scientific reports

Check for updates

OPEN

A bayesian approach for parameterizing and predicting plasmid conjugation dynamics

Sirinapa Kumsuwan, Chanon Jaichuen, Chakachon Jatura & Pakpoom Subsoontorn✉

Population dynamic models that explain and predict the spread of conjugative plasmids are pivotal for understanding microbial evolution and engineering microbiomes. However, prediction uncertainty of these models has rarely been assessed. We adopt a Bayesian approach, employing Markov Chain Monte Carlo (MCMC), to parameterize and model plasmid conjugation dynamics. This approach treats model parameters as random variables whose probability distributions are informed by data on plasmid population dynamics. These distributions allow us to estimate credible intervals of the model's parameters and predictions. We validated this approach using synthetic population dynamic data with known parameter values and experimental population dynamic data of mini-RK2, a miniaturized counterpart of the well-characterized and widely used RK2 conjugation plasmid. Our methodology accurately estimated the parameters of synthetic data, and model predictions were robust across time scales and initial conditions. Incorporating long-term population dynamic data enhances the precision of parameter estimates related to plasmid loss and the accuracy of long-term population dynamic predictions. For experimental data, the model correctly explained and predicted most population dynamic trends, albeit with broader credible intervals. Incorporating long-term data also improves credible ranges of most parameters. However, in some cases, such as with the growth parameter of cells with the conjugative plasmid, the inclusion of long-term data can lead to stronger correlations and potential identifiability issues between key parameters. Overall, our method allows for deeper investigation of plasmid population dynamics and could potentially be generalized to study population dynamics of other mobile genetic elements.

Keywords Conjugation, Plasmid, Bayesian approach, Markov chain Monte Carlo

The study of mobile genetic elements (MGEs), such as phages and conjugative plasmids, is pivotal for understanding Horizontal Gene Transfer (HGT) in microbial communities. These elements are major drivers of microbial evolution, influencing the dissemination and persistence of critical traits, including virulence and antibiotic resistance[1–4]. The ability to accurately predict the dynamics of MGEs spread in microbial populations is essential, not only for understanding microbial evolution but also for addressing public health concerns related to these evolutionary processes[5,6]. Furthermore, this predictive capability is crucial for devising strategies to either disseminate or eliminate specific genes in microbial population, thereby engineering microbiomes with desired features[7–9].

The measurement and modelling of MGEs, particularly conjugative plasmids, within microbial populations have been by extensively explored[10]. A spectrum of modelling approaches, ranging from mass action kinetics to agent-based models, has been devised to explain and predict population dynamics of MGEs[11–20]. These models hinge on accurately determining critical parameters, such as MGEs' transfer and loss rates, alongside their influence on the fitness of host cells. Nonetheless, the simultaneous occurrence of MGE transfer, loss, and cell growth poses significant challenges to precise parameter quantification. Additionally, the sensitivity of parameter measurement techniques to experimental setups introduces further complexity. For instance, the quantification of MGE transfer rates can be influenced by variables such as cell density and growth rates[21,22]. Natural variability in population dynamics and limited reproducibility of experimental measurements make it challenging to assess parameter certainty. As a result, evaluating the reliability of model predictions is a crucial but underexplored aspect. Despite its importance, previous research has often overlooked the reporting and analysis of uncertainty surrounding parameter estimates and model predictions.

Department of Biochemistry, Faculty of Medical Science, Naresuan University, Phitsanulok 65000, Thailand. ✉email: pakpoomsu@nu.ac.th

Parameter estimation from experimental data can fundamentally be tackled via two approaches: the frequentist and the Bayesian approaches[23]. The frequentist approach conceptualizes parameter estimation as an optimization problem, seeking a singular set of parameters that most accurately fits given experimental data. This involves minimizing an error function that quantifies the discrepancy between the experimental observations and the model predictions derived from a specific parameter set. Conversely, the Bayesian approach focuses on estimating the probability distribution of parameters, known as the 'posterior distribution'. This distribution is computed from 'prior distributions', reflecting prior knowledge about the parameters, and the 'likelihood', which is the probability of observing the experimental data assuming that a given parameter set is correct. Despite its higher computational demands, the Bayesian approach offers the advantage of enabling a detailed assessment of the certainty associated with the estimated parameters and the ensuing model predictions. In practice, the posterior distribution can be determined using Markov Chain Monte Carlo (MCMC) techniques, methods that have been pivotal in Bayesian parameter inference across various scientific fields. Originating in the physical sciences, MCMC has expanded into biology, finding applications in the systems biology of gene regulatory networks, epidemiology of infectious diseases, and ecological population dynamics[23–30].

In this study, we adopt a Bayesian approach to investigate the dynamics of MGE spread within microbial populations, with a particular focus on a simplified system consisting of both conjugative and non-mobilizable plasmids. We used mini-RK2 plasmid, a compact version of the extensively studied RK2 plasmid, to represent a conjugative plasmid[31]. Known for its broad host range and efficacy in DNA transfer across a diverse spectrum of microbes, the RK2 plasmid (also known as RP4 plasmid) and its derivatives have been widely used for gene delivery to undomesticated microbes or microbial community[9,32]. The mini-RK2 plasmid, with its relative simplicity and potential for significant applications, thus presents an ideal model for our investigation. We employ MCMC techniques to derive posterior distributions of parameters governing conjugative transfer, plasmid loss, cell growth, and cell death, leveraging both "synthetic data", with known parameters, and "experimental data" from our laboratory involving the mini-RK2 plasmid, where the parameters are undetermined. Our findings not only confirm the utility of MCMC for accurate parameter estimation and dynamic modelling but also highlight the inherent limitations of this approach and the intricate challenges presented by conjugation systems that are not fully addressed by simplistic models.

