Received: 19 July 2020 | Revised: 2 December 2020 | Accepted: 21 January 2021

DOI: 10.1002/bit.27698

ARTICLE

BIOTECHNOLOGY BIOENGINEERING WILEY

Modeling of copy number variability in *Pichia pastoris*

Amos E. Lu[1] | Andrew J. Maloney[1] | Neil C. Dalvie[1,2] | Joseph R. Brady[1,2] | Kerry R. Love[1,2] | J. Christopher Love[1,2] | Richard D. Braatz[1]

[1]Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

[2]Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

Correspondence
Richard D. Braatz, Department of Chemical Engineering, Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, MA 02139-4307, USA.
Email: braatz@mit.edu

Funding information
National Science Foundation, Grant/Award Number: 1122374; Defense Advanced Research Projects Agency and Space and Naval Warfare Systems Command, Grant/Award Number: N66001-13-C-4025

Abstract

Development of continuous biopharmaceutical manufacturing processes is an area of active research. This study considers the long-term transgene copy number stability of *Pichia pastoris* in continuous bioreactors. We propose a model of copy number loss that quantifies population heterogeneity. An analytical solution is derived and compared with existing experimental data. The model is then used to provide guidance for stable operating timescales. The model is extended to consider copy number dependent growth such as in the case of Zeocin supplementation. The model is also extended to analyze a continuous seeding strategy. This study is a critical step towards understanding the impact of continuous processing on the stability of *Pichia pastoris* and the resultant products.

KEYWORDS

continuous biomanufacturing, copy number stability, mechanistic modeling, *Pichia pastoris*, recombinant protein production

1 | INTRODUCTION

Continuous manufacturing of biologics has received significant attention in recent years (Crowell et al., 2018), as a way to reduce costs, increase flexibility, ease scale up, and increase product quality (Konstantinov & Cooney, 2015). However, concerns surrounding the genomic stability of the host (Werner et al., 1992) must be addressed in the extended production periods typically used in continuous culture.

Pichia pastoris (also referred to as *Komagataella phaffii*) is a methylotrophic yeast that has been used to produce a large number of therapeutic proteins (Ahmad et al., 2014) due to its fast growth rate, ability to grow to high cell densities, strongly regulated methanol-inducible AOX1 promoter, and low host cell protein secretion profile (Damasceno et al., 2012; Invitrogen Corporation, 2002). However, transgene copies are unstable in *P. pastoris* (Aw, 2012; Aw & Polizzi, 2013; Curvers et al., 2001; Ohi et al., 1998; Zhu et al., 2009) and have a direct impact on specific productivity (Cos et al., 2005; Marx et al., 2009), which is in turn related to product quality (Cunha et al., 2004; Schenk et al., 2008). While genomic stability in the form of plasmids have been modeled in *Eschericia coli* and *Saccharomyces cerevisiae* (Werbowy et al., 2017), a model for chromosomally integrated DNA does not yet exist for *P. pastoris*.

Here, we propose a mathematical model for analyzing the genomic stability of chromosomally integrated transgene copies in *P. pastoris* by assuming expression cassettes are looped out from tandem arrays (Aw, 2012; Aw & Polizzi, 2013; Schwarzhans et al., 2016). The model is then validated by experimental data in the literature. We then use the model to derive guidelines for stable process operation. Next, the model is expanded to account for antibiotic selection pressure with Zeocin as a case study. We also expand the model to consider continuous seeding as a method to stabilize the production bioreactor.

2 | METHODS

Data for average copy number (ACN) were extracted from figures in the literature (Aw, 2012; Zhu et al., 2009) using WebPlotDigitizer.

Amos E. Lu and Andrew J. Maloney contributed equally to this study.

Differential equations were solved using Matlab R2018a. Model parameters were estimated using least-squares regression.

Samples were collected from bioreactors expressing granulocyte colony stimulating factor (G-CSF)**** (Crowell et al., 2018) at 2 h pre- and 46, 118, and 160 h post-induction. Variant calling was carried out using the GATK Best Practices workflow (Van der Auwera et al., 2013). Alignments were performed using Burrows-Wheeler Aligner (BWA-MEM, version 0.7.5a) (Li & Durbin, 2009). Illumina sequencing reads were mapped to the wildtype (ATCC76473) *K. phaffii* genome (NCBI bioproject accession number PRJNA304977). PCR duplicates were removed using Picard (version 1.94) MarkDuplicates after sorting the sequences using SortSam. Samtools (version 0.1.19) was used for the first round of SNP and Indel calling. These high quality Indels and SNPs were then selected as the input for GATK Best Practice Indel local realignment and base quality recalibration steps (version 3.1.1). Transgene copy number was calculated as the ratio of average cassette coverage to average genome coverage, correcting for the presence of native *K. phaffii* sequence elements in the cassette. Average depth was calculated using Samtools.

3 | RESULTS AND DISCUSSION

Biomass accumulation is conventionally modeled by

$$\frac{dX(t)}{dt} = \mu(S)X(t), \tag{1}$$

where X is the total biomass concentration and the cellular growth rate μ is a function of the limiting substrate concentration S (methanol).

