INFORMATIC APPROACHES TO EVOLUTIONARY SYSTEMS BIOLOGY

A Dissertation

presented to

the Faculty of the Graduate School

at the University of Missouri-Columbia

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

---

by

COREY M. HUDSON

JULY 2013

The undersigned, appointed by the dean of the Graduate School, have examined the dissertation entitled

INFORMATIC APPROACHES TO EVOLUTIONARY SYSTEMS BIOLOGY

presented by Corey M. Hudson, a candidate for the degree of doctor of philosophy, and hereby certify that, in their opinion, it is worthy of acceptance.

Professor Gavin C. Conant

Professor J. Chris Pires

Professor Dmitry Korkin

Professor Jianlin Cheng

ACKNOWLEDGEMENTS

I would like to acknowledge a number of researchers and funding agencies who helped make this work possible. Thanks especially to Dr. Michaël Bekaert and Emily E. Puckett who helped produce and analyze the data in Chapters 3 and 4. The work in these chapters benefited considerably because of their effort. Thanks also to my thesis committee, including Chris Pires, Jianlin Cheng and Dmitry Korkin who helped in numerous ways registering scientific concerns, brainstorming solutions and editing every step of the way. Thanks to Gavin C. Conant who served as an exemplary advisor and helped craft this thesis research. Thanks to the MU Informatics Institute for putting me in the position to do this research. And thanks to the NLM Bioinformatics and Health Informatics Training Fellowship and the Reproductive Biology Group of the Food for the 21st century program at the University of Missouri who funded this research.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF ILLUSTRATIONS

Figure

Additional Files

ABSTRACT

Abstract: The sheer complexity of evolutionary systems biology requires us to develop more sophisticated tools for analysis, as well as more probing and biologically relevant representations of the data. My research has focused on three aspects of evolutionary systems biology. I ask whether a gene's position in the human metabolic network affects the degree to which natural selection prunes variation in that gene. Using a novel orthology inference tool that uses both sequence similarity and gene synteny, I inferred orthologous groups of genes for the full genomes of 8 mammals. With these orthologs, I estimated the selective constraint (the ratio of non-synonymous to synonymous nucleotide substitutions) on 1190 (or 80.2%) of the genes in the metabolic network using a maximum likelihood model of codon evolution and compared this value to the betweenness centrality of each enzyme (a measure of that enzyme's relative global position in the network). Second, I have focused on the evolution of metabolic systems in the presence of gene and genome duplication. I show that increases in a particular gene's copy number are correlated with limiting metabolic flux in the reaction associated with that gene. Finally, I have investigated the proliferative cell programs present in 6 different cancers (breast, colorectal, gastrointestinal, lung, oral squamous and prostate cancers). I found an overabundance of genes that share expression between cancer and embryonic tissue and that these genes form modular units within regulatory, protein-interaction, and metabolic networks. This despite the fact that these genes, as well as the proteins they encode and reactions they catalyze show little overlap among cancers, suggesting parallel independent reversion to an embryonic pattern of gene expression.

CHAPTER 1 INTRODUCTION: YEAST AS A WINDOW INTO CHANGES IN GENOME COMPLEXITY DUE TO POLYPLOIDY

1.1 Introduction

Researchers have found remnants of ancient whole genome duplications (WGDs) preserved in the genomes of many and diverse eukaryotes. This book, in fact, is a testament to that diversity, illustrating the sheer number of independent events in plants as well as the evolutionarily basal events in vertebrates and other, more recent WGDs in teleost fishes and frogs. Although we have still not fully validated Susumo Ohno's claim for the primacy of polyploidization in the generation of new adaptations (Ohno 1970), it is clear that WGD events have had a massive influence on the content and structure of the genomes of their possessors. The next step in exploring Ohno's hypothesis is to link genome evolution to known changes in function. This goal, however, remains challenging, primarily because our knowledge of how genotype links to phenotype remains woefully incomplete (Pigliucci 2010). However, one group of organisms in which we can at least begin to make such associations is in the polyploid yeasts. Our knowledge of the functional genomics of yeast is drawn primarily from *Saccharomyces cerevisiae*, which has a well-annotated genome, decades of biochemical, genetic and cell biology research, a relatively small genome, and a life cycle that lends itself to scalable laboratory analyses. Taken together, these facts have allowed yeast researchers to understand not only the structure of the genome following WGD but also to

experimentally evaluate hypotheses regarding the evolution of particular complex phenotypes. As we will stress throughout this chapter, one of the themes that emerges from all of these analyses is the degree to which the outcome of a WGD depends as much on the interactions *between* genes as on the role of any particular locus. Of equal importance evolutionarily, we now also have data from other polyploid yeast species, which are valuable both as a point of comparison to *S. cerevisiae* and for their own sakes. This wealth of data affords us insight into the mechanisms that drive the preferential loss and preservation of gene duplicates after polyploidy, lead to the functional divergence of genes, and are behind the evolutionary origins of complex phenotypes.

1.2 Evidence for WGD in yeast

Although early analyses of genes potentially created by the vertebrate 2R events used phylogenetic approaches (Hughes 1999; Furlong and Holland 2002), most current studies of WGD rely on one or both of two methods: 1) finding numerous blocks of paralogous genes in multiple chromosomes with similar gene orders and 2) clustering homologs into groups by measuring the rate of synonymous substitutions (K_s or dS). This second method assumes that the gene pairs created by WGD cluster about some mean K_s value (Lynch and Conery 2000). The simultaneous application of both methods have been used to group multiple WGD events within species (e.g., Arabidopsis thaliana and Tetradon nigroviridis; Jaillon, Aury et al. 2004; Van de Peer, Fawcett et al. 2009). However, the yeast genomes present an interesting challenge in this respect because the synonymous substitutions between yeast paralogs produced by WGD (hereafter ohnologs; Wolfe 2000) are often saturated (Byrne and Wolfe 2007). In other words, identical synonymous

positions between two ohnologs occur almost as often due to repeated convergent substitutions as due to common ancestry, a fact pointed out by Smith (1987), who attempted to date histone gene duplicates in yeast. While the genomic structure of the core histone genes suggested that they were all duplicated simultaneously, these genes show considerable variation in the numbers of synonymous substitutions separating them. This led Smith (1987) to hypothesize that *Saccharomyces cerevisiae* underwent a WGD ancient enough that the duplicates surviving from it had saturated. However, it was not until the genome sequence of *S. cerevisiae* became available that this speculation could be confirmed (see below). Another similar but subtler problem in using K_s as a means of dating duplicate genes is the issue of gene conversion. Gene conversion was presumed to be quite common in yeast (Petes and Hill 1988), even prompting some authors (Gao and Innan 2004) to suggest that estimates of duplication rates based on duplicate divergences were inapplicable due to the homogenization of duplicate loci by conversion. Fortunately, although gene conversion is very common among yeast ribosomal proteins, it does not appear to be a general characteristic of the genome (Evangelisti and Conant 2010). Nonetheless, these various issues collectively meant that comparisons of paralogous sequence divergence were deemed unhelpful as a means to detect WGD in yeast.

1.2.1 Synteny based evidence for WGD in *Saccharomyces cerevisiae*

Given that paralogous sequence comparisons were generally unhelpful in finding WGD relics, another tactic was to consider gene order. In fact, even before the *S. cerevisiae* genome was completed in 1996, it was clear to many researchers that it contained numerous, long, homologous clusters of ordered genes (Goffeau, Barrell et al. 1996).

Melnick and Sherman (1993) found ordered homologous gene clusters in chromosomes V and X covering 7.5-kb. Lalo et al. (1993) similarly found ordered homologous gene clusters in chromosomes XIV and III covering 15-kb. When the genome was sequenced, researchers found 18 ordered homologous genes in chromosomes IV and II that covered 120-kb and 170-kb respectively (Goffeau, Barrell et al. 1996). Just how to interpret these redundant regions remained a challenge at that time (Goffeau, Barrell et al. 1996; Oliver 1996), and, in spite of Smith's prior hypothesis (1987), few, if any, of the contemporaneous explanations included an ancient WGD.

However, opinions changed the next year when Wolfe and Shields (1997) presented a thorough, genome-sequence-based, analysis that gave strong evidence for WGD in *S. cerevisiae*. To find syntenic regions, they conducted a BLASTP search of amino acid sequences throughout the yeast genome and made a dot-plot of the results. They then created gene blocks from these data, where each block was required to have at least three homologous pairs with intergenic distances ≤ 50-kb and conservation of gene order and orientation. This analysis yielded 55 duplicated regions containing a total of 376 pairs of ohnologs. The large number of duplicated regions led Wolfe and Shields to posit two explanations: 1) Successive independent gene duplications, and 2) A single duplication of the entire genome, followed by massive gene loss. There were two lines of evidence discounting the first possibility. First, 90% (50/55) of the gene regions shared the same orientation with respect to the centromeres of the duplicated regions when we would expect independent duplications to be instead randomly distributed about the centromeres. Second, there were no examples of triplicated regions in the *S. cerevisiae* genome. If the duplications involved several distinct events separated in time, such a

pattern would be highly unlikely, because it would require that later duplication events *never* overlapped with prior ones. Given these arguments, Wolfe and Shields (1997) argued for a single ancient WGD, which they dated to be hundreds of millions of years old (although attempts to conclusively date this have been difficult, due to a lack of fossils and the previously mentioned saturation of substitutions see Taylor and Berbee 2006; Rolland and Dujon 2011).

1.2.2 Comparative genomics and proof of WGD in *S. cerevisiae*

A number of researchers disputed the claims of Wolfe and Shields (1997), arguing that, because the syntenic regions identified made up only a small part of the genome, independent duplications better explained *S. cerevisiae*'s genomic structure (Coissac, Maillier et al. 1997; Mewes, Albermann et al. 1997; Hughes, Roberts et al. 2000; Llorente, Durrens et al. 2000; Llorente, Malpertuy et al. 2000; Friedman and Hughes 2001; Piskur 2001; Koszul, Caburet et al. 2004). However, this independent duplication hypothesis became untenable following the genome sequencing of other yeasts that proved to lack these syntenic paralog blocks. These sequences were described by three independent groups. The comparison of *S. cerevisiae* with *Kluyveromyces waltii* (Kellis et al. (2004) and the comparison of *S. cerevisiae* with *Ashbya gossypii* (Dietrich et al. 2004) involved different genomes, but effectively made the same argument: that the 2:1 mapping of blocks of paralogs from *S. cerevisiae* to homologous single copy genes in *K. waltii/A.gossypii* could best be explained by WGD. This explanation was particularly striking because the doubly conserved synteny blocks cover 90% of the genome in *K. waltii* (Kellis, Birren et al. 2004) and 96% of that in *A. gossypii* (Dietrich, Voegeli et al.

2004). Furthermore, both studies found a large number of 2:1 pairing of centromeres in the species respective chromosomes. There were 16:8 such pairings between *S. cerevisiae* and *K. waltii* and 14:7 between *S. cerevisiae* and *A. gossypii* with a subsequent break at the expected centromere position in *S. cerevisiae* chromosomes X and XII that are syntenic with regions in *A. gossypii* chromosomes I and III. Finally, and perhaps most strikingly, both groups also showed that the single-copy orthologs of genes from *A. gossypii* or *K. waltii* in the genome of *S. cerevisiae* are interleaved between two paralogous chromosomes in *S. cerevisiae* that nonetheless retain the relative gene order of the single chromosome in the non-WGD yeast (see Figure 1). Such a pattern is only explicable under the hypothesis of a WGD event followed by massive gene losses.

The argument of Dujon et al. (2004) is subtly different. They sequenced and analyzed four other genomes. One genome, that of *Candida glabrata*, shares the genome duplication with *S. cerevisiae*. This was determined by comparing syntenic blocks in *S. cerevisiae* and *C. glabrata* with the other three sequenced genomes, *Kluveromyces lactis*, *Debaryomyces hansenii*, and *Yarrowia lipolytica*. Dujon et al. (2004) found 20 distinct blocks of paralogs shared by both *S. cerevisiae* and *C. glabrata*. These blocks allowed them to map the WGD onto a phylogeny, rather than do a simple pairwise comparison. Mapping this WGD phylogenetically creates distinct hypotheses as to where in the tree we expect to find polyploid yeasts (*c.f.,* Figure 2); predictions that have been confirmed with each of the subsequently sequenced genomes of known phylogenetic position (Wapinski, Pfeffer et al. 2007; Scannell, Zill et al. 2011).

1.2.3 Yeast gene order browser (YGOB)

One of the major benefits of studying the yeast WGD is that the relatively slow rates of gene order change in yeast genomes and the compactness of their genomes means that an exhaustive enumeration of all WGD-produced ohnologs is possible. Just such a project was carried out, with the results presented as the web-based Yeast Gene Order Browser (YGOB; Byrnes, Morris et al. 2006), which illustrates a number of non- and post-WGD yeasts in a graphical framework (Figure 1). This work has been followed by a reconstruction of the set of genes and their relative orders that existed just prior to the WGD (Gordon, Byrne et al. 2011) and by a likelihood-based model of post-WGD duplicate loss that attempts to quantify the orthology inferences made by YGOB (Powell, Conant et al. 2008). On the basis of these three projects, the post-WGD evolutionary history of virtually every locus in the *S. cerevisiae* genome can be traced (Figure 1 is thus illustrative of the predominant pattern seen across the genome).

1.2.4 Additional non-*Saccharomyces*-specific WGDs

In addition to the ancient WGD that characterizes the *Saccharomyces* clade (Figure 2), several cases of allopolyploidy have been discovered in yeasts. Some of these occur in species within the *Saccharomyces sensu stricto* clade (Scannell, Zill et al. 2011), while others are independent.

1.2.4.1 Secondary allopolyploidy in *Saccharomyces pastorianus*

A number of cases of allopolyploidy is known from within *Saccharomyces sensu stricto* (Dequin and Casaregola 2011). One of the most-well studied is that of the lager yeast, *Saccharomyces pastorianus* (syn. *Saccharomyces carlsbergensis*). It has long been known that the polyploid *S. pastorianus* and other members of the complex of related

lager yeasts are allotetraploids of diploid *S. cerevisiae* and some other unknown diploid species (Martini and Kurtzman 1985; Kielland-Brandt, Nilsson-Tillgren et al. 1995). However, aside from the general difficulties facing anyone interested in identifying the origins of hybrid genomes, the debate surrounding the origin of the second parental diploid species was further complicated by a difficulty in delimiting species within these groups (Rainieri, Kodama et al. 2006). The tetraploid *Saccharomyces pastorianus* belongs to a group of yeast species, which until recently, was represented as a phylogenetically unresolved species complex including *S. pastorianus*, *S. monacensis* (*S. pastorianus* strain CBS 1503)*, S. bayanus*, and *S. bayanus* var. *uvarum* (Casaregola, Nguyen et al. 2001; Rainieri, Kodama et al. 2006). This taxonomic confusion has recently been partially resolved through the sequencing of the genomes of both *S. pastorianus* and one of its presumed parental diploid species, *Saccharomyces eubayanus* (Nakao, Kanamori et al. 2009). The genome history that has emerged is a complicated story of allopolyploidy followed by the genomic transformation forming the related species *S. bayanus* (Libkind, Hittinger et al. 2011).As summarized by Libkind et al. (2011) *S. cerevisiae* hybridized with *S. eubayanus* (a species recently recovered in Patagonia) with subsequent genome doubling producing the allotetraploid progenitor of modern *S. pastorianus*. Following domestication, smaller regions of the *S. pastorianus* genome were then apparently transferred into the genome of the diploid parent *S. eubayanus* (which is nonetheless a descendant of the ancient polyploidy). This hybrid form of *S. eubayanus*, with contributions from *S. pastorianus,* then proceeded to interbreed with diploid *S. uvarum* to produce the modern, diploid, *S. bayanus* (Figure 3).

1.2.4.2 *Zygosaccharomyces rouxii* allopolyploidy

Another of the yeast allopolyploids occurs in cultures of the spoilage agent and industrial yeast *Zygosaccharomyces rouxii*. James et al (2005) and Gordon and Wolfe (2008) identified *Z. rouxii* strain ATCC 42981 as an allopolyploid. This hybridization/polyploidy event is significant for two reasons. Firstly, unlike all of the previous examples, it occurs outside of *Saccharomyces stricto sensu*. Secondly, Gordon and Wolfe (2008) determined that most of the paralogs produced by WGD are still present, presumably due to the recentness of the event. Thus, while other yeast genome duplications are ancient and show considerable gene loss and rearrangement (Wolfe and Shields 1997), the *Z. rouxii* genome retains most of the "new" genes produced by its WGD. Since the survival time of ohnologs has been modeled to follow a power-law, most of the duplicates are expected be lost very rapidly (Maere, De Bodt et al. 2005), suggesting that *Z. rouxii* represents an example of the early features of genome evolution following WGD.

1.3 WGD and speciation

An important potential outcome of polyploidy is in altering patterns of speciation. This change can happen in at least two ways. First, the WGD can relax selective constraints resulting in an adaptive radiation by means of ecological speciation. Another, more neutral mechanism, is a special case of the Dobzhansky-Muller (DM) process of speciation, in which species lose reciprocal paralogs following some period of isolation (Lynch and Force 2000). WGD potentially increases the probability of this simply by increasing the number of paired genes in a genome. The fertility of hybrids is 0.75^{n}, where n is the number of reciprocal losses of essential genes between populations (Werth and Windham 1991). Clearly, for any significant number of reciprocal losses (such as

occur after WGD), the number of viable, fertile offspring of a crossing of two such populations is negligible. Both phylogenetic and experimental studies of the DM process after WGD have been carried out in yeast. Scannell et al. (2006) showed that the number of reciprocal gene losses in several species of yeast sharing the *S. cerevisiae* WGD was sufficient to induce such inviability. This observation suggests that a DM mechanism was partly responsible for the multiple speciation events among the *Saccharomyces* species (*e.g., S. cerevisiae*, *S. bayanus*, and *C. glabrata*) following WGD.

An advantage of studying the DM process in yeast is the ability to experimentally create and cross artificial polyploids. This possibility has been highlighted in experimental studies of reproductive isolation. Polyploid yeasts have been allowed to evolve in different selective environments (Dettman, Sirjusingh et al. 2007) and in neutral environments subject to random mutagenesis (Maclean and Greig 2011). These two experiments have shown that moderate reproductive isolation, coupled with reciprocal gene loss results in a clear loss of fitness when independently derived polyploids are crossed. Similarly, Lee et al. (2008) showed that hybrids of *S. cerevisiae* and *S. bayanus* were less fit than their parental phenotypes, due primarily to incompatibility between their nuclear and mitochondrial genomes. Chou et al. (2010) extended this analysis, providing another pair of mitochondrial and nuclear genes and posited nuclear-mitochondrial incompatibility as a common mechanism in species formation. In another twist, Anderson et al. (2010) demonstrated the existence of alleles with depressed hybrid fitness in low-glucose environments, which argues for a model in which neutral changes in paired genes are followed by strong selection, a sequence of events that promotes rapid reproductive isolation. Kao et al. (2010), however, argue against the existence of a small

number of so-called *speciation genes*, instead claiming that genome scans provide no evidence of any single paired dominant or recessive genic incompatabilities. They instead argue that following WGD, many changes in loci of little effect resulted in lowered fitness due, in part, to the rewiring of transcriptional and metabolic networks.

Another debate that has emerged in this field is whether these changes are due primarily to the decrease in the fertility of hybrids (Xu and He 2011) or a decease in their viability (Greig 2008). This question ultimately amounts to a debate about what stage in the yeast life cycle the genetic incompatibilities occur – sporulation or clonal growth, and whether the decrease in fitness is the result of competition for resources or offspring. The discontinuity between these ideas likely represents an opportunity to explain speciation as a process across different genomic and temporal scales, and we would speculate that the process of DM incompatibility induces selection for the evolution of some form of prezygotic barrier.

1.4 Changes in genome content and complexity post-WGD

Duplicate retention and evolutionary models. In addition to such population processes as speciation, WGD also altered many other aspects of the *S. cerevisiae* lifestyle. For instance, several pairs of ohnologs have been shown to have undergone various types of functional divergence, allowing the study of some of the proposed mechanisms of duplicate divergence after duplication (Conant and Wolfe 2008). In an elegant series of experiments, van Hoof (2005) showed that two ohnologs, *ORC1* and *SIR3* have distinct and non-overlapping functions (in DNA replication and gene silencing, respectively). Strikingly, however, the mutual ortholog of these genes from the non-WGD yeast *S.*

kluyveri is able to complement both functions, constituting a clear example of subfunctionalization. An apparently similar case, involving the *S. cerevisiae* ohnolog pair *GAL1* and *GAL3*, which presently function respectively as an enzyme and as a transcriptional regulator, was complicated by the discovery of an adaptive conflict between the shared regulator and enzymatic function of their ortholog in the non-WGD *K. lactis*. Thus, although the *K. lactis GAL1* gene does indeed serve the functions of both *GAL1* and *GAL3* in *S. cerevisiae*, it does so in a suboptimal way, being unable to tune its expression to both roles simultaneously (Libkind, Hittinger et al. 2011). This conflict illustrates an important point about subfunctionalization, namely that the original neutral model of subfunctionalization proposed by Force and coauthors (Force, Lynch et al. 1999) is not the only possible mechanism for such functional partitioning (Des Marais and Rausher 2008). Other examples of divergence among ohnologs where the mechanism of that divergence is less clear include ribosomal proteins (Ni and Snyder 2001; Komili, Farny et al. 2007; Ha, Kim et al. 2009), glucose sensors (Özcan, Dover et al. 1998) and glycolysis enzymes (Boles, Schulte et al. 1997).

The dosage balance hypothesis. In addition to facilitating the above work, the wealth of functional data from *S. cerevisiae* also provides an excellent opportunity to test hypotheses explaining the differences in gene retention patterns after WGD and small scale duplications (hereafter SSD). Chief among these is probably the dosage balance hypothesis (DBH) (Papp, Pal et al. 2003; Freeling and Thomas 2006; Birchler and Veitia 2007; Freeling 2009), which states that, in eukaryotes, there is selection operating to disfavor duplications of central network genes due to the imbalance in network stoichiometry that results. This situation is reversed for WGD because in that case the

loss of a second copy of a gene introduces imbalances relative to the remaining, duplicated, genes. In keeping with the DBH, several classes of genes are over-retained after several evolutionarily ancient WGD events, including that in yeast. They include ribosomal proteins and protein kinases and transcription factors (Seoighe and Wolfe 1999; Blanc and Wolfe 2004; Maere, De Bodt et al. 2005; Aury, Jaillon et al. 2006; Conant and Wolfe 2008). Similarly, genes that tend to have been fixed by WGD are less likely to have undergone SSD in other yeast species (Wapinski, Pfeffer et al. 2007). However, WGD-duplicates produced by genome duplication have more protein interactions (Guan, Dunham et al. 2007; Hakes, Pinney et al. 2007), more phosphorylation sites (Amoutzias, He et al. 2010) and tend to be highly expressed (Seoighe and Wolfe 1999). Although genes retained in duplicate after WGD are rarely essential on an individual basis (Guan, Dunham et al. 2007), this dispensability appears to be due to functional compensation by the other ohnolog (DeLuna, Vetsigian et al. 2008). Thus, it appears that while ohnologs are less likely to be essential than their SSD counterparts today, their ancestral genes were actually at least as essential as current single copy genes (DeLuna, Vetsigian et al. 2008).

System-level changes produced by WGD. Of course, one of the unique features of polyploidy relative to SSD is the possibility of coordinated changes among multiple sets of ohnologs. At the simplest level, we have previously illustrated examples of what appears to be *network* subfunctionalization where a number of ohnologs collectively divided two expression domains amongst themselves (Conant and Wolfe 2006). A more complex and interesting example is the role of the WGD (Piškur, Rozpędowska et al. 2006) in shaping *S. cerevisiae's* propensity for aerobic glucose fermentation (the Crabtree

effect; Geladé, Van de Velde et al. 2003; Johnston and Kim 2005), a novel and somewhat paradoxical phenotype. There is a general association between the presence of the WGD and the Crabtree effect across yeast species (Merico, Sulo et al. 2007). As a result, we and others have argued that dosage effects among the glycolysis enzymes post-WGD helped to increase flux through glycolysis (Blank, Lehmbeck et al. 2005; Kuepfer, Sauer et al. 2005; Conant and Wolfe 2007; Merico, Sulo et al. 2007; van Hoek and Hogeweg 2009). Such increased flux likely could only be accommodated through fermentation pathways, given the complex spatial organization of the competing respiratory pathway (Conant, Wagner et al. 2007). Supporting this hypothesis is an elegant computational analysis by van Hoek and Hogeweg (2009) showing that future WGD events in the modern *S. cerevisiae* could also be expected provide a selective advantage in glucose-rich environments through the preferentially retention of duplicated glycolysis enzymes. Note that the apparently "wasteful" fermentation can actually be selectively advantageous in the context of rich but ephemeral resource patches (Pfeiffer, Schuster et al. 2001; Pfeiffer and Schuster 2005), a phenomenon that has been experimentally confirmed in yeast (MacLean and Gudelj 2006). Such a change in the yeast lifestyle likely led to other, later, changes in the genome. One suggestive example concerns the de-coupling of cytosolic and mitochondrial ribosomal protein expression post-WGD (Ihmels, Bergmann et al. 2005). Prior to WGD, bakers' yeast was likely similar to other yeasts in having a strong association in the expression of the two types of ribosomal proteins. After WGD however, *cis*-regulatory element evolution diverged in the two groups of genes (Ihmels, Bergmann et al. 2005), allowing *S. cerevisiae* to express only cytosolic proteins at high levels during fermentation, an important refinement in a fermentative lifestyle.

Connecting the dosage balance hypothesis to large-scale evolutionary changes following WGD, Conant (2010) and Fusco et al. (2010) found transcriptional regulatory motifs to be over-retained in ohnologs. Modeling network evolution after WGD, these authors find the network enriched for transcription factors and particular network motifs. Duplicated transcription factors still show some relics of the WGD, being more likely to share targets than are random transcription factors, but on the whole show considerable divergence post-WGD (Conant 2010). Given this rapid regulatory evolution (relative to the substitution rates K_a or K_s; Gu et al. 2002), it may not be easy to ascertain the role of WGD in the evolution of the modern *S. cerevisiae* regulatory network. Nonetheless, the retention of many transcription factors that have acquired distinct sets of target genes may imply that the WGD served to "relax" the regulatory complexity of this organism, which may have implications for its future ability to adapt (as seen for the *GAL1/GAL3* example).

1.5. Conclusions

The *S. cerevisiae* WGD has been implicated in a number of evolutionarily complex events. At a minimum, a set of duplicated genes of identical age is a powerful system for exploring duplicate gene evolution (van Hoof 2005; Conant and Wolfe 2006; Fares, Byrne et al. 2006; Kim and Yi 2006). However, we also suggest that, as with the *GAL1/GAL3* example, we will not fully understand the biology of *S. cerevisiae* until we account for how the WGD has altered both the individual roles of particular genes and their relationships to each other. We have outlined some of the areas of yeast biology that we think were altered by this genome-doubling event: there remain others yet to be

discovered. Similarly, the presence of other WGD events, of varying ages, allows us to study how these events unfold over various timescales, including, potentially, on the timescale of laboratory experiments in evolution.

CHAPTER 2 EXPRESSION LEVEL, CELLULAR COMPARTMENT AND METABOLIC NETWORK POSITION ALL INFLUENCE THE AVERAGE SELECTIVE CONSTRAINT ON MAMMALIAN ENZYMES

2. 1 Abstract

2.1.1 Background

A gene's position in regulatory, protein interaction or metabolic networks can be predictive of the strength of purifying selection acting on it, but these relationships are neither universal nor invariably strong. Following work in bacteria, fungi and invertebrate animals, we explore the relationship between selective constraint and metabolic function in mammals.

2.1.2 Results

We measure the association between selective constraint, estimated by the ratio of nonsynonymous (K_a) to synonymous (K_s) substitutions, and several, primarily metabolic, measures of gene function. We find significant differences between the selective constraints acting on enzyme-coding genes from different cellular compartments, with the nucleus showing higher constraint than genes from either the cytoplasm or the mitochondria. Among metabolic genes, the centrality of an enzyme in the metabolic network is significantly correlated with K_a/K_s. In contrast to yeasts, gene expression magnitude does not appear to be the primary predictor of selective constraint in these organisms.

2.1.3 Conclusions

Our results imply that the relationship between selective constraint and enzyme centrality is complex: the strength of selective constraint acting on mammalian genes is quite variable and does not appear to exclusively follow patterns seen in other organisms.

2.2 Background

The rate and manner of evolutionary change has long been a matter of keen interest to biologists (Simpson 1944). Kimura provided theoretical underpinnings to molecular evolution by relating rates of sequence substitution, population parameters and mutation rates (Kimura 1968; Kimura 1969). Thus, Kimura's neutral theory (Kimura 1983) predicts that mutations having no fitness effect will become fixed in a population at a rate equal to the mutation rate. Such neutral mutations therefore provide a standard for measuring the action of natural selection: regions changing more slowly than neutral ones are inferred to be experiencing purifying selection (e.g., selective constraint), those changing more rapidly, adaptive evolution. While the relative contributions of genetic drift, adaptive evolution and purifying selection to population differentiation are still debated, (Nei 2005), there is general agreement that the patterns of selection vary both across species as well as among genes in the same species (Cooper, Brudno et al. 2004).

