This is to certify that the
thesis entitled

RELEASE FROM ANTAGONISTIC PLEIOTROPY AND COEVOLUTION FOLLOWING GENE DUPLICATION IN FUNGAL MITOCHONDRIAL HEAT SHOCK PROTEINS

presented by

Krista Gudrais Reitenga

has been accepted towards fulfillment
of the requirements for the

Master of Science	degree in	Microbiology & Molecular Genetics

Major Professor's Signature

3/20/1

Date

RELEASE FROM ANTAGONISTIC PLEIOTROPY AND COEVOLUTION FOLLOWING GENE DUPLICATION IN FUNGAL MITOCHONDRIAL HEAT SHOCK PROTEINS

By

Krista Gudrais Reitenga

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

MASTER OF SCIENCE

Microbiology & Molecular Genetics

2009

ABSTRACT

RELEASE FROM ANTAGONISTIC PLEIOTROPY AND COEVOLUTION FOLLOWING GENE DUPLICATION IN FUNGAL MITOCHONDRIAL HEAT SHOCK PROTEINS

By

Krista Gudrais Reitenga

SSC1 is a gene that encodes a multifunctional mitochondrial heat shock protein that gave rise to SSQ1 by gene duplication in a subset of yeasts. In contrast to the multiple chaperone functions carried out by most heat shock proteins, Ssq1p is specialized in Fe/S cluster assembly. Ssc1p and Ssq1p both participate in the formation of Fe/S clusters and require interaction with Jac1p. Biochemical experiments and genetic manipulation of *Saccharomyces cerevisiae* have provided evidence that Ssq1p and Jac1p may have coevolved to optimize a specialized interaction. Together, these factors present a unique opportunity to understand how natural selection shapes the functional coevolution of gene duplicates. We hypothesized that the divergence of SSC1 and SSQ1 resulted in the coevolution of the JAC1-SSQ1 pair. Here, we report that, in the presence of a rapidly evolving SSQ1, the average rate of JAC1 evolution has decreased. Our study also supports a burst of adaptive evolution in SSQ1 immediately following its inception. Additionally, both SSC1 and SSQ1 exhibit elevated rates of evolution when co-occurring. When taken together, the signatures of ancestral and present-day selection point to a release from antagonistic pleiotropy that facilitated coevolution between JAC1 and SSQ1. This study offers detailed evidence that the duplication of multifunctional genes allows for the coevolution of interacting proteins to optimize a paired function.

TABLE OF CONTENTS

DISCU

APPE

REFE

LIST OF TABLES

Figur
relati
dupli

Figur
distri

Figur

Figure
constr

Figure
input

Figure
and ta

Figure

Figure

Figure

Figure
SSC1 e

Figure
into a d
topolog

Figure
cerevisi
constra

Figure B

Figure B
alignme

Figure B
alignme

LIST OF FIGURES

SECTI

INTRO

groups

races fo

beyond

their pa

biochen

within c

interdep

can exe

pathway

or gener

fitness a

coevolu

congeni

careful

I

importa

biochem

many kn

the clas

SECTION I. BACKGROUND

INTRODUCTION

Coevolution has long been appreciated as a mechanism that operates between groups of organisms with the potential to create ecological mutualisms and initiate arms races for adaptation. However, coevolution is a pervasive phenomenon which extends beyond macroscopic interactions such as among flowers and their pollinators or hosts and their parasites. Phenotypes that determine ecological fitness are the result of complex biochemical pathways. Coevolution, therefore, also takes place among the molecules within organisms, and at times, may even be responsible for species-level interdependencies and competitive strategies. Through molecular coevolution, proteins can exert a selective influence over interacting partners or components of a biochemical pathway to favor molecular cooperation or antagonism. Proteins may become specialist or generalist as a result. Therefore, the evolutionary success of organisms hinges upon the fitness advantages conferred by molecular components. Additionally, molecular coevolution may influence genetic interactions, which can, among other things, lead to congenital diseases and contribute to the process of speciation. Coevolution thus merits careful study to facilitate our understanding of many fundamental aspects of biology.

Heat shock proteins (Hsps) constitute a group of proteins that are of central importance to nearly all organisms. A great deal of data has been amassed concerning the biochemical and genetic properties of Hsps and has led to detailed understanding of the many known functions of these proteins. While highly conserved and slowly evolving, one class of Hsps exhibits dynamic variation in their gene copy number. In one

inte

asse

deci

inter

exer

char

numb

pathw

Hsp7

increa

kiloDa

a near

throug

absenc

known

transpo

keeping

transcri

integrity

identific

and Hor

interesting case, gene amplification has led to the specialization of an Hsp in Fe/S cluster assembly, an essential pathway for which biochemical mechanisms are only now being deciphered. Furthermore, the multiple functions carried out by Hsps necessitate interaction with a wide variety of protein partners and creates ample potential for Hsps to exert a reciprocal influence on other constituents of networks. Combined, the characteristics of Hsps present a unique opportunity to study how changes in gene copy number affect coevolution of interacting partners within an essential biochemical pathway.

Hsp70s: Characteristic Features

So named for their discovery (Ritossa 1962) as a group of proteins that exhibited increased abundance in cells following heat stress, heat shock proteins of the 70 kiloDalton (kDa) class (Hsp70s) represent a multi-gene family of protein chaperones with a nearly ubiquitous distribution within the tree of life. Homologs have been found throughout the Bacteria and Eukarya, as well as some representatives in Archaea (the absence of Hsp70s in Archaea has been reported by (Gribaldo et al. 1999)). Hsp70s are known to participate in an array of indispensable functions associated with the folding, transport, and degradation of a wide variety of polypeptides. Hsp70s may perform house-keeping functions constitutively under many physiological conditions or exhibit transcriptional up-regulation in response to environmental stresses in order to protect the integrity of polypeptide components of the cell (Boorstein et al. 1994). Since their first identification in heat stressed drosophila cells in the 1970's (Tissieres et al. 1974; Bukau and Horwich 1998) many other stimuli have been demonstrated to trigger increased

synthesis of Hsps, including exposure to ethanol, anoxic conditions, heavy metal ions, and ultraviolet light (Lindquist and Craig 1988). The Hsp70s have an extremely slow rate of evolution and share a common tri-domain protein structure across all three domains of life. The canonical form comprises a 44 kDa amino-terminal ATPase domain, an 18 kDa peptide binding domain (Wang et al. 1993), and a 10 kDa carboxy-terminal domain of variable amino acid composition. Hydrolysis of adenosine-5'-triphosphate (ATP) regulates the induction of a conformational change within the Hsp70s' substrate binding pocket and consequent binding and release of hydrophobic regions of the substrate polypeptide (Bukau and Horwich 1998). The functions of the proteins comprising the Hsp70 family are so well conserved that, when expressed by a mammalian cell, an Hsp70 protein from a fruit fly is able to perform heat stress protection (Pelham 1984).

Hsp70 Phylogenetic Distribution and Gene Family Evolution

Though many Hsp70 homologs have retained equivalent functional abilities across divergent organismal taxa, the number of Hsp70 genes encoded within a genome shows plasticity, a dynamic rife with evolutionary and ecological potential. Comparative sequence analyses have revealed that the eukaryotic Hsp70 genes, all encoded within the nuclear genome, constitute four phylogenetically distinct clades. The clades are characterized by common intracellular localization of the protein products to either the mitochondria, endoplasmic reticulum, plastids, or cytoplasm (Boorstein et al. 1994). Nearly all eukaryotes contain at least three Hsp70 gene copies; the budding yeast *Saccharomyces cerevisiae* possesses 9 cytosolic (cyt), 3 mitochondrial (mt), and 2 endoplasmic reticulum (er) isoforms of Hsp70. However, the number of paralogs

encod

Hsp7

specie

the rac

encodi

establi

the mu

genes f

occurre

genes.

arrange

Nei 20

and Fed

biologic

within g

limit on

Hsp70 g

Hsp70 e

may offe

of a corr

well doc

face of o

encoded by different eukaryotic genomes can vary widely, as exemplified by the 10 Hsp70 genes found in the nematode *Caenorhabditis elegans* and 19 in the closely related species *C. briggsae* (Nikolaidis and Nei 2004). An early gene duplication event prior to the radiation of eukaryotic species gave rise to the cytHsp70s and erHsp70s. The genes encoding Hsp70s of the mitochondria and plastids are likely of bacterial origin. After establishment of the bacterial endosymbionts that are hypothesized to have gaven rise to the mitochondria and plastids in an ancestral eukaryote, lateral transfer of the Hsp70 genes from the organellar genomes to the nuclear chromosome is thought to have occurred (Muhlenhoff and Lill 2000).

Gene duplication is known to play an important role in the amplification of Hsp70 genes. Duplication is likely facilitated by inverted and tandem cytHsp70 gene pair arrangements common to the genomes of the Caenorhabdid nematodes (Nikolaidis and Nei 2004), mosquito (Benedict et al. 1993), rat (Walter et al. 1994), fruit fly (Bettencourt and Feder 2002), fugu (Lim and Brenner 1999), and human (Tavaria et al. 1996). A biological cost-benefit balance may play a role in governing the cytHsp70 copy number within genomes. Cells sustain a cost of deleterious effects on growth, imposing an upper limit on the optimum Hsp70 expression level due to a cost of replicating additional Hsp70 genes, energy required for additional translation, or a toxic effect associated with Hsp70 expression above a certain threshold. Conversely, an increase in Hsp70 expression may offer the benefit of an enhanced ability to survive environmental stresses. Evidence of a correlation between Hsp70 expression level and degree of thermotolerance has been well documented in *Drosophila* (Feder et al. 1996). Thermotolerance and survival in the face of other environmental stressors by Hsp70 buffering therefore constitute ecologically

relevant

gene cor

frequent

Bettenc

concurre

spreadin

reported

within ar

divergen

mutation

of purify

deleterio

As a con

compartr

Hsp70s f

Nei 2004

cytHsp70

selection

process o

1994). Th

divergent

relevant phenotypes on which natural selection may act.

Examination of two cytHsp70 paralog clusters from *Drosophila* revealed that gene conversion between and among groups of physically clustered genes is likely to be a frequent event which contributes to the homogenization of Hsp70 copies within a group (Bettencourt and Feder 2002). Gene conversion maintains sequence similarity, while concurrently enabling a subgroup of cytHsp70s to diverge in a concerted manner by spreading new mutations among copies. Gene conversion among cytHsp70 has also been reported in the nematodes (Nikolaidis and Nei 2004) and has been suspected to occur within angiosperm plants (Renner and Waters 2007). Alternatively, the lack of divergence among a group of Hsp70s may be due to slow evolutionary rates. The bias of mutations exhibited among paralogs toward synonymous changes implies the large role of purifying selection. In conjunction with gene homogenization, the spread of deleterious changes among Hsp70 paralogs is disfavored (Bettencourt and Feder 2002). As a consequence, Hsp70 sequences of proteins localized to the same cellular compartment from distantly related organisms tend to share greater similarity than Hsp70s from different cellular compartments within the same organism. (Nikolaidis and Nei 2004).

Unlike the mechanisms of convergent evolution that characterize many cytHsp70s, the mt- and erHsp70s show evidence of divergent evolution. Diversifying selection is a mechanism which drives divergent evolution and is facilitated by the process of independent gene duplication and loss events among lineages (Ota and Nei 1994). The birth and death of paralogs is a feature of the mt- and erHsp70s. While many divergent eukaryotes, including *Drosophila*, nematodes, and the marine diatom

Thalas

duplica

Arabid

Waters

plastid

to the H

the plas

cyanoba

encode

with 3 h

cyanoba

shown t

dnaK3 l

photosy

localizat

I

detectio

some ta

recogniz

to contro

of the u

The alter

Thalassiosira pseudonana encode a single mtHsp70, the mtHsp70s have undergone duplication in other eukaryotic lineages, with *Saccharomyces cerevisiae* possessing 3, *Arabidopsis thaliana* with 2, and *Plasmodium falciparum* genomes with 1 (Renner and Waters 2007).

Congruent with the hypothesis for the origin of eukaryotic mitochondria and plastids from ancient bacterial endosymbionts, mtHsp70 genes display greatest similarity to the Hsp70 bacterial homologues from representatives of the α-*Proteobacteria*, whereas the plastid Hsp70 genes most closely resemble the heat shock protein genes of cyanobacteria (Boorstein et al. 1994). Within the Bacteria, some organisms may also encode multiple Hsp70s (referred to as dnaK or heat shock cognate, hsc, in the bacteria), with 3 homologs in the *Escherichia coli* genome (Itoh et al. 1999) and the cyanobacterium *Synechococcus* (Ward-Rainey et al. 1997). Bacterial Hsp70s have been shown to display paralog-specific localization patterns. In the case of *Synechococcus,* dnaK3 localizes specifically to the cytosolic thylakoid membrane of an oxygen-producing photosynthetic system (Nimura et al. 1996), analogous to the plastid-specific organellar localization observed in some eukaryotes.

In contrast to the ever-present status of Hsp70 in Eukarya and Bacteria, the detection of gene homologs within Archaea has been patchy, with presence reported in some taxa (Macario et al. 1991; Gupta and Singh 1992, 1994), but absence of recognizable homologs in others (Lange et al. 1997). These observations have given rise to controversy surrounding the origin of the archaeal Hsp70 and challenge the reliability of the use of Hsp70 as a phylogenetic marker with respect to the three domains of life. The alternative hypotheses of lateral acquisition in a subset of lineages (Philippe et al.

1999) and differential gene loss (Gupta 1999) have also been proposed.

Yeast Mitochondrial Hsp70s

The plasticity of gene copy number within the Hsp70 gene family has produced a particularly interesting outcome within the yeast mtHsp70s. Gene duplication has given rise to a functionally specialized protein that can be readily studied in the experimentally tractable model eukaryote, *Saccharomyces cerevisiae*. *S. cerevisiae* encodes three mtHsp70s: Ssc1p, the most abundant Hsp70 that functions within the organelle, plus Ssq1p and Ecm10p, two constitutively present forms of rarer abundance. Included in all three yeast mtHsp70 sequences is a leader sequence that targets the protein products for import into the mitochondria, where they function in the matrix (Craig 1989). In an event independent of the whole genome duplication estimated to have occurred about 150 million years ago in yeast (Langkjaer et al. 2003), SSQ1 arose from SSC1 by gene duplication prior to the most recent common ancestor of *S. cerevisiae* and *Candida albicans* (see Figure 1). Additionally, the paralog SSQ1 has been identified in all descendent fungal taxa studied (Schilke et al. 2006). The duplication of SSC1 is in agreement with the observation that slowly evolving genes in *S. cerevisiae* tend to duplicate, with subsequent retention of paralogs, more frequently than fast evolving genes (Davis and Petrov 2004). Later, ECM10, a third yeast mtHsp70, was generated during the whole genome duplication believed to have occurred in the most recent progenitor of the clade that includes *S.cerevisiae* and *S. castellii* (Kellis et al. 2004) (see Figure 1). While ECM10 now shares 82% amino acid sequence identity with SSC1 of *S. cerevisiae* (Baumann et al. 2000), SSQ1 has undergone greater divergence, particularly

wit

with

chro

effic

cont

evol

funct

Figure 1
selected
modifie
likelihoo
alignmer
supporte
a mtHsp7
which a v

within the substrate-binding domain, sharing an overall amino acid identity of only 52% with SSC1 (Schilke et al. 2006). Each yeast mtHsp70 is located on a separate nuclear chromosome, a feature which has the potential to result in disparate mutation rates and efficiencies of natural selection which act on the three mtHsp70 genes. The genomic context within which the mtHsp70 genes reside can therefore exert an influence on evolutionary rates of these genes independent of their respective protein structure and function (Pal et al. 2006).

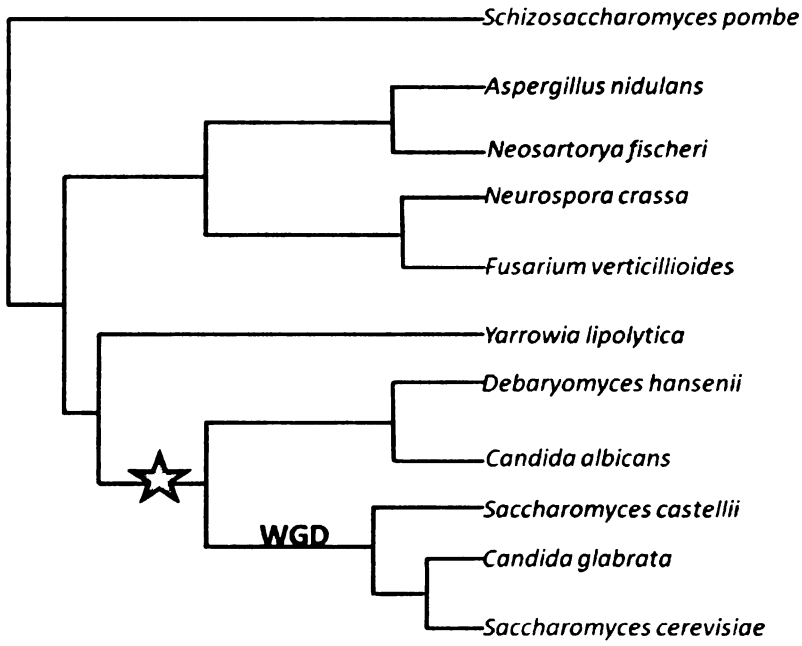


Figure 1: A simplified cladogram representing the evolutionary relationships among selected fungi in relation to mtHsp70 gene duplication events. This cladogram is a modified version of that constructed by Fitzpatrick et al. (2006) using maximum likelihood to infer the organismal relationships among fungi based on a concatenated alignment of 153 universally distributed fungal genes. All branches shown were supported with a bootstrap value of 100. The gray star indicates the lineage within which a mtHsp70 gene duplication gave rise to SSQ1. 'WGD' indicates the lineage within which a whole genome duplication took place, giving rise to ECM10.

SSC

mol

myr

chai

107

prot

cycl

pept

facil

matr

prot

Pept

is ye

term

assoc

is ne

Prote

expe

form

Dure

in the

SSC1: A Multifunctional Mitochondrial Hsp70

Ssc1p is a constitutively expressed, essential protein that functions as the major molecular chaperone within the matrix of the yeast mitochondrion and interacts with a myriad of different peptides. The constitutive chaperone tasks of Ssc1p involve peptide chain folding, unfolding and translocation necessary for mitochondrial biogenesis. About 10% of the Hsp70 protein present in the mitochondria acts as a component of the pre-protein translocase of the inner membrane (TIM) complex. As a TIM constituent, Ssc1p cyclically binds and releases polypeptides to assist the pumping of nuclear-encoded peptide chains across the inner membrane of the mitochondrion. Subsequently, Ssc1p facilitates folding of the chains into their native conformation as they emerge into the matrix (Neupert 1997). Because many proteins translated in the cytosol become folded prior to their import across the mitochondrial membranes, protein unfolding into linear peptide chains appropriate for translocation through the TIM complex is also critical, and is yet another function performed by Ssc1p via interaction with a substrate peptide's N-terminal pre-protein signal sequence (Lim et al. 2001). Ssc1p can also be found associated with mitochondrial ribosomes to fold proteins into their native conformation as newly synthesized peptides emerge during translation.

Under conditions of heat stress, Ssc1p protects the cell from the toxic effects of protein denaturation and aggregation within mitochondria. For instance, Ssc1p is responsible for maintaining Var1p, a subunit of mitochondrial ribosomes, in a soluble form to prevent aggregation or misfolding prior to ribosome assembly, a danger met with increased potential during heat shock (Herrmann et al. 1994). Further, Ssc1p plays a role in the synthesis of mitochondrial DNA as a partner in the Hsp70-Hsp78 mitochondrial

bicha

a part

therm

mtDN

Mip1p

polym

Mip1p

cluster

creatior

linked t

one dep

Iron-Su

ubiquito

process

respirati

protein

mechani

necessar

purine m

shown to

bichaperone system. In yeast, this system is critical to the maintenance and restoration of a particularly thermosensetive enzyme, Mip1p, the mtDNA polymerase, during severe thermal stress. The Hsp70-Hsp78 bichaperone is known to localize within protein-mtDNA complexes known as nucleoids, where the bichaperone may act to quickly refold Mip1p within the nucleoid scaffold leading to protection and reactivation of the mtDNA polymerase. Reactivation of Mip1p is more efficient than importing newly synthesized Mip1p into the mitochondrion (Germaniuk et al. 2002).

In addition to these classical roles as a chaperone, Ssc1p is involved in Fe/S cluster biosynthesis, a function that was encoded by the ancestral mtHsp70 prior to the creation of SSQ1 (Schilke et al. 2006). The process of Fe/S cluster assembly is tightly linked to the mitochondria in eukaryotes and has been appreciated only in recent years as one dependent on an enzyme-mediated biochemical pathway (Zheng et al. 1993).

Iron-Sulfur Cluster Assembly

From a broad perspective, it is no understatement to characterize Fe/S clusters as ubiquitous chemical structures that enable biochemical reactions essential to the processes that drive Earth's ecology, since these units make photosynthesis, cellular respiration, and nitrogen fixation possible. Serving as inorganic cofactors for a variety of proteins, Fe/S clusters participate in substrate binding and dictate many catalytic mechanisms via oxidation and reduction within enzymes. Fe/S cluster proteins are thus necessary for the citric acid cycle, haem biosynthesis, DNA repair, protein synthesis, and purine metabolism (Rouault and Tong 2005). Additionally, Fe/S clusters have been shown to sense oxidative stress and intracellular concentrations of iron to mediate cellular

resp

200

synt

sulf

meta

resid

yet u

Hsp

expe

assis

appr

mapp

struc

Fe S

SSC1

throu

NIF

Pectu

devot

Differ

Biosth

responses, sometimes as Fe/S cluster-containing transcription factors (Kiley and Beinert 2003).

Though many details remain to be clarified, the general mechanism for the synthesis of Fe/S cluster assembly involves an initial step of structurally coordinating sulfur and iron into a cluster on a scaffold protein and the subsequent transfer of the metallocluster to a substrate apoprotein. While enzymatic abstraction from cysteine residues is known to supply the sulfur for Fe/S cluster biogenesis, the source of iron has yet to be elucidated (Lill and Muhlenhoff 2008). Several roles have been proposed for Hsp70 chaperones in the context of Fe/S cluster assembly, though none have been proven experimentally. Hypothesized Hsp70 functions in Fe/S cluster biogenesis include assisting the transfer of assembled Fe/S clusters from the scaffold protein to the recipient apoprotein, or binding to Fe/S assembly proteins and/or substrate apoproteins to prevent inappropriate oxidation of cysteine residues that serve as ligands to coordinate the Fe/S structure (Muhlenhoff and Lill 2000). One certainty that emerges regarding the process of Fe/S cluster assembly is that this multi-step pathway is rife with ample potential for SSC1, SSQ1, and their co-chaperone, JAC1, to interact with many protein players.

As a testament to Fe/S cluster essentiality, three different pathways have arisen throughout the tree of life dedicated to Fe/S cluster biogenesis: the nitrogen fixation (NIF), iron-sulfur cluster (ISC), and sulfur utilization factor (SUF) pathways. The NIF pathway consists of a set of genes highly conserved in azototrophic bacteria and is devoted to the formation of Fe/S clusters exclusively for the maturation of the nitrogenase enzyme. The more general ISC pathway genes interact to assemble Fe/S prosthetic groups onto a variety of apoproteins (Zheng et al. 1998). This second system is

uti

thr

anc

the

200

func

Inste

clust

disco

follow

in con

to be r

al. 20

the SU

assemb

among

compo

All log

that co

the pro

protein

the mu

utilized by a much broader distribution of organisms and shows strong conservation throughout the Bacteria, particularly within the α-*proteobacteria*, the mitochondrial ancestor of which is hypothesized to have bestowed an intact ISC biosynthesis system to the Eukarya with subsequent preservation from yeast to humans (Lill and Muhlenhoff 2008). While the Archaea encode many proteins which rely on Fe/S clusters for their functions, this domain of life lacks homologs of both NIF and ISC assembly systems. Instead, these microbes encode genes homologous to some of the genes of the third Fe/S cluster assembly pathway, SUF. The SUF operon encodes a redundant pathway discovered in *Escherichia coli* when a small degree of Fe/S enzyme activity was retained following deletion of the bacteria's ISC operon (Takahashi and Tokumoto 2002). Later, in contrast to the housekeeping function of the ISC pathway, the SUF pathway was found to be required by *E. coli* under conditions of Fe starvation and oxidative stress (Outten et al. 2004). SUF homologs have also been identified within plastid genomes. Additionally, the SUF system may have served as the origin for the scaffold protein of the ISC cluster assembly pathway in some bacteria (Takahashi and Tokumoto 2002).

Though the NIF, ISC, and SUF pathways function independently, similarities among the systems abound, which have facilitated the identification of the functional components that perform analogous tasks within yeast mitochondria for Fe/S biogenesis. All together, at least 15 proteins have been implicated as Fe/S cluster assembly proteins that cooperate in the mitochondrial matrix (Lill and Muhlenhoff 2008). Though many of the proteins that require Fe/S clusters function within mitochondria, some cytosolic proteins also contain Fe/S clusters and are believed to receive Fe/S clusters exported from the mitochondria, since Fe/S cluster biogenesis has not been demonstrated to occur in the

cytos

mitoc

Fe/S

SSQ1

a loss

bioch

SSQ1

mitoc

conce

invest

conver

enzym

Within

ferredo

holoen

interac

also be

domain

cluster

Fe/S bi

cytosol. The reducing chemical conditions and lower partial pressure of O_2 within the mitochondrial matrix relative to the cytosol may have favored the establishment of the Fe/S cluster biogenesis pathway within this organelle (Muhlenhoff and Lill 2000).

SSQ1: The Mitochondrial Hsp70 Iron-Sulfur Cluster Specialist

SSQ1 has become specialized in the assembly of Fe/S clusters, but at the price of a loss in the multifunctionality displayed by its paralog SSC1. Recent genetic and *in vitro* biochemical experiments offer support of the functional specialization of SSQ1. When SSQ1 was deleted from the *S. cerevisiae* genome, mutants accrued iron within the mitochondrial matrix with a concurrent reduction in Fe/S cluster-containing enzyme concentrations and protection against oxidative agents (Voisine et al. 2000). To further investigate this phenotype, authors of another study used an assay to observe the conversion of ferredoxin, a mitochondrial protein that requires an Fe/S cluster for enzymatic function, from its apo-form to its holo-form within isolated mitochondria. Within mitochondria extracted from an *S. cerevisiae* SSQ1 deletion strain, the majority of ferredoxin failed to mature into a holoenzyme. Ferredoxin that did achieve the holoenzyme state was found to have reduced kinetic character (Lutz et al. 2001). The interaction of Ssq1 with known components of the Fe/S cluster assembly pathway has also been tested, and investigators observed efficient binding of a purified protein binding domain fragment of Ssq1p to a peptide fragment of the scaffold protein involved in Fe/S cluster formation (Schilke et al. 2006). These results suggest that Ssq1p is important for Fe/S biogenesis and is able to physically interact with a key component of the pathway.

Consistent with the hypothesis of specialization and the concomitant loss of

ances

bindi

mtHs

follow

memb

bind t

et al. 2

require

finding

(Schlik

These r

SSQ1 d

matrix v

protectio

expressio

SSC1 an

the same

al. 2006),

of this ov

allows AD

The greate

limit the a

ancestral mtHsp70 function after duplication, Ssq1p was found to have very weak binding specificity to peptides known to be bound by Ssc1p of both pre- and post mtHsp70 duplication (Andrew et al. 2006). Additionally, Ssq1p was not detected following co-immunoprecipitation with Tim44p, a component of the inner mitochondrial membrane peptide translocase, was attempted. This demonstrated that Ssq1p does not bind to Tim44p, in contrast to Ssc1p, which acts as a subunit of the TIM complex (Lutz et al. 2001). Ssq1 has lost the ability to bind the same variety of peptide substrates that require Hsp70s for general protein translocation and folding. In accordance with this finding, the inability of SSQ1 over-expression to complement an SSC1 null mutation (Schilke et al. 1996) is consistent with the loss of general chaperone function by SSQ1. These results support the conclusion that Ssq1p is no longer a generalist mtHsp70.

SSQ1 is dispensable for yeast survival due to some functional overlap of SSC1. SSQ1 deletion mutants have been observed to accrue iron within the mitochondrial matrix with a concurrent reduction in Fe/S cluster-containing enzyme concentrations and protection against oxidative agents, phenotypes that can be partially rescued by the over-expression of SSC1 (Voisine et al. 2000). Furthermore, because the mechanism by which SSC1 and SSQ1 participate in Fe/S cluster biogenesis seems to require interaction with the same conserved motif of Isu, the Fe/S cluster biogenesis scaffold protein (Schilke et al. 2006), SSC1 is likely to assist Fe/S cluster formation in fungi lacking SSQ1. Because of this overlap, Ssc1p and Ssq1p compete for nucleotide exchange factor Mge1p, which allows ADP and P_i to be released from the mtHsp70s and is present in limiting amounts. The greater abundance of Ssc1p in the mitochondrial matrix, compared to Ssq1p, may limit the amount of Mge1p that can interact with Ssq1p to be recycled to its active form.

As a r

a restr

the Hs

yeasts.

obligat

protein

al. 200

have in

and onc

have be

includir

multifur

dedicati

establis

J-protei

J

family a

required

isoforms

represent

structural

As a result, the reduced proportion of activated Ssq1p may only be sufficient to carry out a restricted task load compared to Ssc1p (Schmidt et al. 2001).

The role of Ssq1p in Fe/S cluster formation is analogous to the specialized task of the Hsp70 HscA in bacteria and it appears that, after arising independently in a subset of yeasts, SSQ1 has undergone functional evolution. In the process, Ssq1p has acquired an obligatory protein interaction with the yeast orthologs of the cluster assembly scaffold protein and the co-chaperone proteins with which the bacterial HscA interacts (Schilke et al. 2006). A specialized Hsp70 committed to Fe/S cluster biogenesis therefore appears to have independently arisen twice throughout the course of evolution- once in the bacteria and once in the yeast. The initial discoveries of SSQ1 in *E. coli* and *S. cerevisiae* seem to have been serendipitous; SSQ1 homologs remain undetected in many eukaryotes, including humans (Schilke et al. 2006). Given that most eukaryotes utilize a multifunctional mtHsp70 in the Fe/S cluster biogenesis pathway, the advantage of dedicating a separate mtHsp70 to assist exclusively in this process in yeasts remains to be established.

J-protein Co-chaperones

J-domain protein co-chaperones belong to the 40 kDa heat shock protein (Hsp40) family and engage in an obligate, physical interaction with Hsp70s. J-proteins are required to stimulate the activity and mediate the function of Hsp70s; thus, J-protein isoforms are active in all cellular compartments containing Hsp70s. While the J-proteins represent a disparate group of proteins with little gene sequence conservation or protein structural organization among members, J-proteins do all share a defining feature called

the J-c

contai

activit

protein

Hsp70

interac

by Sah

the abs

differen

general

domain

Jjj3p w

be rescu

Sahi an

not repl

finger d

compon

exclusiv

chaperon

ribosom

Hsp70 S

of severa

I

the J-domain, named for its sequence similarity to the *E. coli* DnaJ protein. All J-domains contain a histadine-proline-aspartic acid motif essential for stimulation of the ATPase activity of the Hsp70 partner (Cheetham and Caplan 1998). Both general and specialist J-proteins exist in yeast, with several unique J-proteins that function to assist general Hsp70 functions or specialized Hsp70 roles, depending on the specific J-protein/Hsp70 interaction. A distinction between generalist and specialist J-proteins was demonstrated by Sahi and Craig (2007) in *S. cerevisiae* when the deleterious growth effect caused by the absence of J-protein Ydj1p was rescued by expressing J-domain fragments of several different J-protein co-chaperones, indicating that Ydj1p is a generalist J-protein. Such general J-proteins may thus work to indiscriminately stimulate the ATPase functional domain common to all Hsp70s. When specialist J-proteins Cwc23p, Sis1p, Jjj1p, and Jjj3p were deleted from the yeast genome, however, the deleterious phenotype could not be rescued by expression of any other gene. Thus, in contrast to the generalist Ydj1p, Sahi and Craig (2007) showed that the J-domain fragment of specialist Jjj3p alone could not replace the function of full-length specialist J-proteins. For Jjj3p, an additional zinc finger domain was shown to be required for the J-protein's specialized role as a component in the diphthamide biosynthesis pathway. Some specialist J-proteins form an exclusive mtHsp70 partnership to perform a single function, as in the case of a chaperone-co-chaperone pair, Ssz1p and Zuo1p, which associates with translating ribosomes to fold newly synthesized peptides. J-protein Zuo1p interacts solely with the Hsp70 Ssz1p, and Ssz1p does not pair with any other J-protein, despite the co-occurrence of several other types of J-proteins.

In some cases, J-proteins have been shown to bind substrate peptides themselves,

independent of the formation of a complex with an Hsp70. For the *E. coli* DnaJ homolog, a zinc finger-like region and the carboxy-terminal region are required for ligand binding function (Han and Christen 2003). Some J-proteins may deliver substrates to the Hsp70 or recruit the Hsp70 to a peptide when they exhibit a ligand binding function. Similar substrate polypeptide binding features in specialized yeast J-proteins may also act to localize the J-protein to a particular site within the cell, thereby sequestering a J-protein and rendering it unavailable to function in place of other J-proteins, thus conferring specificity (Sahi and Craig 2007).

JAC1: The Mitochondrial J-protein Iron-Sulfur Cluster Specialist

JAC1 is an essential gene and encodes one of 22 J-proteins in the *S. cerevisiae* genome. The Jac1p protein contains an N-terminal mitochondrial signal sequence and is imported into the mitochondrial matrix where it serves as a specialized co-chaperone to assist in Fe/S cluster generation (Voisine et al. 2001). Its task is to bind the Fe/S cluster assembly scaffold protein Isup for delivery to a mtHsp70 and stabilize the Isu1p-Hsp70 interaction (Andrew et al. 2006). Jac1p serves as the only known J-protein capable of interaction with mtHsp70 Ssq1p and together, the J-protein/Hsp70 pair has become specialized in the yeast Fe/S cluster assembly pathway. However, because JAC1 and SSC1 orthologs have been preserved together from bacteria to humans as components of the ISC Fe/S cluster formation pathway, they retain the ability to cooperate in yeast, explaining why the effects of deleting SSQ1 from the *S. cerevisiae* genome may be compensated for by the over-expression of JAC1 (Andrew et al. 2006).

Several pieces of evidence demonstrate that Jac1p and Ssq1p are a functional pair

specialized in Fe/S cluster biogenesis and are consistent with the coevolution of the two proteins. When the mitochondria of mutant JAC1 *S. cerevisiae* strains are isolated and manipulated to contain normal Fe concentrations, a decrease in the activity of Fe/S cluster enzymes was reported. Further, the JAC1 mutation created in this study was found to display a negative genetic interaction with deletion of SSQ1, as these double mutants could not be recovered (Voisine et al. 2001). Additionally, Both Jac1p and Ssq1p have been demonstrated by Andrew and colleagues (2006) to bind the C-terminal domain of Isup. This has revealed that Jac1p and Ssq1p interact with a common component of the Fe/S cluster biogenesis pathway. Importantly, although Jac1p also has the potential to pair with the more abundant mtHsp70 SSC1, Jac1p displays a greater degree of in vitro stimulation of Ssq1p ATPase activity compared with the efficiency of the Jac1p – Ssc1p interaction of both pre- and post mtHsp70 duplication yeasts (Schilke et al. 2006).

Recently, new insight into genetic basis of differences that have evolved at the JAC1 locus and are responsible for the increased efficiency of Ssq1p ATPase stimulation have been elucidated, and involve shortening of the J-domain (Marszalek, unpublished). JAC1 from *S. cerevisiae* was engineered to include a section of the J-domain from the pre-duplication yeast *Y. lipolytica*. The elongated J-domain more closely resembled JAC1 sequences from yeasts encoding Ssc1p, but lacking Ssq1p. The ability of the chimeric protein to stimulate Ssc1p in *S. cerevisiae*, relative to native Jac1p, was increased. Therefore, the portion of the J-domain lost in yeasts encoding Ssq1p may be important for interaction with mtHsp70s and the increased affinity of Jac1p for Ssq1p compared to Ssc1p may have been due to this J-domain modification. The functional specialization of the Jac1p - Ssq1p pair emerged through the sequence of events in evolutionary history

that f

diver

molda

Patte

lineag

genom

molec

the ge

multif

novel a

fate of

retenti

diverge

those d

evoluti

redunda

event ha

paralog

ancestra

ogh par

that followed the duplication of an ancient, multifunctional mtHsp70. Conversely, the divergence of paralogs SSC1 and SSQ1 may have shaped the evolution of JAC1 and molded this J-protein into an Fe/S cluster assembly specialist as well.

Patterns and Mechanisms of Gene Evolution Following Duplication

Because SSQ1 and SSC1 originated from a gene duplication event in a yeast lineage, it is important to understand how the presence of paralogous genes within a genome can affect evolutionary divergence. Gene duplication plays a prominent role in molecular evolution as a mechanism of spawning the genetic material needed to generate the genomic variation responsible for biological diversity. When the ancestral, multifunctional mtHsp70 gene duplicated, a potential was created for the development of novel adaptation unattainable in the single gene copy state. However, the evolutionary fate of gene duplicates depends on two distinct types of mechanisms: 1) one of initial retention within a population and 2) one of several alternative modes of paralog divergence. While many models exist to describe the modes of gene duplicate evolution, those described here have emerged to the forefront of research studies (Hurles 2004).

