The genome of the fire ant *Solenopsis invicta*

Yannick Wurm[a,b,1], John Wang[a,c], Oksana Riba-Grognuz[a,b], Miguel Corona[a], Sanne Nygaard[d], Brendan G. Hunt[e], Krista K. Ingram[f], Laurent Falquet[b], Mingkwan Nipitwattanaphon[a], Dietrich Gotzek[g], Michiel B. Dijkstra[a], Jan Oettler[h], Fabien Comtesse[a], Cheng-Jen Shih[i], Wen-Jer Wu[i], Chin-Cheng Yang[i], Jerome Thomas[j], Emmanuel Beaudoing[j], Sylvain Pradervand[j], Volker Flegel[b], Erin D. Cook[e], Roberto Fabbretti[b], Heinz Stockinger[b], Li Long[b], William G. Farmerie[k], Jane Oakey[l], Jacobus J. Boomsma[d], Pekka Pamilo[m], Soojin V. Yi[e], Jürgen Heinze[h], Michael A. D. Goodisman[e], Laurent Farinelli[n], Keith Harshman[j], Nicolas Hulo[o], Lorenzo Cerutti[o], Ioannis Xenarios[b,o,2], DeWayne Shoemaker[p,2], and Laurent Keller[a,2]

[a]Department of Ecology and Evolution, University of Lausanne, 1015 Lausanne, Switzerland; [b]Vital-IT Group, Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland; [c]Biodiversity Research Center, Academia Sinica, Nangang Taipei 115, Taiwan; [d]Centre for Social Evolution, Department of Biology, University of Copenhagen, 2100 Copenhagen, Denmark; [e]Georgia Institute of Technology, School of Biology, Atlanta, GA 30332-0230; [f]Department of Biology, Colgate University, Hamilton, NY 13346; [g]Department of Entomology, National Museum of Natural History, Smithsonian Institution, Washington, DC 20013-7012; [h]Biologie I, Universität Regensburg, 93040 Regensburg, Germany; [i]Department of Entomology, National Taiwan University, Taipei 10617, Taiwan; [j]Lausanne Genomic Technologies Facility, University of Lausanne, 1015 Lausanne, Switzerland; [k]Interdisciplinary Center for Biotechnology Research, University of Florida, Gainesville, FL 32610; [l]Biosecurity Queensland, Brisbane QLD 4108, Queensland, Australia; [m]Department of Biosciences, University of Helsinki, 00014, Helsinki, Finland; [n]Fasteris SA, 1228 Plan-les-Ouates, Switzerland; [o]Swiss Institute of Bioinformatics, 1211 Geneva, Switzerland; and [p]US Department of Agriculture Agricultural Research Service, Center for Medical, Agricultural, and Veterinary Entomology, Gainesville, FL 32608

Edited* by Gene E. Robinson, University of Illinois at Urbana–Champaign, Urbana, IL, and approved November 8, 2010 (received for review July 6, 2010)

Ants have evolved very complex societies and are key ecosystem members. Some ants, such as the fire ant *Solenopsis invicta*, are also major pests. Here, we present a draft genome of *S. invicta*, assembled from Roche 454 and Illumina sequencing reads obtained from a focal haploid male and his brothers. We used comparative genomic methods to obtain insight into the unique features of the *S. invicta* genome. For example, we found that this genome harbors four adjacent copies of vitellogenin. A phylogenetic analysis revealed that an ancestral vitellogenin gene first underwent a duplication that was followed by possibly independent duplications of each of the daughter vitellogenins. The vitellogenin genes have undergone subfunctionalization with queen- and worker-specific expression, possibly reflecting differential selection acting on the queen and worker castes. Additionally, we identified more than 400 putative olfactory receptors of which at least 297 are intact. This represents the largest repertoire reported so far in insects. *S. invicta* also harbors an expansion of a specific family of lipid-processing genes, two putative orthologs to the *transformer/feminizer* sex differentiation gene, a functional DNA methylation system, and a single putative telomerase ortholog. EST data indicate that this *S. invicta* telomerase ortholog has at least four spliceforms that differ in their use of two sets of mutually exclusive exons. Some of these and other unique aspects of the fire ant genome are likely linked to the complex social behavior of this species.

social insect | caste differences | nonmodel organism | de novo genome assembly

The major organizing principle of societies of bees, wasps, termites, and ants is a reproductive division of labor, whereby one or a few individuals (the queens and males) specialize in reproduction and the majority of individuals (the workers and soldiers) participate in cooperative tasks such as building the nest, collecting food, rearing the young, and defending the colony. This social organization provides numerous advantages and forms the basis for the tremendous ecological success of social insects. For example, they are found in almost every type of terrestrial environment, make up as much as 50% of animal biomass in some habitats, and play crucial roles as predators, pollinators, and soil processors (1).

In addition to being critically important members of many terrestrial ecosystems, many ant species are also highly successful invasive pests. One such notorious invasive ant species is the fire ant, *Solenopsis invicta* (Fig. 1*A*). This species was inadvertently introduced to the southern United States from South America in the 1930s (2, 3). *S. invicta* is now of profound economic importance, with annual losses to households, businesses, governments, and institutions of $5,000 million across the United States (4). For example, *S. invicta* aggressively uses its very potent sting, inflicting pain and inducing hypersensitivity reactions in humans (Fig. 1*B*). Furthermore, it forms large colonies at high densities, is capable of damaging agricultural machinery, and thus interfering with crop production and harvesting (5, 6). The many existing methods of fire ant control have failed to halt the spread of this exotic species and have hurt its indigenous competitors. There is thus an urgent need to develop effective and safe alternative management techniques (7). Despite extensive quarantine and extermination efforts, *S. invicta* is now also found in many other countries including Australia, China, and Taiwan (8–10).

The past decade has seen the development of several tools for studying ants at the molecular level (11, 12). Although these tools have provided insights into the genetics of caste differentiation (13, 14) and the effects of social context (15, 16), they are somewhat limited, given that they survey only a small subset of the genome. We therefore undertook whole-genome sequencing and de novo assembly of the fire ant *S. invicta*.

Results and Discussion

Genome Assembly. We report the draft sequence of the genome of the fire ant *S. invicta*, obtained by combined Roche 454 and Illumina technologies for a sequencing cost of approximately $230,000. Our assembly strategy was as follows: We first assembled only Illumina 352-bp insert paired-end reads (Table S1*A*) and subsequently chopped up the resulting assembly into sequences equivalent to the length of Roche 454 reads. These artificial reads then were combined with Roche 454 shotgun-sequenced reads,

Author contributions: Y.W., J.W., O.R.-G., M.C., S.N., B.G.H., M.A.D.G., L. Farinelli, N.H., L.C., I.X., D.S., and L.K. designed research; Y.W., J.W., O.R.-G., M.C., S.N., B.G.H., L. Falquet, F.C., E.D.C., L. Farinelli, N.H., L.C., and I.X. performed research; Y.W., J.W., O.R.-G., M.C., S.N., B.G.H., K.K.I., L. Falquet, M.N., M.B.D., J. Oettler, C.-J.S., W.-J.W., C.-C.Y., J.T., E.B., S.P., V.F., R.F., H.S., L.L., W.G.F., J. Oakey, J.J.B., P.P., S.V.Y., J.H., M.A.D.G., L. Farinelli, K.H., N.H., L.C., I.X., D.S., and L.K. contributed new reagents/analytic tools; Y.W., J.W., O.R.-G., M.C., S.N., B.G.H., K.K.I., L. Falquet, D.G., N.H., L.C., and I.X. analyzed data; and Y.W., O.R.-G., M.C., S.N., B.G.H., K.K.I., N.H., L.C., D.S., and L.K. wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

Data deposition: The sequences reported in this paper have been deposited at the National Center for Biotechnology Information under Genome Project ID 49629.

