Revised version: clean.

Expression pattern of the *Pneumocystis jirovecii* major surface glycoprotein superfamily in patients with pneumonia

Emanuel Schmid-Siegert [1,a], Sophie Richard [2], Amanda Luraschi [2,b], Konrad Mühlethaler [3], Marco Pagni [1]*, Philippe M. Hauser [2]*

* These authors contributed equally to the work.

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

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

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

[a] Present address: Selexis SA, Chemin des Aulx 14, 1228 Plan-les-Ouates, Switzerland

[b] Present address: Resistell AG, Muttenz, Switzerland

Correspondence : Philippe.Hauser@chuv.ch

Running title: *Pneumocystis* surface proteins expression (40 characters)

Abstract : 154 words

Text : 3499 words

Brief summary (40 words)

The six families of surface glycoproteins of the human pathogen *Pneumocystis jirovecii* were studied by RNA sequencing. A number of different isoforms of most families were expressed simultaneously, suggesting that individual cells with distinct surface of properties compose each population.

Competing interests

The authors declare that no competing interests exist.

Funding information

This work was supported by the Swiss National Science Foundation, grant 310030_165825. This Foundation had no role in any steps of the study.

Abstract

Background. The human pathogen *Pneumocystis jirovecii* harbors six families of major surface glycoproteins (MSG) encoded by a single gene superfamily. MSGs are presumably responsible for antigenic variation and adhesion to host cells. The genomic organization suggests that a single member of family I is expressed at a time per cell, whereas members of the other families are simultaneously expressed.

Methods. We analyzed RNA sequences expressed in several clinical samples, using specific weighted profiles for reads sorting and single nucleotide variants calling to estimate the diversity of the expressed genes.

Results. A number of different isoforms of at least four MSG families were expressed simultaneously, including of family I for which confirmation was obtained in the wet laboratory.

Conclusion. These observations suggest that every single *P. jirovecii* population is made of individual cells with distinct surface properties. Our results enhance our understanding of the unique antigenic variation system and cell surface structure of *P. jirovecii*.

Keywords: surface antigenic variation, RNAseq, virulence factor, colonization factor

Introduction

The fungal genus *Pneumocystis* belongs to the subphylum Taphrinomycotina of the Ascomycota [1]. It encompasses extracellular parasites that colonize the lungs of mammals [2]. The species infecting humans is *Pneumocystis jirovecii*, while rats and mice harbor respectively *Pneumocystis carinii* and *Pneumocystis murina*. Should the human immune system decline, *P. jirovecii* can become an opportunistic pathogen that causes severe pneumonia, which can be lethal if not treated (*Pneumocystis* pneumonia, PCP). An *in vitro* method of long-term culture for *Pneumocystis* species is still not available.

The lack of a culture method complicates the study of *P. jirovecii* pathogenicity. Its major surface glycoproteins (MSG) constitute potentially a crucial factor involved in colonization and/or virulence. These proteins are thought to generate surface antigenic variation allowing escape from the human immune system [3-5]. Besides, the MSGs are thought to be involved in adhesion to host cells [6, 7]. MSGs are encoded by six families of hypervariable genes located at all subtelomeric regions of the ca. 20 chromosomes of *P. jirovecii* [8-10]. These families form the largest surface protein superfamily known among fungi with approximately 160 *msg* genes per genome (families I to VI include respectively ca. 85, 20, 10, 20, 20, and 5 genes encoding isoforms; family IV is the only one potentially not anchored in the cell surface because of the lack of a glycosylphosphatidylinositol signal). MSGs of family I (MSG-I) are thought to be the most abundant at the *P. jirovecii* cell surface [11], although only one *msg*-I gene out of ca. 80 present in the genome is probably expressed at a time in every single cell. Indeed, firstly, the latter mutually exclusive expression would rely on the expression of a single gene under the control of a transcription promoter that is present at a single copy per genome (within a upstream conserved sequence, UCS), whereas the other genes of family I have no promoter. Secondly, *Pneumocystis* cells are mostly haploid except transiently during the sexual cycle [12-15]. However, at the population level, several different *msg*-I isoforms linked to the

UCS at the DNA level are observed, and thus are presumably expressed [9, 16]. In *P. carinii*, the diversity of *msg*-I isoforms expressed has been observed also at the RNA level [17], as well at the protein level that revealed a focal distribution of the epitopes within the lung [18]. The exchange of the expressed *msg*-I isoform would occur upon recombination at a sequence of 33 bps that is present both at the end of the UCS including the promoter and at the beginning of each *msg*-I gene (the conserved recombination junction element, CRJE). By contrast, each gene of the other five MSG families II to VI possesses its own promoter allowing potentially independent and simultaneous expression [9]. In addition to the exchange of the *msg*-I isoform expressed, antigenic variation is thought to rely on recombinations between the genes of each MSG family that generate gene mosaicism [3, 9, 19].

We propose a model for the surface antigenic variation system of *P. jirovecii* consisting in the continuous segregation of new subpopulations that are antigenically different (Figure 1). The aim of the present study was to challenge this model by investigating the expression of the different *P. jirovecii* MSG families at the RNA level in clinical samples of patients with PCP. However, the classical approach of mapping RNA sequencing (RNAseq) reads onto genomic sequences was inadequate for the following reasons. Firstly, ambiguous mapping of the RNAseq reads and thus erroneous signals could result from (i) the repetitive nature of these genes because they are made of conserved domains [8-10], and (ii) the presence of identical sequences in several *msg* genes due to the recombinations creating mosaicism by gene conversion. Secondly, the set of subtelomeres present in the cells of each subpopulation probably varies considerably because (i) a different *msg*-I gene may be linked to the UCS, and (ii) recombinations may occur frequently between *msg* genes. Consequently, only the most stable part of the subtelomeres that is present in the majority of the cells can be assembled, at least until single cell sequencing will be adapted to *P. jirovecii*. Thus, a complete set of subtelomeres that could be used as a reference for mapping RNAseq reads does not exist, even

when analyzing the genome assembly and RNAseq reads from the same *P. jirovecii* isolate. In order to circumvent these limitations, we developed dedicated bioinformatics procedures in order to analyze the expression of the MSG genes.

Material and methods

RNA extraction and whole transcriptome amplification

Total RNAs were extracted from the bronchoalveolar lavage fluid specimens (BALFs; see supplementary information) using the RiboPure yeast kit (Ambion). The whole transcriptome was amplified from each RNA preparation using the SeqPlex^{Tm} RNA Amplification Kit (Sigma). The procedure resulted in cDNAs with a mean size of 100 to 150 bps as revealed using a 2100 Bioanalyzer system (Agilent Technologies). According to the manufacturer, such size, *i.e.* below 200-400 bps, indicates degradation of input RNA. Only patient 2's BALF led to a mean cDNA size of 250 bps. RNA degradation was expected because of the uncontrolled period between collection of the BALFs from the patient and arrival in our laboratory, as well as the complex and varying microbiota present in these samples. We previously observed varying RNA degradation in BALFs by amplifying specific transcripts [20, 21]. Because the SeqPlex^{Tm} RNA Amplification Kit generated too small cDNAs for most samples (100 to 150 bps), the absence of genomic DNA in the RNA preparations was checked on larger cDNAs (ca. 800 bps) obtained from the same RNA preparations using the REPLI-g WTA Single Cell Kit (Qiagen) also involving random amplification. This check consisted in (i) the lack of amplification in the absence of reverse transcription (*i.e.* directly on RNA), and (ii) the lack of intron in the PCR product from the unrelated gene encoding *β*-tubulin, as we performed previously [20, 21].

