Science
AAAS

Supplementary Materials for

Rapid genome shrinkage in a self-fertile nematode reveals sperm competition proteins

Da Yin, Erich M. Schwarz*, Cristel G. Thomas, Rebecca L. Felde, Ian F. Korf, Asher D. Cutter, Caitlin M. Schartner, Edward J. Ralston, Barbara J. Meyer, Eric S. Haag*

* correspondence to: ehaag@umd.edu, ems394@cornell.edu

This PDF file includes:

Materials and Methods
Figs. S1 to S9
Tables S1 to S9 (captions only for S3-S5 and S8)
Captions for Datasets S1 to S5

Other Supplementary Materials for this manuscript include the following:

Tables as Excel files:
- Table S3
- Table S4
- Table S5
- Table S8

Datasets as zipped archives:
- Supplementary Dataset S2a
- Supplementary Dataset S2b

Datasets available via internet server:
- Supplementary Dataset S1
- Supplementary Dataset S3
- Supplementary Dataset S4
- Supplementary Dataset S5

Materials and Methods

Strains. The inbred wild-type *C. nigoni* strains JU1421 and JU1422 were obtained from Marie-Anne Félix (Ecole Normale Supérieure) and from the *Caenorhabditis* Genetics Center (CGC). Because each was derived by 25 rounds of inbreeding from the wild isolate strain JU1325 (*46*), we expected that they would be largely homozygous, but with some residual heterozygosity (*30*). Wild-type *C. briggsae* strain AF16 (*47, 48*) and *C. remanei* strains EM464 (*49, 50*) and SB146 (*51, 52*) were also obtained from the CGC. *C. briggsae unc-119(nm67)* (*53*) and *she-1(v35)* (*54*) mutants were derived from AF16. All species were cultured on standard NGM plates (*55*), supplemented with additional agar (to 2.2%) to discourage burrowing. The *C. briggsae* strains RW20025 and JU936 were gifts from Zhongying Zhao (Hong Kong Baptist University) and Marie-Anne Félix, respectively.

Genomic DNA purification. For Illumina sequencing of *C. nigoni* JU1422, genomic DNA was prepared by methods optimized for short-insert paired-end reads (*56*). However, PacBio sequencing of *C. nigoni* JU1422 required genomic DNA of very high molecular weight that was free of contamination by both RNA and sequencing inhibitors (such as EDTA). To avoid fragmenting genomic DNA, wide-bore pipette tips were used throughout, generated by cutting off the ends of normal pipette tips with a sterile razor blade. Worms were grown on NGM plates, washed off with M9 buffer, and rinsed 2-3 times with microcentrifugation (1200 rpm for 1 minute) at 4°C. The worm pellet was washed once in disruption buffer (200 mM NaCl; 50 mM EDTA; 100 mM Tris, pH 8.5) without SDS and again microfuged for 1 minute at 4°C. The pellet was then resuspended in 5 volumes of disruption buffer with a final concentration of 0.5% (w/v) added SDS, refrigerated in a -80°C freezer to soften tissues, and thawed to room temperature. Proteinase K (20 mg/ml stock solution) was added to a final concentration of 100-200 µl/ml, and the mix was incubated for ~5 hours at 68°C until dissolved, with periodic mixing by gentle inversion. Co-purified RNA was digested before phenol-chloroform extraction by adding 2 µl RNase A stock (10 mg/ml) for each 50 µl of sample, and incubating for 30 min. at 37°C. One volume of a 1:1 mixture of mixture of chloroform and buffer-saturated phenol (pH ~8) was added to the sample, and the sample's microfuge tubes were gently inverted until phases were mixed. Tubes were then microfuged at 13,000 RPM at room temperature for 2 minutes. The upper aqueous layer with its genomic DNA was similarly extracted two more times, followed by a final extraction with 24:1 chloroform/isoamyl alcohol. In a fresh tube, 0.1 volumes of 3 M sodium acetate (pH 5.2) and 1 volume of 100% isopropanol were added and mixed by gentle inversion. After incubation at -20°C overnight, DNA was pelleted by microfuging at 13,000 RPM for 10-15 minutes at 4°C. The DNA pellet was washed with 70% ethanol, inverting the sealed tube a few times, and reseating the DNA pellet with a spin at 13,000 RPM for 5 minutes at 4°C. Remaining supernatant was removed by pipetting, and the pellet allowed to air-dry at room temperature, taking care not to overdry it. The genomic DNA pellet was then resuspended in 50 µl EB buffer from Qiagen; to avoid shearing, the high molecular weight DNA was not resuspended by pipetting, but rather was allowed to slowly dissolve overnight at 4°C.

Genomic DNA sequencing. Genomic DNA from JU1422 was subjected to Blue Pippin selection, and sequenced to ~96x coverage (given an initial genome size estimate of 130 Mb) with Pacific Biosciences (PacBio) 20-kb libraries at the Cold Spring Harbor Laboratory genome facility. These data totaled 12,538,321,717 nt in 1,480,550 reads, with a mean length of 8,469 nt (maximum, 47,453 nt; minimum, 50 nt). In addition, we also sequenced JU1422 genomic DNA to ~100x coverage with Illumina paired-end libraries at the UC Berkeley genome facility.

Genome assembly. PacBio reads, despite their high (~15%) error rate, can be self-corrected and assembled if their genomic coverage is sufficiently high (50x or more). We generated an initial genome assembly from our PacBio data by using PBcR-MHAP 8.3rc2 (*57*) (*https://sourceforge.net/projects/wgs-assembler/files/wgs-assembler/wgs-8.3*) to self-correct the longest 40x coverage of reads, and then to assemble the longest 25x coverage of reads. We ran PBcR-MHAP with Java Runtime Environment 1.8.0_31, the parameters "*-pbCNS -noclean -length 500 -partitions 200*", and a .spec file containing the following parameters, designed for assembly of a haploid metazoan genome of moderate size: "*merSize=16; mhap=-k 16 --num-hashes 512 --num-min-matches 3 --threshold 0.04 --weighted; useGrid=0; scriptOnGrid=0; ovlMemory=32; ovlStoreMemory=32000; threads=32; ovlConcurrency=1; cnsConcurrency=8; merylThreads=32; merylMemory=32000; ovlRefBlockSize=20000; frgCorrThreads = 16; frgCorrBatchSize = 100000; ovlCorrBatchSize = 100000; sgeScript = -pe threads 1; sgeConsensus = -pe threads 8; sgeOverlap = -pe threads 15 –l mem=2GB; sgeCorrection = -pe threads 15 –l mem=2GB; sgeFragmentCorrection = -pe threads 16 –l mem=2GB; sgeOverlapCorrection = -pe threads 1 –l mem=16*". This yielded an initial assembly of 143 Mb and 485 contigs, with a contig N50 of 2.2 Mb.

Caenorhabditis are typically grown on monoxenic cultures of *Escherichia coli* OP50 (*58*), and they can also harbor cryptic infections of diverse bacteria such as *Leucobacter* sp. AEAR (*59*) and *Stenotrophomonas maltophilia* (*60*). Despite efforts to clean worms before sequencing, such microbes can and do contaminate nematode genomic assemblies (*61*). We tested our initial genome assembly contigs for bacterial origin by searching a custom database of bacterial genomes with *blastn* from BLAST+ 2.2.31 (*62*). We downloaded 3,000 bacterial genomes at EBI on 11/11/2015 from *ftp://ftp.ebi.ac.uk/pub/databases/fastafiles/embl_genomes/genomes/Bacteria*; to these, we added the sequence of *E. coli* OP50 at NCBI from *http://www.ncbi.nlm.nih.gov/Traces/wgs/?download=ADBT01.1.fsa_nt.gz*, and the sequence of *Leucobacter* sp. AEAR from Lsp41_scaffolds.fa in the PubMed Central archive *http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3583267/bin/bbi-7-2013-055files.zip*. The full bacterial genome set was formatted for *blastn* searches with *makeblastdb* with arguments "*-dbtype nucl -input_type fasta*"). Searching was done with *blastn* with the parameters "*-outfmt 6 -task megablast -perc_identity 50 -evalue 1e-06*". We used a custom Perl script to identify 52 contigs totaling 3.4 Mb that matched bacterial genomic DNA along 50% or more of their lengths, and removed such likely contaminants from the assembly.

Individual PacBio reads can be used to improve existing genome assemblies by linking existing sequences. To do this with our decontaminated *C. nigoni* genome, we used PBJelly2 from PBSuite 15.8.24 (*63*) (*http://sourceforge.net/projects/pb-jelly/files*) and our error-corrected 40x coverage of reads, with the parameters "*-minMatch 50 -minPctIdentity 95 -bestn 5 -nCandidates 20 -maxScore -500 -noSplitSubreads*".

The accuracy of PacBio-based genome assemblies can be driven to final consensus accuracies in excess of 99.999% (QV of >50) by detailed comparison to the raw pulse and base-call information of their original PacBio read files (*64*). To do this with our PBJelly2-scaffolded *C. nigoni* genome, we used Quiver (*64*) from SMRTanalysis 2.3.0 (downloaded from *https://s3.amazonaws.com/files.pacb.com/software/smrtanalysis/2.3.0/smrtanalysis_2.3.0.140936.run* and *https://s3.amazonaws.com/files.pacb.com/software/smrtanalysis/2.3.0/smrtanalysis-patch_2.3.0.140936.p4.run*, and then installed as described in *http://www.pacb.com/wp-content/uploads/2015/09/SMRT-Analysis-Software-Installation-v2.3.0.pdf*). Quiver requires sorted alignments of very large amounts of data; this was done using the divide-and-conquer strategy described in *https://github.com/PacificBiosciences/pbalign/wiki/Tutorial:-How-to-divide-and-conquer-large-datasets-using-pbalign*, along with the programs *pbalign*, *cmph5tools.py*, and *samtools faidx* from the SMRTanalysis 2.3.0 software package. Quiver added 103,264 nt while removing one contig from the *C. nigoni* genome assembly.

PacBio sequencing and assembly, even after revision with Quiver, can have residual errors (e.g., in homopolymeric sequences) that are most easily remedied by non-PacBio sequence data. We thus further corrected our Quiver-polished *C. nigoni* genome with Pilon 1.14 (*65*) and paired-end Illumina sequencing data, after mapping the Illumina data to the genome with bowtie2 2.2.6 (*66*) and converting the mappings to aligned BAM files with SAMTools 1.2 (*67*). To restrict Pilon to small corrections while leaving diploid alleles unchanged, we used the arguments "*--changes --fix bases --chunksize 8000000 --diploid*". Pilon made 10,328 changes to the genome assembly with the net effect of adding 2,898 nt to it.

Genome assemblies of diploid, heterozygous organisms can contain allelic sequences that can be computationally resolved into a single quasi-haploid consensus assembly. To resolve such minor alleles in the Quiver- and Pilon-polished *C. nigoni* genome assembly, we used HaploMerger2 20151106 (*68*) (*http://mosas.sysu.edu.cn/genome/download_softwares.php*) by protocols given in the software documentation. In particular, we generated species-specific transition matrices by comparing the largest 10% of *C. nigoni* contigs to the other 90% with HaploMerger2's *lastz_D_Wrapper.pl*, and repeatmasked the *C. nigoni* assembly with windowmasker (*69*) (*ftp://ftp.ncbi.nlm.nih.gov/pub/agarwala/windowmasker/windowmasker*) before running HaploMerger2 itself. HaploMerger2 entailed three rounds of misjoin detection and removal (with the program *hm.batchA*), one round of creating an initial haploid assembly, refining its single alleles, and creating a final haploid assembly (with *hm.batchB*), and three rounds of removing tandem repeats from the haploid assembly (with *hm.batchD*). For *hm.batchA*, we used the parameter "*identity=80*"; for *hm.batchD*, we used the parameters "*filterAli=4000 minLen=5000*", "*filterAli=2400 minLen=3000*", and "*filterAli=1000 minLen= 1500*".

RNA-seq data. We used biologically triplicated RNA-seq data from male and female adult *C. nigoni*. These data for *C. nigoni* were generated as single-end 100-nt reads from the strain JU1421, which like JU1422 also is an inbred derivative of JU1325 (*46*). We also used RNA-seq data for mixed-sex whole-animal *C. nigoni* that had been previously generated by the modENCODE consortium (*70*); these data were paired-end 100-nt reads with a 231 nt insert size, available at *https://www.ncbi.nlm.nih.gov/sra/?term=SRR241784*.

*RNA-seq data for non-*nigoni Caenorhabditis. We used previously published RNA-seq data to determine sex-biased gene expression and assemble cDNAs for *C. briggsae*, *C. remanei*, *C. brenneri*, and *C. elegans*. The sex-biased RNA-seq data were from Thomas et al. (*11*); other RNA-seq data were from modENCODE (*70*) and (for *C. brenneri*) from Romiguier et al. (*71*). For each species, the SRA accession numbers were as follows. For *C. briggsae*: SRR580389, SRR580390, SRR580391, SRR580392, SRR580393, and SRR580394 (from Thomas et al.). For *C. remanei*: SRR580407, SRR580408, SRR580409, SRR580410, SRR580411, and SRR580412 (from Thomas et al.); SRR149192, SRR149193, SRR357248, SRR357249, SRR357250, SRR357251, SRR357252, SRR357253, SRR357254, SRR357255, SRR357256, SRR357257, SRR357258, SRR357259, SRR357260, SRR357261, SRR357262, SRR357263, SRR357264, SRR357265, SRR357266, and SRR357267 (from modENCODE). For *C. brenneri*: SRR580401, SRR580402, SRR580403, SRR580404, SRR580405, and SRR580406 (from Thomas et al.); SRR352227, SRR352228, SRR352229, SRR352230, SRR352231, SRR352232, SRR352233, SRR352234, SRR352235, SRR352236, SRR352237, SRR352238, SRR352239, SRR352240, SRR359072, and SRR359075 (from modENCODE); SRR1324828, SRR1324829, SRR1324830, SRR1324831, SRR1324832, SRR1324833, SRR1324834, SRR1324835, SRR1324836, and SRR1324837 (from Romiguier et al.). For *C. elegans*: SRR580383, SRR580384, SRR580385, SRR580386, SRR580387, and SRR580388 (from Thomas et al.).

Assembly of cDNA from RNA-seq data. We assembled all of our RNA-seq data into cDNA with Trinity 2.2.0 (*72*) (*https://github.com/trinityrnaseq/trinityrnaseq/releases*), in both genome-guided and ab initio form, with the arguments "*--no_version_check --seqType fq --normalize_reads --CPU 8 --max_memory 80G --trimmomatic --verbose --min_contig_length 100*"; for genome-guided assembly, the argument "*--genome_guided_max_intron 10000*" was also used. The genome-guided cDNA assembly was subsequently used to guide gene parameter construction and gene prediction; to avoid circularity, the non-genome-guided cDNA assembly was subsequently used to enable cDNA scaffolding of the genome assembly.

Genome scaffolding with cDNA and peptide data. We further scaffolded our *C. nigoni* genome assembly with *C. nigoni* non-genome-guided cDNA (assembled from RNA-seq data) and *C. briggsae* peptide sequences (downloaded from WormBase release 254; *ftp://ftp.sanger.ac.uk/pub2/wormbase/releases/WS254/species/c_briggsae/PRJNA10731/c_briggsae.PRJNA10731.WS254.protein.fa.gz*), using L_RNA_scaffolder (*73*) (*http://www.fishbrowser.org/*

software/L_RNA_scaffolder/downloads/L_RNA_scaffolder.tar.gz) and PEP_scaffolder (*74*) (*http://www.fishbrowser.org/software/PEP_scaffolder/downloads/PEP_scaffolder.tar.gz*); both programs also required BLAT 36 and Bioperl 1.6.923. We used arguments of "*-f 1*" for both programs, and an argument of "*-e 100000*" for PEP_scaffolder.

