F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.
16. M. Kaviratne, S. M. Khan, W. Jarra, P. R. Preiser, *Eukaryot. Cell* **1**, 926 (2002).
17. M. Haeggstrom *et al.*, *Mol. Biochem. Parasitol.* **133**, 1 (2004).
18. T. Y. Sam-Yellowe *et al.*, *Genome Res.* **14**, 1052 (2004).
19. J. Gorodkin, L. J. Heyer, S. Brunak, G. D. Stormo, *Comput. Appl. Biosci.* **13**, 583 (1997).
20. Z. Bozdech *et al.*, *PLoS Biol.* **1**, E5 (2003).
21. K. G. Le Roch *et al.*, *Science* **301**, 1503 (2003).
22. A search engine to identify proteins containing the PlasmoHT motif is available at www.haldarlab.northwestern.edu.
23. X.-Z. Su *et al.*, *Cell* **82**, 89 (1995).
24. J. F. Kun *et al.*, *Mol. Biochem. Parasitol.* **85**, 41 (1997).
25. We thank W. Kibbe, L. Zhu, V. Haztimanikatis, A. Vania Apkarian, and A. Chenn for helpful discussion. Supported by American Heart Association fellowship (0215246z to N.L.H.) and the NIH (HL69630, AI39071 to K.H.). PlasmoDB and GenBank identification codes, respectively: PFE1615c: NP_703661; PfHSP40: PFE0055c and NP_703357; PfEMP1 fragment chr4.glm_42. The PfEMP1 used for transmembrane domain and cytoplasmic tail has NCBI identification code AAB09769.1.

Supporting Online Material
www.sciencemag.org/cgi/content/full/306/5703/1934/DC1
Materials and Methods
Figs. S1 to S4
Table S1
Bioinformatic Data

13 July 2004; accepted 19 October 2004
10.1126/science.1102737

A Draft Sequence for the Genome of the Domesticated Silkworm (*Bombyx mori*)

Biology analysis group: Qingyou Xia,[1]*† Zeyang Zhou,[1]* Cheng Lu,[1]* Daojun Cheng,[1] Fangyin Dai,[1] Bin Li,[1] Ping Zhao,[1] Xingfu Zha,[1] Tingcai Cheng,[1] Chunli Chai,[1] Guoqing Pan,[1] Jinshan Xu,[1] Chun Liu,[1] Ying Lin,[1] Jifeng Qian,[1] Yong Hou,[1] Zhengli Wu,[1] Guanrong Li,[1] Minhui Pan,[1] Chunfeng Li,[1] Yihong Shen,[1] Xiqian Lan,[1] Lianwei Yuan,[1] Tian Li,[1] Hanfu Xu,[1] Guangwei Yang,[1] Yongji Wan,[1] Yong Zhu,[1] Maode Yu,[1] Weide Shen,[1] Dayang Wu,[1] Zhonghuai Xiang[1]†
Genome analysis group: Jun Yu,[2,3]*† Jun Wang,[2,3]* Ruiqiang Li,[2]* Jianping Shi,[2] Heng Li,[2] Guangyuan Li,[2] Jianning Su,[2] Xiaoling Wang,[2] Guoqing Li,[2] Zengjin Zhang,[2] Qingfa Wu,[2] Jun Li,[2] Qingpeng Zhang,[2] Ning Wei,[2] Jianzhe Xu,[2] Haibo Sun,[2] Le Dong,[2] Dongyuan Liu,[2] Shengli Zhao,[2] Xiaolan Zhao,[2] Qingshun Meng,[2] Fengdi Lan,[2] Xiangang Huang,[2] Yuanzhe Li,[2] Lin Fang,[2] Changfeng Li,[2] Dawei Li,[2] Yongqiao Sun,[2] Zhenpeng Zhang,[2] Zheng Yang,[2] Yanqing Huang,[2] Yan Xi,[2] Qiuhui Qi,[2] Dandan He,[2] Haiyan Huang,[2] Xiaowei Zhang,[2] Zhiqiang Wang,[2] Wenjie Li,[2] Yuzhu Cao,[2] Yingpu Yu,[3] Hong Yu,[3] Jinhong Li,[3] Jiehua Ye,[3] Huan Chen,[3] Yan Zhou,[3] Bin Liu,[2] Jing Wang,[2] Jia Ye,[3] Hai Ji,[2] Shengting Li,[2] Peixiang Ni,[2] Jianguo Zhang,[2] Yong Zhang,[2] Hongkun Zheng,[2] Bingyu Mao,[2] Wen Wang,[2] Chen Ye,[2] Songgang Li,[2] Jian Wang,[2,3] Gane Ka-Shu Wong,[2,3,4]† Huanming Yang[2,3]†

We report a draft sequence for the genome of the domesticated silkworm (*Bombyx mori*), covering 90.9% of all known silkworm genes. Our estimated gene count is 18,510, which exceeds the 13,379 genes reported for *Drosophila melanogaster*. Comparative analyses to fruitfly, mosquito, spider, and butterfly reveal both similarities and differences in gene content.

Silk fibers are derived from the cocoon of the silkworm *Bombyx mori*, which was domesticated over the past 5000 years from the wild progenitor *Bombyx mandarina* (*1*). Silkworms are second only to fruitfly as a model for insect genetics, owing to their ease of rearing, the availability of mutants from genetically homogeneous inbred lines, and the existence of a large body of information on their biology (*2*). There are about 400 visible phenotypes, and ~200 of these are assigned to linkage groups (*3*). Silkworms can also be used as a bioreactor for proteinaceous drugs and as a source of biomaterials. Here, we present a draft sequence of the silkworm genome with 5.9× coverage.

B. mori has 28 chromosomes. More than 1000 genetic markers have been mapped at an average spacing of 2 cM (~500 kb) (*4*). A physical map is being constructed through the fingerprinting and end sequencing of bacterial artificial chromosome (BAC) clones (*5*). Many expressed sequence tags (ESTs) have been produced (*6*), and a 3× draft sequence has just been announced by the International Lepidopteran Genome Project (*7*). Our project is independent of, but complementary to, that of the consortium. Our sequence has been submitted to the DNA Data Bank of Japan/European Molecular Biology Laboratory/GenBank (project accession number AADK00000000, version AADK01000000) and is also accessible from our Web site (http://silkworm.genomics.org.cn) (*8*). ESTs discussed in this Report can be found at GenBank (accession numbers CK484630 to CK565104).

DNA for genome sequencing is derived from an inbred domesticated variety, *Dazao* (posterior silk gland, fifth-instar day 3, on a mix of 1225 males). A whole-genome shotgun (*9*) technique was used, and our coverage is 5.9×. Including the unassembled reads, the total estimated genome size is 428.7 Mb, or 3.6 and 1.54 times larger than that of fruitfly (*10*) and mosquito (*11*). The N50 contig and scaffold sizes are 12.5 kb and 26.9 kb. Our assembly contains 90.9% of the 212 known silkworm genes (with full-length cDNA sequence), 90.9% of ~16,425 EST clusters, and 82.7% of the 554 known genes from other Lepidoptera. Additional details of our quality analyses are given in the supporting online material (fig. S1 and tables S1 to S6).

We developed a gene-finder algorithm *BGF* (BGI GeneFinder) (fig. S2), based on *GenScan* and *FgeneSH*. To determine a gene count for silkworm, one must correct for erroneous and partial predictions (Table 1). The final corrected gene count for silkworm is 18,510 genes, which far exceeds the official gene count of 13,379 for fruitfly

[1]Southwest Agricultural University, Chongqing Beibei, 400716, China. [2]Beijing Institute of Genomics of Chinese Academy of Sciences, Beijing Genomics Institute, Beijing Proteomics Institute, Beijing 101300, China. [3]James D. Watson Institute of Genome Sciences of Zhejiang University, Hangzhou Genomics Institute, Key Laboratory of Genomic Bioinformatics of Zhejiang Province, Hangzhou 310008, China. [4]University of Washington Genome Center, Department of Medicine, University of Washington, Seattle, WA 98195, USA.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: xiaqy@swau.cq.cn (Q.X.), xzh@swau.cq.cn (Z.X.), junyu@genomics.org.cn (J.Y.), gksw@genomics.org.cn (G.K-S.W.), yanghm@genomics.org.cn (H.Y.)

(our *BGF*-based procedures predict 13,366 genes for fruitfly). We find that 14.9% of predicted genes are confirmed by ESTs (based on aligning the ESTs to the genome and looking for a 100–base pair overlap with the predicted exons); 60.4% and 63.1% are confirmed by similarity to fruitfly genes and GenBank nonredundant proteins (*BlastP* at 10^{-6} E-value). Overall, 69.7% are confirmed by at least one method.

