Genetic exchange with an outcrossing sister species causes severe genome-wide dysregulation in a selfing *Caenorhabditis* nematode

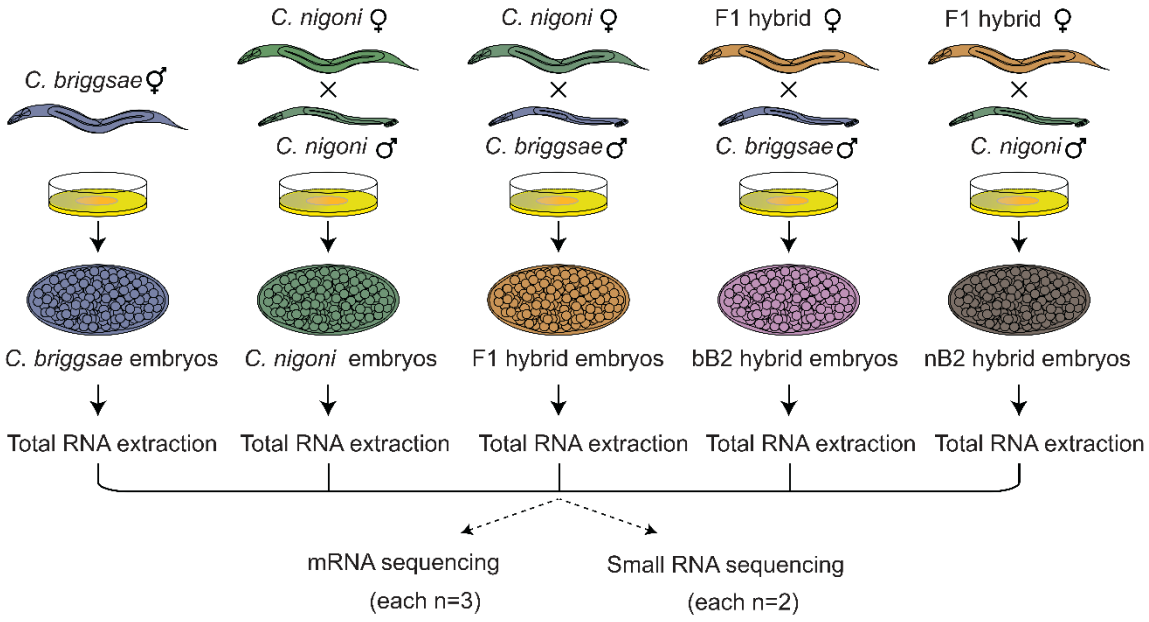
Dongying Xie[1], Pohao YE[1], Yiming Ma[1], Yongbin Li[2], Xiao Liu[2], Peter Sarkies[3], and Zhongying Zhao[1,4*]

[1]Department of Biology, Hong Kong Baptist University, Hong Kong, China; [2]College of Life Sciences, Capital Normal University, Beijing, China; [3] Department of Biochemistry, University of Oxford, United Kingdom; [4]State Key Laboratory of Environmental and Biological Analysis**,** Hong Kong Baptist University**,** Hong Kong, China

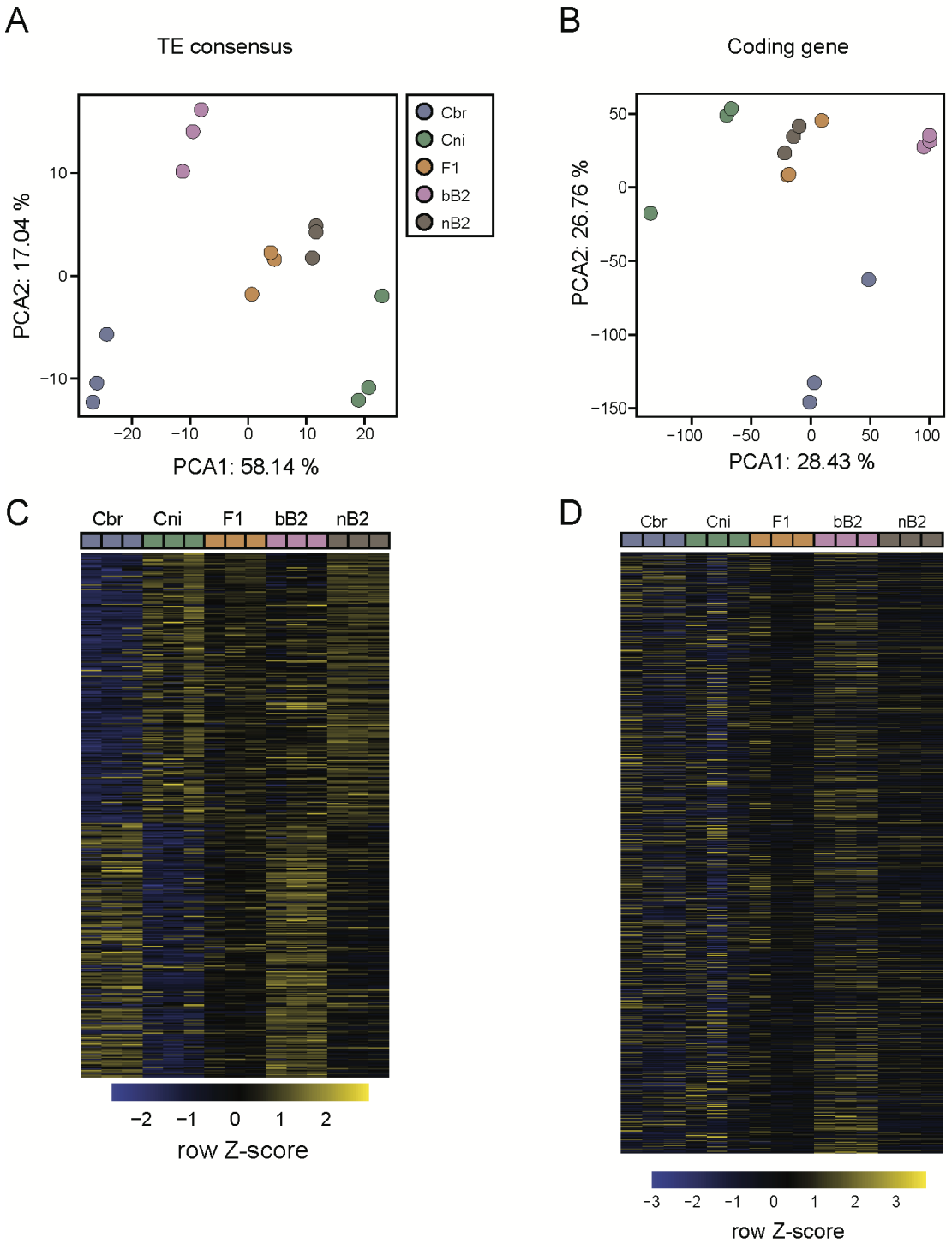
* To whom correspondence should be addressed.

