Genomic analysis of 61 Chlamydia psittaci strains reveals extensive divergence associated with host preference

Konrad Sachse, Martin Hölzer, Fabien Vorimore, Lisa-Marie Barf, Carsten Sachse, Karine Laroucau, Manja Marz, Kevin Lamkiewicz

BMC Genomics

RESEARCH

Open Access

Genomic analysis of 61 *Chlamydia psittaci* strains reveals extensive divergence associated with host preference

Konrad Sachse[1*], Martin Hölzer[2], Fabien Vorimore[3], Lisa-Marie Barf[1], Carsten Sachse[4,5,6], Karine Laroucau[7], Manja Marz[1] and Kevin Lamkiewicz[1,8]

Abstract

Background *Chlamydia (C.) psittaci*, the causative agent of avian chlamydiosis and human psittacosis, is a genetically heterogeneous species. Its broad host range includes parrots and many other birds, but occasionally also humans (via zoonotic transmission), ruminants, horses, swine and rodents. To assess whether there are genetic markers associated with host tropism we comparatively analyzed whole-genome sequences of 61 *C. psittaci* strains, 47 of which carrying a 7.6-kbp plasmid.

Results Following clean-up, reassembly and polishing of poorly assembled genomes from public databases, phylogenetic analyses using *C. psittaci* whole-genome sequence alignment revealed four major clades within this species. Clade 1 represents the most recent lineage comprising 40/61 strains and contains 9/10 of the psittacine strains, including type strain 6BC, and 10/13 of human isolates. Strains from different non-psittacine hosts clustered in Clades 2– 4. We found that clade membership correlates with typing schemes based on SNP types, *ompA* genotypes, multilocus sequence types as well as plasticity zone (PZ) structure and host preference. Genome analysis also revealed that i) sequence variation in the major outer membrane porin MOMP can result in 3D structural changes of immunogenic domains, ii) past host change of Clade 3 and 4 strains could be associated with loss of MAC/perforin in the PZ, rather than the large cytotoxin, iii) the distinct phylogeny of atypical strains (Clades 3 and 4) is also reflected in their repertoire of inclusion proteins (Inc family) and polymorphic membrane proteins (Pmps).

Conclusions Our study identified a number of genomic features that can be correlated with the phylogeny and host preference of *C. psittaci* strains. Our data show that intra-species genomic divergence is associated with past host change and includes deletions in the plasticity zone, structural variations in immunogenic domains and distinct repertoires of virulence factors.

Keywords *Chlamydia psittaci*, Genome analysis, Phylogeny, Host preference, Plasticity zone, Polymorphic membrane proteins, Inclusion proteins, Plasmid

*Correspondence:
Konrad Sachse
konrad.sachse@uni-jena.de
Full list of author information is available at the end of the article

Background

Chlamydia (C.) psittaci is known as the etiological agent of avian chlamydiosis and human psittacosis [1]. Like other chlamydiae, *C. psittaci* can also cause asymptomatic infections. Due to its capability of causing systemic infection with acute to chronic course in poultry, pet birds and some mammals, as well as its worldwide dissemination [2], it is probably the most important veterinary chlamydial pathogen. Besides, the importance of *C. psittaci* as a human pathogen is often underestimated. Although the big outbreaks of "parrot fever" following large shipments of exotic birds from South America to Europe and North America in the period from 1892 to 1929 are now history, the agent still deserves permanent attention. The zoonotic potential of *C. psittaci* is well documented in the literature [3–5]. Typically, individuals with previous contact to birds are affected, but fulminant manifestations in humans usually occur only when efficacious antimicrobials are not administered in time. The course of the human disease ranges from asymptomatic to flu-like to severe systemic illness, with the latter manifesting as pneumonia, myocarditis, encephalitis or sepsis. Mild symptoms are seen in most individuals infected, but immunocompromised persons are more likely to develop clinical signs. Occasionally, also apparently healthy individuals can be severely affected [6, 7]. In the last two decades, cases of zoonotic transmission were reported from psittacine birds [8], as well as ducks [9, 10], turkeys [11] and mixed domestic poultry [7] as the main sources. In addition, human-to-human transmission was shown to be a relevant infection route in a number of cases [12–16].

Like all chlamydial organisms, *C. psittaci* is an obligate intracellular bacterium distinguished by a biphasic developmental cycle comprising extracellular and intracellular stages. In the course of evolution, the genomes of all *Chlamydia* spp. have undergone vast condensation, which was shown to have resulted from genome streamlining rather than degradation [17]. The relatively small genome size of approximately 1 Mbp implies the absence of essential cellular pathways and, consequently, reliance on host cells for nutrients, such as amino acids, nucleotides and lipids [18, 19]. Chlamydiae are assumed to compensate for this deficiency by co-opting suitable cellular pathways [20, 21].

As handling of *Chlamydia* spp. using cell culture requires special expertise and their genetic manipulation is much more difficult than of most other bacteria, analysis of whole-genome sequences can be a viable alternative to characterize strains of interest and provide clues to understand pathogenic properties.

A large number of genome assemblies of varying quality from all *Chlamydia* spp. were published in the last decade. Given the steady advance of sequencing technologies and bioinformatics tools, it is not surprising that these genome assemblies differ significantly in quality, e.g. in scaffold numbers between 1 and 851 and N50 values between 1 Mbp (genome size) and 725 bp. In addition, the quality of gene annotations depends on the quality of the underlying assembly, the used annotation approach, software versions and parameters, as well as the reference database, which still represents a bioinformatics bottleneck even when high-quality genome assemblies are available [22]. These deficiencies in genome quality and annotation status can seriously hamper comparative studies.

Several genomic studies dealt with the comparison of whole-genome sequences among the major *Chlamydia* species [23–25]. Among individual species, genomes of the human pathogen *C. trachomatis* were most frequently analyzed to address genetic diversity, phylogeny and tissue tropism [26–31]. A few more studies dealt with *C. pneumoniae* [32], *C. pecorum* [33], as well as the zoonotic agent *C. abortus* [34]. Concerning *C. psittaci*, Read et al. analyzed the genomes of 20 strains from different hosts to suggest events of host switching and recombination along the timeline of phylogenetic evolution [35]. There are also reports dealing with SNP- and MLST-based phylogeny [36], and analysis of human strains from Australia [37].

In view of the considerable variation in terms of host preference, growth characteristics and pathogenicity observed among *C. psittaci* strains, it seems necessary to study a larger number of field strains to obtain data on intra-species genetic variation at genome level. In a recent comparative study comprising 33 strains of 12 different *Chlamydia* spp., including 10 strains of *C. psittaci*, we distinguished genomic features characteristic for *C. psittaci*, i.e. (i) a relatively short plasticity zone (PZ), (ii) an Inc protein set comprising IncA, B, C, V, X, Y, (iii) the largest chlamydial SinC protein sized 502 amino acids, and (iv) an elevated number of subtype G Pmp proteins ($n = 14$) [24].

In the present paper, we report the findings of a comparative analysis of 61 *C. psittaci* genomes that were deposited in public databases and met our quality requirements. For the latter, we set the maximum number of scaffolds tolerated per genome to 50 with an N50 value of at least 100,000 bp, and required at least one-third of the annotated ORFs to be complete, meaning an alignment length ratio of 1.0 to UniProtKB homologs. Our study aimed at elucidating the extent of intra-species genomic divergence and searching for possible correlations between genomic and phenotypic parameters.

Results

Improvement of genome assemblies and annotation

By February 1, 2021, 71 genome assemblies of *C. psittaci* strains had been uploaded to the NCBI and ENA databases. To ascertain data consistency and comparability, we removed any duplicates and finally included only whole-genome sequences fulfilling the quality criteria stated below. We re-assembled raw sequencing data available for 38 *C. psittaci* strains to achieve a better genome and annotation quality. After all cleaning steps, 11 re-assemblies achieved a better quality and were used instead of the original NCBI genomes in our final genome collection (Supplemental Table S1). We set the maximum number of scaffolds to be tolerated to 50 to ensure consistency among the finally used assemblies, e.g., for detection of the PZ. In addition, we checked assembly metrics such as the N50 value and the number of potentially fragmented ORFs (using IDEEL plots as described in [38]) to finally include genome assemblies of 61 strains in this study. Among them, 37 had completely assembled genome sequences (one sequence), 20 consisted of 2–8 scaffolds and another four genomes had 16, 19, 28 or 44 scaffolds, respectively (Table S1). This is a considerable improvement compared to the assembly states appearing in the respective NCBI and ENA entries. The reassembled genomes are deposited in the OSF repository: https://osf.io/rbca9/.

In Supplemental Figure S1, we compare all re-assembled and original genomes regarding their genome contiguity and N50 values, while the IDEEL plots in the OSF repository show the improvement regarding fragmented ORFs. Moreover, re-annotation of all genomes using recent software versions and reference databases helped to reduce the proportion of non-annotated genome features designated hypothetical proteins to 25% (average 251 of 994 CDS, see Table S1).

Genome size and core genome

The average genome size of the 61 *C. psittaci* strains was 1,166,132 bp. Strains 99DC5 and WS-RT-E30 were found to harbor the largest (1,175,249 bp) and smallest (1,140,789 bp) genomes of the present panel, respectively. Genome sizes of all strains are provided in Table S1. While the pan-genome was composed of 1 126 CDS, the core genome of this strain panel was calculated to be 904 common CDS (using `RIBAP`) [39], which is 90.9% of the average 994 CDS detected. The complete output of `RIBAP` is available at the OSF repository: https://osf.io/rbca9/.

Phylogenetic analysis

To explore the phylogenetic relationship among the strains, we reconstructed a tree based on the alignment of whole-genome sequences (Fig. 1A). Our data shows that the species comprises four major clades. The largest clade, consisting of 40/61 strains, includes the type strain 6BC and will be referred to as Clade 1. All psittacine isolates are on this clade alongside some others from humans, cattle, sheep and horses. Genome sequences of Clade 1 strains tend to be highly similar despite the presence of up to six major recombination sites (Fig. 1D/E, sites are marked grey). Since these strains belong to sequence types that are most common among currently known isolates we regard them as being "typical" *C. psittaci* strains.

Clade 2 contains only four strains, 99DC5, Ful127, Mat116 and WC. Compared to Clade 1 strains, they carry clade-specific SNPs and MLST-relevant sequences that have been rarely encountered so far. Each of these genomes harbors an unusually high number of unique sequences due to recombination events (blue boxes in Fig. 1E).

Clade 3 has seven strains isolated from non-psittacine hosts, such as duck, sheep, cattle and human: (06–1683, 08_2626_L3, AMK, C1/97, GR9, WS-RT-E30, and Rostinovo-70). Their typing parameters are largely atypical, i.e. PZ type 2, SNP type II, *ompA* types EB or C and MLST 28 (Table S1, for SNP typing see [40]).

Clade 4 is formed by eight strains from non-psittacine hosts, mainly pigeons (MN, 09DC77, 09DC78, 09DC79, 09DC80, Frances, CP3, and 01DC12).

