Supporting Information Appendix

Radical remodeling of the Y chromosome in a recent radiation of malaria mosquitoes

Andrew B. Hall, Philippos-Aris Papathanos, Atashi Sharma, Changde Cheng, Omar S. Akbari, Lauren Assour, Nicholas H. Bergman, Alessia Cagnetti, Andrea Crisanti, Tania Dottorini, Elisa Fiorentini, Roberto Galizi, Jonathan Hnath, Xiaofang Jiang, Sergey Koren, Tony Nolan, Diana Radune, Maria V. Sharakhova, Aaron Steele, Vladimir A. Timoshevskiy, Nikolai Windbichler, Simo Zhang, Matthew W. Hahn, Adam M. Phillippy, Scott J. Emrich, Igor V. Sharakhov*, Zhijian Tu*, Nora J. Besansky*

*correspondence to: igor@vt.edu; jake@vt.edu; nbesansk@nd.edu

Supporting Information Appendix

Table of Contents

S1. Genomic and transcriptomic datasets

S1.1. PacBio WGS Sequencing

Twenty adult male offspring of a single pair mating from the *An. gambiae* Pimperena laboratory colony were collected and genomic DNA was extracted. Genomic DNA was quantified by spectrophotometry (Quant-iT PicoGreen, Life Technologies). A library was prepared from this genomic DNA using the Pacific Biosciences recommended protocols and P4-C2 reagents. The resulting sample was sequenced with 58 cells on a Pacific Biosciences RS instrument with 180-minute runtimes. Short insert reads that qualify as circular consensus sequence (CCS) reads were extracted. Further PacBio sequencing was performed with newer P5-C3 chemistry and library preparation techniques. The same genomic DNA from above was sheared by Covaris g-TUBE and a BluePippin instrument was used to select fragments from 15-50 kb in length. A large insert library was prepared using Pacific Biosciences recommended protocols and P5-C3 reagents. The resulting sample was sequenced with 11 cells on a Pacific Biosciences RS sequencing instrument with 180-minute runtimes. The PacBio reads have been submitted to the SRA: SRS667972 (SRX668744 - SRX668812).

S1.2. PacBio WGS error-correction and assembly

A combination of P4-C2 (58 cells) and P5-C3 (11 cells) sequencing produced ~70X and ~35X coverage of autosomes and heterosomes, respectively, with an average read length of 2,479 bp, N50 length of 3,134 bp, and a maximum length of 31,875 bp.

Two sets of corrected reads were produced. The first set includes only the P4-C2 sequences corrected with the Celera Assembler 8.0 PBcR pipeline (1, 2) using BLASR

(3) for overlap detection and AMOS make-consensus (4) for correction. To maximize total bases for Ydb, no quality trimming was performed on the corrected reads. This resulted in a total of 40X corrected sequences with an average read length of 2,449 bp, N50 length of 2,799 bp, and maximum length of 19,353 bp. This set was used to build Ydb (Other Supporting File 1).

The second set of corrected reads includes the longer, but potentially more error-prone P5-C3 sequences. For *de novo* assembly, the combined P4-C2 and P5-C3 reads were corrected with the Celera Assembler 8.1 PBcR pipeline (1, 2). All data was aligned with BLASR (3) to the longest 50X subset (based on a 300 Mb genome size) and consensus correction was performed using PBDAGCON (5). The command to correct the sequences was:

pacBioToCA -length 500 -s pacbio.spec -t 32 -partitions 200 -pbCNS -l pbcns -fastq filtered_subreads.fastq QV=52.5

A total of 25X corrected reads were collected for assembling with an average read length of 2,433 bp, N50 length of 3,161 bp, and a maximum length of 30,325 bp. These sequences were assembled using relaxed parameters to merge haplotypes. The command to subset and assemble the sequences was:

runCA -s asm.spec -p asm -d asm pbcns.frg

Finally, all raw data were used to re-call a final consensus using Quiver (5) by importing the assembly into SMRTportal v2.1.0 and running the re-sequencing protocol for a diploid genome. Assembled contigs with less than 20 sequences or 2X coverage after Quiver were discarded, resulting in an assembly of 313.7 Mb in 8,322 contigs with a maximum contig length of 4,895,868 bp and an N50 length of 110,347 bp. The assembly was screened for bacterial contamination using Kraken v0.10.4-beta (6). A custom database was built, including *Homo sapiens* (GRCh37), *An. gambiae* Pimperena strain (ABKQ02), *An. gambiae* PEST strain (AAAB01), and NCBI Bacteria and Viruses (NCBI Genomes FTP, May 2014). All contigs were classified using the following commands:

```
kraken --db DB --threads 32 --fastq-input --output kraken.hits --preload pacbio.fastq
cat kraken.hits | kraken-filter --db DB --threshold 0 > kraken.filtered.hits
```

A total of 5% (~20 Mb) of contigs were classified as Bacterial/Viral and were removed from the final assembly. A contig associated with the mitochondria was identified by reference mapping and the overlapping region was deleted from the 3′ end to circularize it. The final assembly has 293.7 Mb in 8,000 contigs with a maximum contig length of 4,670,695 bp and an N50 length of 101,465 bp. The assembly has been submitted to NCBI as LCWJ00000000.

Assembled contigs were validated using Illumina sequences derived from male mosquitos of the same *An. gambiae* Pimperena colony. Illumina reads were aligned to the PacBio assembly using Bowtie (7) and variants (both SNPs and short indels) were called using FreeBayes (8). Variants supported by less than 50% of the mapped Illumina reads

were considered putative errors in the PacBio assembly. For all mapped bases with a minimum coverage of 3 Illumina reads, this technique yielded a PacBio consensus accuracy estimate of 99.97% (or average Phred base quality of Q35). Bowtie was run with the parameters (-I 200 -X 800 -p 32 -f -l 25 -e 140 --best --chunkmbs 256 -k 1 -S --un=unaligned.fasta -q asm-ctg -1 <run>.1.fastq -2 <run>.2.fastq) and FreeBayes with the parameters (freebayes -C 2 -0 -O -q 20 -F 0.5 -z 0.02 -E 0 -X -u -p 1 -b asm.sorted.bam -v asm.vcf -f asm.fasta).

All corrected PacBio reads used for the construction of Ydb were aligned to the validated assembly to assess per-read accuracy. In contrast to the assembly, which used quality-trimmed corrected reads, the full length of the reads was considered for this analysis. Using BLASR, 99.6% of the untrimmed PacBio corrected reads aligned to the assembly, totaling 98.4% of the corrected bases with an average per-read identity of 93.8%. BLASR was run with the parameters (-nproc 16 -bestn 10 -minReadLength 200 -maxScore -1000 -maxLCPLength 16 -m 4).

S1.3. Identification and PacBio sequencing of Y chromosome derived BACs

We retrieved BAC-end sequences from the *An. gambiae* Pimperena strain BAC-library from VectorBase (https://www.vectorbase.org/download/anopheles-gambiae-s-pimperenabac-ends2012-12fagz). We identified putative Y-linked BACs with CQs calculated for these BAC-ends using *An. gambiae* G3 strain pooled male and female Illumina data (9). Based on a cutoff of CQ < 0.2, we selected 8 BACs for PacBio sequencing. BAC DNA was purified using an alkaline lysis preparation at the midi-scale

to reduce host DNA contamination. BAC DNA was quantified by spectrophotometry (Quant-iT PicoGreen, Life Technologies). DNA was sheared by Covaris g-TUBE and a large insert library was prepared using Pacific Biosciences recommended protocols and a BluePippin instrument to select library fragments from 15-50 kb in length. The BAC samples were sequenced on a Pacific Biosciences RS II sequencing instrument with 180-minute runtimes using P5-C3 reagents. Due to low sequencing yield, only four BACs were able to be successfully assembled using HGAP3 in PacBio SMRTportal 2.2.0 (5). All four BAC assemblies have high and uniform coverage of mapped reads (**Fig. S1**). We aligned reads from pooled males and females of *An. gambiae* Pimperena and G3 colonies against the assembled BACs and calculated CQ in 100-bp windows over the entire assembly (**Fig. S2**). Three of the four BACs had CQ<0.2 throughout, while no extended region of the fourth BAC passed the CQ<0.2 threshold and therefore it was considered to be a false positive. The sequencing reads for all four BACs have been submitted to the SRA: SRX1012990 (10L19), SRX1012997 (28P09), SRX1013030 (60P19), and SRX1013031 (46I18). The assemblies have been deposited in NCBI: KR610409 (10L19), KR494253 (28P09), KR610410 (60P19), and KR610411 (46I18).

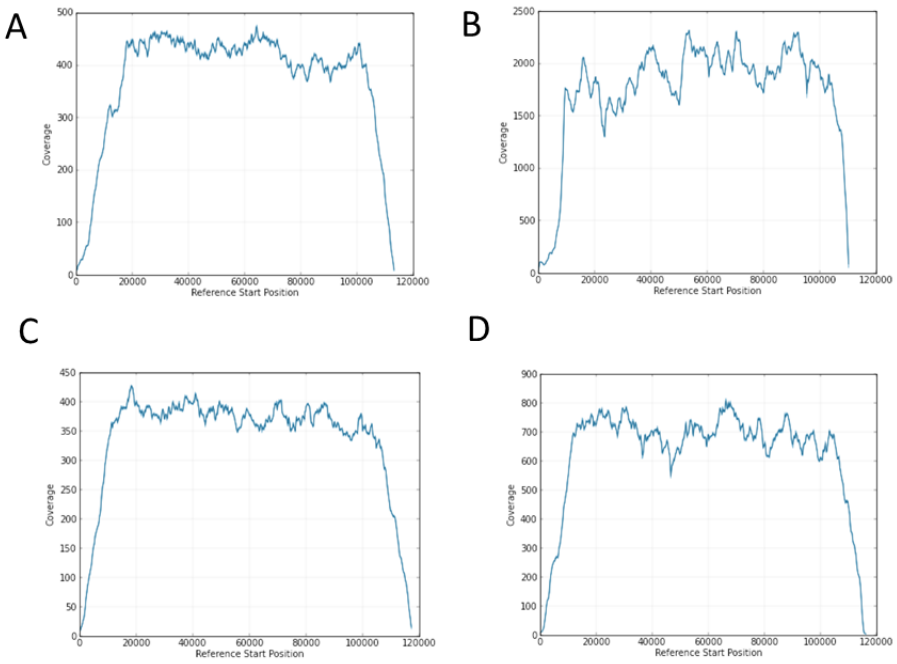

Figure S1. Coverage of the four PacBio sequenced BACs. All four BACs have high and uniform supporting coverage of mapped reads. A, 60P19; B, 10L19; C, 28P09; D, 46I18.

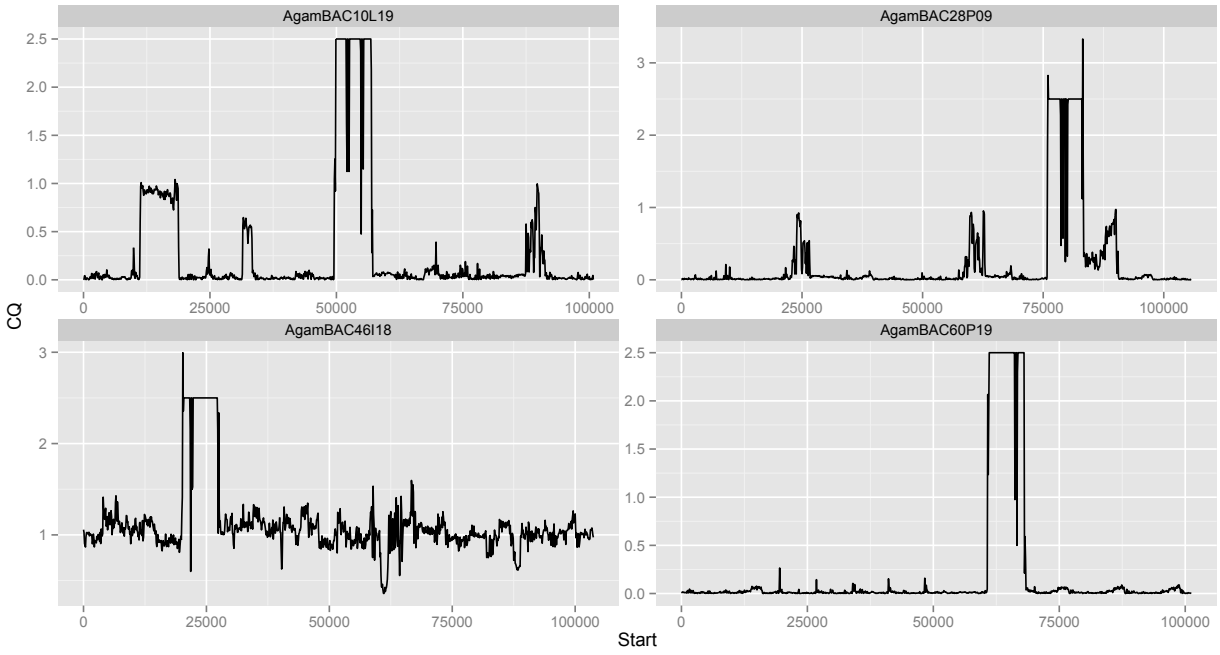

Figure S2. **Ratio of female:male read alignments (CQ) of the four PacBio sequenced BACs in 100-bp non-overlapping windows.** CQ values below the threshold (0.2) are indicative of Y linkage, while CQ=1 (most of BAC46I18) indicates an autosomally-derived BAC. Regions within BACs 10L19, C28P09 and C60P19 whose CQ values do not reflect Y-linkage are repeated on other chromosomes.

S1.4. Male and female Illumina WGS of species in the *An. gambiae* complex

Pools of at least 30 males and 30 females from different strains and species of the *An. gambiae* complex were sequenced (**Table S1**). Genomic DNA was isolated from separate pools of males and females of each strain (species) using the DNeasy Blood and Tissue kit (Qiagen Inc., Valencia, CA, USA), as specified by the manufacturer. Each pool was sequenced with 1 lane of Illumina HiSeq2000 and the data were submitted to the SRA (**Table S1**).

Table S1. Illumina WGS sequencing data from male and female pools.

Dataset	Reads	SRA Accession
An. gambiae G3 Male	52,368,341	SRR534285
An. gambiae G3 Female	59,223,187	SRR534286
An. gambiae Pimperena Male	174,891,782	SRR1509742
An. gambiae Pimperena Female	177,090,932	SRR1508169
An. gambiae Asembo Male	70,209,481	SRR1504990
An. gambiae Asembo Female	61,033,449	SRR1504983
An. merus MAF Male	66,672,308	SRR1504817
An. merus MAF Female	62,883,410	SRR1504857
An. quadriannulatus SANGWE Male	63,214,173	SRR1508191
An. quadriannulatus SANGWE Female	59,554,332	SRR1508190
An. arabiensis Dongola Male	67,965,377	SRR1504818
An. arabiensis Dongola Female	73,915,849	SRR1504792

S1.5. Illumina WGS of field-collected individuals of *An. gambiae*

Collections of indoor resting *An. gambiae* were made by spray catch from five villages in a forest/savanna mosaic zone in the eastern part of Cameroon in Sep-Oct 2009 (10). Mosquitoes were sorted morphologically to *An. gambiae s.l.* and genomic DNA was isolated from individual mosquitoes using the DNeasy Blood and Tissue kit. Molecular identification to species followed an rDNA-based PCR assay (11). Each of 275 samples was individually sequenced by the Wellcome Trust Sanger Institute using 1 lane of Illumina HiSeq2000 (The *Anopheles gambiae* 1000 Genomes Consortium (2015): Ag1000G phase 1 AR3 data release. MalariaGEN. http://www.malariagen.net/data/ag1000g-phase1-AR3). The 40 male and 45 female genomic sequences employed for this study were used by permission from the Ag1000G Consortium, and are available from the SRA (**Table S2**).

Table S2. Male and female *An. gambiae* samples from Cameroon sequenced by the Ag1000G project and used in this study.

SRA Accession	Region	Latitude	Longitude	Sex
ERS224579	Gado-Badzere	5.747	14.442	M
ERS224558	Gado-Badzere	5.747	14.442	M
ERS224453	Gado-Badzere	5.747	14.442	M
ERS224651	Gado-Badzere	5.747	14.442	M
ERS224626	Gado-Badzere	5.747	14.442	M
ERS224430	Gado-Badzere	5.747	14.442	M
ERS224645	Gado-Badzere	5.747	14.442	M
ERS224535	Gado-Badzere	5.747	14.442	M
ERS224493	Gado-Badzere	5.747	14.442	M
ERS224576	Gado-Badzere	5.747	14.442	M
ERS224528	Gado-Badzere	5.747	14.442	M
ERS224494	Gado-Badzere	5.747	14.442	M
ERS224574	Gado-Badzere	5.747	14.442	M
ERS224459	Gado-Badzere	5.747	14.442	M
ERS224519	Gado-Badzere	5.747	14.442	M
ERS224590	Mayos	4.341	13.558	M
ERS224438	Mayos	4.341	13.558	M
ERS224592	Mayos	4.341	13.558	M
ERS224595	Mayos	4.341	13.558	M

ERS224555	Mayos	4.341	13.558	M
ERS224506	Mayos	4.341	13.558	M
ERS224538	Mayos	4.341	13.558	M
ERS224488	Mayos	4.341	13.558	M
ERS224501	Mayos	4.341	13.558	M
ERS224660	Mayos	4.341	13.558	M
ERS224610	Mayos	4.341	13.558	M
ERS224836	Mayos	4.341	13.558	M
ERS224591	Mayos	4.341	13.558	M
ERS224835	Mayos	4.341	13.558	M
ERS224525	Daiguene	4.777	13.844	M
ERS224654	Daiguene	4.777	13.844	M
ERS224655	Daiguene	4.777	13.844	M
ERS224516	Daiguene	4.777	13.844	M
ERS224638	Daiguene	4.777	13.844	M
ERS224560	Daiguene	4.777	13.844	M
ERS224839	Daiguene	4.777	13.844	M
ERS224523	Daiguene	4.777	13.844	M
ERS224619	Daiguene	4.777	13.844	M
ERS224622	Daiguene	4.777	13.844	M
ERS224652	Daiguene	4.777	13.844	M
ERS224618	Zembe-Borongo	5.747	14.442	F
ERS224534	Zembe-Borongo	5.747	14.442	F
ERS224389	Zembe-Borongo	5.747	14.442	F
ERS224552	Zembe-Borongo	5.747	14.442	F
ERS224512	Zembe-Borongo	5.747	14.442	F
ERS224570	Zembe-Borongo	5.747	14.442	F
ERS224587	Zembe-Borongo	5.747	14.442	F
ERS224484	Zembe-Borongo	5.747	14.442	F
ERS224659	Zembe-Borongo	5.747	14.442	F
ERS224428	Zembe-Borongo	5.747	14.442	F
ERS224549	Zembe-Borongo	5.747	14.442	F
ERS224436	Zembe-Borongo	5.747	14.442	F
ERS224449	Gado-Badzere	5.747	14.442	F
ERS224631	Gado-Badzere	5.747	14.442	F
ERS224379	Gado-Badzere	5.747	14.442	F
ERS224395	Gado-Badzere	5.747	14.442	F
ERS224348	Gado-Badzere	5.747	14.442	F
ERS224476	Gado-Badzere	5.747	14.442	F
ERS224390	Gado-Badzere	5.747	14.442	F
ERS224596	Gado-Badzere	5.747	14.442	F
ERS224341	Gado-Badzere	5.747	14.442	F
ERS224522	Mayos	4.341	13.558	F
ERS224514	Mayos	4.341	13.558	F
ERS224641	Mayos	4.341	13.558	F
ERS224577	Mayos	4.341	13.558	F
ERS224585	Mayos	4.341	13.558	F
ERS224644	Mayos	4.341	13.558	F
ERS224557	Mayos	4.341	13.558	F
ERS224599	Mayos	4.341	13.558	F
ERS224544	Mayos	4.341	13.558	F
ERS224503	Mayos	4.341	13.558	F
ERS224578	Mayos	4.341	13.558	F
ERS224666	Daiguene	4.777	13.844	F
ERS224843	Daiguene	4.777	13.844	F
ERS224598	Daiguene	4.777	13.844	F

ERS224662	Daiguene	4.777	13.844	F
ERS224589	Daiguene	4.777	13.844	F
ERS224627	Daiguene	4.777	13.844	F
ERS224647	Daiguene	4.777	13.844	F
ERS224612	Daiguene	4.777	13.844	F
ERS224621	Daiguene	4.777	13.844	F
ERS224613	Daiguene	4.777	13.844	F
ERS224841	Daiguene	4.777	13.844	F
ERS224632	Daiguene	4.777	13.844	F
ERS224844	Daiguene	4.777	13.844	F

S1.6. mRNA-Sequencing (RNA-Seq) datasets

S1.6.1. mRNA sampling

Samples were derived from wild type or transgenic G3 strains of *An. gambiae*. In total our datasets were composed of 22 embryonic, 16 larval and 2 whole adult samples as well as 12 dissected adult samples (**Fig. S3**, **Table S3**). Developmentally staged embryonic samples were derived from the wild type G3 strain or from eggs from crosses of G3 females to transgenic males, bearing the transgene beta2-tubulin I-*PPoI*-11A, which results in embryonic clutches that are ~95% males (12). L1, L2-L3 and L3-L4 larval instar pools were sexed according to the inheritance of an X-chromosome-linked 3xP3: dsRED transgene in a G3 background. Using the transgenic lines allowed us to extract either sexed (for larvae) or heavily sex-biased (for embryos) mRNA from developmental stages, where the sexes are not morphologically distinguishable. Dissected reproductive tissues, comprising testis and accessory glands for males and 48-hr post-blood meal ovaries for females, along with carcass samples lacking reproductive organs, were obtained from 2-3 day old males and non-virgin females of the wildtype G3 strain. Total mRNA was extracted using Tri-Reagent and DNAse treated with TurboDNAse (Ambion). All samples were paired-end sequenced at either 100-bp or 85-bp read lengths using Illumina HiSeq2000. In addition to these datasets we also included in some analyses RNA-Seq data from previously published work (9), which included mixed sex wild type embryos, larvae, pupae and male and female adults.

