ON TWO PROBLEMS IN COMPARATIVE GENOMICS OF EUKARYOTES

APPROVED BY SUPERVISORY COMMITTEE

Nick Grishin, Ph.D. (Mentor)

Joseph Albanesi, Ph.D. (Chair)

Lindsay Cowell, Ph.D.

Zbyszek Otwinowski, Ph.D.

Woodring Wright, M.D., Ph.D.

DEDICATION

To my parents and to Jessica.

ON TWO PROBLEMS IN COMPARATIVE GENOMICS OF EUKARYOTES

by

JEREMY RAYMOND SEMEIKS

DISSERTATION / THESIS

Presented to the Faculty of the Graduate School of Biomedical Sciences

The University of Texas Southwestern Medical Center at Dallas

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

The University of Texas Southwestern Medical Center at Dallas

Dallas, Texas

May, 2013

ON TWO PROBLEMS IN COMPARATIVE GENOMICS OF EUKARYOTES

Publication No._____

Jeremy Raymond Semeiks, Ph.D.

The University of Texas Southwestern Medical Center at Dallas, 2013

Supervising Professor: Nick Grishin, Ph.D.

The recent advent of whole-genome sequencing allows us to use novel comparative methods to explore the genetic bases for traits of interest. Here, I present two case studies of such methods applied to eukaryote genomes.

The first study regards the evolution of longevity in the mammalian proteome. Evolutionary theory suggests that the force of natural selection decreases with age. To explore the extent to which this prediction directly affects protein structure and function, I used computational methods to identify positions of proteins conserved in long-lived but not in short-lived mammal species. I analyzed 7,590 orthologous protein families in 33 mammalian species, accounting for body mass, phylogeny, and species-specific mutation rate. Overall, I found that

the number of longevity-selected positions in the mammalian proteome is much greater than would be expected by chance. Further, these positions are enriched in domains of several proteins that interact with one another in inflammation and other aging-related processes, as well as in organismal development. I present as an example the kinase domain of anti-Müllerian hormone type-2 receptor (AMHR2). AMHR2 inhibits ovarian follicle recruitment and growth, and my results show that its longevity-selected positions cluster near a SNP associated with delayed human menopause. Distinct from its canonical role in development, this region of AMHR2 may function to regulate the protein's activity in a lifespan-specific manner.

The second study concerns the genetic basis for toxin production in the black mold genus *Stachybotrys,* which produces several diverse toxins that can damage human health. Its strains comprise two mutually-exclusive toxin chemotypes, one producing satratoxins (a subclass of trichothecenes) and the other producing the less-toxic atranones. To determine the genetic bases for chemotype-specific differences in toxin production, I sequenced and assembled *de novo* four *Stachybotrys* genomes, including two from atranone strains and two from satratoxin strains. Comparative analysis of these four 35-Mbp genomes revealed several chemotype-specific gene clusters that are predicted to make atranones and satratoxins, based on several lines of evidence. I show that chemotype-specific gene clusters are likely the genetic basis for the mutually-exclusive toxin chemotypes of *Stachybotrys*. I then present a unified biochemical model for *Stachybotrys* toxin production.

TABLE OF CONTENTS

CHAPTER TWO

PRIOR PUBLICATIONS

Semeiks J, Borek D, Otwinowski Z, Grishin NV (2013) Comparative genome sequencing of the toxigenic black mold *Stachybotrys.* Manuscript submitted for publication.

Semeiks J, Grishin NV (2012) A method to find longevity-selected positions in the mammalian proteome. PLoS ONE 7(6): e38595. doi:10.1371/journal.pone.0038595.

Coppe J-P, Amend C, Semeiks J, Baehner FL, Bayani N, Campisi J, Benz CC, Gray JW, Neve RM (2010) ERBB receptor regulation of ESX/ELF3 promotes invasion in breast epithelial cells. Open Cancer J 3:89—100.

Rizki A, Weaver VM, Lee S-Y, Rozenberg GI, Chin K, Myers CA, Bascom JL, Mott JD, Semeiks JR, Grate LR, Mian IS, Borowsky AD, Jensen RA, Idowu MO, Chen F, Chen DJ, Petersen OW, Gray JW, Bissell MJ (2008) A human breast cell model of preinvasive to invasive transition. Cancer Res 68:1378—87.

Kenny PA, Lee GY, Myers CA, Neve RM, Semeiks JR, Spellman PT, Lorenz K, Lee EH, Barcellos-Hoff MH, Petersen OW, Gray JW, Bissell MJ (2007) The morphologies of breast cancer cell lines in three-dimensional assays correlate with their profiles of gene expression. Molecular Oncology 1:84—96.

Semeiks JR, Rizki A, Bissell MJ, Mian IS (2006) Ensemble attribute profile clustering: discovering and characterizing groups of genes with similar patterns of biological features. BMC Bioinformatics 7:147.

Semeiks JR, Grate LR, Mian IS (2005) Text-based analysis of genes, proteins, aging, and cancer. Mech Ageing Dev 126:193—208.

Semeiks JR, Grate LR, Mian IS (2004) Biological information networks of genetic loci and the scientific literature. Proceedings of the 2004 International Conference on Complex Systems.

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

LIST OF ABBREVIATIONS

ATR (eg, *ATR1, ATR2*) – set of genes specific to atranone-producing *Stachybotrys* strains
AC – *Stachybotrys* atranone strain-specific gene cluster
ACVR1, 2A, or 2B – activin receptor type-1, type-2A, or type-2B
AMH – anti-Müllerian hormone
AMHR2 – anti-Müllerian hormone type-2 receptor
BMPR1A or BMPR2 – bone morphogenetic protein receptor type-1A or type-2
BVMO – Baeyer-Villiger monooxygenase
B80 or BLOSUM80 – Blocks Substitution Matrix with minimum 80% identity
CAC – core atranone gene cluster
CDD – NCBI Conserved Domain Database
CTAB – cetyltrimethylammonium bromide
CTC – core trichothecene gene cluster
EPT – 12,13-epoxytrichothec-9-ene
FPP – farnesyl pyrophosphate
GGPP – geranylgeranyl pyrophosphate
kbp – DNA kilobase pair
MSA – multiple sequence alignment
MLS – maximum lifespan
PKS – polyketide synthase
PMDS – persistent Müllerian duct syndrome
PGLS – phylogenetic generalized least squares
SAT (eg, *SAT1, SAT2*) – set of genes specific to satratoxin-producing *Stachybotrys* strains
SC – *Stachybotrys* satratoxin strain-specific gene cluster
SMB – secondary metabolite biosynthesis
SNP – single nucleotide polymorphism
TGFbeta – transforming growth factor beta
TGFBR2 – TGF-beta Receptor Type II
TRI (eg, *TRI3, TRI4, TRI5*) – set of genes required for synthesis of trichothecenes

CHAPTER ONE
A method to find longevity-selected positions in the mammalian proteome

1.1. INTRODUCTION

1.1.1. The problem of aging: an evolutionary perspective

This chapter is adapted from an article conceived and published in the course of my graduate work (Semeiks and Grishin, 2012). It concerns the biomedical problem that most interests me: why do we humans age? More precisely, what are the molecular bases of the human aging phenotype?

For an organism, I define *aging* as *experiencing a set of time-related changes that adversely affect function and increase mortality risk as a function of time*[1]. Advanced age in humans is clearly the greatest risk factor for a large number of diseases, eg stroke, Alzheimer's disease, and many types of cancer (Gorelick, 2004; Yancik and Ries, 1994). Toward preventing such diseases as well as preserving healthy function late in life, I find it worthwhile to ask whether there are common and intelligible mechanisms of aging in humans that might be modified with medical intervention. Contrarily, it is also possible that what we term aging is a completely unintelligible phenomenon that is secondary only to time.

Why do we age? is a question that must be addressed at several levels. Teleologically, the classical evolutionary theory of antagonistic pleiotropy (Williams, 1957) posits that aging arises due to the decrease in selection pressure that occurs after successful reproduction. More specifically, in any population whose lifespans are initially limited by external factors such as predation, maximization of reproductive fitness earlier in life will occur at the expense of

1 Definition from Finch (1990), although he prefers the term *senescence*.

personal fitness later in life, leading to aging as a teleological epiphenomenon. Conversely, and consistently with this theory, lifespan extension has been shown to occur in *Drosophila* when selection pressure is experimentally increased in later age (Rose, 1984).

Mechanistically, the causes of human aging remain elusive. At the level of physiology, many theories have been advanced that propose some single mechanism as the main driver of aging. Most of these theories have not been well-tested, because the experiments would require great time, expense, and in many cases the development of interventions that do not presently exist. One counterexample that has been relatively well-tested and found lacking, at least in mice, flies, and other model organisms, is the mitochondrial free radical theory, which in strong form states that the primary cause of aging is free radical species generated by the mitochondria (critically reviewed by Lapointe and Hekimi, 2010). Although single-cause physiological theories are attractive due to their simplicity, the accepted evolutionary theory provides no reason to favor any single-cause physiological theory over more complex theories.

In the context of my current field of computational biology, it makes more sense to explore aging mechanisms at the level of the genome, broadly defined to include the proteome. Questions of genomics are computationally more tractable than those of physiology, and certainly if common physiological mechanisms of aging exist, they will ultimately have at least partial genomic bases. At the level of the genome as at the level of physiology, if aging is an intelligible process, it may be caused by several

nonexclusive mechanisms. These mechanisms may entail features or changes in the sequence, structure, function, and abundance of DNA, RNA, and proteins. Here, I focus on the relatively well-understood genomic phenomenon of changes in protein coding sequence and structure caused by fixed amino acid substitutions within mammal species.

If I am first interested in the genomics of human aging, then why here do I study the proteomes of many mammal species? In short, because of the way that mammals have evolved, this broader perspective may identify novel proteins and other genomic features that are implicated in human aging. Despite a 10–100-fold difference in maximum lifespan (MLS), most known mammal species show similar phenotypes of aging, in contrast to the more diverse phenotypes exhibited by many other clades (Finch, 1990). Some such phenotypes include vascular lesions, menopause, and wearing of teeth. This commonality suggests that the genetic determinants of mammalian aging and lifespan may have been relatively plastic starting from at least the time of the eutherian radiation 80 million years ago.

Regarding my chosen problem to identify fixed amino acid substitutions related to aging, two recent studies (Jobson, Nabholz, and Galtier, 2010; Li and de Magalhães, 2011) predicted a simple consequence of the evolutionary theory: one might expect that proteins necessary for slower mammalian aging (ie, greater MLS) would be conserved to a greater extent in long-lived versus short-lived species. Thus, the principle underlying these two studies is that it is possible to identify aging-related proteins (more specifically, families of orthologous proteins) by inferring and comparing some measure of

preferential DNA or amino acid substitutions, here called *longevity-selected positions*, among the several dozen mammal species whose proteomes are available. Jobson et al (2010) accomplished this comparison from the perspective of classical genetics, applying to a codon model a measure similar to the well-known dN/dS ratio (reviewed by Yang and Bielawski, 2000). Two pertinent features of Jobson et al's method were that (1) species were binned as “long-lived” or “short-lived” based on MLS and (2) in keeping with the conventional framework, the total estimate of synonymous substitutions in a gene determined the threshold for whether any particular codon position in that gene was called longevity-selected.

1.1.2. Innovations of the present work

In contrast to the genetics-based approach described above, here I have started from the perspective of protein structure to identify longevity-selected positions in the mammalian proteome. More so than classical genetics, structural biology emphasizes two concepts that may be useful to find interesting longevity-selected positions. First, positions within a protein are not interchangeable; whatever the estimated synonymous substitution rate, a single nonsynonymous substitution in a certain structural context can change a protein's function. Second, not all nonsynonymous substitutions are equal; rather, *a priori* one should expect those amino acid substitutions that commonly change biochemical function to matter most for the function of a specific protein. Based on these observations, I present a simple regression-based method to find longevity-selected

positions in orthologous protein families of mammals. My method employs the phylogenetic generalized least squares framework (PGLS, equivalent to phylogenetic independent contrasts; Freckleton, Harvey, and Pagel 2002) to relate, for each position (column) of a protein alignment, the MLS of the species represented in the column to the biochemical divergence of their residues from the residue of a long-lived reference species. Two benefits of PGLS are that (1) it is straightforward to control for common gerontological confounders, including species-specific mutation rate, body mass, and shared phylogeny (Speakman, 2005), and (2) one can naturally fit a continuous variable such as MLS, removing the need to arbitrarily bin species.

I use my method as a starting point to analyze longevity-selected positions, placing emphasis on their structural contexts. My results concern both the proteome as a whole and a specific protein domain identified by our analysis, the kinase domain of anti-Müllerian hormone type-2 receptor (AMHR2; kinase nomenclature per Knighton et al, 1990). AMHR2 is a receptor protein serine/threonine kinase in the TGF-beta Receptor Type II (TGFBR2) subfamily. The canonical role of AMHR2 is to inhibit the Müllerian ducts during development of the male fetus, and mutations cause the rare disease persistent Müllerian duct syndrome (PMDS; Imbeaud et al, 1995). More recently, a role for AMHR2 in ovarian follicle development of the adult female has also been identified (Durlinger, Visser, and Themmen, 2002), and this noncanonical function may relate to my findings.

1.2. MATERIALS AND METHODS

1.2.1. Selection of positions to analyze

I analyzed a selected subset of the OrthoMaM database version 6, which comprises multiple sequence alignments (MSAs) of 11,746 protein ortholog families from 36 mammalian species (Ranwez et al, 2007). Most of the sequences in the database were extracted from low-coverage genomes, so completeness and quality of alignments varied considerably. I first masked nonstandard isoforms and other divergent subsequences using a sliding window-based approach. Specifically, I excluded from further analysis any subsequence of at least 10 residues in which every 10-residue window had at least four residues each with less than 30% sequence identity to the rest of its column. I also excluded the three non-eutherian species due to high sequence divergence. Of the remaining data, I selected for fitting only columns that (1) included at least ten characters overall and (2) specifically included characters for both human and shrew. I refer to this subset as *selected columns*.

1.2.2. Column correction, fitting, and analysis

I define *fit columns* as the subset of selected columns that have at least three characters different from the human reference character, and *conserved columns* as all other selected columns. I independently fit each column in the *fit* subset to a phylogenetic generalized linear model (Freckleton, Harvey, and Pagel, 2002), as implemented in the R package caper, version 0.4 (Orme et al, 2011). Briefly, this framework assumes a Brownian model of trait evolution and uses the method of generalized least squares to perform multiple regression with correction for

global phylogenetic dependence, as indicated by the mammalian supertree. Specifically, for each column I fit the regression model

$$B80_{mut} \sim \log_{10} MLS + \log_{10} mass$$

Here, *MLS* are the maximum lifespans and *mass* the body masses of each species as reported in AnAge version 11 (de Magalhães and Costa, 2009). $B80_{mut}$ are the BLOSUM80 scores for each nonhuman ortholog character in the column versus the human ortholog character (Henikoff and Henikoff, 1992), corrected for mutation rate as described below. I used caper's default parameter values, including fixed nonterminal branch length multiplier (λ) of 1 for phylogenetic correction. As the input phylogeny, I used the mammalian supertree (Bininda-Emonds et al, 2007), which is ultrametric. Fitting completed after running for one day on a standard single processor (2.2 GHz, 16 GB RAM).

To control for each species' overall mutation rate, I used an approach similar to that of Li and de Magalhães (2011). Specifically, to estimate these mutation rates I constructed by maximum likelihood (protml; Adachi and Hasegawa, 1996) a tree whose total branch length, m_s, for each species s was the expected number of amino acid substitutions for s, as estimated from the set of all MSAs. I used this tree to correct each BLOSUM80 score for the mutation rate of s relative to human by adding $\log_{10}(m_s / m_{ref})$

to the score, where m_{ref} was the total branch length of human. I restricted the range of each $B80_{mut}$ score to the standard range of B80 scores attainable by its human character.

Each fit yielded both a longevity-selected slope, b_{MLS}, and p-value, p_{MLS}. I defined as a *longevity-selected position* any column with both $b_{MLS} > 0$ and $p_{MLS} < 0.01$. Conserved columns were assigned $b_{MLS} = 0$ and $p_{MLS} = 1$.

For the large-scale analyses, in the human orthologs I predicted secondary structure with PSIPRED (Jones, 1999); differences in composition were tested with Pearson's χ^2 test. Protein domain definitions and other features were taken from Swiss-Prot (Boeckmann et al, 2003) and mapped to OrthoMaM alignments by aligning each Swiss-Prot sequence to its human counterpart in OrthoMaM. Ontology enrichment analysis was performed using the Functional Annotation Clustering module of DAVID (Huang, Sherman, and Lempicki, 2009) with default parameters, including the human genome as background. I report only Benjamini-corrected p-values. For the rolling-window analysis of positions, I included only contiguous blocks of 10 selected positions.

To construct the randomized control dataset, for each species *A* except human, I swapped MLS, body mass, and phylogenetic label with those of one species *B*, which was randomly selected without replacement. All alignments remained unchanged, meaning that the same set of columns were fit in both the randomized control and real sets.

1.2.3. Homology modeling and structural analysis of AMHR2

I created a homology model of AMHR2 with Modeler (Eswar et al, 2007), using as template the kinase domain of BMPR2 (PDB ID: 3G2F). Homology models made with SWISS-MODEL (Arnold et al, 2005), or using as template the kinase domain of ACVR2B (PDB ID: 2QLU), yielded similar results. Positional conservation was calculated with AL2CO (Pei and Grishin, 2001), using 3G2F as the structural model.

1.3. RESULTS AND DISCUSSION

1.3.1. Due to overall conservation, most positions in the mammalian proteome are not longevity-selected

To identify specific positions (i.e., alignment columns) in the mammalian proteome that are conserved in long-lived but not in short-lived species, I fit a generalized multiple regression model to each position independently. My overall approach followed from the observation that many of the positions I sought were distinguished by high correlation between (1) each species' MLS and (2) functional similarity of each species' residue to that of a long-lived reference species. I used human as the reference species, both because it was the longest-lived mammal whose sequence was available and because I had the best confidence in the accuracy of its sequence. Figure 1-1 shows a simplified conceptual example of my approach. My method also accounted for each species' body mass and overall mutation rate relative to human. I emphasize that I chose multiple regression not for rigorous statistical reasons, but only as a computational tool to help form new biological hypotheses.

In this manner, I fit selected columns among the proteomes of 33 species (Figure 1-2), with MLS ranging from 3 y (shrew) to 90 y (human). Initially, I attempted to fit the entire unfiltered OrthoMaM database. However, this effort yielded many obvious false-positives driven by low sample size and data of questionable quality, including highly divergent sequence at exon-intron boundaries and alternate isoforms. For this reason, I implemented several heuristic filters for selection. In particular, because rodents are over-represented among the short-lived species of OrthoMaM, I selected only columns containing a character for shrew (*Sorex araneus)*, the shortest-lived non-rodent. (Sections 1.2.1 and 1.2.2 give precise definitions of the sets

selected, fit, and *conserved.*) Using these criteria, I verified that most selected positions in the mammalian proteome (80%) are conserved across all species (Table 1-1; also found previously by Jobson, Nabholz, and Galtier [2010] and Li and de Magalhães [2011]). In particular, at least 10 of 261 genes in the GenAge database of aging-related genes (de Magalhães et al, 2009) have protein products that show near-complete conservation in mammals (≥ 90% ratio of conserved positions to total human ortholog length), for example beta-catenin *(CTNNB1)*, valosin-containing protein *(VCP)*, fibroblast growth factor receptor 1 *(FGFR1)*, and lamin A *(LMNA)*. Thus, if these genes contribute to differences in mammalian longevity, it is likely via some mechanism other than structural differences in their protein products.

1.3.2. Among nonconserved positions, longevity-selected positions occur more often than expected by chance

For each selected position, my fitting procedure yielded both a longevity-associated slope (b_{MLS}) and associated p-value (p_{MLS}), as well as corresponding measures for body mass (b_{mass} and p_{mass}). Only those positions with significantly positive slope were called *longevity-selected* ($b_{MLS} > 0$ and $p_{MLS} < 0.01$) or *mass-selected* ($b_{mass} > 0$ and $p_{mass} < 0.01$). There were no positions that were both longevity-selected and mass-selected, suggesting that $p_{MLS} < 0.01$ was a reasonable cutoff in general for my analyses.

Among positions with positive MLS slope, I analyzed the distribution of p-values in order to determine whether proteome-wide trends existed with regard to longevity-

selected positions (Figure 1-3). As a negative control, I also analyzed a matched set of positions whose MLS, body mass, and phylogenetic position had been randomly swapped ("randomized control"). If there were no overall relationship between MLS and amino acid conservation, then one would expect significant p-values to be no more common than nonsignificant p-values after accounting for shared phylogeny and body mass. This is indeed the case for the randomized control (Figure 1-3B). However, for the real data (Figure 1-3A), positions with more significant p-values are clearly overrepresented relative to those with less significant p-values. These results indicate that overall, longevity-selected positions in the mammalian proteome are much more likely than would be expected by chance.

This finding is robust to several perturbations of the data (Figure 1-4), including use of chimp as the reference instead of human (Figure 1-4A) and use of BLOSUM62 scores instead of BLOSUM80 scores (Figures 1-4D and 1-4E). I note that although BLOSUM62 is more commonly used than BLOSUM80, in this case BLOSUM80 is more appropriate because mammalian proteomes typically have 80—90% sequence identity. The result also holds for input in which all nonhuman primates except one (here, rhesus) are omitted (Figure 1-4B), indicating that it is not an artifact of primate over-representation in OrthoMaM. Almost all (7,689/7,723) of the positions called significant in Figure 1-3A have $p_{MLS} < 0.05$ in this control set, suggesting that its relative lack of very significant p_{MLS} values is simply due to fewer data available for each position (n=26 versus 32 species available to fit). Finally, as one would expect the result does not hold

when dog, a shorter-lived species, is used as the reference instead of human (Figure 1-4C).

A plausible biological hypothesis to explain this overall plethora of longevity-selected positions is that the evolution of longer mammalian lifespan requires particular concerted patterns of substitutions throughout the proteome that subtly affect protein properties such as binding affinity, folding, and stability. Consistent with this hypothesis is that relative to mouse (MLS 4 y), the proteome of naked mole rat (a rodent with MLS ~30 y) is more resistant to urea-induced unfolding (Pérez et al, 2009), suggesting increased protein stability in the longer-lived rodent. An analogous process requiring concerted patterns of substitution may be the convergent evolution of hyperthermostability in archaea and bacteria (Suhre and Claverie, 2003).

To determine overall trends in longevity-selected positions with regard to structural features of the proteome, I created and searched databases of both predicted secondary structure and predicted disordered regions for the human proteome. Table 1-2 shows the secondary structure composition of the human genome as a whole, as well as in the longevity-selected positions of both the real fit data and the randomized control data. This table shows two trends. The first trend is that, in both real and randomized longevity-selected positions, random coils are overrepresented, at the expense of α-helices and β-strands, relative to the human proteome as a whole ($\chi^2(2, n=7{,}702) = 265.23$, $p < 2.2e\text{-}16$). This is easily explained by the observation that sequence in random

coils tends to be less conserved than other sequence; thus, fit positions will tend to be overrepresented in these regions.

The second trend is that, relative to randomized longevity-selected positions, real longevity-selected positions are slightly enriched in α-helices at the expense of β-strands ($\chi^2(2, n=7{,}702) = 18.36$, $p = 1.0e\text{-}4$). The significance of this finding is unknown, but it is possible that an abundance of α-helices imparts extra stability to proteins and protein complexes, eg via coiled-coil interactions (Mason and Arndt, 2004).

1.3.3. Longevity-selected positions are enriched in protein domains with known roles in inflammation, development, and other diverse functions

The majority of longevity-selected positions are located in regions of proteins that are unannotated and presumably unstructured. Since these regions are generally of unknown function at present, it is difficult to interpret the biochemical significance of substitutions within them. Thus, to find longevity-selected positions with the best likelihood of causing well-characterized changes to protein structure and function, I next narrowed my focus to known protein domains. Specifically, I compiled a list of all 129 domains that contain at least two longevity-selected positions. The 129 domains are contained in 114 proteins. Appendix A summarizes the proteins and domains, and Appendix B shows data for each longevity-selected position in the domains, including number of characters (ie, species) fit and all slopes and p-values. Five genes for the 114 proteins shown in these tables are present in the GenAge database (de Magalhães et al,

2009): serine-protein kinase ATM, ATM; serine/threonine protein kinase ATR, ATR; breast cancer type 1 susceptibility protein, BRCA1; ATP-dependent DNA helicase Q4, RECQL4; and DNA-dependent protein kinase catalytic subunit, PRKDC. Functional annotation clustering revealed that several of the 114 proteins belong to functional classes that have been associated with aging (Table 1-3). I give three examples. First, leukemia inhibitory factor receptor (LIFR) and interleukin-6 receptor subunit beta (IL6ST) dimerize to form the receptor for leukemia inhibitory factor (LIF; not present in our results), whose signaling is upregulated with age in association with thymic atrophy (Sempowski et al, 2000). Second, arachidonate 15-lipoxygenase (ALOX15) is an enzyme that is upregulated in aging rat brain (Qu, Uz, and Manev, 2000); it functions in production of inflammatory leukotrienes, and may also function nonenzymatically to upregulate NFκB (Manev et al, 2000). Third, several of these proteins function in blood coagulation, markers of which increase with age and may interact with markers of inflammation (Kanapuru and Ershler, 2009). Detailed structural analysis of these domains remains to be performed.

Table 1-3 also indicates that several of our hits are involved in development, and this highlights a limitation of my data set. It is well-known that developmental schedule and longevity have co-evolved in mammals (Finch, 1990; de Magalhães, Costa, and Church, 2007); thus, these positions may specifically be conserved due to their effect on development, or they may pleiotropically affect both development and adult longevity. I did not attempt to correct for developmental schedule when fitting the data. In my

framework, such correction is possible in concept by adding to the regression model a third predictor variable, species age at maturity. However, the present set of species exhibits the problem of multicollinearity: age at maturity is too well-correlated with MLS for correction to be a realistic goal ($R^2 = 0.48$ on n=30 species for age at female maturity after phylogenetic correction). Additionally, no maturity data are available for three important species in our set: *Pteropus, Tarsius,* and *Tupaia.* Thus, it is not possible to distinguish in general whether the positions found by my method are conserved due to their roles in development, adult longevity, or both. But most of these proteins, including the next discussed example AMHR2, do have verified roles in the adult organism and may plausibly affect longevity.

1.3.4. Longevity-selected positions cluster in the kinase domain and C-terminal tail of AMHR2

The domain containing the greatest number of longevity-selected positions (n=7) was the protein kinase domain of AMHR2, a protein introduced in Section 1.1. In human, this domain comprises residues 203—517. Downstream of this domain is the C-terminal tail of the protein, residues 518—573. This cysteine-rich tail is unique to the AMHR2 ortholog family and is predicted to lack secondary structure. On average, the residues in the region 463—573 that do not form secondary structure have p_{MLS} values that are consistently in the top 2% of the 3.6 million selected positions (median $p_{MLS} < 0.568$ by sliding window analysis). But this trend does not hold for AMHR2 overall (median p_{MLS} is 0.800 excluding secondary structure positions), suggesting that specifically the kinase

C-lobe and downstream C-terminal tail of AMHR2 are under lifespan-related selective pressure.

In sequence, five of the seven longevity-selected positions in the kinase domain of AMHR2 concentrate in two regions near the C-terminus of the domain (Figure 1-5). To determine the likely locations of our longevity-selected positions on the structure of AMHR2, I mapped them to the recently-solved structure of the kinase domain of bone morphogenetic protein type 2 receptor (BMPR2). This domain is the closest homolog to the kinase domain of AMHR2, with 40% sequence identity. My mapping (Figure 1-6) shows that these five longevity-selected positions cluster at or near a common surface of the AMHR2 kinase C-lobe. Specifically, they are all located on two loops near the bottom-front face of the domain: the αG—αH loop (Y465, T469, and F473) and the C-terminal loop following αI (E513 and H515). All five side-chains at least partially face solvent.

The side-chain of Y465 forms an intra-loop hydrogen bond with R462, a conserved residue. Thus, in conjunction with P464, the length of the Y465 side-chain constrains the angle of a second conserved arginine, R463, which forms hydrogen bonds to residues on αF and the αF-αG loop. Y465 is conserved in 8/9 species with MLS of at least 30 y, but in 9/19 shorter-lived species it is instead H, F, C, N, or D, none of which (except possibly H) can hydrogen-bond with R463 at the same angle as can Y. I predict that these substitutions would abolish at least the hydrogen bond with R462, destabilizing the αG—αH loop. Although the αG—αH and C-terminal loops are both quite divergent

overall and contain several positions that did not meet the significance criterion for longevity conservation, Y465 is the only nonconserved residue within them whose side-chain is predicted to interact with another residue of AMHR2.

T469 represents a position that may be differentially phosphorylated, but it also highlights some current limitations of my method and data set. T469 is the only predicted phosphorylation site on the αG—αH loop (NetPhos score 0.692; Blom, Gammeltoft, and Brunak, 1999). However, Figure 1-5 shows that this residue is consistently serine or threonine in all species except shrew, where it is alanine. Substitution to a non-phosphorylatable residue would be of biological interest, but it is also possible that this position simply represents a genome assembly error in shrew. This position has p_{MLS} = 0.006, making it the least significant position of the five. My method could exclude it and many similar cases by adding more criteria to the position selection process, but at the cost of decreased sensitivity and increased complexity. I expect that the coming availability of high-coverage *de novo* mammalian genome assemblies will resolve many such cases (eg, Gnerre et al, 2011; Kim et al, 2011).

F473 faces forward in our model. It is conserved hydrophobic (F or L) in all species except guinea pig, rat, mouse, and shrew, where it is S or C. Thus, F473 may be involved in a hydrophobic interaction with a binding partner.

I have low confidence in the exact placement of the short C-terminal loop, because it is not conserved in BMPR2 and lacks contacts with the other elements of our model. However, both E513 and H515 on this loop are preferentially charged in long-

lived species, consistent with differential binding affinity. E513 is conserved in all species except rat, mouse, and shrew, where it is V, G, and A. H515 is conserved positive (H or R) in 16/17 species with MLS at least 16 y, but it is charged (H, R, or D) in only 3/9 shorter-lived species. The preference for hydrophilic residues in long-lived species at these two positions specifically may suggest that flexibility of the C-terminal loop is a longevity-conserved property.

Most of the residues on the αG—αH loop face the solvent, suggesting that they may interact with another domain or protein. If this is the case, then assuming that the gross function of AMHR2 is conserved within mammals, one would expect other residues on a common surface with the αG—αH loop to be conserved. I used positional conservation analysis to determine conservation of the surface residues of two alignments, (1) mammalian AMHR2 orthologs exclusively and (2) a representative set of mammalian orthologs in the TGFBR2 subfamily, including orthologs of TGFBR2, activin receptor type-2A and B (ACVR2A and ACVR2B), BMPR2, and AMHR2 (not shown). This analysis revealed two conserved solvent-facing patches flanking the region of our longevity-selected positions. One patch, mainly comprising the αH—αI loop, is conserved in all TGFBR2 subfamily members, confirming a previous observation (Belville et al, 2009). The other patch, mainly comprising the αF—αG loop, is conserved only in AMHR2 orthologs.

Overall, these findings are consistent with the existence of a large interaction surface conserved in all AMHR2 orthologs whose area, and thus binding affinity, varies

at the αG—αH loop in a species-specific manner. It is likely that this is a novel docking surface involved in the regulation of AMHR2; one possible regulatory binding partner is the C-terminal tail of AMHR2 itself, whose positions also have consistently low p_{MLS} values relative to the proteome overall, as noted above. Since loop αG—αH flanks this conserved patch, and four of the five residues face solvent in my model, it is possible that overall these positions contribute to lifespan-specific binding affinity.

I am unaware of reported mutations specifically in the two loops of AMHR2 that contain the longevity-selected positions described above. However, two prior lines of inquiry are consistent with my docking-surface hypothesis. First, Belville et al (2009) also mapped the AMHR2 kinase domain to a solved structure in order to investigate natural mutations found in PMDS. Although based on a structure of the more distantly-related ACVR2B instead of BMPR2, its details are similar to those of my model, including both overall tertiary structure and specific residue orientation. Of the seven mutated positions they analyzed, four were located in the C-lobe of the kinase domain. One, D491, lies in the conserved αH—αI loop and faces solvent; its mutation to H causes PMDS. This mutation further supports the idea that the solvent-facing bottom of the C-lobe is critical for proper AMHR2 function.

Second, closer to the two loops on the bottom face of the domain is the residue E485, which is in αH and also faces solvent. In two independent population studies of Dutch women, the mutation E485Q was associated with menopause delayed by up to one year (Kevenaar et al, 2007; Voorhuis et al, 2011). In addition to its canonical role in male

fetal development, AMHR2 also plays a second role in adult reproductive function. It is expressed in granulosa cells of adult females, where it seems to act as a feedback inhibitor of follicle recruitment and growth by binding its ligand, AMH, which is secreted in a paracrine manner specifically by more mature follicles (Durlinger, Visser, and Themmen, 2002). Follicle depletion is the cause of menopause (Finch, 1990), and follicular decline or menopause has been observed in most or all mammals studied, including whales (Ward et al, 2009; Foote, 2008), nonhuman primates (Walker and Herndon, 2008), rodents, and others (Finch, 1990; Finch and Gosden, 1986), although admittedly data are lacking for most species in the wild. A reasonable deduction is that a species' rate of follicle depletion scales inversely with its longevity. I speculate that differential regulation of AMHR2 in a lifespan-dependent manner could act as a mechanism that effects this scaling, increasing the probability that a female has used all her reproductive potential before her death. This hypothesis could be viewed as a case of the disposable soma theory of aging (Kirkwood, 1977). It might be tested by relating AMHR2 ortholog sequence to rate of follicular decline across several species.

I also note the unusual cysteine conservation in the C-terminal tail, which is unique to AMHR2 orthologs. There are eight cysteine residues in this region of human AMHR2, all of which are relatively conserved. Multiple regression revealed a specific fit of the number of cysteines conserved to $\log_{10}$ MLS ($p_{MLS} = 0.002$ and $p_{mass} = 0.082$ after phylogenetic correction). It is possible that these residues bind zinc or another metal ion,

thus imparting structure to this region, but the region does not match known zinc finger motifs.

1.3.5. Comparison with previous methods

Generally, I did not observe overlap between the longevity-selected proteins I identified and those identified in previous work (Jobson, Nabholz, and Galtier, 2010; Li and de Magalhães, 2011). But this is not surprising, because I differed in my assumptions, goals, and data sets used, as detailed in Sections 1.1 and 1.2.1. Most notably, I fit a smaller subset of high-quality protein alignments, focused on structured protein regions, and chose to ignore synonymous codon substitutions. Thus, I view my results as complementary, not conflicting, to prior work. Notably, both my method and that of Li and de Magalhães (2011) identified "myosin complex" as an ontology term enriched in longevity-selected proteins (Table 1-3). The two methods also agreed that two proteins were longevity selected, rho guanine nucleotide exchange factor 16 (ARHGEF16) and lymphokine-activated killer T-cell-originated protein kinase (PBK). As both myosin and ARHGEF16 are involved in cell migration (Hiramoto-Yamaki et al, 2010), there may be lifespan-specific differences in this activity, or our concordant findings may simply reflect its standard role in organismal development.

One apparent novelty of my method is that it allows analysis of individual positions in the proteome, not just entire proteins. In fact, this is not novel, as the method of Jobson et al (2010) also entails identification of specific longevity-selected positions, and the genes they called "longevity-selected" and "longevity-relaxed" were simply

genes with statistical over- or under-abundances of such positions. Here, I have emphasized individual positions rather than entire proteins for three reasons. First, I think it plausible, as did Jobson et al (2010), that a mark of a longevity-selected protein is an abundance of longevity-selected positions. Second, a focus on individual positions allows to more precisely determine arbitrary regions of a protein that may be longevity-selected, as exemplified by the C-terminus of AMHR2. In theory, such specific focus can suggest novel biochemical mechanisms. Third, since aging is a complex trait that is under weak selection, it is plausible that some major determinants are subtle general properties of the proteome itself (discussed in Section 1.3.2), rather than the explicit activity of a single protein or even a few functional collections of proteins. Longevity-selected positions are the most obvious markers of such properties, and so may provide clues to identify them, in the same way as they may identify individual longevity-selected proteins. In short, I do not suggest that the positions that I call longevity-selected, in isolation, are major determinants of mammalian longevity. I only suggest that they may mark proteins or proteomic features that *are* such determinants, but are less obvious.

For detecting longevity-selected protein positions, benefits of my method versus standard codon-level methods such as the codeml program of PAML (Yang, 2007) include emphasis on detection of significant biochemical changes that are likely to affect protein structure; straightforward single-position resolution, allowing to easily test hypotheses regarding arbitrary regions of proteins, as described above; simple control for species-specific mutation rate and shared phylogeny and fitting of body mass as an

alternate hypothesis to MLS; avoidance of the need to arbitrarily bin species by MLS; and faster run time. I reiterate that my method in theory is compatible with any quantitative trait, as illustrated by my inclusion of both MLS and body mass, although in practice correlations between predictor variables in the data limit the application of this for the predictors of greatest interest, such as developmental schedule (Section 1.3.3). However, this is a limitation that my approach shares with the others, given the set of species available. Some unique drawbacks of my method, in addition to those described in Section 1.3.4, include its reliance on a single reference proteome and lack of a specific model of protein evolution, implying unsuitability for rigorous statistical hypothesis-testing. It is a task for future work to combine the benefits of this method with those of standard codon model-based methods.

1.4. CONCLUSIONS AND RECOMMENDATIONS

Based on principles of protein structure, I have developed a simple, extensible, and gerontologically-oriented method to find longevity-selected positions in the mammalian proteome. Using this method I found that, surprisingly, longevity-selected positions are much more common in the mammalian proteome than would be expected based on a randomized control. I also used this method to identify specific protein regions that deserve further study in the context of the comparative biology of aging and development, as well as specific aging-related proteins that are likely not lifespan-conserved due to their overall conservation. The protein region I found that is most likely to be lifespan-conserved is the kinase domain and C-terminal tail of AMHR2, in which the longevity-selected residues lie on a common surface of unknown functional significance. Given caveats with my methodology (Section 1.3.5) and also with the dataset I used (Section 1.4.2), I have found my method to be a reasonable starting point for comparative analysis of protein function.

Given my conclusions, the following two topics are best suited for follow-up study.

1.4.1. AMHR2

To determine the functional significance of the longevity-selected surface of AMHR2 identified here, this protein should be further characterized *in vitro.* Currently, the key question regards the identity of AMHR2's binding partners. Of course, the extracellular domain of AMHR2 binds its canonical ligand, AMH (di Clemente et al, 1994). It has also been shown in rodent cells (Belville et al, 2005) that AMHR2 signals through a BMPR-like pathway, indirectly

activating mothers against decapentaplegic homolog 1 (SMAD1) by heterodimerization with the TGFβ Type I receptors activin receptor type-1 (ACVR1) and bone morphogenetic protein receptor type-1A (BMPR1A). To identify other protein binding partners in culture, it may be possible to perform a pulldown experiment with a glutathione S-transferase-tagged AMHR2 fusion construct, or to perform chemical crosslinking followed by fragmentation and mass spectroscopy. As discussed in Section 1.3.4, it is also possible that the cysteine-rich C-terminal tail of AMHR2 binds zinc or another metal ion. This might be tested *in vitro* by mobility assays of a C-terminal tail construct in the presence of various metals. Ultimately, crystallization of the AMHR2 kinase domain and C-terminal tail may be feasible. If C-terminal ligands can be identified, then a crystal structure that includes them would likely reveal a novel mode of kinase regulation, as the longevity-selected surface is distinct from known kinase regulatory regions (Goldsmith et al, 2007). Ideally, all these experiments would be performed using AMHR2 ortholog sequence from both long-lived (eg, human) and short-lived species (eg, shrew or mouse) to contrast their binding partners and structures.

More difficult would be to test my physiological hypothesis that AMHR2 can regulate timing of menopause in a lifespan-specific manner in mammals (Section 1.3.4). Conceptually, a straightforward approach would be to create a transgenic mouse line that replaces the wildtype *Amhr2* with the *AMHR2* ortholog of a long-lived species. Then follicle count and quality of aged transgenic animals could be compared to that of aged wildtypes, in a design similar to that of Perez et al (1999). However, due to this

experiment's substantial time and cost, it should not be attempted until AMHR2 has been characterized further *in vitro,* as above.

An alternate, but purely phenomenological, approach to explore the relationship between AMHR and menopause would be to count follicles throughout the lifespans of a sample of the species listed in Figure 1-5, and then correlate their patterns of decline with features of their AMHR2 sequences. This study would necessarily be poorly controlled, and even with a cross-sectional design the expense would currently be prohibitive due to two factors: the difficulty of obtaining the longest-lived nonhuman species, and lack of a noninvasive method to count total follicle population or otherwise measure ovarian reserve in diverse species. Although ultrasound has been used to count antral follicles noninvasively in species including mouse and cow (Jaiswal, Singh, and Adams, 2009; Burns et al, 2005), antral follicle population is cyclic, and thus cannot represent the total follicle population. The current spatial resolution of ultrasound is too low to resolve primordial follicles (< 100 μm diameter; Palma et al, 2012), which comprise the majority of follicles.

1.4.2. Improvement of mammalian genomic data

To more accurately identify longevity-selected positions in the mammalian proteome, data issues must be addressed at four levels: the initial selection of genomes to sequence, sequencing and assembly *per se*, gene annotation, and finding orthologs in finished proteomes. All four of these levels are interrelated, but the only one specific to

the present project is the conceptual question of which genomes should be sequenced for gerontological analysis, so I focus on that question in this section. My major personal reason for performing the *Stachybotrys* work described in Chapter 2 was to become familiar with the state of the art at the other three levels. Thus, I demonstrate what I have learned about those more technical levels in Chapter 2. However, not all the techniques I describe for fungi will apply to mammalian genomes, which tend to be more complex.

Although OrthoMaM is mainly intended for use by phylogenists (Ranwez et al, 2007), I chose it for this gerontological study because it is curated and contains the highest-quality sequences currently available. However, for any comparative study of mammalian aging, OrthoMaM's selection of species is ultimately inadequate, because it contains relatively few species and also does not represent the full diversity of eutherian lifespans. Relatedly, as detailed in Section 1.3.1, two major problems are that all but one of the shortest-lived species (shrew) is a rodent, and most of the longest-lived species are primates, thus conflating longevity and phylogeny given a naive approach. Table 1-4 lists additional species whose genomes should ideally be included in any gerontological database. Three of these species' genomes have very recently been sequenced, to varying degrees of completeness, by short read-based methods. It should be possible, although labor-intensive, to determine whether the two conclusions of this chapter are robust to addition of these newly-sequenced proteomes to the original OrthoMaM dataset.

In compiling a gerontological genome database for any set of organisms, the primary constraint is the availability of accurate longevity records for each species. This

is because to determine MLS often takes many years and, for the case of a population in captivity, requires that we know how to maintain good health (Finch, 1990). The AnAge database (de Magalhães and Costa, 2009) reports acceptable- or high-quality estimates of MLS for 883 eutherian species in 19 orders (of 21 total) and 97 families (of 108 total), so this should be considered the upper limit on the size of such a gerontological database in the near future. Because the base cost of genome sequencing has dropped dramatically since the introduction of short-read sequencing, given sufficient funding it is possible that a genome database near this size could be assembled within a decade, which would increase the power of the method described here and would certainly prove useful in many other comparative efforts. Given the short read-based sequencing that is most prevalent today, I anticipate that the two biggest areas of challenge in assembling such a genome database would be (1) collection and coordination of DNA samples from zoos and other sources around the world and (2) ensuring high *de novo* assembly completeness and contiguity of these complex and repeat-rich genomes, eg through development of cheaper and less biased methods for construction of jumping libraries.

A final gerontological consideration in deciding which genomes to sequence is that for species of greatest interest, sequencing multiple individuals within the same species may be warranted. Due to the current lack of available data, for non-model organisms it has been implicitly assumed, in this study and many others (eg, Jobson, Nabholz, and Galtier, 2010; Li and de Magalhães, 2011), that the proteome of a single sequenced individual represents all fixed substitutions present in the species. However, at

least a quarter of human genes are known to contain a coding-region SNP (Stetson et al, 2003), and there is no reason *a priori* to assume that other species' genomes are any less diverse. So, given that the mammalian proteome is a large search space (several million positions), this *one individual, one species* assumption could cause false positives for methods with position-level resolution such as mine. A remedy would be to sequence the genomes of several unrelated individuals within each species to get a sense of which divergent positions are truly fixed substitutions and which are only (non-fixed) single-nucleotide polymorphisms (SNPs). When automatic and manual approaches are combined to identify longevity-selected positions, this would be of particular interest for unique species such as those listed in Table 1-4. I expect that this consideration will increase in importance when comparative analysis is extended to noncoding regions, where SNPs occur even more frequently than in coding regions.

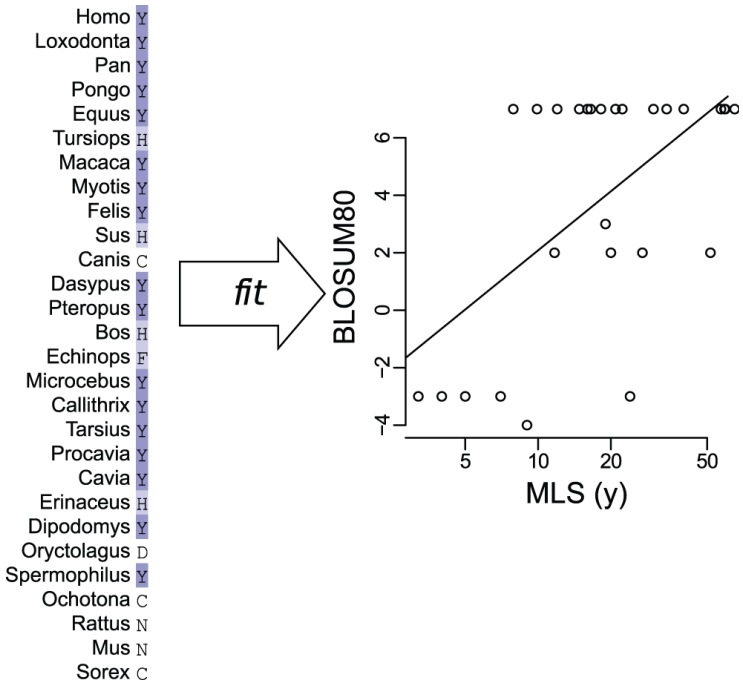

Figure 1-1. Conceptual example of multiple regression method

This example shows my regression method applied to a single aligned amino acid position, Y465 of AMHR2. For full result of this fit, see Appendix B. **Left.** Characters shown ordered by species MLS. For each non-human species, calculate the similarity score (*BLOSUM80*) for the species' amino acid character versus the human character (here Y); eg, this score for *Tursiops* would be the similarity score for H versus Y, which is 2. **Right.** Now, fit the MLS of all non-human species to their similarity scores; eg, *Tursiops'* contribution to this fit is the point (52, 2). Not shown are the steps to correct for mutation rate and shared phylogeny, and the simultaneous fit of body mass. For this column, the data provide relatively strong support for a nonzero slope in the fit of similarity to MLS, even given trends in mutation rate, phylogeny, and body mass, and so this position is assigned a relatively significant p-value (p_{MLS}<0.01).