Finally, strains M56 from a muskrat and NJ1 from a turkey are encountered outside the clades, which is consistent with their atypical classification as SNP type IV, ST31 and 43, as well as *ompA* genotypes M56 and D, respectively.

We additionally constructed alternative phylogenetic trees based on (a) the 904 core genes that emerged from RIBAP, calculated using FastTree (Figure S2), (b) the SNP analysis calculated using RAxML (Figure S3) (c) the extracted PZ of 53 strains (Figure S4), and (d).multiple sequence alignment of all 61 OmpA proteins (Figure S5). The assignment of strains to a particular clade in our phylogenetic tree is largely concordant with the classification schemes derived from SNP typing and MLST analysis, as well as *ompA* genotyping and the newly introduced PZ types (see Table S1). These relationships will be discussed in more detail below.

Comparative analysis of the plasticity zone

The PZ was defined as the segment flanked by genes *accB* (5') and *guaB* (3'). A non-fragmented PZ could be extracted from genome sequences of 53 strains. The zone varied in size from 22,534 nt (WS-RT-E30) to 30,180 nt (C6/98). In the genome of strain 6BC, the PZ is located

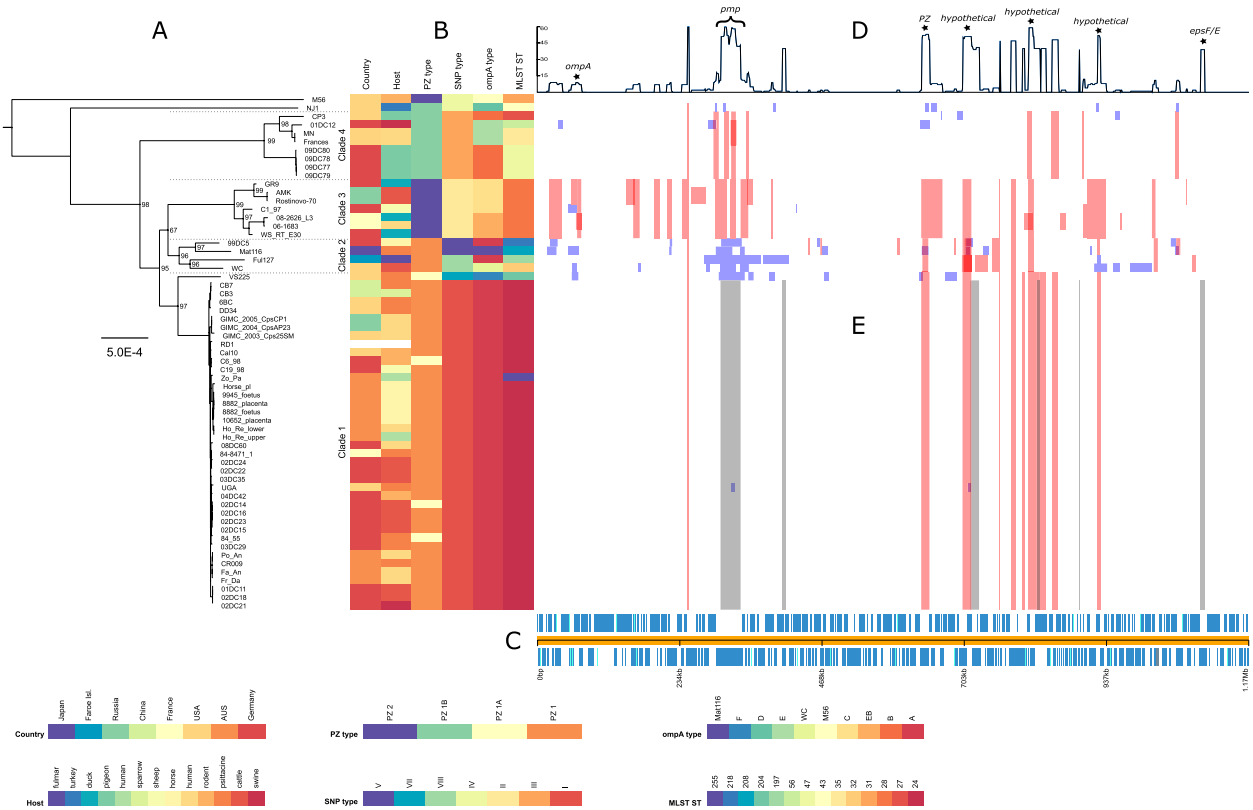

Fig. 1 Global phylogeny based on whole-genome alignment and recombination landscape of 61 *C. psittaci* genomes. The phylogeny (**A**), with associated metadata (**B**) is displayed alongside the linearized chromosome (**C**). Line graph (**D**) shows the number of recombination events affecting individual genes. Colored blocks (**E**) indicate inferred recombination events with blue blocks unique to a single isolate and red blocks shared by multiple strains through common descent. Gray blocks represent major recombination sites. Gene annotations are based on type strain 6BC (CP002549.1). The scale bar of the phylogenetic tree corresponds to 5×10^{-4} substitutions per nucleotide site and bootstrap values indicate stability of the branches based on 1,000 replicates

Table 1 Major open reading frames in the plasticity zone of *C. psittaci* strain 6BC

No	Name	Start	End	Length [nt]	Direction	Comment
1	*accB*	1	501	501	forward	Biotin carboxyl carrier protein of acetyl-CoA carboxylase
2	*accC*	501	1856	1356	forward	Biotin carboxylase (also Biotin acetyl-CoA carboxylase subunit)
3	DUF648 domain-containing protein_1	2008	3438	1431	forward	Domain of Unknown Function
4	DUF648 domain-containing protein_2	3558	5132	1575	forward	Domain of Unknown Function
5	Hypoth. protein ORF 141	5340	6296	957	reverse	
6	Hypoth. protein ORF 137	6664	7605	942	reverse	
7	MAC/perforin family protein	7794	8477	684	reverse	possibly includes phospholipase D domain
8	DUF1389 domain-containing protein	8853	9950	1098	reverse	Domain of Unknown Function
9	putative membrane protein	9880	10,398	519	reverse	
10	*toxB* (LifA/Efa1-related large cytotoxin)	10,434	20,507	10,074	reverse	also: Lymphostatin (in *E.coli*), Cysteine protease, YopT-type domain protein, TcdB toxin N-terminal helical domain protein
11	Hypoth. protein ORF 67	20,788	21,549	762	forward	
12	MAC/perforin	21,962	24,430	2469	forward	Membrane-attack complex/perforin (MACPF) superfamily protein
13	Hypoth. protein G50_0603	24,539	25,060	522	reverse	
14	*ADA*	25,124	26,524	1401	reverse	Adenosine/AMP deaminase
15	*guaA*	26,515	28,053	1539	reverse	GMP synthase
16	*guaB*	28,068	29,144	1077	reverse	Inosine-5′-monophosphate dehydrogenase

between positions 624,296 and 653,440. The major ORFs are compiled in Table 1.

Comparison of PZ structures based on multiple alignment of 53 translated sequences revealed four different types, which we define as types 1, 1A, 1B and 2, as depicted in Fig. 2 (sequence similarity values in Table S2). In PZ type 1, which was encountered in 32 strains, the complete set of 16 ORFs was identified. The main elements include the 5'-terminal biotin modification operon (*accB, accC*), the large cytotoxin (*toxB*) and MAC/perforin in the central region, as well as the purine synthesis and recycling operon (*ADA, guaA, guaB*) at the 3' terminus. PZ type 1A is distinguished by a fragmented or disrupted cytotoxin, e.g. in strain VS225 with four smaller proteins instead of one large molecule, also in strains C6/98, 02DC14, and 84–55 (three fragments). Type 1B can be recognized by the fragmented MAC/perforin (strains MN, 09DC77, 09DC78, 09DC79, 09DC80, Frances, CP3, 01DC12, NJ1). These three PZ types share a highly homologous structure with overall nucleotide sequence identity values above 95%.

PZ type 2, which was encountered in all seven strains of Clade 3 (06–1683, 08-2626_L3, AMK, C1/97, GR9, Rostinovo-70, WS-RT-E30) and M56, represents a reduced PZ version and is characterized by the absence of six ORFs at the 3'-end (nos. 11–16 in Table 1 and Fig. 2) including MAC/perforin and the purine synthesis operon. The PZ of this type is about 20% shorter than in the other strains.

In eight strains, i.e. Fa_An, Fr_Da, CB3, CB7, 9945_foetus, 8882_placenta, 8882_foetus, and 10652_placenta, PZ elements were located on separate contigs, thus precluding extraction of a contiguous PZ. As these strains could not be included in the multiple sequence alignment, the PZ sequence of strain 6BC was mapped to each of these genome sequences. As a result, the presence of all major CDS, including the terminal *accB* and *guaB*, the large cytotoxin and MAC/perforin genes were confirmed, which indicates a type 1 PZ in these strains.

The family of polymorphic membrane proteins (Pmps)

To explore the spectrum of Pmp family members present in the 61 strains, we conducted multiple protein blast analyses using the known Pmp sequences of strain 6BC as queries. The results confirm that all 21 family members are present in all 61 genomes (Table S3). While strains of Clade 1 were found to carry Pmps that are often identical and generally highly similar to their equivalent in strain 6BC, sequence similarity with the rest of the strains was markedly lower. The data in Table 2, which shows representative strains, reveal that the lowest identity percentages were seen in Clade 3 and the outlying strain M56. In addition, some of the Pmps of Clade 4 strains tended to align only partially to the 6BC equivalents, which could mean that they are shorter or contain distinct sequence elements.

Regarding individual Pmps, there is considerable variation among strains in some genomic loci, notably *pmp12-15* and *pmp17-21*, all of which belong to subtype G/I. Pmp15 and Pmp17 are the most variable family members with the average amino acid (aa) identity to the 6BC equivalents being only 75.90%. In contrast, Pmps1-11 as well as 16 and 22 showed less sequence variation with average aa identity values above 95% (Table S4). While amino acid identity values among Pmp2 sequences ranged from 96.7 to 100% (compared to the 6BC reference), Pmps17 varied from 40.9. to 100% among all strains.

The Inc protein family

At least 11 Inc proteins were identified by the annotation pipeline. Nine of them represent different members of the IncA family, which have been arbitrarily designated IncA family protein 1–9. Five of them were found in all 61 strains, while three others were absent in the seven strains of Clade 3. Major inter-strain sequence variation was observed in three proteins, IncA family proteins 3 and 5 as well as IncV, again with Clade 3 strains standing out. The main results are given in Table 3.