For biomass composed of subpopulations with different transgene copy number, we can analogously write

$$\frac{d}{dt}\mathbf{x}(t) = \mu(S)\mathbf{x}(t), \tag{2}$$

where $\mathbf{x}$ is a $n+1$ vector biomass concentration of each subpopulation (i.e., x_i is associated with the subpopulation with copy number i). Instability is introduced through a mass-conserving matrix A that describes the first-order transitions between subpopulations,

$$\frac{d}{dt}\mathbf{x}(t) = \mu(S)\mathbf{x}(t) + \mathbf{A}\mathbf{x}(t), \tag{3}$$

with a mass conservation condition

$$\sum_{i=0}^{n} A_{i,j} = 0, \quad \forall\, j \tag{4}$$

to ensure that the subpopulation model is equivalent to the total biomass model under summation.

Writing $\mathbf{y}$, the fraction of total biomass at each copy number, as[1]

$$\mathbf{y}(t) = \frac{1}{X_T(t)}\mathbf{x}(t) \tag{5}$$

eliminates the dependence on substrate concentration and allows the dynamics of $\mathbf{y}$ to be solely dependent on the matrix $\mathbf{A}$,

$$\frac{d}{dt}\mathbf{y}(t) = \mathbf{A}\mathbf{y}(t). \tag{6}$$

Note that Eequation (6) is independent of the growth rate expression in Equation (1), which is useful as the cellular growth rate can depend on a multitude of factors, all of which are collapsed into the value of α. Assuming that contiguous transgene copies are lost in a loop-out event with a rate constant α (Figure 1),

$$A_{i,j} = \begin{cases} \alpha(j+1), & \text{for} \quad j < i, \\ -\alpha\frac{i(i+1)}{2}, & \text{for} \quad j = i, \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

The $i = j$ condition is equivalent to the sum of integers from 1 to n. The model structure assumes that copies are lost at a rate proportional to the number of ways that contiguous loss could occur (Figure 1). For instance, a cell with four copies can transition from four to three copies in four distinct ways (loss of copy 1, 2, 3, or 4). That same cell can transition from four to two copies in three distinct ways (loss of copies 1 and 2, 2 and 3, or 3 and 4, respectively). Similarly, that cell can transition from four to one copy in two distinct ways (loss of copies 1, 2, and 3 or 2, 3, and 4) and transition from four to zero copies in one way (loss of all copies).

The scalar α is likely to be affected by integration locus, bioreactor operating conditions, and the recombinant gene expressed. The value of α is difficult to predict a priori, and is most readily obtained through parameter estimation.

The vector $\mathbf{y}(t)$ can be solved through eigendecomposition,

$$\mathbf{y}(t) = \mathbf{V}\exp(\alpha t\mathbf{D})\mathbf{V}^{-1}\mathbf{y}_0, \tag{8}$$

where

$$V_{i,j} = \begin{cases} 1, & \text{for} \quad i = j, \\ -1, & \text{for} \quad i = j-1, \\ 0, & \text{otherwise,} \end{cases} \tag{9}$$

$$D_{i,j} = \begin{cases} -\frac{i(i+1)}{2}, & \text{for} \quad i = j, \\ 0, & \text{for} \quad i \neq j, \end{cases} \tag{10}$$

$$V^{-1}_{i,j} = \begin{cases} 1, & \text{for} \quad i \leq j, \\ 0, & \text{otherwise,} \end{cases} \tag{11}$$

and $\mathbf{y}_0 = \mathbf{y}(0)$ is the initial fraction of total biomass at each copy number.

[1]Please see supporting information for the mathematical derivation.

