Fungal Biofilm Morphology Impacts Hypoxia Fitness and Disease Progression

Caitlin H. Kowalski[1], Joshua D. Kerkaert[1], Ko-Wei Liu[1], Matthew C. Bond[2], Raimo Hartmann[3], Carey D. Nadell[2], Jason E. Stajich[4], and Robert A. Cramer[1*]

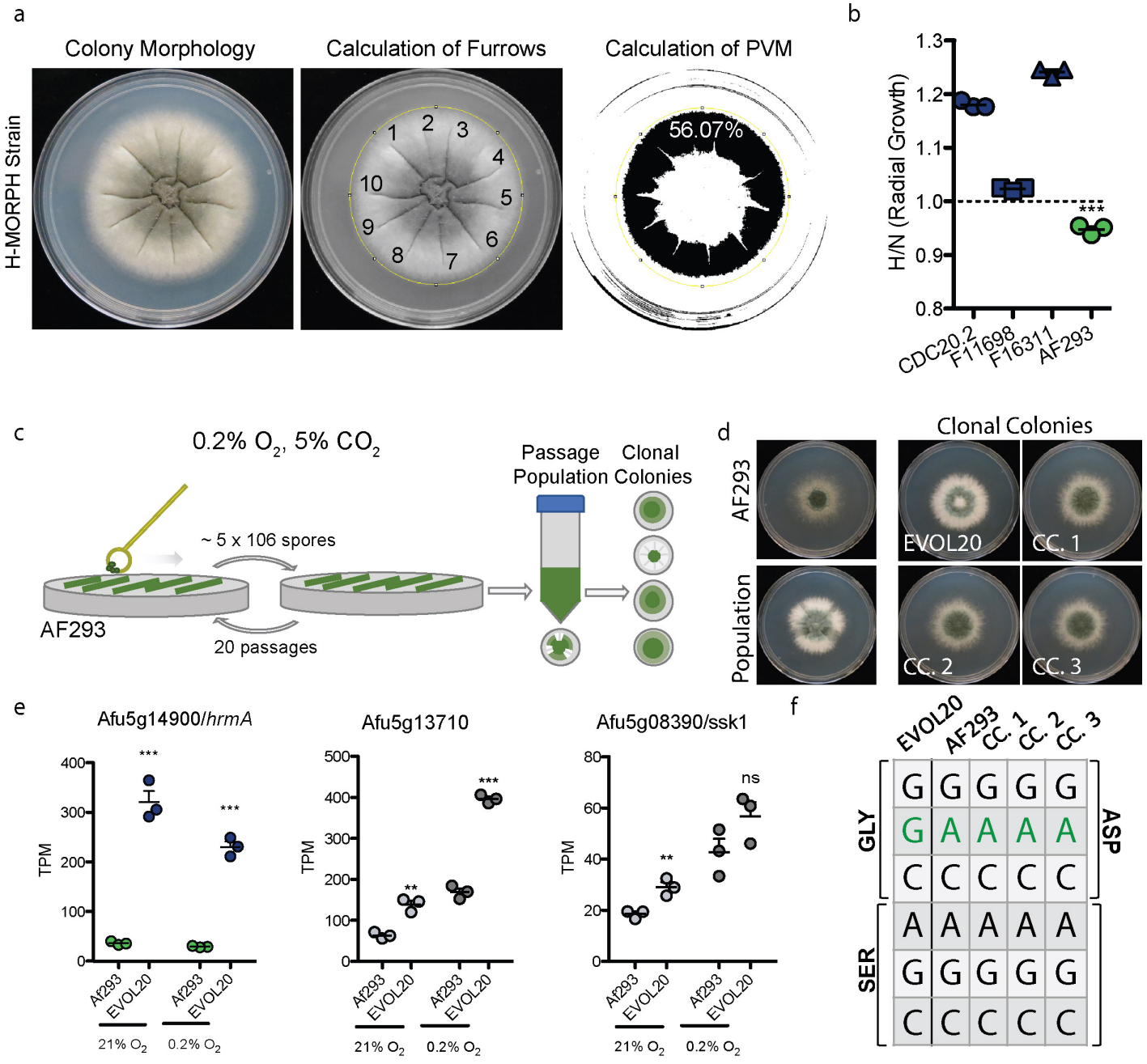
[1]Department of Microbiology and Immunology, Geisel School of Medicine at Dartmouth, Hanover, NH, USA

[2] Department of Biological Science, Dartmouth College, Hanover, NH, USA

[3]No Current affiliation

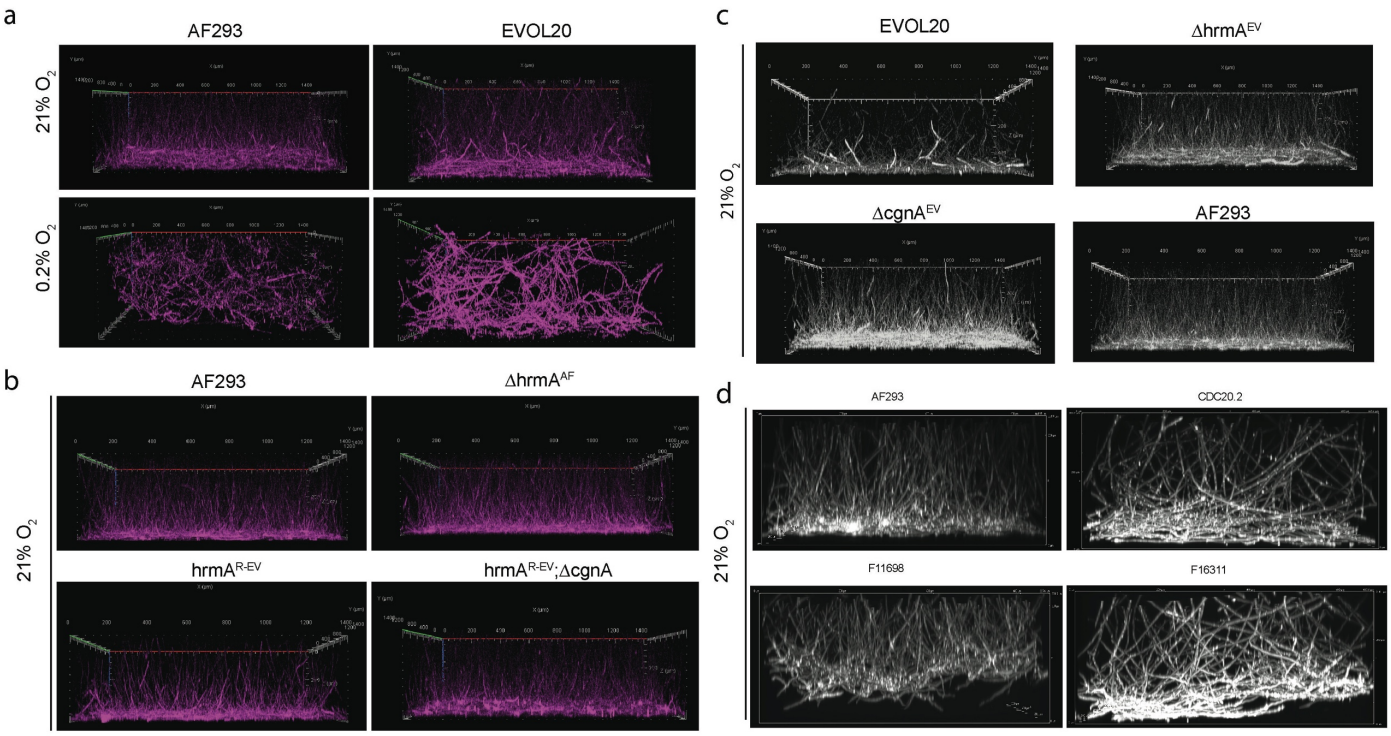
[4] Department of Microbiology and Plant Pathology and Institute for Integrative Genome Biology, University of California-Riverside, Riverside, California, USA

