Re-evaluating evidence for adaptive mutation rate variation

https://doi.org/10.1038/s41586-023-06314-y

Received: 16 June 2022

Accepted: 12 June 2023

Published online: 26 July 2023

Open access

Check for updates

Long Wang[1], Alexander T. Ho[2], Laurence D. Hurst[2]✉ & Sihai Yang[1]✉

ARISING FROM J. G. Monroe et al. *Nature* https://doi.org/10.1038/s41586-021-04269-6 (2022)

Although mutation rates vary within genomes, suggestions[1,2] that more selectively important DNA has a lower mutation rate are contentious not least because unbiased estimation of the mutation rate is challenging[3]. Monroe et al.[4] (hereafter Monroe) also report that in *Arabidopsis* more important sequences have lower mutation rates and, while overlooking similar claims[1], suggest that this challenges "a long-standing paradigm regarding the randomness of mutation"[4]. We find, however, that their mutation calling has abundant sequencing and analysis artefacts explaining why their data are not congruent with well-evidenced mutational profiles. As the key trends associated with sequence importance are consistent with well-described mutation-calling artefacts and are not resilient to reanalysis using the higher-quality components of their data, we conclude that their claims are not robustly substantiated.

In principle, identifying new mutations is simple: one sequences genomes of close relatives and identifies new differences between them. There are, however, multiple pitfalls. For example, incorrect mapping of short reads to the genome can result in erroneous, and commonly clustered, mutation calls. Further, as the rate of sequencing errors is orders of magnitude higher than the rate of mutation, these errors must be excluded. Robust rules for mutation calling, such as requiring multiple independent sequence reads from both strands supporting the same mutation, can obviate many issues.

We expect higher than normal error rates in Monroe as, to identify somatic mutations, they relaxed their previous[5] stringency in mutation calling (Supplementary Methods). To assay the impact of this, we compared their mutation calls to those generated by a conventional pipe. We find only 3.7% (n = 160) concordance with Monroe's 4,322 filtered putative mutations, 61% of which are uncallable. We term the 96.3% of Monroe's mutations that fail conventional filters as low quality (LQ). Their previous data[5] using a more stringent pipe (we dub this Weng data) agree with our analysis: 94.2% agreement vis-à-vis mutations 'confidently' called, only 1.4% uncallable. Prima facie, most of Monroe's mutation calls thus may well be unsafe. This is supported by analysis of the profile of LQ mutations as this is different to the higher-quality calls: they are of a qualitatively different type (intergenic, intronic and so on) compared with high-quality datasets (that mutually agree; Extended Data Fig. 1a,b) and have a different mononucleotide mutational profile (Extended Data Fig. 1c).

Deviations of this magnitude are unlikely to be accounted for by a somatic versus germline difference. Instead, Monroe's data are different largely because they are enriched for sequencing and analysis artefacts. We consider two artefact fingerprints. First, in Illumina sequencing[6,7], a base can be erroneously replaced by a bleeding one within[8,9] (for example, AAAGAAA appears as AAAAAAA) or in the vicinity of[6,7] (for example, AAAAC appears as AAAAA) homopolymeric runs. Second, failure to eliminate poor-quality reads and mapping artefacts will over-report clustering of putative mutations.

More than half (54%) in the LQ data are bleeding-type putative mutations within 5 base pairs of A/T homopolymers (Extended Data Fig. 1d), compared with 24% in the high-quality (HQ) Weng set. Within genic regions (exonic + intronic), the proportion is 67.5% in the LQ data, 91% of which are intronic. Depending on the protocol, the error is biased to bleed-over of either AT or GC residues, but not both[6]. It is then notable that the bias is particular to A/T homomeric runs with 1.5% in HQ Weng and 1.2% in LQ near GC runs (Extended Data Fig. 1d). This AT bias artefact is reported for related Illumina machines[6].

Of the 2,247 bleed mutations near homopolymeric runs, 1,149 (51.1%) are immediately next to or within the runs. Of the remaining 1,098, at least 648 cluster with other bleed errors (for example, AAAAACACACA is read as AAAAAAAAAAA giving three putative mutations). These bear the hallmarks of artefacts: typically only one strand is affected, all of the putative mutations are seen in the same read and their rate decays as a function of distance from the true end of the run. As also expected from the profile of sequencing errors[6], the probability of a mutation being called increases with the length of the homopolymer: in introns, regression of $\log_{10}$(putative mutations per base pair of homopolymeric sequence) predicted by run length, slope = 0.27, Pearson's r^2 = 0.87, P = 0.0006, degrees of freedom (df) = 6.

For Monroe's somatic mutations (recalled from Weng's vcf data) unassociated with homomeric runs (46% of their mutations), most are clustered (2 mutations within 10 base pairs of each, 27.5% of all mutations) or unexpectedly common (>10 samples, 8.7% of all mutations), indicative of mis-mapping issues. Only 2.5% in the Weng HQ data are clustered. Many of Monroe's putative mutations are associated with more than one error: about 34% are associated with A/T homomeric runs and in a tight cluster. As centromeres are prone to mapping errors[10], mis-mapping probably explains why 40.9% of LQ mutations are centromeric (see, for example, Extended Data Fig. 1e) compared with 27.9% in Weng HQ.

We do not suppose these to be all of the errors. Whereas Monroe call 773,141 mutations using our sequence[11], using the same HaplotypeCaller-GVCF calling method[12], with default parameters and without any further filtering, we identify only 31,486 raw indels and 72,516 raw single nucleotide polymorphisms (all but 17 of which are unsafe). This gross

[1]State Key Laboratory of Pharmaceutical Biotechnology, School of Life Sciences, Nanjing University, Nanjing, China. [2]The Milner Centre for Evolution, Department of Biology and Biochemistry, University of Bath, Bath, UK. ✉e-mail: l.d.hurst@bath.ac.uk; sihaiyang@nju.edu.cn

