Making hypotheses precise: applying the physical method and linear modeling to evolutionary genomics

by

David W. Morris

A thesis submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Master of Science

Auburn, Alabama
May 4, 2014

Approved by

Geoffrey E. Hill, Chair, Curator of Birds, Professor of Biological Sciences
Kenneth M. Halanych, Schneller Endowed Chair, Professor of Biological Sciences
Mark R. Liles, Associate Professor of Biological Sciences

Abstract

Hypotheses in evolutionary genomics are commonly tested by qualitative comparisons among variables. However, the number of possible outcomes for such predictions is usually few, which can potentially trivialize or even mislead conclusions. Therefore, I compared the utility of two alternative approaches to testing evolutionary genomic hypotheses. The first is the "physical method", the quantitative approach often used in physical sciences, which I employed to mathematically model the process of gene movement between chromosomes. Using this model, I rejected the hypothesis that sexual conflict has caused biased movement of oxidative phosphorylation (OXPHOS) genes off the X chromosome in mammals. In contrast, the qualitative approach of testing for an under-representation of OXPHOS genes on the X did not yield a clear conclusion. The second alternative is to estimate effect sizes of explanatory variables using linear modelling. I tested the utility of this approach by assessing a prior study in which p-values alone were used to conclude that gene expression level is the only important factor affecting evolutionary constraint on OXPHOS genes in animals. I show that, by considering how effect sizes are estimated, it is clear that at least one additional factor must affect the evolutionary constraint on OXPHOS genes, or that the original data violated at least one assumption of linear modeling. Together, these results suggest quantitative approaches offer a more precise description of hypotheses in evolutionary genomics, which allows for more robust conclusions.

Acknowledgments

I would like to thank my advisor Geoff Hill for his help in all aspects of my thesis and for encouraging me to think independently and creatively; my committee members, Ken Halanych and Mark Liles, for their willingness to help and their encouragement; Jim Johnson for countless thought-provoking discussions; Todd Steury for teaching me how statistics can be used and misused in biology; and the Hill-Hood-Wada lab groups for valuable feedback and support.

Table of Contents

List of Tables

List of Figures

Chapter 1

Chapter 2

List of Abbreviations

EEB	Ecology and Evolutionary Biology
N-mt	Nuclear genes whose proteins are expressed in the mitochondria
nuOXPHOS	Nuclear-encoded genes whose proteins are part of the oxidative phosphorylation complexes
mtOXPHOS	Mitochondrial-encoded genes whose proteins are part of the oxidative phosphorylation complexes
NEW	Study by Nabholz, Ellegren, and Wolfe that is the focus of Chapter 2
d_N/d_S	The ratio of the number of non-synonymous substitutions per non-synonymous site to the number of synonymous substitutions per synonymous site
E	Gene expression level
G	Genome type, either mitochondrial or nuclear
U	The hypothesized unknown factor that affects d_N/d_S

Chapter 1: Assessing the utility of a physics-based quantitative approach to the study of evolutionary genomics: a case study on the movement oxidative phosphorylation genes between chromosomes

Introduction

The use of mathematics and statistics in the study of biological systems is now commonplace, but the approaches to modeling and data analysis differ fundamentally between ecology and evolutionary biology (EEB) and more traditionally quantitative fields like physics (Murray 2000; Berryman 2003; Knapp et al. 2004; Ginzburg et al. 2007; Colyvan and Ginzburg 2010; Evans et al. 2013). Physics employs general equations to predict a wide range of empirical results, while the basic equations of EEB, such as logistic population growth, only predict empirical data in limited cases (Hansson 2003; O'Hara 2005; Ginzburg et al. 2007). As a result, testable predictions are not typically derived from basic equations in EEB as they are in physical sciences; rather, predictions tend to emerge from verbal statements of hypotheses. Consequently, most studies in EEB test hypothesis by comparing qualitative relationships among variables such as means or correlational associations.

In contrast, a quantitative analysis is more robust than a qualitative one because a much larger set of outcomes are possible. For instance, when means are compared qualitatively the first mean is either larger, the same, or smaller than the second. With such constrained predictions, many hypotheses can predict the same qualitative relationship and outcomes become ambiguous. Alternatively, in a quantitative analysis predictions are not restricted to only the direction of difference, but also specify the magnitude of the difference in means. Thus, it is much less likely that two hypotheses will yield the same quantitative prediction. The same arguments apply to

correlational predictions as well. I therefore argue it is desirable to derive and test quantitative instead of qualitative predictions in EEB whenever possible.

To assess the utility of the quantitative approach to the study of EEB systems, I analyzed the process of gene movement between chromosomes. Movement of single genes between chromosomes is relevant to several subfields in evolutionary biology. Gene movement can be an important component of genome reorganization between species (Bhutkar et al. 2007) and has been hypothesized to play a role in speciation (Moyle et al. 2010). Moreover, studies of the patterns of gene movement between the sex chromosome and the autosomes have shed light on the evolution of sexual conflict and sex chromosome meiotic inactivation (Parisi et al. 2003; Bhutkar et al. 2007; Vibranovski et al. 2009; Gallach et al. 2010; Zhang et al. 2010; Ellegren 2011), in addition to the evolutionary origins of the mammalian X chromosome (Potrzebowski et al. 2008; Kaessmann et al. 2009).

In particular, I applied a quantitative model to the question of whether nuclear genes that express in the mitochondria (N-mt genes) exhibit biased gene movement off the X chromosome in mammals. Based on qualitative models, Gallach et al. (2010) and Drown et al. (2012) concluded that such genes are under-represented on X chromosomes in mammals. They interpreted the movement of N-mt genes off of the X to be the result of selection to resolve sexual conflict. Briefly, sexual conflict occurs when a gene product has beneficial effects in one sex but deleterious effects in the other (Vicoso and Charlesworth 2006). Drown et al. hypothesize that since the mitochondrial genome is inherited through the female line almost 100% of the time, mitochondrial genotypes that increase female fitness can become fixed even if they decrease male fitness. If such a fixation occurs, there will be selective pressure on N-mt genes to alleviate the sexual conflict. In mammals, mutations that alleviate the sexual conflict are

more likely to fix on autosomes (present in males 50% of the time) than the X (only 33% in males) because they will experience positive selection in males more often. Therefore, movement of N-mt genes off the X is expected to occur more frequently than movements between other chromosomes and the number of N-mt genes on the X is predicted to be lower than the random expectation. In birds, Drown et al. claim there is little or no selection to move N-mt genes on or off the Z chromosome due to sexual conflict and thus predict no biased representation of N-mt genes in birds. These two predictions are examples of qualitative predictions common in EEB and especially in studies that analyze chromosomal distribution of genes.

While Drown et al. conclude that their observation of significant under-representation of N-mt genes on the X across mammals is in good agreement with their scenario of sexual conflict, some ambiguities exist in their analysis that cast uncertainty on the validity of this conclusion. First, they assume that if selection from sexual conflict is absent, N-mt genes should be distributed randomly on the X. However, N-mt genes are clearly not distributed randomly among the autosomes (Drown et al. 2012, their Figure 1) and their description of sexual conflict makes no prediction of such variation. Therefore, the null hypothesis of a random distribution on the X may not be justified. Second, they justify their assumption that all mammalian lineages are approximately independent by arguing that genomes are very fluid, with rearrangements of genes and chromosomes occurring often. They don't, however, provide any evidence that rearrangements of N-mt genes occur often enough to justify this assumption. Moreover, if the X chromosome just happened to evolve from an autosome that was under-represented in N-mt genes, Drown et al. assume constant selection is required to keep N-mt genes from increasing to the random expectation, although they provide no justification for this.

I ask whether a quantitative approach can yield insight into the ambiguities present in the Drown et al. analysis, thereby allowing more confidence in whatever conclusion the data support. I test the utility of the quantitative approach using oxidative phosphorylation (OXPHOS) genes in mammals and birds. Nuclear-encoded OXPHOS genes (nuOXPHOS) are a subset of all N-mt genes. There are five recognized OXPHOS protein complexes (commonly denoted, Complexes I – V) that participate in electron transfer in the mitochondria and all except Complex II are composed of proteins encoded by both the nuclear and mitochondrial genomes (Bar-Yaacov et al. 2012). Due to this close physical association of nuOXPHOS with mitochondrial-encoded OXPHOS (mtOXPHOS) proteins, it is expected that nuOXPHOS genes will experience the same potential for sexual conflict and thus the same selection for biased movement off of the X if Drown et al.'s hypothesis is correct.

Methods

Calculation of the distribution of nuOXPHOS chromosomal locations

I conducted a comparison similar to the one of Drown et al. I included the same 14 mammal and 2 bird species, plus an additional mammal (*Felis catus*) and bird (*Meleagris gallopavo*) species because their genomes and chromosome maps were available. I also substituted *Pongo abelii* for *Pongo pygmaeus* because genomic data from Ensembl was only available for the former species. The general strategy I employed to test for biased numbers of nuOXPHOS genes on the sex chromosome was to BLAST the genomes in search of orthologs of the well-characterized human nuOXPHOS genes, so that a set of nuOXPHOS could be determined for each species. Orthology determinations were made according to reciprocal BLAST criteria. Briefly, for each human nuOXPHOS gene, only the top BLAST hit was saved

and only if its E-value was less than 10^{-4}. This top hit was then used for a BLASTn against the human genome. If the corresponding top hit in the human genome was the same sequence as the original query, then these two genes were deemed orthologous. This was done for each species in the comparison (Table 1). Because many nuOXPHOS genes are paralogs of each other, reciprocal BLAST struggled to assign orthologs to several genes. In an effort to include more genes in the calculations, additional criteria were used for human genes that failed the reciprocal BLAST. First, if all of the hits generated for a human gene were on the same type of chromosome (i.e. autosomes or sex chromosome), then no matter which sequence was assigned as the ortholog, the chromosome type is the same. In these cases the lowest E-value was assigned as the ortholog; if that hit was already assigned, then the hit with the second lowest E-value was assigned, and so on. If all the hits were already assigned as orthologs, then no orthology assignment was made for that human gene and it was not included in any calculations. Second, if a set of human paralogs each hit the same set of genomic sequences, and the number of human paralogs in that set matched the number of sequences hit, then each human paralog was assigned its lowest E-value hit available. For example, the COX7B and COX7B2 genes failed the reciprocal BLAST in several species. In most of these cases, both genes hit the same two sequences in the genome under search, with one of those genomic sequences located on the X and the other on an autosome. In this case, there is high certainty that one of the orthologs is on the X and the other is on the autosome.

Once orthologous sequences were determined, the number of nuOXPHOS genes on the sex chromosome in each species was counted. The random expectation was calculated by multiplying the total number of nuOXPHOS genes, n, by the proportion, p, of genes in the genome that are located on the sex chromosome. This proportion was estimated using genome

data from Ensemble (ftp://ftp.ensembl.org/pub/release-75/fasta/). To have higher confidence in the estimate of p, an average was taken from three different methods of estimating the number of genes on each chromosome: (1) genes that were considered "known" by Ensembl, (2) gene referred to as "novel" by Ensemble, and (3) genes predicted from *ab initio* algorithms. The number of observed genes on the sex chromosome was then divided the by the random expectation to get the observed/expected (O/E) ratio. A $Bin(n,p)$ distribution was used to test whether the observed number of genes was significantly greater or less than the expected value.

Because the number of nuOXPHOS genes observed and expected to be on the sex chromosome was ~ 2 – 4 in all taxa, there is lower power to detect an O/E ratio as significantly different from one. Thus, an alternative statistical analysis is to deem all species as independent samples, score each O/E as "higher" or "lower" than one, and treat this score as a binomial random variable with $p = 0.5$. The intuitive logic behind this analysis is that, if there is no selection to cause a biased chromosome distribution, we would expect the O/E ratio to be less than and greater than one, each about half of the time. Suppose selection for biased movement off the X does exist; then we would expect more than half the lineages to be scored "lower" than one.

The quantitative model of gene movement

The motivation for developing this model is from analogy with modeling approaches shown to successfully quantitatively predict data. First, I argue that a common strategy in physical sciences is to describe the "interactions" that occur between "components" of a system. For example, Newton's second law describes the forces (the interactions) that occur between idealized "bodies" (the components). Quantitative predictions for a specific system (i.e. the

topology of how the components exert forces on each other and the nature of those forces) can be derived by solving a differential equation. Likewise, I argue the Schrödinger equation describes the energy interactions between particles (the components) and a specific system can be specified by defining the potential. Second, a successful strategy for modelling the enumeration of objects (e.g. concentration of biomolecules) is to postulate a differential equation that gives the rate of change of the number of objects as a function of the rate of increase minus the rate of decrease (e.g. Alon 2006), and then solving for the number of molecules as a function of time. For example, Alon (2006) shows how modeling the rate of change of concentration of a biomolecule in this manner can be used to predict the functions of various types of molecular biological networks.

