Supporting Information

d'Alençon et al. 10.1073/pnas.0910413107

SI Materials and Methods

Construction of BAC Libraries. We used BAC libraries from the two noctuid pests, *H. armigera* (average insert size 114 kb, $n = 332$, genome coverage 10.5×) and *S. frugiperda* (average insert size 125 kb, $n = 80$, genome coverage 10×). The *S. frugiperda* BAC library construction has been described previously (1). The very similar *H. armigera* BAC library was constructed at Clemson University Genomics Institute (CUGI), from high molecular weight DNA from whole pupal cells of the Australian Toowoomba strain, partially digested with HindIII and cloned into plasmid vector pIndigoBAC536. It is available from the Clemson University Genomics Institute (*http://www.genome.clemson.edu/*) as library Ha_Tba.

BAC Selection and Contig Construction. BAC library screening was performed by hybridization according to (1). Probes are listed in Table S1. The genes used as anchors and probes for the selection of BACs ranged from highly conserved genes (acetylcholinesterase 1, *ace-1*; phosphoglucose isomerase, *PGI*; ribosomal proteins *RpPL10A* and *RpL5*; TATA binding protein, *TBP*) to rapidly evolving, lepidopteran-specific genes (P450 genes of the CYP332A and CYP9A families). We used also conserved genes coding for enzymes and receptors involved in insect-specific pathways (aminopeptidases; the ecdysone receptor, *EcR*; Ultraspiracle, *USP*; juvenile hormone acid methyl transferase, *JHAMT*; the olfactory receptor, *Or83b*). When the probe specific for one species was lacking, a heterologous probe was used (consequently, washes of BACs filters after hybridization was limited to a single 2× SSC step). Most of the probes were whole cDNAs obtained by PCR amplification with primer pairs flanking the cloning site in the corresponding vector as described in Table S1, except for the *Sf* TBP probe that was obtained by PCR amplification from genomic DNA as template and the *Sl* Or2 probe that was obtained by restriction of the plasmid of interest. BAC DNA of the positive hit clones was isolated and digested with the restriction enzyme HindIII. The HindIII fingerprints were analyzed for contig construction with the fpc software according to Marra et al. (2) at a tolerance of seven and a cutoff of 10^{-7}. The most central in the contig and longest BAC was chosen for sequencing, usually after checking for the presence of the gene used as probe by PCR.

BAC Sequencing. After mechanical shearing of BAC DNA, 5 kbp DNA fragments were cloned into vector plasmid pcDNA 2.1 (Invitrogen). For two BAC sequences, we used an additional library, with 10-kb inserts into the pCNS vector (pSU18-derived). Vector DNA was purified and end-sequenced using dye terminator chemistry on ABI 3730 sequencers (Applied Biosystems) at approximately 12× coverage. The assemblies were realized using the Phred/Phrap/Consed software package. Primer walks, PCRs, and transposon bombs were needed for the finishing phases. Each BAC was sequenced to a finished level of accuracy.

Gene Annotation and Synteny Analysis. Genes were detected using KAIKOGAAS, a tool adapted to silkworm genes detection from RiceGAAS (3) by T. Shimomura at the National Institute for Agrobiological Sciences of Japan. The KAIKOGAAS gene prediction procedure includes four steps:

1. Normalize the scores for each exon predicted by the six prediction tools (Genscan_Human, Fgenesh_Anopheles, Fgenesh_Celegans, Fgenesh_Drosophila, Fgenesh_Honey bee, and Fgenesh_Tribolium) (4, 5). The normalized score is calculated as follows: (x − average) / SD.
2. Adjust the score for each exon predicted by each prediction tool.
   2.1 The scores are counted for each exon. If the exon in the first prediction tool overlaps with the exon from the second prediction tool, increase the score by 0.6. Compare the resulting exon with the third prediction tool and add 0.6 if there is an overlap. Repeat the same procedure for the remaining prediction tools.
   2.2 If start or stop codon is present, increase the score by 0.4.
   2.3 If a predicted exon position by one predicted tool is included in a hit region of BLASTn searches for cDNAs (ID ≥98%, e-value $\leq e^{-20}$), or if a predicted exon position by one predicted tool is overlapped with a hit region of BLASTx searches for nonredundant protein, increase the score by 0.6.

