Sequencing papaya X and Y^h chromosomes reveals molecular basis of incipient sex chromosome evolution

Jianping Wang[a,1,2], Jong-Kuk Na[a,1,3], Qingyi Yu[b,c,1], Andrea R. Gschwend[a,1], Jennifer Han[a], Fanchang Zeng[a], Rishi Aryal[a], Robert VanBuren[a], Jan E. Murray[a], Wenli Zhang[d], Rafael Navajas-Pérez[e,4], F. Alex Feltus[e,5], Cornelia Lemke[e], Eric J. Tong[c], Cuixia Chen[a,6], Ching Man Wai[c,f], Ratnesh Singh[c], Ming-Li Wang[c], Xiang Jia Min[g], Maqsudul Alam[h], Deborah Charlesworth[i], Paul H. Moore[c], Jiming Jiang[d], Andrew H. Paterson[e], and Ray Ming[a,7]

[a]Department of Plant Biology, University of Illinois at Urbana–Champaign, Urbana, IL 61801; [b]Texas AgriLife Research Center, Department of Plant Pathology and Microbiology, Texas A&M University, Weslaco, TX 78596; [c]Hawaii Agriculture Research Center, Kunia, HI 96759; [d]Department of Horticulture, University of Wisconsin, Madison, WI 53706; [e]Plant Genome Mapping Laboratory, University of Georgia, Athens, GA 30606; [f]Department of Tropical Plants and Soil Sciences, University of Hawaii, Honolulu, HI 96822; [g]Department of Biological Sciences, Youngstown State University, Youngstown, OH 44555; [h]Advanced Studies in Genomics, Proteomics and Bioinformatics, University of Hawaii, Honolulu, HI 96822; and [i]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh EH9 3JT, United Kingdom

Edited by Ralf G Kynast, Royal Botanic Gardens, Kew, Richmond, United Kingdom, and accepted by the Editorial Board July 9, 2012 (received for review May 11, 2012)

Sex determination in papaya is controlled by a recently evolved XY chromosome pair, with two slightly different Y chromosomes controlling the development of males (Y) and hermaphrodites (Y^h). To study the events of early sex chromosome evolution, we sequenced the hermaphrodite-specific region of the Y^h chromosome (HSY) and its X counterpart, yielding an 8.1-megabase (Mb) HSY pseudomolecule, and a 3.5-Mb sequence for the corresponding X region. The HSY is larger than the X region, mostly due to retrotransposon insertions. The papaya HSY differs from the X region by two large-scale inversions, the first of which likely caused the recombination suppression between the X and Y^h chromosomes, followed by numerous additional chromosomal rearrangements. Altogether, including the X and/or HSY regions, 124 transcription units were annotated, including 50 functional pairs present in both the X and HSY. Ten HSY genes had functional homologs elsewhere in the papaya autosomal regions, suggesting movement of genes onto the HSY, whereas the X region had none. Sequence divergence between 70 transcripts shared by the X and HSY revealed two evolutionary strata in the X chromosome, corresponding to the two inversions on the HSY, the older of which evolved about 7.0 million years ago. Gene content differences between the HSY and X are greatest in the older stratum, whereas the gene content and order of the collinear regions are identical. Our findings support theoretical models of early sex chromosome evolution.

Carica papaya | DNA sequencing | molecular evolution | sex chromosomes

Sex chromosomes have evolved independently in diverse lineages of animals and plants, and new dioecious species are still evolving (1, 2). Evidence of homology between nascent sex chromosome pairs in flowering plants and fish (3–6) supports the notion that sex chromosomes evolved from autosomes that gained sex determination genes. The key event in sex chromosome evolution is the suppression of recombination between the sex-determining regions of ancestrally homologous chromosome pairs, which limits one chromosome of the pair to one sex, producing XY (male heterogametic) or ZW (female heterogametic) systems. Evolutionary models predict that a lack of recombination allows for Y- or W-specific characteristics to accumulate, through the reduced efficacy of selection on these chromosomes (7, 8), leading to the Y and W chromosomes accumulating deleterious mutations and transposable elements, and ultimately undergoing genetic degeneration, through the loss of genes or gene functions, as observed in mammals, *Drosophila*, birds, fishes, and snakes (9, 10). In some animals and plants, the greater number of mitotic cell divisions in spermatogenesis than oogenesis also leads to Y chromosomes having a higher mutation rate than autosomes or X chromosomes (11–14) and is predicted to further contribute to greater changes of the evolving Y (or W) chromosome than the X (or Z) chromosome.

To test these predictions of repetitive sequence accumulation, chromosomal rearrangements, gene movement, gene loss, and pseudogenization on the Y versus the X, complete sequences of the sex determining regions are needed. To date, complete sequences of the sex-determining regions of both sex chromosomes are scarce, and complete sequences of one sex chromosome are only available in a handful of species. Human sex chromosomes have complete X and Y sex chromosome sequences and have been thoroughly studied and compared with the X and Y sequences available for the rhesus monkey and chimpanzee (15–20); the chicken Z and *Marchantia polymorpha* Y chromosomes have been completely sequenced (21, 22), but not their W and X chromosomes. Draft Y chromosome sequences are available in *Drosophila*, but not complete sequences (23), although large regions of the recently evolved *Drosophila miranda* neoY have been compared with the homolog (24).

