High rate of translocation-based gene birth on the *Drosophila* Y chromosome

Ray Tobler[a,b,1], Viola Nolte[a], and Christian Schlötterer[a,2]

[a]Institut für Populationsgenetik, Vetmeduni Vienna, 1210 Vienna, Austria; and [b]Vienna Graduate School of Population Genetics, Vetmeduni Vienna, 1210 Vienna, Austria

Edited by Andrew G. Clark, Cornell University, Ithaca, NY, and approved September 26, 2017 (received for review April 21, 2017)

The Y chromosome is a unique genetic environment defined by a lack of recombination and male-limited inheritance. The *Drosophila* Y chromosome has been gradually acquiring genes from the rest of the genome, with only seven Y-linked genes being gained over the past 63 million years (0.12 gene gains per million years). Using a next-generation sequencing (NGS)-powered genomic scan, we show that gene transfers to the Y chromosome are much more common than previously suspected: at least 25 have arisen across three *Drosophila* species over the past 5.4 million years (1.67 per million years for each lineage). The gene transfer rate is significantly lower in *Drosophila melanogaster* than in the *Drosophila simulans* clade, primarily due to Y-linked retrotranspositions being significantly more common in the latter. Despite all Y-linked gene transfers being evolutionarily recent (<1 million years old), only three showed evidence for purifying selection ($\omega \leq 0.14$). Thus, although the resulting Y-linked functional gene acquisition rate (0.25 new genes per million years) is double the longer-term estimate, the fate of most new Y-linked genes is defined by rapid degeneration and pseudogenization. Our results show that Y-linked gene traffic, and the molecular mechanisms governing these transfers, can diverge rapidly between species, revealing the *Drosophila* Y chromosome to be more dynamic than previously appreciated. Our analytical method provides a powerful means to identify Y-linked gene transfers and will help illuminate the evolutionary dynamics of the Y chromosome in *Drosophila* and other species.

Y chromosome | evolution | *Drosophila* | retrocopies | transposition

The heterochromatic, repeat-laden nature of the *Drosophila* Y chromosome makes it difficult to analyze, such that its evolution is still poorly understood. Only 12 Y-linked genes have been discovered on the *Drosophila melanogaster* Y chromosome, all of which arose by transfers from autosomes onto the Y chromosome (1–4). Because only transfers that produce functional Y-linked copies can be detected over long evolutionary timescales, we hypothesized that the underlying primary gene transfer rate may be considerably higher. To investigate this, we developed a method to detect recent gene transfers onto the Y chromosome (GeTYs). We reasoned that mapping short reads from inbred males to a female reference genome would produce polymorphisms in genes that had spawned a Y-linked duplication, whereas the same genomic region should be homozygous for short reads in females from the same inbred strain. Following this idea, we developed a metric for identifying Y-linked transfers (*Methods* and *SI Methods*) and applied it to two inbred strains from three *Drosophila* species: *D. melanogaster*, *Drosophila simulans*, and *Drosophila mauritiana*, which diverged between ~0.24 Mya [*D. simulans* and *D. mauritiana* (5)] and 5.4 Mya [*D. simulans clade* and *D. melanogaster* (6)]. Unlike other methods that exploit sex-specific short read alignments to identify Y chromosome sequences (3, 7), our method does not require preassembled Y contigs.

Results

Y-Linked Transfer Properties and Pipeline Validation. Our method detected numerous putative Y-specific sequences that mapped to feminized reference genomes from three *Drosophila* species (Fig. 1). Consistent with the high repeat content of the Y chromosome, many clusters of SNPs were located in or near heterochromatic regions that overlapped transposable elements (TEs) (Fig. 1). Restricting our analysis to regions that lacked TEs and contained at least five Y-specific SNPs resulted in 66 incipient Y-linked transfers. After combining incipient transfers from closely neighboring regions, we obtained 45 unique Y-linked consensus transfers across the three species (Table 1 and Datasets S1 and S2). Twenty-five of these consensus transfers were GeTYs, including six that were shared between *D. simulans* and *D. mauritiana* (Fig. 2). The set of donor genes underlying the Y-linked transfers were broadly dispersed over the genome and were not enriched with respect to functional category or male-biased expression (Datasets S3 and S4).

To validate our analytical pipeline, we performed a combination of in vitro and in silico tests. First, we estimated the false discovery rate (FDR) of our pipeline by rerunning it in full after reversing the role of the two sexes (*SI Methods*). Because no transfers were detected in this sex-reversed pipeline, the estimated FDR in the present study is indistinguishable from zero. Second, we reasoned that the donor regions of the Y-linked transfers should have significantly elevated coverage in males relative to females of the same strain, after weighting the coverage to account for variation across samples and chromosomes (*SI Methods*). Indeed, the male:female weighted coverage ratio (WCR) was consistently higher in the detected transfer regions than expected according to the empirical WCR distribution (and this was significant for more than half of the incipient transfers; Dataset S1), suggesting our pipeline accurately

Significance

Using a powerful method that uses inexpensive short reads to detect Y-linked transfers, we show that gene traffic onto the *Drosophila* Y chromosome is 10 times more frequent than previously thought and includes the first Y-linked retrocopies discovered in these taxa. All 25 identified Y-linked gene transfers were relatively young (<1 million years old), although most appear to be pseudogenes because only three of these transfers show signs of purifying selection. Our method provides compelling evidence that the *Drosophila* Y chromosome is a highly challenging and dynamic genetic environment that is capable of rapidly diverging between species and promises to reveal fundamental insights into Y chromosome evolution across many taxa.