Neofunctionalization and nonfunctionalization are two models of gene duplicate evolution first put forth by Ohno (1970) to describe the resolution of functionally redundant paralogs. Common to both models is the assumption that a gene duplication event has no effect on organismal fitness because immediately after duplication, the paralogs are equivalent, with each gene copy capable of fulfilling all functions of the ancestral gene equally well. The gene copies are expected to be interchangeable while both paralogs retain high sequence identity, rendering the new duplicate gene immune to

forces

mutatio

loss-of-

gene co

both pr

copy su

exhibiti

duplicat

all func

of delet

or both.

fate of g

longer f

preserva

relaxed

regulato

are thou

novel fu

selective

results i

paralog

forces of selective constraint. Therefore, the duplicate gene is free to accumulate mutations that would have been forbidden in the ancestral single copy state because any loss-of-function that the duplicate gene copy sustains would be rescued by the redundant gene copy. Under this premise, the neofunctionalization and nonfunctionalization models both predict an asymmetry in the evolutionary rates between paralogous genes, with one copy subject to purifying selection to retain ancestral functions and the other copy exhibiting accelerated substitution due to relaxed constraint.

Nonfunctionalization occurs when the period of relaxed constraint on the duplicated gene copy results in the accumulation of deleterious mutations that degenerate all functions of the ancestral gene, without the creation of new functions. Accumulation of deleterious mutations may occur within the protein-coding region, regulatory region, or both, and eventually leads to pseudogenization. This is likely to be the most common fate of gene duplicates (Li 1980). Once a gene has sustained a null mutation and is no longer functional, it is selectively eliminated from the genome and leads to the permanent preservation of the non-mutated paralogs.

Neofunctionalization describes a scenario in which, during the period of initial relaxed selection on the duplicate gene, mutations are acquired in the coding or regulatory sequence that lead to a novel function of the encoded protein. These mutations are thought to be rare, relative to nonfunctionalization. Positive selection to optimize the novel function of the neofunctionalized paralog is then followed by reassertion of selective constraint to preserve the new function. Assuming that neofunctionalization results in the loss of an ancestral gene function, this process too, can lead to non-mutated paralog retention.

tasks

genes.

proces

facilita

subset

distinc

duplica

period

reconst

diverge

novel u

gene co

antagon

opportu

degree

constra

partitio

proceed

been sh

ntilizat

Under a third model of paralog resolution, known as subfunctionalization, the tasks of a multifunctional ancestral gene become partitioned between the two duplicate genes. Duplication-Degeneration-Complementation (DDC) (Force et al. 1999) is one process by which subfunctionalization is thought to occur, where degenerative mutations facilitate the preservation of both paralogs that have become dedicated to complimentary subsets of modular ancestral functions. DDC assumes that the ancestral gene expresses distinct functions ascribed to independent, modular regions of the gene. Following duplication, both paralogs acquire complementary loss-of-function mutations during a period of relaxed selection such that the expression of both paralogs is necessary to reconstitute the repertoire of functions encoded by the ancestral gene.

Escape from adaptive conflict is yet another alternative model of paralog divergence. The premise of this model is that, if an ancestral gene gains an additional novel utility that is in adaptive conflict with the first function, the creation of a duplicate gene could confer an immediate fitness advantage by breaking the ancestral gene free of antagonistic pleiotropy. Through divergent selection, each paralog would have the opportunity to individually specialize in at least one of the ancestral functions to a greater degree than was possible in the ancestral gene. Assuming the ancestral gene was constrained by competing phenotypes conferred by a single gene, the functional partitioning between duplicates or the rise of a novel function after duplication could proceed in a non-neutral manner, driven by an adaptive advantage.

Strong evidence for gene duplicate evolution by escape from adaptive conflict has been shown in the regulatory divergence of paralogs of the *S. cerevisiae* galactose utilization pathway (Hittinger and Carroll 2007). GAL1 and GAL3 arose as duplicates of

a bifu

induc

bifun

promo

transc

to indu

replac

found

these c

GAL3

diverg

promot

maintai

express

sites w

binding

promote

presenc

express

promote

galactok

F

Shape fi

a bifunctional ancestral gene and encode the galactokinase enzyme Gal1p and a co-inducer Gal3p, respectively. While they once shared a common promoter in the bifunctional ancestral gene, near complete subfunctionalization of the upstream promoters between the descendent paralogs has resulted in stringent control of GAL1 transcriptional regulation, contrasting with a more modest GAL3 transcriptional response to induction. The authors swapped the promoter sequences of GAL1 and GAL3, and also replaced native paralog promoters with that of a bifunctional GAL1/GAL3 promoter found in another yeast species, and subsequently evaluated the fitness consequences of these changes. The results from these experiments revealed that switching GAL1 and GAL3 promoters was detrimental, indicating that each promoter had undergone divergence to optimize the expression of GAL1 and GAL3 individually. While the promoter of the bifunctional gene performed well in driving the expression of GAL3 and maintaining yeast fitness, regulation of GAL1 by the bifunctional promoter reduced basal expression and decreased yeast fitness. The spacing of transcriptional activator binding sites was then altered within the promoter sequence of the bifunctional gene to mimic the binding site arrangement of the GAL1 promoter. The manipulated bifunctional gene promoter increased the expression control of the galactokinase function in response to the presence of galactose. Adaptive conflict was therefore proposed to have compromised the expression optimization of galactokinase in the bifunctional ancestral gene with a single promoter. Only after duplication and promoter divergence was the expression of the galactokinase function brought under tighter regulation.

Recently, an additional example of biochemical evidence for gene evolution via escape from adaptive conflict was presented in a study focused on a set of genes involved

in a p

and I

chem

copie

assay

two r

dupli

DFR

dupli

dupli

funct

const

creati

evide

speci

dupli

imme

funda

The i

throu

Kond

of ad

in a pigment biosynthetic pathway in the morning glory, *Ipomeoea purpurea* (Des Marais and Rausher 2008). The dihydroflavonol-4-reductase (DFR) gene, responsible for the chemical reduction of flavonoid precursors of anthocyanin, has given rise to three gene copies, DFR-A, DFR-B, and DFR-C, through two gene duplication events. Biochemical assays for enzymatic reduction of five substrates (three commonly reduced by DFR and two rarely reduced by DFR) by the DFR copies encoded by both pre- and post-duplication species were conducted. Severe reductions in the capacity of post-duplication DFR-A and DFR-C to act on any of the five substrates tested, and an increase in post-duplication DFR-B to reduce all substrates when compared to the activity of pre-duplication DFR enzymes were demonstrated. The authors therefore concluded that the function of the ancestral gene was improved by the DFR-B copy. The release of adaptive constraint, imposed by antagonistic pleiotropy on the ancestral DFR, following the creation of duplicate genes, was consistent with comparative DNA sequence-based evidence of adaptive molecular evolution.

While the above models of paralog evolution offer gene optimization through specialization or the acquisition of novel roles as long-term fitness advantages of gene duplication, short-term benefits must exist to govern the retention of paralogs immediately after gene duplication. The presumed selective neutrality of gene duplication fundamental to Ohno's models fails to offer a short-term benefit of duplication events. The interim retention of duplicate genes on the path to neo- or subfunctionalization through DDC or escape from adaptive conflict also requires a fitness advantage. Kondrashov et al. (2002) have suggested that gene duplication itself may be a mechanism of adaptation by hypothesizing that survival in the face of environmental stresses may

mandate an increase in protein and/or RNA dosage that can be immediately achieved through an increase in gene copy. An environmentally determined optimum copy number may thus exist for each gene under a given set of conditions (Kondrashov et al. 2002). Several studies performed with yeast suggest that environmental conditions may influence gene copy number. For example, when a population of *S. cerevisiae* was experimentally propagated for 450 generations in glucose-limited media, the population evolved the ability to reproduce at a higher cell yield per unit of glucose compared with the ancestral strain (via a glucose transport system with enhanced glucose affinity), resulting from multiple tandem duplications of two hexose transport genes (Brown et al. 1998). The amplification of genes within the Hsp70 gene family may similarly be driven in yeasts as a mechanism to tolerate variations in heat, pH, ethanol, etc., to facilitate the exploration of new environments.

The role that selection plays in determining the evolutionary fates of gene duplicates, from initial retention in the genome to degeneration, neofunctionalization, or subfunctionalization, or other intermediate states of paralog divergence, distinguish the different patterns of gene evolution following duplication described above. Though the previously discussed modes may not be mutually exclusive and no one model of paralog evolution may serve as a general mechanism applicable to all gene duplication events, the ability to characterize the direction and strength of past and present selective forces acting on paralogs are proving to be keys to the elucidation of molecular evolutionary outcomes.

Detectin

A

ultimatel

the resul

within a

via the c

changes

those tha

and trans

are those

mutation

acids to

are pres

assumed

processe

I

site betw

nonsyno

ds. Furth

of the di

number

than the

selection

Detecting Signatures of Selection by Evolutionary Rate Comparisons

At the level of DNA, nucleotide mutations arise in a stochastic manner and ultimately rely on either the forces of natural selection acting on the fitness conferred by the resulting phenotype, or random genetic drift within the population to achieve fixation within a population. Within protein-coding DNA, signatures of selection can be identified via the comparison of the proportion of nonsynonymous to synonymous nucleotide changes that have occurred though time. Nonsynonymous nucleotide substitutions are those that result in the substitution of an amino acid, via a change in both the DNA codon and translated peptide sequence. Synonymous nucleotide substitutions on the other hand, are those that do not alter the amino acid of the corresponding protein. Synonymous mutations exist due to the degeneracy of the genetic code, which allows some amino acids to be specified by several unique nucleotide triplet sequences. Synonymous changes are presumed to be invisible to selection acting on protein phenotypes and are therefore assumed to represent the locus-specific background level of mutations fixed by neutral processes such as population bottleneck events or mutational hitchhiking.

In the context of sequence evolution, the proportion of nucleotide differences per site between two genes that result in nonsynonymous codon changes represents the nonsynonymous substitution rate, d_N, while the synonymous substitution rate is given as d_S. Further, expressing these two rates as the ratio $\omega = d_N/d_S$ can be interpreted as a gauge of the direction and strength of selection. An ω value of less than 1 indicates that the number of mutations resulting in amino acid changes that reach fixation is more restricted than the basal mutation level. Therefore, $\omega < 1$ is indicative of negative or purifying selection, which reduces the rate of fixation of deleterious mutations. When ω is greater

than 1, it may be inferred that positive selection is responsible for the greater-than-expected fixation of amino acid changes. An $\omega = 1$, due to equal rates of nonsynonymous and synonymous nucleotide changes, points to neutral evolution of codons, with amino acid substitutions neither being selected for, nor against. Therefore, the greater the deviation of ω is from 1, the greater the influence of selection. While evidence of negative selection may identify DNA sequence coding for regions of proteins that require strict structural conservation, uncovering signatures of positive selection is of particular importance in the search for evidence of adaptive evolution.

Calculating an average ω for an entire protein-coding gene through pair-wise sequence comparison detects evidence of positive or negative selection throughout an entire gene, evidence found in only a very small proportion of gene sequences. For example, in a large-scale study conducted with 3,595 groups of homologs, comprising 24,832 unique sequences, only 17 gene groups (or 0.45% of the total groups) emerged as candidates of positive selection (Endo et al. 1996). Such estimations of the prevalence and scope of the role of positive selection, however, may be misleading. Since gene-wide mean ω values mask site specific heterogeneity with which natural selection may act, they may not provide an accurate representation of the strength and direction of selective pressures experienced by a gene. Strict interpretations of gene-wide average ω values may overlook the high ω with which a few sites of a gene are evolving, veiled by the low ω values which characterize the evolution of the majority of sites. To bolster this line of reasoning, Yang and Swanson (2002) used several models to estimate the number of codons subjected to positive selection in two gene sequence alignment sets: 192 human class I MHC glycoprotein alleles, and abalone sperm lysin genes from 25 different

spec

neith

abov

appl

signi

gene

'bas

evol

The

to un

diffe

parti

Corr

relev

have

adap

evol

may

inter

and

the

species. For both groups of genes, when ω was averaged across all sites of the sequences, neither the class I MHC glycoproteins, nor the sperm lysin genes yielded an ω value above 1. On the other hand, when a model permitting ω to vary from codon to codon was applied, a number of sites emerged as likely targets of positive selection with ω significantly greater than 1. This result was upheld regardless of whether sites of each gene alignment were partitioned into two groups with different evolutionary rates *a priori* (based on functional information of the encoded structures), or whether sites of different evolutionary rate classes were assumed to be distributed randomly across the sequence. The results of this study highlight the need to account for evolutionary rate heterogeneity to uncover patterns of selection. Several models have been developed to identify differential rates of molecular evolution that can result from selective pressures unique to particular sites and organismal lineages.

Correlated Evolution vs. Co-Adaptation

Differential selective pressures that act on individual sites may be particularly relevant to the detection of coevolution of interacting protein partners. Hakes et al. (2007) have suggested that a distinction must be made between correlated evolution and co-adaptation among protein sites to more specifically describe coevolution. Correlated evolution is the concurrent change among interface residues of interacting proteins that may not necessarily be directly influenced by selective forces due to the protein-protein interaction itself. Co-adaptation, however, is driven by selection to maintain functional and structural integrity of the protein pair to preserve cooperative abilities and results in the compensatory change among interacting protein partners. The compensatory mutation

may be fixed in response to an amino acid substitution in a region of either protein, which serves as the point of contact with the other partner.

Of the proteins investigated by Hakes et al. (2007), an average of only 13% of the total protein sequence was found to correspond to exposed residues directly involved in specific binding activity at the interface of an interacting protein. Patches of proteins that comprise only a minority of residues may experience selective pressure exerted by interacting partners for inter-protein compensatory change. Therefore, correlated increases in the evolutionary rate of whole protein sequences of protein-protein interaction partners do not constitute conclusive evidence for co-adaptation. Instead, correlated coevolution among physically interacting proteins detected by whole gene sequence evolutionary rate analysis may point to other targets of selection unrelated to the interaction of residues at binding surfaces. For instance, gene expression is known to heavily influence the rate of gene evolution (reviewed in Pal et al. 2006). Because cooperative proteins often depend on specific stochiometric ratios of active partners within the cell for efficient interaction, selection for changes in expression of one protein can lead to selective pressure for a corresponding expression change in the other. The resultant expression levels may then be the cause for evolutionary rate changes across the whole protein sequence in both partners, detected as correlated evolution without an underlying adaptation of optimizing inter-protein residue binding. Therefore, evolutionary rate models that account for site-to-site differences are more likely to identify compensatory mutations resulting from co-adaptation.

In addition to acting in a targeted manner within a protein coding gene, adaptive coevolution has been shown to occur in episodic patterns of bursts (Messier and Stewart

1997).

over the

ratios m

informa

history.

may im

selectio

differen

Site-Spe

codon-b

The cod

nucleoti

represen

nucleoti

a sequen

be diffe

simplify

sequenc

such as

the poss

often ge

1997). By effectively averaging any strong, but transient periods of positive selection over the phylogenetic history of two sequences, pair-wise calculation of whole-gene ω ratios may miss evidence for divergent adaptive evolution when lineage-specific information is not taken into account. Events at particular time points in a phylogenetic history, such as environmental changes impacting ecological niches or gene duplication, may impose divergent selective pressures on two protein-coding sequences. Divergent selection is implicated as a cause for ω values of a gene to differ among clades, reflecting different selective pressures influencing different branches of a phylogeny.

Site-Specific Models

Site-specific models define the codon as the unit of evolution and employ a codon-based substitution model to describe site-specific variations in evolutionary rate. The codon substitution model utilizes all of the information encoded within DNA at the nucleotide level, but improves upon the nucleotide substitution model in its representation of molecular evolution by recognizing the amino acids that are encoded as nucleotide triplets. Importantly, considering the amino acid sequence that will result from a sequence of nucleotide codons allows synonymous and nonsynonymous mutations to be differentiated (Goldman and Yang 1994). In employing codon models, several simplifying assumptions must be made. First, the codon model assumes that the DNA sequences under study are protein-coding and does not consider untranslated sequences such as introns. Second, codons which signal translational termination are not included in the possible codons allowed to result from substitution, since these stop codons most often generate a truncated protein and are generally not tolerated within organisms

(Nielse

are ass

mutati

(Goldn

each co

predefi

models

determ

estimat

categor

interpre

conduc

structur

acids pr

do not

When y

class I a

rate clas

among

demons

necessar

(Nielsen and Yang 1998). Lastly, only one of the three nucleotide positions of a codon are assumed to undergo substitution per mutation event (for example, an AGG to CGA mutation would require more than one step under the codon model of evolution) (Goldman and Yang 1994).

The site-specific model of codon substitution assigns a probability with which each codon of a multiple sequence alignment is expected to fall within a particular predefined number of evolutionary rate categories. By conducting this test using nested models of increasing rate categories, the optimum number of rate categories can be determined by statistical tests. The evolutionary rate for each of the rate categories is estimated from the data. Each codon can be assigned to a particular rate class and categorized as evolving under positive or negative selection, and at what magnitude, by interpreting the sign and value of ω.

Maximum likelihood estimation of site-specific rates of evolution can be conducted using fixed-site models or random-sites models. Fixed-site models utilize structural and functional information about a protein of interest to identify specific amino acids predicted to be under equal selective pressures *a priori*, while random-site models do not make any prior assumptions about the evolutionary rate of any particular site. When Yang and Swanson (2002) analyzed the site-specific rates of evolution of MHC class I and sperm lysin genes, the proportion of codons belonging to each evolutionary rate class and the values of ω that were estimated exhibited a high degree of consistency among both fixed- and random-sites models for both gene data sets. The authors demonstrated that partitioning codons into rate classes prior to ω estimation is not necessary; the random-sites model was just as powerful. The residues were classified into

evolutionary

functional or

pressures.

Branch-Spec

Seve

heterogeneit

parameter, w

phylogenetic

multiple line

lineage (with

homogeneou

number of e

model comp

wide ω value

branch-speci

internal bran

may not exis

evolution un

branch-speci

historical ev

source of inc

Yang

evolutionary rate groups corresponding to the groups of residues constituting evolving functional or structural regions of the proteins believed to be under unique selective pressures.

Branch-Specific Models

Several models have also been developed to examine lineage-specific ω value heterogeneity among genes. The simplest lineage-specific model includes only one ω parameter, which assumes the same gene-wide average ω for each branch of a phylogenetic tree. The number of different ω values represented by a gene across multiple lineages may be increased to test whether a gene along one *a priori* identified lineage (with ω_1) is evolving with an overall rate that is significantly different from a homogeneous rate (ω_0) characterizing that gene from all other branches of the tree. The number of estimated branch-specific ω parameters may be increased until maximum model complexity is reached with the "free-ratio" model, in which an independent gene-wide ω value is estimated for each branch of the tree. An important distinction of the branch-specific model compared to the site-specific model is the ability to examine internal branches of a phylogenetic tree. Because known DNA sequence representatives may not exist for internal branches, the ability to detect evidence of ancestral sequence evolution under positive selection makes this model powerful. A test conducted using a branch-specific model allows one to correlate a phylogenetic branch with known historical events, such as ecological shifts or a gene duplication, to hypothesize the source of increased selection.

Yang (1998) used a branch-specific test to demonstrate that a lysozyme gene,

present in t

had a highe

the other ar

Monkey pr

along the b

that the lys

the phyloge

Branch-Site

The

combined t

more specif

particular p

for example

being inves

A "b

may have e

specified *a*

"foreground

branches, w

along the sp

by the null h

when the nu

present in the ancestral primate leading to the divergence of the Hominoid species group, had a higher overall nonsynonymous to synonymous substitution rate ratio compared to the other ancestral and present-day Colobine, Cercopithecine, Hominoid, and New World Monkey primates examined. Furthermore, the average ω of the lysozyme gene inferred along the branch leading to the Hominoids was found to be greater than one, indicating that the lysozyme gene was likely under a divergent positive selection during this time in the phylogenetic history of primates, rejecting a strictly neutral mechanism of evolution.

Branch-Site Models

The principles of site- and branch- specific estimation of ω have also been combined to design models that are used to test gene evolution hypotheses with even more specificity. These methods allow one to gain evidence for hypotheses concerning particular points in evolutionary history. An instance in which ancestral gene duplication, for example, may have given rise to changes in the selective pressures acting on a gene being investigated, could be identified.

A 'branch-site' test allows one to test for the presence of individual codons that may have evolved under positive selection along specified branches. The branch specified *a priori* as that hypothesized to be under positive selection, is denoted the "foreground" branch and is compared to all other branches of the tree, the "background" branches, with respect to site-specific ω distribution. The detection of positive selection along the specified branch relies on the rejection of neutral evolutionary rates predicted by the null hypothesis of a fixed ω = 1 for the gene on the foreground branch. Therefore, when the null hypothesis is rejected, codons are identified along the foreground branch

that exhibit ω both greater than that of the background branch sequences and greater than 1. To increase the rigor with which false positives arise, the subset of positively selected foreground sites are divided among two categories: 1) a class where the ω of background sites is free to vary from $0 < \omega > 1$ and 2) a class where the ω of background sites is fixed at 1. This technique provides a more accurate estimation of the ω of background branches, to which the foreground sites are compared for evidence of positive selection (Zhang et al. 2005).

Clade Models

In addition to the branch-site model, the clade model allows evolutionary rate comparisons to be made simultaneously among codons within a gene and among branches of a phylogeny. The clade model rate test combines patterns of substitution rate heterogeneity across a gene sequence and lineage-dependent rate disparities. However, clade models differ from branch-site models in two important respects: 1) the sequences under analysis must represent at least two clades, defined as a group that includes all of the taxa descended from a common ancestor, a situation described as monophyly, and 2) clade models do not require an $\omega > 1$ to detect a significant difference in evolutionary rates between foreground and background branches. Statistical comparisons of nested clade models can show evolutionary rate accelerations or decelerations that represent a potential increase or relaxation of selective constraint, respectively.

A clade chosen *a priori* is compared with all other clades on the tree with respect to its site-specific ω distribution. These two clades are often called the "foreground" and "background" clades, respectively. Ultimately, individual codons that are evolving at a

different rate in one lineage compared to equivalent codons from another lineage are identified. Individual amino acids may therefore be examined as candidates responsible for functional divergence within a protein.

The clade model was first used to test for divergence in selective pressure between the ε and γ globin genes, paralogs which encode subunits of the hemoglobin oxygen binding protein products in placental mammals (Bielawski and Yang 2003). Following the gene duplication that created the ε and γ globins, selection is thought to be responsible for the divergence in observed expression patterns, leading to delayed, post-embryonic γ globin expression in the simian primate lineage. In contrast, ε globin expression has maintained ancestral gene expression patterns and remains confined to the embryonic life stage of all placental mammals. Under application of the clade model, approximately 16% of the codons common to the ε and γ globins were found to be evolving under divergent selective pressures, with ε globin codons in this rate category evolving under very strong purifying selection ($\omega = 0.008$) and orthologous γ globin codons in the divergent rate category evolving under weak purifying selection ($\omega = 0.79$). The twelve codons that comprised the class of divergently evolving sites among the ε and γ globin clades were subsequently mapped onto three-dimensional globin protein structures to verify that the majority of the encoded residues are part of major structural and functional features of the hemoglobin holoenzyme, one such region being that responsible for oxygen affinity. The authors concluded that, while the majority of globin sites evolve at similar rates when the ε and γ globin clades are compared and display substantial selective constraint, the twelve codons of the divergent ω category are residues likely to have been important for the expression-niche expansion of γ globin to

the fetal developmental stage following gene duplication.

Tools for Evolutionary Rate Analysis

One popular tool that has been developed to model the heterogeneous nature of molecular evolutionary rates is the package of computer programs known collectively as PAML, or Phylogenetic Analysis by Maximum Likelihood. Among other functions, PAML implements maximum likelihood statistical methods in the context of a phylogeny to estimate synonymous and nonsynonymous substitution rates. The estimates can then be used to test hypotheses of site- and lineage-specific ω variation given a sequence alignment and phylogenetic tree topology. Included within PAML is `codeml`, a program that can perform the site-specific, branch-specific, branch-site, and clade model tests. The user inputs a multiple sequence alignment file, a tree topology which describes a hypothesis of evolutionary relationships among the input sequences, and a control file which specifies the model with either initial or fixed parameter values.

A strength of PAML is the ability to optimize parameters that define trends unique to individual data sets of protein-coding sequences through the numerical maximization of the log likelihood value. The likelihood score is indicative of the probability of observing a set of data given a particular model of evolution and phylogenetic tree. Parameters used to describe patterns of sequence change upon which the model and tree are dependent are optimized simultaneously within the likelihood score calculation. Optimized parameters include the transition/transversion rate ratio (κ), and total genetic distance among sequences used to infer branch lengths (t), and nonsynonymous to synonymous substitution rate ratio (ω). Equilibrium codon

frequencies exert an influence on the optimization of κ, t, and ω, and are therefore evaluated by PAML analytically from the sequence alignment.

Recognizing the possibility that multiple local maxima may occur within the distribution of likelihood values (Suzuki and Nei 2001), it is important to allow PAML to optimize parameter values using several different initial parameter input values to ensure that the likelihood space is sufficiently explored. The use of different codon frequency calculation methods is also encouraged to ensure that the parameters are optimized robustly and result in the greatest likelihood score. Ignoring codon bias has been observed to impose an even greater influence on ω estimations than κ, since codon bias is an influential source of unequal substitution rates among codons (Bielawski and Yang 2004b). Additionally, replicate PAML tests should be performed using alternative input tree topologies, if multiple tree topologies exhibit strong statistical support. Because the "true" phylogeny of a set of sequences cannot be known, it is important to show that evolutionary rate analysis results are not dependent on any one tree topology and that test outputs are in agreement with a common conclusion (Bielawski and Yang 2004a).

Outputs obtained from multiple runs can subsequently be compared by their log likelihood scores in a likelihood ratio test, which evaluates the differences between a pair of nested models with different parameters. In this "goodness-of-fit" test, the simpler model represents the null hypothesis. To perform a likelihood ratio test, twice the log likelihood difference between the competing models, defined as the log likelihood test statistic, is first calculated. The log likelihood test statistic is assumed to approximately follow a χ^2 distribution. Therefore, the χ^2 distribution is used to determine an expected value of the log likelihood test statistic, using the number of additional parameters

incorporated into the more complex model relative to the simpler model, as the appropriate degrees of freedom. The null hypothesis is accepted if the log likelihood test statistic falls within the expected distribution (Bielawski and Yang 2004a).

Methodological Limitations

Interpreting the role of selection on a gene through ω estimations of protein-coding regions has the potential to be misleading. For one, the calculation of d_S ignores the cases where a nucleotide substitution that fails to change the encoded amino acid may confer a fitness difference. The value of d_S may therefore be erroneously assumed to be a rate of neutral mutation. For example, biased abundances of iso-accepting tRNAs containing different anticodons, within the cellular pool of tRNAs, may result in differential translational efficiency of sequences containing different nucleotide triplets for the same amino acid. Synonymous substitutions may also violate the assumption of neutrality when a nucleotide is shared between genes, as in the case of genetic material of many viruses (Diamond et al. 1989) for which the mutation is nonsynonymous for an overlapping reading frame. Moreover, nucleotide changes may affect the stability of DNA or RNA molecules if the substitution results in disruption of secondary structure through elimination of a crucial hydrogen bond. Hammerhead ribozymes, for instance, rely on stem-loop features for recognition, binding, and subsequent cleavage of substrates (Tuschl and Eckstein 1993).

In addition, the alignment of DNA and amino acid sequences is implied to be error-free, such that each nucleotide within a 'column' corresponds to the same codon position of all other genes. However, the "true" alignment of a group of sequences is

unknown, and even computer programs using sophisticated algorithms to align sequences can only make an inference of sequence relationships. Assessment of simulated DNA sequence data alignments has shown that the reliability of computer-generated alignments for correctly recognizing homologous sites decreases when the length of sequences that contain insertions and deletions is increased (Nuin et al. 2006). Similarly, the estimation of site or lineage-specific ω values relies on the topology of a cladogram which serves as a description of the ancestral origins and relationships among the sequences in question. However, cladograms represent *hypothesized* phylogenetic relationships; the true phylogenetic history of a set of gene sequences can never be known with certainty.

Furthermore, factors other than positive selection can cause an $\omega > 1$. For instance, the severe reduction in population size caused by a population bottleneck can decrease the effectiveness of purifying selection, allowing deleterious mutations that would otherwise be eliminated, to rise to fixation and oppose selection via drift. In some cases, the random nature of mutation may result in the absence of synonymous substitutions. Thus, a codon may show $\omega > 1$ simply due to the stochastic nature of mutation. Likelihood tests of evolutionary rate heterogeneity do not yet allow such alternative explanations for $\omega > 1$ to be statistically considered (reviewed in Hughes, 2007).

Finally, and perhaps most importantly, recovering molecular signatures indicative of the direction and intensity of selection are not adequate to make conclusions about the phenotypes and subsequent fitness effects of observed mutations. Instead, evolutionary rate analysis should be used as a springboard for the formulation of hypotheses that may directly (i.e. biochemically, at the molecular level) investigate the fitness costs and benefits to organisms conferred by the products of genes evolving at elevated or

decelerated rates. Ultimately, the goal of this line of research should be to seek the ecological origins for evolutionary forces that lead to adaptation.

SECTION II: EXPERIMENTAL STUDY

HYPOTHESES AND PREDICTIONS

We sought to investigate the evolutionary patterns of the mitochondrial heat shock proteins involved in Fe/S cluster biogenesis: the paralogous genes SSC1 and SSQ1, plus their interacting J-protein partner, JAC1. Motivation for this study comes from the observation that SSQ1 represents an example of a heat shock protein that has become specialized in a particular sub-function of its ancestral gene, and interestingly, one unusual to chaperones. In this study, we analyzed sequences of monophyletic fungal groups of comparable within-clade relatedness, two of which diverged from a common ancestor prior to the gene duplication that created SSQ1 (*Aspergillus* and *Fusarium*), and two of which diverged after the duplication event (*Saccharomyces* and *Candida*) (see Figure 2). The presence of JAC1 within each of these clades has given us the opportunity to investigate how the duplication of the ancestral mtHsp70 has influenced the evolutionary paths of SSC1, SSQ1, and JAC1, via extensive comparative analyses of the rate of gene sequence evolution.

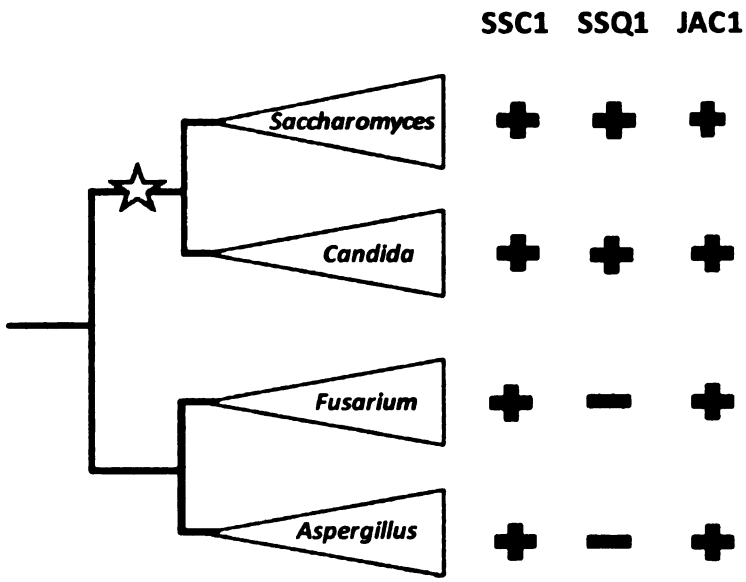


Figure 2: Summary of mitochondrial heat shock protein (mtHsp) distribution among fungal clades. The monophyletic fungal groups examined were the *Saccharomyces*, *Candida*, *Fusarium*, and *Aspergillus* clades. All four clades encode the gene for the multi-functional mtHsp70 SSC1, as well as the interacting mitochondrial J-protein co-chaperone encoded by JAC1 present in all four clades. The "+" and "-" symbols represent the presence or absence of a gene within a clade, respectively. In the lineage indicated by the star, prior to the divergence of the *Saccharomyces* and *Candida* clades, the ancestral mtHsp70 underwent a gene duplication, which gave rise to the gene for the specialized mtHsp70 SSQ1 carried in *Saccharomyces* and *Candida* taxa.

Our objective was to elucidate molecular patterns of selection to test the following hypotheses:

Hypothesis 1: Selective constraint has been relaxed in SSC1 in the presence of its paralog, SSQ1.

H_0: The rate of SSC1 evolution in clades encoding the paralogs SSQ1 is equal to the rate of SSC1 evolution in clades that lack SSQ1

Inability to reject the null would suggest that SSC1 and SSQ1 paralogs are equivalent and therefore, functionally redundant. This outcome seems unlikely, given that evidence indicates SSC1 and SSQ1 cannot replace one another and rules out the functional equivalence of the encoded proteins.

H_1: The rate of SSC1 evolution in clades encoding the paralog SSQ1 is not equal to the rate of SSC1 evolution in clades that lack SSQ1

Evidence to support this hypothesis would be consistent with a relaxation of selective constraint following duplication of the ancestral mtHsp70 gene if SSC1 is evolving at a faster rate than SSQ1. Our *a priori* prediction is that the rate of SSC1 evolution will be elevated in clades encoding SSQ1, versus clades that lack SSQ1. Biochemical evidence for the increased affinity displayed by Jac1p for Ssq1p suggests that Ssq1p may be capable of fulfilling the role of Ssc1p in Fe/S cluster biogenesis. Thus, SSQ1 may have the ability to compensate any loss-of-function mutations affecting SSC1 at sites important for Fe/S cluster biogenesis. SSQ1 would negate the need for SSC1 to maintain sites used for the Fe/S cluster assembly pathway, allowing a greater proportion of mutations to be fixed at these sites in SSC1.

Alternatively, the rate of SSC1 evolution in clades encoding SSQ1 could be decreased relative to the rate of SSC1 evolution in clades that lack SSQ1. Evidence for this result would be consistent with an increase in selective constraint on SSC1 when co-occurring with SSQ1. It is difficult to identify possible sources of increased constraint on SSC1, given that evidence does not exist to suggest that SSC1 has attained a novel function or adaptive peak since the mtHsp70 duplication event.

Hypothesis 2: SSQ1 is under less selective constraint than SSC1 because SSQ1 has fewer encoded functions to maintain.

H_0: SSQ1 and SSC1 evolve at equal rates

Inability to reject the null would be consistent with the conclusion that SSQ1 and SSC1 are not under current divergent selective pressures. One possible explanation for

this result could be that, while functional divergence of Ssq1p and Ssc1p occurred in an ancestral lineage, current selective pressures in extant taxa are now acting with the same direction and magnitude on each paralog. However, the functionally divergent paralogs interact with different groups of substrates and therefore have different sources of possible coevolutionary influence. Thus, it seems unlikely that SSC1 and SSQ1 would be evolving at equal rates.

H_1: SSQ1 and SSC1 evolve at unequal rates

Evidence to support the alternative hypothesis would be consistent with the functional specialization of Ssq1p in Fe/S cluster biogenesis if our *a priori* prediction, that SSQ1 is evolving at an elevated rate compared to SSC1, is observed. According to biochemical experiments, Jac1p stimulates the ATPase activity of Ssq1p to a greater extent than Ssc1p of both pre- and post- mtHsp70 duplication species. Therefore, mutations must have been fixed in SSQ1 since the time of the duplication event to afford functional distinction from SSC1. Additionally, Ssq1p has a diminished functional repertoire. Therefore, SSQ1 must have undergone fixation of mutations that result in loss of function. On the other hand, Ssc1p has not been shown to have gained any novel functions subsequent to the creation of SSQ1. Because Fe/S cluster biogenesis constitutes only one of many roles encoded by SSC1, any loss of performance in Fe/S cluster assembly sustained by SSC1 is predicted to have occurred with a small number of mutations. The number of mutations likely to have occurred in SSQ1, to degenerate the many lost roles in protein folding and translocation, would be comparably large. Therefore, a greater number of mutations are likely to have occurred in SSQ1 than SSC1 since the time of duplication.

Alternatively, SSQ1 could be evolving at a decreased rate compared to SSC1. An increase in selective constraint on SSQ1 would be a possible explanation for this result. However, because fewer functions have been ascribed to Ssq1, relative to Ssc1p, this outcome would support the need to further investigate the functions of Ssq1 to identify additional sources of constraint that could be acting on SSQ1 compared to SSC1.

Hypothesis 3: The rate of JAC1 evolution is positively correlated with the rate of SSQ1 evolution because JAC1 and SSQ1 are coevolving.

H_0: The rate of JAC1 evolution in clades encoding SSQ1 is equal to the rate of JAC1 evolution in clades that lack SSQ1

Inability to reject the null would be consistent with the absence of an influence by SSQ1 on the direction and magnitude of selection acting on JAC1. This outcome does not seem likely, given that Jac1p has been demonstrated to result in different magnitudes of ATPase stimulation for Ssq1p and Ssc1p.Therefore, the selective pressures exerted by Ssq1p and Ssc1p on Jac1p are probably not equivalent. Alternatively, the increased efficiency of the Jac1p – Ssq1 interaction could be due to changes at very few sites in JAC1, or entirely independent of JAC1 evolution, resulting from the specialization of Ssq1 alone.

