Response to Reviewers

Hill-Robertson Interference Reduced Genetic Diversity on a Young Plant Y-chromosome

Josh Hough, Wei Wang, Spencer C. H. Barrett, and Stephen I. Wright

Editor: Bret Payseur
Article Type: Genetics of Sex
TOC Subsection Heading: Population and Evolutionary Genetics
Corresponding Author: Josh Hough
Keywords: Deleterious mutations; Interference Selection; Nucleotide diversity; Suppressed recombination

Editor: **Two experts in the field have reviewed your manuscript, and I have read it as well. My conclusion is that your manuscript would need to be substantially revised to be acceptable for publication in GENETICS.**

We would reconsider a substantially revised manuscript. If you decide to resubmit, please respond in detail to each reviewer concern. Two issues deserve special attention. First, please evaluate the alternative possibility that the reduction in Y-linked nucleotide variation you observed was generated by selective sweeps. Although I appreciate that selective sweep models are more difficult to fit than background selection, I agree with Reviewer 2 that sweeps deserve serious consideration as an explanation for your results. Second, please address potential issues with sequencing coverage and variant calling raised by Reviewer 1.

I look forward to receiving a revised manuscript. I expect it could be submitted within 90 days, but please let me know if you think you will need more time to complete the revision. I may send your resubmission out for review.

Authors: *We thank both of the reviewers and editor for their perceptive comments, which have improved our manuscript considerably. We have revised our manuscript to address the concerns raised, and have conducted each of the additional analyses that they requested (or which were implied from their comments as being important). We have also added a new section in which we provide a more rigorous evaluation of the alternative possibility that the reduction in Y-linked diversity we observed was due to selective sweeps. We also edited the text for clarity, paying particular attention to instances in which the reviewers thought the presentation or nomenclature was unclear (e.g., we have removed the phrase chromosome "race", etc.). Finally, we have updated our supplementary information document to provide a detailed description of our simulation methods and parameters, as well as the code used to run the simulation software. Below, we provide a point-by-point response to the reviewer's comments and highlight the associated changes that are now incorporated into our revised manuscript and SI text.*

Associate Editor Comments:

1. Throughout the paper the authors switch between talking about populations and races. I understand that it is typically in plant genetics to talk about races, but it may be time to

be consistent, and also sensitive to current climate. At a minimum, the authors should be consistent throughout the paper. Preferably, the authors could simply refer to the XY population and the XY1Y2 population.

Authors: *Agreed, this terminology comes from the classic Rumex literature, and we have dropped the term "race" in the revised manuscript.*

2. The authors refer to "gene silencing" in the abstract and on line 126. I think this can be confusing as gene-silencing is also used in reference to dosage compensation. Rather, I think the authors are referring to pseudogenization of genes on the Y chromosome. If this is true, perhaps the authors could use either this, or a more specific word. If I'm not understanding, then perhaps the authors can provide more detail about what kind of silencing (methylation, etc?) they are referring to.

Authors: *The possibilities we are considering here is that most genes on the Y may experience lowered expression early in the process of sex chromosome evolution (by some combination of methylation, heterochromatin formation, and possibly dosage compensation) and therefore are evolving mostly by neutral processes. Under this scenario much of the loss of gene function and degeneration on the Y may occur neutrally rather than due to ongoing linked selection. We have reworded the abstract and Introduction to make clear we are referring to reduced expression on the Y.*

Abstract:
"Given the relatively recent origin of R. hastatulus sex chromosomes, our results imply that Y-chromosome degeneration in the early stages may be largely driven by selective interference rather than by neutral genetic drift of silenced Y-linked genes."

Introduction:
"That is, if Y-linked gene silencing occurs early during sex chromosome evolution due to a combination of dosage compensation, heterochromatin formation, and loss of regulatory elements, then few sites may be under selection, and most Y chromosome degeneration may be driven primarily by genetic drift rather than selective interference."

3. The authors (line 109) refer to the Y chromosome evolving *de novo* from autosomes (it seems in an effort to make a contrast with the X and Y of humans). This is confusing, as typically X and Y chromosomes evolve *de novo* from autosomes, as did the human X and Y. While there are translocations and additions, in most cases (except now the new pillbug-Wolbachia case) we generally expect that sex chromosomes will arise *de novo* from autosomes. Could the authors provide more rationale for this sentence.