Author contributions: C.S. designed research; R.T., V.N., and C.S. performed research; R.T. and V.N. analyzed data; and R.T., V.N., and C.S. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: Sequence data are available from the European Nucleotide Archive (ENA) (accession no. PRJEB22850). Custom scripts used for analyses are available on Dryad (doi:10.5061/dryad.8ph59).

[1]Present address: Australian Centre for Ancient DNA, School of Biological Sciences, University of Adelaide, Adelaide, SA 5005, Australia.

[2]To whom correspondence should be addressed. Email: schlotc@gmail.com.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1706502114/-/DCSupplemental.

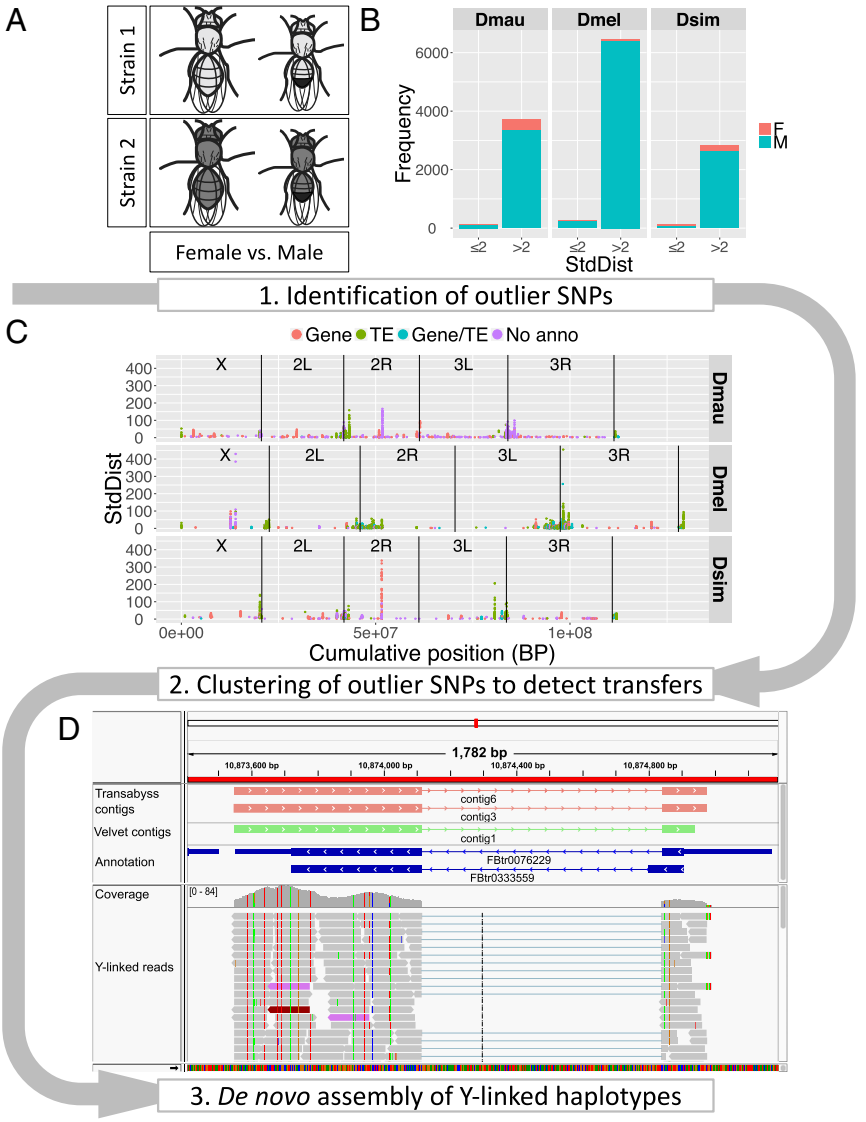

Fig. 1. Schematic overview of the Y-linked transfer detection pipeline. In step 1, two separate Cochran–Mantel–Haenszel (CMH) tests were performed and then combined to identify outlier SNPs; the first CMH test contrasted sexes with the strains (different colored flies) as replicates (*A*), and the second CMH test contrasted the strains with sexes as replicates. For all three species, more outlier SNPs (StdDiff > 2) were detected for male-specific variants (M) than for female-specific variants (F) (*B*), indicating that our pipeline was accurately identifying Y transfers. In step 2, male-specific outlier SNPs are grouped into clusters (*C*). The annotation of genomic regions containing outlier SNPs is indicated by different color codes. Although GeTYs are broadly dispersed across the genome of each species, TE peaks typically cluster within heterochromatic regions. In step 3, reads containing Y-linked variants were used in the de novo assembly of the Y-linked haplotype for each incipient transfer. No anno, no recorded annotation; TE, transposable element. (*D*) The Integrative Genomics Viewer (IGV) screen shot for the read alignment (*Bottom*) and subsequent de novo assembled Y-linked haplotype [green (Velvet) and red (Transabyss) bars; *Top*], relative to the donor gene annotation for GeTY (blue bars; *Top*). The final step of the pipeline involved the iterative aggregation of incipient transfers lying within 200 kb of one another into a single consensus transfer.

identified Y-linked transfers. Finally, we confirmed the existence of all Y-linked transfers from *D. mauritiana* and *D. simulans* by determining that the associated Y-linked sequences generated PCR amplicons in males only (Dataset S5 and *SI Methods*). Notably, our results suggested that a handful of Y-linked transfers had been involved in additional bouts of duplication on the Y chromosome (Table 1 and *SI Methods*). One of these GeTYs (Dsim_2R_9.41) also showed evidence for subsequent transfer onto the autosome or X chromosome (Fig. S1 and *SI Methods*), supporting a previous report that the Y chromosome is also an occasional source for gene transfers to other chromosomes in *Drosophila* (8).

