Sciencexpress Report

Complete Resequencing of 40 Genomes Reveals Domestication Events and Genes in Silkworm (*Bombyx*)

Qingyou Xia,[1,2]* Yiran Guo,[3]* Ze Zhang,[1,2]* Dong Li,[1,3]* Zhaoling Xuan,[3]* Zhuo Li,[3]* Fangyin Dai,[1] Yingrui Li,[3] Daojun Cheng,[1] Ruiqiang Li,[3,4] Tingcai Cheng,[1,2] Tao Jiang,[3] Celine Becquet,[5]† Xun Xu,[3] Chun Liu,[1] Xingfu Zha,[1] Wei Fan,[3] Ying Lin,[1] Yihong Shen,[1] Lan Jiang,[3] Jeffrey Jensen,[5] Ines Hellmann,[5] Si Tang,[5] Ping Zhao,[1] Hanfu Xu,[1] Chang Yu,[3] Guojie Zhang,[3] Jun Li,[3] Jianjun Cao,[3] Shiping Liu,[1] Ningjia He,[1] Yan Zhou,[3] Hui Liu,[3] Jing Zhao,[3] Chen Ye,[3] Zhouhe Du,[1] Guoqing Pan,[1] Aichun Zhao,[1] Haojing Shao,[3] Wei Zeng,[3] Ping Wu,[3] Chunfeng Li,[1] Minhui Pan,[1] Jingjing Li,[3] Xuyang Yin,[3] Dawei Li,[3] Juan Wang,[3] Huisong Zheng,[3] Wen Wang,[3] Xiuqing Zhang,[3] Songgang Li,[3] Huanming Yang,[3] Cheng Lu,[1] Rasmus Nielsen,[4,5] Zeyang Zhou,[1,6] Jian Wang,[3] Zhonghuai Xiang,[1]‡ Jun Wang[3,4]‡

[1]The Key Sericultural Laboratory of Agricultural Ministry, College of Biotechnology, Southwest University, Chongqing 400715, China. [2]Institute of Agronomy and Life Sciences, Chongqing University, Chongqing 400044, China. [3]Beijing Genomics Institute at Shenzhen, Shenzhen 518083, China. [4]Department of Biology, University of Copenhagen, Universitetsparken 15, 2100 Kbh Ø, Denmark. [5]Departments of Integrative Biology and Statistics, UC Berkeley, Berkeley, CA 94720, USA. [6]Chongqing Normal University, Chongqing 400047, China.

*These authors contributed equally to this work.

†Present address: Institute for Human Genetics, University of California, San Francisco, San Francisco, CA 94143–0794, USA.

‡To whom correspondence should be addressed. E-mail: xbxzh@swu.edu.cn (Z.X.); wangj@genomics.org.cn (J.W.)

A single-base-pair resolution silkworm genetic variation map was constructed from 40 domesticated and wild silkworms, each sequenced to ~3X coverage, representing 99.88% of the genome. We identified ~16 million SNPs, many indels and structural variations. We find that the domesticated silkworms are clearly genetically differentiated from the wild ones, but have maintained large levels of genetic variability, suggesting a short domestication event involving a large number of individuals. We also identified signals of selection at 354 candidate genes that may have been important during domestication, some of which have enriched expression in the silk gland, midgut and testis. These data add to our understanding of the domestication processes, and may have applications in devising pest control strategies and advancing use of silkworm as efficient bioreactors.

The domesticated silkworm, *Bombyx mori*, has a mid-range genome size of ~432 Mb (*1*), is the model insect for the order Lepidoptera, has economically important values (e.g., silk and bioreactors production), and has been domesticated for more than 5000 years (*2*). Due to human selection they have evolved complete dependence on humans for survival (*3*) and more than 1000 inbred domesticated strains are kept worldwide (*3*). Archaeological and genetic evidences indicate that the domesticated silkworm originated from the Chinese wild silkworm, *Bombyx mandarina*, which occurs throughout Asia, where modern sericulture and silkworm domestication was initiated.

The origin of the domesticated silkworm is a longstanding question which has not been settled by previous limited biochemical and molecular analyses. Two hypotheses proposed a unique domestication but disagreed on the ancestral variety. One, based on isoenzyme polymorphism, proposed mono-voltinism as ancestral (voltinism represents number of generations per annum), from which bi- and multi-voltine were derived by artificial selection (*4*), whereas the other proposed the reverse path considering evidence from archaeology, history and genetics (*5*). An alternative hypothesis based on random amplification of polymorphic DNA argued that the ancestral domestic silkworm strains were issued not from a unique variety, but from mixed geographic locations and ecological types (*6*). These theories are conflicting likely because they were derived from incomplete genetic information. Consequently, we present here a genome-wide detailed genetic variation map in hope to help reconstruct the silkworm domestication history.