Authors: *Agreed. The sentence was not clear. The point being made was about the time scale over which purifying selection might have a strong effect, which is why we contrast the young*

Rumex sex chromosomes with those in humans. By emphasizing the 'de novo' point, we were trying to distinguish our work from studies of neo-sex chromosomes derived from fusions (e.g. in Drosophila miranda), which also examine early stages of sex chromosome evolution but through fusions onto pre-existing sex chromosomes, where factors like the spread of heterochromatin from ancestral sex chromosomes can come into play. We agree that X and Y chromosomes typically evolve de novo from autosomes, and have done so in humans and plants. We have modified the sentence as follows:

"However, given that human sex chromosomes originally evolved ~200 million years ago (Lahn and Page 1999; Ross et al. 2005), it is not clear whether purifying selection might have had a similarly strong effect on Y chromosomes that arose much more recently - e.g., within the last ~20 MYA in the case of dioecious plants (Charlesworth 2015)."

Reviewer 1: **The authors use transcriptomics to call variants on the X, Y, and autosomes, and use these variant calls to infer levels of selection acting on the Y and X relative to the autosomes. The paper is cleanly written and easy to read. The results are consistent with the evidence presented. I have some questions about the data and methodology, outlined below. I also have few minor comments. My comments are largely for discussion, though the authors may consider some to be worth pursuing.**

1. Was the reference transcriptome of an XX or an XY individual?

Authors: *The reference transcriptome was assembled de novo from pooled paired-end illumina reads from six females (XX) of the XY_1Y_2 sex chromosome system (also referred to as the "North Carolina Race", but we have now removed such language as per the editor's suggestion). We have clarified this in the methods.*

If from an XX individual, what do you expect to be the reduction in mapping quality of sequences from XY individuals. As most of the paper is based on calling fewer variants on the Y chromosome, great care should be taken to make sure the mapping to the sex-linked regions is not the reason for the reduction in Y-linked diversity (or at least part of it).

Authors: *We understand the possible concern about high X-Y sequence divergence leading to reduced mapping quality, and thus fewer ascertained Y-linked polymorphisms. However, there are several lines of evidence that argue against this. First, empirical estimates of mapping quality (in terms of the proportions of mapped and unmapped reads, the proportion of reads mapped with proper-pairing, and MAPQ scores), for XX and XY individuals are very similar on average (see table below), indicating that reads from XY individuals did not experience a major reduction in mapping quality.*

Mapping statistics for male and female alignments to XX reference transcriptome.

	Mapped reads	*Unmapped reads*	*Proper-pair*	*Quality </= 30*	*MAPQ score*
females	*75.7 %*	*24.3 %*	*67.7 %*	*13.5 %*	*median: 60* *mean: 45.5*
males	*75.2 %*	*24.9 %*	*67.1 %*	*13.4 %*	*median: 60* *mean: 42.2*

Furthermore, estimated net X-Y sequence divergence in Rumex is very low (Hough et al. 2014), which suggests that the reduction in mapping quality of Y sequences from males should be minimal for most genes.

However, low Y relative to X expression can indeed affect our ability to properly phase and call Y-specific SNPs; this in fact can have the effect of increasing, rather than decreasing estimates of Y polymorphism, because the Y-specific phased sequence in a subset of individuals may end up falsely incorporating X-specific SNPs due to their higher prevalence in the mapped data, creating false polymorphisms. This was one of the concerns that led us to our phasing requirements, including ensuring that haplotype blocks contained fixed differences between the X and Y sequence, to ensure that Y expression and mapping was high enough for proper phasing. This concern was another reason for careful manual checking of the Y-specific phased SNPs, and we did in fact observe a number of cases where low Y expression led to phasing errors which were removed from the analysis. We have added additional text in the methods to highlight the possible concerns about poor mapping of divergent reads and how our method addresses these concerns (Line X).

2. Line 194 - The authors refer to genes with low coverage and low Phred Quality scores (<20), but I have two questions about this: A) <u>What is the coverage range across genes that were sequenced in each individual</u> - certainly it will vary widely - and why the 10X coverage limit. Given that most Y-linked genes may be expected to have reduced expression relative to X-linked sequences, you might be prematurely limiting your ability to identify Y-linked sequences by removing these "low-coverage" genes.

<u>Authors:</u> *Our 10x coverage requirement was implemented because our phasing approach requires significant coverage of both alleles to work effectively. We have added this point in the text (Line X). While this does mean that we have excluded some low-expression genes, this is unlikely to be biasing our results, since neutral polymorphism on the non-recombining Y should be comparable in high and low expression genes, and our generally high coverage means we have still spanned a broad range of expression levels. Our samples generally have approximately 15,000 genes with a mean coverage above 10x, suggesting that our approach is still enabling the sampling of a large fraction of the transcripts in the genome.*

And B) How do you have a Phred Quality score for a gene? I thought that these scores were given to reads, and that typically you would filter out reads with a Phred score less than "X". Did you instead filter out reads, not genes, with a given Phred score? Or, if it was genes, how was the Phred score per gene decided?

