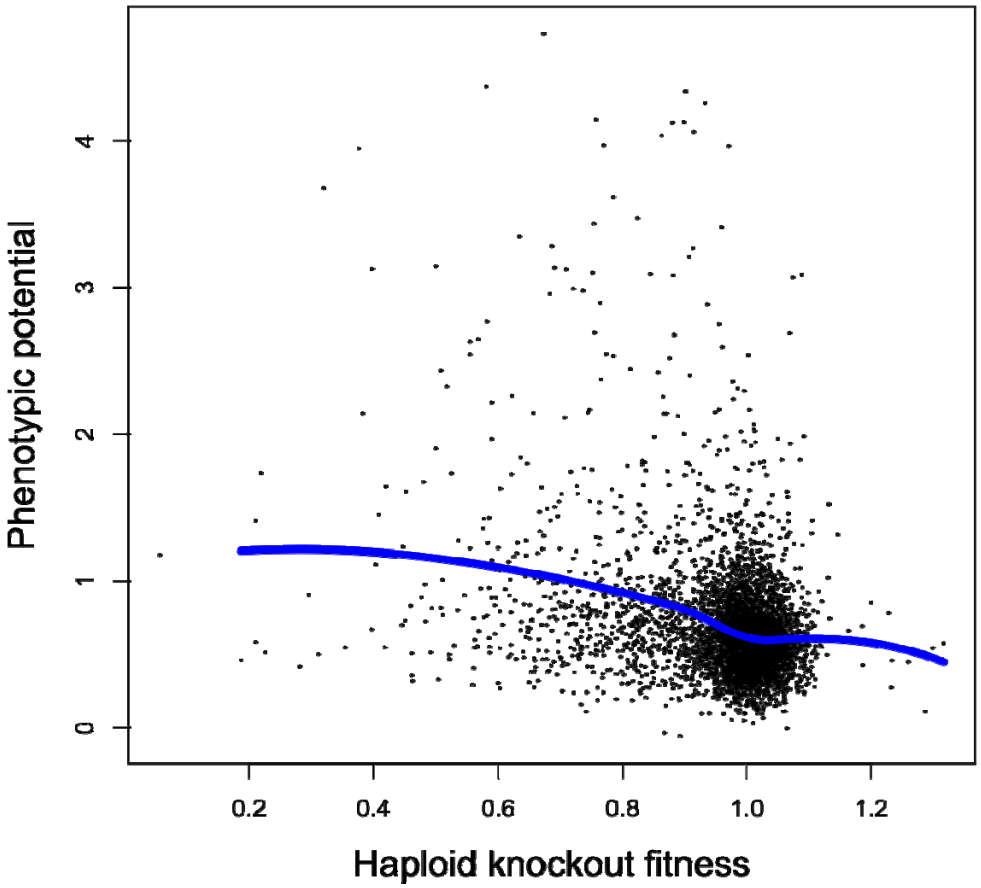


Figure 5-1 Relationship between fitness of the haploid knockout and phenotypic potential (*Pearson's* r=-0.29, p<10^{-15}). The blue line is a Loess curve fitted to the data.

Other variables frequently associated with functional and evolutionary properties of yeast genes are expression level, protein length, and the number of protein-protein interactions. We find that each of these variables shows a weak but statistically significant correlation with phenotypic potential (mRNA expression level: r=0.052, p=0.00011; protein length: r=0.056, p=0.00013; number of protein-protein interactions: r=0.078, p=6.8x10^{-5}).

5.4.2 Other variables do not add to the predictive power of a combined model

We thus have five variables – haploid knockout fitness, synthetic lethal interactions, length, expression level, and protein-protein interactions. However, it is known that many of these variables are correlated among themselves(Pal, Papp, Lercher 2006). Thus, it is possible that only some of these variables are directly connected to phenotypic potential; conversely, the effect of other variables might be due to a confounding variable among the former set. To examine which variables have the strongest explanatory power for variation in phenotypic potential, we employed a linear model.

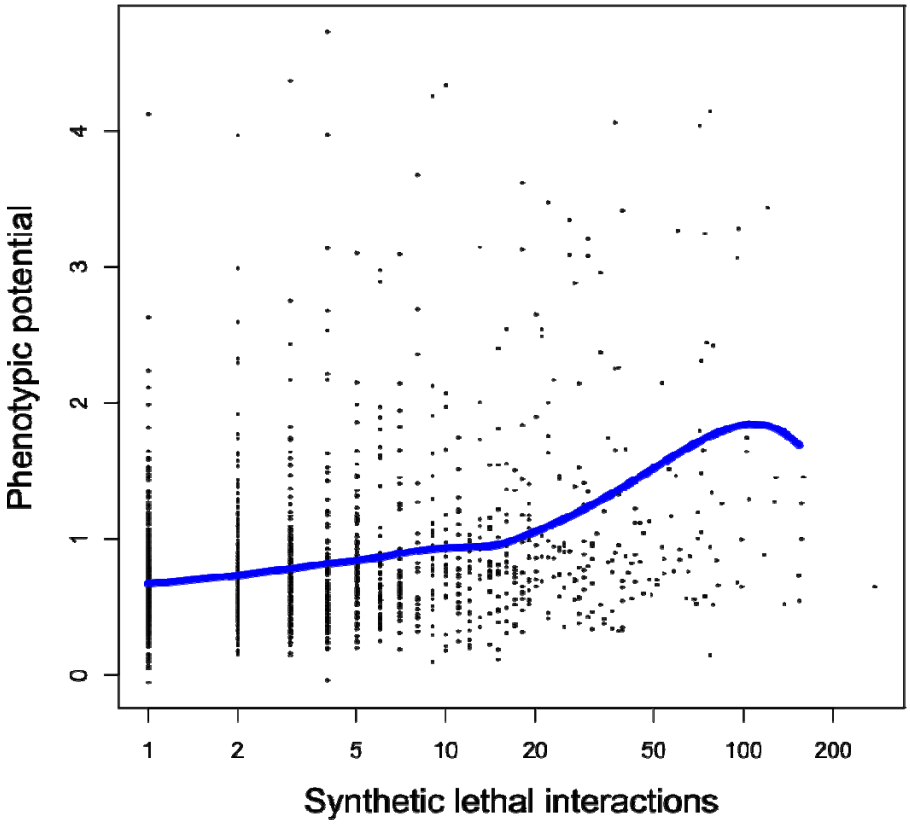


Figure 5-2 Relationship between the number of synthetic lethal interactions and phenotypic potential (*Pearson's r*=0.30, $p<10^{-15}$); only genes with at least one synthetic lethal interaction are used. The blue line is a Loess curve fitted to the data.

The linear model of the form: Phenotypic_potential ~ Haploid_fitness + LC_SLI_degree + length + mRNA_expression + AMS_PPI_degree showed that

only haploid fitness, the connectivity (degree) in the synthetic lethal interaction network, and length contribute independently to the variation of phenotypic potential (Table 5-1); mRNA expression level and protein-protein interaction connectivity do not add any further significant contributions ($p = 0.62$ and $p = 0.21$, respectively). In fact, both haploid fitness and synthetic lethal interactions alone are enough to render the explanatory power of expression level and protein-protein interactions insignificant ($p > 0.18$ in each case).

What is the relative importance of fitness, genetic interactions, and protein length for predicting phenotypic potential? As shown in Table 5-1, haploid fitness and the number of synthetic lethal interactions explain almost 8.8% and 4.6% of the variation in phenotypic potential, respectively, while the contribution of protein length is minute. Thus, the extent to which a mutant increases phenotypic variation is indeed correlated to both the fitness effect of the knockout and to the number of severe negative genetic interactions the mutated protein has, while other tested variables appear to be unimportant.

Table 5-1 Statistical significance and relative importance of the three significant predictor variables for phenotypic potential. 'Relative importance' is a measure of how much variation in phenotypic potential each variable can explain independently (Grömping 2006).

Predictor	*p* (linear model)	Relative importance (95% CI)
Haploid fitness	$<10^{-15}$	8.8% (6.4%-11.9%)
Synthetic lethal interactions	$<10^{-15}$	4.6% (3.0%-6.8%)
Length	0.019	0.3% (0.02%-0.98%)

5.4.3 Protein complexes as phenotypic capacitors

The active cellular agents are often not individual proteins, but rather aggregates of different proteins - protein complexes. Thus, mutations of individual members of a complex are not functionally independent, but might be considered different forms of damage to one functional unit. Consistent with this view, it was found that protein complexes are either biased towards essential or towards non-essential genes (Dezso, Oltvai, Barabasi 2003; Fraser, Plotkin 2007; Hart, Lee, Marcotte 2007); *i.e.*, essentiality is a feature of the complex rather than of the individual gene.

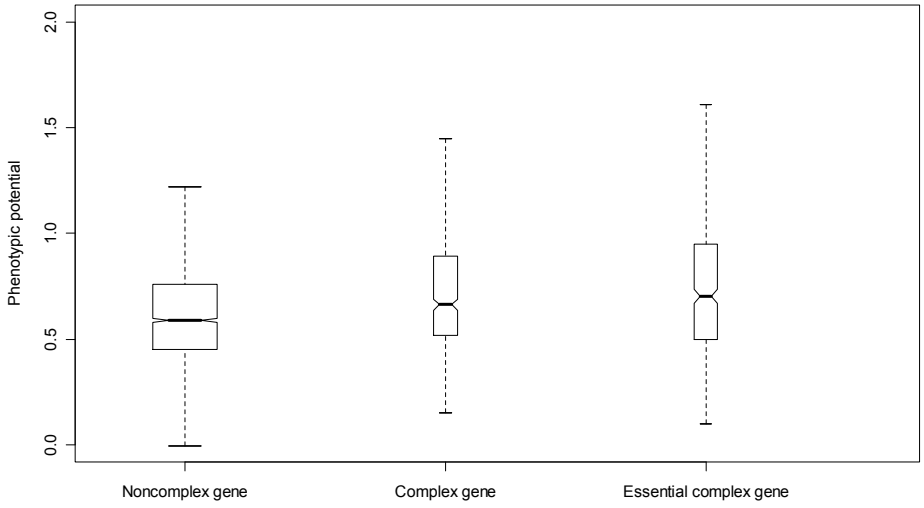


Figure 5-3 Genes in essential protein complexes have higher phenotypic potential than genes not involved in protein complexes or involved in non-essential protein complexes.

Essential complexes as those protein complexes that are lethal when severely impaired in their function. Most essential complexes contain one or more essential genes, and as working definition, we thus term a complex essential if it contains at least one essential gene. The protein complex data we used was downloaded from CYC2008(Pu *et al*. 2009), which contains manually curated complexes, and is an update of the Munich Information Center of Protein Sequences (MIPS) database. We also checked our main results in two recently published algorithm-based protein complex data sets (Hart, Lee, Marcotte 2007; Wang *et al*. 2009) and obtained similar results (not shown).