Results

We established experimental setup, modelling framework and analysis workflow

We used *Escherichia coli* strain DH5α as donors and recipients. The donor cells were equipped with a mini-RK2 conjugative plasmid, termed X61, which harbours genes for green fluorescent protein (GFP) and kanamycin resistance (Km^R). The recipient cells contained a non-mobilizable plasmid, X13, with a red fluorescent protein gene (RFP) and chloramphenicol resistance gene (Cm^R). X61 can self-transfer from donors to recipients, resulting in transconjugants that carry both X61 and X13 plasmids (Fig. 1A). We modelled this system with deterministic mass action kinetics, adapted from Lopatkin et al. (2017)[18] (Fig. 1B, Fig S1). In our model, cell population comprises subpopulations of four distinct cell types: DH5α (DH5a), DH5α with X13 (DH5a-X13), DH5α with X61 (DH5a-X61), and DH5α with both X13 and X61 (DH5a-X13-X61). The dynamics of these subpopulations are governed by eight kinetic parameters: the transfer rate of X61 (η), the growth rate of DH5α (μ), the growth rate of DH5α-X13 (μ_{13}), the growth rate of DH5α-X61 (μ_{61}), the growth rate of DH5α-X13-X61 (μ_{1361}), the loss rate of X13 (κ_{13}), the loss rate of X61 (κ_{61}), and the combined rate of cell death and dilution (D). The model enables the prediction of population trajectories for each cell type over time using ordinary differential equations (ODEs) (Fig S2). In this study, "synthetic data" refers to computationally generated data points derived from our mathematical model, whereas "experimental data" refers to measurements from conjugation experiments in our laboratory.

Our analysis workflow consists of two parts (Fig. 1C). The first part aims to assess the validity and limitations of our MCMC approach for estimating and analyzing kinetic parameters from given datasets (Fig. 1C, top). This part starts with hypothetical 'actual' parameter values, which we used for simulating population time courses. We then added log-normal random noise to these population time courses to generate "synthetic data." This approach reflects how experimental variability affects observed data, resulting in synthetic data that mimics real-world measurements. We generated three synthetic data points for each time point to emulate a triplicate experiment. Part of these synthetic data was designated as a "training set" while the rest is designated as "testing set." Next, we applied the Metropolis MCMC algorithm to the training set of synthetic data to obtain a "parameter ensemble," i.e., a collection of possible parameter sets that can explain the training set data (Fig S3-S5). Finally, the parameter ensemble was used for simulating a collection of population time courses (Fig S6). We can compare how well these simulated population time courses fit the training and testing set of synthetic data. We can also explore distributions and correlations among parameters in the ensembles (Fig S7). These distributions estimate the posterior distribution of parameters. This information tells us how credible our parameter estimation can be as well as how sensitive the population time courses are to changes in parameter values. Moreover, we can compare the 'actual' parameter values to the parameter ensemble to assess how accurate our parameter estimations are.

This second part of the analysis workflow (Fig. 1C, bottom) mirrors the first, with the key difference being the use of "experimental data" derived from actual population measurements in our laboratory. We again divided the data into training and testing sets and employ the Metropolis MCMC algorithm to derive a parameter ensemble from training set. We then analysed population time courses simulated with this ensemble, as well as the posterior distribution and correlation of parameters within the ensemble. Although the actual parameter values remain unknown in this scenario, the posterior distribution can still inform us about the confidence we can place in our estimated parameter values from the given training dataset. Furthermore, any unsuccessful attempts to predict population time courses may indicate that critical mechanisms are missing from our simplified kinetic models.