Variation of OmpA/MOMP sequences in *C. psittaci* strains

Alignment of extracted major outer membrane porin (MOMP) sequences (RIBAP group 879) confirmed that all 61 strains are equipped with this outer membrane porin. While the complete protein of 402 aa was seen in the vast majority of strains, a group of seven strains harbored slightly shortened versions: GR9, C1/97, AMK, Rostinovo-70 (all from Clade 3), VS225 (1), WC (2) and NJ1 (outside). MOMP molecules in that group have lost

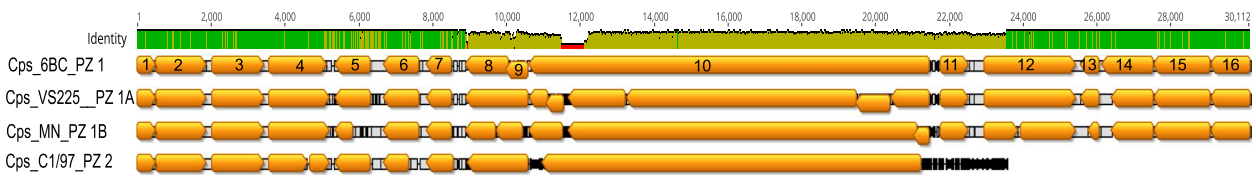

Fig. 2 Alignment of the plasticity zones of strains 6BC, VS225, MN and C1/97 representing PZ types 1, 1A, 1B, and 2, respectively. Numbering of ORFs as in Table 1

Table 2 Amino acid sequence divergence of Pmp repertoires among *C. psittaci* strains from different phylogenetic clades

Strain	**02DC15** (clade 1)		**Mat116** (clade 2)		**C1/97** (clade 3)		**CP3** (clade 4)		**NJ1** (outside)		**M56** (outside)	
Pmp [no._subtype]	% ident	a/q	% ident	a/q	% ident	a/q	% ident	a/q	% ident	a/q	% ident	a/q
Pmp1_B	100.00	1.00	98.72	1.00	99.29	1.00	97.77	1.00	94.56	1.00	87.42	1.00
Pmp2_A	100.00	1.00	99.79	1.00	99.89	1.00	99.68	1.00	99.57	1.00	96.68	1.00
Pmp3_E/F	100.00	1.00	93.41	1.00	86.50	1.00	86.10	1.00	95.10	1.00	70.69	1.00
Pmp4_E/F	99.90	1.00	82.12	*0.66*	94.49	1.00	96.88	1.00	94.07	1.00	72.33	1.00
Pmp5_E/F	100.00	1.00	100.00	1.00	99.44	1.00	99.15	1.00	96.61	1.00	65.73	1.01
Pmp6_H	100.00	1.00	94.01	*0.75*	94.92	1.00	94.82	1.00	93.60	1.00	89.95	1.00
Pmp7_G/I	100.00	1.00	92.56	1.01	93.38	1.00	91.47	1.00	91.66	1.00	86.39	1.00
Pmp8_G/I	100.00	*0.80*	99.74	0.90	99.65	1.00	97.71	*0.46*	95.77	1.00	84.43	*0.81*
Pmp9_G/I	100.00	*0.75*	89.10	*0.46*	93.02	1.00	97.82	*0.34*	79.59	*0.77*	68.05	*0.48*
Pmp10_G/I	100.00	1.00	99.64	1.00	99.26	*0.64*	98.89	*0.64*	97.14	1.00	91.91	1.00
Pmp11_G/I	100.00	1.00	99.53	0.99	99.29	1.00	99.41	1.00	96.82	1.00	90.58	1.00
Pmp12_G/I	100.00	1.00	84.58	1.00	77.54	1.00	84.68	*0.67*	85.71	1.00	81.62	1.00
Pmp13_G/I	99.88	1.00	78.05	1.01	74.76	1.00	77.25	*0.67*	77.62	1.01	84.43	1.01
Pmp14_G/I	100.00	1.00	79.16	1.01	77.27	1.00	81.16	*0.67*	79.93	1.01	80.64	1.00
Pmp15_G/I	83.43	1.00	93.92	1.00	77.90	1.00	96.12	*0.66*	86.45	1.00	80.12	1.00
Pmp16_G/I	100.00	1.00	99.78	1.00	99.57	1.00	98.71	1.01	98.38	1.00	90.82	1.00
Pmp17_G/I	83.43	1.00	93.92	1.00	77.90	1.00	96.12	*0.66*	86.45	1.00	80.12	1.00
Pmp19_G/I	77.66	1.02	77.00	1.02	99.40	1.00	80.42	*0.68*	80.12	1.01	80.83	1.01
Pmp20_G/I	99.88	1.00	81.34	1.00	77.09	1.00	83.63	*0.67*	82.63	1.00	80.02	1.00
Pmp21_G/I	84.68	1.01	86.64	1.01	79.34	1.00	84.30	*0.67*	90.12	1.00	88.92	1.00
Pmp22_D	100.00	1.00	99.28	1.00	87.57	1.00	99.48	*0.88*	97.85	1.00	99.35	1.00

% ident Percentage of identities shared with query sequence of *C. psittaci* strain 6BC

a/q Ratio of aligned target sequence length (blast hit) to query sequence length

Table 3 Presence of Inc proteins in *C. psittaci* strains and variation in atypical strains

Annotation	Size in 6BC [aa]	No. of strains where retrieved	Min. identity among typical strains [%]	Exceptions
IncA family protein 1 (Ribap group 358)	372	61	97.8	
IncA family protein 2 (group 514)	224	61	98.7	
IncA family protein 3 (group 890)	352	61	99.4	Clade 3 strains with low similarity (75% id.)
IncA family protein 4 (group 915)	461	35 (Clades 1 + 2)	99.3	Truncated in strains of Clades 3 + 4
IncA family protein 5 (group 939)	238	61	98.7	Clade 3 strains with low similarity (60% id.); truncated in Mat116, NJ1
IncA family protein 6 (group 698)	382	61	96.6	
IncA family protein 7 (group 886)	342	54	99.4	Missing in Clade 3 strains
IncA family protein 8 (group 931)	479	54	99.4	Missing in Clade 3 strains
IncA family protein 9 (group 923)	810	52 + 2*	99.4	Missing in Clade 3 strains (*also found in strains 8882_placenta and 8882_foetus via mapping)
IncB (group 404)	202	61	97.5	
IncV (group 898)	383	61	95.8	Clade 3 strains with low similarity (85% id.)

up to 11 aa, thus rendering sequence identity to the 6BC counterpart as low as 82.84%. The range of OmpA (or MOMP) sequence diversity among the present 61 strains is illustrated in a RAxML tree in Figure S2.

To better understand the consequences of the observed sequence variation in the framework of the OmpA porin 3D structure, we compared predicted 3D OmpA structures from two "antagonist" strains, 6BC (typical strain) and C1/97 (atypical strain), using `ColabFold` [41, 42]. Overall, OmpA/6BC is an all-β fold protein with an accessory N-terminal α-helix (aa positions 1–22) representing the signal peptide. The

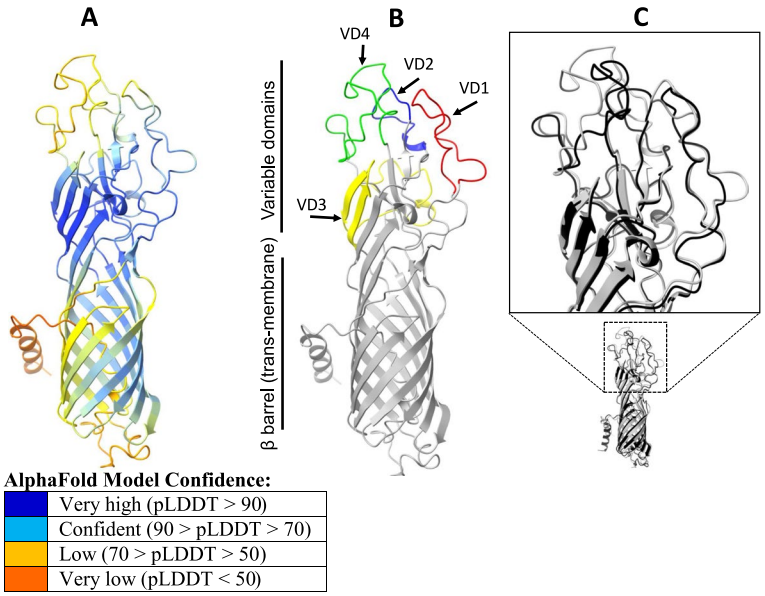

Fig. 3 Predicted 3D structure of OmpA from *C. psittaci* strain 6BC and comparison with strain C1/97. **A** Structure prediction of OmpA from strain 6BC using `AlphaFold`. Color code indicates prediction reliability based on per-residue confidence score (pLDDT). **B** Highlighted locations of variable domains (VD) 1 to 4 in strain 6BC. **C** Superposition of predicted OmpA structures in strains 6BC (gray) and C1/97 (black) with zoomed inset of the VD regions. The 3D structures were rendered using `ChimeraX`

β-barrel of the transmembrane porin fold resides in the bacterial membrane from which the extra-membrane region emanates (Fig. 3A). The core of this region is formed by a short three-stranded β-sheet (159–161,309–312, 352–356), which is surrounded by the solvent-exposed variable domains (VDs) 1, 2, 3 and 4 (Fig. 3B). In OmpA/6BC, the VD1 (83–107), VD2 (164–177) and VD4 (317–350) of the antigen contain three pronounced loop structures without clearly assigned secondary structure, whereas VD3 (228–265) comprises a two-stranded β-sheet. As expected, the superposition of the 6BC and C1/97 structures reveals high overall structural agreement, in particular in the transmembrane β-barrel domain as well as in VD3. Notable differences, however, occur in backbone conformations of the extra-membrane VDs 1, 2 and 4 (Fig. 3C). At the same time, VD1, VD2 and VD4 represent the least reliable regions of the `AlphaFold` predictions due to the lower overall pLDDT score, which is generally recognized as predictor for structural disorder and dynamic flexibility of the protein structure [43]. In pairwise alignment of 6BC and C1/97 OmpA sequences, the latter shows several deletions in VD1 (4-aa gap in pos. 100–103), VD2 (4 aa missing between 167 and 176) and VD4 (3-aa gap 342–344). In addition, VDs contain notable aa variations, i.e. 14 aa in VD1, 6 in VD2 and 12 in VD4. Remarkably, these strain sequence variations are limited to solvent-exposed regions of VD1, VD2 and VD4, but have no effect on the overall structural porin scaffold.

Other potentially important loci

To explore the sequence conservation of presumably non-variable genomic loci that are encoding potential virulence factors, we analyzed the multiple sequence alignments provided in the respective `RIBAP` groups. The findings are summarized in Table 4. None of the seven loci proved 100% identical in all strains. While the CADD and FtsW proteins were found to have only a few variable positions, there was more heterogeneity in the other loci, e.g. SinC, where the strains of Clade 3 harbored a protein variant of lower homology. It is also noteworthy that sequences of the histone-like protein pair HctA/B differed among a number of strains, and HctB was absent in four *C. psittaci* strains, a feature shared with some strains of *C. avium* and *C. gallinacea* [24].

