Are controversies in science driven by randomness and misbeliefs?

Marlena Siwiak Corresp., [1] , **Tomasz Wyszomirski** [2] , **Piotr Zielenkiewicz** Corresp. [1, 3]

[1] Institute of Biochemistry and Biophysics, Polish Academy of Sciences, Warsaw, Poland

[2] Faculty of Biology, Biological and Chemical Research Centre, University of Warsaw, Warsaw, Poland

[3] Faculty of Biology, Laboratory of Plant Molecular Biology, University of Warsaw, Warsaw, Poland

Corresponding Authors: Marlena Siwiak, Piotr Zielenkiewicz
Email address: marlena@ibb.waw.pl, piotr@ibb.waw.pl

Recent years have brought about the realisation of an irreproducibility crisis in science, which may have numerous causes, including common standards of statistical analysis. For decades, the methodological paradigm of null hypothesis significance testing (NHST) has remained under harsh, yet rather ineffective criticism. Here, we show that the vast majority of contradictions between the results of distinct studies may be fictitious, resulting from misbeliefs about NHST. To exemplify how they appear, we provide extensive reanalyses of results from high-profile literature and reveal statistical uncertainties that customarily remained obscured by the NHST paradigm. Widespread awareness of these uncertainties accompanied with quantitative interpretation of the results is the first step in assessing the actual scale of the irreproducibility problem and eradicating it.

Are controversies in science driven by randomness and misbeliefs?

Marlena Siwiak[1*], Tomasz Wyszomirski[2*], and Piotr Zielenkiewicz[1,3]✉

[1]**Institute of Biochemistry and Biophysics, Polish Academy of Sciences, Warsaw, Poland**
[2]**University of Warsaw, Faculty of Biology, Biological and Chemical Research Centre, Warsaw, Poland**
[3]**University of Warsaw, Faculty of Biology, Laboratory of Plant Molecular Biology, Warsaw, Poland**
[*]**These authors contributed equally to this work.**
✉**To whom correspondence should be addressed: piotr@ibb.waw.pl**

ABSTRACT

Recent years have brought about the realisation of an irreproducibility crisis in science, which may have numerous causes, including common standards of statistical analysis. For decades, the methodological paradigm of null hypothesis significance testing (NHST) has remained under harsh, yet rather ineffective criticism. Here, we show that the vast majority of contradictions between the results of distinct studies may be fictitious, resulting from misbeliefs about NHST. To exemplify how they appear, we provide extensive reanalyses of results from high-profile literature and reveal statistical uncertainties that customarily remained obscured by the NHST paradigm. Widespread awareness of these uncertainties accompanied with quantitative interpretation of the results is the first step in assessing the actual scale of the irreproducibility problem and eradicating it.

Keywords: statistics, reproducibility, reproducible-research, uncertainty, null-hypothesis-significance-testing, confidence-intervals

INTRODUCTION

Science is said to study reproducible phenomena. However, in recent years, the recognition that this is not always so has grown, and an "irreproducibility crisis" has been proclaimed (Ioannidis (2005); Lehrer (2010); Nature Publishing Group (2013); Ioannidis (2014); Nature (2016)). In basic science, the term "irreproducibility" is often used in rather vague way, but it essentially refers to replicated experiments in which new results contradict findings of the original. Because practically no replication can be perfect, the criteria for defining a contradiction are of crucial importance. These criteria depend on standards of statistical analyses, which have been proposed as one of possible sources of the irreproducibility crisis (Johnson (2013); Nuzzo (2014); Halsey et al. (2015); Wasserstein and Lazar (2016)).

Null hypothesis significance testing (NHST) is the common currency in many fields of scientific research. Its success is enormous despite harsh critiques over decades that include arguments ranging from logical to ethical (Morrison and Henkel (1970); Harlow et al. (1997); Cumming and Fidler (2011); Fidler (2011); Wasserstein and Lazar (2016)). As a hybrid between disparate statistical approaches (Gigerenzer and Murray (1987)), NHST likely gained its popularity due to over-interpretations. Two of them may be directly responsible for the occurrence of fictitious controversies, as they evoke a false, yet desired (Mullane and Williams (2015)), sense of certainty (Schmidt and Hunter (1997); Cumming (2011)). First, statistically significant results are commonly believed to be "true" in a sense that the true effect size is very close to the point estimate computed from the sample data, and observed effects are often automatically assumed to be scientifically relevant, regardless of their magnitude. Second, non-significant results often serve as support for the tested null hypothesis, i.e., the non-existence of an effect (Maxwell (2004)). Together, these two convictions must lead to fictitious contradictions between results of similar studies– fictitious because they are based only on false beliefs and are otherwise unjustified.

We use a simple model to demonstrate that such fictitious cases may constitute the vast majority of all

contradictions (even up to 90%). We also reanalyse data from three separate sets of high-profile papers to exemplify in detail how fictitious debates emerge in scientific practice. Our intent is neither to settle the subject-matter issues of the traced debates nor to criticise particular papers– they serve merely as examples that might be substituted by many others. To avoid the pitfalls of NHST, we pay attention to effect sizes and adopt thinking in terms of confidence intervals (CIs) (Cumming (2011); Motulsky (2014))– following what we term the ESCI (Effect Size Confidence Interval) approach.

