Supporting information

SI Materials and Methods

BAC clone sequencing and assembly. Each individual BAC clone in the minimum tilling path physical maps for HSY and counterpart X chromosome was fully sequenced using Sanger method. Briefly, BAC DNAs were isolated from single colony culture and randomly sheared to approximately 3 kb fragments. The sheared fragments were end-repaired, size-selected on an agarose gel, purified, and ligated to the pSMART vector (Lucigen, WI 53562 USA). The plasmid DNA of randomly picked electro-transformed clones with 8 to 20 fold coverage of the BAC insert fragment was isolated and sequenced using ABI BigDye Terminator version 3.1on a 3730XL DNA analyzer (Applied Biosystems, http://www.appliedbiosystems.com/).

The Sanger reads were filtered and screened to remove *E.coli* genomic sequences, vector sequences, and low quality or ambiguous sequences. The trimmed reads were sequentially assembled into contigs using Phred/Phrap/Consed (1, 2), CAP3 (3) packages, and Sequencher software (Gene Codes Corporation, Ann Arbor, MI). Individual BAC assemblies were examined manually for sequence regions with signs of miss-assemble and also checked by long range PCR targeting 10kb amplicons. Suspected low quality regions and gaps in the assembly were resolved by removing ambiguous reads, and/or by primer walking, re-sequencing the PCR products, sequencing additional shotgun sub-clones, and/or alignment with neighboring BAC sequences. A BAC was not considered complete until all inconsistent read pairs had been resolved and Consed reported an error rate of less than 1 per 10,000 bases. The GenBank accession numbers for all the sequenced BACs are listed in (Table S14). Based on the physical maps of the HSY and its corresponding X region, the individual BAC sequences were joined into pseudomolecules.

Identification of X and HSY shared collinear blocks. We used Mauve, a genome alignment program, to identify co-linear blocks between the HSY and X progressively from a local co-linear block score (LCBS) of 1K up to 171K. Using a lower LCBS, numerous blocks with rearrangements were detected. By increasing the LCBS, the number of co-linear blocks was decreased and only larger blocks were shown in Fig. S6, and the blocks having lower LCBS were show as solid yellow. The first large Block (Block I) was detected using LCBS 12522. The second large Block was detected using LCBS 78927.

Repeat and CpG islands analysis. Known repeats were identified in the HSY and X pseudomolecule sequences using RepeatMasker (http://www.repeatmasker.org) with a custom repeat library generated by combining Repbase (4), TIGR plant repeats (ftp://ftp.tigr.org/pub/data/ TIGR_Plant_Repeats), and known papaya repeats (5). To identify sex-specific repeats, RepeatScout-v1 (6) was applied to find repeats in both HSY and X sequences. Non-redundant repeats longer than 100bp were compared with known papaya repeats (5) using CD-HIT software (7) with a cut-off of 70% similarity to identify new repeats. The criteria for new repeats were: 1) at least 50% of the consensus sequence length aligned, 2) occurrence at least 10 times in the HSY or X sequences, 3) greater than 75% identity in the alignment with the HSY/X sequence, 4) sex-specific repeats were defined as those with fewer than 10 hits in the complete papaya genome sequence. The tandem repeats were analysed using the Tandem Repeats Finder software (8).

CpG islands were identified in repeat-masked sequences using the Emboss toolbox "newcpgreport" software with the parameters set as follows: window size: 100 bp, minimum GC content: >50%, minimum Obs/Exp ratio: >0.6, and minimum CpG island length: 200. The schematics were drawn using DomainDraw (9).

Annotation of transcription units. The repeat-masked HSY and X sequences were blasted to the papaya EST and gene model databases and also blasted to EST datasets of *Medicago truncatula*, *Oryza*

sativa, *Populus trichocarpa*, *Vitis vinifera*, and *Arabidopsis thaliana*, using tblastx for transcription unit identification. The loci where EST or gene models specifically aligned to HSY sequences with greater than 94% sequence identity were considered HSY-specific transcription units, loci where EST or gene models specifically aligned to X sequences with greater than 98% sequence identity were considered X-specific transcription units, and sequences with greater than 98% identity with X sequences and greater than 94% with HSY sequences were classified as XYtranscription unit pairs. Each predicted transcript was translated in all six reading frames to distinguish protein-coding genes from pseudogenes. Potential functions of protein-coding transcripts were predicted using conserved domains and gene functions of homologs in other plants. Two *de novo* gene prediction programs, Genscan, and Fgenesh, were also used to predict additional genes that may have been missed using the above approaches. Additional genes predicted by all three programs were considered possible genes.

The computationally annotated transcripts were tested by RT-PCR. cDNAs from different developmental stages of flower and leaf were amplified with specific primers designed across (at least one) intron whenever possible. The amplicons were sequenced from at least one tissue of all three sexes were sequenced. The sequences obtained were blasted against the HSY and X pseudomolecules to validate the target amplification and to determine the chromosomal origin of each sequence. Most genes were tested with multiple primer pairs (to test the gene structures predicted by the different prediction program) and the final gene structure was annotated based on the RT product sequences. Most of the genes present in the X chromosome were already annotated in papaya gene models using programs trained for papaya specific transcripts. Transcription units predicted from the papaya gene modal or the papaya EST database were considered to be validated if the expected sized product was obtained from specific primers. RNAseq data obtained from shoot apical meristem cDNA was used to further validate these transcription units.

Additional notes

G+C content of the HSY and X. The G+C contents of the HSY, its X counterpart, and of the entire papaya genome were similar in the unmasked sequences (36.2%. 35.7%, and 35.3%), or the masked ones (35.7%, 34.2% and 33.5%) (Table S15, Fig. S3). CpG islands in repeat-masked sequences are predominantly associated with expressed genes. The density of CpG islands in the HSY and, to a lesser extent, the X counterpart is lower than in the rest of the genome (Table S15 and Fig. S11), probably due to accumulation of repetitive sequences in the regions (Fig. S12).

Identification of Border B. The HSY border B was identified on both the HSY and X physical maps. BACs SH85C03 and SH86B15, which are common to both chromosomes, are immediately adjacent to the genetically defined border (Fig. S2*A*; 10). In BAC SH85C03, a 90 kb sequence overlaps and is identical with X BAC SH30B21, showing that SH85C03 (and also SH86B15) must represent X chromosome sequence; consistent with this, there is a region of 134 kb of overlapping sequence between BAC SH85C03 and the HSY BAC SH58C24, but the sequences differ by 2.1%. Genome-wide BAC end sequence searches revealed a strong candidate for the HSY region corresponding to two of these X BACs, SH85C03 and SH86B15. A BAC clone, SH60M19, with insert size 252 kb, includes an identical overlap of 20.5 kb with HSY BAC 58C24 (Fig. S2*A*). This BAC could be aligned with part of the X BAC SH85C03 and the entire X BAC SH86B15, and it extends 7,110 bp beyond the X BAC SH86B15. Comparison of these two BACs yielded 7 matched pieces, with one uninterrupted piece of 130 kb at the end of SH86B15 furthest from the genetically defined border B. Across the 182 kb region of the X BAC SH86B15 that aligns with HSY sequence, the numbers of single nucleotide polymorphisms (SNPs) and indels decline over the first 80 kb from the border, and the remaining 102 kb included only 6 SNPs and 6 single base indels, close to the sequencing error of 1 per 10^5 nucleotides (Fig. S2*B*). The HSY border B sequence, where homology with the X region becomes high, lies about 277 kb beyond the genetically defined border (Fig. S2*A*).

Computing numbers of genes present in the ancestral chromosome, and estimating the numbers lost from the HSY and the X. It is important to differentiate between (i) transcriptional units present as functional copies on the X and completely missing from the HSY (or vice versa) and (ii) those with a copy recognizable on the HSY, but clearly non-functional (and vice versa). To show our reasoning, consider the example of inferring losses from the HSY, both of these categories suggest that a gene was present in the ancestral chromosome, but the HSY copy was lost (Category i) or pseudogenized (Category ii). Category (i) includes 24 genes (excluding the 4 sequence pairs that are pseudogenes on both the X and the Y); none of these 24 appears to have been moved to the X from an autosomal origin, and so their absence from the HSY implies loss from the HSY. Category (ii) includes the 10 HSY pseudogenes with intact and putatively functional X copies.

Without an outgroup, we cannot, of course, definitively infer losses, or distinguish between losses from the HSY and gains of sequences by the X, or vice versa. However, it seems reasonable to classify the 24 genes in Category (i) genes as losses from the Y, rather than gains by the X, based on the following reasoning. If the genes had been duplicated onto the X (as has been inferred in other sex chromosome systems), we should see the original copy in the autosomal sequences, but we found no traces of these genes. Of course, this does not take into account possible gains of genes by the X where the autosomal source gene is no longer present; however, this seems unlikely in a young sex chromosome system such as papaya (age estimates are described in detail in the manuscript, see also COMMENT 2 below). We thus obtain a total of 24+10=34 genes present in the ancestral chromosome, but lost or pseudogenized in the HSY.

When considering possible losses from the X, a total of 16 HSY genes have no X counterparts, of which 7 may be duplicates of autosomal genes, and 9 may therefore represent losses from the X. The counterpart to Category (ii) includes the 6 X pseudogenes with intact and putatively functional HSY copies. Overall, these two categories give us two alternative total numbers of genes present in the ancestral chromosome, but lost or pseudogenized in the X: either 6+16=22, or 6+9=15.

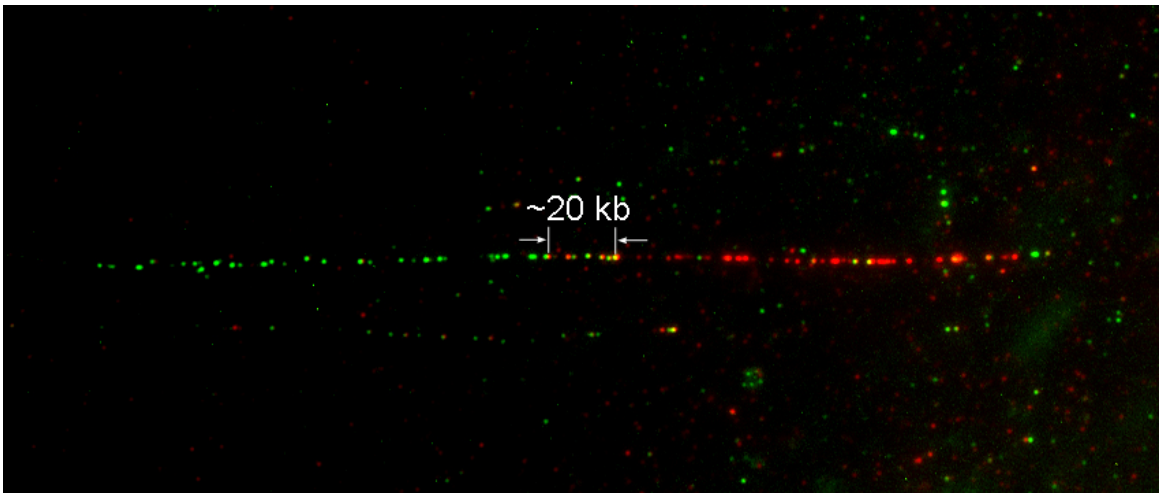
References

1. Ewing B, Hillier L, Wendl M, Green P (1998) Basecalling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* 8:175-185.
2. Gordon D, Abajian C, Green P (1998) Consed: a graphical tool for sequence finishing. *Genome Res* 8:195-202.
3. Huang X, Madan A (1999) Cap3: A DNA sequence assembly program. *Genome Res* 9:868-877.
4. Jurka J, *et al.* (2005) Repbase update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res* 110:462-467.
5. Nagarajan N, *et al.* (2008) Genome-wide analysis of repetitive elements in papaya. *Trop Plant Biol* 1:191-201.
6. Price AL, Jones NC, Pevzner PA (2005) *De novo* identification of repeat families in large genomes. Paper presented at the proceedings of the 13 Annual International conference on Intelligent Systems for Molecular Biology (ISMB-05).
7. Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22:1658-1659.
8. Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* 27:573-580.
9. Fink JL, Hamilton N (2007) DomainDraw: a macromolecular schematic drawing program. *In Silico Biol* 7:14.
10. Na J-K, et al. (2012) Constriction of physical maps for the sex-specific regions of papaya sex chromosomes. *BMC Genomics* 13:176.
11. Zhang W, Wang X, Yu Q, Ming R, Jiang J (2008) DNA methylation and

heterochromatinization in the male-specific region of the primitive Y chromosome of papaya. *Genome Res* 18:1938-1943.
12. Ming R, et al. (2008) The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature* 452:991-996.

Supplemental Figures

A



B

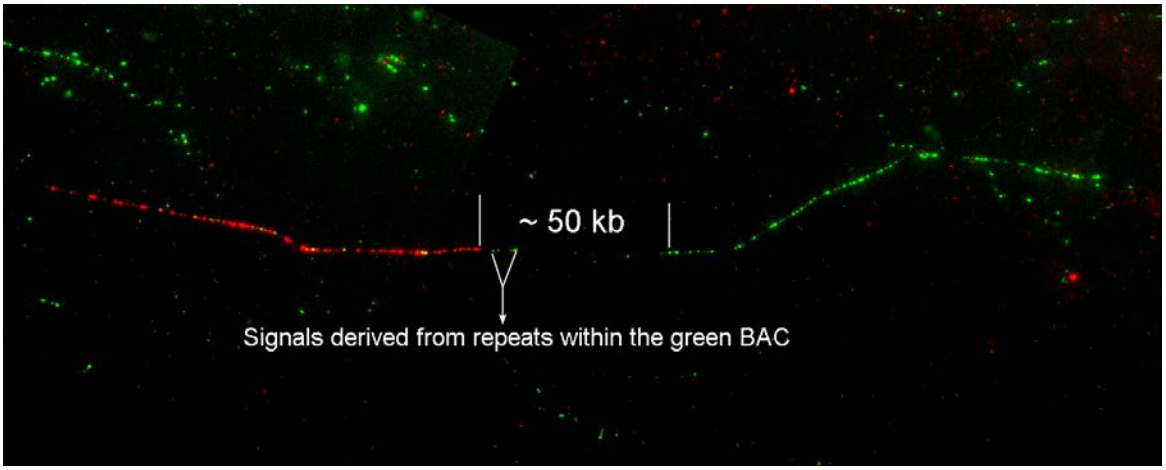


Fig. S1. Confirmation of overlapping BACs and estimation of gap sizes for physical mapping of the HSY and its X counterpart using fiber FISH. (*A*) Fiber-FISH image of two overlapping BAC clones in the HSY of papaya. HSY BAC clone 41F24 (130kb) was labeled as red and HSY BAC 57B23 (155kb) as green. These two BAC clones overlapped approximately 20 kb. (*B*) HSY BAC clone 76M08 (175kb) was labeled as red and HSY BAC 89I15 (180kb) as green, these two BACs are separated by approximately 50 kb. The short green signal within the gap (close to the red signal) may be derived from a repetitive DNA element that is on BAC 89I15.

