Supporting Information

Wurm et al. 10.1073/pnas.1009690108

SI Materials and Methods

Sequencing. Hymenopterans including ants have haplodiploid sex determination: Fertilized (diploid) offspring are female and unfertilized (haploid) offspring are male. We used a single haploid male (e.g., Fig. 1*A*) to facilitate de novo assembly. This focal male had the *Gp-9 B* genotype (1) and was the offspring of a *Gp-9 Bb* queen from a multiple-queen colony originally collected near Athens, Georgia, in 2008. Upon transfer of the colony to the laboratory, the queen was isolated with workers for 1 year to ensure that all progeny in the colony were her offspring.

A Roche 454 shotgun sequencing library and an Illumina paired-end library (insert size estimated from gel: 330 bp) were built from the DNA of the single focal haploid male. The Illumina library provided 160,371,174 pairs of reads with between 36 and 101 bp of read length for a total of 22,063,840,466 bp; the Roche 454 library provided 17,345,989 reads with a mean length of 314.1 bp for a total of 5,448,727,077 bp. These libraries made up 93.8% of our sequencing effort (Table S1*A*).

To permit bridging of repeats that could be longer than the Roche 454 read length, we additionally constructed 8- and 20-kb insert paired-end Roche 454 libraries for the remaining 6.2% of the sequencing effort (1,213,754,187 and 595,410,131 bp, respectively; Table S1*A*). Because large amounts of DNA were required, the 8- and 20-kb libraries were, respectively, constructed from DNA pooled from 10 and 31 brothers of the focal male, all having the *Gp-9 B* genotype (1). The combined haploid genomes of these brothers are representative of the diploid genome of their mother; thus half of the paired-end reads are expected to represent the haploid genome of the focal male.

DNA was isolated using the Qiagen DNeasy Blood and Tissue Kit. Sequencing libraries were constructed according to protocols recommended by the respective manufacturers. All sequence reads were deposited in the National Center for Biotechnology Information (NCBI) Sequence Read Archive (http://www.ncbi.nlm.nih.gov/Traces/sra) under study accession no. SRP002592.

Assembly. Assembly was performed in several steps, beginning with Illumina data. We extracted from each read the longest region (minimum of 30 bp) with a Phred quality score of 5 or greater, resulting in 14,718,896,089 bp of sequence. These reads were assembled using overlap information in SOAPdenovo (release 1.04, 21–12-2009) (2), with the "-R" flag for resolution of small repeats and a k-mer word size of 23. Only contigs longer than 200 bp were retained. Subsequently, 103,982,886 pairs of reads of 2 × 35 bp and with a Phred quality score >4 were extracted from the same initial dataset for a total of 7,278,802,020 bp. These paired reads were used to combine contigs into scaffolds by running SOAPdenovo with minimum 34-bp alignment length. Finally, the 2 × 35-bp paired reads were input to the GapCloser for SOAPdenovo software to extend scaffold edges and fill intrascaffold gaps. This resulted in 124,008 Illumina scaffolds with N50 of 3,658 bp for a total of 206,906,335 bp of sequence.

Preliminary tests indicated that we could combine Illumina with Roche 454 data using the Roche GS De Novo Assembler (Newbler).

The Illumina assembly generated 12,177 scaffolds containing unresolved gaps that were marked by sequences of "N"s. Because incorrect estimation of gap size may create problems during subsequent assembly with 454 data, we split these scaffolds into multiple pieces to eliminate all unresolved gaps. Newbler ignores input sequences that are longer than 1999 bp and does not retain sequences represented by a single read. We thus further split the Illumina scaffolds into subsequences of 300 bp with a 200-bp overlap using EMBOSS splitter (3), resulting in 553,154,535 bp of sequence in FASTA format. The SFF files from Roche 454 shotgun sequencing as well as a FASTA file containing the split Illumina assembly were input to Roche 454 Newbler 2.3 release 091027_1459. Running Newbler with stringent parameters ("-mi 98 -ml 100 -ud -rip -large -m -e 9.5") resulted in 90,446 contigs with an N50 of 12,703 bp for a total of 355 Mb. Subsequently, SFF files from the 8-kb insert Roche 454 library were added with stringent parameters likely to eliminate reads that could introduce ambiguities due to differences from those of the focal haploid male (as above but with "-ml 200"). Finally, SFF files with paired information from the 20-kb insert Roche 454 library were added with the same parameters. The decisions to combine split preassembled Illumina data into the Roche 454 assembly and to increase stringency of Roche 454 assembly led to increases in assembly confidence and quality as determined by N50-related statistics (Table S1*F*). We examined the fate of the sequences that derived from the 12,177 Illumina scaffolds that had been split because of unresolved gaps to determine how much information regarding the formation of Illumina contigs and scaffolds was being lost during the splitting and reassembly procedure. The assembled sequences were within a single scaffold of the final assembly in 92.1% of cases (in 10,436 cases they were within a single contig). The fragments from the remaining 7.9% of split Illumina scaffolds were excluded from the assembly as singletons, outliers, or repeats, were located in different scaffolds, or were located in nonscaffolded contigs.

