















Genomic insights into the host specific adaptation of the *Pneumocystis* genus

Ousmane H. Cissé^{1,17}[✉], Liang Ma^{1,17}[✉], John P. Dekker^{2,3}, Pavel P. Khil^{2,3}, Jung-Ho Youn³, Jason M. Brenchley⁴, Robert Blair⁵⁵, Bapi Pahar⁵⁵, Magali Chabé⁶⁶, Koen K. A. Van Rompay⁷⁷, Rebekah Keesler⁷, Antti Sukura⁸, Vanessa Hirsch⁹⁹, Geetha Kutty¹, Yueqin Liu¹, Li Peng¹⁰, Jie Chen¹⁰¹⁰, Jun Song¹¹¹¹, Christiane Weissenbacher-Lang¹², Jie Xu¹¹¹¹, Nathan S. Upham¹³, Jason E. Stajich¹⁴¹⁴, Christina A. Cuomo¹⁵¹⁵, Melanie T. Cushion¹⁶¹⁶ & Joseph A. Kovacs¹⁶[✉]

Pneumocystis jirovecii, the fungal agent of human *Pneumocystis* pneumonia, is closely related to macaque *Pneumocystis*. Little is known about other *Pneumocystis* species in distantly related mammals, none of which are capable of establishing infection in humans. The molecular basis of host specificity in *Pneumocystis* remains unknown as experiments are limited due to an inability to culture any species in vitro. To explore *Pneumocystis* evolutionary adaptations, we have sequenced the genomes of species infecting macaques, rabbits, dogs and rats and compared them to available genomes of species infecting humans, mice and rats. Complete whole genome sequence data enables analysis and robust phylogeny, identification of important genetic features of the host adaptation, and estimation of speciation timing relative to the rise of their mammalian hosts. Our data reveals insights into the evolution of *P. jirovecii*, the sole member of the genus able to infect humans.

¹ Critical Care Medicine Department, NIH Clinical Center, National Institutes of Health (NIH), Bethesda, MD, USA. ² Bacterial Pathogenesis and Antimicrobial Resistance Unit, National Institute of Allergy and Infectious Diseases (NIAID), NIH, Bethesda, MD, USA. ³ Department of Laboratory Medicine, NIH Clinical Center, National Institutes of Health, Bethesda, MD, USA. ⁴ Laboratory of Viral Diseases, NIAID, NIH, Bethesda, MD, USA. ⁵ Tulane National Primate Research Center, Tulane University, New Orleans, LA, USA. ⁶ Univ. Lille, CNRS, Inserm, CHU Lille, Institut Pasteur de Lille, U1019-UMR 9017-CIIL-Centre d'Infection et d'Immunité de Lille, Lille, France. ⁷ California National Primate Research Center, University of California, Davis, CA, USA. ⁸ Department of Veterinary Pathology, Faculty of Veterinary Medicine, University of Helsinki, Helsinki, Finland. ⁹ Laboratory of Molecular Microbiology, NIAID, NIH, Bethesda, MD, USA. ¹⁰ Department of Respiratory and Critical Care Medicine, the First Affiliated Hospital of Chongqing Medical University, Chongqing, China. ¹¹ Center for Advanced Models for Translational Sciences and Therapeutics, University of Michigan Medical Center, University of Michigan Medical School, Ann Arbor, MI, USA. ¹² Institute of Pathology, Department of Pathobiology, University of Veterinary Medicine Vienna, Vienna, Austria. ¹³ Arizona State University, School of Life Sciences, Tempe, AZ, USA. ¹⁴ Department of Microbiology and Plant Pathology and Institute for Integrative Genome Biology, University of California, Riverside, Riverside-California, Riverside, CA, USA. ¹⁵ Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, MA, USA. ¹⁶ Department of Internal Medicine, College of Medicine, University of Cincinnati, Cincinnati, OH, USA. ¹⁷ These authors contributed equally: Ousmane H. Cissé, Liang Ma. ✉email: ousmane.cisse@nih.gov; mal3@nih.gov; jkovacs@nih.gov

The evolutionary history of *Pneumocystis jirovecii*, a fungus that causes life-threatening pneumonia in immunosuppressed patients such as those with HIV infection, has been poorly defined. *P. jirovecii* is derived from a much broader group of host-specific parasites that infect all mammals studied to date. Until recently, *P. carinii* and *P. murina* (which infect rats and mice, respectively) were the only other species in this genus for which biological specimens suitable for whole-genome sequencing were readily available. Cross-species inoculation studies of *P. jirovecii* and *P. carinii* have found that they can only infect humans and rats, respectively^{1,2}. Further, rats are the only mammals known to be coinfecting by at least two distinct *Pneumocystis* species (*P. carinii* and *P. wakefieldiae*)³. Within the *Pneumocystis* genus, *P. jirovecii* is the only species able to infect and reproduce in humans, although the molecular mechanisms of its host adaptation remain elusive.

Previous efforts to reconstruct the evolutionary history of *Pneumocystis* have estimated the origins of the genus at a minimum of 100 million years ago (mya)⁴. Using a partial transcriptome of *P. sp. macacae* (hereafter referred to as *P. macacae*), the *Pneumocystis* species that infects macaques, we recently estimated that *P. jirovecii* diverged from the common ancestor of *P. macacae* around ~62 mya⁵, which substantially precedes the human-macaque split of ~20 mya⁶. Population bottlenecks in *P. jirovecii* and *P. carinii* at 400,000 and 16,000 years ago, respectively⁵, are also not concordant with population expansions in modern humans (~200,000 years ago⁷) and rats (~10,000 years ago⁸), which suggests a decoupled coevolution between *Pneumocystis* and their hosts. Thus, *Pneumocystis* species may not be strictly coevolving with their mammalian hosts as suggested by ribosomal RNA-based maximum phylogenies⁹. A molecular clock has not been tested in any of these phylogenies. A strict coevolution hypothesis has been further challenged by evidence suggesting a relaxation of the host specificity in *Pneumocystis* infecting rodents^{10,11}. However, the accuracy of speciation times is limited without the complete genomes of additional species including that of *P. macacae*, the closest living sister species to *P. jirovecii* identified to date.

The absence of long-term in vitro culture methods or animal models for most *Pneumocystis* species has precluded obtaining sufficient DNA for full genome sequencing and has hindered investigation of the *Pneumocystis* genus. To date, only the genomes of human *P. jirovecii*^{12,13}, rat *P. carinii*^{13,14}, and mouse *P. murina*¹³, are available. These data have provided important insights into the evolution of this genus, including a substantial genome reduction^{12,13}, the presence of intron-rich genes possibly contributing to transcriptome complexity, and an expansion of a highly polymorphic major surface glycoprotein (*msg*) gene superfamily¹³, some of which are important for immune evasion. However, the lack of whole-genome sequences for many species of this genus (particularly the closely related *P. macacae*) has severely constrained the understanding of the implications of these genome features in *Pneumocystis* evolution and adaptation to hosts.

To explore the evolutionary history of the *Pneumocystis* genus, and investigate *P. jirovecii* genetic factors that support its adaptation to humans, we sequenced 2–6 specimens of four additional species: those that infect macaques (*P. macacae*), rabbits (*P. oryctolagi*), dogs (*P. canis*), and rats (*P. wakefieldiae*).

Results

Direct sequencing of *Pneumocystis*-host mixed samples. We sequenced the genomes of *Pneumocystis* species from infected macaques, rabbits, dogs, and rats (see Methods and Supplementary Methods). Specimens originated from immunosuppressed

animals as a consequence of simian immunodeficiency virus infection in macaques, corticosteroid treatment (rabbits and rats), immunodeficient knockout (rabbits) or possible congenital immunodeficiencies (dogs). For each species, we sequenced multiple samples from 2–6 animals (Supplementary Tables 1 and 2). These data were used to assemble one nearly full-length genome assembly for each species except *P. canis* for which we recovered two nearly full-length assemblies and an additional partial assembly from two separate samples (denoted as A, Ck1, and Ck2). Post assembly mapping revealed a negligible amount of genetic variability among samples, for example the average genome-wide single-nucleotide polymorphism (SNP) diversity among six *P. macacae* isolates excluding highly polymorphic regions such as *Msg* genes is ~0.1%. The genome of *P. macacae* was sequenced using Oxford Nanopore long reads and Illumina short read sequences, whereas the other *Pneumocystis* were sequenced only with Illumina (Supplementary Tables 2 and 3). The genome assemblies range from 7.3 Mb in *P. wakefieldiae* to 8.2 Mb in *P. macacae*. The *P. macacae* and *P. wakefieldiae* genome assemblies consist of 16 and 17 scaffolds, respectively, both of which are highly contiguous and approach the chromosomal level based on similarities with published karyotypes^{3,15} and/or the presence of *Pneumocystis* telomere repeats¹⁶ at the scaffold ends (Supplementary Table 3). The genome assemblies of *P. oryctolagi* and *P. canis* (assemblies A, Ck1, and Ck2) are less contiguous with 38, 33, 78, and 315 scaffolds, respectively. All these assemblies except for the partial assembly of *P. canis* Ck2 have very similar total sizes (7.3–8.2 Mb) comparable to previously sequenced genomes of *P. jirovecii*, *P. carinii*, and *P. murina*, all of which are at or near chromosomal level with a size of 7.4–8.3 Mb (Supplementary Table 3). The genome assemblies are all AT-rich (~71%) and ~3% encode DNA transposons and retrotransposons (Supplementary Table 3). We also assembled complete mitochondrial genomes from all species in this study, which are similar in size (21.2–24.5 kilobases) to published rodent *Pneumocystis* mitogenomes (24.6–26.1 kb)¹⁷ but smaller than that of *P. jirovecii* (~35 kb)¹⁷ (Supplementary Table 3). *P. macacae* has a circular mitogenome similar to *P. jirovecii*¹⁷ whereas all other sequenced species have linear mitogenomes.

Genomic differences among *Pneumocystis* species. To assess the extent of genome structure variations among species, we generated whole-genome alignment of all representative genome assemblies. We found high levels of interspecies rearrangements ranging from 10 breakpoints between *P. wakefieldiae* and *P. murina* to 142 between *P. jirovecii* and *P. oryctolagi* (Fig. 1; Supplementary Table 4). The vast majority of chromosomal rearrangements were inversions, which, for example accounted for 23 out of 29 breakpoints between *P. jirovecii* and *P. macacae* (Supplementary Table 4). Analysis of aligned raw Nanopore and/or Illumina reads back to the assemblies show no evidence of incorrect contig joins around rearrangement breakpoints. There are clearly fewer rearrangements among rodent *Pneumocystis* species (*P. wakefieldiae*, *P. carinii*, and *P. murina*) than among all other species (Fig. 1; Supplementary Table 4), which is likely due to the younger evolutionary ages of rodent *Pneumocystis* (Fig. 2a and c). These rearrangements could have caused incompatibilities between species, thus preventing gene flow for species that infect the same host.

Comparison of pairwise whole-genome alignment identities between species indicates a substantial nucleotide divergence: 14% dissimilarity in aligned regions between *P. jirovecii* and *P. macacae*; 21% between *P. jirovecii* and *P. oryctolagi*; 22% between *P. jirovecii* and *P. canis* Ck1; 15% between *P. wakefieldiae*

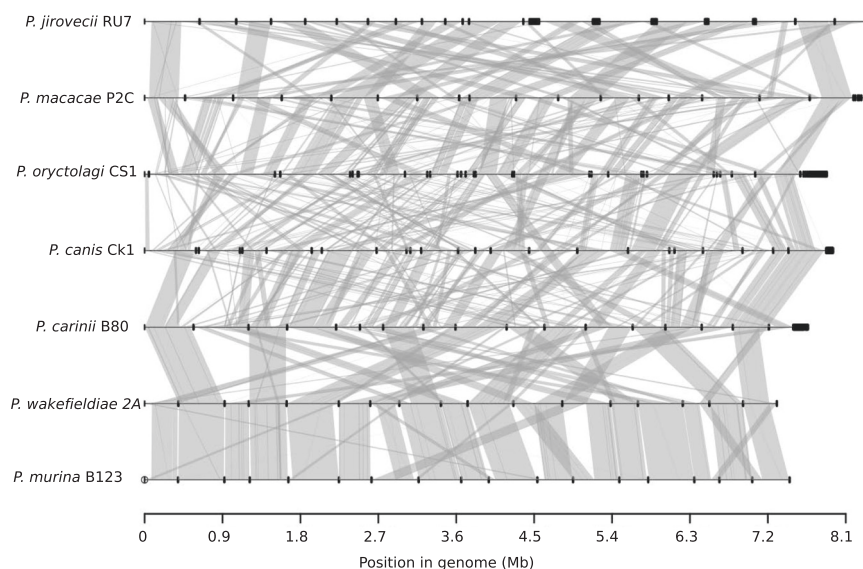


Fig. 1 Whole-genome structure and synteny among *Pneumocystis* species. Species names and their genome assembly identifiers are shown on the left. Horizontal black lines on the right represent sequences of all scaffolds for each genome laid end-to-end, with their nucleotide positions indicated at the bottom. Dark thick squares represent short scaffolds. Syntenic regions between genomes are linked with vertical gray lines.

and *P. carinii*; and 12% between *P. wakefieldiae* and *P. murina* (Supplementary Table 5).