From these sex chromosome sequences, the results of evolution in ancient sex chromosome systems are better understood. Mammalian sex chromosomes evolved about 166 million years ago (MYA) (25). Ninety-five percent of the human Y is a nonrecombining male-specific region (MSY), flanked by two physically small pseudoautosomal regions (15). The MSY has lost most of its gene content relative to the corresponding X chromosome region, which is estimated to have 1,098 genes; the MSY carries 78 protein-coding genes, encoding 27 different proteins, only 16 of which have X-linked homologs representing

Author contributions: Q.Y., M.A., P.H.M., J.J., A.H.P., and R.M. designed research; J.W., J.-K.N., Q.Y., A.R.G., J.H., F.Z., R.A., R.V., J.E.M., W.Z., R.N.-P., F.A.F., C.L., E.J.T., C.C., C.M.W., R.S., M.-L.W., X.J.M., J.J., and R.M. performed research; J.W., J.-K.N., Q.Y., A.R.G., J.H., F.Z., R.A., R.V., J.E.M., W.Z., R.N.-P., F.A.F., C.L., E.J.T., C.C., C.M.W., R.S., M.-L.W., X.J.M., M.A., D.C., P.H.M., J.J., A.H.P., and R.M. analyzed data; and J.W., A.R.G., D.C., and R.M. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. R.G.K. is a guest editor invited by the Editorial Board.

Database deposition: The sequences reported in this paper have been deposited in the GenBank database. See *SI Appendix*, Table S14 for accession nos.

[1]J.W., J.-K.N., Q.Y., and A.R.G. contributed equally to this work.

[2]Present address: Department of Agronomy, University of Florida, Gainesville, FL 32610.

[3]Present address: Department of Molecular Breeding, National Academy of Agricultural Science, Rural Development Administration, Suwon 441-701, Republic of Korea.

[4]Present address: Departamento de Genética, Facultad de Ciencias, Universidad de Granada Campus de Fuentenueva Sin Numero, 18071 Granada, Spain.

[5]Present address: Department of Genetics and Biochemistry, Clemson University, Clemson, SC 29634.

[6]Present address: Department of Agronomy, Shandong Agricultural University, Tai'an, Shandong 271018, China.

[7]To whom correspondence should be addressed. E-mail: rming@life.illinois.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1207833109/-/DCSupplemental.

"X-degenerated" descendents of genes in the ancestral chromosome (26, 27).

No comparisons have yet been possible between completely sequenced sex-determining regions of recently evolved sex chromosome systems. Dioecious plants are better suited than the ancient human or *Drosophila* sex chromosomes for studying the early stages of sex chromosome evolution. In *Silene latifolia*, Y-linked alleles have lower expression than their X homologs and an excess of (presumably often deleterious) amino acid substitutions in their coding sequences, showing genetic degeneration of the Y chromosome is occurring, as predicted (28, 29). *S. latifolia* has heteromorphic sex chromosomes estimated to have evolved 5–10 MYA (30) and are thus slightly older than those of papaya (see below). Gene loss during plant sex chromosome evolution has not yet been assessed, because sex-linked genes have been ascertained largely through genetic mapping of genes discovered from EST sequences; consequently, genes with extant Y-linked alleles are most readily discovered, potentially seriously underestimating the proportion of genes lost during the evolution of Y-linked regions. The *S. latifolia* results just mentioned (28, 29) are based on RNA-Seq, so even though they identify many sex-linked genes, genomic or physical maps of the sex chromosomes and unbiased estimates of gene losses from the Y-linked region of the sex chromosome pair cannot be obtained. Details of sex chromosome repetitive sequences are also unknown, although it is clear that the Y chromosome has a greater abundance than the rest of *S. latifolia*'s large genome (31).

Here we describe the complete sequencing of the papaya Y^h-specific region together with its X counterpart, allowing a comprehensive comparison of the gene content and repetitive sequence content of a plant X/Y system, providing a detailed view of the early evolutionary events of sex chromosome evolution in papaya. Papaya is a trioecious tropical fruit tree with three sex types (female, male, and hermaphrodite) controlled by an XY system. The Y chromosome determines male flower development, and the slightly different Y^h chromosome determines hermaphrodite flower development. DNA sequence divergence between these two Y chromosomes is 1.2% (32). The papaya hermaphrodite-specific region of the Y^h chromosome (HSY), which maps to the middle of chromosome 1, is about 10% of the chromosome's physical size, and is flanked by much larger pseudoautosomal regions (3). Four pairs of papaya X/Y^h genes spanning a region of about 5–6 Mb were previously studied: The Y^h was inferred to have stopped recombining with the X about 2–3 MYA, and no evolutionary strata were found (33). Our analysis of the recently completed sequence of the papaya HSY and its X counterpart now reveals evolutionary strata in the X chromosome, formed by inversions, causing recombination suppression of the sex-determining region. This study details the different contributors to the X–Y size difference and an estimate of gene loss since recombination ceased.

Results

Sequencing HSY and Its X Counterpart. We initiated construction of physical maps of the HSY and its X counterpart by screening bacterial artificial chromosome (BAC) clones with sex cosegregating markers. Positive clones were confirmed by PCR amplification and further verified by florescence in situ hybridization (FISH) to the Y^h or X chromosome (*SI Appendix*, Fig. S1). The BAC ends were then sequenced, and new probes were designed for chromosome walking out from the assembled regions. The two recombining pseudoautosomal regions, which form the bulk of the papaya sex chromosome pair, were sequenced in the draft papaya genome, which used a female SunUp plant, the same cultivar used in our study (34).

The HSY minimum tiling path consists of 68 overlapping BACs with one gap near border A (see below); the corresponding region is continuous in the X physical map. The physical map of the X region corresponding to the HSY region consists of 43 BACs in a minimum tiling path with one gap in the middle, which is contiguous in the HSY physical map (35). Sequencing the highly repetitive HSY and its X counterpart is challenging even using a BAC-by-BAC approach. Each BAC clone in the minimum tiling paths was therefore Sanger sequenced using shotgun libraries at 8- to 20-fold coverage (depending on the complexity of the sequences). After quality trimming, the reads were assembled into contiguous sequences, with manual correction of possible base-call errors (see *Methods*), and joined to form pseudomolecules. The HSY pseudomolecule included a total of 8.1 Mb in 15 contigs. The X pseudomolecule of 5.4 Mb in 12 contigs includes a 1.9-Mb Knob 1 region shared by the X and Y^h and a 3.5-Mb X-specific region (Fig. 1).