Regarding interspecific variation, Lynch and Conery (Lynch and Conery 2003) argue that much of the variation in genome structure and content between species can be attributed to differences in their effective population sizes (N_e). Small effective population sizes limit the efficiency of purifying selection and allow the occasional fixation of mildly deleterious mutations. While some cross-taxa surveys have reported patterns consistent with this hypothesis (Conant 2009; Ellegren 2009; Slotte, Foxe et al. 2010), others have found that if one allows for reasonably frequent directional selection

there is only a weak relationship between N_e and selective constraint (Fay, Wyckoff et al. 2002; Charlesworth and Eyre-Walker 2007; Bachtrog 2008).

The second type of variation in selective constraint, that between genetic loci in the same population, has also been studied (Fitch and Margoliash 1967; Yang 1993; Drummond, Bloom et al. 2005). In particular, considerable effort has gone into identifying factors that predict the selection acting on a particular gene. One critical variable is expression level: mammalian genes expressed in many tissues show stronger selective constraints than do those expressed in only a few tissues (Duret and Mouchiroud 2000). Likewise, in yeast, a high expression level is the primary predictor of strong purifying selection acting on a gene (Drummond, Raval et al. 2006), likely because the selective cost of protein misfolding is especially large for highly translated proteins (Drummond, Bloom et al. 2005).

This association is also in keeping with Wagner's theoretical analyses showing that gene expression is selectively costly in yeast (Wagner 2005). However, as he notes, the fitness cost of mis-expression is likely to be very different in multicellular organisms (Wagner 2005).

The influence of other factors on selective constraint is also debated, with the evidence primarily coming from studies in yeast (Hurst and Smith 1999; Hirsh and Fraser 2001; Fraser, Hirsh et al. 2002; Jordan, Rogozin et al. 2002; Jordan, Wolf et al. 2003; Pál, Papp et al. 2003; Hahn, Conant et al. 2004; Drummond, Raval et al. 2006). The topic is confounded by the intercorrelation of many of these predictors (Drummond, Raval et al. 2006). Thus, some researchers report a significant correlation between the fitness cost of gene knockouts and those genes' selective constraint (Hirsh and Fraser 2001; Jordan,

Rogozin et al. 2002), while others have questioned this association (Hurst and Smith 1999; Pál, Papp et al. 2003). There is similar debate regarding whether the position of a gene or protein in an interaction network influences selective constraint.

Recall that in these networks genes or proteins are nodes; relationships, such as protein interactions or shared metabolites, are represented as edges between nodes. Researchers have studied the association between selective constraint and measures such as node degree (the number of edges for a given node) and betweenness centrality (a more global statistic measuring the number of shortest paths passing through a node ; Freeman 1977; Brandes 2001; Liu, Lin et al. 2007). Significant associations between node importance and selective constraint have been found in regulatory (Jovelin and C 2009), protein interaction (Fraser, Hirsh et al. 2002), coexpression (Jordan, Marino-Ramirez et al. 2004), and metabolic networks (Vitkup, Kharchenko et al. 2006; Greenberg, Stockwell et al. 2008; Wagner 2009). However, at least for protein interaction networks, this association seems to be at best quite weak (Fraser, Hirsh et al. 2002; Jordan, Wolf et al. 2003; Hahn, Conant et al. 2004).

Here we explore to what degree these patterns of constraint extend to mammals. Given the difference in lifestyle and effective population size between humans and yeast, we hypothesized that mammals would have evolved in a manner similar to *Drosophila* (Greenberg, Stockwell et al. 2008), where there is a significant association between enzyme centrality and evolutionary constraint. We asked whether a gene's position in the human metabolic network (Figure 4) predicts the strength of the purifying selection acting on it. Some previous analyses have calculated the protein divergence between two species, using their common divergence to control for the mutation rate (Hahn, Conant et

al. 2004). However, only sampling two sequences offers somewhat limited resolution in the estimation of selective constraint. Here we follow Greenberg, Stockwell and Clark (Greenberg, Stockwell et al. 2008) by estimating the selective constraint acting on each human enzyme by comparing it to its orthologs from seven other eutherian genomes (chimpanzee, macaque, mouse, rat, horse, dog and cow). We find that genes encoding metabolic proteins evolve significantly more slowly than other genes. Among those metabolic genes, the encoded protein's cellular compartment is predictive of selective constraint. We also find a weak, though statistically significant, negative correlation between the betweenness of an enzyme in the metabolic network and constraint.

2.3 Results

Orthology identification. To infer selective constraints for the set of annotated human genes, we identified their orthologs in seven other mammalian genomes using an approach that combines sequence similarity and gene order information (*Methods*). We found 19,416 human genes with at least one ortholog in these genomes. Among those genes, we identified 13,928 sets of orthologs with between 6 and 8 members. Of the 1,496 genes annotated by Duarte et al. (Duarte, Becker et al. 2007) as belonging to the metabolic network, 1,190 are in this ortholog set (Figure 5). A greater percentage of genes in the metabolic network fell into our set of orthologs than did genes from the genome at large ($\chi^2 = 47.9$; $P<0.001$; Figure 2B).

Metabolic and nonmetabolic genes differ in selective constraint. The ratio of nonsynonymous to synonymous substitutions (K_a/K_s: hereafter ω) for each set of orthologous genes was estimated by maximum likelihood using PAML 4.2 (Figure 5A;

Yang 2007). This ratio can be interpreted as a measure of selective constraint: values near 0 indicate strong purifying selection, while values greater than 1.0 suggest directional selection.

We hypothesized that metabolic genes would also be under stronger selective constraint than the average non-metabolic gene, so we performed three statistical tests of this hypothesis. First, using a Mann-Whitney U-test (Wilcoxon two-sample test), we rejected the null hypothesis that the median ω among metabolic genes is no smaller than that of non-metabolic genes (i.e., a one-tailed test, P=0.035; Figure 5). Next, we performed a similar test for unequal *mean* ω values between the two groups. Given that neither distribution in Figure 5A appears normal, we adopted a bootstrapping approach, drawing 1,000,000 samples of size n = 1,190 from the set of non-metabolic genes (n = 12,738) and calculating these samples' means. In no case was the mean value of ω from the bootstrapped samples as small as the observed mean value for metabolic genes ($\omega_{metabolic}$ = 0.1292, $P < 10^{-6}$). We also drew 100,000 samples of sizes n =1,190 and of n = 12,738 and calculated the difference in their means. The absolute differences in the mean values was never as large as that observed between the metabolic and non-metabolic genes ($P < 10^{-5}$; Figure 5C).

Finally, we performed a more general analysis of the distributions of ω in the two gene sets. To do so, we first fit eight common distributions, the normal, gamma, exponential, Cauchy, log-normal, logistic, Weibull, and extreme value distributions, to the overall set of ω values. We then assessed the quality of the fit of each distribution to the data by analyzing the linear correlation between the ranked data and a Q-Q plot (Table 1; see *Methods*). Out of the eight distributions, three, the Weibull, gamma and

exponential provide a visually good fit to the ω values (Additional file 1, Figure S1). For these three distributions, we compared a null model where all genes shared the same distribution parameters to an alternative where the metabolic and non-metabolic genes were allowed to have distinct parameter values for that same distribution. Using a likelihood ratio test, we found that we could reject the null model of identical distributions of ω for the metabolic and non-metabolic genes for all three distributions ($P < 10^{-6}$; chi-square distribution; Table 1). Collectively, these three analyses allow us to firmly conclude that metabolic genes are under greater selective constraint than are arbitrary orthologous genes from these genomes.

Cellular compartments differ in the selective constraint acting on their enzymes. We next investigated whether an enzyme's tolerance for amino acid substitutions depends on its subcellular localization. This analysis is somewhat less straightforward than it might appear both because some reactions (and hence their enzymes) occur in multiple compartments and because some reactions have multiple isoenzymes. As a result, different cellular compartments can contain the same enzyme. However, the set of overlapping enzymes is in general small and thus unlikely to weaken the power of our analysis significantly (Figure 6). For clarity, we defined proteins involved in transport reactions to be their own distinct category: such reactions have their reactants and products in different compartments.

The mean value of ω varies from 0.0935 in the nucleus to 0.1735 in the peroxisome. To determine if the differences in ω values are significant across compartments, we first clustered the compartments by mean (UPGMA; Sneath and Sokal

1973). The resulting three groups, in order of increasing ω, are: the Golgi apparatus and the nucleus, all other compartments except the peroxisome, and finally the peroxisome (Figure 6). We tested for significant pairwise differences between compartments in ω using a Mann-Whitney U-test (Figure 6) at a significance level of a = 0.01 (to account for the inherent multiple testing issues). The tests were conducted in a nested fashion, such that groups for which we could not reject the hypothesis of equal values of ω were compared to their nearest neighbors (c.f., the tree in Figure 6). This procedure allowed us to make seven comparisons, rather than the 56 possible pairwise comparisons. We find that the distributions clustered with low ω values (the nucleus and Golgi apparatus) are statistically indistinguishable ($P = 0.047$). Those in the large intermediate cluster can be split among groups that are statistically indistinguishable including lysozyme and transport compartments ($P = 0.087$), endoplasmic reticulum and external reactions ($P = 0.205$), and the cytosol and mitochondrial compartments, which are both statistically distinct from each other ($P < 0.01$). The peroxisome is also statistically distinct from the remaining compartments ($P < 0.01$).

Network construction. We next explored the role of metabolic network structure in influencing selective constraint, using the metabolic network of Duarte et al. (Duarte, Becker et al. 2007). This network includes information on reaction compartment and directionality that were used to create a semi-directed metabolic network where reactions are nodes. Two nodes are connected by an edge if they share a metabolite. Note that because metabolites are compartment-specific, edges do not connect reactions in differing compartments. Edges are also disallowed if the two reactions in question are

irreversible and the interconnecting metabolite serves as a substrate in both reactions or a product in both. The resulting network has 298,004 edges and 3,741 nodes, of which 2,264 have at least one associated gene (Figure 4).

Removal of currency metabolites. One of the implicit steps in preprocessing metabolic networks is removing currency metabolites, such as water and ATP that participate in numerous reactions. Failing to remove such metabolites prior to analysis can lead to an overestimation of connectedness between reactions.

Rather than introducing an arbitrary cutoff to define currency metabolites, we sought to use to the structure of the network itself to identify them. Other authors have defined and systematically removed currency metabolites from their networks based on their knowledge of the metabolic system (Huss and Holme 2007). Unfortunately the definition of currency metabolites is not consistent in the literature. Therefore, the network statistic we chose to identify currency metabolites is modularity. Newman (Newman 2006) defines a measure of optimal modularity, Q, as the quality of the subdivision of a network (measured as the fraction of vertices within clustered subdivisions minus the expected fraction of vertices with the same subdivisions in a randomly drawn graph) (Newman and Girvan 2004). Huss and Holme (Huss and Holme 2007) introduce ΔQ, which is Q for the empirical network minus the average Q of a number of random networks. As we remove increasingly less common metabolites, the ΔQ of most cellular components has a well-defined maxima (i.e., what modular structure was present in the network is eventually lost as more and more metabolites are removed). Interestingly, when we either consider the network as a whole or the reactions of the

cytoplasm alone, the resulting analysis does not present such a well-defined maximal ΔQ (Figure 7; Additional file 3, Figure S3), and we propose two reasons for this discrepancy. First, the large number of reactions means that removing certain metabolites (such as H^+, responsible for half the edges in the network) dramatically changes the network topology, yielding instability in the modularity measurements (see *Methods*). Second, many of the reactions in the cytoplasm are transporters. Because such transport reactions link distinct modules (i.e., compartments) in the network, it is expected that would they behave suboptimally in a modularity analysis.

Correlations between graph properties and **ω.** We investigated the relationship between two measures of network topology and the selective constraint on the genes associated with network reactions. The measures of reaction importance were the node degree and the betweenness centrality. Interestingly, there is a weak, but statistically significant correlation of betweenness centrality and ω (Figure 8: Spearman's $r = -0.279$, $P < 10^{-4}$), but no significant correlation between node degree and ω (Spearman's $r = -0.029$, $P = 0.075$). The network with currency metabolites included shows no relationship between network position and ω (Spearman's $r_{degree} = -0.03$, $P = 0.118$, $r_{betweenness} = -0.01$, $P = 0.587$).

There could be several sources of error associated with such an analysis of network structure and selective constraint. One obvious one is the compartment-by-compartment differences in average selective constraint already described. To explore the role of compartmentalization on this association, we examined the relationships between centrality and ω on a per-compartment basis (Table 2), finding that four

compartments had statistically significant association between degree and ω and three had significant associations of betweenness and ω. Oddly, we found a significantly positive association between these variables in the lysozyme.

Positive selection among the metabolic genes cannot explain the associations seen. We found 52 sets of orthologous metabolic genes that showed evidence for positive selection, spread across all cellular compartments (ranging from 0.7% of mitochondrial genes to 6.4% of cytoplasmic ones; see *Methods)*. Excluding these genes did not alter our compartment specific estimates of ω, the correlations between network statistics and ω or the significance of the differences in ω between compartments (data not shown).

There is a weak relationship between gene expression and selective constraint. Using 4,105 genes in both our sample and the HUGE Index (Haverty, Weng et al. 2002) we found a weak statistical relationship between ω and maximum expression level (r = -0.081; $P < 10^{-6}$). For metabolic genes this correlation is somewhat stronger (r = -0.089; P = 0.029). This relationship, however, is weaker than the relationship we find between network position and ω in metabolic genes, implying that expression may not the dominant predictor of selective constraint in mammals in the same way it is in yeast (Drummond, Raval et al. 2006).

2.4 Discussion

Our conclusions that gene function, expression, cellular localization and network position influence selective constraint will individually come as little surprise to researchers. This is especially true of our conclusion that purifying selection acts more strongly on

metabolic genes than on genes from the genome at large: function is a known correlate of rate of evolution (De, Lopez-Bigas et al. 2008; Lopez-Bigas, De et al. 2008; Tuller, Kupiec et al. 2009).

While we find a significant correlation between reaction centrality (betweenness) and selective constraint in the metabolic network, this result comes with several important caveats. First, although it is reasonable to interpret K_a/K_s as the level of selective constraint a gene experiences, in fact, this statistic represents an average evolutionary rate: in particular, two genes with the same fraction of amino acid substitutions forbidden by natural selection might have differing values of K_a/K_s if one gene had undergone more adaptive amino acid substitutions. We have partly controlled for this effect by omitting orthologs with evidence of positive selection, but it is not currently possible to completely remove this effect. Another caveat is that the association of betweenness-centrality and (apparent) constraint disappears when the currency metabolites are included. It is also worth noting that node degree on its own is not predictive of constraint in mammals, similar to the lack of association between these variables seen in *E. coli* (Hahn, Conant et al. 2004). We suggest one useful message to take from this result is that the relationship that exists between selective constraint and betweenness centrality is dependent on the manner in which the network is constructed. Special care has been taken in justifying the removal of currency metabolites across networks, however different removal strategies produce different associations of centrality and constraint (*Methods*).

From a more general perspective, it is also important to recall that networks are only computational abstractions of a biological reality. To speak of an association of

betweenness and selection is therefore actually to suggest that betweenness, a measurable quality, also represents an underlying biological feature. In this work, we have not directly demonstrated such a biological association. Likewise, there is a difference between the metabolic network associations seen here and those in protein interaction networks. In protein interaction networks, the pairwise binding of proteins is directly mediated by sequences, and natural selection can act to maintain complementary sequences in two interacting proteins (Bloom and C 2003). In metabolic networks, the relationship is more tenuous; one assumes that central reactions are required for proper function of the metabolic network and hence enzymes catalyzing such reactions will be under greater constraint. Even if this argument holds, the constraint is on function and not specifically on sequence. If an enzyme can maintain this function using differing sequences, there might be no necessary association of sequence constraint and centrality.

When we break the metabolic network down by compartment, we do find associations between network centrality (degree or betweenness) and constraint in some, but not all, compartments (Table 2). One lesson from these complex results is that although it is intuitive to consider the relationship between metabolic network structure and selective constraint at a global level, differences in constraint among compartments may confound global analyses. Likewise, the variation in constraint among these compartment raises interesting questions: it is unclear why enzymes from the Golgi apparatus and nucleus should be more highly conserved than those from the central group of compartments (Figure 6). Strikingly, enzymes implicated in external reactions fall within this central group, and are not distinguished by having a uniquely fast or slow rate of substitution. This result contrasts the findings by Liao et al. (Liao, Weng et al. 2010)

and Julenius and Pedersend (Julenius and Pedersen 2006) that the intra- / extra-cellular localization of a protein is highly predictive of its ω. However, note that these authors considered all genes in a given compartment, as opposed to the strictly metabolic ones analyzed here.

One potential explanation for these differences in constraint between compartments is that those compartments have different tolerances for misfolded proteins. Protein misfolding appears to have a significant fitness cost in yeast (Drummond, Bloom et al. 2005), and it is not unreasonable to hypothesize that the spatial organization of the nucleus (Fraser and Bickmore 2007) might induce a particularly high cost for misfolded proteins. However, one observation that speaks against this hypothesis is the weak association of constraint and expression. Our results thus suggest that although gene expression in some manner constrains mammalian protein evolution, it is less effective at doing so in mammals than in yeast.

2.5 Conclusions

In general, we find that although the position of a mammalian gene's product in the metabolic network and its expression level are both associated with that gene's evolutionary constraint, neither factor is determinative. Thus, unlike yeast, the forces that determine the selective constraint on mammalian protein-coding genes are likely both to be complex and to vary between genes.

2.6 Methods

Orthology identification. Our method for orthology identification first detects homologous genes using sequence similarity and then uses gene order to resolve orthology (orthology and paralogy are reviewed in Koonin 2005). Specifically, we first

conduct a pairwise homology search among all genes in *G1* and *G2* using GenomeHistory (Conant and Wagner 2002). GenomeHistory hits were filtered to exclude those with E-values greater than 10^{-10} (comparisons to chimpanzee and macaque) or 10^{-9} (all other comparisons) and amino acid sequence identity less than 50% (chimpanzee and macaque) or 45% (all others). Cases where two homologous genes are immediate neighbors on a chromosome (e.g., tandem duplicates) are treated as a single locus. An initial ortholog pair *A* and *B* is inferred if three criteria are met:

- *A* is from *G1* and *B* is from *G2*.
- The only homology of *A* in *G2* is *B* and the only homology of *B* in *G1* is *A*.
- The synonymous divergence between *A* and *B* is less than a threshold (K_s<0.5 for the human-chimpanzee and human-macaque comparisons and <0.75 in all other cases).

Many genes in *G1* and *G2* will have multiple homologs and hence not fall into a one-to-one relationship. Instead, we use this smaller set of one-to-one relationships to detect further orthologs. First, define *C* as the immediate (left or right) neighboring gene of *A* and *D* as the neighbor of *B*. If *C* and *D* are homologs, even if they also show homology to other genes, they are defined as orthologs. Importantly, now that *C* and *D* are identified as orthologs, their other homology relationships are deleted. We repeat the procedure for identifying one-to-one pairs, no longer using criterion 3. The entire process is repeated until no further orthologs are identified [9].

Sequence alignment and quality control. We created a data pipeline using Bioperl 1.6 (Stajich, Block et al. 2002). The initial inputs were the set of gene orthologs determined

above: we found 19,416 ortholog sets for the 8 eutherian mammal species. This set is made up of protein-coding genes with no clear evidence of tandem duplication, since duplication and subsequent functional specialization can alter measured selective constraints (Kondrashov, Rogozin et al. 2002). Human genes with orthologs in fewer than 5 other mammals were excluded from analysis.

Given a set of orthologous genes from humans and at least five other mammals, the corresponding amino acid sequences were aligned using MUSCLE v3.6 with default parameters (Edgar 2004). We next performed several filtering steps to assure alignment quality. First, we required that all possible pairs of sequences in each alignment have pairwise percent identity (PID) of $\geq 40\%$. If any pair of sequences had a PID $< 40\%$, then the sequence with the lowest PID to a consensus sequence was removed. The remaining sequences were then realigned and their PID rechecked. Next, we removed gap columns from the finished alignments. In cases where this resulted in fewer than 50 aligned amino acid columns, the sequences with the lowest PID to the consensus sequence was removed and the original sequences were realigned. This was done iteratively as long as there were still more than 5 sequences to align. The result of these filtering steps was the 13,928 multiple sequence alignments used in the remainder of our analyses.

Estimation of selective constraint. Using the above amino acid alignments, we inferred codon-preserving nucleotide alignments and estimated the ratio of nonsynonymous to synonymous rates (K_a/K_s or ω) with the codeml package in PAML 4.2 (Yang 2007). We assumed the sequences had evolved under previously published mammalian phylogenetic

relationships (Murphy, Pevzner et al. 2004; Nishihara, Hasegawa et al. 2006), namely ((((human, chimpanzee), macaque), (mouse, rat)),((horse, dog), cow)).

PAML model M0 was used to estimate the maximum likelihood value of ω (Yang, Nielsen et al. 2000). Recall that the synonymous substitution rate is used as a proxy for the mutation rate: the deficit or surplus of nonsynonymous substitutions relative to this value is then indicative of purifying or diversifying selection.

Identification of metabolic genes under positive selection. With these same data, we used a site-specific model to look for genes under positive selection. We compared PAML models M1a and M2a, nearly-neutral and positive selection models, respectively (Wong, Yang et al. 2005). M1a has two ω parameters: $\omega_0 < 1$ and $\omega_1 = 1$. M2a has three ω parameters: $\omega_0 < 1$, $\omega_1 = 1$, and $\omega_2 > 1$. A likelihood ratio test was used to determine if M2a was a significantly better fit to the data than M1a, given that model M2a has two more free parameters. Genes that had twice the difference in log-likelihood greater than the critical χ^2 value (5.99; $P < 0.05$) were assumed to be under positive selection.

Gene expression analysis. We collected expression levels for 4,105 genes by querying Affimetrix microarray data in the HUman Gene Expression Index (HUGE Index; Haverty, Weng et al. 2002).We then determined the maximum level at which a gene is expressed in the 19 tissues, comprising 59 experiments in the HUGE Index database.

Distributional fits. We tested whether the metabolic and non-metabolic ortholog sets had differing values of ω using a nonparametric Mann-Whitney U-test (Sokal and Rohlf

2000) as implemented in R. Differences in ω between cellular compartments were also analyzed with this test.

As discussed in the *Results*, because the data in question were visually skewed, we sought to confirm the results of the Mann-Whitney test of differences in ω between metabolic and nonmetabolic genes in two ways. First, we bootstrapped samples of size n =1,190 (the number of metabolic orthologs in our dataset) from the set of non-metabolic ω values and compared the sample means to the actual mean ω of the metabolic genes. Second, to compare not only the means of the two sets of genes but also their variability, we fit several probability density functions to these data using the MASS library in R (Venables and Ripley 2002). The distributional parameters were estimated by maximum likelihood: numerical optimization was carried out using Nelder-Mead or Broyden–Fletcher–Goldfarb–Shanno methods for single and multi-parameter distributions, respectively (Venables and Ripley 2002). We used these estimations to calculate the Pearson's correlation coefficient for Q-Q plots of quantile values versus observed frequency (Additional file 1, Figure S1). The result of this analysis was a list of distributional families ranked in terms of the best fit of each distribution to the data (Table 1).

We used a likelihood ratio test (LRT; Sokal and Rohlf 2000) to compare the distributions of metabolic and non-metabolic ω values. The null model requires the set of ω values from both metabolic and non-metabolic genes to be drawn from a single probability distribution. The alternative model allows the metabolic and non-metabolic genes to have differing values of the distribution parameters (while still following the same distribution function). As a result, the alternative model has twice as many free

parameters ($2p$) as does the null model, allowing us to compare the difference in log likelihood between the two models to a χ^2 distribution with p degrees of freedom.

Network properties. Modularity, degree, and betweenness estimates for the metabolic networks were calculated using the igraph library (Csárdi and Nepusz 2006) in R. Modularity was estimated using the Clauset et al. algorithm for detecting community structure (Clauset, Newman et al. 2004). Betweenness was calculated using the Brandes algorithm (Freeman 1977; Freeman 1979; Brandes 2001).

Removal of currency metabolites. As discussed in the *Results*, and following Huss and Holme (Huss and Holme 2007), we defined currency metabolites as metabolites whose removal increased the effective modularity (ΔQ) of the network in question. Modularity is a measure of the degree to which nodes fit into distinct and connected subunits (Newman 2006). Effective modularity is the difference between the maximum modularity of the graph and the average maximum modularity of a number of random graphs (Huss and Holme 2007). Removing currency metabolites increases modularity, since it removes interconnections between distinct subgraphs, while removing non-currency metabolites decreases modularity by isolating reactions from their subgraph modules.

To calculate ΔQ we compared modularity (Q) (Newman 2006) of the graph to the average Q of 1000 randomly rewired graphs. The Q for the graph minus the average Q of these 1000 random graphs is ΔQ (Huss and Holme 2007). We thus ordered the metabolites by the frequency with which they formed edges, removed the most frequent

metabolite and calculated ΔQ. We then reordered the metabolites and repeated this procedure until any further removal of metabolites only decreased ΔQ.

We optimized ΔQ for three different types of networks. First, we used a noncompartmentalized network (i.e., we removed compartmental metabolite designations so that ATP in the cytoplasm is treated equivalently to ATP in the mitochondria, Figure 7A). We next considered a network where we retained compartmental metabolite designations (where a specific metabolite may be removed from one cellular component, but not another) and optimized the global compartmentalized network (Figure 7B; Additional File 2 - Supplemental Table T1). Finally, we optimized the modularity in each cellular compartment individually (Additional File 3 - Supplemental Figure S3). Because the compartmentalized network offered both improved biological intuition and better performance in our modularity analysis, it was used for all of the analyses presented above. We note, however, this noncompartmentalized network shows a weaker association of betweenness and ω than does the compartmentalized one ($r_{betweenness} = -0.07$, $P < 10^{-4}$).

CHAPTER 3 SELECTION FOR HIGHER GENE COPY NUMBER AFTER DIFFERENT TYPES OF PLANT GENE DUPLICATIONS

3. 1 Abstract

The evolutionary origins of the multitude of duplicate genes in plant genomes are still incompletely understood. To gain an appreciation of the potential selective forces acting on these duplicates, we inferred the set of metabolic gene families from ten flowering plant (angiosperm) genomes using a phylogenetic approach. We then compared metabolic fluxes for these families, predicted using the *Arabidopsis thaliana* metabolic network, to the families' duplication propensities. For duplications produced by both small scale (SSD) and genome duplication (WGD), there is a significant association between flux and the tendency to duplicate. Following this global analysis, we made a more fine-scale study of the selective constraints observed on plant sodium and phosphate transporters. We find that different duplication mechanisms give rise to differing selective constraints. However, the exact nature of this pattern varies between gene families, and we argue that duplication mechanism alone does not define a duplicated gene's subsequent evolutionary trajectory. Collectively, our results argue for the interplay of history, function and selection in shaping duplicate gene evolution in plants.

3.2 Introduction

The contribution of gene duplication to evolution has long been a topic of interest in biology (Taylor and Raes 2004), but in the last ten years there has been a resurgence of

interest in the many and varied fates of such duplications (Zhang, Vision et al. 2002; Kondrashov and Koonin 2004; Adams and Wendel 2005; Aury, Jaillon et al. 2006; Rodriguez, Vermaak et al. 2007; Barker, Kane et al. 2008; Liang, Plazonic et al. 2008; Ha, Kim et al. 2009; Innan and Kondrashov 2010; Ramsey 2011). Among those fates, the important role played by genetic drift and simple changes in gene *dosage* is increasingly appreciated. In several contributions, Lynch and coworkers have argued that the relatively small effective population sizes of multicellular eukaryotes could result in the fixation of many gene duplications through non-adaptive processes (Force, Lynch et al. 1999; Lynch and Conery 2003; Lynch 2007). These processes, of course, must coexist with simultaneous actions of natural selection acting on the duplications. For instance, selection may act on gene dosage in one of two ways. First and most obviously, duplication of a gene may increase the rate of transcription and hence translation of the encoded protein, increasing its abundance. We have previously referred to this possibility as selection on *absolute* dosage (Bekaert, Edger et al. 2011). Under circumstances where higher copy number is selectively beneficial (*e.g.,* the duplication of certain genes in mammalian immune systems that decrease disease susceptibility) we expect copy number polymorphisms to become fixated in populations (Blanc and Wolfe 2004; Kondrashov and Kondrashov 2006). Second, if duplications affect only part of the genome (*i.e.*, as in a single gene duplication or due to differential paralog loss after polyploidy), the differences in *relative dosage* between genes that have co-evolved together may introduce selective costs. This second concept is known as the *dosage balance hypothesis* (Freeling 2009) and has been explored by a number of authors (Papp, Pál et al. 2003; Freeling and Thomas 2006; Birchler and Veitia 2007; Edger and Pires 2009). In line with

this distinction, this work focuses on the role of selection on absolute dosage in determining the fate of gene duplications in angiosperms.