To understand the relationship between the intraspecies and interspecies genetic diversity of *Pneumocystis*, we generated additional four *P. jirovecii* and five *P. macacae* genome assemblies from low sequence coverage samples. All data are expressed as the mean \pm standard deviation. The pairwise intraspecies genome divergences among *P. jirovecii* genome assemblies ($0.3 \pm 0.2\%$, $n = 8$) are significantly lower than those obtained when comparing them to *P. macacae* assemblies ($16.1 \pm 0.2\%$, $n = 5$) (two sample *t*-test, p -value = 1.4×10^{-14}) or to other *Pneumocystis* species ($21.6 \pm 0.9\%$, $n = 7$) ($p = 2.9 \times 10^{-10}$). Similarly, mean divergence among *P. macacae* genome assemblies ($0.8 \pm 0.3\%$, $n = 5$) is lower than divergence when they are compared to *P. jirovecii* assemblies ($15.6 \pm 0.2\%$, $n = 8$) ($p = 1.3 \times 10^{-10}$) or other *Pneumocystis* species genome assemblies ($21.8 \pm 0.9\%$) ($p = 7.2 \times 10^{-11}$). The results indicate that interspecies divergence exceeds intraspecies divergence, which is consistent with a complete species separation.

Speciation history of the *Pneumocystis* genus. These new complete genome data enabled us to examine the relationships between different *Pneumocystis* species and to estimate the timing of speciation events that led to the extant species. We inferred a strongly supported phylogeny of *Pneumocystis* species rooted with outgroups from distantly related fungal subphyla. Our phylogenomic analysis of 106 single-copy orthologs inferred from all assemblies including the fragmented Ck2 strongly supports monophyly of *Pneumocystis* species (100% Maximum likelihood bootstrap values; Fig. 2a), Bayesian posterior probabilities (>0.95 ; Supplementary Fig. 1), and highly significant support from the Shimodaira–Hasegawa test¹⁸ ($p < 0.001$; see Methods). An identical phylogeny was recovered using mitochondrial genome data from 33 specimens representing 7 *Pneumocystis* (Supplementary Fig. 2). However, we identified unexpected placements of *P. wakefieldiae*, *P. oryctolagi*, and *P. canis*. First, *P. wakefieldiae* appears as a sister species of *P. murina* instead of *P. carinii* (which also infects rats) (Fig. 2b). This observation is supported by the higher similarity in genome size (Supplementary Table 3), lower sequence divergence (Supplementary Table 4), higher

genome synteny (Fig. 1; Supplementary Table 5) and higher frequencies of supporting genes (0.64 in 1,718 nuclear gene trees examined; Methods) between *P. wakefieldiae* and *P. murina* than between *P. wakefieldiae* and *P. carinii*. These relationships contradict the previous phylogenetic placement of *P. wakefieldiae* as an outgroup of the *P. carinii*/*P. murina* clade⁹ or a sister species of *P. carinii*¹⁹ based on analysis of mitochondrial large and small subunit rRNA genes (mtLSU and mtSSU). Our phylogeny also opposes the prevailing hypothesis for dynamics of host specificity and coevolution within the *Pneumocystis* genus, that is, *P. wakefieldiae* shares with *P. carinii* the same host species (*Rattus norvegicus*) and thus is expected to be more related to *P. carinii* than to *P. murina*.

Similarly, *P. oryctolagi* would be expected to be phylogenetically closer to rodent *Pneumocystis* than to primate *Pneumocystis*, consistent with the closer phylogenetic relationships of rabbits and rodents to each other than to primates²⁰ (Fig. 2a, b). In contrast, *P. oryctolagi* and *P. canis* are more closely related to primate *Pneumocystis* (*P. jirovecii* and *P. macacae*) than rodent *Pneumocystis* (Fig. 2a; Supplementary Figs. 1 and 2; 100% of tree level support in 1718 nuclear genes). The phylogenetic discrepancy between *P. oryctolagi* and its host (rabbit) suggests that host switching may have occurred in their distant history.

From whole-genome Bayesian phylogenetic estimates (see Methods), the common ancestor of all extant species of the genus emerged around 140 mya (confidence intervals: 180–101 mya; Fig. 2c; Supplementary Fig. 1), with a separation of *Pneumocystis* and *Schizosaccharomyces* genera around 512 mya (CI: 822–203 mya), which is consistent with independent estimates of the origin of Taphrinomycota crown group at 530 mya²¹. The *Pneumocystis* genus thereafter divided into two main clades, P1 consisting of *P. jirovecii*, *P. macacae*, *P. oryctolagi*, and *P. canis*, and P2 consisting of species infecting rodents (*P. carinii*, *P. wakefieldiae*, and *P. murina*) (Fig. 2b). Subsequent to the divergence of P1/P2, the clade P1 diversified through a series of speciation events leading either to primate or carnivore species whereas P2 remained localized in rodents. We also found that the divergence time of *Pneumocystis* in the clade P1 predates that of their hosts, that is, the crown of rodent-rabbit-primate

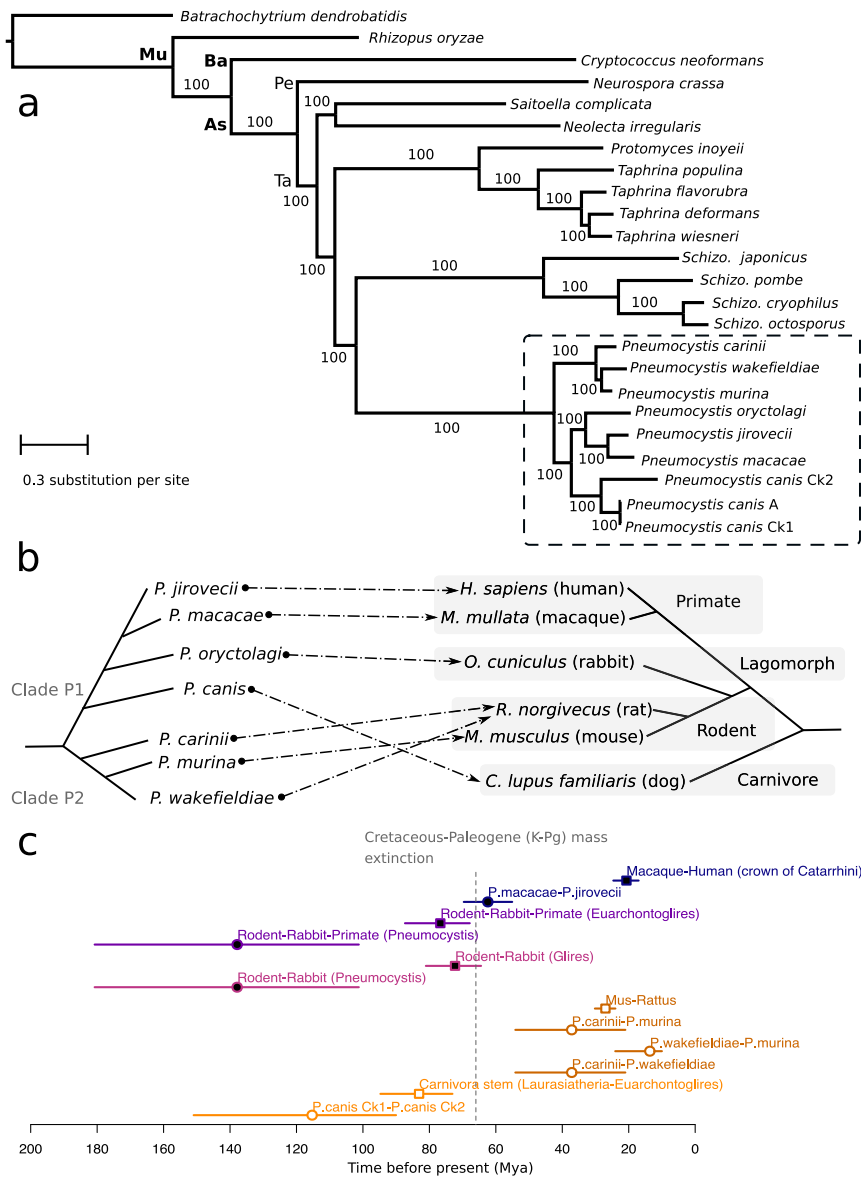


Fig. 2 Phylogeny and divergence times of *Pneumocystis* species. **a** Maximum likelihood phylogeny constructed using 106 single-copy genes based on 1000 replicates from 24 annotated fungal genome assemblies including nine from *Pneumocystis* (highlighted with a dashed box). Only one assembly is shown for each species except there are three for *P. canis* (assemblies Ck1, Ck2, and A). Bootstrap support (%) is presented on the branches. The fungal major phylogenetic phyla and subphyla are represented by their initials: As (Ascomycota), Ba (Basidiomycota), Pe (Pezizomycotina), Mu (Mucoromycota), and Ta (Taphrinomycotina). **b** Schematic representation of species phylogeny and association between *Pneumocystis* species and their respective mammalian hosts. The dashed arrows represent the specific parasite-host relationships. **c** Divergence times of *Pneumocystis* species and mammals ($n = 12$ taxonomic clades analyzed). Divergence time medians are represented as squares for hosts and as circles for *Pneumocystis*, and the horizontal lines represent the 95% confidence interval (CI) error bars, which are color-coded the same for each *Pneumocystis* and its host. Closed elements represent nodes that are different in term of divergence times (nonoverlapping confidence intervals) whereas open elements represent nodes with overlapping confidence intervals. Catarrhini, taxonomic category (parvorder) including humans, great apes, gibbons, and Old-World monkeys. Euarchontoglires, superorder of mammals including rodents, lagomorphs, treeshrews, colugos, and primates. Glires, taxonomic clade consisting of rodents and lagomorphs. Laurasiatheria, taxonomic clade of placental mammals that includes shrews, whales, bats, and carnivorans. Mya, million years ago. K-Pg, Cretaceous-Paleogene. The dotted vertical line representing the K-Pg mass extinction event at 66 mya is included for context only.

Pneumocystis is clearly more ancient than the corresponding superorder of mammals (Euarchontoglires) (Fig. 2c). The pattern in clade P2 is different as the divergence time estimates overlap with those of their hosts (Fig. 2c). On the basis of coalescent estimates, *P. jirovecii* began to split from *P. macacae* at ~62 mya (CI: 69–55 mya), which extended through the Cretaceous-Paleogene mass extinction event at 66 mya, but substantially

predates the crown Catarrhini (human-macaque ancestor) of ~20 mya (CI: 24–17 mya; Fig. 2c; Supplementary Fig. 1).

High levels of population differentiation among *Pneumocystis* genomes support reproductive isolation. To understand the genomic divergence landscape of *Pneumocystis* populations, we performed genome-wide differentiation tests (F_{ST} , relative

population divergence) and nucleotide diversity (π). These analyses used 32 genomic datasets, including 26 publicly available datasets in GenBank for *P. jirovecii*, *P. carinii* and *P. murina* and six datasets generated in this study for other four *Pneumocystis* species (Supplementary Note 1; Supplementary Table 2). We used a trained version LAST²² to account for interspecies divergence during read mapping and ANGSD²³ to derive genotype likelihoods instead of genotypes. Since ANGSD's F_{ST} requires outgroups, we analyzed interspecies divergence between *P. jirovecii*, *P. macacae*, and *P. oryctolagi* populations using a sliding window approach of 5-kb and *P. carinii* as an outgroup species (n samples = 59). *P. murina* genomic divergence relative to *P. carinii* and *P. wakefieldiae* populations was estimated similarly using *P. jirovecii* as an outgroup species ($n = 47$). We found high levels of population differentiation among *Pneumocystis* specimens; 71.9% of the *P. jirovecii* genome had a Fixation index (F_{ST}) > 0.8 compared to the closest species, *P. macacae*, while 90.2% of the genome had a F_{ST} > 0.8 compared to the extant species *P. oryctolagi* (Supplementary Fig. 3). Similarly, 86.3% and 93.7% of the *P. murina* genome had a F_{ST} > 0.8 compared to *P. carinii* and *P. wakefieldiae*, respectively (Supplementary Note 1).