Email: zyzhao@hkbu.edu.hk

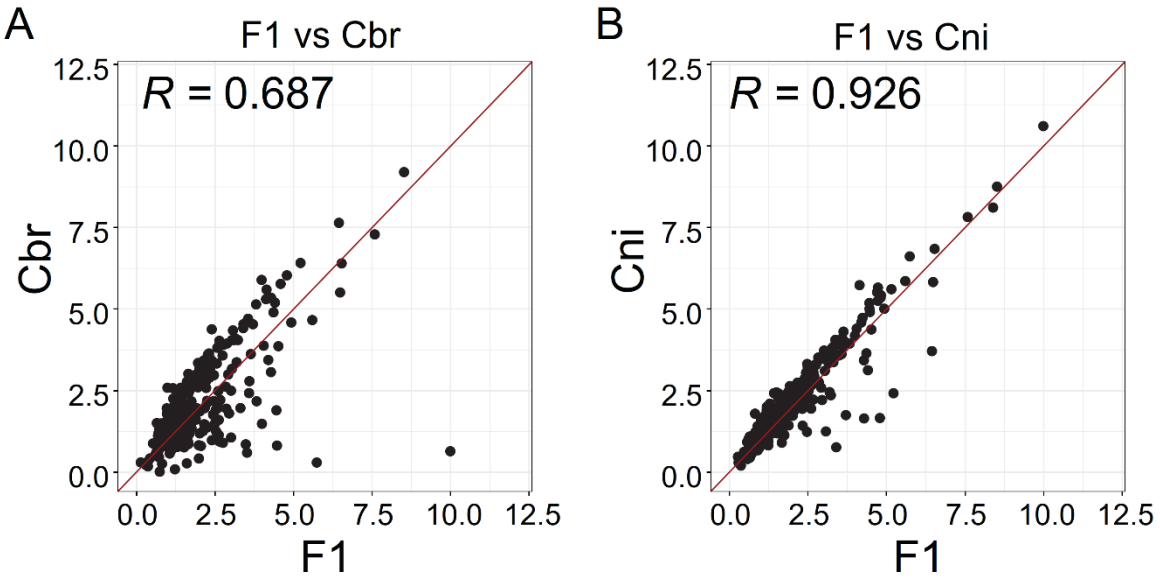
TABLE OF CONTENTS



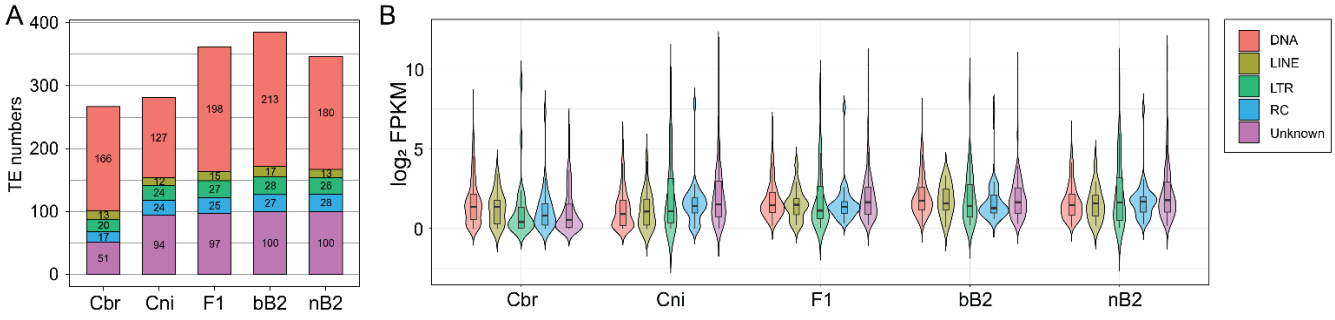
Supplemental Fig. S1 Experimental design. Schematic representation showing crossing strategy for RNA extractions for both mRNA and small RNA sequencing of *C. briggsae* (Cbr), *C. nigoni* (Cni), their hybrid F1 embryos (F1) and backcrossing embryos, i.e., bB2 and nB2, the crossing progenies between the F1 hybrids and *C. briggsae* or *C. nigoni*, respectively. Three and two replicate experiments were conducted for mRNA and small sequencing respectively. The coloring scheme is used throughout the manuscript.