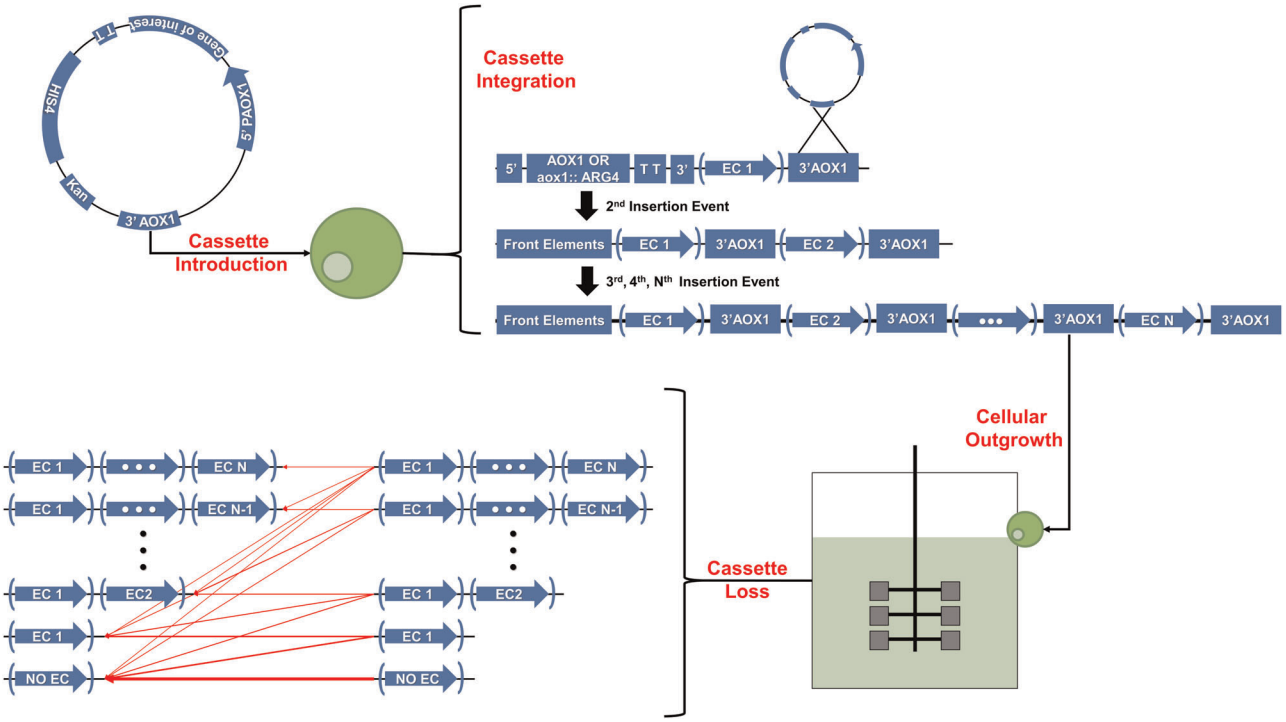

FIGURE 1 Proposed mechanism for copy number generation and loss. Cells gain copy numbers upon cassette introduction and lose copies through loop-out events within bioreactor cultivation. The loss structure is defined in equation 7 of the main text, and the relative rate of loss is dependent on a single parameter, α

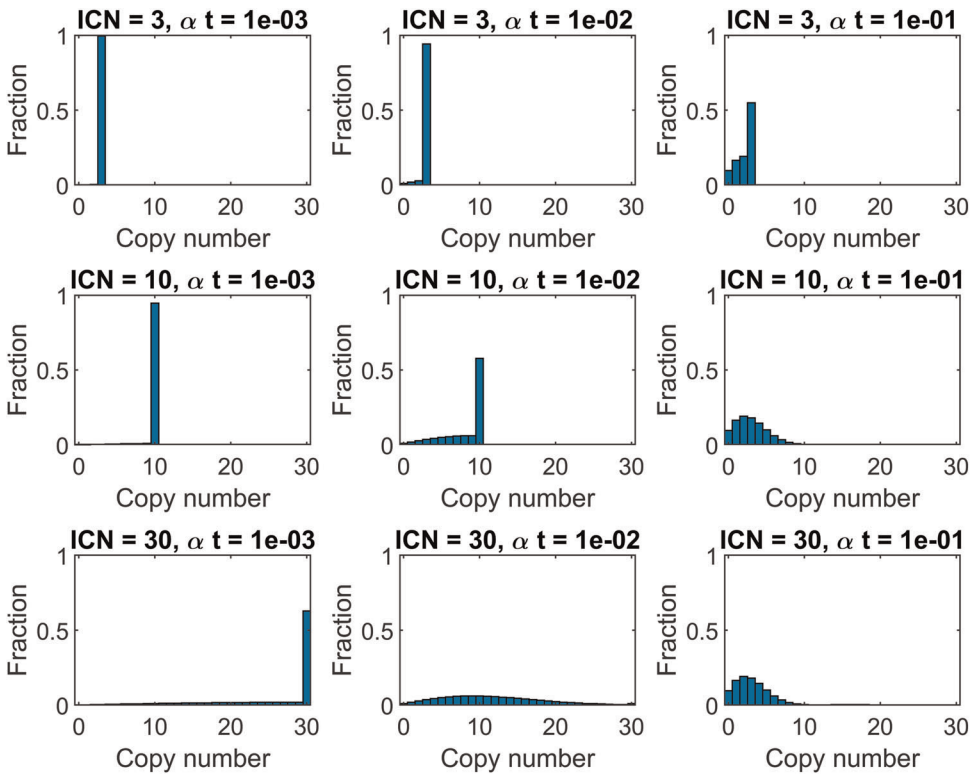

FIGURE 2 Copy number distributions starting from initial copy numbers (ICN) of 3, 10, and 30, at various dimensionless time αt. Low ICN tends to be more stable, taking a longer dimensionless time to decay [Color figure can be viewed at wileyonlinelibrary.com]

By assuming that inoculum is comprised of cells at the highest copy number, $y_n = 1$, the fraction of cells of copy i, y_i, and ACN can be found in terms of a dimensionless time αt,

$$y_i = \begin{cases} \exp\left(-\frac{i(i+1)}{2}\alpha t\right) - \exp\left(-\frac{(i+1)(i+2)}{2}\alpha t\right), & \text{for} \quad i \leq n, \\ \exp\left(-\frac{i(i+1)}{2}\alpha t\right), & \text{for} \quad i = n, \end{cases} \tag{12}$$

$$\text{ACN} = \sum_{i=0}^{n} i y_i, \tag{13}$$

$$= \sum_{i=1}^{n} \exp\left(-\frac{i(i+1)}{2}\alpha t\right). \tag{14}$$

The model is qualitatively consistent with the literature, exhibiting increased stability when starting with lower copy numbers (Figure 2). Quantitatively, the model matches the literature data (Aw, 2012; Zhu et al., 2009) closely with only a single fitting parameter α for each strain (Figure 3). Due to the inability to spontaneously generate additional copies (i.e., **A** is lower triangular), the ACN asymptotically tends toward zero (Figure 4). The behavior of a strain can be condensed into the rate constant α to aid strain selection and process development and characterization.

