Time-resolved microfluidics unravels individual cellular fates during double-strand break repair

Nadia Vertti-Quintero, Ethan Levien, Lucie Poggi, Ariel Amir, Guy-Franck Richard, Charles N Baroud

BMC Biology

RESEARCH ARTICLE

Open Access

Time-resolved microfluidics unravels individual cellular fates during double-strand break repair

Nadia Vertti-Quintero[1], Ethan Levien[2], Lucie Poggi[3], Ariel Amir[4], Guy-Franck Richard[3*] and Charles N. Baroud[1,5*]

Abstract

Background: Double-strand break repair (DSBR) is a highly regulated process involving dozens of proteins acting in a defined order to repair a DNA lesion that is fatal for any living cell. Model organisms such as *Saccharomyces cerevisiae* have been used to study the mechanisms underlying DSBR, including factors influencing its efficiency such as the presence of distinct combinations of microsatellites and endonucleases, mainly by bulk analysis of millions of cells undergoing repair of a broken chromosome. Here, we use a microfluidic device to demonstrate in yeast that DSBR may be studied at a single-cell level in a time-resolved manner, on a large number of independent lineages undergoing repair.

Results: We used engineered *S. cerevisiae* cells in which GFP is expressed following the successful repair of a DSB induced by Cas9 or Cpf1 endonucleases, and different genetic backgrounds were screened to detect key events leading to the DSBR efficiency. Per condition, the progenies of 80–150 individual cells were analyzed over 24 h. The observed DSBR dynamics, which revealed heterogeneity of individual cell fates and their contributions to global repair efficacy, was confronted with a coupled differential equation model to obtain repair process rates. Good agreement was found between the mathematical model and experimental results at different scales, and quantitative comparisons of the different experimental conditions with image analysis of cell shape enabled the identification of three types of DSB repair events previously not recognized: high-efficacy error-free, low-efficacy error-free, and low-efficacy error-prone repair.

Conclusions: Our analysis paves the way to a significant advance in understanding the complex molecular mechanism of DSB repair, with potential implications beyond yeast cell biology. This multiscale and multidisciplinary approach more generally allows unique insights into the relation between in vivo microscopic processes within each cell and their impact on the population dynamics, which were inaccessible by previous approaches using molecular genetics tools alone.

Keywords: Double-strand break repair, Microfluidics, Single-cell, Dynamics

Background

Microsatellites are simple sequence repeats, very common in eukaryotic genomes. They represent 3% of the human genome sequence [1]. Their high mutation rate leads to frequent polymorphisms in the human population [2]. Recurrently, they expand or contract following replication, DNA repair, or homologous recombination

*Correspondence: guy-franck.richard@pasteur.fr; charles.baroud@pasteur.fr

[3] Natural and Synthetic Genome Instabilities Group, Institut Pasteur, CNRS UMR3525, 75015 Paris, France
[5] LadHyX, CNRS, Ecole Poytechnique, Institut Polytechnique de Paris, 91120 Palaiseau, France
Full list of author information is available at the end of the article

(reviewed in [3]). In some unfortunate cases, very large trinucleotide repeat expansions lead to human neurodegenerative disorders such as Huntington disease, myotonic dystrophy type 1, or Friedreich ataxia (reviewed in [4]). The precise molecular mechanism that causes these large expansions is not totally understood but it has been proposed that the propensity of these repeats to form stable secondary structures could trigger such expansion [5, 6].

Shortening expanded repeats to non-pathological lengths—or their complete removal—using highly specific DNA endonucleases has been envisioned as a therapeutic approach [7, 8]. In this context, it is essential to understand the mechanisms and limitations of processing and repairing a double-strand break (DSB) within a repeated and structured DNA sequence.

Given the complexity of genetically manipulating human cells, the budding yeast *Saccharomyces cerevisiae* has been widely adopted as a model suitable for the understanding of cellular processes and protein function in higher eukaryotes. Particularly, budding yeast has been used for decades to study homologous recombination and the fate of a single double-strand break made in its genome using highly specific DNA endonucleases such as HO or I-*Sce* I [9, 10]. More recently, the CRISPR-Cas9 system has stood out because of its favorable properties: it is fast, cheap, accurate, and efficacious in making a DSB at any DNA locus. In such assays, target sequence recognition is based on a complementary guide RNA (gRNA) and on a short sequence called protospacer adjacent motif (PAM), where DSB is induced by an endonuclease associated to this gRNA (reviewed in [11]).

In order to assess double-strand break repair (DSBR) efficacy on repeated and structured DNA, an experimental system was previously designed in *S. cerevisiae*, relying on a bipartite green fluorescent protein gene (*GFP*) interrupted by different microsatellites [12]. Upon targeted DSB induction, both *GFP* moities can recombine with each other to reconstitute a functional *GFP* gene (and thus make correct DSBR), subsequently detectable by in vivo fluorescence of yeast cells. Analysis of whole populations of yeast cells showed that DSBR efficacy was highly variable among microsatellites and endonucleases used to induce the DSB [12]. In this context, essential aspects of a successful DSBR are yet to be fully understood, including the rates of the critical steps in the process, as well as cell-to-cell heterogeneity, which cannot be studied in traditional bulk experiments. Indeed, single-cell assays are required to study individual behaviors of yeast cells within a population, namely to understand whether a small proportion of cells are very efficacious at repairing the break and then propagate within the culture or if all cells are equally competent at repairing. Then, linking the single-cell scale with the dynamics at the scale of the population requires mathematical modeling to bridge them [13].

Previous work has addressed similar questions in yeast cells, using microfluidic devices and mathematical models. An elegant experimental system was setup in which young cells could be separated from older ones in a microfluidic chip [14], and this system was used to study DNA repair following a double-strand break induced by the I-*Sce*I endonuclease. The authors showed that old yeast cells were less efficient to repair the DSB than young ones, indicating an age-associated decline in repair [15]. From a mathematical point of view, the dynamics of a yeast population over time was described using an ordinary differential equation (ODE) model [16] or a stochastic model [17]. In the former case, the model suggests that early repair of DNA damage during the cell life helps to counteract aging caused by damage retention, therefore increasing life span. In the latter case, a stochastic model was used to determine how damage accumulation as well as repair efficacy drastically influence senescence and population fitness. In addition, a stochastic model of genetic activity was presented by Song et al. [18], where changes in cell size, DNA replication, and cell division were taken into account for refining dynamic rate reactions. All these efforts have built up a compendium of mathematical tools for better understanding phenomena in eukaryotic cells at different scales.

In this work, we link the dynamics at the single-cell level with the population-scale efficacy of the gene-editing assay for DSBR in eukaryotic cells. In contrast with the existing literature on single-cell gene network activity, here, we present a simpler approach for screening different combinations of microsatellites and endonucleases for investigating their impact on DSBR efficacy, rather than for describing the single-cell dynamics in a cell lifespan context. To this end, we use the *S. cerevisae* assay previously described, in which a bipartite GFP gene may recombine to form a functional gene, upon successful DSBR [12]. A microfluidic platform [19], in which cells are trapped in an array of cubic compartments of 100 μm edges, enables the identification of successful DSBR in single cells and their follow-up over time. As a result, we obtain time-resolved quantitative observations of biological phenomena happening on small populations stemming from single yeast cells. Molecular measurements of the percentage of cells undergoing DSB after endonuclease induction allow us to formulate an ODE model, capturing the characteristic steps and time scales involved in such process, inferring the growth, breaking, and repair rate of cells. We find that population dynamics from the microfluidic experiments were generally in good agreement with previously published results obtained with

whole cell populations [12] and with the prediction from our ODE model. In addition, the single-cell analysis elucidates the trajectories of individual cells undergoing DSBR and their impact on the global population DSBR efficiency, ultimately leading to the identification of three categories of DSBR: high-efficacy error-free, low-efficacy error-free, and low-efficacy error-prone repair.

Results

Observing cells undergoing DSBR in microfluidic wells

The present work builds on a cellular assay for studying DSBR in yeast cells. The assay relies on a bipartite overlapping *GFP* gene, inserted in a yeast chromosome whose two halves are separated by an intervening 100-bp sequence that contains (CGG)33, (GAA)33, (CTG)33 trinucleotide repeats or a non-repeated sequence (NR) [12]. The different conditions will be hereafter referred to as CGG, GAA, CTG, or NR strains, respectively. A DSB is made within this intervening sequence by either *Streptococcus pyogenes* Cas9 or *Francisella novicida* Cpf1 endonucleases [22, 23] (Fig. 1a). Cas9 is a class 2 type II endonuclease, whereas Cpf1 is a class II type V enzyme [24]. They use different PAMs and different gRNAs and exhibit very different structures and biochemical properties. The endonucleases and gRNAs are carried by different plasmids in modified yeast cells, with the