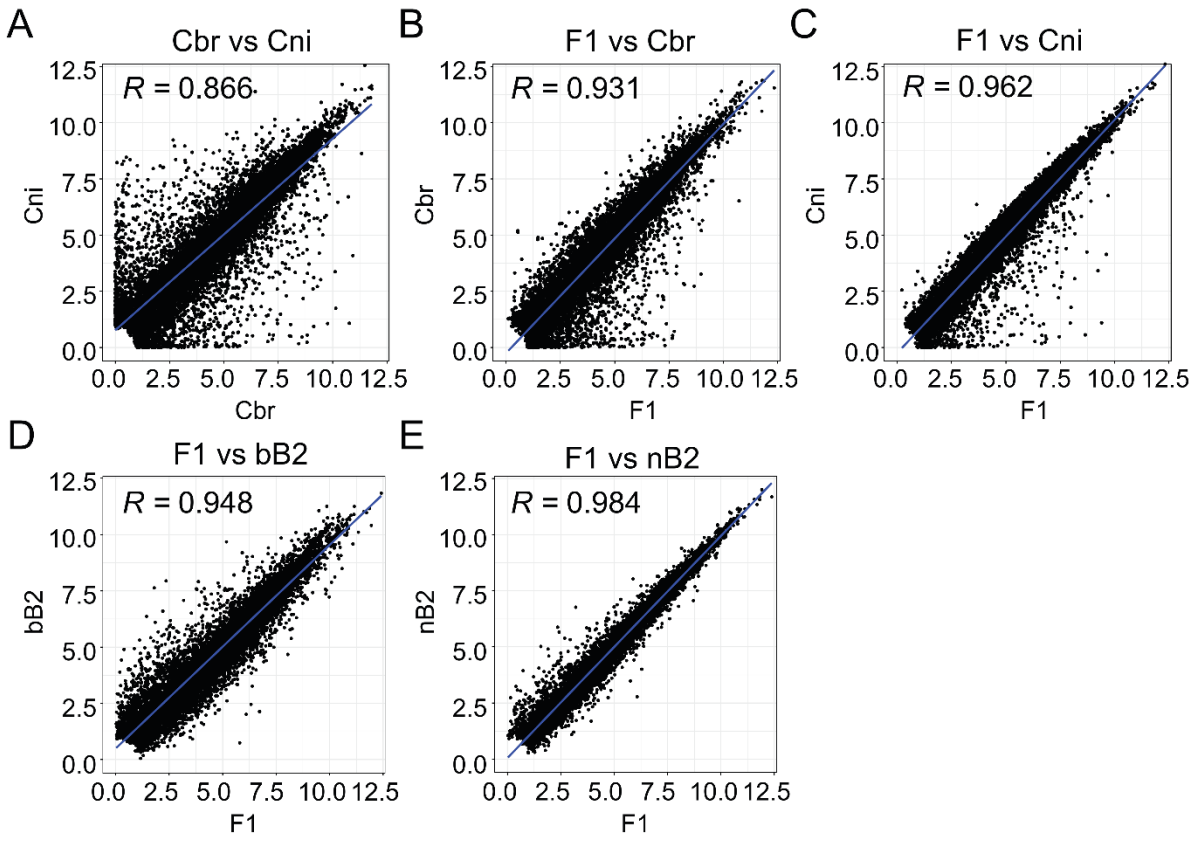
Supplemental Fig. S2 Reproducibility of sequencing reads among replicates. (A-B) Principal Component Analysis (PCA) plots of mRNA sequencing reads from three replicates mapped to the consensus TE library (A) or coding genes (B). Note that reads from three replicates of the same sample tend to cluster together. (C-D) Heatmaps showing the abundance of sequencing reads mapped to TEs (C) or coding genes (D).



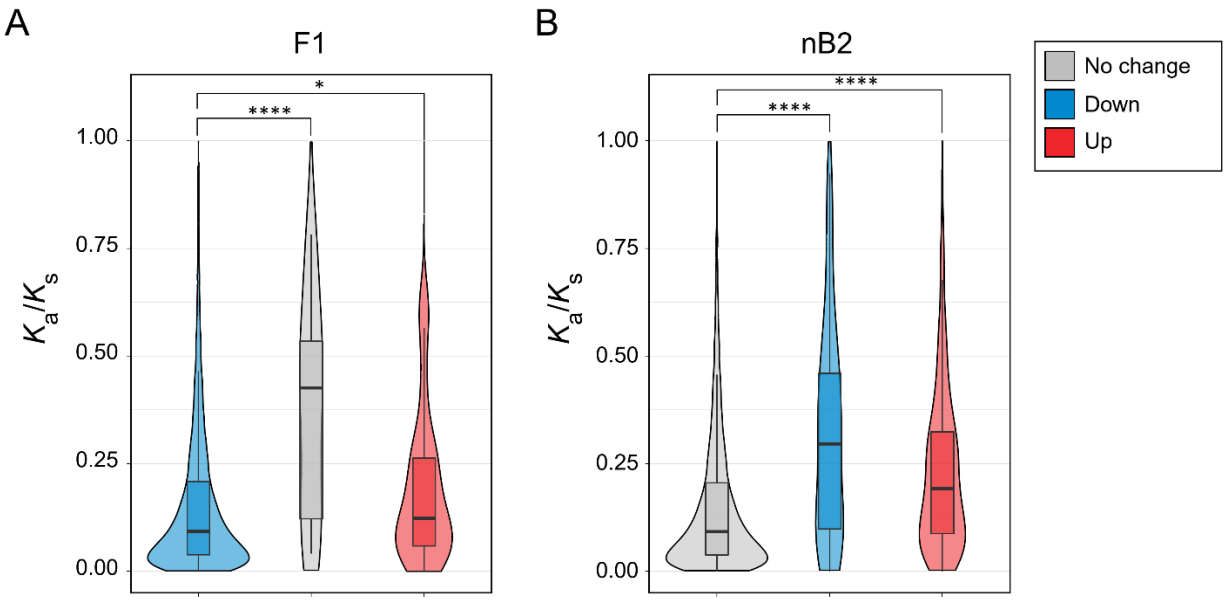
Supplemental Fig. S3 Pairwise comparison of normalized TE read counts between hybrid F1 embryos and *C. briggsae* embryos (A) or *C. nigoni* (B) embryos.



