Tracing the Origin of the Fungal α1 Domain Places Its Ancestor in the HMG-Box Superfamily: Implication for Fungal Mating-Type Evolution

Tom Martin1, Shun-Wen Lu2*†a, Herman van Tilbeurgh3, Daniel R. Ripoll2*nb, Christina Dixielius1, B. Gillian Turgeon2, Robert Debuchy4,5*

1 Department of Plant Biology and Forest Genetics, Uppsala Biocenter, Swedish University of Agricultural Sciences (SLU), Uppsala, Sweden, 2 Department of Plant Pathology and Plant-Microbe Biology, Cornell University, Ithaca, New York, United States of America, 3 Univ Paris-Sud, Institut de Biochimie et de Biophysique Moléculaire et Cellulaire, UMR8619 Univ Paris-Sud CNRS, Orsay, France, 4 Univ Paris-Sud, Institut de Génétique et Microbiologie, UMR8621 Univ Paris-Sud CNRS, Orsay, France, 5 CNRS, Institut de Génétique et Microbiologie, UMR8621 Univ Paris-Sud CNRS, Orsay, France

Abstract

Background: Fungal mating types in self-incompatible Pezizomycotina are specified by one of two alternate sequences occupying the same locus on corresponding chromosomes. One sequence is characterized by a gene encoding an HMG protein, while the hallmark of the other is a gene encoding a protein with an α1 domain showing similarity to the Matα1p protein of Saccharomyces cerevisiae. DNA-binding HMG proteins are ubiquitous and well characterized. In contrast, α1 domain proteins have limited distribution and their evolutionary origin is obscure, precluding a complete understanding of mating-type evolution in Ascomycota. Although much work has focused on the role of the S. cerevisiae Matα1p protein as a transcription factor, it has not yet been placed in any of the large families of sequence-specific DNA-binding proteins.

Methodology/Principal Findings: We present sequence comparisons, phylogenetic analyses, and in silico predictions of secondary and tertiary structures, which support our hypothesis that the α1 domain is related to the HMG domain. We have also characterized a new conserved motif in α1 proteins of Pezizomycotina. This motif is immediately adjacent to and downstream of the α1 domain and consists of a core sequence Y-[LMIF]-x(3)-G-[WL] embedded in a larger conserved motif.

Conclusions/Significance: Our data suggest that extant α1-box genes originated from an ancestral HMG gene, which confirms the current model of mating-type evolution within the fungal kingdom. We propose to incorporate α1 proteins in a new subclass of HMG proteins termed MATα_HMG.

Introduction

Mating types in fungi display highly variable structural and content (Figure 1); in Ascomycota, they consist of dissimilar sequences occupying the same locus on the chromosome. These sequences are termed idiomorphs, to denote that they are not obviously related by structure or common descent [1]. All mating types are not idiomorphic, and there are examples in Zygomycota obviously related by structure or common descent [1]. All mating sequences are termed idiomorphs, to denote that they are not sequences occupying the same locus on the chromosome. These content (Figure 1); in Ascomycota, they consist of dissimilar sequences occupying the same locus on corresponding chromosomes. One sequence is characterized by a gene encoding an HMG protein, while the hallmark of the other is a gene encoding a protein with an α1 domain showing similarity to the Matα1p protein of Saccharomyces cerevisiae. DNA-binding HMG proteins are ubiquitous and well characterized. In contrast, α1 domain proteins have limited distribution and their evolutionary origin is obscure, precluding a complete understanding of mating-type evolution in Ascomycota. Although much work has focused on the role of the S. cerevisiae Matα1p protein as a transcription factor, it has not yet been placed in any of the large families of sequence-specific DNA-binding proteins.

Preparation of the manuscript.

Citation: Martin T, Lu S-W, van Tilbeurgh H, Ripoll DR, Dixielius C, et al. (2010) Tracing the Origin of the Fungal α1 Domain Places Its Ancestor in the HMG-Box Superfamily: Implication for Fungal Mating-Type Evolution. PLoS ONE 5(12): e15199. doi:10.1371/journal.pone.0015199

Editor: Geraldine Butler, University College Dublin, Ireland

Received August 8, 2010; Accepted October 29, 2010; Published December 8, 2010

Copyright: © 2010 Martin et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: TM is supported by Sida SWE-2005-453 (http://www.sida.se) and the Swedish University of Agricultural Sciences (Sveriges lantbruksuniversitet, SLU) (http://www.slu.se/sv/fakulteter/ml). SWL is supported by an NSF grant to BGT (http://www.nsf.gov/). CD is supported by the Swedish University of Agricultural Sciences (Sveriges lantbruksuniversitet, SLU) (http://www.slu.se/sv/fakulteter/ml). RD is supported by contract ANR-05-BLAN-0365 from the Agence nationale de la recherche (ANR) (http://www.agence-nationale-recherche.fr/). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: robert.debuchy@igmors.u-psud.fr
† These authors contributed equally to this work.

‡ Current address: USDA/ARS Cereal Crops Research Unit, Northern Crop Science Laboratory, Fargo, North Dakota, United States of America
§ Current address: BHSAI/MRMC, Frederick, Maryland, United States of America

α1 domain proteins were subsequently found to be ubiquitous in Ascomycotina [4,6]. The constant presence of an α1-box gene in one idiomorph constitutes the basis for mating-type nomenclature in self-incompatible (heterothallic) Pezizomycotina [7]. This gene is called MAT1-1-1 and defines the MAT1-1 idiomorph, while the other idiomorph called MAT1-2, is characterized by the presence of a MAT1-2-1 gene which encodes a transcription factor with a MATA_HMG domain. Although no α1 domain was identified in the mating-type P-specific polypeptide Pc of the fission yeast Schizosaccharomyces pombe (Taphrinomycotina) when its mating-type proteins were described initially [8], nor in subsequent work [9], limited similarity of the Pc protein to the α1 domain has been reported [9], prompting some authors to speculate that Pc might
be an α1-type protein [10]. Currently, αc is annotated as a HMG protein (e.g., Swissprot P10841), although neither the HMG nor the α1 classification has been evaluated rigorously in any publication. The exclusive presence of the α1 genes in MAT loci of Ascomycota (Pezizomycotina, Saccharomycotina and possibly Taphrinomycotina) prompts questions about mechanisms of acquisition and their ancestry.