Probes design

The procedure of enrichment in *P. jirovecii* RNA using bait probes was derived from that described for *Candida albicans* [22, 23]. Our SureSelect capture library included a total of 43,793 biotinylated bait probes of 120-nucleotide that were designed using the eArray software and the 1x tiling option (Agilent Technologies, https://earray.chem.agilent.com/earray/). The

probes were head-to-tail and non-overlapping. They covered completely the length of each ORF. However, no probes was placed when less than 120 nucleotides remained at the end of ORF, so that a small bias was introduced at the 3' end that was more pronounced for small genes. The probes covered a total of 4,135 *P. jirovecii* ORFs, *i.e.* all 3,772 of the reference assembly of the *P. jirovecii* genome Pneu-jiro_RU7_V2 including 181 *msg* genes ([8], 8.4Mb haploid genome; ca. 4.0Mb Orfeome), 83 genes and 25 pseudogenes of all six MSG families that we described previously [9], as well as 255 *msg* genes from previous publications (Table S1). The *msg* genes and pseudogenes were added in order to ensure enrichment in *msg* cDNAs (see text). The UCS including the single copy promoter of MSG family I was also included (locus T551_00002). The mitochondrial ORFs and the single copy rDNA genes were not covered, whereas the ORFs encoding ribosomal proteins were. The probes that mapped onto the human transcriptome present in UCSC resources using BLAT (https://genome.ucsc.edu/) were discarded. There was an average of 10.0 probes for each ORF or pseudogene.

Preparation of RNAseq libraries with enrichment in *P. jirovecii* RNA

RNA libraries for RNAseq were prepared using the Agilent SureSelect^{XT} targeted cDNA Enrichment Kit for multiplexed Illumina sequencing (Agilent Technologies, manufacturer's reference G9611A), using the 200 ng sample preparation procedure without shearing of DNA, and 16 PCR cycles for the step "Indexing and sample processing for multiplex sequencing". The only change to the manufacturer's instructions has been that the samples were dried 3 to 5 min at room temperature rather than at 37°C after all purifications with AMPure XP beads. Briefly, double-stranded cDNA was produced with adapters ligated to both ends of the cDNAs, allowing subsequent amplification using primers matching the adapters. The addition of primers to the cDNAs during the procedure was checked using the Agilent Bioanalyser. Each library received a different index that allowed several libraries to be sequenced together

(multiplexing). Amplified double-stranded cDNA was incubated at 65°C for 24 h with our capture library of biotinylated probes described above. The hybridized sequences were captured with magnetic streptavidin beads. They were next linearly amplified using provided primers and indexed in a new PCR. The non-enriched sample 1NE was sequenced directly after the SeqPlex^{Tm} RNA Amplification Kit, without applying the Agilent SureSelect^{XT} targeted cDNA Enrichment Kit.

RNA sequencing

Libraries resulting from the Agilent SureSelect^{XT} targeted cDNA Enrichment Kit were sequenced on an Illumina MiSeq with a Micro Reagent Kit v2 (300 cycles). Sequencing data were processed using Illumina bcl2fastq2 conversion software v2.20. RNAseq paired Illumina reads were merged using BBMap (v. 37.82). Merged reads mapping onto the human reference genome (grch38) using HiSat2 (v.2.26.0) were discarded, and non-mapping reads were extracted using a combination of samtools (v1.8, parameters: view -h -b -f 4) and bedtools (v2.26.0, parameters: bamtofastq). The obtained human-filtered merged reads were de-duplicated using the program cd-hit-dup from cd-hit (v 4.6.8). The proportion of *P. jirovecii* sequences versus human ones in each samples of RNAseq reads was determined using a splice aware mapper with standard settings (STAR 2.6.0c). The other procedures are described in the supplementary information.

Results

RNAseq of patients' clinical samples with enrichment in *P. jirovecii* RNA

In order to study the expression of the *P. jirovecii* MSG superfamily, we analyzed total RNAs extracted from the bronchoalveolar lavage fluid (BALF) specimens of six patients with PCP. The proportion of *P. jirovecii* RNA in such samples being low (ca. 3%)[24], an enrichment step was required. To that aim, we used the Agilent SureSelect^{XT} targeted cDNA Enrichment Kit relying on hybridization to bait probes covering the whole *P. jirovecii* orfeome (complete set of protein coding sequences). In order to ensure enrichment in *msg* cDNAs (complementary DNA synthetized from RNA), probes were also derived from published *msg* gene sequences (Table S1). The latter approach was based on the presence of conserved motifs within *msg* genes and the fact that the kit allows mismatches in order to detect sequence variants. The enrichment procedure increased the proportion of *P. jirovecii* RNA to 20-60% (Table 1, see enriched ad non-enriched samples of patient 1). After elimination of the human reads, the merged Illumina RNAseq paired sequence reads (150 to 250 nucleotides) were deduplicated in order to avoid biases due to the PCR amplification steps included in the sample preparation. These samples of reads, which characteristics are given in Table 1, were subsequently analyzed.

Assignment of RNAseq reads and MSG expression analysis

Specific weighted profiles (similar to consensus sequences) based on published *msg* gene sequences were generated for each of the six MSG families. Despite that a single gene sequence exists, specific profiles were also generated for eight control genes that can be considered as housekeeping (except superoxide dismutase). These profiles were used to assign RNAseq reads to one MSG family or control gene using a conservative best hit approach (Table 2; see methods). The reproducibility of the procedure was assessed by the similarity of the results

obtained for two independent analyses of patient 2's BALF (samples 2Ea and 2Eb). Importantly, the results of the enriched and non-enriched samples from patient 1 were also similar (samples 1E and 1NE). This latter result validated the procedure of enrichment in *P. jirovecii* RNA, including in *msg* transcripts. All eight samples of reads from the six patients gave comparable results. The vast majority of the assigned reads was from MSG family I (77.1 to 95.0%), the second most important population was from MSG family III (2.8 to 17.7%), and all other MSG families and control genes were less represented (0.1 to 4.2%). Transcripts of MSG family VI, superoxide dismutase, and beta tubulin were not detected. These observations demonstrated that, at the population level, genes of all MSG families are expressed, except possibly those of family VI. Genes of families I and III are expressed at a higher level than the other families and all control genes. The level of expression of the genes of family I was the highest, *i.e.* 20 to 50 times higher than the housekeeping genes investigated.

In silico estimation of the diversity of the MSG genes expressed

The diversity of the genes expressed of each MSG family was estimated by calling single nucleotide variants (SNVs) within a window sliding along the alignment of the RNAseq reads with the specific weighted profile. The optimal size of the window to count haplotypes for all MSG families was determined to be 30 bps (Figure S1A). Because the RNAseq reads were deduplicated, the number of haplotypes (sequences with specific SNVs) obtained can be considered as a surrogate of the number of the different *msg* isoforms expressed. This number depended on the lowest proportion of the reads among those present in the window used to support each haplotype, especially for family I (Figure S1B). In order to avoid detecting sequencing errors while being sensitive enough, the proportion of 0.01 was chosen for all our analyses because it is the usual error rate within Illumina reads [25, 26]. For all MSG families in all samples of RNAseq reads, the number of haplotypes identified was most often

proportional to the read coverage along the profile (for example family I in sample 3E, Figure 2A). The only exceptions were for family I in samples 1E and 2Ea that had more reads than the other samples (Figure 2B and 2C). The peaks of coverage at 3' and especially 5' regions in Figure 2 are likely to result from RNA degradation, as well as to gene-specific degradation pattern [27]. Interestingly, the two samples 1E and 2Ea with sufficient coverage provided drastically reduced numbers of *msg*-I haplotypes at the same four locations along the profile, at positions ca. 100, 2000, 2300, and 2500 bps. These positions might correspond to conserved regions between protein domains where recombinations between these genes occur preferentially. We calculated the median of the numbers of haplotypes obtained along each MSG profile for all samples of RNAseq reads (Table 3; a number of values could not be obtained because of insufficient read coverage). All these values should be considered as minimal because of the conservative parameters used and the dependency on coverage. The observed number of haplotypes varied from three to 21 for MSG family I, and from one to four for the other families. The two samples 1E and 2Ea with a sufficient coverage for family I provided both a value of ca. 20. These results suggested that (i) family I presents the highest diversity of isoforms expressed during *P. jirovecii* infection, and (ii) that a number of different isoforms of each MSG family are expressed, except possibly for families V (only one haplotype detected) and VI (no reads detected).