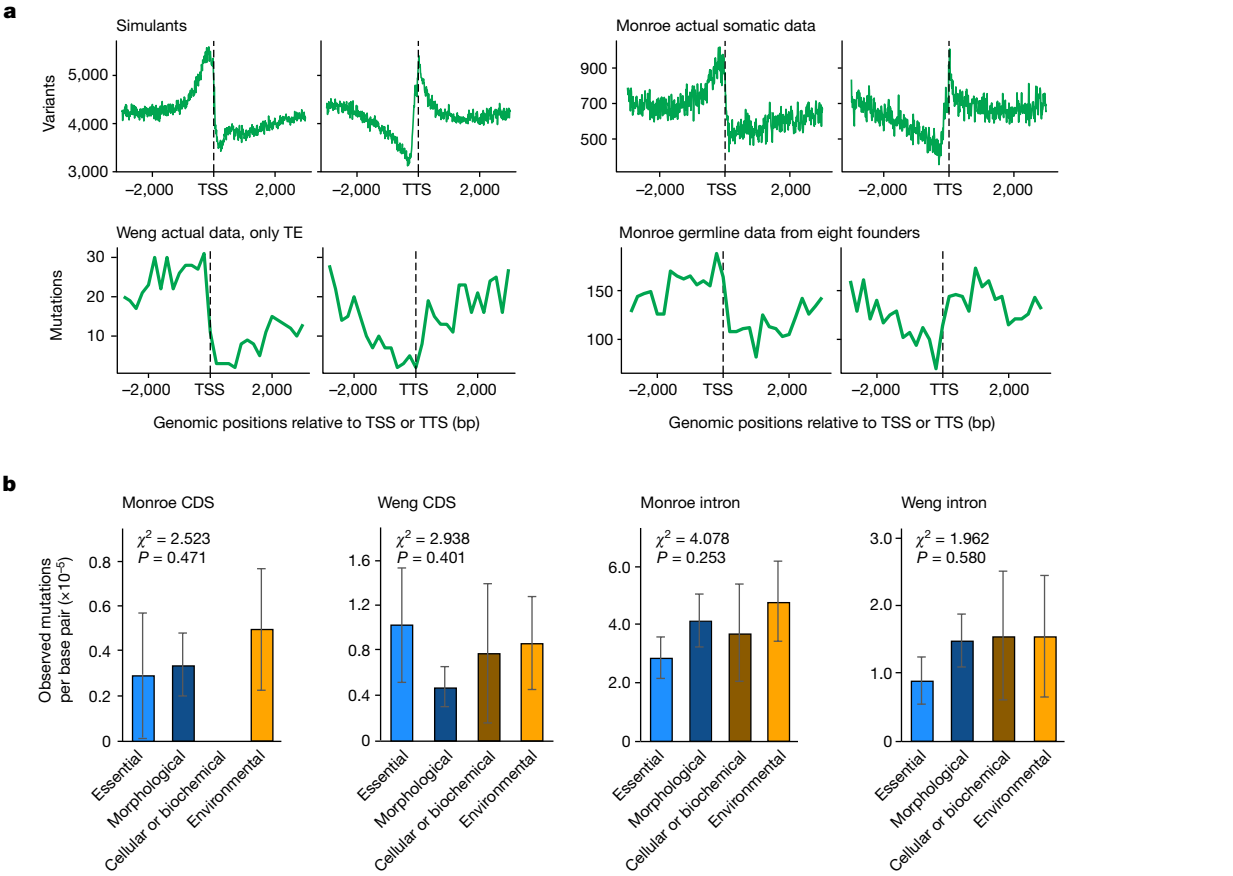


Fig. 1 | Core claims of Monroe et al. are error artefacts. a, Top row: the profile of errors (misascribed as mutations) expected around TSS and TTS attributable to the errors associated with A/T homopolymeric runs. We simulated 2 million errors associated with A/T homopolymeric runs and used the Monroe script to generate the left plot. The data are an exact match to their somatic data calls (reproduced here as the top right plot, figure reused from Monroe et al.[4] under Creative Commons Attribution 4.0 International License). Bottom row: germline mutations in TEs (left panel). Observed mutations in germline unmasked. bp, base pair. **b**, Essential genes do not have a low mutation rate. The second claim of Monroe to substantiate that mutation is lower in more functionally important sequences is that essential genes have the lowest mutation rates (their Fig. 3c). To substantiate this, they seem to have used many thousands of unfiltered calls. We repeat the analysis using filtered data, either Weng or Monroe, including indels. In neither case is there significant heterogeneity (we provide χ^2 values for all comparisons, df = 3, but for Monroe CDS numbers are so small that these calculations are not valid and presented for completeness alone). Tests are one-sided in the sense that we call significance only if there is heterogeneity not if they are more similar than expected. Tests are two-sided in the sense that we ask about deviation from null in any direction. Unification of the two datasets does not alter conclusions: CDS, $\chi^2 = 3.3$; intron, $\chi^2 = 5.76$, $P > 0.05$ for all without multi-test correction. Error bars are s.e.m. across gene samples for which sample sizes are: essential ($n = 719$), morphological ($n = 861$), cellular or biochemical ($n = 297$) and environmental ($n = 522$) for CDS analysis (Monroe CDS and Weng CDS), and essential ($n = 671$), morphological ($n = 789$), cellular or biochemical ($n = 270$) and environmental ($n = 452$) for the intron analysis (Monroe intron and Weng intron). For representation of the underlying data, see Extended Data Fig. 2.

excess of mutations in Monroe is due to an analysis error on their part (see correction from Monroe et al.[13]).

Analysis and sequencing errors explain many of the core claims of Monroe. They report that the mutation rate alters markedly around transcription start sites (TSSs) and stop sites (TTSs), arguing that this provides evidence that gene bodies are mutationally protected. We simulated random errors associated with A/T homopolymeric runs and derived a distribution that is a near-perfect match to their somatic data (Fig. 1a).

Monroe also claimed a low mutation rate in essential genes as evidence for mutational protection for more important sequences (their Fig. 3c). However, to do this, they included orders of magnitude more putative mutations than in their filtered datasets. We repeated their analysis using the Weng data and the Monroe data (their filtering). In neither, nor in the merged dataset, is there heterogeneity in the mutation rate between gene classes in coding sequence (CDS) or intron (Fig. 1b and Extended Data Fig. 2). Indeed, in the best data (Weng), essential genes have the highest mutation rate per base pair of CDS (Fig. 1b and Extended Data Fig. 2).

Artefacts explain other heterogeneities in the mutation rate. Weng's data report a plausible intron to CDS per base pair ratio of about 0.91:1 (paired *t*-test on normalized dinucleotide mutation rates, $P = 0.58$, df = 95), whereas Monroe's data report an unprecedented 5.2 to 1 ratio (paired t, $P = 3.5 \times 10^{-8}$, df = 95). This comparison is especially informative as it controls for transcription-associated mutational effects. Much of this higher intronic rate in Monroe's data is attributable to homopolymeric run artefacts as CDS has fewer, and less error prone, runs (Supplementary Results).

The artefacts are also evident in the profile of mutations called (Extended Data Fig. 1f). Counts of the 96 dinucleotide mutations from the Monroe and Weng data are discordant ($\chi^2 = 1{,}516$, $P < 2 \times 10^{-16}$, df = 95). The most common dinucleotide mutations in the Monroe data end AA or TT as the resolved mutational event, with G/C mutated to the neighbouring A/T being especially discrepant (Extended Data Fig. 1f). The mutational events terminating AA/TT contribute 34.4% of the relative normalized mutations in the Monroe set but only 21.6% in Weng's.

The Monroe data also incorrectly predict the mutational equilibrium frequency of AA/TT dinucleotides compared to observed frequencies.

Using the 16 × 16 normalized mutational matrix for the Monroe and the Weng data individually, we predict mutational equilibrium dinucleotide content[14] and compare with intergenic dinucleotide content. The Weng data are not influenced by AA/TT calls ($P = 0.38$), whereas in the Monroe data AA/TT are over-called outliers ($P = 0.003$; Extended Data Fig. 1g).

