Year : 2023

Study of the surface antigenic variation system and interhuman transmission of the pathogenic fungus Pneumocystis jirovecii

Meier Caroline

Meier Caroline, 2023, Study of the surface antigenic variation system and interhuman transmission of the pathogenic fungus Pneumocystis jirovecii

Originally published at : Thesis, University of Lausanne

Posted at the University of Lausanne Open Archive http://serval.unil.ch
Document URN : urn:nbn:ch:serval-BIB_579290733F273

UNIL | Université de Lausanne
Faculté de biologie et de médecine

Institut de Microbiologie de Lausanne (IMUL)
Département de médecine de laboratoire et pathologie
CHUV

Study of the surface antigenic variation system and interhuman transmission of the pathogenic fungus *Pneumocystis jirovecii*

Thèse de doctorat ès sciences de la vie (PhD)

présentée à la

Faculté de biologie et de médecine
de l'Université de Lausanne

par

Caroline MEIER

Master of Science in Biology
Major in Microbiology and Immunology
of ETH Zürich

Jury

Prof. Patrik Michel, Président
PD Dr. Philippe Hauser, Directeur de thèse
Dr. Marco Pagni, Co-directeur de thèse
Prof. ass. Frédéric Lamoth, Expert
Prof. ass. hon. Michel Monod, Expert
Prof. Joseph Kovacs, Expert

Lausanne
Juin 2023

UNIL | Université de Lausanne
Faculté de biologie et de médecine

Institut de Microbiologie de Lausanne (IMUL)
Département de médecine de laboratoire et pathologie
CHUV

Study of the surface antigenic variation system and interhuman transmission of the pathogenic fungus *Pneumocystis jirovecii*

Thèse de doctorat ès sciences de la vie (PhD)

présentée à la

Faculté de biologie et de médecine
de l'Université de Lausanne

par

Caroline MEIER

Master of Science in Biology
Major in Microbiology and Immunology
of ETH Zürich

Jury

Prof. Patrik Michel, Président
PD Dr. Philippe Hauser, Directeur de thèse
Dr. Marco Pagni, Co-directeur de thèse
Prof. ass. Frédéric Lamoth, Expert
Prof. ass. hon. Michel Monod, Expert
Prof. Joseph Kovacs, Expert

Lausanne
Juin 2023

UNIL | Université de Lausanne
Faculté de biologie et de médecine

Ecole Doctorale
Doctorat ès sciences de la vie

Imprimatur

Vu le rapport présenté par le jury d'examen, composé de

Président·e	Monsieur	Prof.	Patrik	**Michel**
Directeur·trice de thèse	Monsieur	Dr	Philippe	**Hauser**
Co-directeur·trice	Monsieur	Dr	Marco	**Pagni**
Expert·e·s	Monsieur	Prof.	Frédéric	**Lamoth**
	Monsieur	Prof.	Michel	**Monod**
	Monsieur	Prof.	Joseph	**Kovacs**

le Conseil de Faculté autorise l'impression de la thèse de

Caroline Meier

Master of Science ETH in biology, ETHZ Zürich, Suisse

intitulée

Study of the surface antigenic variation system and interhuman transmission of the pathogenic fungus *Pneumocystis jirovecii*

Lausanne, le 14 juillet 2023

pour le Doyen
de la Faculté de biologie et de médecine

P. Michel

Prof. Patrik Michel

CHUV
Dr Patrik Michel
Professeur associé, Médecin chef
Centre cérébrovasculaire
Service de neurologie
CH-1011 Lausanne

Acknowledgements

Many people contributed to the well-going of my thesis.

Firstly, I warmly thank Philippe for giving me the opportunity to complete my thesis in his lab. He was always present when I had any question, doubt or misunderstanding and always made time to talk. His patience, optimism and trust helped me bring this project to what you can read here.

Similarly, I couldn't have done it without the help of Sophie in the lab. Her immense knowledge and experience made each experiment easier. When failing, there would always be a next step and a cheer-up. I also always looked forward to the endless coffee breaks and lunches.

Many thanks to Marco for his help on the development of both the bioinformatics pipeline and my bioinformatics skills.

My deepest gratitude goes to the members of my thesis committee: Prof. Patrik Michel, Prof. ass. Frédéric Lamoth, Prof. Michel Monod, Prof. Joseph Kovacs.

The mycology team also helped me along the way. I want to dearly mention Marion, Jizhou, Danielle and Marine for their numerous in- and out- institute laughs, lunches, drinks and support.

Special thanks go to Melissa for bearing my long complains and always cheering me up by making me laugh ; Amanda who got me back to the gym and followed me with the organization of various murder parties ; Elindi for being the most surprising and breath-taking friendship this project could initiate ; and finally all the IMUL people who impacted my time here only for the better.

I also heartly thank Estelle for tolerating my long calls and vocals and always being there for me, even while doing her own thesis 400 km away.

Of course, I cannot forget Eric, without whom these last few months would have been a lot harder to bear. His tips and endless presence helped me go through the hard times.

Finally, I want to dearly thank my family and mostly my mom. She has always encouraged and helped me to her best and was present at each and every step of the way.

Abstract

Antigenic variation is a mechanism used by various pathogenic microorganisms to evade the host's immune system. The opportunistic fungus *P. jirovecii* is a non-culturable pathogen causing life-threatening pneumonia in immunocompromised individuals. To change the epitopes that are exposed at its cellular surface, it uses hypervariation of the major surface glycoproteins (Msg). These proteins belong to a superfamily composed of six families, which encoding genes are all located within the subtelomeres of all chromosomes of the fungus. Two mechanisms enable antigenic variation: gene mosaicism and mutually exclusive expression of the genes of family I (*msg*-I). The latter consists in the expression of a single gene out of the ca. 80 present in the genome, which can be exchanged overtime. This generates an antigenically heterogeneous population of *P. jirovecii* cells made of subpopulations each expressing a specific gene. In this thesis, all the *msg*-I genes present in each of 29 patients with *Pneumocystis* pneumonia from five cities were sequenced using PacBio circular consensus sequencing (CCS). This determined the repertoires of the *msg*-I genes present in the *P. jirovecii* strain(s) causing each infection. The analysis of these repertoires revealed that translocations of entire genes through intergenic recombinations led to reassortments of the repertoires. Together with evidence of other types of recombinations within the subtelomeres, our observations allowed proposing a model of the mechanisms involved in the antigenic variation system of *P. jirovecii*. We propose in addition that these recombinations are potentially mediated by the DNA triplex that the short imperfect mirror sequences present at the beginning of each *msg*-I gene might form. Besides, our analyses identified three clusters of patients presenting identical or similar *msg*-I repertoires. Two of them were confirmed using a novel genotyping technique based on the PacBio CCS sequencing of the ITS1-5.8S-ITS2 genomic region that we developed. One cluster might have resulted from transmission of a *P. jirovecii* strain harbouring a stable *msg*-I repertoire over time, and/or from infection by a dormant form of the fungus. Another cluster might have resulted from multiple transmission events and enrichment with specific *msg*-I alleles in the concerned geographical area. The last cluster suggested the working hypothesis that the repertoires reassort more or less rapidly according to the underlying disease affecting the patient.

Résumé

La variation antigénique est un mécanisme utilisé par divers micro-organismes pathogènes pour échapper au système immunitaire de l'hôte. Le champignon opportuniste *P. jirovecii* est un pathogène non cultivable qui provoque des pneumonies mortelles chez les personnes immunodéprimées. Pour modifier les épitopes exposés à sa surface cellulaire, il utilise l'hypervariation des major surface glycoproteins (Msg). Ces protéines appartiennent à une superfamille composée de six familles, dont les gènes codants sont tous situés dans les sous-télomères de tous les chromosomes du champignon. Deux mécanismes permettent la variation antigénique : le mosaïcisme génétique et l'expression mutuellement exclusive des gènes de la famille I (*msg*-I). Cette dernière se traduit par l'expression d'un seul gène sur les quelque 80 présents dans le génome, qui peut être échangé au fil du temps. Cela génère une population de *P. jirovecii* antigéniquement hétérogène, composée de sous-populations exprimant chacune un gène spécifique. Dans cette thèse, tous les gènes *msg*-I présents chez 29 patients atteints de pneumocystose et provenant de cinq villes ont été séquencés à l'aide de PacBio circular consensus sequencing (CCS). Cela a permis de déterminer les répertoires des gènes *msg*-I présents dans la ou les souches de *P. jirovecii* à l'origine de des infections. L'analyse de ces répertoires a révélé que les translocations de gènes entiers par recombinaisons intergéniques entraînaient des réarrangements des répertoires. Associées à la mise en évidence d'autres types de recombinaisons au sein des sous-télomères, nos observations ont permis de proposer un modèle des mécanismes impliqués dans le système de variation antigénique de *P. jirovecii*. Nous proposons en plus que ces recombinaisons soient possible grâce au triplex d'ADN formé par les courtes séquences miroirs imparfaites présentes avant chaque gène *msg*-I. Par ailleurs, nos analyses ont permis d'identifier trois groupes de patients avec des répertoires *msg*-I identiques ou similaires. Deux d'entre eux ont été confirmés par une nouvelle technique de génotypage, que nous avons développée, basée sur le séquençage PacBio CCS de la région génomique ITS1-5.8S-ITS2. Un groupe pourrait résulter de la tranmission d'une souche de *P. jirovecii* avec un répertoire *msg*-I stable dans le temps, et/ou d'une infection par une forme dormante du champignon. Un autre groupe pourrait résulter d'événements de transmission multiples et d'un enrichissement de certains allèles *msg*-I dans la zone géographique concernée. Le dernier groupe suggère l'hypothèse que les répertoires se réorganiseraient plus ou moins rapidement en fonction de la maladie sous-jacente du patient.

Table of contents

Abbreviations

BAL Broncho-Alveolar Lavage

BE Bern

BLAST Basic Local Alignment Search Tool

BR Brest

CCS Circular Consensus Sequence

CDS Coding Sequence

CI Cincinnati

DNA Deoxyribonucleic Acid

CRJE Conserved Recombination Junction Element

GPI Glycosylphosphatidylinositol

HIV Human Immunodeficiency Virus

ITS Internal Transcribed Spacer

LA Lausanne

MLST Multilocus Sequence Typing

MSG Major Surface Glycoprotein

PCR Polymerase Chain Reaction

PCP *Pneumocystis jirovecii* Pneumonia

RFLP Restriction Fragment Length Polymorphism

RNA Ribonucleic Acid

SD Standard Deviation

SE Seville

UCS Upstream Conserved Sequence

List of figures

List of tables

List of supplementary figures

List of supplementary tables

General introduction

Pneumocystis spp. history and nomenclature

Pneumocystis was described for the first time in 1909 by Carlos Chagas (Chagas, 1909). During his studies on trypanosome parasites, he observed new forms of cysts and classified them as one of their life stages. A few years later, the suggestion of these cysts being new organisms was emitted by the Delanoë couple who renamed them *Pneumocystis carinii*, in honour of Antonio Carini who provided them with the samples they studied (Delanoë & Delanoë, 1912; Vera & Rueda, 2021). In the following years, descriptions of *Pneumocystis* in various animals were made, but a lack of link with any disease led to a decrease of the interest of the researchers. Around 1940, a new form of pneumonia was described in premature and malnourished children with a large number of cases in Europe (Calderón-Sandubete et al., 2002). The association between this pneumonia and *Pneumocystis* spp. was made later on by Dutch and Czech researchers (Van der Meer & Brug, 1942; Vanek et al., 1953). *Pneumocystis* pneumonia gained importance during the HIV epidemic in the 1980s, during which it was the most frequent opportunistic infection. An important milestone in the *Pneumocystis* history is its attribution from the protozoan to the fungi kingdom thanks to studies of the RNA sequences in the late 1980s (Edman et al., 1988; Stringer et al., 1989). Shortly after, different species were characterized according to the host that they infect. The human pathogen was named *Pneumocystis carinii* forma specialis *hominis* in 1994 (Bartlett et al., 1994) before being officially renamed *Pneumocystis jirovecii* in 1999 (Frenkel, 1999). Nowadays, the *Pneumocystis* genus belongs to the Ascomycetes phylum and the Taphrinomycotina subphylum (Figure 1a), in which they are the only animal pathogens among diverse phytopathogens including the

Schizosaccharomyces genus and the model organism of Eukaryotes Schizosaccharomyces pombe (Eriksson & Winka, 1997).

Host adaptation

Pneumocystis has been found in the lungs of numerous mammal species shortly after its discovery in the early 20th century. Molecular and cross-infection studies revealed the specificity of each *Pneumocystis* species to a unique host species (Figure 1b) (Durand-Joly et al., 2002; Gigliotti et al., 1993). Nowadays, besides *P. jirovecii* in humans, the following species have been characterized: *Pneumocystis murina* in *Mus musculus* (mouse, Keely et al., 2004), *Pneumocystis carinii* and *Pneumocystis wakefieldiae* in *Rattus norvegicus* (rat, Cushion et al., 1993, 2004), *Pneumocystis oryctolagi* in *Oryctolagus cuniculus* (rabbit, Dei-Cas et al., 2006), *Pneumocystis canis* in dogs, and *Pneumocystis macacae* in macaque (Cissé et al., 2021). *Pneumocystis* organisms have also been isolated from numerous other mammal species, such as ferrets, horses, pigs and shrews (Aliouat-Denis et al., 2008; Blasi et al., 2021; Laakkonen et al., 1993; Weissenbacher-Lang et al., 2023). Co-evolution and host adaptation are the most common explanations for host specificity of the *Pneumocystis* species (Aliouat-Denis et al., 2008; Cissé et al., 2021).

P. jirovecii - human relationship

P. jirovecii can cause *Pneumocystis jirovecii* pneumonia (PCP, or PJP) in immunocompromised individuals. Besides HIV positive patients, transplant recipients and individuals with immunosuppressive conditions are the most susceptible to develop PCP. PCP is the second most frequent invasive fungal infections worldwide

with over 500.000 cases annually and a mortality rising to 80% when left untreated (Bongomin et al., 2017; Brown et al., 2012).

The fungus was also found in the lungs of immunocompetent individuals, however without provoking any symptoms. *P. jirovecii* is believed to be able to colonize human lungs without harm (Cushion & Stringer, 2010). Healthcare workers in contact

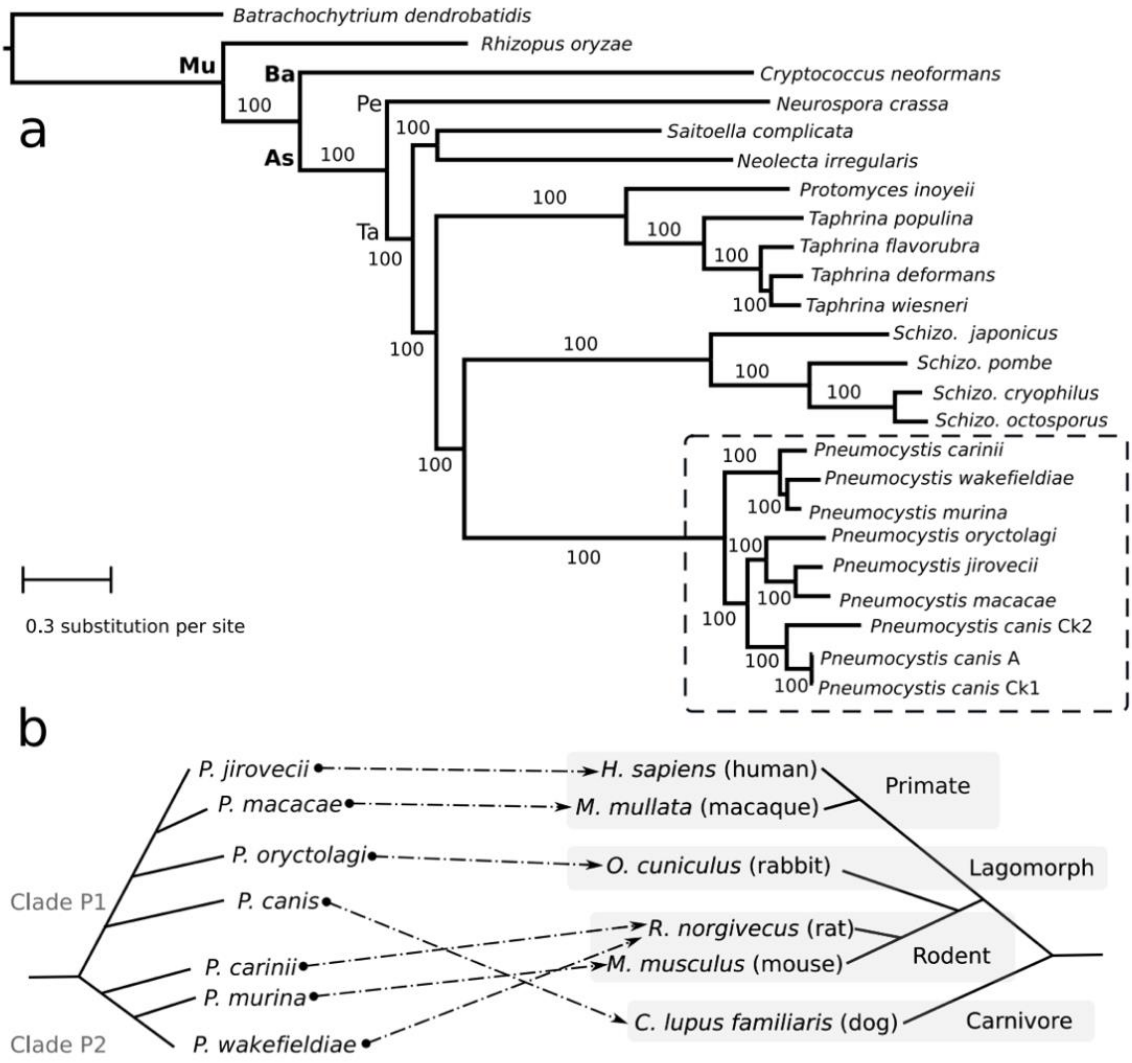


Figure 1. Phylogeny of *P. jirovecii*.

a Maximum likelihood phylogeny of 24 fungal species. The *Pneumocystis* species are within the dashed box. Bootstrap support (%) is presented on the branches. (As : Ascomycota; Ba : Basidiomycota; Pe : Pezizomycotina; Mu : Mucoromycota; Ta : Taphrinomycotina)

b Species phylogeny and association between *Pneumocystis* species and their respective mammalian hosts. The dashed arrows represent the specific parasite-host relationships.

Figures and legends adapted from reference (Cissé et al., 2021)

with PCP patients have been found to be at least transiently colonized with *P. jirovecii* (Miller et al., 2001; Valade et al., 2015). Several studies investigated the presence of the fungus in asymptomatic immunocompetent adults, and the results were diverse, as between 0% and 65% of the tested individuals presented *P. jirovecii* colonization in their lungs (Medrano et al., 2005; Nevez et al., 2006; Ponce et al., 2010; Vera & Rueda, 2021).

Transmission of *P. jirovecii*

The modes of spreading of *P. jirovecii* and the source of infection are still unclear. Two hypothesized transmission routes exist which are not mutually exclusive. First, cells from a previous infection within the host's lungs could reactivate upon a decrease of immunity and cause a new episode of the disease. A primo infection most probably happens very early in life as high proportions of healthy infants possess anti-*P. jirovecii* antibodies by the age of two (Morris et al., 2002; Vargas et al., 2001). High percentages of *P. jirovecii* colonization in pregnant women suggest that transmission from mother to foetus or infant might be a possible transmission route (Montes-Cano et al., 2009; Vargas et al., 2003).

The second hypothesis is *de novo* acquisition of *P. jirovecii* that leads to an infection. The asci and/or the ascospores were identified as the airborne transmission form of the fungus (Cushion et al., 2010; Martinez et al., 2013). However, the transmission's distance might be limited (Cissé et al., 2020). It is unlikely that the fungus has an environmental source as it is an obligate parasite because it lacks many biochemical pathways. This also suggests that mammals are the sole reservoirs of this fungus (Hauser et al., 2010; Ma et al., 2016). Therefore, the host-to-host transmission of *P. jirovecii* is an important spreading mechanism and happens between individuals,

possibly only when in close proximity. Additionally, numerous reports of PCP outbreaks due to interhuman transmission during brief encounters were described (Rabodonirina et al., 2004; Yiannakis & Boswell, 2016). These infection clusters were observed among individuals with the same condition, such as transplant recipients. Besides, one strain was linked to renal transplant recipients within clusters (Sassi et al., 2012).

Immunocompromised patients are known to be a major source of infection of the fungus as they develop PCP. However, several studies have shown, both in human and animal models, that immunocompetent individuals are also key elements of its transmission route (Chabé, Dei-Cas, et al., 2004; Dumoulin et al., 2000; Gigliotti et al., 2003; Menotti et al., 2013). Similarly, colonization of *P. jirovecii* is prevalent among elderly asymptomatic people, suggesting their participation in the transmission cycle of the fungus (Vargas et al., 2010).

Molecular typing of *P. jirovecii*

Most PCP have been shown to be due to multiple co-infecting *P. jirovecii* strains, up to seven were observed in a single patient (Alanio et al., 2016; Azar et al., 2022; Nahimana et al., 2000). Hence, genotyping of *P. jirovecii* is an important aspect for the understanding of the transmission of the fungus. Notably, there is no clear definition of a genotype for this fungus and various molecular typing techniques were described to genotype *P. jirovecii*:

(i) Sanger sequencing of amplified markers was first developed. Markers used included the internal transcribed spacer 1 and 2 (ITS1 and ITS2) of the rRNA operon or genes encoding the beta-tubulin, dihydrofolate reductase or kexin 1 (Ma et al., 2018). The choice of markers is

controversial, as it should stay stable at least for several months to be considered adequate for genotyping (Hauser et al., 1997).

(ii) Multilocus sequence typing (MLST) combines the analysis of multiple markers to increase the discriminatory power. However, MLST has been done with myriads of different marker combinations with varying discriminatory power (Maitte et al., 2013), thus limiting the comparisons between these studies (de Boer et al., 2007; Hauser et al., 1997; Ma et al., 2018; Schmoldt et al., 2008).

(iii) Restriction fragment length polymorphism (RFLP) has been used on various markers as well, rendering comparisons difficult. Recently, RFLP has been done on a fragment of the genes of the family I of the major surface glycoproteins (*msg*-I) which resulted in a high discriminatory power. However, it requires higher amount of start material than other methods (Ripamonti et al., 2009; Sassi et al., 2012).

(iv) Variable-number tandem-repeat takes another approach, as it does not characterize the polymorphisms in marker genes but quantifies the number of copy of short tandem repeats (Ma et al., 2018).

(v) Cutting-edge sequencing techniques enable faster and more accurate sequencing of the same markers as described above. The present study used PacBio Circular Consensus Sequence (CCS), a single-molecule real-time sequencing technique enabling an accuracy of 99.8% and the generation of reads up to 13.5 kb (Wenger et al., 2019). Circularization of the double-stranded DNA template leads to multiple sequencing of the target template molecule, resulting in a highly accurate consensus read (Figure 2).

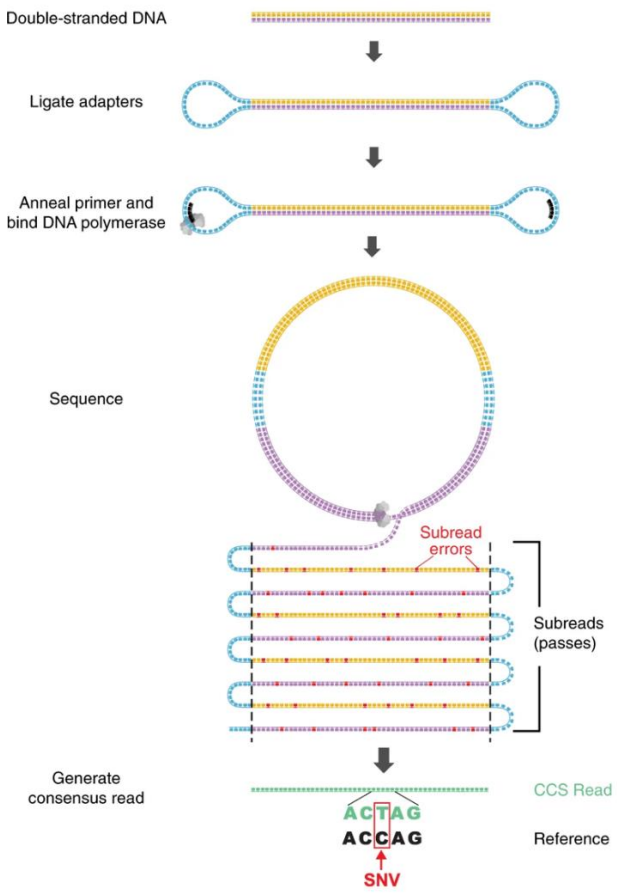


Figure 2. PacBio CCS overview. CCS derives a consensus (CCS read) from multiple sequencing of a single template molecule, producing accurate CCS reads from individual subreads with errors. Adapted figure 1a and legend from reference (Wenger et al., 2019)

Life cycle

The inability of culturing *in vitro* any *Pneumocystis* species is a large drawback for research. Research is ongoing to find a reproducible methodology to cultivate this fungus. The use of polarized CuFi-8, a human epithelial cell line in an air-liquid interface system, was proposed as a culturing method (Schildgen et al., 2014). However, it could not be replicated elsewhere (Liu et al., 2018).

The lack of a long-term culture method restrains considerably the advances in understanding the life and sexual cycles of *P. jirovecii*. Nevertheless, the following section describes the current knowledge (Figure 3). During infections, 98% of the cells within the host's lungs are trophic forms, which are anamorphous without a rigid cell

wall. They are haploid (Aliouat-Denis et al., 2009; Martinez et al., 2011; Wyder et al., 1998) and are believed to replicate asexually either by binary fission or possibly endogeny. They would be able to undergo a sexual cycle, which initiation triggers are still unknown. The fungus uses a primary homothallic reproduction mode (Almeida et al., 2015). This is a self-compatible mating system where the cells co-express both mating types to initiate the sexual cycle without the need for a compatible partner (Hauser, 2021; Hauser & Cushion, 2018; Luraschi et al., 2019; Richard et al., 2018). During its sexual cycle, *P. jirovecii* produces asci each containing eight ascospores. The asci represent a low percentage of all *P. jirovecii* cells present in the infections, ca. 2 to 10% (Shiota et al., 1986). They have a thick cell wall that maintains their spherical shape. The ascospores have also been observed as elongated cells with a condensed cytoplasm (Figure 3, blue cells). Ascospores may exit the asci through a rent at the surface of the ascus within the host's lungs (Hauser & Cushion, 2018). A

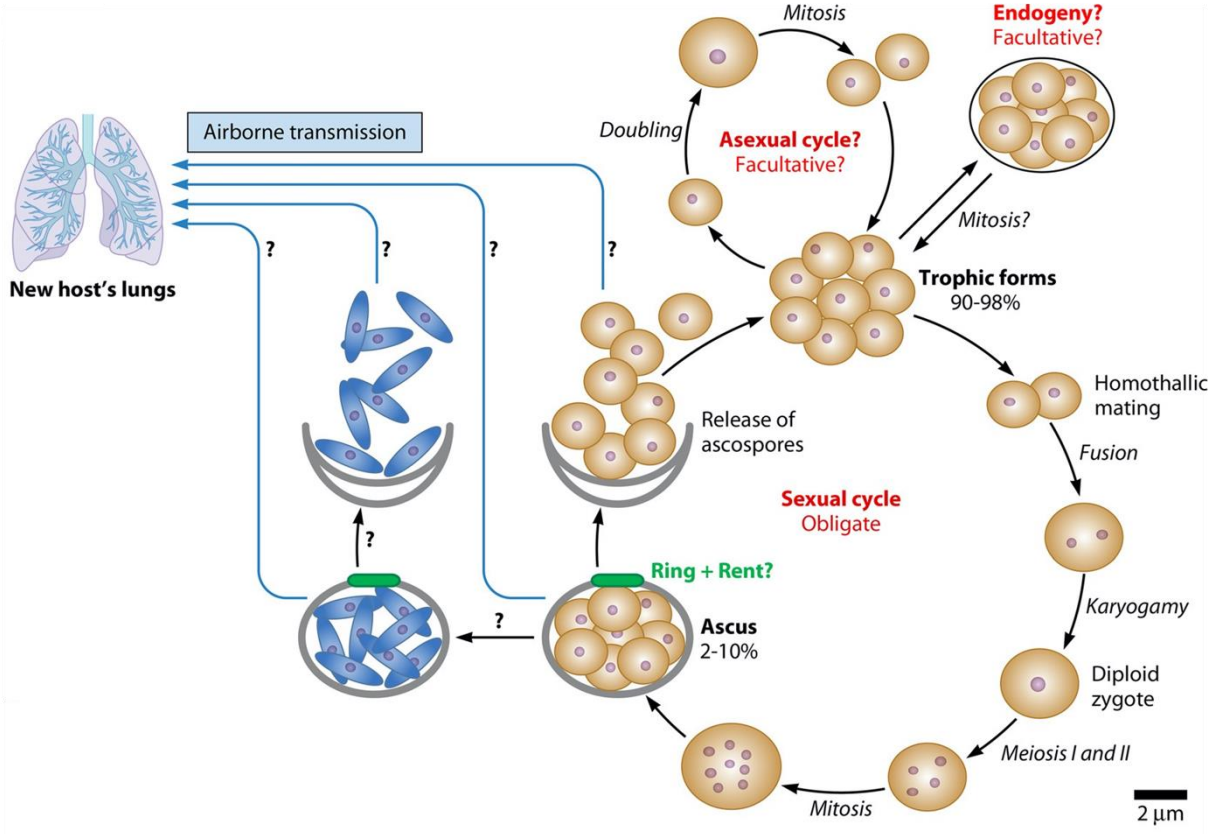


Figure 3. Latest hypothesized life cycle of *P. jirovecii* including one obligate sexual cycle and two potentially facultative asexual cycles. Figure 1A from (Hauser, 2021).

function of the sexual cycle would be to create the airborne transmission form of the fungus. It is still unknown whether it is the asci or the ascospores, or both, that are airborne. Nevertheless, they would exit the host's lungs and spread to another host in which ascospores would transform into trophic cells and restart the cycle.

The systematic presence of asci in each infection, their role in the spread of *P. jirovecii* as well as the presence of genes responsible for various steps of the sexual cycle supports the hypothesis that the sexual cycle is obligate for the fungus' transmission and survival (Hauser, 2021; Hauser & Cushion, 2018; Richard et al., 2018). This contrasts with the requirement of an asexual cycle which occurrence might be facultative, *i.e.* it might depend on the environment or particular conditions of the host (Hauser, 2021; Hauser & Cushion, 2018; Richard et al., 2018).

Genome of *P. jirovecii*

P. jirovecii has a reduced genome and a GC content lower than that of relative organisms. Indeed, its genome is ca. 8Mb on 17 to 20 linear chromosomes with a GC content of 29% versus 13Mb and 42%, respectively 14Mb and 36% for *Taphrina deformans* and *S. pombe* (Cissé et al., 2013; Ma et al., 2016). The whole genome sequencing of *P. jirovecii* was released first by Cissé et al. (2013), who thereupon observed a lack of virulence factors and of most of the enzymes usually dedicated to amino acid biosynthesis in fungi. The reduced genome and loss of essential enzymes support that *P. jirovecii* is an obligate parasite, *i.e.* it is strictly dependent on the host's cells because it scavenges from them numerous essential compounds to survive (Cissé et al., 2013; Hauser et al., 2010). Later, it was hypothesized that the common ancestor of all *Pneumocystis* species lost ca. 40% of its genes (Cissé et al., 2021). Ma et al. (2016) identified additional lacking pathways such as stress responses or

biosynthesis of lipids, as well as biosynthesis of some vitamins and carbohydrates. The latter is particularly interesting because *P. jirovecii* was found to lack critical enzymes for biosynthesis of chitin, which subsequently was confirmed to be absent on the fungus' surface. Chitin is usually considered as an essential component of fungal cell walls and is a common target for antifungals, such as nikkomycins, polyoxins, and plagiochin (Lima et al., 2019). *P. jirovecii* is the first species of the fungal kingdom identified to lack it. The main components of the cell wall of *P. jirovecii* are glycoproteins and β-glucan, however, the latter was only detected on the surface of the cyst form of the fungus as the trophs do not have a cell wall (Kottom & Limper, 2000; Kutty et al., 2015; Ma et al., 2016).

Msg superfamily

The most abundant proteins at the surface of *P. jirovecii* cells are the major surface glycoproteins (Msg; Kutty et al., 2008; Ma et al., 2016, 2020; J. R. Stringer & Keely, 2001). They form a superfamily of proteins including six distinct families, numbered I to VI (Schmid-Siegert et al., 2017). They are linked to the extracellular side of the cell wall by a glycosylphosphatidylinositol (GPI) anchor. Family IV is the only family suspected not to be bound to the fungus' surface as it lacks a GPI anchor at its C terminus (Ma et al., 2020; Schmid-Siegert et al., 2017). Family I is the most abundant, both in terms of protein quantity and number of genes, and is the only family missing a signal peptide at its N-terminus. Families I to III all possess a ST-rich region directly before their GPI anchor, which is a site of oxygen-linked glycosylation supposed to provide adhesive properties and rigidity to the protein (Schmid-Siegert et al., 2017). Another interesting aspect of several Msg families is the presence of cysteine as well as lysin residues at regular intervals in the peptide (Ma et al., 2016).

Cysteines are known to form covalent disulfide bonds between two of their residues. They have been shown to play a key role in the stability of the tertiary structure of extracellular proteins (Sevier & Kaiser, 2002). This suggests their importance in the structure of Msg proteins through intra- and inter-molecular bindings.

The function of the Msg is still largely unclear, but it is believed to be double. Firstly, as surface proteins they are thought to be essential for adhesion to host cells and extracellular matrix proteins, such as fibronectin and vibronectin (Pottratz et al., 1991; Pottratz & Martin, 1990). They also adhere to lung epithelial cells (Kottom, Hebrink, & Limper, 2018), as well as to various C-type lectin receptors present on immune cells (Ezekowitz et al., 1991; Kottom et al., 2017; Kottom, Hebrink, Jenson, et al., 2018; Sassi et al., 2018). Secondly, Msg proteins are thought to be responsible for the antigenic variation system of *P. jirovecii*, which would enable the fungus to escape the host immune system by rapidly changing its surface antigenicity (Keely et al., 2005; Keely & Stringer, 2009; Schmid-Siegert et al., 2017; J. R. Stringer, 2007).

msg genes

The major surface glycoproteins are encoded by multiple genes, each encoding a specific allele with a mean identity of 45% to 83% at the nucleotide sequence level between them, being 70% for family I (Schmid-Siegert et al., 2017). They represent 8% of the fungus' genome. This is surprising considering their highly reduced and compact genome of these fungi. This high proportion underlines the importance of the Msg superfamily (Cissé et al., 2021; Ma et al., 2016). The genes encoding the six families are clustered within the subtelomeres of the ca. 17 to 20 chromosomes of *P. jirovecii* with a conserved order. The genes of family I are located closest to the telomeres, while those of family VI are closest to the genomic genes on the side of the

centromeres. The other families are located in between. The open reading frames (ORF) of all these genes are encoded in the same direction, towards the telomeres. Multiple copies of the genes of each family are present in the subtelomeres, five to 20 according to the family, and 80 for family I. Each gene of the families II to VI are preceded by a promoter which suggests they can be continuously and simultaneously expressed (Schmid-Siegert et al., 2017). The expression levels vary among the families, family I representing ca. 85% of the *msg* expressed genes, *msg*-III ca. 10%, and each of the other families ca. 1% (Schmid-Siegert et al., 2021).

Each *msg*-I gene is ca. 3.1 kb long and multiple allelic genes are present in each subtelomere (Schmid-Siegert et al., 2017), totalling ca. 80 genes per genome. A strong promoter could be identified in a single copy per genome for the *msg*-I family. It is located upstream of the start codon within the upstream conserved sequence (UCS), which also includes the protein start and sequences responsible for post-translational translocation (Edman, 1996; Kutty et al., 2013). The presence of a single copy promoter suggests a mutually exclusive expression system, in which only the allele downstream of the promoter is expressed (Figure 4). Another peculiar aspect of the genes of family I is the presence of a 33-bps sequence called the conserved region junction element (CRJE) in front of each gene and at the end of the UCS. The 3’ end of this CRJE encodes a lysin-arginine sequence, which is a recognition site for kexin proteases. Such an enzyme could be involved in the maturation of the protein by cleaving the CRJE and releasing the constant portion of the protein encoded by the UCS (Keely et al., 2005; Kutty et al., 2001; Ma et al., 2016; Schmid-Siegert et al., 2017).