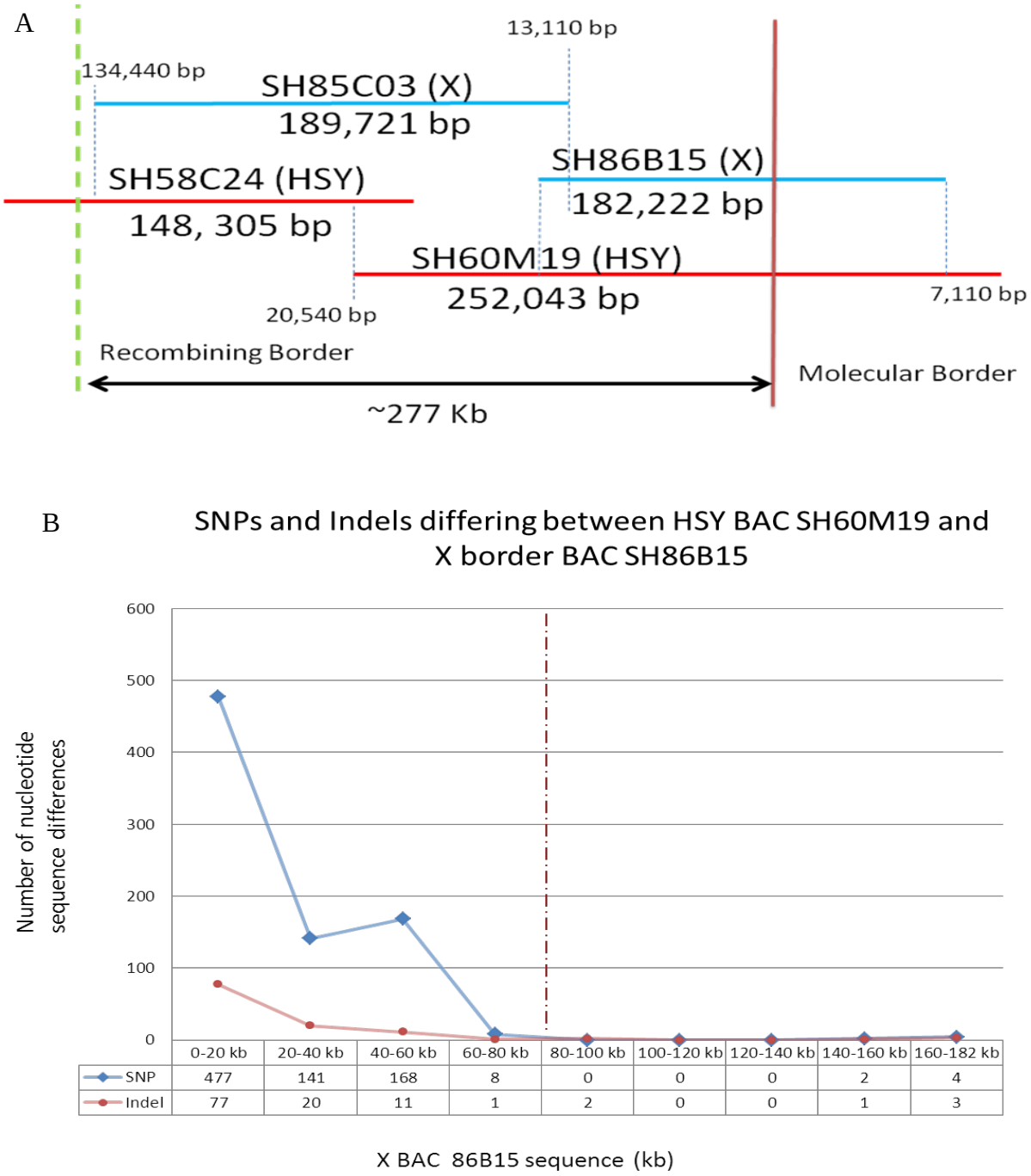
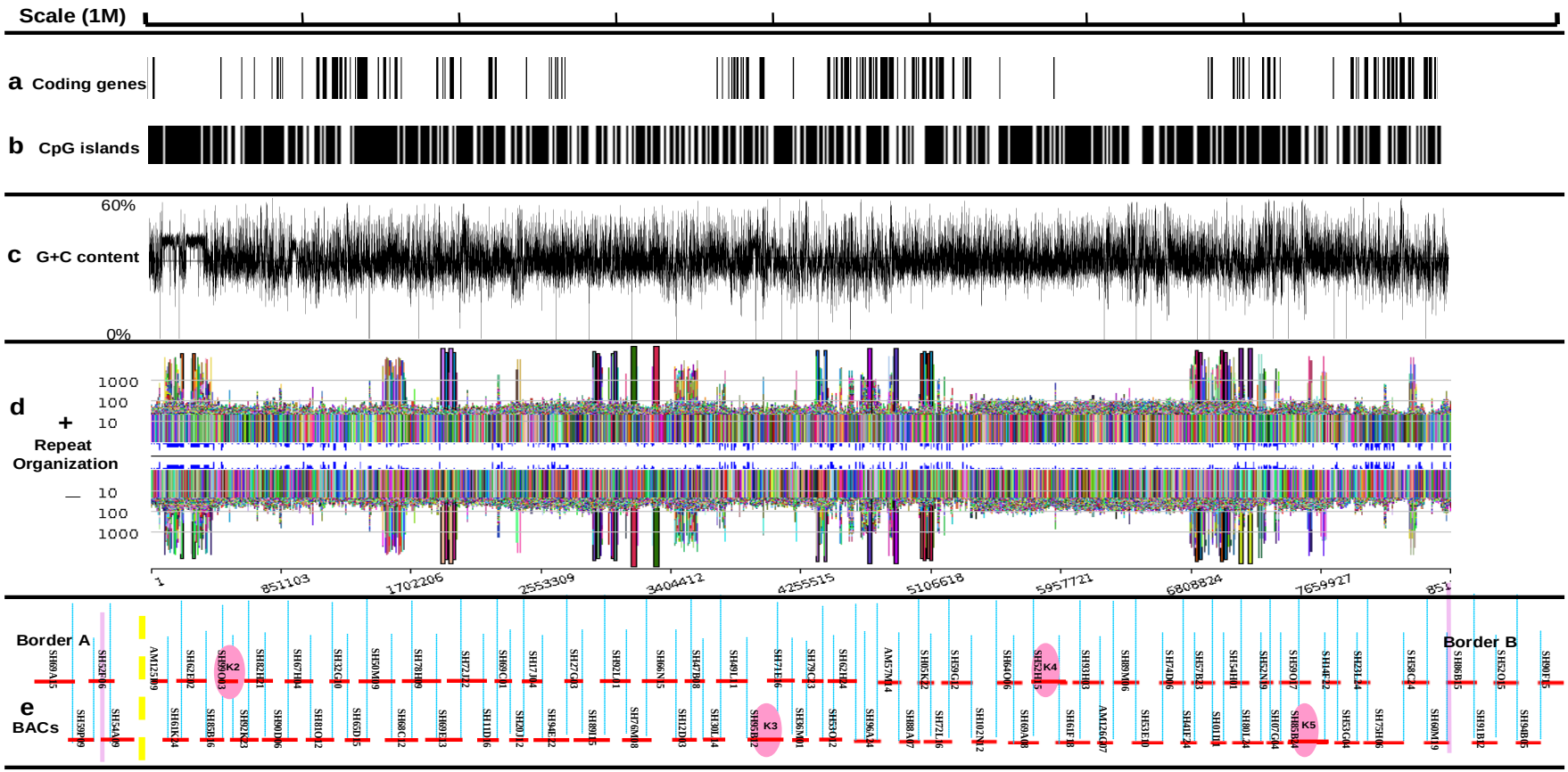


Fig. S2. Resolution of the X/HSY sequences in the Border B region. *(A)* The physical map of Border B. The dashed green line at the left represents the recombination border defined by fine mapping, and the solid maroon line at the right is the border defined by the number of sequence differences (see part B of the figure). The X sequences are in blue and the HSY sequences in red. *(B)* Molecularly defined border B. Numbers of sequence differences (SNPs and indels) between HSY BAC SH60M19 and X border BAC SH86B15. The dashed maroon line represents the molecularly defined border. The data points show number of SNPs and indels per 20 kb using X BAC 86B15 sequence as a reference for the X-axis.

HSY



BAC physical map legends:

- BAC clone
- Border
- Overlap of BAC clones
- Line end points MSY and X corresponding region
- Gap between BAC contigs
- Knob position
- Corresponding position to knobs at MSY

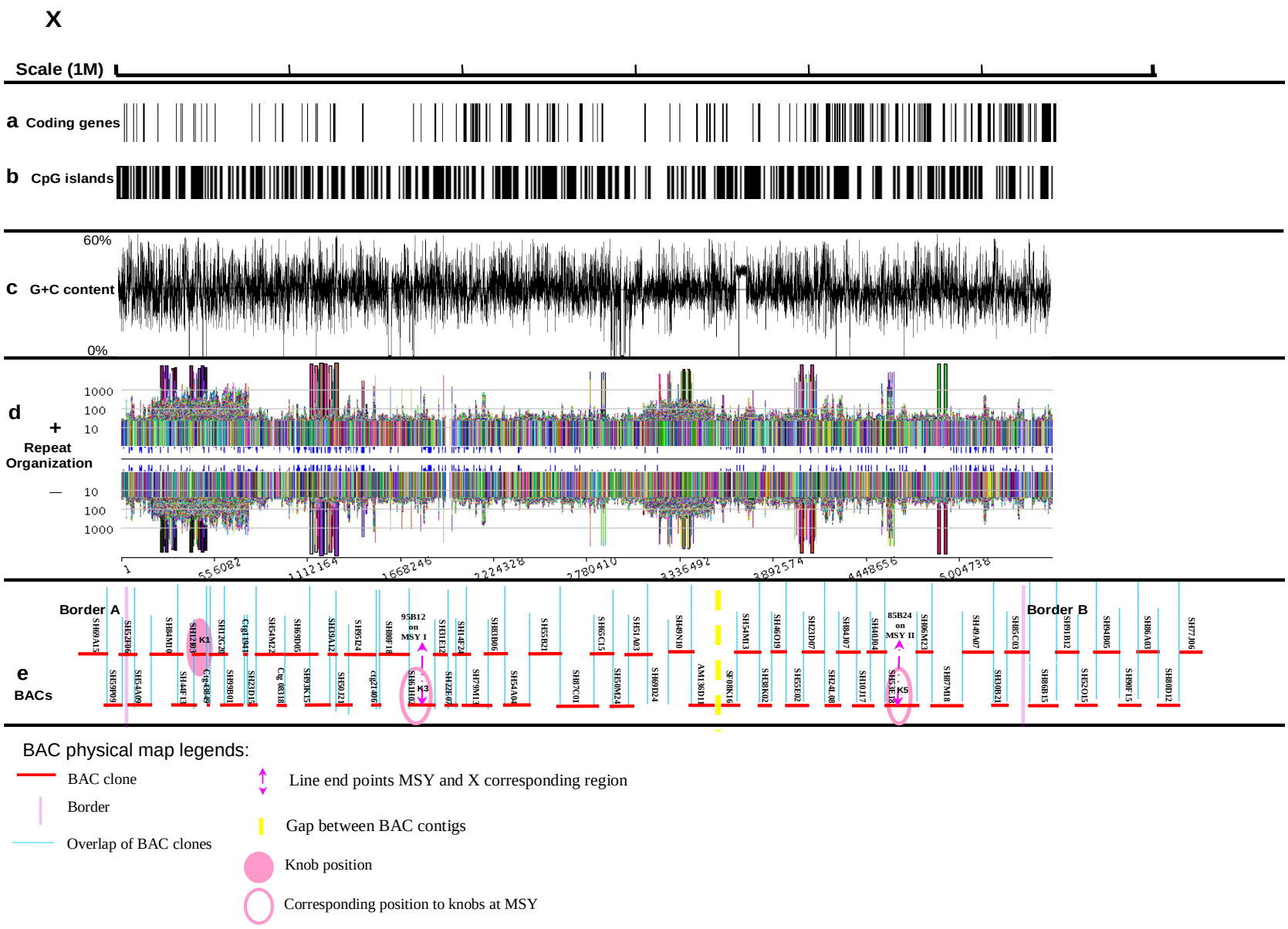

X
Scale (1M)
a Coding genes
b CpG islands
c G+C content
60%
0%
d
+
Repeat
Organization
—
1000
100
10
10
100
1000
1
5560082
1112164
1668246
2224328
2780410
3336492
3892574
4448656
5004738
e
BACs
Border A
Border B
95B12
on
MSY I
85B24
on
MSY II
K1
K3
K5
BAC physical map legends:
BAC clone
Border
Overlap of BAC clones
Line end points MSY and X corresponding region
Gap between BAC contigs
Knob position
Corresponding position to knobs at MSY

Fig. S3. Annotated sequences of the HSY and X on sex chromosome in papaya. Chart of the HSY and X chromosome sequences showing protein-coding genes, CpG islands, G+C content, organization of repeated structures, and sequenced clones. All sequence features and BACs are drawn to scale. (A) The positions of all predicted protein-coding genes are shown in black. (B) CpG islands of the papaya sex chromosome are shown in black. CpG islands, predicted by EMBOSS CpGPlot are shown in black. (C) G+C content is plotted by the ARTEMIS sequence viewer and annotation tool. (D) The organization of repeated structures in the HSY and X counterpart region were *de novo* detected and generated by the Pygram pipeline. Extensive repeats were present on the HSY and X, with the blocks containing the most frequent repeats, are indicated by the small blue boxes proportional in size to the frequency located between the middle black line and the plus(+) and minus (-)strand views at each occurrence of the repeat. The x-axis corresponds to the sequence strand coordinate position (bp). The y-axis, corresponding to repeat size and scale, is logarithmic. Each repeat has its own specific color. All occurrences of the same repeat have the same color on both strands. (E) BACs from the finished sequence of the papaya HSY and X chromosome. One physical gap at HSY border A and one gap on X in the clone map are indicated.