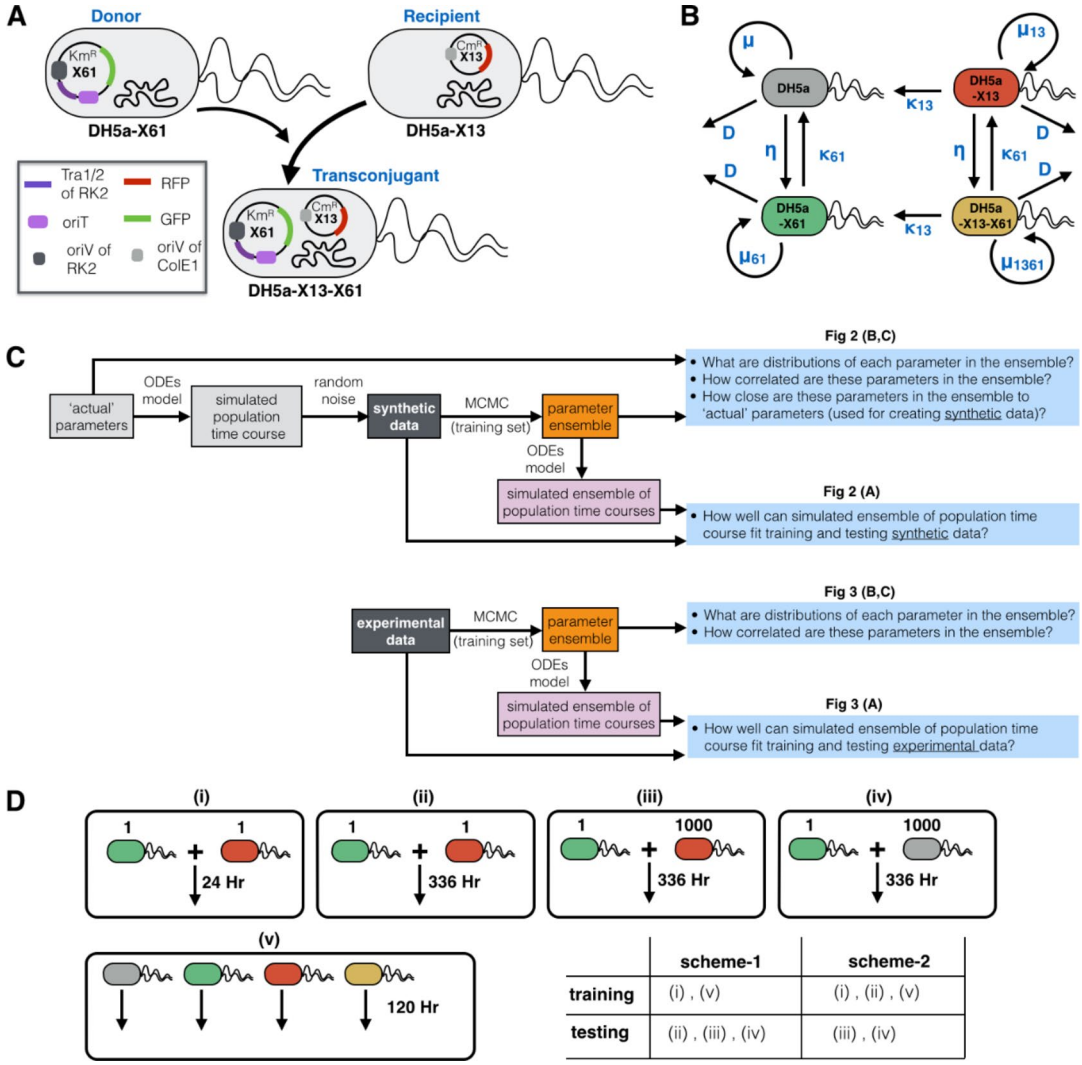

Fig. 1. Conjugative transfer system and parameter analysis workflow. (**A**) A diagram showing conjugative transfer of a plasmid from a donor cell to a recipient cell. (**B**) Modelling population dynamics. Each cell type (DH5a, DH5a-X13, DH5a-X61, and DH5a-X13-X61) could have different growth rate (μ, μ_{13}, μ_{61}, and μ_{1361}, respectively) but have the same cell loss rate due to death and dilution (D). The total population (all cell types combined) is limited to a fixed carry capacity (Nm, not shown in the figure). Plasmid X13 and X61 can be lost from a cell at different rate (κ_{13} and κ_{61}, respectively). Plasmid X61 can be transferred from DH5a-X61 or DH5a-X13-X61 to a cell DH5a or DH5a-X13 at rate η. (**C**) parameter analysis workflows and key questions to answer in this study using synthetic data (top) or experimental data (bottom). (**D**) Five conjugation or growth assays conducted in this study. The number (1 or 1000) above each cell indicates the initial cell ratio at the beginning of the assay. The numbers on the right of the arrows indicate the duration of the assay. The table shows two different schemes for dividing data from these five assays into training and testing datasets.

We used five different conjugation or growth assays to conduct simulations and experiments (Fig. 1D). These assays were: (i) conjugation between DH5a-X61 and DH5a-X13 at a 1:1 ratio for 24 h; (ii) conjugation between DH5a-X61 and DH5a-X13 at a 1:1 ratio for 14 days; (iii) conjugation between DH5a-X61 and DH5a-X13 at a 1:1000 ratio for 14 days; (iv) conjugation between DH5a-X61 and DH5a at a 1:1000 ratio for 14 days; (v) population growth of each cell type (DH5a, DH5a-X13, DH5a-X61, DH5a-X13-X61), cultured separately over five days. We employed two different schemes for splitting the data from these five assays into training and testing dataset. Scheme-1 used (i) and (v) as the training set and (ii), (iii), and (iv) as the testing set; Scheme-2 used (i), (ii), and (v) as the training set and (iii) and (iv) as the testing set. Notice that scheme-1's training set only has short-term conjugation data (e.g., 24 h), whereas scheme-2's training set encompasses data on both short-term and long-term conjugation (e.g., 14 days). Previous studies on conjugation dynamics often rely on short-term conjugation experiments (< 24 h) to determine conjugative transfer rates and assess cell replication rates from growth curves of each strain[18,19]. Therefore, in terms of the data content available for parameter estimation, these studies are akin to our use of the scheme-1 training set. We hypothesize that incorporating

data from long-term conjugation experiments, specifically (ii), could enhance the accuracy and precision of parameter estimation and model predictions.

For synthetic data, the model accurately estimates parameters and predicts datasets with long-term data enhancing credibility of predictions

Using computer simulation, we generated synthetic data sets to emulate the five assays described above (Fig. 2A, data points). To accurately reflect the limitations of measurement in actual experiments, we assumed that not all subpopulation data would be available for analysis. For instance, in conjugation experiments (i) through (iii), data for the DH5a subpopulation time course was unavailable, despite the potential presence of this cell type, due to its lack of a unique selection marker for a colony forming unit assay. Similarly, during growth measurement of DH5a-X13, DH5a-X61, and DH5a-X13-X61, plasmids may be lost from certain cells, leading to the emergence of additional cell types such as DH5a; however, this data was not captured, as our quantification was limited to DH5a-X13, DH5a-X61, and DH5a-X13-X61. Therefore, we utilized only the available data (indicated by data points in Fig. 2A) to derive the parameter ensemble and conduct our analysis. For our selected parameter set, subpopulations harboring X61 (DH5a-X61 and DH5a-X13-X61) eventually dominated the population, even when only 1/1000 of the entire population carried this plasmid at the onset of the simulated conjugation experiments (Fig. 2A, iii – iv). Subpopulations carrying X13 (DH5a-X13 and DH5a-X13-X61) experienced a slight decrease over the course of the simulated conjugation experiments (Fig. 2A, ii-iii).

We used the Metropolis MCMC algorithm to obtain parameter ensembles that could explain training synthetic datasets (Fig S8). Parameter ensembles derived from training datasets were used to simulate collections of population time courses for experiments (i) – (v). The geometric means and 95% credible intervals of these

