Amm_transcriptome_results_update3

We investigated the issue of second cloud of female-biased genes in brain tissues, it turned out that 30% transcripts do not have any BLAST hit (88%) with new XX genome.
We removed the non-frog transcripts, and redone all analysis using the updated transcriptome.

A. Tissue specificity index tau

B. Sex-biased gene expression along developmental stages, across adult tissues

C. Transcription degeneration

D. Coding sequence divergence and Faster-X

A. Tissue specificity index tau

Approach:

1. Define in total 11 tissues, since somatic tissues and early embryonic tissues have very small number of sex-biased genes, we treat them as single tissue regardless of sex, e.g. G23, G27, G31, liver and brain; for late larva tissues and gonad tissues, we treat tissue with separate sexes, e.g. ovary, testis, XX G43, XY G43, XX G46 and XY G46 (see Brown & Bachtrog 2014).
2. Use Kallisto to quantify transcript expression, and then calculate Tau from the generated output of TPM matrix.
3. Use tissue specificity index Tau formula below (Mank et al. 2008; Brown & Bachtrog 2014):

$$\tau = \frac{\sum_{i=1}^{N} 1 - \frac{logE_i}{logE_{max}}}{N-1},$$

i is one tissue, Ei is the expression (non-normalized TPM value) of certain transcript.

4. Keep transcripts which are expressed in at least one of the 11 tissues with TPM >=1.

Figure 1. Histogram distribution of Tau for all transcripts in Ammarnas transcriptome.

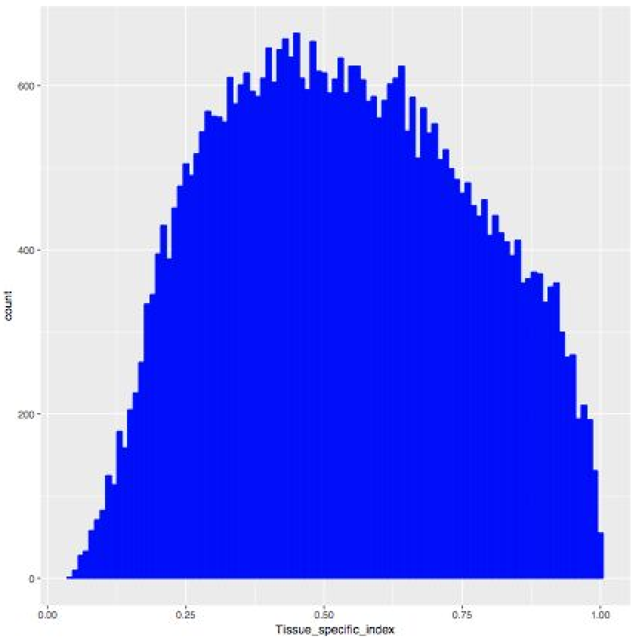


Figure 2. Boxplot of sex-biased genes and the tissue specificity index Tau at stage G46 (gene expression from sex-reversed XX male is not included here).

Sex-biased genes are more tissue specific than unbiased genes (note, male-bias genes have higher tau than female-biased genes).

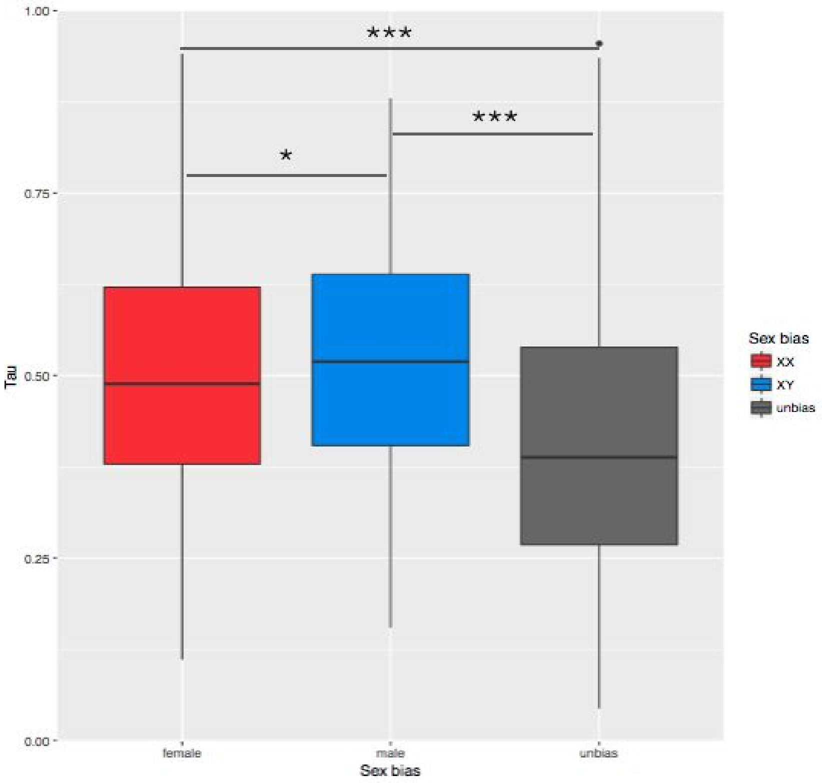
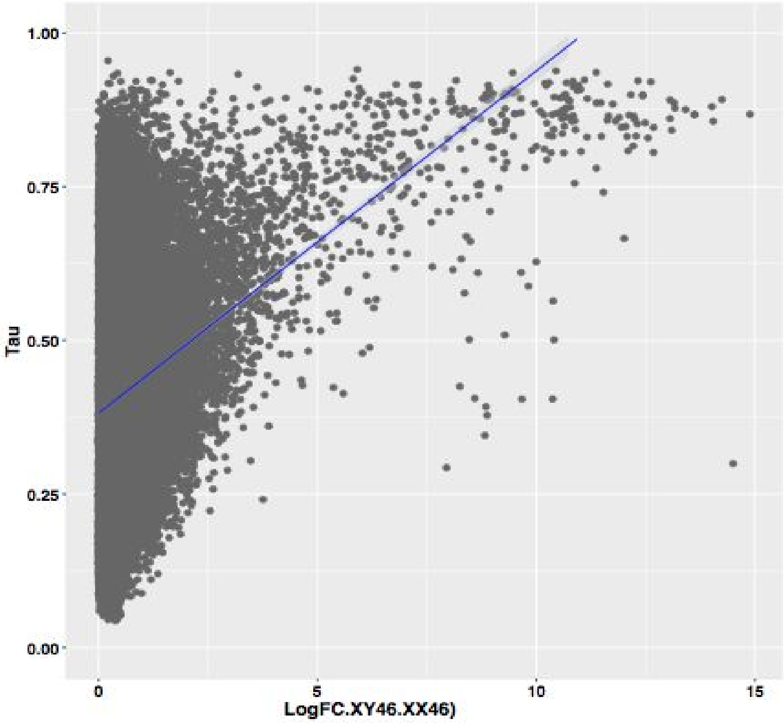


Figure 3. Correlation between absolute values of gene expression ratio Log2(XY46/XX46) (include sex-biased and unbiased genes) and tissue specificity index tau.

```
glm(formula = tau ~ sqrt(abs(g46_tau_sub$logFC.XY46.XX46)) *
    bias, family = binomial, data = g46_tau_sub)
```

Results show tau is significantly correlated with expression ratio, sex bias, as well as the interaction of the two.