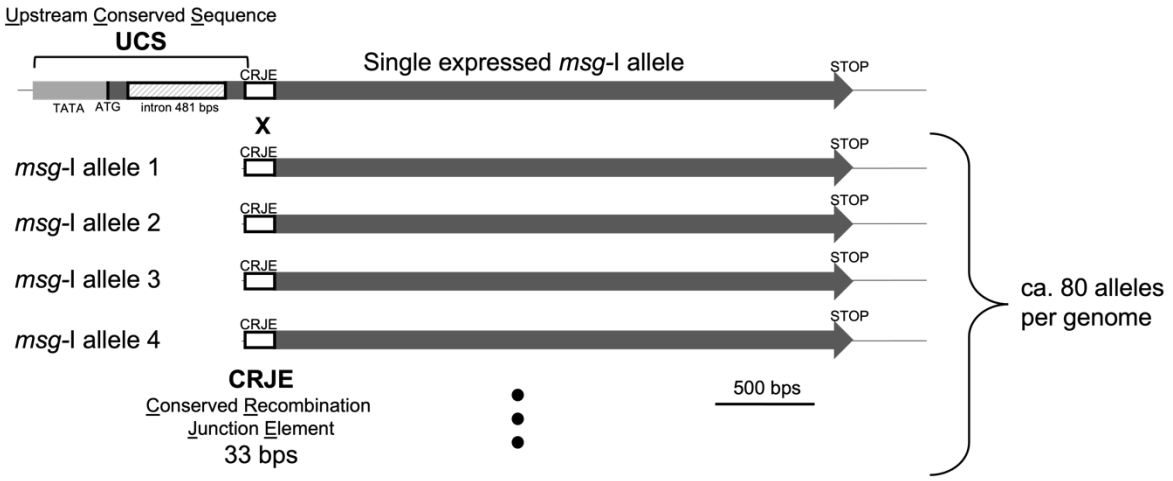


Figure 4 . Scheme of the mutually exclusive expression system of *msg*-I genes located at the subtelomeres of the chromosomes of *P. jirovecii*. A single promoter, located within the UCS, is present in the genome for ca. 80 *msg*-I. The cross represents the putative recombination at the CRJE.

Antigenic variation

Antigenic variation is an important mechanism present in various parasite organisms, such as *Trypanosoma brucei*, *Borrelia burgdorferi* (Vink et al., 2012) and *Plasmodium falciparum* (Freitas-Junior et al., 2000). It enables the pathogen to evade the host immune system by changing the epitopes at its outer surface protein layer. However, the strategies vary significantly between species (Vink et al., 2012). They incorporate genetic, epigenetic and/or expressional mechanisms (Deitsch et al., 2009).

In *P. jirovecii*, it is believed that the antigenic variation system is occurring at the genetic level, mainly using *msg* family I, in order to create a hypervariation of the Msg proteins. Two mechanisms are thought to be involved. Firstly, gene mosaicism of all *msg* families through intragenic recombinations, which results in exchange or conversion of gene fragments within each *msg* family (Keely et al., 2005; Kutty et al., 2008; Schmid-Siegert et al., 2017). Secondly, *msg*-I have a mutually exclusive expression system, as explained above. The small CRJE sequence is believed to have

a crucial role. Recombinations are believed to happen within it, resulting in the exchange of the expressed gene (Kutty et al., 2008).

The frequent exchange of the expressed *msg*-I would generate antigenically heterogeneous populations of *P. jirovecii* composed of several subpopulations of cells, each expressing a different *msg*-I gene and thus with a different Msg exposed at its surface. This diversity would challenge the host's adaptive immune system. However, as *P. jirovecii* probably evolved in immunocompetent individuals, the aim of its antigenic variation system might be the survive of the fungus long enough to spread rather than to overcome the immune system to cause PCP (Keely et al., 2007; J. R. Stringer, 2007). When confronted to a weaker host defence, *P. jirovecii* can evade the impaired immune response, reproduce and accumulate, which can lead to PCP.

Immune system response

Both the innate and the adaptive immune system are required to clear *P. jirovecii* infections (Evans et al., 2016). The innate response is activated by various surface proteins such as β-glucans and Msgs, and the subsequent activation of alveolar macrophages is a major player in the fungus clearance (Limper et al., 1997). The adaptive response uses CD4+ T and B cells with a focus on the interaction between these cells (Beck et al., 1993; Charpentier et al., 2021; Opata et al., 2015). Anti-Msg antibodies produced by the activated B cells are measurable after an infection with *P. jirovecii* (Bishop & Kovacs, 2003; Daly et al., 2006). Additionally, the evasion of the immune system by antigenic variation affects the humoral response by imposing a constant adaptation of the antibodies to the current Msg(s) expressed on the surface of the fungal cells (Deitsch et al., 2009; Vink et al., 2012).

Projects and aims of the PhD thesis

This thesis encompasses two chapters :

Chapter 1: Fungal antigenic variation using mosaicism and reassortment of subtelomeric genes' repertoires, potentially mediated by DNA triplexes

Chapter 2: Clusters of patients with *Pneumocystis jirovecii* pneumonia harbouring similar *msg*-I repertoires

The main aim of this thesis was to better understand the genetic mechanisms involved in the antigenic variation system of *P. jirovecii*. For this purpose, we characterized the *msg*-I gene repertoires present in *P. jirovecii* strains causing active pneumonia in immunocompromised patients from multiple cities, and compared them. Chapter 1 is a manuscript reporting this analysis that is presently submitted for publication.

During the analyses presented in chapter 1, three clusters of patients harbouring similar *P. jirovecii msg*-I repertoires were detected. Chapter 2 is a manuscript in preparation that reports the analysis of these three clusters.

Chapter 1: Fungal antigenic variation using mosaicism and reassortment of subtelomeric genes' repertoires, potentially mediated by DNA triplexes

Summary of the results

The mechanisms involved in the antigenic variation system of *P. jirovecii* believed to allow escape from the host immune system are still poorly understood. Amplification of all entire *msg*-I genes, followed by a cutting edge sequencing method *i.e.* PacBio circular consensus sequencing (CCS), enabled the characterization and comparison of the *msg*-I repertoires among 24 patients from five cities.

The characterization of these repertoires resulted in the identification of 1007 *msg*-I alleles. Ca. 50 % of them were present in multiple cities, whereas 88 % of the other half found in a single city were only present in a single patient, suggesting the need of both inter- and intragenic recombinations to achieve such diverse repertoires. Further observations included (i) the identification of duplicated fragments within *msg*-I genes, which supported an intragenic recombination mechanism; (ii) the detection of a higher identity between intergenic spaces than the genes themselves, suggesting favoured intergenic recombinations relative to intragenic ones; (iii) the detection of interfamily similarities upstream of *msg*-II and *msg*-III genes, which suggested single intergenic recombinations between those two families; and (iv) the study of the structure of the sequence CRJE, present in front of each allele, which has a peculiar mirror repeat sequence suggesting the formation of a DNA triplex, known to promote recombination events.

These observations led to the creation of a more complete model of mechanisms leading to antigenic variation in *P. jirovecii*, including inter- and intra-genic recombinations; possible exchange of the downstream telomere additionally to the genes; and possibly promoted by the CRJE.

Contributions to the manuscript

My contributions to this paper under the supervision of P. Hauser are the following:

- Optimization of the amplification and preparation of the samples for the sequencing
- Development of a bioinformatics pipeline for the analyses of the results (under the supervision of M. Pagni of the Swiss Institute of Bioinformatics)
- Writing of the draft
- Creation of the figures

Fungal antigenic variation using mosaicism and reassortment of subtelomeric genes' repertoires, potentially mediated by DNA triplexes

Caroline S. Meier[1], Marco Pagni[2], Sophie Richard[1], Konrad Mühlethaler[3], Joao M. G. C. F. Almeida[4], Gilles Nevez[5,6], Melanie T. Cushion[7,8], Enrique J. Calderón[9], Philippe M. Hauser[1*]

[1] Institute of Microbiology, Lausanne University Hospital and University of Lausanne, Lausanne, Switzerland

[2] Vital-IT Group, SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland

[3] Institute for Infectious Diseases, University of Bern, Bern, Switzerland

[4] UCIBIO - Applied Molecular Biosciences Unit, NOVA School of Science and Technology, Universidade NOVA de Lisboa, 2829-516 Caparica, Portugal

[5] Laboratoire de Parasitologie et Mycologie, Hôpital de La Cavale Blanche, CHU de Brest, Brest, France

[6] Infections respiratoires fongiques (IFR), Université d'Angers, Université de Brest, France

[7] Department of Internal Medicine, Division of Infectious Diseases, College of Medicine, University of Cincinnati, Cincinnati, OH 45221, USA

[8] Cincinnati VAMC, Medical Research Service, Cincinnati, OH 45220, USA

[9] Instituto de Biomedicina de Sevilla, Hospital Universitario Virgen del Rocío/Consejo Superior de Investigaciones Científicas/Universidad de Sevilla, and Centro de Investigación Biomédica en Red de Epidemiología y Salud Pública, Servicio de Medicina Interna, Hospital Universitario Virgen del Rocío, Facultad de Medicina, Seville, Spain

[*] Corresponding author: Philippe.Hauser@chuv.ch

Running title: Surface antigenic variation of *P. jirovecii*

Main text: 5787 words (introduction 528, results 3121, discussion 2138)

Materials and methods: 2280 words

Abstract: 249 words

Keywords: *Pneumocystis carinii*, major surface glycoprotein, adhesin, subtelomere, PCP, PJP, gene mosaicism, gene conversion, gene exchange, telomere exchange, PacBio sequencing, H-DNA triplex.

Abstract

Surface antigenic variation is crucial for major pathogens that infect humans, e.g. *Plasmodium* (Deitsch & Dzikowski, 2017), *Trypanosoma* (Navarro & Gull, 2001), *Giardia* (Prucca et al., 2008). In order to escape the immune system, they exploit various mechanisms in order to modify or exchange the protein that is exposed on the cell surface, at the genetic, expressional, and/or epigenetic level (Deitsch et al., 2009). Understanding these mechanisms is important to better prevent and fight the deadly diseases caused. However, those used by the fungus *Pneumocystis jirovecii* that causes life-threatening pneumonia in immunocompromised individuals remain poorly understood. Here, though this fungus is currently not cultivable (Liu et al., 2018), our detailed analysis of the subtelomeric sequence motifs and genes encoding surface proteins suggest that the system involves mediation of homologous recombinations during meiosis by DNA triplexes. This leads to the reassortment of the repertoire of ca. 80 non-expressed genes present in each strain, from which single genes are retrieved for mutually exclusive expression within subpopulations of cells (Schmid-Siegert et al., 2017). The recombinations generates also constantly new mosaic genes. Dispersion of the new alleles and repertoires, supposedly by healthy carrier individuals, appears very efficient because identical alleles are observed in patients from all over the world. Our observations reveal a unique strategy of antigenic variation allowing colonization of the non-sterile niche corresponding to lungs of healthy humans. They also highlight the possible role in genome rearrangements of small imperfect mirror sequences forming DNA triplexes (Mirkin & Frank-Kamenetskii, 1994). Such mirror sequences are widespread in eukaryotic genomes (Goñi et al., 2006b), as well as in HIV virus (Lang, 2007), but remain poorly understood so far.

Introduction

The fungus *Pneumocystis jirovecii* is an obligate biotrophic parasite that colonizes specifically the human lungs (Gigliotti et al., 2014). In immunocompromised patients, mostly HIV positive patients and transplant recipients, it causes a life-threatening pneumonia that is among the most frequent invasive fungal infections (Bongomin et al., 2017). Currently, the study of *P. jirovecii* biology is difficult due to the lack of a long-term *in vitro* culturing method.

This fungus lacks chitin, β-glucans, and outer chain N-mannans commonly present in fungal walls, which may help escape the host's immune responses during colonization (Ma et al., 2016). In addition, like other major microbes pathogenic to humans, it possesses a system of surface antigenic variation that appears essential for survival because it represents ca. 8% of its highly compacted genome (Keely & Stringer, 2009; J. R. Stringer & Keely, 2001). The most important player of this system is a superfamily including six families of major surface glycoproteins (Msg-I to VI (Ma et al., 2016; Schmid-Siegert et al., 2017)). These are supposed to be responsible for adherence to various human proteins present in the lungs and on macrophages (Ezekowitz et al., 1991; Kottom, Hebrink, & Limper, 2018; Pottratz et al., 1991; Pottratz & Martin, 1990). All Msgs are believed to cover asci and trophic cells that are present during proliferation, except those of family VI which could be present only at the surface of ascospores, as it is the case in *Pneumocystis murina* infecting specifically mice (Bishop et al., 2018). Family I is the most abundant in number of genes, transcripts, and proteins at the cell surface (Kutty et al., 2013; Ma et al., 2016; Schmid-Siegert et al., 2021). The genes encoding Msgs are located within all subtelomeres of the 17 to 20 chromosomes of *P. jirovecii*, the genes of family I being closest to the telomeres (Ma et al., 2016; Schmid-Siegert et al., 2017). The subtelomeric localization favours

ectopic recombinations within the meiotic bouquet of telomeres, gene silencing, and possibly mutagenesis (Barry et al., 2003). In addition, recombinations of genes in this genomic region presents the advantage to have no or small effect on the overall chromosome structure (J. R. Stringer & Keely, 2001). Genes of families II to VI each have their own promoter and could be constitutively and simultaneously expressed (Schmid-Siegert et al., 2021). On the other hand, a single out of the ca. 80 genes of family I is believed to be expressed at a time in a cell thanks to the presence of a single copy promoter in the genome, within the so called upstream conserved sequence (UCS) (Edman, 1996). At the end of the UCS, a 33 bps-long sequence is present, the conserved recombination junction element (CRJE), which is also existing at the start of each of all *msg*-I alleles (Keely et al., 2007). It is most probably the preferred site of recombination allowing the exchange of the downstream expressed allele. The spontaneous exchange of the single *msg*-I gene expressed per cell is thought to create subpopulations of cells, each expressing a different *msg*-I allele. A second mechanism of antigenic variation relies on intragenic recombinations of the *msg* gene sequences, *i.e.* gene mosaicism (Kutty et al., 2008; Ma et al., 2016, 2020; Schmid-Siegert et al., 2017).

The present work aimed at better understanding the mechanisms involved in the antigenic variation system of *P. jirovecii* by investigating the repertoires of the *msg*-I genes present in patients from different geographical locations. Our observations allow us to complete the model for the surface antigenic variation system of *P. jirovecii*.

The present work was submitted by C. Meier as partial fulfillment of a Ph.D. degree at the Faculty of Biology and Medicine of the University of Lausanne.

Methods

Ethics approval and consent to participate

In Lausanne, Bern, and Seville, the procedure for admittance in the hospital included informed written consent for all patients. The admittance form included the possibility to require their samples not to be used for research. The samples were obtained through the hospital's routine procedure and were anonymized. The research protocol was approved by the Seville Hospital review board and the Swiss institutional review board (Commission Cantonale d'Éthique de la Recherche sur l'Être Humain, http://www.swissethics.ch). The collection and use of archival specimens was approved by the ethics committee of Brest University Hospital on June 24th 2021, and registered with the French Ministry of Research and the Agence Régionale de l'Hospitalisation, No. DC-2008-214. The samples from Cincinnati were anonymized; their use was not for research on human subjects and did not require approval by the Institutional Review Board.

Samples and DNA extraction

Broncho-alveolar lavage (BAL) samples were collected from 24 immunocompromised patients with *Pneumocystis* pneumonia in five different geographical locations over 25 years (Table S1). DNA was extracted from 0.2 ml of each BAL using the QIAamp DNA Blood Mini Kit (Qiagen).

Amplification of the repertoires of *msg*-I genes

Two distinct generic PCRs were used to amplify all entire *msg*-I genes that are expressed (the "expressed repertoire"), or all entire *msg*-I genes of the genome that

are present in the sample, *i.e.* the expressed plus the non-expressed genes (the "complete repertoire"). The expressed repertoire was specifically amplified thanks to a forward primer localized at the end of the UCS sequence, 27 bps upstream of the CRJE (GK135: 5' GACAAGGATGTTGCTTTTGAT 3') (Kutty et al., 2001). The complete repertoire was amplified specifically using a forward primer covering the 3' half of the CRJE sequence (CRJE-for-bis: 5' TGGCGCGGGCGGTYAAG 3'; the underlined Y was introduced because 87 and 13% of the CRJEs harbor respectively T and C at this position). The reverse primer was identical for both PCRs and located in a conserved region of 31 bps ca. 90 bps after the stop codon of the *msg*-I genes (GK452: 5' AATGCACTTTCMATTGATGCT 3'; the underlined M was introduced because ca. 90 and 10% of the sequences harbor respectively T and G at this position) (Kutty et al., 2008). PCR was performed with one microliter of DNA from the BAL in a final volume of 20 µl containing 0.2 µl polymerase (KAPA LongRange HotStart, Roche), the provided buffer, each dNTP at 0.2 mM, each primer at 0.5 µM, and a final $MgCl_2$ concentration of 3 mM. In order to prevent contaminations, PCRs were set up and analysed in physically separate rooms using materials dedicated to each room, and negative controls were systematically carried out at each experiment.

A touchdown PCR procedure was used for both PCRs. To amplify the expressed repertoire, the program began with 10 cycles consisting in a constant decrease of the annealing temperature from 62°C to 55°C, followed by 25 or, if needed to obtain a sufficient amount of PCR product, 30 cycles with annealing at 55°C. For the complete repertoire, the decrease was from 68°C to 58°C in 10 cycles, followed by 20 or 25 cycles with annealing at 58°C. The elongation time was for both PCRs 3 min at 72°C. The reactions finished with a final elongation of 7 min at 72°C. The PCR products were verified by (i) their size of ca. 3100 bps on an agarose gel and (ii)

subcloning some of them using the TOPO cloning kit (Invitrogen) followed by Sanger sequencing. The PCR products were then purified with E-gel CloneWell 0.8% (Invitrogen) according to the manufacturer's instructions, except for the use of 50% glycerol instead of the provided loading buffer to avoid any issues during the sequencing due to its unknown composition. Finally, they were evaporated on a heating block at 50°C with the tube lid open to reach the adequate concentration required for PacBio circular consensus sequencing. Each open tube was covered with a non-airtight lid during evaporation in order to prevent cross-contamination. Absence of the latter was assessed by the presence of a single allele in the plasmid controls upon sequencing.

PacBio circular consensus sequencing (CCS)

Single molecule real time sequencing of the PCR products was performed using the PacBio Sequel II at the Genomic Technologies Facility of the University of Lausanne, Switzerland. Barcodes were added to the amplicons before pooling them. They were then circularized to enable multiple sequencing of the same *msg*-I gene and the generation of circular consensus sequences. This technology provides long and accurate reads of the repetitive sequences present in the *msg*-I genes. The sequencing of 50 samples (48 PCR products from the 24 samples, plus two plasmid controls) resulted in 10'223 to 160'836 reads per sample.

PCR artefacts

The production of artefacts during PCR amplification of mixed alleles results in additional non-biological diversity (Beser et al., 2007). In order to address the occurrence of this problem in our conditions, two plasmids containing each a single

specific *msg*-I allele were mixed before or after PCR amplification, and then sequenced using PacBio CCS. Chimeric amplicons were observed upon mixing before PCR, but not after PCR. These PCR artefacts result probably from a premature detaching of the DNA polymerase during the elongation step that produces incomplete strands. The latter can then be used as primers by annealing to closely related sequences. The putative extremities of the incomplete strands were localized in the first and last 500 bps of the chimeric sequences. These findings led to the decision to trim both ends of the ca. 3.1kb sequences and to keep only the central 2kb for the subsequent analyses (positions 500 to 2500).

Allele identification and quantification

A dedicated bioinformatics pipeline of analysis of the Pacbio CCS raw reads was developed to identify the alleles present in each sample as well as their abundance. Information about the versions of the software and its packages are provided in Table S2. The multistep process consisted in the following sequence of analyses:

1. **hmmer** (http://hmmer.org/, n.d.) was used to identify the reads corresponding to *msg*-I genes by aligning them to a profile-HMM specifically generated using 18 *msg*-I genes previously identified in a pilot experiment and trimmed after the CRJE in order to be in frame. The nhmmer program command was used to filter the reads by alignment length and bit score, which resulted in removal of ca. 5% of the reads and a final number of 9’742 to 156’375 reads per sample for further analysis.
2. **swarm** (version 3.1.0) (Mahé et al., 2014) was used to define cluster seeds for each sample. Each read that is more than 1 bp (option -d 1) different from any

other is defined as a new cluster seed by swarm. The seeds with more than one read in their cluster were kept as cluster seeds for the following clustering step.

3. **cd-hit** (version 4.8.1) (Li & Godzik, 2006) was used to cluster the reads around the cluster seeds defined by swarm. The command cd-hit-est-2d, with options -c 0.99 and -g 1, allowed the allocation of each read to the most similar seed to create clusters of similar reads and determine the abundance of the alleles that were identified among the reads as described in the next section.

The fine-grained clustering obtained so far still accounts for PacBio sequencing errors and an additional step of clustering was required to get rid of this problem. Two plasmids containing each a single specific *msg*-I allele, one with the UCS and the other without the UCS, were amplified by PCR, PacBio CCS sequenced, and clustered as described above, yielding several seed sequences. However, all pairwise sequence identities between the seed sequences from a single plasmid were higher than 99.5%, which matches the expected error rate of the PacBio CCS sequencing (Wenger et al., 2019). An additional round of clustering was performed with an identity threshold set at 99.5%. This yielded clusters of reads that were considered as identical alleles, or alleles that cannot be distinguished by the used sequencing technique. This corresponds to a difference of less than 10 bps in the trimmed 2kb sequences.

Confirmation of the *msg*-I alleles repertoires present patients using specific PCRs

The repertoires observed were supported using PCRs specific to three given alleles among five patients (Table S3, Figure S1). Three alleles were selected that were present in several patients and absent in others. Primers were designed within

these alleles to amplify specifically each of them (Table S3). The specificity of the primer pairs were assessed by blasting them against the 1007 different *msg*-I alleles identified in the present work, as well as against the whole nucleotide collection (nr/nt). The PCRs were performed with the following parameters: 3 min at 94°C followed by 35 cycles consisting of 15 sec at 94°C, 30 sec at 56°C, and 2 min at 72°C, followed by a final extension of 7 min at 72°C. The presence or absence in patients observed by PacBio CCS was confirmed by the specific PCRs (Figure S1). The sequences of the amplicons obtained using the Sanger technology showed 100% identity with those obtained using CCS. The positive DNA control was produced by random amplification of 2.5 µl DNA from the BAL of patient LA2 using the GenomiPhi HY kit (GE Healthcare), followed by a purification step using the columns of the QIAamp DNA Blood Mini Kit (Qiagen). It was also used for the setup of the PCR reactions.

Estimation of the number of *P. jirovecii* strains

The number of *P. jirovecii* strains infecting each patient was estimated by amplifying and sequencing the region comprising the internal transcribed spacers and the 5.8S rRNA gene of the ribosomal RNA operon (ITS1-5.8S-ITS2). The PCRs mixes were identical as for the amplification of the *msg*-I genes with one microliter DNA in a total volume of 20 µl. The PCRs were performed with the following parameters: 3 min at 95°C initial denaturation followed by 35 cycles consisting of 30 sec at 95°C, 30 sec at 62°C and 45 sec at 72°C followed by a final elongation of 1 min at 72°C. The primers were derived from those described by reference (Xue et al., 2019) : 5’ GCTGGAAAGTTGATCAAATTTGGTC 3’ and 5’ TTCGGACGAGACTACTCGCC 3’ (the six underlined bases were added in 5’ to the first primer to adjust the annealing temperature, and the four underlined bases replaced those from *Pneumocystis carinii,*

the species infecting rats, that were in fact within the second primer). The PCR program included 3 min at 95°C of initial denaturation, 30 to 35 cycles consisting of 30 sec at 95°C, 30 sec at 62°C, and 45 sec at 72°C, followed by 1 min of final elongation at 72°C. After purification using the Qiagen PCR column kit, the amplicons were sequenced using PacBio CCS. The bioinformatic pipeline dedicated to the analysis of the ITS1-5.8S-ITS2 reads started with hmmer to identify the correct genes followed by a filtering step by read size between 470 and 500 bps (see above, Allele identification and quantification). Xue et al. (Xue et al., 2019) highlighted that the ITS1-5.8S-ITS2 region includes seven homopolymer stretches that are prone to errors during both amplification and sequencing. Hence, the length of the six homopolymers that showed variation in our data were homogenized (Figure S2). The redundant reads were then clustered using swarm (option d0) and only the clusters comprising more than 1% of the reads present in each patient were analysed.

BALSTn analyses

We searched similarities within the 37 subtelomeres present in the whole *P. jirovecii* genome assembled from a single strain using PacBio sequencing (Schmid-Siegert et al., 2017). We used the BLASTn algorithm on the NCBI website (https://blast.ncbi.nlm.nih.gov/Blast.cgi) with the default parameter values (expected threshold 0.05) and the following settings: database: whole-genome shotgun contigs (wgs), limit by: BioProject, BioProject name: 382815 *Pneumocystis jirovecii* strain E2178, program selection: somewhat similar sequences (BLASTn). The 37 contigs carrying *msg* genes are entire or partial subtelomeres corresponding to most of the 34 to 40 that are present in *P. jirovecii* genome (17 to 20 chromosomes) (Ma et al., 2016). Nevertheless, 17 of the 37 contigs do not carry non-*msg* genes flanking the

subtelomere, so that they could not be attributed to a specific chromosome previously described (Ma et al., 2016). Consequently, some of them could be parts of the same subtelomere, rather than distinct ones. For simplification, each contig is considered as one specific subtelomere in the present manuscript.

Search of mosaicism within the *msg*-I alleles

Duplicated fragments were searched within the 917 alleles identified in the complete repertoires of this paper. An allele chosen randomly was blasted using BLASTn with default parameters to align two or more sequences among the 3.1kb sequences of the 917 alleles (the full size alleles were used because search of duplicated fragments is not impaired by eventual chimera). Its closest allele, *i.e.* that with the highest score, was chosen for the following steps. Alignment of both sequences allowed the visual identification of fragments larger than 100 bps shared by the two alleles. These fragments were then searched among the 917 alleles using BLASTn, identifying the ones containing the fragments completely conserved.

Bioinformatics search for site-specific recombinase genes

Potential genes encoding site-specific recombinases were searched for in the *P. jirovecii* genome (accession number LFWA01000000 by matching this genome against large pools of representative bait sequences by using tBLASTn (NCBI BLAST suite). These pools were recruited through the InterPro annotations IPR011010 (DNA breaking-rejoining enzyme, catalytic core) from a wide range of taxa. To avoid missing candidates because of the use of too stringent conditions, the tBLASTn searches were conducted with relaxed parameters (E value from 1E-4 down to the default value). Each match with a suitable E value was investigated by looking for pre-existing

annotations. If no coding sequence (CDS) annotations were available, the matched region was assessed for putative novel CDSs and their translated sequence were submitted to the InterProScan4 tool to detect the required reference signature matches (Paysan-Lafosse et al., 2023).

Results

Amplification and identification of the *msg*-I alleles

We amplified the *msg*-I alleles present in the BALs of 24 patients with *Pneumocystis* pneumonia from five different geographical locations (Table S1). Generic primers were used in two different PCRs to amplify specifically either all entire *msg*-I genes both expressed and non-expressed (hereafter called the “complete repertoire”), or all entire expressed *msg*-I genes (the “expressed repertoire”). The PCR products were sequenced using PacBio circular consensus sequence (CCS). The reads were processed using a specifically developed bioinformatics pipeline dedicated to the identification of the different biological alleles present, as well as to the determination of their abundance. The known issues represented by chimeric sequences created during PCR amplification and PacBio sequencing errors were specifically addressed and are not believed to affect the results presented below (see methods). The diversity of the alleles reported might however be underestimated.

msg-I alleles identified in the patients

Among the 48 PCR products from the 24 patients, 1007 distinct *msg*-I alleles were identified. They had a mean pairwise sequence identity of 65.7% ± SD 9%. This value is in agreement with that based on 11 alleles from Switzerland that we published previously (71% ± SD 7%) (Schmid-Siegert et al., 2017), as well as with that of 70.5% ± SD 3% among a collection of 80 *msg*-I genes from USA (Ma et al., 2016). If we define pseudogenes as alleles with at least one stop codon, they represented 5.6% of the 1007 alleles. This is consistent with the single previous observation of 11% among 28

genes with a CRJE (Schmid-Siegert et al., 2017). Thus, 94.4% of the 1007 alleles presented a fully open reading frame without any introns.

The 24 complete and 24 expressed repertoires included respectively 917 and 538 distinct alleles that were sorted using hierarchical classification trees (Figure 5). Both trees highlighted the presence of two major and one smaller subgroups of the *msg*-I alleles, which is consistent with previous observations (Kutty et al., 2008; Ma et al., 2016; Schmid-Siegert et al., 2017). The two major subgroups constitute a single family, based on the occurrence of recombinations between them (Kutty et al., 2008; Schmid-Siegert et al., 2017). The smaller one corresponds to outlier *msg* genes that could not be yet classified into family I (Schmid-Siegert et al., 2017). The significance of these subgroups remains unexplained so far.

Repertoires of *msg*-I alleles present in the patients

The alleles of both the expressed and complete repertoires present in each patient were uniformly spread along the trees of the alleles (Figure 5). No repertoire showed clear groups of alleles that would have revealed the presence of related ones. The complete repertoires contained 44 to 185 alleles per patient (104 ± SD 40), whereas 2 to 108 alleles were present in the expressed repertoires (37 ± SD 34, Table S4). Notably, three out of the five samples from Brest (BR1, BR2, BR3) harbored the least diverse expressed repertoires with 2, 3, and 5 alleles. The variation of the number of alleles of the complete repertoires was at least partially explained by the number of *P. jirovecii* strains present in the patients. Indeed, a significant correlation was observed between these numbers (Figure 6). On the other hand, the expressed repertoires showed no correlation, suggesting the involvement of other more important parameters. Figure 6 also shows the consistency of the data. Indeed, the correlation

was 0.74 between the number of alleles of the complete repertoires and that of strains. Moreover, the average of 77.6 (intercept + regression slope) alleles in samples infected by a single strain obtained by regression is compatible with an underestimation of the postulated number of 80 per genome.

Due to the design of the two PCRs, the expressed repertoire should be a subset of the complete repertoire for each patient. Consistently, high proportions of the alleles of the expressed repertoires were also present in the corresponding complete repertoires (85% ± SD 17%, 40 to 100%, “% expressed in complete”, Table S4). Eighteen out of the 24 patients presented a proportion lower than 100% probably because of the limitations that affect the composition in low abundant alleles of the repertoires, and that lead to a slight underestimation of the number of these alleles (Supplementary data 1, Table S5, Figure S3). Reversely and as expected, lower proportions of the alleles of the complete repertoires were present within the corresponding expressed repertoires (33% ± SD 32%, 2 to 100%, “% complete in expressed”, Table S4). These latter proportions are consistent with the single previous estimation of 50% among 28 genes (Schmid-Siegert et al., 2017), they correspond to the alleles that are both non-expressed and expressed.

Similarity of the *msg*-I repertoires between the patients

The complete and expressed *msg*-I repertoires present in the patients were also sorted using hierarchical classification trees (Figure 5). The inspection of these trees reveals that each repertoire was notably different from all the others. This distinctiveness implies the absence of obvious correlation of the repertoires with the year or city of collection of the sample, the underlying disease affecting the patient, or the *P. jirovecii* genotype(s) causing the infection. The complete repertoires of two

samples from Seville, SE1 and SE2, as well as the complete and expressed repertoires of the two samples from Cincinnati, CI1 and CI5, were possible exceptions as they were slightly related.

Although they were all distinct, the repertoires shared many alleles. Indeed, 84% ± SD 7% of the alleles of each complete repertoire were present in at least one complete or expressed repertoire of another patient. The value was 77% ± SD 16% for the expressed repertoires. Figure 7 gives the proportion of alleles shared by each repertoire, and allow visualizing that the overlaps of the repertoires did not differ significantly according the city and continent, as well as of the year of collection.

Distribution of the *msg*-I alleles among the cities and patients

Approximately half of the 917 alleles present in the 24 complete repertoires were found in only one city, of which the vast majority were in a single patient (88.0%, 411 among 467, Table 1a). The remaining half was present in more than one city, 3.6% occurring even in all five cities. The proportion of the alleles found in only one city among the 24 expressed repertoire was higher than in the complete repertoires (72.9%), but a similarly high proportion of which was found in a single patient (88.5%, 347 among 392, Table 1b). A single expressed allele was present in all five cities (0.2%). Figure 8 allows the visualization of these proportions and shows that comparable results were observed in each of the five cities.

In order to understand the parameters influencing the distributions of the alleles among the patients, we performed two simulations experiments *in silico*. First, we investigated the size of the reservoir from which the alleles are retrieved. We simulated the 24 complete repertoires by drawing 24 times the 104 alleles present on average in them out of a simulated reservoir comprising 1’000 to 5’000 alleles. We repeated this

draw 30 times, and then determined the mean number of draws of each allele. The distribution obtained with the reservoir including 2'000 alleles was most similar to that observed in the complete repertoires of our data (Figure S4a). Similarly, drawing 24 times 37 alleles to simulate the 24 expressed repertoires, the distribution with a reservoir of 1'000 alleles resembled that we observed for the expressed repertoires (Figure S4b). In the second simulation experiment, we determined the effect of analysing less than the 24 patients. We drew 30 times 5, 10, 15, or 20 patients randomly, and calculated the mean numbers of alleles observed. Consistently with the first simulation experiment, the numbers of alleles increased regularly, showing that a plateau corresponding to the complete reservoir was not reached with the analysis of 24 patients (Figure S5). These simulations suggest that the high proportion of alleles we observed only once in single patients could be explained by a large reservoir of alleles. However, a not mutually exclusive hypothesis is that a proportion of them corresponds to new alleles created by mosaicism within each patient because this is probably necessary for the survival of the fungus.

Abundance of the *msg*-I alleles in the patients

Each population of *P. jirovecii* cells is expected to be composed of subpopulations, each expressing a distinct *msg*-I allele. The size of these subpopulations, and thus the abundance of the expressed alleles, may vary according to a possible advantage over the host immune system, or to other parameters. To test this hypothesis, the abundance of an allele was defined as its number of reads in percent of the total number of those present in the given repertoire. These abundances were confirmed using subcloning in the wet-lab (Supplementary data 2, Table S6). All the 24 complete repertoires showed a homogenous distribution of the allele

abundances which remains under a maximum of 8% (Figure 9). By contrast, 22 out of the 24 expressed repertoires, *i.e.* except those of LA6 and BE1, showed a heterogeneous distribution of these abundances, with up to 6 highly abundant alleles (≥ 8%), and 107 low abundant alleles (< 8%). The highest abundance was seen in patient LA7 at 71.9%. Among the 62 expressed alleles with an abundance ≥ 8%, 40 were observed in more than one patient (64.5%). This contrasted drastically with the same proportion among all expressed alleles, including those at < 8% (35.5%, 100 minus 64.5%, Table 1b). This difference suggests that the alleles at ≥ 8% might have presented a selective advantage, for example over the host immune system of the individuals that harboured them.

The technical variability of the expressed repertoires characterization is the major limitation of the present study (Supplementary data 1, Table S5, Figure S3). Nevertheless, the heterogeneity of the allele abundances concerned both duplicates of all four expressed repertoires that we analysed (Figure S3b). Moreover, all 11 alleles with an abundance ≥ 8% in at least one duplicate were present in both duplicates, six of them being at < 8% in the other duplicate. Thus, despite the variation of the expressed abundances due to the limitation of the method, the distributions of the abundances differed clearly between the complete and expressed repertoires. This further supports the mutually exclusive expression hypothesis. Indeed, the expressed alleles are likely to correspond to subpopulations expressing each a specific allele, the most frequent alleles by the largest ones.