Analyzing historical hybridization in *Pneumocystis* genus.

Topology-based maximum likelihood analysis of 1718 gene trees using PhyloNet²⁴ found no evidence of gene flow among species of clade P1 (*P. jirovecii*, *P. macacae*, *P. oryctolagi*, and *P. canis*) (Supplementary Fig. 4), which indicates that these lineages were reproductively isolated throughout their evolutionary history, consistent with their isoenzyme diversity²⁵. In contrast, we found strong evidence of ancient hybridization in clade P2, possibly between *P. carinii* and *P. wakefieldiae* (Supplementary Fig. 4), which may then have contributed to the formation of the *P. murina* lineage. We hypothesize that *P. murina* might have originated as a hybrid between ancestors of *P. carinii* and *P. wakefieldiae* in rats, and subsequently shifted to mice, possibly owing to the geographic proximity of ancestral rodent populations (for example in Southern Asia²⁶), which is consistent with the fact that ecological fitting is a major determinant of host switch²⁷. The presumed physiological, cellular and/or immunological similarities among closely related rodent species might also have helped the same *Pneumocystis* species colonizing multiple closely related rodent species^{10,27}.

Gene families and metabolic pathways linked to host specificity.

The predicted protein-coding gene numbers are similar across *Pneumocystis* genomes and range from 3211 in *P. wakefieldiae* to 3476 in *P. canis* strain Ck1 (Supplementary Table 3). Nearly all predicted protein-coding genes in *P. macacae* (96% of 3471) and *P. wakefieldiae* (99% of 3221) genomes have RNA-Seq support. Gene models present a complex architecture with ~6 exons per gene on average. High representation of core eukaryotic genes in *P. macacae*, *P. oryctolagi*, *P. canis* and *P. wakefieldiae* provides evidence that these genomes are nearly complete and comparable in completeness to *P. jirovecii*, *P. murina*, and *P. carinii* genomes: 86.2–93.4% of conserved genes are detectable in all annotated genome assemblies (Supplementary Table 3).

Examination of orthologous genes reveals that ~3100 orthologous clusters had representative genes from all nine analyzed genome assemblies from seven *Pneumocystis* species (Supplementary Table 3). We found a small number of unique genes in each *Pneumocystis* species ranging from 25 in *P. wakefieldiae* to 204 in *P. oryctolagi* (Supplementary Table 3). Unique genes in most species encode for phylogenetically unrelated proteins with unknown function. A striking exception is observed in *P. macacae* in which nearly all unique proteins are part of an

undescribed large protein family ($n = 190$). The members of this family are enriched in arginine and glycine amino-acid residues (denoted RG proteins) (Supplementary Fig. 5a) and have no similarities with transposable elements. While RG motifs are often found in eukaryotic RNA-binding proteins²⁸, *P. macacae* RGs do not possess an RNA-binding domain (Pfam domains PF00076, PF08675, PF05670, PF00035), suggesting a different role. In addition, *P. macacae* RGs lack functional annotation except for two proteins that encode a Dolichol-phosphate mannosyltransferase domain (PF08285) and a leucine zipper domain (PF10259), respectively. Of the 190 RGs, 134 have RNA-Seq based gene expression support, including five among the top highly expressed genes (Supplementary Fig. 5b). Nearly half of RGs are located at subtelomeric regions, often found in close proximity to *msg* genes (Supplementary Data 1). RG proteins can be grouped in three main clusters (based on OrthoFinder clustering; Methods), have a reticulate phylogeny (Supplementary Fig. 5c) and a mosaic gene structure (Supplementary Fig. 5d), which suggest frequent gene conversion events.

To investigate the gene loss patterns in sequenced genomes, we compared *Pneumocystis* gene catalogs to those of related Taphrinomycotina fungi. We found that all sequenced *Pneumocystis* species have lost ~40% of gene families present in other Taphrinomycotina (Supplementary Fig. 6), and that the metabolic pathways are also very similar among *Pneumocystis* species with a few minor (possibly stochastic) differences (Supplementary Note 2). This strongly suggests that *Pneumocystis* ancestry experienced massive gene losses that occurred before the genus diversification.

To investigate changes in gene content that might explain interspecies differences among the seven *Pneumocystis* species, we searched for expansions or contractions in functionally classified gene sets. We identified Pfam domains with significantly uneven distribution among genomes (Wilcoxon signed-rank test $p < 0.05$). Domains associated with *Msg* proteins are enriched in *P. jirovecii* and, to a lesser extent in *P. macacae* compared to other species (Fig. 3a). Domains associated with peptidases (M16) are enriched in *P. carinii*, *P. murina*, and *P. wakefieldiae*. S8 peptidase family (kexin) is expanded in *P. carinii* with 13 copies¹³ whereas all other species have one or three copies (Fig. 3a; Supplementary Fig. 7). Although kexin is localized in other fungi to the Golgi apparatus, and in *Pneumocystis* is believed to be involved in the processing of *Msg* proteins, the expanded copies are predicted to be GPI-anchored proteins, and appear to localize to the cell surface; their function is unknown²⁹. We found that *P. carinii* kexin genes evolved under strong positive selection ($p = 0.008$) whereas *P. wakefieldiae* kexin genes did not ($p = 0.159$).

Phylogenetic analysis of CFEM (common in fungal extracellular membrane) protein domains, which are important for the acquisition of vital compounds in fungi³⁰, suggest that these domains were likely already present in the last common ancestor of *Pneumocystis* and were vertically transmitted (Supplementary Fig. 8).

To investigate changes in enzyme gene content that might account for interspecies differences among *Pneumocystis* species, we searched for enzymes that show clear differences among species, which are represented by Enzyme Commission numbers (ECs) (Fig. 3b). We found 34 ECs, which include 14 that are highly conserved in *P. jirovecii* but have a patchy distribution in other members of clade P1 (*P. macacae*, *P. canis*, and *P. oryctolagi*) and are lost in clade P2 (*P. carinii*, *P. murina*, and *P. wakefieldiae*). Most of these 14 ECs are assigned to the biosynthesis of antibiotics or secondary metabolites and vitamin B6 metabolism according to KEGG pathways. The latter pathway seems only functional in P2 clade (Supplementary Note 2).

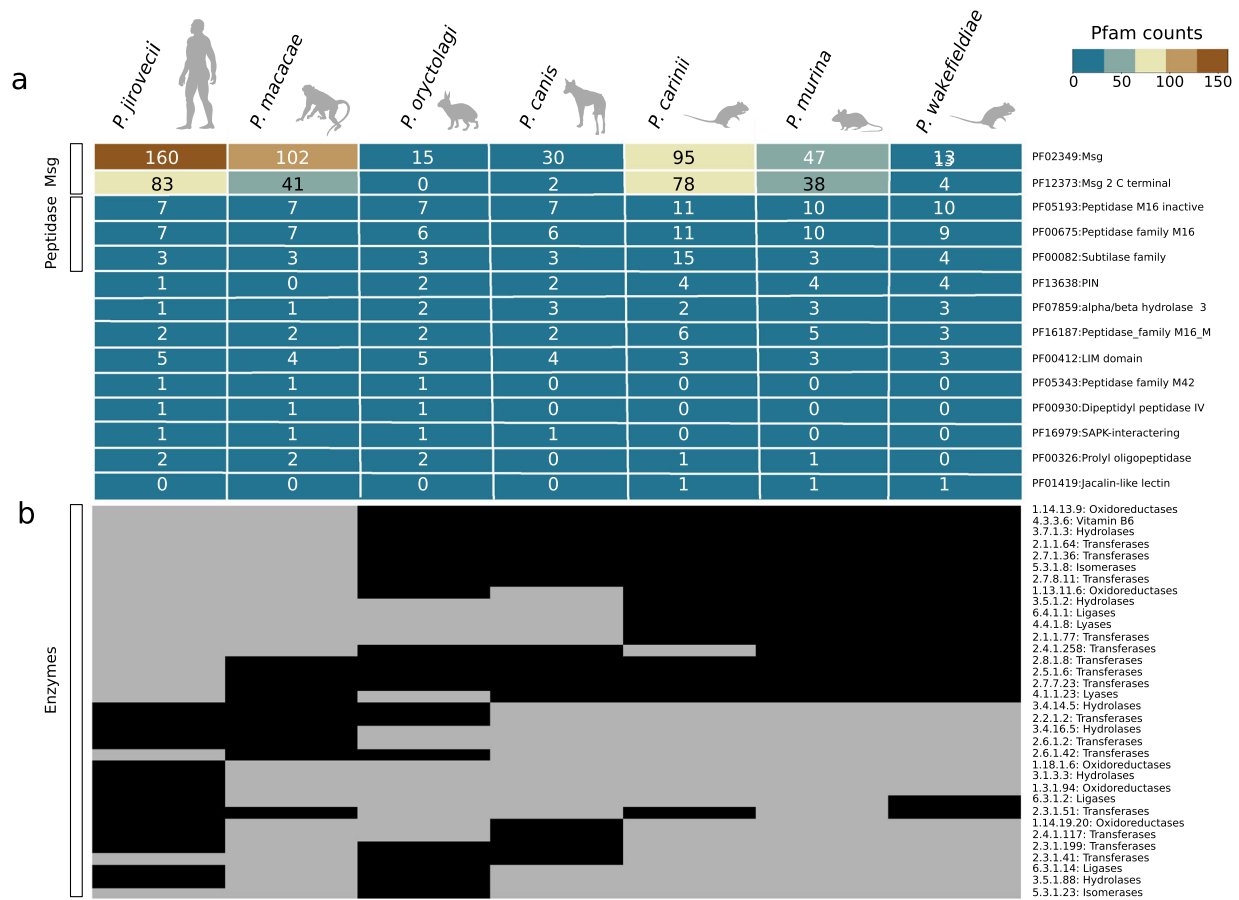


Fig. 3 Distribution of protein families among *Pneumocystis* species. **a** Heatmap of Pfam protein domains with significant differences (Wilcoxon signed-rank test, $p < 0.05$) are included if the domain appears at least once in the following comparisons: primate *Pneumocystis* (*P. jirovecii* and *P. macacae*) versus other *Pneumocystis*, clade P1 (*P. jirovecii*, *P. macacae*, *P. oryctolagi*, and *P. canis* Ck1) versus clade P2 (*P. carinii*, *P. murina*, and *P. wakefieldiae*), primate *Pneumocystis* versus clade P2. The number of proteins containing each domain is indicated within each cell for each species. The heatmap is colored according to a score, as indicated by the key at the upper right corner. **b** Heatmap of distribution of enzymes (represented by Enzyme Commission numbers and their KEGG functional categories), with their presence and absence indicated by black and grey colored cells, respectively. Animal icons were obtained from <http://phylopic.org> under creative commons licenses <https://creativecommons.org/licenses/by/3.0/>: mouse (Anthony Caravaggi; license CC BY-NC-SA 3.0); dog (Sam Fraser-Smith and vectorized by T. Michael Keeseey; license CC BY 3.0); rabbit (by Anthony Caravaggi; license CC BY-NC-SA 3.0), and rat (by Rebecca Groom; license CC BY-NC-SA 3.0). Icons original black color background were modified to light gray.

Intron evolution. We analyzed 1211 one-to-one gene orthologs shared by all sequenced *Pneumocystis* and other Taphrinomycotina fungi (Supplementary Fig. 9a). A total of 9080 homologous sites within 1211 alignments were identified (Supplementary Fig. 9b). While intron densities are similar among *Pneumocystis* species (ranging from 4842 in *P. macacae* to 5289 in *P. murina*), they are markedly more elevated compared to related Taphrinomycotina, including *Neolecta irregularis* ($n = 4202$ introns), *Schizosaccharomyces pombe* ($n = 862$), and *Taphrina deformans* ($n = 639$) (Supplementary Fig. 11b). Predictions of ancestral intron densities show that the *Pneumocystis* common ancestor had at least 5341 introns, of which 37% were not found in other Taphrinomycotina (Supplementary Fig. 9c). This is in contrast to other fungi; ~26% of *Neolecta* introns were independently acquired whereas *S. pombe* and *T. deformans* genomes have experienced intron losses, which is consistent with published studies^{31,32}. These results suggest that the emergence of *Pneumocystis* genus was preceded by gain of introns.

