Meta-Analysis Reveals that Genes Regulated by the Y Chromosome in *Drosophila melanogaster* Are Preferentially Localized to Repressive Chromatin

Timothy B. Sackton*, and Daniel L. Hartl

Department of Organismic and Evolutionary Biology, Harvard University

***Corresponding author:** E-mail: tsackton@oeb.harvard.edu.

Accepted: January 8, 2013

Abstract

The *Drosophila* Y chromosome is a degenerated, heterochromatic chromosome with few functional genes. Despite this, natural variation on the Y chromosome in *D. melanogaster* has substantial *trans*-acting effects on the regulation of X-linked and autosomal genes. It is not clear, however, whether these genes simply represent a random subset of the genome or whether specific functional properties are associated with susceptibility to regulation by Y-linked variation. Here, we present a meta-analysis of four previously published microarray studies of Y-linked regulatory variation (YRV) in *D. melanogaster*. We show that YRV genes are far from a random subset of the genome: They are more likely to be in repressive chromatin contexts, be expressed tissue specifically, and vary in expression within and between species than non-YRV genes. Furthermore, YRV genes are more likely to be associated with the nuclear lamina than non-YRV genes and are generally more likely to be close to each other in the nucleus (although not along chromosomes). Taken together, these results suggest that variation on the Y chromosome plays a role in modifying how the genome is distributed across chromatin compartments, either via changes in the distribution of DNA-binding proteins or via changes in the spatial arrangement of the genome in the nucleus.

Key words: gene expression, heterochromatin, evolution.

Introduction

The *Drosophila* Y chromosome, despite comprising approximately 20% of the male genome in *Drosophila melanogaster*, contains fewer than 20 genes, primarily with specialized male reproductive functions (Gatti and Pimpinelli 1983; Bonaccorsi and Lohe 1991; Carvalho et al. 2000, 2001, 2003; Koerich et al. 2008; Vibranovski et al. 2008; Krsticevic et al. 2010). Most of the chromosome consists of megabase-sized blocks of repetitive DNA, including sequences derived from transposable elements, large microsatellite blocks, and the Y-linked ribosomal DNA (rDNA) array, *bobbed* (Gatti and Pimpinelli 1983; Bonaccorsi and Lohe 1991; Lohe et al. 1993). Although it has long been known that the Y chromosome is essential for male fertility (Brosseau 1960), until recently the *Drosophila* Y chromosome was not thought to have any other significant functions or harbor significant variation among or across populations. In the past decade, however, evidence has emerged that genetic variation on the Y chromosome is associated with variation in a number of traits, including overall male fitness (Chippindale and Rice 2001), sensitivity of spermatogenesis to thermal stress (Rohmer et al. 2004; David et al. 2005), geotaxis (Stoltenberg and Hirsch 1997), and position effect variegation (PEV; Lemos et al. 2010). Despite these observations, a mechanistic basis for widespread phenotypic effects of Y-linked variation has remained elusive.

One potential mechanism is the phenomenon of Y-linked regulatory variation (YRV). Variation on the Y chromosome in both *D. melanogaster* and *D. simulans* is associated with variation in expression of autosomal and X-linked genes (Lemos et al. 2008; 2010; Jiang et al. 2010; Sackton et al. 2011). By introgressing Y chromosomes from a variety of *D. melanogaster* stocks into a common laboratory background, Lemos et al. (2008) demonstrated that hundreds of genes vary in expression in males across lines that differ only in the population of origin of their Y chromosome, whereas no genes vary in expression across females of the same lines. Subsequently, it has been shown that expression of Y-linked protein-coding genes plays at most a minor role in YRV: Because sex

determination in *Drosophila* is based on the number of X chromosomes, individuals with an XX/Y genotype are female, and can be constructed using standard *D. melanogaster* stocks. These females do not transcribe genes from their Y chromosome but still show *trans*-acting effects of Y-linked variation on gene expression (Lemos et al. 2010). YRV is also subject to significant Y-by-background epistatic effects (Jiang et al. 2010) and at least partially attributable to variation in rDNA content on the Y, as YRV is observed among mutant Y chromosomes that vary only in rDNA content (Paredes et al. 2011). Furthermore, Y chromosome divergence between *D. simulans* and *D. sechellia* is also associated with gene expression changes (Sackton et al. 2011).

The phenomenon of YRV implies that the Y chromosome interacts with the rest of the genome in previously unanticipated ways to modify gene expression patterns. However, theoretical predictions and empirical studies suggest that genetic variation on the Y chromosome should be low relative to the autosomes and the X chromosome (Clark 1987, 1990; Clark and Lyckegaard 1990; Bachtrog and Charlesworth 2002; Bachtrog 2005, 2006; Kaiser and Charlesworth 2010), and recent sequencing results in *D. melanogaster* and other species have confirmed this expectation (Zurovcova and Eanes 1999; Kopp, Frank, and Barmina 2006; Kopp, Frank, and Fu 2006; Larracuente and Clark 2012). Some evidence hints that this result may be limited to single-nucleotide polymorphisms, however, and that structural variation may be more prevalent. Multiple cytologically distinguishable forms of the Y chromosome segregate in at least some species of *Drosophila* (Dobzhansky 1935), and variation in rDNA array size and other repetitive sequence blocks exists (Lyckegaard and Clark 1989; Clark and Lyckegaard 1990). Structural variation in the size of the Y chromosome, not single-nucleotide polymorphism, is associated with variation in PEV in strains of *D. melanogaster* with varying amounts of the Y chromosome fused to the X chromosome (Dimitri and Pisano 1989). A reasonable hypothesis, therefore, is that variation across Y chromosomes in the type, amount, and distribution of repetitive DNA has *trans*-acting effects on gene expression in the genome.

However, clear evidence for this hypothesis is difficult to obtain. It is still unclear what varies across Y chromosomes and how exactly that variation mechanistically affects non-Y-linked gene expression. Although characterizing variation, and especially structural variation, on the Y chromosome remains quite challenging, we can gain insight into the basis for YRV by examining the properties of the set of genes that appear to be regulated by Y-linked variation. We have observed YRV in a range of conditions, but both whether a common set of genes regulated by Y chromosome variation across genetic backgrounds exists and the extent to which genes regulated by Y chromosome variation share common sets of genomic correlates (which might predict something about the mechanistic basis for this phenomenon) remain unclear.

To begin to address these questions, we have taken a meta-analytic approach to combine data from a series of published microarray studies (table 1). We estimated robust effect sizes and combined probabilities of Y regulation across studies, which reveal patterns not apparent from the analysis of individual data sets. From this analysis, we first address the extent to which different studies reveal a common set of underlying YRV genes and then address whether there are underlying genomic properties that predict membership in the YRV gene class. We show that, indeed, there is a common class of genes that vary in expression consistently across multiple Y introgression experiments. These genes are more likely than non-YRV genes to be tissue biased in expression, localized to repressive chromatin, and vary in expression within and among species. Taken together, these results provide evidence for the hypothesis that differences among Y chromosomes modify the distribution of genes in active and repressive chromatin across the rest of the genome.

Materials and Methods

Defining a YRV Gene Set

To identify a common set of YRV genes across studies, we first selected experiments to study from the six published surveys of YRV (Lemos et al. 2008, 2010; Jiang et al. 2010; Paredes et al. 2011; Sackton et al. 2011; Zhou et al. 2012). From these surveys, we selected the four sets of experiments where at least three Y chromosomes were compared on a common genetic background in *D. melanogaster* (table 1).

