Describing Genomic and Epigenomic Traits Underpinning Emerging Fungal Pathogens

Rhys A. Farrer[1], **Matthew C. Fisher**
Imperial College London, London, United Kingdom
[1]Corresponding author: e-mail address: rfarrer@broadinstitute.org

Contents

Abstract

An unprecedented number of pathogenic fungi are emerging and causing disease in animals and plants, putting the resilience of wild and managed ecosystems in jeopardy. While the past decades have seen an increase in the number of pathogenic fungi, they have also seen the birth of new big data technologies and analytical approaches to tackle these emerging pathogens. We review how the linked fields of genomics and epigenomics are transforming our ability to address the challenge of emerging fungal pathogens. We explore the methodologies and bioinformatic toolkits that currently exist to rapidly analyze the genomes of unknown fungi, then discuss how these data can be used to address key questions that shed light on their epidemiology. We show how genomic approaches are leading a revolution into our understanding of emerging fungal diseases and speculate on future approaches that will transform our ability to tackle this increasingly important class of emerging pathogens.

Advances in Genetics
ISSN 0065-2660
https://doi.org/10.1016/bs.adgen.2017.09.009

1. INTRODUCTION

The fungal kingdom, diverged from the animal and plant kingdoms around 1.5 million years ago (Wang, Kumar, & Hedges, 1999), is globally ubiquitous and taxonomically diverse with between 1.5 and 5 million species estimated to exist (Blackwell, 2011). Recent phylogenetic classifications (Hibbett et al., 2007; Spatafora et al., 2016) currently group fungi into eight separate phyla, with the zoosporic fungi (Cryptomycota, Chytridiomycota, and Blastocladiomycota) comprising the earliest lineages alongside the Microsporidia. The four remaining phyla include the Zoopagomycota, Mucoromycota, and the "Dikarya higher fungi," comprising the phylum Ascomycota and Basidiomycota. Spanning the breadth of the fungal kingdom are pathogenic fungi that infect animals, plants, and other fungi. Importantly, increasing numbers of fungi are emerging as aetiological agents of disease by either exhibiting newly acquired or increased pathogenicity, or invading new ecological niches (geographically or to new host species), or both (Cushion & Stringer, 2010; Fisher, Gow, & Gurr, 2016; Longo, Burrowes, & Zamudio, 2014).

Emerging fungal pathogens (EFPs) are infections that are rapidly increasing in their incidence, geographic or host range, and virulence (Morse, 1995). This class of pathogens are known to pose an increasing threat to the health of plants, humans, and other animals (Fisher et al., 2012; Fones, Fisher, & Gurr, 2017). Recently highlighted examples include the newly described chytrid fungus *Batrachochytrium salamandrivorans* causing rapid declines of fire salamanders across an expanding region of northern Europe (Martel et al., 2014; Stegen et al., 2017), the basidiomycete fungus *Puccinia graminis* f. sp. *tritici* (Ug99 race) now threatening wheat production and food security worldwide (Singh et al., 2011), the basidiomycete fungus *Cryptococcus gattii* expanding its range into nonendemic environments with a consequential increase of fatal disease in humans (Byrnes et al., 2010; Fraser et al., 2005), and the emergence of *Candida auris* in intensive care units worldwide (Chowdhary, Sharma, & Meis, 2017). The global threat of these and other related diseases is underpinned by fungi harboring complex, recombinogenic and dynamic genomes (Farrer, Henk, Garner, et al., 2013; Fisher et al., 2012). Genomic variability drives rapid macroevolutionary change that can overcome host defenses and allow colonization of new environments. Novel genetic diversity also leads to the genesis of new independently evolving pathogenic lineages. Consequently, there is a clear and

urgent need to understand the mechanisms that drive the evolution of the phenotypic traits that underlie the virulence, pathogenicity, and geographic/host spread of EFPs.

EFPs of wildlife are generally detected following the observation of (initially "enigmatic") mass mortalities and species declines. For instance, population monitoring by ecologists led to the discovery of panzootic chytridiomycosis caused by novel species of *Batrachochytrium*, and bat white-nose syndrome caused by the novel species *Pseudogymnoascus destructans* (Blehert et al., 2009). In contrast, ongoing surveillance and genotyping of crop pathogens are used to detect and map the spread of phytopathogenic fungi and their lineages as they spread via trade and transportation, such as recently occurred with the spatial emergence of wheat blast *Magnaporthe oryzae* in Bangladesh (Islam et al., 2016). Crucially, in both animal and plant systems, rapid genome sequencing is essential to gain a greater understanding of the taxonomy, epidemiology, and evolutionary biology of EFPs and to inform possible mitigation efforts. A growing body of evidence is also meanwhile accumulating to show that epigenenomic processes (such as differential expression (Kuo et al., 2010), nucleosome positioning (Leach et al., 2016), and nucleic acid modifications (Jeon et al., 2015)), alongside genomic processes, influence both host and pathogen phenotypes. For example, in the aggressive phytopathogen *Botrytis cinerea*, small RNAs invade host cells and silence host immunity by hijacking the host RNA interference (RNAi) machinery leading to a virulent host/pathogen interaction (Weiberg et al., 2013). In this review, we discuss the experimental methodologies, and the discoveries they have enabled, that use genome variation within and between populations of EFPs, with a focus on future threats and the genomic resources that are needed to tackle them. Additionally, we discuss the methods and results emerging from experiments characterizing epigenomic variation within and between populations of EFPs, and show how this emerging field will contribute to a more nuanced understanding of the epidemiology of these infections. The toolkits and methodologies that we cover in this review are summarized in Fig. 1.

2. CHARACTERIZING GENOME VARIATION WITHIN AND BETWEEN POPULATIONS OF EFPS

Genome variation ultimately manifests, postsequencing, through the use of bioinformatics, where two or more individuals have subsections of their DNA aligned and compared, revealing single base changes (indicative

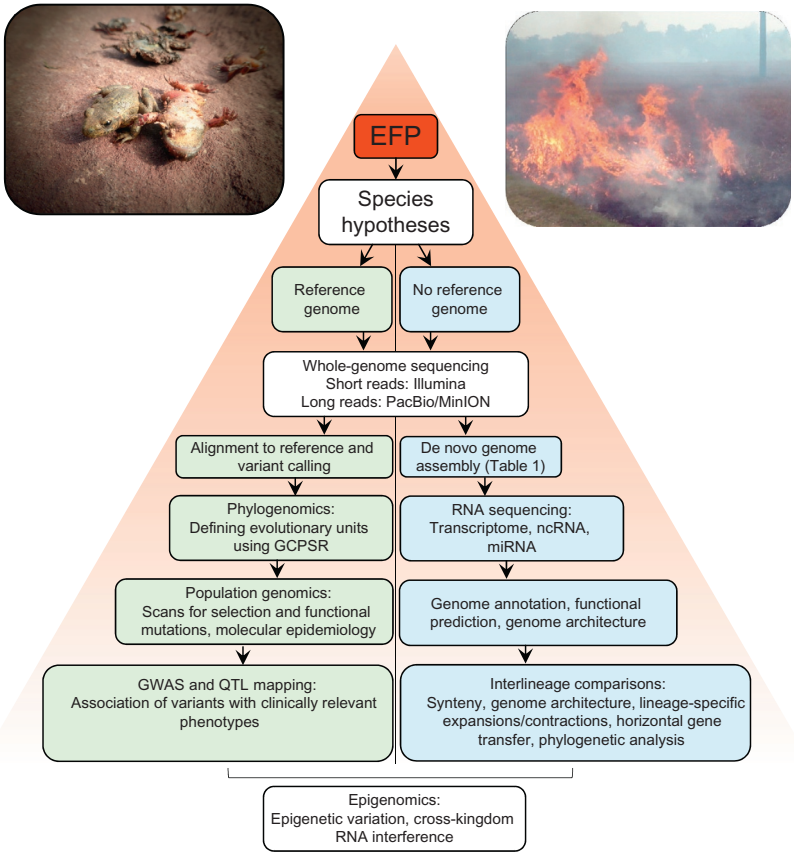

Fig. 1 A generalized workflow detailing the use of genomics to understand the genetic basis that underpins a novel EFP. *Images*: midwife toads with fatal chytridiomycosis caused by *Batrachochytrium dendrobatidis* (M.C. Fisher) and burning of a severely wheat blast (*Magnaporthe oryzae Triticum*) affected field in Meherpur district in Bangladesh, February 2016 (T. Islam, BSMR Agricultural University, Bangladesh).

of point mutations in one or more of the individuals), insertions and deletions (indels), and recombination (shuffling of sequences within and between genomes). Longer alignments and sequencing many times over (such as is often the case with next-generation and third-generation sequencing platforms) are required to identify additional features of genome variation. For example, changes in the depth of sequencing can suggest loss or gain of copy number variation (CNV) for single genes (gene duplication), regions (segmental aneuploidies), or entire chromosomal CNV (chromosomal aneuploidy; Fig. 2B) (Farrer, Henk, Garner, et al., 2013). Other

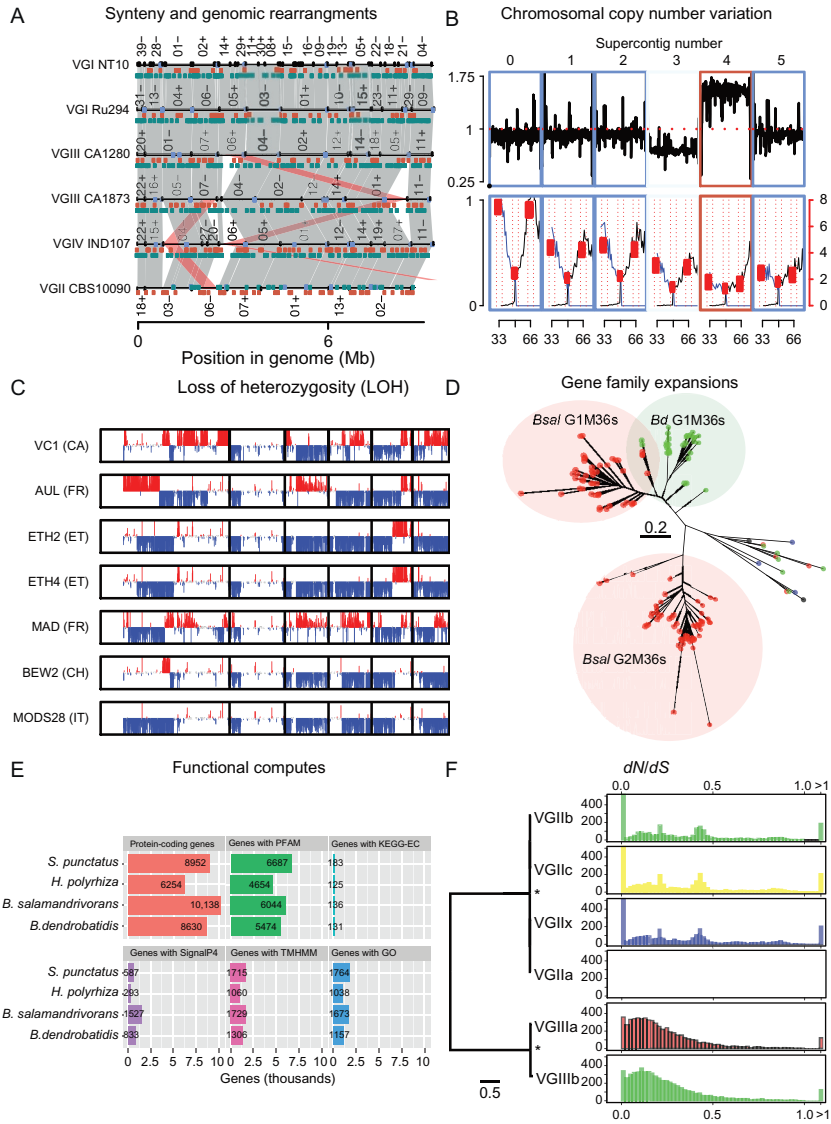

Fig. 2 Examples of genomic features that can be detected in EFPs. (A) Synteny and genomic rearrangements between and within two lineages of *C. gattii* (Farrer et al., 2015), (B) chromosomal copy number variation (CCNV) in *B. dendrobatidis* detected by average read depth from alignments (*top*) and allele frequencies (percent of bases agreeing with reference base vs tally in kilobases) (Farrer et al., 2011), (C) loss of heterozygosity in *B. dendrobatidis* detected using nonoverlapping sliding windows of SNPs minus heterozygous positions (*red*, predominately SNPs; *blue*, predominately heterozygous) (Farrer et al., 2011), (D) gene family expansion in *Batrachochytrium* spp. (Farrer et al., 2017), (E) gene annotation counts of gene types, and functional computes in *Batrachochytrium* spp. (Farrer et al., 2017), (F) measures of selection (i.e., d_N/d_S) across subclades of *C. gattii* using only fixed differences compared to VGIIa (Farrer et al., 2015).

examples include subsections of DNA that are reoriented to one another (inversions), subsections of DNA that occur in different locations in two individual genomes (translocations; Fig. 2A), genetic mosaics of two species in a single isolate (hybridizations) (Rhodes, Desjardins, et al., 2017), reduction of heterozygosity (gene conversion or loss of heterozygosity; Fig. 2C) (Farrer, Henk, Garner, et al., 2013), and changes to the ordering of genes (synteny). From these sources of genomic variation in fungal pathogens, epidemiological features of the outbreak can be inferred, virulence factors identified, and diagnostics and treatments devised.

Each source of genomic variation has unique and cumulative sources of uncertainty. These variants require careful detection and minimization, where possible. First, the quality of the sequence data can be highly variable between experiments, library-building protocols and different sequencing machines, containing low-throughput/depth sequencing, high levels of error in the base calls, or unexpected laboratory contamination (such as bacterial or host DNA). Uncontaminated high-quality samples may then be aligned to distantly related genomes resulting in decreasing accuracy of alignments and base calling. The quality of reference genomes themselves is variable, and they may contain inaccurate reference sequences (misassembled or containing sources of the prementioned errors), which can result in misleading comparisons. Second, variant calling from alignments against reference sequences may contain mistaken assumptions about ploidy, or inaccurately called bases. Alternatively, sequenced reads can be assembled into longer contiguous sections of chromosomes, which themselves can contain inaccurately assembled contigs or scaffolds. These are especially prone to occur over repetitive content or sequencing errors. From these assemblies, gene calling is often performed, which itself may include mistakes in intron/exon boundaries, and often absent or partial 5′ and/or 3′ untranslated regions (UTRs), for example. From comparisons of these gene predictions, analysis of patterns of natural selection could potentially identify unusually evolving genes that are artefacts caused by the aforementioned sources of errors. Fortunately, each of these errors has a range of hallmarks and remedial bioinformatics processes that can be used to ensure their accuracy or minimize those sources of error.

In the following sections, we will discuss the methods for identifying different sources of genomic variation with a focus on EFPs, and the manifestation of this variation within populations (population genetics approaches), and between populations (comparative genomics approaches). Importantly, we will be distinguishing between subgenomic approaches (PCR fingerprinting, microsatellites, restriction fragment length polymorphisms, etc.)

and whole-genomic approaches, focusing entirely on the latter. While subgenomic approaches are undoubtedly useful for characterizing EFPs (e.g., Hsueh et al., 2000; Mohammadi et al., 2015), approaches that are based on using whole genomes are increasingly being used for the detection and rapid characterization of novel pathogens (Hasman et al., 2014; Lecuit & Eloit, 2014). Indeed, beyond the identification of either a known or unknown fungus, the usage of full genome data provides far greater insights into the pathogens evolutionary history, population structure, and repertoire of virulence effectors.

2.1 Assemblies, Alignments, and Annotation

Many EFPs will be initially classified or identified based on their morphological traits, or host species, such as occurred with the amphibian-infecting chytrids *Batrachochytrium dendrobatidis* and *B. salamandrivorans* (Berger et al., 1998; Martel et al., 2013). Initially, subgenomic approaches using a taxonomic marker gene such as analysis of ribosomal DNA (rDNA) (Schoch et al., 2012) against global databases of known fungal sequences such as UNITE (Kõljalg et al., 2013) or the Ribosomal Database Project (RDP; Cole et al., 2014) are needed to define operational taxonomic units (OTUs). OTUs (also called "species hypotheses") are proxies for classically defined species and are used to ordinate the novel EFP taxonomically in the kingdom Fungi—bearing in mind however that the *Microsporidia* do not have the canonical rDNA structure (Dong, Shen, Xu, & Zhu, 2010). Subsequently, genome assembly (assembly de novo) is needed to provide a thorough examination of its genetic makeup and relatedness to known species.

Ideally, assembling a genome de novo will be preplanned, by implementing a long-read technology (such as third-generation sequencing platforms Oxford Nanopore's MinION, or Pacific Biosciences' single molecule real-time sequencing). Alternatively, multiple sequenced paired-end libraries of Illumina with short- and long (also known as "jump") insert sizes can be used by assembly tools optimized for such datasets (such as Allpaths; Butler et al., 2008). Other options for generating a high-quality assembly are the use of Fosmid libraries (fragmenting the genome then cloning into *E. coli*, and sequencing individual libraries separately) or constructing an optical map (a high-resolution restriction map of the genome to aid in assembling subsections of the genome).

Many assembly tools have been developed (Table 1), which may be optimized for different sequencing technologies (e.g., Allpaths for two libraries of paired-end Illumina (Butler et al., 2008), Canu for long reads such as

Table 1 Names, Versions and Descriptions of Popular Genomic Tools Used for Assembly de novo, Pairwise and Multiple Alignment, Gene Annotation and Variant Calling in EFPs

Purpose	Tool	Current Version	Input/Notes	Citations
Assembly	ALLPATHS-LG	v4.7	Two Illumina fragment (paired-end) libraries	Gnerre et al. (2011)
	Canu	v1.4	overlapping for noisy, long reads such as MinION or PacBio	Koren et al. (2017)
	DISCOVAR de novo	N/A	Single Illumina fragment (paired-end) library	Love, Weisenfeld, Jaffe, Besansky, and Neafsey (2016)
	Platanus	v1.2.4	De novo assembly of highly heterozygous genomes	Kajitani et al. (2014)
	SGA	N/A	Memory efficient tool for large genomes	Simpson and Durbin (2012)
	SOAPdenovo	v2	One or more single and/or paired-end libraries	Li et al. (2010)
	SPAdes	v3.5	Single-cell and standard (multicell) libraries, haploid or diploid	Bankevich et al. (2012)
	Trinity	v2.3.2	RNAseq data (optionally genome guided)	Haas et al. (2013)
Alignment	BLAST	Blast+	Fast searches against large databases	Altschul, Gish, Miller, Myers, and Lipman (1990)
	BLAT	N/A	Fast searches and connects homologous hits	Kent (2002)
	Bowtie	v2.3.0	Supports gapped, local, and paired-end alignment modes	Langmead and Salzberg (2012)

Table 1 Names, Versions and Descriptions of Popular Genomic Tools Used for Assembly de novo, Pairwise and Multiple Alignment, Gene Annotation and Variant Calling in EFPs—cont'd

Purpose	Tool	Current Version	Input/Notes	Citations
	BWA-mem	v0.7.15	Low-divergent sequences against a large reference genome	Li and Durbin (2010)
	HISAT2	v2.0.5	DNA or RNA to a population of genomes	Pertea, Kim, Pertea, Leek, and Salzberg (2016)
	MAFFT	v7	High speed multiple sequence alignment program	Katoh and Standley (2013)
	MAVID	v2.0.4	Multiple alignment program for large genomic sequences	Dewey (2007)
	MUMmer	v3.22	Aligns entire genomes	Kurtz et al. (2004)
	MUSCLE	v3.8.31	Multiple sequence alignment	Edgar (2004a)
	STAR	v2.5	Spliced transcripts (RNAseq) to a reference	Dobin et al. (2013)
	TBA MULTIZ	12109	Aligns highly rearranged or incompletely sequenced genomes	Blanchette et al. (2004)
Annotation	Augustus	v2.5.5	Ab initio gene-prediction program for eukaryotes	Stanke et al. (2006)
	EVM	v1.1.1	Combines diverse evidence types into single gene structures	Haas et al. (2008)
	FGENESH	v2.1	HMM-based ab initio gene-prediction program	Salamov and Solovyev (2000)
	GeneID	v1.4.4	Predicts genes in anonymous genomic sequences	Blanco, Parra, and Guigó (2007)

Continued

Table 1 Names, Versions and Descriptions of Popular Genomic Tools Used for Assembly de novo, Pairwise and Multiple Alignment, Gene Annotation and Variant Calling in EFPs—cont'd

Purpose	Tool	Current Version	Input/Notes	Citations
	GenemarkHmmEs	v2.3	Unsupervised training for identifying eukaryotic protein-coding genes	Lukashin and Borodovsky (1998)
	GlimmerHmm	v3.02b	gene finder based on interpolated Markov models (IMMs)	Majoros, Pertea, and Salzberg (2004)
	PASA	v2	spliced alignments of ESTs and RNAseq to model gene structures	Haas et al. (2008)
	RNAmmer	v1.2	Consistent and rapid annotation of ribosomal RNA genes	Lagesen et al. (2007)
	SNAP	2013	Ab initio gene finding program	Korf (2004)
	tRNAscan	v1.3.1	Transfer RNA detection	Lowe and Eddy (1997)
	Wise2 (GeneWise)	v2.4	Predicts gene structure using similar protein sequences.	Birney, Clamp, and Durbin (2004)
Variant calling	Biscap	v0.11	Variants called from Pileup format using binomial probabilities	Farrer, Henk, MacLean, Studholme, and Fisher (2013)
	FreeBayes	v0.9.10	Bayesian haplotype-based polymorphism discovery and genotyping	Garrison and Marth (2012)
	GATK	v3.7	Collection of tools with a focus on variant discovery	McKenna et al. (2010)
	Pilon	v1.5	Corrects draft assemblies and calls sequence variants	Walker et al. (2014)

PacBio or MinION (Koren et al., 2017)), sequencing libraries (e.g., Spades for DNA, Bankevich et al., 2012; Trinity for RNA, Haas et al., 2013); high levels of heterozygosity (e.g., Platinus (Kajitani et al., 2014), estimated ploidies or repeat content, scalability, or computational speeds given differing computational resources. Reviews of methodologies and tools include Assemblethon 2 (Bradnam et al., 2013) and Genome Assembly Gold-standard Evaluations (Salzberg et al., 2012). However, most tools make use of one of two underlying algorithms: Overlap of reads to construct contiguous stretches of sequences, or *k*-mers (subread sequences of length *k*) organized into deBruijn graphs. In both cases, the longest path through the graph is considered correct, and bubbles (loops caused by repeats) cut or removed. Usually the initial reads are organized into contigs, which are separately orientated and connected to one another into scaffolds (connected by Ns, representing ambiguous bases of the estimated length between the two contigs). The finished assembly should be assessed using a variety of metrics, as the result may be suboptimal or inaccurate—thereby negatively impacting any downstream analysis.