[*]Corresponding Author and Lead Contact: Robert.A.Cramer.JR@Dartmouth.edu

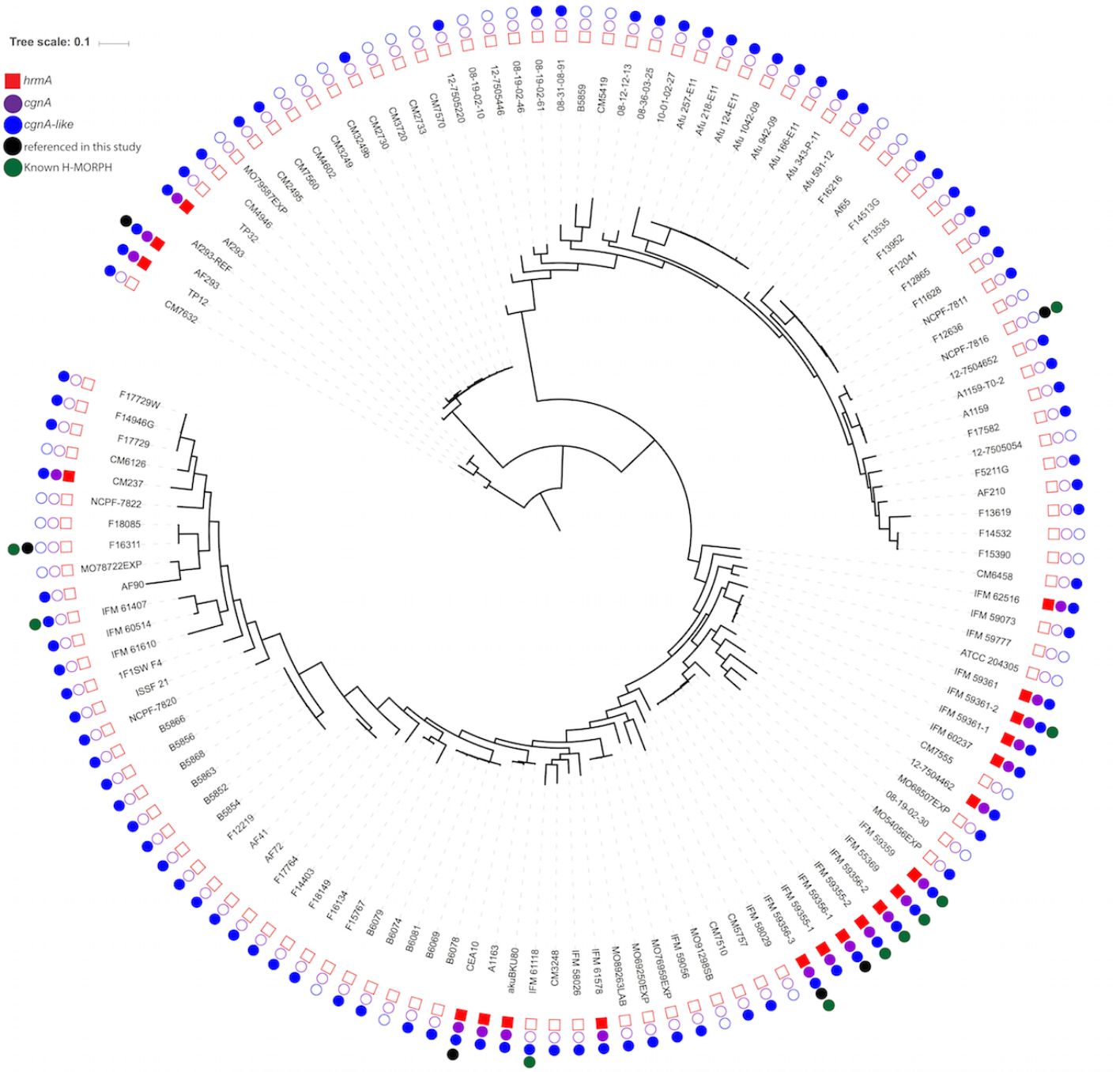


Supplementary Figure 1. Whole genome sequencing and RNA sequencing of the hypoxia-evolved EVOL20 identified three nonsynonymous mutations. **a.** Example identification and quantification of the morphological characteristics of interest: furrowing and PVM**.** This is a representative example of a single analysis that was repeated a three times per sample an averaged. **b**. H-MORPH strains (blue) have significantly increased hypoxia fitness (***: p<0.001 by One-way ANOVA with Dunnett multiple comparison test) compared to AF293. n=3 biologically independent samples. Error bars indicate standard error around the

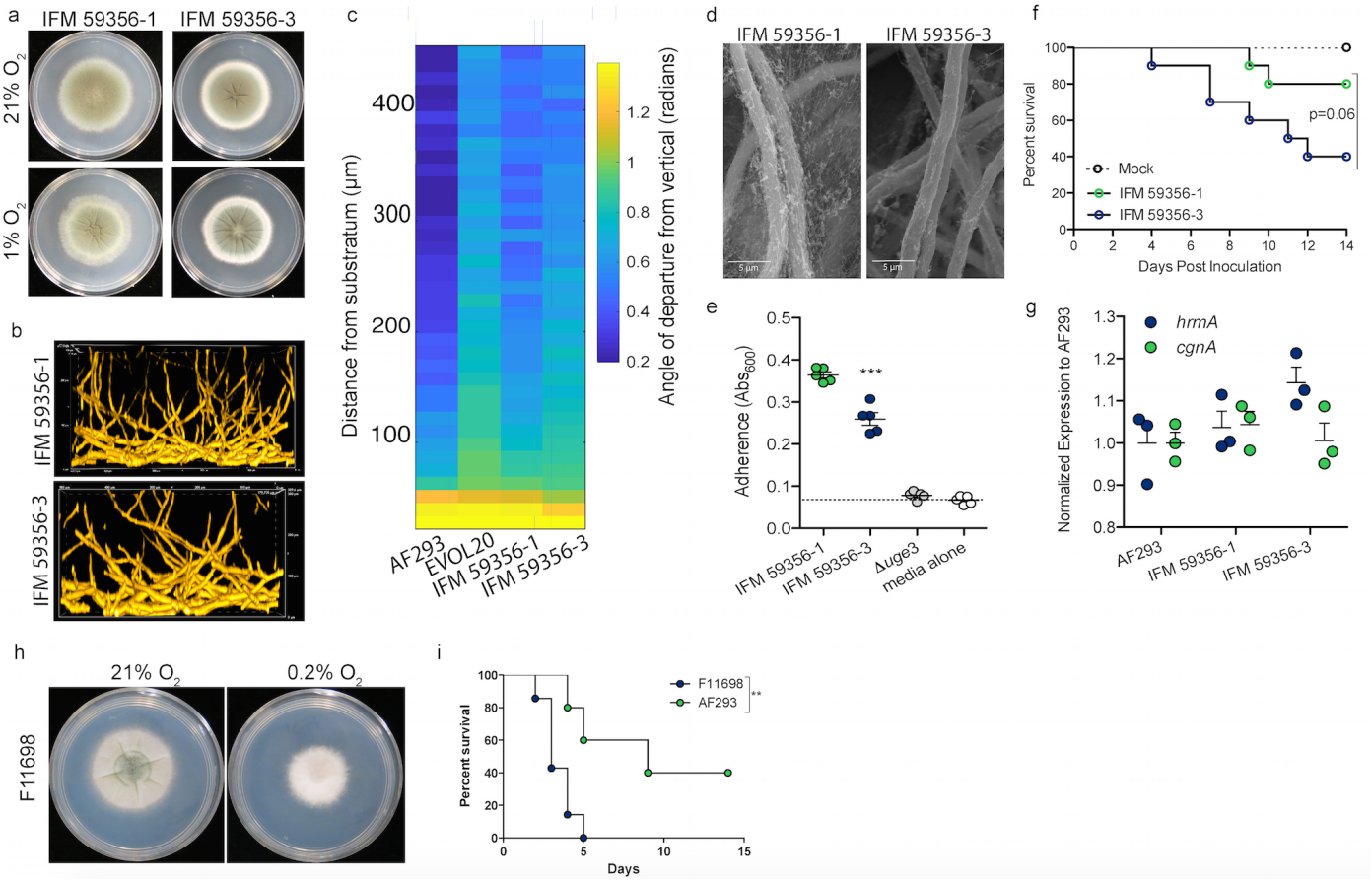
mean (centre). **c.** Experimental evolution schematic. **d**. Following 20 passages in 0.2% O_2 a heterogeneous population was sorted into individual clonal colonies. Images representative of three independent experiments. **e.** Afu5g14900/*hrmA*, Afu5g08390/*ssk1*, Afu5g13710 have non-synonymous mutations in EVOL20 v. AF293. RNA-sequencing shows all three to be oxygen responsive. Two-tailed unpaired t tests show *hrmA* be elevated EVOL20 (***: p<0.0001) as well as AFu5g13710 (**:p=0.0020, ***: p<0.0001) and Afu5g08390/ssk1 (**:p=0.0085, ns=0.1350) .n=3 independent biological samples. **f**. The SNP in *hrmA* (D304G) is only found in the EVOL20 clonal colony.



Supplementary Figure 2. Full field of view biofilm images used in alignment analysis. Full volume view of the submerged fungal biofilms (XZ plane) with constitutively fluorescent fungi or biofilms dyed with calcofluor while. In a, b, c biofilms are 1400 µm X 1400 µm X ~500 µm. In d, biofilms are 665.5 µm X 665.5 µm X ~300 µm. Experiments were repeated a minimum of 3 times showing similar results.

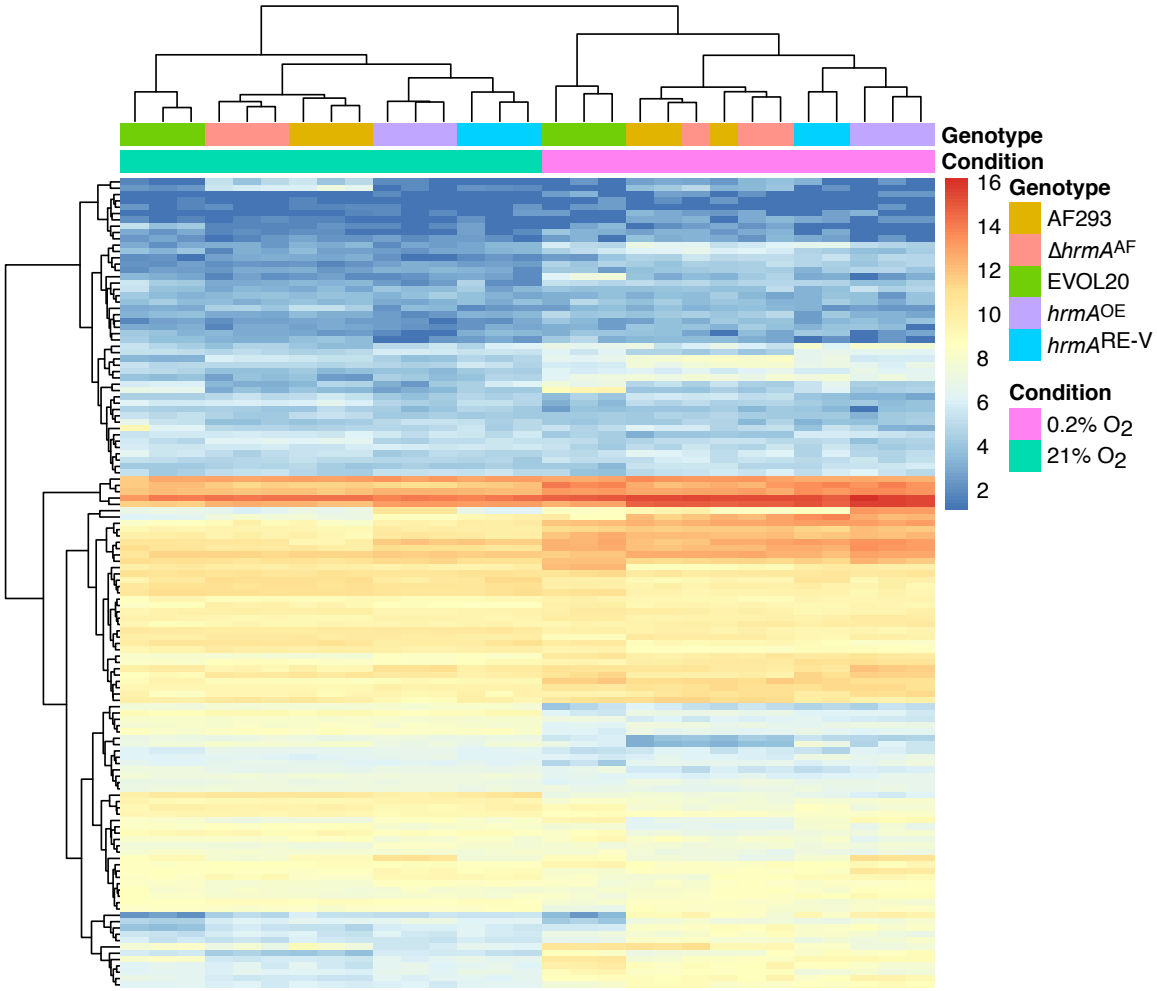


Supplementary Figure 3. *Aspergillus fumigatus* phylogeny of publically available genomes depicting presence and absence of *hrmA* and *cgnA*. *A. fumigatus* species phylogeny indicates the presence of *hrmA* (■, 95% ID) and *cgnA* (●, 95% ID detected within same contig as *hrmA*) in a genetically diverse set of strains, while the majority of strains encode least one *cgnA*-like sequence (●). Strains investigated further in this study are indicated with ●, and known H-morph strains are indicated with ●.