This neighbour base matching affecting both A and T in Monroe's data is an expected bleed artefact with no biological basis. By contrast, we expect CpG>TpG mutations to be common given well-described methylated CpG hyperinstability[15]. In Weng's data, but much less so Monroe's, this is the case (Extended Data Fig. 1f).

Although Monroe's claim that the mutation rates are lower at more functionally important sites seems to be highly influenced by artefacts, nonetheless, the mutation rate is not uniform. In some part this is because transposable elements (TEs) have high mutation rates and TEs are rare in gene bodies. In *Arabidopsis*, cytosine methylation-mediated TE suppression[16] should lead to C instability. In the Weng data, 65% of mutations in TE are CG, CHH or CHG versus 51% in intergenic non-TE (for example, 5.2:1 ratio of CpG>TpG per CG, TE to intergenic non-TE). In (robust) germline data there is a higher mutation rate in TEs than elsewhere, including the best comparator, non-TE intergenic sequence: TE versus non-TE intergenic sequence, mean ratio per dinucleotide = 3.93 (paired *t*-test on normalized dinucleotide rates, $P < 3 \times 10^{-7}$, df = 95). This TE enrichment largely explains why in germline data mutation rates are higher 5′ of TSS and 3′ of TTS, TEs being enriched outside transcribed domains (Fig. 1a).

Although TE mutational enrichment is seen in Monroe HQ data (Extended Data Fig. 1a,b), it is not seen in the Monroe data in toto (paired *t*-test on normalized dinucleotide rates, $P = 0.9$). Given this and the failure to capture well-described methyl C instability[15], although we do not doubt that epigenetic marks such as methylation can affect mutation, the correlations evidenced by Monroe between various marks and mutation rate variation should be treated with the same caution as their claim that mutation is rarer in more important sequences.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-023-06314-y.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The data to replicate the analyses and figures in the paper and the Extended Data are available at https://github.com/wl13/reanalysis1.

Code availability

The code to replicate the analyses and figures in the paper and the Extended Data is available at https://github.com/wl13/reanalysis1.

1. Chuang, J. H. & Li, H. Functional bias and spatial organization of genes in mutational hot and cold regions in the human genome. *PLoS Biol.* **2**, e29 (2004).
2. Martincorena, I., Seshasayee, A. S. N. & Luscombe, N. M. Evidence of non-random mutation rates suggests an evolutionary risk management strategy. *Nature* **485**, 95–98 (2012).
3. Chen, X. Z. & Zhang, J. Z. No gene-specific optimization of mutation rate in *Escherichia coli*. *Mol. Biol. Evol.* **30**, 1559–1562 (2013).
4. Monroe, J. G. et al. Mutation bias reflects natural selection in *Arabidopsis thaliana*. *Nature* **602**, 101–105 (2022).
5. Weng, M. L. et al. Fine-grained analysis of spontaneous mutation spectrum and frequency in *Arabidopsis thaliana*. *Genetics* **211**, 703–714 (2019).
6. Stoler, N. & Nekrutenko, A. Sequencing error profiles of Illumina sequencing instruments. *NAR Genom. Bioinform.* **3**, lqab019 (2021).
7. Nakamura, K. et al. Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res.* **39**, e90 (2011).
8. De Summa, S. et al. GATK hard filtering: tunable parameters to improve variant calling for next generation sequencing targeted gene panel data. *BMC Bioinform.* **18**, 119 (2017).
9. Filtering of variants in homopolymeric regions. *QIAGEN CLC Main Workbench Manual* https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/650/Filtering_variants_in_homopolymeric_regions.html (2022).
10. Naish, M. et al. The genetic and epigenetic landscape of the *Arabidopsis* centromeres. *Science* **374**, eabi7489 (2021).
11. Wang, L. et al. The architecture of intra-organism mutation rate variation in plants. *PLoS Biol.* **17**, e3000191 (2019).
12. Calling variants on cohorts of samples using the HaplotypeCaller in GVCF mode. *GATK Team* https://gatk.broadinstitute.org/hc/en-us/articles/360035890411-Calling-variants-on-cohorts-of-samples-using-the-HaplotypeCaller-in-GVCF-mode (2022).
13. Monroe, J. G. et al. Author Correction: Mutation bias reflectsnatural selection in *Arabidopsis thaliana*. https://doi.org/10.1038/s41586-023-06387-9 (2023).
14. Rice, A. M. et al. Evidence for strong mutation bias toward, and selection against, U content in SARS-CoV-2: implications for vaccine design. *Mol. Biol. Evol.* **38**, 67–83 (2021).
15. Ehrlich, M. & Wang, R. Y. 5-Methylcytosine in eukaryotic DNA. *Science* **212**, 1350–1357 (1981).
16. Wang, Z. & Baulcombe, D. C. Transposon age and non-CG methylation. *Nat. Commun.* **11**, 1221 (2020).

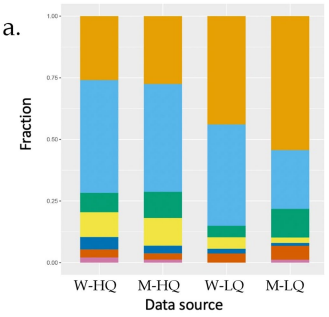
Acknowledgements We thank G. Monroe and D. Weigel for constructive response to issues raised. This work was supported by grants from the National Natural Science Foundation of China (grant numbers 31970236, 32270664 and 32170327).

Author contributions L.W. secured funding, wrote the manuscript and carried out analysis; L.D.H. wrote the manuscript and carried out analysis; A.T.H. carried out analysis; S.Y. secured funding and wrote the manuscript.

Competing interests The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-023-06314-y.
Correspondence and requests for materials should be addressed to Laurence D. Hurst or Sihai Yang.
Reprints and permissions information is available at http://www.nature.com/reprints.
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

a.



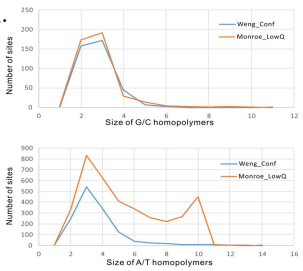
b.

	Weng_HQ	Monroe_HQ	Weng_LQ	Monroe_LQ	Weng_all	Monroe_all
Weng_HQ	-	0.04504	0.19787	**0.37679**	0.01160	**0.36376**
Monroe_HQ	0.85	-	0.18883	**0.35196**	0.04289	**0.33885**
Weng_LQ	0.0059	0.07	-	**0.21735**	0.18627	0.20584
Monroe_LQ	**2.953E-152**	**7.941E-25**	**0.0007***	-	**0.36641**	0.01311
Weng_all	0.99	0.84	0.014	**4.127E-149**	-	**0.35339**
Monroe_all	**1.094E-141**	**2.407E-19**	0.0049	0.86	**1.421E-138**	-

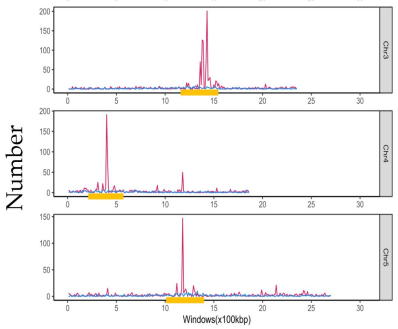
c.

