Pangenome analysis reveals families of ubiquitin-ligase adaptors as key genomic divergence drivers that lead to hybrid incompatibility

Dongying Xie[1,2†], Pohao Ye[1†], Yiming Ma[1], and Zhongying Zhao[1*]

[1]Department of Biology, Hong Kong Baptist University, Hong Kong SAR, China

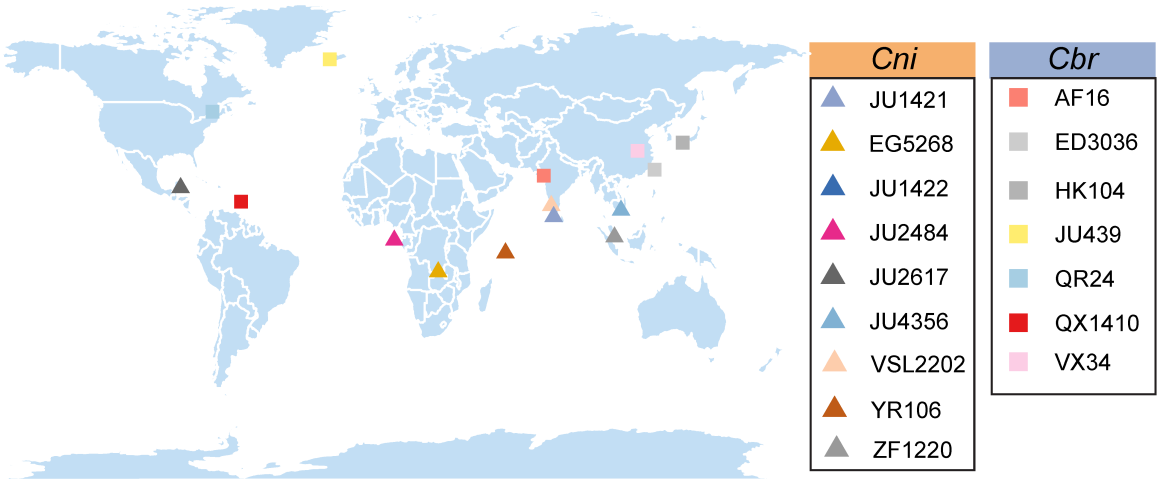
[2]Institute for Research and Continuing Education, Hong Kong Baptist University, Shen Zhen, China

[†] These authors contribute equally to the manuscript.

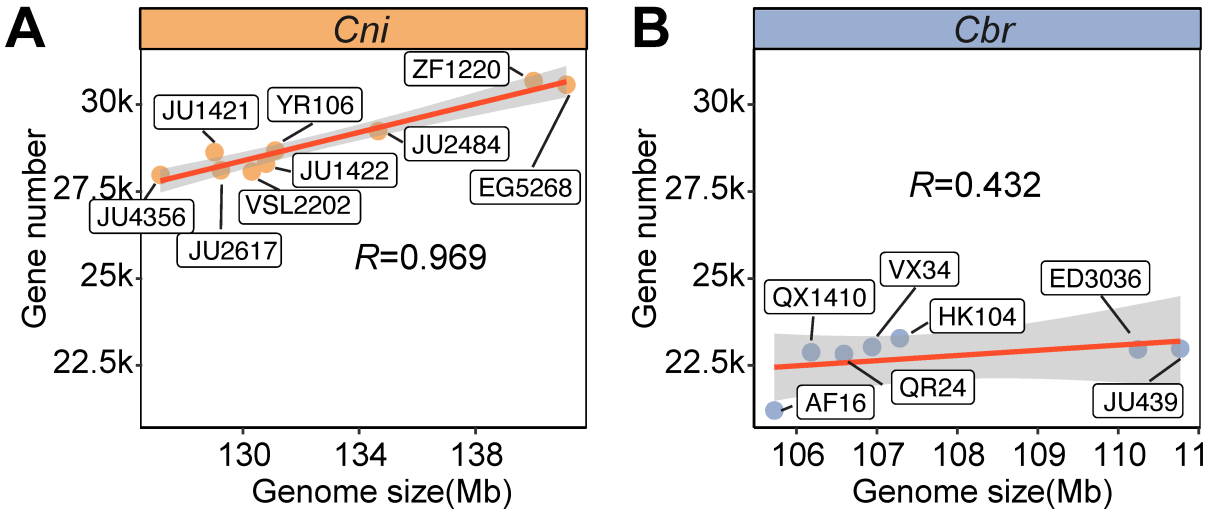
[*] Corresponding author. Email: zyzhao@hkbu.edu.hk

Supplemental Figures (S1-S15)

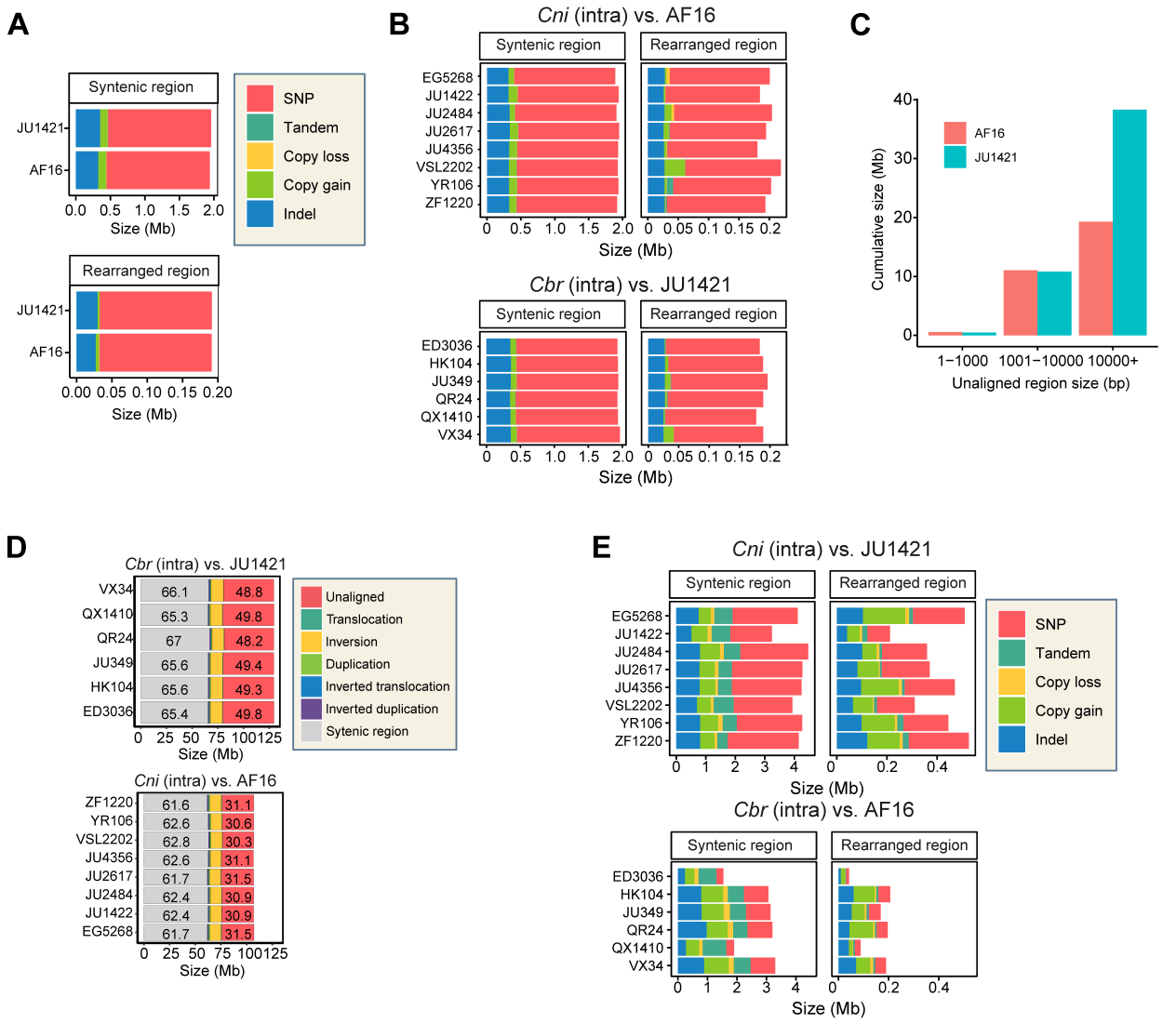
Supplemental Tables (S1-S3)