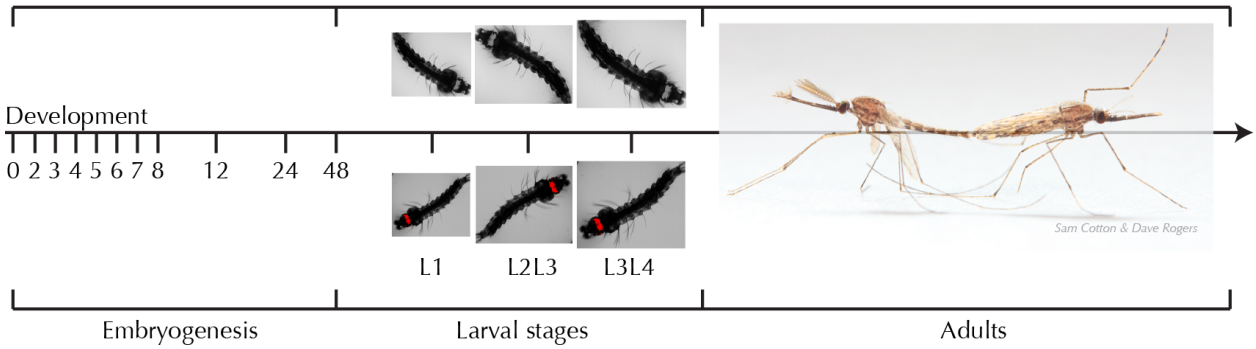

Figure S3. **Developmental stages examined by RNA-Seq.** The numbers below the development bar indicate hours after egg laying, when wild type and *I-PpoI*-11A embryos were collected. The larval instar stages indicated below the fluorescent images illustrate the instars selected and the sexing phenotype. Adults were morphologically distinguishable. Not shown here are the dissected tissues, comprising testes and accessory glands for males and 48-hr post-blood meal ovaries for females, along with corresponding carcass samples lacking reproductive organs.

Table S3. RNA-Seq Illumina samples. All RNA-Seq datasets from the table below can be found under the single study accession number SRP045243 in the NCBI SRA archive.

Dataset	Developmental Stage	Number of reads
1	0hr embryo I-PPoI-11A	46334587
2	0hr embryo wildtype G3	47589058
3	2hr embryo I-PPoI-11A	57212530
4	2hr embryo wildtype G3	50134111
5	3hr embryo I-PPoI-11A	58976089
6	3hr embryo wildtype G3	47345613
7	4hr embryo I-PPoI-11A	36280144
8	4hr embryo wildtype G3	41430533
9	5hr embryo I-PPoI-11A	50632449
10	5hr embryo wildtype G3	50116143
11	6hr embryo I-PPoI-11A	46297153
12	6hr embryo wildtype G3	54887142
13	7hr embryo I-PPoI-11A	45981453
14	7hr embryo wildtype G3	56541072
15	8hr embryo I-PPoI-11A	51380504
16	8hr embryo wildtype G3	53902351
17	12hr embryo I-PPoI-11A	54471610
18	12hr embryo wildtype G3	59193282
19	24hr embryo I-PPoI-11A	42966888
20	24hr embryo wildtype G3	47126528
21	48hr embryo I-PPoI-11A	51472349
22	48hr embryo wildtype G3	34136556
23	Adult female G3	49870018
24	Adult male G3	61473980
25	L1 female Rep1	10057432
26	L1 female Rep2	9479689
27	L1 female Rep3	13791423
28	L1 male Rep1	15140571
29	L1 male Rep2	19026653
30	L1 male Rep3	19056929
31	L2L3 female Rep1	8929475
32	L2L3 female Rep2	10967528
33	L2L3 male Rep1	15365114
34	L2L3 male Rep2	6802972
35	L3L4 female Rep1	13528366
36	L3L4 female Rep2	12465919
37	L3L4 female Rep3	12781905
38	L3L4 male Rep1	14325465
39	L3L4 male Rep2	5787855
40	L3L4 male Rep3	14707969
41	female Carcass1	4235727
42	female Carcass2	4998438
43	female Carcass3	3500175
44	Ovaries 48hrs PBM Rep1	8180622
45	Ovaries 48hrs PBM Rep2	4744888

46	Ovaries 48hrs PBM Rep3	5755372
47	male Carcass1	6249611
48	male Carcass3	9995067
49	male Carcass2	9530566
50	Testis and ACG Rep1	5262842
51	Testis and ACG Rep2	8574538
52	Testis and ACG Rep3	4959052

S1.6.2. mRNA-Seq transcriptome assembly and quantification

Paired reads were processed and aligned (i) against the reference genome of *An. gambiae* (AgamP4) using RNA-STAR (13), or (ii) against transcript sequences using Bowtie2 (14). Allowing only one alignment per read, transcript abundance was quantified using HTSeq (15). Counts were normalized for each sample using an effective library size of concordantly mapping read pairs (16). Biological replicates were collected for larval time points and dissected tissues. We observed good correlations between replicates and also between wild type and I-*PPoI*-11A embryos for similar stages in most cases (**Fig. S4**).

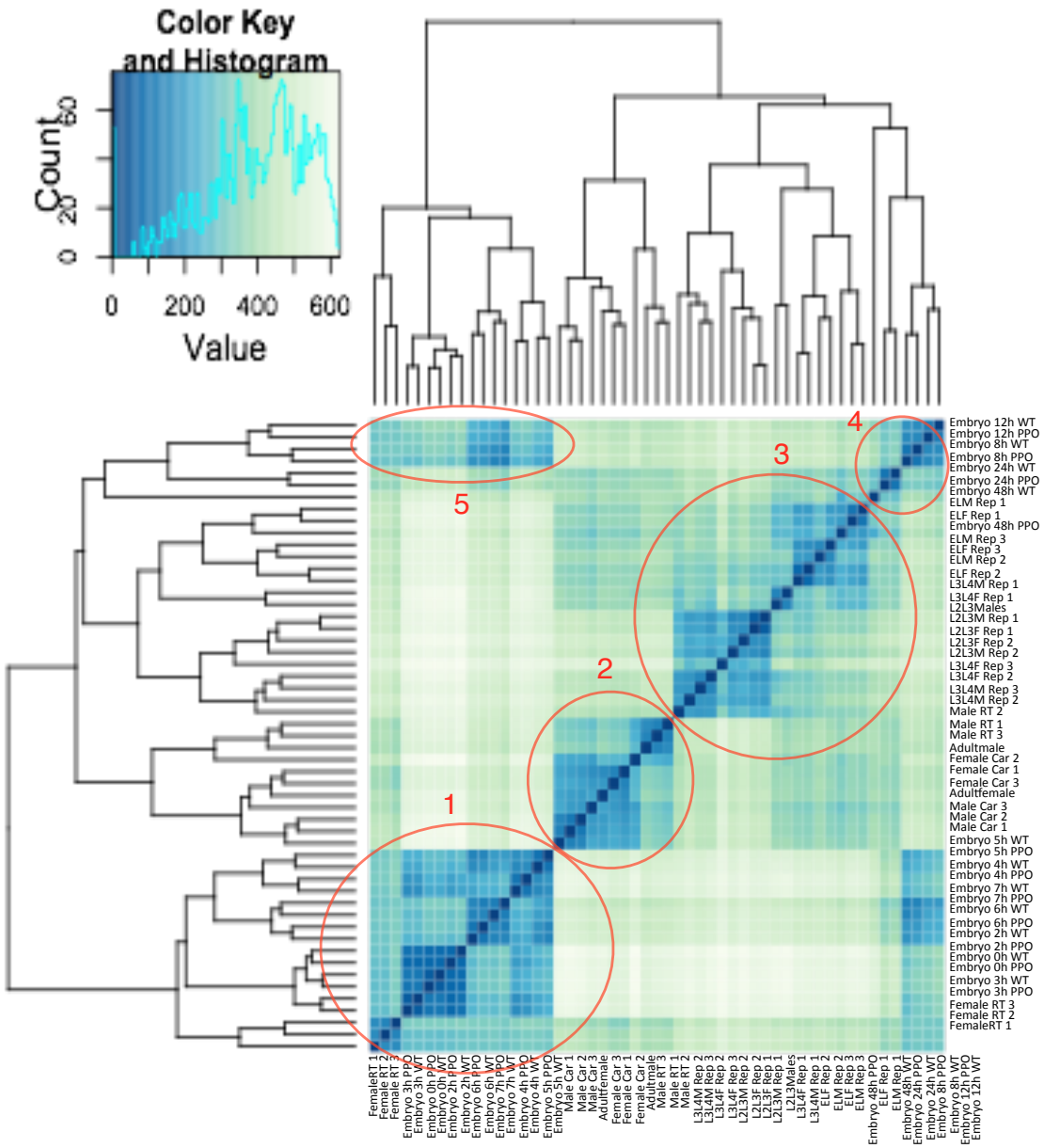

Figure S4. **Correlation analysis between RNA-Seq datasets.** Early embryo and female reproductive samples cluster together at the base of the plot (1), followed by carcass and male reproductive samples (2), larval data (3) and late embryo data (4). Late embryo data also showed a good correlation to early embryo data (5).

S2. Identification of Y sequences from PacBio WGS

S2.1. CQ Methods and Benchmarking

We searched for Y chromosome sequences in the error-corrected WGS PacBio reads using the CQ method as outlined in (9). To calculate CQ, we aligned *An. gambiae* Pimperena male and female Illumina sequences (SRA: SRR1509742, SRR1508169) to the PacBio reads using Bowtie single-end allowing no mismatches over the entire length of the read (-v 0) and reporting all alignments (-a). Stringent alignment parameters reduce misclassification of Y-linkage due to repetitive sequence shared between the Y and other chromosomes. The CQ value for each PacBio read was calculated by dividing the number of female alignments by the number of male alignments over the entire length of the PacBio read. To combat variation in coverage of Illumina sequences, we required a minimum of 30 male alignments to the PacBio reads. To classify a PacBio read as Y-linked, we used an upper-bound CQ threshold of 0.2, imposing the requirement of five times more alignments from male than female sequences over the entire PacBio read.

We have previously used sequences of known chromosome origin from *Homo sapiens*, *Drosophila melanogaster* and *An. gambiae* to estimate the false positive rates of the CQ method using parameters of CQ<0.3 and a minimum of 30 male alignments (9). The false positive rates are 2.44% for *H. sapiens*, 1.85% for *D. melanogaster*, and 0.6% for *An. gambiae*. The more stringent parameters selected in our current study (minimum of 30 male alignments and a CQ threshold of 0.2) should result in even lower rates of false positives.

To estimate the rate of false negatives, we benchmarked the recovery rate of known Y sequences from *H. sapiens*, *D. melanogaster* (9) and *An. gambiae* (17) using

our stringent CQ parameters. As shown in **Table S4**, the recovery rates are 89% for *H. sapiens*, 69% for *D. melanogaster*, and 79% for *An. gambiae*. Thus the CQ method is able to recover a large majority of Y chromosome sequences in all three species.

Table S4. CQ recovery rate of known Y sequences in three species

Species	Number tested	Number Recovered (CQ<0.2)	% Recovery	False Negative rate
An. gambiae	24	19	79%	21%
D. melanogaster	139	97	69%	31%
H. sapiens	2671	2381	89%	11%

For CQ calculations, male and female data were from the same sources listed in ref. (9). The CQ parameters were CQ < 0.2 and male alignments > 29. Only MSY sequences were used for calculation of human CQs.

S2.2. Ydb: an extensive catalog of Y chromosome sequences derived from PacBio sequencing

From the 4,346,630 error-corrected PacBio reads, we identified 79,475 potentially Y-linked reads with more than 29 alignments in males and a CQ value below 0.2. We retrieved these reads and constructed a database we call Ydb (Other Supporting File 1). Ydb contains 246 Mb of predicted Y chromosome sequences with an N50 size of 3,434 bp. Because Ydb is constructed based on a CQ threshold of 0.2, its content is likely derived from the male-limited portion of the Y chromosome that we refer to as the non-recombining Y (NRY). Sequences derived from a putative pseudoautosomal region (PAR) would not be highly represented within Ydb, as these occur in females in equal numbers. The Ydb was screened for bacterial contamination using Kraken (6) as in S1.2.

Only four PacBio reads (EC-read-Y.92, EC-read-Y.53722, EC-read-Y.31930, and EC-read-Y.51121) were identified as bacterial contamination.

Ydb contains a number of satellite and transposable element sequences as well as a small number of genes (**Table S5;** SI Appendix S7). The CQ values of these key features of the NRY are provided in **Table S6**. For cross-species comparisons, which are described in detail in SI Appendix S5, we used relaxed alignment parameters to calculate the relaxed chromosome quotient (RCQ, **Table S7**) of these features.

Table S5. GenBank accession numbers of all sequence features found in *An. gambiae* Ydb.

AgY477[1]	KP666114
AgY373[1]	KP666115
AgY53A	KP666117
AgY53B	KP666118
AgY53D	KP666119
AgY280	KP985635
zanzibar	KP878482
pemba	KP878483
mafia	KP878484
tumbatu	KP878485
chumbe	KP878486
uzi	KP878487
latham	KP878488
mtanga	AF387862
bawe	KU902417
changuu-LCR	KU902416
YG1	KC840350
YG2	KC845524
YG3	KC840349
YG4	KR653310
YG5	KR653309
YG6	KU902415
YG7	AGAP010291-RA
YG8	AGAP008501-RA

[1]Not present in Ydb because it is an X-associated satellite, AgX367 is closely related to AgY373 and AgY477 (**Fig. S6**), and its sequence has been submitted to GenBank under accession KP666116.

All sequence features found in Ydb are described in SI Appendix S2, S3, and S7, and their consensus sequences are provided in Other Supporting File 1. Sequences of candidate Y-linked genes discovered in other *An. gambiae s.l.* species (*An. merus, YG9-YG13; An. quadriannulatus*, *YG14-YG16*) are also provided in Other Supporting File 1 and are described in SI Appendix S7.2.

Table S6. CQs of Y repeats and genes across the *An. gambiae* complex

	An. gambiae G3			*An. gambiae Pimperena*			*An. gambiae Asembo*			*An. quadriannulatus* SANGWE			*An. arabiensis* Dongola			*An. merus* MAF		
	Female	Male	CQ	Female	Male	CQ	Female	Male	CQ	Female	Male	CQ	Female	Male	CQ	Female	Male	CQ
AgY477	2026	181341	0.011	3014	770166	0.004	468	110956	0.004	0	136	0.735	0	0	N/A	0	0	N/A
AgY373	132	104840	0.001	3	63422	0.000	40	41572	0.001	0	0	N/A	0	0	N/A	0	0	N/A
AgY53A	152	100212	0.002	34	471405	0.000	1	740	0.001	0	0	N/A	0	0	N/A	0	0	N/A
AgY53B	646	386405	0.002	24	501201	0.000	53	57031	0.001	0	0	N/A	0	163	0.000	0	0	N/A
AgY53D	21	4834	0.004	14	338607	0.000	0	0	N/A	0	0	N/A	0	0	N/A	0	0	N/A
AgY280	75	713	0.105	11	193912	0.000	19	12	1.583	0	0	N/A	0	0	N/A	0	0	N/A
zanzibar	6356	994335	0.006	12332	3238650	0.004	2460	971679	0.003	4017	458603	0.009	2657	6239	0.426	1569	1531	1.025
mtanga	497	69442	0.007	315	397916	0.001	587	148841	0.004	837	658	1.272	283	272	1.040	378	407	0.929
pemba	261	32537	0.008	1969	314208	0.006	483	136727	0.004	361	88130	0.004	1159	1090	1.063	225	534	0.421
mafia	17752	116144	0.153	40323	597866	0.067	11040	190332	0.058	7427	11358	0.654	6841	5596	1.222	11409	12167	0.938
tumbatu	378	2301	0.164	5030	12947	0.389	1170	5895	0.198	26	18	1.444	424	336	1.262	0	0	N/A
chumbe	1477	37009	0.040	3227	148626	0.022	1177	40633	0.029	1365	1388	0.983	1155	981	1.177	1176	1251	0.940
uzi	5250	13532	0.388	15117	66690	0.227	3190	21822	0.146	1327	1311	1.012	1589	1194	1.331	15	14	1.071
latham	2	688	0.003	0	2126	0.000	0	415	N/A	0	0	N/A	0	0	N/A	0	0	N/A
bawe	912	1057	0.863	1726	2075	0.832	317	355	0.893	88	113	0.779	288	194	1.485	25	16	1.562
changuu	7329	7639	0.959	26882	28920	0.93	9551	10986	0.869	1492	1139	1.310	3841	3471	1.107	3640	4272	0.852
YG1	6	1631	0.004	62	6639	0.009	6	2425	0.002	7	234	0.030	4	494	0.008	0	0	N/A
YG2	2	692	0.003	0	3971	0.000	0	1508	N/A	0	292	0.000	0	567	0.000	0	0	N/A
YG3	34	340	0.100	824	871	0.946	10	8	1.25	0	0	N/A	66	48	1.375	25	30	0.833
YG4	13	80	0.163	0	0	0.000	312	201	1.552	11	10	1.100	61	60	1.017	0	0	N/A
YG5	12	759	0.016	46	8493	0.005	12	95	0.126	39	12	3.250	142	78	1.821	0	0	N/A
YG6	55	20	2.750	204	117	1.744	5	4	1.25	0	0	N/A	0	0	N/A	0	0	N/A
YG7	42	40	1.050	209	145	1.441	45	57	0.789	4	14	0.286	44	39	1.128	0	0	N/A
YG8	64	59	1.085	398	394	1.01	212	252	0.841	22	20	1.100	207	173	1.197	42	41	1.024
YG9	19	16	1.188	900	842	1.069	388	466	0.833	129	101	1.277	309	286	1.08	112	449	0.249
YG10	162	154	1.052	1291	1301	0.992	645	703	0.917	229	252	0.909	417	357	1.168	372	2589	0.144

YG12	172	135	1.274	544	553	0.984	174	232	0.75	59	55	1.073	124	108	1.148	77	511	0.151
YG13	185	109	1.697	893	833	1.072	376	409	0.919	276	337	0.819	383	305	1.256	481	2534	0.190
YG14	178	159	1.119	222	181	1.227	170	183	0.929	10	75	0.133	27	23	1.174	4	5	0.800
YG15	0	0	N/A	0	0	N/A	0	0	N/A	0	540	0.000	0	0	N/A	0	0	N/A
YG16	5	10	0.500	0	0	N/A	0	0	N/A	0	135	0.000	0	0	N/A	0	0	N/A
white	51	35	1.457	1249	674	1.853	543	358	1.517	305	131	2.328	263	101	2.604	229	116	1.974

CQs were calculated with default parameters: Bowtie -v 0 –a, except for the 53-bp satellites with redundancy removed. A longer consensus sequence was used for the 53-bp satellites because our Illumina sequence reads are longer than the satellite consensus. Reads that aligned multiple times to the longer 53-bp consensus sequences were only counted once. AgY477, AgY373, AgY53A, AgY53B, AgY53D, and AgY280 are satellites; *zanzibar*, *mtanga*, *pemba*, *mafia*, *tumbatu*, *chumbe*, *uzi*, *latham*, and *bawe* are TEs from the *zanzibar*-amplified region. *changuu* is a repeat on the Y that is associated with *YG5*. *YG1-YG5* are confirmed Y genes in *An. gambiae*. *YG6-8* are candidate Y genes in *An. gambiae, YG9-YG13* are candidate Y genes in *An. merus*, *YG14-YG16* are candidate genes in *An. quadriannulatus*. *White* is a known single-copy X chromosome gene, for reference. See SI Appendix S7 for details.

Table S7. Relaxed CQs (RCQ) of Y repeats and genes across the *An. gambiae* complex.