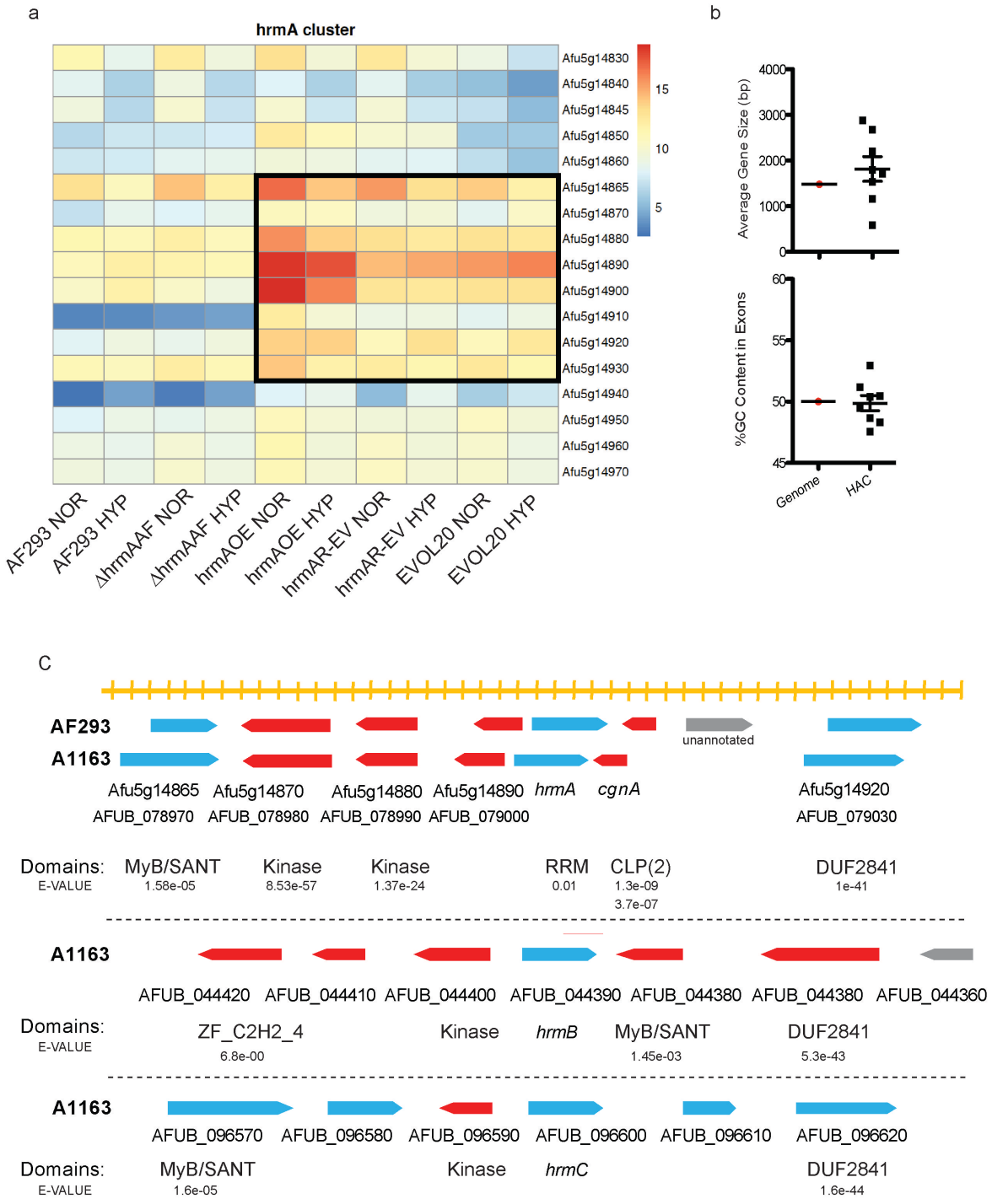


Supplementary Figure 4. Phenotypic characterization of H-MORPH clinical strains. a. Co-isolated clinical strains display N-MORPH and H-MORPH (representative of 3 independent experiments) **b.** IFM 59356-3 biofilm architecture has many horizontal filaments quantified as greater deviation from the vertical compared to the N-MORPH IFM 59356-1 (**c**). **b** and **c** are representative of 3 biologically independent samples from 2 independent experiments. Biofilms in **b** are approximately 500 μm X 300 μm X 200 μm. **d**. IFM 59356-3 has reduced ECM attachment (representatives from 2 independent experiments) than IFM 59356-1. **e.** IFM 59356-3 adheres less well (***:p=0.0002 by two-tailed unpaired t test, n=5 biologically independent samples per group) compared to IFM 59356-1. Error bars indicate standard error around the mean (centre). **f.** IFM 59356-3 (n=10 biologically independent animals) has 40% greater mortality at 14 dpi compared to IFM 59356-1 (n=10 biologically independent animals) (p=0.0651 by Gehan-Breslow-Wilcoxon). **g** *hrmA* nor *cgnA* mRNA are elevated in IFM 59356-3 compared to IFM 59356-1 (n=3 biologically independent samples, error bars indicate standard error around the mean (centre)). **h**. F11698 shows H-MORPH (representative from 3 independent

experiments). **i.** F11698 (n=7 biologically independent animals) is more virulent than AF293 (n=5 biologically independent animals) (*:p=0.0128 by Gehan-Breslow-Wilcoxon).

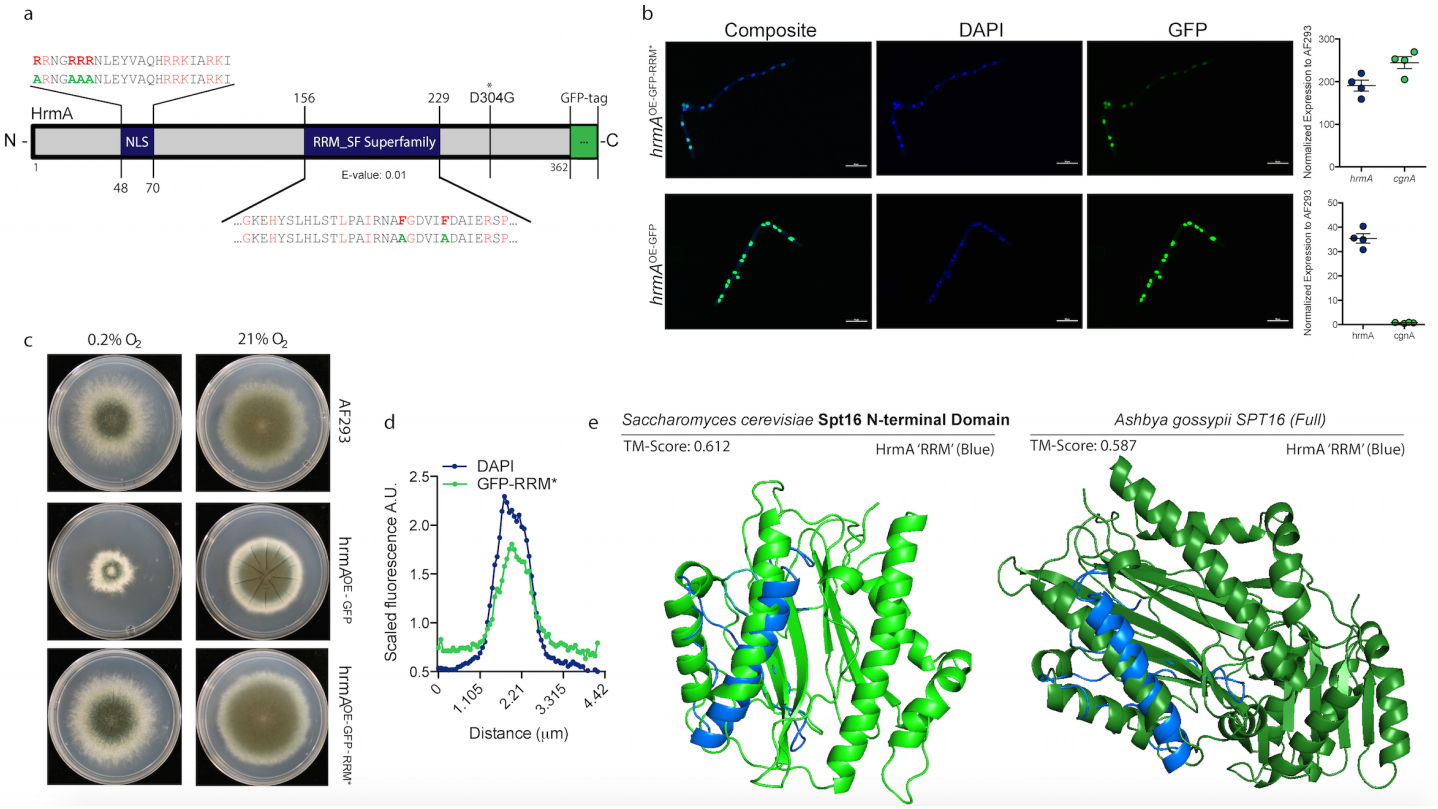


Supplementary Figure 5. Transcriptional profiles cluster overexpression of the reference allele of hrmA (*hrmA*^{OE}) with expression of the hypoxia-evolved allele (*hrmA*$^{R\text{-}EV}$). All genes were plotted that have differential expression across the genotypes with p<0.05 (two-sided) as determined through DESeq2 which utilizes Wald test for differential expression and adjusted for multiple comparisons using Benjamini and Hochberg procedures. Any genes where total FPKM <1 were excluded. These data indicate that AF293 and ΔhrmA^{AF} cluster closely together at 0.2% O2 and 21% O2. The hrmA$^{R\text{-}EV}$ and hrmA^{OE} cluster together in normoxia and hypoxia indicating that these strains have similar transcriptional profiles despite the differences in the hrmA allele. EVOL20 has a more complex genetic background and clusters separately.