For each experiment, we started with raw microarray data available at the NCBI Gene Expression Omnibus (GEO) database and then processed each set identically using limma in Bioconductor (Smyth 2005). We first background-corrected arrays using the "normexp" method (Ritchie et al. 2007) and then normalized with the "loess" method in limma (Smyth 2004). After normalization, we filtered data first by removing

Table 1
Studies Included in Meta-Analysis, with GEO Information and Other Characteristics

Study Name	GEO	Description	Reference
BL08	GSE9457	Original study reporting YRV: compared five geographically disparate Y chromosomes	Lemos et al. (2008)
BL10	GSE23612	YRV in XXY females	Lemos et al. (2010)
SP11	GSE27695	YRV in rDNA deletion lines	Paredes et al. (2011)
JZ12	GSE37068	YRV in mutation accumulation lines (Harwich)	Zhou et al. (2012)

probes that did not have high-quality data in at least 25% of arrays for a given study and second by fitting array weights using the ArrayWeights function in limma (Ritchie et al. 2006) to downweigh lower quality arrays. In most cases, all replication is biological, so we generate fits using the lmfit function in limma; for the JZ12 (GSE37068) data set, technical replicates were fit using the duplicateCorrelation function in limma.

For each normalized expression set, we fit a linear model in limma with a design matrix calculated using the modelMatrix function in limma and including a dye term and then extracted the fit coefficients for all possible pairwise contrasts among Y chromosomes. For each contrast, we calculated Cohen's effect size, *d*, as

$$(2 \times T)/\mathrm{sqrt}(\mathrm{df}), \quad (1)$$

using the degrees of freedom and moderated *T* statistic calculated by the eBayes function in limma. Within a study, all pairwise *d* statistics were averaged to generate an average pairwise *d* for each study. This value is equivalent to the expected difference, in units of standard deviations, between two Y chromosomes drawn at random from the pool of Y chromosomes included in a particular study. We calculated this average *d* statistic for the four *D. melanogaster* studies in table 1 and then averaged the average *d* statistics among the four studies to estimate an overall effect size for each gene. Complete R code for the normalization and analysis steps in limma is available from the authors upon request.

In addition to calculating Cohen's *d*, we also calculated a combined *P* value using Stouffer's method, in which *P* values are first transformed into *Z* values before combining (Stouffer et al. 1949). For each gene and each study, we tested the null hypothesis of equal expression across all Y introgression lines using the *F* statistic calculated by the limma functions lmFit and eBayes. We then combined *P* values for each gene across studies using the R function:

$$\mathrm{pnorm}(\mathrm{sum}(\mathrm{qnorm}(x))/\mathrm{sqrt}(\mathrm{length}(x))), \quad (2)$$

where x represents the vector of *P* values, pnorm is the normal distribution function, and qnorm is the normal quantile function. After combining *P* values, we applied a standard false discovery rate (FDR) multiple test correction in R using the p.adjust function (Benjamini and Hochberg 1995).

We used these combined *P* values to generate our YRV gene set. We are interested in two kinds of YRV genes: those that are common across two or more studies and those that are specific to a single study. Because a combined *P* value can give a significant result either because a test is highly significant in one study but nonsignificant in the remaining studies or because a test is moderately significant in multiple studies, the combined *P* value alone cannot distinguish these. To separate these two classes, we calculated leave-one-out combined *P* values, where we simply drop one of the studies before calculation. We define "specific YRV genes" as those where: 1) the combined *P* value that includes a given study is significant ($q \leq 0.05$), but the combined *P* values that exclude that study are not significant or 2) the individual multiple-test-corrected *P* value for a given study is significant but the combined *P* value is not significant. We define "common YRV genes" as all the remaining genes with a significant combined *P* value (fig. 1).

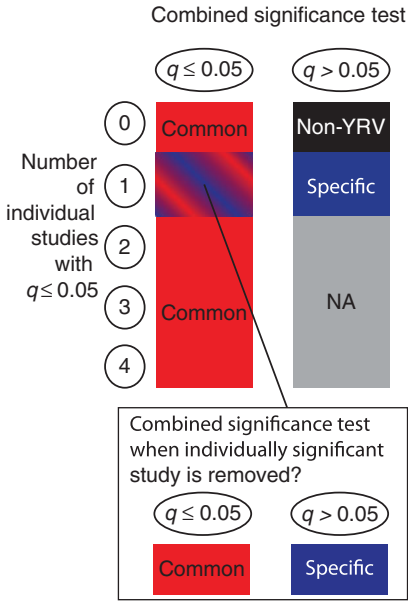

Fig. 1.—Procedure for defining common and specific classes of YRV genes, based on the combined *q* value across studies plus the *q* values of individual studies. No genes with $q \leq 0.05$ in more than one study fail to achieve significance in the combined statistic.

Data Sources for Genomic Correlates

Much of our analysis focuses on the analysis of a wide range of potential correlates to YRV, including components of gene structure, chromatin environment, gene expression and function, and gene evolutionary patterns. The variables included in our analysis, and their sources, are listed in table 2. The full data set, including all covariates and all the calculated meta-analysis statistics from each study included, is provided as supplementary file S1, Supplementary Material online.

Logistic Regression

The fundamental logic of our approach is to use a logistic regression to ask which properties of genes are best at predicting membership in the YRV class, as defined earlier. We used model selection techniques to find the best model among the large number of possible models that include one or more of the terms in table 2, as implemented in the R package glmulti (Calcagno 2012). The basic approach is to fit a series of main effect models using a logistic regression (glm, family = "binomial" in R) and find the best set of models based on an information criteria, here the Akaike Information

Table 2
Variables Used for Model Selection

Variable Name	Description	Source	Reference	Final Model?
expdivtot.m	Expression divergence across seven *Drosophila* species, in males	Custom Nimblegen arrays produced and hybridized by Brian Oliver's laboratory	Larracuente et al. (2008)	Yes
h2m	Mutational variance in expression in *D. melanogaster*	Rifkin et al. (2005)	Rifkin et al. (2005)	Yes
postmei.ratio	Ratio of postmeiotic to max(meiotic, mitotic) expression in spermatogenesis	Vibranovski et al. (2009)	Vibranovski et al. (2009)	Yes
mf.pc1	Principle component representing male/female expression bias	Calculated from modENCODE and SEBIDA data	This article	Yes
tau.avg	Measure of tissue specificity (tau) averaged between adult and larval values	FlyAtlas	This article	Yes
FbtrLen	Transcript length	Calculated from Flybase GFF files	Larracuente et al. (2008)	Yes
het	Binary variable coding presence in repressive chromatin (black, green, or blue states)	Calculated from the data of Filion et al. (2010)	Filion et al. (2010)	Yes
NumInt	Number of introns	Calculated from FlyBase GFF files	Larracuente et al. (2008)	Yes
AveIntLen	Average intron length	Calculated from FlyBase data	Larracuente et al. (2008)	No
window11.length	Length of a window of five upstream and five downstream genes surrounding focal gene	Calculated from FlyBase GFF files	This article	No
FirstIntLen	Length of first intron	Calculated from FlyBase GFF files	Larracuente et al. (2008)	No
exp.pc1	Principle component representing overall, sex-averaged, whole fly transcription level	Calculated from modENCODE male and female expression, FlyAtlas whole fly expression, and codon bias	This article	No
dmsp	Binary variable coding presence in the *D. melanogaster* sperm proteome	Proteomic studies conducted by Tim Karr's laboratory	Wasbrough et al. (2010)	No
transcript.num	Total number of different transcripts for each gene	Calculated from FlyBase GFF files	This article	No
m0w	Omega calculated from PAML model M0	Twelve genomes' annotation	Larracuente et al. (2008)	No
intergenic.dist	Average distance to nearest gene (upstream or downstream)	Calculated from FlyBase GFF files	This article	No
RecomRP	Recombination rate calculated by the RP method	Calculated from FlyBase data	Larracuente et al. (2008)	No
gpmb	Genes per megabase	Calculated from FlyBase GFF files	This article	No
cellular_loc	A simplified gene ontology cellular location	Calculated from FlyBase ontology files	This article	No