In vitro assessment of the diversity of *msg*-I isoforms expressed

We assessed at the DNA level the diversity of *msg*-I isoforms expressed in the *P. jirovecii* population infecting each of the six patients. To that aim, the repertoire of these genes was amplified from each BALF's genomic DNA using primers localized in the UCS containing the single copy promoter and at the end of the genes. The PCR product was subcloned and several subclones were sequenced. The samples of all six patients presented a significant diversity of

250 *msg*-I isoforms expressed, *i.e.* 27 to 80% of 10 to 15 subclones sequenced were unique (Table
251 4). Only two patients shared a single sequence (patients 1 and 2).

Discussion

The human pathogenic fungus *P. jirovecii* harbors most probably a system of surface antigenic variation ensuring presumably both escape from host immunity and adhesion to target cells. This system involves six families of hypervariable surface glycoproteins, the MSGs, family I being under mutually exclusive expression at the individual cell level. In the present study, we analyzed the pattern of expression of the genes encoding these proteins at the RNA level. The results were similar in six patients with PCP. The *msg* transcripts included members of at least five families. The level of expression of families I and III was higher than those of the other MSG families and housekeeping genes. Family I was by far expressed at the highest level. A number of different isoforms of at least four of the six families were expressed. Importantly, the six patients that we investigated were each co-infected with several *P. jirovecii* strains (see methods), so that the results are means from several strains. Nevertheless, the similarity of the results of the six patients suggests that all strains expressed similarly the MSGs.

We did not detect transcripts of the MSG family VI. In *P. murina*, the proteins of this family are present only at the surface of the ascospores, within asci or recently released from the asci [28]. This observation suggested a particular regulation of this family during the cell cycle. Because asci represent generally a minority of ca. 5% in the infecting *P. jirovecii* populations [29], it is possible that the transcripts encoding these proteins were present in amounts too low to be detected by our procedure. Interestingly, the genes of this family are all localized in the distal region of the subtelomeres, *i.e.* most distant from the telomeres and closest to the genomic genes. Moreover, the recombinations between them are less frequent than between the genes of the other MSG families [9]. A relationship between chromosomal location, expression in ascospores, and low frequency of mosaicism is likely to exist.

Our observations support several aspects of our model for the *P. jirovecii* antigenic variation system (Figure 1). The observed expression of several *msg*-I isoforms within each

277 infecting population is consistent with the hypothesis of a continuous segregation of
278 subpopulations expressing each a different single isoform. The value of 20 different isoforms
279 that we observed in two patients is close to the single value of 18 reported so far [9]. The latter
280 value as well as those reported in the present work being all minimal estimations, it is likely
281 that a higher diversity of family I is actually expressed. This hypothesis is also suggested by
282 the peaks up to ca. 30-35 haplotypes that we observed for this family (Figure 2B and 2C, at
283 position 227). As far as the other MSG families are concerned, the mean number of haplotypes
284 observed were always dependent on the read coverage. Nevertheless, peaks up to 10-20
285 haplotypes were observed for all families in sample 1E with the highest coverage, suggesting
286 that all families might also present an important diversity of genes expressed. The expression
287 of several isoforms of all MSG families that we observed, except possibly of family V and VI,
288 is compatible with the postulated independent expression thanks to the promoter that each gene
289 possesses. However, our analyses did not allow assessing if all genes of each family were
290 transcribed. Thus, it remains to determine if these latter genes are constitutively expressed,
291 subject to a regulation during the cell cycle, and/or silenced due to the proximity of the
292 telomeres (by the "telomere position effect")[30].

293 The expression at a very high level of *msg*-I isoforms we observed in *P. jirovecii* is
294 consistent with our model and previous studies at the protein level [8, 11]. This high expression
295 might be due to transcription enhancement driven by the intron present in the UCS that is larger
296 than that present in the other MSG families [9, 10]. On the other hand, the high expression of
297 family III is a new feature. The latter may not be due to the presence of two introns of common
298 size for *P. jirovecii* (40 to 60 bps) close to the promoter of *msg*-III genes because a similar
299 arrangement is present for *msg*-II genes that are not over-expressed.

300 We analyzed immunocompromised patients with active PCP. However, the antigenic
301 surface variation system of *P. jirovecii* has probably evolved in immunocompetent humans

without PCP, and thus be above all a colonization factor. Colonized individuals include potentially several categories of humans, *e.g.* infants experiencing primo-infection, transient carriers such as healthcare workers in contact with PCP patients, patients with chronic lung diseases, pregnant women, and elderly people [31]. In these colonized individuals, the number of subpopulations expressing a different *msg*-I isoform could be reduced by the valid immune system. This hypothesis deserves to be tested in order to better understand the surface antigenic variation system of *P. jirovecii*.

In conclusion, our results enhance the understanding of the mechanisms involved in the surface antigenic variation of *P. jirovecii*, as well as of its cell surface structure. The postulated strategy to produce continuously subpopulations that are antigenically distinct would be unique among human pathogens, and might be associated with the non-sterile niche within lungs [9, 32]. By contrast, pathogens that occupy sterile niches (blood, tissue), such as *Plasmodium* and *Trypanosoma,* rely on cell populations that are antigenically homogeneous. Further work is needed to decipher the unique surface antigenic variation system of *P. jirovecii*.

Acknowledgments

We thank Dr. Michael Walter of the Agilent Diagnostic and Genomics group for his help in the design of the probes and RNA extraction protocol. We also thank Thierry Schuepbach for help in the initial phase of the analyses of the haplotypes numbers. Sequencing was performed at the Lausanne Genomic Technologies Facility, University of Lausanne, Switzerland. Computations were performed at the Vital-IT Center for High-Performance Computing of the Swiss Institute of Bioinformatics (http://www.vital-it.ch).

References

1. Redhead SA, Cushion MT, Frenkel JK, Stringer JR. *Pneumocystis* and *Trypanosoma cruzi*: nomenclature and typifications. J Eukaryot Microbiol **2006**; 53:2–11.
2. Gigliotti F, Limper AH, Wright T. *Pneumocystis*. Cold Spring Harb Perspect Med **2014**; 4:a019828.
3. Keely SP, Renauld H, Wakefield AE, et al. Gene arrays at *Pneumocystis carinii* telomeres. Genetics **2005**; 170, 1589-600.
4. Stringer JR. Antigenic variation in *Pneumocystis.* J Euk Microbiol **2007**; 54:8–13.
5. Keely SP, Stringer JR. Complexity of the MSG gene family of *Pneumocystis carinii*. BMC Genomics **2009**; 10:367.
6. Pottratz ST. *Pneumocystis carinii* interactions with respiratory epithelium. Semin Respir Infect **1998**; 13:323-9.
7. Hoving JC. *Pneumocystis* and interactions with host immune receptors. PLoS Pathogens **2018**; 14:e1006807.
8. Ma L, Chen Z, Huang DW, et al. Genome analysis of three *Pneumocystis* species reveals adaptation mechanisms to life exclusively in mammalian hosts. Nat Commun **2016**; 7:10740.
9. Schmid-Siegert E, Richard S, Luraschi A, Mühlethaler K, Pagni M, Hauser PM. Mechanisms of surface antigenic variation in the human pathogenic fungus *Pneumocystis jirovecii*. mBio **2017**; 8, e01470-17.
10. Ma L, Chen Z, Huang, et al. Diversity and complexity of the large surface protein family in the compacted genomes of various *Pneumocystis* species. mBio **2020**; 11:e02878-19.
11. Kutty G, Shroff R, Kovacs JA. Characterization of *Pneumocystis* major surface glycoprotein gene (*msg*) promoter activity in *Saccharomyces cerevisiae*. Eukaryot Cell **2013**; 12:1349 –355.

