Supplemental Materials

Supplemental Materials include full methods, Supplemental Figures S1-S16, Supplemental Tables S1-S5, and Supplemental Data Files S1-S11.

Full methods

Sample script of the Y-chromosome Genome Scan (*YGS*) method

The program Jellyfish (Marçais and Kingsford 2011) is used to filter the short reads. The program YGS.pl does all other steps of the *YGS* method: extraction of *k*-mers from the filtered short reads, extraction of repetitive *k*-mers from the genomic scaffolds, and the comparison between the genomic and short-read *k*-mers, as exemplified below.

```
# step: filtering D. virilis female Illumina short reads at Phred score of 20
# input file  / format: virfem.fastq.gz / fastq (text file; compressed or
uncompressed)
# output file / format: virfem.jelly / Jellyfish binary file.
# comments: See Methods for choosing k-mer size (-m option). See Jellyfish manual
for other parameters.
zcat virfem.fastq.gz | jellyfish count -m 15 -o virfem.jelly -c 4 -s 10G -t 4 --
both-strands --quality-start=33 --min-quality=20 /dev/fd/0

# step: filtering at a minimum frequency of 5, and production of the short-read
fasta file
# input file /  format: virfem.jelly / Jellyfish binary file.
# output file / format: virfem.fasta.gz / fasta format (text file; compressed)
# comments: see Kelley et al (2010) for choosing minimum frequency cut-off.
jellyfish dump --lower-count=5 virfem.jelly | gzip -c > virfem.fasta.gz

# step: production of the bit-array representing the female k-mers, filtered by
quality and frequency
# input file /  format: virfem.fasta.gz / fasta format (text file; compressed or
uncompressed)
# output file / format: virfem.trace.gz / hexadecimal representation of the bit-
array (text file; compressed)
# comments: output file produced with the "to_Hex" function of the Bit::Vector
module (http://search.cpan.org/dist/Bit-Vector). See its manual for details.
perl YGS.pl kmer_size=15 mode=trace trace=virfem.fasta.gz

# step: production of the validating bit-array (part 1). Its use is optional. We
used Sanger traces for Drosophila, and male short reads for human data.
# input file  / format: drosophila_virilis.*.fastq.gz / fastq (text file;
compressed or uncompressed)
# output file / format: virSanger.jelly / Jellyfish binary file.
zcat drosophila_virilis.*.fastq.gz | jellyfish count -m 15 -o virSanger.jelly -c 4
-s 10G -t 4 --both-strands --min-quality=20 --quality-start=33 /dev/fd/0
# step: production of the validating bit-array (part 2).
# input file  / format: virSanger.jelly / Jellyfish binary file
# output file / format: virSanger.fasta.gz / fasta format (text file; compressed)
jellyfish dump --lower-count=1 virSanger.jelly | gzip -c > virSanger.fasta.gz
# step: production of the validating bit-array (part 3).
```

```
# input file  / format: virSanger.fasta.gz / fasta format (text file; compressed or
uncompressed)
# output file / format: virSanger.trace.gz / hexadecimal representation of the bit-
array (text file; compressed)
# comments: the file virSanger.trace.gz contains the validating bit-array.
perl YGS.pl kmer_size=15 mode=trace trace=virSanger.fasta.gz

# step: production of the bit-array representing the repetitive k-mers of the
genome (i.e., with copy number > 1)
# input file  / format: virCAF12.fasta.gz / genome in fasta format (text file;
compressed or uncompressed)
# output file / format: virCAF12.gen_rep.gz / hexadecimal representation of the
bit-array (text file; compressed)
# comments: very large scaffolds (> ~20 Mbp) are processed slowly.
perl YGS.pl kmer_size=15 mode=contig contig=virCAF12.fasta.gz

# step: final run.
# input files (formats): virCAF12.fasta.gz (fasta, compressed or uncompressed),
virfem.trace.gz (bit-array, compressed text), virCAF12.gen_rep.gz (bit-array,
compressed text), virSanger.trace.gz (bit-array, compressed text).
# output file /  format: virCAF12_virfem_virSanger.final_run / text file. Each line
contains the result of one scaffold, and is printed as processed.
# comments: Run time: 9 hours (Drosophila genome) /  15 days (human genome); see
Methods for details.
perl YGS.pl kmer_size=15 mode=final_run contig=virCAF12.fasta.gz
trace=virfem.trace.gz gen_rep=virCAF12.gen_rep.gz male_trace=virSanger.trace.gz
```

Statistical tests of gene gains and gene losses (*Drosophila* data)

Three statistical procedures were carried out to estimate and compare gene gain and gene loss rates in the *Drosophila* Y chromosome. The first procedure is the "Assumption-free", which is now possible given the availability of two Y chromosomes with well known gene content (*D. virilis* and *D. melanogaster*). Procedure 2 ("Homogeneous Gain Loss") employs the method described in Koerich et al. (2008; with *D. melanogaster* data) to *D. virilis* and to the combined *D. virilis* + *D. melanogaster* data. Procedure 3 ("Approximate Bayesian computation") applies computer simulations, as described in Koerich et al. (2008) with *D. melanogaster*, to the *D. virilis* data. All three approaches address the question "is the number of Y-linked genes increasing or decreasing?" They estimate the unbiased ratio gain rate / loss rate, and statistically test it against the null hypothesis of equal gain and loss rates, with a Poisson regression (procedures 1 and 2) or with approximate Bayesian computation (procedure 3). The two gains of the *kl-5* gene (Fig. 4) are known to be independent events (Koerich et al. 2008), and were counted as such in the three procedures.

Procedure 1: Assumption-free

Under this approach, gene gains and gene losses were estimated in the branches leading to *D. virilis* and *D. melanogaster*, which have Y chromosomes with well known gene content. This is much simpler and does not require the assumption of homogeneous gain and loss rates, which is introduced by the inclusion of species without well known Y chromosomes (see Procedure 2 below); data from these species were used only to identify the ancestral states. For example, the *JYalpha* gene is Y-linked in *D. virilis*, and autosomal in *D. melanogaster*, which could be explained by a autosome-to-Y movement (*i.e.*, a gene gain) in the *D. virilis* lineage, or a Y-to-autosome movement (*i.e.*, a gene loss) in the *D.*

melanogaster lineage. Data from *D. willistoni* and *D. ananassae* (where *JYalpha* is Y-linked) shows that there was a gene loss in the *D. melanogaster* lineage (Fig. 4 and Supplemental Fig. S4).

There were four gene gains and zero gene losses across 63 Myr in the *D. virilis* lineage and the corresponding values in the *D. melanogaster* lineage are seven gene gains and one gene loss across 63 Myr (Fig. 4), yielding a gene gain / gene loss ratio of 11.0 ($P = 0.022$, Poisson regression; 95% confidence interval: 1.4-85.2). The Poisson regression also tests the goodness-of-fit of the model to the data, which here tests the heterogeneity between the *D. virilis* and *D. melanogaster* data. It is not significant ($P = 0.33$) and hence the data from the two *Drosophila* species can be pooled (as we did) to obtain a single gain / loss ratio. We implemented these procedures in R language (R Core Team 2012) with the Supplemental Data Files S3 and S4 (respectively, *Poisson_regression.R* and *AF_vir_mel_data.txt*).

Procedure 2: Homogeneous Gain Loss

This approach was described in Koerich et al. (2008), which should be consulted for details. Briefly, as we discovered Y-linked genes in one reference species (*D. melanogaster* in Koerich et al. 2008; *D. virilis* in the present work), and check for their Y-linkage in the remaining ones, we can only estimate the gene gain rate in the phylogenetic branches connected to the reference species (called "red branches" in Koerich et al. 2008). Conversely, the gene loss rate can only be measured in the other branches ("blue branches"). The ascertainment bias in the gene gain and gene loss rates are then corrected (Koerich et al. 2008). When compared with the "Assumption-free method", this method uses more of the available information (which results in narrower confidence intervals and lower *P* values; see Table 1), but relies on the assumption that the rates of gene gain and gene loss are homogeneous across lineages.

During the present work we found one glitch in the calculations done in Koerich et al. (2008), that did not change any conclusion. Section 1.1.3. of the Supplementary Methods of (Koerich et al. 2008) correctly stated that "The 7 gains in 62.9 Myr we observed is the net gain rate, which does not take into account the genes that were acquired and subsequently lost in the *D. melanogaster* lineage (…) the corrected estimate of the gain rate ν [is obtained by] solving the equation 2 for ν." While applying this in Koerich et al. (2008), we mistakenly include the ancestral genes, whereas the correction was aimed to genes that were acquired in the *D. melanogaster* lineage. Namely, we set N_t to 12 genes and N_0 to 5 genes and obtained $\nu = 0.1201$ and a gain / loss ratio of 10.9 ($P = 0.003$; 95% CI: 2.3 - 52.5; Poisson regression); the appropriate values are $N_t = 7$ and $N_0 = 0$, which yield $\nu = 0.1149$ and a gain / loss ratio of 10.7 ($P = 0.003$; 95% CI: 2.2 - 51.3). The Supplemental Data File S5 (*analytical_vir_mel.xls*) implements the rectified bias corrections for both the *D. melanogaster* data reported in Koerich et al. (2008) and the *D. virilis* data (below).

D. virilis data: When we use the Homogeneous Gain Loss method with *D. virilis* as the reference species, there are four gene gains (the orthologs of *CG11719*, *CG2964*, *kl-5* and *GJ19835*; Fig. 4) by the Y chromosome in 63 million years (Myr) and three genes losses (*PRY*, *Ppr-Y*, and *JYalpha*) in 275 Myr. Both the raw gain rate (0.0636 genes/Myr) and the raw loss rate (0.01090 genes/Myr) are biased, and so is their ratio (5.83); the unbiased values are 0.0669 genes / Myr, 0.01360 genes /Myr, and 4.92, respectively . The Supplemental Data File S5 (*analytical_vir_mel.xls*) implements the bias correction. The statistical significance of the unbiased gain / loss ratio was tested with a Poisson regression, and implemented in R language (R Core Team 2012) with the Supplemental Data Files S3, S6 and S7 (*Poisson_regression.R*, *HGL_vir_data.txt*, and *HGL_mel_data.txt*, respectively). We found that the unbiased gain / loss ratio (4.92) is significantly different from 1 ($P = 0.037$; 95% confidence interval 1.1 - 22.0).

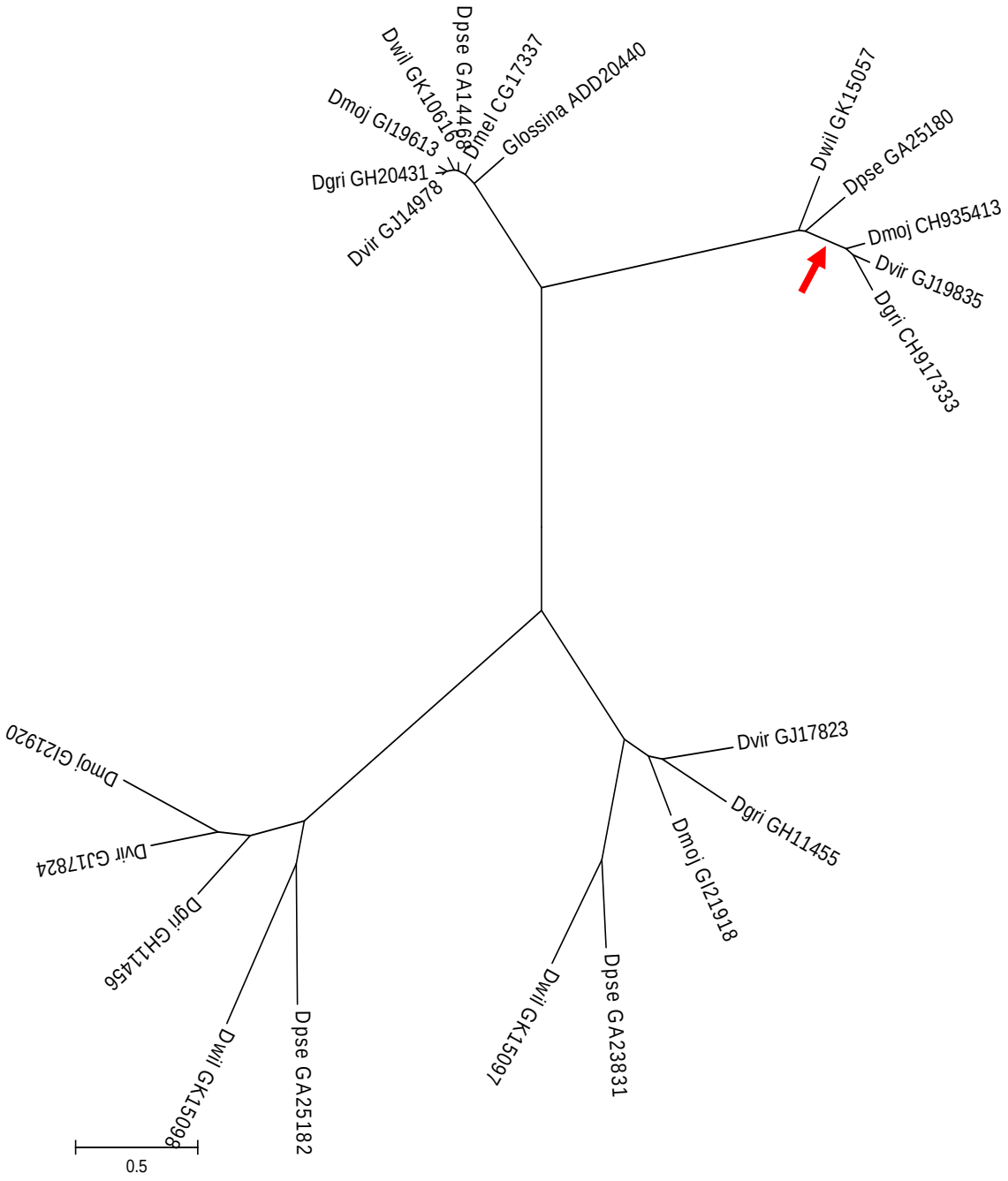
Combined *D. virilis* and *D. melanogaster* data: Given the lack of significant heterogeneity (P = 0.33; section "Procedure 1: Assumption-free"), we combined these data and analyzed them with the Homogeneous Gain Loss method, as follows. We used genes discovered in the two reference species, and checked for their Y-linkage in the remaining ones. As in the case of "*D. virilis* data", we can only estimate the gene gain rate in the phylogenetic branches connected to the reference species ("red branches") but now the same branches provide information for the gene loss rate (e.g., the *JYalpha* gene loss in the *D. melanogaster* lineage; Fig. 4); the other branches ("blue branches") only provide information about gene loss rate. In the combined data there are 11 gene gains in 126 Myr, and three gene losses in 338 Myr. The raw estimates are 0.0874 genes/Myr (gene gain rate), 0.00887 genes/Myr (gene loss rate), and 9.85 (gain / loss ratio). As an approximation of the unbiased estimates, we averaged the *D. melanogaster* and *D. virilis* values, obtaining 0.0909 genes/Myr (gene gain rate), 0.01219 genes / Myr (gene loss rate). These values correspond to a 7.46 gain / loss ratio, which significance was tested with a Poisson regression (P = 0.002; 95% confidence interval 2.1 - 26.7). The Poisson regression was implemented with the Supplemental Data Files S3 and S8 (*Poisson_regression.R* and *HGL_vir_mel_data.txt* , respectively).