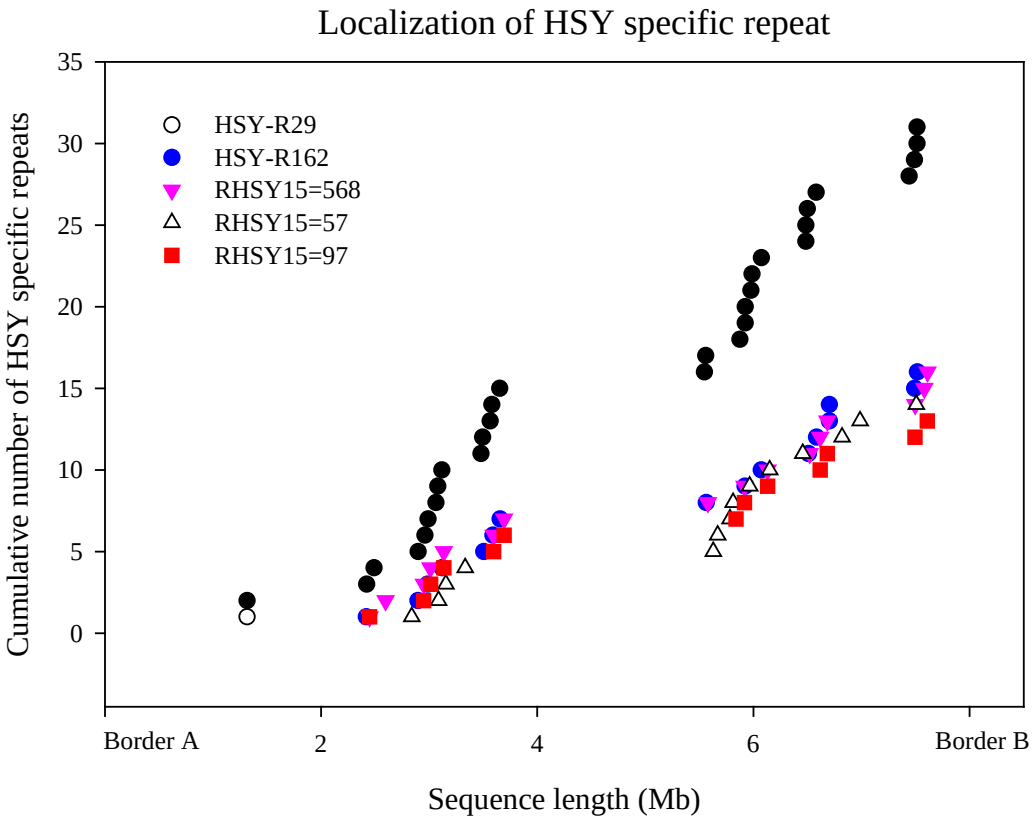


Fig. S4. Distribution of HSY-specific repeats across the HSY pseudomolecule. Each symbol denotes a different HSY-specific repeat (see symbol key).

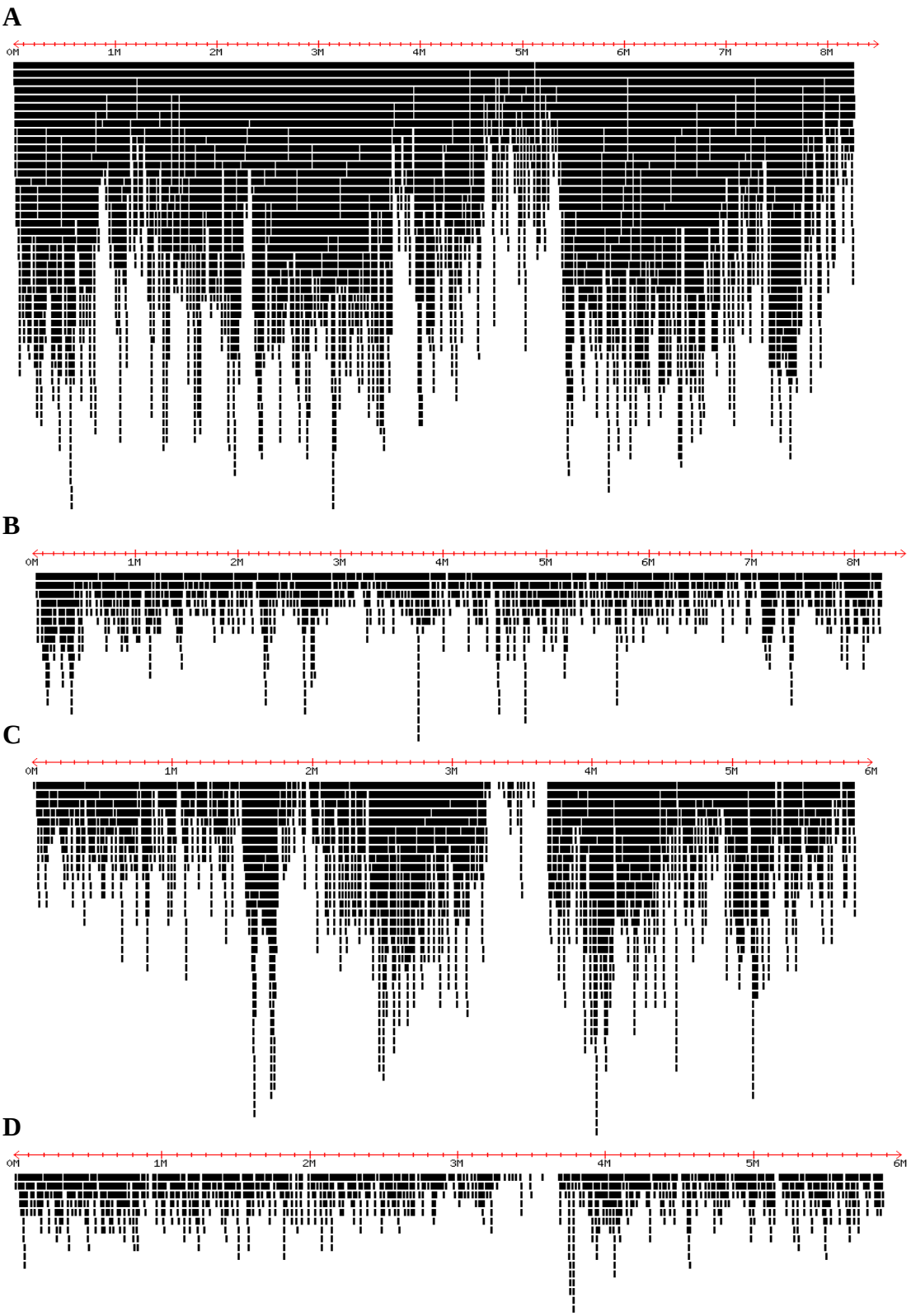


Fig. S5. Distribution of transposable elements and tandem repeats on the HSY and X concatenated pseudomolecules. (*A*) Transposable elements on HSY, (*B*) Tandem repeats on the HSY, (*C*) Transposable elements on the X, and (*D*) tandem repeats on X.

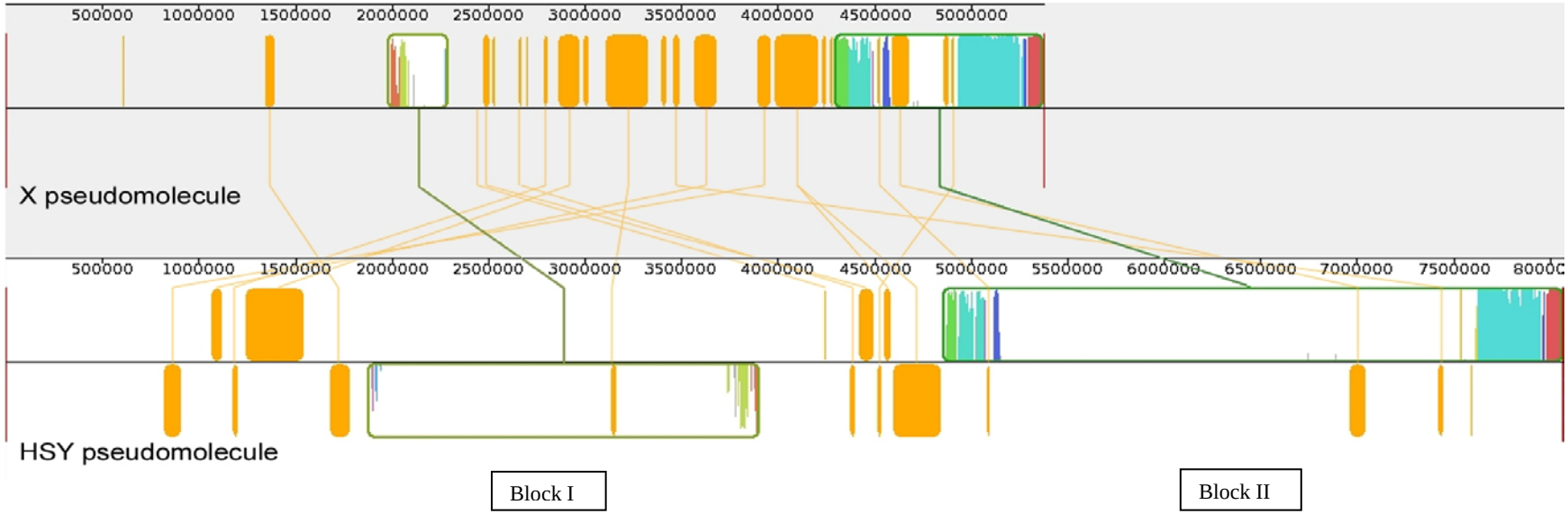


Fig. S6. Expansion of the HSY in comparison with the X counterpart. Mauve, a genome alignment program, performed a series of co-linear block identification comparisons progressively using local co-linear block scores (LCBS) of 1K up to 171K. The homologous segments are shown as colored bars. Block 1 was detected using a LCBS of 12522 and Block 2 was detected using an LCBS of 78927. The average sequence identity in the alignment is proportional to the height of the colored region. The orange blocks are homologous regions between the X and HSY with lower LCB scores. The local collinear blocks are connected across the aligned X and HSY sequences. Blocks shifted downward represent segments that are inverted relative to the reference sequences.

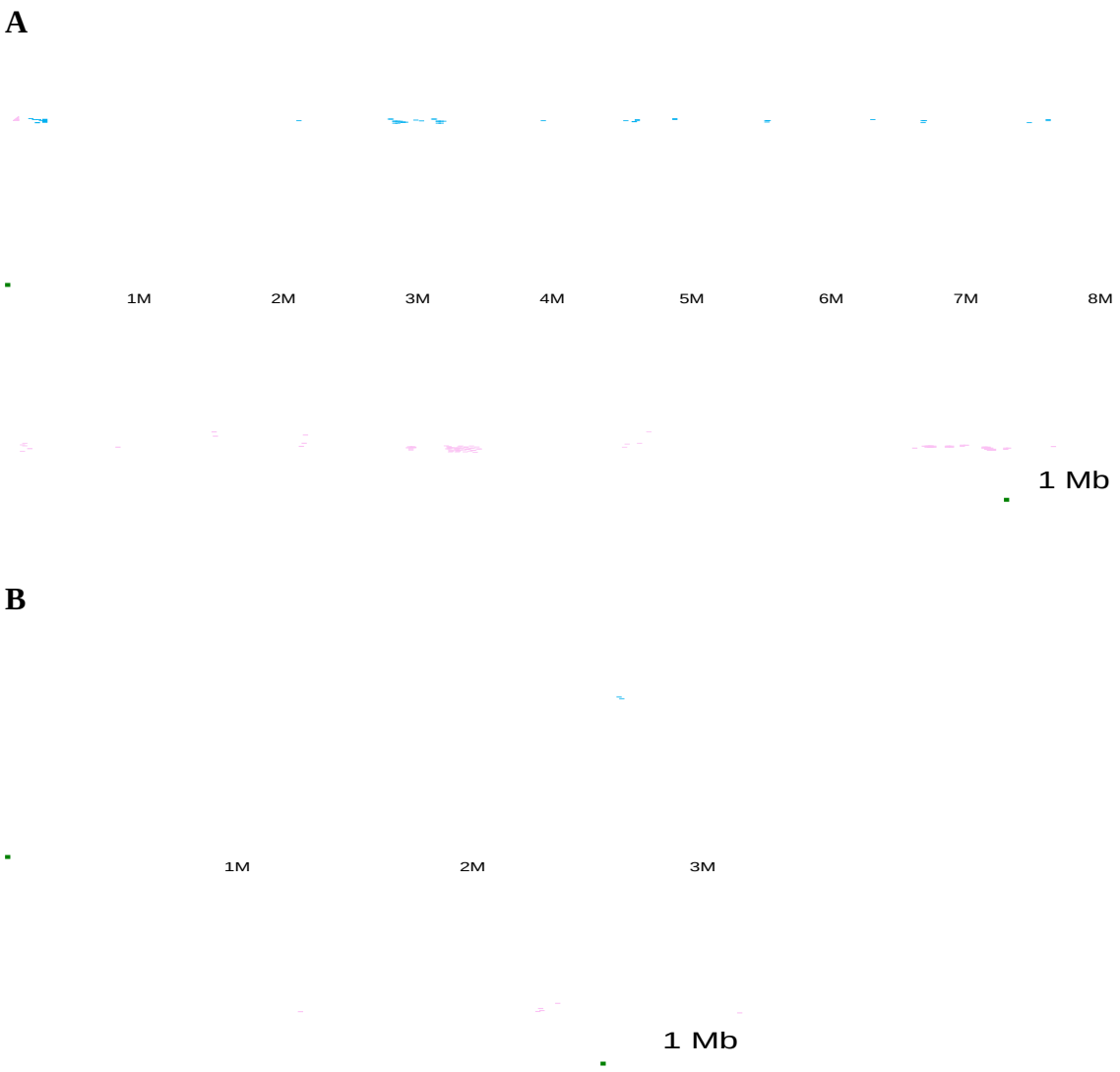


Fig. S7. The intra-chromosomal duplicates in HSY (*A*) and X (*B*) concatenated pseudomolecules. Green bar indicates the chromosome. Above the chromosome are direct duplications in pink bridge lines. Below the chromosome are the inverted duplications in blue bridge lines.