Using the above two approaches as a guide, I attempted to describe the gene movement process mathematically. Gene movement can be thought of as occurring via two separate events: (1) gene "selection" for movement (e.g. excision during transposition (Li 1997), reverse transcription of mRNA (Vibranovski et al. 2009)), and (2) insertion of the copy to a new chromosome. There is empirical evidence from the human and mouse genomes that the number of parent genes (i.e. those genes that have duplicate copies somewhere in the genome) on a chromosome is directly proportional to the total number of genes on the chromosome (Emerson et al. 2004). In addition, the number of copied genes on a chromosome is directly proportional to the size of the chromosome in base pairs (Emerson et al. 2004). This data suggests that the rate of gene movement onto a particular chromosome is proportional to the product of the number of base pairs it has and the number of genes on the other chromosomes in the genome. Likewise, the rate of movement off a particular chromosome would be proportional to the product of the

number of genes it has and the number of base pairs on the other chromosomes. I therefore postulate the following equation to describe gene movement:

$$\frac{dn_i}{dt} = L_i \sum \gamma_j n_j - n_i \sum \gamma_j L_j \quad (1),$$

where the i subscript indicates the chromosome of interest, the j subscript indicates any other chromosome in the system, and the summation is assumed to be taken over all chromosomes except chromosome i. L is the "length" of the chromosome in base pairs, n is the number of genes on the chromosome, γ is a rate coefficient, given in units of $\frac{1}{time*base\ pairs}$, needed to make the units on the right and left-hand sides the same, and $\frac{dn_i}{dt}$ is the instantaneous rate of change of the number of genes on chromosome i. Note that equation (1) assumes a continuous, deterministic process of genes moving between chromosomes and that movement is always accompanied by loss of the parental gene. In reality, gene movements are discrete, occur in a probabilistic fashion, and parent genes often remain in the genome. However, since this is a first analysis of the gene movement process, I argue these simplifications are warranted.

Deriving predictions of the nuOXPHOS system evolving in sexual conflict

I now use equation (1) to specify a genomic system evolving in sexual conflict according to Drown et al. (2012). Since their hypothesis predicts biased gene movement off of the X, the chromosome of interest will be the X. However, no biased movement among autosomes, from the autosomes to the X, or from the X to any particular autosome relative others is predicted. This allows the rate coefficients for each chromosome, γ_i, to be treated as equal. This single γ can now be brought outside the summations. In order to mathematically account for biased movement off the X, I introduce a coefficient, c, which is multiplied by γ in the term describing

movement off of the X. The resulting system of genes evolving in sexual conflict takes a modified form of equation (1):

$$\frac{dn_X}{dt} = \gamma L_X \sum n_j - c\gamma n_X \sum L_j \quad (2).$$

By assuming that all movements of duplicate copies are accompanied by loss of the parent gene, the total number genes in the system, n, is constant. In this analysis, since I am investigating nuOXPHOS genes, n represents the total number of nuOXPHOS genes in the genome and n_X represents the number of nuOXPHOS on the X. It must be true that $n = n_X + \sum n_j$ and $L_G = L_X + \sum L_j$. Solving for $\sum n_j$ and $\sum L_j$ and substituting into equation (2), along with some algebraic manipulation, gives

$$\frac{dn_X}{dt} = \gamma L_X n - D n_X \quad (3),$$

where $D = \gamma(1-c)L_X + c\gamma L_G$ is defined for convenience of notation. Equation (3) is a separable differential equation whose solution is

$$n_X(t) = \frac{np_X}{p_X + (1-p_X)c}(1 - e^{-Dt}) + n_0^X e^{-Dt} \quad (4),$$

where $p_X = L_X/L_G$, the proportion of total base pairs in the genome that are on the X , and n_0^X is the number of nuOXPHOS genes on X at $t = 0$. Equation (4) is the prediction of how many genes should be on chromosome X as a function of time according to sexual conflict. It is interesting to note that

$$\lim_{t\to\infty} n_X(t) = \frac{np_X}{p_X + (1-p_X)c} \equiv n_{st}^X \quad (5),$$

which shows that the number of genes on the X reaches a constant value as time becomes large. For convenience of notation, I denote this value n_{st}^X, where the "st" is short for "steady-state".

Finally, I define the quantity, R_X, as the rate of gene movements involving the X chromosome, that is, the sum of movements on and off the X. This means that R_X is equal to equation (2) with the exception that the second expression is added rather than subtracted. Some algebraic manipulation yields,

$$R_X(t) = \gamma L_X n - D_R n_X(t) \quad (6),$$

where $D_R = \gamma(1 + c)L_X - c\gamma L_G$. The number of movements involving the X in a given time, $M_X(t)$, is found by substituting equation (4) into equation (6) and noting that $M_X(t) = \int R_X(t)dt$ and $M_X(0) = 0$; the result is

$$M_X(t) = (\gamma L_X n - D_R n_{st}^X)t + \left(\frac{D_R}{D}\right)(n_{st}^X - n_0^X)(1 - e^{-Dt}) \quad (7).$$

$M_X(t)$ is a useful quantity because it is easily estimated from genomic data by counting chromosome location differences between pairs of species across a phylogeny.

Results

Chromosomal distribution of nuOXPHOS genes

The comparison of the observed number of nuOXPHOS genes on the sex chromosome versus the random expectation showed the number of nuOXPHOS genes on the sex chromosome was not significantly different from random in any species examined (Table 1). However, due to the low number of expected nuOXPHOS genes on the sex chromosome, there is low power to detect a statistically significant difference. I therefore assumed each species to be an independent binomial trial, exhibiting either "greater" or "less" than the expected number of nuOXPHOS genes. 10 of the 15 mammal species had less nuOXPHOS than expected, but this trend was not significant (binomial test, p = 0.30). For the birds, 2 species had less and 1 more nuOXPHOS

genes than the random expectation; however, this is not a large enough sample to make a statistically meaningful comparison.

Testing the assumption that each species is independent

Drown et al. treat each lineage in their comparison as independent because they assume genomic rearrangements occur frequently enough such that selection is required to keep N-mt genes under-represented on the X. Thus, their assumption predicts that if selection due to sexual conflict is relaxed, the number of genes on the X will increase to the random expectation. My quantitative model of gene movement can be used to test the validity of this assumption. For example, consider the human-chimpanzee ancestor and for simplicity of calculations assume it was in steady-state at the time of their divergence. The number of nuOXPHOS genes on the X is given by equation (5); using reasonable values of parameters, $n = 80, p_X = 0.05, \& \, c = 2$, the theoretical prediction is $n_{st}^X = 2.05$ nuOXPHOS genes. Now assume sexual conflict continues in the chimpanzee but disappears in the human lineage. The new prediction for steady-state in human, now with $c = 1$, is $n_{st}^X = 4$ nuOXPHOS genes. It is clear from Figure 1 that the human X will move toward this new steady-state, which is analogous to the random expectation from the qualitative analysis.

However, until the new steady-state is reached, the human lineage will still be under-represented on the X, despite a lack of selection from sexual conflict. Thus, it follows to ask how long it will take for the human lineage to reach its new steady-state. If this time is longer than the time since divergence, the number of genes on the human and chimpanzee X chromosomes are not independent, as Drown et al. assume. In general, if a lineage has been diverged from all other lineages in the comparison for longer than the time it takes to reach steady-state, it can be

considered independent from all other lineages. Mathematically, steady-state is an asymptote that is never reached exactly. Therefore, an approximate time to steady-state, denoted t_{st}, must be calculated instead. Suppose we want to know the time it takes for the X chromosome to reach a certain proportion, β, of the way to steady-state. If $n_{st}^X > n_0^X$, then $n_X(t_{st}) = \beta(n_{st}^X - n_0^X) + n_0^X$. If $n_{st}^X < n_0^X$, then $n_X(t_{st}) = n_0^X - \beta(n_0^X - n_{st}^X)$. These equations can then be substituted into equation (4) and the time to approximate steady-state for the X, t_{st}^X, can be solved. It turns out that t_{st}^X is the same regardless of the values of n_{st}^X and n_0^X:

$$t_{st}^X = \frac{\ln\left(\frac{1}{1-\beta}\right)}{D} \quad (8).$$

A reasonable value for β might be 0.95 (i.e. the system has moved 95% of the way to steady-state). In this case the numerator is very close to 3. Note that if β is chosen to be anywhere between 0.900 and 0.999, the numerator of equation (8) only varies by a factor of three (~ 2.3 – 6.9). Thus I use the following approximation:

$$t_{st}^X \approx \frac{3}{D} \quad (9).$$

Ideally, t_{st}^X would be compared to divergence times estimated from a phylogeny; however, D contains two parameters, γ and c, that are not easily estimated from data. Therefore, additional analysis is needed to test whether a lineage is in steady-state.

Number of gene movements involving the X chromosome

Although equation (9) cannot be directly test the independence of lineages with respect to nuOXPHOS genes, it can be used, along with equation (7) to gain insight into this question. The quantity, $M_X(t_{st})$, is defined as the number of gene movements involving the X chromosome

occurring in the time it takes to reach steady-state. It can be derived by substituting equation (9) into equation (7), along with manipulating equation (5) into the form $D = \frac{\gamma L_X n}{n_{st}^X}$, to yield:

$$M_X(t_{st}) = 3n_{st}^X\left(1 - \frac{D_R}{D}\right) + \frac{D_R}{D}(n_{st}^X - n_0^X)(1 - e^{-3}) \quad (10),$$

where $\frac{D_R}{D} = \frac{1-c\left(\frac{L_G}{L_X}-1\right)}{1+c\left(\frac{L_G}{L_X}-1\right)}$. Equation (10) gives the number of movements predicted to occur, per lineage, if it is independent. Thus, in the context of my model, Drown et al.'s assumption is that at least $M_X(t_{st})$ movements have occurred, on average, in each mammalian lineage.

Since γ and c (part of the D_R and D terms) cannot be easily estimated, a precise estimate of equation (10) cannot be obtained either. However, a reasonable order of magnitude estimate of $M_X(t_{st})$ is possible without any empirical analysis. For most mammalian genomes $\frac{L_G}{L_X} \approx 20$; using this value and $c = 1$, I find $\frac{D_R}{D} = -0.9$. Since c represents the bias for genes to move off of the X chromosome due to sexual conflict, c is expected to be greater than 1 (if c were equal to 1, there would be no bias because the coefficient γ is multiplied by c). As c is increased above 1, the value of $\frac{D_R}{D}$ moves closer to its limit, -1. Thus it is reasonable to approximate $\frac{D_R}{D} = -1$. For nuOXPHOS genes in mammals, n_{st}^X ranges from 2 to 5 (Table 1) and a conservative range of $n_{st}^X - n_0^X$ is estimated to be ~ -5 to 5. By plugging these values into equation (10), I find $M_X(t_{st}) \sim 10 - 35$. Thus, the number of gene movements involving the X, per lineage, is expected to be at least 10, given Drown et al.'s hypothesis of sexual conflict and steady-state.

Maximum number of observable movements involving the X chromosome

Before proceeding to empirically estimate the number of gene movements involving the X and comparing that value to the order of magnitude prediction, it is important to note that a

gene can only take two meaningful states with respect to the M_X parameter; it can either be (1) on the X, or (2) on the autosomes. Therefore, there is the potential for a phenomenon, similar to "multiple hits" in DNA base pair substitutions (see Nei and Kumar 2000), to occur. The maximum number of gene movements that can be observed, denoted M_{Max}, in one lineage is $n_0^X + n_{st}^X$; this occurs when all of the genes initially on the X are different from all the genes on the X at the present time. The ratio $\frac{M_X(t_{st})}{M_{Max}}$ must be < 1 in order to test the order of magnitude estimate of equation (10) empirically. If equation (10) is divided by $n_0^X + n_{st}^X$ and equation (5) manipulated to the form $n_{st}^X = \frac{np_X\gamma L_G}{D}$, it can be shown that

$$\frac{M_X(t_{st})}{M_{Max}} \approx 6\left(\frac{1}{1 + n_0^X/n_{st}^X}\right) - \left(\frac{1 - n_0^X/n_{st}^X}{1 + n_0^X/n_{st}^X}\right)(0.95) \qquad (11).$$

The form of equation (11) is useful because it is effectively in terms of a single variable, n_0^X/n_{st}^X. The desirable condition is

$$6\left(\frac{1}{1 + n_0^X/n_{st}^X}\right) - \left(\frac{1 - n_0^X/n_{st}^X}{1 + n_0^X/n_{st}^X}\right)(0.95) < 1 \qquad (12),$$

which is solved to yield $n_0^X/n_{st}^X > 81$. Thus, the initial number of genes on the X needs to be more than 81 times the steady-state number, in order for $\frac{M_X(t_{st})}{M_{Max}} < 1$ to hold. Considering that, for the mammals in Table 1, the range of nuOXPHOS genes on the X is 2 – 5 and the total number of nuOXPHOS genes is never greater than 80, it is highly improbable that this inequality holds for any mammalian lineage. For example, in human and chimpanzee, the requirement for inequality (12) to hold is that there must have been at least 243 nuOXPHOS genes on the X at the time of their divergence. Equation (10) therefore cannot realistically be tested empirically.