Criterion (AIC). Because the number of possible models is extremely large for the number of parameters we examine, we used a genetic algorithm to search the space of possible models, implemented in glmulti using the default parameters. The size of the search space also limits our ability to test interactions. To increase our confidence in the output of the genetic algorithm, we ran the entire procedure twice and combined the results into a single consensus output using the consensus function in glmulti. Because the genetic algorithm is not an exhaustive search procedure, this has the effect of increasing the search space and thus increasing our confidence in the results, but the two runs produce similar results when considered individually. From this output, we selected the best model based on the importance of each term, defined as the proportion of the 200 best models in which each given term appears. Full R code is available from the authors upon request.

Cross-Validation

We used a leave-one-out cross-validation approach implemented in the R function cv.glm from R package boot (Canty and Ripley 2012) to validate our model. We define a cost function as:

$$\text{mean}(\text{abs}(\text{observed-predicted}) > 0.5), \quad (3)$$

which is equivalent to an estimate of inaccuracy or the proportion of times the model misclassifies the data point left out from the leave-one-out cross-validation. Because the cross-validation approaches cannot handle data with missing values, we ran this analysis on only a subset of the full data set that excludes missing values; the results of running our full model on this data set are qualitatively identical to the results from running the full model on the original data set.

Clustering Analysis

We used two approaches for clustering analysis. To analyze clustering in the genome, we first divided each chromosome into windows of either 100 kb or 500 kb. For each window, we counted the number of YRV genes in the window and then computed an empirical null distribution by permuting the assignment of YRV genes 10,000 times and, for each permutation, counting the number of shuffled YRV genes in each window. This permutation approach controls for variation in gene density across the genome. To test for clustering specifically, we asked whether the number of windows with a nominally significant (at $\alpha = 0.05$) excess of YRV genes compared with the permutation null distribution is significantly greater than expected by chance. To test for clustering in nuclear space, we used the Hi-C data set from Sexton et al. (2012). This data set is based on an experiment in which DNA in the nucleus was cross-linked and then fragmented and sequenced, such that regions of the genome in close physical proximity are likely to be sequenced in the same fragment. These contacts can then be empirically scored between sliding windows across the genome. Because this contact count is heavily dependent on both chromosomal location and the chromatin context of the interacting pair, Sexton et al. (2012) developed a hierarchical model that corrects for these effects. To isolate the effects of YRV genes above and beyond these factors, we analyze contact counts normalized to the model expectation, rather than raw contact counts.

Results

Defining a Common Set of Genes Regulated by Y-Linked Variation

Over the past 5 years, our laboratory has studied the role of variation on the Y chromosome on gene expression across a variety of contexts and experimental designs. To better understand the commonalities across study designs in the set of genes regulated by Y-linked variation (YRV genes), we used a meta-analysis approach. We focus on two related statistical approaches: effect size as measured by Cohen's *d*, and a combined *P* value based on *Z* scores (Stouffer's method). Effect size allows a comparison of the magnitude of an effect, in standardized units, across many studies; in this case, we calculate an effect size (*d*) that is equivalent to the expression difference in units of standard deviations of a pair of Y chromosomes drawn at random, averaged across all studies. A combined *P* value provides a statistically rigorous approach to use evidence from all studies to test an underlying common null hypothesis; here, a significant combined *P* value indicates statistical support for rejecting the null hypothesis that expression of the gene in question does not vary across Y introgression lines. We focus our analysis on the results from a meta-analysis of the four studies in table 1 and calculate both an effect size and a combined *P* value for all genes.

On the basis of our combined *P* value approach (fig. 1), we identify a total of 678 genes that are susceptible to YRV, which we term YRV genes. Of these, 458 are "common," meaning that evidence for a role of Y-linked variation in their regulation comes from more than one study, and 220 are "specific," meaning that one and only one study supports a role for Y-linked variation in their regulation. Although the "common" set includes a handful of genes with highly significant evidence for YRV in all studies, in most cases these genes are not individually significant after multiple test correction in any studies; rather, the consistency of a trend across all studies provides power for the meta-analysis to identify a role for the Y chromosome (fig. 2). Nonetheless, we believe that these genes are robustly identified by our meta-analysis. Both the "specific" and "common" classes have significantly elevated effect sizes relative to the non-YRV class (fig. 3*A*; median *d* is 0.646 for "common," 0.612 for "specific," and 0.391 for "none," Mann–Whitney *U*, *P* value $< 2 \times 10^{-16}$ for both comparisons). In the case of the "specific" genes, this is

typically driven by a very high effect size in the study where the gene is individually significant, as demonstrated by the high coefficient of variation of Cohen's *d* across studies for the "specific" class (median CV for "specific" = 0.253, compared with 0.0936 for "common" and 0.0682 for "none"; fig. 3*B*).

Although it is possible that some proportion of the genes in the "specific" class could represent genetic background effects, three of the four studies in our analysis use the same background stock. If genetic background effects were a major driver of the "specific" class, we would expect the single study that uses a different genetic background (Lemos et al. 2010) to include a disproportionate number of the "specific" genes, which does not appear to be the case: 22.7% of the genes in the "specific" class are significant in the single study with a different genetic background, which is not significantly different from the 25% expected by chance ($\chi^2 = 0.6061$, df = 1, *P* value = 0.4363). However, we do find a significant excess of "specific" genes in the study by Paredes et al. (2011) (40.9% vs. 25% expected; $\chi^2 = 7.27$, df = 1, *P* value = 0.007), which examined Y chromosomes carrying severe rDNA mutations that might be expected to have disproportionate effects on gene expression. Thus, we suspect that the "specific" effects at least in part represent the fact that the Y chromosomes targeted in each study have quite different properties and thus may contain specific variation that is not observed across all studies; it may especially be the case that the severe mutations screened in the study by Paredes et al. (2011) result in particularly severe distortions of gene expression.

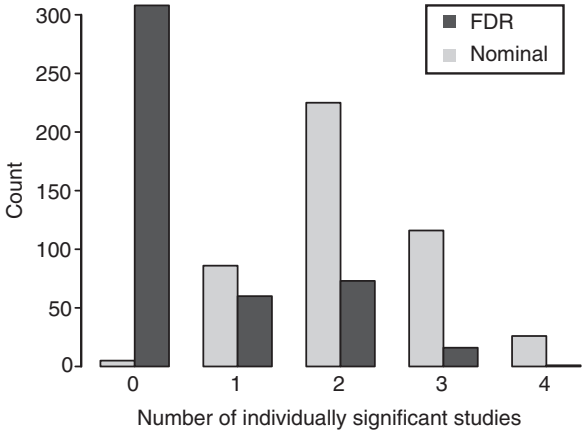

Fig. 2.—The proportion of genes in the common class that are individually significant in 0, 1, 2, 3, or 4 of the underlying studies. Light bars are for a nominal (not multiple-test corrected) *P* value ≤ 0.05. Dark gray bars are for a FDR-corrected *q* value ≤ 0.05.