METHODS

We adopted the commonly-used 95% confidence level. CIs for Spearman correlations were found using the standard R environment. For independent and correlated correlations (Zou (2007)), we used cocor (Diedenhofen (2013)) and bootES (Kirby and Gerlanc (2013)) R packages. To prevent relying on "vibration of effects" due to the choice of the details of the statistical method applied (Ioannidis (2008); Button et al. (2013)), a portion of the results was verified using our own ad-hoc written Fortran programs that computed standard and percentile bootstrap (Manly (1997)) CIs. CIs for odds ratios were checked against CIs obtained from SAS/FREQ procedure (SAS Institute Inc. (2013)) using the exact method. To compare studies (as well as different gene expression measurements in Case I), we found CIs for the appropriate differences of differences (e.g., of correlations, medians); we term them "difference contrasts" or "contrasts" for short in the description of the results. For Case II, the same role is played by the ratio of odds ratios, i.e., relative odds ratio (Suzuki (2006)). We computed all the CIs for contrasts with ad-hoc written Fortran programs and we used standard and percentile bootstrap methods (Manly (1997)). Differences between their results were small enough not to change the overall picture and our conclusions. All details on methods and data sources are described in Supplemental Text S1.

RESULTS

Frequency of fictitious contradictions

To estimate the frequency of fictitious contradictions in science, we contrast two ways of comparing results of identical but independent studies that aim to detect some effect in the same population (Fig. 1). In such a model, no genuine contradictions exist. If the correct method to compare studies is used, the controversy appears solely due to random sampling and its frequency equals to the significance level adopted.

However, if both studies use NHST as the sole benchmark, the comparison is commonly, yet incorrectly, performed by contrasting statistically significant versus non-significant results. In such a case, the vast majority of all contradictions that arise may be fictitious (see below). All of these false controversies are avoided when correct methods of comparison are applied.

For example, assuming $\alpha = 0.05$, 5% of comparisons between studies will yield significant differences between those studies due to random sampling. If the power of each study equals 0.5, in 50% of cases, one study reports a significant result and the other a non-significant result, which produces a potentially fictitious contradiction. Under such a scenario, fictitious contradictions are about ten times more frequent than those that cannot be avoided. Their proportion of about 90%, presented in Fig. 1, is the highest obtainable. However, even very cautious estimations (see Supplemental Text S1) show that for $\alpha = 0.05$ and statistical power between 0.09 and 0.91 this proportion is still higher than 50%. Since in practice of many disciplines statistical power is usually within this range (Jennions and Møller (2003); Smith et al. (2011); Button et al. (2013); Zeggini and Ioannidis (2009); Turner et al. (2013)), and fictitious contradictions may happen quite often also for more extreme values of power, the problem is serious and has been spotted already (Halsey et al. (2015)).

This high frequency results entirely from an erroneously-supposed importance of the difference between statistically significant and non-significant results (Gelman and Stern (2006)). Within the NHST framework, the only way to avoid this problem is to suspend judgement in cases of statistically non-significant results by allowing for a "don't know" category (Dempster (2008)). This leads, however, to the loss of information present in the analysed sample. In contrast, the ESCI framework preserves information and prevents fictitious contradictions at the same time, as illustrated below.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2249v1 | CC BY 4.0 Open Access | rec: 5 Jul 2016, publ: 5 Jul 2016

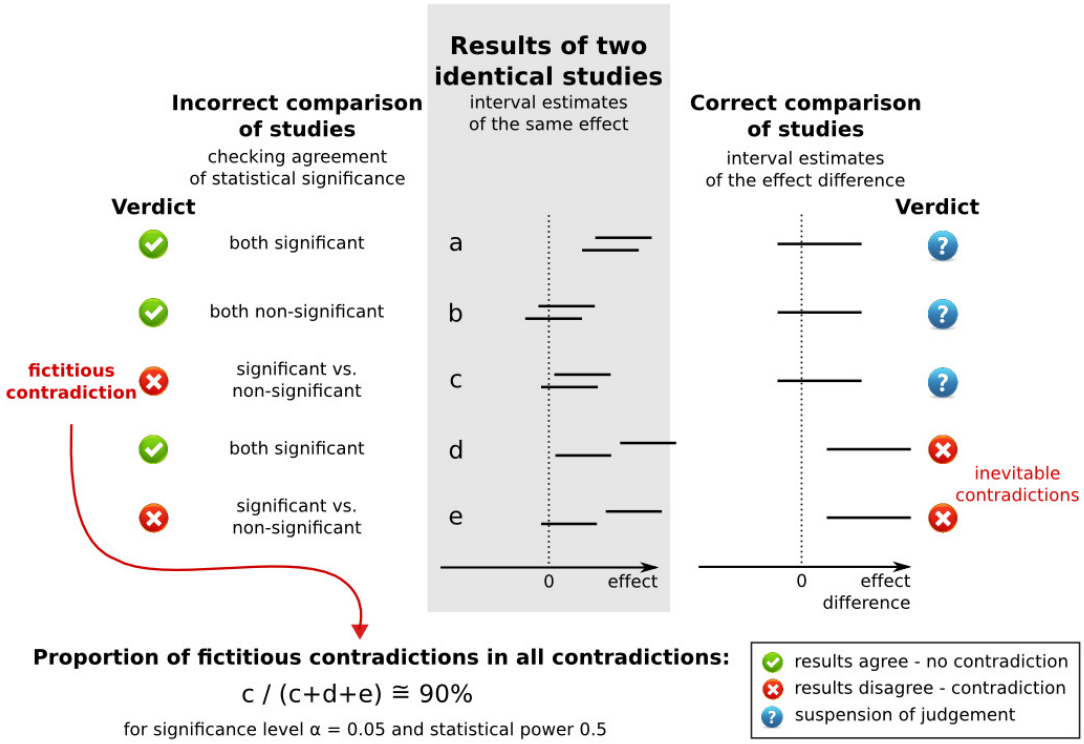

Figure 1. Schematic illustration of correct and incorrect comparisons between two identical studies. The correct comparison points to the ESCI approach. Replacing it with tests, even proper, would bring about the risk of misunderstandings. The incorrect comparison points to absurdities resulting from the NHST approach. As the exemplary results of calculations show, the proportion of fictitious contradictions may be astonishingly high (for derivation and details see Supplemental Text S1.