Procedure 3: Approximate Bayesian computation

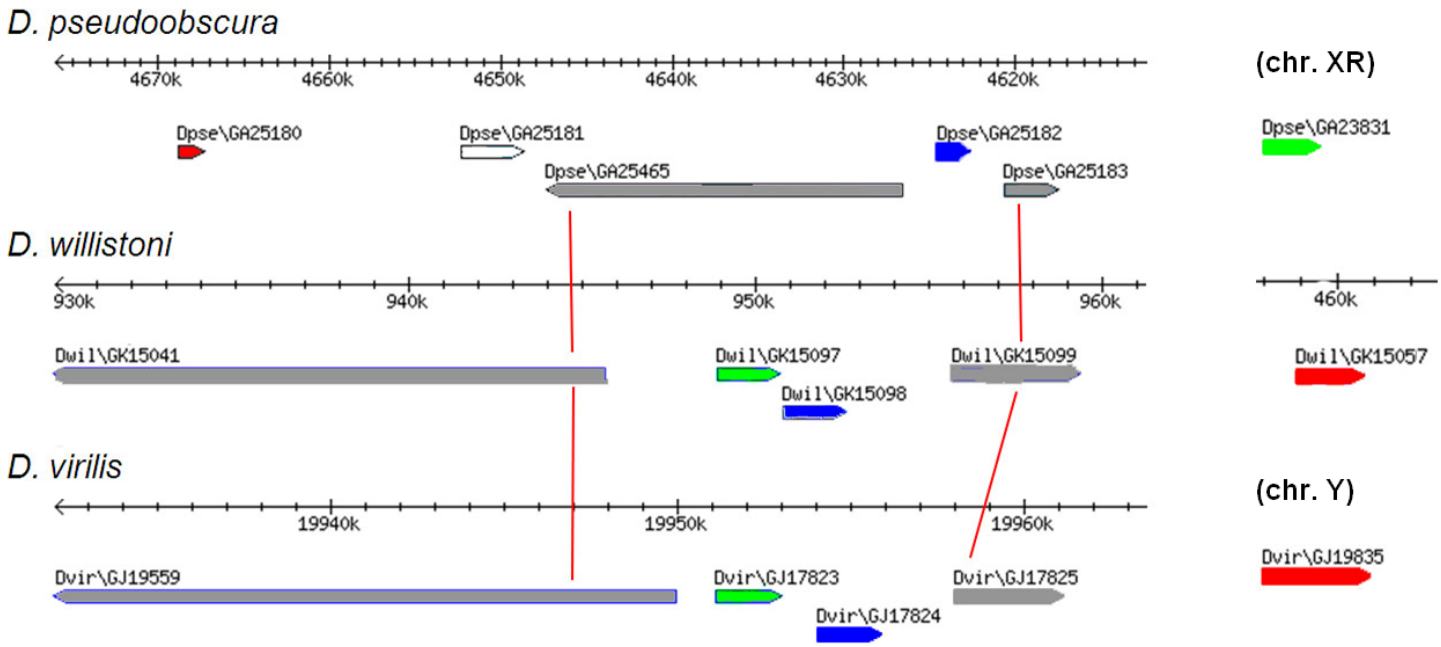
Computer simulations were carried essentially as described in Koerich et al. (2008), which should be consulted for details. Briefly, a Bayesian rejection sampling procedure was applied to yield 1,000 estimates of the rates of gene gain and loss conditional on the observed gains and losses of genes on the *D. virilis* Y chromosome, matching the actual ascertainment of the Y-linked genes. The computer code was written in the statistical language R (R Core Team 2012; Supplemental Data File S9 *ApproxBayes_vir.R*). After drawing random rates of gene gain and loss per gene from an uniform distribution and collecting 1,000 runs that satisfied the rejection criteria (no loss of ancestral genes and four net gains on the *D. virilis* lineage; three losses of known genes on the other branches), approximate Bayesian estimates (Beaumont 2010; Beaumont et al. 2002; Przeworski 2003; Tavare et al. 1997) of the posterior densities of the gain rate, the loss rate per gene, and the net gene gain (gains minus losses) were obtained (Supplemental Fig. S9). In all 1,000 simulations the gains outnumber the losses (Supplemental Fig. S9A and S9B), which again strongly suggest that the Y chromosome on average is gaining genes. The gain rate and the loss rate per gene are the ultimate factors governing gene number dynamics; their joint posterior distributions are shown in Supplemental Fig. S9C. It is likely that as gene number increases the number of gene losses will increase until an equilibrium between gains and losses is attained. The simulations allow us to look at the posterior distribution of the predicted equilibrium gene number (Supplemental Fig. S9D). As expected given the previous result that the gene gains outnumber the losses, nearly all (973 out of 1,000) of the values of equilibrium gene number are above the present Y-linked gene number in *D. virilis* (10 genes). The parameters and values estimated by the simulations agree quite well with the analytical solution. For example, the average ratio of gain rate to loss rate in the simulations is 6.0 (Supplemental Fig. S9B), whereas the analytical value is 4.9 (see above). Perfect agreement is not expected because some assumptions are different. In particular, the simulations allowed variation among samples in the phylogenetic pattern of gains (*i.e.* the rejection criterion focused on counts of gains, not on which branches had gains), whereas this pattern is fixed in the analytical solution.

Supplemental Figures

Supplemental Fig. S1: Panel A



Supplemental Fig. S1: Panel B

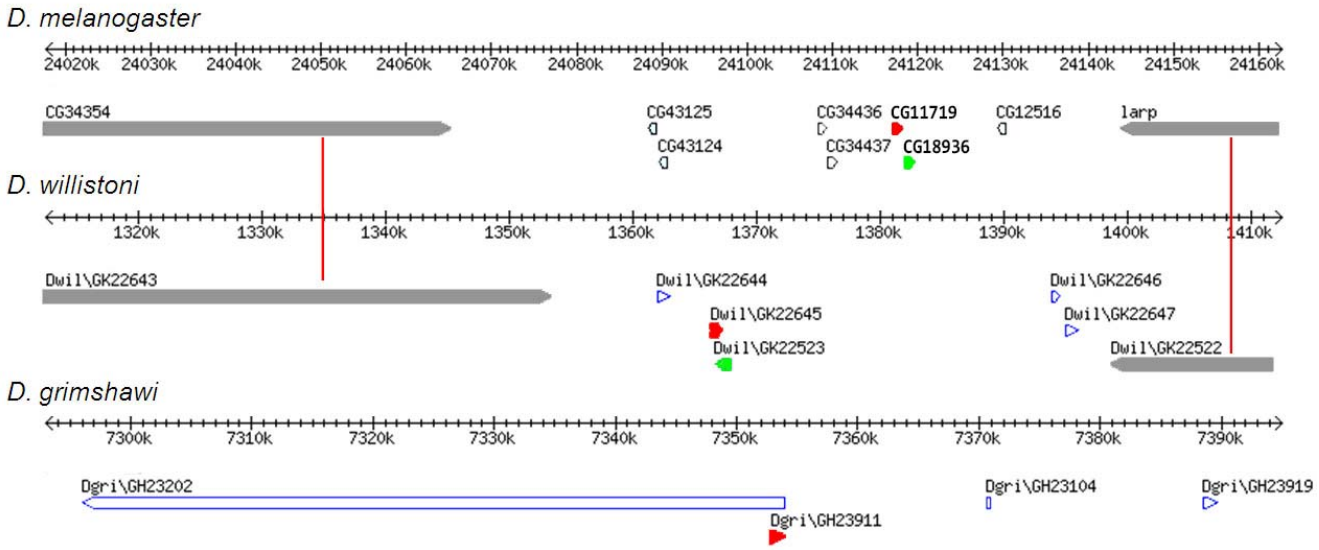


Supplemental Fig. S1. Orthology and synteny analysis of the *GJ19835* gene (M20 dipeptidase). (A) NJ tree of protein sequences of homologs of the *D. virilis GJ19835* gene, which encode dipeptidases of the M20 subfamily (Rawlings 2009). The branch where the gene moved to the Y-chromosome is marked with a red arrow. Note the four clearly delimited phylogenetic clusters. The phylogenetic cluster that contains the *D. melanogaster* gene *CG17337* is well conserved, and has orthologs in all 12 *Drosophila* species (we removed some species of the melanogaster group for the sake of clarity) and in *Glossina morsitans* (Alves-Silva et al. 2010); its *D. melanogaster* member *CG17337* is ubiquitously expressed (www.flybase.org ; McQuilton et al. 2012). The three other clusters do not have orthologs in any species of the melanogaster group (*D. melanogaster*, *D. simulans*, *D. yakuba*, *D. erecta*, *D. ananassae*), or outside the *Drosophila* genus. They are located in imperfect tandem (two genes moved to different locations; see panel B), and most likely were lost in the ancestor of the melanogaster group. The *D. pseudoobscura* ortholog *GA25180* (which is autosomal) of the Y-linked genes is expressed only in males (www.modencode.org ; Celniker et al. 2009; there is no data available for the *D. willistoni* ortholog *GK15057*). Accession numbers of the new sequences reported in this paper: Dvir_GJ19835, BK008736; Dgri_CH917333, BK008737; Dmoj_CH935413, BK008738; Dwil_GK15057, BK008739; all other sequences were taken from FlyBase. (B) Synteny analysis of the three dipeptidase genes (painted in red, blue and green; *CG17337* orthologs were excluded) in a set of representative species (figure modified from www.flybase.org). The red lines show the orthologous genes (painted in grey) that flank the three dipeptidase genes and establish the synteny of this region. These three genes are syntenic in the five species that carry them (*D. pseudoobscura, D. willistoni, D. mojavensis, D. virilis,* and *D. grimshawi*), so they probably originated by gene duplications at this location , and their ancestral location is autosomal. The alternative hypothesis (ancestrally Y-linked) would imply that the ortholog of the Y-linked gene moved in the ancestors of *D. willistoni* and *D. pseudoobscura* from the Y chromosome to the exact location of its paralogs, which is highly improbable (*e.g.*, Koerich et al. 2008). The gene *Dwil\GK15057* is located in the same chromosome, at 500 kb of distance, and *Dpse\GA23831* moved to chromosome XR. The orthologs of the *GJ19835* gene in *D. mojavensis* and *D. grimshawi* have not been annotated before; the *D. willistoni* ortholog (*GK15057*) was partially annotated.

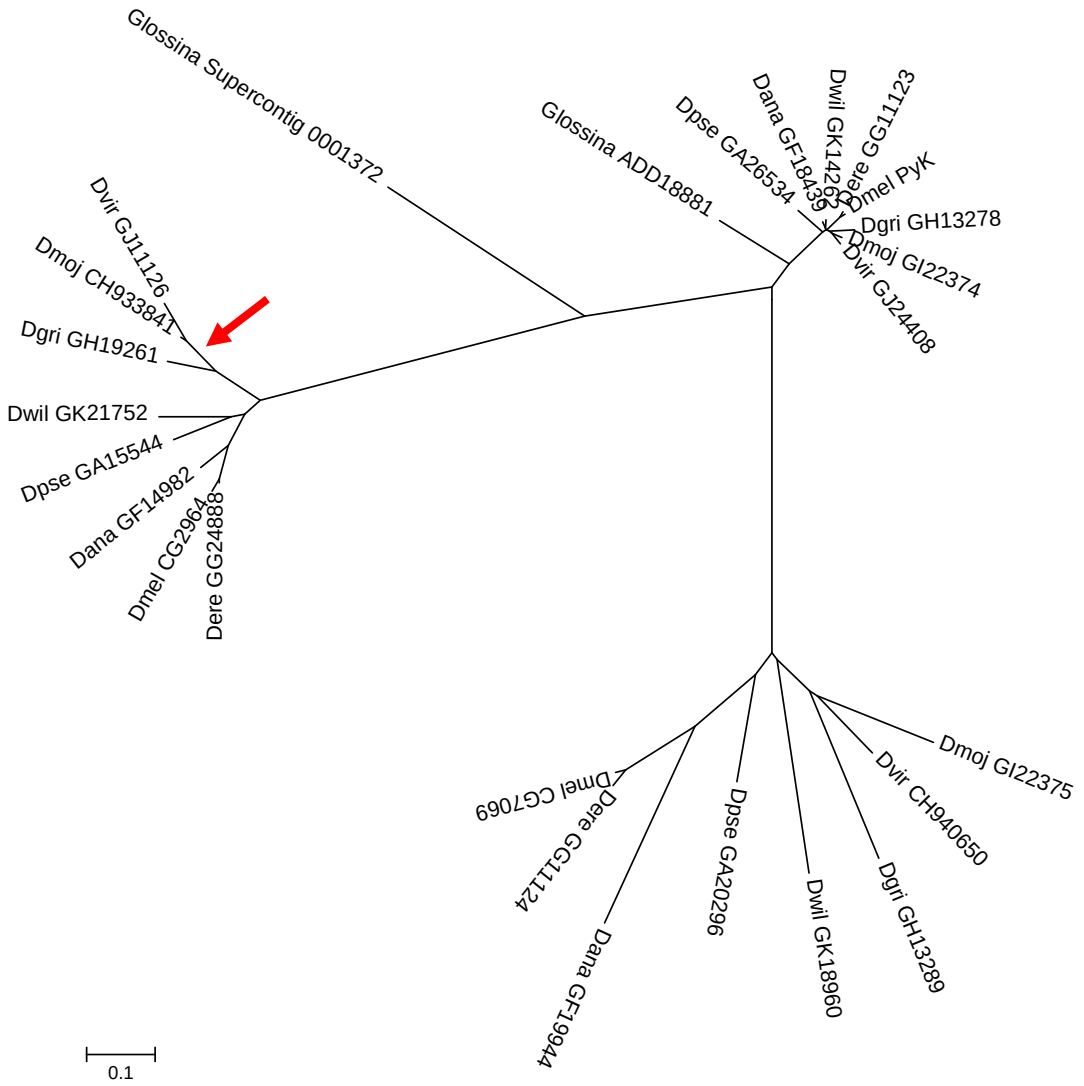
Supplemental Fig. S2, Panel A

```
Dmel_CG11719 MCSPCGPCSPCDPCCGPFECSPKCYNAAQLEALPQCAPRIPPPFPKCITVQQPPRMICKK 60
Dvir_GJ19633 MCSPCGPCSPCDPCCGPFECSPKCYNAAQLEALPQCAPRIPPPFPKCITVQQPPRLICKK 60
Dmel_CG18396 MCSPCGPCSPCDPCCGPFECSPKCYNAAQLEALPQCAPRIPPPFPKCITVQQPPRMICKK 60
Dvir_GJ21217 MCSPCGPCSPCDPCCGPFECSPKCYNAAQLEALPQCAPRIPPPFPKCITVQQPPRLICKK 60
             **************************************************:****

Dmel_CG11719 RVVFTEKIVPEPMVVNRCRQITIPKVVDATRVIKVPKLIWVSQMVREPRVIYYPSMIPDP 120
Dvir_GJ19633 RVVFTEKIVPEPMVVNRCRQITVPKVVDATRVIKVPKLIWVSQMVREPRVIYYPSMIPDP 120
Dmel_CG18396 RVVFTEKIVPEPMVVNRCRQITIPKVVDATRVIKVPKLIWVSQMVREPRVIYYPSMIPDP 120
Dvir_GJ21217 RVVFTEKIVPEPMVVNRCRQITIPKVVDATRVIKVPKLIWVSQMVREPRVIYYPSMIPDP 120
             **********************:***********************************

Dmel_CG11719 YVVCYPKRVCEPREVCQSILCQPKPQTIDIPPPREYCCYPNGPINYKPSAACPPCPIGPC 180
Dvir_GJ19633 YVVCYPKRVCEPREVCQSILCQPKPQTIDIPPPREYCCYPNGPINYKPSAACPPCPIGPC 180
Dmel_CG18396 YVVCYPKRVCEPREVCQSILCQPKPQTIDIPPPREYCCYPNGPINYKPSAACPPCPIGPC 180
Dvir_GJ21217 YVVCYPKRVCEPREVCQSILCQPKPQTIDIPPPREYCCYPNGPINYKPSAACPPCPIGPC 180
             ************************************************************

Dmel_CG11719 APGGPCCPLPCFSTNQYPVAEGRCGPCGPCGPGCG-PCGPRCGPCGPGPCGPCGPRCGPG 239
Dvir_GJ19633 APGGACCPLPCFSTNQYPVCETRCGPCGPCGPCGP-GRCGPCGPGRCGPCGPCSGPCGPC 239
Dmel_CG18396 APGGPCCPLPCFSTNQYPVAEGRCGPCGPCGPG----------CGPCGPCGPWGPGCGPC 230
Dvir_GJ21217 APGGACCPLPCFSTNQYPVCETRCGPCGPCGPGGPCGPCGPCGAGPCGPFGPCGGPCGPC 240
             ****.**************.* **********               ** **  .  ***

Dmel_CG11719 G-PCAVPNCGPCGLTMPSGPFVPAPCGPCAPCGPCGPLGNSPCGPCGPCGPCSPPCPYES 298
Dvir_GJ19633 GGPCAVPNCGPCGLTMPSGPFIPAPCGPCGTG--CGPLGNGPCGPCGPCGPCSPPCPYES 297
Dmel_CG18396 GPCGPCGPFGPGCGPCGPCGPCGPGPCGFGPCGPC------------------------ 263
Dvir_GJ21217 GPGGPCGPCGPTGGPCGPFAPCGPSGPGCYCPCGSL------------------------ 273
             *

Dmel_CG11719 PECGPCYPCAPTPWNTHCGPVGPCGPQVPCGPSGPC 334
Dvir_GJ19633 PECGPCYPCAPTPWNTHCGPVGPCGPQVPCGPCGPC 333
Dmel_CG18396 ----------------------------------- 265
Dvir_GJ21217 ----------------------------------- 276
```

