Cat fleas in flux: Rampant gene duplication, genome size plasticity, and paradoxical *Wolbachia* infection

Timothy P. Driscoll [1,^], Victoria I. Verhoeve [2,^], Joseph J. Gillespie [2,^,*], J. Spencer Johnston [3], Mark L. Guillotte [2], Kristen E. Rennoll-Bankert [2], M. Sayeedur Rahman [2], Darren Hagen [4], Christine G. Elsik [5,6,7], Kevin R. Macaluso [8], Abdu F. Azad [2]

Author details

[1] Department of Biology, West Virginia University, Morgantown, West Virginia

[2] Department of Microbiology and Immunology, University of Maryland School of Medicine, Baltimore, Maryland

[3] Department of Entomology, Texas A&M University, College Station, Texas

[4] Department of Animal and Food Sciences, Oklahoma State University, Stillwater, Oklahoma

[5] Division of Animal Sciences, University of Missouri, Columbia, MO, USA.

[6] Division of Plant Sciences, University of Missouri, Columbia, MO, USA.

[7] MU Informatics Institute, University of Missouri, Columbia, MO, USA.

[8] Vector-borne Disease Laboratories, Department of Pathobiological Sciences, School of Veterinary Medicine, Louisiana State University

* Correspondence to: Joe Gillespie, JGillespie@som.umaryland.edu

[^] Contributed equally.

Running head: Cat fleas have inordinate copy number variation

Key words: *Ctenocephalides felis*; cat flea; genome; Hi-C assembly; PacBio sequencing; *Wolbachia*; gene duplication; copy number variation; parasitism

Abstract

Background: Fleas (Insecta: Siphonaptera) are small flightless parasites of birds and mammals; their blood-feeding can transmit many serious pathogens (i.e. the etiological agents of bubonic plague, endemic and murine typhus). The lack of flea genome assemblies has hindered research, especially comparisons to other disease vectors. Accordingly, we sequenced the genome of the cat flea, *Ctenocephalides felis*, an insect with substantial human health and veterinary importance across the globe.

Results: By combining Illumina and PacBio sequencing with Hi-C scaffolding techniques, we generated a chromosome-level genome assembly for *C. felis*. Unexpectedly, our assembly revealed extensive gene duplication across the entire genome, exemplified by ~38% of protein-coding genes with two or more copies and over 4,000 tRNA genes. A broad range of genome size determinations (433-551 Mb) for individual fleas sampled across different populations supports the widespread presence of fluctuating copy number variation (CNV) in *C. felis*. Similarly broad genome sizes were also calculated for individuals of *Xenopsylla cheopis* (Oriental rat flea), indicating that this remarkable "genome-in-flux" phenomenon could be a siphonapteran-wide trait. Nevertheless, cat flea genomes have not experienced substantial inflation due to rampant CNV, nor do we find a significant reduction in gene families as has been described for other parasites. Quizzically, individual fleas were found to be co-infected with parasitic and symbiotic *Wolbachia*, both of which are novel strains unique to *C. felis*.

46 <u>Conclusion</u>: Rampant CNV in *C. felis* has dire implications for gene-targeting pest control
47 measures and stands to complicate standard normalization procedures utilized in comparative
48 transcriptomics analysis. Coupled with paradoxical co-infection with novel *Wolbachia*
49 endosymbionts, which are potential tools for blocking pathogen transmission, these oddities
50 highlight a unique and underappreciated human disease vector.

51

52 **Background**

53 With over 2,500 described species, fleas (Hexapoda: Siphonaptera) are small (~3 mm) flightless
54 insects that parasitize mainly mammals and birds [1]. Diverging from Order Mecoptera
55 (scorpionflies and hangingflies) in the Jurassic period [2], fleas are one of 11 extant orders of
56 Holometabola, a superorder of insects that collectively go through distinctive larval, pupal, and
57 adult stages. The limbless, worm-like flea larvae contain chewing mouthparts and feed primarily
58 on organic debris, while adult mouthparts are modified for piercing skin and sucking blood.
59 Other adaptations to an ectoparasitic lifestyle include wing loss, extremely powerful hind legs for
60 jumping, strong claws for grasping, and a flattened body that facilitates movement on host fur
61 and feathers.

62 The Oriental rat flea, *Xenopsylla cheopis*, and to a lesser extent the cat flea, *Ctenocephalides*
63 *felis*, transmit *Yersinia pestis*, the causative agent of bubonic plague [3–5]. Fleas that feed away
64 from their primary hosts (black rats and other murids) can introduce *Y. pestis* to humans, which
65 historically has eliminated a substantial fraction of the world's human population; e.g., the
66 Plague of Justinian and the Black Death [5]. Bubonic plague remains a significant threat to
67 human health [6,7] as do other noteworthy diseases propagated by flea infestations, including
68 murine typhus (*Rickettsia typhi*), murine typhus-like illness (*R. felis*), cat-scratch disease

69 (*Bartonella henselae*), and Myxomatosis (myxoma virus) [8,9]. Fleas also serve as intermediate
70 hosts for certain medically-relevant helminths and trypanosome protozoans [10]. In addition to
71 the potential for infectious disease transmission, flea bites are also a significant nuisance and can
72 lead to serious dermatitis for both humans and their companion animals. Epidermal burrowing
73 by the jigger flea, *Tunga penetrans*, causes a severe inflammatory skin disease known as
74 Tungiasis, which is a scourge on many human populations within tropical parts of Africa, the
75 Caribbean, Central and South America, and India [11,12]. Skin lesions that arise from flea
76 infestations also serve as sites for secondary infection. Collectively, fleas inflict a multifaceted
77 human health burden with enormous public health relevance [13].

78 Most flea species reproduce solely on their host; however, their ability to feed on a range of
79 different animals poses a significant risk for humans cohabitating with pets that are vulnerable to
80 flea feeding – which includes most warm-blooded, hairy vertebrates [14]. As such, fleas also
81 have a substantial economic impact from a veterinary perspective [15]. Many common pets are
82 susceptible to flea infestations that often cause intense itching, bleeding, hair loss, and potential
83 development of flea allergy dermatitis, an eczematous itchy skin disease. In the United States
84 alone, annual costs for flea-related veterinary bills tally approximately $4.4 billion, with another
85 $5 billion for prescription flea treatment and pest control [16]. Despite intense efforts to control
86 infestations, fleas continue to pose a significant burden to companion animals and their owners
87 [17].

88 Notwithstanding their tremendous impact on global health and economy, fleas are relatively
89 understudied compared to other arthropod disease vectors [18]. While transcriptomics data for
90 mecopteroids (Mecoptera + Siphonaptera) have proven useful for Holometabola phylogeny
91 estimation [2], assessment of flea immune pathways [19], and analysis of opsin evolution [20],

92 the lack of mecopteroid genomes limits further insight into the evolution of Antliophora
93 (mecopteroids + Diptera (true flies)) and severely restricts comparative studies of disease
94 vectors. Thus, sequencing flea genomes stands to greatly improve our understanding of the
95 shared and divergent mechanisms underpinning flea and fly vectors, a collective lineage
96 comprised of the deadliest animals known to humans [21]. To address this glaring void in insect
97 genomics and vector biology, we sequenced the genome of *C. felis*, a principal vector of *R. typhi*,
98 *R. felis*, and *Bartonella* spp. [22–25] and an insect with substantial human health and veterinary
99 importance across the globe [1].
100

101 **Results**

102 Pooled female fleas from the Elward Laboratory colony (Soquel, California; hereafter EL fleas)
103 were used to generate short (Illumina), long (PacBio), and chromatin-linked (Hi-C) sequencing
104 reads. A total of 7.2 million initial PacBio reads were assembled into 16,622 contigs (773.8 Mb;
105 N50 = 61 Kb), polished with short-read data, then scaffolded using Hi-C into 3,926 scaffolds
106 with a final N50 of 71.7 Mb. A total of 193 scaffolds were identified as arising from microbial
107 sources and removed before gene model prediction and annotation. A large fraction of the total
108 assembly (85.6%, or 654 Mb) was found in nine scaffolds (all greater than 10 Mb, hereafter
109 BIG9), while the remaining 14.4% (119.8 Mb) comprised scaffolds less than 1 Mb in length;
110 therefore, we suggest the *C. felis* genome contains nine chromosomes (**Fig. 1A**), an estimate
111 consistent with previously determined flea karyotypes [26,27]. The 3,724 shorter scaffolds (all
112 less than 1 Mb) mapped back to unique locations on BIG9 scaffolds (**fig. S1A**) but were not
113 assembled into the BIG9 scaffolds via proximity ligation. Comparison of *C. felis* protein-
114 encoding genes to the Benchmarking Universal Single-Copy Orthologues (BUSCO [28]) for

115 eukaryotes, arthropods, and insects indicates our BIG9 assembly is robust and lacks only a few
116 conserved genes (**fig. S1B**). As a result, we focus our subsequent analyses on the BIG9 scaffolds
117 unless otherwise noted.
118

119 **The *C. felis* genome and unprecedented gene duplication**

120 Previous work using flow cytometry estimated the size of the female *C. felis* genome at 465 Mb,
121 while our BIG9 assembly contained 654 Mb total bases (25% larger). Furthermore, BUSCO
122 analysis suggested that roughly 30% of conserved, single-copy Insecta genes in the BUSCO set
123 were duplicated in our assembly (**fig. S1B**). In order to investigate whether this duplication might
124 be widespread across the genome, and thereby account for the larger size of our assembly, we
125 used BLASTP to construct *C. felis*-specific protein families at varying levels of sequence identity
126 from 85-100%. Remarkably, 61% (10,088) of all protein-encoding genes in *C. felis* arise from
127 duplications at the 90% identity threshold or higher (**Fig. 1B**). Over 68% of these comprise true
128 (n=2) duplications, most of which occur as tandem or proximal loci less than 12 genes apart (**Fig.**
129 **1C, fig. S1L**). We observed little change in either the total number of duplications or the
130 distribution at thresholds below 90% identity; consequently, we define "duplications" here as
131 sequences that are 90% identical or higher (see **Methods**).

132 Duplications are on-going and rapidly diverging as evinced by: 1) their high concentration on
133 individual BIG9 scaffolds (**Fig. 1D**, **fig. S1**, **C** to **K**), 2) a lack of increasing divergence with
134 greater distance on scaffolds (**fig. S1L**), and 3) a lack of increasing divergence for duplicate
135 genes found across different scaffolds (**fig. S1M**). Among cellular functions for duplicate genes,
136 certain transposons and related factors (GO:0015074, “DNA integration”) are enriched relative
137 to 6,430 single copy protein-encoding genes (**Fig. 1E**, **table S1**). However, the frequency and

138 distribution of these elements are dwarfed by total duplicate genes (**fig. S1N**). Additionally,
139 transposons and other repeat elements encompass only 10% of the genome (**fig. S1O**), indicating
140 that selfish genetic elements do not contribute significantly to the rampant gene duplication
141 observed. Thus, the *C. felis* genome is remarkable given that genes producing duplications
142 (n=3,863 or ~38% of total protein-encoding genes) are 1) indiscriminately dispersed across
143 chromosomes, 2) not clustered into blocks that would suggest whole or partial genome
144 duplications, and 3) not the product of repeat element-induced genome obesity.

145 The *C. felis* genome also carries an impressive number of tRNA-encoding genes (*n*=4,358 on
146 BIG9 scaffolds) (**Fig. 1A**). While tRNA gene numbers and family compositions vary
147 tremendously across eukaryotes [29], the occurrence of more than 1000 tRNA genes per genome
148 is rare (**Fig. 1F**). Notably, the elevated abundance of tRNA genes in *C. felis* is complemented by
149 an enrichment in translation-related functions among duplicated protein-coding genes (**Fig. 1E,**
150 **table S1**). While this possibly indicates increased translational requirements to accommodate
151 excessive gene duplication, it is more likely a consequence of the indiscriminate nature of the
152 gene duplication process. Relative to tRNA gene frequencies in other holometabolan genomes,
153 *C. felis* has several elevated (Arg, Val, Phe, Thr) and reduced (Gly, Pro, Asp, Gln) numbers of
154 tRNA families (**fig. S1P**); however, *C. felis* codon usage is typical of holometabolan genomes
155 (**fig. S1Q**). Like proliferated protein-encoding genes, the significance of such high tRNA gene
156 numbers is unclear but further accentuates a genome in flux.

157

158 **Genome size estimation**

159 Duplicated regions (including intergenic sequences) account for approximately 227 Mb of the *C.*
160 *felis* genome; when subtracted from the BIG9 assembly (654 Mb), the resulting "core" genome

161 size of 427 Mb is congruous with a previous flow cytometry-based genome size estimate (mean
162 of 465 Mb, range of 32 Mb) for cat fleas previously assayed from a different geographic locale
163 [30]. To determine if EL fleas possess a greater genome size due to pronounced gene
164 duplication relative to other cat fleas, we similarly used flow cytometry to estimate genome sizes
165 for individual EL fleas and compared them to the previous findings. As expected, mean genome
166 size was not significantly different between sex-matched *C. felis* from the two populations (p =
167 0.1299). Remarkably, however, no two individual EL fleas possessed comparable genome sizes,
168 with an overall uniform size distribution and relatively large variability (118 Mb) (**Fig. 2A**).
169 Indeed, the coefficient of variation for *C. felis* (0.13; n = 26) was 3.2X higher than that of either
170 *Drosophila melanogaster* (0.040; n = 26) or *D. viridis* (0.039; n = 26), which were prepared and
171 measured concurrently (**Fig. 2A**, inset), underscoring the extraordinary extent of inter-individual
172 variation in *C. felis*. Genome size estimates for another flea (the rat flea, *X. cheopis*, also sex-
173 matched) show a similar uniform distribution and range across individuals (**Fig. 2A**), pointing to
174 an extraordinary genetic mechanism that may define siphonapteran genomes.

