Riches of phenotype computationally extracted from microbial colonies

Tzu-Yu Liu[a,b,1], Anne E. Dodson[c,1], Jonathan Terhorst[d], Yun S. Song[a,b,d,e,2], and Jasper Rine[c,2]

[a]Department of Mathematics and Department of Biology, University of Pennsylvania, Philadelphia, PA 19104; [b]Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720; [c]Department of Molecular and Cell Biology and California Institute for Quantitative Biosciences, University of California, Berkeley, CA 94720; [d]Department of Statistics, University of California, Berkeley, CA 94720; and [e]Department of Integrative Biology, University of California, Berkeley, CA 94720

Contributed by Jasper Rine, April 7, 2016 (sent for review November 24, 2015; reviewed by Michael B. Elowitz and Andrew W. Murray)

The genetic, epigenetic, and physiological differences among cells in clonal microbial colonies are underexplored opportunities for discovery. A recently developed genetic assay reveals that transient losses of heterochromatic repression, a heritable form of gene silencing, occur throughout the growth of *Saccharomyces* colonies. This assay requires analyzing two-color fluorescence patterns in yeast colonies, which is qualitatively appealing but quantitatively challenging. In this paper, we developed a suite of automated image processing, visualization, and classification algorithms (MORPHE) that facilitated the analysis of heterochromatin dynamics in the context of colonial growth and that can be broadly adapted to many colony-based assays in *Saccharomyces* and other microbes. Using the features that were automatically extracted from fluorescence images, our classification method distinguished loss-of-silencing patterns between mutants and wild type with unprecedented precision. Application of MORPHE revealed subtle but significant differences in the stability of heterochromatic repression between various environmental conditions, revealed that haploid cells experienced higher rates of silencing loss than diploids, and uncovered the unexpected contribution of a sirtuin to heterochromatin dynamics.

heterochromatin dynamics | epigenetics | image segmentation | feature extraction

Microbial colonies arising from single cells have been a workhorse of molecular genetics for decades, yet the genetic and physiological complexity of the population of cells within a colony is often overlooked. For most microbes, the number of cells in a colony is sufficiently large to contain, in some fraction of cells, a loss-of-function mutation in every gene in the genome, and even the majority of possible base-pair changes. The physiology of cells located in different regions of the colony can also vary widely due to limitations of oxygen and nutrient diffusion through the colony (1, 2). Reporter-gene fusions have revealed some of the remarkable differences between cells in the same colony (3). To date, analyzing colony-wide patterns of reporter-gene expression and how they change in response to mutations has been limited to qualitative approaches. For these patterns to serve as a reliable phenotype, however, rigorous quantitation is necessary. In this work, colonial patterns resulting from the dynamic nature of heterochromatin formed the basis upon which to develop a quantitatively robust pattern classifier.

Heterochromatin is a tightly packed state of chromatin that represses, or silences, the expression of genes within it. Furthermore, heterochromatin is an epigenetically heritable form of chromatin structure that helps maintain chromosome segregation fidelity and genome stability. Repression of gene expression in heterochromatin is an important form of gene regulation, but currently little is understood about its dynamics or stability. In *Saccharomyces cerevisiae*, heterochromatin plays an important role in stabilizing the highly repetitive telomeres and ribosomal DNA repeats (4). Heterochromatin also mediates silencing of the cryptic mating-type loci (*HML* and *HMR*) so that only the mating-type allele at the *MAT* locus is expressed (5). *HML* and *HMR* are epigenetically silenced by the Sir (Silent information regulator) proteins Sir1, Sir2, Sir3, and Sir4 (6), which are the structural components of heterochromatin at these loci.

Our recent genetic assay, based on Cre-*loxP* recombination, captures transient losses of gene silencing in *S. cerevisiae* by converting these transient events into a permanent and heritable feature (Fig. 1*A* and ref. 7). In this assay, transient expression of *HML::cre* catalyzes a recombination event that removes a red fluorescent protein (RFP) gene and substitutes a GFP gene in such a way that cells that were red are now green, as are all of their descendants. The Cre-catalyzed changes in genotype and phenotype are permanent and heritable, leading to characteristic two-color fluorescence patterns in yeast colonies (Fig. 1*B*). We hereinafter refer to this method as the Cre-Reported Altered States of Heterochromatin (CRASH) assay.

To date, the only method available for quantifying the dynamics of heterochromatic repression is half-sector analysis (8), whereby the rate of RFP-to-GFP switches per cell division is determined by measuring the frequency of half-sectored colonies. However, half-sector analysis of rare events is laborious and, because it is based upon events confined to the first cell division of colony growth, potentially misses information reflected in the patterns of green spots and sectors throughout colonies. In this paper, we developed a suite of automated image processing, visualization, and classification algorithms to facilitate the analysis of heritable and clonal red-to-green transitions that occurred during the growth of a colony. This suite of programs was built on the basis of mathematical morphological operations, and we refer to it as MORphological PHenotype Extraction (MORPHE). It is freely available at https://sourceforge.net/projects/morphe.

Using MORPHE, we automatically extracted a set of useful features from the observed patterns produced by GFP-expressing

Significance

The genome and physiology of a cell can undergo complex changes among the many cells that make up a growing microbial colony. Genetic and physiological dynamics can be revealed by measuring reporter-gene expression, but rigorous quantitative analysis of colony-wide patterns has been underexplored. Here, we developed a suite of automated image processing, feature extraction, visualization, and classification algorithms to facilitate the analysis of sectoring patterns in *Saccharomyces* colonies. Classification results for various mutants and for colonies grown under different environmental conditions revealed significant differences in sectoring that were not apparent by visual inspection.

Author contributions: T.-Y.L., A.E.D., J.T., Y.S.S., and J.R. designed research; T.-Y.L. and A.E.D. performed research; T.-Y.L., A.E.D., J.T., Y.S.S., and J.R. analyzed data; and T.-Y.L., A.E.D., Y.S.S., and J.R. wrote the paper.

Reviewers: M.B.E., Howard Hughes Medical Institute; and A.W.M., Harvard University.

The authors declare no conflict of interest.

[1]T.-Y.L. and A.E.D. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: jrine@berkeley.edu or yss@berkeley.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1523295113/-/DCSupplemental.

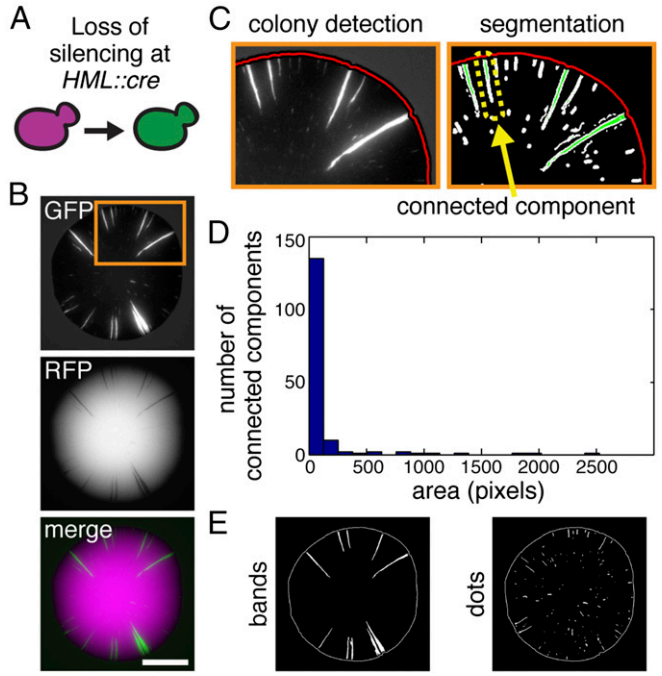

Fig. 1. Design of the CRASH assay and detection of switching events by MORPHE. (*A*) The CRASH assay captures transient losses of silencing at *HML::cre* through a permanent, red-to-green switch in fluorescence. (*B*) Fluorescence of a colony of haploid cells containing *HML::cre* and the fluorescent reporter construct. (Scale bar, 2 mm.) (*C*) Close-up of colony shown in *B* (orange box) following colony detection (*Left*) and segmentation (*Right*). For each contour of a connected component (i.e., the boundary of the connected component, found by edge detection and dilation), we compared the pixel intensities of the interior versus the pixel intensities on the contour. If the interior pixels had higher intensities, the area enclosed by the contour was labeled as a bright region, shown in green. Once each contour was traced, we found the connected components within the colony and computed the area of each connected component. (*D*) Most of the detected connected components had an area of less than 500 pixels. (*E*) Band and dot features, both of which originated from loss-of-silencing events, were classified by thresholding the area of each connected component.

cells in a colony (also referred to below as switching patterns) and performed classification on patterns of colonies from various yeast strains and from wild-type strains grown under various environmental conditions. MORPHE enabled multiple discoveries that had eluded all previous methods by applying quantitative image analysis to the classification of phenotype.