Supplemental Fig. S2, Panel B

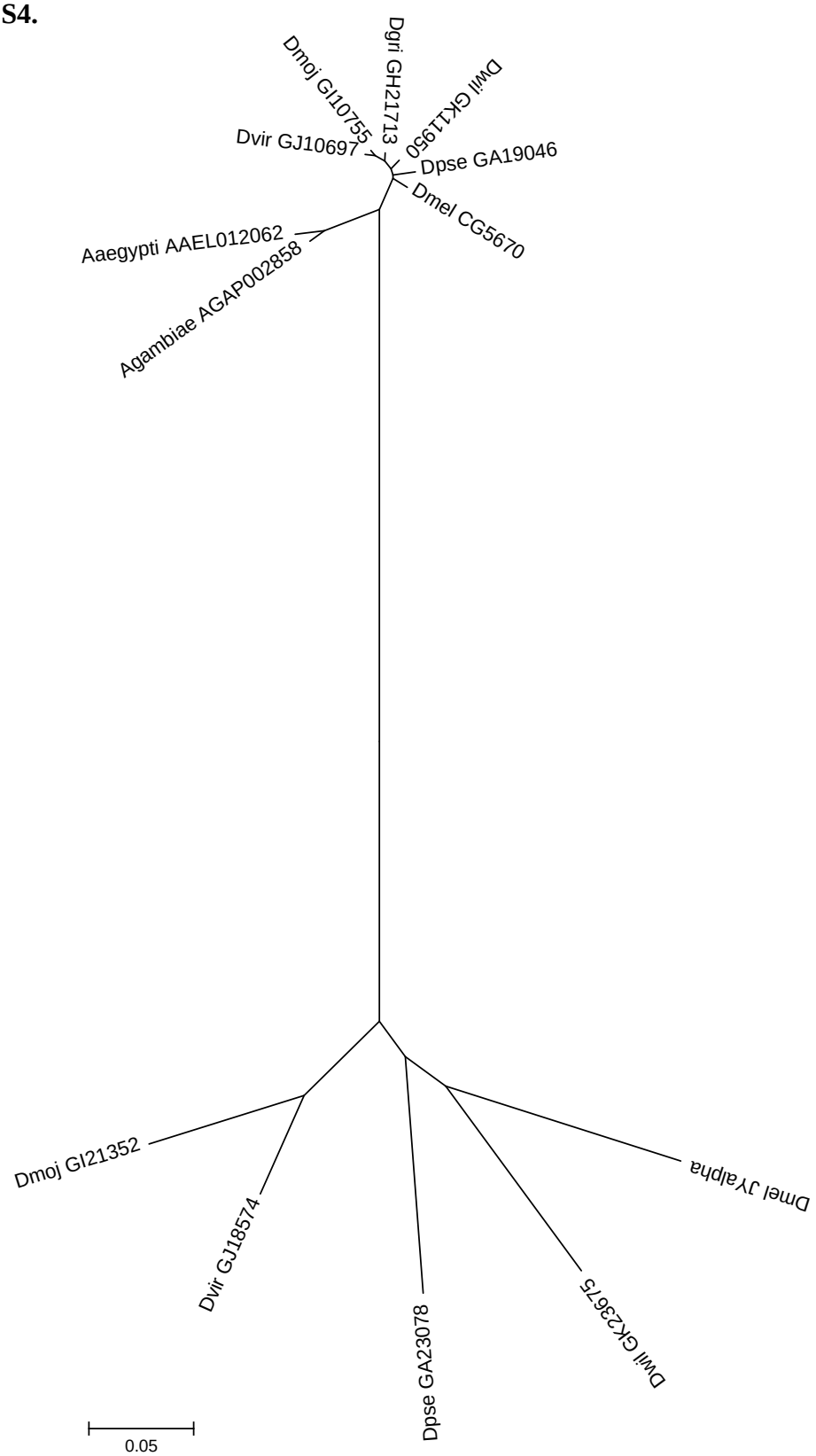


Supplemental Fig. S2. Orthology and synteny analysis of the *GJ19633* gene (*CG11719* ortholog). The *GJ19633* gene is the ortholog of the *D. melanogaster* gene *CG11719*; they have a very similar paralog (*CG18936* / *GJ21217*). Both *D. melanogaster* genes (*CG11719* and *CG18936*) encode well conserved sperm tail structural proteins (Gastmann et al. 1993; Schafer et al. 1993) and are expressed only in males (McQuilton et al. 2012; they are also known as *Mst98Ca* and *Mst98Cb*). (A) Alignment of the *D. melanogaster* and *D. virilis* genes; note the extremely high conservation, except for the repetitive motif Cys-Gly-Pro and for the C-terminus. The orthology relationships are better inferred by the alignment, rather than by a phylogenetic tree because the N-terminus is nearly identical between the paralogs, and the C-terminus is uninformative (it contains gaps and a repetitive motif). All 12 Drosophila species have one "long gene" (with ~340 amino acids; *e.g.*, *CG11719*) and one "short gene" (with ~260 amino acids; *e.g.*, *CG18936*); we found only the long gene in *Glossina*, so the short gene probably originated from the long one (data not shown). The long genes are Y-linked in *D. virilis* (*GJ19633*) and *D. mojavensis* (which has two Y-linked copies, *GI10867* and *GI21836*). (B) Synteny analysis of the *CG11719* and *CG18936* genes (painted in red and green, respectively) in a set of representative species (figure modified from www.flybase.org ; McQuilton et al. 2012). The red lines show the orthologous genes (painted in grey) that flank the two genes and establish the synteny of this region. The two genes are in tandem in all Sophophora, in a conserved synteny block flanked by the genes *CG34354* and *larp* (chromosome 3R; *D. melanogaster* names). Hence the ancestral location of the Y-linked genes is autosomal; the alternative hypothesis imply that the gene moved from the Y to the exact location of its short paralog, which is highly improbable (also, it would had moved twice to an autosome, in the ancestors of Sophophora and of *D. grimshawi*). The synteny is only weakly conserved between the Sophophora and Drosophila subgenera: in *D. grimshawi* the long gene (*GH23911*) is surrounded by genes which *D. melanogaster* orthologs are also located in the 3R chromosome (*GH23302* / *CG5023* and *GH23919* / *Rh3*), suggesting that it stayed in the same chromosome, and was relocated due to a chromosomal inversion (Bhutkar et al. 2007). In *D. mojavensis* and *D. virilis* the gene moved to the Y chromosome. Regarding the short gene, it is syntenic among *D. virilis*, *D. mojavensis* and *D. grimshawi*, but located in different autosome in relation to the Sophophora species (not shown in panel B).

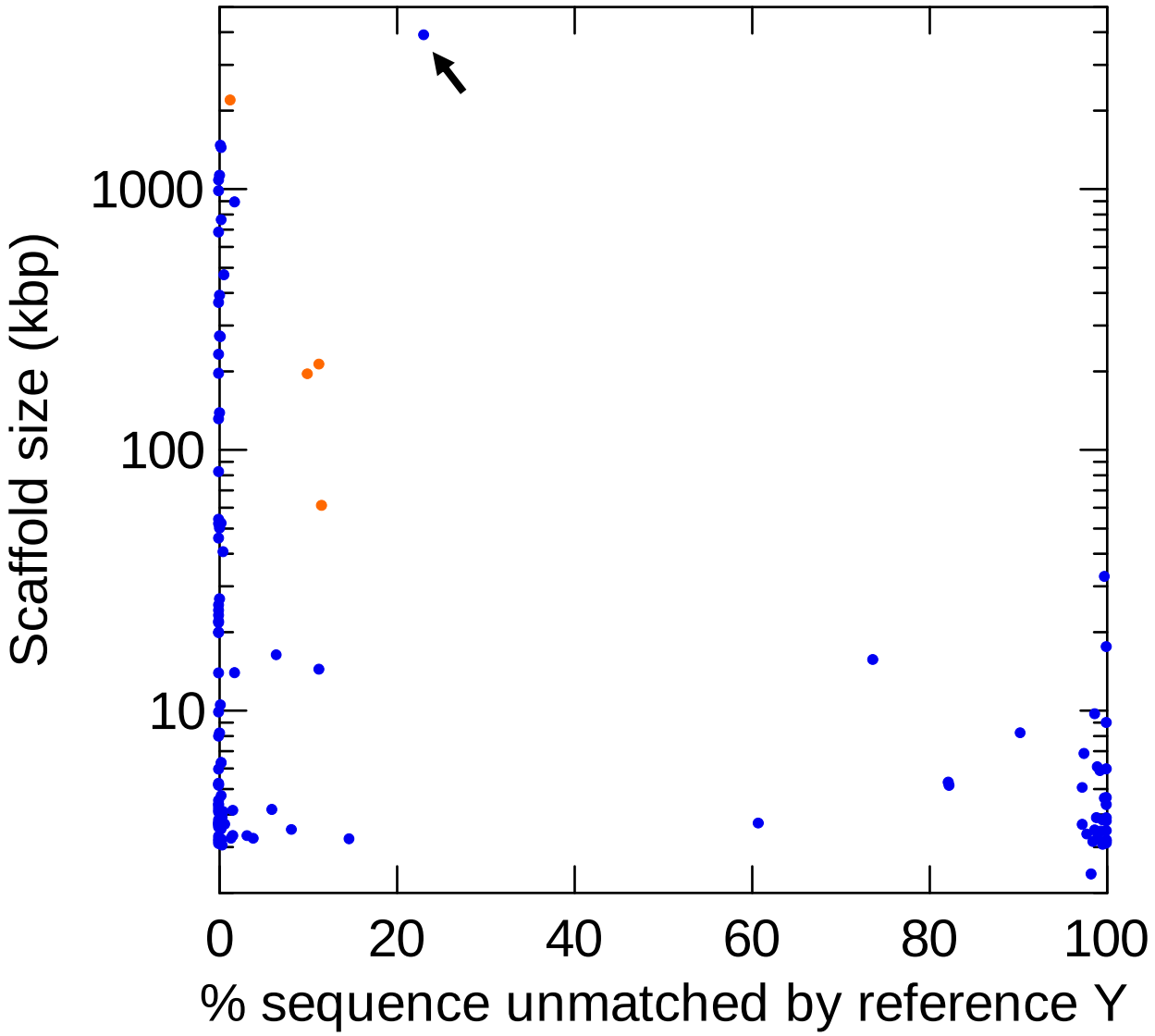


Supplemental Fig. S3. Orthology and synteny analysis of the *GJ11126* gene (*CG2964* ortholog). NJ tree of protein sequences of homologs of the *D. virilis GJ11126* gene, which encode pyruvate kinases. The branch where the gene moved to the Y chromosome is marked with a red arrow. There are three clearly delimited phylogenetic clusters, all with orthologs in the 12 *Drosophila* species. The cluster that contains the *D. melanogaster* gene *PyK* is well conserved and ubiquitously expressed (McQuilton et al. 2012). There is a fast-evolving cluster which is not Y-linked; its *D. melanogaster* representative *CG7069* is expressed only in males. The *D. melanogaster* gene *CG2964*, which belong to the cluster that has Y-linked genes, is expressed only in males (McQuilton et al. 2012) and encodes a pyruvate kinase present in the sperm proteome (Wasbrough et al. 2010). The synteny data is not informative for the origin of the Y-linked genes: *CG2964* is located in chromosome 2L and there is synteny conservation only in the melanogaster group and in *D. pseudoobscura*; the *D. willistoni* ortholog *GK21752* is located in a scaffold devoid of any *D. melanogaster* orthologs, and the *D. grimshawi* ortholog *GH19261* is located in a region with orthologs of *D. melanogaster* genes of chromosome 3R (not shown). The ancestral location in an autosome is established by parsimony; the alternative hypothesis of ancestral Y-linkage would imply in at least two Y-to-autosome movements (in the ancestors of the Sophophora subgenus, and of *D. grimshawi*). The orthologs of the *CG7069* gene in *D. virilis*, and of *CG2964* in *D. mojavensis* and *Glossina* have not been annotated before. Accession numbers of the new sequences reported in this paper: Dvir_GJ11126, BK008740; Dmoj_CH933841, BK008741; all other sequences were taken from FlyBase.