As seen above, phenotypic potential is correlated to the severity of the knockout effect for individual proteins. This suggests that essential complexes may often act as phenotypic capacitors: if their function is impaired (*e.g.*, by the knockout of a non-essential subunit), then this should increase phenotypic variation. Consequently, proteins that form part of an essential complex should on average have higher phenotypic potential than proteins that are part of non-essential protein complexes. As predicted, we found that the mean phenotypic potential of non-essential genes is significantly larger for genes in essential complexes

(0.827 +/- 0.025) compared to either genes in non-essential complex (0.806 +/- 0.025, $p<10^{-15}$ in Brunner-Munzel test) or genes not in any protein complexes (0.651 +/- 0.006, $p<10^{-15}$; see also Figure 5-3).

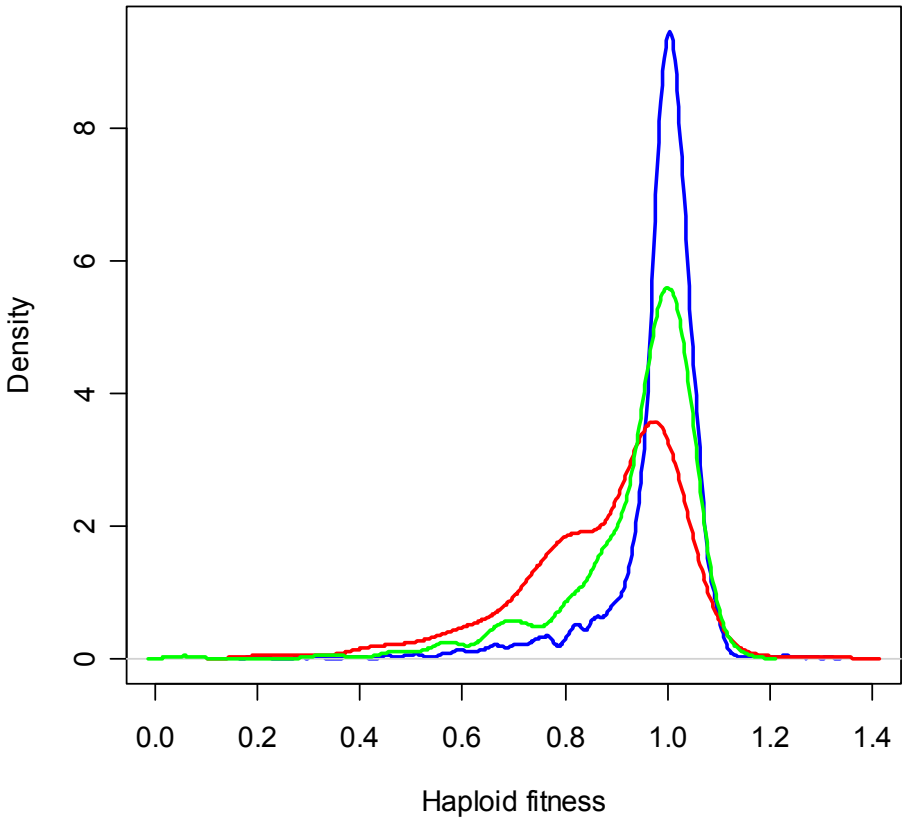


Figure 5-4 The distribution of haploid fitness is shifted towards lower values for genes involved in essential complexes (red) compared to non-essential complexes (green) or genes outside complexes (blue).

What is more, we found that 18% of phenotypic capacitors (genes with high phenotypic potential (Levy, Siegal 2008)) are located in essential rather than non-essential complexes, a fraction more than two times higher than expected from the proportion of non-phenotypic capacitors (7.9%, $p=7.3\text{x}10^{-12}$, Fisher's exact test). More generally, just as essential genes tend to be located in complexes ($p=1.2\text{x}10^{-112}$), the same is true for phenotypic capacitors ($p=5.6\text{x}10^{-21}$). But even excluding phenotypic capacitors, genes in essential complexes still show higher phenotypic potential (0.618+\-0.0112 vs 0.578+/- 0.0031 in non-essential complexes, p = 0.0011 in Brunner-Munzel test). Thus, essential complexes as functional units tend to have high phenotypic potential.

Is this effect of essential complexes independent of knockout fitness and of the number of synthetic lethal interactions, which were shown to be important predictors of phenotypic potential above? The Haploid knockout fitness (Figure 5-4) tend to be higher for genes in essential complexes compared to genes in non-essential complexes (p=3.3x10^{-14}, Brunner-Munzel test) or genes not involved in complexes (p<10^{-15}). However, when we incorporate involvement in an essential complex into the general linear model, we find that both haploid knockout fitness and the number of synthetic lethal interactions alone are sufficient to remove any additional contribution from essential complexes. Thus, it appears that the role of essential complexes is mediated entirely through the severity of knockout effects and through genetic interactions. In sum, while protein complexes form the functional units of the cellular machinery, we can predict their influence on the release of phenotypic variation in the same way as for proteins that act individually.

5.5 Discussion

The knockout of a single gene often leads to an increase of phenotypic (*e.g.*, morphological) variability, measured as phenotypic potential in the data analyzed here (Bergman, Siegal 2003; Levy, Siegal 2008). Consistent with theoretical expectations, we find that increased phenotypic potential is associated with stronger (haploid) fitness effects of the knockout mutant, as well as with an increased number of synthetic lethal interactions. Here, we analyze data for non-genetically caused morphological variability. The connection with synthetic lethal interactions thus suggests that phenotypic capacitors indeed often work on both genetically caused and non-genetically caused phenotypic variability, as predicted from theoretical considerations (Milton *et al*. 2003).

Levy and Siegal (Levy, Siegal 2008) also observed a significant correlation between phenotypic potential and both genetic interactions and haploid growth rate (see their Figure 3E-F). However, their interpretation differed markedly from ours; they did not consider growth rate an important predictor of

phenotypic capacity, stating instead that 'knockouts of these genes do not tend to cause severe decreases in growth rate' (Levy, Siegal 2008). This, however, is not surprising: knockout growth rates tend to be bi-modal, with the majority of mutants being either (nearly) lethal or showing little reduction in fitness. As only non-essential genes can be tested for phenotypic potential, severe decreases in growth rate are expected to be rare.

Based on the function of Hsp90, phenotypic capacitors are often viewed as having buffering functions; *i.e.,* it is (often implicitly) assumed that selection has directly acted on the protein's ability to mask genetic or environmental variation (Masel 2005; Levy, Siegal 2008). Consistent with theoretical (Bergman, Siegal 2003) and experimental (Hermisson, Wagner 2004) results, increased variability may instead be a general consequence of functional impairment in complex cellular networks (Lehar *et al* . 2008). This may be illustrated by a simple example: if flux through a given metabolic pathway can be maintained by increased production of the pathway substrate, then small variations in the efficiency of pathway enzymes can be compensated. If a mutation in an enzyme that feeds the pathway reduces substrate production, then the previously 'cryptic' variation in pathway efficiency will become exposed, resulting in increased phenotypic variability. If a mutation affects the efficiency of important cellular processes, then such 'domino' effects may spread far through the network. If this view is correct, then phenotypic potential is not so much a measure of buffering, but rather of functional importance and of functional centrality in cellular networks. That phenotypic potential is connected with fitness effects of the single-gene knockout supports this notion; in this light, Hsp90 appears as an exception rather than a prototypical phenotypic capacitor.

The subunits of protein complexes are only functional in combination. Accordingly, knockouts of the subunits must be viewed as partial deletions of the complexes, which may in many cases still be partially functional. This view of complexes as evolutionary and genetic modules of course also applies to the concept of phenotypic capacitors. As predicted by this view, we find that proteins in essential complexes have higher phenotypic potential then other

proteins. Again, this enrichment in phenotypic capacitors can be traced to stronger fitness effects and more genetic interactions of these proteins.

Chapter 6 Conclusion

A decade's accumulation of yeast genomic and high throughput interaction data has provided us with an excellent opportunity to study the principles of genome organization and its consequences on networks and phenotypes. Through the analysis of genome-wide data, in this thesis I reported that the co-expression of local genomic gene clusters is not just a result of chromatin effects: even when these gene pairs are separated through genomic rearrangements, they still show weak but significant similarities in their expression profiles. We interpret this as a clear signal of natural selection (Chapter 3).

Furthermore, we show that the underlying principles of local genomic organization go well beyond mere co-expression, but have to a substantial part evolved to optimize noise levels. This phenomenon could potentially explain the existence of 60% of non-coding RNAs in the *Saccharomyces cerevisiae* genome (Chapter 2). The role of non-coding RNAs in noise reduction indicates the importance of noise control in the genome.

Co-expression of genes is not only a local genomic feature, but also an important mechanism of genes from different chromosomes to achieve a specific function or response to environmental stimulations. In chapter 4 we discovered that co-expression relationships also influence constraints on protein evolution: co-expressed partners evolve at similar rates. This constraint is independent of mRNA expression levels and protein abundances; this is in sharp contrast to the influence of protein-protein interactions and metabolic interactions on evolutionary rates.

While cellular interactions can be projected onto different biological networks, it appears that genetic interaction networks are closely connected with the phenotypic variance of the cell: knocking out the hubs of genetic networks releases more phenotypic variation. In addition we also provide evidence that genes in essential protein complexes generally have more phenotypic potential. This potential may be due

to the fact that knockouts in essential complexes – which make the complexes functionally abnormal – almost always cause serious growth defects (chapter 5).

In sum, our analyses of cellular interactions in yeast demonstrate that phenotypic features can be at least partly explained from considerations on the network level. Although each network provides us with a particular view of functional cells, the cell itself is affected by all these networks simultaneously. An important topic for future research is then the integration of all network types, in order to better understand the influence of molecular interactions on protein evolution and cellular phenotypes. Furthermore, bi-directional promoters should represent just one way of controlling noise. There must be more mechanisms that contribute to this phenomenon, particularly at the level of the transcription regulation network. What these mechanisms are, and the phenotypic consequences if these mechanisms fail, also requires further investigation.

Appendix Genes with Bi-directional promoters in *S. cerevisiae*

Table A-1 Protein coding genes which share a promoter with a cryptic unstable transcripts (CUTs) (309), compiled based on (Xu *et al*. 2009).