Evaluating and correcting genomes. To determine the completeness and homozygosity of our successive genome assembly versions, we used CEGMA 2.4 (*21*). To detect possible misjoins in our *C. nigoni* assemblies, we compared them globally to the hard-repeatmasked CB4 genome assembly of *C. briggsae* from WormBase release WS254 (*ftp://ftp.sanger.ac.uk/pub2/wormbase/releases/WS254/species/c_briggsae/PRJNA10731/c_briggsae.PRJNA10731.WS254.genomic_masked.fa.gz*) and looked for clear instances of two chromosomes being linked. We generated graphical and text summaries of these global alignments with the programs *nucmer*, *show-coords*, and *mummerplot* from MUMmer 3.23 (*75*). For *nucmer*, we used the arguments "*--mum --mincluster 100 --maxgap 300*"; for *mummerplot*, we used the arguments "*--filter --large -t png*". In four cases of possible misjoins, we further evaluated them by comparing them (via BlastN) both to our error-corrected individual PacBio reads and to scaffolds from a preliminary *C. nigoni* assembly that we had generated from Illumina sequence data alone. This led us to classify two contigs as probably containing misjoins, which we split manually.

Chromosome tiling. To determine a subset of the *C. nigoni* genome assembly that aligned syntenically (tiled) onto the *C. briggsae* genome, we used *show-tiling* from MUMmer 3.23 with the arguments "*-l 1 -g -1 -i 80.0 -v 1.0 -V 0*". We used the resulting tiling data with custom Perl scripts, along with visual inspection of MUMmer alignments to *C. briggsae*, to generate a chromosomally-aligned and pseudochromosomal version of the *C. nigoni* genome. Within each pseudochromosome, we linked successive contigs or scaffolds with 1000-nt blocks of N residues. This yielded a final assembly with 6 major pseudochromosomal superscaffolds that totaled 118 Mb in length, and a residuum of 150 unaligned contigs that totaled 11.7 Mb.

*Genome sequences for non-*nigoni *Caenorhabditis*. For the reference genome of *C. briggsae*, we used the CB4 assembly in the WS254 release of WormBase (*ftp://ftp.wormbase.org/pub/wormbase/releases/WS254/species/c_briggsae/PRJNA10731/c_briggsae.PRJNA10731.WS254.genomic.fa.gz*), which was generated from the wild-type strain AF16. For *C. remanei*, we used the genome assembly of Fierst et al. from WormBase WS254 (*ftp://ftp.wormbase.org/pub/wormbase/releases/WS254/species/c_remanei/PRJNA248909/c_remanei.PRJNA248909.WS254.genomic.fa.gz*), because it was generated from a *C. remanei* strain that was effectively free of residual heterozygosity (*10*). For *C. brenneri*, we began work with the genome assembly by the Washington University Genome Center from WormBase WS250 (*ftp://ftp.wormbase.org/pub/wormbase/releases/WS250/species/c_brenneri/PRJNA20035/c_brenneri.PRJNA20035.WS250.genomic.fa.gz*). However, this genome is known to have substantial residual heterozygosity (*30*), which we expected would confound analyses of gene homology and similarity with spurious paralogs. Therefore, after removing one sequence that had a strong *blastn* hit to our EBI microbial database, we significantly reduced the likely heterozygosity of the *C. brenneri* assembly *in silico* with HaploMerger 2 (*68*). This HaploMerger2-compressed form of the *C. brenneri* genome assembly was used for gene predictions, that in turn were used for our orthology analyses of *C. nigoni*; this version of the *C. brenneri* genome assembly is provided in **dataset S4**. For *C. elegans*, we used the N2 assembly in the WS254 release of WormBase (*ftp://ftp.wormbase.org/pub/wormbase/releases/WS254/species/c_elegans/PRJNA13758/c_elegans.PRJNA13758.WS254.genomic.fa.gz*).

Identifying repetitive DNA. We used BuildDatabase and RepeatModeler from RepeatModeler-open-1-0-8 (*http://www.repeatmasker.org*) to generate libraries of repetitive DNA elements for the genomes of *C. nigoni* and *C. briggsae*. Both programs were run with the argument "*-engine ncbi*", and were supported by the programs nseg (*ftp://ftp.ncbi.nih.gov/pub/seg/nseg*) (*76*), Tandem Repeats Finder 4.09 (*http://tandem.bu.edu/trf/trf409.legacylinux64.download.html*) (*77*), RepeatScout 1.0.5 (*http://www.repeatmasker.org/RepeatScout-1.0.5.tar.gz*) (*78*), RECON 1.08 (*http://www.repeatmasker.org/RECON-1.08.tar.gz*) (*78*), nhmmscan in HMMER 3.1b2 (*http://eddylab.org/software/hmmer3/3.1b2/hmmer-3.1b2-linux-intel-x86_64.tar.gz*) (*79*), and RMBlast (*ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/2.2.28/ncbi-blast-2.2.28+-x64-linux.tar.gz* and *ftp://ftp.ncbi.nlm.nih.gov/blast/executables/rmblast/2.2.28/ncbi-rmblastn-2.2.28-x64-linux.tar.gz*). RepeatModeler was provided the repeat libraries DFAM 2.0 (*http://www.dfam.org/web_download/Release/Dfam_2.0/Dfam.hmm.gz*) and RMLibrary 20150807 (*http://www.girinst.org/server/RepBase/protected/repeatmaskerlibraries/repeatmaskerlibraries-20150807.tar.gz*). After RepeatModeler had generated the initial *C. nigoni* and *C. briggsae* libraries, we filtered them to remove likely high-copy protein-coding and ncRNA genes as follows. We searched them for ncRNAs with cmsearch from INFERNAL 1.1.2 (using the argument "*--cut_ga*" to impose family-specific significance thresholds) and RFAM 12.1; this detected an rDNA sequence in the initial *C. nigoni* libraries, which we removed. To enable searches for non-repetitive proteins and motifs we used getorf from EMBOSS 6.5.7 (using the argument "-minsize 90") (*80*) to extract peptides of 30 or more residues from the initial repeat libraries. We searched these repeat-encoded peptides with BlastP (from BLAST+ 2.2.31, using the arguments "*-evalue 1e-09 -outfmt 7 -seg yes*") against the *C. elegans* proteome from WormBase (release WS254), and rejected any element that encoded a *C. elegans* proteome match. We again extracted 30+-residue peptides with getorf from the surviving repetitive elements, and searched them with hmmscan from HMMER 3.1b2 (using the argument "*--cut_ga*" to impose family-specific significance thresholds) (*81*) against the protein motif database PFAM (release 30) (*82*). We examined PFAM hits and manually rejected those that were visibly non-repetitive (e.g., motifs for GPCRs). We defined elements passing all of these tests as being genuine repetitive DNA elements, extracted them from the initial libraries, and used them to repeatmask the *C. nigoni* and *C. briggsae* genomes with RepeatMasker-open-4-0-6 (*http://www.repeatmasker.org*). RepeatMasker was run with the arguments "*-e ncbi -s -xsmall -gccalc -gff*".

Gene predictions. We used AUGUSTUS 3.2.2 (*83*) (*http://bioinf.uni-greifswald.de/augustus/binaries/augustus-3.2.2.tar.gz*) to predict protein-coding genes. We began by using the

WebAUGUSTUS server (*84*) (*http://bioinf.uni-greifswald.de/webaugustus/training/create*) to generate initial *C. nigoni*-specific gene parameters, given the syntenically corrected *C. nigoni* genome sequence and our genome-guided Trinity assembly of *C. nigoni* cDNA. This yielded both initial parameters and 4,964 guide gene models. We refined these parameters with *autoAugTrain.pl*, using the arguments "*--optrounds=3 --CRF --useexisting*". To provide hints for gene prediction from genome-guided *C. nigoni* cDNA, we mapped it to the chromosomally-tiled genome assembly with BLAT 36 (*85*) and selected its best alignments with *pslCDnaFilter* from BLAT (argument "*-maxAligns=1*"). We then predicted genes in the chromosomally-aligned genome assembly via AUGUSTUS, using refined parameters and cDNA hints, and with the arguments "*--strand=both --genemodel=partial --noInFrameStop=true --singlestrand= false --maxtracks=3 --alternatives-from-sampling=true --alternatives-from-evidence= true --minexonintronprob=0.1 --minmeanexonintronprob=0.4 --uniqueGeneId=true --protein= on --introns=on --start=on --stop=on --cds=on --codingseq=on --UTR=off --species= caenorhabditis_nigoni --extrinsicCfgFile=[$HOME]/src/augustus-3.2.2/config/extrinsic/ extrinsic.ME.cfg --progress=true --gff3=on*". We used custom Perl scripts to select full-length gene predictions from the resulting GFF3 annotation file, from which we then extracted full-length protein and coding DNA sequences with AUGUSTUS' *getAnnoFasta.pl*.

To improve our ability to compare genomic contents accurately (without biases introduced by different gene-finding methods), we similarly made protein-coding gene predictions for the genomes of *C. briggsae*, *C. remanei*, and *C. brenneri*. These predictions required both cDNA (that could be BLAT-aligned to genomic DNA, in order to provide hints for gene prediction) and species-specific AUGUSTUS parameters. For *C. briggsae*, we used cDNA assembled from publicly available RNA-seq reads, previously generated for WormBase by Gary Williams. For *C. remanei*, we used a combination of cDNA previously assembled from RNA-seq data by Fierst et al. (*10*), and cDNA that we assembled from RNA-seq data generated by Thomas et al. (*11*) and modENCODE (*70*). For *C. brenneri*, we combined cDNAs that we had independently assembled from RNA-seq data generated by Thomas et al. (*11*) and modENCODE (*70*), and from RNA-seq data generated by Romiguier et al. (*71*). For *C. briggsae*, because its reduced hermaphroditic genome might be qualitatively different from that of *C. nigoni*, we created species-specific AUGUSTUS parameters by the same methods that we had used for *C. nigoni*. For *C. remanei* and *C. brenneri*, which are outbreeding species with genome sizes very similar to that of *C. nigoni*, we reused the *C. nigoni* parameters. For all species, AUGUSTUS gene predictions were guided by hints from BLAT-aligned cDNA, and run with general parameters identical to those used for *C. nigoni*.

Previously generated gene predictions. Proteomes, CDS DNA sequences, and GFF3 annotation files generated by others for *Caenorhabditis species* were obtained from the WS254 release of WormBase (for *C. briggsae*, *C. remanei*, *C. elegans*, and *C. japonica*; *ftp://ftp.wormbase.org/pub/wormbase/releases/WS254/species*) or from the *Caenorhabditis* Genomes Project (for *C.* sp. 34 and *C. afra*; *http://download.caenorhabditis.org/v1/sequence*).

RNA-seq expression values and significances. For the *C. nigoni* RNA-seq data, we generated expression values in transcripts per million (TPM) and estimated mapped read counts per gene with Salmon 0.7.0 (*https://github.com/COMBINE-lab/salmon/releases/download/v0.7.0/Salmon-0.7.0_linux_x86_64.tar.gz*) (*86*). For Salmon's *index* program, we used the arguments "*--no-version-check index --kmerLen 31 --perfectHash --type quasi --sasamp 1*"; for Salmon's *quant* program, we used the arguments "*--libType A --seqBias --gcBias --numBootstraps 100 --geneMap* [transcript-to-gene table]", with "*--unmatedReads*" specifying the single-end data (mainly from males or females), or "*--mates1* [first_end_reads] *--mates2* [second_end_reads]" specifying the paired-end modENCODE data. Having generated readcounts for our biological triplicates of male- and female-specific RNA-seq data, we used the exactTest function of edgeR 3.14.0 (*87*) to compute $\log_2$ fold-changes and false discovery rate (FDR) significance values for changes of gene activity between males and females; this test in edgeR has proven to be particularly reliable in computing significant gene expression changes for small numbers of biological replicates (*88*). Equivalent analyses were performed for male-specific and female/hermaphrodite-specific RNA-seq data of *C. briggsae*, *C. remanei*, *C. brenneri*, and *C. elegans*, previously published by Thomas et al. (*11*). In order to detect possible RNA expression of *Cbr-mss-3-ps* in *C. briggsae*, we repeated our *C. briggsae* RNA-seq analyses with Salmon, but used transcripts and gene indexes to which pseudogenic coding exons for *Cbr-mss-3-ps* (predicted by exonerate; see below) had been added.

Protein-coding gene annotations and orthologies. We determined motifs and traits for protein products of our predicted gene set as follows. We predicted signal sequences and transmembrane sequences with Phobius 1.01 (*89*), coiled-coils with NCoils (*90*), and low-complexity domains with PSEG (*91*). We predicted protein motifs from two databases: the PFAM database with hmmscan in HMMER 3.1b2 (*82, 92*), using the argument "*--cut_ga*" to impose family-specific significance thresholds, and the InterPro database with interproscan.sh in InterProScan 5.18-57.0 (*93*) using the arguments "*-dp -hm*". For *mss* and *msrp* gene products, we used NetOGlyc 4.0.0.13 (*94*) to predict possible GalNAc-type O-glycosylation sites via its web server (*http://www.cbs.dtu.dk/services/NetOGlyc*); we also used PredGPI (*95*) to predict possible GPI anchors, via its web server (*http://gpcr.biocomp.unibo.it/predgpi/index.htm*).

We predicted Gene Ontology (GO) terms (*96*) for gene functions with command-line Blast2GO v1.3.3 (*97*). To enable Blast2GO, we generated XML-formatted results of BlastP (from BLAST+ 2.2.31) of *C. nigoni* proteins against key metazoan proteomes from RefSeq (*C. elegans*, *Danio rerio*, *Drosophila melanogaster*, *Homo sapiens*, and *Mus musculus*); BlastP was run with the arguments "*-evalue 1e-5 -max_target_seqs 20 -outfmt 5 -seg yes -show_gis*". These proteomes were selected because they are the most extensively annotated metazoan proteomes in the Gene Ontology database (*http://www.geneontology.org*), and thus most likely to yield informative cross-annotations. We also provided Blast2GO the following GO data files: *http://archive.geneontology.org/full/2017-01-01/go_monthly-assocdb-data.gz* (dated 7 Jan 2017), *ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene_info.gz* (16 Feb 2017), *ftp://ftp.ncbi.nlm.nih.gov/*

gene/DATA/gene2accession.gz (16 Feb 2017), *ftp://ftp.pir.georgetown.edu/databases/idmapping/ idmapping.tb.gz* (18 Jan 2017).

Orthologies between genes were computed with Orthofinder 0.7.1 (*https://github.com/ davidemms/OrthoFinder/releases/download/0.7.1/OrthoFinder-0.7.1.tar.gz*) (*24*), aided by BLAST+ 2.2.31, mcl 11335, and SciPy 0.9.0. For analyses of motifs, GO terms, and orthology, we used only the longest-predicted isoform for each gene in a genome as a representative of that gene.

Comparisons of ortholog family size and protein length. We compared the membership size for each OrthoFinder group (excluding 1-1 orthologs) from both *C. nigoni* and *C. briggsae* (**Figure 3a**), and calculated protein length distributions from the longest isoforms predicted by AUGUSTUS (**Figure 3b**).

Pfam domain enrichment analysis. We ran custom R scripts on our Pfam domain analyses of *C. nigoni* (Supplementary Dataset S1) and *C. briggsae* (Supplementary Dataset S3) to count the genes associated with each Pfam domain in each species, and then to use Fisher's exact test to determine which Pfam domains in *C. nigoni* were significantly overrepresented by comparison with their abundance in *C. briggsae*. We likewise ran custom R scripts on our Pfam domain analyses of *C. elegans*, *C. remanei*, and *C. briggsae* (all in Supplementary Dataset S4) to count the genes associated with each Pfam domain for each species, and to use Fisher's test in pairwise comparisons to find the Pfam domains overrepresented in outcrossing species as compared to the selfing species.

Gene ontology (GO) enrichment analysis. C. nigoni genes were classified by OrthoFinder as either having or lacking *C. briggsae* homologs. To find GO terms enriched in the *C. nigoni* gene set lacking *C. briggsae* homologs, we used the R package topGO (*98*), and calculated the p-values of GO terms with the default method 'weight01'.