A genome assembly will usually aim to represent the nucleotide sequence of a single isolate, separated into individual chromosomes. However, it is always (unless from single-cell sequencing), a consensus from a colony or even population of cells, meaning that the assembly represents a range of individual genotypes. This is especially relevant when fungal cells are heterokaryotic, containing multiple nuclei such as is the case with filamentous ascomycetes and arbuscular mycorrhizal fungi (Pawlowska & Taylor, 2004). Sometimes, such a consensus may even be intentional such as with pangenomes of *Saccharomyces cerevisiae* (Song et al., 2015). In nonhaploid EFPs, including many *Candida* isolates that represent over 50% of human mycoses (Nucci & Marr, 2005), the assembly will consist of a consensus (and arbitrary connection) of haplotypes. Therefore, even a nonerroneous assembly could easily be wrongly interpreted for various downstream analysis including genetic variation, linkage disequilibrium, and recombination.

A simple metric to assess the quality of genome assemblies is the total assembly size—which itself can be informative in identifying contaminants and miss-assemblies (Studholme, 2016). For example, a genome that is far larger or smaller than expected for the genus can indicate multiple sequenced species (i.e., contamination with other organisms), high error rates in the sequencing, highly repetitive sequence, low sequencing depth, or unsuitable parameters. To control quality, an important step is to BLAST (Altschul et al., 1990) the scaffolds against the online or local nonredundant database

in order to identify whether contamination by another species is contributing to an erroneously assembled genome. Such a search, in addition to non-uniform GC content and other assembly metrics, can also be computed by such tools as the Genome Assembly Evaluation Metrics and Reporting (GAEMR) package (http://software.broadinstitute.org/software/gaemr/) or REAPR (Hunt et al., 2013). Given sufficient evidence of contamination, it is often beneficial to reassemble the reads after excluding any reads aligning (and therefore originating from) the source of the contamination—which can lead to improved contiguity and accuracy by excluding erroneous chimeric genomic regions. An assembly should then be assessed for its contiguity; a common measure of assembly contiguity is its N50 (meaning 50% of the assembly is in contigs of this length or larger). Similarly, N90 and N25 are sometimes also reported for assemblies. The N50 can be normalized for comparisons between multiple assemblies by using the estimated genome size instead of total assembly size (denoted NG instead of N (Bradnam et al., 2013)). These metrics can however both be misleading given, for example, a single very long scaffold above the N(G)50 length, which will then ignore the remaining assembly which may occur as highly fragmented scaffolds. Another proposed metric is the proportion of the assembly that has a length of at least the average eukaryotic gene (2.5 kb) (Bradnam et al., 2013) and will therefore be approximately the minimum length necessary for annotation—which may be the primary use for the assembly.

In the past year, genome assemblies from EFPs have included the chytrid fungus *B. salamandrivorans*, which is devastating fire salamanders in Europe (Farrer et al., 2017; Martel et al., 2014). Here, the genome was assembled using Illumina paired-end reads and SPAdes (Bankevich et al., 2012) into a draft assembly, revealing a substantial increase in genome length and expansion of metalloprotease M36 involved in skin destruction compared with its closest relative *B. dendrobatidis* (Farrer et al., 2017) (Fig. 2D). The ascomyceteous fungus *Sarocladium oryzae* is emerging as major threat for rice production (Bigirimana, Hua, Nyamangyoku, & Höfte, 2015) and was assembled by Illumina paired-end reads and SPAdes assembler (Bankevich et al., 2012), revealing a range of expanded gene families including the pathway for steroidal antibiotic helvolic acid thought to be a pathogenicity determinant (Hittalmani, Mahesh, Mahadevaiah, & Prasannakumar, 2016). The ascomyceteous fungus *C. auris* is emerging as a multidrug-resistant human pathogen in intensive care settings across the world and has been Illumina sequenced and assembled using Velvet (Zerbino & Birney, 2008) and scaffolded using SSPACE (Boetzer & Pirovano, 2014) and more recently

using Oxford Nanopore Technology and Illumina (Rhodes et al., 2017). These assemblies are now being used to determine the genetic mechanisms that underpin the multidrug-resistant nature of this species to fluconazole, voriconazole, amphotericin B, and caspofungin (Sharma, Kumar, Meis, Pandey, & Chowdhary, 2015).

Short- or long-read alignments against a presequenced reference are more commonly used for NGS datasets than assemblies, providing that a suitable reference sequence is already available. There are several reasons for opting for alignment over multiple assemblies. First, it is almost always quicker in terms of computation time. Second, alignments negate the necessity to identify orthologous regions of the genome needed to make comparisons. Indeed, orthology finding from assemblies is hindered by the necessity for relaxed sequence similarity thresholds in global sequence alignment algorithms. Furthermore, the steps from alignment to variant calling, gene cataloging, and selection analysis are well established. However, it needs to be born in mind that suboptimal tools, parameters, or quality checks can lead to misleading results.

Sequence alignment tools arrange two (i.e., pairwise) or more (i.e., multiple alignment) nucleic acids or protein sequences, with the intent of identifying regions of similarity that may indicate functional, or evolutionary relationships. Pairwise alignment algorithms are often based on either the Smith–Waterman algorithm (local/subsequences) or the Needleman–Wunsch algorithm (global/complete sequences). Both create a substitution matrix based on a scoring scheme (e.g., +3 for match, −2 for mismatch, −2 for indel) for each nucleotide or amino acid, and then trace back from the highest score. Tools such as EMBOSS' Needle and Water implement these algorithms directly, while others use them for extending seeds (prior screen for short matches), e.g., BWA-mem (Li & Durbin, 2010).

Many heuristic alignment algorithms and tools have been developed to improve on the speed of the Smith–Waterman and Needleman–Wunsch algorithms, such as the Basic Local Alignment Search Tool (BLAST) (Altschul et al., 1990) and the BLAST-like alignment tools (BLAT) (Kent, 2002), which removes low-complexity regions and makes *k*-letter "words" from the query sequence for searching the database of sequences using a scoring matrix. Mash is another recent and promising ultra-fast genomic distance estimation tool (Ondov et al., 2016). BLAST and BLAT are most commonly used for querying against a very large database, while NGS aligners such as BWA-mem (Li & Durbin, 2010) or Bowtie2 (Langmead & Salzberg, 2012, p. 2) are optimized for memory and time-

efficient alignment of a huge number of reads to a genome (Table 1). Global alignment tools for whole genomes include MAVID (Dewey, 2007), MUMMER (Kurtz et al., 2004), and TBA MULTIZ (Blanchette et al., 2004), which generally identify seeds that are joined (or removed) to form anchor regions for the final alignment. Others such as STAR (Dobin et al., 2013) are optimized for alignment of spliced transcripts (e.g., RNAseq data) to a genome.

To determine the overall accuracies of an input read dataset, alignment, and SNP calling method, one method is a comparison of false discovery rates (cFDR) (Farrer, Henk, MacLean, et al., 2013). In this procedure, a specified number of random single base positions in the reference sequence are randomly changes to one of the other three possible nucleotides. Sequence data is then aligned to this modified reference sequence, variant-calling performed, and a comparison made to those known changes to ascertain the overall accuracy. Multiple alignments or variant-calling tools or parameters can be used iteratively to identify the most suitable bioinformatics pipeline. Another metric for assessing the alignment quality is to assess the coverage across the genome by visualization. Some SNP calls such as GATK (McKenna et al., 2010) include the ability to perform local indel realignment (realign reads around indels). Importantly, multiple BAMs relating to related isolates (such as parent and progeny, or those closely related) should be realigned together to avoid newly introduced discrepancies, i.e., regions where one isolate is realigned at a region and another is not. However, this process was recently removed from the best practices of HaplotypeCaller but retained for UnifiedGenotyper. Other tools such as MUMSA (Lassmann & Sonnhammer, 2005) compare multiple alignments using an average overlap score and a multiple overlap score to assess the accuracy of alignments. Both alignments and assemblies can be improved via preprocessing of reads such as by removing low quality reads or 3′ ends.

There are a diversity of potential variant calling tools available, which are mostly used postalignment. Many SNP callers consider homozygous and heterozygous (biallelic) sequences, but often not trialleles, for example, which can be present at low numbers in genomes that exhibit aneuploidy, or polyploidy such as that which marks *B. dendrobatidis* (Farrer, Henk, Garner, et al., 2013). GATK's HaplotypeCaller and UnifiedGenotyper (McKenna et al., 2010) currently require a ploidy to be given as a parameter to inform its genotype likelihood and variant quality calculations. This prior setting is therefore poorly suited for the investigation of aneuploidy in EFPs. Recently the cancer genome variant calling tool MuTect2 (Cibulskis et al.,

2013) has been incorporated into the GATK and allows for a varying allelic fraction for each variant—which could provide a work-around for polyploidy or aneuploidy. Separately, FreeBayes (Garrison & Marth, 2012) calls variants based on a Bayesian statistical framework and is also capable of modeling multiallelic loci in sets of individuals with nonuniform copy number. A computationally inexpensive variant caller is BISCAP (Farrer, Henk, MacLean, et al., 2013), which uses binomial probabilities for an expected error rate following alignment. The tool Pilon (Walker et al., 2014) calls variants of multiple sizes, including very large insertions and deletions, while also able to use them for correcting draft assemblies. Indeed, many other SNP callers have been developed, which may be tailored to data types or expected levels of variation. The number of possible tools and their rate of development make benchmarking an issue that needs to be frequently readdressed to ensure their accuracy and therefore all the downstream analysis based on it.

For any newly sequenced genome including those of EFPs, one of the key features to query will be its gene content, which require gene prediction and annotation protocols. Many tools have been developed (Haas, Zeng, Pearson, Cuomo, & Wortman, 2011) and may be more or less suited to different genomes, genome fragmentation, repeat content, or gene characteristics such as intron lengths. Genomic research institutes such as the Broad Institute of MIT and Harvard, or the U.S. Department of Energy Joint Genome Institute (DOE JGI) automate their pipelines. However, in practice, only partial automation is obtainable due to genomic or gene idiosyncrasies, which require different tools, different parameters, or at a minimum, a manual check of certain gene-prediction outputs. Further, various methods are required for the prediction of protein-coding genes compared to RNA genes, to determine whether the genes are located on the nuclear or mitochondrial genome, and it is usually necessary to separately identify repetitive regions in the genome. However, from a well-annotated genome sequence, numerous aspects of an EFPs biology can be determined such as biological pathways, life cycle, mechanisms of pathogenicity, and its relationship to other species through ortholog detection.

The first step usually taken for gene annotation is repeat identification and masking (replacing sequence with the ambiguity IUPAC code "N"). Repetitive sequences within genomes constitute a range of functional and nonfunctional (in the evolutionary conserved sense) regions of the genomes. For example, if a genome assembly is finished to the level of whole linear chromosomes, the ends will contain tandem (consecutive) repeat sequences

found within telomeres, ranging from 5-mer to 27-mer repeated several thousand times, which both protect the end of the chromosome from deterioration, chromosomal fusion, or recombination, and as a mechanism for senescence and triggering apoptosis. Other tandem repeats are found in centromeres, which are involved in kinetochore formation during mitosis. Centromeres in fungi are diverse sequences ranging from a few kilobases in *Candida albicans* (Sanyal, Baum, & Carbon, 2004) up to 75 kb in *Schizosaccharomyces pombe* (Fishel, Amstutz, Baum, Carbon, & Clarke, 1988), and due to their diverse sequences, are best detected by the binding of the specialized nucleosomes that contain the centromere-specific histone H3, CenH3. Interestingly, *Neurospora* centromeres are composed of degenerate (following repeat-induced point (RIP) mutations) transposons, mostly retrotransposons, and simple sequence repeats (Smith, Galazka, Phatale, Connolly, & Freitag, 2012). Tandem repeats such as those found in telomeres and centromeres are grouped into microsatellites (also known as short tandem repeats or simple sequence repeats), which comprise 2–5mers, and minisatellite repeats comprising 10–50mers. Micro- and minisatellites are useful as genomic markers and are also studied for their role in disease. Tools such as the tandem repeat finder (Benson, 1999) and microsatellite identification tool (Thiel, Michalek, Varshney, & Graner, 2003) can be used to identify tandem repeats, while the Tandem Repeat Database is a public repository of those already identified (Gelfand, Rodriguez, & Benson, 2007).

Repetitive regions of a genome also include mobile DNA elements such as retrotransposons, DNA transposons, and miniature inverted-repeat transposable elements. Transposable element content varies in the fungal kingdom from between 3% (e.g., *Aspergillus nidulans*, *Aspergillus fumigatus*, and *Aspergillus oryzae*) and 10% (e.g., *Neurospora*, *Magnaporthe*) (Galagan, Henn, Ma, Cuomo, & Birren, 2005), but also as much as 76.4% for species such as the ascomycetous *Blumeria graminis f.* sp. *hordei* (Barley powdery mildew) (Amselem, Lebrun, & Quesneville, 2015; Spanu et al., 2010). Retrotransposons usually have long terminal repeats (LTRs) encoding a reverse transcriptase necessary to convert their transcribed RNA back to DNA which is inserted back into a new position of the genome. Others belong to the long interspersed nuclear elements encode a reverse transcriptase and an RNA polymerase II promoter, but lack LTRs. Retrotransposons lacking reverse transcriptase genes and relying on other mobile elements for transposition are called short interspersed elements. DNA transposons, in comparison, do not involve an RNA intermediate and usually encodes

transposase enzymes in order to bind and incorporate itself into a new position in the genome. These genes can be erroneously incorporated into gene models during prediction and provide nonuniform numbers of predicted genes compared with closely related isolates or species.

Following repeat masking by such software as RepeatMasker (http://www.repeatmasker.org/), protein-coding genes are predicted by both ab initio (based on sequence provided only) and homology based (based on similarity to known sequences). Ab initio methods rely on probabilistic models, such as generalized hidden Markov models (GHMMs) or neural networks (NN) to combine information from sequences that indicate the presence of a nearby gene (promoters and other regulatory signals) or protein-coding sequences. Most have individual models to assess, for example, splice donor sites (5′ end of the intron), splice acceptor sites (3′ end of the intron), intron and exon length distributions, open reading frame length, and transcriptional start and stop sites. Programs such as Augustus (Stanke et al., 2006), FGENESH (Salamov & Solovyev, 2000), GeneID (Blanco et al., 2007), GeneMark (Besemer & Borodovsky, 1999), GlimmerHMM (Majoros et al., 2004), and SNAP (Korf, 2004) are used for ab initio gene calling by first generating a training set (taking the highest scoring predictions from their GHMM) and then running across the genome sequence. Others such as GeneMark.hmm-ES (Lukashin & Borodovsky, 1998) is self-training. While any one of these methods could provide a modest initial assessment of gene content, it is worth running a number of tools in order to get a greater range (and therefore sensitivity) of predictions.

Homology-based (empirical) gene finding methods search for sequences that have sequence similarity to previously found genes in other organisms. These methods provide evidence for gene locations, which are both standalone, and compliment ab initio gene finding methods. This requires translating regions of the genomes (ideally in all six possible reading frames), using, for example, a translated BLAST (tblastn) against a database such as the nonredundant sequences from GenBank (Benson et al., 2012) and/or Uniprot (Wu et al., 2006). While this step is very computationally expensive, it provides likely protein hits which can then be assessed more rigorously. One such package for providing spliced gene models from these hits is Wise2 (Birney et al., 2004).

Transcript sequences from the same organism (such as RNAseq, expressed sequence tags/subsequence of cDNA) provide very strong evidence for gene structures in the genome sequence. A common first step

is to assemble de novo the RNAseq reads into longer transcripts, via programs such as Trinity (Haas et al., 2013). The accuracy of Trinity can be improved with strand-specific RNAseq libraries, the genome-guided parameter (both where available), and *k*-mer depth should be increased to at least 2 for improved specificity. Next, tools such as the Program to Assemble Spliced Alignment (PASA) (Haas et al., 2008) map these assembled transcripts (or unprocessed RNAseq reads) to the genome using GMAP (Wu & Watanabe, 2005) or BLAT (Kent, 2002), filtering alignments, grouping alternatively spliced isoforms and output candidate gene structures based on the longest open reading frame (FASTA file and GFF3). In addition, PASA can update prepredicted gene structures.

Finally, tools such as EvidenceModeller (EVM) (Haas et al., 2008) or Maker (Cantarel et al., 2008) assess the evidence provided for gene calls from a range of gene predictions (ab initio, homology-based, transcript data), and output a single set of consensus gene structures. Maker also contains a complete pipeline for identifying repeats, aligning ESTs and proteins to the genome, and ab initio gene prediction, before assessing their evidence. The final gene set following evidence assessment should be given a final check for various issues that may remain (coding length nonmodular 3, genes >50 amino acids, genes with in-frame stops, contain Ns indicative of spanning contig gaps, covering predicted repeats, etc.) prior to finalizing these predicted protein-coding genes with annotation, and gene identifiers such as unique locus tags.

The correct genetic code should be used throughout the entire process of predicting protein-encoding genes. The standard code (which should be the default for most tools) is suitable for most fungal nuclear genomes, although some species such as various *Candida* species in the CTG Clade have CUG codons that encode the amino acid serine instead of leucine (Santos, Keith, & Tuite, 1993), and therefore require the alternative Yeast nuclear code (Osawa, Jukes, Watanabe, & Muto, 1992). Mitochondrial genomes all use separate nonstandard codes, a difference that needs to be accounted for when translating genes in silico as part of the operation of these gene-prediction methods.

Genes that encode tRNA and rRNA are normally found in large numbers throughout a well-assembled genome. rDNA encoding rRNA are usually found entirely occupying large sections of one or more chromosomes, comprising both structural rRNA for small (18S) and large (5S, 5.8S, and 28S) components of ribosomes separated by internal transcribed spacer (ITS) units. These regions of the genome tend to be among the most poorly

resolved due to their repetitive nature—culminating in noncomplete regions that underestimates their number, but result in a region of unusually high depth of coverage following read alignment. Indeed, the rRNA completeness of a genome can be a proxy of genome assembly quality. Separately, ITS1 and ITS2 spacer regions tend to be useful for diagnostic PCR, fungal abundance (qPCR), and even rudimentary phylogenetics due to their ubiquity and genetic diversity in most fungal genomes (however, excluding the microsporidia) (Schoch et al., 2012).

Like protein sequences, RNA families have some level of conserved sequence, but a more highly conserved secondary structure, which is more integral to its function than that imposed by its primary sequence. Unlike protein sequences, ncRNA lack all features apparent from codons and gene structures (e.g., start, termination, codon bias, acceptor and donor splice sites, etc.) that are used for gene prediction, making the structure not only more relevant for its function (or predicted pseudogenization) but also for its prediction. RNA secondary structure arises from base-pairing interactions resulting in stems and loops, e.g., the cloverleaf structure of tRNA comprising several stem-loops, or the pseudoknot also comprising several stem-loops in the RNA component of telomerases (which incidentally is essential for telomerase activity (Chen & Greider, 2005)).

RNA secondary structure prediction based on the lowest free energy structure is a nondeterministic polynomial-time hard (NP-hard) problem (Lyngsø & Pedersen, 2000), and therefore tools based on heuristic algorithms are required for de novo RNA secondary structure (such as HotKnots (Ren, Rastegari, Condon, & Hoos, 2005) and ProbKnot (Bellaousov & Mathews, 2010)). However, prediction in a new genome is usually based on previously identified and characterized ncRNA families, which are often stored in covariance models (CMs) (analogous to hidden Markov models (HMMs)) describing both secondary structure and primary sequence consensus (Eddy & Durbin, 1994). For example, the INFERNAL (Inference of RNA Alignment) software (Nawrocki & Eddy, 2013, p. 1) searches a custom or collection of ncRNA CMs such as the RNA family database (RFam) (Griffiths-Jones, Bateman, Marshall, Khanna, & Eddy, 2003) comprising tRNAs, small nuclear RNAs (snRNA), and small nucleolar RNAs (snoRNAs). snRNAs are involved in splicing and RNA processing, while snoRNAs either methylate (C/D box snoRNA) or pseudouridylate (H/ACA box snoRNAs) other RNAs (rRNA, tRNA and snRNA). Separately, the CM-based tRNAscan-SE (Lowe & Eddy, 1997) can identify tRNA genes with extremely high sensitivity and specificity, and the HMM-based

RNAmmer predicts rRNA genes in the nuclear genome (Lagesen et al., 2007).

It is important to assess the quality of the final gene calls for multiple potential erroneous calls, such as genes that have a length that is not modulus 3 (i.e., sequences not entirely comprised of codons), genes with STOP codons within the sequence, or those ending without a STOP codon are likely errors. Other issues can include very distant exons (e.g., >15 kb) from the remainder of the gene will be likely inaccurate. Gene calls that are supported by only one of multiple gene-prediction methods may also be more dubious than those supported by multiple methods and tools. A simple postannotation metric is the total number of genes predicted. Too many or too few predicted genes for a given genus or species can be indicative of a failed step in the annotation pipeline, or suggest a problem with the genome assembly, e.g., species contamination. In addition to gene count, the completeness of gene sets can be assessed by the coverage of conserved gene sets such as CEGMA (Parra, Bradnam, & Korf, 2007) and BUSCO (Simão, Waterhouse, Ioannidis, Kriventseva, & Zdobnov, 2015), which will give a good indicator of the quality of both the annotation and the assembly protocols. The measure of gene completeness should complement other metrics of genome assembly, and be performed before functional predictions and other downstream analyses are performed.

2.2 Functional Predictions and Gene Family Expansion

Functional genomics describes the relationship between an organism's genome and its phenotype, and is widely used to determine novel pathogenicity-related traits in EFPs. There are numerous experimental ways to study these traits including gene knockouts, gene silencing, transposon or chemical mutagenesis, and QTL mapping. Computational methods for identifying pathogenicity-related gene functions includes Genome Wide Association Studies (GWAS), which commonly compares two large groups of individuals that differ by a pathogenicity-related trait, and to then search for a significant association (low *P*-value from a chi-squared test of the odds ratio). GWAS have been used to successfully identify a wide range of candidate genes and alleles implicated in disease or pathogenicity-related phenotypes, including in a broad range of fungal applications (Plissonneau et al., 2017). For example, a putative Avirulence gene (virulence factors that are detected by the host, and thereby prevent or reduce disease) was recently detected using a GWAS of *Zymoseptoria tritici*, the ascomycetous fungi

responsible for septoria lead blotch in wheat (Hartmann, Sánchez-Vallet, McDonald, & Croll, 2017).

GWAS have several limitations including the necessity for very large sample sizes, which is commonly not available for EFPs, and the need to account for the large numbers of multiple comparisons that inevitably lead to false associations. Furthermore, many populations of fungal pathogens contain a large clonal component to their life cycle—with the consequence that variants are physically linked on the chromosome (high linkage disequilibrium). Clonality therefore impinges on the ability to identify individual variants that are associated with a trait. Finally, specific functions of a protein-coding gene (e.g., those encoding chloride channels) are relatively easy to predict, compared with predicting phenotypes and pathologies linked to mutations or protein misfolding (e.g., those causing cystic fibrosis in humans). This section will focus on ab initio and in silico methods of functional genomics that rely only on a single or very few isolates—such as might be available from the outbreak of an EFP.