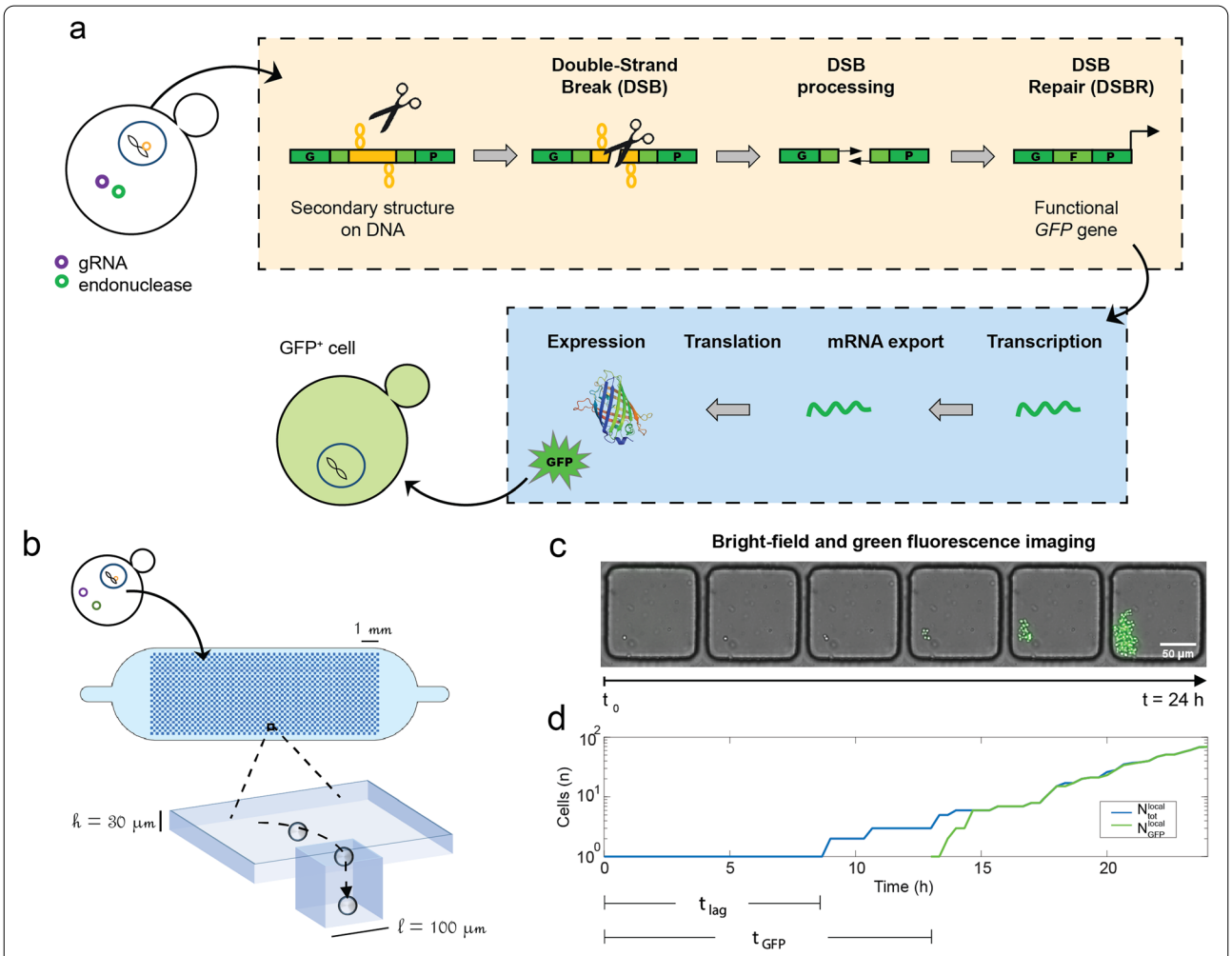

Fig. 1 GFP recombination assay and observation in microfluidic devices. **a** An inactive bipartite GFP gene contains different microsatellites. Upon endonuclease induction, a DSB is made within the repeat, processed and repaired to reassemble a functional *GFP* gene (yellow box). Subsequent downstream processes (blue box) happen until GFP is expressed and cells turn green. **b** Sketch of the microfluidic device containing 1032 cubic traps (100-µm edge). Yeast cells in suspension (concentration: 5 cells/nL) flow into the microfluidic device and sediment into the wells. **c** Cells trapped in wells are monitored over 24 h both in bright-field and epifluorescence. The number of cells in each well and their level of GFP fluorescence are monitored using the plugin *TrackMate* [20] on *ImageJ* [21]. **d** Time series of the number of cells and number of GFP+ cells are obtained for one well

endonuclease being under the control of a galactose-regulatable promoter [25]. Endonuclease expression is induced by switching cells from glucose to a galactose-containing medium. This change produces a metabolic switch, slowing down cell division while switching metabolism to galactose utilization [26].

Once the DSB is induced, a series of events takes place, as shown in Fig. 1a, yellow box. DSB resection—following the break—generates two single-stranded DNA ends whose overlapping halves may anneal with each other, thus reconstituting a functional *GFP* gene. Once the *GFP* gene is reassembled (i.e., completed DSBR), downstream processes are carried out (as shown in Fig. 1a, blue box), including transcription, mRNA export, and translation, until GFP is expressed and the cell becomes green. Due to checkpoint activation following DSB [27], the cell cycle is transiently halted, so that cells cannot divide with a broken chromosome. This assay is functional and has already shown different efficacies of endonucleases on trinucleotide repeats depending on the stability of secondary structures formed by the gRNA [12].

In vivo observations of single yeast cells undergoing DSBR were envisioned to understand different cell aspects: how likely individual cells were to break and repair, and how these steps integrated within the broader cell cycle. This experiment was enabled by the use of a microfluidic device (Fig. 1b): by confining single cells within microfluidic wells, it was possible to observe individual cell divisions and DSBR completion using time-lapse microscopy (Fig. 1c, d). Moreover, by tracking the progeny of each cell, it was possible to link the emergence of these population dynamics with the scale of individual cellular events.

Microfluidic devices have been proposed before for studying individual yeast cells [28]. For example, Jo et al. [29] developed one for analyzing the replicative lifespan of single cells, while Charlebois and collaborators [30] used individual cell traps for observing the expression of a reporter gene on cells upon changes of temperature. In this study, a microfluidic device with similar geometry to the one presented by Amselem et al. [19] was adapted to observe the yeast cells undergoing DSBR in real time. It consisted of a long and wide chamber (6×14 mm) of height $h = 30$ μm, with one inlet and one outlet. The chamber floor was patterned with a two-dimensional array of 1032 cubic wells of $l = 100$ μm edge length. Space between the wells was set to $d = 120$ μm (Fig. 1b).

A typical experiment started by suspending yeast cells at a concentration of 5 cells per nanoliter in a galactose-containing medium (at time t_0), in order to express the endonuclease. This cell suspension was then rapidly introduced into the microfluidic chip, where the individual cells sedimented into the wells. The well occupancy did not have a homogeneous distribution; wells typically contained from 0 to 5 cells. Only populations that started with a single cell were selected for our analysis in order to monitor the lineage of individual cells. The growth of populations and their GFP expression over time was tracked by time-lapse microscopy (Fig. 1c). For each well, the total number of cells ($N_{\mathrm{tot}}^{\mathrm{local}}$) and GFP+ cells ($N_{\mathrm{GFP}}^{\mathrm{local}}$) were counted at each time point, yielding a single growth curve per well, as described in the "Methods" section (Fig. 1d). Measurements were collected on samples ranging from 80 to 150 wells in each microfluidic experiment. We define t_{lag} the moment at which cells start dividing after t_0 and t_{GFP} when they start expressing GFP (Fig. 1d).

The time evolution of DSBR dynamics was studied for 8 different combinations, i.e., two endonucleases (Cas9 and Cpf1) and four target sequences (NR, CGG, GAA, and CTG), using the above analysis pipeline (Fig. 2a and Additional Movie). Generally, individual cells started dividing (t_{lag}) and expressing GFP (t_{GFP}) a few hours after galactose induction (t_0). The data for all the conditions are shown in Fig. 2b, where a variety of dynamics is observed for the different target-endonuclease combinations. Here, the access to the absolute number of cells allowed us to point out some important differences between the different conditions, as observed by the bold curves for the mean behavior in Fig. 2b. To be noted that even in identical experimental conditions, cells started dividing at different t_{lag}, started expressing GFP at different t_{GFP}, and formed populations of different sizes at $t = 24$ h, as can be seen in any subplot of Fig. 2b.

Strikingly, two cases (GAA-Cas9, CTG-Cas9) showed a strong slowing down of the exponential growth, while the cell numbers in most other cases grew exponentially. This slowing down might indicate a loss of fitness that is associated with the DSBR. Another observation concerned the delay between the growth of the population size and the detection of GFP+ cells. This time difference ($t_{\mathrm{GFP}} - t_{\mathrm{lag}}$) was in the range of 4–6 h for most conditions except for the condition CTG-Cpf1, where it was above 15 h, indicating different dynamics between the cell cycle and the DSBR process for these conditions. In the case of CGG-Cpf1, only one GFP+ cell was detected during the course of the experiment.

Compared with the diversity of DSBR efficacy that has been described previously [12], the current measurements highlight the variability of timing in the break and repair processes. This dynamic viewpoint motivated the development of a time-dependent ordinary differential equation (ODE) model, as described next.