Furthermore, we could also ask whether the evolutionary rate of coding sequence can be explained by tissue specificity, or sex bias, or the interaction of the two factors.

```
glm(formula = tau ~ sqrt(abs(g46_tau$logFC.XY46.XX46)) * bias,
    family = binomial, data = g46_tau)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.05176  -0.26724  -0.03895   0.22733   1.22526

Coefficients:
                                       Estimate Std. Error z value Pr(>|z|)
(Intercept)                            -1.66392    0.12150 -13.694  < 2e-16 ***
sqrt(abs(g46_tau$logFC.XY46.XX46))      1.14631    0.08078  14.191  < 2e-16 ***
biasmale                                1.26968    0.30488   4.164 3.12e-05 ***
biasunbias                              0.92699    0.12807   7.238 4.54e-13 ***
```

```
sqrt(abs(g46_tau$logFC.XY46.XX46)):biasmale   -0.79765    0.22160  -3.599 0.000319 ***
sqrt(abs(g46_tau$logFC.XY46.XX46)):biasunbias -0.57202    0.09956  -5.746 9.16e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3473.1  on 24405  degrees of freedom
Residual deviance: 2965.7  on 24400  degrees of freedom
AIC: 30606

Number of Fisher Scoring iterations: 4
```

If removing unbiased genes, the results largely remain, but the interaction does not influence the dNdS anymore.

```
glm(formula = tau ~ sqrt(abs(g46_tau_sub$logFC.XY46.XX46)) *
    bias, family = binomial, data = g46_tau_sub)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.05176  -0.18491  -0.01301   0.17391   0.97283

Coefficients:
                                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                                   -1.66392    0.12150 -13.694  < 2e-16 ***
sqrt(abs(g46_tau_sub$logFC.XY46.XX46))          1.14631    0.08078  14.191  < 2e-16 ***
biasmale                                        1.26968    0.30488   4.164 3.12e-05 ***
sqrt(abs(g46_tau_sub$logFC.XY46.XX46)):biasmale -0.79765    0.22160  -3.599 0.000319 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 568.53  on 4677  degrees of freedom
Residual deviance: 313.42  on 4674  degrees of freedom
AIC: 5657.4

Number of Fisher Scoring iterations: 4
```

#####################
Whether dN/dS can be explained by tau, sex bias, or the interaction.

```
y1a <- lm(sqrt(dNdS) ~ sqrt(tau) * bias, g46_tau_dnds)
Coefficients:
  Estimate Std. Error t value Pr(>|t|)
(Intercept)           0.10758    0.01506   7.143 1.01e-12 ***
  sqrt(tau)            0.23680    0.02136  11.088  < 2e-16 ***
  biasmale              0.10490    0.05529   1.897   0.0578 .
biasunbias           0.10901    0.01620   6.730 1.84e-11 ***
  sqrt(tau):biasmale   -0.09704    0.07573  -1.282   0.2001
sqrt(tau):biasunbias -0.15363    0.02322  -6.617 3.96e-11 ***
  ---
  Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

#######
```

If removing unbiased genes,

```
y2a <- lm(sqrt(dNdS) ~ sqrt(tau) * bias, g46_tau_dnds_sub)
#############
Coefficients:
  Estimate Std. Error t value Pr(>|t|)
(Intercept)         0.10758    0.01503   7.156  1.2e-12 ***
  sqrt(tau)          0.23680    0.02132  11.109  < 2e-16 ***
  biasmale            0.10490    0.05518   1.901   0.0575 .
sqrt(tau):biasmale -0.09704    0.07559  -1.284   0.1993
---
  Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

At stage G46, the coding sequence evolutionary rate can be explained by tissue specificity, sex bias, and also their interaction.

Figure 4. Boxplot of sex-biased genes and the tissue specificity index Tau in gonad tissues. Here again, sex-biased genes are more tissue specific than unbiased genes, but within sex-biased genes, opposite as G46, female-biased genes has higher tau than male-bias.

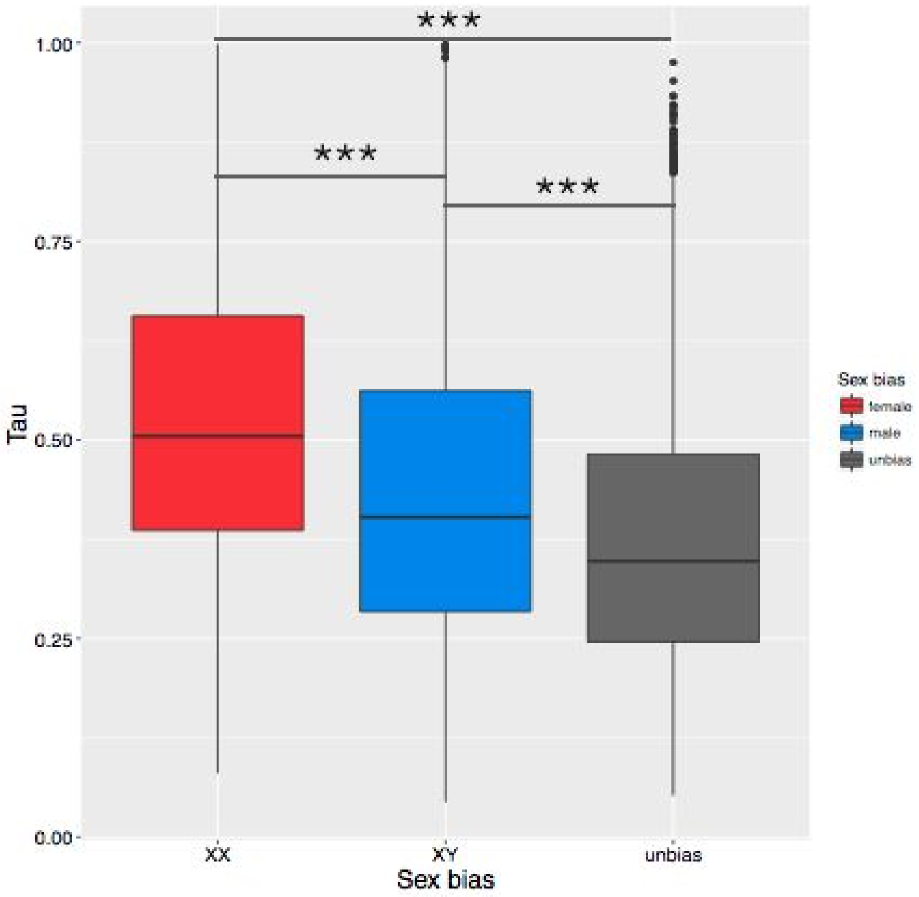
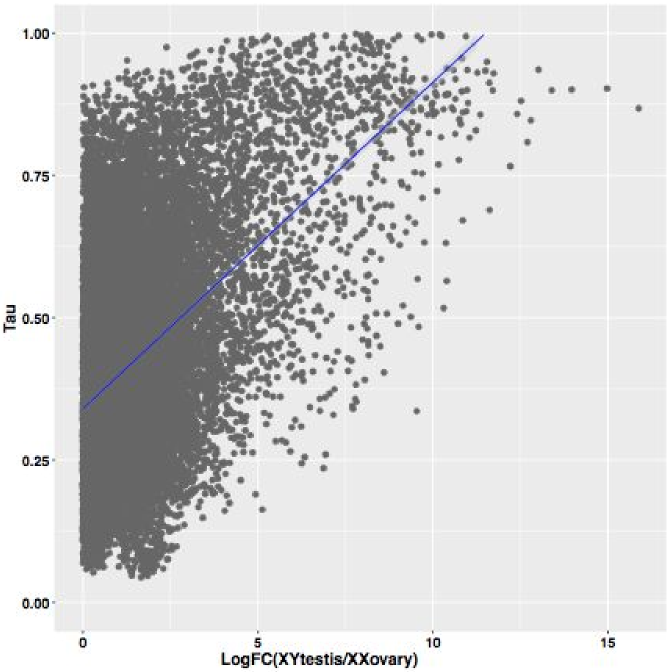


Figure 5. Correlation between absolute values of gene expression ratio Log2(male/female) (only for sex-biased and unbiased genes) and tissue specificity index tau.