The data consisted of 40 samples from 29 phenotypically and geographically diverse domesticated silkworm lines [categorized by geographical regions (*3*): Chinese, Japanese, Tropical, European lineages, and the mutant system], as well as 11 wild silkworms from various mulberry fields in China (table S1). We sequenced each genome at ~threefold

coverage, after creating single- and paired-end (PE) libraries with inserts of PE ranging from 137 to 307 bp (*7*).

Raw short reads were mapped against the refined 432 Mb reference genome from *Dazao* (*1*) with the program SOAP (*8*). We pooled all reads from the 40 complete genomes and identified 15,986,559 SNPs using SoapSNP (*7, 9*) (table S3A). The accuracy of the SNP calling was evaluated with Sequenom genotyping of a representative subset of variants in all 40 varieties, resulting in a 96.7% validation rate (*7*).

We then pooled separately all 29 domesticated strains and all 11 wild varieties and obtained SNP sets for each (*7*). The number of SNPs in the domestic versus wild varieties was 14,023,573 and 13,237,865, respectively (table S3A). To account for the different number of domestic and wild strains, we measured genetic variation using the population size scaled mutation rate θ_S (*10*) (table S3B). We found that $\theta_{S,domesticated}$ (0.0108) was significantly smaller than $\theta_{S,wild}$ (0.0130) [Mann Whitney U (MWU), $P = 1.10 \times 10^{-7}$], which may reflect differences in effective population size and demographic history (including domestication and artificial selection). The rate of heterozygosity in domesticated strains was more than twofold lower than that of wild varieties (0.0032 versus 0.0080, respectively) (MWU, $P = 3.33 \times 10^{-6}$). This reduction in heterozygosity is most likely due to inbreeding or the bottleneck experienced by domesticated lines.

In addition to SNPs, we also identified 311,608 small insertion-deletions (indels) (table S4A), a subset of which were validated with PCR (*7*). The θ_S values for the indels (table S4B) were in agreement with a lower effective population size in domesticated vs. wild varieties. A mate-pair relationship method (*7*, *11*) identified 35,093 structural variants (SVs) among the 40 varieties (table S5). Over three-fourths of the SVs overlapped with transposable elements (TE), suggesting that SV events in silkworm are likely due to TE content (*12*) and mobility (*11*). The SNPs, indels and SVs all contributed to a comprehensive genetic variation map for the silkworm.

In order to elucidate the phylogeny of silkworms beyond previous studies (*6*, *13*, *14*) we used our identified SNPs to estimate a neighbor-joining tree (*7*) on the basis of a dissimilarity measure of genetic distance (Fig. 1A). We note that this tree represents an average of distances among strains, so lineages cannot be directly interpreted as representing phylogenetic relationships. Instead, the distances may reflect gene-flow and other population level processes related to human activities such as ancient commercial trade. Importantly, the unrooted radial relationship reveals a clear split between the domesticated and wild varieties, and the domestic strains cluster into several subgroups (Fig. 1A).

A principle component analysis (PCA) (*7*) had the first four eigenvectors significant (table S6; Tracy-Widom, $P <$ 0.05). The first eigenvector clearly separates the domesticated and wild varieties while the second eigenvector divides the domesticated strains into subgroups correlated with voltinism (Fig. 1B, top). The third principle component separates D01 and D03, which are high-silk producing Japanese domesticated strains, from the other domesticated strains while the fourth separates W01 and W04 from the other wild varieties (Fig. 1B, bottom). Results of population structure analysis (*7*) (fig. S3) confirmed those from the neighbor-joining and PCA analyses. The clear genetic separation between domesticated and wild varieties suggests a unique domestication event and relatively little subsequent gene-flow between the two groups.

One puzzling observation is that, while domesticated strains are clearly genetically differentiated from the wild ones, they still harbor ~83% of the variation observed in the wild varieties. This suggests that the population size bottleneck at domestication only reduced genetic variability mildly (*7*), i.e. a large number of individuals must have been selected for initial domestication or else domestication occurred simultaneously in many places. To quantify this we fit a simple coalescence-based genetic bottleneck model to the SNP frequency spectrum (*7*). The estimated model suggests that the domestication event lead to a 90% reduction in effective population size during the initial bottleneck (fig. S2). We additionally observed no excess of low frequency variants in the domesticated varieties compared to the wild varieties, suggesting that there has not been significant population growth since the domestication event and that the domestic lines likely have had a generally stable effective population size.