[1]To whom correspondence should be addressed. E-mail: yannick.wurm@unil.ch.

[2]I.X., D.S., and L.K. contributed equally to this work.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1009690108/-/DCSupplemental.

Fig. 1. (*A*) *S. invicta* males (larger, with wings) depart on mating flight while workers (smaller, wingless) patrol (photo by Yannick Wurm). (*B*) A fire ant researcher was stung by his study subject (photo by Daniel P. Wojcik, US Department of Agriculture Agricultural Research Service).

resulting in 352.7 Mb of assembled data split among 90,231 contigs with an N50 size of 14,674 bp (N50 is the length such that 50% of the assembled sequence lies in blocks of length N50 or greater). Using 8- and 20-kb insert paired-end Roche 454 reads, 31,250 of the contigs were joined to form 10,543 scaffolds with an N50 size of 720,578 bp (Table 1 and Table S1*F*). These scaffolds represent a total of 352.7 Mb of sequence including 41.3 Mb of undetermined "N" bases found within scaffolds between consecutive contigs. The remaining 58,981 contigs that could not be placed in scaffolds represent a total of 43.4 Mb of sequence and are significantly shorter than those that were placed in scaffolds (maximum size: 2,002 bp). Among these nonscaffolded contigs, 95% were clustered to each other by blastclust as having more than 50% sequence identity over half the sequence length or very significant similarity (blastn $E < 10^{-30}$) to one or more genome scaffolds. These contigs likely are highly repetitive elements (Fig. S1), consistent with the estimation that 23% of the *S. invicta* genome consists of highly repetitive or foldback elements (17).

We assessed the accuracy and completeness of the *S. invicta* assembly by comparing it with an independently sequenced and assembled set of ESTs putatively representing 12,488 genes. Among these putative gene transcripts, 98.2% mapped to the genome assembly (blastn $E < 10^{-50}$). A total of 580 putative transcripts contained two nonoverlapping 200-bp regions that mapped to two different scaffolded contigs. In the 383 cases in which exons from the same putative gene mapped to different contigs within the same scaffold, scaffolding of contigs was consistent with exon order and orientation. In the remaining 197 cases, putative exons mapped to contigs from different scaffolds. We manually inspected 50 of these to determine whether there was evidence for the scaffolds overlapping and whether the 8- and 20-kb insert libraries provided evidence for the scaffolds being adjacent. In 46 of the 50 cases, the scaffolds were probably either adjacent or the smaller scaffold filled a gap in the larger scaffold. Inconsistent mapping of the four remaining putative transcripts possibly reflected problems with EST assembly because three of the four transcripts had highly significant blast similarity to at least two normally unrelated genes. Overall, these results confirm that the genome assembly is essentially complete for gene content and that scaffolding is reliable.

Although the scaffolded *S. invicta* sequence represents 352.7 Mb, unresolved repeats bring the Roche 454 Newbler software estimated genome size to 484.2 Mb. This difference is likely due to the difficulty of resolving repeats. Three conflicting estimates of the haploid genome size of *S. invicta* have been reported previously: 591 Mb (0.62 pg) via reassociation kinetics (17), 753.3 Mb via flow cytometry (18), and 463 Mb via Feulgen image analysis densitometry (0.47 pg reported in ref. 19). The latter estimate is most consistent with our results. Discrepancies in genome size estimations have been previously reported (19, 20). Such variation may be due to technical issues, differences in examined cell types, endoparasitic load, transposon activity, or possibly other genetic differences between individuals or populations.

Gene Content. A combination of ab initio, EST-based, and sequence similarity-based methods generated an official gene set of 16,569 protein-coding genes. There were significant differences in the guanine-cytosine contents of exons (45.0%), introns (30.9%), the 2,000-bp surrounding genes (1,000 bp up- and downstream; 33.5%), and the genome in general (36.1%) (*t* tests, Bonferonni-corrected, all *P* values $<10^{-10}$). These results are similar to those in the honey bee (21). Blastp search of fire ant proteins against protein databases indicate that 47% of *S. invicta* genes have the strongest similarity to apoid sequences, and another 22% have the strongest similarity to *Nasonia* (Fig. 2), which is consistent with ants being more closely related to bees than to chalcidoid wasps (23). An additional 13% of *S. invicta* genes have the highest similarity to nonhymenopteran sequences, suggesting that they may be evolving slowly in *S. invicta* or have been lost in *Apis mellifera* and *Nasonia*. Finally, 18% of *S. invicta* proteins have no significant similarity ($E > 10^{-5}$) to non-*Solenopsis* sequences in the GenBank nonredundant protein database (25), suggesting that they may be fast-evolving or ant-specific. Similarly, 17% of the proteins in the *Nasonia vitripennis* official gene set have no significant similarity to non-*Nasonia* sequences in the nonredundant protein database.

Functional Categories. *S. invicta* appears to harbor a typical insect gene repertoire. For example, the *S. invicta* genome includes a complete set of small RNA-processing genes with orthologs to *Argonaute*, *Drosha*, *Pasha*, *Dicer-1*, *Dicer-2*, *Loquacious*, and *R2D2*. Domain analyses of the *S. invicta*, *N. vitripennis*, *Drosophila melanogaster*, and *A. mellifera* proteomes reveal several putative gene duplications in fire ants (Dataset S1). We highlight here the

Table 1. Genome assembly statistics

	Scaffolds*	Scaffolded contigs	Nonscaffolded contigs	All contigs
Number	10,543	31,250	58,981	90,231
N50 size (bp)	720,578	18,166	983	14,674
Maximum size (bp)	6,355,204	192,021	2,002	192,021
Mean size (bp)	33,452	9,965	735	3,931
Minimum size (bp)	1,997	397	200	200
Total consensus (bp)	352,687,102	311,407,343	43,332,432	354,739,775

*Scaffolds include gaps between adjacent contigs. The estimated lengths of these gaps are included in scaffold size estimations. True sizes likely are slightly different.

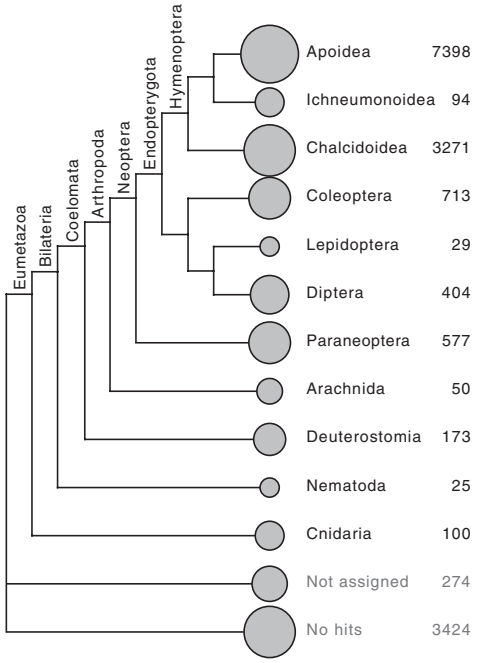

Fig. 2. Taxonomic distribution of best blastp hits of *S. invicta* proteins to the nonredundant (nr) protein database ($E < 10^{-5}$). Results were first plotted using MEGAN software (22) and then branches with fewer than 20 hits were removed, branch lengths were reduced for compactness, and tree topology was adjusted to reflect consensus phylogenies (23, 24).

significance of these duplication events in vitellogenins, odor perception genes, and a family of lipid-processing genes. We also discuss additional features of interest in the fire ant genome relevant to the complex social biology of this species, including sex determination genes, DNA methylation genes, telomerase, and the insulin and juvenile hormone pathways.