Gene Prediction. We combined several data sources and computational tools to establish an Official Gene Set (OGS). First, the MAKER pipeline (4) derived consensus gene models from predictions obtained by Augustus, SNAP, and Exonerate based on built-in models as well as hints from *Solenopsis invicta* ESTs (5, 6) and the proteomes of *Apis mellifera* (prerelease OGS 2), *Nasonia vitripennis* (OGS 1.1), *Drosophila melanogaster* (Biomart download on 26.04.2010) and *Homo sapiens* (UniProt download on 26.04.2010). The longest protein at each genomic locus was retained, resulting in a set of 19,728 gene models. In parallel, we used the homology-based gene structure prediction method GeneWise (7) to detect genes: GeneWise was run using standard parameters on genomic scaffolds together with *N. vitripennis* OGS 1.1. We obtained 10,262 gene models. The MAKER and GeneWise gene sets were then merged. Redundancy was removed by favoring first predictions starting with a methionine and subsequently predictions that were longer. This resulted in a set of 21,552 gene models. Subsequently, we kept sequences that had either support from *S. invicta* ESTs (tblastn $E < 10^{-5}$ with at least 99% identity) or that had a blastp match in *A. mellifera*, *N. vitripennis*, or *D. melanogaster* ($E < 10^{-5}$). Finally, we removed retrotransposon sequences for a final OGS 2.2 of 16,569 genes.

Assembly and Annotation Evaluation. We used the *D. melanogaster* set of 248 conserved core eukaryotic genes (generated by the CEGMA pipeline) (8) to test the quality of our gene models. We found 246 of these 248 CEGMA sequences in the *S. invicta* genome scaffolds. A total of 215 of these are likely complete (the remaining 31 gene models contained either frameshifts or stop codons or their length varied by more than 10% from that of the corresponding *D. melanogaster* sequence).

The Roche 454 technology is prone to insertion and deletion errors in homopolymer runs (i.e., when a single base is repeated multiple times) (9). The Illumina sequencing technology does not have this particular problem (10). Although our assembly approach does not explicitly use Illumina sequence for homopolymer error correction, it is likely that it reduced the proportion of such errors. To determine whether this was the case, we examined the sequences of the single *S. invicta* orthologs to the 246 highly conserved eukaryotic genes of the CEGMA dataset. Consensus *S. invicta* protein sequences were mapped with the Exonerate software (11) to the final *S. invicta* genome assembly as well as to an assembly that was obtained with the same assembly parameters but with only 454 data. Visual inspection of Exonerate output identified 13 frameshift-inducing indels (insertions or deletions) in exons of the 246 CEGMA genes. Pairwise alignments helped characterize the nature of these indels. Six indels of which five are putatively homopolymer-related were specific to the 454-only assembly. These six indels thus had apparently been corrected in the final assembly by the use of Illumina data. One putatively homopolymer-related indel was specific to the final assembly and may reflect stochasticity in the assembly process of Roche 454 Newbler. Finally, six frameshift-inducing indels are shared between the two assemblies.

To estimate the size of inserts in the Illumina paired-end library, we randomly selected 100,000 pairs of reads and mapped them to the genome with Maq software. Maq estimates insert size at 351.9 ± 34.3bp (SD). The gel-extraction estimate of insert size was 330 bp.

Gene-Centered Analyses. To quantify variation in numbers of protein family members, we performed Pfam (version 24.0) (12) and PROSITE profile (13) analyses on *D. melanogaster*, *N. vitripennis*, and *A. mellifera* nonredundant protein datasets (one splice variant per gene) and compared the numbers of matched proteins with those obtained from the *S. invicta* gene set. For genes and gene families discussed in the main text, Apollo-based visual inspection of gene models (14) as well as blast and reciprocal blast analyses were used to help determine orthology relationships. For lipid-processing genes and olfactory receptors, GeneWise (7) helped refine gene models. For putative vitellogenins, insulin-related genes, and telomerase reverse transcriptase, MAKER (4) was rerun locally to improve automated gene predictions, and Apollo (14) was used to manually fine-tune gene models.