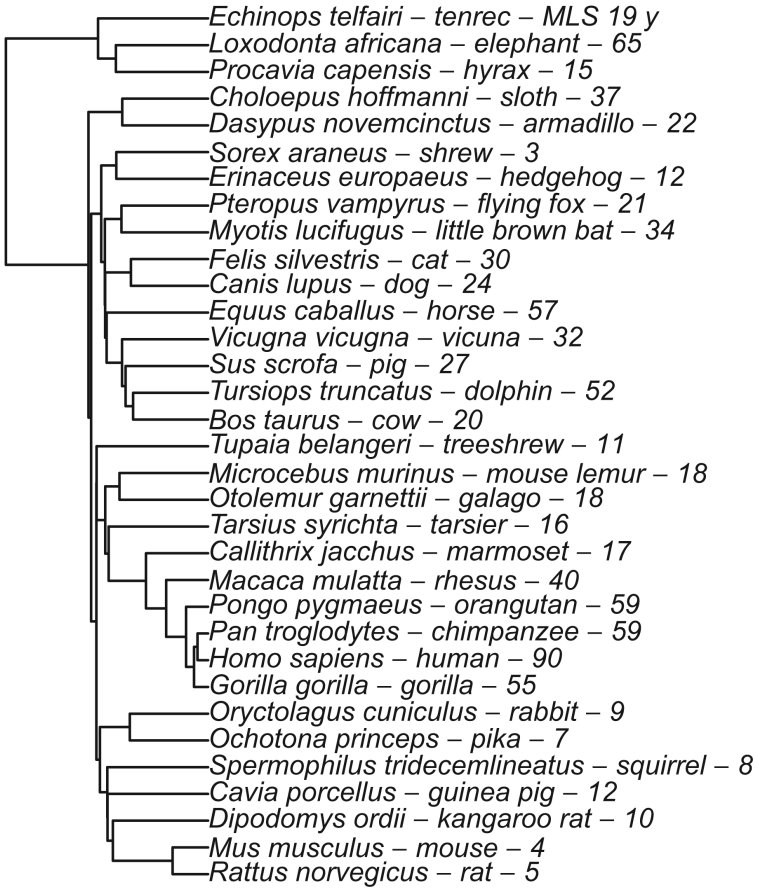

Figure 1-2. Phylogeny of species used in this study

Shown next to each species' node are its binomial, common name, and MLS.

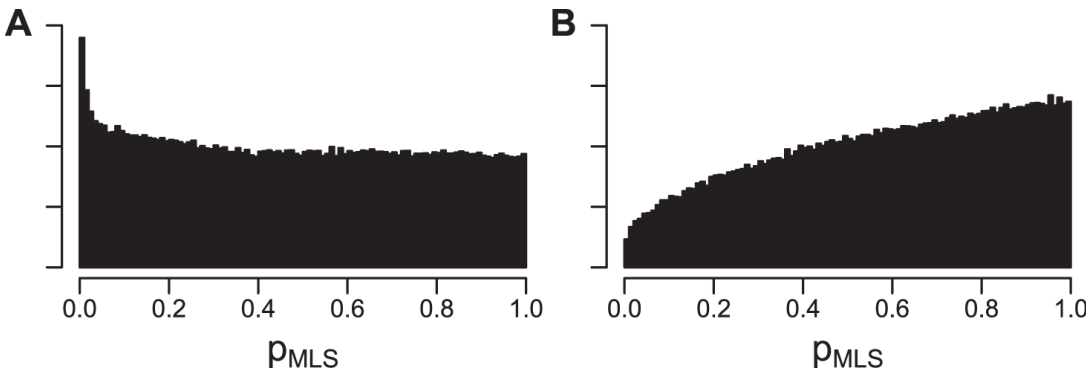

Figure 1-3. Density histograms of p_{MLS} values

A. Real data. **B.** Randomized control. Each histogram shows the n=407,568 fit positions with b_{MLS}>0 in the real dataset. See Section 1.3.2 for details.

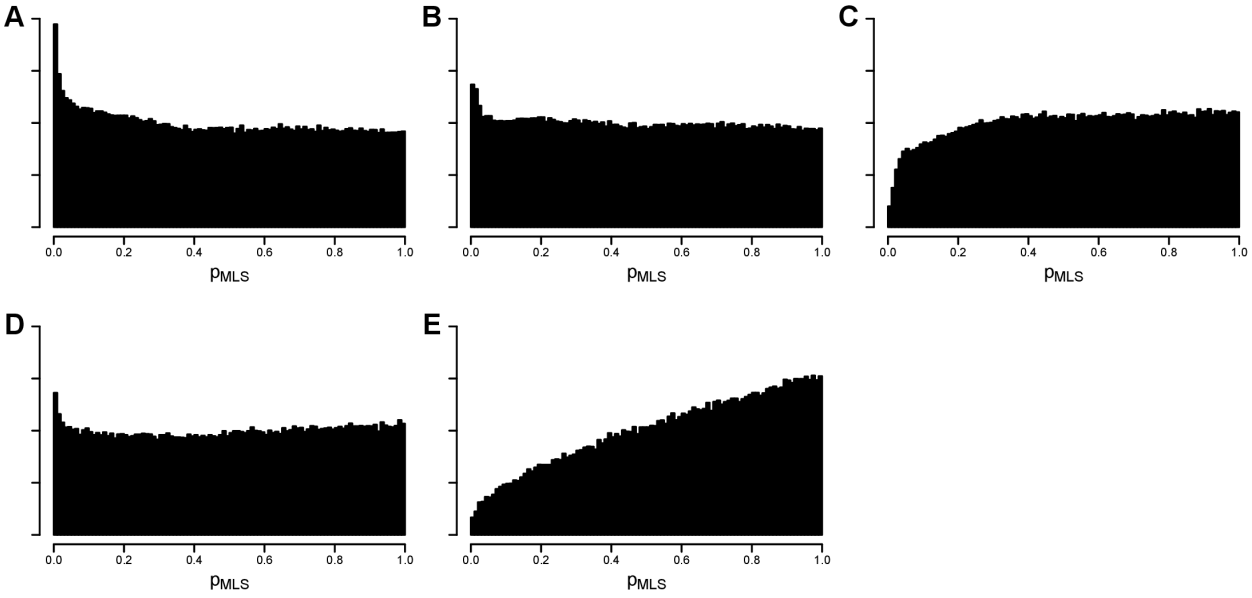

Figure 1-4. Density histograms of p_{MLS} values yielded by fitting alternate input sets

Complete methods for these fits are described in Section 1.2.2. As in Figure 1-3, only fit positions with b_{MLS}>0 are included. **A.** Chimp *(Pan troglodytes)* is the reference, and human is absent. **B.** Rhesus *(Macaca mulatta)* is the only primate fit. Human is the reference. **C.** Dog *(Canis lupus)* is the reference. **D.** Scores taken from BLOSUM62 instead of BLOSUM80. Real data. **E.** Randomized control data for (**D**).

```
                460       470       480       490       500       510
                |    |    |    |    |    |    |    |    |    |    |    |
          Homo  ERRRPYIPSTWRCFATDPDGLRELLEDCWDADPEARLTAECVQQRLAALAHPQESHPF
    Loxodonta   ERRRPYIP-TWHCFAMEPGGLQELLEDCWDADPEARLTAECVQQRLAA-AHP-EAHPC
           Pan  ERRRPYIPSTWRCFATDPDGLRELLEDCWDADPEARLTAECVQQRLAALAHPQESHPF
        Pongo   ERRRPYIPSTWRCFATDPDGLRELLEDCWDADPEARLTAECVQQRLASLAHPQESHSF
        Equus   ERRRPYIPSTWHCFATDPGGLRELLEDCWDADPEARLTAECVQQRLAALAQPQEAHSF
      Gorilla   ----------------DPDGLRELLEDCWDADPEARLTAECVQQRLAALAHPQESHPF
     Tursiops   ERRRPHIPSTWCCFATDPGGLRELLEDCWDADPEARLTAECVQQRLASLARPQEAHPF
       Macaca   ERRRPYIPSTWRCFATDPDGLRELLEDCWDADPEARLTAECVQQRLAALAHPQES-PF
       Myotis   ERRRPYIPPTWCCFATDPGVLREILEDCWDADPEARLTAECVQQRLAALAHPQEAHLF
        Felis   ERRRPYIPSTWHCFTTDPGGLRELLEDCWDADPEARLTAECVQQRLAALARPQEARPF
          Sus   ERRRPHVPSTWSCFATDPGGLRELLEDCWDADPEARLTAACVQQRLAALTHPQEARPF
        Canis   ERRRPCIPSTWHSFTTDPGSLRELLEDCWDADPEARLTAECVQQRLAALAHPQEAQPF
      Dasypus   ERRRPYIPSTWCSFSTEPGGLRELLEDCWDADPEARLTAECVQQRLATLAHPQEVHPL
     Pteropus   ERRRPYIPPTWHCFATDPAGLRELLEDCWDADPEARLTAECVQQRLATLAHPQEAHSF
          Bos   ERRRPHVPSTWCCLTTDPGGLRELLEDCWDADPEARLTAECVQQRLAALASPEEAHLF
     Echinops   ERRRPFVPPTWRFFTAEPGELRELLEDCWDADPEARLTAECVQQRLAXXXXXXXXXXX
  Microcebus    ERRRPYIPSTWCCFVTDPGGLRELLEDCWDADPEARLTAECVQQRLAALAHPQEAHSF
  Callithrix    ERRRPYIPSTWRCFATDPDGLRELLEDCWDADPEARLTAECVQQRLAALAHPQESHPF
     Tarsius    ERRRPYIPSTWYCFAT-----------CWDPDPEARLTAECVQQRLAALAHPQEAHPF
    Macropus    ERRRPXXXXXXXXXXXXXXXXXXELLEDCWDSDPEARLTAECIQHRLXXXXXXXXXXX
    Procavia    ERRRPYIPSTWHCFTTEPGRLRELLEDCWDADPEARLTAECVQQRLAALAHPQETQPC
       Cavia    QRRRPYIPSTWGCSIRDPGGLRELLEDCWDADPEARLTAECVQQRLATLTYPGEADCL
   Erinaceus    ERRRPHVPTTWHCFSTDPGGLRDLLEDCWDADPEARLTAECVQQRLAALASSPETCPA
   Dipodomys    ERRRPYIPPTWNYFATDLSGLRELLEDCWDADPEARLTAECVQQRLAALDHPQELHSF
 Oryctolagus    ERRRPDIPSSWCCFATDPGGLRELLEDCWDADPEARLTAECVQQRLVALVHPQEAQPC
Spermophilus    ERRRPYIPSTWISFVT-----------------------------------------
    Ochotona    ERRRPCIPDSWHCFVTEPGALRELLEDCWDPDPEARLTAECVQQRLAAMLHPPEARPF
      Rattus    ERKRPNIPSSWSCSATDPRGLRELLEDCWDADPEARLTAECVQQRLAALAYPQVASSF
         Mus    ERKRPNIPSTWSCSATDPRGLRELLEDCWDADPEARLTAECVQQRLAALAYPHGASSF
       Sorex    ERRRPCIPPAWHGCPTVPAGLREVLEDCWDADPEARLTAACVQLRLAALGPQQASGPG
```

Figure 1-5. OrthoMaM alignment of the C-terminal region of the AMHR2 kinase domain

Orthologs are ordered by species MLS. The five longevity-selected positions in this region (Y465, T469, F473, E513, and H515) are highlighted in gray. *X* indicates regions that were masked due to excessive divergence (Section 1.2.1). Long regions of gaps are not necessarily real genome deletions, but are more likely to have been missed during genome assembly or annotation.

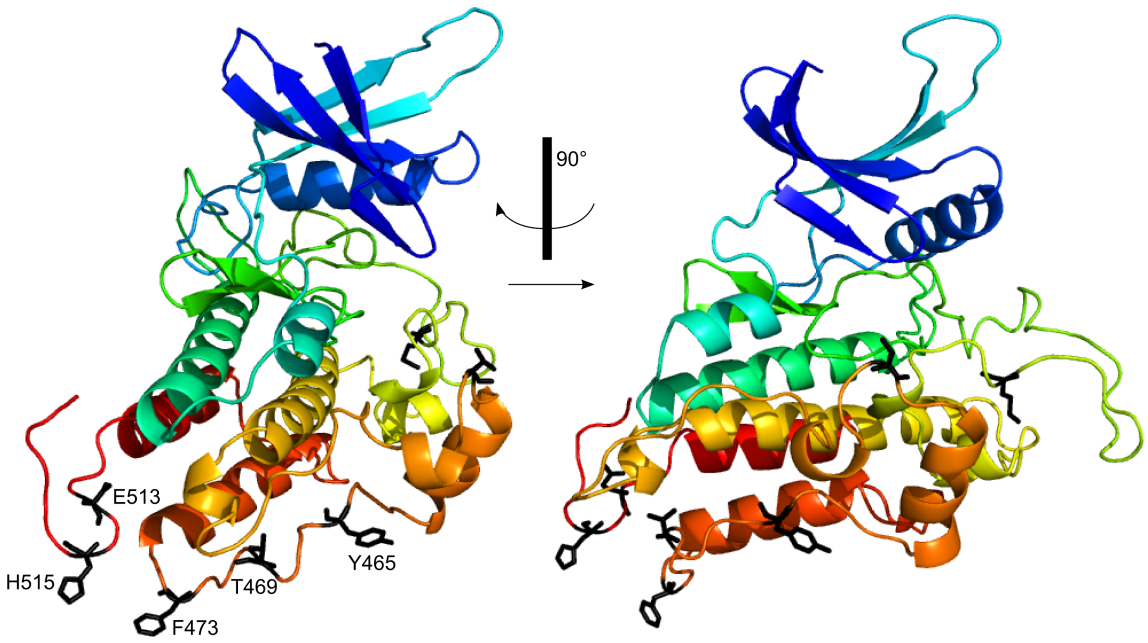

Figure 1-6. AMHR2 kinase domain mapped onto experimental structure of BMPR2 kinase domain

Structure is rainbow-colored by position in sequence, with N-terminus in blue and C-terminus in red. All seven longevity-selected positions found in this domain are shown as black sticks and are further described in Appendix B. The five longevity-selected positions discussed in the text are labeled; they are found on the αG–αH loop (Y465, T469, and F473) and the C-terminal loop following αI (E513 and H515).

subset	**columns ($\times 10^6$)**	**alignments**
include human	7.01	12,746[a]
... selected	3.64	7,708[a]
..... fit	0.73	7,590[a]
..... conserved	2.91	118[b]

[a]Number of alignments that include at least one position in the indicated subset.

[b]Number of alignments that include only conserved positions.

Table 1-1. Number of positions and alignments selected, fit, and conserved

These three terms are defined in Sections 1.2.1 and 1.2.2.

	α-helix	**β-strand**	**random coil**	**total**
human proteome	2,108,348 (29.95%)	956,422 (13.59%)	3,974,432 (56.46%)	7,039,202 (100%)
real data	1,922 (24.95%)	736 (9.56%)	5,044 (65.49%)	7,702 (100%)
randomized control	160 (23.26%)	62 (9.01%)	466 (67.73%)	688 (100%)

Table 1-2. Predicted secondary structure in all positions of the human proteome and in two subsets of longevity-selected positions

functional class	***p***	**examples**
extracellular region	5.1e-5	BTD, CP, LAMA2, LAMA3, PRSS12
cytokine-mediated signaling pathway	0.033	JAK1, IL31RA, IL6ST, LIFR, RIKP1, KIT
protein tyrosine kinase activity	0.018	JAK1, ROS1, MST1R, NIN, OBSCN, KIT, TYK2
multicopper oxidase; copper ion binding	9.4e-3	AFP, CP, F8, HEPHL1
developmental process	0.041	AFP, AMHR2, ALOX15, CFTR, ATR, ATM, BRCA1, RECQL4
motor activity; myosin complex	0.015	KIF18B, KIF20B, KIF22, MYO5C, MYO7B, MYO18A
complement and coagulation cascades; humoral immune response	0.061	F8, F11, CR2, C1R, CFD, LTF, CD83
serine-type endopeptidase activity	0.063	F11, C1R, CFD, KLK6, LTF, PRSS12

Table 1-3. Selected clusters, not mutually exclusive, of ontology terms enriched in top protein domains

binomial	**common name**	**MLS (y)**	**reference (if sequenced)**	**interest**
Balaena mysticetus	bowhead whale	211	Keane, de Magalhães, et al, unpublished	longest-lived mammal
Eubalaena spp	right whale	70		shorter-lived confamilial of bowhead
Blarina brevicauda	short-tailed shrew	4		a second short-lived shrew species
Heterocephalus glaber	naked mole rat	28	Kim et al, 2011	small long-lived rodent
Hystrix brachyura	Old World porcupine	27		large long-lived rodent
Microgale dobsoni	Dobson's long-tailed tenrec	6		shorter-lived confamilial of *E telfairi*
Setifer setosus	greater hedgehog tenrec	14		shorter-lived and larger confamilial of *E telfairi*
Trichechus manatus	Caribbean manatee	56	Broad Institute, unpublished	sirenian that continually replaces teeth
Dugong dugon	dugong	73	Broad Institute, in progress	long-lived sirenian that cannot replace teeth
Cebus capucinus	white-faced capuchin	54		long-lived small primate

Table 1-4. Eutherian species of gerontological interest recommended to include in search for longevity-selected positions

MLS estimates from AnAnge (de Magalhães and Costa, 2009). For sirenian discussion and references, see Finch (1990), p 201.

CHAPTER TWO
Comparative genome sequencing of the toxigenic black mold *Stachybotrys*

2.1. INTRODUCTION

2.1.1. Personal motivation

This chapter is adapted from a manuscript conceived and prepared in the course of my graduate work (Semeiks et al, 2013). The genome, gene, and protein sequences described have been submitted to NCBI, and will be available under Bioproject PRJNA186748.

If my primary interest is the problem of human aging (Section 1.1.1), then why for nearly a year did I choose to study the genome of a mold that is all but certainly irrelevant to the biology of aging? I have two answers, both related to the difficulty of studying the genomics of aging.

The first answer is technical. As an initial step toward correcting the weaknesses I observed in my mammalian gerontological dataset (Section 1.4.2), my goal was to learn how to produce new genome and proteome assemblies *de novo* from start to finish. It is still quite difficult to solve a mammalian genome *de novo*, because the methods required are arcane, costly, and still evolving (eg, Gnerre et al, 2011; Williams et al, 2012). In contrast, fungal genomes are more easily soluble due to their smaller size and reduced complexity; yet building a fungal genome entails many of the same basic techniques in use for mammals. In learning these techniques, I have accomplished my goal, and in Section 2.2 I describe in detail what I have learned.

The second answer is biological. Originally, *Stachybotrys* was suggested (by my collaborator Dominka Borek) as a good fungus to sequence because it was thought to have a small genome and was one of the few known medically-relevant organisms that had remained unsequenced. However, in reading the *Stachybotrys* literature I realized something more: here was an opportunity to design a clean study in comparative genomics that answered a well-defined question of genetics and toxicology. In short, the problem I describe here has many of the desirable scientific qualities that the problem of human aging is currently lacking. It remains an open question whether it is possible to think about the problem of aging in ways that promote these qualities.

2.1.2. The problem of mutually exclusive toxin chemotypes in *Stachybotrys*

Stachybotrys is a genus of filamentous fungi found in soil worldwide (Jarvis, 2003). It can also inhabit damp buildings. It is mainly a saprophyte that feeds by degrading cellulose and other dead plant matter. However, it is related to cellulolytic plant pathogens including *Fusarium* and *Myrothecium*, and there is a report of soybean invasion (Li et al, 2002). *Stachybotrys* does not infect animals, but it does produce a variety of toxins that have killed livestock and sickened humans after contact with contaminated feed, most famously in Ukraine in the 1930s. More recently, several studies have suggested various links between *Stachybotrys*-infested buildings and poor health, but confounders are many (Kuhn and Ghannoum, 2003).

At the basic level, there is a puzzle of genetics in the toxins produced by *Stachybotrys*. Harmful *Stachybotrys* products include both proteins (Shi, Smith, and Miller 2011) and secondary metabolites (Pestka et al, 2008). Of these, the two most well-known classes of secondary metabolite toxins are the trichothecenes and the atranones (Figure 2-1). Both are terpenoids, but they are not otherwise related in structure. The more toxic class, trichothecenes, are strong inhibitors of protein synthesis, and structurally they are further divided into two subclasses, simple and macrocyclic trichothecenes, with the latter subclass including the highly-toxic compounds called satratoxins (intranasal median lethal dose [LD_{50}] ~1 mg/kg in rodents [Jarvis, 2003]). Of the ~200 strains of *Stachybotrys* that have been tested, all can make simple trichothecenes (Andersen, Nielsen, and Jarvis, 2002). However, only a third of these strains can make macrocyclic trichothecenes (eg, satratoxins). Of the other two-thirds, most can make the less-toxic atranones; in fact, they are the only known atranone-producing organisms. Significantly, a strain of *Stachybotrys* that makes both satratoxins and atranones has never been observed, meaning that these chemotypes are likely mutually exclusive. I am not aware of another eukaryote with such a drastic difference in chemotype, and *a priori* there is no apparent biochemical rationale.

To determine the genetic bases for the two chemotypes of *Stachybotrys* and to compare *Stachybotrys* to other trichothecene toxin producers including *Fusarium* and *Trichoderma,* I have sequenced and assembled *de novo* the genomes of four cultured *Stachybotrys* strains. Two of these strains make atranones, and the other two make

satratoxins. I report some global properties of these genomes, most notably an unexpected richness of polyketide synthase (PKS) genes. I then present the core trichothecene cluster (CTC) of *Stachybotrys,* which diverges significantly from the CTCs of other trichothecene producers, with a genomic context that appears to be chemotype-specific. Finally, I use comparative methods to show that toxin chemotype in *Stachybotrys* likely arises from the presence of strain-specific secondary metabolite biosynthesis gene clusters, including three satratoxin-specific clusters and a novel 35-kbp locus that I have named the core atranone cluster (CAC).

2.2. MATERIALS AND METHODS

2.2.1. *Stachybotrys* culture, DNA extraction, and library construction

Stachybotrys strains were kindly provided by Kristian F. Nielsen (Center for Microbial Biotechnology, DTU, Denmark). Initially, fungus was grown on potato dextrose agarose to establish monoclonal populations by single-spore selection; these monoclonal populations were used for all subsequent procedures. Strain identities were verified by PCR-based sequencing of *TRI5* (Andersen et al, 2003). For sequencing libraries, hyphae were grown in 3-ml tubes of potato dextrose broth at 25°C in the dark for 1—2 weeks until confluent. Genomic DNA for sequencing libraries was obtained by a method based on cetyltrimethylammonium bromide (CTAB) disruption and phenol-chloroform extraction that is similar to a previously-described method (Cruse et al, 2002). Fresh hyphae were drained of media and pulverized in liquid N_2. The sample was added to a tube containing hot 2x CTAB buffer and n=3 5-mm glass beads, and then bead-beaten on a vortexer for 1 mn. DNA was extracted with 25:24:1 phenol:chloroform:isoamyl alcohol, treated with Riboshredder (Epicentre) for 30 mn at 37°C, and precipitated with isopropanol.

Multiplexed Illumina DNA fragment libraries were constructed as follows. For each strain, 500—1000 ng genomic DNA was sheared by sonication (Bioruptor, Diagenode) to ~500 bp. Fragments were end-repaired (NEBNext End Repair Module, NEB), dA-tailed (NEBNext dA-Tailing Module, NEB), and ligated (NEBNext Quick Ligation Module, NEB) to custom Y-adapters that included strain-specific 4- or 5-bp barcodes. Each respective reaction product was purified with Agencourt AMPure XP beads (Beckman). Ligated product was size-selected to 350 bp (nominal) by electrophoresis on 2% agarose, excision, and gel extraction (MinElute Gel

Extraction kit, Qiagen) overnight at room temperature. Following size selection, each library was amplified by PCR (Phusion High Fidelity PCR Master Mix with GC Buffer, NEB) using the standard Illumina primers, with 3 ng template, 0.5 μM primers, 12 PCR cycles per reaction, and other reagents and reaction parameters per NEB's instructions. All PCRs in this study were performed on a PCR Express thermal cycler (Thermo Hybaid). PCR product was size-selected as above to remove unreacted primer and adapter dimers. The four libraries were then pooled to 2.5 nM each and submitted to the UT Southwestern Genomics Core for sequencing on a single lane of an Illumina HiSeq 2000.

2.2.2. Genome assembly and resequencing of specific loci

Base-calling of reads from intensity data was accomplished with AYB 2.11 (Massingham and Goldman, 2012). This yielded 394 million paired reads, 72% of which passed purity filtering. Pure reads were demultiplexed and sequencing artifacts (including reads containing adapter and primer sequence) were removed using custom scripts. The remaining reads were end-trimmed to quality 20 or higher. Reads were spectrally corrected with Quake 0.3.0 (Kelley, Schatz, and Salzberg, 2010) and then assembled *de novo* into contigs and scaffolds with SOAPdenovo 1.05 (Li et al, 2010) and AbySS 1.3.4 (Simpson et al, 2009). For each strain and assembler, we produced n=27 (SOAPdenovo) or 10 (AbySS) separate assemblies, in each case iterating K from 31 to 81. I then selected as representative an assembly with a subjectively good combination of total size and N_{50} length; these parameters were generally robust over a wide range of K values. The final

SOAPdenovo assemblies had the following K values: strain 40285, K=43; strain 40288, K=53; strain 40293, K=45; and strain 7711, K=51.

Two loci discussed in Results are each split over two different sequences in our assemblies: the CTC of strain 40293 and the CAC of strain 40288. I verified by Sanger sequencing of PCR amplicons that each of these regions is in fact a single contiguous locus, although in each case the two flanking regions are separated by an estimated 50—100 bp repeat that has not proven possible to sequence by either the parallel or the Sanger method. The PCR primers I used were as follows; they include two independent pairs for each locus. For CTC, primer pair 1: forward TTGGTCGTCTCTTGAGATTCACTGGC, reverse CCAAAGTGGAAGGTTCATGGTTGAGC; primer pair 2: forward TTCCCTTGCTTCCGTACCTTATTCCC, reverse TTATTCCCATCCTTTGTCCGGAGTGG. For CAC, primer pair 1: forward AAGTCTCATCTTGCCTCGGAATCAGG, reverse AGTTCAACCTTCTCTCAGGAACAGGG; primer pair 2: forward CCTGATCTTGGACATTGCTATTCCGC, reverse TTTGCATGAGCTAAACACACCGGG. The CTC was amplified in a 50 µl reaction including 5 µl Accuprime Pfx reaction mix, 0.4 µl Accuprime Pfx DNA polymerase, 0.3 µM each primer, and 3 ng genomic DNA from strain 40293. The CAC was amplified in a 100 µl reaction including 10 µl Accuprime Pfx reaction mix, 0.8 µl Accuprime Pfx DNA polymerase, 0.3 µM each primer, and 3 ng genomic DNA from strain 40288. PCR parameters included 30 (CTC) or 35 (CAC) cycles of denaturation at 95°C for 15 s,

annealing at 55°C (CTC) or 58°C (CAC) for 30 s, and extension at 68°C for 60 s. Before sequencing, both products were gel purified (Minelute Gel Extraction Kit, Qiagen), reamplified with the same PCR parameters as were used for the first reaction, and repurified (Wizard SV kit, Promega).

2.2.3. Proteome assembly

For proteome assembly (ie structural annotation) I used MAKER 2.26 (Holt and Yandell, 2011), which incorporated both homology-based (BLAST 2.2.26 and Exonerate 2.2.0) and *de novo* methods (GeneMark 2.3e and Augustus 2.6.1) and output only transcript models that were supported by both types of evidence. For each strain, MAKER was run twice. The second pass was run in reannotation mode, and included as homology targets all four proteomes output by the first pass. On both passes, other homology targets included the Swissprot database (current build as of 20 Aug 2012) and the three *Fusarium* proteomes (Ma et al, 2010). Full input parameters for MAKER are listed in Appendix G.

To compare the genomes of *Stachybotrys* and *Fusarium*, features of the *F. graminearum* genome were obtained from the *Fusarium graminearum* Genome Database (Wong et al, 2011).

2.2.4. Genetic nomenclature of *Stachybotrys*

In naming *Stachybotrys* genes and proteins, I chose to follow the conventions in use for *E. coli* and *Fusarium.* All gene and protein names are three letters followed by a

number. Gene names are all-uppercase and italicized, eg "*TRI5*". Corresponding protein names are capitalized and in standard face, eg "Tri5".

2.2.5. Protein and rRNA phylogenies

To construct phylogenies, proteins were downloaded from NCBI using the accessions listed in the cited references. Following protein alignment with the L-INS-i method of mafft 6.903b (Katoh and Toh, 2008), any position containing a gap was discarded. Protein phylogenies were inferred with PhyML 20120412 (Guindon et al, 2010), using 100 bootstrap replicates and otherwise default parameters.

2.2.6. Proteome comparisons and PKS inventory

To obtain groups of homologous proteins as described in Results, OrthoMCL 2.0 (Li, Stoeckert, and Roos, 2003) was run on nine proteomes using default parameters, including BLAST E-value cutoff of 1e-5. Protein domains were identified by searching the nine proteomes with RPS-BLAST 2.2.26 against the NCBI Conserved Domain Database (CDD; current as of 2 Aug 2012), and then filtering results using the NCBI Specific Hits algorithm (Marchler-Bauer et al, 2011). Domain enrichment analysis was done using Fisher's exact test with correction for multiple testing, as described in Appendix C. All domain identifiers mentioned in Results are the unique "domain short names" assigned by CDD.

I define a putative *Stachybotrys* PKS as any predicted protein that includes all three of the CDD domains PKS, PKS_AT, and either PKS_PP or PP-binding.

2.2.7. Identification of chemotype-specific gene clusters

I define a chemotype-specific gene cluster as a locus containing at least three genes, all of which are both chemotype-specific and contiguous. Chemotype-specific gene cluster candidates were identified by collating OrthoMCL gene clusters with chemotype-specific loci found by whole-genome alignment with Mugsy 1r2.2 (Angiuoli and Salzberg, 2011); I used custom scripts to process Mugsy's output in this manner. After identification of candidate clusters, OrthoMCL results were analyzed to exclude those clusters that did not meet the above definition of "chemotype-specific gene cluster". Of the 10—20 total candidate clusters, most were easily excluded because Mugsy had missed strain-specific alignments, as verified by BLASTN, and simultaneously there were obvious chemotype-independent orthologs found by OrthoMCL.

2.3. RESULTS AND DISCUSSION

2.3.1. Sequencing and assembly of *Stachybotrys* genomes

The four *Stachybotrys* strains that I sequenced are shown in phylogenetic context in Figure 2-2. The strains include two species, *S. chlorohalonata* (IBT strain 40285) and *S. chartarum* (IBT strains 40288, 40293, and 7711), that are distinguishable by both morphology and molecular markers. Strains 40285 and 40288 make atranones, while strains 40293 and 7711 make satratoxins (Andersen et al, 2003).

The genomes of these four strains were obtained by massive parallel sequencing on an Illumina Hiseq 2000. For each strain, I constructed a separate 300-bp nominal genomic fragment library. I multiplexed these libraries in order to combine them all on a single sequencer lane, which yielded ~70 million 101-bp reads per strain after demultiplexing and error correction. I then assembled each genome independently with SOAPdenovo (Li et al, 2010), followed by structural annotation of each assembly with MAKER (Holt and Yandell, 2011) using a cross-strain iterative strategy.

Table 2-1 summarizes the genome and proteome assemblies, and for comparison also includes a finished assembly of the trichothecene producer *Fusarium graminearum* obtained by Sanger sequencing (Ma et al, 2010). As shown in the table, these five genome and proteome assemblies are similar in size, although those of the *S. chlorohalonata* strain 40285 are slightly smaller than the three *S. chartarum* strains. Except for N_{50} length, the features of all four *Stachybotrys* assemblies, eg their short introns and sparse repeat content, are comparable to the finished *F. graminearum* assembly. This is consistent with the fact that *Fusarium* is one of the closest relations to *Stachybotrys* whose genome has been sequenced.

I independently assembled each strain with ABySS (Simpson et al, 2009) to validate the SOAPdenovo results. While scaffold N_{50} length obtained from ABySS was reduced by 20—80 kbp versus scaffold N_{50} length from SOAPdenovo, total genome sizes were nearly identical. Also, the seven gene clusters I describe below for the SOAPdenovo build were appropriately present in the ABySS assemblies. Specifically, in both the ABySS and SOAPdenovo assemblies, the core trichothecene cluster had identical architecture in all four strains, and the six other novel clusters I describe were consistently atranone- or satratoxin-specific.

2.3.2. Comparative proteome content of *Stachybotrys*

To estimate the completeness of my proteome assemblies and compare them to those of other sequenced fungi, I used two methods. First, I used CEGMA (Parra, Bradnam, and Korf, 2007) to search the *Stachybotrys* genome assemblies for 458 proteins known to be highly conserved in eukaryotes. By this criterion, all four *Stachybotrys* assemblies are 98% complete, with identical completeness found for *F. graminearum* and the two other sequenced *Fusarium* genomes, *F. oxysporum* and *F. verticillioides,* neither of which make trichothecenes. All proteins found by CEGMA were independently found by MAKER in the full *Stachybotrys* proteomes, suggesting that my four genome assemblies are relatively complete.

Second, I identified groups of homologs in the proteome assemblies with OrthoMCL (Li, Stockert, and Roos, 2003). For diversity, I used nine proteomes in total: the four *Stachybotrys* assemblies; the three *Fusarium* proteomes named above (Ma et al,

2010); and two more divergent model fungi, *Aspergillus nidulans* (Arnaud et al, 2012) and *Saccharomyces cerevisiae* (Cherry et al, 1998). OrthoMCL clustered these proteomes into 16,311 groups, each containing at least two proteins. Of these groups, 2,177 contained exactly one orthologous sequence from each of the nine proteomes. Using this subset of proper orthologs, I constructed a robust phylogeny (Figure 2-3) and quantified proteome divergence by calculating pairwise sequence identities (or *proteome identities;* Table 2-2). This phylogeny matches both accepted taxonomy and a previous molecular phylogeny (Wu et al, 2003), grossly validating both my *Stachybotrys* proteomes and my OrthoMCL-based method. As expected given prior analysis of *Stachybotrys* genetic markers (Andersen et al, 2003), the proteome identities indicate that the *S. chlorohalonata* strain 40285 is the most divergent of the four *Stachybotrys* strains, but this divergence is relative: there is 98% proteome identity between 40285 and any *S. chartarum* strain, versus 74% identity between *Stachybotrys* and *Fusarium* and >99% identity within strains of *S. chartarum*.

Figure 2-4 summarizes the distribution of homolog groups in the four genera. Of the 16,311 homolog groups, most included orthologs from *Stachybotrys* (68% of all groups) and *Fusarium* (80%). Many groups were exclusive to *Stachybotrys* (16% of all groups) or *Fusarium* (24%), perhaps reflecting genus-specific secondary metabolites or other phenotypes. As might be expected, most of the proteins in the 2,615 groups exclusive to *Stachybotrys* lack known domains (only 37% contain at least one domain from the Conserved Domain Database [CDD; Marchler-Bauer et al, 2011], vs ~65% of

all *Stachybotrys* proteins). Domain enrichment analysis (full results in Appendix C-1) revealed that of the *Stachybotrys*-exclusive protein domains, those enriched relative to the domains of non-exclusive *Stachybotrys* proteins likely have specialized functions such as mating enforcement (the CDD HET domain), degradation of plant materials (glycosyl hydrolases Glyco_hydro_61 and Glyco_hydro_6; several peptidase domains including M28; the pectate lyase domain Amb_all; and the cellulose-binding domain fCBD), and synthesis of novel secondary metabolites or other products (methyltransferases, acetyltransferases, and cytochrome P450 monooxygenases).

I also compared the whole domain compositions of the *Stachybotrys* and *Fusarium* proteomes, independently of homology considerations. Domain enrichment analysis (Appendix C-2) revealed remarkably few significant differences in gross domain composition between the two genera (only nine domains differentially present of 5,752 tested). Four CDD domains are enriched in the *Stachybotrys* proteome relative to *Fusarium.* Two of them, fCBD and Glyco_hydro_61, are also enriched in the *Stachybotrys*-exclusive proteins described above, and may reflect genus-specific differences in nutrient intake. The other two domains, PKS and PKS_AT, are respectively the ketosynthase and acyltransferase domains that are found constitutively in type I iterative polyketide synthases (PKSs). In fungi, PKSs are large proteins of variable domain architecture that are responsible for producing a diverse array of polyketide secondary metabolites (reviewed by Cox, 2007). It is noteworthy that each strain of *S. chartarum* conservatively encodes 35—37 PKSs, over twice as many as *Fusarium* and

more than any other fungus known, suggesting that a multitude of secondary metabolites from *Stachybotrys* remain uncharacterized (Appendix D). PKSs also appear to play roles in *Stachybotrys's* biosynthesis of trichothecenes and atranones, as discussed next.

2.3.3. The core trichothecene gene cluster of *Stachybotrys* diverges from those of other trichothecene producers

Many fungal secondary metabolites are made by products of genes that are found adjacent to one another in a single contiguous locus (Keller, Turner, and Bennett, 2005). I refer to such genetic loci as secondary metabolite biosynthesis (SMB) clusters. In the simple trichothecene producers *Fusarium graminearum* and *F. sporotrichioides*, a well-studied SMB cluster is the core trichothecene gene cluster (CTC). The *Fusarium* CTC encodes 11—12 genes, most of which are required to catalyze specific steps in trichothecene production (Brown et al, 2004). CTC sequences are also available for *Trichoderma arundinaceum* and *T. brevicompactum* (Cardoza et al, 2011); each contains only seven genes. This divergence of the *Fusarium* and *Trichoderma* CTCs reflects the divergence of trichothecene pathways between the genera. Most prominently, *Fusarium* makes only products modified at backbone position C-3, such as deoxynivalenol and T-2 toxin. In contrast, *Trichoderma* (like *Stachybotrys*) does not modify C-3, but makes exclusively trichothecenes modified at backbone position C-4, including trichodermol (Figure 2-1; McCormick et al, 2011).

Each of the four *Stachybotrys* assemblies includes a complete and identical ~30-kbp locus that I name the *Stachybotrys* CTC. The CTC structure is shown in Figure 2-5,

with more detail in Appendix E-1 (hypothesized protein functions) and Appendix F (full protein sequences). The CTC comprises nine genes, including putative orthologs of seven *Fusarium* and *Trichoderma* genes: the terpene cyclase *TRI5,* the acetyltransferase *TRI3*, the hydroxylases *TRI4* and *TRI11,* the transcription factors *TRI6* and *TRI10,* and a gene of unknown function *TRI14.* The remaining two genes in the *Stachybotrys* CTC are novel, so I name them by convention: the putative PKS *TRI17,* and adjacent to it the *TRI3* paralog *TRI18.*

The products of the *Fusarium, Trichoderma,* and *Stachybotrys* CTCs are consistent with both the divergence of the *Stachybotrys* CTC from that of *Fusarium* and the similar gene content of the *Stachybotrys* and *Trichoderma* CTCs (six genes are shared). However, I was surprised by the divergence in *Stachybotrys* gene order from the *Trichoderma* CTC, since the initial trichothecenes made by *Stachbotrys* and *Trichoderma* are identical. For example, versus *Trichoderma TRI5, Stachybotrys TRI5* is located within the CTC, and I cannot identify any simple rearrangements that would convert one CTC to the other. At least two additional data support the novel CTC architecture of *Stachybotrys*. First is the fact that two independent assemblers yielded the same sequence (Section 2.3.1). Second is the fact that the recently-sequenced CTC of the macrocyclic trichothecene producer *Myrothecium roridum* has a similar architecture, including mostly-conserved gene order and the presence of putative *TRI17* and *TRI18* orthologs (Robert H. Proctor, personal communication). The diversity of the CTC is consistent with

the hypothesis that it is a hotspot for insertion and deletion of enzyme-coding genes, in turn allowing for substantial structural diversity of trichothecenes.

I have identified two *Stachybotrys* loci outside of the CTC that contain paralogs of CTC genes. First, there is the satratoxin-specific cluster SC2, which contains paralogs of *TRI3* and *TRI4;* it is discussed in the satratoxin section below. Second, the assembly of strain 40293 includes a small scaffold that contains only two genes, which I name *TRI19* and *TRI20*, that are respective paralogs of *TRI5* and *TRI6.* With the arguable exception of *TRI8* (Section 2.3.5), I have not observed *Stachybotrys* orthologs of any other known *Fusarium* trichothecene biosynthesis *(TRI)* gene. The most surprising absence is that of the trichothecene exporter *TRI12,* which is present in the CTCs of both *Fusarium* and *Trichoderma* (Cardoza et al, 2011).

2.3.4. The products of the core atranone cluster likely suffice to make all known atranone species

I hypothesized that the two mutually-exclusive chemotypes of *Stachybotrys* were due to the presence of strain-specific SMB clusters. To test this hypothesis computationally, I searched the four *Stachybotrys* genome assemblies for loci that were present in both satratoxin strains but in neither atranone strain, or *vice versa*. The search strategy combined two methods. At the genomic level, I employed four-way whole-genome alignment, using Mugsy (Angiuoli and Salzberg, 2011). At the level of the proteome, I considered the sets of homologs compiled with OrthoMCL as described in Section 2.3.2. Whole-genome alignment was needed to show genomic context, but in

practice Mugsy did not correctly align some locus boundaries, so I manually adjusted its results as described in Methods. The search yielded a formal total of two atranone-specific and four satratoxin-specific gene clusters; I describe the satratoxin-specific clusters in Section 2.3.5.

The larger of the two atranone-specific gene clusters I name the core atranone cluster (CAC, or AC1; Figure 2-6, with details in Appendices E-2 and F). This is a ~35-kbp PKS-based cluster, and it has a nearly-identical architecture of 13—14 genes in both atranone strains. I name these genes *ATR1—ATR14*. The CAC is complete in the sense that both its flanking loci, or their orthologs, are present in all four strains.

I predict that the products of the CAC suffice to catalyze most or all steps of atranone synthesis, starting from geranylgeranyl pyrophosphate (GGPP; Figure 2-1). This prediction is based on two observations. First, the CAC is one of only two clusters exclusive to two relatively divergent strains of *Stachybotrys* (Figure 2-3 and Table 2-2). Second, the predicted CAC products satisfy some key constraints of the chemical model for atranone biosynthesis (Figure 2-7) proposed by Hinkley et al (2000). Specifically, the Hinkley model entails two particular reactions: the initial cyclization of GGPP to dolabellane, and a Baeyer-Villiger oxidation near the end of the scheme to convert atranones D and E to atranones F and G. Putatively, CAC products can catalyze both of these reactions: *ATR13* encodes a terpene cyclase (characteristic terpene cyclase motif DDXXE [Keller, Turner, and Bennett, 2005] and best BLAST hits [E < 1e-40] to related fungal terpene cyclases), while *ATR8* encodes a Baeyer-Villiger monooxygenase

(BVMO; characteristic BVMO motif FXGXXXHXXXWD [Fraaije et al, 2002]; best BLAST hits [E < 1e-65] to the BVMO phenylacetone monooxygenase from fungi, including *Paracoccidioides,* and bacteria). Although terpene cyclases are relatively common in the four *Stachybotrys* proteomes, the BVMO motif is very rare. There is only one other set of homologs that contain the BVMO motif. This second BVMO set has representatives in all four strains; OrthoMCL groups it separately from the atranone-specific pair found in the CAC; and each of its members are located in a chemotype-independent gene cluster that also contains glycosyl hydrolases, consistent with a possible role in feeding rather than secondary metabolism *per se.* Taken together, all these data suggest that the function of the CAC's products is to synthesize atranones.

Of the CAC's other predicted products, the largest is the reducing PKS Atr6. A BLAST search suggests that this protein is related to both fungal and bacterial PKSs, with the best hit to an uncharacterized PKS from *Aspergillus fumigatus.* Some other predicted CAC products include four oxygenases, three short-chain reductases, an esterase, and a methyltransferase. These are all plausibly involved in the various steps of atranone biosynthesis, although their specific roles must await experimental determination, because the types of reactions that they catalyze appear frequently in the Hinkley model (Figure 2-7).

If my predicted function for the CAC is correct, then it remains an open problem how atranone biosynthesis is regulated. Unlike the CTC and satratoxin-specific SMB clusters found in *Stachybotrys,* I have not been able to identify any transcription factors

or other putative regulatory genes within the CAC or nearby; the closest is a chemotype-independent *GAL4*-family gene that is 21 kbp upstream. Another example of a fungal SMB cluster lacking internal regulatory genes is the penicillin cluster of *Aspergillus nidulans* and other species (Spröte et al, 2008). A scan of the 14 putative CAC promoter regions revealed that two-thirds contain both the motif TGTCT and its reverse complement AGACA. (Specifically, 25 of the 28 promoter regions contain AGACA, 24 contain TGTCT, and 20 contain both motifs.) These inverted repeats may be bound by a single transcription factor, as yet unidentified. Alternatively, some CAC products may be widely expressed, with post-transcriptional regulation additionally possible; consistent with this hypothesis is the report that most atranone-producing *Stachybotrys* strains easily produce simple dolabellane derivatives in culture, but do not always produce atranones *per se* (Andersen, Nielsen, and Jarvis, 2002).

From a genetic perspective, it would be ideal to confirm the CAC's predicted function by exogenous expression in either a model organism such as yeast or, better yet, a satratoxin strain of *Stachybotrys*. However, at present this would be difficult. In addition to the uncertain regulation noted above, practical challenges include the large size of the cluster and the fact that, to my knowledge, *Stachybotrys* has not yet been used as a recombinant organism.

The second atranone-specific gene cluster is named AC2; its products are not shown. It is smaller than the CAC, spanning 12 kbp and containing six genes. Unlike the CAC, AC2 is missing one flank in our assemblies, meaning it may be incomplete. Also,

three of its genes are homologous to those of a second distinct locus conserved in all *S. chartarum* strains (on scaffold645 of 40288, scaffold1203 of 40293, and scaffold1305 of 7711). The largest gene in AC2 putatively encodes the phosphate transporter domain PHO4, and another encodes a helix-loop-helix (HLH) transcription factor. Two other genes yielded relatively weak BLAST hits (E ≈ 1e-4 in both cases) to cyclins and arrestins, respectively, suggesting overall that AC2 could be related to environmental phosphate sensing. Because phosphate-substituted compounds are used in synthesis of terpenes, specifically-regulated phosphate transport may be necessary for appropriate production of farnesyl pyrophosphate (FPP) or other atranone precursors.

2.3.5. Gene clusters specific to satratoxin strains of *Stachybotrys*

A general biosynthesis model for the satratoxins has been proposed, based on the known structures of similar molecules that are taken to be intermediates (Figure 2-8, adapted from Degenkolb et al, 2008). In this model, satratoxins and all other macrocyclic trichothecenes derive from trichodermol, first by sequential esterification of two sidechain species to C-4 and C-15 hydroxyl groups on the trichothecene skeleton, and second by condensation of the two sidechains to form the macrocycle. Based on their structures, the sidechains are plausibly polyketide products, although they would need to be modified by external hydroxylases to yield the primary hydroxyl groups that are observed. Optionally, PKS-independent reductases and methyltransferases may also be involved.