```
glm(formula = sqrt(tau) ~ sqrt(abs(logFC.XYtestis.XXovary)) *
    bias, family = binomial, data = gonad_tau_sub)
```

Tau is significantly correlated with expression ratio, but not with sex bias, nor with the interaction



Similaly, for gonad tissues, we could also ask whether the evolutionary rate of coding sequence can be explained by tissue specificity, or sex bias, or the interaction of the two factors.

```
glm(formula = sqrt(tau) ~ sqrt(abs(logFC.XYtestis.XXovary)) *
    bias, family = binomial, data = gonad_tau)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.94063  -0.18875  -0.00767   0.19365   0.92346

Coefficients:
                                        Estimate Std. Error z value Pr(>|z|)
(Intercept)                             -0.45131    0.11497  -3.926 8.65e-05 ***
sqrt(abs(logFC.XYtestis.XXovary))        0.91428    0.07619  11.999  < 2e-16 ***
biasmale                                -0.09787    0.16565  -0.591    0.555
biasunbias                               0.53591    0.12671   4.229 2.34e-05 ***
sqrt(abs(logFC.XYtestis.XXovary)):biasmale   -0.15882    0.10971  -1.448    0.148
sqrt(abs(logFC.XYtestis.XXovary)):biasunbias -0.43344    0.10830  -4.002 6.28e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

If removing unbiased genes, redo the linear regression, the results remain the same.

```
glm(formula = sqrt(tau) ~ sqrt(abs(logFC.XYtestis.XXovary)) *
    bias, family = binomial, data = gonad_tau_sub)
```

```
Deviance Residuals:
    Min        1Q    Median        3Q       Max
-0.94063  -0.17834   0.00147   0.19640   0.82437

Coefficients:
                                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                               -0.45131    0.11497  -3.926 8.65e-05 ***
sqrt(abs(logFC.XYtestis.XXovary))          0.91428    0.07619  11.999  < 2e-16 ***
biasmale                                  -0.09787    0.16565  -0.591    0.555
sqrt(abs(logFC.XYtestis.XXovary)):biasmale -0.15882    0.10971  -1.448    0.148
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

###############
sqrt(dNdS) seems to follow normal distribution.
##############
y4 <- lm(sqrt(dNdS) ~ sqrt(tau) * bias, data=gonad_tau_dnds)
summary(y4)

#######
Coefficients:
  Estimate Std. Error t value Pr(>|t|)
(Intercept)           0.18586    0.01392  13.348  < 2e-16 ***
  sqrt(tau)           0.12870    0.01954   6.585 4.94e-11 ***
  biasmale            0.05603    0.01805   3.104  0.00192 **
  biasunbias          0.02260    0.01618   1.397  0.16253
sqrt(tau):biasmale   -0.08342    0.02576  -3.239  0.00121 **
  sqrt(tau):biasunbias -0.03435    0.02356  -1.458  0.14494
---
  Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
##################

After remove unbiased genes, results remain the same.
##################
y5 <- lm(sqrt(dNdS) ~ sqrt(tau) * bias, gonad_tau_dnds_sub)
summary(y5)
#####
Coefficients:
  Estimate Std. Error t value Pr(>|t|)
(Intercept)         0.18586    0.01388  13.392  < 2e-16 ***
  sqrt(tau)         0.12870    0.01948   6.607 4.55e-11 ***
  biasmale          0.05603    0.01799   3.114  0.00186 **
  sqrt(tau):biasmale -0.08342    0.02567  -3.250  0.00117 **
  ---
  Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#####
```

At stage G46, the coding sequence evolutionary rate can be explained by tissue specificity, but not sex bias. It is interesting to see that here sex bias does not significantly explain the difference in evolutionary rate of coding sequence. I think it will be quite interesting to see the patterns in other tissues too. This is different from pattern in G46.

More analysis will come along the way, now it is meant to get some idea on the patterns and trigger further related questions.

B. Sex-biased gene expression along developmental stages, across adult tissues

Approach: We cleaned the transcriptome by BLAST the transcriptome with the new XX genome, and removed the transcripts which were not found in the XX genome. This transcriptome with removed-non-frog-transcripts is used, we quantify the transcript abundance with Kallisto.

EdgeR is applied to analyze differential or sex-biased gene expression across developmental stages, as well as adult tissues.
The selection criteria to remove low expressed transcripts are:

1) For each transcript, we select the average LogCPM above 0;
2) Additionally, we select LogCPM >1 is present in at least half of the tissues per sex.

Figure 6. Number of sex-biased genes throughout development and adult tissues (FDR < 0.05, Log2 >=1).

Sex-biased genes increase dramatically throughout development, and show stronger difference in adult gonadal tissues, but not in somatic tissues.