Results

In a microbial colony of cells containing a *cre* gene silenced by heterochromatin and a fluorescent reporter cassette, transient failures of silencing produce GFP-expressing sectors or dots (Fig. 1 *A* and *B*) (7). Multiple features of a GFP sector, such as its size and position within the colony, as well as the overall pattern of colony sectoring, inform our understanding of heterochromatin dynamics. To quantitatively analyze these features, we developed an algorithm that detects the sectoring pattern (Fig. 1*C*). Large regions of GFP expression, referred to as bands, typically extended to the periphery of a colony, whereas most small regions of GFP expression, referred to as dots, did not extend to the colony edge, most likely due to neighboring RFP-positive cells randomly taking over the local population (Fig. 1 *D* and *E*). Because both bands and dots arise from a loss of silencing, we combined these two features to determine the overall number of switching events from RFP to GFP expression, which was a direct measure of transient loss-of-silencing (transient failures of heterochromatin) events.

The frequency of switching events was computed in terms of the distance of bands and dots from the center of the colony (see *Methods* for a description of the algorithm). We first labeled the vertex of each band and dot (Fig. 2*A*), because the vertex marks the location of the cell that experienced a loss of silencing at *HML::cre* and therefore marks the origin, or onset, of GFP expression. Next, we determined onset frequencies by confining the analysis to a specific distance from the colony center and quantifying the proportion of vertices within that ring, or annulus (Fig. 2*B*). Because the switch from RFP expression to GFP expression is irreversible and thus GFP-expressing cells do not have the potential to undergo a second switch, GFP-expressing regions other than vertices were excluded from measurements of the total area. This step was repeated for a series of concentric circles with increasing radii, resulting in plots of onset frequency as a function of distance from the colony center (Fig. 2*C*). A summary statistic was obtained by taking the average of the frequency function and was denoted as the mean onset frequency.

In addition to onset frequency, we also extracted a feature defined as the area of GFP fluorescence (Fig. 2*C*). This measurement corresponded to the number of pixels containing GFP signal within the annulus of interest. Together, the onset frequency and area of GFP fluorescence at each given radius provided a set of features by which to compare colonial patterns of GFP expression.

Feature Extraction Detected Obvious Phenotypes, in Agreement with Previous Analyses, As Well As Less Obvious Phenotypes That Escaped Previous Analyses. To test whether the computational method could distinguish patterns that markedly differ by visual inspection, we performed image analysis on both wild-type colonies and *hst3*Δ colonies. Deletion of the *HST3* gene, which encodes a NAD^+-dependent histone deacetylase known to target histone H3K56ac, reduced the stability of heterochromatic repression and thus caused a dramatic increase in the frequency of

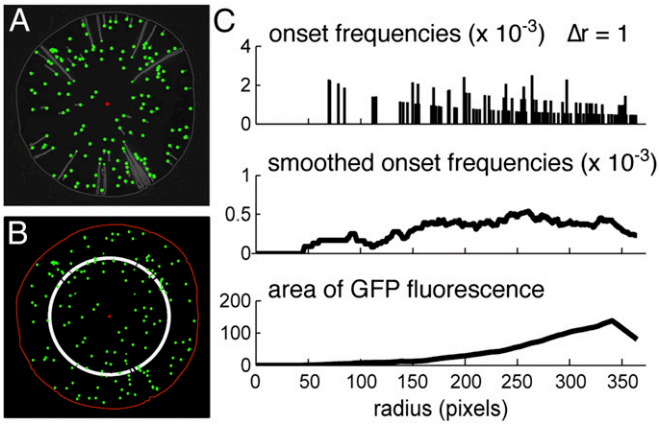

Fig. 2. Features extracted from the fluorescence pattern of a colony. (*A*) The origin of the colony is shown as a red dot. The vertex of each detected connected component is represented by a green circle. Each vertex records a point in time when a loss-of-silencing event occurred. (*B*) The onset frequency was defined as the number of switching events divided by the area in the white ring at each given radius. The difference between the outer radius and inner radius is denoted as Δr (pixels). Because the switch to GFP expression is irreversible, we excluded the area of GFP-expressing regions from the calculation. (*C*) The smoothed onset frequencies were obtained by applying a sliding window across the onset-frequency spike trains and taking the average within the window. The window size was fixed to 50 pixels in this example. We also computed the area of GFP fluorescence in the white ring at each radius.

sectors (Fig. 3*A*), as previously described (7). Consistent with qualitative analysis and half-sector analysis (7), the computational analysis distinguished wild-type colonies from *hst3*Δ colonies with regard to all extracted features (Fig. 3). MORPHE successfully captured the early onsets and the large regions expressing GFP in the *hst3*Δ colonies relative to wild type.

The colonies of cells lacking the NAD^+-dependent deacetylase Hst1, Hst2, or Hst4 exhibited sectoring patterns that were indistinguishable from that of wild type by visual inspection (Fig. 3*A*). However, feature extraction revealed that the mean onset frequency in *hst1*Δ colonies was slightly yet significantly lower than the mean onset frequency in wild-type colonies (Fig. 3*D* and *SI Appendix*, Table S2), suggesting that in wild-type cells, Hst1 destabilized heterochromatic repression to a limited extent. To provide an independent test of the results of the computational analysis, we turned to the traditional half-sector analysis to determine rates of switching events in microbial colonies (8). Consistent with the feature extraction, half-sector measurements showed that deletion of *HST1* caused a subtle reduction (P = 0.004; Student's t test) in the rate of silencing loss. Whereas wild-type cells lose silencing at a rate of 1.58×10^{-3} ($\pm 7 \times 10^{-5}$) per cell division (7), *hst1*Δ cells lost silencing at a rate of 1.2×10^{-3} ($\pm 1 \times 10^{-4}$) per cell division. Thus, MORPHE uncovered a silencing phenotype in *hst1*Δ mutants that previously escaped detection by visual inspection and that could be confirmed, rather laboriously, by the traditional method.

Haploid Cells Exhibited a Lower Switching Rate Relative to Diploid Cells. MORPHE provided a convenient way of visualizing the switching pattern for multiple colonies and was therefore applied to micrographs of diploids lacking one copy of individual *SIR* genes (Fig. 4*A*). In concordance with measurements acquired by traditional half-sector analysis (7), feature extraction showed that diploids containing only one copy of either the *SIR1*, *SIR3*, or *SIR4* gene had higher onset frequencies of switching events and larger GFP-expressing regions compared with wild type (Fig. 4 and *SI Appendix*, Table S3).

A direct comparison of the wild-type haploid and wild-type diploid revealed that haploid colonies exhibited higher onset frequencies of GFP-expressing regions and larger areas of overall GFP expression than diploid colonies (Fig. 5 and *SI Appendix*, Table S4). This observation was consistent with the fourfold increase in the frequency of half-sectored colonies in haploids relative to diploids (7). The diploid strains used in this study were pseudohaploids due to the deletion of one copy of the *MAT* locus. Therefore, the difference between haploids and diploids in red-to-green switching frequencies could not be attributed to any changes in the expression of haploid-specific genes.

In principle, a change in the frequency of red-to-green switches could arise from either a change in the stability of silencing or a change in the efficiency of Cre-*loxP* recombination. Given that the diploid was hemizygous for the RFP-GFP cassette and therefore contained twice the ratio of DNA content to *loxP* sites in comparison with the haploid, we considered the possibility that Cre was less efficient at targeting the *loxP* sites in the diploid. Indeed, increasing the number of RFP-GFP cassettes in the diploid from one copy (JRY10639) to two copies (JRY10640) increased the mean onset frequency, albeit not up to the level of the haploid (Fig. 5 and *SI Appendix*, Table S4). Therefore, the dosage of RFP-GFP cassettes affected the efficiency of Cre-*loxP* recombination in the diploid and contributed in part to the difference in switching frequencies between haploids and diploids.