Vitellogenin Real-Time Quantitative RT-PCR. Quantitative real time RT-PCR (qRT-PCR) was performed on queens and workers with the *Gp-9 BB* genotype from single-queen colonies. Field colonies collected in Athens, Georgia, were returned to the laboratory and reared for 2 months under standard rearing conditions (15). The single mated queen was collected from each of 10 different colonies, and 10 major workers were collected from the foraging area of each of 10 additional colonies. Queens were at least 6 months old, and major workers typically forage at the age of 2 months and die at the age of 4 months (16). RNA extractions were performed using the whole body of the ants and a modified protocol that includes the use of TRIzol (Invitrogen) and the RNeasy extraction kit (Qiagen). For each individual ant, cDNA was synthesized using 200 ng of total RNA, random hexamers, and Applied Biosystems reagents. mRNA quantification of vitellogenins was performed with an ABI Prism 7900 Sequence Detection System, sequence-specific primers (Table S1*G*), and SYBR green. All RT-PCR assays were performed in triplicate and subjected to the heat-dissociation protocol following the final cycle of the RT-PCR to check for amplification specificity. RT-PCR values of *Vg* genes and three housekeeping genes (*RP9*, *RP37*, *H2A*) were tested. The *RP9* gene displayed the least variation among groups and was thus used to normalize the results using the ΔCt method (17).

Transformer/Feminizer Genes. *Transformer/feminizer* sequences were retrieved from GenBank, with accession numbers: ACF08858 (*N. vitripennis*), ABU68668 (*A. mellifera feminizer*), ABU68670 (*A. mellifera CSD*), ABY74329 (*Bombus terrestris*). *Transformer/feminizer* homologs in *S. invicta* were identified via blast similarity, and gene/transcript models manually were inspected and adjusted. CDART (18) was used to identify the SDP_N (cl13684) and Apis_CSD (cl13171) domains on inferred *S. invicta* sequences.

Phylogenetic Trees. Trees shown in Fig. 3*B* and Fig. S2*B* were constructed as follows: Initial protein alignments were performed using ClustalW2 (19) and then edited using Jalview (20). Edited sequences were realigned using ClustalX 2.0.12. Parsimony trees were established using PAUP 4.0 b10 (21) and were rooted using the most divergent sequence in each group as the outgroup. Bootstrap support for internal branches was evaluated from 10,000 full-heuristic searches, and groups with a frequency greater than 50% were retained in the consensus trees. Other trees were constructed as described in the legends of Fig. S2*A* and Fig. S3*C*.

Methylation. The pooled DNA of queen and worker prepupae was subjected to methylated DNA immunoprecipitation (MeDIP) and sequenced to identify putatively methylated regions of the *S. invicta* genome. The Maq program was used to map Illumina sequencing reads to reference genome scaffolds and to extract read-depth information in approximately unique regions within 1-kb windows. As expected, a negative correlation was observed between normalized CpG dinucleotide content ($CpG_{O/E}$) and the coverage of MeDIP-sequencing reads (Table S1 *B* and *C*) (22, 23). In contrast, no correlation was observed with the control statistic $GpC_{O/E}$ (Table S1*B*), which, unlike the measure $CpG_{O/E}$, is not expected to vary as a function of the DNA methylation level. Thus the agreement of MeDIP enrichment and CpG depletion, along with a complete suite of DNA methyltransferase enzymes, provides strong support for the existence of functional methylation in *S. invicta* (24).

A list of genes subject to putatively dense levels of DNA methylation was produced by blastx comparison of windows with read-depth values greater than 15 (~5% of windows with one or more mapped reads) to the *S. invicta* official protein set using a conservative threshold for similarity ($E < 10^{-100}$). This analysis produced a list of 80 putatively methylated genes. The corresponding protein sequences were compared against the nonredunant (nr) protein database using blastp and *D. melanogaster* homologs were identified using InParanoid (25) to test for Gene Ontology biological process term enrichment when compared with a background composed of all *D. melanogaster* genes (23, 26). As is the case in *A. mellifera*, methylated genes were enriched for biological processes related to cellular metabolism and transcription in *S. invicta* (Table S1*E*) (23).