	An. gambiae G3			*An. gambiae Pimperena*			*An. gambiae Asembo*			*An. quadriannulatus*			*An. arabiensis*			*An. merus*		
	Female	Male	RCQ	Female	Male	RCQ	Female	Male	RCQ	Female	Male	RCQ	Female	Male	RCQ	Female	Male	RCQ
AgY477**	1052	171989	0.006	5274	474572	0.011	2314	356263	0.006	0	0	0.000	0	0	0.000	122053	166660	0.732
AgY373***	842	58416	0.014	295	116551	0.003	703	96393	0.007	0	0	0.000	4952	2477	1.999	161973	217364	0.745
AgY53A*	633	337614	0.002	649	958194	0.001	11965	69798	0.171	462	20270	0.023	356	314467	0.001	1233038	1440066	0.856
AgY53B*	615	431081	0.001	128	866397	0.000	1994	278877	0.007	4	546	0.007	799	704568	0.001	366272	341628	1.072
AgY53D*	337	2305	0.146	30	572980	0.000	292	203	1.438	0	0	0.000	0	0	0.000	47	74	0.635
AgY280	335325	322273	1.04	270887	1024730	0.264	87635	217078	0.404	12	15	0.800	758879	394611	1.923	1586135	2122548	0.747
zanzibar	63394	1030442	0.062	209262	3275842	0.064	67643	1077943	0.063	135897	2171018	0.063	84878	101615	0.835	110001	106793	1.030
mtanga	2345	65195	0.036	9220	346467	0.027	2562	129684	0.02	6293	6258	1.006	6202	5642	1.099	5494	5618	0.978
mafia	323215	375454	0.861	1193354	1514771	0.788	434902	602075	0.722	402732	410298	0.982	507381	428743	1.183	400546	384117	1.043
pemba	2745	36161	0.076	19000	285265	0.067	5690	120922	0.047	6611	275828	0.024	17475	16460	1.062	6905	15287	0.452
chumbe	3821	34424	0.111	10140	129520	0.078	4690	41180	0.114	6254	6590	0.949	3450	3065	1.126	4731	5213	0.908
uzi	32972	38045	0.867	104611	151411	0.691	35639	57808	0.617	48592	48638	0.999	36617	30869	1.186	15290	14946	1.023
tumbatu	1305	3284	0.397	10487	16940	0.619	2312	7123	0.325	1085	1135	0.956	1979	1659	1.193	1825	1876	0.973
latham	905	4767	0.19	3930	20670	0.19	1404	8695	0.161	835	880	0.949	1961	1577	1.244	1412	1071	1.318
bawe	11461	10480	1.094	41856	39911	1.049	15340	16141	0.95	10411	11284	0.923	13644	12859	1.061	16189	16117	1.004
changuu	16916	26825	0.631	48031	117963	0.407	17861	20955	0.852	16399	11393	1.439	15163	12997	1.167	13902	15145	0.918
YG1	9562	14100	0.678	21827	28493	0.766	9951	14647	0.679	9556	14515	0.658	9620	11574	0.831	11924	10692	1.115
YG2	2555	3322	0.769	12185	16423	0.742	4080	6830	0.597	5325	8635	0.617	4592	6659	0.690	3900	5120	0.762
YG3	586	859	0.682	2309	2510	0.92	434	610	0.711	736	869	0.847	448	447	1.002	367	361	1.017
YG4	996	1117	0.892	2331	2439	0.956	1851	2140	0.865	870	782	1.113	926	961	0.964	460	505	0.911
YG5	83	1166	0.071	641	14686	0.044	205	410	0.500	1997	1133	1.763	1206	694	1.738	162	180	0.900
YG6	528	202	2.614	3317	1756	1.889	1048	650	1.612	940	501	1.876	942	474	1.987	208	113	1.841
gYG7	513	494	1.038	11332	11390	0.995	2437	2409	1.012	3145	2895	1.086	1906	1769	1.077	1511	1357	1.113
YG8	219	166	1.319	1242	1359	0.914	530	566	0.936	581	565	1.028	652	552	1.181	396	394	1.005
YG9	70	69	1.014	1575	1512	1.042	624	759	0.822	755	721	1.047	787	749	1.051	1023	9416	0.109
YG10	330	303	1.089	2295	2371	0.968	917	985	0.931	1409	1347	1.046	1157	1037	1.116	1649	12736	0.129
YG11	137	124	1.105	706	753	0.938	274	325	0.843	358	405	0.884	338	316	1.070	493	2648	0.186
YG12	289	231	1.251	962	971	0.991	348	448	0.777	519	523	0.992	435	411	1.058	627	4178	0.150
YG13	474	313	1.514	2209	2138	1.033	870	968	0.899	1099	1131	0.972	1106	996	1.110	1281	8166	0.157

YG14	1158	897	1.291	1622	2021	0.803	1622	1650	0.983	420	2445	0.172	1209	935	1.293	764	882	0.866
YG15	748	684	1.094	1166	1030	1.132	1637	1806	0.906	536	1469	0.365	402	427	0.941	288	324	0.889
YG16	335	309	1.084	677	498	1.359	848	983	0.863	171	573	0.298	173	183	0.945	154	166	0.928
white	120	70	1.714	1827	937	1.95	688	465	1.48	861	424	2.031	936	436	2.147	587	291	2.017

If not marked, CQs were calculated with BWA-MEM (18) using male and female Illumina sequences. Satellite sequences AgY477, AgY373, and AgX367 are closely related or even identical over much of their length (see fig. S6). To avoid cross-alignments between AgY477 and AgX367 in the AgY477 analysis, we used a reference sequence that distinguishes the Y-biased from the X-biased satellites (orange shading in fig. S6). However, note that this 90 bp sequence (an insertion from 30-120 bp not found in AgX367, but present in both AgY477 and AgY373) is similar enough to align between the Y-biased satellites. Thus, values for AgY477 in this table reflect the combined AgY373 and AgY477 alignment results. For the AgY373 analysis, the last 85 bp of this satellite monomer were used as a reference for read mapping, as they are unique (fig. S6A, red shading). Due to the short length of the reference sequences for AgY477 and AgY373, BLASTN with default parameters was used to calculate the number of alignments. Reads that aligned more than once were removed. AgY53A and AgY53B were so short, that we found it best to use BLASTN with word_size 7 to calculate the number of alignments. Reads that aligned more than once were removed. AgY477, AgY373, AgY53A, AgY53B, AgY53D, and AgY280 are satellites; *zanzibar*, *mtanga*, *pemba*, *mafia*, *tumbatu*, *chumbe*, *uzi*, *latham*, and *bawe* are sequences from the *zanzibar* amplified region; *changuu* is a repeat on the Y that is associated with *YG5; YG1-YG5* are confirmed Y genes in *An. gambiae*. *YG6-8* are candidate Y genes in *An. gambiae, YG9-YG13* are candidate Y genes in *An. merus*, *YG14-YG16* are candidate genes in *An. quadriannulatus*. *White* is a known single-copy X chromosome gene, for reference. See SI Appendix S7 for details.

S2.3. Comparison of the CQ method to the Y chromosome Genome Scan (YGS) method

The YGS method is an alternate method to identify Y sequences that was designed to compare reference genome assemblies, typically constructed from mixed sexes, to Illumina sequences determined from females (19). It is a kmer-based method that identifies Y-linked sequences (typically scaffolds in an assembly) based on the absence of kmers shared with females, and more specifically, the absence of female kmers that are not repetitive in the female genome. For unambiguous classification of a sequence as Y-linked, YGS requires that the sequence has a large fraction (e.g., >70%) of kmers unmatched in the female Illumina data, under the assumption that Y chromosome sequences are both limited to males and distinctive in sequence from the other chromosomes. We tested the YGS methods using previously known *An. gambiae* Y chromosome sequences ("Y-unplaced" contigs and 3 genes) together with the major repeat features of the Y chromosome (SI Appendix S3; **Table S5**) validated as Y-linked in *An. gambiae* by male-specific PCR and physical mapping (SI Appendix S4). Using a female-derived Illumina assembly of *An. gambiae* to filter out repetitive kmers, we calculated the percent of non-repetitive kmers in these known-Y control sequences that were unmatched in female Illumina data (**Table S8**). None of the 24 Y-unplaced contigs have more than 40% unmatched kmers. This proportion of unmatched kmers is significantly lower than proportions (>70%) that effectively separated Y sequences from X and autosomal sequences in humans and flies (19). Even two known Y genes, *YG1* and *YG2* (9), only had unmatched kmer percentages of 43% and 57%, respectively. Furthermore, none of the major repeat features of the Y identified in Ydb (SI Appendix S3) had unmatched kmer percentages greater than 19%. AgY53A and AgY53B had zero unmatched kmers. While none of these sequences could be identified as Y-linked using the YGS method, all but a few short sequences were identified as Y-linked using the CQ method.

The difference between *D. melanogaster* and *An. gambiae* that causes the YGS method to fail is most likely due to extensive similarity between sequences on the *An. gambiae* Y and sequences elsewhere in the genome. For example, there are other closely-related copies of *zanzibar* and other Y-enriched repeats (SI Appendix S3; **Fig. S11**), scattered throughout other chromosomes. Similarly, satellite sequences AgY477, AgY373, AgY53A, and AgY53B (SI Appendix S3; **Table S12**; **Fig. S11**) are all present on both the X and Y, even though they are far more abundant on the Y chromosome. Therefore, kmers in these sequences match kmers from female Illumina data. The CQ method, which compares the relative abundance of sequences between male and female data, is more effective at identifying Y chromosome sequences in *An. gambiae*.

Table S8. Comparison of YGS and CQ for known Y chromosome sequences

Sequence ID	total k-mers	non-repetitive k-mers	unmatched kmers	% unmatched	Female Alignments	Male Alignments	CQ
Y_unplaced.1	2580	1416	280	20	691	3144	0.220
Y_unplaced.2	4209	2163	780	36	5103	17615	0.290
Y_unplaced.3	35250	19379	7713	40	5527	150334	0.037
Y_unplaced.4	1213	1017	100	10	13	177473	0.000
Y_unplaced.5	3176	1813	461	25	744	3958	0.188
Y_unplaced.6	4563	2786	83	3	21674	1743083	0.012
Y_unplaced.7	2039	1562	458	29	1051	595540	0.002
Y_unplaced.8	1221	880	121	14	35	384665	0.000
Y_unplaced.9	2298	1638	1	0	3506	6927	0.506
Y_unplaced.10	3673	2820	327	12	162	2136161	0.000
Y_unplaced.11	2556	1396	47	3	764	19679	0.039
Y_unplaced.12	2486	1895	252	13	44	402651	0.000
Y_unplaced.13	1435	1044	147	14	17	137871	0.000
Y_unplaced.14	1394	1015	47	5	45	448882	0.000
Y_unplaced.15	1350	956	164	17	24	266163	0.000
Y_unplaced.16	1337	1138	52	5	45	719855	0.000
Y_unplaced.17	1316	650	28	4	692	2052	0.337
Y_unplaced.18	1296	1078	86	8	24411	2506597	0.010
Y_unplaced.19	1262	952	133	14	1002	98047	0.010
Y_unplaced.20	1241	1013	5	0	22453	3156554	0.007
Y_unplaced.21	1218	1026	60	6	20	330155	0.000
Y_unplaced.22	1123	935	59	6	2	14714	0.000
Y_unplaced.23	1102	933	46	5	13	360738	0.000
Y_unplaced.24	1000	729	82	11	12	127402	0.000
YG1	3768	2613	1127	43	62	6639	0.009
YG2	2321	1547	885	57	0	3971	0.000
YG3	850	498	0	0	824	871	0.946
AgY477	266	220	9	4	13146	767543	0.017
AgY373	359	249	6	2	5267	967787	0.005
zanzibar	6745	3867	418	11	12231	3229726	0.004

mtanga	4070	2513	487	19	315	397916	0.001
AgY53D	536	398	13	3	44	1592076	0.000
AgY53B	1106	982	0	0	77	1561798	0.000
AgY53A	258	201	0	0	34	471405	0.000

S2.4. Spatial Mapping

We developed a spatial mapping method to visualize the difference between male and female alignments across all bases in a reference sequence, as compared to CQ which gives an average value over the entire sequence. Illumina WGS sequences from males and females of the colonies and species included in our study were aligned to all Y loci using Bowtie2 (14) using default parameters with the exception of reporting all alignments (-a). We then used bedtools v.2.17.0 genomeCoverageBed (20) to report the number of reads aligned from each sample for every basepair of the Y locus (-d option). Read counts were then normalized by the sample library size to correctly compare male and female samples of each species. The difference between male and female counts was then calculated and female counts were then brought to negative scale. Read counts for male, female and their relative difference were then plotted for every basepair for each Y locus and for each species. Examples of spatial mapping are described in SI Appendix S3 and S7.

S3. The *An. gambiae* Y chromosome contains massively amplified satellites and retrotransposons.

To characterize the major satellite and transposable element (TE) content of the *An. gambiae* Y chromosome (see **Table S5**), we ran RepeatMasker (21) with the default settings against Ydb, using the *An. gambiae* PEST repeat library supplemented with previously identified Y satellites [AgY477, AgY373, AgY53A, and AgY53B; (22)] and TEs [mtanga; (23)]. In total, 98% of Ydb was masked (**Table S9**). Based on the resulting RepeatMasker output we calculated the number of base pairs masked by each repeat in the repeat library for both known and unknown repeat types through iterative clustering and consensus generation. The consensus sequences for satellites AgY477/AgY373, AgY53A and AgY53B were responsible for masking large portions of Ydb. Using the BACs we were able to identify full-length copies matching consensus sequences of many other non-satellite repeats, including *zanzibar*, *pemba*, *mafia*, *chumbe*, *tumbatu*, and *latham*. As a complementary approach, we aligned the male and female Illumina sequences to the PacBio reads using BWA-MEM (18), calculated RCQs, and selected reads with CQ < 0.2. Then we identified the reads with the largest absolute number of alignments. These contained satellites AgY477/AgY373 and amplified TE sequences dominated by *zanzibar* (SI Appendix S3.2), in agreement with the results from repeat-masking (**Table S9**).

Table S9. The proportion of Ydb bases masked by major repetitive features of the *An. gambiae* Y chromosome

Repetitive Sequence	**Length (bp)**	**Percent of Ydb**	**Masked bp**
Satellites			
AgY477/AgY373	477/373	0.437	107472713
AgY53A	53	0.017	3984678
AgY53B	53	0.020	5009893
AgY53D	53	0.004	996214
AgY280	280	0.010	2454154
Amplified Region			
zanzibar	6952	0.2676	65861564
mtanga	4084	0.0443	10909386
pemba	4054	0.0359	8847701
mafia	10689	0.0627	15424206
chumbe	4442	0.0171	4219616
uzi	1929	0.0060	1484017
tumbatu	1832	0.0011	266058
latham	203	0.0007	182957
Other Repeats			
changuu	4838	0.0134	3297609
simple repeats		0.0083	2034466
uncategorized repeats		0.0335	8257520

S3.1. The satellite amplified region (SAR)

S3.1.1. AgY477 and AgY373

AgY477 and AgY373 are closely related, and share 100% identity over part of their length (**Figs. S5-S6**). Therefore, to avoid double-counting the amount of Ydb masked by these sequences, we first masked Ydb with AgY477 and then masked the resulting RepeatMasker output with AgY373. We then combined the percent of Ydb masked by both satellites. Together, AgY477 and AgY373 mask 43.7% of the sequences in Ydb (**Table S9**), and by extension, a large portion of the *An. gambiae* NRY chromosome.

AgY477 and AgY373 also bear sequence similarity to another satellite, AgX367, located on the X chromosome (**Fig. S6**). AgX367 does not share with AgY477 or AgY373 any regions of 100% identity longer than the 100 bp reads used in this study. Therefore, reads from AgX367 will not interfere with CQ calculation because we require 100% identity over the entire length of the read. Using the CQ method, we found that AgY477 and AgY373 are ~100x to ~1000x more abundant in males than in females indicating that they are highly enriched on the Y (tables S6-S7).

Using the error-corrected PacBio reads, we were able to determine how the AgY477 and AgY373 satellites are arranged on the Y chromosome. In Ydb, we identified 33,681 reads containing either AgY477 or AgY373 using BLASTN. We retrieved many of these reads and examined the arrangement of the two satellites, using BLASTN. We found that AgY477 and AgY373 occur in tandem arrays consisting of one or the other satellite monomer (Fig. 2A). Using RepeatMasker, we masked all 33,681 PacBio reads with AgY477 and AgY373 sequences as the repeat library, and found that 95.8% of the total sequence was masked. Many of the bases that remain unmasked are within AgY477 and AgY373 and therefore appear to be from sequence

variation of the satellites or poorly error-corrected regions, rather than from other types of sequences.

Frequently, one or more AgY477 or AgY373 monomer was present in a tandem array dominated by the other satellite. The canonical 477 and 373 monomers differ in their first 30 base pairs, and although they are very similar downstream, they can be distinguished by slight sequence variation. Upon close inspection of the mixed arrays, we found evidence of chimeric monomers, consistent with recombination events. Examples involving AgY477 and AgY373 were found, in which one satellite monomer is present at very low frequency in a tandem array dominated by the other. In such arrays, the rare monomer contains the first 30 base pairs of the dominant satellite, and the adjacent monomer of the dominant satellite carries the first 30 base pairs of the rare satellite monomer (Fig. 2A; **Fig. S5**).

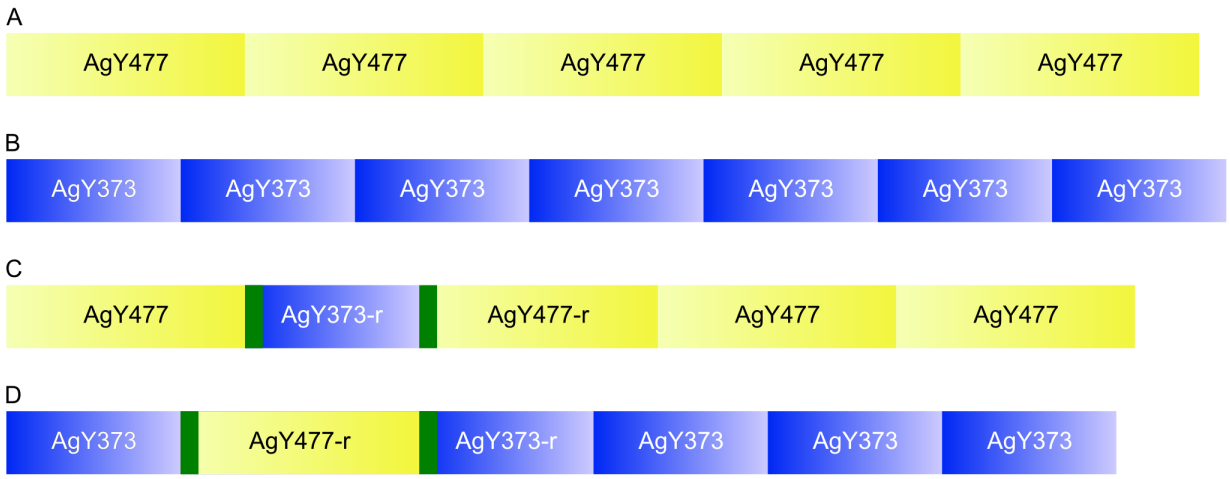

Figure S5. Schematic illustration of the structure of AgY477 and AgY373 satellite arrays. (A) The tandem repeat structure of AgY477. (B) The tandem repeat structure of AgY373. (C) Recombination between AgY477 and AgY373 illustrated by a recombinant AgY373 monomer (AgY373-r) with the first 30 bp of AgY477 in an AgY477-dominated array, and the adjacent recombinant AgY477 monomer (AgY477-r) with the first 30 bp of AgY373. (D) Recombination between AgY477 and AgY373 illustrated by a recombinant AgY477 monomer with the first 30 bp of AgY373 in an AgY373-dominated array, and the adjacent recombinant AgY373 monomer with the first 30 bp of AgY477.

A

```
Y477  TTTGAGCATGTGTTTAAAGGGTAATATGACCCATAAAGGTTAAGCTCAGAGCTTAGGAAC
Y373  AGGAAATTAAAAATTGCGAAGCAAGTTTATCCATAAAGGTAAAGCTCAGAGCAAAGGAAC
X367  TTTGAGCATGTGTTTAAAGGGTAT-----------------------------------

Y477  ATATAGTAAATTGCCTCTAAAGTTGAAGGTTTTGTGGAAAGTCTTCAAATGTGCTTCGGG
Y373  ATATAGTAAATTGCCTCTAAAGTTGAAGGTTTTGTGGAAAGTCTTCAAATGTGCTTCGGG
X367  ------------------------------------------------------------

Y477  GGACTATGACCCAGTATGAAACTTTTTCATCGCCAAGATCCTTGTTATTGTGTCCCAGGG
Y373  GGACTATGACCCAGTATGAAACTTTTTCATCACCAAGATCCTTGTTATTGTGTCCCAGGG
X367  --------------TATGAAACTTTTTCATCGCCAAGATCCTTGTTATTGTGTCCCAGGG
                    ********************* ****************************
Y477  CTTTGATTTGCTTATTCATGAAGCCCCAATGACAAAAGAACGATAATGAATGACCTTGCA
Y373  CTTTGATTTGCTTATTCATGAAGCCCCAATGGCATAAGAACGATATTGAATGACCTTGCA
X367  CTTTGGTTTGCTTATTCATGAAGGCCCAATGACATAAAAACGATAATGGATGACCTTGCA
      ***** ***************** ******* ** ** ***********************
Y477  TTTTCGTCAAACATTCAAGCATGGCCATGGGGACGGATGAGAAAGCTCAAGTGATGTAGT
Y373  TTTTCGTCAAACATTCAAGCATGGCCATGGGGACGGGTGAGAAAGCTGTTTTCAAGAAAA
X367  TTTTCGTCAAAATTTCAACCATGGCCATAGGGACGGGTGAAAAAGCTCAAGTGATGTAGT
      ***********  ***** ********* ******* *** ******
Y477  TGGATGTTC--CTTCAAATGGCCATAACTTCGTAACCATGTGTCCTAGCGTGATGATTGA
Y373  AACACCTCTAACTTTAAACGGCCATAACTTTTGAACTAGAAGAGCTAGAGACATAATCTC
X367  TGGATGTTC--CTTCAAATGGCCATAACTTCGTAGCCATGTGTCCTTGCGTGATGATTAA

Y477  GACAGTTTTGGGAAGGTATTGAAGTGGTCTACAAGATCTGCGCATAGGTTTAAAGATCAG
Y373  TTCACCGTTGGAA-------------------------------------------------
X367  GACAGTTTTGGAAAGGTATTGAAGTGTTCTACAAGATCTGTACAAAGGTTAAAAGATCAA

Y477  AATCACTGGTAGCCTAGTAAATGGCCTCTGAATGCATTGTACTCGGGAAAACCTGTCAA
Y373  ------------------------------------------------------------
X367  ATTCATTTTTAGTCTAGTAAATCGCCTCTGAATGCATTGTACTCGGGAAAACCTGTCAA
```

B

Figure S6. Relationship of AgY477, AgY373, and AgX367. (A) Alignment of AgY477, AgY373, and AgX367, indicating a region that distinguishes the Y-biased satellites (AgY477/AgY373) from AgX367 (orange shading) and the region that distinguishes the two Y-biased satellites (AgY477 and AgY373; red shading). (B) Schematized sequence relationships. The same color indicates similar sequences, while different colors indicate different sequences.

S3.1.2. AgY53A and AgY53B

Two satellites with 53 bp period sizes, AgY53A and AgY53B, mask 1.6 and 2.0 percent of Ydb respectively (**Table S9**). These two satellites appear to occur in long, uninterrupted, tandem arrays. Using BLASTN with an e-value of 1e-100 we identified 2,654 error-corrected PacBio reads containing tandem arrays of AgY53A. With the same parameters we identified 6,702 PacBio reads containing tandem arrays of AgY53B. We identified 34 error-corrected PacBio reads that contained a junction between AgY53A and AgY53B satellite arrays. Therefore, these satellite arrays may occupy neighboring regions of the Y chromosome. AgY53A and AgY53B are some of the most male-biased sequences on the Y. They are ~10,000-20,000 times more abundant in males than in females.

S.3.1.3. AgY53D and AgY280

We noticed that two unannotated satellites were responsible for variation in copy number between individual *An. gambiae* (SI Appendix S5). These satellites have a 53 bp consensus and a 280 bp consensus, respectively, and were named AgY53D and AgY280. AgY280 and AgY53D occur interspersed in the same arrays (Fig. 2C and **Fig. S7**). AgY280 has regions of similarity to AgY373 (**Fig. S8**). Detailed description is provided in SI Appendix S5.

Figure S7. Structure of AgY53D and AgY280 arrays. Two examples illustrating a complex interdigitating relationship of AgY53D and AgY280, likely driven by sequence similarity.

```
AgY373   159   TTCCAACGGTGCAAAAATTATGTCTCTAGCTCTTCTGGTTCCAAAGTTATGGCCGTTTAA
               ||||||||||| | | ||||||||||||||||||| |||| ||||||||||||||||||
AgY280   373   TTCCAACGGTGAAGAGATTATGTCTCTAGCTCTTCTAGTTCAAAAGTTATGGCCGTTTAA

AgY373   219   AGTTAGATGAttttttttCTTCAAAACTGCTTT
               ||||||| |  ||||||||| ||||| |||||
AgY280   313   AGTTAGAGGTGTTTTTTCTTGAAAACAGCTTT
```

Figure S8. Alignment between AgY280 and AgY373, illustrating their sequence similarity. AgY373 is in plus orientation and AgY280 is in minus orientation.