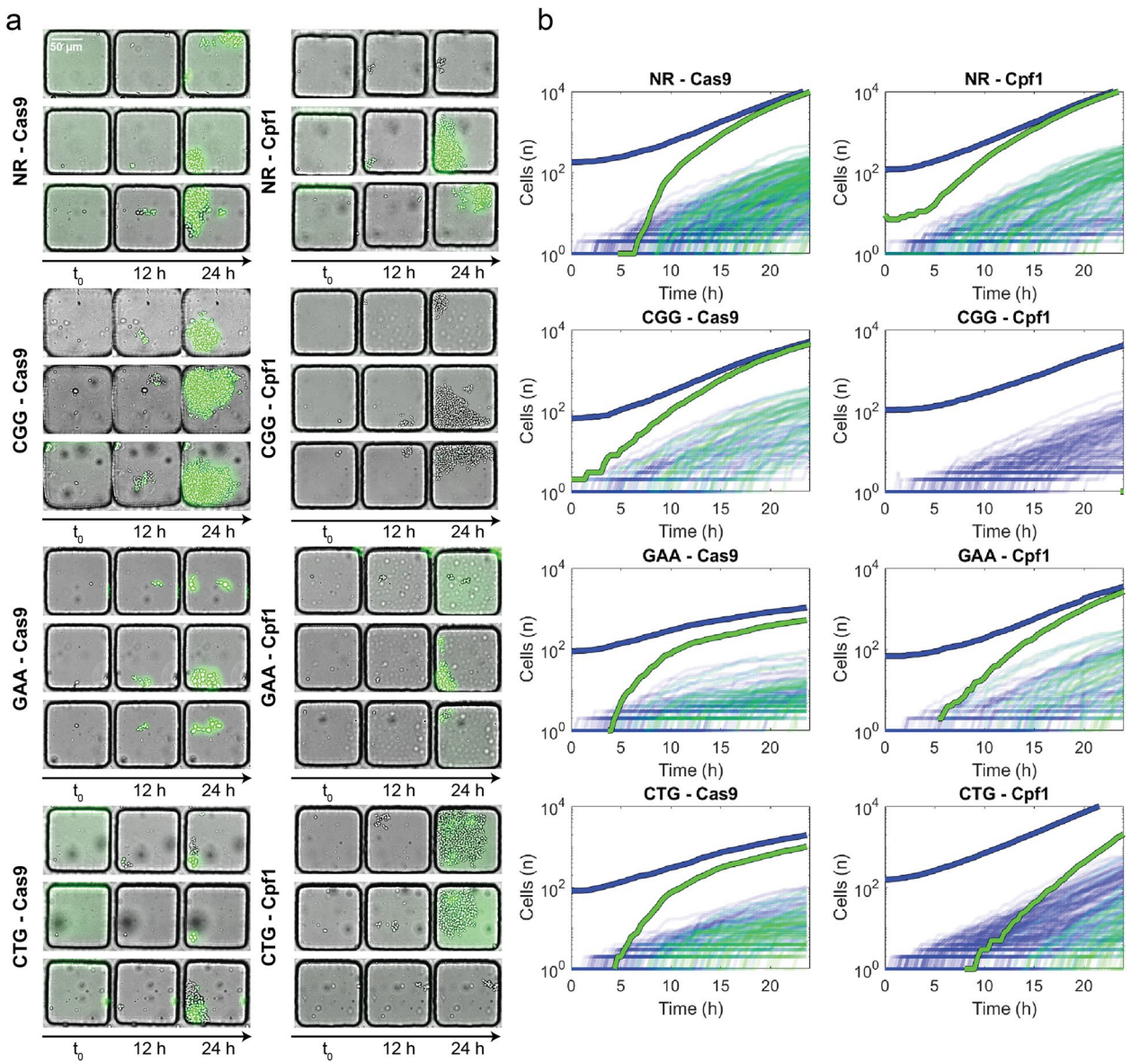

Fig. 2 Microfluidic device yields growth and repair dynamics of populations starting from single cells. **a** Representative images of cells trapped in microfluidic wells at $t = 0$, 12, and 24 h after galactose induction, for the 8 experimental conditions: NR, CGG, GAA, or CTG target sequences with either Cas9 or Cpf1. The scale bar represents 50 µm. **b** Corresponding growth curves for individual wells in each condition. Blue lines represent the number of cells per time point and green lines the number of GFP+ cells. Bold lines correspond to the total number of cells and GFP+ cells at each time point, for each experimental condition

ODE model built on molecular measurements provides rates of break and repair

Successful DSB induction and repair are the result of a series of molecular steps. In order to identify the relevant time scales in the process, we utilized a model which assumes that, upon induction, an initial population of "modified" cells (containing a specific microsatellite or a non-repeated sequence) has a constant per capita division rate α, while cells switch into a non-growing, broken state, at a rate β. The "broken" cells can then become repaired at a rate ρ and once again begin to grow at a rate α (Fig. 3a). All rates in the model can be understood as per unit time probabilities, e.g., βdt is the chance for a modified cell to under the broken state in a time interval dt. The fact that the division rate for modified and repaired cells is the same is consistent with population observations that the repeat does not hinder yeast cell replication [12]. This

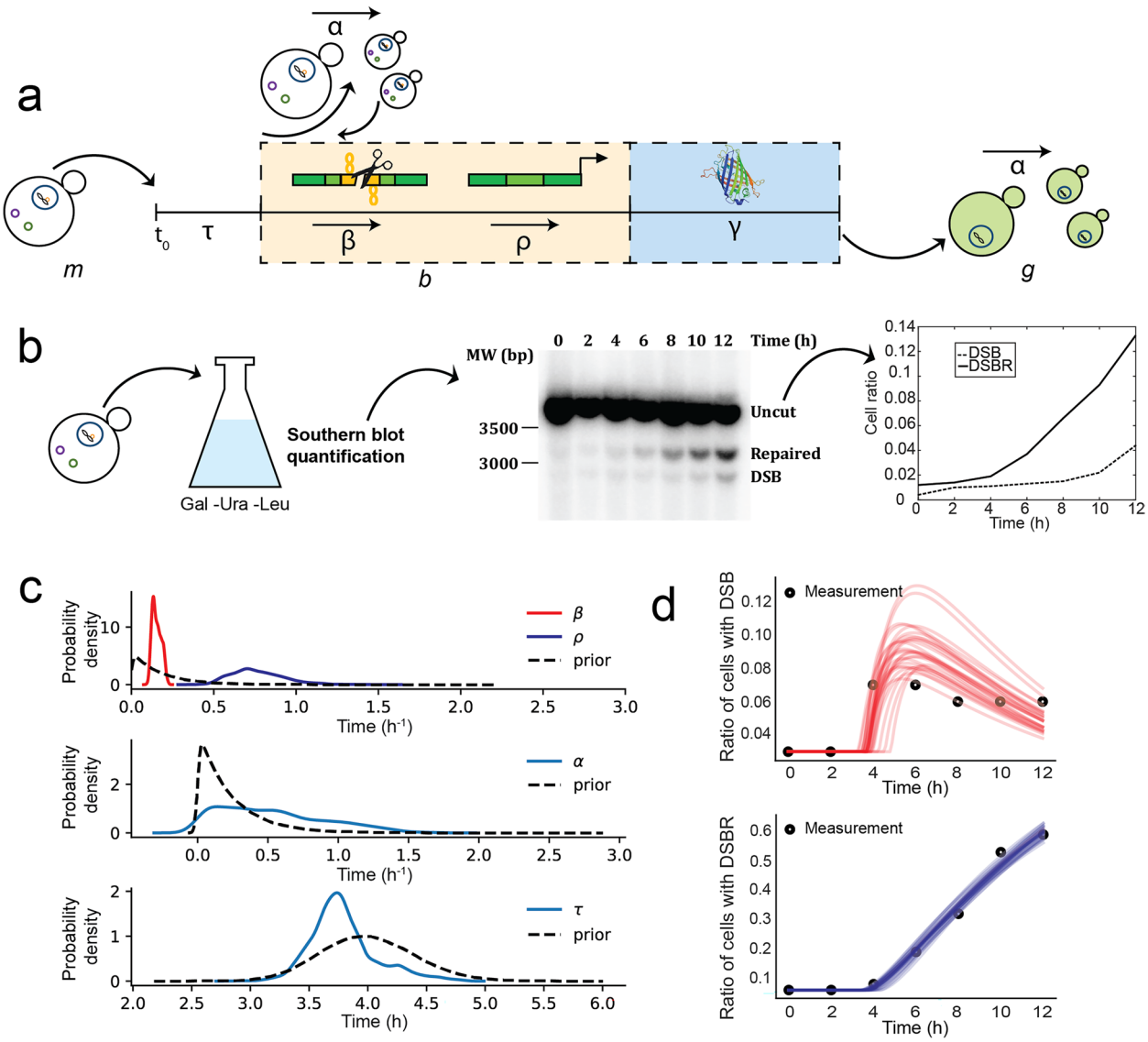

Fig. 3 ODE model captures the characteristic steps in DSBR. **a** Sketch of the parameters of our ODE model that describe the whole DNA repair process, as already illustrated in Fig. 1a: "modified" cells (m) are submitted to metabolism change upon DNA repair induction (for a time period τ) and DSB (at the rate β) and thus becoming "broken" cells (b); they do DSBR (at the rate ρ) and express GFP (after a time period γ), thus becoming GFP+ cells (g); and they divide (at the rate α). **b** Molecular measurements are performed by Southern blotting at regular time intervals (2 h) on samples from a bulk cell culture. They provide information on the fraction of cells in a population that have done DSB, but not yet repaired, and the fraction of cells that have completed repair. **c** Prior and posterior parameter distributions of metabolism change (τ), DSB rate (β), DSBR rate (ρ), and cell division rate (α). **d** Comparison of ODE model predictions with molecular data for the case NR-Cas9. In plots, the black dots represent the molecular data, while the solid lines represent the simulations from the ODE model

model can be written in terms of a Master equation, as described in detail in the Section 5.

We may also write these processes in the form of three coupled equations to describe these dynamics after a lag time τ. Before the lag time, we assume the cells undergo no growth, and therefore no DSBR. Letting m, b, and g be the number of "modified," "broken" (with broken DNA after DSB), and "repaired" (GFP+) cells, we have the system of linear ODEs for the averages:

$$\frac{d}{dt}\langle m\rangle = (\alpha - \beta)\langle m\rangle, \tag{1}$$

$$\frac{d}{dt}\langle b\rangle = \beta\langle m\rangle - \rho\langle b\rangle, \tag{2}$$

$$\frac{d}{dt}\langle g\rangle = \alpha\langle g\rangle + \rho\langle b\rangle. \tag{3}$$

Importantly, due to the linearity of the model, the average $\langle\cdot\rangle$ can be understood either as an ensemble average over many experiments each consisting of a small number of cells or as the large population size limit of a single experiment.

Upon DSBR induction, since all events happen at different moments for different cells in the culture, a sample of the cell population should contain a mixture of the different states: intact cells, cells displaying a broken chromosome, and cells harboring a repaired chromosome. The dynamics of each of these sub-populations can be quantified by molecular analysis on cells sampled at different times in a growing culture, as shown in Fig. 3b. To that end, cells were collected every 2 h after galactose induction and whole genomic DNA was extracted (see Section 5). Hybridization with a probe specific for the *GFP* locus revealed three different types of signals on a Southern blot: a 3544-bp band corresponding to uncut DNA, a 2912-bp band representing the DSB, and a 3162-bp band representing the repaired and functional *GFP* gene (Fig. 3b). Values of the relative abundance of broken and repaired chromosomes are shown in Additional file 1: Fig. S1 and were taken from Poggi et al. [12], except for NR-Cpf1 which was redone here. The fraction of cells that are in the broken state remains low over the 12 h that the measurement is done, since it is a transient state. In contrast, the fraction of cells that have completed DSBR increases over time for almost all conditions, with the notable exception of CGG-Cpf1 and CTG-Cpf1. Note that the NR-Cpf1 case starts already with a comparatively large number of cells that have completed DSBR (40% in comparison to less than 20% for other cases). This is probably due to the leakiness of the Gal promoter that has a more pronounced effect in this strain background [25].