Positive selection footprints in *P. jirovecii* genes. We tested the hypothesis that *P. jirovecii* has adapted specifically to humans after its separation with *P. macacae*, and that there will be footprints of

directional selection in the genome that point to the molecular mechanisms of this adaptation. To infer *P. jirovecii*-specific adaptive changes, we compared the *P. jirovecii* one-to-one orthologs to those of *P. macacae* and *P. oryctolagi* using the branch-site likelihood ratio test³³. Positive selection was identified as an accelerated nonsynonymous substitution rate. The test identified 244 genes (out of 2466) with a signature of positive selection in the human pathogen *P. jirovecii* alone (Bonferroni corrected p -value < 0.05 ; Supplementary Data 2). Gene Ontology enrichment analysis of these genes with accelerate rates identified significant enrichment for the biological process “cellular response to stress” (adjusted using Benjamini–Hochberg p -value = 1.9×10^{-6}) and the molecular function “potassium channel regulator activity” ($p = 2.8 \times 10^{-10}$). Among the 244 genes, 197 are conserved in all *Pneumocystis* genomes available whereas 47 are absent in clade P2 only (*P. carinii*, *P. murina*, and *P. wakefieldiae*; Fig. 2b). While the latter set of genes encode proteins of unknown function, analysis of Pfam domains shows a significant enrichment in the biological process “nucleoside phosphate biosynthetic” process ($p = 9.9 \times 10^{-5}$) and the molecular function “carbon–nitrogen lyase activity” ($p = 2.8 \times 10^{-10}$). Further investigations will be required to determine the precise functions of these genes.

Subtelomeric gene families. Until recently, the only in-depth data on the subtelomeric gene families in *Pneumocystis* have come from the *P. jirovecii*, *P. carinii*, and *P. murina* genomes^{13,34}. These genes, including *msg* and *kexin*, are believed to be important for antigenic variation, and are well represented in the assemblies of *P. macacae*, *P. oryctolagi*, *P. canis*, and *P. wakefieldiae*.

P. macacae subtelomeres encode numerous arrays of *Msg* and *RG* proteins (Supplementary data 1). Phylogenetic analysis of adjacent genes revealed only a few instances of recent paralogs, which suggests that most of the duplications and subsequent positional gene arrangements are ancient. Three *P. macacae* subtelomeric regions have a nearly perfect synteny in *P. jirovecii* with the only difference being the absence of *RG* proteins in *P. jirovecii* (Supplementary Data 1). *P. oryctolagi* subtelomeres tend to be enriched in orphan genes that are not members of the *Msg* superfamily, and are of unknown function. *P. canis* subtelomeres are enriched in *Msg-C* family (see *Msg* section below). *P. wakefieldiae* subtelomeres are rich in *msg* genes, though their types are distinct from those of *P. carinii* and *P. murina*.

Evolution of *msg* genes. Up to 6% of the *Pneumocystis* genomes are comprised of copies of the *msg* superfamily, which are believed to be crucial mediators of pathogenesis through antigenic variation and interaction with the host cells. The superfamily is classified into five families A, B, C, D, and E based on protein domain architecture, phylogeny and expression mode^{13,34,35}. The A family is the largest of the five, has been subdivided into three subfamilies (A1, A2, and A3) and is generally thought to contribute to antigenic variation. Their protein products contain cysteine-rich domain classified as N1 and M1 to M5.

To investigate the origin of *msg* genes, we used previously developed Hidden Markov Models¹³ to search for corresponding gene models in the assemblies of *P. macacae*, *P. oryctolagi*, *P. canis*, and *P. wakefieldiae* and combined these data with published *msg* sequences annotated in *P. jirovecii*, *P. carinii*, and *P. murina* genomes^{13,35}. Of note, in this study only a subset of *msg* genes were assembled for *P. oryctolagi*, *P. canis*, and *P. wakefieldiae* due to difficulties in assembling highly similar short reads from Illumina sequencing exclusively while a potentially complete set of *msg* genes were assembled for *P. macacae* using Illumina and Nanopore reads (Supplementary Table 3). The number of full-length *msg* genes available ranges from 9 in *P. oryctolagi* to 161 in *P. jirovecii*. Sequence-based clustering and phylogenetic analyses of all *msg* genes ($n = 482$) revealed that: (i) there is no evidence of interspecies transfer among *Pneumocystis* species (Fig. 4b to d; Supplementary Fig. 10), (ii) *msg* genes may have a polyphyletic origin, i.e., distinct families were present in most recent ancestors of *Pneumocystis* (Supplementary Fig. 10a), although convergent evolution of *msg* cannot be ruled out; (iii) *msg* genes experienced recombination early in their history as estimated by phylogenetic network analysis (Supplementary Fig. 10b and c).

While some gene expansions are relatively recent (for example, *msg* families A, C, and D) other expansions (*msg* families E and B) occurred before the emergence of *Pneumocystis* genus itself (Supplementary Fig. 11). Subsets of *msg* genes show strong host-specific sequence diversification (Fig. 4a), such as the current A family has emerged relatively recently at 43 mya (CI: 58–28 mya) compared to the emergence of the genus at 140 mya (see Methods; Supplementary Fig. 11). The A1 subfamily displays a substantial expansion in all species (Fig. 4a) and is subject to intraspecies recombination (Fig. 4b to d), which suggest that the most recent *Pneumocystis* ancestor may have developed a pre-*Msg-A* family,

which then evolved through duplication and recombination after the species separation.

The A3 subfamily has expanded only in clade P1 (especially in *P. jirovecii*) whereas A2 has expanded only in clade P2 (*P. carinii*, *P. murina* and to a lesser extent in *P. wakefieldiae*) (Fig. 4a). Although all members of the A family might have a shared deep ancestry, we found no evidence suggesting that the A1, A2, A3 subfamilies are directly derived from one another (Supplementary Fig. 10).

The *msg-B* family underwent a net independent expansion in *P. macacae* ($n = 10$) and *P. jirovecii* ($n = 12$), while being reduced to only one copy in *P. oryctolagi* and *P. canis*, and being completely absent in *P. wakefieldiae*, *P. carinii* and *P. murina* (Fig. 4a). Using Bayesian estimates, we estimated the origin of the B family to be older than the *Pneumocystis* genus itself (~211 vs. 140 mya; Supplementary Fig. 11).

The *msg-D* family is expanded only in *P. macacae* and *P. jirovecii*. The D family emerged at ~69 mya (CI: 109–40 mya) before the split of these two species (Supplementary Fig. 11), thus suggesting a role in adaptation to primates similar to the A3 subfamily. In contrast, the E family, which is conserved in all species, is much more ancient at ~311 mya (CI: 541–158 mya), again preceding the emergence of the genus (Supplementary Fig. 11).

P. jirovecii and *P. macacae* not only have a larger number of *msg*-associated cysteine-rich domains than other *Pneumocystis* species (Fig. 5a) but also a much greater sequence diversity per domain than other *Pneumocystis* species (Fig. 5c). Domain sequences cluster independently, with each cluster containing sequences from all *Pneumocystis* species (Fig. 5b). Domains M1 and M3 are more closely related to each other than other domains, which suggests a relatively recent duplication.

Discussion

Surprisingly, analysis of core genomic regions of the nuclear genomes did not identify clear differences that are suggestive of mechanisms for host-specific adaptation; instead it is the highly polymorphic multicopy gene families (*msgs*) that appear to account for this adaptation. *Msgs*, which provide *Pneumocystis* with a mechanism for antigenic variation and consequent immune evasion, may have been important in allowing *Pneumocystis* organisms to infect mammals successfully, given that an adaptive immune system, which is critical to protection of mammals from exogenous pathogens, arose ~500 mya in the first jawed vertebrates³⁶. *Msgs* likely played a dual role in avoiding the adaptive immunity and in cell adherence.

Based on our analysis, we propose the following series of events for the emergence and adaptation of *P. jirovecii* as a major human opportunistic pathogen (Fig. 6). First, there was a major shift of a pre-*Pneumocystis* lineage (possibly a soil- or plant-adapted organism) to mammals, which led to a genome reduction but with a proliferation of introns and expansions of cysteine-rich domain-containing proteins involved in immune escape and nutrient scavenging from hosts. *Pneumocystis* genomes encode multiple gene families that have experienced a rapid accumulation of mutations favoring fungal replication in mammals. Each *Pneumocystis* species has employed different strategies to adapt to their host including lineage-specific expansions of shared gene families such as *msg* A1, A3, and D in *P. jirovecii* or gain and expansion of *RG* proteins in *P. macacae*. In addition, some shared gene families also have acquired different properties (e.g., transmembrane domain and secreted signals) potentially contributing to host specificity. The absence of a reliable culture method and the inability to genetically manipulate *Pneumocystis* prevents

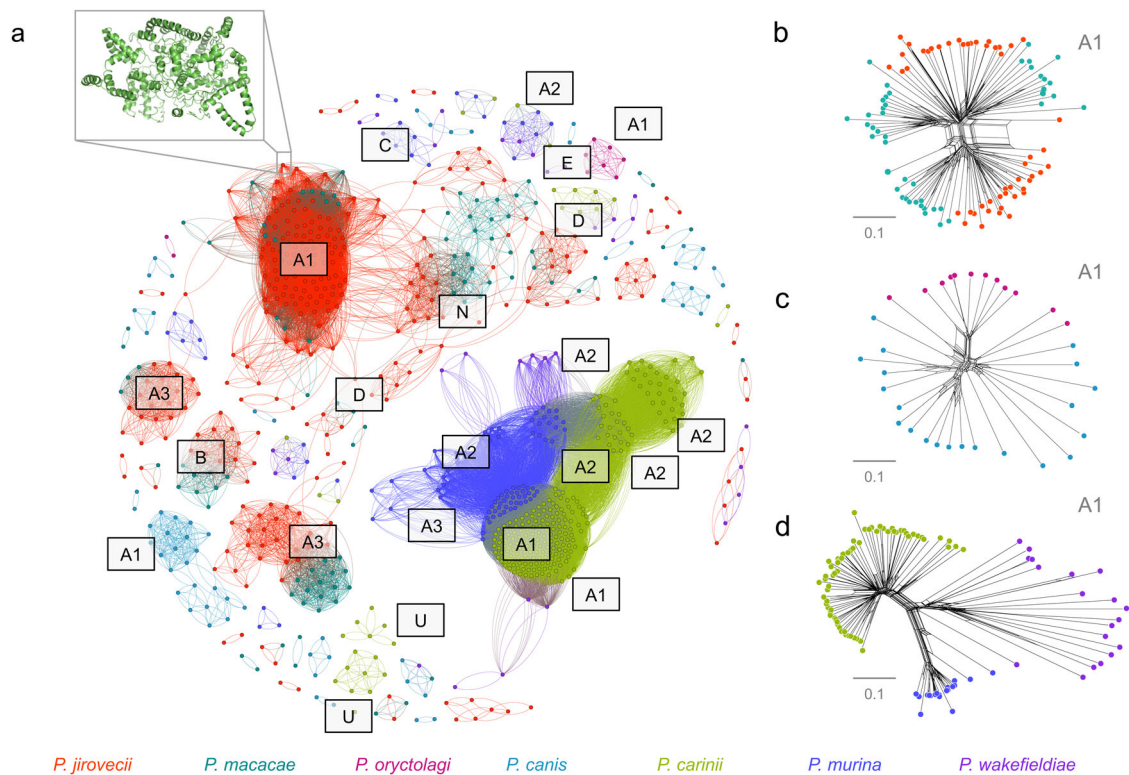


Fig. 4 Clustering of *Pneumocystis* major surface glycoproteins (Msg). **a** Graphical representation of similarity between 482 Msg proteins from seven *Pneumocystis* species generated using the Fruchterman Reingold algorithm. A 3-D model of a representative member of Msg-A1 protein family (NCBI locus tag T551_00910) generated using DESTINI is presented in the upper left insert. Individual protein sequences are shown as dots and color-coded by species as shown at the bottom of the figure. The edge between two dots indicates a global pairwise identity equal or greater than 45%. The letters represent Msg families (A to E) and subfamilies (A1 to A3). N and U letters represent potentially novel Msg sequences (relative to our prior study³⁵) and unclassified sequences, respectively. For sake of clarity only the major clusters were annotated. **b** Phylogenetic network of a subset of Msg-A1 family ($n = 97$) in primate *Pneumocystis* including *P. jirovecii* (red) and *P. macacae* (dark cyan) suggesting recombination events at the root of the network. Nodes with more than two parents represent reticulate events. Bars represent the number of amino acid substitutions per site. **c** Phylogenetic network of Msg family A1 ($n = 33$) in *P. oryctolagi* (red violet) and *P. canis* (light blue). **d** Phylogenetic network of Msg-A1 family ($n = 113$) in rodent *Pneumocystis* including *P. carinii* (green), *P. murina* (dark blue), and *P. wakefieldiae* (blue violet). The network data are available at 10.5281/zenodo.4450766.

directly testing our model. Moreover, for the genes that we have now implicated in the process of host adaptation, only a few have been functionally characterized. Future studies on the role of these genes will be important to elucidate the molecular basis of host-specific adaptation by *Pneumocystis* pathogens.