Following (or as part of) gene prediction, functional annotation can be assigned to each protein-coding gene, and thereby provide a prediction of its function in the organism. Perhaps the most common method to do this is to assign Protein Family (PFAM) domains (Finn et al., 2014), which as of the current v31.0 (10/2016) has defined 16,712 protein families, and to a lesser extent, assigning TIGRFAM domains (Haft, Selengut, & White, 2003), which as of the current v15.0 (10/2014) has defined 4488 protein families—both of which are generated using HMMs. Each protein family is composed of one or more functional regions termed domains—which are found in multiple proteins and protein families. Both PFAM and TIGRFAM databases provide profile HMMs for each protein families, which are built from multiple sequence alignments and are searched either online via web servers (Finn, Clements, & Eddy, 2011) or local copies using the HMMER3 software (Eddy, 2011). Separately, the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Kanehisa & Goto, 2000) can be searched using the predicted gene sequences using a BLASTx (and a suitably stringent *e*-value, i.e., $e < 1 \times 10^{-10}$) to identify various functional information on gene functions, their role in biological pathways and cellular processes. Any successful matches with sufficiently stringent e values can provide compelling evidence toward the function of that gene. However, not all families or domains are contemporaneously informative regarding the function. For example, many domains of unknown function (DUF) are present in the database, which have been identified as a conserved

domain across multiple species, but no known function has yet been identified.

Gene Ontologies (GO) provide a parallel and complimentary gene prediction along with PFAMs/TIGRFAM/etc. assignment. GO terms represent a controlled vocabulary and defined set of relationships between them, as part of the Open Biomedical Ontologies project by the National Center for Biomedical Ontology (NCBO). GO terms cover three domains: cellular components, molecular functions, and biological processes, for which a given gene is often assigned multiple terms, ranging from the very specific (low hierarchical/child terms) such as molecular function GO:0004375 (glycine dehydrogenase (decarboxylating) activity), to the very generic (high hierarchical/parent terms) such as molecular function GO:0003824 (catalytic activity). There are a wide range of tools for working with sets of GO terms, including Blast2GO (Conesa et al., 2005), which uses a stringent BLAST ($e < 1 \times 10^{-10}$) to identify genes with assigned GO terms, which can then be reassigned. Once assigned, GO terms can be very useful for predicted function, grouping genes into functionally relevant categories and ultimately performing enrichment statistical tests between groups of genes.

Some functions such as the secretion signal/peptide found at the N-terminus of newly synthesized proteins destined for the secretion pathway are best predicted by its biochemical properties rather than its poorly conserved primary sequence alone i.e., via sequence similarity or homology. SignalP is a popular tool that predicts the presence of type I signal peptidase cleavage sites from preprotein sequences in bacteria, archaea, fungi, plants, and animals (Petersen, Brunak, von Heijne, & Nielsen, 2011; Tuteja, 2005). Conversely, type II and type IV signal peptidases are restricted to prokaryotes and require prediction by other methods. In SignalP 3.0 and 4.0, type I signal peptidase cleavage sites are detected by neural networks, which are trained on real and negative data from SwissProt (Bairoch & Apweiler, 2000). The two neural networks used in SignalP recognize cleavage sites and determine if a given amino acid belongs to the signal peptide, respectively. SignalP is also informed by filtering propeptide cleavage sites, window length across the protein, and a discrimination score (D-score). The authors of SignalP also assessed the isoelectric point (pH(I); the pH at which the protein carries no net electrical charge) difference between the signal peptide and mature protein, which they found to be distinct in prokaryotes, but not eukaryotes—possibly owing to the much shorter length in eukaryotes. SignalP 4.0 updates the method by distinguishing between signal

peptides and N-terminal transmembrane helices, which can be incorrectly identified. Limitations with SignalP include imperfect sensitivity and specificity (albeit the best method from their own comparisons to other tools and methods).

Some proteins predicted to have a signal peptide may nevertheless be retained intracellularly, i.e., in the endoplasmic reticulum or Golgi. For example, if the protein contains a C-terminal ER retention signal (KDEL or KKXX sequence), or via protein–protein interactions in the Golgi, the protein will not become extracellular (Banfield, 2011; Stornaiuolo et al., 2003). Furthermore, there are additional nonclassical secretion mechanisms in eukaryotes, such as via specific membrane transporters (Nickel & Seedorf, 2008). Tools such as SecretomeP (Bendtsen, Jensen, Blom, Von Heijne, & Brunak, 2004) predict secretory proteins that lack an N-terminal signal peptide. However, this method is tailored primarily to mammalian proteins, and when recently applied to four chytrid genomes, most proteins were identified as being nonclassically secreted (6523/10,128 *B. salamandrivorans* genes; 4478/8644 *B. dendrobatidis* genes; 4581/8952 *Spizellomyces punctatus* genes; 2991/6254 *Homolaphylictis polyrhiza* genes). This finding suggests that the mammalian-trained pipeline is, at least in this case, overpredicting nonclassical secretion motifs and needs to be retrained specifically to the fungal secretome (Farrer et al., 2017).

Secreted proteins are often of paramount importance to pathogens in acquiring nutrients, and interactions with the environment and host. An illuminating example of this are virulence effectors, which are secreted either into the environment or directly into the host, where they selectively bind to a host protein to regulate or modify its intended function (Hogenhout, Van der Hoorn, Terauchi, & Kamoun, 2009). Effectors are produced by a wide range of organisms including many fungal and bacterial pathogens, but also some animals (parasitic nematodes), as well as protists (*Plasmodium* sp. and Oomycetes). For example, effectors may encode proteins that target host defense mechanisms to enable the microbe to gain access to the host cell or avoid detection (either innate or acquired immunity, for example). One example is the gene AVR3a (belonging to a group that have an RXLR or RXLQ motif, and collectively known as RXLRs), which is found in the *Phytophthora* genus. AVR3a (specifically AVR3aKI that contains amino acids C19, K80, and I103) causes suppression of a hypersensitive response (apoptosis) in potatoes that lack the necessary resistance gene R3a (Bos et al., 2006), thereby facilitating its initial biotrophic stage of growth. Changes in the genetic backgrounds upon which virulence effectors are

found can directly drive EFPs, as has been clearly shown by the change of virulence due to the horizontal gene transfer (HGT) of *ToxA* from *Phaeosphaeria nodorum* into *Pyrenophora tritici-repentis* in the 1940s, causing aggressive tan spot disease in wheat (Friesen et al., 2006).

A large number or fraction of secreted genes in the genome can be an indicator that a fungus is pathogenic when compared with their related saprobic relatives; this is clearly the case for the species of *Batrachochytrium* (Rosenblum, Stajich, Maddox, & Eisen, 2008). For example, amplifications of secreted protein repertoires are clearly seen in the genomes of the EFPs *B. salamandrivorans* ($n=1527$) and *B. dendrobatidis* ($n=833$) compared to the related free-living saprobic *S. punctatus* ($n=587$) and *H. polyrhiza* ($n=293$) (Farrer et al., 2017). Here, it was shown experimentally that of the chytrid genes that were significantly upregulated in vivo ($n=550$), a large proportion was unique to *B. salamandrivorans* ($n=327$; 60%), unique to *B. dendrobatidis* ($n=43$; 8%) or unique to the genus *Batrachochytrium* ($n=44$; 8%). Furthermore, around half of the *B. salamandrivorans* and *B. dendrobatidis* upregulated genes were secreted (55% and 47%, respectively). The fact that these secreted proteins are both largely not present in the saprobic chytrids based on ortholog identification, and that they show increased transcription during host colonization, suggests that the transcriptional response is focused on a unique host-interaction strategy in each species.

Separately, several genes from a class called Crinkler and Necrosis (CRN)-like genes can either trigger cell death (such as PsCRN63) or inhibit cell death (such as PsCRN115) when expressed inside plant cells (Liu et al., 2011) by pathogens belonging to the *Phytophthora* and *Lagenidium* genera of Oomycetes (Quiroz Velasquez et al., 2014; Schornack et al., 2010). Crinklers are often located in gene-sparse, repeat rich, regions of the genome in well-studied eukaryotic plant pathogens (Haas et al., 2009). A recent study examined gene-sparse regions of the amphibian-infecting chytrid pathogen *B. dendrobatidis* (Farrer et al., 2017). Notably, it was found regions of low gene density include homologs of CRNs. Chytrid CRNs were identified via BLASTp to those in *Phytophthora infestans* T30-4 (Haas et al., 2009), and CD-hit (Li & Godzik, 2006) under a number of sequence similarity identities, as well as trimming the more divergent C-terminal to 35aa, 40aa, 45aa, and 50aa, followed by, or proceeded by, a MUSCLE alignment (Edgar, 2004b) with or without removing excess gaps using trimAl gappyout (Capella-Gutiérrez, Silla-Martínez, & Gabaldón, 2009). Motif searching was performed using GLAM2 (Frith,

Saunders, Kobe, & Bailey, 2008). Searching all of the sequences together after trimming to 50aa did not yield a convincing single domain. Instead, it was found that manually separating genes with two overrepresented sequences obtained the highest confidence alignments spanning the most number of CRNs (Farrer et al., 2017). This process illustrates the trial-and-error approaches that can be required for investigating and classifying novel protein families in EFPs, particularly those with low sequence similarity or small proteins.

CRN-like genes in *B. dendrobatidis* are characterized by having long intergenic regions that are consistent with a gene-poor repeat-rich environment (averaging 1.4 kb)—a trait shared with *P. infestans* T30-4 (Haas et al., 2009). Farrer et al. (2017) showed that the CRN-like family is more widely distributed among the Chytridiomycota than previously realized. Specifically, this study identified 162 CRN-like genes in *B. dendrobatidis*, 10 in *B. salamandrivorans*, 11 in *H. polyrhiza*, and 6 in *S. punctatus*, many of which ($n=55$) belong to a single subfamily (known as DXX). Besides some sequence similarity, there are multiple differences between CRNs found in the chytrid genomes compared with Oomycete genomes. For example, only two chytrid CRNs had predicted secretion signals (via SignalP4 (Petersen et al., 2011))—one in each of the free-living saprobe chytrids *H. polyrhiza* and *S. punctatus*, which contrasts with CRNs in *Phytophthora* species that are mostly intracellular effectors that target the host nucleus during infection (Stam et al., 2013). Another discrepancy is that CRN-like genes appeared to be downregulated during advanced infection of a susceptible salamander species (*Tylototriton wenxianensis*) compared with in vitro conditions, while many Oomycete CRNs are upregulated *in planta* (Chen, Xing, Li, Tong, & Xu, 2013). In both *B. dendrobatidis* and *B. salamandrivorans*, some CRN-like genes were more highly expressed in the zoospore life stage compared to the sporangia life stage (Farrer et al., 2017). However, incubation of *B. dendrobatidis* zoospores with *T. wenxianensis* tissue for 2 h showed an increased expression of CRN genes, whereas *B. salamandrivorans* zoospores were associated with decreased expression, indicating that CRN genes are possibly of greater interest in the early infection stage of *B. dendrobatidis*, but not *B. salamandrivorans*; the notable expansion of CRN-like genes in *B. dendrobatidis* may suggest that they are of importance; however, their function currently remains unknown.

To ascertain if the secreted proteins in four species of chytrids included any large families (in addition to metalloproteases, for example), clustering

was used to predict secreted genes using the Markov Cluster Algorithm tool (Enright, Van Dongen, & Ouzounis, 2002) with recommended settings (Farrer et al., 2017). Associated PFAM domains were found in all or nearly all members of some tribes, including the second largest, which contained protease M36 domains, or the sixth largest, which contained the peptidase S41 domain. The largest tribe had 105 proteins, and belonged entirely to *B. salamandrivorans*, as did the fourth largest tribe. Many of the members of these secreted tribes were significantly differentially expressed between in vivo and in vitro conditions, including in "Tribe 1" (48% of genes). Furthermore, these tribes are located almost exclusively in nonsyntenic, unique regions of the *B. salamandrivorans* genome. However, this study was unable to identify any sequence similarity to previously described proteins or recognizable functional domains (BLAST, GO terms, PFAM, TIGRFAM, etc.) showing that these putative virulence factors require further work to understand their possible function. This study illustrates an important point: the constellation of virulence factors that lie within the biodiverse fungal kingdom has only been touched on, and future (yet undescribed) EFPs will likely harbor a wealth of undescribed virulence factors that are not represented in today's databases, and need (sometimes urgent) investigation.

A further class of proteins that are often of interest in EFPs are transmembrane (TM) proteins. An HMM-based method for predicting TMs is TMHMM, which purports to correctly predict 97%–98% of the transmembrane helices (Krogh, Larsson, von Heijne, & Sonnhammer, 2001), and with 99% specificity and sensitivity. However, the authors note that the accuracy drops when signal peptides are present. TM proteins function as gateways for substances to move between the environment and intracellular (such as voltage-gated and ligand-gated ion channels), and as such are integral to the functioning of the cell, and at the host–pathogen interface. Toll-like receptors and receptor kinases are examples of conserved TM pattern recognition receptors of the innate immune system of animals and plants, respectively. In *Aspergillus*, the seven-transmembrane domain protein PalH is a putative pH sensor required for virulence on mice (Grice, Bertuzzi, & Bignell, 2013). Another TM protein, TmpL, is necessary for regulation of intracellular ROS levels and tolerance to external ROS, and is required for infection of plants by the necrotroph *Alternaria brassicicola* and for infection of mammals by the human pathogen *A. fumigatus* (Kim et al., 2009).

The repertoire of proteases in EFPs is of interest, owing to their importance in physiology, development, survival, and growth. Furthermore, extracellular serine, aspartic, and metalloproteases are considered

virulence factors in many pathogenic species (Yike, 2011). Proteases can be identified, either by generic databases (e.g., nonredundant BLAST database) or by specialized protease databases (e.g., Merops (Rawlings, Barrett, & Finn, 2016), which as of release 11, contains 447,156 protein sequences of all the peptidases and peptidase inhibitors), both of which can be searched using BLAST.

The metalloproteases of the M36 fungalysin family are important pathogenicity determinants in a number of dermatophytes, which cause cutaneous infections and grow exclusively in the outermost layer of skin, nails, and hair of human and animals. Here, skin-infecting organisms, such as *Trichophyton* spp. that cause Tinea corporis/ringworm, Tinea pedis/athletes foot, secrete M36 proteases that are important for causing disease (Zhang et al., 2014). Again, the chytrid pathogens provide a further good example of M36 proteases and pathogenicity. Metalloproteases are dramatically expanded in *B. salamandrivorans* (Farrer et al., 2017), concordant with the aggressive necrotic pathology that this pathogen causes. Both *B. salamandrivorans* ($n=110$) and *B. dendrobatidis* ($n=35$) have expanded M36 families compared to lower counts in the free-living saprobic chytrids *S. punctatus* and *H. polyrhiza* ($n=2$ and $n=3$, respectively). Phylogenetic analysis revealed a subclass of closely related M36 metalloproteases that are shared across both pathogens that we termed the Batra Group 1 M36s (G1M36) (Fig. 2D).

Species-specific gene family expansion in chytrid pathogens is illustrated by the presence of a novel secreted clade of M36 genes ($n=57$) unique to *B. salamandrivorans*, which were termed the *Bsal* Group 2 M36s (G2M36) (Farrer et al., 2017). These G2M36s are entirely encoded by nonsyntenic regions of the *B. salamandrivorans* genome, supporting a recent species-specific expansion. Although most G1M36s and G2M36s are strongly upregulated in salamander skin, eight *Bsal* G1M36s (19%) appear more highly expressed in vitro, suggesting more complex regulatory circuits underlie this subclass of protease in *B. salamandrivorans*. Furthermore, G1M36s showed greater expression in *B. dendrobatidis* zoospores compared to sporangia, pointing to a crucial role of these proteases during early host colonization in *B. dendrobatidis*, for example, during insertion of their germ tube into the epidermal cells (Van Rooij et al., 2012). In contrast, the low protease activity in *B. salamandrivorans* zoospores, but high activity in the maturing sporangia, suggests a role during later stages of pathogenesis, for example, in breaching the sporangial wall of developing sporangia and subsequent spread to neighboring host cells (Martel et al., 2013).

Carbohydrate-binding modules (CBM), including CBM18, are expanded in *B. dendrobatidis* (Farrer et al., 2017) and been implicated in host–pathogen interactions (Abramyan & Stajich, 2012). To study individual protein families defined by PFAM domains, HMMs can be downloaded from the PFAM database (Finn et al., 2014) that are then used to search through a set of proteins using the HMMER3 (Eddy, 2011) application hmmsearch (with an *e* value cutoff of 0.01 or lower). CBM18s are markedly expanded in both *B. dendrobatidis* and *B. salamandrivorans* compared to the free-living chytrids (Farrer et al., 2017). CBM18 containing proteins are predicted to bind chitin and most copies of these proteins contain secretion signals that will target them to the cell surface or extracellular space. Species-specific differences are notable in the pronounced truncation of the lectin-like CBM18s of *B. salamandrivorans*, suggesting a fundamental difference in capacity to bind some chitin-like molecules. In comparison, CBM18 genes in *B. dendrobatidis* are threefold longer and harbor on average eight CBM18 domains compared with only 2.6 for *B. salamandrivorans*. However, their expression was not significantly altered upon exposure of sporangia to chitinases, suggesting their role in protecting the fungi from host chitinase activity by fencing off the fungal chitin unlikely. Rather, it was hypothesized that the CBM18s play a role in fungal adhesion to the host skin or in dampening the chitin-recognition host response.

CBM18 genes fall into three large groups among chytrids (Abramyan & Stajich, 2012). CBM18s containing carbohydrate esterase 4 (CE4) superfamily mainly includes chitin deacetylases clustered together, and called deacetylase-like. Another group of CBM18s contains a common central domain of tyrosinase, and called tyrosinase-like. A third group consisted of genes with no secondary domains is described as lectin-like. The six *B. dendrobatidis* LL CBM18 had a total of 48 CBM18 modules (averaging 8 per gene), while the six *B. salamandrivorans* lectin-like CBM18s had only 16 CBM18 modules (averaging 2.6 per gene) (Farrer et al., 2017). *B. salamandrivorans* lectin-like CBM18s are also considerably truncated compared with those of *B. dendrobatidis* (mean *B. salamandrivorans* protein length = 606, mean *B. dendrobatidis* protein length = 206). Most *B. dendrobatidis* CBM18s (17/21; 81%) are upregulated in vivo, although mostly nonsignificantly (2 DE, 1TL). In contrast, 7/15 (47%) *B. salamandrivorans* CBM18s are upregulated in vivo. However, five of these genes are significantly upregulated including two tyrosinase-like, two deacetylase-like and one lectin-like. The importance and function of these

genes remain to be fully demonstrated, however, appear to be involved in recognizing and binding host ligands as part of the infection process (Liu & Stajich, 2015).

Identifying gene families in EFPs is a necessary precursor to quantify increases or decreases relative to close relatives, and that may indicate why changes in pathogenicity-related traits have occured. For example, gene family expansions in pathogens compared with closely related nonpathogens can provide candidate virulence determinants. A common way for comparing genes and identifying gene family expansions is to first identify single copy orthologs between two or more species, especially when one of those genomes is well characterized. Recently, substantial investment has been made into developing online fungal genomic resources, such as FungiDB (Stajich et al., 2012) which can be used to assist in the categorization of orthologs. However, the protein-coding genes and gene family expansions make up only one aspect of the EFP's genome, where other aspects such as chromosome number may also be important.

2.3 Chromosomal CNV

Pathogenic fungi often manifest highly plastic genome architecture in the form of variable numbers of individual chromosomes, known as chromosomal copy number variation (CCNV) or aneuploidy. CCNV has been identified across the fungal kingdom in both EFP and nonpathogens alike. For example, among ascomycetous fungi, CCNV has been identified in the generalist plant pathogen *Botrytis cinerea* (Büttner et al., 1994), the human pathogen *Histoplasma capsulatum* (Carr & Shearer, 1998), bakers/brewer's yeast (and an occasional opportunist) *S. cerevisiae* (Sheltzer et al., 2011), and the human pathogen *C. albicans* (Abbey, Hickman, Gresham, & Berman, 2011). The occurrence of stress due to either the host response or exposure to antifungal drugs has been linked to a rapid rate of CCNV in *Candida* spp. (Forche, Magee, Selmecki, Berman, & May, 2009) and, within the Basidiomycota, the human pathogens *Cryptococcus neoformans* and *C. gattii* are both found exhibiting CCNV (Hu et al., 2011; Lengeler, Cox, & Heitman, 2001; Sionov, Lee, Chang, & Kwon-Chung, 2010). Even among the Chytridiomycota, *B. dendrobatidis* shows widespread heterogeneity in ploidy among genomes and among chromosomes within a single genome (Farrer, Henk, Garner, et al., 2013). The mechanism(s) generating chromosomal CCNV in fungi are not yet well understood, but are thought to occur because of nondisjunction following meiotic or mitotic

segregation (Reedy, Floyd, & Heitman, 2009), followed by selection operating to stabilize the chromosomal aneuploidies (Hu et al., 2011).

Dynamic numbers of chromosomes could offer routes to potentially advantageous phenotypic changes via several mechanisms such as overexpression of virulence factors (Hu et al., 2011) or drug efflux pumps (Kwon-Chung & Chang, 2012). CCNV contributes to the maintenance of diversity through homologous recombination (Forche et al., 2008), and increased rates of mutation and larger effective population sizes (Arnold, Bomblies, & Wakeley, 2012). CCNV may also provide the advantage of purging deleterious mutations through nondisjunction during chromosomal segregation (Schoustra, Debets, Slakhorst, & Hoekstra, 2007). Thus, CCNV likely represents an important, yet uncharacterized, source of de novo variation and adaptive potential in many fungi and other nonmodel eukaryote microbial pathogens. By mapping read depth and SNPs across *B. dendrobatidis* genomes, it was discovered that widespread genomic variation occurs in ploidy among genomes and among chromosomes within a single genome (Farrer, Henk, Garner, et al., 2013). Individuals from all three lineages harbored CCNV along with predominantly or even entirely diploid, triploid, and tetraploid genomes. Another study also identified widespread CCNV across diverse lineages of *B. dendrobatidis* recovered largely from infected amphibians in the Americas, including a single haploid chromosome in a global panzootic lineage (GPL) isolate (Rosenblum et al., 2013). This variation may itself, reflect only part of the full diversity in *B. dendrobatidis*, as +2/+3 shifts in ploidy, whole genomes in tetraploid, or chromosomes in pentaploid or greater, may occur and await discovery.

Chromosomal genotype in *B. dendrobatidis* was shown to be highly plastic as significant changes in CCNV occurred in as few as 40 generations in culture (Fig. 2B) (Farrer, Henk, Garner, et al., 2013). It is not known whether other chytrid species such as *B. salamandrivorans* also undergo CCNV, or if this is a unique feature of *B. dendrobatidis*, or even unique of just chytrid pathogens—and hence may be intrinsic to their parasitic mode of life. Currently, CCNV is known to occur in a variety of protist microbial pathogens, including fungi; however, it is currently not known whether this genomic feature is specific to a parasitic lifestyle or is a more general feature of eukaryote microbes; identifying the ubiquity of CCNV or otherwise across nonpathogenic species will therefore be of great interest. Further, the manner by which plasticity of CCNV in *B. dendrobatidis* affects patterns of global transcription and hence the phenotype of each isolate also remains to be studied. However, it is clear from research on yeast, *Candida*

and *Cryptococcus*, that CCNV significantly contributes to generating altered transcriptomic profiles, phenotypic diversity, and rates of adaptive evolution even in the face of quantifiable costs; understanding the relationship between CCNV and the phenotype of *B. dendrobatidis* will therefore likely be key to understanding its patterns of evolution at both micro- and macroscales.