Not only did we find more genes in silkworm than in fruitfly, but we also found larger genes as a result of the insertion of transposable elements (TEs) in introns. For example, in *calcineurin B* (*cnb*), the silkworm gene was 12 times as large as that of fruitfly. To generalize, we compared annotations, found reciprocal best matches, and computed gene size ratios. Because prediction errors are unlikely to be alignable across species, we restricted our analysis to aligned regions, giving us a mean (median) ratio of 2.29 (2.75) (Fig. 1). This combination of more and bigger genes can explain 86% of the factor of 3.67 increase in genome size from fruitfly (116.8 Mb) to silkworm (428.7 Mb). Silkworm genes also had slightly more exons than fruitfly, with a mean (median) ratio of 1.15 (1.12) for number of exons per gene.

As shown by our TE annotations, most of this increase in the genome size of silkworm is relatively recent. Of the 21.1% of the genome that is recognizable as being of TE origins, 50.7% is from a single *gypsy-Ty3*–like retrotransposon (*12*) (table S7). Mean sequence divergence is 7.7%, which dates the initial appearance of this TE to 4.9 million years ago, if we use the fruitfly neutral rate of 15.6×10^{-9} substitutions per year (*13*). Most other TEs are comparably recent in origins (fig. S3). GC-rich regions contain a higher density of TEs, particularly LINEs (long interspersed nuclear elements), which is the exact opposite of what is reported for the human and mouse genomes.

Unlike silkworm, which is a lepidopteran, fruitfly and mosquito are dipterans. The two insect orders diverged about 280 to 350 million years ago (*14*). Comparisons of their genome content were done at the level of InterPro domains. Functional assignments were mapped according to Gene Ontology (GO). Domain clustering (*15*) (table S8) produced 8947 groups, with 2565 shared among insects and 1793 unique to silkworm (Fig. 2). Consistent with the observed TE expansion, domains like reverse transcriptase, integrase, and transposase stand out for their prevalence in silkworm. A complete list of predicted silkworm genes is shown in table S9, with a special indexing table for the genes discussed in this paper.

The silk gland, essentially a modified salivary gland, is a highly specialized organ whose function is to synthesize silk proteins. We identified a set of 1874 annotated genes that are confirmed by silk gland ESTs. Only 45 of these genes had been previously described in *B. mori*. GO function categories for silk gland and 11 other tissue libraries were compared (fig. S4). Several hormone-processing enzymes are active in silk gland, which is of interest because hormones participate in regulation of silk protein genes (*16*). Not counting low expressed genes undetectable at current EST depths, genes found only in silk gland include juvenile hormone (JH) esterase, ecdysone oxidase, and JH-inducible protein 1. Ecdysteroid UDP (uridine 5′-diphosphate)–glucosyl transferase is found in silk gland, testis, and ovary. Fibroin forms the bulk of the cocoon mass. It has two major components, a heavy (350 kD) and a light chain (25 kD). We found 1126 ESTs for the light chain, but only 4 ESTs for the heavy chain, suggesting that the one-to-one ratio for light and heavy chains is maintained at the post-transcription level. The heavy chain has five predominant amino acids: Gly (45.9%), Ala (30.3%), Ser (12.1%), Tyr (5.3%), and Val (1.8%). A complete tRNA gene set (table S10) was detected, including 41 Gly-tRNA and 41 Ala-tRNA, twice as many as in the other two insects and consistent with the requirements for fibroin production.

Another well-studied silk-secreting arthropod is the spider. We compared those 1874 genes expressed in *B. mori* silk gland with all available spider data (1482 from GenBank) and identified 107 homologs, including four *B. mori* counterparts for the major ampullate gland peroxidase in spider, which is involved in silk fiber formation (*17*).

We found 87 neuropeptide hormones, hormone receptors, and hormone-regulation genes. *Drosophila melanogaster* and *Anopheles gambiae* have 101 and 73 such genes, respectively. For *B. mori*, 52 genes were unknown, and 35 others were previously reported. Ecdysone oxidase and ecdysteroid UDP–glucosyl transferase (UGT) are implicated in ecdysone metabolism. We classified 20 UGT genes into five major clades (fig. S5), similar to the 34 UGT genes analyzed for *D. melanogaster* (*18*). Juvenile hormone (JH), ecdysone hormone (EH), and prothoracicotropic hormone (PTTH) work in coordination of ecdysis and metamorphosis. We identified 18 EH-sensitive receptors and receptor-like transcription factors. Four BRC Z4 genes contain intact DNA binding BTB domains. One has two additional zinc finger C2H2 type domains, with a zinc-coordinating cysteine pair and a histidine pair. These are involved in completing the larval-pupal transition, and later morphogenetic defects, or in programmed cell death of larval silk glands (*19*). We found many neuropeptide hormone genes too, like diapause hormone (DH), pheromone biosynthesis activating neuropeptide (PBAN), adipokinetic hormone (AKH), eclosion hormone, and bombyxin (4K-PTTH). In addition, diuretic hormone precursor and its receptor, allatotropin, and allatostatin were found. There was also a homolog to *Lymnaea stagnalis* neuropeptide Y precursor, a gene with pancreatic hormone activity that had not been detected in *D. melanogaster* and other insects and may therefore be new to silkworm.

Developmental genes for *D. melanogaster* have been extensively studied. We focused on 83 genes (*20*) that include 41 maternal genes, 12 gap genes, 9 pair-rule genes, 12 segment polarity genes, and 9 homeotic genes. The maternal genes are subdivided into four groups according to their function in patterning the early embryos (anterior, posterior, terminal, and dorsal-ventral). Only six genes [*oskar*, *swallow*, *trunk*, *fs(1)k10*, *gurken*, and *tube*], all from the maternal group, were not detected in *B. mori*. This confirms that the basic mechanism of development is largely conserved

Table 1. Number of predicted genes from *BGF*. We show the initial count, the number of erroneous predictions, and the gene count after likely errors are removed. There are four successive filters, which include rules to remove TEs and pseudogenes, as described in the SOM Text. The final gene count is computed as row 1 minus the sum of rows 2 to 5. Predictions are classified into single-exon genes, partial genes (no head = no start, no tail = no stop, neither) or complete genes. We correct for partial genes by stipulating that each is worth only half a gene. The final corrected gene count is then 18,510.

	Single exon	No head	No tail	Neither	Complete	All genes	Corrected
Total predicted	10,512	6,366	4,903	550	21,199	43,530	37,621
CDS < 100 bp or max exon score < 0.2	107	974	299	15	84	1,479	835
RepeatMasker TEs or copy number >10	7,334	2,233	2,111	124	7,575	19,377	17,143
Similarity to TE-associated proteins	132	71	68	7	294	572	499
Processed "single-exon" pseudogenes	314	146	179	8	153	800	634
Final annotated	2,625	2,942	2,246	396	13,093	21,302	18,510

across insects. It had been reported that *swallow* and *trunk* have no homologs in *A. gambiae*. We find that *tube* has no homolog in *A. gambiae*. Loss of the other three genes is interesting. Localization of the maternal determinant *oskar* at the posterior pole of the *D. melanogaster* oocyte provides positional information for pole plasm formation (*21*). *Gurken* encodes a ligand for *torpedo* (Egf-r), which triggers dorsal differentiation (*22*), whereas *fs(1)k10* is a probable negative regulator of *gurken* translation.

Lepidopteran wing patterning has stimulated a number of experimental studies. Although domesticated silkworm moths have long lost their ability to fly, as well as their colorful wing patterns, we expected that many of these genes would still be found in the sequence. We detected 18 silkworm homologs of wing-patterning genes from other Lepidoptera, primarily *Junonia coenia*. They include the *Distal-less* homeodomain gene, which affects eyespot number, positions, and sizes (*23*); *Ubx*, which represses *Distal-less* expression and leads to haltere formation in *D. melanogaster*, but may not act in the same manner in butterfly (*24*); Hh signaling pathway genes like *Hh*, *Ci*, *En*, and *Ptc*, which are important in eyespot focus formation; *Wg*, which plays a key role in band formation; and *EcR*, which is expressed in prospective eyespots and is coexpressed with *Distal-less* (*25*). Many of these genes are shared with the Diptera. Of the 323 wing-development genes known in *D. melanogaster*, 300 are found in silkworm. Most are well conserved, in that 87% and 56% align at E-values of better than 10^{-20} and 10^{-50}.