Taken as a whole, these results strongly suggest that YRV is a phenomenon with significant reach. Even assuming that only the "common" class represents robust YRV genes and that this study has uncovered all extant YRV (i.e., no false negatives), we have shown that variation on the Y chromosome affects more than 3% of the protein-coding genes in the *D. melanogaster* genome. However, our analysis only considers the 4,271 genes where we had high-quality expression information across all studies. Thus, we find evidence for a role of YRV in gene expression variation for 15.9% of the genes tested, which, if extrapolated to the entire genome, would suggest that expression of as many as 2,000 genes could be affected by variation on the Y chromosome, either directly or indirectly.

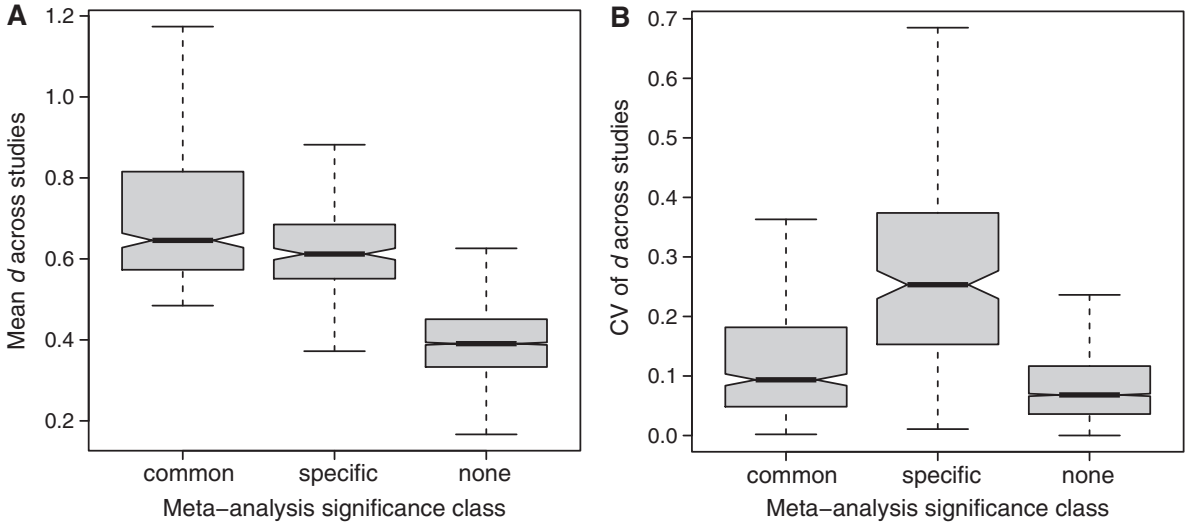

Fig. 3.—(*A*) Boxplot showing the average Cohen's *d* statistic across studies for each significance class, where Cohen's *d* represents the standardized effect size of YRV. Common vs. none, *P* value $< 2 \times 10^{-16}$ (Mann–Whitney *U*); specific vs. none, *P* value $< 2 \times 10^{-16}$ (Mann–Whitney *U*). (*B*) Boxplot of the coefficient of variation (CV) of Cohen's *d* across studies for each significance class. Specific vs. common, *P* value $< 2 \times 10^{-16}$ (Mann–Whitney *U*); specific vs. none, *P* value $< 2 \times 10^{-16}$ (Mann–Whitney *U*); and common vs. none, *P* value $= 8.31 \times 10^{-14}$ (Mann–Whitney *U*).

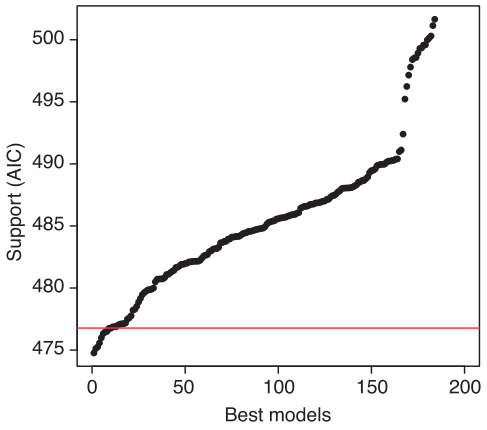

Fig. 4.—Ranked AIC support for the 200 best models. Red line is placed at the AIC of the 10th best model.

Genomic Correlates of YRV Susceptibility

To better understand the basis for susceptibility to YRV, we built a logistic regression model to identify the properties of genes that are predictive of being in the YRV class. We began with a list of 19 variables (table 2). These include parameters representing gene expression, evolutionary history, gene structure, local genomic environment, and chromatin state. To select the best model, we used a genetic algorithm search approach implemented in the R package glmulti (Calcagno 2012), which uses a genetic algorithm to sample a very large number of first-order models (where the terms included in the model are a subset of the full model) and find the one that minimizes the AIC, a measure of the relative goodness of fit of the model. The AIC of the best 200 models is shown in figure 4. Because the best set of models are relatively close in AIC, we select parameters for the final model based on the proportion of times each parameter is present in the best 200 models, rather than the absolute best model (fig. 5). The final model, then, includes all parameters with a model-averaged importance of at least 80%, meaning that those terms appear in 80% or more of the 200 best models, and is:

$$\text{YRV} \sim 1 + \text{het} + \text{FbtrLen} + \text{NumInt} + \text{tau.avg} + \text{mf.pc1} + \text{postmei.ratio} + \text{h2m} + \text{expdivtot.m} \quad \text{(Model 1)}$$

Every one of these model terms is also included in the single best model by AIC, which is:

$$\text{YRV} \sim 1 + \text{het} + \text{FbtrLen} + \text{NumInt} + \text{FirstIntLen} + \text{tau.avg} + \text{mf.pc1} + \text{postmei.ratio} + \text{h2m} + \text{expdivtot.m} \quad \text{(Model 2)}$$

As a further test of the validity of this model, we performed leave-one-out cross-validation using the cv.glm function in the

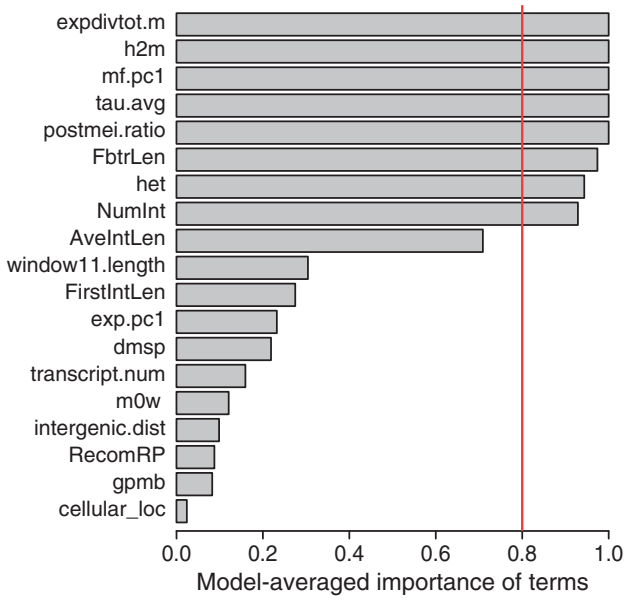

Fig. 5.—Model-averaged importance of each term in the model (table 2), which is defined as the proportion of the 200 best models in which a given term appears. Red line indicates 80% support. Terms with an importance above the red line are included in our final model.