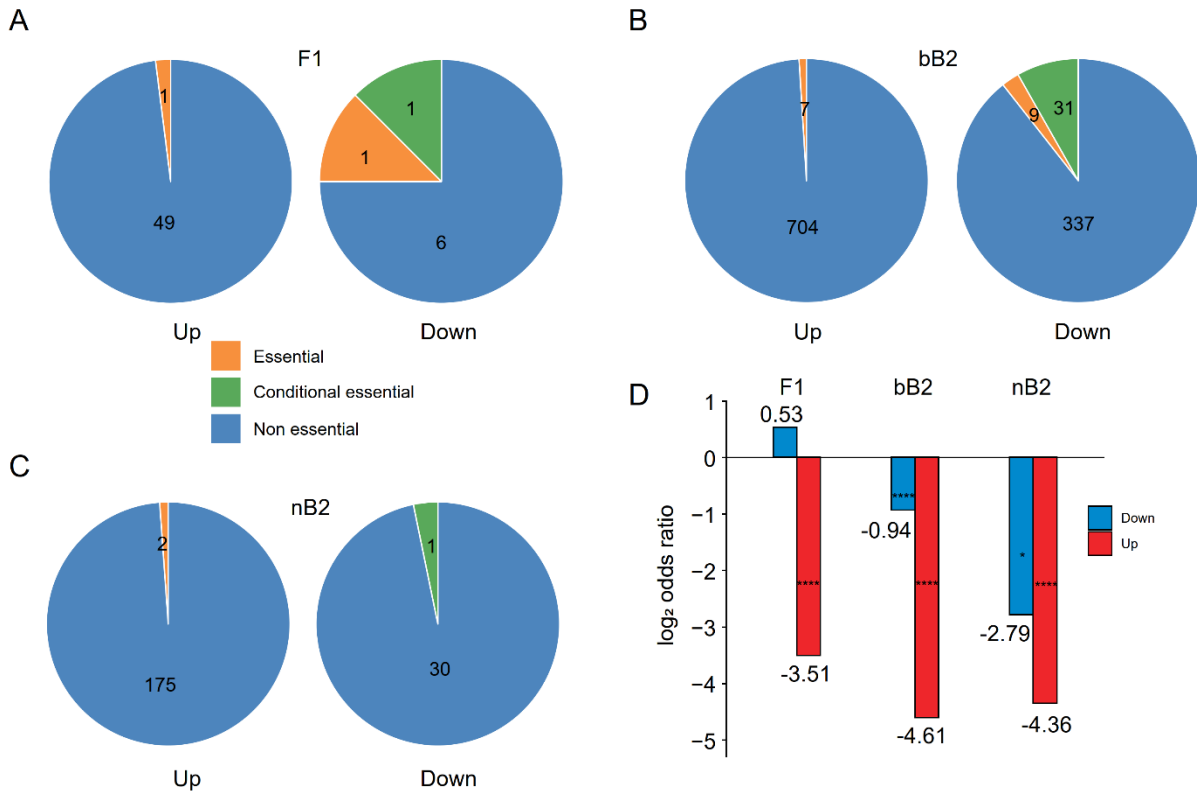
Supplemental Fig. S4 TE expression in the parental, F1, bB2 and nB2 hybrid embryos at class level. (A) Stacked bar plot showing the cumulative number of expressing TE families from various TE classes (differentially color-coded). **(B)** Violin plots with overlaid box plots show the transcript abundance in Fragments Per Kilobase of transcript per Million (FPKM) of each TE class as in **(A)** in the parental, F1, bB2 and nB2 hybrid embryos.