H_1: The rate of JAC1 evolution in clades encoding SSQ1 is not equal to the rate of JAC1 evolution in clades that lack SSQ1

Evidence to support this hypothesis would be consistent with the coevolution of JAC1 with a duplicate mtHsp70 under increased or decreased selective constraint relative to the ancestral pre-duplicate mtHsp70. Our *a priori* prediction is that JAC1 evolves at an increased rate in the presence of SSQ1, compared to JAC1 from clades that lack SSQ1. Because Jac1p must stimulate the ATPase activity of an mtHsp70, evolution of JAC1

would be necessary to accommodate any changes in a mtHsp70 that might hinder the ability of Jac1p to physically interact with the mtHsp70. Molecular coevolution of JAC1 with SSQ1 could account for the specialized interaction that has given rise to the ability of Jac1p to stimulate Ssq1p to a greater extent than Ssc1p. Furthermore, because SSQ1 is a duplicate gene, it is expected to be under relaxed selective constraint compared to mtHsp70s in the single gene state. Therefore, if SSQ1 is evolving at a faster rate, the rate of JAC1 evolution would be expected to accelerate when co-occurring with SSQ1 to maintain the ability to physically interact. This assumes that the faster rate of SSQ1 evolution is due to changes at sites critical to interaction with JAC1. Regardless of whether SSQ1 – JAC1 coevolution was instigated by initial changes in SSQ1 or JAC1, the exertion of reciprocal selective pressures could result in correlated rate acceleration of SSQ1 and JAC1. Thus, JAC1 would be observed to evolve faster in clades encoding SSQ1 compared to clades lacking SSQ1. Conversely, if the evolutionary rate of SSQ1 is observed to be slower than that of SSC1, we predict the rate of JAC1 evolution will be decelerated in the presence of SSQ1.

A negative correlation between JAC1 and SSQ1 evolution would be indicative of antagonistic coevolution. A coevolutionary relationship of this nature could result if either JAC1 or SSQ1 constrain the evolution of the other, such as if the proteins had reached an adaptive peak in their interaction. Alternatively, another factor (perhaps an unidentified component of the Fe/S cluster biosynthesis pathway) could increase selective constraint on JAC1 or SSQ1, while releasing constraint on the other.

Hypothesis 4: SSQ1 has undergone adaptive evolution to optimize the Ssq1 - Jac1p interaction important for Fe/S cluster biogenesis.

H_0: SSQ1 has not evolved under positive selection

confer

relaxat

the div

at sites

selecti

degene

could i

fitness

within

H_1: SS

evolut

functi

in the

assem

benefi

betwe

results

then p

Inability to reject the null would be consistent with the fixation of mutations that confer the functional differences between Ssq1p and Ssc1p to have occurred by a relaxation of selection and/or genetic drift. A possible evolutionary history to account for the divergence of Ssq1p without positive selection would include relaxation of constraint at sites required for protein folding, translocation, and stress responses. A relaxation of selective constraint at those sites would allow deleterious mutations to accumulate and degenerate the encoded functions. The increased ATPase activity in the presence of Jac1p could have arisen in SSQ1 due to the random fixation of beneficial mutations with weak fitness effects. Alternatively, the increased ATPase activity could be due to evolution within JAC1 alone.

H_1: SSQ1 has evolved under positive selection

Evidence to support this hypothesis would be consistent with a period of adaptive evolution in the history of SSQ1. The premise for this proposal is that SSQ1 has become functionally specialized since its divergence from SSC1. SSQ1 shows increased activity in the presence of JAC1, an improvement of ancestral function important for Fe/S cluster assembly. To improve upon the ancestral function, SSQ1 must have acquired mutations beneficial to the Jac1p–Ssq1p interaction, potentially at sites critical to physical contact between the two proteins. If increased efficiency of Ssq1p ATPase stimulation by Jac1p results in an increase in adaptive fitness, perhaps by improving Fe/S cluster biogenesis, then positive selection could drive the new SSQ1 allele to fixation.

METHODS

Fungal Taxa and Gene Sequence Alignments

Gene sequences were retrieved from seven to eight fungal species, from each of four monophyletic clades. The taxa from two of these clades, the *Saccharomyces* and *Candida* groups, encode the duplicate Hsp70, SSQ1, while the *Aspergillus* and *Fusarium* clades lack SSQ1. SSC1, SSQ1, and JAC1 coding region sequences (exons only) were taken from *Saccharomyces cerevisiae* RM111, *Saccharomyces cerevisiae* YJM789, *Saccharomyces paradoxus*, *Saccharomyces mikatae*, *Saccharomyces bayanus*, *Saccharomyces castellii*, and *Candida glabrata* genomes, which comprise the *Saccharomyces* clade, and from *Candida lusitaniae*, *Candida guilliermondii*, *Debaryomyces hansenii*, *Candida parapsilosis*, *Candida tropicalis*, *Candida dubliniensis*, and *Candida albicans* genomes, which comprise the *Candida* clade. SSC1 and JAC1 sequences from the *Aspergillus nidulans*, *Aspergillus niger*, *Aspergillus terreus*, *Aspergillus oryzae*, *Aspergillus flavus*, *Aspergillus clavatus*, *Aspergillus fumigatus*, and *Neosartorya fischeri* genomes comprise the *Aspergillus* clade, while sequences from the *Podospora anserina*, *Trichoderma reesei*, *Fusarium solani*, *Fusarium graminearum*, *Fusarium verticilliodes*, and *Fusarium oxysporum* genomes comprise the *Fusarium* clade. The complete genome of each of the above fungal species has been sequenced and has been made available through sequence databases curated by the Saccharomyces Genome Database, the BROAD Institute, the Joint Genome Institute, the Wellcome Trust Sanger Institute, and Génoscope. Nucleotide and protein sequence BLAST searches were performed to identify orthologs, with reciprocal best BLAST hits used to confirm

orthology and reject paralogy of SSC1 and SSQ1 sequences. The sources and genome coordinates of all sequences are presented in Appendix A.

Because JAC1 is a fast evolving gene and differs by more than 80% at the nucleotide level between fungal clades, JAC1 is too divergent to confidently generate a multiple alignment of JAC1 sequences from all four fungal clades. Therefore, it was necessary to carry out JAC1 sequence alignments, subsequent gene tree construction, and rate analyses, separately for each of the four fungal clades. However, a similar average JAC1 sequence divergence and number of taxa for each clade facilitates comparison of JAC1 sequences among the clades (see Figure 3). Within each clade, any two JAC1 nucleotide sequences differ by approximately 25% to 35%.

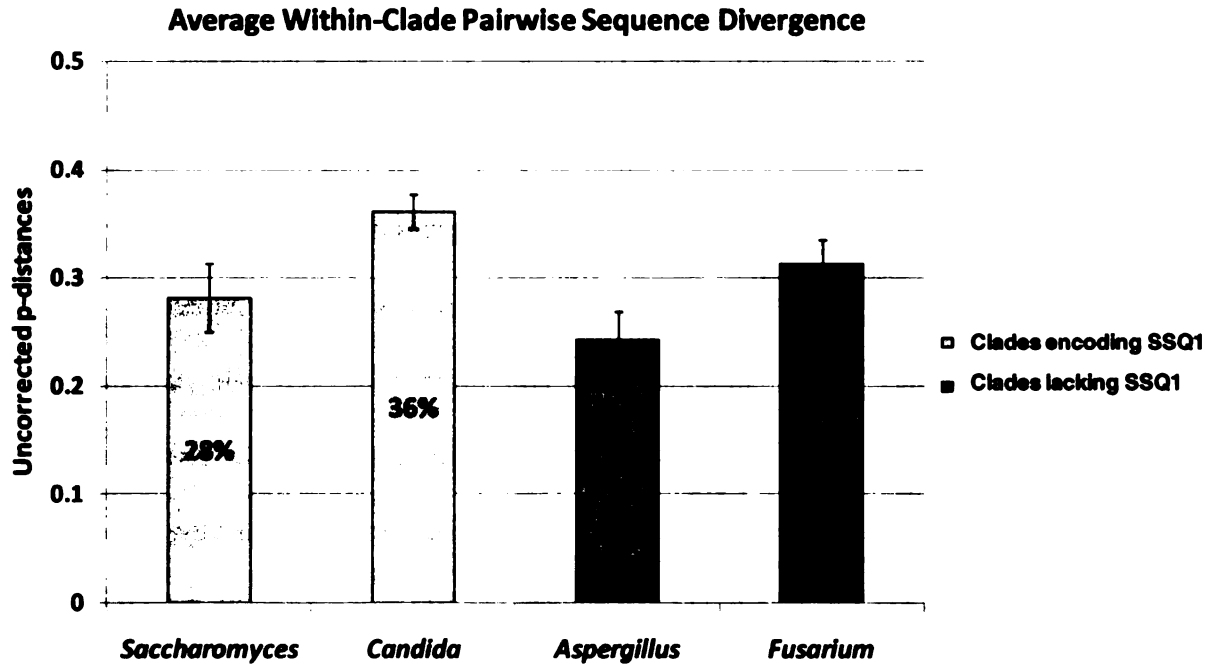


Figure 3: Average within-clade pair-wise sequence divergence of JAC1. The average JAC1 divergence, calculated as the uncorrected p-distance, between any two JAC1 sequences encoded by taxa belonging to the same clade is shown. P-distances are expressed as the percent sequence dissimilarity. Error bars represent standard deviations.

Multiple alignments of translated amino acid sequences were performed using CLUSTAL W (Thompson et al. 1994), with default gap penalties and subsequent manual trimming to remove gaps. The JAC1 alignment of each clade contains the following number of amino acids: *Saccharomyces*: 177, *Candida*: 167, *Fusarium*: 189, and *Aspergillus*: 185. In contrast, the more conserved nature of the mtHsp70 genes permitted alignment of sequences from all taxa. The SSC1 alignment includes 603 amino acid sites from all four fungal clades, and the combined SSC1 and SSQ1 alignment includes 580 amino acid sites from SSC1 of all four fungal clades and SSQ1 from the *Saccharomyces* and *Candida* clades. All amino acid alignments are shown in Appendix B.

Cladogram Construction for PAML Input Trees

Data Partitioning

Figure 4 depicts a graphical summary of gene tree construction. Analysis of separate partitions of data with independent evolutionary models has been demonstrated to fit heterogeneous data better when compared to un-partitioned data. Further, data partitioning may also yield support for alternative tree topologies (DeBry 1999). Analysis of partitioned sequence data is a technique used to accommodate evolutionary heterogeneity within subsets of the sequences. The first and second nucleotide positions of codons within protein coding regions are expected to evolve at a slower rate than third positions, due to the fact that most substitutions at third positions are synonymous. At first positions, however, most substitutions are nonsynonymous, and all substitutions are nonsynonymous at second positions. Therefore, selective constraint is expected to be

weakest for third positions, of intermediate strength at first positions, and strongest for second positions. Thus, the fastest rate of change is expected to take place at third positions and result in greater ability to resolve phylogenetic relationships among closely related or slowly evolving sequences. In such cases where few sequence changes are expected to have accumulated among taxa, first and second positions may not contain sufficient variation to resolve evolutionary histories. First and second positions are often more useful in resolving deep branches of a phylogenetic tree, where the sequences in present-day taxa may be very divergent. Given greater sequence divergence, the chance increases for third positions to become saturated with homoplasies, at which point these nucleotides no longer provide a reliable signal to distinguish basal relationships.

Tree construction of translated amino acid sequences is another method to achieve robust branch resolution, given evolutionary rate variation within genes. A model used to describe patterns of amino acid substitution may be more appropriate than models that use DNA units of evolution, and is useful to complement results of cladogram construction using nucleotide substitution models. While information held within DNA is lost when sequences are examined at the amino acid level due to the degenerate nature of the genetic code, modeling amino acid substitutions releases analyses from biases in nucleotide base composition and mutation more prevalent in nucleotide sequences. For example, unlike peptides, nucleotide evolution is often influenced by structural constraints that favor a particular nucleotide sequence for hairpin or loop regulatory features that result when the DNA is transcribed into RNA. Selection for codon bias also falls into the category of nucleotide compositional bias. Moreover, far more character types make up peptide sequences compared to DNA sequences (there are more types of

amino acids than nucleotide bases), therefore making amino acids less prone to mutations that revert a site back to its ancestral state. Additionally, because amino acid substitutions often require more than one nucleotide substitution, the rate of amino acid evolution is slower than that of nucleotides. Together, the reduced homoplasy and slower evolutionary rate observed at the amino acid level confers the advantage of better phylogenetic resolution of distantly related taxa or fast-evolving genes than might be possible by nucleotides.

Therefore, different subsets of the sequence data were considered here individually. Unrooted gene trees were constructed using the following partitions: first, second, third, first with second, and all nucleotide positions of codons, as well as amino acids.

Maximum Parsimony

Constructing the phylogenetic tree topologies to be used in the estimation of evolutionary rates is a critical initial step that can be accomplished using several different methods of inference. Ideally, given a set of properly aligned sequences, the inferred phylogeny would be identical, regardless of the method used to construct the tree, if the "true" evolutionary history is to be accurately represented. In practice, however, each method of phylogenetic tree construction possesses unique strengths and pitfalls, and therefore, can influence the outcome of phylogenetic analyses. For this reason, it is prudent use more than one method in parallel, and, given a sequence data set, to examine alternative trees in subsequent analyses when possible. Maximum parsimony (MP), maximum likelihood (ML), and the Bayesian inference (BI) methods were used in this

study.

The MP method of phylogenetic inference is a character-based method that seeks to recover tree topologies that minimize the number of evolutionary transitions necessary to explain the distribution of characters among taxa (Hennig 1966). A tree search algorithm is used to evaluate tree topologies according to the minimum number of steps required. The occurrence of convergent evolution, parallel evolution, or character reversals to the ancestral state, may cause two sequences to appear more closely related than they actually are. These are sources of homoplasy; the opportunity for their occurrence increases with the time since divergence from a common ancestor and are assumed to be minimized in the most parsimonious tree. However, MP tree construction has a tendency to erroneously group highly divergent sequences together, particularly when the sequences are distantly related or have undergone very rapid evolution. This problem is known as long-branch attraction (Felsenstein 1978).

MP trees were constructed in PAUP* v 4.0b10 (Swofford 2000), with a heuristic search using the tree-bisection-reconnection (TBR) branch-swapping algorithm and equal weighting for all characters. The TBR method of tree searching starts with an initial tree topology, breaks the tree into two sub-trees, and then reconnects the halves at all possible nodes. Here, the initial topology was generated by the random, stepwise addition of sequences and heuristic tree search proceeded by random addition sequence replications. One hundred bootstrap replacement replicates were performed to determine statistical support for branches of each topology.

Maximum Likelihood

Another commonly used method for inferring phylogenies is the MLmethod introduced by Felsenstein (Felsenstein 1981). ML tree construction can use many different models of sequence substitution in conjunction with the powerful statistical inference of optimizing a likelihood function. This allows the ML method to more efficiently distinguish homoplasy from synapomorphy, an advantage that provides greater accuracy of phylogenetic inference of very divergent taxa or sequences with very different rates of evolution, compared to the MP method. The ML method examines all possible pathways of sequence change possible for a given data set in order to identify the hypothesis most likely for the data. Within the likelihood calculation used to evaluate hypotheses, the tree topology, branch lengths, and evolutionary model components are simultaneously optimized. When these parameters have been optimized to maximize the likelihood, the best evolutionary model and tree have been found (according to ML). This is analogous to reaching a peak in a multi-dimensional parameter landscape. Parameter values at the peak reached in the parameter space are estimated from the data, and therefore do not need to be specified *a priori* by the investigator before examining the data (Holder and Lewis 2003). However, ML method calculations are computationally intensive and may propose an incorrect evolutionary relationship if an inappropriate substitution model is chosen (Huelsenbeck and Crandall, 1997).

Maximum likelihood trees were inferred using PhyML v2.4.4 (Guindon and Gascuel 2003) by applying the general time reversible (GTR) model of nucleotide substitution. The GTR model estimates an independent frequency with which each nucleotide base is observed within a set of sequences and an independent substitution rate

for each pair of nucleotide substitutions. Additionally, each substitution type is assumed to be equally reversible to allow, for instance, G → T and T → G to occur at equal rates. Furthermore, parameters such as the proportion of invariant sites and the gamma shape parameter, used to describe the distribution of substitution rates among sites, account for site-to-site evolutionary patterns. All parameters for the model were estimated from the data, with four discrete categories in the gamma rate distribution. The amino acid model of substitution indicated as the best model for protein evolution by ProtTest v1.4 (Abascal et al. 2005), according to a likelihood ratio test, was specified for each multiple amino acid sequence alignment as follows: *Saccharomyces* JAC1: WAG, *Candida* JAC1: RtREV, *Fusarium* JAC1: WAG, *Aspergillus* JAC1: JTT, SSC1: RtREV, SSC1 and SSQ1 combined: RtREV.

A neighbor-joining tree was generated in PhyML to serve as the starting tree in the tree search. A hill-climbing algorithm was then used to optimize the maximum likelihood. One hundred bootstrap replicates were performed to determine statistical robustness of trees and yielded a bootstrap consensus tree used to assess clade support for both MP and ML methods.

Bayesian Inference

The BI method of phylogenetic reconstruction resembles the ML method in that the BI method can incorporate many different molecular evolutionary models in the search for the best tree. However, the BI method samples from the posterior probability distribution to identify the most probable phylogenetic tree, given a data set. This requires an investigator to assign prior probabilities for all parameters, i.e. predictions

made before examination of the data, a potential source of bias that some consider a disadvantage of the BI method (Felsenstein 2003). In phylogenetic analysis, prior probabilities are usually given an uninformative or "flat" distribution, to regard all possible trees as equal hypotheses until the data are examined.

While the BI method evaluates the likelihood of a hypothesis to calculate the posterior probability, parameters are not optimized as in ML. Instead, the Markov Chain Monte Carlo (MCMC) algorithm is used to estimate the probability distribution of a hypothesis. This algorithm constructs chains to move from one location to the next within a multi-dimensional space of hypotheses, periodically sampling the posterior probability, and moving toward successively greater probability densities. Each "link" within the chain, or location within the tree space, is termed a "generation." The goal is to reach an equilibrium posterior probability distribution, at which time a move to a new location within the tree space does not yield a greater posterior probability. Separate chains running in parallel converge at similar posterior probability values. To avoid becoming stuck in local regions of high posterior probability density and allow more efficient exploration of the hypothesis landscape, the tree and evolutionary model evaluated at a location by one chain may be periodically swapped between other parallel chains (reviewed in Holder and Lewis, 2003). Heated chains are freer to traverse peaks and valleys in the landscape, and are thus useful when posterior probabilities are swapped among parallel chains. The cold chain is more restricted in its movement and is the chain from which sampled posterior probabilities are used as the output for a run. Because the initial locations of the chains (termed the "burn-in") in the tree space is often far from the greatest posterior probability density, a proportion of the first generations are discarded

from the final evaluation of posterior probability distributions.

Bayesian inference trees were constructed in MrBayes v3.1.2 (Ronquist and Huelsenbeck 2003) using the same nucleotide substitution model as described for ML tree construction of nucleotide sequences and mixed model optimization for amino acids. The default assumption of flat prior probability densities was implemented for all parameters. Two parallel Markov chain Monte Carlo processes were initiated, consisting of three hot chains and one cold chain. The chains were run for 1,000,000 generations each, with a sampling frequency of once per 100 generations. The initial 2,500 trees were discarded as the burn-in. Chain parameter and tree convergence within one run and between parallel runs was assessed by likelihood scores. When the likelihood scores of the cold chains were no longer increasing and showed fluctuation within a narrow range, the chain was assumed to have reached stationarity within the parameter space. In addition, plots of generation versus the log posterior probability were also generated for each run to visually detect stationarity via absence of increasing or decreasing posterior probability value trends.

Constructing Composite Input Tree Topologies

For each method of tree construction, the most parsimonious or most likely (as appropriate to the tree method) trees were visually examined for branch resolution on bootstrap consensus trees. Note that in instances where more than one tree topology was returned as the most parsimonious tree by the MP method, computation of the bootstrap consensus negated the need to examine multiple MP trees for each data partition. Bootstraps of $\geq 90\%$ or posterior probability values of ≥ 0.9 were considered sufficiently

well supported. In cases where branch resolution could not be achieved using one data partition, but could with another, branches were manually inserted to produce the best resolved, composite tree. In instances where evolutionary relationships among sequences could not be resolved with high statistical support by any combination of sequence partitions, tree branches were collapsed into polytomies. The number of unique composite tree topologies obtained by each phylogenetic inference method for each gene alignment are as follows: *Saccharomyces* JAC1: 2, *Candida* JAC1: 3, *Fusarium* JAC1: 1, *Aspergillus* JAC1: 3, SSC1: 11, SSC1 and SSQ1 combined: 9. All composite tree topologies used as input trees for `codeml` are displayed in Appendix C.

Figur
rate a
three
Likel
were
the 1
amin
phyl
stron
the F

mtH

SSC

Perf

evol

and

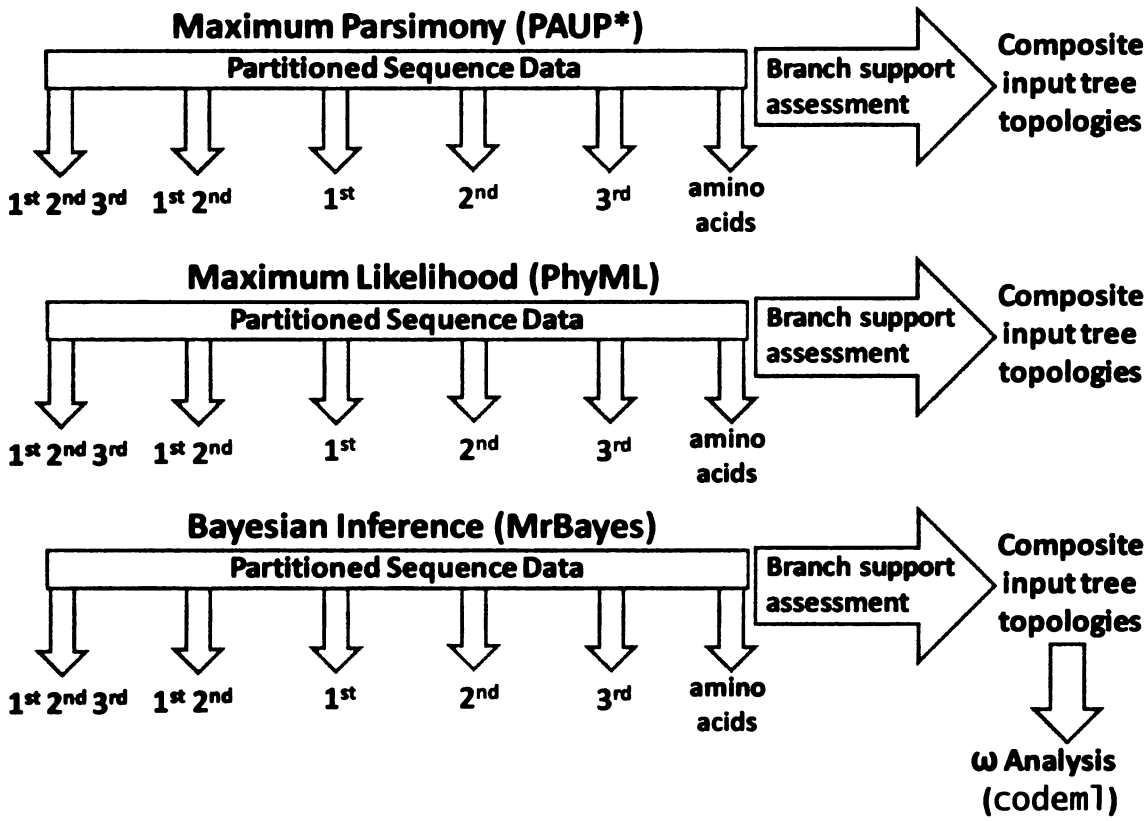


Figure 4: Summary of data partitions and phylogenetic tree construction for evolutionary rate analysis. JAC1, SSC1, and combined SSC1 and SSQ1 gene trees were inferred using three different methods of tree construction: Maximum Parsimony, Maximum Likelihood, and Bayesian Inference. Five different partitions of each sequence alignment were used individually for tree construction: all three nucleotide bases of a triplet codon, the 1st and 2nd nucleotide positions only, each nucleotide position individually, and the amino acid residues. Each tree resulting from each data partition analyzed by the three phylogenetic methods was assessed for branch support to generate all possible unique, strongly supported composite topologies for ω analysis using the `codeml` program of the PAML package.

mtHsp70 Clade Model Rate Comparisons

To investigate the potential role selective constraint has played in the evolution of SSC1 since the inception of SSQ1, as stated in **Hypothesis 1**, a clade model test was performed to examine the influence of the presence of SSQ1 on the rate of SSC1 evolution. The rate of SSC1 codon evolution from taxa possessing SSQ1 (*Saccharomyces* and *Candida* clades) was compared to the rate of SSC1 codon evolution from taxa

lacking SSQ1 (*Aspergillus* and *Fusarium* clades) via application of the clade model rate test. The foreground clade (SSC1 from taxa co-occurring with SSQ1) was distinguished from the background clade (SSC1 from taxa lacking SSQ1) at the node representing the most recent common ancestor of the *Aspergillus* and *Fusarium* sub-clades for each of the eleven input trees (see Figure 5A). Due to a program glitch that we found in the application of model=3 in PAML v. 4.0 (Bielawksi 2008), clade model analyses were carried out in version 3.0 of PAML (Yang 1997).

Tests were conducted under three different clade models that varied in the number of pre-defined rate categories. The null model had one rate category, while the two alternative models were specified to have either two or three rate categories. Likelihood ratio tests were then performed to determine the most appropriate rate test model. Likelihood ratio tests comparing models are shown in Appendix E. For all clade models, initial ω and κ values of 0.5, 1.0, and 1.5 were tested. In addition, two different methods of codon frequency adjustment were applied in all `codeml` tests: 1) codon frequency model F3x4, where 1st, 2nd, and 3rd base frequencies from the data were used to estimate codon frequencies, and 2) a table of codon frequencies observed within the data. Because varying initial ω and κ values and codon frequency models does not alter the number of parameters used by the model, the effect of altering these settings cannot be determined by a likelihood ratio test. However, clade tests conducted with initial values set to 0.5 and 1.0 tended to give slightly higher likelihood scores than with initial values set to 1.5. Initial values of 0.5 and 1.0 gave very similar, and often identical, likelihood scores. Use of the observed codon frequency table always resulted in the highest likelihood scores of all `codeml` tests. The output that yielded the highest likelihood score is reported in the

RESULTS.

To investigate the possibility that SSQ1 is under weaker selective constraint than SSC1, as stated in **Hypothesis 2**, clade model rate analyses were also performed as described above to compare the rate of SSQ1 sequence evolution to that of SSC1. Cladograms constructed for all taxa in a single tree, with gene sequences from both mtHsp70s, were divided into a foreground clade of SSQ1 and background clade of SSC1 (see Figure 5B).

Site-Specific Rate Tests

To investigate the potential role that the mtHsp70 gene duplication played in altering the rate of JAC1 evolution in the presence of SSQ1, as outlined by **Hypothesis 3**, JAC1 site-specific rate tests were conducted. JAC1 sequences were separately evaluated for each of four fungal clades using a site-specific model of gene evolution applied in the `codeml` program of version 4.0 of PAML (Yang 2007). The site-specific rate model is used to estimate ω values for a pre-defined number of rate categories and, subsequently, each codon is assigned to the most likely category. We used this test to look for evidence of increased or decreased selective constraint acting on individual amino acids in sequences derived from clades in which JAC1 co-occurs with both Hsp70 paralogs, SSC1 and SSQ1, compared to JAC1 sequences obtained from clades possessing only SSC1.

Each JAC1 cladogram was subjected to rate analysis using models consisting of either three or ten possible rate categories. High and low initial values (1.3 and 0.3) for ω and κ were tested, and it was found that in all cases the analyses reached convergence under both starting values for both parameters. Codon frequency models were varied as

described above for the clade model rate analyses above. The number of ω categories which best modeled site-specific rates of evolution for each JAC1 clade was determined by likelihood ratio tests (see Appendix E). The use of ten rate categories was found to confer a significantly greater likelihood of predicting the data for the Aspergillus clade when either the BI or ML input trees were used. The results obtained from the simpler model, using three rate categories, was superior in all other cases.

Branch-Site Test

To investigate the possibility that positive selection played an historical role in the adaptation of SSQ1 to Fe/S cluster biogenesis specialization, as stated in **Hypothesis 4**, we conducted a branch-site test to analyze the codon-specific selection pressures of the ancestral SSQ1. The ancestral SSQ1, which existed immediately after the mtHsp70 gene duplication, was defined as the foreground branch (see Figure 5C). We expected sites along the foreground branch to show evidence of positive selection. The model placed each codon into one of four ω rate categories, with restrictions placed on ω values as shown in Table 1. Codons were placed into two classes for which ω was constant among ancestral and descendent sites, and two classes for which ω was variable between ancestral and descendent SSQ1 sites. The value of ω was estimated to be $0 < \omega < 1$ for common rate class 1, while the proportion of sites with $\omega=1.0$ shared among ancestral and descendent sites was estimated for common rate class 2. To test the alternative model of evolution under selection, the estimated ω of ancestral SSQ1 sites was free to vary with $\omega > 1$, while holding descendent SSQ1 sites at $0 < \omega < 1$ for divergent rate class 1. The background ω for divergent rate class 2 was held at 1.0. Posterior probabilities for

site classes were calculated by the Bayes empirical Bayes (BEB) method (Zhang and Yang 2005). The same eleven SSC1 and SSQ1 combined gene trees used for clade model analyses were input into the branch-site test, with a 3X4 codon frequency model and the parameters κ and ω estimated from the data. The results of these tests were compared by likelihood ratio test to the null model under which all sites of the ancestral SSQ1 branch evolving at a divergent rate were modeled with a fixed $\omega = 1$.

Table 1: Evolutionary Rate (ω) Estimation Under the Branch-Site Model

Evolutionary Rate Class	Descendent Lineages (background)	H_0 Ancestral SSQ1 (foreground)	H_1 Ancestral SSQ1 (foreground)
Common rate class 1	$0 < \omega > 1$	$0 < \omega < 1$	$0 < \omega < 1$
Common rate class 2	$\omega = 1$	$\omega = 1$	$\omega = 1$
Divergent rate class 1	$0 < \omega < 1$	$\omega = 1$	$\omega > 1$
Divergent rate class 2	$\omega = 1$	$\omega = 1$	$\omega > 1$

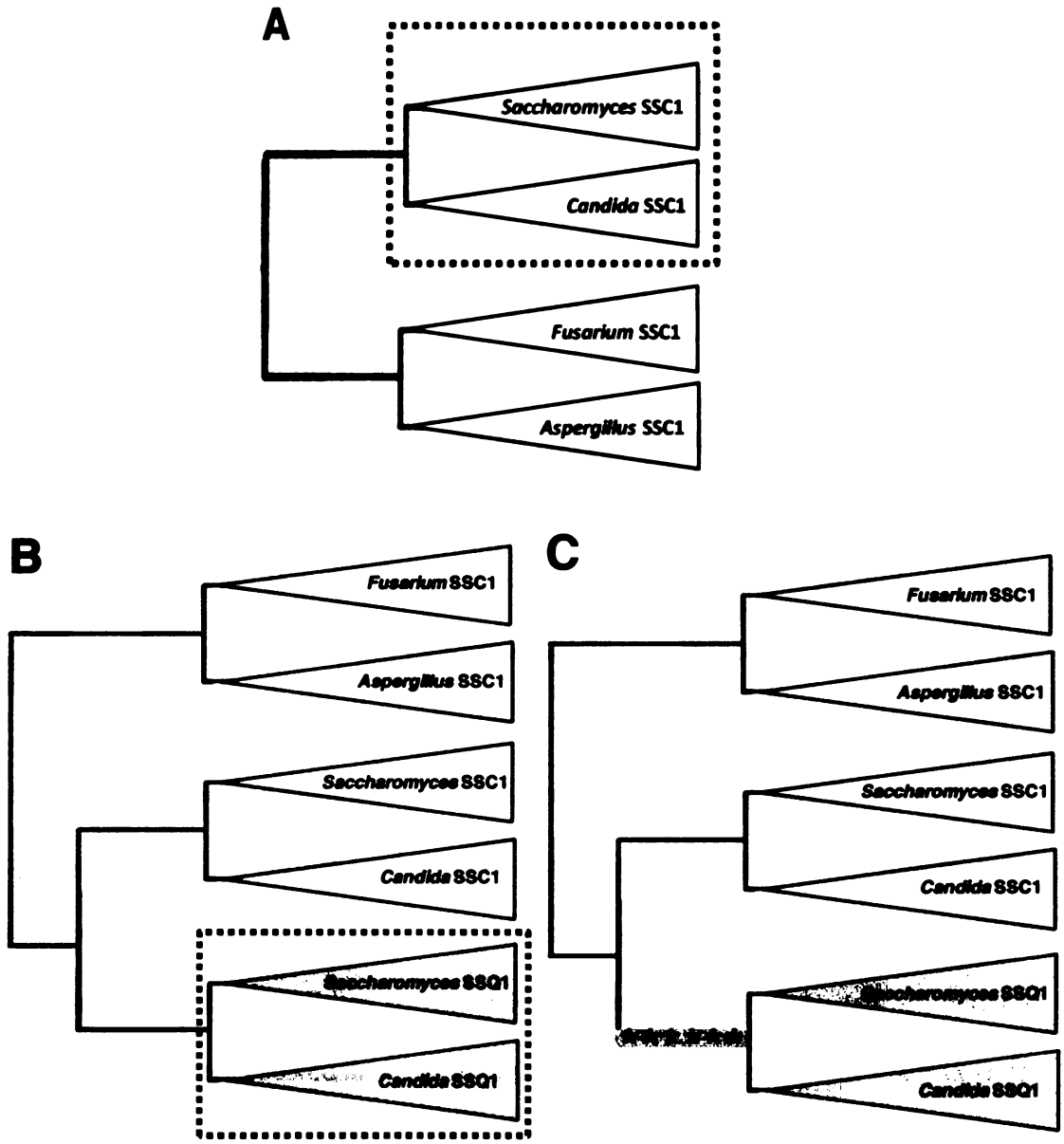


Figure 5: *a priori* defined lineages used for clade and branch-site model input trees. The phylogenetic relationships among SSC1 fungal sequences, shown in dark gray, and SSQ1 fungal sequences, in light gray, are depicted in these simplified schematic trees. Dotted boxes are used to encompass foreground clades in trees A and B. The clade model was used to test for divergent selection pressures among SSC1 of pre- and post- mtHsp70 duplication clades (A), and among SSQ1 and SSC1 (B). The branch-site test was used to look for evidence of positive selection along the highlighted ancestral SSQ1 branch (C). A starred thick gray line is used to indicate the foreground lineage in tree C, representing the ancestral SSQ1 sequence present following the mitochondrial heat shock protein 70 (mtHsp70) gene duplication event and prior to the divergence of the *Saccharomyces* and *Candida* SSQ1 clades

RESULTS

SSC1 evolution accelerated in the presence of SSQ1

The clade model test was conducted to examine whether altered selective constraint affected the evolutionary rate of SSC1 in the presence of SSQ1 (**Hypothesis 1**). Rates of codon evolution were compared between SSC1 DNA sequences derived from fungal clades that differed with respect to the presence of the fungal paralog, SSQ1 (*Candida* and *Saccharomyces* vs. *Aspergillus* and *Fusarium*, harbor the presence and absence of SSQ, respectively). The purpose of this test was to identify the proportion of SSC1 codons evolving at different rates between those SSC1 sequences that co-occur with SSQ1 (foreground clade) and those evolving in the absence of SSQ1 (background clade), and to determine the ω of those sites evolving at differential rates. The results presented were obtained using the SSC1 MLTree 2, the tree that gave the highest likelihood score when used as the input tree. Similar results were attained with all tree topologies tested, and are thus independent of tree topology. More than half (61.6%) of sites in all SSC1 genes exhibited an ω of 0.001 (common rate class 1), and just under a third (28.8%) of sites showed an ω value of 0.038 (common rate class 2), regardless of the presence or absence of the duplicate gene (Figure 6). However, about 9.6% of SSC1 codons differ in their rate of evolution, depending on the presence or absence of SSQ1 (Figure 6). The faster evolving codons, belonging to clades lacking SSQ1, show an ω of 0.107. In contrast, these same SSC1 codons evolved more than twice as fast, with ω = 0.284, in taxa possessing SSQ1 (Figure 6).