S3.2. The *zanzibar* amplified region (ZAR)

S3.2.1. *zanzibar*

We identified a Ty3/Gypsy LTR retrotransposon, *zanzibar*, that represents 26.76% of the base pairs of Ydb using RepeatMasker. The *zanzibar* LTR retrotransposon had not been previously identified as a large component of the Y. We identified many tandem repeats of *zanzibar* in the BACs and error-corrected PacBio reads (Fig. 2A,B). We analyzed these tandem repeats and generated a 6,952 bp *zanzibar* consensus sequence. We analyzed the composition of this consensus sequence using Repbase. In the tandem repeats present on the Y, LTRs do not flank both ends of *zanzibar*. An LTR is only found on the 3' side of the consensus followed directly by the beginning of the next *zanzibar* tandem repeat. In the UNKN chromosome of the PEST assembly, a copy of *zanzibar* was found with LTRs flanking both sides surrounded by non-*zanzibar* sequences. Due to the presence of only one LTR in the *zanzibar* tandem array, we presume that these arrays were not produced by transposition but instead by another mechanism.

To further explore the architecture of the *zanzibar* tandem repeats, we examined all the error-corrected PacBio reads that contained both the beginning and end of the *zanzibar* consensus. We split *zanzibar* into three pieces—which we call beginning, middle, and end—and performed BLASTN against all the error-corrected PacBio reads with parameters: evalue 1e-100 and max_target_seq 10000000. Using these BLASTN results we determined which of the PacBio reads contain either the beginning or end of *zanzibar* more than one time (e.g. reads with a *zanzibar*/*zanzibar* junction). We recovered 4,615 PacBio reads meeting these criteria. When one copy of *zanzibar* ended another invariably started leading us to conclude that *zanzibar* is organized in large tandem arrays we call the *zanzibar* "amplified" region of the *An. gambiae* Y. The *zanzibar* element is associated with a variety of other TEs on the Y chromosome (see

below). Because of this striking association, we have designated these other TEs with the names given to other islands in the Zanzibar archipelago off the coast of East Africa.

S3.2.2. Insertions into *zanzibar* are amplified as part of the ZAR

We noticed that the tandem arrays of *zanzibar* were not composed entirely of *zanzibar*; other sequences are inserted into *zanzibar* copies within the arrays. To systematically identify sequences inserted into the *zanzibar* tandem arrays, we performed BLASTN with the *zanzibar* consensus sequence as the query against Ydb and retrieved all matching reads. Next, we ran RepeatMasker on all these reads using the *zanzibar* consensus sequence as the repeat library. Therefore, the remaining sequence should be sequences associated with *zanzibar*. To identify highly abundant insertions into *zanzibar*, we used cd-hit-est (24) to cluster the sequences. Sequences in the most abundant clusters were subjected to further examination.

S3.2.2.1. *mtanga*

One notable insertion into *zanzibar* is *mtanga*, a previously-identified Y chromosome transposable element (23). *mtanga* is a 4,284 bp long Ty1-copia element with 119 bp LTRs. *mtanga* masks 4.4% of the sequences in Ydb with RepeatMasker (**Table S9**). We noticed that *zanzibar* and *mtanga* co-occurred often in Ydb, so we performed analysis to determine whether *mtanga* inserted into the same site in *zanzibar*. We identified an insertion site of *mtanga* into *zanzibar* and retrieved the flanking sequences on both sides. We used these sequences to perform BLASTN against the error-corrected PacBio reads with evalue 1e-50. In the resulting output, 3,875 error-corrected PacBio reads aligned to *mtanga*. Of these, 2,565 PacBio reads aligned to both *mtanga* and one of the insertion flanking sequences. The remaining 1,310 sequences were

almost exclusively less than the length of *mtanga*. Therefore, when present in Ydb, *mtanga* appears to be associated with the same *zanzibar* insertion site.

To be more specific about the exact insertion site, we chose a reference *zanzibar* sequence with an *mtanga* insertion. Noting the *zanzibar* position where the insertion occurred, we measured the distance from this reference position to the start (5') or end (3') of *mtanga* across the set of reads carrying both elements. If the insertion site of *mtanga* into *zanzibar* is the same or nearly the same as the reference in other copies of *zanzibar*, the measured distances should be both small and consistent. The overwhelming majority of distances to the insertion start (96.2%; 1,221/1268) were within 1-6 bases of the reference. Indeed, the 6-bp distance is the result of a 5 bp insertion/deletion, rather than a different insertion site. Similarly, 98.3% (1,268/1,289) of the distances to the insertion end (3' *zanzibar-mtanga* junction) were 2 bp. We conclude that almost every insertion of *mtanga* into *zanzibar* occupies the exact same site.

S3.2.2.2. *pemba*

Another abundant insertion into *zanzibar* is a 4,054 bp BEL LTR retrotransposon that we have called *pemba*. It is flanked by 200 bp LTRs and has an ORF from 392-3091 bp. We used BLASTN to retrieve 494 PacBio reads from Ydb that contain both the *pemba* consensus and *zanzibar*. We assessed whether the pemba insertion site is consistent between different *zanzibar* copies, using the approach described in SI Appendix S3.2. In 243 of 256 (95%) cases we examined, the *pemba* consensus had the exact same insertion site into *zanzibar*. To confirm Y linkage of *pemba*, we performed PCR on genomic DNA of males and females of the Kisumu and G3 strain.

S3.2.2.3. *mafia*

Another abundant insertion into *zanzibar* is an LTR retroelement we have named *mafia*. The *mafia* consensus sequence, verified as Y-linked by PCR, is 10,698 bp with 408 bp LTRs on both sides. It contains an ORF with similarity to a BEL LTR retrotransposon from 457-6,336 bp. However, as a whole *mafia* appears to be a composite of different elements. The sequences between 6,500 and 10,000 match several MITEs and other repetitive sequences. Nevertheless, the *mafia* composite is characteristic of Ydb and it, too, is associated with *zanzibar*. We identified 230 reads containing *mafia* and *zanzibar* sequences, and detailed analysis revealed that like the other elements, *mafia* also appears to be inserted within a consistent site in *zanzibar*. In 159 of 194 (82%) cases examined, *mafia* insertion sites were identical.

S3.2.2.4. Other insertions into *zanzibar*

Several other less abundant insertions into *zanzibar* were also identified in our systematic analysis. These repeats include: a 4,442 bp GYPSY LTR retrotransposon with 113 bp LTRs that we named *chumba*, a 1,832 bp element of unknown type with 50 bp LTRs that we named *tumbatu*, a 1,929 bp element with similarity to hAT and HARBINGER DNA transposons that we named *uzi*, and a 203 bp element of unknown type without LTRs that we named *latham*.

S3.2.3. Mechanism of amplification

It is unlikely that the pattern of tandem *zanzibar* repeats was created by retrotransposition, principally because of the tandem head-to-tail organization of the repeated copies. Further, for each element that has inserted into *zanzibar* (e.g., *mtanga*, *pemba*, and *mafia*), the insertion site is identical from one *zanzibar* copy to the next in the ZAR (Fig. 2A,B). The most parsimonious

explanation for this arrangement is that the insertions pre-dated *zanzibar* amplification, and were co-amplified with *zanzibar* subsequently through a mechanism unrelated to retrotransposition.

In total the repeats in the ZAR represent 43.5% of sequences in Ydb (**Table S9**). Therefore, the ZAR likely represents a dominant structural feature of the *An. gambiae* NRY.

S3.3. *changuu*

changuu is a repetitive conglomerate element homologous to a region within the segmental duplication (or "Low Copy Repeat"; LCR) that flanks the breakpoints of the 2Rj inversion in the Bamako chromosomal form of *An. gambiae* (25). Each copy of the ~16 kb 2Rj LCR contains two 5.3-kb inverted repeats separated by a 4 kb spacer; *chaguu*-homologous sequences occur in the inverted repeats. *changuu* is highly abundant within the genome of *An. gambiae*, and especially abundant on the Y. Interestingly, because sequences homologous to the 2Rj LCR were only found in the unmapped assembly of PEST, the origin of the 2Rj LCR was initially proposed to be either pericentric heterochromatin or the Y chromosome (25). We identified *changuu* initially in more than 500 PacBio Ydb reads, due to its association with two genes that are homologous to a syntenic gene-pair on the X chromosome (AGAP001087 and AGAP013547). These homologs co-occur with *changuu* in the part of the LCR that flanks 2Rj (**Fig. S9**). Both mean CQ and spatial mapping of genomic reads on AGAP001087, AGAP013547, and their multicopy homologs on other chromosomes showed relatively more reads mapping from males, which suggested Y-linkage (**Fig. S9**). Accordingly, we retrieved Ydb reads that had BLASTN hits to AGAP001087. To investigate the genic neighborhood of AGAP001087 homologs on Ydb reads, we used these PacBio reads as queries in BLAST searches of the full gene set of *An.*

gambiae. Hits to AGAP013547 were found in 59% of the reads, and hits to *YG5* (or AGAP013757, its autosomal homolog) in 11% of the reads, supporting Y-linkage. We confirmed that the homology of *changuu* to the 2Rj LCR extends beyond the coding sequences of genes homologous to AGAP001087 and AGAP013547, by masking the 2Rj LCR with AGAP001087 and AGAP013547 and using the masked sequences in BLAST searches of the set of 1171 PacBio Ydb reads matching AGAP001087 sequence. Of the 1171 reads, 1078 also matched the gene-masked LCRs (93 did not). Thus, there is a strong association of both gene homologs and *changuu* in Ydb reads, and the inclusion of these PacBio reads in Ydb was driven by the presence *changuu*. To test this we masked Ydb with *changuu* and calculated that it accounts for ~1.3% of Ydb bases, suggesting that this conglomerate sequence is highly abundant on the Y chromosome.

We were not able to find any PacBio reads that contained the complete ~16 kb 2Rj LCR. Fragments of the 4-kb spacer that separates the 5.3 kb inverted repeats are present in the PacBio reads, but we were unable to identify any reads that contain the full spacer (**Fig. S9D**). At the 5'-end of the inverted repeat, which contains the AGAP013547 homolog, degradation on the Y has resulted in the loss of the first one-third of this gene, as evident from spatial mapping (**Fig. S9D**). We found the 3'-end of the AGAP013547 homolog (294-411bp) in 536 of 772 Ydb PacBio reads but only 2 of these also contained the 5'-end (1-294bp). By comparison, when looking in the entire PacBio data set, both ends co-occurred much more frequently (in 284 of 615 reads). This suggests that this 5' truncation is typical of the Y-linked copies and not copies found elsewhere in the genome, including 2Rj. RNA-Seq data suggest that AGAP001087, its homologs, and AGAP013547 are not abundantly expressed; we were unable to find any male specific 20-mers.

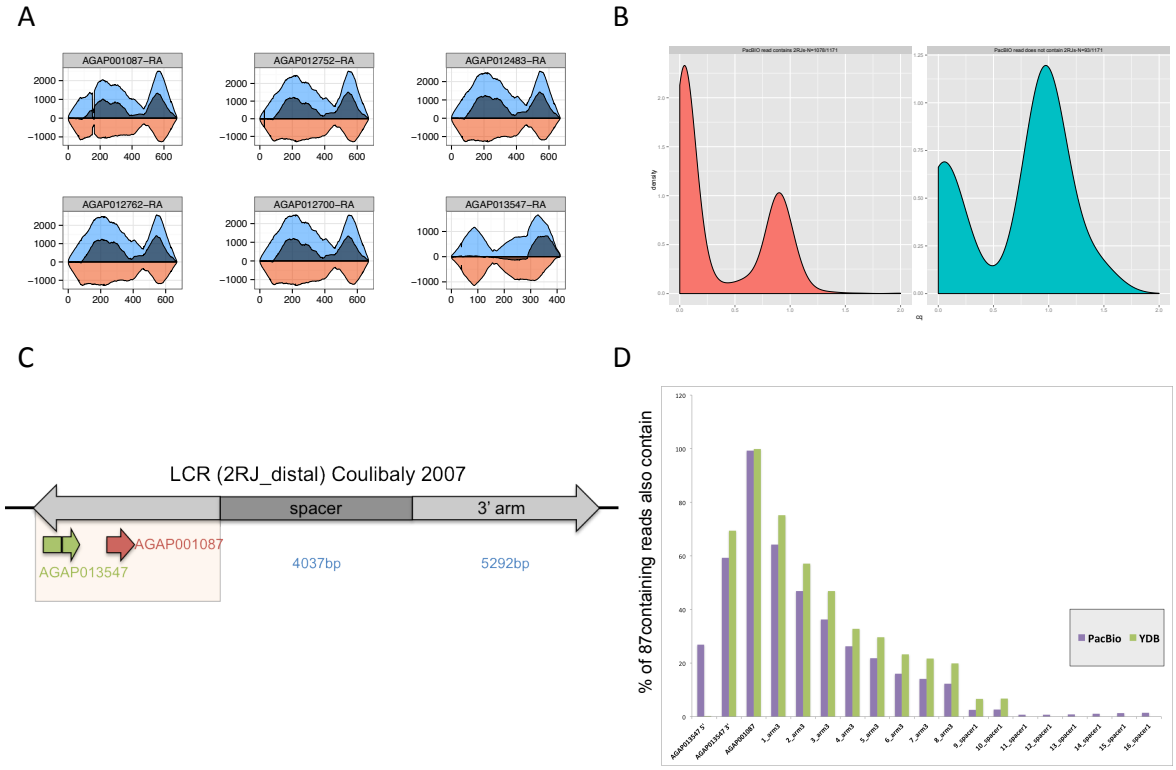

Figure S9. **Computational analysis of *changuu***. (A) Spatial mapping of: AGAP001087, its annotated but unplaced homologs, and AGAP013547. Genomic reads are from males (light blue) and females (light red); the male-female difference is a darker shade of blue or red. (B) Density plot of CQ values for PacBio reads containing hits to both AGAP001087 and the masked 2Rj LCR sequences in red or no significant hits to the LCR in blue, highlighting that when present in PacBio reads, *changuu* drives CQ reductions. (C) Schematic representation of the *2Rj* LCR showing the location of the AGAP001087 and AGAP013547 homologous sequences in the 5'-arm. (The 3'-arm is a mirror image, but the data are not shown). The breakpoint in AGAP013547 that is Y-specific is shown. Sequences upstream of the breakpoint are absent from the Y chromosome as seen in panel A. (D) Results of BLAST searches indicating a distance-dependent reduction in spacer matches in PacBio reads from Ydb (green) versus the entire PacBio data set ("PacBio"; purple). Notice the truncation of the 5'-end of the AGAP013547 homolog, which is specific to Ydb reads.

S4. Fluorescence *in situ* hybridization and estimating size of the Y chromosome

S4.1. Fluorescence *in situ* hybridization

S4.1.1. Methods

Mosquito strains: Laboratory colonies examined for this study were provided by the Malaria Research and Reference Reagent Resource Center (MR4). These included multiple colonies of *An. gambiae* (Pimperena, Asembo, Kisumu, Zanu), and *An. arabiensis* Dongola, *An. quadriannulatus* Sangwe, and *An. merus* MAF. Mosquitoes were reared at 27°C, with 12:12 light:dark cycle and 70% relative humidity.

Chromosome preparation: Most of the physical mapping was done on metaphase and prometaphase mitotic chromosomes sourced from leg and wing imaginal discs of early 4th instar male larvae. However, because the Y and X chromosomes are morphologically similar in *An. merus*, it was more difficult to verify whether imaginal disc preparations were from males (i.e., contained both X and Y chromosomes). Accordingly, for *An. merus* our preparations were derived from testes of late stage pupae, and mapping was done on metaphase I meiotic chromosomes. Chromosomes were prepared following ref. (26). Larvae or pupae were immobilized on ice for 10 minutes. Dissections were performed on microscope slides in a drop of cold freshly made hypotonic solution (0.075M KCl). A fresh drop of hypotonic solution was added to the preparation for 10 minutes, followed by fixation in a drop of modified Carnoy's solution (ethanol:glacial acetic acid, 3:1) for 1 minute. Next, a drop of freshly prepared 50% propionic acid was added and the preparation was immediately covered with a 22x22 mm coverslip. After 5 minutes, the preparation was squashed and dipped in liquid nitrogen for cover slip removal, followed by sequential dehydration steps in 70%, 80% and 100% ethanol. Slides with the highest number of metaphase plates were chosen for FISH.

Probe preparation and FISH: Genomic DNA was isolated from virgin males of the *An. gambiae* Pimperena strain using DNeasy Blood and Tissue Kit (Qiagen Inc., Valencia, CA, USA). Probes to perform FISH were generated by incorporating fluorescent labels during a PCR reaction. Primers for amplifying satellites AgY53A, AgY53B and AgY477 were obtained from ref. (22). For other targets, primers were designed using Primer 3 (27) (**Table S10**). For PCR labeling of satellites or TEs, each 25 μl PCR mix consisted of 35-40 ng genomic DNA, 0.3 U Taq polymerase, 1× PCR buffer, 200 μM each of dATP, dCTP, and dGTP, and 65 μM dTTP, and 0.5 μl Cy3-dUTP or 0.5 μl Cy5-dUTP (*Enzo* Life Sciences, *Inc.,* Farmingdale, NY, USA) or 0.5 μl Fluorescein-dNTP (Thermo Scientific, Waltham, MA, USA). Thermocycling was performed using ImmoMix ^{TM} (Bioline USA Inc., Taunton, MA, USA) beginning with a 95°C incubation for 10 minutes followed by 35 cycles of 95°C for 30 sec, 55°C for 30 sec, 72°C for 30 sec; 72°C for 5 min, and a final hold at 4°C. To prepare the *YG5* probe, a fragment of *YG5* was first amplified by PCR. Primers were then removed using Wizard SV Gel and PCR Clean-Up System (Promega, Madison, WI, USA). The PCR product was labeled by nick-translation: each 50 μl reaction contained 1 μg of DNA; 0.05 mM each of dATP, dCTP, and dGTP, and 0.015 mM of dTTP (Fermentas, Inc., Glen Burnie, MD, USA); 1 μl of Cy3-dUTP or 0.5 μl Cy5-dUTP (Enzo Life Sciences, Inc*.,* Farmingdale, NY, USA; 0.05 mg/ml of BSA (Sigma, St. Louis, MO, USA); 5 μl of 10x nick translation buffer; 20 U of DNA polymerase I (Fermentas, Inc., Glen Burnie, MD, USA); and 0.0012 U of DNAse I (Fermentas, Inc., Glen Burnie, MD, USA). FISH was performed as described previously (26).

Image acquisition and processing: After FISH, chromosomes were counter-stained with DAPI-antifade (Life Technologies, Carlsbad, CA, USA), kept in the dark for at least 2 hours, and visualized on an Olympus BX61 fluorescent microscope using BioView software (BioView Inc.,

Billerica, MA, USA) at 1000x magnification. Individual channels of the same plate were merged and the brightness and contrast were adjusted (applied to an entire image) using Adobe Photoshop.

Laser capture microdissection of Y chromosomes and chromosome painting: Imaginal discs from 4th instar larvae of *An. gambiae* Pimperena were dissected in a drop of hypotonic solution (0.5% sodium citrate) and incubated for ~10 min. Imaginal discs were transferred to methanol:glacial acetic acid (3:1) fixative solution. Fixative solution was replaced by 60% acetic acid for tissue maceration. The solution was pipetted to homogenize the tissue to cell suspension condition. A membrane slide PET (Zeiss USA, Thornwood, NY, USA) was placed on a cold metal plate (-20°C) and ~10 µl of cell suspension was dropped onto the slide, allowing liquid to spread. The slide was placed on a heating table (45°C) to condense the drop. Cooling and heating were repeated five times, and then slides were left on the heating table to complete evaporation of liquid. Slides were dehydrated in 100% ethanol. After air-drying, Giemsa staining solution (1 ml of Giemsa to 10 ml sterile H_2O) was added for 5 min, then excess stain was washed with sterile H_2O. Slides were air-dried and stained in a controlled sterile climate to avoid contamination. Microdissection was performed using the PALM MicroBeam Laser Microdissection system (Zeiss USA, Thornwood, NY, USA) and the PALMRobo software as described previously (28). About 10-15 Y-chromosomes were microdissected from each slide by catapulting individual chromosomes without prior cutting. The captured Y chromosome material was dissolved in 10-15 µl of 1x PBS buffer and processed with the GenomePlex Single Cell WGA4 Kit (Sigma-Aldrich Co. LLC., St. Louis, MO, USA) protocol to produce the first library of amplified DNA. DNA was purified using the Genomic DNA Clean & Concentrator Kit (Zymo Research Corporation, Irvine, CA, USA). Amplified DNA was labeled for FISH using

the GenomePlex WGA3 Reamplification Kit (Sigma-Aldrich Co. LLC., St. Louis, MO, USA) using dNTP mix, described previously for labeling in PCR (28). FISH of labeled microdissected Y chromosomal DNA to mitotic chromosome squash preparations of *An. gambiae* Pimperena was performed according to ref. (26).

Table S10. Primers designed to amplify DNA probes for FISH.

Target	Primer Name	Sequence (5'-3')
mtanga	YAR3F1	CATGGGTTTCATCCTGCTCT
	YAR3R1	TCAATACTGCCCTTCCGAAC
zanzibar	YAR2F2	TTCTTCGATGTTGTGCTGGA
	YAR2R2	ATGGAGAAACAGGGCAACAA
zanzibar	YAR4F1	ATGCATGCTTGGATTCCTTC
	YAR4R1	GGTTTCTATGATCGCCTGGA
zanzibar	YAR5F1	TTGGCATTCATCTGTCCAAA
	YAR5R1	GCACCCTTGATCTCATGTCA
YG5	*YG5*F	GACAGACCGACGGAGTAAGC
	*YG5*R	CATGCCCGAGTGCATAAGTA
AgY53A	AgY53AF2	ATGAAGAATATGGATAATGGAT
	AgY53AR3	ACGGGAGAGAGCAAGAACA
AgY53B	AgY53BF	CCTTTAAACACATGCTCAAATT
	AgY53BR	GTTTCTTCATCCTTAAAGCCTAG
AgY477	AgY477F2	TTTGAGCATGTGTTTAAAGG
	AgY477	AGGTTTTCCCGAGTACAAT

The primers for AgY53A, AgY53B, and AgY477 amplify satellite monomers as described in ref. (22). The PCR probes included fragments for a total size of ~400-500 bp for AgY53A, AgY53B and 477-900 bp for AgY477.

S4.1.2. Results.