12. Stringer JR, Cushion MT. The genome of *Pneumocystis carinii*. FEMS Immunol Med Microbiol **1998**; 22:15–26.
13. Wyder MA, Rasch EM, Kaneshiro ES. Quantitation of absolute *Pneumocystis carinii* nuclear DNA content. Trophic and cystic forms isolated from infected rat lungs are haploid organisms. J Eukaryot Microbiol **1998**; 45:233–9.
14. Martinez A, Aliouat EM, Standaert-Vitse A, Werkmeister E, Pottier M, Pincon C, Dei-Cas E, Aliouat-Denis SCM. Ploidy of cell-sorted trophic and cystic forms of *Pneumocystis carinii*. PLoS ONE **2011**; 6:e20935.
15. Hauser PM, Cushion MT. Is sex necessary for the proliferation and transmission of *Pneumocystis*? PLoS Pathogens **2018**; 14:e1007409.
16. Kutty G, Ma L, Kovacs JA. Characterization of the expression site of the major surface glycoprotein of human-derived *Pneumocystis carinii*. Mol Microbiol **2001**; 42:183-93.
17. Linke MJ, Smulian AG, Stringer JR, Walzer PD. Characterization of multiple unique cDNAs encoding the major surface glycoprotein of rat-derived *Pneumocystis carinii*. Parasitol Res **1994**; 80:478–86.
18. Angus CW, Tu A, Vogel P, Qin M, Kovacs JA. Expression of variants of the major surface glycoprotein of *Pneumocystis carinii.* J Exp Med **1996**;183:1229-34.
19. Kutty G, Maldarelli F, Achaz G, Kovacs JA. Variation in the major surface glycoprotein genes in *Pneumocystis jirovecii*. J Infect Dis **2008**; 198:741-9.
20. Richard S, Almeida JMGCF, Cissé OH, Luraschi A, Nielsen O, Pagni M, Hauser PM. Functional and expression analyses of the *Pneumocystis* MAT genes suggest obligate sexuality through primary homothallism within host lungs. mBio **2018**; 9:e02201-17.
21. Luraschi A, Richard S. Almeida JMGCF, Pagni M, Cushion MT, Hauser PM. Expression and immuno-staining analyses suggest that *Pneumocystis* primary homothallism involves

372 trophic cells displaying both Plus and Minus pheromone receptors. mBio **2019**; 10:e01145-
373 19.
374 22. Amorim-Vaz S, Tran VDT, Pradervand S, Pagni M, Coste AT, Sanglard D. RNA
375 Enrichment method for quantitative transcriptional analysis of pathogens *in vivo* applied
376 to the fungus *Candida albicans.* mBio **2015**; 6:e00942-15.
377 23. Amorim-Vaz S, Sanglard D. Novel approaches for fungal transcriptomics from host
378 samples. Front Microbiol **2016**; 6:1571.
379 24. Cissé OH, PagnI M, Hauser PM. Genome sequence of the uncultivated fungal pathogen
380 *Pneumocystis jirovecii.* mBio **2012**; 4:e00428-12.
381 25. Minoche AE, Dohm JC, Himmelbauer H. Evaluation of genomic high-throughput
382 sequencing data generated on Illumina HiSeq and genome analyzer systems. Genome Biol
383 **2011**; 12:R112.
384 26. Heydari M, Miclotte G, Demeester P, Van de Peer Y, Fostier J. Evaluation of the impact
385 of Illumina error correction tools on *de novo* genome assembly. BMC Bioinformatics
386 **2017**; 18:374.
387 27. Xiong B, Yang Y, Fineis FR, Wang JP. DegNorm: normalization of generalized transcript
388 degradation improves accuracy in RNA-seq analysis. Gen Biol 2019; 20:75.
389 28. Bishop LR, Davis AS, Bradshaw K, Gamez M, Cissé OH, Wang H, Ma L, Kovacs JA.
390 Characterization of p57, a stage-specific antigen of *Pneumocystis murina*. J Infect Dis
391 **2018**; 218:282–90.
392 29. Aliouat-Denis CM, Martinez A, Aliouat E, Pottier M, Gantois N, Dei-Cas E. The
393 *Pneumocystis* life cycle. Mem Inst Oswaldo Cruz **2009**; 104 :419–26.
394 30. Barry JD, Ginger ML, Burton P, McCulloch R. Why are parasite contingency genes often
395 associated with telomeres? Int J Parasitol 2003; 33:29–45.

31. Morris A, Norris KA. Colonization by *Pneumocystis jirovecii* and its role in disease. Clin Microbiol Rev **2012**; 25:297–317.

32. Hauser PM. Is the unique camouflage strategy of *Pneumocystis* associated with its particular niche within host lungs? PLOS Pathogens **2019**; 15:e1007480.

400 # Figure legends:

401 **Figure 1.** Model for the antigenic variation system of *P. jirovecii*. Only four chromosome ends
402 out of ca. 40 are shown in each cell. The fungus segregates continuously new cells expressing
403 each a new *msg*-I isoform, as well as in minority all mosaic isoforms of the other MSG families,
404 except possibly of family VI (see text). Consequently, each *P. jirovecii* population is
405 subdivided into several subpopulations that are antigenically different and that could multiply
406 or not. The single *msg*-I isoform is expressed at a high level, whereas the isoforms of the other
407 families are transcribed at a low level. The proximal location of the *msg*-I genes within the
408 subtelomeres, closest to the telomeres, suggests that the exchange of the expressed isoform
409 might be facilitated by the concomitant exchange of the telomere through a single
410 recombination between two CRJEs. The recombinations between the *msg* genes of each family
411 that generate gene mosaicism contributing to antigenic variation are not figured. CRJE,
412 conserved recombination junction element. UCS, upstream conserved sequence.

413

414 **Figure 2.** *In silico* estimation of the diversity of *msg*-I isoforms present among samples of
415 RNAseq reads. SNVs called within a 30 bps sliding window defined 30 bps haplotypes. The
416 black line shows the number of haplotype identified along the *msg* gene, whereas the red line
417 shows the number of reads analyzed in each window. (A) RNAseq reads sample 3E. (B) sample
418 1E. (C) sample 2Ea.

Supplementary information

Figure S1. Analyses of the parameters for the *in silico* estimation of the median number of MSG haplotypes among RNAseq reads along a specific weighted profile. Sample 1E with the highest number of reads was used (families V and VI are absent because of insufficient read coverage or no reads detected, respectively). (A) sliding window size using a supporting proportion of reads among those present in the window of 0.01. (B) proportion of reads among those present in the window supporting each haplotype using a sliding window of 30 bps.

Table S1. Targeted ORFs/pseudogenes, number of probes, probe sequences.

Supplementary material and methods

Ethics approval and consent to participate

All patients provided informed written consent, which was part of the procedure for admittance in both hospitals. The admittance paperwork included the possibility to ask that their samples not be used for research. The samples were treated anonymously and were collected through a routine procedure at the hospital. This protocol was approved by the institutional review boards (www.Swissethics.ch).

Bronchoalveolar lavage fluid specimens

BALFs positive for *P. jirovecii* were collected at the University Hospitals of Lausanne (patients 1, 3, 4, 6) and Bern (2 and 5) between 2014 and 2017. The diagnostic was made using staining, methenamine silver nitrate in Lausanne [1] and immunofluorescence in Bern (MONOFLUO *Pneumocystis jirovecii* IFA Test kit, BioRad). BALFs were supplemented with 15% (vol/vol) glycerol or with three volumes of RNAlater (Ambion) as quickly as possible upon reception, frozen in liquid nitrogen, and stored at -80°C. This storage occurred the same day as BALF

collection from the patient, except for patients 4 and 6 (one and two days later, respectively). All six BALFs were co-infected with several *P. jirovecii* strains as revealed by multitarget genotyping as we described previously [2].