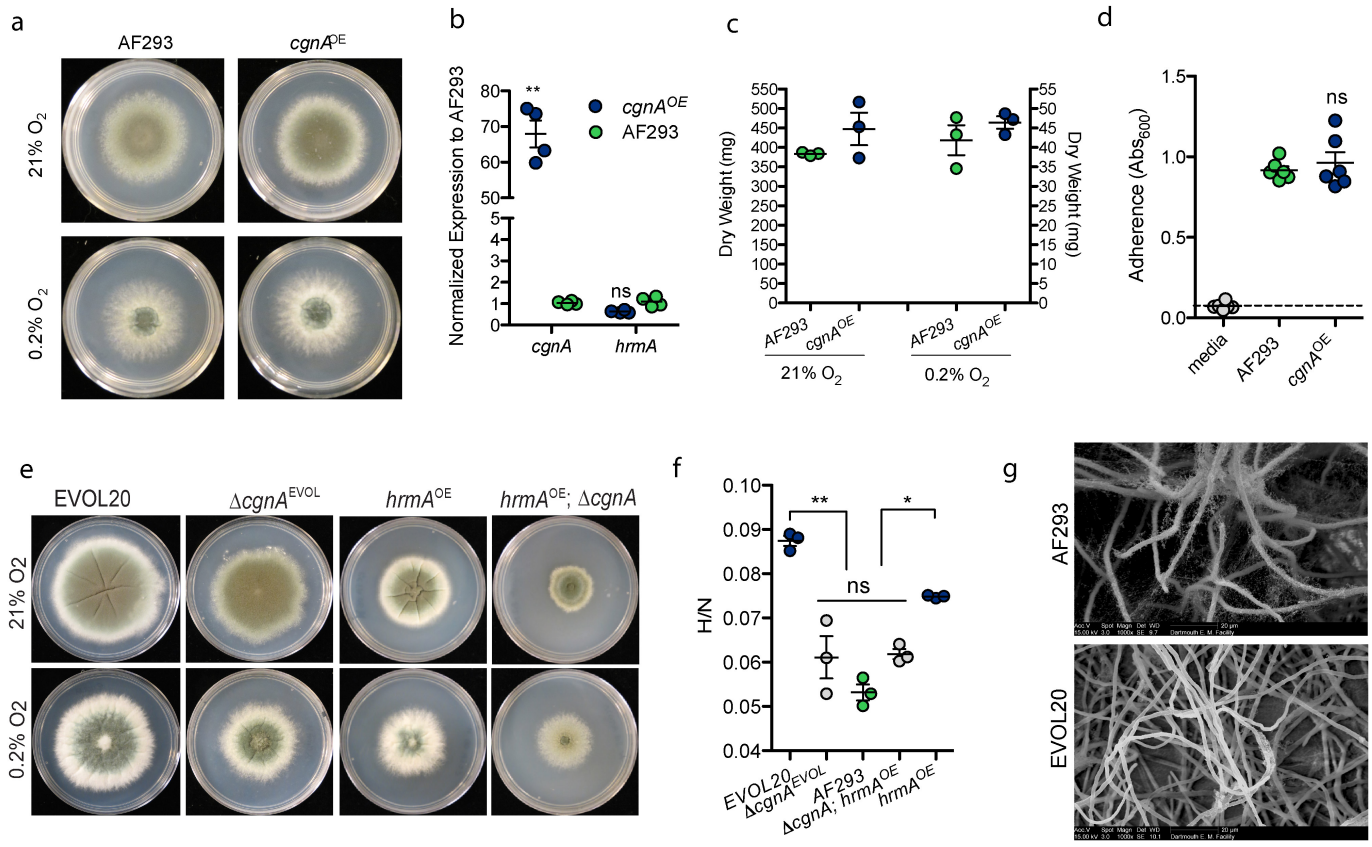


Supplementary Figure 6. HAC is conserved in A1163, a strain that additionally encodes 2 putative orthologous gene clusters. a. Heat map of co-regulated HAC genes from Afu14865 to Afu5g1430. **b.** Comparisons of HAC average gene size and exon GC content with the published averages for the AF293 genome reveal a slightly increased average gene size and a similar GC content to the genome. Error bars

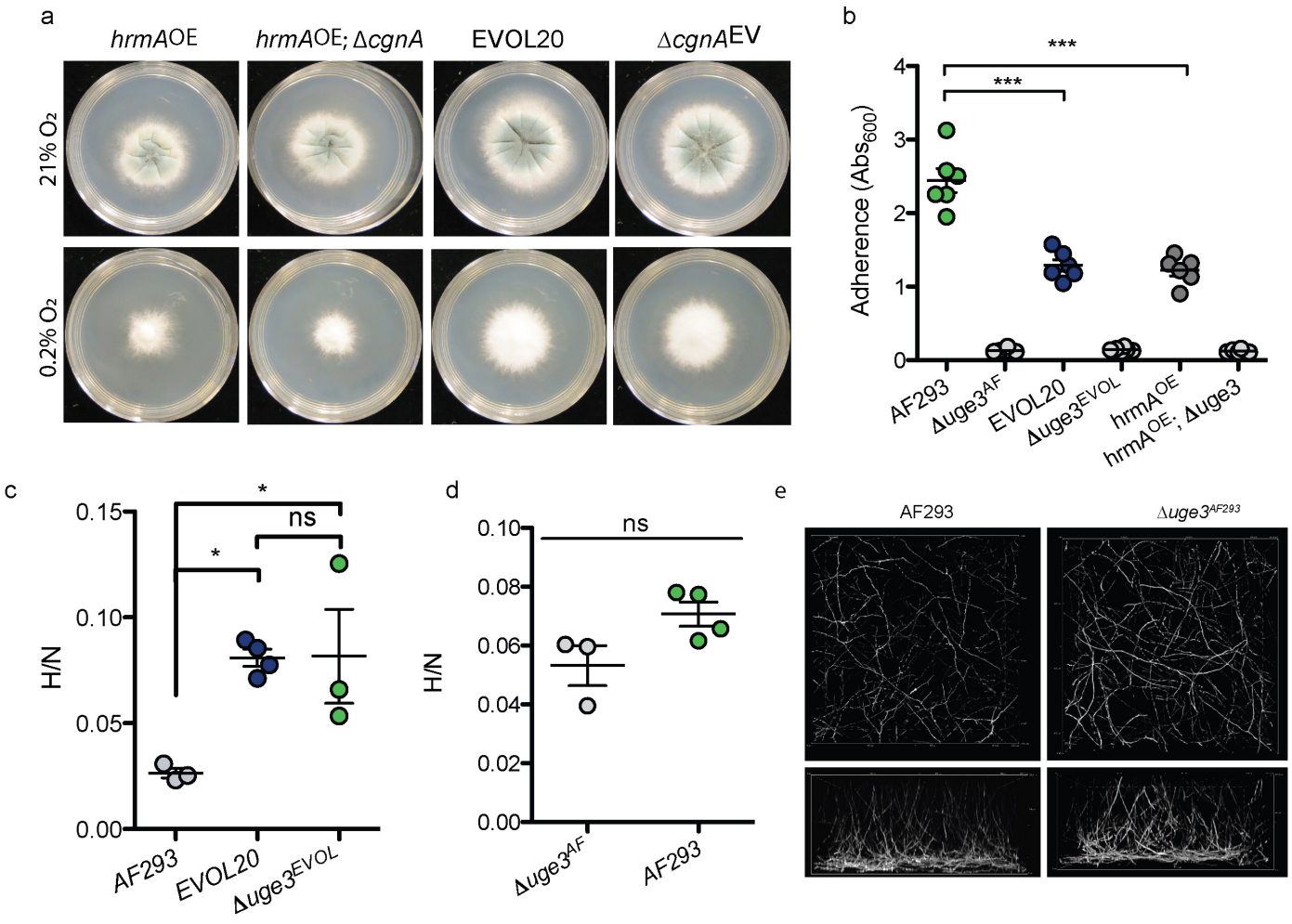
indicate standard error around the mean (centre). **c**. Closely related strains AF293 and A1163 have assembled genomes and reveal synteny within HAC across these two strains. A1163 also encodes two putative orthologous clusters with putative *hrmA* orthologs: AFUB_044390/*hrmB* and AFUB_096600/*hrmC.*



