Supplemental Text S1

Marlena Siwiak, Tomasz Wyszomirski & Piotr Zielenkiewicz

Frequency of fictitious contradictions.

To examine the influence of NHST misinterpretations on scientific debates, we consider two identical and independent studies of an effect that actually exists, i.e., its null hypothesis is false. In each case, some statistical test for the existence of the effect gives either a significant or non-significant result. The results of the two studies may be compared by: (i) a proper statistical test checking whether effects in two cases are the same (correct method), or (ii) by checking whether the studies agree as to the obtained statistical significance (incorrect method).

In the first case, the null hypothesis stating that there is no difference in effects is true because both studies concern the same population. Therefore, a discrepancy of effects will appear by chance with frequency α, which is the significance level adopted when comparing studies. The existence of such false controversies is inevitable.

When the second method is applied, the number of false controversies may be much larger. Assuming the same statistical power M for two studies, they both yield significant results with probability M^2 and non-significant results with probability $(1 - M)^2$. The remaining cases, in which one of the studies gives significant and the other non-significant results, are commonly misinterpreted as contradictions between the results. If such controversies do not result from application of any correct method, we call them fictitious. Overall, significant vs. non-significant contradictions appear with probability $2M(1-M)$ which, for M = 0.5, equals 0.5. Thus, we can expect such contradictions in 50% of comparisons of two identical studies– ten times more frequently than inevitable contradictions if we apply the first method of comparison and take α = 0.05.

Next, we estimate the proportion of all contradictions (i.e., declared by either method) constituted by fictitious contradictions (f). Some but not necessarily all inevitable contradictions may also lead to significant vs. non-significant discrepancies. Let h denote the fraction of such contradictions. Then, the frequency of all contradictions is $2M(1-M)+(1-h)\alpha$, and the frequency of fictitious contradictions is $2M(1-M)-h\alpha$. Therefore,

$$f = \frac{2M(1-M)-h\alpha}{2M(1-M)+(1-h)\alpha}. \tag{1}$$

As is easy to show, f is a strictly decreasing function of h. Thus, for $h \in [0,1]$ this function attains a minimum and maximum at h = 1 and h = 0, respectively. To find what values of f are possible, it is sufficient to consider $f(h)$ for these two extremes of h. Let us assume the conventional significance level, α = 0.05. For statistical power M = 0.5, the values of f are 0.90 and 0.91. For M = 0.8, they are 0.84 and 0.87. Even for a power as high as M = 0.9 (or as low as 0.1), these two values do not decrease dramatically: 0.72 and 0.78.

Above derivation corresponds to the situation in which attention is paid only to statistical significance and not the sign of the effect. If direction of the effect is not omitted, and one of the studies correctly detects the effect and the other meets directional error, i.e., declares the effect of the opposite sign, the pair should be considered an inevitable contradiction, although they both yield significant results. However, it is not straightforward to introduce this into Eq. 1.

For simplicity, assume the same significance level α_* for both studies, such that $\alpha_* = \alpha$. Directional contradictions occur when one study makes a correct directional decision while the other makes a directional error, so that their frequency is the doubled product of probabilities of these events and should be added to the denominator. The probability of directional error is bounded from above by $\alpha/2$, while the probability of correct directional conclusion by 1. Therefore, the upper bound for the frequency of directional contradictions is α and Eq. 1 changes to:

$$f_c > \frac{2M(1-M) - h_c\alpha}{2M(1-M) + (2-h_c)\alpha}. \tag{2}$$

Here, subscript *c* denotes the correction for directional contradictions, and h_c may differ from h but still $h_c \in [0,1]$ and f_c remains a strictly decreasing function of h_c. Thus, by taking $h_c = 1$ one obtains a very cautious lower estimate of the proportion of fictitious contradictions:

$$\underline{f_c} = \frac{2M(1-M) - \alpha}{2M(1-M) + \alpha}. \tag{3}$$

Putting $\alpha = 0.05$ in Eq. 3 and for $M \in [0.09, 0.91]$, more than a half of all contradictions (possibly much more) are fictitious. Since in practice of many disciplines statistical power is usually within this range[1–5], the problem is serious.

To avoid fictitious contradictions resulting from misbelieved properties of NHST, it is necessary to consistently use a "don't know" category, i.e., non-significant result should lead to the suspension of judgement. However, even with non-significant results, one can still obtain some information about population parameters by deeming many values of parameters implausible. To achieve this result, it is necessary to transgress NHST thinking and follow the ESCI approach.

Case study I: Protein translation efficiency determinants.

One of the well-established explanation of biased codon usage states that it increases the efficiency and accuracy of translation[6–8]. A recent study[9] seems to contradict this statement. The authors constructed a library of 158 GFP synonymous sequences and expressed them in an identical regulatory context in *E.coli* cells, gauging the GFP expression level by its fluorescence. Surprisingly, "codon bias did not correlate with expression" (ρ = 0.14, p-value = 0.09), and the authors concluded that codon bias does not have "significant effects" on protein levels. In response, a related study[10] on the endogenous genes of *E.coli* and *S.cerevisiae* reported a statistically significant association between codon bias and protein abundance normalised to mRNA level (ρ = 0.27, p-value = 1.7e-8, and ρ = 0.12, p-value = 1.47e-9, respectively). The authors concluded that codon bias is an important determinant of translation efficiency, and the discrepancy between the studies stems from differences in mRNA's folding energies of synthetic and endogenous genes.