Vitellogenins. In contrast to other insects that mainly have only one or two vitellogenins, the fire ant genome harbors four adjacent copies of vitellogenin (*Vg1-4*) (Fig. 3*A*), all of which are at least partially supported by EST reads. A phylogenetic analysis reveals that an ancestral vitellogenin gene first underwent duplication, followed by possibly independent duplications of each of the daughter vitellogenins, thus giving rise to *Vg1* and *Vg4* and to *Vg2* and *Vg3*. All of these duplications occurred after the ancestor of fire ants split from wasps and bees (Fig. 3*B*). The single vitellogenin found in *A. mellifera* is a multifunctional protein (26) involved in the regulation of life span (27, 28) and division of labor (29). Quantitative RT-PCR shows that *Vg1* and *Vg4* are preferentially expressed in workers and *Vg2* and *Vg3* in queens (Fig. 3*C*, *SI Materials and Methods*, and Table S1*G*). Vitellogenin expression in *S. invicta* workers is surprising because they lack ovaries. Given the superorganism properties of ant societies, the expression patterns suggest that vitellogenins underwent neo- or subfunctionalization after duplication to acquire caste-specific functions.

Odor Perception. Consistent with studies in other insects, we find a single *S. invicta* ortholog to *DmOr83b*, a broadly expressed olfactory receptor (OR) required to interact with other ORs for *Drosophila* and *Tribolium castaneum* olfaction (30–32). Beyond *OR83b*, OR number varies greatly between insect species. Blast searches and GeneWise searches using an HMM profile constructed with aligned ORs from *N. vitripennis* (33) and *Pogonomyrmex barbatus* identified more than 400 loci in the *S. invicta* genome with significant sequence similarity to ORs. Preliminary work on gene model reconstruction identified 297 intact full-length proteins. Many *S. invicta* ORs are in tandem arrays (Fig. S2*A*) and derive from recent expansions. *S. invicta* may thus harbor the largest identified insect OR repertoire because there are 10 ORs in *Pediculus humanus* (34), 60 in *Drosophila*, 165 in *A. mellifera*, 225 in *N. vitripennis* (33), and 259 in *T. castaneum* (32). The large numbers of *N. vitripennis* and *T. castaneum* ORs are thought to be due to current or past difficulties in host and food finding. As has been suggested for *A. mellifera* (35), the large number of *S. invicta* ORs may result from the importance of chemical communication in ants. The odorant-binding proteins (OBPs) are another family of genes also known to play roles in chemosensation in *Drosophila* (36). Intriguingly, the social organization of *S. invicta* colonies is completely associated with se-

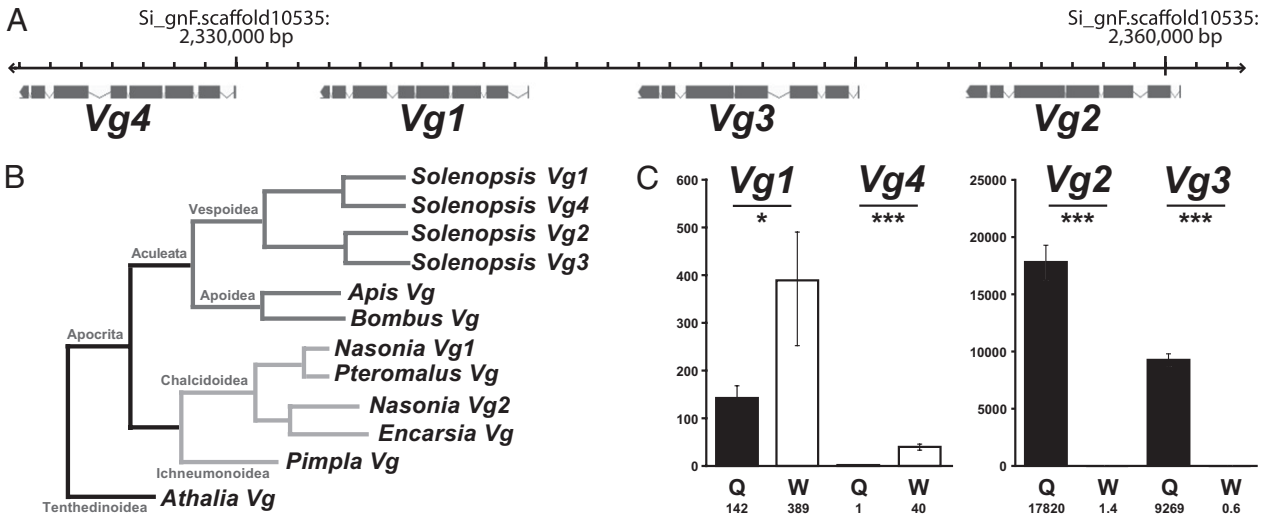

Fig. 3. *S. invicta* vitellogenins. (*A*) Four vitellogenins are located within a single 40,000-bp region of the *S. invicta* genome. (*B*) Parsimony tree of known hymenopteran vitellogenin protein sequences suggests that two rounds of vitellogenin duplication occurred after the split between ants and other hymenopterans including bees and wasps. (*C*) Quantitative RT-PCR of the four putative *S. invicta* vitellogenins on whole bodies of major workers (W) and mated queens (Q) ($n = 10$). The *y* axis indicates mRNA concentrations for the different vitellogenins. Values depicted by each bar are shown below the *x*-axis labels. Error bars represent SEs. Expression differences between queens and workers were significant (Bonferroni-corrected two-tailed *t* tests: $*P < 0.05$, $***P < 10^{-10}$).

quence variation at the OBP gene *Gp-9* (37, 38). We find 12 OBP domains in the *S. invicta* genome, 2 of which are differentially expressed between workers of alternate *Gp-9* genotypes (15). Further analyses will be required to determine the extent to which these genes are directly involved in the morphological and behavioral differences between queens and workers of the two alternate social organizations of *S. invicta*.

Lipid Metabolism. An unusually high number of genes in *S. invicta* have the following protein domains related to fatty-acid metabolism: Ketoacyl-synt (PF00109), Ketoacylsynt_C (PF02801), and Acyl_transf_1 (PF00698) (Dataset S1). Although some are likely pseudogenes, nine *S. invicta* genes are complete and carry all three domains. In comparison, *A. mellifera* and *D. melanogaster* have only two such genes whereas *N. vitripennis* has six (respective odds ratios: 2.3, 3.4, and 0.8). The expansion of fatty-acid metabolism-related genes in *S. invicta* could stem from the fact that young *S. invicta* queens accumulate as much as 60% of their body mass in the form of lipids within the 2 weeks following eclosion from the pupae (39) as a means of rearing a first worker brood for the duration of a claustral phase during which queens neither feed nor forage (40). Alternatively, such lipid-processing genes may help produce the cuticular hydrocarbons that are involved in kin recognition in ants (41).