Aligning genomes and measuring indel sizes, frequencies, inversions, repeats, structural features and gene densities. We used *nucmer* from MUMmer 3.23 with the arguments "*-mincluster 35 -masgap1000 -minmatch 10*" to align the six *C. nigoni* pseudochromosomes to the *C. briggsae* chromosomes, which were themselves generated from a high-resolution recombination map (*99*). We used *dnadiff* from MUMmer 3.23 to generate coordinates for the differences between the two genomes using default parameters. *nucmer* and *dnadiff* generated alignment statistics, gaps, duplications, inversions, etc. We calculated the frequency of indels of different sizes and used the computing language R to generate bar graphs. In addition, we calculated the probability of sequence deletions using a 200-kb non-overlapping sliding window. The probability equals the size of deletions divided by the 200-kb window length. We calculated repeat frequencies based on our analyses of the *C. nigoni* and *C. briggsae* genomes with RepeatModeler and RepeatMasker; inversion frequencies from *dnadiff*; and gene density (using, for each gene, the longest isoforms) from our AUGUSTUS predictions. We plotted these features circularly (**Figure 1**), using the same 200-kb window, with Circos-0.69-3 (*http://circos.ca/software/download/circos*) (*100*). The genome composition for CDS, intron, and intergenic sequences were extracted from the GFF3 file of our AUGUSTUS gene predictions. We used bedtools v2.26.0 (*http://bedtools.readthedocs.io*) (*101*) to intersect the coordinates of genomic features to the coordinates of differences between the genomes, and diagrammed the overlap (**Figure 2**).

Determining and comparing summed exon and intron lengths of orthologous genes. We selected strictly orthologous protein-coding genes for analysis that had the following properties. First, they were identified in *C. nigoni* as being either autosomal or X-chromosomal; we ignored genes whose chromosome in *C. nigoni* was unassigned. Second, they were predicted to have a single, strict ortholog by our OrthoFinder analysis in each of the following proteomes: *C. nigoni*; *C. briggsae* (the official gene prediction set, from WormBase release WS254); *C. briggsae* (our prediction set, "briggsae-alt"); *C. remanei* (the official gene prediction set by Fierst et al., from WS254); *C. remanei*, our prediction set ("remanei_alt"); and *C. elegans* (official gene prediction set from WS254). Imposing this orthology requirement allowed us to make direct comparisons of summed exon and intron size for these genes for any pair of species and predictions. Using our own gene predictions for different species, and comparing these (for *C. briggsae* and *C. remanei*) to earlier gene prediction sets by other investigators, allowed to control for differences that might be due to different gene prediction methods. These criteria yielded 6,404 sets of orthologous autosomal genes and 1,394 sets of orthologous X-chromosomal genes for exon/intron size comparisons.

For each gene in each set, we identified the isoform that encoded the largest protein product, and used that isoform for all exon/intron sums and comparisons. To enforce uniform predictions of exons, we extracted the coordinates of protein-coding sequences (CDS), which were available for all six gene sets and which did not depend on further predictions of non-coding exons (which were available for only some gene sets, and which were likely to be less reliable). For the largest isoform of each gene, we used custom Perl scripts to extract its CDS coordinates from its genome annotation file (in GFF3 format), to infer the coordinates of intervening introns from these CDS coordinates, and to sum up the sizes of both exons and introns. For mass comparisons of exon and intron sets (**table S3**, subtables "exon-intron sums, autosomes" and "exon-intron sums, Xchr") we used the *gsl_fit_linear* and *gsl_stats_correlation* functions of the Math::GSL module in Perl to infer slopes, y-intersects, and Pearson r^2 correlations for sizes of exons and introns. GFF3 annotation files for *C. briggsae*, *C. remanei*, and *C. elegans* were downloaded from WormBase release WS254 (for gene predictions by others; *ftp://ftp.wormbase.org/pub/wormbase/releases/WS254/species/c_briggsae/PRJNA10731/c_briggsae.PRJNA10731.WS254.annotations.gff3.gz*, *ftp://ftp.wormbase.org/pub/wormbase/releases/WS254/species/c_remanei/PRJNA248909/c_remanei.PRJNA248909.WS254.annotations.gff3.gz*, and *ftp://ftp.wormbase.org/pub/wormbase/releases/WS254/species/c_elegans/PRJNA13758/c_elegans.PRJNA13758.WS254.annotations.gff3.gz*) or taken from our AUGUSTUS analyses (for gene predictions by us). Detailed results are provided in **table S3**.

Sex-biased expression analysis. We defined sex-biased expression, or the lack of it, as follows. Having determined gene expression levels by RNA-seq using biological triplicates of males and females, we defined a gene's as male-biased if it was ≥2-fold greater in males than in females (Male.v.Fem.logFC ≥ 1) with a false discovery rate (FDR; corrected for multiple hypothesis testing) of ≤ 0.001. We similarly defined a gene's expression as female-biased if it was ≥2-fold greater in females than in males (Male.v.Fem.logFC ≤ -1) with an FDR ≤ 0.001. All other genes with intermediate expression ratios (-1 < Male.v.Fem.logFC < 1) and for which these intermediate ratios were determined with high significance (FDR ≤ 0.001) were classified as unbiased. These criteria exclude genes whose male-female expression ratios were determined with weaker statistical significance, or for which edgeR assigned no significance at all. We chose these criteria (≥2-fold difference in expression, FDR ≤ 0.001) because they are likely to come closest to accuracy for small numbers of biological replicates (*88*). To detect over- or under-representation of sex-biased expression in conserved versus non-conserved *C. nigoni* genes, we used our OrthoFinder groups to distinguish *C. nigoni* genes with or without *C. briggsae* homologs, compared the frequencies of sex-biased expression within both groups, and determined their significance with Fisher's exact test.

Identification of mss *and* msrp *genes in* Caenorhabditis. Because of their limited sequence conservation, *mss* and *msrp* gene products were difficult to identify reliably with BlastP or psi-BLAST. However, we found that we could identify phylogenetically coherent sets of *mss* and *msrp* genes by iterative sequence analysis. We began with the first *mss* protein sequence that we had identified, Cre-MSS-1/FL81_17790-RA, and carried out psi-BLAST with it against *Caenorhabditis* proteomes that had been selected for proteins ≤200 residues in length. psi-BLAST was run with the arguments "*-evalue 1e-03 -seg no -num_iterations 20 -inclusion_ethresh 1e-03*". By constraining the psi-BLAST search to small proteins, we found that we could increase our ability to detect *mss* homologs while lowering the background rate of hits to large proteins. This highly specialized psi-BLAST search converged on a set of 13 *mss* and *msrp* homologs. We repeated psi-BLAST with each of these homologs; their searches all converged, to yield a cumulative set of 43 *mss* and *msrp* homologs. Having acquired an extensive and carefully defined initial sequence set, we performed several rounds of profile-based sequence searches of *Caenorhabditis* and other nematode proteomes. In each of these rounds, we aligned homologs with MAFFT v7.305b (*102*), used their alignment to construct an HMM with hmmbuild and hmmpress in HMMER 3.1b2, and scanned nematode proteomes with this HMM via hmmsearch from HMMER. MAFFT alignments were run in slow/sensitive mode (L-INS-i), with the arguments "*--maxiterate 1000 --localpair*"; hmmsearch was run with the argument "*--domE 0.5*", to allow weak hits to be examined. For each of these rounds, we selected new hits both for the statistical strength of their alignment to the HMM (e.g., E-values of $\leq 10^{-6}$) and for their general structural characteristics (marginal hits to small domains of large proteins were rejected). Although we did not use sex-biased gene expression data to select genes, we were able to use sex-biased gene expression data for *Caenorhabditis* as a guide to whether the iterative search was detecting credible new hits. We observed that one protein sequence from *C. remanei*, FL81_26171-RA, was identical to FL81_17791 in its first 147 residues, and lacked the last 10 residues of *mss* consensus sequence in its C-terminus; we therefore deleted

FL81_26171-RA as a possible misprediction. Having identified a set of 68 nonredundant *mss* and *msrp* homologs in *C. nigoni*'s closer relatives (*C. briggsae*, *C. sinica*, *C. remanei*, *C. elegans*, and *C. nigoni* itself), we constructed both HMMs for *mss* and *msrp* families and a global 68-sequence HMM, which we used to search the proteomes of three outgroup species, *C. japonica*, *C. afra*, and *C.* sp. 34; genes which gave positive results for both the family-specific HMMs and the global HMM were further examined as possible *mss*/*msrp* homologs. We concluded this search with 81 *mss* and *msrp* homologs from *Caenorhabditis* (**dataset S5**).

Protein sequence alignment and phylogenetic tree for mss and msrp. We used MAFFT to align the protein sequences of all *mss* and *msrp* genes. From the remaining *mss*/*msrp* alignment, we filtered weakly aligned residues having >50% gaps with trimAl v1.4.rev15 (*http://github.com/scapella/trimal*) (*103*), using the argument '*-gt 0.5*'. On the filtered alignment, we used FastTree 2.1.9 (*http://www.microbesonline.org/fasttree*) (*104*) to compute maximum-likelihood trees with posterior probabilities on the tree nodes, with the arguments '*-pseudo -wag*'. We used JalView 2.9.0b2 (*http://www.jalview.org*) (*105*) with ClustalW residue colors to draw the *mss* alignment in **Figure 3d**, and FigTree v1.4.3 (*http://tree.bio.ed.ac.uk/software/figtree*) to draw the *mss*/*msrp* phylogeny.

Predicting exons for the pseudogene Cbr-mss-3-ps *in* C. briggsae. We used exonerate 2.2.0 and the protein sequence of MSS-3 from *C. nigoni* to identify actual or potential exons from both the *Cni-mss-3* genomic locus and the *Cbr-mss-3-ps* genomic locus. In both cases, we ran exonerate with the arguments "*-E -m protein2genome:bestfit --useaatla FALSE -Q protein -q [Cni-mss-3/Cnig_chr_III.g11661.t1 protein sequence] -T dna -t [genomic locus DNA sequence]*". To provide input sequences to exonerate, we used *extractseq* from EMBOSS to extract the predicted region (based on BlastN with the coding sequence of *Cni-mss-3*) and 500 nt of flanking DNA from the appropriate *C. nigoni* or *C. briggsae* genome scaffold. We also ran AUGUSTUS with *C. briggsae*-specific parameters on each region to confirm that there was an observable protein-coding gene in the *Cni-mss-3* region, but no observable protein-coding gene in the *Cbr-mss-3-ps* region (i.e., we checked to make sure that there were no local compensatory mutations that might restore coding capacity to *Cbr-mss-3-ps*).

Identifying the pseudogene Cbr-mss-3-ps *in wild isolates of* C. briggsae. Cutter and coworkers have identified and sequenced genomic DNA from wild isolates of *C. briggsae* worldwide (*27*). Although there were no preexisting genome assemblies for these isolates, their raw genomic sequence data were available. We therefore downloaded genomic sequence reads (generally, though not universally, paired-end) for the following 12 wild isolates (NCBI SRA accession numbers given in parentheses): Hubei_VX0034 (SRR1793007); Kerala_JU1341 (SRR1792996 and SRR1793000); Kerala_JU1348 (SRR1793004); Nairobi_ED3101 (SRR1793002); Quebec_QR24 (SRR1793005); Quebec_QR25 (SRR1793006); Taiwan_NIC19 (SRR1793010); Taiwan_NIC20 (SRR1793012); Temperate_EG4181 (SRR1792978); Temperate_JU516 (SRR1792992); Tropical_JU1399 (SRR1792934); and Tropical_QX1410 (SRR1792974). In particular, we chose Tropical_QX1410 because it is phylogenetically almost identical to the reference wild-type strain AF16, and chose the Kerala isolates because they are currently the most divergent outgroups known in *C. briggsae* (*27, 106*). For each wild isolate's reads, we used ABySS (abyss-pe 2.0.2) to assemble a draft genome, with the arguments "*np=8 j=8 k=51*". The resulting draft genome assemblies for all 12 wild isolates are provided in **dataset S3**. From 11 of the 12 draft assemblies (including one for Kerala), we identified a full-length copy of the *Cbr-mss-3-ps* locus with BlastN, and extracted its first exon with *extractseq* from EMBOSS. We used MAFFT (with arguments "*--maxiterate 1000 --localpair*") to align the first exons for all wild-isolate alleles of *Cbr-mss-3-ps* with *Cni-mss-3*, and to confirm that two mutations inactivating *Cbr-mss-3-ps* in AF16 are also present in all wild isolates.

Quantitative RT-PCR. For *Cre-mss-2* five replicate populations of staged animals were washed off NGM plates in M9 buffer, or individually picked in M9 buffer for male and female preparations. After three washes in RNAse-free water, samples were resuspended in 50 µl of RNAse-free water. 250 µl TRI-Reagent (Molecular Research Center) was added and the samples were frozen at -80˚C. The samples were thawed, pelleted and lysed using a plastic pestle. RNA was purified using manufacturer's instructions. Samples were treated with DNase I (New England Biolabs), phenol/chloroform extracted, isopropanol precipitated, and resuspended in RNase- free water. cDNA was synthesized using 1 µg of total RNA using Superscript III (Invitrogen) in 50 µl according to manufacturer's instructions. 2 µl cDNA were used as template with the Light Cycler 480 SYBR Green I kit (Roche) according to manufacturer's instructions. Primers spanning exon-exon junctions were designed so that the amplicons sizes were between 161 and 220 bp. Primers were CGT057 (5'-GGA TCT TCT GGG GCT TTC GG-3') and CGT058 (5'-GGA TTT CCG ACTCCA CCA TCT G-3'). Control reactions with no template and multiple non-sex-biased transcripts for each primer pair were performed, and all reactions for a gene were run simultaneously on a single 96-well plate on a Roche Light Cycler 480 machine using manufacturer's software. Data were analyzed using the program LinRegPCR (*107, 108*). For the qualitative experiment in **figure S9**, cDNA was synthesized as above, and PCR was carried out using PrimeSTAR Max DNA Polymerase (TAKARA Bio Inc) using primers DY112 (5'-CTG GTC CAT TCA CAG TCA CAG C-3') and DY113 (5'-ATG ATG GTG GTG CTG GAG GC-3') for detection of *mss-1*.

In situ hybridization. Gonads and intestines were dissected from adult *C. remanei,* fixed in a glutaraldehyde-based fixative, and incubated with digoxygenin-labelled antisense ssDNA probes as previously described (*109, 110*). The templates for production of antisense and sense *Cr-mss-1* ssDNA probes via asymmetric PCR was amplified from JU1422 genomic DNA using the following oligonucleotide primers (IDT), where underlined sequence represents the phage T7 promoter. RLM001 (for sense probe PCR and IVT): 5'-TAA TAC GAC TCA CTA TAG GGA GAG CAT TGT TGG CCA CCG-3'. RLM002 (for antisense template PCR): 5'-GCA TTG TTG GCC ACC G-3'. RLM003 (for antisense probe PCR and IVT): 5'-TAA TAC GAC TCA CTA TAG GGA GAT CCT TCG GCT GGT GCT TCT GGT-3'. RLM004 (for sense template PCR): 5'-TCC TTC GGC TGG TGC TTC TGG T-3'.