Y satellites and *zanzibar*. Satellite DNA repeats AgY53A and AgY53B showed similar patterns of hybridization to the *An. gambiae* Y chromosome (Fig. 4 and **Fig. S10**). Both probes were interspersed throughout ~70% of the Y chromosome on FISH images. In *An. arabiensis*, AgY53B showed an *An. gambiae*-like pattern of hybridization to the Y chromosome, but in *An. quadriannulatus*, hybridization was limited to a small region of the Y, and in *An. merus* AgY53B hybridized to multiple heterochromatin blocks that were stained by DAPI. AgY477 also showed an interspersed pattern in *An. gambiae* and *An. arabiensis,* but coverage and signal intensity was less than that observed for AgY53B. All the satellite DNA probes hybridized to both X and Y chromosomes. This can be due to the fact that the X and the Y share the same satellite monomers. However, this result can also be due to sequence similarity between different satellite monomers, which we could not control for in our FISH experiments, as our FISH probes contained the entire satellite monomer (in contrast to our computational analyses; see SI Appendix S5). For example, there is extensive sequence identity between AgY373/AgY477 and the X-biased satellite AgX367 (see **Fig. S6**) which would allow for cross-hybridization. In contrast to the satellite DNA, *zanzibar* hybridized exclusively to the Y chromosome in *An. gambiae* and *An. quadriannulatus*, and the signal covered ~50% of the Y chromosome. In *An. arabiensis*, *zanzibar* hybridized to a much smaller area of the Y chromosome (~20%), and *zanzibar* hybridization was not detected in *An. merus*.

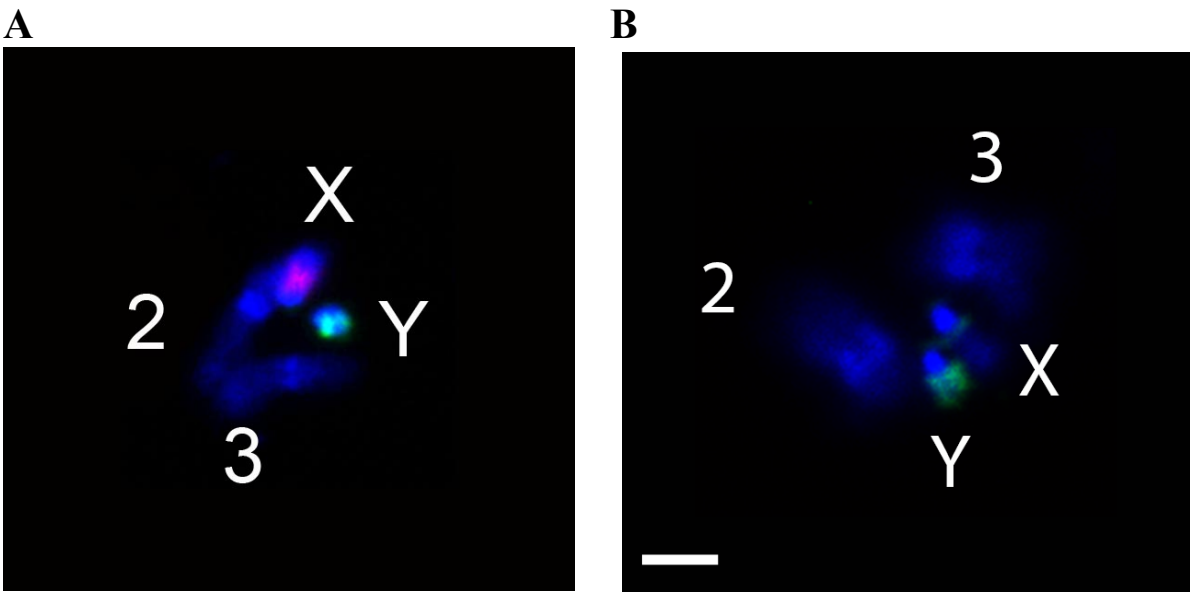

Figure S10. Physical mapping of Y chromosome-biased satellite DNA sequences in *An. gambiae.* (A) Satellite AgY53A (green signal) hybridized to most of the Y chromosome but not much to the X chromosome where the 18S rDNA probe (red signal) was seen in the *An. gambiae* Kisumu strain. (B) Satellite AgY477 (green signal) hybridized to a larger portion of the Y chromosome and to a smaller portion of the X chromosome in the *An. gambiae* Zanu strain. Chromosomes were counterstained with DAPI (blue). The scale bar is 2 μm.

***YG5*.** Our computational evidence was consistent with a Y chromosome location of *YG5* in *An. gambiae,* but an X chromosome location of *YG5* homologous sequences in *An. arabiensis* and *An. quadriannulatus*. FISH with *YG5* gave results consistent with this expectation (Fig. 4). *YG5* exclusively hybridized to a single location on the Y chromosome in *An. gambiae*. In *An. arabiensis* and *An. quadriannulatus, YG5* homologs were detected on the X and chromosome 2, but not on the Y.

Chromosome painting with microdissected Y chromosomes. FISH experiments with *An. gambiae* Pimperena based on microdissected and labeled whole Y chromosome sequence painted virtually the entire Y chromosome (Fig. 5A). Moreover, a substantial portion of X chromosome heterochromatin also was strongly painted. In addition, faint labeling of autosomal heterochromatin could be seen. This result demonstrates a substantial sequence similarity between sex chromosome heterochromatin, which is not shared to the same extent between the Y chromosome and any autosome. The FISH signal is likely driven by repetitive DNA (unique or low copy number sequences generally produce no, or very faint, signals); if so, this result is in good agreement with our computational evidence.

S4.2. Estimated size of the Y chromosome.

We measured sex chromosomes of metaphase plates from *An. gambiae* Pimperena males. Mitotic chromosome spreads were prepared from imaginal discs as described above. Images of DAPI-stained chromosomes obtained from confocal microscope Zeiss LSM 880 (Zeiss USA, Thornwood, NY, USA) were measured using ZenLite software. An important source of chromosome length variation is the degree of X chromosome condensation, which increases from prometaphase to metaphase. In order to minimize this variation, we excluded prometaphase stages from our calculations. Metaphase chromosomes were selected if they had well-separated

homologous chromosomes and identifiable sister chromatids. Prometaphase chromosomes, which had joined homologs and merged chromatids, were discarded.

Our strategy was to estimate the size of the Y chromosome in relation to the size of the assembled portion of the X chromosome. First, we measured the X chromosome from the telomere to the start of the ribosomal locus, which roughly corresponds to the portion (~25 Mb) that is assembled in the *An. gambiae* AgamP4 reference genome. We refer to this measurement as ($X_{assembled}$). The rest of the X chromosome is unassembled, as it is heterochromatic like the Y chromosome. Next, we measured entire (full-length) X and Y chromosomes from each metaphase plate. The formula used to estimate the size of the Y chromosome based on this set of measurements was (25 Mb*Y length in μm)/($X_{assembled}$ length in μm). Our calculations were based on 50 mitotic metaphase plates obtained from 8 *An. gambiae* Pimperena males. We found that the size of metaphase Y is approximately 36.7 Mb on average. The estimated size of the Y chromosome varied from 25.9 Mb to 47.8 Mb (**Table S11**). The size variation could be due to intraspecific polymorphism in the amount of the Y chromosome repeats and/or varying degree of chromosome condensation within metaphase.

Table S11. Estimated size of the *An. gambiae* Y chromosome.

Male No.	Repeated size estimates per male (Mb)										Mean
1	26.62	28.05	29.96	32.73	32.93	33.22	35.17	35.54	35.64		32.21
2	29.27	30.38	36.44	36.97	38.17	44.12					35.89
3	27.42	36.38	37.59	38.56	38.99	41.18					36.69
4	33.48	33.70	35.62	35.94	36.17	47.83					37.12
5	38.87	40.84	42.00	42.45	42.51						41.33
6	43.04	43.17	43.37	45.60	46.85						44.41
7	44.28	47.52	47.61								46.47
8	25.93	27.40	30.14	30.61	31.60	33.22	33.73	33.74	34.68	38.10	31.92
Overall											36.71

S5. Variation of Y repeats within *An. gambiae* and among species in the *An. gambiae* complex

S5.1. Copy number variation in individuals from a natural population of *An. gambiae*

We analyzed Illumina genomic sequences from 40 individual males and 45 individual females to screen for variation in the abundance of genes and repeat features on the Y chromosome in a natural population sample from Cameroon. We aligned the sequences from each of the individual samples to consensus sequences of all Y-linked genes and repeat sequence types in Ydb, using Bowtie and allowing no mismatches over the entire length of the read (-v 0). For reference, we also aligned the individual sequences to a known single-copy gene on the X chromosome, *white* (29). To prevent over-counting, reads that aligned to a single consensus sequence multiple times were only counted once. In addition, due to regions of 100% sequence identity between AgY477 and AgY373 (see **Fig. S6**), the same Illumina read could align to both satellite consensus sequences. Such reads were counted only toward AgY477 alignments in **Tables S12-S14**. Accordingly, counts attributed to AgY477 in these tables (and in **Fig. S11**) should be considered as proxies for the combined AgY477/AgY373. **Table S12** reports the median number of reads mapping to each consensus sequence across the 40 male and 45 female samples, after normalizing for sample library size (to reads per million) and locus length. For each of the individual samples, we report the normalized number of mapped reads (**Tables S13-14; Other Supporting File 2**). There was little variation in the number of alignments among individual male or female samples for genes and most Y chromosome repeats (**Fig. S11**, **Tables S13-14; Other Supporting File 2**). However, we observed that two Y satellite repeats exhibited remarkable variation between individual males in the number of mapping reads (**Fig. S11**, **Table S13; Other Supporting File 2**). Tandem Repeats Finder (30) identified these sequences as

satellite monomers of 53 bp and 280 bp, respectively (AgY53D and AgY280), which occur interspersed in the same arrays (Fig. 2C; **Fig. S7**).

The maximum and minimum number of normalized unique alignments to AgY53D differed by >200x among individual male *An. gambiae* (Fig. 3, **Fig. S11**, **Table S13; Other Supporting File 2**). The median number of reads aligned was 6236 (**Table S12**). Six individual males had alignments in excess of ~5x the median number, and one exceptional individual had ~44x the median number. In this individual male, AgY53D had a comparable number of alignments to AgY477/Ag373 that together comprise >40% of the sequence in Ydb. To rule out that the expansion of AgY53D occurred on the autosomes or X, we also looked for similar patterns of expansion in females. Most females had zero alignments to AgY53D, and the maximum number of alignments to AgY53D in any female sample was 63 (**Table S14; Other Supporting File 2**). Therefore, we conclude that the amplification of AgY53D likely occurred on the Y chromosome. This dramatic diversity of satellite abundance between individual *An. gambiae* males likely underlies the observed polymorphism in Y chromosome size.

Table S12. Median number of reads from male and female *An. gambiae* from Cameroon aligned to reference Y sequences. The number of alignments were normalized to the reads per million from each sample and the length of each feature.

Name	**locus type**	**Median number of reads in males**	**Median number of reads in females**	**CQ**
YG1	gene	413.62	2.05	0.00
YG2	gene	383.42	0.00	0.00
YG3	gene	33.71	16.06	0.48
YG4	gene	0.00	0.00	N/A
YG5	gene	2656.85	7.09	0.00
YG6	gene	11.69	20.58	1.76
YG7	gene	90.33	77.04	0.85
YG8	gene	4816.32	63.03	0.01
zanzibar	TE	121456.28	308.10	0.00
mtanga	TE	26402.57	55.59	0.00
pemba	TE	22108.57	102.23	0.00
mafia	TE	15715.18	695.36	0.04
bawe	TE	414.94	268.16	0.65
tumbatu	TE	1901.88	581.34	0.31
chumbe	TE	9852.54	172.46	0.02
uzi	TE	9585.47	1411.60	0.15
latham	TE	1310.11	0.00	0.00
changuu	TE	1578.17	1599.09	1.01
AgY280	stDNA	378.54	0.00	0.00
AgY373	stDNA	297.30	35.31	0.12
AgY477	stDNA	334960.56	1365.26	0.00
AgY53A	stDNA	193358.34	18.62	0.00
AgY53B	stDNA	72255.83	11.80	0.00
AgY53D	stDNA	6236.39	0.00	0.00

YG9	candidate gene from *merus*	114.62	119.61	1.04
YG10	candidate gene from *merus*	114.39	110.91	0.97
YG11	candidate gene from *merus*	94.04	115.48	1.23
YG12	candidate gene from *merus*	104.19	102.42	0.98
YG13	candidate gene from *merus*	118.55	124.23	1.05
white	reference single copy gene on the X chromosome	73.37	156.75	2.14

Reads were mapped individually for each sample using Bowtie with the k -1 parameter to avoid double counting reads and v -0 to include only reads that map over their entire length with 100% identity. Median values were then calculated for males and females separately. We normalized the data to reads per million for each sample. Read mapping data for all individual males and females are provided in Other Supporting File 2.

Table S13. Normalized number of reads from male *An. gambiae* from Cameroon aligned to a representative subset of Y sequences. The full data are provided in Other Supporting File 2. The number of alignments were normalized to the reads per million from each sample and the length of each feature.

Name	YG2	AgY477	AgY373	AgY53B	AgY53A	AgY53D	AgY280	zanzibar
22	383	387004	8374	87362	364967	3504	137	126317
23	338	281545	23561	52475	153867	1969	804	206160
24	614	315342	248	63755	199759	3497	157	108177
25	375	477715	414	53270	116858	38682	40820	139687
26	384	176528	448	75296	216147	2619	126	108075
27	344	201791	300	85189	192136	6189	1501	120859
28	417	536713	187	70785	109243	3304	961	108351
29	333	699638	162	40784	123324	2590	96	124266
30	368	358883	272	116442	86351	7382	4405	189906
31	377	203951	312	69103	131204	4016	385	75636
32	598	207328	216	122234	356627	4338	168	121977
33	360	210922	295	112867	492091	10272	80	110549
34	343	270550	237	81314	325015	4794	88	101283
35	276	379628	205	110934	183365	1222	177	131071
36	338	238602	260	80143	144874	74589	71080	159490
122	417	391415	359	52899	121259	3651	470	143233
123	423	162161	212	60602	91501	3684	173	133786
124	366	466993	259	43453	28327	30140	27735	140034
125	541	86501	362	103817	406284	100352	117083	127148
126	533	84247	294	99521	389353	96498	113135	122402
127	350	351446	380	75266	262638	8595	185	92348
128	413	792264	469	41398	104764	5256	2042	125937
129	374	230557	231	50466	287037	7810	493	117296
130	337	510385	169	108060	194581	6964	98	127420
131	346	486559	216	64041	195258	6679	336	110380
132	365	191654	349	95036	434640	15112	284	81940
134	426	402457	409	45484	124564	5247	5362	109586
135	382	539585	333	73726	130490	3399	130	128405
136	457	291267	294	44507	288880	6462	946	88547
222	922	321023	203	93830	134524	6847	1753	204880
223	345	509836	339	39750	45237	17907	15898	145958
224	453	401870	544	28384	65987	1867	266	120935
225	422	227596	517	61511	160303	6284	953	111241
226	389	356868	393	86259	238241	6807	841	97658
227	418	124870	280	32446	59464	2296	226	90153
228	367	535110	313	117624	368288	4807	74	106140
229	415	262808	402	48703	308803	6826	372	105492

230	386	145126	238	100741	236432	275718	298089	130867
231	430	413157	274	97647	329902	7491	215	120824
233	390	348898	321	70256	200251	5124	30	125614

Table S14. Normalized number of reads from female *An. gambiae* from Cameroon aligned to a representative subset of Y sequences. The full data are provided in Other Supporting File 2. The number of alignments were normalized to the reads per million from each sample and the length of each feature.

Name	gYG2	AgY477	AgY373	AgY53B	AgY53A	AgY53D	AgY280	zanzibar
64	1	1993	41	15	19	5	14	248
65	0	793	48	0	0	0	0	203
66	0	782	69	0	0	0	0	224
67	0	1388	165	1	0	4	0	164
68	0	1962	14	1	0	0	0	152
69	0	724	6	3	9	0	0	168
70	0	3347	256	12	36	4	7	360
71	0	1884	24	1	0	0	0	378
72	0	2178	19	12	35	0	0	260
74	0	921	8	0	0	0	0	388
75	0	1726	19	0	4	0	0	326
76	0	1096	68	5	19	0	0	271
87	0	1106	31	0	0	0	0	324
88	0	555	35	26	56	14	4	307
89	0	1568	24	25	152	2	0	371
90	0	517	24	0	0	0	0	208
91	0	649	6	0	0	0	0	343
93	1	873	29	19	6	8	5	284
94	0	75400	891	0	0	0	0	237
95	0	963	18	17	38	29	0	227
96	0	2168	7	0	0	63	0	357
191	0	893	27	0	5	0	0	283
192	0	882	82	2	0	0	0	298
193	1	2109	22	27	26	0	0	305
194	1	1113	32	41	105	4	7	328
196	0	847	18	0	0	0	0	317
197	1	1012	49	34	67	2	0	302
198	0	340	77	0	0	0	0	213
199	0	1987	22	49	123	2	4	328
200	0	884	98	44	101	6	4	409
201	0	1084	35	25	92	6	4	298
202	0	1365	27	37	46	10	0	293
280	0	2063	93	70	172	61	58	330
282	0	1336	106	24	65	3	0	287
283	0	1847	84	61	191	42	76	386
284	0	416	54	5	17	0	0	336
285	1	14173	43	36	117	0	0	455
286	0	35564	188	2	0	0	0	308

287	1	14472	65	16	32	30	24	361
288	0	754	32	3	10	2	3	341
290	0	1627	29	0	0	0	0	244
291	0	3286	60	34	139	0	0	313
292	0	44435	232	22	53	2	0	390
294	0	17278	44	57	285	24	41	487
295	0	1901	49	15	48	0	0	312

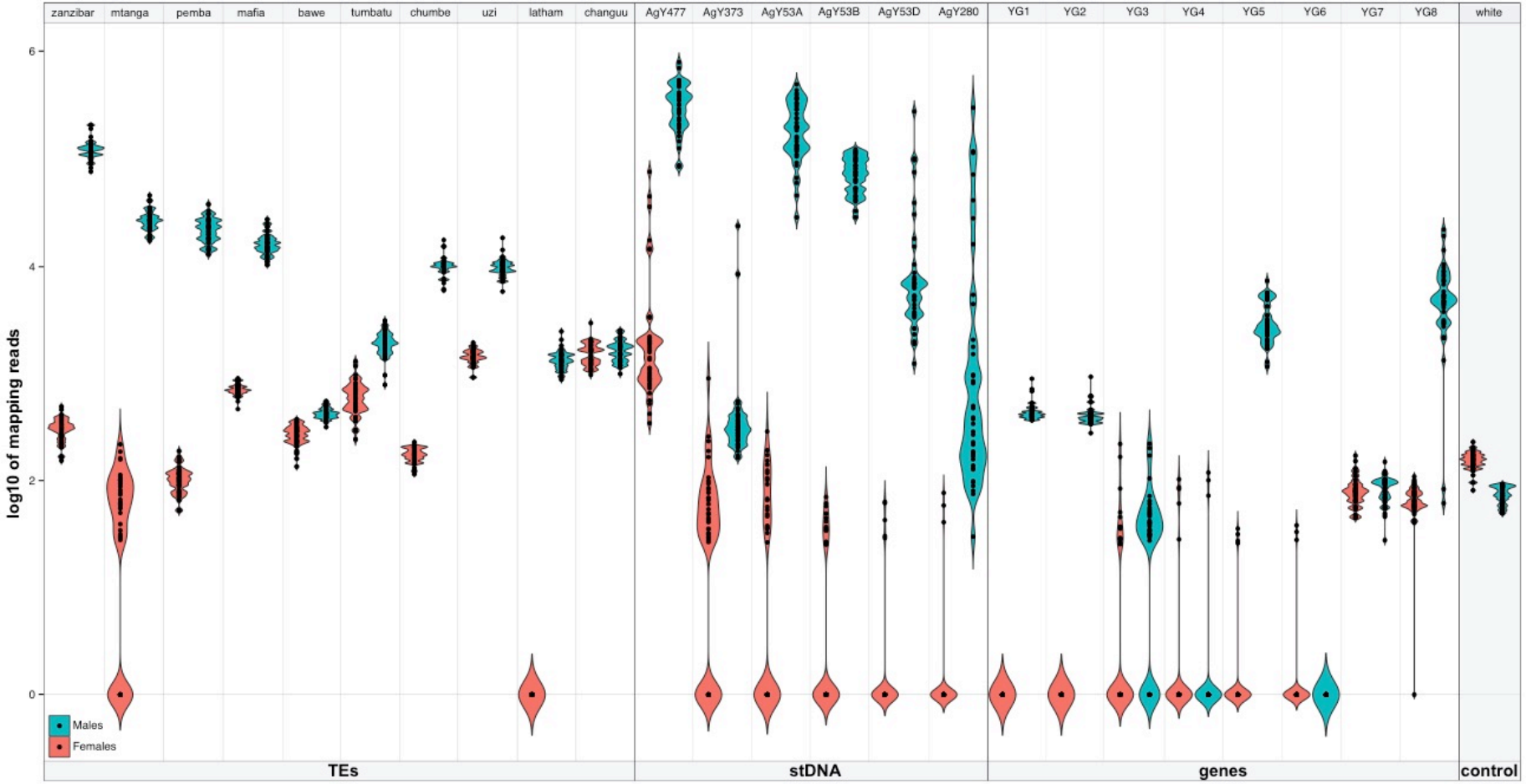

Figure S11. Violin plots of $\log_{10}$ number of Illumina WGS normalized reads from individual male (blue) and female (red) *An. gambiae* mosquitoes from Cameroon mapped to all Y loci described. Reads were mapped individually for each sample using Bowtie with the k -1 parameter to avoid double counting reads and v -0 to include only reads that map over their entire length with 100% identity. Read numbers are normalized by library size and locus length and a minimum cutoff of 25 normalized reads per gene was imposed for a call to be made. Also shown, as a control, are the number of mapping reads against the X-linked single copy *white* gene of *An. gambiae*

S5.2. Repeat variation across the *An. gambiae* complex

Dramatic variation in the presence or absence and abundance of Y repeats was observed across the *An. gambiae* complex. To examine these differences we used strict mapping with Bowtie and relaxed mapping with BWA-MEM using male and female Illumina sequences from three strains of *An. gambiae* and three other sibling species from the *An. gambiae* complex (*An. arabiensis*, *An. quadriannulatus*, and *An. merus*; see *SI Appendix* S1.4-1.5).