The whole-genome comparative method described above revealed four satratoxin-specific gene clusters, three of which encode the types of enzymes just described (Table 2-3; Appendices E-3, E-4, and E-5; and Appendix F). I refer to them as satratoxin clusters (SCs) 1—4, in order of size. The two largest, SC1 and SC2 (Figure 2-9), are classical PKS-based SMB clusters. SC3 (Figure 2-5) is smaller and is not a complete SMB cluster on its own, but it is found adjacent to the CTC. As shown in the figures, all three of these SCs are missing at least one flank in the assembly (ie, they are at the borders of their respective scaffolds), raising the possibility that in fact they are all located close to the CTC and can thus be easily coregulated.

SC1 (Figure 2-9 and Appendix E-3) is a 30-kbp cluster that contains ten genes, *SAT1—SAT10.* The largest genes are *SAT8,* which encodes a putative PKS with a conventional nonreducing architecture (Cox, 2007), and *SAT10*, whose putative product contains four ankyrin repeats (RPS-BLAST prediction) and thus may be involved in protein scaffolding. The putative short-chain reductase Sat3 may assist the PKS in some capacity. Sat6 contains a secretory lipase domain and is similar to the *Fusarium* trichothecene C-15 esterase Tri8 (BLASTP E-value 3e-93, 40% identity, 85% coverage), although it is even more similar to other unstudied proteins from *Fusarium* and *Aspergillus*. The adjacent gene *SAT5* encodes a putative acetyltransferase, and so the two together may effect endogenous protection from toxicity in the same manner as Tri8 and Tri101 of *Fusarium* (McCormick and Alexander, 2002).

SC2 (Figure 2-9 and Appendix E-4) is 20 kbp and contains six genes, *SAT11—SAT16,* the largest of which encodes the putative reducing PKS Sat13. SC2 is unique in that three of its putative products are paralogs of *Stachybotrys* products otherwise encoded exclusively by the CTC. Sat11 is a cytochrome P450 monooxygenase and a Tri4 paralog, while Sat14 and Sat16 are respectively complete and truncated paralogs of the acetyltransferase Tri3. Finally, the cluster may be regulated by the zinc finger protein Sat15, which is most similar to the putative LolU transcription factor reported in an SMB cluster of the grass-endophytic fungus *Neotyphodium* (Spiering et al, 2005). I can find only six putative LolU homologs in *Stachybotrys* 7711, and one also flanks the CTC of *M. roridum* (Robert H. Proctor, personal communication). Taken together with the novel architecture of the *Stachybotrys* CTC, these data indicate that SC2 may have originated as a duplication of the CTC and has subsequently undergone rearrangements and divergence in function.

In contrast to SC1 and SC2, SC3 (Figure 2-5 and Appendix E-5) is a small 10-kbp cluster that contains five genes, *SAT17—SAT21*. Although none of these genes encodes a PKS, the cluster itself is found adjacent to the CTC in satratoxin strains (Figure 2-5), suggesting that the two loci may be coregulated. One SC3 gene, *SAT21,* encodes a putative major facilitator superfamily-type transporter that may function to specifically export macrocyclic trichothecenes, analogously to Tri12 of *Fusarium* (Alexander, McCormick, and Hohn, 1999). The four other putative products of SC3 include the TauD

hydroxylase Sat17, the methyltransferase Sat18, the acetyltransferase Sat19, and the Cys6-type zinc finger Sat20; the latter likely assists in regulation of the cluster.

The smallest satratoxin-specific locus, SC4 (not shown), is in the middle of a chemotype-independent gene cluster that does not appear to encode any of the types of enzymes described above. I have not been able to predict the function of SC4, and so I mention it mainly for completeness. In general, versus the atranone case, I am not aware of any unusual chemistry proposed for the biosynthesis of satratoxins that would more specifically inform as to the relevance of any of these four chemotype-specific loci. Indeed, given the recent divergence of the satratoxin strains relative to the atranone strains (Table 2-2), it is possible that SC4 or other clusters are unrelated to satratoxin biosynthesis. A powerful way to further explore the function of these clusters genomically would be to search for them in the genome of *Myrothecium roridum,* a more divergent macrocyclic trichothecene producer that to our knowledge has not yet been fully sequenced.

2.3.6. Phylogenies for four trichothecene biosynthesis protein families in *Stachybotrys,* and functional implications

Four well-studied CTC proteins are Tri5, Tri4, Tri11, and Tri3. In particular, Tri5 and Tri4 are the earliest known enzymes in the trichothecene pathway: Tri5 cyclizes FPP to trichodiene, and Tri4 multiply hydroxylates trichodiene and its derivatives (McCormick et al, 2011). Both Tri4 and Tri11 are known to catalyze differing reactions

in *Fusarium* versus *Trichoderma* (Cardoza et al, 2011), resulting in two genus-specific series of trichothecenes (C-3 vs non-C-3 substituted, as discussed above). To infer the functions of these genes in *Stachybotrys* and more generally to explore the evolution of the CTC and SC2, I constructed maximum likelihood-based phylogenies of these four proteins and their novel paralogs (Figure 2-10). I used homologs from *Stachybotrys*, *Myrothecium* (only Tri5 and Tri4 are available), *Trichoderma*, and *Fusarium*. Partial 18S rRNA sequences are available for all four genera, and I used these to construct a reference phylogeny (Figure 2-10a). Excluding the *Stachybotrys* SC2 products and other paralogs, the topology of the 18S tree matches that of Tri4 (Figure 2-10c) and Tri3 (Figure 2-10e). However, the 18S tree differs from that of Tri5 (Figure 2-10b), in which *Trichoderma* Tri5 is divergent, and Tri11 (Figure 2-10d), in which *Fusarium* Tri11 is divergent. The Tri5 topology may relate to the fact that, uniquely, *Trichoderma TRI5* is external to the CTC (Cardoza et al, 2011). The Tri11 topology is consistent with *Stachybotrys* Tri11 conserving the function of *Trichoderma* Tri11, which is to hydroxylate the trichothecene skeleton at C-4 to yield trichodermol (Cardoza et al, 2011). While no functional prediction for *Stachybotrys* Tri4 can be made based only on this tree, I assume that (like *Trichoderma* Tri4 and versus *Fusarium* Tri4) it lacks the ability to hydroxylate C-3, since C-3 substituted trichothecenes have not been observed in *Stachybotrys* (McCormick et al, 2011).

That three of these four tree topologies (Tri5, Tri4, and Tri3) mostly match that of 18S supports a single origin for the CTC, at least in part, in the common ancestor of all

four genera. However, the *Stachybotrys* paralogs differ in this regard: while the 18S topology may be conserved for the Tri5 paralog Tri19, 18S topology is not conserved for the Tri4 paralog Sat11 (diverges before *Myrothecium* Tri4), nor for the Tri3 paralogs Tri18 and Sat12 (they form the outgroup to all Tri3 and Sat16). These results are consistent with either gene duplication or independent horizontal transfer events occurring prior to *Stachybotrys* speciation. Further, the clustering of Tri3 with Sat16 on the one hand, and Tri18 with Sat12 on the other, is consistent with my hypothesis that the satratoxin-specific cluster SC2 originated as a duplication of the CTC (Section 2.3.5).

2.3.7. The hard problem: why are the chemotype-specific gene clusters mutually exclusive?

Although I have shown that the presence of certain gene clusters may suffice to produce the strain-specific products observed in *Stachybotrys,* I have not addressed the mechanism or selection pressures by which these clusters have come to be mutually exclusive, and I am not aware of a way to address these issues with computational methods. I do not think that chemotype mutual exclusivity in *Stachybotrys* is well-explained either by chance or by geographic isolation, because the chemotypes of ~200 *Stachybotrys* strains are known (Jarvis 2003), and there is no relationship between chemotype and geographic location. (For example, three of the strains reported here were isolated from the San Francisco Bay Area; two of these, 40285 and 40293, were taken from the same apartment unit [Cruse et al, 2002].) Another hypothesis contradicted by

my results is that both chemotypes in fact have all the machinery needed to produce both atranones and satratoxins, but there is a strain-specific metabolic shunt at work that minimizes production of one type of toxin or the other. It is possible that by unknown mechanisms, presence of the atranone cluster increases a strain's susceptibility to satratoxin toxicity, and vice versa; one way to test this would be to transfect the CAC into a satratoxin strain and observe colony growth, but as noted in the atranone section above, this experiment is not currently feasible. A more interesting, though even more speculative, hypothesis is that there is some novel regulatory mechanism at work that prevents inclusion of both sets of clusters in a single strain.

2.4. CONCLUSIONS AND RECOMMENDATIONS

I summarize my findings with a unified genetic model for atranone and satratoxin biosynthesis in *Stachybotrys* (Figure 2-11) that also incorporates much previous work by biochemists (Cardoza et al, 2011; McCormick et al, 2011; Hinkley et al, 2000; Degenkolb et al, 2008). The main novel feature of this model is that atranones are made by enzyme products of a single gene cluster (viz, the core atranone cluster) found only in atranone strains, while satratoxins are made by enzyme products of up to three gene clusters (viz, satratoxin-specific clusters 1, 2, and 3) found only in satratoxin strains.

Some aspects of this model are speculative, most notably the precise location of the boundary between trichothecenes produced by atranone strains and those produced by satratoxin strains. Although atranone strains are known to make trichodermol, it is unknown whether they can make early macrocyclic trichothecene intermediates such as trichoverrols and trichoverrins. Due to the presence of the chemotype-independent PKS gene *TRI17* within the CTC, I speculate that atranone strains can produce trichoverrols, though perhaps not trichoverrins. Assay of this chemotype in atranone-producing strains is possible by NMR spectroscopy (Jarvis et al, 1986), and this will be critical to more precisely determine the functions of the putative satratoxin-specific enzymes that I have identified in this study.

One toxicological application of this work would be development of a sensitive PCR-based assay to easily distinguish the presence of the two different *Stachybotrys* chemotypes in an infested building. Although satratoxins are extremely toxic in certain experimental settings (Jarvis, 2003), still unknown is their real potential for harm in typical human environments, especially relative to the chronic inflammatory effects of atranones or of other products made by

both strains (Pestka et al, 2008; Shi, Smith, and Miller, 2011). A simple assay to distinguish between chemotypes present in an infested building, in conjunction with rigorous and consistent medical examination of the affected occupants, might help to distinguish the practical medical importance of these types of toxins. However, to find a few informative cases would almost certainly require to test a great number of infested buildings, since both chemotypes are often found together (Cruse et al, 2002), and *Stachybotrys* is usually found alongside other toxigenic fungi due to their shared predilection for dark, damp growth conditions (Kuhn and Ghannoum, 2003).

Other recommendations that follow from this project are found throughout Section 2.3, and so I only summarize them here. To verify the present model of *Stachybotrys* toxin production (Figure 2-11), the products of the core atranone cluster should be exogenously expressed in either a satratoxin-producing strain of *Stachybotrys* or in a model organism such as yeast; this will involve several technical challenges (Section 2.3.4). More easily, the functions of some enzymes discovered here, such as the PKS Tri17 encoded in the core trichothecene cluster (Section 2.3.3), can likely be verified by standard yeast feeding experiments (eg, Cardoza et al, 2011). Another straightforward project would be to sequence *de novo* the genome of the macrocyclic trichothecene producer *Myrothecium roridum* (Section 2.3.5); this result may inform the evolutionary origin of these two mutually-exclusive sets of gene clusters. Finally, the huge potential repertoire of PKSs found in all strains of *Stachybotrys* (Section 2.3.2) should be examined further, as it could be of pharmacological importance. A preliminary

comparison of these PKS sequences with the accepted fungal PKS phylogeny (Kroken et al, 2003) suggested that novel clades of PKSs may be expanded in the *Stachybotrys* lineage.

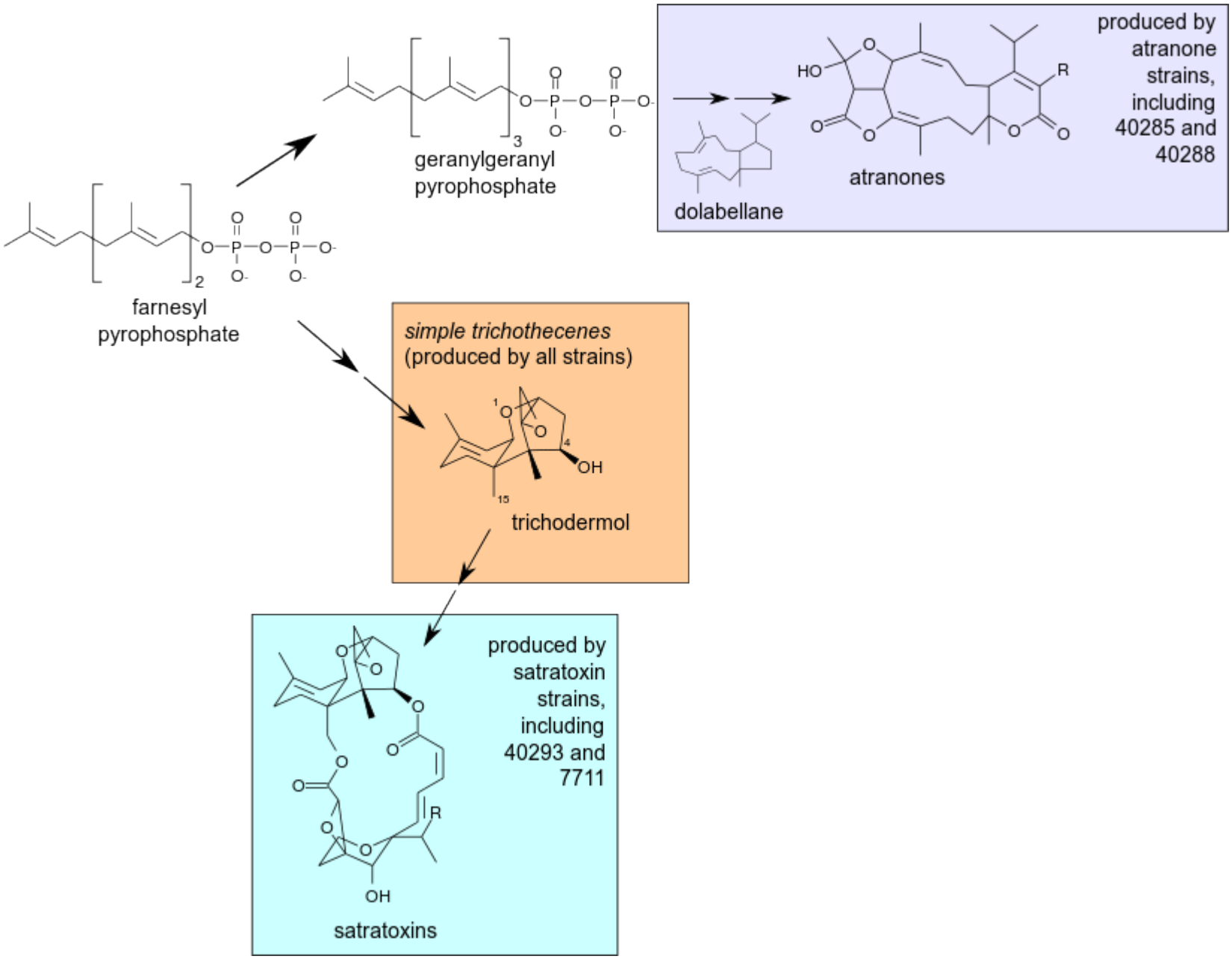
farnesyl
pyrophosphate
geranylgeranyl
pyrophosphate
dolabellane
atranones
produced by
atranone
strains,
including
40285 and
40288
simple trichothecenes
(produced by all strains)
trichodermol
satratoxins
produced by
satratoxin
strains,
including
40293 and
7711

Figure 2-1. The two toxin chemotypes of *Stachybotrys*

Both atranones and satratoxins are terpenoid secondary metabolites thought to derive from the primary metabolite farnesyl pyrophosphate (FPP). Box colors indicate each class of molecule and its specific secondary metabolite precursors: **blue-gray** for atranones, **orange** for simple trichothecenes, and **pale green** for macrocyclic trichothecenes, which include satratoxins. *Atranones* are diterpenoids thought to originate from cyclization of geranylgeranyl pyrophosphate to form dolabellane, which has an unusual eleven-membered ring (Hinkley et al, 2000). Atranones A (R = H) and B (R = O-Me) are shown as representative. *Trichothecenes* are sesquiterpenoids that are products of FPP cyclization. Trichodermol is shown as a representative simple trichothecene, and satratoxins F (R = O) and G (R = OH) as representative satratoxins. The pathway of trichodermol biosynthesis from FPP is known experimentally (Cardoza et al, 2011; Kimura et al, 2007), but there are no experimental data regarding biosynthesis pathways of satratoxins or other trichodermol derivatives, nor of atranones.

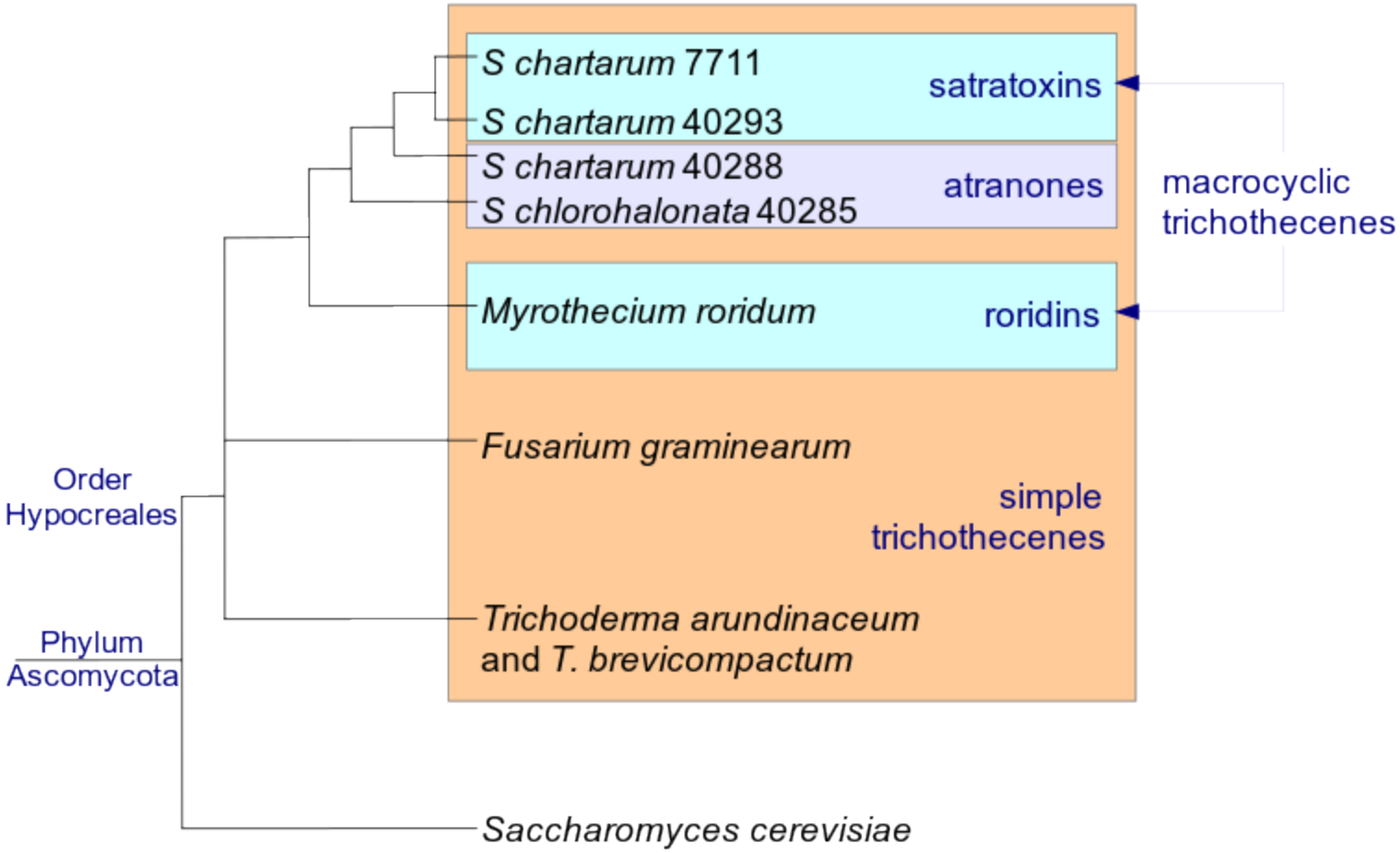

Figure 2-2. *Stachybotrys* strains and other trichothecene producers

This conceptual phylogeny shows the toxin chemotypes of the *Stachybotrys* strains we sequenced in relation to other trichothecene-producing fungi of order *Hypocreales*. *S. cerevisiae* is only distantly related to *Hypocreales* and is shown for context. Topology adapted from Wu et al (2003).

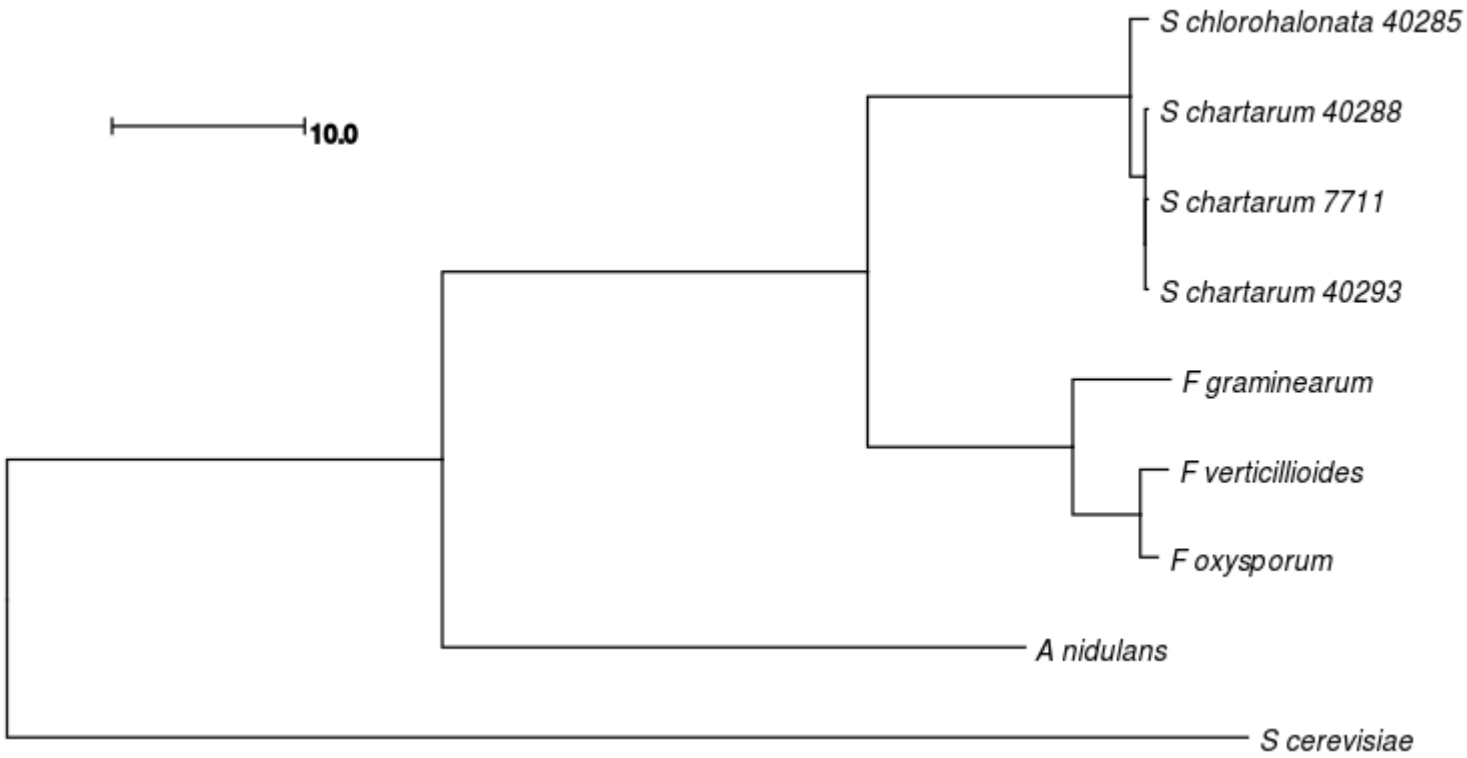

Figure 2-3. Ortholog-based maximum likelihood phylogeny of *Stachybotrys* and other fungi

Phylogeny was constructed from alignment of 2,177 proper protein orthologs identified by OrthoMCL. Scale bar shows number of substitutions per 100 sites. All branches have 100% support.

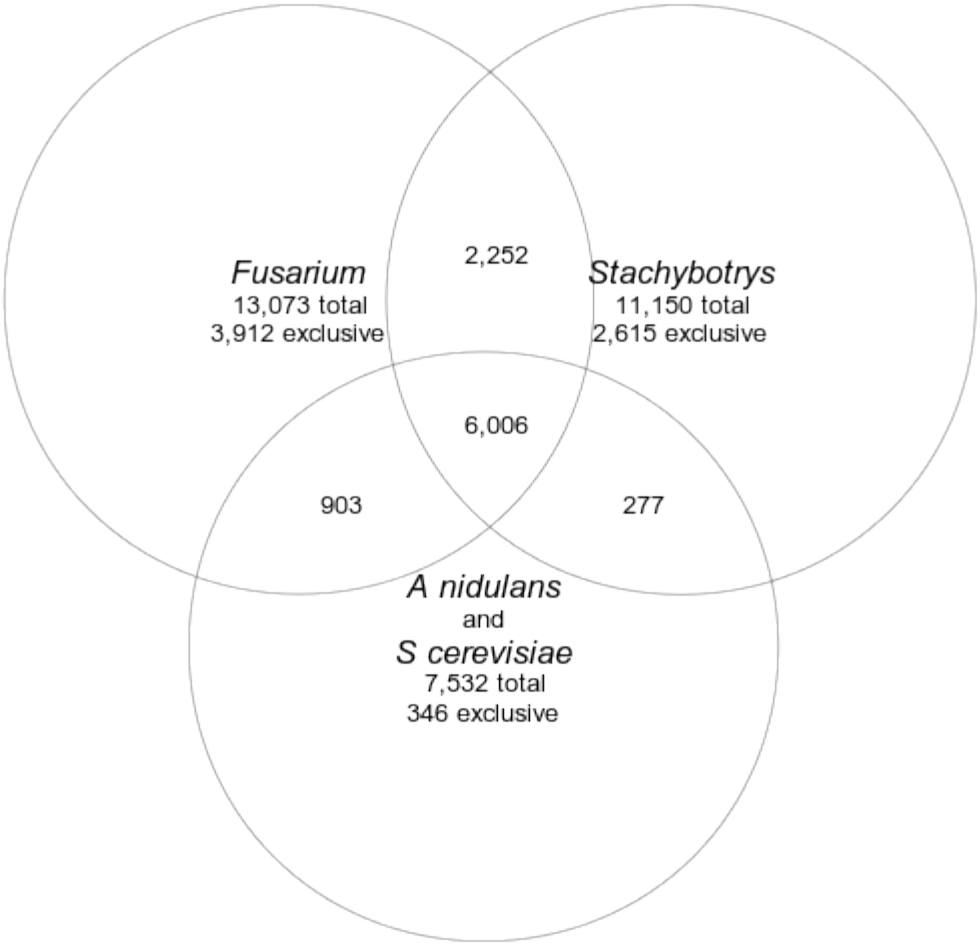

Figure 2-4. Distribution of orthologs of *Fusarium* and *Stachybotrys*

This Venn diagram shows the number of protein homolog groups, of 16,311 total, in each combination of three sets: (1) groups with a homolog in any *Stachybotrys* genome; (2) groups with a homolog in any *Fusarium* genome; and (3) groups with a homolog in *A. nidulans* or *S. cerevisiae,* which are pooled as a single outgroup for simplicity.

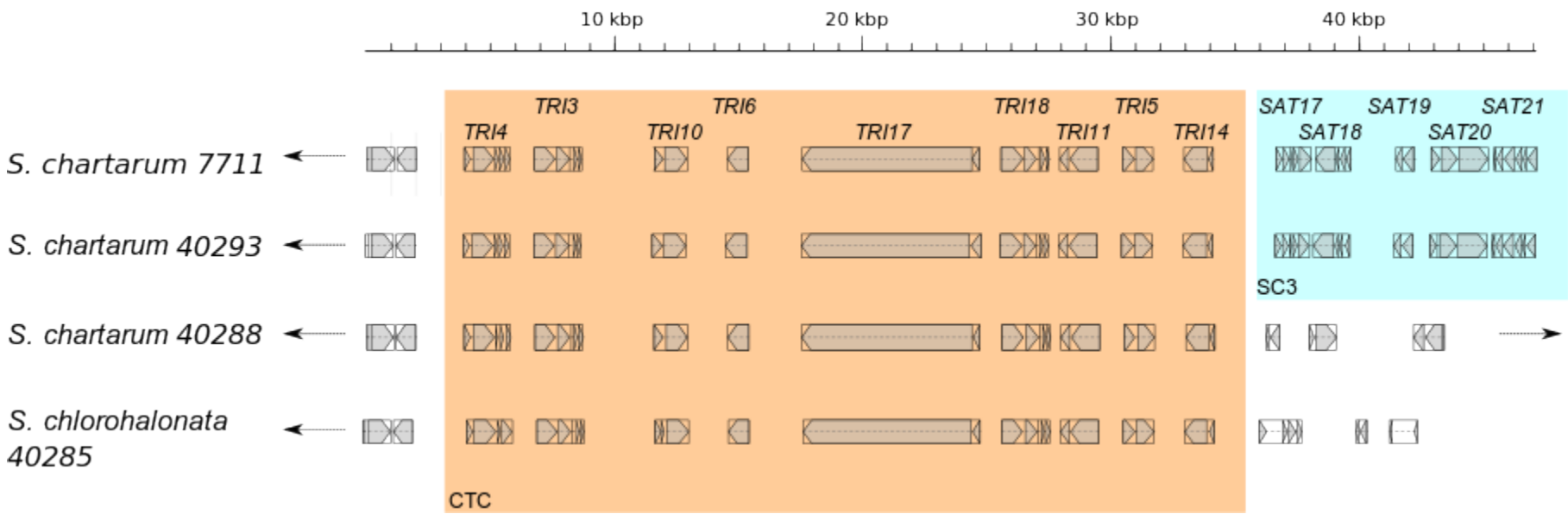

Figure 2-5. The core trichothecene clusters and satratoxin cluster SC3 of each *Stachybotrys* strain

Each rectangle indicates a gene, and each gray arrowhead within indicates an exon and its transcriptional sense. The core trichothecene clusters are shown in the **orange** box, and the adjacent satratoxin cluster SC3 is shown in the **pale green** box. The other genes shown lack similarity to known trichothecene synthesis genes, so they are assumed to be in flanking regions outside these two clusters. A **black arrow** indicates that a scaffold extends to include other genes beyond the region shown, whereas lack of such an arrow indicates a scaffold border. Ruler at top indicates length in kbp.

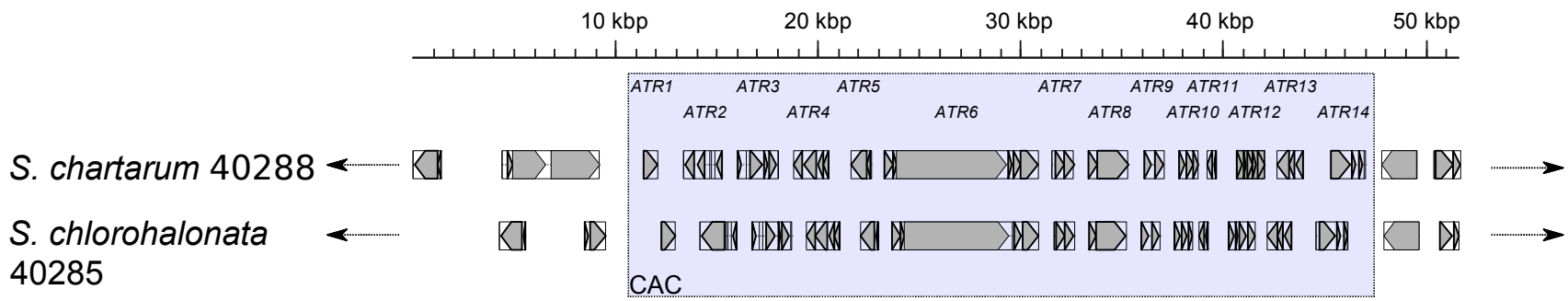

Figure 2-6. The core atranone clusters of the *Stachybotrys* atranone strains

The core atranone clusters are shown in the blue-gray box. The other genes shown are chemotype-independent. Other figure conventions follow those of Figure 2-5. *ATR12* of strain 40288 is shaded to indicate that it is a possible pseudogene; despite its translation having ~90% identity to 40285 Atr12, in our assembly its exon 1 contains an internal stop codon.

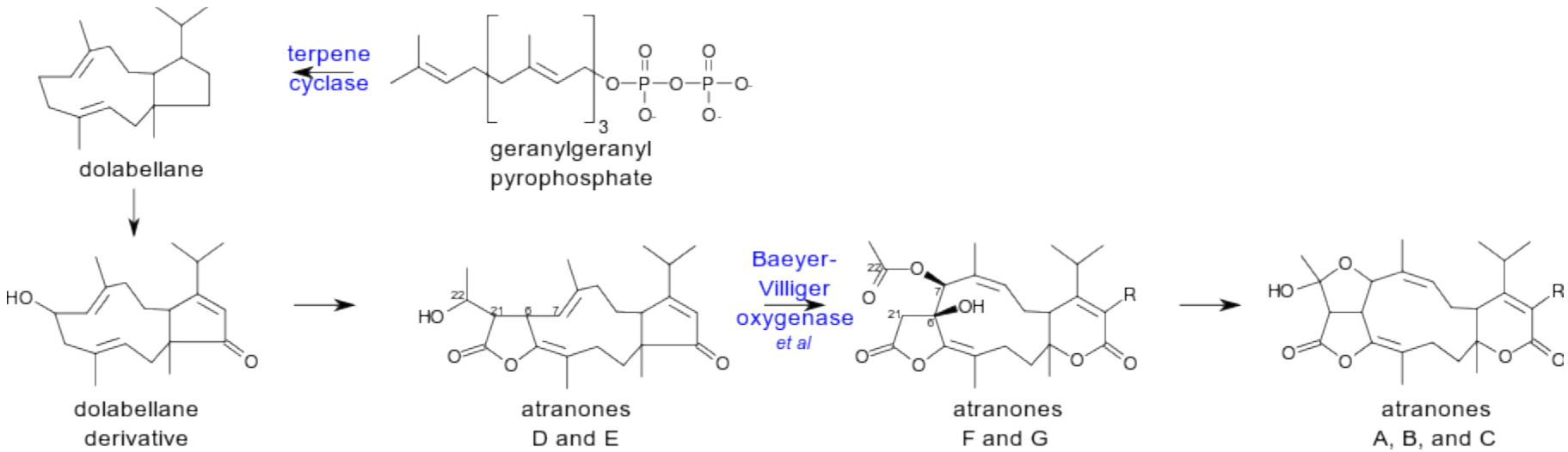

Figure 2-7. Model of atranone biosynthesis

Shown are the structures of all atranones solved by Hinkley et al (2000), as well as types of enzymes capable of catalyzing two postulated reactions in the pathway.

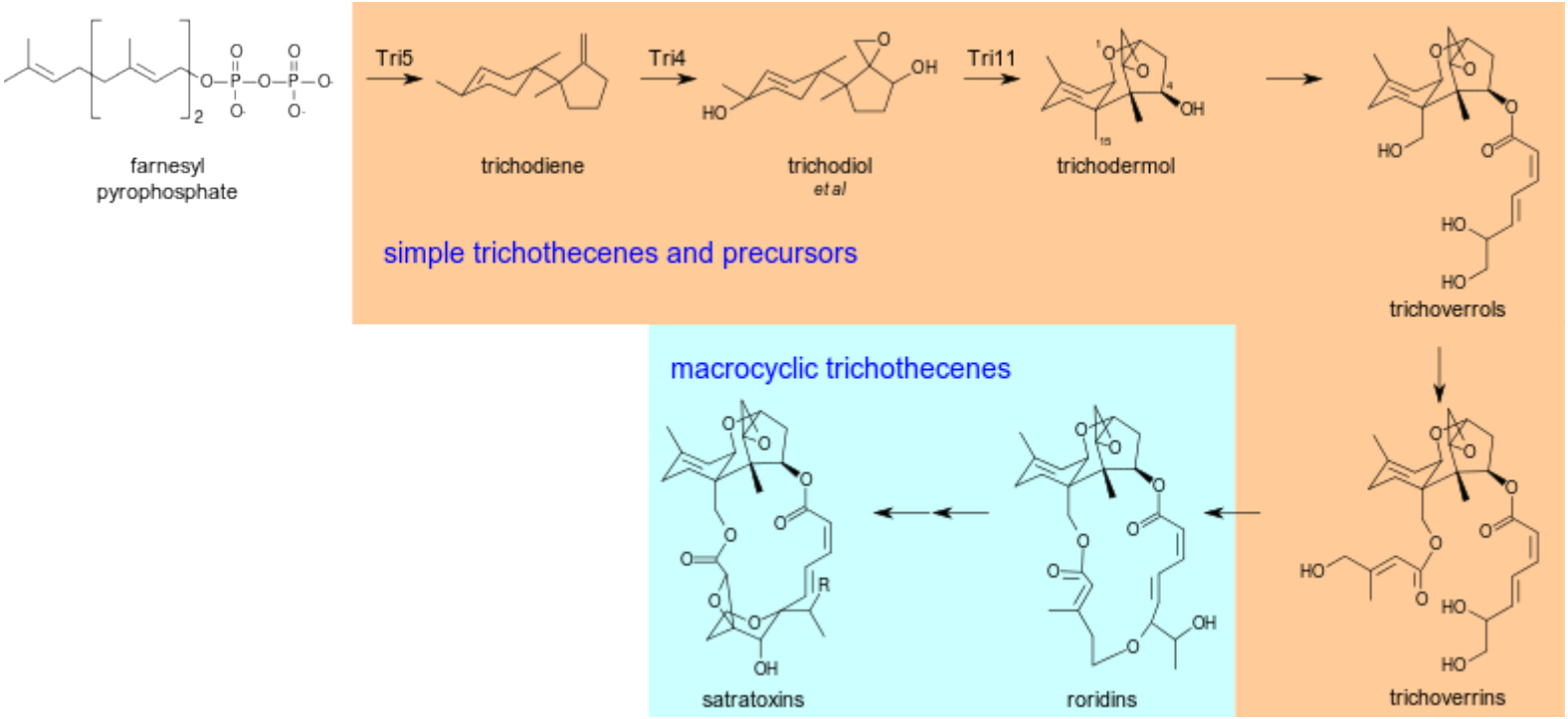

Figure 2-8. Biosynthetic model of satratoxins and other macrocyclic trichothecenes

This is a conceptual pathway adapted from Degenkolb et al (2008) and references therein. It integrates results from several trichothecene producers. Molecule types are color-coded per Figure 2-1. Enzymes shown have been functionally characterized from *Fusarium* (Tri5) or *Trichoderma* (Tri4 and Tri11), but not yet from *Stachybotrys*. Trichodiol is shown to represent several intermediates that undergo both enzymatic hydroxylation and spontaneous rearrangement to form trichodermol, which is the first molecule shown that contains the trichothecene skeleton, ie the tricyclic ring 12,13-epoxytrichothec-9-ene (EPT). In *Fusarium,* trichodermol is not observed; instead, the pathway after trichodiol diverges to a series of products substituted at C-3 of EPT. There are two known trichoverrols (A and B) and two known trichoverrins (A and B), but the respective pairs differ only in the stereochemistry of the C-4 sidechain. The satratoxin F/G skeleton is shown as representative of satratoxins, and roridin E as representative of roridins. Omitted for brevity are the verrucarins (double arrow between roridins and satratoxins).

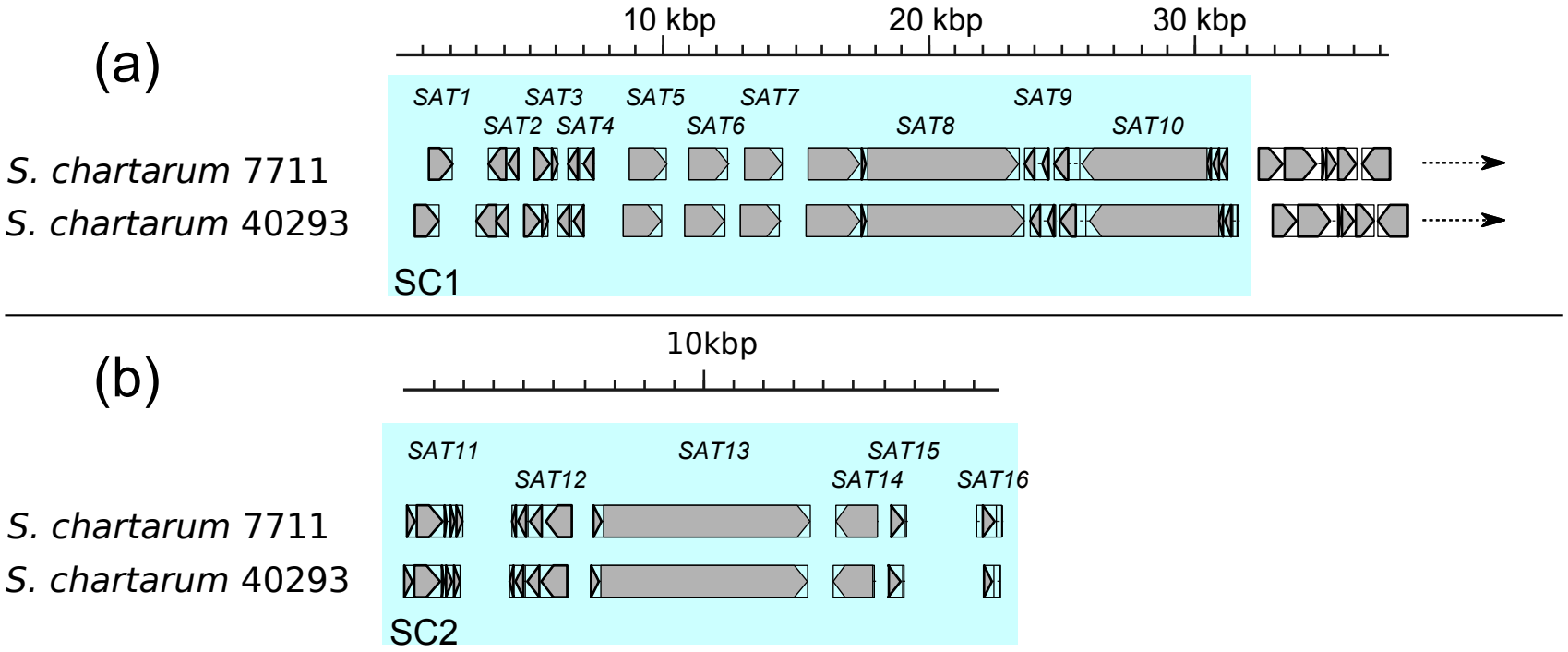

Figure 2-9. Satratoxin-specific clusters SC1 and SC2 of *Stachybotrys*

The satraxin-specific clusters are shown in the pale green boxes. The other genes shown are chemotype-independent. Other figure conventions follow those of Figure 2-5. **(a)** SC1. **(b)** SC2. Additionally, SC3 is shown in Figure 2-5.

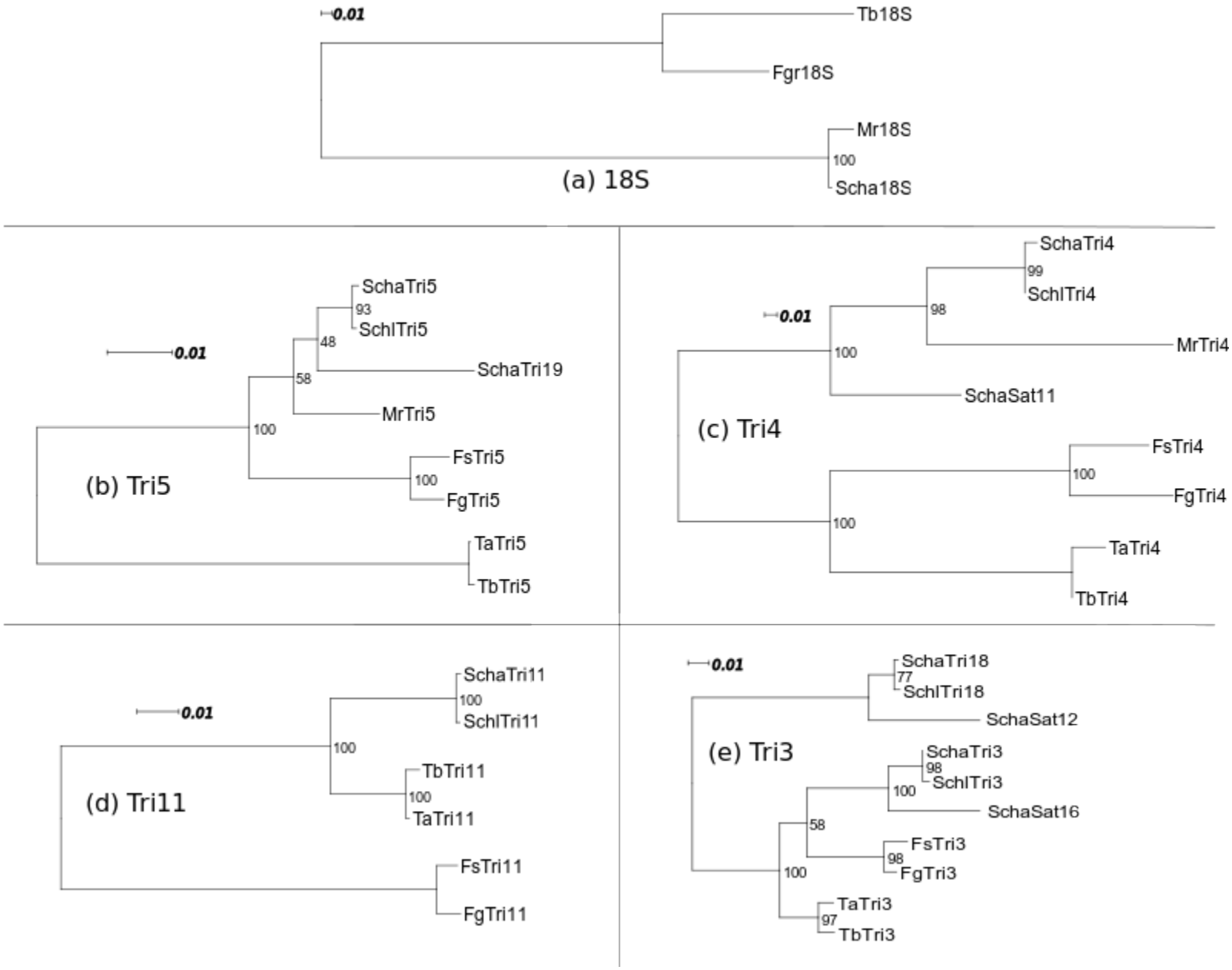
0.01
Tb18S
Fgr18S
Mr18S
100
Scha18S
(a) 18S
0.01
SchaTri5
93
SchlTri5
48
SchaTri19
58
MrTri5
100
FsTri5
100
FgTri5
(b) Tri5
TaTri5
TbTri5
SchaTri4
99
SchlTri4
0.01
98
MrTri4
100
SchaSat11
(c) Tri4
FsTri4
100
FgTri4
100
TaTri4
TbTri4
0.01
SchaTri11
100
SchlTri11
100
TbTri11
100
TaTri11
(d) Tri11
FsTri11
FgTri11
0.01
SchaTri18
77
SchlTri18
SchaSat12
(e) Tri3
SchaTri3
98
SchlTri3
100
SchaSat16
58
FsTri3
98
FgTri3
100
TaTri3
97
TbTri3

Figure 2-10. Maximum likelihood phylogenies of selected Tri homologs

(a) Reference phylogeny made from partial 18S rRNA sequences. **(b)** Tri5, including the paralog Tri19 from strain 40293. **(c)** Tri4, including *Stachybotrys* paralog from SC2. **(d)** Tri11. **(e)** Tri3, including all four *Stachybotrys* paralogs from CTC and SC2. Each phylogeny is rooted at midpoint. Taxon abbreviations are *Scha, S. chartarum* 7711; *Schl, S. chlorohalonata* 40285; *Mr, Myrothecium roridum; Fs*, *Fusarium sporotrichioides; Fg, Fusarium graminearum; Ta, Trichoderma arundinaceum;* and *Tb, Trichoderma brevicompactum.* Branches are labelled with support values of 100 total bootstrap replicates. Scale bars show number of substitutions per site.