Y-Linked Transfer Haplotype Assembly. To facilitate additional analyses of the Y-linked transfers, we reconstructed the Y-linked haplotype for each of the incipient transfers by extracting all reads mapping to the donor region that contained putative Y-linked alleles, then using these reads to de novo assemble the translocated sequence (*SI Methods*). We checked the quality of our de novo assemblies by looking in more detail at the reconstructed haplotype for GeTY Dmel_3R_20.35, for which a 200-kb contig bearing the full translocation is publically available (9) (*SI Methods*). GeTY Dmel_3R_20.35 was previously reported to be a DNA translocation that contains the youngest functional Y-linked gene described for *D. melanogaster* to date (*FDY*; parental ortholog: *vig2*) (9). The original DNA translocation also included additional genes that show evidence of pseudogenization (*Moc2/CG42503*, *Caliban*, and *Bili*) (9). Our four Y-linked haplotypes from this region produced near-perfect alignments with the published 200-kb Y-linked

Table 1. Summary of Y-linked transfers

Species	Transfer ID	Transfer type	Donor(s)	ω	Expression	Age (Ky)
D. mauritiana	Dmau_2L_6.65	Ambig	Hrb27C	1.02,Filt	NA	673 [309,1885]
	Dmau_2L_9.1	Ambig	numb	nORF	NA	460 [211,1287]
	Dmau_2L_12.3	*DNA*	*CG5787;Pih1D1*	*Filt;Filt,Filt*	*NA*	*609 [280,1706]*
	Dmau_2L_15.71	*RNA*	*CG4455*	*1.38*	*NA*	*517 [237,1446]*
	Dmau_2R_0.15*	DNA	NA	NA	NA	429 [197,1201]
	Dmau_2R_4.35	*RNA*	*14–3-3zeta*	*nORF*	*NA*	*383 [176,1072]*
	Dmau_2R_8.65	DNA	L;ttv;LamC	nORF;nORF;0.14†	NA	153(240) [70(110),429(672)]
	*Dmau_2R_9.41**	*RNA*	*SRPK*	*Filt*	*NA*	*429 [197,1201]*
	Dmau_2R_13.28	DNA	CG7229	Filt	NA	80 [37,224]
	Dmau_2R_18.4	RNA	CG3511	0.95	NA	147 [68,413]
	Dmau_2R_19.04	DNA	NA	NA	NA	305 [140,854]
	Dmau_3L_0.16	RNA	CG13876	0.30	NA	158 [73,443]
	Dmau_3L_2.15	DNA	NA	NA	NA	98 [45,273]
	Dmau_3L_7.55	*DNA*	*CG7492;Ank2*	*nORF;Filt*	*NA*	*213 [98,597]*
	Dmau_3L_22.2	DNA	NA	NA	NA	957 [439,2679]
	Dmau_3R_0.34*	DNA	NA	NA	NA	380 [174,1064]
	Dmau_3R_1.23	DNA	Nmdar1;dmau_PG00479	nORF;nORF	NA	282 [130,790]
	Dmau_3R_2.13	DNA	NA	NA	NA	28 [13,78]
	Dmau_3R_13.94	*DNA*	*Tctp*	*0.14*	*NA*	*282 [129,790]*
	Dmau_X_3.07	DNA	CG16781;CG12206	Filt,Filt;0.11†	NA	683(778) [314(357),1913(2180)]
	Dmau_X_8.42*	Ambig	His3.3B	2.83	NA	436 [200,1221]
	Dmau_X_20.09*	DNA	NA	NA	NA	535 [246,1499]
D. melanogaster	Dmel_2L_4.46	DNA	Gs1l;RpL27A	nORF;nORF	NA	45 [21,125]
	Dmel_2L_12.86	DNA	NA	NA	NA	174 [80,487]
	Dmel_2L_19.94	DNA	sick	nORF	NA	559 [256,1564]
	Dmel_2L_22.75	DNA	NA	NA	NA	697 [320,1951]
	Dmel_2R_0.09	DNA	NA	NA	NA	624 [287,1748]
	Dmel_2R_0.57	DNA	NA	NA	NA	303 [139,848]
	Dmel_2R_2.32	DNA	NA	NA	NA	315 [145,883]
	Dmel_3L_23.41	DNA	NA	NA	NA	443 [203,1241]
	Dmel_3L_24.3	DNA	NA	NA	NA	343 [158,962]
	Dmel_3R_17.04	Ambig	CR43975	nORF	NA	516 [237,1444]
	Dmel_3R_20.95	DNA	*vig2;Mocs2/CG42503;Clbn;Bili*	*0.53;0.45;0.64;Filt*	NA	463(497) [213(391),1297(1391)]
	Dmel_X_12.65	DNA	ade5;CG12717	nORF;Filt	NA	725 [333,2031]
	Dmel_X_12.66	DNA	NA	NA	NA	430 [197,1204]
D. simulans	Dsim_2L_11.91	DNA	bru1	Filt	2/11 [3]	169 [77,472]
	Dsim_2L_12.3	*DNA*	*CG5787;Pih1D1*	*Filt;nORF*	*14/17 [4.9];4/6 [1.2]*	*813 [373,2275]*
	Dsim_2L_15.71	*RNA*	*CG4455*	*nORF*	*22/22 [8.7]*	*197 [90,551]*
	Dsim_2L_19.34	DNA	NA	NA	NA	351 [161,983]
	Dsim_2R_0.06	DNA	NA	NA	NA	267 [122,747]
	Dsim_2R_4.35	*RNA*	*14–3-3zeta*	*Filt*	*0/2 [18.5]*	*90 [41,252]*
	*Dsim_2R_9.41**	*RNA*	*SRPK*	*Filt*	*25/29 [575.6]*	*204 [94,573]*
	Dsim_3L_7.55	*DNA*	*CG7492;Ank2*	*nORF;0.63,0.62*	*8/9 [11.2];32/36 [6.6]*	*383 [176,1072]*
	Dsim_3L_10.87	RNA	Sod	0.09†	1/8 [10.4]	182(477) [84(219),509(1335)]
	Dsim_3L_22.2*	DNA	NA	NA	NA	408 [187,1142]
	Dsim_3R_4.18	DNA	NA	NA	NA	161 [74,452]
	Dsim_3R_12.73	DNA	NA	NA	NA	96 [44,269]
	*Dsim_3R_13.94**	*DNA*	*Tctp*	*nORF*	*0/30 [17.4]*	*197 [90,551]*
	Dsim_X_7.6	RNA	Sdt	Filt	11/30 [2.3]	374 [172,1047]
	Dsim_X_15.22	RNA	Cyp1	0.78	0/18 [4.5]	222 [102,623]
	Dsim_X_20.2	DNA	CG17450/CG32819/CG32820	Filt	0/8 [19.5]	268 [123,750]