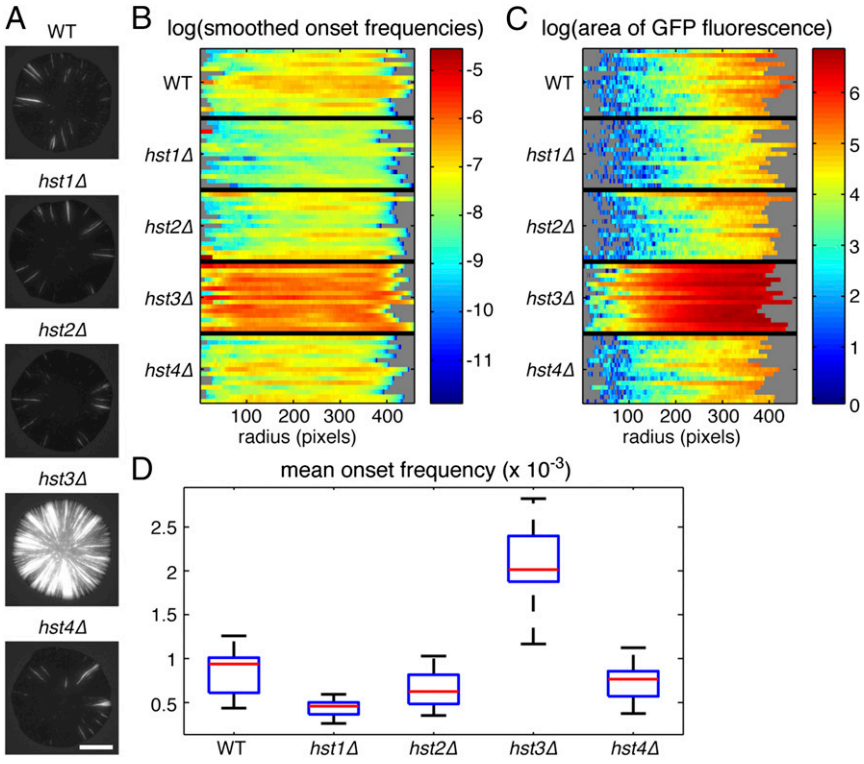

Fig. 3. Feature extraction of haploid colonies. (*A*) GFP fluorescence of representative colonies for haploid strains containing individual deletions of sirtuin genes. (Scale bar, 2 mm.) (*B*) Smoothed onset frequencies of switching events for each genotype. The horizontal axis represents the distance from the origin in pixels, and each row represents a colony. The color bar indicates the natural logarithm of smoothed onset frequencies. (*C*) The area of GFP fluorescence. The color bar indicates the natural logarithm of the area of GFP fluorescence. (*D*) Boxplot of mean onset frequencies. Hereinafter, the red line represents the median, the whiskers extend to the most extreme values that lie within 1.5 times the interquartile range (box edges), and plus signs represent outliers.

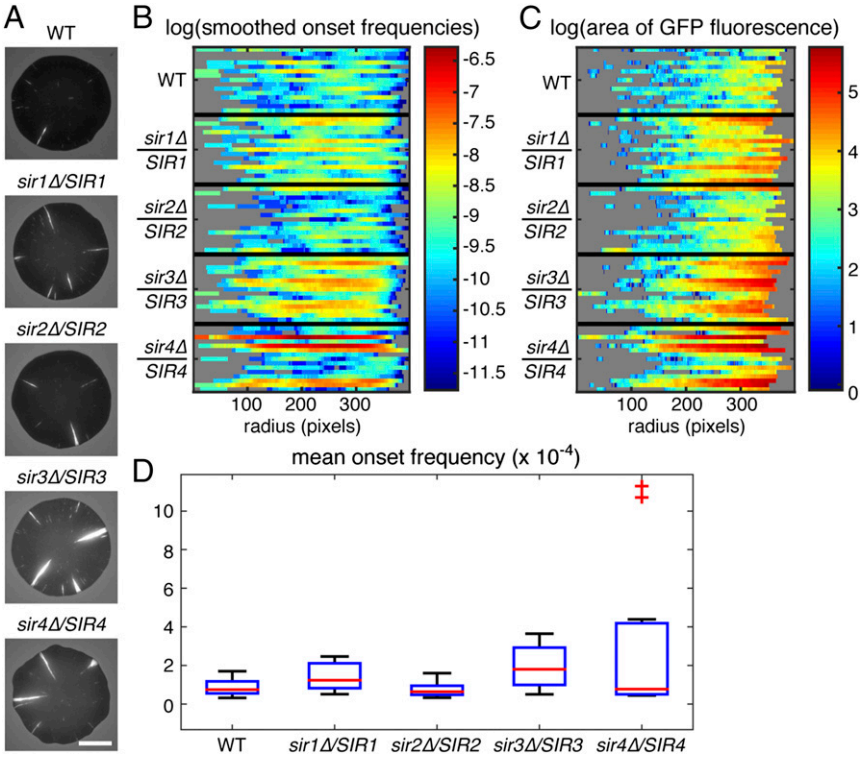

Fig. 4. Feature extraction of diploid colonies. (*A*) GFP fluorescence of representative colonies for diploid strains hemizygous for individual *SIR* genes. (Scale bar, 2 mm.) (*B*) The smoothed onset frequencies of switching events for each genotype. The horizontal axis represents the distance from the origin in pixels, and each row represents a colony. The color bar indicates the natural logarithm of smoothed onset frequencies. (*C*) The area of GFP fluorescence. The color bar indicates the natural logarithm of the area of GFP fluorescence. (*D*) Boxplot of mean onset frequencies.

It is unknown whether transient losses of silencing reflect a local disturbance in heterochromatin or rather a systemic failure in the repression of all heterochromatic loci. To distinguish between these two possibilities, we used MORPHE to compare the frequency of silencing loss between a diploid that was hemizygous for *HML::cre* (JRY10639) and a diploid that was homozygous for *HML::cre* (JRY10641). If instability arose from a locus-specific event, then the two *HML::cre* loci in the homozygote would lose silencing independently of each other and thus double the frequency of red-to-green switches relative to the hemizygote. Alternatively, a cell-wide disruption to heterochromatin would cause a concurrent loss of silencing at both *HML::cre* loci and thus trigger red-to-green switches at the same rate as the hemizygote. MORPHE revealed that the mean frequency of switching events was approximately twofold higher in the *HML::cre* homozygote than in the *HML::cre* hemizygote (Fig. 5 and *SI Appendix*, Table S4). This trend occurred between diploids containing one copy of the RFP-GFP cassette, as well as between diploids containing two copies of the RFP-GFP cassette (Fig. 5). These results, which suggested that the majority of loss-of-silencing events were locus-specific, were consistent with the observation that a *sir1*Δ diploid containing a unique reporter gene at each *HML* locus shows no correlation in expression state between the two *HML* alleles (9). In principle, however, loci that lose silencing in concert could increase sectoring if the level of Cre protein produced during a typical loss-of-silencing event were limiting for recombination efficiency. That is, concurrent losses of silencing could produce higher levels of Cre, which could increase the probability of *loxP* recombination. We cannot rule out this alternative explanation, especially in light of previous studies suggesting that silencing states are a property of the cell, rather than a property of the locus (10, 11).

The Classifiers Distinguished Genotypes That Appeared Similar by Visual Inspection. Beyond uses of the feature extraction method as a visualization tool, we applied classification methods to the extracted features, including the onset frequencies of switching events and the area of GFP fluorescence, to distinguish the classes from one another. Classification, an active area of research in machine learning, has been fruitfully applied in biomedical research (12–16). Briefly, a classifier can be trained on the distribution of labeled feature data to minimize the probability of classification errors. The trained classifier can then be applied to a new sample to predict its label. To prevent differences in colony size from confounding the classification, analysis was restricted to GFP-expressing regions located within a specified distance from the colony center that was equal to the radius of the smallest colony being tested. The results presented here in terms of confusion matrices were obtained using random forest (Fig. 6*A*), an ensemble statistical learning method for classification. We also compared its performance with decision tree and AdaBoost (*SI Appendix*, Fig. S1). (See *Methods*, *Algorithm* for an explanation of these classification methods.) Each row of a confusion matrix represents the true class, and each column represents the predicted class. Hence, the (i,j) entry of the confusion matrix corresponds to the proportion of colonies of type i that got classified by our method as type j. The confusion matrices of all three methods showed that the similarity values of the various genotypes resolved into a two-block structure. One block corresponded to the haploid colonies and the other corresponded to the diploid colonies. Thus, the classifiers distinguished haploids from diploids, consistent with the observation that haploid cells exhibited higher levels of sectoring than diploid cells.