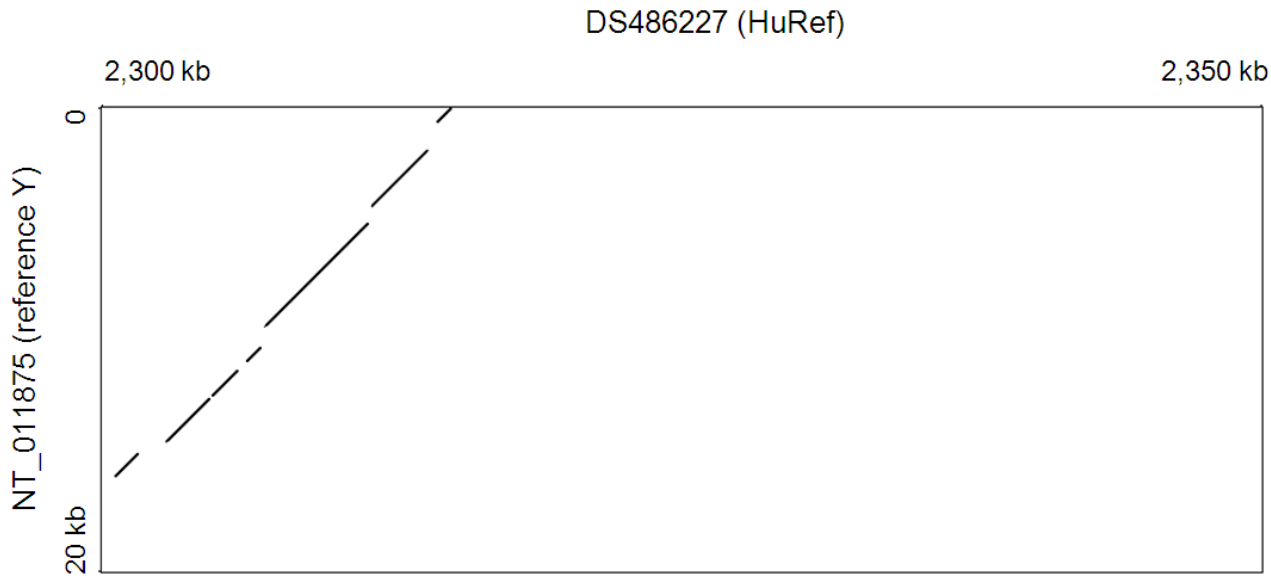
Supplemental Fig. S4.



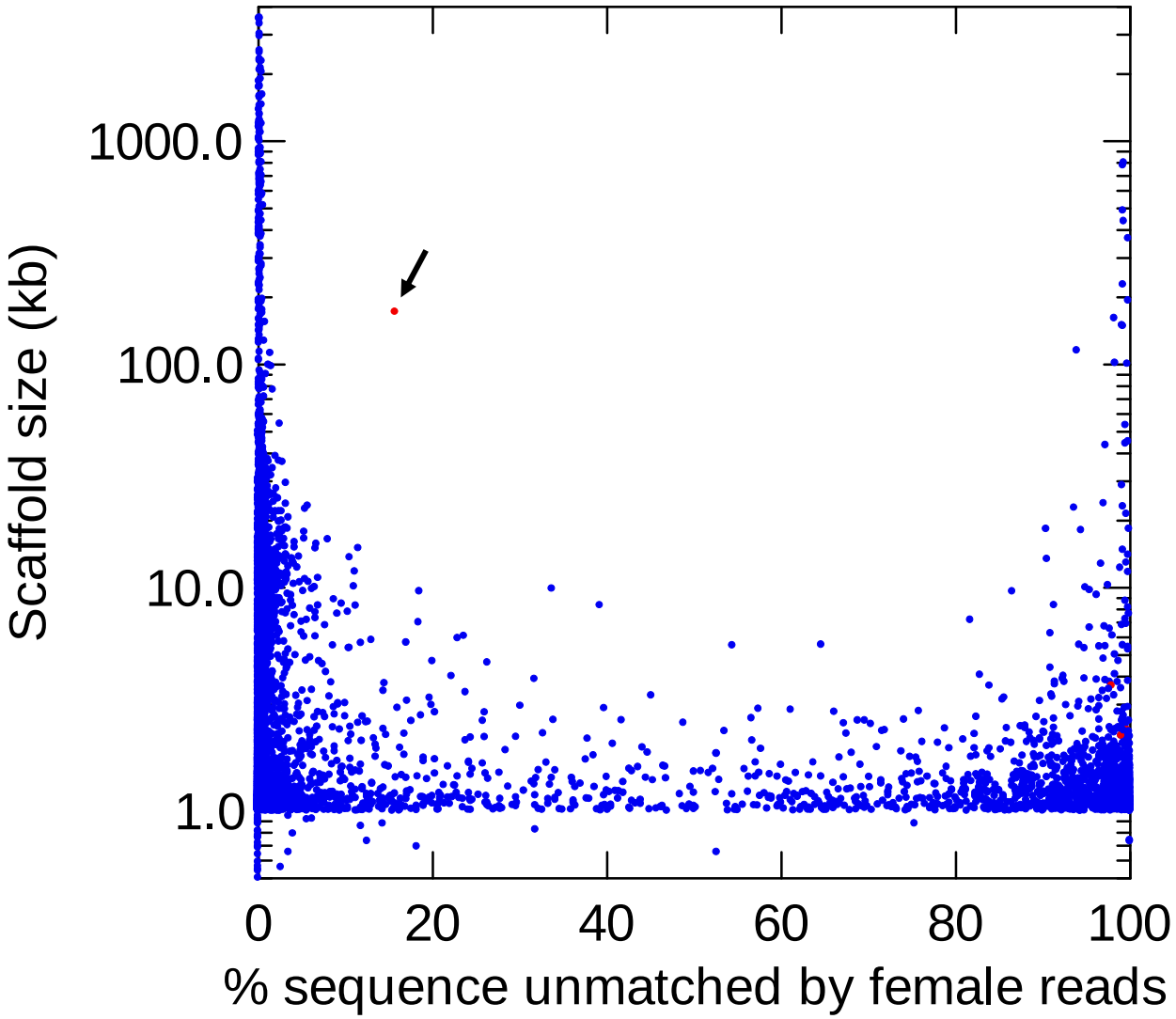
Supplemental Fig. S4. Orthology analysis of the *GJ18574* gene (*JYalpha* ortholog). NJ tree of protein sequences of homologs of the *D. virilis GJ18574* gene, which encode the alpha subunit of Na,K-ATPase (Okamura et al. 2003). This gene belong to the ancestral *Drosophila* Y chromosome and moved to an autosome in the *D. melanogaster* lineage (Supplemental Fig. S8). There are two clusters, both with orthologs in all 12 *Drosophila* species (some orthologs were not shown for the sake of clarity). The cluster that contains the *D. melanogaster* gene *CG5670* is very well conserved, ubiquitously expressed (McQuilton et al. 2012) , and has orthologs in distant Diptera (mosquitoes). The *D. melanogaster* gene *JYalpha*, which belong to the cluster that has Y-linked genes, is expressed only in males (McQuilton et al. 2012) and encodes a sperm-specific subunit of the Na,K-ATPase (Masly et al. 2006); we could not find any ortholog outside the *Drosophila* genus. The synteny data is not informative for this gene given that it is ancestrally Y-linked; the assembly of this chromosome is too fragmented to allow this type of study. The ancestral state of Y-linkage (followed by Y-to-autosome movement within the melanogaster group) is established by parsimony: the alternative hypothesis of ancestral autosomal location would imply in at least three autosome-to-Y movements (in the ancestors of the Drosophila subgenus, of *D. willistoni*, and of *D. ananassae*). As frequently happens, the Y-linked genes were partially annotated during the genome projects (*Dmoj_GI21352*, *Dwil_GK23675*, *Dgri_GH23560*) or completely missed (*D. ananassae* ortholog). Accession numbers of the new sequences reported in this paper: Dvir_GJ18574, BK008742; Dmoj_GI21352, BK008743; Dgri_GH23560, BK008744; all other sequences were taken from FlyBase.



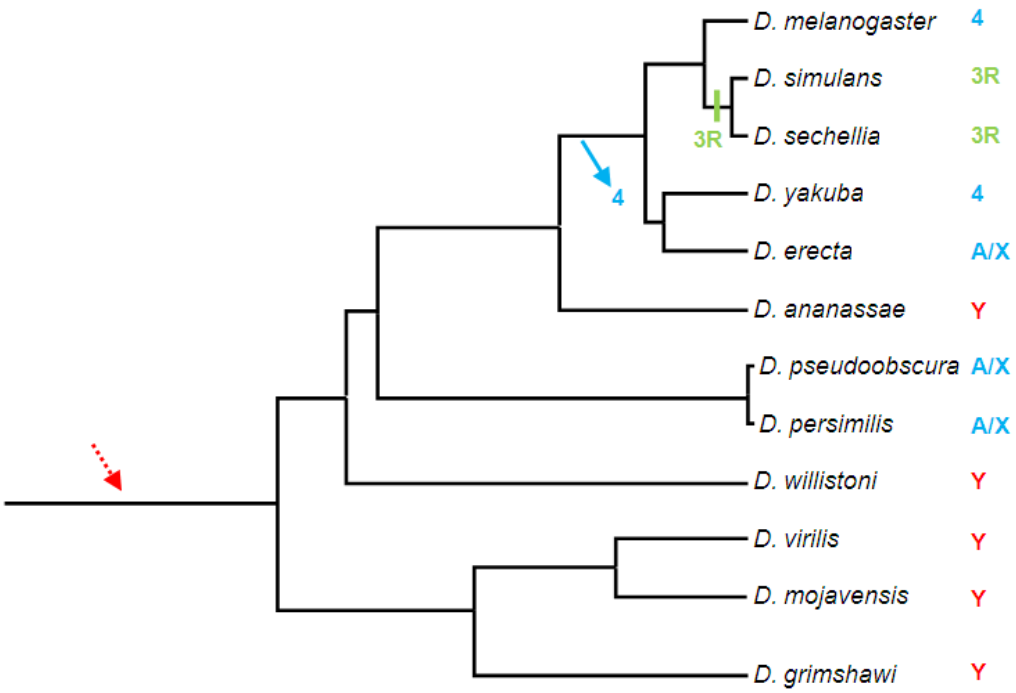
Supplemental Fig. S5. Comparison between the 119 Y-linked HuRef scaffolds and the reference Y chromosome sequence. The comparison aims to detect new sequences (Methods section "Detection and validation of new human Y-linked sequences"). Note the sharp distinction between scaffolds that are mostly composed of new sequence (right side of the graph; 34 scaffolds), and the scaffolds covered by the reference Y sequence (left side of the graph; 85 scaffolds). Four of these 85 scaffolds (orange dots) contain a small amount of new sequence (76 kb total). The arrow points to the DS486171 scaffold, which contains part of the misassembled XTR segmental duplication; it does not contain any new sequence. Abscissa, proportion of the scaffold sequence not covered by the reference Y chromosome (in % unmatched single-copy *k*-mers); ordinate , scaffold size.



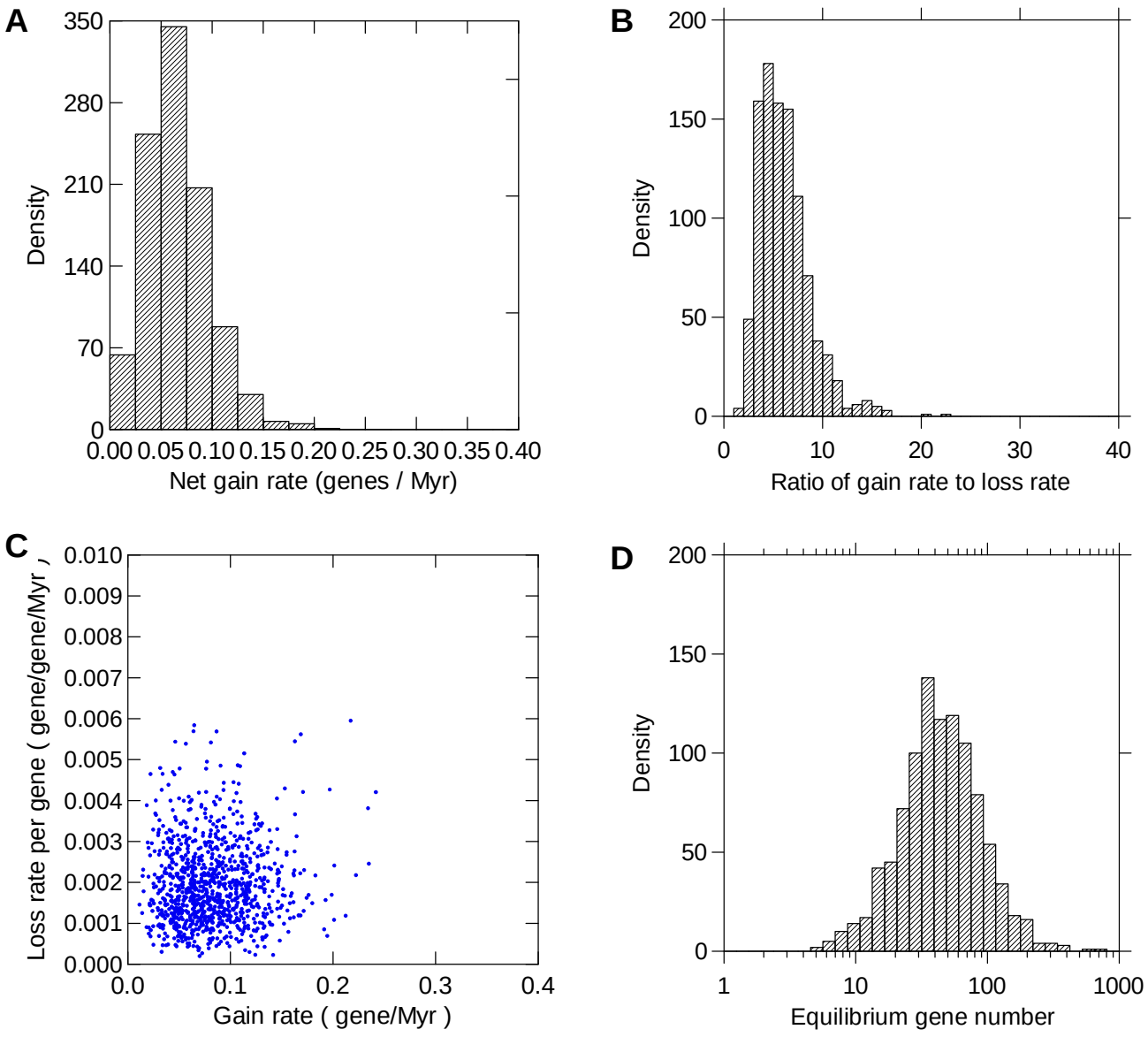
Supplemental Fig. S6. New human Y chromosome sequences. Dot-plot of the alignment of the region shown in Fig. 6 to the reference Y. The scaffold DS486227 extends scaffold NT_011875 to the left, by 32 kb. The new sequence is located in the gap between the reference scaffolds NT_011875 and NT_113819. We found four other regions that extend the reference Y chromosome sequence, totaling 76 kb of unfinished sequence (Supplemental Table S3).