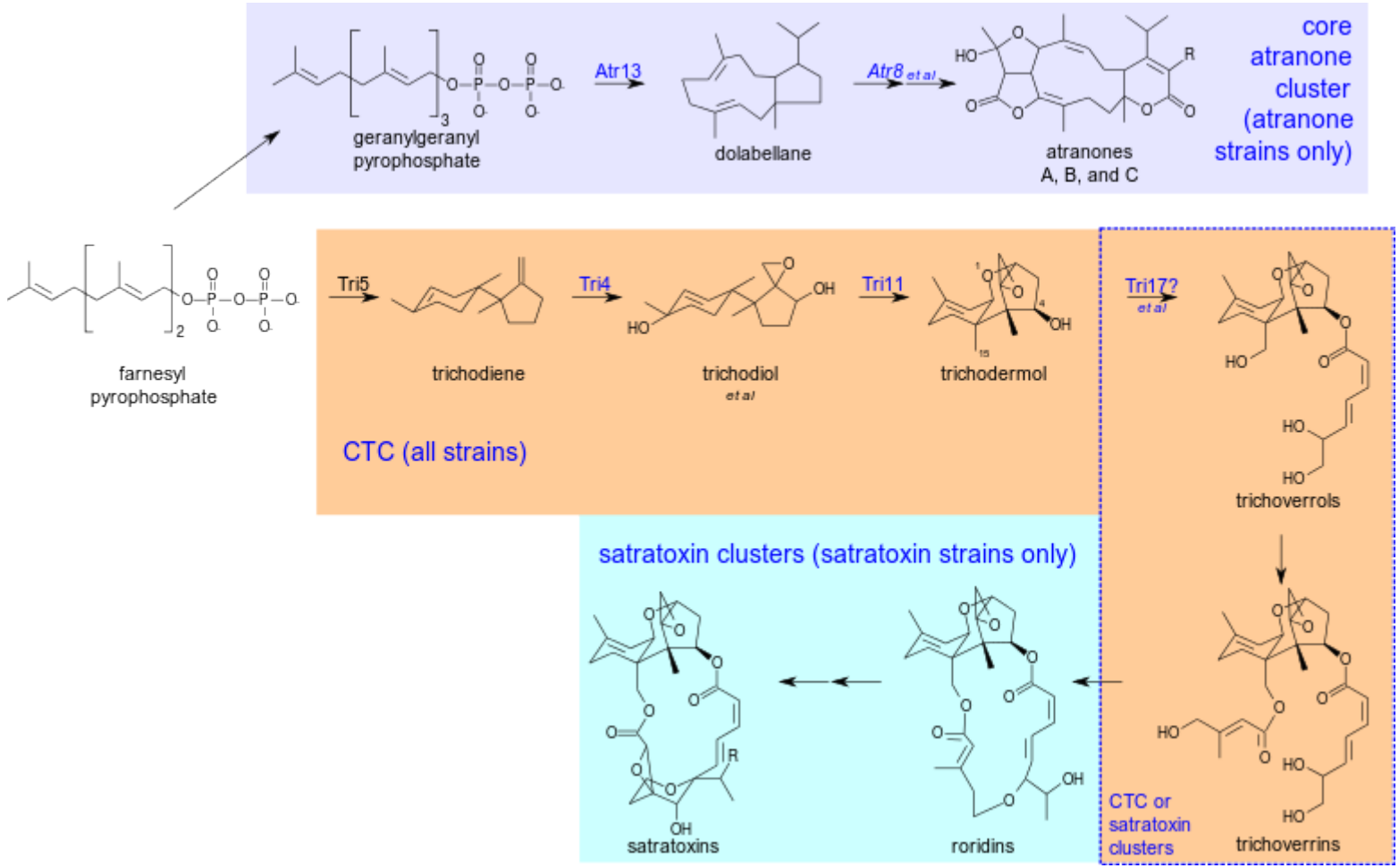

Figure 2-11. Unified genetic model for atranone and satratoxin biosynthesis

Molecules are color-coded per Figure 2-1. **Blue text** indicates that a gene cluster or putative protein was discovered in the present study. Other protein names are in **black**. The **dashed blue box** indicates trichothecenes whose catalysis is uncertain; they may be synthesized by enzyme products of the core trichothecene cluster, by products of satratoxin-specific clusters, or by a mix of both types.

	***S. chlorohalonata* 40285**	***S. chartarum* 40288**	***S. chartarum* 40293**	***S. chartarum* 7711**	***F. graminearum* PH-1**
paired reads (millions)	66.4	58.6	68.8	71.4	NA
assembled sequences	1246	957	826	897	36
assembly size (Mbp)	34.2	36.5	36.1	36.2	36.2
fold coverage	196	162	192	199	10
N_{50} length (kbp)	116	130	214	177	5350
assembly gaps (Mbp)	0.25	0.08	0.16	0.13	0.22
repeat content	1.62%	0.93%	0.93%	1.01%	0.66%
gene content	51.75%	53.42%	53.19%	53.31%	57.18%
predicted coding genes	10866	11719	11532	11543	13332
median gene length (bp)	1357	1377	1380	1379	1259
median protein length (AA)	403	411	412	413	375
mean exons per gene	2.8	2.8	2.8	2.8	2.8
median exon length (bp)	293	296	297	296	255
median intron length (bp)	59	59	59	59	55
predicted products with identified CDD domain	65.87%	65.84%	66.29%	65.94%	61.43%

Table 2-1. Features of *Stachybotrys* genome and proteome assemblies

Stachybotrys assemblies include all contigs and scaffolds of at least 1 kbp. N_{50} is the sequence that includes the middle nucleotide of the assembly when the sequences are ordered by length.

	7711	**40293**	**40288**	**40285**	***Fve***	***Fox***	***Fgr***	***Ani***	***Sce***
7711	100	99.830	99.746	97.701	73.668	73.646	72.995	54.834	39.231
40293		100	99.742	97.707	73.663	73.644	72.998	54.836	39.231
40288			100	97.673	73.663	73.649	73.000	54.836	39.237
40285				100	73.667	73.638	73.011	54.832	39.240
Fve					100	97.174	89.068	55.506	39.796
Fox						100	89.380	55.452	39.742
Fgr							100	54.934	39.373
Ani								100	39.740
Sce									100

Table 2-2. Ortholog-based pairwise proteome identities of *Stachybotrys* and other fungi

The genome abbreviations shown in the row and column headers are as follows. 7711, 40293, 40288, and 40285 are the respective strains of *Stachybotrys* sequenced in this study; *Fve, Fusarium verticillioides; Fox, Fusarium oxysporum; Fgr, Fusarium graminearum; Ani, Aspergillus nidulans; Sce, Saccharomyces cerevisiae.*

putative function	**SC genes**
acetyltransferase	5
hydroxylase	4
regulatory	3
reductase	2
PKS	2
methyltransferase	1
transporter	1
other or unclassified	3
TOTAL	21

Table 2-3. Summary of functions putatively encoded by genes in satratoxin clusters SC1, SC2, and SC3

APPENDIX A (CHAPTER 1)
The 129 protein domains that contain at least two longevity-selected positions

protein	*domain*	*OrthoMaM start*	*OrthoMaM end*	*OrthoMaM length*	*longevity-selected positions*
AMHR2	Protein kinase	203	518	316	7
C3orf48	PP2C-like	171	525	355	4
KIF20B	Kinesin-motor	55	405	351	4
PRSS12	Peptidase S1	631	874	244	4
ALOX15	Lipoxygenase	115	662	548	3
BHMT2	Hcy-binding	11	305	295	3
BTD	CN hydrolase	48	354	307	3
EMID1	Collagen-like	181	370	190	3
IL31RA	Fibronectin type-III 3	234	326	93	3
LIPG	PLAT	347	482	136	3
MST1R	Sema	31	522	492	3
MTMR12	Myotubularin phosphatase	205	643	439	3
NLRP14	NACHT	177	499	323	3
NOV	VWFC	108	174	67	3
PARP14	Macro 3	1040	1211	172	3
PLXNA2	Sema	35	508	474	3
SATL1	N-acetyltransferase	259	379	121	3
SEMA4B	Sema	42	518	477	3
SHBG	Laminin G-like 2	224	390	167	3
STRADB	Protein kinase	58	369	312	3
TMEM56	TLC	44	246	203	3
TRIM26	B30.2/SPRY	295	539	245	3
ABCA6	ABC transporter 1	478	713	236	2
AFP	Albumin 1	19	210	192	2
AFP	Albumin 3	403	601	199	2
ANGPTL3	Fibrinogen C-terminal	237	455	219	2

protein	*domain*	*OrthoMaM start*	*OrthoMaM end*	*OrthoMaM length*	*longevity-selected positions*
ARHGEF16	PH	501	620	120	2
ARHGEF6	PH	443	548	106	2
ATM	FAT	1984	2566	583	2
ATR	FAT	1640	2185	546	2
BRCA1	BRCT 2	1756	1855	100	2
C1R	Peptidase S1	463	701	239	2
CD200	Ig-like V-type	28	138	111	2
CD83	Ig-like V-type	20	114	95	2
CD93	C-type lectin	32	174	143	2
CDCP1	CUB	523	541	19	2
CDON	Ig-like C2-type 5	405	516	112	2
CFD	Peptidase S1	26	253	228	2
CFTR	ABC transmembrane type-1 2	859	1155	297	2
CLCA1	VWFA	306	475	170	2
CP	F5/8 type A 1	20	357	338	2
CP	F5/8 type A 3	730	1061	332	2
CP	Plastocyanin-like 1	20	200	181	2
CR2	Sushi 11	719	775	57	2
CTPS2	Glutamine amidotransferase type-1	300	554	255	2
CUBN	CUB 1	474	586	113	2
CUBN	CUB 12	1738	1850	113	2
CUBN	CUB 17	2336	2448	113	2
CUBN	CUB 4	816	928	113	2
CUBN	EGF-like 4; calcium-binding (Potential)	305	348	44	2
DCHS1	Cadherin 13	1333	1436	104	2
EMB	Ig-like V-type 1	33	120	88	2
ERP27	Thioredoxin	39	152	114	2
F11	Apple 3	200	283	84	2

protein	*domain*	*OrthoMaM start*	*OrthoMaM end*	*OrthoMaM length*	*longevity-selected positions*
F8	F5/8 type A 3	1713	2040	328	2
FAT2	Cadherin 14	1556	1660	105	2
FAT2	Cadherin 15	1661	1758	98	2
FRAS1	VWFC 2	67	127	61	2
FREM1	C-type lectin	2061	2175	115	2
GALNT5	Ricin B-type lectin	804	935	132	2
GDPD1	GDPD	45	304	260	2
HEPHL1	Plastocyanin-like 3	376	558	183	2
IFNAR1	Fibronectin type-III 3	333	424	92	2
IL12RB1	Fibronectin type-III 2	143	236	94	2
IL1RL1	Ig-like C2-type 2	114	197	84	2
IL6ST	Ig-like C2-type	26	120	95	2
IQGAP1	Ras-GAP	985	1218	234	2
ISG20L2	Exonuclease	178	353	176	2
JAK1	FERM	34	420	387	2
KIF18B	Kinesin-motor	4	279	276	2
KIF22	Kinesin-motor	40	299	260	2
KIRREL2	Ig-like C2-type 2	123	222	100	2
KIRREL2	Ig-like C2-type 5	398	501	104	2
KIT	Ig-like C2-type 2	121	205	85	2
KLK6	Peptidase S1	22	242	221	2
LAMA2	Laminin IV type A 2	1176	1379	204	2
LAMA3	Laminin EGF-like 5	514	566	53	2
LETMD1	LETM1	144	359	216	2
LGTN	SUI1	491	564	74	2
LIFR	Fibronectin type-III 2	332	428	97	2
LRIT2	Fibronectin type-III	361	451	91	2
LTF	Transferrin-like 2	364	695	332	2
MAPK13	Protein kinase	25	308	284	2
MCF2	DH	572	752	181	2

protein	*domain*	*OrthoMaM start*	*OrthoMaM end*	*OrthoMaM length*	*longevity-selected positions*
MFI2	Transferrin-like 2	432	706	275	2
MTTP	Vitellogenin	28	659	632	2
MYO18A	Myosin head-like	420	1186	767	2
MYO5C	Myosin head-like	2	755	754	2
MYO7B	FERM 2	2037	2340	304	2
MYO7B	MyTH4 1	983	1186	204	2
NIN	EF-hand 4	219	252	34	2
NLRP12	NACHT	211	528	318	2
NTN5	NTR	345	475	131	2
OBSCN	Ig-like 43	1805	1873	69	2
ORC1L	BAH	45	171	127	2
PBK	Protein kinase	32	322	291	2
PDCD11	S1 motif 9	1036	1109	74	2
PDCD1LG2	Ig-like V-type	21	118	98	2
PGBD1	SCAN box	44	126	83	2
PKD1L1	REJ	377	1274	898	2
PLA2G4A	PLA2c	140	650	511	2
PRKDC	FAT	2882	3538	657	2
PRLR	Fibronectin type-III 2	127	227	101	2
PTPRC	Tyrosine-protein phosphatase 1	651	910	260	2
RASSF6	SARAH	281	328	48	2
RECQL4	Helicase C-terminal	683	850	168	2
RIPK1	Protein kinase	17	289	273	2
ROS1	Fibronectin type-III 3	558	668	111	2
ROS1	Fibronectin type-III 7	1558	1653	96	2
RPGRIP1	C2	801	890	90	2
RPS6KA6	Protein kinase 2	426	683	258	2
SHBG	Laminin G-like 1	45	217	173	2
SHROOM1	ASD2	543	825	283	2
SNX25	PXA	1	164	164	2

protein	***domain***	***OrthoMaM start***	***OrthoMaM end***	***OrthoMaM length***	***longevity-selected positions***
SNX25	RGS	287	401	115	2
SPINK5	Kazal-like 6	361	423	63	2
STK31	Protein kinase	534	842	309	2
SVEP1	Pentaxin	1365	1482	118	2
SVEP1	Sushi 16	2259	2316	58	2
SYPL1	MARVEL	10	219	210	2
TECTA	VWFD 2	712	929	218	2
TFPI2	BPTI/Kunitz inhibitor 2	96	149	54	2
TGFBRAP1	CNH	24	297	274	2
TRIM25	B30.2/SPRY	439	630	192	2
TYK2	FERM	26	431	406	2
USH2A	Fibronectin type-III 6	1954	2051	98	2
USH2A	Fibronectin type-III 7	2052	2138	87	2
USP48	DUSP 2	569	691	123	2
ZNF541	ELM2	1072	1164	93	2

APPENDIX B (CHAPTER 1)
Longevity-selected positions of the protein domains shown in Appendix A

protein	*domain*	*reference (human) character*	*reference OrthoMaM position*	*non-reference characters*	b_{mls}	p_{mls}	b_{mass}	p_{mass}
AMHR2	Protein kinase	M	379	27	1.31	3.97e-3	-0.07	4.58e-1
AMHR2	Protein kinase	T	447	27	3.62	2.63e-3	-0.49	4.47e-2
AMHR2	Protein kinase	Y	465	27	8.56	1.24e-3	-1.14	3.22e-2
AMHR2	Protein kinase	T	469	27	2.09	6.02e-3	-0.20	1.97e-1
AMHR2	Protein kinase	F	473	27	6.00	2.30e-3	-0.61	1.18e-1
AMHR2	Protein kinase	E	513	26	3.85	4.61e-3	-0.17	5.28e-1
AMHR2	Protein kinase	H	515	25	7.84	5.98e-3	-0.62	2.68e-1
C3orf48	PP2C-like	P	224	23	5.47	8.44e-3	-0.20	6.36e-1
C3orf48	PP2C-like	E	405	22	4.54	4.08e-3	-1.08	2.62e-3
C3orf48	PP2C-like	E	411	20	3.02	5.74e-3	-0.63	8.57e-3
C3orf48	PP2C-like	E	459	21	5.34	7.61e-3	-1.35	3.74e-3
KIF20B	Kinesin-motor	S	77	25	1.45	6.42e-3	-0.06	5.31e-1
KIF20B	Kinesin-motor	Q	144	25	1.58	5.50e-3	-0.02	8.31e-1
KIF20B	Kinesin-motor	T	235	28	2.68	8.60e-3	-0.54	1.16e-2
KIF20B	Kinesin-motor	A	259	26	2.53	5.48e-3	-0.27	1.42e-1
PRSS12	Peptidase S1	G	641	26	3.39	7.58e-3	-0.63	2.29e-2
PRSS12	Peptidase S1	Q	710	28	1.81	7.57e-3	-0.26	6.02e-2
PRSS12	Peptidase S1	G	835	31	3.53	3.14e-3	-0.14	5.70e-1
PRSS12	Peptidase S1	K	874	30	2.00	6.46e-3	-0.05	7.26e-1
ALOX15	Lipoxygenase	P	120	15	8.73	7.39e-4	-0.14	7.43e-1
ALOX15	Lipoxygenase	R	205	15	8.37	5.03e-4	-0.28	5.00e-1
ALOX15	Lipoxygenase	K	643	14	4.97	1.98e-4	-0.54	1.59e-2
BHMT2	Hcy-binding	M	61	28	3.21	1.66e-3	-0.15	4.50e-1
BHMT2	Hcy-binding	N	69	28	2.18	8.25e-3	-0.40	2.21e-2
BHMT2	Hcy-binding	D	89	28	6.71	1.97e-3	-1.16	1.11e-2
BTD	CN hydrolase	V	96	24	2.89	2.08e-3	-0.45	2.60e-2
BTD	CN hydrolase	I	221	24	5.82	6.26e-3	-0.92	5.46e-2
BTD	CN hydrolase	S	350	25	5.30	2.77e-4	-0.76	1.30e-2

protein	*domain*	*reference (human) character*	*reference OrthoMaM position*	*non-reference characters*	b_{mls}	p_{mls}	b_{mass}	p_{mass}
EMID1	Collagen-like	S	233	15	3.57	6.27e-3	-0.06	8.16e-1
EMID1	Collagen-like	H	268	15	5.18	3.07e-3	-0.78	2.77e-2
EMID1	Collagen-like	R	346	20	4.37	7.81e-3	-0.03	9.33e-1
IL31RA	Fibronectin type-III 3	E	234	27	5.61	2.26e-3	-0.76	3.89e-2
IL31RA	Fibronectin type-III 3	N	297	24	5.52	8.43e-3	-0.38	3.49e-1
IL31RA	Fibronectin type-III 3	L	300	24	6.66	6.96e-3	-1.21	1.69e-2
LIPG	PLAT	Q	421	26	2.77	1.38e-3	-0.23	1.73e-1
LIPG	PLAT	N	445	26	5.92	2.66e-3	-0.28	4.52e-1
LIPG	PLAT	N	469	27	3.52	1.61e-3	-0.21	3.18e-1
MST1R	Sema	A	257	20	4.29	9.01e-3	-0.24	4.60e-1
MST1R	Sema	S	268	20	6.38	2.43e-3	-1.12	1.03e-2
MST1R	Sema	G	393	19	7.25	9.80e-4	-0.64	1.29e-1
MTMR12	Myotubularin phosphatase	R	416	27	7.56	8.35e-4	-1.37	4.26e-3
MTMR12	Myotubularin phosphatase	Q	509	27	6.17	6.66e-4	-0.05	8.95e-1
MTMR12	Myotubularin phosphatase	S	615	28	3.64	1.93e-3	-0.71	5.15e-3
NLRP14	NACHT	Y	212	22	2.05	5.54e-3	-0.08	5.97e-1
NLRP14	NACHT	L	370	23	6.08	3.99e-4	-0.79	1.55e-2
NLRP14	NACHT	A	432	24	2.02	3.28e-3	-0.07	6.04e-1
NOV	VWFC	S	126	29	3.16	8.25e-4	-0.11	5.42e-1
NOV	VWFC	K	128	29	1.94	2.94e-3	-0.06	6.28e-1
NOV	VWFC	E	162	30	4.69	3.84e-3	-0.49	1.39e-1
PARP14	Macro 3	G	1182	25	6.84	5.90e-4	-0.86	2.90e-2
PARP14	Macro 3	D	1201	24	3.77	7.15e-3	-0.17	5.22e-1
PARP14	Macro 3	V	1202	24	1.74	4.83e-3	-0.10	3.91e-1
PLXNA2	Sema	S	93	27	1.40	8.23e-3	-0.04	6.96e-1
PLXNA2	Sema	S	321	26	3.12	4.03e-4	-0.43	1.41e-2
PLXNA2	Sema	D	332	25	2.41	2.39e-3	-0.13	4.09e-1

protein	*domain*	*reference (human) character*	*reference OrthoMaM position*	*non-reference characters*	b_{mls}	p_{mls}	b_{mass}	p_{mass}
SATL1	N-acetyltransferase	E	271	17	3.26	9.85e-4	-0.29	1.70e-1
SATL1	N-acetyltransferase	Q	317	20	5.79	7.16e-3	-0.96	2.32e-2
SATL1	N-acetyltransferase	A	353	21	3.90	3.82e-3	-0.25	3.65e-1
SEMA4B	Sema	R	363	23	4.49	4.07e-3	-0.73	3.38e-2
SEMA4B	Sema	G	475	25	3.88	5.54e-3	-0.08	7.75e-1
SEMA4B	Sema	I	486	25	1.39	3.54e-3	-0.14	1.65e-1
SHBG	Laminin G-like 2	G	255	20	7.22	6.81e-3	-1.01	9.90e-2
SHBG	Laminin G-like 2	H	264	20	8.37	4.91e-4	-1.98	7.23e-4
SHBG	Laminin G-like 2	G	293	23	6.54	8.80e-4	-0.40	2.85e-1
STRADB	Protein kinase	I	291	27	4.66	1.01e-3	-0.64	2.53e-2
STRADB	Protein kinase	S	320	27	1.82	7.85e-3	-0.05	7.28e-1
STRADB	Protein kinase	S	335	27	1.24	4.59e-3	-0.05	5.54e-1
TMEM56	TLC	I	107	24	1.30	3.31e-3	-0.06	5.12e-1
TMEM56	TLC	Y	109	23	5.57	1.87e-3	-0.26	4.46e-1
TMEM56	TLC	A	163	25	4.42	9.77e-3	-0.57	1.48e-1
TRIM26	B30.2/SPRY	H	347	20	9.19	9.47e-3	-1.88	2.54e-2
TRIM26	B30.2/SPRY	D	415	19	2.67	9.40e-3	-0.25	3.10e-1
TRIM26	B30.2/SPRY	L	434	20	5.25	7.24e-4	-0.55	1.08e-1
ABCA6	ABC transporter 1	E	535	20	6.08	7.99e-3	-1.16	1.91e-2
ABCA6	ABC transporter 1	E	590	21	5.53	1.68e-3	-0.77	3.00e-2
AFP	Albumin 1	F	50	28	5.06	3.98e-3	0.09	8.11e-1
AFP	Albumin 1	E	119	28	4.39	6.99e-4	-0.20	4.35e-1
AFP	Albumin 3	Y	404	29	1.86	5.39e-3	-0.09	5.02e-1
AFP	Albumin 3	T	460	28	4.33	9.19e-3	-0.60	8.19e-2
ANGPTL3	Fibrinogen C-terminal	K	377	27	2.28	5.53e-3	-0.11	4.98e-1
ANGPTL3	Fibrinogen C-terminal	H	405	24	5.55	3.21e-3	-0.33	3.74e-1
ARHGEF16	PH	T	586	20	6.13	3.57e-3	-1.23	9.77e-3
ARHGEF16	PH	H	617	20	9.44	2.51e-3	-1.24	6.12e-2
ARHGEF6	PH	M	457	24	1.65	5.64e-3	-0.08	4.86e-1
ARHGEF6	PH	C	463	24	7.97	7.66e-3	-1.31	3.61e-2

protein	*domain*	*reference (human) character*	*reference OrthoMaM position*	*non-reference characters*	b_{mls}	p_{mls}	b_{mass}	p_{mass}
ATM	FAT	C	2074	27	6.23	2.32e-3	-0.08	8.57e-1
ATM	FAT	Q	2197	27	2.50	6.54e-3	-0.04	8.30e-1
ATR	FAT	S	1876	27	5.65	7.50e-3	-0.67	1.36e-1
ATR	FAT	Q	1877	27	2.07	4.98e-3	-0.04	7.81e-1
BRCA1	BRCT 2	L	1800	26	4.85	5.67e-3	-0.87	2.38e-2
BRCA1	BRCT 2	A	1843	26	1.59	4.89e-3	-0.02	8.72e-1
C1R	Peptidase S1	S	553	22	3.59	5.92e-3	-0.20	4.67e-1
C1R	Peptidase S1	R	669	23	5.66	2.50e-3	-0.27	4.78e-1
CD200	Ig-like V-type	P	43	28	5.33	3.49e-3	-0.35	3.64e-1
CD200	Ig-like V-type	G	114	29	5.09	5.12e-3	-0.65	1.02e-1
CD83	Ig-like V-type	S	99	29	3.91	6.64e-3	-0.14	6.17e-1
CD83	Ig-like V-type	P	112	29	11.57	1.38e-3	-1.68	2.02e-2
CD93	C-type lectin	L	105	25	5.61	8.16e-3	-1.05	3.19e-2
CD93	C-type lectin	V	117	25	1.41	3.93e-3	-0.28	1.20e-2
CDCP1	CUB	R	527	25	4.73	2.93e-3	-0.39	2.80e-1
CDCP1	CUB	G	529	27	8.74	4.95e-3	-2.33	2.94e-3
CDON	Ig-like C2-type 5	R	437	22	3.34	7.66e-3	-0.63	1.98e-2
CDON	Ig-like C2-type 5	Q	492	24	5.17	4.06e-3	-0.66	6.80e-2
CFD	Peptidase S1	D	105	13	6.91	2.63e-3	-0.38	4.38e-1
CFD	Peptidase S1	A	177	16	4.53	7.46e-3	-0.38	3.02e-1
CFTR	ABC transmembrane type-1 2	L	884	26	2.86	1.59e-3	-0.50	1.41e-2
CFTR	ABC transmembrane type-1 2	F	931	25	5.03	8.36e-3	-1.35	1.49e-3
CLCA1	VWFA	N	423	27	5.58	7.66e-3	-0.80	7.81e-2
CLCA1	VWFA	G	473	29	2.70	9.45e-3	-0.47	3.19e-2
CP	F5/8 type A 1	I	54	28	4.24	2.72e-3	-0.21	4.37e-1
CP	Plastocyanin-like 1	I	54	28	4.24	2.72e-3	-0.21	4.37e-1
CP	Plastocyanin-like 1	Q	165	28	4.42	6.80e-3	-0.55	1.21e-1
CP	F5/8 type A 1	Q	165	28	4.42	6.80e-3	-0.55	1.21e-1

protein	*domain*	*reference (human) character*	*reference OrthoMaM position*	*non-reference characters*	b_{mls}	p_{mls}	b_{mass}	p_{mass}
CP	F5/8 type A 3	Q	811	29	3.49	2.64e-3	-0.24	2.95e-1
CP	F5/8 type A 3	L	1026	25	2.09	6.16e-3	-0.47	1.24e-2
CR2	Sushi 11	T	751	21	3.72	8.68e-3	-0.90	6.67e-3
CR2	Sushi 11	H	753	21	8.19	7.39e-3	-1.87	8.39e-3
CTPS2	Glutamine amidotransferase type-1	N	443	27	4.91	8.51e-3	-0.04	9.24e-1
CTPS2	Glutamine amidotransferase type-1	K	458	27	2.86	4.35e-3	-0.24	2.40e-1
CUBN	EGF-like 4; calcium-binding (Potential)	P	321	26	11.49	1.24e-4	-1.17	4.33e-2
CUBN	EGF-like 4; calcium-binding (Potential)	Y	340	25	1.24	5.82e-3	-0.02	8.27e-1
CUBN	CUB 1	H	529	24	4.08	3.58e-3	-0.12	6.73e-1
CUBN	CUB 1	E	556	23	4.49	2.79e-3	-0.25	4.16e-1
CUBN	CUB 4	H	847	24	8.57	1.88e-3	-0.72	2.07e-1
CUBN	CUB 4	E	886	24	4.33	4.79e-3	-0.58	6.87e-2
CUBN	CUB 12	R	1793	29	3.35	3.97e-3	0.09	7.18e-1
CUBN	CUB 12	S	1839	28	3.08	1.55e-3	-0.33	1.15e-1
CUBN	CUB 17	L	2355	27	3.65	8.25e-3	-0.66	2.48e-2
CUBN	CUB 17	S	2371	27	6.68	2.13e-3	-0.85	5.48e-2
DCHS1	Cadherin 13	P	1357	23	10.15	5.05e-4	-1.42	1.36e-2
DCHS1	Cadherin 13	G	1361	22	3.42	5.73e-3	-0.28	2.43e-1
EMB	Ig-like V-type 1	E	70	18	6.78	4.17e-3	-1.92	2.22e-3
EMB	Ig-like V-type 1	I	93	18	1.33	5.58e-4	-0.30	1.62e-3
ERP27	Thioredoxin	A	71	29	3.12	9.70e-3	-0.08	7.32e-1
ERP27	Thioredoxin	T	130	25	3.55	7.56e-3	0.01	9.74e-1
F11	Apple 3	F	210	27	2.73	4.40e-3	-0.16	4.41e-1
F11	Apple 3	S	213	27	6.75	6.32e-3	-0.99	6.87e-2
F8	F5/8 type A 3	A	1741	22	5.11	3.32e-3	-0.49	1.76e-1
F8	F5/8 type A 3	D	1956	27	6.37	1.82e-3	-1.10	9.33e-3
FAT2	Cadherin 14	H	1623	25	6.81	5.26e-4	-0.10	8.05e-1

protein	*domain*	*reference (human) character*	*reference OrthoMaM position*	*non-reference characters*	b_{mls}	p_{mls}	b_{mass}	p_{mass}
FAT2	Cadherin 14	W	1640	25	17.95	2.73e-4	-2.38	2.19e-2
FAT2	Cadherin 15	G	1701	26	2.86	2.15e-3	0.01	9.59e-1
FRAS1	VWFC 2	E	81	26	3.06	8.28e-3	-0.58	2.30e-2
FRAS1	VWFC 2	V	89	27	1.50	8.50e-3	-0.36	5.52e-3
FREM1	C-type lectin	A	2080	29	2.83	2.47e-3	-0.10	6.10e-1
FREM1	C-type lectin	Q	2158	28	7.21	8.85e-4	-0.76	9.62e-2
GALNT5	Ricin B-type lectin	I	807	24	0.97	9.60e-3	0.01	8.57e-1
GALNT5	Ricin B-type lectin	Y	929	28	7.63	9.09e-3	-1.87	4.16e-3
GDPD1	GDPD	N	145	25	3.06	9.99e-3	-0.19	4.55e-1
GDPD1	GDPD	M	244	25	2.02	3.27e-3	-0.09	5.02e-1
HEPHL1	Plastocyanin-like 3	R	443	26	3.94	9.01e-3	-0.38	2.25e-1
HEPHL1	Plastocyanin-like 3	L	444	26	5.07	3.66e-4	-0.32	2.52e-1
IFNAR1	Fibronectin type-III 3	K	413	22	4.38	7.21e-3	-0.98	1.47e-2
IFNAR1	Fibronectin type-III 3	S	417	22	3.96	2.32e-3	-1.00	2.18e-3
IL12RB1	Fibronectin type-III 2	A	152	17	3.66	3.93e-3	-0.01	9.58e-1
IL12RB1	Fibronectin type-III 2	T	161	17	2.21	2.42e-3	-0.03	8.27e-1
IL1RL1	Ig-like C2-type 2	Y	132	24	6.79	4.29e-3	-0.40	4.15e-1
IL1RL1	Ig-like C2-type 2	N	186	26	4.14	8.80e-3	-0.13	7.08e-1
IL6ST	Ig-like C2-type	Y	57	29	9.59	9.30e-4	-1.71	5.04e-3
IL6ST	Ig-like C2-type	L	97	28	4.68	5.79e-3	-0.28	4.41e-1
IQGAP1	Ras-GAP	K	1008	31	1.19	3.02e-3	-0.05	5.17e-1
IQGAP1	Ras-GAP	A	1104	28	2.58	3.01e-3	-0.32	9.08e-2
ISG20L2	Exonuclease	Q	235	27	4.32	8.34e-3	-0.39	2.61e-1
ISG20L2	Exonuclease	P	288	25	11.07	8.16e-3	-1.12	2.27e-1
JAK1	FERM	D	145	25	2.14	1.70e-3	-0.05	7.51e-1
JAK1	FERM	N	309	28	6.69	9.58e-3	-0.71	2.10e-1
KIF18B	Kinesin-motor	V	37	25	0.61	9.63e-3	-0.03	5.51e-1
KIF18B	Kinesin-motor	Q	156	25	5.71	8.84e-3	-0.88	5.58e-2
KIF22	Kinesin-motor	P	41	24	6.57	5.33e-3	-0.50	2.88e-1
KIF22	Kinesin-motor	I	76	26	1.05	5.23e-3	-0.07	3.56e-1

protein	*domain*	*reference (human) character*	*reference OrthoMaM position*	*non-reference characters*	b_{mls}	p_{mls}	b_{mass}	p_{mass}
KIRREL2	Ig-like C2-type 2	L	176	22	4.45	8.17e-3	-0.47	1.99e-1
KIRREL2	Ig-like C2-type 2	T	181	22	2.01	8.79e-3	-0.02	8.89e-1
KIRREL2	Ig-like C2-type 5	R	458	20	6.68	4.39e-3	-0.35	4.62e-1
KIRREL2	Ig-like C2-type 5	S	480	21	4.15	4.38e-3	-0.45	1.43e-1
KIT	Ig-like C2-type 2	K	158	26	2.30	9.55e-3	-0.12	5.33e-1
KIT	Ig-like C2-type 2	D	165	26	2.07	5.94e-3	-0.07	6.75e-1
KLK6	Peptidase S1	S	82	19	6.67	4.36e-4	-1.37	2.56e-3
KLK6	Peptidase S1	T	146	19	5.02	8.92e-4	-1.15	2.05e-3
LAMA2	Laminin IV type A 2	E	1263	27	1.97	6.77e-3	-0.02	8.80e-1
LAMA2	Laminin IV type A 2	Q	1364	24	7.00	8.16e-3	-1.26	2.38e-2
LAMA3	Laminin EGF-like 5	S	527	24	5.01	4.13e-3	-0.59	1.15e-1
LAMA3	Laminin EGF-like 5	H	556	21	5.13	3.42e-3	-0.13	7.12e-1
LETMD1	LETM1	N	260	28	3.77	7.70e-4	-0.44	5.28e-2
LETMD1	LETM1	Q	344	29	1.76	4.33e-3	-0.07	5.65e-1
LGTN	SUI1	R	498	27	2.33	8.01e-3	-0.42	2.68e-2
LGTN	SUI1	I	523	26	4.56	1.85e-3	-0.96	4.91e-3
LIFR	Fibronectin type-III 2	S	392	25	3.10	3.91e-3	-0.43	4.75e-2
LIFR	Fibronectin type-III 2	N	402	25	5.40	2.24e-3	-0.59	8.82e-2
LRIT2	Fibronectin type-III	N	363	18	6.30	8.34e-3	-1.27	3.51e-2
LRIT2	Fibronectin type-III	E	407	22	3.39	2.40e-3	-0.76	5.02e-3
LTF	Transferrin-like 2	D	584	25	4.14	6.46e-3	-1.01	3.06e-3
LTF	Transferrin-like 2	K	675	24	2.01	5.48e-3	-0.21	1.60e-1
MAPK13	Protein kinase	A	175	21	3.30	4.30e-3	-0.62	1.26e-2
MCF2	DH	V	586	25	2.06	1.93e-3	-0.31	2.22e-2
MCF2	DH	S	640	25	5.42	1.56e-3	-1.19	3.16e-3
MFI2	Transferrin-like 2	S	612	24	5.94	2.00e-3	-0.46	2.37e-1
MFI2	Transferrin-like 2	A	699	25	5.49	3.81e-3	-0.73	5.61e-2
MTTP	Vitellogenin	P	107	27	8.76	2.21e-3	-1.03	7.09e-2
MTTP	Vitellogenin	H	181	29	5.76	5.41e-3	-0.25	5.45e-1
MYO18A	Myosin head-like	Y	536	30	6.88	2.81e-4	-0.77	3.90e-2

protein	*domain*	*reference (human) character*	*reference OrthoMaM position*	*non-reference characters*	b_{mls}	p_{mls}	b_{mass}	p_{mass}
MYO18A	Myosin head-like	D	1078	26	3.48	4.60e-3	-0.71	1.01e-2
MYO5C	Myosin head-like	E	55	28	6.49	1.73e-3	-0.23	5.77e-1
MYO5C	Myosin head-like	S	56	28	3.29	5.57e-3	-0.58	2.09e-2
MYO7B	MyTH4 1	A	1025	19	3.50	3.81e-3	-0.75	7.18e-3
MYO7B	MyTH4 1	P	1072	20	11.83	4.44e-3	-2.67	5.12e-3
MYO7B	FERM 2	G	2067	18	14.36	6.89e-3	-2.95	2.26e-2
MYO7B	FERM 2	Q	2328	20	8.16	3.44e-3	-0.80	2.11e-1
NIN	EF-hand 4	E	222	28	3.16	2.59e-3	-0.07	7.32e-1
NIN	EF-hand 4	H	229	24	4.97	5.14e-3	-0.14	6.98e-1
NLRP12	NACHT	K	376	18	1.84	9.21e-3	-0.46	9.69e-3
NLRP12	NACHT	N	394	19	6.45	3.66e-3	-1.23	1.75e-2
NTN5	NTR	N	347	25	4.34	4.00e-3	-0.05	8.72e-1
NTN5	NTR	R	355	24	4.71	6.67e-3	-0.15	6.77e-1
OBSCN	Ig-like 43	R	1818	21	6.13	8.02e-3	-0.45	3.53e-1
OBSCN	Ig-like 43	T	1850	21	6.22	8.00e-3	-1.36	1.09e-2
ORC1L	BAH	Y	117	24	6.61	9.23e-3	-1.41	1.23e-2
ORC1L	BAH	A	119	24	3.65	2.09e-3	-0.52	3.61e-2
PBK	Protein kinase	N	45	24	3.39	2.27e-3	-0.18	4.43e-1
PBK	Protein kinase	K	259	27	2.94	7.55e-3	-0.18	4.21e-1
PDCD11	S1 motif 9	M	1038	28	4.91	5.90e-3	-0.45	2.59e-1
PDCD11	S1 motif 9	I	1046	28	1.24	4.38e-3	-0.12	1.99e-1
PDCD1LG2	Ig-like V-type	V	25	25	2.10	1.16e-3	-0.07	6.00e-1
PDCD1LG2	Ig-like V-type	T	39	25	2.56	2.53e-3	-0.07	6.70e-1
PGBD1	SCAN box	R	46	19	12.11	2.52e-3	-2.17	1.82e-2
PGBD1	SCAN box	T	124	20	10.18	4.67e-4	-2.53	6.23e-4
PKD1L1	REJ	Q	1135	14	11.28	8.29e-3	-3.51	4.00e-3
PKD1L1	REJ	I	1149	14	9.22	4.84e-3	-1.66	4.28e-2
PLA2G4A	PLA2c	E	443	29	8.84	1.41e-4	-1.31	7.34e-3
PLA2G4A	PLA2c	T	512	29	3.96	5.79e-3	-0.66	3.15e-2
PRKDC	FAT	T	3267	26	3.89	4.55e-3	-0.54	4.70e-2

protein	*domain*	*reference (human) character*	*reference OrthoMaM position*	*non-reference characters*	b_{mls}	p_{mls}	b_{mass}	p_{mass}
PRKDC	FAT	T	3483	27	4.15	4.49e-3	-0.59	5.99e-2
PRLR	Fibronectin type-III 2	A	185	29	2.13	8.17e-3	-0.39	2.55e-2
PRLR	Fibronectin type-III 2	S	195	28	2.50	6.79e-3	-0.42	3.50e-2
PTPRC	Tyrosine-protein phosphatase 1	F	711	27	6.97	6.13e-4	-0.72	8.79e-2
PTPRC	Tyrosine-protein phosphatase 1	N	741	26	4.20	5.84e-3	-0.49	1.30e-1
RASSF6	SARAH	Q	308	31	1.14	4.00e-3	-0.04	6.01e-1
RASSF6	SARAH	T	312	30	4.75	3.44e-3	-0.65	5.09e-2
RECQL4	Helicase C-terminal	N	708	27	6.43	2.26e-3	-1.13	1.07e-2
RECQL4	Helicase C-terminal	V	767	27	2.46	4.01e-3	-0.21	2.26e-1
RIPK1	Protein kinase	E	107	29	3.01	2.84e-3	-0.38	5.43e-2
RIPK1	Protein kinase	D	180	30	4.04	6.11e-3	-0.77	1.36e-2
ROS1	Fibronectin type-III 3	H	571	25	5.32	9.02e-3	-0.45	3.33e-1
ROS1	Fibronectin type-III 3	G	631	26	6.78	3.52e-3	-0.97	5.02e-2
ROS1	Fibronectin type-III 7	T	1566	25	3.17	5.97e-4	-0.27	1.29e-1
ROS1	Fibronectin type-III 7	G	1616	23	4.49	2.69e-3	-0.24	4.05e-1
RPGRIP1	C2	A	824	26	3.83	6.08e-3	-0.58	6.10e-2
RPGRIP1	C2	N	843	25	4.23	1.04e-3	-0.48	7.43e-2
RPS6KA6	Protein kinase 2	R	518	26	3.62	3.90e-3	-0.26	2.82e-1
RPS6KA6	Protein kinase 2	M	595	25	3.51	8.33e-4	-0.14	4.70e-1
SHBG	Laminin G-like 1	P	159	18	7.07	9.28e-3	-1.31	3.23e-2
SHBG	Laminin G-like 1	P	166	18	9.16	4.65e-3	-1.05	1.25e-1
SHROOM1	ASD2	G	591	25	7.40	4.13e-3	-1.11	3.71e-2
SHROOM1	ASD2	P	635	25	6.33	8.90e-3	-0.42	3.88e-1
SNX25	PXA	K	3	15	8.69	6.08e-3	-2.35	2.88e-3
SNX25	PXA	K	6	15	8.88	6.71e-3	-2.36	3.63e-3
SNX25	RGS	A	293	25	2.73	3.01e-3	-0.09	6.13e-1
SNX25	RGS	H	300	25	6.21	4.19e-4	-0.66	5.16e-2
SPINK5	Kazal-like 6	V	374	19	7.15	4.82e-3	-1.59	3.42e-3

protein	*domain*	*reference (human) character*	*reference OrthoMaM position*	*non-reference characters*	b_{mls}	p_{mls}	b_{mass}	p_{mass}
SPINK5	Kazal-like 6	R	375	19	4.56	5.30e-4	-0.75	4.00e-3
STK31	Protein kinase	W	605	21	9.91	3.33e-3	-0.21	7.60e-1
STK31	Protein kinase	T	811	19	3.48	5.25e-3	-0.17	4.96e-1
SVEP1	Pentaxin	K	1460	25	0.76	3.70e-3	-0.02	6.97e-1
SVEP1	Pentaxin	K	1466	24	2.64	6.72e-3	-0.59	6.33e-3
SVEP1	Sushi 16	P	2262	30	5.47	1.66e-3	-0.21	5.28e-1
SVEP1	Sushi 16	Q	2305	30	3.74	4.66e-3	-0.03	8.96e-1
SYPL1	MARVEL	V	77	26	4.97	2.48e-3	-0.88	1.36e-2
SYPL1	MARVEL	D	81	26	8.10	3.52e-3	-1.81	3.67e-3
TECTA	VWFD 2	F	735	26	2.88	2.66e-3	-0.04	8.26e-1
TECTA	VWFD 2	E	794	24	1.93	4.89e-3	-0.03	8.33e-1
TFPI2	BPTI/Kunitz inhibitor 2	S	101	21	4.40	3.93e-3	-0.55	9.23e-2
TFPI2	BPTI/Kunitz inhibitor 2	R	132	22	6.59	9.78e-3	-1.23	2.35e-2
TGFBRAP1	CNH	K	67	25	1.16	9.37e-3	-0.20	3.06e-2
TGFBRAP1	CNH	I	148	26	0.65	2.62e-3	-0.04	3.60e-1
TRIM25	B30.2/SPRY	A	446	23	4.42	6.80e-3	-0.66	4.95e-2
TRIM25	B30.2/SPRY	P	450	21	7.38	8.83e-3	-0.41	4.47e-1
TYK2	FERM	D	295	20	4.16	2.11e-3	-0.33	1.99e-1
TYK2	FERM	A	375	22	4.98	5.74e-3	-0.43	2.44e-1
USH2A	Fibronectin type-III 6	T	1959	23	2.95	7.93e-3	-0.81	3.84e-3
USH2A	Fibronectin type-III 6	R	1962	23	7.73	6.45e-3	-1.46	3.09e-2
USH2A	Fibronectin type-III 7	K	2079	26	3.22	2.28e-3	-0.43	5.02e-2
USH2A	Fibronectin type-III 7	Y	2137	23	5.71	1.52e-3	-0.96	9.56e-3
USP48	DUSP 2	D	616	28	3.37	9.12e-3	-0.25	3.50e-1
USP48	DUSP 2	K	669	25	1.78	3.38e-3	-0.12	3.01e-1
ZNF541	ELM2	H	1098	21	5.23	4.28e-3	-0.43	2.71e-1
ZNF541	ELM2	T	1121	22	3.14	5.05e-3	-0.29	2.18e-1

APPENDIX C (Chapter 2)
Domains enriched in the *Stachybotrys* proteome

Lists of *Stachybotrys* protein domains that are significantly enriched (corrected $p < 0.001$) in comparison to control sets, by Fisher's exact test. **(D-1)** Domains that are enriched in proteins exclusive to *Stachybotrys,* relative to *Stachybotrys* domains overall. **(D-2)** Domains that are overrepresented in the entire set of *Stachybotrys* proteins, relative to the entire set of *Fusarium* proteins. Positive $\log_2$ odds ratio indicates that domain is overrepresented in test group relative to control; conversely, negative odds ratio indicates underrepresentation. P-values shown were corrected for multiple testing with the Bonferroni method.