We also applied binary classification to test whether each mutant type could be differentiated from wild type (Fig. 6*B*). The classification performance reached more than 95% accuracy

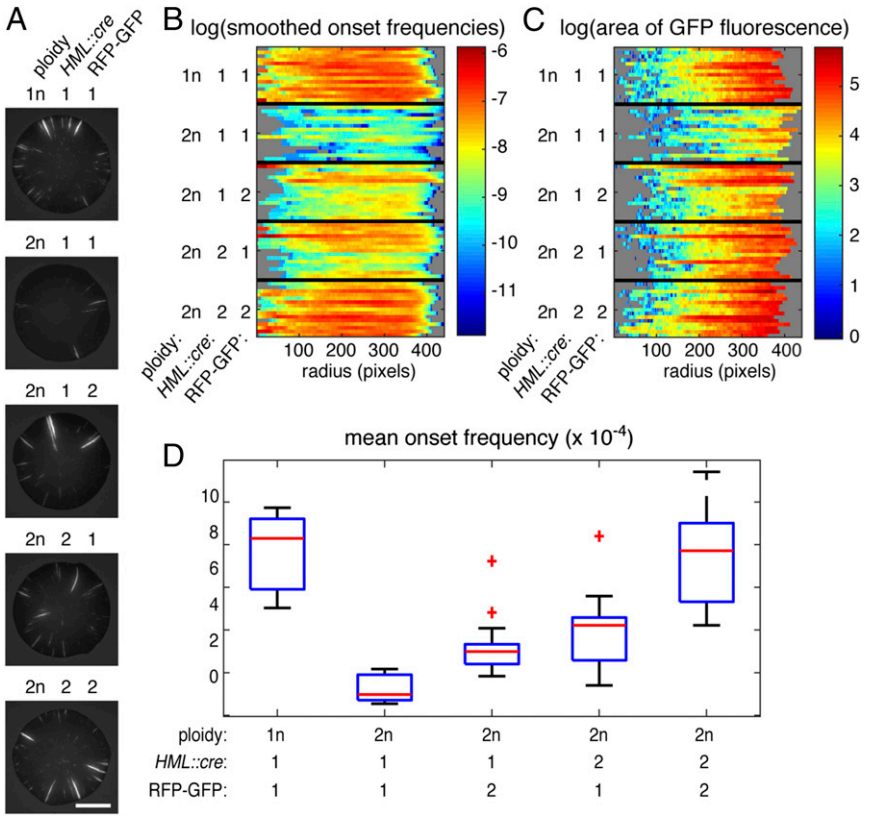

Fig. 5. Feature extraction of colonies containing various copy numbers of *HML::cre* and the RFP-GFP cassette. (*A*) GFP fluorescence of representative colonies for strains containing the specified number of chromosome sets (1n denotes haploidy and 2n denotes diploidy), *HML::cre* alleles, and RFP-GFP cassettes. (Scale bar, 2 mm.) (*B*) The smoothed onset frequencies of switching events for each genotype. The horizontal axis represents the distance from the origin in pixels, and each row represents a colony. The color bar indicates the natural logarithm of smoothed onset frequencies. (*C*) The area of GFP fluorescence. The color bar indicates the natural logarithm of the area of GFP fluorescence. (*D*) Boxplot of mean onset frequencies.

differentiating wild type from the *hst3*Δ mutant, which was the most distinct mutant. The second most distinct mutant was *hst1*Δ, with accuracy over 86%. The *hst2*Δ and *hst4*Δ mutants did not show distinct patterns from wild type. Most of the diploid mutant types could also be differentiated from the wild-type diploid, with accuracy over 75%. This included the *sir4*Δ/*SIR4* mutants, which exhibited relatively large variation in sectoring patterns (Fig. 4). The only exception was the *sir2*Δ/*SIR2* mutant, which was indistinguishable from wild type.

The Classifiers Distinguished Subtle Differences Between Colonies Grown Under Various Environmental Conditions. Treatment of various eukaryotic cells with either ascorbate, more commonly known as vitamin C, or nickel causes changes in gene expression that are thought to occur through modifications to the state of chromatin, reviewed in (17, 18). Therefore, we tested whether varying levels of either agent influenced the stability of heterochromatic repression at *HML*. Whereas visual inspection did not reliably detect a difference in the patterns of GFP expression between colonies grown in the presence versus absence of vitamin C (Fig. 7*A*), MORPHE revealed a slight reduction in onset frequency with increasing concentrations of vitamin C (Fig. 7 and *SI Appendix*, Table S5). At sublethal concentrations, nickel also caused a decrease in onset frequency (Fig. 8 and *SI Appendix*, Table S6). The classifiers distinguished each condition (0 mM, 0.05 mM, and 0.1 mM $NiCl_2$) from the other conditions with at least 80% accuracy. Consistent with the nickel-induced stabilization of silencing at *HML*, $NiCl_2$ treatment of *S. cerevisiae* cells also improves Sir-mediated silencing of a subtelomeric reporter gene (19).

Physiological differentiation during colony development leads to microenvironments within the same colony that differ in metabolite levels. One such metabolite is H_2O_2 (20, 21), a reactive oxygen species that accumulates in cells undergoing respiration. To test whether H_2O_2-induced oxidative stress affects heterochromatin, we extracted features of GFP expression in colonies grown with increasing concentrations of H_2O_2. At the highest concentration tested, H_2O_2 caused a reduction in the onset frequencies, suggesting that H_2O_2 improved the stability of silencing (Fig. 9 and *SI Appendix*, Table S7).

In the laboratory, most experiments are performed with medium containing glucose as a carbon source. In nature, however, *S. cerevisiae* encounters and metabolizes a wide variety of sugars. To determine whether alternative carbon sources affect the dynamics of heterochromatin, we compared the patterns of GFP expression between colonies grown on medium containing 2% (wt/vol) glucose and colonies grown on medium containing either 2% (wt/vol) galactose or 2% (wt/vol) raffinose. In comparison with glucose-grown colonies, colonies grown on the alternative carbon sources exhibited higher onset frequencies and therefore a destabilization of silencing at *HML* (Fig. 10 and *SI Appendix*, Table S8). In addition, silencing was slightly less stable in cells grown on raffinose than in cells grown on galactose. Collectively, these examples indicated that a variety of environmental inputs have the capacity to modify the dynamics of heterochromatin.

Discussion

The sensitivity of the CRASH assay has enabled the identification of genetic and environmental factors that contribute to the stability of heterochromatic repression. However, phenotypic