Authors: *Thanks for pointing this wording error out, we did not have a Phred Quality score for a gene, this was a per SNP filter. Whole genes were removed if they had too much missing data/too few synonymous sites post-filtering. This has been reworded in the manuscript (bottom of page 3)*

Methods:
"We filtered SNPs with quality scores less than 60, and required each sample to have coverage greater than 10x, and individual genotype quality scores (GQ) at SNP sites greater than 30"

3. Will you please provide a citation for the reported SNP error rate on line 283?

Authors: *This was empirically estimated from our family segregation data, where we called SNPs in family data where we knew all Y sequences should be identical. We have clarified the wording. (page 5).*

Methods:
"We further validated the results of HAPCUT's allele phasing by comparing the accuracy of this method with the phasing-by-segregation method that was conducted in (Hough et al. 2014) . To do this, we first phased the sequence data from parents and their progeny using HAPCUT's algorithm (using the same parameters as for the population data), and then identified cases where SNPs were inferred on the Y chromosome by HAPCUT, but where the true level of polymorphism due to family structure (father and sons) was zero. We identified 7% of sex-linked genes that either had phasing errors or genotyping errors. Based on this family-level test for false SNPs, this corresponds to a SNP error rate estimate of 1.7 x 10- 4 . Note that this rate is very low relative to population-based estimates of polymorphism on the X and autosomes (Table 1), and therefore should have minimal effects on our estimation of the X/A ratio."

4. Is it justifiable to add an additional round of filtering to the Y chromosome for false SNPs but not to the autosomes and X chromosome? Presumably this, along with other factors, could have contributed to the greater reduction in Y-linked diversity.

Authors: *We had this worry as well; however, given that the estimated error rate is more than an order of magnitude lower than X polymorphism, and the fact that the errors were primarily due to phasing issues which would not affect X polymorphism estimates in females, our analysis should be robust. Furthermore, estimates of X polymorphism in phased males (see our response to the next query below) are nearly identical to unphased estimates in females, suggesting our X polymorphism estimates are robust to these two different approaches. In*

contrast, Y polymorphism is the same order as the error rate, and so these errors can severely bias diversity estimates on the Y, making it very important to have the additional round of manual filtering for these sequences.

5. When computing X-linked and Y-linked diversity, did the authors compute the X-linked diversity for all individuals? As a check, might it be worth running the X-chromosomes of XY-individuals through the same set of filtering as the Y chromosomes, and computing diversity just in these regions (which will, presumably, show similar sequence coverage to their Y-linked counterparts... unless there is reduction in Y-linked transcription, or mapping bias against Y-linked alleles due to increased divergence).

Authors: *We have calculated X polymorphism on both the phased X chromosomes of males, as well as the unphased X chromosomes of females, and get very comparable estimates. For females our weighted average estimate of synonymous theta is 0.00453, whereas for phased males it is 0.00472. This slight excess is in fact very much in line with the estimated error rate due to phasing on the Y, consistent with our expectation that errors are primarily due to the phasing process and that they are similar on the X. This is now reported in the manuscript (page 5)*

6. Did the authors consider making an alternative reference genome with the fixed X-Y differences, and remapping the sex-linked alleles/genes? I think this could be done just for the X-linked and Y-linked reads (and unmapped reads) from the BAM files, and wouldn't require re-aligning the whole genome. This would increase confidence in Y-linked allele variation.

Authors: *We did indeed consider this; our concern about taking this approach is that the typically low divergence between X and Y sequences (Hough et al 2014) would lead to frequent mapping errors in many cases, and generate large regions where SNPs and invariant sites are difficult to call because of cross-mapping of X and Y sequences. However, to address this question we tried this approach, mapping males and females to a reference transcriptome that includes both the original reference X sequence as well as phased Y sequences. Whereas the males showed greater mapping preference to the Y sequence than females, many reads still mapped to the Y-specific sequence in females, highlighting the low X-Y divergence generating complications with this approach. In particular, we see only a slight increase in the proportion of reads mapping to the Y in males vs. females (0.028 vs. 0.022), highlighting the potential pitfalls of quantifying diversity using simultaneous read mapping to the X and Y.*

7. My biggest concern was the use of transcriptomic data. But after a lot of thought, I'm less worried about it. If I am to think about it correctly, then the use of transcriptomic (transcribed regions) will necessarily mean the authors are analyzing constrained sequences, largely. That being said, the effective population sizes of the X and autosomes are larger than the Y, so natural selection should be more efficient on them.