Note: The score is absolutely operated as a unit, not at a single nucleotide level.

3. Add the scores of the whole exons of a particular gene from each prediction tool and divide the total score of the gene by the number of exons to obtain the average exon score of each prediction tool. Adopt a prediction tool of ORF structure with the highest average exon score irrespective of the number of exons and the length of nucleotide sequence.

Compare the results of the first and second prediction tools. If there is an overlap, choose the higher average (predicted gene) score. Then compare this result with the result of the third prediction tool. If there is an overlap, choose the higher average (predicted gene) score and compare the result with the next prediction tool. Do the same for the other remaining prediction tools. Finally, compare the aforementioned processed result with the results for each of the six prediction tools. If there is no overlap in a predicted gene, add that predicted gene in the above result as a candidate predicted gene. (Or alternatively, choose the highest average exon score for a predicted gene of KAIKOGAAS among six scores.)

4. Insert the predicted internal exon from MZEF (6) into this result if it is compatible with the adopted ORF structure.

The function of a predicted gene model is predicted using GFSelectorK, the Gene function Selector for Kaiko (*http://sgp.dna.affrc.go.jp/tool/index.html*).

To visualize synteny at the nucleotide level, BAC sequences were compared to each other by Zpicture analysis according to Ovcharenko et al. (7) (*http://zpicture.dcode.org/*). The pipeline for definition of the gene categories is outlined in Fig. S1. In brief, we started with the KAIKOGAAS scheme that provides several curated gene categories. "ID" indicates genes with ID that had predicted ORFs similar to known protein by Blastp (threshold 10^{-40}). "UP" indicates unknown proteins that corresponded to genes that scored highly to silkworm EST in the case of *B. mori*, or to the cognate species EST in the case of the noctuid species. The third category was HP. These were further curated manually, and we kept genes with the following criteria: ORF size threshold, 50 aa; existence of a significant match to an EST in Butterflybase (8) or Spodobase (9) or NCBI dbEST library, or a peptide at NCBI or RepBase (10, 11) or Butterflybase, and

protein motif (NCBI). In a final selection step, we also kept a KAIKOGAAS annotated HP when a similar reciprocal best-hit HP was found in synteny in one of the other species. HPs that did not match these criteria were not counted as genes in our analysis. Furthermore, genes that were annotated as reverse transcriptase or transposase were flagged by the repeated sequence pipeline (as described later) and these were kept out of our gene set.

Repeats Detection and Annotation. Given a set of lepidopteran genomic sequences, here BACs from *B. mori*, *S. frugiperda*, and *H. armigera*, we used the REPET pipelines to detect and characterize TEs. The REPET pipeline is divided in two parts: the de novo pipeline and the annotation pipeline. In the first phase, the de novo TE identification and characterization was done by (*i*) searching repeats with BLASTER for an all-by-all genome comparison (9); (*ii*) grouping results using three clustering methods: GROUPER (12), RECON (13), and PILER (14); (*iii*) building one consensus per group with the MAFFT multiple sequence alignment program (15); and (*iv*) classifying each consensus according to structural and coding TE features (10, 16). The output of the de novo part of the REPET pipelines was a de novo TE consensus library for each species.

In the second phase, (i.e., the REPET annotation pipeline), the genome was annotated with the de novo TE consensus library and Repbase Update (10) by (*i*) detecting the TE from the de novo library with the software BLASTER, RepeatMasker, and Censor; (*ii*) finding the simple sequence repeats using the software RepeatMasker, TRF (16), and Mreps (17); and (*iii*) running BLASTx against Repbase Update amino acid sequences (10) to detect TE that had diverged a lot or had not enough copies to build a good consensus. The result of this second phase were GFF3 and/or gameXML files recording the de novo TE consensus library, the Repbase Update reference database, and the simple sequence repeats annotations.

