Genome Analysis

The effect of domestication on the fungal proteome

Antonis Rokas

Vanderbilt University, Department of Biological Sciences, VU Station B 351634, Nashville, TN 37235, USA

The molecular effect of domestication has rarely been examined in fungi. I examined the molecular signature of domestication on *Aspergillus oryzae*, a fungus used to ferment several Japanese food products. The ratio of nonsynonymous to synonymous substitutions and the percentage of estimated deleterious substitutions along the *A. oryzae* lineage were lower than those in the wild sister lineage *Aspergillus flavus*. The patterns of genetic change observed in human-mediated domestication of animals and plants might not be typical in domesticated fungi.

The molecular signature of domestication

The transition to domestication has been frequently associated with a relaxation of selective constraints [1]. Domesticated organisms have usually been selected for beneficial traits conferred by a small number of genetic loci [2,3] and have undergone several rounds of population bottlenecks [4]. Both of these processes could result in genome-wide excesses of nonsynonymous and deleterious substitutions [5]. For example, domesticated rice strains have more nonsynonymous substitutions than wild rice strains and a greater frequency of substitutions toward radical amino acids [1]. Although numerous fungal species have been domesticated by humans for use in food production (e.g. *Saccharomyces* for bread, wine and beer; *Penicillium* for cheeses; see also Ref. [6]), most domestication studies have focused on plant and animal species [2,3,7,8]. I sought to quantify the effect of domestication on the fungal proteome, using the filamentous ascomycete *Aspergillus oryzae* as a case study.

The domestication of *Aspergillus oryzae*

A. oryzae was first isolated in 1876 by Ahlburg from 'koji', the material fermented by *A. oryzae* in solid-state cultivation [9]. Several lines of historical evidence indicate that *A. oryzae* has been used for at least two millennia for the production of *sake* (an alcoholic beverage from rice), *miso* (soy bean paste), *su* (vinegar) and *shoyu* (soy sauce) [9,10], its enzymes having key roles in the degradation of plant starch and proteins. Furthermore, specialized techniques enabling the production of large quantities of *A. oryzae* asexual spores uncontaminated from other microorganisms have been in use for industrial-scale production for hundreds of years [9].

Molecular data are consistent with the historical record. *A. oryzae* seems to have evolved by domestication from *Aspergillus flavus*, a wild-type species commonly found in soil and litter [12], several thousand years ago ([11]; A.R., unpublished). The two species differ in both quantitative (colony color and texture) and qualitative (production of the secondary metabolite aflatoxin) phenotypic characters [11], but are otherwise very similar. This close evolutionary proximity of the two species is reflected in their genomes, which are nearly identical in size and gene content [13,14]. The *A. oryzae* strain sequenced is RIB40 [15], which was isolated in 1950 and exhibits morphological, growth and enzyme production characteristics typical of those in industrial (patented) strains used for sake brewing [16]. The *A. flavus* strain sequenced is NRRL 3357 [13], which was isolated in 1963 (S. W. Peterson, personal communication).

A lower ratio of nonsynonymous to synonymous substitutions in the domesticated lineage

To evaluate the effect of domestication on the *A. oryzae* proteome, I identified all orthologous coding genes shared among *A. oryzae*, *A. flavus* and the outgroup species *A. terreus*, using the reciprocal best-blast hit criterion [17] (e-value cutoff of 1e-06; proteomes were downloaded from www.broad.mit.edu/annotation/genome/aspergillus_group/). The transcripts of the 6365 identified genes were translated into amino acids, aligned using CLUSTALW (http://www.ebi.ac.uk/clustalw) [18] and then reverse-translated back to codon-based alignments (all three steps were run automatically for each transcript using the TRANSALIGN [http://www.uni-oldenburg.de/molekularesystematik/en/34011.html] software [19]). Genes for which $\geq$20% of their sites were missing in any of the three species were excluded, resulting in a dataset of 4484 genes. The percentage of sites with missing data in these 4484 genes was $\leq$3.5% per species (similar results were obtained using a range of cutoff values for the percentage of sites missing).