In biomanufacturing, it is desirable that ACN remains approximately constant across the production window, which can be described as the inequality

$$\frac{\text{ACN}}{\text{ICN}} > \beta, \tag{15}$$

where β is a fraction close to 1. The model can then be used to predict the time at which the ACN decays below the threshold given (Figure 5), which can aid process design.

This model can be extended to account for copy number dependent growth, such as under antibiotic selection pressure or copy-number dependent auxotrophy. This system can be described as

$$\frac{d}{dt}x_i(t) = \mu(S)x_i(t) + \sum_{j=0}^{n} A_{i,j}x_j(t) + \gamma i x_i(t), \tag{16}$$

where the scalar γ is the growth rate change per copy. Antibiotic selection pressure would result in a positive (i.e., copy number dependent) γ. Similarly as before, re-expressing this equation in terms of fractional biomass **y** gives

$$\frac{d}{dt}y_i = \sum_{j=0}^{n} A_{i,j}y_j + \gamma i y_i - \gamma y_i \sum_{j=0}^{n} j y_j. \tag{17}$$

The steady-state distribution of copies can be determined by setting the time derivative to be zero, to give

$$0 = \sum_{j=0}^{n} A_{i,j}y_j + \gamma i y_i - \gamma y_i \sum_{j=0}^{n} j y_j. \tag{18}$$

which can be solved to obtain[2]

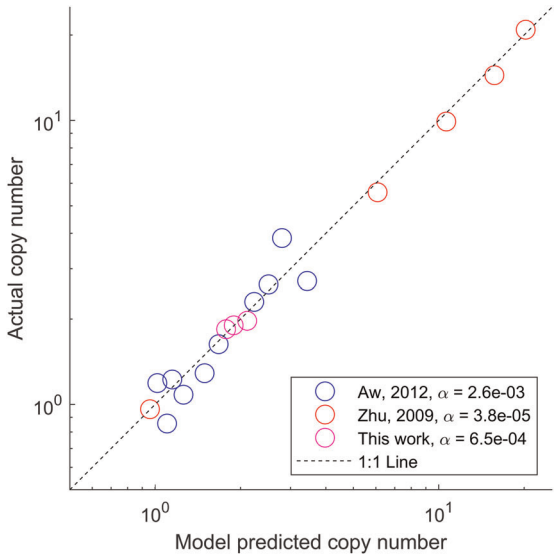

FIGURE 3 Comparison of copy number model predictions with literature and experimental data (Aw, 2012; Zhu et al., 2009), showing excellent agreement [Color figure can be viewed at wileyonlinelibrary.com]

$$\text{ACN} = \text{ICN}\left(1 - \frac{(i+1)\alpha}{2\gamma}\right). \tag{19}$$

Stability, as defined by Equation (15), is a function of the dimensionless group $\frac{\gamma}{\alpha}$ and the initial copy number (Figure 6),

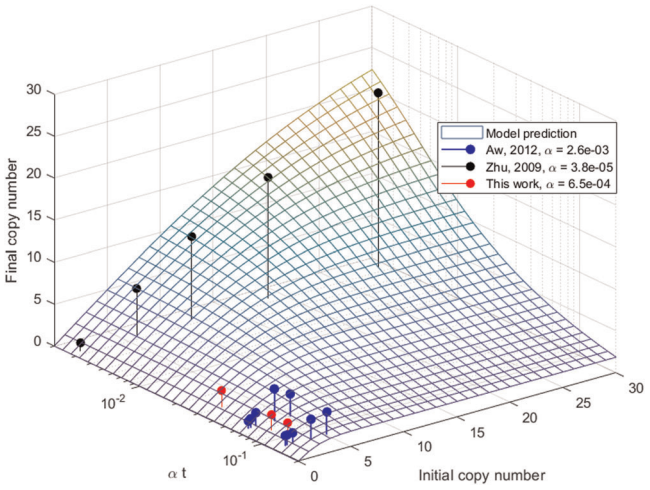

FIGURE 4 Surface plot of average copy number (ACN) as a function of initial copy number and dimensionless time αt. Overlaid on the surface are data from experiments and the literature. The copy number loss rate constant α was obtained by parameter estimation within the data set for each literature source. Under the loss rate of Zhu et al. (2009), $\alpha t = 10^{-1}$, 10^{-2}, and 10^{-3} corresponds to a process time of 16 weeks, 11 days, and 26 h, respectively. The shape of the contour shows that higher initial copy numbers decay rapidly, and that the final average copy number tends toward zero [Color figure can be viewed at wileyonlinelibrary.com]