Our measure of pairwise linkage disequilibrium (LD) (*7*) showed that LD decays rapidly in silkworms, with r^2 decreasing to half of its maximum at a distance of ~46 bp and 7 bp for the domesticated and wild varieties, respectively (fig. S1). The fast decay of LD implies that regions affected by selective sweeps are likely relatively small. To detect regions with significant signatures of selective sweep, we measured SNP variability and frequency spectrum following a genome-wide sliding window strategy (*7*) (Fig. 2A). While the significance of our Z-tests (*7*) cannot be interpreted literally due to correlations in LD and shared ancestral history between the two populations, they suggest differences in frequency spectra and amounts of variability between the two groups. We termed the candidate regions genomic regions of selective signals (GROSS).

We identified a total of 1041 GROSS (*7*), covering 12.5 Mb (2.9%) of the genome, which may reflect genomic footprints left by artificial selection during domestication. A region affected by selective sweep typically has elevated level of LD (*15*, *16*), and in our GROSS, the level of LD among SNP pairs less than 20 Kb apart was 2.3 times higher than

genome average (Fig. 2B), consistent with the hypothesis that selection is affecting these regions. In all these regions, divergence levels (*7*) between the domesticated and wild groups were also elevated (Fig. 2C), confirming the differentiation of the two subpopulations.

B. mori have experienced intense artificial selection, represents a completely domesticated insect (*3*) and has become totally dependent on humans for survival. Artificial selection has also enhanced important economic traits such as cocoon size, growth rate and digestion efficiency (*3*). Moreover, compared to its wild ancestor *B. mandarina*, *B. mori* has gained some representative behavioral characteristics (such as tolerance to human proximity and handling, and extensive crowding) and has lost other traits (such as flight, predators and diseases avoidance). However, to date no genes have been identified as domestication genes under artificial selection. Within GROSS, we identified 354 protein-coding genes that represent good candidates for domestication genes (table S9). Their GO annotation (*17*) showed the most representation in the category of "binding" and "catalytic" in molecular function, and "metabolic" and "cellular" in biological process (fig. S4).

Considering published expression profiles performed on different tissues in fifth-instar day 3 of *Dazao* with genome-wide microarray (*18*), we found that 159 of our GROSS genes exhibit differential expression. Of these, 4, 32, and 54 genes are enriched in tissues of silk gland, midgut, and testis, respectively (fig. S5). Among the genes enriched in the silk gland is silk gland factor-1 (*Sgf-1*), a homolog of a *Drosophila melanogaster Fkh* gene. *Sgf-1* regulates the transcription of the *B. mori* glue protein-encoding *sericin-1* gene and of three fibroin genes encoding fibroin light chain, fibroin heavy chain and *fhx/P25* (*19*, *20*). Another silk gland enriched gene *BGIBMGA005127*, homologous to the *Drosophila sage* gene, was over-expressed fourfold in a high-silk strain compared to *Dazao* (fig. S6). In *Drosophila*, the products of *Fkh* and *sage* genes cooperate to regulate the transcription of the glue genes *SG1* and *SG2*, which are crucial for the synthesis and secretion of glue proteins (*21*, *22*). Additionally, midgut- and testis- enriched genes suggest that genes involved in energy metabolism and reproduction have also been under artificial selection during domestication (*7*). Specifically, we identified three likely candidate for artificial selection: *NM_001130902* is homologous to paramyosin protein in *Drosophila* and may be related to flight (*23*); *NM_001043506* is homologous to fatty-acyl desaturase (*desat1*) in *Drosophila*, which is related to courtship behaviors since mutations in *desat1* can change the pattern of sex pheromones production and discrimination (*24*); and finally *BGIBMGA000972* is homologous to tyrosine-protein kinase *Btk29A* in *Drosophila*, which is involved in male genitalia development (*25*).

In sericulture, silkworms are typically categorized by their geographic origins (*3*). Voltinism, which results from adaption to ecological conditions, as well as geographic systems have been central to previous studies of silkworm origin and domestication (*4–6*). Our findings indicate that a unique domestication event occurred and, while voltinism correlates with genetic distances, major genetically cohesive strains cannot be identified on the basis of voltinism. We observed no correlation between longitudes of the sample origins and any of the principle components, but found a significant correlation between the latitudes and eigenvectors 2 and 4 in the PCA (table S7). Although this correlation might be due to isolation by distance, this result also agrees with previous studies suggesting that climate affects silkworm biology (*2*).