Supplemental Fig. S7. Application of the *YGS* method to a short-read assembly (*D. kikkawai*). As in *D. virilis*, male reads were not available in *D. kikkawai*, so sequencing errors were partially removed by comparison with the mixed-sex reads used to assemble the genome (to avoid the homopolymer errors of 454 reads, we used only the Illumina reads (file SRX097585_sra_data.fastq.gz, downloaded from NCBI)). Red dots are scaffolds previously known to be Y-linked, and blue dots are unmapped or not Y-linked scaffolds. Abscissa, proportion of scaffold sequence not matched by female short reads (in % unmatched single-copy *k*-mers); ordinate , scaffold size. Arrow point to misassembled scaffold that contains part of the Y-linked *kl-2* gene (Supplemental Table S5).

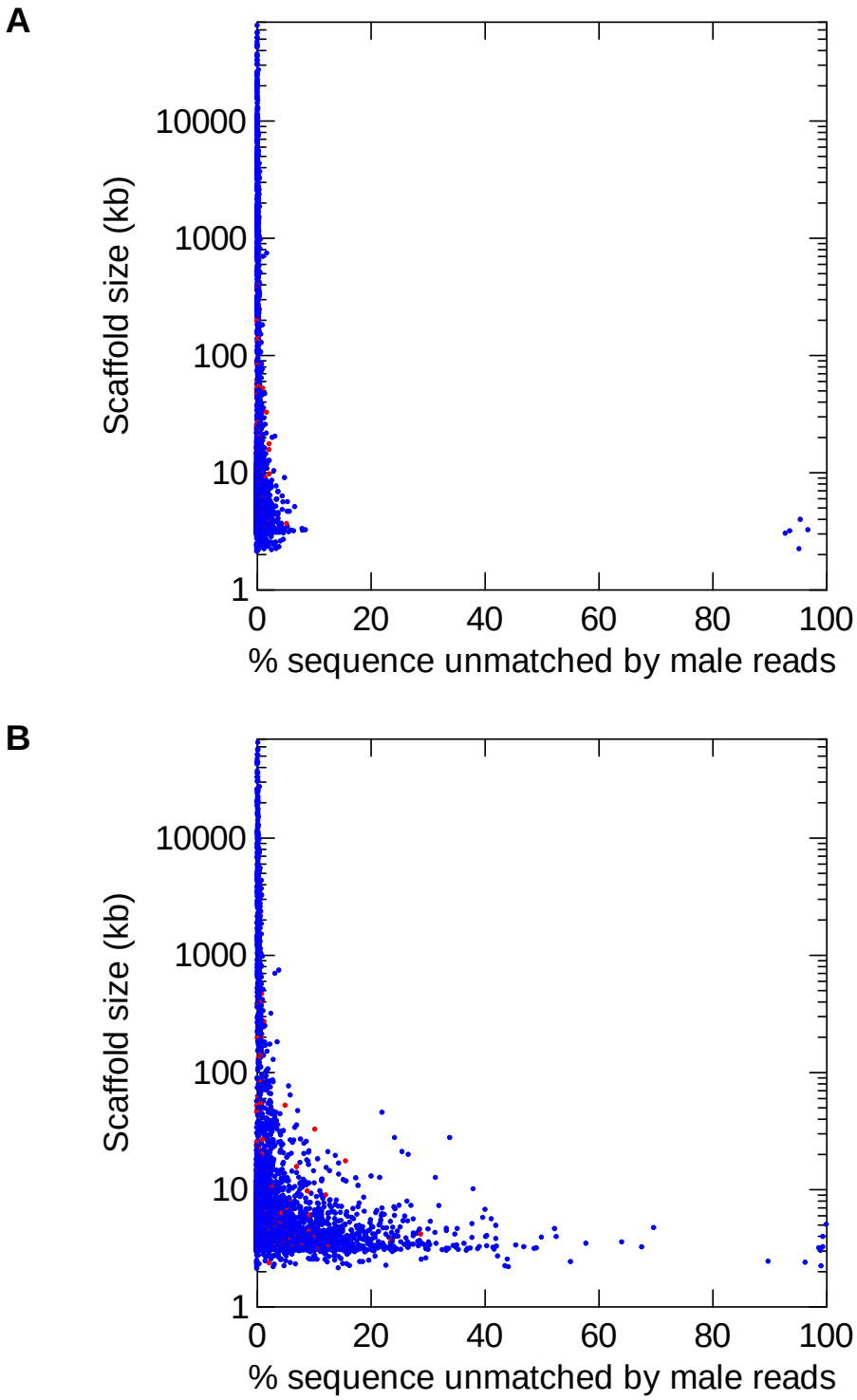


Supplemental Fig. S8. Movements of the *JYalpha* gene in the *Drosophila* genus. *JYalpha* is Y-linked in the Drosophila subgenus, in *D. willistoni*, and *D. ananassae* and autosomal (or X-linked) in all others (Supplemental Table S2). Chromosomal location in *D. melanogaster*, *D. simulans*, *D. sechellia*, and *D. yakuba* were taken from (Masly et al. 2006) , and in the other species was experimentally determined by PCR (Supplemental Table S2). Red dashed arrow, inferred autosome-to-Y movement; blue arrow, Y-to-autosome movement *(i.e.*, a gene loss); green bar, movement within the autosomes. As described in (Carvalho and Clark 2005) the Y chromosome became part of an autosome in the *D. pseudoobscura* lineage so was not considered while measuring individual gene gains and losses.



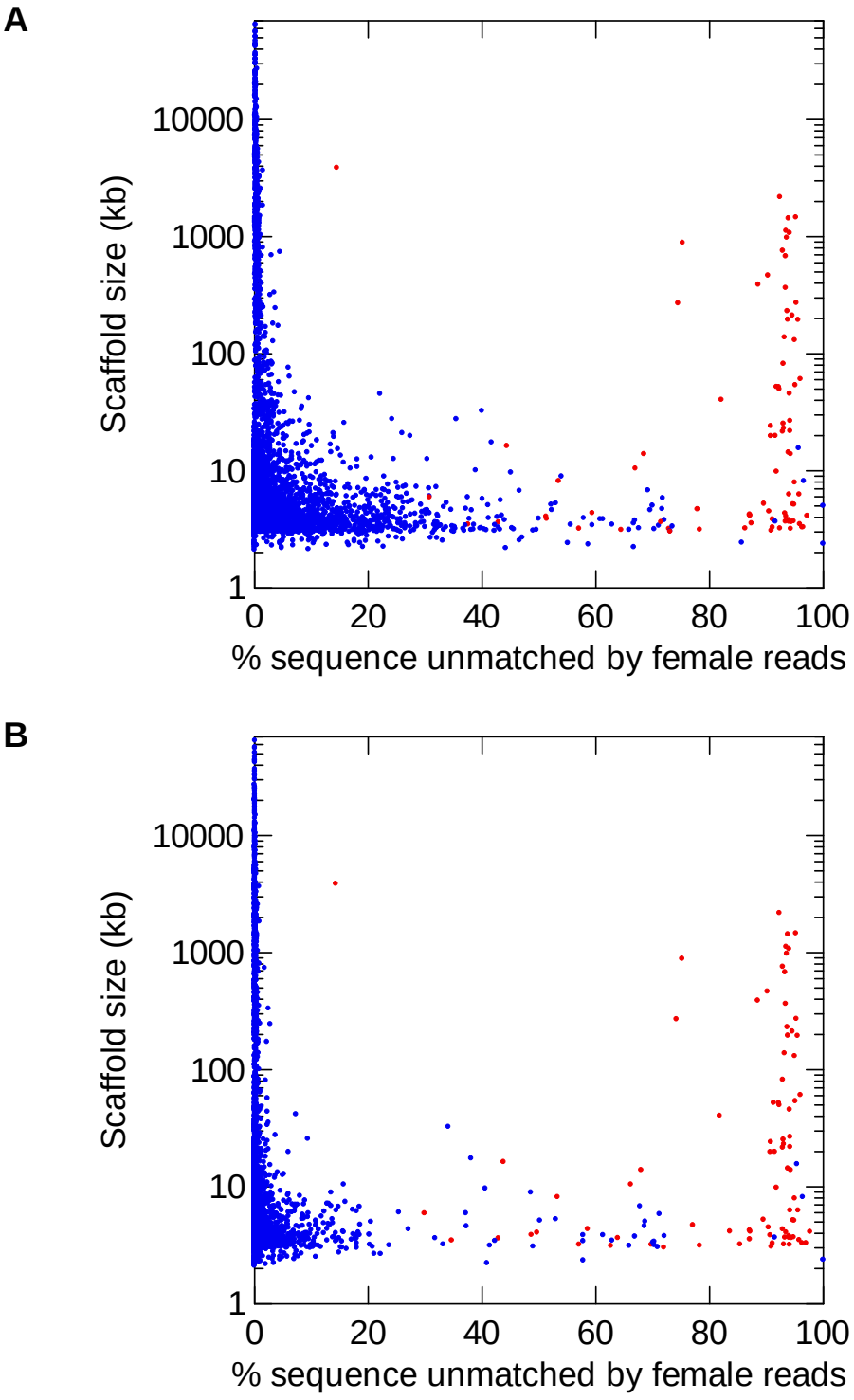
Supplemental Fig. S9. Results of 1,000 computer simulations of gene gain and loss using *D. virilis* data. (A) Posterior density of the net rate of Y-linked gene gain in the *Drosophila* phylogeny. The average net gain rate (gain rate minus loss rate) is 0.066 genes per Myr, and all 1,000 simulations had a higher rate of gene gain than loss (range of net gain rate:0.006 to 0.201). (B) Posterior distribution of the ratio of the rate of gene gain (genes/Myr) to the rate of gene loss (genes/Myr). The average value is 6.0 (range: 1.7 to 22.3; 95% credibility interval: 2.3 - 12.0). (C) Joint posterior distribution of gain rate and loss rate per gene. The average values are 0.0831genes / Myr and 0.0020 genes / gene / Myr, respectively. The uniform distributions used as priors for both parameters had maximums well above the highest accepted values (prior for gain: 0 - 0.5 genes / Myr ; prior for loss: 0 - 0.01 genes / gene /Myr). (D) Posterior distribution of the predicted equilibrium gene number (note the logarithmic scale of the abscissa). The average value is 57 genes (range: 5 to 657). 27 out 1,000 simulations had predicted equilibrium gene number below 10 (the present gene number of the *D. virilis* Y).

Supplemental Fig. S10, panels A and B



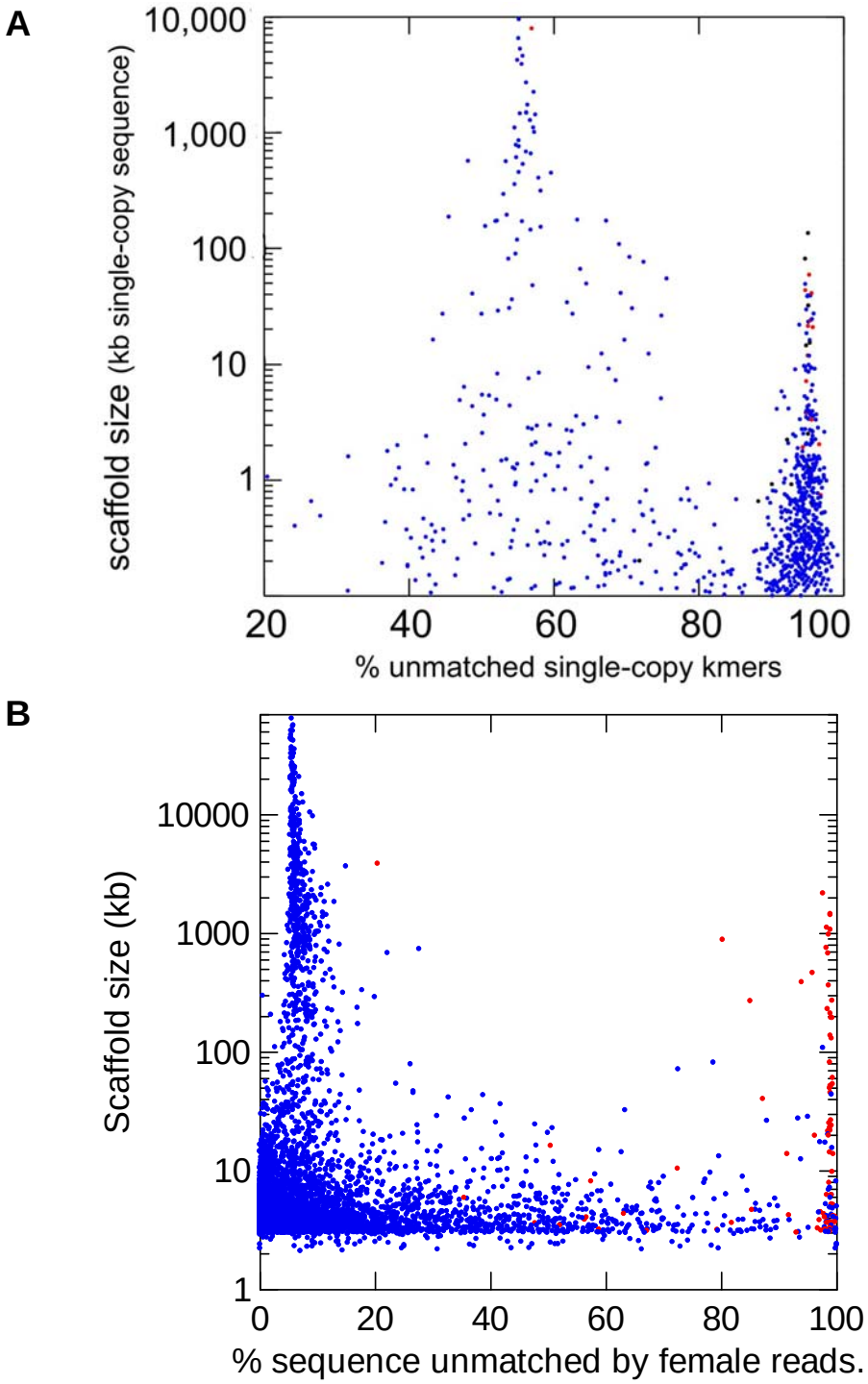
Supplemental Fig. S10. Identification of contaminant scaffolds in the HuRef assembly using male reads. Legitimate genomic sequences should be completely matched by **male** short-reads (barring sequencing errors and rare polymorphisms in the assembled genome), whereas contaminants are expected to get no match (Methods section "Detection of contaminant scaffolds"). (A) Application of this procedure to the HuRef assembly (all *k*-mers, including the repetitive ones). Five scaffolds (out of 4606) stood out; all have bacterial origin and were excluded from all subsequent analysis. (B) Analysis similar to panel A, but using only single-copy *k*-mers. Unmatching *k*-mers are noticeable in the small scaffolds, and probably represent sequencing errors and to a lesser extent rare polymorphisms in the HuRef assembly. Note that single-copy *k*-mers are enriched in sequencing errors and rare polymorphisms, which explains the difference between panels A and B. Red dots are the 119 Y-linked scaffolds. Note also that small scaffolds (below 10 kb) have an increased proportion of unmatched *k*-mers.