175 Accordingly, we propose that our assembly captured a conglomeration of individual flea
176 copy number variations (CNV) that is cumulative for all expansions and contractions of
177 duplicate regions (**Fig. 2B**). The presence of extensive gene duplications is further supported by
178 mapping short read Illumina data to our assembly, which showed a significantly reduced mean
179 read depth across duplicated loci versus single-copy genes (**Fig. 2C**). As an alternative to CNV,
180 we considered that allelic variation could also be contributing to extensive gene duplication in
181 our assembly. To address this concern, we took three approaches. First, polished contigs were
182 scanned for haplotigs using the program *Purge Haplotigs* [31]; no allelic variants were detected.
183 Second, we mapped 1KITE transcriptome reads [2] generated from fleas of an unrelated colony

184 (Kansas State University) to our assembly (**Fig. 2D**). If our sequence duplication is a result of
185 allelic variation within the EL colony, we would expect to see a lack of congruence in the
186 distribution of transcripts mapping to single copy genes versus duplicates (different colonies with
187 different allelic variation). We might also expect to see a significant proportion of transcripts
188 that do not map at all. Instead, 91% of 1KITE reads map to CDS in our assembly, and the
189 distributions of transcripts mapping to single copy and duplicate genes are identical.
190 Third, we reasoned if sequence duplications are the result of misassembled allelic variants,
191 then most duplicate CDS within a cluster would be the same length. Alternatively, if
192 duplications are true CNVs, we would expect a significant number of truncations as a
193 consequence of gene purging associated with unequal crossing over. To assess this, we
194 determined the proportion of duplicate clusters with one or more truncated members, as well as
195 the extent of truncation relative to the longest member of the cluster (**Fig. 2E**). Approximately
196 70% of gene duplications are not comparable in length. In addition, mean extent of truncation is
197 25% or greater across all clusters regardless of % identity. Together with genome size
198 estimations, short read mapping analysis, and transcript mapping to our assembly, these data
199 indicate active gene expansion and contraction underpinning CNV in fleas and dispel allelic
200 variation as a significant contributor to gene duplication. While the cytogenetic mechanisms are
201 unclear, elevated numbers of DNA repair enzymes (GO:0006281) relative to genome size may
202 correlate with excessive CNV (**table S1**).
203

204 **Genome evolution within Holometabola**

205 Despite inordinate gene duplication, the completeness of the *C. felis* proteome as estimated by
206 occurrence of 1,658 insect Benchmarking Universal Single-Copy Orthologues (BUSCOs) is

207 congruous with those of other sequenced holometabolan genomes (**Fig. 3A**). Only one other
208 genome (*Aedes albopictus*) contains greater gene duplication among BUSCOs than *C. felis*;
209 however, this mosquito genome is much larger (~2 Gb) and riddled with repeat elements [32]. A
210 genome-wide analysis of shared orthologs among 53 holometabolan genomes indicates a slight
211 affinity of *C. felis* with Coleoptera, though the divergent nature of Diptera and availability of
212 only a single flea genome likely mask inclusion of fleas with flies (**Fig. 3B**). Overall,
213 phylogenomics analysis reveals that *C. felis* harbors 3,491 orthologs found in at least one other
214 taxon from each holometabolan order (**Fig. 3C**); however, only 577 "core" orthologs were
215 present in all taxa from every order (**Fig. 3C**, yellow bar), reflecting either incomplete genome
216 assemblies or an incredibly patchwork Holometabola accessory genome (**fig. S3A**). Other
217 conserved protein-encoding genes that define higher-generic groups (**Fig. 3C**, inset) will inform
218 lineage diversification within Holometabola (**table S2**). Conversely, 29 protein-encoding genes
219 absent in *C. felis* but conserved in Panorpida species (Antliophora + Lepidoptera (butterflies and
220 moths)) stand to illuminate patterns and processes of flea specialization via reduction (**fig. S3B**,
221 **table S2**). Overall, despite its parasitic lifestyle and reductive morphology, *C. felis* has not
222 experienced a significant reduction in gene families (**fig. S3B, table S2F**) as seen in other host-
223 dependent eukaryotes [33].
224

225 **Unique cat flea genome features**

226 *C. felis* protein-encoding genes that failed to cluster with other Holometabola (4,282 sequences
227 in 2,055 ortholog groups, **Fig. 3C**) potentially define flea-specific attributes. Elimination of
228 divergent "holometabolan-like" proteins, identified with BLASTP against the nr database of
229 NCBI, left 2,084 "unique" *C. felis* proteins (**Fig. 4A, table S3**). These include proteins lacking

230 counterparts in the NCBI nr database (n=766), and proteins with either limited similarity to
231 Holometabola or greater similarity to non-holometabolan taxa (n=1,318). Proteins comprising
232 the latter set were assigned an array of functional annotations (GO, KEGG, InterPro, EC) and
233 stand to guide efforts for deciphering flea-specific innovations (**Fig. 4B**, **table S3**).
234 Two isoforms (A and B) of resilin, an elastomeric protein that provides soft rubber-elasticity
235 to mechanically active organs and tissues, were previously identified in *C. felis* and proposed to
236 underpin tarsal-mediated jumping [34]. Resilins typically have 1) highly repetitive Pro/Gly
237 motifs that provide high flexibility, 2) key Tyr residues that facilitate intermolecular bonds
238 between resilin polypeptides, and 3) a chitin-binding domain (CBD), though *C. felis* isoform B
239 lacks the CBD [34,35]. The *C. felis* assembly has two adjacent genes encoding resilins (gray
240 box, **Fig. 4C**): the larger (680 aa) protein is more similar to both resilin A and B isoforms
241 identified previously (>99 %ID), while the smaller (531 aa) protein is more divergent (98.8
242 %ID). These divergent resilins accentuate the observed CNV in *C. felis* and indicate additional
243 genetic complexity behind flea jumping. Furthermore, a cohort of diverse proteins containing
244 multiple resilin-like features and domains were identified, opening the door for future studies
245 aiming to characterize the molecular mechanisms underpinning the great jumping ability of fleas.
246

247 **The *C. felis* microbiome: evidence for symbiosis and parasitism**

248 Analysis of microbial-like Illumina reads revealed a bacterial dominance, primarily represented
249 by *Proteobacteria* (**Fig. 5A**, **table S4**). Aside from the *Wolbachia* reads (discussed below), none
250 of the bacterial taxa match to species previously detected in environmental [36,37] or colony
251 fleas [38]. Thus, a variable bacterial microbiome exists across geographically diverse fleas and
252 is likely influenced by the presence of pathogens [38]. Strong matches to lepidopteran-

253 associated *Chrysodeixis chalcites* nucleopolyhedrovirus and *Choristoneura occidentalis*
254 granulovirus, as well as *Pandoravirus dulcis*, identify underappreciated viruses that may play
255 important roles in the vectorial capacity of *C. felis*.

256 Remarkably, two divergent *Wolbachia* genomes were assembled, circularized and annotated.
257 Named *w*CfeT and *w*CfeJ, these novel strains were previously identified (using 16S rDNA) in a
258 cat flea colony maintained at Louisiana State University [38–40], which historically has been
259 replenished with EL fleas. Robust genome-based phylogeny estimation indicates *w*CfeT is
260 similar to undescribed *C. felis*-associated strains that branch ancestrally to most other *Wolbachia*
261 lineages [36,41], while *w*CfeJ is similar to undescribed *C. felis*-associated strains closely related
262 to *Wolbachia* supergroups C, D and F [42] (**Fig. 5B**; **table S4**). The substantial divergence of
263 *w*CfeT and *w*CfeJ from a *Wolbachia* supergroup B strain infecting *C. felis* (*w*Cte) indicates a
264 diversity of Wolbachiae capable of infecting cat fleas.

265 *w*CfeT and *w*CfeJ are notable for carrying segments of WO prophage, which are rarely
266 present in genomes of Wolbachiae outside of supergroups A and B [43]. Further, each genome
267 contains features that hint at contrasting relationships with *C. felis*. *w*CfeT carries the unique
268 biotin synthesis operon (**Fig. 5C**), which was originally discovered in *Rickettsia buchneri* by us
269 [44] and later identified in certain *Wolbachia* [45–47], *Cardinium* [48,49] and *Legionella* [50]
270 species. Given that some *Wolbachia* strains provide biotin to their insect hosts [45,51], we posit
271 that *w*CfeT has established an obligate mutualism with *C. felis* mediated by biotin-provisioning.

272 In contrast, *w*CfeJ appears to be a reproductive parasite, as it contains a toxin-antidote (TA)
273 operon that is similar to the CinA/B TA operon of *w*Pip_Pel that induces cytoplasmic
274 incompatibility (CI) in flies [52]. CinA/B operons are analogous to the CidA/B TA operons of
275 *w*Mel and *w*Pip_Pel, which also induce CI in fly hosts [53–55], yet the CinB toxin harbors dual

276 nuclease domains in place of the CidB deubiquitnase domain [56] (**Fig. 5D**). Given that the
277 genomes of many *Wolbachia* reproductive parasites harbor diverse arrays of CinA/B-and
278 CidA/B-like operons [56,57], wCfeJ's CinA/B TA operon might function in CI or some other
279 form of reproductive parasitism. Quizzically, the co-occurrence of *w*CfeJ and *w*CfeT in
280 individual fleas (gel image in **Fig. 5B**) indicates dual forces (mutualism, parasitism) that
281 potentially drive their infection in EL fleas.
282

283 ## Discussion

284 We set out to generate a genome sequence for the cat flea, a surprisingly absent resource for
285 comparative arthropod genomics and vector biology. Our efforts to generate a *C. felis* assembly
286 brought forth an unexpected finding, namely that no two cat fleas share the same genome
287 sequence. We provide multiple lines of evidence supporting flea genomes in flux (**Table 1**).

Table 1. Evidence Supporting Extensive Gene Duplication in Cat Fleas.

Approach	Source	Key Points
Genome size estimation	Fig. 2A, Fig S2	- *C. felis* from two populations have same mean genome size. - Individual cat fleas vary ~118 Mb in estimated genome size. - Individual rat fleas vary ~100 Mb in estimated genome size.
Long read assembly with proximity ligation	Fig. 1 Fig. S1, Table S5	- Nine scaffolds > 10 Mb are littered with gene duplications, which comprise 38% of protein coding genes. - No misassembly of allelic variants in the BIG9 scaffolds.
Transcript mapping	Fig. 2D	- 98% of duplicate genes have transcriptional support in RNA-Seq data from an independent colony (1KITE).
Short read mapping	Fig. 2C	- Short read data map with far greater depth to single-copy genes versus duplicate genes.
Assessment of duplication lengths	Fig. 2E	- 69% of duplications are divergent in length; heterogeneity in length and composition are positively correlated.

288

289 First, genome size estimations for over two dozen individual cat fleas from the EL colony
290 revealed over 150 Mb variation, a result consistent with prior genome size estimates for *C. felis*
291 from a different colony as well as rat fleas. Second, our haplotig-resolved assembly identified
292 rampant gene duplication throughout the genome. Third, RNA-Seq data from an independent
293 colony confirmed the pervasive gene duplication. Finally, ~70% of gene duplications are not
294 comparable in length, indicating active gene expansion and contraction. Since transposons and
295 other repeat elements are relatively sparse in *C. felis* and cannot account for such rampant CNV,
296 and given that no individual flea genome size was estimated to be larger than our BIG9
297 assembly, we posit that unequal crossing over and gene conversion continually create and
298 eliminate large linear stretches of DNA to keep the *C. felis* genome in a fluctuating continuum.
299 We favor this hypothesis over an ancient whole genome duplication event in Siphonaptera
300 provided that the majority of these duplications are tandem or proximal.

301 Ramifications of a genome in flux are readily identifiable. First, as gene duplication is a
302 major source of genetic novelty, extensive CNV likely affords *C. felis* with a dynamic platform
303 for innovation, allowing it to outpace gene-targeting pest control measures. Second, extensive
304 CNV will complicate standard normalization procedures utilized in comparative transcriptomics
305 analysis, requiring a more nuanced interpretation of standard metrics that are based on gene
306 length (i.e. RPKM, TPM, etc.). Furthermore, achieving high confidence with read-mapping to
307 cognate genes will be difficult in the face of neofunctionalization, subfunctionalization and early
308 pseudogenization, as well as dosage-based regulation of duplicate genes. Third, genetic markers
309 typically utilized for evolutionary analyses (e.g., phylotyping, population genetics,
310 phylogeography [58]) may yield erroneous results when applied to *C. felis* and related
311 *Ctenocephalides* species if targeted to regions of CNV (and particularly neofunctionalization).

312 Finally, as a *C. felis* chromosome-level genome assembly was only attainable by coupling
313 Illumina and PacBio sequencing with Hi-C scaffolding techniques, short-read based sequencing
314 strategies will be inadequate for other organisms with high CNV. The ability of the BIG9
315 assembly to serve as a reference genome in future short-read based sequencing efforts for other
316 cat fleas will be determined. Moving forward, newly developed low-input protocols for PacBio
317 sequencing will allow us to query individual fleas to robustly assess the degree of gene
318 duplication.

319 Excessive CNV in *C. felis*, and likely all Siphonaptera, requires the determination of the
320 genetic mechanisms at play. Why extreme gene duplication, when predicted across arthropods
321 using genomic and transcriptomic data [59], was not previously detected in fleas is unclear.
322 Excessive CNV aside, our study provides the first genome sequence for Siphonaptera, which will
323 substantially inform comparative studies on insect vectors of human disease. Furthermore,
324 newly-identified symbiotic (*w*CfeT) and parasitic (*w*CfeJ) *Wolbachia* will be paramount to
325 efforts for biocontrol of pathogens transmitted by cat fleas. The accrued resources and
326 knowledge from our study are timely. A drastic rise of murine typhus cases alone in Southern
327 California (https://www.cdph.ca.gov/Programs/CID/DCDC/Pages/Typhus.aspx) and Galveston,
328 Texas [60], which are directly attributable to fleas associated with increasing population sizes of
329 rodents and opossums, requires immediate and re-focused efforts to combat this serious and
330 underappreciated risk to human health.

331

332 Conclusion

333 Fleas are parasitic insects that can transmit many serious pathogens (i.e. bubonic plague,
334 endemic and murine typhus). The lack of flea genome assemblies has hindered research,

335 especially comparisons to other disease vectors. Here we combined Illumina and PacBio
336 sequencing with Hi-C scaffolding techniques to generate a chromosome-level genome assembly
337 for the cat flea, *Ctenocephalides felis*. Our work has revealed a genome characterized by
338 inordinate copy number variation (~38% of proteins) and a broad range of genome size estimates
339 (433-551 Mb) for individual fleas, suggesting a bizarre genome in flux. Surprisingly, the flea
340 genome exhibits neither inflation due to rampant gene duplication nor reduction due to their
341 parasitic lifestyle. Based on these results, as well as the nature and distribution of the gene
342 duplications themselves, we posit a dual mechanism of unequal crossing-over and gene
343 conversion may underpin this genome variability, although the biological significance remains to
344 be explored. Coupled with paradoxical co-infection with novel *Wolbachia* endosymbionts and
345 reproductive parasites, these oddities highlight a unique and underappreciated human disease
346 vector.
347

348 # Methods

349 **Experimental design**

350 This study was undertaken to generate a high-quality reference genome assembly and annotation
351 for the cat flea, *C. felis*, and represents the first sequenced genome for a member of Order
352 Siphonaptera. Our approach leveraged a combination of long-read PacBio sequencing, short-
353 read Illumina sequencing, and Hi-C (Chicago and HiRise) data to construct a chromosome-level
354 assembly; RNA-seq data and BLAST2GO classifications to assist in gene model prediction and
355 annotation; sequence mapping to address assembly fragmentation and short scaffolds (<1Mb);
356 and ortholog group construction to explore a genetic basis for the cat flea's parasitic lifestyle.
357 Gene duplications were confirmed via orthogonal approaches, including genome size estimates

358 of individual fleas, gene-based read coverage calculations, genomic distance between
359 duplications, and correlation between duplications and repeat elements or contig boundaries.
360

361 **Genome Sequencing and Assembly**

362 Newly emerged (August 2017), unfed female *C. felis* (n = 250) from Elward Laboratories (EL;
363 Soquel, CA) were surface-sterilized for 5 min in 10% NaClO followed by 5 min in 70% C_2H_5OH
364 and 3X washes with sterile phosphate-buffered saline. Fleas were flash-frozen in liquid N_2 and
365 ground to powder with sterile mortar and pestle. High-molecular weight DNA was extracted
366 using the MagAttract HMW DNA Kit (Qiagen; Venlo, Netherlands), quantified using a Qubit
367 3.0 fluorimeter (Thermo-Fisher Scientific; Waltham, MA), and assessed for quality on a 1.5%
368 agarose gel. DNA (50 µg) was submitted to the Institute for Genome Sciences (University of
369 Maryland) for size-selection and preparation of sequencing libraries. Libraries were sequenced
370 on 12 SMRT cells of a PacBio Sequel (Pacific Biosciences; Menlo Park, CA), generating
371 7,239,750 reads (46.7 Gb total). Raw reads were corrected, trimmed, and assembled into 16,622
372 contigs with Canu v1.5 in "pacbio-raw" mode, using an estimated genome size of 465 Mb [30].
373 A second group of newly emerged (January 2016), unfed female EL fleas (n=100) was surface-
374 sterilized and homogenized as above, and genomic DNA extracted using the QIAgen DNeasy®
375 Blood and Tissue Kit (QIAgen, Hilden, Germany). DNA was submitted to the WVU Genomics
376 Core for the preparation of a paired-end 250bp sequencing library with an average insert size of
377 500bp. The library was sequenced on 4 lanes of an Illumina HiSeq 1500 (Illumina Inc.; San
378 Diego, CA), generating 450,132,548 reads which were subsequently trimmed to remove adapters
379 and filtered for length and quality using FASTX-Toolkit v0.0.14 (available from
380 http://hannonlab.cshl.edu/fastx_toolkit/). These short read data were used to polish the Canu

assembly with Pilon v1.1.6 in "fix-all" mode [61], and to determine the composition of the *C. felis* microbiome (see below). Haplotigs in the polished contigs were resolved using purge_haplotigs [31] with coverage settings of 5 (low), 65 (mid), and 180 (high). A third group of newly-emerged (Feburary 2018), unfed female EL fleas (n = 200) were surface-sterilized as above, frozen at -80°C, and submitted for Chicago and Dovetail Hi-C proximity ligation (Dovetail Genomics, Santa Cruz, CA) [62] using the polished Canu assembly as a reference. The resulting scaffolded assembly (3,926 scaffolds) was subjected to removal of microbial sequences as described in the next section.

