How big is your Y? A genome-sequence based estimate of the size of the male-specific region in *Megaselia scalaris*

Kenneth B. Hoehn*§1 and Mohamed A. F. Noor§2

*Department of Zoology, University of Oxford, Oxford OX1 3PS, United Kingdom

§Biology Department, Duke University, Durham, NC 27708

---

1 Oriel College, Oxford, OX1 4EW, United Kingdom
2 French Family Science Center 4244, Box 90338, Durham, NC 27708 USA
Abstract:
The scuttle fly, Megaselia scalaris, is often cited as model in which to study early sex chromosome evolution because of its homomorphic sex chromosomes, low but measurable molecular differentiation between sex chromosomes, and occasional transposition of the male-determining element to different chromosomes in laboratory cultures. Counterintuitively, natural isolates consistently show sex linkage to the second chromosome. Frequent natural transposition of the male-determining element should lead to the loss of male specificity of any non-transposed material on the previous sex-linked chromosome pair. Using next generation sequencing data from a newly-obtained natural isolate of M. scalaris, we show that even highly conservative estimates for the size of the male specific genome are likely too large to be contained within a transposable element. This result strongly suggests that transposition of the male determining region is either extremely rare or has not persisted recently in natural populations, allowing for greater differentiation of the sex-chromosomes of this species.
Theories of sex chromosome evolution suggest that it begins when a cluster of sex-determining genes evolves on the same chromosome as genes conferring sex-specific fitness effects (see Charlesworth and Charlesworth 1978; Bull 1983; Charlesworth et al. 2005; Ellegren 2011). Selection favors reduced recombination among these genes, and sexually antagonistic functions accrue in the linked region, causing the autosomal pair to diverge into proto-X or Y chromosomes. The proto-Y region eventually only exists in heterozygous form, and the absence of recombination causes it to accumulate mutations and diverge from the proto-X region. The non-recombining region grows over time, creating "evolutionary strata" (Lahn and Page 1999) of divergence between the proto-X and Y chromosomes, and may eventually result in two chromosomes that essentially don't exchange alleles. Testing theories of sex-chromosome evolution requires studying proto-X and Y chromosome at different (especially early) stages along this process, such as found in members of the genus Silene (Bergero et al. 2007; Mrackova et al. 2008). However, understanding the very earliest stages necessitates studying the origin of a new sex-determining region immediately after it arises.

The scuttle fly, *Megaselia scalaris*, has been proposed as a model system for the earliest steps of sex chromosome evolution (Traut 1994). This species sports three chromosome pairs, but none of the mitotic metaphase chromosomes appear heteromorphic via microscopy (see review in Traut 2010). Nonetheless, some X-Y chromosome differentiation identified via Southern hybridization (Willhoeft and Traut 1990), RAPD markers (Traut 1994), and DNA sequencing of a 1.8-kb segment (Traut & Wollert 1998). More strikingly, based on linkage to phenotypic markers, the male-determining region transposes among the chromosomes in some laboratory cultures at a detectable rate of 0.05-0.4%
(Main 1964; Traut & Willhoeft 1990; Traut 1994; Traut 2010). In principle, transposition of a male-determining region off the proto-Y would allow the remainder of the proto-Y sequence variants to spread into females, and thus eliminate their male-specificity. Nonetheless, despite the detectably frequent transposition observed in laboratory culture, natural isolates from worldwide collections all have the male-determining region associated with the same chromosome (Traut 2010).

These findings raise the question of whether *Megaselia scalaris* has a large, distinct male-specific region (a true "Y-chromosome"), or whether transposition by the non-recombining male-determining region occurs at a detectable rate in natural populations and the male-specific region is in fact quite small (perhaps encompassing only the transposing fragment). Willhoeft and Traut (1990) identified several male-specific bands in Southern hybridizations of *M. scalaris* DNA, suggesting the male-specific region may be large. In particular, they found that one of eight randomly chosen DNA probes gave a purely male specific band in all strains studied. Assuming these probes are independent of each other, this observation suggests a male specific genome size of 12.5%. However, given only eight probes were used, this is far from a conclusive result.