Sequences flanking the *msg* genes

The *msg*-I genes present short conserved sequences of a length of ca. 30 bps before and after their CDS (coding sequence), the CRJE and 31 bps located after the

stop codon, in which we placed our primers for amplification of the complete repertoires. To investigate if similar conserved sequences flank the *msg* CDSs of the other families, we took advantage of the 37 subtelomeres that we previously assembled from a single strain using the PacBio sequencing technology (Figure S6 and Figure S7) (Schmid-Siegert et al., 2017). These subtelomeres carry 113 distinct *msg* genes of the six *msg* families. We used the 200 bps located immediately up- or downstream of 20 representative CDSs as queries in BLASTn analyses against the PacBio *P. jirovecii* genome assembly. Most of the genes of families I to V produced numerous significant hits with other sequences that flanked a CDS of a gene or pseudogene of the same family, most often both up- and downstream (Table 2, Table S7a). The hits represented variable proportions of the up- or downstream sequences of the same family present in the assembly (20 to 100%). Importantly, for families I and IV, the identities of these hits with the query were ca. 10% higher than those between the genes themselves (Table S7b). On the other hand, these identities were similar between families II, III, and V. Notably, for family I, fewer hits were found for the pseudogenes than for the genes (7 to 38% versus 67 to 79 %). Inspection of the alignments with the hits for families II to V did not identify any conserved sequences flanking the CDSs, contrary to family I. These analyses revealed that the up- and downstream regions of *msg* CDSs are often similar among the genes of the same family, except for family VI, and that those of families I and IV are even more conserved than the genes nearby.

These analyses also revealed that the upstream sequences of the genes of families II and III are often similar, but not their downstream sequences (Table 2). The identities between these hits and their query were much higher than those between the genes nearby (Table S7b, mean difference of 29.8% ± SD 5.9%). These hits

included an important proportion of the upstream sequences of the other family (50 to 78%). Besides, genes no. 3, 37, 55 and 8 of these two families showed no or lower similarities with the other family (Table 2). The latter genes are located centrally in the subtelomeres, whereas the others are located at their extremities (genes no. 7, 25, 34, 53, Figure S8). Furthermore, the similarity between the 200 bps upstream of genes of families II and III proved to extend within the CDS on ca. 100 bps (Figure S8). In conclusion, the upstream regions of the CDSs of *msg* families II and III are often similar and more conserved than the genes nearby, in particular among genes located at the distal tip of the subtelomeres, but not their downstream regions.

Mosaicism of the *msg* genes

The mosaicism of *msg* genes, *i.e.* being composed of fragments potentially originating from other genes, was suggested based on the detection of recombinations and duplicated fragments strictly among members of the same family (Kutty et al., 2008; Schmid-Siegert et al., 2017). We investigated the possible mosaicism of the 917 *msg*-I alleles identified in the complete repertoires during the present study. Inspection of the alignments of three randomly chosen alleles with their closest hit in BLASTn comparison identified in each case four to eight duplicated fragments of a size ≥ 100 bps (total of 16 fragments, one alignment is shown in Figure S9). We then searched each fragment within the 917 alleles using again BLASTn comparison. It appeared that the fragments were conserved among closely related alleles, *i.e.* within the same subgroup of alleles of the tree (Figure 10). They were conserved in several patients, regardless of the year of the pneumonia episode, or city and continent of origin. Moreover, the two fragments < 200 bps were present within two different subgroups of alleles (blue and purple in Figure 10). In the single case of the fragment labelled in

purple, the two subgroups corresponded to very distant alleles because they belonged to the two major subgroups of *msg*-I alleles (see above, section "*msg*-I alleles identified in the patients"). By contrast, the fragment ≥ 200 bps shown in red distributed in only one subgroup. Identical distributions in function of the size of the duplicated fragments were observed for all 16 duplicated fragments, ten ≥ 200 bps and six < 200 bps, two among the latter being nevertheless present only in a single subgroup of alleles.

To understand better the phenomenon, we searched thoroughly duplicated fragments within the 10 representative subtelomeres of the 37 previously assembled (Schmid-Siegert et al., 2017) shown in Figure S6. The results confirm the mosaicism of the *msg* genes and pseudogenes previously reported (Schmid-Siegert et al., 2017), reveal that the intergenic spaces are also concerned, and suggest that no hotspots of recombination exist along the *msg* genes (Supplementary data 3, Table S7 and Table S8, Figure S6, Figure S7 and Figure S10).

Structure of the CRJE sequence present at the beginning of each *msg*-I gene

The sequence CRJE of 33 bps that is conserved at the beginning of each *msg*-I gene is probably the site of recombination allowing the exchange of the expressed allele. It is specific to *P. jirovecii*, and we did not identify any site-specific recombinase that could target it (Supplementary data 4 and 5). We identified two important features of the CRJE.

First, each strand of the CRJE is enriched in purines or pyrimidines that are part of an imperfect mirror repeat (Figure 11a). An AT bp is present between the copies of the mirror repeats (position 14), and four AT bps are organized as inverted repeats at the end of the repeat, at positions 26 to 29 (TTAA). According to a body of literature

(Buske et al., 2011; Frank-Kamenetskii & Mirkin, 1995; Mirkin & Frank-Kamenetskii, 1994; Soyfer & Potaman, 1996), the features of the CRJE suggest that it can form two isomers of a non-canonical H-DNA, so-called *H-DNA (Figure 11b, the symbol * stands for Hoogsteen bonds). *H-DNA is an intramolecular DNA triplex that, in this case, would be made of 11 base triads involving Hoogsteen bonds and presenting a single stranded stretch of 12 bps (Figure S11a and Figure S11b show the structure of the base triads and a 3D model of DNA triplex). Although slightly smaller, the CRJE sequence closely resembles the canonical sequences that have been reported to form *H-DNA (Figure S11c). Consistently, seven out of the 11 base triads potentially formed by the CRJE are one of the two most frequent reported to constitute *H-DNA (CG*G). Two of the 11 triads have been reported only in H-DNA so far (CG*C, GC*G), whereas the two remaining GC*C are non-canonical requiring more energy to be integrated in a DNA triplex but which have been observed *in vitro*. *H-DNA presents sequence requirements much less stringent than H-DNA (Mirkin & Frank-Kamenetskii, 1994), the mirror repeat may even not be present, so that the specificities of the CRJE sequence and the alternate triads that it may form are plausible. DNA triplexes are believed to play a number of roles in the cell (Frank-Kamenetskii & Mirkin, 1995; Soyfer & Potaman, 1996).

The second new feature of the CRJE is that the peptide encoded presents a direct tandem repeat of the motif ARAV, just upstream of the cutting site of the Kexin that is believed to ensure removal of the constant part corresponding to the UCS (Sunkin et al., 1998). Interestingly, one of the four positions that are imperfect in the repeats, the C at position 19, is necessary to encode the arginine (R) within the second copy of the motif ARAV (Figure 11a). RA within ARAV is a motif that is cleaved by a number of peptidase, such as the cathepsins (MEROPS database). Furthermore, the

R residue is a common recognition site for trypsin-like peptidase, for example the transmembrane serine protease present in the human lungs that is involved in the defence system (Uniprot O60235). Thus, the CRJE sequences might also be involved in the proper removal of the constant part of the Msg proteins by host enzymes in order to ensure the variation of each Msg' antigenicity.

As far as the other species of the genus are concerned, *P. carinii* harbours a CRJE sequence with a less symmetrical mirror repeat than that of *P. jirovecii* (Keely et al., 2007), but that could possibly form a DNA triplex (not shown). On the other hand, the one present in *P. murina* presents a much less conserved mirror repeat and symmetry (Keely et al., 2007), so that formation of triplex appears unlikely.

Discussion

We investigated the mechanisms used by the fungus *P. jirovecii* to vary its antigenicity by analyzing the repertoires of the genes encoding its major surface proteins, as well as the motifs and similarities present within the subtelomeres harbouring these genes. Based on our new results, we posit that the antigenic variation system of *P. jirovecii* relies on the three following mechanisms, listed in the order of their importance:

(i) Reassortment of the *msg*-I genes' repertoires and exchange of the expressed allele by translocation of entire genes mediated by DNA triplexes.

(ii) Rearrangement of the subtelomeres through single recombinations.

(iii) Mosaicism of the *msg* genes through intragenic recombinations.

(i) Reassortment of the *msg*-I genes' repertoires and exchange of the expressed allele by translocation of entire genes mediated by DNA triplexes

Although they were all distinct, the complete repertoires of the *P. jirovecii msg*-I alleles overlapped importantly among the patients. This implies that very frequent translocations of entire non-expressed alleles occur among the subtelomeres, which creates new assortments of the alleles and thus new subtelomeres. The latter would be then segregated into different *P. jirovecii* cell lines, in which further translocations of the *msg*-I alleles would occur. This continuous reassortment of alleles would lead to the distinctiveness of each repertoire that we observed independently of the geographic location, the year of the *Pneumocystis* pneumonia episode, or the subgroups of alleles observed by their hierarchical classification. The likely underlying mechanism of these translocations is a combination of two homologous

recombinations, one within the region localized upstream of the *msg*-I CDS and one within the downstream region (Figure 12a). Such events would permit the exchange of the two alleles (reciprocal exchange), but also possibly the replacement of one by the other (conversion, non-reciprocal exchange). Consistently, the up- and downstream regions flanking the *msg*-I CDSs proved to be similar with an identity that was ca. 20% higher than between the genes themselves, probably favouring the postulated recombinations outside of the genes. These recombinations might occur preferentially in the 33 bps CRJE sequence present at the beginning of each CDS and the 31 bps sequence located after the stop codon because these sequences are fully conserved within the up- and downstream regions. We also found that the sequences located immediately up- and downstream of the CDSs in *msg* families II, III, IV, and V are also often similar, particularly among the genes located at the tip of the subtelomeres. Furthermore, those of family IV present also an identity higher than the CDSs themselves, like for family I. These observations suggest that translocation of entire genes may also concern these families, but no data supporting this hypothesis are presently available. The facilitation of the translocation of *msg* genes thanks to intergenic spaces more conserved than the flanking genes themselves has been previously postulated in *P. carinii* (Keely et al., 2005).

Translocations through double recombinations are likely to affect also the expressed *msg*-I alleles because they also present the CRJE upstream and the similarity downstream (Figure 12b). However, most events affect probably the non-expressed alleles because, per genome, the latter are present in ca. 80 copies, whereas there is a single expressed allele per genome. Moreover, translocation of the expressed allele may be reduced by the fact that the similarity upstream of it is only the 33 bps of the CRJE upstream, whereas that of the non-expressed genes is 200

bps long plus 33 bps of the CRJE. As previously postulated (Schmid-Siegert et al., 2017; Sunkin & Stringer, 1996), an alternative mechanism is suggested by the distal location of the *msg*-I genes within the subtelomeres because it may facilitate their exchange through a single recombination. The latter would occur between two CRJE sequences, leading to the exchange of one or several genes together with the telomere linked to them (Figure 12c-d).

The number of expressed alleles per patient varied greatly (2 to 108). Apart from the number of *P. jirovecii* strains infecting the patient, as we evidenced in the present study, one can postulate the three following parameters influencing this number:

(i) The number of cells with distinct *msg*-I repertoires that infected the patient because it can influence the initial number of different expressed alleles.

(ii) The level of immunosuppression of the patient that can eliminate more or less efficiently the cell subpopulations expressing alleles or epitopes previously encountered by the patient.

(iii) The time elapsed between acquisition of the fungus and the *Pneumocystis* pneumonia episode. Indeed, new cell subpopulations expressing a specific allele are presumably segregated continuously and could accumulate over time in immunocompromised patients such as those we analysed.

The conservation of the CRJE sequence *in toto* and in multiple copies in the subtelomeres, precisely at the location where recombinations leading to the exchange

the expressed allele are postulated, very strongly suggest that it plays a crucial role in the antigenic variation system of *P. jirovecii*. The DNA triplexes that these sequences can potentially form because of their distinct motif including an imperfect mirror could be involved. However, despite that they constitute a hallmark representing up to 1% of eukaryotes' genomes (Cox & Mirkin, 1997; Goñi et al., 2006a; Wells, 1988), the function of the imperfect mirror repeats has been difficult to assess and remain putative so far. This situation results from the fact that DNA triplexes are notoriously difficult to tract, mostly because of their putative transient state and that the conditions required for their formation are not reproduced easily *in vitro* (Mirkin et al., 1987; Mirkin & Frank-Kamenetskii, 1994; Soyfer & Potaman, 1996). Nevertheless, a large body of circumstantial evidence (Bacolla et al., 2015), mainly their location within the genome, suggests that they are involved in a number of genetic process: transcription, replication, chromosome folding, structure of chromosome ends, mutational process, instability, rearrangements, translocations, and homologous recombination. In our context, the latter three processes are highly relevant, and homologous recombination is in fact the function that was most recurrently mentioned because mirror repeats have often been reported close to recombination sites (Frank-Kamenetskii & Mirkin, 1995; Rooney & Moore, 1995; Stasiak, 1992; Weinreb et al., 1990; Wells, 1988). Moreover, mediation of homologous recombination is the only potential function that has been supported by an experimental evidence: the presence of a polypurine-polypyrimidine stretch within a plasmid of *Escherichia coli* enhanced homologous recombinations within repetitive sequences present nearby (Kohwi & Panchenko, 1993). Furthermore, nucleic acid triplexes (RNA:DNA hybrids, R-loops) would be involved in (i) the switching of the expressed allele participating to the antigenic variation system of *Trypanosoma brucei* (Saha et al., 2020), and (ii) the switch

recombinations in mammalian immunoglobulins (Roy et al., 2008). Thus, we hypothesize that the DNA triplexes potentially formed by the CRJE sequences mediate, perhaps activate, the recombinations involved in the translocations of the *msg*-I genes, and possibly also the other recombinations involved in the system of antigenic variation.

(ii) Rearrangement of the subtelomeres through single recombinations

We found in the present study that the upstream regions of the genes of families II and III are very similar, but not their downstream sequences. Their identity was ca. 30% higher than between the genes nearby (79.8 versus 50.1%). These similarities may promote exchanges of large parts of subtelomeres through a single homologous recombination between them (Figure 12e). Such intergenic recombinations might be as or more frequent than intragenic recombinations leading to gene mosaicism because the mean identity of the genes within each *msg* family is lower or similar, *i.e.* 66, 83, 83, 72, 66, 45 for respectively families I to VI (Schmid-Siegert et al., 2017). This mechanism would lead to the creation of new subtelomeres that are potentially of varying sizes. This would be compatible with the important variation of the telomere length that we observed (Schmid-Siegert et al., 2017) (Figure S6 and Figure S7). Indeed, the distance between the genomic genes and the *msg*-I genes with a CRJE sequence, which are the genes always closest to the telomere, varied from 6 kb (contig/subtelomere 149, Figure S7a) to 25 kb (54, Figure S6).

(iii) Mosaicism of the *msg* genes through intragenic recombinations

A general phenomenon concerning subtelomeric genes is the occurrence of ectopic recombinations between them, *i.e.* also between non-homologous

chromosomes (Barry et al., 2003; Britten, 1998). These recombinations are likely to occur mostly when the telomeres and subtelomeres are bundled as a “bouquet” by the attachment to the spindle body during the prophase of meiosis (Barry et al., 2003; Oizumi et al., 2021; Yamamoto, 2014). This bouquet is involved in the matching and alignment of the homologous chromosomes by shaking them within the diploid cell. The latter phenomenon concerns most if not all eukaryotes, but has not been documented in the *Pneumocystis* genus so far. The products of two such recombinations are difficult to assess, and may vary according to the organism (Figure 12f). Moreover, a balance between fragment conversions and fragment exchanges may exist in all organisms because the former generally homogenises the alleles, whereas the latter diversifies them (Freitas-Junior et al., 2000). The importance of each process in the balance may depend on the amount of different alleles imported in the system, for example by mating of strains because it leads to the fusion of two sets of alleles. The recombinations responsible for the mosaicism of the alleles could also result from the meiotic reparation of the altered genes through fragment conversions (Chen et al., 2007).

Ectopic recombinations during meiosis are likely to generate the mosaicism of the *P. jirovecii* genes, pseudogenes, and intergenic spaces of the subtelomeres that is observed. However, our present observations suggest that the duplicated fragments we observed among the *P. jirovecii msg*-I alleles have been conserved from ancestral alleles during the diversification of the alleles. Indeed, two thirds (4 out of 6) of the fragments < 200 bps that we investigated were present in two subgroups of distant alleles. One was even in the two main subgroups of *msg*-I alleles, *i.e.* very distant ones. On the other hand, all ten fragments ≥ 200 bps were present in a single subgroup. During the diversification process, the probability for a fragment to be split

is probably greater with a length ≥ 200 than for those < 200. Thus, the larger fragments would more likely be present within a single subgroup than the shorter ones.

An alternative hypothesis is that the duplicated fragments result from conversion events that occurred within subgroups of related alleles, followed by the dissemination of the alleles created among patients. The presence of the same duplicated fragments within two distant subgroups, as we observed, could be explained by the existence of hotspots of recombinations along the gene leading to the same duplicated fragments. However, our analyses of the locations of the duplicated fragments did not reveal the presence of any such hotspots. Therefore, our observations favour the hypothesis that the diversification of the *msg*-I alleles in *P. jirovecii* results primarily from recombinations that split ancestral alleles. These recombinations could be those resulting in fragment exchanges, events we could not detect by our approach, or could be the single ones resulting in the rearrangements of the subtelomeres mentioned here above. This conclusion contrasts with those of previous studies. Indeed, conversion events were suggested by the high content of G+C motifs within *P. jirovecii msg*-I genes (Delaye et al., 2018), and by the analysis of duplicated fragments at the *P. carinii* expression site (J. R. Stringer, 2007).

Putting in perspective

Many different *P. jirovecii msg*-I alleles were observed being expressed in the lungs of the immunosuppressed patients analysed here. In contrast, immunocompetent individuals that are colonized by the fungus might harbour a smaller number of expressed alleles because of their effective immune system, but no data are presently available. The latter immunocompetent individuals constitute probably the niche in which the antigenic variation system of *P. jirovecii* evolved, so

that this system is above all a colonization factor. Thus, to improve our understanding of the latter, it might be useful investigating the immunocompetent transitory carriers such as healthcare workers in contact with patients with *Pneumocystis* pneumonia, or infants experiencing their primo-infection by *P. jirovecii*.

The important overlap of the *msg*-I genes' repertoires implies the existence of a very efficient mean of dissemination of the alleles and strains by frequent contacts between the *P. jirovecii* populations. This might happen through the primo-infections of infants that are frequent events occurring in the general population. Another not mutually exclusive hypothesis is the transient carriage of the fungus by healthy people, a phenomenon that has been frequently hypothesised but not firmly established so far (Chabé, Vargas, et al., 2004; Ma et al., 2018; Menotti et al., 2013).

Our data support that the strategy of *P. jirovecii* is the continuous production of new subpopulations that are antigenically distinct, as we previously proposed (Schmid-Siegert et al., 2017). This strategy relies on recombinations that take place during the prophase of meiosis, within the bouquet of telomeres and subtelomeres. Surface antigenic variation at each generation being probably necessary for survival, it might account for the obligate sexuality of this fungus (Hauser & Cushion, 2018; Richard et al., 2018). Consistently, the latter probably ensures also proliferation of the fungus (Hauser & Cushion, 2018; Richard et al., 2018). However, the frequency of the recombinations and the speed of the subtelomeres evolution remain to be determined. The strategy of *P. jirovecii* is one of a kind among human pathogens and might be adapted to the nonsterile niche constituted by the mammal lungs (Hauser, 2019; Schmid-Siegert et al., 2017). It contrasts with the populations that are antigenically homogenous of *Plasmodium* and *Trypanosoma*, and that vary overtime upon

exchange of the expressed gene (Deitsch et al., 2009). These might be imposed by sterile niches such as blood.

Acknowledgments

Sequencing was performed at the Lausanne Genomic Technologies Facility, University of Lausanne. We thank Patrick Taffé (Unisanté, Division of Biostatistics, University of Lausanne) for input into the simulation experiments. This work was supported by Swiss National Science Foundation grant 310030_ 192802 to P.M.H.. The Swiss National Science Foundation had no role in any steps of the study. All materials are available from the corresponding author.

Data availability statement

PacBio CCS raw reads (accession no. SRR24284242 - SRR24284301) have been deposited in the NCBI Sequence Read Archive linked to BioProject accession no. PRJNA936793 and BioSample accession no. SAMN33368625. The identified msg-I alleles and a table including their relative abundance in each patient are provided in the supplementary data (msg-I_alleles.fasta and Msg-I_alleles_abundance_in_patients.xlsx).

Code availability statement

Computer code used to generate results are available from authors upon request.

Competing interests

The authors declare no competing financial interests.

Author contributions

PMH and MP designed the study. CSM and SR performed lab experiments. CSM performed bioinformatics analyses under the supervision of MP. JMGCFA performed bioinformatics searches. KM, GN, MTC, and EJC prepared the samples and were involved in the writing of the manuscript. PMH and CSM wrote the first draft of the manuscript. All authors reviewed the manuscript.

Tables of Chapter 1

Table 1. Contingency table of the 917 and 538 alleles of the 24 complete (a) and 24 expressed (b) *msg*-I repertoires among the cities and patients.

a

Number of patients	Number of cities					Total	% alleles
	1	2	3	4	5		
1	411	0	0	0	0	411	44.8
2	50	128	0	0	0	178	19.4
3	4	62	40	0	0	106	11.6
4	2	15	36	7	0	60	6.5
5	0	8	16	14	1	39	4.3
6	0	1	20	17	4	42	4.6
7	0	0	8	11	4	23	2.5
8	0	0	6	10	4	20	2.2
9	0	0	1	8	3	12	1.3
10	0	0	2	3	4	9	1.0
11	0	0	0	2	2	4	0.4
12	0	0	0	1	4	5	0.5
13	0	0	0	1	3	4	0.4
14	0	0	0	0	2	2	0.2
15	0	0	0	0	2	2	0.2
Total	467	214	129	74	33	917	
% alleles	50.9	23.3	14.1	8.1	3.6		100

b

Number of patients	Number of cities					Total	% alleles
	1	2	3	4	5		
1	347	0	0	0	0	347	64.5
2	41	61	0	0	0	102	19.0
3	4	27	21	0	0	52	9.7
4	0	7	13	1	0	21	3.9
5	0	3	4	3	0	10	1.9
6	0	0	2	2	0	4	0.7
7	0	0	0	1	0	1	0.2
8	0	0	0	0	1	1	0.2
Total	392	98	40	7	1	538	
% alleles	72.9	18.2	7.4	1.3	0.2		100

Table 2. Homologies to the 200 bps sequences localized immediately up- or downstream of 20 representative *msg* CDSs within *P. jirovecii* subtelomeres from a single strain [a].

Query sequence of 200 bps localized immediately up- or downstream of *msg* CDS			No. of significant BLASTn hits localized immediately up- or downstream of *msg* CDS			
			of the same *msg* family		of another *msg* family	
msg family	*msg* gene no. (contig / subtelomere no.)	Localization relatively to gene	Genes	Pseudo - genes	Genes (family)	Pseudo–genes (family)
I	32 (54)	up [b]	18	1	0	0
		down [c]	11	3	0	0
	6 (18)	up [b]	20	4	0	0
		down [c]	11	3	0	0
II	7 (18)	up	7	2	7 (III)	1 (III)
		down	9	0	0	0
	25 (45)	up	1	0	0	0
		down	9	0	0	0
	3 (6)	up	0	0	0	0
		down	0	0	0	0
	37 (54)	up	1	1	0	0
		down	9	0	0	0
III	34 (54)	up	8	1	8 (II)	2 (II)
		down	4	0	0	0
	53 (74)	up	7	1	8 (II)	2 (II)
		down	4	0	0	0
	55 (74)	up	7	0	7 (II)	1 (II)
		down	4	0	0	0
	8 (18)	up	3	0	0	0
		down	0	0	0	0

Table 2. Continued.

Query sequence of 200 bps localized immediately up- or downstream of *msg* CDS			No. of significant BLASTn hits localized immediately up- or downstream of *msg* CDS			
			of the same *msg* family		of another *msg* family	
msg family	*msg* gene no. (contig / subtelomere no.)	Localization relatively to gene	Genes	Pseudo - genes	Genes no. hits-family	Pseudo–genes no. hits-family
IV	80 (110)	up	4	2	3 [d]	0
		down	3	2	0	0
	63 (95)	up	4	2	0	0
		down	2	0	1 [d]	0
V	54 (74)	up	3	1	1 [d]	0
		down	4	1	1 [d]	0
	42 (55)	up	0	0	0	0
		down	4	0	0	0
	28 (45)	up	2	0	0	0
		down	0	0	0	0
VI	18,39,44,49,73 (26,54,55,59,106)	up	0	0	0	0
	18,49,73	down	0	0	0	0
	39	down	1	0	1 [d]	0
	44	down	2	0	0	0

[a] The 200 bps sequence located immediately up- or downstream of the *msg* CDS was used as query in a BLASTn analysis against the PacBio *P. jirovecii* genome assembly (Schmid-Siegert et al., 2017) (search of somewhat similar sequences). All significant hits with a coverage of the query $\geq$ 40% are listed. The *msg* gene numbers are those used in Figure S6. The identities between hits and query of relevant genes are given in Table S7b.

[b] The CRJE was not included in the 200 bps upstream sequences of *msg*-I genes.

[c] The 200 bps sequence downstream of *msg*-I gene includes the 31 bps fully conserved ca. 90 bps after the stop codon.

[d] These hits were located between and thousands of bps from *msg* genes of different families (Table S7a).

Figures of Chapter 1

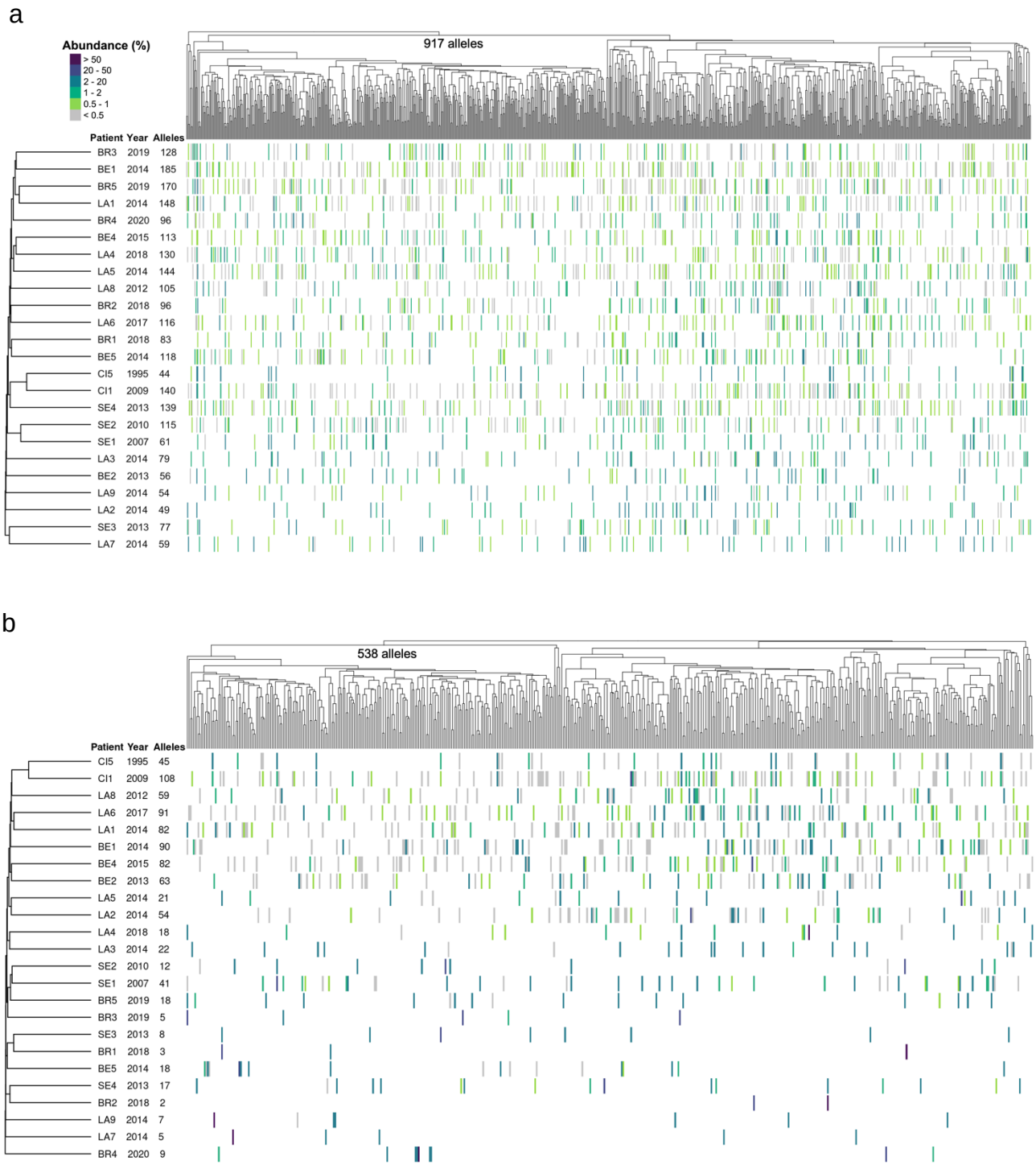


Figure 5. Composition of the complete (a) and expressed (b) *msg*-I repertoires present in the 24 patients. Each vertical line of the heatmap represents an allele present in the given repertoire with the color figuring its abundance in % of all reads composing the repertoire, as indicated at the top left of panel (a). The 917 and 538 distinct alleles identified in the repertoires were sorted using hierarchical classification trees of the multiple alignment of the allele sequences (Fitch distance, average linkage). The patients were sorted using a tree of presence/absence of each allele in their repertoire (binary distance, average linkage). The collection year and the number of alleles present in the patient are indicated next to the patient's name. LA, Lausanne. BE, Bern. BR, Brest. CI, Cincinnati. SE, Seville.

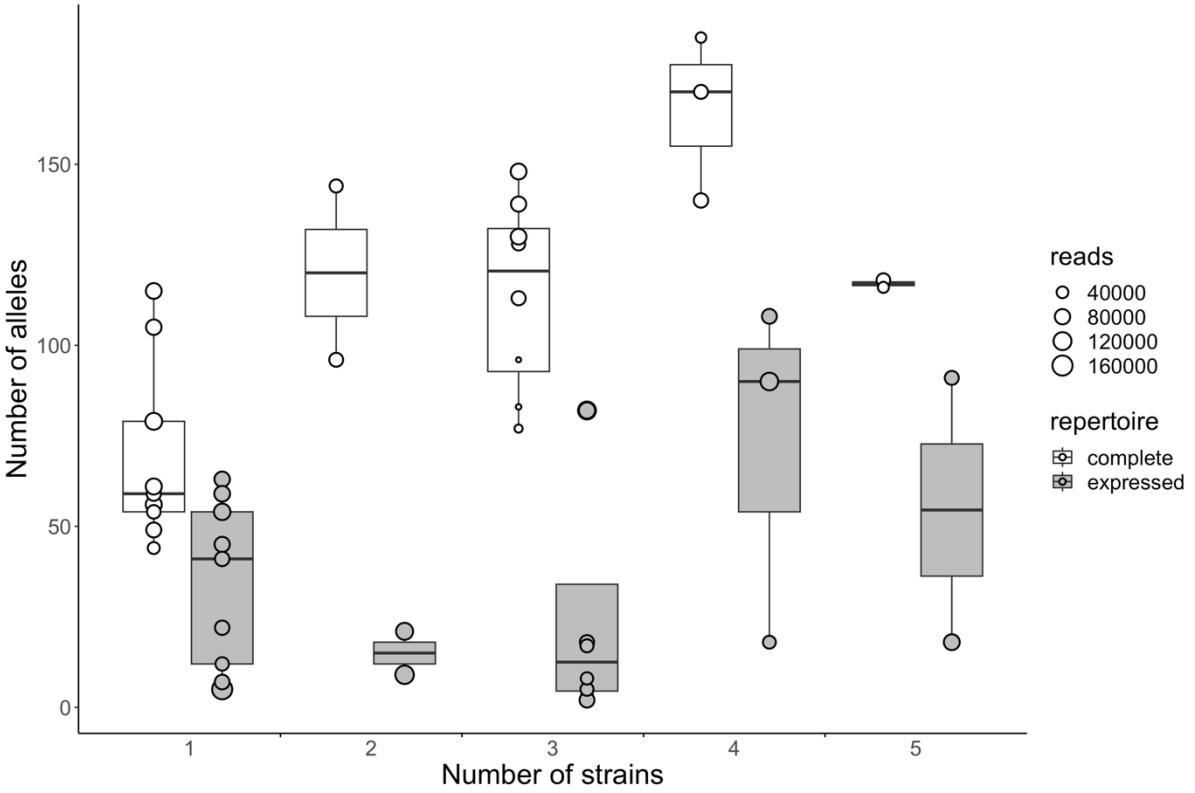
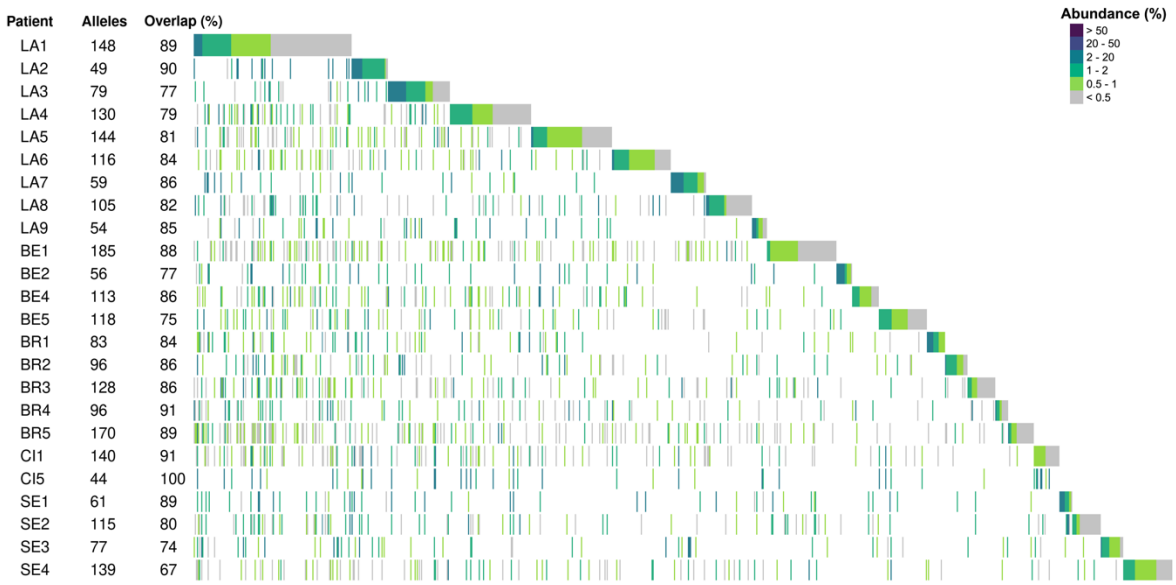


Figure 6. Correlation between the number of *P. jirovecii* strains and the number of alleles present within the 24 complete and 24 expressed repertoires. For the complete repertoire, the correlation between the number of alleles and the number of stains was 0.74 (Pearson correlation weighted by the number of reads, p-value = 4.2x10-5, n=24); the regression slope and intercept were 21.48 and 56.11, respectively. For the expressed repertoire the correlation was 0.32 (p-value = 0.12); the regression slope and intercept were 19.40 and 8.24, respectively. The size of each point indicates the number of reads generated by PacBio CCS for each sample. No correlation was observed between the number of reads and the number of different alleles identified in each sample (Pearson correlation, p-value: 0.20, correlation -0.19, SD 0.14). The center lines of the boxplots represent the median values, the box limits the 25th and 75th percentiles, and the whiskers extend to the largest values no further than 1.5 x inter-quartile range.