(D-1)

CDD domain name	# domains in *Stachybotrys*-exclusive groups	# other domains in *Stachybotrys*-exclusive groups	# domains in *Stachybotrys* proteomes	# other domains in *Stachybotrys* proteomes	log_2 odds ratio	p (corrected)
Mito_carr	0	4038	385	43133	-Inf	8.32E-12
WD40	7	4031	381	43137	-2.35	1.63E-04
RRM	8	4030	413	43105	-2.27	7.49E-05
GAL4	150	3888	919	42599	0.84	4.64E-06
SDR_c	140	3898	483	43035	1.68	3.33E-23
BRLZ	33	4005	108	43410	1.73	2.88E-04
SANT	31	4007	97	43421	1.79	3.09E-04
CypX	158	3880	478	43040	1.87	7.63E-32
ANK	242	3796	717	42801	1.93	1.34E-51
Abhydrolase_5	28	4010	75	43443	2.02	9.07E-05
HET	173	3865	386	43132	2.32	2.39E-49
Inhibitor_I9	22	4016	46	43472	2.37	9.22E-05
Peptidases_S8_PCSK9_ProteinaseK_like	22	4016	46	43472	2.37	9.22E-05
FAD_binding_4	45	3993	94	43424	2.38	5.02E-12
fCBD	66	3972	136	43382	2.41	7.10E-19
Amb_all	27	4011	55	43463	2.41	1.54E-06
Ank_2	32	4006	64	43454	2.44	2.57E-08
GlcD	85	3953	163	43355	2.52	1.29E-26
Acetyltransf_7	17	4021	30	43488	2.62	5.76E-04
CFEM	34	4004	57	43461	2.69	1.25E-10
MPP_Dcr2	24	4014	40	43478	2.70	6.70E-07
Glyco_hydro_61	108	3930	148	43370	3.01	1.87E-44
DUF3632	15	4023	19	43499	3.09	1.37E-04
Glyco_hydro_6	16	4022	20	43498	3.11	4.18E-05
Methyltransf_18	14	4024	17	43501	3.15	2.74E-04
M28	16	4022	16	43502	3.43	4.78E-06

(D-2)

CDD domain name	**# domains in *Stachybotrys* proteomes**	**# other domains in *Stachybotrys* proteomes**	**# domains in *Fusarium* proteomes**	**# other domains in *Fusarium* proteomes**	**log_2 odds ratio**	**p (corrected)**
nitrilases_CHs	0	43518	23	36381	-Inf	8.00E-05
Helitron_like_N	0	43518	23	36381	-Inf	8.00E-05
Dimer_Tnp_hAT	32	43486	81	36323	-1.60	1.27E-04
AA_permease_2	50	43468	103	36301	-1.30	4.69E-04
MFS	1552	41966	1990	34414	-0.64	1.27E-34
PKS	152	43366	46	36358	1.47	5.68E-07
PKS_AT	149	43369	43	36361	1.54	1.75E-07
fCBD	136	43382	35	36369	1.70	4.97E-08
Glyco_hydro_61	148	43370	33	36371	1.91	9.05E-11

APPENDIX D (Chapter 2)
Putative polyketide synthases of *Stachybotrys*

Total counts of putative PKSs found in sequenced fungal genomes. Counts include hybrid NRPS/PKSs. We provided or verified counts for *Stachybotrys*, *Fusarium* spp, and *A. nidulans*; others are reprinted from Table 8 of Kubicek et al (2011).

organism	**PKSs**
Stachybotrys chartarum 40293	37
Stachybotrys chartarum 7711	36
Stachybotrys chartarum 40288	35
Aspergillus oryzae	30
Stachybotrys chlorohalonata 40285	28
Aspergillus nidulans	28
Magnaporthe oryzae	28
Cochliobolus heterostrophus	25
Trichoderma virens	22
Trichoderma atroviridae	19
Botryotinia fuckeliana	19
Fusarium graminearum	15
Aspergillus fumigatus	14
Fusarium oxysporum	13
Fusarium solani	13
Trichoderma reesei	13
Fusarium verticillioides	11
Neurospora crassa	7
Saccharomyces cerevisiae	0

APPENDIX E (Chapter 2)
Summary of selected putative *Stachybotrys* proteins

Each table below represents all proteins located in a single set of orthologous gene clusters among the four *Stachybotrys* strains. Alignment length refers to the length, in amino acids (aa), of each respective alignment of all protein orthologs, as shown in Appendix G. Hypothesized function is based on BLASTP results and conserved domains. See main text for discussion of specific proteins.

F-1. Core trichothecene cluster, 9 products.

symbol	alignment length (aa)	hypothesized function
Tri3	524	acetyltransferase
Tri4	547	hydroxylase
Tri5	383	terpene cyclase
Tri6	277	transcription factor
Tri10	422	transcription factor
Tri11	494	hydroxylase
Tri14	367	
Tri17	2403	reducing polyketide synthase (PKS)
Tri18	541	hydroxylase

F-2. Core atranone cluster, 14 products.

symbol	alignment length (aa)	hypothesized function
Atr1	234	esterase, dehydrogenase, or enol lactonase
Atr2	540	monooxygenase
Atr3	478	monooxygenase
Atr4	524	monooxygenase
Atr5	297	esterase
Atr6	2439	reducing PKS
Atr7	320	short-chain reductase
Atr8	625	BVMO
Atr9	253	short-chain reductase
Atr10	276	short-chain reductase

Atr11	133	polyketide cyclase
Atr12	391	methyltransferase
Atr13	381	terpene cyclase
Atr14	461	short-chain reductase

F-3. Satratoxin cluster 1, 10 products.

symbol	alignment length (aa)	hypothesized function
Sat1	294	hydroxylase
Sat2	354	short-chain dehydrogenase
Sat3	271	ketoacyl reductase
Sat4	268	
Sat5	463	acetyltransferase
Sat6	485	lipase; 40% identity to *Fusarium* Tri8
Sat7	470	hydroxylase
Sat8	2602	nonreducing PKS
Sat9	208	transcription factor
Sat10	1930	

F-4. Satratoxin cluster 2, 6 products.

symbol	alignment length (aa)	hypothesized function
Sat11	526	hydroxylase
Sat12	573	acetyltransferase
Sat13	2383	reducing PKS
Sat14	455	acetyltransferase
Sat15	153	transcription factor
Sat16	184	acetyltransferase

F-5. Satratoxin cluster 3, 5 products.

symbol	alignment length (aa)	hypothesized function
Sat17	393	hydroxylase
Sat18	401	methyltransferase

Sat19	230	acetyltransferase
Sat20	712	transcription factor
Sat21	474	transporter

APPENDIX F (Chapter 2)
Selected draft sequences of putative *Stachybotrys* proteins

Alignments are in a slightly modified version of Phylip interleaved format. Length of each alignment is shown in amino acids (aa). These sequences are the unedited output of MAKER, so some minor errors are present. For example, a gap can indicate either sequence that is truly missing from the genome or simply a region that is missing from the annotation or was incorrectly annotated. Corrected sequence records will be placed in Genbank.

Core trichothecene cluster, 9 products

```
Tri3, 524 aa
S40285     MGSLPELLLP PLTPEIHRWK ITKANPRLAQ RRGVGFEVIV GSEQLNRKGQ
S40288     MGSLPELRLP PLAPEIHRWK ISKTNPRLAQ RRGVGFEVIV GCEQLNRKGQ
S40293     MGSLPELRLP PLAPEIHRWK ISKTNPRLAQ RRGVGFEVIV GCEQLNRKGQ
S7711      MGSLPELRLP PLAPEIHRWK ISKTNPRLAQ RRGVGFEVIV GCEQLNRKGQ

           YDLYLTVTLR MVESSTSTPV SLATLKEKFE LALLVARLEH PECGSSVRWD
           YDLYLIVTLR MVESSTSTPV SLAILKEKFE LALLVARLEH PECGSSVRWD
           YDLYLIVTLR MVESSTSTPV SLAILKEKFE LALLVARLEH PECGSSVRWD
           YDLYLIVTLR MVESSTSTPV SLAILKEKFE LALLVARLEH PECGSSVRWD

           DQPSPIFEYE SPENNEAAIA WAKGIVHALP TSSTAQQVWY DLEQERQKSA
           DQPSPIFEYE SPENNEAALA WAKGIVHALP TSSTAQQVWY DLEQERQKSA
           DRPSPIFEYE SPENNEAALA WAKGIVHALP TSSTAQQVWY DLEQERQKSA
           DQPSPIFEYE SPENNEAALA WAKGIVHALP TSSTAQQVWY DLEQERQKSA

           VSDRKAGKPV EIFLIARTPG KDAQIPQGAS IDVLFHMNHL YWDGIGARIF
           VSDRKAGKPV EIFLVAHTPD KDARLPQGAS IDVLFHMNHL YWDGIGARIF
           VSDRKAGKPV EIFLVAHTPD KDARLPQGAS IDVLFHMNHL YWDGIGARIF
           VSDRKAGKPV EIFLVAHTPD KDARLPQGAS IDVLFHMNHL YWDGIGARIF

           VGYLLRQLNN YIGAAGGQEP PTVHWGSEMS NFHTAQLDAM KVQVQSLGTE
           VGYLLRQLSS YIGAAAGQEP PTVHWGSEMS NFHTAQLDAM KVQVESLGSE
           VGYLLRQLSN YIGAAAGQKP PTVHWGSEMS NFHTAQLDAM KVQVESLGSE
           VGYLLRQLSN YIGAAAGQEP PTVHWGSEMS NFHTAQLDAM KVQVESLGSE

           FEARSHQYVN TLMQSLSCWG MPFKASDEAI PRAHTLTFTP AESTDIIRAV
           FEARSHQYVN TLMQSLSCWG MPFKASDEAI PRAHTLTFTP AESTDIIRAV
           FEARSHQYVN TLMQSLSCWG MPFKASDEAI PRAHTLTFTP AESTDIIRAV
           FEARSHQYVN TLMQSLSCWG MPFKASDEAI PRAHTLTFTP AESTDIIRAV

           KTRLGPHYTI SHLAQAATIV AMLDMYRHTA EILETDSFVA PTAVNARRYL
           KTRLGPQYTI SHLAQAATIV AMLDMYRHTA DILETDSFVA PTAVNARRYL
           KTRLGPQYTI SHLAQAATIV AMLDMYRHTA DILETDSFVA PTAVNARRYL
           KTRLGPQYTI SHLAQAATIV AMLDMYRHTA DILETDSFVA PTAVNARRYL

           RDDLKAGYMA GCVTGAVINV RNLKSLLVSL NDDQDVVVAA LARATKDVKA
           RDDLKADYMA GCVTGAVINV GNLKSLLVSL NDDQDVVVGA LAKATKDVKA
           RDDLKADYMA GCVTGAVINV GNLKSLLVSL NDDQDVVVGA LAKATKDVKA
           RDDLKADYMA GCVTGAVINV GNLKSLLVSL NDDQDVVVGA LAKATKDVKA

           SFDLWIHDQS QLALGLRVHS FEGAMLSKNP MPFDKVSGPF ISSDGINELY
           SFDLWIHDQS QLALGLRVHS FEGAMLSKNP MPFDKTSGPF ISSDGINELY
           SFDLWIHDQS QLALGLRVHS FEGAMLSKNP MPFDKTSGPF ISSDGINELY
           SFDLWIHDQS QLALGLRVHS FEGAMLSKNP MPFDKTSGPF ISSDGINELY

           IPTDISSATT GETFMKTDKF VFLLNQFLPY MALRLDSWKG TSMLTICYND
           IPTDISSATT GETFMKTDKF VFLLNQFLPY MALRLDSWKG TSMLTICYND
           IPTDISSATT GETFMKTDKF VFLLNQFLPY MALRLDSWKG TSMLTICYND
           IPTDISSATT GETFMKTDKF VFLLNQFLPY MALRLDSWKG TSMLTICYND

           GNFSQEETAT YLRAVADFML AFRL
           GNFSQEETAT YLRAVADFML AFRL
           GNFSQEETAT YLRAVADFML AFRL
           GNFSQEETAT YLRAVADFML AFRL
```

```
Tri4, 547 aa
S40285     MPALSDLESI KAVPLWAAAG AVGGLYFVYI LGTCFYNVYL HPLRHIPGSK
S40288     MPALSDLESI KAVPLWAAAG AVGGLYFVYI LGTCFYNVYL HPLRHIPGSK
S40293     MPALSDLESI KAVPLWAAAG AVGGLYFVYI LGTCFYNVYL HPLRHIPGSK
S7711      MPALSDLESI KAVPLWAAAG AVGGLYFVYI LGTCFYNVYL HPLRHIPGSK

           LAVMGPYLEF YHEVIRKGQY LWEIEKMHEK YGPIVRVNPR EIHIKDSSFY
           LAVMGPYLEF YHEVIRKGQY LWEIEKMHEK YGPIVRVNPR EIHIKDSSFY
           LAVMGPYLEF YHEVIRKGQY LWEIEKMHEK YGPIVRVNPR EIHIKDSSFY
           LAVMGPYLEF YHEVIRKGQY LWEIEKMHEK YGPIVRVNPR EIHIKDSSFY

           HTIYAAGSRK TNKDPSAVGA FDVPSSTAAT IDHDQHRARR GYLNPYFSKR
           HTIYAAGSRK TNKDPSAVGA FDVPSSTAAT IDHDQHRARR GYLNPYFSKR
           HTIYAAGSRK TNKDPSAVGA FDVPSSTAAT IDHDQHRARR GYLNPYFSKR
           HTIYAAGSRK TNKDPSAVGA FDVPSSTAAT IDHDQHRARR GYLNPYFSKR

           SLADLEPTIH ERISKLTSRT EKHMTDGDVL TLDGIFSALT ADIICARFYG
           SLADLEPTIH ERISKLTSRT EKHMIDGDVL TLDGIFSALT ADIICARFYG
           SLADLEPTIH ERISKLTSRT EKHMIDGDVL TLDGIFSALT ADIICARFYG
           SLADLEPTIH ERISKLTSRT EKHMIDGDVL TLDGIFSALT ADIICARFYG

           EHFDYLGVPD YHFVVRDGFQ GLTKLYHLAR FVPTLVSALK DLPEQVIRMF
           EHFDYLGVPD YHFVVRDGFQ GLTKLYHLAR FVPTLVSALK DLPEQVIRMF
           EHFDYLGVPD YHFVVRDGFQ GLTKLYHLAR FVPTLVSALK DLPEQVIRMF
           EHFDYLGVPD YHFVVRDGFQ GLTKLYHLAR FVPTLVSALK DLPEQVIRMF

           LPALADLVVM RNEIHANGAS KFTSSQTADA KASALVGALA DKNIPPHERT
           LPALADLVVM RNEIHANGAS KFTSSQTADA KASALVGALA DKNIPPHERT
           LPALADLVVM RNEIHANGAS KFTSSQTADA KASALVGALA DKNIPPHERT
           LPALADLVVM RNEIHANGAS KFTSSQTADA KASALVGALA DKNIPPHERT

           VSRLLDEGTV FLFAGTETTS RTMAITMYYL LTNPECLKKL REELETLPVT
           VSRLLDEGTV FLFAGTETTS RTMAITMYYL LTNPGCLKKL REELETLPVT
           VSRLLDEGTV FLFAGTETTS RTMAITMYYL LTNPECLKKL REELETLPVT
           VSRLLDEGTV FLFAGTETTS RTMAITMYYL LTNPECLKKL REELETLPVT

           EDYKHSLQTL ESLPYLSGVV HEGLRLAFGP ITRSARVPMN KDLQYQDYNI
           EDYKHSLQTL ESLPYLSGVV HEGLRLAFGP ITRSARVPMN KDLQYQNYNI
           EDYKHSLQTL ESLPYLSGVV HEGLRLAFGP ITRSARVPMN KDLQYQNYNI
           EDYKHSLQTL ESLPYLSGVV HEGLRLAFGP ITRSARVPMN KDLQYQNYNI

           PAGTPLSMST YFVHTDAELY PEPEKFKPER WIKAAEDGVP LKKFLTNFSQ
           PAGTPLSMST YFVHTDAELY PEPEKFKPER WIKAAEEGVP LKKFLTNFSQ
           PAGTPLSMST YFVHTDAELY PEPEKFKPER WIKAAEEGVP LKKFLTNFSQ
           PAGTPLSMST YFVHTDAELY PEPEKFKPER WIKAAEEGVP LKKFLTNFSQ

           GSRQCIGIKY VAPCTPRSLE DVLTDLPLSS MSFAEMYLTI SRVARAFDFE
           GSRQCIGI-- ---------- ---------N MSFAEMYLTI SRVARAFDFE
           GSRQCIGI-- ---------- ---------N MSFAEMYLTI SRVARAFDFE
           GSRQCIGI-- ---------- ---------N MSFAEMYLTI SRVARAFDFE

           LFETTAADLD MTYARIVAYP KEIPGKKEGL GEIRVKVTNR NHPVMVE
           LFETTAADLD MTYARIVAYP KEIPGKKEGL GEIRVKVTNR HHPVLVQ
           LFETTAADLD MTYARIVAYP KEIPGKKEGL GEIRVKVTNR HHPVLVQ
           LFETTAADLD MTYARIVAYP KEIPGKKEGL GEIRVKVTNR HHPVLVQ

Tri5, 383 aa
```

```
S40285     METFPTEYFL GTAVRLLENV KYRDSNYTRE ERVENLQYAY NKAAAHFAQE
S40288     MEAFPTEYFL GTAVRLLENV KYRDSNYTRE ERVENLQYAY NKAAAHFAQE
S40293     MEAFPTEYFL GTAVRLLENV KYRDSNYTRE ERVENLQYAY NKAAAHFAQE
S7711      MEAFPTEYFL GTAVRLLENV KYRDSNYTRE ERVENLQYAY NKAAAHFAQE

           RQQQILKVSP KRLEASLRTI VGMVVYSWAK VSKELMADLS IHYTYTLILD
           RQQQILKVSP KRLEASLRTI VGMVVYSWAK VSKELMADLS IHYTYTLILD
           RQQQILKVSP KRLEASLRTI VGMVVYSWAK VSKELMADLS IHYTYTLILD
           RQQQILKVSP KRLEASLRTI VGMVVYSWAK VSKELMADLS IHYTYTLILD

           DSEDDPHPQM LTYFDDLQSG NQQKHPWWML VNEHFPNVLR HFGPFCSLNL
           DSEDDPHPQM LTYFDDLQSG NPQKHPWWML VNEHFPNVLR HFGPFCSLNL
           DSEDDPHPQM LTYFDDLQSG NPQKHPWWML VNEHFPNVLR HFGPFCSLNL
           DSEDDPHPQM LTYFDDLQSG NPQKHPWWML VNEHFPNVLR HFGPFCSLNL

           IRSTLDFFEG CWIEQYNFHG FPGSFDYPGF LRRMNGLGHC VGGSLWPKEN
           IRSTLDFFEG CWIEQYNFHG FPGSFDYPGF LRRMNGLGHC VGGSLWPKEN
           IRSTLDFFEG CWIEQYNFHG FPGSFDYPGF LRRMNGLGHC VGGSLWPKEN
           IRSTLDFFEG CWIEQYNFHG FPGSFDYPGF LRRMNGLGHC VGGSLWPKEN

           FNEQEHFLEI TSAIAQMENW MVWVNDLMSF YKEFDDPRDQ TSLVKNYVVS
           FNEQEHFLEI TSAIAQMENW MVWVNDLMSF YKEFDDPRDQ TSLVKNYVVS
           FNEQEHFLEI TSAIAQMENW MVWVNDLMSF YKEFDDPRDQ TSLVKNYVVS
           FNEQEHFLEI TSAIAQMENW MVWVNDLMSF YKEFDDPRDQ TSLVKNYVVS

           EGITLNQALE KLTQDTLQSS EQMMVVFSQK DPKIMDTIEC FMHGYITWHL
           EGITLNQALE KLTQDTLQSS EQMMVVFSQK DPKIMDTIEC FMHGYITWHL
           EGITLNQALE KLTQDTLQSS EQMMVVFSQK DPKIMDTIEC FMHGYITWHL
           EGITLNQALE KLTQDTLQSS EQMMVVFSQK DPKIMDTIEC FMHGYITWHL

           CDNRYRLKEI YDRTKDIQTE DAMRFRKFYE QAFKVGAIEA TEWAYPTVVE
           CDNRYRLKEI YDRTKDIQTE DAMKFRKFYE QAFKVGAIEA TEWAYPTVVE
           CDNRYRLKEI YDRTKDIQTE DAMKFRKFYE QAFKVGAIEA TEWAYPTVVE
           CDNRYRLKEI YDRTKDIQTE DAMKFRKFYE QAFKVGAIEA TEWAYPTVVE

           RLEQRMAEEQ AERDEQAALA NPEKAQVAQV VLA
           RLEQRKAEEQ AERDEQAALA NPEKAQVAQV VLA
           RLEQRKAEEQ AERDEQAALA NPEKAQVAQV VLA
           RLEQRKAEEQ AEREEQAALA NPEKAQVAQV VLA

Tri6, 277 aa
S40285     MIAKDPPSAR DDGSDAWMAL PLFNRKGSPD HTRNILAWQP PTSIDLDMPN
S40288     MIAKDPSSAR DDGSDAWMAL PLFNRKGSPD HTRNILAWQP PTSIDLDMPN
S40293     MIAKDPSSAR DDRSDAWMAL PLFNRKGSPD HTRNILAWQP PTSIDLDMPN
S7711      MIAKDPSSAR DDRSDAWMAL PLFNRKGSPD HTRNILAWQP PTSIDLDMPN

           LDYGFFDYDI MDYDTLTPSY ASESSRPQSD FAISAFHLNA SSYFADSSRP
           LDYCFFDYDI MDYDTLTPSH TSESSRPQSD FAVSAFHFNA SSYFAHSSRP
           LDYCFFDYDI MDYDTLTPSH TSESSRPQSD FAVSAFHFNA SSYFAHSSRP
           LDYCFFDYDI MDYDTLTPSH TSESSRPQSD FAVSAFHFNA SSYFAHSSRP

           HSASFTPALP YDTPSQSPGI SRPHSVSVSS TVQSDSGYES IIVEQASSSK
           HSASLTPVLP RDTPSQSPDI SRPHSVSVSS TVQSDSGYES IVVEQASSSK
           HSASLTPVLP RGTPSQSPDI SRPHSVSVSS TVQSDSGYES IVVEQASSSK
           HSASLTPVLP RDTPSQSPDI SRPHSVSVSS TVQSDSGYES IVVEQASSSK

           RRSLLNEISD DHIPQPAFTG PCQCTFENCK STTVFTTGRD FRRHYRQHFK
           RRSLLNEISD DHIPQPAFTG PCQCTFENCK STTVFTTGRD FRRHYRQHFK
           RRSLLNEISD DHIPQPAFTG PCQCTFENCK STTVFTTGRD FRRHYRQHFK
```

```
            RRSLLNEISD DHIPQPAFTG PCQCTFENCK STTVFTTGRD FRRHYRQHFK

            RFFCRYEECP QSTDDPGDAG TRGFATRKDR ARHEAKHNPA IKCPWQSRNG
            RFFCRYEECP QSTDDPGDAG TRGFATRKDR ARHEAKHNPA IKCPWQNRNG
            RFFCRYEECP QSTDDPGDAG TRGFATRKDR ARHEAKHNPA IKCPWQNRNG
            RFFCRYEECP QSTDDPGDAG TRGFATRKDR ARHEAKHNPA IKCPWQNRNG

            GTCTRTFSRM DNMRDHYRRI HGKSCKT
            GTCTRTFSRM DNMRDHYRRI HGKSCKT
            GTCTRTFSRM DNMRDHYRRI HGKSCKT
            GTCTRTFSRM DNMRDHYRRI HGKSCKT

Tri10, 422 aa
S40285      MTPITITFPK RTQEKETSLL MHYLDVVFPL QFPVHDRSYE GKREWLLAIL
S40288      MTPITITFPK RTQEKETSLL MHYLDVVFPL QFPVHDRSYE GKREWLLAIL
S40293      MTPITITFPK RTQEKETSLL MHYLDVVFPL QFPVHDRSYE GKREWLLAIL
S7711       MTPITITFPK RTQEKETSLL MHYLDVVFPL QFPVHDRSYE GKREWLLAIL

            TSSRSVYYAT LSLSLLHKES RLDDFES--- ---------- --ESQQLLDQ
            TSSRSVYYAT LSLSLLHKES RLDDFESVWQ SERTRYYILA LQESQQLLDQ
            TSSRSVYYAT LSLSLLHKES RLDDFESVWQ SERTRYYILA LQESQQLLDQ
            TSSRSVYYAT LSLSLLHKES RLDDFESVWQ SERTRYYILA LQESQQLLDQ

            IDTADGVAKL KGNIHALAST VQLISIESSS LSLGDWQVHL LAGKALIPEL
            IDTADGVAKL KGNIHALAST VQLISIESSS LSLGDWQVHL LAGKALIPEL
            IDTADGVAKL KGNIHALAST VQLISIESSS LSLGDWQVHL LAGKALIPEL
            IDTADGVAKL KGNIHALAST VQLISIESSS LSLGDWQVHL LAGKALIPEL

            VHGWTLVPKS GRFTSSVWTV LDAPQFSAAH DEDTLSFEYV GALQFLTNIL
            VHGWTLVPKS GRFTSSVWTV LDAPQFSAAH DEDTLSFEYV GALQFLTNIL
            VHGWTLVPKS GRFTSSVWTV LDAPQFSAAH DEDTLSFEYV GALQFLTNIL
            VHGWTLVPKS GRFTSSVWTV LDAPQFSAAH DEDTLSFEYV GALQFLTNIL

            AMFGIFSCIS IGPAAASFAE YRYLLDQEGM IQMDQIIGCK NWVLLAILEV
            AMFGIFSCIS IGPASASFAE YRYLLDQEGM IQMDQIIGCK NWVLLAILEV
            AMFGIFSCIS IGPASASFAE YRYLLDQEGM IQMDQIIGCK NWVLLAILEV
            AMFGIFSCIS IGPASASFAE YRYLLDQEGM IQMDQIIGCK NWVLLAILEV

            GELDKWKRVE QEHHRLSLKD LGNRATVIEE MIENGLRESS GSALVDLVTS
            GELDKWKRVE QEHHRLSLKD LGNRATVIEE MIETGLRESS GSALVDLVTS
            GELDKWKRVE QEHHRLSLKD LGNRATVIEE MIETGLRESS GSALVDLVTS
            GELDKWKRVE QEHHRLSLKD LGNRATVIEE MIETGLRESS GSALVDLVTS

            IYATSTLTYM HTVVSGLNPN LREVQDSVAA TMVLLKQLPD VRIVKNLVWS
            IYATSTLTYM HTVVSGLNPN LREVQDSVAA TMVLLKQLPD VRIAKNLVWS
            IYATSTLTYM HTVVSGLNPN LREVQDSVAA TMVLLKQLPD VRIAKNLVWS
            IYATSTLTYM HTVVSGLNPN LREVQDSVAA TMVLLKQLPD VRIAKNLVWS

            LVVTGCMASL GQEDFFRGLN AVAGTSVRGL RNCWGLATIW DRTWNMRDTI
            LVVTGCMASL GQEDFFRGLN AVAGTSVRGL RNCWGLATIW DRTWNMRDTI
            LVVTGCMASL GQEDFFRGLN AVAGTSVRGL RNCWGLATIW DRTWNMRDTI
            LVVTGCMASL GQEDFFRGLN AVAGTSVRGL RNCWGLATIW DRTWNMRDTI

            TTTSHTVWED LINGQGPPNL LV
            TTTSHTVWED LINGQGPPNL LV
            TTTSHTVWED LINGQGPPNL LV
            TTTSHTVWED LINGQGPPNL LV
```

```
Tri11, 494 aa
S40285     MTVLLQAAAL AAAVYTLSIP IQSIYNLYFH PLSRIPGPKL WIAFPILGQI
S40288     MTVLLQAAAL AAAAYTLSIP IQSIYNLYFH PLSKIPGPKL WIAFPILGQI
S40293     MTVLLQAAAL AAAAYTLSIP IQSIYNLYFH PLSKIPGPKL WIAFPILGQI
S7711      MTVLLQAAAL AAAAYTLSIP IQSIYNLYFH PLSKIPGPKL WIAFPILGQI

           ARVRGILDSY MCAFHRVYGE AVRYGPDEVS IITEQAWKDI YNHRPNQLER
           ARVRGVLDSY MCAFHRVYGE AVRYGPDEVS IITEQAWKDI YNHRPNQLER
           ARVRGVLDSY MCAFHRVYGE AVRYGPDEVS IITEQAWKDI YNHRPNQLER
           ARVRGVLDSY MCAFHRVYGE AVRYGPDEVS IITEQAWKDI YNHRPNQLER

           NILSTTRRPD IFDAVEVDHD RYRKAMSHAF SPKGLQEQEP IVKGYLDLLI
           NILSTTRRPD IFDAVEVDHD RYRKAMSHAF SPKGLQEQEP IVKGYLDLLI
           NILSTTRRPD IFDAVEVDHD RYRKAMSHAF SPKGLQEQEP IVKGYLDLLI
           NILSTTRRPD IFDAVEVDHD RYRKAMSHAF SPKGLQEQEP IVKGYLDLLI

           ERLNQVAANE GKTDMVQWYN FMLFDTIGDL AFGQSFGGLR DQVLHFSISF
           ERLNQVAANE GKTDMVQWYN FMLFDTIGDL AFGQSFGGLR DQVLHFSISF
           ERLNQVAANE GKTDMVQWYN FMLFDTIGDL AFGQSFGGLR DQVLHFSISF
           ERLNQVAANE GKTDMVQWYN FMLFDTIGDL AFGQSFGGLR DQVLHFSISF

           TFEAFKLLTY MEAGARYPLL LKFLELFTPK SIIEARDRKE EHAEATVKQR
           TFEAFKLLTY MEAGARYPLL LKVLELFTPK SIIEARDRKE EHAEATVKQR
           TFEAFKLLTY MEAGARYPLL LKVLELFTPK SIIEARDRKE EHAEATVKQR
           TFEAFKLLTY MEAGARYPLL LKVLELFTPK SIIEARDRKE EHAEATVKQR

           LENGSMHGRG DFMDAMLRNR GKPGGLNDRE LIANASTLIT AGSETTATIL
           LENGSMHGRG DFMDAMLRNR GKPGGLNDRE LIANASTLIT AGSETTATIL
           LENGSMHGRG DFMDAMLRNR GKPGGLNDRE LIANASTLIT AGSETTATIL
           LENGSMHGRG DFMDAMLRNR GKPGGLNDRE LIANASTLIT AGSETTATIL

           SGMTYWLLRN PDNYKKVVHE VRSAYSSDSE ILMITTTTRL PFMIACFQEA
           SGMTYWLLRN PDNYKKVVHE VRSAYSSDSE ILMITTTTRL PFMIACFQEA
           SGMTYWLLRN PDNYKKVVHE VRSAYSSDSE ILMITTTTRL PFMIACFQEA
           SGMTYWLLRN PDNYKKVVHE VRSAYSSDSE ILMITTTTRL PFMIACFQEA

           FRLYPPVPSC LQRVTPETDM TQISGYDIPP NTKVGVHALA AYTDPRNWHR
           FRLYPPVPSC LQRVTPETGM TQISGYDIPP NTKVGVHALA AYTNPRNWHR
           FRLYPPVPSC LQRVTPETGM TQISGYDIPP NTKVGVHALA AYTDPRNWHR
           FRLYPPVPSC LQRVTPETGM TQISGYDIPP NTKVGVHALA AYTDPRNWHR

           PDEFLPERWL SEAEKNPASP FYKDRRATLQ PFSVGPRSCI GRNMAEQEMR
           PDEFLPERWL PEVEKNPASP FYKDRRATLQ PFSVGPRSCI GRNMAEQEMR
           PDEFLPERWL PEVEKNPASP FYKDRRATLQ PFSVGPRSCI GRNMAEQEMR
           PDEFLPERWL PEVEKNPASP FYKDRRATLQ PFSVGPRSCI GRNMAEQEMR

           LILARLLWNF DLALCPESKN WKEQKTHYLW EKHPLMCNVR RRVF
           LILARLLWNF DLALCPESKD WKEQKTHYLW EKHPLMCSVR RRVF
           LILARLLWNF DLALCPESKD WKEQKTHYLW EKHPLMCSVR RRVF
           LILARLLWNF DLALCPESKD WKEQKTHYLW EKHPLMCSVR RRVF

Tri14, 367 aa
S40285     MLPQVLLSAL GLLGEANCSQ TSGHGSPNHH QCPQLPKGDL IYDVYMGYPE
S40288     MLPQVLLSTL GFLGKASCSQ TSGHGSPNYH QCPELPKGDL IYDVYMGYPE
S40293     MLPQVLLSTL GFLGKASCSQ TSGHGSPNYH QCPELPKGDL IYDVYMGYPE
S7711      MLPQVLLSTL GFLGKASCSQ TSGHGSPNYH QCPELPKGDL IYDVYMGYPE

           NFMWDKRRCV AYVSNLYNAS LSIYDPYESR ILETFQFPGL SHPGNSAIDN
           NFMWDKRRCV AYVSNLYNAS LSIYDPYESR VLETFQFPGL SHPGNSAIDN
```

```
          NFMWDKRRCV AYVSNLYNAS LSIYDPYESR VLETFQFPGL SHPGNSAIDN
          NFMWDKRRCV AYVSNLYNAS LSIYDPYESR VLETFQFPGL SHPGNSAIDN

          PLHTSGLVLR PDSYRAETLE IVVDNGDAFF SNGLNVSGPD YLLTMNLNTK
          PLHTSGLVLR PDSYRAETLE IVVDNGDAFF SNGLNVSGPD YLLTMDLSTK
          PLHTSGLVLR PDSYRAETLE IVVDNGDAFF SNGLNVSGPD YLLTMDLSTK
          PLHTSGLVLR PDSYRAETLE IVVDNGDAFF SNGLNVSGPD YLLTMDLSTK

          EITHQLHLNN GLYAGYADAE LGTDGNTYVL GTYTANILRV TPDKKLSTFY
          EITHQLHLNN GLYAGYADAE LGTDGNTYVL GTYTANILRV TPDKKLSTFY
          EITHQLHLNN GLYAGYADAE LGTDGNTYVL GTYTANILRV TPDKKLSTFY
          EITHQLHLNN GLYAGYADAE LGTDGNTYVL GTYTANILRV TPDKKLSTFY

          VEEPLAPPRL YGFTGITHVG DTLIANNNII GQFVRFSVHD EEGTPIIIKQ
          VEEPLAPPRL YGFTGITHVG DTLIANNNII GQFVRFSVHD DEGTPIIIKQ
          VEEPLAPPRL YGFTGITHVG DTLIANNNII GQFVRFSVHD DEGTPIIIKQ
          VEKPLAPPRL YGFTGITHVG DTLIANNNII GQFVRFSVHD DEGTPIIIKQ

          TPYHNFTTSN VLNLPEKYED TILLATENAT PEHPSGGVGV FRSQDKLFHE
          TPYHNFTTSN VLNLPEKYED TILLATENAT PEHPSGGVGV FRSQDKLFHE
          TPYHNFTTSN VLNLPEKYED TILLATENAT PEHPSGGVGV FRSQDKLFHE
          TPYHNFTTSN VLNLPEKYED TILLATENAT PEHPSGGVGV FRSQDKLFHE

          MEFLGFIPSR LNQALATSAR QMADRIYVVS VYTDGANITV AGHTSKFVFQ
          MEFLGFIPSR LNQALATSAR QMADRIYVVS VYTDGANITV AGHTSKFVFQ
          MEFLGFIPSR LNQALATSAR QMADRIYVVS VYTDGANITV AGHTSKFVFQ
          MEFLGFIPSR LNQALATSAR QMADRIYVVS VYTDGANITV AGHTSKFVFQ

          DFTEEIDALV NAQSDEL
          DFTEEIDALV NTQSDEL
          DFTEEIDALV NTQSDEL
          DFTEEIDALV NTQSDEL

Tri17, 2403 aa
S40285    ---------- ---------- ---------- ---------M SEAQPIPLAI
S40288    ---------- ---------- ---------- ---------M SEADPVPLAI
S40293    PFDYIVLRLI KPPTVFFFPP EPTFQGTITV HTVYTTHTSM SEADPVPLAI
S7711     ---------- ---------- ---------- ---------M SEADPVPLAI

          VGIGCRFPAY ATNPGKLWDL LESGKFTWSK VPADRWNEEA FRRHSVDGIN
          VGIGCRFSGY ATNPGRLWDL LMSGKSTWSK VPADRWNEEA FRHHSPDGIN
          VGIGCRFSGY ATNPGRLWDL LMSGKSTWSK VPADRWNEEA FRHHSPDGIN
          VGIGCRFSGY ATNPGRLWDL LMSGKSTWSK VPADRWNEEA FRHHSPDGIN

          HHQGGHFLCQ DIGRFDAEFF GITPQEAASI ---------- DPQQRLLLET
          NHQGGHFLDQ DIDRFDAELF GITPEEAASI ---------- DPQQRLLLET
          NHQGGHFLDQ DIDRFDAELF GITPEEAASI ---------- DPQQRLLLET
          NHQGGHFLDQ DIDRFDAELF GITPEEAASI NSRADHSSGQ DPQQRLLLET

          TYEALENAGI RQDTIGGSET AVYMALSARD YDHNLDRGAT SVANHCVRSP
          TYEALENAGI RQDTINGSQT AVYTALSARD YDHNVDTGAT SVANPCVRSP
          TYEALENAGI RQDTINGSQT AVYTALSARD YDHNVDTGAT SVANPCVRSP
          TYEALENAGI RQDTINGSQT AVYTALSARD YDHNVDTGAT SVANPCVRSP

          QQDVPANTIS RFFNLTGPSV NIDSGSDGGM AAVTHACQAL RLGRSDVALA
          RQDVPANRIS RLFNLTGPSV NMDSGSDGGM AAITHACQAL RLGRSDVALA
          RQDVPANRIS RLFNLTGPSV NMDSGSDGGM AAITHACQAL RLGRSDVALA
          RQDVPANRIS RLFNLIGPSV NMDSGSDGGM AAITHACQAL RLGRSDVALA
```

```
GASSLILNPD QVNGINDAHG VRIDGCVPRS SSEGSTHSRG EGVAALVIKR
GASNLILNPD QVDGIDDAHG VRVDGRVPLS SSEGSTHSRG EGVAALVIKR
GASNLILNPD QVDGIDDSHG VRVDGRVPLS SSEGSTHSRG EGVAALVIKR
GASNLILNPD QVDGIDDAHG VRVDGRVPLS SSEGSTHSRG EGVAALVIKR

LDDAIRDQNP IRAILHDTNI GEHSYALGSR ENSRRIEICT YDNRGRASAE
LDDAIRDQNP IRAILRDANI GQNSHALGSG EDSRRIEICS NGTPGKGGAE
LDDAIRDQNP IRAILRDANI GQNSHALGSG EDSRRIEICS NGTPGKGGAE
LDDAIRDQNP IRAILRDANI GQNSHALGSG EDSRRIEICS NGTPGKGGAE

LRNTQNVFSQ YQDINLSSFT NQIESTIGDT GCVSAFAAVI ASVLFLENTT
PRNLQSAFSQ HQDINLSSFT DQIKSTIGDT GCVSALAAVI ASVMFLENTT
PRNLQSAFSQ HQDINLSSFT DQIKSTIGDT GCVSALAAVI ASVMFLENTT
PRNLQSAFSQ HQDIKSSSFT DQIKSTIGDT GCVSALAAVI ASVMFLENTT

IPPDKTQNDA AGVVVNAFSS YGTKSHVILE RTTGLLISSS GEEDSSPRLF
ISPDATQNDV AGVVVNAFSS YGTKSHVVLE RTTRPVTTSS GEEESSSRLF
ISPDATQNDV AGVVVNAFSS YGTKSHVVLE RTTRPVTTSS GEEESSSRLF
ISPDATQNDV AGVVVNAFSS YGTKSHVVLE RTTRPVTTSS GEEESSSRLF

IFSASSQTSL VRMLAVYADW TREHAGNAST LRDLSYTLSQ RRSFLPWRFS
VFSASSQTSL VRKLAVHADW TRKHGGNAST LRDLSYTLSQ RRSLLPWRFS
VFSASSQTSL VRKLAVHADW TRKHGGNAST LRDLSYTLSQ RRSLLPWRFS
VFSASSQTSL VRMLAVHADW TRKHGGNAST LRDLSYTMSQ RRSLLPWRFS

CIAENQAELL ESLANGSKKQ DSLVRTTPGA KISFVFTGQG SQWAEMGREL
CIAESQAELL EALASGSKKT DSLVRITPGA KISFIFTGQG SQWAEMGREL
CIAESQAELL EALASGSKKT DSLVRITPGA KISFIFTGQG SQWAEMGREL
CIAESQAELL EALASGSKKT DSLVRITPGA KISFIFTGQG SQWAEMGREL

LLYPAFHDSF QRSREVLQGL GCSWDLVEEI LKTSAESRLH DAELAQPATT
LLYPAFHDSF QRSREILQDL GCSWDLVEET LKTSAESRLH EAELAQPATT
LLYPAFHDSF QRSREILQDL GCSWDLVEET LKTSAESRLH EAELAQPATT
LLYPAFHDSF QRSREILQDL GCSWDLVEET LKTSAESRLH EAELAQPATT

AIQIALVDLA TEWGIVPDSV IGHSSGEVAA AYAAGYLSQH QAIKVAYVQG
AIQIALVDLA TQWGVVPDSV IGHSSGEVAA AYAAGYLSQH QAIKVAYVQG
AIQIALVDLA TQWGVVPDSV IGHSSGEVAA AYAAGYLSQH QAIKVAYVQG
AIQIALVDLA TQWGVVPDSV IGHSSGEVAA AYAAGYLSQH QAIKVAYVQG

FASRLAEERF GKGSMLAVGV GEYGIEPYMD LLSHEGAVVA CQNSPNSTTI
FASRIAEERF GKGSMLAVGV GEYEVEPYMD LLSHEGAVIA CQNSPNSTTI
FASRIAEERF GKGSMLAVGV GEYEVEPYMD LLSHEGAVIA CQNSPNSTTI
FASRIAEERF GKGSMLAVGV GEYEVEPYMD LLSHEGAVIA CQNSPNSTTI

AGDDAAITEL SELVSQESIF NRKLNVDTAY HSHHMQAAAS KMHSALKDII
AGDDAAITEL SELLSRESIF NRKLNVDTAY HSHHMQTAAS MMQSALKDII
AGDDAAITEL SELLSRESIF NRKLNVDTAY HSHHMQTAAS MMQSALKDII
AGDDAAITEL SELLSRESIF NRKLNVDTAY HSHHMQTAAS MMQSALKDII

GAPSAKNGIE VFSTVTGSVK SDAFGADYWI ANLINKVRFC DGLQALCESK
SAPSTNNGIE LFSTVTGSVK SDAFGADYWI ANLINKVRFC DGLQALCESK
SAPSTNNGIE LFSTVTGSVK SDAFGADYWI ANLINKVRFC DGLQALCESK
SAPSTNNGIE LFSTVTGSVK SDAFGADYWI ANLINKVRFC DGLQALCESK

RPSPSCSPET SRIFIELGPH SALAGPIRQC IADMIAPISY SYTSALVRGT
RPSPLCGPES QRIFIELGPH SALAGPIRQC MTDMITPISY CYTSALIRGT
```

```
RPSPLCGPES QRIFIELGPH SALAGPIRQC MTDMITPISY CYTSALIRGT
RPSPLCGPES QRIFIELGPH SALAGPIRQC MTDMITPISY CYTSALIRGT

GAARSALSMA GQVFKQGYSL DLAGVFASYE TSGHMSVMCD LPPYPWDHTR
GAARSTLIMA GQVFNQGYPL NLAAVFASYE TIGYTSVISN LPPYPWDHTR
GAARSTLIMA GQVFNQGYPL NLAAVFASYE TIGYTSVISN LPPYPWDHTR
GASRSTLIMA GQVFNQGYPL NLAAVFASYE TIGYTSVISN LPPYPWDHTR

RYWNESRGSR DYRFRKHPYH DLIGLRMTDN SSIHPSWRHT VSLDNLPWLG
RHWNESRASR DHRFRKHPYH DLLGLRMTDN SSLHPLWRHR VSLGNLPWLG
RHWNESRASR DHRFRKHPYH DLLGLRMTDN SSLHPLWRHR VSLGNLPWLG
RHWNESRASR DHRFRKHPYH DLLGLRMTDN SSLHPLWRHR VSLGNLPWLG

DHIVNHLVLF PCSGYLAMAV EACSQLTGDF HPGKRVERFS LNDVSFLKEL
DHIVNHLVLF PCSGYLAMAV EACSQLIGDY YPGKNVEKFF LKDVSFLKEL
DHIVNHLVLF PCSGYLAMAV EACSQLIGDY YPGKNVEKFF LKDVSFLKEL
DHIVNHLVLF PCSGYLAMAV EACSQLIGDY YPGKNVEKFF LKDVSFLKEL

VIPEDHTSIE IQLCLTSIEH APSQASHTTT KYGFSITAYT SDGQWNEYCH
VIPEDHNSIE IQLCLTSVAH FSSEASHSST RYGFSITAYT TDEQWNEYCH
VIPEDHNSIE IQLCLTSVAH FSSEASHSST RYGFSITAYT TDEQWNEYCH
VIPEDHNSIE IQLCLTSVAH FSSEASHSST RYGFSITAYT TDEQWNEYCH

GTVACEFAAA QPLLVADITQ ADLMHQLDPA LGNLKQAREF YEELGRLGTV
GTIACEFAAG QPLLVADVTQ AHMMHQLDSA SGNLIQARDF YEELGRLGTV
GTIACEFAAG QPLLVADVTQ AHLMHQLDSA SGNLIQARDF YEELGRLGTV
GTIACEFAAG QPLLVADVTQ AHLMHQLDSA SGNLIQARDF YEELGRLGTV

YGSTFTGIEE MTIEGDSAAS YIVVPDVVSA MPSRQLSPHI IHPTTLDILL
YGSTFKGIEE MTVDGDSAAS CIVIPDVVTT MPCRHLSPHI IHPTTLDILL
YGSTFKGIEE MTVDGDSAAS CIVIPDVVTT VPCRYLSPHI IHPTTLDILL
YGSTFKGIEE MTVDGDSAAS CIVIPDVVTT MPCRHLSPHI IHPTTLDILL

HTSMPLVHQK LGAGPVTLAH IKNMCITAAI DNTPGDAFRT VTNLGSSHAN
HTSIPLVHQK LGVGPVTLAH IENMIITAAI DNKPGDAFRT VTNLVSIHSN
HTSIPLVHQK LGVGPVTLAH IENMIITAAI DNTPGDAFRT VTNLVSIHSN
HTSIPLVHQK LGVGPVTLAH IENMIITAAI DNTPGDAFRT VTNLVSIHSN

AAVADIFVFS DKAGASDAPV IYASGIELRS SSPGMDDTDT RDGLPEICYE
AAVADLFVFS EKAGATDAPV LYASGIELRS SSPDMDDTNT PNGLPDICYE
AAVAELFVFS EKAGATDAPV LYASGIELRS SSPDMDDTNT PDGLPDICYE
AAVAELFVFS EKAGATDAPV LYASGIELRS SSPDMDDTNT PDGLPDICYE

MKWVLDERFI SAKRLQPLRP FSVSEDGLAH CCAFLAEYVK HKANKQSGLA
MKWVLDERFI SAKQLQSLRP FSVSEDGLTG CCAFLAEYVK QKANKQSDLA
MKWVLDERFI SAKQLQSLRP FSVSEDGLTG CCAFLAEYVK HKANKQSDLA
MKWVLDERFI SAKQLQSLRP FSVSEDGLTG CCAFLAEYVK HKANKQSDLA

VIELVGPDAA SSATVAFVEA LHSSDARPIV YDFASLSGNF DGIQSALQDQ
VIELAGPDAA SSATVAFLEA LRSSDARPTV YDFASSSGNF DGIQNALHDQ
VIELAGPDAA SSATVAFLEA LRSSDARPTV YDFASSSGNF DGIQNALHDQ
VIELAGPDAA SSATVAFLEA LRSSEARPTV YDFASSSGNF DGIQNALHDQ

DMSVFNFREL DIGADHLDPS FNEHYYNIVL ASNILRDTSN IRSILANARR
DMSVFNFREL DIGADHLDPS FNEHYYNIVL ASNILRETSN IRSILINARR
DMSVFNFREL DIGADHLDPS FNEHYYNIVL ASNILRETSN IRSILTNARR
DMSVFNFREL DIGADHLDPS FNEHYYNIVL ASNILRETSN IRSILTNARR
```

```
LLKPDGVLLL VEDAASSQDS LSIEECADLM LDASFKMHLA ISDNQKRPQC
LLKPDGVLLL VEDATSGQDS LSIEECADLM LDASFKMQLA SPDNQKRPRC
LLKPDGVLLL VEDATSGQDS LSIEECADLM LDASFKMQLA SPDNQKRPRC
LLKPDGVLLL VEDATSGQDS LSIEECADLM LDASFKMQLA SPDNQKRPRC

TFFIARAFTK APTCIPKMIL VSQDSSRHEN FKHLATEMSN TLGRKVAHVV
TFFIARALTK APVLIPKIIL VSQDSSRHEN FKHFATEMSN TLGSNVTQVV
TFFIARALTK APVLIPKIIL VSQDSSRHEN FKHFATEMSN TLGSNVTQVV
TFFIARALTK APVLIPKIIL VSQDSSRHEN FKHFATEMSN TLGSNVTQVV

TELWHEMQPH DANAIYVVID DGAQPLLANV SQDHFRRVVD ILQKPAKVIW
KESWHDMQPH DANAIYVVID DGAQPLLANV SQNRFRHVVD LLQKPAKVIW
KESWHDMQPH DANAIYVVID DGAQPLLANV SQNRFRHVVD LLQKPAKVIW
KESWHDMQPH DANAIYVVID DGAQPLLANV SQNRFRHVVD LLQKPAKVIW

LSVQHNVESR FNPRKHFING VSRTAHKENR DLDMVTIDVQ QTLNQKTKPA
LCVQHSEKSS FNPKKHFING VSRTAHAENR DLDMVTIDVQ QTLDQKTEHA
LCVQHSEKSS FNPKKHFING VSRTAHAENR DLDMVTIDVQ QTLDQKTEHA
LCVQHSEKSS FNPKKHFING VSRTAHAENR DLDMVTIDVQ QTLDQKTEHA

ILQLLSSIFE SFGKKDILRE REYVFNGEDV LVPRLVPHAT LNHQISGKSE
ILQLLSNIVD SFGKKDILRE REYVFKGEDV LVPRLLPHAT LNHQISGKLE
ILQLLSNIVD SFGKKDILRE REYVFKGEDV LVPRLLPHAT LNHQISGKLE
ILQLLSNIVD SFGKKDILRE REYVFKGEDV LVPRLLPHAT LNHQISGKLE

TTIQTMPFSG SPVSLKMVDD KKGFVFVENA SHEQPLLDDF VEIESKAFGI
TTIQTIPFSK SPVSLRMIDD KKGFVFVENA SHGQPLLDNF VEIESKAFGI
TTIQTIPFSK SPVSLRMIDD KKGFVFVENA SHGQPLLDNF VEIESKAFGI
TTIQTIPFSK SPVSLRMIDD KKGFVFVENA SHGQPLLDNF VEIESKAFGI

PASYIKSAQT GNFFNEYTGV ITAVGCQVVA PKIGDRVVTI STESCANRLR
PANYTRSAQS GYFFNEYAGV VAAVGCQVLG PKIGDRVVTI STDSCANRLR
PANYTRSAQS GYFFNEYAGV VAAVGCQVLG PKIGDRVVTI STDSCANRLR
PASYTRSAQS GYFFNEYAGV VAAVGCQVLG PKIGDRVVTI STDSCANRLR

VPAGHVQVIP TQLSFTDAAT LPLAFMAVTH ALVDVANVQA KQMVLVDNAT
VPAGHVQAIP RHLSFTDAAA LPLAFMAAIH ALVDVANIQA KQVLLVDNAT
VPAGHVQAIP RHLSFTDAAA LPLAFMAAIH ALVDVANIQA KQVLLVDNAT
VPAGHVQAIP RHLSFTDAAA LPLAFMAAIH ALVDVANIQA KQVLLVDNAT

SEYGQAALIV ARNLEATVIA AVARVDEASF LQNVFDISPS HIVARDSYLS
SEYGQAALIV ARNLEATVIA AVARVDEAAF LQNVFEIQPS HIVARDSYLS
SEYGQAALIV ARNLEATVIA AVARVDEAAF LQNVFEIQPS HIVARDSYLS
SEYGQAALIV ARNLEATVIA AVARVDEAAF LQNVFEIQPS HIVARDSYLS

HRQMQRSLGL DGGIDVILGC GSTPVTTAVS RMLKPFGTFV SIQPRGGTSE
HRQMQRILGL GGGIDVILGC GSTPVTTAIS RMLKPFGTVV NIQPRGGTSG
HRQMQRILGL GGGIDVILGC GSTPVTTAIS RMLKPFGTVV NIQPRGGTSG
HRQMQRILGL GGGIDVILGC GSTPVTTAIS RMLKPFGTVV NIQPRGGTSG

RHYGATCCSN ATVASFDIDS LLQAKPHKVP VLLQQVVKMV DQGVLLPPRS
HHYGATFCSN ATVATFDIDS LLQARPHKLP VLLEQVVKMV DQGLLLPPRS
HHYGATFCSN ATVATFDIDS LLQARPHKLP VLLEQVVKMV DQGLLLPPRS
HHYGATFCSN ATVATFDIDS LLQARPHKLP VLLEQVVKMV DQGLLLPPRS

TVGLVLNNKL AKELTLIQKQ ENLAKHVIEV QKHSTVKVEK PSYQVPDLDT
TVVLALNNKL EKELNLTQKH GNLVKHIIEV QEHSTVKVEK PSYQVPDLAT
TVVLALNNKL EKELNLTQKH GNLVKHIIEV QEHSTVKVEK PSYQVPDLAT
TVVLALNNKL EKELNLTQKH GNLVKHIIEV QEHSTVKVEK PSYQVPDLAT
```

```
            DATYVVAGGL NDLSQRFLHL MAHAGARYLV TLSSSQVASQ AFHEFGKKLR
            DATYVVAGGL NDLSQRFLLW MAHAGARYLV TLSSSEEASE TFLEFRKKFK
            DATYVVAGGL NDLSQRFLLW MAHAGARYLV TLSSSEEASE TFLEFRKKFK
            DATYVVAGGL NDLSQRFLLW MAHAGARYLV TLSSSEEASE TFLEFRKKFK

            SGNSHCSLVA LNCDVSKVHS VTAALSEIRG HDFPSVKGVI LLNTARNDSA
            DGNSHCSLMT VNCDVSKVHS ITAALSEIRE HGFPSVKGVI LLNTARNDSA
            DGNSHCSLVT VNCDVSKVHS ITAALSEIRG HGFPSVKGVI LLNTARNDSA
            DGNSHCSLMT VNCDVSKVHS ITAALSEIRG HGFPSVKGVI LLNTARNDSA

            LAAMTADTFN TVTNAKVAGV LNLHTVFGSE NLDFFISVSS VTNIIGTEMQ
            LAAMTADTFN AVTNAKVAGV LNLHTVFGRE DLDFFISMSS VTNIIGAEGQ
            LAAMTADTFN AVTNAKVAGV LNLHTVFGRE DLDFFISVSS VTNIIGAEGQ
            LAAMTADTFN AVTNAKVAGV LNLHTVFGRE DLDFFISMSS VTNIIGAEGQ

            ANGNAGDAFQ DALAHFDRDT GCFNMVLNIG GFEGAAHDNG SSIQASPREG
            AIGNAGDAYQ EALAHFDGDT GCFNMVLNIG GFEGAAPDNG PSIQASPREG
            AIGNAGDAYQ EALAHFDGDT GCFNMVLNIG GFEGAAPDNG PSIQASPREG
            AIGNAGDAYQ EALAHFDGDT GCFNMVLNIG GFEGAAPDNG PSIQASPREG

            FDHISDQELT AYLDYALSAN TRRTGCHQSV IGLTPNSIAQ TFATNGAAQT
            FSHISDQELT AYLDYALSAN ARTTRCHQSV IGLMPDIIAR TIASNGAART
            FSHISDQELT AYLDYALSAN ARTTRCHQSV IGLMPDIIAR TIASNGAART
            FSHISDQELT AYLDYALSAN ARTTRCHQSV IGLMPDIIAR TIASNGAART

            SMFTHVRRNA GALMDESEST ARERRFEHIV QEGASNEEIS AFVARSIGDK
            SMFTHVRRNA GALMDENDSA ARERRFEHMV QEGASKEEIS AFVARSIGNK
            SMFTHVRRNA GALMDENDSA ARERRFEHMV QEGASKEEIS AFVARSIGNK
            SMFTHVRRNA GALMDENDSA ARERRFEHMV QEGASKEEIS AFVARSIGNK

            VAEFAAIDPM EVNFDSSILD YGLDSLMAIE LRNWIVRDFD APIRLPEVVD
            VAEFAAIDPM EVKFGSSILD YGLDSLMAIE LRNWMAREFD APIQLPEVVD
            VAEFAAIDPM EVKFGSSILD YGLDSLMAIE LRNWMAREFD APIQLPEVVD
            VAEFAAIDPM EVKFGSSILD YGLDSLMAIE LRNWMAREFD APIQLPEVVD

            SPDIWTLSER VICCSQLTSS RSDTSKSSVI SGSEQDPLST LPTSRANTPE
            SPDIWTLSER VVNCSQLTTS RSDTSKSSVV SGSEQDPLST LPTSRANTPD
            SPDIWTLSER VVNCSQLTTS RSDTSKSSVV SGSEQDPLST LPTSRANTPE
            SPDIWTLSER VVNCSQLTTS RSDTSKSSVV SGSEQDPLST LPTSRANTPE

            VKE
            VKE
            VKE
            VKE

Tri18, 541 aa
S40285      MGSRGQPTAD STRSVPLDAS SPKAEQYAFP LPTLDPAEFR WHPSPKNNSI
S40288      MGSSGQQTAD STRPVPPDAS SPKAEQYAFP LPTLDPAEFR WHPYPKNNSI
S40293      MGSSGQQTAD STRPVPPDAS SPKAEQYAFP LPTLDPAEFR WHPYPKNNSI
S7711       MGSSGQQTAD STRPVPPDAS SPKAEQYAFP LPTLDPAEFR WHPYPKNNSI

            LQRKANGVEA LVGIKDANAV GTYDLYNNIV LRVGDISDLT LPRLKRAFVR
            LQRKANGVEA LVGIKDANAV GTYDLYNNIV LRVGDIPDLT LPRLKRAFVR
            LQRKANGVEA LVGIKDANAV GTYDLYNNIV LRVGDIPDLA LPRLKRAFVR
            LQRKANGVEA LVGIKDANAV GTYDLYNNIV LRVGDIPDLT LPRLKRAFVR

            AMLDARFENP SIACYGVWGQ NKEPHLPHIQ YRPFKSHNEA RTWAYNSIYV
```

```
ALLDARFENP SIACYGVWGQ NKEPHLPHIQ YRPFKSHNEA RAWAYNSIYV
ALLDARFENP SIACYGVWGQ NKEPHLPHIQ YRPFKSHNEA RAWAYNSIYV
ALLDARFENP SIACYGVWGQ NKEPHLPHIQ YRPFKSHNEA RAWAYNSIYV

RATSLTSSEL RAERIEKRRA EAIPKSSNSL DVVISADVAH ERTILEPGTK
RATSLTSSEL RAERIEKRRA EAVPQPSNSL DIVISADVAH ERTILEPGTK
RATSLTSSEL RAERIEKRRA EAVPQPSNSL DIVISADVAH ERTILEPGTK
RATSLTSSEL RAERIEKRRA EAVPQPSNSL DIVISADVAH ERTILEPGTK

LDLMFLFNHL SWDGKARSFT SELVCRATQI LEKGLENTVP AYRWGEEKAR
LDLMFLFNHL SWDGKARSFT SELVHRATQI LEKGLENTVP AYRWGEEKAR
LDLMFLFNHL SWDGKARSFT SELVHRATQI LEKGLENTVP AYRWGEEKAR
VDLMFLFNHL SWDGKARSFT SELVHRATQI LEKGLENTVP AYRWGEEKAR

LDPPILDVML VGMETLGEDY KAVHRRLLES QMQVGLSWGL PVTNHPGDPL
LDPPILDVML VGMETLGDDY KAVHRKLLES QMQVGLSWGL PVTNHPGDPL
LDPPILDVML VGMETLGDDY KAVHRKLLES QMQVGLSWGL PVTNHPGDPL
LDPPILDVML VGMETLGDDY KAVHRKLLES QMQVGLSWGL PVTNHPGDPL

QLRYCMSAEE GKRILNAVKS RLGLKYNIGH LGHAATVLAL LKHHPIPASA
QLRYCMSVEE GKRIANAVKS RLGLKYNIGH LGHAATVLAM LKHHPIPASA
QLRYCMSVEE GKRIANAVKS RLGLKYNIGH LGHAATVLAM LKHHPIPASA
QLRYCMSVEE GKRIANAVKS RLGLKYNIGH LGHAATVLAM LKHHPIPASA

EDTAFLFSPL PVDGRGYLSE DRTTQRYGNA QASAVVEFQK LASWGIKKDD
EDTAFLFSPL PVDGRGYLSE DRTTQRYGNA QASAVVEFQK LASWGIKKDD
EDTAFLFSPL PVDGRGYLSE DRTTQRYGNA QASAVVEFQK LASWGIKKDD
EDTAFLFSPL PVDGRGYLSE DRTTQRYGNA QASAVVEFQK LASWGIKKDD

PEGLKAALDN LAKKIRDDYN FWLGQSDCLL PISVANHNFA SNLIATSSAT
PQGLKTALDK LAKKIRDDYN FWLGQADCLL PISVANHNFA SSLIATSSAT
PQGLKTALDK LAKKIRDDYN FWLGQADCLL PISVANHNFA SNLIATSSAT
PQGLKTALDK LAKKIRDDYN FWLGQADCLL PISVANHNFA SNLIATSSAT

PNVHAPAFCN DGRSENIISY EVLGLTGKKL FEVEDCFMGV EVIGYNAFIR
PKVHAPAFCN DGRSENIISH EVLGLTGKKL FEVEDCFMGV EVIGYNAFIR
PKVHAPAFCN DGRSENIISH EVLGLTGKKL FEVEDCFMGV EVIGYNAFIR
PKVHAPAFCN DGRSENIISH EVLGLTGKKL FEVEDCFMGV EVIGYNAFIR

MDTWKDAIRL TLCYNNGCFS DALAKEYTKD VAEYMLSYAD A
MDTWKDAIRL TLCYNNGCFS DALAKEYTKD VAKYMLEYAD A
MDTWKDAIRL TLCYNNGCFS DALAKKYTKD VAKYMLAYAD A
MDTWKDAIRL TLCYNNGCFS DALAKEYTKD VAKYMLAYAD A
```