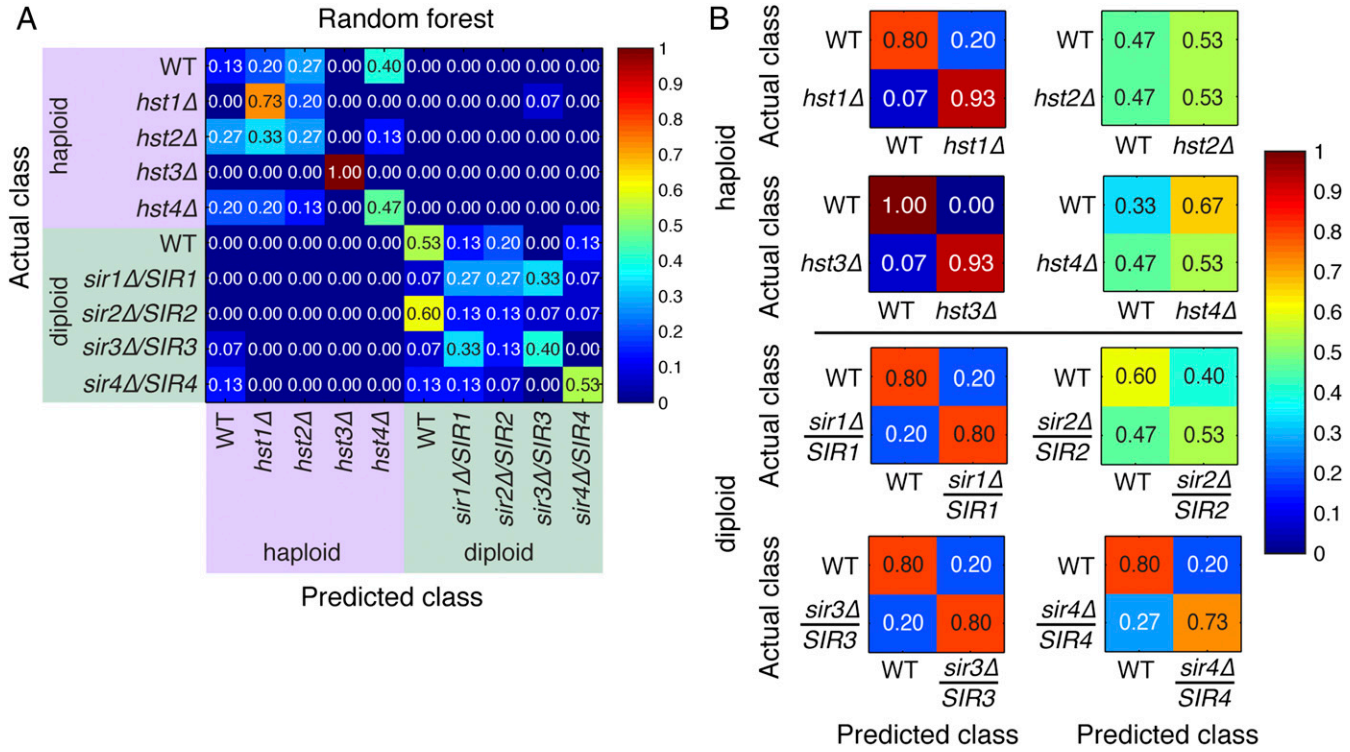

Fig. 6. Classification of genotypes based on the extracted features. (*A*) Confusion matrix by random forest on the multiclass classification of wild type and mutants, including both the haploid and diploid strains. Each row of the confusion matrix represents a different genotype (actual class), and the values within a row show the proportion of colonies that were predicted by the classifier to belong to the genotype specified by each column (predicted class). The color intensity, ranging from 0 to 1, corresponds to the fraction of colonies that were assigned to a particular predicted class. Successful classification results in high values along the diagonal, where each actual genotype intersects with its corresponding predicted genotype. (*B*) Confusion matrices by random forest on the binary classification of wild type versus each mutant. The color intensity, ranging from 0 to 1, corresponds to the fraction of colonies that were assigned to a particular predicted class.

analysis has been limited by the lack of methods available to quantify and distinguish patterns of differential GFP expression in colonies. Here, we present a robust, automated approach to extract the features of GFP expression that inform our understanding of when and how often losses of silencing occur throughout the growth of a colony. The MORPHE software suite allowed quantitative comparisons between known genotypes or conditions and also has the capacity to identify distinct patterns in colonies containing unknown mutations that could arise naturally or from random mutagenesis.

The classification algorithm was reliably able to categorize different patterns of fluorescence that were deemed indistinguishable by visual inspection. Notably, the classifier uncovered a previously unidentified role for the sirtuin Hst1 in antagonizing the stability of silencing. Hst1 is a paralog of the NAD^+-dependent histone deacetylase Sir2, an essential component of heterochromatin at *HML* and *HMR*. Interestingly, Hst1 represses the expression of genes involved in de novo synthesis of NAD^+, and deletion of *HST1* causes a slight increase in cellular levels of NAD^+ (22). One implication of this result was that NAD^+ levels may be limiting for Sir2 activity in wild-type cells, such that deletion of *HST1* would improve the capacity of Sir2 to catalyze the deacetylation reactions necessary for silencing.

The classification algorithm also successfully differentiated cells grown under various environmental conditions by comparing the frequencies of switching events and the area of the GFP-expressing regions. Vitamin C, for example, was shown to slightly increase the stability of heterochromatic repression at *HML*. Interestingly, vitamin C can stimulate in vitro activity of human and murine demethylases that target histones (23, 24), possibly through controlling the oxidation state of Fe located in the active site. Given that *S. cerevisiae* contains homologous histone demethylases (23, 25) and that the methylation status of histones affects silencing at *HML*, vitamin C may stabilize silencing through the enhancement of histone demethylation. Less-direct mechanisms are also possible, however, because vitamin C has the potential to serve as a cofactor for other related enzymes such as Jlp1 and Tpa1 of yeast (26, 27) and also functions as a reducing agent.

Given that certain environmental factors influence silencing and that colonies are environmentally complex, the stability of silencing could, in principle, fluctuate throughout different stages of colony growth or even vary between different microenvironments within the same colony. In certain genetic backgrounds, the rate of silencing loss during the initial stage of colony growth does not correspond to the overall sectoring phenotype of mature colonies (7), suggesting that processes inherent to later stages of colonial growth can modify the stability of silencing. Whereas half-sector analysis is confined to measuring silencing loss during the first cell division of colony growth, the image analysis suite presented here has the potential to quantify heterochromatin dynamics as a function of colony development.

GFP-expressing regions within colonies vary in size and shape due to multiple factors. For example, losses of silencing that occur during the early stages of colony growth have the potential to produce large sectors, whereas losses of silencing that occur during the late stages of colony growth do not. In addition, genetic drift could randomly stunt the growth of a GFP-expressing subclone. In this study, we used MORPHE to quantify the onset frequencies of all GFP-expressing regions, regardless of size; however, the MORPHE software also contains the option to analyze

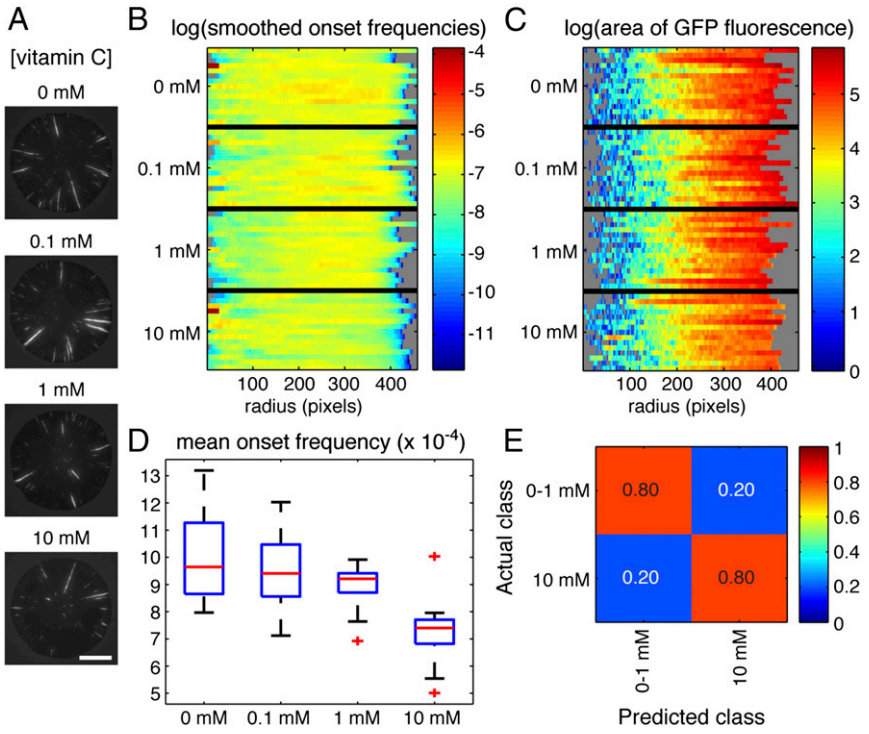

Fig. 7. Feature extraction and classification of colonies grown under various levels of vitamin C. (*A*) GFP fluorescence of representative colonies for haploid strains grown in the presence of vitamin C. (Scale bar, 2 mm.) (*B*) The smoothed onset frequencies of switching, derived by applying a sliding window average. The color bar indicates the natural logarithm of smoothed onset frequencies. (*C*) The area of GFP fluorescence. The color bar indicates the natural logarithm of the area of GFP fluorescence. (*D*) Boxplot of mean onset frequencies. (*E*) Confusion matrix by random forest on classification of colonies grown with different levels of vitamin C. The lower concentrations of vitamin C tested (0.1 mM and 1 mM) were grouped together with the colonies grown without vitamin C. The color intensity, ranging from 0 to 1, corresponds to the fraction of colonies that were assigned to a particular predicted class.

smaller regions (dots) separately from larger regions (bands, or sectors). This option could be useful for the analysis of mutants or conditions that skew the ratio of bands to dots, either through changes in the general growth pattern of colonies, or through effects on the stability of silencing that act during a specific window of colony development.