Silkworm is a female-heterogametic organism (ZZ in male, ZW in female). Sex in *B. mori* is determined by a dominant feminizing factor on W, as compared to the intricate X:A counting system known in *D. melanogaster*. A homolog of the *D. melanogaster* sex-determining gene *dsx* has been isolated in *B. mori*. It is called *Bmdsx*. Although structural features and splice sites are conserved in these two genes, regulatory mechanisms are not (*26*). The splicing regulator *tra* was not identified in *B. mori*. Neither was the TRA/TRA2 binding site for *Bmdsx*, suggesting that the upstream sex-determining cascade for *B. mori* and *D. melanogaster* differ. However, homologs for most known sex-determining factors can be found. Among *daughterless* (*da*), *hermaphrodite* (*her*), *extra macrochaetae* (*emc*), *groucho* (*gro*), *sisterless A* (*sisA*), *scute* (*sc*), *outstretched* (*os*), *deadpan* (*dpn*), and *runt* (*run*) (*27*), homologs for *da*, *emc*, *gro*, *sc*, *dpn*, and *run* were identified in *B. mori*. For *D. melanogaster*, dosage compensation is known to equalize transcription of X-chromosome genes between sexes. At least six genes (*msl-1*, *msl-2*, *msl-3*, *mle*, *mof*, *JIL-1*) are required, and of these, homologs of *mle*, *mof*, and *msl-3* were found in *B. mori*, despite the growing evidence for absence of Z-linked dosage compensation in *B. mori* (*28*). In these and other cases in which insect genes were not found in *B. mori*, we manually checked our automated procedures (see SOM Text). However, further experiments will be needed, given the incompleteness of the genome and the level of homology needed for detection.

Humoral immune factors together with wound healing, homeostasis, and adaptive humoral immune responses are important components of immunity and defense in insects (*29*). We identified a total of 69 such genes, including 34 antibacterial genes, of which 23 appear to be newly identified. They encode the innate immune factors synthesized in fat bodies and hemocytes, which kill bacteria by permeabilizing their membranes. One of them is the Lepidopteran *moricin*, a highly alkaline antibacterial peptide initially isolated from *B. mori*. A new cluster of 8 *moricin* genes was found, with amino acid sequence identities of greater than 90% among members, but only 20% similarity to known *moricins*. *Defensins* specific to Gram-positive bacteria were found, as were *cecropins* (*30*). We detected a previously unknown class of *cecropins*. Other found genes related to insect defense include *lysozymes*, *hemolin*, *lectins*, and *prophenoloxidases*. As a member of the immunoglobulin (Ig) family, *hemolin* is unique to the Lepidoptera. *Lectins* are abundant, with 29 found in *B. mori*, compared to 35 and 22 in *D. melanogaster* and *A. gambiae* (*31*), respectively. We also identified three *prophenoloxidases*, of which two were previously known.

Lepidoptera are unusual because they have holocentric chromosomes with diffuse kinetochores. This characteristic is a potential driver of evolution because of the ability to retain chromosome fragments through many cell divisions. The nematode also has diffuse kinetochores, and five key chromosomal proteins are known (*32*, *33*): *hcp-1*, *hcp-2*, *hcp-3*, *hcp-4*, and *hcp-6*. (The prefix *hcp* stands for "holocentric protein.") *Hcp-3* is detected in all eukaryotic centromeres, similar to histone H3 in its histone-fold domain, but dissimilar in its N-terminal region. It is also known as *Cse4p* in yeast, *Cid* in fruitfly, and *CENP-A* in human. Their proteins are highly diverged. The putative homolog in silkworm has only 23% identity to the histone-fold domain of *hcp-3*, but their lengths are similar: 268 amino acids for silkworm and 288 amino acids for nematode. There are many homologs of *hcp-1* and *hcp-2*—18 and 72, to be specific—making it difficult to determine which ones might be the true orthologs. We could not find a homolog for *hcp-4*, but we did identify a homolog for a related gene that is known as *CENP-C* and was previously found in human, mouse, and chicken. Finally, we were not able to identify the silkworm homolog for *hcp-6*.

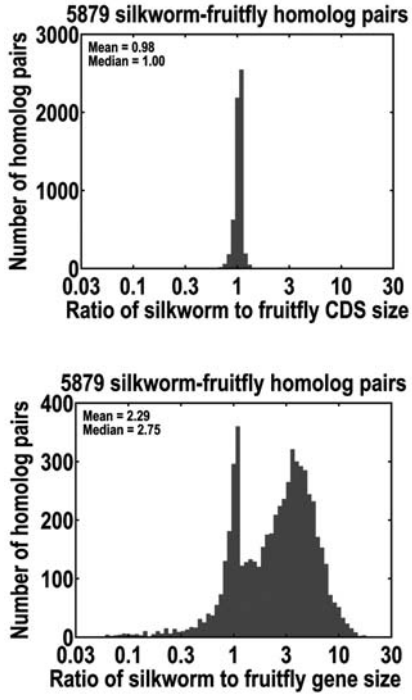

Fig. 1. Comparison of gene size in silkworm-fruitfly orthologs. We use reciprocal best matches, and calculate a ratio over the aligned portion. Size is shown with (gene size) or without (CDS size) introns. The minor peak is due to single-exon alignments.

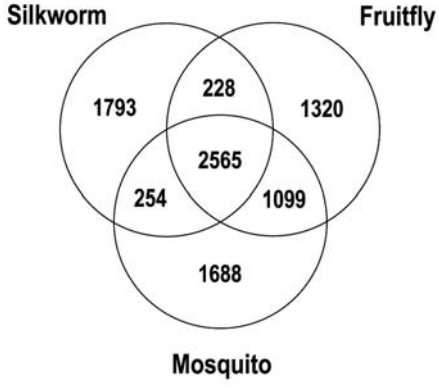

Fig. 2. InterPro domain clusters shared among or unique to all possible combinations of silkworm, fruitfly, and mosquito. Clusters are constructed with the algorithm detailed in table S8, which is based on a similar earlier analysis (*14*).

References and Notes

1. Y. Zhou, *General Entomology* (High Education Publication House, Beijing, ed. 2, 1958).
2. M. R. Goldsmith, in *Molecular Model Systems in the Lepidoptera*, M. R. Goldsmith, A. S. Wilkins, Eds. (Cambridge Univ. Press, Cambridge, 1995), pp. 21–76.

3. H. Doira, H. Fujii, Y. Kawaguchi, H. Kihara, Y. Banno, *Genetic Stocks and Mutations of Bombyx mori* (Institute of Genetic Resources, Kyushu University, Japan, 1992).
4. M. R. Goldsmith, T. Shimada, H. Abe, *Annu. Rev. Entomol.* 10.1146/annurev.ento.50.071803.130456 (2004).
5. C. Wu, S. Asakawa, N. Shimizu, S. Kawasaki, Y. Yasukochi, *Mol. Gen. Genet.* **261**, 698 (1999).
6. K. Mita *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 14121 (2003).
7. K. Mita *et al.*, *DNA Res.* **11**, 27 (2004).
8. J. Wang *et al.*, *Nucleic Acids Res.*, in press.
9. J. Yu *et al.*, *Science* **296**, 79 (2002).
10. M. D. Adams *et al.*, *Science* **287**, 2185 (2000).
11. R. A. Holt *et al.*, *Science* **298**, 129 (2002).
12. H. Abe *et al.*, *Mol. Gen. Genet.* **263**, 916 (2000).
13. W. H. Li, *Molecular Evolution* (Sinauer, Sunderland, MA, 1997).
14. M. W. Gaunt, M. A. Miles, *Mol. Biol. Evol.* **19**, 748 (2002).
15. G. M. Rubin *et al.*, *Science* **287**, 2204 (2000).
16. K. Grzelak, *Comp. Biochem. Physiol. B Biochem. Mol. Biol.* **110**, 671 (1995).
17. N. N. Pouchkina, B. S. Stanchev, S. J. McQueen-Mason, *Insect Biochem. Mol. Biol.* **33**, 229 (2003).
18. T. Luque, D. R. O'Reilly, *Insect Biochem. Mol. Biol.* **32**, 1597 (2002).
19. M. Uhlirova *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 15607 (2003).
20. T. Brody, *Trends Genet.* **15**, 333 (1999); http://flybase.bio.indiana.edu/allied-data/lk/interactive-fly.
21. N. F. Vanzo, A. Ephrussi, *Development* **129**, 3705 (2002).
22. S. Roth, F. S. Neuman-Silberberg, G. Barcelo, T. Schupbach, *Cell* **81**, 967 (1995).
23. P. Beldade, P. M. Brakefield, A. D. Long, *Nature* **415**, 315 (2002).
24. W. O. McMillan, A. Monteiro, D. D. Kapan, *Trends Ecol. Evol.* **17**, 125 (2002).
25. P. B. Koch, R. Merk, R. Reinhardt, P. Weber, *Dev. Genes Evol.* **212**, 571 (2003).
26. M. G. Suzuki, F. Ohbayashi, K. Mita, T. Shimada, *Insect Biochem. Mol. Biol.* **31**, 1201 (2001).
27. C. Schutt, R. Nothiger, *Development* **127**, 667 (2000).
28. M. G. Suzuki, T. Shimada, M. Kobayashi, *Heredity* **81**, 275 (1998).
29. A. B. Mulnix, P. E. Dunn, in *Molecular Model Systems in the Lepidoptera*, M. R. Goldsmith, A. S. Wilkins, Eds. (Cambridge Univ. Press, Cambridge, 1995), pp. 369–395.
30. H. Steiner, D. Hultmark, A. Engstrom, H. Bennich, H. G. Boman, *Nature* **292**, 246 (1981).
31. G. K. Christophides *et al.*, *Science* **298**, 159 (2002).
32. L. L. Moore, M. B. Roth, *J. Cell Biol.* **153**, 1199 (2001).
33. J. H. Stear, M. B. Roth, *Genes Dev.* **16**, 1498 (2002).
34. This project was supported by Chinese Academy of Sciences, National Development and Reform Commission, Ministry of Science and Technology, National Natural Science Foundation of China, Ministry of Agriculture, Chongqing Municipal Government, Beijing Municipal Government, Zhejiang Provincial Government, Hangzhou Municipal Government, and Zhejiang University. Additional funding came from National Human Genome Research Institute (grant 1 P50 HG02351).