Supplemental Figure S1 Geographical distribution of nine *C. nigoni* (triangles) and seven *C. briggsae* (rectangles) strains.



Supplemental Figure S2 Positive correlation between genome sizes and gene numbers is stronger in *C. nigoni* strains than in *C. briggsae* strains.
(A-B) Scatter plots compare the relationship between genome sizes and gene numbers across nine *C. nigoni* strains (A) and seven *C. briggsae* strains (B). The Pearson correlation coefficient (*R*) is also shown.



Supplemental Figure S3 Comparison of sequence divergence between populations of *C. briggsae* and *C. nigoni*.

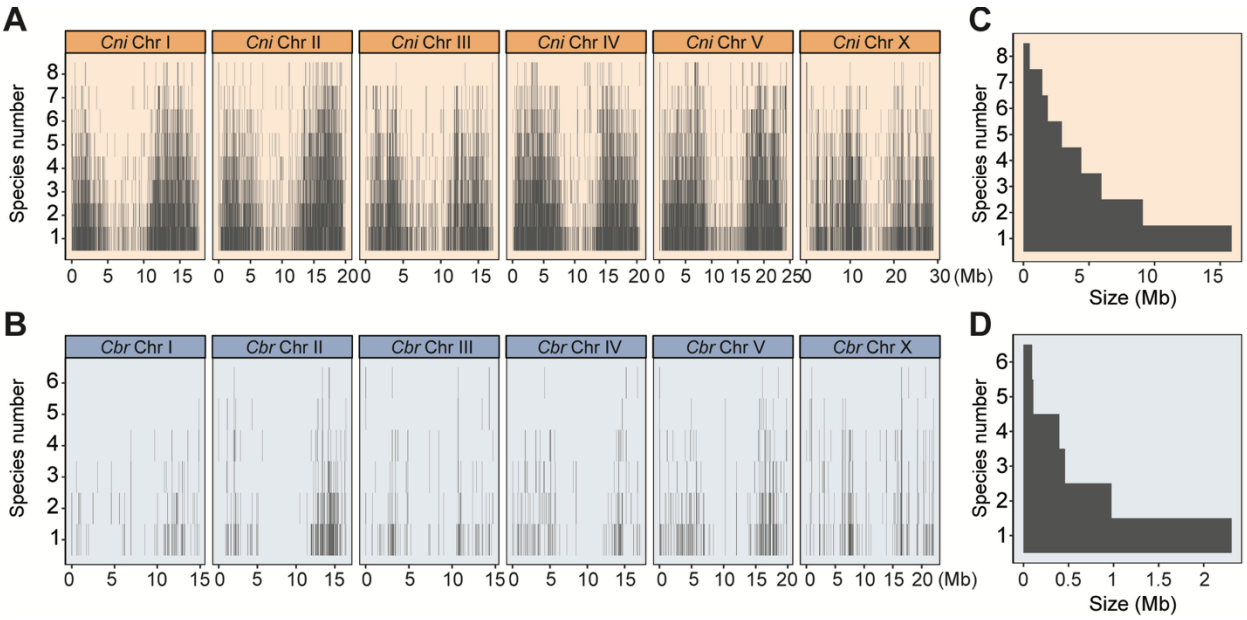
(A) Stacked bar plots illustrating the composition and sizes of sequence variation types within syntenic (top) and rearranged (bottom) regions between the reference strains of *C. nigoni* (JU1421) and *C. briggsae* (AF16). Rearranged regions include all structural variation regions (as shown in Figure 2A), and variation types include SNPs, tandem duplications, copy losses, copy gains, and indels (insertions and deletions).

(B) Stacked bar plots showing the composition and sizes of sequence variation types within syntenic (left) and rearranged (right) regions for wild isolates of *C. nigoni* (top) and *C. briggsae* (bottom) when each is compared to the reference strain of the other species.

(C) Histogram comparing unaligned regions in the reference genomes of *C. briggsae* and *C. nigoni*, stratified into three size categories (1-1,000 bp, 1,001-10,000 bp, and >10,000 bp).

(D) Stacked bar plots showing the composition of intraspecific genome structural variation among wild isolates of *C. briggsae* (top) and *C. nigoni* (bottom) when each is compared to the reference strain of the other species. The sizes of the unaligned and syntenic regions are indicated.