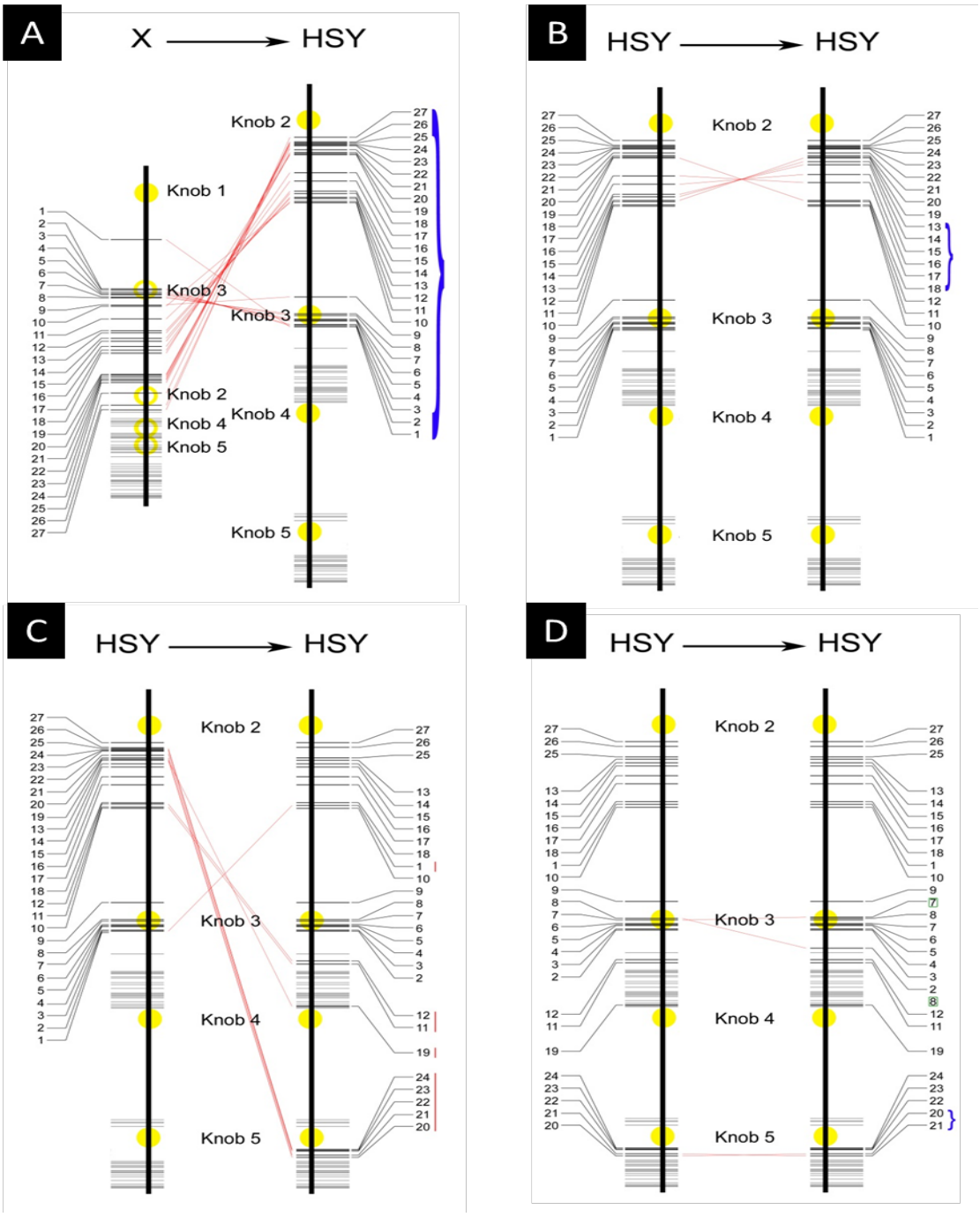


Fig. S8. Proposed steps of chromosomal evolution in inversion 1 between HSY and X based on the 70 homologous transcript-encoding sequence. The vertical black lines indicate the HSY and X sequences; the horizontal lines indicate the 70 homologous sequences, numbered according to their order on the X. The solid yellow circles indicate the heterochromatic knobs (11). Knob 1 is shared between the HSY and X. Knobs 2 - 5 are HSY-specific. The empty yellow circles indicate the estimated positions of knobs in the X chromosome corresponding to each of the numbered HSY knobs. (*A*) Inversion 1 – Genes 1-27 are inverted on the HSY (*B*) Inversion 1 – Genes 13-18 are inverted on the HSY (*C*) Inversion 1 – Genes 1, 12-11, 19, and 24-20 are translocated on the HSY (*D*) Inversion 1 – Genes 7 and 8 are duplicated on the HSY, Genes 20-21 are inverted on the HSY.

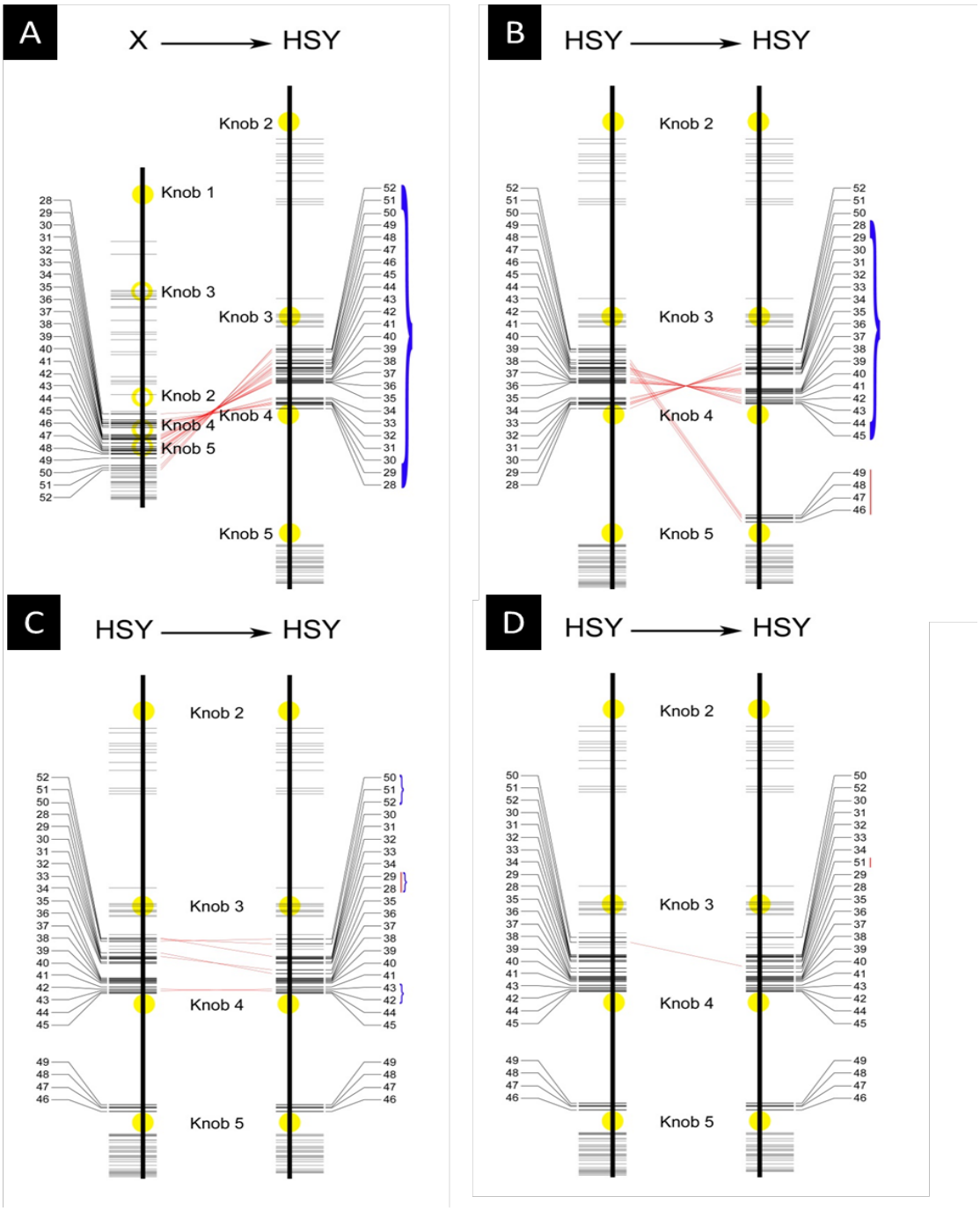
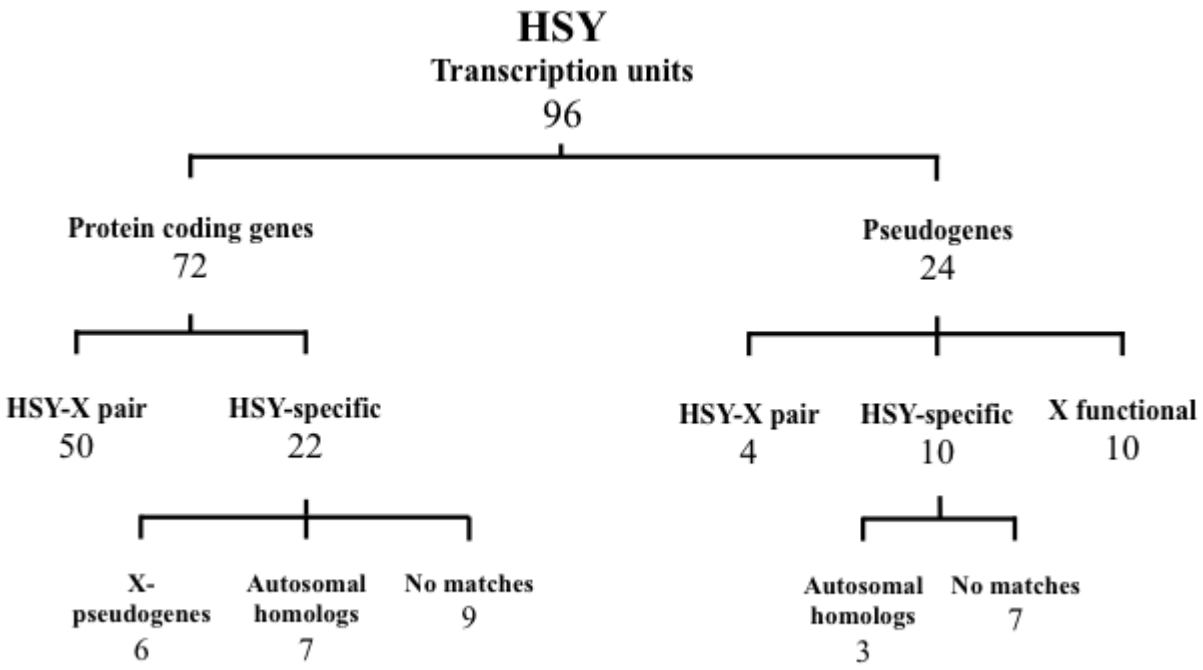


Fig. S9. Proposed steps of chromosomal evolution in inversion 2 between HSY and X based on the 70 homologous transcripts. The vertical black lines indicate the HSY and X sequences; the horizontal lines indicate the 70 homolog transcripts. The 70 homolog transcripts on X are numerically ordered. The solid yellow circles indicate the heterochromatic knobs (11). Knob 1 is shared between HSY and X. Knobs 2 - 5 are HSY specific. The empty yellow circles indicate the estimated corresponding positions of knobs in the X chromosome. (*A*) Inversion 2 – Genes 28-52 are inverted on the HSY (*B*) Inversion 2 – Genes 28-45 are inverted on the HSY and genes 49-46 are translocated on the HSY (*C*) Inversion 2 – Genes 50-52 and 43-42 are inverted on the HSY. Genes 29-28 are inverted and translocated on the HSY (*D*) Inversion 2 – Gene 51 is translocated on the HSY.

A.



B.

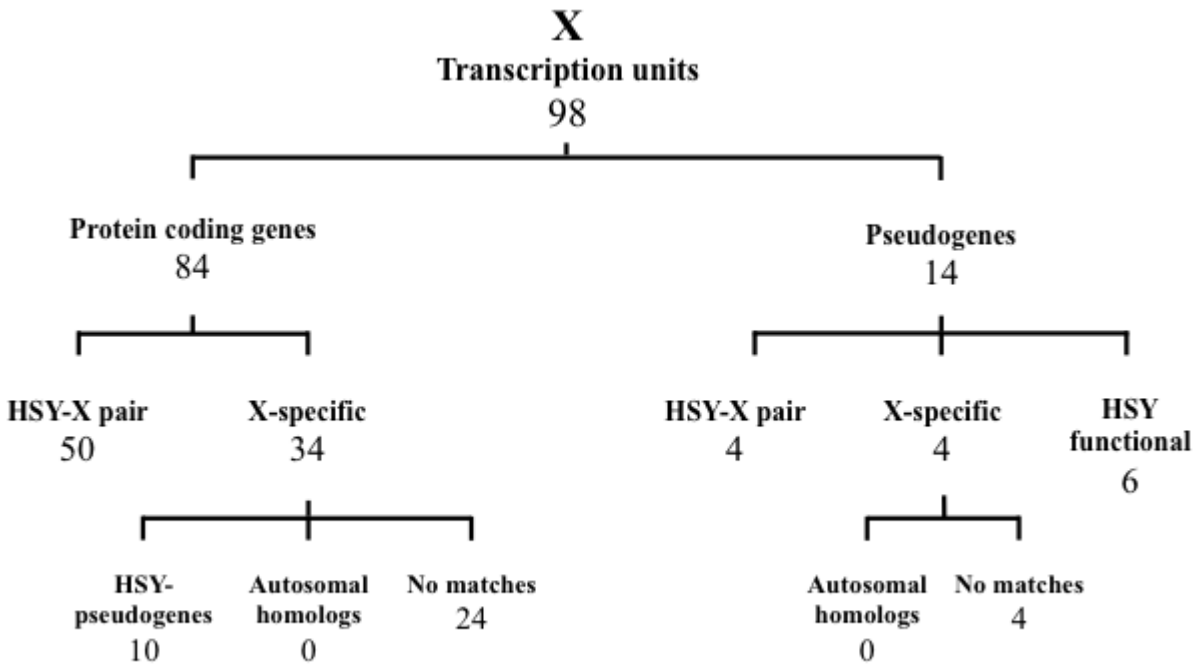


Fig. S10. Distribution of transcript-encoding sequences in the HSY and X. *(A)* The distribution of transcription units on the HSY. The transcription unit total for the HSY includes the protein

coding genes, pseudogenes, and the missing genes (X-specific genes or pseudogenes not found on the HSY or autosome). *(B)* The distribution of transcription units on the X-specific region. The transcription unit total for the X includes the protein coding genes, pseudogenes, and the missing genes (HSY-specific genes or pseudogenes not found on the X or autosome). The X transcription unit total does not include the HSY-specific genes or pseudogenes that have an autosomal homolog.

