A chromosome-level assembly of the cat flea genome uncovers rampant gene duplication and genome size plasticity

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, WV, USA.

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

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

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

[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] Department of Microbiology and Immunology, College of Medicine, University of South Alabama, Mobile, AL, USA.

* 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. Finally, from the *C. felis* sequence reads we also generated closed genomes for two novel strains of *Wolbachia*, one parasitic and one symbiotic, found to co-infect individual fleas.

Conclusion: Rampant CNV in *C. felis* has dire implications for gene-targeting pest control measures and stands to complicate standard normalization procedures utilized in comparative

transcriptomics analysis. Coupled with co-infection by novel *Wolbachia* endosymbionts – potential tools for blocking pathogen transmission – these oddities highlight a unique and underappreciated disease vector.

Background

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

The Oriental rat flea, *Xenopsylla cheopis*, and to a lesser extent the cat flea, *Ctenocephalides felis*, transmit *Yersinia pestis*, the causative agent of bubonic plague [3–5]. Fleas that feed away from their primary hosts (black rats and other murids) can introduce *Y. pestis* to humans, which historically has eliminated a substantial fraction of the world's human population; e.g., the Plague of Justinian and the Black Death [5]. Bubonic plague remains a significant threat to human health [6, 7] as do other noteworthy diseases propagated by flea infestations, including murine typhus (*Rickettsia typhi*), murine typhus-like illness (*R. felis*), cat-scratch disease (*Bartonella henselae*), and Myxomatosis (myxoma virus) [8, 9]. Fleas also serve as intermediate hosts for certain medically-relevant helminths and trypanosome protozoans [10]. In addition to

70 the potential for infectious disease transmission, flea bites are also a significant nuisance and can
71 lead to serious dermatitis for both humans and their companion animals. Epidermal burrowing
72 by the jigger flea, *Tunga penetrans*, causes a severe inflammatory skin disease known as
73 Tungiasis, which is a scourge on many human populations within tropical parts of Africa, the
74 Caribbean, Central and South America, and India [11, 12]. Skin lesions that arise from flea
75 infestations also serve as sites for secondary infection. Collectively, fleas inflict a multifaceted
76 human health burden with enormous public health relevance [13].

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

87 Notwithstanding their tremendous impact on global health and economy, fleas are relatively
88 understudied compared to other arthropod disease vectors [18]. While transcriptomics data for
89 mecopteroids (Mecoptera + Siphonaptera) have proven useful for Holometabola phylogeny
90 estimation [2], assessment of flea immune pathways [19], and analysis of opsin evolution [20],
91 the lack of mecopteroid genomes limits further insight into the evolution of Antliophora
92 (mecopteroids + Diptera (true flies)) and severely restricts comparative studies of disease

93 vectors. Thus, sequencing flea genomes stands to greatly improve our understanding of the
94 shared and divergent mechanisms underpinning flea and fly vectors, a collective lineage
95 comprised of the deadliest animals known to humans [21]. To address this glaring void in insect
96 genomics and vector biology, we sequenced the genome of *C. felis*, a principal vector of *R. typhi*,
97 *R. felis*, and *Bartonella* spp. [22–25] and an insect with substantial human health and veterinary
98 importance across the globe [1]. To overcome the minute body size of individual fleas, we
99 pooled multiple individuals to generate sufficient DNA for sequencing, sampled from an inbred
100 colony to reduce allelic variation, and applied orthogonal informatics approaches to account for
101 challenges arising from the potential misassembly of haplotypes.
102

103 **Results**

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

116 *felis* protein-encoding genes to the Benchmarking Universal Single-Copy Orthologues (BUSCO
117 [28]) for eukaryotes, arthropods, and insects indicates our BIG9 assembly is robust and lacks
118 only a few conserved genes (**Additional file 1: Fig. S1B**). As a result, we focus our subsequent
119 analyses on the BIG9 scaffolds unless otherwise noted.
120

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

122 Previous work using flow cytometry estimated the size of the female *C. felis* genome at 465 Mb,
123 while our BIG9 assembly contained 654 Mb total bases (25% larger). Furthermore, BUSCO
124 analysis suggested that roughly 30% of conserved, single-copy Insecta genes in the BUSCO set
125 were duplicated in our assembly (**Additional file 1: Fig. S1B**). In order to investigate whether
126 this duplication might be widespread across the genome, and thereby account for the larger size
127 of our assembly, we used BLASTP to construct *C. felis*-specific protein families at varying levels
128 of sequence identity from 85-100%. Remarkably, 61% (10,088) of all protein-encoding genes in
129 *C. felis* arise from duplications at the 90% identity threshold or higher (**Fig. 1B**). Over 68% of
130 these comprise true (n=2) duplications, most of which occur as tandem or proximal loci less than
131 12 genes apart (**Fig. 1C, Additional file 1: Fig. S1L**). We observed little change in either the
132 total number of duplications or the distribution at thresholds below 90% identity; consequently,
133 we define "duplications" here as sequences that are 90% identical or higher (see **Methods**).
134 Duplications are on-going and rapidly diverging as evinced by: 1) their high concentration on
135 individual BIG9 scaffolds (**Fig. 1D**, **Additional file 1: Fig. S1C-K**), 2) a lack of increasing
136 divergence with greater distance on scaffolds (**Additional file 1: Fig. S1L**), and 3) a lack of
137 increasing divergence for duplicate genes found across different scaffolds (**Additional file 1:
138 Fig. S1M**). Among cellular functions for duplicate genes, certain transposons and related factors

139 (GO:0015074, "DNA integration") are enriched relative to 6,430 single copy protein-encoding
140 genes (**Fig. 1E**, **Additional file 2: Table S1**). However, the frequency and distribution of these
141 elements are dwarfed by total duplicate genes (**Additional file 1: Fig. S1N**). Additionally,
142 transposons and other repeat elements encompass only 10% of the genome (**Additional file 1:**
143 **Fig. S1O**), indicating that selfish genetic elements do not contribute significantly to the rampant
144 gene duplication observed. Thus, the *C. felis* genome is remarkable given that genes producing
145 duplications (n=3,863 or ~38% of total protein-encoding genes) are 1) indiscriminately dispersed
146 across chromosomes, 2) not clustered into blocks that would suggest whole or partial genome
147 duplications, and 3) not the product of repeat element-induced genome obesity.

148 The *C. felis* genome also carries an impressive number of tRNA-encoding genes (*n*=4,358 on
149 BIG9 scaffolds) (**Fig. 1A**). While tRNA gene numbers and family compositions vary
150 tremendously across eukaryotes [29], the occurrence of more than 1000 tRNA genes per genome
151 is rare (**Fig. 1F**). Notably, the elevated abundance of tRNA genes in *C. felis* is complemented by
152 an enrichment in translation-related functions among duplicated protein-coding genes (**Fig. 1E,**
153 **Additional file 2: Table S1**). While this possibly indicates increased translational requirements
154 to accommodate excessive gene duplication, it is more likely a consequence of the indiscriminate
155 nature of the gene duplication process. Relative to tRNA gene frequencies in other
156 holometabolan genomes, *C. felis* has several elevated (Arg, Val, Phe, Thr) and reduced (Gly,
157 Pro, Asp, Gln) numbers of tRNA families (**Additional file 1: Fig. S1P**); however, *C. felis* codon
158 usage is typical of holometabolan genomes (**Additional file 1: Fig. S1Q**). Like proliferated
159 protein-encoding genes, the significance of such high tRNA gene numbers is unclear but further
160 accentuates a genome in flux.

161

162 **Genome size estimation**

163 Duplicated regions (including intergenic sequences) account for approximately 227 Mb of the *C.*
164 *felis* genome; when subtracted from the BIG9 assembly (654 Mb), the resulting "core" genome
165 size of 427 Mb is congruous with a previous flow cytometry-based genome size estimate (mean
166 of 465 Mb, range of 32 Mb) for cat fleas previously assayed from a different geographic locale
167 [30]. To determine if EL fleas possess a greater genome size due to pronounced gene
168 duplication relative to other cat fleas, we similarly used flow cytometry to estimate genome sizes
169 for individual EL fleas and compared them to the previous findings. As expected, mean genome
170 size was not significantly different between sex-matched *C. felis* from the two populations (p =
171 0.1299). Remarkably, however, no two individual EL fleas possessed comparable genome sizes,
172 with an overall uniform size distribution and relatively large variability (118 Mb) (**Fig. 2A**;
173 **Additional file 3: Fig. S2**). Indeed, the coefficient of variation for *C. felis* (0.13; n = 26) was
174 3.2X higher than that of either *Drosophila melanogaster* (0.040; n = 26) or *D. viridis* (0.039; n =
175 26), which were prepared and measured concurrently (**Fig. 2A**, inset), underscoring the
176 extraordinary extent of inter-individual variation in *C. felis*. Genome size estimates for another
177 flea (the rat flea, *X. cheopis*, also sex-matched) show a similar uniform distribution and range
178 across individuals (**Fig. 2A**), pointing to an extraordinary genetic mechanism that may define
179 siphonapteran genomes.

180 Accordingly, we propose that our assembly captured a conglomeration of individual flea
181 copy number variations (CNV) that is cumulative for all expansions and contractions of
182 duplicate regions (**Fig. 2B**). The presence of extensive gene duplications is further supported by
183 mapping short read Illumina data to our assembly, which showed a significantly reduced mean
184 read depth across duplicated loci versus single-copy genes (**Fig. 2C**). As an alternative to CNV,

185 we considered that allelic variation could also be contributing to extensive gene duplication in
186 our assembly. To address this concern, we took three approaches. First, polished contigs were
187 scanned for haplotigs using the program *Purge Haplotigs* [31]; no allelic variants were detected.
188 Second, we mapped 1KITE transcriptome reads [2] generated from fleas of an unrelated colony
189 (Kansas State University) to our assembly (**Fig. 2D**). If our sequence duplication is a result of
190 allelic variation within the EL colony, we would expect to see a lack of congruence in the
191 distribution of transcripts mapping to single copy genes versus duplicates (different colonies with
192 different allelic variation). We might also expect to see a significant proportion of transcripts
193 that do not map at all. Instead, 91% of 1KITE reads map to CDS in our assembly, and the
194 distributions of transcripts mapping to single copy and duplicate genes are identical.
195 Third, we reasoned if sequence duplications are the result of misassembled allelic variants,
196 then most duplicate CDS within a cluster would be the same length. Alternatively, if
197 duplications are true CNVs, we would expect a significant number of truncations as a
198 consequence of gene purging associated with unequal crossing over. To assess this, we
199 determined the proportion of duplicate clusters with one or more truncated members, as well as
200 the extent of truncation relative to the longest member of the cluster (**Fig. 2E**). Approximately
201 70% of gene duplications are not comparable in length. In addition, mean extent of truncation is
202 25% or greater across all clusters regardless of % identity. Together with genome size
203 estimations, short read mapping analysis, and transcript mapping to our assembly, these data
204 indicate active gene expansion and contraction underpinning CNV in fleas and dispel allelic
205 variation as a significant contributor to gene duplication. While the cytogenetic mechanisms are
206 unclear, elevated numbers of DNA repair enzymes (GO:0006281) relative to genome size may
207 correlate with excessive CNV (**Additional file 2: Table S1**).

208

209 **Genome evolution within Holometabola**

210 Despite inordinate gene duplication, the completeness of the *C. felis* proteome as estimated by
211 occurrence of 1,658 insect Benchmarking Universal Single-Copy Orthologues (BUSCOs) is
212 congruous with those of other sequenced holometabolan genomes (**Fig. 3A**). Only one other
213 genome (*Aedes albopictus*) contains greater gene duplication among BUSCOs than *C. felis*;
214 however, this mosquito genome is much larger (~2 Gb) and riddled with repeat elements [32]. A
215 genome-wide analysis of shared orthologs among 53 holometabolan genomes indicates a slight
216 affinity of *C. felis* with Coleoptera, though the divergent nature of Diptera and availability of
217 only a single flea genome likely mask inclusion of fleas with flies (**Fig. 3B**). Overall,
218 phylogenomics analysis reveals that *C. felis* harbors 3,491 orthologs found in at least one other
219 taxon from each holometabolan order (**Fig. 3C**); however, only 577 "core" orthologs were
220 present in all taxa from every order (**Fig. 3C**, yellow bar), reflecting either incomplete genome
221 assemblies or an incredibly patchwork Holometabola accessory genome (**Additional file 4: Fig.
222 S3A**). Other conserved protein-encoding genes that define higher-generic groups (**Fig. 3C**,
223 inset) will inform lineage diversification within Holometabola (**Additional file 5: Table S2**).
224 Conversely, 29 protein-encoding genes absent in *C. felis* but conserved in Panorpida species
225 (Antliophora + Lepidoptera (butterflies and moths)) stand to illuminate patterns and processes of
226 flea specialization via reduction (**Additional file 4: Fig. S3B**, **Additional file 5: Table S2**).
227 Overall, despite its parasitic lifestyle and reductive morphology, *C. felis* has not experienced a
228 significant reduction in gene families (**Additional file 4: Fig. S3A**, **Additional file 5: Table S2**)
229 as seen in other host-dependent eukaryotes [33].