Defining the HSY Borders at the DNA Level. Borders A and B were defined genetically based on recombination events in 2,920 informative chromosomes from 1,460 F_2 individuals (35). The

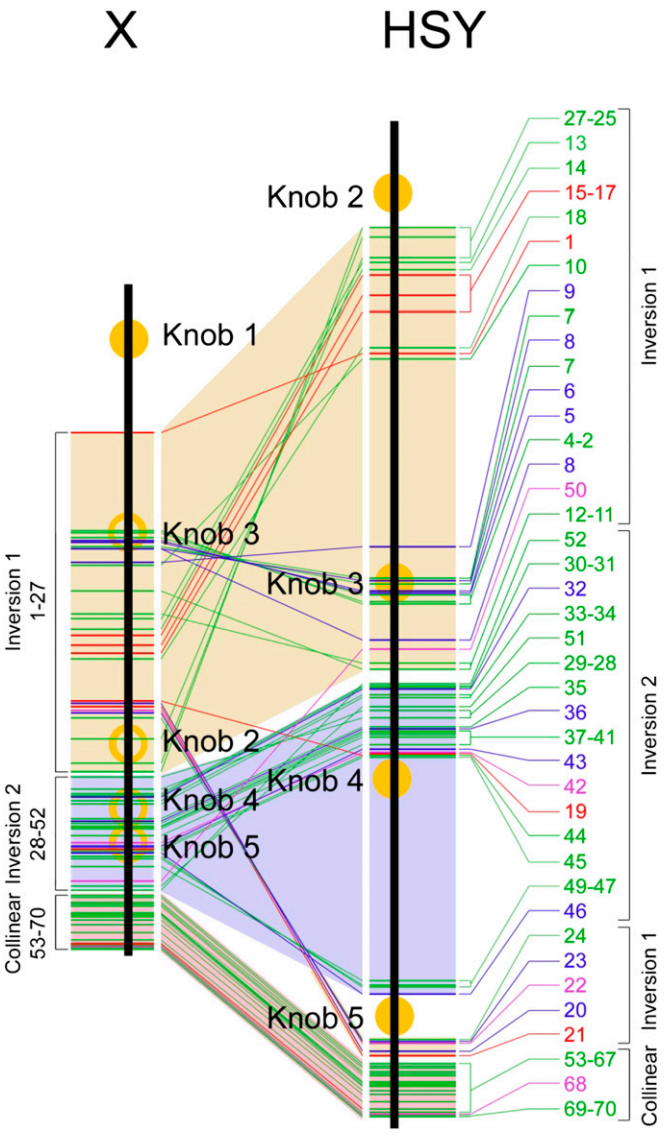

Fig. 1. Comparison between the arrangement of 70 genes on the HSY with homologous copies on the X (HSY-X gene pairs, see text). Each gene's location is indicated by a horizontal line. The vertical black lines indicate the HSY and X sequences, and the sequences encoding the transcripts are numbered according to their order in the X region. The solid yellow circles indicate four heterochromatic knobs (Knobs 2–5) in the HSY (30) and Knob 1 in the X. Knobs 2–5 are specific to the HSY, but their estimated corresponding positions in the X region are indicated by empty yellow circles. Green labels indicate that both transcript copies are functional genes, blue labels indicate that the HSY transcript is a pseudogene, red labels indicate that the X transcript is a pseudogene, and purple labels indicate that both HSY and X copies are pseudogenes. The two inverted regions and the collinear region are marked. Inversion 1 in the HSY was split into two regions due to translocation.

sequence in the nonrecombining HSY region is expected to differ from that carried on the X chromosome, whereas the flanking recombining pseudoautosomal regions should show high homology. The genetically defined border A is adjacent to a heterochromatic knob (Knob 1 in Fig. 1), which is indeed present on both the X and Y^h chromosomes at the same location (36). In the HSY physical map, but not that of the X, there is a gap between border A and Knob 2; Knob 1 could not be mapped in the HSY due to high sequence similarity with Knob 1 in the X (35).

HSY border B was identified on both the HSY and X physical maps. BACs SH85C03 and SH86B15 are immediately adjacent to the genetically defined border (*SI Appendix*, Fig. S2*A*; ref. 35). In BAC SH85C03, a 90-kb sequence overlaps and is identical with X BAC SH30B21, showing that SH85C03 (and also SH86B15) is in the X chromosome. A BAC clone, SH60M19, which has an insert size of 252 kb, includes an identical overlap of 20.5 kb with HSY BAC 58C24 (*SI Appendix*, Fig. S2*A*). SH60M19 aligns with part of the X BAC SH85C03 and the entire X BAC SH86B15, extending 7,110 bp beyond its end. Across the 182-kb region of the X BAC SH86B15 that aligns with HSY sequence, the numbers of single nucleotide polymorphisms (SNPs) and insertion/deletion indels between the X and Y^h decline over the first 80 kb from the genetic border, and the remaining 102 kb furthest from the genetic border includes only six SNPs and six single-base indels, close to the sequencing error of 1 per 10^5 nucleotides (*SI Appendix*, Fig. S2*B*). The molecular HSY border B, where sequence homology with the X region becomes high, lies about 277 kb beyond the genetically defined border; this 277-kb region makes up about 50% of the collinear region and is still recombining, as seen by the fine mapping carried out by Na et al. (35).

Repetitive Sequences and Intrachromosomal Duplications. The 8.1-Mb HSY is more than twice as large as the corresponding 3.5-Mb X region. The difference is mainly due to the repetitive sequence content, which represents 79.3% of the HSY sequence and 67.2% of its X counterpart; both are much higher than the papaya-genome-wide average of 51.9% (*SI Appendix*, Fig. S3 and Table S1; refs. 34 and 37). *Ty3-gypsy* elements are the most abundant repeats in both the HSY and X.