Methods

Experimental design and *Pneumocystis* sample sources. Animal and human subject experimentation guidelines of the National Institutes of Health (NIH) were followed in the conduct of this study. Studies of human and mouse *Pneumocystis* infection were approved by NIH Institutional Review Board (IRB) protocols 99-I-0084 and CCM 19-05, respectively. The collection and processing of a single human *P. jirovecii* sample from China (Pj55) was approved by the IRB of the First Affiliated Hospital of Chongqing Medical University, China (protocol no. 20172901). Written informed consent was obtained from the patient for participation in this study. The authors confirm that personal information was unidentifiable from this report. The National Institute of Allergy and Infectious Diseases (NIAID) Division of Intramural Research Animal Care and Use Program, as part of the NIH Intramural Research Program, approved all experimental procedures pertaining to the macaques (protocol LVD 26). Nonhuman primate study protocols were approved by the Institutional Animal Care and Use Committee of the University of California, Davis (protocol no. 7092), the Tulane National Primate Research Center (TNPRC), and the Institutional Animal Care and Use Committee (IACUC) (protocol no. P0351R). Studies of rabbit *Pneumocystis* infection were reviewed and approved by the Institutional Animal Care and Use Committee of the University of Michigan (protocol no. RO00008218). For rabbit samples obtained from France, the conditions for care of laboratory animals stipulated in European guidelines were followed (See: Council directives on the protection of animals for experimental and other scientific purposes, and J. Off.

Communautés Européennes, 86/609/EEC, 18 December 1986, L358). Samples from *Pneumocystis*-infected dog were collected as diagnostic samples and approved only for research purposes. The owner's consents for using samples and data were obtained on admission of the case and no further ethics permission was required because it was a routine diagnostic case and did not qualify as an animal experiment. Studies of rat *Pneumocystis* infection were approved by the Veteran Affairs animal protocol (VA ACORP #17-12-05-01). Clinical information and demographic data of the groups of individuals are presented in Supplementary Table 1. Three *P. jirovecii* samples were obtained as bronchoalveolar lavage from patients at the NIH Clinical Center in Bethesda, MD, USA and Chongqing Medical University in Chongqing, China. Six *P. macacae* samples were obtained as frozen lung tissues or formalin fixed paraffin embedded (FFPE) tissue sections prepared from SIV-infected rhesus macaques at the NIH Animal Center, Bethesda, Maryland ($n = 2$), the Tulane National Primate Research Center, Covington, Louisiana ($n = 3$), and the UC Davis California National Primate Research Center, Davis, California, USA ($n = 1$).

Four *P. oryctolagi* samples were obtained as frozen lung tissues from one rabbit with severe combined immunodeficiency at the University of Michigan, Ann Arbor, Michigan, USA, or as DNA from two corticosteroid treated rabbits and one rabbit with spontaneous *Pneumocystis* infection at the Institut Pasteur de Lille and the Institut National de la Recherche Agronomique de Tours Pathologie Aviaire et Parasitologie, Tours, France. *P. canis* samples were obtained as DNA from one Cavalier King Charles Spaniel dog at the University of Helsinki, Finland and one Whippet mixed-breed at the University of Veterinary Medicine, Vienna, Austria. The dogs were not laboratory animals. One *P. murina* sample was obtained from a heavily infected CD40L-KO mouse following a short-term in vitro culture. Genomic data obtained from *P. murina* isolates were combined with previously sequenced public data (Supplementary Table 2) and used for population genomics analysis (section "Speciation history of the *Pneumocystis* genus" and Supplementary Note 1). One frozen cell pellet and 4 agarose gel blocks containing *P. wakefieldiae* and *P. carinii* were obtained from immunosuppressed rats (one gel

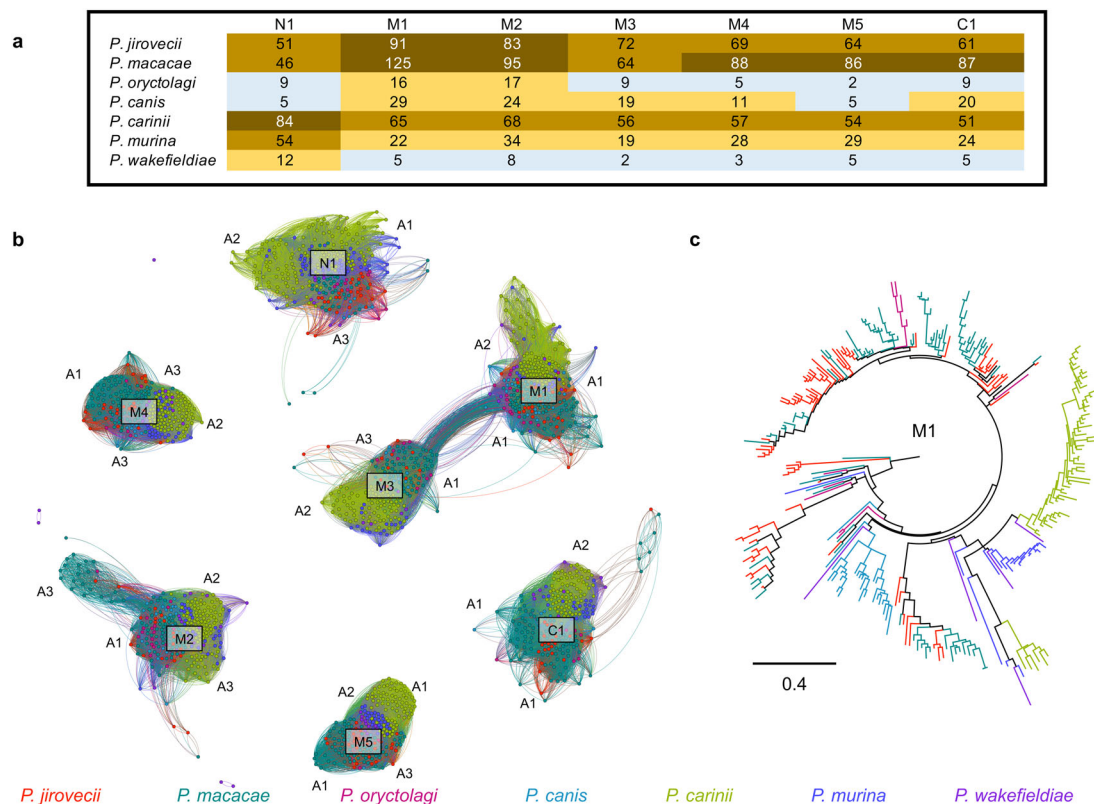


Fig. 5 Evolution of Msg cysteine-rich protein domains in *Pneumocystis*. **a** Heatmap showing the distribution of Msg domains in each *Pneumocystis* species. The color change from blue-orange-brown indicates an increase in the number of domains. **b** Graphical representation of protein similarity between domains, which highlights that the domains were present in the most recent common ancestor and were maintained other than perhaps domains M1 and M3. Domains are clustered by a minimum BLASTp cutoff of 70% protein identity. **c** Maximum likelihood tree of the M1 domain. In both panels **b** and **c**, domains are color-coded by species as shown at the bottom.

block per rat) housed at the Cincinnati VA Medical Center, Veterinary Medicine Unit, Cincinnati, Ohio. Of note, all samples with low sequence coverage were used either in combination with other samples to generate consensus genome assemblies or for population genetic analyses in which variation in sequencing coverage are explicitly accounted for.

Genome sequencing, assembly, and annotation. Genomic DNA in agarose gel blocks was extracted using the ZymoClean Gel DNA Recovery Kit (Zymo Research). Genomic DNA in FFPE sections was extracted using the AllPrep DNA/RNA FFPE Kit (Qiagen). Genomic DNA in frozen lung tissues from two *P. macacae*-infected macaques and a single *P. oryctolagi*-infected rabbit was treated with a sequential collagenase type I and DNase I digestion¹³ to deplete host DNA and extracted using the MasterPure Yeast DNA purification kit (Epicentre Biotechnologies, Madison, WI, USA). Genomic DNA in bronchoalveolar lavage samples from *P. jirovecii*-infected patients was extracted using the MasterPure Yeast DNA purification kit. Total RNAs for *P. macacae*, *P. wakefieldiae* and *P. murina* were isolated using RNeasy Mini kit (Qiagen). For DNA samples with small quantity, including three *P. oryctolagi* samples (RABF, RAB1, and RAB2B) and one *P. jirovecii* sample (RU817), we performed whole-genome amplification prior to Illumina sequencing. Five microliters of each DNA sample were amplified in a 50- μ l reaction using an Illustra GenomiPhi DNA V3 DNA amplification kit (GE Healthcare, United Kingdom). Genomic DNA samples were quantified using Qubit dsDNA HS assay kit (Invitrogen) and NanoDrop (ThermoFisher). RNA integrity and quality were assessed using Bioanalyzer RNA 6000 picosassay (Agilent). The identities of *Pneumocystis* organisms were verified by PCR and Sanger sequencing of mtLSU before high throughput sequencing. No quantitative PCR methods were used. For most of the DNA samples, at least one microgram of each DNA or RNA (depleted of ribosomal RNA using Illumina Ribo-Zero rRNA Removal Kit) sample was sequenced commercially using the Illumina HiSeq2500 platform with 150 or 250-base paired-end libraries (Novogene Inc, USA) or for one DNA sample of *P. jirovecii* from a Chinese patient using a single-read SE50 library using the MGISEq 2000 platform (MGI Tec, China).

Adapters and low-quality reads were discarded using trimmomatic v0.36³⁷ with the parameters “-phred33 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36”. Host DNA and other contaminating sequences were removed by

mapping against host genomes using Bowtie2 v2.4.1³⁸. Filtered Illumina reads were assembled de novo using Spades v3.11.1³⁹. Details for host DNA sequences removal, *Pneumocystis* reads recovery and de novo assembly protocols are presented in Supplementary Methods. Completeness of assemblies was estimated using BUSCO v9⁴⁰, FGMP v1.0.1⁴¹ and CEGMA v2.5⁴². For *P. macacae*, in addition to Illumina sequencing, Nanopore sequencing was performed on *P. macacae* DNA samples prepared from a single heavily infected macaque (P2C) with ~68% *Pneumocystis* DNA based on prior Illumina sequencing (Supplementary Table 2). High molecular weight genomic DNA fragments were isolated using the BluePippin (Sage Science) with the high-pass filtering protocol. A DNA library was prepared using the rapid Sequencing kit (SQK-RAD0004) from Oxford Nanopore Technologies (Oxford, UK) and loaded in the MinION sequencing device. Host reads were removed by mapping to the Rhesus macaque genome (NCBI accession number GCF_000772875.2_Mmul_8.0.1) using Minimap2 v2.10⁴³. Unmapped reads were aligned to the draft version of *P. macacae* assembly built previously using Illumina data (Supplementary Methods) with ngmlr v0.2.7⁴³. A total of 1,633,376 Nanopore reads were obtained, of which ~5% were attributed to *Pneumocystis* (27-fold coverage), which is much less than the 68% based on Illumina data (Supplementary Table 2). This suggests that many *P. macacae* genomic DNA fragments were too short to pass the size selection filter.

Pneumocystis Nanopore reads were assembled using Canu v1.8.0⁴⁴, overlapped with the assembly using Racon v1.3.3⁴⁵ and polished with Pilon v1.22⁴⁶ using the Illumina reads aligned with BWA MEM v0.7.17⁴⁷.