As the first complete genome sequences became available, their patterns of gene duplication were explored, to understand, among other questions, the role of natural selection in duplicate gene fixation (Lynch and Conery 2000; Gu, Cavalcanti et al. 2002; Wagner 2002; Gu, Steinmetz et al. 2003). The duplications themselves have multiple origins including whole-genome duplications (WGD or polyploidy), as well as segmental, tandem and other kinds of single gene duplications (referred to here collectively as small-scale duplications or SSDs; Cannon, Mitra et al. 2004; Thomas, Pedersen et al. 2006; Freeling 2009). The preponderance of polyploids among angiosperms (Wendel 2000) has led plant biologists to focus on understanding the patterns of duplicate gene loss and retention following WGD events (Bowers, Chapman et al. 2003; Blanc and Wolfe 2004; Blanc and Wolfe 2004; De Bodt, Maere et al. 2005; Maere, De Bodt et al. 2005; Pfeil, Schlueter et al. 2005; Sterck, Rombauts et al. 2005; Cui, Wall et al. 2006; Freeling and Thomas 2006; Paterson, Chapman et al. 2006; Schranz and Mitchell-Olds 2006; Town, Cheung et al. 2006; Tuskan, DiFazio et al. 2006; Tang, Wang et al. 2008; Barker, Vogel et al. 2009; Edger and Pires 2009; Soltis, Albert et al. 2009; Wood, Takebayashi et al. 2009; Duarte, Wall et al. 2010; Coate, Schlueter et al. 2011; Jiao, Wickett et al. 2011; Schnable, Pedersen et al. 2011). In this work, we consider more generally a set of gene duplications created by both SSD and WGD that we infer from ten angiosperm genomes: seven dicots (*Arabidopsis*, papaya, soybean, *Medicago truncatula*, poplar, peach, and grape) and three monocots (*Brachypodium distachyon*, rice, and sorghum).

The taxa examined have a long history of polyploidy. Within the eudicots, the oldest genome duplication event, γ, was an ancient hexaploidy that characterizes the Rosidae (sensu Soltis, Smith et al. 2011), if not the core eudicots (Gunneridae sensu Jaillon, Aury et al. 2007; Lyons, Pedersen et al. 2008; Lyons, Pedersen et al. 2008; Ming, Hou et al. 2008; Freeling 2009; Argout, Salse et al. 2011; Jiao, Wickett et al. 2011; Shulaev, Sargent et al. 2011; Soltis, Smith et al. 2011). Comparative genomics suggest that the lineage leading to poplar (*Populus trichocarpa*) had an additional WGD event while that of the thale cress (*Arabidopsis thaliana*) had two: β and α. That these duplications represent independent events in the two taxa is suggested by the lack of more recent events in both grape (*Vitis vinifera*) and papaya (Carica papaya; Figure 9; Jaillon, Aury et al. 2007; Ming, Hou et al. 2008; Tang, Wang et al. 2008; Freeling 2009). Analysis of the non-synonymous substitution rates in the soybean (*Glycine max*) genome has revealed two whole genome duplication events post-WGD-γ: a duplication shared with peanut (*Arachis hypogaea*), a basal legume, and a more recent duplication on the soybean lineage (Bertioli, Moretzsohn et al. 2009; Schmutz, Cannon et al. 2010). The 3:1 ratio of grape to rice (*Oryza sativa*) genomic segments suggests that the γ paleohexaploidy is dicot-specific (Jaillon, Aury et al. 2007). However, cereal monocots also have a whole genome duplication event, ρ, basal to their radiation (Paterson, Bowers et al. 2004); rice, sorghum (*Sorghum bicolor*) and purple false brome (*Brachypodium distachyon*) show no evidence of further WGD events (Throude, Bolot et al. 2009; Vogel, Garvin et al. 2010).

There is mounting evidence that gene duplications created by WGD differ in their ultimate fates from duplicates produced by other mechanisms (Seoighe and Wolfe 1999;

Papp, Pál et al. 2003; Blanc and Wolfe 2004; Blanc and Wolfe 2004; Cannon, Mitra et al. 2004; Aury, Jaillon et al. 2006; Thomas, Pedersen et al. 2006; Hakes, Pinney et al. 2007; Conant and Wolfe 2008; Freeling 2008; Edger and Pires 2009; Freeling 2009; Coate, Schlueter et al. 2011). To cite just one example (relevant to this work), Maere *et al.* (2005) found that ion transporters were over-retained after WGD but under-retained following SSD. Blanc and Wolfe's (2004) study of genome duplication in *Arabidopsis* reached similar conclusions about ion transporters but also found that genes involved in phosphate metabolism are significantly over-represented in duplicate following the most recent (α) polyploidy.

We are interested in the action of natural selection on gene duplicates and in particular whether dosage effects are a strong predictor of duplicate retention. In this work, we have taken both a "high level" phylogenomic approach and a "low-level" single-gene approach to looking for evidence of selection in the process of gene and genome duplications in plants. Our first analysis focuses on metabolism and extends our previous work in *Arabidopsis*, where we found an association between metabolic flux and some, but not all, of the *Arabidopsis* genome duplications (Bekaert, Edger et al. 2011). Specifically, we hypothesize that genes in families with high flux will be, on average, over-duplicated. Given that we have previously found significant differences in duplication propensity between cellular compartments (Bekaert, Edger et al. 2011; Hudson and Conant 2011), we also test whether the relationship between flux and duplication varies by compartment or functional category. Additionally, we hypothesize that WGD-produced and SSD-produced gene duplications will differ in their levels of post-duplication selective constraint. We evaluate this by narrowing our focus to a group

of ion transporters. Such transporters have been implicated by a number of authors as having an outsized influence on metabolic flux (Kacser and Burns notwithstanding; Brown, Todd et al. 1998; Pritchard and Kell 2002). Furthermore, their evolutionary behavior has been shown to be distinct from other metabolic genes after both SSD and WGD (Lin and Li 2010; Bekaert and Conant 2011). Given the complexity of plant genome evolution, limiting our analysis to single gene families in this manner has the additional advantage of allowing us to carefully distinguish WGD and SSD events.

3.3 Results

Computing gene families and flux values

We estimated the flux through each biochemical reaction in the *Arabidopsis* metabolic network using flux-balance analysis (Orth, Thiele et al. 2010), maximizing the production of new cell mass for a fixed input of either light energy (in photosynthetic tissues) or carbohydrates (in non-photosynthetic tissues, see *Methods*). The maximal flux values observed range from 0 to 1,400,360 (arbitrary flux balance units). We then coupled those data to a set of cross-genome gene families identified from the 10 plant genomes (*Methods*). The result was a set of 735 gene families with associated metabolic fluxes. Of these 735 gene families, 463 have absolute flux values greater than zero. These families vary in size from 4 to 306 genes. The number of non-null flux values associated with each family ranges from 1 to 13, with 90% having only one associated flux value and only three having 10 or more flux values. Those three families function as ATP syntases, phospholipid transporters, and cellulose synthases (functional Gene Ontology annotation from TAIR; Swarbreck, Wilks et al. 2008). The number of gene duplications

per family varies from 0 to 210, with a mean of 3.21 duplications per species per gene family. Reactions with no flux can result either from failure to include certain metabolites in the biomass reaction or from a reaction not being used in certain conditions. Because of the potential for error introduced by these two possibilities, we present our results both with and without null-flux reactions.

Correlation between number of duplications and maximum metabolic flux

The correlation between the number of duplications in a gene family and maximal flux is positive and statistically significant whether or not null-flux reactions are included (Spearman's r=0.166 and 0.330, respectively, $P<10^{-9}$; Table 3). Similar results are seen when the number of duplications per species per gene family instead of the total number of duplications is used (Table 3). Likewise, positive associations of flux and duplication are seen for both photosynthetic and non-photosynthetic tissues (Table 3).

Association of flux and duplication is neither taxa nor duplication-mechanism specific.

As described, these species share a history of WGD (Figure 10). We summed the number of duplications on each branch in Figure 10, separating those with lineage-specific whole genome duplications from those without. Duplications in both groups are significantly and positively correlated with maximum flux (WGD: r=0.163, P<0.001; SSD: r=0.169, P<0.001). Of course, the branches containing WGDs will also have some background level of SSD, meaning that the duplications on these branches will not be exclusively due to WGD. However, the similarity in correlations seen between the two types of branch suggests that a more careful accounting of duplicates is unlikely to yield different results.

Similarly, we found significant positive associations of duplication and flux for the monocot subtree as well as the eudicot tree with *A. thaliana* removed ($P<10^{-6}$, $P<0.001$). The similarity of the results for these subtrees implies that our results are not specific to *Arabidopsis*, even though the metabolic network used is from that organism.

Association of flux and duplication extends across compartments and functional annotations

Gene families were associated with GO Slim annotations (Additional File - Supplemental Table T2) for both cellular compartment and function. We found significant Spearman's correlations between flux and duplicate rate for metabolic gene families found in the chloroplast, mitochondria, and endoplasmic reticulum (Table 4). Likewise, gene families that have a role in DNA or RNA binding or metabolism and responses to stimuli or stress had significant correlations between number of duplications and metabolic flux (Table 5).

To determine whether duplication rates differed among compartments or classes, we used Wilcoxon rank-sum test (*Z-scores* in Tables 2 and 3). Although gene families could appear in more than one annotation group, families located in the nucleus, cytosol, plasma membrane, cell wall, and extracellular space were significantly over-duplicated compared to all other gene families (Table 4). None of the functional categories were significantly over-duplicated (Table 5), but gene families involved in DNA or RNA metabolism had significantly fewer duplicates.

Selection on sequence evolution of ion transporters

Phosphate transporters. We narrowed our focus to a number of ion transporters to explore the patterns of post-duplication sequence evolution. Phosphate transporters in *A. thaliana* are divided into four gene families. These families include the high affinity transporters (PHT1; Mudge, Rae et al. 2002; Poirier and Bucher 2002) which import ions across the plasma membrane and the mitochondrial (PHT3; Hamel, Saint-Georges et al. 2004) and chloroplast (PHT4; Guo, Jin et al. 2008) transporters which act in their respective organelles. Finally, low affinity (PHT2) phosphate transporters that are also localized to the chloroplast (Versaw and Harrison 2002).We thus inferred gene phylogenies for the four phosphate transporter families and for one sodium transporter family (see Methods). Although the topology of phosphate transporter gene families is easily reconciled to the species tree, none of the clades contained the 4:2:1 ratio of *A. thaliana* to *P. trichocarpa* to *C. papaya* genes that would be expected if all transporters had been retained following the α, β and *P. trichocarpa* -WGDs (Figures 10A-10B). The average selective constraint (K_a/K_s) for PHT gene families varies considerably from 0.076 in high-affinity transporters to 0.207 in low-affinity transporters (Table 6). In all cases, the branches following gene duplications show significantly higher K_a/K_s than do those following speciation (Table 6; but note that the small size of the low-affinity family limits the strength of our conclusion for that family). We also investigated the possibility that selective constraint was is associated with duplication mechanism by dividing the branches following duplications into those due to WGD and to SSD. Here the difference in selective constraint is less clear: for the high-affinity and chloroplast phosphate transporters, the K_a/K_s values for whole-genome duplicates are not significantly different than those for SSDs. Among the mitochondrial transporters, whole genome duplicates

have significantly higher K_a/K_s than small-scale duplicates, indicating that the selective constraint was weaker following WGD.

Sodium ion transporters. The angiosperm sodium ion transporters (NHX) are a single gene family responsible for keeping Na^+ concentrations at non-toxic levels (Rodríguez-Rosales, Gálvez et al. 2008). The sodium ion transporters have a lower average K_a/K_s than do any of the phosphate transporter families: 0.049 versus 0.076-0.207. Curiously, among the sodium ion transporters, paralogs have significantly lower K_a/K_s values than orthologs, indicating no release in selective constraint after duplication (Table 6). Genes duplicated by WGD seem to be under slightly less selective constraint than gene orthologs; however, small-scale duplicates seem to be under considerably higher selective constraint than either.

3.4 Discussion

Selection on plant gene duplications

Although it has been hypothesized that a substantial fraction of the surviving duplicate genes in the genomes of multicellular eukaryotes might be due to the neutral fixation of duplicates (Lynch and Conery 2003), other potential forces can also be involved in fixation (Kondrashov and Kondrashov 2006; Innan and Kondrashov 2010). Here, we have taken both a low level and a high level approach to looking for evidence of selection in the process of gene and genome duplications in plants.

Selection, sequence evolution and ion transporters

Part of our analysis focused on the patterns of sequence evolution of two families of ion transporters. Transporters sometimes appear to be the limiting step in metabolic pathways (Kacser and Burns notwithstanding; Brown, Todd et al. 1998; Pritchard and Kell 2002), a fact that may partly explain why their evolution after both SSD and WGD is distinct from other metabolic genes (Lin and Li 2010; Bekaert and Conant 2011). Limiting our analysis to single gene families also allows us to carefully distinguish WGD and SSD events as well as to model the selective constraints acting on these genes.

There are two primary hypotheses regarding the expected changes in selective constraint at the sequence level following gene duplication. Predominant neo-functionalization would predict $K_a/K_s > 1.0$ for recently neo-functionalized duplicates (Zhang, Gu et al. 2003; Hahn 2009). On the other hand, sub-functionalization (and likely neutral retention by drift) would suggest that K_a/K_s is elevated after duplication, but not to values exceeding unity (Hughes 1994; Zhang, Rosenberg et al. 1998; Force, Lynch et al. 1999; Lynch and Conery 2000). Importantly, both models predict an elevated value of K_a/K_s after duplication: however, evidence for such increases is mixed. Hughes and Hughes (1993) found no evidence for the relaxation of selective constraint in 17 genes in the tetraploid African clawed frog (*Xenopus laevis*). Kondrashov et al. (2002) found that recent paralogs were under significantly lower selective constraints than orthologs, while others (Lynch and Conery 2000; Kondrashov, Rogozin et al. 2002; Zhang, Gu et al. 2003; Jordan, Wolf et al. 2004) have found evidence for a decrease in selective constraint immediately following duplication. This relaxation appears to be temporary: Jordan et al. (2004) found an increase in the average purifying selection acting on duplicates when compared to non-duplicated genes, presumably following what Innan and Kondrashov

(2010) have referred to as the *fate determining mutation*, which breaks the selective symmetry of the duplicates and sends them down differing paths. Our results parallel those of Jordan et al. (2004) in finding a general relaxation of selective constraint after ion transporter duplication. We also extended our analysis to potential differences in constraint between duplicates produced by SSD and by WGD. We had no *a priori* hypothesis as to whether SSD or WGD would result in a greater relaxation of constraint, and indeed we found that while there were often differences between the duplicates produced by the two mechanisms, the direction of these differences was not consistent.

Associations between duplication propensity and metabolic flux

We also made a very large-scale analysis of the patterns of evolution in the metabolic network. To our knowledge, this analysis represents the first high-level phylogenomic–scale study of gene duplication and metabolism in angiosperms (see Gout, Duret et al. 2009; van Hoek and Hogeweg 2009 for studies of metabolism following WGD in other organisms). By focusing on metabolism, we can ask whether duplications are randomly distributed across the network (as might be expected if drift were the only force at work) or show biases in the patterns of fixation. Notably, we find that there is a statistically significant relationship between duplication propensity across these ten angiosperm genomes and the predicted metabolic flux through the enzymes in question. This analysis follows our work on absolute and relative dosage among *Arabidopsis* WGD duplicates (Bekaert, Edger et al. 2011), where we found that reactions with high flux were enriched for enzymes coded by duplicate genes produced by the ancient b event (but not the more recent a event). Here, we have shown that the relationship between flux and the number

of gene duplicates is not specific to *Arabidopsis* but is rather found across all ten genomes. While it is certainly not the case that all gene duplications are associated with high flux reactions (the association magnitudes found are small), selection for increased gene dosage (Kondrashov and Kondrashov 2006; Conant and Wolfe 2008) is an attractive explanation for the fixation of some of these duplicates. In fact, examples of plant duplications apparently fixed by such selection are well known (van Hoof, Hassinen et al. 2001; Widholm, Chinnala et al. 2001).

Since the association of flux with duplication holds for both SSD and WGD events, we propose that different types of selective environment favor dosage-based duplicates produced by the two duplication mechanisms. Thus, SSD may be useful in situations where the increased dosage would be beneficial at the tips of a pathway or in secondary metabolism: this is likely the case for the copper tolerance duplication in bladder campion (Silene vulgaris van Hoof, Hassinen et al. 2001). However, Kacser and Burns (1981) pointed out that, for most metabolic pathways, it is unlikely that a single reaction is flux-limiting, meaning that a single gene duplication is unlikely to alter the flux in such a pathway. WGD is a potential route to increased flux in such situations, and it appears that such selection may have occurred after a genome duplication in the ancestor of bakers' yeast (Saccharomyces cerevisiae; Conant and Wolfe 2007; Merico, Sulo et al. 2007; van Hoek and Hogeweg 2009).

Taking these analyses to the subcellular level, we find strong correlations between flux and duplication in the mitochondria, chloroplast, and endoplasmic reticulum but not in the cytosol. This result suggests that the general association between flux and

duplication is primarily driven by reactions in these compartments, an unsurprising conclusion given the roles of the chloroplast and the mitochondria as the plant cell's anabolic and energy-yielding centers. These patterns also accord well with our prior analyses of compartmental evolution in the *Arabidopsis* and human metabolic networks (Bekaert, Edger et al. 2011; Hudson and Conant 2011).

Gene and genome duplication, selection and contingency.

Although a WGD that occurs in a particular individual is overall much less likely to be selectively neutral than is a single-gene duplication (Vieta 2005), it does not follow that there should be strong selection on every gene duplicated in such an event. Although this might suggests that WGD produces a large class of duplicate genes that evolve more or less neutrally after WGD, this hypothesis is difficult to reconcile with observations such as the dosage balance hypothesis. To distinguish between these two hypotheses, one might consider the selective constraint of all WGD-produced duplicate genes in a genome and to ask what the sources of variation in this statistic are. In fact, the number of sources of variation in constraint among duplicates at large (Duret and Mouchiroud 2000; Pál, Papp et al. 2003; Drummond, Raval et al. 2006; Vitkup, Kharchenko et al. 2006) suggests the importance of *contingency* in duplicate evolution. In other words, a duplicate's fate will depend on both it's own intrinsic properties (including factors studied here, such as function, cellular compartment and duplication mechanism) as well as the environment it finds itself in at birth.

3.5 Methods

Estimation of metabolic flux

As previously described (Bekaert, Edger et al. 2011), we used the Systems Biology Research Tool v2.0.0 (Wright and Wagner 2008) to perform flux-balance analysis on the *A. thaliana* metabolic network (de Oliveira Dal'Molin, Quek et al. 2010). We estimated the maximal biomass production possible under photosynthetic conditions (a fixed level of photon import allowed, sugar imports forbidden) and non-photosynthetic conditions (photon import forbidden, fixed sugar imports allowed). In each case, we also made every possible reaction knockout whereby a given reaction's flux is constrained to null and the remainder of the network was re-optimized. After knockout, all fluxes were normalized by the value of the biomass flux. Then, for each reaction, we selected the observed maximum flux, across all conditions. By doing so, we find what is essentially an upper bound on the flux of each reaction. It would obviously be desirable to also estimate the sensitivity of the network to changes in flux through each reaction. However, because we do not have kinetic data for the entire network, these values cannot be estimated with flux-balance analysis. Instead, we compared this maximal flux to the duplication status of each reaction node. In cases where there was more than one flux value associated with a gene family, all possible flux values for that family were used in our association analyses, meaning that large gene families will not tend to be biased toward high flux because they encompass more reactions.

Gene family identification

We used the list of *A. thaliana* enzymes from the de Oliveira Dal'Molin et al. (2010) metabolic network to identify enzyme gene families in the genomes of 10 flowering plants (Figure 9; Arabidopsis thaliana, Brachypodium distachyon, Carica papaya, Glycine

max, Medicago trunculata, Oryza satica, Populus trichocarpa, Prunus persica, Sorghum bicolor, and Vitis vinifera The Arabidopsis Genome Initiative 2000; Young, Cannon et al. 2005; Ouyang, Zhu et al. 2006; Tuskan, DiFazio et al. 2006; Jaillon, Aury et al. 2007; Ming, Hou et al. 2008; Paterson, Bowers et al. 2009; Schmutz, Cannon et al. 2010; The International Brachypodium Initiative 2010; The International Peach Genome Initiative 2010; Vogel, Garvin et al. 2010). Homologous relationships were inferred using GenomeHistory (Conant and Wagner 2002), which calculated the non-synonymous substitution rate. or K_a, for all gene pairs with BLAST scores lower than 0.0001 (a rather loose threshold). Gene families were identified by single-linkage clustering with a cutoff in non-synonymous divergence of $K_a \leq 0.20$ for *A. thaliana*/*A. thaliana* comparisons and $K_a \leq 0.30$ for all other comparisons (Powell, Conant et al. 2008). Gene pairs with K_a values below these thresholds were treated as nodes connected by an edge in the provisional gene family networks. These K_a parameters were selected after analyzing the results of using different K_a thresholds. For each threshold, we iteratively removed single edges from the provisional gene families. The chosen K_a thresholds were the largest values that did not cause a noticeable change in the constituency of the provisional gene families when any single edge was removed (data not shown). Families with fewer than 4 member genes were excluded.

Phylogenomics of gene families

Multiple sequence alignments of the protein sequences for each gene family were computed with MUSCLE v3.6 (Edgar 2004) using default parameters. Codon alignments were deduced from those alignments having 50 or more amino acids. We then inferred

maximum likelihood gene trees using RAxML v7.0.4 (Stamatakis, Hoover et al. 2008) with a general time-reversible model and discrete approximation of the gamma distribution (GTR+Γ). Confidence values were assigned to the gene trees from 100 bootstrap replicates. A relatively limited number of replicates were computed because we only wished to use these bootstrap statistics to identify nodes in the phylogeny with low support (<65%) prior to gene tree/species tree reconciliation, We thus reconciled all inferred gene trees with the species tree in Figure 9 (Moore, Bell et al. 2007; Wang H, Moore MJ et al. 2009). To do so, we used NOTUNG v2.6 (Chen, Durand et al. 2000) to infer the most parsimonious pattern of gene duplication and loss. Gene tree nodes with less than 65% bootstrap support were treated as polytomies and allowed to rearrange in order to minimize the number of duplications and/or losses (in practice choosing support value thresholds between 50% and 80% produced similar results; data not shown). Using these parsimony reconstructions, we calculated the number of duplications (and number of duplications per species) for each gene tree.

Manual annotation of transporter gene trees

Coding sequences for the nine annotated PHT1s, one PHT2, three PHT3s, six PHT4s, and eight NHXs of *A. thaliana* were downloaded from TAIR (Swarbreck, Wilks et al. 2008). A BLASTP search of the *A. thaliana* genome with these 19 and 8 sequences identified no further phosphate or sodium transporters in the genome. We then used BLASTP to search for ion transporter homologs in the genomes of papaya and poplar. We retained genes with BLAST E-values less than 10^{-20} as putative members of a given transporter family. Our homology estimation procedure always placed genes from *C.*

papaya and *P. trichocarpa* into only a single *A. thaliana* transporter family. Gene trees were constructed as detailed above. In the case of the NHXs, one *A. thaliana* gene (At2g01980) aligned poorly with the other NHXs and was excluded from the alignment and gene tree.

We manually assigned nodes in these phylogenies as either speciation or duplication events (Figure 9). Nodes connecting genes from the same species were labeled as duplication events where nodes connecting genes from different species were labeled speciation events. Because we were working with only a handful of genes, it was possible to make a more accurate distinction between SSD and WGD genes for these transporters than was possible for the genome-scale analyses. Thus, whole genome duplicates were inferred in cases where the paralogs fit into distinct paralogous synteny blocks from the Plant Whole Genome Duplication Database (PGDD; Tang, Bowers et al. 2008). Nodes connecting gene paralogs that could not be assigned using the PGDD were inferred to be small scale duplicates (SSD; Figure 9). These manual duplication or speciation designations agreed with the automatic assessments of NOTUNG.

Selective constraint following speciation and duplication events in 5 families of ion transporters

The selective constraint (ratio of non-synonymous substitutions to synonymous substitutions i.e., K_a/K_s), for each gene tree was estimated by maximum likelihood under the MG/GY94 codon model (Goldman and Yang 1994; Muse and Gaut 1994): see Conant et al., for details (2007). We tested three nested models of evolution: requiring all branches to have the same value of K_a/K_s (R_Null), allowing different values of K_a/K_s for

branches following a speciation node from those following a duplication node (R_Dupl), and a model with differing values of K_a/K_s for branches following speciation, whole genome, and small-scale duplications (R_WGD). We compared these three models with nested likelihood ratio tests and evaluated statistical significance using the χ^2 distribution, knowing that R_WGD has one more free parameter than R_Dupl, which in turn has one more parameter than R_Null.

Analysis of constraints by GO Slim annotation

Gene ontology slim (GO Slim) annotations were obtained for each *A. thaliana* gene from TAIR. GO Slim categories were further condensed (Supplemental Table T1) and transferred to our gene families. Spearman's rank correlations between flux and number of duplications in each gene family were calculated in SAS (v9.2.1, Cary, NC) for all cellular compartments and cellular functions. Note that gene families could appear in more than one compartment or functional group. We applied a Bonferroni multiple-test correction equal to the number of either compartments or functional groups analyzed, resulting in respective values of α of 0.0055 and 0.0042.

We also used the Wilcoxon rank test (SAS v9.2.1) to ask if the number of duplications per gene family differed for each cellular compartment or cellular function as compared to the reminder of the genome. We used the same Bonferroni multiple test corrections as previously.

CHAPTER 4 PARALLEL, INDEPENDENT REVERSIONS TO AN EMBRYONIC EXPRESSION PHENOTYPE IN MULTIPLE TYPES OF CANCER

4.1 Abstract

Changes in gene expression provide a valuable frame of reference for explaining the development and progression of cancer. Many tissue types radically alter their gene expression profile after becoming oncogenic. We evaluate this change in gene expression in 8 different cancer lines by comparing their expression profiles to that of their associated differentiated tissues as well as profiles for proliferative human embryonic stem cells. We find that, for non-proliferative tissues, the alterations in expression after oncogenesis result in a profile that is significantly more similar to the embryonic expression profile than to the original tissue profile. We also find that the lists of co-similar spots among embryonic and tumor cells are clustered within gene regulatory, protein interaction and metabolic networks. There is however little overlap in these lists between cancer lines and no pattern shared among all cancers in this analysis. We conclude that the manner in which cancers instantiate a proliferative pattern of expression following oncogenesis is diverse and we find no uniform proliferative program among the cancers in this analysis.

4.2 Background

Multicellular organisms maintain numerous systems for controlling the organization and development of their constituent cells (Grosberg and Strathmann 2007). These checks are necessary in organisms that use cell differentiation to build complex organ systems and morphologies (Krakauer and Plotkin 2002). Individual cells are programmed to first

follow a developmental course and then assume particular functions through a combination of genomic control, epigenetic imprinting and various fate-determining signaling pathways (Saitou and Yamaji 2010). As a result, relatively few cells in an adult multicellular organism are programmed to grow and divide without restriction (Campisi and Di Fagagna 2007). However, one or a series of mutations, gene deletions, gene duplications, or epigenetic changes can break this delicate control system, resulting in proliferative cancer cells that follow a program of unrestricted division (You and Jones 2012). In the early stages of this change, it is expected that tumor cells have not evolved a new proliferative cellular program *ad hoc*, but through a series of mutations, primarily in the signaling and regulatory pathways, that return these cell lines to an existing the proliferative program already encoded in the genome, a program that exists to facilitate embryogenesis.

However, while this general picture is reasonable, understanding the precise details by which one or more mutations give rise to the known cancer phenotypes (the genotype to phenotype mapping problem) has proven to be a distinct challenge. Moreover, such knowledge would be more than academic interest, as improving our understanding of this process could allow in predictive phenotyping from tumor resequencing or improved drug design and targeting (Stratton 2011; Al-Lazikani, Banerji et al. 2012; Patel, Halling-Brown et al. 2012).

One approach to the problem has been genetic: the identification of risk alleles for cancer in population. For instance, GWAS studies should help identify loci involved in the original oncogenic transition because individuals with pre-existing variation here would be at higher risk of certain cancers. However, despite their promise, the risk

increase effect sizes in GWAS studies for cancers are low, with very few regions co-occurring across cancers (Freedman, Monteiro et al. 2011). Furthermore, studies of genomic breakpoints in resequenced cancer genomes report highly diverse and non-overlapping patterns among cancers (Malhotra, Lindberg et al. 2013). And the determination of a specific set of genes, that in high or low copy number generally lead to oncogenesis is a current 'dark area' in the data from the massive cancer genome projects (Garraway and Lander 2013).

Strikingly, while the genetics of cancers have proven complex and dissimilar across cancer types (Stratton, Campbell et al. 2009), there are some important common phenotypes observed (Jain, Nilsson et al. 2012). One of the most important of these common changes is tumor cells' switch in their primary mode of sugar metabolism. In particular, while most (resting) cells in the body prefer to respire sugars to carbon dioxide and water using oxidative phosphorylation in the mitochondria, tumor cells are much more likely ferment those sugars using only glycolysis. This change is not minor: oxidative phosphorylation as a primary mode of metabolism appears to have been ubiquitous in the 1-2 billion year historical span covering eukaryotes and may well be the causal explanation for their uniquely complex genomes (Lane and Martin 2010). The precise importance of this *Warburg* effect is still imperfectly understood (Mayfield-Jones, Washburn et al.), but one surprising connection it suggests is to cells in the body that are *supposed* to divide rapidly: embryonic stems. These cells also display Warburg-like phenotypes (Krisher and Prather 2012; Redel, Brown et al. 2012).