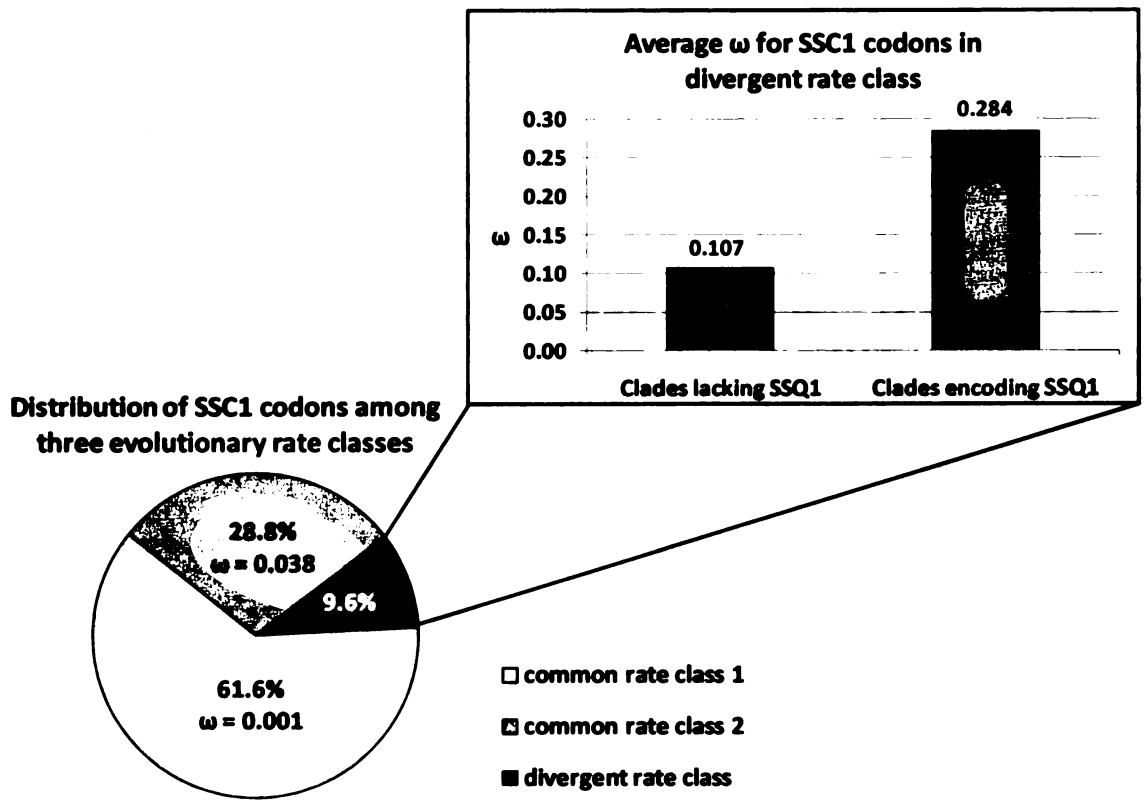


Figure 6: Comparison of SSC1 codon evolution from taxa encoding SSQ1 and taxa lacking SSQ1. The pie graph depicts the distribution of SSC1 codon evolutionary rates. Common rate classes are comprised of codons common to SSC1 from all taxa that evolve at the same rate. Codons of the divergent rate class are those common to SSC1 from all taxa that show two different rates of evolution, corresponding to the co-occurrence or absence of SSQ1. The largest proportion (61.6%) of SSC1 codons belong to common rate class 1, with an $\omega = 0.001$. The second largest proportion (28.8%) of SSC1 codons belong to common rate class 2, with an $\omega = 0.038$. The smallest proportion (9.6%) of SSC1 codons were placed into the divergent rate class. The bar graph depicts the difference in evolutionary rates between SSC1 from clades lacking SSQ1 and clades encoding SSQ1. The codons of the divergent rate class of clades encoding SSQ1 evolve with an $\omega = 0.107$, while the codons of the divergent rate class of clades lacking SSQ1 evolve with an $\omega = 0.284$.

SSQ1 has evolved at a faster rate than SSC1

Additionally, the clade model test was used to examine the rate of SSQ1evolution relative to SSC, in order to determine whether or not there is evidence for an increase or decrease in selective constraint acting on SSQ1 (**Hypothesis 2**). By designating the

monophyletic group formed by all SSQ1 sequences as the foreground clade and the monophyletic group comprised of all SSC1 sequences as the background clade, the clade model test was used to determine the magnitude and direction of selection acting on a proportion of codons evolving at different rates between SSC1 and SSQ1. The results presented were obtained using the SSC1 and SSQ1 combined BI Tree 4, the input tree which yielded the most likely clade model outputs. SSQ1 sequences were found to contain a subset of sites evolving faster than those of SSC1 (Figure 7). Most of the sites conserved between SSC1 and SSQ1 are evolving at equal (slow or intermediate) relative rates, with about 43.5% having an $\omega = 0.002$ and about 40.8% having an $\omega = 0.031$(Figure 7). Approximately 15.6% of codons estimated to have a differential rate ratio of about 0.209 in SSQ1 and about 0.077 in SSC1, which is nearly three times as fast in SSQ1 than in SSC1 (Figure 7). A total of 82 codons comprise the 15.6% of SSC1 and SSQ1 in the divergent rate class. The encoded amino acids are highlighted within the Ssq1p amino acid sequence in Figure 12.

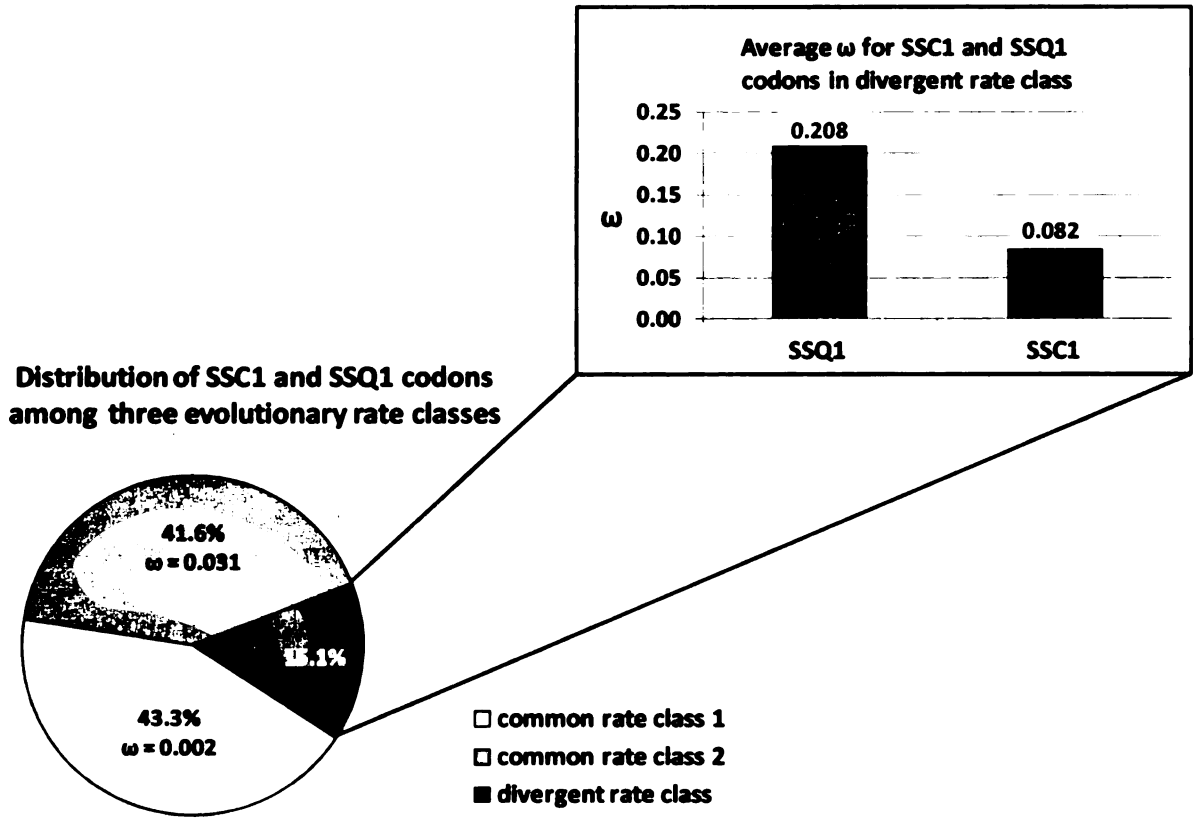


Figure 7: Comparison of SSC1 and SSQ1 codon evolution. The pie graph depicts the distribution of SSC1 and SSQ1 codon evolutionary rates. Common rate classes are comprised of codons common to SSC1 and SSQ1 that evolve at the same rate in all taxa. Codons of the divergent rate class are those common to SSC1 and SSQ1 from all taxa that show two different rates of evolution unique to each paralog. The largest proportion (43.3%) of SSC1 and SSQ1 codons belong to common rate class 1, with an $\omega = 0.002$. A nearly equal proportion (41.6%) of SSC1 and SSQ1 codons belong to common rate class 2, with an $\omega = 0.031$. The smallest proportion (15.1%) of SSC1 and SSQ1 codons were placed into the divergent rate class. The bar graph depicts the difference in evolutionary rates between SSC1 and SSC1. The SSC1 codons of the divergent rate class evolve with an $\omega = 0.208$, while the SSQ1 codons of the divergent rate class evolve with an $\omega = 0.082$.

For both mtHsp70 comparative analyses, the clade model that grouped all codons into one of three rate categories was significantly more likely to predict the data, as indicated by likelihood ratio test, than when only two rate categories were used. Likelihood ratio test results are presented in Appendix E. The null model, with all codons constrained to have evolved at equal rates, was also rejected in every instance by

likelihood ratio tests. Statistical validation of the use of the clade model with three ω categories held among all tree topologies examined (11 input trees for SSC1 and eight input trees for SSC1 and SSQ1 combined). Results of the clade model tests indicated that SSC1 evolved at an elevated rate when co-occurring with the duplicate gene, while SSQ1 evolved faster than SSC1.

JAC1 evolution has decelerated in the presence of SSQ1

A site-specific model was used to examine the direction and strength of selection that acted on individual codons of JAC1 among the *Candida*, *Saccharomyces*, *Aspergillus*, and *Fusarium* fungal clades. Our purpose was to assess possible trends in JAC1 evolution from clades possessing duplicate Hsp70s compared to clades lacking the duplicate mtHsp70 (**Hypothesis 3**).

Figures 8 and 9 show the distribution of codon evolutionary rates across the JAC1 sequences. The alternative tree topologies tested closely agree in the magnitude and location of elevated codon rates for the *Saccharomyces* and *Candida* clades. In the case of the *Aspergillus* clade, examination of alternative, strongly supported tree topologies resulted in some variation in the magnitude, but not location, of elevated codon rates. Only one JAC1 tree topology was used in the analysis of *Fusarium* clade sequences because the topologies generated by each phylogenetic inference method were identical. In the clades containing the duplicate gene, SSQ1, JAC1 shows similar ω values across the gene sequence, rarely rising above 0.1 (Figure 8). In contrast, when the sequences from fungi lacking SSQ1 are examined, the average ω of JAC1 is greater (Figure 9). The average ω across the JAC1 sequence and corresponding standard errors from each clade

were as follows: *Saccharomyces*: 0.0546 ± 0.0028, *Candida*: 0.0348 ± 0.0024, *Aspergillus*: 0.0711 ± 0.0061, and *Fusarium*: 0.0812 ± 0.0056. The variance of ω values estimated for JAC1 from the clades lacking SSQ1 was also greater than from the clades co-occurring with SSQ1 (*Saccharomyces* : 0.0014, *Candida*: 0.0010, *Aspergillus*: 0.0070, and *Fusarium*: 0.0059). Additionally, none of the JAC1 site-specific analyses produced ω estimates of 0, excluding the possibility of the absence of nonsynonymous mutations at a particular site across the sequences of a clade. Thus, the results of our codon-specific rate analysis of JAC1 from four fungal clades has opposed our prediction; the rate of evolution of JAC1, the J-protein co-chaperone specialized in Fe/S cluster assembly, slowed down following the duplication of the mtHsp70.

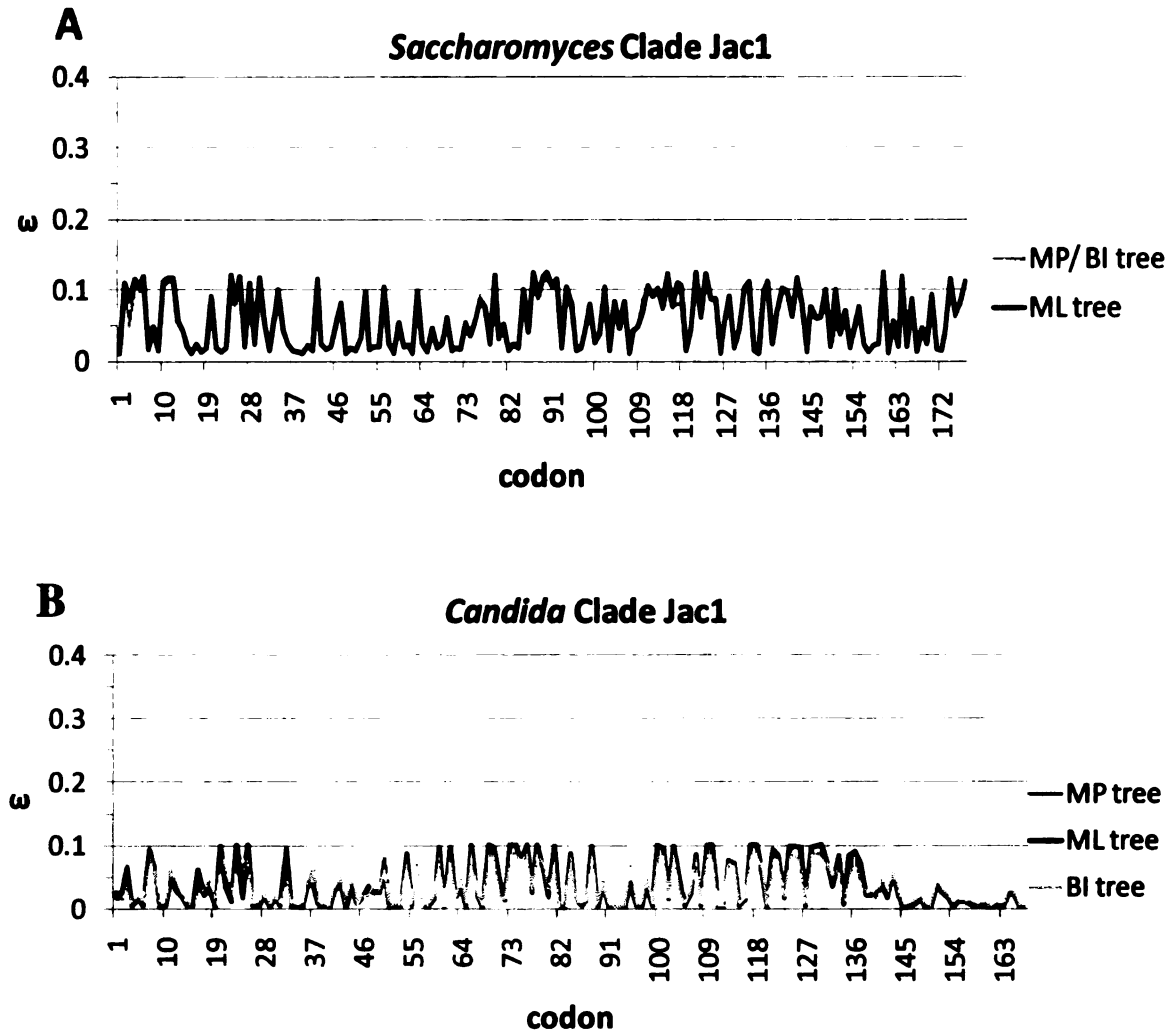


Figure 8: Site-specific ω estimations for JAC1 from clades encoding SSQ1. Evolutionary rates (ω) for JAC1 codons from the *Saccharomyces* and *Candida* clades are shown as a function of codon position within the gene sequence. Codon numbers represent column positions within trimmed nucleotide sequence alignments. Results from each input tree topology are represented: (A) *Saccharomyces* clade, MP/BI tree shown in dark gray, ML tree shown in black, (B) *Candida* clade, MP tree shown in dark gray, ML tree shown in black and BI tree shown in light gray.

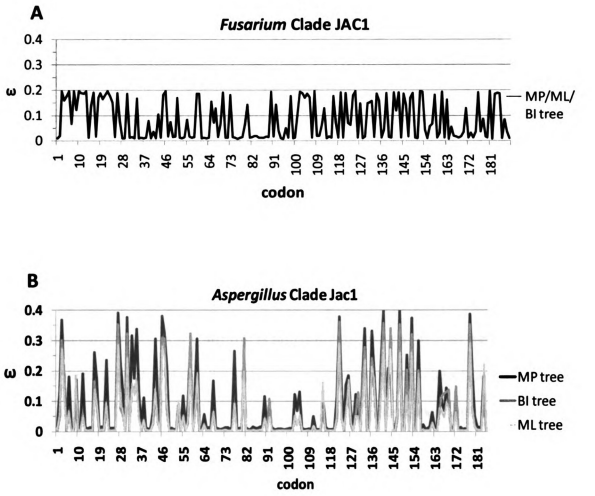


Figure 9: Site-specific ω estimations for JAC1 from clades lacking SSQ1. Evolutionary rates (ω) for JAC1 codons from the *Fusarium* and *Aspergillus* clades are shown as a function of codon position within the gene sequence. Codon numbers represent column positions within trimmed nucleotide sequence alignments. Results from each input tree topology are represented: (A) *Fusarium* clade, MP/ML/BI tree, (B) *Aspergillus* clade, MP tree shown in black, BI tree in light gray, and ML tree shown in dark gray.

SSQ1 has evolved under positive selection

Because JAC1 is evolving slowly in the presence of SSQ1, we suspected that JAC1 and SSQ1 have reached an optimum coevolutionary state among the extant taxa. This suggests that the potential for adaptive coevolution may have occurred between SSQ1 and JAC1 (**Hypothesis 4**). Ideally, we would test for evidence of positive selection

along the ancestral branch of JAC1 corresponding to the lineage in which SSQ1 arose. However, such a JAC1 branch-site test would require a single phylogenetic tree that incorporated sequences from all fungal clades, in order to reconstruct ancestral states at critical points in evolutionary history. Due to our inability to generate the needed multiple sequence alignment, the required tree could not be inferred. However, such tests are possible with SSQ1. Therefore, we conducted a branch-site test to detect evidence of positive selection affecting sites along the tree branch giving rise to SSQ1.

Sites with constant evolutionary rates in both the ancestral SSQ1 branch (inferred sequence of the foreground branch) and all other sequences (background branches) were grouped into two categories (Figure 10). A proportion of codons (81.3%) were estimated to have evolutionary rates of $\omega = 0.034$, representing common rate class 1, and 4.4% exhibited an $\omega = 1.000$, representing common rate class 2. Hence, these rate categories were constant regardless of whether the sequence was that of the ancestral SSQ1 gene or a background gene (Figure 10). For 11.8% of codons, ω was estimated at 1.994 within the ancestral SSQ1 and 0.034 for all other genes, designated divergent rate class 1 (Figure 10). A very small fraction of sites (0.6%) were placed into divergent rate class 2, which evolved at a rate of $\omega = 1.994$ in the ancestral gene, while these same codons evolved at $\omega = 1.000$ in derived sequences. This suggests that 12.4% of ancestral SSQ1 codons, representing codons from both divergent rate classes 1 and 2, were subjected to positive selection immediately following SSC1 gene duplication.

Though the posterior probabilities associated with the placement of each codon into a given rate category varied according to tree topology, five out of the nine tree topologies agreed on four candidate sites for the initial fixation of adaptive mutations

following the birth of SSQ1. These four codons, corresponding to amino acids His^{315}, Lys^{317}, Glu^{338}, and Leu^{346} of the raw SSQ1 sequence from *S. cerevisiae* YJM789, were given a posterior probability of ≥ 0.90 of having an ω of approximately 2 by at least 5 of the tree topologies tested (shown in Figures 11 and 12). Several other residues were given a high probability of having undergone positive selection in ancestral SSQ1 by some tree topologies (see Figure 11). The results obtained using tree topology BI5, however, identified a different set of residues with high probabilities of belonging to a rate category with $\omega > 1$ and did not support evolution under positive selection for the residues shown in Figure 11. The source of this anomaly is unclear, given that the topology of the BI5 tree does not show any large deviations from the other topologies used. All likelihood ratio tests allowed for the rejection of the null model of neutral evolution, validating the branch-site test model incorporating sites evolving under positive selection, as a statistically significantly better fit to model early SSQ1 evolution.

The branch-site test was thus able to detect evidence of positive selection within the ancestral SSQ1 lineage immediately following gene duplication, and thereby rejects evolution by neutrality. Together, the two variable rate categories suggest that adaptive evolution in SSQ1 decelerated in descendent gene sequences after a burst of stronger selection immediately following the inception of SSQ1.

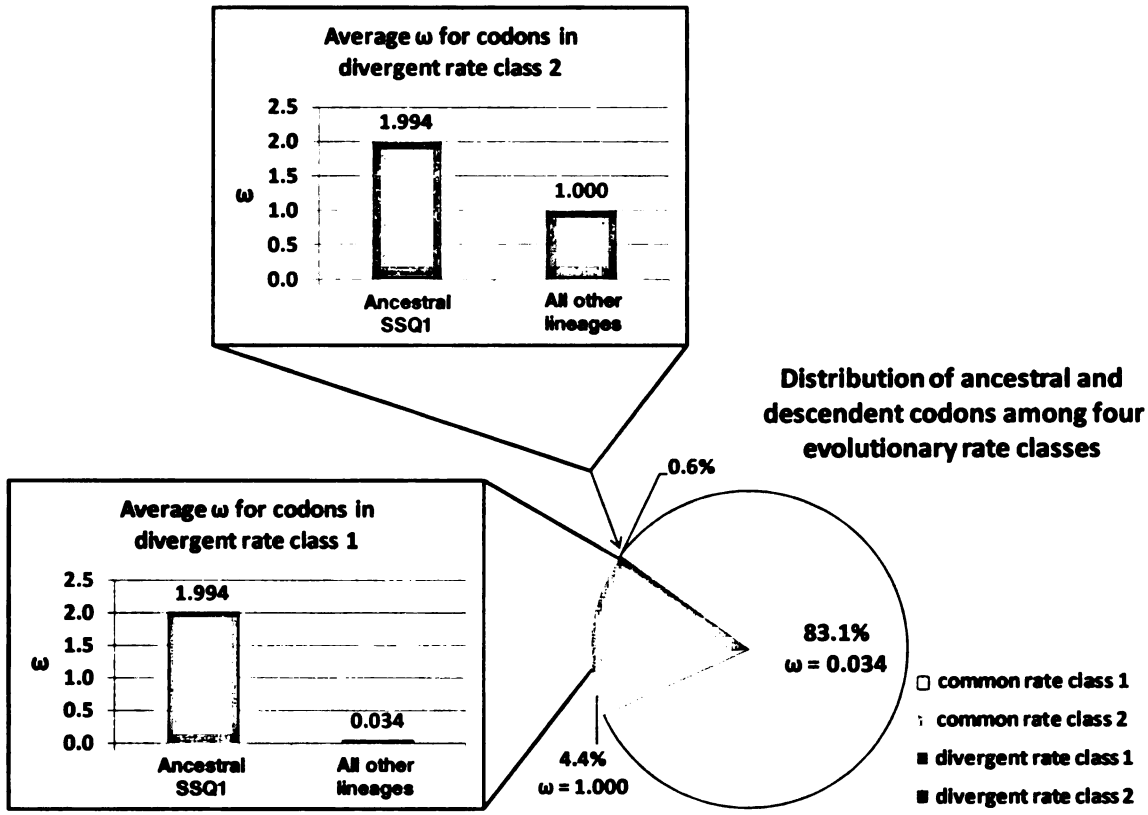


Figure 10: Comparison of ancestral SSQ1 codon evolution to SSQ1 and SSC1 evolution within all other lineages. The pie graph depicts the distribution of SSQ1 codon evolutionary rates. Common rate classes are comprised of codons that evolve at a constant rate. Codons of the divergent rate classes are those common to ancestral and present-day descendent lineages that show two different rates of evolution, for each divergent rate class, unique to the ancestral and descendent lineages. The largest proportion (83.1%) of codons belong to common rate class 1, with an $\omega = 0.002$. A much smaller proportion (4.4%) of codons belong to common rate class 2, with an $\omega = 1.000$. A proportion of 11.8% of codons were placed into divergent rate class 1. The bottom bar graph depicts the difference in evolutionary rates between ancestral SSQ1 codons and those of descendent sequences in divergent rate class 1. The ancestral SSQ1 codons of divergent rate class 1 evolve with an $\omega = 1.994$, while SSC1 and SSQ1 codons from all other lineages of the tree of divergent rate class 1 evolve with an $\omega = 0.034$. The top bar graph depicts the difference in evolutionary rates between ancestral SSQ1 and descendent codons of divergent rate class 2. The ancestral SSQ1 codons of divergent rate class 2 evolve with an $\omega = 1.994$, while the present-day descendent SSQ1 codons of divergent rate class 1 evolve with an $\omega = 1.000$.

		SSQ1 Amino Acid Residues							
		Lys	Asn	His	Lys	Glu	Leu	Arg	Tyr
	Alignment Sequence	171	219	256	258	279	287	327	579
	Raw Sequence	224	274	315	317	338	346	386	649
Input Tree	BI 1	**	**	***	***	***	***	***	**
	B1 2	**	**	***	***	**	**	**	**
	BI 3	**	**	***	***	***	***	**	***
	BI 4	***	***	***	***	***	***	**	***
	BI 5	-	-	-	-	-	-	-	-
	BI 6	**	**	***	***	***	***	**	**
	ML	**	***	***	***	***	***	**	***
	MP 1	-	*	***	**	*	**	**	*
	MP 2	-	-	***	**	*	*	***	*

Figure 11: Comparison of posterior probabilities of placement of sites into a divergent rate class by the branch-site model, among input tree topologies. All residues assigned to divergent rate category 1 or 2, with $\omega > 1$ and a posterior probability of at least 0.9 in at least one of the tested tree topologies is shown. Posterior probabilities for placement in divergent rate class 2 of 0.70-0.79 (*), 0.80-0.89 (**), and 0.90-0.99 (***) are shown, with residues given a posterior probability of less than 0.70 indicated by (-). The His, Lys, Glu, and Leu residues shaded in gray are those residues of ancestral SSQ1 believed to have evolved under positive selection, given that at least five of the nine tree topologies tested resulted in those residues with a posterior probability of 0.90-0.99 of evolving with $\omega > 1$.

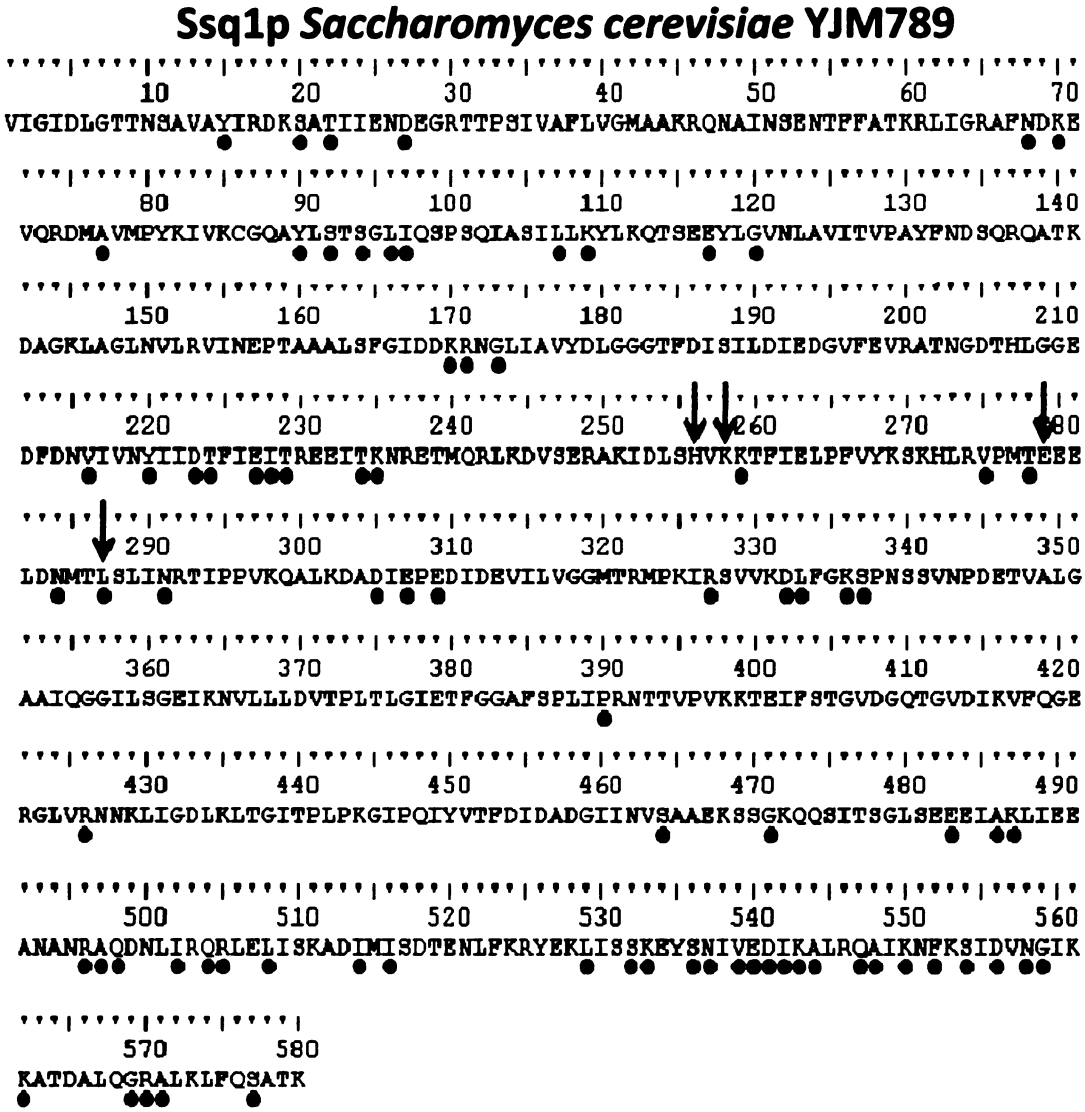


Figure 12: Amino acid sequence of Ssq1 encoded by *Saccharomyces cerevisiae* YJM789 showing sites inferred to exhibit relaxed selective constraint and ancestral positive selection. The 82 amino acids indicated with dots are the amino acids identified as belonging to the divergent rate class in the SSC1 and SSQ1 clade model test (see Figure 7), which evolve at an accelerated rate compared to SSC1. The four residues indicated by the arrows correspond to the sites identified via the branch-site model test as those estimated to have evolved under positive selection in the ancestral SSQ1, immediately following the mitochondrial heat shock protein 70 gene duplication. The Ssq1p sequence shown is from the trimmed SSC1 and SSQ1 combined sequence alignment.

DISCUSSION

Molecular coevolution among interacting proteins can confer fitness consequences to crucial enzymatic pathways and can be initiated by the ubiquitous genetic phenomenon of gene duplication. The findings presented here in the evolutionary rate analyses of the fungal mtHsp70 paralogs SSC1 and SSQ1, and the interacting J-protein co-chaperone JAC1, together with previous observations concerning the functions of the encoded proteins, bear evidence consistent with release from antagonistic pleiotropy following a gene duplication event. Subsequent subfunctionalization has facilitated the coevolution of SSQ1 and JAC1 to optimize a J-protein co-chaperone-mtHsp70 chaperone interaction dedicated to activity in the Fe/S cluster biogenesis pathway in yeast.

The mtHsp70 paralogs investigated here show a history of selection similar to that inferred for morning glory dihydroflavonol-4-reductase (DFR) duplicate genes. PAML rate analyses of the anthocyanin biosynthesis pathway DFR genes were consistent with paralog divergence via escape from adaptive conflict (Des Marais and Rausher 2008). Evidence from clade and branch-specific ω value estimates of codon rate evolution for each of the three DFR copies indicated an ancestral single-copy DFR that was subjected to purifying selection, followed by a relaxation of selective constraint after gene duplication. Evidence for positive selection within the lineage immediately following the second duplication was observed. Positive selection early in the history of the paralogs of the most recent DFR duplication potentially enabled a burst of adaptive mutation fixation within these paralogs. Combined with biochemical evidence of optimization from an

ancestral sub-function of one of the DFR paralogs, and the loss of the ability to perform other ancestral functions in paralogs, the authors concluded that antagonistic pleiotropy enforced selective constraint to prevent full optimization of all ancestral DFR functions in single-copy form. In analogy to the DFR study, one of the fungal mtHsp70s, SSQ1, was found to have undergone positive selection in its ancestral sequence shortly following the gene duplication event from which it was created. Like DFR-B, SSQ1 became specialized in a role performed by the pre-duplication gene, and may have even evolved to outperform its paralog, SSC1, in terms of increased affinity for Jac1p and greater ATPase activation.

SSQ1 shows biochemical evidence of ATPase activity improvement in response to JAC1 stimulation, with the potential to improve Fe/S cluster biogenesis efficiency, an ancestral pre-duplication function. Concomitantly, SSQ1 can no longer perform the ancestral mtHsp70 functions of protein folding and translocation functions, nor provides protection to cellular integrity from environmental stresses. The functional evolution of SSQ1 thus fits the criteria for a case of subfunctionalization. Furthermore, JAC1 evolution resulting in the loss of J-domain residues important for Ssc1p ATPase activation has occurred in yeasts encoding SSQ1. Therefore, an alteration of the J-domain of Jac1p may have been necessary for improved affinity to Ssq1p and may have been evolutionarily favored only in the presence of a mtHsp70 specialized in Fe/S cluster biogenesis. Coevolution of JAC1 with SSQ1 would have thus been a consequence of mtHsp70 paralog evolution following escape from adaptive conflict.

However, evidence is lacking to meet the more stringent criteria of SSC1 and SSQ1 evolution by escape from adaptive conflict. There is no direct proof of a novel

function arising in the pre-duplication mtHsp70 that reduced the ability of the ancestral protein to perform any of its other tasks. This would require the biochemical characterization of the protein translated from an ancestral gene reconstruction. Additionally, future investigation of Ecm10p functions, and the selective forces acting on this third yeast mtHsp70 duplicate, could bolster the case for adaptive conflict in the pre-duplication mtHsp70 if ECM10 has also optimized an ancestral SSC1 function. Finally, it remains to be determined if a more efficient Jac1p-Ssc1p interaction optimizes the Fe/S cluster assembly pathway to increase yeast fitness.

The functional specialization of SSQ1 also resembles the subfunctionalization for optimization of GAL1 and GAL3 functions, after release from antagonistic pleiotropy, by gene duplication (Hittinger and Carroll 2007). While promoter divergence resulted in the evolved phenotypes of differential control over GAL1 and GAL3 transcription, regulatory evolution of the mtHsp70s was not examined in this study. However, previous observations of decreased SSQ1 expression compared to SSC1, within *S. cerevisiae* mitochondria, suggests that SSQ1 and SSC1 have also undergone regulatory divergence.

Another possibility is that the specialized function of SSQ1 hinges on a mutation analogous to a GAL1 Ser-Ala di-peptide identified to be sufficient for galactokinase activity when added to the active site of GAL3, the co-inducer of the galactose uptake pathway. Because deletion of the di-peptide from GAL1, and a pre-duplication GAL1/GAL3 bifunctional protein, did not improve the co-inducer function of the encoded proteins, the Ser-Ala mutation of GAL1 could not be ruled a source of adaptive conflict. The effect of the Ser-Ala mutation on galactokinase function was dependent on the background of residues present at other sites within GAL1. It is possible that

mutations have similarly arisen in SSQ1 that now contribute to functional specialization, but were fixed as compensatory mutations secondary to mutations fixed as a direct result of release from antagonistic pleiotropy.

It is reasonable to hypothesize that opportunity for functional specialization of proteins like pigment biosynthesis enzymes, galactose pathway components, or mtHsp70s, may extend to molecules that participate within a common biological pathway, by coevolution. The release of SSQ1 from antagonistic pleiotropy has influenced the evolution of JAC1, the J-protein partner also specialized in this pathway. JAC1 coevolution with the mtHsp70 paralogs has allowed its interaction with SSQ1 to become more efficient, while decreasing its efficiency of ATPase stimulation in SSC1.

Support for Hypothesis 1: Selective constraint has been relaxed in SSC1 in the presence of its paralog, SSQ1.

An equal rate of SSC1 evolution, in the presence versus absence of SSQ1, was rejected. SSC1 evolved faster in the presence of its paralog, SSQ1.

The result that SSC1 evolved faster when co-occurring with SSQ1 is consistent with the conclusions of Scannell and Wolfe (2008), who found that recent paralogs tend to evolve at an increased rate compared to singleton genes. Here, we suggest that the functional specialization of SSQ1 has relieved SSC1 of the Fe/S cluster biogenesis task, thereby relaxing selective constraint acting on SSC1 for this particular function. The availability of the SSQ1:JAC1 specialized pair could have rendered the SSC1:JAC1 cooperation less important, thus allowing a greater proportion of nonsynonymous codon changes to be tolerated in SSC1, particularly if those sites encode residues that contribute to interaction with JAC1, or other unidentified aspects of Fe/S cluster biogenesis.

In the absence of SSQ1, however, antagonistic pleiotropy would continue to impose evolutionary constraint on SSC1, because SSC1 would be required to perform Fe/S cluster formation, in addition to protein import and folding. While evidence does not yet exist to suggest that SSC1 has improved any other pre-duplication mtHsp70 function in the presence of SSQ1, it could be that escape from adaptive conflict may allow SSC1 to perform a chaperone task, such as peptide translocation across the inner mitochondrial membrane, with greater efficiency if optimization is permitted in the presence of paralogs. This seems plausible if the relaxation of selective constraint on SSC1 among clades that harbor duplicate genes persists for tens of millions of years (Scannell and Wolfe 2008). An extended period of relaxed constraint may have the potential to fix many mutations via drift, and as a composite, could result in an altered phenotype.

Support for Hypothesis 2: SSQ1 is under less selective constraint than SSC1 because SSQ1 has fewer encoded functions to maintain.

Evolution of SSQ1 and SSC1 at equal rates was rejected. SSQ1 evolved faster than SSC1.