Immunohistochemistry and microscopy. After adult *C. remanei* males and mated females were dissected, testis and activated sperm were isolated on ColorFrost Plus slides using a modification of standard protocols (*111*). In brief, after dissection, a cover slip was placed over the sample and then the slide was placed in dry ice for 30 min to freeze and crack. Dissected gonads or sperm were fixed in 4% formaldehyde for 10 min, followed by a 30 min post-fix in 100% methanol at -20°C. After multiple washing steps (PBS containing 0.1% Tween) and blocking (PBS containing 0.1% Tween + 0.75% BSA), incubations with primary antibody were carried out at room temperature in a humid chamber overnight with a 1:100 dilution of Anti-HA-Peroxidaxe-3F10 (Roche). Secondary antibody incubations were carried out for 2 hours with 1:100 Alexa Fluo 555 anti-rat IgG (Invitrogen). Hoechst (Sigma-Aldrich) at 100 μg/ml was used for staining DNA for 15 min. Slides were prepared with Vectashield mounting media. Images were acquired on Leica SP5 X confocal microscope and processed with Zen Lite software (Zeiss). For anti-HA immunoblots, worms were homogenized and boiled in Laemmli sample buffer under standard reducing conditions (*112*). The resulting proteins were fractionated via 15% SDS-PAGE gels and transferred to nitrocellulose. Filters were incubated sequentially with rat anti-HA monoclonal antibody (Santa Cruz), goat anti-rat horseradish peroxidase (HRP)-conjugated secondary (Thermo Fisher), and HRP luminescence detection reagents (Pierce) according to manufacturer's instructions.

Generation of mss knock-out in C. remanei using CRISPR. Cas9 protein containing an NLS (PNA Bio Inc) was reconstituted by dissolving in 40 μl of nuclease-free water, creating a stock at 1250 ng/μl and stored in small aliquots at −80°C. The sgRNAs were transcribed *in vitro* using the Ambion Megascript SP6 kit. We followed a published protocol for steps including annealing oligos, transcribing sgRNA and cleanup of sgRNA, except we used Taq polymerase for fill-in instead of T4 DNA polymerase (*113*). Multiple *in vitro* transcription reactions were set up in order to get very concentrated sgRNA. Prior to the microinjection, *in vitro* DNA cleavage assay was carried out to confirm the effectiveness of the gRNA and Cas9 protein for cutting the target DNA template. In the final injection mixture, the concentration of Cas9 protein was ~800 ng/μl while the concentration of each of gRNAs was ~670 ng/μl. The injection mixture was incubated at 37°C for 10 minutes before being loaded into needles for injection. Young gravid females were injected, put individually on NGM plates, and allowed to lay embryos for 32 hr. After the F1 progeny become adults and mated, the mated females were divided into five worms/plate. After the F2s laid embryos, the F1 females on each plate were picked and genotyped using single-worm PCR to detect any edit. Positives were identified using primers flanking the *mss* genes or the HA epitope tag that generates amplicons of different sizes between edit and non-edit. Once the edit was confirmed by sequencing, the L4s from corresponding plate(s) were used to set up crossings to isolate homozygous mutants. *mss* knock-out mutants were independently generated in both SB146 and EM464 strains. The SB146-derived *mss* knockout strain, CP157, was rendered homozygous for its *nmDf1* deletion by multiple generations of backcrossing. However we were not able to render the EM464 knockout allele, *nmDf2,* homozygous, presumably due to a lethal or harmful allele linked to the *mss* locus.

Therefore, only heterozygous EM464 mutants were generated. To compensate for the effects of inbreeding depression, interstrain progeny were generated by crossing heterozygous EM464 mutants and homozygous SB146 mutants.

HA knock-in. The hemaglutinin (HA) epitope tag was inserted using CRISPR/Cas9 through homologous recombination. The nine amino acid HA epitope was placed between *C. remanei* (EM464*)* MSS-1 residues 22 and 23, one residue downstream of the predicted mature N-terminus after signal peptide cleavage. We designed sgRNA that was complementary to sequence downstream of the signal peptide, so that even a processed peptide would retain the HA epitope at its N-terminus. Edited line CP158 (*Cre-mss-1(nmIs9)*) was identified by PCR (primer DY70 5'-ACG ACG TTC CAG ACT ACG CC-3' and DY41 5'-TGA GTG TCT TTG GGT GCG TT-3') of offspring of injected mothers, after they had participated in pair-wise full-sib mating and laid abundant progeny (see detailed screening method above in the *mss* knock-out section). In the microinjection mixture, the concentration for Cas9 was 900 ng/µl; sgRNA 900 ng/µl; donor (repair) template oligonucleotide, 1000 nM. Repair template: DY68 donor template (for homologous recombination after Cas9 cutting) was a PAGE-purified Ultramer custom oligonucleotide (IDT). DY68 is 89 nt in total. 27 nt (lower case) indicates the HA epitope tag sequence. 31 nt (upper case) is homologous to *C. remanei mss-1* sequence on each side of the Cas9 cleavage site: 5'-CGC ATT GTT GGC CAC CGT CGC TCG TGG AGC Tta ccc ata cga cgt tcc aga cta cgc cGA CGG TGA TAA CGT AGA AGC AGG AGA TGC AC-3'. sgRNA EH75 (*C. remanei mss-1* gene-specific oligo used for generating sgRNA): 5'-att tag gtg aca cta taG ATG CAC AAC TAC CATCAG Ggt ttt aga gct aga aata gca-3'. Lower case DNA on the left is the SP6 promoter; upper case sequence is the gene-specific sequence for making sgRNA; lower case DNA on the right is homologous sequence are anneals with the Constant Oligo during the template fill-in reaction. Constant Oligo: 5'-AAA AGC ACC GAC TCG GTG CCA CTT TTT CAA GTT GAT AAC GGA CTA GCC TTA TTT TAA CTT GCT ATT TCT AGC TCT AAA AC-3'.

C. remanei MSS-1 immunoblots and deglycosylation assay. ~500 *C. remanei* (EM464) adults were washed off plates with M9 buffer, washed twice with more M9, washed once in lysis buffer (150 mM NaCl, 20 mM Tris, pH 8.0, 1 mM EDTA, 0.1% NP40) without protease inhibitors, and then washed twice (and left in) 100 µl lysis buffer with protease inhibitors (Complete EDTA-free, Roche). Worms were homogenized by grinding with a small plastic pestle. 50 µl of the lysate was used as an untreated control. The other 50 µl was used for deglycosylation treatment with a protein glycosidase cocktail (New England Biolabs P6044S, Mix II, used with reducing Buffer 2). We varied the concentration of lysate by having either 5 µl or 3 µl in the final deglycosylation reactions. The mixture was incubated at 75°C for 10 mins, cooled to 15°C, and then the protein deglycosylation mix II was added for a final reaction volume of 50 µl. Digestions were incubated at room temperature for 30 min before incubation overnight (~13 hr) at 37°C. For both treated and untreated control, equal volumes of 2x (Laemmli + BME) solution were added before loading for SDS-polyacrylamide gels, with pre-stained PageRuler Plus (Thermo Scientific) or ExcelBand 3 color Extra Range (SMOBIO). Gel electrophoresis was conducted for 1.5 hours at 150 V, and transfer was conducted overnight at 4°C with 30 V. The membrane was blocked with a 5% BSA+PBST solution and washed with PBST. The primary antibody (rat monoclonal αHA 3F10; Roche) was diluted 300-fold and incubated for 90 mins at room temperature before washing. The secondary antibody (goat anti-rat HRP conjugate; Santa Cruz) was diluted 600-fold and incubated for 1 hour at room temperature. Antibody was detected with the SuperSignal West Pico Chemiluminescent Substrate (Thermo Scientific).

Generation of stable transgenic C. briggsae strains expressing MSS *from C. nigoni*. *mss-1* and *mss-2* from *C. nigoni* were amplified by PrimeSTAR Max DNA Polymerase (TAKARA Bio Inc) using primers flanking the gene family (also including 1,500 nt 5' from the start codon of *mss-1*). *Cnig-mss-1/2* was inserted into a reporter construct pZZ0031 (gift from the laboratory of Zhongying Zhao, Hong Kong Baptist Univ.). The construct has the *Cbr-myo-2* promoter fused with GFP, and *unc-119* rescuing sequence as selection marker for bombardment. *Cbr-unc-119* mutants were bombarded and stable transgenic lines CP161(*nmIs7*) and CP162(*nmIs8*) expressing strong GFP expression were isolated as described (*53, 114*). The primer pair used for amplifying *mss-1* and *mss-2* was as follows. Forward (DY107): 5'-TAG ACT GGG CCC CGA ATT TCC CTG ACG AAT GCT CC-3'. Reverse (DY108): 5'-TAC ATT GGG CCC TTG CGG ACA GAG CCACAG AG-3'. The final insert size of *mss-1* and *mss-2* was 4.5 kb. PrimeSTAR Max DNA Polymerase (TAKARA Bio Inc) was used with a 15-second annealing at 55°C, and a 5-minute extension at 72°C in the PCR. Both primers used above have 6-nt overhangs and GGGCCC ApaI restriction sites added to the 5' end of their *C. nigoni* sequences. The plasmid pZZ0031 was digested with ApaI (New England Biolabs; NEB), followed by dephosphorylation using NEB shrimp alkaline phosphatase (rSAP). ApaI-digested primers were then ligated to the vector using T4 ligase. Transformation was performed on NEB 5-alpha competent *E. coli* and plasmids from colonies were isolated and sequenced to confirm the correct insertion of *mss-1* and *mss-2.*

Sperm competition assay in C. remanei. To create competitor males, heterozygous EM464 *mss(nmDf2)* knockout mutants were crossed to SB146 homozygous *mss(nmDf1)* knockout mutants to create interstrain F1 hybrid progeny. The progeny male genotypes were *nmDf1/+* and *nmDf1/nmDf2*. The males were 24 hr post-L4 virgins. In the MSS- defense scenario, *nmDf1/nmDf2* males were first allowed to mate with 24 hr post-L4 virgin adult females for 4 hours, after which the females were moved to another plate and allowed to mate with wild-type males for 4 hours. In the MSS- offense scenario, the order between the first male and second male were switched. As a control, *nmDf1/+* sibling males were subjected to the same defensive and offensive mating assays. Each mating was set up with one male and one female and allowed to mate for four hours on 3.5-cm NGM plates seeded with 0.5 cm *E. coli* to maximize the chance of contact and mating. Copulation with both males was closely observed under a stereomicroscope to confirm successful transfer of sperm. After mating, each male was genotyped by single-worm PCR. After a total of eight hours of mating, each female was picked into individual plates and allowed to lay embryos for 24 hours. Progeny laid during this period were scored by PCR to score paternity. 24-30 progeny from each plate were scored.

Competition of male sperm against male sperm in C. briggsae. To test CP161(*nmIs7*) MSS+ sperm's defense ability, 20 virgin young adult males [*Cni-mss-1, mss-2, cbr-myo*-2::GFP] were placed with 15 wild-type AF16 hermaphrodites for four hours, after which only hermaphrodites with visible embryos in their uteri were transferred to a new plate with 20 RW20025 (*Cbr-unc-119*, *stIs*20025[*Cbr-HIS-72*::mCherry]) control males for 4 hours. Individual hermaphrodites were then transferred to new plates, and allowed to lay embryos for 18 hr. Individual hermaphrodites were transferred to fresh plates and allowed to lay embryos for another 24-hr period. After two days the progeny were scored under a fluorescent microscope. Progeny sired by CP161(*nmIs7*) MSS+ had green fluorescent protein (GFP) expression in the pharynx, whereas those sired by control male expressed red fluorescent protein (RFP). Since both the GFP and RFP were dominant markers, non-fluorescent progeny were sired by hermaphrodite self sperm. For CP161(*nmIs7*) MSS+ sperm offense, the same experiment was carried out but the hermaphrodites were first mated to RW20025 males, then CP161(*nmIs7*) MSS+ males. Cases in which the second males' mating completely failed, as indicated by 100% cross progeny from the first males, were excluded.

Competition of male sperm vs. hermaphrodite self-sperm. To test the effect of MSS for male sperm precedence over hermaphrodite self-sperm, 20 virgin young adult males of each genotype were placed with 15 AF16 hermaphrodites for 4 hr. The control male genotypes were JU936 *[Ce-lip-1::GFP, Ce-myo-2::GFP],* RW20025 (*Cbr-unc-119*, *stIs*20025 [*Cbr-HIS-72*::mCherry]), and CP161(*nmIs7*) MSS+ experimental male (*Cbr-unc-119*, *Cni-mss-1, mss-2, Cbr-myo-2::*GFP). Individual hermaphrodites were transferred to fresh plates and allowed to lay embryos for a 24-hr period. Progeny were scored for GFP or mCherry after reaching adulthood. The ratio of progeny expressing fluorescent protein to dark progeny was calculated.

Experimental evolution assay for male percentages. 20 L4 males and 20 L4 hermaphrodites were set up for crossing for 24 hours, after which 10 males and 10 mated hermaphrodites (50% ratio of males) were transferred to a new plate as a first-generation population. Both CP161(*nmIs7*) MSS+ and wild-type AF16 strains had five replicates. All populations were grown on 6 cm NGM agar Petri plates seeded with *E. coli* strain OP50 and transferred by chunking to new plates every generation (~3 days/generation). Each chunk had approximately 200-300 L1 stage larvae and a very small number of adults. When the majority of larvae reached adulthood, adult males and hermaphrodites were counted and ratios of males were calculated. The adults were allowed to lay embryos, that were allowed to develop into L1 before each transfer. The whole procedure was repeated for a total of 12 generations. Male ratios were scored at the 4th, 8th and 12th generations. The populations were incubated at 25°C during this assay.

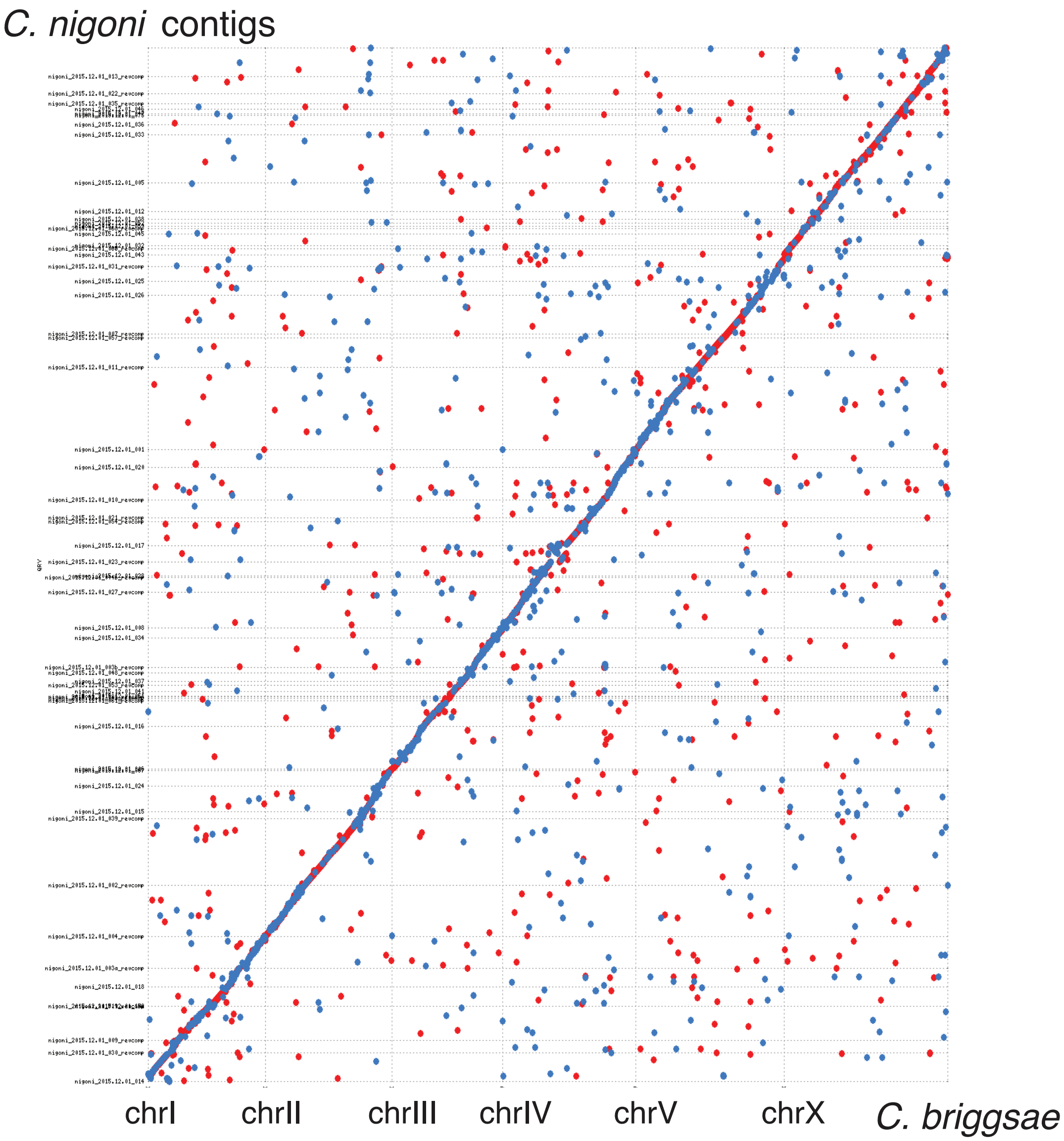


Figure S1 | Tiling of *C. nigoni* contigs against *C. briggsae* chromosomes. MUMMER 3.23 (*75*) was used to align *C. nigoni* contigs (after their full genome assembly, but before their being fused *in silico* into pseudochromosomes) against six chromosomes from the CB4 genome assembly of *C. briggsae* (*99*). Maximal unique matches (MUMs) are depicted as dots: red indicates forward matches; blue indicates reverse matches. Scaffolds from the CB4 assembly that were unassigned to chromosomes were omitted from the alignment.