The probability that each nuOXPHOS gene is on the X

To get around this inability to observe all of the gene movements involving the X when the system is in steady-state, I consider the probability that each nuOXPHOS gene will be present on the X, denoted P_X. This probability is difficult to compute generally, so I instead describe a simplified approach that gives an approximate equation for P_X. Consider, for example, the human lineage. If zero gene movements have occurred from its beginning to the present, then we have the following probability distribution: n_0^X genes have $P_X = 1$ and $n - n_0^X$ genes have $P_X = 0$. Now suppose 2 movements occur, one autosomal gene moves onto the X and one X gene moves to the autosomes. The new probability distribution is that n_0^X genes have $P_X = \frac{n_0^X - 1}{n_0^X}$ and $n - n_0^X$ genes have $P_X = \frac{1}{n - n_0^X}$. If we continue to assume movements only occur in pairs, one on and one off the X, the probability distribution can be computed for every even number of M. When this is done a pattern emerges, suggesting a general equation for the probability distribution after M movements can be computed. However, finding such an equation proved difficult and I was forced to make another approximation. For the probability, $P_{X,X}$, that a gene initially on the X will be on the X after M gene movements, I derived the following equation:

$$P_{X,X} = \left(\frac{n_X - 1}{n_X}\right)^{M/2} + \sum_{k=1}^{f_X} \left[\left(\frac{1}{n_X(n - n_X)}\right)^k \sum_{g=0}^{\frac{M}{2}-2k} (k+g) \left(\frac{n_X - 1}{n_X}\right)^g \left(\frac{n - n_X - 1}{n - n_X}\right)^{\frac{M}{2}-2k-g} \right] \quad (13),$$

where $f_X = \begin{cases} 0, & M = 0, 2 \\ 1, & M = 4, 6 \\ 2, & M = 8, 10 \\ \dots \end{cases}$ and so on. Similarly, I approximate the probability, $P_{X,A}$, that a gene, initially on the autosomes, will be on the X after M gene movements, by

$$P_{X,A} = \sum_{k=1}^{f_A} \left[\left(\frac{1}{n - n_X} \right)^k \left(\frac{1}{n_X} \right)^{k-1} \sum_{g=0}^{\frac{M}{2}+1-2k} k \left(\frac{n_X - 1}{n_X} \right)^g \left(\frac{n - n_X - 1}{n - n_X} \right)^{\frac{M}{2}+1-2k-g} \right] \quad (14),$$

where $f_X = \begin{cases} 0, & M = 0 \\ 1, & M = 2, 4 \\ 2, & M = 6, 8 \\ \ldots \end{cases}$ and so on. Equations (13) and (14) are approximate in that they give the correct value of the summation terms only when $k = 1$ or $\frac{M}{2} - 2k = 1$ for $P_{X,X}$ and $k = 1$ or $\frac{M}{2} + 1 - 2k = 1$ for $P_{X,A}$. One can see the approximation is very good for $M \leq \sim 10$, because these conditions are satisfied for most or all of the summation terms. But as M becomes large these conditions are satisfied for few of the summation terms and $P_{X,X}$ and $P_{X,A}$ are underestimated. However, since $P_{X,X}$ and $P_{X,A}$ are both underestimated, their relative values may be close to those given by the exact equations.

Under Drown et al.'s sexual conflict hypothesis, the value for M in equations (13) and (14) is given by equation (10), the predicted number of movements involving the X at steady-state. I therefore computed $P_{X,X}$ and $P_{X,A}$ when $M = 10$ and $M = 50$, which gives a conservative range of the order of magnitude estimate for equation (10). In addition, visual inspection of the results indicated that most mammalian species had the same nuOXPHOS genes on the X, suggesting little gene movement has occurred. I therefore also computed $P_{X,X}$ and $P_{X,A}$ when $M = 0$ and $M = 2$. The results are shown in Figure 2. The predicted probability of each nuOXPHOS gene to be on the X has a high and low horizontal line, the former representing the genes initially on the X (i.e. $P_{X,X}$), and the latter representing the genes initially on the autosomes (i.e. $P_{X,A}$). It is clear that, as the number of gene movements increases from 0, $P_{X,X}$ decreases considerably and $P_{X,A}$ increases slightly. When $M = 50$, the two probabilities are almost equal (Figure 2D; red line). The distribution of nuOXPHOS genes on the X in mammals is highly

skewed (Figure 2; blue dots). The sequence assigned as orthologous to human NDUFA1 was found on the X in 14/15, NDUFB11 in 12/15, and COX7B in 11/15 mammal species. The sequence deemed orthologous to COX7A2 was found on the X in 2 species, while 6 other genes were each found on the X in 1 species. The remaining ~ 60 – 70 nuOXPHOS genes were found on the autosomes in all 15 mammal species. These results most closely match the predicted distribution when $M = 0$ (Figure 2A) and are somewhat close to the prediction when $M = 2$ (Figure 2B). The predicted distribution when M is within the order of magnitude estimate for equation (10) does not match the data points well (Figures 2C & D).

Genomic data was available for only three avian species, so a comparison with the theoretical distribution is not meaningful. However, it is worth noting that *G. gallus* and *T. guttata* had the same 4 nuOXPHOS genes on the Z chromosome. Although, 0 nuOXPHOS genes were found on the Z in *M. gallopavo*, none of the 4 genes on the Z in *G. gallus* and *T. guttata* could be assigned a human ortholog in *M. gallopavo*. Thus, it cannot be concluded that any gene movement has occurred on or off of the Z in birds.

Discussion

I compared the typical hypothesis testing approach in EEB to the quantitative approach more often used in physical sciences for the specific example of investigating how natural selection due to sexual conflict influences nuOXPHOS gene movement among chromosomes. A verbal hypothesis of sexual conflict (Drown et al. 2012) predicts that nuOXPHOS genes should be under-represented on the X chromosome in mammals. I found no significant under- or over-representation on the X chromosome in any mammalian species. However, it is possible that this failure to reject the null hypothesis could be due to low power. Therefore, I conducted an additional analysis treating whether each species had greater or less than the randomly expected

number of genes on the X as a binomial random variable. While the majority of the mammalian species did have less nuOXPHOS genes on the X than expected, this trend was not statistically significant. Taken together, these results do not support sexual conflict as an important factor in the evolution of nuOXPHOS genes in mammals, although lack of statistical power prevents high-confidence in this conclusion.

For birds, the qualitative approach found no under- or over-representation of nuOXPHOS genes on the Z, which is consistent with Drown et al.'s arguments that sexual conflict is not resolved by movement of N-mt genes in birds. Although there are only three bird species in the analysis, the fact that *G. gallus* and *T. guttata*, a divergence that spans nearly the entire class Aves (Hackett et al. 2008), have the same 4 nuOXPHOS genes on the Z suggests very little movement of nuOXPHOS genes has occurred in birds. Therefore, the results for birds could be explained by a hypothesis of strong selection against nuOXPHOS gene movement or a low overall rate of movement and do not provide strong support for sexual conflict.

In order to compare the above qualitative analysis with a quantitative one, I analyzed the evolution of nuOXPHOS genes by postulating a general model of gene movement between chromosomes. Natural selection on gene movement due to sexual conflict, according to Drown et al., was specified and the resulting differential equation solved to yield the predicted dynamics of nuOXPHOS genes evolving in sexual conflict. The results suggest that the number of nuOXPHOS genes on the X always moves toward a steady-state number, which, when selection is absent, is analogous to the random expectation (Figure 1; equation (5)). This is consistent with Drown et al.'s hypothesis, which assumes that selection is required to keep the number of genes on the X from increasing to its randomly expected value. While at first this result may seem to legitimize Drown et al.'s assumption, my quantitative model allows one to see that a species

cannot be considered independent until it has reached steady-state. Thus, the independence of lineages can be tested by determining the number of movements involving the X predicted to occur in the time to reach steady-state. However, because we cannot realistically observe as many movements as are predicted to occur, the probability that each nuOXPHOS gene will be on the X after M movements must be calculated instead. Figure 2A shows that the observed locations of nuOXPHOS genes are consistent with an average value of $M_X(t_{st})$ close to zero in mammals. Even if $M = 10$, the minimum number of movements predicted to occur in a lineage in steady-state, Figure 2C shows that the nuOXPHOS genes should be nearly evenly distributed on the X, and this is not observed. These results are clearly consistent with the hypothesis that no mammalian lineages have reached steady-state with respect to nuOXPHOS genes. Therefore, for nuOXPHOS genes, Drown et al.'s assumption that gene movement occurs often enough to assume independence of lineages is rejected.

One could argue that sexual conflict has selected for nuOXPHOS genes to move off the X, but the rate coefficient, γ, is so small that steady-state has not been reached. In other words, Drown et al.'s hypothesis of sexual conflict could be correct, even though their assumption of steady-state has been rejected. However, to predict the current distribution (i.e. over- or under-representation) of N-mt genes on the X, equation (4) clearly shows that one cannot simply hypothesize sexual conflict without specifying the initial number of genes on the X and the rate of movement in the absence of selection. For example, if the initial number of N-mt genes on the X was sufficiently high, then, even in the face of selection from sexual conflict, an over-representation on the X could be predicted. Likewise, if the rate of gene movement in the absence of selection is sufficiently low, we would predict the current distribution of genes on the X to be approximately the same as the initial distribution, even though, given enough time,

sexual conflict would cause under-representation on the X. If we accept the notion that testable predictions must derivable from a hypothesis in order for it to be scientific, it follows that the initial distribution on the X and the rate of movement in the absence of selection are not mere assumptions made independently of the sexual conflict hypothesis, but must be part of the hypothesis. Thus, rejection of steady-state is sufficient to reject Drown et al.'s hypothesis for nuOXPHOS genes.

Given this conclusion, it is interesting that the results of Drown et al. were consistent with alleviation of sexual conflict as an important factor in the movement of N-mt genes among chromosomes in mammals. Since they argue the selective pressure for N-mt genes to move off the X arises through interactions with mitochondrial-encoded genes, and N-mt genes represent an average of genes that function in many different mitochondrial pathways, it can be argued that nuOXPHOS genes should experience a greater selective pressure to move off the X than N-mt genes as a whole because nuOXPHOS proteins are physically interacting with mtOXPHOS proteins in the electron transport complexes (Rand et al. 2004). I therefore argue that Drown et al.'s support of sexual conflict in N-mt genes deserves more study. By applying my quantitative model of gene movement one can test whether mammalian lineages are independent with respect to all N-mt genes and thus whether the under-representation of N-mt genes on the X could be due to other factors.