Using the Southern blot measurements (Fig. 3b), we performed Bayesian inference (see the review article [31]) of the parameters α, β, and ρ, which yielded a posterior distribution $P(\boldsymbol{\theta}|\mathbf{X})$. The posterior distribution is defined as the distribution of the model parameters $\boldsymbol{\theta}$ conditioned on the observed data $\mathbf{X}$:

$$P(\boldsymbol{\theta}|\mathbf{X}) \propto P(\mathbf{X}|\boldsymbol{\theta})P(\boldsymbol{\theta}). \tag{4}$$

Here, the likelihood function $P(\mathbf{X}|\boldsymbol{\theta})$ gives the distribution of the data given our parameters, where the data consist of the observed population fractions, $\mathbf{X} = (m/N, b/N, g/N)$ where N is the total number of cells. For each measurement, we assume that the observed fraction is true fraction plus some Gaussian error. The predicted fraction is obtained by solving Eqs. (1), (2), and (3). We further assume that the measurement errors are uncorrelated between different cell states and times. This assumption is, strictly speaking, false, since even if the measurements of m and b are uncorrelated, the measurement errors in their fractions would be correlated. However, numerical experiments with simulated data revealed that the results were robust to this assumption (see Section 5.4 and Figs. S2 and S3).

The distribution $P(\boldsymbol{\theta})$ represents our priors on both the parameters of the ODE model, as well as the measurement error and lag time (τ). With the exception of τ, we place so-called *weakly informative priors* on all parameters, that is, priors that only constrain the parameters to a physically reasonable range, rather than incorporating specific information from previous experiments. The same priors are used for β and ρ, as not to favor either breaking or repair as the limiting process. In the case of τ, the prior is chosen to have a narrow distribution around the known value of the lag. The priors are described in detail in the Section 5.4.

The posterior distributions for the NR-Cas9 condition are shown in Fig. 3b. Comparing the posterior distribution to the prior indicates how much new information about the parameters is obtained from the data. In the case of β and ρ, it can be seen that the data strongly constrain parameter values for many experiments, as evidenced by the fact that the posterior is much narrower than the prior. The value of α however is less-well determined. This may be expected since the measurements provide ratios of the number of cell types, and not absolute numbers. This selectivity on β and ρ is reproducible for all cases, as shown in Additional file 1: Figs. S4 and S5 for all experimental conditions.

Next, the model predictions for the population fractions in the broken or repaired states can be compared with the experimental measurements, as shown in Fig. 3c for the NR-Cas9 case. Good agreement is found for most cases (Fig. S6), where the ODE model gives good predictions of the trends and resolves the time scales of broken and repaired fractions. These observations show that the low time-resolution molecular data are sufficient to estimate the parameters of the proposed ODE model, thus predicting the dynamic behavior of DSB induction and repair.

The posterior values of the breaking rate β and the repair rate ρ are displayed in Fig. S5, for the eight different conditions. From these data, it emerges that the breaking step is rate-limiting for most cases, with the repair happening at a higher rate for all cases. Besides the two very inefficient conditions (CGG-Cpf1 and CTG-Cpf1), three conditions could be described as efficient (NR-Cas9, CGG-Cas9, and NR-Cpf1), with mean values of $\beta > 0.1$ 1/h and a final fraction of repaired cells above

0.6 (Additional file 1: Figs. S5a and S6). The three remaining conditions had intermediate breaking rates $\beta \simeq 0.09$ and a final fraction of repaired cells not exceeding 0.4. In contrast with the breaking rates, which did not show a strong difference between the two nucleases, the repair rates ρ were always faster to repair in the case of Cas9 with respect to Cpf1 (Additional file 1: Fig. S5b).

Global behavior

It is informative to begin by comparing the global behavior in the microfluidic device with the bulk measurements, before studying the lineages of individual cells. This is done by summing the time evolution in each of the individual wells and defining the global measures $N_{\text{tot}}^{\text{global}}$ and $N_{\text{GFP}}^{\text{global}}$, for the total number of cells and the total number of GFP+ cells in each microfluidic experiment. From these two numbers, a global fraction $R^{\text{global}} = N_{\text{GFP}}^{\text{global}}/N_{\text{tot}}^{\text{global}}$ can be computed. This global fraction can be compared with the predictions of the ODE model using the parameter values obtained from the Bayesian fit of the molecular data described above.

The dynamics of R^{global}, obtained by pooling the different cell positions on a single chip, can then be compared with the predictions of the ODE model. The comparison for all eight experimental conditions is shown in Fig. 4, where the black dots show the experimental measurements while the group of cyan lines show the predictions from the ODE model. In this figure, the measurements previously obtained by flow cytometry [12] are indicated with yellow circles, showing mostly a good agreement with the microfluidic and numerical results. Although the values of the parameters α, β, ρ, and τ are obtained from the fitting molecular data from a very different setting, the simulated time evolution of the emergence of GFP+ cells matches the microfluidics experiments in most cases. A table containing the root mean square error (RMSE) between the simulated and experimental data is shown in Additional file 1: Table S1, where a higher RMSE value indicates a larger difference between experimental and simulated data.

In these comparisons, two conditions stand out as matching poorly with the ODE model, as can be seen in the Additional file 1: Table S1. The first concerns the NR-Cpf1 case, which grows faster in the experiments compared with the simulations. This is likely due to a leaky induction of Cpf1, which results in DSB induction before the switch to galactose media at t_0. This mismatch between the beginning of the metabolic switch and the break and repair leads to a reduced delay between t_{lag} and t_{GFP} compared with other conditions, as observed by the early onset of the green curves in Fig. 2b. From a modeling point of view, this complexity would add an additional time scale that is not accounted for in the equations. The other case with a poor fit between the microfluidic experiments and the ODE model is CTG-Cas9. This case corresponds to a condition that has a reduced fitness at later times, as evidenced by the slowed growth of the population numbers. In this case (and also on GAA-Cas9 which poorly matches with the flow cytometry results), some individual cells show an abnormal growth in cell size and atypical shapes mostly correlated with being GFP+, as seen in Fig. 2a and in the Additional Movie. The relation between these morphological changes and their impact on the growth of the populations will be studied in detail below where we study the temporal evolution in individual wells.

DSBR dynamics at the single-lineage scale

The above description treats the microfluidic device as a single population. Further insight can be obtained by looking at the dynamics of the progeny of each one of the yeast cells, which shows individual transition events from the initial state (modified, GFP−) to the repaired state (GFP+). By the same token, studying the individual curves gives access to the heterogeneity that exists between different cells within a single experiment.

Typical measurements from three conditions are shown in Fig. 5. By looking at a few individual traces in the case of NR-Cas9 (A.a–e), two situations are dominant: In some wells, the initial cell divides without any of its daughters becoming GFP+ (Fig. 5A.a). The cell proliferation in locations where the repair does not take place tends to slow down after a few initial divisions, as shown by the slower increase of the blue dots. In other wells, the cells turn green some time after the initial division. The time delay between the initial division and the first detection of a GFP+ cell in each well (t_{GFP}-t_{lag}) is well-distributed around a mean at 5.6 h, as shown in Fig. 5A.b. This delay is consistent with the time required for the cells to translate the new gene and express sufficient GFP molecules to make it detectable. The number of cells turning green after the first detection of a GFP+ cell increases rapidly until it covers all cells within the particular well. This typical behavior is summarized by plotting a few representative curves of the local fraction $R^{\text{local}} = (N_{\text{GFP}}^{\text{local}}/N_{\text{tot}}^{\text{local}})$, as shown in Fig. 5A.c. Here, again some lineages remain with a value of $R^{\text{local}} = 0$ until the end of the experiment but when R^{local} becomes positive it rapidly rises to a value near 1.

Taken together, the measurements of Fig. 5A.a–c indicate that DSB and DSBR take place very early in the lineage tree, possibly in the mother cell or its very first daughters, which explains the low value of the delay and the rapid increase in the number of GFP+ cells. As a result of these dynamics, the distribution of values of R^{local} at the end of the experiment ($t = 24$ h) is strongly

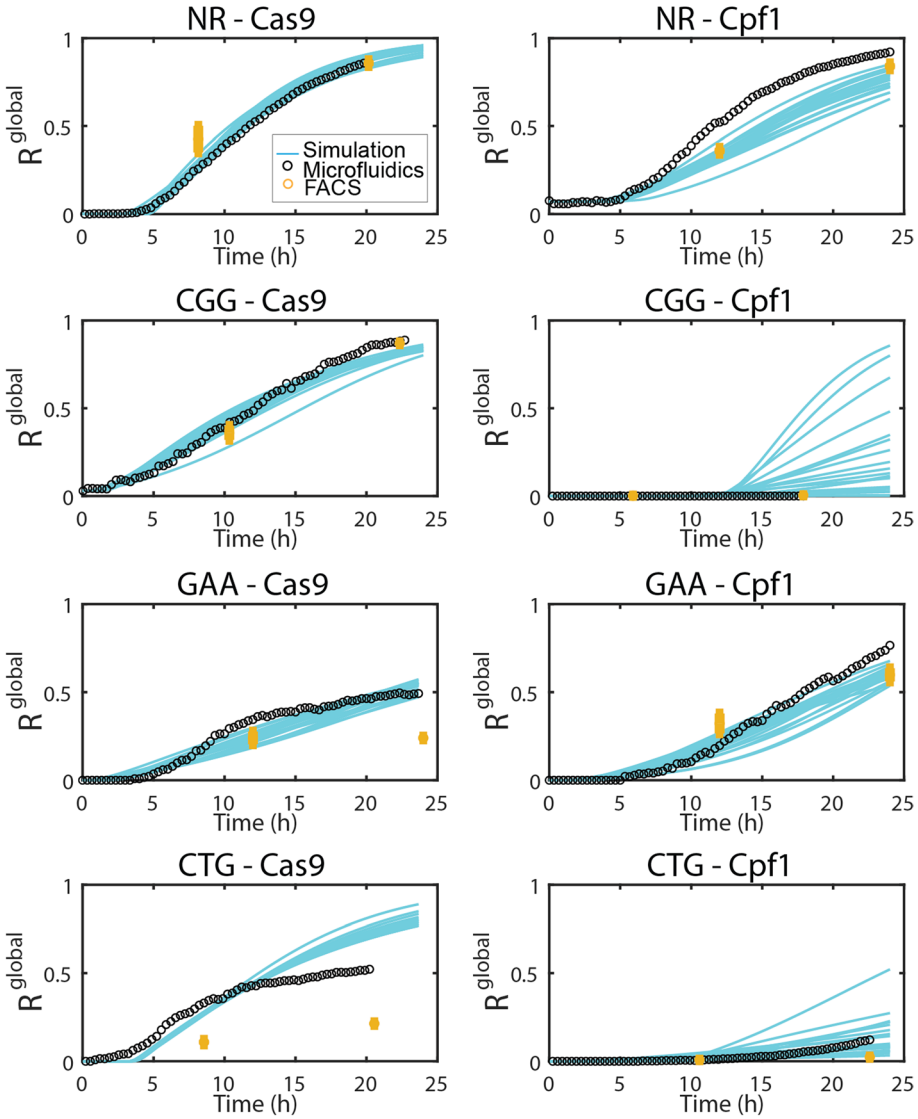

Fig. 4 Population experimental data vs ODE predictions. Comparing Bayesian model predictions (cyan lines) with experimental measurements for the fraction of GFP+ cells to the total number of cells. The black dots represent the measurement obtained with the microfluidic setup, while the yellow dots represent the bulk measurements with flow cytometry ([12])

bimodal. The statistics are dominated by the extreme values of $R^{\text{local}} = 0$ and $R^{\text{local}} \simeq 1$ (Fig. 5A.d). The intermediate values of R^{local} correspond to curves that are in the transition between zero and one at the end of the experiment.