A total of 45 unique Y-linked transfers were detected, arising either as retrocopies (RNA), as DNA translocations (DNA), or via an undetermined mode (Ambig). Twenty-five of the Y-linked transfers harbored at least one gene—i.e., were GeTYs [donor gene names in donor(s) column]—with six GeTYs being shared between species (italicized rows). In all columns, GeTYs comprising several genes have each gene name separated by semicolons, with those having identical gene models being separated by a forward slash. Purifying selection was detected for three GeTYs (ω column; genes having more than one detected ORF being further delineated by a comma), whereas others showed evidence of degeneration (ω column; nORF = no ORF; Filt = Y-linked or donor CDS lacked either a start or stop codon, contained an inactivating mutation, or >10% Y-linked codons were missing in donor alignment; Dataset S6). Two publicly available testes-specific RNA-Seq datasets (23, 24) revealed weak evidence for Y-linked GeTY expression, with most GeTYs having low relative expression (i.e., the fraction of diagnostic exonic sites where the Y-linked allele contributed >1% of the total coverage for that site; fraction shown in expression column) and low absolute expression (i.e., mean coverage of Y-linked alleles mapping to diagnostic exonic sites; values in square brackets in expression column). Point estimates of transfer times revealed evolutionarily recent origins, with the oldest transfer arising around 1 Mya (see age column; error margins in square brackets, age estimates for putatively functional genes after correcting for purifying selection shown in standard parentheses; Dataset S7). For some Y-linked transfers, multiple haplotypes were detected in male-specific short read data suggesting that these transfers likely underwent subsequent duplication on the Y chromosome, with GeTY Dsim_2R_9.41 also showing signs of an additional autosomal/X-linked duplication. Age and expression estimates may be unreliable for these Y-linked transfers.

*Putative duplicated transfers.

†Postmultiple testing correction significance ($q < 0.05$).

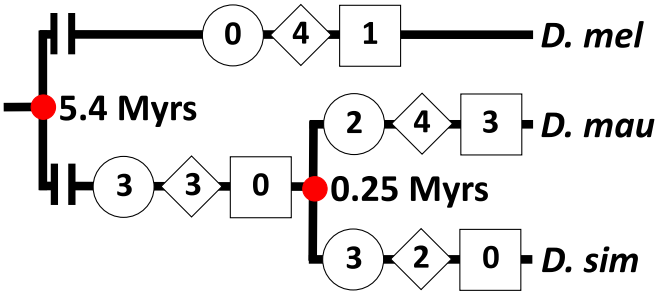

Fig. 2. Origin of Y-linked gene transfers. Retrocopies (circles), DNA translocations (diamonds), and ambiguous transfers (squares) are indicated on the inferred branch of origin in the *D. melanogaster* clade. Divergence times are shown at the red nodes. Shared GeTYs are only found in the *D. simulans* clade. The *D. simulans* clade also contains a significant excess of GeTYs relative to *D. melanogaster*, which appears to be primarily driven by a surplus of retrocopy transfers. Note that the branch lengths are not to scale; both the *D melanogaster* and *D. simulans* clade branches are truncated (depicted by the break points).

contig bearing the full translocation (9) (*SI Methods*), confirming the high quality of our Y-linked haplotype reconstructions.

Divergent Modes of Gene Transfer onto the *Drosophila* Y Chromosome. Y-linked transfers can be generated by two distinct mechanisms: either via a translocation of a genomic region (i.e., DNA translocations) or through the integration of reverse transcribed genes (i.e., retrocopies) (10, 11). Although all nongenic transfers are necessarily DNA translocations, GeTYs may arise from either mechanism. For intron-bearing genes, the distinction between the two mechanisms is straightforward: in the case of retrocopies, alignment of the Y-linked haplotypes to the parental ortholog will show evidence for splicing (should the donor gene contain exons) and will lie within donor gene boundaries. In contrast, DNA translocations will include intronic sequences, and the translocated regions need not coincide with parental gene boundaries. Based on these characteristics, 8 of the 25 GeTYs were Y-linked retrocopy insertions (5 in *D. mauritiana*, 6 in *D. simulans*, 3 shared; 0 in *D. melanogaster*) and 13 were DNA translocations (7 in *D. mauritiana*, 5 in *D. simulans*, 3 shared; 4 in *D. melanogaster*) (Table 1, Figs. S2 and S3, and Dataset S5). The transfer mechanism of the remaining four GeTYs (three in *D. mauritiana* and one *in D. melanogaster*) could not be unambiguously determined (*SI Methods*). Three lines of evidence suggest that the observed Y-linked transfers are probably fixed within each species: (*i*) the effective population size of the Y chromosome is relatively small (25% that of autosomes), (*ii*) two *D. simulans* Y-linked retrocopies analyzed in a PCR assay were fixed in a global sample of 25 males (Dataset S6 and *SI Methods*), and (*iii*) some gene transfers are shared between species.