In this study, we analyze genome sequences from male and female *Megaselia scalaris* to estimate the size of the male-specific region in this species. If the male-specific region is very large, then frequent transposition of the male-determining region and establishment on new chromosomes is improbable in natural populations. Taking the most conservative method possible, we find that the male specific genome is likely too large to be located on a small, frequently transposing element. However, the male specific genome could still represent a fraction of an otherwise large homomorphic sex chromosome.
MATERIALS AND METHODS

The source strain was derived from a single wild caught, fertilized female ("Durham, NC 2"). This strain was initially inbred for 5-7 generations using single full-sibling pairs, and then maintained as a stock in the laboratory for one year. Prior to sample collection, two further rounds of inbreeding were conducted using one single full-sibling pair per round. Sequences were obtained from 30-50 adult males and females separately. For sequencing the females, we used 4 lanes of Illumina/Solexa GA, paired-end, 75bp reads, giving a total of 79,232,896 sequences and 5.9x10^9 bp. The insert lengths for these paired-end Illumina reads are expected to be around 300bp. For the males, we used also used 4 lanes Illumina/Solexa GA, non paired-end, 76 bp reads, giving a total of 111,268,292 sequences and 8.4x10^9 bp. Using an estimated haploid genome size of 500 Mb (Rasmussen and Noor 2009), this gives a coverage of 5.9x10^9/5x10^8 ~ 12x coverage for females, and 8.4x10^9/5x10^8 ~ 17x coverage for males. The sequencing reads from males were assembled using SOAPdenovo2 (k =42, N50=292, Luo et al. 2012). Statistics for the male assembly are available in Supplemental Table S1.

We utilized Y Genome Scan (YGS), a kmer count based method (Carvalho and Clark 2013) to determine which of these contigs were potentially male-specific. This method scans across each contig from a male assembly, storing each unmatched single copy (USC) kmer in each contig and determining if there are any matching kmers in female short read data. From this, plotting the size of each contig against the percentage of USC kmers that are unmatched in the female short read data usually gives two distinct peaks at 100% USC kmers, and 0% USC kmers. These peaks are inferred to be contigs from the Y and autosomal chromosomes, respectively. In our case, we filtered our female short reads using Jellyfish (Marçais and Kingsford 2011, m=15, minquality=20, quality-start=33, lower-count=5).
We ran YGS, as recommended, with 15 mers. We used a Bayesian approach for size inference of the male specific genome. Because both our male and female assemblies were low coverage – which we expect to artificially inflate the number of USC kmers, and thus artificially inflate the number male specific contigs – we took the most conservative approach available and only considered contigs with 100% USC kmers as putatively male specific. We further modeled the probability that a putatively male specific contig was actually male specific as a binomial likelihood parameterized by the number of USC and matched single copy (MSC) kmers. Assuming a beta(a,b) prior, the posterior probability of a putatively male specific contig being male specific is a beta distribution:

\[
P(Male \text{ specific contig} \mid k\text{mers}) = \frac{1}{B(a,b)} (X)^a USC \text{ kmers} (1 - X)^{b + MSC \text{ kmers}}
\]

where \(a / (a + b) = 2 \times 10^{-4}\), giving a prior expected male specific genome size of 0.02%. The sum \(a + b\) is a measure of how heavily the prior is weighted relative to the data. Given a prior weight (\(a + b = PW\)), this system of equations solves to obtain a unique set of prior parameters by the formulas: \(a = PW \times 2 \times 10^{-4}\), and \(b = PW \times (1 - 2 \times 10^{-4})\).

The expected value of this distribution is then

\[
E(P(Male \text{ specific contig} \mid k\text{mers})) = \frac{a + USC \text{ kmers}}{a + b + USC \text{ kmers} + MSC \text{ kmers}}
\]
Choosing a prior weight of 1 means that a 100% USC kmer contig with only one SC kmer has an approximately 50% probability of being male specific. A prior weight of 22 – the median SC kmer count across all contigs in our *Megaselia* assembly – means that 100% USC kmer contigs with 22 SC kmers have a 50% probability of being male specific. Our process for male specific genome size estimation is– using different contig size cutoffs and prior weights – to calculate this expected value for each contig, and multiply by the length of the contig to get a kmer-weighted estimate of the size of the male specific genome. Specifically:

1. Set prior sum a + b, and then solve for $a / (a + b) = 2 \times 10^{-4}$.
2. Remove all contigs smaller than a selected size cutoff.
3. For each contig of 100% USC kmers, calculate the expected probability for male specificity using posterior distribution, multiply this probability by the length of the contig, and sum across all such contigs.
4. Divide the sum from (3) by the total length of all contigs above the size cutoff, giving the expected male specific proportion of the genome.
5. Repeat 1-4 with different size cutoffs and prior weights.