The Lepido-DB Information System. To facilitate the analyses and the exploration of the data generated during this project, an information system was build using open-source software tools from the Generic Model Organism Database including a Chado database (18), a simple but rapid genome browser [Gbrowse (19)], a graphical tool the navigation within multiple maps or genome sequences (Cmap), and an application for the manual curation of the gene [Apollo (20)]. All these tools are available to the scientific community through a Web portal (*http://www.inra.fr/lepidodb*). This Web site also includes a BLAST search and full text search facilities.

1. d'Alencon E, et al. (2004) A genomic BAC library and a new BAC-GFP vector to study the holocentric pest Spodoptera frugiperda. *Insect Biochem Mol Biol* 34: 331–341.
2. Marra MA, et al. (1997) High throughput fingerprint analysis of large-insert clones. *Genome Res* 7:1072–1084.
3. Sakata K, et al. (2002) RiceGAAS: an automated annotation system and database for rice genome sequence. *Nucleic Acids Res* 30:98–102.
4. Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA. *J Mol Biol* 268:78–94.
5. Salamov AA, Solovyev VV (2000) Ab initio gene finding in Drosophila genomic DNA. *Genome Res* 10:516–522.
6. Zhang MQ (1997) Identification of protein coding regions in the human genome by quadratic discriminant analysis. *Proc Natl Acad Sci USA* 94:565–568.
7. Ovcharenko I, Loots GG, Hardison RC, Miller W, Stubbs L (2004) Picture: dynamic alignment and visualization tool for analyzing conservation profiles. *Genome Res* 14: 472–477.
8. Papanicolaou A, Gebauer-Jung S, Blaxter ML, Owen McMillan W, Jiggins CD (2008) ButterflyBase: a platform for lepidopteran genomics. *Nucleic Acids Res* 36(Database issue):D582–D587.
9. Negre V, et al. (2006) SPODOBASE: an EST database for the lepidopteran crop pest Spodoptera. *BMC Bioinformatics* 7:322.
10. Jurka J (2000) Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet* 16:418–420.
11. Jurka J (1998) Repeats in genomic DNA: mining and meaning. *Curr Opin Struct Biol* 8: 333–337.
12. Quesneville H, Nouaud D, Anxolabehere D (2003) Detection of new transposable element families in Drosophila melanogaster and Anopheles gambiae genomes. *J Mol Evol* 57 (suppl 1):S50–S59.
13. Bao Z, Eddy SR (2002) Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res* 12:1269–1276.
14. Edgar RC, Myers EW (2005) PILER: identification and classification of genomic repeats. *Bioinformatics* 21(suppl 1):i152–i158.
15. Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* 30: 3059–3066.
16. Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* 27:573–580.
17. Kolpakov R, Bana G, Kucherov G (2003) mreps: Efficient and flexible detection of tandem repeats in DNA. *Nucleic Acids Res* 31:3672–3678.
18. Mungall CJ, Emmert DB (2007) A Chado case study: an ontology-based modular schema for representing genome-associated biological information. *Bioinformatics* 23:i337–i346.
19. Stein LD, et al. (2002) The generic genome browser: a building block for a model organism system database. *Genome Res* 12:1599–1610.
21. Lewis SE, et al. (2002) Apollo: a sequence annotation editor. *Genome Biol* 3: research0082.1–research0082.14.

Fig. S1. Scheme of annotation. Genes predictions by KAIKOGAAS were curated according to this scheme to classify genes as TE, known genes (ID), unknown protein genes (UP), and HP genes. KAIKOGAAS predicted genes that did not fall into one of these categories were removed as invalid genes. The decision scheme depicted here involved not only BLAST searches but also evidence from the syntenic relationships and therefore manual annotations; it is therefore not an automated pipeline (also see *Materials and Methods* in the main text).