The experimental measurements can be compared with values computed from the stochastic version of the ODE model (see Section 5.5), using the same parameter values obtained from the Bayesian fitting in Section 2.2. A sample of the simulated trajectories is shown in Additional file 1: Fig. S7, while the distribution of final values of R^{local} is shown in Fig. 5A.e. These simulations reproduce well the tendency of the case NR-Cas9 towards $R^{\text{local}} = 1$, as seen by the peak in the histogram. Nevertheless, the simulations fail to reproduce the peak at R^{local}.

The discrepancy between the model and experiments is due to the biological origin of the peak at $R^{\text{local}} = 0$, which corresponds either to cells totally escaping DSB or to broken cells unable to repair the DSB and therefore maintaining cell cycle arrest. This behavior does not

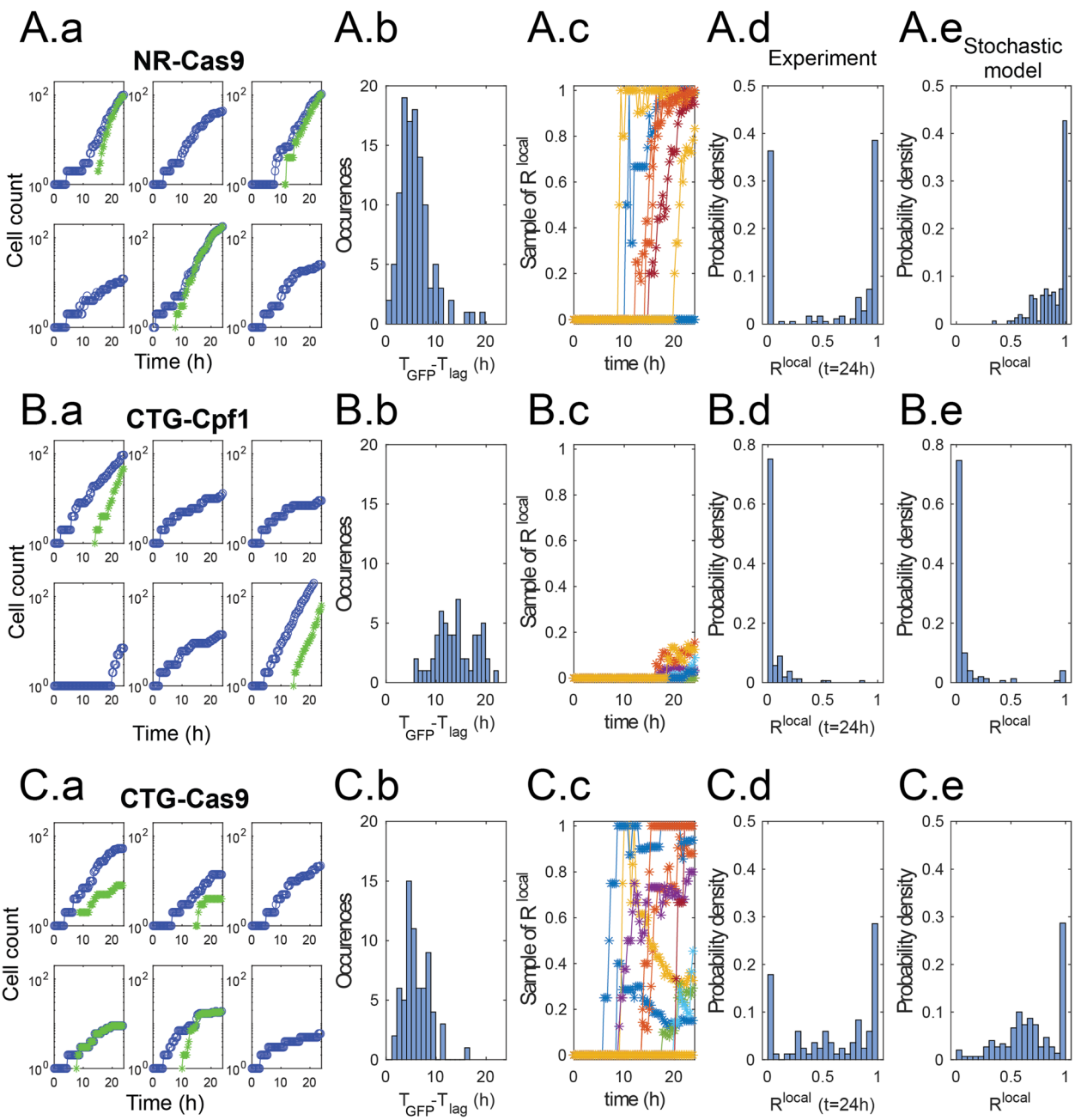

Fig. 5 Dynamics and statistics of individual lineages. **A** NR-Cas9. **a** Dynamics of $N_{\text{tot}}^{\text{local}}$ (blue) and $N_{\text{GFP}}^{\text{local}}$ (green) from six randomly selected wells. Note the diverse dynamics from different positions. **b** Distribution of delay times between first division and first detection of GFP+ cell. **c** Local fraction $R^{\text{local}} = N_{\text{GFP}}^{\text{local}}/N_{\text{tot}}^{\text{local}}$ for 9 randomly selected positions. The transitions from $R^{\text{local}} = 0$ happen at different instants and quickly rise towards $R^{\text{local}} \simeq 1$. Note that in some cases R^{local} remains zero. **d** Distribution of values of R^{local} at the end of the experiment. **e** Distribution of R^{local} at the end of the simulation obtained from the stochastic model using the same parameter values as above. **B** CTG-Cpf1. **C** CTG-Cas9. Same graphs as above

correspond to different values of the parameters (α, β, ρ) but rather to some dynamics that are not included in the theoretical model. Although the unbroken/unrepaired trajectories correspond to about 30% of the wells in the NR-Cas9 case, these positions contribute a small number to the total sum of cells in the experiment since these cells only go through a few division cycles. As a result, they are difficult to observe in the population-scale measurements, which explains the good agreement between the ODE model and global measurements in Fig. 4.

When the same analysis is made for CTG-Cpf1, very different dynamics and statistics are observed (in Fig. 5B.a–e). While the growth of individual lineages from single cells is generally similar to the previous case, the GFP+ cells appear less frequently and much later during the experiment (Fig. 5B.a, b). Indeed, the delay between the first division and the first GFP+ event, when it does occur, is broadly distributed between 5 and 20 h (Fig. 5B.c, Additional file 1: Fig. S7). Moreover, the traces of R^{local} do not rise sharply after the first GFP+ cells. Instead, in both experiments and in simulations, they show a much more gradual increase and only reach a small value at the end of the experiment (Fig. 5B.d, e). In this case, the computed growth curves and histogram of final values of R^{local} are in good agreement with the experimental measurements (Fig. 5B.e). These observations indicate that DSB and DSBR take place in cells long after the first division. As such, these events only affect a fraction of the progeny of the initial cell, which explains the slow rise of R^{local}, while most of the lineage tree maintains an unbroken microsatellite.

Finally, a third type of behavior is observed when considering the CTG-Cas9 condition, as shown in Fig. 5C.a–e. Here, the GFP+ cells appear early after the first division (mean time delay is 6 h) but the increase in the number of GFP+ cells is irregular (Fig. 5C.b, c, Additional file 1: Fig. S7). However, this condition corresponds to more complex biological processes, since GFP+ cells display reduced fitness and division arrest after becoming GFP+ (Fig. 5C.a and Additional Movie). If this arrest occurs after the complete population is repaired, it leads to a value of $R^{\text{local}} = 1$ but on a static population of cells. In other cases, only some of the cells are repaired and slow down their divisions, which leads to a value of R^{local} that initially increases before decreasing again (Fig. 5C.c, Additional file 1: Fig. S7). These dynamics yield a large variety of outcomes for the final value of R^{local}, which covers the whole range between zero and one (Fig. 5C.d, e).

In this last example, the comparison between experimental measurements and simulations from the stochastic model shows good agreement but care must be taken when comparing these two distributions. The peak at $R^{\text{local}} = 0$ is missing for the same reasons as in the NR-Cas9 case above. Moreover, cell cycle arrest of cells that become large is another particularity that is not included in the equations. As such, the model is missing two major specificities of the experiment. Contrary to the two examples discussed previously, the disagreement between the model and the experimental ingredients leads to a poor match in the global ratio (Fig. 4).

Relating the dynamics of individual lineages with the global population behavior

The information shown for three cases in Fig. 5 can be summarized for all conditions by plotting the time dynamics of cell populations as heat maps, as shown in Fig. 6. For each case, three quantities are represented by the color scheme: the number of cells over time ($N_{\text{tot}}^{\text{local}}$), the number of GFP+ cells over time ($N_{\text{GFP}}^{\text{local}}$), and the value of R^{local}. The heat maps are constructed as explained graphically in Additional file 1: Fig. S8: each row represents the time evolution from a single well, with the wells ranked according to the total number of cells at $t = 24$ h. Therefore, rows near the top of the graphs represent small final colonies, while rows near the bottom correspond to the largest colonies at the end of the experiment.