We used BWA-MEM to align male and female Illumina sequences from *An. arabiensis*, *An. quadriannulatus*, and *An. merus* to consensus sequences from the major repeat classes of the *An. gambiae* Y, to assess their presence and relative abundance in this recent species radiation. Species in the *An. gambiae* complex are very closely related, and because BWA-MEM allows for mismatches and gaps, Illumina sequences from other members of the complex should align to the *An. gambiae* consensus sequences. However, in the specific case of satellite sequences, we used BLASTN instead of BWA-MEM for interspecific comparisons. For the 53-bp satellites, we used BLASTN with the parameters: (word_size 7 and evalue 1e-10 and max_target_seqs 100000000). After BLASTN, we removed alignments of less than 60 bp and removed reads that aligned more than once, using the parameters awk and sort –u, respectively.

To evaluate the presence, absence, and relative abundance of Y repeats across the *An. gambiae* complex, we calculated CQ using the BWA-MEM and BLASTN alignments for all the major Y repeat classes. We also used relaxed alignment parameters to compensate for the nucleotide differences between the sibling species. To perform these alignments, we used BWA-MEM with default parameters. BWA-MEM can perform gapped alignments with several mismatches, which is ideal for applications involving less than 100% sequence identity. We call the ratio of female to male alignments with relaxed parameters the “relaxed” chromosome quotient (RCQ), to distinguish this approach from CQ. BWA-MEM did not perform well for the

53-bp satellites or satellite-specific regions of the closely related monomers AgY477/AgY373 because these sequences were very short, so we performed interspecific alignments to these satellites using BLASTN word_size 7.

S5.2.1. AgY477 and AgY373

AgY477 and AgY373 are not Y-enriched in *An. merus*. We identified two segments of AgY477 and AgY373 that together define each satellite monomer and distinguish both from a closely related satellite monomer on the X chromosome, AgX367: (i) a 90-bp insertion found in both AgY477 and AgY373 (blue portion of **Fig. S6B**), and (ii) the last 85 bp of AgY373 (yellow portion of **Fig. S6B**). We used these two segments as our references for AgY477 and AgY373 in these analyses to prevent cross-alignments to sequences that are X-biased in *An. gambiae*. [Note that this approach was not followed in our physical mapping. As probes for FISH were derived from full-length monomers, the Y-biased satellites (AgY477 and AgY373) are expected to hybridize to the X chromosome by virtue of sub-sequences that are shared (at 100% identity) between these two satellites and the X-biased AgX367; see **Fig. S6**]. We found that both AgY477 and AgY373 are extraordinarily male-biased in *An. gambiae*, but present in approximately equal numbers in males and females of *An. merus*. This indicates that while in *An. gambiae* these satellites are primarily located on the Y, in *An. merus* they may be located on both the X and the Y or the autosomes (**Table S15**).

AgY477 and AgY373 are not present in *An. quadriannulatus*. Using BLASTN against the *An. quadriannulatus* Illumina data, we were unable to find sequences that aligned to either the 90 bp insertion into AgY477/AgY373 or the last 85 bp of AgY373 (**Table S15**). All the alignments to AgY477 and AgY373 in *An. quadriannulatus* appear to derive from AgX367-

related sequence outside of these segments—sequences that are not male-biased in *An. quadriannulatus*.

Using BLASTN against the *An. arabiensis* Illumina data, we could not find sequences that aligned to the 90 bp insertion into AgY477 and AgY373 in *An. arabiensis*. Furthermore, the last 85 bp of AgY373 appears to be truncated and alignments to this sequence are not male-biased (**Table S15**).

In summary, it appears that only in *An. gambiae* are AgY477 and AgY373 Y-enriched. In other members of the *An. gambiae* complex they are either absent altogether (*An. quadriannulatus* and *An. arabiensis*) or may be present on both X and Y or even the autosomes (*An. merus*).

S5.2.2. AgY53A and AgY53B

In *An. gambiae* and *An. arabiensis,* AgY53A and AgY53B are extremely male-biased. In contrast, in *An. merus* AgY53A and AgY53B are not Y-enriched; they are present equally in males and females (**Table S15**; **Table S7**). This result is supported by FISH, where AgY53B hybridizes equally both to the X and Y chromosome in *An. merus* (Fig. 4). AgY53B is much less abundant on the *An. quadriannulatus* Y, compared to other members of the *An. gambiae* complex (**Table S15**; **Table S7**). Using BLASTN with the parameters described above there were only 546 alignments to AgY53B from the male *An. quadriannulatus* data. This is in contrast to *An. gambiae*, *An. arabiensis*, and *An. merus* which all have from hundreds of thousands to more than a million alignments to AgY53B (**Table S7**).

S5.2.3. *zanzibar* is only amplified on the Y in *An. gambiae* and *An. quadriannulatus*.

In *An. gambiae* and *An. quadriannulatus*, the *zanzibar* consensus sequence has RCQ values of 0.066 and 0.063 compared to RCQ values of 1.033 and 0.841 in *An. merus* and *An. arabiensis,* respectively. These data suggest that there has been an approximately 10-20x increase in the amount of *zanzibar* on the Y chromosome in *An. gambiae* and *An. quadriannulatus* (**Table S15**; **Table S7**). Moreover, the *pemba* insertion into *zanzibar* appears to be shared by common ancestry between *An. gambiae* and *An. quadriannulatus*. Like *zanzibar*, the *pemba* insertion sequence has CQ values less than 0.07 in both *An. gambiae* and *An. quadriannulatus* (**Tables S6-S7**). Although *pemba* is present on the Y chromosome of *An. arabiensis* (as indicated by RCQ values and validated by genomic PCR), the *pemba* insertion into *zanzibar* appears to be specific to *An. gambiae* and *An. quadriannulatus*. We extracted 120 bp spanning both junctions, between *zanzibar* and the beginning and end of the *pemba* insertion (60 bp from *zanzibar* and 60 bp from either the 5'- or 3'-end of the *pemba* insertion). If a *pemba* insertion into *zanzibar* at this precise location was present in a species, Illumina reads should span these junctions; otherwise no junction-spanning reads should be found. To screen for junction-spanning reads, we filtered for alignment lengths >70 bp. No spanning reads were found in *An. arabiensis* or *An. merus,* yet thousands of spanning reads were found in *An. gambiae* (14,959) and *An. quadriannulatus* (15,400). We conclude that *pemba* is inserted into the exactly same *zanzibar* site in both species, *An. gambiae* and *An. quadriannulatus,* most likely due to an insertion event that preceded the lineage-splitting of these two taxa.

Table S15. Summary of repeat variation between species in the *An. gambiae* complex[1]

	An. gambiae	*An. quadriannulatus*	*An. arabiensis*	*An. merus*
ZAR[2]	Y-biased (Amplified)	Y-biased (Amplified)	Slightly Y-biased	Not Y-biased
AgY477/373[3]	Y-biased	Absent	Absent[3]	RCQ~1[3]
AgY53A[4]	Y-biased	Y-biased	Y-biased	RCQ~1[4]
AgY53B[5]	Y-biased	Mostly absent	Y-biased	Even distribution on X and Y, RCQ~1[5]
AgY53D	Y-biased[6]	Not found	Not found	Low abundance, RCQ ~ 0.6.
AgY280	Y-biased	Not found	Not found	RCQ ~0.7

[1]Summary is based on three types of data: (i) normalized reads mapped from males and females, which indicate presence/absence and relative abundance of a repeat; (ii) (R)CQ data (Tables S6-S7) which indicate Y chromosome bias; and (iii) physical mapping by FISH, to validate the bioinformatic data and rule out the possibility of repeats residing mainly on autosomes when RCQ≈1. Highlighted in red are cases supported by FISH results.

[2]The ZAR is slightly Y-biased in *An. arabiensis* and not Y-biased in *An. merus*. Based on the number of aligned reads ZAR does not appear to be amplified to the same extent in *An. arabiensis* and *An. merus* compared to *An. gambiae* and *An. quadriannulatus*. This conclusion is supported by FISH across all four species (Fig. 4).

[3]Y-biased distribution is confirmed by both RCQ (Table S7) and FISH (Fig. S10) in *An. gambiae*. In *An. arabiensis*, a sequence characteristic of AgY477/AgY373 Y-monomers in *An. gambiae* (the 90 bp insertion into AgY477 and AgY373; fig. S6B) is absent in males, although sequence homology to AgY373 is present in both males and females of this species. In *An. merus*, RCQ≈1, indicating that either AgY477/373 is present on both X and Y, or present on autosomes. Although we have no FISH data to distinguish these possibilities, either case represents radical sequence reorganization between the Y chromosomes of *An. gambiae* Y and *An. merus*.

[4]Y-biased distribution is confirmed by both RCQ (Table S7) and FISH (Fig. S10) in *An. gambiae*. In *An. arabiensis* and *An. quadriannulatus,* Y-biased distribution is indicated by RCQ. In *An. merus*, RCQ≈1suggestiong that either AgY53A is present on both X and Y, or present on autosomes. Although we have no FISH data to distinguish the two possibilities, either case represent radical sequence reorganization between the Y chromosomes of *An. gambiae* Y and *An. merus*.

[5]The pattern of AgY53B distribution is confirmed by both RCQ (Table S7) and FISH (Fig. 4) in all four species. In *An. merus*, RCQ≈1and FISH data suggest equal distribution of AgY53B on X and Y (Fig. 4).

[6]AgY53D is absent in the Asembo strain of *An. gambiae*.

S6. Y chromosome recombination

The Illumina deep genomic sequencing data from 40 and 45 individual males and females were analyzed for evidence of recombination between the X and Y chromosomes. We used Bowtie (-v 0) to perform the alignment as detailed in SI Appendix S5.

As shown in Fig. 5B, in a few individual females the relative abundance of AgY477 was exceptionally high relative to other individuals. The median number of normalized alignments from females to AgY477 was 1365 (**Table S12**). However, there were six females with more than 10x the median number, and one with 55x the median (**Table S14; Other Supporting File 2; Fig. S11**). Those females with elevated numbers of alignments to AgY477 may have acquired extra copies through illegitimate X/Y recombination.

The three 53-bp satellites (AgY53A, AgY53B, and AgY53D) showed an interesting pattern of alignments in the individual females (Fig. 5B, **Fig. S11**). While all males had hundreds or thousands of alignments to all three 53-bp satellites (**Tables S12-S13**) there were 18 female individuals with <10 alignments to all three 53-bp satellites (**Table S14; Other Supporting File 2**). In contrast, 9 individual females had >100 alignments to at least one of the satellite sequences. The most likely scenario to explain the variation in these individuals is that unequal X/Y recombination transferred some of the Y satellites to the X, consistent with additional evidence presented next.

An alternative hypothesis to X/Y recombination could be sperm contamination in the female samples, or laboratory contamination of female with male DNA. Both forms of contamination should affect all Y sequences, yet we do not observe a corresponding increase in female alignments to other Y chromosome satellites (**Table S14**). For example, in individuals with an elevated number of alignments to AgY477, there are not correspondingly elevated

numbers of alignments to the 53-bp satellites. Also, in many cases where one of the 53-bp satellites has an elevated number of alignments, one or both of the other 53-bp satellites do not have an elevated number of alignments. To further rule out contamination of female with male DNA, we examined the number of female alignments to a segment of *YG2* that is exclusive to the Y. All 45 female samples had <1 normalized alignment to *YG2* (**Table S14; Other Supporting File 2**).

Independent evidence of X/Y recombination comes from PacBio reads containing both the predominately X-linked (AgX367) and predominately Y-linked (AgY477/AgY373) satellites (22). Two PacBio reads with alignments of >98% identity to both AgY477/AgY373 and AgX367 were identified (Fig. 5C). Taken together with evidence of arrays containing recombinant monomers composed of portions of X-biased and Y-biased satellites (SI Appendix Text S3.1; **Fig. S5**), these results indicate that occasionally the X-biased satellite AgX367 recombines with the Y-biased satellites AgY477 and AgY373, a conclusion consistent with the high degree of sequence similarity between all three satellite monomers (**Fig. S6**).

S7. Small and labile genic repertoire

S7.1. Identifying candidate Y-linked genes in *An. gambiae*

Y chromosome genes are often located within complex arrays of repetitive DNA, making identification difficult. Gene-finding on the Y is further complicated by the potential absence of Y-linked genes in genome assemblies, and by cases where there is substantial sequence similarity between genes on the Y and homologs present on other chromosomes. To circumvent these complications and provide as comprehensive a catalogue of genes on the Y chromosome as possible, we implemented multiple, complementary strategies. For clarity, we catalogue a locus as a gene, if this locus is actively expressed based on RNA-Seq data and only if the sequence does not bear significant similarities to known anopheline repeats. In a few cases we also reported loci as genes, when expression levels were relatively low, but there was high similarity to an already annotated *An. gambiae* gene in VectorBase. The gene finding strategies we developed differed from each other in at least two important ways: 1) the sequences in which we searched for genes: genome assemblies, RNA-Seq reads, *de novo* transcriptome assemblies, or PacBio reads; and 2) whether we required the mean CQ of the gene to fall below a certain threshold. We were able to validate Y chromosome linkage of some genes using further experimental approaches, such as male specific genomic PCRs, or *in situ* hybridizations to mitotic chromosomes. For some genes however, experimental validation beyond our computational evidence was not possible. These are considered in our work as unconfirmed Y genes. The evidence of Y-linkage and male-biased expression of these genes are summarized in CQ and RCQ tables (**Table S6-S7**), spatial mapping analysis (**Fig. S12**), and expression profile analysis (**Fig. S13**).

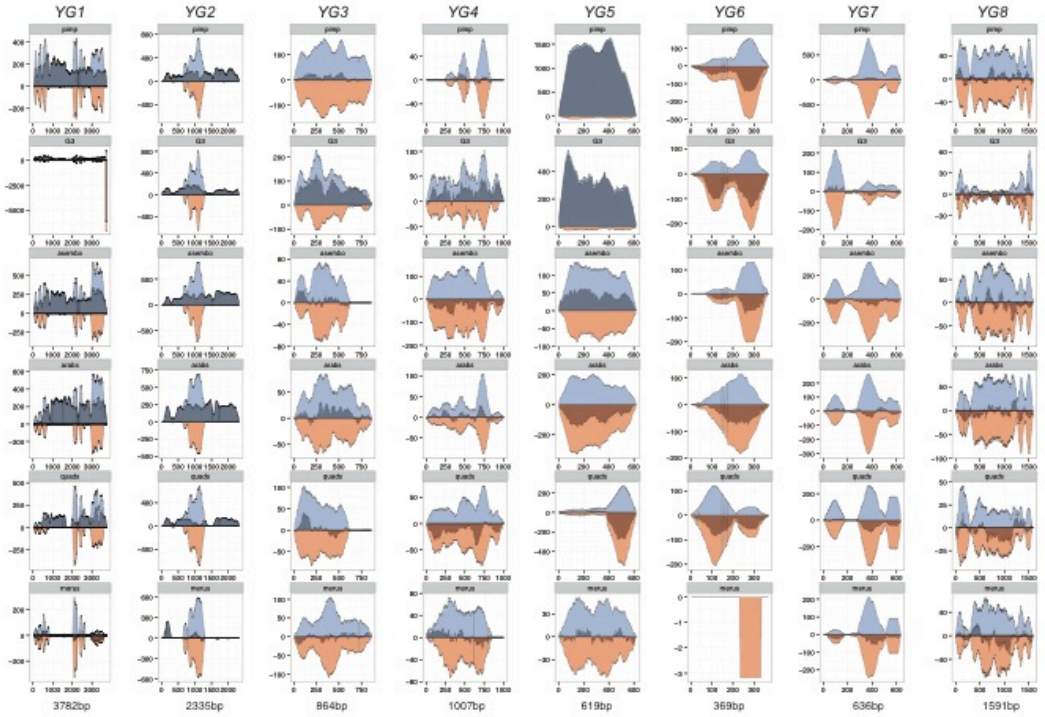

Figure S12. **Spatial mapping of male and female reads to the consensus sequences of the *An. gambiae* Y chromosome genes (plotted start to end on the x axis).** Male reads are shown in light blue in the positive y-axis, female reads are shown in light red in the negative y-axis and the difference between male and female reads is shown in dark blue (for male bias) or dark red (for female bias). For each row, reads from different species or strains is used: pimp: *An. gambiae* Pimperena; G3: *An. gambiae* G3; asembo: *An. gambiae* asembo; arabs: *An. arabiensis*; quads: *An. quadriannulatus*; merus: *An. merus.*

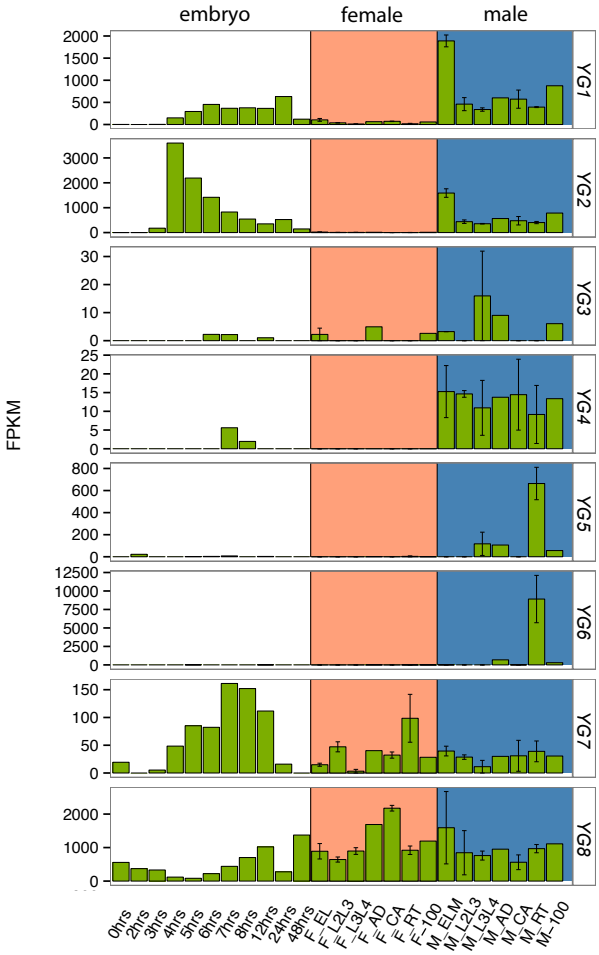

Figure S13. **Expression profile analysis of *An. gambiae* Y chromosome genes.** Normalized counts of reads mapping to each of the verified or candidate Y chromosome genes. Embryonic stages (from wild type embryos) are shown in hours post egg laying, F_EL: female early larvae; F_L2L3: female L2-L3 instar larvae; F_L3L4: female L3-L4 larvae; F_AD: female adults; F_CA: female dissected carcass without ovaries; F_RT: female ovaries; M_EL: male early larvae; M_L2L3: male L2-L3 instar larvae; M_L3L4: male L3-L4 larvae; M_AD: male adults; M_CA: male dissected carcass without ovaries; M_RT: male reproductive tissues (testes and accessory glands). F_100 and M_100 are *in silico* merged datasets of male- and female-only data.

S7.1.1. Methods and results:

S7.1.1.1. CQ-based gene finding:

A given sequence typically displays a mean CQ lower than 0.2 if it is either specific to the Y chromosome (*e.g*., for single copy genes), or if it is present on the Y in many more copies than elsewhere in the genome (for multi-copy genes). Using the criterion of mean CQ ≤ 0.2 (as we have throughout this work) we initially searched for Y-linked genes by a number of strategies briefly described below, and highlight the genes uncovered by each.

To identify Y genes in Ydb, we first ran RepeatMasker using the latest available *An. gambiae* repeats library in VectorBase, updated with the major Y chromosome repeat classes. We split the resulting repeat masked Ydb sequences into sections at masked bases and removed sequences smaller than 200 bp. Then we calculated CQs for the resulting sequences using male and female genomic Illumina data from *An. gambiae* Pimperena and G3, and retrieved sequences with CQ ≤ 0.2 having at least 20 read alignments from males. To identify transcribed loci we then aligned RNA-Seq data from our libraries. The genes *YG1*, *YG2* and *YG5* were identified in several PacBio reads using this method.

To identify genes in transcriptome assemblies, we performed *de-novo* assembly of RNA-Seq samples using Trinity (31). We built an initial set of libraries independently using 14 sequenced samples of male-derived RNA-Seq data, but later repeated this with the entire dataset and our gene-finding results did not differ. Trinity was run using default settings for paired end RNA-Seq except for requiring a minimum contig length of 150bp (-- min_contig_length 150). To identify Y chromosome sequences from these Trinity assemblies we applied the CQ method with Pimperena and G3 Illumina sequences, and collected predicted transcripts with CQ ≤ 0.2 having at least 20 read alignments from males. We identified 1,893 predicted transcripts that

passed this criterion. To distinguish those transcripts that are not TE-like, we first removed sequences that were masked by RepeatMasker using the *An. gambiae* repeats library as above. We also performed BLASTX against the NCBI non-redundant database and removed any sequences with alignments to bacteria or transposable elements. To eliminate repetitive sequences and select only those that are unique to the Y chromosome, we removed sequences with more than 10 alignments with BLASTN to the autosomes or X in the PEST assembly. Finally, we used our RNA-Seq dataset to remove sequences without male-specific expression. While the predicted transcripts were often fragmented in the assemblies, we were able to find *YG1* and *YG2* in all of the transcriptome assemblies made from libraries after 2 hours of embryonic development. Furthermore, we identified *YG4* in transcriptome assemblies of post-embryonic samples, and *YG5* from samples derived from adult males.

To identify genes in the PEST assembly, paired RNA-Seq reads were mapped to the PEST assembly using Tophat (32). To identify novel transcribed regions (NTRs) we ran Cufflinks *de novo* (32) using the assembled PEST genome along with corresponding gene sets (in GTF) as a reference. We generated an expanded merged transcriptome, containing all previously known genes (AGAPs) and non-redundant NTRs. When this analysis was performed, AgamP3.7 was the latest available gene set and it contained 13465 loci and 15322 transcripts with a total length of 27747499 bp. The Cufflinks *de novo* analysis added a large number of novel transcribed regions: 11340 loci, 13974 transcripts and 19829594 bp. Because these NTRs fell mainly under the class of repetitive elements (see below) and the underlying genome assembly remained mostly unchanged, for logistical reasons we did not repeat such a *de novo* assembly with later VectorBase gene set updates. We combined the annotated transcriptome with the NTRs and re-mapped all RNA-Seq datasets independently to identify male-biased or male-

specifically expressed genes using the DESeq package (16). Mean CQ was then calculated for the male-biased merged transcriptome. In total, 79 loci appear both male-biased in expression and had a CQ ≤ 0.2. The majority of these were on the Y-unplaced or the UNKN chromosome, some were in autosomes and none were mapped to the X chromosome. We subsequently performed BLASTN and ran RepeatMasker against the repeat library to eliminate those loci that are likely TEs. This method identified *YG4* (annotated on the UNKN), and *YG5* (as a duplication putatively arising from AGAP13757) and re-identified *YG1* and *YG2* in the Y_unplaced scaffold.