In *S. cerevisiae*, Matα1p is a transcriptional co-activator essential for expression of α-specific genes in haploid α cells including those encoding the mating pheromone α-factor and the receptor for the opposite pheromone factor [11]. Matα1p is a pivotal protein which binds cooperatively with the MADS-box transcription factor [12], Mcm1p, and interacts with Ste12p [13] to activate transcription of α-cell specific genes. It has been suggested that the α1 domain may be involved in the physical interaction of Matα1p with Mcm1p [13]. More recently, the α1 domain has been shown to act as a degradation signal, suggesting that rapid turnover of Matα1p is important during yeast mating-type switching [14]. α1 proteins (MAT1-1-1) of Pezizomycotina are also required for mating-type specific transcription of pheromones and pheromone-receptors [4]. Taken together, these lines of evidence support the idea that α1 proteins are transcription factors which bind to DNA via the conserved α1 domain. To our knowledge, however, the relationship of the α1 domain to other DNA-binding domains has not been documented. As a consequence, it has not yet been placed in any of the large families of sequence-specific DNA-binding proteins. 

---

**Figure 1. Mating-type structure across the fungal kingdom.** α1, genes encoding transcription factors with an α1 domain; PPF, genes encoding proteins with a domain characterized by highly conserved proline and phenylalanine residues [41]; HMG, genes encoding transcription factors with an HMG domain; HD, genes encoding transcription factors with an homeodomain; OTHER, genes encoding proteins not relevant to this study. The standardized nomenclature [7] currently used for Pezizomycotina is indicated below the corresponding domains. +, gene present; +/-, gene present in some species from a group. Mating-type structures were compiled for the following species and corresponding references: *Saccharomyces cerevisiae*, *Kluyveromyces lactis*, *Candida albicans* and *Yarrowia lipolytica* [49,65], *Schizosaccharomyces pombe* [8], *Ustilago maydis* [66], *Phycomyces blakesleeanus* [17] and *Encephalitozoon cuniculi* [46]. The Pc gene from *S. pombe* was placed in the HMG class in agreement with the current classification of Pc protein (P10841) in Swissprot. Mating-type genes from *U. maydis*, *P. blakesleeanus* and *E. cuniculi* were placed arbitrarily in mating type 1 or 2.

doi:10.1371/journal.pone.0015199.g001
binding proteins that are referenced in transcription factor databases (e.g. TRANSFAC [15]) and the α1 domain profile (PDPC51325) in Prosite [16] does not cite a relationship to any well-known DNA binding domain family.

We present sequence comparisons, phylogenetic analyses of mating-type protein domains, and in silico predictions of secondary and tertiary structures, which support our hypothesis that the α1 domain is related to the HMG domain. This finding supports the current model for fungal mating-type evolution which links the appearance of the α1 box to a pre-existing HMG box.

Results and Discussion

The α1 and the HMG domains share conserved sequences

Certain sequence similarities between MATA_HMG and α1 proteins have been noticed previously [3,4,17], however whether this reflects functional analogy was not established. Furthermore, the origin of α1 in HMG has not been explicitly proposed before. Initially, to investigate whether there are similarities between the α1 and MATA_HMG domains, we analyzed a small dataset that included members of each and identified a core region present in both [See Materials and Methods, and Figure S1]. Next, a total of 5,773 sequence sets corresponding to both (See Materials and Methods, and Figure S1). Next, a total of 5,773 sequence sets corresponding to both (See Materials and Methods, and Figure S1). Next, a total of 5,773 sequence sets corresponding to both (See Materials and Methods, and Figure S1). Next, a total of 5,773 sequence sets corresponding to both (See Materials and Methods, and Figure S1). Next, a total of 5,773 sequence sets corresponding to both (See Materials and Methods, and Figure S1). Next, a total of 5,773 sequence sets corresponding to both (See Materials and Methods, and Figure S1). Next, a total of 5,773 sequence sets corresponding to both (See Materials and Methods, and Figure S1). Next, a total of 5,773 sequence sets corresponding to both (See Materials and Methods, and Figure S1). Next, a total of 5,773 sequence sets corresponding to both (See Materials and Methods, and Figure S1). Next, a total of 5,773 sequence sets corresponding to both (See Materials and Methods, and Figure S1). Next, a total of 5,773 sequence sets corresponding to both (See Materials and Methods, and Figure S1). Next, a total of 5,773 sequence sets corresponding to both (See Materials and Methods, and Figure S1). Next, a total of 5,773 sequence sets corresponding to both (See Materials and Methods, and Figure S1). Next, a total of 5,773 sequence sets corresponding to both (See Materials and Methods, and Figure S1). Next, a total of 5,773 sequence sets corresponding to both (See Materials and Methods, and Figure S1). Next, a total of 5,773 sequence sets corresponding to both (See Materials and Methods, and Figure S1). Next, a total of 5,773 sequence sets corresponding to both (See Materials and Methods, and Figure S1). Next, a total of 5,773 sequence sets corresponding to both (See Materials and Methods, and Figure S1). Next, a total of 5,773 sequence sets corresponding to both (See Materials and Methods, and Figure S1). Next, a total of 5,773 sequence sets corresponding to both (See Materials and Methods, and Figure S1). Next, a total of 5,773 sequence sets corresponding to both (See Materials and Methods, and Figure S1). Next, a total of 5,773 sequence sets corresponding to both (See Materials and Methods, and Figure S1). Next, a total of 5,773 sequence sets corresponding to both (See Materials and Methods, and Figure S1). Next, a total of 5,773 sequence sets corresponding to both (See Materials and Methods, and Figure S1). Next, a total of 5,773 sequence sets corresponding to both (See Materials and Methods, and Figure S1). Next, a total of 5,773 sequence sets corresponding to both (See Materials and Methods, and Figure S1). Next, a total of 5,773 sequence sets corresponding to both (See Materials and Methods, and Figure S1). Next, a total of 5,773 sequence sets corresponding to both (See Materials and Methods, and Figure S1). Next, a total of 5,773 sequence sets corresponding to both (See Materials and Methods, and Figure S1). Next, a total of 5,773 sequence sets corresponding to both (See Materials and Methods, and Figure S1). Next, a total of 5,773 sequence sets corresponding to both (See Materials and Methods, and Figure S1). Next, a total of 5,773 sequence sets corresponding to both (See Materials and Methods, and Figure S1). Next, a total of 5,773 sequence sets corresponding to both (See Materials and Methods, and Figure S1). Next, a total of 5,773 sequence sets corresponding to both (See Materials and Methods, and Figure S1). Next, a total of 5,773 sequence sets corresponding to both (See Materials and Methods, and Figure S1). Next, a total of 5,773 sequence sets corresponding to both (See Materials and Methods, and Figure S1). Next, a total of 5,773 sequence sets corresponding to both (See Materials and Methods, and Figure S1). Next, a total of 5,773 sequence sets corresponding to both (See Materials and Methods, and Figure S1). Next, a total of 5,773 sequence sets corresponding to both (See Materials and Methods, and Figure S1). Next, a total of 5,773 sequence sets corresponding to both (See Materials and Methods, and Figure S1). Next, a total of 5,773 sequence sets corresponding to both (See Materials and Methods, and Figure S1).