When the average rate of codon evolution was compared between SSQ1 and SSC1, we were able to conclude that SSQ1 evolved faster than SSC1. This rate asymmetry is consistent with other published analyses of evolutionary rate asymmetry in paralogs (Conant and Wagner 2003; Zhang et al. 2003). An examination of gene duplicates created by a whole genome duplication in yeast revealed that genes with the most dramatic evolutionary rate increase, immediately following duplication, remained the "faster" evolving gene of the two paralogs. Therefore, it is likely that SSQ1 will continue to evolve with a greater ω than SSC1. While evidence for sites under positive

selection (an ω greater than 1) in extant taxa was not identified, the faster rate of SSQ1 evolution compared to SSC1 is interpreted as a result of relaxed constraint, depressed expression level, or both.

The increased rate of evolution for SSQ1 could be due to relaxation of selective constraint that is independent of gene expression in order to allow for specialization on a single function. Relaxed constraint on SSQ1, compared to the ancestral single-copy mtHsp70, likely initially resulted from the ability of SSC1 and SSQ1 to reciprocally compliment one another, and subsequently also provide robustness against deleterious mutations. For example, if a mutation in SSC1 resulted in diminished function as an Fe/S cluster biogenesis chaperone, SSQ1 would have been able to restore this function. We propose that SSQ1 would then have been free to optimize efficiency for its role in Fe/S cluster assembly in the presence of SSC1, which could functionally replace SSQ1 for any of the many sub-functions that may have been compromised during Fe/S cluster assembly optimization. As a result, disproportionately many sites in the protein may now be under relaxed selection and thus evolve at a faster rate compared to the multifunctional SSC1.

Alternatively, gene expression divergence of SSQ1 and SSC1 alone is a viable explanation for the faster rate of SSQ1 evolution. This line of reasoning is supported by the Drummond et al. (2005) study, which concluded that gene expression level is the single greatest determinant of protein evolution, explaining more than half of the variation in nucleotide substitution rates of genes in *S. cerevisiae*. Though gene length, dispensability, and recombination have also been suggested as factors aiding to predict evolutionary rates of genes, these variables seem to play a minor role in the determination of evolutionary rates. In addition, expression levels have been shown to exert control

over these factors, often confounding efforts to link these factors as direct causes. Drummond et al. (2005) revealed that genes with a lower level of expression tend to evolve faster and offer an explanation for this observation independent of selection on protein function. It is known that errors during mRNA translation lead to the accumulation of mis-folded and toxic protein products that impose fitness costs to a cell by disrupting metabolic processes (Bucciantini et al. 2002). It was therefore proposed that selection acts to increase the translational accuracy of a sequence, (i.e. using the most abundant tRNA anticodons for amino acids), and to increase the robustness of a sequence to translational errors. Favoring amino acid sequences that fold into functional proteins, regardless of the generation of missense errors, increases translational robustness (Drummond et al. 2005).

A subsequent study (Drummond and Wilke 2008) identified protein misfolding costs as the underlying selective pressure responsible for the co-variation in evolutionary rates, codon preference, and gene expression within and between genes, observed for model organisms ranging in complexity from *E. coli* to humans. The authors revealed that translational accuracy, translational robustness, the synthesis of full-length peptides, and the tendency to fold properly, all correlate positively with gene expression level. The cost of protein misfolding thus provides a reason for the selective constraint that gives rise to a greater proportion of optimal codons observed at conserved sites within a protein in genes that are most highly expressed.

When a gene is expressed at a higher level, as with SSC1, translation occurs more frequently, increasing the number of opportunities for detrimental errors, so that accuracy and robustness become more influential to the cell's overall fitness. Therefore, by

selecting against protein sequences with toxic characteristics (such as a propensity for aggregation) when translated incorrectly, the same evolutionary forces may indirectly select for a protein structure with enhanced thermodynamic stability. Together, selection which results in an increase in translational accuracy and robustness may have the effect of lowering both the rate of synonymous and nonsynonymous mutation fixation, imposing a form of evolutionary constraint at the sequence level. Higher expression level may therefore bring about increased evolutionary constraint on SSC1, while the relatively decreased level of expression of SSQ1 may result in relaxation of constraint. Divergence in the expression level of paralogous genes could occur as a consequence of accelerated promoter or regulatory region evolution by adaptive or neutral evolution. This has been suggested to be a common phenomenon in eukaryotes (note that evidence of regulatory sequence evolution would go undetected in protein-coding ω analyses) (Zhang, 2003). Alternatively, divergence in paralog expression levels can result from other sequence changes that contribute to mRNA stability or chromatin structure differences between the gene duplicates (Li et al. 2005).

Indeed, the approximately 1000-fold lower concentration of Ssq1 protein present in the mitochondria of *S. cerevisiae* (Voisine et al. 2000) is accompanied by a decreased codon usage bias and an increased overall rate of nucleotide substitution, indicative of relaxed selective constraint. While the codon adaptation index for SSC1 is reported to be 0.521, the codon adaptation index of SSQ1 is much lower, at 0.148 (SGD project, Sept. 2008) and is indicative of less selective constraint acting on third position nucleotides of SSQ1 codons. Less constraint on these nucleotides could allow SSQ1 to tolerate more synonymous substitutions than SSC1. Therefore, the elevated ω of SSQ1 is impressive in

the face of an elevated d_S, as was observed in a gene-wide average of site-specific d_S values estimated across tree branches and compared to SSC1 d_S averages (data not shown).

Lack of support for Hypothesis 3: The rate of JAC1 evolution is positively correlated with the rate of SSQ1 evolution because JAC1 and SSQ1 are coevolving.

An equal rate of JAC1 evolution in clades encoding SSQ1 compared to the rate of JAC1 evolution in clades that lack SSQ1 is rejected. However, JAC1 evolution decelerated after mtHsp70 gene duplication.

Here we have examined the influence of a gene duplication event on the selective forces driving the molecular evolution of protein partners specialized in Fe/S cluster assembly. We have demonstrated that JAC1 evolves faster in the absence of SSQ1. Our proposed explanation is that selective constraint is acting on JAC1 to preserve an optimized, physical interaction with SSQ1, which resulted from the coevolution of JAC1 and the duplicate, specialized mtHsp70. While JAC1 now evolves slowly in the presence of SSQ1, it is conceivable that the rate of evolution of JAC1 was initially accelerated after the mtHsp70 gene duplication that gave rise to SSQ1. Subsequently, JAC1 may have quickly reached an adaptive peak, together with SSQ1, in its ability to facilitate Fe/S cluster assembly. Or, JAC1 was brought under constraint by some other influence. The rapid rate of JAC1 evolution, however, precludes the testing of this hypothesis, as carried out for SSQ1, since we could not reconstruct JAC1 ancestral states. We speculate that, subsequent to initial rate acceleration during a co-adaptive arms race to fix complimentary changes in the sites that physically interact between Jac1p and Ssq1p, Jac1p evolution slowed to maintain efficient cooperation with Ssq1p. An alternative explanation for the faster rate of JAC1 evolution in the *Aspergillus* and *Fusarium* clades

could be a smaller effective population size of representative species compared to the *Candida* and *Saccharomyces* clades, which would in turn result in a reduced efficiency of purifying selection.

Though expression data of JAC1 in the fungal species from which sequences were analyzed is unavailable, it is possible that the expression of JAC1 has been increased in the organisms possessing SSQ1 to balance molecular stoichiometry. Indeed, a higher average codon bias, consistent with higher levels of gene expression (Wang et al. 2005), was observed for JAC1 from clades encoding SSQ1. JAC1 CAI value means and standard errors calculated for each clade were as follows: *Saccharomyces*: 0.273 ± 0.013, *Candida*: 0.249 ± 0.010, *Fusarium*: 0.195 ± 0.013, and *Aspergillus*: 0.183 ± 0.010. Thus, the third nucleotide positions of JAC1 codons from clades encoding SSQ1 are likely to be under stronger selective constraint than third nucleotide positions within JAC1 from clades lacking SSQ1. Increased constraint on third position nucleotides, as well as an overall increase in constraint to preserve translational robustness when gene expression is elevated, may be depressing ω in JAC1 from *Saccharomyces* and *Candida* taxa.

Support for Hypothesis 4: SSQ1 has undergone adaptive evolution to optimize the Ssq1 - Jac1p interaction important for Fe/S cluster biogenesis.

Evolution of SSQ1 in the absence of positive selection is rejected. SSQ1 evolved under positive selection in the lineage immediately following its inception.

The antagonistic pleiotropy that characterized the ancestral mtHsp70 prior to gene duplication may have been broken by positive selection in ancestral SSQ1, immediately following its gene duplication. Positive selection may have enabled a burst of adaptive evolution to optimize the Jac1p-Ssq1p partnership important for Fe/S cluster biogenesis and promoted rapid subfunctionalization of SSQ1, thus relaxing constraint on SSC1 at

sites necessary for interaction with Jac1p. The retention of SSQ1 within the genome following the gene duplication event may be attributable to this subfunctionalization, possibly having involved an adaptive sweep at ancestrally positively selected sites His^{315}, Lys^{317}, Glu^{338}, and Leu^{346} within the ATPase domain. Given that past studies have shown how significant adaptive shifts can be instigated by very few amino acid substitutions (Golding and Dean 1998), rapid mtHsp70 SSQ1 evolution may have been responsible for its coevolution with JAC1 to specialize the J-protein-mtHsp70 pair. Alternatively, the signature of an initial burst of selection detected in the ancestral sequence of SSQ1 may not have been accompanied by functional adaptation at all sites, but instead reflect the fixation of compensatory substitutions to rescue a decrease in fitness arising from deleterious mutations within the gene or even elsewhere within the genome (Pal et al. 2006).

Future Directions

Possible future lines of research include conducting protein structural and functional analyses, via experimental genetics and biochemistry, in order to elucidate the role of particular SSC1, SSQ1, and JAC1 sites in Fe/S cluster biogenesis. Site-directed mutagenesis and reconstruction of inferred ancestral gene sequences, followed by biochemical characterization of 'resurrected' ancestral proteins, is a technique that has been successfully used in the past to gain insight into the fates of paralogous genes following gene duplication (Zhang and Rosenberg 2002). Additional experiments could include mutating SSC1 sites to those corresponding to SSQ1 sites that were identified to have undergone positive selection immediately after the gene duplication. It would be

interesting to determine if those sites from SSQ1improve the efficiency of Ssc1p ATPase activity in the presence of Jac1p, and if so, whether those sites are necessary for direct contact with Jac1p, the nucleotide exchange factor protein, or the nucleotide. SSC1 engineered to encode an ATPase domain that more closely resembles that of SSQ1 might also be predicted to have decreased chaperone and stress mediation functions. Such a result would directly demonstrate the tradeoff between optimization of Jac1p-mediated ATPase activity and loss of performance in other functions within the mtHsp70. The source of antagonistic pleiotropy in the ancestral mtHsp70 would thus be pinpointed within the ATPase domain. Conversely, manipulation of sites in SSQ1, where homologous positions in SSC1 are under relaxed selection, are predicted to be involved in Fe/S cluster biogenesis, as these were the sites predicted to be released from selection by subfunctionalization. On the other hand, independent manipulation of the sequences encoding the substrate binding, ATPase, and variable domains of SSQ1, to contain those sites that are under strict selective constraint in SSC1, should be performed. Such manipulation might lead investigators to attribute the increased Ssq1p ATPase activity to a domain other than the ATPase domain. Identifying sites in JAC1 that have evolved at a fast rate in the *Fusarium* and *Aspergillus* clades, but have evolved at a slower rate in the *Candida* and *Saccharomyces* clades, might also be informative in guiding similar site-specific mutation construction of JAC1.

The role of regulatory sequence evolution should also be explored in the future, perhaps by evaluating the effect of exchanging the promoters of the paralogous mtHsp70s. One expectation might be that replacing the SSC1 promoter with that regulating the transcription of SSQ1 will decrease the expression level of SSC1 within

the mitochondrial matrix. Alteration in the number, orientation, and/or sequence of transcription factor binding sites after mtHsp70 gene duplication might be expected to result in such regulatory differences. Another outcome of mtHsp70 promoter swapping might be that, when under the control of the SSC1 promoter, SSQ1 is increased in its degree of expression. However, the extent to which active Ssq1p is produced may still be less than Ssc1p levels, given that SSQ1 has a lower codon bias and therefore might be more prone to translational errors that result in truncated or misfolded proteins. Such studies would be important to verify that the expression level difference between SSC1 and SSQ1 is due to cis-regulatory evolution and is not an effect of other forms of regulation, such as feed-back inhibition.

The ultimate goal should be to elucidate details of how the mtHsp70 paralogs differ in their interaction with JAC1 and how these changes confer fitness differences via the execution of Fe/S cluster biogenesis. Thus, the direct impact that the increased Ssq1p ATPase stimulation by Jac1p confers upon the level of active Fe/S-containing proteins produced in vivo must be established. Further, the fitness advantage of an optimized Fe/S cluster biosynthesis pathway must be demonstrated by the observation of an adaptive phenotype. This will not be a trivial undertaking, as the advantage of a phenotype often varies under different growth conditions and the presence of ecological competitors. However, as with any molecular process, if we are to advance our understanding of Fe/S cluster biosynthesis, we must study the pathway components in the context of evolutionary and ecological dynamics.

Fun

Appendix A

Fungal Mitochondrial Heat Shock Protein Coding Region DNA Sequence Sources

Table A1: SSC1 Sequence Sources

Taxon	Genome Sequence Source	Genome Coordinates	Additional Sequence References
Saccharomyces cerevisiae RM11-1a	(Wall)		

Table A1: SSC1 Sequence Sources

Taxon	Genome Sequence Source	Genome Coordinates	Additional Sequence References
Saccharomyces cerevisiae RM11 1a	SGD	505092-507059 -, cont. 1.67	GenBank accession AAEG01000002
Saccharomyces cerevisiae YJM789	SGD	176054-178027 -, chrm X, cont. 59	GenBank accession AAFW02000040
Saccharomyces paradoxus NRRL Y-17217	SGD	69197-71152 +, cont. 301	GenBank accession AABY01000022
Saccharomyces mikatae IFO 1815 YM4906	SGD	4110-6089 -, cont. 326	GenBank accession AABZ01000329
Saccharomyces bayanus 623-6C YM4911	SGD	19993-21960 -, cont. 02.670	GenBank accession AACG02000013
Saccharomyces castellii NRRL Y-12630	SGD	12871-14835 +, cont. 560	GenBank accession AACF010000150
Candida glabrata CB8 138	SGD	283299-285239 -, chrm I	GenBank accession NC006032
Candida lusitaniae	BROAD, *Candida* Database	42250-44184 +, supercont. 5	locus CLUG_04122.1
Candida guilliermondii	BROAD, *Candida* Database	350555-352801 +, supercont. 5	locus PGUG_04143.1
Debaryomyces hansenii	BROAD, *Candida* Database	727170-729113 +, chrm 3	locus DEHAOCO8877g
Candida parapsilosis	BROAD, *Candida* Database	154581-156536 -, supercont. 130	locus CPAG_03670.0
Candida tropicalis	BROAD, *Candida* Database	1395100-1397040 +, supercont. 2	locus CTRG_01722.3
Candida dubliniensis CD36	Wellcome Trust Sanger Institute	1509835-1511778, cont. CHR2.070111	
Candida albicans SC5314	BROAD, *Candida* Database	1505513-1507459 +, chrm 2	locus orf19_1896
Aspergillus nidulans FGSCA4	BROAD *Aspergillus* Comparative Database	5484-7604 +, chrm I, cont. 103	locus AN601.3
Aspergillus clavatus NRRL1	SGD	378955-378512 -, cont. 72	locus ACLA_069020
Aspergillus fumigatus Af293	SGD	2551872-2554095 -, chrm 2	locus Afu2g09960
Aspergillus niger CBS 513.88	SGD	25361-27143 +, cont. An16c0180	GenBank accession NW 001594378
Aspergillus terreus NIH2624	BROAD *Aspergillus* Comparative Database	524851-527117 -, supercont. 6	locus ATEG044321
Aspergillus oryzae	BROAD	1645256-1647468 +, supercont. 22	locus AO090011000638
Aspergillus flavus	BROAD *Aspergillus* Comparative Database	1713562-1716206 +, cont. 5	locus AFL26_05388.2
Neosartorya fischeri	BROAD *Aspergillus* Comparative Database	3916569-3916154 -, cont. 508	locus NFIA_085400
Podospora anserina	*Podospora anserina* Genome Project*	1440741-1442600 +, chrm 6_SC2	
Trichoderma reesei	JGI, T. reesei, v2.0	461807-462013 -, 461705-459908 -, scaffold 2	
Fusarium solani	JGI, N. haematococca** v2.0	1561597-1561623 -, 1561337-1561538 -, 1559479-1561285 -, scaffold 18, chrm 2	
Fusarium graminearum	BROAD *Fusarium* Comparative Database	4480983-4483396 +, chrm 3, supercont. 3	locus FGSG_06154.3
Fusarium verticillioides	BROAD *Fusarium* Comparative Database	905127-908495 +, chrm 2, supercont. 7	locus FVEG_06113.3
Fusarium oxysporum f. sp. *lycopersici*	BROAD *Fusarium* Comparative Database	1157999-11 57683 -,chrm 2a, supercont. 10	locus FOXG_08555.2
Neuospora crassa OR74A	SGD	heat shock 70 kDa protein partial mRNA	Genbank accession XM 956660.2
		173415-175269 - (99% identity to mRNA)	Genbank accession NW_047290.1

SGD = *Saccharomyces* Genome Database
BROAD = BROAD Institute
JGI = Joint Genome Institute
*Génoscope
"+" or "-" denote chromosome strand
** The fungus *Nectria haematococca* Mating Population VI (MPVI) is also commonly referred to by its asexual name *Fusarium solani*

Table A2: SSQ1 Sequence Sources

Taxon	Genome Sequence Source	Genome Coordinates	Additional Sequence References
Saccharomyces cerevisiae RM11 1a	SGD	390448-392421 +, cont. 1.75	GenBank accession AAEG01000105
Saccharomyces cerevisiae YJM789	SGD	130933-132906 +, chrm XII, cont. 610	GenBank accession AAFW02000171
Saccharomyces paradoxus NRRL Y-17217	SGD	52252-54225 -, cont. 152	GenBank accession AABY01000020
Saccharomyces mikatae IFO 1815 YM4906	SGD	21530-23494 -, cont. 1173	GenBank accession AABZ01000080
Saccharomyces bayanus 623-6C YM4911	SGD	12085-14061 -, cont. 02.545	GenBank accession AACG02000134
Saccharomyces castellii NRRL Y-12630	SGD	8679-10634 -, cont. 670	GenBank accession AACF01000049
Candida glabrata CB8 138	SGD	470710-472650 +, chrm G	GenBank accession NC 006030
Candida lusitaniae	BROAD, *Candida* Database	1315749- 1317701 -, supercont. 4	locus CLUG_03878.1
Candida guilliermondii	BROAD, *Candida* Database	1055829-1057802 +, supercont. 5	locus PGUG_04519.1
Debaryomyces hansenii	BROAD, *Candida* Database	1354668-1356635 +, supercont. 7	locus DEHA0617688a
Candida parapsilosis	BROAD, *Candida* Database	220038-221975 -, supercont. 139	locus CPAG 04753.0
Candida tropicalis	BROAD, *Candida* Database	759503-161413 -, supercont. 7	locus CTRG 05157.3
Candida dubliniensis CD36	Wellcome Trust Sanger Institute	922410-924326 +, cont. chr7.070112	
Candida albicans SC5314	BROAD, *Candida* Database	866159-868075 +, supercont. 7	locus orf19.7179

SGD = *Saccharomyces* Genome Database
BROAD = BROAD Institute
JGI = Joint Genome Institute
'+' or '-' denote chromosome strand

Table A3: JAC1 Sequence Sources

Table A3: JAC1 Sequence Sources

Taxon	Genome Sequence Source	Genome Coordinates	Additional Sequence References
Saccharomyces cerevisiae RM11 1a	SGD	374283-37837, +, cont. 1.20	GenBank accession AAEG01000073
Saccharomyces cerevisiae YJM789	SGD	54670-55224, -, chrm VII, cont. 154	GenBank accession AAFW02000102
Saccharomyces paradoxus NRRL Y-17217	SGD	2445-2999, +, cont. 17	GenBank accession AABY01000209
Saccharomyces mikatae IFO 1815 YM4906	SGD	3454-4007, +, cont. 2267	GenBank accession AACH01000574
Saccharomyces bayanus 623-6C YM4911	SGD	15477-16030, +, cont. 74	GenBank accession AACA01000186
Saccharomyces castellii NRRL Y-12630	SGD	13997-14563, -, cont. 638	GenBank accession AACF01000083
Candida glabrata CBS 138	SGD	30179-30775, -, chrm A	GenBank accession NC 005967
Candida lusitaniae	BROAD, *Candida* Database	1581070-1581736 +, supercont. 1.1	
Candida guilliermondii	BROAD, *Candida* Database	1569083-1569820 +, supercont. 2	locus PGUG02082.1
Debaryomyces hansenii	BROAD, *Candida* Database	567026-567721 -, chrm A, supercont. 1	locus DEHA0A07062a
Candida parapsilosis	BROAD, *Candida* Database	215628-216281 -, cont. 138	locus CPAG_04593.0
Candida tropicalis	BROAD, *Candida* Database	9596-10228 -, supercont. 2	locus CTRG01083.3
Candida dubliniensis CD36	Wellcome Trust Sanger Institute	39818-40453 +, cont. chr2.070111	
Candida albicans SC5314	BROAD, *Candida* Database	36604-37233 +, chrm 2, supercont. 2	locus orf19.2104
Aspergillus nidulans FGSCA4	SGD	195806-196645, +, chrm VII, cont. 1.22	GenBank accession 101263
Aspergillus clavatus NRRL1	BROAD *Aspergillus* Comparative Database	277406-278230 +, supercont. 17	locus ACLA_057970
Aspergillus fumigatus A1163	SGD	79602-80414 -, cont. 000065	GenBank accession ABDB01000065
Aspergillus niger ATCC 1015	BROAD *Aspergillus* Comparative Database	247056-247844 +, supercont. 18	locus gw1_18.175
Aspergillus terreus NIH2624	SGD	2801-3616 +, cont. 1.2	GenBank accession NW 001471411
Aspergillus oryzae	BROAD *Aspergillus* Comparative Database	812228-813046 -, supercont. 8	locus AO090023000319
Aspergillus flavus	BROAD *Aspergillus* Comparative Database	838541-839359 -, supercont. 4	locus AFL2604175.2
Neosartorya fischeri	BROAD *Aspergillus* Comparative Database	1106608-1107423 +, supercont. 578	locus NFIA097400
Podospora anserina	*Podospora anserina* Genome Project*	1439869-1440723 +, cont. chrm1_SC4	
Trichoderma reesei	JGI, T. reesei, v2.0	1483190-1484089 +, scaffold 3	
Fusarium solani	JGI, N. haematococca * v2.0	1218284-1218499, 1217708-1218262 +, scaffold 1_chr1_3_0	
Fusarium graminearum	BROAD *Fusarium* Comparative Database	3338944-3339729 +, chrm 1, supercont. 1	locus FGSG01028.3
Fusarium verticillioides	BROAD *Fusarium* Comparative Database	3440040-3440822 +, chrm 1, supercont. 1	locus PVEG 01154.3
Fusarium oxysporum f. sp. *lycopersici*	BROAD *Fusarium* Comparative Database	1181211-1181993 -, chrm 1, supercont. 1	locus FOXG00364.2
Neuospora crassa OR74A	SGD	23913-24842 -	GenBank accession NW_047290.1

SGD = *Saccharomyces* Genome Database
BROAD = BROAD Institute
JGI = Joint Genome Institute
*Génoscope
"+" or "-" denote chromosome strand
** The fungus *Nectria haematococca* Mating Population VI (MPVI) is also commonly referred to by its asexual name *Fusarium solani*

Appendix B

Fungal Mitochondrial Heat Shock Protein Multiple Sequence Alignments

Multiple alignments of amino acid sequences translated from protein-coding regions of mitochondrial heat shock proteins (mtHsps) were performed using CLUSTAL W (Thompson et al. 1994) with default gap penalties, and subsequent manual trimming to remove gaps. Alignment columns highlighted in black denote sites sharing 100% identity among all taxa. Taxon name abbreviations used are listed in the table below:

Taxon Abbreviation	Fungal Species	Taxon Abbreviation	Fungal Species
Scer_Y	*Saccharomyces cerevisiae* RM11	Fgra	*Fusarium graminearum*
Scer_R	*Saccharomyces cerevisiae* YJM789	Fver	*Fusarium verticilliodes*
Spar	*Saccharomyces paradoxus*	Fsol	*Fusarium solani*
Smik	*Saccharomyces mikatae*	Ncra	*Neurospora crassa*
Sbay	*Saccharomyces bayanus*	Tree	*Trichoderma reesei*
Scas	*Saccharomyces castellii*	Pans	*Podospora anserina*
Cgla	*Candida glabrata*	Nfis	*Neosartorya fischeri*
Calb	*Candida albicans*	Anid	*Aspergillus nidulans*
Ctro	*Candida tropicalis*	Ater	*Aspergillus terreus*
Cpar	*Candida parapsilosis*	Acla	*Aspergillus clavatus*
Cgui	*Candida guilliermondii*	Afla	*Aspergillus flavus*
Cdub	*Candida dubliniensis*	Anig	*Aspergillus niger*
Clus	*Candida lusitaniae*	Aory	*Aspergillus oryzae*
Dhan	*Debaryomyces hansenii*	Afum	*Aspergillus fumigatus*
Foxy	*Fusarium oxysporum*		

Scer_R
Scer_Y
Cgla
Spar
Smik
Sbay
Scas
Dhan
Calb
Cgui
Ctro
Cpar
Cdub
Clus
Foxy
Fgra
Fver
Ncra
Tree
Pans
Psol
Nfis
Anid
Ater
Acla
Afla
Anig
Aory
Afum

Figure

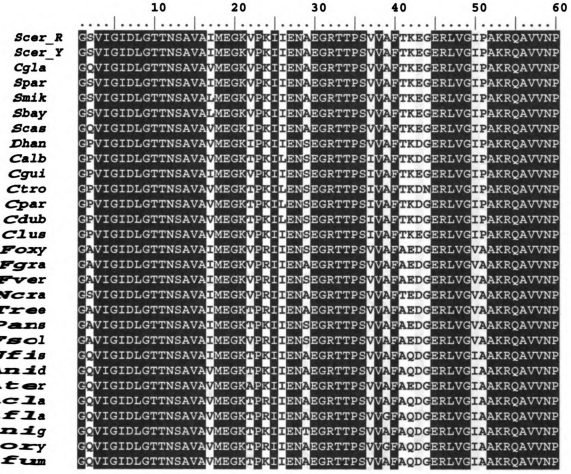
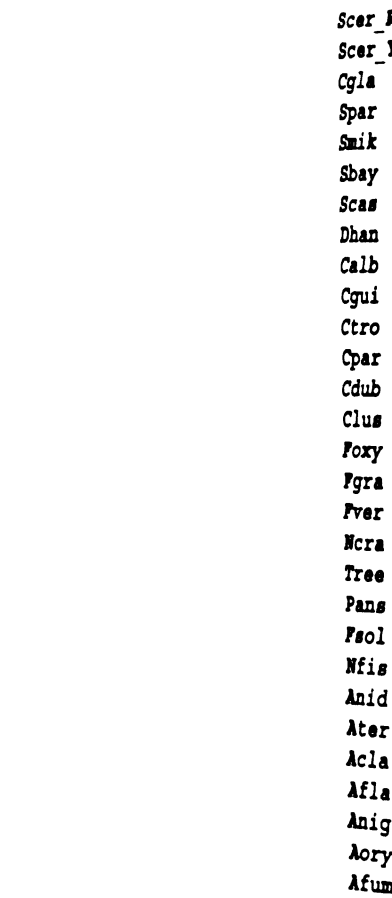


Figure B1: SSC1 amino acid multiple sequence alignment


Scer_
Scer_
Cgla
Spar
Smik
Sbay
Scas
Dhan
Calb
Cgui
Ctro
Cpar
Cdub
Clus
Foxy
Fgra
Fver
Ncra
Tree
Pans
Fsol
Nfis
Anid
Ater
Acla
Afla
Anig
Aory
Afum

Figu

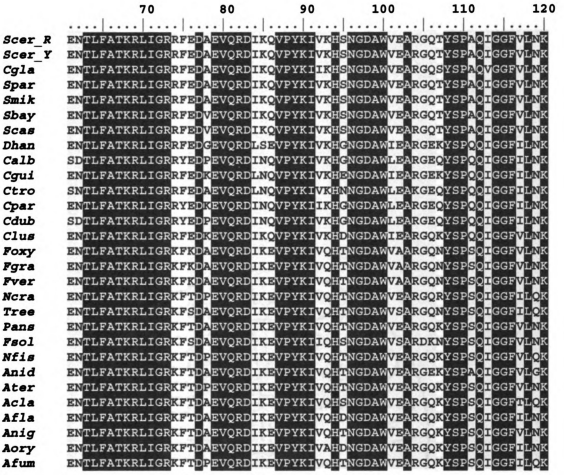


Figure B1: SSC1 amino acid multiple sequence alignment (continued)

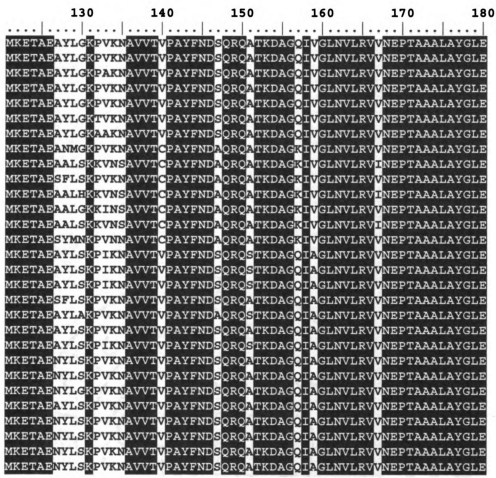
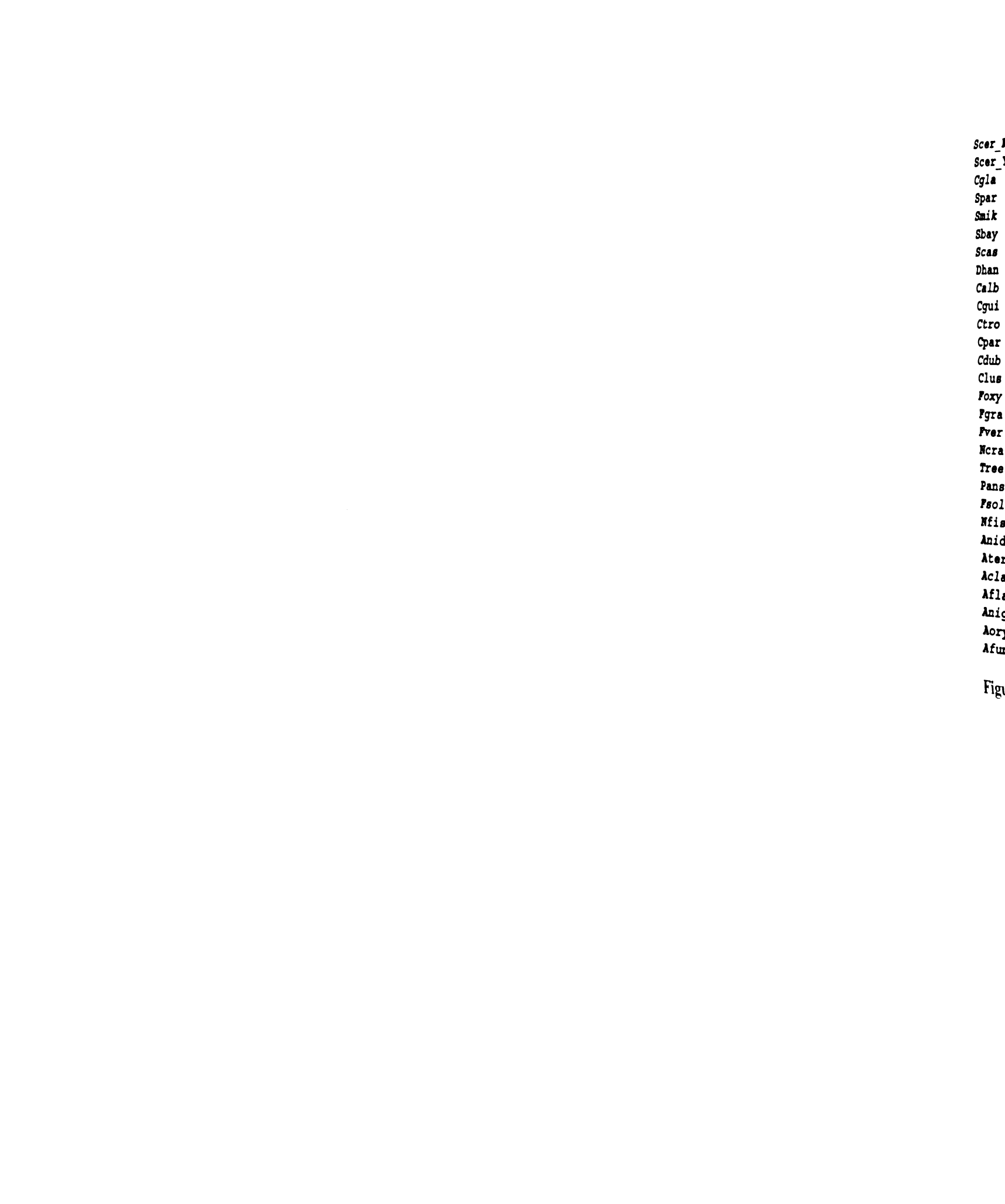


Figure B1: SSC1 amino acid multiple sequence alignment (continued)


Scer_I
Scer_I
Cgla
Spar
Smik
Sbay
Scas
Dhan
Calb
Cgui
Ctro
Cpar
Cdub
Clus
Foxy
Fgra
Fver
Ncra
Tree
Pans
Fsol
Nfis
Anid
Ater
Acla
Afla
Anig
Aory
Afum

Figu

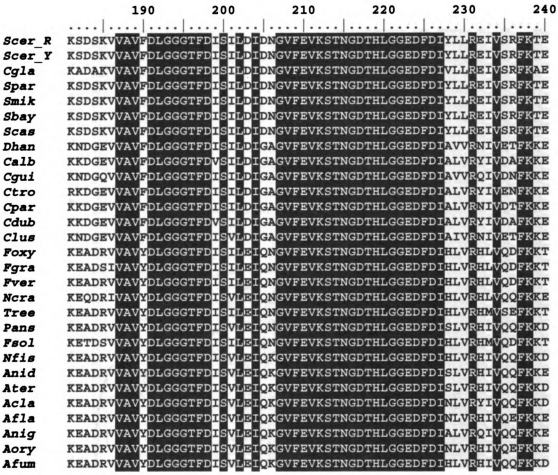


Figure B1: SSC1 amino acid multiple sequence alignment (continued)

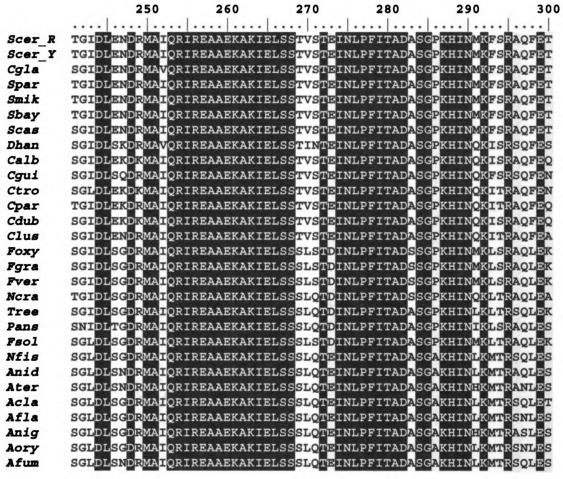


Figure B1: SSC1 amino acid multiple sequence alignment (continued)

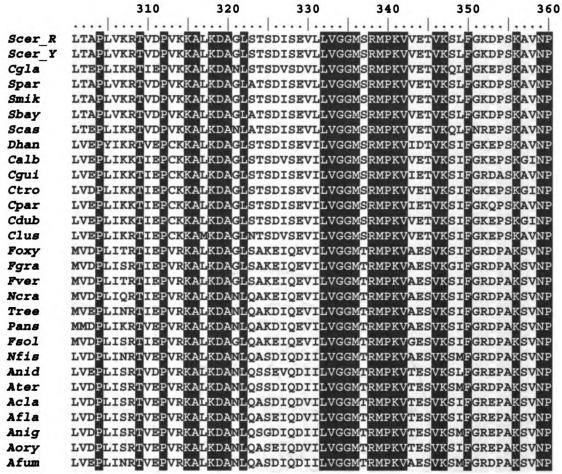


Figure B1: SSC1 amino acid multiple sequence alignment (continued)

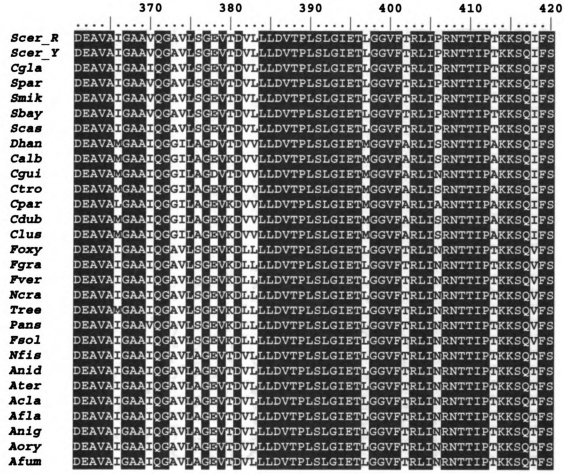


Figure B1: SSC1 amino acid multiple sequence alignment (continued)