Core atranone cluster, 14 products

```
Atr1, 234 aa
S40285     MASIPGLLFS LTKPMDPTIP EAQFNDWYTN KHLVDTVNSG LASLAVRFKN
S40288     MASIPGLLFS LTKPMDPNIP EAQFNDWYTN KHLVDTVNFG LASLAVRFKN

           VNPSHQWPYL ALYRLQDLAK LYNMEFMSSL PTDSPAGWGV PNSKADIRIE
           VNPSHQWPYL ALYRLQDLAK LYDMEFMSSL PTDSPAGWGV PNSKADIRIE

           PRGYQLLTTL ERENAKTGVP KFVLTVEFRE SFMNAEAFVA SCQGLQLDDV
           PRGYQLLTTM ERENANTGVP KFVLTVEFRE STMNAEAFVA SCQSLQFDDV

           GKQPGYRRSM LYQAGRSLVT QEGKAGTEFR SAEQQQPSYL VVHEFDQMPT
           GKQRGYRRSM LYQAGRSLVP QEGKAGTEFR SAEQQQPAYL VVHEFDQMPA

           NTFQEQGASG LWEYMAEYGT GLYRTEPVPV KVYN
           HTFQEEGASG LWEYMAEYGT GLYRTEPVPV KVYN

Atr2, 540 aa
S40285     MAVISLFRII VDKWHVVLAC SACLGALLFQ ALRRQSNSTK DVPFIGMELG
S40288     MASMSLFRII VDEWRVVLAC SACLGALLFQ ALRRQSNSTK DVPFIGMELG

           SAEKRRKAYM TDARSLFRDG YQQFKDRVFG ITTTSENLVV VVPPRFLDEL
           SAEKRRKAYM TDARSLFRDG YQRFKDRVFG ITTTSENLVV VVPPRFLDEL

           GRLPDEVLSA SMAVADISQD KYTKMEITDP IISHAVRGNL TMSLSRLNDA
           GRLPDEVLSA SMAVADISQD KYTKMEITDP IISHAVRRNL TMSLS-----

           ILEELRKALS LLLPTCDEWT SVNISEKLQR IVAVISGRVF VGPELCGSDA
           ---------- ---------- ---------- ---------- ----------

           YLDAAIHIAH EASAAVQSIS TLPPWKRPFL SARLPELRAL RERQDKVHSV
           ---------- --SAAVQSIS TLPPWKRPFL SARLPELRAL RERQEKVYSV

           LRPVLEKRIQ MNEEDRPDDM LTWIISSQKK HGERSIETMA KVQTALHLAA
           LRPVLEKRIQ MNEADRPDDM LTWIISSQKK HGERSIETMA KVQTALHLAA

           IGTTSEMATN AFYNLAAMPE LVPELREEIR TVLEEHDGVV STKSLQAMKK
           IGTTSEMATN AFYNLAAMPE LVPELREEI- ---------- ----------

           LDSFLKETAR LYPPFLCKNF IPMPWWLPGG DNRISDANIV YPKAAFERKV
           ---------- ---------- ---------- ---------- ---PAFERKV

           LRTFTLSNGQ VIPAGALIKV PSQAIMTDPA LFPDPDRFDA FRFYDLQQQK
           LRTFTLSNGQ VIPAGTLIKV PSQAIMTDPM LFPDPDRFDA FRFYDLQQQK

           NILKDGSVSV GASVNQFVNS NKNSLVFGYG RHACPGRFLA ADELKMILVY
           RGLKDGSVSV GPSVNQFVNS NKNSLVFGYG RHACPGRFLA ADELKMILVY

           FLQAYEIRLE EGESRRYRNL EFAAFSIPDP TKTIQMKKLQ
           FLQAYEIRLE EGETQRYRNL EFAAFSIPDP TKTIQMKKL-

Atr3, 478 aa
S40285     METLSQRITS MESVQLQGIA VAFVTASALY YVLPAAISHI QLSALPMLGK
```

```
S40288     METLSQRMTS MESVQLRGIA VAIVTASALY YVLPAAISHI QLSALPMLGK

           TEVVVIPPKL LSELSKSPRT LSAEIAGNEF IAGKYTKVKA LTPILLHSIT
           TEVVVIPPKL LSELSKSPRT LSAEIAGNEP VPELHQDWLS VNIRLV----

           KYLIPSLGRN AVVMSEEVSN AVRLGIPTCN DWTGVNIYPK IMRMVTVSTG
           ------AGRN AVIMSEEVSN AVRLGIPPCN DWTAVKIYPK IMRMVTVSTG

           RFLVGSELNR SEDYIDTVHN YALDVSSAQS AVHKMHPWIR PLLAEWLPEI
           RFLVGSELNR SEEYIDTVHN YALDVSSAQS AVHKMHPWLR PLLAEWLPEI

           RRLRKRTEEA FALFESLIKE RMKMQRELSE SELPDDLLQW MIANRHNYNN
           RRLRKRTEEA FALFESLIKE RMDMQRELSE SELPDDLLQW MIANRHNYNN

           EDAHDLVYSQ LGLTFTANHS TASTITNALY TLATMGDLID VIRDDITQAL
           EDAHDLVYSQ LGLTFTANHS TASTITNALY TLATMGDLID VIRDDITQGL

           AESGGQFTSK ALDSMWKFDS FIKETVRMNP LVMSVAVRKV VEPIKLPSGQ
           EESGGQFTSK ALDSMWKFDS FIKETVRMNP LVMSVAVRKV VEPIKLPSGQ

           VIPTGVTLET PLVAVNLDDQ IFPNADVFDP MRFYNLREKD RKQGDAREAE
           VIPTGVTLET PLVAVNLDDQ IFPNADVFDP MRFYNLREND RKQGDARDAE

           FNQLISSSTS HMSWGFGKHT CPGRAFAAQQ IKMILAHIIL RYDIKLVGDS
           FNQLTSSSTS HMSWGFGKHT CPGRAFAAQQ IKMILAHIIL RYDIKLVGHS

           TDRYENIPKG HLSLPDPTKD ILMKRREI
           TDRYENIPKG HLSLPDPTKD ILMKRREI

Atr4, 524 aa
S40285     MRLDLLGPVA TRIITYLDSL TWVGMALPLF SLCWAISYAR GKAYPTVPGA
S40288     MGLDLLGPAA TRIATYLDSL TWVEIALPLF SLCWAISYAR GKGYPTVPGA

           PVYGYNSRFE PSFMLKSRTY TGFYDILSNG YKMLKDVPFV IPRHDTNINI
           PVYGYNSRFE PSFMLKSRTY TGFYDILSKG YTMLKDVPFV IPRHDTNINI

           LPIKYLDEIR LMPKHILNSH LVLISQMTPK WTWLQPAADS DLVTRVLLTK
           LPIKYLDEIR LMPKHILNSH LVLISQMTPK WTWLQPAADS DLVTRVLLTK

           LNPDLQKYVD ITRLELDSAF KSDFPRHDEE WTEVDFQPLI RRVLTRISAK
           LNPDLQKYVD ITRLELDSAF KSDFPRHDEE WTEVDFQPLI RRVLTRISAK

           IFLGEPACLN EDWLRIAIGY TAGALEVTKD LHKFPSWTHF LVAPLLPSRR
           IFLGEPACLN EDWLRIAIGY TAGALEVTKD LHKFPSWTHF LVAPLLPSRR

           RLRRELDIAM KIVEKQIQLH EQAEKDGLKN YDTLLDWMLD NCSDKESSVE
           RLRRELDIAM KIVEKQIQLH DQAERDGVKN DDTLLDWMLD NCSDKENGVE

           AMTIFQCFIA MASIHTTEFS LANVLFDLCA HPEWFPVLRE ELDEVIRVHG
           AMTIFQCFIA MASIHTTEFS LANVLFDLCA HPEWFPVLRE ELDEVIRAHG

           NIGHRLPAKQ WLQKLEKMDS LLAETLRLCP TMLTSIQRLA LEKVQLKDGT
           HIGEKLPAKQ WLQKLEKMDS LLAETLRLYP TMLTSIQRLA LEKVQLKDGT

           VIPKGSRLAW ASLHHVTDPE VDGTLAAWDP MRNYRKRHSG SGENLTKFVA
```

```
           VIPKGARLAW ASLHHVTDPD VDGTLAAWDP MRNYRKRHSS SGENLNKFVA

           GQINESTLGF GYGNQACPGR YFAVNEIKMM LARLLLEFEF KFPEGKSRPK
           GQINESTLGF GYGNQACPGR YFAVNEIKMM LARLLLEFEF KFPEGKSRPK

           VFFIGEIACL DHDATLMMRN VRTC
           VFFVGEIACL DHDATLMMRK VRTC

Atr5, 297 aa
S40285     MAVMTREDVE FRTMDGLTLR GWLYPSGMRG PALVMTQGFN ASKEYLLADV
S40288     MTVMTREDVE FRTVDGLTLR GWLYPSGKRG PALVMTQGFN ASKEYLLADV

           AVWFQKRGVT SLLVDIRTTG LSDGEPRNDI DLDKQVEDCH DAVSFLSRHP
           AVWFQKRGVT SLLVDIRTTG LSDGEPRNDI DLDKQVEDCH DAVSFLSRHP

           SVDPEMIVYW GYSLGAVISL CAAALDKRAA AVIATAPNTD FIFDPVKRAA
           SVDPEMIVYW GYSLGAVISL CAAALDKRAA AVIATAPNTD FIFDPVKRAA

           TLSLAMRDRL SRLAGNPPLY LKIIGEDGQN PAGWYLGEER RSPEELDALF
           TLSLAMRDRV SRLAGNPPLY LKIIGEDGQN PAGWYLGEER RSPEELDALF

           NSSNIMNQVT IQSYYRLLRW QPFGLMPSVS PTPVMIVTPS DDDLSKPENQ
           NSSNIMNQVT IQSYYRLLRW QPFGLMPSVS PTPVMVVTPS DDDLSKPENQ

           RKLFDIFQEP KEFVLAENKG HMNCISGEDG EQFLQKQLEF MKRMLKF
           RKLFDIFQEP KEFVLAENKG HMNCISGVDG EQFLQKQLEF MKRMLKF

Atr6, 2439 aa
S40285     MDSEAPTPTS SSFALPYAEP IAIVSAACRL PGHIQNPHQL WQFLQAGGIA
S40288     MDSEAPTPTS SSFALPYAEP IAIVSAACRL PGHIQNPHQL WQFLQAGGIA

           TSDVVPESRY NVAGHFDGSG RPGTLKTPGG MFIEDIDLGA FDAPFFHIGK
           TSDVVPESRY NVDGHFDGSG RPGTLKTPGG MFIEDIDLGA FDAPFFHIGK

           SDAVSMDPQQ RQLLEVVYEC LENGGITMQG IDGDQIGCFV ASYSADWHEM
           SDAVSMDPQQ RQLLEVVYEC LENGGITMQG IDGDQIGCFV ASYAADWHEM

           QSRHPASRAP GTTAGTSRAI LSNRISHFFN IKGSSWTIDT ACSGGLVGVD
           QSRHPALRAP GTTAGTSRAI LSNRISHFFN IKGSSWTIDT ACSGGLVGVD

           AACQYLRAGK LNGAIVAAAQ LWMSPEYNEE LGTMRAAASS TGRCHSFDAK
           AACQYLRAGK LNGAIVAAAQ LWMSPEYNEE LGTMRAAASS TGRCHSFDAK

           ADGYCRSEAV NAVYLKRLSD ALRDGDPVRA VIRGTANNSD GRTPGLHSPN
           ADGYCRSEAV NAVYLKRLSD ALRDGDPVRA VIRGTANNSD GRTPGLHSPN

           SDAQAAAIRA AYADAGIDST QYTKTAFMEC HATGTPAGDP SEVRGSASVL
           ADAQAAAIRA AYADAGIDST QYTKTAFMEC HATGTPAGDP SEVKGSASVL

           ASMRPPSDPL IIGTIKSNLG HAEPGAGISG LMKAMMAVEK GIIPGNPTFI
           ASMRPPSDPL IIGTIKSNLG HAEPGAGISG LMKAMMAVEK GIIPGNPTFI

           TPNPNIDFAG LRVRASQRNM RWPQSTKDYR RASVASSGFG GSNAHVVLDN
           TPNPNIDFAG LRVRASQRNM RWPQSTKDYR RASVASSGFG GSNAHVVLDN
```

```
AEHYMQHHFL SVQPQFRTYV SSYAETGDVL SMLSGFGLGA ANSSDKLAPL
AEHYMQHNFL AAQPQTRTYV SSYAESSDVL SMLAGFGLGG ANSSDTPAPL

PNVLVFSAHD ADSLKRQMGA LSAHLVDPRV AIKLSDLSYT LSERRSRHFH
PNVLVFSAHD ADSLKRQMEA LSAHLVDPRV AVKLSDLSYT LSERRSRHFH

RSFIVCRPNK GGNIETLPTD LAKYAMKPTS PVRIGFVFTG QGAQWSGMGA
RSFIVCRPNK GGNIETLPTD LAKYAKKPTS PVRIGFVFTG QGAQWSEMGA

DLIRLFPKTA KAVVDELDAA LQELPADVRP SWSLLAELTE PRSSEHLREP
DLIRLFPQTA KAMVDELDAV LQGLPADIKP SWSLLAELTE PRSSGHLREP

EFSQPLVTAL QLALLAVLKS WNVTADAVVG HSSGEIAAAC SAGLLTPGQA
EFSQPLVTAL QLALLAVLKS WNVTADVVVG HSSGEIAAAC SAGLLTPGQA

ILTAYFRGQA AKQVVMEGSM GMLAVGLGSA GVQKYLEDTS RAGKVVIACY
ILTAYFRGQA AKQVAMEGSM GMLAVGLGSA GVQKYLDDAS RAGKVVIACY

NSPASVTLSG PTSLLSELAQ VIQTDGHFAR LLQVNLPYHS HYMSAIGDRY
NSPASVTLSG PTSLLSELAQ VIQTDGHFAR LLQVNLPYHS HYMSAIGDRY

EKLLLDHGRL DEIQGETATR KIPMISSVST IVLEGSKSCS AAYWKSNMVS
EKLLLDYGRL DETQGETEAR NIPMISSVST AVLEGSESCS AAYWKSNMVS

AVQFDGACKR IVADQDLSAN LLIEIGPSAA LGGPIGQIIK QAGIDNVTYT
AVQFDGACKR TVTDQNLAAN LLIEIGPSAA LGGPIGQIIK EAGIENVTYA

SAAQRGTDSI LALFGVAGQL FLHDCPVSLD HVNTDETALT EPKPAVIIDL
SAAQRGADSI LALFGVAGQL FLHDSSVSLD RVNTDEAALI EPKPAVIIDL

PNYRWNHSTR YWHESLASKD WRFRNFPEHD LLGGKVLGTA WESPSWTKTL
PNYRWNHSTR YWHESLASKD WRFRNFPEHD LLGGKVLGTA WESPSWTKTL

RLEDVPWLRD HKIGSEILFP ASGYIAMAVE AARQATISTA RSQNKAAPSA
RLEDVPWLRD HKIGSEILFP ASGYIAMAVE AARQATISTA RSQNKAVPSA

HAYHYVLRDV HFERGLVLED ETDTTLMLSL APVARLGVKW WVFKVMSLAS
HAYHYVLRDV HFERGLVLED ETDATLMLSL APVARLGVKW WVFKVMSLAS

GGSSSSSDSW IEHSNGLVRL ALNASEPLPR VTPDNYSLPL QYPTPARFWY
GGSSSSSDSW IEHSNGLVRL ALNASEPLPQ ATPDNYSLPL QYPTPARFWY

KAFENAGYGY GPGFQKQSYI ECTEGSFSAR STIMLNPPLS KWEPQPNYPL
KAFDNAGYAY GPGFQKQSDI ECTEGSFSAR STIMLNPPLS KWEPQPDYPL

HPASMESCIQ ATLTSMYRGD RAGINNVLVP NAIDRIILSG DTWRSNEAVS
HPTSMESCIQ STLTSMYRGD RASINNVLVP NAIDRIILSG DIWRSNEAVS

VTTSESSSGI TSKPLSNASL FDPTNGVLII DLRGISMTSV GLQGNVCSFS
VTTSESSSGI TSKPLSNASL FDPTSGGLIV DLRGISMTSV GVQGNICSSP

TYTRVEWKPD ICHLDSDTKI RRAILDLTDG TGDFVQEVLD LAAHKKPNMR
TYTRVEWKPD ICHLDSDAKI RRAIRDLTGG TGDCVQEVLD LAAHKKPNMR

VLEVDLTGGQ PRSLWLSGNE TSRITRAATS EFNYASDRPE SVLSAQDLYS
```

```
VLEVDLTGGQ PKSLWLSGNE TNRITRAATS ELNYASDRPE SVLSAQDLYS

DMSSGYTSRF TLLPITSQSF VAPPELCRSD LVLIRTSQLP SMETASILTR
EISTGYTSQF TLLPITSPSF VTPPEMSKSD LVLIRTSQLA SVETANILAS

NARCLLTEGG TIVLHVLDVS KYSKVGQESL REALSRGKFS KIRQAADGLF
NARCLLTEGG TMVLHVLNYG NDSKVEPESL RKALSQSKFS KIREAADGLF

VAEATDADTA YSQGKSLVVL HFSTSPVFSW SSAVITSLID KGWPITELTL
VAEATVADTA YTQPKSLVVL HFSTSHVSSW STAIITSLMD KGWPITELNL

EEGCRLTELP AKATILVMDE VNRPLFASME EYQLEAIQSI VQRDCSLLWV
EEGRRLTDLP AKATILVMDE VNQPIFASME QYQLETLQSI VQQECSLLWV

TQGSQMHVSS PLKAICHGVF RSVRSMDPNA RIVTLDVDSA AEDQLAKMAD
TQGSQMHVSS PLKAICHGVF RSVRSMEPNA RIVTLDVDSA AENQPEKTAD

ILHTVLLQVR VTPESLPADF EFVERGGLLY ISRLRPAQVD NESRSDGDKD
VLHTALLEVR ATPESLPADS EFVERGGLLY ISRLRPAQSD DDSRSDGDKD

GLQPVPVDLH STESTIGLVS GRPGILDTLH FAELGPGRLL VLGPEDIEVE
RHQPVPVDLH STEATVGLAL GRPGNLDTLH FAELGPGKVS VLDPEEVEVE

IFAASVDDGD YALAKNLDPE DSTRLGYGGA GIVTRTGDSI TDIRAGQRVA
IFAASVDYGD YAVAMNLVPG DPTRLGHGGA GIVTRTGDSI TDVRAGQRVA

LFHGGCVANR IVVARQVVFS VPDTMTFEDA ATLPTAFVPA IYSIYHLAQL
LFHGGSVANR VVVARKVVFP VPDAMTFEDA ATLPTAFVPA IYSIYHLARL

RQGQRVLIHS AANAVGIACV QLCQGLSCKP YVTVDSDEER KFLAEEVGVS
RQDQRVLIHS AANAVGIACV QLCQGLACKP YVTVNSDEEQ KFLAQEIGVS

SDHILLLNSE NFAREMQDSA QNHGFDVIIN TSQHHLPDQG WGVVSPGGVH
PDHILLSNSD NFAREMQHFA QNHGFDVIIN MSQEHLPDQG WDVVSPGGVH

VALGQTINDR SLLPMDYFTN NRSFCSLDIR TLPLDKLA-- ----------
VALCQSIGDR SLLPMDYFAN NRSFCSLDIR TLPLDKLARA CSQLSDLIHG

---------- ---------- ---------- RNVVISTGPD KDVRILVKPE
SCIKPVLPKA IFDYQKIQAA LQSCYGHDRR RSVVISNGPG KDVQILVKTA

KQQPRCTFAP EQTYLLVGKL KGVSGSLALH LARCGAKYLV IMSPKNSENS
EHQPGCTFAP DQPYLLVGKL EGVFGSLALH LARRGAKHLV IMCPRDAENS

ENISRSIRAM GCSLRFFEGD AASIDDMRRC YGQISGPIGG IVHGAAAQSF
ENISRSIRAM GCSLRFFEGD AASIDDMRRC YGQISASIGG IIHGAAAFTA

R------LMS HETYQATLAR SVLSAWNLHT VSLERDDSVP FFIMLSSTAG
RFPFLWPLMS HETYQESLAR NVQSAWNLHT VSLERNGSVP FFIMLSSTAG

VVGDEKQPHH AGSDVFHNAL ATYRCGLGLP STSINLGPIN DDALLPDSEK
VVGDEKQPHH AGSDVFRDAL AKYRCQLGLP STSINLGPIN DDALLLDSEE

TFKTLSSGVW FGVNEAVFRR IIDHSLSREH HGAQRHFELA SQAQIITGIA
TYKTLSSGVW LGINEAVFRR IIDHSLSQKH HGSKRHLELV SEAQIITGIA
```

```
           VPQPGSSDIL HDVRLLGLKL AQSGNSSSAA SGRDDSQNRE MQTFLLCARS
           VPQPESSPIL RDVRLLGLRL AQGGNASSAA SGRDDGQNRE MQTFLLCARS

           TNPDPAVLLS SAVGVLQAQF TKMLRLNELM DPAYPLNTYG MDSLAAAEPR
           TNPDPAVLLS SAVGVLQAQF TKMLRLNELL DPAYPLNTYG MDSLAAAEPR

           SWVRTAFGVQ LTTLDVVNAA SLVVLCQKII SRMGLGKEV
           SWVRTTFGVQ LTTLDVVNAA SLVVLCQKII SRMGLGKEV

Atr7, 320 aa
S40285     MSTLTIDPTS IPPLEGKTAV VTATPLVPGG ASGIGLAAAK IMLQKGATVY
S40288     MTTLTIDLSS IPPLEGKTAV VTATPLVPGG ASGIGLAAAK IMLQKGATVY

           ALDRQEPIEA VPGLKFRRCD VTSWSALREV FDEIQQVHLA FANAGICDKS
           ALDRQEPIEA VPGLKFRRCD VTSWSALRDV FDEMKQVHFA FANAGICDKS

           PESYYDDVCD NGNLQEPDYS MIDVNLKAVL NFVKLARHSM RRHQVQGSIV
           PESYYDDVYD NGNLQEPDYS MIDVNLKAVL NLVKLARHFM RKHQVQGSIV

           ITASSTGLVP EQSAPVYSST KFAVIGLVRT LRSVLIQENI TINAVAPFVT
           ITASSTGLVP EQSAPVYSST KFAVIGLVRT LRSVLIQENI TINAVAPFVT

           TTGMAPAEAM VPLKNLGVQT SPADFVGLAL VYSAVARQTR RVEAYGKETE
           TTGMAPAEAM VPLKNLGVQT SPADFVGLAL VYSAVARQTR RVEAYGKETA

           EDILEHGRWN GRVILTLGDK YTEVEEEFSK SRPLWTGGEV LQSIRLQQAV
           ESILEDGRWN GRVILTLGDK YTEVEEEFSK SRSLWTGGEV LEGIRLQQAV

           LDFRHGGVAI KSNRPSNQLN
           LDFRHGGVAI KSNRPSNQLN

Atr8, 625 aa
S40285     MAVEKVQAFE KVSIPTEKQP GSEDLGFDPA ELQKKYEAER NLRIQNGGVS
S40288     MAVEKVQALE NVSTPAEKEP GSKDLGFDPA ELQKKYEAER NLRIKNGGVS

           QYRSAWKSGF GYYLEDPNAD ANFSRDPISA RYDVVIMGGG FSGLLVAARL
           QYRSAWKSGF GYYLEDPNAD ANFSRDPIDA RYDAVIMGGG FSGLLVAARL

           VQQGITNFTI LDKSADFGGT WYWSRYPGAQ CDVDSTIYLP LLEEVGYIPK
           VQQGITNFAI LDKSADFGGT WYWSRYPGAQ CDVDSTIYLP LLEEVGYIPK

           EKYSFGPEIL EHAQRIAKHF GLYPKALFQT EVKTCHWSEE DSLWTVQTDR
           EKYSFGPEIL EHAQRIAKHF DLYPKALFQT EVKTCHWSEE DSLWTVQTDR

           GDNLRAQFIV SAFGISHMPK LPGISGIENF QGKSFHASRW DYNYTGGDST
           GDNLRAQFIV SAFGISHMPK LPGISGIENF QGKSFHASRW DYNYTGGDST

           GNMTKLADKR VGIIGTGATA IQVVPKLAES AKELYVFQRT PSSVDVRNNR
           GNMTRLADKR VGIIGTGATA IQVVPKLAES AKELYVFQRT PSSVDVRNNR

           PTDAEWAKTL RPGWQQERID NFYAITTGEN VTEDLIDDGW TEIFRLVAAP
           PTNAEWAKTL RPGWQQERID NFYAITTGEN VTEDLIDDGW TEIFRLVAAP

           FFASADIEQS LENRMEQVQI ADFKKMESVR ARVDSLVKDP ATAASLKPWY
```

```
           FFASADVEQS LENRMEQVQI ADFKKMESVR ARVDSLVKDP ATAASLKPWY

           NQFCKRPCFH DEYLQAFNHP NVTLVDTRGH GVDAVTTKGV LAQGKEYELD
           NQFCKRPCFH DEYLQAFNHP NVTLVDTRGH GVDGVTTKGV LAQGKEYELD

           CLIYSTGYEW YTEWEQRTRS QVYGRNGLTI TKKWSQGITT YHGWGVHGFP
           CLIYSTGYEW YTEWEQRTRS QVYGRNGLTI TKKWSQGITT YHGWGVHGFP

           NFMVLSSAQV NNVPNYTHMV GYLSRHLAYI VRTCKDRGIK SVEPTATAES
           NFMVLSSAQV NNVPNYTHMV GYLSRHLAYI IRTCKDRGIK SVEPTANAES

           KWVQQVVEQG AARRDQMKLC TPGYLNHEGD ITEKTDRLYS YNGSGDSKFQ
           KWVQQVVEQG AARRDQMKLC TPGYLNHEGD ITEKTDRLYS YNGSGDTKFQ

           IILDKWRDDG KLVGLSIDCA TEADL
           IILDKWREEG KLVGLSIDGA TEADF

Atr9, 253 aa
S40285     MPTIRGQSIL IIGGSSGIGA AVAKYACGDG VKVSVASSNK GRVEKALKKI
S40288     MPTIRGQSIL IIGGSSGIGA AVAEYACSDG VKVSIASSNR ARVEKAAKKI

           QALVPASEIL GFTVDLSQYD LESRLEKLFK EVVDATGGPL DHVVMTAGTG
           QASVPAAKIL GFTVDLGQYD LESRLEQLLK DVVDATGGPL DHVIVTAGTG

           NMVSLSEYTA KAFQESAPLH FIAPLMVGKV APRFMNRHWK SSITFTSGAF
           NMVSLSEYTA TAFQDCAPLH FIAPLMVGKV APRFMNQHWK SSIIFTSGAF

           GKKPAKGYCV IASAVGALDA ATRALALELA PIRVNAVSPG PTVTEMFGPP
           GKKPAKGYCV IASAVGALDA ATRALALELA PIRVNAVSPG PTVTEMFGPP

           SEALDKAVAA MGAQSLLGKL GRPEDVAEAY IYLMRDANTT GTIVDSNGGA
           SEALDKAVAA MGAQSLLGKL GRPEDVAEAY LYLMRDTNTT GTIVDSNGGA

           FLQ
           FLQ

Atr10, 276 aa
S40285     MARQSAEPPT DDGQSAKEIV VITGGNTGIG FEVARQLLCN YGNRFYVIIG
S40288     MARQSAEPPT NKGQSAKEIV VITGGNTGIG FEVARQLLCN HGDRFYVIIG

           SRTLGKGHTA VAALKQQGYE AVQAVQLDVT KEASIAAAAK IIGEQFGRID
           SRTLAKGHTA VAELKQQGYE AVQAVQLDVT EEASIKAATK TIGEQFGRID

           VLHVNAGVLL EPTDINAKPV PFSETIMETM RTNVAGAAAT VEGFTPLLSI
           VLHVNAGILL EPTDVNGKSV PFSETIMETM RTNVAGAAAT VEGFTPLLSN

           GSNPRVVFMT STAASAQLMH QYSSMTTAPA LSASKAAENI IMIYYYHKYP
           GSNPRVVFMT STAASAQLMH QYSSMTTAPA LSASKAAENI IMIYYYHKYP

           NWKVNACYPG YRDTAMMRRY NASSLSKAYR QPDPVEEGAY NAVRLSLLGK
           NWKVNASYPG YRDTAMMRRY NASSLSKAYR QPDPIEEGAY NAVRLSLLGK

           DGETGTFTEY KGVGEDGQRQ YSALPW
           DGETGTFTEY KGVSEDGQRQ YTTLPW
```

```
Atr11, 133 aa
S40285     MASRKATVQS IVESFNERDI DKMTAPVSKN FVYQLLPKSL ERGPMDVAGF
S40288     MASRKATVQS IVESFNERDI DKMTAPVSDN FVYQLLPKSL ERGPMDVAGF

           RGLFEATKSY FNNFKFEVVD TFEDSAADKM ILWANVTSDT HVGKFATEVM
           RGLFEATKSY FNNFKFEVVD TFEDSAADKM IVWANVTSDT HVGKFATEVM

           LIFYFDTTGK IYKWIEWIDS AVGKEFEQKL QGQ
           LIFYFDKAGK IYRWIEWIDS AVGKEFEQKL QGQ

Atr12, 391 aa
S40285     MASSVATLVK SLDKINAADF ESDEAARVNA IAAAQKMIHR LQSGVERGIE
S40288     MASSVASLVE SLDKINVDDF ESDEAARVNT IVAAQMMIHR LQSGVEWGIE

           LTHQRSTVFP IIDVFEDLGL WEAWASQGHE ISLEGLAQLS NTPLALNLLR
           LTHQRSTVFP IIDVFEDLGL WEAWASQGHE ISLEGLAQLR NTP-------

           RLCRLLTAAD IFEEKSEDCY TPTELSLYMG DKTKGSQVSQ GSAPGWVGSY
           RLCRLLTVAD IFEEKSEDCY MPTELSLYLR DKTKGSQVSQ GSAPGWVGSY

           TNLPIFLKET AYQEPLDPKK SAYSKTAGKS FWEELSQDPL QQENFGRFMS
           TNLPIFLKET SYQEPLDPKK SAYSKTAGKS FWEELSQDPL QQENFGRFMS

           SWAKFKVPWP AFYDTESLVR GAEPGMPILV DIGGNDGTDV ERFLAKHPGV
           SWAKFTVPWP AFYDTESLLR GAKPGKPILV DVGGNDGTDV ERFLAKHPGV

           AAGSLILQDR PAALKLAKVD QKIELMPHDF FTPQPVIGSR AYFFHAVLHD
           ATGSLILQDR PAALKLAKVD QKIDLMPHDF FTTQPVIGSR AYFFHAVLHD

           WDDAHALDIL RNTVPAMRKG YSKLLILDIA IPRTGASLIQ AAMDISMMSL
           WDDAHALDIL KNTVSAMQKG YSKLLILDIA IPRAGASLIQ AAIDISMMSL

           LSSLERPITT WEILLKKAGL KIVKFWPDPR RYETLIEAEL E
           LSSLERPITT WVTLLKKAGL KIVKLWPDPR RYETLIEAEI D

Atr13, 381 aa
S40285     MALEEISERL QVSDFPTLGM AANYDLRRHK FESLANDGSH EMRADVRRWV
S40288     MALDEISERL QVSDFPTLGM AANYDLRRHK FESLANDGSH EMRADVRRWV

           GNPSDFGGCN PINGHIIALT MPMIKPDRVK IAGYIYECWF LYSWDLTTTL
           GNPSDFGGCN PINGHIIALT MPMIKPERVK IAGYIYECWF LYSWDLTTTL

           TGADGFFHDD ILEGTNEGVS DTDAFGLGTA DQDAKARDGR KQIQAKMMYL
           TRVDGFFHDD ILEGTNEGVS DTDAFGLGTA DQDAKARDGR KQIQAKMMYR

           LETTDKACAK HLQKVWSNML VTTIQHKSRD FETLKEYIDF RIRDCGALFG
           LETTDKACAK HLQKVWSNML VTTIQHKSRD FETLEEYIDF RIRDCGALFG

           EGVMLFGMGL ALTEKDREDV ASTIYPCYAA LGLTNDYFSF DREWEEAKRT
           EGVMLFGMGL ALTEKDREDV ASTIYPCYAA LGLTNDYYSF DREWEEAQRS

           GEAKFSNAVR LFMDWQSTGA AAAKEVVRKA IIEYEREFLE LREKFVKANP
           GEAKFSNAVR LFMDWQSTDA ATAKEVVRKA IIEYEREFLE LREKFVKANP

           KAERLHKFLE AMVYQISGHV VWSINCPRYN PSFRYDPNSG VENQVLAERR
           EAERLHQFLE AMVYQISGHV VWSLNCPRYN PSFRYDPNSG VENQLLAERR
```

```
           GKSSSKKPSV MIEEIDEKSH LASETGPAMI A
           GKSSSKKPSA LIEDINEKSH LASESGPTMI A

Atr14, 461 aa
S40285     MNVADIAMDL FRGAKGETIS IFAIAKVTVT G--------- ----------
S40288     MNVSEIAMDL FRGGKGETIS IFAIAKITVT GVSRGLAKLV FGVANQANPA

           ----YVVYSV VSMIYNITLH PLASFPGPVF WGASRWPSIW RLFKGRLVHD
           NLGQYFVYSV VSMLYNITLH PLASFPGPVF WGASRWPSIW RLFKGRLVHD

           VHALHGQYGH VVRIAPNELA FSSAQAWKDI YGHKRGNNSM EEMPKFHKFY
           VHALHGQYGH VVRIAPNELA FSSAQAWKDI YGHKRGNNSM EEMPKFHKFY

           SGISKTPSIV SEPTRDGHRF IRRILSPAFS DKNLRELEPI VQGYISQFID
           SGISKTPSIV SEPTRDGHRF IRRILSPAFS DKNLRELEPI VQGYISQFIN

           QLRSHCEDST GSKVPLDLVS WYNSATFDIV GDLTFGRPFG SLEQGEEDPF
           QLRSRCKDST GSKVPLDLVS WYNSATFDIV GDLTFGRPFG SLEQGEEDPF

           IKDINHFAAV GGAMLIFTSH FPGRGILRFL ASLGKVFQNG QEKHVTKMEE
           IKDINHFAAV GGAMLIFTSH FPGRGILRFL ASLGKVFQNG QEKHVTKMEE

           SLVDRMKNKS SRPDIIDGLV KEKDGFQIDY DRVLENAAAI TMAGSETTAS
           SLVDRMKNKS SRPDIIDGLV KEKDGIQIDY DRVLENAAAI TMAGSETTAS

           QLSGLTALLL QNPNCLERLK KEVRSAFKSD KDITSTSSLV GVWQYSANHS
           QLSGLTALLL QNPTCLETLK KEVRSAFKSD KDITSTSSLV GVWQYSANHS

           PRNFTYPDEF RPDRWLDDRD QKEYEHDHGD AMQPFSVGPR DCPSQK----
           PRNFAYPDEF RPDRWLNDRD QKEYEHDRGD AMQPFSVGPR DCPSKKRSST

           ---------- -
           YNRTAVIQLR Y
```