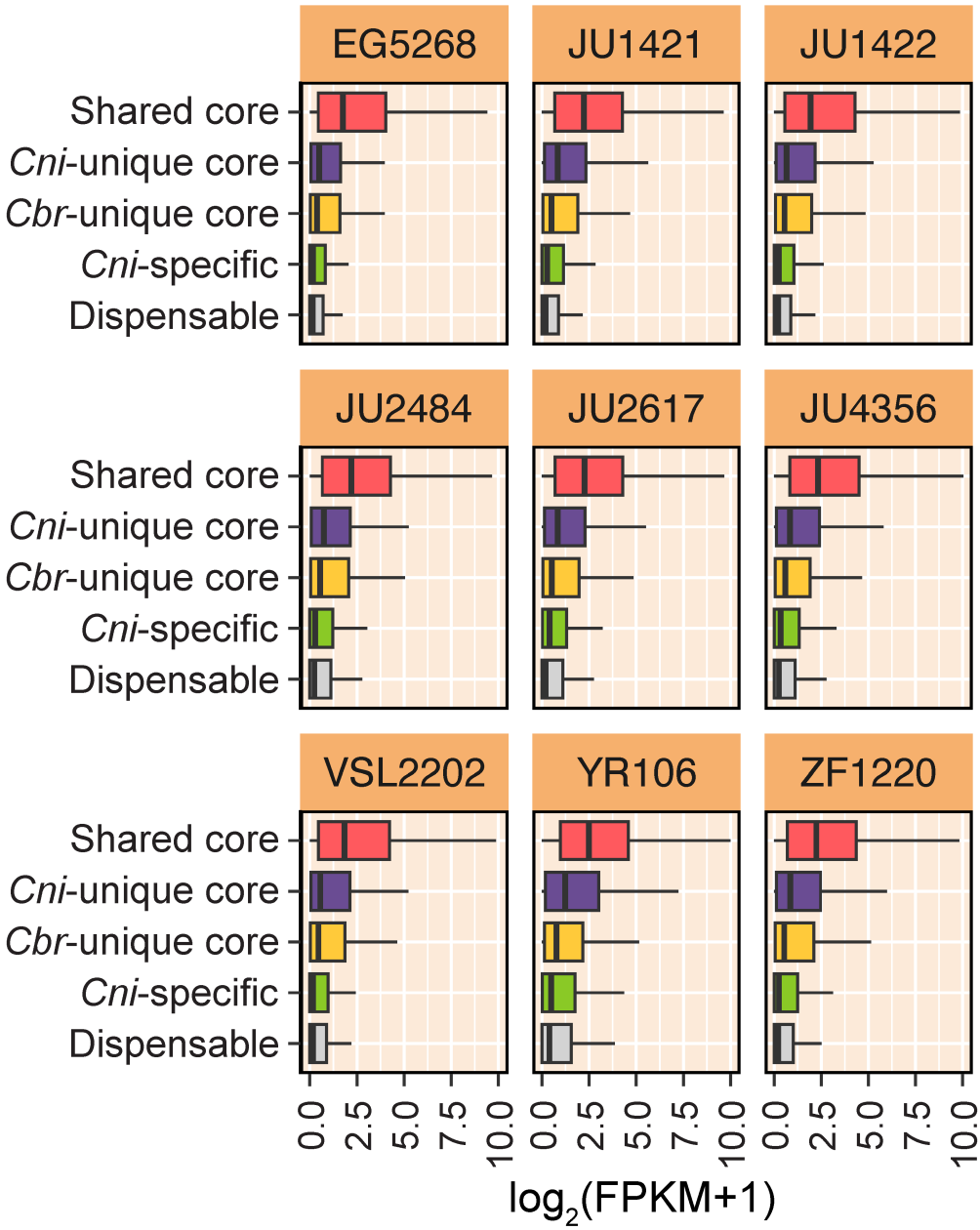
(E) Stacked bar plots illustrating the composition and sizes of sequence variation types within syntenic (left) and rearranged (right) regions for wild isolates of *C. nigoni* (top) and *C. briggsae* (bottom) when each is compared to the reference strain of their respective species.



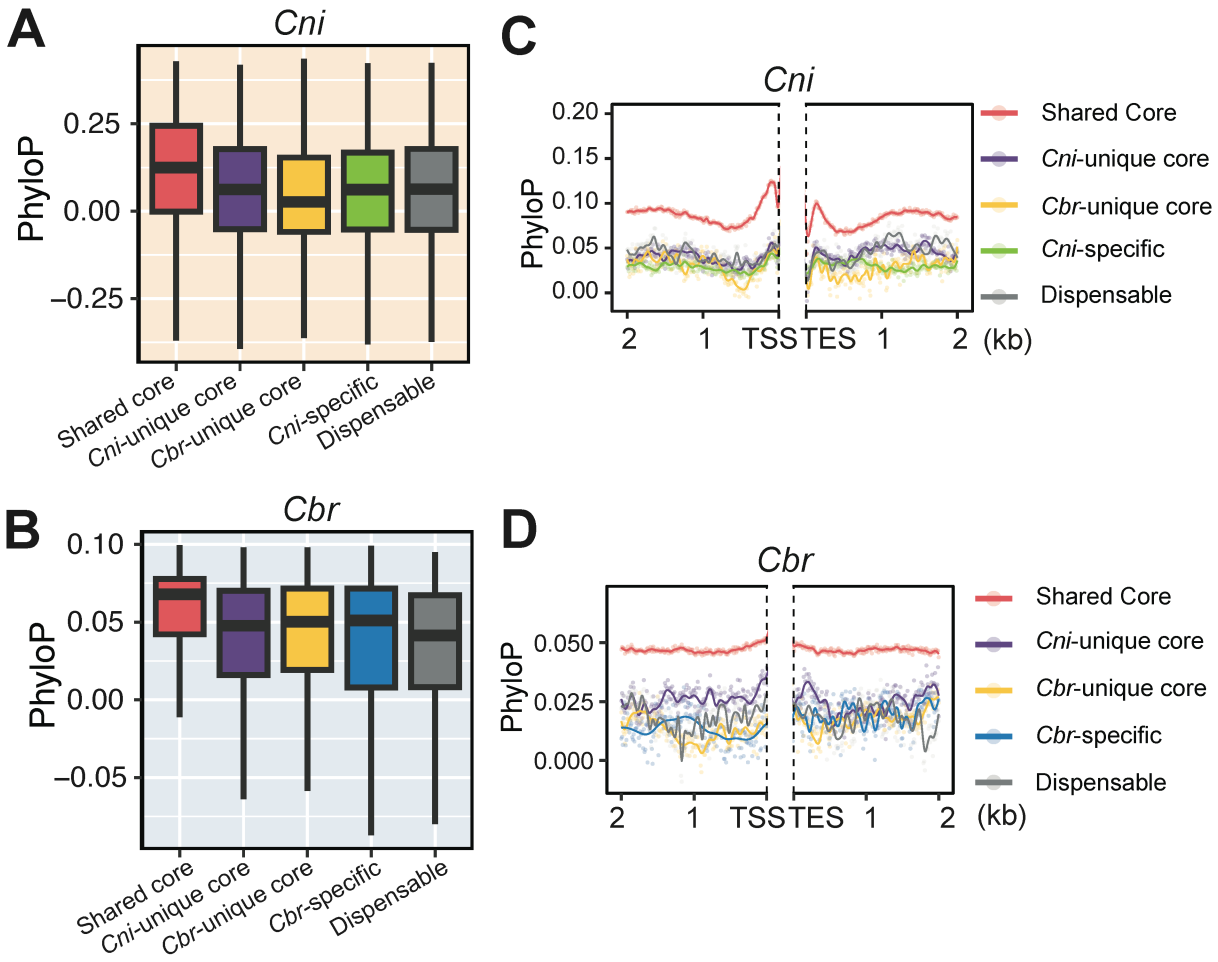
Supplemental Figure S4 Shared intraspecific unalignable regions are rare among wild isolates in both *C. nigoni* and *C. briggsae*.

(A-B) Distribution of unalignable regions across chromosomes in *C. nigoni* (A) and *C. briggsae* (B), stratified by the number of wild isolates in which each region appears. Divergent regions are binned into intervals of 1 kb.