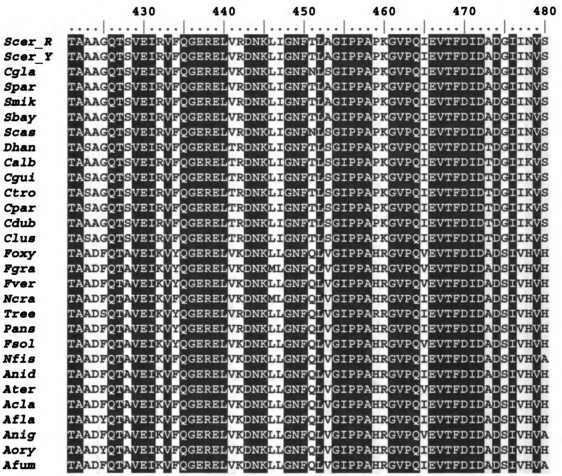


Figure B1: SSC1 amino acid multiple sequence alignment (continued)

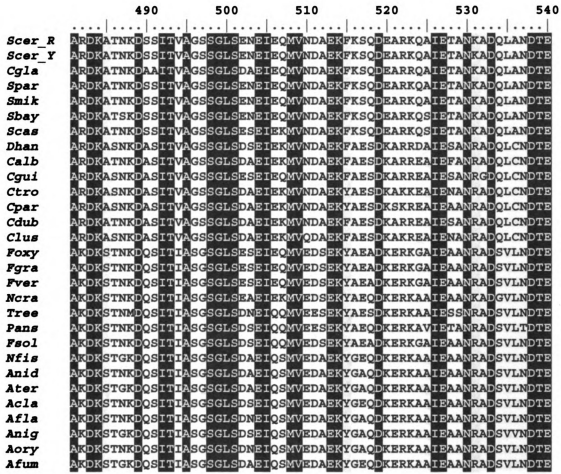


Figure B1: SSC1 amino acid multiple sequence alignment (continued)

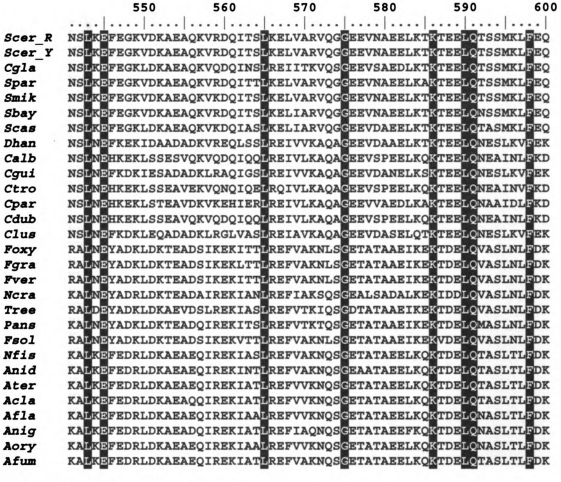


Figure B1: SSC1 amino acid multiple sequence alignment (continued)

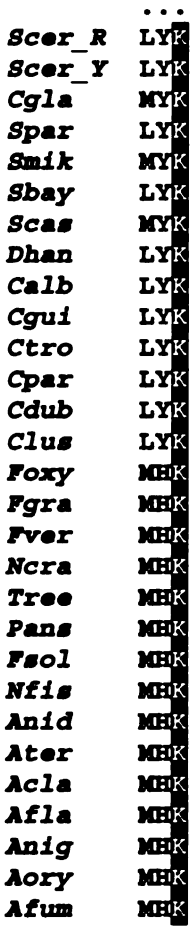


Figure B1: SSC1 amino acid multiple sequence alignment (continued)

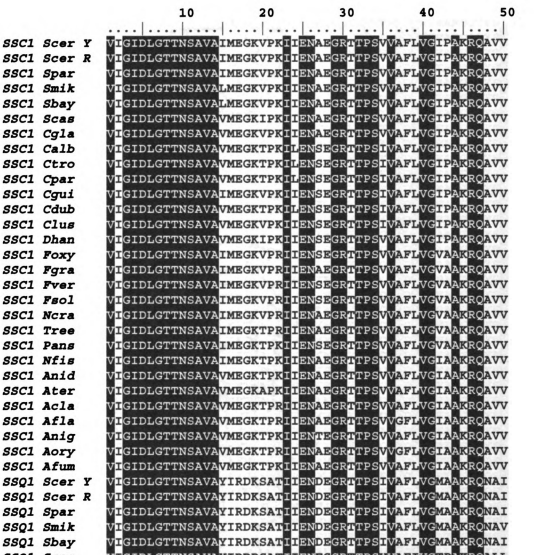


Figure B2: SSC1 and SSQ1 combined amino acid multiple sequence alignment

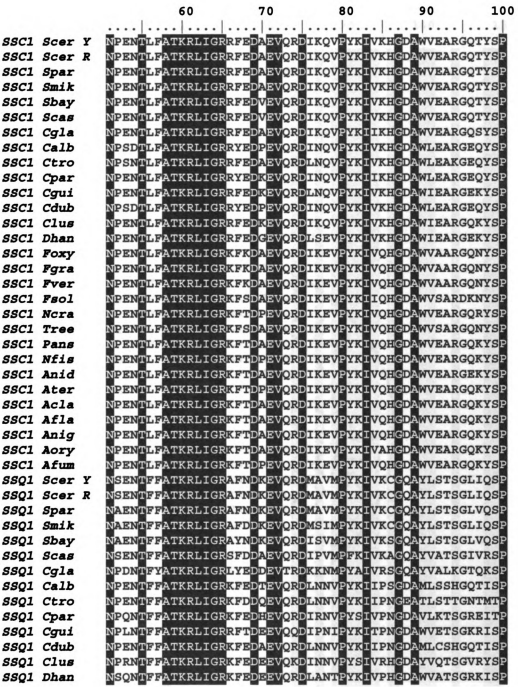


Figure B2: SSC1 and SSQ1 combined amino acid multiple sequence alignment (continued)

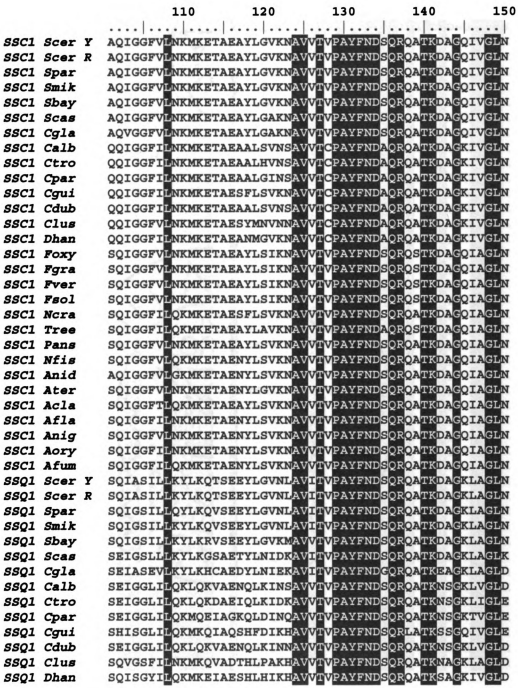


Figure B2: SSC1 and SSQ1 combined amino acid multiple sequence alignment (continued)

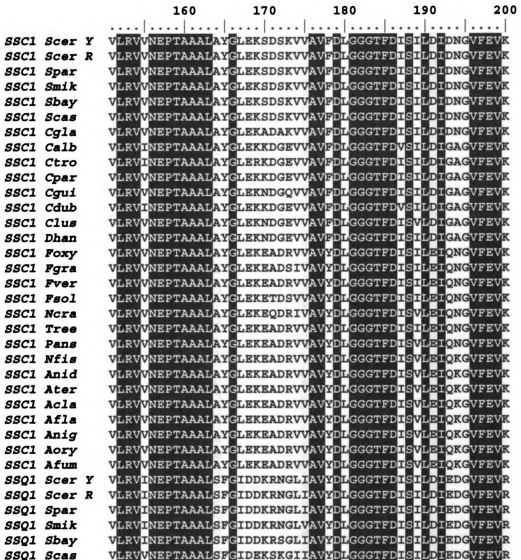


Figure B2: SSC1 and SSQ1 combined amino acid multiple sequence alignment (continued)

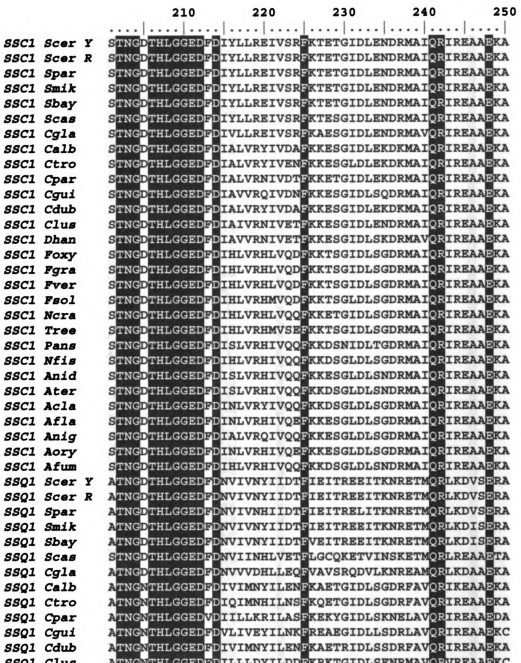


Figure B2: SSC1 and SSQ1 combined amino acid multiple sequence alignment (continued)

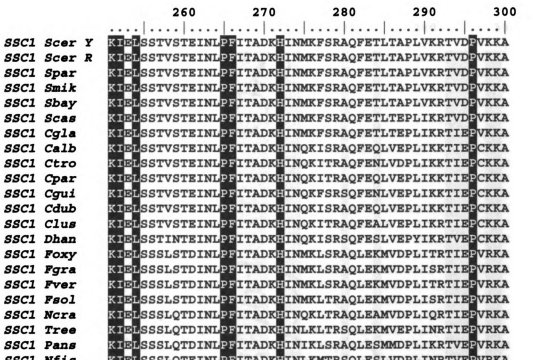


Figure B2: SSC1 and SSQ1 combined amino acid multiple sequence alignment (continued)

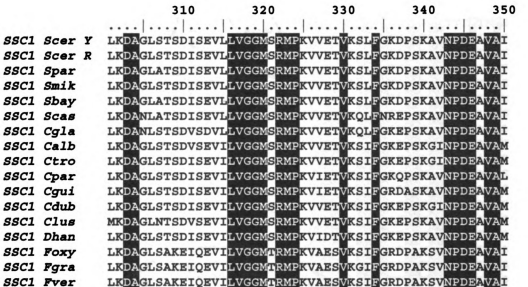


Figure B2: SSC1 and SSQ1 combined amino acid multiple sequence alignment (continued)

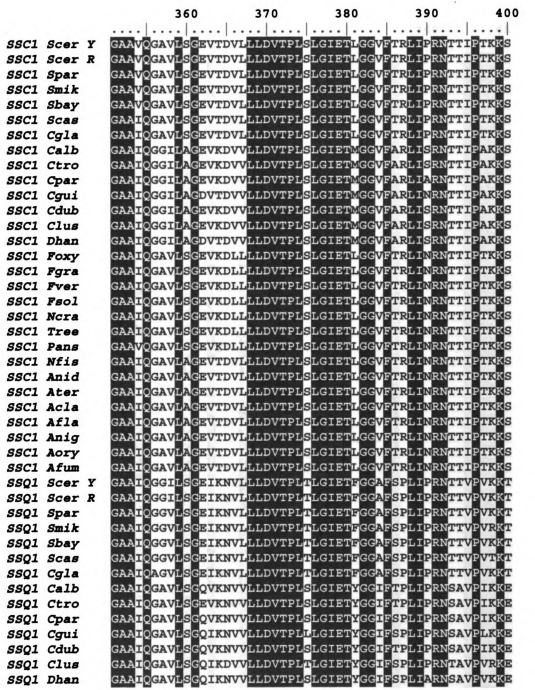


Figure B2: SSC1 and SSQ1 combined amino acid multiple sequence alignment (continued)

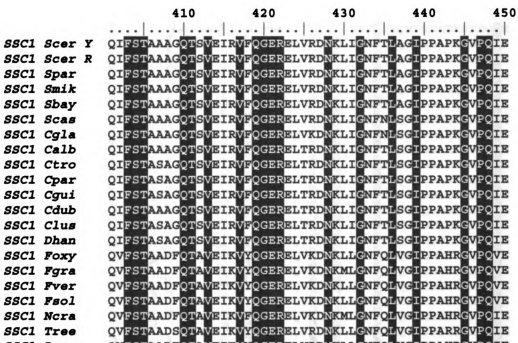


Figure B2: SSC1 and SSQ1 combined amino acid multiple sequence alignment (continued)

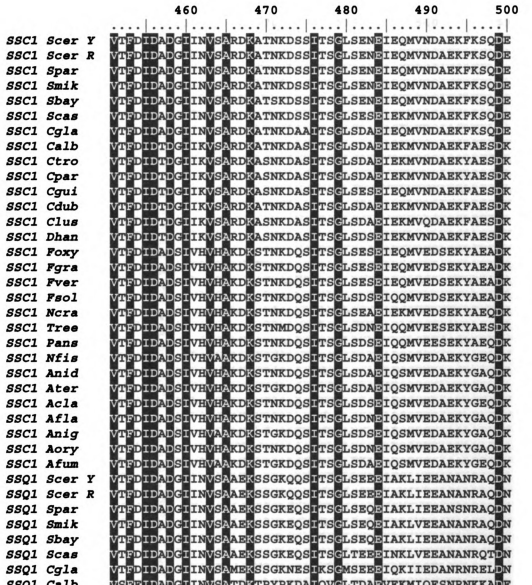


Figure B2: SSC1 and SSQ1 combined amino acid multiple sequence alignment (continued)

Figure B2: SSC1 and SSQ1 combined amino acid multiple sequence alignment (continued)

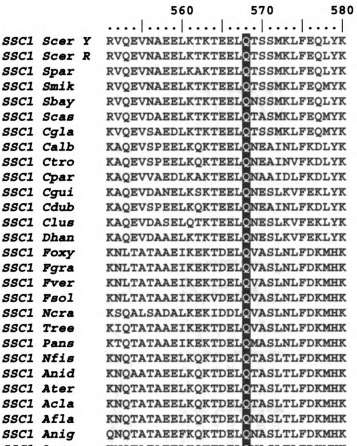


Figure B2: SSC1 and SSQ1 combined amino acid multiple sequence alignment (continued)

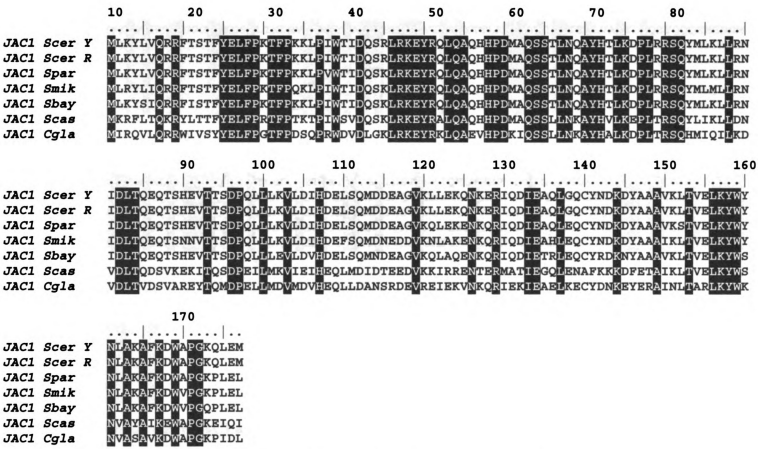


Figure B3: ***Saccharomyces*** **clade JAC1 amino acid multiple sequence alignment**

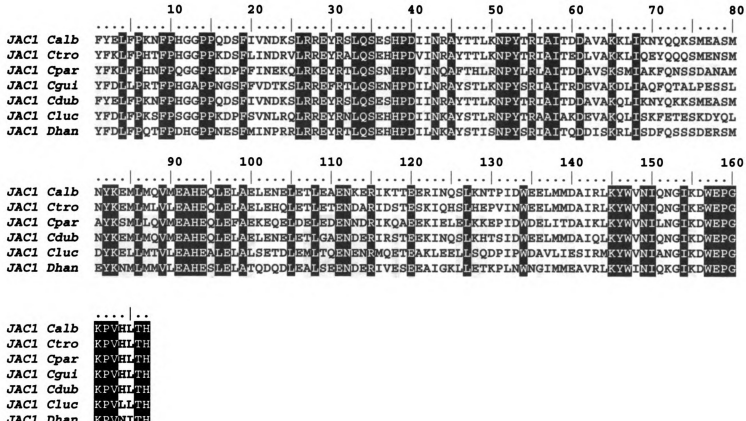


Figure B4: *Candida* clade JAC1 amino acid multiple sequence alignment

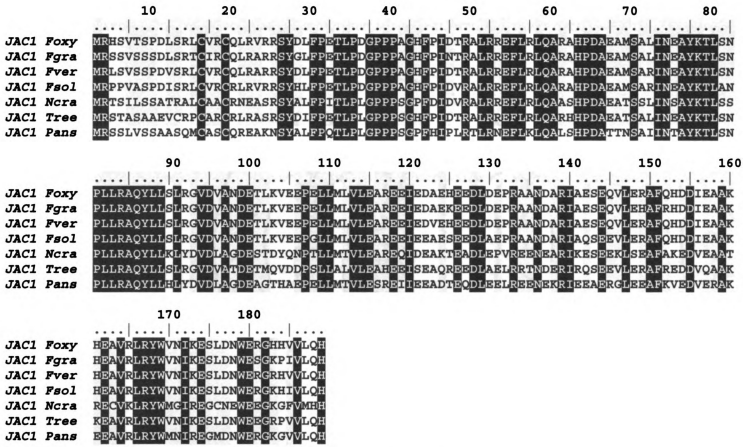


Figure B5: ***Fusarium*** **clade JAC1 amino acid multiple sequence alignment**

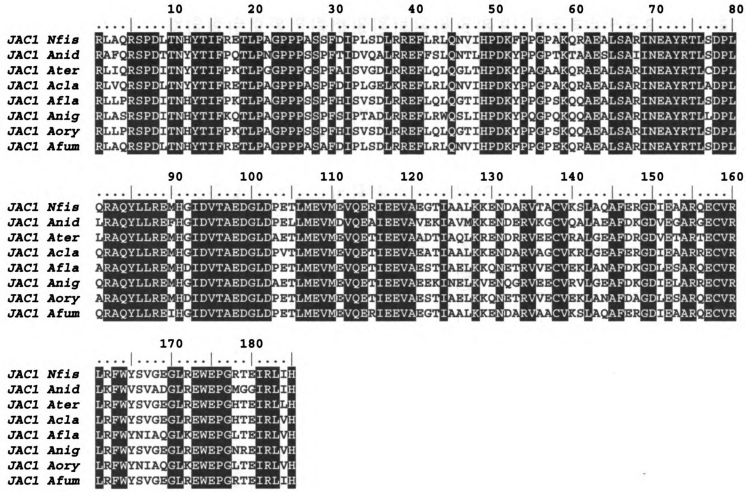


Figure B6: ***Aspergillus*** **clade JAC1 amino acid multiple sequence alignment**

Appendix C

Fungal Mitochondrial Heat Shock Protein Phylogenetic Gene Tree Input Topologies for codeml Evolutionary Rate Analysis

Tree topologies input into codeml represent composite structures of highly supported relationships from trees inferred using the following sequence partitions: 1st, 2nd, 3rd, 1st and 2nd nucleotide positions, all nucleotides, and amino acids. Maximum Parsimony, Maximum Likelihood, and Bayesian Inference methods were used. Branches were manually collapsed if bootstrap support or posterior probabilities were below 90% or 0.9, respectively. All topologically unique composite trees are shown.
Taxon name abbreviations used are listed in the table below:

Taxon Abbreviation	Fungal Species	Taxon Abbreviation	Fungal Species
Scer_Y	*Saccharomyces cerevisiae* RM11	Fgra	*Fusarium graminearum*
Scer_R	*Saccharomyces cerevisiae* YJM789	Fver	*Fusarium verticilliodes*
Spar	*Saccharomyces paradoxus*	Fsol	*Fusarium solani*
Smik	*Saccharomyces mikatae*	Ncra	*Neurospora crassa*
Sbay	*Saccharomyces bayanus*	Tree	*Trichoderma reesei*
Scas	*Saccharomyces castellii*	Pans	*Podospora anserina*
Cgla	*Candida glabrata*	Nfis	*Neosartorya fischeri*
Calb	*Candida albicans*	Anid	*Aspergillus nidulans*
Ctro	*Candida tropicalis*	Ater	*Aspergillus terreus*
Cpar	*Candida parapsilosis*	Acla	*Aspergillus clavatus*
Cgui	*Candida guilliermondii*	Afla	*Aspergillus flavus*
Cdub	*Candida dubliniensis*	Anig	*Aspergillus niger*
Clus	*Candida lusitaniae*	Aory	*Aspergillus oryzae*
Dhan	*Debaryomyces hansenii*	Afum	*Aspergillus fumigatus*
Foxy	*Fusarium oxysporum*		