YPRWtau4	YOL119C	YLR244C	YJL128C	YGR125W	YDR279W	YOR014W
YPR196W	YOL115W	YLR231C	YJL115W	YGR123C	YDR255C	YOR005C
YPR193C	YOL057W	YLR193C	YJL104W	YGR119C	YDR229W	YOR001W
YPR149W	YNR061C	YLR170C	YJL080C	YGR111W	YOL159C-A	YDR184C
YPR135W	YNR054C	YLR167W	YJL061W	YGR097W	YDR183C-A	YOL158C
YPR127W	YNR045W	YLR166C	YJL050W	YGR086C	YDR084C	YOL155C
YPR067W	YNR015W	YLR126C	YJL011C	YGR042W	YDR074W	YOL153C
YPR022C	YNR006W	YLR112W	YJL001W	YGR036C	YDR062W	YOL130W
YPL262W	YNL299W	YLR111W	YIL168W	YGR017W	YDR044W	YJR108W
YPL259C	YNL238W	YLR069C	YIL167W	YGL252C	YDR028C	YJR105W
YPL249C	YNL236W	YLR049C	YIL165C	YGL231C	YDR024W	YJR097W
YPL242C	YNL209W	YLR036C	YIL164C	YGL162W	YDR023W	YJR035W
YPL225W	YNL052W	YLR032W	YIL114C	YGL155W	YDR012W	YJL186W
YPL186C	YMR318C	YLR028C	YIL112W	YGL082W	YDL211C	YJL171C
YPL140C	YMR312W	YLL060C	YIL110W	YGL078C	YDL080C	YJL164C
YPL123C	YMR302C	YLL059C	YIL102C-A	YGL035C	YDL072C	YJL145W
YPL108W	YMR258C	YKRCdelta11	YIL086C	YFR007W	YDL040C	YDR497C
YPL107W	YMR230W	YKR103W	YIL085C	YFL017W-A	YDL025C	YDR473C
YPL106C	YMR211W	YKR094C	YIL074C	YFL013C	YCRCdelta6	YDR440W
YPL059W	YMR177W	YKR088C	YIL038C	YFL004W	YCR090C	YDR397C
YPL051W	YMR126C	YKR014C	YIL035C	YER186C	YCR072C	YDR394W
YPL050C	YMR121C	YKR010C	YHR207C	YER183C	YCR043C	YDR361C
YPL032C	YMR099C	YKL219W	YHR206W	YER176W	YCR028C-A	YDR349C
YOR349W	YMR036C	YKL217W	YJR117W	YML062C	YHL001W	YHR076W
YOR346W	YMR019W	YHR073W-A	YGR283C	YKL037W	YKL178C	YHR027C
YOR316C	YMR013C-A	YKL160W	YER109C	YCL041C	YER012W	YHR073W
YOR310C	YMR013C	YKL129C	YER064C	YBR295W	YER010C	YER163C
YOR309C	YMR002W	YKL126W	YER056C-A	YBR293W	YER006W	YPL007C
YOR285W	YML086C	YKL111C	YER020W	YBR143C	YER002W	YDR520C
YOR250C	YML071C	YKL110C	YER016W	YBR115C	YEL071W	YDR517W
YOR204W	YML067C	YKL085W	YHR123W	YML051W	YDR541C	YDR505C

YJR148W	YML063W	YHL007C	YHR088W	YML028W	YKL218C	YDR318W
YLR395C	YLR376C	YLR332W	YLR326W	YLR268W	YLR262C	YLR261C
YLR249W	YGR282C	YGR266W	YGR240C	YGR238C	YGR225W	YGR204W
YGR158C	YGR145W	YOR203W	YOR179C	YOR136W	YOR130C	YOR114W
YOR108W	YOR047C	YML005W	YLR449W	YLR415C	YCL043C	YMR037C
YDR538W						

Table A-2 Protein coding genes that share a promoter with a transcription unit that contain both coding and non-coding genes (94), compiled based on (Xu *et al*. 2009).

YPR144C	YPR132W	YPR055W	YPL115C	YOR380W	YOR319W	YOR223W
YOL097W-A	YNR027W	YNL142W	YNL059C	YNL025C	YMR117C	YML102W
YLR366W	YLR364C-A	YLR347C	YLR204W	YLR200W	YLR133W	YLR125W
YLL018C	YKR062W	YKR045C	YKL138C	YKL128C	YKL035W	YKL023W
YJL117W	YJL020C	YJL008C	YIR036C	YIR021W	YHR050W	YHR031C
YGR188C	YGR110W	YGR024C	YGL071W	YFR034C	YFL026W	YFL021W
YER033C	YEL044W	YEL029C	YEL012W	YDR527W	YDR525W-A	YDR251W
YDL156W	YDL014W	YCR044C	YCL005W	YBR228W	YBR187W	YBR140C
YOR187W	YOR172W	YOR083W	YOR036W	YHL004W	YGR268C	YGR255C
YML007C-A	YLR455W	YLR435W	YLR401C	YER120W	YER067W	YER040W
YLR124W	YLR123C	YLR122C	YLR074C	YDR181C	YDR163W	YDR021W
YJR016C	YJL214W	YJL198W	YJL189W	YBL071W-A	YAL020C	YBL092W
YHR028C	YER169W	YDR221W				

Table A-3 Protein coding genes that share a promoter with a stable annotated transcripts (SUTs) (276), compiled based on (Xu *et al*. 2009).

YPR199C	YPL188W	YOR070C	YNL241C	YMR139W	YLR264W	YKL115C
YPR198W	YPL172C	YOR048C	YNL216W	YMR131C	YLR196W	YKL100C
YPR183W	YPL159C	YOR034C	YNL206C	YMR123W	YLR175W	YKL088W
YPR166C	YPL157W	YOR031W	YNL205C	YMR068W	YLR129W	YKL079W
YPR165W	YPL139C	YOR030W	YNL192W	YMR067C	YLR064W	YKL072W
YPR154W	YPL042C	YOR027W	YNL151C	YMR065W	YLR026C	YKL045W
YPR137W	YPL038W	YOR018W	YNL141W	YMR064W	YLR024C	YKL027W
YPR114W	YPL030W	YOR004W	YNL138W	YMR063W	YLR022C	YJR152W
YPR084W	YPL028W	YOL148C	YNL130C	YMR049C	YLL020C	YJR093C
YPR080W	YOR370C	YOL116W	YNL116W	YMR044W	YLL019C	YJR085C
YPR075C	YOR353C	YOL097C	YNL101W	YMR038C	YLL018C-A	YJR080C
YPR060C	YOR350C	YOL090W	YNL054W	YMR029C	YLL015W	YJR051W
YPR059C	YOR337W	YOL088C	YNL053W	YMR010W	YKR087C	YJR040W
YPR058W	YOR334W	YOL051W	YNL029C	YML129C	YKR077W	YJR014W
YPR034W	YOR323C	YNR059W	YNL024C-A	YML105C	YKR069W	YJR007W
YPR010C	YOR301W	YNR055C	YMR304W	YML072C	YKR059W	YJL217W
YPR004C	YOR298C-A	YNR053C	YMR300C	YML038C	YKR036C	YJL197W
YPL268W	YOR279C	YNR046W	YMR296C	YLR454W	YKR035C	YJL172W
YPL263C	YOR239W	YNL323W	YMR273C	YLR427W	YKR028W	YJL165C
YPL249C-A	YOR237W	YNL298W	YMR221C	YLR420W	YKR026C	YJL154C
YPL247C	YOR233W	YNL284C	YMR203W	YLR405W	YKR018C	YJL148W
YPL231W	YOR144C	YNL281W	YMR192W	YLR353W	YKL216W	YJL146W
YPL228W	YOR133W	YNL265C	YMR185W	YLR352W	YKL184W	YJL138C
YPL220W	YOR122C	YNL255C	YMR165C	YLR330W	YKL156W	YJL125C
YPL204W	YOR121C	YNL251C	YMR162C	YLR293C	YKL150W	YJL012C
YPL202C	YOR112W	YNL246W	YMR161W	YLR291C	YKL116C	YIR004W
YBR242W	YBR166C	YBR155W	YBR127C	YBR121C-A	YBR121C	YBR086C
YBR052C	YBR044C	YBR031W	YBR028C	YBR027C	YBR021W	YBL095W
YAL007C	YHR030C	YHR012W	YGR262C	YGR260W	YGR252W	YGR235C
YIL158W	YGR107W	YER081W	YDR165W	YGL122C	YDR485C	YDL137W
YIL156W	YGR106C	YER080W	YDR129C	YGL099W	YDR454C	YDL117W
YIL094C	YGR094W	YER077C	YDR119W	YGL092W	YDR449C	YDL092W
YIL075C	YGR061C	YER059W	YDR069C	YGL088W	YDR423C	YDL084W
YIL044C	YGR057C	YER027C	YDR017C	YGL080W	YDR341C	YDL081C
YIL008W	YGL248W	YER026C	YDL240W	YGL040C	YDR338C	YDL058W
YHR201C	YGL234W	YEL002C	YDL236W	YFR052W	YDR337W	YDL056W
YHR169W	YGL195W	YDR529C	YDL224C	YFR048W	YDR319C	YDL020C
YHR084W	YGL129C	YDR494W	YDL192W	YFR022W	YDR314C	YDL019C
YHR072W	YER125W	YDR270W	YBR260C	YFR021W	YDR313C	YDL013W
YHR068W	YER093C	YDR258C	YBR251W	YFR002W	YDR304C	YCR069W
YHR057C	YER082C	YDR169C	YBR247C	YER161C	YDR288W	YCL054W
YHR046C	YBR061C	YBR059C	YBR053C	YER155C	YDR280W	YCL050C
YHR042W	YBL017C	YAR002W	YAL059W	YER129W	YDR275W	YCL009C
YGR166W	YGR141W	YGR135W	YGR127W	YGR124W	YGR122W	YBR069C
YBL037W						

Table A-4 Protein coding genes that share a promoter with another coding gene are not colored (1441), compiled based on (Xu *et al*. 2009).