	Weng_HQ	Monroe_HQ	Weng_LQ	Monroe_LQ	Weng_all	Monroe_all
Weng_HQ	-	0.08224	**0.18090**	**0.27676**	0.01046	**0.26563**
Monroe_HQ	0.38	-	0.18614	**0.31138**	0.08120	**0.29985**
Weng_LQ	**4.494E-06**	0.048	-	0.16335	**0.17043**	0.15337
Monroe_LQ	**1.712E-140**	**3.223E-17**	0.011	-	0.26819	0.01153
Weng_all	0.99	0.46	**6.092E-05**	**5.449E-137**	-	**0.25701**
Monroe_all	**8.741E-132**	**1.557E-15**	0.026	0.99	**4.093E-128**	-

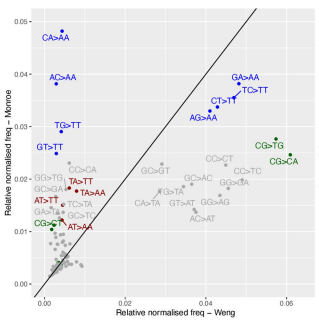
d.



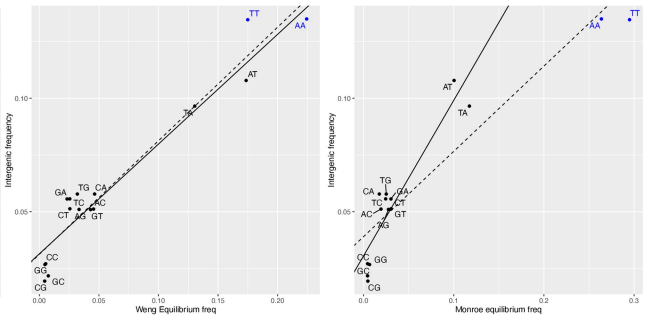
e.



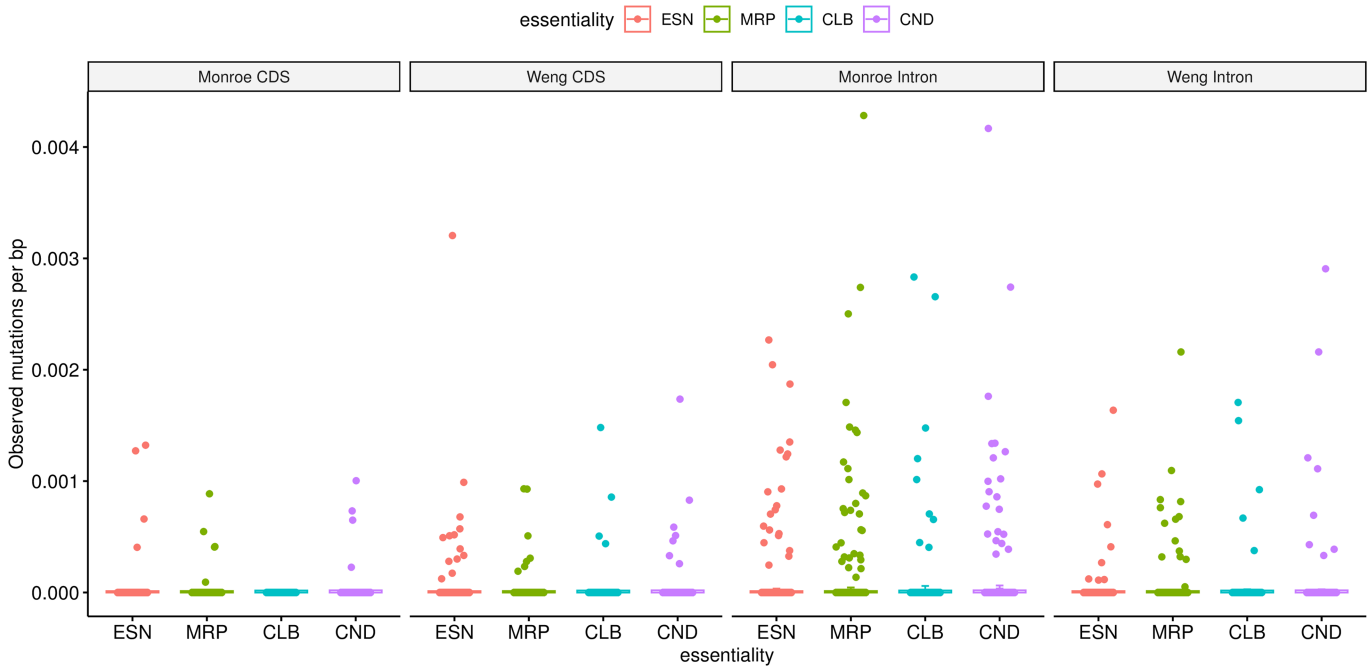
f.



g.



Extended Data Fig. 1 | Mutational properties in different mutation data sets. We consider the data from Weng et al.[5] and Monroe et al.[4]. We reanalysed both datasets from raw files and split the data into confident mutation calls (HQ) and low-quality calls (LQ). The samples sizes are 1743, 160, 107, 4162 in Weng high confident (W-HQ), Monroe high confident (M-HQ), Weng low quality (W-LQ), Monroe low quality (M-LQ). **a** is a visual representation of the frequency of each class of mutation, **b** are the Euclidean distances between the frequency vectors (upper section), including the two full data sets. The values below the diagonal are chi^2 P values with v = 7, based on raw counts omitting any rows where both were zero (indicated *, v = 6). Those significant after Bonferonni correction are highlighted in bold. Tests are one sided in the sense that we call significance only if there is heterogeneity not if they are more similar than expected. Tests are two sided in the sense that we ask about deviation from null in any direction. **c**. Relative rates of different mononucleotide mutations. Above the diagonal, Euclidean distances between the 12 element vectors of relative mutation frequencies. Below the diagnonal, P from chi^2 on raw counts (v = 11), with those significant after Bonferonni correction highlighted in bold. Tests are one sided in the sense that we call significance only if there is heterogeneity not if they are more similar than expected. Tests are two sided in the sense that we ask about deviation from null in any direction. **d**. rates of potential bleed errors in proximity to homopolymeric runs of different length (top panel A or T runs, bottom panel, G or C runs) for Weng HQ (i.e. Confident) and Monroe LQ calls. Y axis is number of homomeric runs with associated bleed type mutations. **e**. Distribution of mutations on chromosomes 3, 4 and 5. Centromere is shown as orange block. Weng HQ data is in blue, Monroe LQ data in red. **f**. Relative normalised dinucleotide mutation frequencies in the Weng et al. and Monroe et al. data. In each data set we determined the absolute number of each dinucleotide-associated mutation. We then determined the normalised rate by dividing observed rates by numbers of each dinucleotide in the genome, this providing a rate per bp. The sum rate for each set was calculated and the normalised rates divided by this sum to provide a relative normalised rate. The line of slope 1 indicates equivalence between the two data sets. In blue and red are all the dinucleotide based events that terminate either AA or TT after mutation. In blue are those mutating C/G residues, in red, A/T residues. CG starting dinucleotides are in green. For clarity most other data points are represented by dots alone. **g**. Predicted and observed dinucleotide frequencies. Observed dinucleotide frequencies are from intergenic sequence. Mutational equilibrium analytically derived as in ref. 14. Left panel, Weng et al. full data, right panel, Monroe et al. full data. To test for AA/TT concordance, we consider slopes from regression of observed and predicted, including (dashed) and omitting (solid) AA and TT. If AA and TT are unduly influential, we expect a significant difference in slopes. Difference in slopes was tested by *t* test with df = 26 (Monroe data, $t = 3.39$, $P = 0.0028$, Weng data, $t = -0.26$, $P = 0.38$). The test is two-sided.



Extended Data Fig. 2 | Representation of Fig. 1b showing underlying data points. Note that for most genes there are no mutations in the reduced data sets hence most data sits at y = 0.

nature portfolio

Corresponding author(s): Hurst, L.D., Yang, S

Last updated by author(s): Mar 21, 2023

Reporting Summary

Nature Portfolio 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 Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Please do not complete any field with "not applicable" or n/a. Refer to the help text for what text to use if an item is not relevant to your study.
For final submission: please carefully check your responses for accuracy; you will not be able to make changes later.

Statistics

For all statistical analyses, confirm that the following items are present in the 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
☐	☒	A statement on 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 statistical parameters 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

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	https://github.com/wl13/reanalysis1
Data analysis	https://github.com/wl13/reanalysis1

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 and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio 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 description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The code to replicate the analyses and figures in the paper and the Extended Data is available at https://github.com/wl13/reanalysis1

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

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

Reply to: Re-evaluating evidence for adaptive mutation rate variation

https://doi.org/10.1038/s41586-023-06315-x

Published online: 26 July 2023

Open access

Check for updates

J. Grey Monroe[1]✉, Kevin D. Murray[2], Wenfei Xian[2], Thanvi Srikant[2], Pablo Carbonell-Bejerano[2], Claude Becker[2], Mariele Lensink[1], Moises Exposito-Alonso[3,4], Marie Klein[1], Julia Hildebrandt[2], Manuela Neumann[2], Daniel Kliebenstein[1], Mao-Lun Weng[5], Eric Imbert[6], Jon Ågren[7], Matthew T. Rutter[8], Charles B. Fenster[9] & Detlef Weigel[2]✉

REPLYING TO L. Wang et al. *Nature* https://doi.org/10.1038/s41586-023-06314-y (2023)

Wang and colleagues[1] argue that our report[2] of lower mutation rates in gene bodies, essential genes and regions marked by H3K4me1 must result from DNA sequencing errors. We appreciate the issues raised by them and by other colleagues[3]. Although we overlooked some sources of errors, these are insufficient to invalidate our conclusions, which are confirmed by more stringent reanalyses of our original data, new analyses restricted to high-confidence germline mutations[4], and direct demonstration of plant DNA repair proteins being recruited to gene bodies, essential genes and H3K4me1, where they reduce local mutation rates[5,6].

Wang and colleagues[1] identify issues with somatic mutation calling, suggesting that homopolymer bleed-through errors in Illumina sequencing are responsible for patterns observed in somatic mutations, and that elevated cytosine deamination in transposable elements is responsible for the patterns in germline mutations. Here we address these concerns.

Consecutive runs of identical nucleotides, or homopolymers, pose challenges to discovering rare mutations because they can lead to Illumina sequencing errors at immediately neighbouring nucleotides through homopolymer bleed-through[7]. At the same time, homopolymer regions have higher true mutation rates even at local but non-adjacent sites (for example, ref. 8). Wang and colleagues[1] found that the distribution of simulated homopolymer errors mirrors the overall distribution of mutations we reported around genes (their Fig. 1a). However, there are several reasons why such homopolymer errors cannot be the source of inferred mutation bias.

Wang and colleagues[1] assume that homopolymer bleed-through errors affect sequences up to five nucleotides away from homopolymers, although these errors occur on modern Illumina platforms at positions immediately adjacent to a run of identical bases[7]. Moreover, their simulation of sequencing errors apparently assumes that 100% of sequencing errors occur as a product of homopolymer bleed-through. By contrast, empirical estimates of sequencing errors report only 0.7 to 5.2% to be the result of homopolymer bleed-through[7]. Across all data in our study, only 12.0% of total single-nucleotide variant calls (10.2% for high-quality germline calls) could be potential homopolymer-adjacent bleed-through errors, and thus on their own cannot explain the approximately 50% mutation rate reduction we observed in gene bodies relative to intergenic regions[2].

More importantly, Wang and colleagues' own analysis[1] reports that the proportion of potential homopolymer bleed-through errors in our data is actually higher in gene bodies (exons plus introns), which should lead to gene body mutation rates being overestimated, not underestimated. We confirm across our datasets that the proportion of potential homopolymer bleed-through errors is not lower in gene bodies (Fig. 1a, left), and differs from the pattern of mutation calls (Fig. 1a, right). Similarly, the proportion of potential homopolymer bleed-through errors is not reduced in essential genes (Fig. 1b). The distribution of potential homopolymer bleed-through errors, therefore, disagrees with the hypothesis of Wang and colleagues[1]. By contrast, the observed pattern is expected if gene bodies and essential genes experienced a reduction in true mutation rates, as noise introduced by sequencing errors should have a proportionally larger effect on regions with truly low mutation rates.

Homopolymeric sequences (but not potential homopolymer bleed-errors) are enriched outside gene bodies, as reported by Wang and colleagues[1]. Thus, the observed mutation rate heterogeneity is consistent with previous evidence that homopolymer-rich regions have higher true mutation rates[8] and that their enrichment in these regions is consistent with the expected long-term evolutionary consequence of lower DNA repair activity, as the expansion of homopolymers is a signature of lower mismatch repair activity (Supplementary Note 3). Moreover, both preferential repair of exons by mismatch repair and higher intronic mutation rates in somatic tissues have been widely documented (Supplementary Note 3). Likewise, considerable differences in mutation rate and spectra between somatic and germline cells are well known, with somatic cells having orders of magnitude higher mutation rates. Indeed, differences between mitotic and meiotic cells have been previously proposed for *Arabidopsis thaliana* by Wang and colleagues[9] (Supplementary Note 3).

Wang and colleagues[1] further suggest that the patterns we observed in germline mutations might result largely from elevated deamination of methylated cytosines (GC-to-AT mutations) in transposable elements. Several findings are inconsistent with this hypothesis: cytosine methylation was included as a covariate in our original models, mutation accumulation experiments consistently indicate that mutation rates are lower in gene bodies relative to non-transposable element intergenic regions in *A. thaliana* (Fig. 2a,b; see below), and removing all GC-to-AT mutations from our original germline dataset does not alter the observed pattern, with H3K4me1 remaining the strongest epigenomic predictor of lower mutation (described in detail recently[4]). The same has been demonstrated for mutation rate variation in rice, in which mutation rates are lower in gene bodies relative to both intergenic regions and transposable elements[6].