230

231 **Unique cat flea genome features**

232 *C. felis* protein-encoding genes that failed to cluster with other Holometabola (4,282 sequences
233 in 2,055 ortholog groups, **Fig. 3C**) potentially define flea-specific attributes. Elimination of
234 divergent "holometabolan-like" proteins, identified with BLASTP against the nr database of
235 NCBI, left 2,084 "unique" *C. felis* proteins (**Fig. 4A, Additional file 6: Table S3**). These
236 include proteins lacking counterparts in the NCBI nr database (n=766), and proteins with either
237 limited similarity to Holometabola or greater similarity to non-holometabolan taxa (n=1,318).
238 Proteins comprising the latter set were assigned an array of functional annotations (GO, KEGG,
239 InterPro, EC) and stand to guide efforts for deciphering flea-specific innovations (**Fig. 4B**,
240 **Additional file 6: Table S3**).

241 Two isoforms (A and B) of resilin, an elastomeric protein that provides soft rubber-elasticity
242 to mechanically active organs and tissues, were previously identified in *C. felis* and proposed to
243 underpin tarsal-mediated jumping [34]. Resilins typically have 1) highly repetitive Pro/Gly
244 motifs that provide high flexibility, 2) key Tyr residues that facilitate intermolecular bonds
245 between resilin polypeptides, and 3) a chitin-binding domain (CBD), though *C. felis* isoform B
246 lacks the CBD [34, 35]. The *C. felis* assembly has two adjacent genes encoding resilins (gray
247 box, **Fig. 4C**): the larger (680 aa) protein is more similar to both resilin A and B isoforms
248 identified previously (>99 %ID), while the smaller (531 aa) protein is more divergent (98.8
249 %ID). These divergent resilins accentuate the observed CNV in *C. felis* and indicate additional
250 genetic complexity behind flea jumping. Furthermore, a cohort of diverse proteins containing
251 multiple resilin-like features and domains were identified, opening the door for future studies
252 aiming to characterize the molecular mechanisms underpinning the great jumping ability of fleas.
253

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

255 Analysis of microbial-like Illumina reads revealed a bacterial dominance, primarily represented
256 by *Proteobacteria* (**Fig. 5A**, **Additional file 7: Table S4**). Aside from the *Wolbachia* reads
257 (discussed below), none of the bacterial taxa match to species previously detected in
258 environmental [36, 37] or colony fleas [38]. Thus, a variable bacterial microbiome exists across
259 geographically diverse fleas and is likely influenced by the presence of pathogens [38]. Strong
260 matches to lepidopteran-associated *Chrysodeixis chalcites* nucleopolyhedrovirus and
261 *Choristoneura occidentalis* granulovirus, as well as *Pandoravirus dulcis*, identify
262 underappreciated viruses that may play important roles in the vectorial capacity of *C. felis*.

263 Remarkably, two divergent *Wolbachia* genomes were assembled, circularized and annotated.
264 Named *w*CfeT and *w*CfeJ, these novel strains were previously identified (using 16S rDNA) in a
265 cat flea colony maintained at Louisiana State University [38–40], which historically has been
266 replenished with EL fleas. Robust genome-based phylogeny estimation indicates *w*CfeT is
267 similar to undescribed *C. felis*-associated strains that branch ancestrally to most other *Wolbachia*
268 lineages [36, 41], while *w*CfeJ is similar to undescribed *C. felis*-associated strains closely related
269 to *Wolbachia* supergroups C, D and F [42] (**Fig. 5B**; **Additional file 7: Table S4**). The
270 substantial divergence of *w*CfeT and *w*CfeJ from a *Wolbachia* supergroup B strain infecting *C.*
271 *felis* (*w*Cte) indicates a diversity of Wolbachiae capable of infecting cat fleas.

272 *w*CfeT and *w*CfeJ are notable for carrying segments of WO prophage, which are rarely
273 present in genomes of Wolbachiae outside of supergroups A and B [43]. Further, each genome
274 contains features that hint at contrasting relationships with *C. felis*. *w*CfeT carries the unique
275 biotin synthesis operon (**Fig. 5C**), which was originally discovered in *Rickettsia buchneri* by us
276 [44] and later identified in certain *Wolbachia* [45–47], *Cardinium* [48, 49] and *Legionella* [50]

277 species. Given that some *Wolbachia* strains provide biotin to their insect hosts [45, 51], we posit
278 that *w*CfeT has established an obligate mutualism with *C. felis* mediated by biotin-provisioning.
279 In contrast, *w*CfeJ appears to be a reproductive parasite, as it contains a toxin-antidote (TA)
280 operon that is similar to the CinA/B TA operon of *w*Pip_Pel that induces cytoplasmic
281 incompatibility (CI) in flies [52]. CinA/B operons are analogous to the CidA/B TA operons of
282 *w*Mel and *w*Pip_Pel, which also induce CI in fly hosts [53–55], yet the CinB toxin harbors dual
283 nuclease domains in place of the CidB deubiquitnase domain [56] (**Fig. 5D**). Given that the
284 genomes of many *Wolbachia* reproductive parasites harbor diverse arrays of CinA/B-and
285 CidA/B-like operons [56, 57], wCfeJ's CinA/B TA operon might function in CI or some other
286 form of reproductive parasitism. Quizzically, the co-occurrence of *w*CfeJ and *w*CfeT in
287 individual fleas (gel image in **Fig. 5B**) indicates dual forces (mutualism, parasitism) that
288 potentially drive their infection in EL fleas.
289

290 ## Discussion

291 We set out to generate a genome sequence for the cat flea, a surprisingly absent resource for
292 comparative arthropod genomics and vector biology. Our efforts to generate a *C. felis* assembly
293 brought forth an unexpected finding, namely that no two cat fleas share the same genome
294 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.

295

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

308 Ramifications of a genome in flux are readily identifiable. First, as gene duplication is a
309 major source of genetic novelty, extensive CNV likely affords *C. felis* with a dynamic platform
310 for innovation, allowing it to outpace gene-targeting pest control measures. Second, extensive
311 CNV will complicate standard normalization procedures utilized in comparative transcriptomics
312 analysis, requiring a more nuanced interpretation of standard metrics that are based on gene

313 length (i.e. RPKM, TPM, etc.). Furthermore, achieving high confidence with read-mapping to
314 cognate genes will be difficult in the face of neofunctionalization, subfunctionalization and early
315 pseudogenization, as well as dosage-based regulation of duplicate genes. Third, genetic markers
316 typically utilized for evolutionary analyses (e.g., phylotyping, population genetics,
317 phylogeography [58]) may yield erroneous results when applied to *C. felis* and related
318 *Ctenocephalides* species if targeted to regions of CNV (and particularly neofunctionalization).
319 Finally, as a *C. felis* chromosome-level genome assembly was only attainable by coupling
320 Illumina and PacBio sequencing with Hi-C scaffolding techniques, short-read based sequencing
321 strategies will be inadequate for other organisms with high CNV. The ability of the BIG9
322 assembly to serve as a reference genome in future short-read based sequencing efforts for other
323 cat fleas will be determined. Moving forward, newly developed low-input protocols for PacBio
324 sequencing will allow us to query individual fleas to robustly assess the degree of gene
325 duplication.

326 Excessive CNV in *C. felis*, and likely all Siphonaptera, requires the determination of the
327 genetic mechanisms at play. Why extreme gene duplication, when predicted across arthropods
328 using genomic and transcriptomic data [59], was not previously detected in fleas is unclear.
329 Excessive CNV aside, our study provides the first genome sequence for Siphonaptera, which will
330 substantially inform comparative studies on insect vectors of human disease. Furthermore,
331 newly-identified symbiotic (*w*CfeT) and parasitic (*w*CfeJ) *Wolbachia* will be paramount to
332 efforts for biocontrol of pathogens transmitted by cat fleas. The accrued resources and
333 knowledge from our study are timely. A drastic rise of murine typhus cases alone in Southern
334 California [60] and Galveston, Texas [61], which are directly attributable to fleas associated with

335 increasing population sizes of rodents and opossums, requires immediate and re-focused efforts
336 to combat this serious and underappreciated risk to human health.
337

338 Conclusion

339 Fleas are parasitic insects that can transmit many serious pathogens (i.e. bubonic plague,
340 endemic and murine typhus). The lack of flea genome assemblies has hindered research,
341 especially comparisons to other disease vectors. Here we combined Illumina and PacBio
342 sequencing with Hi-C scaffolding techniques to generate a chromosome-level genome assembly
343 for the cat flea, *Ctenocephalides felis*. Our work has revealed a genome characterized by
344 inordinate copy number variation (~38% of proteins) and a broad range of genome size estimates
345 (433-551 Mb) for individual fleas, suggesting a bizarre genome in flux. Surprisingly, the flea
346 genome exhibits neither inflation due to rampant gene duplication nor reduction due to their
347 parasitic lifestyle. Based on these results, as well as the nature and distribution of the gene
348 duplications themselves, we posit a dual mechanism of unequal crossing-over and gene
349 conversion may underpin this genome variability, although the biological significance remains to
350 be explored. Coupled with paradoxical co-infection with novel *Wolbachia* endosymbionts and
351 reproductive parasites, these oddities highlight a unique and underappreciated human disease
352 vector.
353

354 Methods

355 Experimental design

356 This study was undertaken to generate a high-quality reference genome assembly and annotation
357 for the cat flea, *C. felis*, and represents the first sequenced genome for a member of Order

358 Siphonaptera. Our approach leveraged a combination of long-read PacBio sequencing, short-
359 read Illumina sequencing, and Hi-C (Chicago and HiRise) data to construct a chromosome-level
360 assembly; RNA-seq data and BLAST2GO classifications to assist in gene model prediction and
361 annotation; sequence mapping to address assembly fragmentation and short scaffolds (<1Mb);
362 and ortholog group construction to explore a genetic basis for the cat flea's parasitic lifestyle.
363 Gene duplications were confirmed via orthogonal approaches, including genome size estimates
364 of individual fleas, gene-based read coverage calculations, genomic distance between
365 duplications, and correlation between duplications and repeat elements or contig boundaries.
366

367 **Genome Sequencing and Assembly**

368 Newly emerged (August 2017), unfed female *C. felis* (n = 250) from Elward Laboratories (EL;
369 Soquel, CA) were surface-sterilized for 5 min in 10% NaClO followed by 5 min in 70% C_2H_5OH
370 and 3X washes with sterile phosphate-buffered saline. Fleas were flash-frozen in liquid N_2 and
371 ground to powder with sterile mortar and pestle. High-molecular weight DNA was extracted
372 using the MagAttract HMW DNA Kit (Qiagen; Venlo, Netherlands), quantified using a Qubit
373 3.0 fluorimeter (Thermo-Fisher Scientific; Waltham, MA), and assessed for quality on a 1.5%
374 agarose gel. DNA (50 µg) was submitted to the Institute for Genome Sciences (University of
375 Maryland) for size-selection and preparation of sequencing libraries. Libraries were sequenced
376 on 12 SMRT cells of a PacBio Sequel (Pacific Biosciences; Menlo Park, CA), generating
377 7,239,750 reads (46.7 Gb total). Raw reads were corrected, trimmed, and assembled into 16,622
378 contigs with Canu v1.5 in "pacbio-raw" mode, using an estimated genome size of 465 Mb [30].
379 A second group of newly emerged (January 2016), unfed female EL fleas (n=100) was surface-
380 sterilized and homogenized as above, and genomic DNA extracted using the QIAgen DNeasy®

381 Blood and Tissue Kit (QIAgen, Hilden, Germany). DNA was submitted to the WVU Genomics
382 Core for the preparation of a paired-end 250bp sequencing library with an average insert size of
383 500bp. The library was sequenced on 4 lanes of an Illumina HiSeq 1500 (Illumina Inc.; San
384 Diego, CA), generating 450,132,548 reads which were subsequently trimmed to remove adapters
385 and filtered for length and quality using FASTX-Toolkit v0.0.14 (available from
386 http://hannonlab.cshl.edu/fastx_toolkit/). These short read data were used to polish the Canu
387 assembly with Pilon v1.1.6 in "fix-all" mode [62], and to determine the composition of the *C.*
388 *felis* microbiome (see below). Haplotigs in the polished contigs were resolved using
389 purge_haplotigs [31] with coverage settings of 5 (low), 65 (mid), and 180 (high). A third group
390 of newly-emerged (Feburary 2018), unfed female EL fleas (n = 200) were surface-sterilized as
391 above, frozen at -80°C, and submitted for Chicago and Dovetail Hi-C proximity ligation
392 (Dovetail Genomics, Santa Cruz, CA) [63] using the polished Canu assembly as a reference.
393 The resulting scaffolded assembly (3,926 scaffolds) was subjected to removal of microbial
394 sequences as described in the next section.
395