Illumina RNA-Seq of the *P. macacae* sample P2C yielded 22 million reads, of which ~92% were attributed to *Pneumocystis* (Supplementary Table 2). Filtered reads were mapped to the *P. macacae* assembly using hisat2 v2.2.0⁴⁸, sorted with SAMtools v1.10⁴⁹ and filtered with PICARD v2.1.1 (<http://broadinstitute.github.io/picard>). De novo transcriptome assembly of filtered reads was performed with Trinity⁵⁰. Quantification of transcript abundance was performed using Kallisto v0.46.1⁵¹. *P. wakefieldiae* (2A) and *P. murina* RNA-Seq data were processed similarly (Supplementary Table 2). DNA transposons, retrotransposons and low complexity repeats were identified using RepeatMasker⁵², RepBase⁵³ and TransposonPSI (<http://transposonpsi.sourceforge.net>). *Pneumocystis* telomere motif “TTAGGG”¹⁶ was searched using “FindTelomere” (available at <https://github.com/JanaSperschneider/FindTelomeres>). The genomes of *P. carinii* strain Ccin¹⁴ and strain SE6¹² were scaffolded with Satsuma⁵⁴ using the *P. carinii* strain

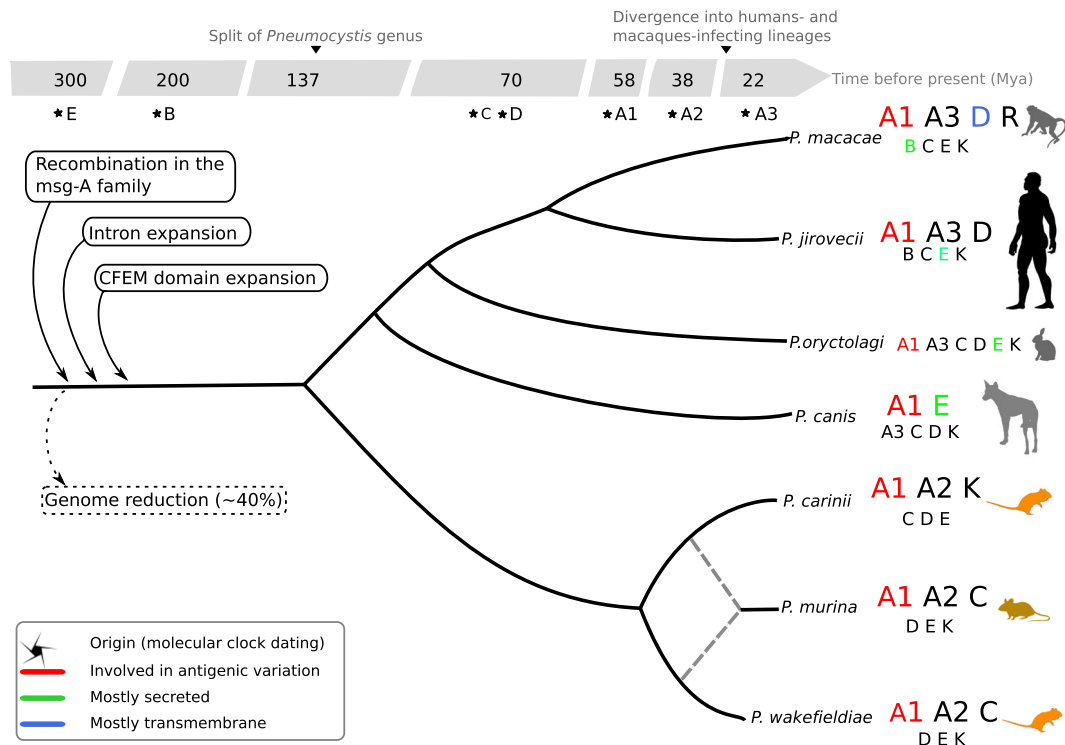


Fig. 6 Overview of the genomic evolution of the *Pneumocystis* genus. Gene families are represented by letters: A to E for the five families of major surface glycoproteins (Msg) with the A family being further subdivided into three subfamilies A1, A2, and A3; K and R for kexins and arginine-glycine rich proteins, respectively. Larger fonts indicate expansions as inferred by maximum likelihood phylogenetic trees and networks. Dashed lines represent ancient hybridization between *P. carinii* and *P. wakefieldiae*. Detailed analysis also reveals distinct phylogenetic clusters within subfamilies. Introns and CFEM (common in fungal extracellular membrane) domains are enriched in *Pneumocystis* genes which indicate that these elements were likely present in the most recent common ancestor of *Pneumocystis* species. Animal icons were obtained from <http://phylopic.org> under creative commons licenses <https://creativecommons.org/licenses/by/3.0/>: mouse (Anthony Caravaggi; license CC BY-NC-SA 3.0); dog (Sam Fraser-Smith and vectorized by T. Michael Keeseey; license CC BY 3.0), rabbit (by Anthony Caravaggi; license CC BY-NC-SA 3.0), and rat (by Rebecca Groom; license CC BY-NC-SA 3.0). Icons original black color background were modified to gray and orange colors.

B80 as reference genome¹³. *P. macacae*, *P. oryctolagi*, *P. canis* Ck1, *P. canis* A, *P. wakefieldiae*, and *P. carinii* (strains Ccin and SE6) genome assemblies were annotated using Funannotate v1.5.3 (<https://doi.org/10.5281/zenodo.1134477>). The homology evidence consists of fungal proteins from UniProt⁵⁵ and BUSCO v9 fungal proteins⁴⁰. For *P. macacae* and *P. wakefieldiae*, RNA-Seq mapping files (BAM) and de novo transcriptome assemblies were used as hints for AUGUSTUS⁵⁶. Ab initio predictions were performed using GeneMark-ES⁵⁷. All evidences were merged using EvidenceModeler⁵⁸. *Taphrina* genomes (*T. deformans*, *T. wiesneri*, *T. flavoruba*, and *T. populina*^{32,59}) and *P. canis* Ck2 genome were annotated using MAKER2⁶⁰ because predicted gene models showed a better quality than those obtained from Funannotate. MAKER2 integrates ab initio prediction from SNAP⁶¹, AUGUSTUS built-in *Pneumocystis* gene models⁶² and GeneMark-ES as well as BLAST-based homology evidences from a custom fungal proteins database. GPI prediction was performed using PredGPI⁶³, big-PI⁶⁴ and KohGPI⁶⁵. Signal peptide leader sequences and transmembrane helices were predicted using Signal-P version 5⁶⁶ and TMPred⁶⁷, respectively. Protein domains were inferred using Pfam database version 3.1⁶⁸ with PfamScan (https://bio.tools/pfamscan_api). PRIAM⁶⁹ release JAN2018 was used to predict ECs using the following options: minimum probability > 0.5, profile coverage > 70%, check catalytic—TRUE and e-value < 10⁻³. *Pneumocystis* mitochondrial genome assembly and annotations are presented in Supplementary Methods. Three dimensional (3D) protein structure prediction of Msg proteins was performed using DESTINI⁷⁰ and visualized with PyMol (www.pymol.org).

Comparative genomics. All genomes were pairwise aligned to the *P. jirovecii* strain RU7 genome NCBI accession GCF_001477535.1¹³ using LAST version 921²² with the MAM4 seeding scheme⁷¹. One-to-one pairwise alignments were created using maf-swap utility of LAST package and merged into a single multispecies whole-genome alignment using LAST's maf-join utility. Pairwise rearrangement distances in terms of minimum number of rearrangements were inferred using GRIMM⁷² and Mauve⁷³. Breakpoints of genomic rearrangements were refined with Cassis⁷⁴ and annotated using BEDtools⁷⁵ 'annotate' command. Average pairwise genome-wide nucleotide divergences were computed with Minimap2⁷⁶. Synteny

visualization was carried out using Synima⁷⁷. Msg protein similarity networks were based on global pairwise identity obtained from pairwise alignments of full-length proteins using Needle⁷⁸ or BLASTp⁷⁹ identity scores for individual protein domains. The networks presented in Figs. 4 and 5 were generated using the Fruchterman Reingold algorithm as implemented in Gephi v0.9.2⁸⁰. To generate genome sequences for low coverage samples, raw Illumina reads were aligned to reference genomes with LAST. Resulting alignments were filtered and sorted with SAMtools. SNPs and indels were identified, normalized, filtered and used to generate consensus genomes using bcftools⁸¹. Pairwise divergence scores were computed using minimap2. Sequence motifs were visualized using WebLogo⁸². Multi-panel figures were assembled in Inkscape (<https://inkscape.org>).

To investigate the evolution of introns in *Pneumocystis* species, we identified unambiguous one-to-one orthologous clusters using reciprocal best Blast hit (e-value of 10⁻¹⁰ as cutoff) in seven *Pneumocystis* species as well as in three other Taphrinomycotina fungi: *S. pombe*, *T. deformans* and *N. irregularis*. Intron position coordinates were extracted from annotated genomes using Replicer⁸³ and projected onto protein multiple alignments using custom scripts. Homologous splice sites in annotated protein sequence alignments were identified using MALIN⁸⁴. We required at least 11 unambiguous splicing sites and five minimal nongapped positions. A potential splice was considered unambiguous if the site has at least five nongap positions in the aligned sequences in both the left and right sides. MALIN uses a rates-across sites markov model with branch specific gain and loss rates to infer evolution of introns. Gain and loss rates were optimized through numerical optimizations. Fungi have a strong tendency to intron loss with few exceptions (e.g., *Cryptococcus*) whereas gain of intron is relatively rare. Thus, we penalized intron gain and set the variation rate to 4/3 for loss and gain levels. Intron evolutionary history was inferred using a posterior probabilistic estimation with 100 bootstrap support values.

Phylogenomics. Orthologous gene families were inferred using OrthoFinder v.2.3.11⁸⁵. In addition to *Pneumocystis* and *Taphrina* species, the predicted proteins for the following species were downloaded from NCBI: *Neolecta irregularis* (accession no. GCA_001929475.1), *S. pombe* (GCF_000002945.1), *S. cryophilus* (GCF_000004155.1),

S. octosporus (GCF_000150505.1), *S. japonicus* (GCF_000149845.2), *Saitoella complicata* (GCF_001661265.1), *Neurospora crassa* (GCF_000182925.2), *Cryptococcus neoformans* (GCF_000149245.1), *Rhizopus oryzae* (GCA_000697725.1) and *Batrachochytrium dendrobatidis* (GCF_000203795.1). Single-copy genes were extracted from OrthoFinder output ($n = 106$) and concatenated into a protein alignment containing 458,948 distinct alignment patterns (i.e., unique columns in the alignment) with a gap proportion of 12.2%. Maximum likelihood tree analysis was performed using RAxML v 8.2.5⁸⁶ with 1000 bootstraps as support values. The LG model⁸⁷ was selected as the best amino-acid model based on the likelihood PROTGAMMAAUTO in RAxML. One hundred and six gene trees were estimated from each of the single-copy genes. The Shimodaira–Hasegawa test¹⁸ was performed on the tree topology for each of the gene trees and the concatenated alignment using IQ-Tree⁸⁸ with 1000 RELL bootstrap replicates. IQ-Tree was run as follows: “iqtree -s [input.phy] -z [input.t] -n 0 -zb [params.n]”, where [input.phy] represents the concatenated alignment of 106 genes in phylip format (24 species; 543,202 amino-acid sites with 14.2076% of constant sites), [input.t] represents a file containing individual maximum likelihood phylogenetic trees for each of the 106 genes, -n 0 parameter avoid tree search and estimate model parameters based on an initial parsimony tree (the best-fit model according to BIC was LG + F + R7) and the -zb option specifies the number of bootstrap replicates for the resampling estimated log-likelihood method (RELL).

To infer the species phylogeny using mitochondrial genomes, protein-coding genes were extracted, aligned using Clustal Omega⁸⁹, and concatenated. The resulting alignment was used to infer phylogeny using IQ-Tree v.1.6⁹⁰ with TVM + F + I + G4 as the Best-fit substitution model and 1000 ultrafast bootstraps and SH-aLRT test. A total of 33 mitogenomes from seven *Pneumocystis* species were used: *P. jirovecii* [$n = 18$ including 3 sequences from this study and 4 from previous studies^{5,12,17}], *P. macacae* ($n = 4$), *P. oryctolagi* ($n = 4$), *P. canis* [$n = 4$, refs. 17,91], *P. carinii* [$n = 2$, refs. 17,91], *P. murina* [$n = 1$, ref. 17], and *P. wakefieldiae* ($n = 1$). Phylogenetic reconciliations of species tree and gene trees were performed using Notung⁹². Ancestral reconstruction of gene family's history was performed using Count⁹³. Phylogenetic network for Msg protein families was inferred using SplitTree⁹⁴. The detection of putative mosaic genes was performed using TOPALI v2.5⁹⁵.

