

A statistical estimator for determining the limits of contemporary and historic phenology

William D. Pearse^{1,2,3*}, Charles C. Davis⁴, David W. Inouye^{5,6}, Richard B. Primack⁷
and T. Jonathan Davies^{1*}

Climate change affects not just where species are found, but also when species' key life-history events occur—their phenology. Measuring such changes in timing is often hampered by a reliance on biased survey data: surveys identify that an event has taken place (for example, the flower is in bloom), but not when that event happened (for example, the flower bloomed yesterday). Here, we show that this problem can be circumvented using statistical estimators, which can provide accurate and unbiased estimates from sparsely sampled observations. We demonstrate that such methods can resolve an ongoing debate about the relative timings of the onset and cessation of flowering, and allow us to place modern observations reliably within the context of the vast wealth of historical data that reside in herbaria, museum collections, and written records. We also analyse large-scale citizen science data from the United States National Phenology Network and reveal not just earlier but also potentially more variable flowering in recent years. Evidence for greater variability through time is important because increases in variation are characteristic of systems approaching a state change.

Anthropogenic climate forcing is likely to increase the global temperature by more than 1.5°C by the end of this century¹. In response to this rapid environmental shift, species must track favourable conditions by moving or altering the timing of their life-history strategies—their phenology—to flower, breed or migrate sooner^{2,3}. However, predicting species' phenological responses is not straightforward: experimental data often do not match observations⁴, and sampling of observational data is frequently limited. Citizen scientists⁵ and historical collections^{6,7} have emerged as valuable sources of ecological data and ongoing efforts to digitise museum and herbarium collections are making an unprecedented wealth of historical records available^{8–11}. Despite their promise, such data present numerous statistical challenges: they are often sparsely sampled spatially and unevenly distributed through time¹², and while they can provide information on the relative timing of events they do not necessarily capture their first occurrence. Compounding this problem, most statistical tools are designed to study changes in species' mean responses, not variation in the onset of events.

Here, we present a method derived from the extinction biology literature¹³ to address these challenges. We also provide three case studies that illustrate the potential of the approach in phenological research. While we focus on plant flowering time, this approach would also be applicable to other systems, such as the phenology of bird migrations and insect emergence, or the limits of other continuous data such as environmental tolerances. First, we revisit an ongoing debate about shifts in the timing of the onset, peak (middle) and cessation of flowering. Second, we show how our approach can reconcile distinct datasets with different sampling (historical collections and field observations), greatly expanding the temporal and climatic ranges across which we can measure change. Third, we apply our method to a sparsely sampled citizen science dataset

and find evidence that climate change is not just altering the timing of plant flowering, but also increasing its variability through time.

Results and discussion

Statistically estimating the start of a process. Estimating the onset of a phenological event is just one instance of the more general problem of determining the absolute limit of a distribution. The tails of distributions are infamously difficult to model because there are fewer data to parameterize them and a single data point can invalidate all previous estimates. This challenge is similar to the 'German tank problem', which was faced by Allied forces during World War II who wanted to estimate the number of German tanks (the limit of the distribution of serial numbers) but only had access to the sequential serial numbers of observed (defeated) tanks¹⁴. Here, we suggest a solution to this problem that parallels methods first described to determine the date a species went extinct¹³. The general approach is to model the distribution of the earliest observations using a (very flexible) Weibull distribution, which provides an estimate of the start of the observed process (for example, plants flowering). The joint distribution of the most recent sightings has approximately the same Weibull form, irrespective of the distribution from which those sightings were sampled¹⁵, making it well-suited to data collected under different sampling regimes. The estimate for the first occurrence of any event is thus the sum of the times of the first k events, weighted in part according to the joint Weibull distribution of all the sightings (following Roberts and Solow¹³ who focused on the last k events). While confidence intervals (CIs) are defined for this estimate, the s.e. must be parametrically bootstrapped as their formula is currently unknown¹⁶. Figure 1 gives an example of how this approach can provide an estimate of when a process (such as flowering) started, even if the very beginning of that process was not directly observed.

¹Department of Biology, McGill University, Montréal, Québec H3A 1B1, Canada. ²Department of Biological Sciences, Université du Québec à Montréal, Montréal, Québec H2X 1Y4, Canada. ³Department of Biology and Ecology Center, Utah State University, 5305 Old Main Hill, Logan, UT 84322, USA. ⁴Department of Organismic and Evolutionary Biology and Harvard University Herbaria, Harvard University, 22 Divinity Avenue, Cambridge, MA 02138, USA. ⁵Department of Biology, University of Maryland, College Park, MD 20742, USA. ⁶Rocky Mountain Biological Laboratory, PO Box 519, Crested Butte, CO 81224, USA. ⁷Biology Department, Boston University, 5 Cummington Street, Boston, MA 02215, USA. *e-mail: will.pearse@usu.edu; j.davies@mcgill.ca

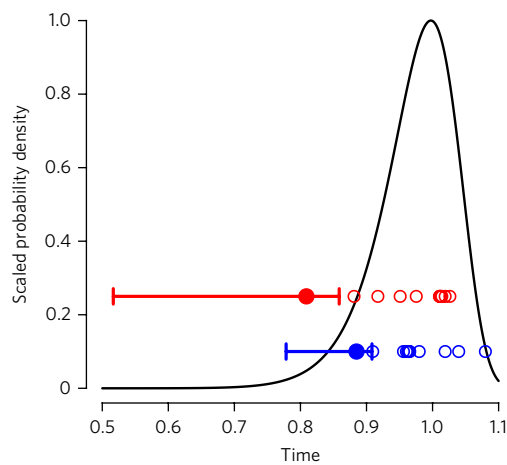


Fig. 1 | Example demonstration of the difference between our method and taking first observations at face value. Two draws of ten samples (open red and blue circles) from the same Weibull distribution (whose probability density is in black) are shown. Our estimates of the lower limit (start) of the distribution are shown in filled circles, with CIs also shown. Two advantages of this new method are clear: (1) the estimates have CIs; and (2) the estimates themselves are closer to the true onset of the process (time 0.5) than the first sample. This results from drawing strength across all observations, not simply the single earliest observation. More details and simulations confirming these intuitive properties are given in the Methods.