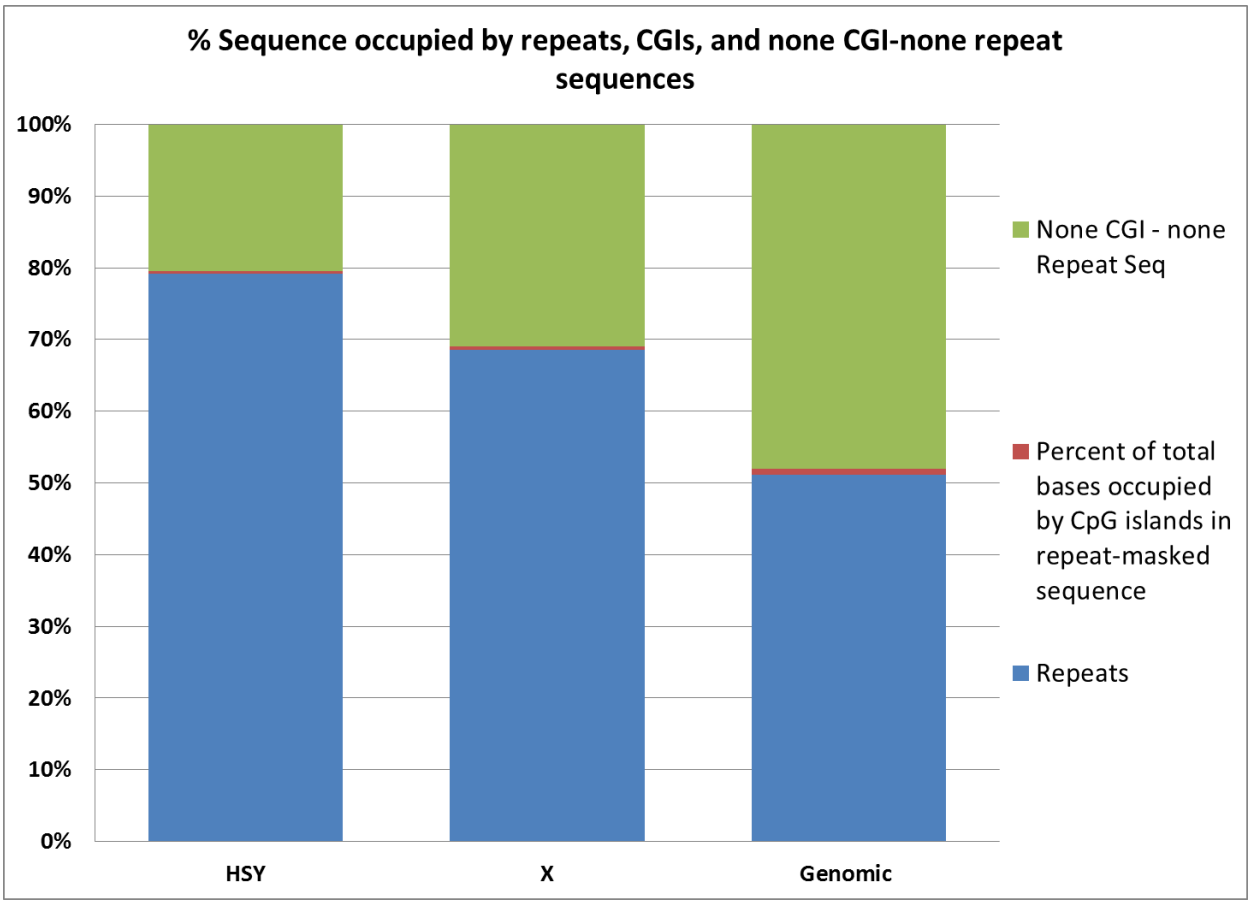


Fig. S11. Abundance of Papaya specific repeats and CpG clusters in non repeat region of HSY, X and genome of papaya. HSY and its X counterpart accumulate proportionally less CpG islands (CGIs) sequences compared to the rest of the genome

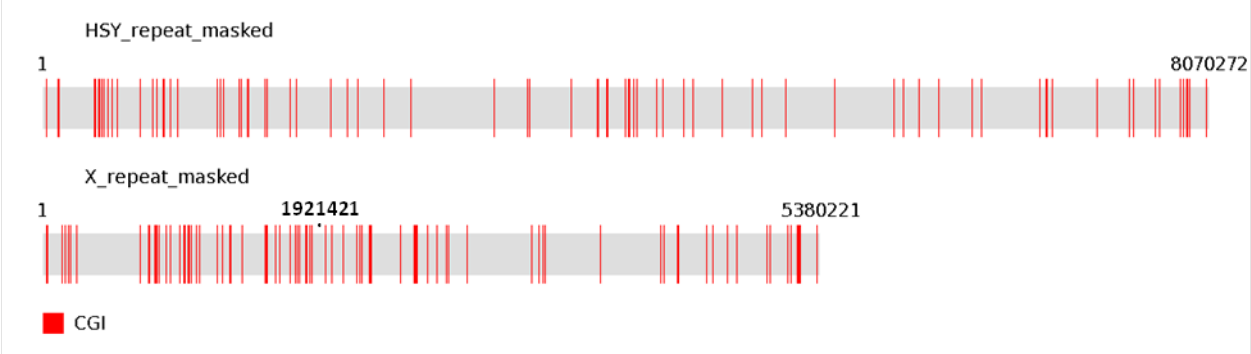


Fig. S12. Distribution of CpG islands on the HSY and X pseudomolecules (including the 1.9-Mb Knob 1 region).

Supplemental Tables

Table S1. Interspersed repeats in the papaya sex chromosomal sequences.

Sequence source	HSY (8,070,272bp)			X (3,415,880bp)			LG1 excluding X specific and knob1 region (19,057,724 bp)		
Repeat class/family	**No. of elements**	**Length occupied (bp)**	**Percentage of sequences (%)**	**No. of elements**	**Length occupied (bp)**	**Percentage of sequences (%)**	**No. of elements**	**Length occupied (bp)**	**Percentage of sequences (%)**
Retroelements:	3962	3,491,027	43.3	1,470	1,544,912	44.7	2,598	2,504,893	18.7
LINEs	54	41330	0.5	55	37074	1.1	203	118,124	0.9
LTR elements:	3908	3,449,697	42.8	1,415	1,507,838	43.6	2,395	2,386,769	17.8
Ty1/Copia	349	358,804	4.5	284	273,411	7.9		403,142	3.0
Ty3/Gypsy	2541	2,404,485	29.8	835	950,748	27.5	1,478	1,674,755	12.5
DNA transposons:	9	6,732	0.1	10	1,357	0	68	8,940	0.1
En-Spm	1	81	0	0	0	0	28	2,061	0.0
MuDR-IS905				1	93	0	4	353	0.0
Unclassified:	5,504	2,902,233	36	1,601	776,581	22.5	2,288	1,004,992	7.5
Total interspersed repeats:	9,475	6,399,992	79.3	3,081	2,322,850	67.2	4,954	3,518,880	26.2

Table S2. Abundances of sex-specific repeats.

Sequence source	HSY	Corresponding X	Corresponding X and knob1 region
Numbers of repeat elements	1,944	376	486
Length occupied (bp)	866,694	152,346	188,550
Percentage of the sequence (%)	10.7	4.4	3.5
Total sequence length (bp)	**8,070,272**	**3,415,880**	**5,850,206**

Table S3. Numbers of tandem repeats in the HSY and its X counterpart.

HSY	Block number	Copy number	Variants	Total length/HSY %
Microsatellites (1-6 bp)				
Mononucleotide	20	777	2	777
Dinucleotide	82	3891	7	7782
Trinucleotide	60	1018.1	15	3054.3
Tetranucleotide	9	119.5	9	478
Pentanucleotide	84	865.2	14	4326
Hexanucleotide	22	550.8	16	3304.8
Total	**277**	**7221.6**	**63**	**19722.1/0.25%**
Minisatellites (7-100 bp)				
7-30 bp	1135	3256.2	779	57230.6
31-50 bp	151	390.4	105	15349.2
51-70 bp	69	201.7	56	12042.3
71-99 bp	34	83.6	26	6904.2
Total	**1389**	**3931.9**	**966**	**91526.3/1.14%**
Satellites (>100 bp)				
100-200 bp	55	114.8	46	15808.8
201-300 bp	24	51.3	20	12964.3
301-400 bp	99	403	27	131416.6
>400 bp	20	73.3	19	33862.2
Total	**198**	**642.4**	**112**	**194051.9/2.43%**
Grand Total	**1864**	**11795.9**	**1141**	**305300.2/3.82%**

X counterpart	Block number	Copy number	Variants	Total length/X %
Microsatellites (1-6 bp)				
Mononucleotide	38	1155	3	1155
Dinucleotide	134	2694	9	5388
Trinucleotide	45	737.8	14	2213.4
Tetranucleotide	8	85.8	7	343.2
Pentanucleotide	47	441.4	15	2207
Hexanucleotide	36	309.3	26	1855.8
Total	**308**	**5423.3**	**74**	**13162.4/0.26%**
Minisatellites (7-100 bp)				
7-30 bp	761	2486.3	531	39493.3
31-50 bp	110	295.3	79	11508.4
51-70 bp	31	71.9	27	4236.8
71-99 bp	45	96.4	72	8125.4
Total	**947**	**2949.9**	**709**	**63363.9/1.27%**
Satellites (>100 bp)				
100-200 bp	71	210.6	55	26236.4
201-300 bp	8	16.5	8	3935
301-400 bp	33	147.7	13	48621.8
>400 bp	1	2	1	962
Total	**113**	**376.8**	**77**	**79755.2/1,6%**
Grand Total	**1368**	**8750**	**860**	**156281.5/3.13%**

Table S4. Transposable elements in the two HSY expansion blocks.

Transposable elements	Block1 (1,370,471 bp)			Block2 (1,921,971 bp)		
	# of repeat elements	Length occupied (bp)	Percentage (%)	# of repeat elements	Length occupied (bp)	Percentage (%)
Retroelements	1,185	883,176	64.4	1,736	1,363,277	70.9
LTR	1,185	883,176	64.4	1,736	1,363,277	70.9
Ty1/Copia	51	29,601	2.2	39	9806	0.5
Gypsy/DIRS1	836	626,311	45.7	1,284	996,331	51.8
DNA transposons	4	3243	0.2	0	0	0
Unclassified	774	273,308	19.9	948	312,015	16.2
Total interspersed repeats	1,963	1,159,727	84.6	2,684	1,675,292	87.2

Table S5. List of XY^h gene and pseudogene pairs in shown in Figures 1, S8 and S9.

Figure 1 & S8-S9 ID	Gene Symbol	Putative Function	Number of Exons		Coding Sequence Length		Intron length		Autosome Homolog		
			X	HSY	X	HSY	X	HSY	Linkage Group	Supercontig	% identity
1	PXCpXY^{h}15◊	---NA---	1	1	965	965	n/a	n/a			
2	CpXY^{h}1	Regulator of Vps4 activity in the MVB pathway	6	6	2280	2265	2890	n/a			
3	CpXY^{h}2	Somatic embryogenesis receptor kinase	11	11	1884	1884	3734	3702			
4	CpXY^{h}3	Exocyst complex subunit SEC6	25	25	2268	2268	40497	7483			
5	PY^{h}CpXY^{h}2	---NA---	1	1	369	380	n/a	n/a			
6	PY^{h}CpXY^{h}3	---NA---	1	2	600	599	n/a	5049			
7	**CpXY^{h}4	Late embryogenesis abundant protein 18	1	1	293	293	n/a	n/a			
8	**PY^{h}CpXY^{h}5	---NA---	1	1	261	260	n/a	n/a			
				1		260	n/a	n/a			
9	PY^{h}CpXY^{h}7	---NA---	7	6	813	579	2034	700			
10	CpXY^{h}5	4-nitrophenylphosphatase	8	8	807	816	2444	2448			
11	CpXY^{h}6	FRA3 inositol or phosphatidylinositol phosphatase	10	10	3336	3336	40354	65434			
12	CpXY^{h}7	Leucine zipper-ef-hand (Ca binding motif)	14	14	2271	2286	10175	19777			
13	CpXY^{h}8	SYN4 (SISTER CHROMATID COHESION 1 PROTEIN 4	15	15	3618	3612	10487	9257			
14	CpXY^{h}9	Transcription regulator	1	1	1161	1161	n/a	n/a			
15	PXCpXY^{h}6‡◊	latex cyanogenic beta glucosidase	11	13	1150	1467	49422	63949			
16	PXCpXY^{h}10◊	RNA recognition splicing factor	14	16	1550	1834	18033	33985			
17	PXCpXY^{h}14	Chromatin/Nucleosome assembly factor	15	19	2420	2736	4299	6282			
18	CpXY^{h}10	Flowering locus t	4	4	525	525	1749	2024			
19	PXCpXY^{h}21	---NA---	1	1	346	363	n/a	n/a	10 & 8	35	90%

Figure 1 & S8-S9 ID	Gene Symbol	Putative Function	Number of Exons		Coding Sequence Length		Intron length		Autosome Homolog		
			X	HSY	X	HSY	X	HSY	Linkage Group	Supercontig	% identity
20	PY^{h}CpXY^{h}22	Vps51/Vps67 family protein	6	5	3240	3045	16497	16473			
21	PXCpXY^{h}22	---NA---	1	1	407	432	n/a	n/a			
22	PXY^{h}CpXY^{h}1	universal stress protein (USP) family protein	4	4	852	840	37086	34346			
23	PY^{h}CpXY^{h}23	Hypothetical protein	1	1	339	326	n/a	n/a			
24	CpXY^{h}11	PWWP domain-containing protein	1	1	3126	3111	n/a	n/a			
25	CpXY^{h}12	ADP-ribosylation factor A1B	5	6	552	558	6244	7526			
26	CpXY^{h}13	R2R3 Myb24 transcription factor (MYB24),	3	3	579	579	2527	2613			
27	CpXY^{h}14	Monodehydroascorbate reductase	7	7	1437	1437	5378	16264			
28	CpXY^{h}15	FU (FUSED); protein serine/threonine kinase	20	20	3804	3828	22887	22993			
29	CpXY^{h}16	Proteasome complex subunit	2	2	237	237	3485	3515	unknown	1914	100
30	CpXY^{h}17	Peptidyl-prolyl cis-trans isomerase (CYP38)	7	7	1347	1347	3097	3088			
31	CpXY^{h}18	Ruptured pollen grain 1	6	6	744	744	2458	2679			
32	PY^{h}CpXY^{h}27	---NA---	2	2	252	239	36	36			
33	CpXY^{h}19	Sequence-specific DNA binding	4	4	819	819	4481	4482			
34	CpXY^{h}20	MBD9, methyl-CpG binding	11	11	6612	6609	22522	37067			
35	CpXY^{h}21	Porin family 3	14	14	1272	1272	13810	13791			
36	PY^{h}CpXY^{h}28	DegP protease precursor	7	7	1029	1029	6420	6612			
37	CpXY^{h}22	MYB family transcription factor	2	2	243	243	166	166			
38	CpXY^{h}23	Carbonate dehydratase, zinc ion binding	9	9	1011	1011	2022	2031			
39	CpXY^{h}24	5'-AMP-activated protein kinase beta-1 subunit-related	10	10	1779	1779	7079	7134			
40	CpXY^{h}25	Hypothetical protein	2	2	549	555	273	273			
41	CpXY^{h}26	Phospholipid-transporting ATPase	9	9	2703	2718	4903	4772			