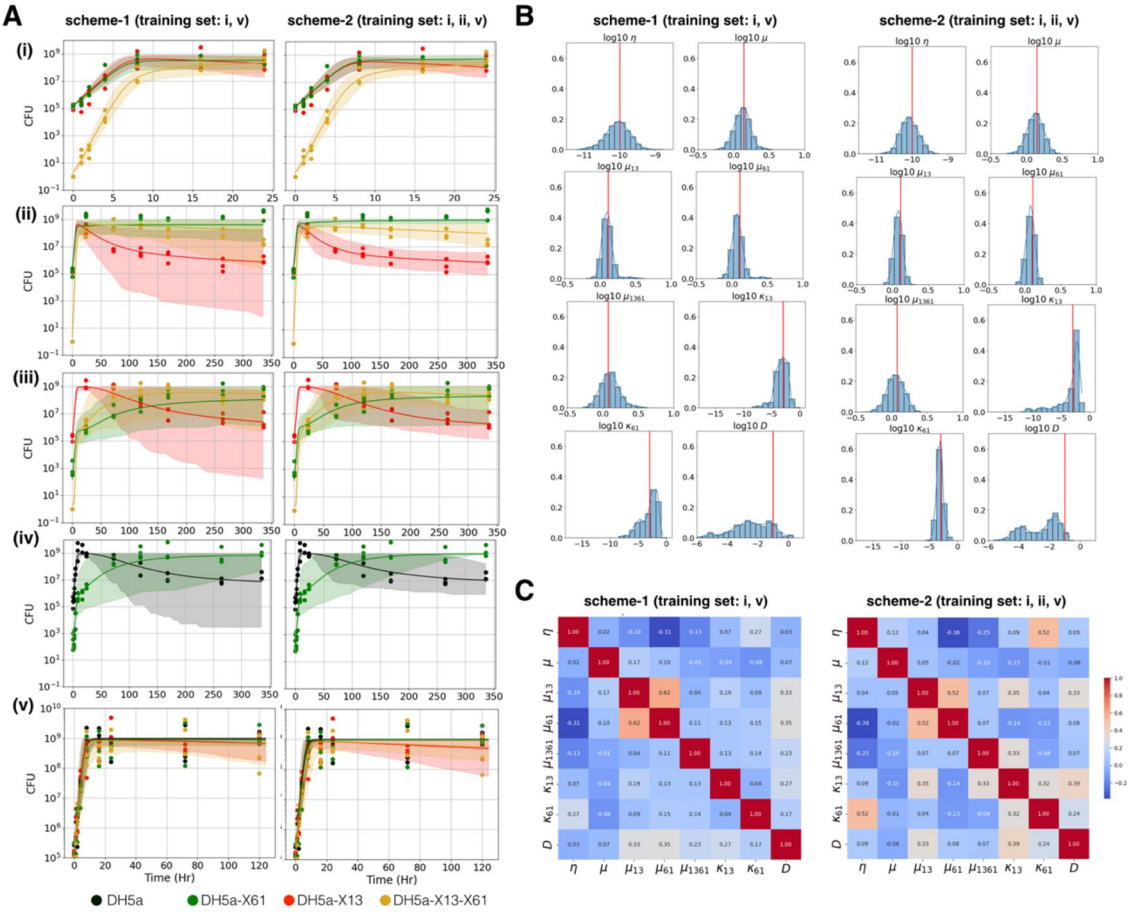

Fig. 2. Utilizing and analysing of parameter ensembles from synthetic data. (**A**) Synthetic data, generated by a computer model, and simulated population time courses using parameter ensembles. Points are synthetic data; curves and shade areas show geometric mean and 95% credible interval, respectively, of simulated ensemble of population time courses. DH5a, DH5a-X13, DH5a-X61 and DH5a-X13-X61 population levels, measured as Colony Forming Unit (CFU), are shown in black, red, green and yellow, respectively. (**i**)- (**iv**) show time course population changes of simulated conjugation experiment. (**v**) shows simulated growth curve of individual cell type cultured in separated environment. Left and right column show the same synthetic data point but different schemes for splitting training and testing datasets. (**B**) Distribution of each parameter in parameter ensembles derived from synthetic data. Red lines show the actual values of parameters used for generating synthetic data (**C**) Correlation among parameters in parameter ensemble derived from synthetic data.

population time courses are depicted with colored lines and shaded areas, respectively, in Fig. 2A. When only short-term conjugation data (i) served as the training set, the geometric mean time courses aligned closely with both the training (Fig. 2A I, v; left) and testing (Fig. 2A ii, iii, iv; left) synthetic data points. Consequently, this parameter ensemble could accurately explain and predict the synthetic data. However, the 95% credible intervals for predictions became significantly wider at later time points, particularly for DH5a and DH5a-X13 (Fig. 2A ii, iii, iv; left), indicating a low predictive precision (i.e., low level of certainty about the predictions) from these parameter ensembles. This outcome was expected, considering the training data encompassed only short-term conjugation information. Conversely, when the training set included both short-term (i) and long-term (ii) conjugation data, the simulated population time courses predicted the long-term testing dataset (iii)-(iv) with greater precision (Fig. 2A, right). Notably, the inclusion of long-term conjugation data in the training set resulted in much narrower 95% credible intervals for the simulation time courses compared to those without it (Fig. 2A, iii-iv, right compared to left).

Posterior distributions of each parameter value illuminate the level of certainty we can attribute to estimated parameter values given specific training datasets. Narrow posterior distributions centered around the 'actual' parameter values used to generate synthetic data suggest a high degree of certainty and accuracy of these estimated parameter values. Whether training sets from (i) + (v) or (i) + (ii) + (v) were used, the estimated conjugative transfer parameters (η) and the growth parameters for all strains (μ, μ_{13}, μ_{61}, and μ_{1361}) closely align with their actual values (Fig. 2B). The shapes and widths of their distributions are almost identical, regardless of the training set used. Therefore, incorporating additional data from a long-term conjugation experiment (ii) does not significantly enhance the accuracy or certainty of these estimated parameter values. On the contrary, with additional data from long-term conjugation experiment (ii), estimated plasmid loss parameters (κ_{13} and κ_{61}) exhibit narrower distribution and shift closer to their actual values. This underscores the utility of long-term conjugation data in refining estimates for these specific parameters. Interestingly, the posterior distributions for the cell loss parameter D span six orders of magnitude and are not centred around the actual parameter value used for generating synthetic data, irrespective of the inclusion of long-term conjugation data (ii) in the training set. This phenomenon could be attributed to the D value in the synthetic data generation parameter set being so low that it has a negligible impact on the simulated time courses, allowing any lower estimated D value to fit the training dataset adequately.

For parameter ensembles derived from both (i) + (v) and (i) + (ii) + (v) training data, we observed similar correlation patterns among parameters. Overall, most parameters exhibited minimal correlation (Fig. 2C). The positive correlation between μ_{13} and μ_{61} indicates that the data are compatible with a range of growth rates for the X13 and X61 strains, as long as their relative ratio within the population remains consistent. The negative correlation between η and μ_{61} suggests that compensatory changes in these parameters are necessary to maintain the prevalence of X61 within the population. The incorporation of data from the long-term conjugation experiment (ii) alters the relationship between η and κ_{61}, shifting from a weak negative correlation to a positive one. This change suggests that the impact of the X61 loss parameter, κ_{61}, becomes more pronounced over extended periods. Therefore, adding long-term experimental data to the training set imposes a further balance between this parameter and the conjugative transfer rate parameter, η.