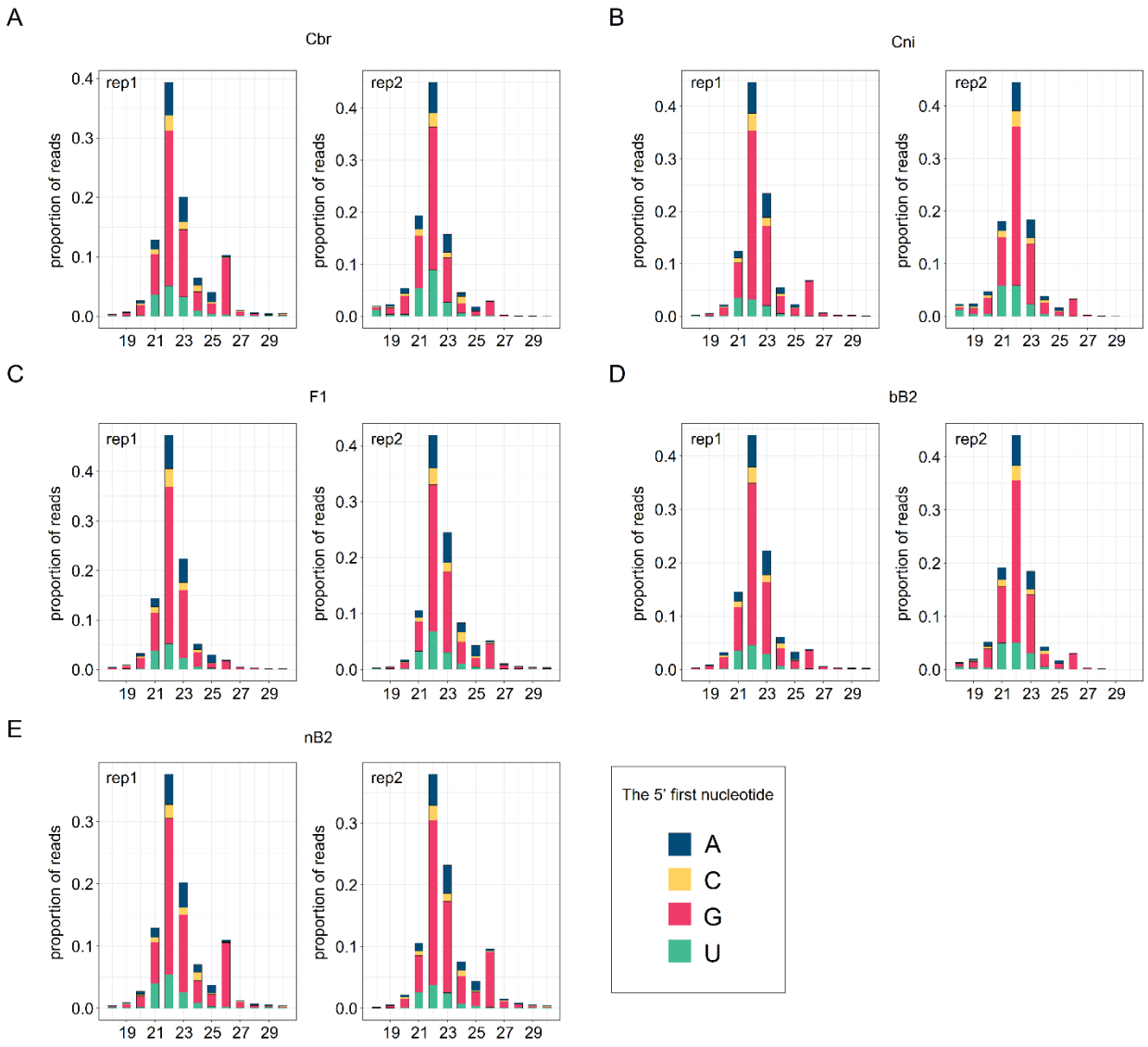
Supplemental Fig. S5 Pairwise comparison of transcript abundances of the one-to-one orthologous coding genes between *C. briggsae* and *C. nigoni* (A), F1 and *C. nigoni* (B), F1 and *C. nigoni* (C), F1 and bB2 (D), F1 and nB2 (E), respectively. Shown are scatter plots of normalized read count in RPKM with Pearson correlation coefficient (*R*) indicated. Linear regression lines are shown in blue.



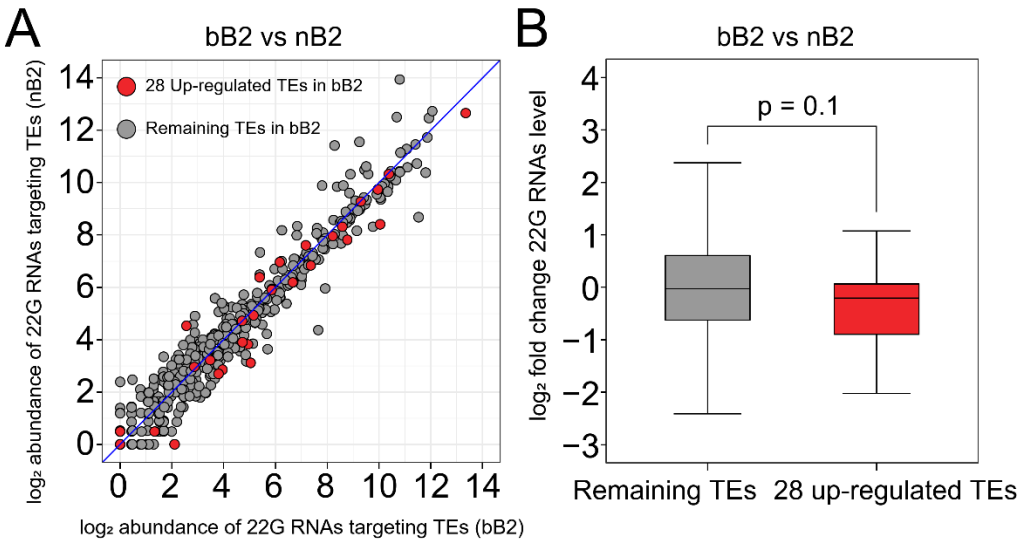
Supplemental Fig. S6 Evolution speed of misregulated genes in the hybrids. Violin plots with overlaid box plots showing the Ka/Ks values of genes with a significant up- or down-regulation, or no significant change in the F1 (A) and nB2 embryos (B) relative to both parents (See Fig. 4A). **** $p<0.0001$ (Wilcoxon ranked sum test).