The silkworm data reported here represents a large body of genome sequences for a lepidopteran species and offers a source of near-relatives in this clade for comparative genomic analysis. We further proposed a set of candidate domestication genes that, in addition to being putatively under artificial selection, also show higher expression levels in tissues important for silkworm economic traits. Because a proportion of the GROSS genes were likely important in domestication, functional verification of these candidate genes may enable a comprehensive understanding of the differences of biological characteristics between *B. mori* and *B. mandarina*. Moreover, domesticated silkworms have been used as bioreactors (*26*, *27*), and such an effort may provide useful clues to help improve the capacity and capability of silkworm to produce foreign proteins (*26*). These findings may also aid in understanding how to enhance traits of interest in other organisms in an environmentally safe manner and, because the wild silkworm is a destructive pest, allow new approaches for pest control.

References and Notes

1. The International Silkworm Genome Consortium, *Insect Biochem. Mol. Biol.* **38**, 1036 (2008).
2. Z. Xiang, J. Huang, J. Xia, C. Lu, *Biology of Sericulture* (China Forestry Publishing House, Beijing, 2005).
3. M. R. Goldsmith, T. Shimada, H. Abe, *Annu. Rev. Entomol.* **50**, 71 (2005).
4. N. Yoshitake, *J. Sericult. Sci. Japan* **37**, 83 (1967).
5. Y. Jiang, *Agric. Archaeol.* **14**, 316 (1987).
6. C. Lu, H. Yu, Z. Xiang, *Agric. Sci. China* **1**, 349 (2002).
7. Materials and methods are available as supporting material on *Science* Online.
8. R. Li, Y. Li, K. Kristiansen, J. Wang, *Bioinformatics* **24**, 713 (2008).
9. R. Li *et al*., *Genome Res.* **19**, 1124 (2009).
10. G. A. Watterson, *Theor. Popul. Biol.* **7**, 256 (1975).
11. J. Wang *et al*., *Nature* **456**, 60 (2008).
12. Q. Xia *et al*., *Science* **306**, 1937 (2004).

13. Q. Xia, Z. Zhou, C. Lu, Z. Xiang, *Acta Entomol. Sinica* **41**, 32 (1998).
14. M. Li *et al*., *Genome* **48**, 802 (2005).
15. R. Nielsen, *Annu. Rev. Genet.* **39**, 197 (2005).
16. M. Slatkin, *Nat. Rev. Genet.* **9**, 477 (2008).
17. J. Ye *et al*., *Nucleic Acids Res.* **34**, W293 (2006).
18. Q. Xia *et al*., *Genome Biol.* **8**, R162 (2007).
19. B. Horard, E. Julien, P. Nony, A. Garel, P. Couble, *Mol. Cell. Biol.* **17**, 1572 (1997).
20. V. Mach *et al*., *J. Biol. Chem.* **270**, 9340 (1995).
21. E. W. Abrams, W. K. Mihoulides, D. J. Andrew, *Development* **133**, 3517 (2006).
22. T. R. Li, K. P. White, *Dev. Cell* **5**, 59 (2003).
23. H. Liu *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **102**, 10522 (2005).
24. F. Marcillac, Y. Grosjean, J. F. Ferveur, *Proc. Biol. Sci.* **272**, 303 (2005).
25. K. Baba *et al*., *Mol. Cell. Biol.* **19**, 4405 (1999).
26. S. Maeda, *Annu. Rev. Entomol.* **34**, 351 (1989).
27. S. Maeda *et al*., *Nature* **315**, 592 (1985).
28. We thank two anonymous referees, L. Goodman, L. Bolund, and K. Kristiansen for providing valuable comments. This work was supported by China (2005CB121000, 2007CB815700, 2006AA10A117, 2006AA10A118, 2006AA02Z177, 2006AA10A121), the Ministry of Education of China (Program for Changjiang Scholars and Innovative Research Team in University, IRT0750), Chongqing Municipal Government, the 111 Project (B07045), the National Natural Science Foundation of China (30725008, 30890032, 90608010), the International Science and Technology Cooperation Project(0806), the Chinese Academy of Science (GJHZ0701-6), the Danish Platform for Integrative Biology, the Ole Rømer grant from the Danish Natural Science Research Council, and the Solexa project (272-07-0196). Raw genome data are deposited in NCBI/SRA with accession SRA009208; silkworm genetic variations, GROSS information, and microarray data can be found in http://silkworm.swu.edu.cn/silkdb/resequencing.html.