The top 96 MeDIP-enriched genes were then selected (as above) for the design of primers to amplify bisulfite-converted genomic DNA using the MethPrimer tool (Table S1*D*) (27). *S. invicta* genomic DNA from workers of mixed developmental stages was bisulfite-converted using the EpiTect Bisulfite Kit (Qiagen). On the basis of PCR amplification efficiency, nine amplicons from distinct genes were cloned using the TOPO TA Cloning Kit for Sequencing (Invitrogen). Between 3 and 14 clones from each amplicon were sequenced (High-Throughput Sequencing Solutions, University of Washington) and analyzed using the QUMA quantification tool for methylation analysis (28). Six of nine genes demonstrated strong evidence of CpG methylation, with CpH (CpA, CpC, or CpT) bisulfite conversion ranging from 91.7% to 100% for each clone (Fig. S4). These results confirm the presence of CpG methylation in *S. invicta* genes.

1. Krieger MJB, Ross KG (2002) Identification of a major gene regulating complex social behavior. *Science* 295:328–332.
2. Li R, et al. (2010) *De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Res* 20:265–272.
3. Rice P, Longden I, Bleasby A (2000) EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet* 16:276–277.
4. Cantarel BL, et al. (2008) MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res* 18:188–196.
5. Wang J, et al. (2007) An annotated cDNA library and microarray for large-scale gene-expression studies in the ant *Solenopsis invicta*. *Genome Biol* 8:R9.
6. Valles SM, et al. (2008) Expressed sequence tags from the red imported fire ant, *Solenopsis invicta*: Annotation and utilization for discovery of viruses. *J Invertebr Pathol* 99:74–81.
7. Birney E, Clamp M, Durbin R (2004) GeneWise and Genomewise. *Genome Res* 14: 988–995.
8. Parra G, Bradnam K, Ning Z, Keane T, Korf I (2009) Assessing the gene space in draft genomes. *Nucleic Acids Res* 37:289–297.
9. Hudson ME (2008) Sequencing breakthroughs for genomic ecology and evolutionary biology. *Mol Ecol Resour* 8:3–17.
10. Chan EY (2009) Next-generation sequencing methods: Impact of sequencing accuracy on SNP discovery. *Methods Mol Biol* 578:95–111.
11. Slater GSC, Birney E (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* 6:31.
12. Finn RD, et al. (2010) The Pfam protein families database. *Nucleic Acids Res* 38(Database issue):D211–D222.
13. Sigrist CJA, et al. (2010) PROSITE, a protein domain database for functional characterization and annotation. *Nucleic Acids Res* 38(Database issue):D161–D166.
14. Lewis SE, et al. (2002) Apollo: A sequence annotation editor. *Genome Biol* 3: research0082.
15. Jouvenaz DP, Allen GE, Banks WA, Wojcik DP (1977) A survey for pathogens of fire ants, *Solenopsis* spp., in the southeastern United States. *Fla Entomol* 60:275–279.
16. Mirenda JT, Vinson SB (1981) Division of labor and specification of castes in the red imported fire ant *Solenopsis invicta* buren. *Anim Behav* 29:410–420.
17. Livak KJ, Schmittgen TD (2001) Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) method. *Methods* 25:402–408.
18. Geer LY, Domrachev M, Lipman DJ, Bryant SH (2002) CDART: Protein homology by domain architecture. *Genome Res* 12:1619–1623.
19. Larkin MA, et al. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* 23: 2947–2948.
20. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ (2009) Jalview Version 2: A multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25: 1189–1191.
21. Swofford DL (2003) *PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods),* Version 4 (Sinauer Associates, Sunderland, MA).
22. Bird AP (1980) DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res* 8:1499–1504.
23. Elango N, Hunt BG, Goodisman MAD, Yi SV (2009) DNA methylation is widespread and associated with differential gene expression in castes of the honeybee, *Apis mellifera*. *Proc Natl Acad Sci USA* 106:11206–11211.
24. Yi SV, Goodisman MAD (2009) Computational approaches for understanding the evolution of DNA methylation in animals. *Epigenetics* 4:551–556.
25. Ostlund G, et al. (2010) InParanoid 7: New algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res* 38(Database issue):D196–D203.
26. Dennis G, Jr., et al. (2003) DAVID: Database for annotation, visualization, and integrated discovery. *Genome Biol* 4:3.
27. Li LC, Dahiya R (2002) MethPrimer: Designing primers for methylation PCRs. *Bioinformatics* 18:1427–1431.
28. Kumaki Y, Oda M, Okano M (2008) QUMA: Quantification tool for methylation analysis. *Nucleic Acids Res* 36:W170–W175.
29. Katoh K, Toh H (2008) Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform* 9:286–298.
30. Eddy SR (2009) A new generation of homology search tools based on probabilistic inference. *Genome Inform* 23:205–211.
31. Howe K, Bateman A, Durbin R (2002) QuickTree: Building huge Neighbour-Joining trees of protein sequences. *Bioinformatics* 18:1546–1547.

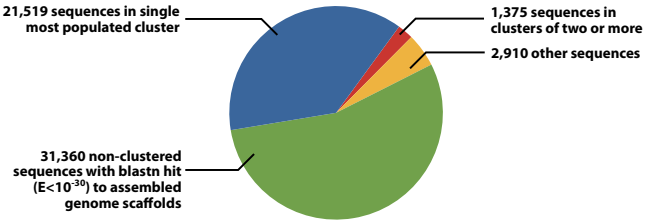

Fig. S1. Genome contigs that were not included in scaffolds largely represent repetitive elements.

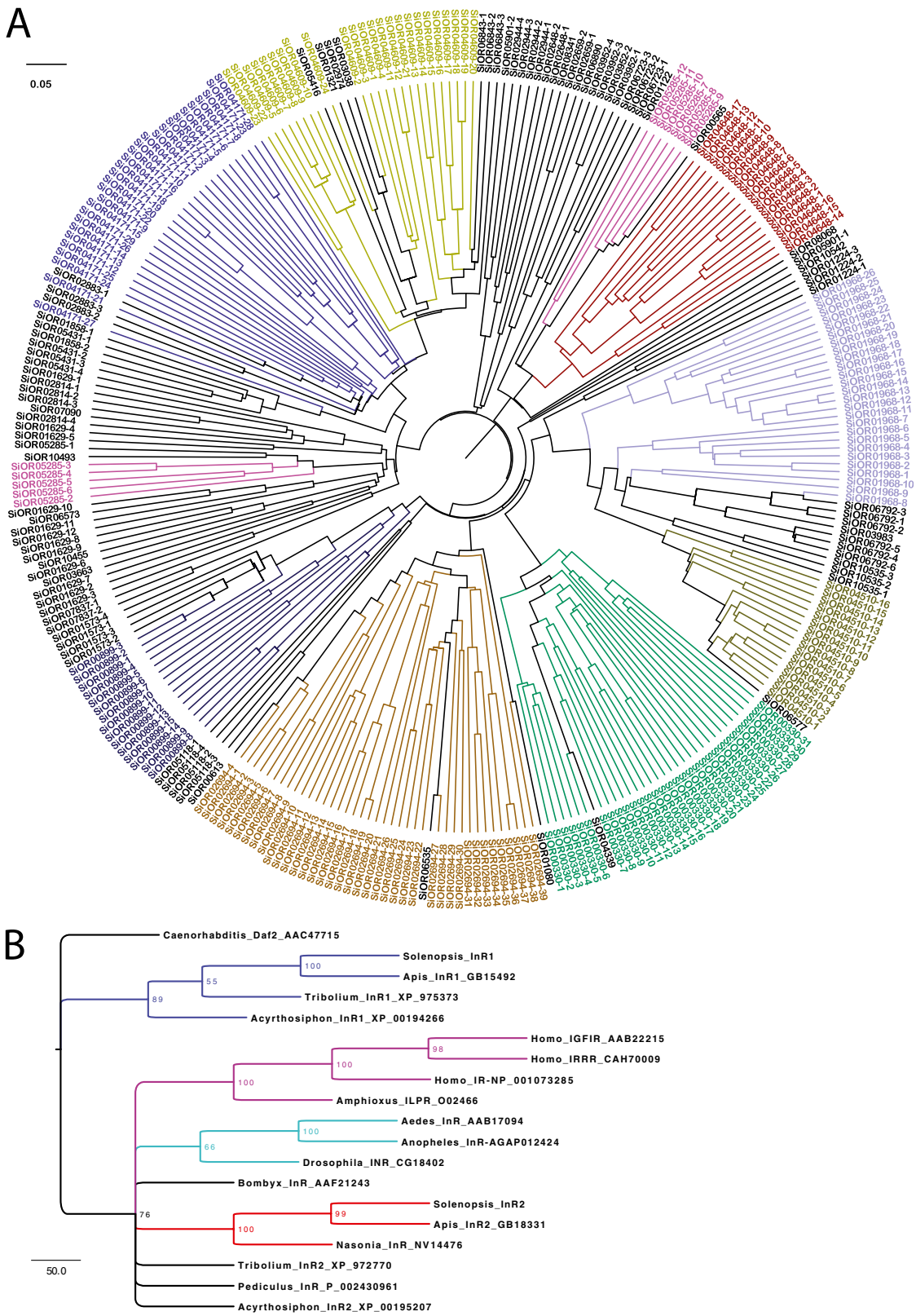

Fig. S2. Phylogenetic trees. (*A*) Phylogenetic relationships between *S. invicta* olfactory receptors (ORs). *Pogonomyrmex barbatus* and *N. vitripennis* OR sequences were aligned with MAFFT (29) and the resulting alignment used to construct an HMM profile (30). This profile was run on the *S. invicta* genome with GeneWise (7) to identify OR proteins. The resulting 475 putative *S. invicta* OR proteins were realigned with the *P. barbatus* ORs. Visual inspection identified 250 putatively complete *S. invicta* ORs. These were aligned with the initial *P. barabatus* and *N. vitripennis* sequences and a new HMM profile was constructed from the alignment. Rerunning GeneWise on the remaining putative *S. invicta* OR regions brought the total number of putatively accurate full-length ORs to 297. These 297 sequences were aligned and then used as input to QuickTree (31) with 1,000 bootstrap samples. The *S. invicta* OR identifiers indicate SignF scaffold number (five digits) and, if a scaffold contains multiple ORs, a unique index number for each OR. Colors highlight scaffolds carrying more than 10 ORs. (*B*) Neighbor-Joining tree of protein sequences of putative insulin receptors from *S. invicta* and other insects (GenBank identifiers shown).

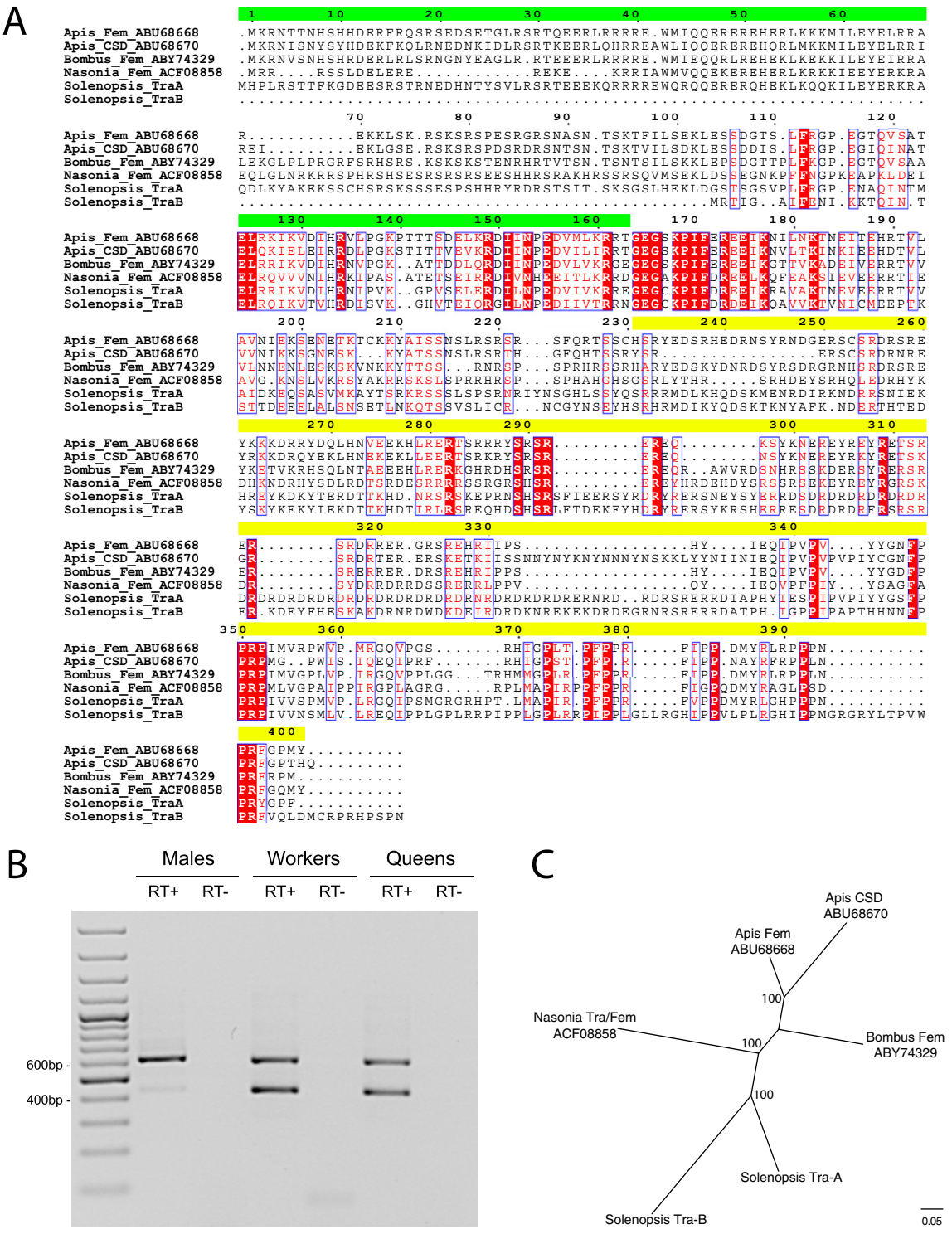

Fig. S3. *Transformer/feminizer* orthologs. (*A*) ClustalW (19) alignment of *transformer/feminizer* orthologs from *A. mellifera*, *B. terrestris*, *N. vitripennis* and *S. invicta*. GenBank accession numbers are provided when available after the second underscore. The *A. mellifera* SDP_N domain is indicated with a green line, and the CSD domain with a yellow line. (*B*) RT-PCR of *S. invicta Tra-B*. RNA was extracted using the Invitrogen TRIzol reagent on three samples: a pool of 7 male pupae, a pool of approximately 30 worker pupae, and a pool of 6 queen pupae. Total RNA was subsequently purified using the Qiagen RNeasy kit. Purified RNA from each sample was subsequently used in parallel reverse transcriptase (RT+) and control (RT−) reactions. Invitrogen SuperScript III reverse transcriptase was used with 9N random primers. PCR amplification with an annealing temperature of 58 °C was subsequently conducted on the three RT+ and three RT− samples with the following primers: 5′-AAGTTTCAGGCTAAATTGATACGTG-3′ and 5′-TGTTCTCTAGAACGCAATCGAATAG-3′. PCR products were resolved by agarose gel electrophoresis. (*C*) A phylogenetic tree was constructed using the alignment from *A* with QuickTree (31) (1,000 bootstrap iterations).