396 **Genome Decontamination**

397 A comparative BLAST-based pipeline slightly modified from our prior work [64] was used to
398 identify and remove microbial scaffolds before annotation. Briefly, polished contigs were
399 queried using BLASTP v2.2.31 against two custom databases derived from the nr database at
400 NCBI (accessed July 2018): (1) all eukaryotic sequences (eukDB), and (2) combined archaeal,
401 bacterial, and viral sequences (abvDB). For each query, the top five unique subject matches (by
402 bitscore) in each database were pooled and scored according to a comparative sequence
403 similarity measure, S_m:

404 $$S_m = bIQ$$

405

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

410 $$C(q,d) = \frac{2(\sum_{n}^{i=1}(n - r_i(q,d)) + 1)}{n(n+1)}$$

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

416 $$P = C(q, eukDB) - C(q, abvDB)$$
417

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

424

425 **Genome Annotation**

426 Assembled and decontaminated scaffolds were annotated with NCBI Eukaryotic Genome
427 Annotation Pipeline (EGAP) v8.1 (https://www.ncbi.nlm.nih.gov/books/NBK143764/). To

428 facilitate gene model prediction, we generated RNA-seq data from 6 biological replicates of
429 pooled *C. felis* females (Heska Corporation, Fort Collins, CO). Briefly, total RNA was isolated
430 and submitted to the WVU Genomics Core for the preparation of paired-end, 100 bp sequencing
431 libraries using ScriptSeq Complete Gold Kit for Epidemiology (Illumina Inc., San Diego, CA).
432 Barcoded libraries were sequenced on 2 lanes of an Illumina HiSeq 1500 in High Throughput
433 mode, yielding approximately 26 million reads per sample (Q > 30). Raw sequencing reads from
434 all 6 samples were deposited in NCBI under the BioProject accession PRJNA484943. In addition
435 to these data, the EGAP pipeline also integrated previously-published *C. felis* expression data
436 from the 1KITE project (accession SRX314844; [2]) and an unrelated EST library (Biosample
437 accession SAMN00161855). The final set of annotations is available as "Ctenocephalides felis
438 Annotation Release 100" at the NCBI.
439

440 **Genome Completeness and Deflation**

441 The distribution of scaffold lengths in our assembly, together with the relatively large number of
442 fleas in our sequenced pool, warranted evaluating short scaffolds as possible sources of genomic
443 heterogeneity among individual fleas. To address this possibility, assembly scaffolds shorter
444 than 1 Mb (n = 3,724) were mapped to scaffolds larger than 1 Mb (n = 9; the BIG9) with BWA-
445 MEM v0.7.12 [65] using default parameters (**Additional file 1: Fig. S1A**). Additionally,
446 genome completeness of the full assembly compared to just the BIG9 scaffolds was assessed
447 with Benchmarking Using Single Copy Orthologs (BUSCO) v3.0.2 [28] in "protein" mode,
448 using the *eukaryota_odb9*, *arthropoda_odb9*, and *insecta_odb9* data sets (**Additional file 1: Fig.
449 S1B**). Isoforms were removed before BUSCO analysis by identifying CDSs that derived from
450 the same protein-coding gene and removing all but the longest sequence.

451

452 **Assessing the Extent of Gene Duplication**

453 Proteins encoded on the BIG9 scaffolds (n = 16,518) were queried against themselves with
454 BLASTP v2.2.31 using default parameters. Pairs of unique sequences that met or exceeded a
455 given amino acid percent identity (%ID) threshold over at least 80% of the query length were
456 binned together. Bins of sequence pairs that shared at least one sequence in common were
457 subsequently merged into clusters. Isoforms were removed after clustering by identifying CDSs
458 in a cluster that derived from the same protein-coding gene and removing all but the longest
459 sequence. This process was used to generate cluster sets at integer %ID thresholds from 90% to
460 100%. These duplicate protein-encoding genes were then mapped onto each of the BIG9
461 scaffolds using Circos [66] (**Additional file 1: Fig. S1C-K**). Cluster diameters were calculated
462 as the number of non-cluster genes that lie between the edges of the cluster (*i.e.*, the two cluster
463 genes that are farthest apart on the scaffold) (**Additional file 1: Fig. S1L**). Clusters that span
464 multiple scaffolds (mapped across all BIG9 scaffolds in **Additional file 1: Fig. S1M**) defy an
465 accurate calculation of diameter and were assigned a cluster diameter of -1. In order to estimate
466 the fraction of our assembly comprising gene duplications, cluster coverages (by %ID threshold)
467 were calculated in three ways. First, the *coverage by CDS* was estimated by comparing the
468 number of single-copy (protein-encoding) genes to the total number of clusters; the latter number
469 is assumed to represent a theoretical set of minimal "seed" sequences. Second, the *coverage by*
470 *gene length* was calculated as the total number of nucleotides encoding the proteins in each
471 cluster (including introns and exons) minus the mean gene length (to account for a hypothetical
472 "ancestor" gene). Finally, the *coverage by genome region* was estimated by adding i*(n-1) to
473 each calculation of coverage by gene length, where n is the number of genes in the cluster and i

474 is the mean intergenic length across all BIG9 scaffolds (17,344 nt). In order to assess possible
475 enrichment of cellular functions among duplicated genes, clusters at the 90% ID level were
476 compared to the remaining BIG9 proteins by Fisher's Exact Test (corrected for multiple testing)
477 which is integrated into the FatiGO package of BLAST2GO (see section "*Functional*
478 *Classification of* C. felis *Proteins*" below). GO categories were reduced to their most specific
479 terms whenever possible.
480

481 **Length Variation Within Gene Duplication Clusters**

482 Variability in intra-cluster CDS length was assessed in two ways. First, the length of each CDS
483 in a cluster was compared to the longest CDS of the cluster, and the proportion of clusters with
484 any truncation (>1 AA) was calculated for each integer %ID threshold between 90 and 100% ID.
485 Second, the mean and distribution of length differences (i.e., the extent of truncation) was
486 calculated across all clusters for each integer %ID threshold between 90 and 100% ID.
487

488 **Analysis of Repeat Regions**

489 The extent and composition of repeat elements in the *C. felis* genome were assessed in two ways.
490 First, proteins annotated in the GO category "DNA Integration GO:0015074" (including
491 retrotransposons) were extracted, plotted by genomic coordinate on each BIG9 scaffold, and
492 assessed for co-localization either with gene duplicates (see above) or near the ends of scaffolds
493 (**Additional file 1: Fig. S1N**). Second, repeat elements were identified on the BIG9 scaffolds
494 with RepeatMasker v4.0.9 (available from http://www.repeatmasker.org/) in "RMBlast" mode
495 (species "holometabola"), using Tandem Repeat Finder v4.0.9 and the Repbase RepeatMasker
496 (October 2018) and Dfam 3.0 databases (**Additional file 1: Fig. S1O**).

497

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

499 Given the relatively large number of tRNA genes in our assembly, and the AT richness of our
500 genome, we were interested in exploring connections between tRNA gene frequencies and codon
501 usage. To this end, tRNA gene abundance on BIG9 scaffolds (n = 4,358) was determined by
502 binning genes into families according to their cognate amino acid and calculating the percent of
503 each family compared to the total number of tRNA genes (**Additional file 1: Fig. S1P**). A
504 similar approach was taken to quantify tRNA gene abundance by anticodon. TA richness of
505 each anticodon was subsequently calculated as the percent of A+T bases in the anticodon
506 corrected for the size of the tRNA family. Codon usage was calculated as the percent of total
507 codons using the coding sequences for genes on the BIG9 scaffolds, with isoforms removed as
508 described previously (**Additional file 1: Fig. S1Q**).

509

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

511 Protein sequences encoded on the BIG9 scaffolds (n = 16,518) were queried with BLASTP
512 v2.2.31 against the nr database of NCBI (accessed July 2018) using a maximum e-value
513 threshold of 0.1. The top 20 matches to each *C. felis* sequence were used to annotate queries
514 with Gene Ontology (GO) categories, Enzyme Classification (EC) codes, and protein domain
515 information using BLAST2GO v1.4.4 [67] under default parameters. A local instance of the GO
516 database (updated February 2019) was used for GO classification, and the online version of
517 InterPro (accessed April 2019) was used for domain discovery, including InterPro, PFAM,
518 SMART, PANTHER, PHOBIUS, and GENE3D domains; PROSITE profiles; SignalP-TM
519 (signal peptide) domains; and TMHMM (transmembrane helix) domains. InterPro data was used

520 to refine GO annotations whenever possible (**Additional file 2: Table S1**). A subset of *C. felis*
521 proteins (n = 153) classified as "DNA repair" (GO:0006281) was identified and all child GO
522 terms of these proteins tabulated (**Additional file 2: Table S1**). Assuming a linear relationship
523 between genome size and number of repair genes [68], we estimate *C. felis* has an enriched
524 repertoire closer to that of a 3 Gb genome.
525

526 **Genome Size Estimation**

527 Estimations for flea genome size largely followed previously reported approaches [69]. For *C.*
528 *felis* individuals, 1/20 of the flea head was combined with two standards: 1/20 of the head of a
529 female (YW) *Drosophila melanogaster* (1C = 175 Mbp) and 1/20 of the head of a lab strain *D.*
530 *virilis* female (1C = 328). The tissues were placed in 1ml of cold Galbraith buffer and ground to
531 release nuclei in a 2ml Kontes Dounce, using 15 strokes of the "A" pestle at a rate of three
532 strokes every two seconds. The resulting solution was strained through a 45µ filter, stained for 3
533 hours in the dark at 4°C with 25µl of propidium iodide, then scored for total red fluorescence
534 using a Beckman-Coulter CytoFLEX flow cytometer. The average channel number of the 2C
535 nuclei of the sample and standards were determined using the CytExpert statistical software.
536 Briefly, the amount of DNA was estimated as the ratio of the average red fluorescence of the
537 sample to the average red fluorescence of the standard multiplied by the amount of DNA (in
538 Mbp) of the standard. The estimates from the two standards were averaged. At least 500 nuclei
539 were counted in each sample and standard peak. The coefficients of variation (CV) for all peaks
540 were < 2.0. Fluorescence activation and gating based on scatter were used to include in each
541 peak only intact red fluorescent nuclei free of associated cytoplasmic or broken nuclear tags.
542 Histograms generated for the largest and smallest determined genome sizes show the minimal

543 change in position for the two standards, demonstrating the significant change in the relative
544 fluorescence (average 2C channel number) between *C. felis* individuals (**Additional file 3: Fig.**
545 **S2**).
546

547 **Characterizing Copy Number Variation**

548 In order to test the hypothesis that our genome assembly represents an agglomeration of
549 individuals with different levels of gene duplication, we used minimap2 [70] to map our short-
550 read sequence data against the full scaffolded assembly. After extracting the mapped reads with
551 samtools v0.1.19 [71], including primary and alternative mapping loci, a vector of sequence
552 depth (in bases) per position was generated with the genomecov function of bedtools v2.25.0
553 [72]. Mean depths for all 16,518 protein-coding genes on the BIG9 scaffolds were calculated as
554 total bases covering each gene divided by gene length. Finally, the mean depth across all
555 duplicated genes was compared to the mean depth across all single-copy genes using a Student's
556 t-test.