Genome Decontamination

A comparative BLAST-based pipeline slightly modified from our prior work [63] was used to identify and remove microbial scaffolds before annotation. Briefly, polished contigs were queried using BLASTP v2.2.31 against two custom databases derived from the nr database at NCBI (accessed July 2018): (1) all eukaryotic sequences (eukDB), and (2) combined archaeal, bacterial, and viral sequences (abvDB). For each query, the top five unique subject matches (by bitscore) in each database were pooled and scored according to a comparative sequence similarity measure, S_m:

$$S_m = bIQ$$

where b is the bitscore of the match; I is the percent identity; and Q is the percent aligned based on the longer of the two sequences. The top 5 scoring matches from the pooled lists of subjects were used to calculate a comparative rank score C for each individual query q against each database d:

404 $$C(q,d) = \frac{2(\sum_{n}^{i=1}(n - r_i(q,d)) + 1)}{n(n+1)}$$

405 where $r_i(q,d)$ is the rank of subject i for query q against database d. For example, if all of the
406 top n matches for query q are in eukDB then $C(q,$eukDB$) = 1$; conversely, if none of the top n
407 matches are in database abvDB then $C(q,$abvDB$) = 0$. Finally, each query q was scored
408 according to a comparative pairwise score P between 1 (purely eukaryotic) and -1 (purely
409 microbial):

410 $$P = C(q, eukDB) - C(q, abvDB)$$
411

412 Scaffolds that contained no contigs with $P > 0.3$ (n = 183), including 5 *Wolbachia*-like
413 scaffolds, were classified "not eukaryotic" and set aside. Scaffolds that contained contigs with a
414 range of P scores (n = 32) were manually inspected to identify and remove scaffolds arising from
415 mis-assembly or contamination (n = 10). The remaining scaffolds (n = 3,733) comprised the
416 initial draft assembly for *C. felis* and were deposited in NCBI under the accession ID
417 GCF_003426905.1.
418

419 **Genome Annotation**

420 Assembled and decontaminated scaffolds were annotated with NCBI Eukaryotic Genome
421 Annotation Pipeline (EGAP) v8.1 (https://www.ncbi.nlm.nih.gov/books/NBK143764/). To
422 facilitate gene model prediction, we generated RNA-seq data from 6 biological replicates of
423 pooled *C. felis* females (Heska Corporation, Fort Collins, CO). Briefly, total RNA was isolated
424 and submitted to the WVU Genomics Core for the preparation of paired-end, 100 bp sequencing
425 libraries using ScriptSeq Complete Gold Kit for Epidemiology (Illumina Inc., San Diego, CA).
426 Barcoded libraries were sequenced on 2 lanes of an Illumina HiSeq 1500 in High Throughput

427 mode, yielding approximately 26 million reads per sample (Q > 30). Raw sequencing reads from
428 all 6 samples were deposited in NCBI under the BioProject accession PRJNA484943. In addition
429 to these data, the EGAP pipeline also integrated previously-published *C. felis* expression data
430 from the 1KITE project (accession SRX314844; [2]) and an unrelated EST library (Biosample
431 accession SAMN00161855). The final set of annotations is available as "Ctenocephalides felis
432 Annotation Release 100" at the NCBI.
433

434 **Genome Completeness and Deflation**

435 The distribution of scaffold lengths in our assembly, together with the relatively large number of
436 fleas in our sequenced pool, warranted evaluating short scaffolds as possible sources of genomic
437 heterogeneity among individual fleas. To address this possibility, assembly scaffolds shorter
438 than 1 Mb (n = 3,724) were mapped to scaffolds larger than 1 Mb (n = 9; the BIG9) with BWA-
439 MEM v0.7.12 [64] using default parameters (**fig. S1A**). Additionally, genome completeness of
440 the full assembly compared to just the BIG9 scaffolds was assessed with Benchmarking Using
441 Single Copy Orthologs (BUSCO) v3.0.2 [28] in "protein" mode, using the *eukaryota_odb9*,
442 *arthropoda_odb9*, and *insecta_odb9* data sets (**fig. S1B**). Isoforms were removed before
443 BUSCO analysis by identifying CDSs that derived from the same protein-coding gene and
444 removing all but the longest sequence.
445

446 **Assessing the Extent of Gene Duplication**

447 Proteins encoded on the BIG9 scaffolds (n = 16,518) were queried against themselves with
448 BLASTP v2.2.31 using default parameters. Pairs of unique sequences that met or exceeded a
449 given amino acid percent identity (%ID) threshold over at least 80% of the query length were

450 binned together. Bins of sequence pairs that shared at least one sequence in common were
451 subsequently merged into clusters. Isoforms were removed after clustering by identifying CDSs
452 in a cluster that derived from the same protein-coding gene and removing all but the longest
453 sequence. This process was used to generate cluster sets at integer %ID thresholds from 90% to
454 100%. These duplicate protein-encoding genes were then mapped onto each of the BIG9
455 scaffolds using Circos [65] (**fig. S1**, **C** to **K**). Cluster diameters were calculated as the number of
456 non-cluster genes that lie between the edges of the cluster (*i.e.*, the two cluster genes that are
457 farthest apart on the scaffold) (**fig. S1L**). Clusters that span multiple scaffolds (mapped across
458 all BIG9 scaffolds in **fig. S1M**) defy an accurate calculation of diameter and were assigned a
459 cluster diameter of -1. In order to estimate the fraction of our assembly comprising gene
460 duplications, cluster coverages (by %ID threshold) were calculated in three ways. First, the
461 *coverage by CDS* was estimated by comparing the number of single-copy (protein-encoding)
462 genes to the total number of clusters; the latter number is assumed to represent a theoretical set
463 of minimal "seed" sequences. Second, the *coverage by gene length* was calculated as the total
464 number of nucleotides encoding the proteins in each cluster (including introns and exons) minus
465 the mean gene length (to account for a hypothetical "ancestor" gene). Finally, the *coverage by*
466 *genome region* was estimated by adding i*(n-1) to each calculation of coverage by gene length,
467 where n is the number of genes in the cluster and i is the mean intergenic length across all BIG9
468 scaffolds (17,344 nt). In order to assess possible enrichment of cellular functions among
469 duplicated genes, clusters at the 90% ID level were compared to the remaining BIG9 proteins by
470 Fisher's Exact Test (corrected for multiple testing) which is integrated into the FatiGO package
471 of BLAST2GO (see section "*Functional Classification of* C. felis *Proteins*" below). GO
472 categories were reduced to their most specific terms whenever possible.

473

474 **Length Variation Within Gene Duplication Clusters**

475 Variability in intra-cluster CDS length was assessed in two ways. First, the length of each CDS
476 in a cluster was compared to the longest CDS of the cluster, and the proportion of clusters with
477 any truncation (>1 AA) was calculated for each integer %ID threshold between 90 and 100% ID.
478 Second, the mean and distribution of length differences (i.e., the extent of truncation) was
479 calculated across all clusters for each integer %ID threshold between 90 and 100% ID.

480

481 **Analysis of Repeat Regions**

482 The extent and composition of repeat elements in the *C. felis* genome were assessed in two ways.
483 First, proteins annotated in the GO category "DNA Integration GO:0015074" (including
484 retrotransposons) were extracted, plotted by genomic coordinate on each BIG9 scaffold, and
485 assessed for co-localization either with gene duplicates (see above) or near the ends of scaffolds
486 (**fig. S1N**). Second, repeat elements were identified on the BIG9 scaffolds with RepeatMasker
487 v4.0.9 (available from http://www.repeatmasker.org/) in "RMBlast" mode (species
488 "holometabola"), using Tandem Repeat Finder v4.0.9 and the Repbase RepeatMasker (October
489 2018) and Dfam 3.0 databases (**fig. S1O**).

490

491 **Codon Usage and tRNA Gene Family Analysis**

492 Given the relatively large number of tRNA genes in our assembly, and the AT richness of our
493 genome, we were interested in exploring connections between tRNA gene frequencies and codon
494 usage. To this end, tRNA gene abundance on BIG9 scaffolds (n = 4,358) was determined by
495 binning genes into families according to their cognate amino acid and calculating the percent of

496 each family compared to the total number of tRNA genes (**fig. S1P**). A similar approach was
497 taken to quantify tRNA gene abundance by anticodon. TA richness of each anticodon was
498 subsequently calculated as the percent of A+T bases in the anticodon corrected for the size of the
499 tRNA family. Codon usage was calculated as the percent of total codons using the coding
500 sequences for genes on the BIG9 scaffolds, with isoforms removed as described previously (**fig.**
501 **S1Q**).
502

503 **Functional Classification of *C. felis* Proteins**

504 Protein sequences encoded on the BIG9 scaffolds (n = 16,518) were queried with BLASTP
505 v2.2.31 against the nr database of NCBI (accessed July 2018) using a maximum e-value
506 threshold of 0.1. The top 20 matches to each *C. felis* sequence were used to annotate queries
507 with Gene Ontology (GO) categories, Enzyme Classification (EC) codes, and protein domain
508 information using BLAST2GO v1.4.4 [66] under default parameters. A local instance of the GO
509 database (updated February 2019) was used for GO classification, and the online version of
510 InterPro (accessed April 2019) was used for domain discovery, including InterPro, PFAM,
511 SMART, PANTHER, PHOBIUS, and GENE3D domains; PROSITE profiles; SignalP-TM
512 (signal peptide) domains; and TMHMM (transmembrane helix) domains. InterPro data was used
513 to refine GO annotations whenever possible (**table S1**). A subset of *C. felis* proteins (n = 153)
514 classified as "DNA repair" (GO:0006281) was identified and all child GO terms of these proteins
515 tabulated (**table S1**). Assuming a linear relationship between genome size and number of repair
516 genes [67], we estimate *C. felis* has an enriched repertoire closer to that of a 3 Gb genome.
517

518 **Genome Size Estimation**

519 Estimations for flea genome size largely followed previously reported approaches [68]. For *C.*
520 *felis* individuals, 1/20 of the flea head was combined with two standards: 1/20 of the head of a
521 female (YW) *Drosophila melanogaster* (1C = 175 Mbp) and 1/20 of the head of a lab strain *D.*
522 *virilis* female (1C = 328). The tissues were placed in 1ml of cold Galbraith buffer and ground to
523 release nuclei in a 2ml Kontes Dounce, using 15 strokes of the "A" pestle at a rate of three
524 strokes every two seconds. The resulting solution was strained through a 45µ filter, stained for 3
525 hours in the dark at 4°C with 25µl of propidium iodide, then scored for total red fluorescence
526 using a Beckman-Coulter CytoFLEX flow cytometer. The average channel number of the 2C
527 nuclei of the sample and standards were determined using the CytExpert statistical software.
528 Briefly, the amount of DNA was estimated as the ratio of the average red fluorescence of the
529 sample to the average red fluorescence of the standard multiplied by the amount of DNA (in
530 Mbp) of the standard. The estimates from the two standards were averaged. At least 500 nuclei
531 were counted in each sample and standard peak. The coefficients of variation (CV) for all peaks
532 were < 2.0. Fluorescence activation and gating based on scatter were used to include in each
533 peak only intact red fluorescent nuclei free of associated cytoplasmic or broken nuclear tags.
534 Histograms generated for the largest and smallest determined genome sizes show the minimal
535 change in position for the two standards, demonstrating the significant change in the relative
536 fluorescence (average 2C channel number) between *C. felis* individuals (**fig. S2**).
537

538 **Characterizing Copy Number Variation**

539 In order to test the hypothesis that our genome assembly represents an agglomeration of
540 individuals with different levels of gene duplication, we used minimap2 [69] to map our short-
541 read sequence data against the full scaffolded assembly. After extracting the mapped reads with

542 samtools v0.1.19 [70], including primary and alternative mapping loci, a vector of sequence
543 depth (in bases) per position was generated with the genomecov function of bedtools v2.25.0
544 [71]. Mean depths for all 16,518 protein-coding genes on the BIG9 scaffolds were calculated as
545 total bases covering each gene divided by gene length. Finally, the mean depth across all
546 duplicated genes was compared to the mean depth across all single-copy genes using a Student's
547 t-test.

548 To evaluate the extent of gene duplication across different *C. felis* populations, reads from
549 the 1KITE transcriptome sequencing project (NCBI Sequence Read Archive accession
550 SRR921588) were mapped to the 3,733 scaffolds from our assembly using HISAT2 v2.1.0 [72]
551 under the --dta and --no_unal options. Mapped reads were sorted with samtools and abundance
552 per gene calculated as transcripts per million reads (TPM) using stringtie v1.3.4d [72]. TPM
553 values were binned and plotted against the number of duplicated (90% aa ID or higher) and
554 single-copy genes in the BIG9 assembly.