Although the number of nongenic transfers is similar across the three species (8 in *D. melanogaster*, 7 in *D. mauritiana*, and 5 in *D. simulans*), the 5 GeTYs in *D. melanogaster* are significantly fewer than the 20 independent transfers observed in the *D. simulans* clade ($P = 0.011$, Poisson test; *Methods*). This discrepancy appears to be largely driven by a significantly elevated Y-linked retrocopy insertion rate in the *D. simulans* clade ($P = 0.004$, Poisson test), a result that is even more remarkable given the absence of evidence for Y-linked retrocopies in *Drosophila* to date. Notably, *D. melanogaster* has the most complete gene annotation of the three species in this study, implying that interspecies differences in the quality and quantity of gene annotations did not bias our Y-linked retrocopy detection. These findings suggest that Y-linked gene transfer rates, and the underlying molecular mechanisms driving the translocations, can undergo significant divergence over relatively brief evolutionary time spans in *Drosophila*.

Y-Linked Gene Transfers Are Recent but Show Limited Evidence of Purifying Selection. The general lack of shared Y-linked transfers across all three species suggests that the observed Y-linked transfers were relatively recent. We estimated the age of each transfer using a method that minimizes the influence of ancestral polymorphisms by ignoring Y-linked substitutions that are still segregating in the donor copy in large female samples (*Methods* and *SI Methods*). Although coarse, our estimates indicate that the Y-linked transfers are evolutionarily recent—with all arising within the past 1 My (Table 1 and Dataset S7)—and that the age distribution of GeTYs and nongenic transfers were not significantly different ($P = 0.58$, two-sided Kolmogorov–Smirnov test; $P = 0.74$ after GeTYs adjusted for the effects of purifying selection and putative duplicated Y-linked transfers removed; *SI Methods*). Although the latter result implies that GeTYs were effectively behaving like neutral loci, the fact that the only shared Y-linked transfers between species were GeTYs suggests they were subject to stronger Y-linked purifying selection than nongenic transfers in general. Further, many of these shared transfers had estimated ages that were younger than the recorded split between the two species, which may have resulted from purifying selection removing new mutations in the Y-linked copies. Thus, we performed two additional analyses to determine if any of the GeTYs were functional.

First, we measured ω, the ratio of nonsynonymous to synonymous substitutions that had accumulated in each Y-linked copy following the transfer, and tested whether this ratio differed from neutral expectations (i.e., ω significantly less than 1; *Methods* and *SI Methods*). To avoid a bias toward high ω values due to incorrect gene models, we performed de novo gene predictions for each GeTY and only retained instances where the predicted Y-linked ORF included the start and stop codons and produced a largely complete (i.e., >90% of the predicted codons could be aligned) and consistent (i.e., contained no frameshifts or stop codons) alignment with the donor copy from the focal species and *Drosophila yakuba*. Our results revealed that many of the GeTYs contained incomplete ORFs or inactivating mutations (Table 1 and Dataset S8), with only three being maintained by purifying selection after transferring to the Y chromosome: two GeTYs in *D. mauritiana* (*LamC* on Dmau_2R_8.73 and *CG12206* on Dmau_X_3.07) and one in *D. simulans* (*Sod* on Dsim_3L_10.87) (all $\omega \leq 0.14$ and $q \leq 0.03$; Table 1 and Dataset S9). Notably, several genes had low to moderate ω values but were not significant (Table 1)—including *FDY*, the recently discovered young Y-linked gene in *D. melanogaster* (9)—suggesting that our selection tests were probably conservative.

As a second test of GeTY functionality, we used testis-specific RNA-Seq data from *D. simulans* to quantify the expression of the GeTYs in this species (*SI Methods*). Several GeTYs showed weak evidence for low levels of expression; however, this did not include the functional GeTY identified in the ω-based tests (Fig. S4 and Dataset S10). Although limitations in our tests may have precluded detection of some functional Y-linked genes, the evidence indicates that purifying selection has played a minor role in maintaining recent gene transfers onto the *Drosophila* Y chromosome.

Discussion

High Rates of Y-Linked Gene Traffic. Our unbiased approach to detect Y-linked gene transfers has uncovered several fundamental aspects of Y chromosome evolution in *Drosophila*. We observe a high transfer rate of primary genetic material from the rest of the genome onto the Y chromosome (1.67 per My; *Methods*), which exceeds the slow accumulation of functional Y-linked genes inferred for the *Drosophila* genus over longer evolutionary times by

an order of magnitude [0.12 per My (12)]. Despite being much higher than previous estimates, the primary Y-linked gene acquisition rate inferred here appears to be up to an order of magnitude lower than for the rest of the genome for *Drosophila* (13–15), although the inclusion of different transfer categories in previous studies (e.g., de novo genes and intrachromosomal transfers) complicates direct comparisons. Further, despite a handful of GeTYs showing evidence for functionality [4/25 = 16%, including *FDY* in *D. melanogaster* (9)]—which lead to a functional Y-linked gene acquisition rate that is approximately double the previous estimate (12) (four functional GeTYs/5.4 My/three species = 0.25 new genes/species/My)—many of the GeTYs did not have complete ORFs, contained frameshift mutations, or showed no evidence of expression. This implies that the majority of the Y-linked transfers reported here have become pseudogenes and that the Y chromosome is a less hospitable genetic environment for new gene evolution than the rest of the genome in *Drosophila*.