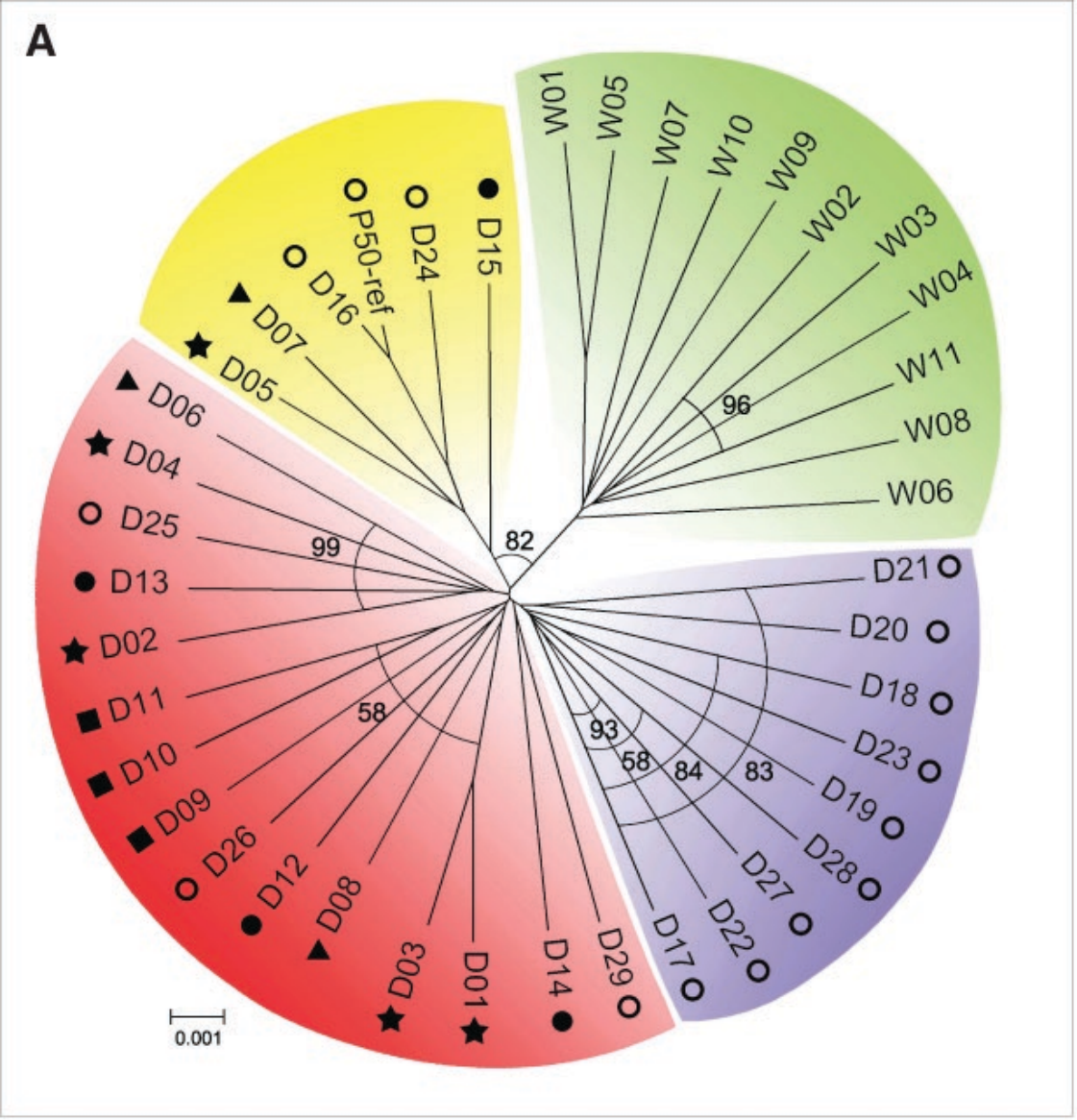
Supporting Online Material

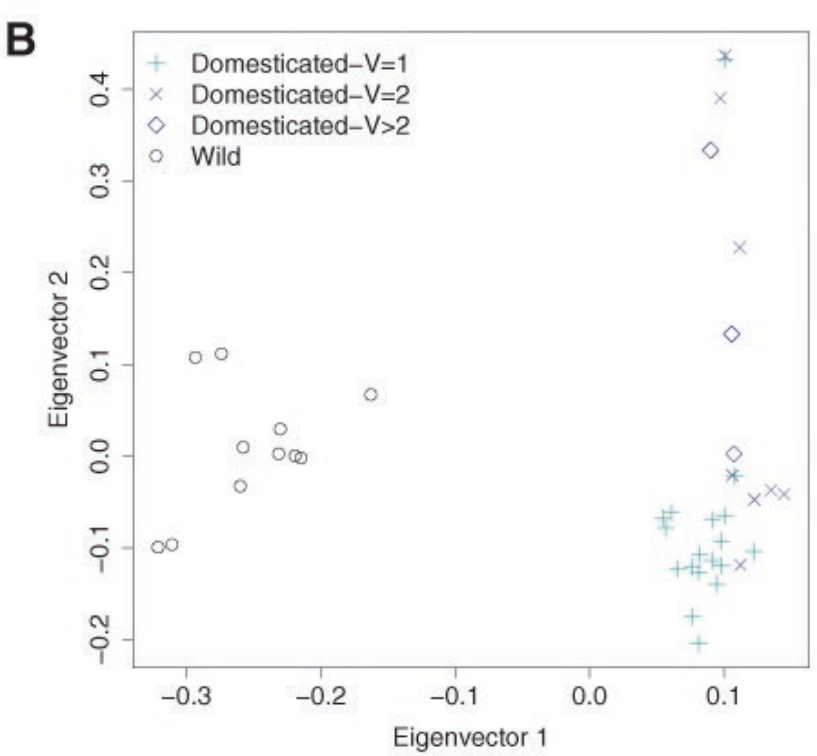
www.sciencemag.org/cgi/content/full/1176620/DC1
Materials and Methods
SOM Text
Figs. S1 to S7
Tables S1 to S10
References