CCNV has also been identified in three isolates belonging to *C. gattii* VGII and VGIII using read coverage (Farrer et al., 2015). Specifically, an additional (disomic) copy of scaffold 13 in VGII veterinary isolate B8828 was identified, and a disomy of scaffold II in VGIII clinical isolate CA1280 (syntenic to the first half of WM276 chromosome cgba). Variation in chromosome copy number has previously been shown to influence the virulence of *Cryptococcus* (Hu et al., 2011) and can further provide resistance to azole drugs by increasing the copy number of the azole drug target (ERG11) or transporter (AFR1) commonly amplified in drug-resistant *Cryptococcus* (Kwon-Chung & Chang, 2012) and across the time-scale of a single infection (Rhodes, Beale, et al., 2017). However, neither gene appears in these aneuploidies, suggesting they are not associated with known drug resistance mechanisms, although may have other effects on those isolates. Separately, a 60 kb intrachromosomal duplication was found in the middle of scaffold 1 of VGII clinical isolate LA55 (also syntenic to WM276 chromosome cgba), which interestingly did not appear in the closely related isolate CBS10090, suggesting it was of a recent origin. This 60 kb region covers 24 protein-coding genes that are not known to influence drug resistance in *Cryptococcus*.

In addition to CCNV, chromosomes of EFPs can fuse, split, and undergo inversions and translocations—which can have a dramatic effect on their phenotype. One method to study this in silico is to identify orthologs, and then to map their synteny. Recently, chromosome structure was compared in detail among the lineages of *C. gattii* (Farrer et al., 2015). Chromosome structure was found to be highly conserved between the four lineages, and very highly conserved within VGII. Almost all syntenic variation was identified among the three closely related lineages, VGI, VGIII, and VGIV (Fig. 2A). In total, 15 large (greater than 100 kb) chromosomal rearrangements were identified, such that on average, only 2.6% of each of the 16 genomes was rearranged with respect to the others. These 15 rearrangements included 10 translocations (7 interchromosomal and 3 intrachromosomal) and 5 scaffold fusions, most of which (13 of the 15) associated with clusters of predicted *Cryptococcus*-specific TcN transposons found

at centromeres (Janbon et al., 2014), suggesting these are primarily whole chromosome arm rearrangements. Four of the rearrangements were supported by multiple isolates, including one chromosomal fusion unique to VGII, two translocations unique to VGIII (700 and 140 kb, respectively), and one 450 kb translocation unique to VGIV. These changes may impact the ability for interlineage genetic exchange, as some crossover events will generate missing chromosomal regions or other aneuploidies and nonviable progeny.

2.4 Natural Selection

The widespread emergence of EFPs is testament to their ability to successfully adapt to infect diverse hosts and ecological niches, suggesting that their genomes are able to respond rapidly to natural selection. Characterizing variants in the genome by the type of selection acting upon them requires population genetics approaches. Some possible scenarios for variants in a population include those that are becoming fixed or rapidly evolving due to positive or diversifying selection, being purged due to purifying selection, being maintained in a population due to stabilizing selection, or accumulating mutations due to relaxed selection. In addition to selection pressures, knowledge of rates of recombination, ploidy, life cycle, population structure, and effective population size are all necessary to accurately assess the processes regulating and influencing allele frequencies in a population. Furthermore, a knowledge of how multiple loci or genes contribute to a given phenotype (epistasis) or are masked by others (pleiotropy), as well as random chance, e.g., genetic drift, gene flow, and HGT between populations all contribute to their genetic makeup.

One approach to study selection from genomic data is to look at patterns of synonymous mutations (those that maintain the amino acid sequence of the protein) and nonsynonymous mutations (those that change the amino acid sequence of the protein). An informative approach is to calculate the number of synonymous mutations per synonymous site (positions in the codon that can undergo synonymous mutations) (d_S) and the number of nonsynonymous mutations per nonsynonymous site (d_N) (Fig. 2F). However, the d_N/d_S ratio was originally developed for distantly diverged sequences, i.e., species, where the differences represent substitutions that have fixed along independent lineages, and is therefore unsuitable for identifying selection within a population (Kryazhimskiy & Plotkin, 2008).

The identification of variants in an alignment can be the result of multiple substitutions (increasingly with age since most recent common ancestor (MRCA)), and therefore substitution models (Markov model) are usually used when calculating d_N and d_S values (also denoted K_a and K_s, respectively). Different substation models may also differentially weight transitions (T_s) with respect to transversions (T_v) as T_s are more common at the third position in the codon, as well as GC and base/codon bias inherent to some genomes. Higher T_s/T_v ratios are also caused by spontaneous or cytidine deaminase-mediated deamination of methylated cytosines (Cooper, Mort, Stenson, Ball, & Chuzhanova, 2010), with differences even between animal mitochondrial genomes compared with their nuclear genomes (Belle, Piganeau, Gardner, & Eyre-Walker, 2005). Finally, suitable substitution models can be used by phylogenetic applications such as PAML (Yang, 2007), which estimates d_N, d_S, and $d_N/d_S=\omega$ by maximum likelihood.

When comparing two sequences (i.e., reference and consensus), any selection detected using d_N/d_S will not reveal where on the phylogenetic tree that selection has occurred, or even which of the two sequences or isolates are under selection. A more comprehensive test is to distinguish between selection on the reference sequence vs selection on the consensus sequence by using the branch-site model (BSM) of selection in the Codeml program of PAML (Yang, 2007) to calculate ω across genes and branches/lineages. Multiple corrections are then used to improve specificity for positive selection (such as Benjamini–Hochberg (Benjamini & Hochberg, 1995), Bonferroni correction (Dunn, 1959) or Storey-Tibshirani (Storey & Tibshirani, 2003)).

Comparing ω values for different gene categories, or individual genes, is indicative of the net selective pressures acting upon these loci. For example, in *Paracoccidioides*, the set of genes evolving under positive selection includes the surface antigen gene GP43, the superoxide dismutase gene SOD3, the alternative oxidase gene AOX, and the thioredoxin gene (Muñoz et al., 2016). Each are virulence-associated genes of fundamental importance in *Paracoccidioides* and other dimorphic fungi. In *Phytophthora* clade 1c, a high proportion of genes annotated as effector genes show signatures of positive selection (300 out of 796) (Raffaele et al., 2010). In *B. dendrobatidis*, CRN-like genes in both *Bd*CAPE and *Bd*CH had the greatest median, upper quartile, and upper tail values of ω compared with other gene categories tested (Farrer, Henk, Garner, et al., 2013). These tests are therefore very useful for identifying selection pressures acting on different genes between populations.

When attempting to understand recent selective processes, alternative methods need to be applied such as the direction of selection (DoS) measure for genes with few substitutions (Stoletzki & Eyre-Walker, 2011). DoS is based on the McDonald–Kreitman test, where the count of fixed synonymous (D_s) and fixed nonsynonymous (D_n) is used in conjunction with the numbers of polymorphisms (in this test defined as sites with any variation within species) and denoted P_n for nonsilent and P_s for silent polymorphisms. Next, using an 2×2 contingency table (McDonald & Kreitman, 1991), deviation from the neutrality index ($\mathrm{NI} = D_s P_n / D_n P_s$ or $(P_n/P_s)/(D_n/D_s)$) can be detected and will indicate positive selection where $D_n/P_n > D_s/P_s$. However, being a ratio of two ratios, the neutrality index is undefined when either D_n or P_s is 0 and tends to be biased and to have a large variance, especially when numbers of observations are small (Stoletzki & Eyre-Walker, 2011). The DoS measure does not have these issues, and so is suitable when the data is sparse. More recently, powerful approaches have been developed that utilize generalized mixed models to estimate selection coefficients for new mutations at a locus and including the synonymous and nonsynonymous mutation rates alongside species divergence times (Eilertson, Booth, & Bustamante, 2012). Such approaches have further been extended to take into account intragenic heterogeneity in the intensity of natural selection (Zhao et al., 2017).

Using the BSM in Codeml, genes with very small Q-values are evidence for positive selection. For example, in *C. gattii*, multiple subclades had low Q values for the cell wall integrity protein SCW1 and the iron regulator 1, while other subclades such as VGI excluding a more divergent isolate had a low Q value for heat shock protein (HSP) 70 (Farrer et al., 2015)—all of which may play roles at the host–pathogen interface. Additionally, two genes (CDR ABC transporter, and ABC-2 type transporter) were independently identified in four subclades of *C. gattii*. Additionally, the PFAM domain "ABC transporter" belonging to a third gene was independently enriched in three of these subclades. Each of these transporters belongs to a single paralog cluster of six genes, which includes the ABC transporter-encoding gene AFR1. This class of gene includes multidrug transporters with azole and fluconozole transporter activity in *C. neoformans* (Sanguinetti et al., 2006), *C. albicans* (Gauthier et al., 2003) and *Penicillium digitatum* (Nakaune, Hamamoto, Imada, Akutsu, & Hibi, 2002). However, the closest *C. gattii* ortholog to AFR1 was not one of the three under selection. While it is likely that selection pressures driving genetic variation in the *C. gattii* population are occurring predominantly in the environment

(*Cryptococcus* is nontransmissable between hosts), they might also result in key pathophysiological differences in humans.

Within-species data on allele frequency spectra are used to detect impacts on natural selection that occur within more recent timeframes. These methods include Tajima's D, which is commonly used to describe genome-wide allele frequency distributions. Tajima's D is a widely used metric that distinguishes between genomic regions that are evolving neutrally (i.e., are under mutation/drift equilibrium) to those that are evolving nonneutrally through the action of selection or demographic processes selection (Tajima, 1989). The biological interpretation of Tajima's D however is not straightforward as divergence from neutral expectations ($D=0$) can be due to different processes that include demographic events alongside the intensity of natural selection. For example, on one hand a negative value of $D<0$ (equating to an excess of rare alleles) can owe to sweeps on a selected polymorphism or population expansion following a genetic bottleneck. On the other hand, a positive value of $D>0$ (equating to a scarcity of rare alleles) can owe to either balancing selection or a demographic contraction. In both cases, correct interpretation of D requires further population genetic analysis. A range of other methods for intrapopulation selection have been developed or used to infer selection, including Fu and Li's D and F, Fay and Wu's H test, long range haplotype test, iHS, LD decay, and F_{ST} (Biswas & Akey, 2006). Different methods may have benefits over others depending on sample size, sequence similarity or distance, population structure, population size, or recombination rates (along with other population-specific traits). Determining the best tools and methods usually requires some benchmarking on the data, testing the effect parameters has on results, and often comparing the results between tests to look for consistency. Ultimately, identifying genes or gene categories that are rapidly changing or are unusually conserved can offer new insights into the biology and pathology of EFPs.

A striking example of the response of fungi to directional selection leading to a novel emerging trait is seen when antifungal drugs are used to treat infectious fungi. Fungicides are an essential component in our armamentarium against fungal disease with sterol demethylation inhibitor (DMI) compounds, such as triazoles and imidazoles, representing the largest class of fungicides that are used in agriculture. These compounds are widely deployed for crop protection with, for instance, over 250,000 kg being used to protect UK crops each year (European Centre for Disease Prevention and Control, 2013) and the global usage being in the thousands of tonnes. In

parallel, azoles are used as frontline drugs to protect humans and other animals against pathogenic fungi. However, the dual-use of DMIs in both clinical and agricultural settings may come at a considerable human cost as in recent years' multiazole resistance in fungi that infect humans has been observed as a widely emerging phenomenon across Europe and beyond. This emergence of resistance has led to the hypothesis that the deployment of azoles in agriculture has led to selection for antifungal resistance not only in target crop pathogens (Cools & Fraaije, 2008; European Centre for Disease Prevention and Control, 2013) but also those fungal species that cooccur in their environment, and that can opportunistically infect humans, specifically the saprophytic genus *Aspergillus* (European Centre for Disease Prevention and Control, 2013). Ergosterol is an essential component of the fungal cell membrane and is the target of triazoles that inhibit its biosynthesis, thereby interfering with the integrity of the fungal cell membrane (Diaz-Guerra, Mellado, Cuenca-Estrella, & Rodriguez-Tudela, 2003). In *Aspergillus*, azole resistance can be an intrinsic phenotype, as it is known to occur in cryptic *Aspergillus* species related to *A. fumigatus*, specifically *A. lentulus* and *A. pseudofischeri* (Van Der Linden, Warris, & Verweij, 2011), whereas wild-type *A. fumigatus* and *A. flavus* are sensitive to these drugs. In *A. fumigatus,* azole resistance is known to be an acquired trait that occurs after azole exposure during medical treatment, or after fungicide exposure in the field where *A. fumigatus* widely occurs in the soil. While a spectrum of resistance mechanisms to azoles has been characterized in *A. fumigatus* (Fraczek et al., 2013; Meneau, Coste, & Sanglard, 2016), azole resistance is frequently the result of mutations in the *cyp51A* gene. Many azole-resistant isolates have nonsynonymous point mutations at codons in this gene, for example, at positions G54, M220, and G138 (Chowdhary, Sharma, Hagen, & Meis, 2014), which are primarily found in patients who have been treated for long periods with azoles (Verweij, Chowdhary, Melchers, & Meis, 2016). However, in addition to mutations that are commonly associated with the de novo acquisition of resistance in the patient, an increasingly large constellation of *cyp51A* mutations are found to occur in "wild" *A. fumigatus*. These mutations are largely characterized by having a tandem repeat (TR) duplication in the promotor region of *cyp51A* linked to structurally important nonsynonymous SNPs (Meis, Chowdhary, Rhodes, Fisher, & Verweij, 2016).

It is now evident that triazole resistance in *Aspergillus* has a global distribution and constitutes a worldwide EFP with important consequences to human health. In some regions, up to 7% of patients are culture-positive

for *Aspergillus* now harbor environmentally associated azole-resistance and azoles are increasingly failing in their role as frontline choices of therapy. Population genomic analysis has been used to show that the most frequently occurring environmental-resistance allele, known as TR_{34}/L98H, occurs on a subset of the observed genetic diversity of *A. fumigatus* with strong linkage disequilibrium being observed, and associations to clonal population sweeps in regions of high azole usage such as India. The balance of evidence suggests that TR_{34}/L98H is a relatively recent and novel evolutionary innovation, and that it is perturbing the natural population genetic structure of *A. fumigatus* in nature as selective sweeps imposed by this allele occur. Fitness costs that are associated with azole-resistance alleles appear to be negligible, and diversification in nature is known to occur as mating occurs leading to the genesis of new combinations of azole-resistance alleles (Abdolrasouli et al., 2015). Thus, strong direction selection through the global usage of azoles appears to have irrevocably perturbed the worldwide population genetic structure of *Aspergillus,* alongside many other plant pathogenic fungi, leading to worldwide breakdown in our ability to use this important class of drugs to secure our health and food security (European Centre for Disease Prevention and Control, 2013).

2.5 Genomic Approaches to Detecting Reproductive Modes, Demographic and Epidemiological Processes in EFPs

2.5.1 Know Your Enemy

Key to the genomic analysis of an EFP is to "know your enemy." Within this context, is the (often novel) EFP a single genotype, a lineage, a species, or a set of species? These distinctions are important as they determine the evolutionary trajectory of the organism by determining the type and rate of evolutionary changes that will occur through time, and how these need to be analyzed within an epidemiological context. Wiley (1978) used an evolutionary concept to define species as "… a single lineage of ancestor descendent populations which maintains its identity from other such lineages and which has its own evolutionary tendencies and historical fate." The evolutionary species concept has been used as the framework that species of fungi can be identified using operational species concepts that use the genealogies inferred from DNA sequences. Of most benefit to analyses of fungal diversity, the system of genealogical concordance phylogenetic species recognition (GCPSR) has been widely used to define evolutionary significant units by identifying the transition from genealogical concordance to conflict (also known as reticulate genealogies) as a means of determining the

limits of species (Dettman, Jacobson, & Taylor, 2003; Taylor et al., 2000). An important use of whole-genome data therefore is to determine the extent that evolutionarily significant units occur within the EFP, be these on a global scale or within a localized outbreak setting.

There are two fundamental means by which fungi and other organisms transmit genes vertically to the next generation, either via clonal reproduction or via mating and recombination (Taylor, Jacobson, & Fisher, 1999). Under a purely clonal reproductive mode, each progeny has as single parent with its genome being an exact mitotic copy of the parental one, and all parts of the parental and progeny genomes share the same evolutionary history. At the other extreme are genetically novel progeny formed by the mating and meiotic recombination of genetically different parental nuclei, events that cause different regions of the progeny genome to have different evolutionary histories. However, many fungi do not fit neatly into these two categories. For instance, on one hand recombination need not be meiotic or sexual because mitotic recombination via parasexuality can mix parental genomes. On the other hand, clonality need not be solely mitotic and asexual, because self-fertilizing or homothallic fungi make meiospores with identical parental and progeny genomes. In addition to the observation that reproductive mode (clonal or recombining) may be uncoupled from reproductive morphology (meiosporic or mitosporic), there is the complication that the same fungus may display different reproductive modes in different localities at different times. These are important distinctions from the point of view of EFPs, as many fungi are flexible in their ability to undergo genetic recombination, hybridization, or HGT (Taylor et al., 1999). This flexibility in life histories allows not only the clonal emergence of pathogenic lineages from their sexual parental species, but can also allow the formation of novel genetic diversity by generating mosaic genomes that may lead to the genesis of new pathogens (Stukenbrock & McDonald, 2008).

Reproductive barriers in fungi are known to evolve more rapidly between sympatric lineages that are in the nascent stages of divergence than between geographically separated allopatric lineages, in a process known as reinforcement (Turner, Jacobson, & Taylor, 2011). As a consequence, the anthropogenic (human-associated) mixing of previously allopatric fungal lineages that still retain the potential for genetic exchange across large genetic distances has the potential to drive rapid macroevolutionary change. Although many outcrossed individuals, or genuine species hybrids, are inviable owing to genome incompatibilities, large phenotypic leaps can be achieved by the resulting "hopeful monsters," potentially leading to host

jumps and increased virulence. Therefore, a nuanced understanding of gene flow within and among fungal lineages is important as recombination is known to novel new interspecific hybrids with novel pathogenic phenotypes as lineages come into contact (Giraud, Gladieux, & Gavrilets, 2010; Inderbitzin, Davis, Bostock, & Subbarao, 2011).

The sequencing of the brewer's yeast *S. cerevisiae* represented a genetic landmark as it was the first fully sequenced eukaryotic genome. From this initial assembled sequence, over a thousand resequenced genomes have now been generated for *S. cerevisiae* and its close relatives leading to an unparalleled genomic description of the evolution of this model globalized fungal species across different spatial and temporal scales (Dujon & Louis, 2017; Liti et al., 2009). Descriptions of global patterns of *S. cerevisiae* genome-wide diversity are now identifying ancestral populations found in South East Asia (Wang, Liu, Liti, Wang, & Bai, 2012) alongside lineages which have undergone global spread through comigrating with humans (Liti et al., 2009). While many genotypes of *S. cerevisiae* are "clean lineages," others show widespread outcrossing that has resulted in gene flow generating mosaic genomes that are characterized by genetic introgressions from other lineages of *S. cerevisiae*, and also via hybridization with other related species of closely wild yeasts such as *Saccharomyces paradoxus*. Therefore, a GCPSR analysis, although not yet formally done, would likely show that the genomes that comprise the *Saccharomyces* clade are evolving in a reticulate manner rather than in a strictly genealogically concordant manner (Dujon & Louis, 2017). Reticulate evolution is likely to be the case for many fungal lineages that we currently recognize as species, and represents a fundamental challenge for the modern fungal taxonomist as well as fungal epidemiologist.

2.5.2 Occurrence of EFPs Caused by Clonal Through to Reticulate Evolution

The correct interpretation of the genetic epidemiology of a fungal outbreak critically depends on understanding how the outbreak isolates are related to the species-wide diversity across the realized global range of the pathogen. A key question is to determine whether the EFP represents the long-distance dispersal of a species resulting in host shifts and the loss of population diversity—clonal evolution, or is a genetic recombinant with novel phenotypic traits—reticulate evolution. Often in the context of an emergence of a novel fungal pathogen, these data can take months, years, or even decades to accrue (but see Islam et al., 2016). However, phylogenomic analysis is likely

to provide crucial understanding of the evolutionary and epidemiological drivers leading to a mycotic outbreak. For example, genetic evidence of the clonal evolution of an EFP following a phylogeographic "leap" from its parental, sexual, population has been forcefully illustrated by the emergence of the aetiological agent of bat white-nose syndrome, *P. destructans* (Blehert et al., 2009). This mycosis emerged in 2006–07 from a single index outbreak site, spreading and devastating multiple species of bats across North America. However, while bats across Europe are infected by this fungus, they appear thus-far unscathed suggesting that European bats have a longer history of coevolution with *P. destructans* compared to their North America conspecifics. Support for this hypothesis initially came from multilocus evidence showing that European isolates of *P. destructans* are highly polymorphic at all loci examined (Leopardi, Blake, & Puechmaille, 2015) and are heterothallic with both mating types coexisting within single bat hibernacula (Palmer et al., 2014). In comparison, recent comparative genome analyses of North American outbreak isolates of *P. destructans* show that they are not only genetically highly homogenous but also comprise a single mating type (Palmer et al., 2014; Trivedi et al., 2017) and show no evidence of recombination. These data strongly support the hypothesis that a single genome of *P. destructans* contaminated North America from a thus-far unidentified location in Europe, followed by clonal amplification and continent-wide spatial expansion of this single genotype.

While the emergence of *P. destructans* presents a dramatic example of a contemporary clonal spatial escape, many other species of EFP show strong similarities to the basic process described earlier. Human-mediated intercontinental trade has been linked clearly to the spread of animal-pathogenic fungi through the transportation of infected vector species. *B. dendrobatidis* has been introduced repeatedly to naive populations worldwide as a consequence of the trade in the infected, yet disease-tolerant species such as North American bullfrogs (*Lithobates catesbeiana*) (Garner et al., 2006) and African clawed frogs (*Xenopus laevis*) (Walker et al., 2008). Recent genome sequencing of a global collection of over 250 genomes of *B. dendrobatidis* has been used to prove that a single genotype, *Bd*GPL, globally emerged in the early 20th century to cause the patterns of amphibian decline seen to date. Analogous to *P. destructans*, population genomic comparisons of sequenced *B. dendrobatidis* isolates show clear patterns of emergence from a defined geographic location, in this case East Asia (O'Hanlon and Fisher, unpublished observation), where isolates of *B. dendrobatidis* show levels of nucleotide diversity that are many fold higher than are seen across

other global regions. However, in contrast to *P. destructans*, the emergence of amphibian chytridiomycosis across over half a century has allowed substantial diversification of the outbreak lineage *Bd*GPL to occur, including the homogenization of large tracts of the polyploid genome through losses of heterozygosity caused by mitotic recombination (Farrer et al., 2011; James et al., 2009). Furthermore, consecutive waves of expansion by *Bd* out of its East Asian home range has allowed globalized lineages to recontact and form recombinant genotypes many decades later (O'Hanlon and Fisher, unpublished observation).