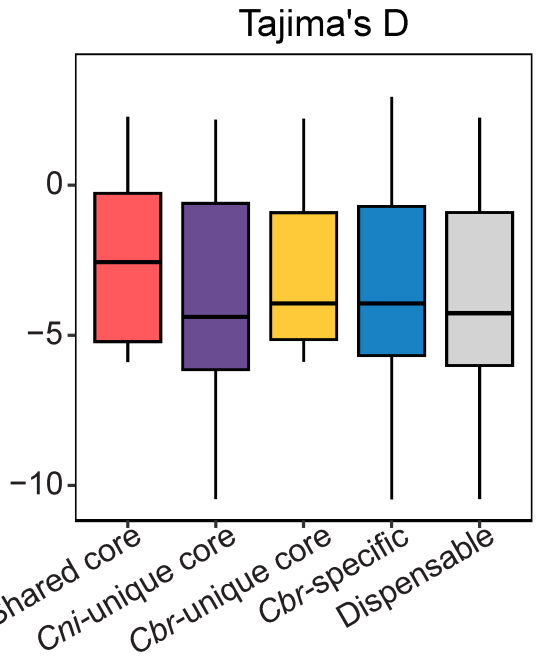
(C-D) Histograms depicting the total sizes of divergent regions shared by varying numbers of wild isolates in *C. nigoni* (C) and *C. briggsae* (D).



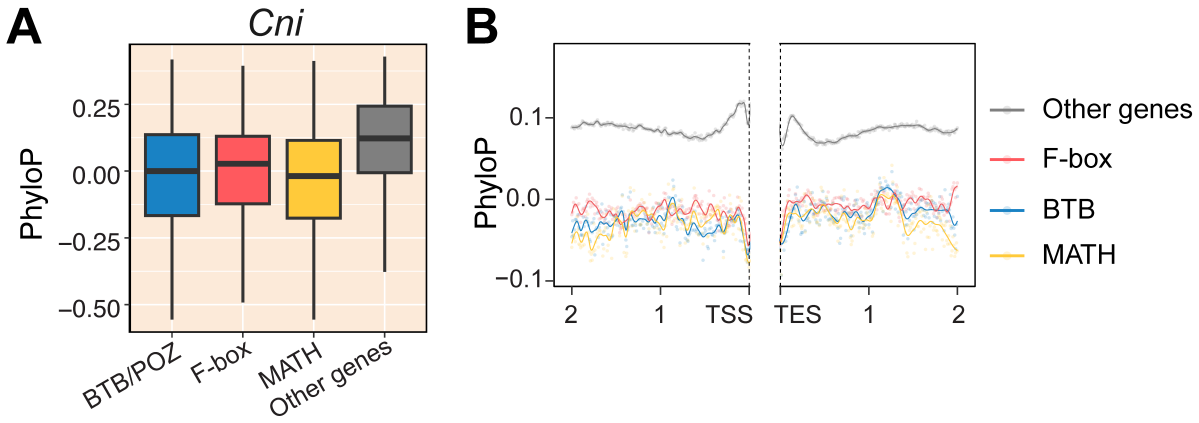
Supplemental Figure S5 *C. nigoni*-specific dispensable genes exhibit lower expression levels compared to shared core genes across all *C. nigoni* strains.
Shown is the comparison of gene expression levels (log_2 FPKM+1) across various gene family groups (as shown in Figure 3B) for *C. nigoni* strains.



Supplemental Figure S6 Species-specific dispensable genes and their flanking genomic sequences exhibit higher nucleotide-level divergence compared to shared core genes.
(A-B) Box plot showing the comparison of average phyloP scores of gene body sequences (From TSS to TES) across various gene family groups (as shown in Figure 3B) for *C. nigoni* (A) and *C. briggsae* (B).
(C-D) Average phyloP scores across 10-bp intervals within 2-kb upstream of transcription start sites (TSS) and 2 kb downstream regions of transcription end sites (TES) for genes in each gene family group (as shown in Figure 3B) for *C. nigoni* reference strain (C) or *C. briggsae* reference strain (D).



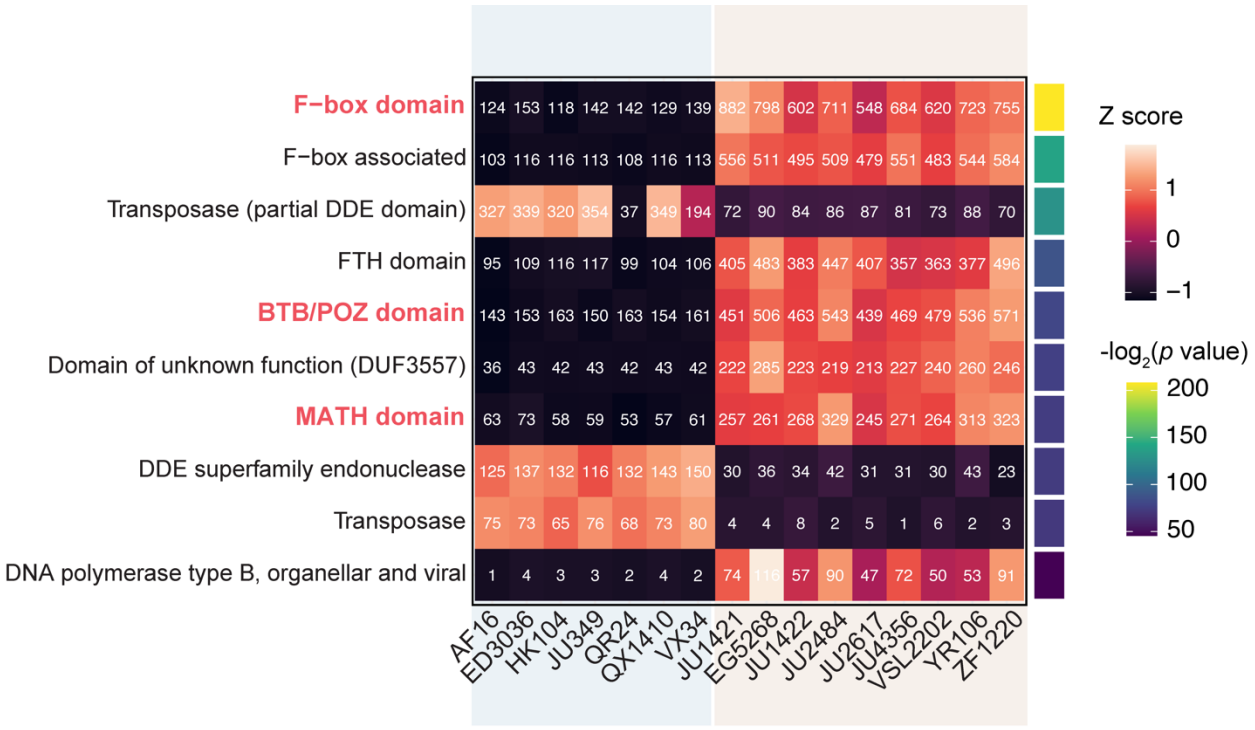
Supplemental Figure S7 Box plot comparing the mean values of Tajima's D for *C. briggsae* genes belonging to each gene family group (as shown in Figure 3B).