R package boot (Canty and Ripley 2012). Leave-one-out cross-validation works by leaving out each data point in turn and predicting YRV membership of the dropped data point using the remaining data. From this procedure, we can calculate a prediction error term defined as the proportion of data points for which we fail to correctly predict the observed data. For the model generated from our model selection procedure (Model 1), the prediction error is 0.1264; for the best model by AIC (Model 2), the prediction error is 0.1335.

The best predictors of membership in the YRV class are measures of the evolutionary rate of change in gene expression (expression divergence and mutational variance of expression), gene expression distribution across tissues and sexes (tau, *M/F* expression ratio, and ratio of postmeiotic to other expression in spermatogenesis), gene size (transcript length, number of introns, and average intron length), and chromatin state (fig. 5). Notably, rate of protein evolution (as measured by d_N/d_S) and overall level of gene expression are not important predictors of membership in the YRV class (fig. 5). To estimate the magnitude and direction of the effect of these predictors on the probability of membership in the YRV class (β parameters), we reran the logistic regression including only those terms deemed important from the model selection procedure (table 2, last column). These results suggest that the typical YRV gene is one that is short, with few introns, in repressive chromatin, tissue specific, rapidly changing in expression both at the population and evolutionary level, and expressed postmeiotically during spermatogenesis (table 3).

Table 3
Model Parameters from the Final Model

Term	β Coefficient
expdivtot.m	0.50137
h2m	0.90826
postmei.ratio	2.28173
mf.pc1	−0.62495
tau.avg	2.92341
FbtrLen	−0.96739
het	0.58295
NumInt	−0.13822

YRV Genes Have Many Characteristics of Repressive Chromatin

One possible mechanistic model for how variation on the Y chromosome regulates non-Y-linked gene expression is via effects on chromatin state induced by changes in the distribution of DNA-binding proteins across the genome. To the extent that Y-linked sequences bind proteins associated with establishing chromatin states, variation in propensity of binding on the Y could impact the distribution of chromatin states across the genome. Thus, we were particularly interested in the possibility that YRV genes are nonrandomly distributed with respect to chromatin state across the genome. To further investigate this possibility, we analyzed chromatin classes based on protein-binding profiles generated by Filion et al. (2010), who defined five chromatin states: GREEN (pericentric heterochromatin), BLUE (Polycomb heterochromatin), BLACK (intercalary heterochromatin), RED (active chromatin), and YELLOW (active chromatin). Using these classes, we compared effect sizes across chromatin states (fig. 6). Genes in the BLUE and BLACK classes, corresponding to Polycomb and intercalary heterochromatin, respectively, have significantly higher average effect sizes than genes in the GREEN (pericentric heterochromatin) class, or either of the active classes (YELLOW and RED) (fig. 6). The BLUE and BLACK classes, in particular, share a high affinity for binding of the Suppressor of Under-Replication (SuUR), Lam, and D1 proteins, suggesting the possibility that these proteins may play a role in YRV.

If this hypothesis is correct, we would expect that YRV genes should be associated with regions of the genome that bind these proteins in independent studies. In *Drosophila*, B-type lamin (Lam) is a primary constituent of the nuclear lamina, the protein network that lines the inner surface of the nuclear envelope; lamins directly interact with chromatin and play a role in transcriptional regulation (reviewed in Marshall 2002). SuUR is associated with late-replicating regions of the genome and may play a role in transcriptional regulation as well (reviewed in Schwaiger and Schübeler 2006). We thus took advantage of two additional data sets that independently addressed binding to the B-type lamin

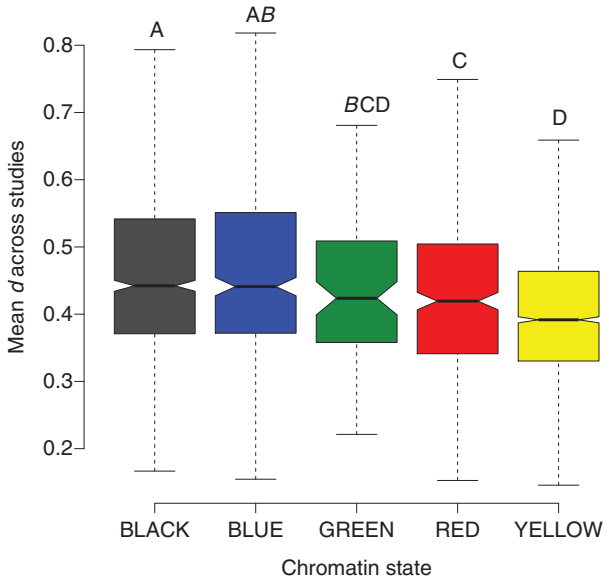

Fig. 6.—Boxplot of Cohen's *d* averaged across studies for each of the five chromatin states defined by Filion et al. (2010). Black, blue, and green are repressive chromatin; red and yellow are active chromatin. Boxplots with the same letter above them are not significantly different ($\alpha = 0.05$; italics indicate difference at $0.05 < \alpha < 0.10$), based on Tukey HSD *P* values, which indicate that: BLACK is greater than GREEN ($P = 0.005$), RED ($P = 0.000016$), and YELLOW ($P < 0.000001$); BLUE is greater than GREEN (marginally so, $P = 0.07$), RED ($P = 0.024$), and YELLOW ($P < 0.000001$); GREEN is not significantly different from RED or YELLOW; RED is greater than YELLOW ($P = 0.0013$).

(Pickersgill et al. 2006) and replication timing across the genome (Schwaiger et al. 2009). Because our logistic regression suggests that YRV genes are preferentially localized to repressive chromatin, we would expect that YRV genes should be both over-represented in the set of genes bound to the nuclear lamina and also over-represented among late-replicating regions of the genome.

We find support for both of these hypotheses. Genes in the YRV class have a significantly higher ratio of lamin bound to lamin unbound signal (YRV = 0.2425, non-YRV = −0.0821, both expressed as $\log_2$ [lamin bound/lamin unbound]; Mann–Whitney *U*, *P* value $< 2 \times 10^{-16}$). Although much of this effect is driven by the bias toward repressive chromatin states for YRV genes, it is notable that YRV genes in active chromatin (RED or YELLOW states) have a significantly higher lamin bound/lamin unbound signal than non-YRV genes in active chromatin, suggesting that even in active chromatin states YRV genes have some repressive-chromatin-like properties (in active chromatin: YRV = −0.07855, non-YRV = −0.226; Mann–Whitney *U*, *P* value = 0.0002; fig. 7). YRV genes are also much more likely to be late-replicating than non-YRV genes (Fisher's exact test $P = 1.87 \times 10^{-7}$, odds ratio = 1.612). Although much of this simply reflects the strong overlap between regions of the genome that are late

replicating and regions of the genome that are in repressive chromatin, it is again notable that, in active regions, YRV genes are later replicating than non-YRV genes (table 4), suggesting that YRV genes in active chromatin have a bias toward showing properties associated with repressive chromatin. However, these patterns are relatively weak compared with the overall bias toward excess YRV genes in late-replicating regions of the genome.

YRV Genes Are Closer than Expected in the Nucleus but Are Not Clustered along Chromosomes

Given the hypothesis that YRV genes are regulated via effects on chromatin state, it is plausible that they might be physically clustered along the chromosome, as chromatin domains are often larger than single genes (e.g., Filion et al. 2010). Although previous studies of individual data sets have occasionally showed relatively modest evidence for clustering along chromosome (e.g., Jiang et al. 2010; Zhou et al. 2012), other individual studies have failed to find this pattern (Paredes et al. 2011), and our reanalysis does not show compelling evidence for clustering along chromosomes.