Using simulations, we demonstrate that our approach has greater power to detect the true onset of a process than existing methods that use only the first observation (see Methods). This is because our approach draws strength from the first k measurements, not just the single earliest observation. This also allows for CI and s.e. to be placed around an estimate, which is impossible when working with the first observation alone. Just as any measure of the central tendency of a distribution (for example, a mean) should not be considered in isolation of the distribution and number of observations underlying it, the same is true of estimates of the limits of a distribution. We also note that attempting to estimate the limit of a distribution by averaging across estimates—as is common in phenological studies—is inherently biased: the average of the two (or more) earliest observations must, by definition, be later than the earliest observation. This has implications not just for generating mean estimates of the onset of flowering, but also for commonly used statistical models that implicitly rely on averages (for example, analysis of variance and multiple regression). The following case studies illustrate the potential of our approach.

Relative change in the onset, peak and cessation of flowering.

First, we re-examined a comprehensive dataset of over two million observations made throughout the past 39 years in the Rocky Mountains of Colorado^{17,18} to explore changes in the onset, peak and cessation of flowering. Previous work on this detailed dataset reported discordance in temporal shifts among phenophases¹⁹. This finding suggests that communities of co-flowering species may be profoundly altered under climate change, with potentially negative consequences for currently co-occurring pollinator and herbivore communities²⁰. Using our approach here, which controls for differences in sampling, we find, surprisingly and to the contrary, a close alignment of change through time among these three aspects of flowering phenology in the same data (Fig. 2). As we are able to measure the confidence in our estimates, our approach allows us to overcome implicit sampling biases in the observation data. For example, there is both theoretical and empirical evidence that greater

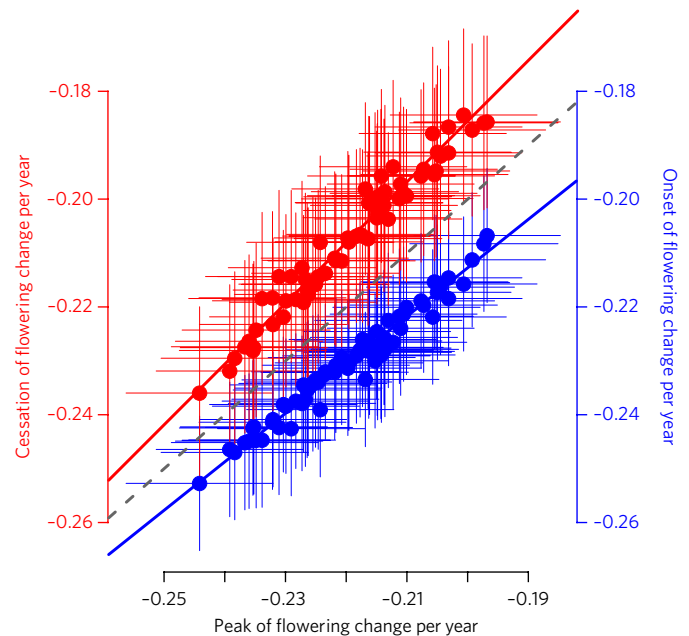


Fig. 2 | Rates of change of the onset, bulk and cessation of flowering through time are tightly correlated in the Rocky Mountain dataset.

This contrasts with a previous analysis not using our approach¹⁹. Each point represents a species' rate of change (per year) of first (blue) and last (red) flowering, plotted as a function of the change in peak flowering (bottom axis). The coloured lines emanating from each point represent the s.e. of each species' change estimate. The thick, solid blue (onset of flowering; slope = 0.99, 95% CI: 0.90-1.08) and red (cessation of flowering; slope = 1.02, 95% CI: 0.91-1.13) lines are best-fit lines from a Deming regression accounting for error in both variables. The grey dashed line is a 1:1 line for reference and is the expectation if the dates of the onset, bulk and cessation of flowering were changing at the same rate in the data. Species estimates are taken from an overall model that accounts for species abundance; each model had an r_{adjusted}^2 greater than 74%.

sampling effort increases the chances of observing an event earlier²¹. Such sampling biases are difficult to avoid when using the first (or last) observation as a measurement, but can be corrected for when working with a statistical estimator derived from sampling theory, as used here. While it is uncertain whether these results hold elsewhere, the unprecedented degree of sampling in this system urges a re-assessment of this controversial aspect of plant phenology.

Reconciling historic herbarium and field observations. Second, we contrasted estimates of first flowering derived from herbarium records with a well-studied historical dataset on flowering times from Massachusetts (USA) initiated by H. D. Thoreau in the 1850s. Despite the age and richness of herbarium data, the records are unevenly sampled through time, making direct comparisons between datasets challenging²¹. While there is a strong correlation between rates of change in herbarium and field observations, herbarium records tend to sample peak flowering better, such that recorded dates of first flowering from the two datasets are not directly comparable²². As we show in Fig. 3, by applying our approach, we directly reconcile estimates of first flowering from these two datasets despite differences in sampling: the two datasets not only show correlated changes through time, but now show dates of flowering that coincide. This is because our approach can use the collection dates of herbarium records to generate a statistical estimate of the onset of flowering, despite having no direct records of the actual onset. This provides hope that our approach can be used to reconcile modern and historical datasets, increasing our power to