To further address concerns with somatic mutation calls in general, we re-called putative somatic mutations in the original 107 lines[10] by mapping reads to an improved reference genome[11] and applying more

[1]University of California Davis, Davis, CA, USA. [2]Max Planck Institute for Biology Tübingen, Tübingen, Germany. [3]Department of Plant Biology, Carnegie Institution for Science, Stanford, CA, USA. [4]Department of Biology, Stanford University, Stanford, CA, USA. [5]Department of Biology, Westfield State University, Westfield, MA, USA. [6]ISEM, University of Montpellier, Montpellier, France. [7]Department of Ecology and Genetics, EBC, Uppsala University, Uppsala, Sweden. [8]Department of Biology, College of Charleston, Charleston, SC, USA. [9]Oak Lake Field Station, South Dakota State University, Brookings, SD, USA. ✉e-mail: gmonroe@ucdavis.edu; weigel@weigelworld.org

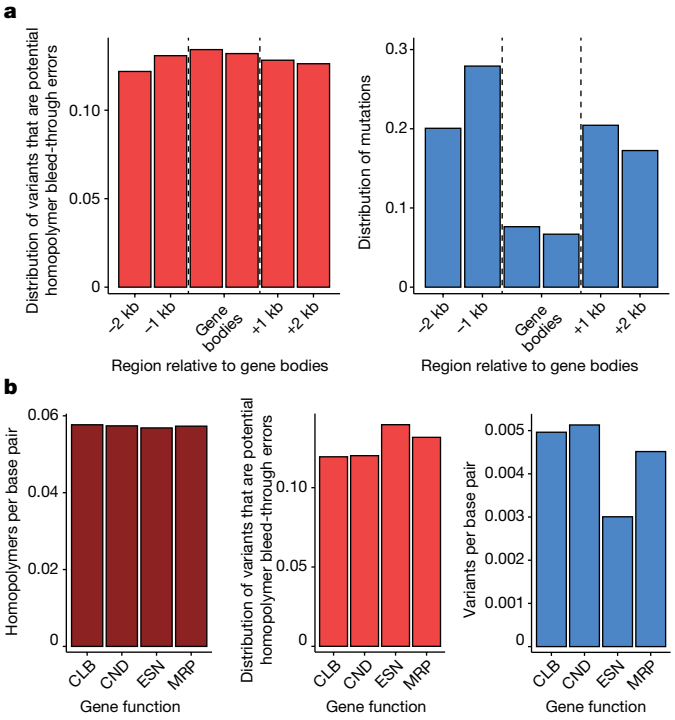


Fig. 1 | Potential homopolymer bleed-through sequencing errors cannot explain differences in mutation rate. a, The proportion of variants that are potential homopolymer bleed-through errors among all mutation calls in our original study[2] is as least as high in gene bodies as in intergenic sequences, and contrasts with the distribution of total mutation calls. kb, kilobase. **b**, Homopolymers and the proportion of variants that are potential homopolymer bleed-through errors in the original study[2] are not lower in essential genes (ESN) compared to genes with environmentally conditional (CND), morphological (MRP) and cellular or biochemical (CLB) functions, and cannot explain the distribution of actual mutation calls.

stringent filtering (Supplementary Note 1). This led to more complete and higher-quality read mapping (Supplementary Fig. 1) and resolved several issues described by Wang and colleagues[1] (for example, high intron-versus-exon mutation ratio and the proportion of potential homopolymer bleed-through errors; Supplementary Fig. 2). These data confirm that gene bodies experience lower mutation rates, including when manually removing potential homopolymer bleed-through errors (Supplementary Note 1). Many of the analyses by Wang and colleagues are affected by unreliable centromeric mutations, which constituted 41% of questioned somatic mutations[1]. These sites, however, could not have affected our conclusions because they were excluded from all of our original analyses (Supplementary Note 2 and Supplementary Fig. 3).

Wang and colleagues[1] examined essential genes with approaches that were not in our original study. They used subsets of our initial datasets, focusing on either about 2,000 germline or about 4,000 somatic single-nucleotide variants, finding that neither dataset directly revealed a statistically significantly lower mutation rate in essential genes. This approach seems underpowered, yielding near-zero values for mutation counts in entire gene classes, an indication that the data are poorly suited for χ^2 approximation (Supplementary Note 5).

In our study[2], we had instead modelled genome-wide mutation rates, and using these models, identified a connection between gene essentiality and mutation rate corresponding to epigenome differences—essential genes are enriched for H3K4me1, for example, which we found to be associated with lower mutation rates. We subsequently tested whether this expectation is met in a very large set of several hundred thousand loosely filtered putative somatic mutations with ample power to compare gene classes. We agree that somatic mutation calling is very difficult, as most real somatic mutations and unrepaired damaged sites (with DNA damage occurring 10,000 to 100,000 times per day per cell; Supplementary Note 3) are expected to be present in only one cell and thus detectable only by a single read. In Supplementary Note 4 and an accompanying Correction[12], we discuss why singletons were called as putative mutations in one of our reanalyses, from 64 leaves[13], owing to inadvertently mapping forward reads twice. However, analyses of variant quality in these data do not support the hypothesis that our results are simply due to higher rates of poor-quality calls in non-genic regions or non-essential genes (Supplementary Note 4 and Supplementary Fig. 4).

Finally, to directly address the possibility that our conclusions reflect unknown sources of bias in inherently uncertain somatic calls, we reanalysed germline mutations from our study[2] along with mutation accumulation experiment data generated in several independent studies (Supplementary Table 1). This meta-analysis of >10,000 germline mutations confirmed the previously reported, nearly universal reduction in single-nucleotide mutation rates in gene bodies, essential genes and regions marked by H3K4me1 (Fig. 2a–c; ref. 4). The notable exception comes from plants lacking the mismatch repair protein MSH2 (Fig. 2a; ref. 5). A similar pattern is seen when somatic mutations were called with very stringent criteria in plants deficient for the MSH2 partner MSH6, using a tool specifically designed for rare somatic mutations[14] (Fig. 2d). This was as predicted from H3K4me1 directly attracting MSH6 to gene bodies[6], confirming that DNA repair in *A. thaliana* is targeted to gene bodies, as is well known in humans (Supplementary Note 3). Finally, analyses of >43,000 experimentally induced de novo germline mutations in rice (previously validated with 99% accuracy) also show that gene bodies, conserved genes, and H3K4me1-marked regions experience lower mutation rates, even when considering only silent (synonymous) mutations[6].

