High-quality genomes reveal significant genetic divergence and cryptic speciation in the model organism Folsomia candida (Collembola)

Yun-Xia Luan[1], Yingying Cui[1], Wan-Jun Chen[2], Jianfeng Jin[3], Ai-Min Liu[4], Cheng-Wang Huang[5], Mikhail Potapov[6], Yun Bu[7], Shuai Zhan[5], Feng Zhang[3], and Sheng Li[1]

[1]South China Normal University
[2]BGI-Shenzhen
[3]Nanjing Agricultural University
[4]South China Agricultural University
[5]Chinese Academy of Sciences
[6]Moscow State Pedagogical University
[7]Shanghai Natural History Museum

December 22, 2021

Abstract

The collembolan Folsomia candida Willem, 1902, is an important representative soil arthropod that is widely distributed throughout the world and has been frequently used as a test organism in soil ecology and ecotoxicology studies. However, it is questioned as an ideal "standard" because of differences in reproductive modes and cryptic genetic diversity between strains from various geographical origins. In this study, we present two high-quality chromosome-level genomes of F. candida, for the parthenogenetic Danish strain (FCDK, 219.08 Mb, N50 of 38.47 Mb, 25,139 protein-coding genes) and the sexual Shanghai strain (FCSH, 153.09 Mb, N50 of 25.75 Mb, 21,609 protein-coding genes). The seven chromosomes of FCDK are each 25–54% larger than the corresponding chromosomes of FCSH, showing obvious repetitive element expansions and large-scale inversions and translocations but no whole-genome duplication. The strain-specific genes, expanded gene families and genes in nonsyntenic chromosomal regions identified in FCDK are highly related to its broader environmental adaptation. In addition, the overall sequence identity of the two mitogenomes is only 78.2%, and FCDK has fewer strain-specific microRNAs than FCSH. In conclusion, FCDK and FCSH have accumulated independent genetic changes and evolved into distinct species since diverging 10 Mya. Our work shows that F. candida represents a good model of rapidly cryptic speciation. Moreover, it provides important genomic resources for studying the mechanisms of species differentiation, soil arthropod adaptation to soil ecosystems, and Wolbachia-induced parthenogenesis as well as the evolution of Collembola, a pivotal phylogenetic clade between Crustacea and Insecta.

High-quality genomes reveal significant genetic divergence and cryptic speciation in the model organism *Folsomia candida*(Collembola)

Yun-Xia Luan[1], Yingying Cui[1], Wan-Jun Chen[2], Jian-Feng Jin[3], Ai-Min Liu[4], Cheng-Wang Huang[5], Mikhail Potapov[6], Yun Bu[7], Shuai Zhan[5], Feng Zhang[3], Sheng Li[1, 8]

[1] Guangdong Provincial Key Laboratory of Insect Development Biology and Applied Technology, Institute of Insect Science and Technology, School of Life Sciences, South China Normal University, Guangzhou, China

[2] BGI-Shenzhen, Shenzhen, China

[3] Department of Entomology, College of Plant Protection, Nanjing Agricultural University, Nanjing, China

[4] Department of Pomology, College of Horticulture, South China Agricultural University, Guangzhou, China

[5] CAS Key Laboratory of Insect Developmental and Evolutionary Biology, CAS Center for Excellence in Molecular Plant Sciences, Chinese Academy of Sciences, Shanghai, China;

[6] Moscow Pedagogical State University, Moscow, Russia

[7] Natural History Research Center, Shanghai Natural History Museum, Shanghai Science & Technology Museum, Shanghai, China

[8] Guangmeiyuan R&D Center, Guangdong Provincial Key Laboratory of Insect Developmental Biology and Applied Technology, South China Normal University, Meizhou, China

Correspondence

Yun-Xia Luan: yxluan@scnu.edu.cn

Sheng Li: lisheng@scnu.edu.cn

Feng Zhang: fzhang@njau.edu.cn

Luan, Cui, Chen and Jin contribute equally to this work.

Running title

Comparative genomics in *Folsomia candida*

Abstract

The collembolan *Folsomia candida* Willem, 1902, is an important representative soil arthropod that is widely distributed throughout the world and has been frequently used as a test organism in soil ecology and ecotoxicology studies. However, it is questioned as an ideal "standard" because of differences in reproductive modes and cryptic genetic diversity between strains from various geographical origins. In this study, we present two high-quality chromosome-level genomes of *F. candida* , for the parthenogenetic Danish strain (FCDK, 219.08 Mb, N50 of 38.47 Mb, 25,139 protein-coding genes) and the sexual Shanghai strain (FCSH, 153.09 Mb, N50 of 25.75 Mb, 21,609 protein-coding genes). The seven chromosomes of FCDK are each 25–54% larger than the corresponding chromosomes of FCSH, showing obvious repetitive element expansions and large-scale inversions and translocations but no whole-genome duplication. The strain-specific genes, expanded gene families and genes in nonsyntenic chromosomal regions identified in FCDK are highly related to its broader environmental adaptation. In addition, the overall sequence identity of the two mitogenomes is only 78.2%, and FCDK has fewer strain-specific microRNAs than FCSH. In conclusion, FCDK and FCSH have accumulated independent genetic changes and evolved into distinct species since diverging 10 Mya. Our work shows that *F. candida*represents a good model of rapidly cryptic speciation. Moreover, it provides important genomic resources for studying the mechanisms of species differentiation, soil arthropod adaptation to soil ecosystems, and *Wolbachia* -induced parthenogenesis as well as the evolution of Collembola, a pivotal phylogenetic clade between Crustacea and Insecta.

KEY WORDS: cryptic species; chromosome-level genome; genome synteny; repetitive element expansion; comparative mitogenomics; miRNA distribution

1 INTRODUCTION

Collembolans have a long evolutionary history (~410 Mya) and are among the most abundant arthropods on Earth (more than 9,000 known species) (Bellinger et al., 1996-2021). Most species in this group consume fungi in soil and leaf litter; they have radiated from the littoral zone to mountaintops and are particularly abundant in epiphytes of tropical rainforests. Collembolans are an integral component of soil ecosystems and are vulnerable to the effects of soil contamination. The abundance, diversity and molecular data of collembolans have been widely used to assess the environmental impacts of a range of pollutants on soils

(Hopkin, 1997; Gunstone et al., 2021). Collembola represents a monophyletic lineage forming an early branch off the line leading to insects. Understanding the evolution of collembolans is pivotal to clarifying the origin of insects and the early diversification of Hexapoda (Nardi et al., 2003; Misof et al., 2014).

The collembolan *Folsomia candida* Willem, 1902, is widely distributed in soils throughout the world and plays an important role in soil ecosystems (Fountain & Hopkin, 2005). The species is easy to maintain in the laboratory and has been used as a "standard" test organism for more than 50 years in the fields of population biology, evolutionary ecology, soil biology and ecotoxicology (ISO 11267:2014). In addition, *F. candida* is increasingly being used in bioassays of soil remediation methods (Lock & Janssen, 2003) and risk assessments of industrial chemicals and genetically modified crops (Huang et al., 2019).

As a cosmopolitan species, several strains of *F. candida* from different geographical origins have been collected and used by many laboratories (Tully et al., 2006; Tully & Potapov, 2015). Most strains of *F. candida* show parthenogenesis, which is probably induced by the endosymbiont *Wolbachia* (Ma et al., 2016). *F. candida*is questioned as an ideal "standard" because of the cryptic genetic diversity revealed in different strains based on COI and COII gene (Frati et al., 2004), RAPD-PCR and 18S/28S rDNA marker analyses (Tully et al., 2006). Tully and Potapov (2015) comprehensively compared the morphological characteristics of eleven clonal strains from Europe and America and one sexual lineage from our lab in Shanghai (FCSH). They identified all 66 individuals (8 to 12 individuals per strain) as the previously recognized *F. candida* morphospecies *sensu stricto* , with characteristic long furca, the same sensillar chaetotaxy and ventral setae on the third thoracic segment (Potapov & Yan, 2012). However, most studied characters vary among individuals and show overlap between strains. FCSH showed the greatest morphological peculiarity, and differed from all parthenogenetic strains by number of setae on thorax, dens and manubrium (Tully & Potapov, 2015). These previous studies indicate the possibility of an interesting evolutionary scenario in the cryptic speciation of *F. candida* .

Despite the importance of *F. candida* in evolution and ecotoxicology studies, the genomic resources of *F. candida* are limited. Faddeeva-Vakhrusheva et al. (2017) reported a reference genome for *F. candida* (FCBL, Berlin strain) of 221.7 Mbp, comprising 162 scaffolds. They found that substantial gene family expansions were linked to the stress response and that over 800 genes related to lignocellulose degradation had been acquired by horizontal gene transfer (HGT). It will be very useful to further obtain and compare high-quality genome information from different strains of *F. candida* to understand its speciation pattern and evolutionary history and to reveal the molecular mechanisms of its response to environmental stressors and reproductive regulation. In this study, by incorporating chromatin conformation (Hi-C) data and sequences obtained from the PacBio and Illumina platforms, we present chromosome-level genome assemblies of the parthenogenetic Danish strain (FCDK) and the sexual Shanghai strain of*F. candida* (FCSH). Interestingly, our comparative genome analyses revealed that FCDK and FCSH separated as early as 10 million years ago (Mya). Despite minor morphological divergences, FCDK and FCSH have accumulated striking genetic differences, including different genome sizes, chromosome structures, gene numbers, mitogenome sequences, and microRNA (miRNA) distributions, suggesting they have evolved into two separate species.

2 MATERIALS AND METHODS

Sample collection and rearing

The parthenogenetic Danish strain FCDK was originally obtained in 2007 from Paul Henning Krogh's lab at Aarhus University, Denmark, and the sexual Shanghai strain of FCSH was collected in 2008 from Shanghai Botanic Garden, China. Both strains have been cultured in our laboratory for over 10 years. They are reared in Petri dishes containing a solidified mixture of plaster of Paris and activated charcoal (9:1 wt/wt, dissolved in distilled water), fed granulated dried baker's yeast and distilled water, and kept in an artificial climate chamber in total darkness (20 ± 1°C, 80% relative humidity).

Comparison of morphological differences

Specimens were mounted on slides in Hoyer's solution. Their morphological characteristics were carefully

studied and compared using a NIKON E600 phase contrast microscope, and photographs were taken with a Nikon DXM1200 digital camera.

Genome sequencing

Total genomic DNA of FCDK and FCSH was extracted from mixed eggs or juveniles, respectively, using a DNeasy Boold & Tissue kit (Qiagen) following the manufacturer's protocol. The Illumina sequencing of FCSH was performed by Geneworks (Suzhou, China) using the Illumina HiSeq X platform. The Pacbio sequencing of FCSH and both the Illumina and PacBio sequencing of FCDK were performed by Macrogen (Korea) using the Illumina HiSeq X and PacBio Sequence RSII (P6C4 chemicals) platforms, respectively. The Hi-C experiments were performed by Frasergen (Wuhan, China) following the standard procedure using the Illumina HiSeq X platform. The detailed statistics of all sequencing data are shown in Table 1.

Genome assembly

PacBio long reads were assembled using Flye v2.7.1 (Kolmogorov et al., 2019), with a minimum overlap between reads of 1,000 and two rounds of self-polishing ('-m 1000 -i 2'). Primary contigs were polished with two iterations of Illumina short reads using NextPolish v1.3.1 (Hu et al., 2020). Quality control was performed for short reads prior to polishing using the 'bbduk.sh' script in BBTools package v38.82 (Bushnell, 2014): quality trimming (> Q20), length filtering (> 15 bp), polymer trimming (> 10 bp) and correction of overlapping paired reads. Redundant haplotypic duplications were removed using Purge_Dups v1.0.1 (Guan et al., 2020) with the default settings. All sequence alignment tasks were performed using Minimap2 v2.17 (Li, 2018) within the above polishing and purging progress. For Hi-C scaffolding, read alignment to the assembly, duplicate removal, and Hi-C contact extractions were executed using Juicer v1.6.2 (Durand et al., 2016) employing BWA v0.7.17 (Li & Durbin, 2009) as the aligner. We then used the 3D-DNA v180922 pipeline (Dudchenko et al., 2017) to anchor contigs to generate pseudochromosomes. Possible assembly errors, such as misjoins, translocations, and inversions, were manually corrected using the Assembly Tools module within Juicebox v1.11.08 (Durand et al., 2016). Potential contaminants were detected using MMseqs2 v11 (Steinegger & Söding, 2017) to perform BLASTN-like searches against the NCBI nucleotide (nt) and UniVec databases. Genome quality was further evaluated based on genome completeness and the mapping rate of raw reads. Genome completeness was assessed using BUSCO v3.0.2 (Waterhouse et al., 2018) against the arthropod gene set (arthropoda_odb10, n = 1,013). Raw PacBio and Illumina reads were aligned to the assembly using Minimap2, with the mapping rate calculated with SAMtools v1.9 (Danecek et al., 2021).