These conflicting results gained a good deal of interest from the research community and had been cited over 900 times at the time of writing. They were discussed in many reviews[8,11,12] and were subjected to further analysis and interpretation[13]. Other researchers[14,15] applied a similar experimental scheme for other genes, and found that neither the mRNA secondary structure nor codon bias correlated with expression, although statistically significant correlations could be spotted for other mRNA features.

We took a step back and calculated 95% CIs for correlations between genes' codon bias and expression level reported originally[9,10]. As reference studies use different methods to gauge codon bias (either Codon Adaptation Index CAI [16] or tRNA Adaptation Index tAI [17]), we performed our calculations for both indices separately. We used the same types of correlations as those calculated in the reference studies: simple (with expression level measured as in the reference studies) and partial (between codon bias and protein level (bacteria and yeast data sets) or fluorescence (GFP data set)) while controlling for the mRNA level. Our analysis confirms previous observations on statistical significance of the obtained correlations (Fig. 2a, main text).

Next, we calculated 95% CIs for the differences of all corresponding pairs of intra-species correlations (Fig. 2b, main text). In the case of the yeast-GFP comparison of simple correlations and also for some *E.coli*-yeast comparisons, the hypothesis that the compared sets have identical correlation coefficients cannot be rejected. For these data, no discrepancy between studies could be found. In case of *E.coli*-GFP comparisons, the true correlation for *E.coli* genes is larger than true correlation for GFP constructs, at least by 0.03.

To decide whether the correlation difference of size 0.03 is large enough to have any practical consequences, we calculated the differences of correlations between codon bias and gene expression, with gene expression inferred by several different, but qualitatively equivalent experiments. Due to the availability of data, we performed this analysis only for the *S.cerevisiae* genes. However, as experimental methods quantifying protein and mRNAs are similar for both prokaryotes and eukaryotes, we expect similar variability in *E.coli*. Originally[10], (local) expression level of a gene was defined as the quotient of its protein and mRNA abundance, measured experimentally by[18] and[19], respectively. The experimental measurements of these quantities were performed independently by several other groups and, as their quality seems similar, it may be assumed that the choice made in the reference study[10] was arbitrary. To check how such choices affect research conclusions, we repeated the calculation of the expression of yeast genes by taking four genome-wide protein-level measurements[18,20–22] and five large scale mRNA measurements[19,23–25], and then calculated protein-mRNA quotients for all possible measurements combinations. Next, we calculated the 95% CIs for correlations between each determined expression level and codon bias inferred by CAI or tAI. We refer to these as alternative correlations. To facilitate comparisons, we limited this analysis to 303 genes common to all nine genome-wide experiments (Fig. 3, main text); however, comparisons over partially overlapping, larger data sets yielded similar results (S1 Fig and S2 Fig). As can be seen from Fig. 3a, main text, the alternative correlations for the same variables (codon bias and expression) and an identical set of genes appear strikingly dissimilar. Some CIs indicate correlations that are stronger than previously reported[10], some are "statistically non-significant", and some even show weak negative correlations.

We quantified the sizes of the differences of correlations by calculating the 95% CIs for the differences between the original correlation[10], and each of the alternative correlations (Fig. 3b, main text). For 8 of 19 cases of tAI vs. expression comparisons, the calculated CIs do not allow us to determine the correlations' difference sign; nevertheless, we observe positive or negative differences in the remaining cases. Examining the extremes of their CIs allows us to conclude that, for these cases, the difference between original and alternative correlation is at least 0.05. However, there are other combinations for which the correlation difference is at least as high as 0.33. Note that some combinations with the "Gr" protein abundance set[22] return negative correlations between codon bias and expression. Even if we assume that these negative correlations must result from data error and reject them as outliers, more than half of our alternative correlations differed from the original correlation by at least 0.05. For confirmation of these results, we also calculated 95% CIs for the correlation differences contrasts (Fig. 3c, main text).

As the calculations were performed on the same set of genes, for which codon bias measures are identical, all the differences between original and alternative correlations must result from the discrepancies in mRNA and protein abundances reported by experimental studies. Such discrepancies may be caused by both imperfections of contemporary experimental techniques, and natural variability of the cell. If this noise yields correlation differences of at least 0.05 in more than half of analysed cases, the discussion of those differences with sizes of at least 0.03 becomes problematic.