Case study I: Protein translation efficiency determinants

The well-established explanation of biased codon usage states that it increases the efficiency and accuracy of translation (Ikemura (1981); Grosjean and Fiers (1982); Plotkin and Kudla (2011)). However, the expression analysis of 158 green fluorescent protein (GFP) synonymous sequences (Kudla et al. (2009)) revealed that codon bias did not have "significant effects" on protein levels. This conclusion was achieved mainly by obtaining a statistically non-significant correlation between GFP's expression levels (as gauged by their fluorescence) and codon bias. In response, a related study (Tuller et al. (2010)) on endogenous genes of *E.coli* and *S.cerevisiae* reported a statistically significant association between codon bias and expression (measured by protein abundance normalised to mRNA level) and concluded that codon bias is an important determinant of translation efficiency. The discrepancy of results was attributed to differences in mRNA's folding energies of synthetic and endogenous genes (Tuller et al. (2010)).

Our analysis demonstrated, however, that due to the smaller number of analysed GFP sequences, CIs for their correlations between codon bias and expression are much wider than for endogenous genes (Fig. 2A), which may be the main cause of statistically non-significant results obtained in the reference study (Kudla et al. (2009)). Additionally, for some GFP–yeast comparisons the hypothesis that correlations are identical for both types of genes cannot be rejected (Fig. 2B); thus, no discrepancy between studies can be declared. In the juxtaposition of *E.coli* with GFP sequences, the true correlation for bacteria was larger by at least 0.03 (Fig. 2B). To analyse the relevance of this effect, we created controlled sets of yeast genes and compared within them several analogous correlations of codon bias and gene expression. The latter was measured by different, yet in principle equivalent, experiments (see Fig. 3, and Fig. S1 – Fig. S2 for other selections of gene subsets). Thus, within identical sets of genes, we obtained alternative correlations for the same variables that appear strikingly dissimilar (Fig. 3B). Approximately half of the

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2249v1 | CC BY 4.0 Open Access | rec: 5 Jul 2016, publ: 5 Jul 2016

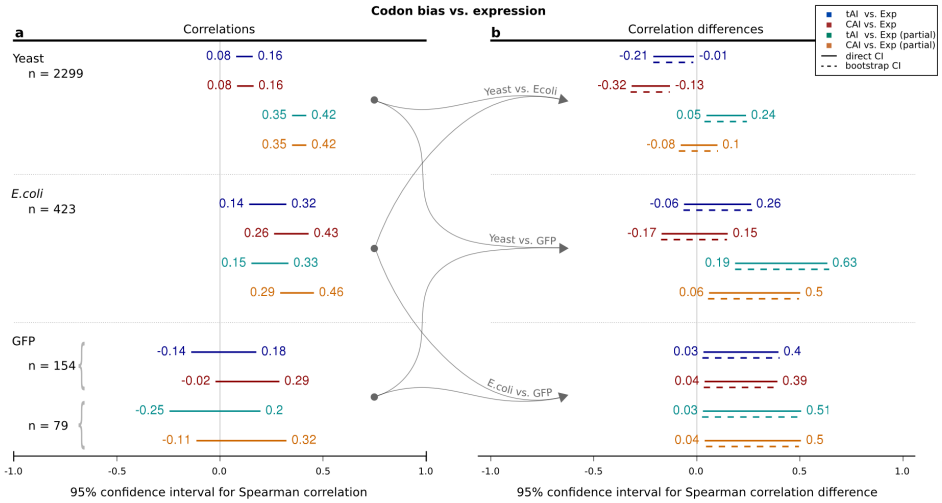

Figure 2. Codon bias vs. expression in three sets of genes. a, 95% CI for Spearman correlation coefficients between codon bias and expression for three sets of genes from Kudla et al. (2009) (GFP), and Tuller et al. (2010) (Yeast and *E.coli*). Correlations are simple or partial (codon bias vs. protein levels when controlled for mRNA levels), n = sample size. **b,** 95% CIs for correlation differences for pairwise comparisons.

alternative correlations differ from the original correlation (calculated as in Tuller et al. (2010)) by at least 0.05, while for the most extreme cases, this difference may be at least as high as 0.32. Thus, either the signs of correlation differences between two contradicting studies by Kudla et al. (2009) and Tuller et al. (2010)– the origin of controversy– cannot be stated, or their sizes make them virtually impossible to distinguish from the noise caused merely by the variability of gene expression measurements in distinct experiments.

To explain this mistakenly-observed discrepancy, one group postulated that folding energy modulates the relation between codon bias and translation efficiency (Tuller et al. (2010)). To demonstrate this claimed phenomenon, sets of analysed *E.coli* and *S.cerevisiae* genes were divided into five equally-sized bins according to the folding energy of their transcripts, a correlation between the codon bias and translation efficiency was computed separately for each bin, and it was observed that the association strength depended on the folding energy levels (Tuller et al. (2010)). These results are reproduced in Fig. S3, but each reported correlation coefficient is supplemented with its CI, and interval estimates of correlation differences for each pair of bins are also provided. The sign of the correlation difference cannot be determined for any single pair of bins, which indicates that these data do not provide evidence that folding energy modulates the association between codon bias and expression. This effect may exist, being too weak to be reliably estimated by the existing means (see Supplemental Text S1 for details).