557 To evaluate the extent of gene duplication across different *C. felis* populations, reads from
558 the 1KITE transcriptome sequencing project (NCBI Sequence Read Archive accession
559 SRR921588) were mapped to the 3,733 scaffolds from our assembly using HISAT2 v2.1.0 [73]
560 under the --dta and --no_unal options. Mapped reads were sorted with samtools and abundance
561 per gene calculated as transcripts per million reads (TPM) using stringtie v1.3.4d [73]. TPM
562 values were binned and plotted against the number of duplicated (90% aa ID or higher) and
563 single-copy genes in the BIG9 assembly.
564

565 **Comparative Genomics**

566 Protein sequences (n=1,077,182) for 51 sequenced holometabolan genomes were downloaded
567 directly from NCBI (n=47) or VectorBase (n=3) or sequenced here (n=1). Isoforms were
568 removed before analysis wherever possible, by identifying CDSs that derived from the same
569 protein-coding gene and removing all but the longest CDS. Genome completeness was
570 estimated with BUSCO v3.0.2 in "protein" mode, using the *insecta_odb9* data set. Ortholog
571 groups (OGs; n=50,118) were constructed in three sequential phases: 1) CD-HIT v4.7 [74] in
572 accurate mode (-g 1) was used to cluster sequences at 50% ID; 2) PSI-CD-HIT (accurate mode,
573 local identity, alignment coverage minimum of 0.8) was used to cluster sequences at 25% ID; 3)
574 clusters were merged using clstr_rev.pl (part of the CD-HIT package). Proteins from *C. felis* that
575 did not cluster into any OG (n=4,282) were queried with BLASTP v2.2.31 against the nr
576 database of NCBI (accessed July 2018). Queries (n=2,170) with a top hit to any Holometabola
577 taxon, at a minimum %ID of 25% and query alignment of 80%, were manually added to the
578 original set of ortholog groups where possible (n=2,142) or set aside where not (n=28). The
579 remaining queries with at least one match in nr (n=1,318) were grouped by GO category level 4
580 and manually inspected; these included queries with top hits to Holometabolan taxa that did not
581 meet the minimum %ID or query coverage thresholds. Finally, *C. felis* proteins with no match in
582 nr (n=766) were binned by query length. These last two sets (n=2,084) comprise the set of
583 proteins unique to *C. felis* among all other Holometabola (**Additional file 6: Table S3**).
584 Congruence between OG clusters and taxonomy was determined by calculating a distance
585 (Euclidean) between each pair of taxa based on the number of shared OGs. The resulting matrix
586 was scaled by classic multidimensional scaling with the cmdscale function of R v3.5.1 [75], and
587 visualized using the ggplot package in R. Finally, pan-genomes were calculated for several key
588 subsets of Holometabola: 1) *C. felis* alone (Siphonaptera); 2) Antliophora (Siphonaptera and

589 Diptera); 3) Panorpida (Siphoanptera, Diptera, and Coleoptera); 4) all taxa except Hymenoptera;
590 and 5) all Holometabola (**Additional file 5: Table S2**). In order to account for differences in
591 genome assembly quality and taxon sampling bias, we define the pan-genome here as the set of
592 all OGs that contain at least one protein *from at least one taxon* in a given order. These
593 intersections were visualized as upset plots using UpSetR v1.3.3 [76]. Intersections of various
594 holometabolous taxa that lack *C. felis* were computed to gain insight on possible reductive
595 evolution in fleas (**Additional file 4: Fig. S3**, **Additional file 5: Table S2**).
596

597 **Microbiome Composition**

598 A composite *C. felis* microbiome was estimated using Kraken Metagenomics-X v1.0.0 [77], part
599 of the Illumina BaseSpace toolkit. Briefly, 105,256,391 PE250 reads from our short read data set
600 were mapped against the Mini-Kraken reference set (12-08-2014 version), resulting in 2,390,314
601 microbial reads (2.27%) that were subsequently assigned to best possible taxonomy (**Additional**
602 **file 7: Table S4**).
603

604 **Assembly of *Wolbachia* Endosymbiont Genomes**

605 Corrected reads from the Canu assembly of *C. felis* were recruited using BWA-MEM v0.7.12
606 (default settings) to a set of concatenated closed *Wolbachia* genome sequences (n=15)
607 downloaded from NCBI (accessed February 2018). Reads that mapped successfully were
608 extracted with samtools v0.1.19 and assembled separately into seed contigs (n=22) with Canu
609 v1.5 using default settings. Gene models on these seed contigs were predicted using the Rapid
610 Annotation of Subsystems Technology (RAST) v2.0 server [78], yielding two small subunit
611 (16S) ribosomal genes that were queried with BLASTN against the nr database of NCBI to

612 confirm the presence of two distinct *Wolbachia* strains. Seed contigs were further analyzed by
613 %GC and top BLASTN matches in the nr database of NCBI, and binned into three groups: *C.*
614 *felis* mitochondrial (n=1), *C. felis* genomic (n=6), and *Wolbachia*-like (n=15) contigs. The
615 *Wolbachia*-like contigs were subsequently queried with BLASTN against the full *C. felis*
616 assembly (before decontamination). A single *Wolbachia*-like contig (tig00000005; wCfeJ)
617 containing one of the two distinct 16S genes was retrieved intact from the full assembly. It was
618 removed from the primary assembly and manually closed by aligning the contig ends with
619 BLASTN. Gaps in the aligned regions were resolved by mapping our short read data to the
620 contig with BWA-MEM (default settings) and manually inspecting the read pileups. Six
621 additional contigs were also retrieved intact from the full assembly; these were likewise removed
622 and manually stitched together using end-alignment and short read polishing, resulting in a
623 second closed *Wolbachia* genome (*w*CfeT). The remaining *Wolbachia*-like contigs (n=8) were
624 found to be fractions of much longer flea-like contigs; these were left in the primary *C. felis*
625 assembly. Both *w*CfeJ and *w*CfeT sequences were submitted to the RAST v2.0 server for gene
626 model prediction and functional annotation.
627

628 **Phylogenomics of *Wolbachia* Endosymbionts**

629 Protein sequences (n=66,811) for 53 sequenced *Wolbachia* genomes plus 5 additional
630 Anaplasmataceae (*Neorickettsia helminthoeca* str. Oregon, *Anaplasma centrale* Israel, *A.*
631 *marginale* Florida, *Ehrlichia chaffeensis* Arkansas, and E. *ruminantium Gardel*) were either
632 downloaded directly from NCBI (n=30), retrieved as genome sequences from the NCBI
633 Assembly database (n=13), contributed via personal communication (n=8; Michael Gerth,
634 Oxford Brookes University), or sequenced here (n=2) (**Additional file 7: Table S4**). For

635 genomes lacking functional annotations (n=15), gene models were predicted using the RAST
636 v2.0 server (n=12) or GeneMarkS-2 v1.10_1.07 (n=3; [79]). Ortholog groups (n=2,750) were
637 subsequently constructed using FastOrtho, an in-house version of OrthoMCL [80], using an
638 expect threshold of 0.01, percent identity threshold of 30%, and percent match length threshold
639 of 50% for ortholog inclusion. A subset of single-copy families (n=47) conserved across at least
640 52 of the 58 genomes were independently aligned with MUSCLE v3.8.31 [81] using default
641 parameters, and regions of poor alignment were masked with trimal v1.4.rev15 [82] using the
642 "automated1" option. All modified alignments were concatenated into a single data set (10,027
643 positions) for phylogeny estimation using RAxML v8.2.4 [83], under the gamma model of rate
644 heterogeneity and estimation of the proportion of invariant sites. Branch support was assessed
645 with 1,000 pseudo-replications. Final ML optimization likelihood was -183020.639712.
646

647 **Confirmation of the presence of *Wolbachia* in *C. felis***

648 To assess the distribution of *w*CfeJ and *w*CfeT in *C. felis*, individual fleas from the sequenced
649 strain (EL) and a separate colony (Heska) not known to be infected with *Wolbachia* were pooled
650 (n=5) by sex and colony, surface-sterilized with 70% ethanol, flash-frozen, and ground in liquid
651 N_2. Genomic DNA was extracted using the GeneJET Genomic DNA Extraction Kit (Thermo-
652 Fisher Scientific; Waltham, MA), eluted twice in 50µl of PCR-grade H_2O, and quantified by
653 spectrophotometry with a Nanodrop 2000 (Thermo-Fisher Scientific; Waltham, MA). 100ng of
654 DNA from each pool was used as template in separate 25 µl PCR reactions using AmpliTaq
655 Gold 360 (Thermo-Fisher Scientific; Waltham, MA) and primer pairs (400 nmoles each) specific
656 for: 1) a 76nt fragment of the *cinA* gene specific to *w*CfeJ (Fwd: 5'-
657 AGCAACACCAACATGCGATT-3'; Rev: 5'- GAACCCCAGAGTTGGAAGGG-3'); 2) a 75nt

658 fragment of the *apaG* gene specific to *w*CfeT (Fwd: 5'- GCCGTCACTGGCAGGTAATA-3';
659 Rev: 5'- GCTGTTCTCCAATAACGCCA-3'); or 3) a 122nt fragment of *Wolbachia* 16S rDNA
660 (Fwd: 5'- CGGTGAATACGTTCTCGGGTY-3'; Rev: 5'- CACCCCAGTCACTGATCCC-3').
661 Primer specificities were confirmed with BLASTN against both the *C. felis* assembly and the nr
662 database of NCBI (accessed June 2018). Reaction conditions were identical for all primer sets:
663 initial denaturation at 95°C for 10 min, followed by 40 cycles of 95°C for 30 sec, 60°C for 30
664 sec, and 72°C for 30 sec, and a final extension at 72°C for 7 min. Products were run on a 2%
665 agarose gel and visualized with SmartGlow Pre Stain (Accuris Instruments; Edison, NJ).
666 Primers were tested before use by quantitative real-time PCR on a CFX Connect (Bio-Rad
667 Laboratories; Hercules, CA).
668

669 **Statistical Analysis**

670 Statistical analyses were carried out in R v3.5.1. Mean coverages across duplicated (n=7852) and
671 single-copy (n=7061) genes at the 90% ID threshold were compared for significance using a
672 Welch Two Sample t-test (unpaired, two-tailed) with 12,930 degrees of freedom and a p-value <
673 2.2×10^{-16}. Mean coverage of duplicated genes at %ID thresholds from 85-100% were compared
674 for significance using one-way Analysis of Variance (ANOVA) with 15 degrees of freedom and
675 a p-value = 0.2. A similar ANOVA was used to compare single-copy genes at 85-100% ID
676 thresholds, with a p-value < 2.2×10^{-16}.
677

678 **Data and Scripts**

679 Data generated for this project that is not published elsewhere, including BLAST2GO
680 annotations and OG assignments, as well as custom analysis scripts, are provided on GitHub in
681 the "cfelis_genome" repository available at https://www.github.com/wvuvectors/cfelis_genome.
682

683 ## Declarations

684 ***Ethics approval and consent to participate.*** Not applicable.

685 ***Consent for publication.*** Not applicable.

686 ***Availability of data and materials.*** All of the sequence data generated for this work are available
687 at the NCBI under Bioproject accessions PRJNA489463 (genome sequence and annotation) and
688 PRJNA484941 (RNA-seq data used to support annotation). Additional tables with GO
689 annotations, ortholog groups, and microbiome data, as well as scripts used to generate data
690 visualizations can be accessed at https://www.github.com/wvuvectors/cfelis_genome. Sequences
691 for *w*CfeT and *w*CfeJ are available on NCBI under Bioproject PRJNA622233.

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

693 ***Funding.*** Research reported in this publication was supported by the National Institute of Health
694 (NIH)/National Institute of Allergy and Infectious Diseases (NIAID) grants R01AI017828 and
695 R01AI126853 to AFA, R21AI26108 and R21AI146773 to JJG & MSR, and R01AII122672 to
696 KRM. KER-B and MLG were supported in part by the NIH/NIAID Grants T32AI095190
697 (Signaling Pathways in Innate Immunity) and T32AI007540 (Infection and Immunity). TPD and
698 VIV were supported by start-up funding provided to TPD by West Virginia University. The
699 content is solely the responsibility of the authors and does not necessarily represent the official
700 views of the funding agencies. The funders had no role in study design, data collection and
701 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.

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

References

1. Rust MK, Dryden MW. The Biology, Ecology, and Management of the Cat Flea. Annu Rev

725 Entomol. 1997;42:451–73.

726 2. Misof B, Liu S, Meusemann K, Peters RS, Donath A, Mayer C, et al. Phylogenomics resolves
727 the timing and pattern of insect evolution. Science (80-). 2014;346:763–7.

728 3. Leulmi H, Socolovschi C, Laudisoit A, Houemenou G, Davoust B, Bitam I, et al. Detection of
729 Rickettsia felis, Rickettsia typhi, Bartonella Species and Yersinia pestis in Fleas (Siphonaptera)
730 from Africa. PLoS Negl Trop Dis. 2014;8.

731 4. Eisen RJ, Gage KL. Transmission of flea-borne zoonotic agents. Annu Rev Entomol.
732 2012;57:61–82.

733 5. Perry RD, Fetherston JD. Yersinia pestis--etiologic agent of plague. Clin Microbiol Rev.
734 1997;10:35–66.

735 6. Nikiforov V V., Gao H, Zhou L, Anisimov A. Plague: Clinics, Diagnosis and Treatment. In:
736 Advances in experimental medicine and biology. 2016. p. 293–312.

737 7. Stenseth NC, Atshabar BB, Begon M, Belmain SR, Bertherat E, Carniel E, et al. Plague: past,
738 present, and future. PLoS Med. 2008;5:e3.