The α1 domain groups with the MATA_HMG domain group in phylogenetic analyses

A maximum likelihood phylogram was constructed using a selection of α1 and HMG core domains from representative taxa (Figure 3). LG+G and LG+I+G models [23] were found to best fit the data and produced almost identical phylogenetic trees. The α1 sequences clustered in a monophyletic clade (A in Figure 3) within the MATA_HMG domain sequence branch (B and E in Figure 3) (LR-ELW edge support = 85). The α1 and MATA_HMG domains clustered separately from SOX (C in Figure 3) and HMGB domains (D in Figure 3) (LR-ELW edge support = 76). Topology tests [24,25] also supported the proposed tree (KH P=1, SH P=1). This places the α1 core sequence specifically closer to fungal MATA_HMG sequences than to the other members of the HMG family. The sequence of the putative α1 domain of S. pombe Pc (Sch9p) did not group with α1 sequences but instead grouped with the Dothideomycete MATA_HMG sequences with extremely high support (LR-ELW edge support = 99). Sequences of Sordariomycetes and Leotiomycetes MAT1-1-3 proteins formed a subgroup (E in Figure 3) within MATA_HMG. The Dothideomycete MATA_HMG sequences were closer to MAT1-1-3 sequences (LR-ELW edge support = 74) than to MAT1-2-1 sequences. Interestingly, the Zygomycete P. blakesleeanus sexM (Phybl8) and sexP (Phybl9) sequences grouped with SOX and MATA_HMG, respectively, while the microsporidia sequences (F in Figure 3) grouped with HMGB (D in Figure 3).

Overall these data support the hypothesis that the genes encoding α1 and MATA_HMG proteins are evolutionarily related. The HMG domain is found in all eukaryotes with the HMGB, SOX and MATA_HMG domains all sharing a common ancestor [26]. The HMG domain was hypothesized to be the oldest with the SOX and MATA_HMG domain lineages arising later and confined to Metazoa and Fungi, respectively [26]. This places the root of all HMG domains within the HMGB group and allows us to map a direction of time onto the phylogram. MATA_HMG is not a monophyletic group without the inclusion of α1, therefore, because α1 is a subgroup of MATA_HMG we infer that MATA_HMG gave rise to α1.

Secondary and tertiary structure prediction of the α1 domain suggests it is a HMG domain

Sequence conservation between the α1 and HMG domains suggests that they may have similar secondary and tertiary structure. We first examined secondary structure predictions for the MATA_HMG domains from MAT1-2-1 and MAT1-1-3 mating-type proteins with Jpred3 [27]. The three alpha helices that characterize HMG domains [28,29,30] were predicted (Figure 4). We then analyzed secondary structures of α1 domains. All α1 domains tested displayed three alpha helices that coincide in position with those obtained with Sox2 (Figure 4), but α1 domains are characterized by a shorter helix 1 and 3, and a fourth alpha helix at the C-terminus. The α1 domain of the S. cerevisiae Matα1p also displayed these four alpha helices, in agreement with previous secondary structure prediction [14]. The putative α1 domain of S. pombe Pc also contained the four helices, however the second has no confidence support (see Figure 4).

Next, the proteins used for secondary structure prediction were submitted to Phyre for fold recognition [31]. As expected, the best matching templates for pezizomycotin MATA_HMG mating-type proteins (MAT1-2-1 and MAT1-1-3, see Figure 1) were known HMG template structures (Table 1). The α1 proteins also had best matching templates in HMG protein structures (Table 1). Likelihood of the homology is good (95%) and all tested α1
domains had the HMG family fold descriptor. Moreover, for all α1 proteins indicated in Table 1, the top ten highest scoring matches were to known HMG structures (see Table S2 for P. anserina FMR1, N. crassa mat A-1 and C. heterostrophus MAT1-1-1). These results strongly suggest that α1 has HMG structure. Although S. pombe Pc protein is classified as an HMG protein in

Figure 2. Conserved sequence of α1 and HMG domains. (A), WebLogo [19] representation of conserved sequences in α1, MATA-HMG, SOX and HMG domains respectively. The x-axis represents amino acid position from the N to C terminal. The amino acid labeled as ‘1’ is located at position 11-48 and 1-2 in the α1 and HMG domains, respectively (NCBI Conserved Domain Database accession numbers: pfam04769 and cd00084). Logos represent an ~40 amino acid core sequence of the DNA binding domain from 300 α1 domains, and 257 MATA_HMG, 3054 SOX_HMG and 2162 HMG_UFB_HMG-box domains. (B) Consensus core sequences produced from conserved amino acids in A. α1 protein domains divided into those of Pleosporales (α1-a) and Pezizomycotina without Pleosporales (α1-b). α1-a and α1-b are considered as one for identity scoring. Three or more identical amino acids among sequences are coloured blue while two or more identical or similar amino acids are coloured grey. Conservation among the five sequences is shown; a letter is used to represent three or more identical amino acids and an asterisk (*) for two identical or similar amino acids. (C) Ancestral core region for α1 and MATA_HMG. Core regions from 300 α1 domains and 257 MATA_HMG sequences were used.