Phylogenetic. Single-copy orthogroup nucleotide sequences were aligned using MACS v0.9b1⁹⁶. Highly polymorphic *msg* sequences were excluded using BLASTn⁷⁹ with an e-value of 10^{-5} as cutoff against 479 published *msg* sequences¹³. We inferred the divergence timing using two datasets: (1) 24 single-copy nuclear gene orthologs shared by all *Pneumocystis* and *S. pombe*; and (2) 568 nuclear genes found in all *Pneumocystis* species. BEAST inputs were prepared using BEAUTi v2.5.1⁹⁷. Unlinked relaxed lognormal molecular clock models^{98,99} and calibrated birth-death tree priors¹⁰⁰ were used to estimate the divergence times and the credibility intervals. The substitution site model HKY was applied¹⁰¹. Three secondary calibration priors were used: (i) *P. jirovecii*/*P. macacae* divergence with a median time of 65 mya as 95% highest posterior density (HPD)⁵, (ii) the emergence of the *Pneumocystis* genus at a minimum age of 100 mya⁴, and (iii) the *Schizosaccharomyces*–*Pneumocystis* split at ~467 mya¹⁰². We sampled from various priors and found a minor difference between the marginal posterior distribution on rate and the marginal prior distribution. This indicates that the posterior simply reflects the prior. For the dataset 2, the 568 gene alignments were concatenated in a super alignment with 568 partitions, with each partition defined by one gene. Gene partitions were collapsed using PartitionFinder v2.1.1¹⁰³ with the “greedy” search to find optimal partitioning scheme. The alignment was split in three partitions in BEAST. Three independent runs for each dataset were performed separately for 60 million generations using random seeds. Run convergence was assessed with Tracer v1.7.1 (minimum effective sampling size of 200 with a burn-in of 10%). Trees were summarized using TreeAnnotator v.2.5.1 (<http://beast.bio.ed.ac.uk/treeannotator>) and visualized using FigTree v.1.4.4 (<http://tree.bio.ed.ac.uk/software/figtree>) to obtain the means and 95% HPD. Host divergences were obtained from the most recent mammal tree of life⁶, available at <http://verlife.org/data/mammals>. The dating of fungal gene families was performed similarly. Phylogenetic trees with geological time scale were visualized using strap version 1.4¹⁰⁴.

Population genomics. Sequence data sources and primary statistics are presented in Supplementary Table 2. Adapter sequences and low-quality headers of base sequences were removed using Trimmomatic³⁷. Interspecies read alignment was performed using LAST²² with the MAM4 seeding scheme⁷¹. Alignments were processed by last-split utility to allow interspecies rearrangements, sorted using SAMtools v1.10⁴⁹. Duplicates were removed using PICARD v2.1.1. To compute the F_{ST} and nucleotide diversity (Watterson, pairwise, FuLi, fayH, L), we calculated the unfolded site frequency spectra for each population using the Analysis of Next Generation Sequencing Data (ANGSD)²³. Site frequency spectra was estimated per base site allele frequencies using ANGS^{23,105}. Hierarchical clustering was performed using ngsCovar¹⁰⁶. All data were formatted to fit a sliding windows of 1–10 kb using BEDTools¹⁰⁷. For each window, an average value of the statistics was calculated using custom scripts.

Gene flow inference. To infer a phylogenetic network, we used 1718 one-to-one orthologs from gene catalogs of seven *Pneumocystis* species using reciprocal best BLASTp hit with an e-value of 10^{-10} as cutoff. Sequences from each orthologous group were aligned using Muscle¹⁰⁸. Alignments with evidence of intragenic recombination were filtered out using PhiPack¹⁰⁹ with a p -value of 0.05 as cutoff. For each aligned group a maximum likelihood (ML) tree was inferred using RAxML-ng¹¹⁰ with GTR + G model and 100 bootstrap replicates, and Bayesian tree was generated using BEAST2⁹⁷. ML trees were filtered using the following criteria: 0.9 as the maximum proportion of missing data, 100 as the minimum number of parsimony-informative sites, 50 as the minimum bootstrap node-support value and 0.05 as the minimum p -value for rejecting the null hypothesis of no recombination within the alignment. BEAST trees with an effective sampling size <200 were removed. Filtered trees were summarized using Treannotator (<https://www.beast2.org/treeannotator/>). Summary trees with an average posterior support inferior to 0.8 were discarded. Species network was inferred using PhyloNet option “InferNetwork_MPL”²⁴ with prior reticulation events ranging from 1 to 4. Phylogenetic networks were visualized using Dendroscope 3¹¹¹.

The highest probability network inferred a hybridization between *P. carinii* and *P. wakefieldiae* leading to *P. murina* followed by a backcrossing between *P. murina* with *P. wakefieldiae* (log probability = -12759.4). Analysis of tree topology frequencies revealed that 64% of the trees were consistent with the topology of (*P. carinii*, (*P. murina*, *P. wakefieldiae*)), 28% with the topology of (*P. wakefieldiae*, (*P. carinii*, *P. murina*)) and 8% with the topology of (*P. murina*, (*P. carinii*, *P. wakefieldiae*)).

Detection of positive selection. To search for genes that have been subjected to positive selection in *P. jirovecii* alone after the divergence from *P. macacae*, we used the branch-site test³³ as implemented in PAML¹¹², which detects sites that have undergone positive selection in a specific branch of the phylogenetic tree (foreground branch). A set of 2466 orthologous groups between *P. jirovecii*, *P. macacae* and *P. oryctolagi* was used for the test. d_N/d_S ratio estimates per branch per gene were obtained using Codeml (PAML v4.4c) with a free ratio model of evolution. This process identified 244 genes with a signal of positive selection only in *P. jirovecii* ($d_N/d_S > 1$).

Statistics and reproducibility. All analyses were conducted in R version 3.3.2¹¹³. Statistical details and experimental design for different data analyses are presented in the respective results and methods sections. No sample-size calculation was performed for genome sequencing. The assessment of protein domains associated Gene Ontologies (GO) term enrichment was performed using hypergeometric test as implemented in dGor version 1.0.6¹¹⁴ (p -values adjusted by Benjamini–Hochberg method). The statistical significance of differences among groups was determined using Wilcoxon signed-rank test.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The datasets generated during and/or analyzed during the current study are available at NCBI BioProject portal (<https://www.ncbi.nlm.nih.gov/bioproject/>): *Pneumocystis macacae* strain P2C (no. PRJNA632025); *P. oryctolagi* strain CS1 (PRJNA632560); *P. canis* strain Ck1 (PRJNA632556); *P. canis* strain Ck2 (PRJNA632878); *P. canis* strain A (PRJNA636786); *P. wakefieldiae* strain 2A (PRJNA632570); *P. jirovecii* strain 55 (PRJNA647920), *P. jirovecii* strain 54c (PRJNA648092), *P. jirovecii* strain 46 (PRJNA648096), *P. macacae* strain CJ36 (PRJNA648103), *P. macacae* strain ER17 (PRJNA648108), *P. macacae* strain UC86 (PRJNA648112), and *P. macacae* strain GL92 (PRJNA648115).

Code availability

All custom bioinformatic analyses were conducted using Perl v5.26.0 (<http://www.perl.org/>) or Python v.3.6 (<http://www.python.org>) scripts. Pipelines were written with Snakemake v5.11.2¹¹⁵. Custom codes generated for this project are available at GitHub: https://github.com/ocisse/pneumocystis_evolution and a stable released version is available at <https://doi.org/10.5281/zenodo.4450766>.

Received: 17 September 2020; Accepted: 4 February 2021;

Published online: 08 March 2021

References

- Durand-Joly, I. et al. *Pneumocystis carinii* f. sp. hominis is not infectious for SCID mice. *J. Clin. Microbiol.* **40**, 1862–1865 (2002).
- Gigliotti, F., Harmsen, A. G., Haidaris, C. G. & Haidaris, P. J. *Pneumocystis carinii* is not universally transmissible between mammalian species. *Infect. Immun.* **61**, 2886–2890 (1993).

3. Cushion, M. T., Keely, S. P. & Stringer, J. R. Molecular and phenotypic description of *Pneumocystis wakefieldiae* sp. nov., a new species in rats. *Mycologia* **96**, 429–438 (2004).
4. Keely, S. P., Fischer, J. M. & Stringer, J. R. Evolution and speciation of *Pneumocystis*. *J. Eukaryot. Microbiol.* **50**, 624–626 (2003).
5. Cisse, O. H. et al. Comparative population genomics analysis of the mammalian fungal pathogen *Pneumocystis*. *mBio* **9**, e00381–00318 (2018).
6. Upham, N. S., Esselstyn, J. A. & Jetz, W. Inferring the mammal tree: species-level sets of phylogenies for questions in ecology, evolution, and conservation. *PLoS Biol.* **17**, e3000494 (2019).
7. McDougall, I., Brown, F. H. & Fleagle, J. G. Stratigraphic placement and age of modern humans from Kibish, Ethiopia. *Nature* **433**, 733–736 (2005).
8. Suzuki, Y., Tomozawa, M., Koizumi, Y., Tsuchiya, K. & Suzuki, H. Estimating the molecular evolutionary rates of mitochondrial genes referring to Quaternary ice age events with inferred population expansions and dispersals in Japanese Apodemus. *BMC Evol. Biol.* **15**, 187 (2015).
9. Guillot, J. et al. Parallel phylogenies of *Pneumocystis* species and their mammalian hosts. *J. Eukaryot. Microbiol. Suppl.* 113S–115S <https://doi.org/10.1111/j.1550-7408.2001.tb00475.x> (2001).
10. Latinne, A. et al. Genetic diversity and evolution of *Pneumocystis* fungi infecting wild Southeast Asian murid rodents. *Parasitology* **145**, 885–900 (2018).
11. Petruzela, J. et al. Evolutionary history of *Pneumocystis* fungi in their African rodent hosts. *Infect. Genet. Evol.* **75**, 103934 (2019).
12. Cisse, O. H., Pagni, M. & Hauser, P. M. De novo assembly of the *Pneumocystis jirovecii* genome from a single bronchoalveolar lavage fluid specimen from a patient. *mBio* **4**, e00428–00412 (2012).
13. Ma, L. et al. Genome analysis of three *Pneumocystis* species reveals adaptation mechanisms to life exclusively in mammalian hosts. *Nat. Commun.* **7**, 10740 (2016).
14. Slaven, B. E. et al. Draft assembly and annotation of the *Pneumocystis carinii* genome. *J. Eukaryot. Microbiol.* **53**, S89–S91 (2006).
15. Lundgren, B., Cotton, R., Lundgren, J. D., Edman, J. C. & Kovacs, J. A. Identification of *Pneumocystis carinii* chromosomes and mapping of five genes. *Infect. Immun.* **58**, 1705–1710 (1990).
16. Underwood, A. P., Louis, E. J., Borts, R. H., Stringer, J. R. & Wakefield, A. E. *Pneumocystis carinii* telomere repeats are composed of TTAGGG and the subtelomeric sequence contains a gene encoding the major surface glycoprotein. *Mol. Microbiol.* **19**, 273–281 (1996).
17. Ma, L. et al. Sequencing and characterization of the complete mitochondrial genomes of three *Pneumocystis* species provide new insights into divergence between human and rodent *Pneumocystis*. *FASEB J.* **27**, 1962–1972 (2013).
18. Shimodaira, H. & Hasegawa, M. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol.* **16**, 1114–1116 (1999).
19. Aliouat-Denis, C. M. et al. *Pneumocystis* species, co-evolution and pathogenic power. *Infect. Genet. Evol.* **8**, 708–726 (2008).
20. Kitazoe, Y. et al. Robust time estimation reconciles views of the antiquity of placental mammals. *PLoS ONE* **2**, e384 (2007).
21. Shen, X. X. et al. Genome-scale phylogeny and contrasting modes of genome evolution in the fungal phylum Ascomycota. *Sci. Adv.* <https://doi.org/10.1126/sciadv.abd0079> (2020).
22. Kielbasa, S. M., Wan, R., Sato, K., Horton, P. & Frith, M. C. Adaptive seeds tame genomic sequence comparison. *Genome Res.* **21**, 487–493 (2011).
23. Korneliusson, T. S., Albrechtsen, A. & Nielsen, R. ANGSD: analysis of next generation sequencing data. *BMC Bioinformatics* **15**, 356 (2014).
24. Yu, Y. & Nakhleh, L. A maximum pseudo-likelihood approach for phylogenetic networks. *BMC Genomics* **16**, S10 (2015).
25. Mazars, E. et al. Isoenzyme diversity in *Pneumocystis carinii* from rats, mice, and rabbits. *J. Infect. Dis.* **175**, 655–660 (1997).
26. Aghova, T. et al. Fossils know it best: Using a new set of fossil calibrations to improve the temporal phylogenetic framework of murid rodents (Rodentia: Muridae). *Mol. Phylogenet. Evol.* **128**, 98–111 (2018).
27. Araujo, S. B. et al. Understanding host-switching by ecological fitting. *PLoS ONE* **10**, e0139225 (2015).
28. McBride, A. E., Conboy, A. K., Brown, S. P., Ariyachet, C. & Rutledge, K. L. Specific sequences within arginine-glycine-rich domains affect mRNA-binding protein function. *Nucleic Acids Res.* **37**, 4322–4330 (2009).
29. Russian, D. A. et al. Characterization of a multicopy family of genes encoding a surface-expressed serine endoprotease in rat *Pneumocystis carinii*. *Proc. Assoc. Am. Physicians* **111**, 347–356 (1999).
30. Bairwa, G., Hee Jung, W. & Kronstad, J. W. Iron acquisition in fungal pathogens of humans. *Metallomics* **9**, 215–227 (2017).
31. Stajich, J. E., Dietrich, F. S. & Roy, S. W. Comparative genomic analysis of fungal genomes reveals intron-rich ancestors. *Genome Biol.* **8**, R223 (2007).
32. Cisse, O. H. et al. Genome sequencing of the plant pathogen *Taphrina deformans*, the causal agent of peach leaf curl. *mBio* **4**, e00055–00013 (2013).
33. Zhang, J., Nielsen, R. & Yang, Z. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol. Biol. Evol.* **22**, 2472–2479 (2005).
34. Schmid-Siegert, E. et al. Mechanisms of surface antigenic variation in the human pathogenic fungus *Pneumocystis jirovecii*. *mBio* **8**, e01470–01417 (2017).
35. Ma, L. et al. Diversity and complexity of the large surface protein family in the compacted genomes of multiple *Pneumocystis* species. *mBio* <https://doi.org/10.1128/mBio.02878-19> (2020).
36. Flajnik, M. F. & Kasahara, M. Origin and evolution of the adaptive immune system: genetic events and selective pressures. *Nat. Rev. Genet.* **11**, 47–59 (2010).
37. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
38. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
39. Bankevich, A. et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
40. Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
41. Cisse, O. H. & Stajich, J. E. FGMP: assessing fungal genome completeness. *BMC Bioinformatics* **20**, 184 (2019).
42. Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061–1067 (2007).
43. Sedlazeck, F. J. et al. Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**, 461–468 (2018).
44. Koren, S. et al. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
45. Vaser, R., Sovic, I., Nagarajan, N. & Sikic, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).
46. Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **9**, e112963 (2014).
47. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* **1303**, 3997 (2013).
48. Pertea, M., Kim, D., Pertea, G. M., Leek, J. T. & Salzberg, S. L. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat. Protoc.* **11**, 1650–1667 (2016).
49. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
50. Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
51. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
52. Smit, A. F. A., Hubley, R. & Green, P. RepeatMasker Open-4.0. (2013–2015). <http://www.repeatmasker.org/asmtpapers.html>.
53. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **6**, 11 (2015).
54. Grabherr, M. G. et al. Genome-wide synteny through highly sensitive sequence alignment: Satsuma. *Bioinformatics* **26**, 1145–1151 (2010).
55. UniProt Consortium, T. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **46**, 2699 (2018).
56. Stanke, M. & Waack, S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19**, ii215–ii225 (2003).
57. Ter-Hovhannisyann, V., Lomsadze, A., Chernoff, Y. O. & Borodovsky, M. Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. *Genome Res.* **18**, 1979–1990 (2008).
58. Haas, B. J. et al. Automated eukaryotic gene structure annotation using EVIDENCEModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008).
59. Tsai, I. J. et al. Comparative genomics of *Taphrina* fungi causing varying degrees of tumorous deformity in plants. *Genome Biol. Evol.* **6**, 861–872 (2014).
60. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491 (2011).
61. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
62. Hauser, P. M. et al. Comparative genomics suggests that the fungal pathogen *Pneumocystis* is an obligate parasite scavenging amino acids from its host's lungs. *PLoS ONE* <https://doi.org/10.1371/journal.pone.0015152> (2010).
63. Pierleoni, A., Martelli, P. L. & Casadio, R. PredGPI: a GPI-anchor predictor. *BMC Bioinformatics* **9**, 392 (2008).