YPR191W	YOR196C	YNL078W	YLR274W	YJR144W	YHR065C	YFR027W
YPR190C	YOR195W	YNL074C	YLR273C	YJR143C	YHR059W	YFR026C
YPR189W	YOR194C	YNL073W	YLR267W	YJR141W	YHR058C	YFR019W
YPR188C	YOR186C-A	YNL062C	YLR266C	YJR140C	YHR049C-A	YFR018C
YPR187W	YOR171C	YNL061W	YLR248W	YJR135W-A	YHR045W	YFR004W
YPR186C	YOR169C	YNL057W	YLR247C	YJR135C	YHR044C	YFR003C
YPR180W	YOR168W	YNL049C	YLR246W	YJR112W	YHR040W	YFL056C
YPR179C	YOR167C	YNL048W	YLR245C	YJR111C	YHR039C-A	YFL055W
YPR178W	YOR165W	YNL041C	YLR243W	YJR092W	YHR035W	YFL050C
YPR176C	YOR164C	YNL040W	YLR242C	YJR091C	YHR034C	YFL049W
YPR175W	YOR160W	YNL022C	YLR240W	YJR089W	YHR032W-A	YFL048C
YPR174C	YOR159C	YNL021W	YLR239C	YJR088C	YHR032W	YFL047W
YPR153W	YOR158W	YNL011C	YLR234W	YJR084W	YHR028W-A	YFL034W
YPR152C	YOR157C	YNL010W	YLR233C	YJR083C	YFL034C-A	YHR025W
YPR140W	YOR147W	YNL002C	YLR226W	YJR068W	YHR024C	YFL033C
YPR139C	YOR145C	YPR133W-A	YLR225C	YJR067C	YHR020W	YFL032W
YNL001W	YOR142W	YMRWdelta16	YLR219W	YJR066W	YHR019C	YFL031W
YPR133C	YOR141C	YMR314W	YLR218C	YJR065C	YHR017W	YFL027C
YPR131C	YOR140W	YMR313C	YLR210W	YJR063W	YHR016C	YFL024C
YPR130C	YOR138C	YMR288W	YLR209C	YJR062C	YHR010W	YFL023W
YPR113W	YOR132W	YMR287C	YLR206W	YJR059W	YFL021C-A	YHR009C
YPR112C	YOR131C	YMR286W	YLR205C	YJR058C	YHL029C	YFL010C
YPR108W	YOR127W	YMR285C	YLR203C	YJR050W	YHL028W	YFL009W
YPR107C	YOR126C	YMR276W	YLR202C	YJR049C	YHL019C	YFL002C
YPR100W	YOR099W	YMR275C	YLR199C	YJR046W	YHL018W	YFL001W
YPR098C	YOR098C	YMR269W	YLR198C	YJR045C	YHL017W	YER173W
YPR097W	YOR094W	YMR268C	YLR190W	YJR042W	YHL016C	YER172C
YPR096C	YOR093C	YMR266W	YLR189C	YJR041C	YHL015W	YER168C
YPR095C	YOR082C	YMR265C	YLR185W	YJR039W	YHL006C	YER157W
YPR094W	YOR081C	YMR262W	YLR183C	YJR036C	YHL005C	YER156C
YPR093C	YOR080W	YMR261C	YLR182W	YJR032W	YHL003C	YER148W
YPR089W	YOR079C	YMR236W	YLR181C	YJR031C	YHL002W	YER147C
YPR088C	YOR077W	YMR235C	YLR180W	YJR022W	YGRWdelta31	YER143W
YPR086W	YOR076C	YMR228W	YLR179C	YJR021C	YGRCdelta20	YER142C
YPR085C	YOR069W	YMR227C	YLR177W	YJR020W	YGR278W	YER140W
YPR083W	YOR067C	YMR223W	YLR176C	YJR018W	YGR277C	YER139C
YPR082C	YOR065W	YMR222C	YLR173W	YJR013W	YGR275W	YER127W
YPR062W	YOR064C	YMR219W	YLR172C	YJR012C	YGR274C	YER126C
YPR061C	YOR061W	YMR218C	YLR148W	YJR005W	YGR270W	YER123W
YPR053C	YOR060C	YMR217W	YLR147C	YJR004C	YGR269W	YER122C
YPR051W	YOR059C	YMR216C	YLR145W	YJLWtau4	YER119C-A	YGR256W
YPR049C	YOR039W	YMR213W	YLR144C	YJLWdelta10	YGR217W	YER119C
YPR046W	YOR038C	YMR212C	YLR132C	YJL216C	YGR216C	YER100W
YPR045C	YOR035C	YMR210W	YLR131C	YJL215C	YGR211W	YER099C
YPR043W	YOR034C-A	YMR209C	YLR107W	YJL204C	YGR210C	YER090W
YPR042C	YOR025W	YMR198W	YLR106C	YJL203W	YGR208W	YER089C

YPR040W	YOR023C	YMR197C	YLR096W	YJL199C	YER088C-A	YGR207C
YPR039W	YOR017W	YMR191W	YLR095C	YJL192C	YER066C-A	YGR186W
YPR037C	YOR016C	YMR190C	YLR090W	YJL191W	YGR185C	YER051W
YPR026W	YOL150C	YMR189W	YLR089C	YJL190C	YGR174W-A	YER050C
YPR025C	YOL147C	YMR188C	YLR088W	YJL185C	YER039C-A	YGR174C
YPR024W	YOL146W	YMR184W	YLR087C	YJL180C	YGR173W	YER039C
YPR023C	YOL145C	YMR183C	YLR086W	YJL179W	YER038W-A	YGR172C
YPL272C	YOL144W	YMR182W-A	YLR085C	YJL142C	YGR149W	YER035W
YPL271W	YOL143C	YMR180C	YLR079W	YJL141C	YGR148C	YER034W
YPL246C	YOL142W	YMR167W	YLR078C	YJL140W	YGR131W	YER032W
YPL245W	YOL138C	YMR166C	YLR075W	YJL127C	YGR130C	YER031C
YPL244C	YOL137W	YMR160W	YLR072W	YJL126W	YGR129W	YER030W
YPL243W	YOL135C	YMR159C	YLR071C	YJL123C	YGR128C	YER029C
YPL236C	YOL133W	YMR158W	YLR068W	YJL122W	YGR105W	YER019W
YPL235W	YOL114C	YMR157C	YLR067C	YJL119C	YGR104C	YER018C
YPL234C	YOL113W	YMR143W	YLR066W	YJL103C	YGR103W	YER009W
YPL233W	YOL111C	YMR142C	YLR065C	YJL102W	YGR102C	YER008C
YPL217C	YOL110W	YMR138W	YLR052W	YJL087C	YGR099W	YER004W
YPL216W	YOL108C	YMR137C	YLR051C	YJL086C	YGR098C	YER003C
YPL214C	YOL107W	YMR128W	YLR041W	YJL085W	YGR096W	YEL046C
YPL213W	YOL099C	YMR127C	YLR021W	YJL074C	YGR095C	YEL045C
YPL212C	YOL098C	YMR115W	YLR020C	YJL073W	YGR082W	YEL039C
YPL211W	YOL094C	YMR114C	YLR019W	YJL072C	YGR081C	YEL038W
YPL209C	YOL093W	YMR090W	YLR018C	YJL071W	YGR072W	YEL028W
YPL208W	YOL087C	YMR089C	YLR017W	YJL063C	YGR071C	YEL025C
YPL199C	YOL086W-A	YMR079W	YLR016C	YJL062W-A	YGR048W	YEL024W
YPL198W	YOL080C	YMR078C	YLR015W	YJL056C	YGR047C	YEL023C
YPL184C	YOL078W	YMR061W	YLR014C	YJL055W	YGR031W	YEL022W
YMR060C	YOL077C	YPL183W-A	YLR009W	YJL047C	YGR030C	YEL019C
YPL176C	YOL076W	YMR054W	YLR008C	YJL046W	YGR026W	YEL018W
YPL175W	YOL073C	YMR053C	YLR005W	YJL035C	YGR025W	YEL017W
YPL174C	YOL072W	YMR052C-A	YLR004C	YJL034W	YEL017C-A	YGR021W
YPL173W	YOL070C	YMR048W	YLL062C	YJL031C	YGR020C	YEL016C
YPL169C	YOL069W	YMR047C	YLL061W	YJL030W	YGR016W	YEL015W
YPL168W	YOL062C	YMR042W	YLL036C	YJL019W	YGR015C	YEL005C
YPL167C	YOL061W	YMR041C	YLL035W	YJL016W	YGR006W	YEL004W
YPL166W	YOL060C	YMR035W	YLL034C	YIR037W	YGR005C	YDR531W
YPL137C	YOL059W	YMR034C	YLL033W	YIR036W-A	YGR003W	YDR530C
YPL135W	YOL038W	YMR027W	YLL022C	YIR033W	YDR524C-B	YGR002C
YPL119C	YOL038C-A	YMR026C	YLL021W	YIR032C	YDR524C-A	YGL246C
YPL118W	YOL031C	YMR022W	YLL017W	YIR020C	YGL245W	YDR515W
YPL114W	YOL030W	YMR021C	YLL016W	YIR019C	YGL242C	YDR514C
YPL105C	YOL018C	YML128C	YLL009C	YIR012W	YGL241W	YDR513W
YPL104W	YOL017W	YML127W	YLL008W	YIR011C	YGL229C	YDR512C
YPL096W	YOL016C	YML106W	YLL007C	YIR009W	YGL228W	YDR501W
YML107C	YOL015W	YPL096C-A	YLL006W	YIR008C	YGL226W	YDR500C
YPL094C	YOL012C	YML099C	YKR101W	YIR003W	YGL226C-A	YDR492W
YPL093W	YOL011W	YML098W	YKR100C	YIR002C	YGL221C	YDR490C