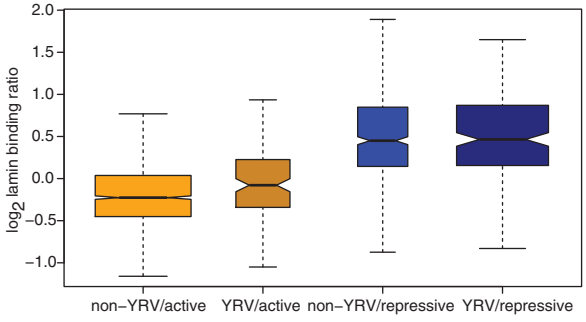

Fig. 7.—Lamin-binding ratio on a $\log_2$ scale for YRV and non-YRV genes sorted into active and repressive chromatin domains. Box widths are proportional to the share of each YRV class that is in active vs. repressive chromatin: Approximately two-thirds of YRV genes are in repressive chromatin, whereas the opposite is true of non-YRV genes. Although within repressive chromatin, YRV and non-YRV genes are not significantly different (Mann–Whitney *U*, $P = 0.9689$), in active chromatin, YRV genes have significantly higher lamin-binding ratios (Mann–Whitney *U*, $P = 0.0002$).

We calculated, based on 1,000 random permutations of the distribution of YRV genes in the genome, the probability that windows of either 100 kb or 500 kb have a significant excess of YRV genes. For 500-kb windows, we do not find an excess of nominally significant windows beyond that expected by chance, and for 100-kb windows, we actually find a significant deficit of windows with excess YRV genes (500-kb windows: 8/244 significant tests, $\chi^2 = 1.522$, df = 1, *P* value = 0.217; 100-kb windows: 26/1,208 significant tests, $\chi^2 = 20.6232$, df = 1, *P* value = 5.59×10^{-16}).

Recent technological innovations (Hi-C) have made it possible to measure the physical proximity in the nucleus of DNA segments genome wide (Sexton et al. 2012). Because the physical conformation of DNA in the nucleus appears to have large impacts on the local accessibility of DNA sequence, we used this Hi-C data to ask whether YRV genes are closer to each other than non-YRV genes are to each other in nuclear space. To do this, we classified each pair of contacts identified by Sexton et al. (2012) as being between two YRV genes, being between a YRV gene and a non-YRV gene, or not involving YRV genes. Because physical, three-dimensional proximity estimated from Hi-C approaches is predicted by both proximity along a chromosome (adjacent loci on a chromosome are likely to be adjacent in the nucleus), and on whether a locus lies in a more tightly packed repressive domain or a less tightly packed active domain (Sexton et al. 2012), we normalized the observed contact counts for each pair to a hierarchical domain model (Sexton et al. 2012), which takes into account both linear sequence distance and different trends within active and repressive domains (Sexton et al. 2012). This normalized distance then represents the relative excess or deficit of contacts observed compared with the model expectation and controls for both proximity along the chromosome and broad-scale chromatin context. Sexton et al. (2012) calculate contacts by dividing the genome into bins ranging in size from 20 to 160 kb and then counting observed contacts by mapping reads to each bin. We focus on the 80 kb bin size for simplicity, but substantially similar results are obtained for all bin sizes.

The fraction of YRV/YRV pairs with an excess of observed contacts (i.e., the number of observed contacts is higher

Table 4
Counts of YRV and Non-YRV Genes that Are in Early- or Late-Replicating Regions of the Genome, by Chromatin State

	Active Chromatin		Repressive Chromatin	
	Non-YRV	YRV	Non-YRV	YRV
Early replicating (geneRT > 0)	1,996	204	521	207
Late replicating (geneRT < 0)	269	40	623	196
	Fisher's exact test (active)	$P = 0.0507$ Odds ratio = 1.45	Fisher's exact test (repressive)	$P = 0.0485$ Odds ratio = 0.792

than the model prediction) is significantly higher than the fraction of non-YRV/non-YRV pairs with an excess of observed contacts (47.3% vs. 42.1%, $\chi^2 = 1961.175$, df = 1, P value $< 2 \times 10^{-16}$). Average normalized contacts between YRV genes are moderately but significantly higher than between non-YRV genes, indicating more observed pairs than expected under the model (median normalized YRV/YRV contacts, on a $\log_2$ scale $= -0.05$, median normalized non-YRV/non-YRV contacts on a $\log_2$ scale $= -0.16$, Mann–Whitney U, P value $< 2 \times 10^{-16}$). Furthermore, YRV genes are on average closer to other YRV genes than to non-YRV genes (median normalized YRV/YRV contacts, on a $\log_2$ scale $= -0.05$, median normalized YRV/non-YRV contacts, on a $\log_2$ scale $= -0.121$, Mann–Whitney U, P value $< 2 \times 10^{-16}$). Together, these results imply that YRV genes are in closer proximity in the nucleus than non-YRV genes on average, over and above the tighter packing predicted by the tendency of YRV genes to fall into repressive chromatin.

Discussion

Despite growing evidence that the Y chromosome plays a significant role in regulating gene expression across many genes, and the implication that this phenomenon may underlie the role of variation on the Y chromosome in phenotypic variation for traits such as thermal tolerance of spermatogenesis (David et al. 2005), male fitness (Chippindale and Rice 2001), male fecundity (Sackton et al. 2011), and geotaxis (Stoltenberg and Hirsch 1997), we still have little understanding of what kinds of genes are susceptible to YRV. Do YRV genes share common properties that implicate a shared mechanistic basis, or are they idiosyncratic? Can the properties of the genes regulated by variation on the Y shed any light on the mechanistic basis for this phenomenon?

In this study, we use meta-analysis approaches to bring together the previous work done on YRV with new genomic data sets available as a result of the modENCODE project and other large-scale genomic screens. These analyses reveal a core set of common properties that distinguish YRV genes: These genes are more tissue specific, diverge more rapidly in expression in intra- and inter-specific contexts, are more likely to be located in repressive chromatin, and tend to be shorter with fewer introns than the average gene. YRV genes are also clustered in physical space in the nucleus but not clearly so (or only weakly so) along the chromosome.

Although some of these traits are of obvious interest (chromatin state, which we address in the next section), others are harder to interpret. It is not entirely obvious why genes affected by variation on the Y chromosome would be more tissue specific, except possibly as byproduct of a potential role of chromatin state in regulating expression of nonhousekeeping genes (many tissue-specific genes are in repressive chromatin; Filion et al. 2010; Kharchenko et al. 2011). The observation that YRV genes are more likely to be expressed postmeiotically in spermatogenesis provides a tantalizing link to our previous observations that Y chromosome divergence between *D. simulans* and *D. sechellia* seems to have strongly affected the regulation of a number of postmeiotic spermatogenesis genes (Sackton et al. 2011). A possible link between high mutation rates of Y-linked repetitive DNA (Lohe and Roberts 2000, 1990) and YRV could be invoked to interpret the connection between mutational variance, YRV, and gene expression divergence.

Of particular interest is the observation that YRV genes are more likely to be located in repressive chromatin than non-YRV genes. This observation is supported by other properties associated with repressive chromatin: YRV genes in general are later replicating than non-YRV genes, which is a common correlate of repressive chromatin. Similar to other regions of repressive chromatin, YRV genes are also more likely to be bound to the nuclear lamina than non-YRV genes.