Supporting Online Material
www.sciencemag.org/cgi/content/full/306/5703/1937/DC1
SOM Text
Figs. S1 to S5
Tables S1 to S10

1 July 2004; accepted 20 October 2004
10.1126/science.1102210

By Carrot or by Stick: Cognitive Reinforcement Learning in Parkinsonism

Michael J. Frank,[1]* Lauren C. Seeberger,[2] Randall C. O'Reilly[1]*

To what extent do we learn from the positive versus negative outcomes of our decisions? The neuromodulator dopamine plays a key role in these reinforcement learning processes. Patients with Parkinson's disease, who have depleted dopamine in the basal ganglia, are impaired in tasks that require learning from trial and error. Here, we show, using two cognitive procedural learning tasks, that Parkinson's patients off medication are better at learning to avoid choices that lead to negative outcomes than they are at learning from positive outcomes. Dopamine medication reverses this bias, making patients more sensitive to positive than negative outcomes. This pattern was predicted by our biologically based computational model of basal ganglia–dopamine interactions in cognition, which has separate pathways for "Go" and "NoGo" responses that are differentially modulated by positive and negative reinforcement.

Should you shout at your dog for soiling the carpet or praise him when he does his business in the yard? Most dog trainers will tell you that the answer is both. The proverbial "carrot-and-stick" motivational approach refers to the use of a combination of positive and negative reinforcement: One can persuade a donkey to move either by dangling a carrot in front of it or by striking it with a stick. Both carrots and sticks are important for instilling appropriate behaviors in humans. For instance, when mulling over a decision, one considers both pros and cons of various options, which are implicitly influenced by positive and negative outcomes of similar decisions made in the past. Here, we report that whether one learns more from positive or negative outcomes varies with alterations in dopamine levels caused by Parkinson's disease and the medications used to treat it.

To better understand how healthy people learn from their decisions (both good and bad), it is instructive to examine under what conditions this learning is degraded. Notably, patients with Parkinson's disease are impaired in cognitive tasks that require learning from positive and negative feedback (*1*–*3*). A likely source of these deficits is depleted levels of the neuromodulator dopamine in the basal ganglia of Parkinson's patients (*4*), because dopamine plays a key role in reinforcement learning processes in animals (*5*). A simple prediction of this account is that cognitive performance should improve when patients take medication that elevates their dopamine levels. However, a somewhat puzzling result is that dopamine medication actually worsens performance in some cognitive tasks, despite improving it in others (*6*, *7*).

Computational models of the basal ganglia–dopamine system provide a unified account that reconciles the above pattern of results and makes explicit predictions about the effects of medication on carrot-and-stick learning (*8*, *9*). These models simulate transient changes in dopamine that occur during positive and negative reinforcement and their differential effects on two separate pathways within the basal ganglia system. Specifically, dopamine is excitatory on the direct or "Go" pathway, which helps facilitate responding, whereas it is inhibitory on the indirect or "NoGo" pathway, which suppresses responding (*10*–*13*). In animals, phasic bursts of dopamine cell firing are observed during positive reinforcement (*14*, *15*), which are thought to act as "teaching signals" that lead to the learning of rewarding behaviors (*14*, *16*). Conversely, choices that do not lead to reward [and aversive events, according to some studies (*17*)] are associated with dopamine dips that drop below baseline (*14*, *18*). Similar dopamine-dependent processes have been inferred to occur in humans during positive and negative reinforcement (*19*, *20*). In our models, dopamine bursts increase synaptic plasticity in the direct pathway while decreasing it in the indirect pathway (*21*, *22*), supporting Go learning to reinforce the good choice. Dips in dopamine have the opposite effect, supporting NoGo learning to avoid the bad choice (*8*, *9*).

A central prediction of our models is that nonmedicated Parkinson's patients are impaired at learning from positive feedback (bursts of dopamine; "carrots"), because of reduced levels of dopamine. However, the

[1]Department of Psychology and Center for Neuroscience, University of Colorado Boulder, Boulder, CO 80309–0345, USA. [2]Colorado Neurological Institute Movement Disorders Center, Englewood, CO 80113, USA.

*To whom correspondence should be addressed. E-mail: frankmj@psych.colorado.edu (M.J.F.); oreilly@psych.colorado.edu (R.C.O.).

Supplement on data quality

Prior to assembly, we remove potential contaminations by randomly selecting two sequences from each plate and comparing these to the other known genome sequences in GenBank, as well as all sequences from the different organisms that have been sequenced at our institute. After assembly, and before submission to GenBank, we remove scaffolds smaller than 2-Kb because, although most of them are of silkworm origin, some might be contaminants that were not removed at the plate level. The logic is that contaminants will have low coverage, and be largely unassembled.

Assembly of the raw sequence reads is done using an updated version of our *RePS* software (1,2), incorporating some recent ideas from *Phusion* (3). The crucial point is that *RePS* uses the *Phred*/*Phrap* system (4-6) to handle its detailed assembly, and thus reliable estimates of the error rate are available for every base. These estimates are represented by a quality Q equal to $-10 \cdot \log_{10}$(error rate). Table S1 shows the raw data in our assembly, and Table S2 shows the resultant contig and scaffold sizes. Note that all low quality bases (below Q20) are removed from the contig ends. Coverage is 5.9x, based on the number of high quality bases (above Q20) in the non-repeated parts of the contigs bigger than 5-Kb. In our subsequent analyses, we exclude unassembled reads and assembled pieces smaller than 2-Kb. But here, to estimate genome size we include them. What we find is a genome size of 428.7-Mb, smaller than the previously estimated size of 530-Mb, but that estimate was based on CoT analysis, which is not as precise. Contig and scaffold sizes are 12.5-Kb and 26.9-Kb, based on N50 statistics, where N50 is that size above which half of the total length of the data set is found. We believe that most of the breaks in the assembly are due to TEs, as opposed to sampling statistics. To get larger scaffolds, one would need linking information on a scale that is comparable to 26.9-Kb, perhaps based on fosmid end-pairs, as their inserts are biologically constrained to be roughly 40-Kb.

We measure the quality of our WGS assembly based on completeness of coverage and single-base error rates. First, we made a list of the 212 silkworm genes with complete

sequence in GenBank. Of these, 90.9% can be found in our WGS assembly, although not necessarily always in one scaffold. Then, we sequenced 80,470 ESTs from tissues shown in Table S3. These collapse to 16,425 UniGene clusters (7), and 90.9% are found in our WGS assembly. To confirm that silkworm is a legitimate model for other lepidoptera, we searched for homology to 554 GenBank genes from other lepidoptera, and we find 82.7% of them. A summary is given in Table S4 and the full set of genes is listed in Table S5. Based on the *Phred*/*Phrap* error estimates provided by *RePS*, we can state that 96.0% and 89.6% of the WGS assembly has an error rate of better than 10^{-3} and 10^{-4}. A cumulant for the estimated error rates is depicted in Figure S1.

Finally, we compared to BACs in GenBank. Although the BACs are from *Dazao*, which is supposedly inbred, there is genetic diversity of about 1.3×10^{-3} between different individuals of *Dazao* (8). Even if this number is a slight over-estimate, resulting from the low quality of the ESTs, perfect concordance should not be expected. Thus, we perform 3 comparisons, for sequences with error estimates of better than 10^{-2} (Q20), 10^{-3} (Q30), and 10^{-4} (Q40). These are summarized through Table S6. Two BACs from chromosome Z are clearly not finished, because their GenBank sequences exist as multiple pieces. One BAC on chromosome 2 has an exceptionally high repeat content, based on known transposable elements and on 20-mers of high copy number. Not surprisingly, the alignments are more fragmented and there are more mismatches. The most representative BACs are the two on chromosomes 11 and 13. For these, coverages range from 91.2 to 92.8%, similar to above estimates based on gene content. Mismatches are 0.045% in the Q20 table and 0.030% in the Q40 table. We believe most of these mismatches are due to polymorphisms. In fact, if we sum over all the *Phred*/*Phrap* error estimates, they would predict a 0.012% difference in mismatch rates between Q20 and Q40, much as is observed.