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

An association between multiple sclerosis (MS) and the 77G allele of the PTPRC gene was claimed on the basis of comparison of the allele frequencies in MS patients and controls (Jacobsen et al. (2000)). In similar studies by Vorechovsky et al. (2001) and Barcellos et al. (2001) no statistically significant difference of allele frequencies was found, and the authors concluded no link between the 77G allele and disease. To explain this discrepancy, we compared all nine pairs of patient groups by computing CIs for the relative odds ratios for bearing the mutated 77G allele (Fig. 4). On this basis, the American and Hannover populations from Jacobsen et al. (2000) and all populations from Vorechovsky et al. (2001) and Barcellos et al. (2001) are practically indistinguishable from each other. Only two Marburg populations from Jacobsen et al. (2000) show a clearly higher odds ratio than the others. This result may indicate a genuine difference between populations, but we have identified another probable cause. In two

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2249v1 | CC BY 4.0 Open Access | rec: 5 Jul 2016, publ: 5 Jul 2016

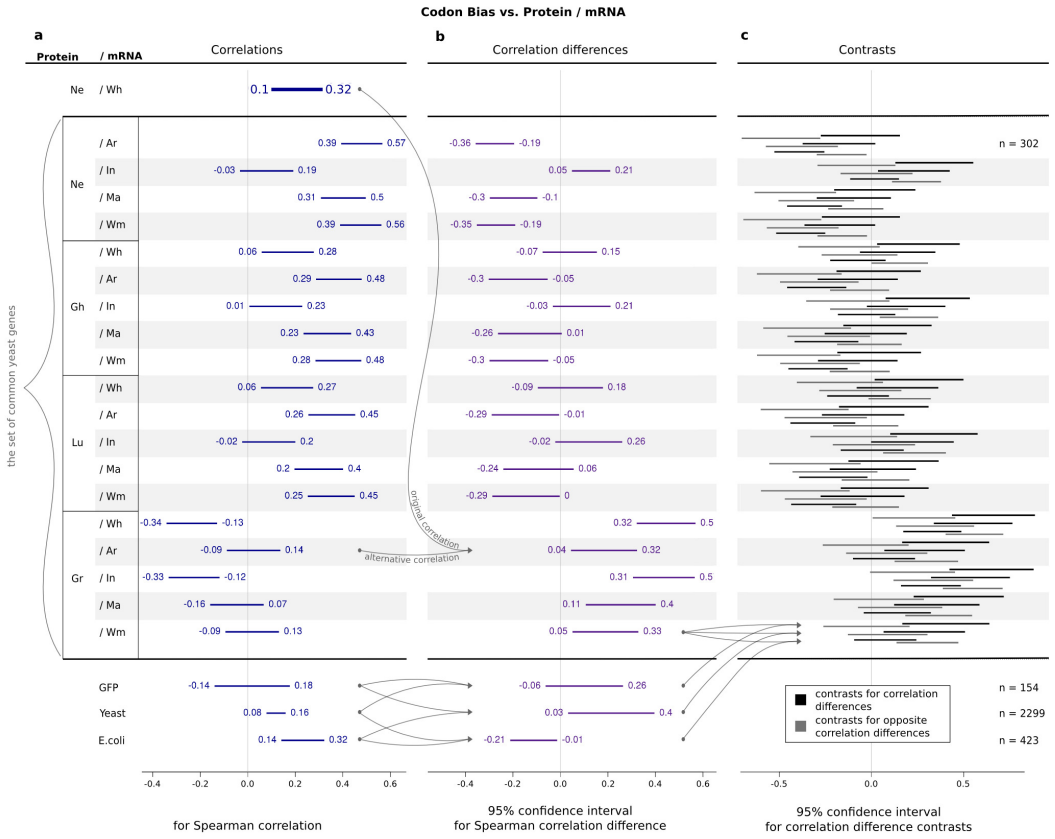

Figure 3. Codon bias vs. expression gauged by different experiments. a, 95% CIs for Spearman correlation coefficients between codon bias (tAI) and gene expression. Main part: correlations for 302 yeast genes with expression defined as a quotient of any possible combination of protein and mRNA abundances gauged from experimental data marked by two-letter shortcuts (see Supplemental Text S1). The Ne/Wh combination was used originally by Tuller et al. (2010). **b,** 95% CIs for correlation differences between the original and alternative correlations derived from other data sources. **c,** 95% CIs for contrasts between each correlation difference shown in the top of panel b and three correlation differences from Fig. 2.

Marburg populations, the recruitment procedure for the control groups was highly restrictive compared to the remaining cases– in particular, it excluded healthy donors with a family history of MS (for whom the probability of having the 77G allele is higher if the tested hypothesis is true). This effect would explain the lack of 77G bearers in control groups of the Marburg studies and could exaggerate the observed association between 77G and disease. Slight contamination of these two control groups with 3 and 5 allele bearers (out of 117 and 194 control patients, respectively) is sufficient to make the observed effect statistically non-significant (Fig. 4, "modified data") and eliminate the observed discrepancy between this (Jacobsen et al. (2000)) and remaining studies by Vorechovsky et al. (2001); Barcellos et al. (2001) (see Supplemental Text S1 for details).

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

Under certain conditions X-chromosome loci are expected to have higher rates of adaptive evolution than those located on the autosomes (Charlesworth et al. (1987)). To test this hypothesis in *Drosophila*, several groups examined the evolutionary rates of X-linked and autosomal genes, and checked whether their average divergence– (synonymous (dS) and non-synonymous (dN))– "differ significantly". Some stated "no difference" in divergence (Betancourt et al. (2002); Begun and Whitley (2000)), while others reported that it was "significantly higher" (Thornton and Long (2002)). These discrepancies were attributed to

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2249v1 | CC BY 4.0 Open Access | rec: 5 Jul 2016, publ: 5 Jul 2016

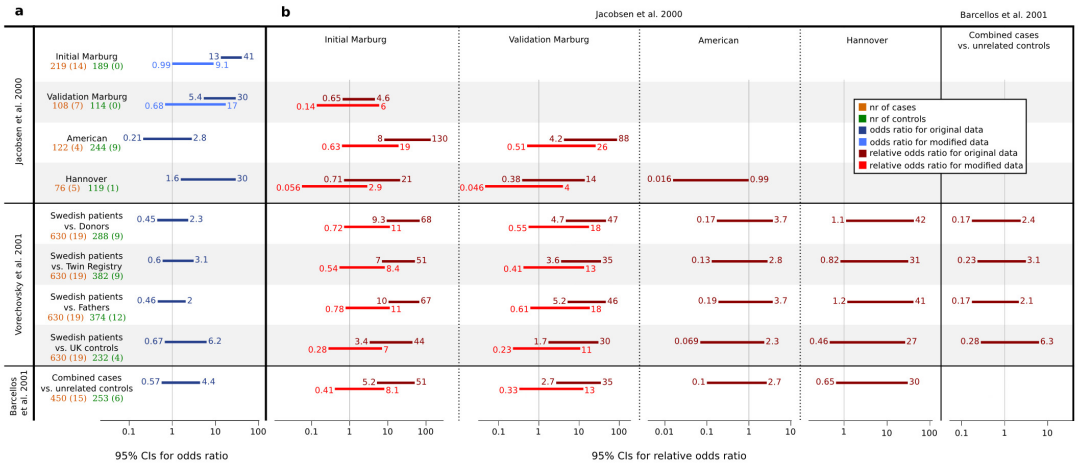

Figure 4. Odds ratios and relative odds ratios for multiple sclerosis. a, 95% CI for the ratio of odds of disease given the 77G allele compared to 77C allele for study groups from three contradicting studies by Jacobsen et al. (2000); Vorechovsky et al. (2001); Barcellos et al. (2001). Numbers of 77G bearers are in brackets. The CIs for modified data (for explanation see text) do not allow us to determine whether disease odds are higher or lower given the 77G allele. **b,** 95% CIs for the relative odds ratios calculated between pairs of studies. When the supposedly-missing carriers are added to Marburg control groups (modified data), the results of studies become indistinguishable.

different types of analysed genes or to unrepresentative sample sizes. Indeed, the hypothesis of faster evolution of sex chromosomes remained disputable until the sequencing of several fly genomes, which enabled testing among samples even 500 times larger (Begun et al. (2007)).

We complemented the results of contradicting studies (Betancourt et al. (2002); Begun and Whitley (2000); Begun et al. (2007)) with CIs for median and median divergence differences between each analysed pair of X-linked and autosomal sets of genes (Fig. 5, Fig. S4). Only with a considerable increase in sample sizes in the genome-wide study of Begun et al. (2007) do the differences between X-linked and autosomal loci become detectable by significance tests. As CIs for appropriate contrasts show (panels c), signs of the differences between results of particular studies cannot be determined. This indicates that, based on these data sets, the compared X-autosome divergence differences from two smaller studies by Betancourt et al. (2002); Begun and Whitley (2000) and one genome-wide study by Begun et al. (2007) cannot be distinguished, and no controversy between them may be claimed.

Regarding the faster-X evolution hypothesis itself, most of the statistically significant X-autosome divergence differences reported by Begun et al. (2007) appear only slightly greater than zero after inspecting CIs. Moreover, some are negative, which suggests the opposite effect: faster autosome evolution. Despite their modest size, the possibility cannot be ruled out that these values may have noticeable biological and evolutionary consequences. For comparison, we examined divergence differences between subsets of autosomal loci (2nd vs. 3rd autosome and left vs. right arms of autosomes 2 and 3), for which any hypothesis concerning differing speeds of evolution has probably never been proposed. For both pairwise and lineage specific dN and dS, the analysed data set does not provide evidence that the faster-X effect is larger than the inter- or intra-autosomal divergence variability (Fig. S5 and Fig. S6). A similar conclusion arises for X-linked and autosomal introns and intergenic regions (Fig. S7 and Fig. S8), with the exception of lineage-specific divergence in *D.melanogaster* introns. For these introns, the X-autosome difference was negative and larger (in absolute value) than any of the inter- and intra-autosomal differences, suggesting the possibility of a biologically-relevant effect of faster evolution of autosomes. UTRs (Fig. S9 and Fig. S10) constitute the only case that may support the faster-X evolution hypothesis. Apart from 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. This result still does not prove that the higher X divergence in UTRs is biologically significant, but at least there is cause to consider the

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2249v1 | CC BY 4.0 Open Access | rec: 5 Jul 2016, publ: 5 Jul 2016