The extent to which this intimate connection between the metabolism of cancer and embryonic cells is the result of an epiphenomenal coincidence or a necessary

functional convergence driven by natural selection pressure is unknown (Ashrafian 2006). Several studies have drawn conclusions about this relationship through the comparison of a limited number of cancers to normal tissues, but, to our knowledge, none has directly made the requisite three-way comparison of tumor, tissue and embryonic cells using the surfeit of next-generation sequencing and gene expression data now available from multiple cancers.

Here, we seek to develop a model-based comparison, evaluating the expression profiles of various cancers with the expression profile of embryonic stem cells and adopt and explicitly network-based approach. Our goal is to evaluate the hypothesis that many tumor cell undergo a reversion to an embryonic pattern of gene expression. In principle, such a change might result from parallel changes in expression in particular genes or by convergence at a higher organization level.

4.3 Methods

Microarray Data Collection

There are 3 cell classes for which we used gene expression in this analysis 1) human stem cell expression data, 2) human tissue expression data and 3) associated tumor expression data. Expression data were collected from Affymetrix microarrays. To standardize the analysis, only experiments on the HG-U133_Plus_2 (NCBI: GPL570) platform were used. Gene expression for proliferative stem cells involved 7 human embryonic stem cell lines, 8 human induced pluripotent cell lines, and 2 fibroblast cell lines (NCBI GSE23402 Guenther, Frampton et al. 2010). To minimize cross-lab experimental error, only studies with tumor and associated-tissue expression experiments were selected. This resulted in 8

distinct cancer types (gastrointestinal cancer *GSE13911*, oral squamous cell carcinoma *GSE30784*, pancreatic cancer *GSE16515*, prostate cancer *GSE17951*, colorectal cancer *GSE23878*, leukemia *GSE15061*, breast cancer *GSE10780* and lung cancer *GSE19198*). Each experiment had a sizeable number of independent replicates from different individuals (16-134 individuals produced the normal tissue samples and tumorous tissues were drawn from 35-181 different individuals). Affymetrix microarray experiments are prone to particular kinds of visualization errors (i.e., smears). Because of this, we manually inspected each experimental CEL file to discount the presence of smears and smudges using the affy package in Bioconductor (Irizarry, Bolstad et al. 2003).

Statistical comparison of expression profile distance

Each microarray experiment was normalized and error corrected using a robust multi-array average (Irizarry, Hobbs et al. 2003). To allow values to be comparable among arrays the value for each spot intensity was then transformed by taking the intensity of $spot_i$ and dividing it by the sum of the intensity of all spots in that experiment:

$$transformed_i = \frac{spot_i}{\sum_j spot_j}$$

For each probe id in each class of experiment (tumor, normal and proliferative), a 3-way pairwise comparison was made using a Kolmorogov distance measure. Kolmorogov distance was used because it has statistical properties that do not assume the underlying distribution is known in advance. For each cancer type (gastric, oral, pancreatic, prostate, breast, lung, leukemia and colorectal), 3 distances have been produced: cancer-normal,

cancer-proliferative and normal-proliferative for each of the 54,675 probe ids in the Affymetrix HG-U133_Plus_2 microarray platform.

P-values for lists of co-similar genes

We would like to know if the lists of co-similar embryonic and tumor genes (which we will hereafter refer to as spots) are higher than would be expected. Ideally, this list would contain all the genes that share an embryonic and tumor expression profile, without any spurious spots. One of the challenges in comparing distances among experimental classes of different size and an unknown underlying distribution is in choosing p-values for significance in the difference in distances (a). In the presence of multiple tests, the least conservative approach is to set a to 0.05 or 0.01. Given the number of statistical tests (k=54,675), per dataset, there is a high likelihood of generating false-positives. One way of to reduce the number of potential false-positives is the Bonferroni-correction where $a'=1-(1-a)^{1/k}$, the value of which is exceedingly low for this set of experiments, of the order a' = 1.83e-07. There is a high likelihood of generating false-negatives under this strategy. To minimize the trade-off between missing coexpressed genes and spuriously reported coexpression among embryonic and tumor expression profiles, we randomly reshuffled the three cell classes, meaning that we created 1000 randomized datasets where the 3 class identity (cancer, normal, and proliferative) was randomly reassigned for each sample. For each dataset, we used a 2-sample Wilcoxon-test of difference to compare the randomly reassigned "embryonic" and "cancerous" cell classes to the "normal" class. We then sought to determine the highest a-value that resulted in no pair of randomly reassigned genes being judged significant (see Figure 11). This a-value was

then used for each cancer-normal paired dataset. In cases in which the cancer and embryonic expression values were found to be closer than the cancer and normal expression sets and normal and embryonic expression sets, a 2-sample Wilcoxon-test (using the previously determined a-value for significance) was used to compare the embryonic and tumor expression with the normal expression values. The genes that significantly differ in distribution were then assumed to be co-similar (Additional File - Supplemental Figure S4).

Network evaluation

We used four networks in our evaluation of tumor/embryonic co-expression (protein-protein (Pérez-Bercoff, McLysaght et al. 2011), gene regulatory (Schaefer, Schmeier et al. 2011), metabolic (Duarte, Becker et al. 2007), and functional annotation (Huang, Sherman et al. 2009)). The goal of our network analysis is to ascertain if the shared genes for a pair of cell classes also cluster in these networks (Eisen, Spellman et al. 1998). Since protein-interactions, metabolic reactions and gene regulation all work in concert to form the cells underlying machinery (Joyce and Palsson 2006), we also evaluated the combination of the protein interaction (PPI), metabolic (MN) and regulatory networks(GRN). This *combined network* (hereafter CN) is formally defined formally as

$$G(v,e) = \bigcup edges \supseteq \{PPI, MN, GRN\}$$

We used several methods to evaluate the clustering in these networks. We measured the transitivity (also known as the average clustering coefficient (Watts and Strogatz 1998)) for the CN, PPI, MN, and GRNs. We also measured the number of connected components. The statistical significance of these values were evaluated by bootstrapping 10^7 random iterations of the network and recalculating these statistics.

Fully random networks tend to be a poor representation of real-world networks (Chung and Lu 2002). One of the primary characteristics of real networks are their power-law degree distribution. Our randomization preserved the number of interactions for each node, while randomized which nodes interacted. This allowed us to retain each networks power-law degree distribution while still randomizing the edges (Viger and Latapy 2005).

In addition to measuring transitivity and the number and size of connected components (all of which can be measured directly), we also evaluated the fit of these networks into highly interconnected *communities* (Newman 2006). The methods for detecting these are not exact, since the fit of vertices into is known to be NP-hard (Brandes, Delling et al. 2006). The strength of communities was evaluated using a modularity statistic, which essentially measures the number of edges within communities versus the number of edges between communities. There were several classes of heuristics used in this approximation of maximum modularity (Lancichinetti and Fortunato 2009). Since we are essentially choosing among heuristics we implemented several of these classes, including *iterative removal of edges based on betweenness* (Newman and Girvan 2004), *greedy modularity maximization* (Clauset, Newman et al. 2004), *label propagation* (Raghavan, Albert et al. 2007) *,* and *random walk* (Pons and Latapy 2005) methods. The statistical significance of these was evaluated by generating 10^7 randomized degree-preserving networks and calculating the maximum modularity for using each heuristic for both the observed and random networks. Network analysis was conducted using the igraph package in R (Csardi and Nepusz 2006).

Functional analysis of network neighbors

We took the list of coexpressed spots for the combined network for each cancer type and evaluated the over-representation of functional classes among the largest 3 communities from the using the DAVID Bioinformatics Resource (Huang, Sherman et al. 2009). We limited the annotations to the Gene Ontology Biological Process and Metabolic Function annotations, and KEGG Pathway annotations. We ranked and evaluated the significance of annotations using Benjamini p-values, which are robust to multiple tests, false positives and hierarchical annotations and evaluated the 10 highest ranking annotation clusters (Sherman and Lempicki 2009). The statistical significance for any given community in the network was evaluated by taking the number of edges within the community for each node and the number of edges between communities for each node and calculating a Wilcox rank sum statistic.

4.4 Results:

Statistical comparison of expression profile distances

We found that expression distances between cancers and embryos were closer than expression distances between normal tissues and embryos for most genes in almost all the cancers (excluding pancreatic cancer: see Table 7). This trend suggests a pattern of shared expression between cancer and embryo for most genes. To evaluate the statistical significance of this trend, we used a binomial test with the null hypothesis that the tumor and normal tissue cell classes were equally likely to have genes that were close to the embryonic pattern (e.g., 50% of the time the tumor would be closer and 50% the normal tissue). For tissues that can be said to be proliferative in their healthy tissue state (white-blood cell and pancreatic B-cell) the proportion of spots where the expression distance

between embryo and cancer is less than the expression distance between embryo and healthy tissue varies between 0.481 and 0.543. For cancers in which the associated healthy tissues are non-proliferative (colorectal, oral squamous, prostate, gastrointestinal, breast, and lung cancers) these proportions range from 0.568 to 0.664 and are all statistically greater than 0,5 (i.e., genes are more likely to be similar in expression between tumor and embryonic cell than normal and embryonic cells $P<0.001$, Table 7).

Lists of co-similar genes

For each of the non-proliferative tissues, the empirically determined α-values, that evaluate whether similarity in expression between two cell classes is statistically significant, are of a similar order (from 2.07e-05 to 5.77e-05). They are all also roughly 2 orders of magnitude higher than the Bonferroni-corrected α'-values (1.83e-07). The number of genes that were found to be co-similar between the cancer cells and embryonic cells varies between 5514 and 9972 (Table 8 and Figure 11). The expected number of genes in these lists is <1 ($P<0.001$) and are based on 1000 random reassignments of cancer, normal and embryonic expression values.

Gene overlap

Given the similarities between six different non-proliferative cancers and the embryonic cell samples, one might expect that a common set of genes would have changed in expression across these six cancers. However, our results do not illustrate this trend: for the 6 sets of experiments reported in this study, no one gene was shared across the embryonically-similar sets of all size cancers. Thus, 19,210 cancer probes are enriched

for an embryonic expression profile in at least 1 experiment. This includes 8,916 that are more highly expressed than in the normal tissue and 10,294 that are less expressed in the normal tissue. The overlap among experiments is considerably lower. There are 2465 spots sharing expression between 2 or more experiments, 285 sharing expression between 3 or more experiments, 36 sharing expression between 4 or more experiments, and 0 sharing expression in 5 or more experiments. The results are similar when the Affymetrix probes are mapped onto Uniprot protein ids. When the Uniprot enzymes and transporters are mapped to the *H. sapiens Recon 1* metabolic model (Duarte, Becker, et al. 2007), the overlap decreases similarly, with the exception that 2 reactions (K+-Cl- cotransport and 3',5'-cyclic-nucleotide phosphodiesterase) which overlap expression in 5 different cancers (see Figure 12).

Cancer networks

For each of the 6 cancers in non-proliferative tissues, the combined networks, protein-interaction networks and metabolic networks have higher than expected average transitivity (P<10e-06), meaning that the co-similar spots in these networks form tight-knit interacting clusters (see Table 9). All of the combined networks, protein-interaction networks and metabolic networks also have a higher than expected number of clusters (P<0.01). This suggests a small number of large, highly interacting clusters that change in expression upon conversion to an oncogenic phenotype (Figure 13).

Unlike the previous three networks, the gene regulatory networks behave very differently. In particular, the gene regulatory networks have either non-significant or

lower than expected transitivity and a lower than expected number of gene clusters (P<10e-06).

The source of this difference may lie in the structure differences between regulatory networks and the other types of networks considered. Thus, it appears that regulatory networks are seldom highly interconnected (Guelzim, Bottani et al. 2002) because, unlike protein-interaction and metabolic networks that have interacting functional modules, gene regulatory networks instead show a strongly hierarchical structure in addition to being modular. For these networks modularity refers to a distinct and non-overlapping groups of co-regulated genes and their shared regulators.

Each of the 4 heuristics for estimating modularity (*greedy*, *edge betweenness, label propagation,* and *random walk*) strongly support the hypothesis of modularity across the 4 network types (Supplemental Table T3). In other works, each type of network, whatever their other differences in structure, tend to consist of distinct units with few interconnections between those units.

Annotation of network features

The 3 largest combined network clusters share many of the same annotation categories across all 6 cancers in non-proliferative tissues (see Supplemental Table T4). All of the 3 largest network clusters are statistically significant (within cluster edges > between cluster edges: Wilcox Test: P<2.2e-16). Taking the 10 highest scoring annotation clusters (based on Benjamini p-value), there are 44 categories shared between all 6 types. These fall broadly into the categories: transcription, nucleic acid metabolism, regulation of biosynthesis, and ATP binding; all of which are primary cellular functions. There are also

21 categories that are shared by 5 cancer types, which fall broadly into the categories: apoptosis, mitotic cell cycle, and phosphorylation. There are also 115 categories unique to each cancer. This includes categories like “negative regulation of DNA binding”, which is a specialization of transcription and DNA binding; uninformative categories like “spliceosome”; and categories like “mTOR signaling” which are expected to be important in both oncogenesis and embryonic stem cell differentiation (Zhou, Su et al. 2009; Dowling, Topisirovic et al. 2010).

4.5 Discussion

In this study, we found that breast, colorectal, gastrointestinal, lung, oral squamous and prostate cancers showed a distinct expression pattern similar to the expression pattern in embryonic cells. This strategy gives us a window into the genetic underpinnings of proliferative behavior in cancer. We find that the genes that share expression between cancer and embryonic cells form distinct clusters. This occurs in terms of gene regulation, protein interaction, and metabolism and suggests that these clusters are functionally significant. Despite this similarity in the formation of gene clusters, the clusters themselves and the genes of similar expression underlying them show very little overlap between different types of cancers. It is unknown whether this lack of overlap is due to the random nature of oncogenic events (e.g., mutation, gene duplication, gene deletion, and epigenetic changes) the selective microenvironment in which the cell resides; or the limited overlap in expression among the original associated tissue. However, each of these cancers express a large set of genes in patterns similar to those in embryos and also functionally distinct. Each of them does so in their own unique way.

This is not solely a function of scale, since we consider variation in gene, protein, and metabolic reaction. At each of these scales, the overlap sometimes shared across two or more cancer types but rarely across more than that.

We assert that cancer cells are individuals, from an evolutionary point of view (Merlo, Pepper et al. 2006), and that cancer phenotypes are, at that scale, not only functional, but potentially selectively advantageous (Bignell, Greenman et al. 2010). This presents something of a paradox. This year millions of people will get cancer. Yet, the manner in which cancer emerges is due to complex interactions between a large set of heterogeneous external factors (smoking, solar rays, pollutants, etc.) and various internal genetic predispositions. Importantly, the initial cancer or tumor development takes a relatively short time-period (as measured in numbers of cell divisions) and hence occurs in a small population of cells. Given this relatively limited space of evolution to operate, it may be surprising that cancers are so often able to dramatically change their expression profiles and phenotypes.

One possible explanation for why cancers do rapidly evolve and share so many aspects of their phenotype (the so-called *hallmarks of cancer*) is that cancer is the result of a small and simple set of aberrant genetic/protein/metabolic changes. Our results argue against this, as do the low effect sizes among GWAS studies. We find very little overlap in gene expression among multiple cancers, whether we consider individual spots, proteins, or metabolic reactions. We hope further analysis will be able to follow up this work and evaluate the extent to which the similarities between the programs of proliferation in embryos and tumors are superficial or causal.

FIGURE / ADDITIONAL FILE LEGEND

FIGURE 1. ***Yeast Genome Order Browser (YGOB).*** Yeast Gene Order Browser (YGOB) screenshots with a window size of six. Each box represents a gene; each color, a chromosome. The gene in focus, the *A. gossypii* gene *ABR086W*, is highlighted by an orange border. Each vertical column ("pillar") represents a single gene prior to the WGD (hence all genes in a column are homologs and the paired upper and lower genes, when present, are paralogs). The ancestral order of these genes (pink boxes) just prior to the WGD has also been exhaustively inferred (Gordon, Byrne et al. 2011). Connectors join nearby genes: a solid bar for adjacent genes, two bars for loci less than five genes apart, and one bar for loci <20 genes apart. The connectors are extended in gray over intervening space. The end of a chromosome or contig is denoted by a brace. Arrows denote transcriptional orientation. The browser also includes a control panel that allows users to select the window size and the gene to focus on. This panel also has buttons for running BLAST searches against YGOB's database, outputting YGOB data in tabular format, obtaining pairwise K_a and K_s values among genes, and computing multiple sequence alignments and phylogenetic trees of individual pillars. Species names for each track are labeled at *right* (Byrne and Wolfe 2005).

FIGURE 2. ***Consensus view of the evolutionary relationships between the yeast taxa discussed.*** Black branches indicate relationships described by both Kurtzman and Robnett (2003) and Fitzpatrick et al. (2006). Red branches indicate conflicts between the two phylogenies, in which case Fitzpatrick et al. (2006) is presented. Curved grey branches illustrate the allopolyploidy events between 2 species (*S. cerevisiae* and *S. bayanus, Z. rouxii* and *Z. pseudorouxii*). Taxa in blue are reported in text (e.g., *S. pastorianus* and *Z. rouxii ATCC 42981*). Stars mark whole genome duplications. Note that genus names are an imperfect guide to the relationships.

FIGURE 3. ***Genome evolution in Saccharomyces pastorianus.*** A model of the formation of *S. pastorianus* and the hybrid strains of *S. bayanus*. First, wild *S. eubayanus* and ale-type *S. cerevisiae* hybridized to form an allotetraploid that became the ancestor of the modern (doubly paleopolyploid) *S. pastorianus*. Second, domestication imposed strong selective pressure for strains with the most desirable brewing properties. Third, in the brewing vats with high densities of *S. pastorianus*, cell lysis releases large DNA fragments that occasionally transform, fourth, contaminating wild strains of *S. eubayanus* (which possesses only the ancient WGD shared with *S. cerevisiae*) because of the lack of pure culture techniques. Fifth, multiple hybridization events between *S. eubayanus* and wild strains of *S. uvarum* gave rise to CBS 380T and NBRC 1948. This model does not exclude prior or parallel involvement of *S. uvarum* in brewing or contamination. *Reprinted from Libkind et al. (2011).*

FIGURE 4. The human metabolic network. A) The full reaction network used in this analysis. Colors correspond to the compartment in which each reaction occurs. The network was visualized with Gephi 0.7 using the Force-Atlas layout algorithm (Bastian et al. 2009). **B)** Schematic view of network construction. Reactions that share a metabolite are joined by edges. Each reaction in **A** may also be associated with one or more genes; it is these genes for which we calculate the selective constraint.

FIGURE 5. Human metabolic genes are under greater selective constraint than other orthologous genes. A) The distribution ω for metabolic genes (red, n_m = 1,190), non-metabolic genes (black, n_o = 12,738), and the total set of orthologs (white, n_t $=n_m+n_o$= 13,922). The x-axis gives ω: the ratio of non-synonymous synonymous substitutions per site (i.e., our proxy for selective constraint, see main text). The y-axis is the number of genes with a given ω value for each gene set. We can reject the hypothesis that the distribution of metabolic genes is not significantly smaller than the distribution of non-metabolic genes (P = 0.035; Mann-Whitney U-test). **B)** The white portion of the circle graph shows the relative proportions of genes for which we cannot identify orthologs for the metabolic and non-metabolic genes (red and black, respectively). These proportions are significantly different (χ^2 = 11.98, P < .001). **C)** In no case does the mean ω for metabolic genes (red line) occur in the 1,000,000 resampled means for non-metabolic genes ($P < 10^{-6}$).

FIGURE 6. Exclusion of currency metabolites by maximizing effective modularity. Effective modularity (ΔQ; y-axis) is maximized by iteratively removing common metabolites (x-axis). The dashed horizontal line indicates the line corresponding to the ΔQ based on 1000 randomizations of the full network (ΔQ = 0). The vertical line corresponds to the points where the graph is maximized. **A)** When compartmental designations are included for each metabolite the ΔQ for the graph is maximized after the 24 most frequently occurring metabolites are removed. **B)** When compartmental designations are not included for each metabolite the ΔQ for the graph is maximized after the 20 most frequently occurring metabolites are removed. **C)** Results of modularity maximization for the individual cellular compartments.

FIGURE 7. Hierarchical clustering of cellular compartments based on selective constraint. Genes are split into nine groups based on their subcellular location (see main text). The box plots show the distribution of ω values for each compartment. Using the mean ω values, we created a phenogram using the UPGMA algorithm (branch lengths are arbitrary). Each branch is colored gold if a Mann-Whitney U-test found that the distributions were significantly different at $P \leq 0.01$, and blue otherwise. For example, the cytoplasm and mitochondria distributions are significantly different (P = 0.010), but the lysozyme and transport groups show no significant difference (P = 0.087). However,

when the group formed by the cytoplasm and mitochondria is compared to that formed by the lysozyme and transporters, there is a statistically significant difference ($P < 0.001$). The 4 Venn diagrams show the proportional degree of overlap in genes among groups (sizes are not comparable across nodes in the tree). In none of these cases is any one set of reactions a superset of the other set of reactions.

FIGURE 8. A negative association of betweenness centrality and selective constraint. The $\log_{10}$ transformed betweenness centrality for each reaction (*x*-axis) is plotted against the estimated selective constraint (ω) for its associated genes. The correlation between these two variables is negative and weak (Pearson's $r = -0.222$, Spearman's $r = -0.279$), but highly significant ($P < 0.0001$).

FIGURE 9. Plant species used in reconciling gene trees. The phylogenetic relationships among these species have been described previously (Moore, Bell et al. 2007; Paterson, Bowers et al. 2009). The branch histograms depict the number of duplications per gene family (*y*-axis) at each branch (branch lengths are arbitrary). The *x*-axis of these histograms are natural log-scaled and is consistent across histograms. Black circles indicate whole genome duplication events.

FIGURE 10. Ion transporter gene trees used in this study. Branches demarcating speciation events are colored orange, whole genome duplication events green and non-whole genome duplications purple. A) High-affinity phosphate transporters AtPHT1;1 – AtPHT1;9 with 16 *P. trichocarpa* and 7 *C. papaya* homologs. B) Low-affinity phosphate transporters AtPHT2;1 with 2 poplar and 1 papaya homologs. C) Mitochondrial phospate transporters AtPHT3;1-ATPHT3;3 with 6 poplar and 1 papaya homologs. D) Chloroplast phosphate transporters AtPHT4;1-AtPHT4;6 with 10 poplar and 6 papaya homolog. E) Sodium ion transporters AtNHX1-AtNHX6, AtNHX8 with 6 poplar and 4 papaya homologs.

FIGURE 11. *P-value* minimization for 2-sample Wilcoxon-test. Embryonic and tumor vs. normal gastrointestinal tissue. The *x-axis* corresponds to the various *p-values* chosen for this analysis, the *y-axis* corresponds to the number of probes in the sample. The black line and boxplots illustrate a 1000 sample reshuffling of embryonic, normal and cancer classes. The green line shows the obserged number of probes at given *p-values*. A log-log linear regression ($r^2 = 0.999$, P<10e-7) of the random bootstrapped samples shows the number of probes by random error (the false-positive count) is expected to be < 1 at P<3.497e-05. At this p-value the observed number of probes is 6351.

FIGURE 12. Overlap in significant probes among different cancer types for genes, proteins, and reactions. The size of spots corresponds to the number of probes shared between two experiments. Gene expression spots are from Affymetrix HG-U133_Plus_2 Microarrays. Protein values have been mapped onto Uniprot IDs. Reaction values have been mapped onto the *H. sapiens Recon 1* metabolic model.

FIGURE 13. Visualization of each network in gastrointestinal cancer. These visualizations show the fundamental features of each of these four networks. The gene regulatory network is has a small number of clusters and is not very highly interconnected. The protein interaction network has one large very highly interconnected cluster and many small satellite clusters (mostly made of pairs of proteins). The metabolic network has very few clusters, which are highly modular and highly interconnected. The combined network has one large highly interconnected cluster and many small satellite clusters.

ADDITIONAL FILES

ADDITIONAL FILE 1, SUPPLEMENTARY FIGURE S1 - Q-Q plots are for 8 common distributions. Weibull, gamma, exponential, logistic, normal, extreme value, log-normal and Cauchy. The X-Y line is y = x. The *x*-axis plots the theoretical quartiles for a statistical population from one of the 8 distributions, while the *y*-axis plots the data. Values that lie on the line y = x are a good fit between the theoretical distribution and data. The Weibull, gamma, and exponential distributions provide close visual fits to the data (see Table 1 for the correlations).

ADDITIONAL FILE 2, SUPPLEMENTARY FIGURE S2 - Ortholog identification. Homologous genes within and between genomes are first identified based on a lack of within-genome paralogs in both genomes. We then identify each pair of genes that are immediate neighbors of a pair of orthologs and are also homologous. Because these genes have other homologs in the other genome, they were not part of the initial ortholog list. We now define them as orthologs, and at the same time, remove any orphan genes that no longer show homology to genes in the other genome not already in orthologous pairs. Using the new pairs, we repeat the process until no further orthologs are located.

ADDITIONAL FILE 3, SUPPLEMENTAL FIGURE S3 - Maximum effective modularity for each compartment and for the total cellular metabolic network. Effective modularity (ΔQ) on the *y*-axis is maximized for each of the subcellular compartments, including organelles, external reactions, and the cytoplasm by iteratively

removing each of the most common metabolites (the number on the *x*-axis). The dashed horizontal line indicates the line corresponding to the ΔQ based on 1000 random iterations (ΔQ = 0). The vertical line corresponds to the points where the graph is maximized.

ADDITIONAL FILE 4, SUPPLEMENTARY TABLE T1 - Compartment specific currency metabolites removed from total network.

ADDITIONAL FILE 5, SUPPLEMENTARY TABLE T2 - Condensed Gene Ontology (GO Slim) annotations from TAIR and the cellular compartment and functional group categories they were condensed into.

ADDITIONAL FILE 6, SUPPLEMENTARY FIGURE S4 - Pipeline for statistical analysis of tumor / embryo co-similarity. d_1, d_2, d_3 correspond to Kolmorogov distance between the expression vectors for tumors, embryos and associated healthy cells. If d_1 is less than d_2 and is less than d_3 perform a Wilcox test. This test is for both embryo and cancer expression vs. normal expression. If the value is less than α (*see* Figure 11), the expression is co-similar between tumor and embryo.

TABLES

TABLE 1 - Log-likelihoods of a linear fit between all ω values and each of 8 common distributions, with likelihood ratio tests for the differences in distributions calculated for the 3 best distributional fits.

Distribution	Pearson's r[a]	k[b]	LRT[c]	df	*P*-value[d]
Weibull	.999	2	186.39	2	$<10^{-6}$
Gamma	.999	2	201.67	2	$<10^{-6}$
Exponential	.998	1	28.29	1	$<10^{-6}$
Logistic	.924	2	-[e]	-	-
Normal	.923	2	-	-	-
Extreme Value	.903	3	-	-	-
Log-Normal	.854	2	-	-	-
Cauchy	.163	2	-	-	-

a. Pearson's *r* is the linear correlation of the data to the quartiles, based on the maximum likelihood inferred parameters for each family of distributions.
b. k is the number of free parameters in the distribution.
c. Likelihood ratio test: LRT = 2 * (log-likelihood of metabolic ω values + log-likelihood of non-metabolic ω values) – (log-likelihood for all ω values).
d. Distributed χ^2.
e. Likelihood ratio test only performed for the best distributional fits.

TABLE 2 - Correlations between ω and the graph properties (degree and betweenness) for each compartment, including the number of reactions and edges in each compartment.

Compartment	r_{degree}/ω [a]	$r_{betweenness}/\omega$ [b]	# of reactions	# of edges
Nucleus	0.231	0.024	149	969
Endoplasmic reticulum	0.112	0.045	301	5706
External	-0.093	0.082	986	10279
Golgi apparatus	-0.130	0.103	343	1502
Cytoplasm	-0.168**	-0.193**	2095	196319
Mitochondria	-0.312**	0.043	594	23501
Lysozyme	-0.213*	0.294**	216	7440
Peroxisome	-0.331*	-0.455**	175	1662

a. Spearman's *r* rank correlation of degree and ω.
b. Spearman's *r* rank correlation of betweenness-centrality and ω.
* $P < 0.05$
** $P < 0.001$

TABLE 3 - Correlations between duplication and flux by gene family.

	All flux values		Excluding null-flux[a]	
	r[b]	***P***[c]	***r***	***P***
Duplications per gene family				
All conditions	0.166	$<10^{-9}$	0.330	$<10^{-9}$
Leaves	0.157	$<10^{-5}$	0.318	$<10^{-16}$
Roots	0.174	$<10^{-10}$	0.376	$<10^{-16}$
Duplications per species per gene family[d]				
All conditions	0.126	$<10^{-3}$	0.181	10^{-8}
Leaves	0.124	$<10^{-5}$	0.147	$<10^{-5}$
Roots	0.130	$<10^{-6}$	0.227	$<10^{-10}$

[a] Flux values equaling 0 can have confounding biological and computational meanings
[b] Spearman's *r*
[c] Correlations and statistical significance calculated in R.
[d] Number of duplication events per gene family divided by the number of species in that family.

TABLE 4 - Duplication status per gene family split by cellular compartment. Bold values are significant at a Bonferroni corrected $\alpha = 0.0055$.