Fig. S1. (DOC)

Fig. S2. Schemes of synteny organization of all three species triplets. Red, genes used as probes; gray, genes with an ID predicted by KAIKOGAAS; pink, TE encoded genes; yellow, other genes; UP, unknown protein (presence of a match to an EST with a threshold of 10^{-40} by BlastN). Synteny links between orthologues in one to one correspondence (as defined in *Materials and Methods* in the main text) are shown as black lines. When a gene is duplicated in one or the two other species, the best paralogues (presumed orthologues, defined by a phylogenetic analysis as defined in *Materials and Methods* in the main text) are joined by a black line and the other paralogues are joined by dotted black lines. *H. armigera* BACs are shown twice for a better visualization of synteny. Synteny blocks are shown as blue squares at the edge of genes boxes and brown squares when genes are inverted. The symbols B1 to B10 refer to the block sizes (gene numbers). Red letters D, I, and T stand for duplication, inversion, and transposition, respectively.

Fig. S2. (PDF)

Fig. S3. Example of ancient duplications, the *APN* cluster. Arrows labeled *APN-1* through *APN-7* represent orthologues of the seven *APN* genes described in *H. armigera* (1) found in exactly the same order and orientation in the three lepidopteran species. The additional *APN* gene discovered in the course of this study corresponds to *APN-8*. *APN-M* corresponds to a member of the more distantly related protease m1 zinc metalloproteases found in insects and vertebrates. The maximum likelihood phylogenetic tree under the cluster shows the duplications history. It was constructed from a Clustal 2 sequence alignment of these proteins (*http://mobyle.pasteur.fr/cgi-bin/portal.py*).

Fig. S3. (PDF)

1. Angelucci C, et al. (2008) Diversity of aminopeptidases, derived from four lepidopteran gene duplications, and polycalins expressed in the midgut of Helicoverpa armigera: identification of proteins binding the delta-endotoxin, Cry1Ac of Bacillus thuringiensis. *Insect Biochem Mol Biol* 38:685–696.

Fig. S4. Example of a large inversion, the *RpL5a* region. The linear organization of the genes in the *Rpl5A* region in the three species is shown in Fig. S2. The nucleotide sequence of the *H. armigera* BAC 64D15, *S. frugiperda* BAC 82A05 and *B. mori* scaffold nscaf2888_4350001_4550000 were aligned by the Zpicture software. The dot plots reveal the large inversion between *B. mori* and the noctuid species. Dot plots between *H. armigera* and *S. frugiperda* (*Left*), *S. frugiperda* and *B. mori* (*Right*).

Fig. S4. (PDF)