555

556 **Comparative Genomics**

557 Protein sequences (n=1,077,182) for 51 sequenced holometabolan genomes were downloaded
558 directly from NCBI (n=47) or VectorBase (n=3) or sequenced here (n=1). Isoforms were
559 removed before analysis wherever possible, by identifying CDSs that derived from the same
560 protein-coding gene and removing all but the longest CDS. Genome completeness was
561 estimated with BUSCO v3.0.2 in "protein" mode, using the *insecta_odb9* data set. Ortholog
562 groups (OGs; n=50,118) were constructed in three sequential phases: 1) CD-HIT v4.7 [73] in
563 accurate mode (-g 1) was used to cluster sequences at 50% ID; 2) PSI-CD-HIT (accurate mode,
564 local identity, alignment coverage minimum of 0.8) was used to cluster sequences at 25% ID; 3)

565 clusters were merged using clstr_rev.pl (part of the CD-HIT package). Proteins from *C. felis* that
566 did not cluster into any OG (n=4,282) were queried with BLASTP v2.2.31 against the nr
567 database of NCBI (accessed July 2018). Queries (n=2,170) with a top hit to any Holometabola
568 taxon, at a minimum %ID of 25% and query alignment of 80%, were manually added to the
569 original set of ortholog groups where possible (n=2,142) or set aside where not (n=28). The
570 remaining queries with at least one match in nr (n=1,318) were grouped by GO category level 4
571 and manually inspected; these included queries with top hits to Holometabolan taxa that did not
572 meet the minimum %ID or query coverage thresholds. Finally, *C. felis* proteins with no match in
573 nr (n=766) were binned by query length. These last two sets (n=2,084) comprise the set of
574 proteins unique to *C. felis* among all other Holometabola (**table S3**). Congruence between OG
575 clusters and taxonomy was determined by calculating a distance (Euclidean) between each pair
576 of taxa based on the number of shared OGs. The resulting matrix was scaled by classic
577 multidimensional scaling with the cmdscale function of R v3.5.1 [74], and visualized using the
578 ggplot package in R. Finally, pan-genomes were calculated for several key subsets of
579 Holometabola: 1) *C. felis* alone (Siphonaptera); 2) Antliophora (Siphonaptera and Diptera); 3)
580 Panorpida (Siphoanptera, Diptera, and Coleoptera); 4) all taxa except Hymenoptera; and 5) all
581 Holometabola (**table S2**). In order to account for differences in genome assembly quality and
582 taxon sampling bias, we define the pan-genome here as the set of all OGs that contain at least
583 one protein *from at least one taxon* in a given order. These intersections were visualized as upset
584 plots using UpSetR v1.3.3 [75]. Intersections of various holometabolous taxa that lack *C. felis*
585 were computed to gain insight on possible reductive evolution in fleas (**fig. S3**, **table S2**).
586

587 **Microbiome Composition**

588 A composite *C. felis* microbiome was estimated using Kraken Metagenomics-X v1.0.0 [76], part
589 of the Illumina BaseSpace toolkit. Briefly, 105,256,391 PE250 reads from our short read data set
590 were mapped against the Mini-Kraken reference set (12-08-2014 version), resulting in 2,390,314
591 microbial reads (2.27%) that were subsequently assigned to best possible taxonomy (**table S4**).
592

593 **Assembly of *Wolbachia* Endosymbiont Genomes**

594 Corrected reads from the Canu assembly of *C. felis* were recruited using BWA-MEM v0.7.12
595 (default settings) to a set of concatenated closed *Wolbachia* genome sequences (n=15)
596 downloaded from NCBI (accessed February 2018). Reads that mapped successfully were
597 extracted with samtools v0.1.19 and assembled separately into seed contigs (n=22) with Canu
598 v1.5 using default settings. Gene models on these seed contigs were predicted using the Rapid
599 Annotation of Subsystems Technology (RAST) v2.0 server [77], yielding two small subunit
600 (16S) ribosomal genes that were queried with BLASTN against the nr database of NCBI to
601 confirm the presence of two distinct *Wolbachia* strains. Seed contigs were further analyzed by
602 %GC and top BLASTN matches in the nr database of NCBI, and binned into three groups: *C.*
603 *felis* mitochondrial (n=1), *C. felis* genomic (n=6), and *Wolbachia*-like (n=15) contigs. The
604 *Wolbachia*-like contigs were subsequently queried with BLASTN against the full *C. felis*
605 assembly (before decontamination). A single *Wolbachia*-like contig (tig00000005; wCfeJ)
606 containing one of the two distinct 16S genes was retrieved intact from the full assembly. It was
607 removed from the primary assembly and manually closed by aligning the contig ends with
608 BLASTN. Gaps in the aligned regions were resolved by mapping our short read data to the
609 contig with BWA-MEM (default settings) and manually inspecting the read pileups. Six
610 additional contigs were also retrieved intact from the full assembly; these were likewise removed

611 and manually stitched together using end-alignment and short read polishing, resulting in a
612 second closed *Wolbachia* genome (*w*CfeT). The remaining *Wolbachia*-like contigs (n=8) were
613 found to be fractions of much longer flea-like contigs; these were left in the primary *C. felis*
614 assembly. Both *w*CfeJ and *w*CfeT sequences were submitted to the RAST v2.0 server for gene
615 model prediction and functional annotation.
616

617 **Phylogenomics of *Wolbachia* Endosymbionts**

618 Protein sequences (n=66,811) for 53 sequenced *Wolbachia* genomes plus 5 additional
619 Anaplasmataceae (*Neorickettsia helminthoeca* str. Oregon, *Anaplasma centrale* Israel, *A.*
620 *marginale* Florida, *Ehrlichia chaffeensis* Arkansas, and E. *ruminantium Gardel*) were either
621 downloaded directly from NCBI (n=30), retrieved as genome sequences from the NCBI
622 Assembly database (n=13), contributed via personal communication (n=8; Michael Gerth,
623 Oxford Brookes University), or sequenced here (n=2) (**table S4**). For genomes lacking
624 functional annotations (n=15), gene models were predicted using the RAST v2.0 server (n=12)
625 or GeneMarkS-2 v1.10_1.07 (n=3; [78]). Ortholog groups (n=2,750) were subsequently
626 constructed using FastOrtho, an in-house version of OrthoMCL [79], using an expect threshold
627 of 0.01, percent identity threshold of 30%, and percent match length threshold of 50% for
628 ortholog inclusion. A subset of single-copy families (n=47) conserved across at least 52 of the
629 58 genomes were independently aligned with MUSCLE v3.8.31 [80] using default parameters,
630 and regions of poor alignment were masked with trimal v1.4.rev15 [81] using the "automated1"
631 option. All modified alignments were concatenated into a single data set (10,027 positions) for
632 phylogeny estimation using RAxML v8.2.4 [82], under the gamma model of rate heterogeneity

633 and estimation of the proportion of invariant sites. Branch support was assessed with 1,000
634 pseudo-replications. Final ML optimization likelihood was -183020.639712.
635
636 **Confirmation of the presence of *Wolbachia* in *C. felis***
637 To assess the distribution of *w*CfeJ and *w*CfeT in *C. felis*, individual fleas from the sequenced
638 strain (EL) and a separate colony (Heska) not known to be infected with *Wolbachia* were pooled
639 (n=5) by sex and colony, surface-sterilized with 70% ethanol, flash-frozen, and ground in liquid
640 N_2. Genomic DNA was extracted using the GeneJET Genomic DNA Extraction Kit (Thermo-
641 Fisher Scientific; Waltham, MA), eluted twice in 50µl of PCR-grade H_2O, and quantified by
642 spectrophotometry with a Nanodrop 2000 (Thermo-Fisher Scientific; Waltham, MA). 100ng of
643 DNA from each pool was used as template in separate 25 µl PCR reactions using AmpliTaq
644 Gold 360 (Thermo-Fisher Scientific; Waltham, MA) and primer pairs (400 nmoles each) specific
645 for: 1) a 76nt fragment of the *cinA* gene specific to *w*CfeJ (Fwd: 5'-
646 AGCAACACCAACATGCGATT-3'; Rev: 5'- GAACCCCAGAGTTGGAAGGG-3'); 2) a 75nt
647 fragment of the *apaG* gene specific to *w*CfeT (Fwd: 5'- GCCGTCACTGGCAGGTAATA-3';
648 Rev: 5'- GCTGTTCTCCAATAACGCCA-3'); or 3) a 122nt fragment of *Wolbachia* 16S rDNA
649 (Fwd: 5'- CGGTGAATACGTTCTCGGGTY-3'; Rev: 5'- CACCCCAGTCACTGATCCC-3').
650 Primer specificities were confirmed with BLASTN against both the *C. felis* assembly and the nr
651 database of NCBI (accessed June 2018). Reaction conditions were identical for all primer sets:
652 initial denaturation at 95°C for 10 min, followed by 40 cycles of 95°C for 30 sec, 60°C for 30
653 sec, and 72°C for 30 sec, and a final extension at 72°C for 7 min. Products were run on a 2%
654 agarose gel and visualized with SmartGlow Pre Stain (Accuris Instruments; Edison, NJ).

655 Primers were tested before use by quantitative real-time PCR on a CFX Connect (Bio-Rad
656 Laboratories; Hercules, CA).
657
658 **Statistical Analysis**
659 Statistical analyses were carried out in R v3.5.1. Mean coverages across duplicated (n=7852) and
660 single-copy (n=7061) genes at the 90% ID threshold were compared for significance using a
661 Welch Two Sample t-test (unpaired, two-tailed) with 12,930 degrees of freedom and a p-value <
662 $2.2x10^{-16}$. Mean coverage of duplicated genes at %ID thresholds from 85-100% were compared
663 for significance using one-way Analysis of Variance (ANOVA) with 15 degrees of freedom and
664 a p-value = 0.2. A similar ANOVA was used to compare single-copy genes at 85-100% ID
665 thresholds, with a p-value < $2.2x10^{-16}$.
666
667 **Data and Scripts**
668 Data generated for this project that is not published elsewhere, including BLAST2GO
669 annotations and OG assignments, as well as custom analysis scripts, are provided on GitHub in
670 the "cfelis_genome" repository (available from https://github.com/driscollmml/cfelis_genome).
671
672 **Declarations**
673 ***Ethics approval and consent to participate.*** Not applicable.
674 ***Consent for publication.*** Not applicable.
675 ***Availability of data and materials.*** All of the sequence data generated for this work are available
676 at the NCBI under Bioproject accessions PRJNA489463 (genome sequence and annotation) and
677 PRJNA484941 (RNA-seq data used to support annotation). Additional tables with GO

annotations, ortholog groups, and microbiome data, as well as scripts used to generate data visualizations can be accessed at https://www.github.com/driscollmml/cfelis_genome. Sequences for *w*CfeT and *w*CfeJ are available on NCBI under Bioproject PRJNA622233.

Competing interests. The authors declare that they have no competing interests.

Funding. Research reported in this publication was supported by the National Institute of Health (NIH)/National Institute of Allergy and Infectious Diseases (NIAID) grants R01AI017828 and R01AI126853 to AFA, R21AI26108 and R21AI146773 to JJG & MSR, and R01AII122672 to KRM. KER-B and MLG were supported in part by the NIH/NIAID Grants T32AI095190 (Signaling Pathways in Innate Immunity) and T32AI007540 (Infection and Immunity). TPD and VIV were supported by start-up funding provided to TPD by West Virginia University. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Authors' contributions. TPD, VIV, JJG, KRM, and AFA conceived this study and developed the overall experimental framework. TPD, VIV, and KRM isolated genomic DNA from fleas. MLG, KER-B, and MSR performed flea dissections and isolated RNA from flea midgut tissues. TPD, VIV, JJG, DH, and CGE conceptualized the strategies for assembly and annotation. TPD, VIV, and JJG performed quality control at various stages of the assembly, executed the overall analyses of flea annotation, performed the phylogenomics analyses, and analyzed the flea microbiome. JSJ estimated genome sizes for individual cat and rat fleas. All authors contributed to writing and review of the final manuscript, with TPD, VIV, JJG, JSJ, KRM, and AFA playing the key roles. All authors read and approved the final manuscript.

700 ***Acknowledgements.*** We would like to thank Luke Tallon and Lisa Sadzewicz (Institute for
701 Genome Sciences, University of Maryland, Baltimore) for assistance with PacBio sequencing,
702 Ryan Percifield (Genomics Core, West Virginia University) for Illumina sequencing, and Mark
703 Daly, Elizabeth Fournier, Shaune Hall, and Thomas Swale (Dovetail Genomics) for facilitating
704 Hi-C assembly. We acknowledge the generous gift of unpublished *Wolbachia* endosymbiont of
705 *Zootermopsis nevadensis* (wZoo), *Wolbachia* endosymbiont of *Mengenilla moldrzyki* (wMen),
706 and *Wolbachia* endosymbiont of *Ctenocephalides felis* (wCte) contigs from Michael Gerth
707 (University of Liverpool). We thank Dr. Joe Hinnebusch (National Institute of Allergy and
708 Infectious Diseases) for providing the rat fleas used for genome size estimation. We are grateful
709 to Dr. John Beckmann (Auburn University) for critical discussion regarding *Wolbachia* biology
710 and Magda Beier-Sexton (University of Maryland, Baltimore) for administrative support.
711

712 **References**

713 1. Rust MK, Dryden MW. The Biology, Ecology, and Management of the Cat Flea. Annu Rev
714 Entomol [Internet]. 1997 [cited 2019 Jan 20];42:451–73. Available from:
715 http://www.ncbi.nlm.nih.gov/pubmed/9017899
716 2. Misof B, Liu S, Meusemann K, Peters RS, Donath A, Mayer C, et al. Phylogenomics resolves
717 the timing and pattern of insect evolution. Science (80-) [Internet]. 2014 [cited 2017 Feb
718 2];346:763–7. Available from: http://www.sciencemag.org/cgi/doi/10.1126/science.1257570
719 3. Leulmi H, Socolovschi C, Laudisoit A, Houemenou G, Davoust B, Bitam I, et al. Detection of
720 Rickettsia felis, Rickettsia typhi, Bartonella Species and Yersinia pestis in Fleas (Siphonaptera)
721 from Africa. Vinetz JM, editor. PLoS Negl Trop Dis. Public Library of Science; 2014;8:e3152.
722 4. Eisen RJ, Gage KL. Transmission of flea-borne zoonotic agents. Annu Rev Entomol.

723 2012;57:61–82.