Sex Determination. Hymenopterans, including wasps, bees, and ants, have a haplo-diploid sex determination system whereby diploid eggs normally develop into females, and haploid eggs develop into males. In *N. vitripennis*, female development is initiated by maternally derived *transformer/feminizer* mRNA in the zygote (42). In contrast, sex is determined by the *complementary sex determiner* (*csd*) gene in *A. mellifera* (43, 44): Eggs that are heterozygous at this locus develop into females, whereas hemizygous haploids and homozygous diploids develop into males. The *csd* gene is thought to be a recent *Apis* innovation (43), having arisen through a duplication of the *transformer/feminizer* gene. The sex determination mechanism in ants is unknown, but the occurrence of diploid males in some *S. invicta* populations suggests a *csd*-like mechanism (45, 46). The genome of *S. invicta* contains two linked sequences with similarity to *transformer/feminizer* (Fig. S3*A*), but unlike the *A. mellifera* sex-determining locus, the *S. invicta* genes are coded on opposite strands. Members of the Apis transformer/feminizer protein family contain two distinct domains: An N-terminal SDP_N domain and a C-terminal Apis_CSD domain. One of the *S. invicta* sequences (*Tra-A*) contains both domains (SDP_N: $E = 3.5 \times 10^{-11}$; Apis_CSD: $E = 1.4 \times 10^{-7}$). The other (*Tra-B*) contains a partial SDP_N domain ($E = 9.5 \times 10^{-4}$) and a nonsignificant match to Apis_CSD. Alternative splicing of *transformer/feminizer* mRNA is known to play a crucial role in sex determination for many insects (47). Intriguingly, the *S. invicta Tra-B* transcript appears to have two different spliceforms, with only one spliceform including the SDP_N domain. This longer form appears to be the dominant transcript in males, whereas both forms are equally expressed in queens and workers (Fig. S3*B*). A phylogenetic analysis of the *transformer/feminizer* homologs from several hymenopterans shows that the *S. invicta* sequences cluster together, consistent with independent *transformer/feminizer* duplication in the ant and honey bee lineages (Fig. S3*C*).

Methylation. *S. invicta* harbors a complete set of genes known to be involved in DNA methylation, maintenance of methylation patterns, and tRNA methylation in eukaryotes, including DNMT3, DNMT1, and TRDMT1 (previously known as DNMT2) (48). A negative correlation between $CpG_{O/E}$, a statistical measure of DNA methylation, and enrichment of sequence obtained after methylated DNA immunoprecipitation (MeDIP) from a pool of queen and worker prepupae suggested the existence of functional methylation in *S. invicta* (Table S1 *B* and *C*). DNA methylation was confirmed by sequencing of bisulfite-converted genomic DNA from nine genes (*SI Materials and Methods*, Fig. S4, Table S1*D*). DNA methylation is hypothesized to play a key role in developmental responsiveness to environmental factors and may play an important role in social insect caste determination (49). However, the primary targets of DNA methylation in insects appear to be genes with ubiquitous expression across tissues and alternate phenotypes (50–52). Putatively methylated genes identified from MeDIP analysis in *S. invicta* were enriched for biological processes related to cellular metabolism and transcription (*SI Materials and Methods*, Table S1*E*), as is the case with methylated genes in *A. mellifera* (50).

Telomerase Reverse Transcriptase. Ants show remarkable intraspecific life-span variation with queens of some species living to the astonishing age of more than 20 y and workers typically dying within several months or at most a few years and males within a few months (53). Aging is associated with a decline in telomere repair and consequent telomere shortening (54) in many animals. Similarly, in the long-lived ant *Lasius niger* somatic tissues of the short-lived males have dramatically shorter telomeres than those of the much longer-lived queens and workers (55). The telomere sequence of most ants and other insects is composed of TTAGG repeats (56, 57). Consistent with this, the ends of *S. invicta* chromosomes showed strong hybridization signal to labeled $(TTAGG)_n$ probe (Fig. S5). Furthermore, *S. invicta* sequences harbor many more degenerate TTAGG repeats than vertebrate-like TTAGGG repeats. Finally, in contrast with dipteran insect species that lack telomerase, but similarly to other nondipteran insect species whose genomes have been sequenced, *S. invicta* has a single putative telomerase ortholog with RNA-binding (TRBD) and reverse transcriptase (TERT) domains. Interestingly, EST data derived from mixed-stage, mixed-caste, whole-body libraries indicate that this *S. invicta* telomerase ortholog has at least four strongly supported spliceforms that differ in their use of two sets of mutually exclusive exons. These alternative spliceforms may permit fine-tuning of telomerase activity, perhaps in caste- or tissue-specific manners.

Insulin/Insulin-Like Growth Factor Signaling. Insulin and insulin-like growth factor (IGF) signaling is a key integrative pathway regulating aging and fertility in animals (58, 59). In *A. mellifera*, insulin and IGF signaling are involved in the regulation of caste determination (60, 61), division of labor (62), and queen longevity (28) and may play similar roles in other social insects. The family of insulin-like peptides (ILPs) underwent many clade and species-specific duplications, leading to 37 members in *Caenorhabditis elegans*, 27 in *Bombyx mori*, and 7 in *Drosophila* and *Anopheles*. In contrast, *S. invicta* and *A. mellifera* have only two orthologous ILPs, one of which also occurs in *N. vitripennis* (Fig. S2*B*). Both *S. invicta* and *A. mellifera* also have two insulin/IGF1 receptors. Phylogenetic analyses suggest that these two receptors resulted from an ancient duplication with subsequent losses in Diptera and *Nasonia*. Interestingly, we find that one of the putative insulin/IGF1 receptors belongs to our list of genes putatively subjected to dense methylation in *S. invicta* (*SI Materials and Methods*).

Juvenile Hormone. Juvenile Hormone (JH) regulates larval molting and reproductive development in many insects (63). Increases in JH titer have also been shown to induce soldier-caste differentiation in termites (64) and behavioral changes in *A. mellifera* workers (65, 66). Interestingly, *S. invicta* harbors >30 putative juvenile hormone binding protein (JHBPs; PF06585, Dataset S1) encoding genes, more than half of which are located in a single 600,000-bp region. By contrast, there are only 16 such JHBP domains in *Nasonia* and 19 in *A. mellifera*. Similarly, the number of genes that encode juvenile hormone epoxide hydrolases (JHEHs), enzymes that degrade JH, is much higher in *S. invicta* than in

A. mellifera (one) and *Nasonia* (two). Four of the six *S. invicta* JHEH encoding genes are adjacent, suggesting recent duplications. Because JH titer determines fecundity of *S. invicta* queens (65), the expansions of both JHBP and JHEH gene families in *S. invicta* may reflect strong selection occurring after the death of the mother queen with many nonreproductive queens competing to reproduce quickly and become "replacement" queens (40, 67).