Creation of MSG family specific profiles

We created weighted profiles specific for each MSG family [3]. The profiles were based on the CDS (coding DNA sequences) that we published previously [2]. For each family, CDS were aligned using MAFFT (v7.310, parameters:--globalpair --maxiterate 1000). The alignments were manually curated and translated into weighted profiles using pftools (v2.3.5.d, parameters: pfmake -2 -m). Specific profiles were also derived for the eight control genes based on the single *P. jirovecii* mRNA sequence present in GenBank (actin 1, T551_01287; alpha tubulin T551_01836; beta tubulin, T551_02025; dihydrofolate reductase, T551_02883; dihydropteroate synthase, U66281; elongation factor 2, T551_00854; elongation factor 3, T551_02306; superoxide dismutase, T551_02671). Weighted profiles were aligned against a random shuffled DNA sequence collection and scores were used to calibrate the profiles with pftools (v.3, parameters: --method evd_full --heuristic-db profile --profile-sampling 25 --mode 1).

MSG expression analysis

The obtained human-filtered and de-duplicated reads were aligned against all six MSG families and control genes profiles with pftools in both senses (v3, parameters: -N -o 4). The results were further filtered to obtain a conservative best hit by removing reads with a calibrated score below 10, as well as those that had a difference between the first and second best family scores below 20% of the score for the first best hit.

Haplotype estimation

Consistently, SNVs observed in enriched and non-enriched samples 1E and 1NE were almost identical both in position and proportion (not shown). SNVs were characterized from the obtained alignments (psa-format) in a sliding-window approach. Only reads spanning the entire analyzed window were considered and reads containing identical SNVs grouped as potential haplotypes. Each windows needed 20 reads or more to be considered "callable" and only haplotypes supported by a minimum four reads with identical SNVs and 0.01 frequency (number of haplotype supporting reads/total reads in window) were considered. The median number of haplotypes for each MSG family was calculated over "callable" regions, whereas regions with insufficient reads coverage were ignored. The window size was optimized based on the sample 1E and calling haplotypes with 10 bps increasing windows sizes (10 to 100 bps, frequency cut-off of 0.01, Figure S1A). Optimization of the frequency cut-off was similarly conducted on sample 1E with a fixed window size of 30 bps and frequencies of 50, 10, 5, 1, 0.5 and 0.01 (Figure S1B). Curves of median number of haplotypes and coverage were plotted using GraphPad Prism (v8.0.1).

PCR amplification of the expressed *msg*-I isoforms repertoires

In order to avoid contaminations, PCRs were set up and analyzed in separate rooms, and negative controls were systematically performed at each experiment. PCRs were performed using the High Fidelity Expand polymerase (Roche) on genomic DNA extracted from the BALF using the QIAamp DNA minikit (Qiagen). The primers were GPI-rev (AAA TCA TGA ACG AAA TMA YCA TTG C) and GK135 (GAC AAG GAT GTT GCT TTT GAT; Kutty et al., 2001), generating a PCR product of ca. 3050 bps covering the entire *msg*-I expressed genes. Primers were synthesized by Microsynth (Balgach, Switzerland). The final concentration of $MgCl_2$ in the reaction was 3 mM. Each reaction began with denaturation for 3 min at 94 °C,

494 followed by 35 cycles of 15 sec at 94°C, 30 sec at the annealing temperature of 55°C, and 3
495 min 30 sec of elongation at 72 °C. The reactions ended with extension of 10 minutes at 72°C.
496 The realtime PCR used to determine the number of *P. jirovecii* copies in BALF given in Table
497 1 has been described previously [4].
498
499 **PCR product subcloning and sequencing**
500 Subcloning of the PCR products was carried out using the TOPO TA cloning Kit (Life
501 Technologies, Inc.). Sanger sequencing of ca. 850 bps of the 5' end of the cloned *msg*-I gene
502 was performed on minipreparation of subclone plasmid DNA [5] using primer M13 forward
503 (GTA AAA CGA CGG CCA GT). When the insert was in the inadequate direction, the
504 subclone was sequenced again using primer M13 reverse (AGC GGA TAA CAA TTT CAC
505 ACA GG). Because only one strand was sequenced, and thus that few sequencing errors were
506 possible, subclones with one or two bps different were considered identical. This concerned
507 only two to five subclones in five of the six patients studied. After alignment and trimming, the
508 identities between subclones' sequences were determined using the MAFFT identity matrix
509 tool (https://www.ebi.ac.uk/Tools/msa/mafft/).
510
511 **Sequence data accession number**
512 The RNA-seq data without human sequences analyzed in this work are accessible under the
513 BioProject PRJNA608830 with bioSample accession numbers SAMN14209915 to
514 SAMN14209922.
515
516 **References**
517 1. Musto L, Flanigan M, Elbadawi A. Ten-minute silver stain for *Pneumocystis carinii* and
518 fungi in tissue sections. Arch Pathol Lab Med **1982**; 106:292–4.

519 2. Schmid-Siegert E, Richard S, Luraschi A, Mühlethaler K, Pagni M, Hauser PM.
520 Mechanisms of surface antigenic variation in the human pathogenic fungus *Pneumocystis*
521 *jirovecii*. mBio **2017**; 8, e01470-17.
522 3. Bucher P, Bairoch A. A generalized profile syntax for biomolecular sequence motifs and
523 its function in automatic sequence interpretation. Proc Int Conf Intel Syst Mol Biol **1994**;
524 2:53– 61.
525 4. Richard S, Almeida JMGCF, Cissé OH, Luraschi A, Nielsen O, Pagni M, Hauser PM.
526 Functional and expression analyses of the *Pneumocystis* MAT genes suggest obligate
527 sexuality through primary homothallism within host lungs. mBio **2018**; 9:e02201-17.
528 5. Birnboim HC, Doly J. A rapid alkaline extraction procedure for screening recombinant
529 plasmid DNA. Nucl Ac Res 1979; 7:1513–23.

Table 1. Characteristics of the sets of RNAseq reads from BALFs of six patients with PCP.

	RNAseq reads sample [a]							
	1E	**1NE**	**2Ea**	**2Eb**	**3E**	**4E**	**5E**	**6E**
Underlying disease [b]	HIV+	HIV+	KT	KT	ALL	PNET	MM	HIV+
P. jirovecii copies in BALF by realtime PCR (x10^6/ml)	22.9	22.9	0.7	0.7	1,111	3.1	4.2	57.4
Read pairs (x10^6)	12.6	8.1	2.7	1.0	1.0	11.0	0.5	2.7
Merged read pairs (x10^6)	5.5	4.3	2.0	0.9	0.9	1.0	0.4	2.5
P. jirovecii reads in merged read pairs (%)	55	0.2	21	23	62	26	52	57
Human reads in merged read pairs (%)	23	89	50	57	7	39	13	5
Deduplicated merged read pairs without human (x10^3)	1’499	464	703	126	183	153	73	383

[a] Code sample name: 1 to 6, patient number. E, enriched in *P. jirovecii* cDNA using Sureselect kit. NE, non-enriched. a and b, duplicate from the same patient’s BALF (patient 2).

[b] KT, Kidney transplantation. ALL, Acute lymphocytic leukemia. PNET, primitive neuroectodermal tumor. MM, multiple myeloma.