As the rate of interlineage recombination between fungi will be proportional to their contact rates, a prediction is that the globalization of pathogenic fungi will increase the frequency that recombinant genotypes are generated. Confirming this hypothesis, outcrossing to generate novel mosaic genomes among lineages is now increasingly observed for sequenced isolates of *B. dendrobatidis* in regions where lineages are found to occur in sympatry. The process of recombination through secondary contact is potentially important in an epidemiological context as theory and experimentation have shown that virulent lineages can have a competitive advantage that results in increased transmission (de Roode et al., 2005; Karvonen, Rellstab, Louhi, & Jokela, 2012). This implies that the generation of novel genotypes with varied virulence phenotypes may force the epidemiological characteristics of a disease system as well as allowing the generation of novel interlineage recombinant mosaic genomes with novel phenotypes. A case in point here is the formation of a novel pathogen of triticale, *B. graminis triticale*, which evolved through the hybridization of two *formae specialis* from wheat and rye hosts (Menardo et al., 2016) clearly demonstrating that new evolutionarily significant units, and thus EFPs, can be generated through outcrossing.

The use of population genomics is increasingly widely used to map phylogeographic escapes that have led to outbreaks of EFPs. Owing to its ability to cause severe disease in humans, the basidiomycete yeasts *C. neoformans* and *C. gattii* have been subjected to detailed genomic scrutiny. Both species show the existence of strong genetic subdivision into lineages with high statistical support (Farrer et al., 2016; Rhodes, Desjardins, et al., 2017). Whether these represent evolutionary species or not is currently a subject of wide debate as genome-wide tests (Hagen et al., 2015; Menardo et al., 2016) of genealogical concordance have not been performed to date (Hagen et al., 2015; Kwon-Chung et al., 2017). Certainly, the realized potential for interlineage recombination is apparent as hybrid ancestry is readily detected based on the detection of large blocks of shared ancestry

among all three lineages of *C. neoformans* var. *grubii* (lineages VNI, VNII, and VNB) (Rhodes, Desjardins, et al., 2017) and interlineage hybrids between *C. neoformans* and *C. gattii* have been described (Bovers et al., 2006; Engelthaler et al., 2014). However, superimposed upon this background of mosaic lineages, population genomic analysis of both species show very clear evidence of clonal expansions that are associated with clinical disease. *C. neoformans* lineage VNI appears to have expanded globally (likely anciently) due to widening avian host distributions (Litvintseva et al., 2011; Rhodes, Desjardins, et al., 2017), and the emergence of *C. gattii* lineage VGIIa in the Pacific Northwest has recently caused a widely studied outbreak of aggressive clinical disease (Engelthaler et al., 2014; Fraser et al., 2005; Hagen et al., 2013). For fungi that cause disease in plants, clonal expansion causing epidemic outbreaks following long-distance dispersal of infectious propagules has relentlessly attacked agriculturally important crops and damages our ability to safely feed humanity on an annual basis (Fisher et al., 2016, 2012; Fones et al., 2017). Examples here are many (Fisher et al., 2016, 2012; Fones et al., 2017; McDonald & Stukenbrock, 2016) and include the recent emergence of wheat blast caused by a clonal outbreak of *M. oryzae* vectored from South America to Bangladesh with associated catastrophic losses (Islam et al., 2016). This study is notable in that the team was able to sequence and assemble an open-access genome-wide dataset of SNPs derived from a broad global set of isolates within a matter of months in order to identify the likely geographic source of the Bangladesh outbreak, thus illustrating how the future of rapid population genomic analysis of EFPs may unfold.

Beyond describing the spatiotemporal phylodynamic aspects that underpin EFPs, population genomics is leading to an increasingly nuanced understand of how fungi acquire novel pathogenicity traits through the process of HGT. HGT is a special case of hybridization, where a defined genetic locus is transferred between large genetic distances that range from interspecies through to inter-kingdom transfers. An arresting example of a locus-specific HGT leading to the evolution of an EFP was determined through sequencing the genome of the wheat pathogen *P. nodorum* where a gene encoding a host-specific protein toxin (*ToxA*) was identified by homology to a known toxin from another wheat pathogen *P. tritici-repentis*. It is now known that *ToxA* jumped from *P. nodorum* into *P. tritici-repentis* through close genetic linkage to a retrotransposon, sometime in the 1940s resulting in the rapid emergence of aggressive tan spot disease of wheat caused by *P. tritici-repentis* (Friesen et al., 2006). More recent advances in other species have further

detailed the acquisition of novel virulence-associated loci via HGT in *Fusarium pseudograminearum* where horizontal transfers from bacterial and other fungal species were discovered that were clearly associated with virulence in this EFP (Gardiner et al., 2012).

2.5.3 Mutation Rates, Molecular Clocks, and EFPs

A key question that needs to be asked early on when analyzing an outbreak of an EFP is to determine when the genotype (or phenotype) of interest evolved. This question is currently being addressed for a wide variety of EFPs including *Batrachochytrium* and *Cryptococcus* sp. For the latter, comparative genomics has been used to compare orthologous-coding regions in order to determine the proportion of nucleotide sites that have undergone substitutions. Such analyses were recently used to show that ~17% of sites were polymorphic when representative genomes of *C. gattii* and *C. neoformans* were compared against one another. If fungal mutation rates lie between 0.9×10^{-9} and 16.7×10^{-9} substitutions per nucleotide per year as has been calculated across a range of filamentous fungi (Kasuga, White, & Taylor, 2002; Sharpton, Neafsey, Galagan, & Taylor, 2008), then the divergence time between these species would lie between 5.2×10^{6} and 96.7×10^{6} years ago, which is concordant with the breakup of the Pangean supercontinent causing allopatric speciation of *C. neoformans* and *C. gattii* through a model of vicariance (Casadevall, Freij, Hann-Soden, & Taylor, 2017). However, a cautionary note needs to be interjected here: Accurate estimates of substitution rates are crucial in order to investigate the evolutionary history of virtually any species. It becoming increasingly apparent that "the molecular clock" is not a one-size-fits-all and in fact can vary by two orders of magnitude even within a single lineage. A case in point here are recent investigations into the population genomics of microevolution in serially collected isolates of *C. neoformans* from HIV/AIDs patients with cryptococcal meningitis in South Africa. While comparisons revealed a clonal relationship for most pairs of isolates recovered before and after relapse of the original infection, one pair of isolates manifested a substitution rate that was greatly inflated above that of the others. Further investigation showed the occurrence of nonsense mutations in DNA mismatch repair pathways leading to the evolution of a hypermutator phenotype (Rhodes, Beale, et al., 2017).

The occurrence of hypermutators in fungal populations is now being described more widely, not only in *Cryptococcus* (Boyce et al., 2017; Rhodes, Beale, et al., 2017) but also species of *Candida* (Healey et al.,

2016). This means that there is a real need to carefully scrutinize the range of substitution rates within and between species, and to not assume a "one-size-fits-all" approach as this is almost certainly incorrect. A further complexity is that nuclear genomes that have undergone recombination are mosaics of gene genealogies with varied evolutionary histories, which can have the effect of creating a false signal of mutation. Therefore, in order to accurately estimate substitution rates, efforts need to be made to control for the effects of recombination, either by directly partitioning the data around recombining sites as was done to date the origin of the *Batrachochytrium* hypervirulent lineage *Bd*GPL to the 20th century (Farrer et al., 2011) or by choosing a nonrecombining section of the genome, such as the mitochondrial DNA.

Once appropriate genomic regions have been identified, then the most direct approach is to use root-to-tip estimations of substitution rates for collections, where the MRCAs are known from either a fossil record or time-dated biological events such as date of isolation. Critically, for root-to-tip estimations of rates to work, studies need to be able to access time-stamped genomic data that is measurably evolving through time (Rieux & Balloux, 2016). If time-calibrated phylogenies that are measurably evolving can be constructed, then sophisticated analyses of demographic histories can be inferred including the estimation of effective population sizes through time, implemented in coalescent-based algorithms such as BEAST (Drummond & Rambaut, 2007). Such analytical approaches have proven critical to understanding pandemics of viruses such as HIV (Faria et al., 2014) and the spread of bacterial pathogens (Croucher & Didelot, 2015). However, beyond the single example of our attempt to understand the date of *Bd*GPLs origin (Farrer et al., 2011), we are unaware of serious attempts to analyze EFPs using modern tip-calibrated approaches to estimating fungal molecular clocks with rigor.

3. EPIGENOMIC VARIATION WITHIN AND BETWEEN POPULATIONS OF EFPs

Phenotypic traits of EFPs are determined by their genomes, the environment, and their interactions. Epigenetics was a name given by Conrad Waddington to "the branch of biology which studies the causal interactions between genes and their products, which bring the phenotype into being" (Goldberg, Allis, & Bernstein, 2007). However, the term has since been used to describe a range of processes: for example, the temporal/spatial control of

gene activity during animal development (Holliday, 1990), and changes in phenotype caused without alterations in the DNA sequence, that are either not necessarily heritable (Bernstein et al., 2010), or are exclusively heritable (Berger, Kouzarides, Shiekhattar, & Shilatifard, 2009). The latter definition (and the others listed) includes a wide range of processes ranging from base modifications such as cytosine methylation and cytosine hydroxymethylation, as well as histone posttranslational modifications, nucleosome positioning, and ncRNA regulating gene expression.

Epigenetic processes often culminate in differential expression, e.g., nucleosome occupancy negatively correlating with gene expression (Leach et al., 2016). Indeed, detecting expression values between conditions, or between isolates or even orthologous genes between species remains a key question for many EFPs and has been discussed in some detail in the previous sections. Many tools have been made available for detecting levels of expression and expression differences. A key normalized metric from RNAseq is the "reads per kilobase of transcript model per million reads" (RPKM). RPKM can be calculated by several tools such as EdgeR (Robinson, McCarthy, & Smyth, 2010) or Cufflinks (Trapnell et al., 2012). Alternatively, RPKM can be calculated simply by (1) counting the total number of reads in a sample divided by 1 million to give the "per million scaling factor" (PMSF), (2) dividing the number of reads aligned to a gene by the PMSF, and dividing that by the length of the gene in kilobases. A slightly updated metric is FPKM that looks at the number of fragments (the number of paired or individual reads that aligned). For single-end reads, FPKM equals RPKM. Finally, the transcripts per kilobase million (TPM) normalizes for the gene length first (rather than the scaling factor) and provides a relative abundance of transcripts. FPKM and RPKM can be further normalized using the trimmed mean of *M*-values (TMM) (Robinson & Oshlack, 2010), which includes additional scaling factors on the upper and lower expression values of the data, as is implemented in tools such as EdgeR (Robinson et al., 2010). Although each of these expression value metrics is designed to normalize RNAseq across samples or datasets, each may ultimately have a bias for longer or small gene families, library preparation or GC content, which should be identified during an analysis of differential expression.

Gene expression values (i.e., TMM, TPKM, or TPM) across multiple isolates or experiments are usually compared during differential expression analysis, which can require up to 12 biological replicates for the greatest accuracy rates (Schurch et al., 2016), although in practice usually only three are generated due to cost. Tools such as EdgeR (Robinson et al., 2010) and

Deseq2 (Love, Huber, & Anders, 2014) identify differentially expressed transcripts based on a generalized linear model for each gene assuming a negative binomial distribution and includes several other steps to eliminate bias in long genes or minimize "noisy" expression data. Other tools include DEGseq (Wang, Feng, Wang, Wang, & Zhang, 2010) which is also an R bioconductor package and assumes a poisson model which is appropriate for technical replicates but may overestimate expression differences between conditions. MMseq (multimapping RNA-seq analysis) (Turro, Astle, & Tavaré, 2014) to detect allele or isoform-specific expression and Cuffdiff (Cufflinks' method for estimating differential expression) (Trapnell et al., 2012) using quartile-based normalization are additional tools that may provide comparable or better results, and LOX to examine differential expression across multiple experiments, time points, or treatments (Zhang, López-Giráldez, & Townsend, 2010). Ultimately, studies usually have a defined cutoff, e.g., log fold changes between conditions, and/or FDR rates to identify genes that are changing most rapidly. Plots such as Volcano and MA plots can show the distribution of expression values for all genes, and those that are considered differentially expressed, thereby highlighting biases of those methods, e.g., bias of low average counts of reads/transcripts per million.

Numerous examples of differential expression have been discussed in the previous chapter, such as the secreted clade of G2M36 genes ($n = 57$) unique to *B. salamandrivorans*, which are mostly upregulated in salamander skin (Farrer et al., 2017). Notably, the study also generated a transcriptome of the Wenxian knobby newt (*T. wenxianensis*) to identify host genes that were differentially expressed during infection. Emerging fungal diseases are often nonmodel organisms, as is the case for *B. salamandrivorans*, and will themselves infect nonmodel organism hosts. To effectively study the genomics and epigenomics of these diseases, and their effect on the host, it is essential to move away from model-based systems and generate resources such as draft genome assemblies and gene sets for the growing repertoire of EFP's and their hosts.

The associations of mutations and changes in fitness, as well as transcriptional regulation, during pathogenicity are beginning to be characterized within a multitude of eukaryotic pathogens, e.g., *Cryptococcus* (Magditch, Liu, Xue, & Idnurm, 2012; Panepinto & Williamson, 2006). However, the modifications of both DNA and histones that play a key role in transcriptional regulation are to date largely uncharacterized in EFPs. In eukaryotes, histones assemble into octomers called nucleosomes, which wrap around

approximately 147 base pairs of DNA (Stroud et al., 2012). While the position of each histone can be mapped independently by ChIP-seq, including variants of each type, a single type may be used as a proxy for nucleosome positions. Variation in histone-binding sites is found between isolates of fungal pathogens, as well as varying upon condition such as *C. albicans* during heat shock (Leach et al., 2016). Furthermore, nucleosome levels in *C. albicans* decrease near to the transcription factor-binding sites of key pathogenicity genes, allowing activation by transcription factors and RNA polymerase (Leach et al., 2016).

Histones undergo posttranslational modifications on their N-terminal tails that alter their interactions with the DNA and other proteins that they bind. Modifications can be made to any of the four types of histones at several amino acid sites and can include acetylation, phosphorylation, methylation, deamination/citrullination (arginine ➔ citrulline), β-*N*-acetylglucosamination, ADP ribosylation, ubiquitination and small ubiquitin-like modifier (SUMO)-lyation, tail-clipping, and proline isomerization (Bannister & Kouzarides, 2011). These modifications ultimately alter the chromatin structure, which can manifest into changes in transcription, repair, replication, and recombination. For example, acetylation of lysine residues on H3 and H4 by protein complexes involving histone acetyltransferases (HATs) is associated with active transcription for several fungal pathogens (Jeon, Kwon, & Lee, 2014). Notably, the Rtt109 HAT is responsible for acetylation of H3K56 and contributes to pathogenicity of *C. albicans* in mouse macrophages (Lopes da Rosa, Boyartchuk, Zhu, & Kaufman, 2010). Another family of HATs are the Gcn5-related *N*-acetyltransferases (GNAT), including the GCN5 protein implicated in *C. neoformans* growth rates at high temperatures, capsule attachment, and tolerance of oxidative stress (O'Meara, Hay, Price, Giles, & Alspaugh, 2010).

DNA methylation is another important mechanism for epigenetic changes regulating gene expression and transposon silencing (Lister et al., 2009). Whole-genome bisulfite sequencing and methylated DNA immunoprecipitation are methods to profile the methylation of cytosine (carbon 5) to 5-methylcytosine (5-meC) in eukaryotes, generally within cytosine-rich genomic islands (CpG, CpHpG, and CpHpH) (Lou et al., 2014). DNA methylation is achieved via a number of DNA methyltransferase (DNMT), dependent on the species, resulting in 5-meC that can be heritable (e.g., via DNMT1 and UHRF1) (Law & Jacobsen, 2010). 5-meC is widespread in bacteria, plants, and mammalian cells, but differentially

conserved across the fungal kingdom. Notably, 5-meC appears to be absent in a number of fungal genera including *Saccharomyces* and *Pichia* (Capuano, Mülleder, Kok, Blom, & Ralser, 2014). However, in *Neurospora* 5-meC within CpG islands is located in ex-transposons targeted by RIP mutations, where its presence is dependent on a single DNMT named DIM-2, directed by a histone H3 methyltransferase (Selker et al., 2003).

In *C. neoformans* isolate H99, Huff et al. have identified DNMT5 as a CG-specific DNMT and show that knockouts appear to completely remove 5-meC (Huff & Zilberman, 2014, p. 1). Separately, DNMT5 has been implicated in infection in mice (Liu et al., 2008), where knockouts show significantly reduced virulence. 5-meC in *Cryptococcus* is primarily associated with transposable elements, and the methylation directly disfavors nucleosome binding (Huff & Zilberman, 2014, p. 1) (determined using micrococcal nuclease (MNase) to digest chromatin followed by sequencing). Huff et al. show that 5-meC is negatively associated with nucleosome positions, but it remains to be shown how the patterns and associations with nucleosomes varies between isolates or during infection, and as suggested by Liu et al., it may reveal insights into the mechanisms of infection (Liu et al., 2008).

Epigenomics in fungal pathology remains an active of area of research that compliments the larger field of genomics (i.e., DNAseq) in identifying new genotypic features of EFPs and particularly dynamic changes associated with virulence traits. However, since most fungal pathogens remain unculturable, and some (such as *Microsporidia*) are obligate intracellular pathogens—obtaining high quality and sufficient depth of coverage for RNAseq, let alone ChIPseq or Methylseq remains an obstacle. An increase in sampling across fungal pathogens and their nonpathogenic relatives, especially for generating new high-quality genomes for comparison, but also transcriptomics is likely to improve our understanding of fungal pathogenesis. Sampling nonpathogenic relatives will require a move away from focusing solely on outbreak strains, and also looking for fungal relatives in host populations that are not experiencing population declines may yield novel-related isolates. The field of metagenomics also promises to identify new locations and relatives for EFPs.

ncRNA such as miRNA and siRNA of the RNAi pathways are prominent epigenetic components found throughout the fungal kingdom, where they function to silence or downregulate gene expression via complimentary sequences to mRNA targets (Pasquinelli, 2012) or gene promoters (Chu, Kalantari, Dodd, & Corey, 2012). RNAi is achieved via either microRNA

(miRNA) derived from single-stranded RNA transcripts that fold to form ~70 nt hairpins, or small interfering RNAs (siRNAs; short interfering RNA; silencing RNA) that derive from longer regions of double-stranded RNA. siRNA ultimately cleaves sequence-specific mRNAs, compared with miRNA that has reduced specificity and therefore may target a wider range of mRNAs (Lam, Chow, Zhang, & Leung, 2015). Both miRNA and siRNA are cleaved by the RNase III endoribonuclease Dicer (Dicer-1 and Dicer-2, respectively) prior to being incorporated into either the cytoplasmic RNA-induced silencing complex (RISC) or the nuclear RNA-induced transcriptional silencing (RITS) complex, where the RNA (guide strand) binds target mRNA (such as miRNA response elements; MRE; found in 3′ UTRs), which is cleaved by the PIWI domain of a catalytic Argonaute protein (a major component of both the RISC and RITS) thereby causing degradation of the transcript. Another category of RNA silencing molecules is the Dicer-independent PIWI-associated/interacting RNAs (piRNAs), some of which are classified as repeat-associated small interfering RNA (rasiRNA)—however, both sets are thought to be absent in the fungal kingdom.

RNAi silencing machinery (in contrast to piRNA and rasiRNA) is prominent throughout the fungal kingdom, especially filamentous fungi, although is lost sporadically in some species of both yeasts and filamentous fungi (Dang, Yang, Xue, & Liu, 2011). Excitingly, exogenous/artificial (in addition to endogenous) miRNA and siRNA derived from double-stranded RNA or hairpin RNA with complementary sequence to target gene promoters (Chu et al., 2012) or mRNA targets (Pasquinelli, 2012) are being increasingly used for therapeutics against fungi that cause disease in plants (Duan, Wang, & Guo, 2012) and humans (Khatri & Rajam, 2007). For example, a synthetic 23-nucleotide siRNA was designed with complementary base pairs to the sequence of a key polyamine biosynthesis gene (ornithine decarboxylase) in *A. nidulans* required for normal growth, resulting in a reduction in mycelial growth, target mRNA titers, and cellular polyamine concentrations (Khatri & Rajam, 2007). Despite their important role in endogenous gene control (especially transposons), and their potential therapeutic role, endogenous miRNAs and siRNAs (and their respective targets) are not routinely predicted from the genome sequence, despite various in silico strategies existing (i.e., Bengert & Dandekar, 2005). Currently, the extent that RNAi has a role on gene regulation in many fungal pathogens including EFPs is unclear. However, as described later, the study of RNAi is a rapidly emerging field that holds great promise not only as a tool for

understanding fungal virulence but also as a novel approach to disrupt fungal pathogenicity.

RNAi has been shown in numerous biological roles across the fungal kingdom. For example, *N. crassa* can initiate potent RNAi-mediated gene silencing to defend against viral and transposon invasion (Dang et al., 2011). Other functions of RNAi include sex-induced silencing in *C. neoformans*, which is mediated by RNAi via sequence-specific small RNAs (Wang, Hsueh, et al., 2010). Interestingly, one lineage of the related *C. gattii* (VGII) is missing PAZ, Piwi, and DUF1785 domains, all of which are components of the RNAi machinery. This loss of RNAi has been hypothesized to contribute to increased genome plasticity in this lineage that may have contributed to specific hypervirulent traits in VGII (D'Souza et al., 2011; Farrer et al., 2015; Wang, Hsueh, et al., 2010).

The discovery that communication between host and pathogen can occur through the transfer of extracellular microvesicles (ExMVs) has opened a new research field into the horizontal transfer of bioactive molecules in cell-to-cell communication (Ratajczak & Ratajczak, 2016). It has now been well documented that horizontal transfer of miRNAs occurs between fungal and host cells occurs via the action of ExMVs, that this transfer is bidirectional, and that the transfer of miRNAs can result in RNAi that induces host susceptibility to a pathogen (Wang et al., 2016). RNAi that is mediated via such "cross-kingdom" transfer of ExMVs has been show to occur in the aggressive pathogenic fungus *Blumeria cinerea*, where selective silencing of host plant immune genes occurs by the introduction of miRNA virulence effectors (Weiberg et al., 2013). The characterization of miRNAs in EFPs using high-throughput RNA sequencing approaches followed by identification of matching host sequences therefore offers an opportunity to identify potential RNA-based virulence effectors. Moreover, the recognition that virulence can be mediated epigenomically has opened up new opportunities to control fungal diseases using nonfungicide means. For instance, recent work has shown that in *B. cinerea*, the majority of miRNA effectors are derived from retrotransposon LTRs which, when miRNA production is knocked-down through deletion of the key component of the *B. cinerea* RNAi pathway Dicer, that virulence is abrogated *in planta* (Wang et al., 2016). This seminal result was then extended to show that engineering the host plant, in this case *Arabidopsis*, to express the anti-Dicer RNAi conferred resistance against *B. cinerea* demonstrating that host-induced gene silencing of the pathogen occurs. Finally, it was then demonstrated that the simple application of synthetic environmental anti-Dicer RNAi to the fungus, while in the act of infecting the host, resulted in

the attenuation of virulence as the fungus took up the RNAi constructs via ExMVs. Studies such as these showing that pathogenic fungi can be epigenomically silenced through nonfungicide-based means, and by the simple application of a nontoxic and highly specific RNAi construct, are clearly a disruptive approach that has broad applicability to a broad span of the nonmodel EFPs that we have discussing here and shows much promise.