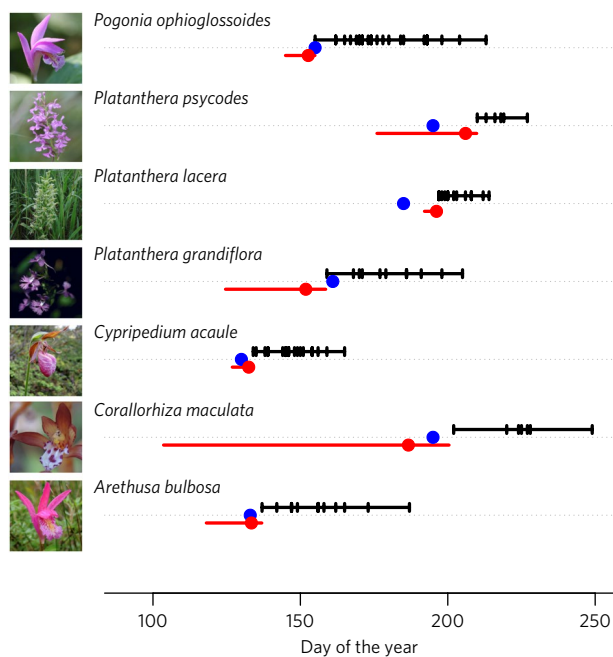


Fig. 3 | Reconciling flowering phenology in two historic datasets. Data collected between 1858 and 1902 as part of two datasets^{21,22} were used, corresponding to the period of greatest overlap when A. Hosmer and H. D. Thoreau were collecting phenological data. The horizontal black line represents the range of herbarium records and the vertical ticks represent each observation. The blue circles represent the earliest field observation and the red circles our modelled estimate of onset from the herbarium records (with 95% CIs also in red). Our approach produces estimates that are, on average, almost 4.96 days closer to the true onset of flowering, as recorded by H. D. Thoreau and A. Hosmer, than the earliest herbarium record (paired test comparing differences between the earliest observation and modelled onset: $t_e = -2.61$, $P = 0.0399$). We acknowledge that this approach does not account for variation across years, which is mainly driven by annual temperature variation^{35,39}. Species with fewer than six herbarium or field observations were excluded from the analyses; see Methods for more details. Credit: images of plant species *Pogonia ophioglossoides*, *Platanthera lacera*, *Cypripedium acaule*, *Corallorhiza maculata* and *Arethusa bulbosa*, Steven J. Baxter; *Platanthera psycodes*, Rob Routledge/Sault College; *Platanthera grandifolia*, Arnold T. Drooz/USDA Forest Service.

detect whether current conditions differ from those in the past and mitigating shifting baseline syndrome²³. In addition, by leveraging the vast wealth of data in herbaria, our method allows us to dramatically expand the climate space within which we can study plant phenological responses²², which is currently strongly biased towards northern temperate biomes²⁴.

Increased variation in flowering phenology across the USA.

Third, we applied our method to phenological observations from the National Phenology Network (NPN; <https://www.usanpn.org/>)—one of the largest citizen science monitoring schemes, with more than one million records spanning continental USA over the past decade. In parallel with the increasing appreciation and use of collections data, citizen science has emerged as a powerful tool for collecting large volumes of data across broad taxonomic and spatial scales⁵. However, like herbarium records, such data often suffer from poor sampling for rare or difficult to identify events, potentially biasing estimates for those species most at risk from climate change. As our method requires relatively few samples (see Methods), it is well-suited for such cases. For our analysis,

Table 1 | Modelled estimates of first flowering date in the NPN data

Selected model coefficients ^a	0.5%	2.5%	Median	97.5%	99.5%	s.d.
Overall mean (μ)	94.80	96.99	105.03	113.58	116.33	4.27
Annual change	-5.35	-4.61	-2.68	-0.99	-0.48	0.93
Temperature	-1.34	-1.08	-0.29	0.45	0.67	0.38
Precipitation	-0.10	-0.07	0.03	0.13	0.18	0.05
Overall variation	22.92	23.30	24.48	25.86	26.30	0.66
Annual variation change (σ^2)	-0.82	-0.59	0.16	0.90	1.13	0.39
Pooling estimates ^b	Mean	Median	s.e.	s.d.		
Species	1.00	1.00	0.013	0.00187		
State × Year	0.58	0.54	0.234	0.03482		
State	0.12	0.13	0.052	0.00650		
State × Year	0.75	0.81	0.238	0.03001		
r^2	0.53	0.53	0.010	0.00018		

See Fig. 4 for plots of the model output through time. We modelled the onset of flowering as a function of species-specific responses and environmental conditions (see Methods). All coefficients are summaries of Bayesian credible intervals (not frequentist CIs) taken from 3,200 samples across 16 Markov chain Monte Carlo runs with all $n_{\text{iter}} > 3,000$ and $\hat{R} = 1$ (see Methods for more details and all model coefficients).^aModel coefficients taken from the posterior distribution of the model (see Methods for all coefficients). The first four rows describe changes in the date of flowering through time, whereas the last two rows describe how variation about the average flowering date changes through time. These provide support for earlier flowering in hotter years and locations, along with more variable flowering through time.^bEstimates of the degree of pooling²⁷ for species and state means and changes through time ('Year') in the data. Pooling indicates the extent to which estimates at each level within a multi-level model vary; values close to 0 indicate variation, whereas values close to 1 indicate no variation. Thus, these results suggest that individual species' flowering times varied independently, but state-level effects did not vary to the same extent.

we calculated an estimate of first flowering for each species, in each year and state, with more than five records. As the potential for sampling error in such a broad dataset is high, we used a hierarchical Bayesian approach that allowed us to propagate error clearly throughout every stage of the analysis. Such models are robust to over-parameterisation²⁶, and so we can model each species with a hierarchically drawn intercept and slope of change through time.

Our model has two main components: (1) systematic variation in the date of first flowering as a function of the species, state where it was observed, and year of an observation; and (2) estimated variation in the date of first flowering (full details are presented in Methods). Our model finds increases in the first flowering date of 2.49 days from 2009 to 2015 on average within New York (the state with the most data in our model; see Table 1, 'Selected model coefficients' and Methods), but average rates of change mask significant variation among species (Table 1, 'Pooling estimates'). The flowering date is negatively associated with temperature—warmer temperatures result in earlier flowering—however, estimates of the pooling of the overall mean date of first flowering among species and states suggest that, once climate is accounted for, species' flowering dates are relatively invariant among states (see Table 1, 'Pooling estimates'). Taken together, these results indicate that species are responding consistently to climate across continental USA.