64. Eisenhaber, B., Schneider, G., Wildpaner, M. & Eisenhaber, F. A sensitive predictor for potential GPI lipid modification sites in fungal protein sequences and its application to genome-wide studies for *Aspergillus nidulans*, *Candida albicans*, *Neurospora crassa*, *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*. *J. Mol. Biol.* **337**, 243–253 (2004).
65. Fankhauser, N. & Maser, P. Identification of GPI anchor attachment signals by a Kohonen self-organizing map. *Bioinformatics* **21**, 1846–1852 (2005).
66. Almagro Armenteros, J. J. et al. SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.* **37**, 420–423 (2019).
67. Stoffel, K. H. A. W. TMbase—a database of membrane spanning proteins segments. *Biol. Chem. Hoppe Seyler* **374**, 166 (1993).
68. El-Gebali, S. et al. The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432 (2019).
69. Claudel-Renard, C., Chevalet, C., Faraut, T. & Kahn, D. Enzyme-specific profiles for genome annotation: PRIAM. *Nucleic Acids Res.* **31**, 6633–6639 (2003).
70. Gao, M., Zhou, H. & Skolnick, J. DESTINI: a deep-learning approach to contact-driven protein structure prediction. *Sci. Rep.* **9**, 3514 (2019).
71. Frith, M. C. & Noe, L. Improved search heuristics find 20,000 new alignments between human and mouse genomes. *Nucleic Acids Res.* **42**, e59 (2014).
72. Tesler, G. GRIMM: genome rearrangements web server. *Bioinformatics* **18**, 492–493 (2002).
73. Darling, A. E., Mau, B. & Perna, N. T. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE* **5**, e11147 (2010).
74. Baudet, C. et al. Cassis: detection of genomic rearrangement breakpoints. *Bioinformatics* **26**, 1897–1898 (2010).
75. Quinlan, A. R. BEDTools: the Swiss-Army tool for genome feature analysis. *Curr. Protoc. Bioinformatics* **47**, 11.12.11–11.12.34 (2014).
76. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
77. Farrer, R. A. Synima: a Synteny imaging tool for annotated genome assemblies. *BMC Bioinformatics* **18**, 507 (2017).
78. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276–277 (2000).
79. Altschul, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
80. Bastian, M., Heymann, S. & Jacomy, M. In *International AAAI Conference on Weblogs and Social Media* (2009).
81. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetic parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
82. Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res.* **14**, 1188–1190 (2004).
83. Seton Bocco, S. & Csuros, M. Splice sites seldom slide: intron evolution in oomycetes. *Genome Biol. Evol.* **8**, 2340–2350 (2016).
84. Csuros, M. Malin: maximum likelihood analysis of intron evolution in eukaryotes. *Bioinformatics* **24**, 1538–1539 (2008).
85. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**, 157 (2015).
86. Stamatakis, A. RAXML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
87. Le, S. Q. & Gascuel, O. An improved general amino acid replacement matrix. *Mol. Biol. Evol.* **25**, 1307–1320 (2008).
88. Minh, B. Q. et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
89. Sievers, F. & Higgins, D. G. Clustal Omega, accurate alignment of very large numbers of sequences. *Methods Mol. Biol.* **1079**, 105–116 (2014).
90. 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 (2015).
91. Sesterhenn, T. M. et al. Sequence and structure of the linear mitochondrial genome of *Pneumocystis carinii*. *Mol. Genet. Genomics* **283**, 63–72 (2010).
92. Stolzer, M. et al. Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. *Bioinformatics* **28**, i409–i415 (2012).
93. Csuros, M. Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics* **26**, 1910–1912 (2010).
94. Huson, D. H. & Bryant, D. Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* **23**, 254–267 (2006).
95. McGuire, G. & Wright, F. TOPAL 2.0: improved detection of mosaic sequences within multiple alignments. *Bioinformatics* **16**, 130–134 (2000).
96. Ranwez, V., Douzery, E. J. P., Cambon, C., Chantret, N. & Delsuc, F. MACSE v2: toolkit for the alignment of coding sequences accounting for frameshifts and stop codons. *Mol. Biol. Evol.* **35**, 2582–2584 (2018).
97. Bouckaert, R. et al. BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* **10**, e1003537 (2014).
98. Drummond, A. J., Ho, S. Y., Phillips, M. J. & Rambaut, A. Relaxed phylogenetics and dating with confidence. *PLoS Biol.* **4**, e88 (2006).
99. Gernhard, T. The conditioned reconstructed process. *J. Theor. Biol.* **253**, 769–778 (2008).
100. Heled, J. & Drummond, A. J. Calibrated birth-death phylogenetic time-tree priors for bayesian inference. *Syst. Biol.* **64**, 369–383 (2015).
101. Hasegawa, M., Kishino, H. & Yano, T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22**, 160–174 (1985).
102. Beimforde, C. et al. Estimating the Phanerozoic history of the Ascomycota lineages: combining fossil and molecular data. *Mol. Phylogenet. Evol.* **78**, 386–398 (2014).
103. Lanfear, R., Frandsen, P. B., Wright, A. M., Senfeld, T. & Calcott, B. PartitionFinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Mol. Biol. Evol.* **34**, 772–773 (2017).
104. Bell, M. A. & Graeme, T. L. strap: an R package for plotting phylogenies against stratigraphy and assessing their stratigraphic congruence. *Palaeontology* **58**, 379–389 (2015).
105. Korneliussen, T. S., Moltke, I., Albrechtsen, A. & Nielsen, R. Calculation of Tajima's D and other neutrality test statistics from low depth next-generation sequencing data. *BMC Bioinformatics* **14**, 289 (2013).
106. Fumagalli, M. et al. Quantifying population genetic differentiation from next-generation sequencing data. *Genetics* **195**, 979–992 (2013).
107. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
108. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
109. Bruen, T. C., Philippe, H. & Bryant, D. A simple and robust statistical test for detecting the presence of recombination. *Genetics* **172**, 2665–2681 (2006).
110. Kozlov, A. M., Darriba, D., Flouri, T., Morel, B. & Stamatakis, A. RAXML-NG: A fast, scalable, and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* <https://doi.org/10.1093/bioinformatics/btz305> (2019).
111. Huson, D. H. & Scornavacca, C. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. *Syst. Biol.* **61**, 1061–1067 (2012).
112. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
113. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/> (2019).
114. Fang, H. dcGOR: an R package for analysing ontologies and protein domain annotations. *PLoS Comput. Biol.* **10**, e1003929 (2014).
115. Koster, J. & Rahmann, S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520–2522 (2012).

Acknowledgements

This work has been funded in whole or in part with federal funds from the Intramural Research Program of the US National Institutes of Health (NIH) Clinical Center and the National Institute of Allergy and Infectious Diseases (NIAID). This study used the Office of Cyber Infrastructure and Computational Biology (OCIB) High Performance Computing (HPC) cluster at the National Institute of Allergy and Infectious Diseases (NIAID), Bethesda, MD. This study also utilized the high-performance computational capabilities of the Biowulf Linux cluster at the NIH, Bethesda, MD (<http://biowulf.nih.gov>). The content of this publication neither necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government. M.T.C. is a VA Senior Research Career Scientist supported by 51K6BX005232. Her lab is funded by support from the NIH R01HL146266 and VA Grant I01BX004441. J.E.S. and C.A.C. are CIFAR Fellows in the program Fungal Kingdom: Threats and Opportunities. Animal icons used in Figs. 3 and 6 were obtained from <http://phylopic.org> under creative commons licenses <https://creativecommons.org/licenses/by/3.0/>.

Author contributions

O.H.C., L.M. and J.A.K. conceived the project and designed all the experiments. L.M., O.H.C., C.W.L., J.B., J.X., J.S., R.B., B.P., K.V.R., R.K., A.S., M.C., V.H., J.C., L.P., M.T.C., G.K., Y.L. and J.A.K. performed the laboratory work to obtain samples for sequencing. O.H.C., L.M., J.P.D., P.P.K. and J.L. developed and implemented methods for sample processing, library preparation and sequencing. O.H.C., L.M., J.E.S., C.A.C. and N.S.U. analyzed the data. O.H.C., L.M. and J.A.K. drafted the manuscript, which was revised by all authors.

Funding

Open Access funding provided by the National Institutes of Health (NIH).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s42003-021-01799-7>.

Correspondence and requests for materials should be addressed to O.H.C., L.M. or J.A.K.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

This is a U.S. government work and not under copyright protection in the U.S.; foreign copyright protection may apply 2021