PZ, SNP, ompA and MLST types vs. host preference

The origins and typing results of all 61 strains are given in Supplemental Tables S1 and S5. We used Fisher's exact test to detect possible association between the various genotypes and the animal host (see sec. 4 Methods). We detected a significant association of the pigeon host with PZ type 1B, SNP type III and *ompA* genotype

Table 4 Presence of potential virulence factors in *C. psittaci* strains and variation in atypical strains

Gene product	Size in 6BC [aa]	No. of strains where retrieved	Min. identity among typical strains [%]	Exceptions
CADD: CADD family putative folate metabolism protein (Ribap group 207)	245	61	99	
CPAF:Protease-like activity factor (group 684)	605	61	95	M56 only 91.9% id. to other strains
FtsW: Putative lipid II flippase (group 619)	384	61	99	
HctA: Histone H1-like protein Hc1 (group 702)	123	61	99	117-aa variant in Clade 4 members plus Ful127; M56 only 95.1% id
HctB: Histone H1-like DNA-binding protein Hc2 (group 867)	197	57	99	154-aa variant (78.2% id.) in C19/98, C6/98; missing in CP3, M56, NJ1, VS225
SinC: Secreted inner nuclear membrane-associated Chlamydia protein	502	61	99	503-aa variant (86.5% id.) in Clade 3 members
TarP: Type III secretion system actin-recruiting effector (group 742)	874	61	99	M56 only 88.6–89.1% id., truncated in Mat116 (826 aa)

B ($n=5$, p-value 1.6e-7), as well as of the duck host with PZ type 2, SNP type II and *ompA* genotype EB ($n=2$, p-value < 0.005). PZ type 1, which is associated with Clade 1, did not present an association with any particular host, but was associated with the supergroup of human, psittacine and cattle hosts ($n=22$, p-value < 0.005).

Plasmid analysis

A plasmid sized between 7487 bp (strain 84/55) and 7677 bp (strain MN) was identified in 47 of the 61 strains examined. Basic characteristics are given in Table S5. A phylogenetic tree based on the alignment of all sequences was reconstructed and is shown in Figure S6. The plasmid tree visibly differs from the whole-genome tree, for instance Clade 1 strains appear to be less homogeneous. On the other hand, there are a few common features, such as the joint clustering of strains belonging to Clades 3 and Clade 4, respectively, and the outlying positions of strains M56 and NJ1.

Nevertheless, the lowest nucleotide sequence similarity value among all plasmids is still as high as 96.48% (M56 vs. MN, see Table S6). This confirms investigations from other *Chlamydia* spp. showing the high level of conservation among plasmids of a species [44]. Except for Ful127, all strains were found to contain eight CDS, which were annotated as virulence plasmid proteins pGP2-D, pGP3-D, pGP4-D, and pGP6-D, ParA family protein pGP5-D, integrases pGP7-D and pGP8-D, as well as putative plasmid replicative DNA helicase.

Discussion

Creating a comparable sequence dataset

To ensure accurate annotation and correct genome analysis our initial efforts we initially improved the assembly status of those whole-genome sequences that were poorly assembled or still in draft state. The limitations posed by incompletely assembled genomes are well known [45, 46], particularly their negative impact on annotation quality [22]. Using state-of-the-art algorithms we decontaminated and re-assembled the raw reads of those chlamydial genomes that were of insufficient quality. The resulting harmonized genome set was validated using QUAST, Icarus, and IDEEL plots (see Figure S1 and the OSF repository). Subsequent annotation of this dataset yielded better comparability among the genomes included in this study. To have a closer look at the divergence of specific loci in *C. psittaci*, a refined core gene set was calculated, which also considered orthologous genes of low sequence similarity (≥ 60%).

Genomic regions of lower synteny

The PZ of *C. psittaci* is of smaller size than that of *C. suis* (82,505 nt), but larger than in *C. avium* (5,694 nt) [24], and in contrast to *C. caviae*, *C. felis* and *C. pecorum*, the tryptophan biosynthesis operon is missing [47, 48]. Although variations in size and contents are more extensive among chlamydial species the diversity within *C. psittaci* is still considerable, thus justifying the introduction of four different structural types, i.e. 1, 1A, 1B, and 2. Among the strains examined here, C6/98 had the largest (30,180 nt) and WS-RT-E30 the shortest (22,534 nt) PZ.

The most prominent locus in the PZ is the *toxB* gene encoding the large cytotoxin. The fact that its annotation is still a challenge for most algorithms may be due to its large size (3,358 aa in 6BC, 3,159 aa in C1/97) and/or multifunctional domain content. The *toxB* gene product is an ortholog of lymphostatin/EFA-1, a toxin known from *E. coli* (EPEC and EHEC) and also *Citrobacter rodentium*. This protein carries three enzymatic activities attributed to motifs of glycosyltransferase

(D-X-D), protease (C, H, D) and aminotransferase (TMGKALSASA) motifs [49]. Therefore, the molecule is often annotated as LifA/Efa1-related large cytotoxin [50] or cysteine protease YopT-type domain protein [51]. Additionally, based on homology to *Clostridioides difficile*, the 150-aa segment near the N-terminus carrying glycosyltransferase activity is designated TcdB toxin N-terminal helical domain protein in the U. database [52].

While these features render the large cytotoxin a straightforward candidate for virulence factor, experimental evidence in a chlamydial context has yet to be obtained. Reports showing clostridial cytotoxins being capable of Ras superfamily inactivation [53, 54] and host cytoskeleton disassembly [51, 55] could be an incentive to pursue this path. In the present study, we observed truncated versions of the *toxB* gene in strains VS225, C6/98, 02DC14, and 84/55 classified as PZ type 1A. This is the first report of truncated cytotoxin genes in *C. psittaci* strains. The only similar observation from other chlamydiae is from *C. gallinacea*, where a premature STOP codon in the *toxB* gene of one strain was observed [56]. Notably, disruption of this gene in certain *C. psittaci* strains did not coincide with different host tropism or other obvious distinctions compared to strains harboring the full-size *toxB*. This is contrasting our findings on strains carrying genes encoding another PZ protein that contains a membrane attack complex/perforin (MACPF) domain. These immune effectors are part of eukaryotic defense mechanisms and can induce cell killing through targeting microbial or host membranes. In the course of co-evolution with the host, chlamydiae have acquired their own MACPF-domain protein. It is assumed that this has enabled chlamydial organisms to partially resist MACPF effector mechanisms from the host and to facilitate their own infection [57]. The orthologous MACPF-domain protein of *C. trachomatis* was suggested to be able to permeabilize the inclusion membrane [58]. We found that all genomes having a type 1 or 1A PZ harbored an ORF encoding MACPF, which was sized 2,469 nt in strain 6BC (ORF12 in Fig. 2). The translated protein of 822 aa has a phospholipase D domain in its N-terminal region, however without homologs outside the *Chlamydia* spp. A recent study also showed differential expression of the gene in cell culture [59]. In addition, another ORF in the PZ was annotated as MAC/perforin family protein (ORF8 in Fig. 2) in a number of strains. Our finding that presence or absence of the intact MACPF gene coincides with host tropism underlines the importance of this factor and will be discussed below.

The data obtained in this study raises important questions on the functionality of MACPF as a potential virulence factor, as well as the annotation and identity of the ORF8 product, all of which should be addressed in future laboratory studies.

The Pmp family consists of autotransporters with surface-exposed and membrane-translocated domains. Members of this highly variable and complex protein family are regarded as virulence factors [60] and/or adhesins and immune modulators [61]. In the present study, Pmps were annotated as autotransporter domain-containing proteins. It was known from previous genome studies that strain 6BC possessed 21 different Pmps [24, 60], and the study by Wolff et al. [62] provided some insight into *pmp* locus divergence among 12 *C. psittaci* strains. These authors found that subtype G Pmps had the highest degree of divergence in *C. psittaci* genomes. The same observation was made later in an analysis of *C. abortus, C. avium, C. caviae, C. felis, C. gallinacea,* and *C. pecorum* genomes [24].

The present data indicate that eight of 14 subtype G/I Pmps are subject to considerable strain-to-strain sequence divergence, i.e. Pmps 12–15, 17 and 19–21 (Table 2). As can be seen from Figure S7, these variable loci are arranged in two clusters located in separate genomic regions, from pos. 319,411 to 335,394 and 707,548 to 715,415, respectively. In contrast, the most conserved family members, Pmps 1, 2 and 22 (subtypes B, A and D), are located outside the two variable genome clusters. Interestingly, Pmp17, the most variable representative, was suggested to be a key player in host adaptation [63]. Given the extent of strain-to-strain sequence variation observed with some of the G/I subtype Pmps it is important to note that we found representatives of all 21 family members in all 61 strains. In addition to the complete set of 21 Pmps, our multiple protein blast analysis identified a large number of possibly truncated low-similarity hits in all strains. The significance of these Pmp-like elements is unknown and will require future studies.

The family of inclusion membrane (Inc) proteins utilizes approximately 4% of the coding capacity in chlamydial genomes [64] and is rather heterogeneous in terms of sequence similarity, which represents a real challenge to annotation tools. At the same time, Incs share a common structural feature, i.e. they are inserted in the inclusion membrane via type III secretion. If exposed to the cytosol, some of them are among the major immunogens of *Chlamydia* spp. [25, 65]. Therefore, strain-to-strain differences in Inc protein sequences could result in different pathogenic properties and host tropism.

In analogy to the nomenclature in *C. trachomatis*, the presence of six Inc protein family subtypes was suggested in *C. psittaci*, IncA, B, C, V, X, and Y, the latter three only provisionally assigned [24]. Our analysis revealed 11 individual Incs, which is probably only the tip of the iceberg,

since up to 59 family members have been predicted for *C. trachomatis* and 92 for *C. pneumoniae* [64]. The main characteristics of the items identified here are i) highly conserved sequences of all identified Incs among the typical strains (Clades 1, 2 and partly 4), ii) low-similarity Inc variants in four strains, all belonging to Clade 3, and iii) three individual Incs absent in Clade 3 strains. Our finding that strains of Clade 3 carry a markedly distinct set of Inc proteins is remarkable in the light of a paper on *C. trachomatis*, where Lutter et al. suggested that some of the more divergent Incs were associated with clinical groupings (LGV, ocular and urogenital) and could contribute to tissue tropism [66].

Outer membrane porin (OmpA/MOMP)

The *ompA* gene locus is one of the genomic sites with the highest recombination rate in *Chlamydia* spp. [26]. Structural variation in the OmpA antigen, also called major outer membrane protein, was the basis of *C. psittaci* serotyping introduced in the 1990s [67]. Later on, the corresponding *ompA* genotypes were defined and became a more practicable equivalent to serotypes [68]. Previous attempts of correlating *ompA* genotypes with host preference revealed tendencies, but remained tentative [4]. Due to high sequence variation among *C. psittaci* strains (see Fig. 1B), the *ompA* locus is often missed as a core gene. However, based on our analysis, which also considers genes of lower sequence similarity, we were able to reconstruct a core genome including *ompA* (see `RIBAP` group 879 in the OSF supplement). The fact that the RAxML tree inferred from OmpA protein sequence alignment of all 61 strains (Figure S2) shows the same grouping of strains as in the trees based on genomes and PZ, respectively, indicates that this locus could also be used as a marker of host tropism, which is in accordance with recent studies on *C. trachomatis* [29, 69].