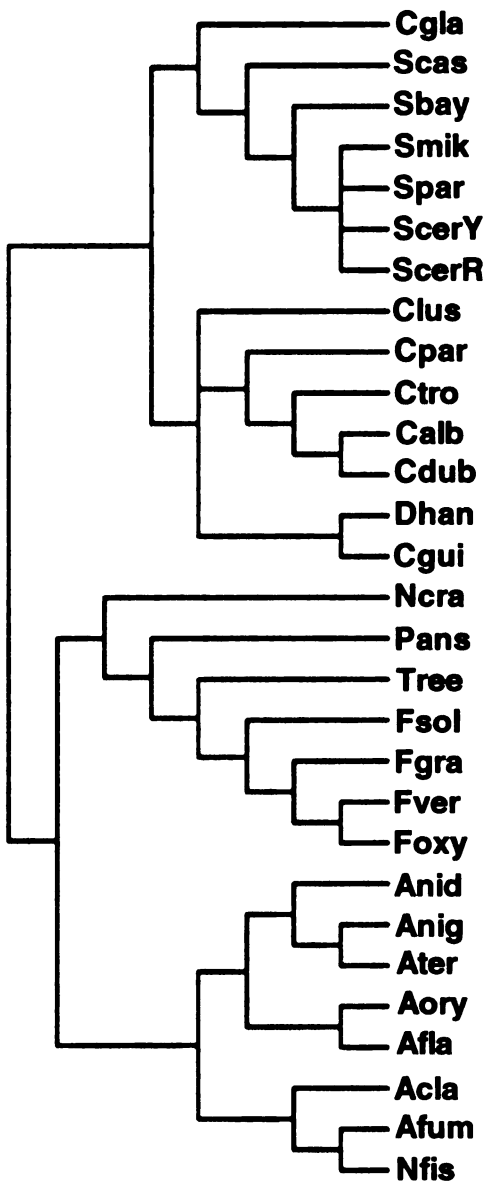


Figure C1: SSC1 Bayesian Inference Tree 1

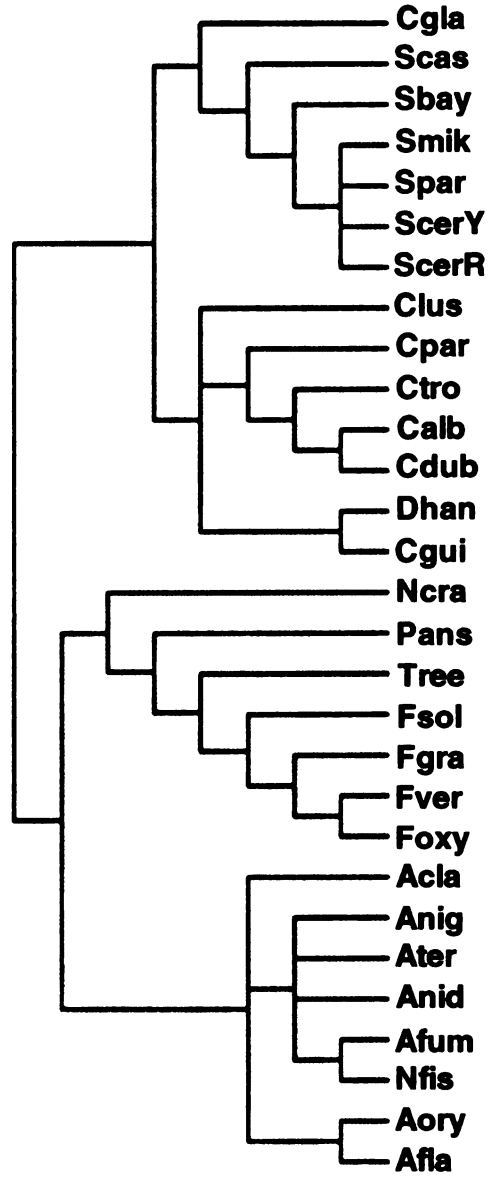


Figure C2: SSC1 Bayesian Inference Tree 2

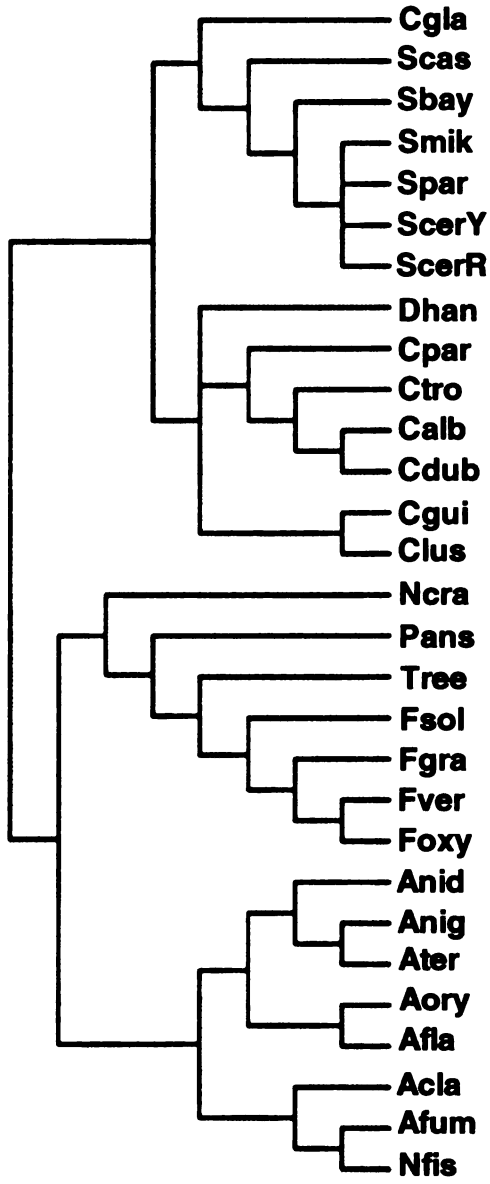


Figure C3: SSC1 Bayesian Inference Tree 3

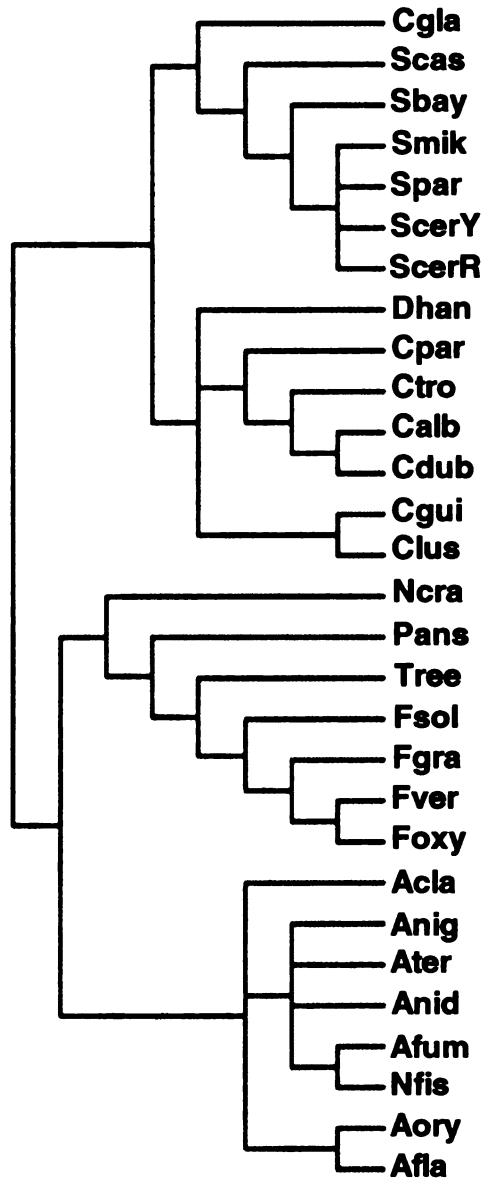


Figure C4: SSC1 Bayesian Inference Tree 4

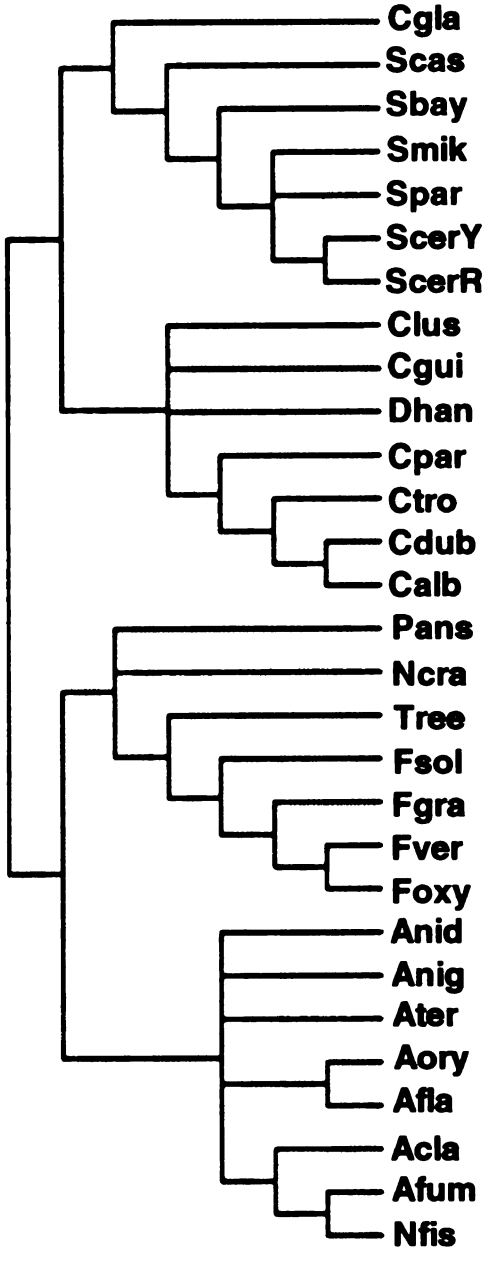


Figure C5: SSC1 Maximum Likelihood Tree 1

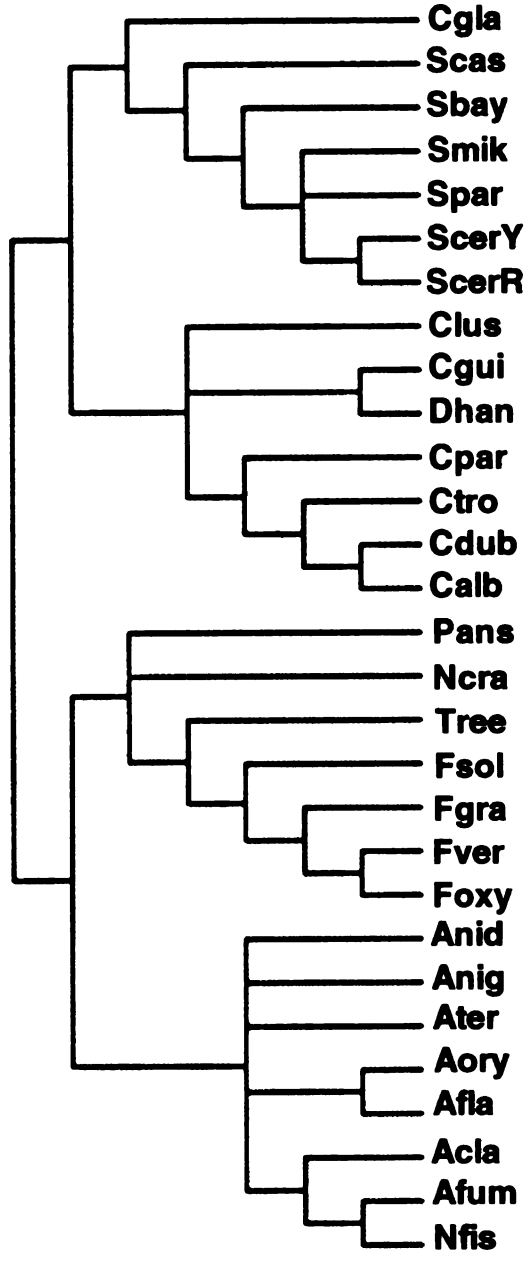


Figure C6: SSC1 Maximum Likelihood Tree 2

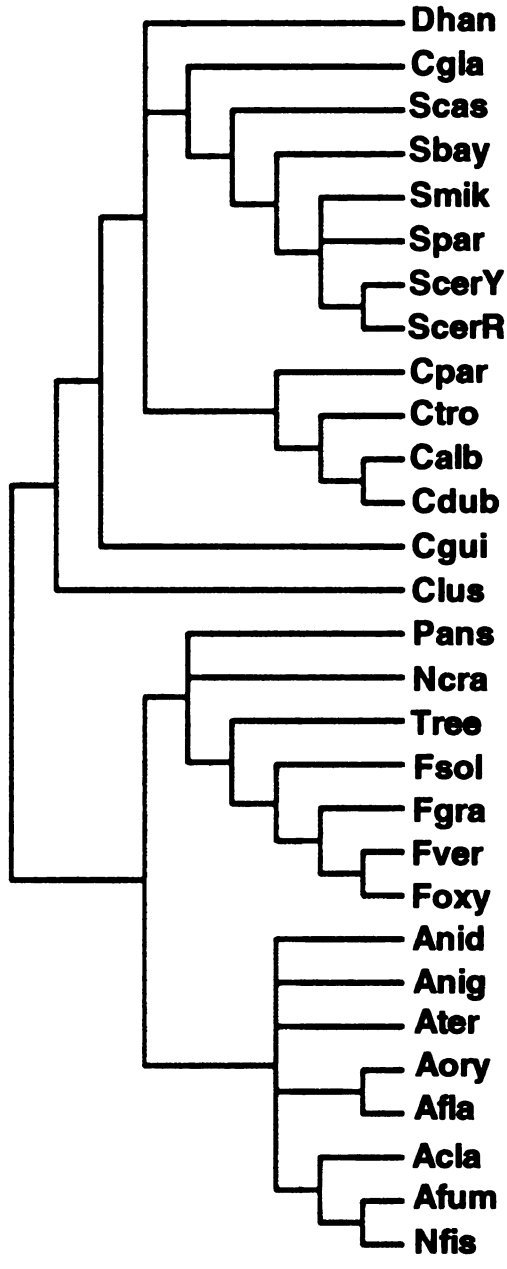


Figure C7: SSC1 Maximum Likelihood Tree 3

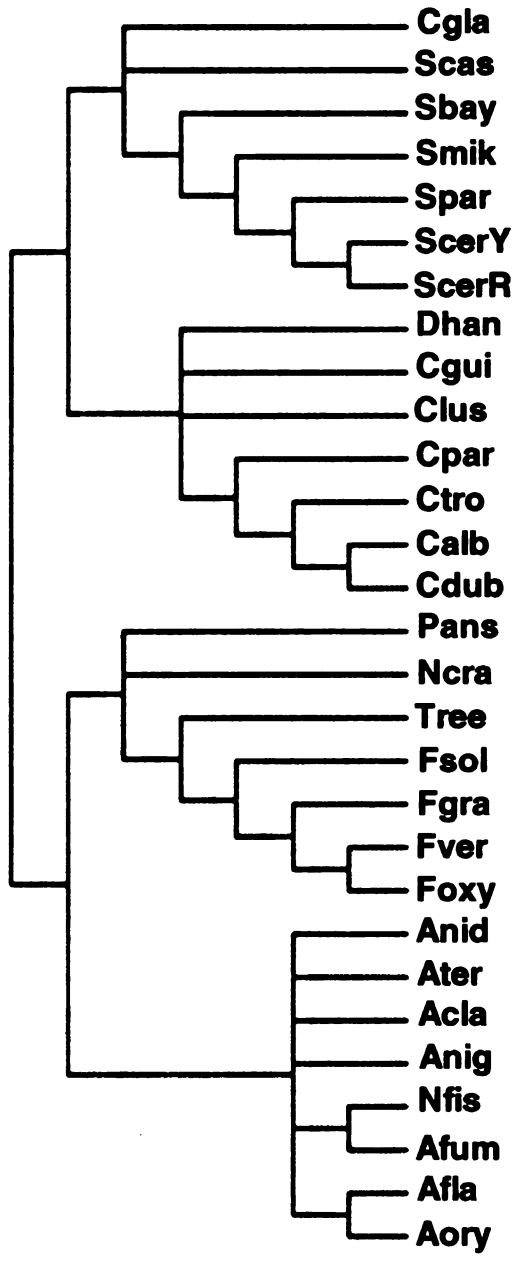


Figure C8: SSC1 Maximum Parsimony Tree 1

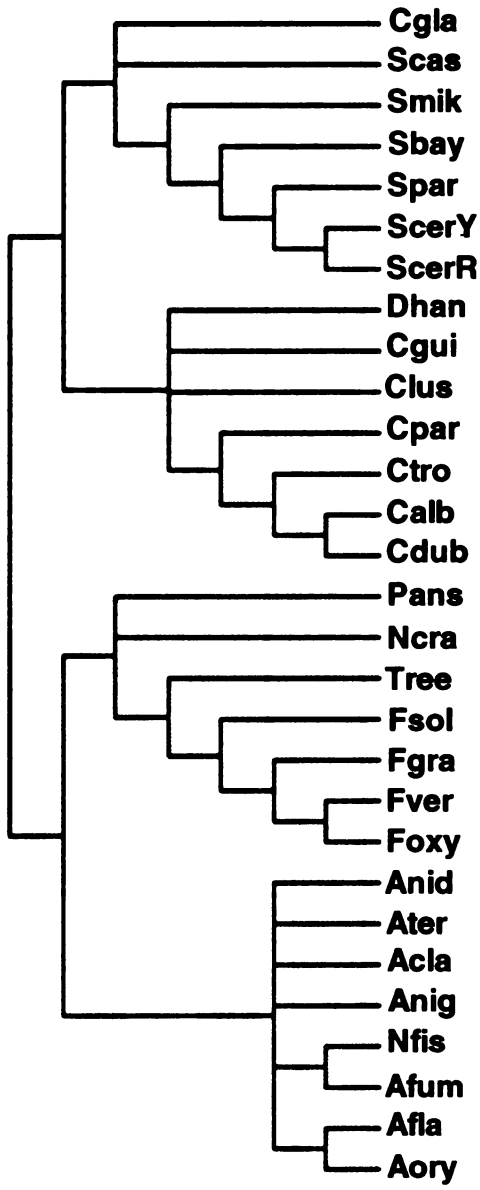


Figure C9: SSC1 Maximum Parsimony Tree 2

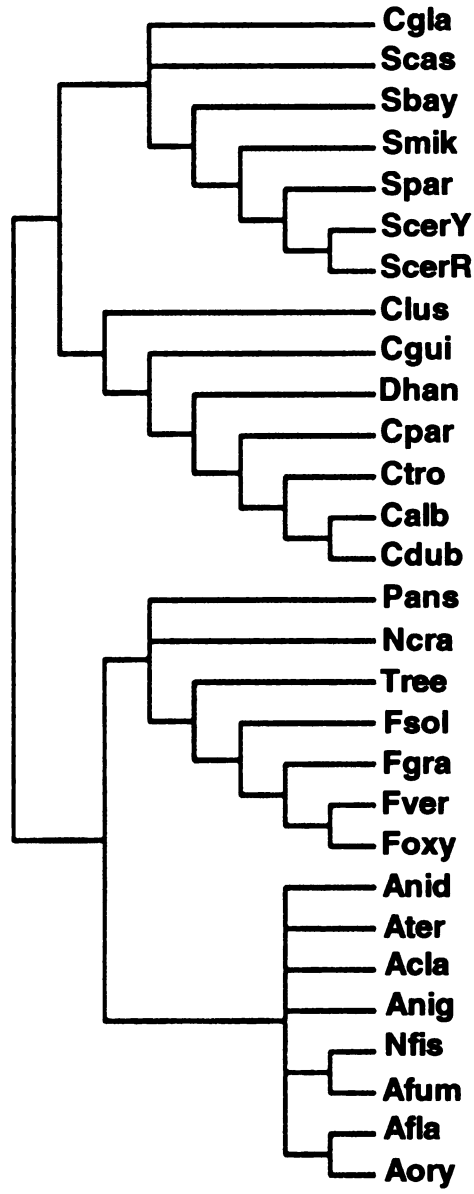


Figure C10: SSC1 Maximum Parsimony Tree 3

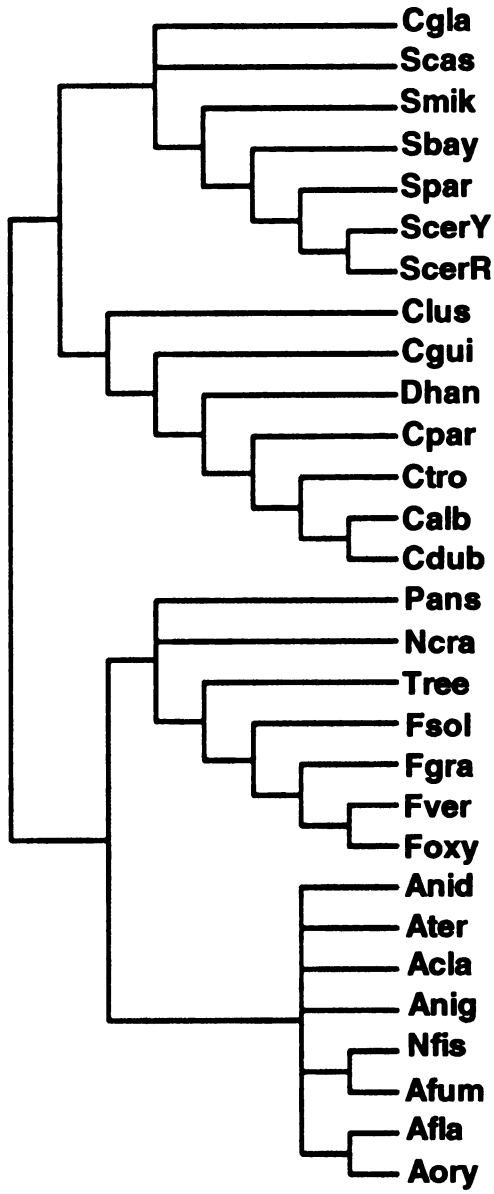


Figure C11: SSC1 Maximum Parsimony Tree 4

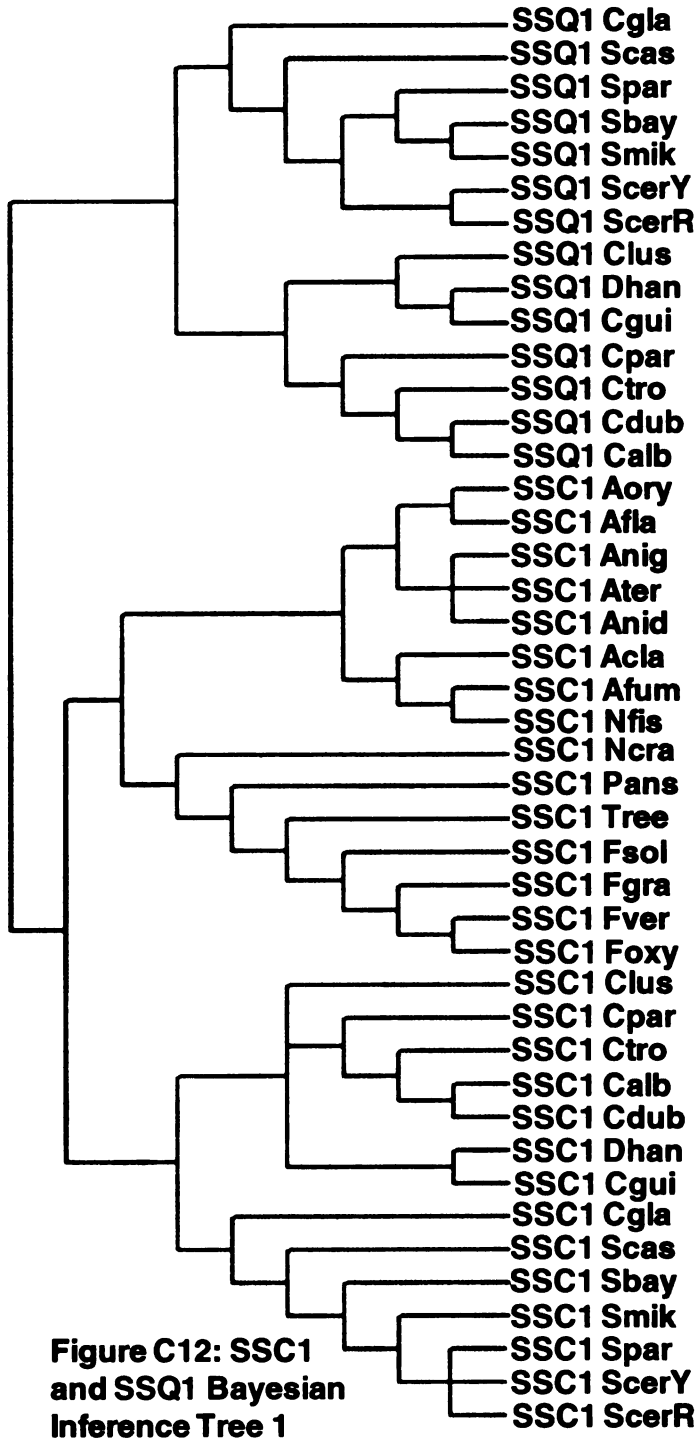


Figure C12: SSC1 and SSQ1 Bayesian Inference Tree 1

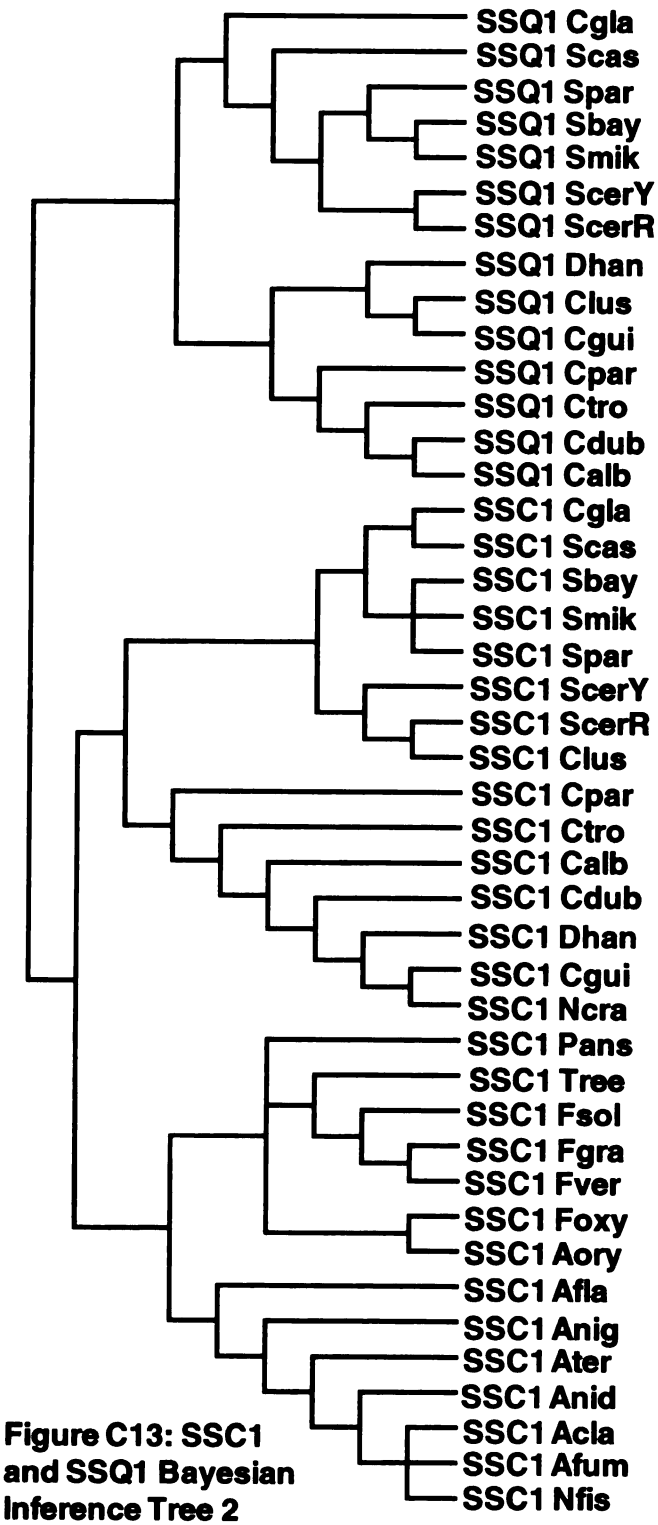


Figure C13: SSC1 and SSQ1 Bayesian Inference Tree 2

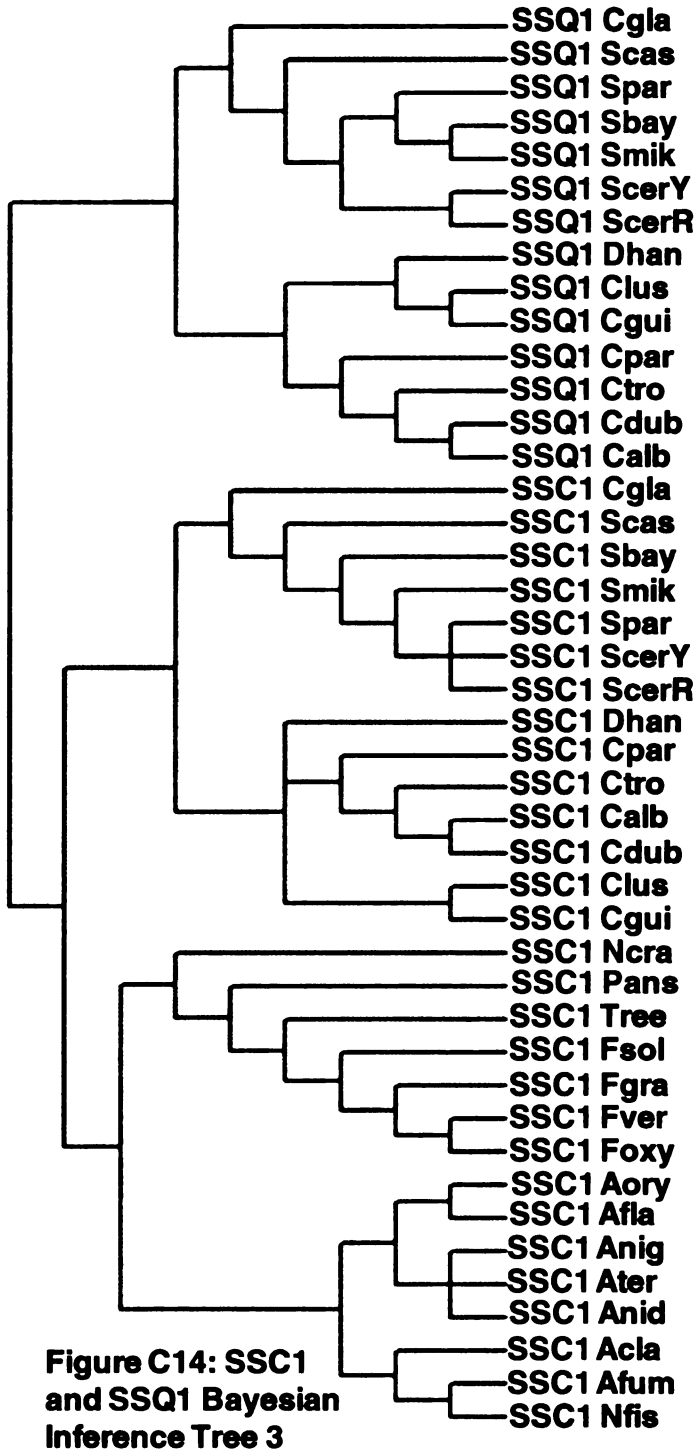


Figure C14: SSC1 and SSQ1 Bayesian Inference Tree 3

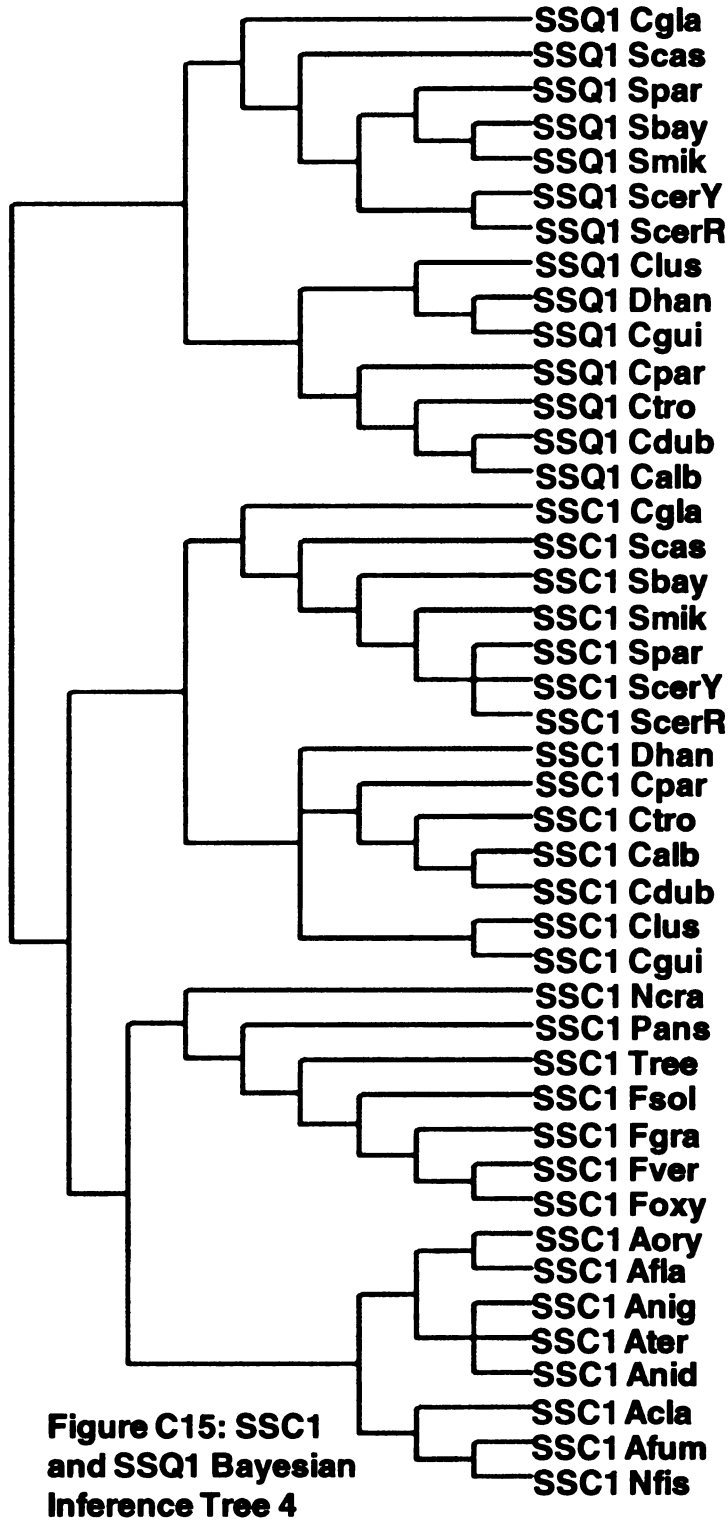


Figure C15: SSC1 and SSQ1 Bayesian Inference Tree 4

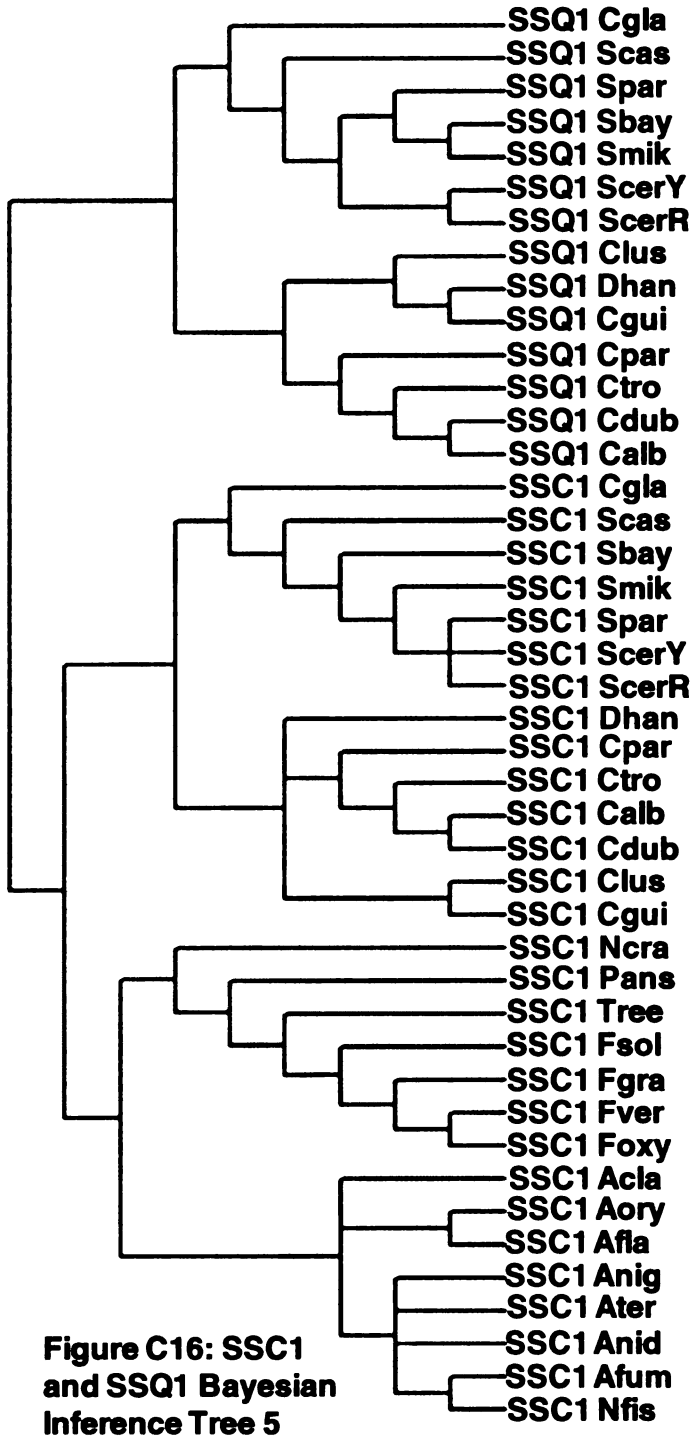


Figure C16: SSC1 and SSQ1 Bayesian Inference Tree 5

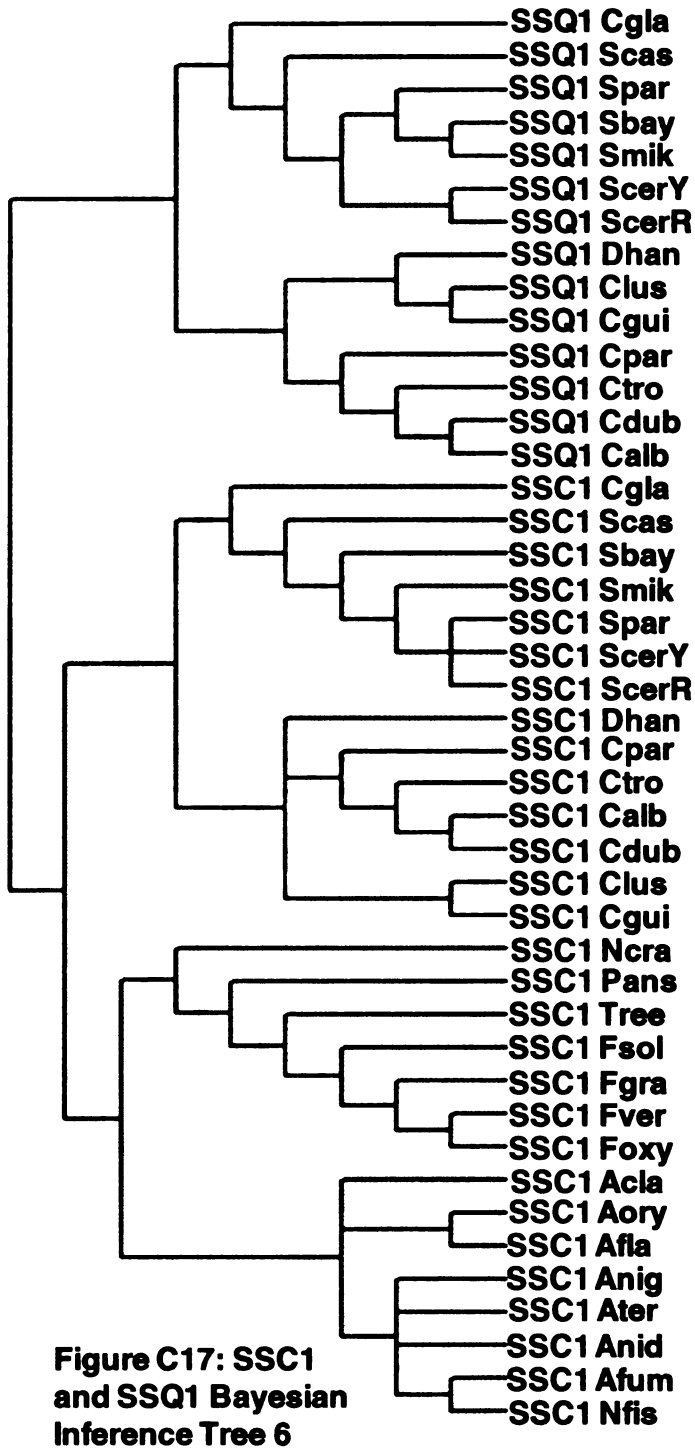


Figure C17: SSC1 and SSQ1 Bayesian Inference Tree 6

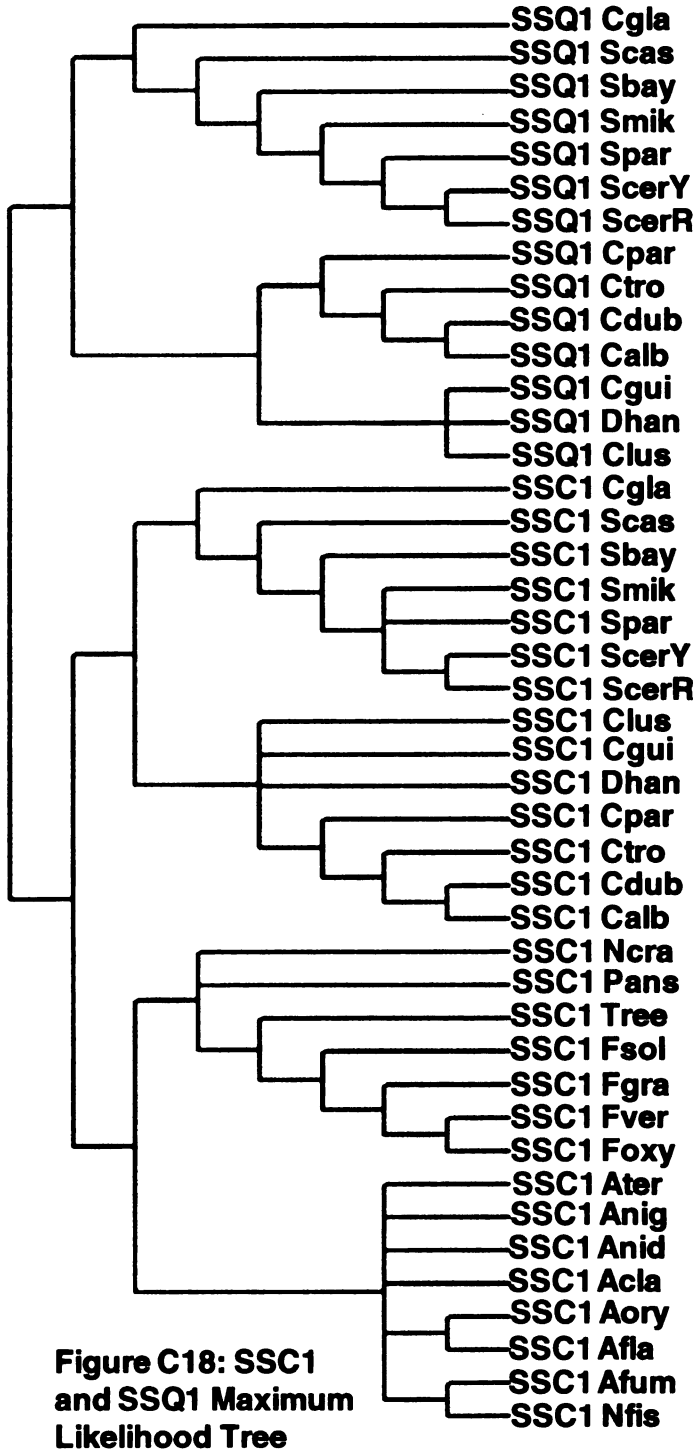


Figure C18: SSC1 and SSQ1 Maximum Likelihood Tree

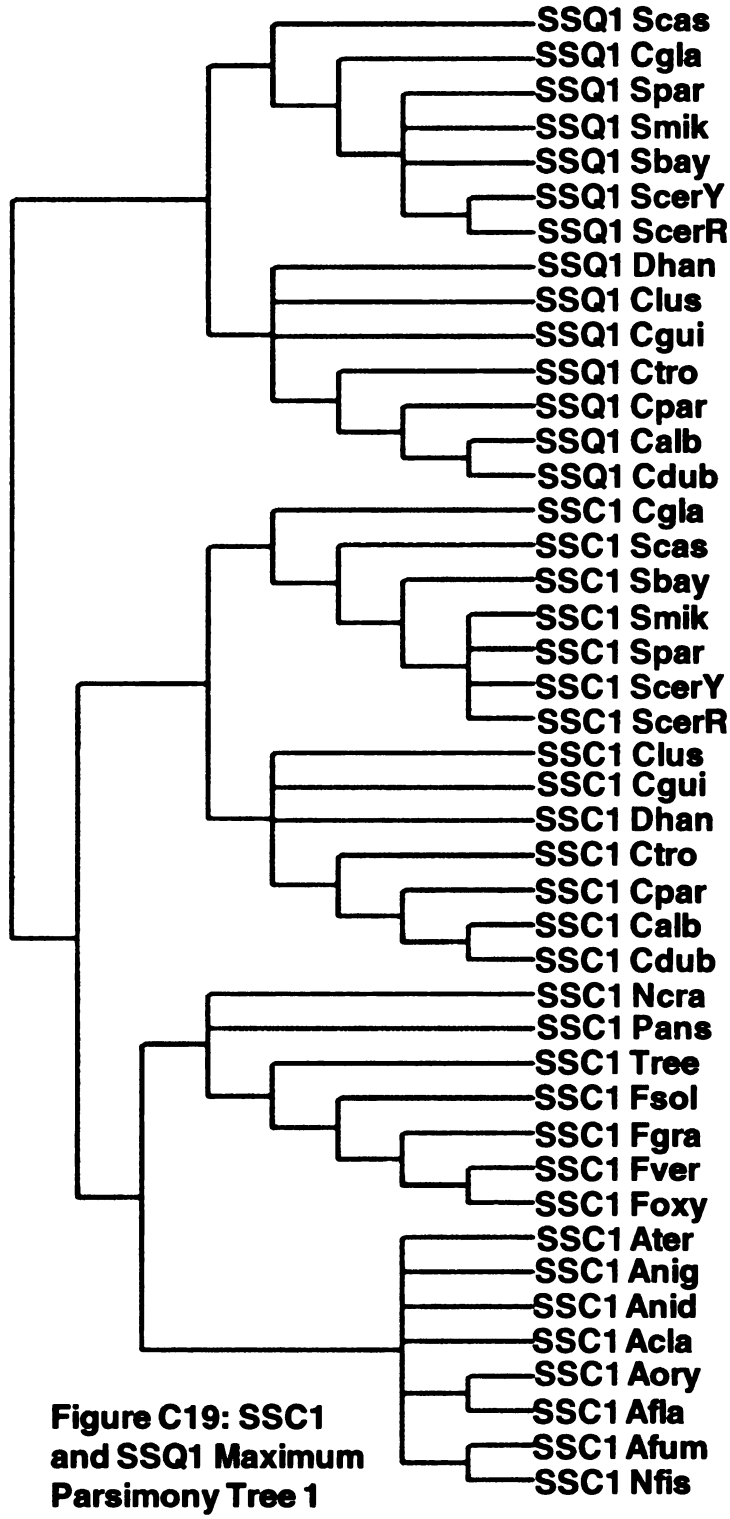


Figure C19: SSC1 and SSQ1 Maximum Parsimony Tree 1

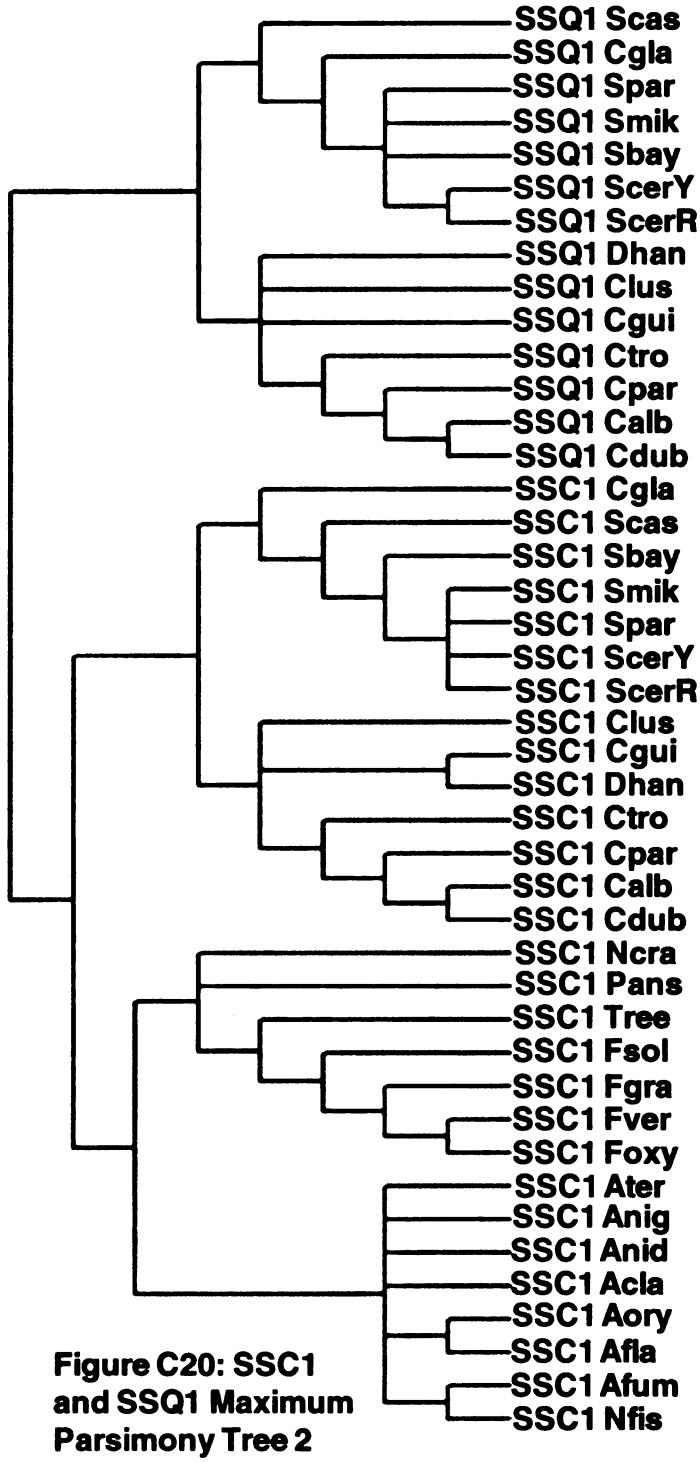


Figure C20: SSC1 and SSQ1 Maximum Parsimony Tree 2

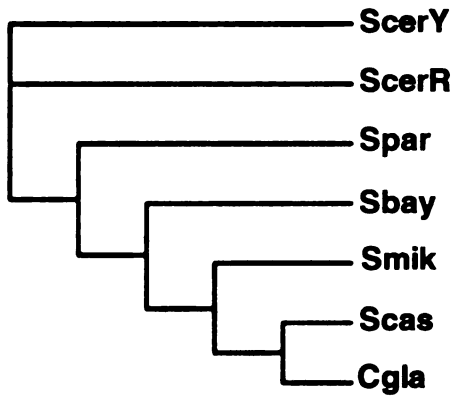


Figure C21: JAC1 *Saccharomyces* Bayesian Inference/ Maximum Parsimony Tree

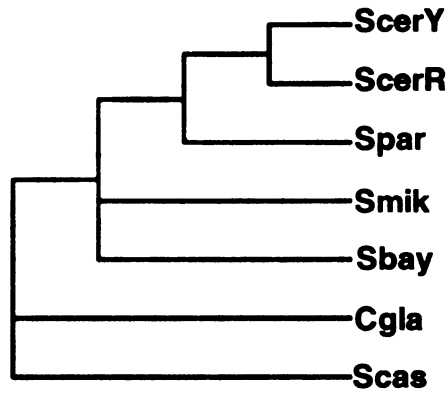


Figure C22: JAC1 *Saccharomyces* Maximum Likelihood Tree

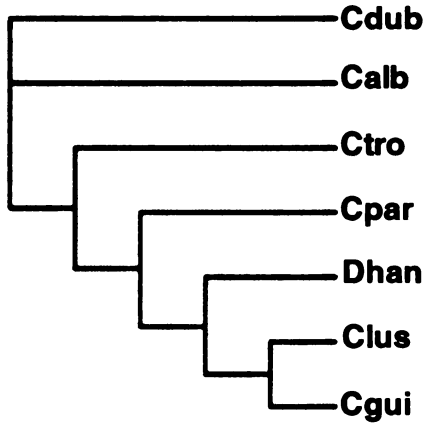


Figure C23: JAC1 *Candida* Bayesian Inference Tree

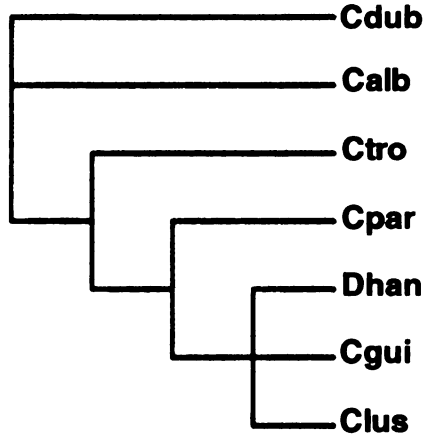


Figure C24: JAC1 *Candida* Maximum LikelihoodTree

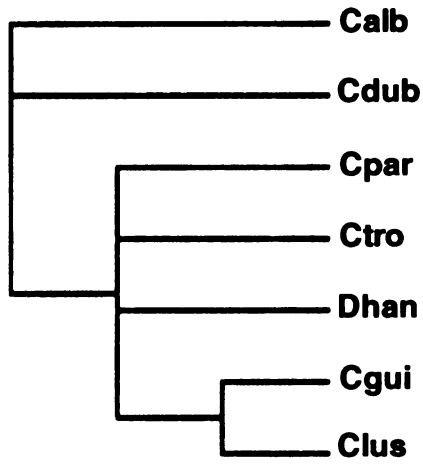


Figure C25: JAC1 *Candida* Maximum Parsimony Tree

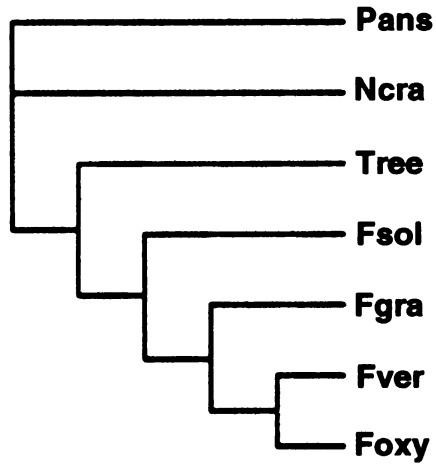


Figure C26: JAC1 *Fusarium* Bayesian Inference/ Maximum Likelihood / Maximum Parsimony Tree