There are two reasons to be cautious when interpreting the magnitude of these flowering responses to temperature through time. First, we only used data covering the period 2009–2015, so our model may not capture the decadal dynamics of flowering responses. However, our model is consistent with independent data across the period 2001–2008 (Fig. 4) whose mean date of first flowering is later than that of 2009–2015 (as predicted by our model;

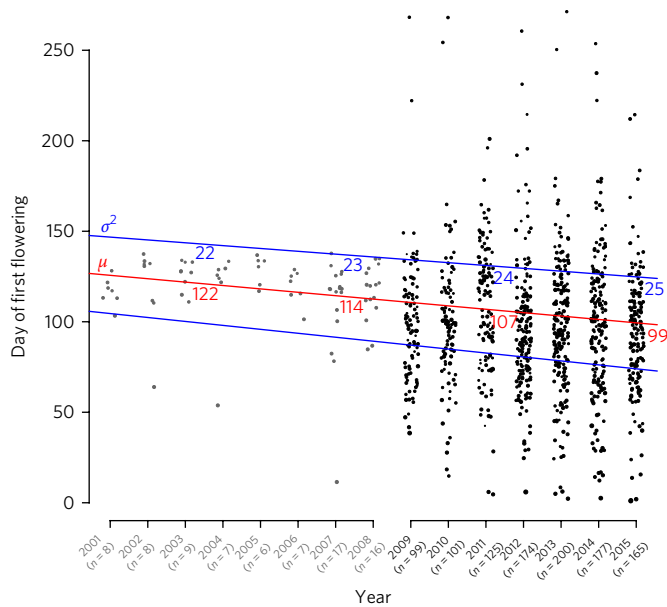


Fig. 4 | Annual variation in flowering phenology throughout North America in the NPN data. On the vertical axis, we plot the estimated date of first flowering, with the point size inversely proportional to the s.e. of the estimate. The red line is the average estimate of flowering time through time (μ in Table 1), whereas the blue upper and lower lines are the modelled variance of flowering through time (σ^2 in Table 1). Estimates for particular years are labelled on the graph. The trend for earlier flowering through time is shown, as well as the increase in the variability of the first flowering date through time. We plot data from 2001 to 2008 that were not used to parameterize the model in grey, to show the predictive power of the model for novel data.

$t_{77} = 4.30$, $P < 0.0001$). Second, our model suggests an increase in the variability of the date of first flowering through time (Table 1), which is also visible in Fig. 4. This increase probably obscures the degree of phenological change we are already experiencing in North America. Conservatively comparing our modelling results for 2011 and 2015, the variation in the first flowering date has increased by 13% (coefficient of variation ($\frac{\sigma^2}{\mu}$); see Fig. 4).

That variability is increasing through time is important as increases in the unpredictability of, and variation in, a system are thought to be indicative of a system approaching a regime shift^{27,28}. There is accumulating evidence that species are approaching the limit of their capacity to adapt their phenology to climate change^{29–31}, and we suggest that our results are consistent with species being pushed to their limits of phenological adaptation. Using a Bayesian approach to model fitting, we were able to estimate the relative support for our hypothesis and found that it is twice as likely that the variance is increasing through time than decreasing (on the basis of posterior densities; see Table 1 and Methods). It is possible that the expansion of the NPN scheme through time might have contributed to this pattern. However, we found a similar tendency for increasing variation in the more detailed and consistently sampled Rocky Mountain dataset, with much greater confidence (99.15% probability of increase through time; see Methods). Detecting such an increase in variation through time would be difficult, if not impossible, in studies using space-for-time substitutions or lacking a hierarchical modelling framework such as ours.

Conclusion

The dual approach presented here—accounting for uncertainty around estimates and using a modelling framework that allows uncertainty to percolate through into predictions—allows for

more robust understanding of climate-driven phenological shifts. By drawing information from the sampled distribution of records and not simply the first observation, our approach accurately estimates the timing of first events from sparsely collected data. We show how this has far-reaching consequences for our understanding of flowering phenology and allows us to marry historic and modern datasets, vastly increasing the temporal and climatic range over which we can study phenological change. Applying our method to one intensively studied field dataset and another continental-scale citizen science dataset, we find tentative evidence for an increase in the variability of phenology through time. Increases in variation may have profound implications for ecosystems, and additional research is urgently needed to examine whether these patterns generalize beyond the North American continental and local-scale botanical systems presented here.

Methods

All analyses were conducted in R version 3.3.2 (ref. 32).

A new approach to estimating the start of a process. Roberts and Solow¹³ gave formulae to produce an estimate for the end of a process, as well as CIs for that estimate. These same formulae can be used to estimate the beginning of a process if the values are sorted ascendingly. In the accompanying supplement, we provide code (headers.R) to perform these calculations that re-creates the exact values as reported by Roberts and Solow¹³ in their original manuscript. Figure 1 gives a graphical example of the difference between our approach and that of taking the first observation at face value.

We were unable to find an analytical solution for the s.e. of the onset or end of events, and so used a parametric bootstrap to estimate its error (code also in headers.R). For this, we estimated the shape parameter of the joint Weibull distribution of sighting times, drew 100 samples of the same size as our observed sample from a distribution parameterized by the estimated shape parameter and calculated the s.d. of the samples. Note that, as is clear from Fig. 1, the CIs generated from this approach are not symmetrical; we therefore caution against the uncritical use of the width of the CIs as an estimate of error of an estimate.

Our approach cannot be used when all observations are made at exactly the same time, or when the first and last onset and cessation observations are exactly identical, so our code removes all such duplicates and issues a warning. When measurements have been made on only two or fewer unique dates or times there can be no estimate of the onset or end, so our code returns a ‘not applicable’ value and again issues a warning.

Finally, we note that very large samples of observations are not as informative as might be expected using this method, as the standard gamma distribution on which it is based greatly weakens the influence of observations far from the tail of the distribution being estimated. This makes a degree of intuitive sense: when estimating the onset of a process, the end of it has very little information content (and vice-versa). In our experience, the weakening is such that examining more than the earliest or latest 30 observations is unnecessary; the influence of such values is so low that it can go beyond the numerical precision of some R instances and cause errors. In all of the analyses below, we used a maximum of the 50 earliest observations; concerned users can alter this using the k parameter in our code.

Our approach versus the first observation. To examine our power to detect the true onset of a process, we examined type I error rates—when the two-tailed 95% CIs of our estimate overlapped the true value of the onset of the process. Fifty times each, we drew n samples from a uniform distribution ranging from 0 and m across all combinations of n and m , where n was 4, 5, ..., 49, 50, and m 20, 21, ..., 349, 350. We consider these ranges and sampling regimes to reflect the kinds of phenological data frequently used (that is, the sample size of observations and the day of the year on which flowering was first observed). For these simulations, 0 was the true onset of the process, even if a sample was not drawn with a value of 0 (that is the statistical limit of the uniform distribution from which we were sampling). When using 95% CIs (α_{95}), we would typically expect an 80% chance of producing CIs that encompass the true value (that is, a statistical power ‘ β' ’ of 80%); we exceeded this expectation in 93% of the parameter combinations. As Supplementary Fig. 1 shows, the overwhelming majority of cases in which we had poorer power were when we had fewer than ten samples (the left-hand side of the figure). We thus consider our approach to have high power.