4. CONCLUDING REMARKS

Presently, phylogenomic, comparative genomic, and epigenomic methods are becoming the *modus operandi* for detection and characterization of virulence determinants and epidemiological parameters among EFPs (Hasman et al., 2014; Lecuit & Eloit, 2014), which are themselves increasingly taking center stage for contemporaneous epidemics of plants, humans, and other animals (Fisher et al., 2016). Testaments to the success of this approach are the many examples of traits underpinning EFP that have been identified using these methods. While the full scope and potential of these experimental techniques and resulting compendiums of data are being realized, many challenges remain. Importantly, the continuing adoption of best practices, repeatable protocols, standardizations, and data storage need to be developed to guide future studies working with these new data types and developing powerful new experimental designs. The rapidity of disease outbreaks far outpaces current systems for genomic/epigenomic data acquisition and distribution. Expeditious evaluation and disease mitigation require collaborative research groups that can contribute and coordinate the varied expertise and skills that are needed to tackle new outbreaks. Given the pace and scope of genomics and epigenomic techniques, these fields will likely continue to shape our understanding of pathogen evolution and provide additional approaches to combatting the increasing threat that EFPs pose to biodiversity and ecosystem health.

ACKNOWLEDGMENTS

R.A.F. is supported by an MIT/Wellcome Trust Fellowship. M.C.F. is supported by the UK Natural Research Council, the Leverhulme Trust, and the Morris Animal Foundation.

REFERENCES

Abbey, D., Hickman, M., Gresham, D., & Berman, J. (2011). High-resolution SNP/CGH microarrays reveal the accumulation of loss of heterozygosity in commonly used Candida albicans strains. *G3 (Bethesda, Md)*, *1*, 523–530. https://doi.org/10.1534/g3.111.000885.

Abdolrasouli, A., Rhodes, J., Beale, M. A., Hagen, F., Rogers, T. R., Chowdhary, A., et al. (2015). Genomic context of azole resistance mutations in Aspergillus fumigatus

determined using whole-genome sequencing. *mBio6*. e00536. https://doi.org/10.1128/mBio.00536-15.

Abramyan, J., & Stajich, J. E. (2012). Species-specific chitin-binding module 18 expansion in the amphibian pathogen Batrachochytrium dendrobatidis. *mBio 3*. e00150-112. https://doi.org/10.1128/mBio.00150-12.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*, 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2.

Amselem, J., Lebrun, M.-H., & Quesneville, H. (2015). Whole genome comparative analysis of transposable elements provides new insight into mechanisms of their inactivation in fungal genomes. *BMC Genomics*, *16*, 141. https://doi.org/10.1186/s12864-015-1347-1.

Arnold, B., Bomblies, K., & Wakeley, J. (2012). Extending coalescent theory to autotetraploids. *Genetics*, *192*, 195–204. https://doi.org/10.1534/genetics.112.140582.

Bairoch, A., & Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research*, *28*, 45–48.

Banfield, D. K. (2011). Mechanisms of protein retention in the Golgi. *Cold Spring Harbor Perspectives in Biology*, *3*(8), a005264. https://doi.org/10.1101/cshperspect.a005264.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., et al. (2012). SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, *19*, 455–477. https://doi.org/10.1089/cmb.2012.0021.

Bannister, A. J., & Kouzarides, T. (2011). Regulation of chromatin by histone modifications. *Cell Research*, *21*, 381–395. https://doi.org/10.1038/cr.2011.22.

Bellaousov, S., & Mathews, D. H. (2010). ProbKnot: Fast prediction of RNA secondary structure including pseudoknots. *RNA*, *16*, 1870–1880. https://doi.org/10.1261/rna.2125310.

Belle, E. M. S., Piganeau, G., Gardner, M., & Eyre-Walker, A. (2005). An investigation of the variation in the transition bias among various animal mitochondrial DNA. *Gene*, *355*, 58–66. https://doi.org/10.1016/j.gene.2005.05.019.

Bendtsen, J. D., Jensen, L. J., Blom, N., Von Heijne, G., & Brunak, S. (2004). Feature-based prediction of non-classical and leaderless protein secretion. *Protein Engineering, Design & Selection: PEDS*, *17*, 349–356. https://doi.org/10.1093/protein/gzh037.

Bengert, P., & Dandekar, T. (2005). Current efforts in the analysis of RNAi and RNAi target genes. *Briefings in Bioinformatics*, *6*, 72–85.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B: Methodological*, *57*, 289–300.

Benson, G. (1999). Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Research*, *27*, 573–580.

Benson, D. A., Karsch-Mizrachi, I., Clark, K., Lipman, D. J., Ostell, J., & Sayers, E. W. (2012). GenBank. *Nucleic Acids Research*, *40*, D48–D53. https://doi.org/10.1093/nar/gkr1202.

Berger, S. L., Kouzarides, T., Shiekhattar, R., & Shilatifard, A. (2009). An operational definition of epigenetics. *Genes & Development*, *23*, 781–783. https://doi.org/10.1101/gad.1787609.

Berger, L., Speare, R., Daszak, P., Green, D. E., Cunningham, A. A., Goggin, C. L., et al. (1998). Chytridiomycosis causes amphibian mortality associated with population declines in the rain forests of Australia and Central America. *Proceedings of the National Academy of Sciences of the United States of America*, *95*, 9031–9036.

Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., Milosavljevic, A., Meissner, A., et al. (2010). The NIH roadmap epigenomics mapping consortium. *Nature Biotechnology*, *28*, 1045–1048. https://doi.org/10.1038/nbt1010-1045.

Besemer, J., & Borodovsky, M. (1999). Heuristic approach to deriving models for gene finding. *Nucleic Acids Research*, *27*, 3911–3920.

Bigirimana, V. d. P., Hua, G. K. H., Nyamangyoku, O. I., & Höfte, M. (2015). Rice sheath rot: An emerging ubiquitous destructive disease complex. *Frontiers in Plant Science*, *6*, 1066. https://doi.org/10.3389/fpls.2015.01066.

Birney, E., Clamp, M., & Durbin, R. (2004). GeneWise and genomewise. *Genome Research*, *14*, 988–995. https://doi.org/10.1101/gr.1865504.

Biswas, S., & Akey, J. M. (2006). Genomic insights into positive selection. *Trends in Genetics: TIG*, *22*, 437–446. https://doi.org/10.1016/j.tig.2006.06.005.

Blackwell, M. (2011). The fungi: 1, 2, 3 ... 5.1 million species? *American Journal of Botany*, *98*, 426–438. https://doi.org/10.3732/ajb.1000298.

Blanchette, M., Kent, W. J., Riemer, C., Elnitski, L., Smit, A. F. A., Roskin, K. M., et al. (2004). Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Research*, *14*, 708–715. https://doi.org/10.1101/gr.1933104.

Blanco, E., Parra, G., & Guigó, R. (2007). Using geneid to identify genes. *Current Protocols in Bioinformatics* chapter 4, Unit 4.3https://doi.org/10.1002/0471250953.bi0403s18.

Blehert, D. S., Hicks, A. C., Behr, M., Meteyer, C. U., Berlowski-Zier, B. M., Buckles, E. L., et al. (2009). Bat white-nose syndrome: An emerging fungal pathogen? *Science*, *323*, 227. https://doi.org/10.1126/science.1163874.

Boetzer, M., & Pirovano, W. (2014). SSPACE-LongRead: Scaffolding bacterial draft genomes using long read sequence information. *BMC Bioinformatics*, *15*, 211. https://doi.org/10.1186/1471-2105-15-211.

Bos, J. I. B., Kanneganti, T.-D., Young, C., Cakir, C., Huitema, E., Win, J., et al. (2006). The C-terminal half of Phytophthora infestans RXLR effector AVR3a is sufficient to trigger R3a-mediated hypersensitivity and suppress INF1-induced cell death in Nicotiana benthamiana. *The Plant Journal: For Cell and Molecular Biology*, *48*, 165–176. https://doi.org/10.1111/j.1365-313X.2006.02866.x.

Bovers, M., Hagen, F., Kuramae, E. E., Diaz, M. R., Spanjaard, L., Dromer, F., et al. (2006). Unique hybrids between the fungal pathogens Cryptococcus neoformans and Cryptococcus gattii. *FEMS Yeast Research*, *6*, 599–607. https://doi.org/10.1111/j.1567-1364.2006.00082.x.

Boyce, K. J., Wang, Y., Verma, S., Shakya, V. P. S., Xue, C., & Idnurm, A. (2017). Mismatch repair of DNA replication errors contributes to microevolution in the pathogenic fungus Cryptococcus neoformans. *mBio*, *8*(3), pii: e00595-17. https://doi.org/10.1128/mBio.00595-17.

Bradnam, K. R., Fass, J. N., Alexandrov, A., Baranay, P., Bechner, M., Birol, I., et al. (2013). Assemblathon 2: Evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience*, *2*, 10. https://doi.org/10.1186/2047-217X-2-10.

Butler, J., MacCallum, I., Kleber, M., Shlyakhter, I. A., Belmonte, M. K., Lander, E. S., et al. (2008). ALLPATHS: De novo assembly of whole-genome shotgun microreads. *Genome Research*, *18*, 810–820. https://doi.org/10.1101/gr.7337908.

Büttner, P., Koch, F., Voigt, K., Quidde, T., Risch, S., Blaich, R., et al. (1994). Variations in ploidy among isolates of Botrytis cinerea: Implications for genetic and molecular analyses. *Current Genetics*, *25*, 445–450.

Byrnes, E. J., III, Li, W., Lewit, Y., Ma, H., Voelz, K., Ren, P., et al. (2010). Emergence and pathogenicity of highly virulent Cryptococcus gattii genotypes in the Northwest United States. *PLoS Pathogens 6*. e1000850. https://doi.org/10.1371/journal.ppat.1000850.

Cantarel, B. L., Korf, I., Robb, S. M. C., Parra, G., Ross, E., Moore, B., et al. (2008). MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Research*, *18*, 188–196. https://doi.org/10.1101/gr.6743907.

Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (2009). trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics (Oxford, England)*, *25*, 1972–1973. https://doi.org/10.1093/bioinformatics/btp348.

Capuano, F., Mülleder, M., Kok, R., Blom, H. J., & Ralser, M. (2014). Cytosine DNA methylation is found in Drosophila melanogaster but absent in Saccharomyces cerevisiae, Schizosaccharomyces pombe, and other yeast species. *Analytical Chemistry*, *86*, 3697–3702. https://doi.org/10.1021/ac500447w.

Carr, J., & Shearer, G. (1998). Genome size, complexity, and ploidy of the pathogenic fungus Histoplasma capsulatum. *Journal of Bacteriology*, *180*, 6697–6703.

Casadevall, A., Freij, J. B., Hann-Soden, C., & Taylor, J. (2017). Continental drift and speciation of the *Cryptococcus neoformans* and *Cryptococcus gattii* species complexes. *mSphere2*. e00103-17. https://doi.org/10.1128/mSphere.00103-17.

Chen, J.-L., & Greider, C. W. (2005). Functional analysis of the pseudoknot structure in human telomerase RNA. *Proceedings of the National Academy of Sciences of the United States of America*, *102*, 8080–8085. https://doi.org/10.1073/pnas.0502259102.

Chen, X.-R., Xing, Y.-P., Li, Y.-P., Tong, Y.-H., & Xu, J.-Y. (2013). RNA-Seq reveals infection-related gene expression changes in Phytophthora capsici. *PLoS One8*. , https://doi.org/10.1371/journal.pone.0074588.

Chowdhary, A., Sharma, C., Hagen, F., & Meis, J. F. (2014). Exploring azole antifungal drug resistance in Aspergillus fumigatus with special reference to resistance mechanisms. *Future Microbiology*, *9*, 697–711. https://doi.org/10.2217/fmb.14.27.

Chowdhary, A., Sharma, C., & Meis, J. F. (2017). Candida auris: A rapidly emerging cause of hospital-acquired multidrug-resistant fungal infections globally. *PLoS Pathogens13*. e1006290. https://doi.org/10.1371/journal.ppat.1006290.

Chu, Y., Kalantari, R., Dodd, D. W., & Corey, D. R. (2012). Transcriptional silencing by hairpin RNAs complementary to a gene promoter. *Nucleic Acid Therapeutics*, *22*, 147–151. https://doi.org/10.1089/nat.2012.0360.

Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., et al. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology*, *31*, 213–219. https://doi.org/10.1038/nbt.2514.

Cole, J. R., Wang, Q., Fish, J. A., Chai, B., McGarrell, D. M., Sun, Y., et al. (2014). Ribosomal database project: Data and tools for high throughput rRNA analysis. *Nucleic Acids Research*, *42*, D633–642. https://doi.org/10.1093/nar/gkt1244.

Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., & Robles, M. (2005). Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics (Oxford, England)*, *21*, 3674–3676. https://doi.org/10.1093/bioinformatics/bti610.

Cools, H. J., & Fraaije, B. A. (2008). Are azole fungicides losing ground against Septoria wheat disease? Resistance mechanisms in Mycosphaerella graminicola. *Pest Management Science*, *64*, 681–684. https://doi.org/10.1002/ps.1568.

Cooper, D. N., Mort, M., Stenson, P. D., Ball, E. V., & Chuzhanova, N. A. (2010). Methylation-mediated deamination of 5-methylcytosine appears to give rise to mutations causing human inherited disease in CpNpG trinucleotides, as well as in CpG dinucleotides. *Human Genomics*, *4*, 406–410.

Croucher, N. J., & Didelot, X. (2015). The application of genomics to tracing bacterial pathogen transmission. *Current Opinion in Microbiology*, *23*, 62–67. https://doi.org/10.1016/j.mib.2014.11.004.

Cushion, M. T., & Stringer, J. R. (2010). Stealth and opportunism: Alternative lifestyles of species in the fungal genus Pneumocystis. *Annual Review of Microbiology*, *64*, 431–452. https://doi.org/10.1146/annurev.micro.112408.134335.

D'Souza, C. A., Kronstad, J. W., Taylor, G., Warren, R., Yuen, M., Hu, G., et al. (2011). Genome variation in Cryptococcus gattii, an emerging pathogen of immunocompetent hosts. *mBio 2*. e00342-310. https://doi.org/10.1128/mBio.00342-10.

Dang, Y., Yang, Q., Xue, Z., & Liu, Y. (2011). RNA interference in fungi: Pathways, functions, and applications. *Eukaryotic Cell*, *10*, 1148–1155. https://doi.org/10.1128/EC.05109-11.

de Roode, J. C., Pansini, R., Cheesman, S. J., Helinski, M. E. H., Huijben, S., Wargo, A. R., et al. (2005). Virulence and competitive ability in genetically diverse malaria infections. *Proceedings of the National Academy of Sciences of the United States of America*, *102*, 7624–7628. https://doi.org/10.1073/pnas.0500078102.

Dettman, J. R., Jacobson, D. J., & Taylor, J. W. (2003). A multilocus genealogical approach to phylogenetic species recognition in the model eukaryote Neurospora. *Evolution; International Journal of Organic Evolution*, *57*, 2703–2720.

Dewey, C. N. (2007). Aligning multiple whole genomes with Mercator and MAVID. *Methods in Molecular Biology (Clifton, NJ)*, *395*, 221–236.

Diaz-Guerra, T. M., Mellado, E., Cuenca-Estrella, M., & Rodriguez-Tudela, J. L. (2003). A point mutation in the 14alpha-sterol demethylase gene cyp51A contributes to itraconazole resistance in Aspergillus fumigatus. *Antimicrobial Agents and Chemotherapy*, *47*, 1120–1124.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)*, *29*, 15–21. https://doi.org/10.1093/bioinformatics/bts635.

Dong, S., Shen, Z., Xu, L., & Zhu, F. (2010). Sequence and phylogenetic analysis of SSU rRNA gene of five microsporidia. *Current Microbiology*, *60*, 30–37. https://doi.org/10.1007/s00284-009-9495-7.

Drummond, A. J., & Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, 7, 214. https://doi.org/10.1186/1471-2148-7-214.

Duan, C.-G., Wang, C.-H., & Guo, H.-S. (2012). Application of RNA silencing to plant disease resistance. *Silence*, *3*, 5. https://doi.org/10.1186/1758-907X-3-5.

Dujon, B. A., & Louis, E. J. (2017). Genome diversity and evolution in the budding yeasts (Saccharomycotina). *Genetics*, *206*, 717–750. https://doi.org/10.1534/genetics.116.199216.

Dunn, O. J. (1959). Estimation of the medians for dependent variables. *Annals of Mathematical Statistics*, *30*, 192–197. https://doi.org/10.1214/aoms/1177706374.

Eddy, S. R. (2011). Accelerated profile HMM searches. *PLoS Computational Biology* 7. e1002195. https://doi.org/10.1371/journal.pcbi.1002195.

Eddy, S. R., & Durbin, R. (1994). RNA sequence analysis using covariance models. *Nucleic Acids Research*, *22*, 2079–2088.

Edgar, R. C. (2004a). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*, 1792–1797. https://doi.org/10.1093/nar/gkh340.

Edgar, R. C. (2004b). MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, *5*, 113. https://doi.org/10.1186/1471-2105-5-113.

Eilertson, K. E., Booth, J. G., & Bustamante, C. D. (2012). SnIPRE: Selection inference using a poisson random effects model. *PLoS Computational Biology 8*. e1002806. https://doi.org/10.1371/journal.pcbi.1002806.

Engelthaler, D. M., Hicks, N. D., Gillece, J. D., Roe, C. C., Schupp, J. M., Driebe, E. M., et al. (2014). Cryptococcus gattii in North American Pacific Northwest: Whole-population genome analysis provides insights into species evolution and dispersal. *mBio 5*. e01464-14. https://doi.org/10.1128/mBio.01464-14.

Enright, A. J., Van Dongen, S., & Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, *30*, 1575–1584.

European Centre for Disease Prevention and Control. (2013). *Risk assessment on the impact of environmental usage of triazoles on the development and spread of resistance to medical triazoles in Aspergillus species*. Stockholm: European Centre for Disease Prevention and Control.

Faria, N. R., Rambaut, A., Suchard, M. A., Baele, G., Bedford, T., Ward, M. J., et al. (2014). HIV epidemiology. The early spread and epidemic ignition of HIV-1 in human populations. *Science*, *346*, 56–61. https://doi.org/10.1126/science.1256739.

Farrer, R. A., Desjardins, C. A., Sakthikumar, S., Gujja, S., Saif, S., Zeng, Q., et al. (2015). Genome evolution and innovation across the four major lineages of Cryptococcus gattii. *mBio*, *6*(5), e00868-15. https://doi.org/10.1128/mBio.00868-15.

Farrer, R. A., Henk, D. A., Garner, T. W. J., Balloux, F., Woodhams, D. C., & Fisher, M. C. (2013). Chromosomal copy number variation, selection and uneven rates of recombination reveal cryptic genome diversity linked to pathogenicity. *PLoS Genetics 9*. e1003703. https://doi.org/10.1371/journal.pgen.1003703.

Farrer, R. A., Henk, D. A., MacLean, D., Studholme, D. J., & Fisher, M. C. (2013). Using false discovery rates to benchmark SNP-callers in next-generation sequencing projects. *Scientific Reports*, *3*, 1512. https://doi.org/10.1038/srep01512.

Farrer, R. A., Martel, A., Verbrugghe, E., Abouelleil, A., Ducatelle, R., Longcore, J. E., et al. (2017). Genomic innovations linked to infection strategies across emerging pathogenic chytrid fungi. *Nature Communications 8*. 14742. https://doi.org/10.1038/ncomms14742.

Farrer, R. A., Voelz, K., Henk, D. A., Johnston, S. A., Fisher, M. C., May, R. C., et al. (2016). Microevolutionary traits and comparative population genomics of the emerging pathogenic fungus Cryptococcus gattii. *Philosophical Transactions of the Royal Society of London Series B, Biological Sciences 371*. 20160021. https://doi.org/10.1098/rstb.2016.0021.

Farrer, R. A., Weinert, L. A., Bielby, J., Garner, T. W. J., Balloux, F., Clare, F., et al. (2011). Multiple emergences of genetically diverse amphibian-infecting chytrids include a globalized hypervirulent recombinant lineage. *Proceedings of the National Academy of Sciences of the United States of America*, *108*, 18732–18736. https://doi.org/10.1073/pnas.1111915108.

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., et al. (2014). Pfam: The protein families database. *Nucleic Acids Research*, *42*, D222–D230. https://doi.org/10.1093/nar/gkt1223.

Finn, R. D., Clements, J., & Eddy, S. R. (2011). HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Research*, *39*, W29–W37. https://doi.org/10.1093/nar/gkr367.

Fishel, B., Amstutz, H., Baum, M., Carbon, J., & Clarke, L. (1988). Structural organization and functional analysis of centromeric DNA in the fission yeast Schizosaccharomyces pombe. *Molecular and Cellular Biology*, *8*, 754–763.

Fisher, M. C., Gow, N. A. R., & Gurr, S. J. (2016). Tackling emerging fungal threats to animal health, food security and ecosystem resilience. *Philosophical Transactions of the Royal Society of London Series B, Biological Sciences 371*. 20160332. https://doi.org/10.1098/rstb.2016.0332.

Fisher, M. C., Henk, D. A., Briggs, C. J., Brownstein, J. S., Madoff, L. C., McCraw, S. L., et al. (2012). Emerging fungal threats to animal, plant and ecosystem health. *Nature*, *484*, 186–194. https://doi.org/10.1038/nature10947.

Fones, H. N., Fisher, M. C., & Gurr, S. J. (2017). Emerging fungal threats to plants and animals challenge agriculture and ecosystem resilience. *Microbiology Spectrum*, *5*(2). https://doi.org/10.1128/microbiolspec.FUNK-0027-2016.

Forche, A., Alby, K., Schaefer, D., Johnson, A. D., Berman, J., & Bennett, R. J. (2008). The parasexual cycle in Candida albicans provides an alternative pathway to meiosis for the formation of recombinant strains. *PLoS Biology 6*. e110. https://doi.org/10.1371/journal.pbio.0060110.

Forche, A., Magee, P. T., Selmecki, A., Berman, J., & May, G. (2009). Evolution in Candida albicans populations during a single passage through a mouse host. *Genetics*, *182*, 799–811. https://doi.org/10.1534/genetics.109.103325.