Genome annotation

We annotated repetitive elements, noncoding RNAs (ncRNAs) and protein-coding genes (PCGs) in both genomes. A *de novo* repeat library was constructed using RepeatModeler v2.0.1 (Flynn et al., 2020), and an additional LTR discovery pipeline was also applied ('-LTRStruct'). We then combined the *de novo* library with the Dfam 3.1 (Hubley et al., 2016) and RepBase-20181026 databases (Bao et al., 2015) to construct a custom repeat library, which was employed to mask repeats in the genome using RepeatMasker v4.0.9 (Smit et al., 2013–2015). We scanned ncRNAs using Infernal v1.1.3 (Nawrocki & Eddy, 2013) and tRNAscan-SE v2.0.7 (Chan & Lowe, 2019); low-confidence tRNAs were filtered with the tRNAscan-SE built-in script 'EukHighConfidenceFilter'.

We predicted PCGs using the MAKER v3.01.03 pipeline (Holt & Yandell, 2011), which included the EVidenceModeler (EVM, Haas et al., 2008) module. *Ab initio* , transcriptome and protein homology-based evidence were employed to support the predicted gene models. *Ab initio* gene models were generated using BRAKER v2.1.5 (Brůna et al., 2021) (a combination of two *ab initio* predictors, Augustus v3.3.4 (Stanke et al., 2004) and GeneMark-ES/ET/EP 4.68_lic (Brůna et al., 2020)); BRAKER can simultaneously incorporate transcriptome and protein evidence to improve prediction accuracy. The input transcriptome alignments were produced using HISAT2 v2.2.0 (Kim et al., 2019), and arthropod proteins were mined from the OrthoDB10 v1 database (Kriventseva et al., 2019) as a reference. Transcriptome evidence (transcripts) was assembled using the genome-guided assembler StringTie v2.1.4 (Kovaka et al., 2019). Protein sequences of *Drosophila melanogaster* , *Tribolium castaneum* , *Bombyx mori* ,*Apis mellifera* and *Daphnia magna* were

downloaded from NCBI and fed to MAKER as evidence of protein homology. EVM was also activated with the weights of *ab initio* prediction, transcripts and proteins set to 1, 2 and 8, respectively. We assigned gene functions using Diamond v2.0.8 (Buchfink et al., 2021) to search the UniProtKB database; the more sensitive mode and an e-value of 1e-5 were used ('–more-sensitive -e 1e-5'). We further identified protein domains and assigned Gene Ontology (GO) and pathway (KEGG, Reactome) annotations using eggNOG-mapper v2.0.1 (Huerta-Cepas et al., 2017) and InterProScan 5.47–82.0 (Finn et al., 2017). Five databases were included in the InterProScan analyses: Pfam (El-Gebali et al., 2019), SMART (Letunic & Bork, 2018), Superfamily (Wilson et al., 2009), Gene3D (Lewis et al., 2018), and CDD (Marchler-Bauer et al., 2017).

Inter- and intra-genomic synteny

To reveal chromosomal evolution between FCDK and FCSH, syntenic blocks were detected using MCScanX (Wang et al., 2012) with an e-value of 1e-10 and a minimum syntenic size of 5 ('-b 2 -s 5 -e 1e-10'). Protein sequence alignments were generated by MMseqs2 BLASTP-like searching in the sensitive mode (-s 7.5) with an e-value of 1e-5. Large chromosomal regions unique to FCDK (no syntenic blocks were defined in FCSH) were surveyed. GO and KEGG pathway enrichment analyses were performed for PCGs located in these regions using clusterProfiler v3.14.3 (Yu et al., 2012). The significance level was controlled to the default, with a p-value of the hypergeometric distribution of 0.01 and a q-value for the multiple comparison FDR of 0.05.

Whole-genome duplication

Li et al. (2018) inferred the occurrence of a whole-genome duplication (WGD) event in *F. candida* , but Roelofs et al. (2020) refuted this assumption and proposed a large number of small-scale gene duplications. We tested the WGD hypothesis in both genomes using three methods: 1) examination of patterns of collinearity within genomes: intraspecific genomic collinear blocks were identified as in the above syntenic analyses except that the mode was intraspecific rather than interspecific ('-b 1'); 2) evaluation of the paranome distributions of synonymous distances (Ks, Blanc & Wolfe, 2004): the paralogue Ks distribution and WGD signal were calculated from coding sequences using wgd v1.1.1 (Zwaenepoel & Van de Peer, 2019), and kernel density estimation (KDE) and BGMM mixture modelling were used to fit Ks distributions; and 3) observation of the number of HOX gene clusters (Ferrier & Minguillón, 2003): Hox genes were manually annotated by TBLASTN-like MMseqs2 searches, with a sensitive mode and an e-value of 1e-10, using reference HOX protein sequences of *F. candida* mined from the NCBI. Synteny, circus and HOX distribution figures were visualized using TBtools v1.095 (Chen et al., 2020).

Orthology identification and phylogenetic inference

We inferred PCG sequence orthology across eleven arthropod species: one crustacean (*D. magna*), one dipluran (*Catajapyx aquilonaris*), four insects (*Zootermopsis nevadensis* , *T. castaneum* , *A. mellifera* , *D. melanogaster*), and five collembolans (*Sinella curviseta* , *Orchesella cincta* ,*Holacanthella duospinosa* , FCDK, FCSH). Protein sequences of*C. aquilonaris* and *H. duospinosa* were downloaded from i5K, those of *S. curviseta* were obtained from FigShare (https://doi.org/10.6084/m9.figshare.7286231.v2), and other data were procured from NCBI. After removing redundant isoforms, orthogroups (gene families) were inferred using OrthoFinder v2.5.2 (Emms & Kelly, 2019), and Diamond was employed for sequence alignment in ultrasensitive mode ('-S diamond_ultra_sens').

Single-copy orthologues estimated with OrthoFinder were used to infer phylogeny and divergence times. We aligned the protein sequences of each orthologue using MAFFT v7.394 (Katoh & Standley, 2013) with the high-accuracy L-INS-I method, trimmed unreliable homologous sites using BMGE v1.12 (Criscuolo & Gribaldo, 2010) with stringent parameters ('-m BLOSUM90 -h 0.4'), and concatenated individual alignments into a matrix. We then estimated substitution models and partitioning schemes and reconstructed the phylogeny using IQ-TREE v2.0.7 (Minh et al., 2020); genes that violated SRH (stationary, reversible and homogeneous) assumptions were excluded ('–symtest-remove-bad –symtest-pval 0.10'); to reduce the computational burden, the model was restricted to LG ('–mset LG'), and the top 10% of partitioning schemes were considered

('–rclusterf 10'); ultrafast bootstrap and SH-like approximate likelihood ratio tests were calculated to assess node support ('-B 1000 –alrt 1000'). We estimated divergence times using MCMCTree within the PAML v4.9j package (Yang 2007); the JC69 substitution model, the independent rate clock model, and the approximate likelihood calculation and ML estimation of branch lengths were applied. We repeated the runs at least twice to ensure convergence, and each ran for 60,000 generations, with the first 10,000 considered burn-in. Five fossils from the PBDB database (*https://www.paleobiodb.org/navigator/*) were applied for node calibration: one Branchiopoda (<541 Mya), one Hexapoda (<485.4 Mya), the most recent common ancestor (MRCA) of Diplura and Insecta (>407.6 Mya), one Holometabola (315.2-382.7 Mya), and the MRCA of Coleoptera and Diptera (>295.5 Mya).

Gene family evolution

We inferred gene family evolution (expansion and contraction) based on the dating tree using CAFE v4.2.1 (Han et al., 2013), and the mode of the single birth-death parameter 'lambda' was employed with a significance level of 0.01. PCGs from significantly expanded and species-specific gene families were further enriched for GO and KEGG terms using clusterProfiler with the default parameters (p-value of 0.01, q-value of 0.05).

Annotation of xenobiotic detoxification-related gene families

In contrast to insects, xenobiotic detoxification-related gene families are greatly expanded in Collembola, possibly due to their adaptations to complex soil environments (Faddeeva-Vakhrusheva et al., 2017; Manni et al., 2020). The copy numbers of these families of FCDK and FCSH may be different, since the parthenogenetic strains show a wider distribution and better adaptability than the sexual strains of *F. candida* . We annotated the genes of five detoxification-related families, including the cytochrome P450 (CYP), ATP-binding cassette transporter (ABC), carboxyl/cholinesterase (CCE), UDP-glycosyltransferase (UGT), and glutathione-S-transferase (GST) families, using the BITACORA v1.3 (Vizueta et al., 2020) pipeline, and we further manually checked them. BITACORA performed initial BLASTP searches of the annotated proteins generated via the automatic MAKER pipeline and TBLASTN analyses in the genome assembly and confirmed the gene models with protein domains in each family via HMMER searches (Altschul, 1997; Eddy, 2011). Reference protein sequences of *D. melanogaster* , *B. mori* and *F. candida* for the ABC, CCE, GST and UGT families were obtained from the NCBI RefSeq database, whereas CYP sequences were mined from Dermauw et al. (2020). HMM profiles of each family were downloaded from the PFAM database: ABC (PF00005), CCE (PF00135), GST (PF14497, PF02798), CYP (PF00067), and UGT (PF00201). A cut-off e-value of 1e-5 was applied for BLAST and HMM searches. A close proximity algorithm was used to predict novel genes from TBLASTN alignments with a maximum intron length of 15,000 bp. The resulting CYP sequences were manually examined based on conserved protein structures, which were characterized by a four-helix bundle (D, E, I and L), helices J and K, two sets of β sheets and a coil 'meander'. The functions of predicted proteins were checked via online BLASTP analysis in the nonredundant protein database (nr). The classification of each family and possible sequence errors were assisted by constructing phylogenetic trees. To construct the phylogenies of five gene families, the amino acid sequences of each family were aligned using MAFFT via the L-INS-I method and trimmed using trimAl v1.4.1 (Capella-Gutiérrez et al., 2009) with the 'gappyout' mode strategy. Phylogenetic trees were constructed using IQ-TREE, with automatic model selection and 1,000 ultrafast bootstrap replicates. Tree figures were enhanced using online EvolView v3 (Subramanian et al., 2019).

Comparative analysis of mitochondrial genomes

The mitochondrial genomes of FCDK (GenBank accession No. OL672753) and FCSH (GenBank accession No. OL672754) were *de novo* assembled from Illumina raw reads (approximately 6 GB raw data) using NOVOPlasty v2.7.2 (Dierckxsens et al., 2017). The assembled mitochondrial genomes were annotated via the MITOS2 web portal (Donath et al. 2019) and manually checked. The circular representation of the mitochondrial genome was plotted using the CGview Server (Grant & Stothard, 2008). The alignment of the two whole mitochondrial genomes was built using the Mauve plugin for Geneious Prime (Kearse et al. 2012), and detailed sequence identities were calculated with custom Python scripts.

Comparative analysis of miRNAs

The evolutionary divergence of miRNAs may contribute to the diversification of species (Plasterk et al., 2006). Liu et al. (2020) analysed the miRNA data of FCDK based on genome and small RNA (sRNA) data from mixed embryos, juveniles and adults (GenBank accession No. SRP132383). In this study, we *de novo* sequenced sRNAs of mixed individual specimens of FCSH, including living embryos, juveniles and adults in different developmental stages (GenBank accession No. SRP308782). Based on the sRNA data and chromosomal-level genome assemblies of FCDK and FCSH, known miRNAs with homologous genes in the miRBase 21 database (Kozomara & Griffiths-Jones, 2014) were identified, and novel miRNAs were predicted according to the procedure described in Liu et al. (2020). By mapping miRNA precursors to the corresponding genome using MapChart (Voorrips, 2002), the locations of miRNAs were recorded and compared between FCDK and FCSH. The miRNA nomenclature followed previous studies (Fromm et al., 2015; Liu et al., 2020). The prefix MIR- indicates family names, and Mir- indicates gene names. Duplicated paralogous genes were identified with the letter 'p' and a number.

3. RESULTS

Morphological differences between FCDK and FCSH

We checked more than 50 individuals of each strain and found two morphological characteristics that could distinguish between adults of FCDK and FCSH (Figure 1a-h): the numbers of seta on the retinaculum (2 or 3 for FCSH vs. only 1 for FCDK), which was not included by Tully & Potapov (2015), and the numbers of ventral setae on the third segment of the thorax (3(4-6)+3(4-6) for FCSH vs. 2(3)+2(3) for FCDK), which is consistent with the observation of character 11 in Tully & Potapov (2015). However, the numbers of setae on dens and manubrium vary greatly between individuals, and we have not confirmed the differences corresponding characters 12 and 17 in Tully & Potapov (2015).