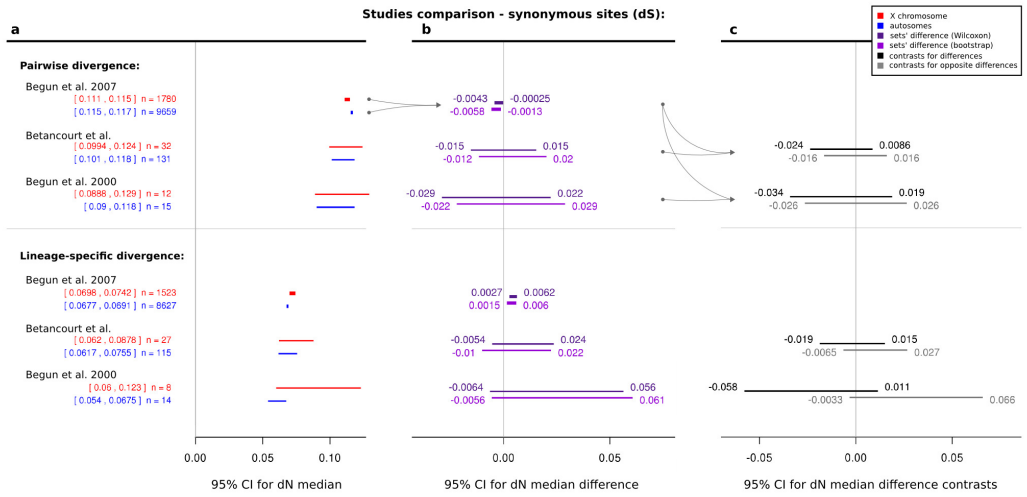

Figure 5. Median dS, median X-autosome dS differences and contrasts in three studies of *Drosophila*. a, 95% CIs for median divergence in dS for X-linked and autosomal loci for three contradicting studies by Begun et al. (2007); Betancourt et al. (2002); Begun and Whitley (2000). Bootstrap CIs are shown in brackets, n = sample size. **b,** 95% CIs for the median dS difference between X-linked and autosomal loci for each research. **c,** 95% CIs for contrasts, i.e., the differences of the X-autosome dS differences for genes analysed by two contradicting studies. Their signs cannot be identified, and thus no controversy between the studies can be claimed.

possibility.

In the 17 of 18 lineage-specific cases that (according to Begun et al. (2007)) confirm the faster-X evolution, only 5 remained after our revision. For pairwise divergence, the success rate dropped from 4 of 6 to 2 of 6 confirming cases. For most gene elements, the magnitude of the "faster-X effect" was indistinguishable from the quite unexpected magnitude of the "faster-2nd-autosome effect" or "faster-left-autosomal-arms effect" (see Supplemental Text S1 for details).

Summary of case studies

All three examples demonstrate that contradictions often disappear when a "significant vs. non-significant" approach is replaced with a proper method (preferably ESCI) of comparing the results of the studies; therefore, these contradictions are fictitious. We stress, however, that if the studies do not differ significantly, it does not mean that they agree perfectly but only that their plausible differences lie within the calculated confidence interval. Additionally, even if the differences of results remain statistically significant, they often cannot be declared large enough in comparison with the variability caused by the details, protocols and approaches of research procedures– a phenomenon called "vibration of effects" (Ioannidis (2008); Button et al. (2013)). The choice of the measurement method (Case I) and the selection of individuals for controls (Case II) are examples. Unexpected variation unrelated to the hypothesis that is tested (Case III) is even more challenging. Focusing on statistical significance alone hides these problems because it draws attention from quantitative questions. In none of the analysed papers does the matter of quantitative importance of an effect appear at the foreground.

Further examples from the literature

Although the prevalence of fictitious controversies stemming from NHST misinterpretations is difficult to assess, our analysis is not isolated. Similar examples, causing serious scientific and practical problems, were described in medicine and psychology (Fidler (2011); McCormack et al. (2013)). Two studies (Knape and de Valpine (2012); Osenberg et al. (2002)) concerning key issues in ecology stress the importance of the uncertainties acknowledgement, one of them (Osenberg et al. (2002)) arrives at conclusions strictly

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2249v1 | CC BY 4.0 Open Access | rec: 5 Jul 2016, publ: 5 Jul 2016

consonant with ours. In genetics, a recent critique of the ENCODE project by Graur et al. (2013) demonstrates how "absurd conclusion" may be reached when statistical significance is exalted above the magnitude of the effect.

The latest examples come from the recent effort of Open Science Collaboration to reproduce 100 psychological findings (Collaboration (2015)). One of the criterion used to evaluate replication was an "intuitively appealing" and "consistent with common heuristics" test whether the replication shows a statistically significant (p-value < 0.05) effect with the same direction as the original study, however, some attempts to involve the ESCI framework were also made. The authors came to conclusion that only 36-47% of the original studies could be successfully replicated. A debate emerged soon after, whether this number is (Gilbert et al. (2016)) or is not (Anderson et al. (2016)) in agreement with by-chance-alone expectations, and does it justify the proclamation of the reproducibility crisis in psychology (Maxwell et al. (2015); Baker (2016)). Similarly confusing indicators of replicability were used by Camerer et al. (2016) to evaluate the laboratory experiments in economics. As a result, a significant effect in the same direction as in the original study was found in only 61% of experiment replications, while measures partially based on ESCI methods scored 67-78% of successfully replicated studies. The only indicator that yielded 83-89% was the one that acknowledged sampling variation in both the original and replicated studies by counting how many replicated effects lay in 95% prediction intervals derived from original studies (Leek et al. (2015)). This methodology is close to ours and does not lead to fictitious contradictions, however, it was used rather as a supplementary method and its less exciting results were not even mentioned in the abstract. When the same method was used to evaluate the psychological findings of the Open Science Collaboration (Collaboration (2015)), the fraction of successfull replicates increased from 36-47% to 70-77% (Leek et al. (2015)). All this indicates that replicability is still poorly understood and before proclaiming its new, haunting crisis, we should first define its effective, ESCI-based and ready-to-use measures.

DISCUSSION & CONCLUSIONS

Controversies and debates are at the core of scientific progress, but it is hardly believable that controversies about random events play a constructive role. As long as statistical significance is a benchmark of research results, it also serves as a basis for their comparisons (Miller (2009); Collaboration (2015); Camerer et al. (2016)).

Consequently, fictitious contradictions are inevitable (Leek et al. (2015)) and cannot be differentiated from real contradictions without labour-intensive reanalyses, as exemplified here. Even if the original raw data are not available, approximations of confidence intervals may be obtained by "reverse engineering" from the reported tabular and graphical information. Such work is worth doing, as it would help to estimate the scale of the problem and it would promote awareness of statistical uncertainties and quantitative interpretation of results. Our analysis demonstrates that the question whether separate studies agree or disagree is not a proper one, as it is exposed to the pitfalls of NHST. It is better to ask, following the ESCI framework, about the plausible ranges of differences between results of studies, i.e., to consider the problem in quantitative terms taking statistical uncertainties into account.