In addition to sexual conflict, there are at least two alternative hypotheses that might explain the distribution of N-mt genes among chromosomes, the coadaptation (CA) hypothesis (Wade and Goodnight 2006; Brandvain and Wade 2009; Drown et al. 2012) and the coevolution (CE) hypothesis (Rand et al. 2004; Hill 2013). Briefly, theory predicts that coadapted protein complexes should evolve more efficiently by epistatic selection if the genes are inherited

together. For OXPHOS protein complexes the highest probability of cotransmission occurs when nuOXPHOS genes are on the X in mammals and the autosomes in birds. Thus, the CA hypothesis predicts that nuOXPHOS genes should be over-represented on the X and under-represented on the Z. However, an issue with cotransmission of genes in protein complexes is the so-called Hill-Robertson effect (Comeron et al. 2007), which is the phenomenon that the efficacy of selection on a given locus is reduced when a linked locus also has polymorphisms under selection. For example, Rand et al. (2004) hypothesize that slightly deleterious mutations in mtOXPHOS genes become fixed due to drift and this creates positive selection for compensatory mutations in nuOXPHOS. Hill-Robertson effects would reduce the efficacy of positive selection on such compensatory mutations by an amount directly related to the degree of linkage between nuOXPHOS and mtOXPHOS genes. Therefore, the CE hypothesis implies that nuOXPHOS genes will be selected to move to chromosomes that have the lowest degree of linkage with the mitochondrial genome, which will allow maximum opportunity for coevolution between nuOXPHOS and mtOXPHOS genes. The corresponding predictions of the CE hypothesis are that nuOXPHOS genes will be under-represented on the X and over-represented on the Z (Hill 2013). Obviously, since Table 1 shows no significant over- or under-representation in any mammal or bird species, neither of these two hypotheses is supported by the qualitative approach. Moreover, even though I did not use equation (3) to derive quantitative predictions for either of these hypotheses, the data in Figure 2 suggest any hypothesis that proposes nuOXPHOS genes move primarily due to positive selection is unlikely because there have been close to 0 gene movements per lineage, on average. Only a scenario in which γ is very low can overall positive selection on nuOXPHOS gene movement be consistent with the data in Figure 2.

Finally, it is important to note that I have specified a system of genes evolving in sexual conflict only according to the description of Drown et al. For example, I assumed that once a duplicate copy moves to a new chromosome, the parent gene is lost; however a scenario of sexual conflict can be described in which parent genes are retained in the genome. Modelling this requires a different solution to differential equation (3) because the total number of genes in the system, n, is now a function of time, and thus equation (3) is no longer a separable differential equation. In this case, the equation for $n_X(t)$ would be different and thus the predicted distribution of nuOXPHOS genes on the X, will also be different, at least quantitatively (it is possible that the qualitative form could be very similar to that in Figure 2). Moreover, if the parent gene is not lost, evolutionary theory predicts selection favors the evolution of sex-biased gene expression (Vicoso and Charlesworth 2006), which, once evolved, could reduce the selection pressure for the parent gene to move off the X. This would be specified mathematically by making c in equation (3) a function of time and would lead to another distinct equation for $n_X(t)$. These considerations demonstrate an important point about the distinction between the qualitative and quantitative approaches. Drown et al. imply that their data support sexual conflict regardless of whether gene movement occurs with retention or loss of the parent gene. However, the quantitative approach shows these two processes do not lead to the same quantitative predictions and therefore a more precise description of the underlying gene movement process is necessary to understand the role of sexual conflict in specific systems of genes.

In conclusion, I find that, for the system of nuOXPHOS genes, the qualitative approach finds no support for the sexual conflict hypothesis according to Drown et al., but this conclusion is ambiguous due to low statistical power. Alternatively, I argue the quantitative analysis allows for a robust rejection of the hypothesis (Figure 2) and gives insight into whether selection is

required to keep nuOXPHOS genes under-represented on the X (Figure 1). These findings demonstrate the usefulness of my model of gene movement (equation (1)) to understanding properties of systems of genes that share common features. This model could potentially be used to investigate any such system. The prescription for such an analysis is simply to specify the pattern of the γ_i and biased movement due to selection, c, for each of the j chromosomes. The result is an equation, analogous to equation (2), (which is specific to sexual conflict) that can be solved to yield a testable quantitative prediction. It is important to note, however, that every such analysis, is dependent upon equation (1) being an accurate description of the gene movement process. Therefore, future research should be directed toward testing equation (1) empirically and also assessing the degree to which stochastic variation can affect the robustness of the theoretical predictions. Beyond the question of the validity of the model of gene movement I have proposed, I maintain that the quantitative analysis I have described is a general way in which scientists can gain deeper insight into EEB systems. I recommend that theoretical research in EEB be directed to proposing descriptions of interactions of system components from which testable quantitative predictions can be derived for specific systems.

Table 1. Comparison of observed and expected numbers of nuOXPHOS genes on the sex chromosome.

Species	n[a]	p[b]	Expected[c]	Observed	Obs/Exp	p-value
Mammals						
H. sapiens	80	0.037	2.92 ± 1.7	3	1.03	0.56
P. troglodytes	78	0.042	3.28 ± 1.8	3	0.91	0.58
G. gorilla	75	0.046	3.45 ± 1.8	3	0.87	0.55
P. abelii	77	0.043	3.30 ± 1.8	3	0.91	0.58
M. mulatta	76	0.051	3.87 ± 1.9	3	0.78	0.46
C. jacchus	58	0.042	2.45 ± 1.5	4	1.63	0.23
M. musculus	78	0.055	4.30 ± 2.0	2	0.46	0.19
R. norvegicus	78	0.042	3.30 ± 1.8	2	0.61	0.35
O. cuniculus	43	0.051	2.18 ± 1.4	3	1.37	0.37
C. familiaris	79	0.050	3.96 ± 1.9	5	1.26	0.36
B. taurus	76	0.046	3.53 ± 1.8	3	0.85	0.53
S. scrofa	75	0.059	4.46 ± 2.0	3	0.67	0.34
E. caballus	75	0.052	3.90 ± 1.9	3	0.77	0.45
M. domestica	69	0.025	1.73 ± 1.3	2	1.15	0.52
F. catus	75	0.049	3.66 ± 1.9	3	0.82	0.50
Birds						
G. gallus	64	0.065	4.13 ± 2.0	4	0.97	0.60
T. guttata	58	0.059	3.41 ± 1.8	4	1.17	0.45
M. gallopavo	38	0.051	1.93 ± 1.4	0	0.00	0.14

[a] total number of nuOXPHOS genes
[b] proportion of total genes in the genome on the sex chromosome
[c] random expectation of nuOXPHOS genes on the sex chromosome; given as mean ± 1 st. dev.

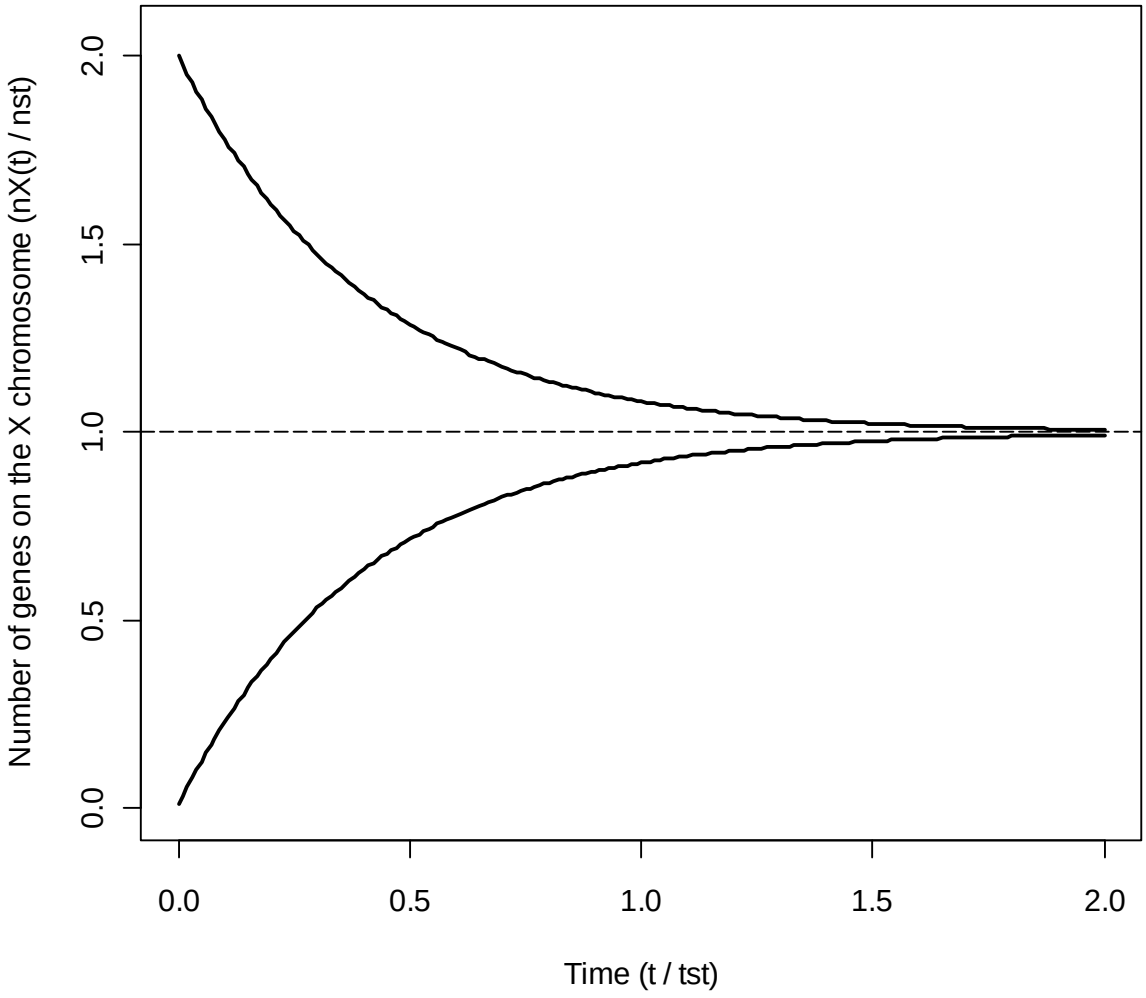

Figure 1. The predicted gene movement dynamics on the X chromosome, according to equation (4). It is clear that no matter the initial number of nuOXPHOS genes on the X, the system eventually reaches a steady-state equilibrium.

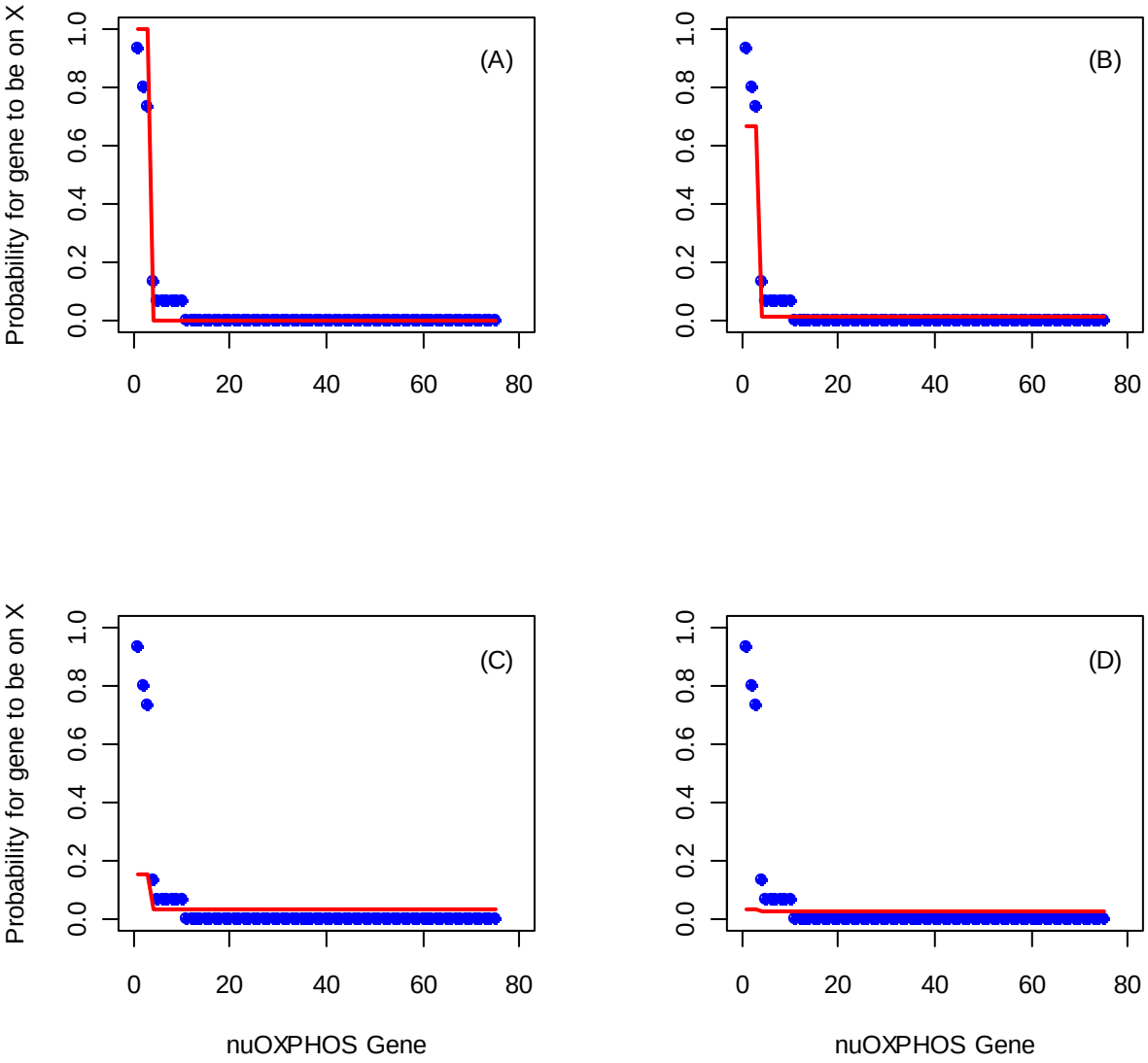

Figure 2. Comparison of the predicted probability distribution (red line) that nuOXPHOS genes will be on the X chromosome to the observed proportion of species for which each nuOXPHOS gene was found on the X (blue dots) in mammals. Each panel has the same data points. The theoretical prediction was computed using equations (13) and (14) when (A) $\boldsymbol{M = 0}$, (B) $\boldsymbol{M = 2}$, (C) $\boldsymbol{M = 10}$, and (D) $\boldsymbol{M = 50}$. According to the sexual conflict hypothesis, the value of M at steady-state is between 10 and 50 for nuOXPHOS genes.