Satratoxin cluster 1, 10 products

```
Sat1, 294 aa
S40293     MWASFPRPEA IPPGYVGETQ HQHRSIMLLN GKSRSWIFLY ERLPAPSHDR
S7711      MWASFPRPEA IPPGYVGETQ HQHRSIMLLN GKSRSWIFLY ERLPAPSHDR

           VKCIAEDVIE FADSFADWSI WNNTKLEDVV DHSTAGMSNL EEGIVKNFSH
           VKCIAEDVIE FADSFADWSI WNNTKLEDVV DHSTAGMSNL EEGIVKNFSH

           GRIVLVGDAC HKFTSNAGLG LNNGIQDIVA GCNSIRKVVT ESGFDLPDVK
           GRIVLVGDAC HKFTSNAGLG LNNGIQDIVA GCNSIRKVVT ESGFDLPDVK

           ALEATFKTYY EMRLGPFNDD FIHSKMMTRM QAWANTWYFL FTRYLFFIFS
           ALEATFKTYY EMRLGPFNDD FIHSKMMTRM QAWANTWYFL FTRYLFFIFS

           EWILFGFTML RRVCIGLVLT MHLAKSRLLA LLNGFIRSGY HSFIWISAIR
           EWILFRFTML RRVCTGLVLT MHLAKSRLVA LLNGFIRSGY HSFIWISAIR

           PCNDQLLNQL LAGAIQIEIL CSLNTWRISS CRAPLLLVFD QARG
           PCNDQLLNQL LAGAIQIEIL CSLNTWRISS CRAPLLLVFD QARG

Sat2, 354 aa
S40293     MPSLQVIRAA VAELPQGSPI VAAVAGGTTG IGSYLAKALA TTFASHGSKL
S7711      MPSLQVIRAA VAELPQGSSI VAAVAGGTTG IGSYLAKALA TTFASHGSKL

           RVYIVGRNAG RAKTVISECQ KISPGSDWRF IHATDLALIS EVDKSSAEII
           RVYIVGRNAG RAKTVISECQ KISPGSDWRF IHATDLALIS EVDKSSAEII

           KQETEAPFHG ELARLDLLYM THAIPILGHK RTTEEGLDAL ESTIYYSRIR
           KQETEAPFHG ELARLDLLYM THAIPILGHK RTTEEGLDAL ESTIYYSRIR

           FILQLLPLLT ASPRVAHVIS VYAGGMENGV KPDEEPIGFV PAEIYHFNTV
           FILQLLPLLT ASPRVAHVIS VYAGGMENGV KPDEEPIGFV PAEIYHFNTV

           RKYTTFMKTF VFEELAEKHA ERLSLIHIYP GLVDGPGFTQ MPRWFRVLFT
           RKYTTFMKTF VFEELAEKYA ERLSLIHIYP GLVDGPGFTQ MPRWFRVLFT

           LMKPLTSLYM TRSEDCGMVM AYLATSRFSA KGSGQDAPTS TDTLAPKSSL
           LMKPLTSLYM TRSEDCGMVM AYLATSRFSA KGSGQDAPTS TDTLAPKSSL

           GVVGGGAYSL GQRADSQTPQ IMFEKSRKPD TSKKAWDHTI RTLDDIAKKN
           GVVGGGAYSL GQRADSQTPQ IMFEKSRKPD TSKKAWDHTI RTLDDIAKKN

           ATIA
           ATIA

Sat3, 271 aa
S40293     MSEALIGGGA KKVYILSRRR DVLESAAAKH EGILIPIQCD VTSKASLQSA
S7711      MSEALIGGGA KKVYILSRRR DVLESAAAKH EGILIPIQCD VTSKASLQSA

           VDIVTKDSGY VNLLIANSGT LGPTNRLDHD LSIHELRKNV FDNVSFEDFN
           VDIVTKDSGY VNLLIANSGT LGPTNRLDHD LSIHELRKNV FDNVSFEDFN

           NTLSVNTTGA YFTMLAFLEL LDAGNKNALK GGFGGPSTEG GAPSIQSQVI
```

```
           NTLSVNTTGA YFTMLAFLEL LDAGNKNALK GGFGGPSTEG GAPSIQSQVI

           FTSSLGAYSR DRLSPPAYSA SKSALSHLAK HASTNLAKYG IRVNVLAPGL
           FTSSLGAYSR DRLSPPAYSA SKSALSHLAK HASTNLAKYG IRVNVLAPGL

           FPSEIATLMT ANRDPATENL GDRMFIPARK FGGAEEMGGT VLYLASRAGS
           FPSEIATLMT ANRDPATENL GDRMFIPARK FGGAEEMGGT VLYLASRAGS

           YCNGLILVND GGRLSVMLSE Y
           YCNGLILVND GGRLSVMLSE Y

Sat4, 268 aa
S40293     MNGIYALQQT FVKFSLLALY HRLFWVNRHF VRSVWLVGIV QGCWGIAILL
S7711      MNGIYALQQT FVKFSLLALY HRLFWVNRHF VRSVWLVGIV QGCWGIAILL

           VHIFLCTPME KIWTPWMVEG TCVDVNTLFA IYEALNSVLD FIVAGLAIWM
           VHIFLCTPME KIWTPWMVEG TCVDVNTLFA IYEALNSVLD FIVAGLAIWM

           LPSLQIRKST RWHLAGLFVL GAFSGFIGII KIVEAYDSAQ RNFQAVIWNV
           LPSLQIRKST RWHLAGLFVL GAFSGFIGII KIVEAYDSAQ RNFQAVIWNV

           VQMSISIICC CAPIYRSILP KMGMSSIPSW ASWSLRGSSR RSKAVASTAD
           VQMSISIICC CAPIYRSILP KMGMSSIPSW ASWSLRGSSR RSKAVASTAD

           GTSKFSMRSY QGEGKAGGTS VSGNWINLDG SSQRALAWVD AESHGKDQST
           GTSKFSMRSY QGEGKAGGTS VSGNWINLDG SSQRALAWVD AESHGKDQST

           YQDIPMGRMK VERSVEVI
           YQDIPMGRMK VERSVEVI

Sat5, 463 aa
S40293     MSTMAKSPEA NNLHQDVIAQ FPILNGYTHT VGAFSQPLNV SRLFIIDEIQ
S7711      MSTMAKSPEA NNLHQDVIAQ FPILNGYTHT VGAFSQPLNV SRLFIIDEIQ

           TAYDELRVQI PWLAHQVVVV DAGPGKSGYI TTAPWPSSAP PNDVTYEEKD
           TAYDELRVQI PWLAHQVVVV DAGPGKSGYI TTAPWPSSAP PNDVTYEEKD

           DAFPSLNTLI KSGGSFLATK DLVGYPGLPE PHGLHPTPVA TIRLVFITGG
           DAFPSLNTLI KSGGSFLATK DLVGYPGLPE PHGLHPTPVA TIRLVFITGG

           VLVVLSTHHN IVDGIGLMQM WDYLDILMGG GAISRQDARS ANADRARVLP
           VLVVLSTHHN IVDGIGLMQM WDYLDILMGG GAISRQDARS ANADRARVLP

           LIAPGEPVKD YSHLIRPDPW PLPPPPKTEW RLFKMHPWAL AEIRSRARDG
           LIAPGEPVKD YSHLIRPNPW PLPPPPKTEW RLFKMHPWAL AEIRSRARDG

           TDQRASARPA SSDDALTAFC WQRVSAMRLA SGRVTGDQVS KFGRAVNGRS
           TDQRASARPA SSDDALTAFC WQRVSAMRLA SGRVTGDQVS KFGRAVNGRS

           AMGLDSSYLF HMMLHTETRL PIEQIARSTL AELSTQLRKD LDAARTEWSV
           AMGLDSSYLF HMMLHTETRL PIEQIARSTL AELSTQLRKD LDAARTEWSV

           RSYATFLAGV ADKTRLLYGG ITNPQTDLGG TSTMHWASRR PIRLGLLGDC
           RSYATFLAGV ADKTRLLYGG ITNPQTDLGG TSTMHWASRR PIRLGLLGDC
```

```
           HLIRKPEGMP LPGCLYFMPS GGTSGVVQLL LCLPKEELDA LQEDAEWKHY
           HLIRKPEGMP LPGCLYFMPS GGTSGVVQLL LCLPKEELDA LQEDAEWKHY

           TESGGRRVDG PRL
           TESGGRRVDG PRL

Sat6, 485 aa
S40293     MPQDPNTTLQ MSSSKPSLSD LSVSADPVLG KADNQVRDSL ALPSIEGGEE
S7711      MPQDPNTTLQ MSSSKPSLSD LSVSADPVLG KADNQVRDSL ALPSIEGGEE

           GVMRPLAWLF GLCAIQQASG ATLLRNDVST VEPLPPTQDP WYRAPPGFEK
           GVMRPLAWLF GLCAIQQASG ATLLRNDVST VEPLPPTQDP WYRAPPGFEK

           KQPGDVLRIR QAPGNLTTVV SNSSAAFHIL FRTTNARSEP AWAVTTLFLP
           KQPGDVLRIR QAPGNLTTVV SNSSAAFHIL FRTTNARSEP AWAVTTLFLP

           KKLYRAPSRN AALLSFQLAD NSANPDSAPS LGLYWRLAQD NPMLGLRSDT
           KKLYRAPSRN AALLSFQLAD NSANPDSAPS LGLYWRLAQD NPMLGLRSDT

           SFISNLLSEG WLVNIPDQSG PEAAFGASRQ AGHATIDAIR AIQHLCSLTG
           SFISNLLSEG WLVNIPDQSG PEAAFGASRQ AGHATIDAIR AIQHLCSLTG

           ATGINAAIWG YSGGTFATGA AAELMPTYAP NINIVGAVLG GMVTDVSGGF
           ATGINAAIWG YSGGTFATGA AAELMPTYAP NINIVGAVLG GMVTDVSGGF

           DSLNRSPIAA TIIATLLGVT AQFPEERAYL ESRLVPETRD EFMSVLDINV
           DSLNRSPIAA TIIATLLGVT AQFPEERAYL ESRLVPETRD EFMSVLDINV

           FDALVHFAGR DIYAFFIDGA ADIEAPILQN LFEAQSRIGF GDIPPMPMFI
           FDALVHFAGR DIYAFFIDGA ADIEAPILQN LFEAQSRIGF GDIPPMPMFI

           YKAIADEVVP IGPTDVTVQR WCDGGADITY ERNTVGGHIA EIENGKPRAI
           YKAIADEVVP IGPTDVTVQR WCDGGADITY ERNTVGGHIA EIENGKPRAI

           QWLWSIFDES YSAQSPECRI RDVTVEVPVQ VVGRV
           QWLWSIFDES YSAQSPECRI RDVTVEVPVQ VVGRV

Sat7, 470 aa
S40293     MSTSTSEPGA IAGLPLGAEV RTDGDATGLS VAIVGGGIVG IALALGLVER
S7711      MSTSTSEPGA IAGLPLGAEV RTDGDATGLS VAIVGGGIVG IALALGLVER

           GVRVSVYERA QELPEIGVGF AFNGAARKSM ARLSPLVIAA LERVANENEQ
           GVRVSVYERA QELPEIGVGF AFNGAARKSM ARLSPLVMAA VERVANENEQ

           AYDNYWDGYT STAEDDESST ASKRGKLLFR MPNSNMAWWS CLRSQFLNEM
           AYDNYWDGYT STAEDDESST ASKRGKLLFR MPNSNMAWWS CLRSQFLNEM

           LQALPPGTVT FGKELDSYND PFDTSDPVRL RFTDGTTAAA NVLIGSDGLR
           LQALPPGTVT FGKELDSYDD PFDTSDPVRL RFTDGTTAAA NVLIGSDGLR

           SRVRQQLFAT SHPEVCNPTY THKTCYRAVI PMAAAESAMG LSKPHNHCMH
           SRVRQQLFAT SHPEVCNPTY THKTCYRAVI PMAAAESAMG LSKPHNHCMH

           TGPRAHVLSY PIAQHKLVNV VLFVTHDEPW VDGTGDEAIS VPRMTRPGDK
           TGPRAHVLSY PIAQHKLVNV VLFVTHDEPW VDGTGDEAIS VPRMTRPGDK
```

```
            KVLQNRLADW RPEVRNLVAQ LPDAPTAWGI FDTAEHPVPF YAAGRVGLVG
            KVLQNRLADW RPEVRNLVAQ LPDAPTAWGI FDTAEHPVPF YAAGRVGLVG

            DAAHASSPHH GAGAGFGVED ALALAVALGM ATEKKQSVAA ALQAFNDVRY
            DAAHASSPHH GAGAGFGVED ALALAVALGM ATEKKQSVAA ALQAFNDVRY

            DRTQWLIRSS KETGDIYEWK HFGVGGDPVK IRAELEGRQK TIWDYDVDAM
            DRTQWLIRSS KETGDIYEWK HFGVGGDPVK IRAELEGRQK TIWDYDVDAM

            AEEVKTRYEA RVTSETTAHQ
            AEEVKTRYEA RVTSETTAHQ

Sat8, 2602 aa
S40293      MATTLLLFGP QAASMSKQSI TQLQVALRDQ EWAFDALSNV QPIIQRASTS
S7711       MATTLLLFGP QAASMSKQSI TQLQVALRDQ EWAFDALSNV QPIIQRASTS

            ISGLDQISLD ERLADLTRWL KHGPKDQEEL AEIPNIMLAP LTTLSHLVQY
            ISGLDQISLD ERLADLTRWL KHGPKDQEEL AEIPNIMLAP LTTLSHLVQY

            RRYIERHYPN ESDAHAALLQ QKPVATLGFC NGLLAAFATT SSATLNDWER
            RRYIERHYPN ESDAHAALLQ QKPVATLGFC NGLLAAFATT SSATLNDWER

            YAAVATRLAL LVGAVIDAAD ELQPHGPAAS YGVSWRDIDG ARQLEQILSP
            YAAVATRLAL LVGAVIDAAD ELQPHGPAAS YGVSWRDIDG ARQLEQILSP

            FPGDAYVSVW YDRSRATVTV SKHLVRTVLH LVEAAGMAVV PVRLRGRYHS
            FPGDAYVSVW YDRSRATVTV SKHLVRTVLH LVEAAGMAVV PVRLRGRYHS

            RQHAEVAEAL IRLCDAEPDL LALPDARNLC LPTYSNVGHG EVVREGRLHE
            RQHAEVAEAL IRLCDAEPDL LALPDARNLC LPTYSNVGHG EVVREGRLHE

            IALQAMLVQQ CDWYSTLSGI TDESQVQVVC LSEVSTLPPS LTFKLKPQME
            IALQAMLVQQ CDWYSTLSGI TDESQVQVVC LSEVSTLPPS LTFKLKPQME

            YFAPLEEKTA PKDNFSGRAD GGSQFSFSML ENSTSPPSPA ATSSNSHCEY
            YFAPLEEKTA PKDNFSGRAD GGSQFSFSML ENSTSPPSPA ATSSNSHCEY

            SVDPRDIAIV GMSVKVAGAD DVVEYESILR GGVSQHQQVR KNRVPFGYNS
            SVDPRDIAIV GMSVKVAGAD DVVEYESILR GGVSQHQQVR KNRVPFGYNS

            FRPEEPGHKW YGNFVRDVDA FDHKFFRKSS RESAAMDPQQ RLVLQAAYQA
            FRPEEPGHKW YGNFVRDVDA FDHKFFRKSS RESAAMDPQQ RLVLQAAYQA

            VEQSGYYASG TEPDQHIGVY LGTCATDYEQ NANCHAPGAF TVTGLLRGFI
            VEQSGYYASG TEPDQHIGVY LGTCATDYEQ NANCHAPGAF TVTGLLRGFI

            AGRISHFFGW TGPAMTYDTA CSGSAVAIHS AVQALVSGEC SAALAGGVNT
            AGRISHFFGW TGPAMTYDTA CSGSAVAIHS AVQALVSGEC SAALAGGVNT

            IGNEVWFQNL AGAQFLSPTG QCKPFDDAAD GYCRGEGIAC VVLKPMAKAI
            IGNEVWFQNL AGAQFLSPTG QCKPFDDAAD GYCRGEGIAC VVLKPMAKAI

            ADGNQIFGRI ASSAVHQSVN CTPLFVPNVP SLSRLFGDVM RQARLEPHDI
            ADGNQIFGRI ASSAVHQSVN CTPLFVPNVP SLSRLFGDVM RQARLEPHDI
```

SFVEAHGTGT PVGDPVEYES IRAILGGPLR DKPLSLGSAK GLVGHTESTS
SFVEAHGTGT PVGDPVEYES IRAILGGPLR DKPLSLGSAK GLVGHTESTS

GVVSLVKVLL MMQSGFIPPQ ASYSKLSHRI APSASAMIQV STTLQPWTDS
GVVSLVKVLL MMQSGFIPPQ ASYSKLSHRI APSASAMIQV STTLQPWTDS

YKAALINNYG ASGSNAAMVV TQGPAQTARS PRGEADGAHL PFWIPGFDSA
YKAALINNYG ASGSNAAMVV TQGPAQTARS PRGEADGAHL PFWIPGFDSA

RIAAYCARLS AFIEANRSTI HLADIAYNIS RQSNRTLSHA LLFRCNSIDS
RIAAYCARLS AFIEANRSTI HLADIAYNIS RQSNRTLSHA LLFRCNSIDS

LVGQLSSAAA PQTVQVKPSR PVILCFGGQM STFVGLNREI YDSSPILRDH
LVGQLSSAAA PQTVQVKPSR PVILCFGGQM STFVGLNREI YDSSPILRDH

LSQCDAAIRA LGFGSIFPSI FATIPIEDTV LLQTVLFSFQ YACAKSWIDC
LSQCDAAIRA LGFGSIFPSI FATIPIEDTV LLQTVLFSFQ YACAKSWIDC

GVRPTAVVGH SFGEITALCI AEVLSLDDTI KLVTRRAKVV RDSWGADRGV
GVRPTAVVGH SFGEITALCI AEVLSLDDTI KLVTRRAKVV RDSWGADRGV

MMAVEGEVDQ VERLLEEANK DLDTHSPASI ACYNGLRNFT LAGSTLAMER
MMAVEGEVDQ VERLLEEANK DLDTHSPASI ACYNGLRNFT LAGSTLAMER

VALALSSSAY ASIRGKKLNV TNAFHSALVD PLLQELEQAG SDLTFNKARI
VALALSSSAY ASIRGKKLNV TNAFHSALVD PLLQELEQAG SDLTFNKARI

KVERATKEST TGEPCAPKFV GEHMRNPVFF RQAAQRLARD NPSAVWIEAG
KVERATKEST TGEPCAPKFV GEHMRNPVFF RQAAQRLARD NPSAVWIEAG

SATKITAMAR RSLDSNAESH FQGITVTGED GLDKLTEATL SLWKQGLNVA
SATKITAMAR RSLDSNAESH FQGITVTGED GLDKLTEATL SLWKQGLNVA

FWAHHGPQAT RDHQPLLLPP YQFEKSRHWL DVKAPPVMLA DTAQGDNGPL
FWAHHGPQAT RDHQPLLLPP YQFEKSRHWL DVKAPPVMLA DTAQGDNGPL

FGLLTFVGFQ DAEERRARFK INTESERYKS LVIPHIIART APICPATLEY
FGLLTFVGFQ DAEERRARFK INTESERYKS LVIPHIIART APICPATLEY

SLAIQALLTL RDHKHFESRD MHPVIRDMRN DAPLCLNSDQ STWLDLEANK
SLAIQALLTL RDHKHFESRD MHPVIRDMRN DAPLCLNSDQ STWLDLEANK

TSPRSLVWKV FTAPVSRQLD SHNDSDETLC AQGKLDLLSS SETTEFAQYE
TSPRSLVWKV FTAPVSRQLD SHNDSDETLC AQGKLDLLSS SETTEFAQYE

QLATYDACVS LLQDDDGDVS GLQGRSVYRS LADVVDYGVH YQKVRRVSGR
QLATYDACVS LLQDDDGDVS GLQGRSVYRS LADVVDYGVH YQKVRRVSGR

NSESAGIVRG ASGGNWLSDL PIIDSFSQVA GVWVNCLADR TPGSDDLFLA
NSESAGIVRG ASGGNWLSDL PIIDSFSQVA GVWVNCLADR TPGSDDLFLA

TGCETIMTSP TFLHADRGGK SWHVWAKHHR ESERSYRTDV LVFDATNGQL
TGCETIMTSP TFLHADRGGK SWHVWAKHHR ESERSYRTDV LVFDATNGQL

ADVFLGIAYT RIPRHSMTRL LSKLSEPSAL QAQAALPSST GHEGLTAKTA
ADVFLGIAYT RIPRHSMTRL LSKLSEPSAL QAQAALPSST GHEGLTAKTA

SSQRLGQDTL KQTVGQIIAS LSGVEAAQIT DESALADIGI DSLAGMELAR
SSQRLGQDTL KQTVGQIIAS LSGVEAAQIT DESALADIGI DSLAGMELAR

DIESVLGCKL DLEELLFTHD TFGAFVRYIS KVVNGEDDLG TPSHSDNDSH
DIESVLGCKL DLEELLFTHD TFGAFVRYIS KVVNGEDDLG TPSHSDNDSH

VTGTTATPNS SSASSDTHHG NSKLQIAVAQ SSQADASSSS PLPPQHVISS
VTGTTATPNS SSASSDTHHG NSKLQIAVAQ SSQADASSSS PLPPQHVISS

FEQVKLSTDQ RIREEKADNT DDIIVSRSNL LCVALVVEAF EQLGCPLRGV
FEQVKLSTDQ RIREEKADNT DDIIVSRSNL LCVALVVEAF EQLGCPLRGV

PAGEALKRIQ HAPQHARLVD WLYRFLEDEA RLINTEGTLI LRTSNGAPNK
PAGEALKRIQ HAPQHARLVD WLYRFLEDEA RLINTEGTLI LRTSNGAPNK

TSQAIFQDLE HANDRWIESH RLANYAGKNL ADVVSGKKEG IHVLFGSAEG
TSQAIFQDLE HANDRWIESH RLANYAGKNL ADVVSGKKEG IHVLFGSAEG

RELVRGLYSG LPFNCLFYKQ VRDTISLIVE KVKDDFQGPL RILEMGAGTG
RELVRGLYSG LPFNCLFYKQ VRDTISLIVE KVKDDFQGPL RILEMGAGTG

GTTQVLAPFL ATLDIPVEYT MTDLSPYMVA QAQRSFGTKY PFMHFAVHDI
GTTQVLAPFL ATLDIPVEYT MTDLSPYMVA QAQRSFGTKY PFMHFAVHDI

EKPPAESLLG TQHIIVASNA VHATANLADS AANIRSTLRP DGILLLVEMT
EKPPAESLLG TQHIIVASNA VHATANLADS AANIRSTLRP DGILLLVEMT

ESLPFVDIVF GLLEGWWRFA DGREHAIVPA EQWEARLRDA GYGHVDWTDG
ESLPFVDIVF GLLEGWWRFA DGREHAIVPA EQWEARLRDA GYGHVDWTDG

VFSENRLQKV ILAMASELPD GLPVSSGVPE PVQPALEVTT TVAREANAEA
VFSENRLQKV ILAMASELPD GLPVSSGVPE PVQPALEVTT TVAREANAEA

YVTQYSADFT YDGESSGNIE AHEAQDSRIV VVTGATGSLG SHMVASFAES
YVTQYSADFT YDGESSGNIE AHEAQDSRIV VVTGATGSLG SHMVASFAES

PSVTSVVCIN RRNSGKATAL ERQQEAFTSR GITLSPDAFG KLRVFATDTA
PSVTSVVCIN RRNSGKATAL ERQQEAFTSR GITLSPDAFG KLRVFATDTA

QPQLGLPLEE YEWLVTHATH IVHNAWPMSA SRPIQAFQPQ FKTMSRLLDL
QPQLGLPLEE YEWLVTHATH IVHNAWPMSA SRPIQAFQPQ FKTMSRLLDL

AAAIAQQSTS RCVVFQLISS IGVVGSAPMI DTRVPERRVP VSYTLPNGYC
AAAIAQQSTS RFVVFQLISS IGVVGSAPMI DTRVPERRVP VSYTLPNGYC

EAKWVCEQLL NETLHQYPER FRAMVVRPGQ IAGSSVNGVW NPVEHFPALV
EAKWVCEQLL NETLHQYPER FRAMVVRPGQ IAGSSVNGVW NPVEHFPALV

RSSQALRAFP ALGGTLQWIP VDVAAGTVAD LALNQQAGEP VYHIDNPVGQ
RSSQALRAFP ALGGTLQWIP VDVAAGTVAD LALNQQAGEP VYHIDNPVGQ

SWSDMVPILA DELNIPGERI IPLGEWVRKV KRSSLLETEN PASRLPDFFE

```
           SWSNMVPILA DELNIPGERI IPLGEWVRKV KRSSLLETEN PASRLPDFFE

           QHFERMSCGG LILDVALATK RSGTLAAQGA VSADTARKYI QTWKDMKFLD
           QHFERMSCGG LILDVALATK RSGTLAAQGA VSADTARKYI QTWKDMKFLD

           RY
           RY

Sat9, 208 aa
S40293     MASAALFYDS RATIARRLTL HCQYMARAVL FAQCRSAETM LTFILDLSWL
S7711      MASAALFYDS RATIARRLTL HCQYMARAVL FAQCRSAETM LTFILDLSWL

           FPDENGMRDN TCCYIVATIT MALDLLRDRV LAVAVSLDVE LLPYVYIVVT
           FPDENGMRDN TCCYIVATIT MALDLLRDRV LAVAVSLDVE LLPYVYIVVT

           HTRLHLLASL LNYPPVHLDV RRMVSEAALH SACEVLRAAV RGEDQLKYIP
           HTRLHLLASL LNYPPVHLDV RRMVSEAALH SACEVLRAAV RGEDQLKYIP

           NNLVIMICYA SCISHVPKVR RLVAQLLQFS WTMQGLFTHC VSLLVATLAA
           NNLVIMICYA SCISHVPKVR RLVAQLLQFS WTKQGLFTHC VSLLVATLAA

           KSGSQNSR
           KSGSQNSR

Sat10, 1930 aa
S40293     MSIQFFSYDV GGLIVKE--- ---------- ---------- -ALRIASSDK
S7711      MSIQFFSYDV GGLIVKEVRL LILWRRSALP NGRTLFARMT QALRIASSDK

           KYRSIFDRTS LLAFFATPHR STQHQSPESV ALALLNQCYC GLISPWISIF
           KYRSIFDRTS LLAFFATPHR STQHQSPESV ALALLNQCYC GLISPWISIF

           PPNFSKVVAR SEVEFRPPTR AHILNVFQDP GPASPIKDSV VVHKSCAVLG
           PPNFSKVVAR SEVEFRPPTR AHILNVFQDP GPASPIKDSV VVHKSCAVLG

           VDGEVLIGLD CSHYTLARLL KARDKRYLLR QASHAALRHG KAFQQVVGLF
           VDGEVLIGLD CSHYTLARLL KARDKRYLLR QASHAALRHG KAFQQVVGLF

           FLDSIRTFDA EGPKFECRKV VSQLASLSQL QSWRQGSVDS RVLWFGTPAM
           FLDSIRTFDA EGPKFECRKV VSQLASLPQL QSWRQGSVDS RVLWFGTPAM

           LDPTSLFRTL RSQIQEENQL GDPIFIRVDS TLHRHNELSE PQILASMCQQ
           LDPTSLFRTL RSQIQEENQL GDPIFIRVDS TLHRHNELSE PQILASMCQQ

           ILRQQPQLTS ALQDLLLNVE DAAVGSRDCW KQRTLWNCLL VLLYHPKDAE
           ILRQQPQLTS ALQDLLLNVE DAAVGSRDCW KQRTLWNCLL VLLYHPKDAE

           TFCFIDATSS LQTKNLAGQL DSVMKESEMP LRLIVSCRST AKQTPEASTQ
           TFCFIDATSS LQTKNLAGQL ESVMKESEMP LRLIVSCRST AKQTPEASTQ

           VDVDLSDAGF DEPLRQDLEE WIRESLECGL TDSTLREALL TQILSSGDFH
           VDVDLSDAGF DEPLRQDLEE WIRESLECGL TDSTLREALL TQILSSGDFH

           LARHALEFFA STGSWLTKWS VPSVWAMLSK QSAAELFIED NIRRHGQWLL
           LARHALEFFA STGSWLTKWS VPSVWAMLSK QSAAELFIED NIRRHGQWLL
```

```
IPMLWALEAF EPMQVDEIDV ALMLEDDGVA GQVEDFINLL PGISTIRRGV
IPMLWALEAF EPMQVDEIDV ALMLEDNGVA GQVEDFINLL PGISTIRRGV

FVIADHARPA FDYLWGKYFA TYQHHVYLAK SCVAALRHHL RVTPAIPQPR
FVIADHVRPA FDYLWGKYFA TYQHHVYLAK SCVAALRHHL RVTPAIPQLR

EDKSSDAAAR LCTYAAMNWV RHFSLQTNSE TTKYVPHPTI ITANETDPGS
EDKSSDAAAR LCTYAAMNWV RHFSLQRNSE TTKYVPHPTI ITANETDPGS

PNEHAGVSEA FLEDPELSRL WITHLRRALD LDGLDEELGH LILPETLGSR
PNEHAGVSEA FLEDPELSRL WITHLRRALD LDGLDEELGH LILPETLGSR

LGICTGWALR ISRQLASMRL SSERDVTNSL LVIGSETDDL AMVQSCLSAD
LGICTGWAIR ISRQLASMRL SSERDVTNSL LVIGSETDDL AMVQSCLSAD

PSPTEHTLGY ALAAASDPIK EKLLQQVGEP SDEFLYRALL SSICFGNVSV
PSPTEHTLGY ALAAASDPIK EKLLQQVGEP SDEFLYRALL SSICFGNVSV

TKDLLVRITD KVRVAQVTPL EGSKLQWPQA NESSESAETF RHTPLGVATA
TEDLLVRITD KVRVAQVTPL EGSKLQWPQA NESSESAETF RHTPLGVATA

YGDADVIDLL INHDISWWDL EERSPPPGTW NALHHAALGG QRNIMCKLLL
YGDADVIDLL INHDISWWDL EERSPPPGTW NALHHAALGG QRNIMCKLLL

QQRKGVLNSS ARIPNTVTES GNTPLILAAS RGFHKIVALL LEDGSMRGYG
QQRKGVLNSS ARIPNTVTES GNTPLILAAS RGFHKIVALL LEDGSMRGYG

VDVNIQNEQR SSALLAAARY GFSQTLEMLL TYEGIDYSKT DSNGASILHL
VDVNIQNEQR SSALLAAARY GFSQTLEMLL TYEGIDYSKT DSNGASILHL

ALVNDREAAA LQILAHKDIF SNEMEYQEAN MEANPVNKFE DDDSFSESSV
ALVNDREAAA LQILAHKDIF SNEMEYQEAN MEANSVNNFE DDDSFSESSV

DTTDVVYTVP ARPRISLHQK DGSGLTSLTI AIWRNLKSIV EILIAMDADA
DTTDVVYTVP ARPRISLHQK DGSGLTSLTI AIWRNLKSIV EILIAMDADA

NGPEGEFEAP LVAAAEVGSF ELFTMFTKIG ATKTEAALNT ISTGRTRPLH
NGPEGEFEAP LVAAAEVGSF ELFTMFTKIG ATKTEAALNT ISTGRTRPLH

AACAMGHLEV VRELLKDSVT QLSHTDSNQR TPLCAAISRD QNHVISVLLD
AACAMGHLEV VRELLKDSVT QLSHTDSNQR TPLCAAISRD QNHVISVLLD

RETETGLQEG LWEAARSGKA HILDQLLRRG AEINAQDEYG NTALQWASYY
RETETGLQEG LWEAARSGKA HILDQLLRRG AEINAQDEYG NTALQWASYY

NKPRCVERLL LGGARLDLLD CDNVNALGDA ARSGSAEPLK LLVDVGVDVN
NKPRCVERLL LGGARLDLLD CDNVNALGDA ARSGSAEPLK LLVDVGVDVN

AEAGGDTALC RAIWAEEVEC VSVLLQGGAK FILSSAQSRF ENLLTFAVQV
AEAGGDTALC RAIWAEEVEC VSVLLQGGAK FILSSAQSRF ENLLTFAVQV

SSPEILRLLL KAPEERDLAP TLRSACAMQS TSQLEVLLEF YDPAKVDLGS
SSPEILRLLL KAPEERDLAP TLRSACAMQS TSQLEVLLEF YDPAKVDLGS

GWTILHLAAV HGTLAGLTKV LDHATGRAAL NYGPKKVGTP FEMAAFSSKE
```