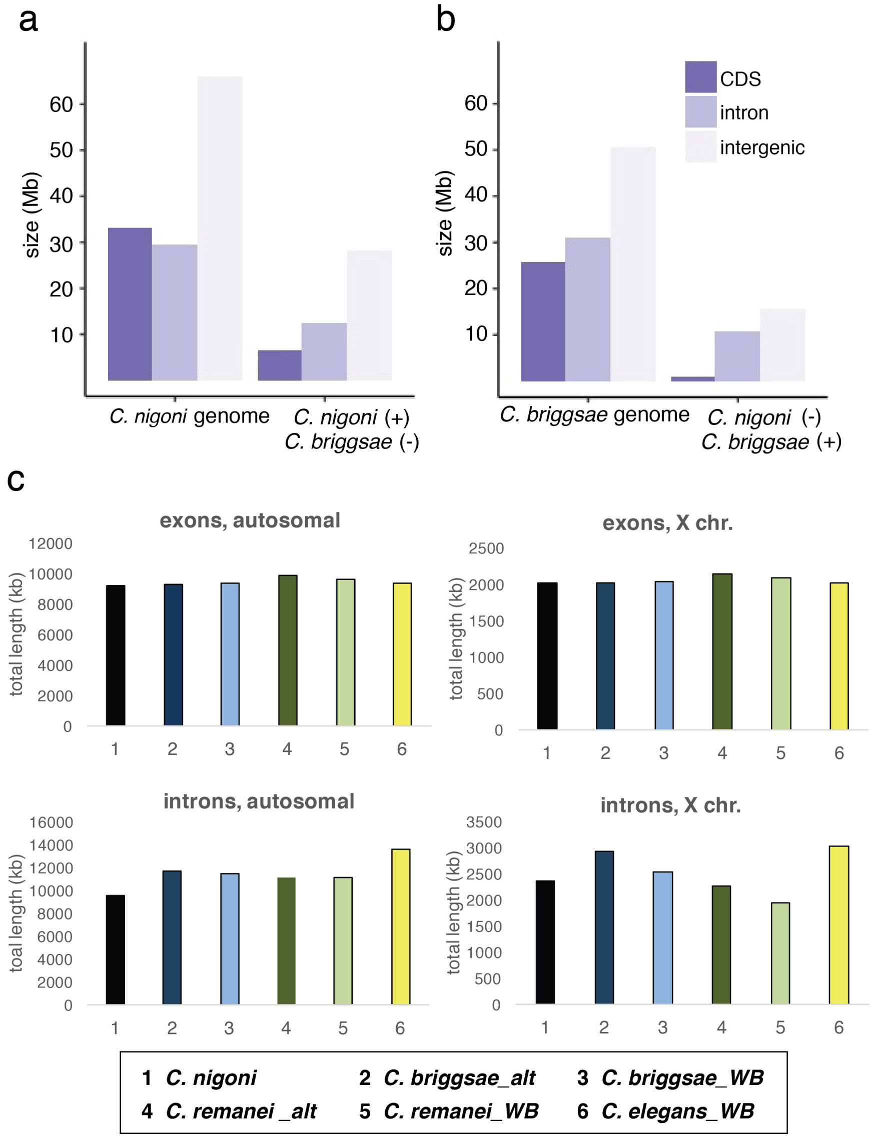


Figure S2 | Coding, intronic, and intergenic proportions of the *C. nigoni* and *C. briggsae* genomes. Values are presented for the *C. nigoni* (**a**) and *C. briggsae* (**b**) genomes overall and for their species-specific sequences. **c**, summary of total exon and intron lengths extracted from predicted protein-coding genes with strict orthology (1:1 between all taxa and gene predictions) for *C. nigoni, C. briggsae, C. remanei,* and *C. elegans*. Gene sets with dark bars (including those with the "alt" suffix; 1, 2, 4) were predicted in this study by consistent methods. Gene sets with pastel bars (3, 5, 6) were taken from WormBase (*www.wormbase.org,* release WS254). Underlying data can be found in **table S3**. X-linked *C. briggsae* (alt) orthologs have significantly more intron content than do those of *C. nigoni* or *C. remanei* (alt). For autosomal introns, *C. briggsae* (alt) and *C. remanei* (alt) orthologs have significantly greater intron content than *C. nigoni* than do their *C. nigoni* counterparts. Intron comparisons used a Wilcoxon rank-sum test with Bonferroni correction, $p < 0.05$.

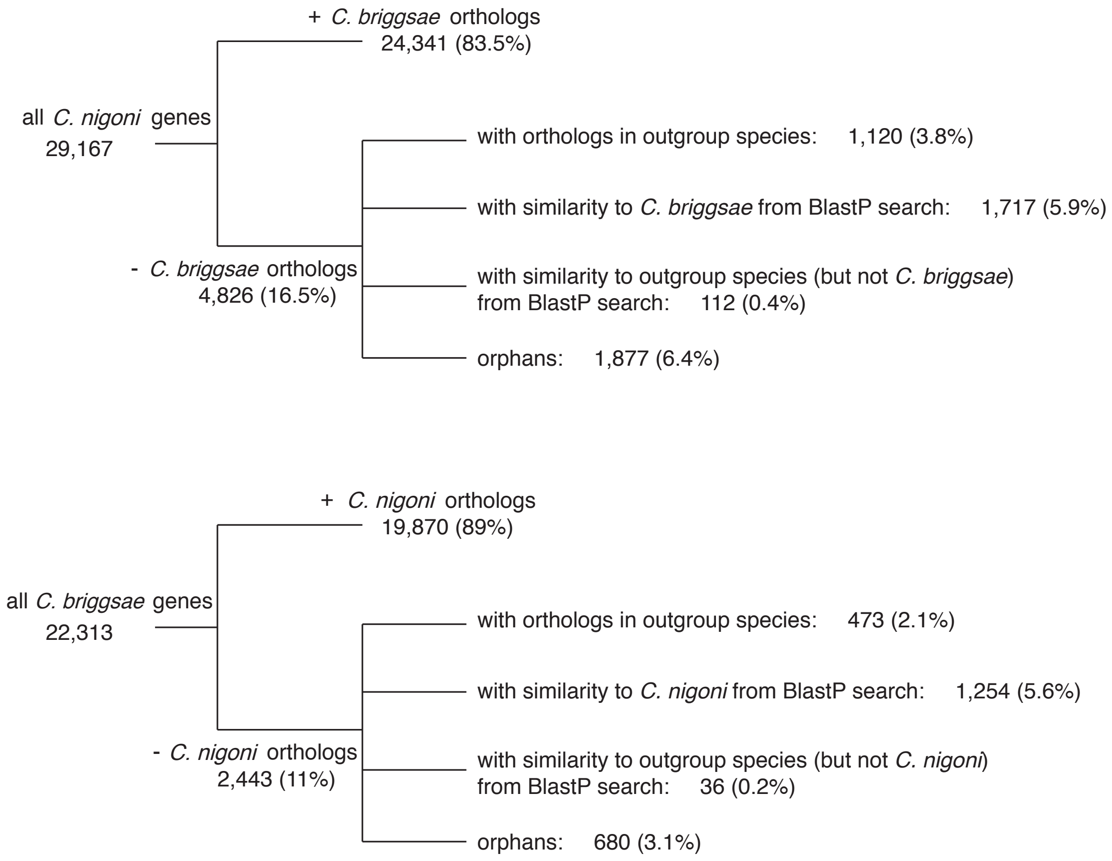
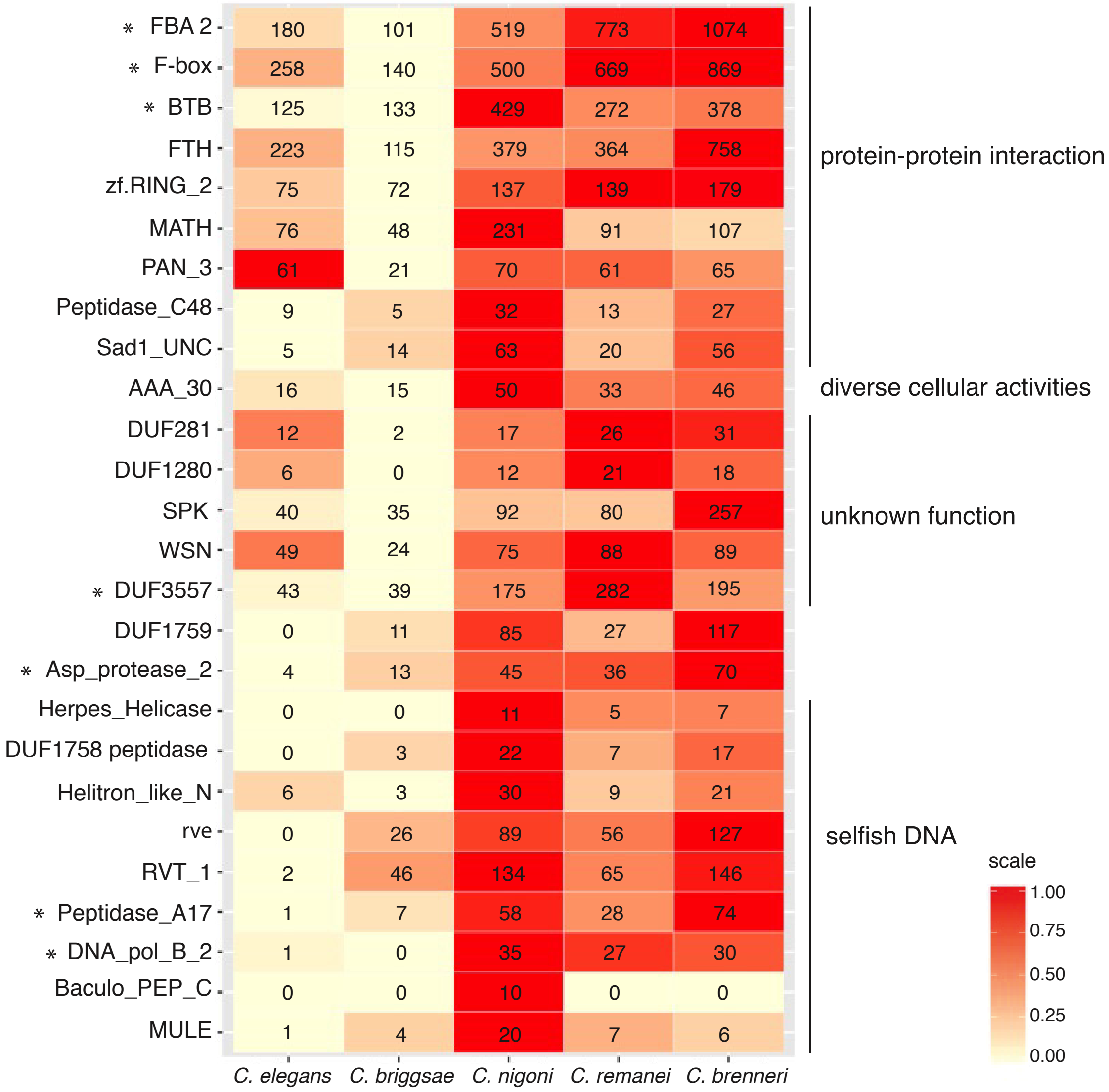

+ C. briggsae orthologs
24,341 (83.5%)
all C. nigoni genes
29,167
with orthologs in outgroup species: 1,120 (3.8%)
with similarity to C. briggsae from BlastP search: 1,717 (5.9%)
- C. briggsae orthologs
4,826 (16.5%)
with similarity to outgroup species (but not C. briggsae)
from BlastP search: 112 (0.4%)
orphans: 1,877 (6.4%)
+ C. nigoni orthologs
19,870 (89%)
all C. briggsae genes
22,313
with orthologs in outgroup species: 473 (2.1%)
with similarity to C. nigoni from BlastP search: 1,254 (5.6%)
- C. nigoni orthologs
2,443 (11%)
with similarity to outgroup species (but not C. nigoni)
from BlastP search: 36 (0.2%)
orphans: 680 (3.1%)

Figure S3 | Gene homology categorization

A detailed classification of gene homologies is shown. For *C. nigoni*: 4,826 genes that lacked *C. briggsae* orthologs (16.5%) may represent both recent losses in *C. briggsae* and recent gains in *C. nigoni*. Within this group, 1,120 genes (3.8% of the total) had OrthoFinder orthologs in outgroup species, and therefore are very likely to represent losses in the *C. briggsae* lineage. Many of the remaining 3,706 genes specific to *C. nigoni* may also represent cases of *C. briggsae* gene loss, including 1,717 (5.9%) that exhibited some similarity to non-orthologous *C. briggsae* genes in BlastP searches ($E \leq 10^{-5}$) (*115*) and 112 (0.4%) that lacked *C. briggsae* BlastP similarity but had BlastP similarity to other *Caenorhabditis*. Finally, 1,877 *C. nigoni* genes (6.4%) lacking *C. briggsae* homologs had neither homologs nor similarities to other species, and were classed as orphans. Conversely, for *C. briggsae*: 2,443 genes (16.5%) lacked *C. nigoni* orthologs; 473 genes (2.1%) had other *Caenorhabditis* orthologs; 1,254 genes (5.6%) had some similarity to *C. nigoni*; 36 genes (0.2%) had similarity to other *Caenorhabditis*; and 680 genes (3.1%) were orphans.



	C. elegans	C. briggsae	C. nigoni	C. remanei	C. brenneri
* FBA 2	180	101	519	773	1074
* F-box	258	140	500	669	869
* BTB	125	133	429	272	378
FTH	223	115	379	364	758
zf.RING_2	75	72	137	139	179
MATH	76	48	231	91	107
PAN_3	61	21	70	61	65
Peptidase_C48	9	5	32	13	27
Sad1_UNC	5	14	63	20	56
AAA_30	16	15	50	33	46
DUF281	12	2	17	26	31
DUF1280	6	0	12	21	18
SPK	40	35	92	80	257
WSN	49	24	75	88	89
* DUF3557	43	39	175	282	195
DUF1759	0	11	85	27	117
* Asp_protease_2	4	13	45	36	70
Herpes_Helicase	0	0	11	5	7
DUF1758 peptidase	0	3	22	7	17
Helitron_like_N	6	3	30	9	21
rve	0	26	89	56	127
RVT_1	2	46	134	65	146
* Peptidase_A17	1	7	58	28	74
* DNA_pol_B_2	1	0	35	27	30
Baculo_PEP_C	0	0	10	0	0
MULE	1	4	20	7	6

Figure S4 | Enriched Pfam domains in *C. nigoni* relative to *C. briggsae.* A heatmap of all 26 Pfam domains encoded by significantly more genes in *C. nigoni* than in *C. briggsae* (*P* < 0.01, Fisher's exact test) is shown, with broad functional categories indicated to the right. By an equivalent criterion (pairwise Fisher's exact tests with p < 0.01, followed by Holm-Bonferroni correction), seven of these Pfam domains are encoded by significantly more genes in all three of the outcrossing species (*C. nigoni*, *C. remanei*, and *C. brenneri*) than in both of the selfing species (*C. briggsae* and *C. elegans*); these seven domains are marked with asterisks. For each domain, the numbers of genes encoding it in a given species are indicated in the boxes. For the scale color of the heatmap, each domain had its percentage of genes in a given genome divided by the total gene number in that genome; the species with the highest fraction of genes for that domain was then assigned a scale value of 1, and the fractions in other species were scaled down in order to draw the heatmap.