In such a case, the explanation that the observed discrepancies result from differences in folding energies of endogenous and artificial genes[10] seems highly questionable. In support of the hypothesis that folding energy modulates the relation between codon bias and translation efficiency, the authors divided the sets of analysed *E.coli* and *S.cerevisiae* genes into five equal size bins according to the folding energy of their transcripts. In each bin, they separately calculated the correlation between codon bias and translation efficiency. They observed that "the strength of association between codon bias and local translation efficiency is dependent on the levels of folding energy" by finding that "the most significant correlation between codon bias and local translation efficiency is in the bin corresponding to very high folding energy (-1.2 mean folding energy)". As p-values are not, by definition, estimates of the association strength, we assume that "the most significant" was supposed to refer to the strongest correlation, not the lowest p-value. We reproduce the results presented in Fig. 3.[10] and supplement each reported correlation coefficient with its 95% CI (S3 Fig, panels a and c). Next, for each pair of bins we calculated the 95% CI for the difference of their correlations. As shown in S3 Fig (panels b and d), we cannot determine the difference between any single pair of bins. This means that, on the basis of these data, we are not able to answer whether folding energy modulates the association between codon bias and expression. However, such an effect may exist, but it is relatively weak and possibly too weak to be measured by existing techniques (compared to the correlation differences calculated above).

Case study II: PTPRC (CD45) association with the development of multiple sclerosis.

In 2000, a 77G allele of the gene PTPRC encoding protein tyrosine phosphatase, receptor-type C (or CD45), was associated with multiple sclerosis (MS)[26]. One line of evidence for the increased susceptibility to MS caused by the 77C→G polymorphism was based on the analysis of allele frequencies in MS patients and controls. The authors claimed that "in three of four independent case-control studies, we demonstrated an association of the mutation with MS", which was achieved by obtaining p-values < 0.05 in an exact Fisher test. In the two following studies[27,28], the authors performed a similar analysis for different case-control groups, but they "did not find a significant difference between allele frequencies in (...) MS patients and controls ($P > 0.05$)"[27].

We reanalysed the data from all three contradicting publications. To determine whether the 77G allele is a risk factor for the occurrence of multiple sclerosis, we first calculated the 95% CIs for the ratio between the odds for the occurrence of disease among carriers of the mutated (77G) and regular (77C) allele (Fig. 4a, main text). In all study groups, except the Initial and Validation Marburg studies and the Hannover study from[26], the obtained CIs for log odds ratio (dark blue) did not allow us to determine whether the 77G allele is associated with higher or lower odds of disease.

Next, we compared the odds ratios obtained for reference studies[26–28] by calculating 95% CIs for the relative odds ratios (i.e., ratio of odds ratios) between all possible pairs of study groups

(Fig. 4b, main text). As all four groups from[27] share the same cohort of MS patients and thus, are not independent, they were not compared with each other. Hence, the odds ratios for the American study group from[26] and for all study groups from[27,28] cannot be distinguished between each other based on the analysed data. In case of the Hannover group from[26] and its comparisons with groups from[27,28], all CIs are slightly right shifted, but the elevated odds ratio may be stated only in two cases. For the Initial and Validation Marburg groups from[26], the odds ratios are clearly elevated in relation to groups from[27,28] and also in relation to the American group from the same study[26]. This confirms the initial controversy between the analysed studies.

Why do the Initial and Validation Marburg groups[26] differ so greatly? A closer look at the Marburg controls raises supposition that the number of 77G carriers among them may be slightly underestimated. By simple proportion, and assuming equal frequency of the 77G allele in all analysed populations of healthy people, one should expect approximately 2 to 5 carriers in Marburg control groups. The cause of this inconsistency may be found in the Methods section in[26], which explains, “We personally interviewed all healthy donors with respect to neurological symptoms. Healthy donors with a history of neurological symptoms or a family history for MS or related diseases were excluded [from the Marburg controls ed.]”.

This approach requires further thought, as comparing two Marburg groups[26] with studies that do not impose such restrictions may be not appropriate. If the 77G allele indeed occurs more often among MS patients and an examined healthy donor has a family history of MS, it is more likely that some of his relatives are carriers of the 77G allele rather than 77C; hence it is more likely that he is a 77G carrier himself. Excluding such persons from the control group results in a decreased number of 77G carriers in Marburg controls. This exaggerates the initial disproportion in the mutated allele frequencies, which finally may become statistically significant in an exact Fisher test or may result in elevated odds of disease, as shown above.

To examine how strongly the interview procedure influences the final outcome, we determined how many 77G carriers must be excluded from the control group of the Initial and Validation Marburg studies to obtain at least somewhat higher odds of disease given the mutated allele (i.e., to obtain statistically significant odds ratio). We found that if the interview procedure had overlooked only 5 and 3 77G carriers in the Initial and Validation Marburg study, respectively, the elevated odds for disease could not have been stated and the exact Fisher test performed in[26] would not have resulted in a statistically significant outcome. Note that both numbers of 77G carriers in controls (5/194 and 3/117) are in agreement with expectations based on proportions observed in other analysed populations. Marburg study groups with modified controls, i.e., enlarged by 5 (Initial study) and 3 (Validation study) healthy 77G carriers, were further used to calculate new CIs for odds ratios and relative odds ratios, which are presented in the main text Fig. 4a (light blue) and 4b (light red), respectively. The obtained CIs for odds ratios do not allow us to state whether there are higher or lower odds of disease among 77G carriers in both modified Marburg studies. Panel b shows that results for the modified Marburg data are consistent with those reported for other study groups, as none of the relative odds ratios may be shown to be different than one (in the linear scale) with 95% confidence.