739 8. Bertagnoli S, Marchandeau S. Myxomatosis. Rev Sci Tech. 2015;34:549–56, 539–47.

740 9. McElroy KM, Blagburn BL, Breitschwerdt EB, Mead PS, McQuiston JH. Flea-associated
741 zoonotic diseases of cats in the USA: bartonellosis, flea-borne rickettsioses, and plague. Trends
742 Parasitol. 2010;26:197–204.

743 10. Votýpka J, Suková E, Kraeva N, Ishemgulova A, Duží I, Lukeš J, et al. Diversity of
744 Trypanosomatids (Kinetoplastea: Trypanosomatidae) Parasitizing Fleas (Insecta: Siphonaptera)
745 and Description of a New Genus Blechomonas gen. n. Protist. 2013;164:763–81.

746 11. Feldmeier H, Heukelbach J, Ugbomoiko US, Sentongo E, Mbabazi P, von Samson-
747 Himmelstjerna G, et al. Tungiasis—A Neglected Disease with Many Challenges for Global

748 Public Health. PLoS Negl Trop Dis. 2014;8:e3133.

749 12. Feldmeier H, Keysers A. Tungiasis – A Janus-faced parasitic skin disease. Travel Med Infect

750 Dis. 2013;11:357–65.

751 13. Millán J. Comments on the manuscript by Bitam et al., 'Fleas and flea-borne diseases.' Int J

752 Infect Dis. 2011;15:e219.

753 14. Krasnov BR. Functional and evolutionary ecology of fleas : a model for ecological

754 parasitology. https://www.cambridge.org/vi/academic/subjects/life-

755 sciences/entomology/functional-and-evolutionary-ecology-fleas-model-ecological-

756 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. Encyclopedia of Entomology. Dordrecht: Springer

760 Netherlands; 2008. p. 797–801.

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. 2014;30:228–33.

763 18. Rust M. The Biology and Ecology of Cat Fleas and Advancements in Their Pest

764 Management: A Review. Insects. 2017;8:118.

765 19. Rennoll SA, Rennoll-Bankert KE, Guillotte ML, Lehman SS, Driscoll TP, Beier-Sexton M,

766 et al. The cat flea (Ctenocephalides felis) immune deficiency signaling pathway regulates

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

768 20. Böhm A, Meusemann K, Misof B, Pass G. Hypothesis on monochromatic vision in

769 scorpionflies questioned by new transcriptomic data. Sci Rep. 2018;8:9872.

770 21. Tolle MA. Mosquito-borne Diseases. Curr Probl Pediatr Adolesc Health Care. 2009;39:97–

771 140.

772 22. Glickman LT, Moore GE, Glickman NW, Caldanaro RJ, Aucoin D, Lewis HB. Purdue
773 University-Banfield National Companion Animal Surveillance Program for emerging and
774 zoonotic diseases. Vector Borne Zoonotic Dis. 2006;6:14–23.

775 23. Bouhsira E, Franc M, Boulouis H-J, Jacquiet P, Raymond-Letron I, Liénard E. Assessment
776 of persistence of Bartonella henselae in Ctenocephalides felis. Appl Environ Microbiol.
777 2013;79:7439–44.

778 24. Nogueras MM, Pons I, Ortuño A, Miret J, Pla J, Castellà J, et al. Molecular detection of
779 Rickettsia typhi in cats and fleas. PLoS One. 2013;8:e71386.

780 25. Angelakis E, Mediannikov O, Parola P, Raoult D. Rickettsia felis: The Complex Journey of
781 an Emergent Human Pathogen. Trends Parasitol. 2016;32:554–64.

782 26. Kichijo H. A note on the chromosomes of the flea, Ctenocephalus canis. Japanese J Genet.
783 1941;17.3:122–3.

784 27. Thomas C, Prasad RS. Chromosome variations in Xenopsylla astia Rothschild, 1911
785 (Siphonaptera). A preliminary report. Experientia. 1978;34:1440–1.

786 28. Seppey M, Manni M, Zdobnov EM. BUSCO: Assessing Genome Assembly and Annotation
787 Completeness. In: Methods in molecular biology (Clifton, N.J.). 2019. p. 227–45.

788 29. Chan PP, Lowe TM. GtRNAdb 2.0: an expanded database of transfer RNA genes identified
789 in complete and draft genomes. Nucleic Acids Res. 2016;44:D184–9.

790 30. Hanrahan SJ, Johnston JS. New genome size estimates of 134 species of arthropods.
791 Chromosom Res. 2011;19:809–23.

792 31. Roach MJ, Schmidt SA, Borneman AR. Purge Haplotigs: allelic contig reassignment for
793 third-gen diploid genome assemblies. BMC Bioinformatics. 2018;19:460.

794 32. Chen X-G, Jiang X, Gu J, Xu M, Wu Y, Deng Y, et al. Genome sequence of the Asian Tiger
795 mosquito, *Aedes albopictus* , reveals insights into its biology, genetics, and evolution. Proc Natl
796 Acad Sci. 2015;112:E5907–15.
797 33. Poulin R, Randhawa HS. Evolution of parasitism along convergent lines: from ecology to
798 genomics. Parasitology. 2015;142:S6–15.
799 34. Lyons RE, Wong DCC, Kim M, Lekieffre N, Huson MG, Vuocolo T, et al. Molecular and
800 functional characterisation of resilin across three insect orders. Insect Biochem Mol Biol.
801 2011;41:881–90.
802 35. Su RS-C, Kim Y, Liu JC. Resilin: protein-based elastomeric biomaterials. Acta Biomater.
803 2014;10:1601–11.
804 36. Vasconcelos EJR, Billeter SA, Jett LA, Meinersmann RJ, Barr MC, Diniz PPVP, et al.
805 Assessing Cat Flea Microbiomes in Northern and Southern California by 16S rRNA Next-
806 Generation Sequencing. Vector-Borne Zoonotic Dis. 2018;18:491–9.
807 37. Lawrence AL, Hii S-F, Chong R, Webb CE, Traub R, Brown G, et al. Evaluation of the
808 bacterial microbiome of two flea species using different DNA-isolation techniques provides
809 insights into flea host ecology. FEMS Microbiol Ecol. 2015;91:fiv134.
810 38. Pornwiroon W, Kearney MT, Husseneder C, Foil LD, Macaluso KR. Comparative
811 microbiota of Rickettsia felis-uninfected and -infected colonized cat fleas, Ctenocephalides felis.
812 ISME J. 2007;1:394–402.
813 39. Sunyakumthorn P, Bourchookarn A, Pornwiroon W, David C, Barker SA, Macaluso KR.
814 Characterization and growth of polymorphic Rickettsia felis in a tick cell line. Appl Environ
815 Microbiol. 2008;74:3151–8.
816 40. Gillespie JJ, Driscoll TP, Verhoeve VI, Utsuki T, Husseneder C, Chouljenko VN, et al.

817 Genomic Diversification in Strains of Rickettsia felis Isolated from Different Arthropods.
818 Genome Biol Evol. 2015;7:35–56.
819 41. González-Álvarez VH, de Mera IGF, Cabezas-Cruz A, de la Fuente J, Ortega-Morales AI,
820 Almazán C. Molecular survey of Rickettsial organisms in ectoparasites from a dog shelter in
821 Northern Mexico. Vet Parasitol Reg Stud Reports. 2017;10:143–8.
822 42. Casiraghi M, Bordenstein SR, Baldo L, Lo N, Beninati T, Wernegreen JJ, et al. Phylogeny of
823 Wolbachia pipientis based on gltA, groEL and ftsZ gene sequences: clustering of arthropod and
824 nematode symbionts in the F supergroup, and evidence for further diversity in the Wolbachia
825 tree. Microbiology. 2005;151:4015–22.
826 43. Bordenstein SR, Bordenstein SR. Eukaryotic association module in phage WO genomes
827 from Wolbachia. Nat Commun. 2016;7:13155.
828 44. Gillespie JJ, Joardar V, Williams KP, Driscoll TP, Hostetler JB, Nordberg E, et al. A
829 Rickettsia genome overrun by mobile genetic elements provides insight into the acquisition of
830 genes characteristic of an obligate intracellular lifestyle. J Bacteriol. 2012;194:376–94.
831 45. Nikoh N, Hosokawa T, Moriyama M, Oshima K, Hattori M, Fukatsu T. Evolutionary origin
832 of insect-Wolbachia nutritional mutualism. Proc Natl Acad Sci U S A. 2014;111:10257–62.
833 46. Gerth M, Bleidorn C. Comparative genomics provides a timeframe for Wolbachia evolution
834 and exposes a recent biotin synthesis operon transfer. Nat Microbiol. 2017;2:16241.
835 47. Balvín O, Roth S, Talbot B, Reinhardt K. Co-speciation in bedbug Wolbachia parallel the
836 pattern in nematode hosts. Sci Rep. 2018;8:8797.
837 48. Penz T, Schmitz-Esser S, Kelly SE, Cass BN, Müller A, Woyke T, et al. Comparative
838 Genomics Suggests an Independent Origin of Cytoplasmic Incompatibility in Cardinium hertigii.
839 PLoS Genet. 2012;8:e1003012.

840 49. Zeng Z, Fu Y, Guo D, Wu Y, Ajayi OE, Wu Q. Bacterial endosymbiont Cardinium cSfur
841 genome sequence provides insights for understanding the symbiotic relationship in Sogatella
842 furcifera host. BMC Genomics. 2018;19:688.
843 50. Ríhová J, Nováková E, Husník F, Hypša V. Legionella Becoming a Mutualist: Adaptive
844 Processes Shaping the Genome of Symbiont in the Louse Polyplax serrata. Genome Biol Evol.
845 2017;9:2946–57.
846 51. Ju J-F, Bing X-L, Zhao D-S, Guo Y, Xi Z, Hoffmann AA, et al. Wolbachia supplement
847 biotin and riboflavin to enhance reproduction in planthoppers. ISME J. 2019;:1–12.
848 52. Chen H, Ronau JA, Beckmann JF, Hochstrasser M. A Wolbachia Nuclease and Its Binding
849 Partner Comprise a Novel Mechanism for Induction of Cytoplasmic Incompatibility. 2019.
850 53. Beckmann JF, Ronau JA, Hochstrasser M. A Wolbachia deubiquitylating enzyme induces
851 cytoplasmic incompatibility. Nat Microbiol. 2017;2:17007.
852 54. LePage DP, Metcalf JA, Bordenstein SR, On J, Perlmutter JI, Shropshire JD, et al. Prophage
853 WO genes recapitulate and enhance Wolbachia-induced cytoplasmic incompatibility. Nature.
854 2017;543:243–7.
855 55. Beckmann JF, Fallon AM. Detection of the Wolbachia protein WPIP0282 in mosquito
856 spermathecae: Implications for cytoplasmic incompatibility. Insect Biochem Mol Biol.
857 2013;43:867–78.
858 56. Gillespie JJ, Driscoll TP, Verhoeve VI, Rahman MS, Macaluso KR, Azad AF. A Tangled
859 Web: Origins of Reproductive Parasitism. Genome Biol Evol. 2018;10:2292–309.
860 57. Beckmann JF, Bonneau M, Chen H, Hochstrasser M, Poinsot D, Merçot H, et al. The Toxin–
861 Antidote Model of Cytoplasmic Incompatibility: Genetics and Evolutionary Implications. Trends
862 Genet. 2019.