Supplemental Fig. S11, panels A and B



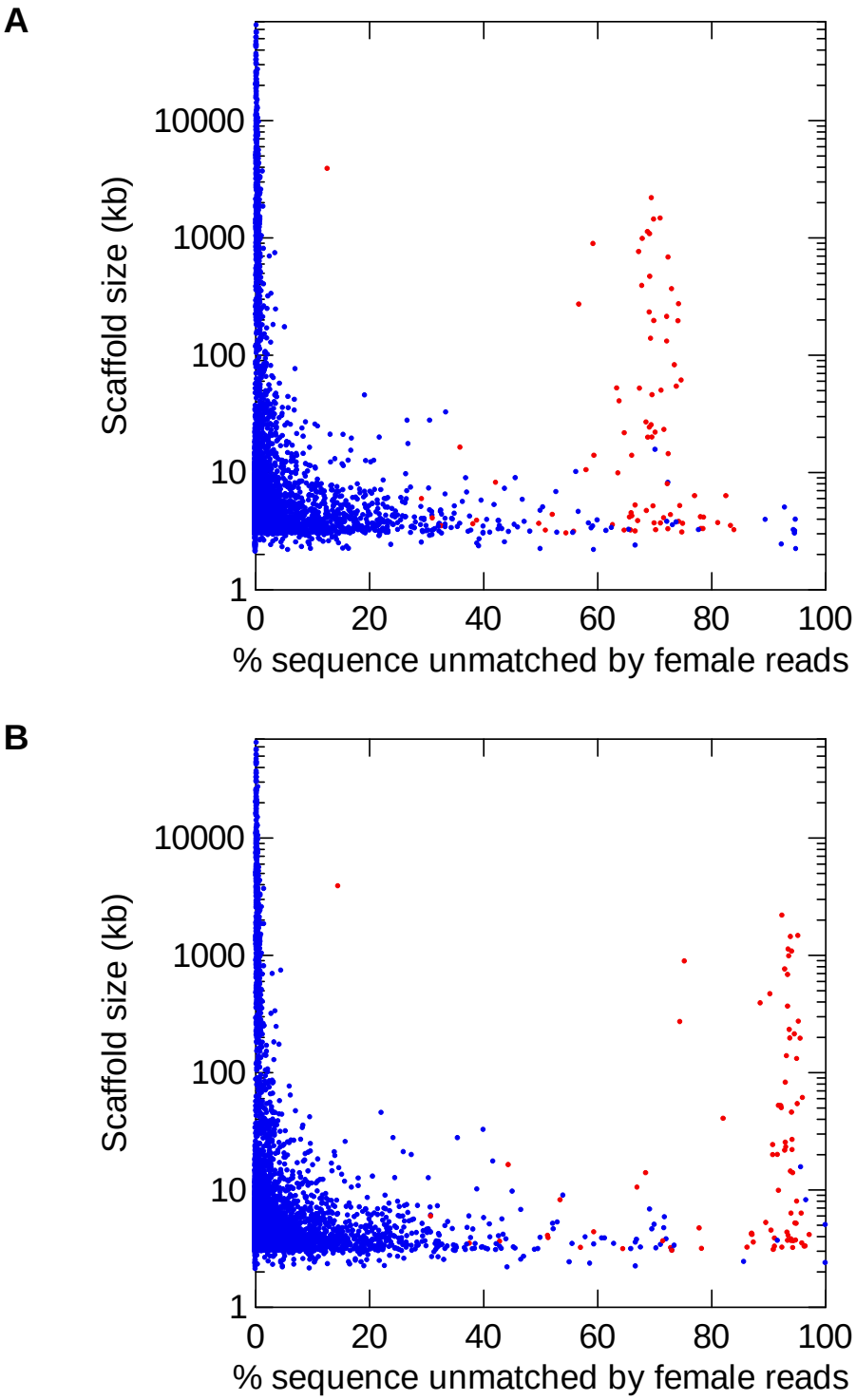
Supplemental Fig. S11. Removal of errors in the HuRef assembly and the identification of Y-linked scaffolds. (A) Analysis of the HuRef assembly without validation from male short reads. (B) Same dataset, using validation from male short reads (this is a copy of Fig. 3B, reproduced here for comparison with panel A. Note that validation eliminates many "intermediate scaffolds" that resulted either from sequencing errors or rare polymorphisms in the HuRef assembly (these "intermediate scaffolds" moved to the autosomal / X-linked peak). The few remaining intermediate scaffolds are Y-linked and resulted from misassembled segmental duplications. Red dots are scaffolds previously known to be Y-linked. Abscissa, proportion of scaffold sequence not matched by female short reads (in % unmatched single-copy *k*-mers); ordinate , scaffold size.

Supplemental Fig. S12, panels A and B



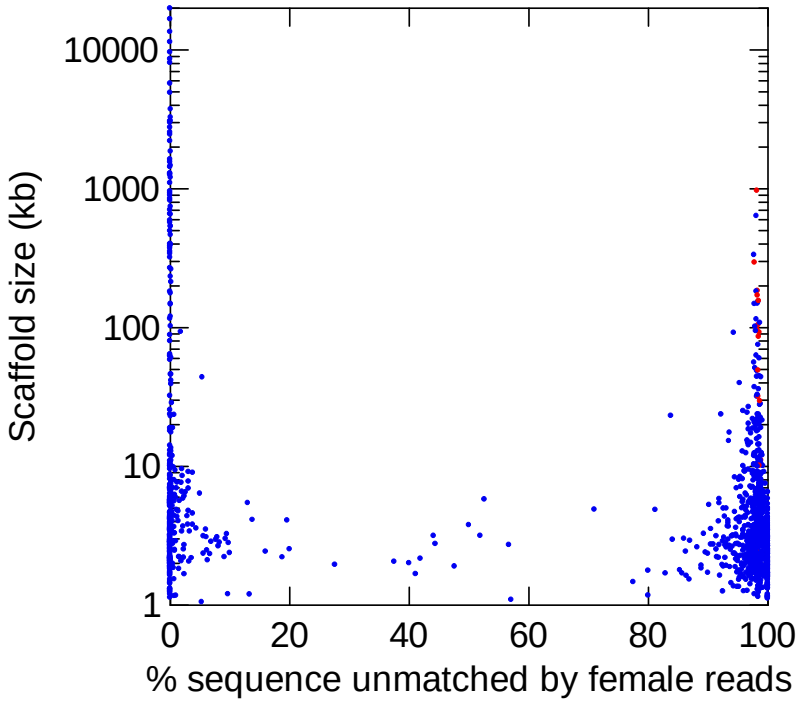
Supplemental Fig. S12. Identification of Y-linked scaffolds with low coverage female short reads. (A) Pilot experiment with *D. virilis* done in 2007 (Carvalho and Clark 2008). The non-Y scaffolds have ~55% unmatched *k*-mers, and hence the effective coverage of the genome by the female reads was 0.6-fold. Red dots are scaffolds previously known to be Y-linked, blue dots are unmapped or not-Y scaffolds, and black dots are scaffolds tested for Y-linkage with PCR (all are Y-linked). The preliminary *D. virilis* assembly used contain a misassembled piece of the Y chromosome (the red dot in the large scaffold). Abscissa, proportion of scaffold sequence not matched by female short reads (in % unmatched single-copy *k*-mers); ordinate , scaffold size. (B) Human data, using 5-fold coverage from a single female, and without removal of sequencing errors in the HuRef assembly (*i.e.*, not using male reads). There are many small, intermediate scaffolds, due to a combination of binomial sampling error and unremoved sequencing errors / polymorphisms in the HuRef assembly. Note that nearly all of them are eliminated by size cut-off of 10kb, which removes 0.6% of the sequence. Abscissa, proportion of scaffold sequence not matched by female short reads (in % unmatched single-copy *k*-mers); ordinate , scaffold size.