		Duplication vs Flux[a]		Duplication[b]	
Cellular Compartment	**n**	**r[c]**	***P***	**Z[d]**	***P***
Nucleus	56	0.267	0.046	3.481	**0.0005**
Cytosol	74	0.174	0.138	4.910	**<0.0001**
Chloroplast and Plastid	273	0.260	**<0.0001**	-0.646	0.518
Mitochondria	134	0.443	**<0.0001**	1.012	0.311
Plasma Membrane	97	0.107	0.297	7.371	**<0.0001**
Endoplasmic Reticulum	44	0.459	**0.002**	-1.161	0.246
Golgi Apparatus	12	0.606	0.037	2.280	0.023
Cell Wall	52	0.359	0.009	3.210	**0.001**
Extracellular	51	0.130	0.363	5.115	**<0.0001**

[a] Duplications per gene family versus the maximum flux (excluding null values)
[b] Wilcoxon rank test of difference across compartments (positive values: over-duplication; negative values: under-duplication)
[c] Spearman's r, calculated in SAS (v 9.2.2, Cary, NC)
[d] Wilcoxon's Z, calculated in SAS (v 9.2.2, Cary, NC)

TABLE 5 - Duplication status per gene family split by functional annotation. Bold values are significant at a Bonferroni corrected $\alpha = 0.0042$.

		Duplication vs Flux[a]		Duplication[b]	
Function	**n**	**r[c]**	***P***	**Z[d]**	***P***
Cell Organization and Biogenesis	29	0.138	0.476	1.010	0.312
Developmental Processes	20	-0.055	0.817	1.693	0.090
DNA or RNA Binding or Metabolism	26	0.722	**<0.0001**	-2.284	0.022
Electron Transport	7	0.598	0.156	0.460	0.645
Hydrolase Activity	114	0.251	0.007	-1.661	0.097
Kinase Activity	62	-0.216	0.092	1.167	0.243
Nucleic Acid or Nucleotide Binding	94	0.066	0.529	0.011	0.991
Protein Binding or Metabolism	121	0.145	0.113	1.463	0.143
Signal Transduction	13	-0.128	0.676	2.450	0.014
Stimulus or Stress Response	199	0.248	**0.0004**	2.650	0.008
Transferase Activity	166	0.178	0.021	-0.833	0.405
Transporters or Transport	56	0.062	0.651	1.755	0.079

[a] Duplications per gene family versus the maximum flux (excluding null values)
[b] Wilcoxon rank test of difference across compartments (positive values: over-duplication; negative values: under-duplication)
[c] Spearman's r, calculated in SAS (v 9.2.2, Cary, NC)
[d] Wilcoxon's Z, calculated in SAS (v 9.2.2, Cary, NC)

TABLE 6. Selective constraint estimated with three models of gene evolution for ion transporters of *A. thaliana*, *C. papaya*, and *P. trichocarpa*.

		PHT1-High affinity phosphate transporter		**PHT2-**Low affinity phosphate transporter		**PHT3-**Mitochondrial phosphate transporter		**PHT4-**Chloroplast phosphate transporter		**NHX-**Sodium ion transporter	
Model	**Branches**	**K_a/K_s**	**-lnL**	**K_a/K_s**	**-lnL**	**K_a/K_s**	**-lnL**	**K_a/K_s**	**-lnL**	**K_a/K_s**	**-lnL**
R_Null	All	0.076		0.207		0.114		0.148		0.049	
			15379.5		3577.0		5898.3		24869.3		11210.1
R_Dupl	Speciation	**0.063**[a]		**0.156**[a]		**0.080**[a]		**0.123**[a]		**0.062**[a]	
	Duplication	**0.082**[a]		**0.415**[a]		**0.133**[a]		**0.249**[a]		**0.031**[a]	
			15376.7		3570.2		5894.1		24847.0		11188.5
R_WGD	Speciation	0.063		-[b]		**0.080**[a]		0.123		**0.061**[a]	
	WGD[c]	0.081		-		**0.190**[a]		0.233		**0.067**[a]	
	SSD[d]	0.085		-		**0.112**[a]		0.265		**0.018**[a]	
			15376.6		-		5891.2		24846.7		11175.3

[a] Significant improvement over the model immediately above at $P<0.05$; nested likelihood ratio test (distributed χ^2, $P<0.05$, degrees of freedom=1)

[b] No small scale duplications in PHT2, so model R_Dupl is equivalent to model R_WGD

[c] Whole genome duplication: determined by syntenic paralogy using the Plant Genome Duplication Database (Tang, Bowers et al. 2008)

[d] Small-scale duplication: determined either by a lack of syntenic paralogy and/or by tandem duplication status.

TABLE 7. Distance measures between cancer, normal and embryonic cells among proliferative (i.e., leukemia and pancreatic) associated tissue and non-proliferative associated tissues (i.e., colorectal, oral squamous, prostate, gastro-intestinal, breast and lung).

Cancer type	Number of cases of tumor and embryonic expression profiles being closest	Number of cases of tissue and embryonic expression profiles being closest	Proportion (tumor/tissue)
Colorectal	31549**	19650	0.616
Oral squamous	34683**	17946	0.659
Prostate	32978**	18184	0.645
Gastro-intestinal	27933**	21176	0.568
Breast	28418**	18598	0.604
Lung	35580**	17950	0.664
Leukemia	*28665***	*24152*	*0.543*
Pancreatic	*24345*	*26269*	*0.481*

** statistically significant for Binomial difference in equal proportions (proportion = 0.5) at 0.001

TABLE 8. The number of probes with co-similar expression between tumor and embryonic tissue for each cancer type at *P-values* determined to have fewer than 1 false positive.

Cancer type	Significant probes	Empirically determined *P-values*
Gastrointestinal	6351	3.49e-05
Oral squamous	6394	2.07e-05
Colorectal	6625	5.77e-05
Prostate	8959	3.25e-05
Breast	5514	3.27e-05
Lung	9972	3.07e-05

TABLE 9. Network statistics for each cancer type and biological network.

Cancer type	Graph	Vertices	Edges	Transitivity	Modularity	Number of Clusters
Breast	**Combined**	**1313**	**3388**	**0.312****	**0.686****	*33*
	Gene-Regulatory	302	421	*0.053*	**0.669****	**12****
	Protein-Protein	1007	1651	**0.081****	**0.611****	*44*
	Metabolic	303	1557	**0.527****	**0.711****	*2*
Colorectal	**Combined**	**1694**	**4169**	**0.221****	**0.641****	*44*
	Gene-Regulatory	493	688	*0.015*	**0.616****	**7****
	Protein-Protein	1402	2362	**0.047****	**0.597****	*58*
	Metabolic	292	1396	**0.553****	**0.703****	*6*
Gastrointestinal	**Combined**	**1490**	**3375**	**0.296****	**0.668****	*46*
	Gene-Regulatory	321	398	0.030	**0.684****	**34***
	Protein-Protein	1190	1760	**0.046****	**0.600****	*57*
	Metabolic	309	1458	**0.523****	**0.710****	*5*
Lung	**Combined**	**2371**	**6824**	**0.293****	**0.633****	*45*
	Gene-Regulatory	579	803	0.030	**0.652****	**16****
	Protein-Protein	1964	3694	**0.056****	**0.542****	*64*
	Metabolic	331	838	**0.523****	**0.685****	*4*
Oral squamous	**All**	**1444**	**3993**	**0.312****	**0.681****	*40*
	Gene-Regulatory	279	304	0.022	**0.780***	**24****
	Protein-Protein	1141	1868	**0.068****	**0.686****	*64*
	Metabolic	340	1962	**0.482****	**0.718****	*2*
Prostate	**Combined**	**2020**	**5331**	**0.216****	**0.625****	*44*
	Gene-Regulatory	488	631	**0.038****	**0.711****	**18****
	Protein-Protein	1668	3062	**0.057****	**0.533****	*56*
	Metabolic	363	1939	**0.441****	**0.732****	*5*

REFERENCES

Adams, K. and J. Wendel (2005). "Allele-specific, bidirectional silencing of an alcohol dehydrogenase gene in different organs of interspecific diploid cotton hybrids." Genetics **171**: 2139-2142.

Al-Lazikani, B., U. Banerji, et al. (2012). "Combinatorial drug therapy for cancer in the post-genomic era." Nat Biotech **30**(7): 679-692.

Amoutzias, G. D., Y. He, et al. (2010). "Posttranslational regulation impacts the fate of duplicated genes." Proceedings of the National Academy of Sciences, U.S.A. **107**(7): 2967-2971.

Anderson, J. B., J. Funt, et al. (2010). "Determinants of Divergent Adaptation and Dobzhansky-Muller Interaction in Experimental Yeast Populations." Current Biology **20**(15): 1383-1388.

Argout, X., J. Salse, et al. (2011). "The genome of *Theobroma cacao*." Nat Genet **43**(2): 101-108.

Ashrafian, H. (2006). "Cancer's sweet tooth: the Janus effect of glucose metabolism in tumorigenesis." The Lancet **367**(9510): 618-621.

Aury, J.-M., O. Jaillon, et al. (2006). "Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*." Nature **444**(7116): 171-178.

Bachtrog, D. (2008). "Similar rates of protein adaptation in *Drosophila miranda* and *D. melanogaster*, two species with different current effective population sizes." BMC Evol Biol **8**: 334.

Barker, M. S., N. C. Kane, et al. (2008). "Multiple paleopolyploidizations during the evolution of the Compositae reveal parallel patterns of duplicate gene retention after millions of years." Mol Biol Evol **25**(11): 2445-2455.

Barker, M. S., H. Vogel, et al. (2009). "Paleopolyploidy in the Brassicales: analyses of the Cleome transcriptome elucidate the history of henome duplications in Arabidopsis and other Brassicales." Genome Biol Evol **1**: 391-399.

Bekaert, M. and G. C. Conant (2011). "Copy number alterations among mammalian enyzmes cluster in the metabolic network." Mol Biol Evol **28**: 1111-1121.

Bekaert, M., P. P. Edger, et al. (2011). "Two-phase resolution of polyploidy in the Arabidopsis metabolic network gives rise to relative followed by absolute dosage constraints." Plant Cell **23**: 1-10.

Bertioli, D. J., M. C. Moretzsohn, et al. (2009). "An analysis of synteny of *Arachis* with *Lotus* and *Medicago* sheds new light on the structure, stability and evolution of legume genomes." BMC Genomics **10**.

Bignell, G. R., C. D. Greenman, et al. (2010). "Signatures of mutation and selection in the cancer genome." Nature **463**(7283): 893-898.

Birchler, J. A. and R. A. Veitia (2007). "The gene balance hypothesis: from classical genetics to modern genomics." Plant Cell **19**(2): 395-402.

Blanc, G. and K. H. Wolfe (2004). "Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution." Plant Cell **16**(7): 1679-1691.

Blanc, G. and K. H. Wolfe (2004). "Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes." Plant Cell **16**(7): 1667-1678.
Blank, L. M., F. Lehmbeck, et al. (2005). "Metabolic-flux and network analysis of fourteen hemiascomycetous yeasts." FEMS Yeast Research **5**: 545-558.
Bloom, J. D. and A. C (2003). "Apparent dependence of protein evolutionary rate on number of interactions is linked to biases in protein-protein interactions data sets." BMC Evol Biol **3**: 21.
Boles, E., F. Schulte, et al. (1997). "Characterization of a glucose-repressed pyruvate kinase (Pyk2p) in Saccharomyces cerevisiae that is catalytically insensitive to fructose-1, 6-bisphosphate." Journal of bacteriology **179**(9): 2987-2993.
Bowers, J., B. Chapman, et al. (2003). "Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events." Nature **422**: 433 - 438.
Brandes, U. (2001). "A faster algorithm for betweenness centrality." J of Math Sociol **25**(2): 163-177.
Brandes, U., D. Delling, et al. (2006). "Maximizing modularity is hard." arXiv preprint physics/0608255.
Brown, C. J., K. M. Todd, et al. (1998). "Multiple duplications of yeast hexose-transport genes in response to selection in a glucose-limited environment." Mol Biol Evol **15**(#8): 931-942.
Byrne, K. P. and K. H. Wolfe (2007). "Consistent patterns of rate asymmetry and gene loss indicate widespread neofunctionalization of yeast genes after whole-genome duplication." Genetics **175**(3): 1341-1350.
Byrnes, J. K., G. P. Morris, et al. (2006). "Reorganization of adjacent gene relationships in yeast genomes by whole-genome duplication and gene deletion." Mol Biol Evol **23**(6): 1136-1143.
Campisi, J. and F. D. A. Di Fagagna (2007). "Cellular senescence: when bad things happen to good cells." Nature Reviews Molecular Cell Biology **8**(9): 729-740.
Cannon, S., A. Mitra, et al. (2004). "The roles of segmental and tandem gene duplication in the evolution of large gene families in *Arabidopsis thaliana*." BMC Plant Biol **4**(1): 10.
Casaregola, S., H. V. Nguyen, et al. (2001). "Analysis of the constitution of the beer yeast genome by PCR, sequencing and subtelomeric sequence hybridization." International Journal of Systematic and Evolutionary Microbiology **51**(4): 1607-1618.
Charlesworth, J. and A. Eyre-Walker (2007). "The other side of the nearly neutral theory, evidence of slightly advantageous back-mutations." Proc Natl Acad Sci U S A **104**: 16992-16997.
Chen, K., D. Durand, et al. (2000). "NOTUNG: a program for dating gene duplications and optimizing gene family trees." Journal of Computational Biology **7**(3-4): 429-447.
Chou, J.-Y., Y.-S. Hung, et al. (2010). "Multiple Molecular Mechanisms Cause Reproductive Isolation between Three Yeast Species." PLoS Biol **8**(7): e1000432.
Chung, F. and L. Lu (2002). "Connected components in random graphs with given expected degree sequences." Annals of combinatorics **6**(2): 125-145.

Clauset, A., M. E. J. Newman, et al. (2004). "Finding community structure in very large networks." Physical review E **70**(6): 066111.
Clauset, A., M. E. J. Newman, et al. (2004). "Finding community structure in very large networks." Phys Rev E **70**(6): 066111.
Coate, J., J. Schlueter, et al. (2011). "Comparative evolution of photosynthetic genes in response to polyploid and nonpolyploid duplication." Plant Physiol **155**(2081-2095).
Coissac, E., E. Maillier, et al. (1997). "A comparative study of duplications in bacteria and eukaryotes: the importance of telomeres." Molecular Biology and Evolution **14**(10): 1062-1074.
Conant, G. (2009). "Neutral evolution on mammalian protein surfaces." Trends Gen **25**: 377-381.
Conant, G. C. (2010). "Rapid reorganization of the transcriptional regulatory network after genome duplication in yeast." Proceedings of the Royal Society B **277**: 869-876.
Conant, G. C. and A. Wagner (2002). "GenomeHistory: A software tool and its application to fully sequenced genomes." Nucleic Acids Res **30**: 3378-3386.
Conant, G. C., A. Wagner, et al. (2007). "Modeling amino acid substitution patterns in orthologous and paralogous genes." Mol Phylogenet Evol **42**(2): 298-307.
Conant, G. C. and K. H. Wolfe (2006). "Functional partitioning of yeast co-expression networks after genome duplication." PLoS Biol **4**: e109.
Conant, G. C. and K. H. Wolfe (2007). "Increased glycolytic flux as an outcome of whole-genome duplication in yeast." Mol Biol Evol **3**: 129.
Conant, G. C. and K. H. Wolfe (2008). "Probabilistic cross-species inference of orthologous genomic regions created by whole-genome duplication in yeast." Genetics **179**: 1681-1692.
Conant, G. C. and K. H. Wolfe (2008). "Turning a hobby into a job: How duplicated genes find new functions." Nat Rev Genet **9**(12): 938-950.
Cooper, G. M., M. Brudno, et al. (2004). "Characterization of evolutionary rates and constraints in three mammalian genomes." Genome Res **14**(4): 539-548.
Csardi, G. and T. Nepusz (2006). "The igraph software package for complex network research." InterJournal, Complex Systems **1695**: 38.
Csárdi, G. and T. Nepusz (2006). "The igraph software package for complex network research." InterJournal, Complex Systems: 1695.
Cui, L., P. K. Wall, et al. (2006). "Widespread genome duplications throughout the history of flowering plants." Genome Res **16**(6): 738-749.
De Bodt, S., S. Maere, et al. (2005). "Genome duplication and the origin of angiosperms." Trends Ecol Evol **20**(11): 591-597.
de Oliveira Dal'Molin, C. G., L.-E. Quek, et al. (2010). "AraGEM, a Genome-Scale Reconstruction of the Primary Metabolic Network in Arabidopsis." Plant Phys **152**(2): 579-589.
De, S., N. Lopez-Bigas, et al. (2008). "Patterns of evolutionary constraints on genes in humans." BMC Evol Biol **8**: 275.
DeLuna, A., K. Vetsigian, et al. (2008). "Exposing the fitness contribution of duplicated genes." Nature Genetics **40**(5): 676-681.

Dequin, S. and S. Casaregola (2011). "The genomes of fermentative *Saccharomyces*." Comptes Rendus Biologies **334**(8-9): 687-693.
Des Marais, D. L. and M. D. Rausher (2008). "Escape from adaptive conflict after duplication in an anthocyanin pathway gene." Nature **454**(7205): 762-765.
Dettman, J. R., C. Sirjusingh, et al. (2007). "Incipient speciation by divergent adaptation and antagonistic epistasis in yeast." Nature **447**(7144): 585-588.
Dietrich, F. S., S. Voegeli, et al. (2004). "The Ashbya gossypii Genome as a Tool for Mapping the Ancient Saccharomyces cerevisiae Genome." Science **304**(5668): 304-307.
Dowling, R. J., I. Topisirovic, et al. (2010). "Dissecting the role of mTOR: lessons from mTOR inhibitors." Biochimica et biophysica acta **1804**(3): 433.
Drummond, D., A. Raval, et al. (2006). "A single determinant dominates the rate of yeast protein evolution." Mol Biol Evol **23**: 327-337.
Drummond, D. A., J. D. Bloom, et al. (2005). "Why highly expressed proteins evolve slowly." Proc Natl Acad Sci U S A **102**(40): 14338-14343.
Drummond, D. A., A. Raval, et al. (2006). "A single determinant dominates the rate of yeast protein evolution." Mol Biol Evol **23**(2): 327-337.
Duarte, J., P. K. Wall, et al. (2010). "Identification of shared single copy nuclear genes in Arabidopsis, Populus, Vitis and Oryza and their phylogenetic utility across various taxonomic levels." BMC Evol Biol **10**(1): 61.
Duarte, N. C., S. A. Becker, et al. (2007). "Global reconstruction of the human metabolic network based on genomic and bibliomic data." Proc Natl Acad Sci U S A **104**(6): 1777-1782.
Duarte, N. C., S. A. Becker, et al. (2007). "Global reconstruction of the human metabolic network based on genomic and bibliomic data." Proceedings of the National Academy of Sciences **104**(6): 1777-1782.
Dujon, B., D. Sherman, et al. (2004). "Genome evolution in yeasts." Nature **430**(6995): 35-44.
Duret, L. and D. Mouchiroud (2000). "Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate." Mol Biol Evol **17**: 68-85.
Duret, L. and D. Mouchiroud (2000). "Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate." Mol Biol Evol **17**: 68-85.
Edgar, R. (2004). "MUSCLE: multiple sequence alignment with high accuracy and high throughput." Nucleic Acids Res **32**(5): 1792-1797.
Edgar, R. C. (2004). "MUSCLE: multiple sequence alignment with high accuracy and high throughput." Nucleic Acids Res **32**(5): 1792-1797.
Edger, P. and J. C. Pires (2009). "Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes." Chromosome Res **17**: 699-717.
Eisen, M. B., P. T. Spellman, et al. (1998). "Cluster analysis and display of genome-wide expression patterns." Proceedings of the National Academy of Sciences **95**(25): 14863-14868.
Ellegren, H. (2009). "A selection model of molecular evolution incorporating the effective population size." Evolution **63**: 301-305.

Evangelisti, A. M. and G. C. Conant (2010). "Nonrandom Survival of Gene Conversions among Yeast Ribosomal Proteins Duplicated through Genome Doubling." Genome Biology and Evolution **2**: 826-834.
Fares, M. A., K. P. Byrne, et al. (2006). "Rate asymmetry after genome duplication causes substantial long-branch attraction artifacts in the phylogeny of *Saccharomyces* species. ." Molecular Biology and Evolution **23**: 245-253.
Fay, J. C., G. J. Wyckoff, et al. (2002). "Testing the neutral theory of molecular evolution with genomic data from Drosophila." Nature **415**: 1024-1026.
Fitch, W. M. and E. Margoliash (1967). "Construction of phylogenetic trees." Science **155**: 279-284.
Fitzpatrick, D., M. Logue, et al. (2006). "A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis." BMC Evolutionary Biology **6**: 99.
Force, A., M. Lynch, et al. (1999). "Preservation of duplicate genes by complementary, degenerative mutations." Genetics **151**(4): 1531-1545.
Fraser, H. B., A. E. Hirsh, et al. (2002). "Evolutionary rate in the protein interaction network." Science **296**: 750-752.
Fraser, P. and W. Bickmore (2007). "Nuclear organization of the genome and the potential for gene regulation." Nature **447**: 413-417.
Freedman, M. L., A. N. A. Monteiro, et al. (2011). "Principles for the post-GWAS functional characterization of cancer risk loci." Nature genetics **43**(6): 513-518.
Freeling, M. (2008). "The evolutionary position of subfunctionalization, downgraded." Genome Dyn **4**: 25-40.
Freeling, M. (2009). "Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition." Ann Rev Plant Biol **60**: 433-453.
Freeling, M. and B. C. Thomas (2006). "Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity." Genome Res **16**: 805-814.
Freeman, L. C. (1977). "A set of measures of centrality based on betweenness." Sociometry **40**: 35-41.
Freeman, L. C. (1979). "Centrality in social networks I: conceptual clarificaiton." Social Networks **1**: 215-239.
Friedman, R. and A. L. Hughes (2001). "Gene Duplication and the Structure of Eukaryotic Genomes." Genome Research **11**(3): 373-381.
Furlong, R. F. and P. W. H. Holland (2002). "Were vertebrates octoploid?" Philosophical Transactions of the Royal Society of London B **357**: 531-544.
Fusco, D., L. Grassi, et al. (2010). "Ordered structure of the transcription network inherited from the yeast whole-genome duplication." BMC Systems Biology **4**: 77.
Gao, L.-z. and H. Innan (2004). "Very low gene duplicaiton rate in the yeast genome." Science **306**: 1367-1370.
Garraway, L. A. and E. S. Lander (2013). "Lessons from the cancer genome." Cell **153**(1): 17-37.

Geladé, R., S. Van de Velde, et al. (2003). "Multi-level response of the yeast genome to glucose." Genome Biology **4**: 233.
Goffeau, A., B. Barrell, et al. (1996). "Life with 6000 genes." Science **274**: 562-567.
Goldman, N. and Z. Yang (1994). "A codon-based model of nucleotide substitution for protein-coding DNA sequences." Mol Biol Evol **11**(5): 725-736.
Gordon, J. L., K. P. Byrne, et al. (2011). "Mechanisms of chromosome number evolution in yeast." PLoS Genet **7**: e1002190.
Gordon, J. L. and K. H. Wolfe (2008). "Recent allopolyploid origin of Zygosaccharomyces rouxii strain ATCC 42981." Yeast **25**(6): 449-456.
Gout, J.-F., L. Duret, et al. (2009). "Differential retention of metabolic genes following whole-genome duplication." Mol Biol Evol **26**: 1067-1072.
Greenberg, A. J., S. R. Stockwell, et al. (2008). "Evolutionary constraint and adaptation in the metabolic network of Drosophila." Mol Biol Evol **25**: 2537-2546.
Greig, D. (2008). "Reproductive isolation in Saccharomyces." Heredity **102**(1): 39-44.
Grosberg, R. K. and R. R. Strathmann (2007). "The evolution of multicellularity: a minor major transition?" Annu. Rev. Ecol. Evol. Syst. **38**: 621-654.
Gu, Z., A. Cavalcanti, et al. (2002). "Extent of gene duplication in the genomes of *Drosophila*, nematode, and yeast." Mol Biol Evol **19**(3): 256-262.
Gu, Z., L. M. Steinmetz, et al. (2003). "Role of duplicate genes in genetic robustness against null mutations." Nature **421**: 63-66.
Guan, Y., M. J. Dunham, et al. (2007). "Functional analysis of gene duplications in *Saccharomyces cerevisiae*." Genetics **175**(2): 933-943.
Guelzim, N., S. Bottani, et al. (2002). "Topological and causal structure of the yeast transcriptional regulatory network." Nature genetics **31**(1): 60-63.
Guenther, M., G. Frampton, et al. (2010). "Chromatin structure and gene expression programs of human embryonic and induced pluripotent stem cells. ." Cell **7**(2): 249-257.
Guo, B., Y. Jin, et al. (2008). "Functional analysis of the Arabidopsis PHT4 family of intracellular phosphate transporters." New Phyto **177**(4): 889-898.
Ha, M., E.-D. Kim, et al. (2009). "Duplicate genes increase expression diversity in closely related species and allopolyploids." Proc Natl Acad Sci U S A **106**(7): 2295-2300.
Hahn, M. W. (2009). "Distinguishing amoung evolutionary models for the maintenance of gene duplicates." J Heredit **100**(5): 605-617.
Hahn, M. W., G. C. Conant, et al. (2004). "Molecular evolution in large genetic networks: Connectivity does not equal constraint." J Mol Evol **58**(2): 203-211.
Hakes, L., J. Pinney, et al. (2007). "All duplicates are not equal: the difference between small-scale and genome duplication." Genome Biol **8**(10): R209.
Hamel, P., Y. Saint-Georges, et al. (2004). "Redundancy in the function of mitochondrial phosphate transport in *Saccharomyces cerevisiae* and *Arabidopsis thaliana*." Mol Microbiol **51**(307-317).
Haverty, P., Z. Weng, et al. (2002). "HugeIndex: a database with visualization tools for high-density oligonucleotide array data from normal human tissues." Nucleic Acids Res **30**(1): 214-217.

Hirsh, A. E. and H. B. Fraser (2001). "Protein dispensability and rate of evolution." Nature **411**: 1046-1049.
Huang, D., B. Sherman, et al. (2009). "Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources." Nature Protocols **41**(1): 44-57.
Hudson, C. and G. C. Conant (2011). "Expression level, cellular compartment and metabolic network position all influence the average selective constraint on mammalian enzymes." BMC Evol Biol **11**: 89.
Hughes, A. (1994). "The evolution of functionally novel proteins after gene duplication." Proc Royal Soc B **256**: 119-124.
Hughes, A. L. (1999). "Phylogenies of developmentally important proteins do not support the hypothesis of two rounds of genome duplication early in vertebrate history." Journal of Molecular Evolution **48**: 565-576.
Hughes, M. and A. Hughes (1993). "Evolution of duplicate genes in a tetraploid animal, *Xenopus laevis*." Mol Biol Evol **10**: 1360-1369.
Hughes, T. R., C. J. Roberts, et al. (2000). "Widespread aneuploidy revealed by DNA microarray expression profiling." Nature Genetics **25**: 333-337.
Hurst, L. D. and N. G. Smith (1999). "Do essential genes evolve slowly?" Curr Biol **9**: 474-450.
Huss, M. and P. Holme (2007). "Currency and commodity metabolites: their identification and relation to the modularity of metabolic networks." IET Systems Biology **1**: 280-285.
Ihmels, J., S. Bergmann, et al. (2005). "Rewiring of the yeast transcriptional network through the evolution of motif usage." Science **309**(5736): 938-940.
Innan, H. and F. Kondrashov (2010). "The evolution of gene duplications: classifying and distinguishing between models." Nat Rev Genet **11**(2): 97-108.
Irizarry, R. A., B. M. Bolstad, et al. (2003). "Summaries of Affymetrix GeneChip probe level data." Nucleic Acids Research **31**(4): e15.
Irizarry, R. A., B. Hobbs, et al. (2003). "Exploration, normalization, and summaries of high density oligonucleotide array probe level data." Biostatistics **4**(2): 249-264.
Jaillon, O., J.-M. Aury, et al. (2004). "Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype." Nature **431**(7011): 946-957.
Jaillon, O., J. M. Aury, et al. (2007). "The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla." Nature **449**(7161): 463-U465.
Jain, M., R. Nilsson, et al. (2012). "Metabolite Profiling Identifies a Key Role for Glycine in Rapid Cancer Cell Proliferation." Science **336**(6084): 1040-1044.
James, S. A., C. J. Bond, et al. (2005). "Molecular evidence for the existence of natural hybrids in the genus *Zygosaccharomyces*." FEMS Yeast Research **5**(8): 747-755.
Jiao, Y., N. J. Wickett, et al. (2011). "Ancestral polyploidy in seed plants and angiosperms." Nature **473**: 97-100.
Johnston, M. and J.-H. Kim (2005). "Glucose as a hormone: Receptor-mediated glucose sensing in the yeast *Saccharomyces cerevisiae*." Biochemical Society Transactions **33**: 247-252.
Jordan, I., Y. Wolf, et al. (2004). "Duplicated genes evolve slower than singletons despite the initial rate increase." BMC Evol Biol **4**: 22.