Chromosome-level genome assembly and annotation

We produced 121.19 Gb and 62.21 Gb of total sequencing data for the genome assemblies of FCDK and FCSH, respectively (Table 1). After primary assembly, polishing, redundancy removal, Hi-C scaffolding, and contaminant detection, we generated two highly contiguous, nearly complete chromosome-level genomes. Detailed assembly statistics are summarized in Table 1. FCDK and FCSH had genome sizes of 219.08 Mb and 153.90 Mb, scaffold N50 lengths of 38.47 Mb and 25.75 Mb, and GC contents of 37.49% and 38.54%, respectively. More than 97% of the genome was anchored to seven pseudochromosomes in both strains (Figure 1i, Table 1). Each chromosome of FCDK was 24.65% - 54.11% longer than the corresponding homologous chromosome of FCSH (Figure 1j, Table S1). High integrity was revealed by high ratios of single-copy BUSCO genes (97.3% and 97.0%), very low ratios of duplicated genes (0.8% and 0.9%) indicated no obvious redundancy in the assemblies, and high mapping ratios of long and short reads (> 94%) confirmed the high quality of the two assemblies. In addition, a scaffold of ˜0.5 Mb in size corresponding to a *Wolbachia*endosymbiont was detected in the FCDK assembly, showing great similarity (99.9%) to *Wolbachia* sequences assembled from the FCBL strain (Faddeeva-Vakhrusheva et al., 2007).

We masked repetitive regions with 22.61% (49.53 Mb) and 10.03% (15.43 Mb) of the genomes of FCDK and FCSH, respectively (Table 1, Table S2 and Table S3). DNA, LINE, LTR and unclassified transposon elements were significantly enriched at the FCDK-specific regions of Chr1, 3, 4, and 7 (Figure 1i). Relative to FCSH, many repeat families of FCDK were obviously expanded (Figure 1k), particularly families such as TcMar-Tc1, CMC-EnSpm, Penelope, Gypsy, and Pao.

Using the Infernal and tRNAscan-SE automatic prediction pipelines, 396 and 334 ncRNAs were identified in FCDK and FCSH, respectively (Table 1, Table S4 and Table S5). Both strains possessed 21 isotypes of tRNA and lacked Supres. With regard to snRNAs, FCDK/FCSH exhibited 33/21 spliceosomal RNAs (U1, U2, U4, U5, U6, and U11), 3/3 minor spliceosomal RNAs (U4atac, U6atac, and U12), 6/7 C/D box small nucleolar RNAs (snoRNAs), 2/3 H/ACA box snoRNAs, and 1/1 other snoRNA (SCARNA8).

We predicted 25,139 and 21,609 PCGs for FCDK and FCSH, respectively (Table 1 and Table S6). The

annotation statistics of the two strains were very similar in terms of the mean lengths of genes (˜4,000 bp), exons (˜250 bp) and CDSs (˜200 bp) and the mean numbers of exons (˜8) and introns (˜6.5) per gene. However, FCDK showed a longer mean intron length than FCSH (312.1 vs. 263.5 bp). The BUSCO completeness of predicted proteins exceeded 97% for both strains. In addition, the distribution patterns of PCGs and transposons on chromosomes showed the opposite trends; i.e., chromosomal regions of high gene density usually showed a low transposon density and vice versa (Figure 1i). Protein domains of approximately 2/3 of the genes of both strains were identified by InterProScan, and nearly 1/2 of the predicted genes were annotated in GO and KEGG pathways (Table S6).

Intergenomic synteny and whole-genome duplication

All seven corresponding chromosomes between the FCDK and FCSH genomes exhibit conserved syntenic relationships (Figure 2a). MCScanX identified 212 syntenic blocks containing 27,559 (59.55%) PCGs from the two genomes. Both chromosomes 1 and 4 of FCDK showed perfect matches with the corresponding chromosomes of FCSH; however, some obvious large-scale chromosomal structural variations were present between other corresponding chromosomes of the two strains. Large-scale inversions occurred at least one, three, two, and four times on corresponding chromosomes 2, 3, 5 and 7, respectively, of FCDK and FCSH. Meanwhile, many translocation events occurred on corresponding chromosomes 6 and 7, respectively of the two strains.

Several large distal regions of FCDK chromosomes 1 (˜10.51 Mb), 3 (˜3.76 Mb), 4 (˜7.71 Mb) and 7 (˜9.13 Mb) did not share any syntenic blocks with FCSH (Figure 2a). The top enrichment GO terms of ˜2,300 PCGs located in these FCDK-specific regions were mainly related to nutrient responses, immunity, stress tolerance and protein conformation, etc. (Figure 2b), which may associate with the better adaptation of this strain to the environment relative to FCSH. Most of other enriched GO terms (Figure 2b) and most enriched KEGG pathways (Figure 2c) were associated with viruses, bacteria and fungi as well as interactions between these organisms and FCDK. A large number of genes may have been generated from ancient horizontal gene transfers (HGTs), as revealed by Faddeeva-Vakhrusheva et al. (2017) in the FCBL strain.

We tested the WGD hypotheses using within-genome collinear patterns, paranome Ks distributions, and the estimated number of HOX gene clusters. First, 14 syntenic blocks (198 genes) were identified within the FCDK genome, and nine (131 genes) were found within the FCSH genome (Figure 1i), among which only nine FCDK blocks were distributed on different chromosomes. Most of the blocks occurred in genomic regions with a high density of transposons. Second, KDEs fit the distributions very well without obvious peaks in the Ks distributions (Figure 2d). When BGMM mixture components were fit to Ks distributions, none of the four default hypotheses (1˜4 components/WGDs) was found to be predominant (Figure S2). Third, only a single, complete HOX gene cluster was identified on chromosome 5 (Figure 1i), with a length of approximately 5 Mb, in both the FCDK and FCSH genomes. The Hox gene cluster showed the same order of *Scr* , *Ftz* , *Antp* ,*Ubx* , *lab* , *pb* , *Hox3* , *Dfd* , *Abd-A*and *Abd-B* reported in a previous study (Faddeeva-Vakhrusheva et al., 2007). Therefore, three lines of evidence (i.e., few within-genome syntenic blocks, an absence of obvious peaks in Ks distribution plots, and single HOX clusters in the genome) absolutely supported an absence of WGD events in *F. candida* .

Orthology inference, phylogeny and gene family evolution

We inferred PCG orthology across ten insects and one crustacean and clustered 90.1% of the genes into 21,942 orthogroups (gene families) (Figure 3a, Table S7). Among these groups, 4,009 orthogroups, including 1,372 single-copy orthogroups, were present in all species, and 4,640 orthogroups, containing 20,304 genes, were species-specific. A total of 526 orthogroups, containing 4,601 genes, were unique to Collembola. FCDK exhibited 13,970 gene families, 572 (2,285 genes) of which were unique; FCSH exhibited 13,675 gene families, including 204 unique families (820 genes) (Table S7).

After aligning, trimming and filtering, a final matrix of 499,899 amino acid sites from 1,304 single-copy genes was used for phylogenetic inference and dating estimates. The identified phylogenetic relationships were consistent with recent phylogenomic studies (Misof et al., 2014): Collembola was located at the base

of Hexapoda, and Diplura was sister to insects (Figure 3a). *Folsomia* (Isotomoidea) separated from Entomobryoidea in the Late Triassic-Early Jurassic (186.6-216.8 Mya). FCDK and FCSH diverged in the middle Neogene (11.7-14.7 Mya), indicating that the two strains had been separated for long enough to be considered independent species.

Gene family evolution analyses with CAFE identified a large number of significantly expanded families in most collembolan species (i.e., Entomobryomorpha). Among these families, a total of 224 and 81 were expanded in FCDK and FCSH, respectively (Figure 3a, Table S8). The two strains shared large expanded cytochrome P450, ABC transporter, zinc finger, Sec14, pickpocket, lactase-phlorizin hydrolase, chitin-binding type-2 domain-containing protein, F-box, and ionotropic receptor families (Figure 3b, c; Table S9), which have generally been found to be expanded in two other Entomobryomorpha species (Faddeeva-Vakhrusheva et al., 2017; Zhang et al., 2019) and play an important role in the adaptive evolution of Collembola. Relative to FCSH, FCDK showed additional expansions of histone, glutathione S-transferase, lytic polysaccharide monooxygenase (LPMO), exoskeleton protein, bacillopeptidase F, beta-lactamase, tenascin, ATP-dependent DNA helicase, down syndrome cell adhesion molecule-like protein, chymotrypsinogen, gustatory receptor and neuroligin sequences, which are related to genetic modification, detoxification, cuticle and nervous system development, digestion, chemosensation, antibiotic biosynthesis and lignocellulose degradation (Figure 3b, c). In addition to terms related to the regulation of translation and transcription factors, GO and KEGG enrichment analyses of expanded and species-specific families involved in symbiotic interactions and related biological processes was performed (Figure S3a-d). However, these terms were absent or generally received few annotations in FCSH (Figure S3e-h).

Annotation of detoxification-related gene families

The detoxification of xenobiotics is essential for the environmental adaptation of arthropods, including collembolans. The expansion of detoxification-related gene families is often observed in highly polyphagous species. We manually annotated ABC, CCE, CYP, GST and UGT genes in the two *F. candida*strains (Table 2, Figure 4, Figure S4). In contrast to *D. melanogaster* , the ABC, CCE, CYP and GST families were greatly expanded in the two *F. candida* strains. FCDK showed greater expansion of the ABC, CYP and GST families than FCSH, consistent with the wider distribution of the former strain.

ABC transporters function as primary active transporters that transport diverse substrates across lipid membranes and include eight subfamilies, A-H (Dermauw & Van Leeuwen, 2014). The gene numbers of ABCB, ABCD, ABCE and ABCF subfamilies are similar among FCDK, FCSH and *D .melanogaster* (Table 2, Figure 4a). *D. melanogaster* shows the fewest ABCG and ABCH genes but the greatest number of ABCA genes, which are related to lipid trafficking (Wenzel et al., 2007). ABCG genes exhibit diverse functions in eye colouration determination (Ewart et al., 1994; Mackenzie et al., 1999), moulting hormone regulation (Hock et al., 2000; Broehan et al., 2013) and xenobiotic resistance (Labbe et al., 2011). In contrast to FCSH, FCDK shows a major expansion of ABCC subfamily, which are functionally diverse, playing roles in processes comprising ion transport, cell-surface receptor activity and the translocation of a broad range of substrates (Dean et al., 2001; Kruh & Belinsky, 2003; Moreau et al., 2005).

CCEs hydrolysing carboxylic esters can be classified into three larger classes: dietary/detoxification, hormone/semiochemical processing and neurodevelopmental functions (Oakeshott et al., 2005). Dietary/detoxification CCEs accounted for 2/3 of all annotated CCEs in FCDK and FCSH. The two *F. candida* strains possessed five times as many dietary/detoxification-related and two times as many hormone/semiochemical CCEs as *D. melanogaster* (Table 2). No significant difference in CCE copy numbers was observed between FCDK and FCSH (Figure S4a).

The CYP superfamily is one of the most groups of important xenobiotic metabolism enzymes in arthropods and is usually classified into CYP2, CYP3, CYP4 and mitochondrial clans (Feyereisen 2006, 2012). For the CYPs identified in both *F. candida* strains, approximately 1/2 of the genes were clustered into the CYP2 clan, whereas few genes were included in the mitochondrial clan (Table 2, Figure 4b). The large number

of CYP2 genes found in *F. candida* was similar to that reported in Chelicerata (Grbić et al., 2011; Fan et al., 2021) but different from that in insects. CYP2 enzymes are associated with the detoxification and bioactivation of certain exogenous chemicals (Feyereisen 2006). An obvious expansion of the clan CYP3 was also observed in FCDK and FCSH; these enzymes have been found to be major participants associated with xenobiotic metabolism and insecticide resistance across a wide range of artificial or natural chemicals. FCSH showed a contraction in the CYP4 clan relative to FCDK and *D. melanogaster* .

GST enzymes play a crucial role in the detoxification of a wide range of exogenous and endogenous compounds, particularly in insecticide resistance (Enayati et al., 2005; Li et al., 2007). GSTs can be classified into three superfamilies: cytosolic (13 classes), mitochondrial (kappa), and microsomal (MAPEG) (Oakley, 2011). Few microsomal GSTs were observed, and mitochondrial GSTs were completely lacking in the three taxa. Seven cytosolic classes were observed, among which the epsilon class was absent in FCDK and FCSH and mu was absent in*D. melanogaster* (Table 2, Figure S4b). The GST class pattern observed in the two *F. candida* strains strongly conformed to that observed in noninsect arthropods (Labbé et al., 2011; Roncalli et al., 2015; Fan et al., 2021). *Folsomia* showed a great expansion of the sigma class, which accounted for more than half of the annotated GSTs. Sigma GSTs have anti- and proinflammatory functions in mammals, immune response functions in helminth parasites, and lipid oxidation product detoxification functions in insects (Flanagan & Smythe, 2011).