To contrast our approach with assuming the first observed value as the onset of a process, we also recorded the least (in our context, earliest) observation while performing the same simulations as above. Supplementary Fig. 2 shows the percentage error of the estimate ($\frac{\text{estimate} - \text{true value}}{\text{range}} \times 100$). Note that it is impossible to perform a direct quantitative comparison of these two approaches: our method produces a statistical estimator with an associated degree of error, while the first observation is a single observation for which there is no meaningful estimate of confidence.

The first estimate under-estimates the onset of flowering in many cases; a log-unit increase or decrease of the range in the sampling results in a log-unit increase in the percentage error (Supplementary Fig. 2). Thus, as the duration of a process increases, the amount of sampling required to estimate the true onset accurately increases. That uncritical use of the first observation is biased is uncontroversial; it is well-known that the first observation of a flower in bloom is strongly affected by sampling effort²¹. Even while keeping the variance of a distribution constant, sampling it more times gives more opportunity for a more extreme event, by chance, to be sampled—the limits of most statistical distributions are infinite. Our approach, which produces a statistical estimator, can account for this, which is not possible when working with the first estimate.

We also note that the accuracy of this method has been empirically verified by Clements et al.³³, who examined its ability to detect local extinction accurately under different sampling regimes and experimental conditions.

Colorado Rocky Mountains data. Data are from CaraDonna et al.¹⁹ and consist of regular surveys carried out in the Colorado Rocky Mountains (USA); from 1974 to 2012, 30 square 4 m² plots were surveyed and the number of flowers was counted on each individual every two days. Following CaraDonna et al.¹⁹, we restricted our analyses to those species for which there were records in at least half of the dataset (19 years). Estimates for each species were calculated for each plot within each year; if such a grouping had fewer than ten measurements we excluded that measurement as we wished to model changes in variability and we did not want to include less precise estimates that could inflate variation. Our power analyses (see above) suggested that ten samples was sufficient to estimate the onset of a process with reasonable confidence. We included log-transformed abundance as a factor in our analyses.

Colorado Rocky Mountains—onset versus peak versus cessation. The models presented in the Results and Discussion regress onset and cessation of flowering against peak (median) flowering, ignoring variation among species and abundance. To account for these factors following an earlier analysis of this dataset¹⁹, we fitted full linear models incorporating species' identities and their interaction with year, and a separate additive effect of abundance. The model results for the shifts in the onset, peak and cessation of flowering can be seen in Supplementary Tables 1, 2 and 3, respectively, and each model had an r^2_{adjusted} value greater than 74%. We then performed Deming regressions of species-level changes in onset (slope = 0.99, 95% CI: 0.90–1.08) and cessation (slope = 1.02, 95% CI: 0.91–1.13) of flowering through time as a function of peak flowering. Deming regressions were performed using *deming*³⁴ and account for error in estimates of change in both predictor and response variables.

Historical comparisons of phenology. Data were taken from Davis et al.²² and consist of herbarium records and direct field observations from the surroundings of Concord (MA, USA). These historical collections reflect four main periods of sampling: records collected by Thoreau (1852–1858), Hosmer (1878 and 1888–1902), Miller-Rushing and Primack (2003–2006) and Davis and Connolly (2011–2013)^{35,36}. We restricted ourselves to only those samples collected before 1903, as this time period overlapped best with the collection of herbarium specimens and it was the comparison between these two sets of observations that we were most interested in. The herbarium data themselves were extracted from the Harvard University Herbaria, New York Botanical Garden's William and Lynda Steere Herbarium, Yale University Herbarium and University of Connecticut's George Safford Torrey Herbarium by Davis et al.²². A specimen was recorded as flowering if over 75% of its flowers were open (if multiple flowers were present in a specimen); for more details see Davis et al.²². We analysed species that were common to both datasets and that had (at a minimum) six dated herbarium records. We estimated the onset of flowering and CIs in these data as described above and plotted the results in Fig. 3.

NPN data. Data were downloaded from the NPN, including observations only from 1 January 2001 to 13 February 2017 (the date of download); the species functional type was set to 'deciduous broadleaf' and the phenophase category to 'leaves, flowers', focusing on data collected from continental USA. Only events referring to flowers were retained for analysis; specifically, those with 'flower' and 'bloom' (but not 'end' or 'pollen') in their phenophase descriptions. Observations were split according to species, state and year, and estimates of first flowering (and their s.e.) calculated across these groupings were the basis of the analysis. Temperature and precipitation data were taken from the University of East Anglia's Climatic Research Unit high-resolution gridded historical datasets (version 3.24.01; ref. ³⁷) and annual mean values for each state were calculated on the basis of state outlines taken from the Global Administrative Areas dataset (version 2.8; <http://www.gadm.org/>). Since these temperature data are currently only available from 1901 to 2015, we restricted our analyses to estimates of first flowering between 1 January 2005 and 31 December 2015.

In the analyses presented in the Results and Discussion, we (conservatively) limited our analyses to species–site–year estimates with at least five observations, and excluded species with fewer than ten species–site–year estimates. This provided 1,041 observations across a total of 63 species in 45 states, covering the

period 2009–2015. All parameter estimates from these analyses are presented in Supplementary Table 4. Here, we also present results from a model fit to all data from 2009 to 2015 (1,249 observations of 150 species in 46 states) and show that the results are qualitatively identical (Supplementary Table 5). In addition, as the coverage of the data is markedly increased after 2009 (see Fig. 4), we fit models to data collected from 2001 to 2015. Results from the 2001 to 2015 data limited to species–site–year estimates with at least five observations and excluding species with fewer than ten species–site–year estimates (1,119 observations of 63 species in 45 states) are given in Supplementary Table 6. Results from all data from 2001 to 2015 (1,327 observations of 150 species in 46 states) are given in Supplementary Table 7. All year, temperature, precipitation, longitude (of state centroid) and latitude (also of state centroid) data were scaled to have a mean of 0 and s.d. of 1 to make model coefficients directly comparable (following ref. ²⁶). Model coefficients were back-transformed to their original scales in the Results and Discussion, but not in the Supplementary Tables.