Supplement on data analysis

To find genes in this and other genomes, we developed an *ab initio* program, *BGF* (BGI GeneFinder), based on *GenScan* (9) and *FgeneSH* (10). We did it because *GenScan* does not make its source code freely available for further customizations, and *FgeneSH* is now commercial. We make no claims for originality, just convenience. Our program was tested against fruitfly, on a set of 6,667 complete cDNA-to-genomic alignments, with the methods from a recent review (11). This is shown in Figure S2. *BGF* compares favorably with *GenScan* on per-amino-acid false positive (FP) and false negative (FN) rates. *BGF* is slightly better in its ability to stop at the end of a gene, instead of over-predicting exons in regions outside the gene. For silkworm, the number of cDNA-to-genomic alignments that can be used to validate the program is much smaller. Even after including fruitfly cDNAs with obvious alignments to silkworm, we had only 238 genes. Averaged over this test set, FP=0.06 and FN=0.07. Over-predicted exons appear in 22% of the genes (18% at 5'-ends and 5% at 3'-ends), and erroneous exons overshoot the correct start/stop codon sites by a mean of 1188-bp (1326-bp at 5'-ends and 484-bp at 3'-ends).

Two factors must be considered in arriving at a gene count: partial predictions and erroneous predictions. *BGF* flags partial genes as no head (missing start), no tail (missing stop), or neither (missing start and stop). These can arise from any of a number of factors, including lack of contiguity, failures in the gene finder, and pseudogenes. In any case, the simplest way to fix the gene count is to treat partial predictions as half a gene. To remove erroneous gene predictions, we apply four successive filters. First, we remove predictions where coding regions (CDS) are smaller than 100-bp or maximum *BGF* exon confidences are less than 0.2. Second, we remove likely TEs with more than 50% repeats in the CDS, where by repeats we mean *RepeatMasker* TEs or 20-mers of copy number over 10. About 90% of all erroneous predictions are removed by this filter. Third, additional putative TEs are removed by searching for similarity to TE-associated genes with GenBank descriptors like retrotransposon, transposase, and reverse transcriptase. Fourth, we remove processed

pseudogenes where 75% of the CDS is in a single exon, and it has 90% identity over 80% of its length to another multi-exon silkworm gene.

To identify transposable elements (TEs), we constructed our own repeat library by merging silkworm TEs in GenBank with fruitfly/invertebrate TEs in *RepBase* (12). These library entries are used by *RepeatMasker* (13) to generate Table S7. Of the library entries that are usable, 82, 118, and 60 come from silkworm, fruitfly, and invertebrates. It should be noted that identifiable TE content is necessarily an underestimate, as the repeat library is incomplete, and the largest repeats are not assembled by *RePS*. Indeed, if we collect all contiguous blocks of mathematically perfect repeats with a 20-mer copy number over 10, *RepeatMasker* would fail to identify half of these sequences as TEs. For the 21.1% of the genome that is identified as TEs, 50.7% are from a single *gypsy-Ty3*-like retrotransposon (14). The mean divergence is 7.7%, dating the origins of this TE to 4.9 million years ago, using the fruitfly neutral rate of 15.6×10^{-9} substitutions per year (15). Additional plots for age and GC content distributions are depicted in Figure S3.

InterPro domains (16) are annotated by *InterProScan* Release 7.0. Since the exact positions of the domains are not kept in the databases, we reran *InterProScan* on all three insects. Gene Ontology (GO) (17) assignments are derived from this. To compare insects, we apply the clustering algorithm (18) detailed by Table S8. Briefly, a set of n domains is said to form a cluster when the number of acceptable homologs exceeds some fraction f (default is 0.25) of the theoretical maximum $(n-1)!$. Acceptable is when the homologous region exceeds 50% of the domain or 100 amino acids, for *BlastP* expectation value 10^{-6}. Finally, we assess the strength of evolutionary selection through Ka/Ks, where Ka and Ks are substitutional rates per non-synonymous and synonymous site. As expected from their known evolutionary relatedness, selectional forces are stronger for fruitfly-mosquito than for silkworm-fruitfly or for silkworm-mosquito.

Sequence conservation between closely related species is an increasingly popular method that is used to identify putatively functional non-coding motifs. Unfortunately, in the case of silkworm, neither fruitfly nor mosquito is sufficiently close for this purpose. If

we compare the silkworm genome to the fruitfly genome with *BlastZ* (19), only 7.7% can be aligned. In contrast, for mouse-human comparisons, 40.5% align. Looking upstream of 2902 orthologous gene pairs in silkworm-fruitfly identifies at most 60 conserved regions, even with a relatively liberal criterion of 60% sequence identity over a 30-bp region. This is disappointing given, for example, that regulatory elements for chorion gene expression are known to be conserved between the silkworm and fruitfly (20). The problem need not be a lack of conservation *per se*. It is just as likely that the conserved motifs are too small for the existing cross-species alignment software to detect.

For the comparative analyses, we use *BDGP* Release 3.1 for fruitfly and *Ensembl* Release 16.2 for mosquito. Unless otherwise stated all other sequences are from GenBank Release R137 October 2003. To establish if a gene is "newly discovered", we consider if the sequence (or even a part of the sequence) is present in GenBank or one of the species-specific databases for the sequenced organisms like fruitfly, mosquito, etc. The homology searches use *BlastP*, at an initial E-value threshold of 10^{-6}. When the automated searches fail, we repeat the searches manually, to ensure that weaker but still valid homologies are not being rejected by our automation code. For example, we check for partial alignments to regions containing known domains, especially when it is known from the literature that that domain is not well conserved. We also check to ensure that the identified homolog is not more similar (*i.e.* with a better E-value) to another gene with a different function. Any additional criteria are specified in the captions.

References

1. J. Wang, G.K. Wong, P. Ni, Y. Han, X. Huang, et al. RePS: a sequence assembler that masks exact repeats identified from the shotgun data. *Genome Res.* **12**, 824-831 (2002).

2. L. Zhong, K. Zhang, X. Huang, P. Ni, Y. Han, et al. A statistical approach designed for finding mathematically defined repeats in shotgun data and determining the length distribution of clone-inserts. *Geno. Prot. & Bioinfo.* **1**, 43-51 (2003).

3. J.C. Mullikin, Z. Ning. The phusion assembler. *Genome Res.* **13**, 81-90 (2003).

4. B. Ewing, L. Hillier, M.C. Wendl, P. Green. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8**, 175-185 (1998).

5. B. Ewing, P. Green. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8**, 186-194 (1998).

6. P. Green. http://www.phrap.org.

7. M.S. Boguski, G.D. Schuler. ESTablishing a human transcript map. *Nat. Genet.* **10**, 369-371 (1995).

8. T.C. Cheng, Q.Y. Xia, J.F. Qian, C. Liu, Y. Lin, X.F. Zha, Z.H. Xiang. Mining single nucleotide polymorphisms from EST data of silkworm, Bombyx mori, inbred strain Dazao. *Insect Biochem. Mol. Biol.* **34**, 523-530 (2004).

9. C. Burge, S. Karlin. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78-94 (1997). http://genes.mit.edu/GENSCAN.html.

10. A.A. Salamov, V.V. Solovyev. Ab initio gene finding in Drosophila genomic DNA. *Genome Res.* **10**, 516-522 (2000). http://www.softberry.com/berry.phtml.

11. J. Wang, S. Li, Y. Zhang, H. Zheng, Z. Xu, et al. Vertebrate gene predictions and the problem of large genes. *Nature Rev. Genet.* **4**, 741-749 (2003).

12. J. Jurka. Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet.* **16**, 418-420 (2000). http://www.girinst.org.

13. A.F. Smit, P. Green. http://repeatmasker.genome.washington.edu/cgi-bin/RepeatMasker.

14. H. Abe, F. Ohbayashi, T. Shimada, T. Sugasaki, S. Kawai, et al. Molecular structure of a novel gypsy-Ty3-like retrotransposon (Kabuki) and nested retrotransposable elements on the W chromosome of the silkworm Bombyx mori. *Mol. Gen. Genet.* **263**, 916-924 (2000).

15. W.H. Li. *Molecular Evolution* (Sinauer, Sunderland, 1997).

16. N.J. Mulder, R. Apweiler, T.K. Attwood, A. Bairoch, D. Barrell, et al. The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.* **31**, 315-318 (2003). http://www.ebi.ac.uk/interpro.

17. E. Camon, M. Magrane, D. Barrell, D. Binns, W. Fleischmann, et al. The Gene Ontology Annotation (GOA) Project: Implementation of GO in SWISS-PROT, TrEMBL, and InterPro. *Genome Res.* **13**, 662-672 (2003). http://www.geneontology.org.

18. G.M. Rubin, M.D. Yandell, J.R. Wortman, G.L. Gabor Miklos, C.R. Nelson, et al. Comparative genomics of the eukaryotes. *Science* **287**, 2204-2215 (2000). See footnote 63 for domain clustering algorithm.

19. S. Schwartz, W.J. Kent, A. Smit, Z. Zhang, R. Baertsch, et al. Human-mouse alignments with BLASTZ. *Genome Res.* **13**, 103-107 (2003).

20. F.C. Kafatos, G. Tzertzinis, N.A. Spoerel, H.T. Nguyen, in *Molecular model systems in the Lepidoptera*, M.R. Goldsmith, A.S. Wilkins, Eds. (Cambridge University Press, Cambridge, 1995) pp. 181-215.

Figure captions

Figure S1: *Phrap* quality cumulant, depicting percentage of assembled sequence with a quality Q less than the indicated abscissa. Q is related to the estimated single nucleotide substitution error rate by equation $Q = -10 \cdot \log_{10}(\text{error rate})$.

Figure S2: Gene prediction by *BGF* and *GenScan*, tested on 6,667 cDNA-defined fruitfly genes. Genomic size refers to the unspliced transcript, including introns, from the start to stop codons. False positive (FP) and false negative (FN) rates are computed on a per-aa (per amino acid) basis, meaning that we require the reading frame to be correctly called. Our definition of FP counts erroneously predicted exons only in the region of the genome defined by the cDNA. When the prediction goes past the start/stop codon, we call that an over-prediction, not an FP. Here, we show the probability of such an over-prediction, and the genomic extent of these over-predictions, at both 5' and 3' ends.

Figure S3: Age and GC content for silkworm TEs. Divergence (age) is defined relative to the consensus used by *RepeatMasker* to identify that TE, and y-axis is normalized to the size of the silkworm genome. GC content is computed with a 2-Kb window, and y-axis is normalized to the amount of sequence at each GC content.

Figure S4: Gene Ontology classifications for EST-confirmed genes in silk gland (on third day of fifth larva instar) compared to 11 other libraries. We were able to classify 46.6% of 1,874 genes in silk gland and 37.2% of 9859 genes in the other libraries.

Figure S5: *ClustalW* phylogeny for 20 UDP-glucosyl transferase (UGT) genes, based on conserved C-terminal region (250 aa). Horizontal lengths are drawn to scale and indicate sequence divergence. The scale bar is a divergence of 5%.

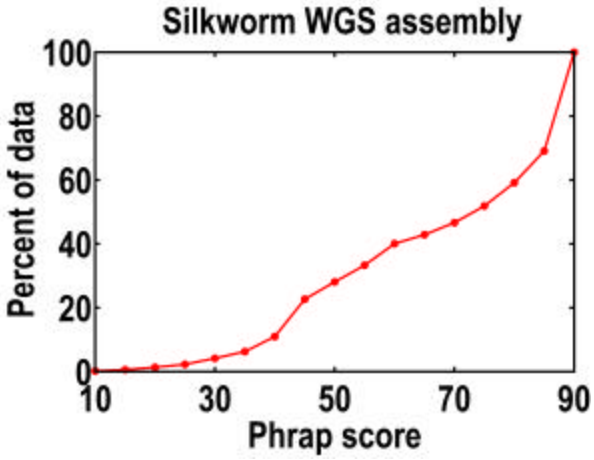
Silkworm WGS assembly
Percent of data
100
80
60
40
20
0
10
30
50
70
90
Phrap score

Figure S1.

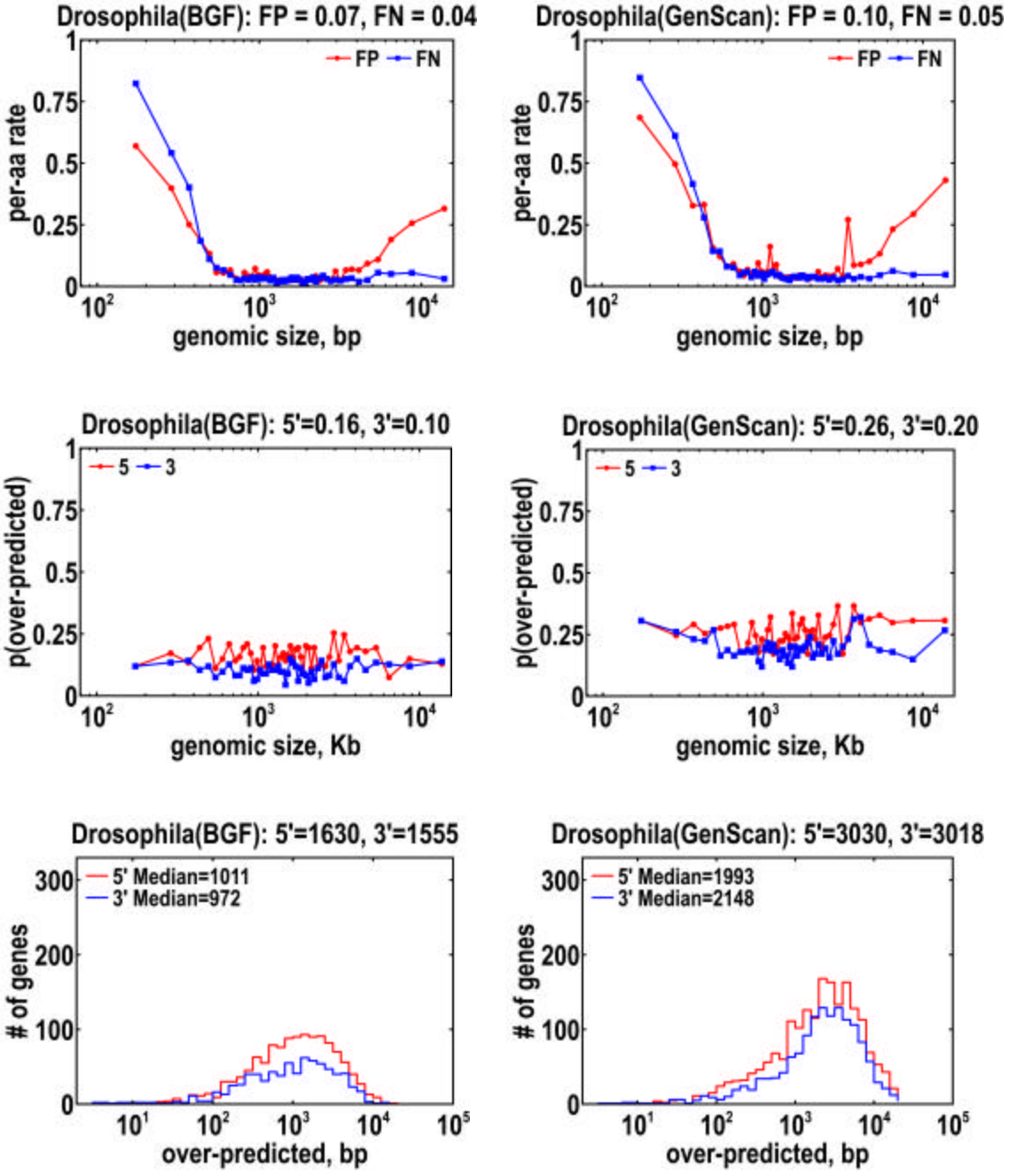
Drosophila(BGF): FP = 0.07, FN = 0.04
Drosophila(GenScan): FP = 0.10, FN = 0.05
FP
FN
per-aa rate
genomic size, bp
Drosophila(BGF): 5'=0.16, 3'=0.10
Drosophila(GenScan): 5'=0.26, 3'=0.20
5
3
p(over-predicted)
genomic size, Kb
Drosophila(BGF): 5'=1630, 3'=1555
Drosophila(GenScan): 5'=3030, 3'=3018
5' Median=1011
3' Median=972
5' Median=1993
3' Median=2148
# of genes
over-predicted, bp

Figure S2.