YPL077C	YOL005C	YML097C	YKR099W	YIL154C	YGL220W	YDR489W
YPL076W	YOL004W	YML096W	YKR098C	YIL153W	YGL219C	YDR488C
YPL071C	YNR043W	YML066C	YKR090W	YIL149C	YGL216W	YDR483W
YPL070W	YNR041C	YML065W	YKR089C	YIL148W	YGL213C	YDR482C
YPL067C	YNR040W	YML061C	YKR086W	YIL143C	YGL212W	YDR469W
YPL066W	YNR039C	YML060W	YKR085C	YIL142W	YGL206C	YDR468C
YPL064C	YNR038W	YML049C	YKR082W	YIL131C	YGL205W	YDR462W
YPL063W	YNR037C	YML048W	YKR081C	YIL130W	YDR461C-A	YGL200C
YPL053C	YNR032W	YML037C	YKR064W	YIL127C	YGL198W	YDR460W
YPL052W	YNR031C	YML036W	YKR063C	YIL126W	YGL193C	YDR459C
YPL049C	YNR030W	YML035C	YKR056W	YIL107C	YGL191W	YDR448W
YPL048W	YNR029C	YML034W	YKR055W	YLR387C	YKL106W	YOR275C
YPL046C	YNR026C	YML032C	YKR054C	YLR386W	YKL106C-A	YOR262W
YPL045W	YNR025C	YML031W	YKR050W	YLR385C	YKL099C	YOR261C
YPL040C	YNR023W	YML025C	YKR048C	YLR384C	YKL098W	YOR260W
YPL039W	YNR022C	YML024W	YKR047W	YLR383W	YKL074C	YOR259C
YML023C	YNR021W	YPL038W-A	YKR044W	YLR382C	YKL073W	YOR257W
YPL023C	YNR020C	YML022W	YKR043C	YLR375W	YKL052C	YOR256C
YPL022W	YNR008W	YML021C	YKR025W	YLR373C	YKL051W	YOR252W
YPL019C	YNR007C	YML020W	YKR024C	YLR371W	YKL048C	YOR251C
YPL018W	YNR004W	YML015C	YKR023W	YLR370C	YKL047W	YOR244W
YPL017C	YNR003C	YML014W	YKR022C	YLR363W-A	YKL036C	YOR243C
YPL016W	YNL315C	YML011C	YKR020W	YLR363C	YKL024C	YOR222W
YPL013C	YNL314W	YML010W	YKR019C	YLR349W	YKL023C-A	YOR221C
YPL012W	YNL313C	YML009W-B	YKR007W	YLR324W	YKL018W	YOR220W
YPL011C	YNL312W	YML007W	YKR006C	YLR323C	YKL018C-A	YOR219C
YPL010W	YNL310C	YLRCdelta27	YKR002W	YLR320W	YKL016C	YOR217W
YDR447C	YBR250W	YDR182W-A	YDL033C	YBR130C	YDR291W	YCL002C
YDR182W	YBR249C	YDR444W	YDL031W	YBR108W	YDR289C	YCL001W
YDR443C	YBR246W	YDR179W-A	YDL016C	YBR107C	YDR265W	YBR289W
YDR179C	YBR245C	YDR433W	YDL015C	YBR103W	YDR264C	YBR288C
YDR176W	YBR237W	YDR432W	YDL012C	YBR102C	YDR254W	YBR284W
YDR175C	YBR236C	YDR430C	YDL010W	YBR096W	YDR253C	YBR283C
YDR167W	YBR235W	YDR425W	YDL002C	YBR095C	YDR250C	YBR278W
YDR166C	YBR234C	YDR424C	YDL001W	YBR071W	YDR249C	YBR276C
YDR162C	YBR233W	YCR095W-A	YDR412W	YBR070C	YDR244W	YBR274W
YDR152W	YBR231C	YDR411C	YCR095C	YBR030W	YDR243C	YBR273C
YDR151C	YDR405W	YBR230W-A	YCR093W	YBR029C	YDR220C	YBR271W
YDR140W	YBR230C	YDR404C	YCR092C	YBR005W	YDR219C	YBR270C
YDR139C	YBR227C	YDR392W	YCR079W	YBR004C	YDR204W	YBR265W
YDR124W	YBR226C	YDR391C	YCR077C	YBR003W	YDR202C	YBR264C
YDR123C	YBR225W	YCR075W-A	YDR388W	YBR002C	YDR201W	YBR263W
YDR121W	YBR224W	YCR073W-A	YCR075C	YBL103C	YDR200C	YBR262C
YDR120C	YBR223C	YDR118W-A	YDR387C	YBL102W	YDR197W	YBR259W
YDR382W	YBR217W	YDR381C-A	YCR073C	YBL093C	YDR196C	YBR258C
YDR118W	YBR216C	YDR376W	YCR060W	YBL091C	YDR191W	YBR255W
YDR117C	YBR212W	YDR375C	YCR059C	YBL090W	YDR190C	YBR254C
YDR110W	YBR211C	YDR373W	YCR048W	YBL072C	YIL106W	YGL175C

YDR109C	YBR206W	YDR372C	YCR047C	YBL071C-B	YIL093C	YGL174W
YDR108W	YBR205W	YCR045W-A	YDR371W	YBL061C	YIL092W	YGL173C
YDR107C	YBR204C	YDR370C	YCR036W	YBL060W	YIL091C	YGL172W
YDR103W	YBR199W	YDR363W	YCR035C	YBL057C	YIL090W	YGL167C
YDR101C	YBR198C	YDR362C	YCR020W-B	YBL056W	YIL077C	YGL166W
YDR082W	YBR175W	YDR358W	YCR020C-A	YBL051C	YIL076W	YGL161C
YDR081C	YBR173C	YDR357C	YCR020C	YBL050W	YIL065C	YGL160W
YDR080W	YBR171W	YDR351W	YCR003W	YBL047C	YIL064W	YGL154C
YDR079W	YBR170C	YDR079C-A	YCR002C	YBL046W	YIL051C	YGL153W
YDR350C	YBR165W	YDR347W	YCL064C	YBL039C-A	YIL050W	YGL150C
YDR078C	YBR164C	YDR346C	YCL063W	YBL039C	YIL047C-A	YGL148W
YDR068W	YBR151W	YDR334W	YCL059C	YBL038W	YIL047C	YGL142C
YDR067C	YBR150C	YDR333C	YCL058W-A	YBL033C	YIL046W-A	YGL141W
YDR057W	YBR142W	YDR330W	YCL052C	YBL032W	YIL042C	YGL140C
YDR056C	YDR329C	YBR141W-A	YCL051W	YBL028C	YIL041W	YGL139W
YDR031W	YBR141C	YDR321W	YCL037C	YBL027W	YIL020C-A	YGL131C
YDR030C	YBR139W	YDR320C-A	YCL036W	YBL021C	YIL020C	YGL130W
YDR020C	YBR138C	YDR316W	YCL029C	YBL020W	YIL019W	YGL127C
YDR019C	YBR135W	YDR315C	YCL028W	YBL014C	YHR200W	YGL126W
YDR007W	YBR133C	YDR296W	YCL008C	YBL013W	YHR199C-A	YGL112C
YDR006C	YBR131W	YDR295C	YCL005W-A	YBL010C	YHR199C	YGL111W
YDR002W	YAL013W	YHR148W	YGL043W	YBL009W	YHR194W	YGL107C
YDR001C	YAL005C	YHR147C	YGL031C	YAR008W	YHR193C	YGL106W
YDL238C	YAL003W	YHR134W	YGL030W	YAR007C	YHR192W	YGL094C
YDL237W	YLR423C	YHR133C	YGL023C	YAR003W	YHR191C	YGL093W
YDL234C	YLR422W	YHR132W-A	YGL022W	YAR002C-A	YHR189W	YGL091C
YDL233W	YLR421C	YHR132C	YGL020C	YAL043C	YHR188C	YGL090W
YDL231C	YLR419W	YHR119W	YGL019W	YAL042W	YHR187W	YGL087C
YDL230W	YLR418C	YHR118C	YGL018C	YAL036C	YHR186C	YGL086W
YDL226C	YLR410W	YHR117W	YGL017W	YAL035W	YHR167W	YGL072C
YDL225W	YLR409C	YHR116W	YGL013C	YAL034C	YHR166C	YGL070C
YDL220C	YLR403W	YHR115C	YGL012W	YAL033W	YHR159W	YGL068W
YDL219W	YLR402W	YHR113W	YGL011C	YAL031C	YHR158C	YGL064C
YDL213C	YLR390W	YHR112C	YGL010W	YAL030W	YHR154W	YGL063W
YDL212W	YLR389C	YHR108W	YGL003C	YAL029C	YHR153C	YGL061C
YDL209C	YLR388W	YHR107C	YGL002W	YAL028W	YHR152W	YGL060W
YDL208W	YKL176C	YHR083W	YFR042W	YAL019W-A	YHR151C	YGL048C
YDL205C	YKL175W	YHR082C	YFR041C	YAL019W	YHR150W	YGL047W
YDL204W	YKL174C	YHR078W	YFR040W	YAL014C	YHR149C	YGL044C
YDL203C	YKL173W	YHR077C	YFR039C	YDL097C	YDL095W	YDL090C
YDL202W	YDL197C	YDL195W	YDL190C	YDL189W	YDL187C	YDL171C
YDL170W	YDL166C	YDL165W	YDL160C	YDL159W	YDL158C	YDL157C
YDL153C	YDL150W	YDL148C	YDL147W	YDL144C	YDL143W	YDL142C
YDL141W	YDL121C	YDL120W	YDL103C	YDL102W	YDL100C	YDL099W
YDL089W	YDL083C	YDL082W	YDL076C	YDL075W	YDL074C	YDL073W
YDL065C	YDL064W	YDL061C	YDL060W	YDL052C	YDL051W	YNL181W
YNL166C	YNL165W	YNL162W-A	YNL163C	YNL159C	YNL158W	YNL156C
YNL155W	YNL154C	YNL153C	YNL152W	YNL148C	YNL147W	YNL143C

YNL140C

References

Bahar, R, CH Hartmann, KA Rodriguez, *et al* . 2006. Increased cell-to-cell variation in gene expression in ageing mouse heart. Nature 441:1011-1014.

Bailey, JA, EE Eichler. 2006. Primate segmental duplications: crucibles of evolution, diversity and disease. Nat Rev Genet 7:552-564.

Balaji, S, MM Babu, LM Iyer, NM Luscombe, L Aravind. 2006. Comprehensive analysis of combinatorial regulation using the transcriptional regulatory network of yeast. J Mol Biol 360:213-227.

Bar-Even, A, J Paulsson, N Maheshri, M Carmi, E O'Shea, Y Pilpel, N Barkai. 2006. Noise in protein expression scales with natural protein abundance. Nat Genet 38:636-643.

Basehoar, AD, SJ Zanton, BF Pugh. 2004. Identification and distinct regulation of yeast TATA box-containing genes. Cell 116:699-709.

Batada, NN, LD Hurst. 2007. Evolution of chromosome organization driven by selection for reduced gene expression noise. Nat Genet 39:945-949.

Batada, NN, AO Urrutia, LD Hurst. 2007. Chromatin remodelling is a major source of coexpression of linked genes in yeast. Trends Genet 23:480-484.

Becskei, A, BB Kaufmann, A van Oudenaarden. 2005. Contributions of low molecule number and chromosomal positioning to stochastic gene expression. Nat Genet 37:937-944.

Bergman, A, ML Siegal. 2003. Evolutionary capacitance as a general feature of complex gene networks. Nature 424:549-552.

Birnbaum, K, DE Shasha, JY Wang, JW Jung, GM Lambert, DW Galbraith, PN Benfey. 2003. A gene expression map of the Arabidopsis root. Science 302:1956-1960.

Blake, WJ, G Balazsi, MA Kohanski, FJ Isaacs, KF Murphy, Y Kuang, CR Cantor, DR Walt, JJ Collins. 2006. Phenotypic consequences of promoter-mediated transcriptional noise. Mol Cell 24:853-865.

Blake, WJ, KA M, CR Cantor, JJ Collins. 2003. Noise in eukaryotic gene expression. Nature 422:633-637.