Jordan, I. K., L. Marino-Ramirez, et al. (2004). "Conservation and coevolution in the scale-free human gene coexpression network." Mol Biol Evol **21**(11): 2058-2070.
Jordan, I. K., I. B. Rogozin, et al. (2002). "Essential genes are more evolutionarily conserved than are nonessential genes in bacteria." Genome Res **12**: 962-968.
Jordan, I. K., Y. I. Wolf, et al. (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. and P. P. C (2009). "Evolutionary rates and centrality in the yeast gene regulatory network." Genome Biol **10**: R35.
Joyce, A. R. and B. Ø. Palsson (2006). "The model organism as a system: integrating'omics' data sets." Nature Reviews Molecular Cell Biology **7**(3): 198-210.
Julenius, K. and A. G. Pedersen (2006). "Protein evolution is faster outside the cell." Mol Biol Evol **22**: 2039-2048.
Kacser, H. and J. A. Burns (1981). "The molecular basis of dominance." Genetics **97**(3-4): 639-666.
Kao, K. C., K. Schwartz, et al. (2010). "A Genome-Wide Analysis Reveals No Nuclear Dobzhansky-Muller Pairs of Determinants of Speciation between S. cerevisiae and S. paradoxus, but Suggests More Complex Incompatibilities." PLoS Genet **6**(7): e1001038.
Kellis, M., B. W. Birren, et al. (2004). "Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae." Nature **428**(6983): 617-624.
Kielland-Brandt, M. C., T. Nilsson-Tillgren, et al. (1995). Genetics of brewing yeasts. The Yeasts. 2nd Ed. Vol. 6. A. H. Rose, A. E. Wheals and J. S. Harrison. London, Academic Press**:** 223-254.
Kim, S.-H. and S. V. Yi (2006). "Correlated asymmetry of sequence and functional divergence between duplicate proteins of *Saccharomyces cerevisiae.* ." Molecular Biology and Evolution **23**: 1068-1075.
Kimura, M. (1968). "Evolutionary rate at the molecular level." Nature **217**: 624-626.
Kimura, M. (1969). "The rate of molecular evolution considered from the standpoint of population genetics." Proc Natl Acad Sci U S A **63**: 1181-1188.
Kimura, M. (1983). The neutral theory of molecular evolution. Cambridge, Cambridge University Press.
Komili, S., N. G. Farny, et al. (2007). "Functional specificity among ribosomal proteins regulates gene expression." Cell **131**(3): 557-571.
Kondrashov, F., I. Rogozin, et al. (2002). "Selection on the evolution of gene duplicates." Genome Biol **3**.
Kondrashov, F. A. and A. S. Kondrashov (2006). "Role of selection in fixation of gene duplications." J Theor Biol **239**(2): 141-151.
Kondrashov, F. A. and E. V. Koonin (2004). "A common framework for understanding the origin of genetic dominance and evolutionary fates of gene duplications." Trends Genet **20**(7): 287-290.
Kondrashov, F. A., I. B. Rogozin, et al. (2002). "Selection in the evolution of gene duplications." Genome Biol **3**: research0008.

Koonin, E. (2005). "Orthologs, paralogs, and evolutionary genomics." Annual Review of Genetics **39**: 309-338.
Koszul, R., S. Caburet, et al. (2004). "Eucaryotic genome evolution through the spontaneous duplication of large chromosomal segments." EMBO J **23**(1): 234-243.
Krakauer, D. C. and J. B. Plotkin (2002). "Redundancy, antiredundancy, and the robustness of genomes." Proceedings of the National Academy of Sciences **99**(3): 1405-1409.
Krisher, R. L. and R. S. Prather (2012). "A role for the Warburg effect in preimplantation embryo development: Metabolic modification to support rapid cell proliferation." Molecular reproduction and development **79**(5): 311-320.
Kuepfer, L., U. Sauer, et al. (2005). "Metabolic functions of duplicate genes in Saccharomyces cerevisiae." Genome Research **15**(10): 1421-1430.
Kurtzman, C. and C. Robnett (2003). "Phylogenetic relationships among yeasts of the '*Saccharomyces* complex' determined from multigene sequence analyses." FEMS Yeast Research **3**: 417-432.
Lalo, D., S. Stettler, et al. (1993). "Two yeast chromosomes are related by a fossil duplication of their centromeric regions." Comptes rendus de l'Académie des sciences **316**: 367-373.
Lancichinetti, A. and S. Fortunato (2009). "Community detection algorithms: A comparative analysis." Physical review E **80**(5): 056117.
Lane, N. and W. Martin (2010). "The energetics of genome complexity." Nature **467**(7318): 929-934.
Lee, H.-Y., J.-Y. Chou, et al. (2008). "Incompatibility of Nuclear and Mitochondrial Genomes Causes Hybrid Sterility between Two Yeast Species." Cell **135**(6): 1065-1073.
Liang, H., K. R. Plazonic, et al. (2008). "Protein under-wrapping causes dosage sensitivity and decreases gene duplicability." PLoS Genet **4**(1): e11.
Liao, B., M. Weng, et al. (2010). "Impact of extracellularity on the evolutionary rate of mammalian proteins." Genome Biol Evol **2**: 39-43.
Libkind, D., C. T. Hittinger, et al. (2011). "Microbe domestication and the identification of the wild genetic stock of lager-brewing yeast." Proceedings of the National Academy of Sciences **108**(35): 14539-14544.
Lin, Z. and W. H. Li (2010). "Expansion of hexose transporter genes was associated with the evolution of aerobic fermentation in yeasts." Mol Biol Evol **28**(1): 131-142.
Liu, W.-c., W.-h. Lin, et al. (2007). "A network perspective on the topological importance of enzymes and their phylogenetic conservation." BMC Bioinformatics **8**: 121.
Llorente, B., P. Durrens, et al. (2000). "Genomic Exploration of the Hemiascomycetous Yeasts: 20. Evolution of gene redundancy compared to Saccharomyces cerevisiae." FEBS Letters **487**(1): 122-133.
Llorente, B., A. Malpertuy, et al. (2000). "Genomic Exploration of the Hemiascomycetous Yeasts: 18. Comparative analysis of chromosome maps and synteny with Saccharomyces cerevisiae." FEBS Letters **487**(1): 101-112.

Lopez-Bigas, N., S. De, et al. (2008). "Functional protein divergence in the evolution of *Homo sapiens*." Genome Biol **9**: R33.
Lynch, M. (2007). "The evolution of genetic networks by non-adaptive processes." Nat Rev Genet **8**(10): 803-813.
Lynch, M. and J. S. Conery (2000). "The evolutionary fate and consequences of duplicate genes." Science **290**: 1151-1154.
Lynch, M. and J. S. Conery (2003). "The evolutionary demography of duplicate genes." J Struct Funct Genomics **3**: 35 - 44.
Lynch, M. and J. S. Conery (2003). "The origins of genome complexity." Science **302**: 1401-1404.
Lynch, M. and A. G. Force (2000). "The Origin of Interspecific Genomic Incompatibility via Gene Duplication." The American Naturalist **156**(6): 590-605.
Lyons, E., B. Pedersen, et al. (2008). "Finding and comparing syntenic regions among Arabidopsis and the outgroups papaya, poplar, and grape: CoGe with Rosids. ." Plant Physiol **148**(4): 1772-1781.
Lyons, E., B. Pedersen, et al. (2008). "The value of nonmodel genomes and an example using SynMap within CoGe to dissect the hexaploidy that predates the rosids." Trop Plant Biol **1**(3): 181-190.
Maclean, C. J. and D. Greig (2011). "RECIPROCAL GENE LOSS FOLLOWING EXPERIMENTAL WHOLE-GENOME DUPLICATION CAUSES REPRODUCTIVE ISOLATION IN YEAST." Evolution **65**(4): 932-945.
MacLean, R. C. and I. Gudelj (2006). "Resource competition and social conflict in experimental populations of yeast." Nature **441**: 498-501.
Maere, S., S. De Bodt, et al. (2005). "Modeling gene and genome duplications in eukaryotes." Proceedings of the National Academy of Sciences of the United States of America **102**(15): 5454-5459.
Maere, S., S. De Bodt, et al. (2005). "Modeling gene and genome duplications in eukaryotes." Proc Natl Acad Sci U S A **102**: 5454-5459.
Malhotra, A., M. R. Lindberg, et al. (2013). "Breakpoint profiling of 64 cancer genomes reveals numerous complex rearrangements spawned by homology-independent mechanisms." Genome research.
Martini, A. V. and C. P. Kurtzman (1985). "Deoxyribonucleic Acid Relatedness among Species of the Genus Saccharomyces Sensu Stricto." International Journal of Systematic Bacteriology **35**(4): 508-511.
Mayfield-Jones, D., J. D. Washburn, et al. Watching the grin fade: Tracing the effects of polyploidy on different evolutionary time scales, Elsevier.
Melnick, L. and F. Sherman (1993). "The gene clusters ARC and COR on chromosomes 5 and 10, respectively, of *Saccharomyces cerevisiae* share a common ancestry." Journal of Molecular Biology **233**(3): 372-388.
Merico, A., P. Sulo, et al. (2007). "Fermentative lifestyle in yeasts belonging to the *Saccharomyces* complex." FEBS J **274**: 976-989.
Merlo, L. M. F., J. W. Pepper, et al. (2006). "Cancer as an evolutionary and ecological process." Nature Reviews Cancer **6**(12): 924-935.
Mewes, H., K. Albermann, et al. (1997). "Overview of the yeast genome." Nature **387**: 7-65.

Ming, R., S. Hou, et al. (2008). "The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus)." Nature **452**(7190): 991-996.
Moore, M., C. Bell, et al. (2007). "Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms." Proc Natl Acad Sci U S A **104**(49): 19363-19368.
Mudge, S. R., A. L. Rae, et al. (2002). "Expression analysis suggests novel roles for members of the Pht1 family of phosphate transporters in Arabidopsis." Plant J **31**(3): 341-353.
Murphy, W. J., P. A. Pevzner, et al. (2004). "Mammalian phylogenomics comes of age." Trends Gen **20**(12): 631-639.
Muse, S. V. and B. S. Gaut (1994). "A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome." Mol Biol Evol **11**(5): 715-724.
Nakao, Y., T. Kanamori, et al. (2009). "Genome Sequence of the Lager Brewing Yeast, an Interspecies Hybrid." DNA Research **16**(2): 115-129.
Nei, M. (2005). "Selectionism and neutralism in molecular evolution." Mol Biol Evol **22**: 2318-2342.
Newman, M. E. J. (2006). "Modularity and community structure in networks." Proc Natl Acad Sci U S A **103**: 8577-8582.
Newman, M. E. J. (2006). "Modularity and community structure in networks." Proceedings of the National Academy of Sciences **103**(23): 8577-8582.
Newman, M. E. J. and M. Girvan (2004). "Finding and evaluating community structure in networks." Physical Review E **69**: 026113.
Newman, M. E. J. and M. Girvan (2004). "Finding and evaluating community structure in networks." Physical review E **69**(2): 026113.
Ni, L. and M. Snyder (2001). "A Genomic Study of the Bipolar Bud Site Selection Pattern inSaccharomyces cerevisiae." Molecular biology of the cell **12**(7): 2147-2170.
Nishihara, H., M. Hasegawa, et al. (2006). "Pegasoferae, an unexpected mammalian clade revealed by tracking ancient retroposon insertions." Proc Natl Acad Sci U S A **103**(26): 9929-9934.
Ohno, S. (1970). Evolution by gene duplication. New York, Springer-Verlag.
Oliver, S. G. (1996). "From DNA sequence to biological function." Nature **379**: 597-600.
Orth, J., I. Thiele, et al. (2010). "What is flux balance analysis?" Nature Biotechnology **28**: 245-248.
Ouyang, S., W. Zhu, et al. (2006). "The TIGR Rice Genome Annotation Resource: improvements and new features." Nucleic Acids Res **35**(suppl 1): D883-D887.
Özcan, S., J. Dover, et al. (1998). "Glucose sensing and signaling by two glucose receptors in the yeast Saccharomyces cerevisiae." The EMBO journal **17**(9): 2566-2573.
Pál, C., B. Papp, et al. (2003). "Rate of evolution and gene dispensability." Nature **421**: 496-497.
Pál, C., B. Papp, et al. (2003). "Rate of evolution and gene dispensability." Nature **421**: 496-497.

Papp, B., C. Pal, et al. (2003). "Evolution of cis-regulatory elements in duplicated genes of yeast." Trends Genet **19**(8): 417-422.
Papp, B., C. Pál, et al. (2003). "Dosage sensitivity and the evolution of gene families in yeast." Nature **424**(6945): 194-197.
Patel, M. N., M. D. Halling-Brown, et al. (2012). "Objective assessment of cancer genes for drug discovery." Nature Reviews Drug Discovery **12**(1): 35-50.
Paterson, A., J. Bowers, et al. (2009). "The *Sorghum bicolor* genome and the diversification of grasses." Nature **457**: 551-556.
Paterson, A. H., J. E. Bowers, et al. (2004). "Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics." Proc Natl Acad Sci U S A **101**(26): 9903-9908.
Paterson, A. H., B. A. Chapman, et al. (2006). "Many gene and domain families have convergent fates following independent whole-genome duplication events in Arabidopsis, Oryza, Saccharomyces and Tetraodon." Trends Genet **22**(11): 597-602.
Pérez-Bercoff, Å., A. McLysaght, et al. (2011). "Patterns of indirect protein interactions suggest a spatial organization to metabolism." Molecular BioSystems **7**(11): 3056-3064.
Petes, T. and C. Hill (1988). "Recombination between repeated genes in microorganisms." Annual Review of Genetics **22**: 147-168.
Pfeiffer, T. and S. Schuster (2005). "Game-theoretical approaches to studying the evolution of biochemical systems." Trends in Biochemical Sciences **30**(1): 20-25.
Pfeiffer, T., S. Schuster, et al. (2001). "Cooperation and competition in the evolution of ATP-producing pathways." Science **292**: 504-507.
Pfeil, B. E., J. A. Schlueter, et al. (2005). "Placing paleopolyploidy in relation to taxon divergence: a phylogenetic analysis in legumes using 39 gene families." Syst Biol **54**(3): 441-454.
Pigliucci, M. (2010). "Genotype-phenotype mapping and the end of the 'genes as blueprint' metaphor." Philosophical Transactions of the Royal Society of London B **365**: 557-566.
Piskur, J. (2001). "Origin of the duplicated regions in the yeast genomes." Trends in Genetics **17**(6): 302-303.
Piškur, J., E. Rozpędowska, et al. (2006). "How did< i> Saccharomyces</i> evolve to become a good brewer?" Trends in Genetics **22**(4): 183-186.
Poirier, Y. and M. Bucher (2002). Phosphate transport and homeostasis in Arabidopsis. The Arabidopsis book. C. Somerville and E. M. Meyerowitz. Rockville, MD, USA, American Society of Plant Biologists**:** 1-35.
Pons, P. and M. Latapy (2005). Computing communities in large networks using random walks. Computer and Information Sciences-ISCIS 2005, Springer**:** 284-293.
Powell, A., G. C. Conant, et al. (2008). "Altered patterns of gene duplication and differential gene gain and loss in fungal pathogens." BMC Genomics **9**: 147.
Pritchard, L. and D. B. Kell (2002). "Schemes of flux control in a model of *Saccharomyces cerevisiae* glycolysis." Euro J Biochem **269**: 3894-3904.
Raghavan, U. N., R. k. Albert, et al. (2007). "Near linear time algorithm to detect community structures in large-scale networks." Physical review E **76**(3): 036106.

Rainieri, S., Y. Kodama, et al. (2006). "Pure and Mixed Genetic Lines of Saccharomyces bayanus and Saccharomyces pastorianus and Their Contribution to the Lager Brewing Strain Genome." Applied and Environmental Microbiology **72**(6): 3968-3974.

Ramsey, J. (2011). "Polyploidy and ecological adaptation in wild yarrow." Proc Natl Acad Sci U S A **108**(17): 7096-7101.

Redel, B. K., A. N. Brown, et al. (2012). "Glycolysis in preimplantation development is partially controlled by the Warburg Effect." Molecular reproduction and development **79**(4): 262-271.

Rodriguez, M. A., D. Vermaak, et al. (2007). "Species-specific positive selection of the male-specific lethal complex that participates in dosage compensation in Drosophila." Proc Natl Acad Sci U S A **104**(39): 15412-15417.

Rodríguez-Rosales, M. P., F. J. Gálvez, et al. (2008). "Plant NHX cation/proton antiporters." Plant Signal Behav **4**(4): 265-276.

Rolland, T. and B. Dujon (2011). "Yeasty clocks: Dating genomic changes in yeasts." Comptes Rendus Biologies **334**: 620-628.

Saitou, M. and M. Yamaji (2010). "Germ cell specification in mice: signaling, transcription regulation, and epigenetic consequences." Reproduction **139**(6): 931-942.

Scannell, D. R., K. P. Byrne, et al. (2006). "Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts." Nature **440**(7082): 341-345.

Scannell, D. R., O. A. Zill, et al. (2011). "The Awesome Power of Yeast Evolutionary Genetics: New Genome Sequences and Strain Resources for the Saccharomyces sensu stricto Genus." G3: Genes, Genomes, Genetics **1**(1): 11-25.

Schaefer, U., S. Schmeier, et al. (2011). "TcoF-DB: dragon database for human transcription co-factors and transcription factor interacting proteins." Nucleic Acids Research **39**(suppl 1): D106-D110.

Schmutz, J., S. B. Cannon, et al. (2010). "Genome sequence of the palaeopolyploid soybean." Nature **463**(7278): 178-183.

Schnable, J., B. Pedersen, et al. (2011). "Dose–sensitivity, conserved non-coding sequences, and duplicate gene retention through multiple tetraploidies in the grasses." Frontiers Plant Sci **2**: 2.

Schranz, M. E. and T. Mitchell-Olds (2006). "Independent ancient polyploidy events in the sister families Brassicaceae and Cleomaceae." Plant Cell **18**(5): 1152-1165.

Seoighe, C. and K. H. Wolfe (1999). "Yeast genome evolution in the post-genome era." Curr Opin Microbiol **2**(5): 548-554.

Sherman, B. T. and R. A. Lempicki (2009). "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists." Nucleic Acids Research **37**(1): 1-13.

Shulaev, V., D. J. Sargent, et al. (2011). "The genome of woodland strawberry (*Fragaria vesca*)." Nat Genet **43**(2): 109-116.

Simpson, G. G. (1944). Tempo and mode in evolution. New York, Columbia University Press.

Slotte, T., J. P. Foxe, et al. (2010). "Genome-wide evidence for efficient positive and purifying selection in *Capsella grandiflora*, a plant species with a large effective population size." Mol Biol Evol **27**: 1813-1821.
Smith, M. (1987). "Molecular evolution of the *Saccharomyces cerevisiae* histone gene loci." Journal of Molecular Evolution **24**: 252-259.
Sneath, P. H. A. and R. R. Sokal (1973). Numerical Taxonomy. San Francisco, W. H. Freeman & Co.
Sokal, R. and F. J. Rohlf (2000). Biometry, 3rd Edition. New York, W. H. Freeman and Company.
Soltis, D. E., V. A. Albert, et al. (2009). "Polyploidy and angiosperm diversification." Am J Bot **96**(1): 336-348.
Soltis, D. E., S. A. Smith, et al. (2011). "Angiosperm phylogeny: 17 genes, 640 taxa." Am J Bot **98**(4): 704-730.
Stajich, J. E., D. Block, et al. (2002). "The Bioperl Toolkit: Perl modules for the life sciences." Genome Res **12**(10): 1611-1618.
Stamatakis, A., P. Hoover, et al. (2008). "A Fast Bootstrapping Algorithm for the RAxML Web-Servers." Syst Biol **57**(5): 758-771.
Sterck, L., S. Rombauts, et al. (2005). "EST data suggest that poplar is an ancient polyploid." New Phyto **167**(1): 165-170.
Stratton, M. R. (2011). "Exploring the Genomes of Cancer Cells: Progress and Promise." Science **331**(6024): 1553-1558.
Stratton, M. R., P. J. Campbell, et al. (2009). "The cancer genome." Nature **458**(7239): 719-724.
Swarbreck, D., C. Wilks, et al. (2008). "The Arabidopsis Information Resource (TAIR): gene structure and function annotation." Nucleic Acids Res **36**.
Tang, H., J. E. Bowers, et al. (2008). "Synteny and collinearity in plant genomes." Science **320**: 486-488.
Tang, H., X. Wang, et al. (2008). "Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps." Genome Res **18**(12): 1944-1954.
Taylor, J. S. and J. Raes (2004). "Duplication and divergence: The evolution of new genes and old ideas." Ann Rev Genet **38**: 615-643.
Taylor, J. W. and M. L. Berbee (2006). "Dating divergences in the Fungal Tree of Life: review and new analyses." Mycologia **98**(6): 838-849.
The Arabidopsis Genome Initiative (2000). "Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*." Nature **408**(6814): 796-815.
The International Brachypodium Initiative (2010). "Genome sequencing and analysis of the model grass *Brachypodium distachyon*." Nature **463**(7282): 763-768.
The International Peach Genome Initiative (2010). "Peach Genome."
Thomas, B., B. Pedersen, et al. (2006). "Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes." Genome Res **16**: 934-946.
Throude, M., S. Bolot, et al. (2009). "Structure and expression analysis of rice paleo duplications." Nucleic Acids Res **37**(4): 1248-1259.

Town, C., F. Cheung, et al. (2006). "Comparative genomics of *Brassica oleracea* and *Arabidopsis thaliana* reveal gene loss, fragmentation, and dispersal after polyploidy." Plant Cell **18**(6): 1348-1359.
Tuller, T., M. Kupiec, et al. (2009). "Co-evolutionary networks of genes and cellular processes across fungal species." Genome Biol **10**: R48.
Tuskan, G. A., S. DiFazio, et al. (2006). "The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray)." Science **313**(5793): 1596-1604.
Van de Peer, Y., J. Fawcett, et al. (2009). "The flowering world: a tale of duplications." Trends in Plant Science **14**(12): 680-688.
van Hoek, M. J. and P. Hogeweg (2009). "Metabolic adaptation after whole genome duplication." Mol Biol Evol **26**(11): 2441-2453.
van Hoof, A. (2005). "Conserved functions of yeast genes support the duplication, degeneration and complementation model for gene duplication. ." Genetics **171**: 1455-1461.
van Hoof, N. A., V. H. Hassinen, et al. (2001). "Enhanced copper tolerance in Silene vulgaris (Moench) Garcke populations from copper mines is associated with increased transcript levels of a 2b-type metallothionein gene." Plant Physiol **126**(4): 1519-1526.
Venables, W. N. and B. D. Ripley (2002). Modern Applied Statistics with S. New York, Springer.
Versaw, W. K. and M. J. Harrison (2002). "A chloroplast phosphate transporter, PHT2;1, influences allocation of phosphate within the plant and phosphate-starvation responses." Plant Cell **14**: 1751-1766.
Vieta, R. (2005). "Paralogs in polypoids: one for all and all for one?" Plant Cell **17**: 4-11.
Viger, F. and M. Latapy (2005). Efficient and simple generation of random simple connected graphs with prescribed degree sequence. Computing and Combinatorics, Springer**:** 440-449.
Vitkup, D., P. Kharchenko, et al. (2006). "Influence of metabolic network structure and function on enzyme evolution." Genome Biol **7**: R39.
Vitkup, D., P. Kharchenko, et al. (2006). "Influence of metabolic network structure and function on enzyme evolution." Genome Biol **7**: R39.
Vogel, J. P., D. F. Garvin, et al. (2010). "Genome sequencing and analysis of the model grass *Brachypodium distachyon*." Nature **463**(7282): 763-768.
Wagner, A. (2002). "Asymmetric functional divergence of duplicate genes in yeast." Mol Biol Evol **19**(10): 1760-1768.
Wagner, A. (2005). "Energy constraints on the evolution of gene expression." Mol Biol Evol **22**(6): 1365-1374.
Wagner, A. (2009). "Evolutionary constraints permeate large metabolic networks." BMC Evol Biol **9**: 231.
Wang H, Moore MJ, et al. (2009). "Rosid radiation and the rapid rise of angiosperm-dominated forests." Proc Natl Acad Sci U S A **10**(106): 3853-3858.
Wapinski, I., A. Pfeffer, et al. (2007). "Natural history and evolutionary principles of gene duplication in fungi." Nature **449**(7158): 54-61.
Watts, D. J. and S. H. Strogatz (1998). "Collective dynamics of 'small-world' networks." Nature **393**(6684): 440-442.

Wendel, J. F. (2000). "Genome evolution in polyploids." Plant Mol Biol **42**(1): 225-249.
Werth, C. R. and M. D. Windham (1991). "A Model for Divergent, Allopatric Speciation of Polyploid Pteridophytes Resulting from Silencing of Duplicate-Gene Expression." The American Naturalist **137**(4): 515-526.
Widholm, J. M., A. R. Chinnala, et al. (2001). "Glyphosate selection of gene amplification in suspension cultures of 3 plant species." Physiol Plant **112**(4): 540-545.
Wolfe, K. H. (2000). "Robustness-it's not where you think it is." Nature Genetics **25**: 3-4.
Wolfe, K. H. and D. Shields (1997). "Molecular evidence for an ancient duplication of the entire yeast genome." Nature **387**: 708-713.
Wong, W., Z. Yang, et al. (2005). "Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites." Genetics **168**: 1041-1051.
Wood, T., N. Takebayashi, et al. (2009). "The frequency of polyploid speciation in vascular plants." Mol Biol Evol **19**: 1464-1473.
Wright, J. and A. Wagner (2008). "The Systems Biology Research Tool: evolvable open-source software." BMC Syst Biol **2**.
Xu, M. and X. He (2011). "Genetic Incompatibility Dampens Hybrid Fertility More Than Hybrid Viability: Yeast as a Case Study." PLoS ONE **6**(4): e18341.
Yang, Z. (1993). "Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites." Mol Biol Evol **10**: 1396-1401.
Yang, Z. (2007). "PAML 4: Phylogenetic analysis by maximum likelihood." Mol Biol Evol **24**(8): 1586-1591.
Yang, Z., R. Nielsen, et al. (2000). "Codon-substitution models for heterogeneous selection pressure at amino acid sites." Genetics **155**(1): 431-449.
You, J. S. and P. A. Jones (2012). "Cancer Genetics and Epigenetics: Two Sides of the Same Coin?" Cancer Cell **22**(1): 9-20.
Young, N. D., S. B. Cannon, et al. (2005). "Sequencing the genespaces of *Medicago truncatula* and *Lotus japonicus*." Plant Physiol **137**: 1174-1181.
Zhang, J., Z. Gu, et al. (2003). "Different evolutionary patterns between young duplicate genes in the human genome." Genome Biol **4**.
Zhang, J., H. Rosenberg, et al. (1998). "Positive Darwinian selection after gene duplication in primate ribonuclease genes." Proc Natl Acad Sci U S A **95**: 3708-3713.
Zhang, L., T. J. Vision, et al. (2002). "Patterns of nucleotide substitution among simultaneously duplicated gene pairs in *Arabidopsis thaliana*." Mol Biol Evol **19**(9): 1464-1473.
Zhou, J., P. Su, et al. (2009). "mTOR supports long-term self-renewal and suppresses mesoderm and endoderm activities of human embryonic stem cells." Proceedings of the National Academy of Sciences **106**(19): 7840-7845.