a



b

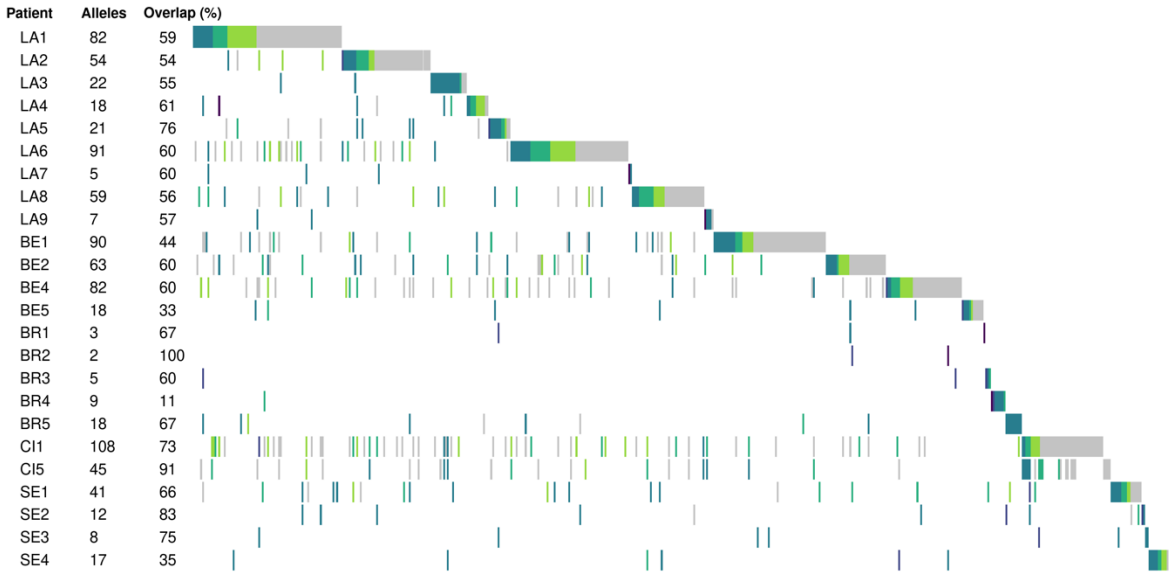
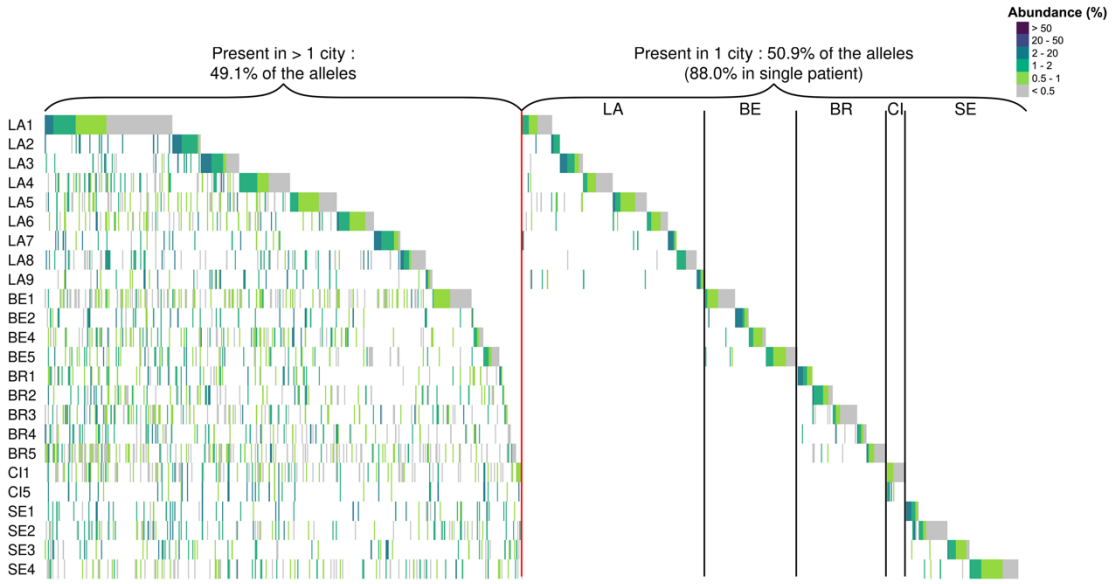


Figure 7. Complete (a) and expressed (b) repertoires of the 24 patients ordered by city in order to visualize the alleles that the patients share. The order of the cities was arbitrarily chosen. Each vertical line represents an allele present in the given repertoire with the color figuring its abundance in % of all reads composing the repertoire, as indicated at the top right of panel (a). For each patient, the alleles that are stacked on the right correspond to those not shared with the patient(s) that are placed above. The number of alleles present in each repertoire and the proportion of alleles shared with other patient(s), *i.e.* the overlaps, are indicated next to the patient's name. LA, Lausanne. BE, Bern. BR, Brest. CI, Cincinnati. SE, Seville.

a



b

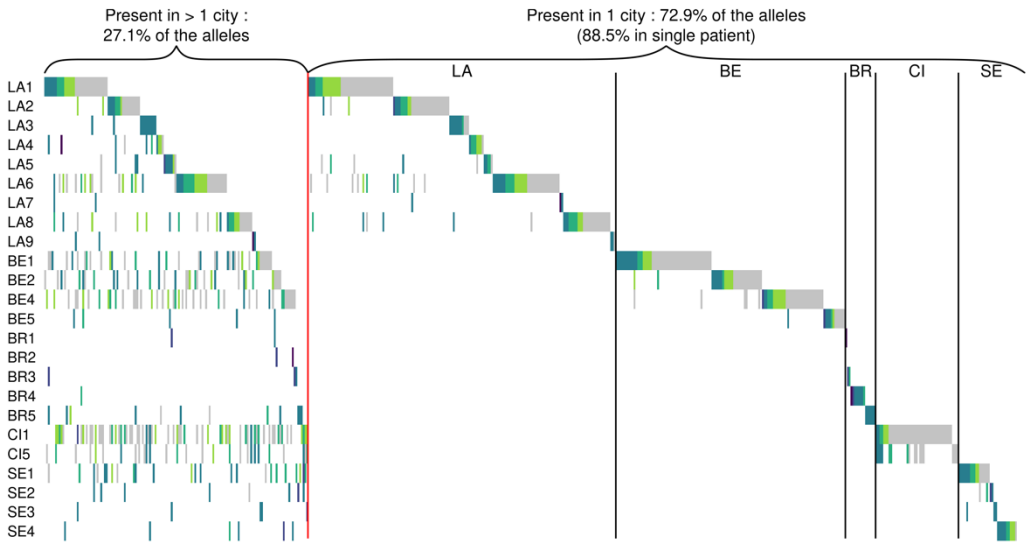


Figure 8. Complete (a) and expressed (b) repertoires of the 24 patients ordered by city and presence of the alleles in one or multiple cities. The order of the cities was arbitrarily chosen. Each vertical line represents an allele present in the given repertoire with the color figuring its abundance in % of all reads composing the repertoire, as indicated at the top right of panel (a). LA, Lausanne. BE, Bern. BR, Brest. CI, Cincinnati. SE, Seville.

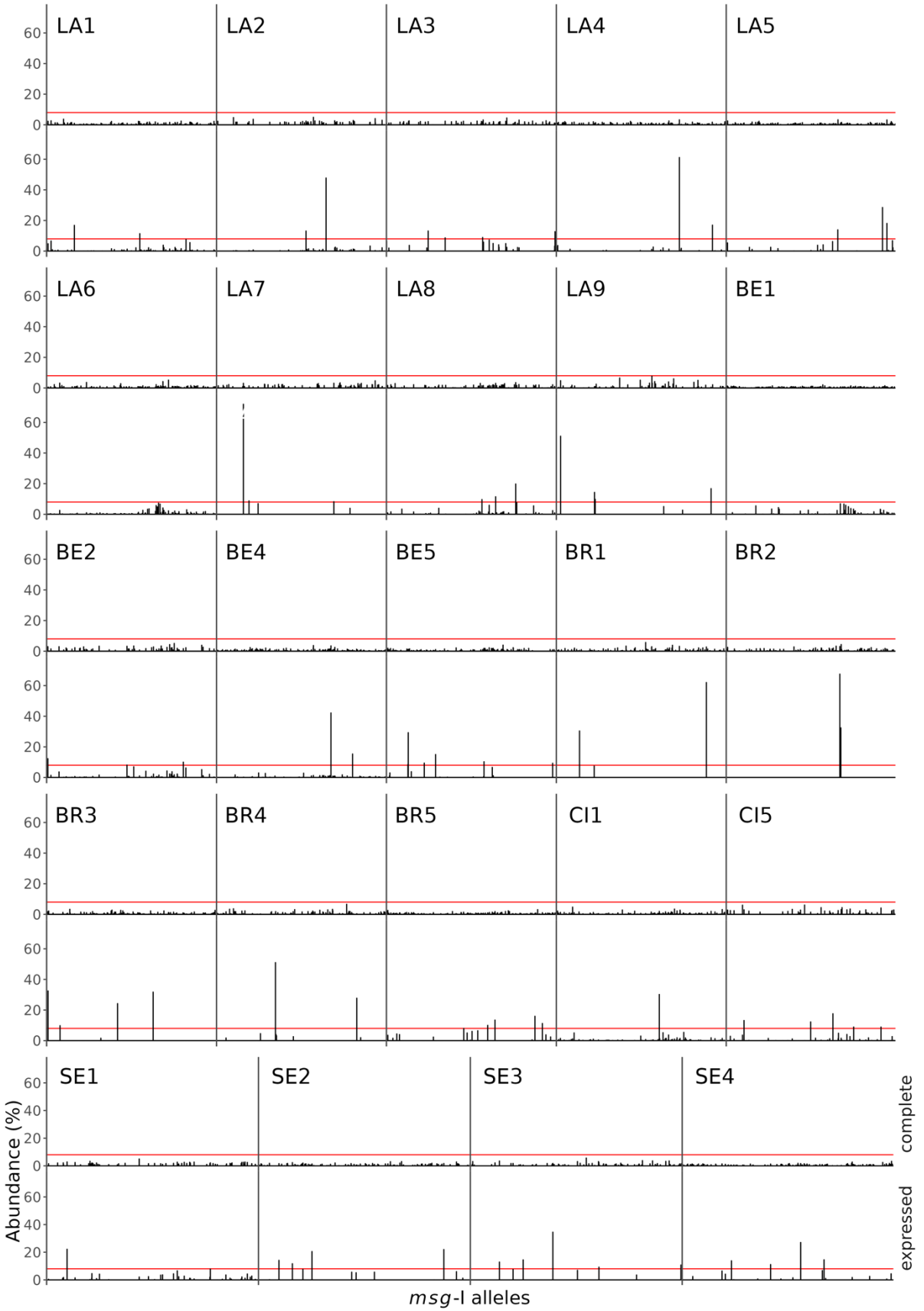


Figure 9. Abundance of the alleles present in the complete (top row) and expressed (bottom row) repertoires. The alleles were sorted using a hierarchical classification tree of a multiple alignment of all allele sequences found in the patient (not shown). The red lines indicate an abundance of 8.0%. LA, Lausanne. BE, Bern. BR, Brest. CI, Cincinnati. SE, Seville

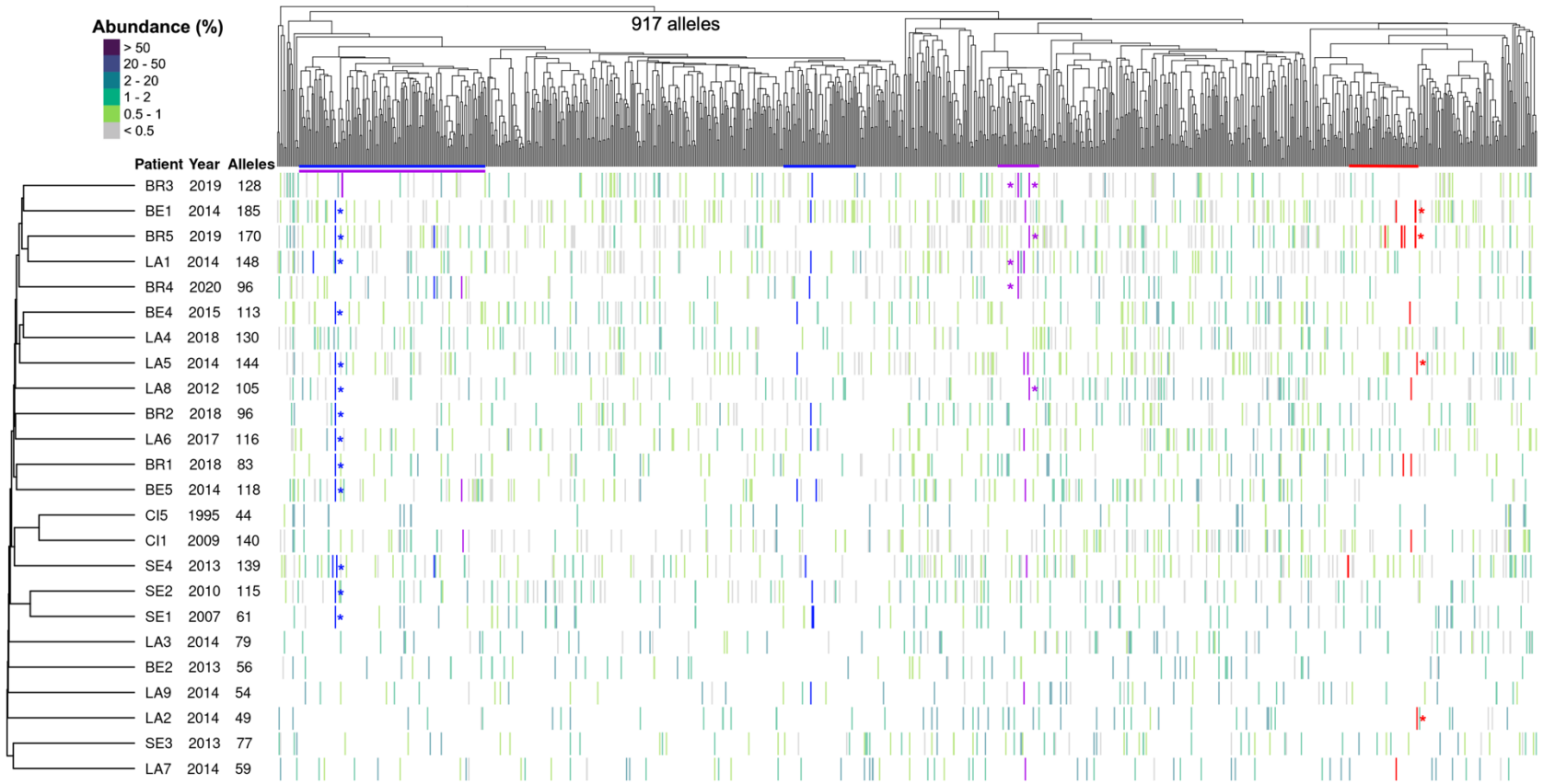


Figure 10. Conservation of fragments within subgroups of alleles among the complete *msg*-I repertoires of the 24 patients. Data are added onto Figure 1. Each vertical line of the heatmap represents an allele present in the given repertoire with the color figuring its abundance in % of all reads composing the repertoire, as indicated at the top left. The presence of each of the three fragments conserved in different alleles is represented by the coloration of the vertical line in blue, purple or red (red corresponds to the fragment C of 420 bps shown in the alignment of Figure S9, blue and purple correspond to fragments of respectively 133 and 104 bps, each from one of the two other alignments analysed). The stars indicate the two alleles initially aligned in order to identify the duplicated fragments. The colored horizontal lines below the tree show the subgroups of the hierarchical classification tree in which each of the three fragments are present. BE, Bern. BR, Brest. CI, Cincinnati. LA, Lausanne. SE, Seville.

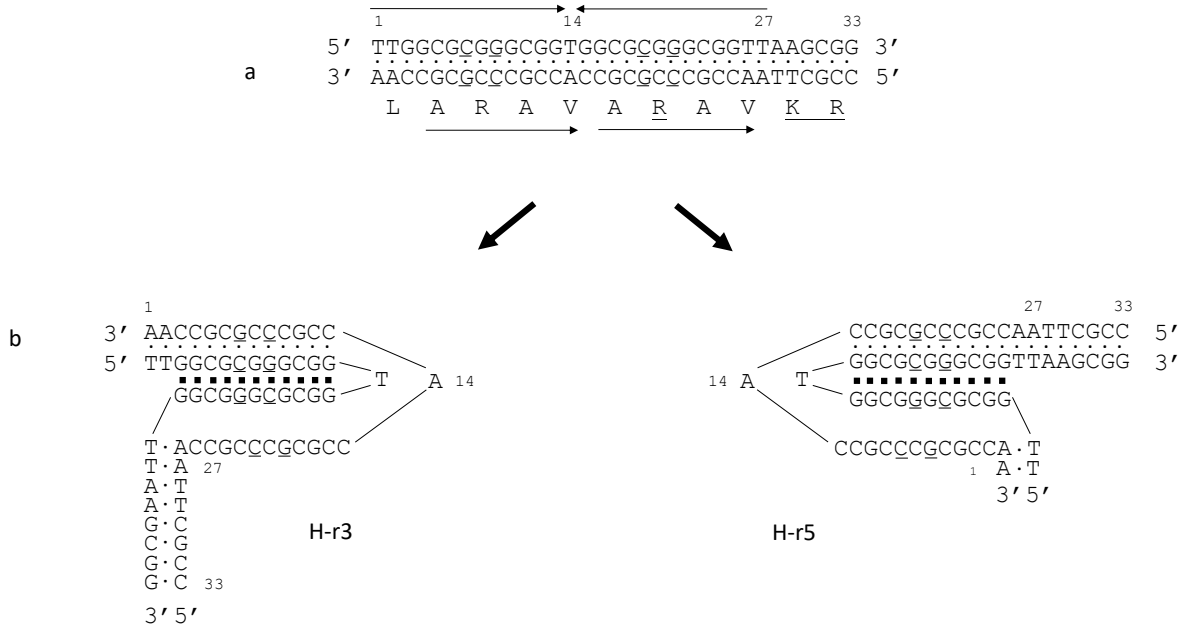
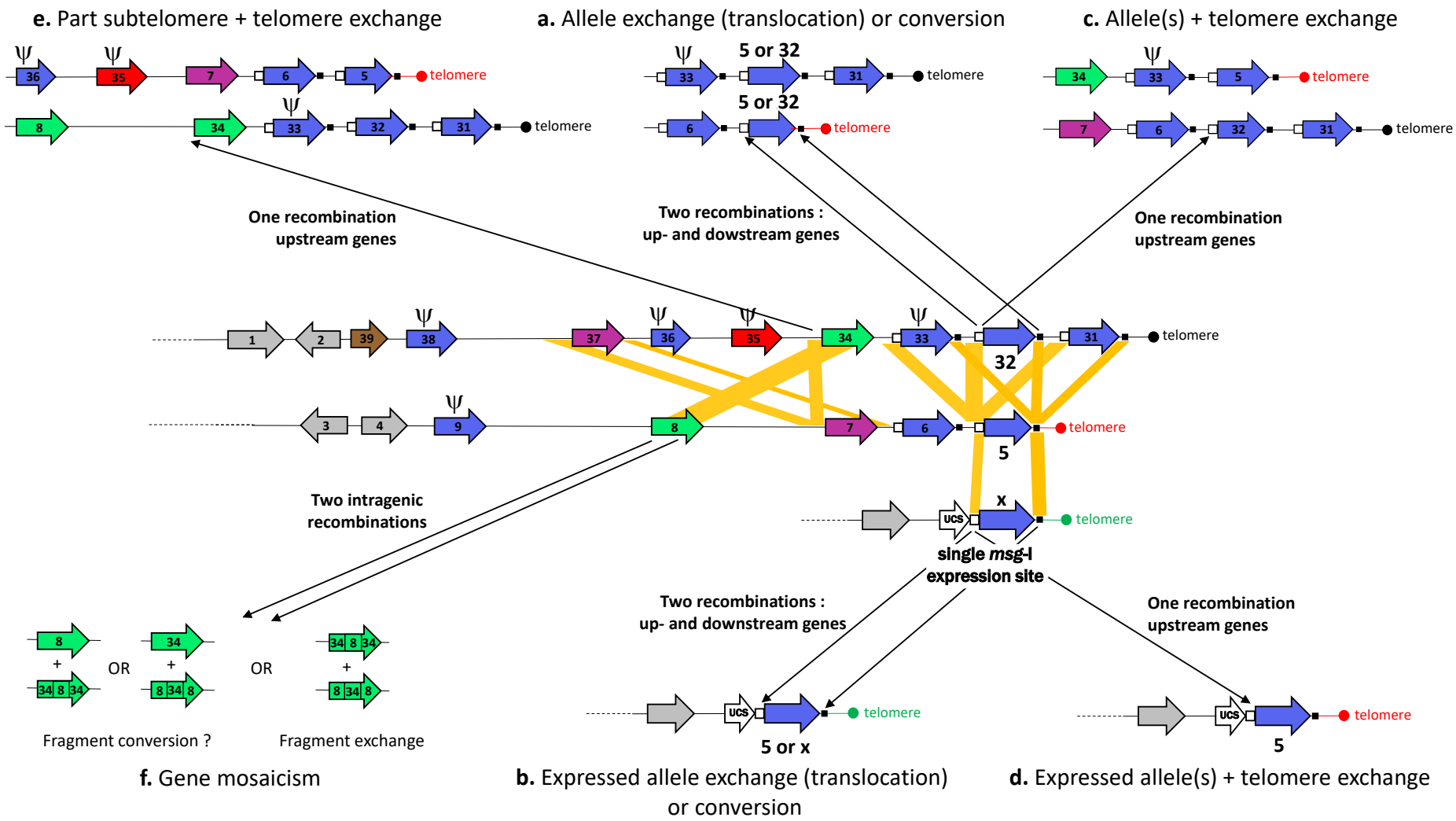


Figure 11. Structure of the conserved Recombination Junction Element (CRJE) and the *H-DNA triplex potentially formed.

a. Each strand of the CRJE is enriched in purines or pyrimidines that are part of an imperfect mirror repeats over 27 bps (symbolized by the convergent arrows). The imperfect positions 7, 9, 19 and 21 are underlined. The strand enriched in purines encodes the peptide shown underneath that is part of the Msg protein (amino acids' symbols are positioned at the center of the codon). This peptide presents a direct tandem repeat of the motif ARAV (symbolized by arrows pointing to the right). The imperfect C at position 19 of the DNA leads to the underlined R residue in the protein. The repeated ARAV is located before the recognition site KR of the kexin that is underlined (the cleavage by the kexin is believed to remove the constant part of the Msg protein, the UCS, during the maturation (Sunkin et al., 1998)). Approximately 13% of the CRJEs presents a transition T to C at position 27 of the DNA, not shown, which is a silent polymorphism.

b. The CRJE can potentially form two isomers of *H-DNA triplex made of base 11 triads involving each 2 or 3 Watson-Crick (points) and 2 Hoogsteen (squares) hydrogen bonds. In the two *H-DNA triplexes, a portion of the strand enriched in pyrimidines is single-stranded over 12 bps. The isomer in which the 5' half of the purines repeat is used as third stand (right, H-r5), rather than 3' (left, H-r3), is commonly less frequently formed (Mirkin & Frank-Kamenetskii, 1994).


e. Part subtelomere + telomere exchange
a. Allele exchange (translocation) or conversion
c. Allele(s) + telomere exchange
telomere
5 or 32
One recombination upstream genes
Two recombinations : up- and dowstream genes
One recombination upstream genes
32
5
x
Two intragenic recombinations
single msg-I expression site
Two recombinations : up- and downstream genes
One recombination upstream genes
5 or x
5
OR
Fragment conversion ?
Fragment exchange
f. Gene mosaicism
b. Expressed allele exchange (translocation) or conversion
d. Expressed allele(s) + telomere exchange

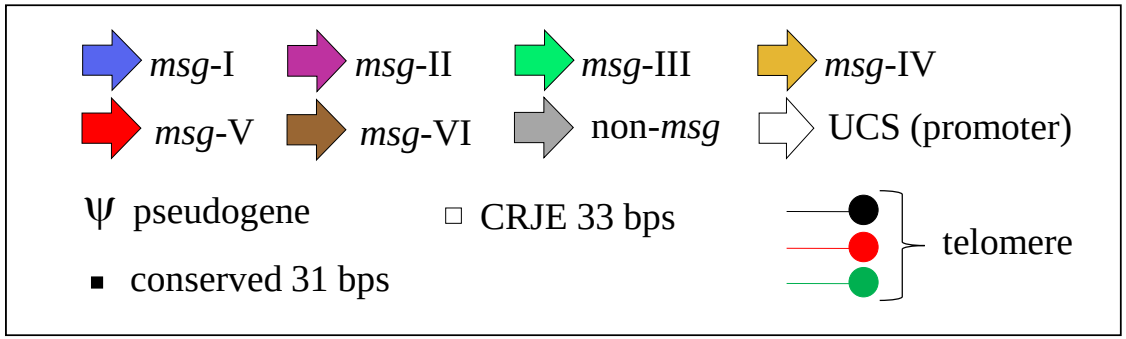


Figure 12. Model for the antigenic variation system of *P. jirovecii*. The subtelomeres shown are adapted from Fig. 3 of reference (Schmid-Siegert et al., 2017) (*i.e.* contigs 54 and 18). The subtelomere carrying the UCS (upstream conserved sequence) contains the promoter present at a single copy per genome (*i.e.* contig 72). Relevant similarities between the three subtelomeres are shown by the yellow parallelograms (between up- and downstream regions, and between *msg*-III genes). Recombinations within these similarities lead to the mechanisms indicated by the black arrows. The mechanisms a to f are described in the text. The occurrence of fragment conversions is not supported by the present study (mechanism f). DNA triplexes potentially formed by the CRJE (conserved recombination junction element) sequences might mediate the homologous recombinations.

Chapter 2: Clusters of patients with *Pneumocystis jirovecii* pneumonia harbouring similar *msg*-I repertoires

Introduction

The characterization of the *P. jirovecii msg*-I repertoires described in the chapter 1 of this thesis was initially applied to five additional patients with PCP. These five patients proved to be part of two clusters made of three and four patients harbouring identical or similar complete *msg*-I repertoires. The epidemiological analysis of these clusters is described in the present chapter 2. A third cluster with two patients harbouring weakly related repertoires was also analysed. These analyses suggest that a *P. jirovecii* strain with a stable *msg*-I repertoire was transmitted through a chain of multiple interhuman transmission events. Besides, the working hypothesis of the existence of infectious dormant cells is formulated.

Methods

All methods used for this chapter 2 are identical to those used in chapter 1. In particular, the *msg*-I repertoires and the ITS1-5.8S-ITS2 genotypes of the *P. jirovecii* strains infecting patients were determined using PacBio circular consensus sequencing (CCS). The variations to the methods described in chapter 1 are mentioned hereafter.

Patients

In this chapter 2, we analysed the broncho-alveolar lavage (BAL) samples of 29 immunocompromised patients with PCP from five different geographical locations over 26 years (Table 3). These 29 patients include the 24 analysed in chapter 1.

Correspondences of the genotypes with previous publications

In order to compare the ITS1-5.8S-ITS2 genotypes observed among the 29 patients to those reported in previous studies by other researchers, each sequence was searched within the NCBI GenBank database using BLASTn and the default settings. Six homopolymer stretches were ignored because of their artefactual variation (see section “Estimation of the number of *P. jirovecii* strains” of chapter 1, and Figure S2 of Annex 1). Only 100% identical hits were considered (Table 4).

Results

Detection of clusters of patients with *P. jirovecii* pneumonia by their identical or similar complete *msg*-I repertoires

The characterization of the complete repertoires of *P. jirovecii msg*-I genes present in 29 patients with PCP highlighted that only 20 repertoires were clearly distinct. The remaining nine formed three clusters, one including three patients from Switzerland with nearly identical repertoires, another including four patients from North America (Cincinnati) with similar repertoires, and a third including two patients from Spain (Seville) with weakly related repertoires (Figure 13a, the clusters are emphasized by binary hierarchical clustering of the patients shown by the tree on the left).

Two of the three patients forming the Swiss cluster were from Bern (BE2 and BE3), both were renal transplant recipients and their PCP infection occurred seven months apart in 2013 and 2014. The third patient (LA10) was diagnosed with PCP in Lausanne in 2021, *i.e.* seven years after the two other patients. The underlying disease of LA10 is unknown (Table 3). The three patients harboured repertoires that shared 54 *msg*-I alleles, while only two or three additional were present in low abundance in respectively BE2 and BE3 relatively to LA10. These differences might result from the main limitation of our approach, *i.e.* the non-amplification in each PCR of the low abundant alleles because of the stochastic variations of their amount serving as template (see Annex 1, supplementary data 1 of chapter 1).

The second cluster included four patients diagnosed in Cincinnati (CI1 to CI4). The four infections happened in 2009, but none of the underlying diseases are known (Table 3). The complete repertoires present in the patients were much less similar than

those of the Swiss cluster as their numbers of *msg*-I alleles varied between 89 and 144 (Figure 13a). Nevertheless, there was a clear similarity between these repertoires.

The third cluster included two HIV positive patients diagnosed in Seville in 2007 and 2010. The patients harboured weakly similar *msg*-I complete repertoires, *i.e.* much less similar than those of the American cluster (Figure 13a).

Similarity between the expressed *msg*-I repertoires present in the clusters

We also characterized the *P. jirovecii* expressed *msg*-I repertoire present in each of the 29 patients (Figure 13b). All expressed repertoires were distinct one from another, their alleles being spread evenly over the whole of their classification tree, as seen for the complete repertoires (Figure 13a). Nevertheless, those of the Swiss and American clusters were clearly related, whereas those of the Spanish cluster were only weakly related.

The Swiss cluster presented expressed repertoires that were related but not as identical as their complete ones (Figure 13a and Figure 13b). The number of expressed alleles was very different between the patients, ranging from 13 to 63. Additionally, most of the alleles expressed in LA10 and BE3, respectively 100 and 92%, were among the 63 expressed in BE2. Hence, the expressed repertoires of LA10 and BE3 were both subsets of that of BE2, who was the first patient diagnosed with PCP among the cluster. Interestingly, the identities of the expressed alleles were largely different between LA10 and BE3 as only four alleles were expressed in both patients (20% and 31% of the alleles). The latter observation is noteworthy because the expressed alleles were supposedly retrieved from identical or nearly identical complete repertoires for mutually exclusive expression.

The complete and expressed repertoires of the American cluster shared many alleles and presented a comparable level of similarity as evidenced by the binary hierarchical trees sorting the patients (Figure 13a and 10b).

On the other hand, the expressed repertoires of the Spanish cluster were very weakly similar, less than their complete ones.

Of note, BE2 is one of the three patients out of the 29 with a larger number of alleles in its expressed repertoire compared to its complete one (with CI5 and LA2, Figure 13a and 10b). This is probably explained by the main limitation of our experimental approach (see above).

No laboratory mix-up of the samples

The legitimate assumption of a laboratory mix-up of samples that would have created artificially the three clusters was addressed by analysing the possibility of cross-contamination between the samples during the whole experimental procedure. For the Swiss cluster, the expressed repertoires of BE2 and BE3 were processed together for all steps, whereas their complete repertoires only during DNA extraction and PCR amplification. Additionally, the PCR product of the complete repertoire of LA10 was purified with that of BE3. For the American cluster, seven out of the eight samples of the four patients were handled in parallel during all steps of the procedure. The exception was the complete repertoire of CI1 that was processed alongside other samples. For the Spanish cluster, all the samples were handled in parallel for all steps of the procedure.

Nevertheless, cross-contaminations most probably did not occur because of four reasons:

(i) The two samples of patient CI5, who is not a member of the American cluster, were processed together with the four patients of this cluster during the whole procedure. However, CI5 presented complete and expressed repertoires very different from those observed among the cluster (Figure 13a and 10b).

(ii) The complete repertoire of CI1 as well as the expressed repertoire of LA10 were never processed with any other sample of their respective cluster, and thus could not have been contaminated.

(iii) The six samples of the three patients forming the Swiss cluster were processed, at least for one step of the procedure, together with four to seven samples of patients presenting repertoires very different from those of the cluster (*i.e.* BE1, BE4, BE5, LA2, LA3, LA5, LA8, LA9, BR2, CI1).

(iv) The two control plasmids carrying a single allele that we analysed to set up the procedure (see chapter 1) have not been contaminated by the four samples containing many different alleles with which they were processed for the evaporation of their PCR product before sequencing (*i.e.* BR1, BR4, SE3, SE4).

A novel genotyping method using PacBio CCS

In order to further assess the three clusters detected using characterization of the *msg*-I repertoires, we developed a novel method to genotype *P. jirovecii* relying on the polymorphism of the region ITS1-5.8S-ITS2 of the nuclear rRNA operon (500 bps long). The latter was amplified using a generic PCR and sequenced with PacBio CCS. The advantage of PacBio CCS is the possibility to process many samples at once and

to detect efficiently co-infections with several genotypes that occur in most patients. However, we faced the problem that the ITS1 and ITS2 genotypes differ often by a single nucleotide polymorphism, which is very similar to the error rate of PacBio CCS of 0.2% to 0.5% (Wenger et al., 2019; determined in chapter 1). To find the threshold of sequence identity and procedure to use for proper genotyping, we analysed three control plasmids carrying each a single ITS1-5.8S-ITS2 allele and searched how to identify a single genotype upon processing the CCS reads. Accordingly, we defined a *P. jirovecii* ITS1-5.8S-ITS2 genotype as a group of CCS sequences that are fully identical, and we considered only the groups that represented more than 1% of the total reads present in at least one patient. This procedure resulted in 31% to 59% of the reads being included within the genotypes, according to the patient (mean 47%, SD 7%). It is possible that several genotypes representing less than the threshold of 1% are true biological co-infecting genotypes that are ignored using our approach, but our results contain the genotypes present in majority, which are likely to have the highest biological significance.

ITS1-5.8S-ITS2 PacBio CCS genotyping of the strains infecting the 29 patients

We identified 28 different genotypes among the 29 patients. Ten patients were infected with a single genotype, while 20 were co-infected with two to five genotypes (Figure 14). Table 4 shows the correspondences of these genotypes with the numbering of Xue et al. (2019), as well as with their accession numbers in the GenBank database. Figure S12 in annex 2 shows the alignment of their sequences. The two most common genotypes (Pj01 and Pj02) were present in respectively 16 and 12 patients, 11 other genotypes were present in 2 to 5 patients (Pj03 to Pj13), while 15

genotypes were present only in single patients (Pj14 to Pj28). All patients harboured at least one out of the six most prevalent genotypes (Pj01 to Pj06), *i.e.* the genotypes that were present in at least four different patients (Figure 14).

The Swiss cluster identified by the *msg*-I repertoires was confirmed by the ITSs genotyping (Figure 14). Indeed, all three patients were infected by genotype Pj01 as the most abundant one. Both BE2 and LA10 harboured only this single genotype while BE3 had the additional genotype Pj05, though at a low abundance of 1.96% (versus 48.51% for Pj01). Nevertheless, because of this low abundance, the supplementary ITSs genotype could have been missed in the other patients.

On the other hand, the American cluster showed a different pattern. Indeed, three patients were co-infected each with four different genotypes, while the fourth harboured five. Moreover, two subclusters were identified within this cluster, CI1 and CI2 on one side, CI3 and CI4 on the other (Figure 14). The latter subcluster was infected by the same four genotypes (Pj01, Pj04, Pj11, Pj13). The former had three genotypes in common (Pj02, Pj07, Pj12), but Pj01 was also present in patient CI1, and Pj09 plus Pj23 in CI2. Thus, notably, the two subclusters shared almost no genotype. These two subclusters were not identified by the determination of the complete *msg*-I repertoires (Figure 13a). However, the subcluster composed of CI3 and CI4 identified by ITSs genotyping was also present among the expressed *msg*-I repertoires (Figure 13b).

The patients of the Spanish cluster were both infected with the single genotype Pj01, confirming the relationship suggested by the complete *msg*-I repertoires.

In contrast, most of the remaining 20 patients harboured a distinct single ITSs genotype or a distinct combination of co-infecting ITSs genotypes (Figure 14). Nevertheless, three patients from Lausanne (LA3, LA8 and LA9) were infected with

the single genotype Pj01 between 2012 and 2014. Two of these patients were HIV positive, whereas the underlying condition of LA3 is unknown. However, the complete and expressed repertoires harboured by these patients were not related (Figure 13a and 10b).

Discussion

The characterization of the complete repertoires of *P. jirovecii msg*-I genes present in 29 patients detected three clusters involving nine patients with identical or similar repertoires. Genotyping *P. jirovecii* by sequencing the ITS1-5.8S-ITS2 region with PacBio CCS confirmed two of these clusters. The latter analysis detected in addition one small cluster of patients harbouring an identical ITSs genotype. The four clusters observed have drastically different characteristics, leading to the distinct interpretations that are exposed hereafter.

The three patients forming the Swiss cluster harboured nearly identical complete *msg*-I repertoires and ITSs genotypes. However, their expressed *msg*-I repertoires varied considerably. Two of them were subsets of the one of the first diagnosed patient, *i.e.* the “index patient”. This suggests that only some of the *P. jirovecii* sub-populations present in the index patient, each expressing a specific *msg*-I allele, were transmitted to and/or succeeded to proliferate in each of the other two patients of the cluster. Of note, the latter observation suggests the original hypothesis that transmission of a diversity of expressed alleles might be crucial to infect successfully new hosts. Thus, a single strain, although expressing different *msg*-I repertoires, infected the three patients of this cluster, implying transmission between them. This interpretation is supported by the fact that two of the patients were kidney transplant recipients from the same city that were diagnosed with PCP only seven months apart. Indeed, many clusters among transplant recipients due to interhuman transmission have been described, and were explained through encounters during visits at the hospital (Rabodonirina et al., 2004; Yiannakis & Boswell, 2016). Nevertheless, the third patient of the cluster has an unknown condition and was diagnosed with PCP in another Swiss city. Moreover, the diagnosis occurred seven

years after those of the other two patients. This long period of time raises the question of how the strain was transmitted to this third patient. We propose the two following not mutually exclusive hypotheses.