Figure 1 & S8-S9 ID	Gene Symbol	Putative Function	Number of Exons		Coding Sequence Length		Intron length		Autosome Homolog		
			X	HSY	X	HSY	X	HSY	Linkage Group	Supercontig	% identity
42	$PXY^{h}CpXY^{h}2$	---NA---	2	2	306	306	54	54			
43	$PY^{h}CpXY^{h}30$	---NA---	1	1	684	682	n/a	n/a			
44	$CpXY^{h}27$	protein kinase	19	19	2052	2052	8744	8696			
45	$CpXY^{h}28$	Proteasome assembly chaperone 3	5	5	264	264	1019	1015			
46	$PY^{h}CpXY^{h}31$	Hypothetical protein	2	1	465	290	157	n/a	unknown	3104	94
47	$CpXY^{h}29$	N-acetylglucosamine-phosphate mutase	9	7	1686	1362	32248	19588			
48	$CpXY^{h}30$	NUC (nutcracker); nucleic acid binding	1	1	209	312	n/a	n/a			
49	$CpXY^{h}31$	Asymmetric leaves 2	1	1	618	618	n/a	n/a			
50	$PXY^{h}CpXY^{h}3$	chloroplast ribosomal protein L12 and tRNA-Protein	1	1	722	652	n/a	n/a			
51	$CpXY^{h}32$	Ribonuclease P subunit	3	3	747	747	503	503	unknown	45	89-92
52	$CpXY^{h}33$	Protein kinase atmrk1	4	4	504	681	1559	1559			
53	$CpXY^{h}34$	Metal-dependent phosphohydrolase HD domain-containing protein	19	18	1413	1413	7477	7339			
54	$CpXY^{h}35$	Calcium homeostasis regulater-like protein	3	3	819	819	873	883			
55	$CpXY^{h}36$	Formate dehydrogenase	6	6	1149	1149	1496	1497			
56	$CpXY^{h}37$	ERA1 (ENHANCED RESPONSE TO ABA 1); farnesyltranstransferase	14	14	1299	1299	5840	5859			
57	$CpXY^{h}38$	Integral membrane protein	3	3	903	903	1285	1285			
58	$CpXY^{h}39$	Exonuclease family protein	6	6	1017	1017	718	718			
59	$CpXY^{h}40$	Kinesin light chain	2	2	1845	1845	80	80			
60	$CpXY^{h}41$	Carboxilic ester hydrolase	5	5	1149	1150	1369	1368			
61	$CpXY^{h}42$	AO (L-ASPARTATE OXIDASE); oxidoreductase	7	7	1956	1956	3349	3349			
62	$CpXY^{h}43$	ADP-ribosylation factor	10	10	1287	1287	15535	15247			
63	$CpXY^{h}44$	Protein phosphatase 4 core regulatory subunit R2	9	9	897	897	29605	31227			

Figure 1 & S8-S9 ID	Gene Symbol	Putative Function	Number of Exons		Coding Sequence Length		Intron length		Autosome Homolog		
			X	HSY	X	HSY	X	HSY	Linkage Group	Supercontig	% identity
64	CpXY^{h}45	Guanosine-3',5'-bis(diphosphate) 3'-pyrophosphohydrolase	11	11	1419	1419	41434	41537			
65	CpXY^{h}46	---NA---	1	1	498	498	n/a	n/a			
66	CpXY^{h}47	exportin 1a	35	35	3627	3615	17326	17187			
67	CpXY^{h}48	Mitochondrial NADH ubiquinone oxidoreductase 13kd-like subunit	2	2	174	174	158	158			
68	PXY^{h}CpXY^{h}4	---NA---	1	1	905	909	n/a	n/a			
69	CpXY^{h}49	Pentatricopeptide repeat-containing protein	10	10	3354	3354	3148	3156			
70	CpXY^{h}50	Hypothetical protein	2	2	309	309	81	81			

PXY^{h}CpXY^h = Both X and Y^h transcript units are pseudogenes
PXCpXY^h = X transcript is a pseudogene
PY^{h}CpXY^h = Y^h transcript is a psuedogene
**Gene is duplicated
NA- Not Available

Table S6. Estimated synonymous and non-synonymous nucleotide divergence between X-HSY gene pairs.

Gene ID	No. Sites (bp)				No. Mutations		Ks	Ka	Ka/Ks	Average Ks Value ± SE	
	Total Site	Total Coding sites	Syn sites	Non-Syn sites	Syn Mutations	Non-Syn mutations					
PXCpXY^{h}15◊	965	373	84.17	284.83	0	0	0.000	0.000	n/a		
CpXY^{h}1	5170	2265	530.08	1734.92	22	35	0.044	0.021	0.475		
CpXY^{h}2	5618	1884	454.75	1429.25	24	4	0.055	0.003	0.051		
CpXY^{h}3	2268	2265	520.25	1744.75	30	6	0.060	0.003	0.058		
PY^{h}CpXY^{h}2	369	366	86.75	279.25	10	23	0.125	0.087	0.697		
PY^{h}CpXY^{h}3	600	591	127.25	463.75	18	27	0.157	0.061	0.387		
CpXY^{h}4	294	291	65.75	225.25	15	9	0.272	0.041	0.151		
PY^{h}CpXY^{h}5	260	255	61.17	193.83	2	7	0.033	0.037	1.107		
PY^{h}CpXY^{h}7	813	573	133.00	440.00	11	10	0.088	0.023	0.263		
CpXY^{h}5	3251	807	190.17	616.83	12	15	0.066	0.025	0.374		
CpXY^{h}6	43584	3336	770.83	2565.17	55	61	0.076	0.024	0.321		
CpXY^{h}7	12446	2271	502.58	1768.42	32	36	0.068	0.021	0.310		
CpXY^{h}8	12792	3603	816.67	2786.33	55	93	0.071	0.034	0.484	0.094 ± 0.011	Inversion 1
CpXY^{h}9	1161	1158	239.83	918.17	22.5	22.5	0.100	0.025	0.249		
PXCpXY^{h}6	50608	1186	263.67	918.33	0	0	0.000	0.000	n/a		
PXCpXY^{h}10	19304	1271	269.83	981.17	29	39	0.116	0.041	0.352		
PXCpXY^{h}14	2420	2406	492.92	1913.08	33	70	0.070	0.038	0.535		
CpXY^{h}10	2274	525	128.08	396.92	14	5	0.118	0.013	0.108		
PXCpXY^{h}21	363	322	70.25	244.75	12.5	23.5	0.203	0.103	0.506		
PY^{h}CpXY^{h}22	3240	3033	704.75	2328.25	55.5	65.5	0.083	0.029	0.345		
PXCpXY^{h}22	407	405	85.92	310.08	14.25	33.75	0.187	0.118	0.627		
PXY^{h}CpXY^{h}1	852	834	177.67	656.33	10	34	0.059	0.054	0.917		
PY^{h}CpXY^{h}23	339	326	60.25	263.75	7	28	0.126	0.114	0.907		
CpXY^{h}11	3078	3063.00	676.67	2386.33	65.6	93.5	0.104	0.040	0.388		
CpXY^{h}12	6796	552	123.50	428.50	8	1	0.068	0.002	0.035		
CpXY^{h}13	3064	579	125.50	0.01	13	6	0.111	0.013	0.121		
CpXY^{h}14	6815	1437	357.08	1079.92	26	16	0.077	0.015	0.196		
CpXY^{h}15	26691	3804	901.33	2902.67	23	15	0.026	0.005	0.200		
CpXY^{h}16	3723	237	35.00	202.00	1	0	0.029	0.000	0.000		
CpXY^{h}17	4444	1347	312.00	1035.00	4	6	0.013	0.006	0.452		
CpXY^{h}18	3202	744	174.50	569.50	1	4	0.006	0.007	1.233		
PY^{h}CpXY^{h}27	252	237	55.00	182.00	1	3	0.018	0.017	0.905		
CpXY^{h}19	5300	819	183.17	635.83	0	0	0.000	0.000	n/a		
CpXY^{h}20	6612	6609	1497.33	5111.67	18	17	0.012	0.003	0.275		
CpXY^{h}21	15082	1272	290.17	981.83	6	1	0.021	0.001	0.049		
PY^{h}CpXY^{h}28	7440	1020	258.33	758.67	5	2	0.020	0.003	0.135		
CpXY^{h}22	409	243	51.83	191.17	0	0	0.000	0.000	n/a		
CpXY^{h}23	3033	1011	245.67	765.33	0	0	0.000	0.000	n/a		
CpXY^{h}24	8858	1779	413.75	1365.25	7	15	0.017	0.011	0.648	0.031 ± 0.007	Inversion 2
CpXY^{h}25	822	549	119.83	429.17	1	1	0.008	0.002	0.280		
CpXY^{h}26	7606	2703	621.42	2081.58	13	11	0.021	0.005	0.250		
PXY^{h}CpXY^{h}2	360	306	71.17	231.83	0	1	0.000	0.004	n/a		
PY^{h}CpXY^{h}30	684	675	147.00	528.00	7	9	0.049	0.017	0.350		
CpXY^{h}27	11123	2070	464.08	1605.92	4	7	0.009	0.004	0.504		
CpXY^{h}28	1284	264	63.33	200.67	3	1	0.049	0.005	0.104		
PY^{h}CpXY^{h}31	465	286	59.33	225.67	6	10	0.109	0.046	0.421		
CpXY^{h}29	32248	1680	395.75	1284.25	14	23	0.036	0.018	0.500		
CpXY^{h}30	308	308	62.00	244.00	6	16	0.104	0.069	0.662		
CpXY^{h}31	618	618	144.33	470.67	12	1	0.088	0.002	0.024		
PXY^{h}CpXY^{h}3	722	648	160.00	488.00	16	18	0.107	0.038	0.352		
CpXY^{h}32	1260	747	168.25	578.75	3	13	0.018	0.023	1.270		
CpXY^{h}33	2066	504	124.67	379.33	2	1	0.016	0.003	0.163		

Gene ID	No. Sites (bp)				No. Mutations		Ks	Ka	Ka/Ks	Average Ks Value ± SE	
	Total Site	Total Coding sites	Syn sites	Non-Syn sites	Syn Mutations	Non-Syn mutations					
CpXY^{h}34	8970	1413	312.25	1100.75	6	3	0.019	0.003	0.140	0.007 ± 0.002	Collinear
CpXY^{h}35	1693	819	188.67	630.33	0	0	0.000	0.000	n/a		
CpXY^{h}36	2629	1149	267.50	881.50	0	0	0.000	0.000	n/a		
CpXY^{h}37	7139	1299	300.17	998.83	2	5	0.007	0.005	0.753		
CpXY^{h}38	2188	903	197.17	705.83	1	0	0.005	0.000	0.000		
CpXY^{h}39	1743	1017	226.00	791.00	0	0	0.000	0.000	n/a		
CpXY^{h}40	1925	1845	421.50	1423.50	2	2	0.005	0.001	0.296		
CpXY^{h}41	2518	1149	270.67	878.33	1	2	0.004	0.002	0.619		
CpXY^{h}42	5296	1956	468.17	1487.83	2	1	0.004	0.001	0.157		
CpXY^{h}43	16795	1287	267.00	1020.00	3	5	0.013	0.005	0.411		
CpXY^{h}44	30503	897	192.92	704.08	4	2	0.021	0.003	0.136		
CpXY^{h}45	42869	1419	340.17	1078.83	2	2	0.006	0.002	0.315		
CpXY^{h}46	498	495	130.00	365.00	1	0	0.008	0.000	0.000		
CpXYh47	3615	3609	808.92	2800.08	4	1	0.005	0.000	0.073		
CpXY^{h}48	333	174	38.83	135.17	0	0	0.000	0.000	n/a		
PXY^{h}CpXY^{h}4	906	906	196.67	709.33	4	5	0.021	0.007	0.343		
CpXY^{h}49	6503	3354	747.67	2606.33	3	10	0.004	0.004	0.958		
CpXY^{h}50	391	309	63.83	245.17	0	0	0.000	0.000	n/a		

Table S7. Estimated age of divergence X-HSY gene pairs.

Gene ID	Silent Sites	Silent mutations	Silent divergence (K_{sil})	Estimated age (MYA)	Average Estimated age (MYA) ± SE	
CpXY^{h}1	3412.08	51	0.015	1.906	7.033 ± 0.853	Inversion 1
CpXY^{h}2	4153.75	127	0.031	3.903		
CpXY^{h}5	2615.17	154	0.061	7.666		
CpXY^{h}6	38859.83	7551	0.225	---		
CpXY^{h}7	10088.58	629	0.065	8.144		
CpXY^{h}8	9999.70	608	0.063	7.925		
CpXY^{h}10	1872.08	93	0.051	6.425		
CpXY^{h}12	6293.50	454	0.076	9.475		
CpXY^{h}13	2613.50	186	0.075	9.348		
CpXY^{h}14	5627.08	366	0.068	8.505		
CpXY^{h}15	23714.33	721	0.031	3.880	1.931 ± 0.447	Inversion 2
CpXY^{h}16	3516.00	42	0.012	1.505		
CpXY^{h}17	3376.00	30	0.009	1.118		
CpXY^{h}18	2618.50	46	0.018	2.223		
CpXY^{h}19	4667.17	1	0.000	0.025		
CpXY^{h}21	13994.17	741	0.055	6.864		
CpXY^{h}22	220.83	0	0.000	0.000		
CpXY^{h}23	2261.67	13	0.006	0.721		
CpXY^{h}24	7478.75	117	0.016	1.976		
CpXY^{h}25	395.83	1	0.003	0.316		
CpXY^{h}26	5393.42	121	0.023	2.848		
CpXY^{h}27	9451.08	131	0.014	1.750		
CpXY^{h}28	1082.33	20	0.019	2.339		
CpXY^{h}29	25702.75	2335	0.097	---		
CpXY^{h}32	674.25	11	0.016	2.061		
CpXY^{h}33	1686.67	18	0.011	1.344		
CpXY^{h}34	7617.25	77	0.010	1.273	0.619 ± 0.121	Collinear
CpXY^{h}35	1065.67	0	0.000	0.000		
CpXY^{h}36	1749.50	3	0.002	0.215		
CpXY^{h}37	6137.17	35	0.006	0.715		
CpXY^{h}38	1485.17	1	0.001	0.084		
CpXY^{h}39	952.00	0	0.000	0.000		
CpXY^{h}40	504.50	2	0.004	0.496		
CpXY^{h}41	1642.67	7	0.004	0.534		
CpXY^{h}42	3807.17	20	0.005	0.659		
CpXY^{h}43	15490.00	122	0.008	0.994		
CpXY^{h}44	29686.92	134	0.005	0.566		
CpXY^{h}45	41708.17	292	0.007	0.879		
CpXY^{h}48	200.83	1	0.005	0.625		
CpXY^{h}49	3899.67	17	0.004	0.546		
CpXY^{h}50	148.83	2	0.014	1.695		

We excluded pseudogenes because of uncertainties in their alignments. Genes with incomplete BAC sequences in intron regions and single exon genes were also excluded from the analysis. Genes CpXYh6 and CpXYh29 were unable to be accurately aligned due to extreme differences in intron sizes, therefore an accurate divergence age cannot be estimated.