The addition of 5 and 3 77G carriers to control groups made the initial controversy disappear. This example shows, that NHST based reasoning exhibits chaotic behaviour, as small changes in the input data may result in radically different outcomes, namely, a lack or presence of statistical significance, that is immediately interpreted as the lack or presence of the association between the analysed factor and disease. This chaos originated from the reluctance to quantitative interpretations and it is apparent in the poorly framed research problem. The question is not whether there is an association between the 77G allele and disease, but rather how strong this association is and, further, whether it is strong enough to have any practical consequences on multiple sclerosis prevention and treatment.

Case study III: Divergence of X-linked and autosomal genes in *Drosophila*.

According to a popular hypothesis, if certain conditions are fulfilled, loci on the X chromosome are expected to have higher rates of adaptive evolution than those located on the autosomes[29]. When the number of known gene sequences was small, several groups tested this hypothesis, examining the evolutionary rates of X-linked and autosomal genes in *Drosophila*[30,31]. Using known DNA sequences of *D.melanogaster* and *D.simulans*, they estimated non-synonymous (dN) and synonymous divergence (dS) for tens of genes associated either with X or autosomal chromosomes. However, the comparison of their mean divergences with a t-test did not support the hypothesis, as the authors found "no difference in the rate of amino acid substitutions between X-linked and autosomal loci (using dN or dN/dS)"[31]. Similar results for dS were obtained elsewhere using a rank test[30], and the authors concluded: "we find (...) no difference between arms [X chromosome vs. right arm of the 3rd autosome - ed.] for silent divergence between species". This result led to the conclusion that "there is no evidence for faster X evolution, at least in the present dataset"[31]. This observation seemed also incoherent with a previous finding that X-linked paralogs have "significantly higher" rates of divergence (i.e., dN/dS) than other gene duplicates in *D.melanogaster*[32]. The authors ascribed the observed disparity to the differences in the distribution of dominance effects between single-copy and duplicate genes[31]. At the same time, however, they emphasised that their dataset might be unrepresentative, and a faster X effect might be revealed, if tested among all potential X-linked and autosomal targets of selection.

Indeed, as soon as the sequences of several *Drosophila* genomes became available, and mean divergence could be compared between sets of genes even approximately 500 times larger than previously, "statistical support for greater divergence of X-linked versus autosomal genes"[33] was finally found by obtaining p-values $< 0.05^*$. To demonstrate how such support could be achieved, we compared previous results[30,31] with genome-wide analysis[33]. To unify the calculations between studies, for all genes we used gene divergence estimates and chromosome associations from[33] (see Supplementary Methods, below).

We calculated the 95% CI for the median dN and dS among X-linked and autosomal genes of *D.melanogaster* that were analysed by each of the reference studies. As shown in Fig. 5, main

[*] ... or even much less. One of the reported p-values in this work equalled 1.8e-275– the smallest we have ever seen. This value is incomprehensible: the ratio of the Planck length to the Observable Universe diameter is only 1.8e-62. However, neither p-value size nor precision of its estimate guarantee a biologically significant effect. A small p-value only indicates that there is an extremely strong evidence for the sign of the observed difference or correlation.

text, and S4 Fig, with the increase in the sample sizes in the genome-wide study[33], the CIs for median X-autosomal divergence differences narrow and finally most become statistically significant (panels b). Panels c show 95% CIs for the contrasts between X-autosome divergence differences calculated for genes from conflicting studies. In this analysis both types of inter-studies differences were considered, namely: the genome-wide study[33] vs. smaller studies[30,31] and *vice versa*, smaller studies vs. the genome-wide study, but the given data did not allow us to identify any discrepancy between studies (i.e., the sign of contrasts could not be stated).

Therefore, as far as statistical significance is the yardstick, these studies lead to opposite conclusions, but actually no conflict between their results may be claimed. To solve this issue, the question of the X-autosome divergence difference must be reconsidered, e.g., by the ESCI method, which is illustrated in S5 Fig and S6 Fig. For instance, the X-autosome difference for pairwise dN ranges from 0.00022 to 00012. However, when the CI limits are calculated by other tools (bootstrap instead of the Wilcoxon test), statistical significance vanishes. A slight shift of CIs caused by different statistical approaches is not surprising but may dramatically change the NHST-based conclusions, especially when the biological significance of an effect is ignored. Note also that in case of dS, lineage-specific median divergence is larger in X-linked genes, and the difference ranges from 0.0015 to 0.006 (bootstrap CIs). A similar size difference was found for the median pairwise dS (0.0013–0.0058), but in this case the autosomal genes have higher dS. This lack of coherence may suggest that the observed X-autosomal divergence differences are fluctuations that manifested as statistically significant after the increase of sample sizes. The only way to prove that X-linked genes evolve faster than autosomal loci in *Drosophila* is to demonstrate that the observed X-autosomal divergence difference is positive and large enough to have biological and evolutionary consequences. We suspect that this question may be difficult even for population genetics experts, as direct interpretation of divergence values does not seem to be a common practice. In fact, none of papers we discuss attempted to pose this question.