First, the strain was genetically stable over a long period of time, including at its subtelomeres carrying the *msg*-I repertoire, and was transmitted to the third patient through a chain of transmission involving unknown infected individuals. The antigenic variation system of *P. jirovecii* is believed to allow escaping the human immune system mostly through reassortment of the repertoires, mutually exclusive expression, and mosaicism of the *msg*-I genes (Ma et al., 2016; Schmid-Siegert et al., 2017, Meier et al., submitted). The dynamics of these mechanisms is unknown, and *P. jirovecii* strains may have stable *msg*-I repertoires over long period of time, which could explain our observations. Such a strain with stable *msg*-I repertoires has been previously described among two clusters of renal transplant recipients that occurred ca. 300 km apart in a period of four years (Sassi et al., 2012). This observation relied on identical patterns observed by restriction analysis of all the 3' halves of the *msg*-I genes present in each patient, a highly discriminant genotyping method analysing repertoires that is similar to the analysis used in the present study. The authors concluded that this strain was adapted to renal transplant recipients. They called it the European Renal Transplant (ERT) strain because they observed another strain with a stable but different *msg*-I restriction pattern in Japan. The ERT strain was later reported in a third cluster that occurred during the same period of time ca. 400 km away from the two other clusters (Hauser et al., 2013). Genotyping using multilocus of four genomic loci showed that the strain involved in the Swiss cluster reported in the present study is different from the ERT strain (results not shown). The latter observation confirms that different strains may present stable *msg*-I repertoires.

The second not mutually exclusive hypothesis to explain the seven years gap is that a dormant state of *P. jirovecii* exists. The stability of the complete *msg*-I repertoire would be accounted for by the reactivation of such dormant cells after a long period of time. *P. jirovecii* has been detected in air samples collected around PCP patients in hospital and at their homes (Bartlett et al., 1997; Le Gal et al., 2015). Asci might be disseminated over long distances in such a dormant state, which would fit their detection in the air within a rural location far from any cities (Wakefield, 1996). This hypothesis is supported by the fact that *P. jirovecii* asci contain glucans within their wall that are supposed to confer resistance to the damaging physical constraints present in the environment. Such long-distance dispersal of *P. jirovecii* would also account for or contribute to the dissemination of the *msg*-I alleles all around the world during the process of repertoires reassortment that we report (chapter 1; Meier et al., submitted). Nevertheless, only *P. jirovecii* and no other *Pneumocystis* species was present in human exposome metagenomic data sets, *i.e.* from particles collected in the close proximity of individuals. This suggested that local transmission of *P. jirovecii* is strongly favoured over long distance ones (Cissé et al., 2020). It might be that both types of transmission coexist, but that the local one is more frequent. It is possible that asci can remain in a dormant state in some environmental locations, for example within dust, where *P. jirovecii* has been recently detected (Gantois et al., 2021). Moreover, it cannot be excluded presently that such dormant forms might remain within the lungs of humans, after a PCP episode or upon carriage of the fungus. Indeed, the detection methods might not be sensitive enough to detect these cells because of their extremely low number. Notably, the chains of transmission of the ERT strain between the patients of the three clusters were continuous, *i.e.* linked by encounters allowing potential

transmission events, and thus did not need to include a dormant state of the source of infection (Gianella et al., 2010; Pliquett et al., 2012; Schmoldt et al., 2008).

The American cluster contained four patients that harboured complete and expressed *msg*-I repertoires with a comparable similarity and large numbers of alleles. Such numerous alleles were consistent with the co-infections with several ITSs genotypes observed in these patients. The presence of multiple genotypes might be due to their accumulation through several interhuman transmission events, possibly between these patients, which is not excluded because they were diagnosed in the same city during the same year. Surprisingly, the two very distinct subclusters of ITSs genotypes composing this cluster were not observed among the complete *msg*-I repertoires classification, and only partially among the expressed *msg*-I ones. A tentative hypothesis is that the multiple transmission events between the patients coincided with an enrichment with specific alleles in the city or geographical area. Nevertheless, this cluster remains epidemiologically unsolved.

The two HIV positive patients of the Spanish cluster harboured complete and expressed *msg*-I repertoires that were weakly related, but an identical ITSs genotype. Given that this cluster was observed over a period between two and three years, interhuman transmission appears possible, although transmission through an intermediary patient or a dormant source of infection is not excluded. The weak relationship of the *msg*-I repertoires might result from a relatively rapid reassortment of the repertoires within a stable ITSs genotype, in contrast to the stability observed in the Swiss cluster. Indeed, one cannot exclude that these dynamics of reassortment vary according to conditions encountered by the fungus. For example, it may depend on the pressure exerted by the immune system within the lungs. Thus, it might depend on the status of the host, immunocompetent or immunocompromised, and/or the

underlying immune impairing condition. Accordingly, a drastic decrease in the rate of repertoires reassortment may result from specific parameters present in the environment of lungs of transplant recipients. Consequently, the cause of the difference between the Swiss and Spanish clusters would be crucial to determine, because it might be related to the HIV positive rather than transplant recipient status of the patients. However, alternatively, given that only two patients are involved, one cannot exclude that the weak relationship between the Spanish repertoires is simply due to an enrichment with specific alleles in the city or geographical area, and that the presence of the same ITSs genotype is a coincidence. However, this cluster remains epidemiologically unclear.

ITSs genotyping identified one additional small cluster in Lausanne consisting in two HIV positive patients and one patient with unknown underlying disease. All were infected by the same single genotype but harboured distinct complete and expressed *msg*-I repertoires. The period of two to three years between the diagnoses implies again that interhuman transmission has been possible, and that intermediary patients or a dormant source might have been involved. This cluster also remains unclear, a coincidence of the same ITSs genotype is also possible.

In conclusion, our observations confirm that strains with a stable repertoire of *msg*-I genes are involved in clusters of transplant recipients, spanning over a few years. They also suggest the hypothesis that another mean than direct transmission of the fungus between individuals may exist, *i.e.* a dormant form as an environmental source of the infection. Finally, they lead to the working hypothesis that the speed of evolution of the *msg*-I repertoires may vary according to conditions or environment, possibly according to the underlying disease of the host.

Tables of Chapter 2

Table 3. BAL samples from 29 immunocompromised patients included in this study. Both clusters are indicated in colors (Swiss cluster in green, American cluster in violet, Spanish cluster in orange)

City	Country	Patient name [a]	Collection year	Underlying disease
Lausanne	Switzerland	LA1	2014	HIV
		LA2	2014	HIV
		LA3	2014	unknown
		LA4	2018	unknown
		LA5	2014	unknown
		LA6	2017	unknown
		LA7	2014	HIV
		LA8	2012	HIV
		LA9	2014	HIV
		LA10	2021	unknown
Bern	Switzerland	BE1	2014	HIV
		BE2	2013	kidney transplant
		BE3	2014	kidney transplant
		BE4	2015	cancer
		BE5	2014	cancer
Brest	France	BR1	2018	giant cell arteritis
		BR2	2018	cancer
		BR3	2019	HIV
		BR4	2020	psoriasis (methotrexate)
		BR5	2019	cancer
Cincinnati	United States	CI1	2009	unknown
		CI2	2009	unknown
		CI3	2009	unknown
		CI4	2009	unknown
		CI5	1995	unknown
Seville	Spain	SE1	2007	HIV
		SE2	2010	HIV
		SE3	2013	HIV
		SE4	2013	HIV

[a] LA, Lausanne. BE, Bern. BR, Brest. CI, Cincinnati. SE, Seville.

Table 4. Genotypes identified among the 29 patients of this study.

PacBio ITS1-5.8S-ITS2 genotype	Genotype number of Xue et al. (2019)[a]	GenBank accession number	Frequency among the 29 patients
Pj01	1	JQ365709	16
Pj02	17	AB481410	12
Pj03	59	MK300661	5
Pj04	10	JQ365725	4
Pj05	nd [b]	JQ365722	4
Pj06	12	AB481406	4
Pj07	62	MK300664	3
Pj08	27	JQ365707	2
Pj09	nd	nd	2
Pj10	4	KC470776	2
Pj11	nd	nd	2
Pj12	nd	nd	2
Pj13	nd	nd	2
Pj14	nd	nd	1
Pj15	nd	nd	1
Pj16	nd	nd	1
Pj17	nd	nd	1
Pj18	9	JQ365723	1
Pj19	nd	nd	1
Pj20	nd	nd	1
Pj21	nd	nd	1
Pj22	nd	nd	1
Pj23	nd	nd	1
Pj24	nd	nd	1
Pj25	nd	nd	1
Pj26	nd	nd	1
Pj27	nd	nd	1
Pj28	nd	nd	1

[a] The sequence of the 5.8S gene was not considered by Xue et al. (2019).
[b] nd, not described.

Figures of Chapter 2

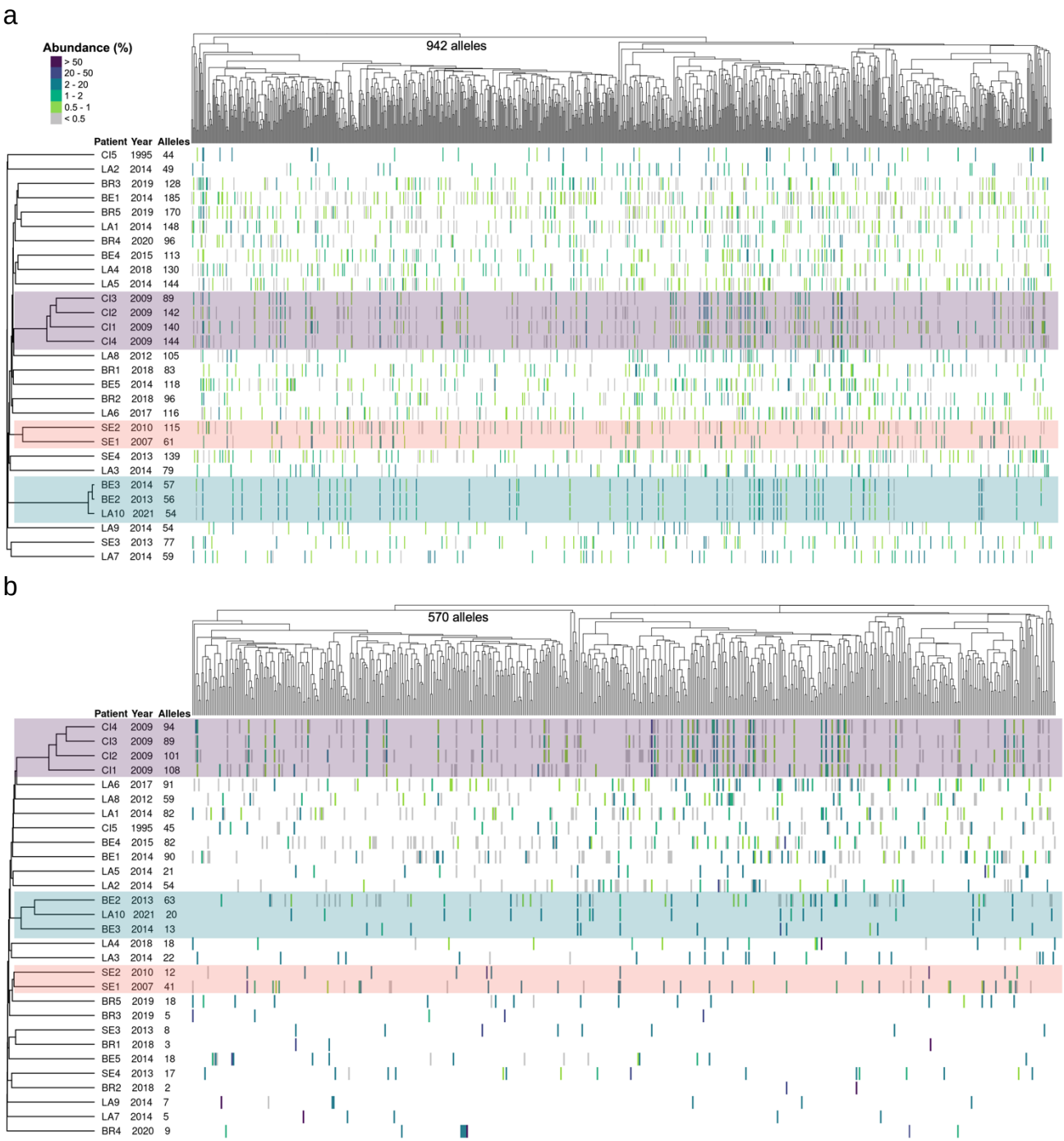


Figure 13. Composition of the complete (a) and expressed (b) *msg*-I repertoires present in the 29 patients. The clusters are indicated by the colored backgrouds (green: Swiss cluster; violet: American cluster, orange: Spanish cluster). Each vertical line of the heatmap represents an allele present in the given repertoire with the color figuring its abundance in % of all reads composing the repertoire, as indicated at the top left of panel a. The 942 and 570 distinct alleles identified in the repertoires were sorted using hierarchical classification trees of the multiple alignment of the allele sequences (Fitch distance, average linkage). The patients were sorted using a binary hierarchical clustering tree of presence/absence of each allele in their repertoire (binary distance, average linkage). The collection year and the number of alleles present in the patient are indicated next to the patient's name. LA, Lausanne. BE, Bern. BR, Brest. CI, Cincinnati. SE, Seville

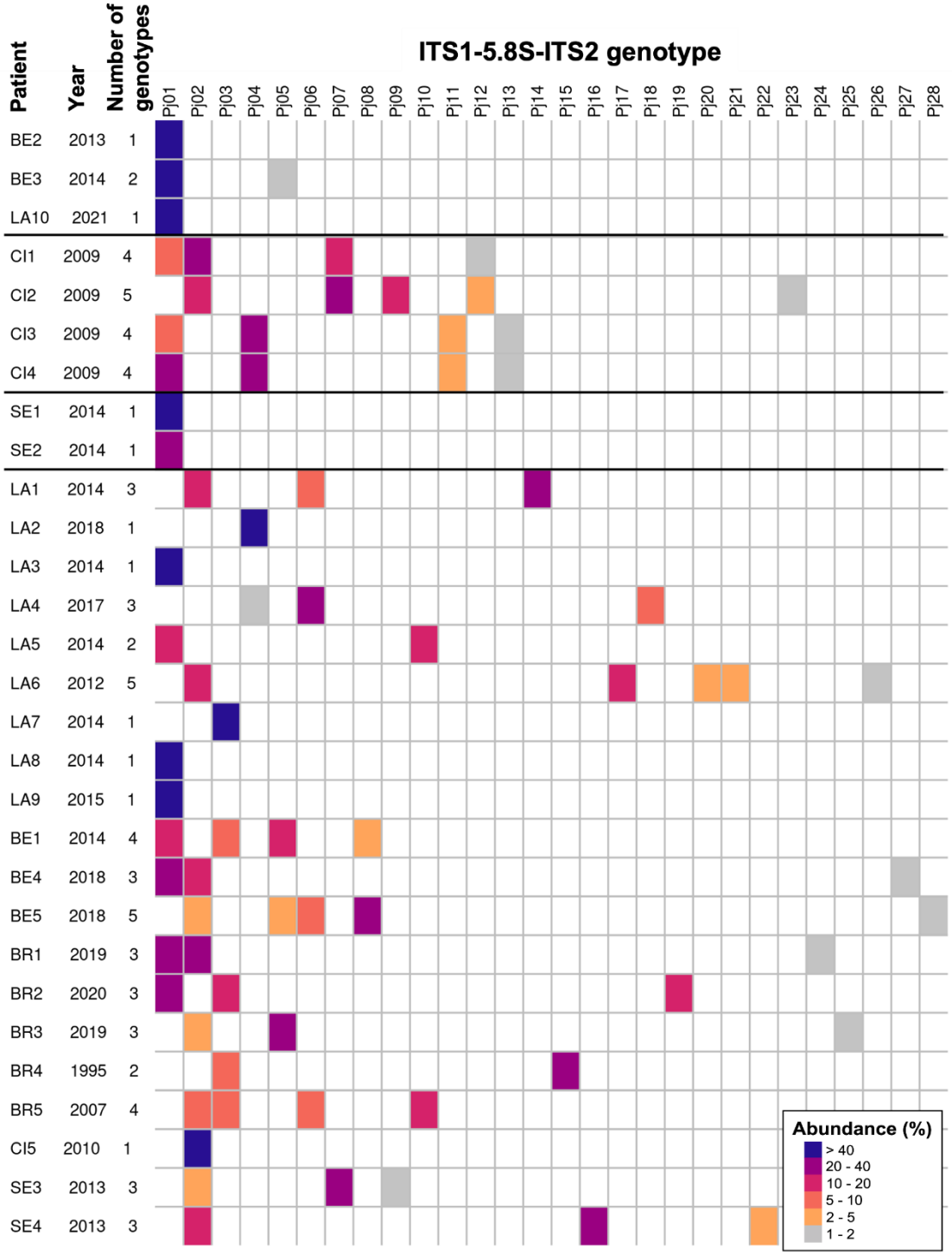


Figure 14. Identity and abundance of the *P. jirovecii* genotypes present in each of the 29 patients. The Swiss, American and Spanish clusters are shown at the top of the figure. The genotypes are ordered in decreasing abundance. Each colored rectangle of the heatmap represents a genotype present in the given patient, with the color figuring its abundance in % of all reads analysed for the patient, as indicated at the bottom right of the figure. The total abundance of the genotype(s) in each patient is below 100% because those under 1% were not considered (see text). The collection year and the number of genotypes present in the patient are indicated next to the patient's name. LA, Lausanne. BE, Bern. BR, Brest. CI, Cincinnati. SE, Seville.

General conclusions

Based on our observations, we propose in chapter 1 an overview of the potential mechanisms involved in the antigenic variation system of *P. jirovecii* (Figure 15). Intergenic recombinations would allow the reassortment of the *msg*-I genes' repertoires through translocations of entire alleles, and possibly also the exchange of the expressed alleles (Figure 15a-b). Single recombinations would lead to the rearrangement of the subtelomeres (Figure 15c-d) and pairs of intragenic recombinations would create new mosaic *msg* genes (Figure 15f). Finally, the recombinations might be mediated by DNA triplexes. Previous studies had shown the importance of mosaicism and exchange of the expressed alleles (Keely et al., 2005; Kutty et al., 2008; Schmid-Siegert et al., 2017). However, we show here for the first time that intergenic recombinations are responsible for the translocation of entire alleles. We also observed for the first time similarities between intergenic spaces close to genes of two different *msg* families, suggesting a new type of subtelomeric rearrangements (Figure 15e).

Various human pathogenic microorganisms possess a mutually exclusive expression system to vary their antigenicity (Deitsch et al., 2009; Vink et al., 2012). This is the case for example of the VSG system of *Trypanosoma brucei* (Faria et al., 2022) and the VAR system of *Plasmodium falciparum* (Frank et al., 2008; Freitas-Junior et al., 2000). However, *P. jirovecii* displays a unique antigenic variation system compared to these organisms. Indeed, it would rely on antigenically heterogeneous populations of cells, whereas the others have homogeneous ones. A major difference between these pathogens is their niche. That of *P. jirovecii* is the human lungs, an environment harbouring its own microbiota because it is in contact with the outside air and thus with

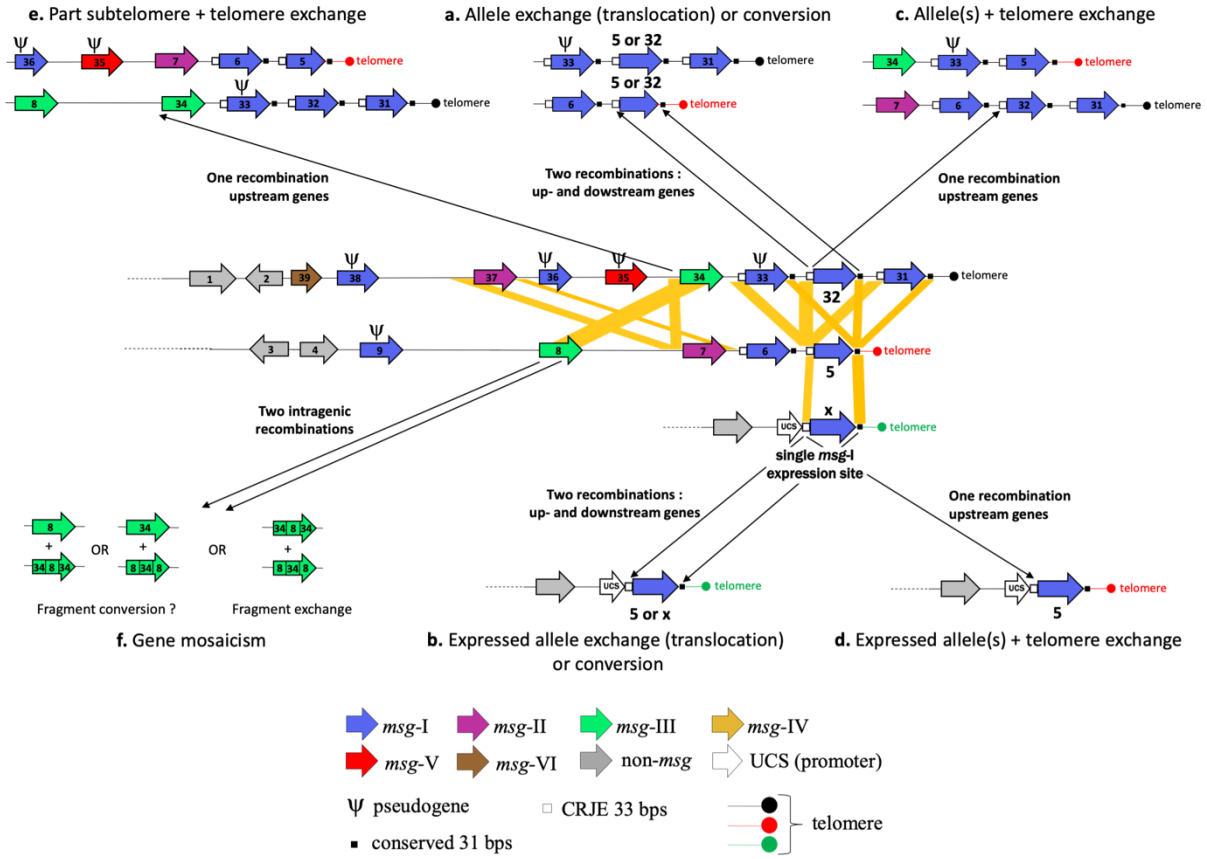


Figure 15. Model of the various recombination mechanisms involved in the antigenic variation system of *P. jirovecii*. The subtelomeres shown are adapted from Fig. 3 of reference (Schmid-Siegert et al., 2017) (*i.e.* contigs 54 and 18). The subtelomere carrying the UCS (upstream conserved sequence) contains the promoter present at a single copy per genome (*i.e.* contig 72). Relevant similarities between the three subtelomeres are shown by the yellow parallelograms. Recombinations within these similarities lead to the mechanisms indicated by the black arrows. DNA triplexes potentially formed by the CRJE (conserved recombination junction element) sequences might mediate the homologous recombinations.

The following recombination mechanisms are illustrated in this figure:

- **a.** Two homologous recombinations, one within the conserved CRJE and one within the conserved downstream region, would lead to the translocation of the entire *msg*-I genes.
- **b.** Two homologous recombinations, one within the CRJE next to the UCS and one within the downstream region, would lead to the exchange or conversion of the entire expressed *msg*-I gene.
- **c.** A single recombination between two CRJE sequences would lead to the exchange of one or several genes together with the telomere linked to them.
- **d.** A single recombination between the CRJE next to the UCS and a second CRJE sequence would lead to the exchange of the expressed gene as well as all downstream genes and the telomere linked to them.
- **e.** Similarities between the upstream regions of families II and II of the *msg* genes may promote exchanges of large parts of subtelomeres through a single homologous recombination at these locations.
- **f.** Intragenic recombinations would lead to gene mosaicism due to fragment exchange or conversion.

Figure and legend adapted from Meier et al., submitted.

a myriad of organisms, while that of *T. brucei* and *P. falciparum* is the bloodstream and tissues, which are sterile environments. In the lungs, it might be more advantageous to present multiple epitopes, such as the antigenically heterogeneous populations of *P. jirovecii*, that mimic the low abundance of fungi present in the lung microbiome. On the other hand, the homogeneity of the population might be advantageous within sterile niches because the pathogen faces there a very strong immune reaction due to the intolerance of any organisms (Hauser, 2019).

Three clusters of patients with PCP were detected by their identical or similar complete *msg*-I repertoires. The genotyping of the strains infecting the patients involved and the epidemiological analysis of the clusters suggested putative distinct transmission routes. The transplant recipients who form a cluster that lasts a relatively short timeframe of a few years are likely to have been infected through direct interhuman transmission during encounters at the hospital. A stability of the *msg*-I repertoire within a cluster might reflect the involvement of unknown carriers or infected individuals in a chain of transmission, and/or the reactivation of a dormant state of the fungus present within the environment or, possibly, the host's lungs. One of our observations suggests that this stability might not be occurring among patients with other underlying conditions, suggesting the working hypothesis that the speed of the evolution of the repertoires might vary due to the environment surrounding *P. jirovecii*, such as the pressure exerted by the immune system.

The dormancy state we hypothesized might be a form where most cellular processes and metabolism are stopped (Maire et al., 2020). Many fungal species have known dormant states, such as *Saccharomyces cerevisiae*, *Schizosaccharomyces.*

pombe and *Aspergillus fumigatus* (Dworkin & Shah, 2010). The extant of the metabolic halt can vary among species. The production of the ascospores is a hallmark of the Ascomycetes phylum, to which *P. jirovecii* belongs. The asci and/or the ascospores they contain are the putative dormant form in this phylum. The elongated ascospores with a condensed cytoplasm (Figure 3, blue cells) are good candidates because they are similar to the dehydrated bacterial spores that are known to remain dormant up to thousands of years. Such dormant cells are formed in response to stress (Hatanaka & Shimoda, 2001) and are able to survive over a long period of time in presence of various environmental insults (Neiman, 2005). Hence, as such dormant states are found in various fungi from the phylum that includes *Pneumocystis* species, and as asci and ascospores are produced by *P. jirovecii*, the presence of a similar dormant state of this fungus appears plausible.

Perspectives

Although the study of the *msg*-I repertoires and genotypes harboured by 29 patients enabled a better understanding of the mechanisms involved in the antigenic variation of *P. jirovecii* and modes of interhuman transmission, there are several perspectives that could take further the project to understand these issues.

Adding more patients in this analysis may increase the number of *msg*-I distinct alleles to the reservoir studied. This would improve the understanding of the importance of the size of the reservoir versus the creation of new alleles in the antigenic variation system (mosaicism). Additionally, the analysis of more patients may detect more clusters of patients with similar or identical repertoires and thus help increasing our knowledge of the modes of transmission of *P. jirovecii*.

Studying immunocompetent individuals colonized with *P. jirovecii* would give insights into the impact of the immune system on the reassortment and diversity of the *msg*-I repertoires. The evolution in the number of alleles present in each repertoire, especially in the expressed ones, as well as the number of genotypes present in the infection, may be impacted by the immune status of the host.

A wet-lab approach to study CRJE is necessary to confirm the hypothesis emitted here, *i.e.* that the DNA triplex potentially formed by the CRJE mediates the recombination events in the subtelomeric regions, possibly particularly those affecting the *msg*-I genes due to their proximity. As *P. jirovecii* is presently not cultivable, inserting the CRJE sequence into a plasmid vector of a model organism, such as *Escherichia coli*, *S. cerevisiae* or *S. pombe*, might reveal the effects of this sequence

on different cellular or metabolic parameters, possibly on recombination. RNA-Seq could also be used to link this sequence with specific pathways through differential gene expression.

The novel ITSs genotyping technique using PacBio CCS developed here requires more validation by its comparison to an established genotyping method relying on a different sequencing technique, such as multilocus sequence typing (MLST). Subcloning of the PCR products followed by Sanger sequencing would confirm the ITSs genotypes observed by PacBio CCS, as would the amplification of given ITS-5.8S-ITS2 alleles with specific primers, similarly to the controls we performed in chapter 1 for the *msg*-I alleles.

References

Alanio, A., Hauser, P. M., Lagrou, K., Melchers, W. J. G., Helweg-Larsen, J., Matos, O., Cesaro, S., Maschmeyer, G., Einsele, H., Donnelly, J. P., Cordonnier, C., Maertens, J., Bretagne, S., Agrawal, S., Kibbler, C., Pagliuca, A., Ward, K., Akova, M., Herbrecht, R., … Wood, C. (2016). ECIL guidelines for the diagnosis of *Pneumocystis jirovecii* pneumonia in patients with haematological malignancies and stem cell transplant recipients. *Journal of Antimicrobial Chemotherapy*, *71*(9), 2386–2396. https://doi.org/10.1093/jac/dkw156

Aliouat-Denis, C. M., Chabé, M., Demanche, C., Aliouat, E. M., Viscogliosi, E., Guillot, J., Delhaes, L., & Dei-Cas, E. (2008). Pneumocystis species co-evolution and pathogenic power. *Infection, Genetics and Evolution*, *8*(5), 708–726. https://doi.org/10.1016/j.meegid.2008.05.001

Aliouat-Denis, C. M., Martinez, A., Aliouat, E. M., Pottier, M., Gantois, N., & Dei-Cas, E. (2009). The Pneumocystis life cycle. *Memórias Do Instituto Oswaldo Cruz*, *104*(3), 419–426. https://doi.org/10.1590/S0074-02762009000300004

Almeida, J. M. G. C. F., Cissé, O. H., Fonseca, Á., Pagni, M., & Hauser, P. M. (2015). Comparative genomics suggests primary homothallism of *Pneumocystis* species. *MBio*, *6*(1). https://doi.org/10.1128/mBio.02250-14

Auguie, B. (2017). *gridExtra: Miscellaneous functions for "grid" graphics*. https://cran.r-project.org/package=gridExtra

Azar, M. M., Cohen, E., Ma, L., Cissé, O. H., Gan, G., Deng, Y., Belfield, K., Asch, W., Grant, M., Gleeson, S., Koff, A., Gaston, D. C., Topal, J., Curran, S., Kulkarni, S., Kovacs, J. A., & Malinis, M. (2022). Genetic and epidemiologic analyses of an outbreak of *Pneumocystis jirovecii* pneumonia among kidney transplant recipients in the United States. *Clinical Infectious Diseases*, *74*(4), 639–647. https://doi.org/10.1093/cid/ciab474

Bacolla, A., Wang, G., & Vasquez, K. M. (2015). New perspectives on DNA and RNA triplexes as effectors of biological activity. *PLOS Genetics*, *11*(12), e1005696. https://doi.org/10.1371/journal.pgen.1005696

Barry, J. D., Ginger, M. L., Burton, P., & McCulloch, R. (2003). Why are parasite contingency genes often associated with telomeres? *International Journal for Parasitology*, *33*(1), 29–45. https://doi.org/10.1016/S0020-7519(02)00247-3

Bartlett, M., Cushion, M. T., Fishman, J. A., Kaneshiro, E. S., Lee, C. H., Leibowitz, M. J., Lu, J. J., Lundgren, B., Peters, S. E., & Smith, J. . (1994). Revised nomenclature for Pneumocystis carinii. The *Pneumocystis* workshop. *The Journal of Eukaryotic Microbiology*, *41*(5), 121S-122S. http://www.ncbi.nlm.nih.gov/pubmed/7804215

Bartlett, M., Vermund, S. H., Jacobs, R., Durant, P. J., Shaw, M. M., Smith, J. W., Tang, X., Lu, J. J., Li, B., Jin, S., & Lee, C. H. (1997). Detection ofPneumocystis carinii DNA in air samples: likely environmental risk to susceptible persons. *Journal of Clinical Microbiology*, *35*(10), 2511–2513. https://doi.org/10.1128/jcm.35.10.2511-2513.1997

Beck, J. M., Warnock, M. L., Kaltreider, H. B., & Shellito, J. E. (1993). Host defenses against *Pneumocystis carinii* in mice selectively depleted of CD4+ lymphocytes.

Chest, *103*(2), 116S-118S. https://doi.org/10.1378/chest.103.2_Supplement.116S-a

Beser, J., Hagblom, P., & Fernandez, V. (2007). Frequent *in vitro* recombination in internal transcribed spacers 1 and 2 during genotyping of *Pneumocystis jirovecii*. *Journal of Clinical Microbiology*, *45*(3), 881–886. https://doi.org/10.1128/JCM.02245-06

Bishop, L. R., Davis, A. S., Bradshaw, K., Gamez, M., Cissé, O. H., Wang, H., Ma, L., & Kovacs, J. A. (2018). Characterization of p57, a stage-specific antigen of *Pneumocystis murina*. *The Journal of Infectious Diseases*, *218*(2), 282–290. https://doi.org/10.1093/infdis/jiy099

Bishop, L. R., & Kovacs, J. A. (2003). Quantitation of anti– *Pneumocystis jiroveci* antibodies in healthy persons and immunocompromised patients. *The Journal of Infectious Diseases*, *187*(12), 1844–1848. https://doi.org/10.1086/375354

Blasi, B., Sipos, W., Knecht, C., Dürlinger, S., Ma, L., Cissé, O. H., Nedorost, N., Matt, J., Weissenböck, H., & Weissenbacher-Lang, C. (2021). *Pneumocystis* spp. in pigs: A longitudinal quantitative study and co-infection assessment in austrian farms. *Journal of Fungi*, *8*(1), 43. https://doi.org/10.3390/jof8010043

Bongomin, F., Gago, S., Oladele, R., & Denning, D. (2017). Global and multi-national prevalence of fungal diseases—estimate precision. *Journal of Fungi*, *3*(4), 57. https://doi.org/10.3390/jof3040057

Britten, R. J. (1998). Precise sequence complementarity between yeast chromosome ends and two classes of just-subtelomeric sequences. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(11), 5906–5912. https://doi.org/10.1073/pnas.95.11.5906

Brown, G. D., Denning, D. W., Gow, N. A. R., Levitz, S. M., Netea, M. G., & White, T. C. (2012). Hidden killers: Human fungal infections. *Science Translational Medicine*, *4*(165), 1–10. https://doi.org/10.1126/scitranslmed.3004404

Buske, F. A., Mattick, J. S., & Bailey, T. L. (2011). Potential *in vivo* roles of nucleic acid triple-helices. *RNA Biology*, *8*(3), 427–439. https://doi.org/10.4161/rna.8.3.14999

Calderón-Sandubete, E. J., Varela-Aguilar, J. M., Medrano-Ortega, F. J., Nieto-Guerrero, V., Respaldiza-Salas, N., De la Horra-Padilla, C., & Dei-Cas, E. (2002). Historical perspective on *Pneumocystis carinii* infection. *Protist*, *153*(3), 303–310. https://doi.org/10.1078/1434-4610-00107

Chabé, M., Dei-Cas, E., Creusy, C., Fleurisse, L., Respaldiza, N., Camus, D., & Durand-Joly, I. (2004). Immunocompetent hosts as a reservoir of Pneumocystis organisms: histological and RT-PCR data demonstrate active replication. *European Journal of Clinical Microbiology & Infectious Diseases*, *23*(2), 89–97. https://doi.org/10.1007/s10096-003-1092-2

Chabé, M., Vargas, S. L., Eyzaguirre, I., Aliouat, E. M., Follet-Dumoulin, A., Creusy, C., Fleurisse, L., Recourt, C., Camus, D., Dei-Cas, E., & Durand-Joly, I. (2004). Molecular typing of *Pneumocystis jirovecii* found in formalin-fixed paraffin-embedded lung tissue sections from sudden infant death victims. *Microbiology*, *150*(5), 1167–1172. https://doi.org/10.1099/mic.0.26895-0

Chagas, C. (1909). Nova tripanozomiaze humana: estudos sobre a morfolojia e o

ciclo evolutivo do *Schizotrypanum cruzi* n. gen., n. sp., ajente etiolojico de nova entidade morbida do homem. *Memórias Do Instituto Oswaldo Cruz*, *1*(2), 159–218. https://doi.org/10.1590/S0074-02761909000200008

Charif, D., & Lobry, J. R. (2007). Seqin{R} 1.0-2: a contributed package to the {R} project for statistical computing devoted to biological sequences retrieval and analysis. In U. Bastolla, M. Porto, H. E. Roman, & M. Vendruscolo (Eds.), *Structural approaches to sequence evolution: Molecules, networks, populations* (pp. 207–232). Springer Verlag.

Charpentier, E., Ménard, S., Marques, C., Berry, A., & Iriart, X. (2021). Immune response in *Pneumocystis* infections according to the host immune system status. *Journal of Fungi*, *7*(8), 625. https://doi.org/10.3390/jof7080625

Chen, J. M., Cooper, D. N., Chuzhanova, N., Férec, C., & Patrinos, G. P. (2007). Gene conversion: Mechanisms, evolution and human disease. *Nature Reviews Genetics*, *8*(10), 762–775. https://doi.org/10.1038/nrg2193

Cissé, O. H., Ma, L., Dekker, J. P., Khil, P. P., Youn, J.-H., Brenchley, J. M., Blair, R., Pahar, B., Chabé, M., Van Rompay, K. K. A., Keesler, R., Sukura, A., Hirsch, V., Kutty, G., Liu, Y., Peng, L., Chen, J., Song, J., Weissenbacher-Lang, C., … Kovacs, J. A. (2021). Genomic insights into the host specific adaptation of the *Pneumocystis* genus. *Communications Biology*, *4*(1), 305. https://doi.org/10.1038/s42003-021-01799-7

Cissé, O. H., Ma, L., Jiang, C., Snyder, M., & Kovacs, J. A. (2020). Humans are selectively exposed to *Pneumocystis jirovecii*. *MBio*, *11*(2). https://doi.org/10.1128/mBio.03138-19

Cissé, O. H., Pagni, M., & Hauser, P. M. (2013). De novo assembly of the *Pneumocystis jirovecii* genome from a single bronchoalveolar lavage fluid specimen from a patient. *MBio*, *4*(1), 1–4. https://doi.org/10.1128/mBio.00428-12

Cox, R., & Mirkin, S. M. (1997). Characteristic enrichment of DNA repeats in different genomes. *Proceedings of the National Academy of Sciences of the United States of America*, *94*(10), 5237–5242. http://www.jstor.org/stable/42480

Cushion, M. T., Keely, S. P., & Stringer, J. R. (2004). Molecular and phenotypic description of *Pneumocystis wakefieldiae* sp. nov., a new species in rats. *Mycologia*, *96*(3), 429. https://doi.org/10.2307/3762163

Cushion, M. T., Linke, M. J., Ashbaugh, A., Sesterhenn, T., Collins, M. S., Lynch, K., Brubaker, R., & Walzer, P. D. (2010). Echinocandin treatment of *Pneumocystis* Pneumonia in rodent models depletes cysts leaving trophic burdens that cannot transmit the infection. *PLoS ONE*, *5*(1), e8524. https://doi.org/10.1371/journal.pone.0008524

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

Cushion, M. T., Zhang, J., Kaselis, M., Giuntoli, D., Stringer, S. L., & Stringer, J. R. (1993). Evidence for two genetic variants of *Pneumocystis carinii* coinfecting laboratory rats. *Journal of Clinical Microbiology*, *31*(5), 1217–1223. https://doi.org/10.1128/jcm.31.5.1217-1223.1993

Daly, K. R., Huang, L., Morris, A., Koch, J., Crothers, K., Levin, L., Eiser, S., Satwah, S., Zucchi, P., & Walzer, P. D. (2006). Antibody response to *Pneumocystis jirovecii* major surface glycoprotein. *Emerging Infectious Diseases*, *12*(8), 1231–1237. https://doi.org/10.3201/eid1708.060230

Dayn, A., Samadashwily, G. M., & Mirkin, S. M. (1992). Intramolecular DNA triplexes: unusual sequence requirements and influence on DNA polymerization. *Proceedings of the National Academy of Sciences*, *89*(23), 11406–11410. https://doi.org/10.1073/pnas.89.23.11406

de Boer, M. G. J., Bruijnesteijn van Coppenraet, L. E. S., Gaasbeek, A., Berger, S. P., Gelinck, L. B. S., van Houwelingen, H. C., van den Broek, P., Kuijper, E. J., Kroon, F. P., & Vandenbroucke, J. P. (2007). An outbreak of *Pneumocystis jiroveci* pneumonia with 1 predominant genotype among renal transplant recipients: interhuman transmission or a common environmental source? *Clinical Infectious Diseases*, *44*(9), 1143–1149. https://doi.org/10.1086/513198

Dei-Cas, E., Chabé, M., Moukhlis, R., Durand-Joly, I., Aliouat, E. M., Stringer, J. R., Cushion, M., Noël, C., Sybren de Hoog, G., Guillot, J., & Viscogliosi, E. (2006). *Pneumocystis oryctolagi* sp. nov., an uncultured fungus causing pneumonia in rabbits at weaning: review of current knowledge, and description of a new taxon on genotypic, phylogenetic and phenotypic bases. *FEMS Microbiology Reviews*, *30*(6), 853–871. https://doi.org/10.1111/j.1574-6976.2006.00037.x

Deitsch, K. W., & Dzikowski, R. (2017). Variant gene expression and antigenic variation by malaria parasites. *Annual Review of Microbiology*, *71*(1), 625–641. https://doi.org/10.1146/annurev-micro-090816-093841

Deitsch, K. W., Lukehart, S. A., & Stringer, J. R. (2009). Common strategies for antigenic variation by bacterial, fungal and protozoan pathogens. *Nature Reviews Microbiology*, *7*(7), 493–503. https://doi.org/10.1038/nrmicro2145

Delanoë, P., & Delanoë, M. (1912). Sur les rapports des kystes de Carini du poumon des rats avec le Trypanosoma lewisi. *CR Acad Sci. Paris*, *155*, 658–660.

Delaye, L., Ruiz-Ruiz, S., Calderon, E., Tarazona, S., Conesa, A., & Moya, A. (2018). Evidence of the red-queen hypothesis from accelerated rates of evolution of genes involved in biotic interactions in *Pneumocystis*. *Genome Biology and Evolution*, *10*(6), 1596–1606. https://doi.org/10.1093/gbe/evy116

Dumoulin, A., Mazars, E., Seguy, N., Gargallo-Viola, D., Vargas, S., Cailliez, J. C., Aliouat, E. M., Wakefield, A. E., & Dei-Cas, E. (2000). Transmission of Pneumocystis carinii disease from immunocompetent contacts of infected hosts to susceptible hosts. *European Journal of Clinical Microbiology & Infectious Diseases*, *19*(9), 671–678. https://doi.org/10.1007/s100960000354

Durand-Joly, I., Aliouat, E. M., Recourt, C., Guyot, K., François, N., Wauquier, M., Camus, D., & Dei-Cas, E. (2002). *Pneumocystis carinii* f. sp. *hominis* is not infectious for SCID mice. *Journal of Clinical Microbiology*, *40*(5), 1862–1865. https://doi.org/10.1128/JCM.40.5.1862-1865.2002

Dworkin, J., & Shah, I. M. (2010). Exit from dormancy in microbial organisms. *Nature Reviews Microbiology*, *8*(12), 890–896. https://doi.org/10.1038/nrmicro2453

Edman, J. C. (1996). A single expression site with a conserved leader sequence regulates variation of expression of the *Pneumocystis carinii* family of major

surface glycoprotein genes. *DNA and Cell Biology*, *15*(11), 989–999. https://doi.org/10.1089/dna.1996.15.989

Edman, J. C., Kovacs, J. A., Masur, H., Santi, D. V, Elwood, H. J., & Sogin, M. L. (1988). Ribosomal RNA sequence shows *Pneumocystis carinii* to be a member of the fungi. *Nature*, *334*(6182), 519–522. https://doi.org/10.1038/334519a0

Eriksson, O., & Winka, K. (1997). Supraordinal taxa of Ascomycota. *Myconet*.

Ezekowitz, R. A. B., Williams, D. J., Koziel, H., Armstrong, M. Y. K., Warner, A., Richards, F. F., & Rose, R. M. (1991). Uptake of *Pneumocystis carinii* mediated by the macrophage mannose receptor. *Nature*, *351*(6322), 155–158. https://doi.org/10.1038/351155a0

Faria, J., Briggs, E. M., Black, J. A., & McCulloch, R. (2022). Emergence and adaptation of the cellular machinery directing antigenic variation in the African trypanosome. *Current Opinion in Microbiology*, *70*, 102209. https://doi.org/10.1016/j.mib.2022.102209

Frank-Kamenetskii, M. D., & Mirkin, S. M. (1995). Triplex DNA structures. *Annual Review of Biochemistry*, *64*(1), 65–95. https://doi.org/10.1146/annurev.bi.64.070195.000433

Frank, M., Kirkman, L., Costantini, D., Sanyal, S., Lavazec, C., Templeton, T. J., & Deitsch, K. W. (2008). Frequent recombination events generate diversity within the multi-copy variant antigen gene families of *Plasmodium falciparum*. *International Journal for Parasitology*, *38*(10), 1099–1109. https://doi.org/10.1016/j.ijpara.2008.01.010

Freitas-Junior, L. H., Bottius, E., Pirrit, L. A., Deitsch, K. W., Scheidig, C., Guinet, F., Nehrbass, U., Wellems, T. E., & Scherf, A. (2000). Frequent ectopic recombination of virulence factor genes in telomeric chromosome clusters of *P. falciparum*. *Nature*, *407*(6807), 1018–1022. https://doi.org/10.1038/35039531

Frenkel, J. K. (1999). *Pneumocystis* pneumonia, an immunodeficiency-dependent disease (IDD): a critical historical overview. *The Journal of Eukaryotic Microbiology*, *46*(5), 89S-92S. http://www.ncbi.nlm.nih.gov/pubmed/10519262

Galili, T. (2015). *dendextend: an R package for visualizing, adjusting, and comparing trees of hierarchical clustering*. https://doi.org/10.1093/bioinformatics/btv428

Gantois, N., Lesaffre, A., Durand-Joly, I., Bautin, N., Le Rouzic, O., Nseir, S., Reboux, G., Scherer, E., Aliouat, E. M., Fry, S., Gosset, P., & Fréalle, E. (2021). Factors associated with *Pneumocystis* colonization and circulating genotypes in chronic obstructive pulmonary disease patients with acute exacerbation or at stable state and their homes. *Medical Mycology*, *60*(1). https://doi.org/10.1093/mmy/myab070

Gianella, S., Haeberli, L., Joos, B., Ledergerber, B., Wüthrich, R. P., Weber, R., Kuster, H., Hauser, P. M., Fehr, T., & Mueller, N. J. (2010). Molecular evidence of interhuman transmission in an outbreak of Pneumocystis *jirovecii pneumonia* among renal transplant recipients. *Transplant Infectious Disease*, *12*(1), 1–10. https://doi.org/10.1111/j.1399-3062.2009.00447.x

Gigliotti, F., Harmsen, A. G., Haidaris, C. G., & Haidaris, P. J. (1993). *Pneumocystis carinii* is not universally transmissible between mammalian species. *Infection and Immunity*, *61*(7), 2886–2890. https://doi.org/10.1128/iai.61.7.2886-

2890.1993

Gigliotti, F., Harmsen, A. G., & Wright, T. W. (2003). Characterization of transmission of *Pneumocystis carinii* f. sp. *muris* through immunocompetent BALB/c mice. *Infection and Immunity*, *71*(7), 3852–3856. https://doi.org/10.1128/IAI.71.7.3852-3856.2003

Gigliotti, F., Limper, A. H., & Wright, T. (2014). *Pneumocystis*. *Cold Spring Harbor Perspectives in Medicine*, *4*(12), a019828. https://doi.org/10.1101/cshperspect.a019828

Goñi, J. R., Vaquerizas, J. M., Dopazo, J., & Orozco, M. (2006a). Exploring the reasons for the large density of triplex-forming oligonucleotide target sequences in the human regulatory regions. *BMC Genomics*, *7*, 1–10. https://doi.org/10.1186/1471-2164-7-63

Goñi, J. R., Vaquerizas, J. M., Dopazo, J., & Orozco, M. (2006b). Exploring the reasons for the large density of triplex-forming oligonucleotide target sequences in the human regulatory regions. *BMC Genomics*, *7*(1), 63. https://doi.org/10.1186/1471-2164-7-63

Hatanaka, M., & Shimoda, C. (2001). The cyclic AMP/PKA signal pathway is required for initiation of spore germination in *Schizosaccharomyces pombe*. *Yeast*, *18*(3), 207–217. https://doi.org/10.1002/1097-0061(200102)18:3<207::AID-YEA661>3.0.CO;2-I

Hauser, P. M. (2019). Is the unique camouflage strategy of *Pneumocystis* associated with its particular niche within host lungs? *PLOS Pathogens*, *15*(1), e1007480. https://doi.org/10.1371/journal.ppat.1007480

Hauser, P. M. (2021). *Pneumocystis* mating-type locus and sexual cycle during infection. *Microbiology and Molecular Biology Reviews*, *85*(3). https://doi.org/10.1128/mmbr.00009-21

Hauser, P. M., Burdet, F. X., Cissé, O. H., Keller, L., Taffé, P., Sanglard, D., & Pagni, M. (2010). Comparative genomics suggests that the fungal pathogen *Pneumocystis* is an obligate parasite scavenging amino acids from its host's lungs. *PLoS ONE*, *5*(12), e15152. https://doi.org/10.1371/journal.pone.0015152

Hauser, P. M., & Cushion, M. T. (2018a). Is sex necessary for the proliferation and transmission of *Pneumocystis*? *PLOS Pathogens*, *14*(12), 1–7. https://doi.org/10.1371/journal.ppat.1007409

Hauser, P. M., & Cushion, M. T. (2018b). Is sex necessary for the proliferation and transmission of Pneumocystis? *PLOS Pathogens*, *14*(12), 1–7. https://doi.org/10.1371/journal.ppat.1007409

Hauser, P. M., Francioli, P., Bille, J., Telenti, A., & Blanc, D. S. (1997). Typing of Pneumocystis carinii f. sp. hominis by single-strand conformation polymorphism of four genomic regions. *Journal of Clinical Microbiology*, *35*(12), 3086–3091. https://doi.org/10.1128/jcm.35.12.3086-3091.1997

Hauser, P. M., Rabodonirina, M., & Nevez, G. (2013). *Pneumocystis jirovecii* genotypes involved in *pneumocystis* pneumonia outbreaks among renal transplant recipients. *Clinical Infectious Diseases*, *56*(1), 165–166. https://doi.org/10.1093/cid/cis810

http://hmmer.org/. (n.d.). *Hmmer* (3.3.2). http://hmmer.org/

Keely, S. P., Fischer, J. M., Cushion, M. T., & Stringer, J. R. (2004). Phylogenetic identification of *Pneumocystis murina* sp. nov., a new species in laboratory mice. *Microbiology*, *150*(5), 1153–1165. https://doi.org/10.1099/mic.0.26921-0

Keely, S. P., Linke, M. J., Cushion, M. T., & Stringer, J. R. (2007). *Pneumocystis murina* MSG gene family and the structure of the locus associated with its transcription. *Fungal Genetics and Biology*, *44*(9), 905–919. https://doi.org/10.1016/j.fgb.2007.01.004

Keely, S. P., Renauld, H., Wakefield, A. E., Cushion, M. T., Smulian, A. G., Fosker, N., Fraser, A., Harris, D., Murphy, L., Price, C., Quail, M. A., Seeger, K., Sharp, S., Tindal, C. J., Warren, T., Zuiderwijk, E., Barrell, B. G., Stringer, J. R., & Hall, N. (2005). Gene Arrays at *Pneumocystis carinii* Telomeres. *Genetics*, *170*(4), 1589–1600. https://doi.org/10.1534/genetics.105.040733

Keely, S. P., & Stringer, J. R. (2009). Complexity of the MSG gene family of *Pneumocystis carinii*. *BMC Genomics*, *10*, 367. https://doi.org/10.1186/1471-2164-10-367

Kohwi, Y., & Panchenko, Y. (1993). Transcription-dependent recombination induced by triple-helix formation. *Genes & Development*, *7*(9), 1766–1778. https://doi.org/10.1101/gad.7.9.1766

Kottom, T. J., Hebrink, D. M., Jenson, P. E., Marsolek, P. L., Wüthrich, M., Wang, H., Klein, B., Yamasaki, S., & Limper, A. H. (2018). Dectin-2 Is a C-type lectin receptor that recognizes *Pneumocystis* and participates in innate immune responses. *American Journal of Respiratory Cell and Molecular Biology*, *58*(2), 232–240. https://doi.org/10.1165/rcmb.2016-0335OC

Kottom, T. J., Hebrink, D. M., Jenson, P. E., Nandakumar, V., Wüthrich, M., Wang, H., Klein, B., Yamasaki, S., Lepenies, B., & Limper, A. H. (2017). The Interaction of *Pneumocystis* with the C-type lectin receptor Mincle exerts a significant role in host defense against infection. *The Journal of Immunology*, *198*(9), 3515–3525. https://doi.org/10.4049/jimmunol.1600744

Kottom, T. J., Hebrink, D. M., & Limper, A. H. (2018). Binding of *Pneumocystis carinii* to the lung epithelial cell receptor HSPA5 (GRP78). *Journal of Medical Microbiology*, *67*(12), 1772–1777. https://doi.org/10.1099/jmm.0.000864

Kottom, T. J., & Limper, A. H. (2000). Cell Wall Assembly by *Pneumocystis carinii*. *Journal of Biological Chemistry*, *275*(51), 40628–40634. https://doi.org/10.1074/jbc.M002103200

Kutty, G., Davis, A. S., Ma, L., Taubenberger, J. K., & Kovacs, J. A. (2015). *Pneumocystis* encodes a functional endo-β-1,3-glucanase that is expressed exclusively in cysts. *The Journal of Infectious Diseases*, *211*(5), 719–728. https://doi.org/10.1093/infdis/jiu517

Kutty, G., Ma, L., & Kovacs, J. A. (2001). Characterization of the expression site of the major surface glycoprotein of human-derived *Pneumocystis carinii*. *Molecular Microbiology*, *42*(1), 183–193. https://doi.org/10.1046/j.1365-2958.2001.02620.x

Kutty, G., Maldarelli, F., Achaz, G., & Kovacs, J. A. (2008). Variation in the major surface glycoprotein genes in *Pneumocystis jirovecii*. *Journal of Infectious Diseases*, *198*(5), 741–749. https://doi.org/10.1086/590433

Kutty, G., Shroff, R., & Kovacs, J. A. (2013). Characterization of *Pneumocystis* major

surface glycoprotein gene (msg) promoter activity in *Saccharomyces cerevisiae*. *Eukaryotic Cell*, *12*(10), 1349–1355. https://doi.org/10.1128/EC.00122-13

Laakkonen, J., Sukura, A., Haukisalmi, V., & Henttonen, H. (1993). *Pneumocystis carinii* and helminth parasitism in shrews *Sorex araneus* and *Sorex caecutiens*. *Journal of Wildlife Diseases*, *29*(2), 273–277. https://doi.org/10.7589/0090-3558-29.2.273

Lang, D. M. (2007). Imperfect DNA mirror repeats in the gag gene of HIV-1 (HXB2) identify key functional domains and coincide with protein structural elements in each of the mature proteins. *Virology Journal*, *4*(1), 113. https://doi.org/10.1186/1743-422X-4-113

Le Gal, S., Pougnet, L., Damiani, C., Fréalle, E., Guéguen, P., Virmaux, M., Ansart, S., Jaffuel, S., Couturaud, F., Delluc, A., Tonnelier, J.-M., Castellant, P., Le Meur, Y., Le Floch, G., Totet, A., Menotti, J., & Nevez, G. (2015). *Pneumocystis jirovecii* in the air surrounding patients with *Pneumocystis* pulmonary colonization. *Diagnostic Microbiology and Infectious Disease*, *82*(2), 137–142. https://doi.org/10.1016/j.diagmicrobio.2015.01.004

Lemon, J. (2006). Plotrix: a package in the red light district of R. *R-News*, *6*(4), 8–12.

Li, W., & Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, *22*(13), 1658–1659. https://doi.org/10.1093/bioinformatics/btl158

Lima, S. L., Colombo, A. L., & de Almeida Junior, J. N. (2019). Fungal cell wall: emerging antifungals and drug resistance. *Frontiers in Microbiology*, *10*. https://doi.org/10.3389/fmicb.2019.02573

Limper, A. H., Hoyte, J. S., & Standing, J. E. (1997). The role of alveolar macrophages in *Pneumocystis carinii* degradation and clearance from the lung. *Journal of Clinical Investigation*, *99*(9), 2110–2117. https://doi.org/10.1172/JCI119384

Liu, Y., Fahle, G. A., & Kovacs, J. A. (2018). Inability to culture *Pneumocystis jirovecii*. *MBio*, *9*(3). https://doi.org/10.1128/mBio.00939-18

Luraschi, A., Richard, S., Almeida, J. M. G. C. F., Pagni, M., Cushion, M. T., & Hauser, P. M. (2019). Expression and immunostaining analyses suggest that *Pneumocystis* primary homothallism involves trophic cells displaying both plus and minus pheromone receptors. *MBio*, *10*(4). https://doi.org/10.1128/mBio.01145-19

Ma, L., Chen, Z., Huang, D. W., Cissé, O. H., Rothenburger, J. L., Latinne, A., Bishop, L., Blair, R., Brenchley, J. M., Chabé, M., Deng, X., Hirsch, V., Keesler, R., Kutty, G., Liu, Y., Margolis, D., Morand, S., Pahar, B., Peng, L., … Kovacs, J. A. (2020). Diversity and complexity of the large surface protein family in the compacted genomes of multiple *Pneumocystis* species. *MBio*, *11*(2), 1–20. https://doi.org/10.1128/mBio.02878-19

Ma, L., Chen, Z., Huang, D. W., Kutty, G., Ishihara, M., Wang, H., Abouelleil, A., Bishop, L., Davey, E., Deng, R., Deng, X., Fan, L., Fantoni, G., Fitzgerald, M., Gogineni, E., Goldberg, J. M., Handley, G., Hu, X., Huber, C., … Kovacs, J. A. (2016). Genome analysis of three *Pneumocystis* species reveals adaptation mechanisms to life exclusively in mammalian hosts. *Nature Communications*, *7*.

https://doi.org/10.1038/ncomms10740

Ma, L., Cissé, O. H., & Kovacs, J. A. (2018). A molecular window into the biology and epidemiology of *Pneumocystis* spp. *Clinical Microbiology Reviews*, *31*(3), 1–49. https://doi.org/10.1128/CMR.00009-18

Mahé, F., Rognes, T., Quince, C., de Vargas, C., & Dunthorn, M. (2014). Swarm: Robust and fast clustering method for amplicon-based studies. *PeerJ*, *2014*(1), 1–13. https://doi.org/10.7717/peerj.593

Maire, T., Allertz, T., Betjes, M. A., & Youk, H. (2020). Dormancy-to-death transition in yeast spores occurs due to gradual loss of gene-expressing ability. *Molecular Systems Biology*, *16*(11). https://doi.org/10.15252/msb.20199245

Maitte, C., Leterrier, M., Le Pape, P., Miegeville, M., & Morio, F. (2013). Multilocus sequence typing of *Pneumocystis jirovecii* from clinical samples: how many and which loci should be used? *Journal of Clinical Microbiology*, *51*(9), 2843–2849. https://doi.org/10.1128/JCM.01073-13

Martinez, A., Aliouat, E. M., Standaert-Vitse, A., Werkmeister, E., Pottier, M., Pinçon, C., Dei-Cas, E., & Aliouat-Denis, C. M. (2011). Ploidy of cell-sorted trophic and cystic forms of *Pneumocystis carinii*. *PLoS ONE*, *6*(6), e20935. https://doi.org/10.1371/journal.pone.0020935

Martinez, A., Halliez, M. C. M., Aliouat, E. M., Chabé, M., Standaert-Vitse, A., Fréalle, E., Gantois, N., Pottier, M., Pinon, A., Dei-Cas, E., & Aliouat-Denis, C. M. (2013). Growth and airborne transmission of cell-sorted life cycle stages of Pneumocystis carinii. *PLoS ONE*, *8*(11), 1–8. https://doi.org/10.1371/journal.pone.0079958

Medrano, F. J., Montes-Cano, M., Conde, M., de la Horra, C., Respaldiza, N., Gasch, A., Perez-Lozano, M. J., Varela, J. M., & Calderon, E. J. (2005). *Pneumocystis jirovecii* in general population. *Emerging Infectious Diseases*, *11*(2), 245–250. https://doi.org/10.3201/eid1102.040487

Menotti, J., Emmanuel, A., Bouchekouk, C., Chabe, M., Choukri, F., Pottier, M., Sarfati, C., Aliout, E. M., & Derouin, F. (2013). Evidence of airborne excretion of *Pneumocystis carinii* during infection in immunocompetent rats. Lung involvement and antibody response. *PLoS ONE*, *8*(4), e62155. https://doi.org/10.1371/journal.pone.0062155

Miller, R. F., Ambrose, H. E., & Wakefield, A. E. (2001). *Pneumocystis carinii* f. sp. *hominis* DNA in immunocompetent health care workers in contact with patients with *P. carinii* pneumonia. *Journal of Clinical Microbiology*, *39*(11), 3877–3882. https://doi.org/10.1128/JCM.39.11.3877-3882.2001

Mirkin, S. M. (2008). Discovery of alternative DNA structures: a heroic decade (1979-1989). *Frontiers in Bioscience*, *13*(13), 1064. https://doi.org/10.2741/2744

Mirkin, S. M., & Frank-Kamenetskii, M. D. (1994). H-DNA and related structures. *Annual Review of Biophysics and Biomolecular Structure*, *23*(1), 541–576. https://doi.org/10.1146/annurev.bb.23.060194.002545

Mirkin, S. M., Lyamichev, V. I., Drushlyak, K. N., Dobrynin, V. N., Filippov, S. A., & Frank-Kamenetskii, M. D. (1987). DNA H form requires a homopurine-homopyrimidine mirror repeat. *Nature*, *330*(6147), 495–497. https://doi.org/10.1038/330495a0

Montes-Cano, M. A., Chabe, M., Fontillon-Alberdi, M., de la Horra, C., Respaldiza, N., Medrano, F. J., Varela, J. M., Dei-Cas, E., & Calderon, E. J. (2009). Vertical transmission of *Pneumocystis jirovecii* in humans. *Emerging Infectious Diseases*, *15*(1), 125–127. https://doi.org/10.3201/eid1501.080242

Morgan, M. (2021). *BiocManager: Access the Bioconductor project package repository*. https://cran.r-project.org/package=BiocManager

Morris, A., Ben Beard, C., & Huang, L. (2002). Update on the epidemiology and transmission of *Pneumocystis carinii*. *Microbes and Infection*, *4*(1), 95–103. https://doi.org/10.1016/S1286-4579(01)01514-3

Nahimana, A., Blanc, D. S., Francioli, P., Bille, J., & Hauser, P. M. (2000). Typing of *Pneumocystis carinii* f. sp. *hominis* by PCR-SSCP to indicate a high frequency of co-infections. *Journal of Medical Microbiology*, *49*(8), 753–758. https://doi.org/10.1099/0022-1317-49-8-753

Navarro, M., & Gull, K. (2001). A pol I transcriptional body associated with VSG mono-allelic expression in *Trypanosoma brucei*. *Nature*, *414*(6865), 759–763. https://doi.org/10.1038/414759a

Neiman, A. M. (2005). Ascospore formation in the yeast *Saccharomyces cerevisiae*. *Microbiology and Molecular Biology Reviews*, *69*(4), 565–584. https://doi.org/10.1128/MMBR.69.4.565-584.2005

Nevez, G., Magois, E., Duwat, H., Gouilleux, V., Jounieaux, V., & Totet, A. (2006). Apparent absence of *Pneumocystis jirovecii* in healthy subjects. *Clinical Infectious Diseases*, *42*(11), e99–e101. https://doi.org/10.1086/503908

Oizumi, Y., Kaji, T., Tashiro, S., Takeshita, Y., Date, Y., & Kanoh, J. (2021). Complete sequences of *Schizosaccharomyces pombe* subtelomeres reveal multiple patterns of genome variation. *Nature Communications*, *12*(1), 3–8. https://doi.org/10.1038/s41467-020-20595-1

Opata, M. M., Hollifield, M. L., Lund, F. E., Randall, T. D., Dunn, R., Garvy, B. A., & Feola, D. J. (2015). B lymphocytes are required during the early priming of CD4+ T cells for clearance of *Pneumocystis* infection in mice. *The Journal of Immunology*, *195*(2), 611–620. https://doi.org/10.4049/jimmunol.1500112

Pagès, H., Aboyoun, P., Gentleman, R., & DebRoy, S. (2019). *Biostrings: Efficient manipulation of biological strings*.

Paysan-Lafosse, T., Blum, M., Chuguransky, S., Grego, T., Pinto, B. L., Salazar, G. A., Bileschi, M. L., Bork, P., Bridge, A., Colwell, L., Gough, J., Haft, D. H., Letunić, I., Marchler-Bauer, A., Mi, H., Natale, D. A., Orengo, C. A., Pandurangan, A. P., Rivoire, C., … Bateman, A. (2023). InterPro in 2022. *Nucleic Acids Research*, *51*(D1), D418–D427. https://doi.org/10.1093/nar/gkac993

Pliquett, R. U., Asbe-Vollkopf, A., Hauser, P. M., Presti, L. L., Hunfeld, K. P., Berger, A., Scheuermann, E. H., Jung, O., Geiger, H., & Hauser, I. A. (2012). A *Pneumocystis jirovecii* pneumonia outbreak in a single kidney-transplant center: role of cytomegalovirus co-infection. *European Journal of Clinical Microbiology & Infectious Diseases*, *31*(9), 2429–2437. https://doi.org/10.1007/s10096-012-1586-x

Ponce, C. A., Gallo, M., Bustamante, R., & Vargas, S. L. (2010). *Pneumocystis*

colonization Is highly prevalent in the autopsied lungs of the general population. *Clinical Infectious Diseases*, *50*(3), 347–353. https://doi.org/10.1086/649868

Pottratz, S. T., & Martin, W. J. (1990). Role of fibronectin in *Pneumocystis carinii* attachment to cultured lung cells. *Journal of Clinical Investigation*, *85*(2), 351–356. https://doi.org/10.1172/JCI114445

Pottratz, S. T., Paulsrud, J., Smith, J. S., & Martin, W. (1991). *Pneumocystis carinii* attachment to cultured lung cells by *pneumocystis* gp120, a fibronectin binding protein. *Journal of Clinical Investigation*, *88*(2), 403–407. https://doi.org/10.1172/JCI115318

Prucca, C. G., Slavin, I., Quiroga, R., Elías, E. V., Rivero, F. D., Saura, A., Carranza, P. G., & Luján, H. D. (2008). Antigenic variation in *Giardia lamblia* is regulated by RNA interference. *Nature*, *456*(7223), 750–754. https://doi.org/10.1038/nature07585

Rabodonirina, M., Vanhems, P., Couray-Targe, S., Gillibert, R.-P., Ganne, C., Nizard, N., Colin, C., Fabry, J., Touraine, J.-L., van Melle, G., Nahimana, A., Francioli, P., & Hauser, P. M. (2004). Molecular evidence of interhuman transmission of *Pneumocystis* pneumonia among renal transplant recipients hospitalized with HIV-infected patients. *Emerging Infectious Diseases*, *10*(10), 1767–1773. https://doi.org/10.3201/eid1010.040453

Richard, S., Almeida, J. M. G. C. F., Cissé, O. H., Luraschi, A., Nielsen, O., Pagni, M., & Hauser, P. M. (2018). Functional and expression analyses of the *Pneumocystis* mat genes suggest obligate sexuality through primary homothallism within host lungs. *MBio*, *9*(1), 1–10. https://doi.org/10.1128/mBio.02201-17

Ripamonti, C., Orenstein, A., Kutty, G., Huang, L., Schuhegger, R., Sing, A., Fantoni, G., Atzori, C., Vinton, C., Huber, C., Conville, P. S., & Kovacs, J. A. (2009). Restriction fragment length polymorphism typing demonstrates substantial diversity among *Pneumocystis jirovecii* isolates. *The Journal of Infectious Diseases*, *200*(10), 1616–1622. https://doi.org/10.1086/644643

Rooney, S. M., & Moore, P. D. (1995). Antiparallel, intramolecular triplex DNA stimulates homologous recombination in human cells. *Proceedings of the National Academy of Sciences*, *92*(6), 2141–2144. https://doi.org/10.1073/pnas.92.6.2141

Roy, D., Yu, K., & Lieber, M. R. (2008). Mechanism of R-Loop formation at immunoglobulin class switch sequences. *Molecular and Cellular Biology*, *28*(1), 50–60. https://doi.org/10.1128/MCB.01251-07

Saha, A., Nanavaty, V. P., & Li, B. (2020). Telomere and subtelomere R-loops and antigenic variation in Trypanosomes. *Journal of Molecular Biology*, *432*(15), 4167–4185. https://doi.org/10.1016/j.jmb.2019.10.025

Sassi, M., Kutty, G., Ferreyra, G. A., Bishop, L. R., Liu, Y., Qiu, J., Huang, D. W., & Kovacs, J. A. (2018). The major surface glycoprotein of *Pneumocystis murina* does not activate dendritic cells. *The Journal of Infectious Diseases*, *218*(10), 1631–1640. https://doi.org/10.1093/infdis/jiy342

Sassi, M., Ripamonti, C., Mueller, N. J., Yazaki, H., Kutty, G., Ma, L., Huber, C., Gogineni, E., Oka, S., Goto, N., Fehr, T., Gianella, S., Konrad, R., Sing, A., &

Kovacs, J. A. (2012). Outbreaks of *Pneumocystis* pneumonia in 2 renal transplant centers linked to a single strain of *Pneumocystis*: implications for transmission and virulence. *Clinical Infectious Diseases*, *54*(10), 1437–1444. https://doi.org/10.1093/cid/cis217

Schildgen, V., Mai, S., Khalfaoui, S., Lüsebrink, J., Pieper, M., Tillmann, R. L., Brockmann, M., & Schildgen, O. (2014). *Pneumocystis jirovecii* can be productively cultured in differentiated CuFi-8 airway cells. *MBio*, *5*(3). https://doi.org/10.1128/mBio.01186-14

Schmid-Siegert, E., Richard, S., Luraschi, A., Mühlethaler, K., Pagni, M., & Hauser, P. M. (2017). Mechanisms of surface antigenic variation in the human pathogenic fungus *Pneumocystis jirovecii*. *MBio*, *8*(6), 1–17. https://doi.org/10.1128/mBio.01470-17

Schmid-Siegert, E., Richard, S., Luraschi, A., Mühlethaler, K., Pagni, M., & Hauser, P. M. (2021). Expression pattern of the *Pneumocystis jirovecii* major surface glycoprotein superfamily in patients with pneumonia. *Journal of Infectious Diseases*, *223*(2), 310–318. https://doi.org/10.1093/infdis/jiaa342

Schmoldt, S., Schuhegger, R., Wendler, T., Huber, I., Söllner, H., Hogardt, M., Arbogast, H., Heesemann, J., Bader, L., & Sing, A. (2008). Molecular evidence of nosocomial *Pneumocystis jirovecii* transmission among 16 patients after kidney transplantation. *Journal of Clinical Microbiology*, *46*(3), 966–971. https://doi.org/10.1128/JCM.02016-07

Schulz, A. (2021). *pBrackets: Plot brackets*. https://cran.r-project.org/package=pBrackets

Sevier, C. S., & Kaiser, C. A. (2002). Formation and transfer of disulphide bonds in living cells. *Nature Reviews Molecular Cell Biology*, *3*(11), 836–847. https://doi.org/10.1038/nrm954

Shiota, T., Yamada, M., & Yoshida, Y. (1986). Morphology, development and behavior of *Pneumocystis carinii* observed by light-microscopy in nude mice. *Zentralblatt Für Bakteriologie, Mikrobiologie Und Hygiene. Series A: Medical Microbiology, Infectious Diseases, Virology, Parasitology*, *262*(2), 230–239. https://doi.org/10.1016/S0176-6724(86)80024-4

Soyfer, V. N., & Potaman, V. N. (1996). *Triple-helical nucleic acids*. Springer New York. https://doi.org/10.1007/978-1-4612-3972-7

Stasiak, A. (1992). Three-stranded DNA structure; is this the secret of DNA homologous recognition? *Molecular Microbiology*, *6*(22), 3267–3276. https://doi.org/10.1111/j.1365-2958.1992.tb02194.x

Stringer, J. R. (2007). Antigenic variation in *Pneumocystis*. *The Journal of Eukaryotic Microbiology*, *54*(1), 8–13. https://doi.org/10.1111/j.1550-7408.2006.00225.x

Stringer, J. R., & Keely, S. P. (2001). Genetics of surface antigen expression in *Pneumocystis carinii*. *Infection and Immunity*, *69*(2), 627–639. https://doi.org/10.1128/IAI.69.2.627-639.2001

Stringer, S. L., Hudson, K., Blase, M. A., Walzer, P. D., Cushion, M. T., & Stringer, J. R. (1989). Sequence from ribosomal RNA of *Pneumocystis carinii* compared to those of four fungi suggests an ascomycetous affinity. *The Journal of Protozoology*, *36*(1), 14S-16S. https://doi.org/10.1111/j.1550-

7408.1989.tb02670.x

Sunkin, S. M., Linke, M. J., McCormack, F. X., Walzer, P. D., & Stringer, J. R. (1998). Identification of a putative precursor to the major surface glycoprotein of *Pneumocystis carinii*. *Infection and Immunity*, *66*(2), 741–746. https://doi.org/10.1128/IAI.66.2.741-746.1998

Sunkin, S. M., & Stringer, J. R. (1996). Translocation of surface antigen genes to a unique telomeric expression site in *Pneumocystis carinii*. *Molecular Microbiology*, *19*(2), 283–295. https://doi.org/10.1046/j.1365-2958.1996.375905.x

Valade, S., Azoulay, E., Damiani, C., Derouin, F., Totet, A., & Menotti, J. (2015). *Pneumocystis jirovecii* airborne transmission between critically ill patients and health care workers. *Intensive Care Medicine*, *41*(9), 1716–1718. https://doi.org/10.1007/s00134-015-3835-9

Van der Meer, G., & Brug, S. L. (1942). *Infection à* Pneumocystis *chez l'homme et chez les animaux*.