Fraczek, M. G., Bromley, M., Buied, A., Moore, C. B., Rajendran, R., Rautemaa, R., et al. (2013). The cdr1B efflux transporter is associated with non-cyp51a-mediated itraconazole resistance in Aspergillus fumigatus. *The Journal of Antimicrobial Chemotherapy*, *68*, 1486–1496. https://doi.org/10.1093/jac/dkt075.

Fraser, J. A., Giles, S. S., Wenink, E. C., Geunes-Boyer, S. G., Wright, J. R., Diezmann, S., et al. (2005). Same-sex mating and the origin of the Vancouver Island Cryptococcus gattii outbreak. *Nature*, *437*, 1360–1364. https://doi.org/10.1038/nature04220.

Friesen, T. L., Stukenbrock, E. H., Liu, Z., Meinhardt, S., Ling, H., Faris, J. D., et al. (2006). Emergence of a new disease as a result of interspecific virulence gene transfer. *Nature Genetics*, *38*, 953–956. https://doi.org/10.1038/ng1839.

Frith, M. C., Saunders, N. F. W., Kobe, B., & Bailey, T. L. (2008). Discovering sequence motifs with arbitrary insertions and deletions. *PLoS Computational Biology*, *4*(5), e1000071. https://doi.org/10.1371/journal.pcbi.1000071.

Galagan, J. E., Henn, M. R., Ma, L.-J., Cuomo, C. A., & Birren, B. (2005). Genomics of the fungal kingdom: Insights into eukaryotic biology. *Genome Research*, *15*, 1620–1631. https://doi.org/10.1101/gr.3767105.

Gardiner, D. M., McDonald, M. C., Covarelli, L., Solomon, P. S., Rusu, A. G., Marshall, M., et al. (2012). Comparative pathogenomics reveals horizontally acquired novel virulence genes in fungi infecting cereal hosts. *PLoS Pathogens 8*. e1002952. https://doi.org/10.1371/journal.ppat.1002952.

Garner, T. W., Perkins, M. W., Govindarajulu, P., Seglie, D., Walker, S., Cunningham, A. A., et al. (2006). The emerging amphibian pathogen Batrachochytrium dendrobatidis globally infects introduced populations of the North American bullfrog, Rana catesbeiana. *Biology Letters*, *2*, 455–459. https://doi.org/10.1098/rsbl.2006.0494.

Garrison, E., & Marth, G. (2012). *Haplotype-based variant detection from short-read sequencing*. ArXiv12073907 Q-Bio.

Gauthier, C., Weber, S., Alarco, A.-M., Alqawi, O., Daoud, R., Georges, E., et al. (2003). Functional similarities and differences between Candida albicans Cdr1p and Cdr2p transporters. *Antimicrobial Agents and Chemotherapy*, *47*, 1543–1554. https://doi.org/10.1128/AAC.47.5.1543-1554.2003.

Gelfand, Y., Rodriguez, A., & Benson, G. (2007). TRDB—The tandem repeats database. *Nucleic Acids Research*, *35*, D80–D87. https://doi.org/10.1093/nar/gkl1013.

Giraud, T., Gladieux, P., & Gavrilets, S. (2010). Linking the emergence of fungal plant diseases with ecological speciation. *Trends in Ecology & Evolution*, *25*, 387–395. https://doi.org/10.1016/j.tree.2010.03.006.

Gnerre, S., Maccallum, I., Przybylski, D., Ribeiro, F. J., Burton, J. N., Walker, B. J., et al. (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences of the United States of America*, *108*, 1513–1518. https://doi.org/10.1073/pnas.1017351108.

Goldberg, A. D., Allis, C. D., & Bernstein, E. (2007). Epigenetics: A landscape takes shape. *Cell*, *128*, 635–638. https://doi.org/10.1016/j.cell.2007.02.006.

Grice, C. M., Bertuzzi, M., & Bignell, E. M. (2013). Receptor-mediated signaling in Aspergillus fumigatus. *Frontiers in Microbiology*, *4*, 26. https://doi.org/10.3389/fmicb.2013.00026.

Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., & Eddy, S. R. (2003). Rfam: An RNA family database. *Nucleic Acids Research*, *31*, 439–441.

Haas, B. J., Kamoun, S., Zody, M. C., Jiang, R. H. Y., Handsaker, R. E., Cano, L. M., et al. (2009). Genome sequence and analysis of the Irish potato famine pathogen Phytophthora infestans. *Nature*, *461*, 393–398. https://doi.org/10.1038/nature08358.

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., et al. (2013). De novo transcript sequence reconstruction from RNA-Seq: Reference generation and analysis with Trinity. *Nature Protocols*, *8*, 1494–1512. https://doi.org/10.1038/nprot.2013.084.

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., et al. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the program to

assemble spliced alignments. *Genome Biology*, *9*, R7. https://doi.org/10.1186/gb-2008-9-1-r7.

Haas, B. J., Zeng, Q., Pearson, M. D., Cuomo, C. A., & Wortman, J. R. (2011). Approaches to fungal genome annotation. *Mycology*, *2*, 118–141. https://doi.org/10.1080/21501203.2011.606851.

Haft, D. H., Selengut, J. D., & White, O. (2003). The TIGRFAMs database of protein families. *Nucleic Acids Research*, *31*, 371–373.

Hagen, F., Ceresini, P. C., Polacheck, I., Ma, H., van Nieuwerburgh, F., Gabaldón, T., et al. (2013). Ancient dispersal of the human fungal pathogen Cryptococcus gattii from the Amazon rainforest. *PLoS One 8*. e71148. https://doi.org/10.1371/journal.pone.0071148.

Hagen, F., Khayhan, K., Theelen, B., Kolecka, A., Polacheck, I., Sionov, E., et al. (2015). Recognition of seven species in the Cryptococcus gattii/Cryptococcus neoformans species complex. *Fungal Genetics and Biology: FG & B*, *78*, 16–48. https://doi.org/10.1016/j.fgb.2015.02.009.

Hartmann, F. E., Sánchez-Vallet, A., McDonald, B. A., & Croll, D. (2017). A fungal wheat pathogen evolved host specialization by extensive chromosomal rearrangements. *The ISME Journal*, *11*, 1189–1204. https://doi.org/10.1038/ismej.2016.196.

Hasman, H., Saputra, D., Sicheritz-Ponten, T., Lund, O., Svendsen, C. A., Frimodt-Møller, N., et al. (2014). Rapid whole-genome sequencing for detection and characterization of microorganisms directly from clinical samples. *Journal of Clinical Microbiology*, *52*, 139–146. https://doi.org/10.1128/JCM.02452-13.

Healey, K. R., Zhao, Y., Perez, W. B., Lockhart, S. R., Sobel, J. D., Farmakiotis, D., et al. (2016). Prevalent mutator genotype identified in fungal pathogen Candida glabrata promotes multi-drug resistance. *Nature Communications* 7. 11128. https://doi.org/10.1038/ncomms11128.

Hibbett, D. S., Binder, M., Bischoff, J. F., Blackwell, M., Cannon, P. F., Eriksson, O. E., et al. (2007). A higher-level phylogenetic classification of the Fungi. *Mycological Research*, *111*, 509–547. https://doi.org/10.1016/j.mycres.2007.03.004.

Hittalmani, S., Mahesh, H. B., Mahadevaiah, C., & Prasannakumar, M. K. (2016). De novo genome assembly and annotation of rice sheath rot fungus Sarocladium oryzae reveals genes involved in Helvolic acid and Cerulenin biosynthesis pathways. *BMC Genomics*, *17*, 271. https://doi.org/10.1186/s12864-016-2599-0.

Hogenhout, S. A., Van der Hoorn, R. A. L., Terauchi, R., & Kamoun, S. (2009). Emerging concepts in effector biology of plant-associated organisms. *Molecular Plant-Microbe Interactions: MPMI*, *22*, 115–122. https://doi.org/10.1094/MPMI-22-2-0115.

Holliday, R. (1990). DNA methylation and epigenetic inheritance. *Philosophical Transactions of the Royal Society of London Series B, Biological Sciences*, *326*, 329–338. https://doi.org/10.1098/rstb.1990.0015.

Hsueh, P.-R., Teng, L.-J., Hung, C.-C., Hsu, J.-H., Yang, P.-C., Ho, S.-W., et al. (2000). Molecular evidence for strain dissemination of Penicillium marneffei: An emerging pathogen in Taiwan. *The Journal of Infectious Diseases*, *181*, 1706–1712. https://doi.org/10.1086/315432.

Hu, G., Wang, J., Choi, J., Jung, W. H., Liu, I., Litvintseva, A. P., et al. (2011). Variation in chromosome copy number influences the virulence of Cryptococcus neoformans and occurs in isolates from AIDS patients. *BMC Genomics*, *12*, 526. https://doi.org/10.1186/1471-2164-12-526.

Huff, J. T., & Zilberman, D. (2014). Dnmt1-independent CG methylation contributes to nucleosome positioning in diverse eukaryotes. *Cell*, *156*, 1286–1297. https://doi.org/10.1016/j.cell.2014.01.029.

Hunt, M., Kikuchi, T., Sanders, M., Newbold, C., Berriman, M., & Otto, T. D. (2013). REAPR: A universal tool for genome assembly evaluation. *Genome Biology*, *14*, R47. https://doi.org/10.1186/gb-2013-14-5-r47.

Inderbitzin, P., Davis, R. M., Bostock, R. M., & Subbarao, K. V. (2011). The ascomycete Verticillium longisporum is a hybrid and a plant pathogen with an expanded host range. *PLoS One 6*. e18260. https://doi.org/10.1371/journal.pone.0018260.

Islam, M. T., Croll, D., Gladieux, P., Soanes, D. M., Persoons, A., Bhattacharjee, P., et al. (2016). Emergence of wheat blast in Bangladesh was caused by a South American lineage of Magnaporthe oryzae. *BMC Biology*, *14*, 84. https://doi.org/10.1186/s12915-016-0309-7.

James, T. Y., Litvintseva, A. P., Vilgalys, R., Morgan, J. A. T., Taylor, J. W., Fisher, M. C., et al. (2009). Rapid global expansion of the fungal disease chytridiomycosis into declining and healthy amphibian populations. *PLoS Pathogens 5*. e1000458. https://doi.org/10.1371/journal.ppat.1000458.

Janbon, G., Ormerod, K. L., Paulet, D., Iii, E. J. B., Yadav, V., Chatterjee, G., et al. (2014). Analysis of the genome and transcriptome of Cryptococcus neoformans var. grubii reveals complex RNA expression and microevolution leading to virulence attenuation. *PLoS Genetics 10*. e1004261. https://doi.org/10.1371/journal.pgen.1004261.

Jeon, J., Choi, J., Lee, G.-W., Park, S.-Y., Huh, A., Dean, R. A., et al. (2015). Genome-wide profiling of DNA methylation provides insights into epigenetic regulation of fungal development in a plant pathogenic fungus, Magnaporthe oryzae. *Scientific Reports*, *5*, 8567. https://doi.org/10.1038/srep08567.

Jeon, J., Kwon, S., & Lee, Y.-H. (2014). Histone acetylation in fungal pathogens of plants. *Plant Pathology Journal*, *30*, 1–9. https://doi.org/10.5423/PPJ.RW.01.2014.0003.

Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., et al. (2014). Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Research*, *24*, 1384–1395. https://doi.org/10.1101/gr.170720.113.

Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, *28*, 27–30.

Karvonen, A., Rellstab, C., Louhi, K.-R., & Jokela, J. (2012). Synchronous attack is advantageous: Mixed genotype infections lead to higher infection success in trematode parasites. *Proceedings of the Biological Sciences*, *279*, 171–176. https://doi.org/10.1098/rspb.2011.0879.

Kasuga, T., White, T. J., & Taylor, J. W. (2002). Estimation of nucleotide substitution rates in eurotiomycete fungi. *Molecular Biology and Evolution*, *19*(12), 2318–2324.

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, *30*, 772–780. https://doi.org/10.1093/molbev/mst010.

Kent, W. J. (2002). BLAT—The BLAST-like alignment tool. *Genome Research*, *12*, 656–664. https://doi.org/10.1101/gr.229202. Article published online before March 2002.

Khatri, M., & Rajam, M. V. (2007). Targeting polyamines of Aspergillus nidulans by siRNA specific to fungal ornithine decarboxylase gene. *Medical Mycology*, *45*, 211–220. https://doi.org/10.1080/13693780601158779.

Kim, K.-H., Willger, S. D., Park, S.-W., Puttikamonkul, S., Grahl, N., Cho, Y., et al. (2009). TmpL, a transmembrane protein required for intracellular redox homeostasis and virulence in a plant and an animal fungal pathogen. *PLoS Pathogens 5*. e1000653. https://doi.org/10.1371/journal.ppat.1000653.

Kõljalg, U., Nilsson, R. H., Abarenkov, K., Tedersoo, L., Taylor, A. F. S., Bahram, M., et al. (2013). Towards a unified paradigm for sequence-based identification of fungi. *Molecular Ecology*, *22*, 5271–5277. https://doi.org/10.1111/mec.12481.

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: Scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*, *27*(5), 722–736. https://doi.org/10.1101/gr.215087.116.

Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinformatics*, *5*, 59. https://doi.org/10.1186/1471-2105-5-59.

Krogh, A., Larsson, B., von Heijne, G., & Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *Journal of Molecular Biology*, *305*, 567–580. https://doi.org/10.1006/jmbi.2000.4315.

Kryazhimskiy, S., & Plotkin, J. B. (2008). The population genetics of dN/dS. *PLoS Genetics*, *4*(12), e1000304. https://doi.org/10.1371/journal.pgen.1000304.

Kuo, D., Tan, K., Zinman, G., Ravasi, T., Bar-Joseph, Z., & Ideker, T. (2010). Evolutionary divergence in the fungal response to fluconazole revealed by soft clustering. *Genome Biology*, *11*, R77. https://doi.org/10.1186/gb-2010-11-7-r77.

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., et al. (2004). Versatile and open software for comparing large genomes. *Genome Biology*, *5*, R12. https://doi.org/10.1186/gb-2004-5-2-r12.

Kwon-Chung, K. J., Bennett, J. E., Wickes, B. L., Meyer, W., Cuomo, C. A., Wollenburg, K. R., et al. (2017). The case for adopting the "Species Complex" nomenclature for the etiologic agents of cryptococcosis. *mSphere*, *2*(1), pii: e00357-16. https://doi.org/10.1128/mSphere.00357-16.

Kwon-Chung, K. J., & Chang, Y. C. (2012). Aneuploidy and drug resistance in pathogenic fungi. *PLoS Pathogens*, *8*(11), e1003022. https://doi.org/10.1371/journal.ppat.1003022.

Lagesen, K., Hallin, P., Rødland, E. A., Staerfeldt, H.-H., Rognes, T., & Ussery, D. W. (2007). RNAmmer: Consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research*, *35*, 3100–3108. https://doi.org/10.1093/nar/gkm160.

Lam, J. K. W., Chow, M. Y. T., Zhang, Y., & Leung, S. W. S. (2015). siRNA versus miRNA as therapeutics for gene silencing. *Molecular Therapy Nucleic Acids, 4*, e252. https://doi.org/10.1038/mtna.2015.23.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*, 357–359. https://doi.org/10.1038/nmeth.1923.

Lassmann, T., & Sonnhammer, E. L. L. (2005). Automatic assessment of alignment quality. *Nucleic Acids Research*, *33*, 7120–7128. https://doi.org/10.1093/nar/gki1020.

Law, J. A., & Jacobsen, S. E. (2010). Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nature Reviews Genetics*, *11*, 204–220. https://doi.org/10.1038/nrg2719.

Leach, M. D., Farrer, R. A., Tan, K., Miao, Z., Walker, L. A., Cuomo, C. A., et al. (2016). Hsf1 and Hsp90 orchestrate temperature-dependent global transcriptional remodelling and chromatin architecture in Candida albicans. *Nature Communications7*, 11704. https://doi.org/10.1038/ncomms11704.

Lecuit, M., & Eloit, M. (2014). The diagnosis of infectious diseases by whole genome next generation sequencing: A new era is opening. *Frontiers in Cellular and Infection Microbiology*, *4*, 25. https://doi.org/10.3389/fcimb.2014.00025.

Lengeler, K. B., Cox, G. M., & Heitman, J. (2001). Serotype AD strains of Cryptococcus neoformans are diploid or aneuploid and are heterozygous at the mating-type locus. *Infection and Immunity*, *69*, 115–122. https://doi.org/10.1128/IAI.69.1.115-122.2001.

Leopardi, S., Blake, D., & Puechmaille, S. J. (2015). White-nose syndrome fungus introduced from Europe to North America. *Current Biology: CB*, *25*, R217–219. https://doi.org/10.1016/j.cub.2015.01.047.

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, *26*, 589–595. https://doi.org/10.1093/bioinformatics/btp698.

Li, W., & Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics (Oxford, England)*, *22*, 1658–1659. https://doi.org/10.1093/bioinformatics/btl158.

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., et al. (2010). De novo assembly of human genomes with massively parallel short read sequencing. *Genome Research*, *20*, 265–272. https://doi.org/10.1101/gr.097261.109.

Lister, R., Pelizzola, M., Dowen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J., et al. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, *462*, 315–322. https://doi.org/10.1038/nature08514.

Liti, G., Carter, D. M., Moses, A. M., Warringer, J., Parts, L., James, S. A., et al. (2009). Population genomics of domestic and wild yeasts. *Nature*, *458*, 337–341. https://doi.org/10.1038/nature07743.

Litvintseva, A. P., Carbone, I., Rossouw, J., Thakur, R., Govender, N. P., & Mitchell, T. G. (2011). Evidence that the human pathogenic fungus Cryptococcus neoformans var. grubii may have evolved in Africa. *PLoS One 6*. e19688. https://doi.org/10.1371/journal.pone.0019688.

Liu, O. W., Chun, C. D., Chow, E. D., Chen, C., Madhani, H. D., & Noble, S. M. (2008). Systematic genetic analysis of virulence in the human fungal pathogen Cryptococcus neoformans. *Cell*, *135*, 174–188. https://doi.org/10.1016/j.cell.2008.07.046.

Liu, P., & Stajich, J. E. (2015). Characterization of the carbohydrate binding module 18 gene family in the amphibian pathogen Batrachochytrium dendrobatidis. *Fungal Genetics and Biology: FG & B*, 77, 31–39. https://doi.org/10.1016/j.fgb.2015.03.003.

Liu, T., Ye, W., Ru, Y., Yang, X., Gu, B., Tao, K., et al. (2011). Two host cytoplasmic effectors are required for pathogenesis of Phytophthora sojae by suppression of host defenses. *Plant Physiology*, *155*, 490–501. https://doi.org/10.1104/pp.110.166470.

Longo, A. V., Burrowes, P. A., & Zamudio, K. R. (2014). Genomic studies of disease-outcome in host–pathogen dynamics. *Integrative and Comparative Biology*, *54*, 427–438. https://doi.org/10.1093/icb/icu073.

Lopes da Rosa, J., Boyartchuk, V. L., Zhu, L. J., & Kaufman, P. D. (2010). Histone acetyltransferase Rtt109 is required for Candida albicans pathogenesis. *Proceedings of the National Academy of Sciences of the United States of America*, *107*, 1594–1599. https://doi.org/10.1073/pnas.0912427107.

Lou, S., Lee, H.-M., Qin, H., Li, J.-W., Gao, Z., Liu, X., et al. (2014). Whole-genome bisulfite sequencing of multiple individuals reveals complementary roles of promoter and gene body methylation in transcriptional regulation. *Genome Biology*, *15*, 408. https://doi.org/10.1186/s13059-014-0408-0.

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*, 550. https://doi.org/10.1186/s13059-014-0550-8.

Love, R. R., Weisenfeld, N. I., Jaffe, D. B., Besansky, N. J., & Neafsey, D. E. (2016). Evaluation of DISCOVAR de novo using a mosquito sample for cost-effective short-read genome assembly. *BMC Genomics*, *17*, 187. https://doi.org/10.1186/s12864-016-2531-7.

Lowe, T. M., & Eddy, S. R. (1997). tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research*, *25*, 955–964.

Lukashin, A. V., & Borodovsky, M. (1998). GeneMark.hmm: New solutions for gene finding. *Nucleic Acids Research*, *26*, 1107–1115.

Lyngsø, R. B., & Pedersen, C. N. (2000). RNA pseudoknot prediction in energy-based models. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, *7*, 409–427. https://doi.org/10.1089/106652700750050862.

Magditch, D. A., Liu, T.-B., Xue, C., & Idnurm, A. (2012). DNA mutations mediate microevolution between host-adapted forms of the pathogenic fungus Cryptococcus neoformans. *PLoS Pathogens 8*. e1002936. https://doi.org/10.1371/journal.ppat.1002936.

Majoros, W. H., Pertea, M., & Salzberg, S. L. (2004). TigrScan and GlimmerHMM: Two open source ab initio eukaryotic gene-finders. *Bioinformatics (Oxford, England)*, *20*, 2878–2879. https://doi.org/10.1093/bioinformatics/bth315.

Martel, A., Blooi, M., Adriaensen, C., Van Rooij, P., Beukema, W., Fisher, M. C., et al. (2014). Wildlife disease. Recent introduction of a chytrid fungus endangers Western Palearctic salamanders. *Science*, *346*, 630–631. https://doi.org/10.1126/science.1258268.

Martel, A., Spitzen-van der Sluijs, A., Blooi, M., Bert, W., Ducatelle, R., Fisher, M. C., et al. (2013). Batrachochytrium salamandrivorans sp. nov. causes lethal chytridiomycosis in amphibians. *Proceedings of the National Academy of Sciences of the United States of America*, *110*, 15325–15329. https://doi.org/10.1073/pnas.1307356110.

McDonald, J. H., & Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. *Nature*, *351*, 652–654. https://doi.org/10.1038/351652a0.

McDonald, B. A., & Stukenbrock, E. H. (2016). Rapid emergence of pathogens in agro-ecosystems: Global threats to agricultural sustainability and food security. *Philosophical Transactions of the Royal Society of London Series B, Biological Sciences*, *371*(1709), pii: 20160026. https://doi.org/10.1098/rstb.2016.0026.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*, 1297–1303. https://doi.org/10.1101/gr.107524.110.

Meis, J. F., Chowdhary, A., Rhodes, J. L., Fisher, M. C., & Verweij, P. E. (2016). Clinical implications of globally emerging azole resistance in Aspergillus fumigatus. *Philosophical Transactions of the Royal Society of London Series B, Biological Sciences*, *371*(1709), pii: 20150460. https://doi.org/10.1098/rstb.2015.0460.

Menardo, F., Praz, C. R., Wyder, S., Ben-David, R., Bourras, S., Matsumae, H., et al. (2016). Hybridization of powdery mildew strains gives rise to pathogens on novel agricultural crop species. *Nature Genetics*, *48*, 201–205. https://doi.org/10.1038/ng.3485.

Meneau, I., Coste, A. T., & Sanglard, D. (2016). Identification of Aspergillus fumigatus multidrug transporter genes and their potential involvement in antifungal resistance. *Medical Mycology*, *54*, 616–627. https://doi.org/10.1093/mmy/myw005.

Mohammadi, R., Badiee, P., Badali, H., Abastabar, M., Safa, A. H., Hadipour, M., et al. (2015). Use of restriction fragment length polymorphism to identify Candida species, related to onychomycosis. *Advanced Biomedical Research*, *4*, 95. https://doi.org/10.4103/2277-9175.156659.