Blomme, T, K Vandepoele, S De Bodt, C Simillion, S Maere, Y Van de Peer. 2006. The gain and loss of genes during 600 million years of vertebrate evolution. Genome Biol 7:R43.

Bonnefoy, N, C Remacle, TD Fox. 2007. Genetic transformation of Saccharomyces cerevisiae and Chlamydomonas reinhardtii mitochondria. Methods Cell Biol 80:525-548.

Bonnet, E, M Tatari, A Joshi, T Michoel, K Marchal, G Berx, Y Van de Peer. 2010. Module network inference from a cancer gene expression data set identifies microRNA regulated modules. PLoS One 5:e10162.

Breitkreutz, BJ, C Stark, T Reguly, *et a l*. 2008. The BioGRID Interaction Database: 2008 update. Nucleic Acids Res 36:D637-640.

Brosius, J. 2005. Waste not, want not--transcript excess in multicellular eukaryotes. Trends Genet 21:287-288.

Brunet, FG, H Roest Crollius, M Paris, JM Aury, P Gibert, O Jaillon, V Laudet,

M Robinson-Rechavi. 2006. Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. Mol Biol Evol 23:1808-1816.

Brunner, E, U Munzel. 2000. The nonparametric Behrens-Fisher problem: Asymptotic theory and a small-sample approximation. Biometrical Journal 42:17-25.

Byrne, KP, KH Wolfe. 2005. The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. Genome Res 15:1456-1461.

Carlson, MR, B Zhang, Z Fang, PS Mischel, S Horvath, SF Nelson. 2006. Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks. BMC Genomics 7:40.

Chang, HH, M Hemberg, M Barahona, DE Ingber, S Huang. 2008. Transcriptome-wide noise controls lineage choice in mammalian progenitor cells. Nature 453:544-547.

Chen, WH, J de Meaux, MJ Lercher. 2010. Co-expression of neighbouring genes in Arabidopsis: separating chromatin effects from direct interactions. BMC Genomics 11:178.

Cherry, JM, C Adler, C Ball, *et al* . 1998. SGD: Saccharomyces Genome Database. Nucleic Acids Res 26:73-79.

Cheung, J, X Estivill, R Khaja, JR MacDonald, K Lau, LC Tsui, SW Scherer. 2003. Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. Genome Biol 4:R25.

Choi, JK, SC Kim, J Seo, S Kim, J Bhak. 2007. Impact of transcriptional properties on essentiality and evolutionary rate. Genetics 175:199-206.

Choi, JK, YJ Kim. 2008. Epigenetic regulation and the variability of gene expression. Nat Genet 40:141-147.

Choi, JK, YJ Kim. 2009. Intrinsic variability of gene expression encoded in nucleosome positioning sequences. Nat Genet 41:498-503.

Choi, PJ, L Cai, K Frieda, XS Xie. 2008. A stochastic single-molecule event triggers phenotype switching of a bacterial cell. Science 322:442-446.

Cohen, BA, RD Mitra, JD Hughes, GM Church. 2000. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. Nat Genet 26:183-186.

Collins, PJ, Y Kobayashi, L Nguyen, ND Trinklein, RM Myers. 2007. The ets-related transcription factor GABP directs bidirectional transcription. PLoS Genet 3:e208.

Conant, GC, KH Wolfe. 2007. Increased glycolytic flux as an outcome of whole-genome duplication in yeast. Mol Syst Biol 3:129.

Cook, DL, AN Gerber, SJ Tapscott. 1998. Modeling stochastic gene expression: implications for haploinsufficiency. Proc Natl Acad Sci U S A 95:15641-15646.

Cooper, TF, AP Morby, A Gunn, D Schneider. 2006. Effect of random and hub gene disruptions on environmental and mutational robustness in Escherichia coli. BMC Genomics 7:237–47.

Core, LJ, JJ Waterfall, JT Lis. 2008. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science 322:1845-1848.

Costanzo, M, A Baryshnikova, J Bellay, *et al*. The genetic landscape of a cell. Science 327:425-431.

Costanzo, M, A Baryshnikova, J Bellay, *et al*. 2010. The genetic landscape of a cell. Science 327:425-431.

Datta, A, S Jinks-Robertson. 1995. Association of increased spontaneous mutation rates with high levels of transcription in yeast. Science 268:1616-1619.

de Lichtenberg, U, LJ Jensen, S Brunak, P Bork. 2005. Dynamic complex formation during the yeast cell cycle. Science 307:724-727.

Deutschbauer, AM, DF Jaramillo, M Proctor, J Kumm, ME Hillenmeyer, RW Davis, C Nislow, G Giaever. 2005. Mechanisms of haploinsufficiency revealed by genome-wide profiling in yeast. Genetics 169:1915-1925.

Dezso, Z, ZN Oltvai, AL Barabasi. 2003. Bioinformatics analysis of experimentally determined protein complexes in the yeast Saccharomyces cerevisiae. Genome Res 13:2450-2454.

Dixon, SJ, M Costanzo, A Baryshnikova, B Andrews, C Boone. 2009. Systematic mapping of genetic interaction networks. Annu Rev Genet 43:601-625.

Dixon, SJ, Y Fedyshyn, JL Koh, *et a l*. 2008. Significant conservation of synthetic lethal genetic interaction networks between distantly related eukaryotes. Proc Natl Acad Sci U S A 105:16653-16658.

Drummond, DA, JD Bloom, C Adami, CO Wilke, FH Arnold. 2005. Why highly expressed proteins evolve slowly. Proc Natl Acad Sci U S A 102:14338-14343.

Drummond, DA, A Raval, CO Wilke. 2006. A single determinant dominates the rate of yeast protein evolution. Mol Biol Evol 23:327-337.

Drummond, DA, CO Wilke. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell 134:341-352.

Ebisuya, M, T Yamamoto, M Nakajima, E Nishida. 2008. Ripples from neighbouring transcription. Nat Cell Biol 10:1106–1113.

Edwards, JS, BO Palsson. 1999. Systems properties of the Haemophilus influenzae Rd metabolic genotype. J Biol Chem 274:17410-17416.

Elowitz, MB, AJ Levine, ED Siggia, PS Swain. 2002. Stochastic gene expression in a single cell. Science 297:1183-1186.

Falb, M, K Muller, L Konigsmaier, T Oberwinkler, P Horn, S von Gronau, O Gonzalez, F Pfeiffer, E Bornberg-Bauer, D Oesterhelt. 2008. Metabolism of halophilic archaea. Extremophiles 12:177-196.

Fall, S, A Mercier, F Bertolla, A Calteau, L Gueguen, G Perriere, TM Vogel, P Simonet. 2007. Horizontal Gene Transfer Regulation in Bacteria as a "spandrel "of DNA Repair Mechanisms. PLoS One 2:e1055.

Feist, AM, BO Palsson. 2008. The growing scope of applications of genome-scale metabolic reconstructions using Escherichia coli. Nat Biotechnol 26:659-667.

Field, Y, N Kaplan, Y Fondufe-Mittendorf, IK Moore, E Sharon, Y Lubling, J Widom, E Segal. 2008. Distinct modes of regulation by chromatin encoded through nucleosome positioning signals. PLoS Comput Biol 4:e1000216.

Florens, L, MP Washburn, JD Raine, *et al* . 2002. A proteomic view of the Plasmodium falciparum life cycle. Nature 419:520-526.

Force, A, M Lynch, FB Pickett, A Amores, YL Yan, J Postlethwait. 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151:1531-1545.

Forster, J, I Famili, P Fu, BO Palsson, J Nielsen. 2003. Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. Genome Res 13:244-253.

Fraser, HB, AE Hirsh, G Giaever, J Kumm, MB Eisen. 2004. Noise minimization in eukaryotic gene expression. PLoS Biol 2:e137.

Fraser, HB, AE Hirsh, LM Steinmetz, C Scharfe, MW Feldman. 2002. Evolutionary rate in the protein interaction network. Science 296:750-752.

Fraser, HB, JB Plotkin. 2007. Using protein complexes to predict phenotypic effects of gene mutation. Genome Biol 8:R252.

Fraser, HB, DP Wall, AE Hirsh. 2003. A simple dependence between protein evolution rate and the number of protein-protein interactions. BMC Evol Biol 3:11.

Friedman, R, AL Hughes. 2001. Gene duplication and the structure of eukaryotic genomes. Genome Res 11:373-381.

Fukuoka, Y, H Inaoka, IS Kohane. 2004. Inter-species differences of co-expression of neighbouring genes in eukaryotic genomes. BMC Genomics 5:4.

Ghaemmaghami, S, WK Huh, K Bower, RW Howson, A Belle, N Dephoure, EK O'Shea, JS Weissman. 2003. Global analysis of protein expression in yeast. Nature 425:737-741.

Goffeau, A, BG Barrell, H Bussey, *et al*. 1996. Life with 6000 genes. Science 274:546, 563-547.

Gordon, JL, KP Byrne, KH Wolfe. 2009. Additions, losses, and rearrangements on the evolutionary route from a reconstructed ancestor to the modern Saccharomyces cerevisiae genome. PLoS Genet 5:e1000485.

Grömping, U. 2006. Relative Importance for Linear Regression in R: The Package relaimpo. J Statist Soft 17:1-27.

Gu, X, Z Zhang, W Huang. 2005. Rapid evolution of expression and regulatory divergences after yeast gene duplication. Proc Natl Acad Sci U S A 102:707-712.

Gu, Z, LM Steinmetz, X Gu, C Scharfe, RW Davis, WH Li. 2003. Role of duplicate genes in genetic robustness against null mutations. Nature 421:63-66.

Hahn, MW, AD Kern. 2005. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. Mol Biol Evol 22:803-806.

Hakes, L, SC Lovell, SG Oliver, DL Robertson. 2007. Specificity in protein interactions and its relationship with sequence diversity and coevolution. Proc Natl Acad Sci U S A 104:7999-8004.

Hart, GT, I Lee, ER Marcotte. 2007. A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality. BMC Bioinformatics 8:236.

Hartman, JLt, B Garvik, L Hartwell. 2001. Principles for the buffering of genetic variation. Science 291:1001-1004.

Hartwell, LH, JJ Hopfield, S Leibler, AW Murray. 1999. From molecular to modular cell biology. Nature 402:C47-52.

Hermisson, J, GP Wagner. 2004. The population genetic theory of hidden variation and genetic robustness. Genetics 168:2271-2284.