Table S1. List of probes and BAC accession numbers

Target	*H. armigera*			*S. frugiperda*			*B. mori* scaffold no.
	Probe sequence name	Accession no. (when known)	BAC selected coordinates; accession no.	Probe sequence name	Accession no. (when known)	BAC selected coordinates; accession no.	
APN1	HaC5S01562	—	11C07; FP340421	Sf2M00699-5–1	—	41F05; FP340413	Nscaf2889_860001-1060000
APN2	HaC5S00486	—	64G08; FP340425	Sf2M04758-5–1	—	21C22; FP340406	Nscaf2889_860001-1060000
APN3	HaC5S01766	—	80K05; FP340434	—	—	—	Nscaf2889_860001-1060000
APN4	HaC5S00969	—	31C13; FP340430	—	—	—	Nscaf2889_860001-1060000
CYP9A	HaC5S00277	—	02A22; FP340423	Sf1H00841-3–1 ; Sf1H00937-5–1	—	10N12; FP340410	Nscaf2766_300001_500000
CYP332A1	HaC5S00630	—	35D18; FP340428	Sf2M00867-5–1	—	87A24; FP340417	Nscaf2888_3150001-3350000
CYP4L	HaC5S02881	—	92N13; FP340433	Sf1M04650-3–1 ; Sf1M04746-5–1	—	41I04; FP340412	Nscaf1898_13150001-13350000
CYP4M	HaC5S00687	—	33M07; FP340427	Sf2M00436-5–1	—	78G03; FP340419	Nscaf2674_4000000-4200000
idem	—	—	–	–	—	67K19; FP340411	
CYP6B	HaC5S02428	—	12E11; FP340431	Sf2M00205-5–1	—	83A13; FP340416	Nscaf2136_5100001-5300000
Ace 1	—	DQ064790	41B01; FP340437	Sf Ace1	—	70A06; FP340420	Nscaf2655_2400001-2600000
PGI	PGI Noct	—	23L11; FP340435	PGI Noct	—	49G12; FP340409	Nscaf2780_1-200000
TBP	—	L22538	94B11; FP340432	—	L22538	72F01; FP340415	Nscaf2998_900001-1100000
USP	—	AF411255	26P10; FP340422	—	AF411255	22I20; FP340407	Nscaf2847_6350001-6550000
EcR	—	AF411254	11C15; FP340426	—	AF411254	15B14 ; FP340404	Nscaf2855_5900001-6100000
Or83b	SlOR2	—	41P10; FP340424	Sl OR2	—	83A24; FP340418	Nscaf3058_4650001-4850000
JHAMT	Sl Jham	—	25P08; FP340436	Sl Jham	—	33G08; FP340408	Nscaf2993_6100001-6300000
RpL5	cx0071	—	64D15; FP340429	SF9L02022	—	82A05; FP340414	Nscaf2888_4350001-4550000
RpL10A	cx0272	—	43E14; FP340438	SF9L06387	—	60F14; FP340405	Nscaf2655_1400001-1600000

The list of probes used to identify the *H. armigera* and *S. frugiperda* BACs sequenced in this study is provided as well as the corresponding *Bombyx mori* scaffolds. In the case of the APN genes, our study revealed that these were clustered, and we obtained overlapping BACs in each noctuid species, all corresponding to a single *B. mori* scaffold. Similarly, two overlapping BACs were obtained for the CYP4M probe in *S. frugiperda*. The 15 genomic regions were found to be homologous in the three species, as evidenced by the presence of several 1:1 orthologous genes (Fig. S2). In the case of the Or83b region, the evidence for orthology of the *B. mori* region (in addition to the Or83b gene itself) was found on the same scaffold but beyond the 200-kb region chosen for analysis (Fig. S2). JHAMT, juvenile hormone acid methyl transferase.

Table S2. General features of the genomic regions compared in the three species

Feature	*H. armigera*	*S. frugiperda*	*B. mori*
Sequence analyzed, Mb	1.963	2.042	3.000
GC content, %	36.3	35.9	32.7
Exons, %	14.4	14.5	12.0
Introns, %	30.4	29.7	47.1
Intergenic, %	55.1	55.8	40.9
Number of genes*	201	274	502
Genes with a synteny link[†]	135	137	141
Average gene length, bp[†]	4,855	4,894	6,501
Average exon length, bp	239	241	208
Average intron length, bp	602	608	942
Average exon number	6.5	6.5	6.6
Average peptide length, AA	507	475	451
Subset of unique genes[§]	18	18	18
Average gene length, bp[‡]	4,054	4,333	7,199
Average exon length, bp	224	226	225
Average intron length, bp	752	804	1,336
Average exon number	6.4	6.4	6.4
Average peptide length, AA	478	482	480
Subset of genes in cluster[§]	25	28	16
Average gene length, bp[‡]	3,094	4,566	8,029
Average exon length, bp	180	180	180
Average intron length, bp	516	761	1,338
Average exon number	7.0	7.0	7.0
Average peptide length, AA	420	420	420

*Genes (including TE genes) were detected by KAIKOGAAS. Hypothetical genes were validated according to size, presence of a functional domain, a match to an EST or a known protein, or a synteny link.
[†]Gene length is meant as the distance between start and stop codon.
[‡]In one or both of the other species.
[§]Accurately annotated thanks to availability of a cDNA sequence.

Other Supporting Information

Dataset S1 (XLS)
Dataset S2 (XLS)
Dataset S3 (XLS)