To overcome this problem, we compared the obtained X-autosomes dN and dS differences with a control, i.e., divergence differences between several subsets of autosomal genes that are not assumed to have significant consequences on evolution and population genetics in *Drosophila*. Following the steps from[33], we performed the analysis of lineage specific divergence separately for three *Drosophila* species, and pairwise divergence between *D.melanogaster* and *D.simulans*. Based on their original data, we calculated 95% CIs for the weighted mean divergence of four subsets of autosomal genes associated with chromosome 2, chromosome 3, the left arms of chromosomes 2 and 3, and the right arms of chromosomes 2 and 3 (S5 Fig and S6 Fig). Next, we calculated the 95% CIs for the median divergence differences between autosomes 2 and 3 (referred to as "inter-autosomal difference") and between the left and right arms of autosomes (referred to as "intra-autosomal difference"), as shown in panels b. The choice of weighted means for panels a and medians for panels b was made with respect to the methodology applied to generate Tables S1 and S3 by[33]. Finally, panels c show 95% CIs for the contrasts between X-autosome and both types of autosomal divergence differences. Again, the biological significance of these values should be taken into consideration. Nevertheless, in this particular situation, a distinction between smaller and larger divergence differences is sufficient for our purposes, i.e., the elimination of effects (X-autosome divergence differences) indistinguishable from naturally-occurring divergence fluctuations within autosomal loci. More precisely, we recognise a biological significance of an effect as indemonstrable if it cannot be distinguished from at least one of the inter- or intra-autosomes divergence differences calculated within the same set of genes. All possible types of inter- and intra-autosomes differences are considered, namely 2nd autosome vs. 3rd, 3rd vs. 2nd, left (L) vs. right (R) autosomal arms, and R vs. L.

Hence, for the three analysed species of *Drosophila*, and both pairwise and lineage-specific divergence, dN differences between X-linked and autosomal genes cannot be distinguished from differences observed between genes located on two autosomes or the left and right autosomal arms, for which a hypothesis of faster evolution has never been proposed (S6 Fig). The same observation holds for X-autosomal dS differences, even though some of them are statistically significant (S5 Fig)[†].

Except the nonsynonymous and synonymous sites in coding sequences, the reference study[33] compared sequence divergence in four additional genetic elements. Following their methodology, we prepared four more figures, analogous to S5 Fig, with similar analysis for introns, intergenic regions, 3’ and 5’ UTRs.

[†] Note that despite smaller sample sizes, all lineage-specific dS differences between the 2nd and 3rd autosome shown in this figure are also statistically significant. Maybe population genetics should revise its hypotheses to include one of faster evolution of even chromosomes? Even if it does not make biological sense, we may claim that such a hypothesis has been well “statistically supported”.

For introns (S7 Fig) and intergenic regions (S8 Fig), the X-autosome divergence differences are again indistinguishable from inter- and intra-autosomal differences. The only exception is the case of lineage-specific divergence in *D.melanogaster* introns. Not only is the X-autosome divergence difference negative, but it is also larger (in absolute value) than any of the inter- or intra-autosomal differences. This finding suggests the possibility of a biologically relevant effect, but of faster evolution of autosomal introns. As we are aware that the biological significance threshold used in our study is not very demanding and that such an outcome is limited to only one case, we shall not discuss its evolutionary consequences. Consequently, it seems that only evidence from the UTR case may support the faster-X evolution hypothesis. As shown in S9 Fig and S10 Fig, besides the lineage-specific divergence in *D.melanogaster*, all X-autosomal divergence differences are positive and at least somewhat larger than inter- and intra- autosomal differences. Although it still does not prove that the X-autosome divergence difference in UTRs is biologically important, we may at least discuss this possibility.

Finally, we compared the above results with conclusions reached previously[33] after the analysis of the same data. The authors summarised their findings as follows: "(...) of the 18 lineage divergence estimates (six site types and three lineages) only one, *D.simulans* synonymous sites, failed to show faster-X evolution (Table 1)." The reference table shows that, indeed, there are 17 positive X-autosome differences in divergence weighted means, more precisely, in their point estimates. Surprisingly, their next sentence notes "not all classes of site/lineages showed statistically significant faster X-evolution". The results of their rank tests and obtained p-values are presented schematically in Table S3[33]. Visual inspection immediately provided two cases of “failure” (marked as X < A and indicating negative X-autosome divergence difference). Additionally, we found four cases of statistically non-significant X-autosome divergence difference that were misleadingly marked as X = A. We distinctly disagree with the authors and claim that such cases also do not show faster-X evolution, as it is not possible to recognise the difference as positive if measurements are not sufficiently precise to state its sign. Among the 12 remaining lineage-specific comparisons, 7 show X-autosome divergence differences indistinguishable from divergence fluctuations observed within and between autosomes. Upon recalculation, the spectacular outcome of 17 of the 18 cases confirming the faster-X evolution is reduced to 5 of 18 cases and is limited only to the UTRs. Similarly, from 4 out of 6 confirming comparisons of pairwise divergence, only 2 remained after our revision. This reduction was achieved solely by replacing p-values with CI calculations and applying one of the weakest criteria for the elimination of biologically insignificant results.