NPN hierarchical modelling. We computed our model using *rstan*³⁸ in each dataset, running a total of 16 chains for 20,000 iterations, sampling every 50 iterations and discarding the first 10,000 iterations as burn-in. All models were checked graphically for convergence and mixing, and $\hat{\rho}$ values were all equal to 1. In the Results and Discussion, we report that it is twice as likely that the variation in the date of first flowering is increasing through time than it is not (that is, that $\epsilon_{\beta} > 0$; see below for definitions); we base this on the observation that 66.67% of the posterior distribution of ϵ_{β} was greater than 0.

In Fig. 4 we show for reference points from 2001 to 2008 that were not used to fit models to data. Visual posterior predictive checks were also performed on all model results to ensure model validity. We draw the reader's attention to the greater support for our main result (increased variance through time as measured with the parameter ϵ_{β} ; see below) in the model fitted to the longer time series (Supplementary Table 6); we consider it more conservative, and so preferable, to present the more modest coefficients in the main text of the manuscript.

The general structure of our model is described in the Results and Discussion; here, we present it more formally. Specifically, the higher-level structure of the model is as follows:

$$\text{DOY} \sim N\left(\alpha_0 + \mu_{\text{sp}} + \mu_{\text{env}} + \mu_{\text{space}} + \mu_{\text{space-time}}, \epsilon\right) \quad (1)$$

where DOY is the estimated 'day of year' of first flowering and α_0 is the overall first flowering date. For ease of presentation, we have grouped the model parameters together: the terms μ_{sp} and μ_{env} describe species and environmental effects, μ_{space} and $\mu_{\text{space-time}}$ account for spatial and temporal auto-correlation, and ϵ describes changes in the variance of DOY through time. We describe each below.

Species-specific changes through time, μ_{sp} is defined as:

$$\mu_{\text{sp}} = \alpha_i + \beta_0 \times \text{Year} + \beta_j \times \text{Year} \quad (2)$$

where α_i is the difference from the overall mean (α_0) for each species (i), Year is the year of an observation, β_0 is the slope of the overall change in DOY through time and β_j is the difference in that slope for each species.

The environmental determinant of DOY, μ_{env} is defined as:

$$\mu_{\text{env}} = \tau \times \text{Temp}_{j,k} + \pi \times \text{Precip}_{j,k} \quad (3)$$

where τ quantifies the effect of the mean annual temperature ($\text{Temp}_{j,k}$) of an observation's state (j) in a given year (k), and π the effect of the mean annual precipitation ($\text{Precip}_{j,k}$) of an observation's state in a given year.

Each state's residual variation in DOY, both overall (μ_{space}) and through time ($\mu_{\text{space-time}}$), is expressed similarly. μ_{state} is defined as:

$$\mu_{\text{state}} = \alpha_j + x_{\alpha} \times \text{Long}_j + y_{\alpha} \times \text{Lat}_j + z_{\alpha} \times \text{Long}_j \times \text{Lat}_j \quad (4)$$

where α_j is the difference from the overall mean (μ_0) for each state, and x_{α} and y_{α} measure variation in DOY longitudinally (Long) and latitudinally (Lat), respectively. z_{α} captures the interaction of latitude and longitude. Note that each state's (j) latitude and longitude is measured as the centroid of a state, as described above. The influence of each state may also vary through time, as captured in the definition of $\mu_{\text{space-time}}$:

$$\mu_{\text{space-time}} = \text{Year} \times \left(\beta_j + x_{\beta} \times \text{Long}_j + y_{\beta} \times \text{Lat}_j + z_{\beta} \times \text{Long}_j \times \text{Lat}_j\right) \quad (5)$$

where β_j is the difference from the overall change through time (β_0) for each state, and x_{β} and y_{β} measure variation in DOY longitudinally and latitudinally through time, respectively. z_{β} captures the interaction of latitude and longitude through time.

Finally, but importantly, the term ϵ measures the overall variance of DOY:

$$\epsilon = \epsilon_0 + \beta_{\epsilon} \times \text{Year} \quad (6)$$

where ε_0 is the overall variance (error) in our data and β_i is the change in that variance through time.

The species-specific parameters were drawn from previous distributions centred at 0 with estimated variances. Specifically:

$$\alpha_i \sim \text{Normal}(0, \sigma_{\alpha_i}) \quad (7)$$

$$\alpha_j \sim \text{Normal}(0, \sigma_{\alpha_j}) \quad (8)$$

$$\beta_i \sim \text{Normal}(0, \sigma_{\beta_i}) \quad (9)$$

$$\beta_j \sim \text{Normal}(0, \sigma_{\beta_j}) \quad (10)$$

Other parameters were given normal priors with wide distributions so as to be uninformative, specifically:

$$\alpha_0, \beta_0, x, y, z, x_p, y_p, z_p \sim \text{Normal}(0, 1,000) \quad (11)$$

With the exception of the variance parameters, for which our priors were:

$$\varepsilon_0, \sigma_{\alpha_i}, \sigma_{\alpha_j}, \sigma_{\beta_i}, \sigma_{\beta_j} \sim \text{Uniform}(0.0001, \text{Infinity}) \quad (12)$$

$$\varepsilon_{\beta} \sim \text{Uniform}(-10, 10) \quad (13)$$

Colorado Rocky Mountains—hierarchical modelling. Within the Results and Discussion, we refer to a hierarchical model of the onset of species' flowering times in the Rocky Mountain dataset, which we describe here in full.

We computed our model using *rstan*³⁸ in each dataset, running a total of 16 chains for 20,000 iterations, sampling every 50 iterations and discarding the first 10,000 iterations as burn-in. All models were checked graphically for convergence and mixing, and \hat{r} values were all equal to 1.

The structure of our model, which is comparable to that of the NPN model above, is as follows:

$$\text{DOY} \sim N(\alpha_i + \beta_i \times \text{Year} + \gamma \times \text{Abundance}, \varepsilon_0 + \varepsilon_{\beta} \times \text{Year}) \quad (14)$$

where DOY is the estimated 'day of year' of first flowering, α_i is the mean DOY for each species (i), β_i is the slope of annual change of DOY for each species, γ is a slope accounting for abundance-driven changes, ε_0 is the mean variance of DOY and ε_{β} is the rate of change of variance through time. The terms 'Year' and 'Abundance' represent the recorded year and abundance of species within each plot, respectively. These terms are similar to those used for the NPN model (described above).