Variation in selection pressure along branches of the three-species tree was examined using the CODEML module from PAML (http://abacus.gene.ucl.ac.uk/software/paml.html) [20]. Importantly, because the *A. oryzae* lineage seems to be monophyletic and nested within *A. flavus* [11], any gene from the three species will exhibit the same tree topology, although its branch lengths can differ. I have accounted for this difference in coalescence depth by estimating branch lengths for each gene, when possible. I first examined whether the ratio ω of nonsynonymous (dN) to synonymous (dS) substitutions was higher in the 4484 *A. oryzae* genes than in their *A. flavus* counterparts under the free-ratio model, where ω was free to vary along each lineage [21]. I found that the value for *A. oryzae* (ω = 0.45) was actually lower than the value for *A. flavus* (ω = 0.51). I next examined whether the ω ratio per gene was higher for *A. oryzae* genes than for *A. flavus* genes under the free-ratio model. Because *A. oryzae* and *A. flavus*

Corresponding author: Rokas, A. (antonis.rokas@vanderbilt.edu).

are very closely related, 3852 of the 4484 genes exhibited no synonymous substitutions in either of the two lineages (2471 genes in *A. oryzae* and 2546 in *A. flavus*) and were excluded. An additional, much smaller, number of genes also had to be excluded either owing to saturation of dS values (dS >1) or to unreliable ω values in either of the species (ω >10), resulting in a gene set of 529 genes (similar results were obtained by increasing both the dS and ω by an order of magnitude). Both median and average ω ratios were slightly higher for *A. oryzae* than for *A. flavus* (*A. oryzae*: median ω = 0.5, average ω = 1.21; *A. flavus*: median ω = 0.41, average ω = 0.96; $t = 2.05$, *d.f.* = 528, $p = 0.04$). Examination of the distribution of ω ratios for the 529 genes from the two species also indicates a lower number of *A. oryzae* genes with small ω values and a higher number of genes with large ω values.

I next examined whether a greater number of genes in *A. oryzae* were under relaxed purifying selection relative to its wild close relative *A. flavus*. Specifically, I evaluated the log likelihood of three alternative hypotheses relative to the null hypothesis H_0, under which all three lineages exhibited the same ω ratio (Figure 1). The first alternative hypothesis (H_1) stated that the ω ratio along the *A. oryzae* lineage was different from that in the other two lineages (Figure 1). To discriminate between genes that were consistent with a different ω ratio only along the *A. oryzae* lineage from genes that were consistent with distinct ω ratios in all three lineages, I also tested the second alternative hypothesis (H_2), in which each lineage exhibited its own ω ratio (i.e. the free-ratio model discussed in the previous paragraph), against the H_0 hypothesis (Figure 1). Finally, to evaluate whether the number of *A. oryzae* genes with a distinct ω was larger than would be expected in wild species, I also tested the third alternative hypothesis H_3, in which the ω ratio differed only across the *A. flavus* lineage, against the H_0 hypothesis (Figure 1).

The majority of the 4484 genes did not reject the null hypothesis H_0 of a uniform ω across all three lineages of the phylogeny for any of the three alternative hypotheses (all tests were done at $p = 0.01$ significance). 926 genes rejected H_0 for the H_1 hypothesis of elevated ω ratios only in the *A. oryzae* lineage, whereas 837 genes rejected H_0 for the H_3 hypothesis of elevated ω ratios only in the *A. flavus* lineage. The gene numbers for both lineages were larger than the number of genes expected to be significant due to random chance (4484 genes × 0.01 ≈ 45). However, the majority of these 926 and 837 genes also rejected H_0 in favor of the free-ratio model supported by H_2 hypothesis. Specifically, only 93 of the 926 genes supported H_1 over H_0, but rejected both H_2 and H_3 over H_0, and only 82 of the 837 genes supported H_3 over H_0, but rejected H_1 and H_2 over H_0. There was no significant difference in the numbers of genes that had significantly elevated ω ratios along *A. oryzae* versus those along *A. flavus* ($\chi^2 = 0.58$, $p = 0.45$). The very small number of genes with significantly elevated ω ratios only along *A. oryzae*, as well as the similar number of genes that exhibits significantly elevated ω ratios along *A. flavus*, are both inconsistent with the hypothesis that the *A. oryzae* proteome underwent a relaxation of selective constraints as a consequence of domestication.

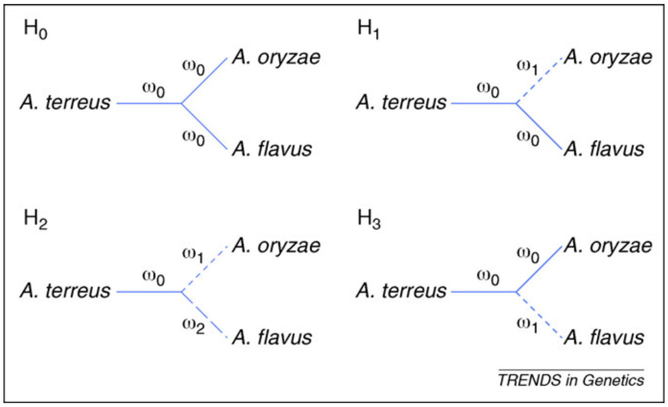

Figure 1. The four alternative hypotheses designed to test whether a greater number of genes in *A. oryzae* evolved under relaxed purifying selection. Under hypothesis H_0, all three lineages exhibit the same ω ratio of nonsynonymous to synonymous substitutions. Under hypothesis H_1, the ω ratio along the *A. oryzae* lineage is different from that in the other two lineages. Under hypothesis H_2, each lineage exhibits its own ω ratio. Finally, under hypothesis H_3, the ω ratio along the *A. flavus* lineage is different from that in the other two lineages.