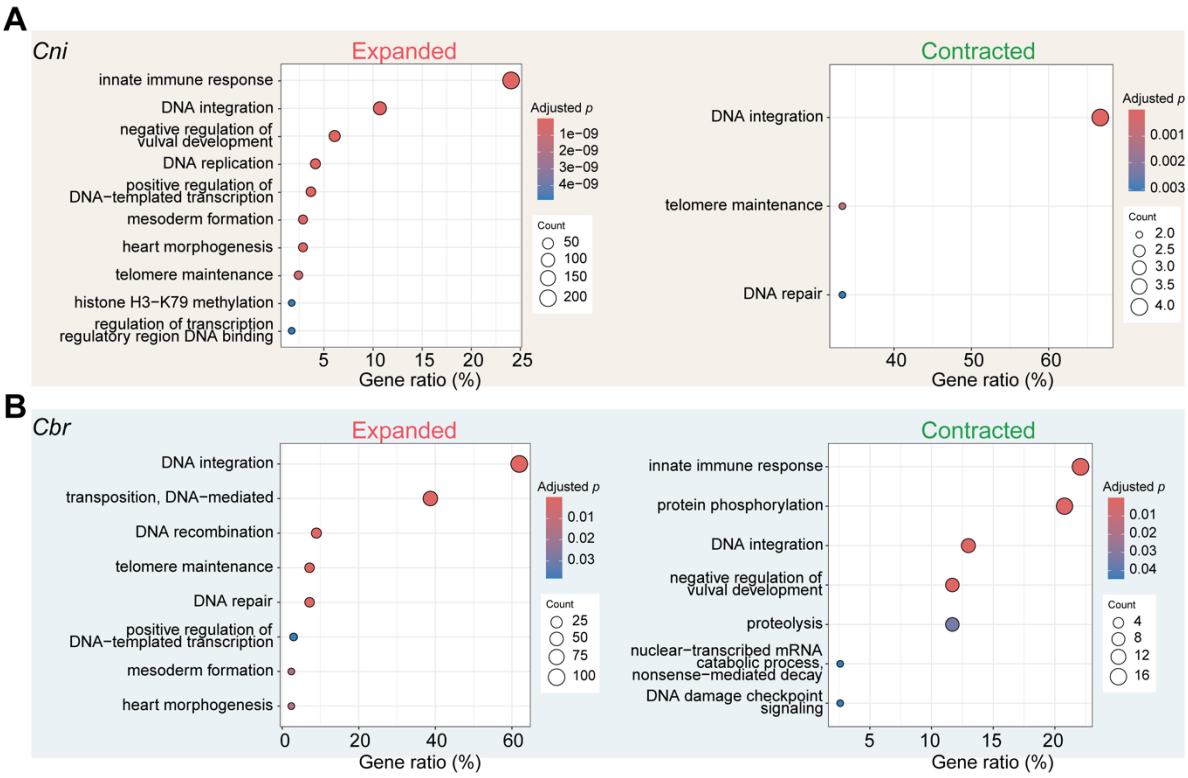
Supplemental Figure S8 E3 ubiquitin-ligase adaptor genes contribute to genomic DNA sequence divergence in *C. nigoni* populations.

(A) Box plot showing the comparison of average phyloP values for *C. nigoni* gene body sequences (from TSS to TES) among genes encoding proteins containing BTB/POZ, F-box, or MATH domains, as well as all other genes in JU1421.

(B) Average phyloP scores across 10-bp intervals within 2-kb upstream of TSS and 2-kb downstream regions of TES for genes encoding proteins with BTB/POZ, F-box, or MATH domains, and all other genes in JU1421.

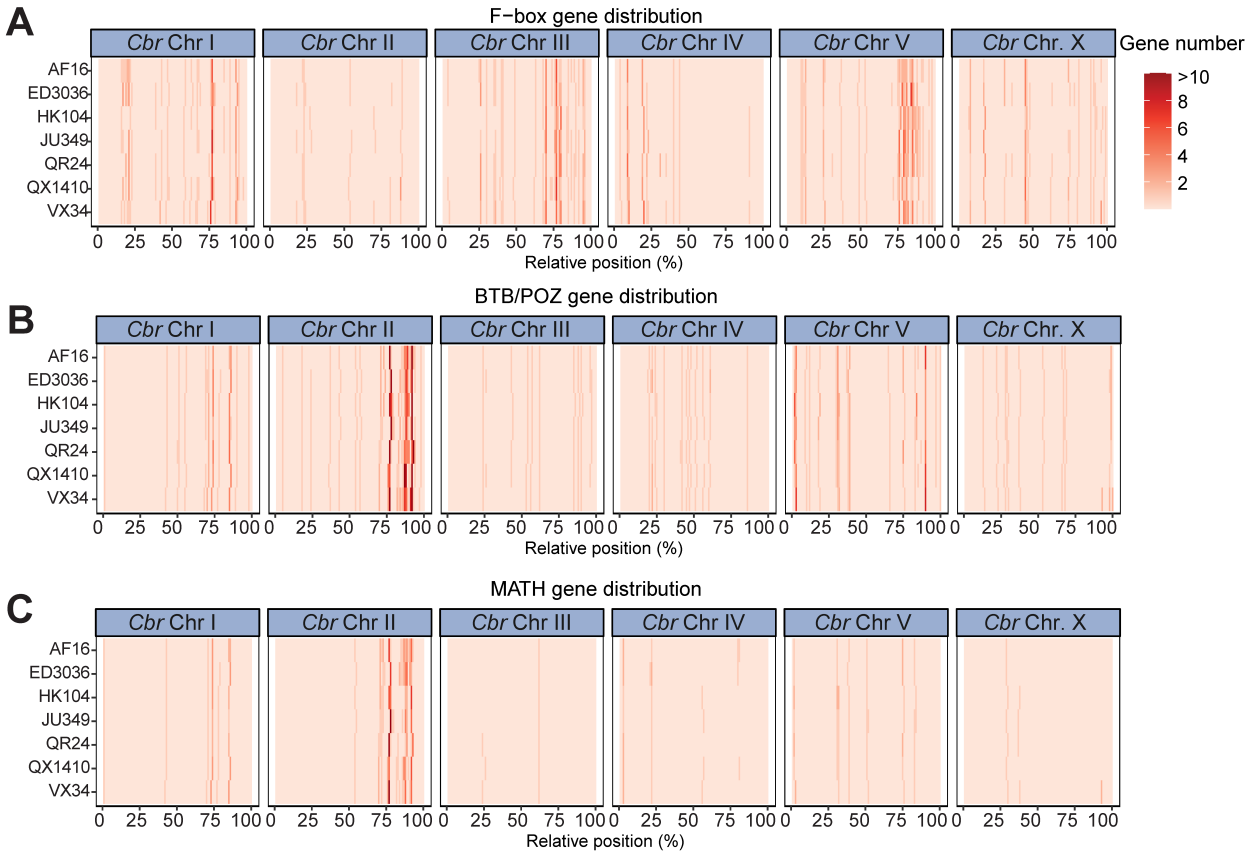


Supplemental Figure S9 E3 ubiquitin-ligase-adaptor genes are significantly enriched in *C. nigoni* strains compared to *C. briggsae* strains.
The heatmap illustrates differences in the number of genes encoding protein domains between *C. nigoni* and *C. briggsae* strains. The total number of genes encoding proteins with each domain from all strains of *C. nigoni* and *C. briggsae* was used for the comparison. Protein domains with significant differences in gene counts between the two species ($P < 0.01$, Fisher's exact test) are included. Only the top ten protein domains with the most significant *P* values are displayed. The heatmap's color gradient represents Z-score normalization values for each row. The total number of genes encoding each protein domain is also indicated for all strains of *C. nigoni* and *C. briggsae*.