S7.1.1.2. Non-CQ based methods: identifying non-unique Y genes in Ydb

All pipelines for identifying Y linked loci described above require that mean CQ falls at or below our selected cutoff of 0.2. This is effective for finding loci that are either relatively specific to the Y (see below), or for those that are more abundant on the Y than in other chromosomal locations (e.g. *YG5*). It is important to note that the CQ method was successfully used to identify Y-linked genes that have autosomal homologs of >95% nucleotide identity because 100% identical alignments of short reads can differentiate highly similar sequences (9). Nonetheless, to identify possible Y-linked genes that have nearly identical or a large number of non-Y paralogs, which may be missed by CQ analysis, we took advantage of our long reads database, with the expectation that neighboring regions of such loci could provide sufficient evidence for their Y linkage, a strategy of "guilt by association". We searched using BLASTN within Ydb for homologs of known *An. gambiae* transcripts (AgamP4.4) with a cutoff e-value of 1e-50 and a minimum sequence identity of 80%. To avoid BLAST hits being predominantly mediated by poor annotation that includes TEs, we first masked transcripts using the *An. gambiae* repeat library. This filter eliminated 19 of the 66 transcripts with alignments to Ydb. We used two

criteria to consider genes as good candidates for inclusion. First, we counted the number of unique Ydb reads per transcript and only transcripts that had hits to more than five Ydb sequences were considered (*n*=16, **Table S16**). Second, a good Y chromosome candidate would be a transcript that commonly co-occurs on the same PacBio read with major Y chromosome repeat features in Ydb, but does so much less frequently across the entire PacBio dataset.

Of the 16 annotated transcripts, three are homologous to *YG1* and/or *YG2* (AGAP004894-RA; AGAP001079-RA; AGAP001078-RA) and one is homologous to *YG5* (AGAP013757-RA); these four were not characterized further. Among the 12 remaining annotated transcripts, we discovered two cases of redundancies arising from multi-gene families, different members of which shared sequence similarity to the same PacBio reads. While all members of these families were included in subsequent analysis, for simplicity we refer to these two families (comprising 4 and 6 annotated genes) by the names of their two Y-linked members (*YG6* and *changuu,* respectively). Through this analysis, we identified three more candidate genes, *YG6-YG8*, and the *changuu* repeat (**Table S16**).

Table S16. Sixteen transcripts with more than 5 hits to Ydb after repeat masking

	Y locus	number of times in YDB	paralogue	Chr	length	%blast length	Read CQ_C	Read CQ_Rel	Gene CQ_C	Gene CQ_Rel	Spatial Mapping	Match to Repeats	%length repeat
1	*YG5*	>500	AGAP013757	3R	618	100.97	0.32	0.08	0.26	0.06	male_biased	#N/A	#N/A
2	*changuu*	>500	AGAP012762	UNKN	668	71.71	0.07	0.32	0.62	0.59	male_biased	Agam_m8bp_Ele13	6.59
3		>500	AGAP012752	UNKN	669	71.9	0.18	0.41	0.87	0.6	male_biased	Agam_m8bp_Ele13	6.58
4		>500	AGAP012700	UNKN	668	71.86	0.18	0.41	0.76	0.61	male_biased	Agam_m8bp_Ele13	6.59
5		>500	AGAP012483	UNKN	668	71.71	0.07	0.32	0.62	0.59	male_biased	Agam_m8bp_Ele13	6.59
6		>500	AGAP001087	X	680	72.21	0.07	0.32	1.36	0.62	male_biased	Agam_m8bp_Ele13	6.47
7		358	AGAP013547	X	411	62.04	0.24	0.68	1.25	0.81	male_biased	#N/A	#N/A
8	*YG1 and YG2*	76	AGAP004894	2L	726	27	0	0.26	0.99	0.99	non_biased	#N/A	#N/A
9		72	AGAP001079	X	727	26.96	0	0.26	1.08	0.99	non_biased	#N/A	#N/A
10		72	AGAP001078	X	727	26.96	0	0.26	1.08	0.99	non_biased	#N/A	#N/A
11	*YG7*	16	AGAP010291	3R	636	72.64	0.01	0.07	1.44	0.96	non_biased	#N/A	#N/A
12	*YG6*	7	AGAP013235	X	540	83.89	0.19	0.71	2.32	2.03	female_biased	#N/A	#N/A
13		7	AGAP012536	UNKN	757	71.47	0.23	0.8	2.04	1.89	female_biased	#N/A	#N/A
14		6	AGAP013428	X	582	88.66	0.19	0.71	1.7	1.92	female_biased	#N/A	#N/A
15		4	AGAP013444	X	570	83.68	0.1	0.7	2.25	2.12	female_biased	#N/A	#N/A
16	*YG8*	6	AGAP008501	3R	1591	60.28	0.12	0.68	1.01	0.94	non_biased	#N/A	#N/A

S7.1.2. Description of Y Genes

YG1 and ***YG2*** were previously described; both RNA-Seq and RT-PCR indicate early-embryonic expression starting from 2-4 hours after oviposition (9). The two genes are overlapping and expressed from opposite strands so unique query sequences or primers were used to differentiate the expression patterns between the two genes. Using our new and expansive RNA-Seq datasets we confirmed that these two genes are abundantly expressed in males throughout development, both in somatic and reproductive tissues. *YG2* expression begins at 3 hours after egg-laying and reaches its maximal expression in the male at 4 hours of embryonic development. For *YG1*, expression begins at 4 hours after egg-laying, one hour later than *YG2*, but remains low during embryogenesis. Expression peaks at the transition between embryo and larvae (**Fig. S13**).

There are a number of sequences similar to *YG1* and *YG2* throughout the PEST assembly including some annotated genes (9). To gain an understanding of chromosomal distribution of sequence similarity between these Y-linked sequences and their counterparts on other chromosomes, we performed BLASTN analysis against the PEST assembly and the assemblies of the other three species (**Fig. S14**). We also performed CQ, RCQ and spatial mapping with these genes (**Table S6-S7; Fig. S12**). We find that Y-specificity of these sequences is not uniform throughout the length of the transcripts (**Fig. S12, S14**). The first two exons of *YG2* occur uniquely on the Y chromosome, which overlaps with the PCR amplicon sequences we used for phylogenetic reconstruction (SI Appendix S8). CQ, RCQ and spatial mapping using male and female WGS Illumina sequences from the other species revealed that *YG2*—and importantly the exons that encode the putative male-determining factor—are present on, and unique to, the Y chromosome in all species examined. *YG1* is present on the Y of *An. arabiensis* and *An. quadriannulatus,* but evidence of Y-linkage is ambiguous in *An. merus* (**Fig. S12**).

We have evidence that *YG2* exists on the Y in multiple copies nearly identical in sequence. In the multiple sequence alignment of a 755 bp segment of *YG2* (PCR-amplified from individual male mosquitoes sampled from natural populations and colonies, and used for phylogenetic reconstruction (SI Appendix S8; Other Supporting File 3), two positions—335 and 427—were polymorphic in Cameroon samples of *An. gambiae* and the same positions also were segregating for the same polymorphsims in our RNA-Seq data and Illumina sequences from different sources (*An. gambiae* G3 and Pimperena), indicating that the polymorphisms are unlikely to be sequencing error or PCR artifacts, and suggesting that they are transcribed. Next, we retrieved 73 Ydb PacBio reads overlapping the 755 bp *YG2* segment, and added them to the alignment. We observed both polymorphic sites represented multiple times among the PacBio reads. **Table S17** summarizes the observed frequencies in all our data sets.

Remarkably, sites 335 and 427 were highly correlated in the 73 PacBio sequences, visible by eye (**Fig. S15**) and tested statistically (after discarding minor alleles, $D = 0.16$; Correlation test $P < 2.2e\text{-}16$; Fisher exact test $P = 5.03e\text{-}11$). The powerful advantage of the PacBio Ydb reads is not merely their added length, but also the fact that they are derived from a single paternal Y chromosome, as the template DNA for PacBio sequencing was derived from sons descended from a single pair mating of *An. gambiae*. Accordingly, each PacBio read is not only one haplotype, but each haplotype observed necessarily represents a different copy of the *YG2* locus, assuming no Y chromosome mutational events between generations and mitotic replications.

It is important to note that these sequence differences at position 335 and 427 between *YG2* copies are downstream from the presumptive 56-aa coding region, and therefore would not be expected to affect the encoded protein sequence. However, both mutations appear to affect

5'-splice sites, thus they potentially cause alternative splicing patterns. The overall very high degree of nucleotide similarity between these copies of *YG2* suggests that gene conversion may act to correct mutational differences and preserve function, as observed for amplified genes on human and ape Y chromosomes (33).

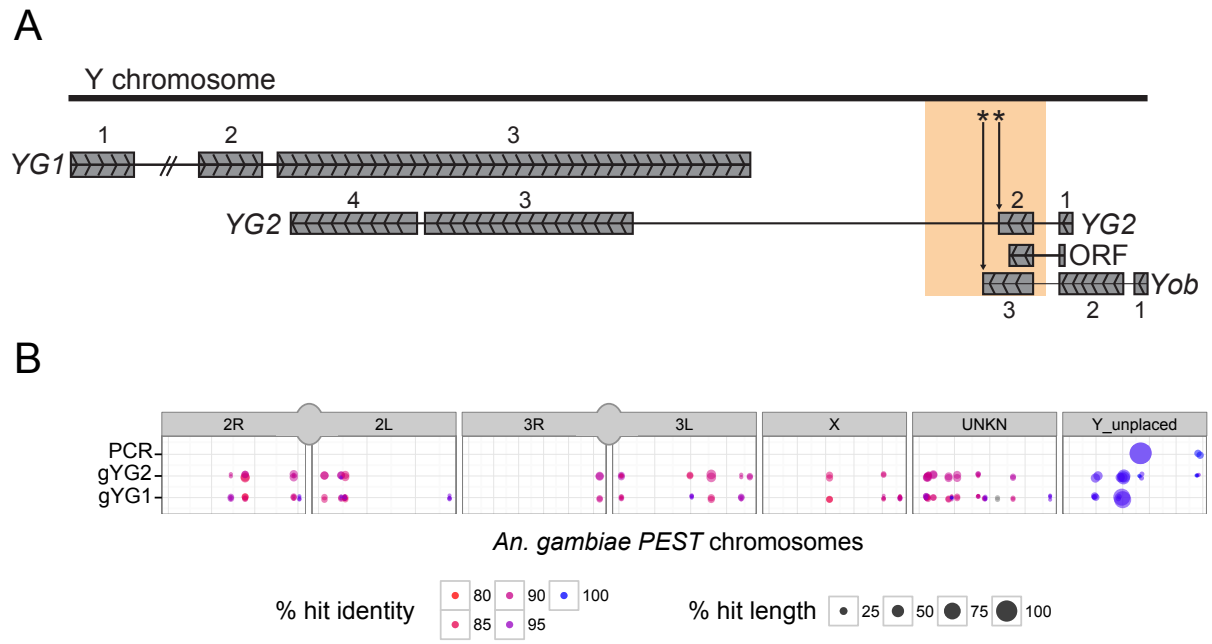

Figure S14. **BLASTN analysis of *YG1* and *YG2* against the *An. gambiae* PEST assembly.** A) Overlapping locus organization of *YG1* and *YG2* highlighting the location of the ORF and the PCR amplicon used for phylogenetic reconstruction (orange box). Numbers on top of transcripts indicate exon numbers. The asterisks indicate positions of polymorphisms that lead to alternative splicing. B) Results from BLASTN hits of the *YG1*, *YG2* transcript sequences and the sequence of the PCR amplicon used for phylogenetic reconstruction against the *An. gambiae* PEST assembly. The results show that the region used for PCR is Y specific, while partial but never complete regions of *YG1* and *YG2* have hits elsewhere in the genome.

Table S17. Polymorphisms in *An. gambiae YG2* sequence observed in 4 data sets

	***YG2* Alignment Position**[1]	
***An. gambiae YG2* sequence source**	**335**	**427**
PCR amplicons (Cameroon field samples)	3 C, 14 T	4 G, 13 A
Illumina genomic sequence (Pimperena)	33 C, 38 T	62 G, 45 A
RNA-Seq (G3)	88 C, 131 T	18 G, 30 A
Ydb PacBio reads (Pimperena)	52 C, 21 T	54 G, 19 A

[1] *YG2* sequence alignment provided as Other Supporting File 3.

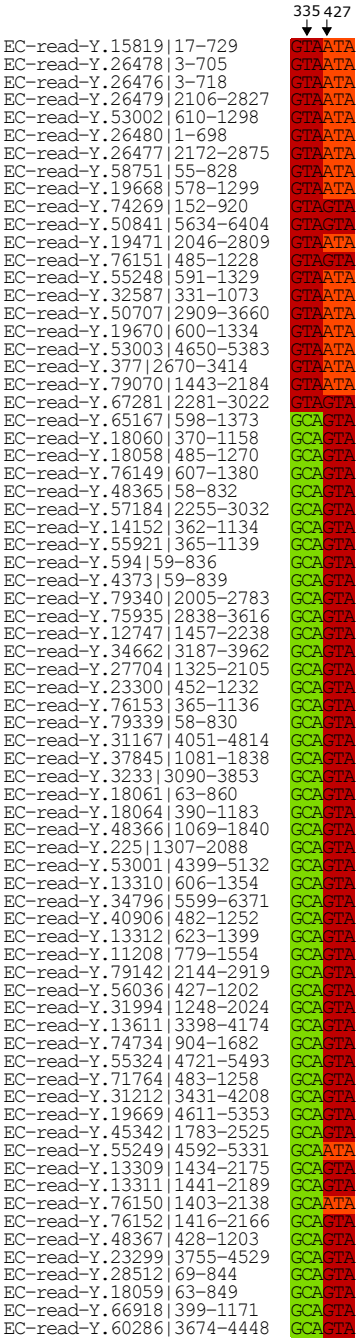

Figure S15. **Sequence alignment of 73 Ydb PacBio reads, showing only the two non-adjacent *YG2* polymorphic sites, positions 335 and 427.** The color-coded columns display alignment positions 334-336 and 427-429, respectively, with respect to the *YG2* sequence alignment given in Other Supporting File 3.

YG3 was originally reported in ref. (9). Interestingly, *YG3* is on the Y chromosome of *An. gambiae* only in the G3 colony, based on evidence from CQ and spatial mapping. Similar computational results did not support Y linkage of this gene in the Pimperena and Asembo colonies (**Tables S6-S7**). Unsurprisingly, given the Pimperena origin of Ydb, querying this database with *YG3* by BLASTN returned no significant hits. The gene does not encode any apparent ORFs and is 864 bp in length. Expression analysis using RNA-Seq data from the G3 strain indicates that this locus is expressed at low levels, detectable mostly in male L3-L4 instar larvae and whole adults. We did not detect expression in our dissected tissues (carcass and reproductive tissues) so it is currently unclear where within the male this gene is expressed. Male-specific genomic and RT-PCR confirmed the presence and expression of *YG3* on the Y of the G3 strain (9).

YG4 was identified with two CQ-based approaches: in the pool of Trinity-predicted transcripts and in the loci of the PEST assembly that displayed low CQ and male biased expression. The Trinity-predicted transcripts corresponding to *YG4* were inaccurately split into two fragments, so we anchored the predictions made by Cufflinks on the PEST assembly (Unknown_Chr: 14402957-14404138) to assemble the gene model for *YG4*. Similar to *YG3*, *YG4* appears to be Y-linked only in the G3 strain of *An. gambiae* based on CQ and spatial mapping (**Tables S6-S7**; **Fig. S12**). Therefore BLASTN against Ydb with *YG4* as a query recovered no significant hits. RNA-Seq data supported male-specific expression of *YG4* in all post-embryonic developmental stages, albeit at weak levels compared to other Y-linked genes (**Fig. S13**).

YG5 was identified with all gene-finding approaches. *YG5* is homologous to the autosomal gene AGAP013757 on chromosome 3. Because orthologs of AGAP013757 have preserved synteny with neighboring genes in the assemblies of *An. gambiae*, *An. arabiensis, An. quadriannulatus* and *An. merus* (**Fig. S16**), we suggest that AGAP013757 is the ancestral copy of this gene and that *YG5* arose through duplication onto the Y chromosome. *YG5* likely encodes a DNA-binding protein based on the presence of topoisomerase and zinc-RING finger domains. Both *YG5* and its autosomal homolog AGAP013757 are expressed exclusively in the male testis based on RNA-Seq data and RT-PCR (Fig. 1, **Fig. S13**), indicating that *YG5* may be important for male fertility, possibly by regulating chromatin condensation during spermatogenesis and male meiosis. Compared to AGAP013757, the *YG5* sequence contains a number of male-specific SNPs that could not be detected in PCR products amplified from females using primers that anneal to AGAP013757 (and also match *YG5*; data not shown). We used these *YG5*-specific SNPs to confirm male-specific expression, by sequencing RT-PCR products from male testis (data not shown). *In situ* hybridization of *YG5* to mitotic male chromosomes also confirmed Y linkage, and revealed that *YG5* is located at the tip of the Y chromosome (Fig. 1, **Fig. S17**).

CQ and spatial mapping analysis of the region surrounding AGAP013757 on chromosome 3 suggested that the duplication event that copied AGAP013757 to the Y chromosome (*YG5*) also resulted in the copying of 1791 base pairs flanking the ORF, likely the regulatory regions that regulate sperm-specific expression. Within the subset of 898 Ydb reads that contain *YG5*, we calculated the most frequent associations with all other Y-linked loci and found that *changuu* (the Y-linked LCR-related element) co-occurred in 53% of the reads (Fig. 1B). Within these 898 reads we also found at least three other TEs from the *An. gambiae* repeats library, namely Agam_CR1_Ele15 (in 121 reads), Agam_Pao_Bel_Ele46 (in 179 reads) and

Agam_CR1_Ele7 (in 396 reads). To characterize the organization of *YG5* we selected reads containing both *YG5* and *changuu* and aligned male and female Illumina sequences from the Pimperena strain, to compare relative copy number, investigate Y-linkage, and characterize the structure and frequency of *YG5* and *changuu* junctions on representative PacBio reads. We conclude that the original duplication event resulted in the integration of *YG5* into the *changuu* locus, followed by amplification of the combined sequences on the Y chromosome. It is possible that multiple copies of *YG5* supplement gene expression, and/or help preserve *YG5* sequences from mutational inactivation on the Y chromosome as suggested above.

Interestingly, while *YG5* is Y-linked in *An. gambiae*, analysis by CQ, RCQ and spatial mapping analysis (**Tables S6-S7**; **Fig. S12**) indicates X-linkage and amplification of *YG5*-homologous sequences in *An. quadriannulatus* and *An. arabiensis.* In *An. merus,* computational evidence was inconsistent with sex chromosome linkage or amplification. We confirmed X-linkage of *YG5* homologs in *An. quadriannulatus* and *An. arabiensis* by *in situ* hybridization to male mitotic chromosomes (**Fig. S17**). To test whether the X chromosome homologs of *YG5* in *An. quadriannulatus* and *An. arabiensis* and *YG5* in *An. gambiae* are derived from a single duplication event that occurred prior to the species split or arose independently later, we searched by BLAST for *YG5* within the genome assemblies of *An. quadriannulatus* and *An. arabiensis*. We found two genes within two different (non-redundant) contigs, homologous to *YG5* orthologue in the *An. arabiensis* assembly, namely AARA005061 on contig APCN01002478 and AARA010330 on contig APCN01001040. Based on synteny we concluded that scaffold APCN01002478 and the gene within it AARA005061, represent the original autosomal position of AGAP013757. CQ analysis of these two contigs using male and female Illumina sequences from *An. arabiensis* was consistent with X-linkage of contig APCN01001040

(contig CQ=1.7). At the 3' end of this *An. arabiensis* contig we found sequences homologous to *YG5* and its junction with *changuu* present that occur on the Y chromosome of *An. gambiae.* Although CQ, RCQ and spatial mapping confirmed that *changuu* is female-biased in *An. quadriannulatus* as well as *An. arabiensis* (**Tables S6-S7**), we were not able to find contigs containing *YG5* homologs in BLAST searches of the *An. quadriannulatus* assembly. To test whether the *YG5* homolog on the X chromosome of *An. quadriannulatus* is also associated with *changuu* as it is in *An. gambiae* and *An. arabiensis*, we extracted a 120 bp sequence from Ydb reads representing the junction between the 3' end of *YG5* and *changuu* on the Y chromosome of *An. gambiae*. We then used this junction sequence as a query in BLAST searches of Illumina sequences from males and females of all three species, and calculated CQ (**Table S18**). In *An. gambiae* the *YG5-changuu* junction had matches exclusively from males confirming specificity to the Y-chromosome. In contrast, in both *An. quadriannulatus* and *An. arabiensis* the CQ of reads mapping from males and females were indicative of X linkage. We obtained the same result, when using the homologous 120 bp junction sequence from the *An. arabiensis* contig APCN01001040 (**Table S18**). This confirmed that the *YG5* homolog is located next to *changuu* also in *An. quadriannulatus.* We concluded that the insertion of a *YG5* homolog into *changuu* preceded radiation of the species complex, on an ancestral X or Y chromosome.