Further, Y-linked alleles are, as the authors rightly discuss, affected by interfering selection and nearly completely linked (or expected to be). So, one presumes that, unless this species is extremely odd, the diversity can be expected to increase as one moves outside of genes on the X chromosome and autosomes, but should not be expected to increase on the Y chromosome as one moves away from genes. Thus, the broad conclusions of the paper would still hold. That said, Table 1 is very odd (as it seems to assume expected levels of *neutral* diversity in these very constrained regions). The X/A ratio is already quite high, and if patterns in non-Rumex species are to be believed, the X/A ratio will increase as one moves outside of, and away from, genes. This suggests, that the 0.85 X/A ratio observed is already quite high - higher than neutral expectations - and will be much higher after moving into less-constrained regions. It may be worth a sentence or two discussing the limitations of transcriptome sequence, and the predicted observations as one is able to analyze different genomic regions.

Authors: *This is an interesting point we hadn't thought about. We have added this into the discussion, and will be an interesting one to follow up on with more genomic information. Our observed X/A diversity ratio is in line with neutral expectations given our sex ratio in natural populations, suggesting that perhaps in our species the differential effects of hitchhiking are not as severe as observed in humans, but clearly another interpretation of this, like the reveiwer's is that 'neutral' X/A is very high. More investigation is needed to examine this possibility in more detail. We added an additional paragraph at the end of the discussion making these points:*

"One important caveat to our conclusions is that we have not considered the possible effects of background selection and selective sweeps in the recombining regions of the X and autosomes. Studies in a number of taxa, particularly in humans and maize (Hammer et al. 2010; Beissinger et al. 2016), have shown that diversity on recombining chromosomes increases as a function of distance from genes, likely due to the increasing effects of linked selection near regions with a high density of functional sites. Since the non-recombining Y chromosome will not experience this escape from hitchhiking away from genes, our conclusion about diversity loss on the Y may be conservative, and the true loss of diversity may be more extreme in unconstrained regions. In particular, if genome structure and selection parameters in Rumex are comparable to humans, we would predict the most gene-rich regions of recombining chromosomes to experience an approximately 60% reduction in diversity (McVicker et al. 2009), implying that our ratio of Y-linked to autosomal diversity in unconstrained regions may be as low as 0.011, in comparison to the ratio of 0.018 estimated here. While this may imply possibly even stronger effects of background selection and selective sweeps than inferred here, we note that correcting for this may provide an even better fit to the purifying selection model with our independently estimated number of selected sites (Figure 3). Future work integrating whole genome polymorphism analyses and estimates of the number of selected sites will enable an important examination of the possibility that the ratio of both Y/A and X/A diversity changes as a function of distance from genes (see Hammer et al. 2010)."

Reviewer 2:

In this study the authors have complied genomic data from 12 samples (RNAseq data) in *R. hastatulus* to study the early stages of sex chromosome evolution. The main result is a severe reduction of Y-chromosome diversity compared with the X chromosome and autosomes. The authors study several possible explanations and suggest that background selection explains the observed diversity on the Y chromosome (after ruling out biased sex ratios and variance in male reproductive success).

1. The authors show that background selection could explain Y-chromosome diversity. What is the evidence that selective sweeps are not the cause of this reduced diversity? The authors need to estimate parameters of background selection, selective sweeps or a combination of both modes of linked selection compatible with the data.

Authors: *We have now added positive selection to our forward simulations in order to explore the possible role of selective sweeps (and interference with beneficial mutations) on diversity loss. We used models that incorporated both positive and negative selection; since a model with positive selection but zero negative selection on the Y seemed implausible, we varied the proportion of selected sites subject to positive and negative selection, and the total number of sites under selection to evaluate a range of possible models. When we simulate using the independently estimated number of sites under selection (1.3 MB) from cytological data and estimates of the fraction of degenerated genes, we find our observed data is most consistent with a model of purifying selection alone, without the action of linked positive selection. However, if we reduce the number of selected sites, we find that models with positive selection are more likely, although do not show a statistical improvement over the purifying selection only models. Thus, our new simulations demonstrate that the relative importance of the contribution of positive and negative selection to the observed diversity loss depends in important ways on the number of sites under selection, which may be further refined with ongoing genomic efforts in this system. Given our prior estimate, our original conclusion that purifying selection alone can explain the detail is upheld, but we demonstrate the range of parameter space where a role for positive selection is also plausible.*

2. The authors cite Hough et al.(2014) to indicate that differences in diversity are not due to differences in mutation rates. Hough et al.(2014) also showed that rates of Y-linked genes have a dn/ds less than three-fold higher compared with X-linked genes. Is this very limited increase in dn/ds compatible with the changes in Ne proposed here? It seems to me that a 95-97% reduction in Ne necessarily predicts a much higher increase in dn/ds thus favoring recent selective sweeps as explanation. I believe that the authors need to

include dn/ds (divergence) data to study the causes of the observed reduction in Y-chromosome diversity.