Supplemental Fig. S13, panels A and B

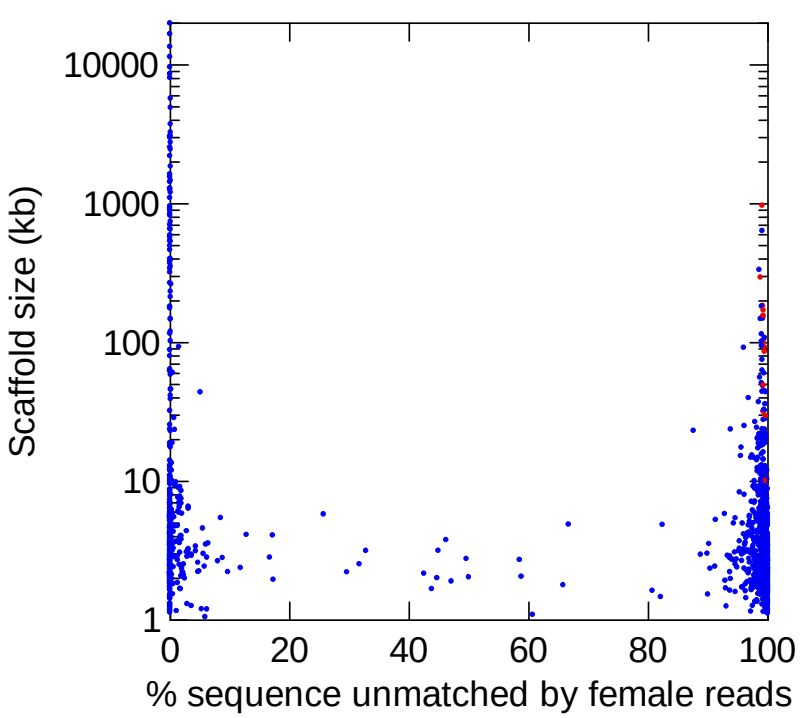


Supplemental Fig. S13, panels C and D

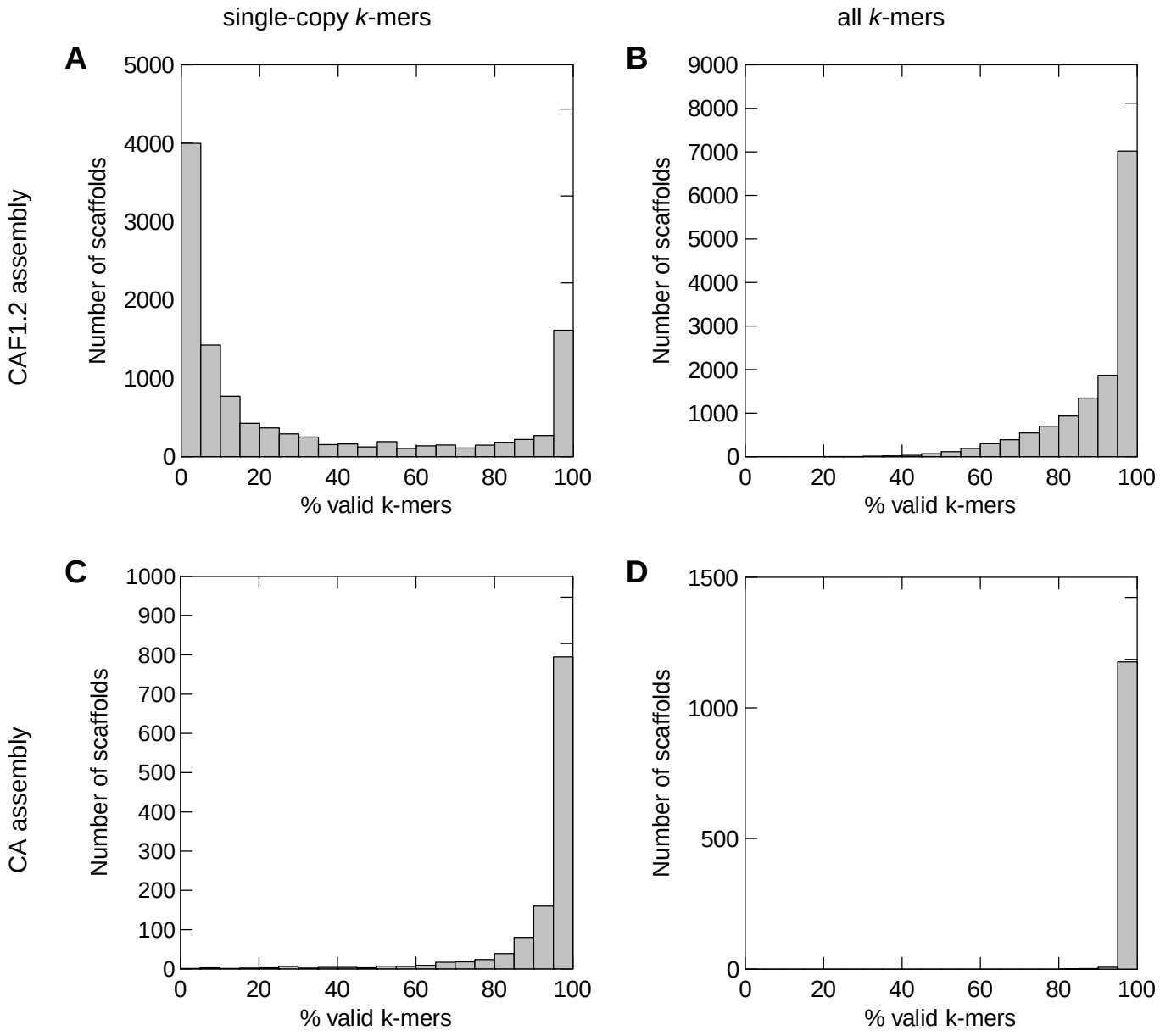
C



D

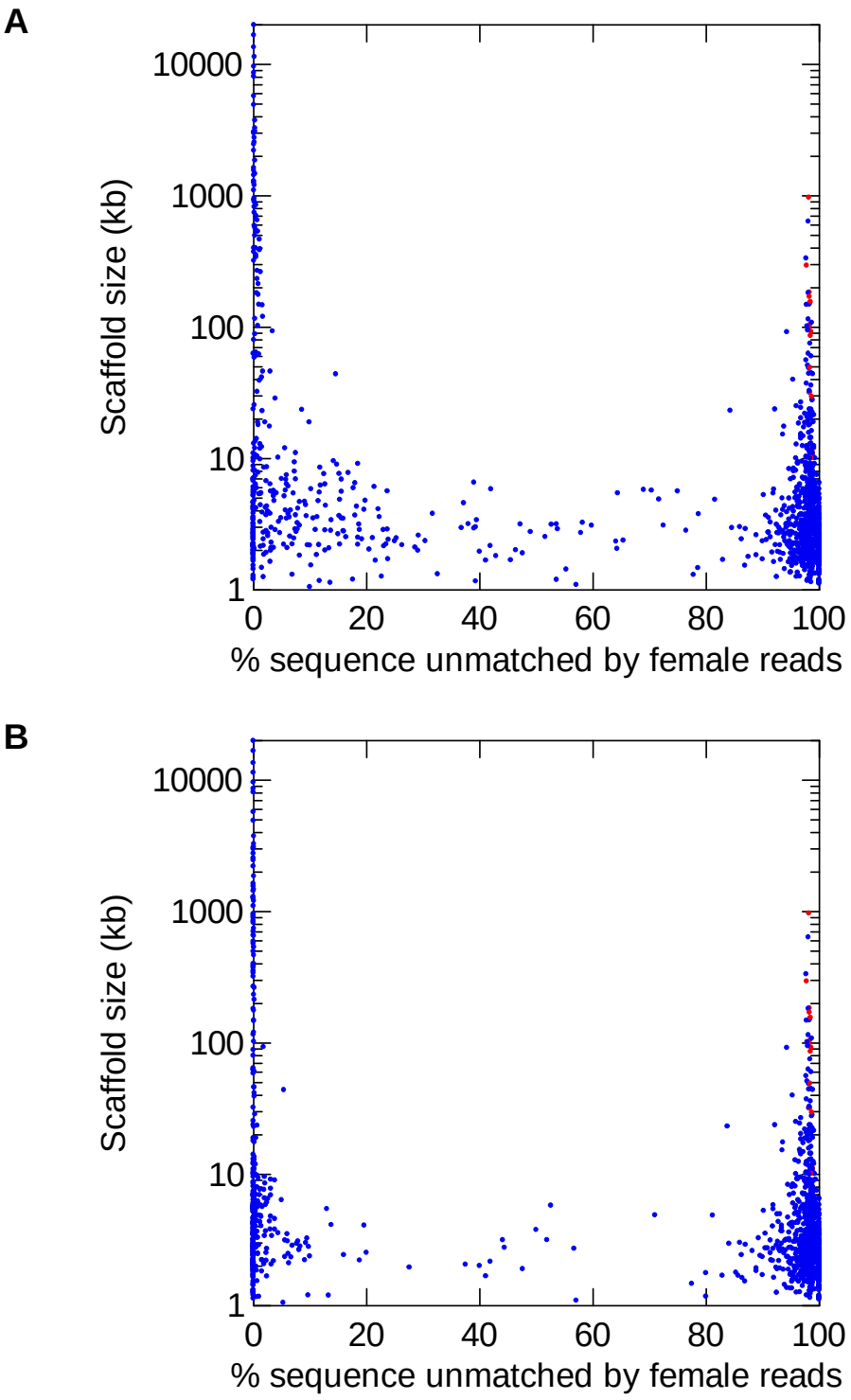


Supplemental Fig. S13. Appropriate *k*-mer sizes for the *YGS* method. (A) Human genome (HuRef assembly) using *k*-mer size of 16, without validation from male reads. (B) Same data, except for the use of *k*-mer size of 18 (this is a copy of Supplemental Fig. S11A, reproduced here for comparison). Note the worse separation between Y and not-Y scaffolds when 16-mers are used. (C) *D. virilis* genome (CA assembly) using *k*-mer size of 15 (this is a copy of Fig. 3A, reproduced here for comparison). (D) Same data, except for the use of *k*-mer size of 17. Note that the larger *k*-mer size causes little improvement in this case. Abscissa, proportion of scaffold sequence not matched by female short reads (in % unmatched single-copy *k*-mers); ordinate , scaffold size.



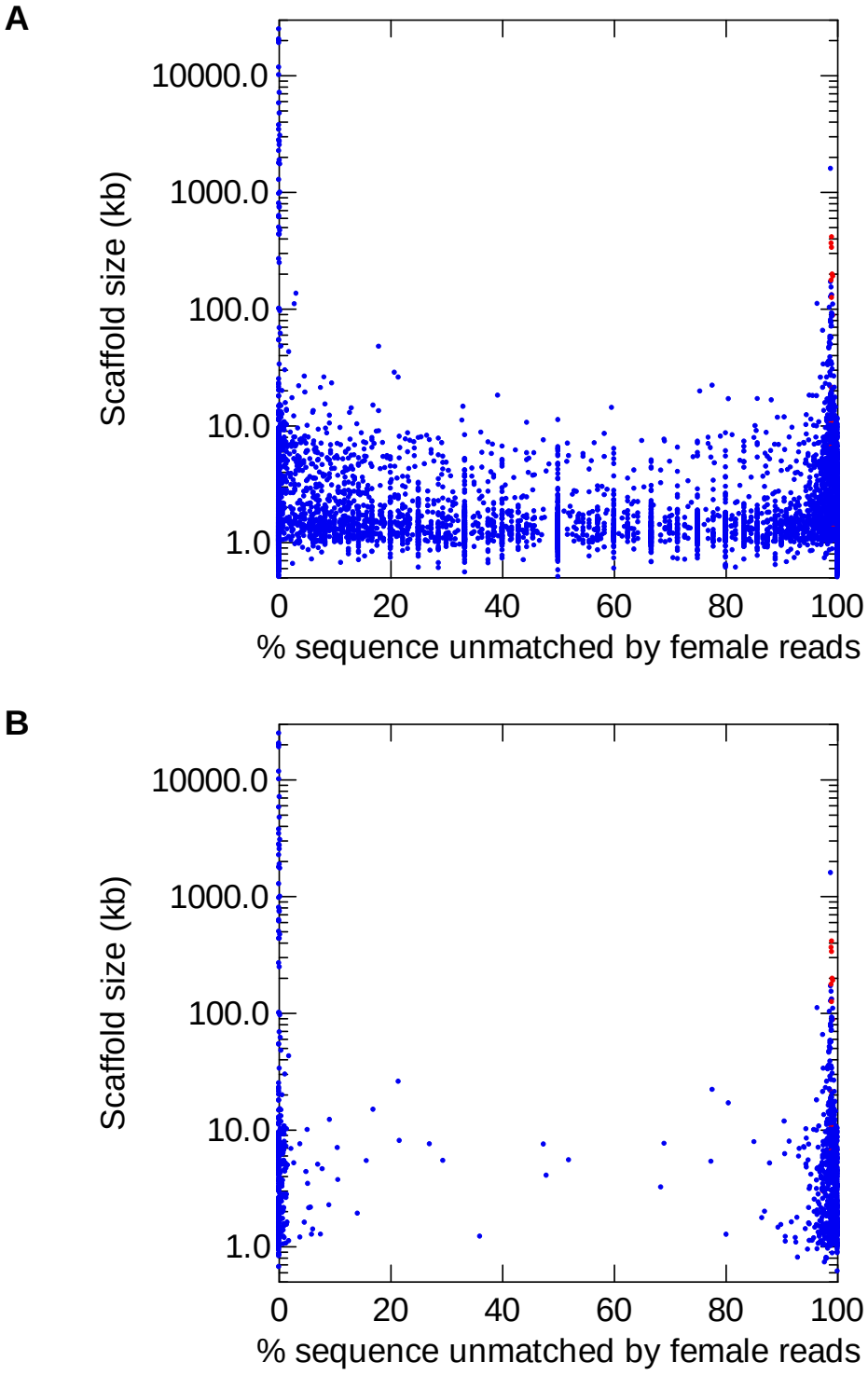
Supplemental Fig. S14. Quality assessment of two *D. virilis* genome assemblies. Two assemblies, "CAF1.2" (panels A and B) and "CA" (panels C and D) were compared to the Sanger traces (filtered at *Phred* score > 20) used to build them, in order to detect low quality scaffolds. Invalid *k*-mers are those absent from the filtered Sanger traces, and correspond to sequencing or assembling errors (see Methods section "Removal of sequencing errors in the assembled genomes"). Note the large fraction of low quality scaffolds in the CAF1.2 assembly, specially visible in the single-copy *k*-mers (panel A), but also clear when considering all *k*-mers (panel B), and their nearly absence in the CA assembly (panels C and D). Abscissa, proportion of valid *k*-mers in the scaffolds; ordinate, number of scaffolds.

Supplemental Fig. S15, panels A and B



Supplemental Fig. S15. Removal of errors in the *D. virilis* genome (CA assembly) and the identification of Y-linked scaffolds. (A) Analysis of the *D. virilis* CA assembly without validation from Sanger traces. (B) Same dataset, using validation from Sanger traces (this is a copy of Fig. 3A, reproduced here for comparison with panel A). Note that validation nearly eliminates "intermediate scaffolds" that resulted from sequencing errors (they moved to the autosomal / X-linked peak). Validation was performed as described in the Methods section "Removal of sequencing errors in the assembled genomes". Red dots are scaffolds previously known to be Y-linked, and blue dots are unmapped or not Y-linked scaffolds. Abscissa, proportion of scaffold sequence not matched by female short reads (in % unmatched single-copy *k*-mers); ordinate , scaffold size.

Supplemental Fig. S16, panels A and B



Supplemental Fig. S16. Identification of Y-linked scaffolds in the *D. virilis* CAF1.2 assembly. (A) Analysis of the *D. virilis* CAF1.2 assembly without removal of low quality scaffolds. Many "intermediate" scaffolds are low-quality and have a very small number of single-copy *k*-mers. (B) Analysis of the *D. virilis* CAF1.2 assembly after filtering low quality scaffolds (Methods section "Comparison of the *D. virilis* assemblies"). Note that the vast majority of the scaffolds that produce ambiguous results have low quality. Both panels used *k*-mer validation from Sanger traces. Abscissa, proportion of scaffold sequence not matched by female short reads (in % unmatched single-copy *k*-mers); ordinate , scaffold size.

Supplemental Tables

Supplemental Table S1. Candidate scaffolds tested experimentally for Y-linkage in *D. virilis*. Y-linkage was tested by standard PCR with male and female DNA as templates (Carvalho et al. 2000). All 15 tested candidates are Y-linked (scaffold CH940663 is very large and was tested twice). Scaffold names (CH940733, etc) are from the CAF1.2. assembly. Disruptive mutations were confirmed by Sanger sequencing in the 11 genes labeled as pseudogenes.

Scaffold	% unmatched *k*-mers	PCR primers	oC	PCR size	Gene content*	Comments
CH940733	98.3	GCTTGTTTTCGCTTCGGGAGTG CTTTGGTGCGCCTTGCTCTTTC	56	560	*CG2964*	functional
CH940663	98.8	ACCCGATGAGCCGCCTTTTTG ACGCCGCCTGCTTCTGTGC	57	900	M20 dipeptidase	functional
CH940726	98.9	TTTAAAAAGGAAGTAGAAACAGATAACCACAAGAT TGGAAGTATGAAAACAATCCTGTAACAACC	55	210	*JYalpha*	functional
CH940676	99.0	AATGCAGCGCAGTTGGAAGCAT GCACAATATGGACTGGCAAACTTCTCTA	58	347	*CG11719*	functional
CH940728	99.0	TGCTTGTGCTGAACGGTGTTCTC GGTCGATTATCCCTGAAAACAGTGG	60	~800	*SP2637*	pseudogene
CH940729	98.7	TCCTTTTGTGAGAGCAATAACAATAATAATCT GATCGAGGGCGGCATCG	53	190	*Dok*	pseudogene
CH940693	99.0	CAGCTTGAGCGCGTGTATAAGTGTAAT TTTTGTGTTTTCTTTTGGTCGCTGAA	57	289	*CG4542*	pseudogene
CH940686	98.8	GGCTGCTTGGGCTTAAAACTATTGTTA GTGCCCAATTGCCAGGGC	60	~1346	*kelch*	pseudogene
CH940663	98.8	GTCGACCATTACCCAAATGAT CGGAGATTCTCACTATCAAATAACA	52	210	*CG15012*	pseudogene
CH940761	98.3	TATTAGCGATGCCAGCAATGTGCAT CGATATTCGTAGAGATTGAATTTCTTCCATACC	60	814	*CG2146*	pseudogene
CH940721	98.7	AGCCACCGCCACTGCGAA GGAAGCGTTGGTAAATCGCAAGTAA	61	654	*CG3969*	pseudogene
CH940691	99.2	GTTTTTTTCAGAAGCCCGATTTATTAAATT TGGGAAACTGCGTGCGTGTG	57	436	*CG11001*	pseudogene
CH942352	99.7	TCTAATGTGGCGCCGCAGAGA GTGGATGCAGCTCCTCCTCCTT	63	1136	*CG13425*	pseudogene
CH940717	98.1	GGTGCCCGCTCAAAAGATCG CAGCTGCATTAATCCATAATGAGCCT	61	~750	*CG11100*	pseudogene
CH941566	99.7	GGCCGTCGAGCGCAAGG GAAAGTCTGAGTGTAGTAATAAATGCTAGGTTTGT	60	~690	*CG10308*	pseudogene
CH942814, many	98.6	ATGATAACAAATTTACTTTCTGTATTCAACCA GCAGATTCAAGAACTAATAAAGCAATTAGG	53	568	*ATPase 6* (mitoch.)	?

* *D. melanogaster* ortholog

Supplemental Table S2. Y-linkage across the 12 *Drosophila* species of the four new Y-linked genes of *D. virilis*. Y-linkage was tested by standard PCR with male and female DNA as templates (Carvalho et al. 2000). Unabridged species names (in the order of appearance) are: *D. melanogaster*, *D. simulans*, *D. sechellia*, *D. erecta*, *D. yakuba*, *D. ananassae*, *D. pseudoobscura*, *D. persimilis*, *D. willistoni*, *D. mojavensis*, *D. virilis*, and *D. grimshawi*. "+" Y-linked gene ; " - " autosomal or X-linked gene; "0" gene absent from the genome.

Gene	*mel* ortholog	*mel*	*sim* *sec*	*ere* *yak*	*ana*	*pse* *per**	*wil*	*moj*	*vir*	*gri*
GJ19835	-	0†	0	0	0	-	-	+	+	+
GJ19633	*CG11719*	-	-	-	-	-	-	+	+	-
GJ11126	*CG2964*	-	-	-	-	-	-	+	+	-
GJ18574	*JYalpha*	-	-	-	+	-	+	+	+	+

* As described in Carvalho and Clark (2005) the Y chromosome became part of an autosome in the *D. pseudoobscura* lineage.

† The ortholog of the *GJ19835* gene was lost in the ancestor of the melanogaster group of species (Supplemental Fig. S1).

.

Supplemental Table S3. New euchromatic sequences in the human Y.

Scaffold (GenBank accession #)	Size (kb)	New sequence coordinates	Size (kb)*	Location in the reference Y	Main BlastN hits (% identity)	Main BlastX hits
DS486628	66	1-6,390	6.4	gap between NT_011875 and NT_011903 (extends NT_011903 to left)	chrY (79%), chr5 (86%)	TCEB1, UQCC
DS486519	216	200,399-end	13.2	gap between NT_011875 and NT_011903 (extends NT_011875 to right)	chrY (79%), chr8 (84%), chr10 (85%), chr4 (85%)	SAMD4B, UTY, LOC100507556
DS486512	228	208,122-end	16.8	gap between NT_011896 and NT_086998 (extends NT_086998 to left)	chrY (94%), many chromosomes [4, 8, 9, 10, *etc.*]	UBE2D4, EBPL, LOC100128274
DS486227	2,350	1,628,501-1,637,498	7.4	internal insertion in NT_011875 (position 692637)	chrX (91%), chr6 (87%), chr13 (79%)	ZNF793, FLJ13197, SLC37A2, ARSD
DS486227	2,350	2,315,007-end	32.3	gap between NT_113819 and NT_011875 (extends NT_011875 to left)	32% simple repeats + satellite + low complexity	PCSK5

* Sizes do not include gaps. For example, the internal insertion in DS486227 has 9.4 kb including a gap and 7.4 kb excluding it. Its real size in the finished sequence (BAC clone AC245170) is 11 kb.

Supplemental Table S4. Summary data of the 119 HuRef Y-linked scaffolds.