ACKNOWLEDGMENTS

We thank Volker Grimm and Marian Siwiak for critical discussions and comments on the manuscript; Geoff Cumming, the author of the acronym ESCI, for permission to assign a new meaning to it; and Grzegorz Kudla for supplementary data access.

REFERENCES

Anderson, C. J., Bahnik, ., Barnett-Cowan, M., Bosco, F. A., Chandler, J., Chartier, C. R., Cheung, F., Christopherson, C. D., Cordes, A., Cremata, E. J., Della Penna, N., Estel, V., Fedor, A., Fitneva, S. A., Frank, M. C., Grange, J. A., Hartshorne, J. K., Hasselman, F., Henninger, F., van der Hulst, M., Jonas, K. J., Lai, C. K., Levitan, C. A., Miller, J. K., Moore, K. S., Meixner, J. M., Munafo, M. R., Neijenhuijs, K. I., Nilsonne, G., Nosek, B. A., Plessow, F., Prenoveau, J. M., Ricker, A. A., Schmidt, K., Spies, J. R., Stieger, S., Strohminger, N., Sullivan, G. B., van Aert, R. C., van Assen, M. A., Vanpaemel, W., Vianello, M., Voracek, M., and Zuni, K. (2016). Response to Comment on "Estimating the reproducibility of psychological science". *Science*, 351(6277):1037.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2249v1 | CC BY 4.0 Open Access | rec: 5 Jul 2016, publ: 5 Jul 2016

Baker, M. (2016). Psychology's reproducibility problem is exaggerated – say psychologists. *Nature*.

Barcellos, L. F., Caillier, S., Dragone, L., Elder, M., Vittinghoff, E., Bucher, P., Lincoln, R. R., Pericak-Vance, M., Haines, J. L., Weiss, A., Hauser, S. L., and Oksenberg, J. R. (2001). PTPRC (CD45) is not associated with the development of multiple sclerosis in U.S. patients. *Nat. Genet.*, 29(1):23–4.

Begun, D. J., Holloway, A. K., Stevens, K., Hillier, L. W., Poh, Y., Hahn, M. W., Nista, P. M., Jones, C. D., Kern, A. D., Dewey, C. N., Pachter, L., Myers, E., and Langley, C. H. (2007). Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. *PLoS Biol.*, 5(11):doi:10.1371/journal.pbio.0050310.

Begun, D. J. and Whitley, P. (2000). Reduced X-linked nucleotide polymorphism in *Drosophila simulans*. *Proc. Natl. Acad. Sci. U.S.A.*, 97(11):5960–5.

Betancourt, A. J., Presgraves, D. C., and Swanson, W. J. (2002). A test for faster X evolution in *Drosophila*. *Mol. Biol. Evol.*, 19(10):1816–1819.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., and Munafo, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.*, 14(5):365–376.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., and Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280):1433–1436.

Charlesworth, B., Coyne, J. A., and Barton, N. H. (1987). The relative rates of evolution of sex chromosomes and autosomes. *Am. Nat.*, 130(1):113–146.

Collaboration, O. S. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716+.

Cumming, G. (2011). *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis*. Routledge Academic, New York City, USA.

Cumming, G. and Fidler, F. (2011). From hypothesis testing to parameter estimation: An example of evidence-based practice in statistics. In Panter, A. and Sterba, S., editors, *Handbook of Ethics in Quantitative Methodology*, pages 293–312. Routledge, New York City, USA.

Dempster, A. P. (2008). The Dempster-Shafer calculus for statisticians. *Int. J. Approx. Reasoning*, 48(2):365–377.

Diedenhofen, B. (2013). *cocor: Comparing correlations*. (Version 0.01-4).

Fidler, F. (2011). Ethics and statistical reform: lessons from medicine. In Panter, A. and Sterba, S., editors, *Handbook of Ethics in Quantitative Methodology*, pages 445–462. Routledge, New York City, USA.

Gelman, A. and Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *Amer. Statist.*, 60(4):328–331.

Gigerenzer, G. and Murray, D. (1987). *Cognition as Intuitive Statistics*. Lawrence Erlbaum Associates Inc, Mahwah, USA.

Gilbert, D. T., King, G., Pettigrew, S., and Wilson, T. D. (2016). Comment on "Estimating the reproducibility of psychological science". *Science*, 351(6277):1037.

Graur, D., Zheng, Y., Price, N., Azevedo, R. B. R., Zufall, R. A., and Elhaik, E. (2013). On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. *Genome Biol. and Evol.*, 5(3):578–590.

Grosjean, H. and 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(3):199–209.

Halsey, L. G., Curran-Everett, D., Vowler, S. L., and Drummond, G. B. (2015). The fickle P value generates irreproducible results. *Nature Methods*, 12(3):179–185.

Harlow, L., Mulaik, S., and Steiger, J., editors (1997). *What If There Were No Significance Tests?* Lawrence Erlbaum Associates Inc, Mahwah, USA.

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(3):389–409.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Med.*, 2(8):doi:10.1371/journal.pmed.0020124.

Ioannidis, J. P. (2008). Why most discovered true associations are inflated. *Epidemiology*, 19(5):640–648.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2249v1 | CC BY 4.0 Open Access | rec: 5 Jul 2016, publ: 5 Jul 2016

Ioannidis, J. P. A. (2014). How to make more published research true. *PLoS Med.*, 11(10):doi:10.1371/journal.pmed.1001747.