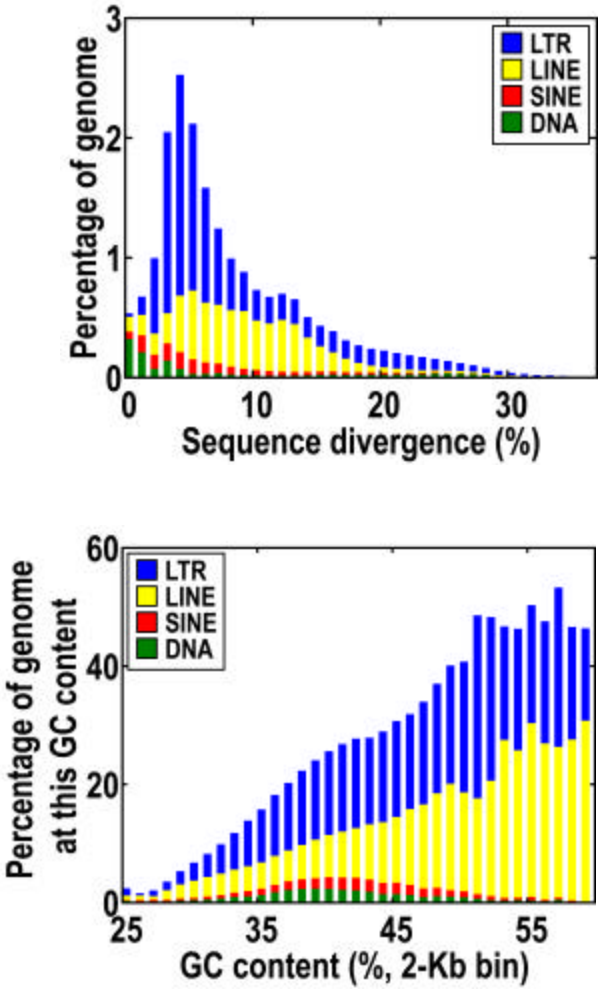
Percentage of genome
Sequence divergence (%)
LTR
LINE
SINE
DNA
0
1
2
3
0
10
20
30
Percentage of genome at this GC content
GC content (%, 2-Kb bin)
LTR
LINE
SINE
DNA
0
20
40
60
25
35
45
55

Figure S3.

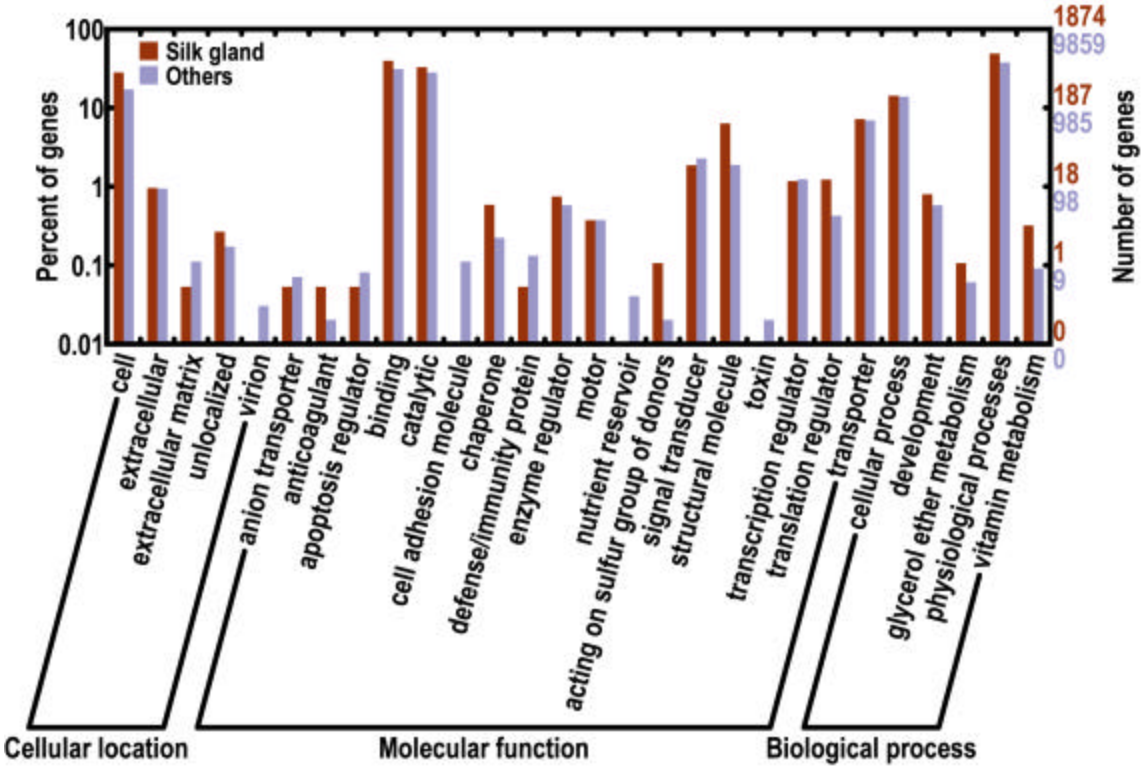
Silk gland
Others
Percent of genes
Number of genes
100
10
1
0.1
0.01
1874
9859
187
985
18
98
1
9
0
0
cell
extracellular
extracellular matrix
unlocalized
virion
anion transporter
anticoagulant
apoptosis regulator
binding
catalytic
cell adhesion molecule
chaperone
defense/immunity protein
enzyme regulator
motor
nutrient reservoir
acting on sulfur group of donors
signal transducer
structural molecule
toxin
transcription regulator
translation regulator
transporter
cellular process
development
glycerol ether metabolism
physiological processes
vitamin metabolism
Cellular location
Molecular function
Biological process

Figure S4.

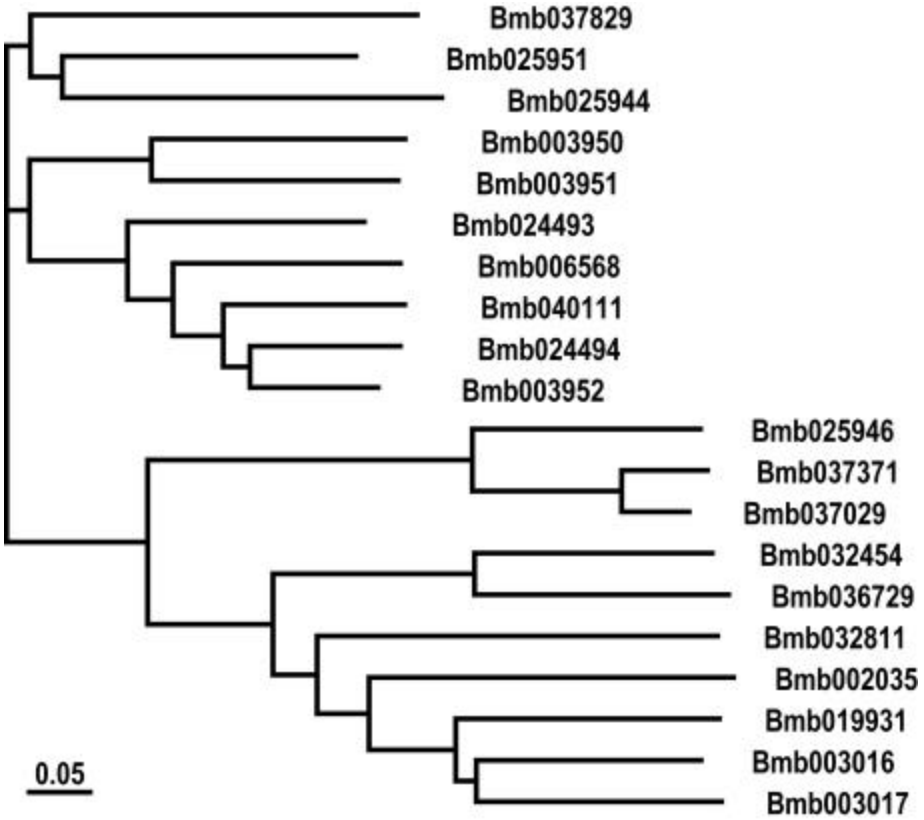
Bmb037829
Bmb025951
Bmb025944
Bmb003950
Bmb003951
Bmb024493
Bmb006568
Bmb040111
Bmb024494
Bmb003952
Bmb025946
Bmb037371
Bmb037029
Bmb032454
Bmb036729
Bmb032811
Bmb002035
Bmb019931
Bmb003016
Bmb003017
0.05

Figure S5.

Table captions

Table S1: Raw data in assembly. Clone insert sizes are given for 10^{th} to 90^{th} percentiles. Read lengths count Q20 bases with error rates below 10^{-2}. Effective coverage is defined by Q20 bases in non-repeated region of contigs over 5-Kb.

Table S2: Summary of assembled contigs and scaffolds. N50 is that size over which half of the total length of the sequence set is found. Equivalent size for unassembled reads is computed as number of Q20 bases divided by effective coverage of 5.9.

Table S3: Description of tissues sampled by expressed-sequence-tags (ESTs). We give the number of tags in each library, including redundancies.

Table S4: Completeness of assembly. Here, we search the WGS for silkworm full-length cDNAs, silkworm UniGene-EST clusters, and homologs of genes from other lepidoptera. For comparison within silkworm, we compute the fraction of the gene set (by length) that is aligned to the WGS, with a 95% match criterion. For comparison between lepidoptera, we use *TblastN* to search the WGS in all six reading frames at expectation values of 10^{-6} and count the number of genes with similarity over 50% of their length.

Table S5: List of genes searched in Table S4. For comparison within silkworm, similarity is based on a 95% match. For comparison between lepidoptera, similarity is based on E-values of 10^{-6}, and we give the identity in the *TblastN* hits.

Table S6: Comparisons to sequenced BACs in GenBank. We depict 3 tables, for subset of WGS sequence with error estimates better than 10^{-2} (Q20), 10^{-3} (Q30), and 10^{-4} (Q40). Mismatch rates are computed for aligned regions with sizes above 500-bp. We compute repeat content in regions with 20-mer copy numbers greater than 10 or 50, and in known transposable elements (TEs). We consider the entire BAC, and unaligned regions within each BAC. As a baseline, we show that 40.2% and 30.0% of the whole genome shotgun reads have 20-mer copy numbers greater than 10 or 50, respectively.

Table S7: Transposable elements (TEs) identified with *RepeatMasker*. Classes are LTR, LINE, SINE, or DNA. Each class is further subdivided into families, like *copia* and *gypsy*. Within each family, we show the number of TEs used to train *RepeatMasker*, their mean size, the number of bases identified from that family, and the fraction of the total genome or identified repeats attributed to the TEs from that family.

Table S8: Domain clustering procedure. We use pairwise comparisons, and require that the size of the homologous region exceeds 50% of the domain or 100 amino acids. A set of *n* domains is said to be a cluster if the number of acceptable homolog pairs exceeds a fraction *f* of the theoretical maximum, $(n-1)!$. Ideally, every cluster would correspond to a single InterPro category. In practice, this is not achievable, and we always find domain clusters with two or more InterPro categories, and InterPro categories scattered over two or more domain clusters. The best compromise parameter is *f*=0.25.

Table S9: Complete list of silkworm genes with similarity to existing genes or proteins in the databases, and highlighting genes discussed in text. A small number of genes that were not predicted by *BGF* but were identified through a *TblastN* homology search of the silkworm genome have been included. These are named "Bmp000001" to "Bmp000010". DNA and protein sequences are also provided, but as separate files.

Table S10: Summary of tRNA genes found by *tRNAScan*-SE. Abundances of Gly and Ala tRNAs in silkworm are consistent with fibroin production.

Table S1.

	Library 1	Library 2	Total data
insert size range	1.76k--2.61k	2.20k--7.97k	
sequenced reads	3,493,976	1,409,313	4,903,289
plasmid end pairs	1,763,694	721,995	2,485,689
mean Q20 length	520	514	518
shotgun coverage	4.24	1.69	5.93

Table S2.

	Number	N50 size (Kb)	Total size (Mb)
unassembled			31.0
contigs >2Kb	41,283	12.5	365.3
scaffolds >2Kb	23,155	26.9	397.7
genome size			428.7

Table S3.

	# of ESTs
Embryo (72 hours)	4,411
Embryo (nondiapause)	5,825
Embryo (unfertilized)	7,696
Fat body (f)	6,078
Fat body (m)	6,480
Fat body (pupa)	5,922
Hemocyte (f)	6,113
Hemocyte (m)	4,728
Midgut	7,214
Ovary	8,267
Silk gland	9,420
Testis	8,316
Total	**80,470**

Table S4.

	# of genes	% in WGS
silkworm full-length cDNA	212	90.9%
silkworm UniGene clusters	16,425	90.9%
other Lepidopteran genes	554	82.7%

Table S5. Attached as EXCEL files Silkworm-Functional_Coverage_Details.xls and OtherLp_Functional_Coverage_Details.xls.

Table S6.

GenBank BACs			# of pieces		alignments		the entire BAC			unaligned regions			
Accession	chr	size (bp)	in BAC	in WGS	coverage	mismatch	copyN >10	copyN >50	known TEs	copyN >10	copyN >50	known TEs	
AB090307	Z	151,992	2	13	77.8%	0.045%	27.2%	20.2%	16.8%	25.0%	18.7%	16.3%	
AB090308	Z	155,952	3	11	88.8%	0.078%	30.2%	20.9%	17.1%	24.4%	13.3%	9.6%	
AB159445	2	205,107	1	38	79.1%	0.102%	58.1%	44.1%	40.6%	51.6%	40.9%	45.7%	
AB159446	11	149,562	1	12	92.3%	0.030%	37.4%	26.3%	18.0%	48.2%	35.3%	22.6%	
AB159447	13	124,898	1	15	91.2%	0.063%	38.6%	27.3%	20.3%	50.3%	39.2%	27.9%	Q20

GenBank BACs			# of pieces		alignments		the entire BAC			unaligned regions			
Accession	chr	size (bp)	in BAC	in WGS	coverage	mismatch	copyN >10	copyN >50	known TEs	copyN >10	copyN >50	known TEs	
AB090307	Z	151,992	2	13	77.8%	0.040%	27.2%	20.2%	16.8%	25.0%	18.7%	16.3%	
AB090308	Z	155,952	3	11	89.0%	0.073%	30.2%	20.9%	17.1%	23.2%	11.8%	8.0%	
AB159445	2	205,107	1	38	80.5%	0.095%	58.1%	44.1%	40.6%	52.3%	41.3%	45.9%	
AB159446	11	149,562	1	12	92.8%	0.028%	37.4%	26.3%	18.0%	45.4%	37.7%	23.8%	
AB159447	13	124,898	1	15	91.2%	0.053%	38.6%	27.3%	20.3%	50.2%	39.1%	27.9%	Q30

GenBank BACs			# of pieces		alignments		the entire BAC			unaligned regions			
Accession	chr	size (bp)	in BAC	in WGS	coverage	mismatch	copyN >10	copyN >50	known TEs	copyN >10	copyN >50	known TEs	
AB090307	Z	151,992	2	13	79.7%	0.029%	27.2%	20.2%	16.8%	23.9%	17.1%	14.9%	
AB090308	Z	155,952	3	11	89.0%	0.055%	30.2%	20.9%	17.1%	23.2%	11.8%	8.5%	
AB159445	2	205,107	1	38	81.6%	0.082%	58.1%	44.1%	40.6%	52.4%	40.9%	46.1%	
AB159446	11	149,562	1	12	92.8%	0.020%	37.4%	26.3%	18.0%	45.3%	37.6%	23.8%	
AB159447	13	124,898	1	15	91.5%	0.041%	38.6%	27.3%	20.3%	48.9%	39.3%	27.1%	Q40

Table S7.

TE class	TE family	Number	Mean (bp)	Identified (bp)	% of genome	% of repeats
LTR	*copia*-like	2	4,653	34,290	0.0%	0.0%
LTR	*gypsy*-like	15	5,659	42,592,265	10.7%	50.8%
LTR	*pao*-like	6	3,357	1,263,367	0.3%	1.5%
LTR	others	52	4,998	553,112	0.1%	0.7%
LINE	LINE	71	3,954	26,724,468	6.7%	31.8%
SINE	SINE	4	319	5,730,723	1.4%	6.8%
DNA	*mariner*-like	47	1,068	6,686,025	1.7%	8.0%
DNA	*Tc*-like	4	1,555	140,867	0.0%	0.2%
DNA	others	19	1,704	127,158	0.0%	0.2%
unclassified		40	2,207	66,616	0.0%	0.1%
Total		260	3,205	83,918,891	21.1%	100.0%

Table S8.

Clustering parameter	Domain clusters	Clusters with >=2 categories	Max # of category per cluster	InterPro categories	Categories with >=2 clusters	Max # of cluster per category
0	6825	238	139	2740	986	499
0.25	8947	338	9	2740	1083	700
0.5	10123	378	7	2740	1179	736
0.75	12140	446	7	2740	1372	782

Table S9. Attachment is EXCEL file Silkworm-FromBiologySection.xls, along with Silkworm-PredictionsRelease.cds, Silkworm-PredictionsRelease.pep, Silkworm-TblastN-Homologs.cds, Silkworm-TblastN-Homologs.pep for the predicted genes and *TblastN* homologs that are cited therein.

Table S10.

Amino acid	*B.mori*	*D.melanogaster*	*A.gambiae*
Ala	**41**	**17**	**26**
Arg	28	23	22
Asn	25	8	12
Asp	26	14	20
Cys	9	7	5
Gln	11	12	15
Glu	25	16	26
Gly	**41**	**20**	**24**
His	13	5	21
Ile	12	12	14
Leu	26	23	23
Lys	1	19	27
Met	26	12	18
Phe	12	8	0
Pro	19	17	28
Ser	**24**	**20**	**22**
Thr	16	17	15
Trp	8	8	6
Tyr	**12**	**9**	**22**
Val	20	15	63