Since the well-known OmpA sequence variations among *C. psittaci* strains have not yet been addressed at the 3D structural level, we compared the protein structures of two antagonist strains, 6BC (Clade 1) and C1/97 (Clade 3). They belong to different *ompA* genotypes, A and C, and originate from different hosts, parrot and sheep, respectively. Our in silico analysis was facilitated by the use of `AlphaFold`, a new AI system predicting 3D protein structures with high accuracy [41, 70]. While the predicted structures exhibit the expected hallmarks of a porin, it also reveals strain-to-strain differences in the loops formed by VDs 1, 2 and 4. The presence of flexible VD1, VD2 and VD4 loops in OmpA and the strain-to-strain sequence variations are likely a result of adaptation due to interaction with complementary protein surfaces. The analyzed 3D structure and sequence properties of VD1, VD2 and VD4 provide a rationale to the required structural plasticity when the antigen is facing antibodies from different host immune responses. Thus we were able to show that 3D structural differences between OmpAs from strains of different origin are indeed detectable.

Although 3D structure models of *C. trachomatis* OmpA were previously obtained [71], this is the first study to visualize 3D OmpA structural variations between chlamydial strains using a state-of-the-art in silico machine learning approach. We are aware that it is still a singular finding, but the observation could be potentially important for the study of host–pathogen interaction. Verification of this intriguing speculation will require systematic investigation including wet-lab experiments.

Phylogeny and host tropism

Although the topic of host preference of *C. psittaci* was addressed in previous studies [24, 62], many questions have remained unanswered to date. It seems certain that there is no single genomic locus determining host predilection, but rather a panel of genes or gene products.

In the present study, we observed a remarkable similarity in the topology of different phylogenetic trees reconstructed from genomic data, i.e. whole genomes, core genes, SNPs, PZ and OmpA (Figs. 1A, S2, S3, S4, S5). With a few exceptions, membership of the major clades or lineages proved stable among these trees, thus indicating that all five datasets reflect phylogeny.

In contrast, analysis of plasmid sequences revealed some strain-to-strain relationships reminiscent of genome-based phylogeny, such as the clustering of Clades 3 and 4, but also marked differences among Clade 1 strains. In this context, the methodological limitation underlying phylogenetic tree reconstruction from plasmid sequences must be mentioned. As the sequences are short (~7 kb) and very similar, the tree derived from multiple sequence alignment will not be as robust as in the case of whole genomes, particularly at the leaf level. Apart from this, the limited conformity between plasmid and genome-based tree could indicate that the phylogeny of the extrachromosomal element included some, but not all, phases of the chromosome-derived phylogeny. While it is widely accepted that chlamydial plasmids and genomes have co-evolved [44, 72], this does not mean synchronized evolution and leaves the possibility of individual loci evolving independently. Jones et al. [73] were able to demonstrate that each of the eight CDS in the *C. trachomatis* plasmid was distinguished by its own specific SNP frequency, which implies a high potential for sequence variation in the course of evolution. Szabo et al., who included 16 strains of *C. psittaci* in their study (14 of which are also part of our study), defined three plasmid genotypes M, N and O of this species [72].

Regarding these 14 strains, our outliers M56 and NJ1 were assigned genotype M, whereas CP3 and WC were N, and 6BC, Cal10, DD34, RD1, 84/55, CB7, VS225, WS-RT-E30, Frances, and MN were genotype O. The results of our own comparative analysis (Figure S5) are in line with the plasmid genotype classification by Szabo et al., but future studies will show whether the introduction of more plasmid genotypes of *C. psittaci* is necessary.

As only whole-genome sequences deposited in public databases could be included in our study, the present choice of *C. psittaci* strains is arbitrary, so that our data cannot comprehensively reveal the geographical distribution and genetic divergence among all naturally occurring strains. Nevertheless, our phylogenetic analysis includes the largest number of *C. psittaci* strains examined so far and allows some interesting insight into the history of this species and the dynamics of its evolution. Based on analysis of 20 genomes, Read et al. [35] suggested a high evolution rate of 1.68×10^{-4} mutation/site/yr (175 SNPs per year) for the species of *C. psittaci*. Their findings would be consistent with the assumption that the most recent common ancestor can be dated to the era of the colonization of South America in the 16th to 18th centuries. However, Hadfield et al. argued that this evolution rate is probably an overestimation [26]. Likewise, Branley et al. [37] predicted a lower mutation rate (6.301×10^{-7}) and the emergence of the common ancestor of 6BC clade strains 2000 years ago.

While our study did not include tools calculating phylogenetic timelines the overall genetic homogeneity among Clade 1 strains could suggest more recent emergence, which would be consistent with reports of large psittacosis outbreaks in the decades before and after 1900.

The limited genetic divergence on this clade could also imply that the isolates from non-avian hosts were previously acquired from birds. While isolated from five different avian (four psittacine birds and one sparrow) and seven non-avian host species including humans, cattle and horses, it is remarkable that 90% of all psittacine strains are on Clade 1 (another one only on adjacent Clade 2). In analogy to the clade of LGV strains in *C. trachomatis* [26], members of Clade 1 include more virulent strains than the other clades.

The only psittacine strain on Clade 2, Mat116, is from Japan, which could explain its genome being distinct from Clade 1 because of the large geographical distance from most of the other strains.

Clade 3 represents the most genetically diverse lineage of the species. It seems to carry both mammalian and avian strains, however no psittacine strains. In view of the low number of members ($n=7$) it is not yet certain that psittacine hosts can be excluded, but if so, a switching event from psittacine to non-psittacine host could have initiated the development of this lineage. Notably, the three strains isolated from ducks are encountered on this clade. Ducks are the main *C. psittaci* host among domestic poultry, and these strains seem to be pathogenic to humans as cases of transmission were reported regularly [4, 74–76].

Clade 4, was probably the first to separate from the most recent common ancestor. Similar to members of Clade 3, these eight strains from non-psittacine hosts are distinguished by the loss of MACPF in the PZ, as well as aberrant repertoires of Incs and Pmps.

In contrast to the existing typing systems, the four-clade scheme presented here is based on whole-genome sequences rather than one or several genomic sites. This renders it a more robust and comprehensive tool for classifying chlamydial strains.

Conclusions

Overall, the present study provides novel insights into the genetic diversity within the species *C. psittaci*. Our data show that intra-species genomic divergence is associated with past host change and includes deletions in the plasticity zone, 3D structural variations in immunogenic domains of the outer membrane porin OmpA, and distinct repertoires of virulence factors, such as proteins of the Pmp and Inc families. Future experimental laboratory studies can investigate the phenotypic consequences of these divergent features to explain strain virulence and characteristic clinical courses. Therefore, the findings of this comparative genome analysis will improve our understanding of potential links between genomic features and phenotypic traits.

Methods

Chlamydial strains

We included *C. psittaci* strains whose genome sequences, including raw data, were available from the NCBI or ENA databases on February 1, 2022 and had less than 50 scaffolds after reassembly using our own procedure described below. Strain designations, host organisms and assembly states are given in Table S7. A more detailed overview with basic characteristics and database accession numbers of all 61 *C. psittaci* strains is given in Table S1.

Reassembly, clean-up and normalization of genome sequences

To reduce technical bias, the same tools and software versions were used to quality-control and de novo assemble all single and paired short-read data sets. For paired-end data sets, the respective input parameter settings were adjusted for each tool. First, reads were quality-trimmed

using fastp v0.20.1 [77] with parameters `-5 -3 -W 4 -M 20 -l 15 -x -n 5 -z 6`. Subsequently, de novo assembly was performed using SPAdes v3.14.1 [78] with parameters `--careful --cov-cutoff auto`. Afterwards, the initial assemblies were subjected to two additional polishing rounds using Pilon v1.23 [79]. We used BWA-MEM v0.7.17 [80] to map the quality-controlled short reads to each respective assembly as input for the polishing steps. Finally, we used Bandage v0.8.1 [81] to examine the assembly graphs and their contiguity.

We used the decontamination workflow Clean (https://github.com/hoelzer/clean v0.2.0) before and after reassembly to remove foreign sequences from the genomes in public databases, with sequences of *Homo sapiens, Chlorocebus sabeus, Gallus gallus, Macaca nemestrina, phiX* and *Chlamydia psittaci* plasmids being considered to be potential foreign contaminants. After reassembly, all strains with more than 50 scaffolds were discarded. We justify this cut-off based on our additional assessment of assembly quality. First, we used QUAST v5.2.0 [82] to calculate various metrics between the original and reassembled genomes and in particular visualized contig N50 and assembly contiguity using the built-in Icarus tool [83]. Secondly, to assess ORF quality and compare it in the original and re-assembled genomes, we ran a Snakemake pipeline3 implementing an approach called IDEEL as previously presented [38]. Briefly, annotated ORFs were translated into proteins and compared against a reference database to plot their completeness.

To facilitate genome comparison, whole-genome sequences of all finally included 61 strains were normalized using a custom python script (located as helper script in the RIBAP repository; https://github.com/hoelzer-lab/ribap/tree/master/bin/helper/rearrange), so that the *hemB* (delta-aminolevulinic acid dehydratase) gene appeared in the initial position.

Annotation and processing of genome sequences

All refined genome sequences were processed using the Roary ILP Bacterial Annotation Pipeline (`RIBAP`) v0.6.2 (https://github.com/hoelzer-lab/ribap), which was recently developed in our group ([24]. This pipeline performs annotation, core gene set calculation, alignments and phylogenetic reconstructions of homologous genes in `RIBAP` groups.

As an initial operation, annotation was conducted using Prokka v1.14.5 [84] with the `--proteins` option and the annotated GenBank file of *C. psittaci* 6BC (CP002549.1) as reference to ensure consistency in gene denominations. After calculating an initial core gene set with `Roary` v3.13.0 [85], `RIBAP` performs a less stringent all-versus-all MMSeqs2 v10.6d92c [86] approach to find potential homologous genes missed by `Roary`. Based on all pairwise comparisons, two `Roary` clusters are merged into one `RIBAP` group if the majority shows sufficient homology. Any gene from the core gene set has to be present in all input genomes. The output of a `RIBAP` run includes a searchable and interactive HTML table featuring the final `RIBAP` groups. This includes gene designation, gene description, a color-coded heat map based on `Roary` assignments at different similarity thresholds, as well as a phylogenetic tree based on multiple sequence alignment (MSA) of the CDS contained in the respective `RIBAP` group.

Whole-genome SNP analysis

Each assembly file was processed with the `ART-MountRainier` simulation tool, which generates synthetic paired-end reads with coverage of 50 [87]. These reads were aligned and mapped against the reference sequence of *C. psittaci* 6BC (CP002549.1) using the BWA algorithm implemented in `BioNumerics` v7.6.1 (bioMèrieux, Applied Maths, Sint-Martens-Latem, Belgium) with minimum sequence identity of 90%. SNPs were identified using the `BioNumerics` wgSNP module and then filtered using the following conditions: minimum 5× coverage to call a SNP, removal of positions with at least one ambiguous base, one unreliable base or non-informative SNP and minimum inter-SNP distance of 25 bp.

Phylogenetic and recombination analyses

A whole-genome alignment (WGA) was performed on the 61 *C. psittaci* genomes using `SKA` v1.0 (https://github.com/simonrharris/SKA). Prior to that alignment, fragmented genomes with more than one scaffold were mapped on the 6BC reference genome (NC_015470.1) using minimap2 v2. 17 [88]. Variants were called and a consensus sequence was reconstructed using samtools/bcftools v1.9 [89].