Analysis of these heat maps allows us to classify the behavior of DSB and DSBR according to three typical cases.

1. **High-efficacy error-free repair, normal cell growth.** The four conditions labeled with the star in Fig. 6a follow a high-efficacy situation. These conditions display first a strong correlation between the moment of the first division and the size at $t = 24$ h, as shown by the sideways slant of the pink border describing $N_{\text{tot}}^{\text{local}}$. The second observation is the relatively narrow delay between the first division and the first GFP+ cell, as seen by the small distance between the black line and the left edge of the pink region in the middle heat map. This small delay indicates that the first repair takes place when there are only a few cells in the well. Finally, $R^{\text{local}} \simeq 1$ for the bottom part of the heat maps, indicating that the largest individual populations are also the best repaired. This type of behavior is observed in four conditions: NR-Cas9, NR-Cpf1, CGG-Cas9, and GAA-Cpf1, and their progeny trees would resemble the ones illustrated in Fig. 6b, first panel.
2. **Low-efficacy error-free repair, normal cell growth.** The two conditions labeled with a circle in Fig. 6a follow a scenario that is consistent with a late breaking of the microsatellite. In both of these conditions, the cell division begins in a similar fashion to the high-efficacy cases described above, with a strong correlation between the first division event and the final size of the colony, as seen in the shape of the $N_{\text{tot}}^{\text{local}}$ heat maps. However, the time for the first GFP+ detection is very long compared with the high-efficacy cases. This long delay is an indication that the break and repair events happen after several cell divisions, as shown schematically in the middle panel of each condition. It is possible that the repair step is also poorly

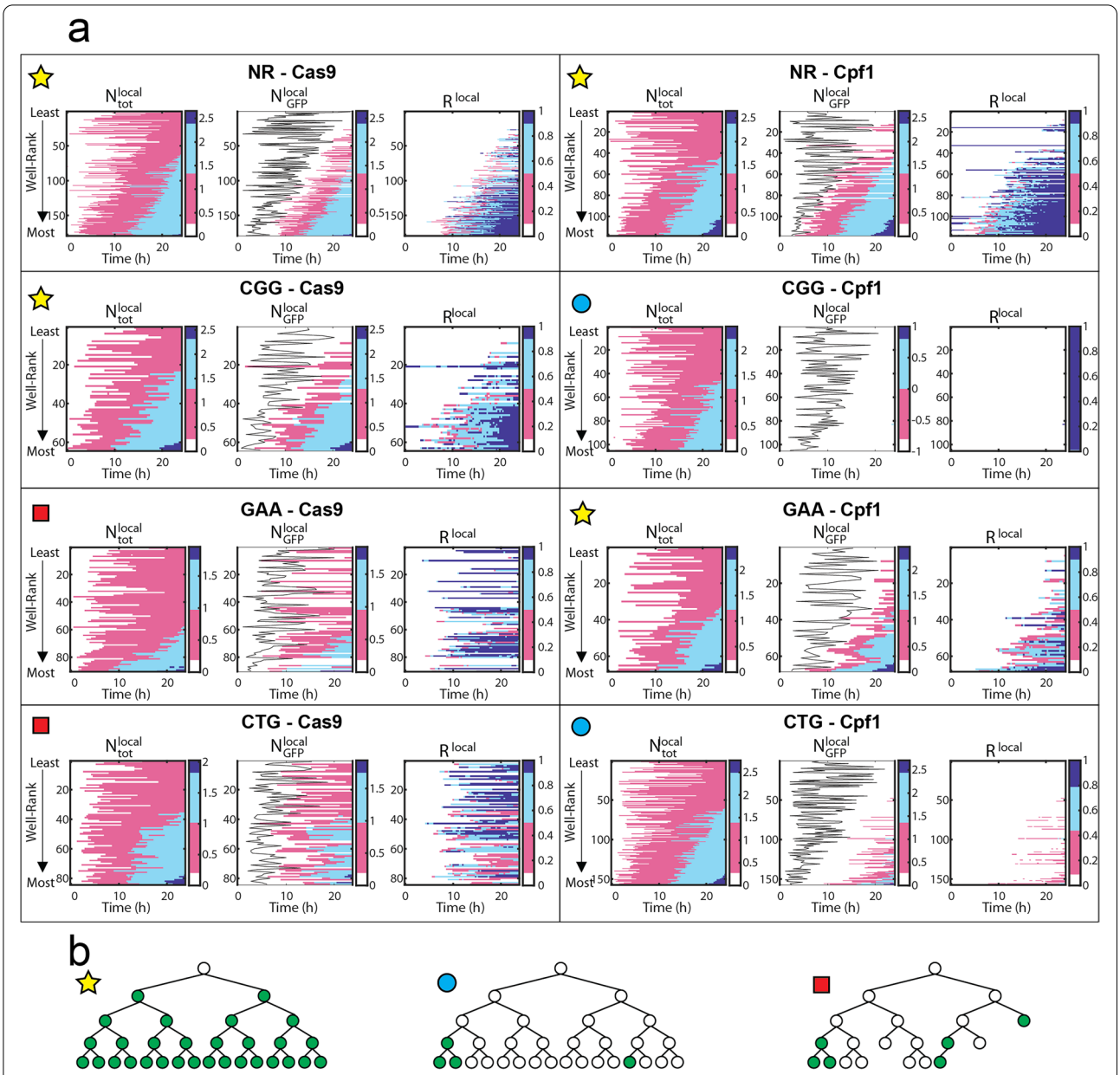

Fig. 6 Identification of DSBR cases by capturing single-cell variability. **a** Heat maps of individual wells per experimental condition: Each row per map represents data from an individual well, and the wells are ranked vertically in each map from those with least to most of cells at $t = 24$ h (see arrows on the left side). For each condition, from left to right: $\log_{10}(N_{\mathrm{tot}}^{\mathrm{local}})$, $\log_{10}(N_{\mathrm{GFP}}^{\mathrm{local}})$, and R^{local}. The color scales highlight extreme values near the 10th and 90th percentiles. In N_{local} maps, the black line indicates the first division time t_{lag}. **b** Cartoons of the three identified behaviors: high-efficacy error-free repair, normal cell growth (first panel); low-efficacy error-free repair, normal cell growth (middle panel); and low-efficacy error-prone repair, impaired cell growth (third panel)

performed by the cells, although it is not possible to confirm this from the current experiments. As a result of this long delay, the values of R^{local} all remain small at $t = 24$ h, in line with the low value of R^{global} (Fig. 4). CGG-Cpf1 and CTG-Cpf1 show this type of behavior, and their progeny tree would be similar to the one illustrated in Fig. 6b, second panel.

3. **Error-prone repair, impaired cell growth.** Different dynamics are evidenced by the analysis on the final two conditions, marked with the square in Fig. 6a. The appearance of GFP+ cells here is followed by a loss of fitness, marked by the slowing down or stopping of cell division. A consequence of this behavior is the broad distribution of wells that reach $R^{\mathrm{local}} \simeq 1$, both for small

and large final colony sizes. In contrast with the previous cases, the well with a high value of R^{local} is distributed throughout the whole range of colony sizes. This is also the only condition for which the value of R^{local} is not monotonically increasing but sometimes decreases.

Discussion

Bulk experiments, traditionally used to study DSBR, provide the ratios of broken or repaired cells to the total number of cells within a population. Such measurements are sometimes repeated during the course of an experiment to provide values at early, intermediate, and late time points, thus estimating the repair dynamics. It is nevertheless difficult to interpret the significance of the cell ratios. For example, it is not possible to know if the repaired cells at any time point constitute the progeny of a small number of efficient mother cells or if they are the result of a large number of independent repair events. Moreover, in the case of poor efficacy, it is not possible to determine if that is due to poor breaking, poor repair, or loss of fitness. Here, we addressed these issues by combining traditional molecular measurements with a dynamical ODE model and with time-resolved microfluidic imaging experiments.

From the ODE model, we were able to estimate the break and repair rates (β and ρ, respectively) and show that their distributions vary among conditions. Remarkably, the values of ρ are larger when Cas9 is induced than when Cpf1 is induced (Additional file 1: Fig. S5), suggesting that Cas9 DSB are repaired more quickly than Cpf1 DNA breaks. This difference may be related to the nature of the breaks produced by each of the nucleases: Cas9 makes blunt DSB [32], whereas Cpf1 makes staggered cuts, leaving 4–5 nucleotides 5′ overhangs [23, 33], that need to be resected for processing and repair of the break [34]. It is therefore possible that blunt DSB left by Cas9 are correctly processed by the cell, whereas 5′ overhangs left by Cpf1 are poorly resected, hindering effective DSBR. This would explain the longer repair time observed with Cpf1.

Even though the values of β and ρ were estimated from molecular measurements on populations of cells, the dynamics predicted by the ODE model matched remarkably well with the microfluidic measurements in most cases. Cases for which the match between model and experiment was not good yielded insights into additional biological mechanisms that were not suspected in advance. In particular, the microscopy enabled the detection of changes in cell morphology for the GAA-Cas9 and CTG-Cas9 conditions. Both of these conditions exhibited non-exponential growth after DSBR, suggesting that deleterious off-target mutations could have been induced by the Cas9 endonuclease. Poggi et al. [12] showed that Cas9 indeed induced frequent off-target mutations in the *LEO1* gene and less frequent ones in the *CLB5* gene when GAA were targeted and in the *YMR124w* gene when CTG were targeted. *LEO1* is involved in general transcription elongation whereas *CLB5* is a B-type cyclin involved in DNA replication. A null mutation causes slow growth, delayed progression through S and G1 phases of the cell cycle, and increased cell size, phenotypes that are recapitulated in the present experiments. *YMR124w* (also called *EPO1*) is involved in endoplasmic reticulum metabolism and interacts with *CRM1*, an essential gene encoding a nuclear export factor. The defects observed in our experiments could therefore be a direct or indirect effect of mutations in *YMR124w*.