Note: Due to the comparatively short generation time found in Arabidopsis compared to papaya and the perennial nature of papaya, it is possible the divergence age reported here for papaya may somewhat vary from the actual divergence age.

Table S8. Comparisons of K_S and K_{silent} between the three regions identified by the inversions-

	K_{silent}		K_S	
Comparison	Z value	P value	Z value	P value
Gene pairs				
Inversion 1 versus inversion 2 genes	-3.24	0.002	-4.16	<0.0001
Inversion 2 versus collinear region genes	-2.84	0.004	-2.47	0.013
All sequence pairs, including single exon genes and pseudogenes				
Inversion 1 versus inversion 2 genes	—	—	3.67	0.0002
Inversion 2 versus collinear region genes	—	—	-3.10	0.0018

Table S9. Pairs of genes found in both the X and Y^h sequences, and their positions in the X and HSY pseudomolecules.

Gene Symbol	Putative Function	# of Exons		Coding Sequence Length		Intron length		Autosome Homolog			BAC				Inversion
								Linkage Group #	Superc ontig #	% identity	Position	ID	Position	ID	
		X	HSY	X	HSY	X	HSY				X		HSY		
$CpXY^h1$	Regulator of Vps4 activity in the MVB pathway	6	6	2280	2265	2890	5378				77370-82539	SH61H02	97148-104790	SH95B12	1
$CpXY^h2$	Somatic embryogenesis receptor kinase	11	11	1884	1884	3734	3702				69931-75548	SH61H02	89764-95349	SH95B12	1
$CpXY^h3$	Exocyst complex subunit SEC6	25	25	2268	2268	40497	7483				66454-91181	SH31E12	1-34368	SH95B12	1
											1-32799	SH61H02	168590-173463	SH49L11	1
**$CpXY^h4$	Late embryogenesis abundant protein 18	1	1	293	293	n/a	n/a				30641-30933	SH31E12	103645-103938	SH49L11	1
													73364-73657	SH49L11	1
$CpXY^h5$	4-nitrophenylphosphatase	8	8	807	816	2444	2448				66241-69491	SH79M13	95427-98690	SH69E13	1
$CpXY^h6$	FRA3 inositol or phosphatidylinositol phosphatase	10	10	3336	3336	40354	65434				54810-98499	SH54A04	6542-75311	SH96A24	1
$CpXY^h7$	Leucine zipper-ef-hand (Ca binding motif)	14	14	2271	2286	10175	19777				59838-72283	SH55B21	113904-135966	SH96A24	1
$CpXY^h8$	SYN4 (SISTER CHROMATID COHESION 1 PROTEIN 4	15	15	3618	3612	10487	9257				97208-111312	SH55B21	73441-86309	SH81O12	1
$CpXY^h9$	Transcription regulator	1	1	1161	1161	n/a	n/a				187518-188678	SH55B21	129678-130838	SH81O12	1
$CpXY^h10$	Flowering locus t	4	4	525	525	1749	2024				87468-89741	SH65C15	55397-57945	SH78H09	1
$CpXY^h11$	PWWP domain-containing protein	1	1	3126	3111	n/a	n/a				110760-113885	SH69D24	95120-98230	SH53G04	1
$CpXY^h12$	ADP-ribosylation factor A1B	5	6	552	558	6244	7526				159681-166476	SH54M13	32224-40307	SH81O12	1
$CpXY^h13$	R2R3 Myb24 transcription factor (MYB24),	3	3	579	579	2527	2613				74982-78087	SH46O19	135292-138484	SH90D06	1
$CpXY^h14$	Monodehydroascorbate reductase	7	7	1437	1437	5378	16264				22083-28897	SH46O19	93935-111635	SH90D06	1
$CpXY^h15$	FU (FUSED); protein serine/threonine kinase	20	20	3804	3828	22887	22993				33225-59915	SH55E02	56523-83343	SH72L16	2
$CpXY^h16$	Proteasome complex subunit	2	2	237	237	3485	3515	n/a	1914	100	136991-140712	SH55E02	130400-134151	SH72L16	2
$CpXY^h17$	Peptidyl-prolyl cis-trans isomerase (CYP38)	7	7	1347	1347	3097	3088				114633-119076	SH23D07	36026-40460	AM57M14	2

Gene Symbol	Putative Function	# of Exons		Coding Sequence Length		Intron length		Autosome Homolog			BAC				Inversion
								Linkage Group #	Supercontig #	% identity	Position	ID	Position	ID	
		X	HSY	X	HSY	X	HSY				X		HSY		
CpXY^{h}18	Ruptured pollen grain 1	6	6	744	744	2458	2679				128038-131239	SH23D07	15924-19346	SH88A07	2
CpXY^{h}19	Sequence-specific DNA binding	4	4	819	819	4481	4482				24017-29316	SH24L08	51193-56493	SH88A07	2
CpXY^{h}20	MBD9, methyl-CpG binding	11	11	6612	6609	22522	37067				166625-195755	SH84J07	91204-134880	SH05K22	2
CpXY^{h}21	Porin family 3	14	14	1272	1272	13810	13791				111656-126737	SH84J07	29895-44957	SH59G12	2
CpXY^{h}22	MYB family transcription factor	2	2	243	243	166	166				87116-87524	SH84J07	69860-70268	SH59G12	2
CpXY^{h}23	Carbonate dehydratase, zinc ion binding	9	9	1011	1011	2022	2031				63692-66724	SH84J07	77740-80781	SH102N12	2
CpXY^{h}24	5'-AMP-activated protein kinase beta-1 subunit-related	10	10	1779	1779	7079	7134				51917-60774	SH84J07	67453-76365	SH102N12	2
CpXY^{h}25	Hypothetical protein	2	2	549	555	273	273				43556-44377	SH84J07	58388-59215	SH102N12	2
CpXY^{h}26	Phospholipid-transporting ATPase	9	9	2703	2718	4903	4772				2579-10170	SH10J17	120094-127583	SH64O06	2
CpXY^{h}27	protein kinase	20	20	2091	2088	9032	8988				19577-30699	SH53E18	34767-45842	SH64O06	2
CpXY^{h}28	Proteasome assembly chaperone 3	5	5	264	264	1019	1015				31445-32728	SH53E18	32709-33988	SH64O06	2
CpXY^{h}29	N-acetylglucosamine-phosphate mutase	9	7	1686	1362	32248	19588				86982-119229	SH53E18	87673-107260	SH59O17	2
CpXY^{h}30	NUC (nutcracker); nucleic acid binding	1	1	209	312	n/a	n/a				92981-93289	SH53E18	92916-93227	SH59O17	2
CpXY^{h}31	Asymmetric leaves 2	1	1	618	618	n/a	n/a				201739-202356	SH53E18	135383-136000	SH59O17	2
CpXY^{h}32	Ribonuclease P subunit	3	3	747	747	503	503	n/a	45	89-92	46758-48017	SH87M18	44020-45269	SH05K22	2
CpXY^{h}33	Protein kinase atmrk1	4	4	504	681	1559	1559				200558-202622	SH49A07	57804-60043	AM57M14	2
CpXY^{h}34	Metal-dependent phosphohydrolase HD domain-containing protein	19	18	1413	1413	7477	7339				155987-164916	SH49A07	151017-159769	SH75H06	Collinear
CpXY^{h}35	Calcium homeostasis regulater-like protein	3	3	819	819	873	883				150401-152093	SH49A07	145436-147128	SH75H06	Collinear
CpXY^{h}36	Formate dehydrogenase	6	6	1149	1149	1496	1497				147308-149953	SH49A07	142343-144988	SH75H06	Collinear
CpXY^{h}37	ERA1 (ENHANCED RESPONSE TO ABA 1); farnesyltranstransferase	14	14	1299	1299	5840	5859				100081-107219	SH49A07	95081-102238	SH75H06	Collinear
CpXY^{h}38	Integral membrane protein	3	3	903	903	1285	1285				93334-	SH49A07	88332-	SH75H06	Collinear

Gene Symbol	Putative Function	# of Exons		Coding Sequence Length		Intron length		Autosome Homolog			BAC				Inversion
								Linkage Group #	Superc ontig #	% identity	Position	ID	Position	ID	
		X	HSY	X	HSY	X	HSY				X		HSY		
											95521		90519		
CpXY^{h}39	Exonuclease family protein	6	6	1017	1017	718	718				76141-77880	SH49A07	71136-72875	SH75H06	Collinear
CpXY^{h}40	Kinesin light chain	2	2	1845	1845	80	80				12067-13991	SH49A07	7038-8962	SH75H06	Collinear
CpXY^{h}41	Carboxilic ester hydrolase	5	5	1149	1150	1369	1368				8529-11046	SH49A07	1176-3693	SH58C24	Collinear
CpXY^{h}42	AO (L-ASPARTATE OXIDASE); oxidoreductase	7	7	1956	1956	3349	3349				4217-5228	SH49A07	6994-12298	SH58C24	Collinear
CpXY^{h}43	ADP-ribosylation factor	10	10	1287	1287	15535	15247				167950-184772	SH85C03	23653-40187	SH58C24	Collinear
CpXY^{h}44	Protein phosphatase 4 core regulatory subunit R2	9	9	897	897	29605	31227				125940-156442	SH85C03	51396-83520	SH58C24	Collinear
CpXY^{h}45	Guanosine-3',5'-bis(diphosphate) 3'-pyrophosphohydrolase	11	11	1419	1419	41434	41537				71785-114637	SH85C03	94730-137686	SH58C24	Collinear
CpXY^{h}46	---NA---	1	1	498	498	n/a	n/a				62018-62515	SH85C03	146943-147440	SH58C24	Collinear
CpXY^{h}47	exportin 1a	35	35	3627	3615	17326	17187				19289-40143	SH86B15	84828-105629	SH60M19	Collinear
CpXY^{h}48	Mitochondrial NADH ubiquinone oxidoreductase 13kd-like subunit	2	2	174	174	158	158				43977-44309	SH86B15	109287-109619	SH60M19	Collinear
CpXY^{h}49	Pentatricopeptide repeat-containing protein	10	10	3354	3354	3148	3156				51972-58474	SH86B15	117258-123768	SH60M19	Collinear
CpXY^{h}50	Hypothetical protein	2	2	309	309	81	81				65923-66313	SH86B15	131211-131601	SH60M19	Collinear
PXY^{h}CpXY^{h}1	universal stress protein (USP) family protein	4	4	852	840	37086	34346				69753-107690	SH69D24	57117-92302	SH53G04	1
PXY^{h}CpXY^{h}2	---NA---	2	2	306	306	54	54				73287-73646	SH40J04	73712-74071	SH64O06	2
PXY^{h}CpXY^{h}3	chloroplast ribosomal protein L12 and tRNA-Protein	1	1	722	652	n/a	n/a				12839-13560	SH06M23	72741-94883	SH62H24	2
PXY^{h}CpXY^{h}4	---NA---	1	1	905	909	n/a	n/a				45572-46477	SH86B15	110870-111775	SH60M19	Collinear

PXY^{h}CpXY^h = Both X and Y^h transcript units are pseudogenes. They are numbered in consecutive order as they appear on the X physical map.
** duplication
‡Gene spans 2 BACs
NA- Not Available

Table S10. List of Y^h-specific genes plus pseudogenes in the HSY sequence with papaya autosomal homologs

Gene Symbol	Putative Function	Number of Exons	Coding Sequence Length	Intron length	Autosome Homolog			BAC			Inversion
					Super-contig #	Linkage Group	Percent Identity	Position	ID	Accession Number	
CpY^h-1	hypothetical protein	2	273	103	368	NA	100%	35858-36233	AM125I09	AC239252	1
CpY^h-2	---NA---	1	198	NA	1657	NA	100%	45118-45315	AM125I09	AC239252	1
CpY^h-3	---NA---	1	213	NA	10	5	81%	82486-82698	SH67H04	AC238615	1
CpY^h-4	---NA---	1	294	NA				146858-147151	SH81O12	AC239165	1
CpY^h-5	---NA---	2	153	2824				166701-169678	SH81O12	AC239165	1
$PXCpXY^h$6‡◊	latex cyanogenic beta glucosidase	13	1467	63949				156303-178595, 40004-83126	SH81O12, SH32G10	AC238596	1
CpY^h-7**	---NA---	1	237	NA				184093-184330	SH32G10	AC238596	1
								77447-77683	SH32G10	AC238596	1
CpY^h-8	---NA---	1	105	NA	29	2	87%	47046-47150	SH32G10	AC238596	1
CpY^h-9	---NA---	1	318	NA	29	2	81%	46506-46824	SH32G10	AC238596	1
$PXCpXY^h$10◊	RNA recognition motif-containing protein-splicing factor	16	1834	35818				43145-78963	SH65D15	AC238767	1
CpY^h-11	---NA---	3	420	1231				38621-40271	SH65D15	AC238767	1
CpY^h-12	---NA---	1	288	NA				23717-24004	SH65D15	AC238767	1
CpY^h-13	---NA---	1	261	NA				22245-22705	SH65D15	AC238767	1
$PXCpXY^h$14	Chromatin/Nucleosome assembly factor	19	2736	6282				144247-153264	SH50M09	AC238602	1
$PXCpXY^h$15◊	---NA---	1	965	NA				45539-46503	SH69E13	AC238619	1
CpY^h-16	---NA---	1	336	NA				82847-83182	SH72J22	AC238620	1
CpY^h-17	---NA---	2	348	173				31224-31744	SH71E16	AC239163	1
CpY^h-18**	---NA---	2	936	37	211	6	100%	29807-30779	SH71E16	AC239163	1

Gene Symbol	Putative Function	Number of Exons	Coding Sequence Length	Intron length	Autosome Homolog			BAC			Inversion
					Super-contig #	Linkage Group	Percent Identity	Position	ID	Accession Number	
								25944-26916	SH71E16	AC239163	1
CpY^{h}-19†	MADS_box protein	7	554	5369	55	2	93%	7864-13719, 22092-22158	SH71E16	AC239163	1
CpY^{h}-20	---NA---	2	396	216				10200-10811	SH71E16	AC239163	1
PXCpXY^{h}21	---NA---	1	363	NA	35	10 & 8	90%	59558-59920	SH64O06	AC239158	2
PXCpXY^{h}22	---NA---	1	432	NA				75917-76651	SH23L24	AC238593	Collinear
PCpY^{h}-1	zinc finger-homeodomain protein	2	701	34	55	2	92%	75917-76651	SH85B16	AC238630	N/A
PCpY^{h}-2	---NA---	1	1062	NA				139702-140763	SH81O12	AC239165	1
PCpY^{h}-3	---NA---	1	2366	NA				71306-73658	SH32G10	AC238596	1
PCpY^{h}-4	photosystem II protein D1	2	490	28	3	10 & 8	95%	13836-14353	SH32G10	AC238596	1
PCpY^{h}-5	---NA---	1	234	NA				114992-115225	SH65D15	AC238767	1
PCpY^{h}-6◊	---NA---	1	237	NA				86266-86502	SH65D15	AC238767	1
PCpY^{h}-7	---NA---	1	447	NA				84427-84873	SH65D15	AC238767	1
PCpY^{h}-8	---NA---	2	375	29917				157607-157839	SH50M09	AC238602	1
PCpY^{h}-9	---NA---	1	320	NA				86998-87317	SH72J22	AC238620	1
PCpY^{h}-10	ATP synthase CFO subunit 1 protein	2	434	524	30	11 & 9	95%	63737-64694	SH20J12	AC239142	1

Pseudogenes are denoted by a "P" and both HSY-specific genes and pseudogenes are numbered in consecutive order as they appear on the HSY physical map.