Hirsh, AE, HB Fraser, DP Wall. 2005. Adjusting for selection on synonymous

sites in estimates of evolutionary distance. Mol Biol Evol 22:174-177.
Holstege, FC, EG Jennings, JJ Wyrick, TI Lee, CJ Hengartner, MR Green, TR Golub, ES Lander, RA Young. 1998. Dissecting the regulatory circuitry of a eukaryotic genome. Cell 95:717-728.
Houston, P, PJ Simon, JR Broach. 2004. The Saccharomyces cerevisiae recombination enhancer biases recombination during interchromosomal mating-type switching but not in interchromosomal homologous recombination. Genetics 166:1187-1197.
Hui, W, YR Gel, JL Gastwirth. 2008. lawstat: An R Package for Law, Public Policy and Biostatistics. J Stat Soft 28:1-26.
Hurst, LD, C Pal, MJ Lercher. 2004. The evolutionary dynamics of eukaryotic gene order. Nat Rev Genet 5:299-310.
Hurst, LD, EJ Williams, C Pal. 2002. Natural selection promotes the conservation of linkage of co-expressed genes. Trends Genet 18:604-606.
Ihaka, R, R Gentleman. 1996. R: A Language for Data Analysis and Graphics. J Comput Graph Stat 5:299-314.
Ihmels, J, S Bergmann, M Gerami-Nejad, I Yanai, M McClellan, J Berman, N Barkai. 2005. Rewiring of the yeast transcriptional network through the evolution of motif usage. Science 309:938-940.
Ito, T, T Chiba, R Ozawa, M Yoshida, M Hattori, Y Sakaki. 2001. A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci U S A 98:4569-4574.
Jaillon, O, JM Aury, B Noel, *et al* . 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449:463-467.
Jensen, LJ, M Kuhn, M Stark, *et al*. 2009. STRING 8--a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res 37:D412-416.
Jordan, IK, YI Wolf, EV Koonin. 2003. No simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors tend to evolve slowly. BMC Evol Biol 3:1.
Jovelin, R, PC Phillips. 2009. Evolutionary rates and centrality in the yeast gene regulatory network. Genome Biol 10:R35.
Joy, MP, A Brock, DE Ingber, S Huang. 2005. High-betweenness proteins in the yeast protein interaction network. J Biomed Biotechnol 2005:96-103.
Juan, D, F Pazos, A Valencia. 2008. Co-evolution and co-adaptation in protein networks. FEBS Lett 582:1225-1230.
Kaern, M, TC Elston, WJ Blake, JJ Collins. 2005. Stochasticity in gene expression: from theories to phenotypes. Nat Rev Genet 6:451-464.
Kafri, R, A Bar-Even, Y Pilpel. 2005. Transcription control reprogramming in genetic backup circuits. Nat Genet 37:295-299.
Kann, MG, BA Shoemaker, AR Panchenko, TM Przytycka. 2009. Correlated evolution of interacting proteins: looking behind the mirrortree. J Mol Biol 385:91-98.
Kar, S, WT Baumann, MR Paul, JJ Tyson. 2009. Exploring the roles of noise in the eukaryotic cell cycle. Proc Natl Acad Sci U S A 106:6471-6476.
Keller, PJ, M Knop. 2009. Evolution of mutational robustness in the yeast genome: a link to essential genes and meiotic recombination hotspots. PLoS Genet 5:e1000533.

Kellis, M, BW Birren, ES Lander. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. Nature 428:617-624.

Kensche, PR, M Oti, BE Dutilh, MA Huynen. 2008. Conservation of divergent transcription in fungi. Trends Genet 24:207-211.

Kim, TH, B Ren. 2006. Genome-wide analysis of protein-DNA interactions. Annu Rev Genomics Hum Genet 7:81-102.

Kitano, H. 2004. Biological robustness. Nature Rev Genet 5:826-837.

Kovacs, K, LD Hurst, B Papp. 2009. Stochasticity in protein levels drives colinearity of gene order in metabolic operons of Escherichia coli. PLoS Biol 7:e1000115.

Kruglyak, S, H Tang. 2000. Regulation of adjacent yeast genes. Trends Genet 16:109-111.

Kussell, E, S Leibler. 2005. Phenotypic diversity, population growth, and information in fluctuating environments. Science 309:2075-2078.

Lander, ESLM LintonB Birren, *et al*. 2001. Initial sequencing and analysis of the human genome. Nature 409:860-921.

Lehar, J, A Krueger, G Zimmermann, A Borisy. 2008. High-order combination effects and biological robustness. Mol Syst Biol 4:215.

Lehner, B. 2008. Selection to minimise noise in living systems and its implications for the evolution of gene expression. Mol Syst Biol 4:170.

Lercher, MJ, T Blumenthal, LD Hurst. 2003. Coexpression of neighbouring genes in Caenorhabditis elegans is mostly due to operons and duplicate genes. Genome Res 13:238-243.

Lercher, MJ, LD Hurst. 2006. Co-expressed yeast genes cluster over a long range but are not regularly spaced. J Mol Biol 359:825-831.

Lercher, MJ, AO Urrutia, LD Hurst. 2002. Clustering of housekeeping genes provides a unified model of gene order in the human genome. Nat Genet 31:180-183.

Levy, SF, ML Siegal. 2008. Network Hubs Buffer Environmental Variation in Saccharomyces cerevisiae. PLoS Biol 6:2588-2604.

Li, B, M Carey, JL Workman. 2007. The role of chromatin during transcription. Cell 128:707-719.

Liang, H, WH Li. 2009. Lowly expressed human microRNA genes evolve rapidly. Mol Biol Evol 26:1195-1198.

Liao, BY, J Zhang. 2008. Coexpression of linked genes in Mammalian genomes is generally disadvantageous. Mol Biol Evol 25:1555-1565.

Lin, JM, PJ Collins, ND Trinklein, Y Fu, H Xi, RM Myers, Z Weng. 2007a. Transcription factor binding and modified histones in human bidirectional promoters. Genome Res 17:818-827.

Lin, YS, WL Hsu, JK Hwang, WH Li. 2007b. Proportion of solvent-exposed amino acids in a protein and rate of protein evolution. Mol Biol Evol 24:1005-1011.

Liti, G, DM Carter, AM Moses, *et al*. 2009. Population genomics of domestic and wild yeasts. Nature 458:337-341.

Lopez-Maury, L, S Marguerat, J Bahler. 2008. Tuning gene expression to changing environments: from rapid responses to evolutionary adaptation. Nat Rev Genet 9:583-593.

Lovdok, L, K Bentele, N Vladimirov, A Muller, FS Pop, D Lebiedz, M Kollmann, V Sourjik. 2009. Role of translational coupling in robustness

of bacterial chemotaxis pathway. PLoS Biol 7:e1000171.
Luo, F, Y Yang, J Zhong, H Gao, L Khan, DK Thompson, J Zhou. 2007. Constructing gene co-expression networks and predicting functions of unknown genes by random matrix theory. BMC Bioinformatics 8:299.
Lynch, M, JS Conery. 2000. The evolutionary fate and consequences of duplicate genes. Science 290:1151-1155.
Lynch, M, A Force. 2000. The probability of duplicate gene preservation by subfunctionalization. Genetics 154:459-473.
Maamar, H, A Raj, D Dubnau. 2007. Noise in gene expression determines cell fate in Bacillus subtilis. Science 317:526-529.
Maere, S, S De Bodt, J Raes, T Casneuf, M Van Montagu, M Kuiper, Y Van de Peer. 2005. Modeling gene and genome duplications in eukaryotes. Proc Natl Acad Sci U S A 102:5454-5459.
Martens, JA, L Laprade, F Winston. 2004. Intergenic transcription is required to repress the Saccharomyces cerevisiae SER3 gene. Nature 429:571-574.
Masel, J. 2005. Evolutionary capacitance may be favored by natural selection. Genetics 170:1359-1371.
Masel, J, ML Siegal. 2009. Robustness: mechanisms and consequences. Trends Genet 25:395-403.
Michalak, P. 2008. Coexpression, coregulation, and cofunctionality of neighbouring genes in eukaryotic genomes. Genomics 91:243-248.
Milton, CC, B Huynh, P Batterham, SL Rutherford, AA Hoffmann. 2003. Quantitative trait symmetry independent of Hsp90 buffering: distinct modes of genetic canalization and developmental stability. Proc Natl Acad Sci U S A 100:13396-13401.
Mitchell-Olds, T, CA Knight. 2002. Evolution. Chaperones as buffering agents? Science 296:2348-2349.
Murphy, KF, G Balazsi, JJ Collins. 2007. Combinatorial promoter design for engineering noisy gene expression. Proc Natl Acad Sci U S A 104:12726-12731.
Neil, H, C Malabat, Y d'Aubenton-Carafa, Z Xu, LM Steinmetz, A Jacquier. 2009. Widespread bidirectional promoters are the major source of cryptic transcripts in yeast. Nature 457:1038-1042.
Newman, JR, S Ghaemmaghami, J Ihmels, DK Breslow, M Noble, JL DeRisi, JS Weissman. 2006. Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. Nature 441:840-846.
Niida, A, AD Smith, S Imoto, H Aburatani, MQ Zhang, T Akiyama. 2009. Gene set-based module discovery in the breast cancer transcriptome. BMC Bioinformatics 10:71.
Novick, A, M Weiner. 1957. Enzyme induction as an all-or-none phenomenon. Proc. Natl Acad. Sci. USA 43:553-566.
Ohno, S. Evolution by Gene Duplication (George Allen and Unwin, London, 1970).
Ozbudak, EM, M Thattai, I Kurtser, AD Grossman, A van Oudenaarden. 2002. Regulation of noise in the expression of a single gene. Nat Genet 31:69-73.
Pal, C, LD Hurst. 2003. Evidence for co-evolution of gene order and recombination rate. Nat Genet 33:392-395.
Pal, C, B Papp, LD Hurst. 2001. Highly expressed genes in yeast evolve slowly. Genetics 158:927-931.

Pal, C, B Papp, MJ Lercher. 2006. An integrated view of protein evolution. Nat Rev Genet 7:337-348.

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

Papp, B, B Teusink, RA Notebaart. 2009. A critical view of metabolic network adaptations. HFSP J 3:24-35.

Paterson, AH, M Freeling, H Tang, X Wang. 2010. Insights from the comparison of plant genome sequences. Annu Rev Plant Biol 61:349-372.

Piskur, J. 2001. Origin of the duplicated regions in the yeast genomes. Trends Genet 17:302-303.

Poyatos, JF, LD Hurst. 2007. The determinants of gene order conservation in yeasts. Genome Biol 8:R233.

Preker, P, J Nielsen, S Kammler, S Lykke-Andersen, MS Christensen, CK Mapendano, MH Schierup, TH Jensen. 2008. RNA exosome depletion reveals transcription upstream of active human promoters. Science 322:1851-1854.

Preker, P, J Nielsen, MH Schierup, TH Jensen. 2009. RNA polymerase plays both sides: vivid and bidirectional transcription around and upstream of active promoters. Cell Cycle 8:1106-1107.