Consistent with HSY expansion, in the overlapping regions, we identified 20 HSY-specific repeat units, which were not found in the publicly available repeat databases, (10.7% of its length, see *SI Appendix*, Table S2) versus only a single X-specific repetitive sequence (3.5% of its sequence). The HSY repeats mostly accumulated in two regions, one estimated to be between 2.3 and 3.8 Mb (on the x-axis of *SI Appendix*, Fig. S4) and the other at 5.5–7.8 Mb. We also scanned the sequences for tandem repeats and estimate that these form 3.1% of the X-region sequences and 3.8% of the HSY (*SI Appendix*, Fig. S5 and Table S3).

Large-scale homologous comparisons of unmasked sequences between the X and HSY revealed that two blocks (1.7 and 2.4 Mb) account for a great portion of the larger physical size of the HSY compared with the X region. These regions are highly repetitive, with 84.6% and 87.2% repetitive sequences, respectively, and correspond to heterochromatic Knob 3 in the first block and Knobs 4 and 5 in the second block (36), forming two large gaps between HSY gene islands (*SI Appendix*, Fig. S6). HSY-specific repeats are enriched in these blocks, of which most are *Ty3-gypsy* elements (*SI Appendix*, Table S4).

Intrachromosomal similarity analysis of the HSY pseudomolecule revealed that 96 sequences (a combined length of 315.9 kb, or 3.9% of the total) are direct (noninverted) duplications, with fragment lengths ranging from 1 kb (the cutoff used) to 25 kb (*SI Appendix*, Fig. S7). Sequence divergence between the duplicates is 0–2%, and the two members are located from 6 kb to 5.6 Mb apart within the HSY. In addition, 61 sequence fragments are inverted duplications with 0–2% sequence divergence (cumulatively about 157 kb of the HSY sequence, or 1.9% of the total) (*SI Appendix*, Fig. S7). Altogether, the 157 pairs of duplicated segments in the HSY constitute a length of 473 kb, 5.8% of the HSY sequence. In contrast, the 3.5 Mb of X counterpart sequence contains only 8 duplications (totaling 17.3 kb, or 0.5% of the X region sequence); 6 are direct duplications (11.8 kb total length, 0.3% of the X sequence), and 2 are inverted (5.5 kb, or 0.2% of the X sequence) (*SI Appendix*, Fig. S7).

Without an outgroup, we cannot definitively infer which changes occurred in the X and which changes occurred in the HSY, but the fact that the HSY differs most from other genome regions suggests that it has expanded. The companion article to our study, with analyses of a portion of the fully sex-linked region of the X chromosome and an outgroup (38), supports this conclusion.

Two Evolutionary Strata in the X Corresponding to Two Inversions in the HSY. There are 70 transcription units that are alignable between the X and HSY; these units are found in three subregions (Fig. 1 and *SI Appendix*, Table S5): The first HSY region, with coding sequence pairs 1–27, is inverted with respect to the X order; pairs 28–52 are inverted and rearranged in a second HSY region; and pairs 53–70 (region 3) are in a collinear region next to border B of the HSY and X regions.

To estimate the time since recombination stopped between the HSY and X across the sex-specific region, we analyzed divergence for 70 X/HSY nonduplicated sequence pairs with known X physical map positions. We estimated the numbers of synonymous substitutions per synonymous site (K_s) and silent (synonymous or intron site) substitutions per silent site (K_{sil}) in these sequences, which reflect the divergence times of the gene pairs (*SI Appendix*, Tables S6 and S7). The values clearly reveal two distinct evolutionary strata. These strata correspond perfectly with the two inversions identified above (Fig. 2); divergence values for sequence pairs in inversions 1 and 2 differ

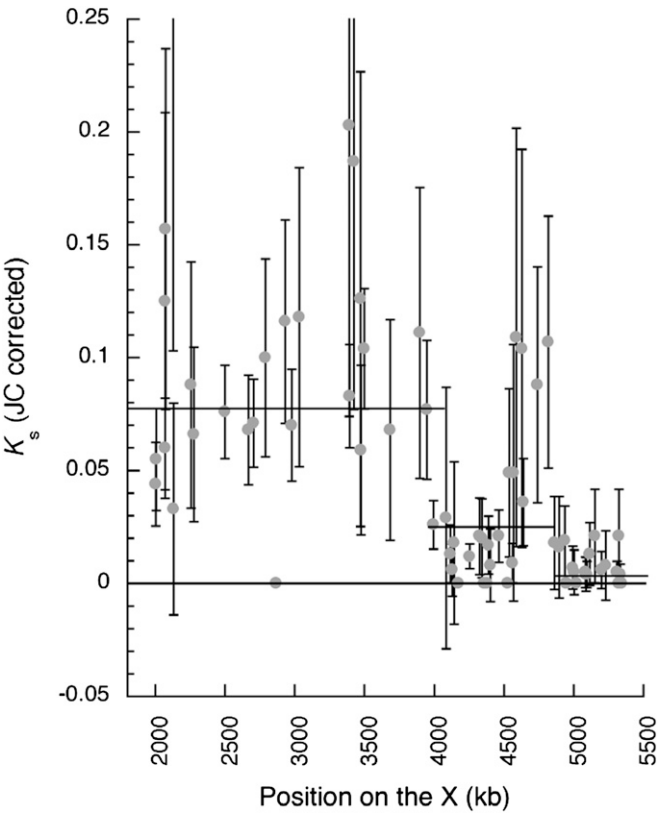

Fig. 2. Plot of synonymous site divergence (K_s) of 70 paired X/HSY genes arranged according to the location on the X chromosome. Horizontal lines indicate the mean K_s value of the genes found in each region (inversion 1, inversion 2, and collinear). Divergence values are significantly different between inversion 1, inverion 2, and the collinear region ($P < 0.01$). Gene pair names for each stratum are given in *SI Appendix*, Table S6.

GENETICS

significantly for either divergence measure (K_s or K_{sil} $P < 0.01$ by Mann–Whitney U tests, *SI Appendix*, Table S8).