Due to the enormous size of the reference study[33], we could not repeat all its analyses. Nevertheless, even our limited revision revealed that the majority of the analysed statistically significant outcomes could not be demonstrated to have biologically relevant effects. These results are further utilised as support for generalisations such as “the X chromosome (...) evolves faster” or “faster-X effect is likely to be general for *Drosophila* but vary in magnitude across lineages and site types”[33], even though its magnitude is indistinguishable from the magnitude of the “faster-2nd-autosome effect” or “faster-left-autosomal-arms effect” for most studied gene elements, including coding sequences, introns and intergenic regions. Even the promising case of UTRs requires further examination because the biological significance threshold we applied was one of the lowest possible. Despite its impact and over 200 citations, the genome-wide study[33] did not eliminate the faster-X effect debate[34–36]. Instead, it is further fuelled by new NHST-based conclusions. It seems that as long as measurement precision estimates and quantitative interpretations are omitted, many similar conflicts are likely to emerge and nestle in biological sciences, especially now when the analysis of large samples– which facilitates the obtaining of statistically significant outcomes– is easier than ever.

Supplementary Methods

Case study I: Protein translation efficiency determinants.

Coding sequences. The sequences of GFP constructs were taken from[9]. The sequences of *E.coli* and *S.cerevisiae* genes were taken from the same sources as in[10], i.e., from the National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/ftp/), accessed November 2012, and from[37], respectively.

Translation efficiency. Translation efficiency of 423 GFP constructs gauged by fluorescence and the mRNA levels of 79 GFP constructs gauged by Northern Blotting, were taken from[9]. As in the reference study[10], protein and mRNA abundances for *E.coli* genes were taken from[21]. Similarly, yeast protein abundances were taken from[18] (referred to as “Ne”), and yeast mRNA levels were taken from[19] (although they originally come from[38], referred to as “Wh”).

Alternative correlations. Alternative correlations were calculated taking protein levels of yeast genes from the following studies: “Gh”[20], “Gr”[22], and “Lu”[21] (YEPD medium). The mRNA levels were taken from: “Ar”[23], “In” [25], “Ma”[24], and “Wm”[19] (micro-array experiment). The two letter codes are the same as those used in Fig. 3, main text, and S1 Fig and S2 Fig to refer to these studies. In case of the “In” study[25], the relative values of mRNA levels gauged by RNAseq were transformed to the mRNA copy number in the cell, as shown by[39]. The remaining studies provide information on absolute abundances and did not require further curation.

Codon Bias. The CAI was computed using the seqinr R package[40]. The tAI was computed as shown in[10]. The values of the relative adaptiveness of a codon (w) required to calculate tAI were taken directly from Table S2[10].

Correlations, correlations differences and contrasts. All correlations reported in our analysis are the non-parametric Spearman correlations. The 95% CIs for correlation coefficients were calculated using standard tools from the R environment. Correlation differences CIs were calculated with help of cocor[41] and bootES[42] R packages, and verified using our own ad-hoc written Fortran programs, implementing both standard and percentile bootstrap[43].

The correlation differences CIs presented in the main text Fig. 2b (solid line), in the bottom panels of Fig. 3 (main text), S1 Fig, and S2 Fig were calculated by the cocor.indep.groups function, zou2007 method[44]. As the analysed sets of GFP, Yeast and *E.coli* genes do not intersect, correlations within them may be treated as independent. The same method was used to calculate CIs presented in the S3 Fig (panels b and d). The bootES[42] R package was used to calculate correlation difference CIs presented in the main text Fig. 2b (dashed lines).

The correlation difference CIs presented in the main text Fig. 3b were calculated by the cocor.dep.groups.overlap function from the cocor package[41], using the zou2007 method[44]. This function is used to compare dependent (correlated) correlations (i.e., those calculated over the same sample), that have a common variable (i.e., codon bias gauged by tAI). Note, however, that the original and alternative sets of genes overlap only partially, which raises doubts whether correlations within them should be treated as completely dependent. Although performing the analysis over the set of 303 common genes fulfils the definition of dependence, it strongly reduces the sample size and widens the obtained CIs. For this reason, we decided to repeat the calculations of correlations and correlation differences separately for each of the alternative studies. We created 19 sets of genes common for the original study and each of the 19 alternative studies used in our analysis. For each set we calculated the 95% CIs for correlation between codon bias and translation efficiency. Next, we estimated correlation differences between each of these sets and the original study with help of the cocor package[41]. If codon bias in alternative correlations was gauged by tAI, the cocor.dep.group.overlap was used; if codon bias was gauged by CAI, i.e., there was no variable in common, the cocor.dep.group.nonoverlap function was applied. As seen from the S1 Fig, such relaxation of constrains does not affect the conclusions drawn from the main text Fig. 3. This observation also holds, when all compared correlations within original and alternative sets are treated as completely independent and their differences are estimated by the cocor.indep.group function[41] (S2 Fig).

