Corrected: Publisher Correction

Selective effects of heterozygous protein-truncating variants

To the Editor — Cassa et al.[1] have recently presented an interesting analysis of the selection coefficients against heterozygous carriers of protein-truncating variants (PTVs) in several human populations, concluding that the mean selection coefficient against such a mutation when heterozygous is approximately 0.05, with a wide distribution around the mean (p. 809 and Fig. 1 in ref. [1]). With random mating in a large population, selection against the heterozygous carriers of strongly deleterious mutations is the predominant selective force, because homozygotes are very rare. The equilibrium frequency of mutant alleles at a locus, q^*, is then equal to u/s_{het}, where u is the mutation rate of a deleterious allele, and s_{het} is the decrease in fitness experienced by heterozygous carriers, measured relative to the fitness of normal individuals[2].

For this purpose, it is reasonable to assume that, for a given gene, u is the net rate of mutation to all possible PTVs that can be generated for the gene in question and that s_{het} is the same for all the PTV mutations in the gene. If the mutations are sufficiently severe in their fitness effects that they are destined for rapid elimination from the population, the mean frequency of mutant alleles over the probability distribution of q generated by random genetic drift is approximately equal to q^* (ref. [3]), thus apparently justifying the assumption of mutation-selection equilibrium. In their analysis, Cassa et al. assumed that the observed number of copies of a mutant allele for a given gene in a set of N alleles sampled from a population is drawn from a Poisson distribution with mean Nq^*. For this assumption to be valid, the fluctuations in q around q^* produced by drift must be negligible. Cassa et al. justified this assumption through a heuristic argument (first section in Methods in ref. [1]).

We believe that this assumption is questionable, as can be seen by considering the probability density of q, $\phi(q)$, at the stationary state among mutation, selection and drift in a randomly mating population with effective size N_e, first studied by Wright[4] (formally, the existence of the stationary state requires a small amount of back mutation from mutant to wild type, but this has a trivial effect and can be ignored). Nei[3] has shown that $\phi(q)$ for a strongly selected mutation with a heterozygous selection coefficient s_{het} is well approximated by a gamma distribution, with a mean of q^* and shape parameter $\theta = 4N_e u$. Poisson sampling from a gamma distribution generates a negative binomial distribution[5] for the number of copies i of a mutant allele in a sample of N alleles:

$$P(i) = \binom{i+\theta-1}{i}\left(\frac{z}{z+1}\right)^{\theta}\left(\frac{1}{z+1}\right)^{i}$$

where $z = 4N_e s_{het}/N$.

The mean and variance of the distribution are Nq^* and $\theta(1+z)/z^2$, respectively. The ratio of the coefficient of variation of this distribution to that for a Poisson distribution with the same mean is $\sqrt{(1+z^{-1})}$. It follows that, if $z << 1$, there is a much wider spread in the sampling distribution of the observed numbers of copies of mutant alleles across different genes than was assumed by Cassa et al. For example, with $N = 60{,}000$, $s_{het} = 0.05$, and $N_e = 10{,}000$ (a frequently used estimate for the species effective population size of humans[6]), the ratio is equal to 5.57.

This result implies that there may be a substantial upward bias in the spread of the distribution of s_{het} values estimated by the method of Cassa et al. We recognize that it is probably not appropriate to use the above value of $N_e = 10{,}000$–$20{,}000$ for humans, which is obtained from putatively neutral DNA sequence diversity and reflects the harmonic mean of the species effective population size over several hundred thousand years in the past[6]. Mutations destined for loss persist in a population for only a few tens of generations at most[7], and so the N_e relevant for PTVs is likely to reflect the much larger population sizes characteristic of the last few hundred years, thus decreasing the size of the bias. For example, with $N_e = 100{,}000$, the ratio of the coefficients of variation becomes 2.

In addition, the elimination of strongly deleterious alleles is also affected by population subdivision, and their fate is then strongly determined by the local effective population size, as shown by the classic studies of Dobzhansky and Wright[8] on the allelism of lethal mutations in populations of *Drosophila pseudoobscura*. Even for the simple case of an island model with an infinite number of demes, the expression for $\phi(q)$ becomes more complex than a gamma distribution, and the mean allele frequency can depart substantially from q^* (ref. [9]). A detailed analysis of the effects of drift on the frequency distribution of the numbers of deleterious mutations with the demographies characteristic of the populations used in their study would be needed to determine whether the conclusions reached by Cassa et al. concerning the width of distribution of the heterozygous selection coefficient are valid. In addition, their estimates of the selection coefficients for individual genes were based on the inferred distribution of s_{het}, thus also prompting questions about their accuracy.

In response to these comments, the authors[10] have conducted an analysis that includes a model of recent population-size change for Europeans. The results appear to substantiate their previous conclusions, notwithstanding the approximations made in their original study.

Reporting summary

Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper. ❐

Brian Charlesworth* and William G. Hill*
Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Edinburgh, UK.
e-mail: Brian.Charlesworth@ed.ac.uk; W.G.Hill@ed.ac.uk

Published online: 26 November 2018
https://doi.org/10.1038/s41588-018-0291-9

References

1. Cassa, C. A. et al. *Nat. Genet.* **49**, 806–810 (2017).
2. Haldane, J. B. S. *Proc. Camb. Philos. Soc.* **23**, 838–844 (1927).
3. Nei, M. *Proc. Natl. Acad. Sci. USA* **60**, 517–524 (1968).
4. Wright, S. *Genetics* **16**, 97–159 (1931).
5. Fisher, R. A. *Ann. Eugen.* **11**, 182–187 (1941).
6. Charlesworth, B. *Nat. Rev. Genet.* **10**, 195–205 (2009).
7. Kimura, M. & Ota, T. *Genetics* **63**, 701–709 (1969).
8. Wright, S., Dobzhansky, T. & Hovanitz, W. *Genetics* **27**, 363–394 (1942).
9. Glémin, S. *Genet. Res.* **86**, 41–51 (2005).
10. Weghorn, D. et al. Preprint at https://www.biorxiv.org/content/early/2018/10/03/433961 (2018).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-018-0291-9.

nature research

Corresponding author(s): Brian Charlesworth

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed
☒	☐ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐ An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐ A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☒	☐ Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	None was used
Data analysis	No data were involved

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

No data were generated.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Not applicable
Data exclusions	Not applicable
Replication	Not applicable
Randomization	Not applicable
Blinding	Not applicable

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging