Supplementary methods

**Ka/Ks calculation**

Proteomes from *Pneumocystis jirovecii*, *P. carinii*, *Taphrina deformans*, *Schizosaccharomyces pombe*, *Schizosaccharomyces japonicus*, *S. octosporus* and *S. cryophilus* were clustered using Orthomcl with an inflation parameter of 1.5 and a BLASTP expected of $10^{-5}$ as cutoff. We recovered 898 single copy ortholog groups. For each of these 898 annotated gene pairs, we calculated the Ka/Ks ratios using the KaKs calculator (Zhang et al. 2006).

**Intergenic sequences**

Annotated genomes were downloaded from NCBI Shotgun Assembly Sequence section: *Pneumocystis jirovecii* (NCBI accession no. CAKM01000001), *Taphrina deformans* (CAHR02000001 to CAHR02000508), *Schizosaccharomyces pombe* (NC_003423, NC_003424, NC_003421), *Sacharomyces cerevisiae* (NC_001133 to NC_001148), *Encephalitozoon cuniculi* (NC_003229 to NC_003238, NC_003242) *P. carinii* intergenic sequences were extracted from previously annotated genome (Cisse et al. 2012), and correspond to genome data published elsewhere (Slaven et al. 2006). Intergenic sequences were extracted using FeatureExtract (Wernersson 2005). Sequence lengths statistics were calculated using R (http://www.R-project.org.).These results are summarized in the supplementary Figure S4 and table S6.

**Repeat estimation**

Repeat density was estimated in genomes using RepeatMasker with the fungi option (http://www.repeatmasker.org/).

**Pseudogenes searches**

A total of 4,004 *Pneumocystis* annotated genes were retrieved from EBI (http://www.ebi.ac.uk/ena/). The EBI annotation pipeline does not allow coding regions (CDS) to contain frameshifts. The *P. jirovecii* genome obtained from NCBI archive database (see above) was masked with coding regions using cross_match (Ewing and Green 1998; Ewing et al. 2002). The protein sequences from SWISSPROT ((UniProt Consortium 2014), number of protein entries: 544, 996) and Uniref90 database ((Suzek et al. 2007); release 2014; number of protein clusters: 21, 300, 339 cluster proteins) were searched in the masked genome using TBLASTN (Altschul et al. 1997) with an e-value

of $10^{-10}$ as cutoff. For each alignment generated by TBLASTN, the protein sequence and corresponding genomic location were extracted, and re-aligned with GeneWise (Birney and Durbin 2000), allowing to report frameshitfs and in-frame stop codons. A total of 1.8% and 0.84% of SWISSPROT and Uniref90 proteins were mapped to the masked *P. jirovecii* genome, respectively. To investigate whether these putative pseudogenes could result from errors in assembly, raw reads of the *P. jirovecii* genome were downloaded from NCBI Shotgun Assembly Sequence section (SRA: ERS145933; BioSample: SAMEA1496007; sample name: BALE8). These reads were re-mapped onto the masked genome using Roche Newbler gsMapper and visualized inspected using Tablet (Milne 2013).

Supplementary figures legends

**Fig. S1. Comparison of *P. jirovecii* proteome to those of other fungi.**

The phylogenetic profile of each species was determined as described in material and methods. Thirty-one percent (*n* = 1,200) of the 3,898 *P. jirovecii* proteins have orthologs in all species considered here (shown as green slices). Another 37% (*n* = 1,444) are shared, *i.e.* they are present in *P. jirovecii* and at least one other species (blue slices). About 1% (*n* = 32) and 0.2% (*n* = 7) are in Ascomycota and Taphrinomycotina specific clusters, respectively (dark green and mallow slices, respectively). About 4% (*n* = 165) are specific to the *Pneumocystis* genus (red slices). About 2% are in single taxa clusters, correspond to *P. jirovecii* specific genes (*n* = 319, light grey slices), and encode mostly for proteins with unknown function. Dark grey slices in other fungi represent genes having orthologs in other fungi, but not in *P. jirovecii*. White slices represent genes that lack sufficient homology for clustering. The asterisk in *P. carinii* figures that the gene subset is likely to be overestimated because of genome fragmentation. Genes specific to the *Schizosaccharomyces* spp. are shown light red. Genes specific to *Saccharomycotina* and *Pezizomycotina* subphyla are indicated in yellow and dark mallow, respectively.

**Fig. S2. Features of *Pneumocystis* spp. genomes versus other fungi.**

(A) Ka/Ks ratios for 898 orthologous gene pairs shared by *Pneumocystis* and *Schizosaccharomyces* spp.. (B) Intergenic region lengths of a panel of fungi with various lifestyle strategies.

**Fig. S3. Phylogenetic tree of *Pneumocystis* spp. and representative fungi.**

The maximum likelihood phylogeny of *P. jirovecii* and other fungi was described previously (Cissé et al., 2012). The scale bar represents 0.1 amino acid changes per site and all nodes have ≥ 99% bootstrap support. The bootstrap values are shown on the nodes and lengths are indicated on the branches. *Rhizopus delemar* was used as the outgroup.

**Supplementary Literature Cited**


Altschul SF, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389-3402.

Birney E, Durbin R. 2000. Using GeneWise in the *Drosophila* annotation experiment. Genome Res.,10: p. 547-8.

Cisse OH, Pagni M, Hauser PM. 2012. *De novo* assembly of the *Pneumocystis jirovecii* genome from a single bronchoalveolar lavage fluid specimen from a patient. mBio 4: e00428-00412.

Cisse OH, et al. 2013. Genome sequencing of the plant pathogen *Taphrina deformans*, the causal agent of peach leaf curl. MBio 4: e00055-00013.

Ewing B, Green P. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res, 1998. 8: p. 186-94.

Ewing B, Hiller L, Wendi MC, Green P. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res, 1998. 8: p. 175-85.

Milne I, et al. 2013. *Using Tablet for visual exploration of second-generation sequencing data.* Brief Bioinform, 14: p. 193-202.

Slaven BE, et al. 2006. Draft assembly and annotation of the Pneumocystis carinii genome. J Eukaryot Microbiol 53 Suppl 1: S89-91.

Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. 2007. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics, 23: p. 1282-8.

The UniProt Consortium. 2014. Activities at the Universal Protein Resource (UniProt). Nucleic Acids Res, 2014. 42: p. D191-8.

Wernersson R. 2005. FeatureExtract--extraction of sequence annotation made easy. Nucleic Acids Res. 33: p. W567-9.

Zhang Z, et al. 2006. KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. Genomics Proteomics Bioinformatics. 4: p. 259-63.

Supplementary figures

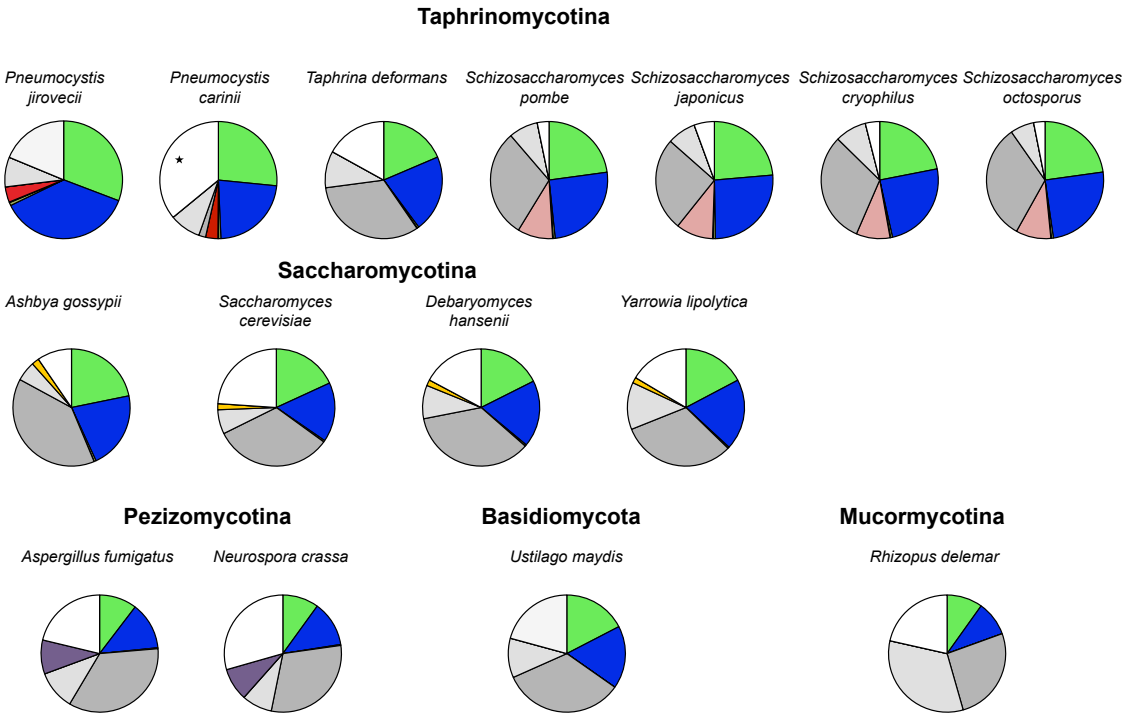


**Fig. S1. Comparison of *P. jirovecii* proteome to those of other fungi.**

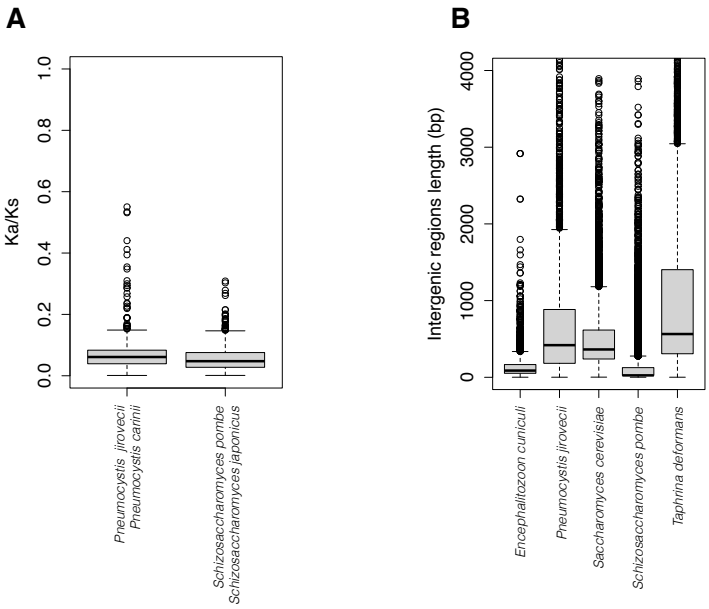


**Fig. S2. Features of *Pneumocystis* spp. genomes versus other fungi.**

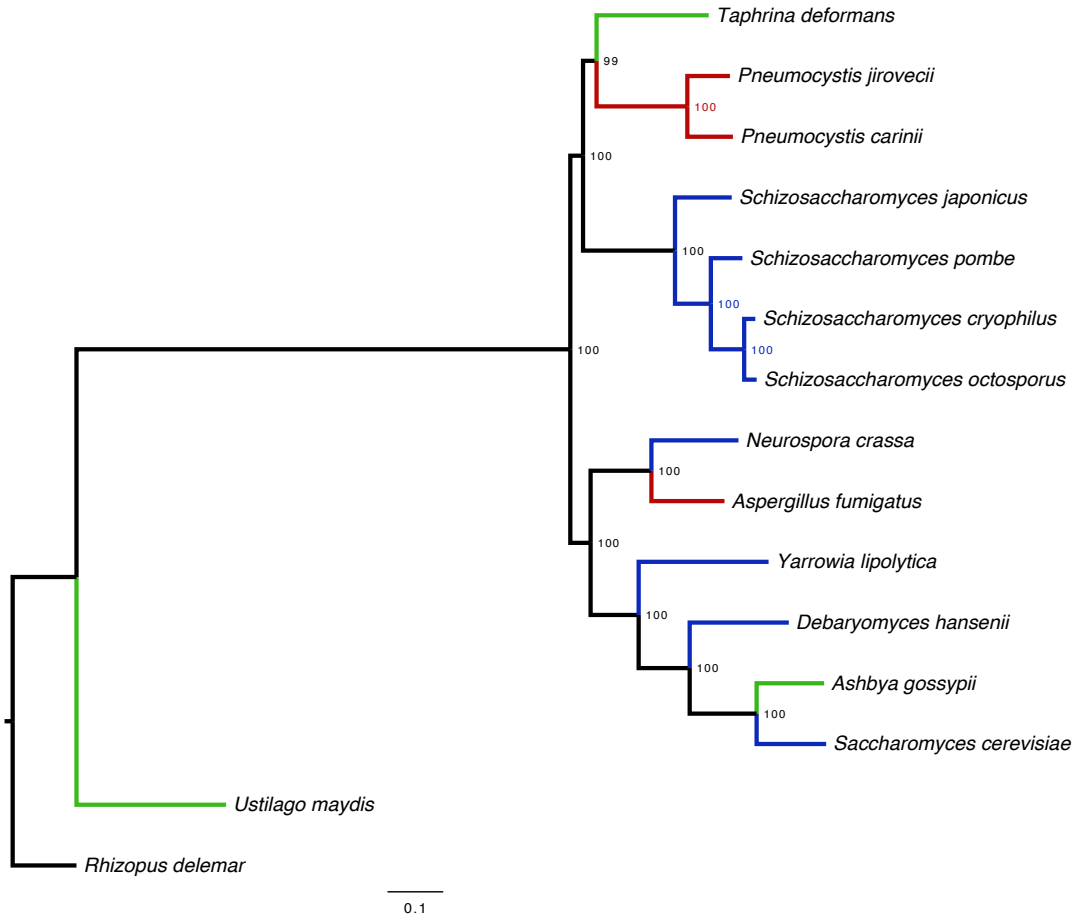


**Fig. S3. Phylogenetic tree of *Pneumocystis* spp. and representative fungi.**