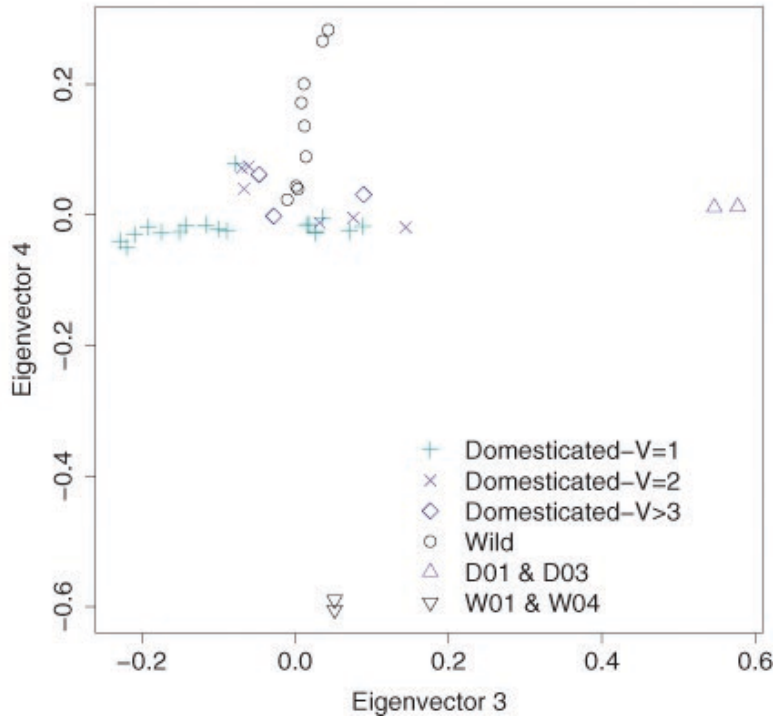
21 May 2009; accepted 12 August 2009
Published online 27 August 2009; 10.1126/science.1176620
Include this information when citing this paper.