FIGURE 1

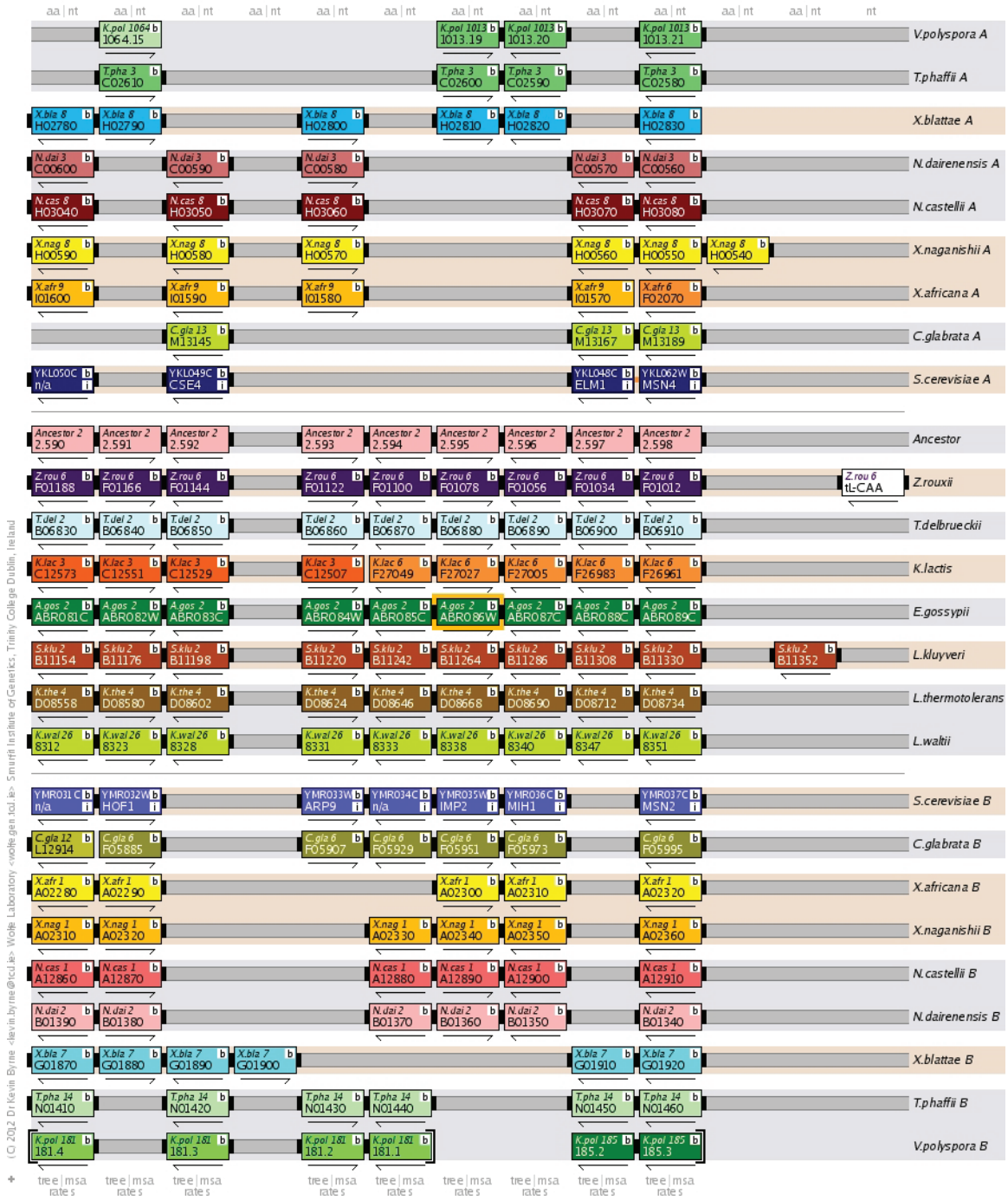

aa | nt
V.polyspora A
T.phaffii A
X.blattae A
N.dairenensis A
N.castellii A
X.naganishii A
X.africana A
C.glabrata A
S.cerevisiae A
Ancestor
Z.rouxii
T.delbrueckii
K.lactis
E.gossypii
L.kluyveri
L.thermotolerans
L.waltii
S.cerevisiae B
C.glabrata B
X.africana B
X.naganishii B
N.castellii B
N.dairenensis B
X.blattae B
T.phaffii B
V.polyspora B
tree | msa
rate s
(C) 2012 Dr Kevin Byrne <kevin.byrne@ucd.ie> Wolfe Laboratory <wolfe.gen.tcd.ie> Smurfit Institute of Genetics, Trinity College Dublin, Ireland

FIGURE 2

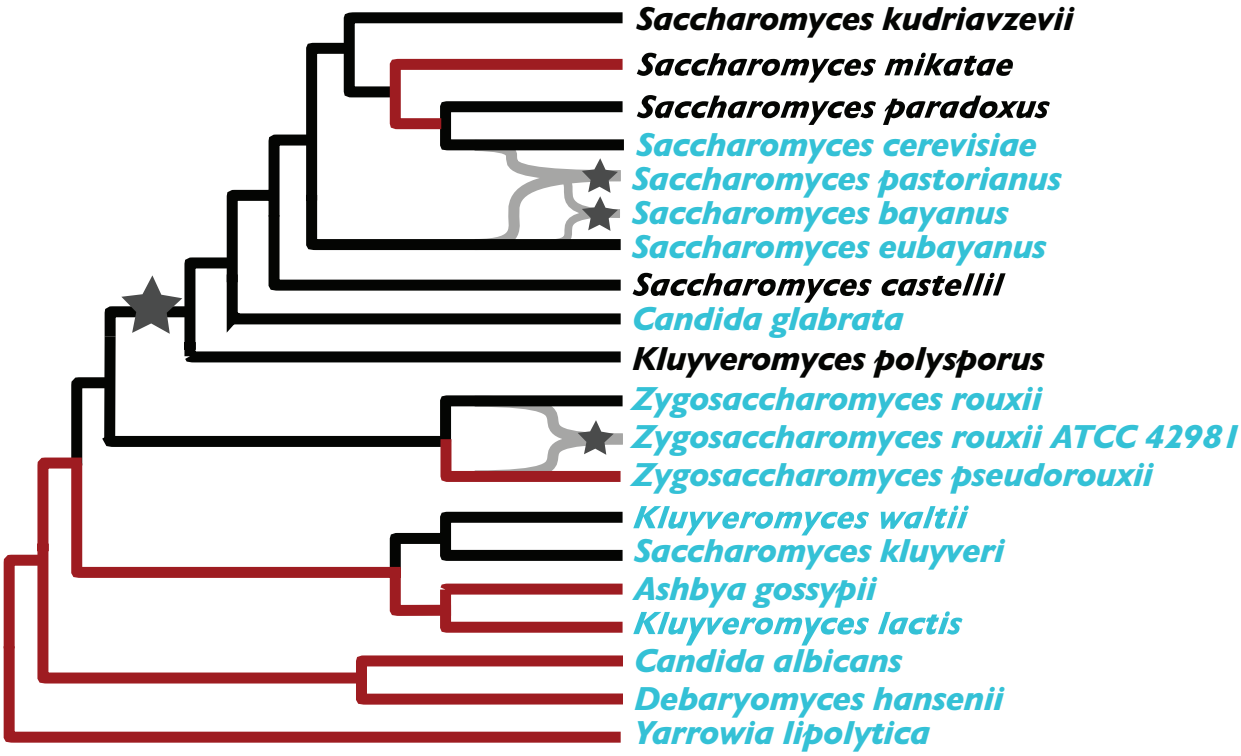

Saccharomyces kudriavzevii
Saccharomyces mikatae
Saccharomyces paradoxus
Saccharomyces cerevisiae
Saccharomyces pastorianus
Saccharomyces bayanus
Saccharomyces eubayanus
Saccharomyces castellii
Candida glabrata
Kluyveromyces polysporus
Zygosaccharomyces rouxii
Zygosaccharomyces rouxii ATCC 42981
Zygosaccharomyces pseudorouxii
Kluyveromyces waltii
Saccharomyces kluyveri
Ashbya gossypii
Kluyveromyces lactis
Candida albicans
Debaryomyces hansenii
Yarrowia lipolytica

FIGURE 3

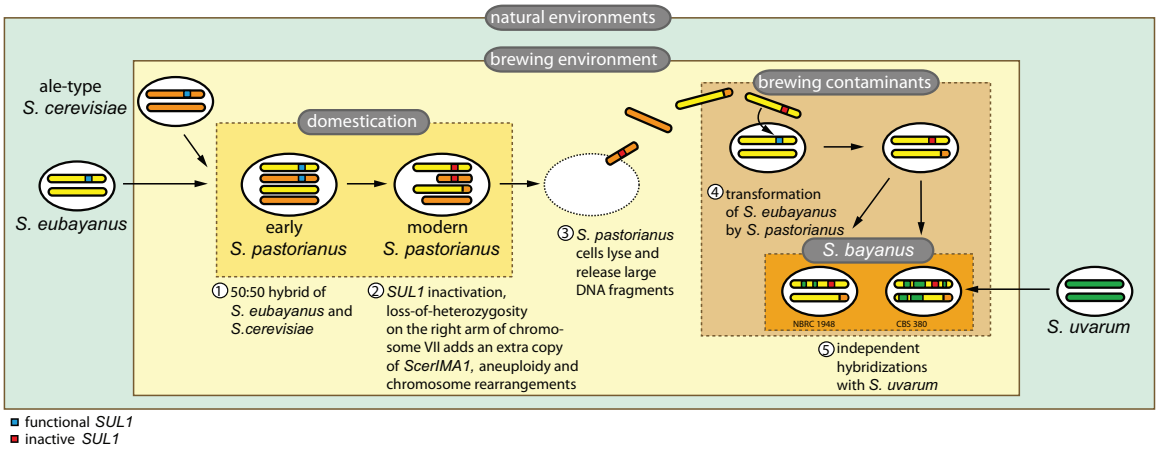

natural environments
brewing environment
ale-type
S. cerevisiae
S. eubayanus
domestication
early
S. pastorianus
modern
S. pastorianus
① 50:50 hybrid of S. eubayanus and S.cerevisiae
② SUL1 inactivation, loss-of-heterozygosity on the right arm of chromosome VII adds an extra copy of ScerIMA1, aneuploidy and chromosome rearrangements
③ S. pastorianus cells lyse and release large DNA fragments
brewing contaminants
④ transformation of S. eubayanus by S. pastorianus
S. bayanus
NBRC 1948
CBS 380
⑤ independent hybridizations with S. uvarum
S. uvarum
functional SUL1
inactive SUL1

FIGURE 4

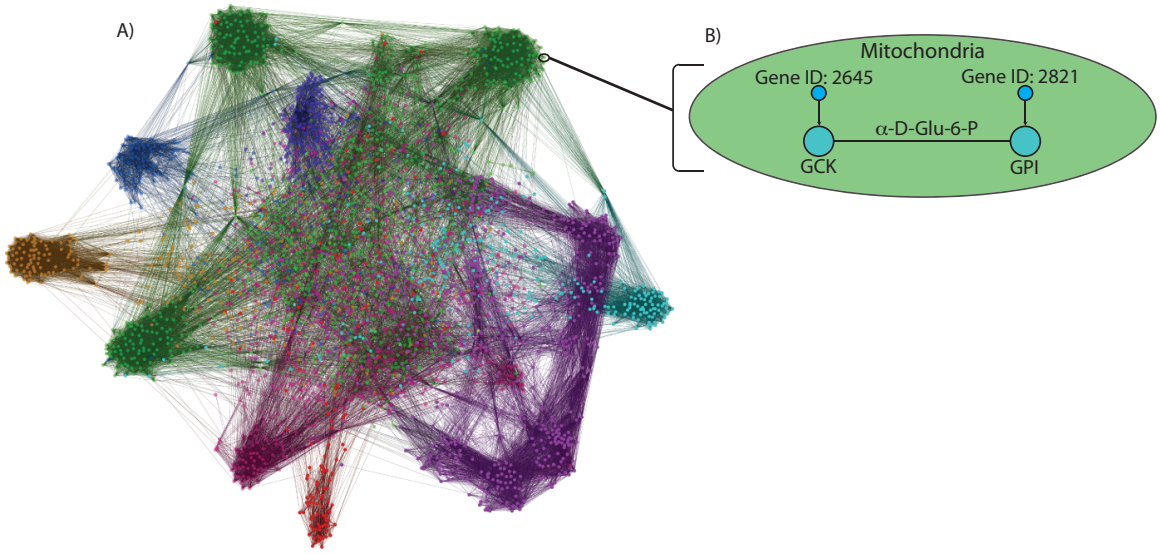


FIGURE 5

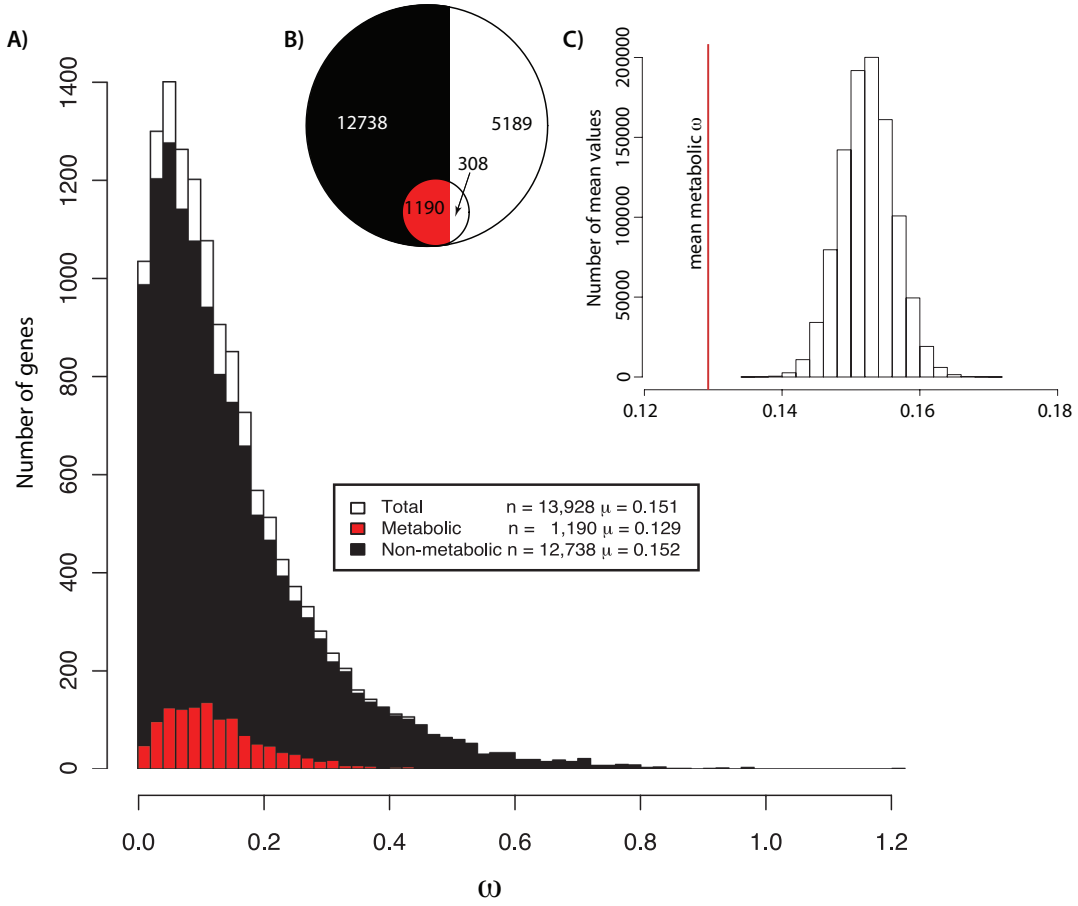
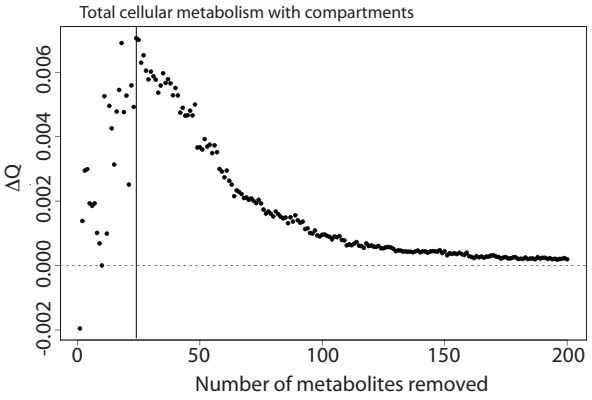
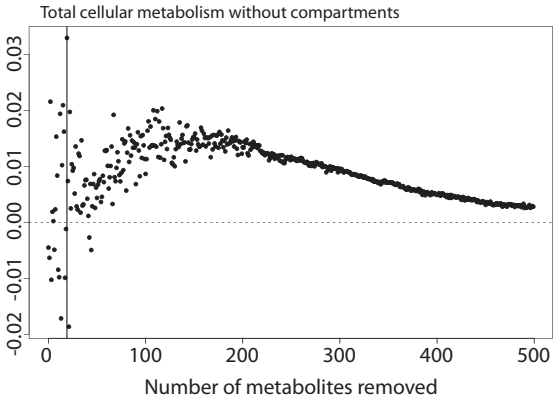

A)
B)
C)
Number of genes
0
200
400
600
800
1000
1200
1400
0.0
0.2
0.4
0.6
0.8
1.0
1.2
ω
12738
5189
308
1190
Total n = 13,928 μ = 0.151
Metabolic n = 1,190 μ = 0.129
Non-metabolic n = 12,738 μ = 0.152
Number of mean values
0
50000
100000
150000
200000
mean metabolic ω
0.12
0.14
0.16
0.18

FIGURE 6

A)



B)



C)

Compartment	Number of metabolites removed to maximize ΔQ
Lysozyme	2
External	5
Golgi apparatus	5
Peroxisome	8
Nucleus	9
Endoplasmic reticulum	20
Mitochondria	21
Cytoplasm	44
Total cellular metabolism with compartments	24
Total cellular metabolism without compartments	20

FIGURE 7

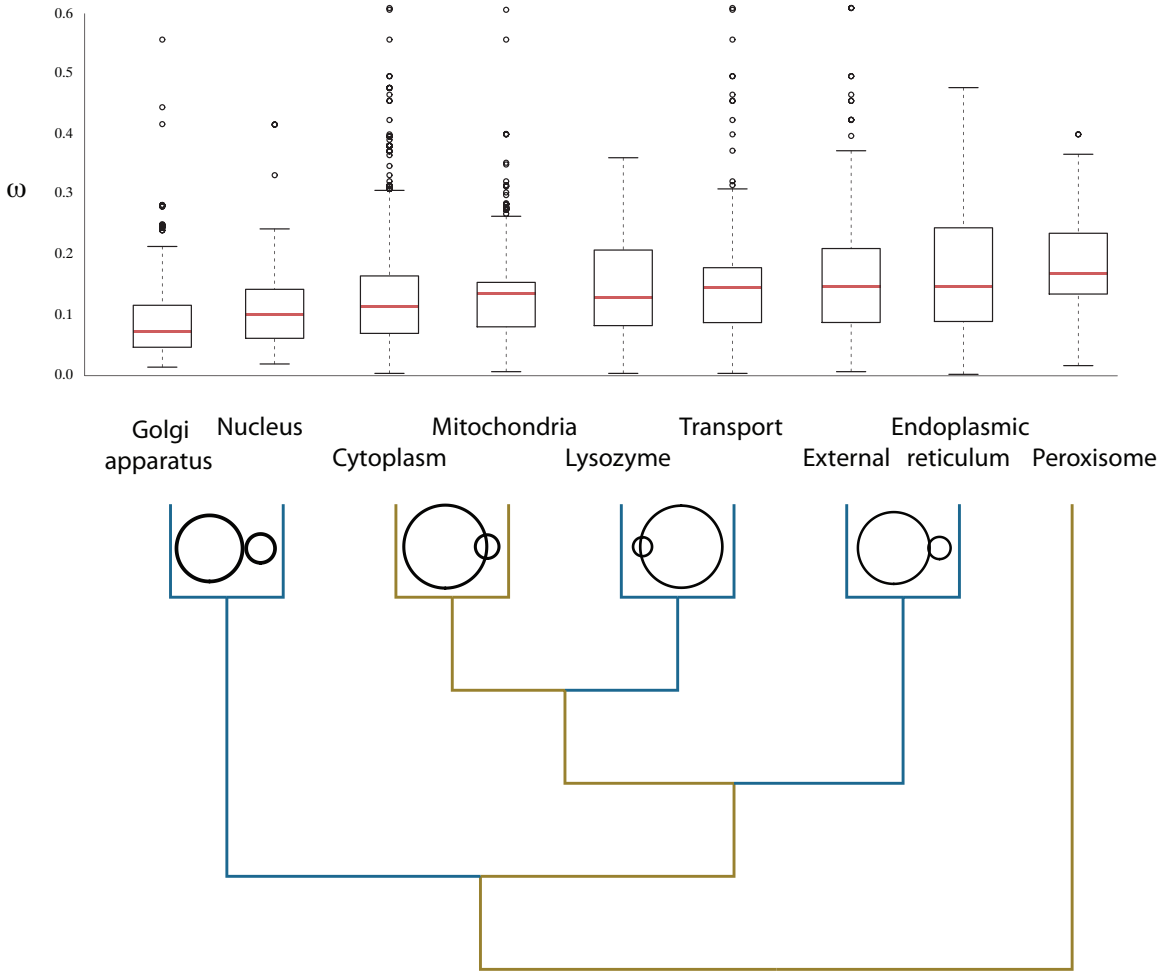

ω
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Golgi apparatus
Nucleus
Cytoplasm
Mitochondria
Lysozyme
Transport
External
Endoplasmic reticulum
Peroxisome

FIGURE 8

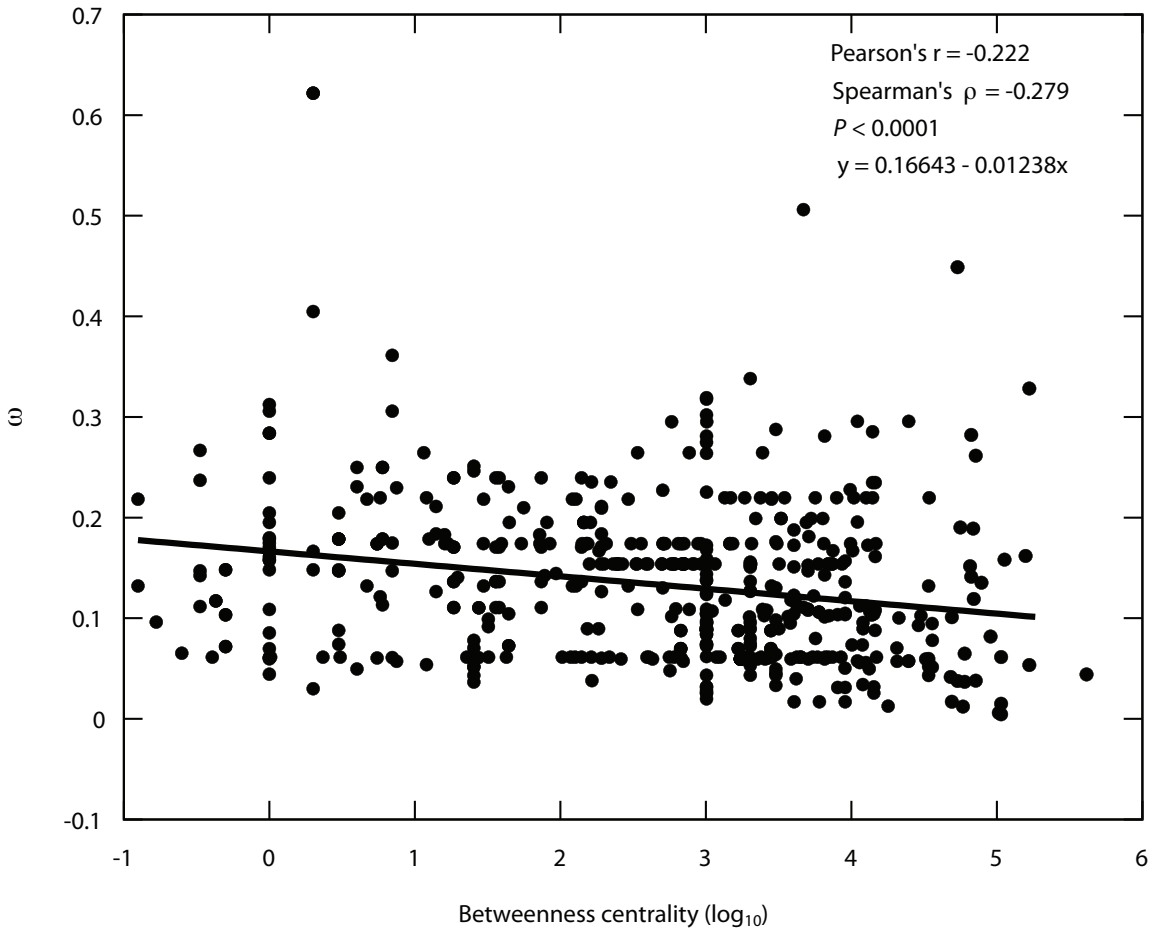

Pearson's r = -0.222
Spearman's ρ = -0.279
P < 0.0001
y = 0.16643 - 0.01238x
ω
Betweenness centrality (log10)
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-0.1
-1
0
1
2
3
4
5
6

FIGURE 9

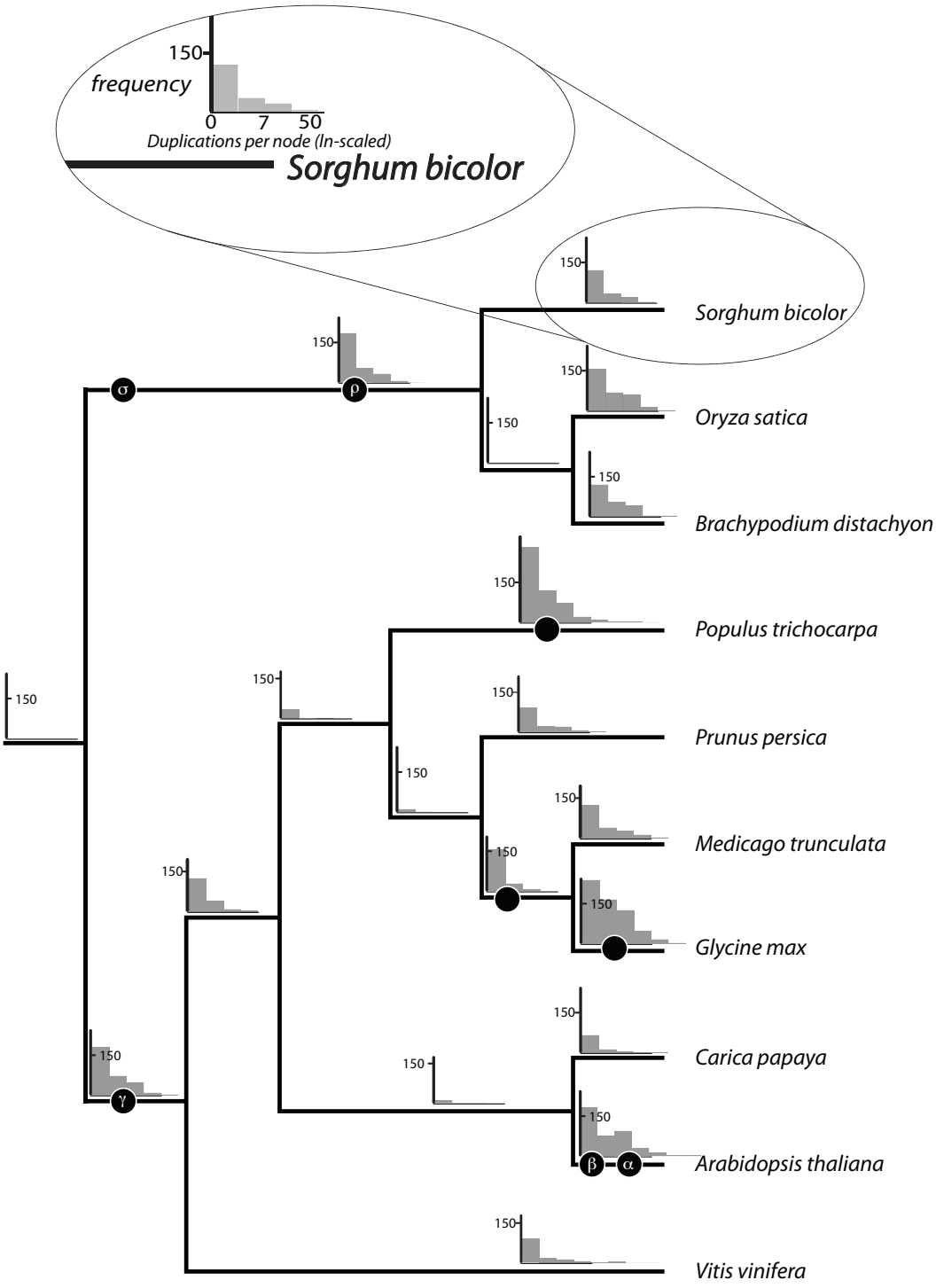

150
frequency
0 7 50
Duplications per node (ln-scaled)
Sorghum bicolor
Sorghum bicolor
Oryza satica
Brachypodium distachyon
Populus trichocarpa
Prunus persica
Medicago trunculata
Glycine max
Carica papaya
Arabidopsis thaliana
Vitis vinifera
σ
ρ
γ
β
α