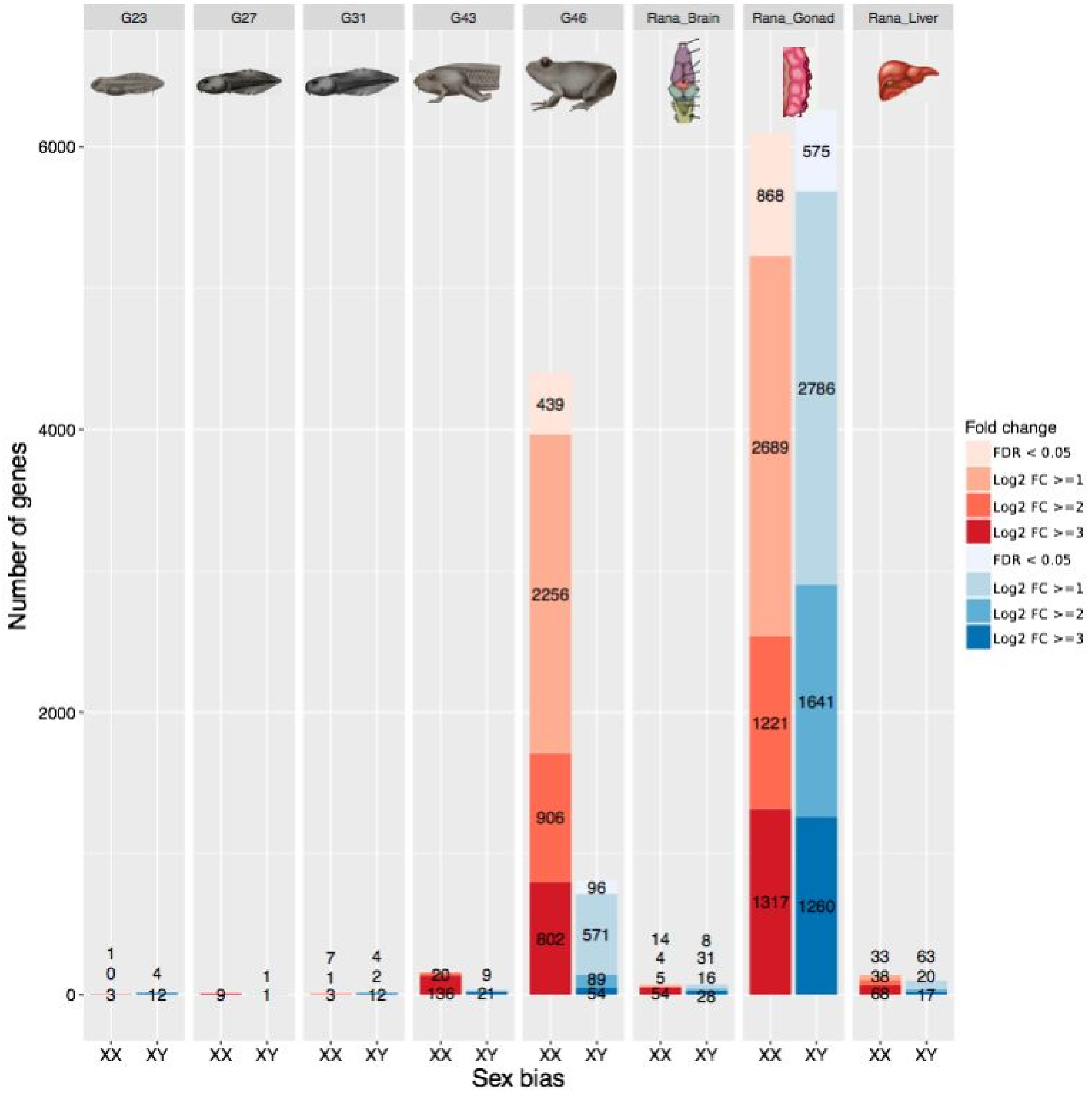


Figure 7. Venn diagram on sex-biased genes throughout development. (a) Shared sex-biased genes, (b) shared female-biased genes and (c) shared male-biased genes throughout development stages.

There is none shared sex-biased genes across five developmental stages, but there are certain number of shared in adjacent stages.

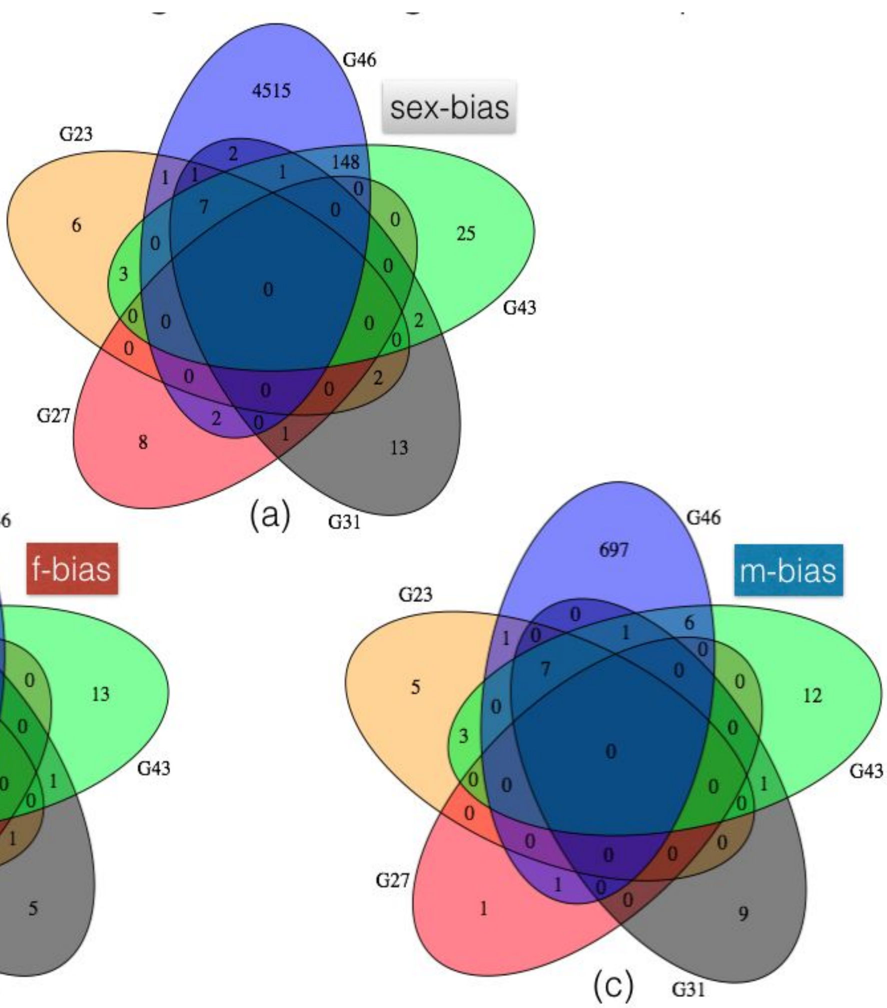


Figure 8. Venn diagram on sex-biased genes among adult tissues. (a) Shared sex-biased genes, (b) shared female-biased genes and (c) shared male-biased genes among three adult tissues, brain, gonad and liver.