Just as MORPHE complements the limitations of half-sector analysis, half-sector analysis also complements the limitations of

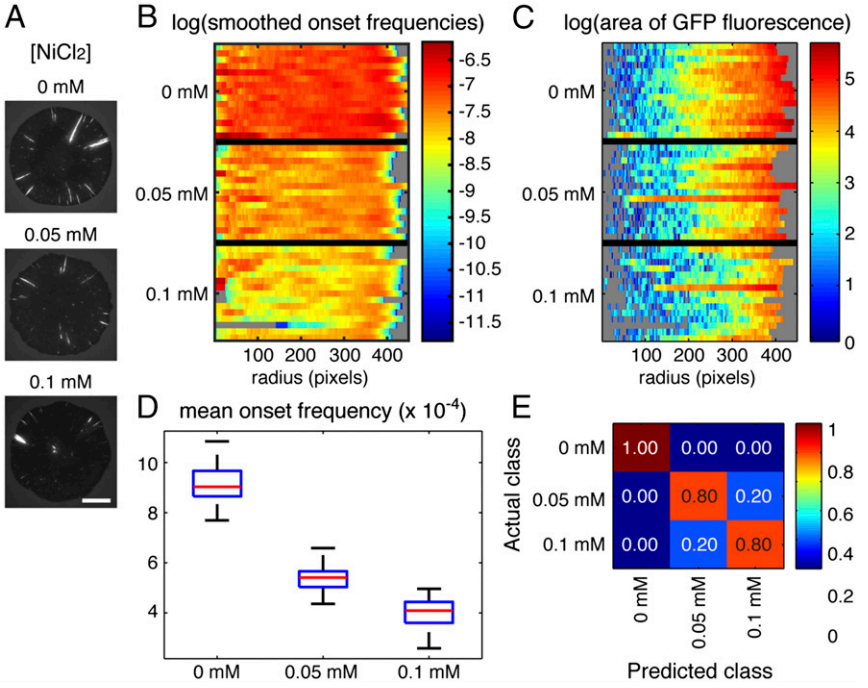

Fig. 8. Feature extraction and classification of colonies grown under various levels of $NiCl_2$. (*A*) GFP fluorescence of representative colonies for haploid strains grown in the presence of $NiCl_2$. (Scale bar, 2 mm.) (*B*) The smoothed onset frequencies of switching, derived by applying a sliding window average. The color bar indicates the natural logarithm of smoothed onset frequencies. (*C*) The area of GFP fluorescence. The color bar indicates the natural logarithm of the area of GFP fluorescence. (*D*) Boxplot of mean onset frequencies. (*E*) Confusion matrix by random forest on classification of colonies grown with the specified doses of $NiCl_2$. The color intensity, ranging from 0 to 1, corresponds to the fraction of colonies that were assigned to a particular predicted class.

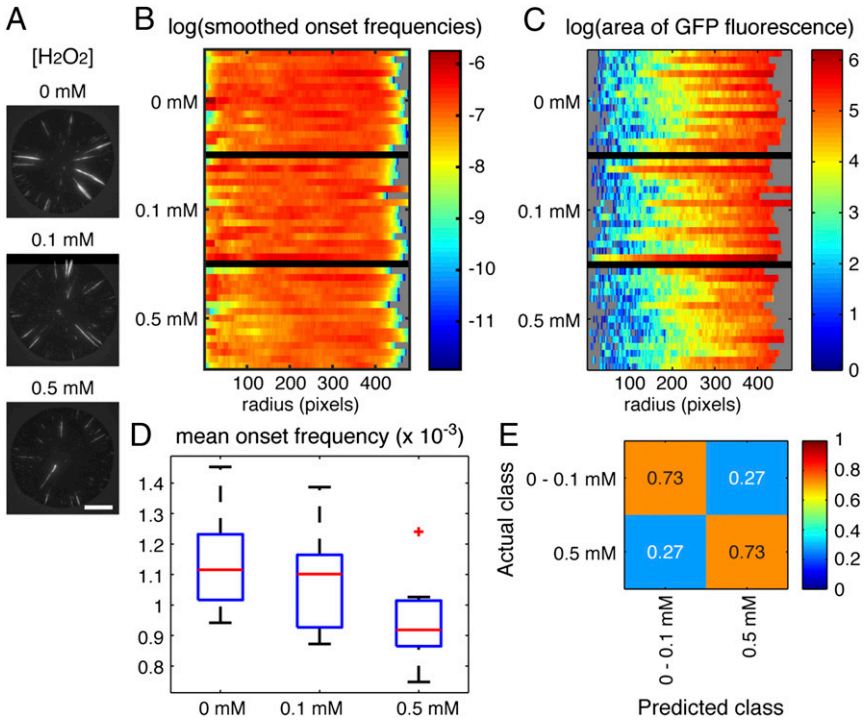

Fig. 9. Feature extraction and classification of colonies grown under various levels of H_2O_2. (*A*) GFP fluorescence of representative colonies for haploid strains grown in the presence of H_2O_2. (Scale bar, 2 mm.) (*B*) The smoothed onset frequencies of switching events, derived by applying a sliding window average. The color bar indicates the natural logarithm of smoothed onset frequencies. (*C*) The area of GFP fluorescence. The color bar indicates the natural logarithm of the area of GFP fluorescence. (*D*) Boxplot of mean onset frequencies. (*E*) Confusion matrix by random forest on classification of colonies grown with the specified doses of H_2O_2. The lowest concentration of H_2O_2 tested (0.1 mM) was grouped together with the colonies grown without H_2O_2. The color intensity, ranging from 0 to 1, corresponds to the fraction of colonies that were assigned to a particular predicted class.

MORPHE. In principle, high levels of sectoring would impair the ability of MORPHE to distinguish individual GFP-expressing regions. Given that MORPHE successfully analyzed the variety of sectoring patterns presented here, we anticipate that the software will perform well on colonies of cells that lose silencing at a rate ranging anywhere from that of wild-type diploids (3.7×10^{-4} per cell division) to that of *hst3*Δ mutants (1.1×10^{-2} per cell division) (7). For rates of silencing loss that occur above this range, however, half-sector analysis may be a more suitable method of quantification. In addition, half-sector analysis provides a measurement of the absolute number of red-to-green switches per cell division, whereas the measurements made by MORPHE are relative.

Notably, the application of MORPHE extends beyond the measurement of heterochromatin dynamics. Analyses of several other phenomena leading to differential gene expression in colonies, such as telomere position effect (28), will benefit from the quantitative method described here. Moreover, this approach may be applicable to the study of any generator of diversity within microbial colonies, from phase variation (29) and antigen switching (30) to genome rearrangements and mutagenesis.

Methods

Yeast Strains. All strains used in this study were derived from W303 and were previously described (7). See *SI Appendix*, Table S1 for a description of each genotype.

Colony Growth and Imaging. Strains were initially streaked onto solid medium containing G418 (Geneticin; Life Technologies) to select for cells expressing RFP, which were then grown to midlog phase in liquid Complete Supplement Mixture (CSM) – Trp (Sunrise Science Products) under nonselective conditions (no G418). Following 10-fold serial dilutions in 1× PBS, cells were spread onto CSM – Trp, 1% agar plates at a density of ∼10 cells per plate and were grown for 6 d at 30 °C.

The resulting colonies were imaged with a Zeiss Axio Zoom.V16 microscope equipped with ZEN software (Zeiss), a Zeiss AxioCam MRm camera, and a PlanApo Z 0.5× objective. For each experiment, the magnification and exposure times remained constant across all genotypes or conditions. Micrographs were assembled using Photoshop (Adobe Systems).

All experiments testing the effects of environment on the stability of silencing at *HML* were performed using JRY9628. For the comparison of different carbon sources, cells were grown on CSM – Trp, 1% agar plates containing either D-glucose (Fisher Scientific), D-galactose (Sigma-Aldrich), or D-raffinose pentahydrate (Sigma-Aldrich) at a concentration of 2% (wt/vol). To test the effects of other metabolites on the stability of silencing, aqueous stock solutions were first made as follows: L-ascorbic acid 2-phosphate (Sigma-Aldrich) was at 0.1 M, 30% hydrogen peroxide (BDH Chemicals) was at 0.1 M, and nickel(II) chloride hexahydrate (Sigma-Aldrich) was at 0.01 M. Each stock solution was filter-sterilized and then mixed in with freshly autoclaved CSM – Trp, 1% agar to achieve the specified concentrations.

Algorithm. We developed the analysis package MORPHE to extract informative features for characterizing the sectoring pattern resulting from heterochromatin dynamics in *S. cerevisiae*. The method was divided into five sections: (*i*) colony segmentation; (*ii*) switching events detection; (*iii*) onset detection; (*iv*) onset-frequency estimation; and (*v*) colony classification based on these features.