Table 2. Proportion (%) of RNAseq reads assigned to one MSG family or control gene. [a]

	RNAseq reads sample [b]							
	1E 182,389 [c]	**1NE** 1,213	**2Ea** 61,248	**2Eb** 10,641	**3E** 7,324	**4E** 7,106	**5E** 8,948	**6E** 49,497
MSG family I (A1) [d]	79.4	87.4	82.5	78.3	95.0	77.1	87.6	86.1
MSG family II (A3)	3.3	2.5	3.1	2.6	1.1	2.1	3.4	3.3
MSG family III (A3)	6.0	5.4	8.2	14.5	2.8	17.7	7.1	6.1
MSG family IV (B)	1.6	1.1	2.1	2.4	0.2	3.0	0.8	4.3
MSG family V (D)	0.3	0.2	1.4	0.1	0.2	0	0	0
MSG family VI (E)	0	0	0	0	0	0	0	0
Actin 1	0.3	0.3	0.1	0	0	0	0	0
Alpha tubulin	4.2	1.8	1.8	1.4	0.2	0	0	0
Beta tubulin	0	0	0	0	0	0	0	0
Dihydrofolate reductase	1.8	0.4	0.5	0.7	0.5	0.1	0	0.1
Dihydropteroate synthase	0.3	0	0	0	0	0	0	0
Elongation factor 2	1.4	0.5	0.3	0.7	0	0	1.2	0
Elongation factor 3	1.3	0.5	0.2	0	0	0	0	0.1
Superoxide dismutase	0	0	0	0	0	0	0	0

[a] RNAseq reads were assigned to one MSG family or gene using specific weighted profiles.

[b] Code sample name: 1 to 6, patient number. E, enriched in *P. jirovecii* cDNA using Sureselect kit. NE, non-enriched. a and b, duplicates from the same patient's BALF (patient 2).

[c] Total number of merged reads (without human, de-duplicated) assigned to one MSG family or control gene.

[d] The nomenclature of the MSG families of Ma et al. [8] is in parentheses.

Table 3. Median number of expressed haplotypes for each MSG family. [a]

	RNAseq reads sample [b]							
	1E	**1NE**	**2Ea**	**2Eb**	**3E**	**4E**	**5E**	**6E**
	165,244 [c]	1,172	59,594	10,418	7,273	7,099	8,850	49,398
I (A1) [d]	18	nd [e]	21	4	3	nd	2	6
II (A3)	4	nd	4	nd	nd	nd	nd	nd
III (A3)	1	nd	4	nd	nd	nd	nd	nd
IV (B)	2	nd	3	nd	nd	nd	nd	nd
V (D)	nd	nd	nd	1	nd	nd	nd	nd
VI (E)	nd	nd	nd	nd	nd	nd	nd	nd

[a] The median number of expressed haplotypes was determined for each MSG family using SNVs calling within the RNAseq reads within a sliding window.

[b] Code sample name: 1 to 6, patient number. E, enriched in *P. jirovecii* cDNA using Sureselect kit. NE, non-enriched. a and b, duplicates from the same patient's BALF (patient 2).

[c] Total number of MSG RNAseq reads analyzed.

[d] The nomenclature of the MSG families of Ma et al. [8] is in parentheses.

[e] not determined because of insufficient read coverage (or no reads for family VI).

Table 4. Diversity of *msg*-I isoforms expressed. [a]

	Patient no.					
	1	**2**	**3**	**4**	**5**	**6**
No. subclones sequenced	12	11	10	10	10	15
% unique subclones (no.)	58 (7)	27 (3)	80 (8)	70 (7)	30 (3)	33 (5)
% mean identity of subclones (range)	69 (66-81)	72 (68-75)	68 (63-80)	69 (63-78)	98 (97-98)	73 (65-88)

[a] The repertoire of expressed *msg*-I genes (*i.e.* linked to and thus under the control of the single copy promoter present in the UCS) was amplified by PCR from the genomic DNA of the patient's BALF, the PCR product was subcloned in a plasmid, and ca. 850 bps of the 5' end of each subclone were sequenced.

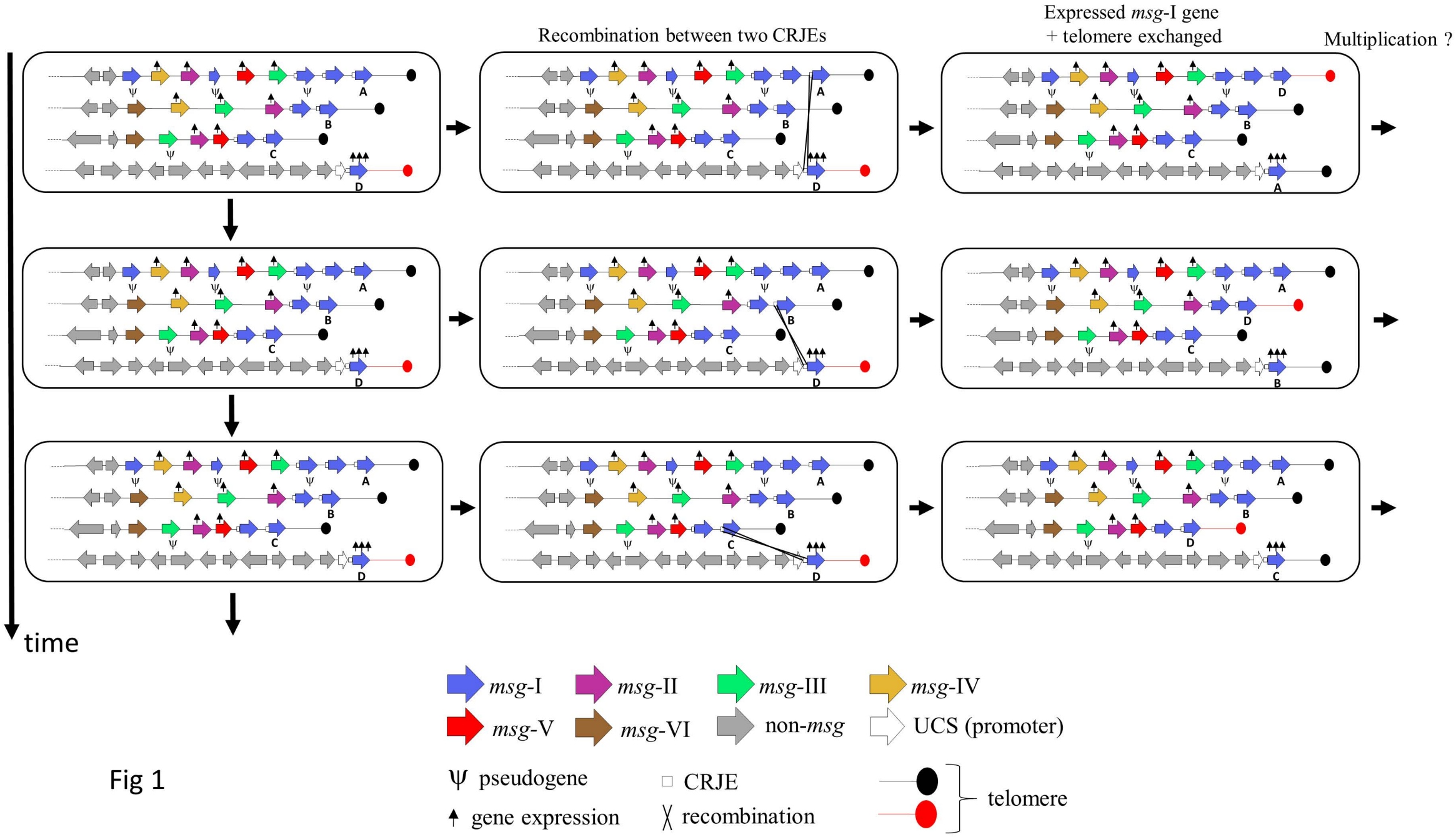
Recombination between two CRJEs
Expressed msg-I gene
+ telomere exchanged
Multiplication ?
time
msg-I
msg-II
msg-III
msg-IV
msg-V
msg-VI
non-msg
UCS (promoter)
Ψ pseudogene
CRJE
gene expression
recombination
telomere