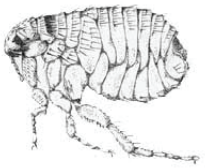
863 58. Lawrence AL, Webb CE, Clark NJ, Halajian A, Mihalca AD, Miret J, et al. Out-of-Africa,
864 human-mediated dispersal of the common cat flea, Ctenocephalides felis: The hitchhiker's guide
865 to world domination. Int J Parasitol. 2019;49:321–36.
866 59. Li Z, Tiley GP, Galuska SR, Reardon CR, Kidder TI, Rundell RJ, et al. Multiple large-scale
867 gene and genome duplications during the evolution of hexapods. Proc Natl Acad Sci.
868 2018;115:201710791.
869 60. California Department of Public Health.
870 https://www.cdph.ca.gov/Programs/CID/DCDC/Pages/Typhus.aspx.
871 61. Blanton LS, Idowu BM, Tatsch TN, Henderson JM, Bouyer DH, Walker DH. Opossums and
872 Cat Fleas: New Insights in the Ecology of Murine Typhus in Galveston, Texas. Am J Trop Med
873 Hyg. 2016;95:457–61.
874 62. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an
875 integrated tool for comprehensive microbial variant detection and genome assembly
876 improvement. PLoS One. 2014;9:e112963.
877 63. Putnam NH, O'Connell BL, Stites JC, Rice BJ, Blanchette M, Calef R, et al. Chromosome-
878 scale shotgun assembly using an in vitro method for long-range linkage. Genome Res.
879 2016;26:342–50.
880 64. Driscoll TP, Gillespie JJ, Nordberg EK, Azad AF, Sobral BW. Bacterial DNA sifted from the
881 Trichoplax adhaerens (Animalia: Placozoa) genome project reveals a putative rickettsial
882 endosymbiont. Genome Biol Evol. 2013;5:621–45.
883 65. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform.
884 Bioinformatics. 2009;25:1754–60.
885 66. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an

886 information aesthetic for comparative genomics. Genome Res. 2009;19:1639–45.
887 67. Gotz S, Garcia-Gomez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, et al. High-
888 throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res.
889 2008;36:3420–35.
890 68. Voskarides K, Dweep H, Chrysostomou C. Evidence that DNA repair genes, a family of
891 tumor suppressor genes, are associated with evolution rate and size of genomes. Hum Genomics.
892 2019;13:26.
893 69. Johnston JS, Bernardini A, Hjelmen CE. Genome size estimation and quantitative
894 cytogenetics in insects. In: Brown SJ, Pfrender ME, editors. Insect Genomics. New York:
895 Humana Press; 2019. p. 15–26.
896 70. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics.
897 2018;34:3094–100.
898 71. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence
899 Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.
900 72. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features.
901 Bioinformatics. 2010;26:841–2.
902 73. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of
903 RNA-seq experiments with HISAT, StringTie and Ballgown. Nat Protoc. 2016;11:1650–67.
904 74. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation
905 sequencing data. Bioinformatics. 2012;28:3150–2.
906 75. R Core Team. R: A Language and Environment for Statistical Computing. 2018.
907 https://www.r-project.org/.
908 76. Gehlenborg N. UpSetR: A More Scalable Alternative to Venn and Euler Diagrams for

909 Visualizing Intersecting Sets. 2017. https://cran.r-project.org/package=UpSetR.
910 77. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact
911 alignments. Genome Biol. 2014;15:R46.
912 78. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, et al. The RAST Server:
913 Rapid Annotations using Subsystems Technology. BMC Genomics. 2008;9:75.
914 79. Lomsadze A, Gemayel K, Tang S, Borodovsky M. Modeling leaderless transcription and
915 atypical genes results in more accurate gene prediction in prokaryotes. Genome Res.
916 2018;28:1079–89.
917 80. Li L, Stoeckert CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic
918 genomes. Genome Res. 2003;13:2178–89.
919 81. Edgar RC. MUSCLE: Multiple sequence alignment with high accuracy and high throughput.
920 Nucleic Acids Res. 2004;32:1792–7.
921 82. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment
922 trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25:1972–3.
923 83. Stamatakis A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large
924 phylogenies. Bioinformatics. 2014;30:1312–3.
925 84. Giraldo-Calderón GI, Emrich SJ, MacCallum RM, Maslen G, Dialynas E, Topalis P, et al.
926 VectorBase: an updated bioinformatics resource for invertebrate vectors and other organisms
927 related with human diseases. Nucleic Acids Res. 2015;43 Database issue:D707-13.
928
929
930
931 **Figures, tables and additional files**

932 **Fig. 1. *C. felis* genome characteristics**. (**A**) Summary statistics for long-read sequencing,
933 assembly and gene annotation. (**B**) Of 16,518 total protein-encoding genes (BIG9 scaffolds),
934 10,088 are derived from gene duplications (6,225 duplication events within 3,863 OGs at a
935 threshold of 90% aa identity). (**C**) Assessment of the number of genes per duplication (*left*) and
936 the relative distances between duplicate genes (*right*). Distances were computed only for true
937 duplications (n=2 genes) at a threshold of 90% aa identity. (**D**) Gene duplications are enriched
938 within BIG9 scaffolds (tandem and proximal, red numbers) versus across scaffolds (dispersed,
939 black numbers). (**E**) Enriched cellular functions of duplicate genes relative to single-copy genes.
940 (**F**) *C. felis* belongs to a minimal fraction of eukaryotes containing abundant tRNA genes. tRNA
941 gene counts are shown for disease vectors (VectorBase [84]) and eukaryotes carrying over 1000
942 tRNA genes (GtRNAdb [29]; ratios show number of genomes with > 1000 tRNA genes per
943 taxon.
944

A



C. felis

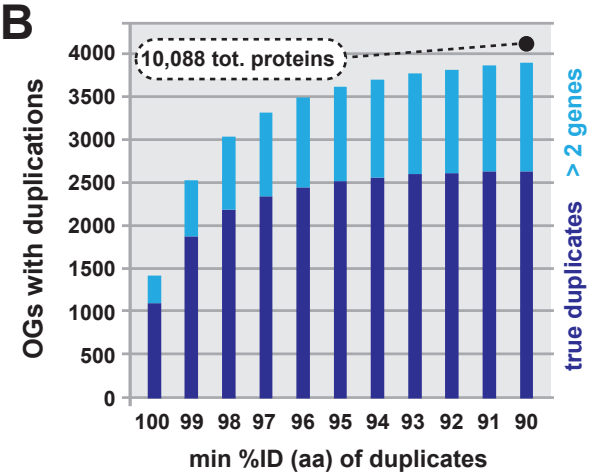
LONG-READ SEQUENCING

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

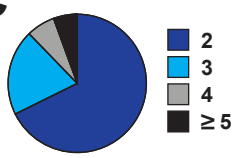
ASSEMBLY	ALL	BIG9
Tot. seq. length (Mb)	773.8	654.0
Contigs assembled	16,622	12,348
Contig N50 (Mb)	0.061	0.082
Longest contig (Mb)	1.9	1.9
Mean contig %GC	30.2	29.2
Scaffolds constructed	3,926	9
Scaffold N50 (Mb)	71.7	86.1
Scaffold L50	4	3
Longest scaffold (Mb)	185.5	185.5
No. scaffolds > 10Mb	9	9

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)

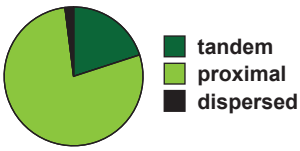
B



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

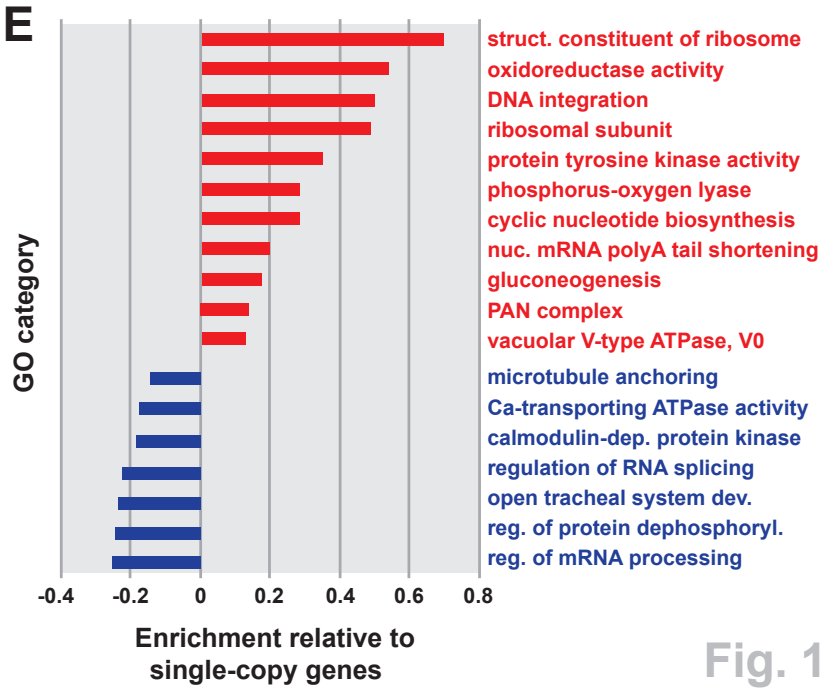
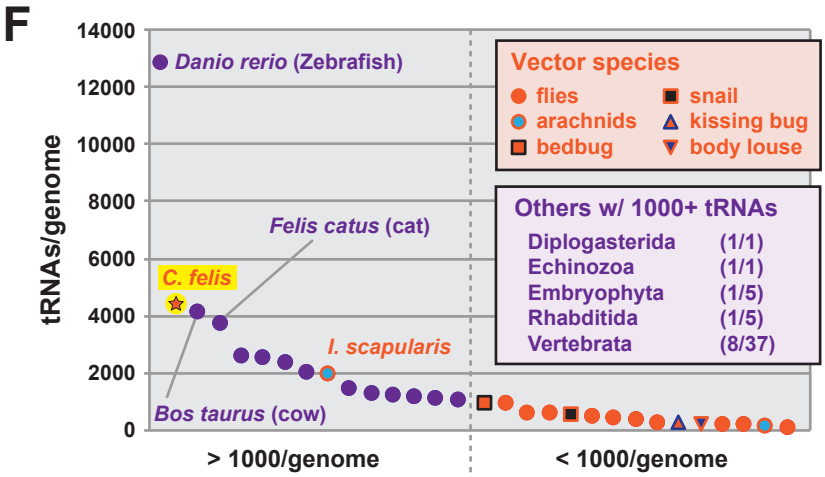


Fig. 1

F



946 **Fig. 2. Evidence for excessive copy number variation in the *C. felis* genome.** (**A**) Flea
947 genome size estimates. Flow cytometer-based estimates were performed for male and female
948 individuals of *X. cheopis* (Texas), *C. felis* (Texas), and for female *C. felis* EL from the sequenced
949 colony (see **Additional file 3: Fig. S2**). The inset (top right) depicts the coefficients of variation
950 in measured fluorescence (relative fluorescence units; RFU) for *Drosophila melanogaster*
951 (n=26), *D. viridis* (n=26), and *C. felis* EL (n=26) females prepared and analyzed simultaneously.
952 (**B**) Graphic depiction of assembling CNV. Two theoretical individual fleas are shown with
953 different CNVs for loci A and B. Regions unique to each individual genome are shown by the
954 red dashed boxes. Only reads concordant between individuals are included in the conglomerate
955 assembly. (**C**) Comparison of Illumina read coverage-mapping between duplicate genes (blue)
956 and single-copy genes (green) at different %ID thresholds. Reads that mapped to multiple
957 locations (alternative mappings) were included. Asterisks indicate statistically significant
958 difference (Welch Two-Sample t-test, $p < 2.2e-16$) between mean coverage of single-copy and
959 duplicate genes at the 90 %ID threshold. (**D**) Transcriptional support for *C. felis* EL genes
960 within the 1KITE transcriptomic data. Counts of transcripts per million reads (TPM) were
961 mapped (Hisat2 & Stringtie), binned, and plotted against the number of duplicated (blue) and
962 single-copy (green) genes in the BIG9 assembly. (**E**) Extent of truncation within clusters of
963 duplicated genes in *C. felis*. The number of clusters with truncated members at each integer %ID
964 threshold (left) was calculated as the proportion of total clusters at that threshold (center). The
965 distribution of length differences in these clusters (relative to the longest member in each cluster)
966 is plotted as a violin plot (right); black diamonds represent the mean length difference at each
967 %ID threshold.