Chapter 2: The role of gene expression in the evolutionary constraint of OXPHOS genes: a case study of the use of regression in evolutionary genomics

Introduction

OXPHOS proteins are directly involved in electron transport and proton pumping, both of which are crucial to the survival of most eukaryotes (Lane 2011). They have also been hypothesized to play an important role in several evolutionary processes, such as hybrid breakdown (Burton et al. 2006), mate choice and sexual selection (Hill and Johnson 2012; Hill and Johnson 2013), and the evolution of sexual reproduction (Hadjivasiliou et al. 2012). Thus, understanding the evolution of OXPHOS genes is potentially important to understanding central questions in evolutionary biology.

Nabholz, Ellegren, and Wolfe (Nabholz et al. 2013; hereafter NEW) studied patterns of evolutionary constraint, as measured by the ratio of non-synonymous substitutions per non-synonymous site to synonymous substitutions per synonymous site (d_N/d_S), in OXPHOS genes across several vertebrate and invertebrate taxa. They found that d_N/d_S was about 4 times lower in mtOXPHOS than in nuOXPHOS genes, on average. This pattern is unexpected if the nearly neutral theory of molecular evolution is largely responsible for the variation in d_N/d_S among OXPHOS genes because the probability of fixation of a slightly deleterious mutation is predicted to be higher when the effective population size (N_e) is lower (Kimura 1962). Mitochondrial genes have a much lower N_e than most nuclear genes due mainly to their haploid nature and lack of recombination (Dowling et al. 2008). Therefore, if mutations in mtOXPHOS and nuOXPHOS have the same distribution of selection coefficients, mtOXPHOS should have a higher d_N/d_S than nuOXPHOS genes, because of their lower N_e. The pattern of d_N/d_S NEW observed

therefore suggests that mutations in mtOXPHOS have, on average, a more negative selection coefficient than in nuOXPHOS genes.

There are two hypotheses often cited in the literature that attempt to explain how selection affects variation in d_N/d_S among genes. The first is that the more "functionally important" (commonly measured by relative growth rates of single-gene deletion yeast strains) a protein is, the more evolutionarily constrained its encoding gene will be. The degree of functional importance is therefore predicted to be negatively correlated with d_N/d_S. This prediction has been upheld, but the strength of the correlation is rather weak (Wang and Zhang 2009). The second hypothesis is that protein misfolding (Drummond and Wilke 2008; Drummond and Wilke 2009) or protein misinteraction (Yang et al. 2012) reduces fitness by an amount directly related to the number of proteins translated. Assuming that relative translation and transcription rates are similar among genes, this hypothesis predicts a negative correlation between gene expression level and d_N/d_S, which has been found in several studies (Drummond et al. 2006).

To determine the degree to which these two hypotheses can explain the lower d_N/d_S in mtOXPHOS genes, NEW gathered a large dataset consisting of several functional parameters, such as hydrophobicity score, proportion of transmembrane helices, and number of protein-protein interactions, as well as estimates of gene expression level. In addition, NEW presented GC content and substitution rate as potential non-selective causes of the variation in d_N/d_S. After eliminating most of the possible explanatory variables, NEW built regression models to analyze which of the remaining candidate variables influenced d_N/d_S. These variables were genome type (G; mitochondrial or nuclear), gene expression level (E; in units of $\log_2$RPKM), hydrophobicity score (H), and the interaction term between E and H. The results of these

regression models (summarized in NEW Table 1), strongly suggest that E is an important explanatory variable of d_N/d_S. H is borderline significant in all models in which it is included and may therefore have some importance, too. From these results, NEW concluded that the low d_N/d_S of mtOXPHOS relative to nuOXPHOS genes exists primarily because mtOXPHOS genes have higher E, on average.

However, inspection of NEW Table 1 reveals an interesting feature of the regression results. Despite NEW's conclusion that E is the only important factor, their Table 1 shows G has a highly significant effect on d_N/d_S in every model in which it is included. This is interesting considering G itself is merely a label and cannot have a true biological effect. Such variables can have a statistical effect for at least two reasons. First, G could be correlated with at least one variable that does have a real biological effect on d_N/d_S. NEW demonstrated that H and E are both correlated with G (NEW Table 2); thus it could be that the effect of G is caused solely by its collinear relationship with these variables. However, simulation studies of regression with collinearity among explanatory variables show that, when all variables with real effects are included in the model, variables that are labels should have zero effect (Freckleton 2002; Freckleton 2011). When NEW included all of their variables in the regression, G still had a highly significant effect, suggesting the existence of at least one unidentified factor that affects d_N/d_S and is collinear with G. Second, it is possible that E truly is the only important factor, but the data violate the assumptions of linear regression, causing the statistical effect of G. One regression assumption is that all explanatory variables are measured without error. NEW speculate that uncertainty in their values of E could cause its effect to be underestimated, in which case the effect of G is expected to be overestimated (Freckleton 2011), possibly by enough to become significant.

I used simulations to test whether NEW's conclusion, that E is the only major factor causing mtOXPHOS to have a lower d_N/d_S than nuOXPHOS genes (hereafter, the NEW hypothesis), is consistent with the significant effect of G, or whether the addition of an unknown factor, U, that affects d_N/d_S and is correlated with G (hereafter, the U hypothesis) is necessary to reproduce this effect. In addition, to determine whether violation of a regression assumption could explain the results, I simulate data, according to both hypotheses, in which E is measured with uncertainty. I then discuss my results in the context of other genomic studies that use regression as a tool to test evolutionary hypotheses.

Methods

I simulated data for the variables, d_N/d_S, E, G, and the unknown factor U. H was ignored because its effect on d_N/d_S is minor, if it has any at all (NEW Table 1). NEW also considered a "clade" variable, but since the trend of lower d_N/d_S in mtOXPHOS was found for each clade (NEW Figure 2A), I ignored this variable as well. I chose a sample size of 93, after the 80 nuOXPHOS and 13 mtOXPHOS genes in human. G was assigned a value of 0 for the nuclear genes and 1 for the mitochondrial genes. The remaining variables were defined according to the following equations:

$$E = \beta_E^0 + \beta_G G + e_x \quad (1)$$

$$U = \beta_U^0 + \beta_G G + e_x \quad (2)$$

$$\frac{d_N}{d_S} = \beta_\omega^0 + \beta_E E + \beta_U U + e_y \quad (3).$$

The parameter values for each intercept, coefficient, and error term, were chosen somewhat arbitrarily, but in an attempt to make the simulated data as close to NEW's data as possible. The qualitative conclusions of the simulation should not depend on the specific parameter values except in extreme cases. β_E^0, β_U^0, and β_ω^0 are the y-intercepts, and were given the values, 3, 2, and 0.25, respectively. β_G is the effect of G on both E and U and was set to 2. β_E is the effect of E on d_N/d_S and was set to -0.025. For the data simulated according to the NEW hypothesis, β_U, the effect of U on d_N/d_S, was given a value of 0. For the data simulated according to the U hypothesis, β_U was given a value of -0.025. The error terms, e_x and e_y, were simulated according to a normal distribution centered on zero, with standard deviation 0.05 and 0.001, respectively.

In order to test whether measurement error in E could account for the effect of G in NEW's results, I also simulated data according to the same equations above, except that in the regression models, E was replaced with E_{obs}, where

$$E_{obs} = E + e_{obs} \quad (4).$$

The measurement error, e_{obs}, was simulated according to a normal distribution centered on zero, with standard deviation 0.25. This value was chosen because preliminary simulations suggested values much higher or lower could not conceivably result in both E and G having significant effects. All simulations were run with 500 replicates.

For both the NEW and U hypotheses, I ran two linear regressions on each replicate dataset: (1) $\frac{d_N}{d_S} \sim E$, and (2) $\frac{d_N}{d_S} \sim E + G$. NEW Table 1 showed that E is significant in the first regression and both E and G are significant in the second regression. These are the results that any hypothesis of which factors affect d_N/d_S must be able to replicate.

Results

I first analyzed the data simulated according to the NEW hypothesis. For the regression $\frac{d_N}{d_S} \sim E$, I found the histogram of the effects of E to be centered on the specified true value, that is, the estimate was unbiased (Figure 1A). Additionally, E was statistically significant in 100% of the replicates. For the regression $\frac{d_N}{d_S} \sim E + G$, the estimate of the effect of E was unbiased (Figure 2A) and E was statistically significant in 98.6% of the replicates. The effect of G was also unbiased, close to its specified value of zero (Figure 2B). Accordingly, G was significant only in 6.4% of the simulations, which is not significantly more often than expected from type I error rates (binomial test, p=0.15).

I then analyzed the data simulated according to the U hypothesis. For the regression $d_N/d_S \sim E$, the average effect of E was about twice that of the true value, that is, the estimate was biased (Figure 1B), but was significant 100% of the time. For the regression $\frac{d_N}{d_S} \sim E + G$, the effect of E was unbiased (Figure 2C) and statistically significant in 97.6% of the replicates. However, the effect of G was biased (Figure 2D), significant greater than 0 in 97.2% of the replicates.

To test if violating the assumption of linear regression that all explanatory variables are measured without error could explain NEW's results, I ran the same regressions as above, but accounted for measurement uncertainty in E (equation (4)). For data simulated according to both the NEW and U hypotheses, the regression $\frac{d_N}{d_S} \sim E_{obs}$ yielded biased estimates of the effect for E (Figure 3). For both hypotheses, E_{obs} was significant in 100% of the simulations. For the regression $\frac{d_N}{d_S} \sim E_{obs} + G$, biased estimates of effect of E and G were also found for both hypotheses (Figure 4). For the NEW hypothesis, E_{obs} and G were statistically significant in

88.4% and 100% of the simulations, respectively. For the U hypothesis, E_{obs} and G were significant in 84.8% and 100% of the simulations, respectively. Thus, the effect of G was significantly greater than 0, for all 500 simulations of both hypotheses.

Discussion

After analyzing a large dataset of d_N/d_S, gene expression level, and hydrophobicity scores, for nuOXPHOS and mtOXPHOS genes in diverse animal taxa, NEW (Nabholz et al. 2013) concluded that gene expression level, E, was the primary factor causing d_N/d_S to be about 4 times lower in mtOXPHOS than in nuOXPHOS genes. I tested whether this conclusion is consistent with the significant effect of G (NEW Table 1) by simulating data according to two hypotheses, one in which E is the only factor affecting d_N/d_S (the NEW hypothesis), and another in which d_N/d_S is affected by an unknown factor, U, that is collinear with G, in addition to E (the U hypothesis).

The goal was to test which, if either, of these two hypotheses could reproduce the following two results from NEW Table 1: for the regression $\frac{d_N}{d_S} \sim E$, E had a significant effect on d_N/d_S; for the regression $\frac{d_N}{d_S} \sim E + G$, both E and G had a significant effect on d_N/d_S. For both the NEW and U hypotheses, the regression $\frac{d_N}{d_S} \sim E$ reproduced the significant effect of E, in all simulation replicates (Figures 1 and 3). Thus, each hypothesis is consistent with this part of NEW's results. For the NEW hypothesis, the regression $\frac{d_N}{d_S} \sim E + G$, reproduced the effect of E (Figure 2A), but did not reproduce the effect of G in more replicates than expected from type I error rates (Figure 2B). This demonstrates that NEW's hypothesis is not consistent with their

regression results, and suggests E is not the only major factor that has caused the low d_N/d_S in mtOXPHOS genes.