For experimental data, the model accurately predicts dataset trends but encounters wide credible intervals as well as complex parameter distributions and correlations

We gathered "experimental data" analogous to the "synthetic data" discussed in the previous section. Specifically, we conducted four distinct conjugation experiments on filter, each under different initial conditions or time scales (Fig. 3A i-iv, data points). Additionally, we measured the growth kinetics of each strain separately (Fig. 3A v, data points). All four strains exhibited similar growth kinetics, with carrying capacities (the maximum total cell density in the system) approximately 1E + 9 CFU. The non-mobilizable plasmid X13 was rapidly lost from the population. Notably, a decline in subpopulations carrying X13 (DH5a-X13 and DH5a-X13-X61) was observed within the first 24 h without antibiotic selection (Fig. 3A, i). X13 was entirely absent from the population within a week, even when half of the population initially contained this plasmid (Fig. 3A, ii). Over an extended timeframe, we noted significant variability in the DH5a-X13 and DH5a-X13-X61 populations during growth experiments (Fig. 3A, v). This variability likely stems from the stochastic loss of X13 at the experiment's early stages, which, over time, may be amplified into more pronounced variability. In contrast, X61 rapidly proliferated within the population in experiments (i) – (iv). Specifically, in experiments (i) and (ii), the proportion of cells containing X61 (DH5a-X61 and DH5a-X13-X61) increased from 50% to nearly 100% within 24 h; in experiment (iii-iv), the proportion rose from 0.1% to nearly 100% within five days.

To obtain parameter ensembles, we used experimental data from either (i) + (v) or (i) + (ii) + (v) as the training dataset. We used Metropolis MCMC algorithm to obtain parameter ensembles that could explain training experimental data sets (Fig S9). These ensembles were then used to simulate the experimental outcomes, with the geometric mean and credible intervals of the simulated time courses presented as colored lines and shaded areas, respectively (Fig. 3A). To provide a clearer comparison between the experimental data and the simulated time courses, we also present experimental data without the credible intervals, alongside simulated time courses derived from both the geometric mean of the Bayesian ensemble and the best-fit parameter set (Fig S10).

We observed that parameter ensembles could account for the training data from the short-term experiment (i). Specifically, the geometric means of the simulated time courses closely matched the experimental data points, and the credible intervals remained relatively narrow, regardless of whether we used (i) + (v) or (i) + (ii) + (v) as the training data (Fig. 3A). For the training datasets, the simulated time courses from the best-fit parameter set were similar to the geometric means of the simulated time courses from the parameter ensembles (Fig S10, row-i and row-ii). However, the predictive capability of these ensembles for the testing experimental data was less robust. Notable discrepancies arose between the geometric means of the simulated time courses and the

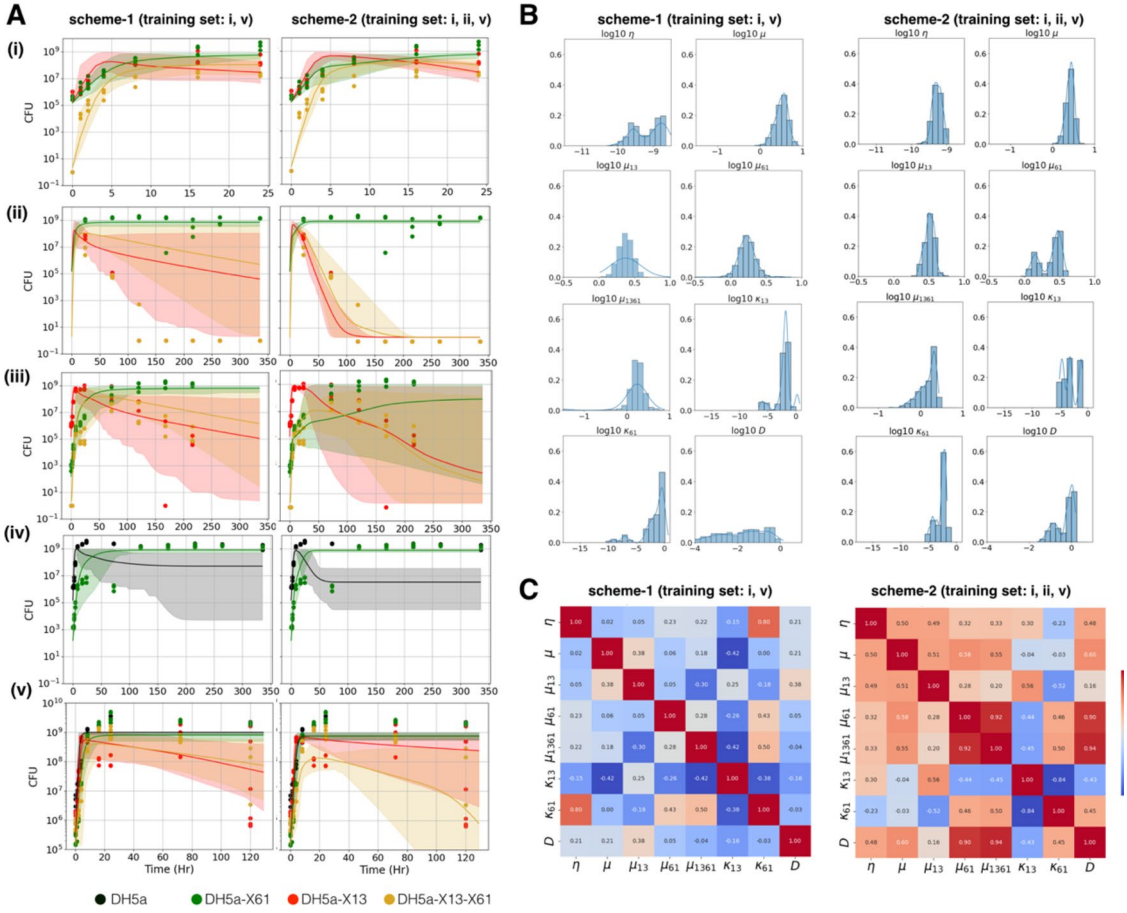

Fig. 3. Utilizing and analysing of parameter ensembles from experimental data. (**A**) Experimental data and simulated population time courses using extracted parameter ensembles. Points are experimental data; curves and shade areas show geometric mean and 95% credible interval, respectively. DH5a, DH5a-X13, DH5a-X61 and DH5a-X13-X61 population levels, measured as Colony Forming Unit (CFU), are shown in black, red, green and yellow, respectively. (**i**)- (**iv**) show time-course population changes of conjugation experiment. (v) shows growth curve of individual cell type cultured in separated environment. Note that in experiment (**iv**), it is difficult to reliably determine the population level of DH5a when its level is significantly lower than that of DH5a-X61. This difficulty arises because DH5a cannot be specifically selected and must be calculated by subtracting DH5a-X61 from the total population, as there is no selection marker to isolate DH5a from DH5a-X61. After day 5 of this experiment, we found that CFU on LB agar are within same order of magnitude as CFU on LB agar with kanamycin (selector for X61). This could imply that most cells have received X61 plasmid. (**B**) Distribution of each parameter in parameter ensembles derived from experimental data. (**C**) Correlation among parameters in parameter ensemble derived from experimental data.