There are two boundary cases for what is meant by “expected male specific proportion of the genome” (hereafter MSP). In the first case, if there is very low genetic diversity in the individual sequenced—such as from a highly inbred line – we expect each locus in the genome to be represented by no more than one contig, and thus,

$$MSP_h = \frac{Y}{X + Y + \text{Autosomes}}$$
which is approximately the MSP of the haploid genome. In the opposite most extreme case, if parental autosomes are very divergent in the individual sequenced, we expect small, diverse contigs covering the same locus to not assemble together, meaning each locus will be represented by two contigs. Thus,

\[ MSP_d = \frac{Y}{X + Y + 2 \times \text{Autosomes}} \]

which is the MSP of the diploid genome. MSP\(_d\) is the more immediately usable and interpretable. To give the most conservative lower bound for MSP\(_d\), we assume that the MSP we measure from step 4) is MSP\(_h\) and then divide by two to obtain a lower bound for MSP\(_d\). From there, we can give a lower bound for the absolute size of the male specific genome by multiplying MSP\(_d\) by the diploid genome size.

To test this technique, we also tried it using a low coverage female genome assembly tested under Carvalho and Clark (2013). However, in this case the male specific peak was shifted below 100%, and as such we considered contigs with > 80% USC kmers.

**RESULTS**

Our application of the YGS method to locating putatively male specific contigs showed what we interpret as a successful separation of male specific and autosomal contigs. A heatmap (Supplemental Figure S1) plotting log(contig size) vs percent USC kmers shows two clear peaks at 0% USC kmers and 100% USC kmers. Like Carvalho and Clark (2013), we interpret these to be the result of male specific contigs not matching female short read kmers. Because we anticipate low coverage assemblies
to bias results towards higher percentages of USC kmers, we take the most conservative approach and only consider contigs with 100% USC kmers as putatively male specific.

We applied the procedure of estimating male specific genome size using contig size cutoffs of 0bp, 1000bp, 2000bp, and 3000bp. We further used different prior weights (a + b) between 1 and 70. This altered how heavily we weighted the expectation of small (0.02%) male genome size. From these results, we obtained a generally consistent estimation of the MSP for *M. scalaris*.

A prior weight of one gives a size estimate between 9.9% and 5.3%. Increasing the prior weight to 22 – the median SC kmer count across all contigs – lowered this to between 3.5% and 2.4%. Increasing the prior rates further caused a convergence of MSP, the lowest estimation was at a prior weight of 70, which returned between 1.5% and 1.2% MSP. A prior weight of 70 is likely far too conservative. These results are summarized in Supplemental Figure S2. Taking these results, we divide the percentage by two to be sure the estimate is below the MSP for the diploid genome (MSP$_d$) and multiply by the diploid genome size of ~1,000Mb (Rasmussen and Noor 2009). From this, we estimate a lower bound MSP$_d$ for the prior weight of 1 as 26.5Mb, 12Mb for the median prior weight, and 5.5Mb for the highly conservative weight of 70.

As a control, we applied this method to *Drosophila virilis* as well (also used in the Carvalho and Clark 2013 study), which differed from our *Megaselia* data set in that it had significantly higher coverage in its male assembly, and substantially fewer contigs. Furthermore, in this case there was a distinct peak of % USC kmers not at 100% but between 80% and 100%, so we considered all contigs in this range to be putatively male specific. In this analysis, we applied size cutoffs between 0 and 10Kb, and, because
the contigs had a kmer count median of 259, and a 90th percentile of 4971, we applied prior weights
between 1 and 10,000. These all gave a steady lower bound estimate of 6.6% for the least conservative,
6.1% for the median, and 3.6% for the most conservative. These results are summarized in
Supplemental Figure S3.