Alternatively, the U hypothesis reproduces the significant effects of both E and G for the $\frac{d_N}{d_S} \sim E + G$ regression in nearly all replicates (Figure 2C and 2D). Thus, the U hypothesis is supported, suggesting at least one unknown factor, which is correlated with G, has also contributed to the low d_N/d_S in mtOXPHOS genes.

The above conclusions apply only to data that satisfy the assumptions of linear regression. One such assumption is that all explanatory variables are measured with negligible uncertainty. NEW acknowledge that the values of E in their dataset likely have a non-negligible amount of uncertainty. I therefore repeated the above analysis, but modelled uncertainty in E. For the amount of measurement error specified, I found that both hypotheses predict E to have a significant effect for the regression $\frac{d_N}{d_S} \sim E_{obs}$ (Figure 3). In addition, for the regression $\frac{d_N}{d_S} \sim E_{obs} + G$, both hypotheses predict E and G to each have significant effects (Figure 4). Thus, when the amount of measurement error in E is of suitable magnitude, both hypotheses can be consistent with NEW's results. Note, however, that preliminary data suggests the amount of measurement error influences whether both E and G will have significant effects; when the error is too small, only E has a significant effect, but when the error is too large, only G has a significant effect.

NEW also argue that if the effect of E on d_N/d_S is different among genome types (i.e. there is a statistical interaction between E and G), this indicates E alone cannot explain the variation in d_N/d_S. NEW did not find a significant $E{:}\,G$ interaction and concluded this may offer support for E as the lone factor affecting d_N/d_S. While it may be true that a significant $E{:}\,G$

interaction precludes E from being the only explanatory variable, it does not follow that a lack of an interaction offers evidence that E is the only significant factor. To demonstrate this, I ran regressions with the interaction $E{:}\,G$ for both the NEW and U hypotheses. The average effect of $E{:}\,G$ on d_N/d_S was 0 in all cases, regardless of which hypothesis or whether measurement error was considered (Figure 5). Therefore, the lack of an interaction between E and G cannot be used as support for the NEW hypothesis.

In sum, my simulation results suggest NEW's conclusion that E is the only important factor affecting d_N/d_S cannot account for the effect of G observed in their Table 1. However, NEW admit measurement error in E is likely, so perhaps one could argue their conclusion is consistent with the data because Figure 4B shows a significant effect of G when measurement error was included. I argue that stating E causes the low d_N/d_S of mtOXPHOS genes and is measured with error, is not a valid hypothesis because its predictions, namely whether E and G have significant effects on d_N/d_S, depend on the degree of measurement error. This illustrates the point that one cannot formulate (or make a conclusion about) a valid scientific hypothesis independent of the assumptions necessary to derive its predictions. In the case of measurement error in E, a quantitative value of the degree of error must be specified as part of any hypothesis developed from the regression results. The simplest way to accomplish this is to assume the measurement error is zero. Of course, this can never be satisfied completely, but when an explanatory variable is measured with error that is negligible relative to the uncertainty in the response variable, approximately unbiased estimates of effect can be obtained (Taylor 1996; Freckleton 2011). Under the assumption of zero measurement error, my simulation data clearly show that at least one unknown explanatory variable is a major contributor to the low d_N/d_S of mtOXPHOS genes.

However, it is desirable to determine whether the measurement error is negligible, rather than to just assume it so. Based on patterns in my simulated results, this may be possible using regression results. When data was simulated according the NEW hypothesis without measurement error, the average effect of E was -0.02493 for $\frac{d_N}{d_S} \sim E$ and -0.02498 for $\frac{d_N}{d_S} \sim E + G$, which are effectively identical. Alternatively, when measurement error was modelled, the average effect of E changed from -0.022 for $\frac{d_N}{d_S} \sim E$ to -0.007 for $\frac{d_N}{d_S} \sim E + G$. When data was simulated according to the U hypothesis with no measurement error, the average effect of E changed from -0.0495 for $\frac{d_N}{d_S} \sim E$ to -0.0252 for $\frac{d_N}{d_S} \sim E + G$, and when measurement error was included, the average effect of E changed from -0.0436 to -0.007. Thus, the only case in which the effect of E doesn't change when G is added to the regression is the NEW hypothesis with no measurement error. One may therefore be able to reject this hypothesis with NEW's real data, although the other three hypotheses are indistinguishable. Unfortunately, NEW did not present the estimates of effect from their regressions, so I could not carry out this test. It is also important to note that these changes in the effect of E are an average of 500 replicate datasets. With real regression data, usually only a single dataset is possible; therefore, more research is needed to test the reliability of using change in the effect of E from a single dataset to distinguish between such hypotheses. Alternatively, measurement error can be quantified if repeated measurements of the same quantity can be taken (Taylor 1996). Once measurement error is estimated, model-fitting procedures exists that unbiasedly estimate the effects of the explanatory variables, although these methods are more obscure, if not more complicated, than linear regression (Freckleton 2011).

It is interesting to speculate on the identity of the unknown factor, U, if it exists. NEW compared d_N/d_S between pairs of mtOXPHOS and nuOXPHOS genes that they claimed to be the closest functional matches possible. They made the following comparisons: d_N/d_S of nuclear Cytochrome c1 vs. mitochondrial Cytochrome b and the average d_N/d_S of nuclear COX5A and COX11 vs. the average of mitochondrial COX1, COX2, and COX3. In all taxa examined, d_N/d_S of the mitochondrial genes was lower, which NEW used as evidence that the higher functional constraint of mtOXPHOS genes cannot be explained by differences in functional importance between mtOXPHOS and nuOXPHOS proteins. However, I argue that even if these comparisons truly are the closest functional matches possible between mtOXPHOS and nuOXPHOS genes, it does not follow that their functions must be similar enough to experience similar selective constraints. In other words, even among the genes NEW compared, mtOXPHOS proteins may be sufficiently more functionally constrained to cause at least part of the lower d_N/d_S in mtOXPHOS genes. Hence, U could be some difference in functional importance between mtOXPHOS and nuOXPHOS proteins. One hypothesis is that differences in functional importance exist because mtOXPHOS proteins are located at the structural and functional core of the OXPHOS complexes. For example, in Complex IV, mt-COI is thought to form a channel that can deliver protons used in the reduction of oxygen and through which protons are pumped to maintain the inner mitochondrial membrane potential (Castoe et al. 2008). If, in the other complexes, and especially in Complex I because it has 7 mtOXPHOS genes, mtOXPHOS proteins also make-up the core parts of the protein involved in proton or electron transfer, then perhaps this could contribute significantly to the lower d_N/d_S of mtOXPHOS genes.

Apart from the evolution of OXPHOS genes, I argue that my simulation results provide a case study of the general implications of misuse of regression in genomic studies. The large

amount of genomic datasets available allow for the comparison of many molecular evolutionary parameters between large numbers of genes and species. Regression is a tool often chosen by researchers to help them draw conclusions from such comparisons. As with any method of analysis, it is desirable to understand how tools can be misused in a way that causes misleading conclusions. For example, some investigations into the causes of variation in d_N/d_S among genes in which the explanatory variables were suspected to be collinear with each other, used partial correlations (Drummond et al. 2006; Liao et al. 2010). This process involves running regressions with pairwise combinations of explanatory variables to determine the significance of each variable independent of the others. However, simulations suggest this strategy will give biased estimates of effect for any explanatory variable that is collinear with more than one of the other explanatory variables (Freckleton 2002; Drummond et al. 2006). In addition, non-parametric correlations or linear regressions without effect sizes presented are often run with a single explanatory variable (Flynn et al. 2010; Yang et al. 2010; Yang et al. 2012). If the explanatory variable in such a regression is collinear with any other variable that affects the response, biased results will be obtained. Moreover, my simulation results show how consideration of estimates of effect, as well as p-values, can potentially help distinguish between hypotheses when collinear and perhaps measurement error are present in the data. This is in addition to the many other reasons to include estimates of effect, rather than only p-values, that have been argued previously (Johnson 1999; Anderson and Burnham 2002). Thus, the issue of accounting for collinearity among explanatory variables is certainly not restricted to NEW's analysis. In sum, I have shown that when the effects of collinearity and assumptions of regression are not considered in the analysis of genomic datasets, a hypothesis can be given

undeserved support. More scrutiny should be applied when using regression to analyze genomic datasets.

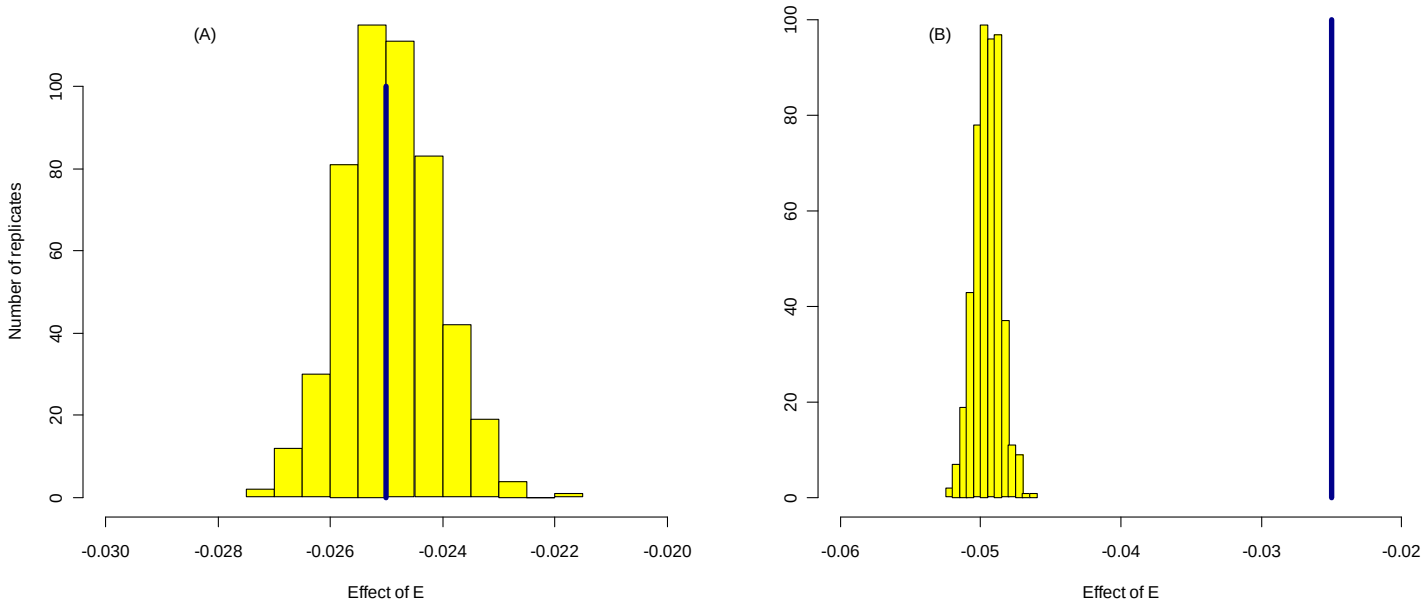

Figure 1. Comparison of the histogram of estimated effects of $\boldsymbol{E}$ from 500 simulations (yellow bars) with the specified true value (dark blue line) for the regression $\frac{\boldsymbol{d_N}}{\boldsymbol{d_S}} \sim \boldsymbol{E}$. Data simulated according to (A) the NEW hypothesis, and (B) the U hypothesis with no measurement error.