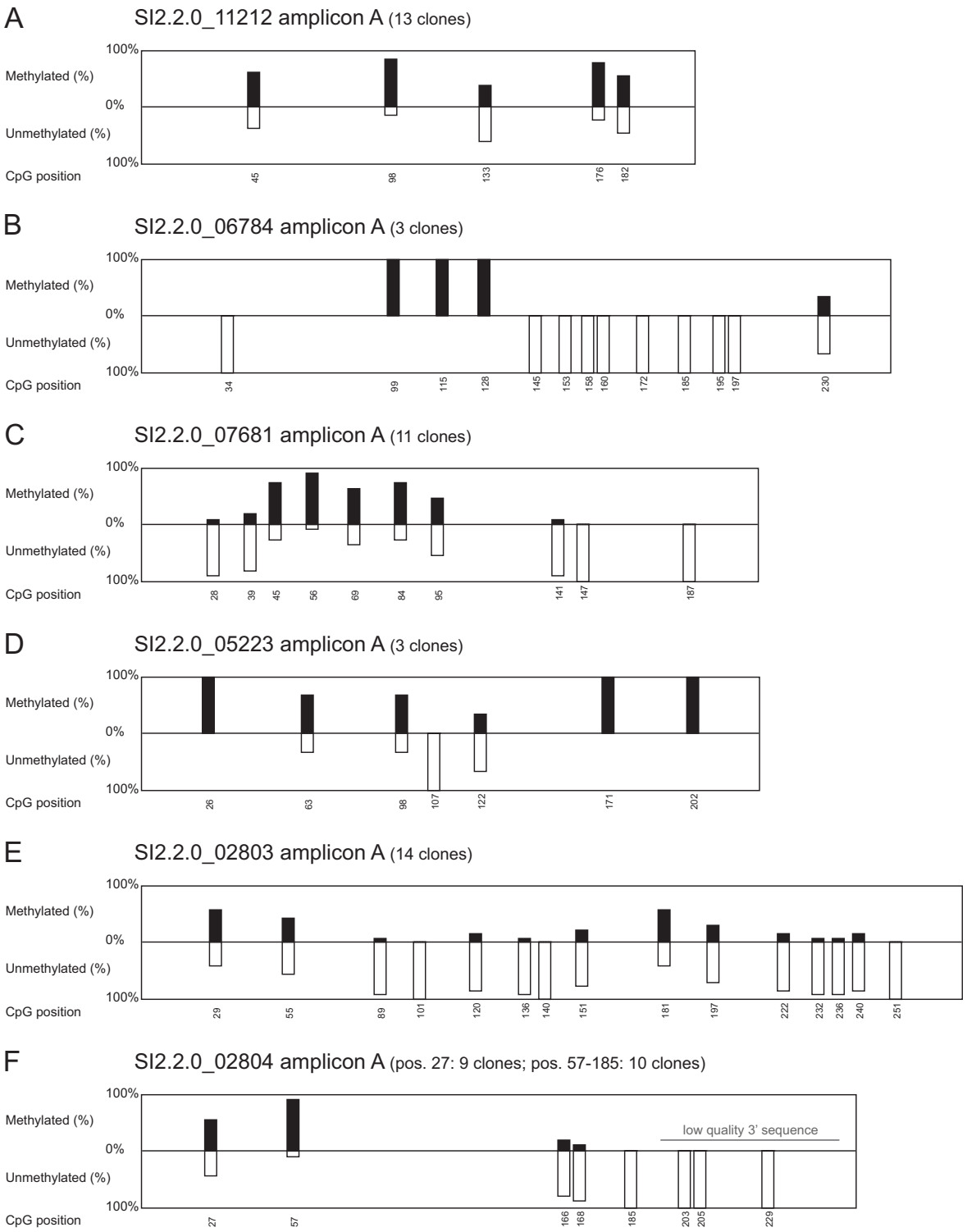

Fig. S4. CpG methylation in *S. invicta* genes is confirmed by the sequencing of bisulfite-converted amplicons. (*A–F*) Proportions of methylated Cs for each CpG site are provided for each gene.

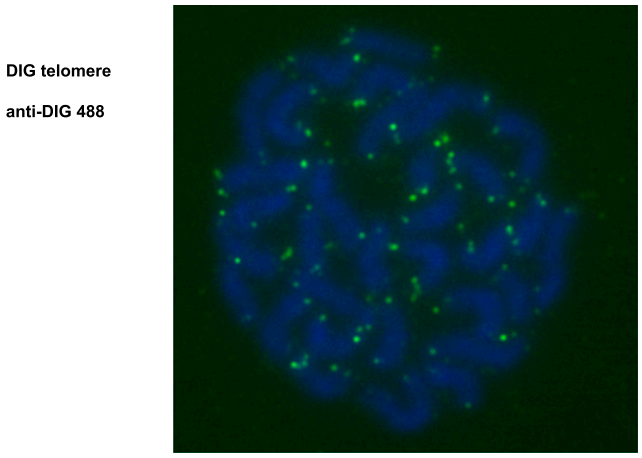

Fig. S5. Strong hybridization of fluorescently labeled $(TTAGG)_n$ probes to chromosome ends in an *S. invicta* larval nucleus suggests that *S. invicta* telomeres consist of $(TTAGG)_n$ repeats.

Table S1*A*. Summary of obtained sequence data

Table S1*B*. Mean $CpG_{O/E}$ according to MeDIP approximately unique read depth

Table S1*C*. Correlation between MeDIP approximately unique read depth and $CpG_{O/E}$

Table S1*D*. Primers used for the amplification of bisulfite-converted DNA

Table S1*E*. Gene Ontology biological process enrichment among putatively methylated genes

Table S1*F*. Comparison of N50 sizes and numbers of contigs and scaffolds larger than N50

Table shows that the three following decisions improve the genome assembly: combining large-insert paired-end libraries, using separately assembled Illumina data, and using stricter-than-default Roche 454 assembly parameters. The assembly used for analyses in the main text is highlighted in blue.

Table S1*G*. Primers used for qRT-PCR of vitellogenins and control genes

Table S1 *A–G* (DOC)

Dataset S1. Counts of PFAM domains identified in *S. invicta*, *N. vitripennis*, *D. melanogaster* and *A. mellifera* highlight several expansions specific to *S. invicta*

Dataset S1 (XLS)