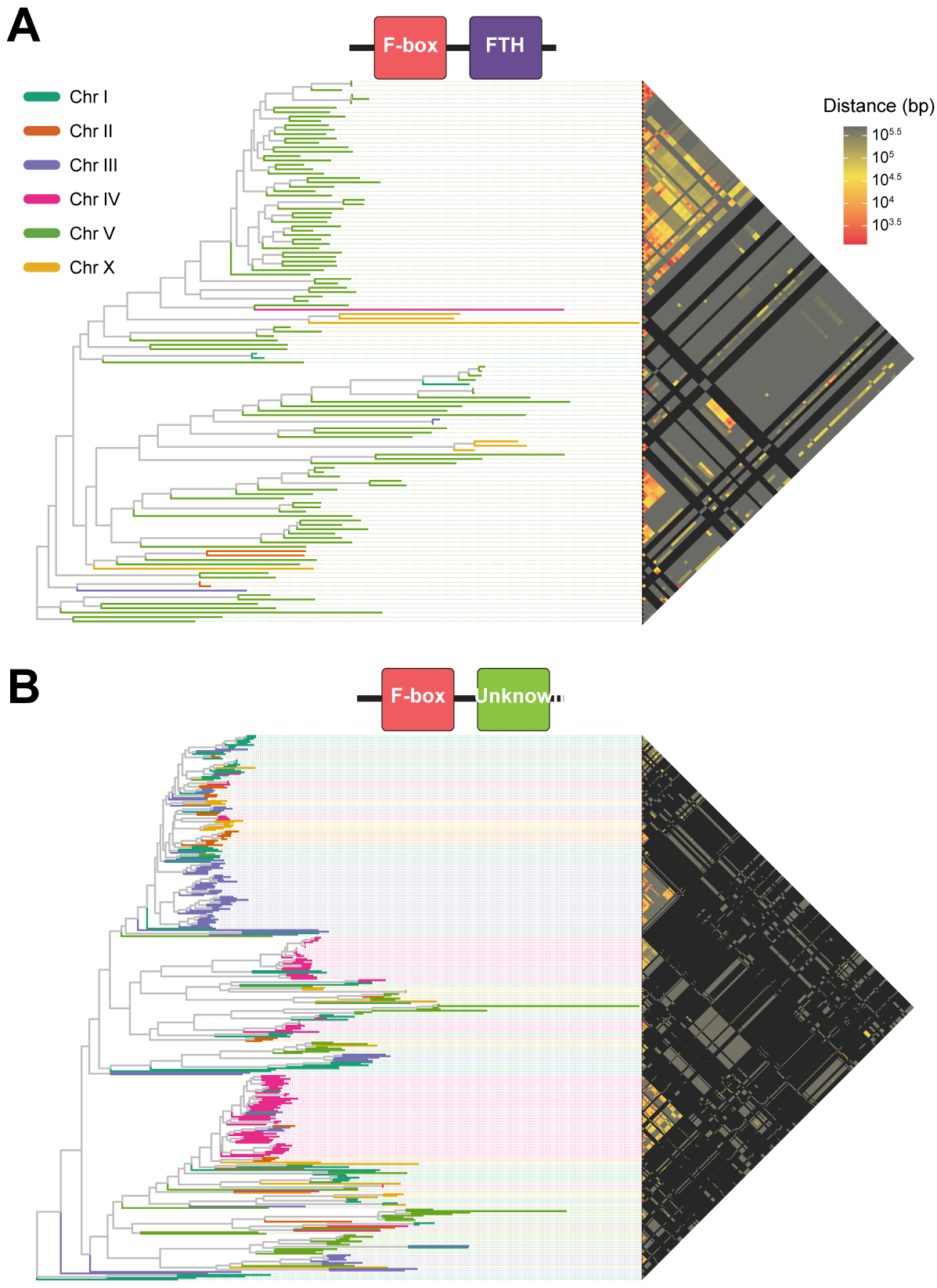


Supplemental Figure S10 Expanded gene families in *C. nigoni* are enriched for the innate immune response.

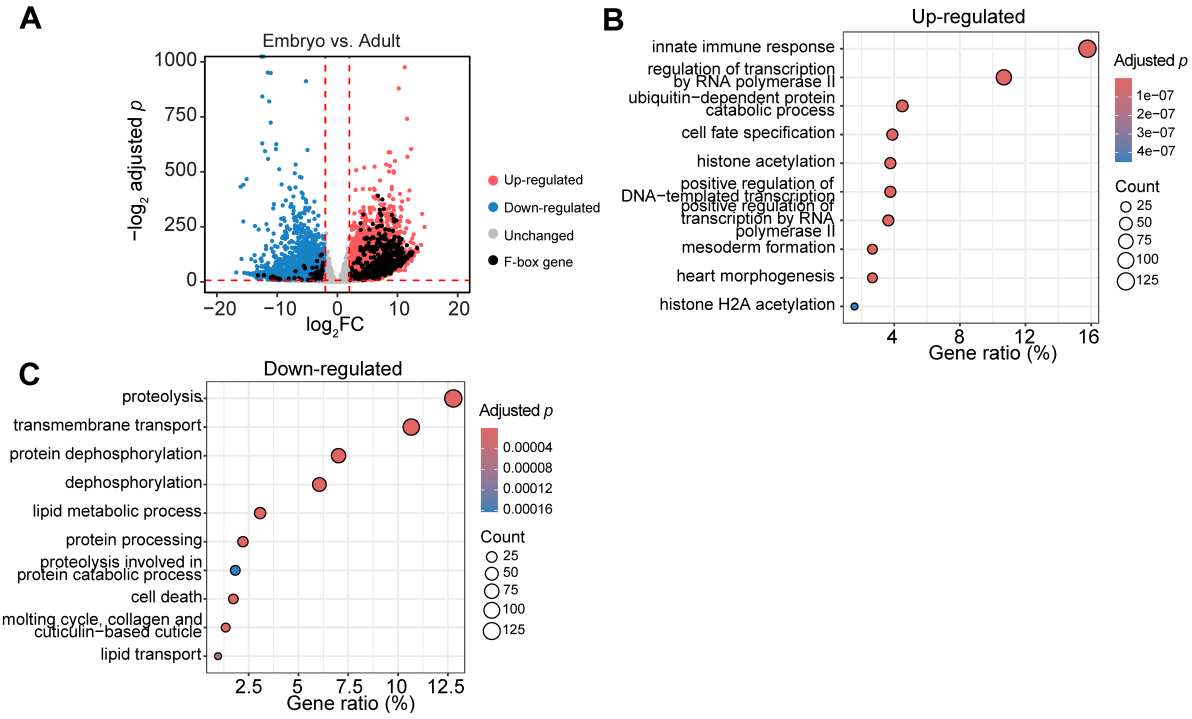
(A-B) Bubble plot showing the Gene Ontology (GO) analysis of biological processes for genes in the expanded (left) and contracted (right) gene families of *C. nigoni* (A) or *C. briggsae* (B). Note that the innate immune response genes are the most significantly enriched in the expanded genes of *C. nigoni*, whereas they are most significantly enriched in the contracted gene families of *C. briggsae*.