The Dynamic and Challenging Genetic Environment of the *Drosophila* Y Chromosome. A recent study revealed that *D. melanogaster* has more Y-linked genes than *Drosophila virilis*, primarily due to the higher number of gene gains in the former since the two species last shared a common ancestor (3). This result suggests that the elevated rate of functional Y-linked gene acquisition reported here may reflect a general acceleration in this rate across the *Drosophila* subgroup relative to their sister taxa. The mechanistic driver behind this putative lineage-specific change remains unknown, but possible factors include increased accessibility of the Y chromosome to transfers (e.g., due to more relaxed chromatin conformation) or improved efficacy of Y-linked selection (e.g., due to increased effective population size for the Y chromosome), among others. Alternatively, the rate of functional GeTY acquisition reported here could be a transient phenomenon, whereby the short-term rate (over ∼1 My) eventually converges with the slower long-term rates (over ∼60 My). The efficacy of selection on weakly deleterious mutations is reduced on the *Drosophila* Y chromosome relative to other chromosomes (16), whereby many of the newly transferred genes, including those currently under selection, could become pseudogenized over longer time periods. Consistent with this idea, many of the GeTYs displaying evidence for low levels of expression in *D. simulans* were also present in *D. mauritiana*; however, none of these shared GeTYs displayed significant purifying selection in either species. Additional testing on more *Drosophila* species is ultimately required to determine the temporal stability the Y-linked gene transfer rate, although the acquisition rate of duplicated genes on the autosomes and X chromosome appears to be relatively stable over long periods in *Drosophila* (15), particularly for retrocopies (14). Regardless of the underlying cause of the temporal disparity in Y-linked gene gains reported here, when combined with the significant differences in Y-linked retrocopy traffic across the *D. melanogaster* subgroup, our study reveals that the *Drosophila* Y chromosome is an even more dynamic genetic environment than previously appreciated and is capable of undergoing significant changes over relatively short evolutionary time scales.

Conclusion

In contrast to many other species, the *Drosophila* Y chromosome is a highly dynamic genetic environment. For example, in *D. pseudoobscura* the Y chromosome is not homologous to the ancestral *Drosophila* Y chromosome but has arisen de novo (17). We have shown that Y-linked gene acquisition over the past million years is a highly dynamic feature of the *Drosophila* Y chromosome, with 10 times more gene traffic and twice the number of functional gene gains than are expected given the Y-linked gene acquisition rate recorded over the past 63 My (12). In addition to heterogeneous Y-linked gene acquisition dynamics, our method has revealed previously unknown properties of the Y chromosome, i.e., frequent retrocopy traffic onto the Y chromosome, which appears to have lineage-specific dynamics. Further research is required to determine whether this pattern reflects an actual elevation in the functional gene acquisition rate or represents the short-term evolutionary dynamics of the Y chromosome, which will eventually converge to the slower long-term rate. Similarly, we still do not know what the ancestral Y-linked retrocopy transfer rate was and how this is evolving in general across the *Drosophila* complex. These questions and many others can be empirically tested by applying the present method to the growing number of *Drosophila* species with reference genomes. Moreover, because our method can determine Y-linked gene transfers using inexpensive short reads and does not depend on a preassembled Y chromosome or associated contigs, it holds the potential reveal fundamental details of Y chromosome evolution in many other species at a hitherto unmatched level of resolution.

Methods

Y-Linked Transfer Identification and Haplotype Assembly. Using Illumina paired end reads, we sequenced males and females of two strains from all three species and used Burrows-Wheeler Aligner (BWA) (18) to map reads on reference genomes lacking Y chromosomes. SNPs with large allele frequency differences between the two sexes from the same strain were determined (see Dataset S11 for a list of all diagnostic SNPs) and then grouped into larger regions according to inter-SNP distances and gene boundaries. Haplotypes for the resulting regions were de novo assembled with trans-ABySS (Assembly by Short Sequences) (19) combining the subset of reads containing Y-linked alleles from both strains. A detailed explanation of the analytical pipeline and haplotype assembly is provided in *SI Methods*.

Estimating GeTY Age and Function. The age of each GeTY was estimated using the Y-specific nucleotide divergence from the parental ortholog scaled by the *D. melanogaster* base substitution rate (2.8×10^{-9}; ref. 20). Tests for purifying selection were performed by using AUGUSTUS (21) to predict Y-linked ORFs for each GeTY, then using codon-based phylogenetic analyses implemented in PAML (Phylogenetic Analysis by Maximum Likelihood) (22) to estimate ω for each ORF relative to the reconstructed ancestral donor sequence and using likelihood ratio tests to determine whether these ω estimates significantly differed from 1. More details on the aging and function tests are provided in *SI Methods*.