The 18 intact gene pairs in the collinear region have significantly lower nucleotide divergence than the 25 gene pairs in stratum 2. The differences between inversion 2 and the collinear region probably reflect the fact that variants at sites closely linked to a site under balancing selection are expected to have variability elevated above that at unlinked sites (39, 40).

We estimated the divergence times for the gene pairs in the two strata by applying a molecular clock rate estimated for the closely related family Brassicaceae to our K_{sil} estimates, which yielded lower and upper bounds of 1.9–9.5 MY for stratum 1 (mean 7.0 MY) and 0–6.9 MY (mean 1.9 MY) for stratum 2. Because of the correspondence between the strata and the physical organization difference between the HSY and the X, we conclude that inversion 1 occurred first and that the HSY-specific region expanded when the second inversion occurred. HSY stratum 1 is therefore probably younger than the nonrecombining region of the Y chromosome in *S. latifolia*, where the oldest stratum formed about 10 MYA (30).

The gene order within the two inversions in the HSY is no longer simply the reversed X gene order; other intrachromosomal rearrangements have clearly occurred after the initial inversion events (Fig. 1 and *SI Appendix*, Figs. S8 and S9 and Table S5). Rearrangements involving regions that contain X/Y^h gene pairs are clearly detectable; indels and other rearrangements in noncoding sequences are too numerous to count, as shown by direct comparisons of paired X and HSY BACs (33). We estimate that at least nine rearrangements are required to reach the current gene order in the older inversion 1, including three inversions, four translocations, and two duplications. In the younger inversion 2, we detect at least seven chromosomal rearrangements, including four inversions, two translocations, and one inverted translocation. In the portion of the sex-specific region where an outgroup can be compared, the X is colinear with the outgroup (38), suggesting that the rearrangements occurred in the HSY.

Gene Contents of the X Region and Its HSY Counterpart. After several rounds of gene prediction and validation by RT-PCR, 72 protein-coding genes were annotated in the HSY and 84 protein-coding genes were annotated in its X counterpart, showing lower gene content in the HSY (Table 1). There are 106 combined protein-coding genes in the HSY and X, including 50 X/HSY paired genes, 22 HSY-specific genes, and 34 X-specific genes (*SI Appendix*, Tables S9, S10, and S11). The average gene density is one gene per 112.5 kb in the HSY, much lower than the average of one gene per 43.2 kb in the X counterpart, which is still lower than the genome-wide average of one gene per 16 kb (34).

We classified coding sequences with premature stop codons, frame shift mutations, or truncated proteins as pseudogenes. Of the 96 transcription units annotated in the HSY, 72 are protein-coding genes as described above and 24 (25%) are pseudogenes (Table 1 and *SI Appendix*, Fig. S10*A* and Table S12). In the X counterpart, 84 of the 98 are protein-coding genes and 14 (14%) are pseudogenes (*SI Appendix*, Fig. S10*B*). Four pseudogene pairs are found in both the X and the HSY. There are 124 combined transcription units, including 70 X/HSY paired, 26 HSY-specific, and 28 X-specific units. The 70 paired transcription units include 50 paired genes, 10 X genes paired with HSY pseudogenes, 6 HSY genes paired with X pseudogenes, and 4 paired X/HSY pseudogenes. There are 28 X-specific transcription units, but 34 X-specific genes. This appears not to make sense at the first glance; the cause of this discrepancy is the 10 X-gene–HSY pseudogene pairs. There are 24 X-specific genes without any homologous sequences (genes or pseudogenes) and 4 X-specific pseudogenes. For the same reason, the 26 HSY-specific transcription units include 16 (not 22) HSY-specific genes without homologous sequences and 10 HSY-specific pseudogenes.

What proportion of genes present in the ancestral chromosome region have copies whose functions have been lost since recombination stopped? Although a slightly higher proportion of the HSY than X sequences are pseudogenes (25% versus 14%), the difference is nonsignificant. The sequences classified as pseudogenes have intact sequences, suggesting that their functions were only recently lost. HSY–X divergence per nonsynonymous site (K_a) in nonpseudogene sequences is considerably lower than their K_{sil} or synonymous divergence, K_s. The K_a/K_s average is 0.40 for the 27 genes in the older inversion, similar to values in the other two regions (*SI Appendix*, Table S6).

When a gene copy exists in the HSY and also in an autosome, but not the X, gain from the autosomes (i.e., movement from a heterologous chromosome) seems more likely than loss from the X. Using this reasoning, we estimated that the HSY acquired 10 transcripts from autosomes, but the X region acquired none (Table 1 and *SI Appendix*, Fig. S10 *A* and *B*). Of the 124 transcription units annotated in either the HSY or the X region, 114 were therefore probably present in the ancestral chromosome. Although the ancestral gene number is not estimated definitively, these values suggest somewhat greater gene and gene function loss from the HSY than the X.

The gene content differs to a surprising degree between the HSY and X region. For example, 28 transcript-encoding sequences (corresponding to 24 X genes and 4 pseudogenes) are X-specific, with no homologs on either the HSY or the autosomes, whereas 16 transcripts (9 HSY genes and 7 pseudogenes) are HSY-specific and are absent from the X and autosomes (Table 1 and *SI Appendix*, Table S12). In total, the HSY has no functional copies of 52 genes (the 28 X-specific genes and pseudogenes plus 24 HSY pseudogenes), whereas the X region lacks functional copies of 30 genes (14 X pseudogenes plus the 16 HSY-specific genes); this difference is significant ($P = 0.014$ by a two-tailed Fisher's exact test).