GWTILHLAAV HGTLAGLTKV LDHATGRAAL NYGPKKVGTP FEMAAFSSKE

SLSKVEHLYS NAALPGLVQP SSRFGTALNA ASYRLNDPVV VYLLQKMQLE
SLSKVEHLYS NAALPGLVQP SSRFGTALNA ASYRLNDPVV VYLLQKMQLE

DINASGARYG NAIQNMLASA WMDTERSLKL LGILLEAGVS LTPTSADRHG
DINASGARYG NAIQNMLASA WMDTERSLKL LGILLEAGVS LTPTSADRHG

TALHTAALFS PKPLVEKVLE TSRMLADERD GEGRLPIHLS ALQEEWASMM
TALHTAALFS PKPLVEKVIE TSRMLADERD GEGRLPIHLS ALQEEWASMM

LLSTTTSTIR SVDKMGRNAV HLAAAAGARS VLEKIFEVEE NEDLLLEADF
LLSTTTSTIR SVDKMGRNAV HLAAAAGARS VLEKIFEVEE NEDLLLEADF

DGWTPFHWAC RGEDDDCARF LIETARKIFD SKWDSMKHEL VTTDEKTWTP
DGWTPFHWAC RGEDDDCARF LIEKARKIFD SKWDSMKHEL VTTDEKTWTP

LDVARFHQRR EVELLLSLGM TTSDAENWMP DQSQNLGSYC DSCACQIWHE
LDVARFHQRR EVELLLSLGM TTSDAENWMP DQSQNLGSYC DSCACQIWHE

EHHSSALHCK RLYLRFNGWS RMCQLAPTAK RKYTPTLLGY KTCFKVADNL
EHHSSALHCK KLYLRFNGWS RMCQLAPTAK RKYTPTLLGY KTCFKVADNL

MKARLAKVLV PHSQGPLQRK GRGKGEGEED EEGIATVENY PVSQSCGVEM
VKARLAKVLV PHSQGPLQRK GRGKGEGEED EEGIATVENY PVSQSCGVEM

IRAFVIDDRR IQPDATRNLN HEVSVGSEDQ TQMSLGTGRL FVQSVDSPLE
IRAFVIDDRR IQPDATRNLN HEVSVGSEDQ TQMSLGTGRL FVQSVDSPLE

MVMDVAAIFL SSPKSAHKQD ETTRAFATAK
MVMDVAAIFL SSPKSAHKQD ETTRAFATAK

Satratoxin cluster 2, 6 products

```
Sat11, 526 aa
S40293     MTIPTSFEML KGMHIKDAFL LIAMLYLGYL LCICFYNIYL HPLRHIPGSK
S7711      MTIPTSFEML KGMHIKDAFL LIAMLYLGYL LCICFYNIYL HPLRHIPGSK

           LAVMGPYLEF YHEVIRQGQY LWEIEKMHDK YGPIVRVNER EIHIRDSSYY
           LAVMGPYLEF YHEVIRQGQY LWEIEKMHDK YGPIVRVNER EIHIRDSSYY

           HTIYAAGSRK TNKDAATVGA FDVPNSTAAT VDHDQHRARR GYLNPYFSKR
           HTIYAAGSRK TNKDAATVGA FDVPNSTAAT VDHDQHRARR GYLNPYFSKR

           SLANLEPTIH ERISKLLNRL EQHQNNDDII TLDGIFSALT ADVICSRFYG
           SLANLEPTIH ERISKLLNRL EQHQNNDDII TLDGIFSALT ADVICSRFYG

           KHFDYLSIPD YHFVVRDGFQ GLTKLYHLGR FLPTLVTILK CLPQQIIRLI
           KHFDYLSIPD YHFVVRDGFQ GLTKLYHLGR FLPTLVTILK CLPQQIIRLI

           LPNLADLIVM RDEIQANGIA QFTSSQTADS KASALVGALG DKNIPPHERT
           LPNLADLIVM RDEIQANGIA QFTSSQTADS KASALVGALG DKNIPPHERT

           VARLLDEGTV FLFAGTETTS RTLAVTMFYL LTNPDCLKKL RAELDTLPST
           VARLLDEGTV FLFAGTETTS RTLAVTMFYL LTNPDCLKKL RAELDTLPST

           EDYQHSLSTL ESLPYLSGVV HEGLRLAFGP ITRSARVPMN VDLQYKEYTI
           EDYQHSLSTL ESLPYLSGVV HEGLRLAFGP ITRSARVPMN VDLQYKEYTI

           PAGTPLSMST YFVHTDKELY PEPEKFKPER WIQAAEENIP LKKFLTNFSQ
           PAGTPLSMST YFVHTDKELY PEPEKFKPER WIQAAEENIP LKKFLTNFSQ

           GSRQCIGISM SFAEMYLTIS RVARAYNFEL YETTAADLDM TYARIVPYPK
           GSRQCIGISM SFAEMYLTIS RVARAYNFEL YETTAADLDM TYARIVPYPK

           EIPGKTEGLG EIRVKIVGKN HSQIEE
           EIPGKTEGLG EIRVKIVGKN HSQIEE

Sat12, 573 aa
S40293     ---------- ---MSNASSR EASITSRSSS TSGNNSLPED RGAVVQLPTL
S7711      MLDDDCSPTS SSEMSNASSR EASITSRSSS TSGNNSLPED RGAVVQLPTL

           NPSDYRWHPF PGDSSVLQRK AIGVEALVGI RDANSRGEYD FYNNIVLRVG
           NPSDYRWHPF PGDSSVLQRK AIGVEALVGI RDANSRGEYD FYNNIVLRVG

           NALELTLTRL KRAFVKAMLD ARFENPSIAC YGVWGQNKEQ YLPHIQYKSF
           NALELTLTRL KRAFVKAMLD ARFENPSIAC YGVWGQNKEQ YLPHIQYKSF

           KSQSEALAWA NNCIIIQATS LTGSELRAER LKKRRAQAVP QPSNPLDIII
           KSQSEALAWA NNCIIIQATS LTGSELRAER LKKRRAQAVP QPSNPLDIII

           YADVANQRNR LEPGTEVNIL FLFNHLIWDG KGRYFTSELV QRATTILDQE
           YADVANQRNR LEPGTEVNIL FLFNHLIWDG KGRYFTSELV QRATTILDQE

           KENIMPTHRW GEEKSRLDPP ILDVMLVNLD KMGPDYDLAH RKLLNSQLQV
           KENIMPTHRW GEEKSRLDPP ILDVMLVNLD KMGPDYDLAH RKLLNSQLQV
```

```
           GLSWGLPLTR NPGEPLQIRH CISREDSTKI TDAVRARLGP KYNIGHLGHA
           GLSWGLPLTR NPGEPLQIRH CISREDSTKI TDAVRARLGP KYNIGHLGHA

           ATVLSLLKNN PIPPSTQDTA FLFSPLPVDG RPYLLEERKT PRYGNAQAAA
           ATVLSLLKNN PIPPSTQDTA FLFSPLPVDG RPYLLEERKT PRYGNAQAAA

           VVELQKLASW GIKSDNLNGV KVALDDLAKK VKEDYDYWLT NLVAWRFKSS
           VVELQKLASW GIKSDNLNGV KVALDDLAKK VKEDYDYWLT NLVAWRFKSS

           CTSEFIAFGS AIYQTPYLDP GAPKVKVGTG TSTDMVFLKA FCNDGRAESI
           CTSEFIAFGS AIYQTPYLDP GAPKVKVGTG TSTDMVFLKA FCNDGRAESI

           IAYTMHGPSG KELFQVDDCF GGVDVLGSNA FIRMDTWKDA IRLTLCYNSG
           IAYTMHGPSG KELFQVDDCF GGVDVLGSNA FIRMDTWKDA IRLTLCYNSG

           CFSDAVANSF TTDVAQYMLA YSW
           CFSDAVANSF TTDVAQYMLA YSW

Sat13, 2383 aa
S40293     MSGPNPVPLA IVGIACRFPG DATNPEKFWD LLANARSGWS RVPNDRWNEE
S7711      MSGPNPVPLA IVGIACRFPG DATNPERFWD LLANARSGWS RVPNDRWNEE

           AFWHPDPDDT NGTNNHMGGH FLNQDLARFD AGFFNVTPQE AASMDPQQRL
           AFWHPDPDDT NGTNNHMGGH FLNQDLARFD AGFFNVTPQE AASMDPQQRL

           LLETTYEALE SAGIPQEHIR GSNTAAYMAM FTRDYDRNVY KDMMSIPKYH
           LLETTYEALE SAGIPQEHIR GSNTAAYMAM FTRDYDRNVY KDMMSIPKYH

           VTGTGDAILA NRISHLFDLR GPSVTMDTGC SGGLTAISHA CQALRSGLSD
           VTGTGDAILA NRISHLFDLR GPSVTMDTGC SGGLTAISHA CQALRSGLSD

           IGLAGAVNLI LTPDHMVGMS NLHMLNVNGR SFSFDSRGAG YGRGEGVATL
           IGLAGAVNLI LTPDHMVGMS NLHMLNVNGR SFSFDSRGAG YGRGEGVATL

           VIKRLDDAIR DKDPVRAILR DAAINQDGYT AGITLPSGRA QQALERRVWD
           VIKRLDDAIR DKDPVRAILR DAAINQDGYT AGITLPSGRA QQALERRVWD

           VLNLDPATVG YVEAHGTGTL AGDSAELEGI SKIFCENRDH GSPLIVGSVK
           VLNLDPATVG YVEAHGTGTL AGDSAELEGI SKIFCENRDH GSPLIVGSVK

           SNIGHTECVS GIAAVIKSTL ILENGTIPPN INFEQPRESL DLRNKKIKVP
           SNIGHTECVS GIAAVIKSTL ILENGTIPPN INFEQPRESL DLRNKKIKVP

           NALMPWPQTT GTARISVNSF GYGGTNAHAV LERAERVIDT TCPEEDDAPQ
           NALMPWPQTT GTARISVNSF GYGGTNAHAV LERAERVIDT TCPQEDDAPQ

           LFIFSAASQT SLLGMLAANR DWVSENRERA WVMRDLAYTL SQRRSLLPWR
           LFIFSAASQT SLLGMLAANR DWVSENRERA WVMRDLAYTL SQRRSLLPWR

           FSCVAANRSE LLETLSSVPQ NANSIARITP GSRISFIFTG QGAQWAGMGR
           FSCVAANRSE LLETLSSVPQ NANSIARITP GSRISFIFTG QGAQWAGMGR

           ELLSMPTFNS SLQRSNEILQ DLGCSWDLIE EVSKQKPESR LHEPELSQPL
           ELLSMPTFNS SLQRSNEILQ DLGCSWDLIE EVSKQKPESR LHEPELSQPL
```

```
TTAIQIALVD LFREWGIVPD SVIGHSSGEI GAAYTAGHIA HCQAIKVAYF
TTAIQIALVD LFREWGIVPD SVIGHSSGEI GAAYTAGHIA HCQAIKVAYF

RGFSSAWAAQ AHKRGAMLAV GLGEYDVEPY LEQLEQGHAS IACQNSPNST
RGFSSAWAAQ AHKRGAMLAV GLGEYDVEPY LEQLGQGHAS IACQNSPNST

TVSGDDAAIS ELSEILTKES IFNRKLNITV AYHSHHMQTA ACQYKAALEP
TVSGDDAAIS ELSEILTKES IFNRKLNITV AYHSHHMQTA ACQYKAALEP

LLTNPSLDTG IEMFSTVTGS IKKDAFNSNY WVENLVSKVR FCDGLQALCE
LLTNPSLDTG IEMFSTVTGS IKKDAFNSNY WVENLVSKVR FCDGLQALCE

STQASPLGSS KAERIFIEIG PHSALAGPTR QCIADLITPL PYSYTSGLLR
STQASPLGSS KAERIFIEIG PHSALAGPTR QCIADLITPL PYSYTSGLLR

ETGAVKSALA MVGHIFNRGY SLNLAAISAS NKTSQYATVL SNLPSYHWDH
ETGAVKSALA MVGHIFNRGY SLNLAAISAS NKTSQYATVL SNLPSYHWDH

TRRHWNESRI SREYRFRKHP YHDLLGLRMT EVSPLRPSWR HMIGTKGLPW
TRRHWNESRI SREYRFRKHP YHDLLGLRMT EVSPLRPSWR HMIGTKGLPW

LADHVVDDLV IFPGSGYLAM AIEACSQLAD DRYPGREIER FSLNDIFFLK
LADHVVDDLV IFPGSGYLAM AIEACSQLAD DRYPGREIER FSLNDIFFLK

GLIIPDDGAR VEVQLSLNPI EPADKDTRMN VMQHEFSVTA FTDEARWNEH
GLIIPDDGAR VEVQLSLNPI EPADKDTRMN VMQHEFSVTA FTDEARWNEH

CRGNIVVVFK TSSATERLVA NGFTRGDMAA QLDPVSGKLT HAGQLYPELR
CRGNIVVVFK TSSATERLVA NGFTRGDMAA QLDPVSGKLT HAGQLYPELR

KAGNSYGLTF NGIQRMKIGA DSASSDVIIP DVVSRMPACH MRPHIIHPTT
KAGNSYGLTF NGIQRMKIGA DSASSDVIIP DVVSRMPACH MRPHIIHPTT

LDILLHTTLP LVHQKLGVGS VMPVHIRNMD VSADIESTPR KMFRVVTTLT
LDILLHTTLP LVHQKLGVGS VMPVHIRNMD VSADIESTPR KMFRVVTTLT

SSHARAADTE LFVFSEEGHV DDTPVVSAAG MELRSFVARD SNDAGSSDGH
SSHARAADTE LFVFSEEGHV DDTPVVSAAG MELRSFVARD SNDAGSSDGH

RDICSELKWI PDERFITAKH LQVLQPSILT KDALARCYAL MAQYLKQMAI
RDICSELKWI PDERFITAKH LQVLQPSILT KDALARCYAL MAQYLKQMAI

KHSDLSVLEL GGDDTTSGAT KTFLEVFHAG GTAPAMYDFC TSLKDFDVIQ
KHSDLSVLEL GGDDTTSGAT KTFLEVFHAG GTAPAMYDFC TSLKDFDVIQ

RKLEAFDCEK VHKVEMKRIE LDAVSENRYD VVLSCNTIYN AADVKSVLSH
RKLEAFDCEK VHKMEMKRIE LDAVSENRYD VVLSCNTIYN AADVKSVLSH

ARKLLKLDGV LLFVEDMSSR ESRSSSEWSK LMSEASFKMQ LAVTDNDATR
ARKLLKLDGV LLFVEDMSSR ESRSSSEWSK LMSEASFKMQ LAVTDNDATR

QLTFFATRAV EDAIASVHNV SIVSGCNLPL HIQNFLPQIE SELGSKGMQV
QLTFFATRAV EDAIASVHNV SIVSGCNLPL HIQNFLPQIE SELGSKGMQV
```

```
TRSCWDKLPP NGTDIYIIVD DGSRPILSGI NQDRFRIVTG LLQKTARIIW
TRSRWDKLPP NGTDIYIIVD DGSRPILSGI NQDRFRIVTG LLQKTARIIW

LSVQDDETFR FNPRKHLITG LSRTAHAENE GLDMVTIDVQ ETLNQKTQPE
LSVQDDETFR FNPRKHLITG LSRTAHAENE GLDMVTIDVQ ETLNQKTQPE

VIGFLSQVVG LFDCKHITRE REYVYNGTDI LIPRLIPHQR LNLQVSGKIG
VIGFLSQVVG LFDCKHITRE REYVYNGTDI LIPRLIPHQR LNLQVSGKIG

TSIEAMAFTN SSVPLKLSDG QNRLVFVENM DHKQALCHDY VEIETKAVGL
TSIEAMAFTN SSVPLKLSDG QNRLVFVENM DHKQALCHDY VEIETKAVGL

PPGFNGVQSG NTVYEYAGII IAVGSEVSTL KAGDRAVAYS STPCANVLRV
PPGFNGVQSG NTVYEYAGII IAVGSEVSTL KAGDRAVAYS STPCANVLRV

PAIQAQLIPS NLSFKDAAAM PRALMAVSHA LVHIANVQPG QVVFVDDAAT
PAIQAQLIPS NLSFKDAAAM PRALMAVSHA LVHIANVQPG QVVFVDDAAT

EIGLAAICVA QNLGSTLIAA VSTKEEAAFI KNTFKVPSRH IVPRDSYFGQ
EIGLAAICVA QNLGSTLIAA VSTKEEAAFI KNTFKVPSRH IVPRDSYFGQ

RQVRTLVRPN GGLDVILGCG KSPVTAVTSE LLKPFGLLVH VRNRASDPKR
RQVRTLVRPN GGLDVILGCG KSPVTAVTSE LLKPFGLLVH VRNRASDPKR

YDGTGYPPNL TVASFDIDSL LQASTKNSAE LFQKVMEMVN RGMIPPSQSI
YDGTGYPPNL TVASFDIDSL LQASTKNSAE LFQKVMEMVN RGMIPPSQSI

VAIEAGIKIE EAISLAQKQG SMKKCVLEFN ENSIVNVETS FHHIPSLKPH
VAIEAGIKIE EAISLAQKQG SMKKCVLEFN ENSIVNVETS FHHIPSLKPH

ATYVVAGGLG DLGQRLLRLM AQAGARHLVS LSRKGAGSKE FRGLEKELKG
ATYVVAGGLG DLGQRLLRLM AQAGARHLVS LSRKGAGSKE FRGLEKELKG

VHPGCSLLAI DCDILREESV SAALAEIKQQ GFPTVKGVVQ SAVILKDATL
VHPGCSLLAI DCDILREESV SAALAEIKQQ GFPTVKGVVQ SAVILKDATL

DSMTAELFNS VVSVKAEGTL NLHRVFIQEE LAFFISFSSV MSIIGGKAQA
DSMTAELFNS VVSVKAEGTL NLHRVFIQEE LAFFISFSSV MSIIGGKAQA

NYNAGNAVQD AFAQFERRNP HCFYMSLNIG GIKDAAVNND AIVQSIRRQG
NYNAGNAVQD AFAQFERRNP HCFYMSLNIG GIKDAAVNND AIVQSIRRQG

LTQISHEELS SYLKYAFSDD ARKTGCKQPV IGFTAETIVS TTAVNGTAHT
LTQISHEELS SYLKYAFSDD ARKTGCKQPV IGFTAETIVS TTAVNGTAHT

PMFTHVRQKP TAKTTVGNVN EKRSFKDVVN SGTNKGEISE FVARSICDKI
PMFTHVRQKP TAKTTVGNVN EKRSFKDIVN SGTNKGEISE FVARSICDKI

ADLTGIDLAE VNLDSGISDY GLDSLVSIEL RNWLMREFDS PIQSSEVLDS
ADLTGIDLAE VNLDSGISDY GLDSLVSIEL RNWLMREFDS PIQSSEVLDS

HGIRDLAQKV VSRSRLVTTE TDVVHTVNGE APT
HGIRDLAQKV VSRSRLVTTE TDVVHTVNGE APT
```

Sat14, 455 aa

```
S40293     MATIPIRLQA LATDQTVLKL PHPYKTEFAV RKASNASTKL PVYNLVPKPF
S7711      MATIPIRLQA LATDQTVLKL PHPYKTEFAV RKASKASTKL PVYNLVPKPF

           PTRPLPFELH NDHLVFTDAI HLKSSELPPD SNNGAWARAR RAPCVTLYWD
           PTRPLPFELH NDHLVFTDAI HLKSSELPPD SNNGAWARAR RAPCVTLYWD

           GVEVPTLKQA WLVVYAFFTM RPGMDSFRLE LDGNSAANLA RQIKDVLLGI
           GVEVPTLKQA WLVVYAFFTM RPGMDSFRLE LDGNSAANLA RQIKDVLLGI

           DHPIKARQQQ EPCAKTKENT LLILRSTFWQ GAGCPFGPRP VWCPQESPSS
           DHPIKARQQQ EPCAKTKENT LLILRSTFWQ GAGCPFGPRP VWCPQESPSS

           LLPSTCLSSF PLAPFHRTST ISLAGDPEDF DRCQQSWHPI RPAKPAPGSI
           LLPSTCLSSF PLAPFHRTST ISLAGDPEDF DRCQQSWHPI RPAKPAPGSI

           IYSRWIPYLG EMFSMVALDP EDSEHVRLFH EWQSDPRVLQ GWTETKTLDQ
           IYSRWIPYLG EMFSMVALDP EDSEHVRLFH EWQSDPRVLQ GWTETKTLDQ

           HRRYLEALHK DPHQLTVLAK WDDSPFAYFE LYWAKENRLG GYIDAGDFDR
           HRRYLEALHK DPHQLTVLAK WDDSPFAYFE LYWAKENRLG GYIDAGDFDR

           GRHSFVGDVR FRGPLRVSAW WSSLMHYLFL DDPRTMYIVG EPRDTHSTVL
           GRHSFVGDVR FRGPLRVSAW WSSLMHYLFL DDPRTMHIVG EPRDTHSTVL

           MYDFIHGFGL DRFIDLPSKR SAFMRCSRDR FFQSFPLEDS EKVIGGTSIR
           MYDFIHGFGL DRFIDLPSKR SAFMRCSRDR FFQSFPLEDS EKVIGGTSIR

           VVQKL
           VVQKL

Sat15, 153 aa
S40293     MTTPPGWKVS GQNEISRPFD ILEAWFHRIV GGGNLTRERD SFGSNYVVKL
S7711      MTTPPGWKVS GQNEISRPFD ILEAWFHRIV GGGNLTRERD SFGSNYVVKL

           GFPGSVADPI PYLRRAWLVT RYLHPQLGAT YSSKSLDDLR YIIRPLDEQI
           GFPGSVADPI PYLRRAWLVT RYLHPQLGAT YSSKSLDDLR YIIRPLDEQI

           WLQTTFFVEQ GPSATYSSAE DAVSKYLSKS TTTAHWIPAT SEFMISPTAS
           WLQTTFFVEQ GPSATYSSAE DAVSKYLSKS TTTAHWIPAT SEFMISPTAS

           SPL
           SPL

Sat16, 184 aa
S40293     ---------- ---------- ---------- ---------- ----------
S7711      MVELNPVTIT NDNATPRGHV LKLILVESRD IIRAVKTRLC PQYTISYLAQ

           ------MLDT YSSKSELSKP DFFVALTAVN GRRYLREDLE SNYLAGYVTG
           AATVIAMLDT YSSKSELSKP DFFVALTVVN GRRYLREDLE SNYLAGYVTG

           APIKIEKLRS LLVSLDDSKD IIVSALEKAA KDAKRRLDMW IYDQSQLATG
           APIKIEKLRS LLVSLDDSKD IIVSALEKAA KDAKRRLDMW IYDQSQLATG

           FRIHSFKGAM SSENPELFIK TAVPYLSSYG INEV
           FRIHSFKGAM SSENPELFKK TAVPYLSSYG INEV
```

Satratoxin cluster 3, 5 products

```
Sat17, 393 aa
S40293     MAASPVLATT SHPIGHEAAV VTDADLDRHY AVKLAGKLND EMAWVGQQFT
S7711      MAASPVLATT SHPIGHEAAV VTDADLDRHY AVKLAGKLND EMAWVGQQFT

           GEEDFVVCLS EADVAEVNAA LTAFQGMFLK PDYYTGLKPG YLSPETFKLP
           GEEDFVVCLS EADVAEVNAA LTAFQD---- ----TGLKPG YLSPETFKLP

           KLGPKLRLLS QRIHEQEGFI VLRGLQPWRY RRLENTIVFT GIASYIGNRR
           KLGPKLRLLS QRIHEQEGFI VLRGLQPWRY RRLENTIVFT GIASYIGNRR

           GVQCADGPVM THIFDYSTEV EEKEKLNDGY LGHANRTSYL PFHTDDGHII
           GVQCADGPVM THIFDYSTEV EEKEKLNDGY LGHANRTSYL PFHTDDGHII

           SLYCLQAADI GGRTLLASSH AIYNHLLETR PDVIETLKEE WIWDSFIPEK
           SLYCLQAADI GGRTLLASSH AIYNHLLETR PDVIETLKEE WIWDSFIPEK

           PSFIRPLLLE QDGKLICNYR IRPFLGTPGY PRNAALGPLP AHQEEALNTV
           PSFIRPLLLE QDGKLICNYR IRPFLGTPGY PRNAALGPLP AHQEEALNTV

           AEIAEKLSLK FEFKTGDIQF LNNLSILHAR EEFHCAKGDT TRRHLLRLVQ
           AEIAEKLSLK FEFKTGDIQF LNNLSILHAR EEFHCAKGDT TRRHLLRLVQ

           MDDELAWRLP PGLSKDMDKM FQHDLEEEKF IWSPEPL--- ---
           MDDELAWRLP PGLSKDMDKM FQHDLEEEKF IWSPEPLPYV IGQ

Sat18, 401 aa
S40293     MKLVEIAEDI LSKANAYTNN TGLTSSQRFQ LREEIRYQAN GILSAIDGPE
S7711      MKLVEIAEDI LSKANAYTNN TGLTSSQRFQ LREEIRYQAN GILSAIDGPE

           QTMKAIARSY TTCTALKVCV DLKLASHLPL SDARSLSQLA QICGCDSRVL
           QTMKAIARSY TTCTALKVCV DLKLASHLPL SDARSLSQLA QICGCDSLVL

           RPMLRLLAKN GIFEQVDAET WQHTELSAVM AQPPFQALEE KYRSVAHLPR
           RPMLRLLAKN GIFEQVDAET WQHTELSAVM AQPPFQALEE KYRSVAHLPR

           LLQAVSHQFP TPGRTAFNQV YCTSLDFYTY SNELDHAAAR NFAFSMKELA
           LLQAVSHQFP TPGRTAFNQV YCTSLDFYTY SNELDHAAAR NFAFSMKELA

           RNQIPFVQQS YPLETIDPES HFIDVAGGVG YLSFFLAGSF PKATFEVQDH
           RNQIPFVQQS YPLETIDPES HFIDVAGGVG YLSFFLAGSF PKATFEVQDH

           PFIIEEAHSV CPSELRDRIT FRAHNILHPQ PEIAKEINGR LVFLVKIILH
           PFIIEEAHSV CPSELRDRIT FRAHNILHPQ PEIAKEINGR LVFLVKIILH

           DHGDDDCRLM LRNLVSVMKQ GDRILIIDTV IPETGGSLSS ANSDIIIMSM
           DHGDDDCRLM LRNLVSVMKQ GDRILIIDTV IPETGGSLSS ANSDIIIMSM

           FGSGHRTLEE FRALIHHCGE DLVIETFASG DEEYDGMMVI EVRKAEPVLD
           FGSGHRTLEE FRALIHRCGE DLVIETFASG DEEYDGMMVI EVRKAEPVLD
```

```
           N
           N

Sat19, 230 aa
S40293     MATATPPPQD FPPYPPFTSL RLAAARDVAQ MANLSVQGFK DSEIFRYERP
S7711      MATATPPPQD FPPYPPFTSL RLAAARDVAQ MANLSVQGFK DSEIFRYERP

           GHDQYPEDAV AYFANLYRDR LEDPRAVVIV AEDWDGAERV VVGVGCWILP
           GHDQYPEDAV AYFANLYRDR LEDPRAVVIV AEDWDGAERV VVGVGCWILP

           QDSPRTGQFV VPCVGDREPA LDRDLCCRRL ELFNAVTKAT EERYLDGKVI
           QDSPRTGQFV VPCVGDREPA LDRDLCCRRL ELFNAVTKAT EERYLDGKVI

           CDKFVVHPSY QRRGHGTAML RWSLRLCTQD TVDQGVIPSH VGEPVYLSLG
           CDKFVVHPSY QRRGHGTAML RWSLRLCTQD TVDQGVIPSH VGEPVYLSLG

           FEVIGEMHVP DEGDTQGFTQ RVAVYKARQT
           FEVIGEMHVP DEGDTQGFTQ RVAVYKARQT

Sat20, 712 aa
S40293     MPNLPGSSDS TQRHQRNPGI DELVCSSTKP NAAQENADTE LAQEKHPQLL
S7711      MPNLPGSSDS TQRHQRNPGT DELVCSSTKP NAAQENADTE LAQEKHPQLL

           SPQTDIPPVC SQPNVSFAQW WDQNFFLDAA LTG------- --------HM
           SPQTDIPPVC SQPNVSFAQC WDQNFFLDAA LTGQNSMTLL GVFDLPNSHM

           NLTETLSSQL GLEPSQNAFG QSSFDPFFPS SEPSVHSTDN PWPSLPTQLA
           NLTETLSSQL GLEPSQNAFG QGSFDPFVPS SEPSVHSTDN PWPSLPTQLA

           FQPTNNNTSS LALLGPDPEQ LSTLPSPWTG PLEEWPPLDL GQDFAALLSP
           FQPTNNNTSS LALLGPDPEQ LSTLPSSWTG PLDEWPPLDL GQDFAALLSP

           TYQALEATPR HTHAHAHQPP RATRHITSQQ SPEPYVQVHS TAKIVDRFLV
           TYQAFEATPR HTHAHAHQPP RATRHITSQQ SPEPYVQVHS TAKIVDRFLV

           PIAPKPILVE RDGPVSGASL PSNPPTALSS TGTRKRKRFN KADRERVNQM
           PIAPKPILVE RDGPVSGASL PSNPPTALSS TGTRKRKRFN KADRERVNQM

           RKLGSCFRCR MYKENCDPGL PCKNCMRVQV TRRTFFGPCI RIKWEEVHTF
           RKLGSCFRCR MYKENCDPGL PCKNCMRVQV TRRTFFGPCI RIKWEEVHTF

           RAGDGDLGQI RATLQTFQWT LGGQVKSIDV QWPFRDDKVK PPILSIECQQ
           RAGDGDLGQI RATLQTFQWT LGGQVKSIDV QWPFRDDKVK PPILSIECQQ

           FLPKHEHVAE EYSVAGQAYK ILLPPWACSN TKAASKKVEA FVRQCQAPLE
           FLPKHEHVAE EYSVAGQAYK ILLPPWACSN TKAASKKVEA FVRQCQAPLE

           EEIRHTLNDP ILLLTLDEAR RYRNETGSKL VATALEIYAG AMMNSRYPAS
           EEIRHTLNDP ILLLTLDEAR RYRNETGSKL VATALEIYAG AMMNSRYPAS

           TESDIFGVVD QLHTPYFFDK VPLPPQLTCQ IQIMVAQVML DKQKNALKRL
           TESDIFGVVD QLHTPYFFDK VPLPPQLTCQ IQIMVAQVML DKQKNALKRL

           QERALSKNRH KVWYECYLTI FILLATIELV YQVQLRFVKA KQGVSDRNAT
           QERALSKNRH KVWYECYLTI FILLATIELV YQVQLRFVKA KQGVSDRNAT
```

```
          NLSYVTQYMI EEWEESILTL VGLFHCVMNG GLPFTQSWED GGENHRLTEL
          NLSYVTQYMI EEWEESILTL VGLFHCVMNG GLPFTQSWED GGENHRLTEL

          DDKALVYVRS LKAEIEQRRG ELIALRNRRG RWRYEQPLAA ICQLFLPSQD
          DDKALVYVRS LKAEIEQRRG ELIALRNRRG RWRYEQPLAA ICQLFLPSQD

          GDKGEGRAAP PS
          GDKGEGRAAP PS

Sat21, 474 aa
S40293    MLRNTHLVVP FILYLLFRLS HFLLEVPTVR MIELAACHQH LRLDHGPLNE
S7711     MLRNTHLVLP FILYLLFRLS HFLLEVPTVR MIELAACHQH LRLDHGPLNE

          AACKTPPVQE HVSLVVGWKM TFDSIPGLMS ILYFGTLADK SGHRAILRLC
          AACKTPPVQE HVSLVVGWKM TFDSIPGLMS ILYFGTLADK SGHRAILRLC

          CVGYLLAILW VLITCLFHQV FPVELVLLSS LFLFIGGGQL VFAAVITAFV
          CVGYLLAILW VLITCLFHQV FPVELVLLSS LFLFIGGGQL VFAAVITAFV

          ADLFPPPSRT KFLFLLAAMP HMDKVASPAL ATKLMEQNLF LPSLVSMAIV
          ADLFPPPSRT KFLFLLAAMP HMDKVASPAL ATKLMEQNLF LPSLVSMAIV

          VICVALLQMS DVGRETAASK VVGSTSDQTE PFLRSSSNSS QESGTAAPAI
          VICVALLQMS DVGRETAASK VVGSTSDQTE PFLRSSSNSS QESGTAAPAI

          DPEQARGPFR QLKNIICWVY REPVLFICYL CFFLKSNAMA SEAFIFQYLS
          DPEQARGPFR QLKNIICWVH REPVLFICYL CFFLKSNAMA SEAFIFQYLS

          EKFGWPLRET TVMRLALSSG AVISTLIICP LANATLHNRG VASARINIGA
          EKFGWPLRET TVMRLALSSG AVISTLIICP LANATLHNRG VASARINIGA

          VHASSIVLVA SFIMAWQASS STAFIFSMLA AGFGEGLEPA LQGVLAAASQ
          VHASSIVLVA SFIMAWQASS STAFIFSMLA AGFGEGLEPA LQGVLAAASQ

          TKAKGSIFAL MCTCSLLGDM TGGPLMSALM SIGRGGNGVS DGYCFLASAL
          TKAKGSIFAL MCTCSLLGDM TGGPLMSALM SIGRGGNGVS DGYCFLASAL

          VFGAVIVLAH LLWALGAEEM LGED
          VFGAVIVLAH LLWALGAEEM LGED
```

APPENDIX G (Chapter 2)
Parameters used in *Stachybotrys* genome annotation

This appendix shows the two parameter files, maker_opts.ctl and maker_bopts.ctl, that were used by MAKER during the second and final pass of our annotation. These specific files were used for strain 7711, but parameters were the same for the other three assemblies.

maker_opts.ctl

```
#-----Genome (these are always required)
genome=$HOME/sch/genomes03/S7711-1e3.fa #genome sequence (fasta file or
fasta embeded in GFF3 file)
organism_type=eukaryotic #eukaryotic or prokaryotic. Default is
eukaryotic

#-----Re-annotation Using MAKER Derived GFF3
maker_gff=/home/jrs/sch/makers14.1/S7711/genome2.gff #MAKER derived
GFF3 file
est_pass=0 #use ESTs in maker_gff: 1 = yes, 0 = no
altest_pass=0 #use alternate organism ESTs in maker_gff: 1 = yes, 0 =
no
protein_pass=0 #use protein alignments in maker_gff: 1 = yes, 0 = no
rm_pass=0 #use repeats in maker_gff: 1 = yes, 0 = no
model_pass=1 #use gene models in maker_gff: 1 = yes, 0 = no
pred_pass=0 #use ab-initio predictions in maker_gff: 1 = yes, 0 = no
other_pass=0 #passthrough anyything else in maker_gff: 1 = yes, 0 = no

#-----EST Evidence (for best results provide a file for at least one)
est= #set of ESTs or assembled mRNA-seq in fasta format
altest= #EST/cDNA sequence file in fasta format from an alternate
organism
est_gff= #aligned ESTs or mRNA-seq from an external GFF3 file
altest_gff= #aligned ESTs from a closly relate species in GFF3 format

#-----Protein Homology Evidence (for best results provide a file for at
least one)
protein=/home/jrs/sch/one.fa:S40285,/home/jrs/sch/two.fa:S40288,/home/j
rs/sch/three.fa:S40293,/home/jrs/fusarium/fgdb/FGDB_v32.prot:f_graminea
rum,/home/jrs/fusarium/fusarium_oxysporum_f._sp._lycopersici_4287_2_pro
teins.fasta:f_oxysporum,/home/jrs/fusarium/fusarium_verticillioides_760
0_3_proteins.fasta:f_verticillioides,/home/jrs/h4/db/uniprot_sprot.fast
a:swiss  #protein sequence file in fasta format (i.e. from mutiple
oransisms)
protein_gff=  #aligned protein homology evidence from an external GFF3
file

#-----Repeat Masking (leave values blank to skip repeat masking)
model_org=all #select a model organism for RepBase masking in
RepeatMasker
rmlib= #provide an organism specific repeat library in fasta format for
RepeatMasker
repeat_protein=/home/jrs/maker-2.26-beta/data/te_proteins.fasta
#provide a fasta file of transposable element proteins for RepeatRunner
rm_gff= #pre-identified repeat elements from an external GFF3 file
prok_rm=0 #forces MAKER to repeatmask prokaryotes (no reason to change
this), 1 = yes, 0 = no
```

```
softmask=1 #use soft-masking rather than hard-masking in BLAST (i.e.
seg and dust filtering)

#-----Gene Prediction
snaphmm= #SNAP HMM file
gmhmm=$HOME/sch/gm01/mod/es.mod #GeneMark HMM file
augustus_species=fusarium_graminearum #Augustus gene prediction species
model
fgenesh_par_file= #FGENESH parameter file
pred_gff= #ab-initio predictions from an external GFF3 file
model_gff= #annotated gene models from an external GFF3 file
(annotation pass-through)
est2genome=0 #infer gene predictions directly from ESTs, 1 = yes, 0 =
no
protein2genome=0 #infer predictions from protein homology, 1 = yes, 0 =
no
unmask=0 #also run ab-initio prediction programs on unmasked sequence,
1 = yes, 0 = no

#-----Other Annotation Feature Types (features MAKER doesn't recognize)
other_gff= #extra features to pass-through to final MAKER generated
GFF3 file

#-----External Application Behavior Options
alt_peptide=C #amino acid used to replace non-standard amino acids in
BLAST databases
cpus=1 #max number of cpus to use in BLAST and RepeatMasker (not for
MPI, leave 1 when using MPI)

#-----MAKER Behavior Options
max_dna_len=1000000000 #length for dividing up contigs into chunks
(increases/decreases memory usage)
min_contig=1 #skip genome contigs below this length (under 10kb are
often useless)

pred_flank=200 #flank for extending evidence clusters sent to gene
predictors
pred_stats=1 #report AED and QI statistics for all predictions as well
as models
AED_threshold=1 #Maximum Annotation Edit Distance allowed (bound by 0
and 1)
min_protein=0 #require at least this many amino acids in predicted
proteins
alt_splice=0 #Take extra steps to try and find alternative splicing, 1
= yes, 0 = no
always_complete=0 #extra steps to force start and stop codons, 1 = yes,
0 = no
map_forward=1 #map names and attributes forward from old GFF3 genes, 1
= yes, 0 = no
```

```
keep_preds=1 #Concordance threshold to add unsupported gene prediction
(bound by 0 and 1)

split_hit=10000 #length for the splitting of hits (expected max intron
size for evidence alignments)
single_exon=0 #consider single exon EST evidence when generating
annotations, 1 = yes, 0 = no
single_length=250 #min length required for single exon ESTs if
'single_exon is enabled'
correct_est_fusion=0 #limits use of ESTs in annotation to avoid fusion
genes

tries=2 #number of times to try a contig if there is a failure for some
reason
clean_try=0 #remove all data from previous run before retrying, 1 =
yes, 0 = no
clean_up=0 #removes theVoid directory with individual analysis files, 1
= yes, 0 = no
TMP=/home/jrs/h9/tmp #specify a directory other than the system default
temporary directory for temporary files
```

maker_bopts.ctl

```
#-----BLAST and Exonerate Statistics Thresholds
blast_type=ncbi+ #set to 'ncbi+', 'ncbi' or 'wublast'

pcov_blastn=0.8 #Blastn Percent Coverage Threhold EST-Genome Alignments
pid_blastn=0.85 #Blastn Percent Identity Threshold EST-Genome Aligments
eval_blastn=1e-10 #Blastn eval cutoff
bit_blastn=40 #Blastn bit cutoff
depth_blastn=0 #Blastn depth cutoff (0 to disable cutoff)

pcov_blastx=0.2 #Blastx Percent Coverage Threhold Protein-Genome
Alignments
pid_blastx=0.2 #Blastx Percent Identity Threshold Protein-Genome
Aligments
eval_blastx=1e-06 #Blastx eval cutoff
bit_blastx=30 #Blastx bit cutoff
depth_blastx=0 #Blastx depth cutoff (0 to disable cutoff)

pcov_tblastx=0.8 #tBlastx Percent Coverage Threhold alt-EST-Genome
Alignments
pid_tblastx=0.85 #tBlastx Percent Identity Threshold alt-EST-Genome
Aligments
eval_tblastx=1e-10 #tBlastx eval cutoff
bit_tblastx=40 #tBlastx bit cutoff
depth_tblastx=0 #tBlastx depth cutoff (0 to disable cutoff)

pcov_rm_blastx=0.5 #Blastx Percent Coverage Threhold For Transposable
Element Masking
pid_rm_blastx=0.4 #Blastx Percent Identity Threshold For Transposbale
Element Masking
eval_rm_blastx=1e-06 #Blastx eval cutoff for transposable element
masking
bit_rm_blastx=30 #Blastx bit cutoff for transposable element masking

ep_score_limit=20 #Exonerate protein percent of maximal score threshold
en_score_limit=20 #Exonerate nucleotide percent of maximal score
threshold
```

BIBLIOGRAPHY

Alexander NJ, McCormick SP, Hohn TM (1999) TRI12, a trichothecene efflux pump from Fusarium sporotrichioides: gene isolation and expression in yeast. Mol Gen Genet 261:977–984.

Arnaud MB, Cerqueira GC, Inglis DO, Skrzypek MS, Binkley J, et al (2012) The Aspergillus Genome Database (AspGD): recent developments in comprehensive multispecies curation, comparative genomics and community resources. Nucleic Acids Res 40:D653–659.

Arnold K, Bordoli L, Kopp J, Schwede T (2005) The SWISS-MODEL workspace: A web-based environment for protein structure homology modelling. Briefings Bioinf 22:195–201.

Adachi J, Hasegawa M (1996) MOLPHY version 2.3: Programs for molecular phylogenetics based on maximum likelihood. Computer Science Monographs. Tokyo: Institute of Statistical Mathematics, 28:1–150.

Andersen B, Nielsen KF, Jarvis BB (2002) Characterization of Stachybotrys from water-damaged buildings based on morphology, growth and metabolite production. Mycologia 94:392–403.

Andersen B, Nielsen KF, Thrane U, Szaro T, Taylor JW, Jarvis BB (2003) Molecular and phenotypic descriptions of Stachybotrys chlorohalonata sp. nov. and two chemotypes of Stachybotrys chartarum found in water-damaged buildings. Mycologia 95:1227–1238.

Angiuoli SV, Salzberg SL (2011) Mugsy: fast multiple alignment of closely related whole genomes. Bioinformatics 27:334–342.

Belville C, Jamin SP, Picard J-Y, Josso N, di Clemente N (2005) Role of type I receptors for anti-Müllerian hormone in the SMAT-1 Sertoli cell line. Oncogene 24:4984–4992.

Belville C, Maréchal J-D, Pennetier S, Carmillo P, Masgrau L, et al (2009) Natural mutations of the anti-Müllerian hormone type II receptor found in persistent Müllerian duct syndrome affect ligand binding, signal transduction and cellular transport. Hum Mol Genet 18:3002–3013.

Bininda-Emonds OR, Cardillo M, Jones KE, MacPhee RD, Beck RM, et al. (2007) The delayed rise of present-day mammals. Nature 446:507–512.

Blom N, Gammeltoft S, Brunak S (1999) Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. J Mol Biol 294:1351–1362.

Boeckmann B, Bairoch A, Apweiler R, Blatter M-C, Estreicher A, et al. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nuc Acids Res 31:365–370.

Brown DW, Dyer RB, McCormick SP, Kendra DF, Plattner RD (2004) Functional demarcation of the Fusarium core trichothecene gene cluster. Fungal Genet Biol 41:454–462.

Burns DS, Jimenez-Krassel F, Ireland JL, Knight PG, Ireland JJ (2005) Numbers of antral follicles during follicular waves in cattle: evidence for high variation among animals, very high repeatability in individuals, and an inverse association with serum follicle-stimulating hormone concentrations. Biol Reprod 73:54–62.

Cardoza RE, Malmierca MG, Hermosa MR, Alexander NJ, McCormick SP, et al (2011) Identification of loci and functional characterization of trichothecene biosynthesis genes in filamentous fungi of the genus Trichoderma. Appl Environ Microbiol 77:4867.

Cherry JM, Adler C, Ball C, Chervitz SA, Dwight SS, et al (1998) SGD: Saccharomyces Genome Database. Nucleic Acids Res 26:73–79.

Cox RJ (2007) Polyketides, proteins and genes in fungi: programmed nano-machines begin to reveal their secrets. Org Biomol Chem 5:2010–2026.

Cruse M, Telerant R, Gallagher T, Lee T, Taylor JW (2002) Cryptic species in Stachybotrys chartarum. Mycologia 94:814–822.

Degenkolb T, Dieckmann R, Nielsen KF, Gräfenhan T, Theis C, et al (2008) The Trichoderma brevicompactum clade: a separate lineage with new species, new peptaibiotics, and mycotoxins. Mycol Progress 7:177–219.

de Magalhães JP, Budovsky A, Lehmann G, Costa J, Li Y, et al (2009) The Human Ageing Genomic Resources: Online databases and tools for biogerontologists. Aging Cell 8:65–72.

de Magalhães JP, Costa J (2009) A database of vertebrate longevity records and their relation to other life-history traits. J Evol Biol 22:1770–1774.

de Magalhães J, Costa J, Church G (2007) An Analysis of the Relationship Between Metabolism, Developmental Schedules, and Longevity Using Phylogenetic Independent Contrasts. J Gerontol A Biol Sci Med Sci 62:149–160.

di Clemente N, Wilson C, Faure E, Boussin L, Carmillo P, et al (1994) Cloning, expression, and alternative splicing of the receptor for anti-Müllerian hormone. Mol Endocrinol 8:1006–1020.

Durlinger ALL, Visser JA, Themmen APN (2002) Regulation of ovarian function: The role of anti-Müllerian hormone. Reproduction 124:601–609.

Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, et al (2007) Comparative protein structure modeling using MODELLER. Curr Protoc Protein Sci pp 2.9.1–2.9.31.

Huang DW, Sherman BT, Lempicki RA (2009) Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. Nuc Acid Res 37:1–13.

Finch CE (1990) Longevity, Senescence, and the Genome. Chicago: University of Chicago Press. 922 p.

Finch CE, Gosden RG (1986) Animal models for the human menopause. In: Mastroianni L, Paulsen CA, editors. pp 3–34. New York: Plenum.

Foote A (2008) Mortality rate acceleration and post-reproductive lifespan in matrilineal whale species. Biol Lett 4:189–191.

Fraaije MW, Kamerbeek NM, van Berkel WJH, Janssen DB (2002) Identification of a Baeyer-Villiger monooxygenase sequence motif. FEBS Lett 518:43–47.

Freckleton RP, Harvey PH, Pagel M (2002) Phylogenetic analysis and comparative data: A test and review of evidence. Am Naturalist 160:712–726.

Gnerre S, MacCallum I, Przybylski D, Ribeiro FJ, Burton JN, et al (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Nat Acad Sci USA 108:1513–1518.

Goldsmith EJ, Akella R, Min X, Zhou T, Humphreys JM (2007) Substrate and docking interactions in serine/threonine protein kinases. Chem Rev 107:5065–5081.

Gorelick PB (2004) Risk factors for vascular dementia and Alzheimer disease. Stroke 35:2620–2622.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. Systematic Biol 59:307–321.

Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc Nat Acad Sci USA 89:10915–10919.

Hinkley SF, Mazzola EP, Fettinger JC, Lam Y-F, Jarvis BB (2000) Atranones A-G, from the toxigenic mold Stachybotrys chartarum. Phytochemistry 55:663–673.

Hiramoto-Yamaki N, Takeuchi S, Ueda S, Harada K, Fujimoto S, et al. (2010) Ephexin4 and EphA2 mediate cell migration through a RhoG-dependent mechanism. J Cell Biol 190: 461–477.

Holt C, Yandell M (2011) MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics 12:491.

Imbeaud S, Faure E, Lamarre I, Mattei M-G, di Clemente N, et al (1995) Insensitivity to anti-mullerian hormone due to a mutation in the human anti-mullerian hormone receptor. Nature Genet 11:382–388.

Jaiswal RS, Singh J, Adams GP (2009) High-resolution ultrasound biomicroscopy for monitoring ovarian structures in mice. Reprod Biol Endocrinol 7:69.

Jarvis BB, Lee YW, Cömezoglu SN, Yatawara CS (1986) Trichothecenes produced by Stachybotrys atra from Eastern Europe. Appl Environ Microbiol 51:915–918.

Jarvis BB (2003) Stachybotrys chartarum: a fungus for our time. Phytochemistry 64:53–60.

Jobson RW, Nabholz B, Galtier N (2010) An evolutionary genome scan for longevity-related natural selection in mammals. Mol Biol Evol 27:840–847.

Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292:195–202.

Kanapuru B, Ershler WB (2009) Inflammation, coagulation, and the pathway to frailty. Am J Med 122:605–613.

Katoh K, Toh H (2008) Recent developments in the MAFFT multiple sequence alignment program. Brief Bioinf 9:286–298.

Keller NP, Turner G, Bennett JW (2005) Fungal secondary metabolism -- from biochemistry to genomics. Nat Rev Microbiol 3:937–947.

Kelley DR, Schatz MC, Salzberg SL (2010) Quake: quality-aware detection and correction of sequencing errors. Genome Biol 11:R116.

Kevenaar ME, Themmen APN, Rivadeneira F, Uitterlinden AG, Laven JSE, et al (2007) A polymorphism in the AMH type II receptor gene is associated with age at menopause in interaction with parity. Hum Reprod 22:2382–2388.

Kim EB, Fang X, Fushan AA, Huang Z, Lobanov AV, et al. (2011) Genome sequencing reveals insights into the physiology and longevity of the naked mole rat. Nature 479:223–227.

Kirkwood T (1977) Evolution of ageing. Nature 270:301–304.

Knighton DR, Zheng J, TenEyck LF, Ashford VA, Xuong N-H, et al (1991) Crystal structure of the catalytic subunit of cyclic adenosine monophosphate-dependent protein kinase. Science 253:407–414.

Kroken S, Glass NL, Taylor JW, Yoder OC, Turgeon BG (2003) Phylogenomic analysis of type I polyketide synthase genes in pathogenic and saprobic ascomycetes. Proc Nat Acad Sci USA 100:15670–15675.

Kubicek CP, Herrera-Estrella A, Seidl-Seiboth V, Martinez DA, Druzhinina IS, et al (2011) Comparative genome sequence analysis underscores mycoparasitism as the ancestral life style of Trichoderma. Genome Biol 12:R40.

Kuhn DM, Ghannoum MA (2003) Indoor mold, toxigenic fungi, and Stachybotrys chartarum: infectious disease perspective. Clin Microbiol Rev 16:144–172.

Lapointe J, Hekimi S (2009) When a theory of aging ages badly. Cell Mol Life Sci 67:1–8.

Li L, Stoeckert CJ Jr, Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13:2178–2189.

Li R, Zhu H, Ruan J, Qian W, Fang X, et al (2010) De novo assembly of human genomes with massively parallel short read sequencing. Genome Res 20:265–272.

Li S, Hartman GL, Jarvis BB, Tak H (2002) A Stachybotrys chartarum isolate from soybean. Mycopathologia 154:41–49.

Li Y, de Magalhães J (2011) Accelerated protein evolution analysis reveals genes and pathways associated with the evolution of mammalian longevity. AGE:1–14. doi:springerlink. pp. 10.1007/s11357–011-9361-y.

Ma L-J, van der Does HC, Borkovich KA, Coleman JJ, Daboussi M-J, et al (2010) Comparative genomics reveals mobile pathogenicity chromosomes in Fusarium. Nature 2010, 464:367–373.

Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, et al (2011) CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Res 39:D225–229.

Mason JM, Arndt KM (2004) Coiled coil domains: Stability, specificity, and biological implications. ChemBioChem 5:170–176.

Massingham T, Goldman N (2012) All Your Base: a fast and accurate probabilistic approach to base calling. Genome Biol 13:R13.

Manev H, Uz T, Sugaya K, Qu T (2000) Putative role of neuronal 5-lipoxygenase in an aging brain. FASEB J 14:1464–1469.

McCormick SP, Alexander NJ (2002) Fusarium Tri8 encodes a trichothecene C-3 esterase. Appl Environ Microbiol 68:2959–2964.

McCormick SP, Stanley AM, Stover NA, Alexander NJ (2011) Trichothecenes: from simple to complex mycotoxins. Toxins (Basel) 3:802–814.

Orme D, Freckleton R, Thomas G, Petzoldt T, Fritz S, et al (2011) Accessed 19 Feb 2013. caper: Comparative Analyses of Phylogenetics and Evolution in R. Available:http://cran.r-project.org/web/packages/caper/index.html.

Palma GA, Argañaraz ME, Barrera AD, Rodler D, Mutto AA, Sinowatz F (2012) Biology and Biotechnology of Follicle Development. Sci World J 2012:938138. doi:10.1100/2012/938138.

Parra G, Bradnam K, Korf I (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 23:1061–1067.

Pei J, Grishin NV (2001) AL2CO: Calculation of positional conservation in a protein sequence alignment. Bioinformatics 17:700–712.

Pérez VI, Buffenstein R, Masamsetti V, Leonard S, Salmon AB, et al (2009) Protein stability and resistance to oxidative stress are determinants of longevity in the longest-living rodent, the naked mole-rat. Proc Nat Acad Sci USA 106:3059–3064.

Pestka JJ, Yike I, Dearborn DG, Ward MDW, Harkema JR (2008) Stachybotrys chartarum, trichothecene mycotoxins, and damp building-related illness: new insights into a public health enigma. Toxicol Sci 104:4–26.

Qu T, Uz T, Manev H (2000) Inflammatory 5-LOX mRNA and protein are increased in brain of aging rats. Neurobiol Aging 21:647–652.

Ranwez V, Delsuc F, Ranwez S, Belkhir K, Tilak M, et al (2007) OrthoMaM: A database of orthologous genomic markers for placental mammal phylogenetics. BMC Evol Biol 7:241.

Rose MR (1984) Laboratory evolution of postponed senescence in Drosophila melanogaster. Evolution 38:1004–1010.

Semeiks J, Grishin NV (2012) A method to find longevity-selected positions in the mammalian proteome. PLoS ONE 7(6):e38595. doi:10.1371/journal.pone.0038595.

Semeiks J, Borek D, Otwinowski Z, Grishin NV (2013) Comparative genome sequencing of the toxigenic black mold *Stachybotrys.* Manuscript submitted for publication.

Sempowski GD, Hale LP, Sundy JS, Massey JM, Koup RA, et al (2000) Leukemia inhibitory factor, Oncostatin M, IL-6, and stem cell factor mRNA expression in human thymus increases with age and is associated with thymic atrophy. J Immunol 164:2180–2187.

Shi C, Smith ML, Miller JD (2011) Characterization of human antigenic proteins SchS21 and SchS34 from Stachybotrys chartarum. Int Arch Allergy Immunol 155:74–85.

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol I (2009) ABySS: a parallel assembler for short read sequence data. Genome Res 19:1117–1123.

Speakman JR (2005) Correlations between physiology and lifespan – two widely ignored problems with comparative studies. Aging Cell 4:167–175.

Spiering MJ, Moon CD, Wilkinson HH, Schardl CL: Gene clusters for insecticidal loline alkaloids in the grass-endophytic fungus Neotyphodium uncinatum. Genetics 2005, 169:1403–1414.

Spröte P, Hynes MJ, Hortschansky P, Shelest E, Scharf DH, et al (2008) Identification of the novel penicillin biosynthesis gene aatB of Aspergillus nidulans and its putative evolutionary relationship to this fungal secondary metabolism gene cluster. Mol Microbiol 70:445–461.

Stetson PD, Ball EV, Mort M, Phillips AD, Shiel JA, et al (2003) Human Gene Mutation Database (HGMD): 2003 update. Hum Mutat 21:577–581.

Suhre K, Claverie J-M (2003) Genomic correlates of hyperthermostability, an update. J Biol Chem 278:17198–17202.

Voorhuis M, Broekmans FJ, Fauser BCJM, Onland-Moret NC, van der Schouw YT (2011) Genes involved in initial follicle recruitment may be associated with age at menopause. J Clin Endo Metab 96:E473–E479.

Walker M, Herndon J (2008) Menopause in nonhuman primates? Biol Reproduction 79:398–406.

Ward E, Parsons K, Holmes E, Balcomb III K, Ford J (2009) The role of menopause and reproductive senescence in a long-lived social mammal. Front Zool 6:4.

Williams GC (1957) Pleiotropy, natural selection, and the evolution of senescence. Evolution 11:398–411.

Williams LJ, Tabbaa DG, Li N, Berlin AM, Shea TP, et al (2012) Paired-end sequencing of Fosmid libraries by Illumina. Genome Res 22:2241–2249.

Wong P, Walter M, Lee W, Mannhaupt G, Münsterkötter M, et al (2011) FGDB: revisiting the genome annotation of the plant pathogen Fusarium graminearum. Nucl Acids Res 39:D637–D639.

Wu Z, Tsumura Y, Blomquist G, Wang X-R (2003) 18S rRNA gene variation among common airborne fungi, and development of specific oligonucleotide probes for the detection of fungal isolates. Appl Environ Microbiol 69:5389–5397.

Yancik R, Ries LA (1994) Cancer in older persons. Magnitude of the problem-how do we apply what we know? Cancer 74:1995–2003.

Yang Z (2007) PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol Biol Evol 24:1586–1591.

Yang Z, Bielawski JP (2000) Statistical methods for detecting molecular adaptation. Trends Ecol Evol 15:496–503.