Fig 1

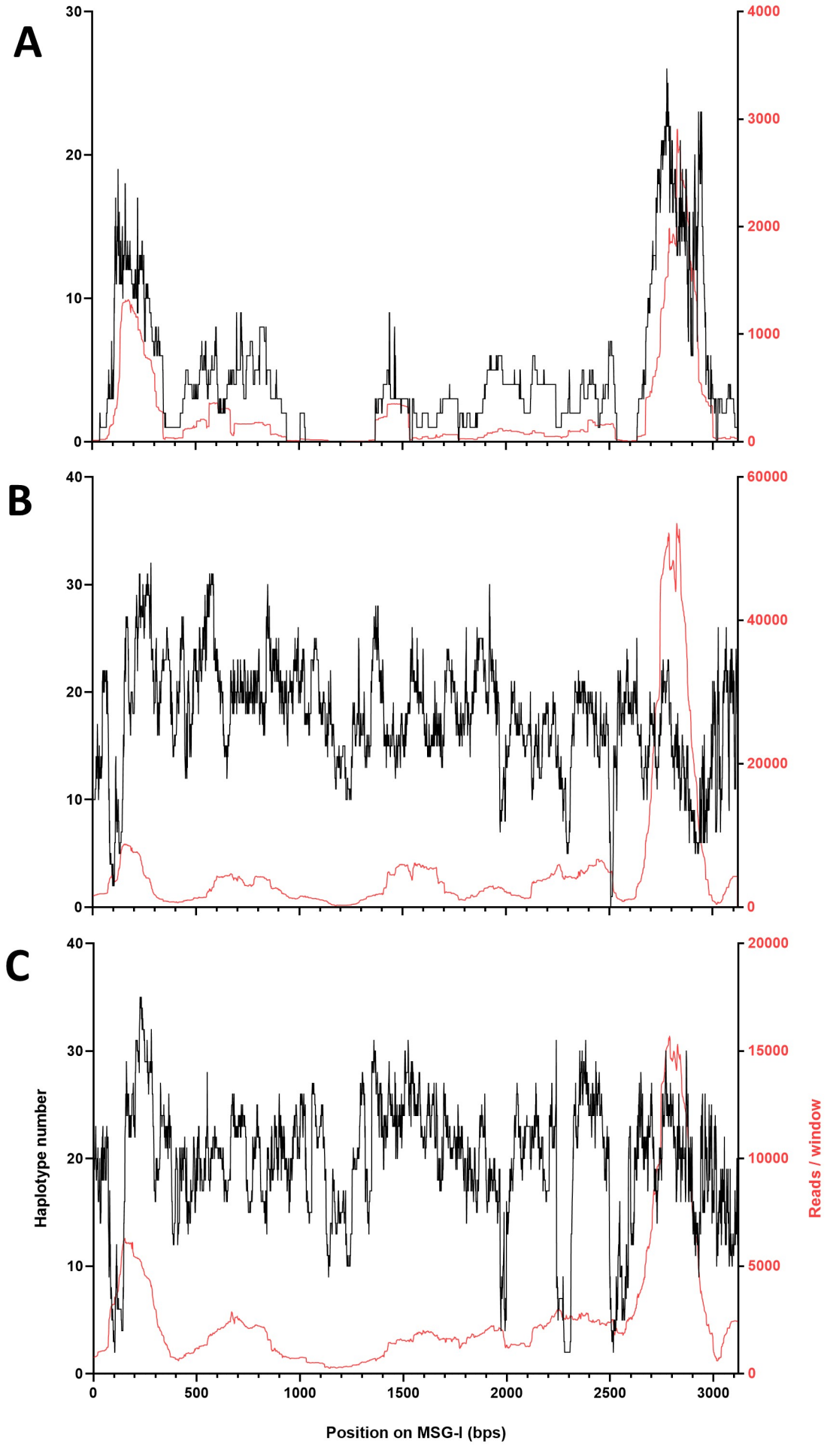
A
B
C
Haplotype number
Reads / window
Position on MSG-I (bps)
0
500
1000
1500
2000
2500
3000
0
10
20
30
40
0
1000
2000
3000
4000
0
20000
40000
60000
0
5000
10000
15000
20000

Fig 2

Supplementary materials

Figure S1. Analyses of the parameters for the *in silico* estimation of the median number of MSG haplotypes among RNAseq reads along a specific weighted profile. Sample 1E with the highest number of reads was used (families V and VI are absent because of insufficient read coverage or no reads detected, respectively). (A) sliding window size using a supporting proportion of reads among those present in the window of 0.01. (B) proportion of reads among those present in the window supporting each haplotype using a sliding window of 30 bps.

Table S1. Targeted ORFs/pseudogenes, number of probes, probe sequences.

Supplementary material and methods

Ethics approval and consent to participate

All patients provided informed written consent, which was part of the procedure for admittance in both hospitals. The admittance paperwork included the possibility to ask that their samples not be used for research. The samples were treated anonymously and were collected through a routine procedure at the hospital. This protocol was approved by the institutional review boards (www.Swissethics.ch).

Bronchoalveolar lavage fluid specimens

BALFs positive for *P. jirovecii* were collected at the University Hospitals of Lausanne (patients 1, 3, 4, 6) and Bern (2 and 5) between 2014 and 2017. The diagnostic was made using staining, methenamine silver nitrate in Lausanne [1] and immunofluorescence in Bern (MONOFLUO *Pneumocystis jirovecii* IFA Test kit, BioRad). BALFs were supplemented with 15% (vol/vol) glycerol or with three volumes of RNAlater (Ambion) as quickly as possible upon reception, frozen in liquid nitrogen, and stored at -80°C. This storage occurred the same day as BALF

collection from the patient, except for patients 4 and 6 (one and two days later, respectively). All six BALFs were co-infected with several *P. jirovecii* strains as revealed by multitarget genotyping as we described previously [2].

Creation of MSG family specific profiles

We created weighted profiles specific for each MSG family [3]. The profiles were based on the CDS (coding DNA sequences) that we published previously [2]. For each family, CDS were aligned using MAFFT (v7.310, parameters:--globalpair --maxiterate 1000). The alignments were manually curated and translated into weighted profiles using pftools (v2.3.5.d, parameters: pfmake -2 -m). Specific profiles were also derived for the eight control genes based on the single *P. jirovecii* mRNA sequence present in GenBank (actin 1, T551_01287; alpha tubulin T551_01836; beta tubulin, T551_02025; dihydrofolate reductase, T551_02883; dihydropteroate synthase, U66281; elongation factor 2, T551_00854; elongation factor 3, T551_02306; superoxide dismutase, T551_02671). Weighted profiles were aligned against a random shuffled DNA sequence collection and scores were used to calibrate the profiles with pftools (v.3, parameters: --method evd_full --heuristic-db profile --profile-sampling 25 --mode 1).

MSG expression analysis

The obtained human-filtered and de-duplicated reads were aligned against all six MSG families and control genes profiles with pftools in both senses (v3, parameters: -N -o 4). The results were further filtered to obtain a conservative best hit by removing reads with a calibrated score below 10, as well as those that had a difference between the first and second best family scores below 20% of the score for the first best hit.

Haplotype estimation

Consistently, SNVs observed in enriched and non-enriched samples 1E and 1NE were almost identical both in position and proportion (not shown). SNVs were characterized from the obtained alignments (psa-format) in a sliding-window approach. Only reads spanning the entire analyzed window were considered and reads containing identical SNVs grouped as potential haplotypes. Each windows needed 20 reads or more to be considered "callable" and only haplotypes supported by a minimum four reads with identical SNVs and 0.01 frequency (number of haplotype supporting reads/total reads in window) were considered. The median number of haplotypes for each MSG family was calculated over "callable" regions, whereas regions with insufficient reads coverage were ignored. The window size was optimized based on the sample 1E and calling haplotypes with 10 bps increasing windows sizes (10 to 100 bps, frequency cut-off of 0.01, Figure S1A). Optimization of the frequency cut-off was similarly conducted on sample 1E with a fixed window size of 30 bps and frequencies of 50, 10, 5, 1, 0.5 and 0.01 (Figure S1B). Curves of median number of haplotypes and coverage were plotted using GraphPad Prism (v8.0.1).