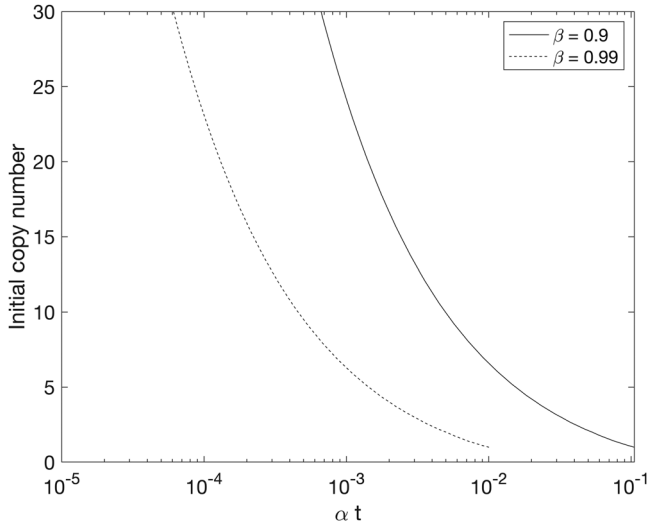

FIGURE 5 Dimensionless time for average copy number to decrease to a fraction β (defined in Equation 5 in text) of the initial copy. Under the loss rate of Zhu et al. (2009), $\alpha t = 10^{-1}, 10^{-2}$, and 10^{-3} corresponds to a process time of 16 weeks, 11 days, and 26 h, respectively

$$\frac{\mathrm{ICN}+1}{2}\frac{\alpha}{\gamma} < 1-\beta. \tag{20}$$

This model can be used to estimate γ based on experimental data. For instance, using data for copy loss in *P. pastoris* under 2000 μg/ml Zeocin selection pressure (Aw, 2012) and choosing γ to minimize the sum-of-square errors between model prediction and data, a low value of $\gamma = 3.4 \times 10^{-2}\mathrm{h}^{-1}$ and $\frac{\gamma}{\alpha} = 13.4$ is obtained. In a similar fashion, this model can be applied to other sources of selection pressure to evaluate the potential for culture stabilization.

Another approach to stability extends the concept of seed trains (Frahm, 2014) by continuously injecting high-copy cells from a continuous glycerol-fed seed reactor into the production bioreactor to replenish lost copies (Figure 7). The seed reactor is in turn stable, because copy loss does not occur under glycerol (Zhu et al., 2009).

Leveraging a bioreactor material balance as shown in chapter 9 of Bailey and Ollis (2006), the model can be rewritten as

$$\frac{d}{dt}\mathbf{x}(t) = \mu(S)\mathbf{x}(t) + \mathbf{A}\mathbf{x}(t) + \frac{q_{\mathrm{in}}}{V}\mathbf{x}_{\mathrm{in}}(t). \tag{21}$$

where V is the cultivation volume and q_{in} and $\mathbf{x}_{\mathrm{in}}$ are the flowrate and the vector of biomass concentration of each subpopulation in the seed stream, respectively.

If the injected biomass is assumed to have the highest copy number in the system,

$$x_{\mathrm{in},i} = \begin{cases} X_{\mathrm{seed}}, & \text{for} \quad i = n, \\ 0, & \text{otherwise}, \end{cases} \tag{22}$$

[2]Please see supporting information for the mathematical derivation.

while maintaining a constant seeding ratio $\frac{q_{\mathrm{in}}X_{\mathrm{seed}}}{VX_T} = \phi$, we can use the fractional form to derive

$$\frac{d}{dt}\mathbf{y}(t) = \mathbf{A}\mathbf{y}(t) + \frac{q_{\mathrm{in}}}{VX_T}\mathbf{x}_{\mathrm{in}}(t) - \frac{q_{\mathrm{in}}}{VX_T}\sum_{i=0}^{n} x_{\mathrm{in},i}\mathbf{y}(t), \tag{23}$$

followed by eigendecomposition to obtain

$$\mathbf{y}(t) = \mathbf{V}\exp(y\mathbf{D})\mathbf{V}^{-1}(\mathbf{y}_0 - \mathbf{y}_f) + \mathbf{y}_f, \tag{24}$$

where the matrices $\mathbf{V}$ and $\mathbf{V}^{-1}$ are in Equations (9) and (11), respectively, and

$$D_{i,i} = \begin{cases} -\alpha\frac{i(i+1)}{2} - \phi, & \text{for} \quad i = j, \\ 0, & \text{for} \quad i \neq j, \end{cases} \tag{25}$$

$$\mathbf{y}_f = \begin{cases} \frac{2\phi/\alpha}{(i)(i+1)+2\phi/\alpha} - \frac{2\phi/\alpha}{(i+1)(i+2)+2\phi/\alpha}, & \text{for} \quad i < n, \\ \frac{2\phi/\alpha}{(i)(i+1)+2\phi/\alpha}, & \text{for} \quad i = n. \end{cases} \tag{26}$$

The ACN is given by

$$\mathrm{ACN}(t) = \sum_{i=1}^{n}\exp\left[\left(-\alpha\frac{i(i+1)}{2} - \phi\right)t\right]\left[1 - \frac{2\phi/\alpha}{i(i+1)+2\phi/\alpha}\right] + \frac{2\phi/\alpha}{i(i+1)+2\phi/\alpha} \tag{27}$$

and the steady-state ACN is

$$\mathrm{ACN} = \sum_{i=1}^{n}\frac{2\phi/\alpha}{i(i+1)+2\phi/\alpha}. \tag{28}$$