**Gene is duplicated

† Gene spans 2 unordered BAC contigs, which is reflected in the BAC position, and the actual coding and intron length are larger than what is listed in the table.

‡Gene spans 2 BACs

◊Gene is fragmented on the X due to gaps in the X BAC sequence. Whether this gene is truly a pseudogene on the X is unknown.

NA- Not Available

Table S11. List of papaya X-specific genes and pseudogenes in the papaya X sequence

Gene Symbol	Putative function	# of Exons	Coding Sequence Length	Intron length	Autosome Homolog			BAC position	BAC ID	Accession Number	Inversion
					Linkage Group #	Super-contig #	% identity				
CpX-1	---NA---	1	355	NA				124509-124863	SH61H02	EF661023	1
PY^{h}CpXY^{h}2	---NA---	1	369	NA				11583-11951	SH61H02	EF661023	1
PY^{h}CpXY^{h}3	---NA---	1	600	NA				4966-5565	SH61H02	EF661023	1
CpX-4	Late embryogenesis abundant protein 18(lEA6)	1	294	NA				30640-30933	SH31E12	AC238595	1
PY^{h}CpXY^{h}5	---NA---	1	261	NA				26912-27172	SH31E12	AC238595	1
CpX-6	---NA---	2	351	71				9356-9777	SH31E12	AC238595	1
PY^{h}CpXY^{h}7	---NA---	7	813	2034				87299-90145	SH14F24	FJ429365	1
CpX-8	---NA---	4	1104	3143				606-4852	SH79M13	AC239204	1
CpX-9	---NA---	3	468	140				33662-34269	SH79M13	AC239204	1
CpX-10	---NA---	2	432	532				51828-52791	SH79M13	AC239204	1
CpX-11	---NA---	1	207	NA				58107-58313	SH79M13	AC239204	1
CpX-12	---NA---	1	183	73				140453-140708	SH79M13	AC239204	1
CpX-13	Beta-hexosaminidase	4	375	1564				145210-147148	SH79M13	AC239204	1
CpX-14	Beta-hexosaminidase	2	273	67				70136-70475	SH83B06	AC238628	1
CpX-15	---NA---	1	775	NA				76137-76911	SH83B06	AC238628	1
CpX-16	---NA---	1	246	NA				125817-126062	SH83B06	AC238628	1
CpX-17	---NA---	1	186	NA				75754-75939	SH54A04	AC239201	1
CpX-18	---NA---	1	249	NA				109919-110167	SH54A04	AC239201	1
CpX-19	Defensin protein	1	178	NA				140283-140460	SH54A04	AC239201	1
CpX-20	---NA---	1	249	NA				133351-133599	SH55B21	FJ429367	1
CpX-21	---NA---	1	474	NA				157918-158391	SH55B21	FJ429367	1
PY^{h}CpXY^{h}22	Vps51/Vps67 family protein	6	3240	16497				178754-180952	SH51A03	AC238603	1
PY^{h}CpXY^{h}23	Hypothetical protein	1	339	NA				86822-87160	SH69D24	AC239162	1

Gene Symbol	Putative function	# of Exons	Coding Sequence Length	Intron length	Autosome Homolog			BAC position	BAC ID	Accession Number	Inversion
					Linkage Group #	Super-contig #	% identity				
CpX-24	---NA---	1	206	NA				3337-3542	SF08K16	AC239251	1
CpX-25	---NA---	1	195	NA				48016-48210	SH54M13	AC238609	1
CpX-26	---NA---	1	387	NA				74095-74481	SH38K02	AC239149	1
PY^{h}CpXY^{h}27	---NA---	2	252	36				142159-142446	SH23D07	AC238592	2
PY^{h}CpXY^{h}28	DegP protease precursor	7	1029	6420				100325-107773	SH84J07	AC238629	2
CpX-29	---NA---	1	465	NA				20405-20869	SH40J04	AC238598	2
PY^{h}CpXY^{h}30	---NA---	1	684	NA				83840-84523	SH40J04	AC238598	2
PY^{h}CpXY^{h}31	Hypothetical protein	2	465	157	NA	3104	94	52581-53202	SH53E18	EF661026	2
CpX-32	---NA---	1	170	NA				125470-125639	SH53E18	EF661026	2
CpX-33	---NA---	2	135	45				197429-197608	SH53E18	EF661026	2
CpX-34	---NA---	1	315	NA				242470-242784	SH53E18	EF661026	2
PCpX-1	Aspartic proteinase family protein	1	1185	NA				119238-120423	SH61H02	EF661023	1
PCpX-2	Angio-associated migratory cell protein/transducin family protein	3	376	465				117541-118381	SH61H02	EF661023	1
PCpX-3	---NA---	1	153	NA				109764-109916	SH54A04	AC239201	1
PCpX-4	---NA---	1	352	NA				65772-66123	SH51A03	AC238603	1

Pseudogenes are denoted by a "P" and are numbered in consecutive order as they appear on the physical map.
The X-specific genes are numbered as they appear on the X physical map. For a compiled list of the paired, HSY, and X genes, numbered in order, see Table S5.

NA- Not Available

Table S12. Summary of the numbers of transcription units found on the HSY and the corresponding X chromosomal region, showing three breakdowns: (i) into sequences that encode potentially functional products versus pseudogenes, (ii) into genes with potential functions, based on sequences in other plants and the analyses described above, versus ones with unknown functions, and (iii) into the genes' locations. The data are summarized from Tables S10, S11 and S12, as detailed in the footnotes.

	Total genes		Known genes		New genes (unknown functions)		Locations of genes: Stratum 1		Stratum 2		Collinear region	
	HSY	**X**	**HSY**	**X**	**HSY**	**X**	**HSY**	**X**	**HSY**	**X**	**HSY**	**X**
Protein coding transcription units in HSY or in X	**72**	**84**	54	57	18	27	20	26	1	8	1	0
HSY/X pairs of transcription units	**50**		49		1		14		19		17	
Autosomal origins	2		2		0		—		—		—	
Intact genes on HSY or X with pseudogene in the homolog	16		8		8		4	6	1	4	1	0
X pseudogene-HSY intact[1]	—	**6**	—	3	—	3	4		1		1	
HSY pseudogene-X intact[2]	**10**	—	5	—	5	—		6		4		0
HSY- or X-specific transcription units	**16**[3]	**24**[4]	2	4	14	20	16	20	0	4	0	0
Autosomal origins	7	0	2	0	5	0	—		—		—	
Unknown origins	9	24	0	4	9	20	—		—		—	
Pseudogenes with no functional counterpart in either the X or HSY	**14**	**8**	5	4	9	4	1		2		1	
Pseudogene pairs (in both HSY and X)	**4**		2		2		1		2		1	
Pseudogenes specific to HSY or X	**10**[5]	**4**[6]	3	2	7	2	10	4	0	0	0	0
Autosomal origins	3	0	3	0	0	0	—		—		—	
Unknown origins	7	4	0	2	7	2	—		—		—	

[1]Denoted by PXCpXY^h in Table S10
[2]PY^{h}CpXY^h in Table S11
[3]CpY^h in Table S10
[4]CpX in Table S11
[5]PCpY^h in Table S10
[6]PCpX in Table S11

.

Table S13. Gene densities and repetitive element contents of the inversions and the collinear region of the HSY and the homologous region of the X chromosome.

Location	Sequence length (kb)	Number of genes		Gene density (kb/1 gene)	Repetitive elements
HSY					
Inversion 1	3680	HSY-specific	20	175	**80.7%**
		Pair	14	307	
		Total	**33**	**112**	
Inversion 2	3097	HSY-specific	1	1548	**80.2%**
		Pair	19	163	
		Total	**21**	**147**	
Collinear region	545	HSY-specific	1	272	**48.2%**
		Pair	17	34	
		Total	**18**	**30**	
X chromosome homologous region					
Inversion 1	1915	X-specific	26	74	**76.8%**
		Pair	14	147	
		Total	**39**	**49**	
Inversion 2	912	X-specific	8	114	**60.5%**
		Pair	19	51	
		Total	**26**	**35**	
Collinear region	454	X-specific	0	-	**45.5%**
		Pair	17	28	
		Total	**16**	**28**	
Genome wide average	**372000**	**Total**	**22934**	**16**	**51.9%**

Pseudogenes are excluded from the gene totals in this table. The genome wide data are from Ming et al.,2008 (12).

Table S14. GenBank accession numbers of the papaya HSY and X BACs.

BAC ID	GenBank Accession No.	Insert size (Kb)	Chromosome
SH01I11	AM778099	107	HSY
SH05K22	AC239135	96	HSY
SH07G04	AC238760	106	HSY
SH11D16	AC238589	106	HSY
SH12D03	AC239139	98	HSY
SH14F22	AC239140	93	HSY
SH17J04	AC238761	95	HSY
SH20J12	AC239142	109	HSY
SH23L24	AC238593	102	HSY
SH27G03	AC239146	80	HSY
SH30L14	AC238594	94	HSY
SH32G10	AC238596	89	HSY
SH36M01	AC239148	69	HSY
SH41F24	EF625817	119	HSY
SH47B08	AC238600	97	HSY
SH49L11	AC238601	173	HSY
SH50M09	AC238602	196	HSY
SH52H15	AC238762	242	HSY
SH52N19	AC238763	63	HSY
SH53E10	AC238764	198	HSY
SH53G04	AC238607	166	HSY
SH53O12	AC238608	129	HSY
SH54H01	EF625819	189	HSY
SH57B23	AC238765	145	HSY
AM57M14	AC239203	108	HSY
SH58C24	AC239153	148	HSY
SH59G12	AC238611	102	HSY
SH59O17	AC239154	163	HSY
SH60M19	JN088492	255	HSY
SH61F18	AC238766	202	HSY
SH61K24	AC239155	138	HSY
SH62E'02	AC239156	185	HSY
SH62H24	AC239157	176	HSY
SH64O06	AC239158	139	HSY
SH65D15	AC238767	124	HSY
SH66N15	AC238768	224	HSY
SH67H04	AC238615	178	HSY
SH68C12	AC238616	146	HSY
SH69A08	AC238769	179	HSY
SH69C01	AC238618	42	HSY
SH69E13	AC238619	217	HSY
SH71E16	AC239163	163	HSY
SH72J22	AC238620	173	HSY
SH72L16	AC238621	183	HSY
SH74D06	AC238770	154	HSY
SH75H06	AC238771	222	HSY
SH76M08	EF625820	133	HSY
SH78H09	AC238624	223	HSY
SH79C23	AC238625	105	HSY

SH80L24	AC239164	195	HSY
SH81O12	AC239165	186	HSY
SH82H21	AC238627	188	HSY
SH85B16	AC238630	150	HSY
SH85B24	EF661024	294	HSY
SH88A07	AC239168	156	HSY
SH89I15	AC238633	180	HSY
SH89M06	AM778096	192	HSY
SH90D06	AC238772	221	HSY
SH92K23	AC238635	184	HSY
SH92L01	AC239169	163	HSY
SH93H03	AC238636	177	HSY
SH94E22	EF625821	184	HSY
SH95B12	EF661025	146	HSY
SH96A24	AC239171	189	HSY
SH99O03	FJ429372	128	HSY
SH102N12	FJ429373	84	HSY
AM125I09	AC239252	80	HSY
AM126G07	AC239198	98	HSY
SH06M23	AC239136	68	X
SF08K16	AC239251	70	X
SH10J17	AC239138	106	X
SH12G20	GU799580	95	X
SH12I03	AC238590	76	X
SH14F24	FJ429365	104	X
SH22E’07	AC239143	76	X
SH23D07	AC238592	141	X
SH23D15	AC239144	20	X
SH24L08	AC239145	76	X
SH30B21	AC239147	101	X
SH31E12	AC238595	91	X
SH38K02	AC239149	97	X
SH39A12	AC238597	49	X
SH40J04	AC238598	103	X
SH42B05	EF625818	91	X
SH46O19	AC238599	103	X
SH49A07	AC239150	203	X
SH50J21	AC239200	343	X
SH50M24	JN088493	148	X
SH51A03	AC238603	184	X
SH53E18	EF661026	252	X
SH54A04	AC239201	200	X
SH54A09	AC239152	158	X
SH54M13	AC238609	216	X
SH54M22	AC239202	168	X
SH55B21	FJ429367	198	X
SH55E02	FJ429368	159	X
SH61H02	EF661023	168	X
SH65C15	AC239160	153	X
SH69D05	AC239161	190	X
SH69D24	AC239162	144	X
SH79M13	AC239204	153	X

SH80F18	AC238626	183	X
SH83B06	AC238628	141	X
SH84J07	AC238629	222	X
SH84M10	AC239166	259	X
SH85C03	AC238631	190	X
SH86B15	AC239167	182	X
SH87C01	NA	239	X
SH87M18	AC238632	81	X
SH93K15	AC239170	186	X
SH95I24	GU799579	218	X
SH99B01	AC238637	157	X
AM136D11	AC239253	89	X

Table S15. GC content and CpG island in the HSY, the corresponding X sequence, and the complete papaya genome sequence.

	HSY (8.1 Mb)		**X (3.5 Mb)**		**X Knob 1 (1.9 Mb)**		**Genome**	
	Total	**Masked**	**Total**	**Masked**	**Total**	**Masked**	**Total**	**Masked**
Sequence length	8070272	1675370	3458800	1086276	1921421	765855	273570756	133701001
Bases occupied by CpG islands	253397	27967	124448	16125	97812	19353	10984317	2315247
%GC	36.19	35.71	35.73	34.17	35.99	34.6	35.32	33.45
No. of CpG islands	838	91	400	50	307	58	34680	6923
No. of CpG islands/Mb of total sequence	103.84	11.3	115.65	14.5	159.78	30.2	126.77	25.3
CpG island size range (bp)	200-1433	200-1433	200-959	200-771	200-1316	201-1316	200-3037	200-1702