Y-Linked Gene Transfer Rate. The transfer rate was calculated as the average number of GeTYs observed across the three lineages divided by the estimated time of divergence between *D. melanogaster* and the *D. simulans* clade. To ensure phylogenetic independence in the *D. simulans* clade, we counted the transfers shared between species only once and added this number to the average of the remaining species-specific transfers in this clade [i.e., 6 + (5 + 9)/2 = 13]. Thus, the effective number of transfers, N_{eff}, is equal to N_{mel} + N_{sim_clade} = 5 + 13 = 18. To get the average number of Y-linked transfers per lineage, we divided N_{eff} by the number of distinct lineages, L, and divided this value by the divergence time, d, to, derive the average transfer rate: $(N_{eff}/L)/d$ = (18/2)/5.4 My = 1.67 novel GeTYs per lineage per My. Note that this serves as a lower bound to the true GeTY rate because transfers will be unobserved if they have degraded sufficiently to prevent read alignment or because they lack the required number of diagnostic divergent SNPs to be determined in our pipeline (i.e., ≤5 SNPs; *SI Methods*).

Lineage-Specific Rate Tests. We modeled gene transfers on the Y chromosome as a Poisson process where λ is the Y-linked gene transfer rate. Differences in the Y-linked gene transfer rate between the *D. melanogaster* and *D. simulans* lineages were tested by determining the probability of observing at most the number of *D. melanogaster*-specific transfers given the average number of transfers specific to the two species in the *D. simulans* clade. Phylogenetic independence was accounted for in the same way as for the estimation of the gene transfer rate (see above). This method was applied to all gene transfers, and DNA translocations and retrocopies separately, resulting in the following probabilities: Poisson($X \leq 5 \mid \lambda = 13$) = 0.011 for all gene transfers, Poisson($X \leq 4 \mid \lambda = 6$) = 0.285 for DNA translocations, and Poisson($X = 0 \mid \lambda = 5.5$) = 0.004 for retrocopies. Note that the latter

EVOLUTION

test remained significant when treating the ambiguous GeTY in *D. melanogaster* as a retrocopy: Poisson($X \leq 1 \mid \lambda = 5.5$) = 0.027. No significant differences were observed between *D. melanogaster* and the *D. simulans* clade for the detected functional GeTYs: Poisson($X \leq 1 \mid \lambda = 1.5$) = 0.56.

ACKNOWLEDGMENTS. We thank E. Hellmich, C. Pegueroles-Queralt, R. Sommer, B. Ballard, A. Paaby, B. Sebnem Onder, J. F. Garcia, M. Ofner, M. Puchinger, T. Little, M. Imhof, and C. Niessinger for collecting or providing the fly stocks used in this study. We are grateful to Nicola Palmieri for providing updated versions of the *D. simulans* and *D. mauritiana* reference genomes and the team at the Vienna BioCenter Core Facilities (VBCF) Next Generation Sequencing (NGS) Unit (www.vbcf.ac.at/home/) for performing part of the Illumina sequencing for this study. We thank Andy Clark for helpful discussions on data interpretation and acknowledge the insightful comments from two anonymous reviewers that led to further improvements in the manuscript. We thank the National Human Genome Research Institute (NHGRI)-funded Encyclopedia of DNA Elements Consortium for providing some of the testis data used in this study and the Brian Oliver laboratory for generating these data. This work was supported by a PhD fellowship (to R.T.) from the Vetmeduni Vienna, the Austrian Science Fund (W1225-B20), and the European Research Council grant ArchAdapt (to C.S.).