Insect UGTs have diverse functions, including roles in processes such as detoxification (Smith, 1955), insecticide resistance (Lee et al., 2005), pigmentation (Hopkins & Ahmad, 1991), and sclerotization (Hopkins, 1992). FCDK and FCSH each had 48 UGTs, which was slightly greater than the number in *D. melanogaster* (Table 2). Most *Folsomia*UGTs were clustered into two clades with node support of 100 and were isolated from *D. melanogaster* UGTs, indicating an independent origin of *Folsomia* UGT families (Figure S4c).

Comparative analysis of mitochondrial genomes between FCDK and FCSH

The mitochondrial genomes of both FCDK and FCSH share the ancestral gene order found in Pancrustracea. The genome sizes of the two strains were slightly different, at 15,141 bp for FCSH and 15,177 bp for FCDK, with a 36 bp insertion in the FCDK mitochondrial sequence between trnQ and trnM (Figure 5a, b). Fourteen out of 37 mitochondrial genes shared the same gene length in FCDK and FCSH. The overall sequence identity between the two mito-genomes was 78.2% (Figure 5c). The average pairwise identity was 77.13% for 13 CDSs, 87.0% for 22 tRNAs, 82.6% for 2 rRNAs, and 60.2% for the noncoding regions. The pairwise identity of the standard barcoding region (a 658 bp sequence in the 5' region of the *cox1* gene) in these two sibling species was 80.9%. Such great sequence divergence would be unlikely to be attributable to intraspecies variation. Strengthened by the mitochondrial genome analyses, FCDK and FCSH should be regarded as two distinct species.

miRNA comparison between FCDK and FCSH

FCDK and FCSH shared 91 miRNAs from 64 miRNA gene families (Figure 6a), including 78 known miRNAs from 53 families conserved in crustaceans and hexapods, and 13 miRNAs from 10 predicted families (MIR-fca1 to MIR-fca10) originated in the ancestor of FCDK and FCSH, since none of the miRNAs in the latter group were found in other collembolan genome data by the BLAST algorithm. Notably, seven known conserved miRNAs (MIR-bg5, MIR-927, MIR-3049, MIR-6012, MIR-971, MIR-11, and MIR-995a) were annotated in the sRNA data of FCSH but not in the sRNA or genomic data of FCDK. These miRNAs have probably been lost in FCDK.

By mapping miRNA precursors to the corresponding chromosomes, it was found that the locations of most shared miRNAs in FCDK and FCSH showed evolutionarily conserved synteny (Figure 6b). Only three *F. candida* -specific miRNAs (fca6-p1, fca6-p2 and fca4-p4) were located on different chromosomes of FCDK and FCSH.

It is of note that there were many more specific miRNAs in FCSH than in FCDK. Thirty-nine miRNAs from 15 predicted families (MIR-fcaSH-n1 to MIR-fcaSH-n15) were unique to FCSH, but only 9 miRNAs from 6

predicted families (MIR-fcaDK-n1 to MIR-fcaDK-n6) were unique to FCDK. We think that these miRNAs are strain specific since they could not be found in other collembolan genomes. Most FCDK-specific miRNA families had only one copy, with the exception of MIR-fcaDK-n3. However, many FCSH-specific miRNA families showed multiple duplications (Figure 6a) and tended to form clusters on chromosome 7 (Figure 6b). The remarkable differences in the miRNA distribution in FCSH and FCDK confirm that they have experienced independent evolutionary processes and evolved into different species.

4 DISCUSSION

Based on the comparative genomic analyses of the two strains of *F. candida* , we found significant differences between them in terms of both genome structure and gene distributions. The genome of FCDK (219.08 M) was 65.18 M (42.4%) larger than that of FCSH (153.90 M); half of this difference consisted of repetitive elements (34.1 M); and each chromosome of FCDK was 25-54% longer than the corresponding chromosome of FCSH (Figure 1). The seven pairs of chromosomes showed obvious large-scale chromosomal inversions and translocations, and several large distal regions of FCDK chromosomes 1, 3, 4 and 7 did not share any syntenic blocks with FCSH (Figure 2). There were 3,530 more PCGs annotated in FCDK than in FCSH (Table 1). The numbers of FCDK-specific gene families and genes (572/2285) were far greater than the numbers of FCSH-specific sequences (204/820) (Table S7). In addition, FCDK and FCSH showed obvious divergence between their mitochondrial genomes and miRNA distributions. The overall sequence identity of the two mitogenomes was only 78.2%, and the pairwise identity of their standard DNA barcodes (658 bp of COI) was 80.9%, which is much lower than the similarity observed in individuals from the same insect species (Hebert et al., 203). Although 91 common miRNAs of FCDK and FCSH showed conserved synteny on their homologous chromosomes, there were many more strain-specific miRNAs in FCSH (39 miRNAs from 15 predicted families) than in FCDK (9 miRNAs from 6 predicted families). Moreover, most FCDK-specific miRNA families had only one copy, but many FCSH-specific miRNA families showed multiple duplications, which tended to form clusters on chromosome 7 (Figure 6). All of these genetic differences demonstrated that after 10 million years of divergence (Figure 3), FCDK and FCSH have evolved into distinct species.

In *F. candida* , parthenogenetic strains usually show a wider distribution and stronger environmental adaptability than bisexual strains (Tully & Potapov, 2015). The genomic differences revealed in this study demonstrated obvious evolutionary advantages of the parthenogenetic FCDK genome. Relative to bisexual FCSH, FCDK showed more strain-specific and expanded gene families, many of which were involved in resistance to environmental xenobiotics, such as the ABC, CYP and GST families (Table 2). Interestingly, histone H2A, one of the core histones, showed obvious expansion in FCDK (Figure 3b). The sequences and expression patterns of core histones can profoundly alter chromatin properties, and their posttranslational modifications, such as acetylation, methylation, and ubiquitination, are involved in the regulation of nucleosome dynamics (Lawrence et al., 2016). In particular, variants of histone H2A are associated with DNA repair and gametogenesis (Hanson et al., 2013) as well as environmental responses (Talbert & Henikoff, 2014). The potential roles of expanded H2A genes in FCDK deserve further study. In addition, FCDK genes exhibited longer introns than FCSH genes (Table 1). Introns can provide selective advantages to cells by regulating alternative splicing, gene expression, chromatin assembly, etc. (Jo & Choi, 2015), and they might play roles in determining species-specific characteristics and complexities (Carvunis et al., 2012).

The genetic drivers of species differentiation have always been a hot research topic. Our study reveals several possible factors contributing to the genetic divergence between FCDK and FCSH. First, the increase in transposon copies is a very important driving force leading to the enlargement of the genome (Chalopin et al., 2015; Talla et al., 2017), and transposable elements may contribute significantly to divergent chromatin structures (Diehl et al., 2020). In this study, we found that transposons were significantly enriched in the FCDK-specific regions of Chr1, Chr3, Chr4, and Chr7 (Figure 2), and many repeat families were obviously expanded in FCDK (Figure 3b). Most chromosomal regions with a high transposon density showed a low gene density and high recombination (Figure 1). Second, a large number of genes acquired through HGT located in FCDK-specific chromosome regions probably play important roles in the spread of antibiotic resistance and genome evolution (Faddeeva-Vakhrusheva et al., 2017). Significantly, *Wolbachia*infection may induce or

accelerate speciation and contribute to reproductive isolation (Werren, 1998), which is probably associated with increased recombination (Singh, 2019). Although we did not find any obvious genomic characteristics related to the differences in reproductive mechanisms between FCDK and FCSH based on comparative genomic analyses, our high-quality genome data provide a basis for further mechanistic studies. In addition, strain-specific miRNAs could drive the adaptive diversification of the genomic regulatory mechanisms of FCDK and FCSH. Novel miRNAs could be gained and lost relatively rapidly across closely related groups, especially when these groups undergo extensive adaptive diversification (Xiong et al., 2019). It would be interesting to further study how miRNA genes evolve and contribute to speciation.

In conclusion, we generated two high-quality chromosome-level genomes of the *Folsomia candida* DK and SH strains, revealed their genomic differences, and proposed that they have differentiated into two separate species. Our work provides important genomic resources for studying speciation and the effect of *Wolbachia* on host reproduction and genetic differentiation, as well as the basis for mechanistically understanding soil arthropods in the context of soil ecology, and the evolution links between insects and crustaceans.

ACKNOWLEDGEMENT

This study is supported by the National Natural Science Foundation of China (31970438, 31772510 and 31620103917), National Key R&D Program of China (2019YFD1002102) and the National Science & Technology Fundamental Resources Investigation Program of China (2018FY100300).

CONFLICT OF INTEREST

The authors declare that they have no competing interests.

AUTHOR CONTRIBUTIONS

Y.X.L., S.L. and F.Z. conceived and designed the project. W.J.C., C.W.H. and A.M.L. prepared samples for sequencing. C.W.H., M.P. and Y.B. performed morphological studies. F.Z., Y.X.L., Y.C, J.F.J., W.J.C. and A.M.L. performed data analyses. Y.X.L., F.Z., Y.C. and W.J.C. wrote the manuscript. S.L. and S.Z. provided critical revisions. All authors read and approved the final manuscript.

DATA ACCESSIBILITY AND BENEFIT-SHARING SECTION

Data Accessibility Statement

The sequencing reads are deposited at NCBI (SRR13452034–SRR13452037 and SRR13435515-SRR13435517) under BioProject PRJNA723214 and PRJNA686204. The genome assemblies are deposited at GenBank under the accessions JAEMPH000000000 and JAEMPG000000000.

Benefit-Sharing Statement

Benefits Generated: The research presents a good model of cryptic speciation. Benefits from this research accrue from the sharing of our data and results on public databases as described above.

REFERENCES

Altschul, S. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research* , 25(17), 3389-3402. https://doi.org/10.1093/nar/25.17.3389

Bao, W., Kojima, K.K., & Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* . 6(11), 1-6. https://doi.org/10.1186/s13100-015-0041-9

Bellinger, P. F., Christiansen, K. A. & Janssens, F. 1996-2021. Checklist of the Collembola of the World. http://www.collembola.org

Blanc G., & Wolfe K. H. (2004). Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *The Plant Cell* , 16(7), 1667-1678. https://doi.org/10.1105/tpc.021345

Broehan, G., Kroeger, T., Lorenzen, M., & Merzendorfer H. (2013). Functional analysis of the ATP-binding cassette (ABC) transporter gene family of *Tribolium castaneum* . *BMC Genomics* , 14, 6. https://doi.org/10.1186/1471-2164-14-6

Brůna, T., Lomsadze, A., & Borodovsky, M. (2020). GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins. *NAR Genomics and Bioinformatics* , 2(2), lqaa026. https://doi.org/10.1093/nargab/lqaa026

Brůna, T., Hoff, K. J., Lomsadze, A., Stanke, M., & Borodovsky, M. (2021). BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genomics and Bioinformatics* , 3(1), lqaa108. https://doi.org/10.1093/nargab/lqaa108

Buchfink, B., Reuter, K., & Drost, H. G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods* , 18, 366-368. https://doi.org/10.1038/s41592-021-01101-x

Bushnell, B. (2014). BBTools software package. https://sourceforge.net/projects/bbmap

Carvunis, A. R., Rolland, T., Wapinski, I., Calderwood, M.A., Yildirim, M. A., Simonis, N., ... Vidal, M. (2012). Proto-genes and de novo gene birth. *Nature,* 487(7407), 370-374. https://doi.org/10.1038/nature11184

Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* , 25(15), 1972-1973. https://doi.org/10.1093/bioinformatics/btp348

Chalopin, D., Naville, M., Plard, F., Galiana, D., & Volff, J. N. (2015). Comparative analysis of transposable elements highlights mobilome diversity and evolution in vertebrates. *Genome Biology and Evolution* , 7(2), 567-580. https://doi.org/10.1093/gbe/evv005

Chan, P. P., & Lowe, T. M. (2019). tRNAscan-SE: searching for tRNA genes in genomic sequences. *Methods in Molecular Biology* , 1962, 1-14. https://doi.org/10.1007/978-1-4939-9173-0_1

Chen, C., Chen, H., Zhang, Y., Thomas, H. R., Frank, M. H., He, Y., & Xia, R. (2020). TBtools: An integrative toolkit developed for interactive analyses of big biological data. *Molecular Plant* , 13(8), 1194-1202. https://doi.org/10.1016/j.molp.2020.06.009

Criscuolo, A., & Gribaldo, S. (2010). BMGE (Block Mapping and Gathering with Entropy): selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evolutionary Biology* , 10(1), 210. https://doi.org/10.1186/1471-2148-10-210