A lower fraction of estimated deleterious substitutions in the domesticated lineage

If domestication resulted in the relaxation of purifying selection on the *A. oryzae* proteome, then I would expect to observe a higher frequency of deleterious substitutions and a lower frequency of neutral substitutions in the *A. oryzae* lineage. To test this hypothesis, I identified all 2753 orthologs across the genomes of eight *Aspergillus* species (*A. oryzae*, *A. flavus*, *A. terreus*, *A. niger*, *A. fumigatus*, *A. fischeri* and *A. nidulans*), using *Coccidioides immitis* as an outgroup [14,22]. To infer all derived substitutions to the *A. oryzae* and *A. flavus* lineages, I employed maximum likelihood ancestral state reconstruction as implemented in PAML [20]. I identified a total of 3551 substitutions that occurred with ≥90% probability between the *A. flavus*–*A. oryzae* last common ancestor and *A. oryzae* and 5540 substitutions that occurred with more than 90% probability between the *A. flavus*–*A. oryzae* last common ancestor and *A. flavus*. For each of these substitutions, I evaluated whether the derived amino acid was likely to have a neutral or deleterious effect on the function of the protein using the amino acid substitution prediction software ALIGN-GVGD (http://agvgd.iarc.fr/agvgd_input.php) [23]. ALIGN-GVGD takes advantage of the amino acid variation present at a particular site in an amino acid alignment to evaluate whether a particular amino acid substitution is likely to be neutral in its effects on protein function (e.g. a conservative substitution) or deleterious (e.g. a radical substitution). I found that 52% and 38% of substitutions occurring along the *A. oryzae* lineage were deleterious and neutral, respectively, with the remaining 10% of substitutions unclassified (Table 1). Interestingly, the percentage of substitutions classified as deleterious was higher in the *A. flavus* lineage (59%) and the percentage of neutral substitutions slightly lower (35%) than *A. oryzae* (Table 1) ($\chi^2 = 0.29$, $p = 0.59$). These data indicate that the *A. oryzae* proteome has not fixed a higher fraction of deleterious substitutions as a consequence of its domestication.

Comparisons of the *A. oryzae* proteome with the proteomes of the distantly related species *A. fumigatus* and

Table 1. Functional classification of derived substitutions in *A. oryzae* and *A. flavus*

Classification	*A. oryzae* substitutions (%)	*A. flavus* substitutions (%)
Deleterious	1863 (52%)	3254 (59%)
Neutral	1348 (38%)	1919 (35%)
Unclassified	340 (10%)	367 (6%)
Total	3551	5540

A. nidulans have revealed that *A. oryzae* has a substantially larger array of proteins involved in metabolism and secondary metabolism, a feature that probably contributes to the fermentation abilities of *A. oryzae* and that is proposed to have been important in its domestication [15]. However, virtually the same gene complement is found in *A. flavus*, the wild relative of *A. oryzae* [13,14], indicating that these gene set changes must have long preceded the domestication of *A. oryzae*. I further examined whether the 91 fast-evolving proteins in the *A. oryzae* lineage were potentially directly involved in its domestication, by evaluating their functional classification according to the FunCat scheme (http://mips.gsf.de/projects/funcat) [24]. The frequency of the fast-evolving *A. oryzae* proteins belonging to 20 different functional categories was not significantly different than the background frequency of all other proteins in the genome, including metabolism (Table S1 in the supplementary material online).

Explaining the molecular signature of fungal domestication

Domestication can lead to a relaxation of selective constraints. Examination of the proteome of *A. oryzae* indicates that more than two millennia of domestication have left no visible molecular signature. Intriguingly, similar conclusions have been reached in population genetic studies of the baker's yeast *Saccharomyces cerevisiae*. The 'domesticated' laboratory strain does show an elevated substitution rate [4,25], although comparison with wild strains shows that other strains exist that are faster-evolving than the laboratory one [25]. Furthermore, although genetic diversity is low in strains associated with wine and sake production, this is not so for other domesticated strains [26]. Finally, there are no clear patterns of association between the processes for which yeast strains have been domesticated and their genotypic relationships, with several wild isolates grouping within domesticated strain clusters [26], and multiple origins of domesticated strains [26,27] (see also Carter *et al.*; http://hdl.handle.net/hdl:10101/npre.12008.11988.10101).