Most of the HSY- and X-specific genes are in the older inversion 1, which includes 20 of 22 (91%) of the HSY-specific protein coding genes (Table 1 and *SI Appendix*, Table S13). Seven of the 22 genes, six of which are in inversion 1, match paralogs in the autosomes, suggesting possible gains by the HSY. Nine genes that have no matching sequences in the X counterpart or autosomes, are all in inversion 1; these could either represent additions to the HSY or losses from the X. The remaining six HSY-specific genes have homologous X-pseudogenes: four in inversion 1, one in inversion 2, and one in the collinear region. The X-specific genes are also mostly in the older stratum (26 of 34 of the X-specific protein-coding genes = 76%), whereas X/Y^h sequence pairs are rarest in this region (14 of 50 = 28%). In contrast, the region inverted later (inversion 2) contains only one (5%) HSY-specific and eight (24%) X-specific genes. The collinear region contains one Y-specific gene (5%) and none that are specific to the X. Of the four HSY/X pseudogene pairs, one is also in inversion 1.

Discussion

Sequencing the papaya HSY and its X counterpart provides a complete picture of the sex-specific regions of a plant sex chromosome pair and supports theoretical models of the early stages of sex chromosome evolution (41, 42). First, our results directly indicate that inversions in the HSY region caused recombination suppression with the X, initiating sex chromosome evolution and permitting numerous intrachromosomal rearrangements that are predicted to become fixed on the Y after recombination ceases (43). Rearrangements were even detected in the recent inversion that occurred only 1.9 MYA. The high number of rearrangements suggests that they may have involved ectopic recombination events between similar transposable element sequences. The involvement of two inversions in the suppression of recombination, resulting in two strata, contributes to the generality of "evolutionary strata" observed in mammals, birds, and the plant *S. latifolia* (30, 44–46), supporting the notion that suppressed recombination between sex chromosomes generally evolves in multiple events (42). Our results reveal that the inversions are not the sole contributor to suppression of recombination. We found that about half of the collinear region has stopped recombining. This finding is consistent with the involvement of sexually antagonistic mutations, which benefit one sex but harm the other, as predicted by sex-chromosome evolution models (47, 48). If such mutations occur at loci closely

Table 1. Summary of the transcription units in the HSY and corresponding X chromosomal region

Types of sequences	Total no. of sequences	
	HSY	X
Combined transcription units in HSY and X	124	
Total transcription units	96	98
Transcription unit pairs	70	
HSY- or X-specific transcription units*	26	28
Combined intact protein-coding transcription units	106	
Intact protein-coding transcription units	72	84
Genes present in the HSY and X	50	
Copy present elsewhere in the papaya whole genome	2	
Genes specific to the HSY or X[†]	22	34
Protein-coding transcripts specific to the HSY or X	16	24
Copy present elsewhere in the papaya whole genome	7	0
Sequences of unknown origin	9	24
Protein coding genes in one region and pseudogenes in the other[‡]	6	10
Combined pseudogene transcription units	34	
Total pseudogene transcription units	24	14
Transcripts that are pseudogenes	4	
Combined pseudogene–gene pairs	16	
Pseudogenes in one region and protein coding genes in the other[‡]	10ψ	6ψ
Pseudogenes specific to the HSY or X	10	4
Copy present elsewhere in the papaya whole genome	3	0
Sequence of unknown origin	7	4

*HSY- or X-specific transcription units include HSY- and X-specific protein coding transcripts (16 and 24 transcripts, respectively) and HSY- and X-specific pseudogenes (10 and 4 transcripts, respectively).

[†]Genes specific to the HSY or X include 16 HSY-specific and 24 X-specific protein coding transcripts as well as the 6 and 10 paired gene–pseudogene transcripts, respectively.

[‡]These pseudogenes are included twice in the table, once in the "Protein coding genes in one region and pseudogenes in the other" category and once in the "Pseudogenes in one region and protein coding genes in the other" category. There are 6 transcripts that are functional genes on the HSY and pseudogenes on the X and 10 transcripts that are functional genes on the X and pseudogenes on the HSY. They are only counted once in the combined transcript unit total.

linked to a male-specific region and establish a polymorphism, this leads to selection for suppressed recombination (49).

Our results also support the predicted early accumulation of transposable elements in the initial stages of sex-chromosome evolution after recombination stops, leading to the physical expansion of the sex-determining regions (42). 80.2% of the younger inversion 2 sequence is already repetitive, as is at least 80.7% of inversion 1 (probably an underestimate, because elements degenerate after insertions and become unrecognizable). The physical size of the younger inversion 2 in the HSY is more than double that of its X chromosome counterpart, due to accumulation of repetitive sequences since the suppression of recombination (80.2% in HSY vs. 60.5% in the X, Fig. 1). However, the older inversion 1 region is similar in size in the HSY and X chromosome regions, and they have similar repetitive sequence content (80.7% and 76.5%, respectively). The papaya X counterpart also has a higher repetitive content than the genome-wide average, as has been found in other organisms (50). The lower gene density and higher repetitive sequences in the X region corresponding to inversion 1 suggest that the X chromosome might be expanding (see the companion paper, ref. 38). In addition, the X chromosomes recombine in females, but the X region does not recombine with the Y or Y^h in males or hermaphrodites. To a lesser extent than the Y, the X region will thus have a lower effective population size than the autosomes and a reduced efficacy of purifying selection; this may explain the higher level of repetitive sequences in the X region corresponding to the older inversion 1 compared with the inversion 2 (*SI Appendix*, Table S12; ref. 38). The X expansion is further supported by comparing part of the X with the orthologous autosome in *Vasconcellea monoica*, a related monoecious species that does not have sex chromosomes (38).

The HSY has fewer functional genes than the X region. Although it is possible that genes have been added to the X chromosome (e.g., perhaps mediated through retrotransposon activity), we found no clear examples of this and so we infer that the differences in gene content between the HSY and the X region mostly represent the loss or pseudogenization of genes present on the ancestral chromosome. Gene loss and pseudogenization, along with transposable element accumulation, are consistent with predictions of the beginning of Y degeneration in the early stages of sex-chromosome evolution (42).