A phylogenetic tree was then reconstructed from the WGA with RAxML v8.2.12 [90] using the `GTRGAMMA` model and 1000 bootstrap replicates. The tree was rooted using midpoint.

Recombination analysis was performed from the same WGA using `Gubbins` v3.1.6 [91] with default settings. Figure 1 was generated using `Phandango` v1.3.0 [92].

The tree based on the 904 common genes of 61 *C. psittaci* strains (Figure S2) was constructed based on the core gene alignment produced by `RIBAP` and calculated using `FastTree` v2.1.10 [93].

A phylogenetic tree from SNP analysis was built using `RAxML` version 8.2.9 with the `GTRGAMMA` model and 1000 bootstrap replicates based on the filtered SNP matrix (4011 SNPs) from `BioNumerics`.

Multiple sequence alignments

Alignments of nucleotide and amino acid sequences were obtained using the `ClustalW` algorithm in `Geneious` 10.2.4 (Biomatters Ltd., Auckland, New Zealand). The `RAxML` tree based on OmpA sequences was constructed using protein model `GAMMA BLOSUM62` with the Rapid hill-climbing algorithm [90].

Extraction of the PZ

The plasticity zone was identified by conducting a `BLAST` v2.10.0 + search against the gene *accB* as the 5'-terminus of the PZ and the gene *guaB* as the 3'-end. In case of a truncated PZ the MAC/perforin gene was taken as the 3'-end. Subsequently, genome fragments located within these boundaries were extracted using `bedtools` v2.7.1 `getfasta`.

3D structures of OmpA

OmpA protein structures were obtained using `ColabFold` [42], a free platform for prediction of 3D structures from amino acid sequences based on `AlphaFold` v2.1.0 [41, 70].

Graphics were generated using the molecular visualization program `ChimeraX` v1.4.dev202201150102, which was developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco, with support from National Institutes of Health R01-GM129325 and the Office of Cyber Infrastructure and Computational Biology, National Institute of Allergy and Infectious Diseases [94].

Multiple blast to identify homologs of Pmp and SinC proteins

All amino acid sequences of Pmp family members of *C. psittaci* strain 6BC were compiled in a multi-FASTA file and blasted against the proteome sequences of all 61 strains. The resulting hits were filtered to obtain those with the best bitscore for each individual Pmp in each strain. Likewise, the amino acid sequence of SinC of *C. psittaci* Cal10 (EGF85279.1) was blasted against all 61 proteomes.

Statistical methods

We used Fisher's exact test to characterize possible relationships between isolation source/host and typing parameters of the *C. psittaci* strains. For each host and each genotype, we constructed a two-by-two contingency table counting the number of strains with this combination. We then tested for the enrichment of some hosts with a particular genotype.

Plasmid analysis

First, we selected all contigs annotated as plasmids from all originally downloaded FASTA files (NCBI GenBank). We checked that all contigs had a length of ~ 7 kb and discarded too fragmented sequences, resulting in 28 plasmids. Second, we built a Blast database from these 28 sequences to query our re-assemblies for additional plasmids that might not have been successfully assembled before or were discarded when the original assemblies were uploaded to the NCBI database. By that, we discovered 19 additional plasmid sequences for strains, where we did not find a plasmid sequence on NCBI, resulting in a total set of 47 plasmids for our 61 *C. psittaci* strains included in the study. However, during a first analysis and annotation round, we found four plasmids obtained from NCBI with a differing number of annotated CDS in comparison to the usually observed pattern in our *C. psittaci* plasmid collection. While for the majority of plasmids we found 8 CDSs and homology-based functionality with high sequence similarity to reference sequences, we observed discrepancies for the following plasmid sequences uploaded to NCBI: 1) strain MN: 10 CDSs, 1 pseudogene, and 2 hypothetical proteins; 2) strain Ho_Re_lower: 9 CDSs; 3) strain Ho_Re_upper: 9 CDSs and 1 pseudogene; 4) strain Ful127: 9 CDSs and 2 pseudogenes. As we suspected potential assembly errors in those plasmid sequences, we screened again for corresponding plasmids in our own re-assemblies of these strains. In the case of strain Ful127, we did not find any raw read data and did not reconstruct a new assembly. For the other three strains, we found plasmids in our re-assemblies. Our analysis showed that these three plasmids were in better agreement with all the other *C. psittaci* plasmids in our data set (based on annotation and alignment). Thus, we used the re-assembled plasmids to replace the original NCBI plasmids for strains MN, Ho_Re_lower, and Ho_Re_upper. More details (pairwise alignments between the NCBI and re-assembled plasmids, differences in the annotation) can be found in the OSF repository (https://osf.io/rbca9). Finally, our plasmid panel comprised 25 plasmids from the NCBI database and 22 plasmids from our re-assemblies. Three re-assembled plasmids (MN, Ho_Re_lower, Ho_Re_upper) replaced the NCBI plasmids due to better assembly quality and annotation agreement.

We used pLannotate v1.2.0 [95] to obtain gene annotation and confirmed the selected sequences as plasmids. Subsequently, we screened the pLannotate annotations and selected GP3D_CHLT2, a CDS found in all 47 plasmids, to rearrange them in the same strand orientation and order (see Figure S8) using a custom python script (https://github.com/MarieLataretu/rearrangeFasta).

Next, we performed a more detailed functional annotation of the re-arranged plasmid sequences using Bakta v1.6.0 [96] with parameters –genus Chlamydia –species psittaci. Finally, we used MAFFT v7.508 [97] with the 'linsi' option to calculate a multiple sequence alignment followed by RAxML v8.2.12 [90] with 1000 bootstrap replicates and the GTRGAMMA model to construct a plasmid tree. We used strain M56 (NC_018635) as an outgroup to arrange a topology comparable to the genome-based phylogeny. We visualized the tree using the Newick Utilities software package [98] and finalized the plasmid tree figure using Inkscape. Please note that we removed low bootstrap values at the leaves of the tree where no robust arrangement of identical or highly similar plasmid sequences was possible.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-023-09370-w.

Additional file 1: Figure S1. Assembly contiguity and size of 38 re-assembled genomes. For n = 38 NCBI genomes we were able to find Illumina short-read sequencing data to reassemble these strains. The Icarus plot (output of Quast v5.2.0 run with default parameters) shows the assembly contiguity and size for all 38 re-assemblies together with the corresponding genomes downloaded from NCBI. In addition, re-assemblies and NCBI genomes were decontaminated using CLEAN. The top eleven re-assemblies were finally selected to be integrated in our study and replaced the original NCBI genomes due to higher assembly contiguity and better N50 values. All contigs are sorted by length, starting with the longest contigs on the left and decreasing in length to the right. Thus, it is possible to mark the contigs where 50% (90%) of all the nucleotides in an assembly are covered by this contig and all longer contigs as a measure of assembly contiguity and quality. The purple bars mark contigs where a certain Nx (N50 or N90) is reached in an assembly.

Additional file 2: Figure S2. Phylogenetic tree based on the 904 common genes of 61 *C. psittaci* strains. The tree was reconstructed based on the concatenated core gene alignments at protein level produced by RIBAP and calculated using FastTree.

Additional file 3: Figure S3. SNP-based tree determined from 61 *C. psittaci* genomes. The eight distinct lineages defined by Vorimore *et al.* [37],i.e. group I_psittacine, group II_duck, group III_pigeon, group IV_turkey,group V_Mat116, group VI_M56, group VII_VS225, and group VIII_WC, are represented by colored circles. The tree was built using RAxML version 8.2.9 with the GTRGAMMA model and 1000 bootstrap replicates based on thefiltered SNP matrix (4011 SNPs) from BioNumerics.

Additional file 4: Figure S4. Phylogenetic tree based on nucleotide sequences of the extracted PZ of 53 *C. psittaci* strains used in this study. Sequences of 8 strains, where this region was located on several scaffolds, were not included here. The tree was constructed using RAxML v8.2.11 with GTRGAMMA nucleotide model and Rapid hill-climbing algorithm.

Additional file 5: Figure S5. RAxML tree of the alignment of OmpA amino acid sequences from 61 *C. psittaci* strains as processed by RIBAP (Group 879). Bootstrap values are indicated at inner nodes. For identical taxa, bootstrap values were discarded, due to the interchangeability of corresponding gene sequences. The colored bar on the right indicates the respective *ompA* genotypes.

Additional file 6: Figure S6. Phylogenetic tree reconstructed from the alignment of re-assembled and re-arranged plasmid sequences from 47 *C. psittaci* strains. The tree was built using RAxML version 8.2.12 with the GTRGAMMA model and 1000 bootstrap replicates.

Additional file 7: Figure S7. Location of genes encoding polymorphic membrane proteins in the genome of *C. psittaci* strain 6BC.

Additional file 8: Figure S8. Here, we exemplarily show the results of our re-assembly and re-arrangement efforts for the plasmid of *C. psittaci* strain MN. A) CDS annotation achieved with pLannotate for plasmid sequence directly obtained from NCBI. Note that genes GP5D_CHLPS and GP2D_CHLPS were only found with 68 and 52 % sequence similarity, respectively, which is marked by white arrows. B) Re-assembled plasmid sequence using corresponding raw-read data of strain MN and after re-arrangement using GP3D_CHLT2 as marker gene (orange frame). In the re-assembled plasmid, GP5D_CHLPS and GP2D_CHLPS achieved a sequence similarity of 99 and 100 %, respectively. Further details and results for all other plasmids can be found in the OSF (https://osf.io/rbca9).

Additional file 9: Table S1. Basic characteristics, assembly state and typing data of all 61 strains included in this study. **Table S2.** Genetic distances (% identities) calculated from the nucleotide sequence alignment of complete plasticity zones of 53 *Chlamydia psittaci* strains. **Table S3.** Multiple blast search for Pmp homologs to strain 6BC in 61 *Chlamydia psittaci* genomes (best hits). **Table S4.** Amino acid sequence variation among individual Pmps from 61 *Chlamydia psittaci* strains. **Table S5.** Basic parameters of plasmids in *C. psittaci* strains. **Table S6.** Genetic distances (% identities) calculated from the nucleotide sequence alignment of 47 *C. psittaci* plasmids. **Table S7.** *Chlamydia psittaci* strains and host organisms.

Acknowledgements

We thank Hugues Richard, Robert Koch Institute, MF1, for his help with the statistical analysis. We are also grateful to Marie Lataretu for her assistance with the plasmid analysis.

Ethical guidelines

Not applicable.

Authors' contributions

KS, MH and MM designed the study. LMB, FV and KeL conducted the processing and analysis of raw sequence data. CS contributed the structural analysis of OmpA. MH, KS, KeL and KaL conducted comparative analysis of genomic sites. KeL, MH and KS did the plasmid analysis. KS, MH, KeL and MM wrote the manuscript. All authors read and approved the revised version of the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL. The study was partially funded by Deutsche Forschungsgemeinschaft (DFG), FZT 118 – iDiv 202548816 (Kevin Lamkiewicz).