Relationships between histone modifications, DNA repair, and mutation rate are widely known (Supplementary Note 3). Our work[2] considered the evolutionary implication of these relationships. We had leveraged models of the drift-barrier hypothesis to discover that natural selection could favour mechanisms linking DNA repair to widely distributed epigenomic features, such as H3K4me1, which is not only enriched in gene bodies and essential genes in *A. thaliana*, but also the histone modification most strongly associated with lower mutation rates in our data[2]. An important higher-order test of our conclusions is therefore whether they are mechanistically supported. Since publication of our article[2], it has been demonstrated that plant DNA repair proteins are recruited by H3K4me1 to gene bodies and essential genes. These repair proteins, which contain Tudor 'reader' domains that bind H3K4me1, include PDS5C, involved in homology-directed repair, and MSH6, which functions as a dimer with MSH2 in the mismatch repair pathway and recruits MutY of the base-excision repair pathway[15]. The genome-wide distribution of PDS5C, as measured by chromatin immunoprecipitation followed by sequencing[4,6,16], confirms that regions subject to elevated repair protein activity coincide with features at which we detected lower spontaneous mutation rates[4,6,16].

We conclude that the reported relationships between epigenomic features and mutation rates[2] are well supported mechanistically (Fig. 2e). We agree that there are issues and inherent uncertainties with somatic mutation calling, which make it difficult to know the accuracy of individual calls in the very large set of loosely filtered somatic variants[2]. However, the proposal that the observed patterns result only from sequencing errors is inconsistent with multiple lines of evidence from the original study, independent analyses and emerging parallel work.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information,

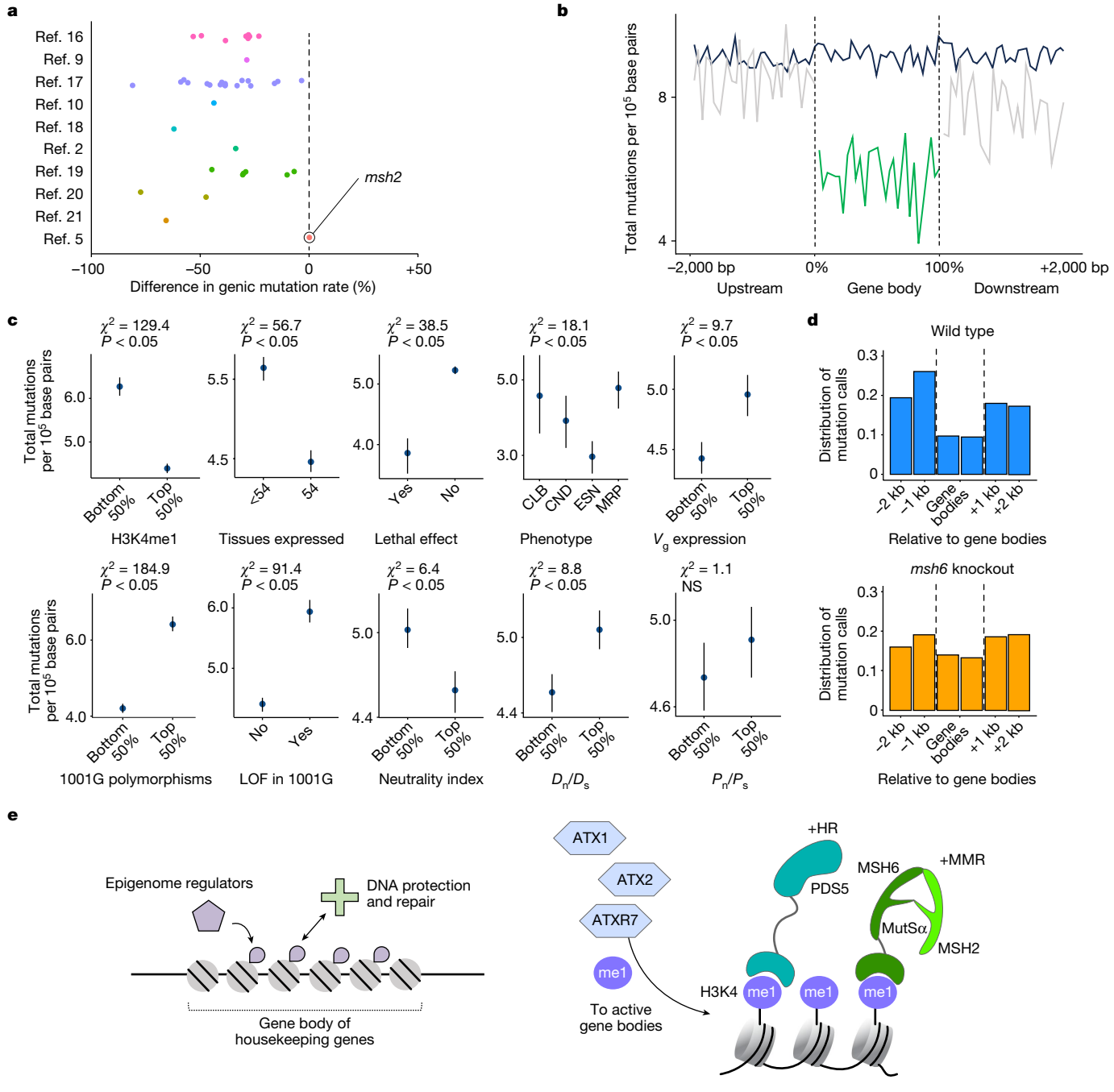


Fig. 2 | Joint analyses of germline mutations in several published *A. thaliana* mutation accumulation studies align with mechanistic models of mutation bias. a, Reduction in genic single-nucleotide germline mutation rates compared against genomic background in multiple *A. thaliana* datasets (Supplementary Table 1). For our original study[2], only new mutations from 400 mutation accumulation lines are shown; the other mutations in that paper were already described[10] and are shown separately here. Mutation rate reduction in genic regions is eliminated in *msh2* DNA repair mutants[5]. bp, base pair. **b**, Mutation rates around gene bodies (grey and green lines). Black line indicates randomly selected windows based on gene lengths. **c**, Mutation rates in genes classified by functional category, rates of sequence evolution, patterns of expression and estimates of selection. Significance tested with χ^2 test, $n = 27{,}206$ genes, with raw P values tested against $\alpha = 0.05$ (unadjusted for multiple comparisons). Data show mean values for groups ± error bars reflecting 95% confidence intervals from bootstrapping. V_g, genetic variance of gene expression; 1001G, 1001 Genomes Project; LOF, loss of function; D_n, non-synonymous divergence; D_s, synonymous divergence; P_n, non-synonymous polymorphism; P_s, synonymous polymorphism; NS, not significant. **d**, Somatic mutations identified with very stringent criteria and using a caller specifically designed for rare somatic mutations, Strelka2, are reduced in gene bodies of wild-type plants, but not *msh6* mutants[6]. **e**, Left, general mechanism proposed in ref. 2. Right, new knowledge regarding biochemical mechanisms underlying reduced mutation rates in gene bodies established by recent discoveries in plants and synthesized in ref. 6. HR, homology-directed repair; MMR, mismatch repair[17–22].

acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-023-06315-x.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The TAIR10 *A. thaliana* reference genome can be found at https://arabidopsis.org/download. The more recent, improved *A. thaliana* reference genome can be found at https://github.com/schatzlab/Col-CEN. Sequencing reads for 107 *A. thaliana* mutation accumulation lines are stored in the National Center for Biotechnology Information Short Read Archive, accession number SRP133100. Additional mutation datasets were downloaded from publications cited in Supplementary Table 1.