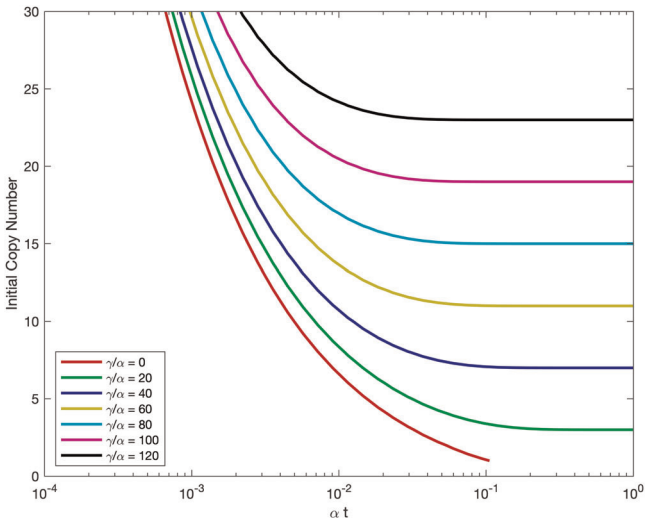

FIGURE 6 Constant $\beta = 0.9$ (defined in equation 5 in text) contours for various γ, αt, and initial copy numbers. Increasing γ increases the value of the initial copy number at infinite time. Under the loss rate of Zhu et al. (2009), $\alpha t = 10^{-1}, 10^{-2}$, and 10^{-3} corresponds to a process time of 16 weeks, 11 days, and 26 h, respectively [Color figure can be viewed at wileyonlinelibrary.com]

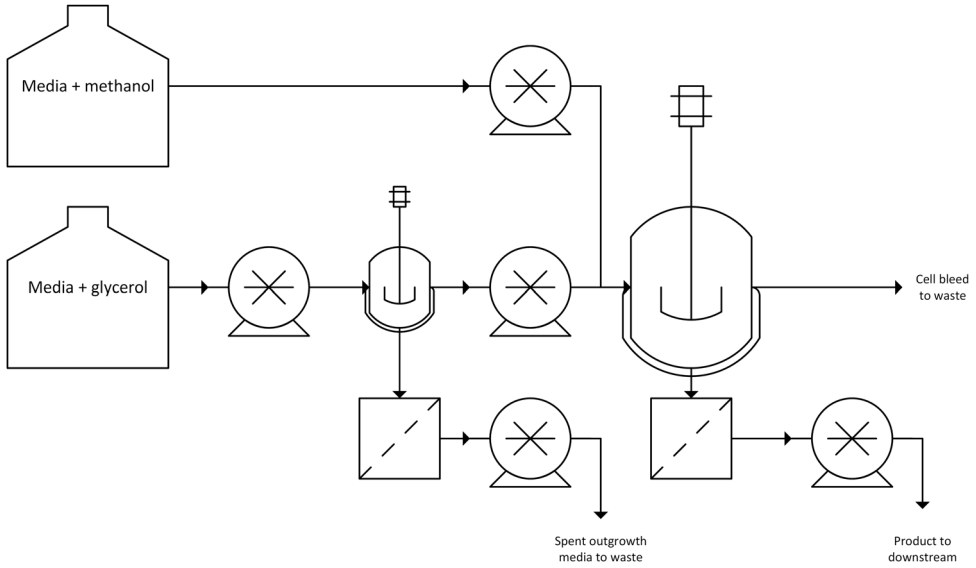

FIGURE 7 Process flow diagram for potential biomass supplementation scheme to maintain copy number stability

Continuous seeding improves stability of processes (Figure 8) at $\frac{\phi}{\alpha} > 100$. At steady state, we can express the ratio of volumes of the seed and production reactors as

$$\frac{V_s}{V} = \frac{\phi X_T}{\mu X_n}, \tag{29}$$

where V_s is the volume of the seed reactor and μ is the growth rate in the seed reactor. If the seed and production reactors operate at equal density, with a copy loss rate of Zhu et al. (2009) ($\alpha = 3.8 \times 10^{-5} h^{-1}$) and a growth rate of $\mu = 0.2 h^{-1}$, the seed reactor is 19 ml/L of production reactor (1.9% working volume), which highlights the relatively high stability of this host.

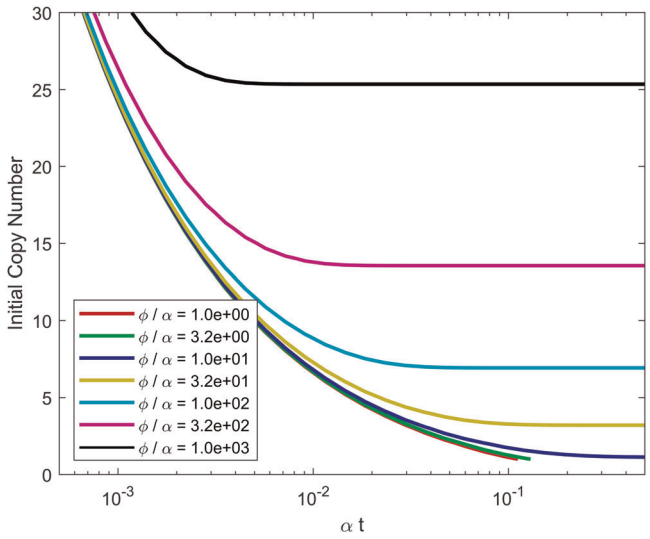

FIGURE 8 Dimensionless time αt required for decay of copies to 0.9 of initial copies over a range of dimensionless seeding rates $\frac{\phi}{\alpha}$. Under the loss rate of Zhu et al. (2009), $\alpha t = 10^{-1}$, 10^{-2}, and 10^{-3} corresponds to a process time of 16 weeks, 11 days, and 26 h, respectively [Color figure can be viewed at wileyonlinelibrary.com]

This method has manifold advantages. It utilizes the same materials as regular fed-batch or perfusion outgrowth, reducing potential scale-up and regulatory risks. While *P. pastoris* has a simplified implementation due to stability under non-induction outgrowth conditions, there may be an added difficulty in this scheme due to methanol adaptation upon seeding (Invitrogen Corporation, 2002). This can be addressed by having a second methanol adaptation vessel in the seed train, or the process can be run with a slightly lower specific productivity. This consideration applies more broadly to extended perfusion processing of any unstable organism, including organisms that utilize constitutive expression. Experimental confirmation of these novel processing schemes would be beyond the scope, but This study demonstrates the theoretical feasibility of such an approach which can later be validated experimentally.

4 | CONCLUSIONS

This study considers genomic stability of transgene copies of a recombinant host, *P. pastoris*, for continuous manufacturing. Model predictions, based on first-order loss in copies, compare well to experimental data in the literature. The stability of strains are dependent on a strain-specific copy number loss rate, α, which can be empirically determined from cultivation data. An extended model accounted for the impact of selection pressure on copy number stability. The selection pressure and stabilization potential of Zeocin was quantified. Another extension to the model analyzed the continuous feeding of high-copy cells grown under a stable, glycerol-fed seed reactor. This approach was shown to stabilize the reactor with a small seed reactor volume.

The quantitative tools developed here provide predictive tools for process design and intensification. Copy number losses obtained from short reactor runs can be used to determine loss rates, which can in turn predict the genomic stability of the strain over longer production periods. Additionally, alternative methods to provide selection pressure or to inject high copy cells can be quantitatively evaluated to determine the level of stabilization provided. Copy number loss in production may be unavoidable but need not be feared, so long the losses can be predicted, managed, and controlled over time.

ACKNOWLEDGMENTS

This study was supported by the Defense Advanced Research Projects Agency (DARPA) and SPAWAR System Center Pacific (SSC Pacific) under contract no. N66001-13-C-4025 and the National Science Foundation (NSF) Graduate Research Fellowship Program under Grant No. 1122374. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author (s) and do not necessarily reflect the views of DARPA, SSC Pacific, or the NSF.

AUTHOR CONTRIBUTIONS

Amos E. Lu and Andrew J. Maloney developed, solved, and analyzed the model framework. Neil C. Dalvie and Joseph R. Brady created strains and performed experimental studies. Amos E. Lu and Andrew J. Maloney authored and edited the first draft. All authors contributed to editing and internal reviews of the work. J. Christopher Love and Richard D. Braatz acquired funding and supervised work.

ORCID

Andrew J. Maloney https://orcid.org/0000-0002-9744-6433
Joseph R. Brady https://orcid.org/0000-0002-2284-3872
Richard D. Braatz https://orcid.org/0000-0003-4304-3484

REFERENCES

Ahmad, M., Hirz, M., Pichler, H., & Schwab, H. (2014). Protein expression in *Pichia pastoris*: Recent achievements and perspectives for heterologous protein production. *Applied Microbiology and Biotechnology*, *98*(12), 5301–5317. https://doi.org/10.1007/s00253-014-5732-5

Aw, R. (2012). Factors affecting the specific productivity of *Pichia pastoris* (Ph.D. thesis). Imperial College London.

Aw, R., & Polizzi, K. M. (2013). Can too many copies spoil the broth? *Microbial Cell Factories*, *12*(1), 1–9. https://doi.org/10.1186/1475-2859-12-128

Bailey, J. E., & Ollis, D. F. (2006). *Biochemical engineering fundamentals*. New York: McGraw-Hill.