Flow cytometry studies (Gregory and Johnston 2008) suggest that the Y chromosome of *D. virilis* is
approximately as large as the X, and that the X is approximately as large as each of the four autosomes
(FlyBase: http://flybase.org/reports/FBsp00000251.html). This gives an approximate expected MSP\(_d\) of
1/10 = 0.1, and MSP\(_h\) of 1/6 = 0.167. The technique clearly acts, as we anticipated, as a very
conservative lower bound. Because the male assembly for *D. virilis* is much more complete than
*Megaselia*, it makes sense that the least conservative estimates of MSP would be the most accurate
estimation because we anticipate far fewer false positive scaffolds. In *Megaselia*, we expect a large
portion of our 100% USC contigs, particularly those of small size, to be false positives, so a more
conservative approach is warranted. However, the MSP of *Megaselia* may still be much larger than
estimated through this conservative method. Further, because our assembly is at such a lower coverage,
we expect the MSP measured through our method to be closer to MSP\(_d\) than to MSP\(_h\), which, if true,
would effectively double our estimates of the lower bound size of the male specific genome of
*Megaselia*.

**DISCUSSION**

*Megaselia scalaris* has long been cited as an example of early sex chromosome evolution. Early
karyotypic studies of this fly's chromosomes revealed homomorphic sex chromosomes (Traut et al.
and molecular studies revealed some differentiation between the two chromosomes (Wilhoeft
1990) and Traut 1990). However, Traut and Wilhoeft (1990) showed that the male determining region (M)
regularly transposes from the second chromosome to the third in the laboratory. Followup studies on
the same lines (Traut 1994) suggested that this transfer did not occur as a translocation at the
chromosome ends, but as a comparatively small, complex transposable element.

The frequent transposition of the male determining region is seemingly at odds with observations that
natural isolates appear to always show sex linkage to the second chromosome (Traut 2010). Frequent
and persistent transposition in natural populations would rapidly lead to the decay of male specificity of
previously male specific parts of the genome. Hence, if transposition were frequent, we would expect
the size of the male specific region to be not much larger than the size of known transposable elements.
However, all of our estimates of the lower bound size of the male-specific genome – even those taken
from a highly conservative approach – are multiple orders of magnitude larger than the lengths of
known transposable elements, the largest of which, “Mavericks,” are on the order of 20-50kb
(Feschotte and Pritham 2007). Our findings indicate that location of the male determining region in this
natural isolate (Durham 2) has been persistent long enough in natural populations for a large male
specific genomic region to evolve. However, these results do not necessarily conflict with observations
of transposition in the lab, and may be reconciled through population cage experiments (reviewed in
Traut 2010) showing sex linkage to the second chromosome gives a selective advantage in well-aerated
cages, but is lost under crowded cage conditions. Hence, more generally, these findings could suggest
that the male-determining region is able to transpose in natural populations, but that it is rapidly
selected against when linked to the first or third chromosome.
Ultimately, while our results suggest that transposition of the male determining element of *Megaselia scalaris* has not occurred or persisted recently in natural populations, our results are not wholly inconsistent with the hypothesis that *Megaselia scalaris* is currently at an early stage of sex chromosome evolution. The full size of the male specific region, and an accurate estimate of the age of the transposition event which established it, are beyond the scope of this analysis. However, we were able to show that a highly conservative estimate of the size of the male specific genome reveals that it is far too large to be contained within a known transposable element, indicating that the male determining element has remained linked to its current position long enough for a male specific genome of at least several million bases to be established.

**ACKNOWLEDGEMENTS**

KBH was supported by an A.B. Duke undergraduate scholarship and Marshall scholarship while completing this project. Sequencing was supported by NIH grant GM086445. Short-read sequences are available at [http://www.ncbi.nlm.nih.gov/sra/SRX208940%5Baccn%5D](http://www.ncbi.nlm.nih.gov/sra/SRX208940%5Baccn%5D), and a draft assembly is available at: [http://metazoa.ensembl.org/Megaselia_scalaris/Info/Index](http://metazoa.ensembl.org/Megaselia_scalaris/Info/Index)
REFERENCES