724 5. Perry RD, Fetherston JD. Yersinia pestis--etiologic agent of plague. Clin Microbiol Rev
725 [Internet]. 1997 [cited 2019 Jan 17];10:35–66. Available from:
726 http://www.ncbi.nlm.nih.gov/pubmed/8993858

727 6. Nikiforov V V., Gao H, Zhou L, Anisimov A. Plague: Clinics, Diagnosis and Treatment. Adv
728 Exp Med Biol [Internet]. 2016 [cited 2019 Jan 17]. p. 293–312. Available from:
729 http://www.ncbi.nlm.nih.gov/pubmed/27722868

730 7. Stenseth NC, Atshabar BB, Begon M, Belmain SR, Bertherat E, Carniel E, et al. Plague: past,
731 present, and future. PLoS Med [Internet]. 2008 [cited 2019 Jan 17];5:e3. Available from:
732 https://dx.plos.org/10.1371/journal.pmed.0050003

733 8. Bertagnoli S, Marchandeau S. Myxomatosis. Rev Sci Tech [Internet]. 2015 [cited 2019 Jan
734 17];34:549–56, 539–47. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26601455

735 9. McElroy KM, Blagburn BL, Breitschwerdt EB, Mead PS, McQuiston JH. Flea-associated
736 zoonotic diseases of cats in the USA: bartonellosis, flea-borne rickettsioses, and plague. Trends
737 Parasitol [Internet]. 2010 [cited 2019 Jan 17];26:197–204. Available from:
738 http://www.ncbi.nlm.nih.gov/pubmed/20185369

739 10. Votýpka J, Suková E, Kraeva N, Ishemgulova A, Duží I, Lukeš J, et al. Diversity of
740 Trypanosomatids (Kinetoplastea: Trypanosomatidae) Parasitizing Fleas (Insecta: Siphonaptera)
741 and Description of a New Genus Blechomonas gen. n. Protist [Internet]. 2013 [cited 2019 Jan
742 17];164:763–81. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24113136

743 11. Feldmeier H, Heukelbach J, Ugbomoiko US, Sentongo E, Mbabazi P, von Samson-
744 Himmelstjerna G, et al. Tungiasis—A Neglected Disease with Many Challenges for Global
745 Public Health. PLoS Negl Trop Dis [Internet]. 2014 [cited 2017 Jun 22];8:e3133. Available

746 from: http://www.ncbi.nlm.nih.gov/pubmed/25356978
747 12. Feldmeier H, Keysers A. Tungiasis – A Janus-faced parasitic skin disease. Travel Med Infect
748 Dis [Internet]. 2013 [cited 2019 Jan 17];11:357–65. Available from:
749 http://www.ncbi.nlm.nih.gov/pubmed/24211240
750 13. Millán J. Comments on the manuscript by Bitam et al., 'Fleas and flea-borne diseases.' Int J
751 Infect Dis [Internet]. 2011 [cited 2019 Jan 17];15:e219. Available from:
752 http://www.ncbi.nlm.nih.gov/pubmed/21193340
753 14. Krasnov BR. Functional and evolutionary ecology of fleas : a model for ecological
754 parasitology [Internet]. [cited 2019 Jan 17]. Available from:
755 https://www.cambridge.org/vi/academic/subjects/life-sciences/entomology/functional-and-
756 evolutionary-ecology-fleas-model-ecological-parasitology?format=HB
757 15. Mullen GR, Durden LA. Medical and veterinary entomology. Elsevier; 2009.
758 16. Hinkle NC, Koehler PG. Cat Flea, Ctenocephalides felis felis Bouché (Siphonaptera:
759 Pulicidae). In: Capinera JL, editor. Encycl Entomol [Internet]. Dordrecht: Springer Netherlands;
760 2008. p. 797–801. Available from: https://doi.org/10.1007/978-1-4020-6359-6_536
761 17. Halos L, Beugnet F, Cardoso L, Farkas R, Franc M, Guillot J, et al. Flea control failure?
762 Myths and realities. Trends Parasitol [Internet]. 2014 [cited 2019 Jan 18];30:228–33. Available
763 from: http://www.ncbi.nlm.nih.gov/pubmed/24661796
764 18. Rust M. The Biology and Ecology of Cat Fleas and Advancements in Their Pest
765 Management: A Review. Insects [Internet]. 2017 [cited 2018 Oct 14];8:118. Available from:
766 http://www.ncbi.nlm.nih.gov/pubmed/29077073
767 19. Rennoll SA, Rennoll-Bankert KE, Guillotte ML, Lehman SS, Driscoll TP, Beier-Sexton M,
768 et al. The cat flea (Ctenocephalides felis) immune deficiency signaling pathway regulates

769 Rickettsia typhi infection. Infect Immun. 2018;86.

770 20. Böhm A, Meusemann K, Misof B, Pass G. Hypothesis on monochromatic vision in
771 scorpionflies questioned by new transcriptomic data. Sci Rep [Internet]. 2018 [cited 2019 Jan
772 18];8:9872. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29959337

773 21. Tolle MA. Mosquito-borne Diseases. Curr Probl Pediatr Adolesc Health Care [Internet].
774 2009 [cited 2019 Jan 20];39:97–140. Available from:
775 http://www.ncbi.nlm.nih.gov/pubmed/19327647

776 22. Glickman LT, Moore GE, Glickman NW, Caldanaro RJ, Aucoin D, Lewis HB. Purdue
777 University-Banfield National Companion Animal Surveillance Program for emerging and
778 zoonotic diseases. Vector Borne Zoonotic Dis [Internet]. 2006 [cited 2019 Jan 14];6:14–23.
779 Available from: https://www.liebertpub.com/doi/10.1089/vbz.2006.6.14

780 23. Bouhsira E, Franc M, Boulouis H-J, Jacquiet P, Raymond-Letron I, Liénard E. Assessment
781 of persistence of Bartonella henselae in Ctenocephalides felis. Appl Environ Microbiol
782 [Internet]. 2013 [cited 2019 Jan 14];79:7439–44. Available from:
783 http://aem.asm.org/lookup/doi/10.1128/AEM.02598-13

784 24. Nogueras MM, Pons I, Ortuño A, Miret J, Pla J, Castellà J, et al. Molecular detection of
785 Rickettsia typhi in cats and fleas. Munderloh, Ph.D. UG, editor. PLoS One [Internet]. 2013 [cited
786 2019 Jan 14];8:e71386. Available from: https://dx.plos.org/10.1371/journal.pone.0071386

787 25. Angelakis E, Mediannikov O, Parola P, Raoult D. Rickettsia felis: The Complex Journey of
788 an Emergent Human Pathogen. Trends Parasitol [Internet]. 2016 [cited 2019 Jan 14];32:554–64.
789 Available from: https://linkinghub.elsevier.com/retrieve/pii/S1471492216300125

790 26. Kichijo H. A note on the chromosomes of the flea, Ctenocephalus canis. Japanese J Genet.
791 1941;17.3:122–3.

792 27. Thomas C, Prasad RS. Chromosome variations in Xenopsylla astia Rothschild, 1911
793 (Siphonaptera). A preliminary report. Experientia [Internet]. Birkhäuser-Verlag; 1978 [cited
794 2019 Jun 10];34:1440–1. Available from: http://link.springer.com/10.1007/BF01932337
795 28. Seppey M, Manni M, Zdobnov EM. BUSCO: Assessing Genome Assembly and Annotation
796 Completeness. Methods Mol Biol [Internet]. 2019 [cited 2019 May 7]. p. 227–45. Available
797 from: http://www.ncbi.nlm.nih.gov/pubmed/31020564
798 29. Chan PP, Lowe TM. GtRNAdb 2.0: an expanded database of transfer RNA genes identified
799 in complete and draft genomes. Nucleic Acids Res [Internet]. 2016 [cited 2019 Jan
800 29];44:D184–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26673694
801 30. Hanrahan SJ, Johnston JS. New genome size estimates of 134 species of arthropods.
802 Chromosom Res [Internet]. 2011 [cited 2017 Feb 7];19:809–23. Available from:
803 http://www.ncbi.nlm.nih.gov/pubmed/21877225
804 31. Roach MJ, Schmidt SA, Borneman AR. Purge Haplotigs: allelic contig reassignment for
805 third-gen diploid genome assemblies. BMC Bioinformatics [Internet]. 2018 [cited 2019 Mar
806 29];19:460. Available from:
807 https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-018-2485-7
808 32. Chen X-G, Jiang X, Gu J, Xu M, Wu Y, Deng Y, et al. Genome sequence of the Asian Tiger
809 mosquito, *Aedes albopictus* , reveals insights into its biology, genetics, and evolution. Proc Natl
810 Acad Sci [Internet]. 2015 [cited 2019 Jun 17];112:E5907–15. Available from:
811 http://www.ncbi.nlm.nih.gov/pubmed/26483478
812 33. Poulin R, Randhawa HS. Evolution of parasitism along convergent lines: from ecology to
813 genomics. Parasitology [Internet]. 2015 [cited 2019 May 17];142:S6–15. Available from:
814 http://www.ncbi.nlm.nih.gov/pubmed/24229807

815 34. Lyons RE, Wong DCC, Kim M, Lekieffre N, Huson MG, Vuocolo T, et al. Molecular and
816 functional characterisation of resilin across three insect orders. Insect Biochem Mol Biol
817 [Internet]. 2011 [cited 2019 Feb 27];41:881–90. Available from:
818 https://linkinghub.elsevier.com/retrieve/pii/S0965174811001603
819 35. Su RS-C, Kim Y, Liu JC. Resilin: protein-based elastomeric biomaterials. Acta Biomater
820 [Internet]. 2014 [cited 2019 Mar 30];10:1601–11. Available from:
821 https://linkinghub.elsevier.com/retrieve/pii/S1742706113003310
822 36. Vasconcelos EJR, Billeter SA, Jett LA, Meinersmann RJ, Barr MC, Diniz PPVP, et al.
823 Assessing Cat Flea Microbiomes in Northern and Southern California by 16S rRNA Next-
824 Generation Sequencing. Vector-Borne Zoonotic Dis [Internet]. 2018 [cited 2018 Oct 14];18:491–
825 9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29893631
826 37. Lawrence AL, Hii S-F, Chong R, Webb CE, Traub R, Brown G, et al. Evaluation of the
827 bacterial microbiome of two flea species using different DNA-isolation techniques provides
828 insights into flea host ecology. Olson J, editor. FEMS Microbiol Ecol [Internet]. 2015 [cited
829 2017 Feb 9];91:fiv134. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26542076
830 38. Pornwiroon W, Kearney MT, Husseneder C, Foil LD, Macaluso KR. Comparative
831 microbiota of Rickettsia felis-uninfected and -infected colonized cat fleas, Ctenocephalides felis.
832 ISME J [Internet]. 2007 [cited 2017 Feb 9];1:394–402. Available from:
833 http://www.ncbi.nlm.nih.gov/pubmed/18043659
834 39. Sunyakumthorn P, Bourchookarn A, Pornwiroon W, David C, Barker SA, Macaluso KR.
835 Characterization and growth of polymorphic Rickettsia felis in a tick cell line. Appl Environ
836 Microbiol [Internet]. 2008 [cited 2018 Oct 14];74:3151–8. Available from:
837 http://aem.asm.org/cgi/doi/10.1128/AEM.00025-08

838 40. Gillespie JJ, Driscoll TP, Verhoeve VI, Utsuki T, Husseneder C, Chouljenko VN, et al.
839 Genomic Diversification in Strains of Rickettsia felis Isolated from Different Arthropods.
840 Genome Biol Evol [Internet]. 2015 [cited 2018 Apr 27];7:35–56. Available from:
841 http://www.ncbi.nlm.nih.gov/pubmed/25477419
842 41. González-Álvarez VH, de Mera IGF, Cabezas-Cruz A, de la Fuente J, Ortega-Morales AI,
843 Almazán C. Molecular survey of Rickettsial organisms in ectoparasites from a dog shelter in
844 Northern Mexico. Vet Parasitol Reg Stud Reports [Internet]. 2017;10:143–8. Available from:
845 http://www.sciencedirect.com/science/article/pii/S2405939017301442
846 42. Casiraghi M, Bordenstein SR, Baldo L, Lo N, Beninati T, Wernegreen JJ, et al. Phylogeny of
847 Wolbachia pipientis based on gltA, groEL and ftsZ gene sequences: clustering of arthropod and
848 nematode symbionts in the F supergroup, and evidence for further diversity in the Wolbachia
849 tree. Microbiology [Internet]. 2005 [cited 2018 Nov 19];151:4015–22. Available from:
850 http://www.ncbi.nlm.nih.gov/pubmed/16339946
851 43. Bordenstein SR, Bordenstein SR. Eukaryotic association module in phage WO genomes
852 from Wolbachia. Nat Commun [Internet]. 2016 [cited 2017 Aug 3];7:13155. Available from:
853 http://www.ncbi.nlm.nih.gov/pubmed/27727237
854 44. Gillespie JJ, Joardar V, Williams KP, Driscoll TP, Hostetler JB, Nordberg E, et al. A
855 Rickettsia genome overrun by mobile genetic elements provides insight into the acquisition of
856 genes characteristic of an obligate intracellular lifestyle. J Bacteriol [Internet]. 2012 [cited 2015
857 Jan 15];194:376–94. Available from:
858 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3256634&tool=pmcentrez&renderty
859 pe=abstract
860 45. Nikoh N, Hosokawa T, Moriyama M, Oshima K, Hattori M, Fukatsu T. Evolutionary origin

861 of insect-Wolbachia nutritional mutualism. Proc Natl Acad Sci U S A [Internet]. 2014 [cited
862 2016 Aug 23];111:10257–62. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24982177
863 46. Gerth M, Bleidorn C. Comparative genomics provides a timeframe for Wolbachia evolution
864 and exposes a recent biotin synthesis operon transfer. Nat Microbiol [Internet]. 2017 [cited 2019
865 Jan 10];2:16241. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28005061
866 47. Balvín O, Roth S, Talbot B, Reinhardt K. Co-speciation in bedbug Wolbachia parallel the
867 pattern in nematode hosts. Sci Rep [Internet]. 2018 [cited 2019 Jan 22];8:8797. Available from:
868 http://www.ncbi.nlm.nih.gov/pubmed/29891919
869 48. Penz T, Schmitz-Esser S, Kelly SE, Cass BN, Müller A, Woyke T, et al. Comparative
870 Genomics Suggests an Independent Origin of Cytoplasmic Incompatibility in Cardinium hertigii.
871 Moran NA, editor. PLoS Genet [Internet]. 2012 [cited 2017 May 5];8:e1003012. Available from:
872 http://www.ncbi.nlm.nih.gov/pubmed/23133394
873 49. Zeng Z, Fu Y, Guo D, Wu Y, Ajayi OE, Wu Q. Bacterial endosymbiont Cardinium cSfur
874 genome sequence provides insights for understanding the symbiotic relationship in Sogatella
875 furcifera host. BMC Genomics [Internet]. 2018 [cited 2019 Jan 22];19:688. Available from:
876 http://www.ncbi.nlm.nih.gov/pubmed/30231855
877 50. Ríhová J, Nováková E, Husník F, Hypša V. Legionella Becoming a Mutualist: Adaptive
878 Processes Shaping the Genome of Symbiont in the Louse Polyplax serrata. Genome Biol Evol
879 [Internet]. 2017 [cited 2018 Dec 13];9:2946–57. Available from:
880 http://academic.oup.com/gbe/article/9/11/2946/4562435
881 51. Ju J-F, Bing X-L, Zhao D-S, Guo Y, Xi Z, Hoffmann AA, et al. Wolbachia supplement
882 biotin and riboflavin to enhance reproduction in planthoppers. ISME J [Internet]. Nature
883 Publishing Group; 2019 [cited 2020 Jan 17];1–12. Available from:

884 http://www.nature.com/articles/s41396-019-0559-9

885 52. Chen H, Ronau JA, Beckmann JF, Hochstrasser M. A Wolbachia Nuclease and Its Binding
886 Partner Comprise a Novel Mechanism for Induction of Cytoplasmic Incompatibility. 2019;

887 53. Beckmann JF, Ronau JA, Hochstrasser M. A Wolbachia deubiquitylating enzyme induces
888 cytoplasmic incompatibility. Nat Microbiol [Internet]. 2017 [cited 2017 Jun 30];2:17007.
889 Available from: http://www.ncbi.nlm.nih.gov/pubmed/28248294

890 54. LePage DP, Metcalf JA, Bordenstein SR, On J, Perlmutter JI, Shropshire JD, et al. Prophage
891 WO genes recapitulate and enhance Wolbachia-induced cytoplasmic incompatibility. Nature
892 [Internet]. 2017 [cited 2017 Jun 30];543:243–7. Available from:
893 http://www.ncbi.nlm.nih.gov/pubmed/28241146

