ABSTRACT
The post-genome sequencing era heralds new avenues to explore the biology of organisms hitherto considered genetically intractable. While genomics is not typically thought of as hypothesis-driven science, comparative genomics presents an opportunity to generate many new hypotheses. Genome annotation predicts where the functional elements of the genome are located and provides a starting point for directed experiments. Comparative genomics is a filter to identify regions of the genome that have undergone particular evolutionary changes and helps to establish the genetic bases of phenotypic differences.

The kingdom Fungi is comprised of organisms that thrive in diverse habitats such as decaying organic matter or as plant and animal pathogens. Comparative genomics has helped the study of fungal evolution and provided a framework for identifying patterns of change in gene content and structure among species. A comparison of different fungi across lineages would illustrate how adaptive changes to habitats or lifestyles are manifested in the genome. Comparative genomics can also identify lineage-specific gene family expansions indicative of particular mechanisms for species adaptation. With the advancement of genomic technologies sequencing of around 80 fungal species are in the public domain and ~50 more sequences covering all the lineages are in various stages of completion. Several related fungal genomes have been sequenced so as to enable comparative analysis across a range of evolutionary distances. These clusters also include related organisms that differ in terms of specific physiological traits (i.e., pathogenicity), thus allowing these traits to be explored through comparison. The growing complement of fungal genome sequences enables other strategies for investigating fungal infection. Comparing genomes from nonpathogenic species to related pathogenic organisms can identify genetic differences
that contribute to infection and disease, while the comparison between strains with different host specificity may help clarify the genomic basis for differences in virulence and host interactions. Comparative analyses of these sorts are possible due to the sequencing of clusters of related genomes often centering on a pathogenic fungus, but including related nonpathogenic fungi as in the case of *Candida* spp. and *Saccharomyces* spp.

Based on morphology, physiology and nuclear ribosomal DNA (rDNA) sequences and orthologous protein sequences, the fungal kingdom has been classified into one subkingdom, Dikarya, and six other phyla, Neocallimastigomycota, Blastocladiomycota, Chytridiomycota, Glomeromycota, Microsporidia and Zygomycota. The Dikarya contains two phyla, Ascomycota and Basidiomycota. Ascomycota is a large and diverse group of species which includes a vast majority of pathogens of both plants and animals and many industrially useful organisms. A subclass of this group includes Hemiascomycetes, a monophyletic group, which have predominant yeast like morphology. Hemiascomycetes includes the *Candida* clade organisms that decode CUG codon as serine. The *Candida* clade includes opportunistic human pathogens like *Candida albicans*, *Candida dubliniensis*, *Candida tropicalis* and *Candida parapsilosis* and non-pathogenic organisms like *Debaryomyces hansenii* and *Lodderomyces elongisporus*. Phylogenetic analysis of *Candida* clade organisms showed these organisms are evolutionarily very close, however, their biology and virulence have diverged. Comparative analysis using genomic tools can be helpful to dissect the factors governing the differences and similarities among the clade organisms. The availability of ~80 fully sequenced fungal genomes provides an exciting opportunity to study the distribution of protein families across diverse set of organisms.
Transcription factors are modular proteins which have distinct domains for DNA binding, protein interaction, activation or repression of transcription and nuclear localization. Based on DNA binding domain sequences, these regulatory proteins are grouped into various families such as bZIP, bHLH, Zinc coordinating fingers and MYB. We constructed a local database, the DikarylProteome database by extracting translated protein sequences from 62 sequenced fungal genomes. The statistical representation of ~300 TF families, also known as hidden Markov models, was used to query the DikarylProteome database to identify known and diverged transcription factor families. The 47 TF families identified were not universally present in all genomes. We identified several TF families ABF1, BESS, CTA and GCM1 that are exclusive to certain genomes. A number of families not known to be present in fungal genomes were also identified. Metazoan-specific TF families like LAG1, CP2, GCM, SAM_PNT, the plant-specific WRKY and the bacterial-specific HTH_PSQ were identified from our database.

The TF family ABF1, represented by *S. cerevisiae* ARS binding factor, was selectively present in the *Saccharomyces* clade organisms and no orthologs were found in the other dikaryl proteomes examined. Similar to ABF1 family, the VHR1 family was found exclusively in Hemiascomycetes fungi and absent from other fungal lineages. However in the *Candida* clade organisms only one of the two paralogous VHR1 sequence was identified. An interesting TF group is of the WRKY-GCM1 superfamily which consists of AFT1, FAR1, WRKY, GCM1 and FLYWCH families of TFs. The AFT1 family is present only in Hemiascomycetes genomes, FAR1 and WRKY are known to be plant-specific families, while FLYWCH and GCM1 are metazoan families. These families share an unusual C2H2 zinc finger and the family members are known to have conserved four-stranded fold. Our analyses of this
superfamily identified AFT1, FAR1, WRKY and GCM1 TFs in the dikaryl genomes examined. AFT1 like domains were identified in the Saccharomyces and Candida genomes, while WRKY and GCM1 were mainly confined to the Zygomycetes lineage. Interestingly the FAR1-like TFs were found across several fungal lineages except the Saccharomyces clade organisms. Surprisingly we identified nine FAR1 domain containing sequences in the C. albicans genome. Comparative analysis showed that even highly related Candida dubliniensis differed with respect to the number of FAR1 proteins and sequence conservation in the domain. Only one C. albicans FAR1 domain containing protein orf19.4959 was present in all Candida clade organisms. Our analysis of FAR1 identified orf19.5888/CaRBF1 as a hit and multiple sequence analysis showed that this protein likely carried a WRKY-GCM1 fold. Phylogenetic analysis of FAR1 domain sequences shows that orf19.5888 and its orthologs in Candida clade clustered together and were connected to orf19.4959 group. CaRBF1 is a known transcription regulator in Candida albicans and is known to regulate the filamentation pathway.