Methods

Gene Annotation. The repeat-masked HSY and X sequences were blasted to the papaya EST and gene model databases, and EST datasets of *Medicago truncatula*, *Oryza sativa*, *Populus trichocarpa*, *Vitis vinifera*, and *Arabidopsis thaliana*, using tblastx for transcription unit identification. Two gene-prediction programs, Genscan and Fgenesh, were also used to predict additional genes that may have been missed by the former approach. Each predicted transcript was translated in all six reading frames to distinguish protein-coding ones from pseudogenes. Potential functions of protein-coding transcripts were predicted using conserved domains and homologous gene functions.

BAC Clone Sequencing and Assembly. Individual BAC clones were sequenced using Sanger sequencing method. *Escherichia coli* genomic sequences, vector sequences, low quality sequences, and ambiguous sequences were removed from the Sanger reads. The trimmed sequences were assembled using Phred/Phrap/Consed (51, 52), CAP3 (53) packages, and Sequencher software (Gene Codes). GenBank accession numbers of sequenced BACs are listed in *SI Appendix*, Table S14.

Repetitive Sequence 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, TIGR plant repeats (http://plantrepeats.plantbiology.msu.edu/index.html), and papaya repeats (37). Sex-specific repeats in the HSX and X were identified using RepeatScout-V1. Nonredundant repeats longer than 100 bp were compared with papaya repeats using CD-HIT software (54). Tandem repeats were analyzed using the Tandem Repeats Finder software (55). The schematics were drawn using DomainDraw (56).

Inversion Analysis. Concatenated repeat-masked pseudomolecules of HSY and its X counterpart were aligned using Mauve, a genome alignment tool, using the default settings (http://asap.ahabs.wisc.edu/mauve/index.php) (57). Different local collinear block (LCB) weights (nucleotide alignment scores) ranging from 1 to 300,000 were applied to visualize the sequence rearrangements that differ between the HSY and the X regions.

Intrachromosomal Duplication. Concatenated repeat-masked pseudomolecules of HSY and the corresponding X region were blasted against each other to identify intrachromosomal duplications, using a cutoff of 98% sequence identity in alignments of 1,000 bp.

Sequence Divergence Analysis of Coding Sequences. The X and HSY transcript pairs were sequentially aligned using BioEdit. Exon and intron junctions of the aligned transcript pairs were designated based on the predicted gene structure. The substitution rates were estimated for synonymous (K_s), nonsynonymous (K_a), and silent (K_{sil}) sites by using the method of Nei and Gojobori (58) implemented in DnaSP v5. Divergence times for the transcript pairs were determined according to the methods described by Li (59), using the synonymous substitution rate of 4×10^{-9} substitutions per synonymous site per year determined for *Arabidopsis* relatives (60).

ACKNOWLEDGMENTS. We thank Andrew Wood and Andrea Blas for technical assistance and Judith Mank for constructive comments and suggestions. This work was supported by National Science Foundation (NSF) Plant Genome Research Program Awards DBI0553417 and DBI-0922545 (to R.M., Q.Y., P.H.M., J.J., and A.H.P.) and a grant from the University of Illinois Research Board.