894 55. Beckmann JF, Fallon AM. Detection of the Wolbachia protein WPIP0282 in mosquito
895 spermathecae: Implications for cytoplasmic incompatibility. Insect Biochem Mol Biol [Internet].
896 2013 [cited 2018 Jan 16];43:867–78. Available from:
897 http://www.ncbi.nlm.nih.gov/pubmed/23856508

898 56. Gillespie JJ, Driscoll TP, Verhoeve VI, Rahman MS, Macaluso KR, Azad AF. A Tangled
899 Web: Origins of Reproductive Parasitism. Eric B, editor. Genome Biol Evol [Internet]. 2018
900 [cited 2018 Dec 14];10:2292–309. Available from:
901 http://www.ncbi.nlm.nih.gov/pubmed/30060072

902 57. Beckmann JF, Bonneau M, Chen H, Hochstrasser M, Poinsot D, Merçot H, et al. The Toxin–
903 Antidote Model of Cytoplasmic Incompatibility: Genetics and Evolutionary Implications. Trends
904 Genet [Internet]. 2019 [cited 2019 Jan 29]; Available from:
905 http://www.ncbi.nlm.nih.gov/pubmed/30685209

906 58. Lawrence AL, Webb CE, Clark NJ, Halajian A, Mihalca AD, Miret J, et al. Out-of-Africa,

907 human-mediated dispersal of the common cat flea, Ctenocephalides felis: The hitchhiker's guide
908 to world domination. Int J Parasitol [Internet]. 2019 [cited 2019 May 22];49:321–36. Available
909 from: http://www.ncbi.nlm.nih.gov/pubmed/30858050
910 59. Li Z, Tiley GP, Galuska SR, Reardon CR, Kidder TI, Rundell RJ, et al. Multiple large-scale
911 gene and genome duplications during the evolution of hexapods. Proc Natl Acad Sci [Internet].
912 2018 [cited 2019 Feb 5];115:201710791. Available from:
913 http://www.ncbi.nlm.nih.gov/pubmed/29674453
914 60. Blanton LS, Idowu BM, Tatsch TN, Henderson JM, Bouyer DH, Walker DH. Opossums and
915 Cat Fleas: New Insights in the Ecology of Murine Typhus in Galveston, Texas. Am J Trop Med
916 Hyg [Internet]. 2016 [cited 2017 Feb 13];95:457–61. Available from:
917 http://www.ajtmh.org/cgi/doi/10.4269/ajtmh.16-0197
918 61. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an
919 integrated tool for comprehensive microbial variant detection and genome assembly
920 improvement. Wang J, editor. PLoS One [Internet]. 2014 [cited 2019 Mar 27];9:e112963.
921 Available from: https://dx.plos.org/10.1371/journal.pone.0112963
922 62. Putnam NH, O'Connell BL, Stites JC, Rice BJ, Blanchette M, Calef R, et al. Chromosome-
923 scale shotgun assembly using an in vitro method for long-range linkage. Genome Res [Internet].
924 2016 [cited 2019 Mar 27];26:342–50. Available from:
925 http://www.ncbi.nlm.nih.gov/pubmed/26848124
926 63. Driscoll TP, Gillespie JJ, Nordberg EK, Azad AF, Sobral BW. Bacterial DNA sifted from the
927 Trichoplax adhaerens (Animalia: Placozoa) genome project reveals a putative rickettsial
928 endosymbiont. Genome Biol Evol [Internet]. 2013 [cited 2014 Dec 22];5:621–45. Available
929 from:

930 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3641634&tool=pmcentrez&renderty
931 pe=abstract
932 64. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform.
933 Bioinformatics [Internet]. 2009 [cited 2019 May 24];25:1754–60. Available from:
934 http://www.ncbi.nlm.nih.gov/pubmed/19451168
935 65. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an
936 information aesthetic for comparative genomics. Genome Res [Internet]. 2009 [cited 2019 Apr
937 16];19:1639–45. Available from: http://genome.cshlp.org/cgi/doi/10.1101/gr.092759.109
938 66. Gotz S, Garcia-Gomez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, et al. High-
939 throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res
940 [Internet]. 2008 [cited 2019 May 24];36:3420–35. Available from:
941 https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkn176
942 67. Voskarides K, Dweep H, Chrysostomou C. Evidence that DNA repair genes, a family of
943 tumor suppressor genes, are associated with evolution rate and size of genomes. Hum Genomics
944 [Internet]. 2019 [cited 2019 Jun 18];13:26. Available from:
945 https://humgenomics.biomedcentral.com/articles/10.1186/s40246-019-0210-x
946 68. Johnston JS, Bernardini A, Hjelmen CE. Genome size estimation and quantitative
947 cytogenetics in insects. In: Brown SJ, Pfrender ME, editors. Insect Genomics. New York:
948 Humana Press; 2019. p. 15–26.
949 69. Li H. Minimap2: pairwise alignment for nucleotide sequences. Birol I, editor. Bioinformatics
950 [Internet]. Narnia; 2018 [cited 2019 May 29];34:3094–100. Available from:
951 https://academic.oup.com/bioinformatics/article/34/18/3094/4994778
952 70. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence

953 Alignment/Map format and SAMtools. Bioinformatics [Internet]. 2009 [cited 2019 May
954 24];25:2078–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19505943
955 71. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features.
956 Bioinformatics [Internet]. Narnia; 2010 [cited 2019 May 29];26:841–2. Available from:
957 https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btq033
958 72. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of
959 RNA-seq experiments with HISAT, StringTie and Ballgown. Nat Protoc. Nature Publishing
960 Group; 2016;11:1650–67.
961 73. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation
962 sequencing data. Bioinformatics [Internet]. Oxford University Press; 2012 [cited 2019 May
963 29];28:3150–2. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23060610
964 74. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna,
965 Austria; 2018. Available from: https://www.r-project.org/
966 75. Gehlenborg N. UpSetR: A More Scalable Alternative to Venn and Euler Diagrams for
967 Visualizing Intersecting Sets [Internet]. 2017. Available from: https://cran.r-
968 project.org/package=UpSetR
969 76. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact
970 alignments. Genome Biol [Internet]. 2014 [cited 2017 Feb 8];15:R46. Available from:
971 http://www.ncbi.nlm.nih.gov/pubmed/24580807
972 77. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, et al. The RAST Server:
973 Rapid Annotations using Subsystems Technology. BMC Genomics. 2008;9:75.
974 78. Lomsadze A, Gemayel K, Tang S, Borodovsky M. Modeling leaderless transcription and
975 atypical genes results in more accurate gene prediction in prokaryotes. Genome Res [Internet].

2018 [cited 2019 May 24];28:1079–89. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29773659

79. Li L, Stoeckert CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res [Internet]. 2003 [cited 2017 Apr 23];13:2178–89. Available from: http://www.genome.org/cgi/doi/10.1101/gr.1224503

80. Edgar RC. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32:1792–7.

81. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics [Internet]. 2009 [cited 2019 Jan 10];25:1972–3. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp348

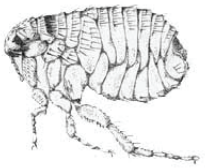
82. Stamatakis A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30:1312–3.

83. Giraldo-Calderón GI, Emrich SJ, MacCallum RM, Maslen G, Dialynas E, Topalis P, et al. VectorBase: an updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases. Nucleic Acids Res [Internet]. 2015 [cited 2018 Oct 15];43:D707-13. Available from: https://academic.oup.com/nar/article/43/D1/D707/2437683

Figures, tables and additional files

998 **Fig. 1. *C. felis* genome characteristics**. (**A**) Summary statistics for long-read sequencing,
999 assembly and gene annotation. (**B**) Of 16,518 total protein-encoding genes (BIG9 scaffolds),
1000 10,088 are derived from gene duplications (6,225 duplication events within 3,863 OGs at a
1001 threshold of 90% aa identity). (**C**) Assessment of the number of genes per duplication (*left*) and
1002 the relative distances between duplicate genes (*right*). Distances were computed only for true
1003 duplications (n=2 genes) at a threshold of 90% aa identity. (**D**) Gene duplications are enriched
1004 within BIG9 scaffolds (tandem and proximal, red numbers) versus across scaffolds (dispersed,
1005 black numbers). (**E**) Enriched cellular functions of duplicate genes relative to single-copy genes.
1006 (**F**) *C. felis* belongs to a minimal fraction of eukaryotes containing abundant tRNA genes. tRNA
1007 gene counts are shown for disease vectors (VectorBase [83]) and eukaryotes carrying over 1000
1008 tRNA genes (GtRNAdb [29]; ratios show number of genomes with > 1000 tRNA genes per
1009 taxon.
1010
1011

A



C. felis

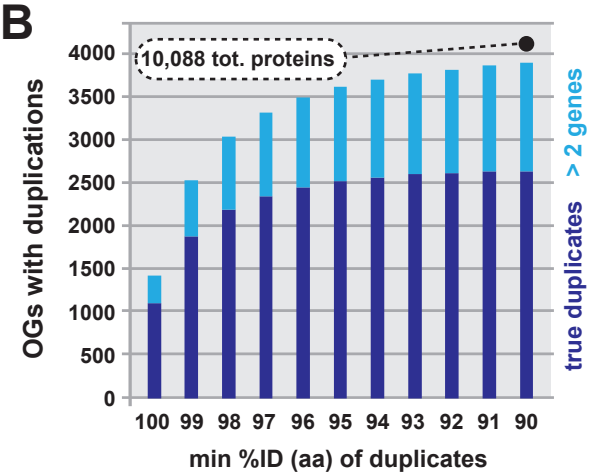
LONG-READ SEQUENCING

Tot. reads generated	7,239,750
Error-corrected reads	1,719,943
Estimated coverage	25X

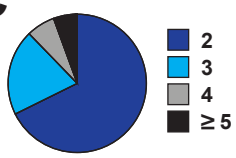
ASSEMBLY	ALL	BIG9
Tot. seq. length (Mb)	773.8	654.0
Contigs assembled	16,622	12,348
Contig N50 (Mb)	0.061	0.082
Longest contig (Mb)	1.9	1.9
Mean contig %GC	30.2	29.2
Scaffolds constructed	3,926	9
Scaffold N50 (Mb)	71.7	86.1
Scaffold L50	4	3
Longest scaffold (Mb)	185.5	185.5
No. scaffolds > 10Mb	9	9

ANNOTATION	ALL	BIG9
Scaffolds annotated	3,733	9
Tot. seq annotated (Mb)	763.8	654.0
No. total genes	26,844	23,558
No. protein coding genes	18,878	16,518
No. rRNA genes	466	184
No. tRNA genes	5,847	4,358
Mean tRNA length	74 (66-85)	74 (66-85)

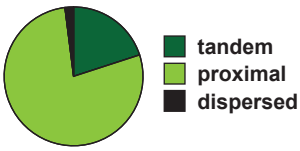
B



C



Genes per duplicate family



Duplicate gene proximity

D

Scaffold	9	8	7	6	5	4	3	2	1
9 (NW_020539727)	304								
8 (NW_020539724)	2	360							
7 (NW_020539726)	4	2	444						
6 (NW_020537758)	3	7	18	573					
5 (NW_020537646)	22	15	44	34	909				
4 (NW_020539725)	6	19	23	14	42	1120			
3 (NW_020537324)	5	2	45	19	76	21	1135		
2 (NW_020536999)	4	13	48	32	64	43	68	1320	
1 (NW_020538040)	23	5	91	47	137	52	81	137	1314

E

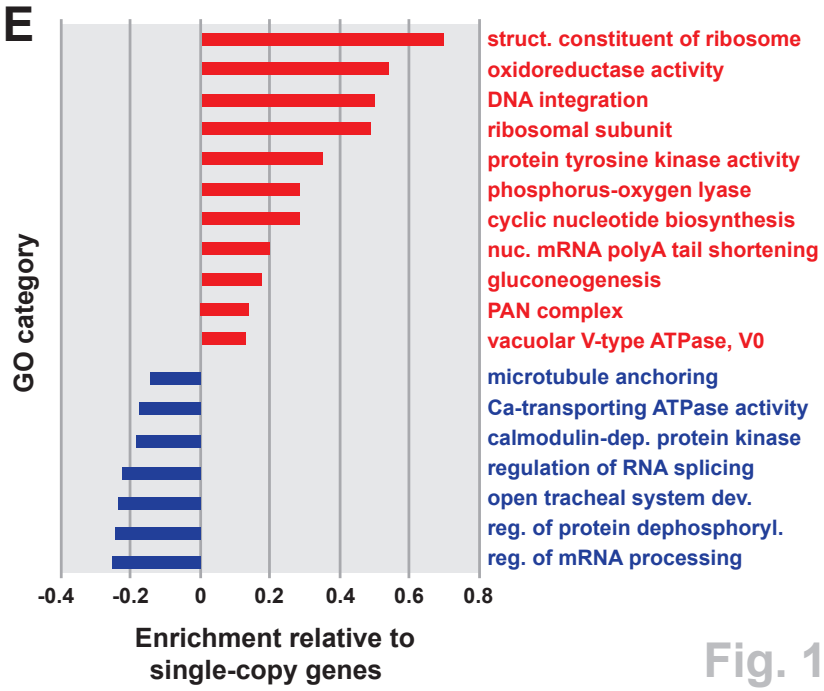
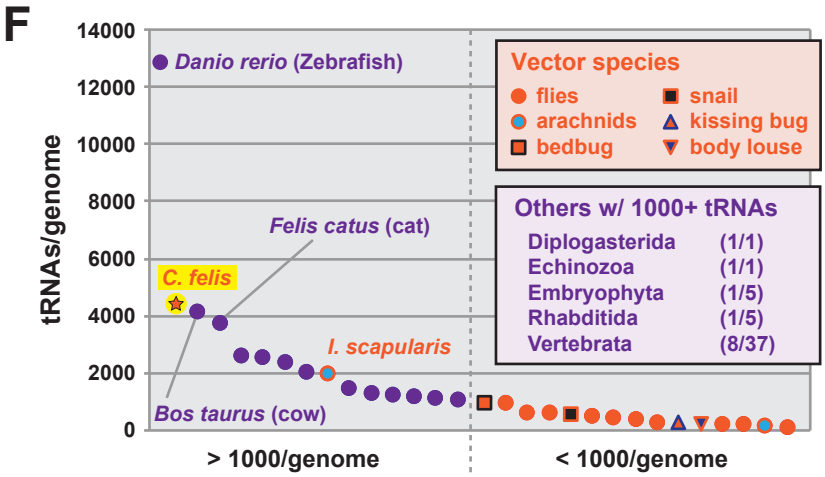