experimental data points when only short-term conjugation data served as the training dataset (Fig. 3A, ii-iv, left; Fig S10, ii-iv first column). The incorporation of long-term conjugation data into the training set improved the congruence between simulated and experimental time courses (Fig. 3A, ii-iv, right; Fig S10, ii-iv third column), yet the credible intervals of these predictions remained exceedingly wide (Fig. 3A, iii, right). Furthermore, the inclusion of long-term conjugation data appeared to widen the credible intervals for the growth curves (Fig. 3A, v, right versus left). Overall, our simplified model faced challenges in simultaneously explaining the short-term conjugation (i), long-term conjugation (ii), and growth kinetics (v). Notably, for the testing datasets, the simulated time courses from the best-fit parameter set differed significantly from the geometric means of the simulated time courses from the ensembles (Fig S10, row-iii, third and forth column), with the geometric means appearing to align better with the experimental data.

For the synthetic data discussed in the previous section, the posterior distributions of parameters typically approached unimodal distributions, except for the cell loss rate constant D (Fig. 2B). However, with the experimental data, we observed that posterior distributions were more prone to deviations from unimodal forms (Fig. 3B). For instance, the posterior distributions for the X61 transfer rate (η) and the X61 loss rate (κ_{61}) appeared bimodal when using (i) + (v) and (i) + (ii) + (iv) as training sets, respectively. We hypothesize that this discrepancy arises from a more complex landscape of posterior values in the experimental data compared to the synthetic data, potentially leading to multiple pronounced local maxima that trap the MCMC random walks. Additionally, contrary to the synthetic data where the addition of long-term conjugation data scarcely influenced

the posterior distribution of the cell loss parameter (D), the inclusion of long-term data in the experimental context significantly narrowed and shifted the distribution of D towards higher values. This shift indicates that a higher value of D is instrumental in elucidating the rapid decline of DH5a-X13 and DH5a-X13-X61 observed in long-term conjugation data (ii). Distinct differences were also evident in the parameter correlation patterns when long-term conjugation data were excluded versus included in the training dataset (Fig. 3C, left compared to right). In particular, a more pronounced positive correlation among parameters emerged with the inclusion of long-term data. The growth parameters for the subpopulation carrying X61 (μ_{61}, and μ_{1361}) and the cell loss parameter (D) became strongly correlated. We postulate that such correlations enable the model to sustain the level of X61 in the population while accelerating the attrition of the X13-carrying subpopulations, consistent with the trends suggested by the training data from the long-term conjugation experiment (ii).

With (i) + (ii) + (iv) as the training dataset, we observed bimodal distributions for two key parameters: the growth rate of DH5α cells carrying the X61 plasmid (μ_{61}) and the combined death and dilution parameter (D) (Fig. 3B). To explore the implications of these bimodal posteriors, we selected parameter sets from each of the two peaks and performed additional simulations to compare the model's predictive accuracy. We ran separate simulations using parameter sets from each peak to evaluate how well they captured the dynamics of both the training datasets (i, ii, v) and the testing datasets (iii–iv). The results, illustrated in Figure S11, indicate that while both parameter sets adequately fit the training data, their predictive accuracy varied for the testing data.

For μ_{61}, the predicted time courses using the parameter set from the lower peak were a better match to the testing dataset (iii). Specifically, this parameter set correctly predicted that X13-harboring subpopulations would go extinct, while the DH5a-X61 subpopulation would reach nearly 100% within a few days (Fig S11, iii, first column). In contrast, the parameter set from the higher peak of the μ_{61} distribution incorrectly predicted that the majority of the population would still consist of DH5a-X13, at least until the end of the 14-day experiment (Fig S11, iii, second column). For D, the parameter sets from both peaks produced similar prediction trends but with different timings. The parameter set from the lower peak predicted a faster extinction of X13-harboring subpopulations and a quicker dominance of DH5a-X61 (Fig S11, iii, third column). Conversely, the parameter set from the higher peak predicted a slower extinction of X13-harboring subpopulations and a slower dominance of DH5a-X61, which seemed to better align with the actual experimental data (Fig S11, iii, fourth column). Notably, this parameter set and prediction trend are similar to those of the best-fit parameter set (Fig S10, iii, fourth column).

Discussion

Population dynamics and parameters of the conjugative plasmid

Our work presents the first comprehensive study of mini-RK2 population dynamics. The mini-RK2 plasmid was derived from the widely utilized and well-studied RK2, which is renowned for its broad-host-range and extensive applications in gene delivery to microbial hosts and microbiomes[9]. Silbert et al. (2021)[33] and Aparicio et al. (2022)[31]miniaturized RK2 into mini-RK2, reducing its size from 60 kb to 25 kb by removing genetic components extraneous to DNA transfer. This streamlining potentially simplifies the understanding and engineering of this plasmid, opening doors to various applications in the fundamental science of IncP plasmid group and the field of genetic engineering. Although these studies quantified the conjugation efficiency and demonstrated the plasmid's ability to infiltrate and persist in complex microbial communities, they did not explore the intricate details of the plasmid's propagation, loss, and impact on host fitness. Our investigation bridges this knowledge gap by scrutinizing both the short-term and long-term dynamics of the plasmid under varied initial conditions, extracting key parameters, and rigorously testing our model's predictive power. Such in-depth analyses of conjugative plasmid behaviors are scarce in the literature, even for well-characterized plasmids like the original RK2, F, and R388[10].