We were particularly intrigued to note that, based on the chromatin classification scheme of Filion et al. (2010), YRV genes are primarily biased toward the two classes of nonpericentric heterochromatin (BLACK and BLUE). These two classes of chromatin share high levels of binding of three proteins (D1, SuUR, and LAM) that in turn distinguish them from other chromatin states. Although we were not able to find completely independent verifications of D1 or SuUR binding, we were able to confirm that YRV genes share a significant excess of binding to LAM in an independent data set.

The protein D1 is an AT-hook protein, containing a structural motif (the AT hook) that is known to bind to AT-rich sequences (Levinger 1985; Aulner et al. 2002). D1 has been shown in vitro and in vivo to bind to AT-rich satellite motifs in *D. melanogaster*, including the SATI and SATIII repeats that are localized to, among other regions, the Y chromosome (Aulner et al. 2002; Monod et al. 2002; Blattes et al. 2006). Binding of D1 is not, however, limited to the repetitive sequences, as recent high-throughput studies demonstrate D1 binding to dispersed euchromatin regions of the genome (Filion et al. 2010). Although an exact function for D1 is unknown, it is hypothesized to be a general transcriptional regulator (Levinger 1985; Smith and Weiler 2010).

These findings suggest a possible hypothesis for the basis of YRV, which we refer to as the heterochromatic sink model: If variation on the Y chromosome exists for the extent of D1 binding, this could change the genomic distribution of the D1 protein between Y introgression lines. Given the potential role of D1 in transcriptional regulation, this alone may be sufficient to influence gene expression. The implication is that the binding of chromosomal associated proteins to the Y chromosome alters their binding elsewhere, which in turn modifies gene expression profiles. In support of this model, we find that YRV genes have significantly more AT-rich upstream regions than non-YRV genes (YRV: 56.9% AT, non-YRV: 54.4% AT,

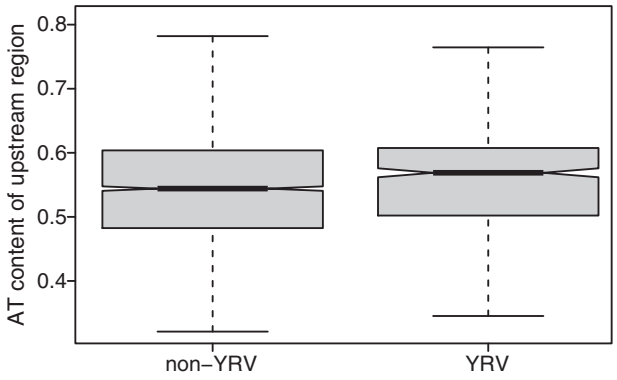

Fig. 8.—Boxplot of AT content of upstream regions adjacent to YRV and non-YRV genes. Mann–Whitney *U*, *P* value = 8.44×10^{-5}.

Mann–Whitney *U*, *P* value = 8.44×10^{-5}; fig. 8), implying that targets of D1 binding may be more prevalent proximate to YRV genes. The obvious implication of this is that we should be able to detect a difference in D1 binding in vivo across Y introgression lines, and future experiments are underway to test exactly this. It is important to note that although D1 is an obvious candidate given its binding properties, it is likely that other DNA-binding proteins may play an important role in this model.

An alternate, and potentially complementary, hypothesis is suggested by the observation that YRV genes both tend to occupy a particular place in nuclear space (near the nuclear envelope) and that YRV genes are significantly more clustered in physical nuclear space than linear space along the chromosome. In this model, which we call the spatial arrangement model, variation on the Y chromosome impacts the packing of chromosomes into the nucleus and thus the physical propensity for YRV genes to be in accessible or inaccessible regions of the genome.

In both the heterochromatin sink model and the spatial arrangement model, which are not mutually exclusive, the Y chromosome plays a role in modifying how the genome is distributed across chromatin compartments. This may be particularly important in the case of genes that are only expressed in limited contexts (postmeiotically in spermatogenesis, in single tissues), as expression of these genes may be particularly sensitive to small changes in the propensity to shift from silent to active chromatin contexts. Further work is needed, however, to explicitly test this hypothesis in an experimental context.

Supplementary Material

Supplementary file S1 is available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The authors thank Jun Zhou, Julien Ayroles, and Russ Corbett-Detig for valuable comments on the manuscript. This work was supported by the National Institutes of Health grant GM084236 to D.L.H.

Literature Cited

Aulner N, et al. 2002. The AT-hook protein D1 is essential for *Drosophila melanogaster* development and is implicated in position-effect variegation. Mol Cell Biol. 22:1218–1232.

Bachtrog D. 2005. Sex chromosome evolution: molecular aspects of Y-chromosome degeneration in *Drosophila*. Genome Res. 15: 1393–1401.

Bachtrog D. 2006. A dynamic view of sex chromosome evolution. Curr Opin Genet Dev. 16:578–585.

Bachtrog D, Charlesworth B. 2002. Reduced adaptation of a non-recombining neo-Y chromosome. Nature 416:323–326.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B. 57:289–300.

Blattes R, et al. 2006. Displacement of D1, HP1 and topoisomerase II from satellite heterochromatin by a specific polyamide. EMBO J. 25: 2397–2408.

Bonaccorsi S, Lohe A. 1991. Fine mapping of satellite DNA sequences along the Y chromosome of *Drosophila melanogaster*: relationships between satellite sequences and fertility factors. Genetics 129: 177–189.

Brosseau GE. 1960. Genetic analysis of the male fertility factors on the Y chromosome of *Drosophila melanogaster*. Genetics 45: 257–274.

Calcagno V. 2012. CRAN—package glmulti [WWW Document], glmulti: model selection and multimodel inference made easy. [cited 2012 Dec 1] Available from: http://cran.r-project.org/web/packages/glmulti/index.html.

Canty A, Ripley BD. 2012. CRAN—package boot. [Internet]. boot: bootstrap R (S-Plus) functions. [cited 2012 Dec 1] Available from http://cran.r-project.org/web/packages/boot/index.html.

Carvalho AB, Dobo BA, Vibranovski MD, Clark A. 2001. Identification of five new genes on the Y chromosome of *Drosophila melanogaster*. Proc Natl Acad Sci U S A. 98:13225–13230.

Carvalho AB, Lazzaro BP, Clark A. 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.

Carvalho AB, et al. 2003. Y chromosome and other heterochromatic sequences of the *Drosophila melanogaster* genome: how far can we go? Genetica 117:227–237.

Chippindale AK, Rice WR. 2001. Y chromosome polymorphism is a strong determinant of male fitness in *Drosophila melanogaster*. Proc Natl Acad Sci U S A. 98:5677–5682.

Clark A. 1987. Variation in Y chromosome segregation in natural populations of *Drosophila melanogaster*. Genetics 115:143–151.

Clark A. 1990. Two tests of Y chromosomal variation in male fertility of *Drosophila melanogaster*. Genetics 125:527–534.

Clark A, Lyckegaard E. 1990. Two neutrality test of Y-linked rDNA variation in *Drosophila melanogaster*. Evolution 44:2106–2112.

David JR, et al. 2005. Male sterility at extreme temperatures: a significant but neglected phenomenon for understanding *Drosophila* climatic adaptations. J Evol Biol. 18:838–846.

Dimitri P, Pisano C. 1989. Position effect variegation in *Drosophila melanogaster*: relationship between suppression effect and the amount of Y chromosome. Genetics 122:793–800.