Availability of data and materials

All used genomes including the reassembled ones, the `QUAST` and `IDEEL` quality results, as well as the complete core gene `RIBAP` output are deposited in an OSF repository:https://osf.io/rbca9/.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]RNA Bioinformatics and High-Throughput Analysis, Friedrich Schiller University Jena, 07743 Jena, Germany. [2]Methodology and Research Infrastructure, Bioinformatics, Robert Koch Institute, 13353 Berlin, Germany. [3]Laboratory for Animal Health, Identypath, ANSES Maisons-Alfort, Paris-Est University, 94706 Paris, France. [4]Ernst-Ruska Centre 3 / Structural Biology, Forschungszentrum Jülich, Wilhelm-Johnen-Straße, 52425 Jülich, Germany. [5]Institute

for Biological Information Processing 6 / Structural Cellular Biology, Forschungszentrum Jülich, Wilhelm-Johnen-Straße, 52425 Jülich, Germany. [6]Department of Biology, Heinrich Heine University, Universitätsstr. 1, 40225 Düsseldorf, Germany. [7]Laboratory for Animal Health, Bacterial Zoonosis Unit, ANSES Maisons-Alfort, Paris-Est University, 94706 Paris, France. [8]JRG Analytical Micro-Bioinformatics, Friedrich Schiller University Jena, 07743 Jena, Germany.

Received: 20 September 2022 Accepted: 9 May 2023
Published online: 29 May 2023

References

1. Knittler MR, Sachse K: Chlamydia psittaci: update on an underestimated zoonotic agent (Minireview). FEMS Pathogens and Disease 2014, 73: https://doi.org/10.1093/femspd/ftu1007.
2. Kaleta EF, Taday EM. Avian host range of Chlamydophila spp. based on isolation, antigen detection and serology. Avian Pathol. 2003;32(5):435–61.
3. Beeckman DS, Vanrompay DC. Zoonotic Chlamydophila psittaci infections from a clinical perspective. Clin Microbiol Infect. 2009;15(1):11–7.
4. Sachse K, Laroucau K, Vanrompay D. Avian Chlamydiosis. Curr Clin Microbiol Rep. 2015;2:10–21.
5. Borel N, Sachse K. Zoonotic Transmission of Chlamydia spp.: Known for 140 Years, but Still Underestimated, in Zoonoses: Infections Affecting Humans and Animals. Sing A, Editor. Springer; 2023. https://doi.org/10.1007/978-3-030-85877-3_53-1.
6. Arenas-Valls N, Chacon S, Perez A, Del Pozo R. Atypical Chlamydia psittaci pneumonia Four related cases. Arch Bronconeumol. 2017;53(5):277–9.
7. Gaede W, Reckling KF, Dresenkamp B, Kenklies S, Schubert E, Noack U, Irmscher HM, Ludwig C, Hotzel H, Sachse K. Chlamydophila psittaci infections in humans during an outbreak of psittacosis from poultry in Germany. Zoonoses Public Health. 2008;55(4):184–8.
8. Ferreira VL, Silva MV, Bassetti BR, Pellini ACG, Raso TF. Intersectoral action for health: preventing psittacosis spread after one reported case. Epidemiol Infect. 2017;145(11):2263–8.
9. Vorimore F, Thebault A, Poisson S, Cleva D, Robineau J, de Barbeyrac B, Durand B, Laroucau K. Chlamydia psittaci in ducks: a hidden health risk for poultry workers. Pathog Dis. 2015;73(1):1–9.
10. Hinton DG, Shipley A, Galvin JW, Harkin JT, Brunton RA. Chlamydiosis in workers at a duck farm and processing plant. Aust Vet J. 1993;70(5):174–6.
11. Van Droogenbroeck C, Beeckman DS, Verminnen K, Marien M, Nauwynck H, Boesinghe Lde T, Vanrompay D. Simultaneous zoonotic transmission of Chlamydophila psittaci genotypes D, F and E/B to a veterinary scientist. Vet Microbiol. 2009;135(1–2):78–81.
12. McGuigan CC, McIntyre PG, Templeton K. Psittacosis outbreak in Tayside, Scotland, December 2011 to February 2012. Eurosurveillance. 2012;17(22):20186.
13. Wallensten A, Fredlund H, Runehagen A: Multiple human-to-human transmission from a severe case of psittacosis, Sweden, January-February 2013. Euro Surveill 2014;19(42). https://doi.org/10.2807/1560-7917.es2014.19.42.20937.
14. Hughes C, Maharg P, Rosario P, Herrell M, Bratt D, Salgado J, Howard D. Possible nosocomial transmission of psittacosis. Infect Control Hosp Epidemiol. 1997;18(3):165–8.
15. Ito I, Ishida T, Mishima M, Osawa M, Arita M, Hashimoto T, Kishimoto T. Familial cases of psittacosis: possible person-to-person transmission. Intern Med. 2002;41(7):580–3.
16. Fischer N, Rohde H, Indenbirken D, Gunther T, Reumann K, Lutgehetmann M, Meyer T, Kluge S, Aepfelbacher M, Alawi M, et al. Rapid metagenomic diagnostics for suspected outbreak of severe pneumonia. Emerg Infect Dis. 2014;20(6):1072–5.
17. Sigalova OM, Chaplin AV, Bochkareva OO, Shelyakin PV, Filaretov VA, Akkuratov EE, Burskaia V, Gelfand MS. Chlamydia pan-genomic analysis reveals balance between host adaptation and selective pressure to genome reduction. BMC Genomics. 2019;20(1):710.
18. Kuo CC, Grayston JT. Amino acid requirements for growth of Chlamydia pneumoniae in cell cultures: growth enhancement by lysine or methionine depletion. J Clin Microbiol. 1990;28(6):1098–100.
19. Moulder JW. Interaction of chlamydiae and host cells in vitro. Microbiol Rev. 1991;55(1):143–90.
20. Ouellette SP, Dorsey FC, Moshiach S, Cleveland JL, Carabeo RA. Chlamydia species-dependent differences in the growth requirement for lysosomes. PLoS One. 2011;6(3):e16783.
21. Saka HA, Valdivia RH. Acquisition of nutrients by Chlamydiae: unique challenges of living in an intracellular compartment. Curr Opin Microbiol. 2010;13(1):4–10.
22. Salzberg SL. Next-generation genome annotation: we still struggle to get it right. Genome Biol. 2019;20(1):92.
23. Collingro A, Tischler P, Weinmaier T, Penz T, Heinz E, Brunham RC, Read TD, Bavoil PM, Sachse K, Kahane S, et al. Unity in variety–the pan-genome of the Chlamydiae. Mol Biol Evol. 2011;28(12):3253–70.
24. Holzer M, Barf LM, Lamkiewicz K, Vorimore F, Lataretu M, Favaroni A, Schnee C, Laroucau K, Marz M, Sachse K. Comparative genome analysis of 33 Chlamydia strains reveals characteristic features of Chlamydia psittaci and closely related species. Pathogens. 2020;9(11):899.
25. Nunes A, Gomes JP. Evolution, phylogeny, and molecular epidemiology of Chlamydia. Infect Genet Evol. 2014;23:49–64.
26. Hadfield J, Harris SR, Seth-Smith HMB, Parmar S, Andersson P, Giffard PM, Schachter J, Moncada J, Ellison L, Vaulet MLG, et al. Comprehensive global genome dynamics of Chlamydia trachomatis show ancient diversification followed by contemporary mixing and recent lineage expansion. Genome Res. 2017;27(7):1220–9.
27. Eder T, Kobus S, Stallmann S, Stepanow S, Kohrer K, Hegemann JH, Rattei T. Genome sequencing of Chlamydia trachomatis serovars E and F reveals substantial genetic variation. Pathog Dis. 2017;75(9):ftx120.
28. Seth-Smith HMB, Benard A, Bruisten SM, Versteeg B, Herrmann B, Kok J, Carter I, Peuchant O, Bebear C, Lewis DA, et al. Ongoing evolution of Chlamydia trachomatis lymphogranuloma venereum: exploring the genomic diversity of circulating strains. Microb Genom. 2021;7(6):000599.
29. Versteeg B, Bruisten SM, Pannekoek Y, Jolley KA, Maiden MCJ, van der Ende A, Harrison OB. Genomic analyses of the Chlamydia trachomatis core genome show an association between chromosomal genome, plasmid type and disease. BMC Genomics. 2018;19(1):130.
30. Suchland RJ, Carrell SJ, Ramsey SA, Hybiske K, Debrine AM, Sanchez J, Celum C, Rockey DD. Genomic analysis of MSM rectal Chlamydia trachomatis isolates identifies predicted tissue-tropic lineages generated by intraspecies lateral gene transfer-mediated evolution. Infect Immun. 2022;90(11):e0026522.
31. Alkhidir AAI, Holland MJ, Elhag WI, Williams CA, Breuer J, Elemam AE, El Hussain KMK, Ournasseir MEH, Pickering H. Whole-genome sequencing of ocular Chlamydia trachomatis isolates from Gadarif State, Sudan. Parasit Vectors. 2019;12(1):518.
32. Roulis E, Bachmann NL, Myers GS, Huston W, Summersgill J, Hudson A, Dreses-Werringloer U, Polkinghorne A, Timms P. Comparative genomic analysis of human Chlamydia pneumoniae isolates from respiratory, brain and cardiac tissues. Genomics. 2015;106(6):373–83.
33. Jelocnik M, Bachmann NL, Kaltenboeck B, Waugh C, Woolford L, Speight KN, Gillett A, Higgins DP, Flanagan C, Myers GS, et al. Genetic diversity in the plasticity zone and the presence of the chlamydial plasmid differentiates Chlamydia pecorum strains from pigs, sheep, cattle, and koalas. BMC Genomics. 2015;16:893.
34. Seth-Smith HMB, Buso LS, Livingstone M, Sait M, Harris SR, Aitchison KD, Vretou E, Siarkou VI, Laroucau K, Sachse K, et al. European Chlamydia abortus livestock isolate genomes reveal unusual stability and limited diversity, reflected in geographical signatures. BMC Genomics. 2017;18(1):344.
35. Read TD, Joseph SJ, Didelot X, Liang B, Patel L, Dean D. Comparative analysis of Chlamydia psittaci genomes reveals the recent emergence of a pathogenic lineage with a broad host range. MBio. 2013;4(2):e00604-12.
36. White RT, Anstey SI, Kasimov V, Jenkins C, Devlin J, El-Hage C, Pannekoek Y, Legione AR, Jelocnik M. One clone to rule them all: Culture-independent genomics of Chlamydia psittaci from equine and avian hosts in Australia. Microb Genom. 2022;8(10):mgen000888.
37. Branley J, Bachmann NL, Jelocnik M, Myers GS, Polkinghorne A. Australian human and parrot Chlamydia psittaci strains cluster within the highly virulent 6BC clade of this important zoonotic pathogen. Sci Rep. 2016;6:30019.
38. Stewart RD, Auffret MD, Warr A, Walker AW, Roehe R, Watson M. Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. Nat Biotechnol. 2019;37(8):953–61.