How do our findings about mini-RK2 plasmid dynamics and parameters compare to previous studies? Our short-term conjugation results dynamics align well with the findings of Aparicio et al. (2022)[31]. Both their study and ours showed that, starting with a 1:1 donor-to-recipient ratio on agar, mini-RK2 can infect most recipient cells within one day. Aparicio et al. also demonstrated that mini-RK2 could persist in a soil microbiome for at least 35 days, albeit at a very low frequency of 1E-7 transconjugants per recipient, even though their experiments began with a 10:1 donor-to-recipient ratio. This low level of infected cells could be explained by the fact that most soil microorganisms may be less susceptible to mini-RK2 than *E. coli*. In contrast, in our long-term experiments where all the recipients were *E. coli*, mini-RK2 successfully infected nearly all recipient cells. Regarding RK2 and its other relatives, most long-term plasmid population studies were conducted in microbiomes[34–37]. These studies reported plasmid persistence but did not provide sufficient relevant data for direct comparison with our results. The most relevant long-term study on RK2 in a simple *E. coli* system was by Lopatkin et al. (2017)[18]. This study explored the conjugation dynamics of RK2 in an *E. coli* population over 21 days. Unlike our study, Lopatkin et al.[18] performed their experiments in liquid culture at an extremely low cell density in each growth cycle (~ 400 cells per mL). Despite these conditions, RK2 spread throughout the population within 10 days.

Lopatkin et al.[18] reported a conjugative transfer rate (η) for RK2 of 1.76E-12 cells/hour, which is about 3 orders of magnitude lower than our best-fit conjugative transfer rate of 5.89E-9 cells/hour (Fig S10). This discrepancy likely stems from the higher efficiency of conjugative plasmids with rigid pili, like RK2, on filter compared to liquid culture. Indeed, a meta-analysis by Sheppard et al. (2020)[38] reported that the conjugative transfer rate of such plasmids on agar is 2–3 orders of magnitude higher than in liquid culture. Notably, Lopatkin et al.[18] also reported a negative growth burden for RK2, where plasmid-bearing cells grew faster than those without it, even without antibiotic selection. In our case, we observed a bimodal posterior distribution for μ_{61} (Fig S11). One parameter set indicated slower growth for plasmid-bearing cells ($\mu_{61} < \mu$), while the other showed slightly faster growth for plasmid-bearing cells ($\mu_{61} > \mu$). Although both sets fit the training data, only the set with $\mu_{61} < \mu$ accurately predicted the testing data. Given that mini-RK2 still contains over 20 constitutively expressed

genes, it is more plausible for it to impose a growth burden. In contrast, the natural RK2 has intricate regulatory networks controlling gene expression to minimize resource burdens on the host[39]. The loss of these networks in mini-RK2 likely contributes to its higher resource demand.

What is the relative importance of each parameter in the model's performance for explaining the data? In this study, the posterior distributions of the plasmid transfer rate (η) and growth parameters (μ, μ_{13}, μ_{61}, μ_{1361}) were narrowly spread within an order of magnitude, indicating that these parameters are critical for explaining the data (Fig. 3B, training set i + ii + v). In contrast, the death and dilution parameter (D) had a broader distribution, spanning about two orders of magnitude, while the plasmid loss parameters (κ_{13} and κ_{61}) showed even wider distributions, spanning approximately five orders of magnitude. This suggests that plasmid loss contributes very little to the overall population dynamics. Lopatkin et al. (2017)[18] also explored the roles of these parameters in determining whether the plasmid would persist in the population. The authors derived a criterion for plasmid persistence, showing that when the death and dilution parameter is significantly higher than the plasmid loss parameter, the impact of plasmid loss becomes negligible. Therefore, our findings are consistent with their conclusions.

Training vs. testing, synthetic vs. experimental, short vs. long-term conjugation

The second aspect of novelty in this study lies in the methodological approach to modeling and validating plasmid conjugation dynamics. While modeling efforts date back to the 1970s, previous studies have predominantly aimed at predicting the steady-state outcomes of conjugative plasmid populations, such as their persistence or extinction[11–13,18,40]. Our study advances beyond this traditional objective by seeking to chart the temporal dynamics of subpopulations carrying the plasmid. We used distinct training and testing datasets to assess predictive power of the model, eschewing the common practice of merely adjusting the model to fit the data. While a study by Malwade et al. (2017)[19] present the prediction of short-term dynamics over a five-hour window, it did not extend to the long-term dynamics that are central to our analysis. Our work, therefore, showcases the model's predictive strength, successfully predicting the behavior of the conjugation system under various initial conditions and timescales that diverge from the scenarios presented in the training datasets. We also leveraged both synthetic data, where underlying mechanisms and parameters are fully known and can be manipulated, and experimental data, which reflect real biological complexity but are more challenging to interpret. This comprehensive approach allowed us to validate the model with a clear understanding of its behavior under controlled conditions, while also confirming its applicability to real-world scenarios.

We investigated how the choice of training datasets affects our ability to determine parameter distributions and make predictions. Specifically, we examined the extent to which including long-term conjugation data (ii) in the training dataset improves parameter estimation and predictive accuracy for testing datasets. For synthetic data, even without long-term conjugation data, we could accurately estimate parameter values and correctly predict population time course trends. Including long-term data had minimal impact on most parameter posterior distributions. However, it significantly narrowed the credible intervals of the predicted population time courses. For experimental data, incorporating long-term conjugation data clearly altered the posterior distributions of most parameters. Additionally, without long-term data in the training set, the model failed to predict the rapid loss of populations harbouring X13 (Fig. 3A, ii–iii, left). This finding is particularly significant, as many previous studies have relied solely on short-term conjugation data, which may not provide sufficient information for accurate predictions across different conditions and timescales[10]. Researchers should exercise caution when interpreting results based only on short-term data, as it may lead to incomplete or misleading conclusions about plasmid dynamics over extended periods.

For experimental data, incorporating long-term data improves the precision of certain parameters but introduces potential challenges related to parameter identifiability. For instance, the growth parameter (μ_{61}, μ_{1361}) and death parameter (D) exhibited stronger correlations with other model parameters when long-term conjugation data were included. This increased correlation suggests that the additional time-course information may impose tighter constraints on the model, leading to interdependencies between certain parameters. Such correlations can make it difficult to distinguish the individual contributions of each parameter to the observed population dynamics, thereby complicating their precise estimation. This is not the case for our synthetic data where there is less dramatic changes cell population over long time scale (comparing Fig. 2 A-ii vs. Fig. 3 A-ii data points; Fig. 2 C vs. Fig. 3 C correlation heat map). These findings indicate the need for careful interpretation of posterior distributions when correlations between parameters are high, and point towards a potential limitation in the identifiability of parameters in long-term datasets.