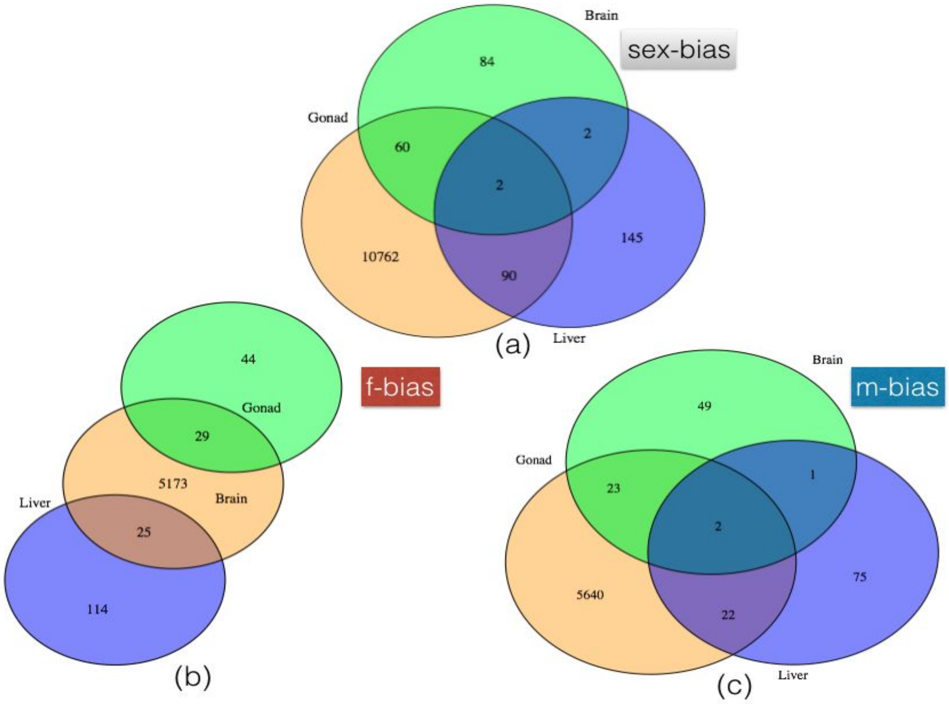


Figure 9. Venn diagram on sex-biased genes between G46 and gonad tissues. (a) Shared sex-biased genes, (b) shared female-biased genes and (c) shared male-biased genes between G43 and gonad tissues which have high sex-biased genes.

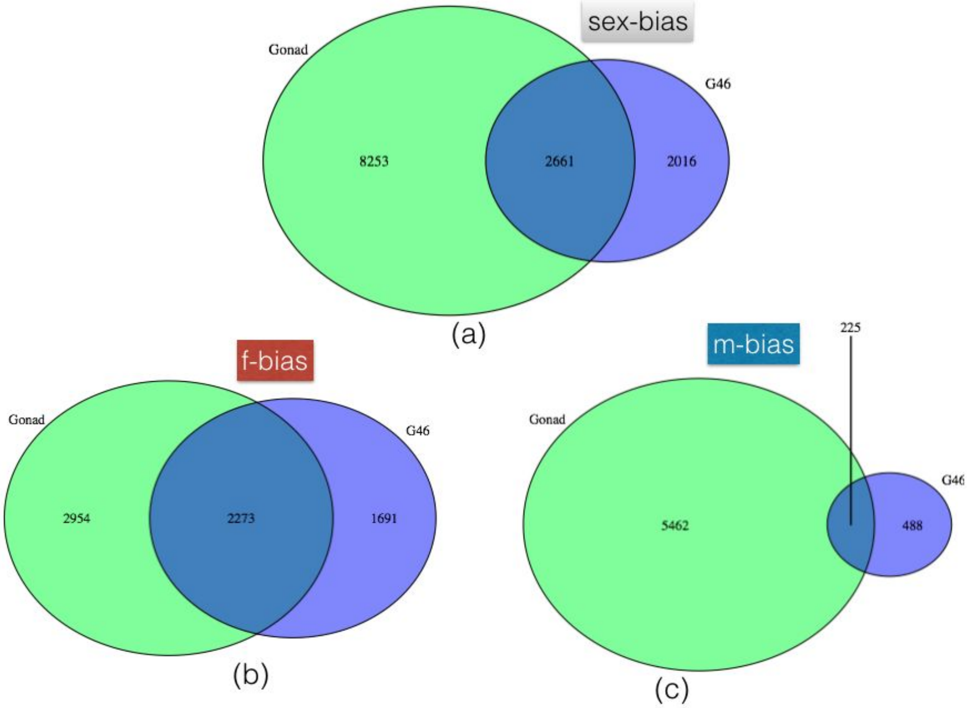
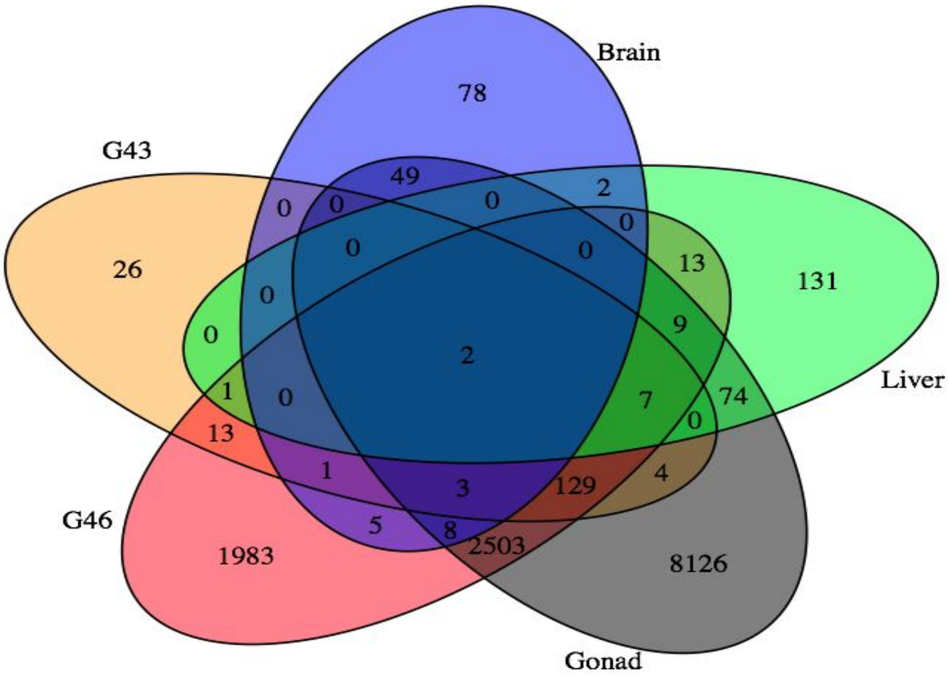


Figure 10. Shared sex-biased genes between tissues of G43, G46, and adult gonad, liver and brain tissues.