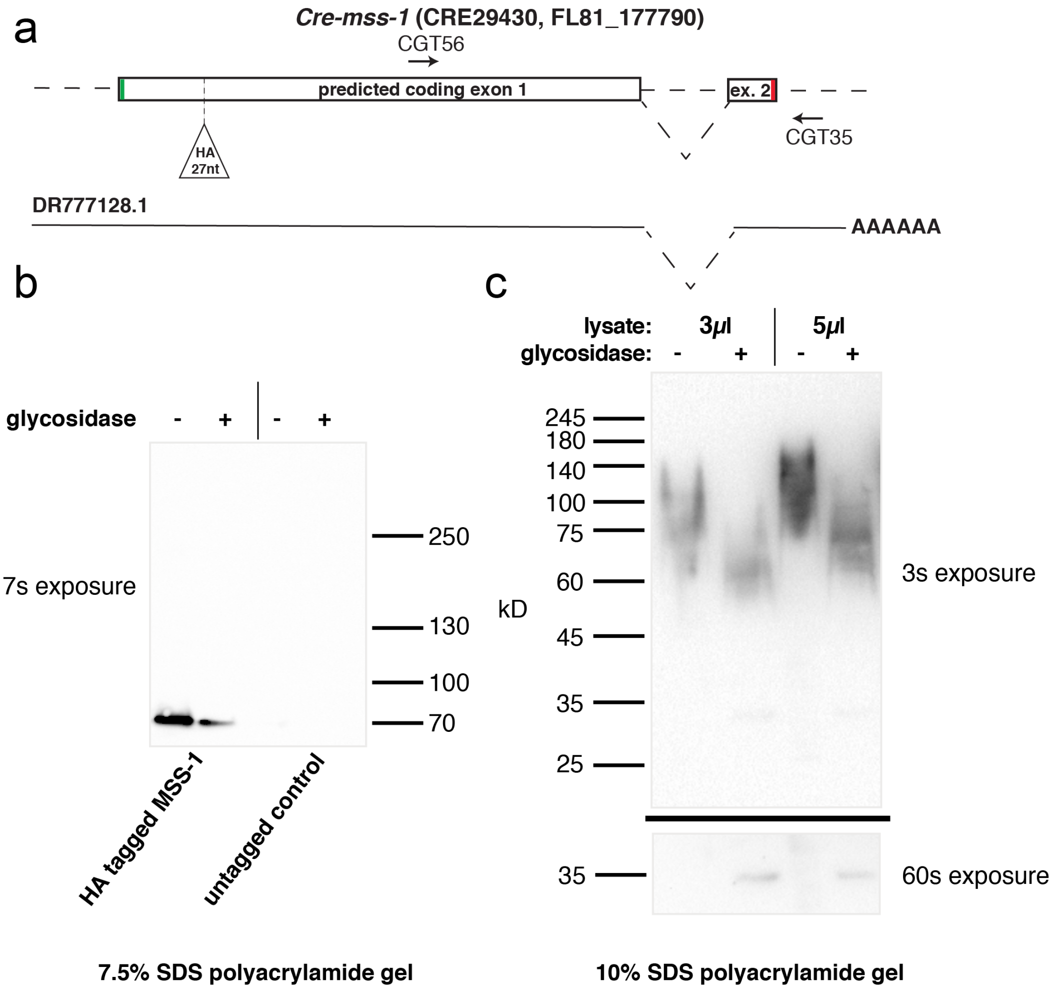


Figure S5 | Cre-MSS-1 is post-translationally glycosylated. a, Gene model. Two independent predictions of protein-coding genes for the genome assemblies of *C. remanei* are shown. CRE29430 is from the Washington University Genome Center assembly of *C. remanei* strain PB4641 (available at *http://www.wormbase.org*). FL81_17790 is from the assembly of *C. remanei* strain PX356 by Fierst et al. (*10*). Both of these are derivatives of the strain EM464, which was first described as "*C. vulgaris*" (*49*) and later found to be synonymous with *C. remanei* (*116*). These gene predictions are 100% identical. Predicted start (green) and stop (red) codons are indicated. The intron location and part of the 3' UTR was confirmed by RT-PCR and sequencing using the primers CGT56 and CGT35, as indicated. The EST DR777128 is derived from the strain SB146, and reveals the extent of the 5' and 3' untranslated regions, including the site of 3' cleavage and polyadenylation. It differs from the EM464-derived models in some codons in the hypervariable domain. The site of insertion of the HA epitope tag just downstream of the signal peptide by CRISPR-Cas9 editing is indicated. **b** and **c,** Western analysis of HA-tagged Cre-MSS-1. Proteins from 500 *C. remanei* homozygous for the HA-tagged MSS-1 allele as well as untagged negative controls were resolved by SDS-PAGE on a

7.5% gel (**b**) and a 10% gel (**c**), then detected with chemiluminescence with different exposures. The image merges the anti-HA fluorescent signal (black) with a reflected light image of pre-stained markers. On the 7.5% gel (**b**), MSS-1 runs near the dye front. On the 10% gel (**c**), deglycosylated (+) and untreated (-) HA-tagged MSS-1 show distinct mobilities. Either 5 µl or 3 µl of lysate was used in the final deglycosylation reactions. Upon longer exposure (**c**, bottom view), a tight band with an apparent mass of 34 kD appears in glycosidase-treated samples.

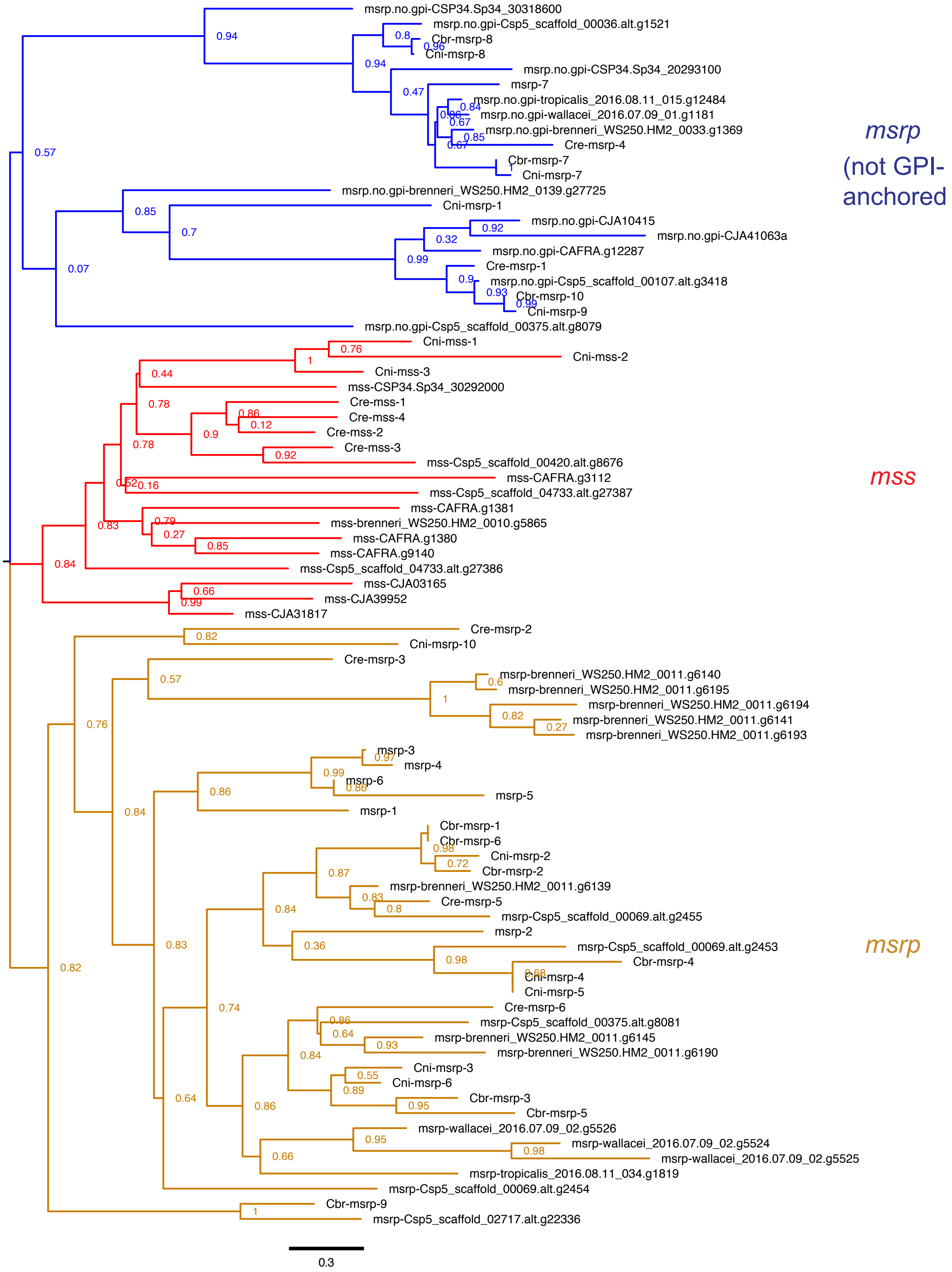

msrp.no.gpi-CSP34.Sp34_30318600
msrp.no.gpi-Csp5_scaffold_00036.alt.g1521
Cbr-msrp-8
Cni-msrp-8
msrp.no.gpi-CSP34.Sp34_20293100
msrp-7
msrp.no.gpi-tropicalis_2016.08.11_015.g12484
msrp.no.gpi-wallacei_2016.07.09_01.g1181
msrp.no.gpi-brenneri_WS250.HM2_0033.g1369
Cre-msrp-4
Cbr-msrp-7
Cni-msrp-7
msrp.no.gpi-brenneri_WS250.HM2_0139.g27725
Cni-msrp-1
msrp.no.gpi-CJA10415
msrp.no.gpi-CJA41063a
msrp.no.gpi-CAFRA.g12287
Cre-msrp-1
msrp.no.gpi-Csp5_scaffold_00107.alt.g3418
Cbr-msrp-10
Cni-msrp-9
msrp.no.gpi-Csp5_scaffold_00375.alt.g8079
Cni-mss-1
Cni-mss-2
Cni-mss-3
mss-CSP34.Sp34_30292000
Cre-mss-1
Cre-mss-4
Cre-mss-2
Cre-mss-3
mss-Csp5_scaffold_00420.alt.g8676
mss-CAFRA.g3112
mss-Csp5_scaffold_04733.alt.g27387
mss-CAFRA.g1381
mss-brenneri_WS250.HM2_0010.g5865
mss-CAFRA.g1380
mss-CAFRA.g9140
mss-Csp5_scaffold_04733.alt.g27386
mss-CJA03165
mss-CJA39952
mss-CJA31817
Cre-msrp-2
Cni-msrp-10
Cre-msrp-3
msrp-brenneri_WS250.HM2_0011.g6140
msrp-brenneri_WS250.HM2_0011.g6195
msrp-brenneri_WS250.HM2_0011.g6194
msrp-brenneri_WS250.HM2_0011.g6141
msrp-brenneri_WS250.HM2_0011.g6193
msrp-3
msrp-4
msrp-6
msrp-5
msrp-1
Cbr-msrp-1
Cbr-msrp-6
Cni-msrp-2
Cbr-msrp-2
msrp-brenneri_WS250.HM2_0011.g6139
Cre-msrp-5
msrp-Csp5_scaffold_00069.alt.g2455
msrp-2
msrp-Csp5_scaffold_00069.alt.g2453
Cbr-msrp-4
Cni-msrp-4
Cni-msrp-5
Cre-msrp-6
msrp-Csp5_scaffold_00375.alt.g8081
msrp-brenneri_WS250.HM2_0011.g6145
msrp-brenneri_WS250.HM2_0011.g6190
Cni-msrp-3
Cni-msrp-6
Cbr-msrp-3
Cbr-msrp-5
msrp-wallacei_2016.07.09_02.g5526
msrp-wallacei_2016.07.09_02.g5524
msrp-wallacei_2016.07.09_02.g5525
msrp-tropicalis_2016.08.11_034.g1819
msrp-Csp5_scaffold_00069.alt.g2454
Cbr-msrp-9
msrp-Csp5_scaffold_02717.alt.g22336
0.94
0.8
0.96
0.94
0.47
0.84
0.86
0.67
0.85
0.67
1
0.57
0.85
0.7
0.92
0.32
0.99
0.9
0.93
0.99
0.07
1
0.76
0.44
0.78
0.86
0.12
0.9
0.92
0.78
0.52
0.16
0.83
0.79
0.27
0.85
0.84
0.66
0.99
0.82
0.57
0.6
1
0.82
0.27
0.76
0.97
0.99
0.86
0.86
0.84
0.98
0.72
0.87
0.83
0.8
0.84
0.36
0.98
0.68
0.83
0.74
0.86
0.64
0.93
0.84
0.55
0.89
0.95
0.64
0.86
0.95
0.98
0.66
0.82
1
0.3
msrp
(not GPI-
anchored
mss
msrp

Figure S6 | Phylogenetic relationships of the MSS family and MSS-related proteins (MSRPs). Beginning with *mss* homologs (**Figure 3d**), iterative searches were made to identify other, more distantly related protein sequences encoded by *Caenorhabditis* genomes. Alignment and maximum-likelihood phylogenetic analysis of the final homolog set (**table S8, dataset S5**) revealed three distinct, major clades. Two clades contain distinct subsets of *msrp* genes, which differed with respect to the presence or absence of predicted GPI lipid anchor modification sites. The third clade encompasses all *mss* genes. The scale bar (for horizontal branch lengths) represents amino acid substitutions per aligned site, as estimated in the phylogeny.

a

```
      Cnig_chr_III.g11661.t1 (Cni-mss-3 ORF) vs. Cbr-mss-3-pseudo gDNA

  1 :  M  L  H  K  T  T  L  L  F  L  A  L  A  L  I  A  V  A  F  G  E :   21
      !!:|||||||||||||||||||||||||||||||||||||||||||||||||||||||||!
       I  L  H  K  T  T  L  L  F  L  A  L  A  L  I  A  V  A  F  G  A
  1 : ATTCTCCACAAAACGACCTTGCTCTTTTTGGCTCTTGCACTGATCGCTGTAGCTTTTGGAGC :   60

 22 :   D  N  D  G  G  P  G  A  A <-><-> K  S  V  V  Q  T  G ---- N  :   38
      !:!!|||||||||||| !!|||!.!|||       ! |||  ! !!! !!.!..!####|||
        N  N  D  G  G  A  G  G  A  Y  D  H  S  R  F  P  N  S #### N
 61 : AAATAATGATGGTGGAGCTGGGGGGGCTTACGACCATTCCCGTTTTCCAAATTCTTCCAAAT :  121

 39 :  F  T  A  V  E  D  M  T  T  T  T  K  A  K  S  T  A  T  F  V  K :   59
      ! !:!!.!! !!  !|||:!!|||  ! !! !!|||!.!|||||||||:!!||| !!|||||
       S  S  T  F  I  D  V  T  D  P  P  K  G  K  S  T  S  T  V  V  K
122 : TCTTCCACATTTATAGATGTGACAGATCCACCCAAGGGTAAAAGTACATCAACTGTTGTGAA :  184

 60 :   Y  G  I { } >>>> Target Intron 1 >>>>  {G } A  S  L  V  L    :   69
      ||||||||||{|}           538 bp           {||}|||||||||!.!||||
        Y  G  I { }++                        ++{G } A  S  L  A  L
185 : GTATGGAATT{G}gt........................ag{GT}GCATCATTGGCCCTTC  :  252

 70 : L  A  A  L  :   72
      |||||||||||
      L  A  A  L
253 : TGGCTGCTCTC :  263
```