Scaffold (GenBank accession #)	Size (kb)	% unmatched by female reads*	Mapping to reference Y†	% unmatched by reference Y‡	Differential male reads hits¶	Comments
DS486171	3874	14.3	yes	23.1	327150	XTR region
DS486227	2177	92.3	yes	1.3	20436	new sequence (40 kb)
DS486351	886	75.2	yes	1.8	7303	XTR region
DS486286	1432	93.8	yes	0.3	5598	new sequence (several small patches < 2kb)§
DS486512	211	94.6	yes	11.3	5215	new sequence (17 kb)
DS486519	194	95.5	yes	10.0	4149	new sequence (13 kb)
DS486428	466	90.2	yes	0.6	2386	new sequence (few small patches < 2kb)§
DS486628	61	96.0	yes	11.6	2029	new sequence (6 kb)
DS486288	1459	95.2	yes	0.2	1069	new sequence (2kb)§
DS486319	1120	93.5	yes	0.1	1037	possible misassembly (2kb)‖
DS487109	14	93.8	yes	11.3	952	new sequence (1.5kb)§
DS486381	757	92.9	yes	0.3	727	possible misassembly (2 kb)‖
DS486902	16	43.8	yes	6.5	439	XTR region
DS486345	978	93.6	yes	0.0	269	
DS489675	4	34.7	yes	8.2	233	XTR region
DS486616	40	81.8	yes	0.5	169	
DS486329	1075	94.0	yes	0.0	131	
DS486460	270	74.2	yes	0.2	124	
DS487533	3	69.8	yes	0.0	122	
DS486478	271	95.3	yes	0.1	84	
DS486864	14	68.0	yes	1.8	64	
DS486395	679	93.3	yes	0.0	62	
DS487476	3	57.1	yes	14.7	55	XTR region
DS486497	231	93.7	yes	0.0	35	
DS486457	364	93.4	yes	0.0	33	
DS486451	388	88.5	yes	0.1	31	
DS488948	4	87.2	yes	0.0	19	
DS486553	138	93.2	yes	0.1	16	
DS490089	3	93.1	yes	3.9	13	
DS486601	82	92.9	yes	0.0	12	

Supplemental Table S4 (continuation)

DS486531	195	93.8	yes	0.0	7
DS486555	131	95.0	yes	0.0	6
DS488783	4	92.9	yes	0.0	5
DS487199	8	53.3	yes	0.1	4
DS486664	52	92.2	yes	0.0	3
DS489232	4	90.7	yes	0.3	3
DS489480	4	63.9	yes	0.7	2
DS489939	3	97.0	yes	1.6	2
DS487528	3	91.1	yes	0.0	1
DS488791	4	97.7	yes	1.6	1
DS489942	3	96.3	yes	3.2	1
DS486658	54	95.1	yes	0.0	-
DS486663	52	91.3	yes	0.3	-
DS486667	50	92.3	yes	0.1	-
DS486686	46	94.1	yes	0.0	-
DS486820	27	94.2	yes	0.1	-
DS486836	25	93.0	yes	0.0	-
DS486847	24	90.8	yes	0.0	-
DS486863	23	93.1	yes	0.0	-
DS486884	22	94.2	yes	0.0	-
DS486887	22	92.9	yes	0.0	-
DS486922	20	91.5	yes	0.0	-
DS486926	20	90.7	yes	0.0	-
DS487120	10	66.2	yes	0.2	-
DS487149	14	94.3	yes	0.0	-
DS487291	6	95.6	yes	0.3	-
DS487408	4	49.7	yes	0.0	-
DS487638	10	91.8	yes	0.0	-
DS487772	8	95.0	yes	0.0	-
DS488005	6	94.2	yes	0.3	-
DS488082	6	29.9	yes	0.0	-
DS488338	5	94.7	yes	0.0	-
DS488350	5	94.9	yes	0.0	-
DS488544	5	77.1	yes	0.3	-
DS488678	5	90.4	yes	0.0	-
DS488767	4	58.6	yes	0.0	-
DS488861	4	87.1	yes	0.0	-
DS488997	4	93.5	yes	0.5	-
DS489296	4	94.0	yes	0.0	-
DS489402	4	94.8	yes	0.0	-

Supplemental Table S4 (continuation)

DS489436	4	93.3	yes	0.4	-	
DS489438	4	94.0	yes	0.0	-	
DS489473	4	94.4	yes	0.0	-	
DS489579	4	87.1	yes	0.0	-	
DS489648	4	95.8	yes	0.3	-	
DS490084	3	85.4	yes	1.4	-	
DS490135	3	94.1	yes	0.3	-	
DS490240	3	78.3	yes	0.0	-	
DS490267	3	62.7	yes	0.3	-	
DS490374	3	90.9	yes	0.0	-	
DS490483	3	72.0	yes	0.4	-	
DS487465	4	48.7	yes	0.4	-	XTR region
DS487515	4	42.9	yes	0.0	-	XTR region
DS488309	5	89.5	yes	0.0	-	XTR region
DS488926	4	83.6	yes	6.0	-	XTR region
DS489331	4	66.9	no	100.0	249442	Yq12; DZY2 satellite
DS490348	3	49.0	no	100.0	205296	DZY2 satellite
DS487367	5	53.0	no	82.2	148188	Yq12; DYZ1 satellite
DS487401	5	50.2	no	82.3	103119	Yq12; DYZ1 satellite
DS487348	3	70.3	no	99.0	49060	Yq12; DYZ1 satellite
DS488597	5	68.6	no	100.0	21072	Yq12; DZY2 satellite
DS488387	5	68.7	no	97.3	20753	similar to satellite III
DS490176	3	70.3	no	100.0	14871	Yq12; DYZ1 satellite
DS490265	3	65.9	no	98.5	13581	similar to satellite III
DS488106	6	71.2	no	99.3	12884	similar to satellite III
DS487228	7	67.8	no	97.5	12563	similar to satellite III
DS487584	4	62.9	no	98.7	10517	similar to satellite III
DS489220	4	61.3	no	98.9	10341	DYZ1 satellite
DS486638	33	34.1	no	99.8	8682	
DS490420	3	70.9	no	99.6	7151	
DS487536	4	72.1	no	99.6	6602	
DS487569	3	70.1	no	97.8	5502	
DS490587	2	57.8	no	98.3	5211	
DS486680	16	95.4	no	73.7	4768	
DS486878	8	96.4	no	90.3	3290	
DS488053	6	25.4	no	99.0	1931	
DS488608	5	37.3	no	99.8	1758	
DS489752	3	57.8	no	99.4	1629	

Supplemental Table S4 (continuation)

DS487648	10	40.6	no	98.7	1217	
DS489488	4	31.8	no	97.3	986	
DS486918	17	38.1	no	100.0	960	olfactory receptor
DS487107	9	48.6	no	100.0	906	
DS490088	3	33.2	no	99.0	903	
DS488076	6	37.2	no	100.0	844	
DS489434	4	91.5	no	60.8	674	
DS490239	3	41.4	no	100.0	644	
DS489225	4	57.8	no	100.0	627	
DS489734	3	42.3	no	100.0	600	
DS488780	4	27.1	no	100.0	481	

* Proportion of sequence not matched by female short reads (in % unmatched single-copy *k*-mers). Data plotted in Fig. 3B.

† Mapping to reference Y was inferred from column five. 85 HuRef scaffolds are largely or totally included into the reference Y sequence ("yes") and 34 scaffolds contain mostly new sequence ("no").

‡ Proportion of sequence not matched by the reference Y chromosome sequence (in % unmatched single-copy *k*-mers). Data plotted in Supplemental Fig. S5 (Methods section "Detection and validation of new human Y-linked sequences").

¶ Number of hits from male short reads that failed to align to the reference Y sequence (Methods section "Detection and validation of new human Y-linked sequences").

§ These small regions were not further studied.

‖Misassembly suspected due to hits from female short reads completely overlapping hits from male short reads.

Supplemental Table S5. Application of the *YGS* method to a short-read assembly (*D. kikkawai*): result for known Y-linked and autosomal genes.

Scaffold	Gene content*	Known location†	% unmatched *k*-mers
KB458921	*kl-2* (N-term)	Y	97.9
AFFH02000026	*kl-2* (middle)	Y	99.8
KB459848	*kl-2* (C-term)	Y	15.7
AFFH02002333	*PPr-Y* (N-term)	Y	99.5
AFFH02002060	*PPr-Y* (middle)	Y	98.7
AFFH02002372	*PPr-Y* (C-term)	Y	98.9
AFFH02002301	*PRY* (N-term)	Y	99.0
AFFH02002310	*PRY* (C-term)	Y	99.7
AFFH02001073	*WDY* (N-term)	Y	100.0
AFFH02002318	*WDY* (C-term)	Y	99.6
KB459683	*CG2964*	autosome	0.2
KB459690	*CG11719*	autosome	0.2
KB459791	*Hex-A*	autosome	0.4
KB459688	*Adh*	autosome	0.1

* *D. melanogaster* ortholog

† Experimentally determined location (PCR test).

Captions for Supplemental Data Files S1 to S10

Supplemental Data File S1 (separate file)
Illumina_reads.txt This plain text file lists the male and female Illumina fastq files used to detect Y-linked sequences in the HuRef human assembly (data came from the 1000 Genomes Project).

Supplemental Data File S2 (separate file)
YGS.pl This is a program written in PERL that implements the main steps of the *YGS* method of identification of Y-linked scaffolds.

Supplemental Data File S3 (separate file)
Poisson_regression.R This is a program written in R language that implements the Poisson regression that tests the statistical significance of the gene gain / gene loss ratio for the *D. virilis* and *D. melanogaster* data.

Supplemental Data File S4 (separate file)
AF_vir_mel_data.txt This is the combined data file (*D. virilis* + *D. melanogaster*) used by the Poisson_regression.R program (Assumption-Free method).

Supplemental Data File S5 (separate file)
analytical_vir_mel.xls This MS-EXCEL file implements the analytical treatment of the ascertainment bias, and estimates the unbiased ratio of the gain rate to loss rate, for the *D. virilis*, *D. melanogaster*, and for the combined data of both species, as described in the Supplemental Material section "Statistical tests of gene gains and gene losses" (Homogeneous Gain Loss method).

Supplemental Data File S6 (separate file)
HGL_vir_data.txt This is the *D. virilis* data file used by the Poisson_regression.R program (Homogeneous Gain Loss method).

Supplemental Data File S7 (separate file)
HGL_mel_data.txt This is the *D. melanogaster* data file used by the Poisson_regression.R program (Homogeneous Gain Loss method).

Supplemental Data File S8 (separate file)
HGL_vir_mel_data.txt This is the combined data file (*D. virilis* + *D. melanogaster*) used by the Poisson_regression.R program (Homogeneous Gain Loss method).

Supplemental Data File S9 (separate file)
ApproxBayes_vir.R This is a program written in R language that implements the computer simulations of the gain and loss of genes in the *Drosophila* Y chromosome using the *D. virilis* data, as described in the Supplemental Material section "Statistical tests of gene gains and gene losses". It produces approximate Bayesian estimates of the posterior densities of the rates of gene gain and loss. The run time is approximately 12 hours in a 2.33 GHz computer.

Supplemental Data File S10 (separate file)
virCAF12_scafs_classification.pdf. This file lists all 13,530 scaffolds of the virCAF1.2 assembly, along with their classification (Y-linked, not Y-linked, intermediate, contaminant, low-quality, and small).

Supplemental References:

Alves-Silva, J., J.M.C. Ribeiro, J.V.D. Abbeele, G. Attardo, Z. Hao, L.R. Haines, M.B. Soares, M. Berriman, S. Aksoy, and M.J. Lehane. 2010. An insight into the sialome of *Glossina morsitans morsitans*. *BMC Genomics* **11:** 213.

Beaumont, M.A. 2010. Approximate Bayesian Computation in Evolution and Ecology. *Annual Review of Ecology, Evolution, and Systematics* **41:** 379-406.

Beaumont, M.A., W. Zhang, and D.J. Balding. 2002. Approximate Bayesian computation in population genetics. *Genetics* **162:** 2025-2035.

Bhutkar, A., S. Russo, T. Smith, and W. Gelbart. 2007. Genome-scale analysis of positionally relocated genes. *Genome Research* **17:** 1880-1887.

Carvalho, A.B. and A.G. Clark. 2005. Y chromosome of *D. pseudoobscura* is not homologous to the ancestral *Drosophila* Y. *Science* **307:** 108-110.

Carvalho, A.B. and A.G. Clark. 2008. Efficient identification of Drosophila Y-chromosome sequences by short-read sequencing. *The 49th Annual Drosophila Research Conference*, abstract 112, pp. 124. The Genetics Society of America, San Diego, USA.

Carvalho, A.B., B.P. Lazzaro, and A.G. Clark. 2000. Y chromosomal fertility factors kl-2 and kl-3 of Drosophila melanogaster encode dynein heavy chain polypeptides. *Proc. Natl. Acad. Sci. U.S.A.* **97:** 13239-13244.

Celniker, S.E., L.A. Dillon, M.B. Gerstein, K.C. Gunsalus, S. Henikoff, G.H. Karpen, M. Kellis, E.C. Lai, J.D. Lieb, D.M. MacAlpine et al. 2009. Unlocking the secrets of the genome. *Nature* **459:** 927-930.

Gastmann, O., P. Burfeind, E. Gunther, H. Hameister, C. Szpirer, and S. Hoyer-Fender. 1993. Sequence, expression, and chromosomal assignment of a human sperm outer dense fiber gene. *Mol Reprod Dev* **36:** 407-418.

Koerich, L.B., X. Wang, A.G. Clark, and A.B. Carvalho. 2008. Low conservation of gene content in the *Drosophila* Y chromosome. *Nature* **456:** 949-951.

Marçais, G. and C. Kingsford. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics* **27:** 764-770.

Masly, J.P., D. Jones, M.A. Noor, J. Locke, and H.A. Orr. 2006. Gene transposition as a cause of hybrid sterility in *Drosophila*. *Science* **313:** 1448-1450.

McQuilton, P., S.E. St Pierre, and J. Thurmond. 2012. FlyBase 101--the basics of navigating FlyBase. *Nucleic Acids Res* **40:** D706-714.

Okamura, H., J.C. Yasuhara, D.M. Fambrough, and K. Takeyasu. 2003. P-type ATPases in *Caenorhabditis* and *Drosophila*: implications for evolution of the P-type ATPase subunit families with special reference to the Na,K-ATPase and H,K-ATPase subgroup. *J Membr Biol* **191:** 13-24.

Przeworski, M. 2003. Estimating the time since the fixation of a beneficial allele. *Genetics* **164:** 1667-1676.

R Core Team. 2012. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rawlings, N.D. 2009. A large and accurate collection of peptidase cleavages in the MEROPS database. *Database: the journal of biological databases and curation* **2009**.

Schafer, M., D. Borsch, A. Hulster, and U. Schafer. 1993. Expression of a gene duplication encoding conserved sperm tail proteins is translationally regulated in *Drosophila melanogaster*. *Molecular and Cellular Biology* **13:** 1708.

Tavare, S., D.J. Balding, R.C. Griffiths, and P. Donnelly. 1997. Inferring coalescence times from DNA sequence data. *Genetics* **145:** 505-518.

Wasbrough, E.R., S. Dorus, S. Hester, J. Howard-Murkin, K. Lilley, E. Wilkin, A. Polpitiya, K. Petritis, and T.L. Karr. 2010. The *Drosophila melanogaster* sperm proteome-II (DmSP-II). *J Proteomics* **73:** 2171-2185.