Vanek, J., Jirovec, O., & Lukes, J. (1953). Interstitial plasma cell pneumonia in infants. *Annales Paediatrici. International Review of Pediatrics*, *180*(1), 1–21. http://www.ncbi.nlm.nih.gov/pubmed/13051050

Vargas, S. L., Hughes, W. T., Santolaya, M. E., Ulloa, A. V., Ponce, C. A., Cabrera, C. E., Cumsille, F., & Gigliotti, F. (2001). Search for primary infection by *Pneumocystis carinii* in a cohort of normal, healthy infants. *Clinical Infectious Diseases*, *32*(6), 855–861. https://doi.org/10.1086/319340

Vargas, S. L., Pizarro, P., López-Vieyra, M., Neira-Avilés, P., Bustamante, R., & Ponce, C. A. (2010). *Pneumocystis* colonization in older adults and diagnostic yield of single versus paired noninvasive respiratory sampling. *Clinical Infectious Diseases*, *50*(3), e19–e21. https://doi.org/10.1086/649869

Vargas, S. L., Ponce, C. A., Sanchez, C. A., Ulloa, A. V., Bustamante, R., & Juarez, G. (2003). Pregnancy and asymptomatic carriage of *Pneumocystis jiroveci*. *Emerging Infectious Diseases*, *9*(5), 605–606. https://doi.org/10.3201/eid0905.020660

Vera, C., & Rueda, Z. V. (2021). Transmission and colonization of *Pneumocystis jirovecii*. *Journal of Fungi*, *7*(11), 1–16. https://doi.org/10.3390/jof7110979

Vink, C., Rudenko, G., & Seifert, H. S. (2012). Microbial antigenic variation mediated by homologous DNA recombination. *FEMS Microbiology Reviews*, *36*(5), 917–948. https://doi.org/10.1111/j.1574-6976.2011.00321.x

Wakefield, A. E. (1996). DNA sequences identical to *Pneumocystis carinii* f. sp. *carinii* and *Pneumocystis carinii* f. sp. *hominis* in samples of air spora. *Journal of Clinical Microbiology*, *34*(7), 1754–1759. https://doi.org/10.1128/jcm.34.7.1754-1759.1996

Warnes, G. R., Bolker, B., Bonebakker, L., Gentleman, R., Huber, W., Liaw, A., Lumley, T., Maechler, M., Magnusson, A., Moeller, S., Schwartz, M., & Venables, B. (2020). *gplots: Various R programming tools for plotting data*. https://cran.r-project.org/package=gplots

Weinreb, A., Collier, D. A., Birshtein, B. K., & Wells, R. D. (1990). Left-handed Z-DNA and intramolecular triplex formation at the site of an unequal sister chromatid

exchange. *The Journal of Biological Chemistry*, *265*(3), 1352–1359. http://www.ncbi.nlm.nih.gov/pubmed/2104839

Weissenbacher-Lang, C., Blasi, B., Bauer, P., Binanti, D., Bittermann, K., Ergin, L., Högler, C., Högler, T., Klier, M., Matt, J., Nedorost, N., Silvestri, S., Stixenberger, D., Ma, L., Cissé, O. H., Kovacs, J. A., Desvars-Larrive, A., Posautz, A., & Weissenböck, H. (2023). Detection of *Pneumocystis* and morphological description of fungal distribution and severity of infection in thirty-six mammal species. *Journal of Fungi*, *9*(2), 220. https://doi.org/10.3390/jof9020220

Wells, R. D. (1988). Unusual DNA structures. *Journal of Biological Chemistry*, *263*(3), 1095–1098. https://doi.org/10.1016/s0021-9258(19)57268-4

Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P. C., Hall, R. J., Concepcion, G. T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N. D., Töpfer, A., Alonge, M., Mahmoud, M., Qian, Y., Chin, C. S., Phillippy, A. M., Schatz, M. C., Myers, G., DePristo, M. A., … Hunkapiller, M. W. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*, *37*(10), 1155–1162. https://doi.org/10.1038/s41587-019-0217-9

Wickham, H. (2007). Reshaping data with the {reshape} package. In *Journal of Statistical Software* (Vol. 21, Issue 12). http://www.jstatsoft.org/v21/i12/

Wickham, H. (2015). stringr: Simple, consistent wrappers for common string operations. In *R package version*. https://cran.r-project.org/package=stringr

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Wright, E. S. (2016). Using DECIPHER v2.0 to analyze big biological sequence data in R. *The R Journal*, *8*(1), 352. https://doi.org/10.32614/RJ-2016-025

Wyder, M. A., Rasch, E. M., & Kaneshiro, E. S. (1998). Quantitation of absolute *Pneumocystis carinii* nuclear DNA content. Trophic and cystic forms isolated from infected rat lungs are haploid organisms. *The Journal of Eukaryotic Microbiology*, *45*(3), 233–239. https://doi.org/10.1111/j.1550-7408.1998.tb04531.x

Xue, T., Ma, Z., Liu, F., Du, W.-Q., He, L., Ma, L., An, C.-L., & Land Geoffrey, A. (2019). Genotyping of *Pneumocystis jirovecii* by use of a new simplified nomenclature system based on the internal transcribed spacer regions and 5.8S rRNA gene of the rRNA operon. *Journal of Clinical Microbiology*, *57*(6), e02012-18. https://doi.org/10.1128/JCM.02012-18

Yamamoto, A. (2014). Gathering up meiotic telomeres: A novel function of the microtubule-organizing center. *Cellular and Molecular Life Sciences*, *71*(11), 2119–2134. https://doi.org/10.1007/s00018-013-1548-1

Yiannakis, E. P., & Boswell, T. C. (2016). Systematic review of outbreaks of *Pneumocystis jirovecii* pneumonia: evidence that *P. jirovecii* is a transmissible organism and the implications for healthcare infection control. *Journal of Hospital Infection*, *93*(1), 1–8. https://doi.org/10.1016/j.jhin.2016.01.018

Annex 1 : Supplementary data of chapter 1

1. Reproducibility of the determination of the repertoires

In order to investigate the reproducibility of the whole methodology, eight samples were analysed twice in independent experiments using identical conditions and protocol, except for the DNA extraction that was performed only once. The alleles identified and their abundance were highly similar in the duplicates of four complete repertoires with 73 to 98% of alleles in common (Table S5, Figure S3a). Moreover, all alleles identified in only one of the duplicates were low abundant (less than 1% of all reads composing the repertoire, grey and light green lines in Figure S3a). These alleles were presumably not amplified in one duplicate. This can be explained by the stochastic variation in the number of copies of the low abundant genes used as template in the PCR amplification. In other words, low abundant alleles are not amplified each time in repeated PCRs.

As far as the expressed repertoires are concerned, there were more differences between the duplicates than for the complete repertoires with only 33 to 53% of alleles in common. For two samples (BE1 and LA1), most alleles (97 to 100%) that were not present in both duplicates had an abundance lower than 1%, as observed for the complete repertoires (Table S5). On the other hand, the values for samples LA3 and LA4 were only 33 and 50%, respectively. This might result from the lower number of alleles in these two latter samples (14 to 22 versus 82 to 140 in BE1 and LA1). The reduced reproducibility for the expressed repertoires might be due to the greater variation in the allele abundances than in the complete repertoires (see Results, section “Abundance of the *msg*-I alleles in the patients”), a phenomenon that could be increased when the number of alleles present decreases.

2. Confirmation of the allele abundance within patients using amplicon subcloning

The last step of the bioinformatics pipeline provided the abundance of each allele in each sample in percentage of all reads composing the repertoire. A wet-lab approach was used to verify these abundance values. Amplicons from the expressed repertoire of patients LA7, BR3 and SE3 used for PacBio CCS were subcloned using the TOPO cloning kit (Invitrogen). These samples were selected because the low numbers of alleles composing their expressed repertoires facilitates the estimation of abundances by subcloning. Despite the low accuracy of the subcloning approach due to the small number of subclones analysed (8 to 19), the abundances obtained are consistent with those observed using Pacbio CCS followed by the bioinformatics pipeline (Table S6).

3. Search of duplicated fragments within the subtelomeres of a single strain

We searched for all duplicated fragments of size ≥ 100 bps present within the 10 representative subtelomeres assembled from a single strain (Figure S6). We used BLASTn comparisons of the 35 genes, 15 pseudogenes, and 50 intergenic spaces covering integrally these subtelomeres to the *P. jirovecii* PacBio genome assembly. The significant hits obtained for each gene corresponded to genes and pseudogenes of the same family, whereas pseudogenes generated few hits. The hits obtained for the intergenic spaces were intergenic spaces of the subtelomeres, which were upstream of genes or pseudogenes of the same family as for the query. However, families II and III constituted an exception because some of their upstream sequences produced reciprocal hits with a low score or coverage of the query, which was

consistent with the observations made in the section “Sequences flanking the *msg* genes” of the Results. Visual inspection of all alignments of the queries with their hits identified 61 and 38 duplicated fragments involving respectively 14 genes, 5 pseudogenes and 17 intergenic spaces, with a length up to 1142 bps (Table S8). We did not detect any duplicated fragments between the genes and intergenic spaces of family VI, nor between those of families II and III. Alignments of a number of mosaic genes, for example *msg*-I no. 45 and 94 (Figure S10), suggested that the regions between the shared fragments presented a sequence identity close to those observed on average between the members of each *msg* family (66 to 83%) (Schmid-Siegert et al., 2017). The presence of at least one of these duplicated fragments suggests that 22 to 100% of genes, pseudogenes, or intergenic spaces of families I to V are mosaic. These proportions are in fair agreement with those we previously observed (Schmid-Siegert et al., 2017) for genes and pseudogenes, including the absence of mosaicism in family VI (Table S8). Five out of the 15 pseudogenes investigated were also concerned by the phenomenon. These observations confirm the mosaicism of *msg* genes and pseudogenes, and reveal that the intergenic spaces are also concerned.

The 99 duplicated fragments detected are represented on the 10 representative subtelomeres that we analysed (Figure S6), as well on the 27 supplementary subtelomeres (Figure S7). Inspection of Figure S6 revealed the following features:

(i) Most subtelomeres share fragments with many other subtelomeres.

(ii) Some genes and their upstream space presented many duplicated fragments (for example, gene *msg*-II no.13 in subtelomere 26, and *msg*-III no. 53 in subtelomere 74). Sequence alignments revealed that these fragments

sometimes overlap or are identical. This is also revealed by the locations of the shared fragments within the query (Table S7c and Table S7d). The important variation of the latter locations suggest that no hotspots of recombination are not present along the *msg* genes.

(iii) The entirety of the partial *msg*-I gene no. 52 at the end of subtelomere 74 (Figure S6) is duplicated in subtelomere 95 (subtelomere/contig 95, gene no. 61, Figure S7b), suggesting a whole gene duplication.

4. Specificity of the CRJE sequence for *P. jirovecii*

Because of the peculiarity of the CRJE, we wondered if it was specific to *P. jirovecii* by using it as query in a BLASTn analysis against the whole nucleotide collection (nr/nt). The full CRJE sequence of 33 bps is present only in *P. jirovecii*, and only at the beginning of *msg*-I genes. Nevertheless, the 25 bps sequence covering the mirror repeats, *i.e.* positions 2 to 26 in Fig. 7a, was present in few copies and few other organisms (2 copies in 1 bacterial species, and 1 to 4 copies in 7 different butterfly species).

5. Absence of site-specific recombinase targeting the *P. jirovecii* CRJE sequence

To investigate if of a recombinase that recognizes specifically the CRJE sequence exists in *P. jirovecii*, we performed extensive homology searches using BLASTn involving site-specific recombinases as bait from various organisms (see Methods). These analyses did not detect any recombinases of interest.

Supplementary tables

Table S1. BAL samples from 24 immunocompromised patients analysed in this study.

City	Country	Patient code [a]	Collection year	Underlying disease
Lausanne	Switzerland	LA1	2014	HIV
		LA2	2014	HIV
		LA3	2014	unknown
		LA4	2018	unknown
		LA5	2014	unknown
		LA6	2017	unknown
		LA7	2014	HIV
		LA8	2012	HIV
		LA9	2014	HIV
Bern	Switzerland	BE1	2014	HIV
		BE2	2013	kidney transplant
		BE4	2015	cancer
		BE5	2014	cancer
Brest	France	BR1	2018	giant cell arteritis
		BR2	2018	cancer
		BR3	2019	HIV
		BR4	2020	psoriasis (methotrexate)
		BR5	2019	cancer
Cincinnati	US	CI1	2009	unknown
		CI5	1995	unknown
Seville	Spain	SE1	2007	HIV
		SE2	2010	HIV
		SE3	2013	HIV
		SE4	2013	HIV

[a] *LA, Lausanne. BE, Bern. BR, Brest. CI, Cincinnati. SE, Seville.*

Table S2. R packages used with the R version 4.1.0 (2021-05-18).

R package	Version	Description	Reference
BiocManager	1.30.16	access bioconductor package repository	Morgan (2021)
biostrings	2.62.0	manipulation of biological strings	Pagès et al. (2019)
DECIPHER	2.22.0	manage biological sequences	Wright (2016)
dendextend	1.15.2	dendogram manipulation	Galili (2015)
ggplot2	3.3.5	figures and plots	Wickham (2016)
gplots	3.1.1	create heatmaps	Warnes et al. (2020)
gridExtra	2.3	arrange plots	Auguie (2017)
pBrackets	1.0.1	bracket elements in plot	Schulz (2021)
plotrix	3.8-2	plot options	Lemon (2006)
reshape2	1.4.4	reshape data	Wickham (2007)
seqinr	4.2-8	manipulation of sequences	Charif & Lobry (2007)
stringr	1.4.0	string operations	Wickham (2015)

Table S3. Primers used for the control PCRs specific to given alleles.

Allele	Name allele PacBio	Present in patients [a]	Primer	Position (nt)	Sequence (5' - 3')
A	CI3c106432837	LA2, LA8, BE1, CI3	A-for	198-224	CCAAGTTGTAAAGGTAGTAAATGTAGC
			A-rev	1909-1929	TAAACGACTTCGTGTCTTTTG
B	LA2cd106037614	LA2, CI3	B-for	52-70	GTCCACCTCTAGTGCAAGC
			B-rev	1627-1651	TTCTTGACATTTCCATTTTTATTAG
C	LA7c108528291	LA2, BE1, CI3	C-for	57-80	GAAAAGACATGTGAGAACCTTATG
			C-rev	1791-1812	GCTTTTCTAGTTCTTTTCTTGC

[a] BE, Bern. CI, Cincinnati. LA, Lausanne.

Table S4. Characteristics of the expressed and complete *msg*-I gene repertoires observed in the 24 patients.

Patient [a]	Number of alleles in		Number of alleles in common between the two repertoires	% expressed in complete (alleles in common / total number in expressed)	% complete in expressed (alleles in common / total number in complete)	Number of *P. jirovecii* strains
	expressed repertoire	complete repertoire				
LA1	82	148	72	88	49	3
LA2	54	49	44	81	90	1
LA3	22	79	20	91	25	1
LA4	18	130	14	78	11	3
LA5	21	144	19	90	13	2
LA6	91	116	82	90	71	5
LA7	5	59	5	100	8	1
LA8	59	105	51	86	49	1
LA9	7	54	7	100	13	1
BE1	90	185	59	66	32	4
BE2	63	56	56	89	100	1
BE4	82	113	79	96	70	3
BE5	18	118	9	50	8	5
BR1	3	83	3	100	4	3
BR2	2	96	2	100	2	3
BR3	5	128	2	40	2	3
BR4	9	96	5	56	5	2
BR5	18	170	17	94	10	4
CI1	108	140	100	93	71	4
CI5	45	44	32	71	73	1
SE1	41	61	37	90	61	1
SE2	12	115	10	83	9	1
SE3	8	77	8	100	10	3
SE4	17	139	17	100	12	3

[a] LA, Lausanne. BE, Bern. BR, Brest. CI, Cincinnati. SE, Seville.

Table S5. Duplicated analyses of eight samples.

Repertoire	Patient [a]	Number alleles			Common alleles in duplicates		% of different alleles with abundance <1%
		duplicate 1	duplicate 2	Total distinct	number	% (alleles in common / total number distinct alleles)	
complete	LA2	49	48	49	48	98	100
	LA1	149	134	154	129	84	100
	BE1	185	211	212	184	87	100
	LA3	79	80	92	67	73	100
expressed	BE1	90	121	138	73	53	100
	LA1	82	140	148	74	50	97
	LA3	22	19	28	13	46	33
	LA4	18	14	24	8	33	50

[a] LA, Lausanne. BE, Bern.

Table S6. Abundance of alleles determined using PacBio CCS and subcloning.

			Subcloning	
Patient [a]	**Allele name** [b]	**Abundance PacBio (%)**	**Abundance (%)**	**Nb of clones**
BR3	BR3u100403135	33	74	14
	BR3u100532918	32	0	0
	BR3u103089526	24	21	4
	BR3u100992345	10	5	1
	BR3u119802438	2	0	0
LA7	LA7u47056848	72	60	6
	LA7u100272069	9	0	0
	LA7u10029856	8	10	1
	LA7u101910862	7	20	2
	LA7u100074959	4	10	1
SE3	SE3u100009239	35	38	3
	SE3u100271002	14	13	1
	BR1u102369376 [b]	13	13	1
	SE3u100008413	11	13	1
	CI1u100600372 [b]	9	0	0
	SE3u117048602	8	25	2
	SE3u105186188	7	0	0
	SE3u110953735	4	0	0

[a] LA, Lausanne. BR, Brest. CI, Cincinnati. SE, Seville.

[b] The name of these alleles results from their high abundance in samples BR1 or CI1.

Table S7. See separate excel file.

Table S8. Duplicated fragments ≥ 100 bps within the 10 *P. jirovecii* representative subtelomeres from a single strain a.

	msg family	No. of genes or intergenic space upstream of genes (of which pseudogenes) used as query	No. of genes or intergenic space upstream of genes (of which pseudogenes) with duplicated fragments ≥ 100 bps	Total no. of duplicated fragments ≥ 100 bps detected [b] (of which between genes and pseudogenes)	Mean size (bps) of duplicated fragments (range)	% potential mosaic genes or intergenic space (pseudogenes included)	% potential mosaic genes in reference (Schmid-Siegert et al., 2017)[c]
Genes and pseudogenes	I	20 (8)	5 (2)	7 (2)	513 (109-670+1142) [d]	25	42
	II	9 (3)	5 (0)	32 (0)	249 (117-710)	56	28
	III	5 (0)	3 (0)	12 (0)	120 (100-153)	60	40
	IV	2 (2)	2 (2)	4 (4)	165 (113-221)	100	22
	V	7 (1)	4 (1)	7 (1)	183 (102-363)	57	7
	VI	4 (0)	0 (0)	0 (0)	0	0	0
	outlier	3 (1)	0 (0)	0 (0)	0	0	-
	Total	**50** (15)	**19** (5)	**61** (7)			
Intergenic spaces	I	20 (8)	9 (1)	12 (1)	148 (105-224)	45	-
	II	9 (3)	2 (0)	6 (0)	314 (131-787)	22	-
	III	5 (0)	2 (0)	6 (0)	233 (132-320)	40	-
	IV	2 (2)	1 (1)	2 (2)	129 (115, 142)	50	-
	V	7 (1)	3 (1)	12 (3)	178 (115-334)	43	-
	VI	4 (0)	0 (0)	0 (0)	0	0	-
	outlier	3 (1)	0 (0)	0 (0)	0	0	-
	Total	**50** (15)	**17** (3)	**38** (6)			

Footnotes of Table S8 :

[a] All *msg* genes, pseudogenes, and intergenic spaces composing the 10 representative subtelomeres shown in Figure S6 were used as query in BLASTn analyses (search of somewhat similar sequences) against the PacBio *P. jirovecii* genome assembly (Schmid-Siegert et al., 2017). Each upstream intergenic space extended up to the end of the gene located upstream (all genes are oriented towards the telomere within the subtelomeres, see Figure S6). Visual inspection of all alignments of the query with the significant hits identified the duplicated fragments. Outlier genes are those that could not be attributed to one of the six *msg* families (Schmid-Siegert et al., 2017).

[b] The duplicated fragments detected twice because of reciprocal BLASTn analyses were counted only once.

[c] Using various numerical bioinformatics tools to detect recombination events between *msg* genes and pseudogenes (Schmid-Siegert et al., 2017).

[d] The duplicated fragment of 1142 bps corresponds to the entirety of the partial gene *msg*-I no. 52 at the end of contig 74 (Figure S6).

Supplementary figures

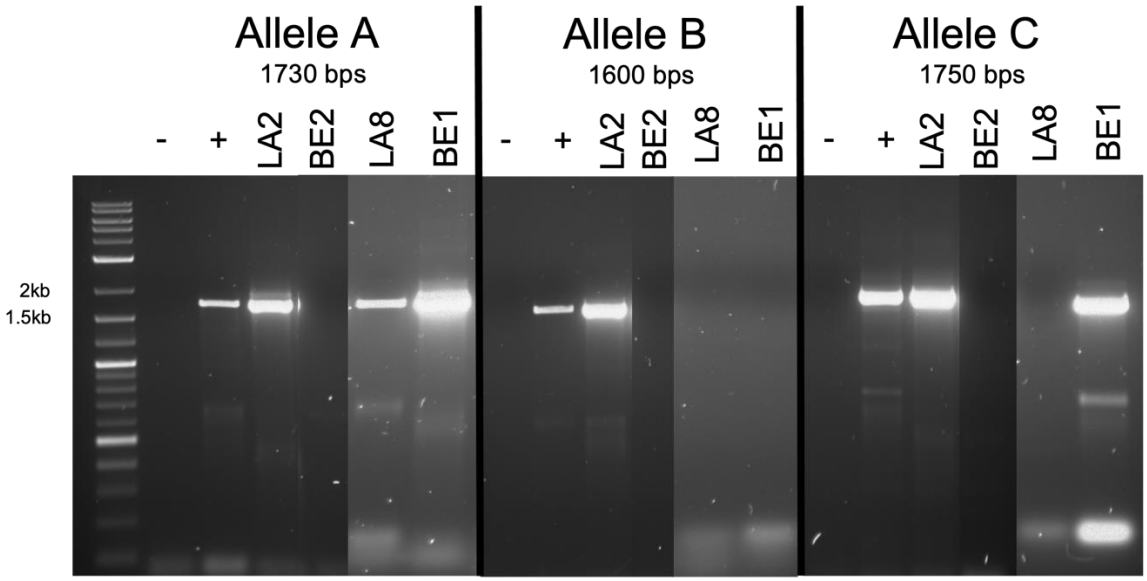


Figure S1. Analysis in agarose gel of the PCRs using primers specific to fragments of *msg*-I alleles A, B and C. The size of the PCR product from each allele is indicated. The positive control (+) is randomly amplified DNA of patient LA2 that contains all three alleles. The primers used and alleles are described in Table S3.

GAAAATTCAGCTTAAACACTTCCCTAGTGTTTTAGCATTTTTCAAACATCTGTGAA

10xT 6xT

TTTTTTTTTTGTTTGGCGAGGAGCTGGCTTTTTTGCTTGCCTCGCCAAAGGTGTT

TATTTTTAAAATTTTAAATTGAATTTCAGTTTTAGAATTTTTTAAAAACTTTCAACAA

TGGATCTCTTGGCTCTCGCGTCGATGAAGAACGTGGCAAAATGCGATAAGTAGT

GTGAATTGCAGAATTTAGTGAATCATCGAATTTTTGAACGCATCTTGCGCTCCTT

5xT

AGTATTCTAGGGAGCATGCCTGTTTGAGCGTTATTTTTAAGTTCCTTTTTTCAAG

5xA 3xA

CAGAAAAAAGGGGATTGGGCTTTGCAAATATAATTAGAATAAAATAATTATATGC

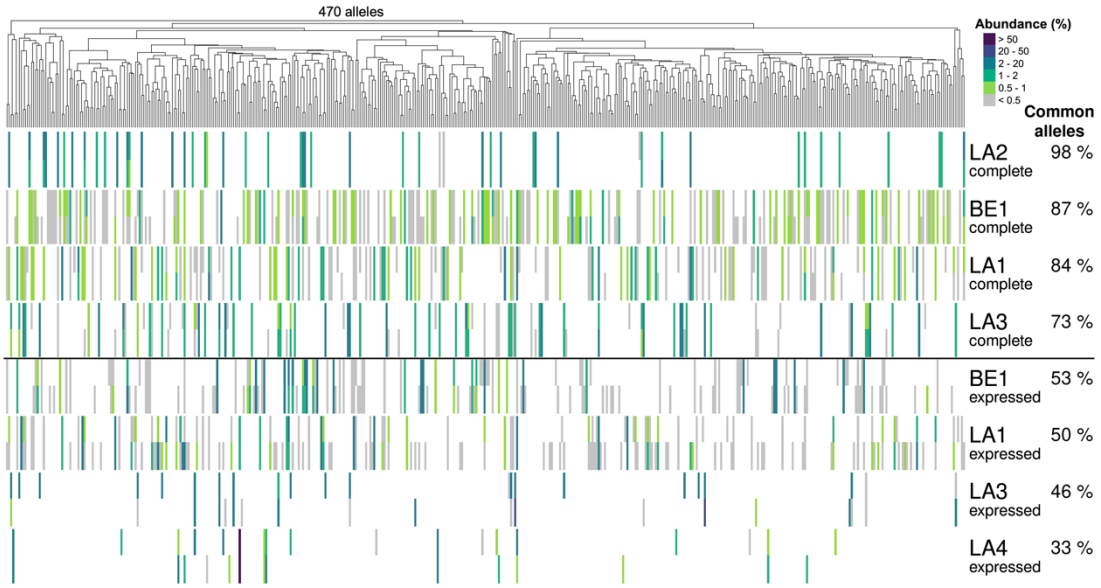
ATGCTAGTCTGAAATTCAAAAGTAGCTTTTTTTCTTTGCCTAGTGTCGTAAAAATT

4xT

CGCTGGGAAAGAAGGAAAAAAGCTTTTATAAATACAAGAATT

Figure S2. *P. jirovecii* ITS1-5.8S-ITS2 sequence (JQ365709.1). Highlighted in yellow are the six homopolymers that were homogenized in the present study, i.e. they were replaced in all sequences by the number of nucleotides indicated in red.

a



b

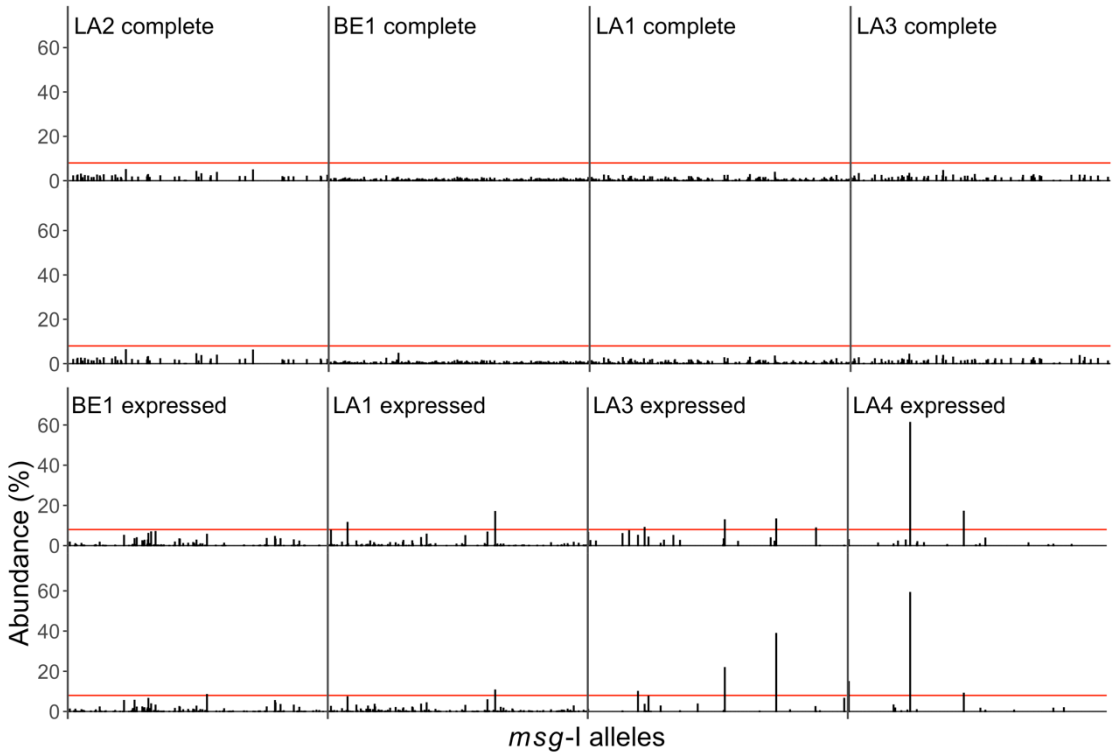


Figure S3. Duplicate analyses to evaluate the reproducibility of the whole methodology. LA, Lausanne. BE, Bern.

a. Composition of the complete and expressed *msg*-I repertoires observed. Each vertical line of the heatmap represents an allele present in the given repertoire with the color figuring its abundance in % of all reads composing the repertoire, as indicated at the top right of the figure. The 470 alleles observed were sorted using hierarchical classification trees of the multiple alignments of the allele sequences (Fitch distance, average linkage).The percentage of common alleles between the duplicates are indicated next to the patient's name.

b. Abundance of the alleles within the complete and expressed repertoires (see Results, section "Abundance of the *msg*-I alleles in the patients"). The alleles are sorted using the same tree as in panel a. The red lines indicate an abundance of 8.0%.

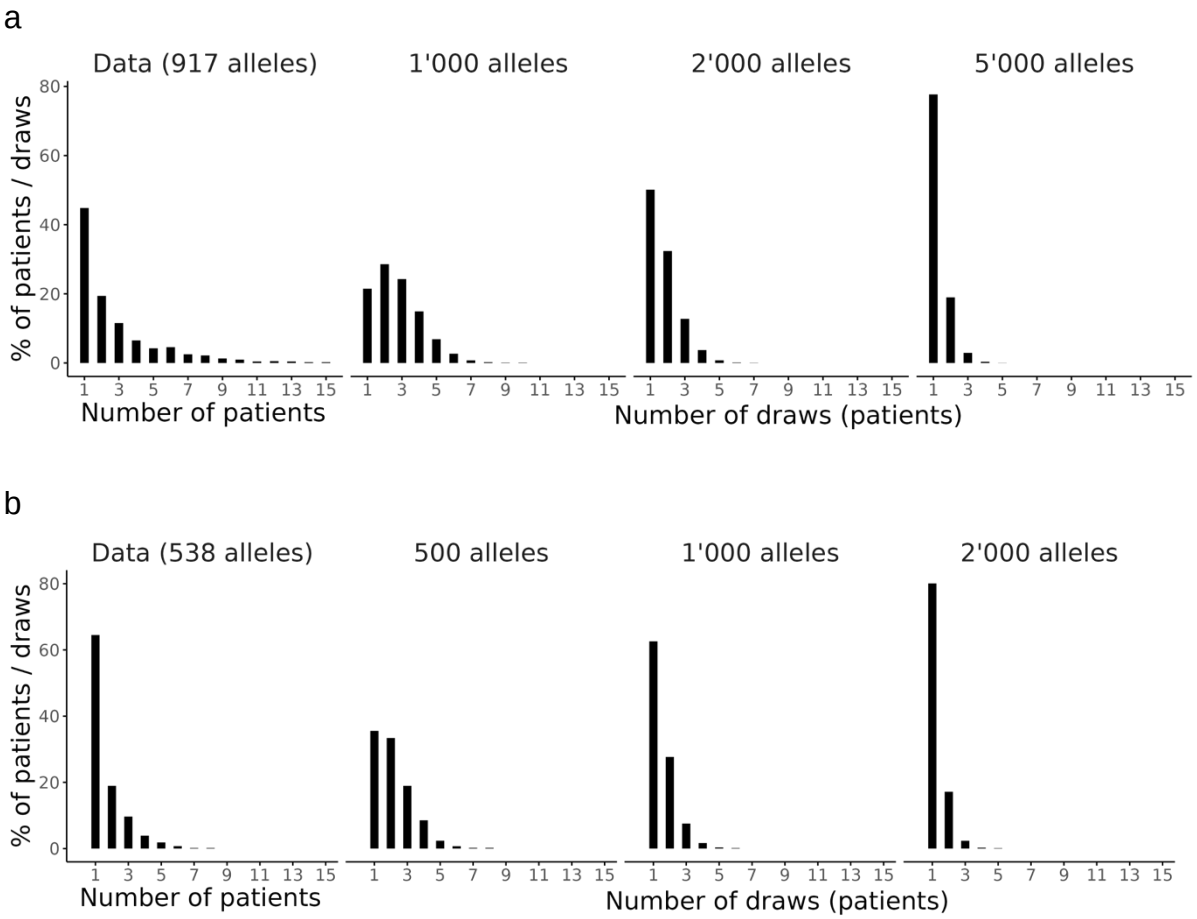


Figure S4. Comparison of the distribution of the alleles observed in the complete (a) and expressed (b) repertoires of the 24 patients with those obtained by simulating reservoirs of alleles of increasing size (see text). The software R was used.

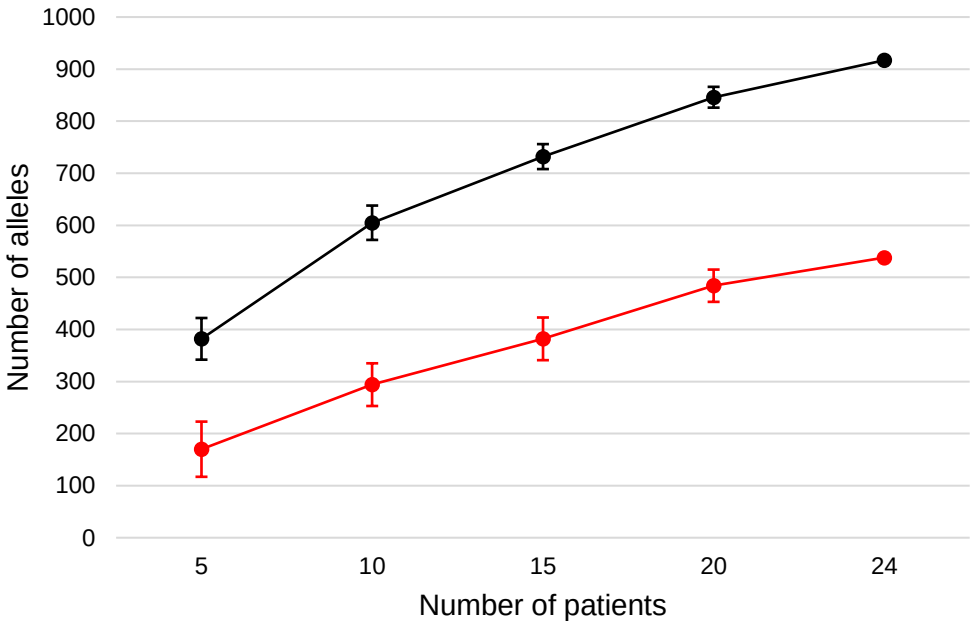


Figure S5. Number of alleles observed in the complete (black) and expressed (red) repertoires in function of the simulated number of patients analysed (see text). SD are shown. The software R was used.