Jacobsen, M., Schweer, D., Ziegler, A., Gaber, R., Schock, S., Schwinzer, R., Wonigeit, K., Lindert, R. B., Kantarci, O., Schaefer-Klein, J., Schipper, H. I., Oertel, W. H., Heidenreich, F., Weinshenker, B. G., Sommer, N., and Hemmer, B. (2000). A point mutation in PTPRC is associated with the development of multiple sclerosis. *Nat. Genet.*, 26(4):495–9.

Jennions, M. D. and Møller, A. P. (2003). A survey of the statistical power of research in behavioral ecology and animal behavior. *Behav. Ecol.*, 14(3):438–445.

Johnson, V. E. (2013). Revised standards for statistical evidence. *Proc. Natl. Acad. Sci. U.S.A.*, 110(48):19313–19317.

Kirby, K. N. and Gerlanc, D. (2013). BootES: An R package for bootstrap confidence intervals on effect sizes. *Behav. Res. Methods.*, 45(4):905–927.

Knape, J. and de Valpine, P. (2012). Are patterns of density dependence in the Global Population Dynamics Database driven by uncertainty about population abundance? *Ecol. Lett.*, 15(1):17–23.

Kudla, G., Murray, A. W., Tollervey, D., and Plotkin, J. B. (2009). Coding-sequence determinants of gene expression in *Escherichia coli*. *Science*, 324(5924):255–258.

Leek, J. T., Patil, P., and Peng, R. D. (2015). A glass half full interpretation of the replicability of psychological science. *arXiv:1509.08968 stat.AP*.

Lehrer, J. (2010). The truth wears off. *The New Yorker*.

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

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: causes, consequences, and remedies. *Psychol. Methods*, 9(2):147–63.

Maxwell, S. E., Lau, M. Y., and Howard, G. S. (2015). Is psychology suffering from a replication crisis? what does "failure to replicate" really mean? *American Psychologist*, 70(6):487–498.

McCormack, J., Vandermeer, B., and Allan, G. M. (2013). How confidence intervals become confusion intervals. *BMC Med. Res. Methodol.*, 13:134.

Miller, J. (2009). What is the probability of replicating a statistically significant effect? *Psychon. Bull. Rev.*, 16(4):617–40.

Morrison, D. and Henkel, R., editors (1970). *The Significance Test Controversy.* Aldine Publ., Chicago, USA.

Motulsky, H. (2014). *Intuitive Biostatistics, 3rd edition*. Oxford University Press, New York City, USA.

Mullane, K. and Williams, M. (2015). Unknown unknowns in biomedical research: does an inability to deal with ambiguity contribute to issues of irreproducibility? *Biochemical Pharmacology*.

Nature (2016). Reality check on reproducibility. *Nature*, 533(7604):437.

Nature Publishing Group (2013). Announcement: Reducing our irreproducibility. *Nature*, 496:398.

Nuzzo, R. (2014). Scientific method: statistical errors. *Nature*, 506(7487):150–152.

Osenberg, C., St Mary, C., R.J., S., Holbrook, S., Chesson, P., and Byrne, B. (2002). Rethinking ecological inference: density dependence in reef fishes. *Ecol. Lett.*, 5:17–23.

Plotkin, J. B. and Kudla, G. (2011). Synonymous but not the same: the causes and consequences of codon bias. *Nat. Rev. Genet.*, 12(1):32–42.

SAS Institute Inc. (2013). *SAS/STAT 12.3 User's Guide.*

Schmidt, F. and Hunter, J. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In Harlow, L., Mulaik, S., and Steiger, J., editors, *What If There Were No Significance Tests?*, pages 37–64. Lawrence Erlbaum Associates.

Smith, D. R., Hardy, I. C. W., and Gammell, M. P. (2011). Power rangers: no improvement in the statistical power of analyses published in Animal Behaviour. *Anim. Behav.*, 81(1):347–352.

Suzuki, S. (2006). Conditional relative odds ratio and comparison of accuracy of diagnostic tests based on 2 x 2 tables. *J. Epidemiol.*, 16(4):145–153.

Thornton, K. and Long, M. (2002). Rapid divergence of gene duplicates on the *Drosophila melanogaster* X chromosome. *Mol. Biol. Evol.*, 19(6):918–25.

Tuller, T., Waldman, Y. Y., Kupiec, M., and Ruppin, E. (2010). Translation efficiency is determined by both codon bias and folding energy. *Proc. Natl. Acad. Sci. U.S.A.*, 107(8):3645–50.

Turner, R. M., Bird, S. M., and Higgins, J. P. T. (2013). The impact of study size on meta-analyses: examination of underpowered studies in Cochrane reviews. *PLoS One*,

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2249v1 | CC BY 4.0 Open Access | rec: 5 Jul 2016, publ: 5 Jul 2016

8(3):doi:10.1371/journal.pone.0059202.

Vorechovsky, I., Kralovicova, J., Tchilian, E., Masterman, T., Zhang, Z., Ferry, B., Misbah, S., Chapel, H., Webster, D., Hellgren, D., Anvret, M., Hillert, J., Hammarstrom, L., and Beverley, P. C. (2001). Does 77C→G in PTPRC modify autoimmune disorders linked to the major histocompatibility locus? *Nat. Genet.*, 29(1):22–3.

Wasserstein, R. L. and Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician*.

Zeggini, E. and Ioannidis, J. P. (2009). Meta-analysis in genome-wide association studies. *Pharmacogenomics*, 10(2):191–201.

Zou, G. Y. (2007). Toward using confidence intervals to compare correlations. *Psychol. Methods*, 12(4):399–413.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2249v1 | CC BY 4.0 Open Access | rec: 5 Jul 2016, publ: 5 Jul 2016