Fig. 1

F



1012 **Fig. 2. Evidence for excessive copy number variation in the *C. felis* genome.** (**A**) Flea
1013 genome size estimates. Flow cytometer-based estimates were performed for male and female
1014 individuals of *X. cheopis* (Texas), *C. felis* (Texas), and for female *C. felis* EL from the sequenced
1015 colony (see **fig. S2**). The inset (top right) depicts the coefficients of variation in measured
1016 fluorescence (relative fluorescence units; RFU) for *Drosophila melanogaster* (n=26), *D. viridis*
1017 (n=26), and *C. felis* EL (n=26) females prepared and analyzed simultaneously. (**B**) Graphic
1018 depiction of assembling CNV. Two theoretical individual fleas are shown with different CNVs
1019 for loci A and B. Regions unique to each individual genome are shown by the red dashed boxes.
1020 Only reads concordant between individuals are included in the conglomerate assembly. (**C**)
1021 Comparison of Illumina read coverage-mapping between duplicate genes (blue) and single-copy
1022 genes (green) at different %ID thresholds. Reads that mapped to multiple locations (alternative
1023 mappings) were included. Asterisks indicate statistically significant difference (Welch Two-
1024 Sample t-test, $p < 2.2e-16$) between mean coverage of single-copy and duplicate genes at the 90
1025 %ID threshold. (**D**) Transcriptional support for *C. felis* EL genes within the 1KITE
1026 transcriptomic data. Counts of transcripts per million reads (TPM) were mapped (Hisat2 &
1027 Stringtie), binned, and plotted against the number of duplicated (blue) and single-copy (green)
1028 genes in the BIG9 assembly. (**E**) Extent of truncation within clusters of duplicated genes in *C.*
1029 *felis*. The number of clusters with truncated members at each integer %ID threshold (left) was
1030 calculated as the proportion of total clusters at that threshold (center). The distribution of length
1031 differences in these clusters (relative to the longest member in each cluster) is plotted as a violin
1032 plot (right); black diamonds represent the mean length difference at each %ID threshold.
1033
1034

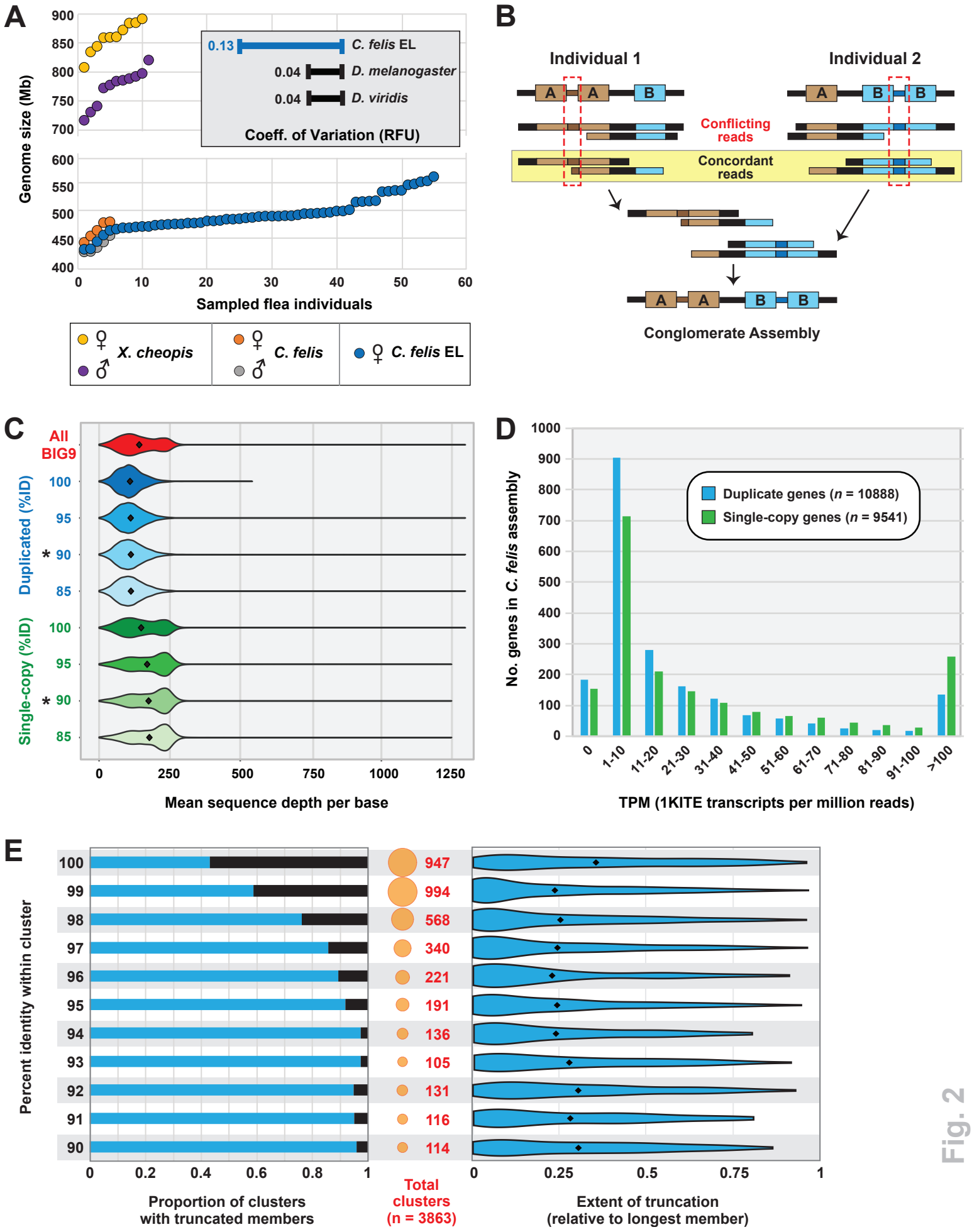
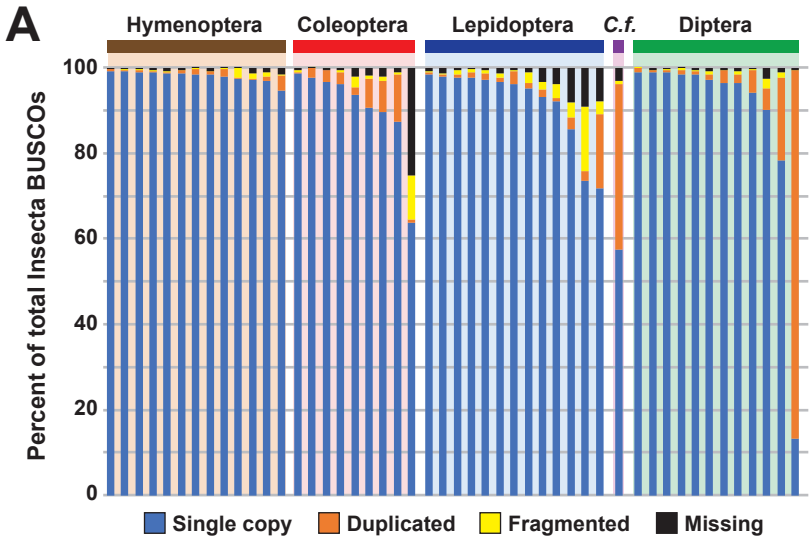
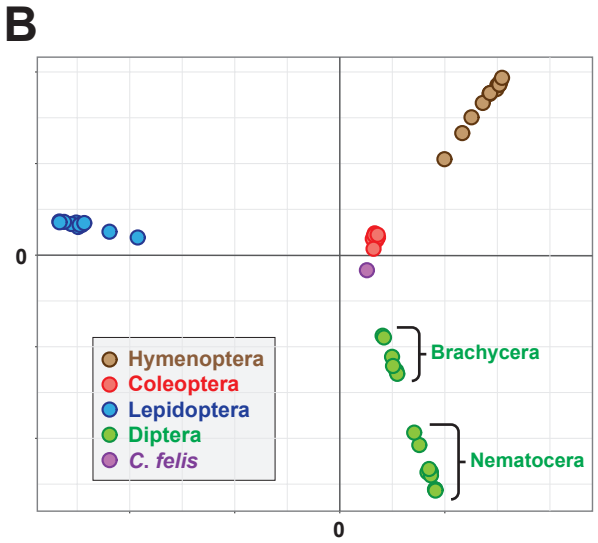

A
Genome size (Mb)
900
850
800
750
700
600
550
500
450
400
0.13 C. felis EL
0.04 D. melanogaster
0.04 D. viridis
Coeff. of Variation (RFU)
0
10
20
30
40
50
60
Sampled flea individuals
♀ ♂ X. cheopis
♀ ♂ C. felis
♀ C. felis EL
B
Individual 1
Individual 2
A
A
B
A
B
B
Conflicting reads
Concordant reads
A
A
B
B
Conglomerate Assembly
C
All BIG9
Duplicated (%ID)
100
95
* 90
85
Single-copy (%ID)
100
95
* 90
85
0
250
500
750
1000
1250
Mean sequence depth per base
D
No. genes in C. felis assembly
1000
900
800
700
600
500
400
300
200
100
0
Duplicate genes (n = 10888)
Single-copy genes (n = 9541)
0
1-10
11-20
21-30
31-40
41-50
51-60
61-70
71-80
81-90
91-100
>100
TPM (1KITE transcripts per million reads)
E
Percent identity within cluster
100
99
98
97
96
95
94
93
92
91
90
947
994
568
340
221
191
136
105
131
116
114
0
0.2
0.4
0.6
0.8
1
Proportion of clusters with truncated members
Total clusters (n = 3863)
0
0.25
0.5
0.75
1
Extent of truncation (relative to longest member)

Fig. 2

1035 **Fig. 3. Phylogenomics analysis of the *C. felis* genome**. (**A**) Assessing completeness and
1036 conservation of select holometabolan genomes using insect (n=1,658) Benchmarking Universal
1037 Single-Copy Orthologues (BUSCOs) [28]. (**B**) Multidimensional scaling plots gauging within-
1038 and across-order similarity of protein orthologous groups. Inset show color scheme for
1039 holometabolous orders. (**C**) Upset plot illustrating *C. felis* protein orthologous groups that
1040 intersect with other holometabolous insects. Inclusion criteria: one protein from at least one
1041 genome/order must be present. Yellow bar, 577 proteins found in all analyzed genomes. Inset,
1042 redrawn phylogeny estimation of Holometabola [2]; numbers indicate *C. felis* unique protein
1043 groups or higher-generic monophyletic groups (see **table S2**).
1044
1045


A
Hymenoptera
Coleoptera
Lepidoptera
C.f.
Diptera
Percent of total Insecta BUSCOs
100
80
60
40
20
0
Single copy
Duplicated
Fragmented
Missing