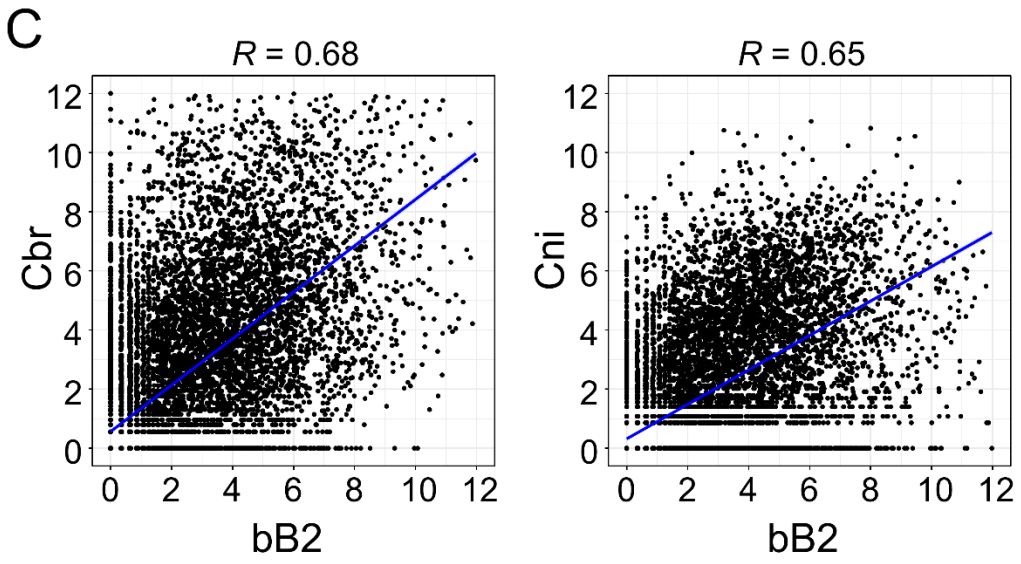


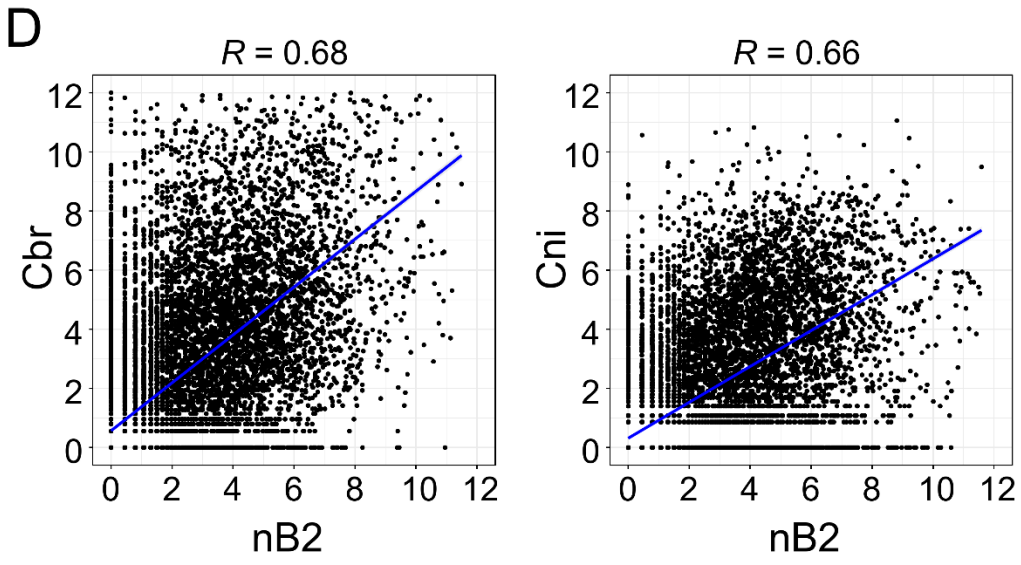
Supplemental Fig. S7 Most misregulated genes in the hybrid embryos are non-essential ones. (A-C) Pie charts showing the number of up- or down-regulated essential, conditional essential and non-essential genes in the F1 **(A)**, bB2 **(B)** and nB2 embryos **(C)**. Note that only the *C. elegans* orthologues of misregulated genes were counted. **(D)** Enrichment analysis of the up- or down-regulated essential/conditionally essential genes in the hybrid embryos relative to all the essential/conditionally essential genes. Numbers shown below are the log2 odds ratio. * $p < 0.05$, *** $p < 0.001$ (Fisher's extract test).



Supplemental Fig. S8 Distributions of length and 5' first nucleotide of small RNAs in the parental, F1, bB2 and nB2 embryos. Stacked bar plots show the proportions of small RNA reads with length ranging from 18 to 30 nucleotides for the two replicates of *C. briggsae* **(A)**, *C. nigoni* **(B)**, F1 **(C)**, bB2 **(D)** and nB2 embryos **(E)**. The 5' first nucleotides are differentially color coded.