Colony segmentation. The first step was to segment the raw colony images. A literature review of recent segmentation methods and some of their results (*SI Appendix*, Fig. S2) on our images can be found in *SI Appendix*. Motivated by the segmentation examples in refs. 31 and 32, we developed a pipeline tailored to our application: A Canny edge detector (33) was first applied to the raw image, and then the detected edges were dilated to form a closed boundary surrounding a given colony. By restricting all subsequent analysis to this enclosed region, we could remove background noise and detect the switching events specific to the colony of interest. Our segmentation process is illustrated in *SI Appendix*, Fig. S3 and a comparison with other segmentation methods can be found in *SI Appendix*, Fig. S2. Our method was a combination of feature detection and morphological filtering, both of which are widely used for other purposes (34, 35). The underlying assumption that the Petri dish background was homogeneous compared with the object of interest (i.e., the colonies) held, and our tailored method

outperformed classic methods based on normalized cuts (36) or energy minimization (37).

Switching events detection. After segmenting the colonies, we detected the bands and dots within each colony for subsequent featurization. Separating features from the colony background by intensity thresholding was precluded by the lack of a universal intensity threshold that worked for all images, because the pixel intensity distribution varied widely between images. To eliminate the need for manual tuning, we processed the data using the aforementioned edge detection and dilation approach restricted to each colony found by the segmentation procedure described earlier. Then, for each connected component detected (Fig. 1*C*), that is, a region of pixels in which any two pixels are connected to each other by the edge detection and edge dilation, we applied the Moore-Neighbor tracing algorithm (31, 38) modified by Jacob's stopping criteria (31) to extract the boundary contour. The pixel intensities in the enclosed area of a contour were compared with those of the exterior region. If the inner pixels had intensity values larger than the exterior, we labeled the enclosed area as a feature. This step enabled us to detect the boundaries of the features of interest, overcoming the sensitivity of edge detection to the change of pixel intensities. However, the interior pixels of the features may be homogeneous and hence could be missed by edge detection. By comparing the inner and outer pixels, we successfully detected the band and dot features, as illustrated in Fig. 1*C*.

Onset detection. Next, we determined the distance from the colony centroid at which the band and dot features first appeared. (In what follows, we translated each colony centroid to the origin of a Cartesian coordinate system.) This step provided a measure of the timing of the genetic switching event that gave rise to each feature. Fig. 2*A* shows the vertex, that is, the point closest to the origin in Euclidean distance, of each connected component found by switching-event detection.

Onset-frequency estimation. Next, we estimated the frequency of feature formation (termed "onset frequency") per unit area using the learned onset features. For a fixed radius r, a naive estimate of this frequency was simply the number of observed switching events divided by the total area (excluding GFP-expressing regions) in the annulus of width Δr (Fig. 2*B*), that is, the radii of the outer ring being $r+\Delta r$ and the radii of the inner ring being r. However, because switching events are rare, this estimator had high variance, inspiring further processing of these raw onset-frequency estimates.

We applied kernel smoothing to recover the shape of the underlying probability density function from the discrete events observed. This method is widely used in time-series analysis that extracts an underlying continuous function from limited discrete events (e.g., estimating the firing rate from the train of action potentials in neuroscience) (39). We let the sequence of onset frequencies of the i_{th} colony be $\{r_{i,j}, q^f_{i,j}\}_{j=1}^{R}$, in which $r_{i,j}$ is the distance from the origin of the Cartesian coordinate system, $q^f_{i,j}$ is the corresponding onset frequency of switching events, and R is the radius of the colony in terms of number of pixels. Then, the estimated smoothed onset frequency function of switching events was

$$s^f_i(r) = \frac{\sum_{j=1}^{R} k_\lambda(r, r_{i,j})\, q^f_{i,j}}{\sum_{j=1}^{R} k_\lambda(r, r_{i,j})}.$$

The kernel k_λ has a window parameter λ. Fig. 2*C* shows the sequence of onset frequencies, along with the smoothed functions by applying kernel smoothing. We adopted the sliding-window averaging method as the kernel function with a window size of $\lambda = 50$ pixels for illustration.

Colony classification. In addition to the onset of switching events, we also introduced the area of GFP fluorescence, denoted as $s^a_i(r)$, as an additional feature. These derived features provided an efficient way of visualizing heterochromatin dynamics and could also be used as discriminative features in classification, which could provide insights into the different switching patterns that were not obvious by visual inspection. We let the extracted features and labels be $\{\boldsymbol{x}_i, y_i\}_{i=1}^{N}$, where $\boldsymbol{x}_i = [\boldsymbol{s}_i(r_1), \boldsymbol{s}_i(r_2), \ldots, \boldsymbol{s}_i(r_R)]$ and $\boldsymbol{s}_i(r) = [s^f_i(r),\ s^a_i(r)]$. That is, $\boldsymbol{x}_i \in \mathbb{R}^{1\times 2R}$ is a vector consisting of the smoothed onset frequencies of switching events and area of GFP fluorescence evaluated at R radii and $y_i \in \{1,2,\ldots,K\}$ represents the class label, which could be the specific yeast strain or the environmental condition. For convenience, we denoted $\boldsymbol{X} = [\boldsymbol{x}_1; \boldsymbol{x}_2; \ldots; \boldsymbol{x}_N]$ and $\boldsymbol{Y} = [y_1; y_2; \ldots; y_N]$. In this paper, we presented the performance of decision trees and ensemble learning methods, including AdaBoost and random forest. The decision tree is a greedy algorithm that recruits splitting nodes to the tree by defining half planes $P_1 = \{z_j | z_j \le s\}$ and $P_2 = \{z_j | z_j > s\}$, in which z_j is the splitting variable from one of the features and s is the splitting point. At each candidate node, one computes an error measure such as the misclassification rate, Gini index, or the cross-entropy (40). Then the splitting nodes are selected and added sequentially to improve performance. Each terminal node (or a leaf) represents a class label and provides a prediction if one follows the path from the root to that leaf, and a decision is made by taking the majority vote among the leaf nodes.

AdaBoost (41) and random forest (42) are common ensemble methods, in which a committee is formed by combining the outputs of many weak learners (40), and the committee performs a majority vote to decide the predicted class labels. The details of these algorithms are described in *SI Appendix*. All of the classification results were obtained with the leave-one-out test, and

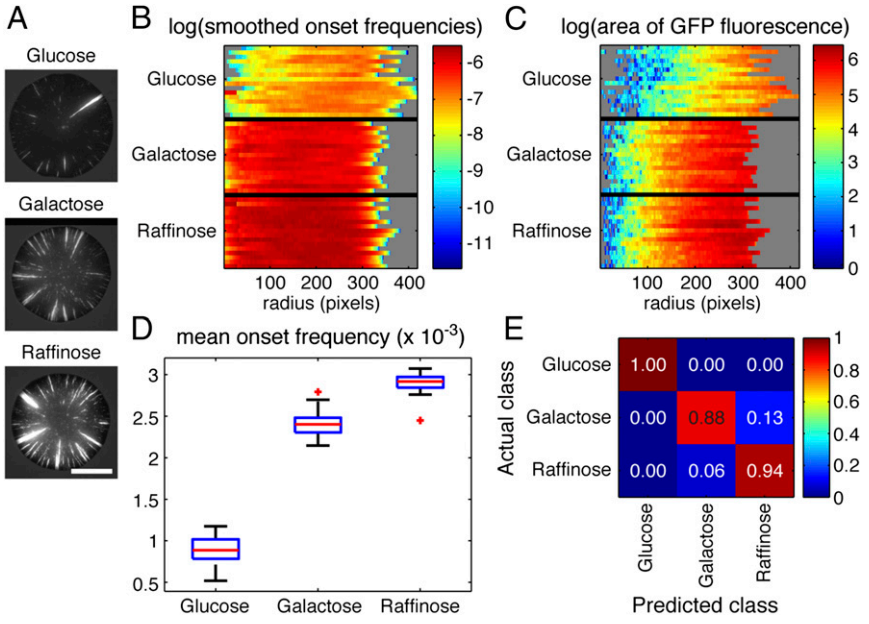

Fig. 10. Feature extraction and classification of colonies grown with different sugars. (*A*) GFP fluorescence of representative colonies for haploid strains grown in the presence of the indicated carbon sources. (Scale bar, 2 mm.) (*B*) The smoothed onset frequencies of switching events, derived by applying a sliding window average. The color bar indicates the natural logarithm of smoothed onset frequencies. (*C*) The area of GFP fluorescence. The color bar indicates the natural logarithm of the area of GFP fluorescence. (*D*) Boxplot of mean onset frequencies. (*E*) Confusion matrices by random forest on classification of colonies grown with the specified sugar supply. The color intensity, ranging from 0 to 1, corresponds to the fraction of colonies that were assigned to a particular predicted class.

we applied stratified sampling such that the number of samples for each class was the same in the training set. The AdaBoost or AdaBoost.M2 for multiclass was applied with 200 learning cycles; the random forest used the square root of the total number of features at random at each split, and 200 trees were generated. When the colonies had different sizes, we limited all of the samples to the smallest size observed.