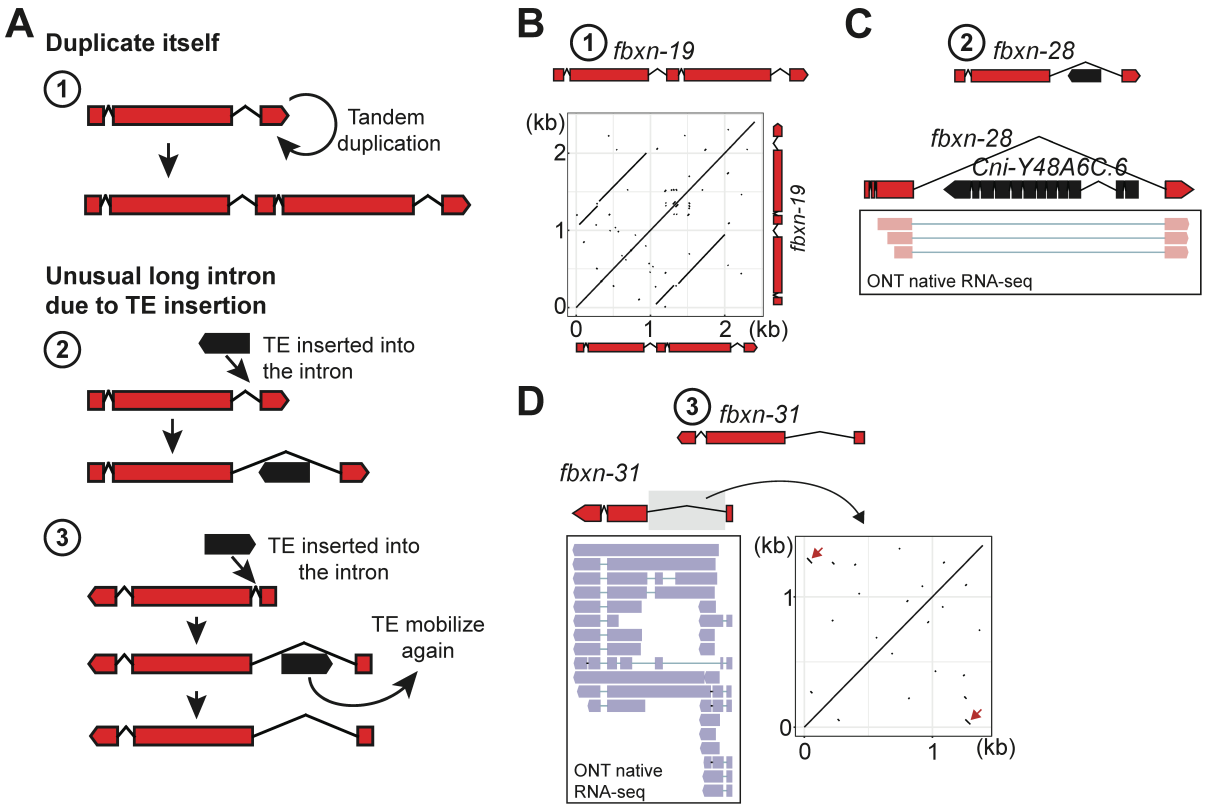
Supplemental Figure S11 Chromosomal distribution of genes encoding proteins with F-box domains (A), BTB/POZ domains (B), and MATH domains (C) across all the *C. briggsae* strains.
Relative chromosomal positions are used to account for differences in chromosome lengths across the strains.


A
F-box
FTH
Chr I
Chr II
Chr III
Chr IV
Chr V
Chr X
Distance (bp)
$10^{5.5}$
10^{5}
$10^{4.5}$
10^{4}
$10^{3.5}$
B
F-box
Unknow

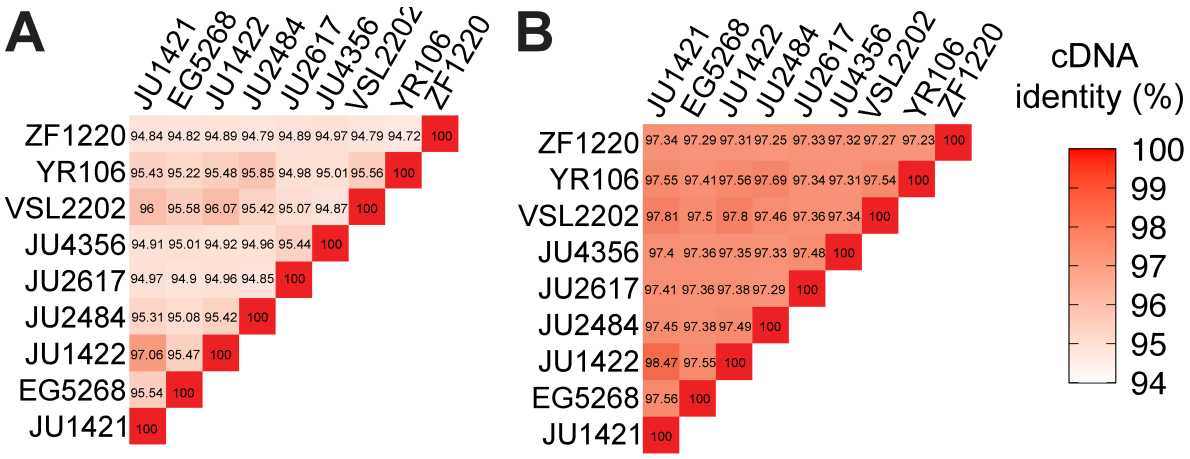
Supplemental Figure S12 The physical and phylogenetic distances of the F-box genes are largely correlated.
(A-B) Left: Protein phylogenetic trees of two types of F-box genes i.e., those containing an FTH domain (A) or an unknown domain (B) (as shown in Figure 5E) in the *C. nigoni* reference strain (JU1421) using the maximum likelihood method. Chromosomal location of the genes is differentially color-coded. Right: Pairwise physical distances between each type of the F-box genes.



Supplemental Figure S13 F-box gene expression is predominantly up-regulated during the embryonic stage compared to the adult stage.
(A) Volcano plot illustrating differentially expressed genes between embryos and adults of *C. nigoni (JU1421)*. Genes exhibiting a two-fold change with an adjusted *P*-value < 0.01 were considered significantly dysregulated. Note that 536 and 74 (out of 882) F-box genes are up-regulated and down-regulated during the embryonic stage, respectively.
(B-C) Gene Ontology (GO) analysis of biological processes for genes up-regulated (B) or down-regulated (C) in embryos compared to adults, as shown in panel A. Note that innate immune response genes are significantly enriched among the up-regulated genes.