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience* , 10(2), giab008. https://doi.org/10.1093/gigascience/giab008

Dean, M., Rzhetsky, A., & Allikmets R. (2001). The human ATP-binding cassette (ABC) transporter superfamily. *Genome Research* , 11(7), 1156-1166. https://doi.org/10.1101/gr.184901

Dermauw, W., & Van Leeuwen, T. (2014). The ABC gene family in arthropods: comparative genomics and role in insecticide transport and resistance. *Insect Biochemistry and Molecular Biology* , 45, 89-110. https://doi.org/10.1016/j.ibmb.2013.11.001

Dermauw, W., Van Leeuwen, T., & Feyereisen, R. (2020). Diversity and evolution of the P450 family in arthropods. *Insect Biochemistry and Molecular Biology* , 127, 103490. https://doi.org/10.1016/j.ibmb.2020.103490

Diehl, A. G., Ouyang, N., & Boyle, A. P. (2020). Transposable elements contribute to cell and species-specific chromatin looping and gene regulation in mammalian genomes. *Nature Communication* , 11, 1796. https://doi.org/10.1038/s41467-020-15520-5

Dierckxsens, N., Mardulyn, P., & Smits, G. (2017). NOVOPlasty: de novo assembly of organelle genomes from whole genome data. *Nucleic Acids Research* , 45(4), e18. https://doi.org/10.1093/nar/gkw955

Donath, A., Jühling, F., Al-Arab, M., Bernhart, S.H., Reinhardt, F., Stadler, P.F., Middendorf, M., & Bernt, M. (2019). Improved annotation of protein-coding genes boundaries in metazoan mitochondrial genomes. Nucleic Acids Research, 47(20), 10543-10552.

https://doi.org/10.1093/nar/gkz833

Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N.C., ... Aiden, E. L. (2017). De novo assembly of the*Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* , 356(6333), 92-95. https://doi.org/10.1126/science.aal3327

Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S. P., Huntley, M., H., Lander, E. S., & Aiden, E. L. (2016). Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Systems* , 3(1), 95-98. https://doi.org/10.1016/j.cels.2016.07.002

Eddy, S. R. (2011). Accelerated profile HMM searches. *PLoS Computational Biology* , 7(10), e1002195. https://doi.org/10.1371/journal.pcbi.1002195

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., ... Finn, R. D. (2019). The Pfam protein families database in 2019. *Nucleic Acids Research* , 47(D1), D427–D432. https://doi.org/10.1093/nar/gky995

Emms, D. M., & Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology* , 20(1), 238. https://doi.org/10.1186/s13059-019-1832-y

Enayati, A. A., Ranson, H., & Hemingway, J. (2005). Insect glutathione transferases and insecticide resistance. *Insect Molecular Biology* , 14, 3-8. https://doi.org/10.1111/j.1365-2583.2004.00529.x

Ewart, G. D., Cannell, D., Cox, G. B., & Howells A. J. (1994). Mutational analysis of the traffic ATPase (ABC) transporters involved in uptake of eye pigment precursors in *Drosophila melanogaster* . Implications for structure-function relationships. *The Journal of Biological Chemistry* , 269, 10370-10377.

Faddeeva-Vakhrusheva, A., Kraaijeveld, K., Derks, M. F. L., Anvar, S. Y., Agamennone, V., Suring, W., ... Roelofs, D. (2017). Coping with living in the soil: the genome of the parthenogenetic springtail*Folsomia candida* . *BMC Genomics* , 18, 493. https://doi.org/10.1186/s12864-017-3852-x

Fan, Z., Yuan, T., Liu, P., Wang, L. Y., Jin, J. F., Zhang, F., & Zhang, Z. S. (2021) A chromosome-level genome of the spider*Trichonephila antipodiana* reveals the genetic basis of its polyphagy and evidence of an ancient whole-genome duplication event.*Gigascience* , 10(3), giab016. https://doi.org/10.1093/gigascience/giab016

Ferrier, D. E., & Minguillón, C. (2003). Evolution of the Hox/ParaHox gene clusters. *The International Journal of Developmental Biology* , 47 (7-8), 605-611.

Feyereisen, R. (2006). Evolution of insect P450. *Biochemical Society Transactions* , 34(6), 1252-1255.

Feyereisen, R. (2012). Insect CYP genes and P450 enzymes. In: L. I. Gilbert (Ed.), *Insect Molecular Biology and Biochemistry* (pp. 236-316). Academic Press.

Finn, R. D., Attwood, T. K., Babbitt, P. C., Bateman, A., Bork, P., Bridge, A. J., ... Mitchell, A. L. (2017). InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Research* , 45(D1), D190-D199. https://doi.org/10.1093/nar/gkw1107

Flanagan, J. U., & Smythe, M. L. (2011). Sigma-class glutathione transferases. *Drug Metabolism Reviews* , 43(2), 194-214. https://doi.org/10.3109/03602532.2011.560157

Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., & Smit A. F. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences* , 117(17), 9451-9457. https://doi.org/10.1073/pnas.1921046117

Fountain, M. T. & Hopkin, S. P. (2005).*Folsomia candida* (Collembola): a "standard" soil arthropod.*Annual Review of Entomology* , 50, 201-222. https://doi.org/10.1146/annurev.ento.50.071803.130331

Frati, F., Negri, I., Fanciulli, P. P., Pellecchia, M., De Paola, V., Scali, V., & Dallai, R. (2004). High levels of genetic differentiation between *Wolbachia* -infected and non-infected populations of*Folsomia candida* (Collembola, Isotomidae). *Pedobiologia* , 48, 461-468. https://doi.org/10.1016/j.pedobi.2004.04.004

Fromm, B., Billipp, T., Peck, L. E. Johansen, M., Tarver, J. E., King, B. L., Newcomb, J. M., Sempere, L. F., Flatmark, K., Hovig, E., & Peterson, K. J. (2015). A uniform system for the annotation of vertebrate microRNA genes and the evolution of the human microRNAome.*Annual Review of Genetics* , 49, 213-242. https://doi.org/10.1146/annurev-genet-120213-092023

Grant, J. R., & Stothard, P. (2008). The CGView Server: a comparative genomics tool for circular genomes, *Nucleic Acids Research* , 36(suppl_2), W181–W184. https://doi.org/10.1093/nar/gkn179

Grbić, M., Van Leeuwen, T., Clark, R. Rombauts, S., Rouzé, P., Grbić, V., Osborne, E. J., . . . Van de Peer, Y. (2011). The genome of*Tetranychus urticae* reveals herbivorous pest adaptations.*Nature* , 479, 487-492 https://doi.org/10.1038/nature10640

Guan, D., McCarthy, S. A., Wood, J., Howe, K., Wang, Y., & Durbin, R. (2020). Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* , 36(9), 2896-2898. https://doi.org/10.1093/bioinformatics/btaa025

Gunstone, T., Cornelisse, T., Klein, K., Dubey, A. & Donley. N. (2021). Pesticides and soil invertebrates: a hazard assessment. *Frontiers in Environmental Science* , 9, 643847. https://doi.org/10.3389/fenvs.2021.643847

Haas, B. J., Salzberg, S. L., Zhu, W. Pertea, M., Allen, J. E., Orvis, J., White, O., Buell, C. R., & Wortman, J. R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology* , 9, R7. https://doi.org/10.1186/gb-2008-9-1-r7

Han, M. V., Thomas, G., Lugo-Martinez, J., Hah, M. W. (2013). Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. *Molecular Biology and Evolution* , 30 (8), 1987-1997. https://doi.org/10.1093/molbev/mst100

Hanson, S. J., Stelzer, C. P., Welch, D. B. M. & Logsdon, J. M. (2013). Comparative transcriptome analysis of obligately asexual and cyclically sexual rotifers reveals genes with putative functions in sexual reproduction, dormancy, and asexual egg production. *BMC Genomics* , 14, 412. https://doi.org/10.1186/1471-2164-14-412

Hebert, P. D., Cywinska, A., Ball, S. L., deWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society B, Biological Science* , 270(1512), 313-321.*https://doi.org/10.1098/rspb.2002.2218*

Hock, T., Cottrill, T., Keegan, J., Garza D. (2000). The E23 early gene of *Drosophila* encodes an ecdysone-inducible ATP-binding cassette transporter capable of repressing ecdysone-mediated gene activation.*Proceedings of the National Academy of Sciences* , 97(17), 9519-9524. https://doi.org/10.1073/pnas.160271797

Holt, C., & Yandell, M. (2011). MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects.*BMC Bioinformatics* , 12(1), 491. https://doi.org/10.1186/1471-2105-12-491

Hopkins, T. (1992). Insect cuticle sclerotization. *Annual Review of Entomology* , 37, 273-302. https://doi.org/10.1146/annurev.ento.37.1.273

Hopkins, T. L., & Ahmad, S. A. (1991). Flavonoid wing pigments in grasshoppers. *Experientia* , 47, 1089-1091. https://doi.org/10.1007/BF01923349

Hopkin, S. (1997). *Biology of the Springtails (Insecta: Collembola)* . Oxford University Press, Oxford.

Hu, J., Fan, J., Sun, Z., & Liu S. (2020). NextPolish: a fast and efficient genome polishing tool for long read assembly.*Bioinformatics* , 36(7), 2253-2255. https://doi.org/10.1093/bioinformatics/btz891

Huang, C., Chen, W., Ke, X., Li, Y., & Luan, Y. (2019). A multi-generational risk assessment of Cry1F on the non-target soil organism *Folsomia candida* (Collembola) based on whole transcriptome profiling. *PeerJ* , 7, e6924. https://doi.org/10.7717/peerj.6924

Hubley, R., Finn, R. D., Clements, J., Eddy, S. R., Jones, T. A., Bao, W., Smit, A. F. A., & Wheeler, T. J. (2016). The Dfam database of repetitive DNA families. *Nucleic Acids Research* , 44(D1): D81-D89. https://doi.org/10.1093/nar/gkv1272

Huerta-Cepas, J., Forslund, K., Coelho, L. P., Szklarczyk, D., Jensen, L. J., von Mering, C., & Bork, P. (2017). Fast genome-wide functional annotation through orthology assignment by eggNOG-Mapper.*Molecular Biology and Evolution* , 34(8), 2115-2122. https://doi.org/10.1093/molbev/msx148

ISO (International Organization for Standardization). (2014). *Soil quality - Inhibition of reproduction of Collembola (* Folsomia candida*) by soil contaminants* (ISO Standard No. 11267:2014). https://www.iso.org/standard/57582.html

Jo, B. S., & Choi, S. S. (2015). Introns: the functional benefits of introns in genomes. *Genomics and Informatics* , 13(4), 112-118. https://doi.org/10.5808/GI.2015.13.4.112

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability.*Molecular Biology and Evolution* , 30(4), 772-780. https://doi.org/10.1093/molbev/mst010

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P., & Drummond, A. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* , 28(12), 1647–1649. https://doi.org/10.1093/bioinformatics/bts199

Kim, D., Paggi, J. M., Park, C., Bennett C. & Salzberg S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology* , 37(8):907-915. https://doi.org/10.1038/s41587-019-0201-4

Kolmogorov, M., Yuan, J., Lin, Y., Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology* , 37(5), 540-546. https://doi.org/10.1038/s41587-019-0072-8

Kovaka S, Zimin, A. V., Pertea, G. M., Razaghi, R., Salzberg S. L., & Pertea M. (2019). Transcriptome assembly from long-read RNA-seq alignments with StringTie2, *Genome Biology* , 20(1), 278. https://doi.org/10.1186/s13059-019-1910-1

Kozomara, A. & Griffiths-Jones, S. (2014). miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Research* , 42(D1), D68-D73. https://doi.org/10.1093/nar/gkt1181

Kriventseva, E. V., Kuznetsov, D., Tegenfeldt, F., Manni, M., Dias, R., Simão, F. A., & Zdobnov, E. M. (2019). OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Research* , 47(D1), D807-D811. https://doi.org/10.1093/nar/gky1053.