doi:10.1371/journal.pone.0015199.g002
Figure 3. Unrooted phylogram for the HMG superfamily and the α1 domain core amino acid sequences. Clustering of core amino acid sequences using maximum-likelihood and model LG+G [67]. Labelling is as follows: α1 (A, green), MATA_HMG (B, yellow), SOX (C, orange), HMGB (D, blue), MAT1-1-3 subgroup of MATA_HMG (E, white), Microsporidia MAT sex locus HMG (F, white), Phycomyces blakesleeanus (Zygomycota) sexM (Phybl8) and sexP (Phybl9) are circled in purple. LR-ELW values above 70% are shown. Abbreviations: Ailme, Ailuropoda melanoleuca; Ajeca, Ajellomyces capsulatus; Altl, Alternaria alternata; Albr, Alternaria brassicicola; Anoga, Anopheles gambiae; Antlo, Antonospora locustae; Arath, Arabidopsis thaliana; Aspfu, Aspergillus fumigatus; Aspni, Aspergillus nidulans; Bipsa, Bipolaris sacchari; Botfu, Botrytis fuckeliana; Caee, Caenorhabditis elegans; Canal, Candida albicans; Cerel, Cervus elaphus yarkandensis; Ciosa, Ciona savignyi; Coche, Cochliobolus heterostrophus; Crypa, Cryphonectria parasitica; Culqu, Culex quinquefasciatus; Danre, Danio rerio; Dotpi, Dothistroma pini; Drome, Drosophila melanogaster; Enccu, Encephalitozoon cuniculi; Entbi, Enterocytozoon bieneusi; Fusac, Fusarium acaciae-mearnsii; Gibfu, Gibberella fujikuroi; Gibze, Gibberella zeae; Homxa,
Swissprot (P10841) and our phylogenetic analysis placed it closest to Dothideomycete MATA_HMG, the Pc protein has no significant support as an HMG domain (Table 1). We conclude that classification of Pc as an α1 or HMG protein sensu stricto is uncertain, although a relationship to HMG (and therefore to α1) is suggested by the phylogram (Figure 3). Additional examples from taphrinomycotinan species are needed to determine if they encode a new class of HMG-box genes.

To further search for structural homologs of the α1 domain we submitted the *N. crassa* α1 protein (mat A-1) sequence to the I-Tasser Structure Prediction Meta Server [32]. All best scoring templates for the α1 domain were structures of HMG proteins. When we iterated this search using Rosetta [33] and FUGUE [34], both predicted that the α1 domain has an HMG-like architecture (data not shown). In Figure 5 we show a model of the mat A-1 α1 domain superimposed upon the HMG domain of the
than those of the canonical HMG domain and thus may therefore be maintained upon DNA binding. Complexes have shown that the structure of the HMG-core is helix one is bent. Various structures of HMG-domain DNA helices are about the same length but the third one is much longer.

Non-structured peptide extensions are usually present at the N- and C-terminal ends. These peptides become ordered upon other. For complete names and accession numbers, see Table S4. 

transcription factor Sox2 in a ternary complex with an oligonucleotide and the POU DNA-binding domain of the OCT1 transcription factor [35]. HMG-box proteins have an L-shaped fold, comprising three alpha helices, stabilized by a hydrophobic core. Helix 3 and the N-terminal strand form the shaped fold, comprising three alpha helices, stabilized by a hydrophobic core. Helix 3 and the N-terminal strand form the OCT1 transcription factor [35]. HMG-box proteins have an L-domain should be able to bind DNA in a manner similar to canonical HMG domains. In support of this, we note that the DNA-binding core motifs for the N. crassa MATA_HMG mat a-1 and S. cerevisiae Mat1p are CAAAG [36] and CAATG [12], respectively.

3D-structures for a number of mammalian HMG-DNA complexes have been determined, including Sox2 [28] used in Figure 4, HMG-D [37], LEF-1 [38] and SRY [30]. In all cases, the HMG domain binds to the minor groove of DNA and introduces severe bends toward the major groove. Side chains from residues of helix 1 and helix 2 are inserted between base-pair stacks of the recognition sequence. However, the C-terminal region of each of these proteins interacts differently with its DNA target. For instance, for HMG-D, which binds DNA without sequence specificity, the C-terminal helix does not interact, while for LEF-1 it lies in the compressed major groove and stabilizes the bend conformation. Sequence specific HMG domains intercalate a hydrophobic residue between two bases of the \( A/T \)\( /A/T \)CAAAG [39] recognition sequence. These residues are either Met, Ile or Val (position 9 in Figures 2A and 4) and these are

| Query name (domain) | Fungus* | Template* (identity) | E-value * | Estimated precision | Fold/PDB descriptor |
|---------------------|---------|----------------------|-----------|---------------------|---------------------|
| FPR1 (HMG) | P. anserina | d2lefa (24%) | 2.8e-14 | 100% | HMG |
| mat a-1 (HMG) | N. crassa | d2lefa (18%) | 1.5e-14 | 100% | HMG |
| MAT1-2-1 (HMG) | C. heterostrophus | d2lefa (19%) | 5.6e-14 | 100% | HMG |
| MAT1-2-1 (HMG) | M. graminicola | d2lefa (30%) | 8.7e-15 | 100% | HMG |
| SMR2 (HMG) | P. anserina | d2lefa (25%) | 1.1e-14 | 100% | HMG |
| mat A-3 (HMG) | N. crassa | d2lefa (20%) | 9.9e-14 | 100% | HMG |
| MAT1-1-3 (HMG) | G. zeae | d2lefa (19%) | 1.8e-13 | 100% | HMG |
| MAT1-1-3/phb1 (HMG) | P. brassicae | d2lefa (23%) | 4.9e-15 | 100% | HMG |
| FMRI (x1) | P. anserina | d1qra (12%) | 0.005 | 95% | HMG |
| mat A-1 (x1) | N. crassa | d1qra (11%) | 0.028 | 95% | HMG |
| SMT A-1 (x1) | S. macrospora | d1qra (11%) | 0.0043 | 95% | HMG |
| MAT1-1 (x1) | M. oryzae | d1qra (14%) | 0.026 | 95% | HMG |
| MAT1-1 (x1) | C. parasitica | d1qra (10%) | 0.017 | 95% | HMG |
| MAT1-1 (x1) | D. sp | d2gza2 (14%) | 0.0022 | 95% | HMG |
| MAT1-1 (x1) | G. fujikuroi | d1qra (18%) | 0.014 | 95% | HMG |
| MAT1-1 (x1) | G. zeae | d1qra (15%) | 0.0052 | 95% | HMG |
| MAT1-1/pad1 (x1) | P. brassicae | d1qra (18%) | 0.0025 | 95% | HMG |
| MAT1-1 (x1) | A. fumigatus | d1qra (15%) | 0.012 | 95% | HMG |
| MAT1-1/MATB (x1) | A. nidulans | d1qra (15%) | 0.0016 | 95% | HMG |
| MAT1-1 (x1) | H. capsulatum | d1qra (14%) | 0.014 | 95% | HMG |
| MAT1-1 (x1) | C. heterostrophus | d1qra (15%) | 0.0013 | 95% | HMG |
| MAT1-1 (x1) | M. graminicola | d1qra (19%) | 0.0059 | 95% | HMG |
| Mat1p (x1) | S. cerevisiae | d1k99a (12%) | 0.0086 | 95% | HMG |
| Pc (HMG) | S. pombe | d2lefa (14%) | 4 | 45% | HMG |