In conclusion, this study reveals that a combination of Roche 454 and Illumina sequencing can be used to obtain a good quality genome even when the genome is relatively large and contains a high proportion of repetitive elements. Comparison with other genomes shows that the fire ant genome has many unique properties probably associated with the complex social life of this species. Finally, the sequencing of the fire ant genome provides the foundation for future evolutionary, biomedical, sociogenetic, and pest-management studies of this important pest species and facilitates comparisons with other social species.

Materials and Methods

Computation was performed at the Vital-IT (http://www.vital-it.ch) center for high-performance computing of the Swiss Institute of Bioinformatics. Analyses were assisted by custom Ruby/Bioruby (68, 69), Perl (70), and R (71) scripts. The details of the sequencing, assembly, annotation, and analyses are given in *SI Materials and Methods*.

ACKNOWLEDGMENTS. We thank Christophe Lucas and members of the Marc Robinson-Rechavi laboratory for advice and discussion; Catherine Berney and Christine La Mendola for help in the laboratory; Kenneth G. Ross (Department of Entomology, University of Georgia, Athens, GA), for providing the ants; Jérôme Notari, Laélia Maumary, and Jézaëlle Rufener for help rearing the ants; and Hugh M. Robertson for *Pogonomyrmex* olfactory receptor sequences. We thank Timothy Harkins, Pascal Bouffard, Linda Burton and 454 Life Sciences, a Roche Company (Branford, CT) for sequencer access, construction of 454 sequencing libraries and access to Newbler software, the Interdisciplinary Center for Biotechnology Research at the University of Florida for Roche 454 sequencing, and the Lausanne Genomic Technologies Facility and Fasteris SA for Illumina sequencing. This work was supported by grants from the European Union-funded Network of Excellence Lifespan FP6 036894, the Swiss National Science Foundation, the University of Lausanne, the Rectorate of the University of Lausanne, an ERC Advanced grant and the Herbette Foundation (to L.K.); US Department of Agriculture Agriculture and Food Research Initiative Award 2009-35302-05301 (to D.S. and J.W.); by an Infectigen grant (to L.K., D.S., and Y.W.); the Academy of Finland Grant 122210 (to P.P.); grants from the European Union-funded Network of Excellence Lifespan FP6 036894, the Swiss National Science Foundation, the University of Lausanne, and the Rectorate of the University of Lausanne (to L.K.); by National Science Foundation Grants DEB-0640690 (to M.A.D.G. and S.V.Y.), MCB-0950896 (to S.V.Y. and M.A.D.G.), and DEB-1011349 (to B.G.H., S.V.Y., and M.A.D.G.); by grants from the Danish National Research Foundation (to S.N. and J.J.B.); by Grant 98AS-9.1.2-BQ-B1(1) from the Bureau of Animal and Plant Health Inspection and Quarantine, Council of Agriculture, Executive Yuan, Taiwan (to W.-J.W.); by the Swiss National Science Foundation project Grant 315230B-131083 (to L.C. and N.H.); and by the Picker Interdisciplinary Science Institute, Colgate University (K.K.I.).

1. Hölldobler B, Wilson EO (1990) *The Ants* (The Belknap Press of Harvard University Press, Cambridge, MA).
2. Shoemaker DD, DeHeer C, Krieger MJB, Ross KG (2006) Population genetics of the invasive fire ant *Solenopsis invicta* in the U.S.A. *Ann Entomol Soc Am* 99:1213–1233.
3. Ross KG, Shoemaker DD (2008) Estimation of the number of founders of an invasive pest insect population: The fire ant *Solenopsis invicta* in the USA. *Proc Biol Sci* 275: 2231–2240.
4. McDonald M (2006) Reds under your feet. *New Sci* 189:50–51.
5. Lard CF, et al. (2006) An economic impact of imported fire ants in the United States of America. PhD thesis (Texas A&M University, College Station, TX).
6. Vinson SB (1986) *Economic Impact and Control of Social Insects* (Praeger, New York).
7. Williams DF, deShazo RD (2004) Biological control of fire ants: An update on new techniques. *Ann Allergy Asthma Immunol* 93:15–22.
8. Henshaw MT, Kunzmann N, Vanderwoude C, Sanetra M, Crozier RH (2005) Population genetics and history of the introduced fire ant, *Solenopsis invicta* Buren (Hymenoptera: Formicidae), in Australia. *Aust J Entomol* 44:37–44.
9. Yang CC, Shoemaker DD, Wu WJ, Shih CJ (2008) Population genetic structure of the red imported fire ant, *Solenopsis invicta*, in Taiwan. *Insectes Soc* 55:54–65.
10. Xu Y-J, Huang J, Lu Y-Y, Zeng L, Liang G-W (2009) Observation of nuptial flights of the red imported fire ant, *Solenopsis invicta* (Hymenoptera: Formicidae) in mainland China. *Sociobiology* 54:831–840.
11. Wurm Y, et al. (2009) Fourmidable: A database for ant genomics. *BMC Genomics* 10:5.
12. 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.
13. Goodisman MA, Isoe J, Wheeler DE, Wells MA (2005) Evolution of insect metamorphosis: A microarray-based study of larval and adult gene expression in the ant *Camponotus festinatus*. *Evolution* 59:858–870.
14. Gräff J, Jemielity S, Parker JD, Parker KM, Keller L (2007) Differential gene expression between adult queens and workers in the ant *Lasius niger*. *Mol Ecol* 16:675–683.
15. Wang J, Ross KG, Keller L (2008) Genome-wide expression patterns and the genetic architecture of a fundamental social trait. *PLoS Genet* 4:e1000127.
16. Wurm Y, Wang J, Keller L (2010) Changes in reproductive roles are associated with changes in gene expression in fire ant queens. *Mol Ecol* 19:1200–1211.
17. Li J, Heinz KM (2000) Genome complexity and organization in the red imported fire ant *Solenopsis invicta* Buren. *Genet Res* 75:129–135.
18. Johnston JS, Ross LD, Beani L, Hughes DP, Kathirithamby J (2004) Tiny genomes and endoreduplication in Strepsiptera. *Insect Mol Biol* 13:581–585.
19. Ardila-Garcia AM, Umphrey GJ, Gregory TR (2010) An expansion of the genome size dataset for the insect order Hymenoptera, with a first test of parasitism and eusociality as possible constraints. *Insect Mol Biol* 19:337–346.
20. Tsutsui ND, Suarez AV, Spagna JC, Johnston JS (2008) The evolution of genome size in ants. *BMC Evol Biol* 8:64.
21. Honeybee Genome Sequencing Consortium (2006) Insights into social insects from the genome of the honeybee *Apis mellifera*. *Nature* 443:931–949.
22. Huson DH, Auch AF, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. *Genome Res* 17:377–386.
23. Sharkey MJ (2007) Phylogeny and classification of Hymenoptera. *Zootaxa* 1668: 521–548.
24. Wiegmann BM, et al. (2009) Single-copy nuclear genes resolve the phylogeny of the holometabolous insects. *BMC Biol* 7:34.
25. Pruitt KD, Tatusova T, Maglott DR (2005) NCBI Reference Sequence (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* 33(Database issue):D501–D504.
26. Amdam GV, Norberg K, Hagen A, Omholt SW (2003) Social exploitation of vitellogenin. *Proc Natl Acad Sci USA* 100:1799–1802.
27. Seehuus SC, Norberg K, Gimsa U, Krekling T, Amdam GV (2006) Reproductive protein protects functionally sterile honey bee workers from oxidative stress. *Proc Natl Acad Sci USA* 103:962–967.
28. Corona M, et al. (2007) Vitellogenin, juvenile hormone, insulin signaling, and queen honey bee longevity. *Proc Natl Acad Sci USA* 104:7128–7133.
29. Nelson CM, Ihle KE, Fondrk MK, Page RE, Amdam GV (2007) The gene vitellogenin has multiple coordinating effects on social organization. *PLoS Biol* 5:e62.
30. Larsson MC, et al. (2004) Or83b encodes a broadly expressed odorant receptor essential for *Drosophila* olfaction. *Neuron* 43:703–714.
31. Benton R, Sachse S, Michnick SW, Vosshall LB (2006) Atypical membrane topology and heteromeric function of *Drosophila* odorant receptors *in vivo*. *PLoS Biol* 4:e20.
32. Engsontia P, et al. (2008) The red flour beetle's large nose: An expanded odorant receptor gene family in *Tribolium castaneum*. *Insect Biochem Mol Biol* 38:387–397.
33. Robertson HM, Gadau J, Wanner KW (2010) The insect chemoreceptor superfamily of the parasitoid jewel wasp *Nasonia vitripennis*. *Insect Mol Biol* 19(Suppl 1):121–136.
34. Kirkness EF, et al. (2010) Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle. *Proc Natl Acad Sci USA* 107:12168–12173.
35. Robertson HM, Wanner KW (2006) The chemoreceptor superfamily in the honey bee, *Apis mellifera*: Expansion of the odorant, but not gustatory, receptor family. *Genome Res* 16:1395–1403.
36. Xu P, Atkinson R, Jones DNM, Smith DP (2005) *Drosophila* OBP LUSH is required for activity of pheromone-sensitive neurons. *Neuron* 45:193–200.
37. Ross KG, Keller L (1998) Genetic control of social organization in an ant. *Proc Natl Acad Sci USA* 95:14232–14237.
38. Keller L, Ross KG (1998) Selfish genes: A green beard in the red fire ant. *Nature* 394: 573–575.
39. Keller L, Passera L (1989) Size and fat content of gynes in relation to the mode of colony founding in ants (Hymenoptera; Formicidae). *Oecologia* 80:236–240.
40. Tschinkel WR (2006) *The Fire Ants* (The Belknap Press of Harvard University Press, Cambridge, MA).
41. Blomquist GJ, Bagnères AG (2010) Insect Hydrocarbons: Biology, Biochemistry, and Chemical Ecology (Cambridge University Press, Cambridge, UK).
42. Verhulst EC, Beukeboom LW, van de Zande L (2010) Maternal control of haplodiploid sex determination in the wasp *Nasonia*. *Science* 328:620–623.
43. Hasselmann M, et al. (2008) Evidence for the evolutionary nascence of a novel sex determination pathway in honeybees. *Nature* 454:519–522.
44. Gempe T, et al. (2009) Sex determination in honeybees: Two separate mechanisms induce and maintain the female pathway. *PLoS Biol* 7:e1000222.
45. Glancey BM, Romain MKS, Crozier RH (1976) Chromosome numbers of the red and black imported fire ants, *Solenopsis invicta* and *S. richteri*. *Ann Entomol Soc Am* 69: 469–470.
46. Ross KG, Fletcher DJC (1985) Genetic origin of male diploidy in the fire ant, *Solenopsis invicta* (Hymenoptera: Formicidae), and its evolutionary significance. *Evolution* 39: 888–903.
47. Verhulst EC, van de Zande L, Beukeboom LW (2010) Insect sex determination: It all evolves around *transformer*. *Curr Opin Genet Dev* 20:376–383.
48. Jurkowski TP, et al. (2008) Human DNMT2 methylates tRNA(Asp) molecules using a DNA methyltransferase-like catalytic mechanism. *RNA* 14:1663–1670.
49. Kucharski R, Maleszka J, Foret S, Maleszka R (2008) Nutritional control of reproductive status in honeybees via DNA methylation. *Science* 319:1827–1830.