B
0
0
Hymenoptera
Coleoptera
Lepidoptera
Diptera
C. felis
Brachycera
Nematocera

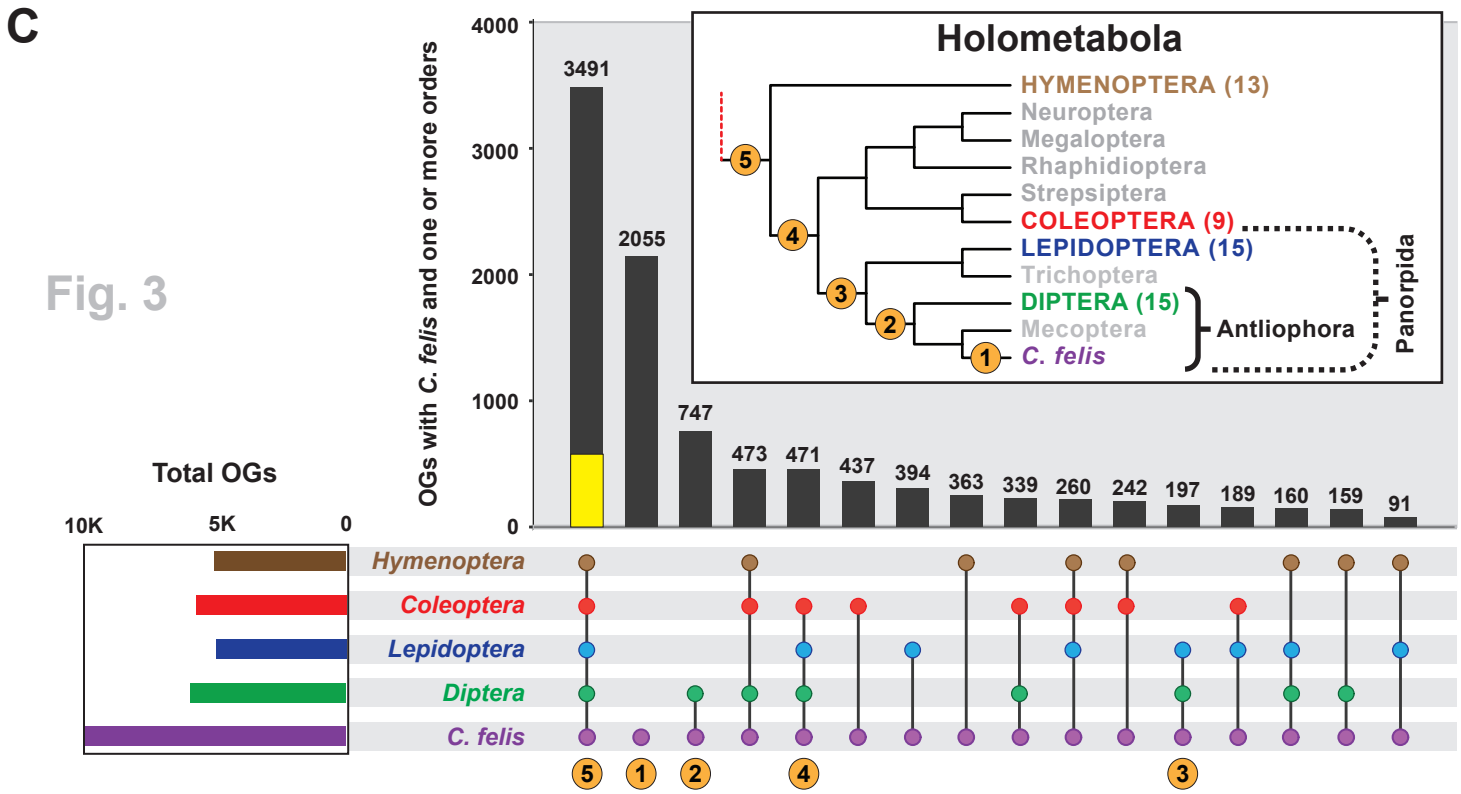
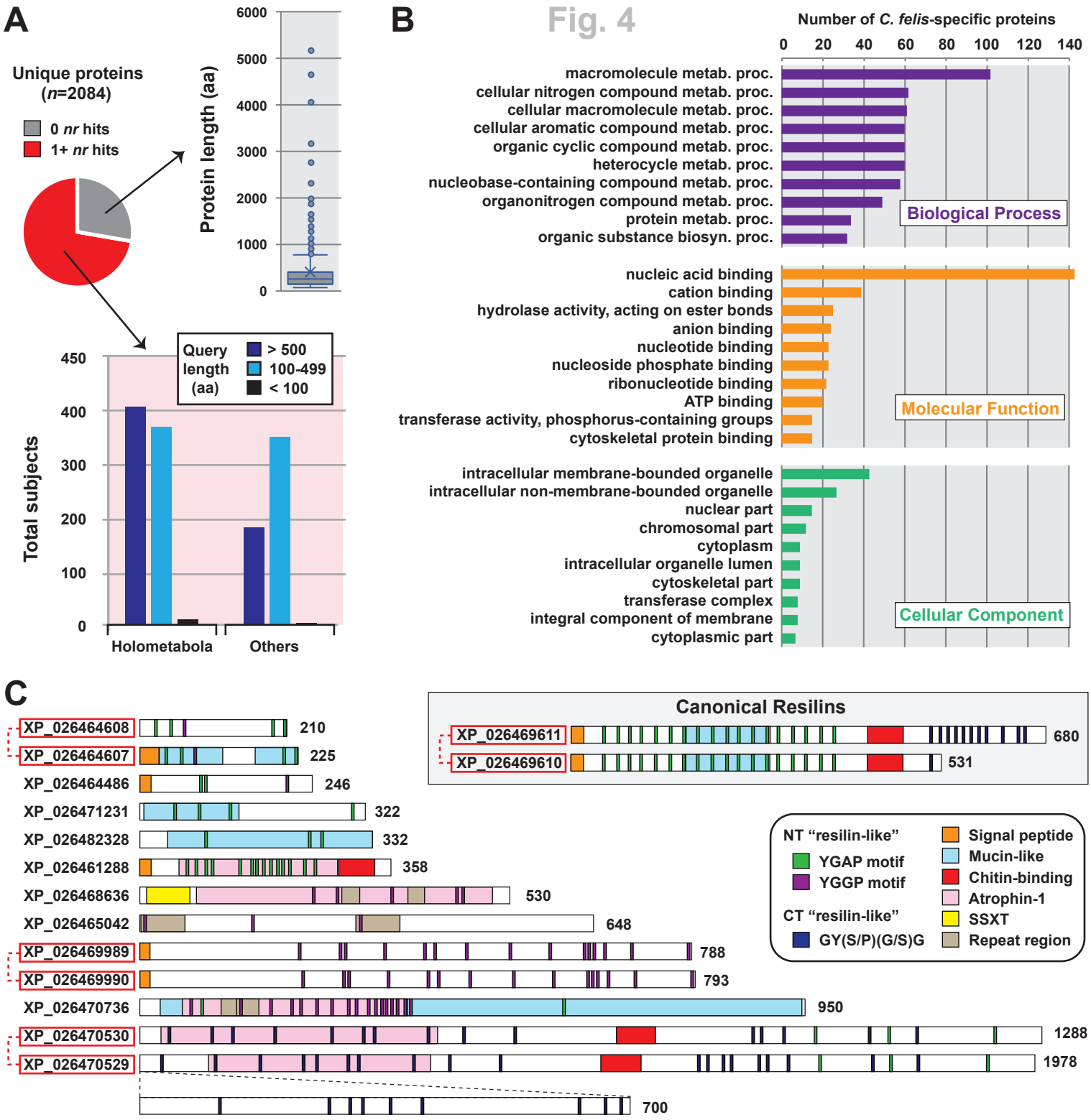
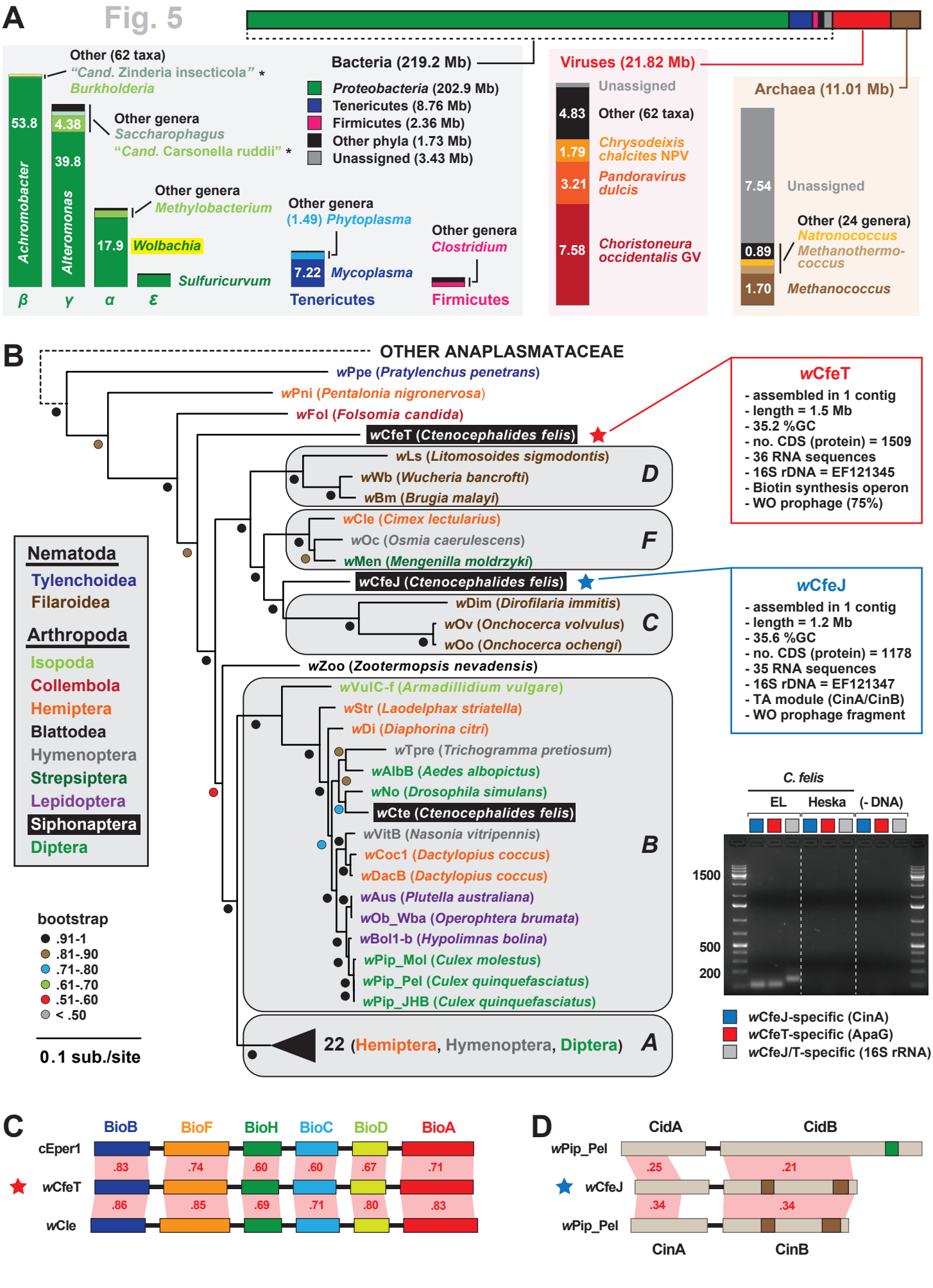

C
Holometabola
HYMENOPTERA (13)
Neuroptera
Megaloptera
Rhaphidioptera
Strepsiptera
COLEOPTERA (9)
LEPIDOPTERA (15)
Trichoptera
DIPTERA (15)
Mecoptera
C. felis
Antliophora
Panorpida
OGs with C. felis and one or more orders
4000
3000
2000
1000
0
3491
2055
747
473
471
437
394
363
339
260
242
197
189
160
159
91
Total OGs
10K
5K
0
Hymenoptera
Coleoptera
Lepidoptera
Diptera
C. felis

Fig. 3

1046 **Fig. 4. Identifying *C. felis*-specific genes.** (**A**) *C. felis* proteins failing to cluster with
1047 counterparts in other holometabolan genomes were determined to lack (top) or possess limited
1048 similarity to (bottom) proteins from holometabolan or other genomes (bottom). (**B**) For 1,318
1049 proteins, Gene Ontologies and Interpro domains were included in annotation and clustering into
1050 broad cellular function categories. (**C**) *C. felis* carries tandemly-arrayed resilin homologs (gray
1051 inset) as well as a cohort of other proteins containing resilin-like features. Red boxes indicate
1052 other tandemly-arrayed genes.
1053
1054


A
Unique proteins
(n=2084)
0 nr hits
1+ nr hits
Protein length (aa)
6000
5000
4000
3000
2000
1000
0
Query length (aa)
> 500
100-499
< 100
Total subjects
450
400
300
200
100
0
Holometabola
Others
B
Fig. 4
Number of C. felis-specific proteins
0
20
40
60
80
100
120
140
macromolecule metab. proc.
cellular nitrogen compound metab. proc.
cellular macromolecule metab. proc.
cellular aromatic compound metab. proc.
organic cyclic compound metab. proc.
heterocycle metab. proc.
nucleobase-containing compound metab. proc.
organonitrogen compound metab. proc.
protein metab. proc.
organic substance biosyn. proc.
Biological Process
nucleic acid binding
cation binding
hydrolase activity, acting on ester bonds
anion binding
nucleotide binding
nucleoside phosphate binding
ribonucleotide binding
ATP binding
transferase activity, phosphorus-containing groups
cytoskeletal protein binding
Molecular Function
intracellular membrane-bounded organelle
intracellular non-membrane-bounded organelle
nuclear part
chromosomal part
cytoplasm
intracellular organelle lumen
cytoskeletal part
transferase complex
integral component of membrane
cytoplasmic part
Cellular Component
C
Canonical Resilins
XP_026469611
680
XP_026469610
531
XP_026464608
210
XP_026464607
225
XP_026464486
246
XP_026471231
322
XP_026482328
332
XP_026461288
358
XP_026468636
530
XP_026465042
648
XP_026469989
788
XP_026469990
793
XP_026470736
950
XP_026470530
1288
XP_026470529
1978
700
NT "resilin-like"
YGAP motif
YGGP motif
CT "resilin-like"
GY(S/P)(G/S)G
Signal peptide
Mucin-like
Chitin-binding
Atrophin-1
SSXT
Repeat region

1055 **Fig. 5. The microbiome of EL fleas.** (**A**) Breakdown of the *C. felis* (EL fleas) microbiome. Bar
1056 at top graphically depicts the taxonomic distribution of non-flea Illumina reads across Bacteria,
1057 viruses and Archaea. Each group is further classified, with the major taxa (genus-level in most
1058 cases) and compiled read size (Mb) provided. Taxa with asterisks are AT-rich genomes that
1059 were later determined to match to *C. felis* mitochondrial reads. (**B**) *Wolbachia* genome-based
1060 phylogeny estimation. *Wolbachia* supergroups are within gray ellipses. *C. felis*-associated
1061 Wolbachiae are within black boxes. Red (*w*CfeT) and blue (*w*CfeJ) stars depict the two novel
1062 Wolbachiae infecting *C. felis*, with assembly information for each genome provided at right.
1063 Inset: color scheme for nematode and arthropod hosts. For tree estimation see **Methods**. Gel
1064 image (unaltered) depicts PCR results using 100ng of flea template DNA (quantified via
1065 nanodrop) in separate reactions with gene-specific primers. (**C**) *w*CfeT contains the unique biotin
1066 synthesis operon carried by certain obligately host-associated microbes. Schema follows our
1067 previous depiction of the unique *bio* gene order [44], with all proteins drawn to scale (as a
1068 reference, *w*CfeT BioB is 316 aa). Comparisons are made to the *bio* proteins of *Cardinium*
1069 endosymbiont of *Encarsia pergandiella* (cEper1, CCM10336-CCM10341) and *Wolbachia*
1070 endosymbiont of *Cimex lectularius* (*w*Cle, BAP00143-BAP00148). Red shading and numbers
1071 indicate % identity across pairwise protein alignments (blastp). (**D**) *w*CfeJ contains a CinA/B
1072 operon. Comparisons are made to the CidA/B (top, CAQ54390/1) and CinA/B (bottom) operons
1073 of *Wolbachia* endosymbiont of *Culex quinquefasciatus* Pel (wPip_Pel, CAQ54402/3). Green,
1074 CE clan protease; brown, PD-(D/E)XK nuclease. All proteins are drawn to scale (as a reference,
1075 *w*CfeJ CinB is 777 aa). Red shading and numbers indicate % identity across pairwise protein
1076 alignments (blastp).
1077


A
Fig. 5
Bacteria (219.2 Mb)
Proteobacteria (202.9 Mb)
Tenericutes (8.76 Mb)
Firmicutes (2.36 Mb)
Other phyla (1.73 Mb)
Unassigned (3.43 Mb)
Viruses (21.82 Mb)
Unassigned
4.83 Other (62 taxa)
1.79 Chrysodeixis chalcites NPV
3.21 Pandoravirus dulcis
7.58 Choristoneura occidentalis GV
Archaea (11.01 Mb)
7.54 Unassigned
Other (24 genera)
0.89 Natronococcus
Methanothermo-coccus
1.70 Methanococcus
Other (62 taxa)
"Cand. Zinderia insecticola" *
Burkholderia
53.8 Achromobacter
4.38
Other genera
Saccharophagus
"Cand. Carsonella ruddii" *
39.8 Alteromonas
Other genera
Methylobacterium
17.9 Wolbachia
Sulfuricurvum
β γ α ε
Other genera
(1.49) Phytoplasma
7.22 Mycoplasma
Tenericutes
Other genera
Clostridium
Firmicutes
B
OTHER ANAPLASMATACEAE
wPpe (Pratylenchus penetrans)
wPni (Pentalonia nigronervosa)
wFol (Folsomia candida)
wCfeT (Ctenocephalides felis)
wLs (Litomosoides sigmodontis)
wWb (Wucheria bancrofti)
wBm (Brugia malayi)
D
wCle (Cimex lectularius)
wOc (Osmia caerulescens)
wMen (Mengenilla moldrzyki)
F
wCfeJ (Ctenocephalides felis)
wDim (Dirofilaria immitis)
wOv (Onchocerca volvulus)
wOo (Onchocerca ochengi)
C
wZoo (Zootermopsis nevadensis)
wVulC-f (Armadillidium vulgare)
wStr (Laodelphax striatella)
wDi (Diaphorina citri)
wTpre (Trichogramma pretiosum)
wAlbB (Aedes albopictus)
wNo (Drosophila simulans)
wCte (Ctenocephalides felis)
wVitB (Nasonia vitripennis)
wCoc1 (Dactylopius coccus)
wDacB (Dactylopius coccus)
wAus (Plutella australiana)
wOb_Wba (Operophtera brumata)
wBol1-b (Hypolimnas bolina)
wPip_Mol (Culex molestus)
wPip_Pel (Culex quinquefasciatus)
wPip_JHB (Culex quinquefasciatus)
B
22 (Hemiptera, Hymenoptera, Diptera)
A
wCfeT
- assembled in 1 contig
- length = 1.5 Mb
- 35.2 %GC
- no. CDS (protein) = 1509
- 36 RNA sequences
- 16S rDNA = EF121345
- Biotin synthesis operon
- WO prophage (75%)
wCfeJ
- assembled in 1 contig
- length = 1.2 Mb
- 35.6 %GC
- no. CDS (protein) = 1178
- 35 RNA sequences
- 16S rDNA = EF121347
- TA module (CinA/CinB)
- WO prophage fragment
Nematoda
Tylenchoidea
Filaroidea
Arthropoda
Isopoda
Collembola
Hemiptera
Blattodea
Hymenoptera
Strepsiptera
Lepidoptera
Siphonaptera
Diptera
bootstrap
.91-1
.81-.90
.71-.80
.61-.70
.51-.60
< .50
0.1 sub./site
C. felis
EL
Heska
(- DNA)
1500
500
200
wCfeJ-specific (CinA)
wCfeT-specific (ApaG)
wCfeJ/T-specific (16S rRNA)
C
BioB
BioF
BioH
BioC
BioD
BioA
cEper1
.83
.74
.60
.60
.67
.71
wCfeT
.86
.85
.69
.71
.80
.83
wCle
D
CidA
CidB
wPip_Pel
.25
.21
wCfeJ
.34
.34
wPip_Pel
CinA
CinB