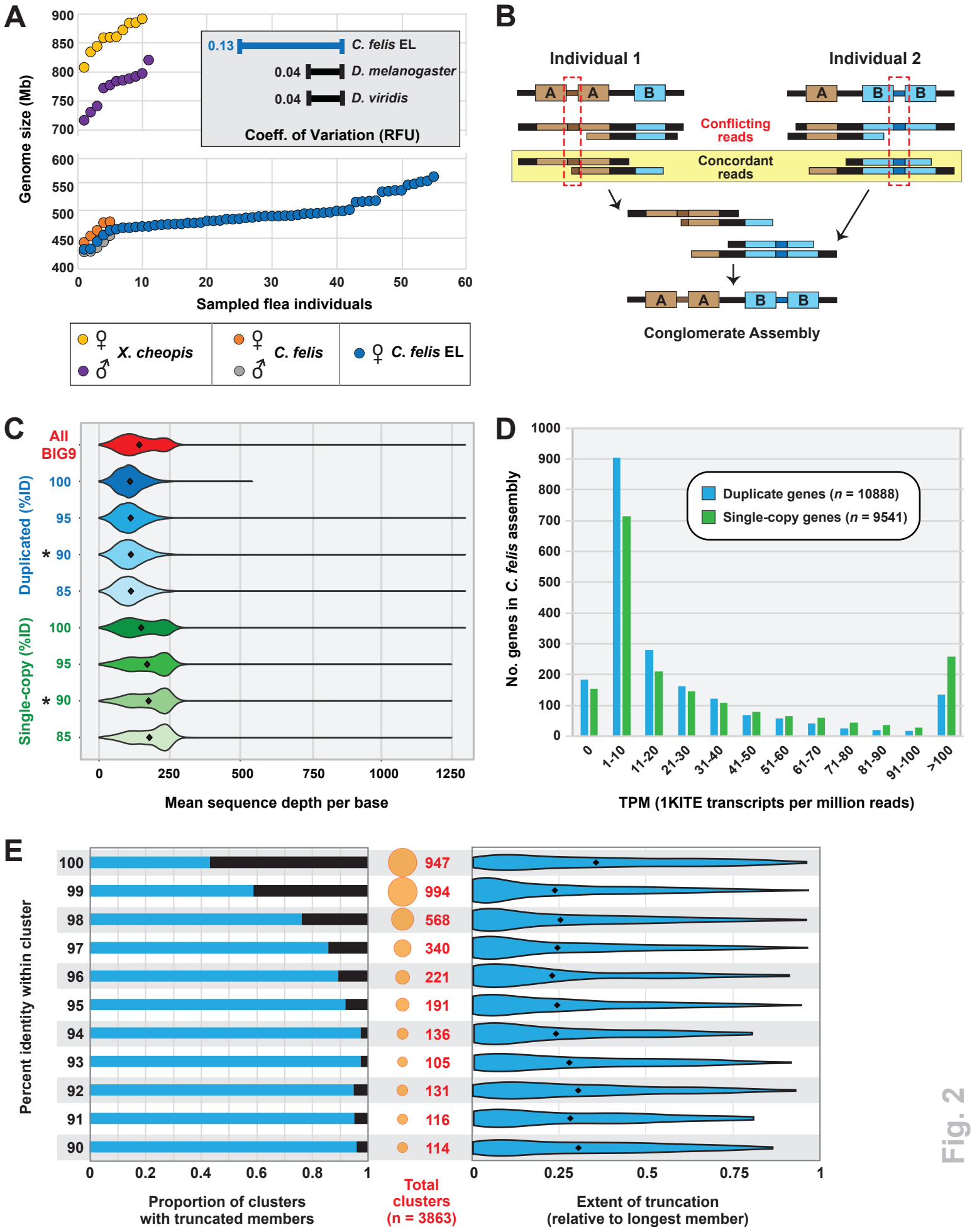
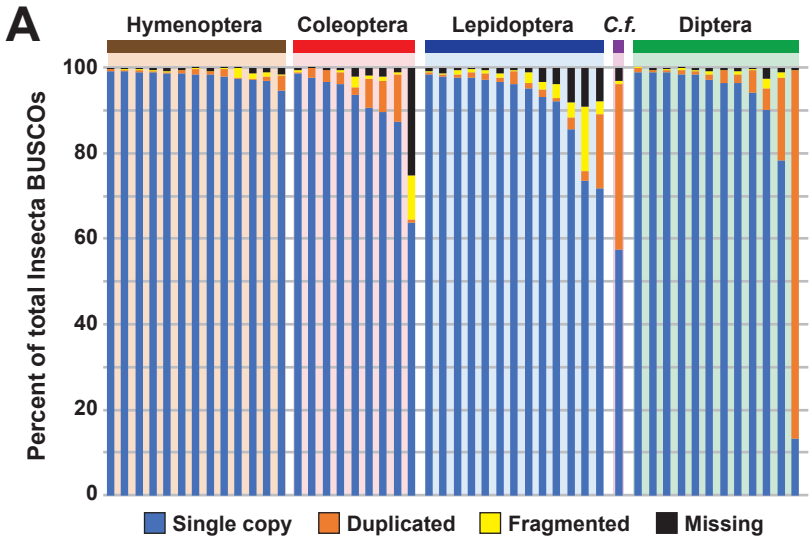
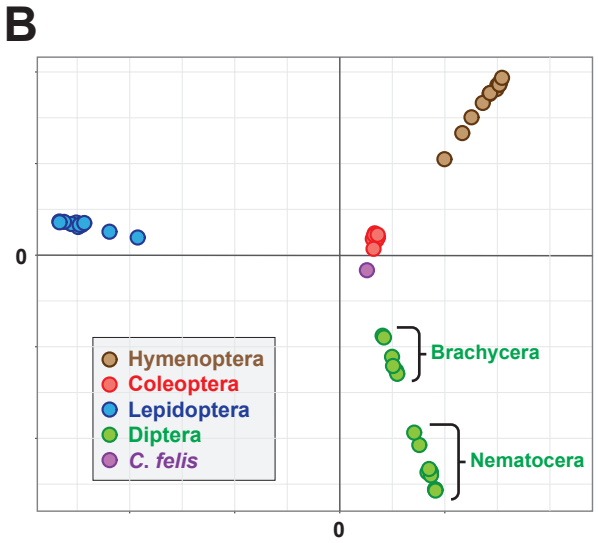

A
Genome size (Mb)
900
850
800
750
700
600
550
500
450
400
0.13 C. felis EL
0.04 D. melanogaster
0.04 D. viridis
Coeff. of Variation (RFU)
0
10
20
30
40
50
60
Sampled flea individuals
♀ ♂ X. cheopis
♀ ♂ C. felis
♀ C. felis EL
B
Individual 1
Individual 2
A
A
B
A
B
B
Conflicting reads
Concordant reads
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

969 **Fig. 3. Phylogenomics analysis of the *C. felis* genome**. (**A**) Assessing completeness and
970 conservation of select holometabolan genomes using insect (n=1,658) Benchmarking Universal
971 Single-Copy Orthologues (BUSCOs) [28]. (**B**) Multidimensional scaling plots gauging within-
972 and across-order similarity of protein orthologous groups. Inset show color scheme for
973 holometabolous orders. (**C**) Upset plot illustrating *C. felis* protein orthologous groups that
974 intersect with other holometabolous insects. Inclusion criteria: one protein from at least one
975 genome/order must be present. Yellow bar, 577 proteins found in all analyzed genomes. Inset,
976 redrawn phylogeny estimation of Holometabola [2]; numbers indicate *C. felis* unique protein
977 groups or higher-generic monophyletic groups (see **Additional file 5: Table S2**).
978


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

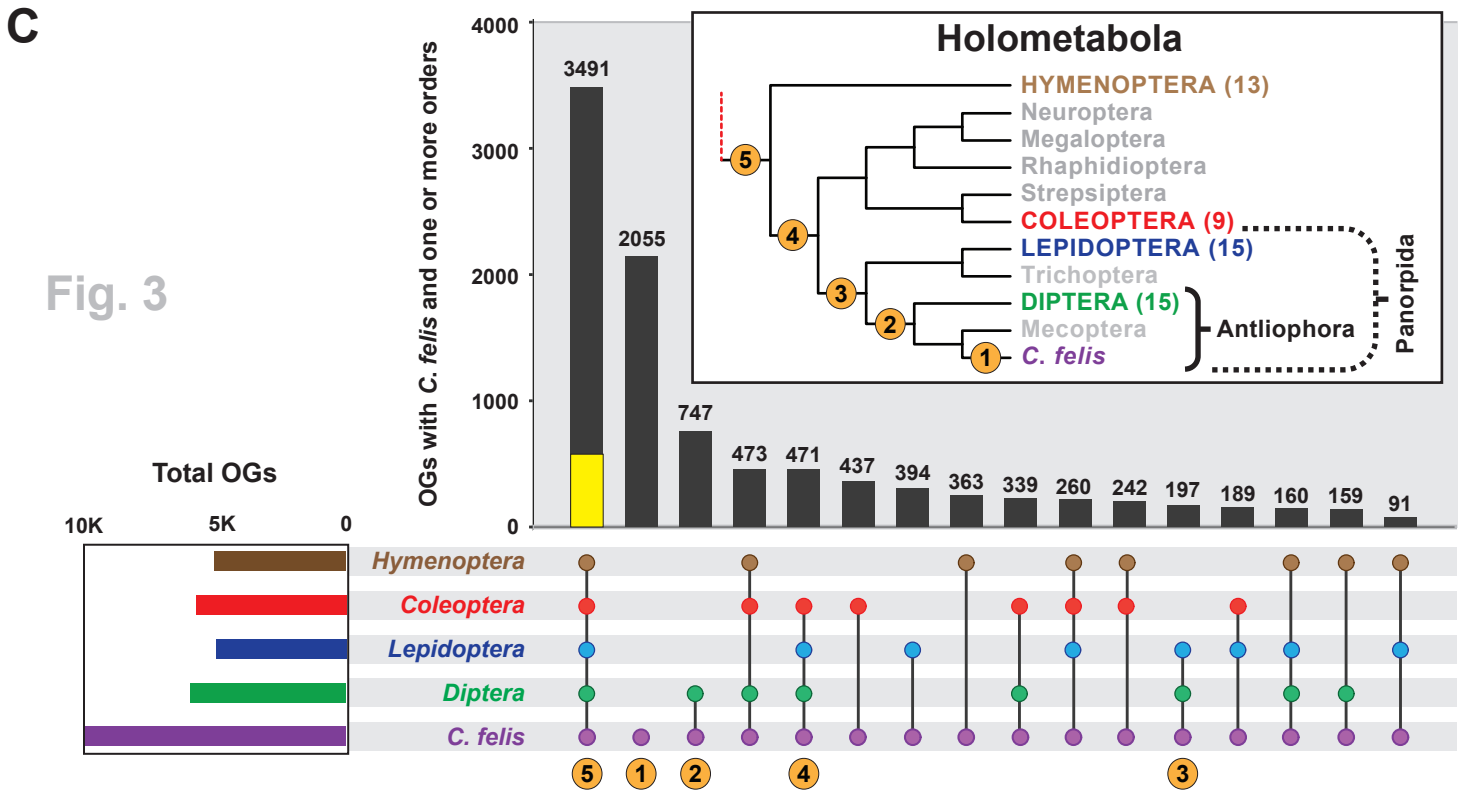
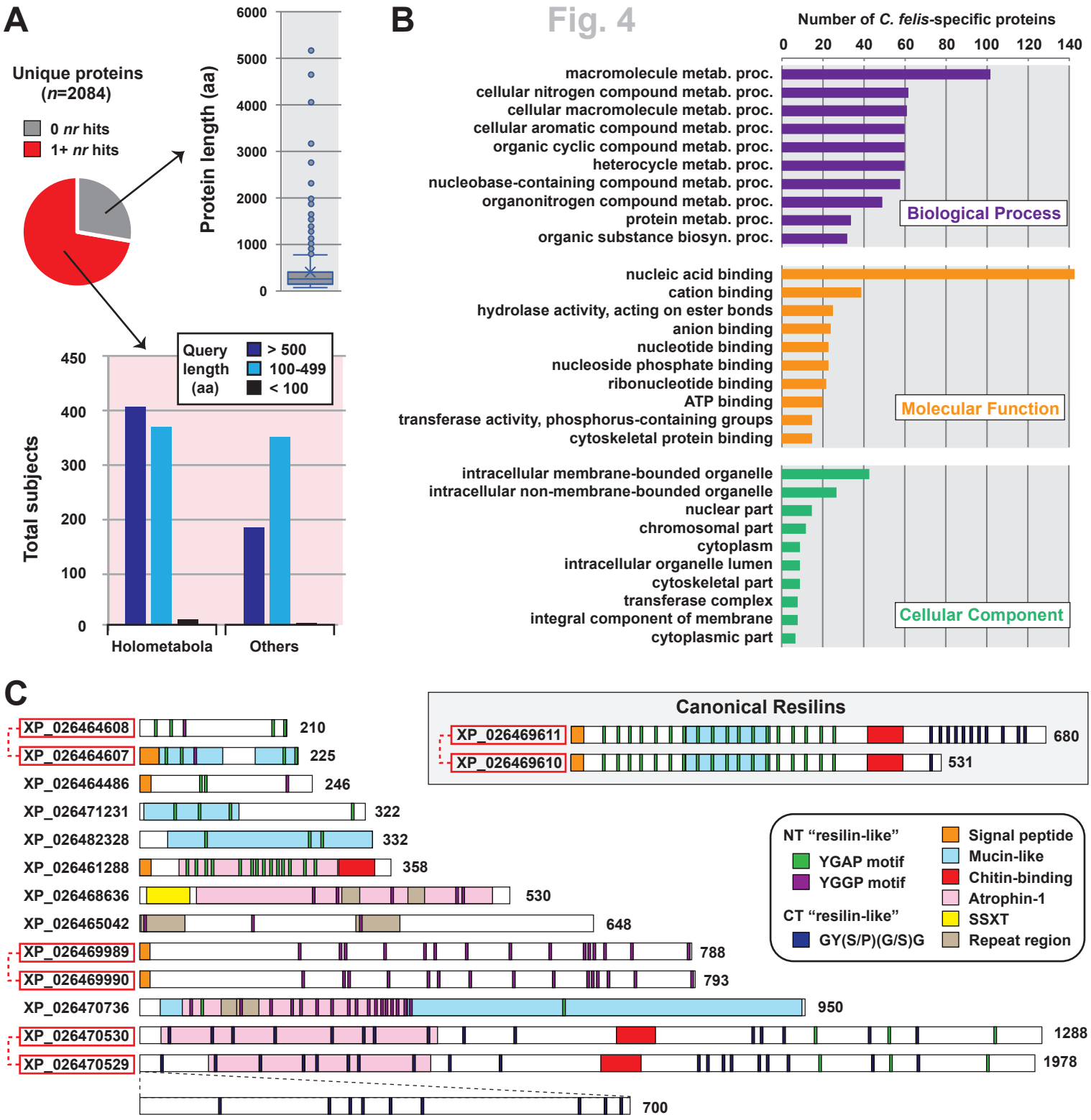
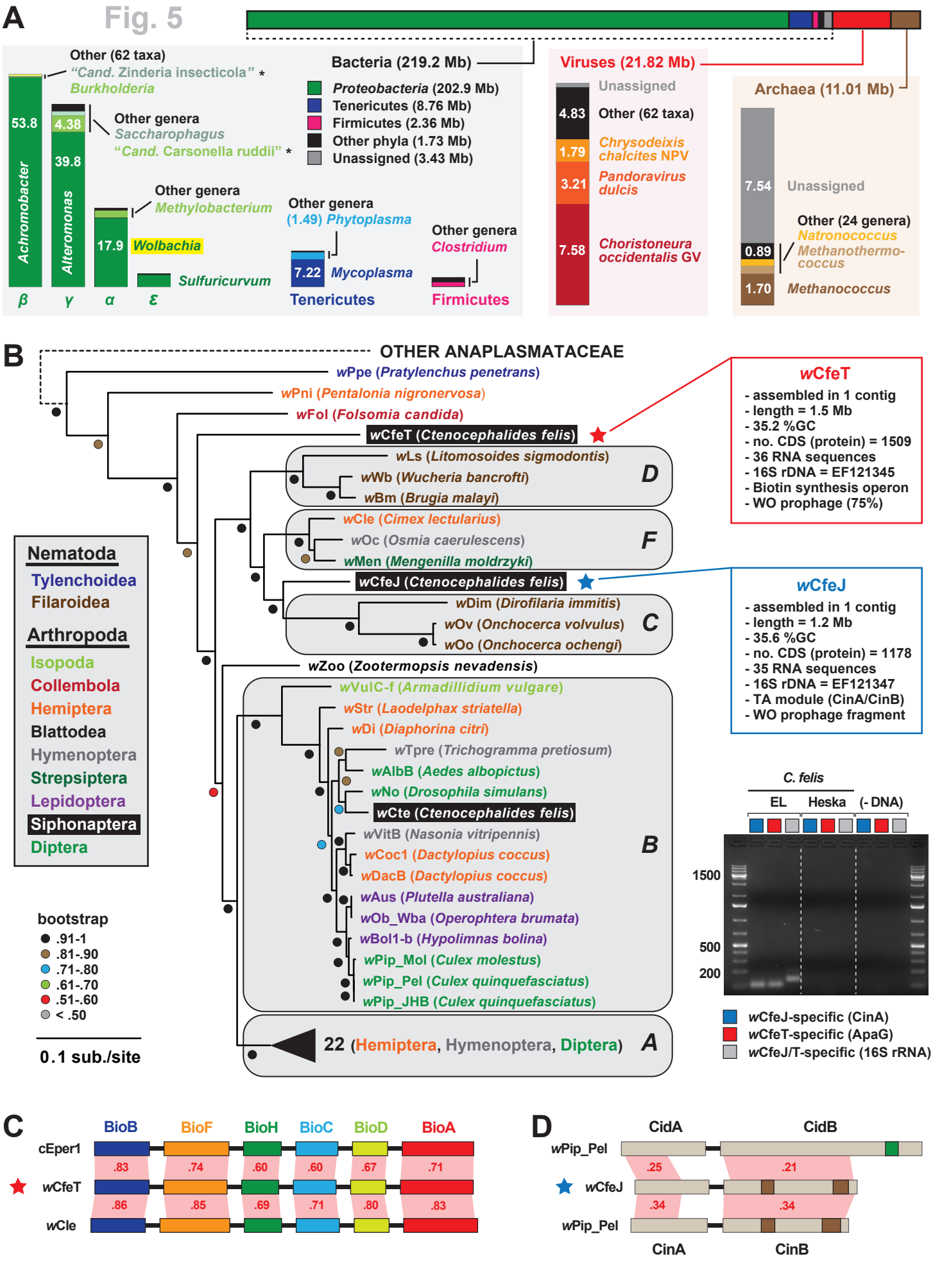