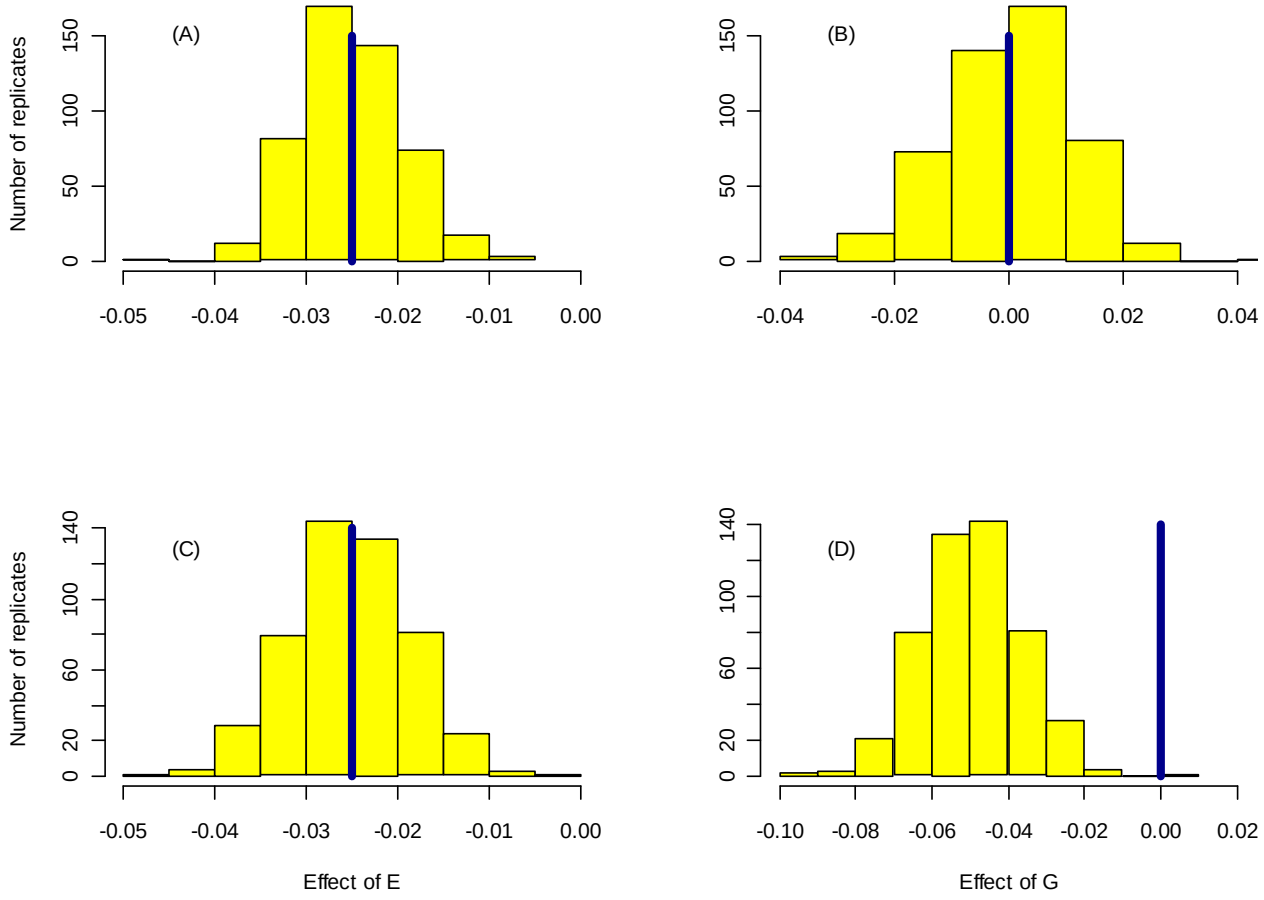

Figure 2. Comparison of the histogram of estimated effects of $\boldsymbol{E}$ and $\boldsymbol{G}$ from 500 simulations (yellow bars) with the specified true value (dark blue line) for the regression $\frac{\boldsymbol{d_N}}{\boldsymbol{d_S}} \sim \boldsymbol{E} + \boldsymbol{G}$. Data was simulated according to the NEW hypothesis, (A) and (B), and the U hypothesis, (C) and (D), with no measurement error.

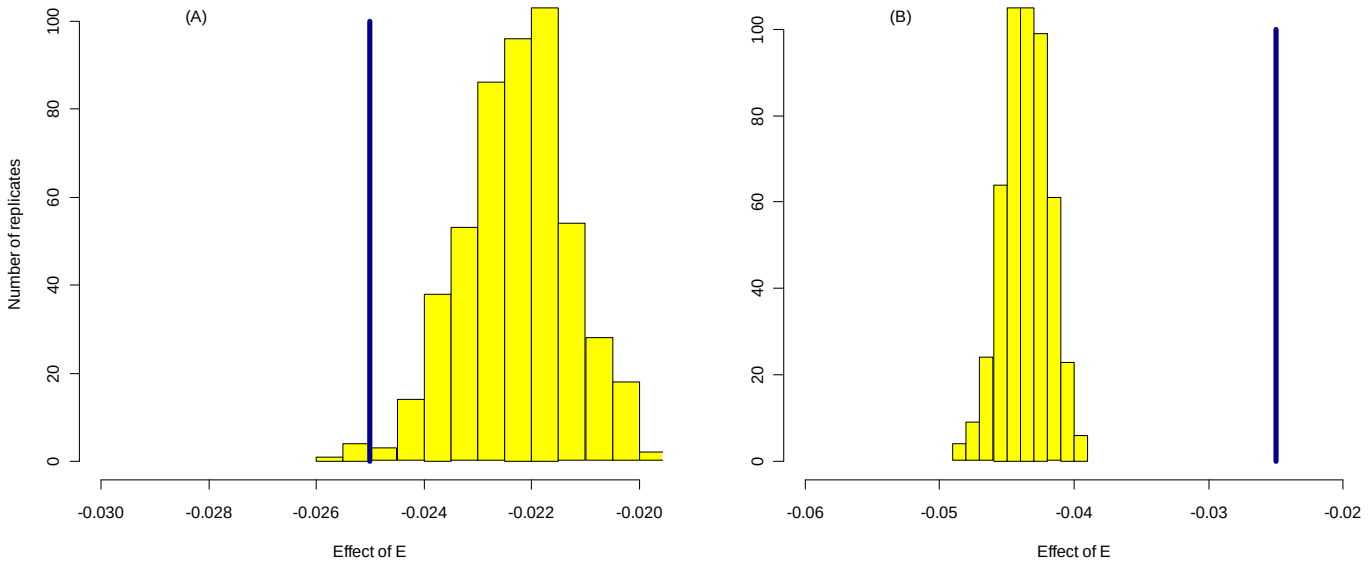

Figure 3. Comparison of the histogram of estimated effects of $\boldsymbol{E}$ from 500 simulations (yellow bars) with the specified true value (dark blue line) for the regression $\frac{\boldsymbol{d_N}}{\boldsymbol{d_S}} \sim \boldsymbol{E_{obs}}$. Data simulated according to (A) the NEW hypothesis, and (B) the U hypothesis with measurement error.

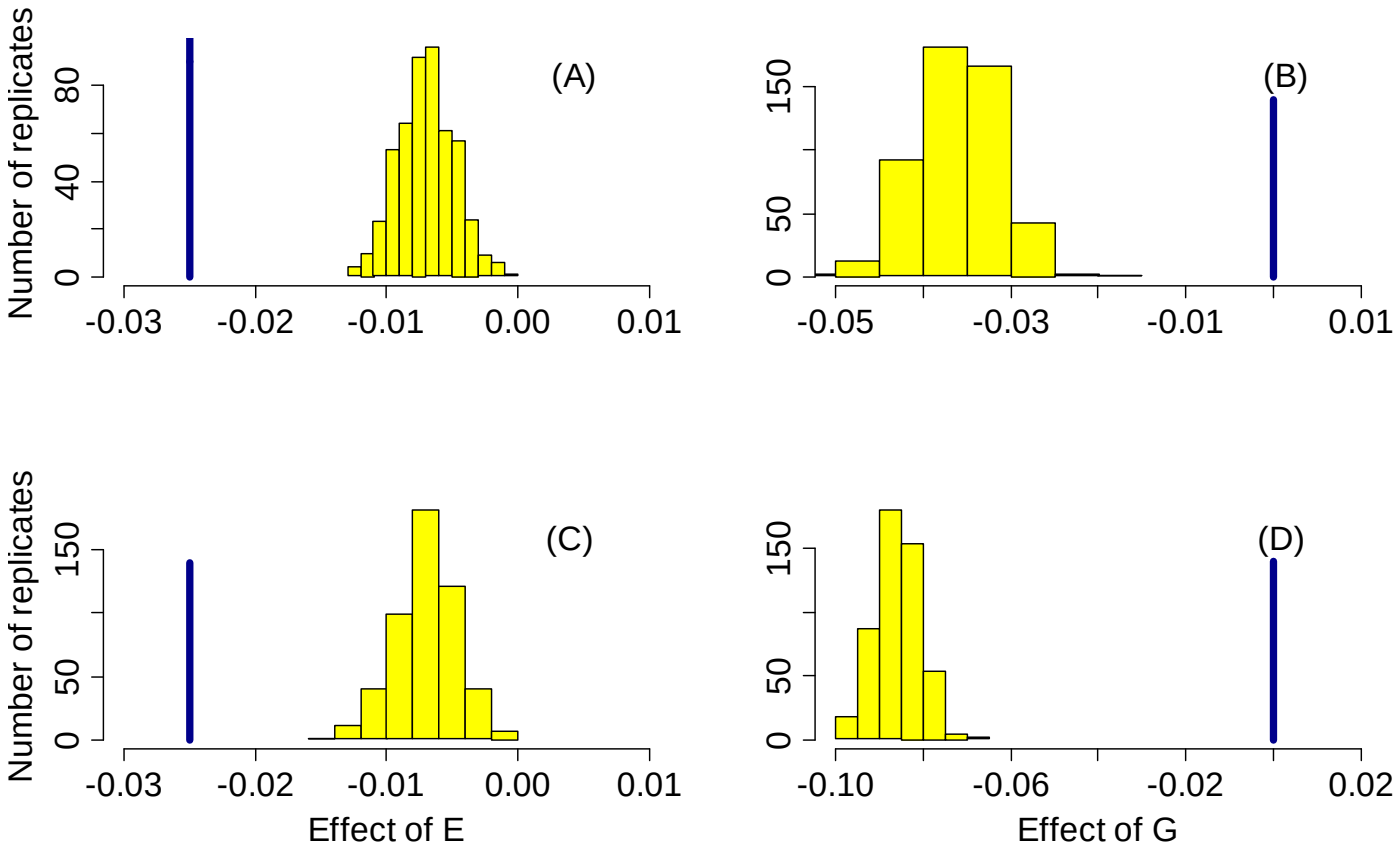

Figure 4. Comparison of the histogram of estimated effects of $\boldsymbol{E}$ and $\boldsymbol{G}$ from 500 simulations (yellow bars) with the specified true value (dark blue line) for the regression $\frac{\boldsymbol{d_N}}{\boldsymbol{d_S}} \sim \boldsymbol{E_{obs}} + \boldsymbol{G}$. Data was simulated according to the NEW hypothesis, (A) and (B), and the U hypothesis, (C) and (D) with measurement error.