The CIs for correlation differences contrasts were calculated using our own, ad-hoc written Fortran programs, employing IMSL subroutines and implementing both standard and percentile bootstrap[43], and treating correlations as independent or dependent as above. Since computing contrasts between correlations is quite non-standard, we paralelly computed CIs for contrasts of z-transformed correlations to test them for non-zero values of contrasts considered. The results obtained from all these alternative variants (not shown) were very similar and the overall picture remained the same.

Case study II: PTPRC (CD45) association with the development of multiple sclerosis.

Odds ratio and relative odds ratio. The numbers of MS patients and controls, as well as the numbers of 77G carriers were taken directly from the reference works of [26–28]. The 95% CIs for odds ratios and relative odds ratios were found by standard and percentile bootstrap[43] using our own, ad-hoc written Fortran programs. Results for odds ratios were checked against those from SAS/FREQ procedure employing the exact method[45] and were close to them, making no difference in the overall picture.

Case study III: Divergence of X-linked and autosomal genes in *Drosophila*.

Data. The lists of *D.melanogaster* genes analysed by[30,31] were taken from the reference studies and mapped to the data set of gene divergence estimates from[33]. Due to annotation changes and nomenclature inconsistencies, only the subsets of genes analysed previously were found in the divergence data set, diminishing the size of the original set of[30] from 40 to 27 genes, and the set of[31] from 245 to 164 genes. One additional gene was removed from the latter due to disagreement of the X-autosome allocation between studies.

Divergence comparisons. For Fig. 5 (main text) and S4-S10 Fig, the 95% CIs for median/weighted mean divergence (panels a), median divergence differences (panels b), and divergence difference contrasts (panels c) were obtained using standard and percentile bootstrap[43], implemented in our own, ad-hoc written Fortran programs. Whenever possible, their results were checked against calculations performed with help of a simpleboot R package[46], with the CI limits gauged by both normal approximation and a percentile method. All calculations variants returned similar results (for larger samples, almost identical), and thus only the results of the percentile bootstrap are shown in the figures. For easier comparison with the reference studies, the 95% CIs for median divergence differences were also calculated with help of a wilcox.test method from the stats R package and are shown in panels b.

References

1. Jennions MD, Møller AP (2003) A survey of the statistical power of research in behavioral ecology and animal behavior. Behav Ecol 14: 438–445.
2. Smith DR, Hardy ICW, Gammell MP (2011) Power rangers: no improvement in the statistical power of analyses published in Animal Behaviour. Anim Behav 81: 347–352.

3. Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, et al. (2013) Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci 14: 365–376.

4. Zeggini E, Ioannidis JP (2009) Meta-analysis in genome-wide association studies. Pharmacogenomics 10: 191–201.

5. Turner RM, Bird SM, Higgins JPT (2013) The impact of study size on meta-analyses: examination of underpowered studies in Cochrane reviews. PLoS One 8: doi:10.1371/journal.pone.0059202.

6. Ikemura T (1981) Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. J Mol Biol 151: 389–409.

7. Grosjean H, Fiers W (1982) Preferential codon usage in prokaryotic genes: the optimal codon-anticodon interaction energy and the selective codon usage in efficiently expressed genes. Gene 18: 199–209.

8. Plotkin JB, Kudla G (2011) Synonymous but not the same: the causes and consequences of codon bias. Nat Rev Genet 12: 32–42.

9. Kudla G, Murray AW, Tollervey D, Plotkin JB (2009) Coding-sequence determinants of gene expression in *Escherichia coli*. Science 324: 255–258.

10. Tuller T, Waldman YY, Kupiec M, Ruppin E (2010) Translation efficiency is determined by both codon bias and folding energy. Proc Natl Acad Sci USA 107: 3645–50.

11. Gingold H, Pilpel Y (2011) Determinants of translation efficiency and accuracy. Mol Syst Biol 7: 481.

12. Angov E (2011) Codon usage: nature's roadmap to expression and folding of proteins. Biotechnol J 6: 650–9.

13. Supek F, Šmuc T (2010) On relevance of codon usage to expression of synthetic and natural genes in *Escherichia coli*. Genetics 185: 1129–34.

14. Agashe D, Martinez-Gomez NC, Drummond DA, Marx CJ (2013) Good codons, bad transcript: large reductions in gene expression and fitness arising from synonymous mutations in a key enzyme. Mol Biol Evol 30: 549–60.

15. Welch M, Govindarajan S, Ness JE, Villalobos A, Gurney A, et al. (2009) Design parameters to control synthetic gene expression in *Escherichia coli*. PLoS ONE 4: doi:10.1371/journal.pone.0007002.

16. Sharp PM, Li WH (1987) The codon Adaptation Index–a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 15: 1281–95.

17. dos Reis M, Savva R, Wernisch L (2004) Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res 32: 5036–44.

18. Newman JRS, Ghaemmaghami S, Ihmels J, Breslow DK, Noble M, et al. (2006) Single-cell proteomic analysis of *S. cerevisiae* reveals the architecture of biological noise. Nature 441: 840–6.