Code availability

Code for this work uses functions maintained in https://github.com/greymonroe/polymorphology, with additional scripts and data for analyses and figures in https://github.com/greymonroe/mutation_bias_analysis2.

1. Wang, L., Ho, A. T., Hurst, L. D. & Yang, S. Re-evaluating evidence for adaptive mutation rate variation. *Nature* https://doi.org/10.1038/s41586-023-06314-y (2023).
2. Monroe, J. G. et al. Mutation bias reflects natural selection in *Arabidopsis thaliana*. *Nature* **602**, 101–105 (2022).
3. Liu, H. & Zhang, J. Is the mutation rate lower in genomic regions of stronger selective constraints? *Mol. Biol. Evol.* **39**, msac169 (2022).
4. Monroe, J. G. et al. Report of mutation biases mirroring selection in *Arabidopsis thaliana* unlikely to be entirely due to variant calling errors. Preprint at *bioRxiv* https://doi.org/10.1101/2022.08.21.504682 (2022).
5. Belfield, E. J. et al. DNA mismatch repair preferentially protects genes from mutation. *Genome Res.* **28**, 66–74 (2018).
6. Quiroz, D. et al. The H3K4me1 histone mark recruits DNA repair to functionally constrained genomic regions in plants. Preprint at *bioRxiv* https://doi.org/10.1101/2022.05.28.493846 (2022).
7. Stoler, N. & Nekrutenko, A. Sequencing error profiles of Illumina sequencing instruments. *NAR Genom. Bioinform.* **3**, lqab019 (2021).
8. Tran, H. T., Keen, J. D., Kricker, M., Resnick, M. A. & Gordenin, D. A. Hypermutability of homonucleotide runs in mismatch repair and DNA polymerase proofreading yeast mutants. *Mol. Cell. Biol.* **17**, 2859–2865 (1997).
9. Yang, S. et al. Parent–progeny sequencing indicates higher mutation rates in heterozygotes. *Nature* **523**, 463–467 (2015).
10. Weng, M.-L. et al. Fine-grained analysis of spontaneous mutation spectrum and frequency in *Arabidopsis thaliana*. *Genetics* **211**, 703–714 (2019).
11. Naish, M. et al. The genetic and epigenetic landscape of the *Arabidopsis* centromeres. *Science* **374**, eabi7489 (2021).
12. Monroe, J. G. et al. Author Correction: Mutation bias reflects natural selection in *Arabidopsis thaliana*. *Nature* https://doi.org/10.1038/s41586-023-06387-9 (2023).
13. Wang, L. et al. The architecture of intra-organism mutation rate variation in plants. *PLoS Biol.* **17**, e3000191 (2019).
14. Kim, S. et al. Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods* **15**, 591–594 (2018).
15. Gu, Y. et al. Human MutY homolog, a DNA glycosylase involved in base excision repair, physically and functionally interacts with mismatch repair proteins human MutS homolog 2/human MutS homolog 6. *J. Biol. Chem.* **277**, 11135–11142 (2002).
16. Niu, Q. et al. A histone H3K4me1-specific binding protein is required for siRNA accumulation and DNA methylation at a subset of loci targeted by RNA-directed DNA methylation. *Nat. Commun.* **12**, 3367 (2021).
17. Zhu, X. et al. Non-CG DNA methylation-deficiency mutations enhance mutagenesis rates during salt adaptation in cultured *Arabidopsis* cells. *Stress Biol.* **1**, 12 (2021).
18. Willing, E.-M. et al. UVR2 ensures transgenerational genome stability under simulated natural UV-B in *Arabidopsis thaliana*. *Nat. Commun.* **7**, 13522 (2016).
19. Ossowski, S. et al. The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science* **327**, 92–94 (2010).
20. Lu, Z. et al. Genome-wide DNA mutations in *Arabidopsis* plants after multigenerational exposure to high temperatures. *Genome Biol.* **22**, 160 (2021).
21. Jiang, C. et al. Environmentally responsive genome-wide accumulation of de novo *Arabidopsis thaliana* mutations and epimutations. *Genome Res.* **24**, 1821–1829 (2014).
22. Belfield, E. J. et al. Thermal stress accelerates *Arabidopsis thaliana* mutation rate. *Genome Res* **31**, 40–50 (2021).

Acknowledgements Research was conducted at the University of California, Davis, which is located on land that was the home of the Patwin people for thousands of years.

Author contributions J.G.M., K.D.M., W.X., T.S., P.C.-B. and D.W. contributed to data analysis. J.G.M., K.D.M., W.X., T.S., P.C.-B. and D.W. contributed to the writing. J.G.M., K.D.M., W.X., T.S., P.C.-B., C.B., M.L., M.E.-A., M.K., J.H., M.N., D.K., M.-L.W., E.I., J.Å., M.T.R., C.B.F. and D.W. contributed to the interpretation of the results. K.D.M. and W.X., who were not part of the study by J.G.M. et al.[2], carried out analyses to validate the impact of an improved genome reference sequence on reducing sequencing errors.

Competing interests The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-023-06315-x.

Correspondence and requests for materials should be addressed to J. Grey Monroe or Detlef Weigel.

nature portfolio

Corresponding author(s): Grey Monroe
Detlef Weigel

Last updated by author(s): Apr 28, 2023

Reporting Summary

Nature Portfolio 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 Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the 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
☒	☐	A statement on 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 statistical parameters 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

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	GATK V4, bwa 0.7, Strelka2 2.9.10
Data analysis	Analyses were conducted with R version 4.2.2, github.com/greymonroe/mutation_bias_analysis2, github.com/greymonroe/polymorphology

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 and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio 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 description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The TAIR10 A. thaliana reference genome is found at Arabidopsis.org/download. The more recent improved A. thaliana reference genome is found at https://github.com/schatzlab/Col-CEN. Sequencing reads for 107 A. thaliana mutation accumulation lines are stored in the NCBI Short Read Archive (SRA) accession number SRP133100. Additional mutation datasets were downloaded from publications cited in Supplemental Table 1.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is 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/documents/nr-reporting-summary-flat.pdf

Life sciences study design

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

Sample size	Sample sizes (number of studies examined) were based on the total published literature reporting mutation rates in Arabidopsis thaliana. Citations of these publications are found in Supplemental Table 1.
Data exclusions	Data were filtered according to descriptions in the manuscript. For example, variants were filtered to remove cases of being potentially caused by homopolymer bleed-through.
Replication	We tested whether mutation biases reported were replicated by 10 independent mutation accumulation datasets. In all cases, there was an observation of lower gene body mutation rates, with the exception of MSH2 knockout lines (Belfield et al. 2018) which indicates that MSH2 preferentially repairs gene bodies.
Randomization	NA: Data were observational measures of mutation rate variation across the genome from mutation accumulation experiments.
Blinding	NA: Data were observational measures of mutation rate variation across the genome from mutation accumulation experiments.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

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