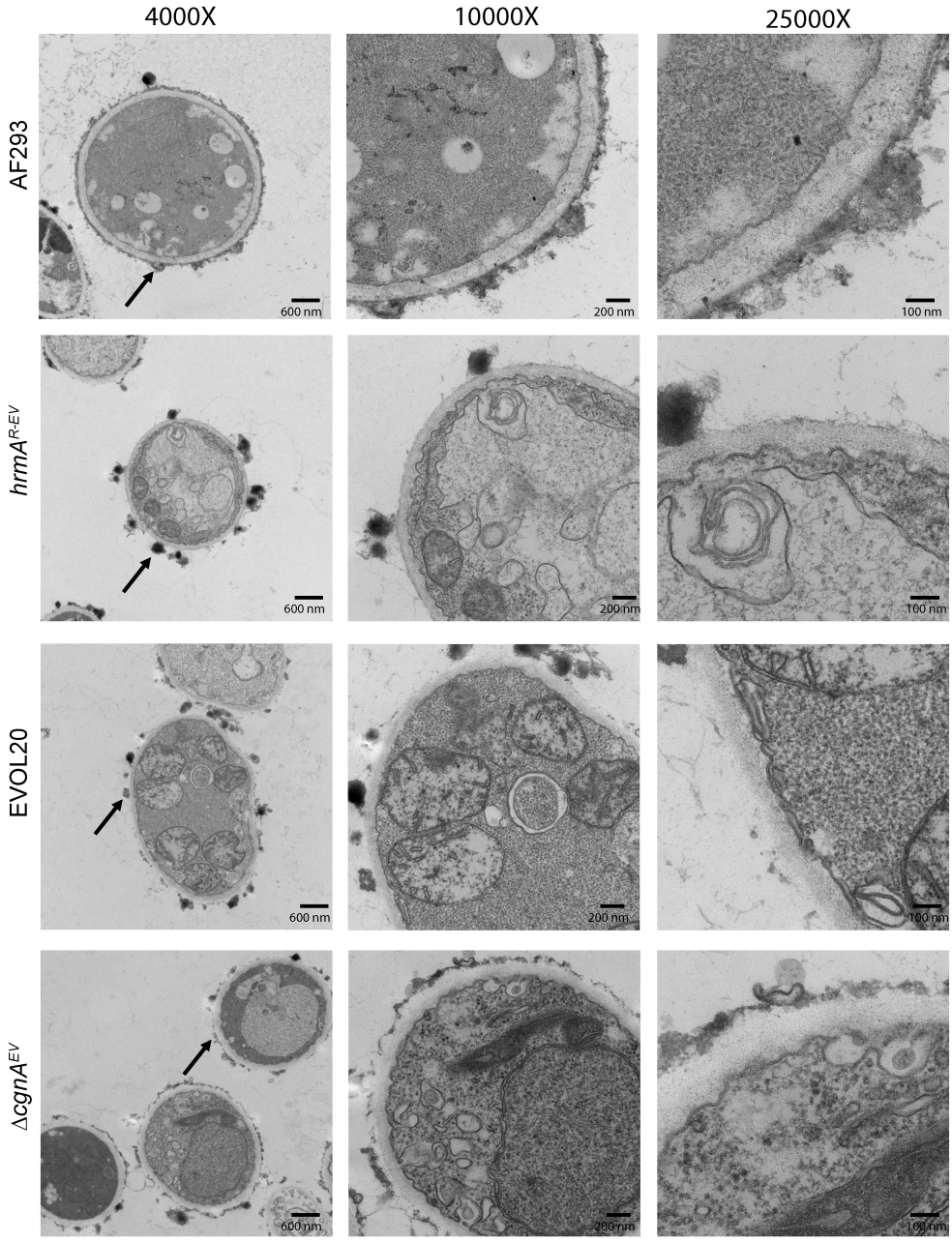
Supplementary Figure 7. HrmA has a RNA recognition motif (RRM) with two conserved phenylalanine residues necessary for H-MORPH. a. Model of *hrmA* with an N-terminal bipartite NLS, a weakly predicted RRM domain (E-VALUE 0.01), EVOL20 SNP locale, and the GFP tag. Green residues are the result of the site directed mutagenesis. **b**. Two conserved phenylalanine residues with the RRM domain are not necessary for nuclear localization (images: n=10 biologically independent samples from 3 independent experiments), but are necessary for induction of *cgnA* (n=3 biologically independent samples). Scale bars are 10 μm. Error bars indicate standard error around the mean (centre). **c**. The RRM mutant no longer adopts the H-MORPH characteristics of elevated PVM and furrowing (representative of 3 independent experiments). **d.** Quantification of the localization signal (representative of 5 independent experiments). **e.** iTASSER protein modeling of the RRM domain of HrmA (156-229 residues) maps the structure of the chromatin remodeling protein Spt16 in *Saccharomyces cerevisiae* and *Ashbya gossypii.*



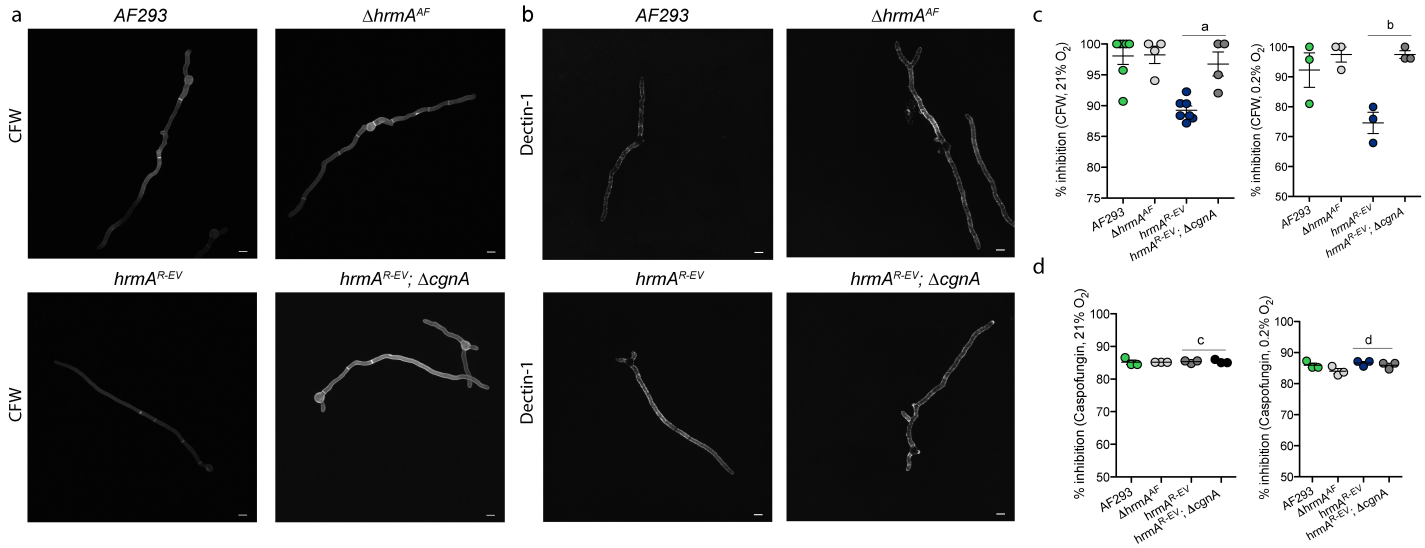
Supplementary Figure 8. Overexpression of *cgnA* alone is not sufficient to generate H-MORPH. a. Ectopic over expression of *cgnA* does not generate H-MORPH in AF293 (representative of 3 independent experiments), does not impact *hrmA* RNA levels (n=4 biologically independent samples, **: p=0.0014 and ns: p=0.3960 by two-tailed unpaired t test) (**b**), does not alter fungal growth at 21% O_2 or 2% O_2 (n=3 biologically independent samples) (**c**), and does not impact surface adherence of AF293 (n=6 biologically independent samples, ns: p= 0.5347 by unpaired t test) (**d**). **e.** Loss of *cgnA* in the HAC inducing strains EVOL20 and *hrmA*^{OE} results in a loss of H-MORPH (representative of 3 independent experiments) and a reduction in hypoxia fitness (n=3 biologically independent samples, **: p<0.0001, *: p=0.0134, ns: p=1103, p=0.1579, by One-way ANOVA with Sidak's multiple comparisons test) (**f**). **g.** ECM detachment from EVOL20 but not AF293 (representative of 2 independent experiments). Scale bars: 20 μm. All error bars indicate the standard error around the mean (centre).



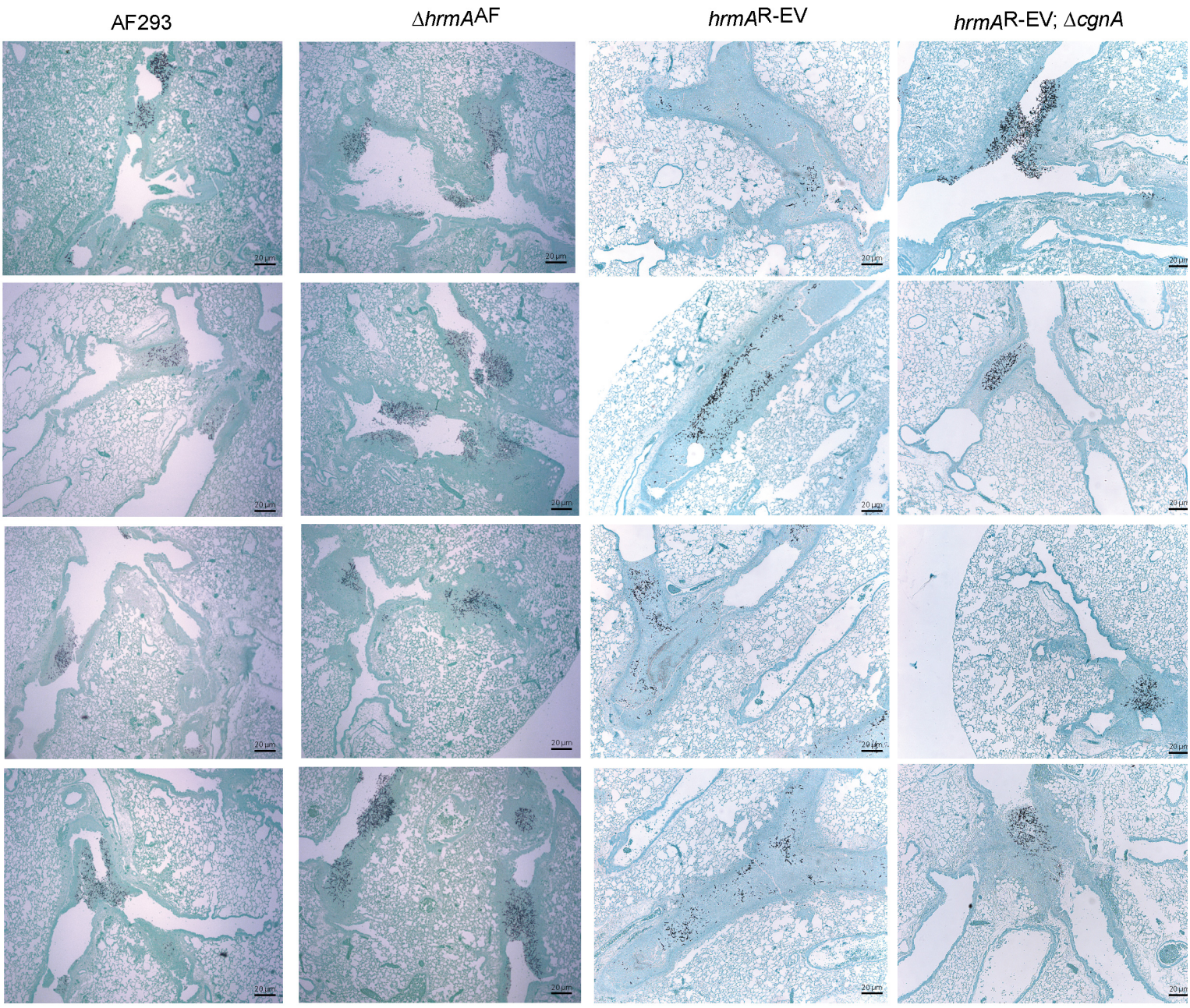
Supplementary Figure 9. Loss of *uge3* does not impact H-MORPH or hypoxia fitness. A. Deletion of *uge3* from *hrmA*^{OE} or EVOL20 does not impact the macroscopic H-MORPH (representative of 3 independent experiments) (**a**), but as expected reduces adherence of the strains (n=6 independent biological samples, ***: p<0.0001 One-way ANOVA with Dunnett multiple comparisons test) (**b**)**. c.** Δ*uge3*^{EVOL} does not impact hypoxia fitness of EVOL20 (n=3 biologically independent samples, a: p=0.0294, b: p=0.0370, ns: p=0.9988 by One-way ANOVA with Tukey's Multiple Comparisons test) (**d**) nor hypoxia fitness of AF293 (n=3 biologically independent samples, ns: p=0.0672 by two-tailed unpaired t test). **e.** Loss of *uge3* in AF293 does not impact biofilm architecture (representative of 3 independent experiments). All error bars indicate standard error around the mean (centre).



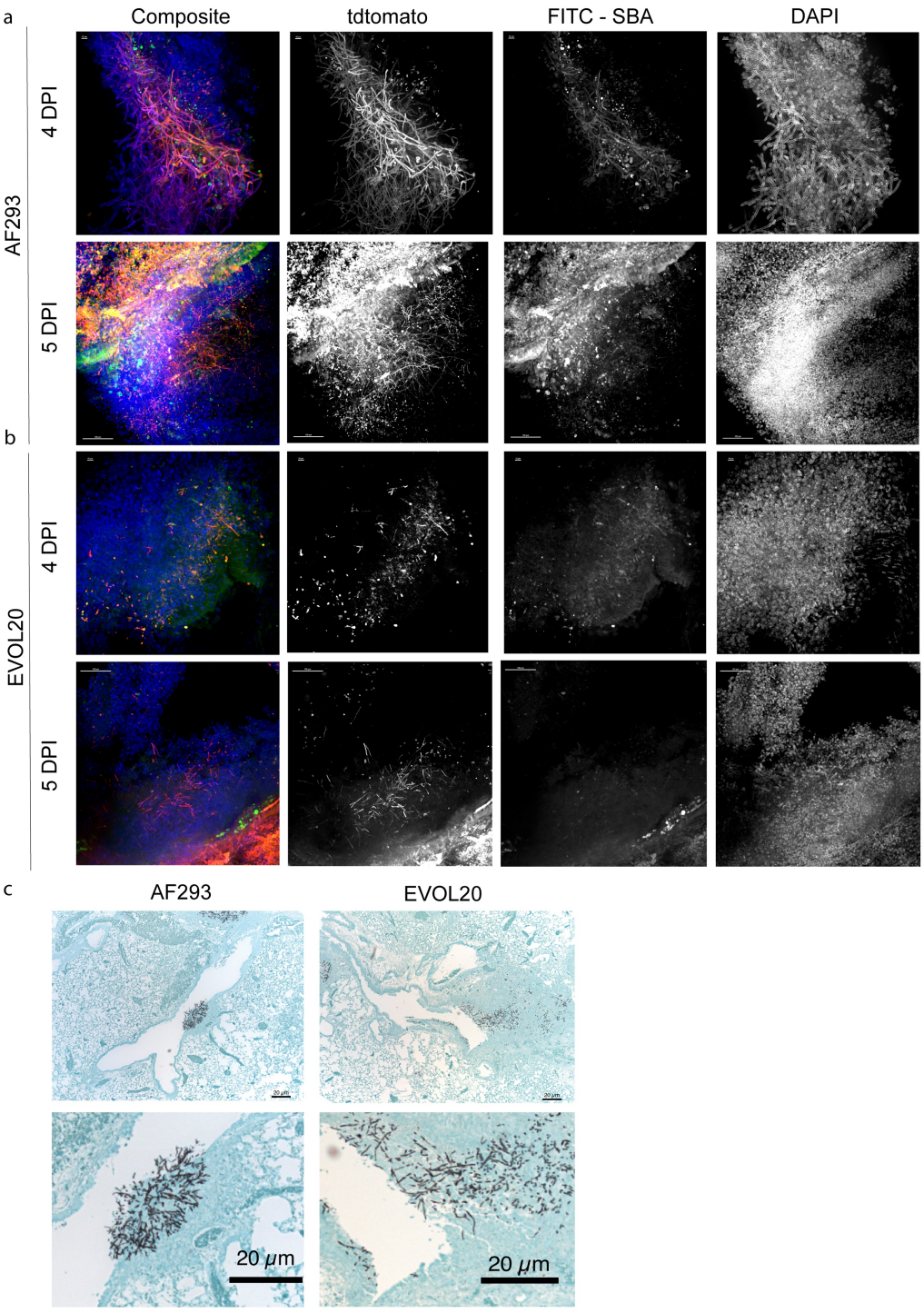
Supplementary Figure 10. Representative TEM images used for cell wall thickness analysis. Black arrows indicate electron-dense material presumed to be the extracellular matrix that is aggregated in the *hrmA*$^{R\text{-}EV}$ and EVOL20 H-MORPH strains. Scale bars for 4000X images are 600 μm, for 10000X images are 200 μm, and for 25000X images are 100 μm. Images are representative of 13 biologically independent samples from 2 biologically independent experiments.



Supplementary Figure 11. Cell wall staining and Cell wall perturbing agents. a. Calcofluor white (CFW) staining of hyphae and processed as SUM projections. **b.** Soluble Dectin-1 staining for detection of β- glucan and processed as SUM projections. Scale bars are 5 μm. Images representative from 3 biologically independent experiments. **c.** At 21% or 0.2% O_2, H-MORPH *hrmA*$^{R\text{-}EV}$ (n=7 biologically independent samples) is more sensitive to growth on CFW than *hrmA*$^{R\text{-}EV}$*;*Δ*cgnA* (n=4 biologically independent samples) (a: p=0.0016, b: p=0.0037 by two-tailed unpaired t test). **d.** In contrast, there is no impact of *hrmA/cgnA* on fungal sensitivity to the Echinocandin Caspofungin (n=3 biologically independent samples, c: p=0.9313, d: p=0.4550 by two-tailed unpaired t test). All error bars indicate standard error around the mean (centre).

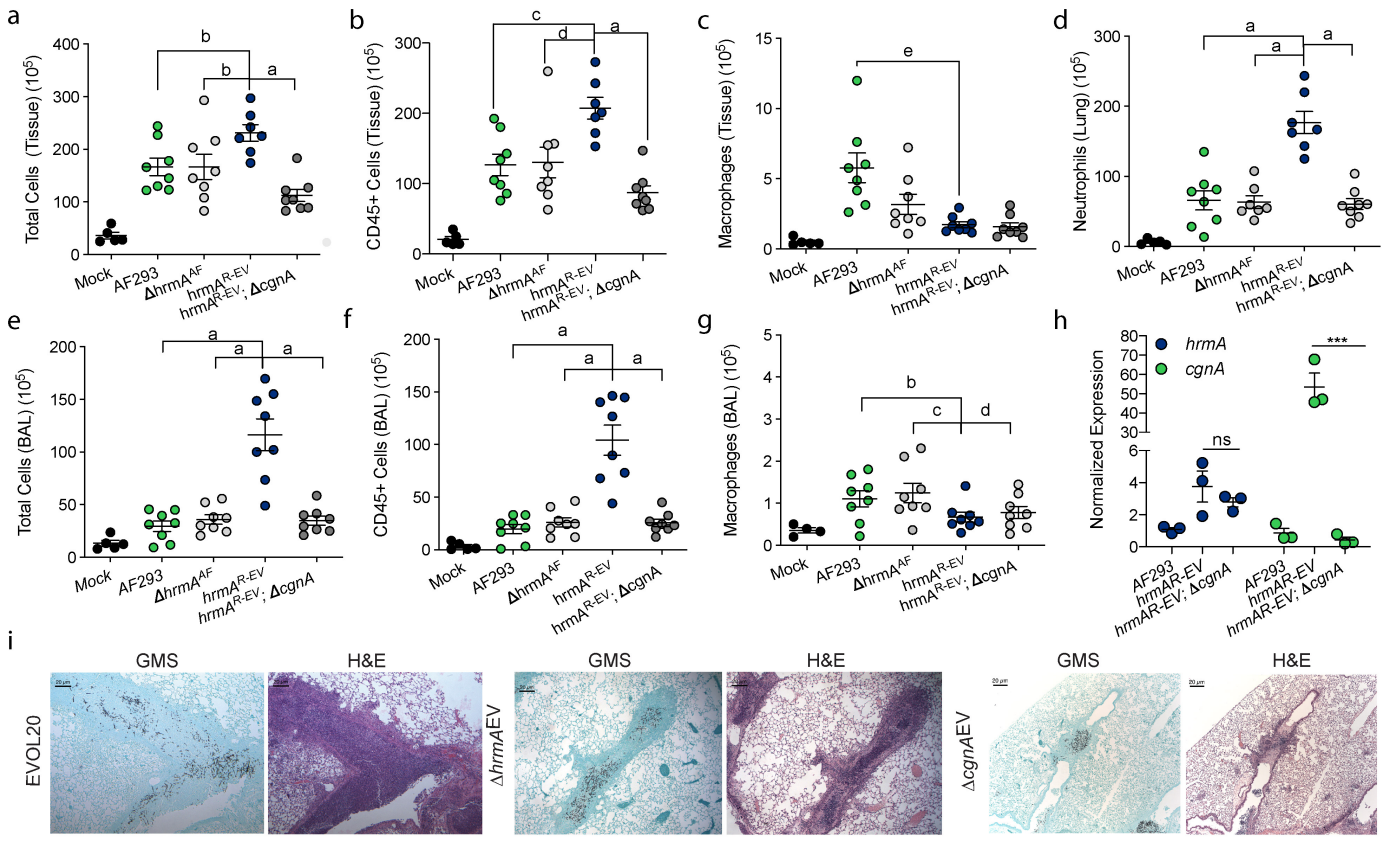


Supplementary Figure 12. Example histopathology slides utilized in the analysis of lesion diffuseness. Four representative images of GMS stained fungal lesions within the murine lung per strain group. Airways inoculated with *hrmA*$^{R\text{-}EV}$ have more diffuse lesions that spread across the span of the airway and are not localized in a single foci of infection, as observed with AF293, Δ*hrmA*^{AF}, and Δ*hrmA*^{EV}. Scale bars: 20 μm. Images are representative of a minimum of 5 biologically independent animals from a 2 independent experiments.



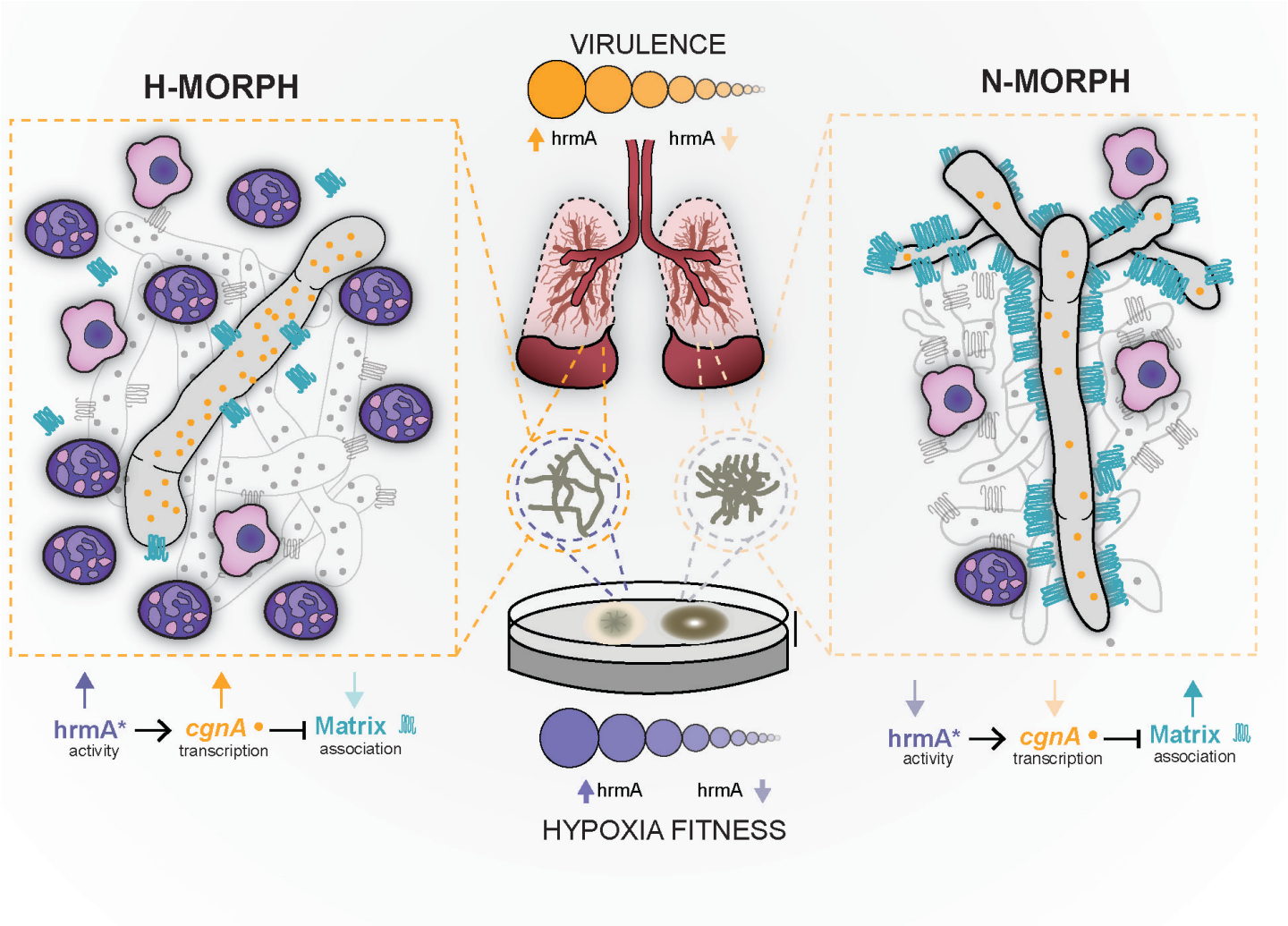
Supplementary Figure 13. FunPact images reflect lesion morphology from histopathology. a. A293 lesions show complex interconnection between fungal filaments. **b.** EVOL20 lesions show diffuse fungal lesions at both day 4 and 5 post inoculation. **c.** Af293 lesions are compact in the airways while EVOL20 lesions

are more diffuse. Scale bars - 4 DPI: 10 μm, 5 DPI: 100 μm, histopathology: 20 μm. All images are representative from 5 biologically independent animals from 2 independent sample preparations.

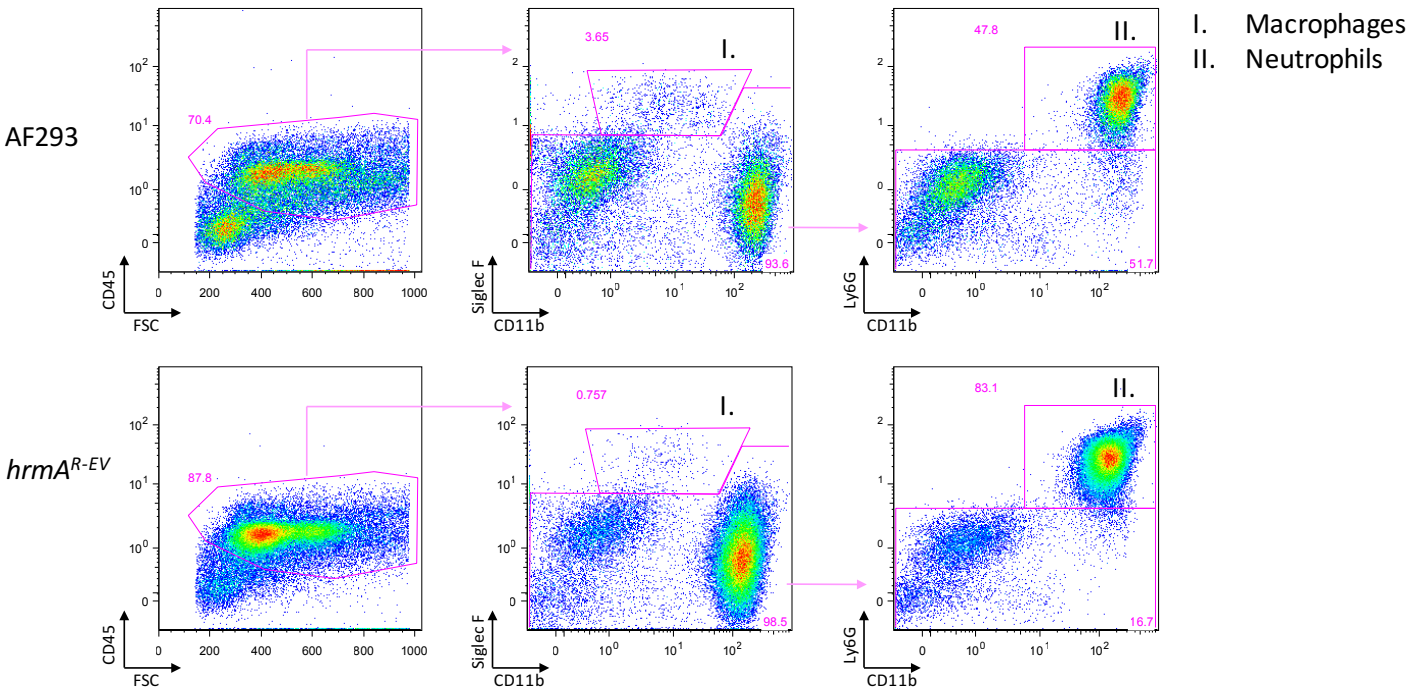


Supplementary Figure 14. Cellularity from tissue and histopathology. In murine lung tissue total cells (a: p<0.0001, b: p=0.0384 by One-way ANOVA with Dunnett's Multiple Comparison test) (**a**) and CD45+ leukocytes (a: p<0.0001, c: p=0.0033, d: p=0.0049 One-way ANOVA with Dunnett's Multiple Comparison test) (**b**) are elevated at 60 hpi with *hrmA*$^{R\text{-}EV}$ (n=7 biologically independent animals) relative to the other strains (n=8 biologically independent animals per group). **c.** Macrophages are significantly reduced between AF293 (n=8 biologically independent animals) and *hrmA*$^{R\text{-}EV}$ (n=7 biologically independent animals) in the lung tissue (e:p=0.0002 by One-way ANOVA with Dunnett's Multiple Comparison test), but (**d**) *hrmA*$^{R\text{-}EV}$ neutrophils (n=7 biologically independent animals) are significantly elevated relative to all groups (n=8 biologically independent animals per group) (a: p<0.0001 by One-way ANOVA with Dunnett's Multiple Comparison test). In BAL, total cells (a: p<0.0001 by One-way ANOVA with Dunnett's Multiple Comparison test) (**e**) and CD45+ leukocytes (a: p<0.0001 One-way ANOVA with Dunnett's Multiple Comparison test) (**f**) are elevated in *hrmA*$^{R\text{-}EV}$ (n=8 biologically independent animals per group), but not macrophages (b: p=0.2356, c: p=0.0755, d: p=0.9774, by

One-way ANOVA with Dunnett's Multiple Comparison test) (**g**). **h**. Deletion of *cgnA* in $hrmA^{R\text{-}EV}$ does not impact the *hrmA* mRNA *in vitro* (n=3 biologically independent samples) (ns: p=0.4508, **: p=0.0002, One-way ANOVA with Dunnett's Multiple Comparison test). All error bars indicate standard error around the mean (centre). **i**. GMS and H&E histopathology of airway lesions of EVOL20, $\Delta hrmA^{EV}$, and $\Delta cgnA^{EV}$ (Scale bars: 20 μm). Images are representative of 5 biologically independent animals from 2 independent experiments.