Dobzhansky T. 1935. The Y chromosome of *Drosophila pseudoobscura*. Genetics 20:366–376.
Filion GJ, et al. 2010. Systematic protein location mapping reveals five principal chromatin types in *Drosophila* cells. Cell 143:212–224.
Gatti M, Pimpinelli S. 1983. Cytological and genetic-analysis of the Y-chromosome of *Drosophila melanogaster*. 1. Organization of the fertility factors. Chromosoma 88:349–373.
Jiang P-P, Hartl DL, Lemos B. 2010. Y not a dead end: epistatic interactions between Y-linked regulatory polymorphisms and genetic background affect global gene expression in *Drosophila melanogaster*. Genetics 186:109–118.
Kaiser VB, Charlesworth B. 2010. Muller's ratchet and the degeneration of the *Drosophila miranda* neo-Y chromosome. Genetics 185: 339–348.
Kharchenko PV, et al. 2011. Comprehensive analysis of the chromatin landscape in *Drosophila melanogaster*. Nature 471:480–485.
Koerich LB, Wang X, Clark A, Carvalho AB. 2008. Low conservation of gene content in the *Drosophila* Y chromosome. Nature 456: 949–951.
Kopp A, Frank A, Fu J. 2006. Historical biogeography of *Drosophila simulans* based on Y-chromosomal sequences. Mol Phylogenet Evol. 38:355–362.
Kopp A, Frank AK, Barmina O. 2006. Interspecific divergence, intrachromosomal recombination, and phylogenetic utility of Y-chromosomal genes in *Drosophila*. Mol Phylogenet Evol. 38:731–741.
Krsticevic FJ, Santos HL, Januário S, Schrago CG, Carvalho AB. 2010. Functional copies of the Mst77F gene on the Y chromosome of *Drosophila melanogaster*. Genetics 184:295–307.
Larracuente AM, Clark AG. 2012. Surprising differences in the variability of Y chromosomes in African and cosmopolitan populations of *Drosophila melanogaster*. Genetics 2013;193:201–214.
Larracuente AM, et al. 2008. Evolution of protein-coding genes in *Drosophila*. Trends Genet. 24:114–123.
Lemos B, Araripe LO, Hartl DL. 2008. Polymorphic Y chromosomes harbor cryptic variation with manifold functional consequences. Science 319: 91–93.
Lemos B, Branco AT, Hartl DL. 2010. Epigenetic effects of polymorphic Y chromosomes modulate chromatin components, immune response, and sexual conflict. Proc Natl Acad Sci U S A. 107: 15826–15831.
Levinger LF. 1985. D1 protein of *Drosophila melanogaster*. Purification and AT-DNA binding properties. J Biol Chem. 260:14311–14318.
Lohe AR, Hilliker AJ, Roberts PA. 1993. Mapping simple repeated DNA sequences in heterochromatin of Drosophila melanogaster. Genetics 134:1149–1174.
Lohe AR, Roberts PA. 1990. An unusual Y chromosome of *Drosophila simulans* carrying amplified rDNA spacer without rRNA genes. Genetics 125:399–406.
Lohe AR, Roberts PA. 2000. Evolution of DNA in heterochromatin: the *Drosophila melanogaster* sibling species subgroup as a resource. Genetica 109:125–130.
Lyckegaard E, Clark A. 1989. Ribosomal DNA and Stellate gene copy number variation on the Y chromosome of *Drosophila melanogaster*. Proc Natl Acad Sci U S A. 86:1944–1948.
Marshall WF. 2002. Order and disorder in the nucleus. Curr Biol. 12: R185–R192.
Monod C, Aulner N, Cuvier O, Käs E. 2002. Modification of position-effect variegation by competition for binding to *Drosophila* satellites. EMBO Rep. 3:747–752.
Paredes S, Branco AT, Hartl DL, Maggert KA, Lemos B. 2011. Ribosomal DNA deletions modulate genome-wide gene expression: "rDNA-Sensitive" genes and natural variation. PLoS Genet. 7: e1001376.
Pickersgill H, et al. 2006. Characterization of the *Drosophila melanogaster* genome at the nuclear lamina. Nat Genet. 38:1005–1014.
Rifkin SA, Houle D, Kim J, White KP. 2005. A mutation accumulation assay reveals a broad capacity for rapid evolution of gene expression. Nature 438:220–223.
Ritchie M, et al. 2006. Empirical array quality weights in the analysis of microarray data. BMC Bioinformatics 7:261.
Ritchie ME, et al. 2007. A comparison of background correction methods for two-colour microarrays. Bioinformatics 23:2700–2707.
Rohmer C, David JR, Moreteau B, Joly D. 2004. Heat induced male sterility in *Drosophila melanogaster*: adaptive genetic variations among geographic populations and role of the Y chromosome. J Exp Biol. 207:2735–2743.
Sackton TB, Montenegro H, Hartl DL, Lemos B. 2011. Interspecific Y chromosome introgressions disrupt testis-specific gene expression and male reproductive phenotypes in *Drosophila*. Proc Natl Acad Sci U S A. 108:17046–17051.
Schwaiger M, Schübeler D. 2006. A question of timing: emerging links between transcription and replication. Curr Opin Genet Dev. 16: 177–183.
Schwaiger M, et al. 2009. Chromatin state marks cell-type- and gender-specific replication of the *Drosophila* genome. Genes Dev. 23:589–601.
Sexton T, et al. 2012. Three-dimensional folding and functional organization principles of the *Drosophila* genome. Cell 148: 458–472.
Smith MB, Weiler KS. 2010. *Drosophila* D1 overexpression induces ectopic pairing of polytene chromosomes and is deleterious to development. Chromosoma 119:287–309.
Smyth GK. 2004. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol. 3:e3.
Smyth GK. 2005. Limma: linear models for microarray data. In: Gentleman R, Carey V, Dudoit S, Irizarry R, Huber W, editors. Bioinformatics and computational biology solutions using R and bioconductor. New York: Springer. p. 397–420.
Stoltenberg SF, Hirsch J. 1997. Y-chromosome effects on *Drosophila* geotaxis interact with genetic or cytoplasmic background. Anim Behav. 53:853–864.
Stouffer SA, Suchman EA, Devinney LC, Star SA, Williams RM Jr. 1949. The American soldier: adjustment during army life. (Studies in social psychology in World War II, Vol. 1.). Oxford: Princeton University Press.
Vibranovski MD, Koerich LB, Carvalho AB. 2008. Two new Y-linked genes in *Drosophila melanogaster*. Genetics 179:2325–2327.
Vibranovski MD, Lopes HF, Karr TL, Long M. 2009. Stage-specific expression profiling of *Drosophila* spermatogenesis suggests that meiotic sex chromosome inactivation drives genomic relocation of testis-expressed genes. PLoS Genet. 5:e1000731.
Wasbrough ER, et al. 2010. The *Drosophila melanogaster* sperm proteome-II (DmSP-II). J Proteomics. 73:2171–2185.
Zhou J, et al. 2012. Y chromosome mediates ribosomal DNA silencing and modulates the chromatin state in *Drosophila*. Proc Natl Acad Sci U S A. 109:9941–9946.
Zurovcova M, Eanes WF. 1999. Lack of nucleotide polymorphism in the Y-linked sperm flagellar dynein gene *Dhc-Yh3* of *Drosophila melanogaster* and D. *simulans*. Genetics 153:1709–1715.

Associate editor: Soojin Yi