Kruh, G. D., & Belinsky, M. G. (2003). The MRP family of drug efflux pumps. *Oncogene* , 22, 7537-7552. https://doi.org/10.1038/sj.onc.1206953

Labbé, R., Caveney, S., & Donly, C. (2011). Genetic analysis of the xenobiotic resistance-associated ABC gene subfamilies of the Lepidoptera. *Insect Molecular Biology* , 20, 243-256. https://doi.org/10.1111/j.1365-2583.2010.01064.x

Lee, S.-W., Shono, T., Tashiro, S., & Ohta, K. (2005). Metabolism of pyraclofos in housefly, *Musca domestica* . *Journal of Asia-Pacific Entomology* , 8, 387-392. https://doi.org/10.1016/S1226-8615(08)60261-7

Lawrence, M., Daujat, S., & Schneider, R. (2016). Lateral thinking: how histone modifications regulate gene expression. *Trends in Genetics* , 32, 42-56. https://doi.org/10.1016/j.tig.2015.10.007

Letunic, L., & Bork, P. (2018). 20 years of the SMART protein domain annotation resource. *Nucleic Acids Research* , 46(D1), D493-D496. https://doi.org/10.1093/nar/gkx922

Lewis, T. E., Sillitoe, I., Dawson, N., Lam, S. D., Clarke, T., Lee, D., Orengo, C. & Lees, J. (2018). Gene3D: extensive prediction of globular domains in proteins. *Nucleic Acids Research* , 46(D1), D435-D439. https://doi.org/10.1093/nar/gkx1069

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences.*Bioinformatics* , 34(18), 3094-3100. https://doi.org/10.1093/bioinformatics/bty191

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* , 25(14), 1754-1760. https://doi.org/10.1093/bioinformatics/btp324

Li, X. C., Schuler, M. A., Berenbaum, M. R. (2007). Molecular mechanisms of metabolic resistance to synthetic and natural xenobiotics.*Annual Review of Entomology* , 52, 231-253. https://doi.org/10.1146/annurev.ento.51.110104.151104

Li, Z., Tiley, G. P., Galuska, S. R., Reardon, C. R., Kidder, T. I., Rundell, R. J., & Barker, M. S. (2018). Multiple large-scale gene and genome duplications during the evolution of hexapods. *Proceedings of the National Academy of Sciences* , 115, 4713-4718. https://doi.org/10.1073/pnas.1710791115

Liu, A., Chen, W., Huang, C., Qian, C., Liang, Y., Li, S., Zhan, S., & Luan, Y.-X. (2020). MicroRNA evolution provides new evidence for a close relationship of Diplura to Insecta. *Systematic Entomology* , 45, 365-377. https://doi.org/10.1111/syen.12401

Lock, K., & Janssen, C. R. (2003). Effect of new soil metal immobilizing agents on metal toxicity to terrestrial invertebrates. Environmental Pollution, 121(1), 123-27. https://doi.org/10.1016/s0269-7491(02)00202-6

Ma, Y., Chen, W. J., Li, Z. H., Zhang, F., Gao, Y., & Luan, Y. X. (2017). Revisiting the phylogeny of *Wolbachia* in Collembola.*Ecology and Evolution* , 7(7), 2009-2017. https://doi.org/10.1002/ece3.2738

Mackenzie, S. M., Brooker, M. R., Gill, T. R., Cox, G. B., Howells, A. J., & Ewart, G. D. (1999). Mutations in the white gene of*Drosophila melanogaster* affecting ABC transporters that determine eye colouration. *Biochimica Biophysica Acta* , 1419(2), 173-185. https://doi.org/10.1016/s0005-2736(99)00064-4

Manni, M., Simao, F. A., Robertson, H. M., Gabaglio, M. A., Waterhouse, R. M., Misof, B., Niehuis, O., Szucsich, N. U., Zdobnov, E. M. (2020). The genome of the blind soil-dwelling and ancestrally wingless dipluran*Campodea augens* : a key reference hexapod for studying the emergence of insect innovations. *Genome Biology and Evolution* , 12(1), 3534-3549. https://doi.org/10.1093/gbe/evz260

Marchler-Bauer, A., Bo, Y., Han, L., He, J., Lanczycki, C. J., ... Bryant, S. H. (2017). CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Research* , 45(D1), D200-D203. https://doi.org/10.1093/nar/gkw1129.

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., Lanfear, R. (2020). IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era.*Molecular Biology and Evolution* , 37(5), 1530-1534. https://doi.org/10.1093/molbev/msaa015.

Misof, B., Liu, S., Meusemann, K., Peters, R. S., Donath, A., Mayer, C., ... Zhou, X. (2014). Phylogenomics resolves the timing and pattern of insect evolution. *Science* , 7, 346(6210), 763-767. https://doi.org/10.1126/science.1257570

Moreau, C., Prost, A. L., Derand, R., & Vivaudou, M. (2005). SUR, ABC proteins targeted by KATP channel openers. *Journal of Molecular and Cellular Cardiology* , 38, 951-963. https://doi.org/10.1016/j.yjmcc.2004.11.030

Nardi, F., Spinsanti, G., Boore, J., Carapelli, A., Dallai, R., & Frati, F. (2003). Hexapod origins: monophyletic or paraphyletic?*Science* , 299(5614), 1887-1889. https://doi.org/10.1126/science.1078607

Nawrocki, E. P., & Eddy, S. R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* , 29(22), 2933-2935. https://doi.org/10.1093/bioinformatics/btt509

Oakeshott, J. G., Claudianos, C., Campbell, P. M., Newcomb, R. D., & Russell, R. J. (2005). Biochemical genetics and genomics of insect esterases. In L. I. Gilbert, K. Iatrou, S. S. Gill, (Eds.),*Comprehensive molecular insect science* (pp. 309-381). Oxford: Elsevier. https://doi.org/10.1016/B0-44-451924-6/00073-9

Oakley, A. (2011). Glutathione transferases: a structural perspective.*Drug Metabolism Reviews* , 43 (2), 138-151. https://doi.org/10.3109/03602532.2011.558093

Plasterk, R. H. A. (2006). Micro RNAs in animal development.*Cell* , 124(5), 877-881. https://doi.org/10.1016/j.cell.2006.02.030

Potapov, M., & Yan, G. (2012). *Folsomia* of China I—*fimetaria* group (Collembola: Isotomidae). *Annales de la Société Entomologique de France* , 48 (1-2), 51-56. https://doi.org/10.1080/00379271.2012.10697750

Roncalli, V., Cieslak, M. C., Passamaneck, Y., Christie, A. E., & Lenz, P.H. (2015). Glutathione S-Transferase (GST) gene diversity in the crustacean Calanus finmarchicus – Contributors to Cellular Detoxification. *PLOS ONE* , 10(5), e0123322. https://doi.org/10.1371/journal.pone.0123322

Roelofs, D., Zwaenepoel, A., Sistermans, T. Nap, J., Kampfraath, A. A. Van de Peer, Y., Ellers J., & Kraaijeveld, K. (2020). Multi-faceted analysis provides little evidence for recurrent whole-genome duplications during hexapod evolution. *BMC Biology* , 18, 57. https://doi.org/10.1186/s12915-020-00789-1

Singh, N. D. (2019). *Wolbachia* infection associated with increased recombination in *Drosophila* . *G3 (Bethesda)* , 9(1), 229-237. https://doi.org/10.1534/g3.118.200827

Smit, A. F. A., Hubley, R., & Green, P. 2013-2015. Repeat Masker Open-4.0. Available from http://www.repeatmasker.org

Smith, J. N. (1955). Detoxification mechanisms in insects.*Biological Reviews* , 30, 455-475. https://doi.org/10.1111/j.1469-185X.1955.tb01548.x

Stanke, M., Steinkamp, R., Waack, S., Morgenstern, B. (2004). AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Research* , 32(suppl 2), W309-W312. https://doi.org/10.1093/nar/gkh379

Steinegger, M., & Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology* , 35, 1026-1028. https://doi.org/10.1038/nbt.3988

Subramanian, Bm., Gao, S., Lercherm M, J., Hu, S., & Chen, W. H. (2019). Evolview v3: a webserver for visualization, annotation, and management of phylogenetic trees. *Nucleic Acids Research* , 47(W1), W270-W275. https://doi.org/10.1093/nar/gkz357

Talbert, P. B., & Henikoff, S. (2014). Environmental responses mediated by histone variants. *Trends in Cell Biology* , 24(11), 642-650. https://doi.org/10.1016/j.tcb.2014.07.006

Talla, V., Suh, A., Kalsoom, F., Dincă, V., Vila, R., Friberg, M., Wiklund, C., Backström, N. (2017). Rapid increase in genome size as a consequence of transposable element hyperactivity in wood-white (*Leptidea*) butterflies, *Genome Biology and Evolution* , 9(10), 2491-2505. https://doi.org/10.1093/gbe/evx163

Tully, T., D'Haese, C. A., Richard, M., & Ferriere, R. (2006). Two major evolutionary lineages revealed by molecular phylogeny in the parthenogenetic Collembola species *Folsomia candida* .*Pedobiologia* , 50(2), 95-104. https://doi.org/10.1016/j.pedobi.2005.11.003

Tully, T., & Potapov, M. (2015). Intraspecific phenotypic variation and morphological divergence of strains of *Folsomia candida* (Willem) (Collembola: Isotomidae), the "standard" test springtaill. *PLoS ONE* , 10(9), e0136047. https://doi.org/10.1371/journal. pone.0136047

Vizueta, J., Sánchez-Gracia, A., & Rozas, J. (2020). BITACORA: A comprehensive tool for the identification and annotation of gene families in genome assemblies. *Molecular Ecology Resources* , 20, 1445-1452. https://doi.org/10.1111/1755-0998.13202

Voorrips, R. E. (2002). MapChart: Software for the graphical presentation of linkage maps and QTLs. *The Journal of Heredity* , 93 (1), 77-78. https://doi.org/10.1093/jhered/93.1.77

Wang, Y., Tang, H., DeBarry, J. D., Tan, X., Li, J., Wang, X., Lee, T. H., Jin, H., Marler, B., Guo, H., Kissinger, J. C., Paterson, A. H. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Research* , 40(7), e49.https://doi.org/10.1093/nar/gkr1293

Waterhouse, R. M., Seppey, M., Simão, F. A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E. V., & Zdobnov, E. M. (2018). BUSCO applications from quality assessments to gene prediction and phylogenomics. *Molecular Biology and Evolution* , 35, 543-548. https://doi.org/10.1093/molbev/msx319

Wenzel, J. J., Piehler, A., & Kaminski, W. E. (2007). ABC A-subclass proteins: gatekeepers of cellular phospho- and sphingolipid transport.*Frontiers in Bioscience-Landmark* , 12, 3177-3193. https://doi.org/10.2741/2305

Werren, J. H. (1998). *Wolbachia* and speciation. In D. Howard and S. Berlocher (eds.), *Endless forms: Species and speciation* (pp. 245-260). Oxford University Press.

Wilson, D., Pethica, R., Zhou, Y., Talbot, C., Vogel, C., Madera, M., Chothia, C., & Gough J. (2009). SUPERFAMILY—sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Research* , 37(Suppl_1), D380-D386. https://doi.org/10.1093/nar/gkn762

Xiong, P., Schneider, R. F., Hulsey, C. D., Meyer, A., & Franchini, P. (2019). Conservation and novelty in the microRNA genomic landscape of hyperdiverse cichlid fishes. *Scientific reports* , 9(1), 13848. https://doi.org/10.1038/s41598-019-50124-0

Yang, Z. H. (2007). PAML 4: phylogenetic analysis by maximum likelihood.*Molecular Biology and Evolution* , 24(8), 1586-1591. https://doi.org/10.1093/molbev/msm088

Yu, G., Wang, L. G., Han, Y., & He, Q. Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS, 16(5), 284-287. https://doi.org/10.1089/omi.2011.0118

Zhang, F., Ding, Y., Zhou, Q. S., Wu, J., Luo, A., & Zhu, C. D. (2019). A High-quality draft genome assembly of *Sinella curviseta* : a soil model organism (Collembola). *Genome Biology and Evolution* , 11(2), 521-530. https://doi.org/10.1093/gbe/evz013

Zwaenepoel, A., & Van de Peer, Y. (2019). Inference of ancient whole-genome duplications and the evolution of gene duplication and loss rates. *Molecular Biology and Evolution* , 36, 1384-404. https://doi.org/10.1093/molbev/msz088

Tables

Table 1 Genome sequencing, assembly and annotation statistics of two *Folsomia candida*strains

Content	*Folsomia candida* DK	*Folsomia candida* SH
Sequencing		

Content	*Folsomia candida* DK	*Folsomia candida* SH
Illumina (Gb)	59.47 (270X)	13.74 (90X)
PacBio (Gb)	12.87 (58X)	11.79 (76X)
Hi-C (Gb)	48.85 (222X)	36.72 (238X)
Genome assembly		
Assembly size (Mb)	219.08	153.90
Number of pseudo-chromosomes (sizes)	7 (216.64 Mb)	7 (150.53 Mb)
Number of scaffolds/contigs	75/321	250/850
Longest scaffold/contig (Mb)	44.87/13.22	32.28/8.88
N50 scaffold/contig length (Mb)	38.47/2.51	25.75/0.79
GC content (%)	37.49	38.54
BUSCO completeness (%)	97.3	97.0
Mapping ratio of Illumina reads (%)	97.14	94.71
Mapping ratio of PacBio reads (%)	95.52	97.02
Protein-coding genes		
Number	25,139	21,609
Mean gene length (bp)	3,973.0	3,835.2
Exons/CDS/introns per gene	7.7/7.4/6.4	8.1/7.8/6.8
Mean exon/CDS/introns length	253.9/197.7/312.1	248.6/198.3/263.5
BUSCO completeness (%)	97.0	97.6
Repetitive elements		
Size (Mb)	49.53 (22.61%)	15.43 (10.03%)
DNA transposons (Mb)	9.38 (4.28%)	2.56 (1.67%)
SINEs (kb)	15.59 (0.01%)	30.74 (0.02%)
LINEs (Mb)	3.21 (1.47%)	0.60 (0.39%)
LTRs (Mb)	4.82 (2.20%)	2.19 (1.43%)
Unclassified transposons (Mb)	29.31 (13.38%)	7.40 (4.81%)
Number of ncRNA*	396	334
rRNA	70	39
miRNA	34	28
snRNA	54	42
tRNA	115	104

*The number was estimated by the Infernal and tRNAscan-SE automatic prediction pipeline.