Supplemental Figure S14 The *fbxn* genes frequently undergo rearrangement during their expansion.
(A) Schematics showing the rearranged *fbxn* gene types in *C. nigoni* reference strain JU1421. Top: An *fbxn* gene can undergo tandem duplication of a partial sequence spanning both exons and introns. Middle: A transposon can insert into the intron of an *fbxn* gene, leading to an exceptionally long intron. Bottom: A transposon can initially insert into an intron and subsequently mobilize, leaving an exceptionally long intron.
(B) *fbxn-19* is the *fbxn* gene that has undergone tandem duplication. Dot plot showing the alignment of *fbxn-19* genomic sequences.
(C) *fbxn-28* is the *fbxn* gene with a transposon (orthologous to *Y48A6C.6* in *C. elegans*) inserted into its second exon. Shown are the gene models and read tracks from Oxford nanopore native RNA-seq for *fbxn-28*. Note that no RNA-seq reads support the transposon, indicating that it is transcriptionally silenced.
(D) *fbxn-31* is the *fbxn* gene possibly with a transposon that was inserted into its first exon and subsequently mobilized. Shown are the gene models and read tracks of Oxford nanopore native RNA-seq for *fbxn-31*. Also shown is the dot plot of the first intron. Note that a pair of inverted terminal repeats is present, supporting the remnants of a transposon insertion.



Supplemental Figure S15 Single-copy orthologous genes located within the unaligned regions exhibit lower conservation compared to those outside the unaligned regions.
(A-B) Heatmap showing the average cDNA sequence identity of single-copy orthologous genes located either within the *C. nigoni* intraspecific unaligned regions (as shown in Fig. 2B) (A) or outside these regions (B) across *C. nigoni* strains. Note that the genes within the unaligned regions show markedly lower sequence identity compared to those outside.

Supplemental Table S1 Statistics of the assembled genomes for *C. nigoni* and *C. briggsae* strains.

Strain name	Genome size (bp)	Scaffold No.	Scaffold N50 (bp)	Gene No.	Max gene length (bp)	Average gene length (bp)	Average CDS length (bp)	Average intron length (bp)	BUSCO	Unassigned contigs No.	Heterozygosity (%)
***C. nigoni* strains**											
JU1421	129,020,397	45	20,457,831	28,630	71,268	2,838	221	289	99.1	39	0.18
EG5268	141,128,490	607	21,012,276	30,569	60,114	2,832	220	294	99.2	601	0.514
JU1422	130,793,008	359	20,228,028	28,300	66,548	2,831	219	291	99.1	353	0.233
JU2484	134,652,797	568	20,398,616	29,236	62,116	2,788	219	290	99.1	562	0.23
JU2617	129,239,438	467	19,863,965	28,110	52,689	2,819.	221	289	98.8	461	0.256
JU4356	127,162,494	571	19,174,480	27,976	58,783	2,782	219	282	98.3	565	0.266
VSL2202	130,304,040	449	20,048,174	28,076	73,798	2,819	220	293	99.2	443	0.117
YR106	131,113,396	536	20,400,464	28,676	71,988	2,790	218	286	99.0	530	0.152
ZF1220	140,001,514	693	20,001,117	30,675	56,438	2,776	218	290	99.0	687	0.492
***C. briggsae* strains**											
AF16	105,723,858	25	17,218,686	21,206	88,188	2,956	217	306	98.7	19	0.09
ED3036	105,663,136	62	17,201,145	22,961	50,818	2,869	215	288	96.5	56	NA
HK104	107,196,466	193	17,174,180	23,282	74,188	2,856	214	291	96.5	187	NA
JU349	106,946,627	187	17,141,370	22,985	49,917	2,882	216	290	96.9	181	NA
QR24	106,543,070	176	17,157,184	22,830	79,268	2,899	213	290	96.8	170	NA
QX1410	106,181,468	6	17,080,301	22,879	69,838	2,900	216	289	96.9	0	NA
VX34	106,938,796	6	17,304,409	23,031	106,358	2,880	216	289	96.9	0	NA

Supplemental Table S2 Statistics of the repeat contents for *C. nigoni* and *C. briggsae* wild isolates.

Strain name	Repeat (%)	Repeat total size (bp)	Retrotransposon (%)	Retrotransposon total size (bp)	DNA transposon (%)	DNA transposon total size (bp)	Small RNA (%)	Small RNA total size	Rolling circle transposon (%)	Rolling circle transposon total size (bp)	Simple repeat (%)	Simple repeat total size (bp)	Unclassified (%)	Unclassified total size (bp)
***C. nigoni* strains**														
JU1421	20.81	26,852,664	0.06	73,261	5.83	7,517,062	0.05	65,008	2.34	3,025,012	3.47	4,100,894	9.07	11,696,462
EG5268	20.06	28,308,050	0.60	848,457	5.37	7,574,132	0.04	58,129	1.84	2,602,637	3.48	4,505,998	8.73	12,313,801
JU1422	21.06	27,549,933	0.07	87,202	6.10	7,976,358	0.04	53,681	1.23	1,609,302	4.49	5,323,401	9.14	11,9528,67
JU2484	20.91	28,157,937	0.16	217,691	6.08	8,181,634	0.02	24,367	1.04	1,397,388	4.34	5,397,512	9.28	12,496,390
JU2617	20.97	27,104,607	0.04	49,228	6.01	7,769,020	0.06	71,129	2.13	2,756,402	3.72	4,333,768	9.01	11,647,878
JU4356	19.86	25,254,009	0.02	25,481	6.34	8,062,789	0.05	65,800	1.99	2,535,344	3.25	3,780,528	8.21	10,433,730
VSL2202	20.32	26,477,119	0.04	49,935	5.60	7,293,599	0.06	77,747	1.75	2,286,723	3.67	4,374,932	9.20	11,992,655
YR106	20.42	26,771,024	0.08	104,003	5.92	7,767,174	0.03	43,218	1.62	2,118,218	3.97	4,830,261	8.79	11,529,673
ZF1220	21.06	29,490,410	0.05	63,678	6.11	8,547,251	0.06	80,421	1.62	2,266,576	3.64	4,477,627	9.60	13,438,104
***C. briggsae* strains**														
AF16	20.55	21,7305,84	1.00	1,052,341	5.54	5,860,008	0.03	34,911	1.55	1,640,223	2.81	2,624,925	9.62	10,166,249
ED3036	20.13	22,194,884	0.15	165,501	5.57	6,137,039	0.04	44,578	1.02	1,125,532	2.64	2,588,486	10.71	11,806,290
HK104	21.27	22,819,709	0.26	283,554	5.65	6,057,366	0.03	30,927	2.41	2,588,223	3.56	3,253,221	9.36	10,040,662
JU349	20.01	22,161,117	0.15	164,197	6.37	7,054,038	0.04	39,533	1.39	1,538,227	3.42	3,446,423	8.64	9,571,860
QR24	20.53	21,885,646	0.14	149,378	5.87	6,254,336	0.03	35,243	1.42	1,511,425	3.60	3,457,612	9.47	10,090,526
QX1410	20.22	21,473,328	0.09	91,539	5.79	6,144,346	0.03	30,622	1.51	1,605,064	3.42	3,295,468	9.38	9,957,770
VX34	20.49	21,914,140	0.09	92,422	5.97	6,387,333	0.10	111,456	2.37	2,539,715	3.60	3,491,311	8.35	8,929,235

Supplemental Table S3 the size distribution of indels between the *C. briggsae* (AF16) and *C. nigoni* (JU1421) reference strains.

Indel size (bp)	Indel number
1-10	274,601
10-100	2,662
100-300	132
300-1000	104
>1000	48