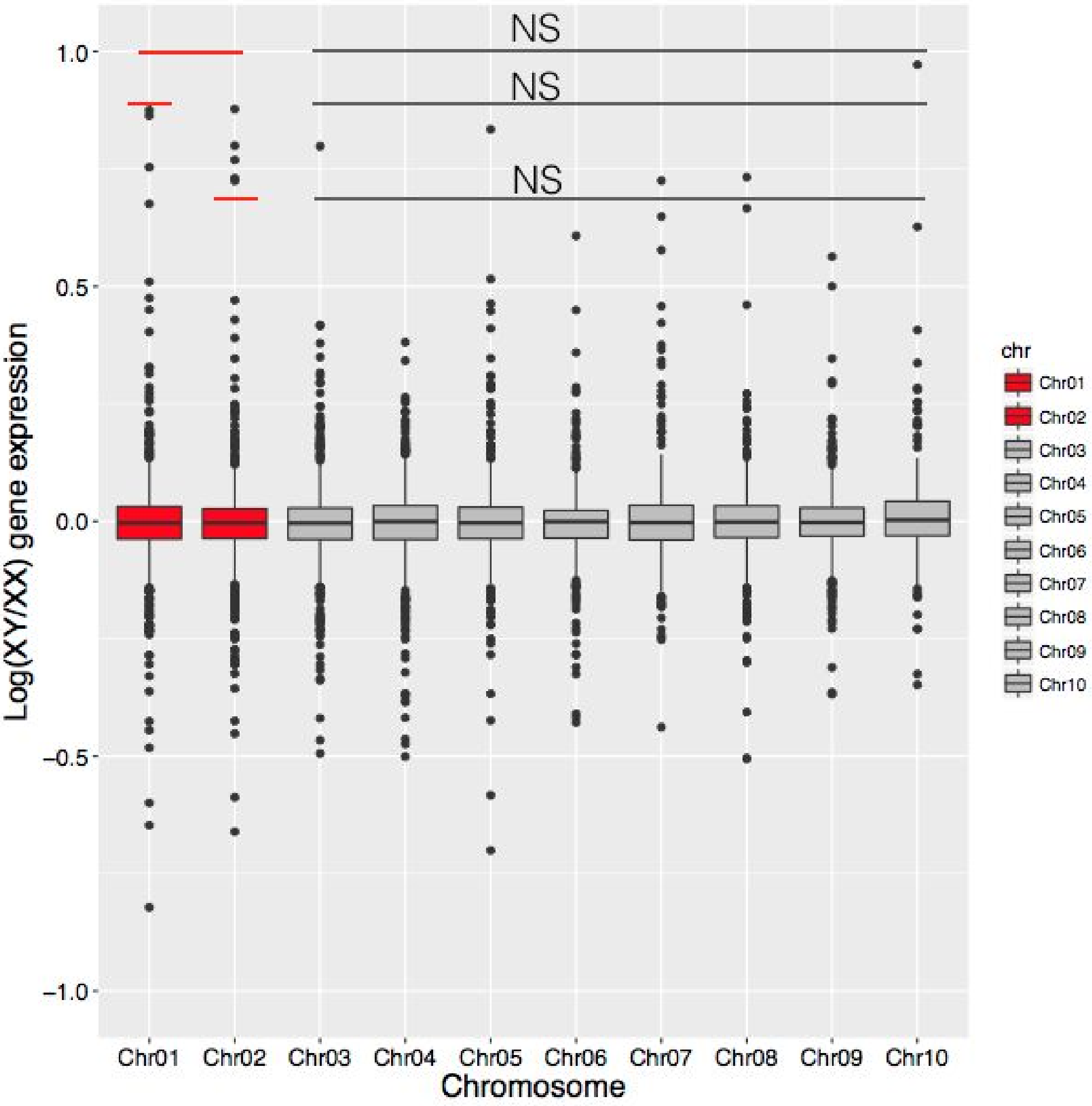
It is interesting to know the proportion of sex-biased genes in certain comparisons, will add that later on.

C. Transcription degeneration

To investigate possible transcriptional degeneration on the Y chromosome at early sex chromosome evolution, we compare gene expression ratio of XY individuals with testis and XX individuals with testis, Log2(XY/XX).

Figure 10. Boxplot of gene expression ratio between individuals of XY with testis and XX with testis. Note, chromosome 1 and 2 are both shown to have male specific haplotypes.

No significant differences between gene expression ratio on chr 1 and autosomes, or chr 2 and autosomes, or ch1+chr2 and autosomes.



**In total, there are 134 differentially expressed genes between XY males and XX with testis. Out of the 134 transcripts, 11 out of 33 orthologs are identified on sex chromosomes (chr1 + chr2), chisq.test shows it is not significant.

D. Coding sequence divergence and Faster-X

Approach: find the longest ORFs per transcript, use 1:1 ortholog with *X.tropicalis* to locate the genome locations for each transcript, finally use PLINK to perform coden alignment and finally calculate dN, dS, dN/dS with the Codeml model in PAML.

Figure 11. dN/dS ratio of sex-biased and unbiased genes throughout development and three adult tissues. The sample number are included in each category.
M= male-bias, F= female-bias, U= unbias, Signif. codes: 0 ‘***’, 0.001 ‘**’, 0.01 ‘*’, 0.05 ‘.’.