*aFor complete names and accession numbers, see Table S4.

*bHighest scoring template to the query. Templates are known structures from the PHYRE fold library; d2lefa, lymphoid enhancer-binding factor, LEF1 from Mouse (Mus musculus); d1qra, HMG-D from Drosophila melanogaster; d2gza2, SRY from Human (Homo sapiens); d1k99a, nucleolar transcription factor 1 (Upstream binding factor 1, UBF-1) from Human (H. sapiens). The percentage sequence identity between the query and template is displayed in brackets. This is calculated relative to the shortest sequence.

*Hlikelihood of structural homology.

doi:10.1371/journal.pone.0015199.t001 |
MAT1-1-1 proteins contain a second conserved region in addition to the α1 domain

The alignment of the MATα1_HMG proteins reveals a conserved region spanning approximately 60 residues, immediately adjacent to and downstream of the fourth alpha helix of the MATα1_HMG domain in pezizomycotinid proteins (Figure 4). The region consists of a core conserved motif Y-[LMIF]-x(3)-G-[WL], and less conserved residues covering a larger region (Figure 6). S. cerevisiae, Pichia angusta and Candida albicans MATα1_HMG proteins stop 7, 14 and 15 residues, respectively, after the end of the MATα1_HMG domain and therefore do not include this 60 residue conserved region. Alignment of the 59 and 88 residues downstream of MATα1_HMG domain from Kluyveromyces lactis and Yarrowia lipolytica, respectively, failed to reveal the conserved region in these species (data not shown). Moreover, ScanProsite [40] did not detect the Y-[LMIF]-x(3)-G-[WL] motif in MATα1_HMG proteins of S. cerevisiae, P. angusta, C. albicans, K. lactis or Y. lipolytica. Taken together, these observations support the idea that this conserved region is specific to Pezizomycotina. Analysis of currently available MATα1_HMG proteins from Diaporthales indicates that the core consensus Y-[LMIF]-x(3)-G-[WL] is either modified or lost in this group, although the larger conserved region is present (Figure 6). Screening of entire Diaporthale sp. MATα1_HMG proteins [41] with ScanProsite failed to detect the core consensus motif. A similar search performed on C. paralisca protein [42] revealed the motif Y-L-N-L-A-G-T starting at position 106. Additional examples from diaporthale mating types are needed to determine a possible new core consensus motif. Conservation of this region was noted previously (and designated as HMGb) by Turgeon and Lu and reported in [43,44]. These authors hypothesized that it resembles an HMG domain. Prediction of HMG secondary structures with Jpred3 [27] and modelling with the I-Tasser Structure Prediction Meta Server [32], however, does not reveal the characteristic secondary and tertiary structures of HMG domains (data not shown). Further analyses will be necessary to establish the structure and origin of this region. Data obtained from mutations in the MATα1_HMG-box gene of N. crassa (mat A-1) suggest that this conserved region is necessary for male, but not female, fertility [45]. For the MATα1_HMG protein of C. heterostrophus, changing the conserved tryptophan (W) residue to alanine or arginine in the Y-L-N-L-A-G-T motif affects the number and development of pseudothecia, supporting the importance of this region for protein function (unpublished, Liu and Turgeon).

Mating-type evolution in the fungal kingdom

Idnurm and co-workers proposed that HMG domain proteins might represent the ancestral fungal sex determinant based on the discovery of HMG-box genes at the MAT locus in early diverged branches of fungi [17,46]. This model and subsequent analyses [47,48], however, do not explain the acquisition of α1-box genes in ascomycotan mating types. Low similarities between α1 and HMG domains have been noticed previously and a relationship suggested [3,4,17], although this contention has not been carefully examined. Sequence and phylogenetic analyses and structural modelling presented here substantiate the hypothesis that the evolutionary origin of α1 is in the HMG domain, thus providing a clue to the origin of the α1-box genes. This hypothesis is in agreement with the model proposed by Idnurm and co-workers [17]. However this model is strengthened by data which reveal linkage conservation of certain genes flanking the mating-type locus in Microsporidia and Ascomycota. A gene encoding a DNA lyase is immediately adjacent to MAT of many Ascomycota [43,49,50] (Figure 7). Remarkably, the analysis of the environment...
of the putative mating-type locus of Encephalitozoon cuniculi (Microsporidia) reveals the presence of an homolog of the DNA lyase encoding genes [46]. This gene is 7 kb away from the E. cuniculi putative MAT locus [51] (Figure 7) and analysis with FUNGIpath [52] confirmed that it is an ortholog of the DNA lyases genes adjacent to MAT loci in Ascomycotina. Although synteny sensu stricto is not conserved between Microsporidia and Ascomycota mating types, the presence of these orthologous DNA lyase encoding genes in the vicinity of the mating-type locus in Microsporidia and Ascomycota is highly significant and strongly supports a common origin.