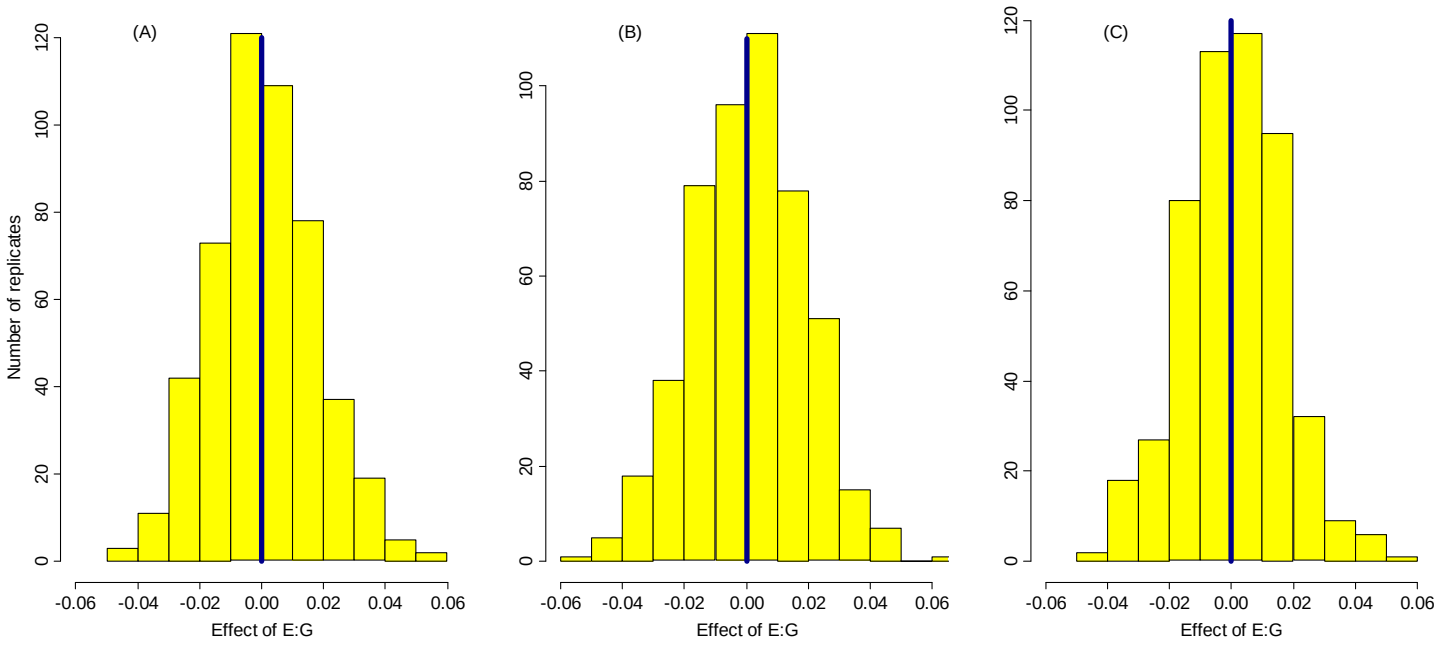

Figure 5. Comparison of the histogram of estimated effects of $\boldsymbol{E}:\boldsymbol{G}$ from 500 simulations (yellow bars) with the specified true value (dark blue line). The regression models were (A) $\frac{\boldsymbol{d_N}}{\boldsymbol{d_S}} \sim \boldsymbol{E} + \boldsymbol{G} + \boldsymbol{E}:\boldsymbol{G}$, for data simulated according to the NEW hypothesis, and (B) $\frac{\boldsymbol{d_N}}{\boldsymbol{d_S}} \sim \boldsymbol{E} + \boldsymbol{G} + \boldsymbol{E}:\boldsymbol{G}$, and (C) $\frac{\boldsymbol{d_N}}{\boldsymbol{d_S}} \sim \boldsymbol{E} + \boldsymbol{G} + \boldsymbol{U} + \boldsymbol{E}:\boldsymbol{G}$, for data simulated according the U hypothesis. The estimates of the effect of $\boldsymbol{E}:\boldsymbol{G}$ for data simulated with measurement error are almost identical (not shown).

References

Alon U. 2006. An Introduction to Systems Biology: Design Principles of Biological Circuits. CRC press

Anderson DR, Burnham KP. 2002. Avoiding pitfalls when using information-theoretic methods. J. Wildl. Manage. 66:912–918.

Bar-Yaacov D, Blumberg A, Mishmar D. 2012. Mitochondrial-nuclear co-evolution and its effects on OXPHOS activity and regulation. Biochim. Biophys. Acta [Internet] 1819:1107–1111. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22044624

Berryman AA. 2003. On principles, laws and theory in population ecology. Oikos 103:695–701.

Bhutkar A, Russo SM, Smith TF, Gelbart WM. 2007. Genome-scale analysis of positionally relocated genes. Genome Res. [Internet] 17:1880–1887. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2099595&tool=pmcentrez&rendertype=abstract

Brandvain Y, Wade MJ. 2009. The functional transfer of genes from the mitochondria to the nucleus: the effects of selection, mutation, population size and rate of self-fertilization. Genetics [Internet] 182:1129–1139. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2728854&tool=pmcentrez&rendertype=abstract

Burton RS, Ellison CK, Harrison JS. 2006. The Sorry State of F 2 Hybrids : Consequences of Rapid Mitochondrial DNA Evolution in. Am. Nat. 168:S14–S24.

Castoe T a, Jiang ZJ, Gu W, Wang ZO, Pollock DD. 2008. Adaptive evolution and functional redesign of core metabolic proteins in snakes. PLoS One [Internet] 3:e2201. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2376058&tool=pmcentrez&rendertype=abstract

Colyvan M, Ginzburg LR. 2010. Analogical thinking in ecology: looking beyond disciplinary boundaries. Q. Rev. Biol. [Internet] 85:171–182. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20565039

Comeron JM, Williford A, Kliman RM. 2007. The Hill-Robertson effect: evolutionary consequences of weak selection and linkage in finite populations. Heredity (Edinb). [Internet]:1–13. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17878920

Dowling DK, Friberg U, Lindell J. 2008. Evolutionary implications of non-neutral mitochondrial genetic variation. Trends Ecol. Evol. [Internet] 23:546–554. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18722688

Drown DM, Preuss KM, Wade MJ. 2012. Evidence of a paucity of genes that interact with the mitochondrion on the X in mammals. Genome Biol. Evol. [Internet] 4:763–768. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3509887&tool=pmcentrez&rendertype=abstract

Drummond DA, Raval A, Wilke CO. 2006. A single determinant dominates the rate of yeast protein evolution. Mol. Biol. Evol. [Internet] 23:327–337. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16237209

Drummond DA, Wilke CO. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell [Internet] 134:341–352. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2696314&tool=pmcentrez&rendertype=abstract

Drummond DA, Wilke CO. 2009. The evolutionary consequences of erroneous protein synthesis. Nat. Rev. Genet. [Internet] 10:715–724. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2764353&tool=pmcentrez&rendertype=abstract

Ellegren H. 2011. Emergence of male-biased genes on the chicken Z-chromosome : Sex-chromosome contrasts between male and female heterogametic systems. Genome Res.:2082–2086.

Emerson JJ, Kaessmann H, Betrán E, Long M. 2004. Extensive gene traffic on the mammalian X chromosome. Science (80-.). [Internet] 303:537–540. Available from: http://www.ncbi.nlm.nih.gov/pubmed/14739461

Evans MR, Grimm V, Johst K, et al. 2013. Do simple models lead to generality in ecology? Trends Ecol. Evol. [Internet] 28:578–583. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23827437

Flynn KM, Vohr SH, Hatcher PJ, Cooper VS. 2010. Evolutionary rates and gene dispensability associate with replication timing in the archaeon Sulfolobus islandicus. Genome Biol. Evol. [Internet] 2:859–869. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3000693&tool=pmcentrez&rendertype=abstract

Freckleton RP. 2002. On the misuse of residuals in ecology: regression of residuals vs. multiple regression. J. Anim. Ecol.:542–545.

Freckleton RP. 2011. Dealing with collinearity in behavioural and ecological data : model averaging and the problems of measurement error. Behav. Ecol. Sociobiol. 65:91–101.

Gallach M, Chandrasekaran C, Betrán E. 2010. Analyses of nuclearly encoded mitochondrial genes suggest gene duplication as a mechanism for resolving intralocus sexually

antagonistic conflict in Drosophila. Genome Biol. Evol. [Internet] 2:835–850. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2995371&tool=pmcentrez&rendertype=abstract

Ginzburg LR, Jensen CXJ, Yule J V. 2007. Aiming the "unreasonable effectiveness of mathematics" at ecological theory. Ecol. Modell. [Internet] 207:356–362. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0304380007002876

Hackett SJ, Kimball RT, Reddy S, et al. 2008. A phylogenomic study of birds reveals their evolutionary history. Science (80-.). [Internet] 320:1763–1768. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18583609

Hadjivasiliou Z, Pomiankowski A, Seymour RM, Lane N. 2012. Selection for mitonuclear co-adaptation could favour the evolution of two sexes. Proc. R. Soc. B [Internet] 279:1865–1872. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3297446&tool=pmcentrez&rendertype=abstract

Hansson L. 2003. Why ecology fails at application: should we consider variability more than regularity? Oikos 100:624–627.

Hill GE, Johnson JD. 2012. The vitamin A-redox hypothesis: a biochemical basis for honest signaling via carotenoid pigmentation. Am. Nat. [Internet] 180:E127–50. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23070328

Hill GE, Johnson JD. 2013. The mitonuclear compatibility hypothesis of sexual selection The mitonuclear compatibility hypothesis of sexual selection. Proc. Roy. Soc. B.

Hill GE. 2013. Sex linkage of nuclear-encoded mitochondrial genes. Heredity (Edinb). [Internet]:1–2. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24346499

Johnson DH. 1999. The Insignificance of Statistical Significance Testing. J. Wildl. Manage. 63:763–772.

Kaessmann H, Vinckenbosch N, Long M. 2009. RNA-based gene duplication: mechanistic and evolutionary insights. Nat. Rev. Genet. 10:19–31.

Kimura M. 1962. On the probability of fixation of mutant genes in a population. Genetics 47:713–719.

Knapp AK, Smith MD, Collins SL, et al. 2004. Generality in ecology: testing North American grassland rules in South African savannas. Front. Ecol. Environ. 2:483–491.

Lane N. 2011. Mitonuclear match: optimizing fitness and fertility over generations drives ageing within generations. Bioessays [Internet] 33:860–869. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21922504

Li W-H. 1997. Molecular Evolution. Sinauer Associates

Liao B-Y, Weng M-P, Zhang J. 2010. Impact of extracellularity on the evolutionary rate of mammalian proteins. Genome Biol. Evol. [Internet] 2:39–43. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2839354&tool=pmcentrez&rendertype=abstract

Moyle LC, Muir CD, Han M V, Hahn MW. 2010. The contribution of gene movement to the "two rules of speciation". Evolution (N. Y). [Internet] 64:1541–1557. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20298429

Murray BG. 2000. Universal laws and predictive theory in ecology and evolution. Oikos 89:403–408.

Nabholz B, Ellegren H, Wolf JBW. 2013. High levels of gene expression explain the strong evolutionary constraint of mitochondrial protein-coding genes. Mol. Biol. Evol. [Internet] 30:272–284. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23071102

Nei M, Kumar S. 2000. Molecular Evolution and Phylogenetics. Oxford University Press

O'Hara RB. 2005. The anarchist's guide to ecological theory. Or, we don't need no stinkin' laws. Oikos 110:390–393.

Parisi M, Nuttall R, Naiman D, Bouffard G, Malley J, Eastman S, Oliver B. 2003. Paucity of Genes on the Drosophila X Chromosome Showing Male-Biased Expression. Science (80-.). 299:697–700.

Potrzebowski L, Vinckenbosch N, Marques AC, Chalmel F, Jégou B, Kaessmann H. 2008. Chromosomal gene movements reflect the recent origin and biology of therian sex chromosomes. PLoS Biol. [Internet] 6:e80. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2276528&tool=pmcentrez&rendertype=abstract

Rand DM, Haney R a, Fry AJ. 2004. Cytonuclear coevolution: the genomics of cooperation. Trends Ecol. Evol. [Internet] 19:645–653. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16701327

Taylor JR. 1996. An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements. University Science Books

Vibranovski MD, Zhang Y, Long M. 2009. General gene movement off the X chromosome in the Drosophila genus. Genome Res. [Internet] 19:897–903. Available from:

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2675978&tool=pmcentrez&rendertype=abstract

Vicoso B, Charlesworth B. 2006. Evolution on the X chromosome: unusual patterns and processes. Nat. Rev. Genet. [Internet] 7:645–653. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16847464

Wade MJ, Goodnight CJ. 2006. CYTO-NUCLEAR EPISTASIS : TWO-LOCUS RANDOM GENETIC DRIFT IN HERMAPHRODITIC AND DIOECIOUS SPECIES CYTO-NUCLEAR EPISTASIS : TWO-LOCUS RANDOM GENETIC DRIFT IN. Evolution (N. Y). 60:643–659.

Wang Z, Zhang J. 2009. Why is the correlation between gene importance and gene evolutionary rate so weak? PLoS Genet. [Internet] 5:e1000329. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2605560&tool=pmcentrez&rendertype=abstract

Yang J-R, Liao B-Y, Zhuang S-M, Zhang J. 2012. Protein misinteraction avoidance causes highly expressed proteins to evolve slowly. Proc. Natl. Acad. Sci. U. S. A. [Internet] 109:E831–40. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3325723&tool=pmcentrez&rendertype=abstract

Yang J-R, Zhuang S-M, Zhang J. 2010. Impact of translational error-induced and error-free misfolding on the rate of protein evolution. Mol. Syst. Biol. [Internet] 6:doi:10.1038/msb.2010.78. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2990641&tool=pmcentrez&rendertype=abstract

Zhang YE, Vibranovski MD, Landback P, Marais G a B, Long M. 2010. Chromosomal redistribution of male-biased genes in mammalian evolution with two bursts of gene gain on the X chromosome. PLoS Biol. [Internet] 8. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2950125&tool=pmcentrez&rendertype=abstract