FIGURE 10

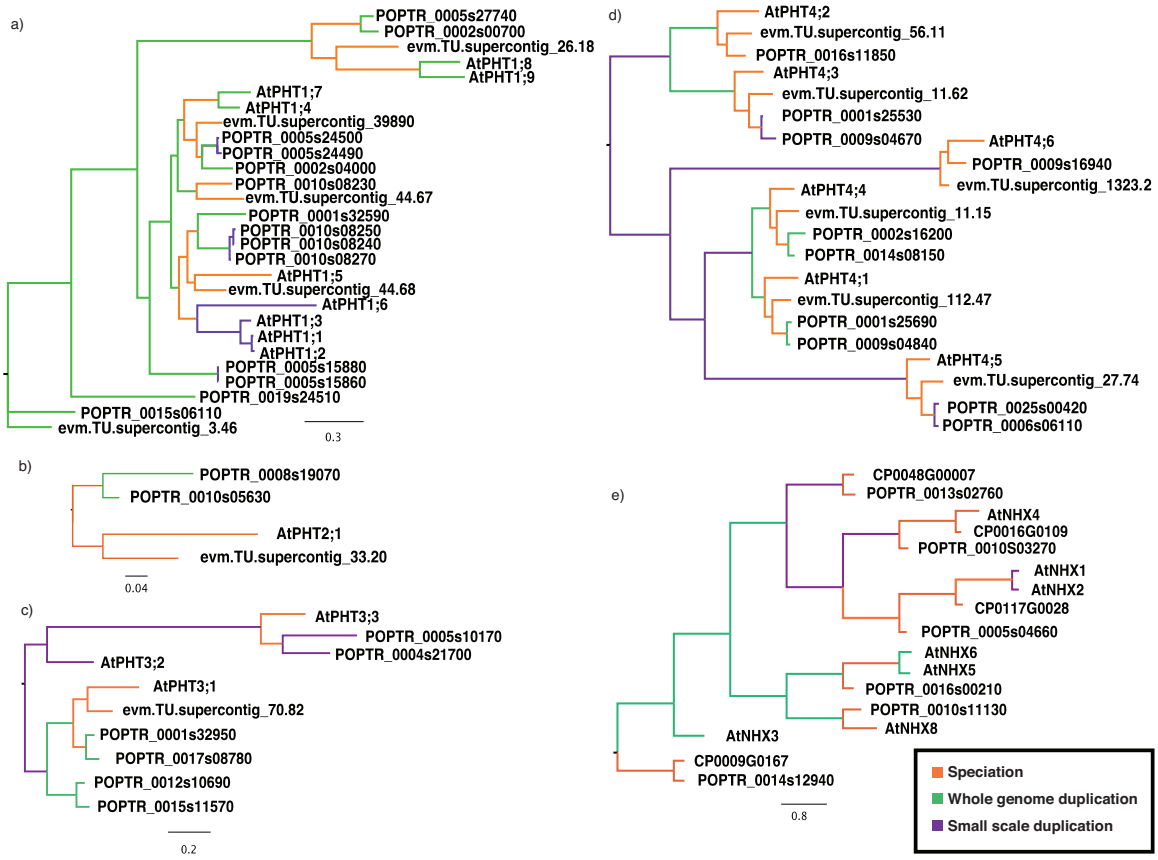

a)
POPTR_0005s27740
POPTR_0002s00700
evm.TU.supercontig_26.18
AtPHT1;8
AtPHT1;9
AtPHT1;7
AtPHT1;4
evm.TU.supercontig_39890
POPTR_0005s24500
POPTR_0005s24490
POPTR_0002s04000
POPTR_0010s08230
evm.TU.supercontig_44.67
POPTR_0001s32590
POPTR_0010s08250
POPTR_0010s08240
POPTR_0010s08270
AtPHT1;5
evm.TU.supercontig_44.68
AtPHT1;6
AtPHT1;3
AtPHT1;1
AtPHT1;2
POPTR_0005s15880
POPTR_0005s15860
POPTR_0019s24510
POPTR_0015s06110
evm.TU.supercontig_3.46
0.3
b)
POPTR_0008s19070
POPTR_0010s05630
AtPHT2;1
evm.TU.supercontig_33.20
0.04
c)
AtPHT3;3
POPTR_0005s10170
POPTR_0004s21700
AtPHT3;2
AtPHT3;1
evm.TU.supercontig_70.82
POPTR_0001s32950
POPTR_0017s08780
POPTR_0012s10690
POPTR_0015s11570
0.2
d)
AtPHT4;2
evm.TU.supercontig_56.11
POPTR_0016s11850
AtPHT4;3
evm.TU.supercontig_11.62
POPTR_0001s25530
POPTR_0009s04670
AtPHT4;6
POPTR_0009s16940
evm.TU.supercontig_1323.2
AtPHT4;4
evm.TU.supercontig_11.15
POPTR_0002s16200
POPTR_0014s08150
AtPHT4;1
evm.TU.supercontig_112.47
POPTR_0001s25690
POPTR_0009s04840
AtPHT4;5
evm.TU.supercontig_27.74
POPTR_0025s00420
POPTR_0006s06110
e)
CP0048G00007
POPTR_0013s02760
AtNHX4
CP0016G0109
POPTR_0010S03270
AtNHX1
AtNHX2
CP0117G0028
POPTR_0005s04660
AtNHX6
AtNHX5
POPTR_0016s00210
POPTR_0010s11130
AtNHX8
AtNHX3
CP0009G0167
POPTR_0014s12940
0.8
Speciation
Whole genome duplication
Small scale duplication

FIGURE 11

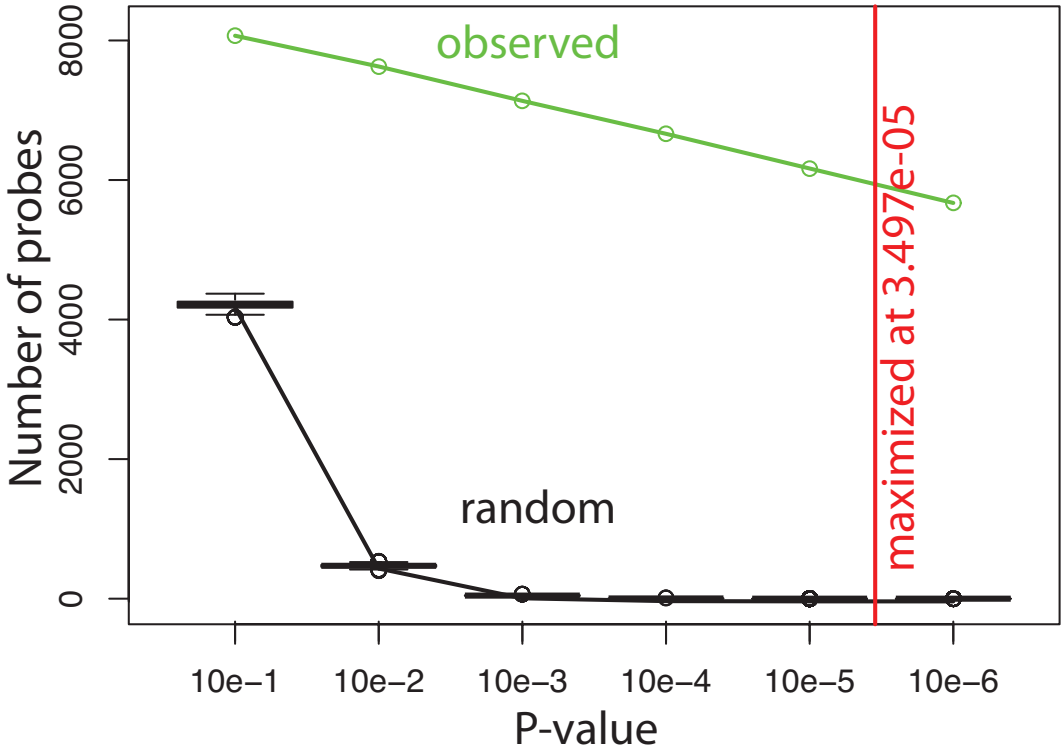

observed
random
maximized at 3.497e-05
Number of probes
0
2000
4000
6000
8000
10e−1
10e−2
10e−3
10e−4
10e−5
10e−6
P-value

FIGURE 12


Shared between
Gene expression spots
Proteins
Reactions
1 or more
19210
10924
1292
2 or more
2465
1603
322
3 or more
285
193
75
4 or more
36
18
12
5 or more
0
0
2

FIGURE 13

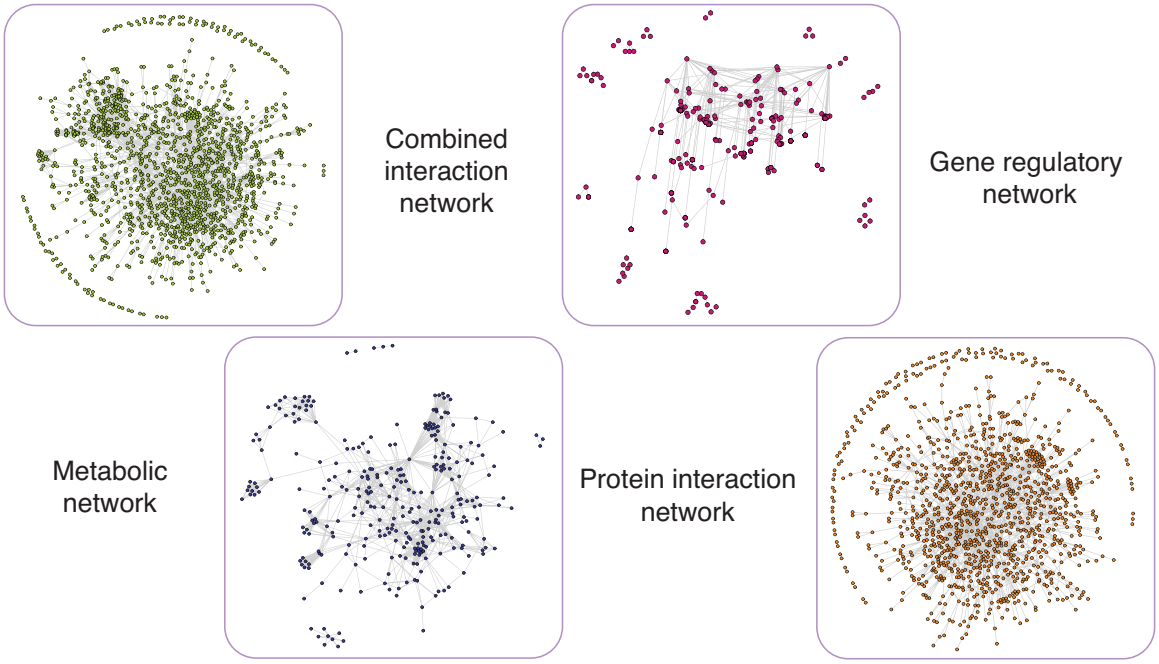

Combined interaction network
Gene regulatory network
Metabolic network
Protein interaction network

FIGURE S1

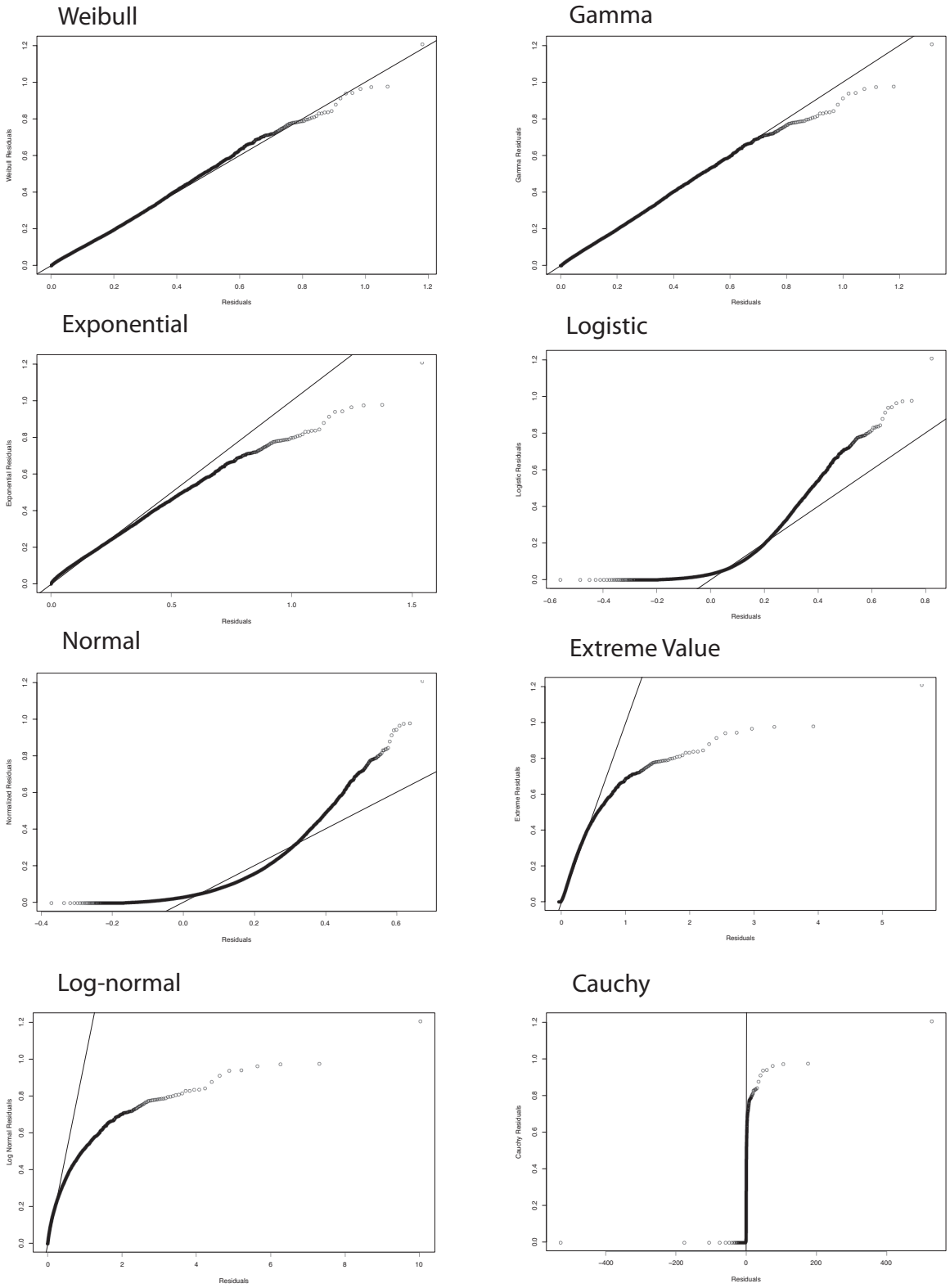

Weibull
Weibull Residuals
Residuals
Gamma
Gamma Residuals
Residuals
Exponential
Exponential Residuals
Residuals
Logistic
Logistic Residuals
Residuals
Normal
Normalized Residuals
Residuals
Extreme Value
Extreme Residuals
Residuals
Log-normal
Log Normal Residuals
Residuals
Cauchy
Cauchy Residuals
Residuals

FIGURE S2

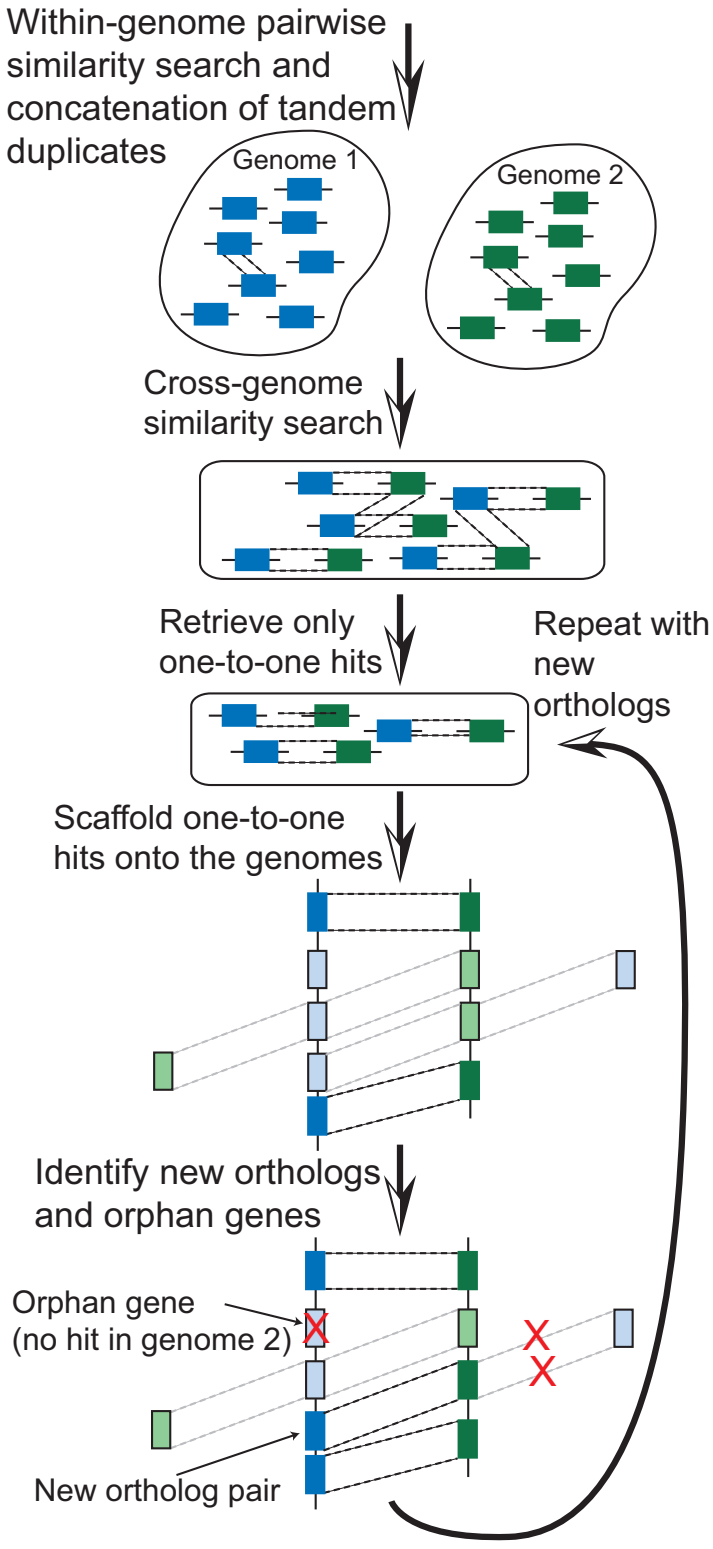

Within-genome pairwise similarity search and concatenation of tandem duplicates
Genome 1
Genome 2
Cross-genome similarity search
Retrieve only one-to-one hits
Repeat with new orthologs
Scaffold one-to-one hits onto the genomes
Identify new orthologs and orphan genes
Orphan gene (no hit in genome 2)
New ortholog pair

FIGURE S3

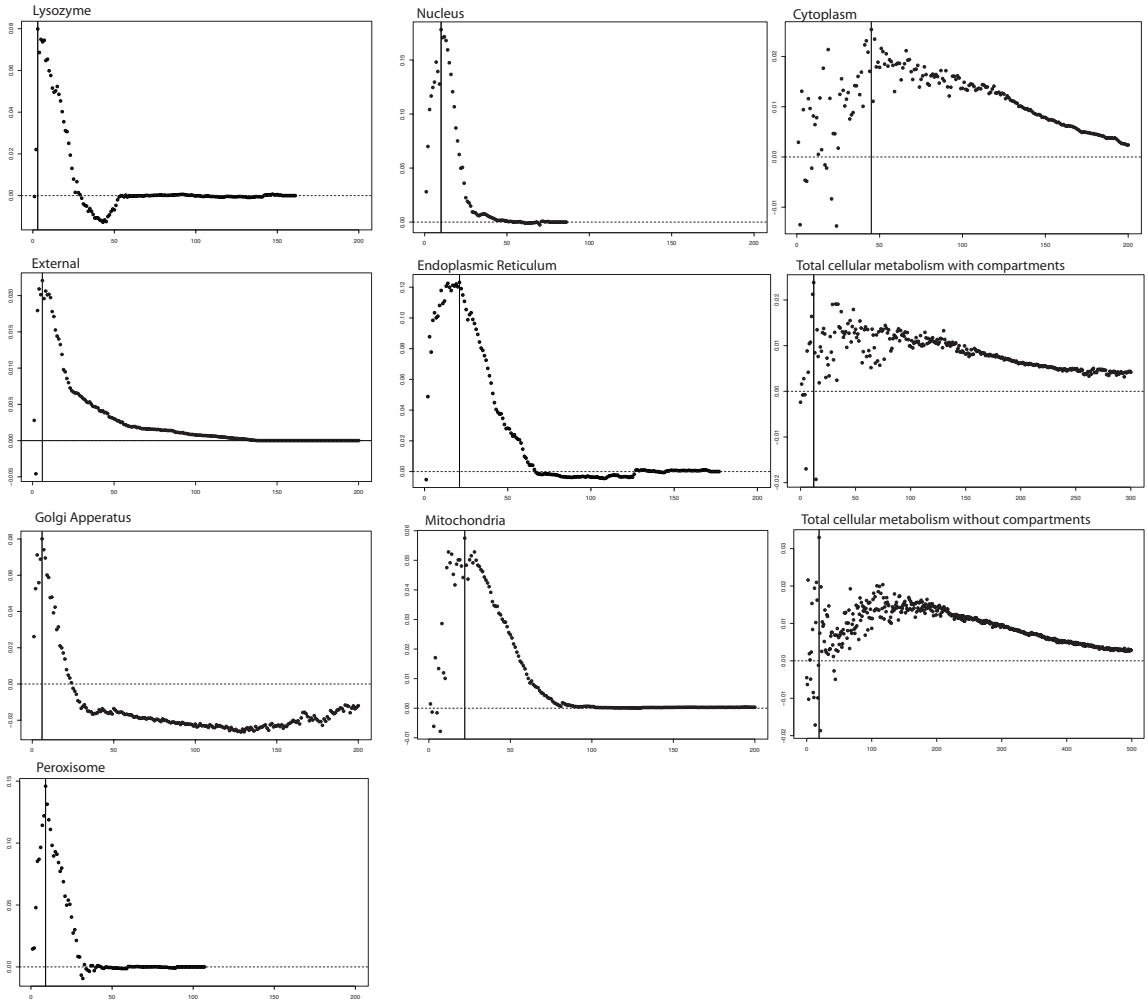

Lysozyme
Nucleus
Cytoplasm
External
Endoplasmic Reticulum
Total cellular metabolism with compartments
Golgi Apperatus
Mitochondria
Total cellular metabolism without compartments
Peroxisome

TABLE T1

Removed Metabolite	Compartment							
	Cytoplasm	Mitochondria	Nucleus	Endoplasmic Reticulum	Golgi apparatus	External	Lysozyme	Peroxisome
M_h[a]	Yes	Yes	No	Yes	Yes	Yes	Yes	No
M_h2o	Yes	Yes	No	Yes	No	No	Yes	No
M_atp	Yes	No	No	No	No	No	No	No
M_na1	Yes	No	No	No	No	No	No	No
M_adp	Yes	Yes	No	No	No	No	No	No
M_coa	Yes	Yes	No	No	No	No	No	No
M_pi	Yes	No	No	No	No	No	No	No
M_nad	Yes	No	No	No	No	No	No	No
M_amp	Yes	No	No	No	No	No	No	No
M_nadp	Yes	No	No	No	No	No	No	No
M_ppi	Yes	No	No	No	No	No	No	No
M_fad	No	Yes	No	No	No	No	No	No
M_o2	Yes	No	No	No	No	No	No	No
M_glu_DASH_L	Yes	No	No	No	No	No	No	No

a. Metabolite nomenclature following Duarte et al. [1]

References:

1. Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD, Srivas R, Palsson Bò: **Global reconstruction of the human metabolic network based on genomic and bibliomic data.** *Proc Natl Acad Sci U S A* 2007, **104:**1777-1782.

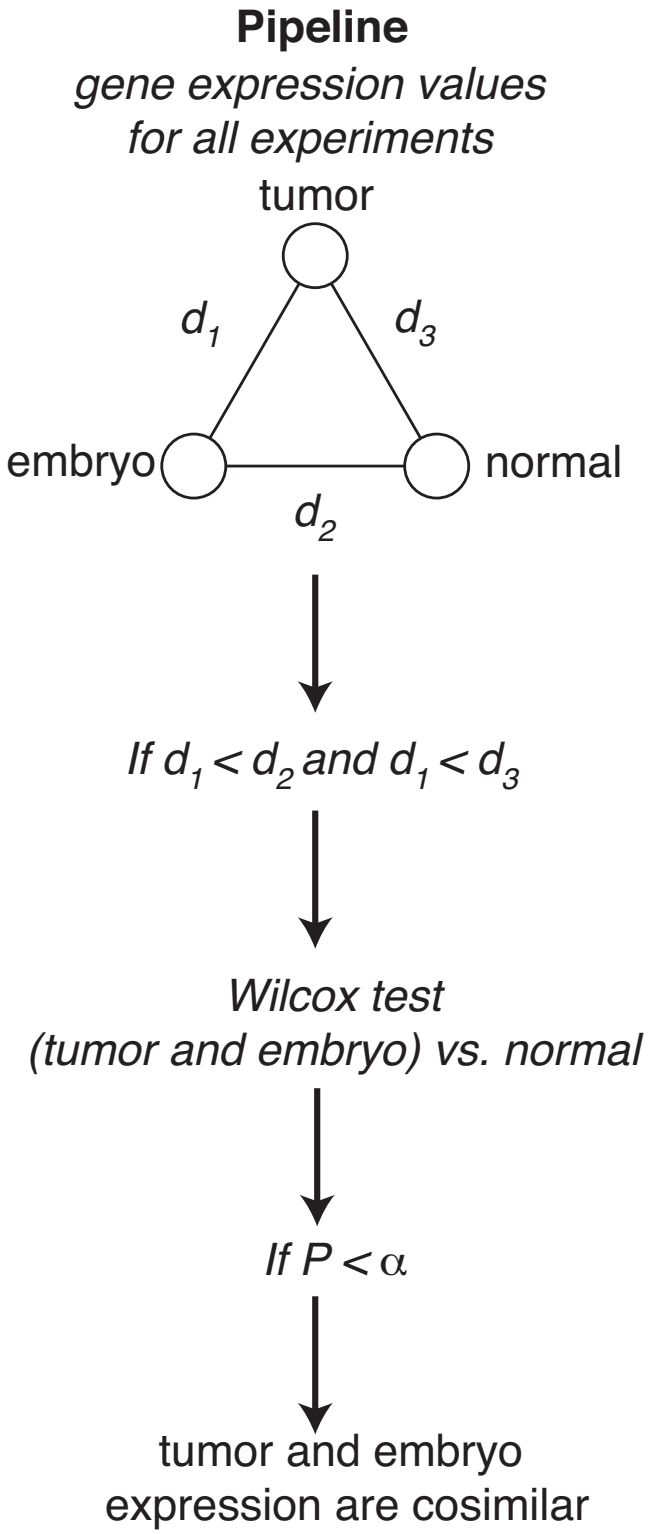
TABLE T2

Condensed Category	All GO Slim Categories within the Condensed Category
Cell Organization and Biogenesis	cell organization and biogenesis
Cell Wall	cell wall
Chloroplast and Plastid	chloroplast
	plastid
	plastid | other cytoplasmic components | other intracellular components
	plastid | other membranes | other cytoplasmic components | other intracellular components
	plastid | chloroplast | other cytoplasmic components | other intracellular components
	plastid | chloroplast | other cytoplasmic components | other intracellular components | plastid | chloroplast | other cytoplasmic components | other intracellular components
	plastid | chloroplast | other membranes | other cytoplasmic components | other intracellular components
Cytosol	cytosol
	cytosol | other cytoplasmic components | other intracellular components
	cytosol | ribosome | other cytoplasmic components | other intracellular components
Developmental Processes	developmental processes
	developmental processes | cell organization and biogenesis
	developmental processes | other cellular processes
	developmental processes | other cellular processes | cell organization and biogenesis
	developmental processes | response to abiotic or biotic stimulus
DNA or RNA Binding or Metabolism	DNA or RNA binding
	DNA or RNA metabolism
	DNA or RNA metabolism | other cellular processes | other metabolic processes
Electron Transport	electron transport or energy pathways
Endoplasmic Reticulum	ER
	ER | other membranes | other cytoplasmic components | other intracellular components
Extracellular	extracellular
Golgi Apparatus	Golgi apparatus

	Golgi apparatus | other cytoplasmic components | other intracellular components
	Golgi apparatus | other membranes | other cytoplasmic components | other intracellular components | Golgi apparatus | other membranes | other cytoplasmic components | other intracellular components
Hydrolase Activity	hydrolase activity
	hydrolase activity | other enzyme activity
	hydrolase activity | transferase activity
	hydrolase activity | transporter activity
Kinase Activity	kinase activity
	kinase activity | transferase activity
	kinase activity | transferase activity | receptor binding or activity
Mitochondria	mitochondria
	mitochondria | other cytoplasmic components | other intracellular components
	mitochondria | other membranes | other cytoplasmic components | other intracellular components
Nucleic Acid or Nucleotide Binding	nucleic acid binding
	nucleotide binding
	other binding | nucleotide binding
Nucleus	nucleus
	nucleus | nucleus | nucleus
	nucleus | other intracellular components
Plasma Membrane	plasma membrane
	plasma membrane | plasma membrane
Protein Binding or Metabolism	protein binding
	protein binding | protein binding
	protein metabolism
	protein metabolism | other cellular processes | other metabolic processes
	protein metabolism | other cellular processes | other metabolic processes | cell organization and biogenesis
Signal Transduction	signal transduction
	signal transduction | other cellular processes
Stimulus or Stress Response	response to abiotic or biotic stimulus
	response to stress | response to abiotic or biotic stimulus
	response to stress | response to abiotic or biotic stimulus | response to stress | response to abiotic or biotic stimulus
	response to stress | response to abiotic or biotic stimulus | response to stress | response to abiotic or biotic stimulus | response to stress | response to abiotic or biotic stimulus

	response to stress
	response to stress | cell organization and biogenesis
	response to stress | other cellular processes
	response to stress | other cellular processes | other metabolic processes
	response to stress | other cellular processes | other metabolic processes | response to abiotic or biotic stimulus
	response to stress | other metabolic processes
Transferase Activity	transferase activity
	transferase activity | other enzyme activity
Transporters or Transport	transport
	transport | DNA or RNA metabolism | other cellular processes
	transport | other cellular processes
	transport | other cellular processes | cell organization and biogenesis
	transport | other cellular processes | other metabolic processes | electron transport or energy pathways
	transport | transport
	transporter activity

FIGURE S4


Pipeline
gene expression values
for all experiments
tumor
d_1
d_3
embryo
normal
d_2
If $d_1 < d_2$ and $d_1 < d_3$
Wilcox test
(tumor and embryo) vs. normal
If $P < \alpha$
tumor and embryo
expression are cosimilar

VITA

Corey Hudson is a bioinformaticist and computational biologist working generally on the topic of the evolution of complex systems. Specifically, he works on how metabolic structure constrains the material for natural selection, and conversely how evolutionary processes structure metabolic systems. He has published on several systems including yeast, eutherian mammals, Angiosperm plants and cancer.