Conclusions

In summary, we show here how to detect the DSBR dynamics at the single-cell level, by combining genetically modified cells with microfluidics and time-lapse microscopy. Then, by following the progeny of hundreds of individual cells, we provide a new framework to bridge the scales between the single-cell behavior and population dynamics. The link between these scales is further strengthened by a three-state coupled ODE model that coarse-grains this highly regulated process, involving dozens of proteins acting in a defined successive order. The mathematical model provides a quantitative basis to compare the dynamics observed in microfluidics with molecular and bulk measurements. The remarkable agreement between these different experimental approaches confirms that the microfluidic format does not introduce any artifactual bias. Instead, the ability to observe departures from the quantitative agreement, in combination with single-cell imaging, serves as a basis to distinguish between different repair scenarios: low-efficacy error-free and error-prone repair cases. Even though these scenarios are difficult to distinguish in bulk experiments, they correspond to widely different cellular histories and distribution of cell states. Ongoing work will then use this multiscale platform to identify specific events during the break and repair processes, which will help decipher differences in cell-to-cell response to DNA damage.

Methods

Biological protocols

Yeast plasmids and strains are described in Poggi et al. [12].

Time courses of DSB inductions Cells were transformed using standard lithium-acetate protocol [35] with both sgRNA and endonuclease and selected on 2% glucose synthetic complete, uracyl, leucine (SC-UR-LEU) plates and grown for 36 h. Each colony was seeded into 2 mL of 2% glucose SC-URA-LEU for 24 h and then diluted into 10 mL of 2% glucose SC-URA-LEU for 24 h as a pre-culture step. Cells were washed twice in water and diluted at ca. 7×10^6 cells/mL in 2% galactose SC-URA-LEU,

before being harvested at each time point (0h, 2h, 4h, 6h, 8h, 10h, 12h) for subsequent DNA extractions. The same cultures were used for cytometry analyses.

Southern blots For each Southern blot, 3–5 μg of genomic DNA digested with Eco RV and Ssp I were loaded on a 1% agarose gel and electrophoresis was performed overnight at 1 V/cm. The gel was manually transferred overnight in 20X SSC, on a Hybond-XL nylon membrane (GE Healthcare), according to manufacturer recommendations. Hybridization was performed with a 302 bp ^{32}P-randomly labeled CAN1 probe amplified from primers CAN133 and CAN135 [36]. Each probe was purified on a G50 column (ProbeQuant G50 microcolumn, GE Healthcare) and specific activities were verified to be above 2.4 × 108 cpm/μg. The membrane was exposed 3 days on a phosphor screen and quantifications were performed on a FujiFilm FLA-9000 phosphorimager, using the Multi Gauge (v. 3.0) software. Percentages of DSB and recombinant molecules were calculated as the amount of each corresponding band divided by the total amount of signal in the lane, after background subtraction. Note that DSB and repaired values were taken from Poggi et al. [12] for each strain, except for NR-Cpf1 for which two additional time courses and Southern blots were run.

Microfluidics and microfabrication

Master molds for the microfluidic devices were created using photolitography techniques by adapting the methods described in Ref. [37]: Briefly, designs were created with *CleWin* software and printed onto high-resolution polymer photomasks. Master molds were then fabricated with negative photoresist SU8 onto silicon wafers, following a double-layer procedure in order to obtain different specific heights for the wells and the chamber. Microfluidic devices were created using two pieces of polydimethylsiloxane (PDMS): one thin (∼ 300 μm) layer patterned by the master mold described before and a second blank thick (∼ 8 mm) slab where inlet and outlets were forged. The whole device was assembled, using plasma oxygen, as follows (from bottom to top): a glass slide, the patterned PDMS layer facing up, and the blank PDMS slab closing the microfluidic chamber.

In each experiment, cells were introduced into the microfluidic chip, at 5 μL/min, controlled by a syringe pump system (Nemesys cetoni) and were allowed to settle on the bottom of the device for 5 min. Subsequently, the culture medium was supplied at 10 μL/min for at least 10 min in order to remove non-trapped cells. The well occupancy did not follow a homogeneous distribution: wells typically contained from 0 to 5 cells. Only populations that started with a single cell were selected for our analysis in order to monitor the lineage of individual cells. In this context, wells that were contaminated by cells that were not stemming from the original trapped cells were discarded, as well as wells disturbed by air bubbles at some point of the time lapse. Cells were cultured inside the microfluidic device with a culture medium continuously supplied at low flow rates (0.1 μL/min) over 24 h in order to ensure viability and favorable growth conditions. The chip and the syringe pump were maintained at 30° on a temperature-controlled box (Oko lab) mounted on top of an inverted microscope (Nikon eclipse) for 24 h.

Image acquisition and analysis

The whole microfluidic chip was imaged with a 20× objective, every 20 min both in bright-field and in green epi fluorescence. On such an imaging routine, a rectangular lattice was followed by the motorized stage in order to obtain 176 (22 × 8) fields of view (each 600 μm × 600 μm). The images were processed with the open-source software *ImageJ* [21]. First, only those image sets with wells that contain one single cell at the beginning of the experiment were selected and cropped. Such image sets were structured into hyperstacks of 73 (73 time points) per two (two color channels: bright-field and green fluorescence) images. Using the *ImageJ* plugin *TrackMate* [20], the number of cells at every experimental time point both in bright-field and green fluorescence channels was computed: Using the bright-field channel in each time point, round elements (with a specific size ∼ 3.5 μm diameter) inside the region of interest (ROI) were detected and segmented. Such selection was then applied to both channels in order to measure the mean intensity value in the selected circles. In this manner, we could determine if a cell (contained in the circle selection) was expressing GFP (GFP+) by comparing its mean intensity value (measured in the green fluorescence channel) to the background mean intensity value. If the measured value of green fluorescence in the cell was more than 1.5× the background level, then the cell was considered GFP+. This method provided a time-resolved quantification of both proliferation of cells and their GFP expression upon DNA repair, as shown in Fig. 1c, d.

The delay between recent repaired *GFP* gene and the GFP detection on single yeast cells was estimated to be 3 h. This value was estimated by comparing Southern blot data and expression curves obtained with the microfluidic setup and the image processing here explained. This delayed would correspond to the parameter γ on Fig. 3a.

Bayesian inference

Prior selection

Bayesian inference of the ODE model parameters from the Southern blot measurements was performed

Table 1 ODE model parameters

Parameter	Description	Prior mean	Prior CV
α	Growth rate of cells	$\ln(2)/3\,\mathrm{h}^{-1}$	1
β	DNA break rate	$\ln(2)/6\,\mathrm{h}^{-1}$	2
ρ	DNA repair rate	$\ln(2)/6\,\mathrm{h}^{-1}$	2
τ	Lag time	4 h	0.1
σ	Measurement noise	$\ln(2)/3$	2

in Julia using Markov Chain Monte Carlo simulations using the Turing.jl library [38]. Our prior distributions were independent gamma distributions for each parameter. The gamma distribution is parameterized by a shape and scale parameters, denoted α and θ, respectively. The mean and coefficient of the gamma distribution are given by $\mu = \alpha\theta$ and $\mathrm{CV} = 1/\sqrt{\alpha}$, respectively. For each parameter, we selected a mean and a CV which constrained the parameters within some physical reasonable range. In particular, we know that the time scale for double-strand breaks to appear in the population is less than the length of the experiment, so β is not likely less than $\ln(2)/12\,\mathrm{h}^{-1}$. On the other hand, broken cells do not appear instantly, so it is not likely to be more than $\ln(2)\,\mathrm{h}^{-1}$. Similarly, for α, we know that the doubling time is on the order of 3 h, but it could be as large as 6 or as small as 1; thus, we take priors with a mean of ln(2) and CV of 1. We believe that τ is the same for each experiment; thus, we infer τ from a single condition and used an approximation of the resulting posterior as priors for all other experiments. Table 1 lists the mean and variance we used for each parameter.

These are so-called weakly informative priors, meaning they are not meant to incorporate specific information we have, e.g., from a previous experiment, but rather make parameter values which are physically implausible highly unlikely.

Diagnostics on simulated data

We first tested the Bayesian inference on simulated data from the ODE model, with uncorrelated Gaussian errors added to the species fractions. Additional file 1: Fig. S2 shows a pair plot with the joint posterior distribution of each parameter pair, along with the true parameter values used to generate the simulated data for the fraction of modified, broken, and repaired cells.

In order for the parameters extracted from the Bayesian inference to be biologically meaningful, the inference should be robust to violations in the model assumptions. Thus, we next tested that the Bayesian inference can still resolve the parameters when the Gaussian error model is incorrect. To generate non-Gaussian errors, we assumed that the Southern blot measurements themselves, rather than the fractions, are corrupted by Gaussian noise. Additional file 1: Fig. S3 shows a pair plot for this simulated data. Code to reproduce the Bayesian inference can be found at https://github.com/elevien/yeast_dna_repair.

Stochastic model

The ODE model describes the evolution of cell numbers when there is a sufficiently large number of cells to neglect small number, or demographic, fluctuations. Invalid for the microfluidic experiments, however, we must consider a stochastic model which treats the events of cell division, DNA break, and repair probabilistically. There are many ways to do this, but we adapt a simple approach of assuming all events occur at exponentially distributed times with rate parameters α, β, and ρ, respectively. As a result of this assumption, the stochastic process for (m, g, r) is Markovian, meaning that it is not necessary to have knowledge of how long each of the cells has been in a given state to predict the future evolution. The probability distribution $P(m, g, r)$ can be shown to obey the Master equation [39]

$$\begin{aligned}\frac{d}{dt}P(m,g,r,t) =& \alpha(m-1)P(m-1,g,r,t) + \beta(m+1)P(m+1,g-1,r,t)\\ &+ \rho(g+1)P(m,g+1,r-1,t) + \alpha(r-1)P(m,g,r-1,t)\\ &- P(m,g,r,t)[\alpha(m+r) + \beta m + \rho g].\end{aligned} \tag{5}$$

In our stochastic simulation samples, paths of the process (m, g, r) are generated using the Gillespie Algorithm [39].

It should be noted that while the assumption that events occur with a constant probability per unit time is strictly speaking false, as we know, cell division does not happen at a constant rate per unit time, but for making qualitative predictions about the fluctuations, it is sufficient.