PCR amplification of the expressed *msg*-I isoforms repertoires

In order to avoid contaminations, PCRs were set up and analyzed in separate rooms, and negative controls were systematically performed at each experiment. PCRs were performed using the High Fidelity Expand polymerase (Roche) on genomic DNA extracted from the BALF using the QIAamp DNA minikit (Qiagen). The primers were GPI-rev (AAA TCA TGA ACG AAA TMA YCA TTG C) and GK135 (GAC AAG GAT GTT GCT TTT GAT; Kutty et al., 2001), generating a PCR product of ca. 3050 bps covering the entire *msg*-I expressed genes. Primers were synthesized by Microsynth (Balgach, Switzerland). The final concentration of $MgCl_2$ in the reaction was 3 mM. Each reaction began with denaturation for 3 min at 94 °C,

followed by 35 cycles of 15 sec at 94°C, 30 sec at the annealing temperature of 55°C, and 3 min 30 sec of elongation at 72 °C. The reactions ended with extension of 10 minutes at 72°C. The realtime PCR used to determine the number of *P. jirovecii* copies in BALF given in Table 1 has been described previously [4].

PCR product subcloning and sequencing

Subcloning of the PCR products was carried out using the TOPO TA cloning Kit (Life Technologies, Inc.). Sanger sequencing of ca. 850 bps of the 5' end of the cloned *msg*-I gene was performed on minipreparation of subclone plasmid DNA [5] using primer M13 forward (GTA AAA CGA CGG CCA GT). When the insert was in the inadequate direction, the subclone was sequenced again using primer M13 reverse (AGC GGA TAA CAA TTT CAC ACA GG). Because only one strand was sequenced, and thus that few sequencing errors were possible, subclones with one or two bps different were considered identical. This concerned only two to five subclones in five of the six patients studied. After alignment and trimming, the identities between subclones' sequences were determined using the MAFFT identity matrix tool (https://www.ebi.ac.uk/Tools/msa/mafft/).

Sequence data accession number

The RNA-seq data without human sequences analyzed in this work are accessible under the BioProject PRJNA608830 with bioSample accession numbers SAMN14209915 to SAMN14209922.

References

1. Musto L, Flanigan M, Elbadawi A. Ten-minute silver stain for *Pneumocystis carinii* and fungi in tissue sections. Arch Pathol Lab Med **1982**; 106:292–4.
2. Schmid-Siegert E, Richard S, Luraschi A, Mühlethaler K, Pagni M, Hauser PM. Mechanisms of surface antigenic variation in the human pathogenic fungus *Pneumocystis jirovecii*. mBio **2017**; 8, e01470-17.
3. Bucher P, Bairoch A. A generalized profile syntax for biomolecular sequence motifs and its function in automatic sequence interpretation. Proc Int Conf Intel Syst Mol Biol **1994**; 2:53– 61.
4. Richard S, Almeida JMGCF, Cissé OH, Luraschi A, Nielsen O, Pagni M, Hauser PM. Functional and expression analyses of the *Pneumocystis* MAT genes suggest obligate sexuality through primary homothallism within host lungs. mBio **2018**; 9:e02201-17.
5. Birnboim HC, Doly J. A rapid alkaline extraction procedure for screening recombinant plasmid DNA. Nucl Ac Res **1979**; 7:1513–23.

Sequence data accession number 463

The RNA-seq data without human sequences analyzed in this work are accessible under the BioProject PRJNA608830 with bioSample accession numbers SAMN14209915 to SAMN14209922 in the European Nucleotide Archive (https://www.ebi.ac.uk/ena).

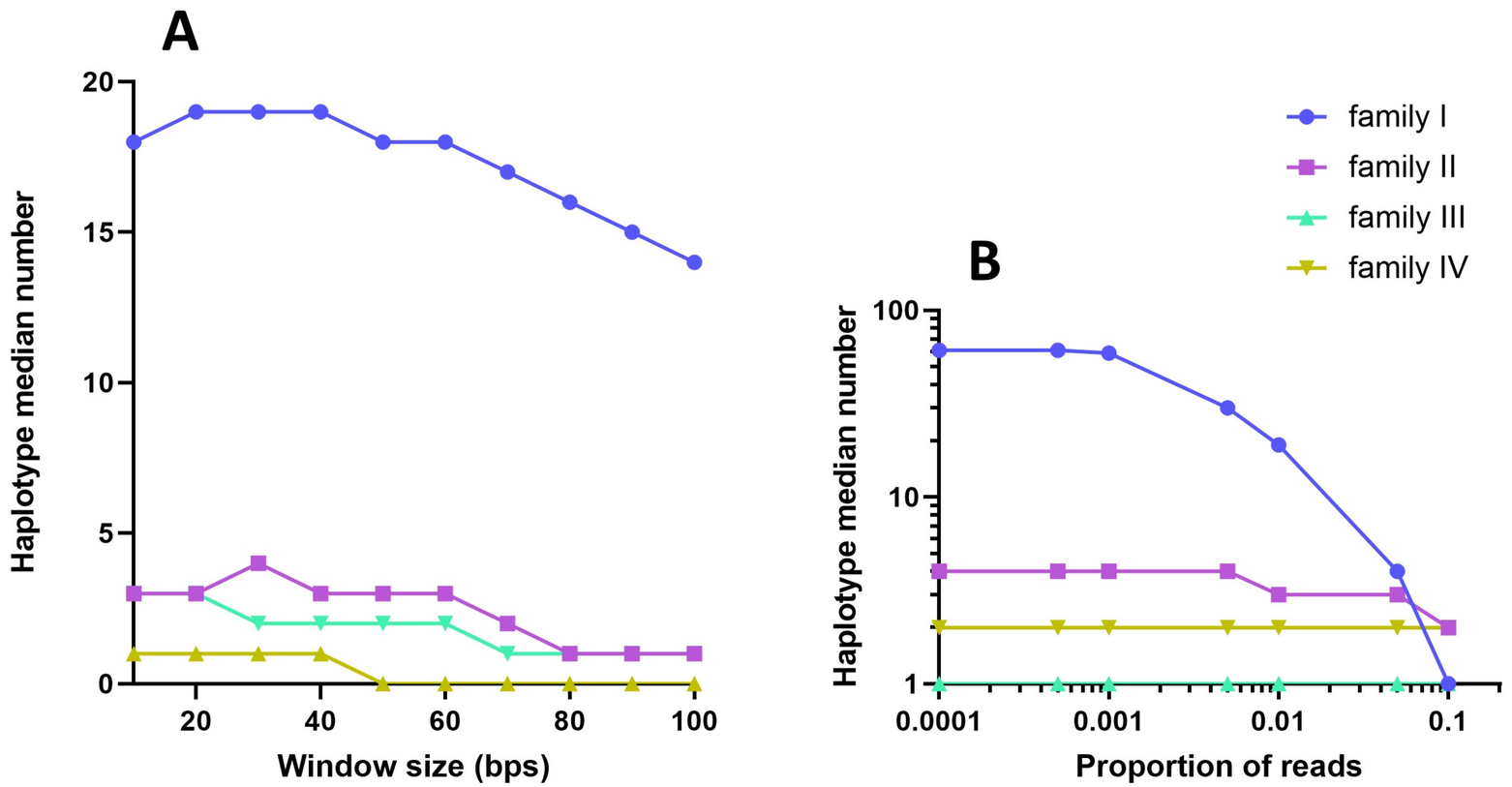
A
Haplotype median number
20
15
10
5
0
20
40
60
80
100
Window size (bps)
B
Haplotype median number
100
10
1
0.0001
0.001
0.01
0.1
Proportion of reads
family I
family II
family III
family IV

Fig S1