19. Wang Y, Liu CL, Storey JD, Tibshirani RJ, Herschlag D, et al. (2002) Precision and functional specificity in mRNA decay. Proc Natl Acad Sci U S A 99: 5860–5.

20. Ghaemmaghami S, Huh WK, Bower K, Howson RW, Belle A, et al. (2003) Global analysis of protein expression in yeast. Nature 425: 737–41.

21. Lu P, Vogel C, Wang R, Yao X, Marcotte EM (2007) Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. Nat Biotechnol 25: 117–24.

22. Greenbaum D, Colangelo C, Williams K, Gerstein M (2003) Comparing protein abundance and mRNA expression levels on a genomic scale. Genome Biol 4: 117.

23. Arava Y, Wang Y, Storey JD, Liu CL, Brown PO, et al. (2003) Genome-wide analysis of mRNA translation profiles in *Saccharomyces cerevisiae*. Proc Natl Acad Sci USA 100: 3889–94.

24. Garcia-Martinez J, Aranda A, Perez-Ortin J (2004) Genomic run-on evaluates transcription rates for all yeast genes and identifies gene regulatory mechanisms. Mol Cell 15: 303–313.

25. Ingolia N, Ghaemmaghami S, Newman J, Weissman J (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science 324: 218–23.

26. Jacobsen M, Schweer D, Ziegler A, Gaber R, Schock S, et al. (2000) A point mutation in PTPRC is associated with the development of multiple sclerosis. Nat Genet 26: 495–9.

27. Vorechovsky I, Kralovicova J, Tchilian E, Masterman T, Zhang Z, et al. (2001) Does 77C→G in PTPRC modify autoimmune disorders linked to the major histocompatibility locus? Nat Genet 29: 22–3.

28. Barcellos LF, Caillier S, Dragone L, Elder M, Vittinghoff E, et al. (2001) PTPRC (CD45) is not associated with the development of multiple sclerosis in U.S. patients. Nat Genet 29: 23–4.

29. Charlesworth B, Coyne JA, Barton NH (1987) The relative rates of evolution of sex chromosomes and autosomes. Am Nat 130: 113–146.

30. Begun DJ, Whitley P (2000) Reduced X-linked nucleotide polymorphism in *Drosophila simulans.* Proc Natl Acad Sci USA 97: 5960–5.

31. Betancourt AJ, Presgraves DC, Swanson WJ (2002) A test for faster X evolution in *Drosophila*. Mol Biol Evol 19: 1816–1819.

32. Thornton K, Long M (2002) Rapid divergence of gene duplicates on the *Drosophila melanogaster* X chromosome. Mol Biol Evol 19: 918–25.

33. Begun DJ, Holloway AK, Stevens K, Hillier LW, Poh Y, et al. (2007) Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. PLoS Biol 5: doi:10.1371/journal.pbio.0050310.

34. Vicoso B, Haddrill PR, Charlesworth B (2008) A multispecies approach for comparing sequence evolution of X-linked and autosomal sites in *Drosophila*. Genet Res (Camb) 90: 421–31.

35. Hu TT, Eisen MB, Thornton KR, Andolfatto P (2013) A second-generation assembly of the *Drosophila simulans* genome provides new insights into patterns of lineage-specific divergence. Genome Res 23: 89–98.

36. Meisel RP, Connallon T (2013) The faster-X effect: integrating theory and data. Trends Genet 29: 537–44.

37. Man O, Pilpel Y (2007) Differential translation efficiency of orthologous genes is involved in phenotypic divergence of yeast species. Nat Genet 39: 415–21.

38. Holstege FC, Jennings EG, Wyrick JJ, Lee TI, Hengartner CJ, et al. (1998) Dissecting the regulatory circuitry of a eukaryotic genome. Cell 95: 717–28.

39. Siwiak M, Zielenkiewicz P (2010) A comprehensive, quantitative, and genome-wide model of translation. PLoS Comput Biol 6: doi:10.1371/journal.pcbi.1000865.

40. Charif D, Lobry J (2007) SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In: Bastolla U, Porto M, Roman E, Vendruscolo M, editors, Structural Approaches to Sequence Evolution: Molecules, Networks, Populations, New York City, USA: Springer Verlag. pp. 207–232.

41. Diedenhofen B (2013) cocor: Comparing correlations. URL `http://r.birkdiedenhofen.de/pckg/cocor/`. (Version 0.01-4).

42. Kirby KN, Gerlanc D (2013) BootES: An R package for bootstrap confidence intervals on effect sizes. Behav Res Methods 45: 905–927.

43. Manly B (1997) Randomization, Bootstrap and Monte Carlo Methods in Biology. London, UK: Chapman & Hall.

44. Zou GY (2007) Toward using confidence intervals to compare correlations. Psychol Methods 12: 399–413.

45. SAS Institute Inc (2013) SAS/STAT 12.3 User's Guide. URL `http://support.sas.com/documentation`.

46. Peng RD (2008) simpleboot: Simple bootstrap routines. URL `http://cran.r-project.org/web/packages/simpleboot/index.html`. (Version 1.1-3).