Table 2 Comparison of detoxification-related gene families among the two *Folsomia* strains and *Drosophila melanogaster*(Dmel).

Family	Clan	FCSH	FCDK	Dmel
ABC		**118**	**132**	**56**
	ABCA	2	2	10
	ABCB	7	10	8
	ABCC	15	25	14
	ABCD	3	3	2
	ABCE	1	1	1
	ABCF	3	3	3
	ABCG	69	69	15
	ABCH	17	19	3
CCE		**117**	**120**	**36**
	Dietary/Detoxification	81	82	14

Family	Clan	FCSH	FCDK	Dmel
	Hormone/Semiochemical	20	20	8
	Neuro-developmental	16	18	14
CYP		**183**	**203**	**87**
	Mitochondrial	4	6	12
	CYP2	100	104	7
	CYP3	56	59	36
	CYP4	23	34	32
GST		**70**	**79**	**40**
	Microsomal	5	5	3
	Delta	4	4	11
	Epsilon	0	0	14
	Mu	13	13	0
	Omega	5	4	5
	Sigma	39	48	1
	Theta	2	3	4
	Zeta	2	2	2
UGT		**48**	**48**	**35**

FIGURE LEGENDS

FIGURE 1 Morphological and genomic comparison between FCDK and FCSH. (a) Adult female FCDK. (b) FCDK furca with 1 well-developed apical seta. (c) Four ventral setae on the third thoracic segment of FCDK. (d) Adult female FCSH. (e) Adult male FCSH. (f-g) FCSH furca with 3 or 2 well-developed apical setae. (h) Seven ventral setae on the third thoracic segment of FCSH. (i) Circos graph of the genome characteristics of FCDK and FCSH. Element distributions in 100 kb sliding windows from outer to inner circles: chromosome length, GC content, density of protein-coding genes, DNA transposons, SINE/LINE/LTR retrotransposons, unclassified (other) transposons and simple repeats. Hox gene clusters are masked on chromosome 5. For collinear blocks, reddish links are intraspecific, and links of other colours are interspecific. (j) Comparison of chromosome lengths of both strains. The number of protein-coding genes on chromosomes is shown in brackets. (k) Comparison of major expanded repeat families between the two strains. Only repeat families larger than 100 kb are shown.

FIGURE 2 Chromosomal synteny, gene enrichment and Ks frequency distributions. (a) Syntenic links among chromosomes, arrows indicate nonhomologous regions; (b)–(c) GO (b) and KEGG pathways (c) of protein-coding genes located in nonhomologous regions of FCDK chromosomes 1, 3, 4 and 7. (d) Ks frequency distributions in the FCDK and FCSH genomes.

FIGURE 3 Phylogeny and gene family evolution. (a) Dating tree and orthologue statistics. Node values represent the 95% highest probability densities of divergence times (unit, 100 Mya). Numbers of significantly expanded and contracted gene families are labelled at terminals. '1:1:1' represents shared single-copy orthologues, 'N:N:N' represents multicopy orthologues shared by all species, 'Collembola' represents orthologues unique to Collembola, 'Others' represents unclassified orthologues, and 'Unassigned' represents orthologues that cannot be assigned to any orthogroups. (b)-(c) Significantly expanded gene families and corresponding orthologue numbers in FCDK (b) and FCSH (c).

FIGURE 4 Massive expansion of *Folsomia* ABC (a) and CYP gene families (b). Ultrafast bootstrap values are indicated at the nodes. Dk, FCDK (red name); Sh, FCSH (blue name); Dm, *Drosophila melanogaster* (green name).

FIGURE 5 Gene organizations and alignment of two mitochondrial genomes. (a) FCDK (15,177 bp). (b) FCSH (15,141 bp). CDS, tRNA, rRNA and rep_origin sequences are shown in the figure. The names of the

majority-strand-encoded genes are labelled in the outer circle, the minority-strand-encoded genes are labelled in the inner circle, and the gene order, GC content, and GC skew of both strand tracks are located between labels. (c) Mauve alignment with a horizontal track to show the global similarity of the two mitochondrial genomes. The degree of DNA sequence similarity is indicated by the height of the red-coloured regions. The box-like diagrams show the annotation of coding regions (white boxes), tRNA genes (green), and rRNA genes (red).

FIGURE 6 Comparison of the miRNA distribution between FCDK and FCSH. (a) Distribution of miRNA genes and families. (I) Conserved miRNAs in both strains; (II) FCDK-specific miRNAs; (III) FCSH-specific miRNAs. (b) The chromosomal distribution of miRNAs in FCDH (red bars) and FCSH (blue bars). The serial numbers of chromosomes (chr1-chr7) are shown on the top of each chromosome. The miRNAs are labelled to the right of the bar, corresponding to their gene locations on the chromosomes. Brown lines indicate miRNAs shared by the two strains (black names). Red names represent FCDK-specific miRNAs, and blue names represent FCSH-specific miRNAs.

Supporting Information

Table S1 Chromosome lengths (bp) and number of protein-coding genes for each chromosome of the two *Folsomia candida* strains.

Table S2 Repetitive elements in the genome of the *Folsomia candida* DK strain.

Table S3 Repetitive elements in the genome of the *Folsomia candida* SH strain.

Table S4 Noncoding RNAs in the genome of the *Folsomia candida* DK strain.

Table S5 Noncoding RNAs in the genome of the *Folsomia candida* SH strain.

Table S6 Annotation statistics of protein-coding genes of the two*Folsomia* strains.

Table S7 General statistics of orthology inference for eleven species.

Table S8 Gene family evolution inferred from CAFÉ.

Table S9 Significantly expanded gene families for two *Folsomia*strains.

Figure S1. Hi-C interaction maps of the two *Folsomia candida*strains.

Figure S2. WGD results obtained using BGMM mixture modelling inferred from the wgd pipeline. (a) FCDK; (b) FCSH. The first column plots Ks frequencies, the second transforms frequencies to log-normal distributions, the third column shows the probability of belonging to a particular component of the mixture shown in the second column, and the fourth column plots the mixture with associated weights for each component.

Figure S3. GO and KEGG enrichment for significantly expanded and species-specific gene families of FCDK (a-d) and FCSH (e-h).

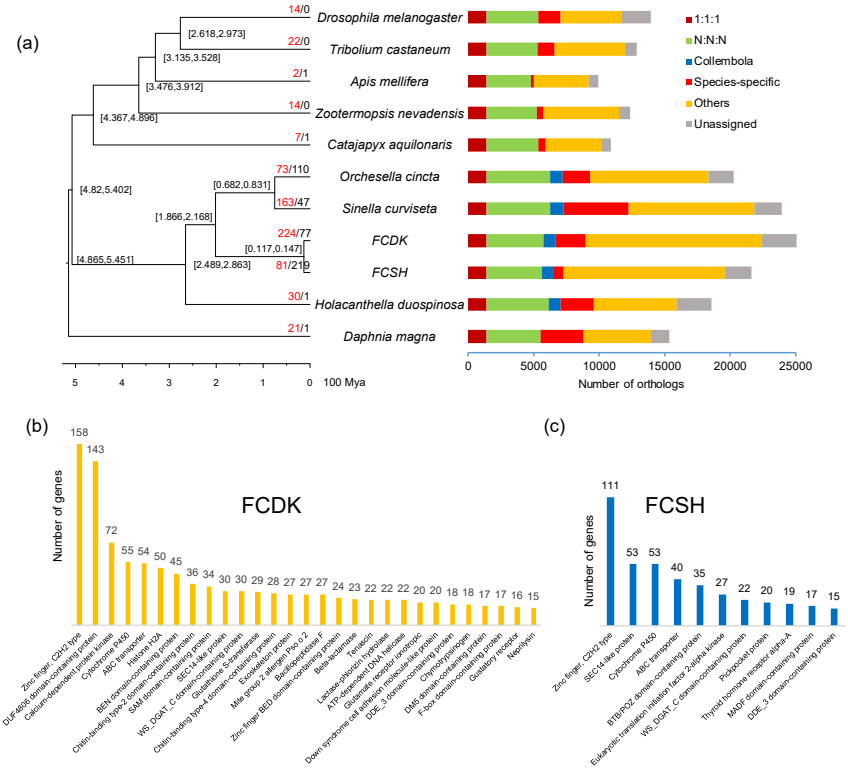
Figure S4. Phylogeny and expansion of CCE, GST and UGT gene families. Ultrafast bootstrap values are shown at the nodes. Dk, FCDK; Sh, FCSH; Dm, *Drosophila melanogaster* . (a) CCE; (b) GST; (c) UGT.

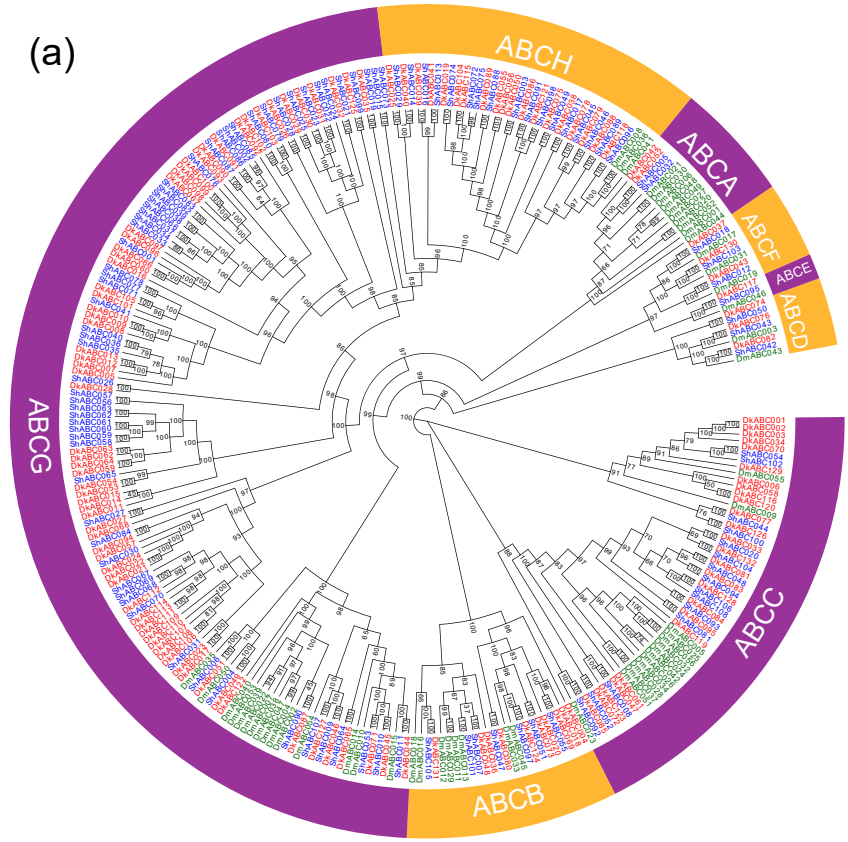
Hosted file

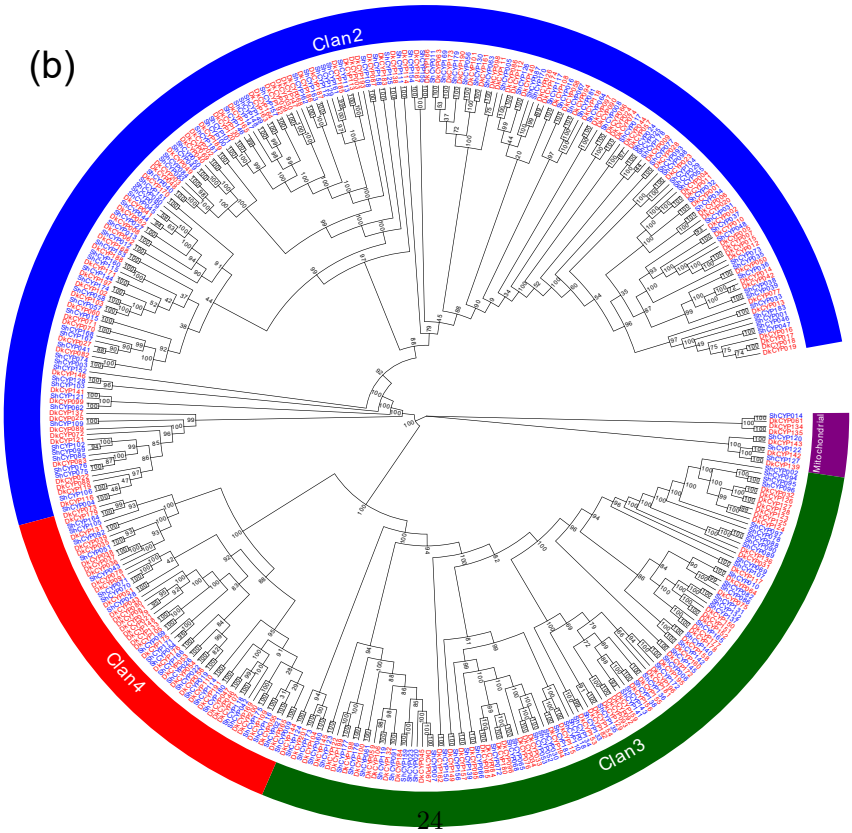
`Figure 1.pdf` available at https://authorea.com/users/452296/articles/550402-high-quality-genomes-reveal-significant-genetic-divergence-and-cryptic-speciation-in-the-model-organism-folsomia-candida-collembola