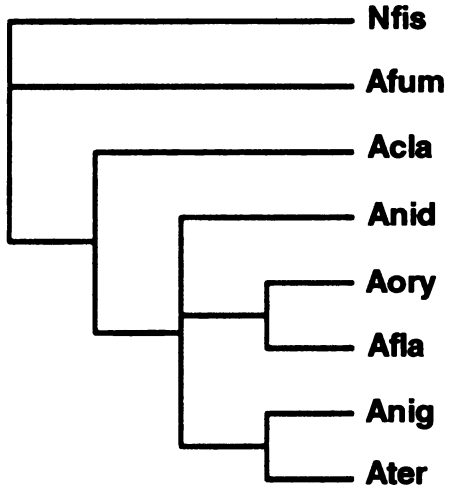


Figure C27: JAC1 *Aspergillus* Bayesian Inference Tree

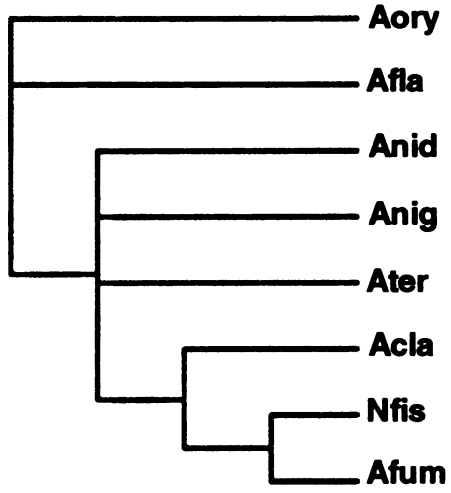


Figure C28: JAC1 *Aspergillus* Maximum Likelihood Tree

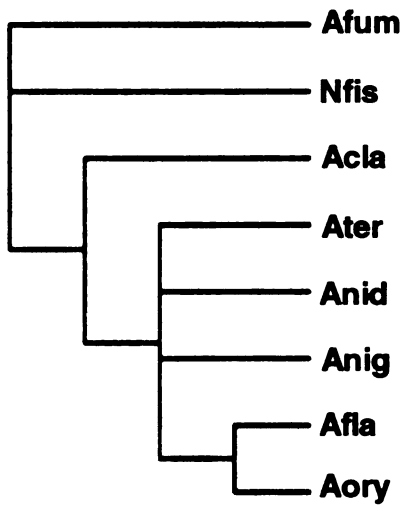


Figure C29: JAC1
***Aspergillus* Maximum**
Parsimony Tree

Appendix D

Evolutionary Rate Test Specifications Used in Control Files Used to Run `codeml` of PAML

Site-Specific Model
Model = 0
Nsites = 7 (ω distribution approximated as a beta distribution)
ncatG = 3 or 10 (# of categories pre-defined in the ω distribution)

Branch-Site Model: ***Model A as defined by Zhang et al. (2005)***
Model = 2
Nsites = 2 (ω distribution includes sites under positive selection)
ncatG = 3 (# of categories pre-defined in the ω distribution)
fix_kappa = 0 (kappa to be estimated)
fix_omega = 0 (omega to be estimated)

null model for branch-site test
Model = 2
Nsites = 2 (ω distribution includes sites under positive selection)
ncatG = 3 (# of categories pre-defined in the ω distribution)
fix_kappa = 1 (kappa fixed)
kappa = 1 (fixed value of kappa)
fix_omega (omega fixed)
omega = 1 (fixed value of omega)

Clade Model: ***Model D as defined by Bielawski and Yang (2004)***
Model = 3
Nsites = 3 (discrete ω distribution)
ncatG = 3 (# of categories pre-defined in the ω distribution)

null model for clade test
Model = 0 (ω distribution and estimated values apply to all branches of the tree)
Nsites = 0 (one gene-wide average ω estimated)
ncatG = 1 (# of categories pre-defined in the ω distribution)

Appendix E

Likelihood Ratio Tests of `codeml` Evolutionary Rate Analyses

Table E1: Likelihood Ratio Test Comparison of SSC1 Clade Model Test Outputs

PAML output negative log-likelihood scores (-lnL)

BI Tree 1

M0	MD (NcatG=2)	MD (NcatG=3)
-16792.46	-16286.11	-16213.79

BI tree 2

M0	MD (NcatG=2)	MD (NcatG=3)
-16840.70	-16322.97	-16248.26

BI tree 3

M0	MD (NcatG=2)	MD (NcatG=3)
-16808.21	-16295.06	-16221.94

BI tree 4

M0	MD (NcatG=2)	MD (NcatG=3)
-16855.95	-16331.81	-16255.92

Likelihood ratio test statistics

BI Tree 1

	2Δ(-lnL)	df	P-value	
M0 vs. MD (NcatG=2)	1012.71	3	3.15E-219	*
M0 vs. MD (NcatG=3)	1157.35	5	5.07E-248	*
MD (NcatG=2) vs. MD (NcatG=3)	144.64	2	3.90E-32	*

BI tree 2

	2Δ(-lnL)	df	P-value	
M0 vs. MD (NcatG=2)	1035.46	3	3.66E-224	*
M0 vs. MD (NcatG=3)	1184.88	5	5.53E-254	*
MD (NcatG=2) vs. MD (NcatG=3)	149.42	2	3.57E-33	*

BI tree 3

	2Δ(-lnL)	df	P-value	
M0 vs. MD (NcatG=2)	1026.28	3	3.57E-222	*
M0 vs. MD (NcatG=3)	1172.54	5	2.60E-251	*
MD (NcatG=2) vs. MD (NcatG=3)	146.26	2	1.74E-32	*

BI tree 4

	2Δ(-lnL)	df	P-value	
M0 vs. MD (NcatG=2)	1048.27	3	6.07E-227	*
M0 vs. MD (NcatG=3)	1200.06	5	2.85E-257	*
MD (NcatG=2) vs. MD (NcatG=3)	151.78	2	1.10E-33	*

Table E1 (continued): Likelihood Ratio Test Comparison of SSC1 Clade Model Test Outputs

PAML output negative log-likelihood scores (-lnL)

ML tree 1

M0	MD (NcatG=2)	MD (NcatG=3)
-16829.47	-16297.18	-16217.32

ML tree 2

M0	MD (NcatG=2)	MD (NcatG=3)
-16804.35	-16283.83	-16205.78

ML tree 3

M0	MD (NcatG=2)	MD (NcatG=3)
-16900.94	-16332.84	-16247.00

MP tree 1

M0	MD (NcatG=2)	MD (NcatG=3)
-16874.22	-16327.90	-16248.96

Likelihood ratio test statistics

ML tree 1

	2Δ(-lnL)	df	P-value	
M0 vs. MD (NcatG=2)	1064.59	3	1.75E-230	*
M0 vs. MD (NcatG=3)	1224.31	5	1.60E-262	*
MD (NcatG=2) vs. MD (NcatG=3)	79.86	2	4.56E-18	*

ML tree 2

	2Δ(-lnL)	df	P-value	
M0 vs. MD (NcatG=2)	1041.04	3	2.25E-225	*
M0 vs. MD (NcatG=3)	1197.14	5	1.23E-256	*
MD (NcatG=2) vs. MD (NcatG=3)	156.09	2	1.27E-34	*

ML tree 3

	2Δ(-lnL)	df	P-value	
M0 vs. MD (NcatG=2)	1136.20	3	5.10E-246	*
M0 vs. MD (NcatG=3)	1307.88	5	1.25E-280	*
MD (NcatG=2) vs. MD (NcatG=3)	171.68	2	5.25E-38	*

MP tree 1

	2Δ(-lnL)	df	P-value	
M0 vs. MD (NcatG=2)	1092.65	3	1.43E-236	*
M0 vs. MD (NcatG=3)	1250.53	5	3.33E-268	*
MD (NcatG=2) vs. MD (NcatG=3)	157.88	2	5.21E-35	*

Table E1 (continued): Likelihood Ratio Test Comparison of SSC1 Clade Model Test Outputs

PAML output negative log-likelihood scores (-lnL)

MP tree 2

M0	MD (NcatG=2)	MD (NcatG=3)
-16873.00	-16326.23	-16248.43

MP tree 3

M0	MD (NcatG=2)	MD (NcatG=3)
-16856.38	-16321.07	-16244.34

MP tree 4

M0	MD (NcatG=2)	MD (NcatG=3)
-16855.30	-16319.30	-16243.82

Likelihood ratio test statistics

MP tree 2

	2Δ(-lnL)	df	P-value	
M0 vs. MD (NcatG=2)	1093.53	3	9.22E-237	*
M0 vs. MD (NcatG=3)	1249.14	5	6.67E-268	*
MD (NcatG=2) vs. MD (NcatG=3)	155.61	2	1.62E-34	*

MP tree 3

	2Δ(-lnL)	df	P-value	
M0 vs. MD (NcatG=2)	1070.62	3	8.60E-232	*
M0 vs. MD (NcatG=3)	1224.08	5	1.79E-262	*
MD (NcatG=2) vs. MD (NcatG=3)	153.45	2	4.77E-34	*

MP tree 4

	2Δ(-lnL)	df	P-value	
M0 vs. MD (NcatG=2)	1072.00	3	4.32E-232	*
M0 vs. MD (NcatG=3)	1222.97	5	3.11E-262	*
MD (NcatG=2) vs. MD (NcatG=3)	150.97	2	1.65E-33	*

M0 is the null model; MD is the clade model described in Appendix C

NcatG' = the number of ω categories

* Denotes P-values significant at $P < 0.05$

Black boxes indicate the clade model significantly most likely to predict the data

Negative log-likelihood values shaded in gray indicate the best likelihood score among all SSC1 and SSQ1 clade model tests

Table E2: Likelihood Ratio Test Comparison of SSC1 and SSQ1 Clade Model Test Outputs

PAML output negative log-likelihood scores (-lnL)

BI Tree 1

M0	MD (NcatG=2)	MD (NcatG=3)
-31237.76	-30363.17	-30191.26

BI tree 2

M0	MD (NcatG=2)	MD (NcatG=3)
-31235.79	-30367.17	-30188.94

BI tree 3

M0	MD (NcatG=2)	MD (NcatG=3)
-31254.46	-30380.36	-30197.37

BI tree 4

M0	MD (NcatG=2)	MD (NcatG=3)
-31256.25	-30376.25	-30199.58

BI tree 5

M0	MD (NcatG=2)	MD (NcatG=3)
-31303.93	-30418.17	-30238.49

Likelihood ratio test statistics

BI Tree 1

	2Δ(-lnL)	df	P-value	
M0 vs. MD (NcatG=2)	1749.17	3	0.00E+00	*
M0 vs. MD (NcatG=3)	2092.98	5	0.00E+00	*
MD (NcatG=2) vs. MD (NcatG=3)	343.82	2	2.19E-75	*

BI tree 2

	2Δ(-lnL)	df	P-value	
M0 vs. MD (NcatG=2)	1737.23	3	0.00E+00	*
M0 vs. MD (NcatG=3)	2093.69	5	0.00E+00	*
MD (NcatG=2) vs. MD (NcatG=3)	356.46	2	3.94E-78	*

BI tree 3

	2Δ(-lnL)	df	P-value	
M0 vs. MD (NcatG=2)	1748.20	3	0.00E+00	*
M0 vs. MD (NcatG=3)	2114.18	5	0.00E+00	*
MD (NcatG=2) vs. MD (NcatG=3)	365.98	2	3.38E-80	*

BI tree 4

	2Δ(-lnL)	df	P-value	
M0 vs. MD (NcatG=2)	1760.00	3	0.00E+00	*
M0 vs. MD (NcatG=3)	1056.67	5	3.22E-226	*
MD (NcatG=2) vs. MD (NcatG=3)	353.34	2	1.88E-77	*

BI tree 5

	2Δ(-lnL)	df	P-value	
M0 vs. MD (NcatG=2)	1771.53	3	0.00E+00	*
M0 vs. MD (NcatG=3)	2130.89	5	0.00E+00	*
MD (NcatG=2) vs. MD (NcatG=3)	359.36	2	9.25E-79	*

Table E2 (continued): Likelihood Ratio Test Comparison of SSC1 and SSQ1 Clade Model Test Outputs

PAML output negative log-likelihood scores (-lnL)

BI tree 6

M0	MD (NcatG=2)	MD (NcatG=3)
-31302.21	-30422.29	-30236.43

ML tree

M0	MD (NcatG=2)	MD (NcatG=3)
-31759.55	-30475.41	-30284.90

MP tree 1

M0	MD (NcatG=2)	MD (NcatG=3)
-32025.64	-30650.50	-30447.31

MP tree 2

M0	MD (NcatG=2)	MD (NcatG=3)
-31988.10	-30629.28	-30432.39

Likelihood ratio test statistics

BI tree 6

	2Δ(-lnL)	df	P-value	
M0 vs. MD (NcatG=2)	1759.83	3	0.00E+00	*
M0 vs. MD (NcatG=3)	2131.54	5	0.00E+00	*
MD (NcatG=2) vs. MD (NcatG=3)	371.71	2	1.92E-81	*

ML tree

	2Δ(-lnL)	df	P-value	
M0 vs. MD (NcatG=2)	2568.28	3	0.00E+00	*
M0 vs. MD (NcatG=3)	2949.30	5	0.00E+00	*
MD (NcatG=2) vs. MD (NcatG=3)	381.02	2	1.83E-83	*

MP tree 1

	2Δ(-lnL)	df	P-value	
M0 vs. MD (NcatG=2)	2750.29	3	0.00E+00	*
M0 vs. MD (NcatG=3)	3156.67	5	0.00E+00	*
MD (NcatG=2) vs. MD (NcatG=3)	406.38	2	5.70E-89	*

MP tree 2

	2Δ(-lnL)	df	P-value	
M0 vs. MD (NcatG=2)	2717.65	3	0.00E+00	*
M0 vs. MD (NcatG=3)	3111.42	5	0.00E+00	*
MD (NcatG=2) vs. MD (NcatG=3)	393.77	2	3.12E-86	*

M0 is the null model; MD is the clade model described in Appendix C

NcatG' = the number of ω categories

* Denotes P-values significant at $P < 0.05$

Black boxes indicate the clade model significantly most likely to predict the data

Negative log-likelihood values shaded in gray indicate the best likelihood score among all SSC1 clade model tests

Table E3: Likelihood Ratio Test Comparison of JAC1 Site-Specific Model Test Output

PAML output negative log-likelihood scores (-lnL)			Likelihood ratio test statistics			
		***Saccharomyces* clade**				
MP/BI tree			**MP/BI tree**			
Number of ω Categories						
3	10		2Δ(-lnL)	df	P-value	
-2248.93	-2249.55		1.25	7	9.90E-01	
ML tree			**ML tree**			
Number of ω Categories						
3	10		2Δ(-lnL)	df	P-value	
-2253.53	-2283.81		60.55	7	1.17E-10	*
		***Candida* clade**				
BI tree			**BI tree**			
Number of ω Categories						
3	10		2Δ(-lnL)	df	P-value	
-2767.48	-2765.66		3.63	7	8.21E-01	
ML tree			**ML tree**			
Number of ω Categories						
3	10		2Δ(-lnL)	df	P-value	
-2797.53	-2796.00		3.06	7	8.79E-01	
MP tree			**MP tree**			
Number of ω Categories						
3	10		2Δ(-lnL)	df	P-value	
-2791.03	-2787.95		6.16	7	5.21E-01	
		***Fusarium* clade**				
MP/BI/ML tree						
Number of ω Categories			2Δ(-lnL)	df	P-value	
3	10		14.00	7	5.12E-02	
-2786.48	-2789.61					

Table E3 (continued) : Likelihood Ratio Test Comparison of JAC1 Site-Specific Model Test Outputs

PAML output negative log-likelihood scores (-lnL)		Likelihood ratio test statistics			
***Aspergillus* clade**					
MP Tree		**MP/BI/ML tree**			
Number of ω Categories					
3	10	2Δ(-lnL)	df	P-value	
-2508.72	-2508.73	0.03	7	1.00E+00	*
BI tree		**BI tree**			
Number of ω Categories					
3	10	2Δ(-lnL)	df	P-value	
-2801.61	-2790.07	23.08	7	1.65E-03	*
ML tree		**ML tree**			
Number of ω Categories					
3	10	2Δ(-lnL)	df	P-value	
-2856.26	-2847.00	18.51	7	9.88E-03	

* Denotes P-values significant at $P < 0.05$

Black boxes indicate the number of ω categories in the site-specific model that was significantly most likely to predict the data

Negative log-likelihood values shaded in gray indicate the overall best likelihood score for the given clade obtained among all site-specific tests

Tabe E4: Likelihood Ratio Test Comparison of SSQ1 Branch-Site Model Test Outputs

PAML output negative log-likelihood scores (-lnL)		Likelihood ratio test statistics			
BI tree 1		**BI tree 1**			
Model A, estimated ω	Model A, fixed ω = 1	Model A, estimated ω vs. Model A, fixed ω = 1			
-31029.34	-31129.88	2Δ(-lnL)	df	P-value	
		201.07	2	2.18E-44	*
BI tree 2		**BI tree 2**			
Model A, estimated ω	Model A, fixed ω = 1	Model A, estimated ω vs. Model A, fixed ω = 1			
-31023.33	-31128.51	2Δ(-lnL)	df	P-value	
		210.35	2	2.11E-46	*
BI tree 3		**BI tree 3**			
Model A, estimated ω	Model A, fixed ω = 1	Model A, estimated ω vs. Model A, fixed ω = 1			
-31037.84	-31139.44	2Δ(-lnL)	df	P-value	
		203.21	2	7.48E-45	*
BI tree 4		**BI tree 4**			
Model A, estimated ω	Model A, fixed ω = 1	Model A, estimated ω vs. Model A, fixed ω = 1			
-31043.68	-31124.20	2Δ(-lnL)	df	P-value	
		161.03	2	1.08E-35	*
BI tree 5		**BI tree 5**			
Model A, estimated ω	Model A, fixed ω = 1	Model A, estimated ω vs. Model A, fixed ω = 1			
-34199.23	-34292.48	2Δ(-lnL)	df	P-value	
		186.50	2	3.18E-41	*

PAML output negative log-likelihood scores (-lnL)		Likelihood ratio test statistics			
BI tree 6		**BI tree 6**			
Model A, estimated ω	Model A, fixed ω = 1	Model A, estimated ω vs. Model A, fixed ω = 1			
-31088.51	-31166.09	2Δ(-lnL)	df	P-value	
		155.15	2	2.04E-34	*
ML tree		**ML tree**			
Model A, estimated ω	Model A, fixed ω = 1	Model A, estimated ω vs. Model A, fixed ω = 1			
-31158.86	-31254.19	2Δ(-lnL)	df	P-value	
		190.67	2	3.95E-42	*
MP tree 1		**MP tree 1**			
Model A, estimated ω	Model A, fixed ω = 1	Model A, estimated ω vs. Model A, fixed ω = 1			
-31374.70	-31463.91	2Δ(-lnL)	df	P-value	
		178.43	2	1.80E-39	*

Table E4 (continued): Likelihood Ratio Test Comparison of SSQ1 Branch-Site Model Test Outputs

PAML output negative log-likelihood scores (-lnL)		Likelihood ratio test statistics			
MP tree 2		MP tree 2			
Model A, estimated ω	Model A, fixed ω = 1	Model A, estimated ω vs. Model A, fixed ω = 1			
-31351.81	-31458.31	2Δ(-lnL)	df	P-value	
		213.00	2	5.59E-47	*

* Denotes P-values significant at P < 0.05

Black boxes indicate the significantly model most likely to predict the data

Negative log-likelihood values shaded in gray indicate the overall best likelihood score obtained among all branch-site PAML tests

REFERENCES

Abascal F, Zardoya R, Posada D (2005) ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21(9): 2104-2105.

Andrew AJ, Dutkiewicz R, Knieszner H, Craig EA, Marszalek J (2006) Characterization of the interaction between the J-protein Jac1p and the scaffold for Fe-S cluster biogenesis, Isu1p. J Biol Chem 281(21): 14580-14587.

Baumann F, Milisav I, Neupert W, Herrmann JM (2000) Ecm10, a novel hsp70 homolog in the mitochondrial matrix of the yeast Saccharomyces cerevisiae. FEBS Lett 487(2): 307-312.

Benedict MQ, Cockburn AF, Seawright JA (1993) The Hsp70 heat-shock gene family of the mosquito Anopheles albimanus. Insect Mol Biol 2(2): 93-102.

Bettencourt BR, Feder ME (2002) Rapid concerted evolution via gene conversion at the Drosophila hsp70 genes. J Mol Evol 54(5): 569-586.

Bielawski JP, Yang Z (2003) Maximum likelihood methods for detecting adaptive evolution after gene duplication. J Struct Funct Genomics 3(1-4): 201-212.

Bielawski JP, Yang Z (2004a) Maximum Likelihood Methods for Detecting Adaptive Protein Evolution.

Bielawski JP, Yang Z (2004b) A maximum likelihood method for detecting functional divergence at individual codon sites, with application to gene family evolution. J Mol Evol 59(1): 121-132.

Boorstein WR, Ziegelhoffer T, Craig EA (1994) Molecular evolution of the HSP70 multigene family. J Mol Evol 38(1): 1-17.

Brown CJ, Todd KM, Rosenzweig RF (1998) Multiple duplications of yeast hexose transport genes in response to selection in a glucose-limited environment. Mol Biol Evol 15(8): 931-942.

Bucciantini M, Giannoni E, Chiti F, Baroni F, Formigli L et al. (2002) Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases. Nature 416(6880): 507-511.

Bukau B, Horwich AL (1998) The Hsp70 and Hsp60 chaperone machines. Cell 92(3): 351-366.

Cheetham ME, Caplan AJ (1998) Structure, function and evolution of DnaJ: conservation and adaptation of chaperone function. Cell Stress Chaperones 3(1): 28-36.

Conant GC, Wagner A (2003) Asymmetric sequence divergence of duplicate genes. Genome Res 13(9): 2052-2058.

Craig EA (1989) Essential roles of 70kDa heat inducible proteins. Bioessays 11(2-3): 48-52.

Davis JC, Petrov DA (2004) Preferential duplication of conserved proteins in eukaryotic genomes. PLoS Biol 2(3): E55.

DeBry RW (1999) Maximum likelihood analysis of gene-based and structure-based process partitions, using mammalian mitochondrial genomes. Syst Biol 48(2): 286-299.

Des Marais DL, Rausher MD (2008) Escape from adaptive conflict after duplication in an anthocyanin pathway gene. Nature 454(7205): 762-765.

Diamond ME, Dowhanick JJ, Nemeroff ME, Pietras DF, Tu CL et al. (1989) Overlapping genes in a yeast double-stranded RNA virus. J Virol 63(9): 3983-3990.

Drummond DA, Wilke CO (2008) Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell 134(2): 341-352.

Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH (2005) Why highly expressed proteins evolve slowly. Proc Natl Acad Sci U S A 102(40): 14338-14343.

Endo T, Ikeo K, Gojobori T (1996) Large-scale search for genes on which positive selection may operate. Mol Biol Evol 13(5): 685-690.

Feder ME, Cartano NV, Milos L, Krebs RA, Lindquist SL (1996) Effect of engineering Hsp70 copy number on Hsp70 expression and tolerance of ecologically relevant heat shock in larvae and pupae of Drosophila melanogaster. J Exp Biol 199(Pt 8): 1837-1844.

Felsenstein J (1978) Cases in which parsimony or compatibility methods can be positively misleading. Syst Zool 27: 401-419.

Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17(6): 368-376.

Felsenstein J (2003) Inferring Phylogeneies. United States: Sinauer Associates.

Fitzpatrick DA, Logue ME, Stajich JE, Butler G (2006) A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. BMC Evol Biol 6: 99.

Force A, Lynch M, Pickett FB, Amores A, Yan YL et al. (1999) Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151(4): 1531-1545.

Germaniuk A, Liberek K, Marszalek J (2002) A bichaperone (Hsp70-Hsp78) system restores mitochondrial DNA synthesis following thermal inactivation of Mip1p polymerase. J Biol Chem 277(31): 27801-27808.

Golding GB, Dean AM (1998) The structural basis of molecular adaptation. Mol Biol Evol 15(4): 355-369.

Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11(5): 725-736.

Gribaldo S, Lumia V, Creti R, de Macario EC, Sanangelantoni A et al. (1999) Discontinuous occurrence of the hsp70 (dnaK) gene among Archaea and sequence features of HSP70 suggest a novel outlook on phylogenies inferred from this protein. J Bacteriol 181(2): 434-443.

Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52(5): 696-704.

Gupta RS (1999) Hsp70 sequences and the phylogeny of prokaryotes. Mol Microbiol 31(3): 1007-1009.

Gupta RS, Singh B (1992) Cloning of the HSP70 gene from Halobacterium marismortui: relatedness of archaebacterial HSP70 to its eubacterial homologs and a model for the evolution of the HSP70 gene. J Bacteriol 174(14): 4594-4605.

Gupta RS, Singh B (1994) Phylogenetic analysis of 70 kD heat shock protein sequences suggests a chimeric origin for the eukaryotic cell nucleus. Curr Biol 4(12): 1104-1114.

Hakes L, Lovell SC, Oliver SG, Robertson DL (2007) Specificity in protein interactions and its relationship with sequence diversity and coevolution. Proc Natl Acad Sci U S A 104(19): 7999-8004.

Han W, Christen P (2003) Mechanism of the targeting action of DnaJ in the DnaK molecular chaperone system. J Biol Chem 278(21): 19038-19043.

Hennig W (1966) Phylogenetic Systematics. Urbana, IL: Univ. of Illinois Press.

Herrmann JM, Stuart RA, Craig EA, Neupert W (1994) Mitochondrial heat shock protein 70, a molecular chaperone for proteins encoded by mitochondrial DNA. J Cell Biol 127(4): 893-902.

Hittinger CT, Carroll SB (2007) Gene duplication and the adaptive evolution of a classic genetic switch. Nature 449(7163): 677-681.

Holder M, Lewis PO (2003) Phylogeny estimation: traditional and Bayesian approaches. Nat Rev Genet 4(4): 275-284.

Hurles M (2004) Gene duplication: the genomic trade in spare parts. PLoS Biol 2(7): E206.

Itoh T, Matsuda H, Mori H (1999) Phylogenetic analysis of the third hsp70 homolog in Escherichia coli; a novel member of the Hsc66 subfamily and its possible co-chaperone. DNA Res 6(5): 299-305.

Kellis M, Birren BW, Lander ES (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. Nature 428(6983): 617-624.

Kiley PJ, Beinert H (2003) The role of Fe-S proteins in sensing and regulation in bacteria. Curr Opin Microbiol 6(2): 181-185.

Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV (2002) Selection in the evolution of gene duplications. Genome Biol 3(2): RESEARCH0008.

Lange M, Macario AJ, Ahring BK, Conway de Macario E (1997) Heat-shock response in Methanosarcina mazei S-6. Curr Microbiol 35(2): 116-121.

Langkjaer RB, Cliften PF, Johnston M, Piskur J (2003) Yeast genome duplication was followed by asynchronous differentiation of duplicated genes. Nature 421(6925): 848-852.

Li WH (1980) Rate of gene silencing at duplicate loci: a theoretical study and interpretation of data from tetraploid fishes. Genetics 95(1): 237-258.

Li WH, Yang J, Gu X (2005) Expression divergence between duplicate genes. Trends Genet 21(11): 602-607.

Lill R, Muhlenhoff U (2008) Maturation of iron-sulfur proteins in eukaryotes: mechanisms, connected processes, and diseases. Annu Rev Biochem 77: 669-700.

Lim EH, Brenner S (1999) Short-range linkage relationships, genomic organisation and sequence comparisons of a cluster of five HSP70 genes in Fugu rubripes. Cell Mol Life Sci 55(4): 668-678.

Lim JH, Martin F, Guiard B, Pfanner N, Voos W (2001) The mitochondrial Hsp70-dependent import system actively unfolds preproteins and shortens the lag phase of translocation. Embo J 20(5): 941-950.

Lindquist S, Craig EA (1988) The heat-shock proteins. Annu Rev Genet 22: 631-677.

Lutz T, Westermann B, Neupert W, Herrmann JM (2001) The mitochondrial proteins Ssq1 and Jac1 are required for the assembly of iron sulfur clusters in mitochondria. J Mol Biol 307(3): 815-825.

Macario AJ, Dugan CB, Conway de Macario E (1991) A dnaK homolog in the archaebacterium Methanosarcina mazei S6. Gene 108(1): 133-137.

Messier W, Stewart CB (1997) Episodic adaptive evolution of primate lysozymes. Nature 385(6612): 151-154.

Muhlenhoff U, Lill R (2000) Biogenesis of iron-sulfur proteins in eukaryotes: a novel task of mitochondria that is inherited from bacteria. Biochim Biophys Acta 1459(2-3): 370-382.

Neupert W (1997) Protein import into mitochondria. Annu Rev Biochem 66: 863-917.

Nielsen R, Yang Z (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148(3): 929-936.

Nikolaidis N, Nei M (2004) Concerted and nonconcerted evolution of the Hsp70 gene superfamily in two sibling species of nematodes. Mol Biol Evol 21(3): 498-505.

Nimura K, Yoshikawa H, Takahashi H (1996) DnaK3, one of the three DnaK proteins of cyanobacterium Synechococcus sp. PCC7942, is quantitatively detected in the thylakoid membrane. Biochem Biophys Res Commun 229(1): 334-340.

Nuin PA, Wang Z, Tillier ER (2006) The accuracy of several multiple sequence alignment programs for proteins. BMC Bioinformatics 7: 471.

Ohno S (1970) Evolution by gene duplication. New York: Springer-Verlag.

Ota T, Nei M (1994) Divergent evolution and evolution by the birth-and-death process in the immunoglobulin VH gene family. Mol Biol Evol 11(3): 469-482.

Outten FW, Djaman O, Storz G (2004) A suf operon requirement for Fe-S cluster assembly during iron starvation in Escherichia coli. Mol Microbiol 52(3): 861-872.

Pal C, Papp B, Lercher MJ (2006) An integrated view of protein evolution. Nat Rev Genet 7(5): 337-348.

Pelham HR (1984) Hsp70 accelerates the recovery of nucleolar morphology after heat shock. Embo J 3(13): 3095-3100.

Philippe H, Budin K, Moreira D (1999) Horizontal transfers confuse the prokaryotic phylogeny based on the HSP70 protein family. Mol Microbiol 31(3): 1007-1010.

Renner T, Waters ER (2007) Comparative genomic analysis of the Hsp70s from five diverse photosynthetic eukaryotes. Cell Stress Chaperones 12(2): 172-185.

Ritossa F (1962) A new puffing pattern induced by heat shock and DNP in Drosophila. Experientia 18: 571-573.

Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19(12): 1572-1574.

Rouault TA, Tong WH (2005) Iron-sulphur cluster biogenesis and mitochondrial iron homeostasis. Nat Rev Mol Cell Biol 6(4): 345-351.

Sahi C, Craig EA (2007) Network of general and specialty J protein chaperones of the yeast cytosol. Proc Natl Acad Sci U S A 104(17): 7163-7168.

Scannell DR, Wolfe KH (2008) A burst of protein sequence evolution and a prolonged period of asymmetric evolution follow gene duplication in yeast. Genome Res 18(1): 137-147.

Schilke B, Williams B, Knieszner H, Pukszta S, D'Silva P et al. (2006) Evolution of mitochondrial chaperones utilized in Fe-S cluster biogenesis. Curr Biol 16(16): 1660-1665.

Schilke B, Forster J, Davis J, James P, Walter W et al. (1996) The cold sensitivity of a mutant of Saccharomyces cerevisiae lacking a mitochondrial heat shock protein 70 is suppressed by loss of mitochondrial DNA. J Cell Biol 134(3): 603-613.

Schmidt S, Strub A, Rottgers K, Zufall N, Voos W (2001) The two mitochondrial heat shock proteins 70, Ssc1 and Ssq1, compete for the cochaperone Mge1. J Mol Biol 313(1): 13-26.

Suzuki Y, Nei M (2001) Reliabilities of parsimony-based and likelihood-based methods for detecting positive selection at single amino acid sites. Mol Biol Evol 18(12): 2179-2185.

Swofford DL (2000) PAUP*, Phylogenetic Analysis Using Parsimony, Version 4.0b10.

Takahashi Y, Tokumoto U (2002) A third bacterial system for the assembly of iron-sulfur clusters with homologs in archaea and plastids. J Biol Chem 277(32): 28380-28383.

Tavaria M, Gabriele T, Kola I, Anderson RL (1996) A hitchhiker's guide to the human Hsp70 family. Cell Stress Chaperones 1(1): 23-28.

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22(22): 4673-4680.

Tissieres A, Mitchell HK, Tracy UM (1974) Protein synthesis in salivary glands of Drosophila melanogaster: Relation to chromosome puffs. J Mol Biol 85(3): 389-398.

Tuschl T, Eckstein F (1993) Hammerhead ribozymes: importance of stem-loop II for activity. Proc Natl Acad Sci U S A 90(15): 6991-6994.

Voisine C, Schilke B, Ohlson M, Beinert H, Marszalek J et al. (2000) Role of the mitochondrial Hsp70s, Ssc1 and Ssq1, in the maturation of Yfh1. Mol Cell Biol 20(10): 3677-3684.

Voisine C, Cheng YC, Ohlson M, Schilke B, Hoff K et al. (2001) Jac1, a mitochondrial J-type chaperone, is involved in the biogenesis of Fe/S clusters in Saccharomyces cerevisiae. Proc Natl Acad Sci U S A 98(4): 1483-1488.

Walter L, Rauh F, Gunther E (1994) Comparative analysis of the three major histocompatibility complex-linked heat shock protein 70 (Hsp70) genes of the rat. Immunogenetics 40(5): 325-330.

Wang R, Prince JT, Marcotte EM (2005) Mass spectrometry of the M. smegmatis proteome: protein expression levels correlate with function, operons, and codon bias. Genome Res 15(8): 1118-1126.

Wang TF, Chang JH, Wang C (1993) Identification of the peptide binding domain of hsc70. 18-Kilodalton fragment located immediately after ATPase domain is sufficient for high affinity binding. J Biol Chem 268(35): 26049-26051.

Ward-Rainey N, Rainey FA, Stackebrandt E (1997) The presence of a dnaK (HSP70) multigene family in members of the orders Planctomycetales and Verrucomicrobiales. J Bacteriol 179(20): 6360-6366.

Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13(5): 555-556.

Yang Z (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol 15(5): 568-573.

Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24(8): 1586-1591.

Yang Z, Swanson WJ (2002) Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. Mol Biol Evol 19(1): 49-57.

Zhang J, Rosenberg HF (2002) Complementary advantageous substitutions in the evolution of an antiviral RNase of higher primates. Proc Natl Acad Sci U S A 99(8): 5486-5491.

Zhang J, Nielsen R, Yang Z (2005) Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol 22(12): 2472-2479.

Zhang P, Gu Z, Li WH (2003) Different evolutionary patterns between young duplicate genes in the human genome. Genome Biol 4(9): R56.

Zheng L, Cash VL, Flint DH, Dean DR (1998) Assembly of iron-sulfur clusters. Identification of an iscSUA-hscBA-fdx gene cluster from Azotobacter vinelandii. J Biol Chem 273(21): 13264-13272.

Zheng L, White RH, Cash VL, Jack RF, Dean DR (1993) Cysteine desulfurase activity indicates a role for NIFS in metallocluster biosynthesis. Proc Natl Acad Sci U S A 90(7): 2754-2758.