Pu, S, J Wong, B Turner, E Cho, SJ Wodak. 2009. Up-to-date catalogues of yeast protein complexes. Nucleic Acids Res 37:825-831.

Queitsch, C, TA Sangster, S Lindquist. 2002. Hsp90 as a capacitor of phenotypic variation. Nature 417:618-624.

Raj, A, CS Peskin, D Tranchina, DY Vargas, S Tyagi. 2006. Stochastic mRNA synthesis in mammalian cells. PLoS Biol 4:e309.

Raj, A, A van Oudenaarden. 2008. Nature, nurture, or chance: stochastic gene expression and its consequences. Cell 135:216-226.

Raj, A, A van Oudenaarden. 2009. Single-molecule approaches to stochastic gene expression. Annu Rev Biophys 38:255-270.

Ramsey, S, A Ozinsky, A Clark, KD Smith, P de Atauri, V Thorsson, D Orrell, H Bolouri. 2006. Transcriptional noise and cellular heterogeneity in mammalian macrophages. Philos Trans R Soc Lond B Biol Sci 361:495-506.

Raser, JM, EK O'Shea. 2004. Control of stochasticity in eukaryotic gene expression. Science 304:1811-1814.

Rattray, AJ, JN Strathern. 2003. Error-prone DNA polymerases: when making a mistake is the only way to get ahead. Annu Rev Genet 37:31-66.

Reymond, A, V Marigo, MB Yaylaoglu, *et al*. 2002. Human chromosome 21 gene expression atlas in the mouse. Nature 420:582-586.

Roy, PJ, JM Stuart, J Lund, SK Kim. 2002. Chromosomal clustering of muscle-expressed genes in Caenorhabditis elegans. Nature 418:975-979.

Rozas, J, R Rozas. 1999. DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics 15:174-175.

Ruan, J, AK Dean, W Zhang. 2010. A general co-expression network-based approach to gene expression analysis: comparison and applications. BMC Syst Biol 4:8.

Rutherford, SL, S Lindquist. 1998. Hsp90 as a capacitor for morphological evolution. Nature 396:336-342.

Saeed, R, CM Deane. 2006. Protein protein interactions, evolutionary rate, abundance and age. BMC Bioinformatics 7:128.

Sangster, TA, N Salathia, S Undurraga, R Milo, K Schelienberg, S Lindquist, C Queitsch. 2008. HSP90 affects the expression of genetic variation and developmental stability in quantitative traits. Proc Natl Acad Sci U S A 105:2963-2968.

Scannell, DR, KP Byrne, JL Gordon, S Wong, KH Wolfe. 2006. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. Nature 440:341-345.

Scannell, DR, AC Frank, GC Conant, KP Byrne, M Woolfit, KH Wolfe. 2007. Independent sorting-out of thousands of duplicated gene pairs in two yeast species descended from a whole-genome duplication. Proc Natl Acad Sci U S A 104:8397-8402.

Seila, AC, JM Calabrese, SS Levine, GW Yeo, PB Rahl, RA Flynn, RA Young, PA Sharp. 2008. Divergent transcription from active promoters. Science 322:1849-1851.

Seila, AC, LJ Core, JT Lis, PA Sharp. 2009. Divergent transcription: A new feature of active promoters. Cell Cycle 8:2557-2564.

Semon, M, L Duret. 2006. Evolutionary origin and maintenance of coexpressed gene clusters in mammals. Mol Biol Evol 23:1715-1723.

Seoighe, C, KH Wolfe. 1999. Yeast genome evolution in the post-genome era. Curr Opin Microbiol 2:548-554.

Shi, Z, CK Derow, B Zhang. 2010. Co-expression module analysis reveals biological processes, genomic gain, and regulatory mechanisms associated with breast cancer progression. BMC Syst Biol 4:74.

Sigal, A, R Milo, A Cohen, N Geva-Zatorsky, Y Klein, Y Liron, N Rosenfeld, T Danon, N Perzov, U Alon. 2006. Variability and memory of protein levels in human cells. Nature 444:643-646.

Singer, GA, AT Lloyd, LB Huminiecki, KH Wolfe. 2005. Clusters of co-expressed genes in mammalian genomes are conserved by natural selection. Mol Biol Evol 22:767-775.

Somel, M, P Khaitovich, S Bahn, S Paabo, M Lachmann. 2006. Gene expression becomes heterogeneous with age. Curr Biol 16:R359-360.

Sopko, R, D Huang, N Preston, *et al*. 2006. Mapping pathways and phenotypes by systematic gene overexpression. Mol Cell 21:319-330.

Spellman, PT, GM Rubin. 2002. Evidence for large domains of similarly expressed genes in the Drosophila genome. J Biol 1:5.

Sugino, RP, H Innan. 2005. Estimating the time to the whole-genome duplication and the duration of concerted evolution via gene conversion in yeast. Genetics 171:63-69.

Suyama, M, P Bork. 2001. Evolution of prokaryotic gene order: genome rearrangements in closely related species. Trends Genet 17:10-13.

Suzuki, Y, HF Nijhout. 2006. Evolution of a polyphenism by genetic accommodation. Science 311:650-652.

Taft, RJ, EA Glazov, N Cloonan, *et a l*. 2009. Tiny RNAs associated with transcription start sites in animals. Nat Genet 41:572-578.

Tarassov, K, V Messier, CR Landry, S Radinovic, MM Serna Molina, I Shames, Y Malitskaya, J Vogel, H Bussey, SW Michnick. 2008. An in vivo map of the yeast protein interactome. Science 320:1465-1470.

Teichmann, SA, MM Babu. 2004. Gene regulatory network growth by

duplication. Nat Genet 36:492-496.
Tillier, ER, RL Charlebois. 2009. The human protein coevolution network. Genome Res 19:1861-1871.
Tirosh, I, N Barkai. 2008. Two strategies for gene regulation by promoter nucleosomes. Genome Res 18:1084-1091.
Tong, AH, G Lesage, GD Bader, *et al* . 2004. Global mapping of the yeast genetic interaction network. Science 303:808-813.
Trinklein, ND, SF Aldred, SJ Hartman, DI Schroeder, RP Otillar, RM Myers. 2004. An abundance of bidirectional promoters in the human genome. Genome Res 14:62-66.
Tsai, HK, PY Huang, CY Kao, D Wang. 2009. Co-Expression of Neighbouring Genes in the Zebrafish (Danio rerio) Genome. Int J Mol Sci 10:3658-3670.
Tuskan, GAS DifazioS Jansson, *et al*. 2006. The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). Science 313:1596-1604.
Uetz, P, L Giot, G Cagney, *et al* . 2000. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature 403:623-627.
van Dam, TJ, B Snel. 2008. Protein complex evolution does not involve extensive network rewiring. PLoS Comput Biol 4:e1000132.
van Noort, V, B Snel, MA Huynen. 2004. The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. EMBO Rep 5:280-284.
Vassar, R, J Ngai, R Axel. 1993. Spatial segregation of odorant receptor expression in the mammalian olfactory epithelium. Cell 74:309-318.
Venter, JCMD AdamsEW Myers, *et a l*. 2001. The sequence of the human genome. Science 291:1304-1351.
Vinogradov, AE. 2010. Systemic factors dominate mammal protein evolution. Proc Biol Sci 277:1403-1408.
Vitkup, D, P Kharchenko, A Wagner. 2006. Influence of metabolic network structure and function on enzyme evolution. Genome Biol 7:R39.
von Mering, C, R Krause, B Snel, M Cornell, SG Oliver, S Fields, P Bork. 2002. Comparative assessment of large-scale data sets of protein-protein interactions. Nature 417:399-403.
Wagner, GP. 2003. Evolutionary genetics: The nature of hidden genetic variation unveiled. Current Biology 13:R958-R960.
Wang, H, B Kakaradov, SR Collins, L Karotki, D Fiedler, M Shales, KM Shokat, TC Walther, NJ Krogan, D Koller. 2009. A complex-based reconstruction of the Saccharomyces cerevisiae interactome. Mol Cell Proteomics 8:1361-1381.
Wapinski, I, A Pfeffer, N Friedman, A Regev. 2007. Natural history and evolutionary principles of gene duplication in fungi. Nature 449:54-61.
Warnecke, T, LD Hurst. 2010. GroEL dependency affects codon usage--support for a critical role of misfolding in gene evolution. Mol Syst Biol 6:340.
Wilgenbusch, JC, D Swofford. 2003. Inferring evolutionary trees with PAUP*. Curr Protoc Bioinformatics Chapter 6:Unit 6.4.
Williams, EJ, DJ Bowles. 2004. Coexpression of neighbouring genes in the genome of Arabidopsis thaliana. Genome Res 14:1060-1067.
Williams, EJ, LD Hurst. 2002. Clustering of tissue-specific genes underlies much of the similarity in rates of protein evolution of linked genes. J

Mol Evol 54:511-518.

Wolfe, KH, DC Shields. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. Nature 387:708-713.

Xu, Z, W Wei, J Gagneur, F Perocchi, S Clauder-Munster, J Camblong, E Guffanti, F Stutz, W Huber, LM Steinmetz. 2009. Bidirectional promoters generate pervasive transcription in yeast. Nature 457:1033-1037.

Yang, L, J Yu. 2009. A comparative analysis of divergently-paired genes (DPGs) among Drosophila and vertebrate genomes. BMC Evol Biol 9:55.

Yin, S, P Wang, W Deng, H Zheng, L Hu, LD Hurst, X Kong. 2009. Dosage compensation on the active X chromosome minimizes transcriptional noise of X-linked genes in mammals. Genome Biol 10:R74.

Yosef, N, M Kupiec, E Ruppin, R Sharan. 2009. A complex-centric view of protein network evolution. Nucleic Acids Res 37:e88.

Yu, H, P Braun, MA Yildirim, *et al* . 2008. High-quality binary protein interaction map of the yeast interactome network. Science 322:104-110.

Zelle, RM, E de Hulster, WA van Winden, P de Waard, C Dijkema, AA Winkler, JM Geertman, JP van Dijken, JT Pronk, AJ van Maris. 2008. Malic acid production by Saccharomyces cerevisiae: engineering of pyruvate carboxylation, oxaloacetate reduction, and malate export. Appl Environ Microbiol 74:2766-2777.

Zhang, B, S Horvath. 2005. A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol 4:Article17.

Zhang, J. 2003. Evolution by gene duplication: an update. Trends in Ecology & Evolution 18:292-298.

Zhang, Z, W Qian, J Zhang. 2009. Positive selection for elevated gene expression noise in yeast. Mol Syst Biol 5:299.