Abbreviations

DSBR: Double-strand break repair; GFP: Green fluorescent protein; DSB: Double-strand break; PAM: Protospacer adjacent motif; ODE: Ordinary differential equation; PDMS: Polydimethylsiloxane; RMSE: Root mean square error; SC: Synthetic complete; URA: Uracyl; LEU: Leucine.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12915-022-01456-3.

Additional movie: Time-lapse microscopy of microfluidic wells. **Table S1:** RMSE of experimental population vs simulated data on Fig. 4. **Figure S1:** Southern blot quantification of DSB and DSBR for all experimental conditions. **Figure S2:** Test of Bayesian inference on simulated data with correctly specified model. **Figure S3:** Test of Bayesian inference on simulated data with non-Gaussian errors. **Figure S4:** Prior and posterior distributions of β, ρ, τ and α for all experimental conditions. **Figure S5:** Posterior distributions of breaking (β) and repair (ρ) rates for all experimental conditions. **Figure S6:** Comparison of ODE model predictions to Southern blot quantification of DSB and DSBR for all experimental conditions. **Figure S7:** Examples of individual R^{local} trajectories from stochastic simulations. Figure S8: Explanatory schematic for reading the heat map on Fig. 6.

Acknowledgements
We would like to thank the Microfluidics and Biomaterials core facility at Institut Pasteur, where we made our microfluidic devices. The authors acknowledge the guidance of the Image Analysis Hub of the Institut Pasteur for the image processing, as well as useful discussions with Erik Maikranz.

Authors' contributions
N.V.-Q., G.-F.R., and C.N.B. contributed to the conception of the work. N.V.-Q. and L.P. performed the acquisition of experimental data, including the design, development, and optimization of the microfluidic platform. E.L. performed Bayesian analysis and mathematical modeling. A.A., G.-F.R., and C.N.B. contributed to the interpretation of data. N.V.-Q., G.-F.R., and C.N.B. have drafted the work and revised it. The authors read and approved the final manuscript.

Funding
This project was partially funded by the Inception program of Institut Pasteur. Equipment was partially founded by DIM ELICIT. LP was supported by a PhD thesis fellowship from Fondation Blanchecape and Association Française contre les Myopathies (AFM). Work in the GFR laboratory is supported by the Institut Pasteur and the Centre National de la Recherche Scientifique (CNRS). We acknowledge funding support from NSF Grant DMS-1902895 (EL).

Availability of data and materials
All data generated or analyzed during this study are included in this published article and its supplementary information files. Experimental data and code to reproduce the Bayesian inference can be found at https://github.com/elevien/yeast_dna_repair.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Physical Microfluidics and Bioengineering Unit, Institut Pasteur, 75015 Paris, France. [2]Mathematics Department, Dartmouth College, 03755 Hanover, NH, USA. [3]Natural and Synthetic Genome Instabilities Group, Institut Pasteur, CNRS UMR3525, 75015 Paris, France. [4]John A. Paulson School of Engineering and Applied Sciences, Harvard University, 02138 Cambridge, MA, USA. [5]LadHyX, CNRS, Ecole Poytechnique, Institut Polytechnique de Paris, 91120 Palaiseau, France.

Received: 10 May 2022 Accepted: 31 October 2022
Published online: 05 December 2022

References

1. International Human Genome Sequencing Consortium, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409(6822):860–921.
2. Ellegren H. Microsatellites: simple sequences with complex evolution. Nat Rev Genet. 2004;5:435–45. https://doi.org/10.1038/nrg1348.
3. Richard GF, Kerrest A, Dujon B. Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. Microbiol Mol Biol Rev. 2008;72(4):686–727.
4. Orr HT, Zoghbi HY. Trinucleotide repeat disorders. Annu Rev Neurosci. 2007;30:575–621.
5. McMurray CT. DNA secondary structure: a common and causative factor for expansion in human disease. Proc Natl Acad Sci. 1999;96(5):1823–5.
6. Poggi L, Richard GF. Alternative DNA structures in vivo: molecular evidence and remaining questions. Microbiol Mol Biol Rev. 2021;85(1).
7. Mosbach V, Poggi L, Richard GF. Trinucleotide repeat instability during double-strand break repair: from mechanisms to gene therapy. Curr Genet. 2019;65(1):17–28.
8. Richard GF. Shortening trinucleotide repeats using highly specific endonucleases: a possible approach to gene therapy? Trends Genet. 2015;31(4):177–86. https://doi.org/10.1016/j.tig.2015.02.003.
9. Haber JE. In vivo biochemistry: physical monitoring of recombination induced by site-specific endonucleases. Bioessays. 1995;17(7):609–20.
10. Plessis A, Perrin A, Haber J, Dujon B. Site-specific recombination determined by I-SceI, a mitochondrial group I intron-encoded endonuclease expressed in the yeast nucleus. Genetics. 1992;130(3):451–60.
11. Doudna JA, Charpentier E. The new frontier of genome engineering with CRISPR-Cas9. Science. 2014;346(6213).
12. Poggi L, Emmenegger L, Descorps-Declère S, Dumas B, Richard GF. Differential efficacies of Cas nucleases on microsatellites involved in human disorders and associated off-target mutations. Nucleic Acids Res. 2021;49(14):8120–34.
13. Charlebois DA, Balázsi G. Modeling cell population dynamics. In Silico Biol. 2019;13(1–2):21–39.
14. Liu P, Young TZ, Acar M. Yeast replicator: a high-throughput multiplexed microfluidics platform for automated measurements of single-cell aging. Cell Rep. 2015;13(3):634–44.
15. Young DJ, Guydosh NR. Hcr1/eIF3j is a 60S ribosomal subunit recycling accessory factor in vivo. Cell Rep. 2019;28(1):39–50.
16. Schnitzer B, Borgqvist J, Cvijovic M. The synergy of damage repair and retention promotes rejuvenation and prolongs healthy lifespans in cell lineages. PLoS Comput Biol. 2020;16(10):e1008314.
17. Song J, Peng W, Wang F. A random walk-based method to identify driver genes by integrating the subcellular localization and variation frequency into bipartite graph. BMC Bioinformatics. 2019;20(1):1–17.
18. Song R, Peng W, Liu P, Acar M. A cell size-and cell cycle-aware stochastic model for predicting time-dynamic gene network activity in individual cells. BMC Syst Biol. 2015;9(1):1–11.
19. Amselem G, Guermonprez C, Drogue B, Michelin S, Baroud CN. Universal microfluidic platform for bioassays in anchored droplets. Lab Chip. 2016;16(21):4200–11.
20. Tinevez JY, Perry N, Schindelin J, Hoopes GM, Reynolds GD, Laplantine E, et al. TrackMate: an open and extensible platform for single-particle tracking. Methods. 2017;115:80–90.
21. Schneider CA, Rasband WS, Eliceiri KW. NIH Image to ImageJ: 25 years of image analysis. Nat Methods. 2012;9(7):671–5.
22. Sternberg SH, Redding S, Jinek M, Greene EC, Doudna JA. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. Nature. 2014;507(7490):62–7.
23. Świat MA, Dashko S, den Ridder M, Wijsman M, van der Oost J, Daran JM, et al. Fn Cpf1: a novel and efficient genome editing tool for Saccharomyces cerevisiae. Nucleic Acids Res. 2017;45(21):12585–98.
24. Makarova KS, Wolf YI, Iranzo J, Shmakov SA, Alkhnbashi OS, Brouns SJ, et al. Evolutionary classification of CRISPR-Cas systems: a burst of class 2 and derived variants. Nat Rev Microbiol. 2020;18(2):67–83.
25. DiCarlo JE, Norville JE, Mali P, Rios X, Aach J, Church GM. Genome engineering in Saccharomyces cerevisiae using CRISPR-Cas systems. Nucleic Acids Res. 2013;41(7):4336–43.

26. Guarente L, Yocum RR, Gifford P. A GAL10-CYC1 hybrid yeast promoter identifies the GAL4 regulatory region as an upstream site. Proc Natl Acad Sci. 1982;79(23):7410–4.
27. Usui T, Ogawa H, Petrini JH. A DNA damage response pathway controlled by Tel1 and the Mre11 complex. Mol Cell. 2001;7(6):1255–66.
28. Hersen P, McClean MN, Mahadevan L, Ramanathan S. Signal processing by the HOG MAP kinase pathway. Proc Natl Acad Sci. 2008;105(20):7165–70.
29. Jo MC, Liu W, Gu L, Dang W, Qin L. High-throughput analysis of yeast replicative aging using a microfluidic system. Proc Natl Acad Sci. 2015;112(30):9364–9.
30. Charlebois DA, Hauser K, Marshall S, Balázsi G. Multiscale effects of heating and cooling on genes and gene networks. Proc Natl Acad Sci. 2018;115(45):E10797–806.
31. van de Schoot R, Depaoli S, King R, Kramer B, Märtens K, Tadesse MG, et al. Bayesian statistics and modelling. Nat Rev Methods Prim. 2021;1(1):1–26.
32. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science. 2012;337(6096):816–21.
33. Zetsche B, Gootenberg JS, Abudayyeh OO, Slaymaker IM, Makarova KS, Essletzbichler P, et al. Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. Cell. 2015;163(3):759–71.
34. Cejka P, Symington LS. DNA end resection: mechanism and control. Annu Rev Genet. 2021;55:285–307.
35. Gietz RD, Schiestl RH, Willems AR, Woods RA. Studies on the transformation of intact yeast cells by the LiAc/SS-DNA/PEG procedure. Yeast. 1995;11(4):355–60.
36. Viterbo D, Marchal A, Mosbach V, Poggi L, Vaysse-Zinkhöfer W, Richard GF. A fast, sensitive and cost-effective method for nucleic acid detection using non-radioactive probes. Biol Methods Protocol. 2018;3(1):bpy006.
37. Amselem G, Sart S, Baroud CN. Universal anchored-droplet device for cellular bioassays. Methods Cell Biol. 2018;148:177–99.
38. Ge H, Xu K, Ghahramani Z. Turing: a language for flexible probabilistic inference. In: International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain; 2018. p. 1682–1690. http://proceedings.mlr.press/v84/ge18b.html. Accessed 28 Feb 2022.
39. Bressloff PC. Stochastic processes in cell biology. Berlin: Springer; 2014.

Publisher's Note