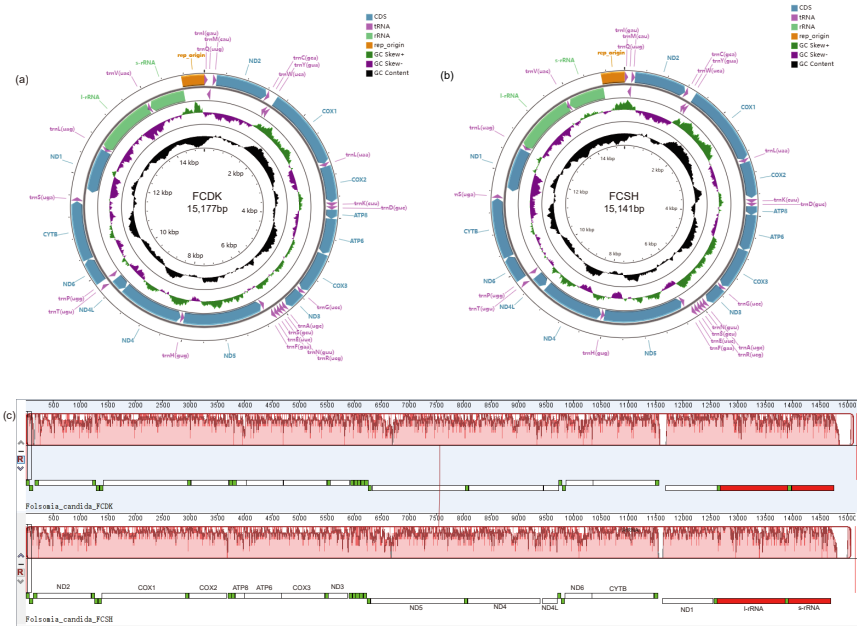
Hosted file

`Figure 2.pdf` available at https://authorea.com/users/452296/articles/550402-high-quality-genomes-reveal-significant-genetic-divergence-and-cryptic-speciation-in-the-model-organism-folsomia-candida-collembola









(a)

I

family	gene	FCDK	FCSH
bantam	bantam	*/o	*/o
let-7	let-7	*/~	*/~
MIR-1	Mir-1	*/o	*/o
MIR-10	Mir-10	*/o	*/o
	Mir-100	*/o	*/o
	Mir-125	*/o	*/~
	Mir-993	*/o	*/o
MIR-1000	Mir-1000-p1	*/o	*/o
MIR-1175	Mir-1175-p1	*/o	*/o
MIR-12	Mir-12	*/o	*/o
MIR-124	Mir-124	*/o	*/o
MIR-133	Mir-133	*/o	*/o
MIR-137	Mir-137	*/o	*/o
MIR-14	Mir-14	*/o	*/o
MIR-153	Mir-153	*/o	*/o
MIR-184	Mir-184	*/o	*/o
MIR-190	Mir-190	*/o	*/o
MIR-193	Mir-193	*/o	*/o
MIR-2	Mir-13-p1	*/o	*/o
	Mir-13-p2	*/o	*/o
	Mir-2944	*/o	*/o
	Mir-2-p1	*/o	*/o
	Mir-2-p2	*/o	*/o
MIR-2001	Mir-2001	*/o	*/o
MIR-210	Mir-210	*/~	*/o
MIR-216	Mir-216-p1	*/~	*/o
	Mir-216-p2	*/o	*/o
MIR-219	Mir-219-p1	*/o	*/o
	Mir-219-p2	*/o	*/o

family	gene	FCDK	FCSH
MIR-22	Mir-22	*/o	*/o
MIR-252	Mir-252a	*/o	*/o
	Mir-252b	*/o	*/o
MIR-275	Mir-275	*/o	*/o
MIR-276	Mir-276	*/o	*/o
MIR-2765	Mir-2765	*/o	*/o
MIR-277	Mir-277-p1	*/o	*/o
	Mir-277-p2	*/~	*/o
MIR-278	Mir-278	*/~	*/o
MIR-279	Mir-279-p1	*/o	*/o
	Mir-279-p2	*/o	*/o
	Mir-279-p3	*/o	*/o
	Mir-279-p4	*/o	*/o
	Mir-279-p5	*/o	*/o
MIR-2796	Mir-2796-p	*/o	*/o
MIR-281	Mir-281	*/o	*/o
MIR-282	Mir-282	*/o	*/o
MIR-29	Mir-29	*/o	*/o
MIR-3	Mir-36-p1	*/o	*/o
	Mir-36-p2	*/o	*/o
MIR-305	Mir-305	*/o	*/o
MIR-306	Mir-306	*/o	*/o
MIR-31	Mir-31	*/o	*/o
MIR-316	Mir-316	*/~	*/o
MIR-317	Mir-317	*/o	*/o
MIR-33	Mir-33	*/o	*/o
MIR-34	Mir-34	*/o	*/o
MIR-375	Mir-375	*/o	*/o
MIR-67	Mir-67-p1	*/o	*/o

family	gene	FCDK	FCSH
MIR-7	Mir-7-p1	*/o	*/o
	Mir-7-p2	*/o	*/o
MIR-71	Mir-71	*/o	*/o
MIR-750	Mir-750-p1	*/o	*/o
	Mir-750-p2	*/o	*/o
MIR-76	Mir-76	*/o	*/o
MIR-8	Mir-8	*/o	*/o
MIR-87	Mir-87-p1	*/o	*/o
	Mir-87-p2	*/o	*/o
MIR-9	Mir-9-p1	*/o	*/o
	Mir-9-p2	*/o	*/~
	Mir-9-p3	*/o	*/o
MIR-92	Mir-92-p1	*/o	*/o
	Mir-92-p2	*/o	*/o
MIR-965	Mir-965	*/o	*/o
MIR-iab-4	Mir-iab-4	*/o	*/o
	Mir-iab-8	*/o	*/o
MIR-pan1	Mir-pan1	*/o	*/o
MIR-fca1	Mir-fca1-p1	*/o	*/o
	Mir-fca1-p2	*/o	*/o
MIR-fca2	Mir-fca2-p1	*/o	*/o
	Mir-fca2-p2	*/o	*/o
MIR-fca3	Mir-fca3	*/o	*/~
MIR-fca4	Mir-fca4-p4	*/~	*/o
MIR-fca5	Mir-fca5-p1	*/o	*/o
MIR-fca6	Mir-fca6-p1	*/~	*/o
	Mir-fca6-p2	*/~	*/o
	Mir-fca6-p3	*/~	*/o

family	gene	FCDK	FCSH
MIR-fca7	Mir-fca7	*/~	*/o
MIR-fca8	Mir-fca8	*/~	*/o
MIR-fca9	Mir-fca9	*/~	*/o
MIR-fca10	Mir-fca10	*/~	*/o

II

family	miRNA gene	FCDK
MIR-1000	Mir-1000-p2	*/~
MIR-2796	Mir-2796-p2	*/~
MIR-67	Mir-67-p2	*/~
MIR-fca1	Mir-fca1-p3	*/o
	Mir-fca1-p4	*/~
	Mir-fca1-p5	*/~
	Mir-fca1-p6	*/~
	Mir-fca1-p7	*/~
	Mir-fca1-p8	*/~
	Mir-fca1-p9	*/~
	Mir-fca1-p10	*/~
MIR-fcaDK-n1	Mir-fcaDK-n1	*/o
MIR-fcaDK-n2	Mir-fcaDK-n2	*/o
MIR-fcaDK-n3	Mir-fcaDK-n3-p1	*/o
	Mir-fcaDK-n3-p2	*/o
	Mir-fcaDK-n3-p3	*/o
	Mir-fcaDK-n3-p4	*/o
MIR-fcaDK-n4	Mir-fcaDK-n4	*/o
MIR-fcaDK-n5	Mir-fcaDK-n5	*/o
MIR-fcaDK-n6	Mir-fcaDK-n6	*/o

III

family	gene	FCSH
MIR-bg5	Mir-bg5	~/o
MIR-827	Mir-827	~/o
MIR-3049	Mir-3049	~/o
MIR-6012	Mir-6012	~/o
MIR-971	Mir-971	~/o
MIR-2	Mir-11	~/o
MIR-29	Mir-995a	*/o
MIR-1175	Mir-1175-p2	*/o
MIR-fca2	Mir-fca2-p2	*/o
MIR-fca4	Mir-fca4-p1	*/~
	Mir-fca4-p2	*/~
	Mir-fca4-p3	*/~
	Mir-fca4-p5	*/~
	Mir-fca4-p6	*/~
	Mir-fca4-p7	*/o
	Mir-fca4-p8	*/~
MIR-fca5	Mir-fca5-p1	*/o
	Mir-fca5-p3	*/~
	Mir-fca5-p4	*/o
MIR-fcaSH-n1	Mir-fcaSH-n1-p1	*/o
	Mir-fcaSH-n1-p2	*/o
	Mir-fcaSH-n1-p3	*/o
	Mir-fcaSH-n1-p4	*/o
	Mir-fcaSH-n1-p5	*/o
	Mir-fcaSH-n1-p6	*/o
	Mir-fcaSH-n1-p7	*/o
MIR-fcaSH-n2	Mir-fcaSH-n2-p1	*/o
	Mir-fcaSH-n2-p2	*/o
	Mir-fcaSH-n2-p3	*/o

family	gene	FCSH
	Mir-fcaSH-n2-p4	*/o
	Mir-fcaSH-n2-p5	*/o
	Mir-fcaSH-n2-p6	*/o
	Mir-fcaSH-n2-p7	*/o
MIR-fcaSH-n3	Mir-fcaSH-n3-p1	*/o
	Mir-fcaSH-n3-p2	*/~
	Mir-fcaSH-n3-p3	*/~
	Mir-fcaSH-n3-p4	*/~
	Mir-fcaSH-n3-p5	*/~
	Mir-fcaSH-n3-p6	*/~
MIR-fcaSH-n4	Mir-fcaSH-n4-p1	*/o
	Mir-fcaSH-n4-p2	*/o
	Mir-fcaSH-n4-p3	*/o
	Mir-fcaSH-n4-p4	*/o
MIR-fcaSH-n5	Mir-fcaSH-n5-p1	*/o
	Mir-fcaSH-n5-p2	*/o
	Mir-fcaSH-n5-p3	*/o
MIR-fcaSH-n6	Mir-fcaSH-n6-p1	*/o
	Mir-fcaSH-n6-p2	*/o
MIR-fcaSH-n7	Mir-fcaSH-n7-p1	*/o
	Mir-fcaSH-n7-p2	*/o
MIR-fcaSH-n8	Mir-fcaSH-n8	*/o
MIR-fcaSH-n9	Mir-fcaSH-n9	*/o
MIR-fcaSH-n10	Mir-fcaSH-n10	*/o
MIR-fcaSH-n11	Mir-fcaSH-n11	*/o
MIR-fcaSH-n12	Mir-fcaSH-n12	*/o
MIR-fcaSH-n13	Mir-fcaSH-n13	*/o
MIR-fcaSH-n14	Mir-fcaSH-n14	*/o
MIR-fcaSH-n15	Mir-fcaSH-n15	*/o

* precursor evidence from genome data; ○ sRNA expression evidence; ~ not found

(b)