Conclusion
The model proposed by Lee et al. [33] for early steps of mating type formation should result in idiomorphic or allelic sequences of a given mating-type locus containing phylogenetically related genes. The presence of MATa_HMG and MATA_HMG-box genes in ascomycotan opposite mating types (Figure 1) is in agreement with this model. Only a few mating types are an exception to this rule; ironically, the most prominent example is S. cerevisiae MAT, one of the most thoroughly characterized loci in terms of MAT regulation. It lacks the MATa2 (MATA_HMG-box) gene [49] (Figure 1), but has evolved alternative transcriptional circuits ensuring appropriate mating-type target gene expression [54].

The identification of the MATa_HMG structure is an additional example of a study confirming that protein spatial structure is more conserved than amino acid sequences (reviewed in [55]), as suggested first by Lesk and Chothia [56]. Functional conservation acts as a strong restraint limiting sequence and, even more, structural divergence [57]. It must be noted, however, that there are some differences between the predicted MATa_HMG structure and SOX2 folding, in particular the presence of a fourth alpha helix. Experimental determination of crystal structure of the MATa_HMG domain is in progress and should help in understanding the function of this additional structure.
helix. It is surprising that the MATA and MATα_HMG sequences are so much divergent, especially when paralogous MATA and MATα_HMG proteins encoded by opposite idiomorphs are considered. It is worth noting that the term of idiomorph was indeed proposed by Metzenberg and Glass in 1990 to denote that mating-type sequences “are not obviously related by structure or common descent” [1]. Further investigations will be necessary to identify the factors that favored MATα_HMG divergence and have thwarted the determination of its origin for such a long time.

**Figure 7.** Mating-type loci and DNA lyase gene position in representative species of Ascomycota. The DNA lyase orthologs are indicated only when confirmed by sequencing. The physical linkage of the MAT locus and the DNA lyase gene may be relaxed, as exemplified by Cochliobolus heterostrophus, where the two genes are separated by 181 kb. Orthology of DNA lyase genes was determined by FUNGIpath [52]. Mating-type structures were compiled for the following species and corresponding references: Cochliobolus heterostrophus [68], Leptosphaeria maculans [69], Mycosphaerella graminicola [70], Aspergillus fumigatus [71], Coccidioides immitis [72], Neurospora crassa [49,50], Podospora anserina [43], Magnaporthe oryzae [43], Gibberella fujikuroi [73], Gibberella zeae [50], Cordyceps takaomontana [74], Yarrowia lipolytica [49], Encephalitozoon cuniculi [46]. Circled figures on the left: 1: Dothideomycetes; 2: Eurotiomycetes; 3: Sordariomycetes; 4: Saccharomycetales; 5: Microsporidia. Linkage of C. heterostrophus MAT1-1 to DNA lyase gene (ESTEXT_GENEWISE1PLUS.C_40361) was determined from the sequence data produced by the US Department of Energy Joint Genome Institute http://www.jgi.doe.gov/. Linkage of G. fujikuroi (Fusarium verticillioides) MAT1-1-1 to DNA lyase gene (FVEG_02488) was determined from the version 1 sequence data produced by the Broad Institute http://www.broadinstitute.org/annotation/genome/fusarium_group/MultiHome.html. doi:10.1371/journal.pone.0015199.g007
Materials and Methods

Sequence acquisition

Initially, we retrieved and aligned ~200 residues from five α1 and ~75 residues from five MATA_HMG domains, from selected Ascomycetes (Figure S1). Alignment with Kalign [58] revealed a core region of ~40 amino acids with conserved signatures starting at position 1-2 and 11-48 in the MATA_HMG and α1 sequences, respectively (Figure S1). Sequences annotated as α1 (MAT_Alpha1) or HMG (MATA_HMG, SOX-TCF_HMG, or HMGB-UBF_HMG) in the NCBI database were collected. The core region of ~40 amino acids was aligned for all sequence sets using Muscle [18]; sequences with less than 80% coverage of the core were removed. HMGB-UBF HMG-domain sequences contained a small section of varying size within the core region that was removed to create a compact alignment with conserved sections only. The resulting core region dataset consisted of 300 α1 (Dataset S1), 257 MATA_HMG (Dataset S2), 3,054 SOX_HMG (Dataset S3) and 2,162 HMGB_HMG sequences (Dataset S4).

Identifying consensus amino acids

Conserved amino acids were estimated with WebLogo [19] using core region data sets. The resultant logos were taken as the consensus sequence for each of the domains. The α1 domain consensus was divided into two; one corresponded to the Plosporales and the second to α1 in all other Pezizomycotina. COMPASS was used for profile-profile analysis [21].

E-value computing

Alignments were performed using the NCBI BLASTP suite-2 tool [59].

Ancestral sequence prediction

Input for this were sequences corresponding to ascomycete α1 and MATA_HMG domains. The datasets contained domains from Sordariomycetes, Leotiomycetes, Eurotiomycetes, Dothideomycetes, Pezizomycetes, Saccharomycotina and Taphrinomycotina and represented a broad range of species. Sequences were input as independent HMG and α1 datasets. The predicted ancestral amino acid sequences of the ascomycete α1 and HMG domains were determined using the Ancescon ancestral protein predictor [22]. Statistical alignments were performed using the NCBI BLASTP suite-2 sequences [59].

Phylogenetic analysis

Randomly selected and certain selected core sequences from the α1 and HMG core region datasets were aligned using Kalign [58]. ProtTest v2.4 identified LG+G and LG+I+G as the best models for the data [23]. Trees were produced using both models with TREEFINDER using maximum likelihood, selected models and 10,000 replicates producing concurrent trees with the LG+G tree shown [60]. Phylogenograms were viewed using TreeView 1.6.6 [61]. Local rearrangement of expected likelihood weights (LR-ELW) edge support were used as confidence in configuration of branches [62]. Alternative topologies were tested using the KH and SH tests in TREEFINDER [24,25].

Structure prediction

Sequence alignments were obtained with ClustalW2 [63], colours with Jalview [64] and structure prediction with I-Tasser [27]. These tools were provided by EBI on http://www.ebi.ac.uk/services/. Fold recognitions, 3D structure predictions and motif searches were performed with Phyre [31], I-Tasser Structure Prediction Meta Server [32] and ScanProsite [40], respectively.