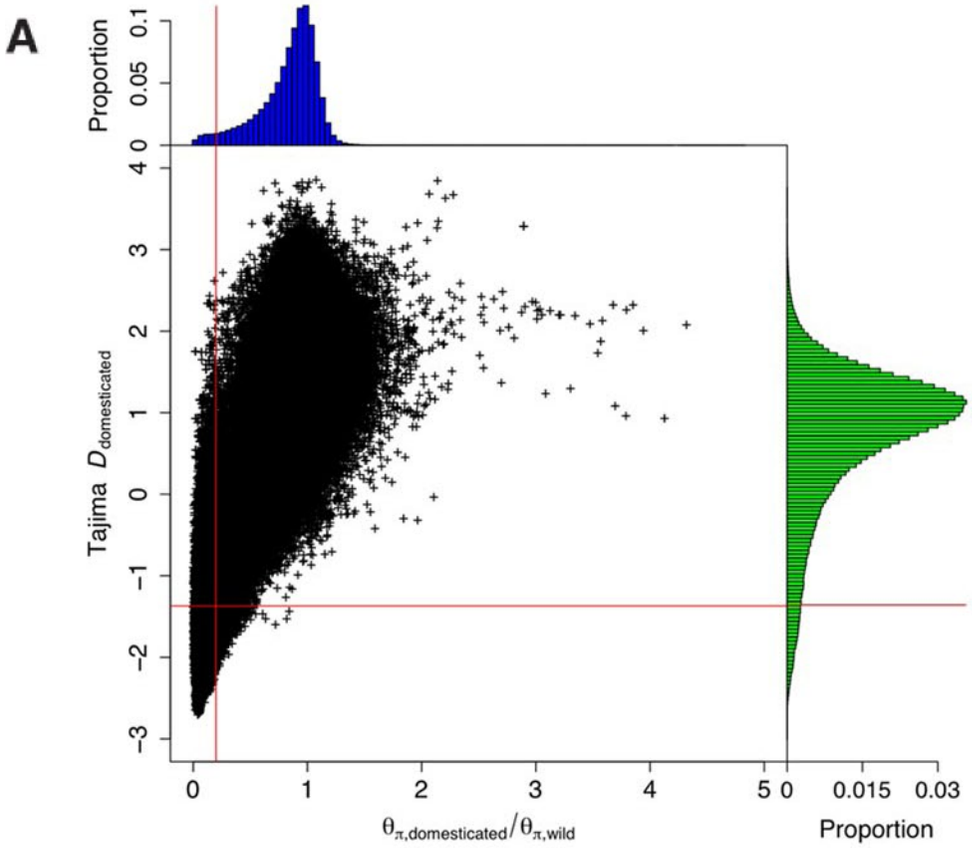
Fig. 1. Silkworm phylogeny and population structure from PCA. (**A**) A neighbor-joining tree from genomic SNPs, bootstrapped with 1000 replicates (bootstrap values less than 100 are shown on arcs, and those equal to 100 are not shown): green for all wild varieties; others are domesticated strains separated into three groups (purple, red, and yellow). Domesticated strains are denoted by a combination of symbols representing silkworm systems (hollow circles for Chinese, star for Japanese, triangles for Tropical, boxes for European and filled circles for the mutant system) and sample IDs ("D01" to "D29," and "P50-ref" for the reference genome of *Dazao*). Wild varieties are indicated by their IDs ("W01" to "W12"). Frequencies of base-pair differences are shown as a scale bar at the bottom left. (**B**) PCA results of the first four significant components. (Upper panel) The first eigenvector separating domesticated and wild varieties, and the second dividing the domesticated strains into subgroups. (Lower panel) The third eigenvector separating the high-silk production Japanese domesticated strains D01 and D03 from the other domesticated strains, and the fourth separating the wild varieties W01 and W04 from the other wild varieties.

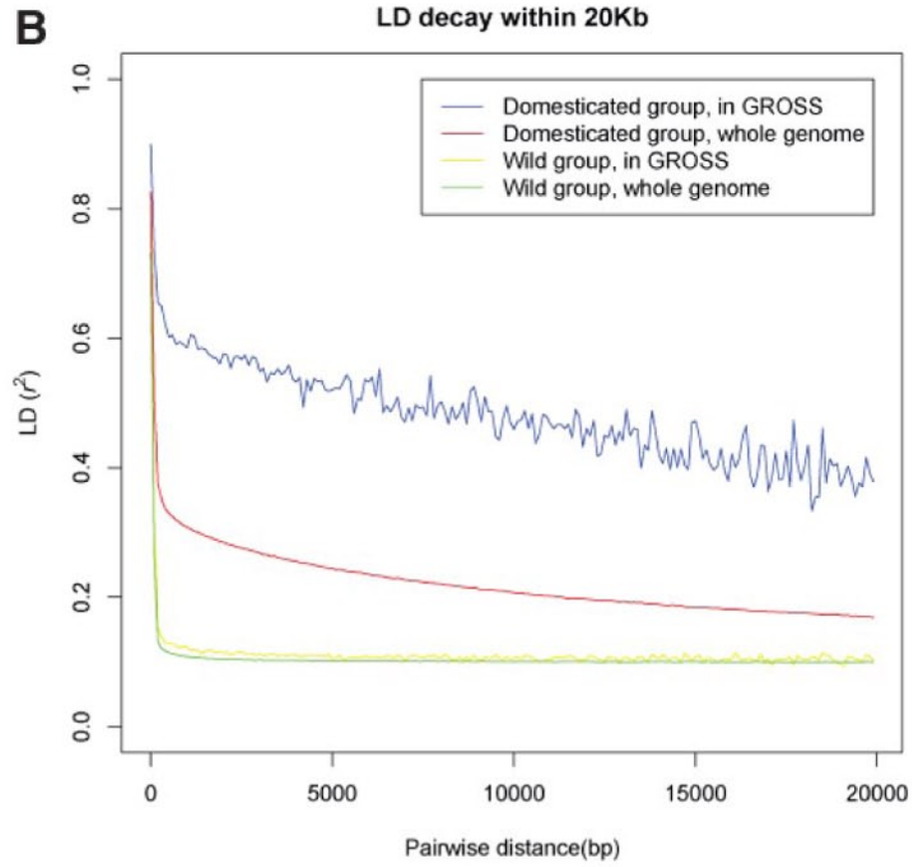
Fig. 2. Genomic regions of selective signals (GROSS). (**A**) 2D distribution for $\theta_{\pi,\text{domesticated}}/\theta_{\pi,\text{wild}}$ and Tajima's *D* for domesticated silkworms. 5-Kb windows, data points of which locate to the left of the vertical red line (corresponding to Z-test $P < 0.005$) and below the horizontal red line (also Z-test $P < 0.005$) were picked out as building blocks of GROSS. (**B**) Linkage disequilibrium (LD) in GROSS. For domesticated silkworms, LD decays much slower in GROSS than in the whole genome, whereas for wild varieties, no obvious change in the pattern was observed. (**C**) Distribution of F_{st} between domesticated and wild groups in GROSS versus whole genome.