A
bB2 vs nB2
28 Up-regulated TEs in bB2
Remaining TEs in bB2
log₂ abundance of 22G RNAs targeting TEs (nB2)
log₂ abundance of 22G RNAs targeting TEs (bB2)
B
bB2 vs nB2
p = 0.1
log₂ fold change 22G RNAs level
Remaining TEs
28 up-regulated TEs


C
R = 0.68
Cbr
bB2
R = 0.65
Cni
bB2


D
R = 0.68
Cbr
nB2
R = 0.66
Cni
nB2

Supplemental Fig. S9 Expression of 22G RNAs targeting TEs or coding genes. (A) Pairwise comparison of normalized read count of 22G small RNA targeting TEs between the bB2 and nB2 hybrid embryos. The up-regulated (see Fig. 2C) and non-up-regulated TEs in bB2 are color-coded with red and gray, respectively. **(B)** Box plots showing the normalized log2 fold changes in read abundance of 22G small RNAs (bB2 over nB2) targeting the 28 up-regulated TEs (see Fig. 2C) (red) or the non-up-regulated TEs (gray). Wilcoxon ranked sum test p value is indicated. **(C-D)** Pairwise comparisons of normalized read abundances of 22G RNAs targeting coding genes between bB2 (A) or nB2 (B) and parental embryos. Also shown is Pearson correlation coefficient (*R*). Linear regression lines are shown in blue.

Supplemental Table legends

Supplemental Table 1 (separate file)

Sequencing read statistics.

Supplemental Table 2 (separate file)

The log2 normalized read counts of RNAseq reads mapped to TE consensus in the parental and hybrid embryos.

Supplemental Table 3 (separate file)

log2FC(bB2 vs parental species) of the 28 up-regulated TEs in bB2 embryos.

Supplemental Table 4 (separate file)

The one-to-one orthologous genes of *C. briggase* and *C. nigoni*.

Supplemental Table 5 (separate file)

log2FC of coding genes (F1 vs parental strains).

Supplemental Table 6 (separate file)

log2FC of coding genes (bB2 vs parental strains).

Supplemental Table 7 (separate file)

log2FC of coding genes (nB2 vs parental strains).

Supplemental Table 8 (separate file)

K_a/K_s value for all the one-to-one orthologs of *C. briggase* and *C. nigoni*.

Supplemental Table 9 (separate file)

Normalized RPM of *C. briggsae* piRNAs in the parental and hybrid embryos.

Supplemental Table 10 (separate file)

Normalized RPM of *C. nigoni* piRNAs in the parental and hybrid embryos.

Supplemental Table 11 (separate file)

log2 normalized read counts of 22G RNAs mapping to the TE consensus library.

Supplemental Table 12 (separate file)

log2 normalized read counts reads counts of 22G RNAs mapping to the coding genes.

Supplemental Table 13 (separate file)

log2 normalized reads counts of miRNAs.

Supplemental Methods and References

Analysis of mRNA-seq reads for coding gene expression

Raw RNA-seq reads were trimmed and filtered as described for TE analysis. The processed reads from all the samples were mapped against a combined genome of cb4 (WS280) (Ross et al. 2011) and cn2 (Ren et al. 2018) using STAR (version 2.7.0) (Dobin et al. 2013) with the parameter "--outSAMmultNmax 1", to assign only one alignment record to each read. Similarly, a combined gene annotation file from cb4 and cn2 was manually concatenated to map the processed reads. As the multiple aligned reads were randomly assigned by STAR, ambiguous reads that could be mapped to both the *C. briggsae* and *C. nigoni* genome were randomly assigned and only counted once. The read counts were summarized using featureCounts (v2.0.1) (Liao et al. 2014) with the parameters "-O -M -p -t gene -f -g ID". One-to-one orthologs between *C. briggsae* and *C. nigoni* were inferred by the mutual best hit of the protein sequences of the longest isoform of each gene using BLASTP (v2.7.1), and are listed in Table S4. The read count for each member of a given ortholog pair was combined to represent the expression level of the pair, named after the *C. briggsae* gene. *C. briggsae*- and *C. nigoni*-specific genes were excluded from subsequent analysis. For expression inheritance profile analysis, the one-to-one orthologs were subjected to read normalization and the fold change of expression levels was calculated using the edgeR package (v 3.34.1) (Robinson et al. 2010). The gene inheritance categories were classified similarly to that described previously (McManus et al. 2010). Namely, a change in the normalized read count less than 1.3-fold was treated as no change in the expression level. The expression distance of each gene in the hybrids, defined as the absolute Euclidean distances of the expression differences between the hybrids and the two parental strains, was calculated as described (Sánchez-Ramírez et al. 2021). For differentially expressed genes (DEGs) analysis,

the significantly dysregulated genes (up- or down-regulation) were defined as those that showed a change in the expression level of at least two-fold with FDR corrected *p* value output by the edgeR being < 0.05.

The K_a/K_s value of each gene was calculated using paraAT (Zhang et al. 2012) and KaKs_calculator (Zhang et al. 2006). The Gene Ontology analysis was performed using the clusterProfiler package (v 4.0.5) (Yu et al. 2012). Only enriched biological processes with FDR corrected $p<0.05$ were used in generating bubble plots.