50. 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.
51. Foret S, Kucharski R, Pittelkow Y, Lockett GA, Maleszka R (2009) Epigenetic regulation of the honey bee transcriptome: Unravelling the nature of methylated genes. *BMC Genomics* 10:472.
52. Hunt BG, Brisson JA, Yi SV, Goodisman MAD (2010) Functional conservation of DNA methylation in the pea aphid and the honeybee. *Genome Biol Evol* 2:719–728.
53. Keller L, Genoud M (1997) Extraordinary lifespans in ants: A test of evolutionary theories of ageing. *Nature* 389:958–960.
54. Mariani E, et al. (2003) Different rates of telomere shortening and telomerase activity reduction in CD8 T and CD16 NK lymphocytes with ageing. *Exp Gerontol* 38:653–659.
55. Jemielity S, et al. (2007) Short telomeres in short-lived males: What are the molecular and evolutionary causes? *Aging Cell* 6:225–233.
56. Lorite P, Carrillo JA, Palomeque T (2002) Conservation of (TTAGG)(n) telomeric sequences among ants (Hymenoptera, Formicidae). *J Hered* 93:282–285.
57. Frydrychová R, Grossmann P, Trubac P, Vítková M, Marec F (2004) Phylogenetic distribution of TTAGG telomeric repeats in insects. *Genome* 47:163–178.
58. Finch CE, Ruvkun G (2001) The genetics of aging. *Annu Rev Genomics Hum Genet* 2: 435–462.
59. Dillin A, Crawford DK, Kenyon C (2002) Timing requirements for insulin/IGF-1 signaling in *C. elegans*. *Science* 298:830–834.
60. Wheeler DE, Buck N, Evans JD (2006) Expression of insulin pathway genes during the period of caste determination in the honey bee, *Apis mellifera*. *Insect Mol Biol* 15: 597–602.
61. de Azevedo SV, Hartfelder K (2008) The insulin signaling pathway in honey bee (*Apis mellifera*) caste development: Differential expression of insulin-like peptides and insulin receptors in queen and worker larvae. *J Insect Physiol* 54:1064–1071.
62. Ament SA, Corona M, Pollock HS, Robinson GE (2008) Insulin signaling is involved in the regulation of worker division of labor in honey bee colonies. *Proc Natl Acad Sci USA* 105:4226–4231.
63. Riddiford LM (2008) Juvenile hormone action: A 2007 perspective. *J Insect Physiol* 54: 895–901.
64. Scharf ME, Buckspan CE, Grzymala TL, Zhou X (2007) Regulation of polyphenic caste differentiation in the termite *Reticulitermes flavipes* by interaction of intrinsic and extrinsic factors. *J Exp Biol* 210:4390–4398.
65. Robinson GE, Vargo EL (1997) Juvenile hormone in adult eusocial Hymenoptera: Gonadotropin and behavioral pacemaker. *Arch Insect Biochem Physiol* 35:559–583.
66. Sullivan JP, Fahrbach SE, Robinson GE, Robinson GE (2000) Juvenile hormone paces behavioral development in the adult worker honey bee. *Horm Behav* 37:1–14.
67. Brent CS, Vargo EL (2003) Changes in juvenile hormone biosynthetic rate and whole body content in maturing virgin queens of *Solenopsis invicta*. *J Insect Physiol* 49:967–974.
68. Aerts J, Law A (2009) An introduction to scripting in Ruby for biologists. *BMC Bioinformatics* 10:221.
69. Goto N, et al. (2010) BioRuby: Bioinformatics software for the Ruby programming language. *Bioinformatics* 26:2617–2619.
70. Stajich JE, et al. (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res* 12:1611–1618.
71. R Development Core Team (2007) *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna).