Why might the fungal molecular signature of domestication be different to that of plants and animals? Consideration of the population genetics and ecology of fungal domestication events offers alternative, yet complementary, explanations to the observed lack of selective relaxation. From a population genetic perspective, it might be expected that domestication-associated bottleneck events in animals and plants have almost always been more severe than those in fungi. This is so because, in both plants and animals, one can select a single or very few individuals to be the parents of the next generation, whereas in fungi a population culture consisting of thousands to millions of individuals is usually selected. Furthermore, whereas reproductive isolation between domesticated lines and wild ones can easily be established in most plants and animals, such isolation is probably much more difficult to establish when dealing with fungal, and more generally in microbial, populations. It should also be noted that *A. oryzae* is among the 20% of fungal species [28] that are morphologically asexual, although the presence of mating loci, recombination and meiosis-associated genes all indicate that an as yet unidentified sexual generation exists [29,30]. *Saccharomyces* yeasts do have a sexual generation, but the overwhelming majority of their generations are asexual [29]. If the frequency of sexual reproduction in fungi is much more rare than in plants or animals, the efficacy of selection in fungal lineages would be reduced [30].

Alternatively, it could be reasoned that if the environments encountered in plant crop fields or in animal barns were more favorable than the environments encountered in the wild, a larger number of deleterious mutations could be tolerated in domesticated versus wild populations. Thus, the observed differences between fungal and plant or animal domestication events could be theoretically explained by underlying differences in domestication ecologies rather than by population genetics. Alternative explanations to the observed lack of selective relaxation notwithstanding, our findings indicate that the molecular patterns associated with the domestication of fungi might differ from the paradigm established by studies on domesticated animals and plants. Given the large number of fungal species that have been domesticated by humans for food production [6], further study of domesticated fungi promises to yield additional insights on the signature and impact of the domestication process on the fungal genome.

Acknowledgements

I thank Sean B. Carroll for bringing to my attention the work describing the Align-GVGD software, John G. Gibbons for running the analyses of the FunCat classification, and three anonymous referees for insightful comments. This work was conducted in part using the resources of the Advanced Computing Center for Research and Education at Vanderbilt University. Research in the Rokas lab is supported by the Searle Scholars Program and Vanderbilt University.

Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.tig.2008.11.003.

References

1 Lu, J. *et al.* (2006) The accumulation of deleterious mutations in rice genomes: a hypothesis on the cost of domestication. *Trends Genet.* 22, 126–131

2 Burger, J.C. *et al.* (2008) Molecular insights into the evolution of crop plants. *Am. J. Bot.* 95, 113–122

3 Doebley, J.F. *et al.* (2006) The molecular genetics of crop domestication. *Cell* 127, 1309–1321

4 Gu, Z. *et al.* (2005) Elevated evolutionary rates in the laboratory strain of *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. U. S. A.* 102, 1092–1097

5 Olson, M.V. (1999) When less is more: gene loss as an engine of evolutionary change. *Am. J. Hum. Genet.* 64, 18–23

6 Hesseltine, C.W. (1965) A Millennium of fungi food and fermentation. *Mycologia* 57, 149–197

7 Dobney, K. and Larson, G. (2006) Genetics and animal domestication: new windows on an elusive process. *J. Zool.* 269, 261–271

8 Cruz, F. *et al.* (2008) The legacy of domestication: accumulation of deleterious mutations in the dog genome. *Mol. Biol. Evol.* 25, 2331–2336

9 Machida, M. *et al.* (2008) Genomics of *Aspergillus oryzae*: learning from the history of Koji mold and exploration of its future. *DNA Res.* 15, 173–183

10 Baker, S.E. and Bennett, J.W. (2008) An overview of the genus Aspergillus. In *The Aspergilli: Genomics, Medical Applications, Biotechnology, and Research Methods* (Goldman, G.H. and Osmani, S.A., eds), pp. 3–13, CRC Press

11 Geiser, D.M. *et al.* (1998) Cryptic speciation and recombination in the aflatoxin-producing fungus *Aspergillus flavus*. *Proc. Natl. Acad. Sci. U. S. A.* 95, 388–393

12 Klich, M.A. (2002) Biogeography of *Aspergillus* species in soil and litter. *Mycologia* 94, 21–27

13 Payne, G.A. *et al.* (2006) Whole genome comparison of *Aspergillus flavus* and *A. oryzae*. *Med. Mycol.* 44 (Suppl), 9–11