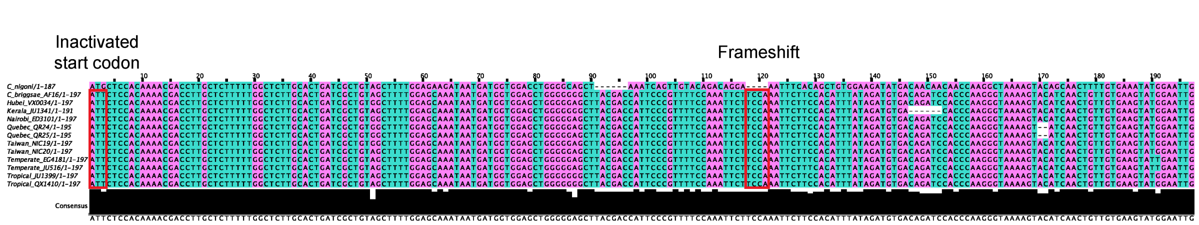


Figure S7 | Alignment of the *Cbr-mss-3-ps* pseudogene across world diversity of *C. briggsae*. a, An alignment, generated by exonerate (*117*), of homologous codons in the predicted gene *Cni-mss-3* (top) versus the predicted pseudogene *Cbr-mss-3-ps* (middle; from the *C. briggsae* strain AF16 reference genome), with the genomic DNA sequence of the latter (bottom). A 10-nt repeat present two and half times in the frame-shifted region of *Cbr-mss-3-ps* is underlined. **b,** Alignment of *Cbr-mss-3-ps* alleles found in wild isolates of *C. briggsae* versus the AF16 reference allele, with start codon and frameshift mutations indicated. Homologous fragments were identified in genome assemblies produced independently for a collection of *C. briggsae* wild isolates representing all known global diversity (*27*).

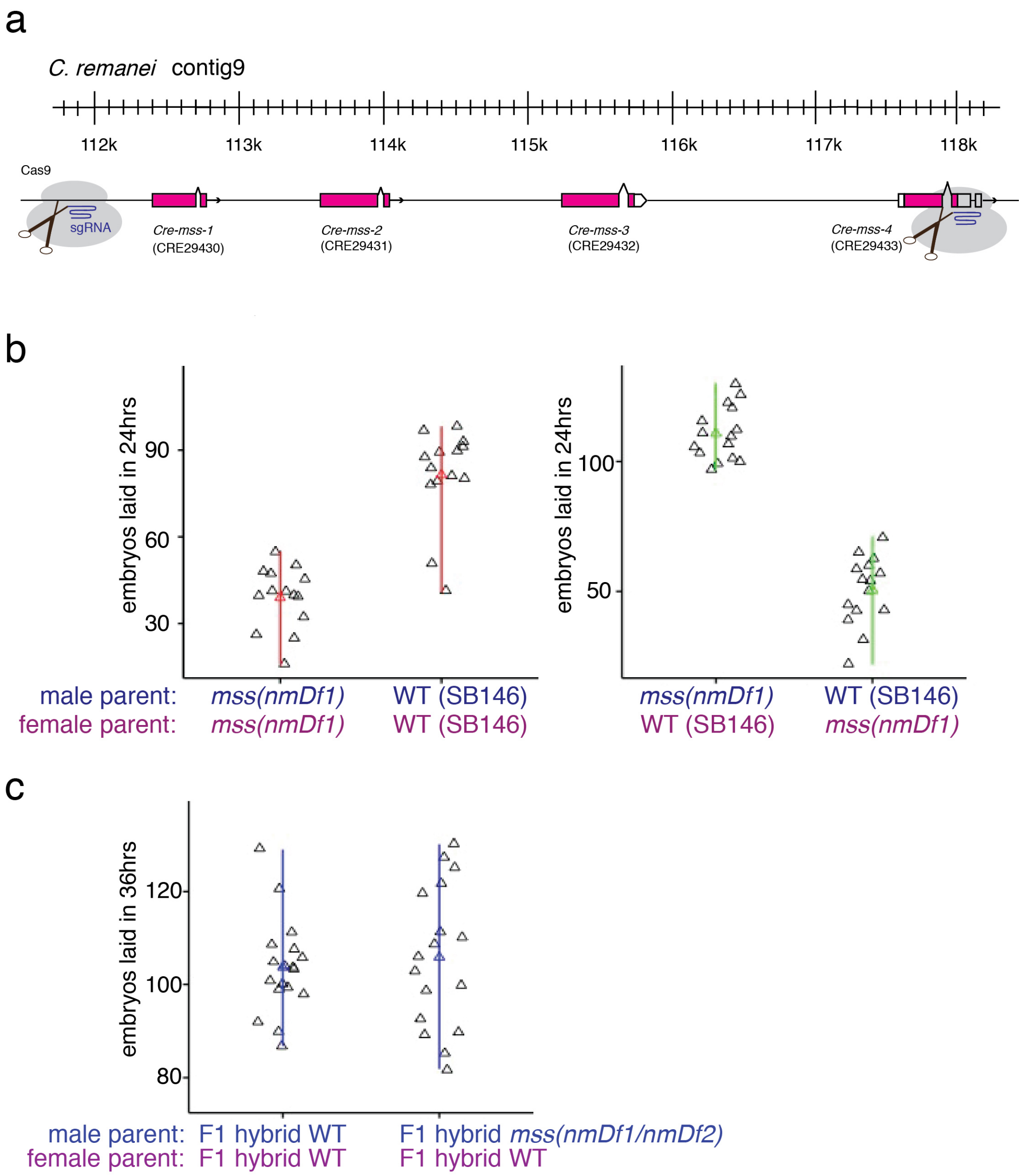


Figure S8 | MSS does not affect intrinsic fertility when inbreeding depression is eliminated. **a** Strategy for CRISPR-mutagenic deletion of the entire tandem array of *mss* paralogs from the genome of *C. remanei*, which yielded the deletion alleles *Cre-mss*(*nmDf1*) and *Cre-mss*(*nmDf2*). **b** (left), *C. remanei* strain SB146 *mss(nmDf1)* parents produced fewer embryos in a 24 hr period than wild-type *C. remanei* SB146 parents. **b** (right), *C. remanei* SB146 *mss(nmDf1)* males fertilized wild-type females and produced more embryos than when wild-type males and *C. remanei* SB146 *mss(nmDf1)* females were the parents ($P < 0.001$, two-sample Kolmogorov–Smirnov test). **c,** Hybrid (SB146/EM464) wild-type parents produced embryos in a 36 hr period that were very similar in number to hybrid *mss(nmDf1/nmDf2)* mutant males and wild-type females (statistically non-significant, two-sample Kolmogorov–Smirnov test).

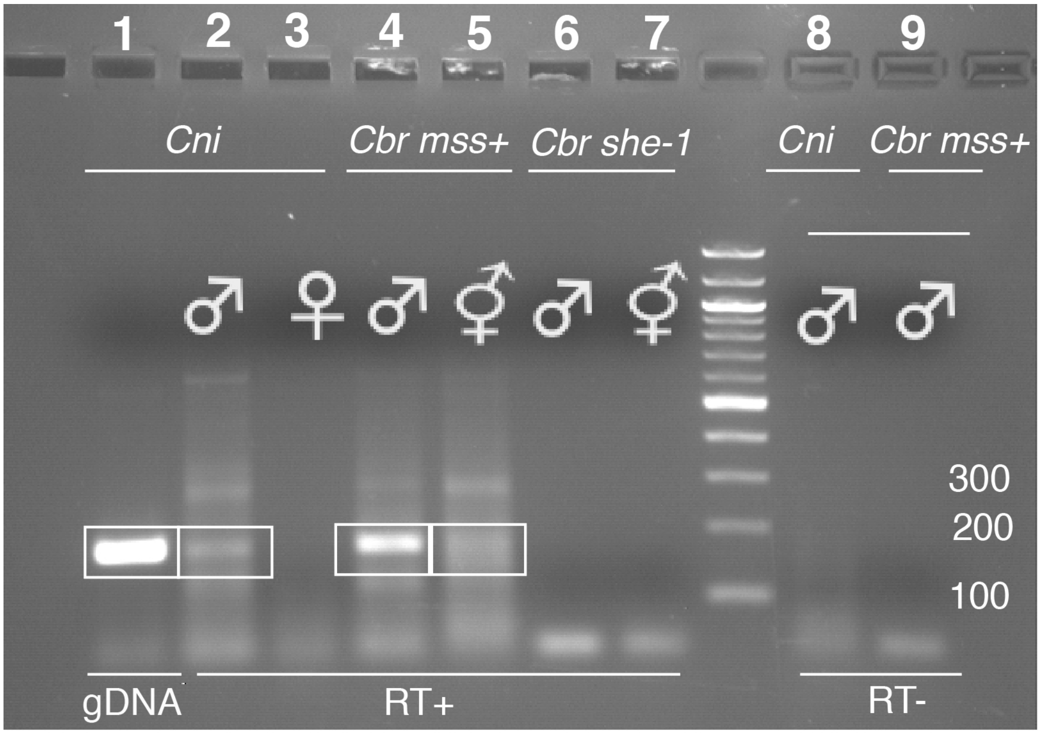


Figure S9 | *Cni-mss-3* expression in transgenic *C. briggsae.* *Cni-mss-3* transcripts from transgenic *C. briggsae* males (strong expression) and hermaphrodites (weak expression) were detected by RT-PCR (lanes 4, 5). Positive controls were performed using *C. nigoni* male genomic DNA and cDNA (lanes 1, 2). Negative controls were performed using *C. nigoni* female (lane 3); *C. briggsae she-1(v35)* mutant females (lanes 6, 7) and RNA from *C. nigoni* and transgenic *C. briggsae* without reverse transcription (lanes 8, 9).

Table S1. Genome statistics for *C. nigoni* and other *Caenorhabditis* in this study. Non-*nigoni* genome data sources are detailed in Methods. For *C. nigoni* and *C. briggsae*, repetitive DNA analyses were from this study; for *C. remanei* and *C. elegans*, they were taken from WormBase release WS254 (*118*). "% repetitive" was computed with respect to the number of non-scaffolding (non-N) residues, not the total assembly size. Protein-coding genes were computed using our methods or taken from WormBase WS254 (in the cases of *C. briggsae* and *C. remanei*, from both sources). CEGMA scores were taken from complete, full-length matches to 248 single-copy eukaryotic genes (*21*): they denote the number of genes detected, the equivalent completeness score, and the average number of hits for these 248 genes in the genome. The average number in male-female species is similar to that in hermaphroditic species, indicating that heterozygosity in all of these genomes is low.

Species:	***C. nigoni***	***C. briggsae***	***C. remanei***	***C. brenneri***	***C. elegans***
Total nt:	129,488,540	108,384,165	118,549,266	147,122,675	100,286,401
Scaffolds:	155	367	1,591	1,859	7
Contigs:	213	6,724	48,142	8,900	7
ACGT nt:	129,435,387	105,416,539	112,435,984	133,694,091	100,286,401
N-res. nt:	53,153	2,967,626	6,113,282	13,428,584	0
% non-N:	99.96	97.26	94.84	90.87	100.00
% GC:	37.75	37.35	37.89	38.58	35.44
% repetitive:	27.31	26.04	16.99	n/d	21.95
Scaffold N50 nt:	20,390,332	17,485,439	1,522,088	760,442	17,493,829
Scaffold N90 nt:	15,535,478	14,578,851	98,048	82,237	13,783,801
Scaf. max. nt:	23,648,458	21,540,570	18,579,143	4,147,112	20,924,180
Scaf. min. nt:	1,315	1,378	864	538	13,794
Contig N50 nt:	3,254,670	41,490	14,669	37,413	17,493,829
Contig N90 nt:	562,841	9,261	2,027	8,044	13,783,801
Contig max. nt:	9,436,569	516,571	207,691	448,409	20,924,180
Contig min. nt:	1,315	1	1	1	13,794
Protein-coding genes (this study):	29,167	22,313	26,960	34,049	[none]
Protein-coding genes (WB):	[none]	21,814	26,226	[none]	20,257
CEGMA genes	247	247	238	247	244
CEGMA % completeness	99.60	99.60	95.97	99.60	98.39
CEGMA Average	1.19	1.13	1.19	1.24	1.12

Table S2. Chromosome sizes and species-specific sequence content. Sizes, total unalignable (species-specific) sequences, and the subset of species-specific sequences larger than 1 kb for a given chromosome are shown for *C. nigoni* (top) and *C. briggsae* (bottom).

C. nigoni	**chrI**	**chrII**	**chrIII**	**chrIV**	**chrV**	**chrX**
Size (Mb)	16.7	19.2	15.5	20.4	22.3	23.6
Species-specific DNA	5.2 (31%)	6.7 (35%)	5.1 (33%)	8.1 (40%)	7.3 (33%)	8.3 (35%)
Species-specific DNA (>1 kb)	2.1 (13%)	2.8 (15%)	2.0 (13%)	3.7 (18%)	3.2 (15%)	3.8 (16%)
C. briggsae	**chrI**	**chrII**	**chrIII**	**chrIV**	**chrV**	**chrX**
Size (Mb)	15.5	16.6	14.6	17.5	19.5	21.5
Species-specific DNA	3.7 (24%)	4.2 (25%)	3.7 (25%)	5.1 (29%)	4.4 (23%)	5.6 (26%)
Species-specific DNA (>1 kb)	0.67 (4.3%)	0.80 (4.8%)	0.68 (4.7%)	1.1 (6.3%)	0.80 (4.1%)	1.2 (5.4%)
***C. nigoni* size difference**	8.3%	15.7%	6.6%	16.6%	14.3%	9.8%

Table S3. Comparison of intron and exon content of orthologous genes. Available online as an Excel file. Subtables 1 and 2 present summaries of all pairwise regressions of exon and intron sums for 1:1 orthologs of *C. nigoni, C. briggsae, C. remanei,* and *C. elegans* (defined in orthology groups in the OFind_5spp_Summary and OFind_5spp data columns of **table S4**), for autosomes (subtable 1) and X chromosomes (subtable 2). Subtables 3 and 4 present the exon and intron size data for the longest predicted isoform of each orthologous gene, from which the summaries were prepared, for autosomes (subtable 3) and X chromosomes (subtable 4). For *C. briggsae* and *C. remanei,* alternative predictions generated by the same methods used for *C. nigoni* are indicated with the "alt" suffix. Results of Wilcoxon rank-sum tests for differences in the comparably predicted *C. nigoni, C. briggsae*, and *C. remanei* exon and intron sums are also presented.

Table S4. Protein-coding gene predictions and annotations for *C. nigoni*. Available online as an Excel file. Its data columns are as follows:

Gene: a given predicted protein-coding gene in the *C. nigoni* genome assembly. All further data columns are pertinent to that particular gene.

Prot_size: this shows the full range of sizes for all protein products from a gene's predicted isoforms.

Max_prot_size: the size of the largest predicted protein product.

Phobius: this denotes predictions of signal and transmembrane sequences made with Phobius 1.01 (*89*). 'SigP' indicates a predicted signal sequence, and 'TM' indicates one or more transmembrane-spanning helices, with N helices indicated with '(Nx)'. Varying predictions from different isoforms are listed.