Small RNA-seq analysis

Small (18-36 nt) RNAs were aligned to the same combined genome as used for the coding gene analysis , using bBowtie (Langmeadet al. 2009), allowing for 0 mismatches and reporting only one match per small RNA sequence. piRNA annotations in *C. briggsae* were taken from previous annotation (Beltran et al. 2019). The upstream sequences taken from Beltran et al., define the piRNA motif in *C. briggsae* and were used as an input to MEME (Bailey et al. 2015). The resulting position weight matrix was used to scan the *C. nigoni* genome and 21U-RNAs were predicted on this basis. Notably 21U-RNA sequences were almost entirely distinct between the two species, enabling unambiguous mapping. Counts for 21URNAs were obtained using BEDTools intersect. 22G RNAs were identified and those mapping antisense to either TE consensus or protein-coding genes were identified using BEDTools (Quinlan andHall 2010) intersect. To account for the fact that orthologous protein-coding genes will have highly similar sequences between *C. briggsae* and *C. nigoni*, reads mapping to either *C. briggsae* or *C. nigoni* orthologs were summed together to produce a single total count for each gene. miRNAs in *C.*

nigoni were identified by taking the sequences of *C. briggsae* miRNAs (Kozomara et al. 2019) and aligning them to the *C. nigoni* genome using Bowtie, allowing for up to 1 mismatch to account for potential divergence in miRNA sequences. Counts for miRNAs were then obtained using BEDTools intersect and summed for orthologous miRNAs to produce a single value for each miRNA. To perform data analysis, piRNAs were normalized to the total 22G-RNA size factors calculated using DESeq2 (Love et al. 2014). 22G-RNA counts were normalized to the total 22G-RNA size factors using DESeq2. miRNA counts were normalized to the total miRNA size factors using DESeq2.

References

BaileyTL, JohnsonJ, GrantCE, NobleWS. 2015. The MEME Suite. *Nucleic Acids Res* **43**: W39–W49. https://pubmed.ncbi.nlm.nih.gov/25953851/ (Accessed July7, 2022).

BeltranT, BarrosoC, BirkleTY, StevensL, SchwartzHT, SternbergPW, FradinH, GunsalusK, PianoF, SharmaG, et al. 2019. Comparative Epigenomics Reveals that RNA Polymerase II Pausing and Chromatin Domain Organization Control Nematode piRNA Biogenesis. *Dev Cell* **48**: 793-810.e6. https://pubmed.ncbi.nlm.nih.gov/30713076/ (Accessed July7, 2022).

DobinA, DavisCA, SchlesingerF, DrenkowJ, ZaleskiC, JhaS, BatutP, ChaissonM, GingerasTR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15–21. https://pubmed.ncbi.nlm.nih.gov/23104886/ (Accessed June22, 2022).

KozomaraA, BirgaoanuM, Griffiths-JonesS. 2019. miRBase: from microRNA sequences to function. *Nucleic Acids Res* **47**: D155–D162. https://pubmed.ncbi.nlm.nih.gov/30423142/ (Accessed July7, 2022).

LangmeadB, TrapnellC, PopM, SalzbergSL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**. https://pubmed.ncbi.nlm.nih.gov/19261174/ (Accessed July7, 2022).

LiaoY, SmythGK, ShiW. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**: 923–930. https://pubmed.ncbi.nlm.nih.gov/24227677/ (Accessed June22, 2022).

LoveMI, HuberW, AndersS. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**: 1–21.

https://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0550-8 (Accessed July7, 2022).

McManusCJ, CoolonJD, DuffMO, Eipper-MainsJ, GraveleyBR, WittkoppPJ. 2010. Regulatory divergence in Drosophila revealed by mRNA-seq. *Genome Res* **20**: 816–25. http://www.ncbi.nlm.nih.gov/pubmed/20354124 (Accessed June7, 2018).

QuinlanAR, HallIM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842. https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btq033 (Accessed March3, 2021).

RenX, LiR, WeiX, BiY, HoVWS, DingQ, XuZ, ZhangZ, HsiehC-L, YoungA, et al. 2018. Genomic basis of recombination suppression in the hybrid between Caenorhabditis briggsae and C. nigoni. *Nucleic Acids Res*. http://www.ncbi.nlm.nih.gov/pubmed/29325078 (Accessed January16, 2018).

RobinsonMD, McCarthyDJ, SmythGK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**: 139–140. http://www.ncbi.nlm.nih.gov/pubmed/19910308 (Accessed June11, 2018).

RossJA, KoboldtDC, StaischJE, ChamberlinHM, GuptaBP, MillerRD, BairdSE, HaagES. 2011. Caenorhabditis briggsae recombinant inbred line genotypes reveal inter-strain incompatibility and the evolution of recombination. *PLoS Genet* **7**: e1002174.

Sánchez-RamírezS, WeissJG, ThomasCG, CutterAD. 2021. Widespread misregulation of inter-species hybrid transcriptomes due to sex-specific and sex-chromosome regulatory evolution. *PLOS Genet* **17**: e1009409. https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1009409 (Accessed April27, 2022).

YuG, WangLG, HanY, HeQY. 2012. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**: 284–287. https://pubmed.ncbi.nlm.nih.gov/22455463/ (Accessed June29, 2022).

ZhangZ, LiJ, ZhaoXQ, WangJ, WongGKS, YuJ. 2006. KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. *Genomics Proteomics Bioinformatics* **4**: 259–263. https://pubmed.ncbi.nlm.nih.gov/17531802/ (Accessed June29, 2022).

ZhangZ, XiaoJ, WuJ, ZhangH, LiuG, WangX, DaiL. 2012. ParaAT: a parallel tool for constructing multiple protein-coding DNA alignments. *Biochem Biophys Res Commun* **419**: 779–781. https://pubmed.ncbi.nlm.nih.gov/22390928/ (Accessed June29, 2022).