14 Rokas, A. *et al.* (2007) What can comparative genomics tell us about species concepts in the genus *Aspergillus*? *Stud. Mycol.* 59, 11–17

15 Machida, M. *et al.* (2005) Genome sequencing and analysis of *Aspergillus oryzae*. *Nature* 438, 1157–1161

16 Abe, K. *et al.* (2006) Impact of *Aspergillus oryzae* genomics on industrial production of metabolites. *Mycopathologia* 162, 143–153

17 Koonin, E.V. (2005) Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.* 39, 309–338

18 Thompson, J.D. *et al.* (1994) Clustal-W - improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22, 4673–4680

19 Bininda-Emonds, O.R. (2005) transAlign: using amino acids to facilitate the multiple alignment of protein-coding DNA sequences. *BMC Bioinformatics* 6, 156

20 Yang, Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* 13, 555–556

21 Yang, Z. (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol. Biol. Evol.* 15, 568–573

22 Rokas, A. and Galagan, J.E. (2008) The *Aspergillus nidulans* genome and a comparative analysis of genome evolution in *Aspergillus*. In *The Aspergilli: Genomics, Medical Applications, Biotechnology, and Research Methods* (Goldman, G.H. and Osmani, S.A., eds), pp. 43–55, CRC Press

23 Mathe, E. *et al.* (2006) Computational approaches for predicting the biological effect of p53 missense mutations: a comparison of three sequence analysis based methods. *Nucleic Acids Res.* 34, 1317–1325

24 Ruepp, A. *et al.* (2004) The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Res.* 32, 5539–5545

25 Ronald, J. *et al.* (2006) Genomewide evolutionary rates in laboratory and wild yeast. *Genetics* 174, 541–544

26 Fay, J.C. and Benavides, J.A. (2005) Evidence for domesticated and wild populations of *Saccharomyces cerevisiae*. *PLoS Genet.* 1, 66–71

27 Legras, J.L. *et al.* (2007) Bread, beer and wine: *Saccharomyces cerevisiae* diversity reflects human history. *Mol. Ecol.* 16, 2091–2102

28 Taylor, J.W. *et al.* (2000) Phylogenetic species recognition and species concepts in fungi. *Fungal Genet. Biol.* 31, 21–32

29 Tsai, I.J. *et al.* (2008) Population genomics of the wild yeast *Saccharomyces paradoxus*: Quantifying the life cycle. *Proc. Natl. Acad. Sci. U. S. A.* 105, 4957–4962

30 Lynch, M. and Deng, H.W. (1994) Genetic slippage in response to sex. *Am. Nat.* 144, 242–261

0168-9525/$ – see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.tig.2008.11.003 Available online 10 December 2008

Genome Analysis

Methylation and deamination of CpGs generate p53-binding sites on a genomic scale

Tomasz Zemojtel*, **Szymon M. Kielbasa***, **Peter F. Arndt, Ho-Ryun Chung and Martin Vingron**

Department of Computational Molecular Biology, Max-Planck-Institute for Molecular Genetics, Ihnestrasse 73, D-14195 Berlin, Germany

The formation of transcription-factor-binding sites is an important evolutionary process. Here, we show that methylation and deamination of CpG dinucleotides generate *in vivo* p53-binding sites in numerous Alu elements and in non-repetitive DNA in a species-specific manner. In light of this, we propose that the deamination of methylated CpGs constitutes a universal mechanism for *de novo* generation of various transcription-factor-binding sites in Alus.

Methylated TEs as a source of transcription-factor-binding sites

The mobility of transposable elements (TEs) has been proposed to have an important role in spreading regulatory elements throughout the genome [1]. In mammals, most TEs are rendered silent by DNA methylation of cytosines in the context of CpG dinucleotides. Methylated cytosines can easily be converted to thymine residues via deamination and this mutational process has the highest rate among all base substitutions [2]. Therefore, it becomes an attractive hypothesis that these silenced TEs are a source of transcription-factor-binding sites generated by means of cytosine deamination-driven mutagenesis.

In vivo p53-binding sites in Alus

Until recently, no efficient technique was available for identification of transcription-factor-binding sites residing in TEs on a genome-wide scale. However, a recently proposed approach that combines chromatin immunoprecipitation (ChIP) with paired end tag (PET) sequencing made it possible to detect p53-binding sites in the repetitive portion of the human genome [3]. This study has

Corresponding author: Zemojtel, T. (zemojtel@molgen.mpg.de).
* These authors contributed equally