39. Lamkiewicz K, Barf L-M, Sachse K, Hölzer M. Pangenome calculation beyond the species level using RIBAP: A comprehensive bacterial core genome annotation pipeline based on Roary and pairwise ILPs. https://doi.org/10.1101/2023.05.05.539552, https://www.biorxiv.org/content/10.1101/2023.05.05.539552v1.
40. Vorimore F, Aaziz R, de Barbeyrac B, Peuchant O, Szymanska-Czerwinska M, Herrmann B, Schnee C, Laroucau K. A new SNP-based genotyping method for C. psittaci: Application to field samples for quick identification. Microorganisms. 2021;9(3):625.
41. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Zidek A, Potapenko A, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583–9.
42. Mirdita M, Schutze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: making protein folding accessible to all. Nat Methods. 2022;19(6):679–82.
43. Akdel M, Pires DEV, Porta Pardo E, Jänes J, Zalevsky AO, Mészáros B, Bryant P, Good LL, Laskowski RA, Pozzati G et al: A structural biology community assessment of AlphaFold 2 applications. bioRxiv 2021:2021.2009.2026.461876.
44. Seth-Smith HM, Harris SR, Persson K, Marsh P, Barron A, Bignell A, Bjartling C, Clark L, Cutcliffe LT, Lambden PR, et al. Co-evolution of genomes and plasmids within Chlamydia trachomatis and the emergence in Sweden of a new variant strain. BMC Genomics. 2009;10:239.
45. Schmid M, Frei D, Patrignani A, Schlapbach R, Frey JE, Remus-Emsermann MNP, Ahrens CH. Pushing the limits of de novo genome assembly for complex prokaryotic genomes harboring very long, near identical repeats. Nucleic Acids Res. 2018;46(17):8953–65.
46. Ricker N, Qian H, Fulthorpe RR. The limitations of draft assemblies for understanding prokaryotic adaptation and evolution. Genomics. 2012;100(3):167–75.
47. Voigt A, Schofl G, Saluz HP. The Chlamydia psittaci genome: a comparative analysis of intracellular pathogens. PLoS One. 2012;7(4):e35097.
48. Bonner CA, Byrne GI, Jensen RA. Chlamydia exploit the mammalian tryptophan-depletion defense strategy as a counter-defensive cue to trigger a survival state of persistence. Front Cell Infect Microbiol. 2014;4:17.
49. Klapproth JM. The role of lymphostatin/EHEC factor for adherence-1 in the pathogenesis of gram negative infection. Toxins (Basel). 2010;2(5):954–62.
50. Belland RJ, Scidmore MA, Crane DD, Hogan DM, Whitmire W, McClarty G, Caldwell HD. Chlamydia trachomatis cytotoxicity associated with complete and partial cytotoxin genes. Proc Natl Acad Sci USA. 2001;98(24):13984–9.
51. Read TD, Myers GS, Brunham RC, Nelson WC, Paulsen IT, Heidelberg J, Holtzapple E, Khouri H, Federova NB, Carty HA, et al. Genome sequence of Chlamydophila caviae (Chlamydia psittaci GPIC): examining the role of niche-specific genes in the evolution of the Chlamydiaceae. Nucleic Acids Res. 2003;31:2134–47.
52. UniProt [https://www.uniprot.org/uniprot/A0A656ISD9].
53. Reinert DJ, Jank T, Aktories K, Schulz GE. Structural basis for the function of Clostridium difficile toxin B. J Mol Biol. 2005;351(5):973–81.
54. Busch C, Hofmann F, Selzer J, Munro S, Jeckel D, Aktories K. A common motif of eukaryotic glycosyltransferases is essential for the enzyme activity of large clostridial cytotoxins. J Biol Chem. 1998;273(31):19566–72.
55. Carabeo R. Bacterial subversion of host actin dynamics at the plasma membrane. Cell Microbiol. 2011;13(10):1460–9.
56. Guo W, Jelocnik M, Li J, Sachse K, Polkinghorne A, Pannekoek Y, Kaltenboeck B, Gong J, You J, Wang C. From genomes to genotypes: molecular epidemiological analysis of Chlamydia gallinacea reveals a high level of genetic diversity for this newly emerging chlamydial pathogen. BMC Genomics. 2017;18(1):949.
57. Keb G, Fields KA. An ancient molecular arms race: Chlamydia vs. membrane attack complex/perforin (MACPF) domain proteins. Front Immunol. 2020;11:1490.
58. Taylor LD, Nelson DE, Dorward DW, Whitmire WM, Caldwell HD. Biological characterization of Chlamydia trachomatis plasticity zone MACPF domain family protein CT153. Infect Immun. 2010;78(6):2691–9.
59. Beder T, Saluz HP. Virulence-related comparative transcriptomics of infectious and non-infectious chlamydial particles. BMC Genomics. 2018;19(1):575.
60. Vasilevsky S, Stojanov M, Greub G, Baud D. Chlamydial polymorphic membrane proteins: regulation, function and potential vaccine candidates. Virulence. 2016;7(1):11–22.
61. Molleken K, Schmidt E, Hegemann JH. Members of the Pmp protein family of Chlamydia pneumoniae mediate adhesion to human cells via short repetitive peptide motifs. Mol Microbiol. 2010;78(4):1004–17.
62. Wolff BJ, Morrison SS, Pesti D, Ganakammal SR, Srinivasamoorthy G, Changayil S, Weil MR, MacCannell D, Rowe L, Frace M, et al. Chlamydia psittaci comparative genomics reveals intraspecies variations in the putative outer membrane and type III secretion system genes. Microbiology. 2015;161(7):1378–91.
63. Favaroni A, Trinks A, Weber M, Hegemann JH, Schnee C. Pmp repertoires influence the different infectious potential of avian and mammalian Chlamydia psittaci strains. Front Microbiol. 2021;12:656209.
64. Mital J, Miller NJ, Dorward DW, Dooley CA, Hackstadt T. Role for chlamydial inclusion membrane proteins in inclusion membrane structure and biogenesis. PLoS One. 2013;8(5):e63426.
65. Sachse K, Rahman KS, Schnee C, Muller E, Peisker M, Schumacher T, Schubert E, Ruettger A, Kaltenboeck B, Ehricht R. A novel synthetic peptide microarray assay detects Chlamydia species-specific antibodies in animal and human sera. Sci Rep. 2018;8(1):4701.
66. Lutter EI, Martens C, Hackstadt T. Evolution and conservation of predicted inclusion membrane proteins in chlamydiae. Comp Funct Genomics. 2012;2012:362104.
67. Andersen AA. Serotyping of Chlamydia psittaci isolates using serovar-specific monoclonal antibodies with the microimmunofluorescence test. J Clin Microbiol. 1991;29(4):707–11.
68. Vanrompay D, Butaye P, Sayada C, Ducatelle R, Haesebrouck F. Characterization of avian Chlamydia psittaci strains using omp1 restriction mapping and serovar-specific monoclonal antibodies. Res Microbiol. 1997;148(4):327–33.
69. Borges V, Cordeiro D, Salas AI, Lodhia Z, Correia C, Isidro J, Fernandes C, Rodrigues AM, Azevedo J, Alves J, et al. Chlamydia trachomatis: when the virulence-associated genome backbone imports a prevalence-associated major antigen signature. Microb Genom. 2019;5(11):e000313.
70. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, Yuan D, Stroe O, Wood G, Laydon A, et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res. 2022;50(D1):D439–44.
71. Feher VA, Randall A, Baldi P, Bush RM, de la Maza LM, Amaro RE. A 3-dimensional trimeric beta-barrel model for Chlamydia MOMP contains conserved and novel elements of Gram-negative bacterial porins. PLoS One. 2013;8(7):e68934.
72. Szabo KV, O'Neill CE, Clarke IN. Diversity in Chlamydial plasmids. PLoS One. 2020;15(5):e0233298.
73. Jones CA, Hadfield J, Thomson NR, Cleary DW, Marsh P, Clarke IN, O'Neill CE. The nature and extent of plasmid variation in Chlamydia trachomatis. Microorganisms. 2020;8(3):373.
74. Laroucau K, de Barbeyrac B, Vorimore F, Clerc M, Bertin C, Harkinezhad T, Verminnen K, Obeniche F, Capek I, Bebear C, et al. Chlamydial infections in duck farms associated with human cases of psittacosis in France. Vet Microbiol. 2009;135(1–2):82–9.
75. Cong W, Huang SY, Zhang XY, Zhou DH, Xu MJ, Zhao Q, Song HQ, Zhu XQ, Qian AD. Seroprevalence of Chlamydia psittaci infection in market-sold adult chickens, ducks and pigeons in north-western China. J Med Microbiol. 2013;62(Pt 8):1211–4.
76. Lugert R, Gross U, Masanta WO, Linsel G, Heutelbeck A, Zautner AE. Seroprevalence of Chlamydophila psittaci among employees of two German duck farms. Eur J Microbiol Immunol (Bp). 2017;7(4):267–73.
77. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34(17):i884–90.
78. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19(5):455–77.
79. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 2014;9(11):e112963.

80. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. aRxiv. 2013. https://doi.org/10.48550/arXiv.1303.3997.
81. Wick RR, Schultz MB, Zobel J, Holt KE. Bandage: interactive visualization of de novo genome assemblies. Bioinformatics. 2015;31(20):3350–2.
82. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013;29(8):1072–5.
83. Mikheenko A, Valin G, Prjibelski A, Saveliev V, Gurevich A. Icarus: visualizer for de novo assembly evaluation. Bioinformatics. 2016;32(21):3321–3.
84. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;14:2068–9.
85. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, Fookes M, Falush D, Keane JA, Parkhill J. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics. 2015;31(22):3691–3.
86. Steinegger M, Soding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 2017;35(11):1026–8.
87. Huang W, Li L, Myers JR, Marth GT. ART: a next-generation sequencing read simulator. Bioinformatics. 2012;28(4):593–4.
88. Li H. New strategies to improve minimap2 alignment accuracy. Bioinformatics. 2021;37(23):4572-4.
89. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021;10(2):giab008.
90. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30(9):1312–3.
91. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, Parkhill J, Harris SR. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Res. 2015;43(3):e15.
92. Hadfield J, Croucher NJ, Goater RJ, Abudahab K, Aanensen DM, Harris SR. Phandango: an interactive viewer for bacterial population genomics. Bioinformatics. 2018;34(2):292–3.
93. Price MN, Dehal PS, Arkin AP. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol. 2009;26(7):1641–50.
94. Pettersen EF, Goddard TD, Huang CC, Meng EC, Couch GS, Croll TI, Morris JH, Ferrin TE. UCSF ChimeraX: Structure visualization for researchers, educators, and developers. Protein Sci. 2021;30(1):70–82.
95. McGuffie MJ, Barrick JE. pLannotate: engineered plasmid annotation. Nucleic Acids Res. 2021;49(W1):W516–22.
96. Schwengers O, Jelonek L, Dieckmann MA, Beyvers S, Blom J, Goesmann A. Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. Microb Genom. 2021;7(11):000685.
97. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30(4):772–80.
98. Junier T, Zdobnov EM. The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. Bioinformatics. 2010;26(13):1669–70.

Publisher's Note