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.

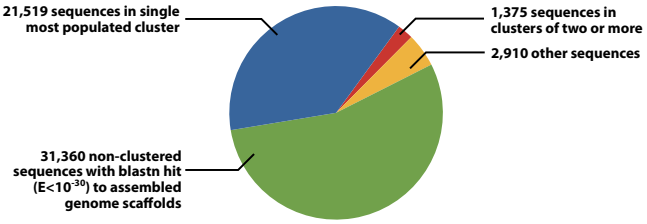

Fig. S1. Genome contigs that were not included in scaffolds largely represent repetitive elements.

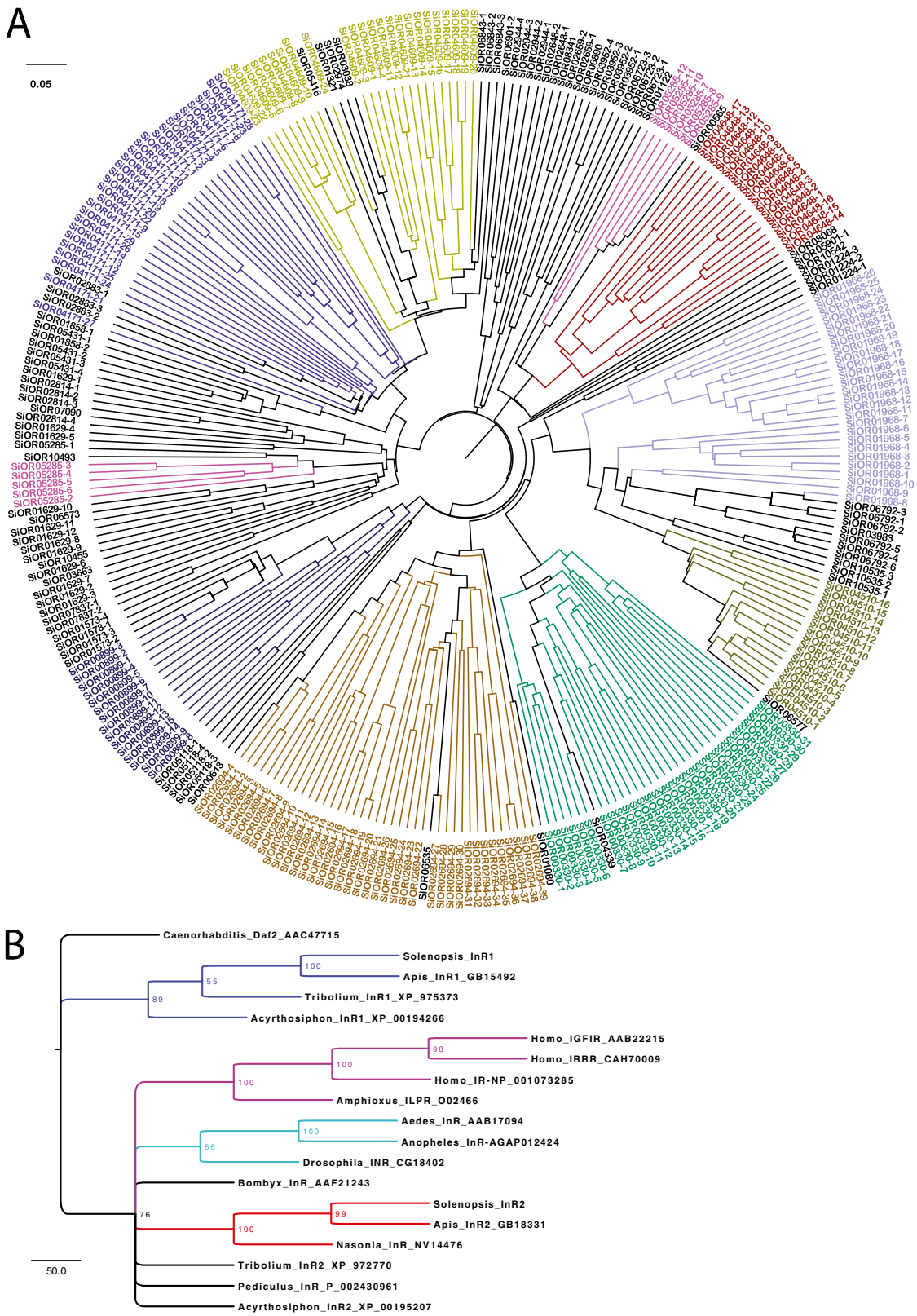

Fig. S2. Phylogenetic trees. (*A*) Phylogenetic relationships between *S. invicta* olfactory receptors (ORs). *Pogonomyrmex barbatus* and *N. vitripennis* OR sequences were aligned with MAFFT (29) and the resulting alignment used to construct an HMM profile (30). This profile was run on the *S. invicta* 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).