Supplementary Figure 15. Model for the impact of macroscopic morphology of disease progression of *A. fumigatus.* Hypoxic environments, such as the infected host lung, provide an adaptive pressure to the obligate aerobic mold *Aspergillus fumigatus*. We suspect this selective pressure leads to the adoption of certain hypoxia-typic morphological features – furrowing, elevated PVM, and that strains where these morphological traits occur regardless of oxygen tensions have a fitness advantage in low oxygen and simultaneously an increase in virulence. Our model suggests that the macroscopic morphology (N-MORPH v. H-MORPH) is an indicator of biofilm architecture and hyphal surface characteristics that provide a loose disorganized biofilm and highly inflammatory diffuse fungal lesions. Here we describe a sub-telomeric gene cluster regulator HrmA that promotes expression of CLP *cgnA* that alters the fungal surface preventing matrix attachment as one mechanism, of likely many, for the generation of H-MORPH characteristics.



Supplementary Figure 16. Representative pseudocolor plot of the Flow Cytometry gating. The neutrophils were identified as $CD45^{+}SiglecF^{-}Ly6G^{+}CD11b^{+}$ cells and alveolar macrophages as $CD45^{+}SiglecF^{+}CD11b^{dim}$ cells. The cells were first gated on singlet live cells.

Supplementary Video Legends

Supplementary Video 1: AF293 Biofilm at 21% O_2 N-MORPH strain AF293 cultured at 21% O_2 forms an organized biofilm with straight filaments growing toward the air-liquid interface. This video is representative of a minimum of three independent experiments.

Supplementary Video 2: AF293 Biofilm at 0.2% O_2 N-MORPH strain AF293 cultured at 0.2% O_2 forms a high disorganized biofilm with many filaments that divert from the vertical. This video is representative of a minimum of three independent experiments.

Supplementary Video 3. EVOL20 Biofilm at 21% O_2 H-MORPH strain EVOL20 cultured at 21% O_2 forms a disorganized biofilm with many filaments that divert from the vertical. This video is representative of a minimum of three independent experiments.

Supplementary Video 4. EVOL20 Biofilm at 0.2% O_2 H-MORPH strain EVOL20 cultured at 21% O_2 forms a disorganized biofilm with filaments that divert from the vertical. This video is representative of a minimum of three independent experiments.

Supplementary Video 5. funPACT AF293 Lesion Day 5 Post Inoculation. AF293 20X lesion from Figure S11 shown in 3D video. The fungal filaments are red (tdtomato), the host nuclei are blue (DAPI), and SBA staining of *N*-acetyl-galactosamine are green (FITC-SBA). This video is representative of a minimum of five independent samples from two independent sample preparations.

Supplementary Video 6. funPACT EVOL20 Lesion Day 5 Post Inoculation. EVOL20 20X lesion from Figure S11 shown in 3D video. The fungal filaments are red (tdtomato), the host nuclei are blue (DAPI), and SBA staining of *N*-acetyl-galactosamine are green (FITC-SBA). This video is representative of a minimum of five independent samples from two independent sample preparations.

Supplemental Methods

Whole genome sequencing and analysis

Identification of mutations accumulated during the passaging of EVOL20 strain were determined by sequencing parent strains on the Illumina MiSeq in the Genomics Core facility, Institute for Integrative Genome Biology at the University of California-Riverside. DNA was prepared from strains grown on MEYE agar followed by DNA extraction using 1000 fungal genomes modified CTAB protocol [1]; (http://1000.fungalgenomes.org/home/protocols/high-quality-genomic-dna-extraction/). Sequencing libraries were constructed with Neoprep TruSeq Nano DNA automated preparation, multiplexed on a single flow cell and 2x150 bp sequenced, a total of 3.03 M reads for the Af293 parent and 1.42 M for EVOL20. Reads were aligned to the Af293 genome with BWA mem (v 0.7.12) [2] followed by Picard (v2.6.01) to sort, assign read groups, mark duplicates (tools CleanSam, AddOrReplaceReadGroups, MarkDuplicates) followed by realignment around Indels and variant calling using GATK (v3.6-0-g89b7209) following Best Practices (RealignerTargetCreator, IndelRealign, HaplotypeCaller, GenotypeGVCFs, VariantFiltration) [3] and archived in repository (https://github.com/stajichlab/Afum_EVOL20; BioProject identifier PRJNA417720]. High quality SNPs (296) were identified using GATK VariantFiltration criteria to remove variants which met any criteria "--clusterWindowSize 10; QD < 2 ; MQ <40; QUAL<100; MQRankSum < -12.5; SOR > 4 ; FS > 60; ReadPosRankSum < -8". EVOL20 specific SNPs (32) were further identified using VCTools (v0.1.13) [4] and a Perl script (find_EVOL_only.pl) followed by testing for overlap with exons annotation with BEDtools (v2.26.0) [5] to identify 3 candidate non-synonymous mutations in genes Afu5g08390, Afu5g13710, and Afu5g14900.

Genome assembly of strains

Public sequence data of *A. fumigatus* strains were downloaded from NCBI Sequence Read Archive by querying for genomic datasets containing *A. fumigatus*. Scripts and download procedures available https://github.com/stajichlab/Afum_popgenome. Assemblies were generated with AAFTF (http://github.com/stajichlab/AAFTF) which quality and contaminant filters reads with BBmap (v 38.16; https://sourceforge.net/projects/bbmap/), assembles with SPAdes[6] (v 3.12.1), screens for vector sequence with

BLASTN[7] (v 2.7.1+), and polishes assemblies with Pilon[8] (v 1.22). Strain assemblies were assessed for quality with BUSCO[9] (v 2.0) using the eurotiomycetes_odb9 database of 3,959 conserved loci. Assemblies of strains which had > 90% BUSCO scores were considered high quality and retained for *hrmA* cluster analysis and population genomics analyses.

hrmA Cluster identification

To identify the presence of hrmA, cgnA and cgnA-like copies we searched the assembled strain genomes with TFASTX[10] with hrmA (Afu5g14900) and cgnA (Afu5g14910) protein sequences, followed by alignment polishing with exonerate[10]. Pipeline scripts, resulting alignments and reports are archived at https://github.com/stajichlab/Afum_hrmA_cluster_evolution .

Phylogenetic tree of strains

Phylogenetic tree of *A. fumigatus* strains was constructed by first identifying Single Nucleotide Variants among strains using the reference genome *A. fumigatus* strain Af293 from FungiDB version 39 release. Variant calling was performing using GATK by first aligning sequence reads to genome with BWA[11] (v 0.7.17-r1188) followed by MarkDuplicates tool in Picard (v 2.18.3; http://broadinstitute.github.io/picard/), and Indel Realigner tool in GATK[3,12] (v 3.7-0-gcfedb67) to produce a BAM file. The BAM files were stored in CRAM format to save space. When multiple sequence runs were produced for the same strain and BioSample the alignment files were combined into a single file for the strain. BAM files were generated on the fly from the CRAM files during the genotyping pipeline in order to minimize disk space requirements. Genotyping proceeded by running GATK HaplotypeCaller (v3.8-0-ge9d806836) with parameters "-ERC GVCF -ploidy 1" on each strain BAM file to produce a .gvcf file for each strain. The joint VCF was generated from these generated gvcf files was computed using the GATK tool GenotypeGVCFs. This VCF file was processed GATK SelectVariants to separate genotypes into two files one for biallelic SNP calls and one for INDEL, MIXED, and MNPs. These files were each further filtered with parameters to filter SNPs as "MQ < 40.0", "QUAL < 100", "MQRankSum < -12.5", "SOR > 3.0", "FS > 60.0", "ReadPosRankSum < -8.0" and "--missing-values-evaluate-as-failing --cluster-window-size 10". The

INDELS were further filtered with parameters "MQRankSum < -12.5", "SOR > 4.0", "FS > 200.0", "InbreedingCoeff < -0.8", "ReadPosRankSum < -20.0". The variants which had a value of PASS for all these filters were kept using the GATK SelectVariants tool.

An alignment was generated from genotypes extracted from the VCF file with bcftools (v 1.9; https://samtools.github.io/bcftools/), selecting only SNPs which did not overlap genomic regions identified as repetitive as masked by RepeatMasker (v 4-0-7; Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-4.0. 2013-2015 <http://www.repeatmasker.org>) implemented in pipeline steps in https://github.com/stajichlab/Afum_popgenome. A phylogenetic tree was constructed from the alignment with the ultrafast bootstrap[13] using IQ-TREE[14] with the parameters "-alrt 1000 -bb 1000" to find best tree and model fit from the sequences. Strain tree and gene cluster status was rendered with iTOL[15].

Supplementary References

1 Murray, M. G. & Thompson, W. F. Rapid isolation of high molecular weight plant DNA. *Nucleic Acids Res* **8**, 4321-4325 (1980).

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

3 Van der Auwera, G. A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* **43**, 11 10 11-33, doi:10.1002/0471250953.bi1110s43 (2013).

4 Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156-2158, doi:10.1093/bioinformatics/btr330 (2011).

5 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842, doi:10.1093/bioinformatics/btq033 (2010).

6 Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* **19**, 455-477, doi:10.1089/cmb.2012.0021 (2012).

7 Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-3402, doi:10.1093/nar/25.17.3389 (1997).

8 Walker, B. J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963, doi:10.1371/journal.pone.0112963 (2014).

9 Seppey, M., Manni, M. & Zdobnov, E. M. BUSCO: Assessing Genome Assembly and Annotation Completeness. *Methods Mol Biol* **1962**, 227-245, doi:10.1007/978-1-4939-9173-0_14 (2019).

10 Pearson, W. R., Wood, T., Zhang, Z. & Miller, W. Comparison of DNA sequences with protein sequences. *Genomics* **46**, 24-36, doi:10.1006/geno.1997.4995 (1997).

11 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

12 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303, doi:10.1101/gr.107524.110 (2010).

13 Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol Biol Evol* **35**, 518-522, doi:10.1093/molbev/msx281 (2018).

14 Nguyen, L. T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* **32**, 268-274, doi:10.1093/molbev/msu300 (2015).

15 Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res*, doi:10.1093/nar/gkz239 (2019).