Authors: *This is an important point, and in fact is a useful way to address whether the observed signal of weakened selection can be explained by the Ne reduction or not. The answer to this question will depend strongly on the shape of the distribution of deleterious fitness effects (DFE). To address this point, we used our estimated DFE of deleterious mutations from the X chromosome data. In particular, the inferred DFE on the X implies that 15.7 % of nonsynonymous sites have a strength of selection (Ns) <1, and are in the nearly neutral zone. If we use the estimated shape of the gamma distribution from the X, and take the observed reduction in diversity on the Y relative to the X of 0.021 to readjust the expected mean Ns for this chromosome, we predict that the Y chromosome should have 42.4% of mutations as effectively neutral. Thus, we expect a 2.7-fold increase in the fraction of effectively neutral deleterious mutations, which is very close* to our observed increase of 2.57 in dN/dS. We have added this point into the discussion:

"Using our observed loss of diversity on the Y, we can ask whether the inferred reduction in the effective population size explains the observed signal of reduced selection efficacy on the Y compared with the X in Hough et al. (2014). In particular, our previous work revealed a 2.57-fold increase in the rate of nonsynonymous relative to synonymous divergence. To address the rate of increased divergence expected on the Y, we can use our estimated DFE of deleterious mutations from the X chromosome data. In particular, the inferred DFE on the X implies that 15.7 % of nonsynonymous sites have a strength of selection (Ns) <1, and are in the nearly neutral zone. If we use the estimated shape of the gamma distribution from the X, and take the observed reduction in diversity on the Y relative to the X of 0.021 to readjust the expected mean Ns for this chromosome, we predict that the Y chromosome should have 42.4% of mutations as effectively neutral. Thus, we expect a 2.7-fold increase in the fraction of effectively neutral deleterious mutations, which is very close to our observed increase of 2.57 in dN/dS. This result provides further evidence to suggest that selective interference, rather than relaxed selection per se, is the primary factor driving Y chromosome degeneration in Rumex."

3. Please, report and discuss the frequency spectra of neutral mutations.

Authors: *The frequency spectra are comparable across all X, Y and autosomes. We have added mean Tajima's D values into the text (bottom of page 9): "The frequency spectra of variant sites of X, Y and autosomal loci, as quantified by mean Tajima's D at synonymous sites are similar (X, -0.43; Y, -0.34; Autosomal, -0.27)". In addition, we also calculated a second version of our likelihood from the simulations, using a summary that incorporates a summary of SNP frequencies (number of singletons), this is shown as the gray surface in Figure 4.*

4. The authors conduct forward simulations using SFSCODE but there is almost no information on important parameters. Please include population size, parameters

of the gamma distribution, number of replicates, etc.

Authors: *We apologize for this omission, these details are now in the Methods, and we have include the exact simulation commands in the supplementary material.*

Methods:
To study the effects of purifying selection on expected levels of Y-chromosome diversity, we conducted forward-time simulations of haploid Y chromosomes using the software SFS_CODE (Hernandez 2008) . We first estimated the distribution of fitness effects of deleterious amino acid mutations from our polymorphism data for X-linked genes using the method of (Keightley and Eyre-Walker 2007) , which fits a gamma distribution of negative selection coefficients to the observed frequency distribution of nonsynonymous and synonymous polymorphisms. This analysis estimated a mean strength of selection (Ns) of 493, and a shape parameter of 0.258 for the X chromosome. Rescaling the effective population size for the Y chromosome assuming a sex ratio of 0.6 gives a mean Ns for the Y of 493 x 0.259=128. We then used this estimated gamma distribution to parameterize the simulations, initializing them with our estimated θ_{aut} (Table 1), adjusted to reflect the neutrally expected N_{eY} for a sex ratio of r = 0.6, 0.0055 x 0.21=0.0011. To match our sample size and the number of synonymous sites sampled from our data (see Supporting Information), the simulations sampled 6 haploid chromosomes, and the genome sequence contained 45,331 bp of linked neutral sequence from which we calculated silent site diversity, π_s. To increase computational speed, we simulated a population of 500 haploid chromosomes. For each parameter set, we conducted replicate simulations (purifying selection alone: 50,000 reps; purifying and positive selection: 20,000 reps) and calculated πs for each replicate simulation.