C
Holometabola
HYMENOPTERA (13)
Neuroptera
Megaloptera
Rhaphidioptera
Strepsiptera
COLEOPTERA (9)
LEPIDOPTERA (15)
Trichoptera
DIPTERA (15)
Mecoptera
C. felis
Antliophora
Panorpida
OGs with C. felis and one or more orders
4000
3000
2000
1000
0
3491
2055
747
473
471
437
394
363
339
260
242
197
189
160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

Fig. 3

980 **Fig. 4. Identifying *C. felis*-specific genes.** (**A**) *C. felis* proteins failing to cluster with
981 counterparts in other holometabolan genomes were determined to lack (top) or possess limited
982 similarity to (bottom) proteins from holometabolan or other genomes (bottom). (**B**) For 1,318
983 proteins, Gene Ontologies and Interpro domains were included in annotation and clustering into
984 broad cellular function categories. (**C**) *C. felis* carries tandemly-arrayed resilin homologs (gray
985 inset) as well as a cohort of other proteins containing resilin-like features. Red boxes indicate
986 other tandemly-arrayed genes.
987


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

989 **Fig. 5. The microbiome of EL fleas.** (**A**) Breakdown of the *C. felis* (EL fleas) microbiome. Bar
990 at top graphically depicts the taxonomic distribution of non-flea Illumina reads across Bacteria,
991 viruses and Archaea. Each group is further classified, with the major taxa (genus-level in most
992 cases) and compiled read size (Mb) provided. Taxa with asterisks are AT-rich genomes that
993 were later determined to match to *C. felis* mitochondrial reads. (**B**) *Wolbachia* genome-based
994 phylogeny estimation. *Wolbachia* supergroups are within gray ellipses. *C. felis*-associated
995 Wolbachiae are within black boxes. Red (*w*CfeT) and blue (*w*CfeJ) stars depict the two novel
996 Wolbachiae infecting *C. felis*, with assembly information for each genome provided at right.
997 Inset: color scheme for nematode and arthropod hosts. For tree estimation see **Methods**. Gel
998 image (unaltered) depicts PCR results using 100ng of flea template DNA (quantified via
999 nanodrop) in separate reactions with gene-specific primers. (**C**) *w*CfeT contains the unique biotin
1000 synthesis operon carried by certain obligately host-associated microbes. Schema follows our
1001 previous depiction of the unique *bio* gene order [44], with all proteins drawn to scale (as a
1002 reference, *w*CfeT BioB is 316 aa). Comparisons are made to the *bio* proteins of *Cardinium*
1003 endosymbiont of *Encarsia pergandiella* (cEper1, CCM10336-CCM10341) and *Wolbachia*
1004 endosymbiont of *Cimex lectularius* (*w*Cle, BAP00143-BAP00148). Red shading and numbers
1005 indicate % identity across pairwise protein alignments (blastp). (**D**) *w*CfeJ contains a CinA/B
1006 operon. Comparisons are made to the CidA/B (top, CAQ54390/1) and CinA/B (bottom) operons
1007 of *Wolbachia* endosymbiont of *Culex quinquefasciatus* Pel (wPip_Pel, CAQ54402/3). Green,
1008 CE clan protease; brown, PD-(D/E)XK nuclease. All proteins are drawn to scale (as a reference,
1009 *w*CfeJ CinB is 777 aa). Red shading and numbers indicate % identity across pairwise protein
1010 alignments (blastp).
1011


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

1012 **Table 1. Evidence Supporting Extensive Gene Duplication in Cat Fleas.**
1013

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 >10Mb are littered with gene duplications, which comprise 38% of protein coding genes. - No misassembly of allelic variants in the BIG 9 scaffolds.
Transcript mapping	Fig. 2D	- 98% duplicate genes have transcriptional support from RNA-Seq data from an independent study (1KITE).
Short read mapping	Fig. 2C	- Illumina reads maps 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.

1014 1015 1016 1017 1018 1019 1020 1021 1022 1023 1024 1025 1026 1027 1028 1029 1030 1031 1032 1033

1034 **Additional file 1: Fig. S1.** Assessing assembly fragmentation, gene duplication and repeat
1035 elements within the *C. felis* assembly. (**A**) Evaluating assembly fragmentation via mapping of
1036 scaffolds shorter than 1 Mb (n = 3,724) to scaffolds larger than 1 Mb (n = 9, "BIG9 scaffolds").
1037 All but 2 short scaffolds mapped to a BIG9 scaffold at least once; confidence intervals are based
1038 on the probability of mapping to a single unique location. (**B**). Assessing the "genome
1039 completeness" of the *C. felis* full assembly and BIG9 scaffolds through comparison to eukaryote,
1040 arthropod and insect BUSCOs. (**C**) Tandem and proximal duplicate gene locations on BIG9
1041 scaffold 1, (**D**) BIG9 scaffold 2, (**E**) BIG9 scaffold 3, (**F**) BIG9 scaffold 4, (**G**) BIG9 scaffold 5,
1042 (**H**) BIG9 scaffold 6, (**I**) BIG9 scaffold 7, (**J**) BIG9 scaffold 8, (**K**) BIG9 scaffold 9. (**L**)
1043 Duplications by proximity. Only true duplications (n=2) are shown. Red bars (*) depict
1044 "dispersed" clusters that span multiple scaffolds. (**M**) Dispersed duplicate gene locations across
1045 BIG9 scaffolds. (**N**) Distribution across BIG9 scaffolds of *C. felis* proteins annotated as "DNA
1046 integration" (GO:0015074, see **Additional file 2: Table S1** for specific accession numbers) and
1047 their relation to gene duplications. (**O**) Compilation of retroelements, DNA transposons and
1048 other repeat elements predicted across the BIG9 scaffolds. Overall totals are highlighted yellow.
1049 (**P**) tRNA gene abundances and (**Q**) codon usage/amino acid for select Holometabola.
1050

1051 **Additional file 2: Table S1.** Functional predictions and enrichment analysis of *C. felis* proteins.
1052 <click for link to Table S1>
1053

1054 **Additional file 3: Fig. S2.** Representative histograms produced by flow cytometry showing the
1055 peak positions of the 2C nuclei of *Drosophila melanogaster* (left) and *D. virilis* (center) female
1056 standards, and individual *C. felis* females (right) from the sequenced EL strain. (**A**) A 434 Mb

1057 flea. (**B**) A 553 Mb flea. All peaks have CV < 1.5 and > 500 nuclei under the statistical gates
1058 (red lines spanning the 2C peaks).
1059
1060 **Additional file 4: Fig. S3.** Phylogenomics analysis of select Holometabola. (**A**) Assessment of
1061 holometabolan accessory genomes. (**B**) *Top:* Identification of conserved protein families present
1062 in select taxa from each holometabolan order but absent from *C. felis*. *Bottom:* Protein families
1063 conserved across all sequenced holometabolan genomes except *C. felis* (see **Additional file 5:**
1064 **Table S2**). Four assemblies were identified as particularly patchy (*Oryctes borbonicus*,
1065 *Operophtera brumata*, *Heliothis virescens*, and *Plutella xylostella*) and 100% conservation
1066 ("perfect") was also relaxed to exclude these taxa. Inset, redrawn phylogeny estimation of
1067 Holometabola [2].
1068
1069 **Additional file 5: Table S2.** Pan-genomes across sequenced Holometabola.
1070 <click for link to Table S2>
1071
1072 **Additional file 6: Table S3.** Analysis of *C. felis* proteins that did not cluster with other
1073 Holometabola.
1074 <click for link to Table S3>
1075
1076 **Additional file 7: Table S4.** Elements of the *C. felis* microbiome and associated *Wolbachia*
1077 phylogeny estimation.
1078 <click for link to Table S4>
1079

1080 **Additional file 8: Table S5.** Coverage of corrected PacBio reads against all 16,622 polished

1081 assembly contigs.

1082 <click for link to Table S5>

1083