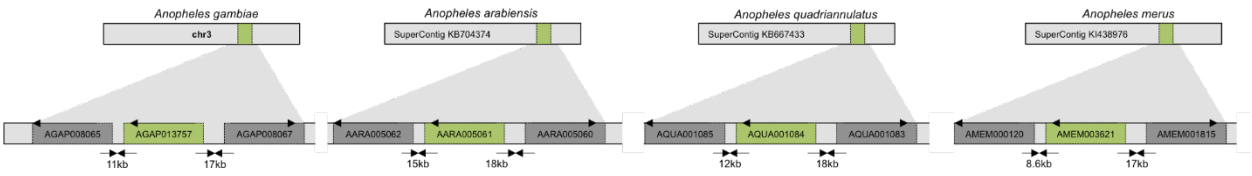

Figure S16. **Preservation of synteny of AGAP0013757, the autosomal *YG5* homolog, in genome assemblies of the *An. gambiae* species complex.**

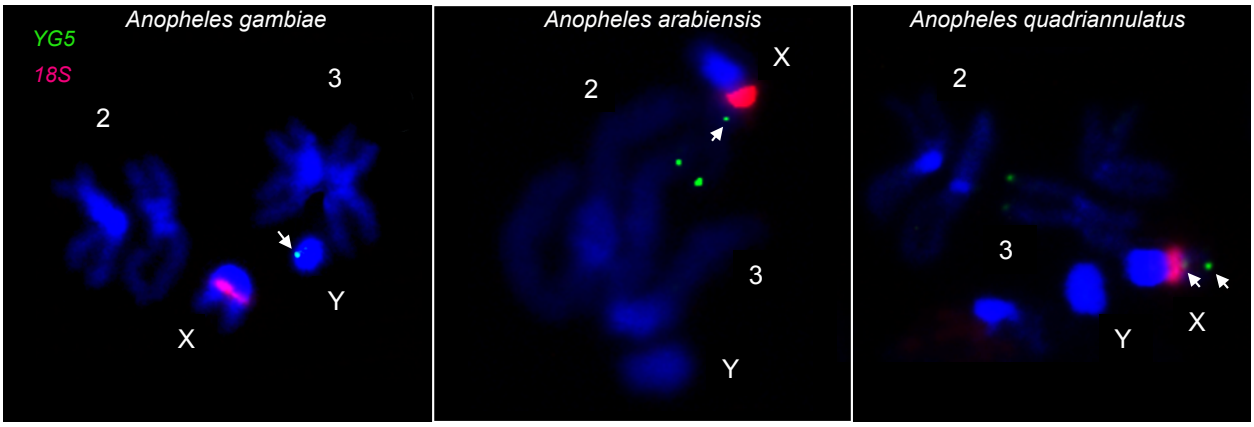

Figure S17. **Physical mapping of *YG5*.** FISH with *YG5* (green label) confirms Y chromosome linkage in *An. gambiae* and X chromosome linkage in *An. arabiensis* and *An. quadriannulatus*. The 18S ribosomal rDNA gene is shown in red.

Table S18. Quantification of *YG5-changuu* junctions indicating number of normalized reads matching to the 120-bp sequences in each sex and species with CQ values.

	An. gambiae Pimperena			***An. Arabiensis***			***An. Quadriannulatus***		
	Male	Female	CQ	Male	Female	CQ	Male	Female	CQ
Junction *YG5-changuu*	1171.31	0.00	0.00	54.38	88.00	1.62	52.01	121.00	2.33
Junction AARA010330-*changuu*	1172.32	0.00	0.00	56.55	87.00	1.54	50.95	121.00	2.37

YG6: Based on our RNA-Seq data, all six annotated transcripts from the gene family related to *YG6* (AGAP012356, AGAP012998, AGAP013104, AGAP013235, AGAP013428 and AGAP013444) are expressed exclusively in male reproductive tissues (**Fig. S18**). Of the six annotated genes, five have assigned chromosomal locations in a repetitive region of the X chromosome at ~17 Mb, consistent with mean gene CQ of 1.7 and 2.7 in the *Pimperena* and G3 strains, respectively (**Table S6**). We only found seven Ydb reads containing *YG6*. The Y copy is flanked by a single *zanzibar* insertion (which drives inclusion into Ydb). To examine whether the Y-linked homolog was expressed, two *YG6*-containing Ydb PacBio reads were compared to the RNA-Seq data. Matching sequences were split into 20-mers using jellyfish (34). Equivalent kmers were generated by chopping male and female WGS and RNA-Seq Illumina sequences that aligned to the two Ydb PacBio reads. We used relaxed mapping parameters with Bowtie2 at default settings and required that each kmer occur > 50 times to be included in the analysis. We found 215 male-specific kmers (mskmers) in the RNA-Seq data, but these RNA-mskmers were not represented in genomic-mskmers. As all *YG6* paralogs of this gene are expressed exclusively in male reproductive tissues, it is possible that the 215 RNA-mskmers are in fact derived from non-Y chromosome homologs of *YG6* and that *YG6* itself is not actively expressed. Alternatively, *YG6* is not sufficiently diverged from its non-Y homologs to provide genomic mskmers, which require that a kmer does not occur in female genomic sequence.

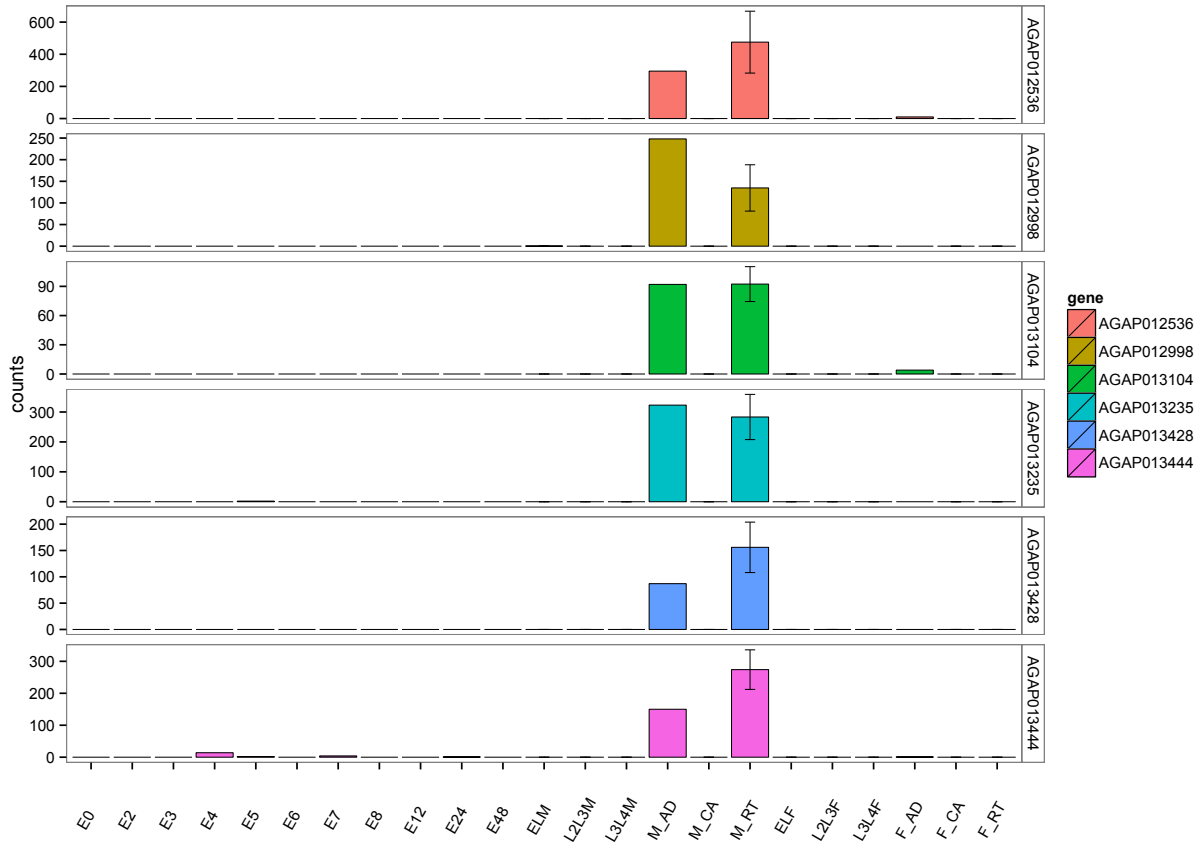

Figure S18. **Expression profile analysis of six X chromosome or autosomal homologs of *YG6*.** *YG6* homologs are expressed exclusively in male reproductive tissues across development. En: embyo sample and hours post egg laying; ELM: Early male larvae; L2L3M: L2-L3 instar male larvae; L3L4M: L3-L4 instar male larvae; M_AD: male adult; M_Ca: Male dissected carcass; M_RT: Male reproductive tissues; ELF: Early female larvae; L2L3F: L2-L3 instar female larvae, L3L4F: L3-L4 instar female larvae; F_AD: female adult; F_Ca: Female dissected carcass; F_RT: Ovaries 48hrs post bloodfeeding.

YG7 is homologous to AGAP010291, located on chromosome 3R with no known annotated function. Our RNA-Seq data indicates weak *YG7* expression throughout development, with a peak of expression during embryogenesis. Our evidence for Y chromosome linkage in the Pimperena strain is co-occurrence in 16 Ydb reads with *zanzibar* (out of 467 PacBio reads genome-wide that contain *YG7* homologs). Like *YG6*, there are no shared mskmers between RNA-Seq and genomic sequences, indicating either that *YG7* is not expressed from the Y chromosome in the stages of development sampled, its expression cannot be detected at the selected sequencing depth, or its sequence has not significantly diverged from non-Y copies.

YG8 is homologous to AGAP008501 which encodes a ubiquitously expressed glutaryl-CoA dehydrogenase. Evidence for Y-linkage is the presence of the gene in six Ydb PacBio reads that also contain *zanzibar* and the *mafia* TEs. In the entire PacBio dataset there are 67 reads that contain the gene, of which only 15 also contain *zanzibar*. The autosomal assembly of AGAP008501 does not contain gaps, and *zanzibar* is not present. Similar to *YG6* and *YG7,* there are no shared mskmers between RNA-Seq and genomic sequences, indicating that *YG8* is either not expressed sufficiently to be detected given the available sequencing depth, is not expressed from the Y chromosome in the stages of development we have selected, or that its sequence has not significantly diverged from copies not located on the Y chromosome. Further evidence for Y-linkage comes from deep WGS Illumina sequencing of 40 individual *An. gambiae* males from Cameroon, in which there are approximately 76x more reads of *YG8* in male samples compared to females (**Tables S12-S14**).

S7.2. Identification of Y chromosome genes in other members of the *An. gambiae* complex.

S7.2.1. Using RNA-Seq data from the *Anopheles* 16 genome project to identity Y genes in other members of the *An. gambiae* complex.

We used RNA-Seq data from the *Anopheles* 16 genome project (35) to search for Y chromosome genes in *An. merus* (SRX200223), *An. arabiensis* (SRX004365) and *An. quadriannulatus* (SRX096314). We assembled the RNA-Seq data with Trinity. Trinity was run in with the default setting for paired-end RNA-Seq data, except for the following parameters: (--min_contig_length 150). To identify Y chromosome sequences from these Trinity assemblies we used the CQ method and our male and female pooled Illumina data from these species to identify sequences with CQ $\leq$ 0.2.

To identify non-repetitive Y genes we first removed sequences that were masked by RepeatMasker using the repeats library for each species. We also performed BLASTX against the NCBI nr database and removed any sequences with alignments to bacteria or transposable elements. To further eliminate repetitive sequences we removed sequences with more than 10 alignments with BLASTN to the female-derived genome assemblies from the three species.

Using this approach we identified *YG1* in *An. arabiensis* and *An. quadriannulatus,* and *YG2* in all three species. In *An. arabiensis*, no novel Y chromosome genes were identified. In *An. quadriannulatus*, two novel candidate genes named *YG15* and *YG16* were identified. These may be exons of the same gene because they align to nearby regions of the X chromosome.

S7.2.2. Using *An. gambiae* transcripts to identity Y genes in other members of the *An. gambiae* complex.

Using BWA-MEM we mapped the male and female reads from *An. arabiensis*, *An. quadriannulatus*, and *An. merus* to *An. gambiae* transcripts. RCQs were calculated based on

these results. No *An. gambiae* transcripts with CQ ≤ 0.2 were identified in *An. arabiensis*. We identified five *An. gambiae* transcripts with CQ ≤ 0.2 in *An. merus* (*YG9-YG13*) (**Table S19**), and one *An. gambiae* transcript with CQ ≤ 0.2 in *An. quadriannulatus* (*YG14*) (**Table S20**). Due to the number of alignments in the other *An. gambiae* complex members (**Tables S6-S7**), these genes are likely the result of duplications to the Y followed by amplifications on the Y.

Table S19. RCQs of *An. gambiae* transcripts based on *An. merus* Illumina sequences

Transcript	Length	An. merus female alignments	An. merus male alignments	RCQ
YG9 (AGAP009631-RA)	2118	1540	12639	0.122
YG10 (AGAP009632-RA)	3387	2928	21514	0.136
YG11 (AGAP009633-RA)	883	676	3500	0.193
YG12 (AGAP009636-RA)	1210	983	6428	0.153
YG13 (AGAP009637-RA)	2169	1821	11713	0.155

Table S20. RCQs of *An. gambiae* transcripts based on *An. quadriannulatus* Illumina sequences

Transcripts	Length	An. quad female alignments	An. quad male alignments	RCQ
YG14 (AGAP010306-RA)	1000	493	2677	0.184

S8. Phylogeny reconstruction and coalescent simulations.

Overlapping fragments of *YG2* were PCR-amplified using male-specific primers (**Table S21**). Genomic template DNA was sourced from individual adult male and female mosquitoes sampled from colonies and natural populations, as follows: *An. gambiae* NDKO colony and Cameroon field collections in 2007 (36); *An. coluzzii* SUCAM colony; *An. arabiensis* Dongola colony and Cameroon field collections in 2007 (36); *An. quadriannulatus* SANGWE colony and Zimbabwe field collections in 1986 (37); *An. merus* MAF and Ophansi colonies. Each 25µl PCR reaction included 1x PCR buffer, 0.2mM each dNTP, 1.4mM MgCl2, 0.4µM each forward and reverse primer, 1U Taq polymerase, and 1µl of genomic DNA. Gel electrophoresis (2% agarose) revealed amplification of expected Y fragments in males but not in females. PCR products for each individual male were purified using USB ExoSAP-IT and directly sequenced on both strands using an Applied Biosystems 3730xl DNA Analyzer and Big Dye Terminator v3.1 chemistry. The CQ values of the amplicons (**Table S22**) are consistent with their Y-linkage. Sequences were concatenated and aligned using the online tool Clustal Omega (www.ebi.ac.uk/Tools/msa/clustalo/). The aligned sequences were imported into SeaView v4 (38) for manual inspection of the alignment. Maximum likelihood phylogenetic estimation was performed under a generalized time-reversible model of sequence evolution (39), using PhyML (40) through SeaView, with node support estimated by nonparametric bootstrap (100 replicates). Due to the absence of an outgroup, midpoint rooting was employed.

To assess the relative probability that the observed gene tree on the Y chromosome is due to incomplete lineage sorting versus introgression, we ran coalescent simulations without introgression. Employing the species tree and divergence times for *An. gambiae, arabiensis,*

quadriannulatus and *merus* that were defined in ref. (41), we used the program "ms" (42) to generate 1000 gene trees using the following command line arguments:

ms 4 1000 -T -I 4 1 1 1 1 -ej 1.28 3 2 -ej 1.84 4 2 -ej 1.85 1 2

In this species tree *An. gambiae* and *An. arabiensis* are not sister taxa. Therefore, to assess the probability that a gene tree placing them together (as observed for the Y chromosome) is due to ILS alone, for each simulated gene tree we asked whether these two species were sister to one another. In 62 of the 1000 gene trees *gambiae* and *arabiensis* were placed together, which indicates that the probability of seeing this topology on the Y chromosome by chance (*i.e.* ILS alone) is 0.062.

Table S21. Primers used to amplify Y-specific regions of *YG2*

Primer	Sequence (5'-3')
P345-F	CGGCGAGTGATACAGAACCC
P345-R	GAGAAGAAATCATCCAGCCATGTT
P339-F	CGATCAATAATGCGGCAGCTC
P339-R	GTTGCGGTCTGCGAAGAGAA
YC4-F	AATAATGGTTGCGTGCTGGTG
YC4-R	GTTCTGTATCACTCGCCGGT

Table S22. CQs of representative *YG2* sequences used for phylogenetic reconstruction across the *An. gambiae* complex. CQs were calculated with default parameters (Bowtie -v 0 –a) against sequenced male-specific PCR products

	An. gambiae G3			***An. gambiae Pimperena***			***An. gambiae Asembo***			***An. quadriannulatus***			***An. arabiensis***			***An. merus***		
	F	**M**	**CQ**	**F**	**M**	**CQ**	**F**	**M**	**CQ**	**F**	**M**	**CQ**	**F**	**M**	**CQ**	**F**	**M**	**CQ**
ndko	0	390	0	0	166	0	0	156	0	0	0	N/A	0	28	0	0	0	N/A
sucam	0	443	0	0	203	0	0	179	0	0	0	N/A	0	28	0	0	0	N/A
gamb1892	0	240	0	0	43	0	0	24	0	0	0	N/A	0	16	0	0	0	N/A
gamb1889	0	926	0	0	301	0	0	304	0	0	79	0	0	105	0	0	0	N/A
arab	0	136	0	0	42	0	0	19	0	0	10	0	1	391	0.0025	0	0	N/A
arab0590	0	168	0	0	52	0	0	25	0	0	44	0	1	460	0.0021	0	0	N/A
arab0592	0	145	0	0	45	0	0	19	0	0	10	0	1	412	0.0024	0	0	N/A
quad	0	23	0	0	17	0	0	0	N/A	0	970	0	0	8	0	0	0	N/A
quad0027	0	66	0	0	28	0	0	6	0	0	1133	0	0	11	0	0	0	N/A
quad0046	0	66	0	0	28	0	0	6	0	0	1162	0	0	11	0	0	0	N/A
maf	0	0	N/A	0	0	N/A	0	0	N/A	0	0	N/A	0	0	N/A	2	651	0.003
ophs0001	0	0	N/A	0	0	N/A	0	0	N/A	0	0	N/A	0	0	N/A	2	574	0.003

SI Appendix References

1. Koren S, *et al.* (2013) Reducing assembly complexity of microbial genomes with single-molecule sequencing. *Genome Biology* 14:R101.
2. Koren S, *et al.* (2012) Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat. Biotechnol.* 30:693-700.
3. Chaisson M & Tesler G (2012) Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* 13:238.
4. Treangen TJ, Sommer DD, Angly FE, Koren S, & Pop M (2002) Next Generation Sequence Assembly with AMOS. *Current Protocols in Bioinformatics*, (John Wiley & Sons, Inc.).
5. Chin CS, *et al.* (2013) Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods* 10:563-569.
6. Wood D & Salzberg S (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology* 15:R46.
7. Langmead B, Trapnell C, Pop M, & Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10:R25.
8. Garrison E & Marth G (2012) Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907 [q-bio.GN]*
9. Hall AB, *et al.* (2013) Six novel Y chromosome genes in *Anopheles* mosquitoes discovered by independently sequencing males and females. *BMC Genomics* 14:273.
10. Lobo NF, *et al.* (2010) Breakpoint structure of the Anopheles gambiae 2Rb chromosomal inversion. *Malar J* 9:293.
11. Santolamazza F, Della Torre A, & Caccone A (2004) Short report: A new polymerase chain reaction-restriction fragment length polymorphism method to identify *Anopheles arabiensis* from *An. gambiae* and its two molecular forms from degraded DNA templates or museum samples. *Am. J. Trop. Med. Hyg.* 70:604-606.
12. Galizi R, *et al.* (2014) A synthetic sex ratio distortion system for the control of the human malaria mosquito. *Nat Commun* 5:3977.
13. Dobin A, *et al.* (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29:15-21.
14. Langmead B & Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9:357-359.
15. Anders S, Pyl PT, & Huber W (2015) HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31:166-169.

16. Anders S & Huber W (2010) Differential expression analysis for sequence count data. *Genome Biol* 11:R106.
17. Krzywinski J, Nusskern D, Kern M, & Besansky NJ (2004) Isolation and characterization of Y chromosome sequences from the African malaria mosquito *Anopheles gambiae*. *Genetics* 166:1291-1302.
18. Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv:1303.3997v2 [q-bio.GN]*.
19. Carvalho AB & Clark AG (2013) Efficient identification of Y chromosome sequences in the human and Drosophila genomes. *Genome Res.* 23:1894-1907.
20. Quinlan AR & Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26:841-842.
21. Smit AFA, Hubley R, & Green P (2013-2015) (RepeatMasker Open-4.0, http://www.repeatmasker.org).
22. Krzywinski J, Sangare D, & Besansky NJ (2005) Satellite DNA from the Y chromosome of the malaria vector Anopheles gambiae. *Genetics* 169:185-196.
23. Rohr CJ, Ranson H, Wang X, & Besansky NJ (2002) Structure and evolution of *mtanga*, a retrotransposon actively expressed on the Y chromosome of the African malaria vector *Anopheles gambiae*. *Mol. Biol. Evol.* 19:149-162.
24. Fu L, Niu B, Zhu Z, Wu S, & Li W (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28:3150-3152.
25. Coulibaly MB*, et al.* (2007) Segmental duplication implicated in the genesis of inversion 2Rj of *Anopheles gambiae*. *PLoS ONE* 2:e849.
26. Timoshevskiy VA, Sharma A, Sharakhov IV, & Sharakhova MV (2012) Fluorescent in situ hybridization on mitotic chromosomes of mosquitoes. *JoVE* 67:e4215.
27. Rozen S & Skaletsky HJ (2000) Primer3 on the WWW for general users and for biologist programmers. *Bioinformatics Methods and Protocols: Methods in Molecular Biology*, eds Krawetz S & Misener S (Humana Press, Totowa, NJ), pp 365-386.
28. George P, Sharma A, & Sharakhov IV (2014) 2D and 3D chromosome painting in malaria mosquitoes. *JoVE* 83:e51173.
29. Besansky NJ*, et al.* (1995) Cloning and characterization of the white gene from Anopheles gambiae. *Insect Mol. Biol.* 4:217-231.
30. Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* 27:573-580.
31. Grabherr MG*, et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29:644-652.

32. Trapnell C, *et al.* (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols* 7:562–578.
33. Rozen S, *et al.* (2003) Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. *Nature* 423:873-876.
34. Marçais G & Kingsford C (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27:764-770.
35. Neafsey DE, *et al.* (2015) Highly evolvable malaria vectors: the genomes of 16 *Anopheles* mosquitoes. *Science* 347:1258522.
36. Cheng C, *et al.* (2012) Ecological genomics of *Anopheles gambiae* along a latitudinal cline: a population-resequencing approach. *Genetics* 190:1417-1432.
37. Collins FH, *et al.* (1988) Comparison of DNA-probe and isoenzyme methods for differentiating *Anopheles gambiae* and *Anopheles arabiensis* (Diptera: Culicidae). *J. Med. Entomol.* 25:116-120.
38. Gouy M, Guindon S, & Gascuel O (2010) SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol. Biol. Evol.* 27:221-224.
39. Yang Z (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* 39:306-314.
40. Guindon S, *et al.* (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* 59:307-321.
41. Fontaine MC, *et al.* (2015) Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science* 347:1258524.
42. Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18:337-338.