Morse, S. S. (1995). Factors in the emergence of infectious diseases. *Emerging Infectious Diseases*, *1*, 7–15. https://doi.org/10.3201/eid0101.950102.

Muñoz, J. F., Farrer, R. A., Desjardins, C. A., Gallo, J. E., Sykes, S., Sakthikumar, S., et al. (2016). Genome diversity, recombination, and virulence across the major lineages of paracoccidioides. *mSphere 1*. e00213-16. https://doi.org/10.1128/mSphere.00213-16.

Nakaune, R., Hamamoto, H., Imada, J., Akutsu, K., & Hibi, T. (2002). A novel ABC transporter gene, PMR5, is involved in multidrug resistance in the phytopathogenic fungus Penicillium digitatum. *Molecular Genetics and Genomics*, *267*, 179–185. https://doi.org/10.1007/s00438-002-0649-6.

Nawrocki, E. P., & Eddy, S. R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, *29*, 2933–2935. https://doi.org/10.1093/bioinformatics/btt509.

Nickel, W., & Seedorf, M. (2008). Unconventional mechanisms of protein transport to the cell surface of eukaryotic cells. *Annual Review of Cell and Developmental Biology*, *24*, 287–308. https://doi.org/10.1146/annurev.cellbio.24.110707.175320.

Nucci, M., & Marr, K. A. (2005). Emerging fungal diseases. *Clinical Infectious Diseases*, *41*, 521–526. https://doi.org/10.1086/432060.

O'Meara, T. R., Hay, C., Price, M. S., Giles, S., & Alspaugh, J. A. (2010). Cryptococcus neoformans histone acetyltransferase Gcn5 regulates fungal adaptation to the host. *Eukaryotic Cell*, *9*, 1193–1202. https://doi.org/10.1128/EC.00098-10.

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., et al. (2016). Mash: Fast genome and metagenome distance estimation using MinHash. *Genome Biology*, *17*, 132.

Osawa, S., Jukes, T. H., Watanabe, K., & Muto, A. (1992). Recent evidence for evolution of the genetic code. *Microbiological Reviews*, *56*, 229–264.

Palmer, J. M., Kubatova, A., Novakova, A., Minnis, A. M., Kolarik, M., & Lindner, D. L. (2014). Molecular characterization of a heterothallic mating system in Pseudogymnoascus destructans, the Fungus causing white-nose syndrome of bats. *G3 (Bethesda, Md)*, *4*, 1755–1763. https://doi.org/10.1534/g3.114.012641.

Panepinto, J. C., & Williamson, P. R. (2006). Intersection of fungal fitness and virulence in Cryptococcus neoformans. *FEMS Yeast Research*, *6*, 489–498. https://doi.org/10.1111/j.1567-1364.2006.00078.x.

Parra, G., Bradnam, K., & Korf, I. (2007). CEGMA: A pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics (Oxford, England)*, *23*, 1061–1067. https://doi.org/10.1093/bioinformatics/btm071.

Pasquinelli, A. E. (2012). MicroRNAs and their targets: Recognition, regulation and an emerging reciprocal relationship. *Nature Reviews Genetics*, *13*, 271–282. https://doi.org/10.1038/nrg3162.

Pawlowska, T. E., & Taylor, J. W. (2004). Organization of genetic variation in individuals of arbuscular mycorrhizal fungi. *Nature*, *427*, 733–737. https://doi.org/10.1038/nature02290.

Pertea, M., Kim, D., Pertea, G. M., Leek, J. T., & Salzberg, S. L. (2016). Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nature Protocols*, *11*, 1650–1667. https://doi.org/10.1038/nprot.2016.095.

Petersen, T. N., Brunak, S., von Heijne, G., & Nielsen, H. (2011). SignalP 4.0: Discriminating signal peptides from transmembrane regions. *Nature Methods*, *8*, 785–786. https://doi.org/10.1038/nmeth.1701.

Plissonneau, C., Benevenuto, J., Mohd-Assaad, N., Fouché, S., Hartmann, F. E., & Croll, D. (2017). Using population and comparative genomics to understand the genetic basis of effector-driven fungal pathogen evolution. *Frontiers in Plant Science*, *8*, 119. https://doi.org/10.3389/fpls.2017.00119.

Quiroz Velasquez, P. F., Abiff, S. K., Fins, K. C., Conway, Q. B., Salazar, N. C., Delgado, A. P., et al. (2014). Transcriptome analysis of the entomopathogenic Oomycete Lagenidium giganteum reveals putative virulence factors. *Applied and Environmental Microbiology*, *80*, 6427–6436. https://doi.org/10.1128/AEM.02060-14.

Raffaele, S., Farrer, R. A., Cano, L. M., Studholme, D. J., MacLean, D., Thines, M., et al. (2010). Genome evolution following host jumps in the Irish potato famine pathogen lineage. *Science*, *330*, 1540–1543. https://doi.org/10.1126/science.1193070.

Ratajczak, M. Z., & Ratajczak, J. (2016). Horizontal transfer of RNA and proteins between cells by extracellular microvesicles: 14 years later. *Clinical and Translational Medicine*, *5*, 7. https://doi.org/10.1186/s40169-016-0087-4.

Rawlings, N. D., Barrett, A. J., & Finn, R. (2016). Twenty years of the MEROPS database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Research*, *44*, D343–350. https://doi.org/10.1093/nar/gkv1118.

Reedy, J. L., Floyd, A. M., & Heitman, J. (2009). Mechanistic plasticity of sexual reproduction and meiosis in the Candida pathogenic species complex. *Current Biology: CB*, *19*, 891–899. https://doi.org/10.1016/j.cub.2009.04.058.

Ren, J., Rastegari, B., Condon, A., & Hoos, H. H. (2005). HotKnots: Heuristic prediction of RNA secondary structures including pseudoknots. *RNA*, *11*, 1494–1504. https://doi.org/10.1261/rna.7284905.

Rhodes, J., Abdolrasouli, A., Farrer, R. A., Cuomo, C. A., Aanensen, D. M., Armstrong-James, D., et al. (2017). Rapid genome sequencing for outbreak analysis of the emerging human fungal pathogen. *Candida auris bioRxiv*. https://doi.org/10.1101/201343.

Rhodes, J., Beale, M. A., Vanhove, M., Jarvis, J. N., Kannambath, S., Simpson, J. A., et al. (2017). A population genomics approach to assessing the genetic basis of within-host microevolution underlying recurrent Cryptococcal meningitis infection. *G3 (Bethesda, Md)*, 7(4), 1165–1176. https://doi.org/10.1534/g3.116.037499.

Rhodes, J., Desjardins, C. A., Sykes, S. M., Beale, M. A., Vanhove, M., Sakthikumar, S., et al. (2017). Tracing genetic exchange and biogeography of Cryptococcus neoformans var. grubii at the global population level. *Genetics*, *207*, 327–346. https://doi.org/10.1534/genetics.117.203836.

Rieux, A., & Balloux, F. (2016). Inferences from tip-calibrated phylogenies: A review and a practical guide. *Molecular Ecology*, *25*, 1911–1924. https://doi.org/10.1111/mec.13586.

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*, 139–140. https://doi.org/10.1093/bioinformatics/btp616.

Robinson, M. D., & Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, *11*, R25. https://doi.org/10.1186/gb-2010-11-3-r25.

Rosenblum, E. B., James, T. Y., Zamudio, K. R., Poorten, T. J., Ilut, D., Rodriguez, D., et al. (2013). Complex history of the amphibian-killing chytrid fungus revealed with genome resequencing data. *Proceedings of the National Academy of Sciences of the United States of America*, *110*, 9385–9390. https://doi.org/10.1073/pnas.1300130110.

Rosenblum, E. B., Stajich, J. E., Maddox, N., & Eisen, M. B. (2008). Global gene expression profiles for life stages of the deadly amphibian pathogen Batrachochytrium dendrobatidis. *Proceedings of the National Academy of Sciences of the United States of America*, *105*, 17034–17039. https://doi.org/10.1073/pnas.0804173105.

Salamov, A. A., & Solovyev, V. V. (2000). Ab initio gene finding in Drosophila genomic DNA. *Genome Research*, *10*, 516–522.

Salzberg, S. L., Phillippy, A. M., Zimin, A., Puiu, D., Magoc, T., Koren, S., et al. (2012). GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Research*, *22*, 557–567. https://doi.org/10.1101/gr.131383.111.

Sanguinetti, M., Posteraro, B., La Sorda, M., Torelli, R., Fiori, B., Santangelo, R., et al. (2006). Role of AFR1, an ABC transporter-encoding gene, in the in vivo response to fluconazole and virulence of Cryptococcus neoformans. *Infection and Immunity*, *74*, 1352–1359. https://doi.org/10.1128/IAI.74.2.1352-1359.2006.

Santos, M. A., Keith, G., & Tuite, M. F. (1993). Non-standard translational events in Candida albicans mediated by an unusual seryl-tRNA with a 5'-CAG-3' (leucine) anticodon. *The EMBO Journal*, *12*, 607–616.

Sanyal, K., Baum, M., & Carbon, J. (2004). Centromeric DNA sequences in the pathogenic yeast Candida albicans are all different and unique. *Proceedings of the National Academy of Sciences of the United States of America*, *101*, 11374–11379. https://doi.org/10.1073/pnas.0404318101.

Schoch, C. L., Seifert, K. A., Huhndorf, S., Robert, V., Spouge, J. L., Levesque, C. A., et al. (2012). Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences of the United States of America*, *109*, 6241–6246. https://doi.org/10.1073/pnas.1117018109.

Schornack, S., van Damme, M., Bozkurt, T. O., Cano, L. M., Smoker, M., Thines, M., et al. (2010). Ancient class of translocated oomycete effectors targets the host nucleus. *Proceedings of the National Academy of Sciences of the United States of America*, *107*, 17421–17426. https://doi.org/10.1073/pnas.1008491107.

Schoustra, S. E., Debets, A. J. M., Slakhorst, M., & Hoekstra, R. F. (2007). Mitotic recombination accelerates adaptation in the fungus Aspergillus nidulans. *PLoS Genetics*, *3*(4), e68. https://doi.org/10.1371/journal.pgen.0030068.

Schurch, N. J., Schofield, P., Gierliński, M., Cole, C., Sherstnev, A., Singh, V., et al. (2016). How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA*, *22*, 839–851. https://doi.org/10.1261/rna.053959.115.

Selker, E. U., Tountas, N. A., Cross, S. H., Margolin, B. S., Murphy, J. G., Bird, A. P., et al. (2003). The methylated component of the Neurospora crassa genome. *Nature*, *422*, 893–897. https://doi.org/10.1038/nature01564.

Sharma, C., Kumar, N., Meis, J. F., Pandey, R., & Chowdhary, A. (2015). Draft genome sequence of a fluconazole-resistant candida auris strain from a Candidemia patient in India. *Genome Announcements*, *3*(4), pii: e00722-15. https://doi.org/10.1128/genomeA.00722-15.

Sharpton, T. J., Neafsey, D. E., Galagan, J. E., & Taylor, J. W. (2008). Mechanisms of intron gain and loss in Cryptococcus. *Genome Biology*, *9*(1), R24.

Sheltzer, J. M., Blank, H. M., Pfau, S. J., Tange, Y., George, B. M., Humpton, T. J., et al. (2011). Aneuploidy drives genomic instability in yeast. *Science*, *333*, 1026–1030. https://doi.org/10.1126/science.1206412.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics (Oxford, England)*, *31*, 3210–3212. https://doi.org/10.1093/bioinformatics/btv351.

Simpson, J. T., & Durbin, R. (2012). Efficient de novo assembly of large genomes using compressed data structures. *Genome Research*, *22*, 549–556. https://doi.org/10.1101/gr.126953.111.

Singh, R. P., Hodson, D. P., Huerta-Espino, J., Jin, Y., Bhavani, S., Njau, P., et al. (2011). The emergence of Ug99 races of the stem rust fungus is a threat to World wheat production. *Annual Review of Phytopathology*, *49*, 465–481. https://doi.org/10.1146/annurev-phyto-072910-095423.

Sionov, E., Lee, H., Chang, Y. C., & Kwon-Chung, K. J. (2010). Cryptococcus neoformans overcomes stress of azole drugs by formation of disomy in specific multiple chromosomes. *PLoS Pathogens 6*. e1000848. https://doi.org/10.1371/journal.ppat.1000848.

Smith, K. M., Galazka, J. M., Phatale, P. A., Connolly, L. R., & Freitag, M. (2012). Centromeres of filamentous fungi. *Chromosome Research*, *20*, 635–656. https://doi.org/10.1007/s10577-012-9290-3.

Song, G., Dickins, B. J. A., Demeter, J., Engel, S., Dunn, B., & Cherry, J. M. (2015). AGAPE (Automated Genome Analysis PipelinE) for pan-genome analysis of Saccharomyces cerevisiae. *PLoS One 10*. e0120671. https://doi.org/10.1371/journal.pone.0120671.

Spanu, P. D., Abbott, J. C., Amselem, J., Burgis, T. A., Soanes, D. M., Stüber, K., et al. (2010). Genome expansion and gene loss in powdery mildew fungi reveal tradeoffs in extreme parasitism. *Science*, *330*, 1543–1546. https://doi.org/10.1126/science.1194573.

Spatafora, J. W., Chang, Y., Benny, G. L., Lazarus, K., Smith, M. E., Berbee, M. L., et al. (2016). A phylum-level phylogenetic classification of zygomycete fungi based on genome-scale data. *Mycologia*, *108*, 1028–1046. https://doi.org/10.3852/16-042.

Stajich, J. E., Harris, T., Brunk, B. P., Brestelli, J., Fischer, S., Harb, O. S., et al. (2012). FungiDB: An integrated functional genomics database for fungi. *Nucleic Acids Research*, *40*(D1), D675–D681.

Stam, R., Jupe, J., Howden, A. J. M., Morris, J. A., Boevink, P. C., Hedley, P. E., et al. (2013). Identification and characterisation CRN Effectors in Phytophthora capsici Shows modularity and functional diversity. *PLoS One 8*. e59517. https://doi.org/10.1371/journal.pone.0059517.

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., & Morgenstern, B. (2006). AUGUSTUS: Ab initio prediction of alternative transcripts. *Nucleic Acids Research*, *34*, W435–W439. https://doi.org/10.1093/nar/gkl200.

Stegen, G., Pasmans, F., Schmidt, B. R., Rouffaer, L. O., Van Praet, S., Schaub, M., et al. (2017). Drivers of salamander extirpation mediated by Batrachochytrium salamandrivorans. *Nature*, *544*, 353–356. https://doi.org/10.1038/nature22059.

Stoletzki, N., & Eyre-Walker, A. (2011). Estimation of the neutrality index. *Molecular Biology and Evolution*, *28*, 63–70. https://doi.org/10.1093/molbev/msq249.

Storey, J. D., & Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America*, *100*, 9440–9445. https://doi.org/10.1073/pnas.1530509100.

Stornaiuolo, M., Lotti, L. V., Borgese, N., Torrisi, M.-R., Mottola, G., Martire, G., et al. (2003). KDEL and KKXX retrieval signals appended to the same reporter protein determine different trafficking between endoplasmic reticulum, intermediate compartment, and Golgi complex. *Molecular Biology of the Cell*, *14*, 889–902. https://doi.org/10.1091/mbc.E02-08-0468.

Stroud, H., Otero, S., Desvoyes, B., Ramírez-Parra, E., Jacobsen, S. E., & Gutierrez, C. (2012). Genome-wide analysis of histone H3.1 and H3.3 variants in Arabidopsis thaliana. *Proceedings of the National Academy of Sciences of the United States of America*, *109*, 5370–5375. https://doi.org/10.1073/pnas.1203145109.

Studholme, D. J. (2016). Genome update. Let the consumer beware: Streptomyces genome sequence quality. *Microbial Biotechnology*, *9*, 3–7. https://doi.org/10.1111/1751-7915.12344.

Stukenbrock, E. H., & McDonald, B. A. (2008). The origins of plant pathogens in agro-ecosystems. *Annual Review of Phytopathology*, *46*, 75–100. https://doi.org/10.1146/annurev.phyto.010708.154114.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, *123*, 585–595.

Taylor, J. W., Jacobson, D. J., & Fisher, M. C. (1999). THE EVOLUTION OF ASEXUAL FUNGI: Reproduction, speciation and classification. *Annual Review of Phytopathology*, *37*, 197–246. https://doi.org/10.1146/annurev.phyto.37.1.197.

Taylor, J. W., Jacobson, D. J., Kroken, S., Kasuga, T., Geiser, D. M., Hibbett, D. S., et al. (2000). Phylogenetic species recognition and species concepts in fungi. *Fungal Genetics and Biology: FG & B*, *31*, 21–32. https://doi.org/10.1006/fgbi.2000.1228.

Thiel, T., Michalek, W., Varshney, R. K., & Graner, A. (2003). Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.). *TAG Theoretical and Applied Genetics Theoretische und Angewandte Genetik*, *106*, 411–422. https://doi.org/10.1007/s00122-002-1031-0.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., et al. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols*, 7, 562–578. https://doi.org/10.1038/nprot.2012.016.

Trivedi, J., Lachapelle, J., Vanderwolf, K., Misra, V., Willis, C. K. R., Ratcliffe, J., et al. (2017). Fungus causing white-nose syndrome in bats accumulates genetic variability in North America with no sign of recombination. *mSphere*, *2*(4), e00271–17. https://doi.org/10.1128/mSphereDirect.00271-17.

Turner, E., Jacobson, D. J., & Taylor, J. W. (2011). Genetic architecture of a reinforced, postmating, reproductive isolation barrier between Neurospora species indicates evolution via natural selection. *PLoS Genetics* 7. e1002204. https://doi.org/10.1371/journal.pgen.1002204.

Turro, E., Astle, W. J., & Tavaré, S. (2014). Flexible analysis of RNA-seq data using mixed effects models. *Bioinformatics (Oxford, England)*, *30*, 180–188. https://doi.org/10.1093/bioinformatics/btt624.

Tuteja, R. (2005). Type I signal peptidase: An overview. *Archives of Biochemistry and Biophysics*, *441*, 107–111. https://doi.org/10.1016/j.abb.2005.07.013.

Van Der Linden, J. W. M., Warris, A., & Verweij, P. E. (2011). Aspergillus species intrinsically resistant to antifungal agents. *Medical Mycology*, *49*(Suppl. 1), S82–89. https://doi.org/10.3109/13693786.2010.499916.

Van Rooij, P., Martel, A., D'Herde, K., Brutyn, M., Croubels, S., Ducatelle, R., et al. (2012). Germ tube mediated invasion of Batrachochytrium dendrobatidis in amphibian skin is host dependent. *PLoS One 7*. e41481. https://doi.org/10.1371/journal.pone.0041481.

Verweij, P. E., Chowdhary, A., Melchers, W. J. G., & Meis, J. F. (2016). Azole resistance in Aspergillus fumigatus: Can we retain the clinical use of mold-active antifungal azoles? *Clinical Infectious Diseases*, *62*, 362–368. https://doi.org/10.1093/cid/civ885.

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., et al. (2014). Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One 9*. e112963. https://doi.org/10.1371/journal.pone.0112963.

Walker, S. F., Bosch, J., James, T. Y., Litvintseva, A. P., Oliver Valls, J. A., Piña, S., et al. (2008). Invasive pathogens threaten species recovery programs. *Current Biology: CB*, *18*, R853–854. https://doi.org/10.1016/j.cub.2008.07.033.

Wang, L., Feng, Z., Wang, X., Wang, X., & Zhang, X. (2010). DEGseq: An R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics (Oxford, England)*, *26*, 136–138. https://doi.org/10.1093/bioinformatics/btp612.

Wang, X., Hsueh, Y.-P., Li, W., Floyd, A., Skalsky, R., & Heitman, J. (2010). Sex-induced silencing defends the genome of Cryptococcus neoformans via RNAi. *Genes & Development*, *24*, 2566–2582. https://doi.org/10.1101/gad.1970910.

Wang, D. Y., Kumar, S., & Hedges, S. B. (1999). Divergence time estimates for the early history of animal phyla and the origin of plants, animals and fungi. *Proceedings of the Biological Sciences*, *266*, 163–171. https://doi.org/10.1098/rspb.1999.0617.

Wang, Q.-M., Liu, W.-Q., Liti, G., Wang, S.-A., & Bai, F.-Y. (2012). Surprisingly diverged populations of Saccharomyces cerevisiae in natural environments remote from human activity. *Molecular Ecology*, *21*, 5404–5417. https://doi.org/10.1111/j.1365-294X.2012.05732.x.

Wang, M., Weiberg, A., Lin, F.-M., Thomma, B. P. H. J., Huang, H.-D., & Jin, H. (2016). Bidirectional cross-kingdom RNAi and fungal uptake of external RNAs confer plant protection. *Nature Plants*, *2*, 16151, https://doi.org/10.1038/nplants.2016.151.

Weiberg, A., Wang, M., Lin, F.-M., Zhao, H., Zhang, Z., Kaloshian, I., et al. (2013). Fungal small RNAs suppress plant immunity by hijacking host RNA interference pathways. *Science*, *342*, 118–123. https://doi.org/10.1126/science.1239705.

Wiley, E. O. (1978). The evolutionary species concept reconsidered. *Systematic Zoology*, *27*, 17–26. https://doi.org/10.2307/2412809.

Wu, C. H., Apweiler, R., Bairoch, A., Natale, D. A., Barker, W. C., Boeckmann, B., et al. (2006). The universal protein resource (UniProt): An expanding universe of protein information. *Nucleic Acids Research*, *34*, D187–191. https://doi.org/10.1093/nar/gkj161.

Wu, T. D., & Watanabe, C. K. (2005). GMAP: A genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics (Oxford, England)*, *21*, 1859–1875. https://doi.org/10.1093/bioinformatics/bti310.

Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, *24*, 1586–1591. https://doi.org/10.1093/molbev/msm088.

Yike, I. (2011). Fungal proteases and their pathophysiological effects. *Mycopathologia*, *171*, 299–323. https://doi.org/10.1007/s11046-010-9386-2.

Zerbino, D. R., & Birney, E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, *18*, 821–829. https://doi.org/10.1101/gr.074492.107.

Zhang, Z., López-Giráldez, F., & Townsend, J. P. (2010). LOX: Inferring Level Of eXpression from diverse methods of census sequencing. *Bioinformatics*, *26*, 1918–1919. https://doi.org/10.1093/bioinformatics/btq303.

Zhang, X., Wang, Y., Chi, W., Shi, Y., Chen, S., Lin, D., et al. (2014). Metalloprotease genes of Trichophyton mentagrophytes are important for pathogenicity. *Medical Mycology*, *52*, 36–45. https://doi.org/10.3109/13693786.2013.811552.

Zhao, Z.-M., Campbell, M. C., Li, N., Lee, D. S. W., Zhang, Z., & Townsend, J. P. (2017). Detection of regional variation in selection intensity within protein-coding genes using DNA sequence polymorphism and divergence. *Molecular Biology and Evolution*. https://doi.org/10.1093/molbev/msx213.