NCoils: this shows coiled-coil domains, predicted by ncoils (*90*). Both the proportion of such sequence (ranging from 0.01 to 1.00) and the exact ratio of coiled residues to total residues are given. Proteins with no predicted coiled residues are blank.

Psegs: this shows what fraction of a protein is low-complexity sequence, as detected by pseg (*91*). As with Ncoils, the relative and absolute fractions of each protein's low-complexity residues are shown.

PFAM: predicted protein domains from Pfam 31 (*82*), with an E-value of $\leq 10^{-5}$.

InterPro: predicted protein domains from InterPro, predicted with InterProScan 5.18-57.0 (*93*).

GO_terms: Gene Ontology (GO) terms, generated with Blast2GO v1.3.3 (*97*).

Orphan/non-orphan: for each *C. nigoni* gene, its homology or similarity to non-*nigoni* genes is categorized as follows. Genes are classified as having non-*nigoni* homologs if they belong to OrthoFinder groups containing non-*nigoni* members, and the species of these members are noted. Genes without homologs are classified as having similarities, if they have BlastP matches to non-*nigoni* gene products; again, the species of these BlastP matches are noted. Finally, genes lacking either OrthoFinder homologies or BlastP similarities are classified as pure orphans.

OFind_5spp_Summary and **OFind_5spp:** the results for our OrthoFinder analysis, encompassing five *Caenorhabditis* (*C. nigoni*, *C. briggsae*, *C. remanei*, *C. brenneri*, and *C. elegans*). For two of these species (*C. briggsae* and *C. remanei*), two different proteomes were included: the official predicted proteome in WormBase WS254 (labeled 'briggsae' and 'remanei'), and our own independent predictions using computational methods equivalent to the ones used for *C. nigoni* (labeled 'briggsae_alt' and 'remanei_alt'). Two different views of these results are given: the summary lists taxa and gene counts, while the full results give individual gene names.

Male_Female_log2FC: the fold-changes of gene expression between males and females, expressed as $\log_2$ values, and with positive values representing greater male expresssion. The values listed here are only those were computed to be significant using edgeR 3.14.0 (*87*), with three biological RNA-seq replicates per sex, and with significant results annotated for individual genes.

Male_Female_FDR: the false discovery rate (FDR) for gene expression changes between males and females, annotated for individual genes. The FDR for a given set of positive results is defined as that significance threshold which, if accepted, will lead to the entire set of positives having a collective false-positive rate no greater than the FDR; it therefore provides a way to correct for testing multiple hypotheses without rejecting excessive numbers of true positives. As with Male_Female_log2FC, only changes that were computed to be significant by edgeR are listed.

Male_modENCODE_log2FC: the fold-changes of gene expression between males and mixed-sex whole-animal *C. nigoni* that had been previously characterized by the modENCODE consortium (*70*); changes are expressed as $\log_2$ values, and with positive values representing greater male expression.

Male_modENCODE_FDR: the FDR for gene expression changes between males and modENCODE, annotated for individual genes.

Female_modENCODE_log2FC: the fold-changes of gene expression between females and mixed-sex whole-animal *C. nigoni* that had been previously characterized by the modENCODE consortium (*70*); changes are expressed as $\log_2$ values, and with positive values representing greater female expression.

Female_modENCODE_FDR: the FDR for gene expression changes between females and modENCODE, annotated for individual genes.

[Sample]_TPM: gene expression levels in TPM, computed for individual RNA-seq data sets by Salmon (*86*). RNA-seq samples include biological triplicates for males (Male_1 through Male_3), biological triplicates for females (Female_1 through Female_3), and modENCODE data.

[Sample]_reads: numbers of mapped RNA-seq reads per gene, computed for individual RNA-seq data sets by Salmon, with fractional values rounded down to integers.

Table S5. Pfam domain and Gene ontology (GO) term enrichment analysis. Available online as an Excel file. Subtable "Pfam" lists 26 Pfam domains that are statistically overrepresented in *C. nigoni* versus *C. briggsae*, and 13 Pfam domains that are overrepresented in *C. briggsae* versus *C. nigoni*. Subtable "Pfam multispecies comparison" provides multispecies statistical analysis of the 26 overrepresented Pfam domains, showing seven of these domains to be generally overrepresented in male-female versus hermaphroditic *Caenorhabditis*. Subtable "GO terms and InterPro families" includes tables for the GO terms enriched in the set of *C. nigoni* genes that lack *C. briggsae* homologs, and for the OrthoFinder gene families with the highest proportion of *C. nigoni* genes to *C. briggsae* genes and their associated functional annotations, including Pfam, InterPro and GO. Other subtables provide details of individual Pfam domains that are overrepresented in *C. nigoni*, such as F_box.

Table S6. Relationship of protein size to sibling-species conservation for *C. nigoni* and *C. briggsae* proteins. For all *C. nigoni* genes, the longest isoforms were divided into different size classes and categorized as either lacking or having *C. briggsae* homologs ("*C. briggsae* [-]" and "*C. briggsae* [+]"); the same was also done for *C. briggsae* genes with respect to *C. nigoni* homologs. Equivalent methods were used to predict both gene sets, and homologs were determined by OrthoFinder. In the two smallest size classes (66-99 and 100-199 residues), *C. nigoni* proteins without *C. briggsae* homologs are significantly overrepresented with respect to the total proteome (Fisher's two-tailed exact test, confidence level 0.99); conversely, in the median and two largest size classes (200-399, 400-599 and 600+ residues), *C. nigoni* proteins with *C. briggsae* homologs are significantly overrepresented. The same pattern is observed for *C. briggsae* proteins.

***C. nigoni* prots.:**	**66-99 aa**	**100-199 aa**	**200-399 aa**	**400-599 aa**	**600+ aa**	**Total**
***C. briggsae* [-]**	1,455	1,905	934	250	282	5,626
***C. briggsae* [+]**	1,436	5,031	9,264	4,511	4,099	23,541
Total	2,891	6,936	10,198	4,761	4,381	29,167
Bias	[-]	[-]	[+]	[+]	[+]	n/a
p-value	<2.2e-16	<2.2e-16	<2.2e-16	<2.2e-16	<2.2e-16	n/a

***C. briggsae* prots.:**	**66-99 aa**	**100-199 aa**	**200-399 aa**	**400-599 aa**	**600+ aa**	**Total**
***C. nigoni* [-]**	826	1,043	418	92	64	2,443
***C. nigoni* [+]**	1,105	4,290	7,294	3,764	3,417	19,870
Total	1,931	5,333	7,712	3,856	3,481	22,313
Bias	[-]	[-]	[+]	[+]	[+]	n/a
p-value	<2.2e-16	<2.2e-16	<2.2e-16	<2.2e-16	<2.2e-16	n/a

Table S7. Statistics for sex-biased genes in *Caenorhabditis* genomes. Protein-coding genes, their observed sex-biased RNA-seq expression, and their XX sex are listed for the genomes of five *Caenorhabditis*. Sources of genomic and RNA-seq data for the non-*nigoni* species are listed in Methods. For *C. briggsae* and *C. remanei,* alternative predictions of protein-coding genes, produced by the same methods used for *C. nigoni*, are indicated with "(alt.)". Details of gene expression data are given in **table S4** (for *C. nigoni*), **dataset S3** (for *C. briggsae*) and **dataset S4** (for *C. remanei*, *C. brenneri*, and *C. elegans*). Notably, these were used to provide the sex-biased gene expression data for *mss* and *msrp* genes in **table S8**.

Species	**Total genes**	**XO-biased**	**XX-biased**	**XX sex**
C. nigoni	29,167	3,895	1,707	female
C. briggsae (alt.)	22,313	4,589	2,215	hermaphrodite
C. briggsae	21,814	4,334	2,100	hermaphrodite
C. remanei (alt.)	26,960	5,442	3,898	female
C. remanei	26,226	5,334	3,724	female
C. brenneri	34,049	6,689	4,486	female
C. elegans	20,257	5,166	3,448	hermaphrodite

Table S8. Definition and expression of *mss* and *msrp* gene families. Available online as an Excel file. This table matches the gene names used in **Figure 3d** and **figure S6** with gene prediction names from this study and WormBase. Its data columns are as follows:

Gene: a given predicted protein-coding gene in a *Caenorhabditis* species. All further data columns are pertinent to that particular gene.

***mss*/*msrp* family:** the phylogenetic class to which the gene belongs, as defined in **figure S6**.

Locus name: the classical gene name (where one has been given). These names have been approved by WormBase, and correspond with genes that have official WormBase gene identification numbers.

WBgene ID: the official WormBase identification for a gene (where one exists).

***Caeno*. sp.:** the *Caenorhabditis* species encoding the gene.

Comment: notes on synteny with other genes (in cases where such synteny can be detected).

XO.vs.XX_log2FC: the fold-changes of gene expression between XO animals (uniformly males) and XX animals (depending on the *Caenorhabditis* species, either females or hermaphrodites), expressed as $\log_2$ values, and with positive values representing greater male expresssion. The values listed here are only those that were computed to be significant using edgeR 3.14.0 (*87*), with three biological RNA-seq replicates per sex, and with significant results annotated for individual genes.

XO.vs.XX_FDR: the false discovery rate (FDR) for gene expression changes between XO animals (uniformly males) and XX animals (depending on the *Caenorhabditis* species, either females or hermaphrodites), annotated for individual genes. The FDR for a given set of positive results is defined as that significance threshold which, if accepted, will lead to the entire set of positives having a collective false-positive rate no greater than the FDR; it therefore provides a way to correct for testing multiple hypotheses without rejecting excessive numbers of true positives (*119*). As with Male_Female_log2FC, only changes that were computed to be significant by edgeR are listed.

XO.vs.XX log2FC (Thomas): for *C. japonica* genes, the fold-changes of gene expression between XO animals (uniformly males) and XX animals (uniformly females), expressed as $\log_2$ values, and with positive values representing greater male expresssion. The values listed here are only those that were computed to be significant by Thomas et al. (*11*), with three biological RNA-seq replicates per sex, and with significant results annotated for individual genes. Data were downloaded from the NCBI GEO archive at *ftp://ftp.ncbi.nlm.nih.gov/geo/series/GSE41nnn/GSE41367/suppl/GSE41367%5Fjaponica%5Fgns%2Eexpr%2Etxt%2Egz*.

p-value (Thomas): for *C. japonica* genes, the statistical significance for differences of expression between XO and XX animals, as computed by Thomas et al. (*11*); these data, like those for "XO.vs.XX log2FC (Thomas)", were taken from the GEO archive.

Phobius: this denotes predictions of signal and transmembrane sequences made with Phobius 1.01 (*89*). 'SigP' indicates a predicted signal sequence, and 'TM' indicates one or more transmembrane-spanning helices, with N helices indicated with '(Nx)'. Varying predictions from different isoforms are listed.

NetOGlyc_4.0: this denotes predictions of GalNAc-type O-glycosylation sites made with NetOGlyc 4.0.0.13 (*94*) via its web server (*http://www.cbs.dtu.dk/services/NetOGlyc*); the isoform on which predictions were made is listed in brackets.

PredGPI: this denotes predictions of GPI anchor sites made with PredGPI (*95*) via its web server (*http://gpcr.biocomp.unibo.it/predgpi/index.htm*); these used the same isoforms as NetOGlyc.

Omega_site: this denotes predicted locations of omega sites for GPI anchoring made with PredGPI.

Table S9. Proportions of *C. nigoni* and *C. briggsae* genes with male-biased, unbiased, and female-biased expression. For each species, we identified genes that exhibited highly significant expression data (FDR ≤ 0.001) for differential RNA-seq expression in comparisons between males (XO chromosomes) and either females or hermaphrodites (XX chromosomes, for *C. nigoni* and *C. briggsae*). Note that this filters out roughly two-thirds of the genes in each species, but selects for the one-third of genes whose expression could be sharply distinguished between male-biased, female/hermaphrodite-biased, or unbiased. We defined male-biased as those genes exhibiting ≥2-fold higher expression in males than in females/hermaphrodites, and female-/hermaphrodite-biased as those genes exhibiting ≥2-fold higher expression in females/hermaphrodites than in males; genes falling in between these two sets were classified as unbiased. We further stratified genes by whether they had homologs in their sibling species or not (e.g., "*C. briggsae* [+]" and "*C. briggsae* [-]"). Both absolute numbers and percentages of genes within a homology class are given. For both *C. nigoni* and *C. briggsae*, there was an increased proportion of genes with male-biased expression among genes lacking homologs in their sibling species, but the proportional increase in such genes was far greater in *C. nigoni* than in *C. briggsae*.

***C. nigoni* genes:**	Any/no homology	+ *C. briggsae* [+]	+ *C. briggsae* [-]
FDR ≤ 0.001	7,409 (100%)	6,804 (100%)	605 (100%)
+ male-biased	3,895 (52.6%)	3,466 (50.9%)	429 (70.9%)
+ unbiased	1,807 (24.4%)	1,728 (25.4%)	79 (13.6%)
+ female-biased	1,707 (23.0%)	1,610 (23.7%)	97 (15.0%)

***C. briggsae* genes:**	Any/no homology	+ *C. nigoni* [+]	+ *C. nigoni* [-]
FDR ≤ 0.001	7,696 (100%)	7,308 (100%)	388 (100%)
+ male-biased	4,589 (59.6%)	4,349 (59.5%)	240 (61.9%)
+ unbiased	892 (11.6%)	871 (11.9%)	21 (5.4%)
+ hermaph.-biased	2,215 (28.8%)	2,088 (28.6%)	127 (32.7%)

Dataset S1. Genomic sequences, gene predictions, and repetitive DNA analyses for *C. nigoni*. Data files are available at the OSF (*https://osf.io/dkbwt* and doi:10.17605/osf.io/dkbwt).

Dataset S2. Positions of *C. briggsae* and *C. nigoni* sequences that are not alignable with the other species ("species-specific sequences") shown in Figure 2a. Text files are available online. Both files were generated by the *nucmer* and *dnadiff* program of MUMmer.

Dataset S3. Gene predictions, repetitive DNA analyses, OrthoFinder homologies, male-female RNA-seq analyses, other gene annotations, and assemblies of wild isolate genomes for *C. briggsae*. Data files are available at the OSF (*https://osf.io/a4e8g* and doi:10.17605/osf.io/a4e8g).

Dataset S4. Gene predictions and RNA-seq analyses for *C. remanei*, *C. brenneri*, and *C. elegans*, with a heterozygosity-reduced genome sequence for *C. brenneri*. Data files are available at the OSF for *C. remanei* (*https://osf.io/hxszb* and doi:10.17605/osf.io/hxszb), *C. brenneri* (*https://osf.io/674un* and doi:10.17605/osf.io/674un), and *C. elegans* (*https://osf.io/ze6kt* and doi:10.17605/osf.io/ze6kt).

Dataset S5. Sequence, alignment, and predicted phylogeny of *mss* and *msrp* sequences. Data files are available at the OSF (*https://osf.io/g397t* and doi:10.17605/osf.io/g397t). Most files give details of *mss*/*msrp* protein sequences, alignment, and phylogeny. One file (*mss_genomic_DNA_alignment_2017.08.30.pdf*) shows an annotated alignment of DNA sequences for the *mss* genomic loci of *C. nigoni* and *C. briggsae*; note that these are the same loci shown in a schematic overview in **Figure 3e**. "Cbr" denotes *C. briggsae* genomic sequence, and "Cni" that of *C. nigoni.* Exons and introns are noted above, with the terminal arrow indicating the direction of transcription and the reading frame of translation indicated in parentheses. Forward translation frames are shown below each sequence (reverse frames are omitted to save space). Note that *Cni-mss-3* and the orthologous *C. briggsae* pseudogene are transcribed in reverse orientation. A 10-nt repeat sequence associated with a frameshift in the *Cbr-mss-3-ps* pseudogene is also indicated.