Orthologous gene analysis

The orthology of DNA lyase proteins was determined with FUNGIpath [52].

Supporting Information

Figure S1 Initial alignment of MATA_HMG and α1 domains used to identify a conserved core region. ClustalW2 [63] alignment of complete α1 and HMG domains from five α1 and five MATA_HMG sequences. Identical amino acids across all sequences are coloured blue, >5 identical or similar amino acids are coloured grey. Core region indicated with *. Accession numbers for MATA_HMG: Pyrenopeziza brassicae MAT1-2-1/phb2 (CAA06843), Neurospora crassa MAT1-2-1/mat a-1 (AAA33590), Mucor rouxii 1-2-1 (AAL30836), Podospora anserina MAT1-1-3/SMR2 (CAAS2051), Cochliobolus heterosporus MAT1-2-1 (CAAG4644). Accession numbers for α1: Podospora anserina FMR1 (CAAS519), N. crassa mat A-1 (AAC37478), Alternaria alternata (O94160), Cochliobolus ellisi (O99867), Fusarium oxysporum (OS93051).

Table S1 Accession numbers for proteins of Figure 3.

Table S2 Top ten scoring with PHYRE for selected α1 domains.

Table S3 Color scheme used for Jalview.

Table S4 Accession numbers for proteins of Figure 4, Table 1 and Table S2.

Dataset S1 α1 sequences used for α1 core region determination.

Dataset S2 MATA_HMG sequences used for HMG core region determination.

Dataset S3 SOX_HMG sequences used for HMG core region determination.

Dataset S4 HMGB_HMG sequences used for HMG core region determination.
Acknowledgments

We thank Dr. J. Sohlb erg for helpful suggestions on the phylogenet ic analysis and Evelyne Coppin for critical reading of the manuscript. The C achi to bas is heterostro ph us sequence data used in this study were produced by the US Department of Energy Joint Genome Institute http://www.jgi.doe. gov/in collaboration with the user community. The Gi bb erella fujikuri i (Fusarium verticilloides) DNA lyase gene was mapped with the sequence data provided by the Broad Institute http://www.broadinstitute.org/.

Author Contributions

Conceived and designed the experiments: CD BGT RD. Performed the experiments: TM SWL HvT DRR. Analyzed the data: TM SWL HvT DRR. Wrote the paper: TM HvT CD BGT RD.

References

1. Metzenberg RL, Glass NL (1990) Mating type and mating strategies in Neurospora. Bioessays 12: 53-59.
2. Casselton LA, Feldbruge M (2010) Mating and sexual morphogenesis in basidiomycete fungi. In: Borwick KA, Ebbole DJ, eds. Cellular and Molecular Biology of Filamentous Fungi. Washington, DC: ASM Press. pp 536-555.
3. Lee SC, Ni M, Li W, Sherotz C, Heiman J (2010) The evolution of sex: a feature RC (found) Mitgial IE, a multiple structure prediction method. Mol Biol Evol 74: 290-340.
4. Debuchy R, Ber teaux-Lecer lier V, Silar P (2010) Mating systems and sexual morphogenesis in Ascomycetes. In: Borwick KA, Ebbole DJ, eds. Cellular and Molecular Biology of Filamentous Fungi. Washington, DC: ASM Press. pp 573-587.
5. Astell CR, Ahlstrom-Jonasson L, Smith M, Tatchell K, Nasmyth KA, et al. (1981) The sequence of the DNA coding for the mating-type loci of Saccharomyces cerevisiae. Cell 27: 15-23.
6. Glass NL, Grothee nhein J, Metzenberg RL (1990) Neurospora crass a A mating-type region. Proc Natl Acad Sci U S A 87: 4912-4916.
7. Turg eon BG, Yoder OC (2000) Proposed nomenclature for mating type genes of filamentous ascomycetes. Fungal Genet Biol 31: 1-5.
8. Kelly M, Burke SA, Smith M, Klar A, Beach D (2000) Four mating-type gene control sexual differentiation in the fusion yeast. EMBO J 7: 1537-1547.
9. Niels en O, Friis T, Jerr ull S (1996) The Schizosaccharomyces pombe mat1 gene encodes an SRF/MCM1-related protein required for Pcll specific gene control. EMBO J 15: 385-395.
10. Coppin E, Debuchy R, Arnais e S, Picard M (1997) Mating types and sexual development in filamentous ascomycetes. Microbiol Mol Biol Rev 61: 411-420.
11. Tsong AE, Brian BT, John son AD (2007) Rewiring transcriptional circuitry: mating-type regulation in Saccharomyces cerevisiae and Candida albicans as a model for evolution. In: Hei tman J, Kron stadt JW, Taylor JW, Casselton LA, eds. Sex in Fungi, molecular determination and evolutionary implications. Washington, DC: ASM Press. pp 73-89.
12. Hagen DC, Bruhn L, Westby CA, Sprague GF, Jr. (1993) Transcription of alpha-specific genes in Saccharomyces cerevisiae DNA sequence requirements for activity of the coregulator alpha 1. Mol Cell Biol 13: 6866-6875.
13. Yuan YO, Stroke IL, Fields S (1993) Coupling of cell identity to signal response activity of the coregulator alpha 1. Mol Cell Biol 13: 6866-6875.
14. Kishino H, Hasegawa M (1989) Evaluation of the maximum likelihood estimate order in hominoidea. J Mol Evol 29: 170-179.
15. Sadreyev RI, Tang M, Klar A, Beach D (2001) Identification and cloning of TCF-1, a T lymphocyte-specific transcription factor containing a PROSITE signature match and ProRule-associated functional and structural residues in proteins. Nucleic Acids Res 34: W362-363.
16. Kanematsu S, Adachi Y, Ito T (2007) Mating-type loci of heterothallic Diaporth a: homologous genes are present in opposite mating-types. Curr Genet 52: 11-22.
17. McGuire CA, Marra RE, Turg eon BG, Milgroom MG (2001) Analysis of mating-type genes in the chestnut blight fungus, Cryphonectria parasitica. Fungal Genet Biol 34: 131-144.
18. Debuchy R, Turg eon BG (2006) Mating-Type Structure, Evolution and Function in Ascomycetes. In: Ku¨es U, Fischer R, eds. The Mycota I. 2 ed. BerlinHeidelberg: Springer-Verlag, pp 1-20.
19. Turg eon BG, Debuchy R (2007) C achi to bas is heterostro ph us and Podospora: mechanism of sex determination and the evolution of reproductive lifestyle. In: Heitman J, Kron stadt JW, Taylor JW, Casselton LA, eds. Sex in Fungi, molecular determination and evolutionary implications. Washington, DC: ASM Press. pp 93-121.
20. de Castro E, Sigrist CJ, Gustier A, Bulliard V, Land engendijk-Genevaux PS, et al. (2006) ScanProsite: detection of PROSITE signature matches and ProRule-assoc iated functional and structural residues in proteins. Nucleic Acids Res 34: W362-363.
21. Mcgu re CA, Marra RE, Turg eon BG, Milgroom MG (2001) Analysis of mating-type genes in the chestnut blight fungus, Cryphonectria parasitica. Fungal Genet Biol 34: 131-144.
22. Debuchy R, Turg eon BG (2006) Mating-Type Structure, Evolution and Function in Ascomycetes. In: Ku¨es U, Fischer R, eds. The Mycota I. 2 ed. BerlinHeidelberg: Springer-Verlag, pp 1-20.
23. Turg eon BG, Debuchy R (2007) C achi to bas is heterostro ph us and Podospora: mechanism of sex determination and the evolution of reproductive lifestyle. In: Heitman J, Kron stadt JW, Taylor JW, Casselton LA, eds. Sex in Fungi, molecular determination and evolutionary implications. Washington, DC: ASM Press. pp 93-121.
24. de Castro E, Sigrist CJ, Gustier A, Bulliard V, Land engendijk-Genevaux PS, et al. (2006) ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. Nucleic Acids Res 34: W362-363.
53. Lee SC, Corradi N, Doan S, Dietrich FS, Keeling PJ, et al. (2010) Evolution of the sex-related locus and genomic features shared in microsporidia and fungi. PLoS One 5: e10539.