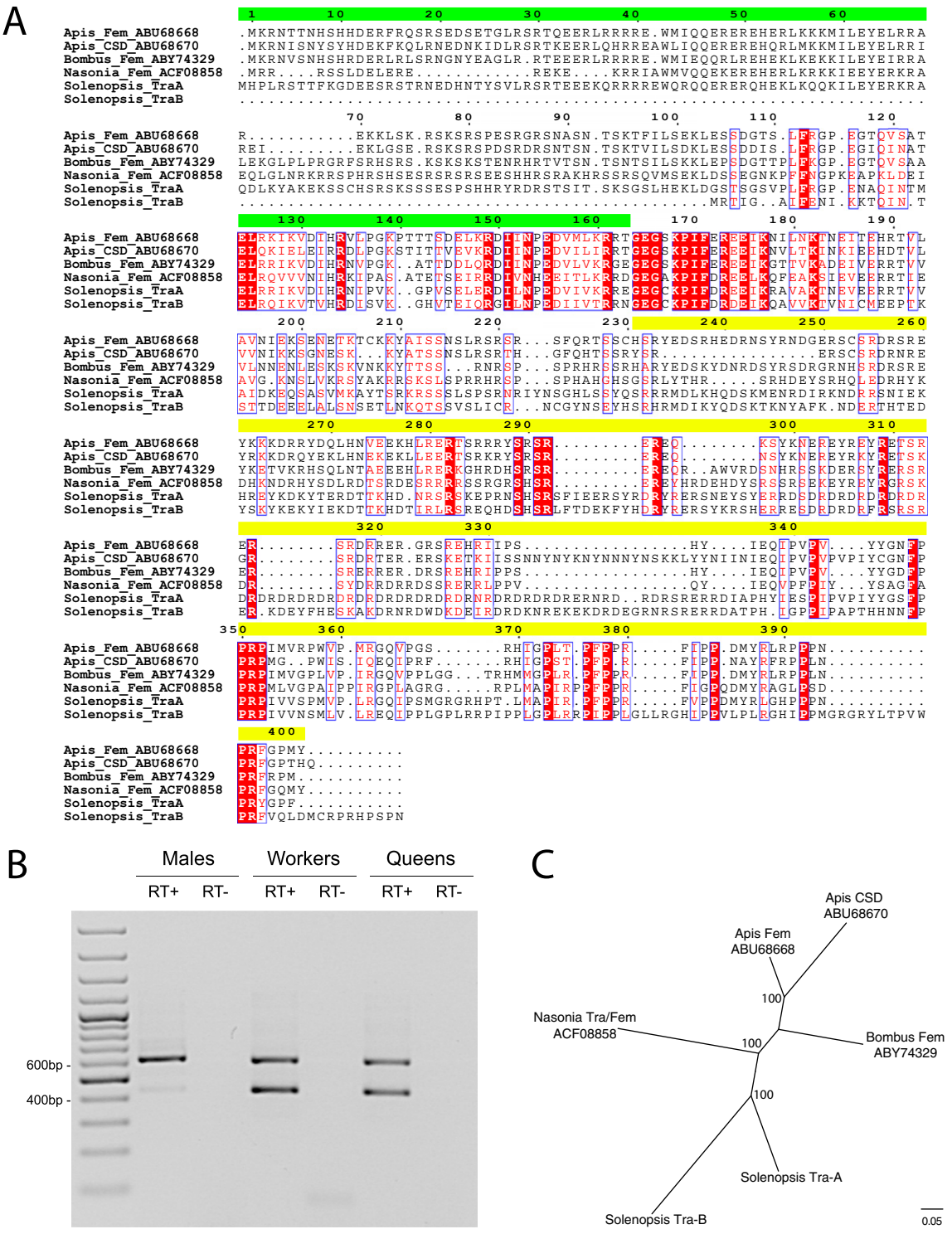

Fig. S3. *Transformer/feminizer* orthologs. (*A*) ClustalW (19) alignment of *transformer/feminizer* orthologs from *A. mellifera*, *B. terrestris*, *N. vitripennis* and *S. invicta*. GenBank accession numbers are provided when available after the second underscore. The *A. mellifera* SDP_N domain is 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).

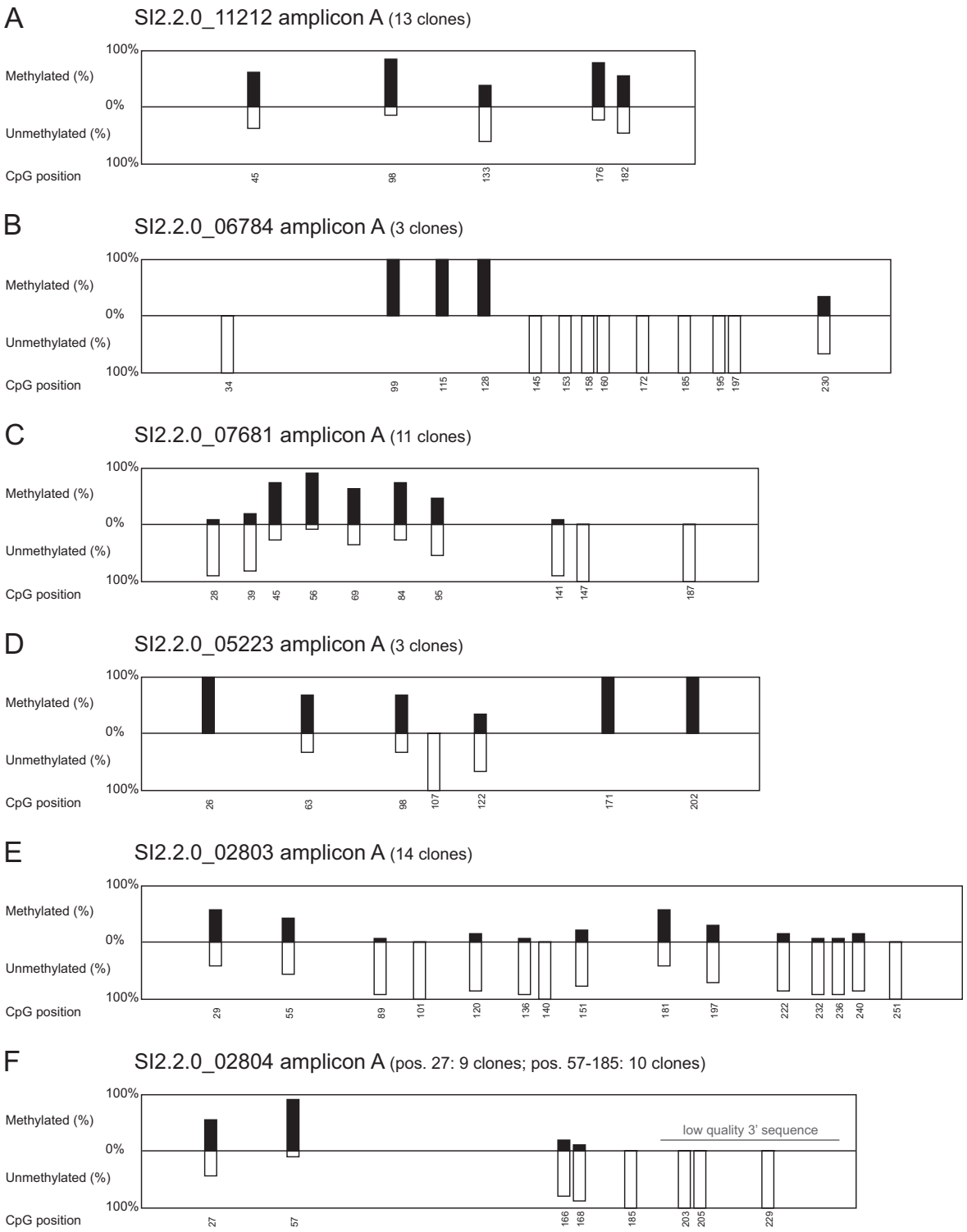

Fig. S4. CpG methylation in *S. invicta* genes is confirmed by the sequencing of bisulfite-converted amplicons. (*A–F*) Proportions of methylated Cs for each CpG site are provided for each gene.

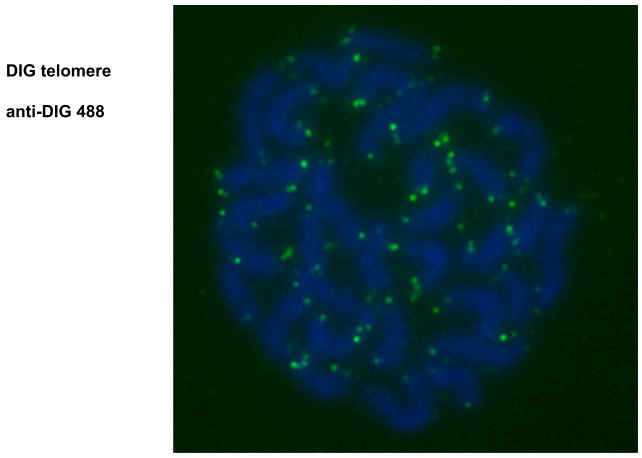

Fig. S5. Strong hybridization of fluorescently labeled $(TTAGG)_n$ probes to chromosome ends in an *S. invicta* larval nucleus suggests that *S. invicta* telomeres consist of $(TTAGG)_n$ repeats.

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)