Software Availability. Our software MORPHE, which stands for MORphological PHenotype Extraction, is freely available at https://sourceforge.net/projects/morphe. The implementation includes a graphical user interface (*SI Appendix*, Figs. S4 and S5). The details can be found in *SI Appendix*.

ACKNOWLEDGMENTS. We thank Ryan Janke and Gavin Schlissel for insightful discussion and the Hallatschek laboratory for the use of their microscope. This work was supported by National Science Foundation (NSF) CAREER Grant DBI-0846015, a Packard Fellowship for Science and Engineering, and a Math+X Research Grant from the Simons Foundation (to Y.S.S.) and by Grant GM 31105 from the National Institutes of Health (to J.R.) and NSF Graduate Research Fellowship DGE 1106400 (to A.E.D.).

1. Meunier JR, Choder M (1999) Saccharomyces cerevisiae colony growth and ageing: Biphasic growth accompanied by changes in gene expression. *Yeast* 15(12): 1159–1169.
2. Váchová L, Kucerová H, Devaux F, Ulehlová M, Palková Z (2009) Metabolic diversification of cells during the development of yeast colonies. *Environ Microbiol* 11(2):494–504.
3. Shapiro JA (1984) The use of Mudlac transposons as tools for vital staining to visualize clonal and non-clonal patterns of organization in bacterial growth on agar surfaces. *J Gen Microbiol* 130(5):1169–1181.
4. Gottlieb S, Esposito RE (1989) A new role for a yeast transcriptional silencer gene, SIR2, in regulation of recombination in ribosomal DNA. *Cell* 56(5):771–776.
5. Grunstein M, Gasser SM (2013) Epigenetics in Saccharomyces cerevisiae. *Cold Spring Harb Perspect Biol* 5(7):1–28.
6. Rine J, Herskowitz I (1987) Four genes responsible for a position effect on expression from HML and HMR in Saccharomyces cerevisiae. *Genetics* 116(1):9–22.
7. Dodson AE, Rine J (2015) Heritable capture of heterochromatin dynamics in Saccharomyces cerevisiae. *eLife* 4:e05007.
8. Hieter P, Mann C, Snyder M, Davis RW (1985) Mitotic stability of yeast chromosomes: A colony color assay that measures nondisjunction and chromosome loss. *Cell* 40(2): 381–392.
9. Xu EY, Zawadzki KA, Broach JR (2006) Single-cell observations reveal intermediate transcriptional silencing states. *Mol Cell* 23(2):219–229.
10. Pillus L, Rine J (1989) Epigenetic inheritance of transcriptional states in S. cerevisiae. *Cell* 59(4):637–647.
11. Sussel L, Vannier D, Shore D (1993) Epigenetic switching of transcriptional states: cis- and trans-acting factors affecting establishment of silencing at the HMR locus in Saccharomyces cerevisiae. *Mol Cell Biol* 13(7):3919–3928.
12. Chen X, Zhou X, Wong STC (2006) Automated segmentation, classification, and tracking of cancer cell nuclei in time-lapse microscopy. *IEEE Trans Biomed Eng* 53(4): 762–766.
13. Segal E, et al. (2007) Decoding global gene expression programs in liver cancer by noninvasive imaging. *Nat Biotechnol* 25(6):675–680.
14. Held M, et al. (2010) CellCognition: Time-resolved phenotype annotation in high-throughput live cell imaging. *Nat Methods* 7(9):747–754.
15. Jones TR, et al. (2008) CellProfiler Analyst: Data exploration and analysis software for complex image-based screens. *BMC Bioinformatics* 9:482.
16. Sommer C, Straehle C, Kothe U, Hamprecht FA (2011) Ilastik: Interactive learning and segmentation toolkit. *Proceedings of the International Symposium on Biomedical Imaging* (IEEE, Piscataway, NJ), pp 230–233.
17. Sun H, Shamy M, Costa M (2013) Nickel and epigenetic gene silencing. *Genes (Basel)* 4(4):583–595.
18. Young JI, Züchner S, Wang G (2015) Regulation of the epigenome by vitamin C. *Annu Rev Nutr* 35:545–564.
19. Broday L, Cai J, Costa M (1999) Nickel enhances telomeric silencing in Saccharomyces cerevisiae. *Mutat Res* 440(2):121–130.
20. Váchová L, Palková Z (2005) Physiological regulation of yeast cell death in multicellular colonies is triggered by ammonia. *J Cell Biol* 169(5):711–717.
21. Cáp M, Váchová L, Palková Z (2009) Yeast colony survival depends on metabolic adaptation and cell differentiation rather than on stress defense. *J Biol Chem* 284(47): 32572–32581.
22. Bedalov A, Hirao M, Posakony J, Nelson M, Simon JA (2003) NAD+-dependent deacetylase Hst1p controls biosynthesis and cellular NAD+ levels in Saccharomyces cerevisiae. *Mol Cell Biol* 23(19):7044–7054.
23. Tsukada Y, et al. (2006) Histone demethylation by a family of JmjC domain-containing proteins. *Nature* 439(7078):811–816.
24. Wang T, et al. (2011) The histone demethylases Jhdm1a/1b enhance somatic cell reprogramming in a vitamin-C-dependent manner. *Cell Stem Cell* 9(6):575–587.
25. Tu S, et al. (2007) Identification of histone demethylases in Saccharomyces cerevisiae. *J Biol Chem* 282(19):14262–14271.
26. Hogan DA, Auchtung TA, Hausinger RP (1999) Cloning and characterization of a sulfonate/alpha-ketoglutarate dioxygenase from Saccharomyces cerevisiae. *J Bacteriol* 181(18):5876–5879.
27. Kim HS, et al. (2010) Crystal structure of Tpa1 from Saccharomyces cerevisiae, a component of the messenger ribonucleoprotein complex. *Nucleic Acids Res* 38(6): 2099–2110.
28. Gottschling DE, Aparicio OM, Billington BL, Zakian VA (1990) Position effect at S. cerevisiae telomeres: Reversible repression of Pol II transcription. *Cell* 63(4):751–762.
29. Henderson IR, Owen P, Nataro JP (1999) Molecular switches–the ON and OFF of bacterial phase variation. *Mol Microbiol* 33(5):919–932.
30. Vink C, Rudenko G, Seifert HS (2012) Microbial antigenic variation mediated by homologous DNA recombination. *FEMS Microbiol Rev* 36(5):917–948.
31. Gonzalez R, Woods R, Eddins S (2004) *Digital Image Processing Using MATLAB* (Pearson Education, Upper Saddle River, NJ).
32. Mathworks (2015) MATLAB and Image Processing Toolbox Release 2015a (The MathWorks, Inc., Natick, MA).
33. Canny J (1986) A computational approach to edge detection. *IEEE Trans Pattern Anal Mach Intell* 8(6):679–698.
34. Meijering E (2012) Cell Segmentation: 50 Years Down the Road. *IEEE Signal Process Mag* 29(5):140–145.
35. Pham DL, Xu C, Prince JL (2000) Current methods in medical image segmentation. *Annu Rev Biomed Eng* 2:315–337.
36. Shi J, Malik J (2000) Normalized cuts and image segmentation. *IEEE Trans Pattern Anal Mach Intell* 22(8):888–905.
37. Delong A, Osokin A, Isack HN, Boykov Y (2012) Fast approximate energy minimization with label costs. *Int J Comput Vis* 96(1):1–27.
38. Moore E (1962) Machine models of self-reproduction. *Mathematical Problems in the Biological Sciences*, ed Bellman R (Am Mathematical Soc, Providence, RI), pp 17–33.
39. Dayan P, Abbott LF (2001) *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems* (MIT Press, Cambridge, MA).
40. Hastie T, Tibshirani R, Friedman J (2003) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer, New York).
41. Schapire RE (1999) A short introduction to boosting. *IJCAI International Joint Conference on Artificial Intelligence* (Morgan Kauffman, San Francisco), Vol 2, pp 1401–1406.
42. Breiman L (2001) Random forests. *Mach Learn* 45(1):5–32.