54. Tsong AE, Tuch BB, Li H, Johnson AD (2006) Evolution of alternative transcriptional circuits with identical logic. Nature 443: 413–420.

55. Grishin NV (2001) Fold change in evolution of protein structures. J Struct Biol 134: 167–183.

56. Le Corre AM, Chothia C (1980) How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. J Mol Biol 136: 211–230.

57. Pascual-Garcia A, Abia D, Mendez R, Nido GS, Bastolla U. Quantifying the evolutionary divergence of protein structures: the role of function change and function conservation. Proteins 78: 181–196.

58. Lassmann T, Sonnhammer EL (2006) Kalign, Kalignvu and Mumsa: web servers for multiple sequence alignment. Nucleic Acids Res 34: W596–599.

59. Absalou SF, Madden TL, Schaller AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389–3402.

60. Jobb G, von Haeseler A, Strimmer K (2004) TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. BMC Evol Biol 4: 18.

61. Page RD (1996) TreeView: an application to display phylogenetic trees on personal computers. Comput Appl Biosci 12: 357–358.

62. Strimmer K, Rambaut A (2002) Inferring confidence sets of possibly misspecified gene trees. Proc Biol Sci 269: 137–142.

63. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, et al. (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23: 2947–2948.

64. Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ (2009) Jalview Version 2-a multiple sequence alignment editor and analysis workbench. Bioinformatics 25: 1189–1191.

65. Butler G, Rasmussen MD, Lin MF, Santos MA, Sakakihara S, et al. (2009) Evolution of pathogenicity and sexual reproduction in eight Candida genomes. Nature 459: 657–662.

66. Kahmann R, Rometis T, Bolker M, Kamper J (1995) Control of mating and development in Ustilago maydis. Curr Opin Genet Dev 5: 559–564.

67. Le SQ, Gascuel O (2008) An improved general amino acid replacement matrix. Mol Biol Evol 25: 1307–1320.

68. Wiese S, Houthwaite B, Yamauchi K, Yoder OC, Turgeon BG (1998) Single mating type-specific genes and their 3' UTRs control mating and fertility in Cochliobolus heterosporus. Mol Gen Genet 259: 272–281.

69. Cozijn AJ, Howlett BJ (2005) Characterisation of the mating-type locus of the plant pathogenic ascomycete Leptosphaeria maculans. Curr Genet 43: 351–357.

70. Waalwijk C, Mendes O, Verstappen EG, de Waard MA, Kema GH (2002) Isolation and characterization of the mating-type idiomorphs from the wheat leaf blotch fungus Mycosphaerella graminicola. Fungal Genet Biol 35: 277–286.

71. Rydhholm G, Dyer PS, Latoula F (2007) DNA sequence characterization and molecular evolution of MAT1 and MAT2 mating-type loci of the self-compatible ascomycete mould Neurospora fischeri. Eukaryot Cell 6: 868–874.

72. Fraser JA, Stajich JE, Tarcha EJ, Cole GT, Ingls DO, et al. (2007) Evolution of the mating type locus: insights gained from the dimorphic primary fungal pathogens Humanaspora capsulatus, Coccidioides immitis, and Coccidioides posadasii. Eukaryot Cell 6: 622–629.

73. Yun SH, Arie T, Kaneko I, Yoder OC, Turgeon BG (2000) Molecular organization of mating type loci in heterothallic, homothallic, and asexual Gibberella/Fusarium species. Fungal Genet Biol 31: 7–20.

74. Yokoyama E, Yamagishi K, Hara A (2003) Structures of the mating-type loci of Cordyceps takaomontana. Appl Environ Microbiol 69: 5019–5022.

75. DeLano WL (2002) Unraveling hot spots in binding interfaces: progress and challenges. Curr Opin Struct Biol 12: 14–20.

76. Palasingam P, Jauch R, Ng CK, Kolatkar PR (2009) The structure of Sex17 bound to DNA reveals a conserved bending topology but selective protein interaction platforms. J Mol Biol 388: 619–630.