Cos, O., Serrano, A., Montesinos, J. L., Ferrer, P., Cregg, J. M., & Valero, F. (2005). Combined effect of the methanol utilization (Mut) phenotype and gene dosage on recombinant protein production in *Pichia pastoris* fed-batch cultures. *Journal of Biotechnology*, *117*(1), 321–335. https://doi.org/10.1016/j.jbiotec.2004.12.010

Crowell, L. E., Lu, A. E., Love, K. R., Stockdale, A., Timmick, S. M., Wu, D., Wang, Y., Doherty, W., Bonnyman, A., Vecchiarello, N., Goodwine, C. A., Bradbury, L., Brady, J. R., Clark, J. J., Colant, N. A., Cvetkovic, A., Dalvie, N. C., Liu, D., Liu, Y., … Love, J. C. (2018). On-demand manufacturing of clinical-quality biopharmaceuticals. *Nature Biotechnology*, *36*(10). https://doi.org/10.1038/nbt.4262

Cunha, A. E., Clemente, J. J., Gomes, R., Pinto, F., Thomaz, M., Miranda, S., Pinto, R., Moosmayer, D., Donner, P., & Carrondo, M. J. T. (2004). Methanol induction optimization for scFv antibody fragment production in *Pichia pastoris*. *Biotechnology and Bioengineering*, *86*(4), 458–467. https://doi.org/10.1002/bit.20051

Curvers, S., Linnemann, J., Klauser, T., Wandrey, C., & Takors, R. (2001). Recombinant protein production with *Pichia pastoris* in continuous fermentation—Kinetic analysis of growth and product formation. *Chemie Ingenieur Technik*, *73*(12), 1615–1611. https://doi.org/10.1002/1522-2640(200112)73:12%3C1615::AID-CITE1615%3E3.0.CO;2-6

Damasceno, L. M., Huang, C. J., & Batt, C. A. (2012). Protein secretion in *Pichia pastoris* and advances in protein production. *Applied Microbiology and Biotechnology*, *93*(1), 31–39. https://doi.org/10.1007/s00253-011-3654-z

Frahm, B. (2014). Seed train optimization for cell culture. *Methods in Molecular Biology*, *1104*(S6), 355–367. https://doi.org/10.1007/978-1-62703-733-4_22

Invitrogen Corporation (2002). *Pichia fermentation process guidelines* (Tech. Rep.). Carlsbad, California: Author.

Kim, N. S., Kim, S. J., & Lee, G. M. (1998). Clonal variability within dihydrofolate reductase-mediated gene amplified chinese hamster ovary cells: Stability in the absence of selective pressure. *Biotechnology and Bioengineering*, *60*(6), 679–688. https://doi.org/10.1002/(sici)1097-0290(19981220)60:6%3C679::aid-bit5%3E3.0.co;2-q

Konstantinov, K. B., & Cooney, C. L. (2015). White paper on continuous bioprocessing. *Journal of Pharmaceutical Sciences*, *104*(3), 813–820. https://doi.org/10.1002/jps.24268

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Marx, H., Mecklenbräuker, A., Gasser, B., Sauer, M., & Mattanovich, D. (2009). Directed gene copy number amplification in *Pichia pastoris* by vector integration into the ribosomal dna locus. *FEMS Yeast Research*, *9*(8), 1260–1270. https://doi.org/10.1111/j.1567-1364.2009.00561.x

Ohi, H., Okazaki, N., Uno, S., Miura, M., & Hiramatsu, R. (1998). Chromosomal DNA patterns and gene stability of *Pichia pastoris*. *Yeast*, *14*, 895–903. https://doi.org/10.1002/(SICI)1097-0061(199807)14:10%3C895::AID-YEA288%3E3.0.CO;2-9

Schenk, J., Balazs, K., Jungo, C., Urfer, J., Wegmann, C., Zocchi, A., Marison, I. W., & VonStockar, U. (2008). Influence of specific growth rate on specific productivity and glycosylation of a recombinant avidin produced by a *Pichia pastoris* Mut$^+$ strain. *Biotechnology and Bioengineering*, *99*(2), 368–377. https://doi.org/10.1002/bit.21565

Schwarzhans, J.-P., Wibberg, D., Winkler, A., Luttermann, T., Kalinowski, J., & Friehs, K. (2016). Integration event induced changes in recombinant protein productivity in *Pichia pastoris* discovered by whole genome sequencing and derived vector optimization. *Microbial Cell Factories*, *15*(1), 84. https://doi.org/10.1186/s12934-016-0486-7

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., DelAngel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S., & DePristo, M. A. (2013). From FastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics*, *43*(1), 11–10. https://doi.org/10.1002/0471250953.bi1110s43

Werbowy, O., Werbowy, S., & Kaczorowski, T. (2017). Plasmid stability analysis based on a new theoretical model employing stochastic simulations. *PLoS ONE*, *12*(8), 1–21. https://doi.org/10.1371/journal.pone.0183512

Werner, R. G., Walz, F., Noé, W., & Konrad, A. (1992). Safety and economic aspects of continuous mammalian cell culture. *Journal of Biotechnology*, *22*(1-2), 51–68. https://doi.org/10.1016/0168-1656(92)90132-s

Zhu, T., Guo, M., Sun, C., Qian, J., Zhuang, Y., Chu, J., & Zhang, S. (2009). A systematical investigation on the genetic stability of multi-copy *Pichia*

pastoris strains. *Biotechnology Letters*, *31*(5), 679–684. https://doi.org/10.1007/s10529-009-9917-4

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Lu A. E., Maloney A. J., Dalvie N. C., et al. Modeling of copy number variability in *Pichia pastoris*. *Biotechnology and Bioengineering*. 2021;118:1832–1839. https://doi.org/10.1002/bit.27698