1. Graves JAM, Shetty S (2001) Sex from W to Z: Evolution of vertebrate sex chromosomes and sex determining genes. *J Exp Zool* 290:449–462.
2. Ming R, Bendahmane A, Renner SS (2011) Sex chromosomes in land plants. *Annu Rev Plant Biol* 62:485–514.
3. Liu Z, et al. (2004) A primitive Y chromosome in papaya marks incipient sex chromosome evolution. *Nature* 427:348–352.
4. Peichel CL, et al. (2004) The master sex-determination locus in threespine sticklebacks is on a nascent Y chromosome. *Curr Biol* 14:1416–1424.
5. Spigler RB, Lewers KS, Main DS, Ashman TL (2008) Genetic mapping of sex determination in a wild strawberry, *Fragaria virginiana*, reveals earliest form of sex chromosome. *Heredity (Edinb)* 101:507–517.
6. Yin T, et al. (2008) Genome structure and emerging evidence of an incipient sex chromosome in *Populus*. *Genome Res* 18:422–430.
7. Bachtrog D, Charlesworth B (2002) Reduced adaptation of a non-recombining neo-Y chromosome. *Nature* 416:323–326.
8. Bachtrog D (2003) Adaptation shapes patterns of genome evolution on sexual and asexual chromosomes in *Drosophila*. *Nat Genet* 34:215–219.
9. Charlesworth B, Charlesworth D (2000) The degeneration of Y chromosomes. *Philos Trans R Soc Lond B Biol Sci* 355:1563–1572.
10. Bachtrog D (2008) The temporal dynamics of processes underlying Y chromosome degeneration. *Genetics* 179:1513–1525.
11. Haldane JBS (1935) The rate of spontaneous mutation of a human gene. *J Genet* 31:317–326.
12. Li WH, Yi S, Makova K (2002) Male-driven evolution. *Curr Opin Genet Dev* 12:650–656.
13. Filatov DA, Charlesworth D (2002) Substitution rates in the X- and Y-linked genes of the plants, *Silene latifolia* and *S. dioica*. *Mol Biol Evol* 19:898–907.
14. Qiu S, Bergero R, Forrest A, Kaiser VB, Charlesworth D (2010) Nucleotide diversity in Silene latifolia autosomal and sex-linked genes. *Proc Roy Soc B-Biol Sci* 277:3283–3290.
15. Skaletsky H, et al. (2003) The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* 423:825–837.
16. Ross MT, et al. (2005) The DNA sequence of the human X chromosome. *Nature* 434:325–337.
17. Mikkelson TS, et al.; Chimpanzee Sequencing and Analysis Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437:69–87.
18. Gibbs RA, et al. Rhesus Macaque Genome Sequencing and Analysis Consortium (2007) Evolutionary and biomedical insights from the rhesus macaque genome. *Science* 316: 222–234.
19. Hughes JF, et al. (2010) Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content. *Nature* 463:536–539.
20. Hughes JF, et al. (2012) Strict evolutionary conservation followed rapid gene loss on human and rhesus Y chromosomes. *Nature* 483:82–86.
21. Bellott DW, et al. (2010) Convergent evolution of chicken Z and human X chromosomes by expansion and gene acquisition. *Nature* 466:612–616.
22. Yamato KT, et al. (2007) Gene organization of the liverwort Y chromosome reveals distinct sex chromosome evolution in a haploid system. *Proc Natl Acad Sci USA* 104:6472–6477.
23. Koerich LB, Wang X, Clark AG, Carvalho AB (2008) Low conservation of gene content in the *Drosophila* Y chromosome. *Nature* 456:949–951.
24. Bachtrog D, Hom E, Wong KM, Maside X, de Jong P (2008) Genomic degradation of a young Y chromosome in *Drosophila miranda*. *Genome Biol* 9:R30.
25. Veyrunes F, et al. (2008) Bird-like sex chromosomes of platypus imply recent origin of mammal sex chromosomes. *Genome Res* 18:965–973.
26. Delbridge ML, Lingenfelter PA, Disteche CM, Graves JAM (1999) The candidate spermatogenesis gene RBMY has a homologue on the human X chromosome. *Nat Genet* 22:223–224.
27. Bhowmick BK, Satta Y, Takahata N (2007) The origin and evolution of human ampliconic gene families and ampliconic structure. *Genome Res* 17:441–450.
28. Bergero R, Charlesworth D (2011) Preservation of the Y transcriptome in a 10-million-year-old plant sex chromosome system. *Curr Biol* 21:1470–1474.
29. Chibalina MV, Filatov DA (2011) Plant Y chromosome degeneration is retarded by haploid purifying selection. *Curr Biol* 21:1475–1479.
30. Bergero R, Forrest A, Kamau E, Charlesworth D (2007) Evolutionary strata on the X chromosomes of the dioecious plant *Silene latifolia*: Evidence from new sex-linked genes. *Genetics* 175:1945–1954.
31. Kejnovsky E, Hobza R, Cermak T, Kubat Z, Vyskot B (2009) The role of repetitive DNA in structure and evolution of sex chromosomes in plants. *Heredity (Edinb)* 102:533–541.
32. Yu Q, et al. (2008) Recent origin of dioecious and gynodioecious Y chromosomes in papaya. *Trop Plant Biol* 1:49–57.
33. Yu Q, et al. (2008) Low X/Y divergence in four pairs of papaya sex-linked genes. *Plant J* 53:124–132.
34. Ming R, et al. (2008) The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature* 452:991–996.
35. Na J-K, et al. (2012) Construction of physical maps for the sex-specific regions of papaya sex chromosomes. *BMC Genomics* 13:176.
36. 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.
37. Nagarajan N, et al. (2008) Genome-wide analysis of repetitive elements in papaya. *Trop Plant Biol* 1:191–201.
38. Gschwend AR, et al. (2012) Rapid divergence and expansion of the X chromosome in papaya. *Proc Natl Acad Sci USA*, 10.1073/pnas.1121096109.
39. Charlesworth B, Nordborg M, Charlesworth D (1997) The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. *Genet Res* 70:155–174.
40. Kirkpatrick M, Guerrero RF, Scarpino SV (2010) Patterns of neutral genetic variation on recombining sex chromosomes. *Genetics* 184:1141–1152.
41. Charlesworth B (1991) The evolution of sex chromosomes. *Science* 251:1030–1033.
42. Charlesworth D, Charlesworth B, Marais G (2005) Steps in the evolution of heteromorphic sex chromosomes. *Heredity (Edinb)* 95:118–128.
43. Charlesworth B, Langley CH (1989) The population genetics of Drosophila transposable elements. *Annu Rev Genet* 23:251–287.
44. Lahn BT, Page DC (1999) Four evolutionary strata on the human X chromosome. *Science* 286:964–967.
45. Sandstedt SA, Tucker PK (2004) Evolutionary strata on the mouse X chromosome correspond to strata on the human X chromosome. *Genome Res* 14:267–272.
46. Handley L-JL, Ceplitis H, Ellegren H (2004) Evolutionary strata on the chicken Z chromosome: implications for sex chromosome evolution. *Genetics* 167:367–376.
47. Charlesworth D, Charlesworth B (1980) Sex differences in fitness and selection for centric fusions between sex-chromosomes and autosomes. *Genet Res* 35:205–214.
48. Rice WR (1987) The accumulation of sexually antagonistic genes as a selective agent promoting the evolution of reduced recombination between primitive sex-chromosomes. *Evolution* 41:911–914.
49. Ellis N, Yen P, Neiswanger K, Shapiro LJ, Goodfellow PN (1990) Evolution of the pseudoautosomal boundary in Old World monkeys and great apes. *Cell* 63:977–986.
50. Bergero R, Forrest A, Charlesworth D (2008) Active miniature transposons from a plant genome and its nonrecombining Y chromosome. *Genetics* 178:1085–1092.
51. Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using *phred*. I. Accuracy assessment. *Genome Res* 8:175–185.
52. Gordon D, Abajian C, Green P (1998) *Consed*: A graphical tool for sequence finishing. *Genome Res* 8:195–202.
53. Huang X, Madan A (1999) CAP3: A DNA sequence assembly program. *Genome Res* 9: 868–877.
54. 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.
55. Benson G (1999) Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res* 27:573–580.
56. Fink JL, Hamilton N (2007) DomainDraw: A macromolecular schematic drawing program. *In Silico Biol* 7:14.
57. Darling AC, Mau B, Blattner FR, Perna NT (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res* 14:1394–1403.
58. Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* 3:418–426.
59. Li WH (1997) *Molecular Evolution* (Sinauer Associates, Sunderland, MA).
60. Beilstein MA, Nagalingum NS, Clements MD, Manchester SR, Mathews S (2010) Dated molecular phylogenies indicate a Miocene origin for *Arabidopsis thaliana*. *Proc Natl Acad Sci USA* 107:18724–18728.