α_i and β_i are species-specific parameters and are drawn from distributions parameterized as follows:

$$\alpha_i \sim \text{Normal}(\alpha_0, \sigma_{\alpha_i}) \quad (15)$$

$$\beta_i \sim \text{Normal}(\beta_0, \sigma_{\beta_i}) \quad (16)$$

Most parameters were given normal priors with wide distributions so as to be uninformative, specifically:

$$\alpha_0, \beta_0, \gamma_0, \varepsilon_0, \varepsilon_{\beta} \sim \text{Normal}(0, 1,000) \quad (17)$$

The only exceptions to this were our hyper-parameters of variance, for which such priors would be inappropriate (negative variances are impossible). Our hyper-parameter priors were:

$$\sigma_{\alpha_i}, \sigma_{\beta_i}, \varepsilon_0 \sim \text{Uniform}(0.0001, \text{Infinity}) \quad (18)$$

All parameter estimates from this model are given in Supplementary Table 8. In the manuscript (section 'Increased variation in flowering phenology across the USA') we refer to evidence that the variance in the onset of flowering in the Rocky Mountain dataset has been increasing through time. This is supported by the estimates of ε_{β} in Supplementary Table 8, whose high-credibility intervals (and s.e. and deviations) suggest a positive (non-zero) change through time. In the Results and Discussion, we report a 99.15% probability that the variation in the date of first flowering is increasing through time (that is, that $\varepsilon_{\beta} > 0$); we base this on the observation that 99.15% of the posterior distribution of ε_{β} was greater than 0.

Variation among species in flowering time. There is growing evidence that early-flowering species are changing their phenology more strongly in response to climate change. One of the advantages of our hierarchical approach is that it permits the examination of variation among species' responses, while propagating uncertainty for each species' response through into the final analysis. In Supplementary Figs. 3 and 4, we plot the species-level changes in flowering phenology through time as a function of overall first flowering data for both the Rocky Mountain and NPN data, respectively. We provide these data as a test of the overall validity of our approach and note that the Rocky Mountain data show some support for two kinds of flowering regime (early versus late).

Life Sciences Reporting Summary. Further information on experimental design and reagents is available in the Life Sciences Reporting Summary.

Data availability. All the data we have analysed are publicly available at the references we provide above. The Colorado data are archived through the Open Science Framework at <https://osf.io/jt4n5/>.

Received: 22 June 2017; Accepted: 20 September 2017;

Published online: 06 November 2017

References

1. IPCC *Climate Change 2014: Synthesis Report* (eds Core Writing Team, Pachauri, R. K. & Meyer L. A.) (Cambridge Univ. Press, Cambridge, 2015).
2. Menzel, A. et al. European phenological response to climate change matches the warming pattern. *Glob. Change Biol.* **12**, 1969–1976 (2006).
3. Parmesan, C. et al. Ecological and evolutionary responses to recent climate change. *Annu. Rev. Ecol. Syst.* **37**, 637–669 (2006).
4. Wolkovich, E. M. et al. Warming experiments underpredict plant phenological responses to climate change. *Nature* **485**, 494–497 (2012).
5. Silvertown, J. A new dawn for citizen science. *Trends Ecol. Evol.* **24**, 467–471 (2009).
6. Pyke, G. H. & Ehrlich, P. R. Biological collections and ecological/environmental research: a review, some observations and a look to the future. *Biol. Rev.* **85**, 247–266 (2010).
7. Robbirt, K. M., Davy, A. J., Hutchings, M. J. & Roberts, D. L. Validation of biological collections as a source of phenological data for use in climate change studies: a case study with the orchid ophrys sphegodes. *J. Ecol.* **99**, 235–241 (2011).
8. Baird, R. C. Leveraging the fullest potential of scientific collections through digitisation. *Biodivers. Informatics* **7**, 130–136 (2010).
9. Beaman, R. S. & Cellinese, N. Mass digitization of scientific collections: new opportunities to transform the use of biological specimens and underwrite biodiversity science. *ZooKeys* **209**, 7–17 (2012).
10. Balke, M. et al. Biodiversity into your hands: a call for a virtual global natural history 'metacollection'. *Front. Zool.* **10**, 55 (2013).
11. Willis, C. G. et al. Crowdcurio: an online crowdsourcing platform to facilitate climate change studies using herbarium specimens. *New Phytol.* **215**, 479–488 (2017).
12. Meyer, C., Weigelt, P. & Kreft, H. Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecol. Lett.* **19**, 992–1006 (2016).
13. Roberts, D. L. & Solow, A. R. Flightless birds: when did the dodo become extinct? *Nature* **426**, 245–245 (2003).
14. Ruggles, R. & Brodie, H. An empirical approach to economic intelligence in World War II. *J. Am. Stat. Assoc.* **42**, 72–91 (1947).
15. Cooke, P. Optimal linear estimation of bounds of random variables. *Biometrika* **67**, 257–258 (1980).
16. Weissman, I. Confidence intervals for the threshold parameter. *Commun. Stat. Theory Methods* **10**, 549–557 (1981).
17. Inouye, D. W. Effects of climate change on phenology, frost damage, and floral abundance of montane wildflowers. *Ecology* **89**, 353–362 (2008).
18. Aldridge, G., Inouye, D. W., Forrest, J. R., Barr, W. A. & Miller-Rushing, A. J. Emergence of a mid-season period of low floral resources in a montane meadow ecosystem associated with climate change. *J. Ecol.* **99**, 905–913 (2011).
19. CaraDonna, P. J., Iler, A. M. & Inouye, D. W. Shifts in flowering phenology reshape a subalpine plant community. *Proc. Natl Acad. Sci. USA* **111**, 4916–4921 (2014).
20. Visser, M. E. & Both, C. Shifts in phenology due to global climate change: the need for a yardstick. *Proc. R. Soc. B* **272**, 2561–2569 (2005).
21. Miller-Rushing, A. J., Inouye, D. W. & Primack, R. B. How well do first flowering dates measure plant responses to climate change? The effects of population size and sampling frequency. *J. Ecol.* **96**, 1289–1296 (2008).
22. Davis, C. C., Willis, C. G., Connolly, B., Kelly, C. & Ellison, A. M. Herbarium records are reliable sources of phenological change driven by climate and provide novel insights into species phenological cueing mechanisms. *Am. J. Bot.* **102**, 1599–1609 (2015).