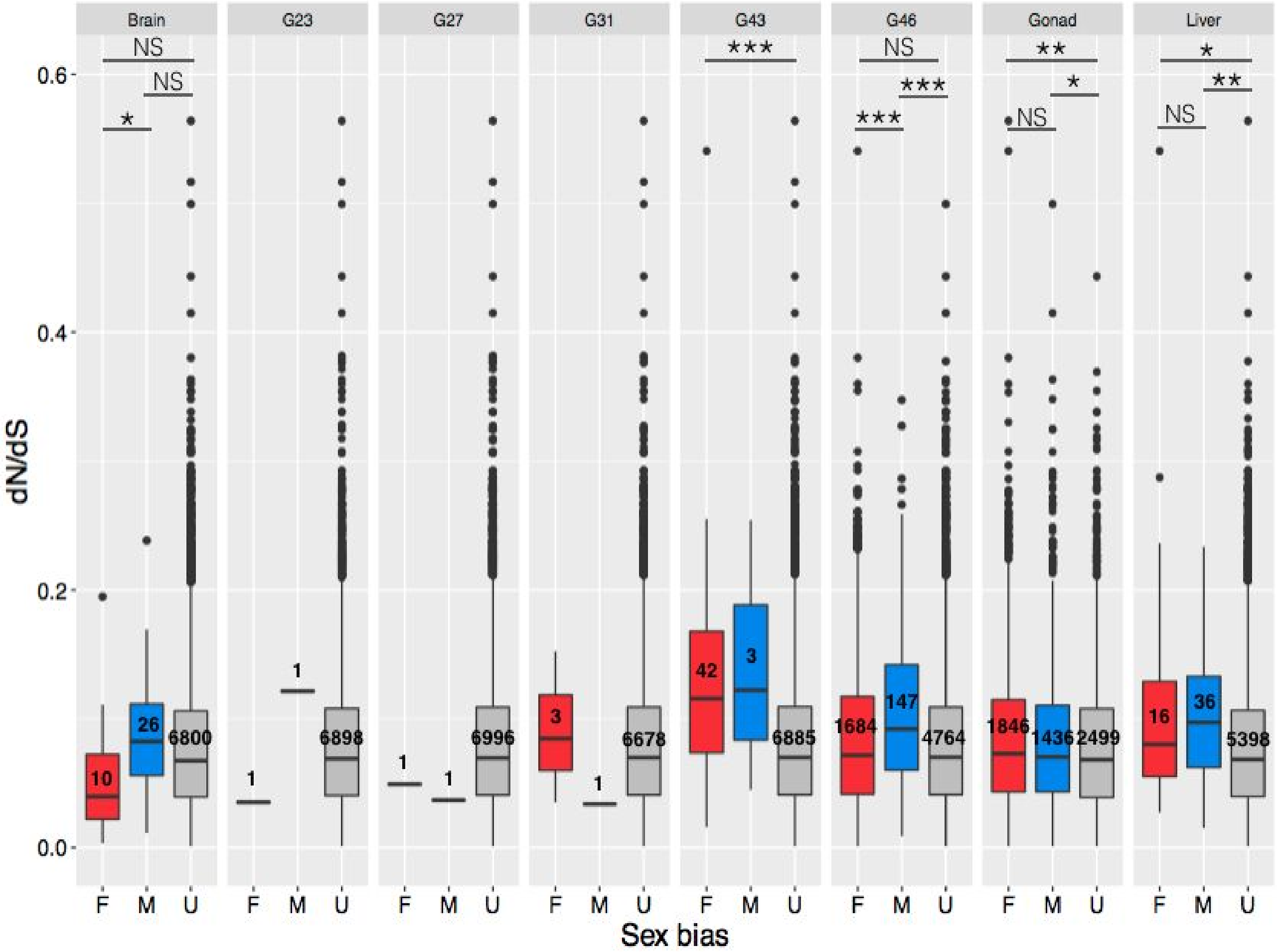
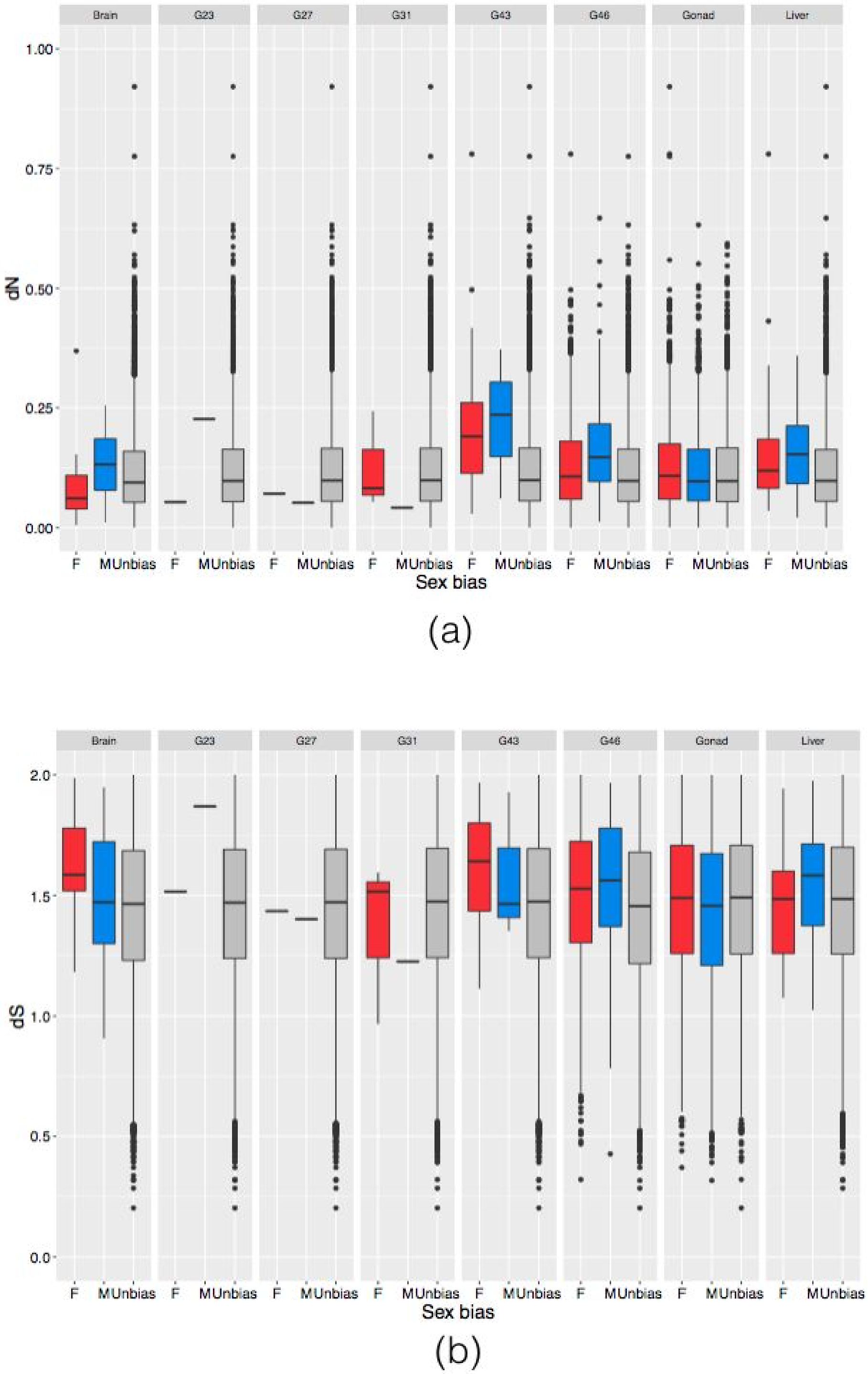


Figure 12. dN (a) and dS (b) values of sex-biased and unbiased genes throughout development and three adult tissues.
This is for supplementary materials i think, need to add the significance levels.



(a)

(b)

Figure 12. dN/dS ratio of genes on sex chromosomes (chromosome 1 and chromosome 2) and autosomes.