1. Bernardo Carvalho A, Koerich LB, Clark AG (2009) Origin and evolution of Y chromosomes: *Drosophila* tales. *Trends Genet* 25:270–277.
2. Carvalho AB (2002) Origin and evolution of the *Drosophila* Y chromosome. *Curr Opin Genet Dev* 12:664–668.
3. Carvalho AB, Clark AG (2013) Efficient identification of Y chromosome sequences in the human and *Drosophila* genomes. *Genome Res* 23:1894–1907.
4. Bachtrog D (2013) Y-chromosome evolution: Emerging insights into processes of Y-chromosome degeneration. *Nat Rev Genet* 14:113–124.
5. Garrigan D, et al. (2012) Genome sequencing reveals complex speciation in the *Drosophila simulans* clade. *Genome Res* 22:1499–1511.
6. Tamura K, Subramanian S, Kumar S (2004) Temporal patterns of fruit fly (*Drosophila*) evolution revealed by mutation clocks. *Mol Biol Evol* 21:36–44.
7. Hall AB, et al. (2013) Six novel Y chromosome genes in *Anopheles* mosquitoes discovered by independently sequencing males and females. *BMC Genomics* 14:273.
8. Dyer KA, White BE, Bray MJ, Piqué DG, Betancourt AJ (2011) Molecular evolution of a Y chromosome to autosome gene duplication in *Drosophila*. *Mol Biol Evol* 28: 1293–1306.
9. Carvalho AB, Vicoso B, Russo CAM, Swenor B, Clark AG (2015) Birth of a new gene on the Y chromosome of *Drosophila melanogaster*. *Proc Natl Acad Sci USA* 112: 12450–12455.
10. Long M, VanKuren NW, Chen S, Vibranovski MD (2013) New gene evolution: Little did we know. *Annu Rev Genet* 47:307–333.
11. Chen S, Krinsky BH, Long M (2013) New genes as drivers of phenotypic evolution. *Nat Rev Genet* 14:645–660.
12. Koerich LB, Wang X, Clark AG, Carvalho AB (2008) Low conservation of gene content in the *Drosophila* Y chromosome. *Nature* 456:949–951.
13. Zhou Q, et al. (2008) On the origin of new genes in *Drosophila*. *Genome Res* 18: 1446–1455.
14. Bai Y, Casola C, Feschotte C, Betrán E (2007) Comparative genomics reveals a constant rate of origination and convergent acquisition of functional retrogenes in *Drosophila*. *Genome Biol* 8:R11.
15. Zhang YE, Vibranovski MD, Krinsky BH, Long M (2010) Age-dependent chromosomal distribution of male-biased genes in *Drosophila*. *Genome Res* 20:1526–1533.
16. Singh ND, Koerich LB, Carvalho AB, Clark AG (2014) Positive and purifying selection on the *Drosophila* Y chromosome. *Mol Biol Evol* 31:2612–2623.
17. Carvalho AB, Clark AG (2005) Y chromosome of D. pseudoobscura is not homologous to the ancestral Drosophila Y. *Science* 307:108–110.
18. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754–1760.
19. Robertson G, et al. (2010) De novo assembly and analysis of RNA-seq data. *Nat Methods* 7:909–912.
20. Keightley PD, Ness RW, Halligan DL, Haddrill PR (2014) Estimation of the spontaneous mutation rate per nucleotide site in a *Drosophila melanogaster* full-sib family. *Genetics* 196:313–320.
21. Stanke M, Diekhans M, Baertsch R, Haussler D (2008) Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* 24: 637–644.
22. Yang Z (2007) PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24:1586–1591.
23. Miller SA, Dykes DD, Polesky HF (1988) A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res* 16:1215.
24. Palmieri N, Nolte V, Chen J, Schlötterer C (2014) Genome assembly and annotation of a *Drosophila simulans* strain from Madagascar. *Mol Ecol Resour* 15:372–381.
25. Brizuela BJ, Elfring L, Ballard J, Tamkun JW, Kennison JA (1994) Genetic analysis of the brahma gene of *Drosophila melanogaster* and polytene chromosome subdivisions 72AB. *Genetics* 137:803–813.
26. Nolte V, Pandey RV, Kofler R, Schlötterer C (2013) Genome-wide patterns of natural variation reveal strong selective sweeps and ongoing genomic conflict in *Drosophila mauritiana*. *Genome Res* 23:99–110.
27. Pandey RV, Schlötterer C (2013) DistMap: A toolkit for distributed short read mapping on a Hadoop cluster. *PLoS One* 8:e72614.
28. Li H, et al.; 1000 Genome Project Data Processing Subgroup (2009) The sequence alignment/map format and SAMtools. *Bioinformatics* 25:2078–2079.
29. Kofler R, Pandey RV, Schlötterer C (2011) PoPoolation2: Identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics* 27:3435–3436.
30. Agresti A (2002) Categorical Data Analysis (John Wiley & Sons, Hoboken, NJ).
31. Smit A, Hubley R, Green P (2013) RepeatMasker Open-4.0. 2013–2015. Available at repeatmasker.org. Accessed May 2, 2015.
32. Tobler R, et al. (2014) Massive habitat-specific genomic response in *D. melanogaster* populations during experimental evolution in hot and cold environments. *Mol Biol Evol* 31:364–375.
33. Orozco-terWengel P, et al. (2012) Adaptation of *Drosophila* to a novel laboratory environment reveals temporally heterogeneous trajectories of selected alleles. *Mol Ecol* 21:4931–4941.
34. Zerbino DR, Birney E (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* 18:821–829.
35. Wu TD, Nacu S (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* 26:873–881.
36. Kent WJ (2002) BLAT—The BLAST-like alignment tool. *Genome Res* 12:656–664.
37. Dennis G, Jr, et al. (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol* 4:P3.
38. Huang W, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 4:44–57.
39. Assis R, Zhou Q, Bachtrog D (2012) Sex-biased transcriptome evolution in *Drosophila*. *Genome Biol Evol* 4:1189–1200.
40. Roy S, et al.; modENCODE Consortium (2010) Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science* 330:1787–1797.
41. Katz Y, et al. (2015) Quantitative visualization of alternative exon expression from RNA-seq data. *Bioinformatics* 31:2400–2402.
42. Thorvaldsdóttir H, Robinson JT, Mesirov JP (2013) Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Brief Bioinform* 14: 178–192.
43. Innan H, Kondrashov F (2010) The evolution of gene duplications: Classifying and distinguishing between models. *Nat Rev Genet* 11:97–108.
44. Hill T, Schlötterer C, Betancourt AJ (2016) Hybrid dysgenesis in *Drosophila simulans* associated with a rapid invasion of the P-element. *PLoS Genet* 12:e1005920.
45. Kofler R, Nolte V, Schlötterer C (2015) Tempo and mode of transposable element activity in *Drosophila*. *PLoS Genet* 11:e1005406.
46. Slater GSC, Birney E (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* 6:31.
47. Loytynoja A (2014) Phylogeny-aware alignment with PRANK. *Methods Mol Biol*, 1079, pp 155–170.
48. Löytynoja A, Goldman N (2005) An algorithm for progressive multiple alignment of sequences with insertions. *Proc Natl Acad Sci USA* 102:10557–10562.
49. Storey JD (2002) A direct approach to false discovery rates. *J R Stat Soc Series B Stat Methodol* 64:479–498.
50. Rogers RL, Shao L, Sanjak JS, Andolfatto P, Thornton KR (2014) Revised annotations, sex-biased expression, and lineage-specific genes in the *Drosophila melanogaster* group. *G3 (Bethesda)* 4:2345–2351.
51. Chen ZX, et al. (2014) Comparative validation of the *D. melanogaster* modENCODE transcriptome annotation. *Genome Res* 24:1209–1223.
52. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489:57–74.