23. Papworth, S., Rist, J., Coad, L. & Milner-Gulland, E. Evidence for shifting baseline syndrome in conservation. *Conserv. Lett.* **2**, 93–100 (2009).
24. Cook, B. I. et al. Sensitivity of spring phenology to warming across temporal and spatial climate gradients in two independent databases. *Ecosystems* **15**, 1283–1294 (2012).
26. Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. *Bayesian Data Analysis* 3rd edn (CRC Press, Boca Raton, 2014).
27. Scheffer, M., Carpenter, S., Foley, J. A., Folke, C. & Walker, B. Catastrophic shifts in ecosystems. *Nature* **413**, 591–596 (2001).
28. Scheffer, M. et al. Early-warning signals for critical transitions. *Nature* **461**, 53–59 (2009).
29. Körner, C. & Basler, D. Phenology under global warming. *Science* **327**, 1461–1462 (2010).
30. Cook, B. I., Wolkovich, E. M. & Parmesan, C. Divergent responses to spring and winter warming drive community level flowering trends. *Proc. Natl Acad. Sci. USA* **109**, 9000–9005 (2012).
31. Brooks, S. J., Self, A., Toloni, F. & Sparks, T. Natural history museum collections provide information on phenological change in british butterflies since the late-nineteenth century. *Int. J. Biometeorol.* **58**, 1749–1758 (2014).
32. R Core Team *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, 2016); <https://www.R-project.org/>
33. Clements, C. F. et al. Experimentally testing the accuracy of an extinction estimator: Solow's optimal linear estimation model. *J. Anim. Ecol.* **82**, 345–354 (2013).
34. Therneau, T. *deming: Deming, Thiel-Sen and Passing-Bablok Regression*. R package version 1.0-1 (R Foundation for Statistical Computing, Vienna, 2014); <https://CRAN.R-project.org/package=deming>.
35. Miller-Rushing, A. J. & Primack, R. B. Global warming and flowering times in Thoreau's Concord: a community perspective. *Ecology* **89**, 332–341 (2008).
36. Willis, C. G., Ruhfel, B., Primack, R. B., Miller-Rushing, A. J. & Davis, C. C. Phylogenetic patterns of species loss in Thoreau's woods are driven by climate change. *Proc. Natl Acad. Sci. USA* **105**, 17029–17033 (2008).
37. Harris, I., Jones, P., Osborn, T. & Lister, D. Updated high-resolution grids of monthly climatic observations—the CRU TS3. 10 dataset. *Int. J. Climatol.* **34**, 623–642 (2014).
38. Carpenter, B. et al. Stan: a probabilistic programming language. *J. Stat. Softw.* **76**, 1–32 (2017).
39. Ellwood, E. R., Temple, S. A., Primack, R. B., Bradley, N. L. & Davis, C. C. Record-breaking early flowering in the eastern United States. *PLoS ONE* **8**, e53788 (2013).
40. Gelman, A. & Pardoe, I. Bayesian measures of explained variance and pooling in multilevel (hierarchical) models. *Technometrics* **48**, 241–251 (2006).

Acknowledgements

We thank the volunteers of the NPN for data collection. W.D.P. and T.J.D. were funded by Fonds de Recherche Nature et Technologies grant number 168004. The Colorado data collection was supported by National Science Foundation grants DEB 75-15422, DEB 78-07784, BSR 81-08387, DEB 94-08382, IBN 98-14509, DEB 02-38331 and DEB 09-22080 to D.W.I. We are grateful to D. Roberts and A. Solow for helping to calculate the s.e. of estimates of timing, and E. L. Wolkovich and B. G. Waring for comments on the paper. We also thank the photographers whose images we used in Fig. 3.

Author contributions

W.D.P., T.J.D. and C.C.D. conceived of the study. W.D.P. analysed the data. All authors wrote the manuscript and interpreted the results.

Competing interests

The authors declare no competing financial interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41559-017-0350-0>.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to W.D.P. or T.J.D.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see [Reporting Life Sciences Research](#). For further information on Nature Research policies, including our [data availability policy](#), see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

▶ Experimental design

1. Sample size

Describe how sample size was determined.

This is an observational study, and all data on plant flowering times as collected by the National Phenology Network were used.

2. Data exclusions

Describe any data exclusions.

No data were excluded.

3. Replication

Describe whether the experimental findings were reliably reproduced.

No experiments were conducted; this was an observational study.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

No experiments were conducted; this was an observational study.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

No experiments were conducted; this was an observational study.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

- | n/a | Confirmed |
|-------------------------------------|--|
| <input type="checkbox"/> | <input checked="" type="checkbox"/> The <u>exact sample size</u> (n) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.) |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A statement indicating how many times each experiment was replicated |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section) |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A description of any assumptions or corrections, such as an adjustment for multiple comparisons |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> The test results (e.g. P values) given as exact values whenever possible and with confidence intervals noted |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A clear description of statistics including <u>central tendency</u> (e.g. median, mean) and <u>variation</u> (e.g. standard deviation, interquartile range) |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> Clearly defined error bars |

See the web collection on [statistics for biologists](#) for further resources and guidance.

► Software

Policy information about [availability of computer code](#)

7. Software

Describe the software used to analyze the data in this study.

R; functions within R that were written for this study and are released in the supplementary materials.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* [guidance for providing algorithms and software for publication](#) provides further information on this topic.

► Materials and reagents

Policy information about [availability of materials](#)

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

N/A

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

N/A

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

N/A

b. Describe the method of cell line authentication used.

N/A

c. Report whether the cell lines were tested for mycoplasma contamination.

N/A

d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by [ICLAC](#), provide a scientific rationale for their use.

N/A

► Animals and human research participants

Policy information about [studies involving animals](#); when reporting animal research, follow the [ARRIVE guidelines](#)

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.

N/A

Policy information about [studies involving human research participants](#)

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

N/A