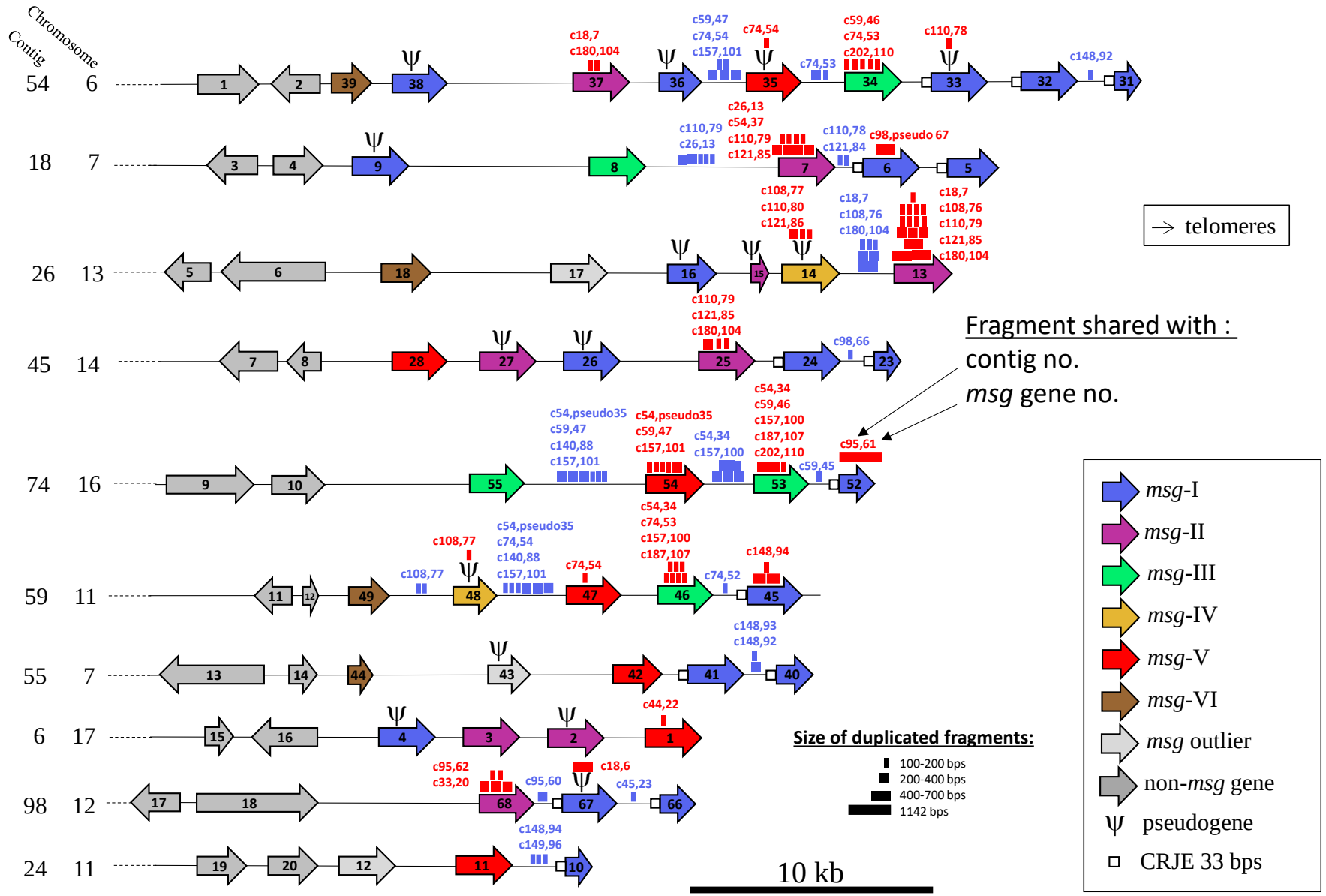

Chromosome
Contig
54 6
18 7
26 13
45 14
74 16
59 11
55 7
6 17
98 12
24 11
→ telomeres
Fragment shared with :
contig no.
msg gene no.
msg-I
msg-II
msg-III
msg-IV
msg-V
msg-VI
msg outlier
non-msg gene
ψ pseudogene
CRJE 33 bps
Size of duplicated fragments:
100-200 bps
200-400 bps
400-700 bps
1142 bps
10 kb

Figure S6. Ten representative subtelomeres (i.e. contigs, see Methods) which genes, pseudogenes, and intergenic spaces were used as queries against the whole genome assembled from a single *P. jirovecii* strain using PacBio sequencing (Schmid-Siegert et al., 2017). The symbols represent the size of the duplicated fragments identified within genes and pseudogenes (red symbols), or intergenic spaces (blue symbols). The position of the symbols is not precise. The positions of the duplicated fragments within the query and the subtelomeres are given in Table S7c and Table **S7**d. Adapted from Fig. 3 of reference (Schmid-Siegert et al., 2017).

a

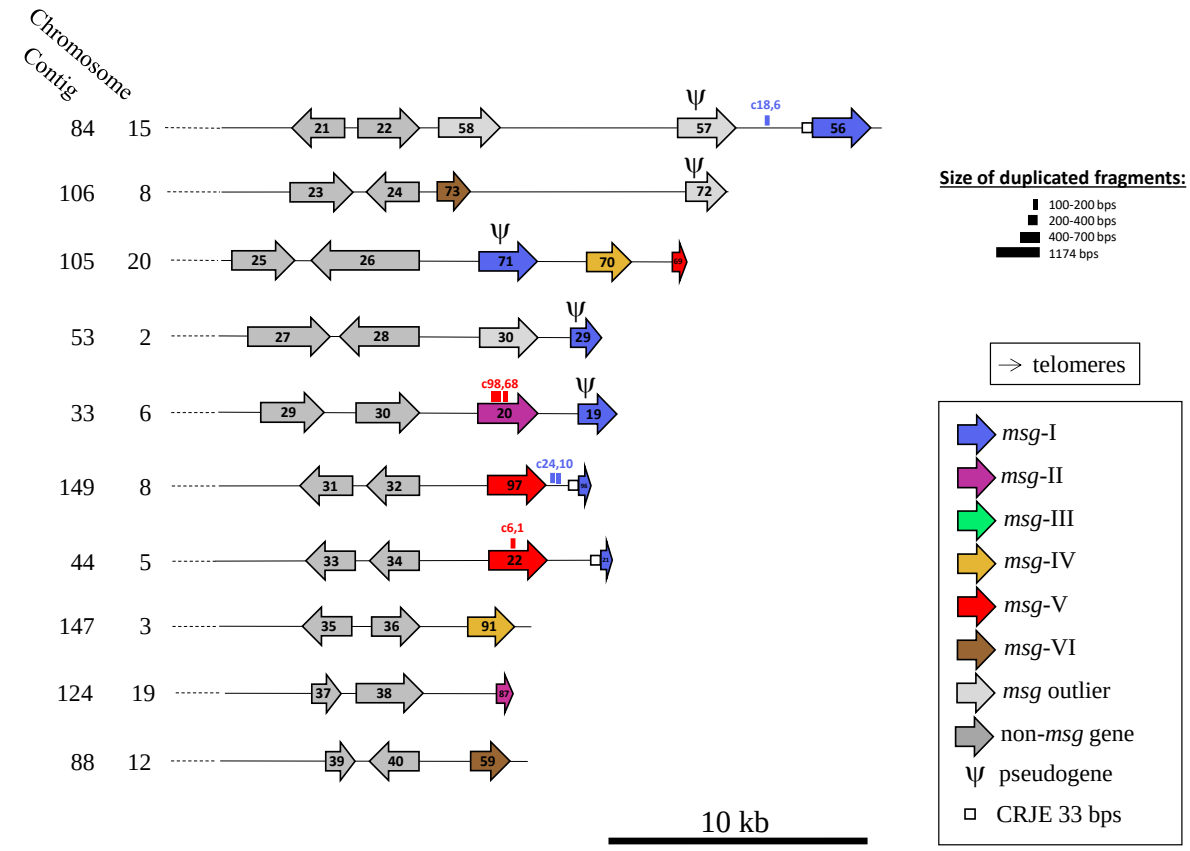

Chromosome
Contig
84 15
106 8
105 20
53 2
33 6
149 8
44 5
147 3
124 19
88 12
21 22 58 57 56
23 24 73 72
25 26 71 70
27 28 30 29
29 30 20 19
31 32 97
33 34 22
35 36 91
37 38
39 40 59
c18,6
c98,68
c24,10
c6,1
10 kb
Size of duplicated fragments:
100-200 bps
200-400 bps
400-700 bps
1174 bps
telomeres
msg-I
msg-II
msg-III
msg-IV
msg-V
msg-VI
msg outlier
non-msg gene
pseudogene
CRJE 33 bps

Figure S7.

b

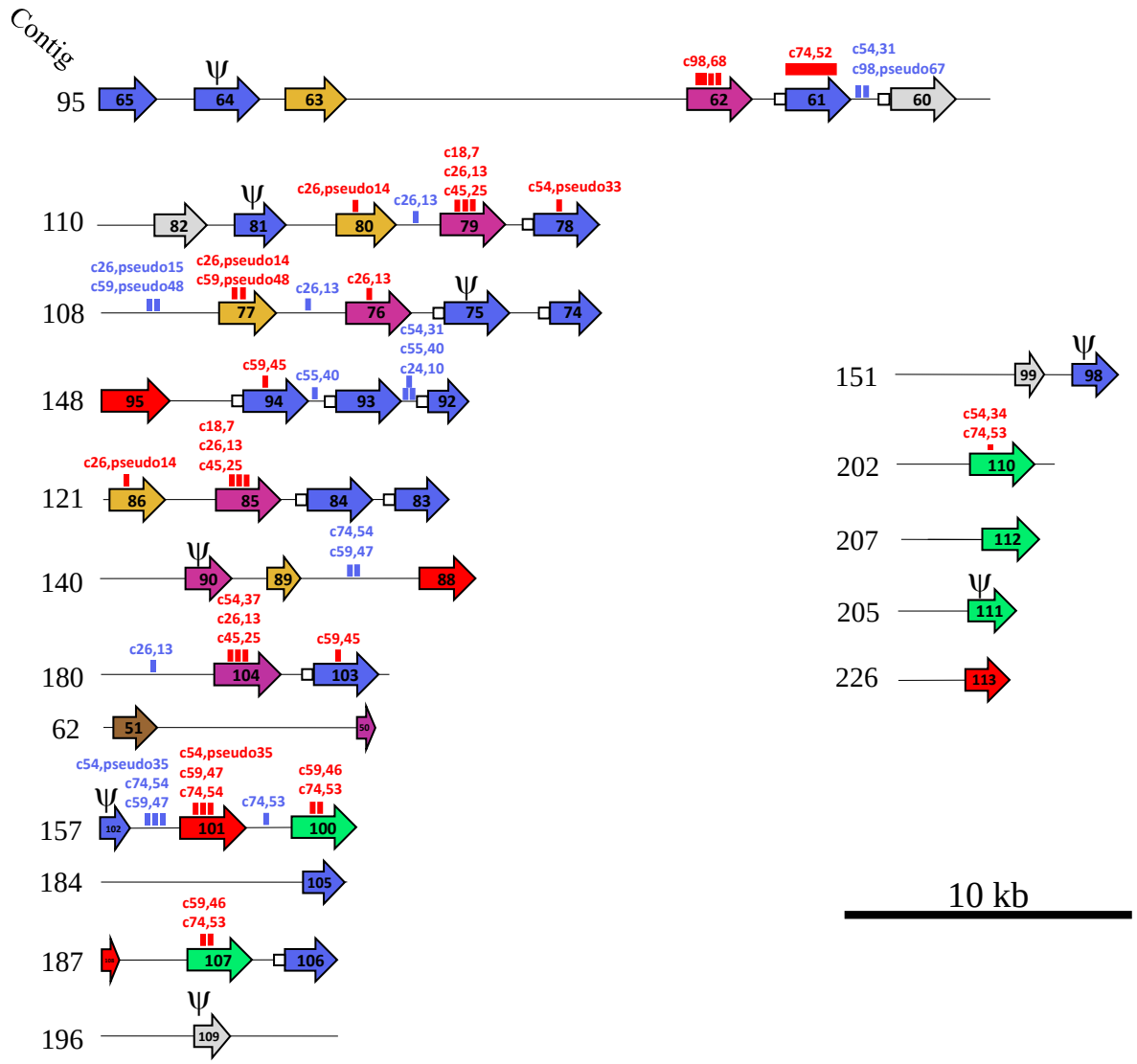


Figure S7. Supplementary subtelomeres (i.e. contigs, see methods) from the same single strain as the 10 representative ones shown Figure S6. The symbols represent the size of the duplicated fragments identified within genes and pseudogenes (red symbols), or intergenic spaces (blue symbols). Their positions within the query and the subtelomeres are given in Table S7c and Table **S7**d. Adapted from Figure S6 of reference (Schmid-Siegert et al., 2017).

a. 10 subtelomeres harboring genomic genes (in gray) that allowed their attribution to a specific chromosome from reference (Ma et al., 2016).

b. 17 subtelomeres not harboring genomic genes, preventing their attribution to a chromosome.

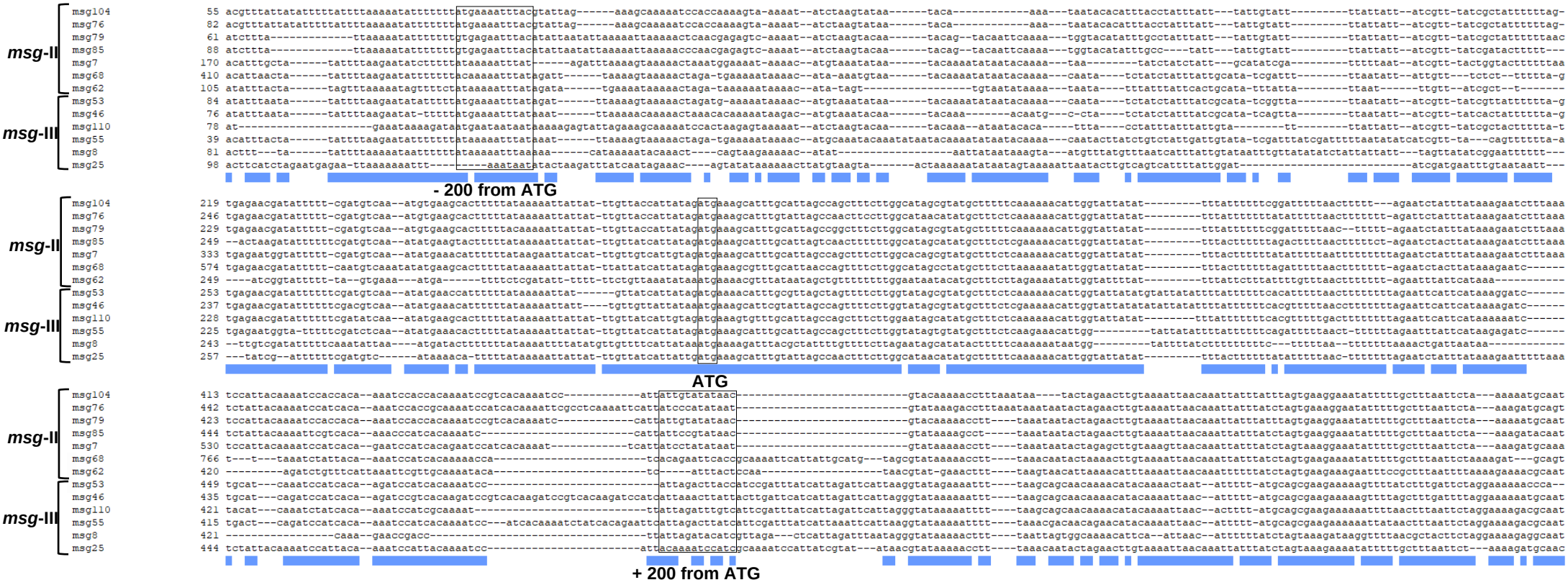


Figure S8. Alignments of the region surrounding the ATG of seven *msg*-II and six *msg*-III genes presenting significant similarity in the 200 bps upstream of their CDS. The blue bars underneath indicate areas of significant similarity. Such areas are particularly long around the start codon ATG of the CDS. The ATG start codons and the regions encompassing the – and + 200 bps positions relatively to the ATG of sequences are boxed.

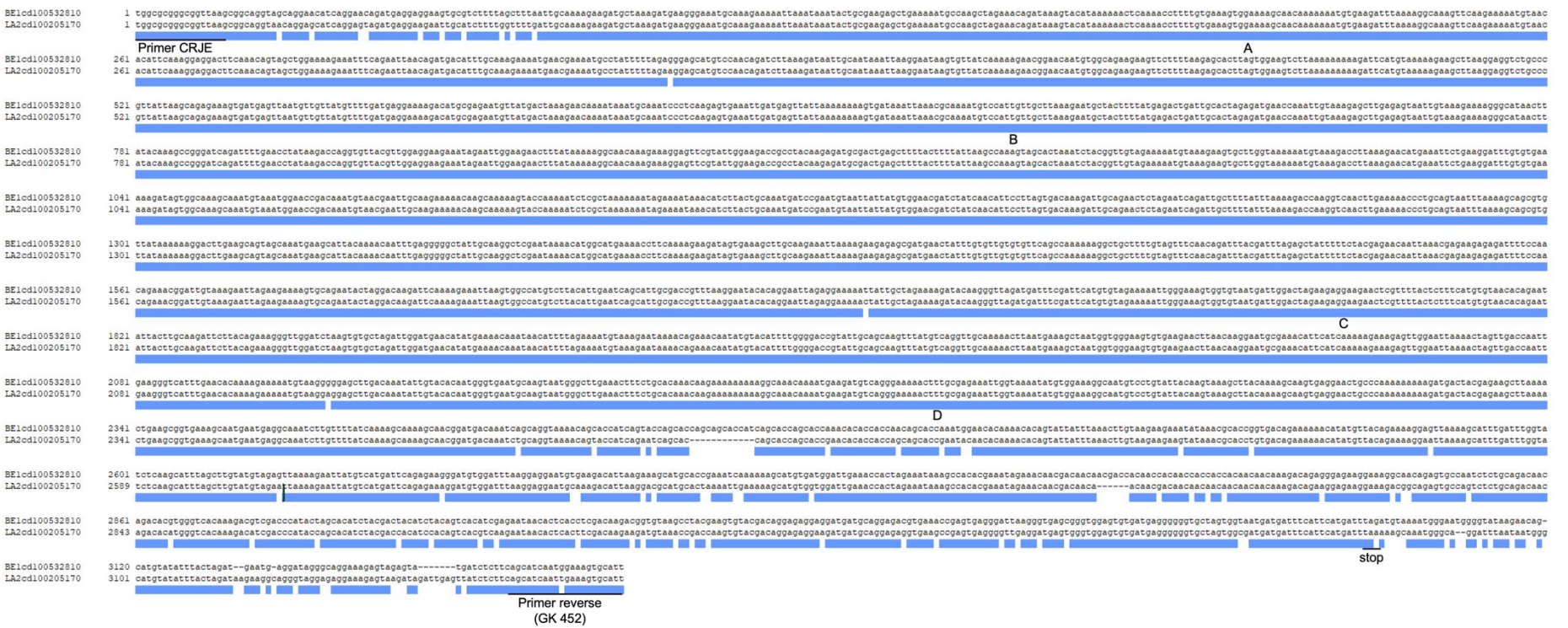


Figure S9. Alignment of two mosaic *msg*-I genes identified in this study (the names of the alleles are given on the left). The blue bars indicate areas of full identity. The four fragments of > 100 bps shared are indicated (A to D, respectively of 284, 1335, 420, and 294 bps). Fragment C is shown in red within Figure 6. Clone Manager 9 Professional Edition software version 9.51 (Sci Ed Software LLC) and the alignment format “similarity summary bars” were used.

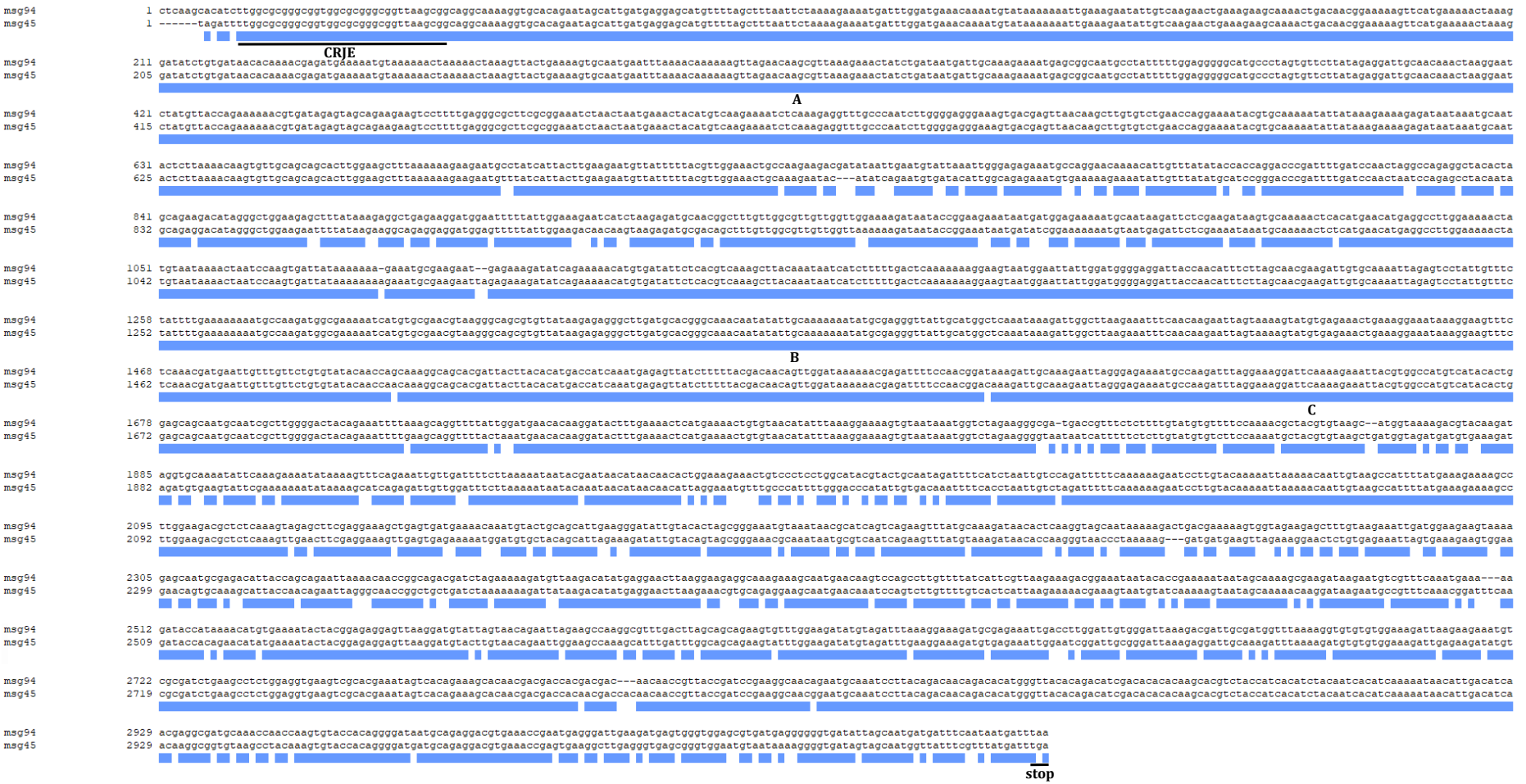


Figure S10. Alignment of the mosaic *msg*-I genes no. 45 and 94. The blue bars indicate areas of full identity. The three fragments of > 100 bps shared are numbered (respectively 670, 427, and 118 bps). Clone Manager 9 Professional Edition software version 9.51 (Sci Ed Software LLC) and the alignment format “similarity summary bars” were used.

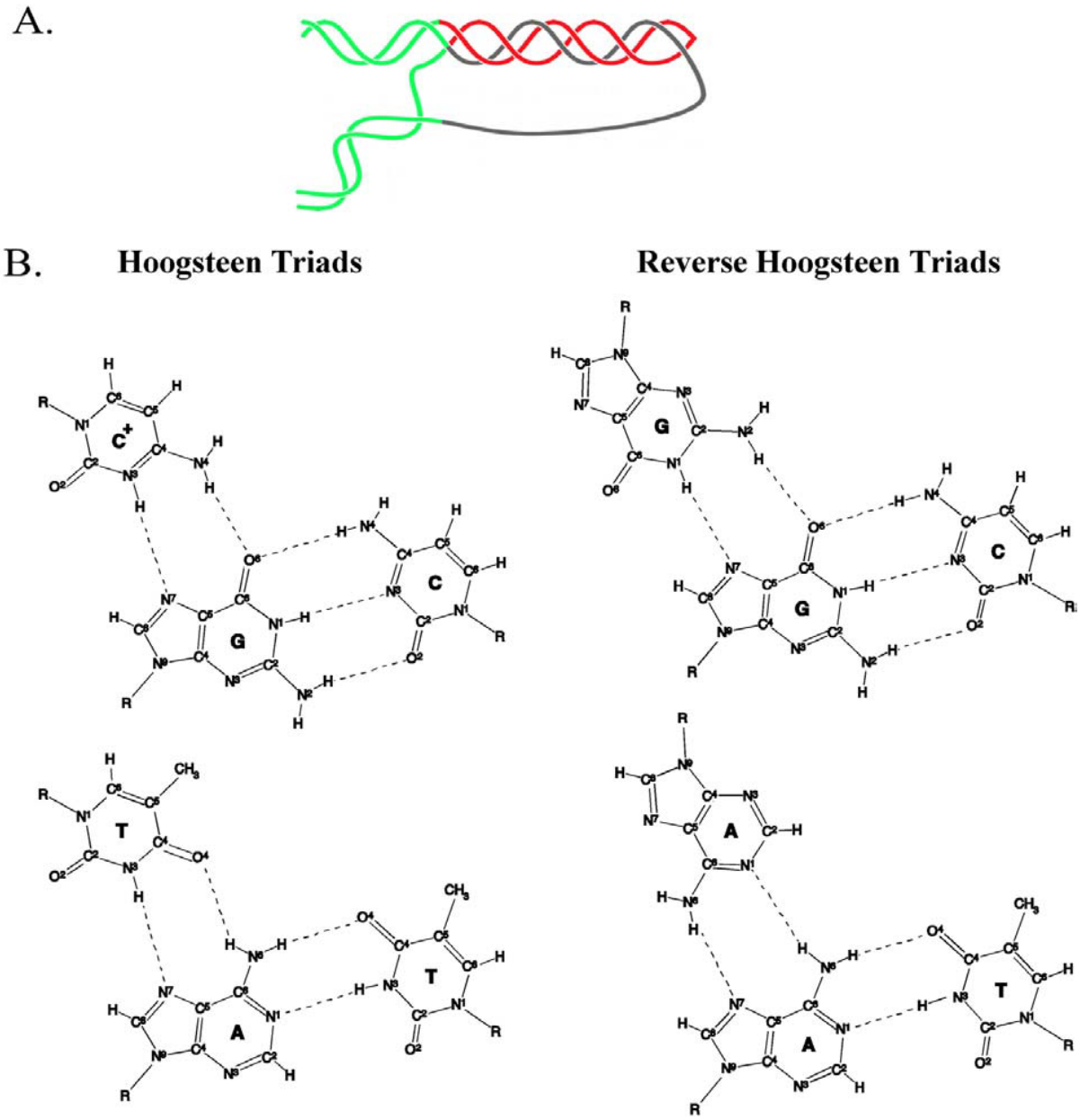


Figure S11.

C.

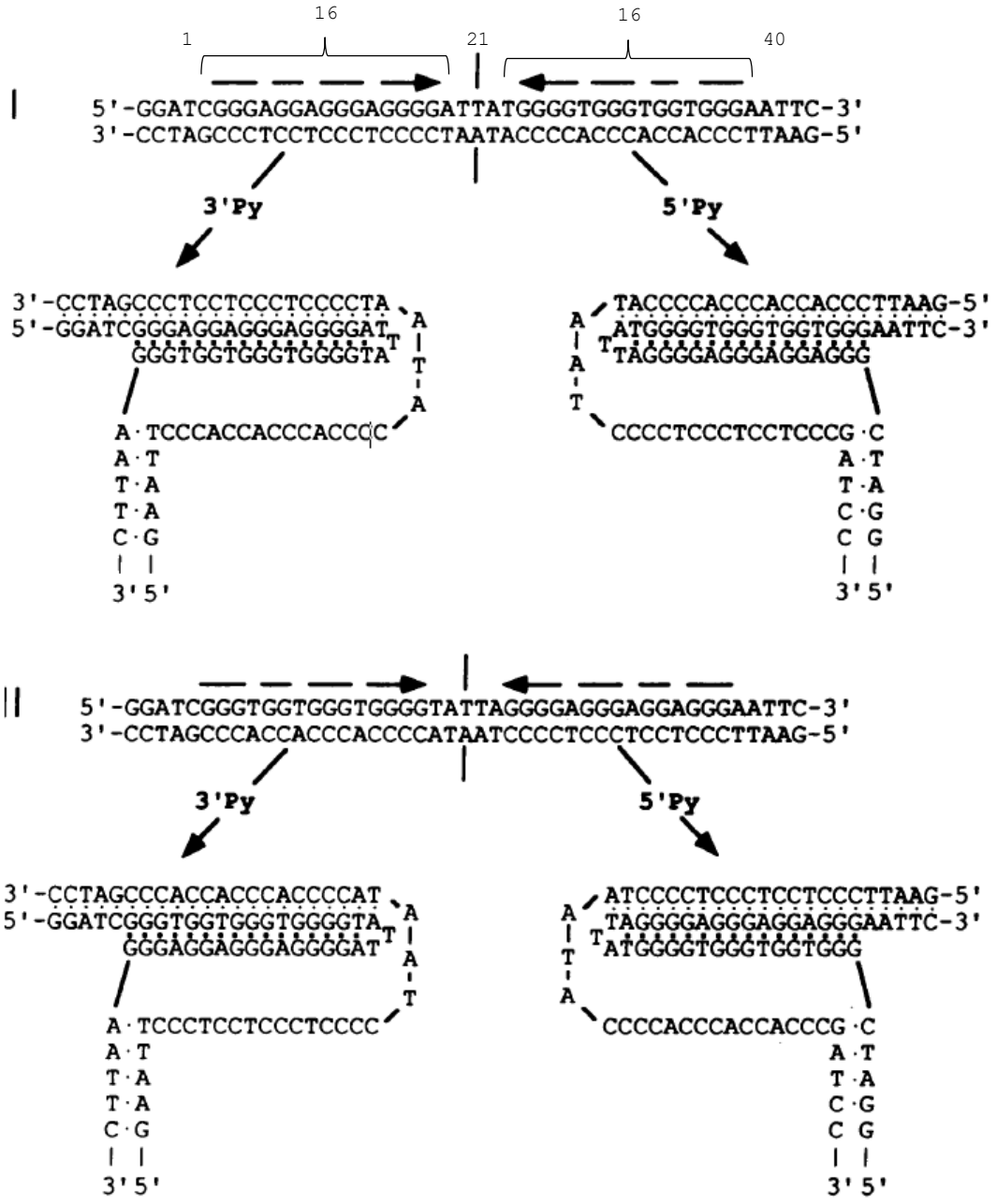


Figure S11. Features of H-DNA and *H-DNA triplexes. A. and B.: Figure 3 and its legend from reference (Mirkin, 2008). C.: Figure 1 and its legends from reference (Dayn et al., 1992), with added numbering. All three panels are reproduced with permission from corresponding author and copyright holder S. M. Mirkin.

A. The structure of an intramolecular triplex. The two complementary strands of a homopurine-homopyrimidine repeat are colored in red and gray, while flanking DNA is colored green. The structure is called H-y when the red strand is homopyrimide, and H-r if when it is homopurine. One can see that the red and gray strands in this structure are not linked, i.e. formation of H-DNA is topologically equivalent to an unwinding of the entire homopurine-homopyrimidine repeat.

B. H-y form is built from TA*T and CG*C+ triads, in which pyrimidines in the third strand form Hoogsteen hydrogen bonds with the purines of the duplex. H-r form and is built of CG*G and TA*A triads, where purines from the third strand form reverse Hoogsteen hydrogen bonds with the purines in the duplex.

C. Intramolecular triplexes consisting of GG*C and TA*T triads. In both sequences I and II, GC base pairs are arranged as mirror images (shown by arrows; the pseudosymmetry axis is shown by a vertical line), whereas AT base pairs are arranged as inverted repeats. Points, Watson-Crick hydrogen bonds; squares, Hoogsteen hydrogen bond

Annex 2 : Supplementary data of chapter 2

Multiple alignment of ITS1-5.8S-ITS2 sequences

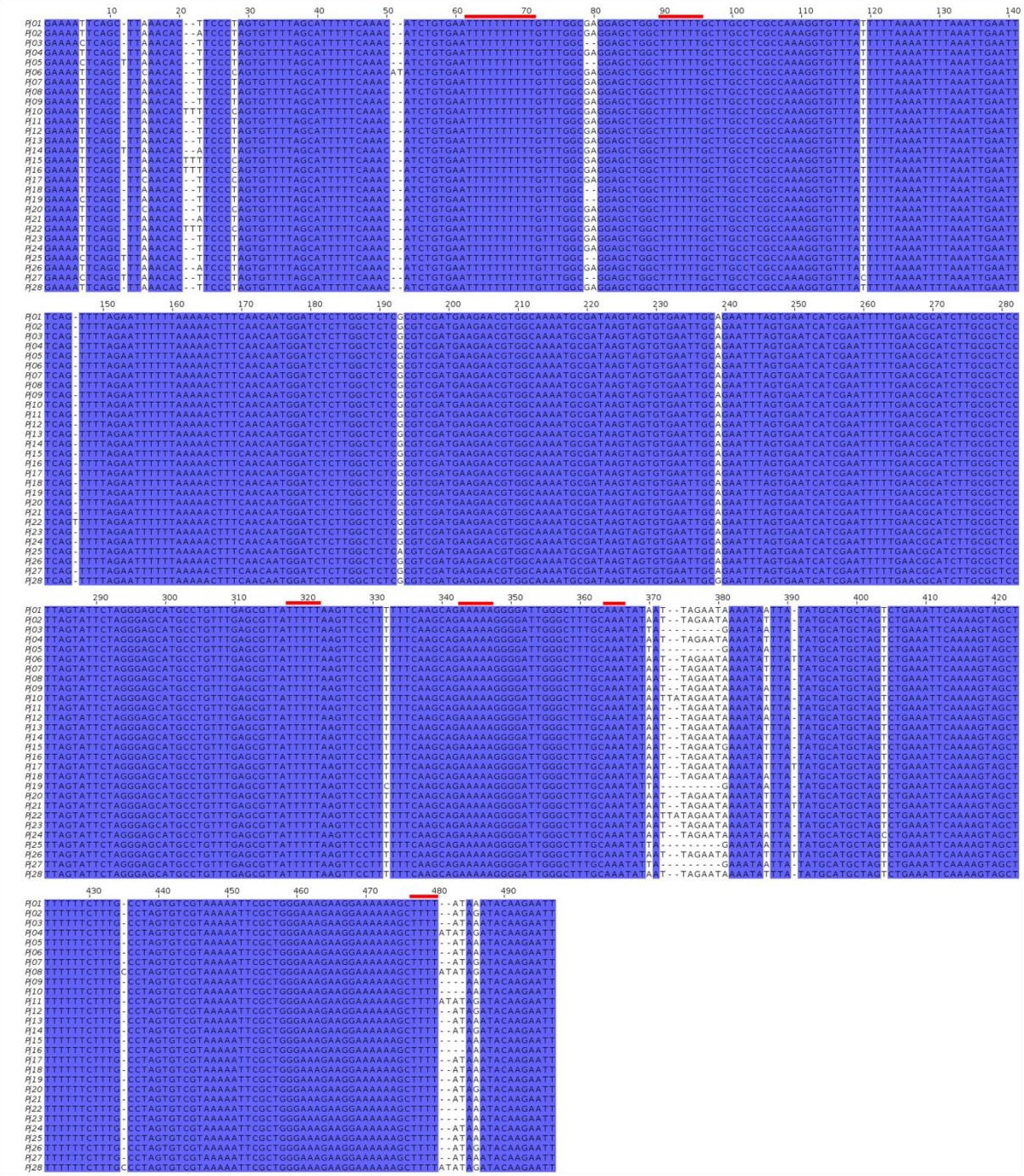


Figure S12. Multiple alignment of all the identified ITS1-5.8S-ITS2 sequences (Pj01-Pj28). Blue background indicates a 100% identity between all sequences and the red lines show the six homopolymer stretches that were ignored.