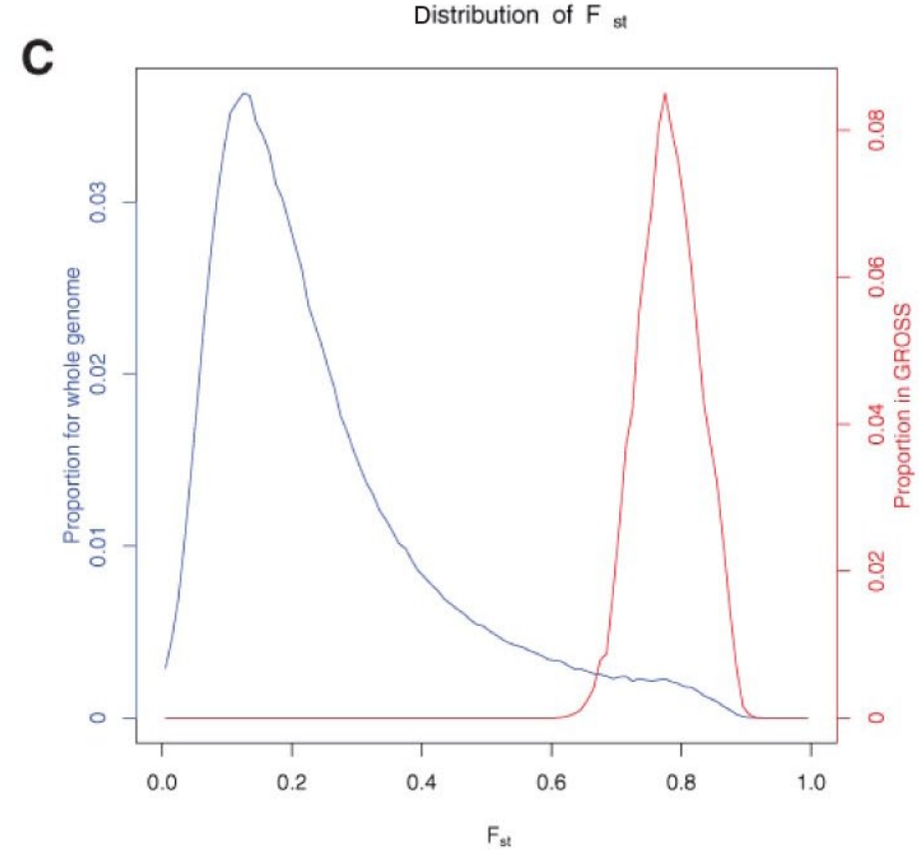

A
W01
W05
W07
W10
W09
W02
W03
W04
W11
W08
W06
D21
D20
D18
D23
D19
D28
D27
D22
D17
D29
D14
D01
D03
D08
D12
D26
D09
D10
D11
D02
D04
D06
D25
D13
D05
D07
D16
P50-ref
D24
D15
96
82
99
58
93
58
84
83
0.001


B
Domesticated–V=1
Domesticated–V=2
Domesticated–V>2
Wild
Eigenvector 2
Eigenvector 1
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.2
−0.1
0.0
0.1


Eigenvector 4
Eigenvector 3
0.2
0.0
−0.2
−0.4
−0.6
−0.2
0.0
0.2
0.4
0.6
Domesticated–V=1
Domesticated–V=2
Domesticated–V>3
Wild
D01 & D03
W01 & W04


A
Proportion
Tajima $D_{domesticated}$
$\theta_{\pi,domesticated}/\theta_{\pi,wild}$
Proportion


B
LD decay within 20Kb
Domesticated group, in GROSS
Domesticated group, whole genome
Wild group, in GROSS
Wild group, whole genome
LD (r^2)
Pairwise distance(bp)


C
Distribution of F_{st}
Proportion for whole genome
Proportion in GROSS
F_{st}