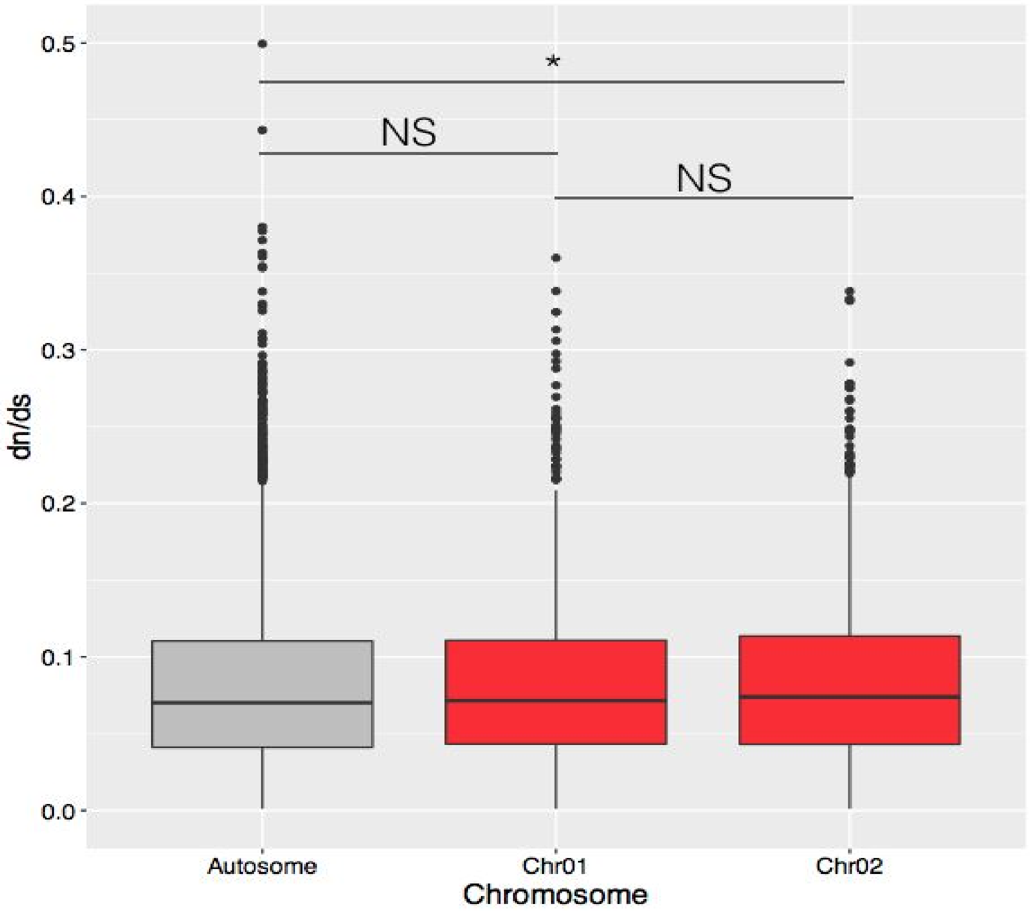
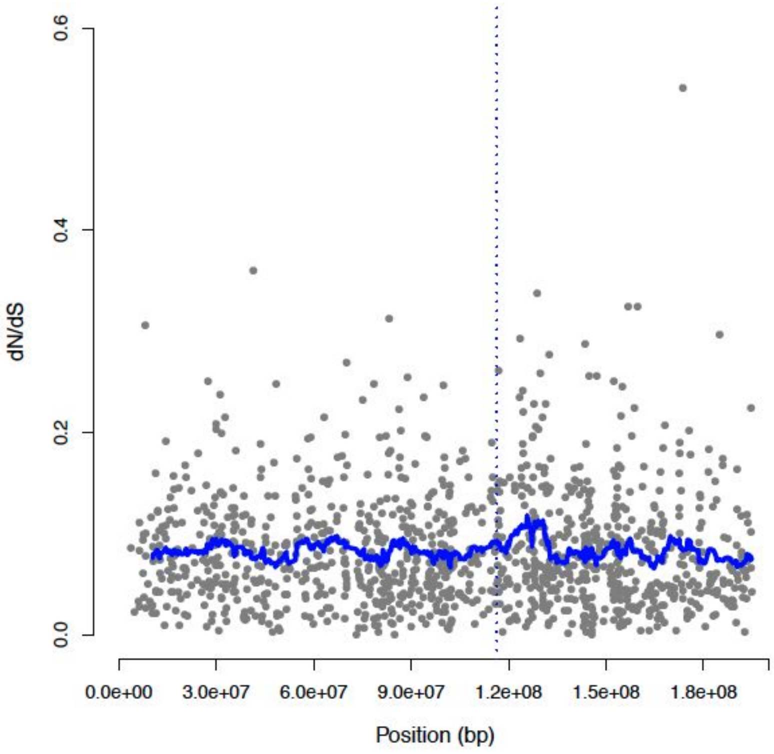
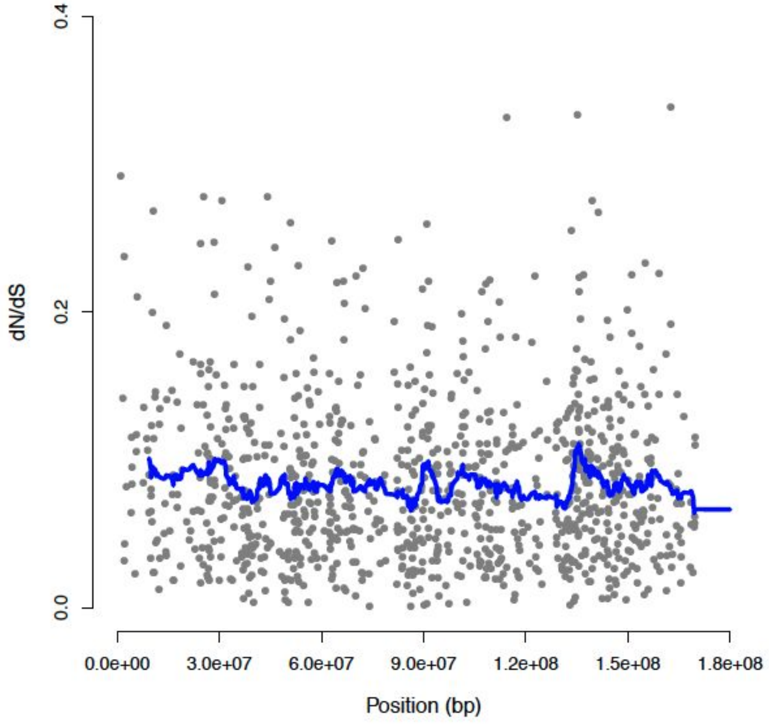


Figure 13. dN/dS ratio of genes along sex chromosomes, chr1 (a) and chr2 (b), with a sliding window of 40 genes.


0.6
0.4
0.2
0.0
dN/dS
0.0e+00
3.0e+07
6.0e+07
9.0e+07
1.2e+08
1.5e+08
1.8e+08
Position (bp)

(a)


0.4
0.2
0.0
dN/dS
0.0e+00
3.0e+07
6.0e+07
9.0e+07
1.2e+08
1.5e+08
1.8e+08
Position (bp)

(b)

Here we update the results with transcriptome without non-frog transcripts, we still need to do get XX and XY transcriptome, by replacing SNPs of current transcriptome by mapping XX individual and XY individual respectively.