The TF repertoire of Candida clade organisms were compared in detail using reciprocal best hit analysis. Comparative analysis between C. albicans and S. cerevisiae TFs showed that 189 are shared, which is ~59.6% of C. albicans and ~71% of S. cerevisiae total TFs. The analysis of very closely related C. dubliniensis showed that its TF repertoire is 99.7% identical to that of C. albicans. The number of TF families identified in C. albicans was same among other clade members however when compared to S. cerevisiae the Candida clade seemed to gain few families like CTA, FAR1, HTH_PSQ, IRO1 etc with the loss of ABF1 family only. It was striking to note a major expansion of CENPB, CTA and FAR1 in the genome of C. albicans, but not in other Candida clade organisms.
Detailed analysis of basic helix loop helix family identified two *C. albicans* factors orf19.6824 and orf19.921 to be *Candida* clade-specific, as it had no orthologs in other fungal genomes. The analysis of GATA factor family in the genomes of *Candida* clade and the *Saccharomyces* clade organisms showed the presence of eleven subfamilies. The families ECM23/SRD1 and GAT3/GAT4 were completely absent from the genomes of *Candida* clade organisms. Four families GAT5, GAT8, GAT9 and SFU1 were present only in the *Candida* clade organisms. The *Candida* clade family SFU1 (Ca orf19.4869 and its orthologs) belong to two-domain GATA factors. The *C. albicans* orf19.4869/Sfu1p was shown to be involved in regulation of iron homeostasis and its orthologs are present in other Ascomycetes and Basidiomycetes genomes. The families GAT5 (Caorf19.1150 and its orthologs) and GAT8 (Caorf19.4056 and its ortholog) represent canonical uncharacterized GATA TFs (Table V.2). The subfamily GAT9 consists of *C. albicans* Caorf19.4301/GAT9 and its orthologs in *Candida* clade organisms. The GATA DNA binding domain of GAT9 subfamily is non-canonical in that it contains a large insert in the zinc finger domain (see Appendix III). Like other GATA domains, the GAT9 family DBD has four conserved cysteine residues. However unlike canonical GATA factors, the insert between first and second cysteines was four amino acids (instead of two), while the insert region between second and third cysteines was 40 amino acids (instead of 17–20 residues). PSI-BLAST analysis with whole protein identified *S. cerevisiae* Spt21p and AMS2 and AMS2-like proteins from the genomes of *S. pombe, A. nidulans, M. grisea* and many other genomes. This analysis showed that the homology region between all the hits was confined to the N-terminal region only. Interestingly both ScSpt21p and SpAms2p are known to be involved in regulation of histone gene expression. It is also
known that SpAms2p recruits CenpA protein (a variant H3 histone) at the core centromere for maintaining functional centromeres.

Our bioinformatic analysis identified Gat9p as unique TF in *C. albicans* genome. We decided to characterize and identify the function of this gene in *C. albicans* using molecular genetic approaches. The *GAT9* gene was found to be essential for growth of *C. albicans*, because attempts to create null mutants persistently lead to trisomy. To examine the function of Gat9p depletion, we prepared *C. albicans* strains bearing regulatable *GAT9* alleles. We replaced the native promoter of *GAT9* with the maltose-inducible and glucose-repressible *MAL2* promoter. We found that the *GAT9* mRNA and protein levels in these strains are highly induced by maltose and repressed by glucose. The depletion of Gat9p resulted in pseudohyphae under conditions that favour normal yeast-form growth and consequently the colonies exhibited rough and wrinkled morphology. We examined the phenotypes of the strains in glucose and maltose media under different stress conditions. Our phenotype analysis showed that Gat9p is required for high temperature tolerance and specifically cadmium resistance. Recent studies in *S. cerevisiae* and humans showed that cadmium caused excessive DNA repair errors, in particular, the Okazaki fragment repair during DNA replication. These data suggested that Gat9p likely functions as a sensor of DNA damage or repair. Analysis of histone gene transcription showed that the expression of histone H2A and H2B genes may not be under control of Gat9p. It would be interesting to study the role of this protein in centromere organization and to understand the regulatory role, if any, in *C. albicans* cell cycle. Additional clues to the function of Gat9p would be obtained by examining protein-protein interactions of Gat9p and from whole genome expression profiling.