For experimental data, the observed shift towards higher death and dilution rates (D) when incorporating long-term data raises questions about the biological realism of these estimates. One possible explanation for this shift is that long-term dynamics capture additional population-level phenomena such as nutrient depletion, waste accumulation, or changes in cellular stress responses, which are not accounted for in our current model. These environmental changes could lead to an increased rate of cell death over extended periods, which the model interprets as a higher death parameter (D). Another possible explanation is the rapid decline in the subpopulation containing the X13 plasmid observed in long-term conjugation experiments (Fig. 3A-ii). This decline could result from actual cell death or the loss of the plasmid, causing cells to revert to the DH5α strain without the plasmid. Since the data on the subpopulation of DH5α without the plasmid is not available in this training dataset, the model may compensate by increasing the cell death parameter (D), when in reality, the decline could be attributed to plasmid loss. Further experimental validation and model refinement are necessary to assess the biological plausibility of these parameter estimates and to improve the model's representation of long-term population dynamics.

Application of a bayesian approach to study plasmid population dynamics

The third and perhaps most significant novelty of this study is the adoption of MCMC and Bayesian approach for estimating and employing parameters in modeling plasmid conjugation dynamics. Historically, parameters were measured individually, often without quantifying the level of confidence, and models that used these parameters seldom reported the certainty of their predictions. The use of MCMC and Bayesian inference allows for simultaneous extraction of parameters from experimental data along with a quantifiable degree of certainty[23]. This method introduces greater flexibility in experimental design, allowing any time course data to inform the determination of parameter posterior distributions. Furthermore, the parameter distribution data informs the selection of the most informative experimental setups for parameter estimation and model prediction. Our study shares this core methodological idea with that of Herman et al. (2011)[41], who applied MCMC-based Bayesian inference to study the molecular-scale regulation of RK2 plasmid genes. While their study explored the effects of parameter uncertainty on gene regulation dynamics, our focus on population dynamics demonstrates that the MCMC approach is versatile and valuable across multiple biological scales. To our knowledge, this is the first application of MCMC to implement a Bayesian approach for studying plasmid population dynamics. Our approach's ability to concurrently estimate the distribution of all parameters and make predictions from any given dataset increases the versatility in experimental design and data utilization. For instance, it could facilitate parameter estimation from in situ conjugation data where researchers might have limited control over the experimental conditions[34–37]. Understanding the interplay between raw data and parameter estimation also reveals the system's robustness to parametric variations, indicating that even substantial fluctuations in parameters like plasmid and cell loss rates may have minimal impact on the observable dynamics of subpopulations. Thus, we may conclude that the system exhibits resilience to changes in certain parameter values, suggesting that microscopic alterations, such as mutations affecting plasmid conjugation, might exert negligible effects on the broader subpopulation dynamics.

While both Bayesian and Frequentist approaches can estimate parameters to fit training datasets, the Bayesian approach offers distinct advantages in capturing the complexities of biological systems. As illustrated in Figure S10, the geometric mean of the Bayesian ensemble closely aligns with experimental data for testing datasets, particularly in scenarios involving long-term dynamics. There are at least a few possible explanations for superior performance of Bayesian ensemble parameters over a single best-fit parameter set. First, the ensemble accounts for parameter uncertainty by averaging predictions from multiple plausible parameter sets, which helps to generalize the model and prevent overfitting to training data. This model averaging leads to more reliable predictions in testing conditions. Second, the ensemble approach provides flexibility in handling complex and multimodal parameter distributions, capturing diverse biological dynamics that a single best-fit parameter set might miss, especially in scenarios with long-term or varying experimental conditions. As illustrated in Figure S11, the appearance of bimodal posterior distributions for parameters μ_{61} points to two sets of parameter combinations that are both compatible with the training data but lead to different predicted dynamics under testing conditions. Similar findings were reported by Herman et al. (2011)[41], where bimodal distributions of parameters in their model of RK2 plasmid regulation corresponded to qualitatively different predictions for the biological system. Following their approach, we explored the two parameter sets associated with each peak, and our results indicate that the lower μ_{61} values align better with long-term population dynamics. This suggests that one set of parameter values is more biologically reasonable than the other. This finding underscores a key advantage of the Bayesian approach: its ability to identify and analyze multiple plausible parameter sets, rather than converging on a single best-fit value[23]. Such flexibility is crucial when dealing with complex biological systems that may exhibit multiple, competing behaviors under varying conditions.

Limitations and future directions

Despite the advantages of the Bayesian approach, several important limitations must be considered. One key challenge is the significantly higher computational power and time required for MCMC sampling and analysis of resulting parameter distributions. This complexity makes the Bayesian approach less practical for models with numerous parameters or datasets that require frequent updates. Additionally, if the MCMC algorithm does not thoroughly explore the entire parameter space, the resulting posterior distributions may be incomplete or biased, leading to potentially misleading conclusions. Regardless of whether one uses Bayesian or traditional frequentist approaches, the quality and quantity of input data and the appropriateness of the model structure are crucial. In this study, a key limitation in our training data was the inability to measure certain subpopulations, resulting in gaps in the data. Furthermore, there was significant variability in population levels across experimental replicates, particularly in growth dynamics data. Our model is also highly simplified, as it does not account for changes in growth and conjugation rates over time as cultures progress through different growth phases[42]. These limitations are not specific to the Bayesian approach but collectively restrict our ability to accurately estimate parameter values and make reliable predictions.

One important consideration in using mass action kinetics to model conjugation is the assumption of a well-mixed, spatially homogeneous environment. However, in our study, conjugation experiments were performed using a filter mating assay, where donor and recipient populations were spatially structured on a surface. This introduces localized interactions that deviate from traditional well-mixed assumptions. Simonsen (1990)[43] and Malwade et al. (2017)[19] demonstrated that, under certain conditions, spatially structured conjugation on filter can still be reasonably well-approximated by mass action kinetic models, especially at high cell densities. These studies found that when bacterial cells form a confluent layer on the filter surface, the dynamics of plasmid transfer closely resemble those observed in well-mixed liquid cultures. Nevertheless, caution must be exercised when interpreting our model's results, as spatial effects can significantly influence conjugation rates, especially at lower cell densities or in non-confluent regions. It is possible that at such a low donor-to-recipient ratio as in some of our experiments, traditional mass-action kinetics may be insufficient for elucidating the mechanisms of

solid-phase conjugation on agar surfaces[17,43]. Thus, while mass action kinetics provide a useful approximation, more detailed spatial models may be required to capture all aspects of conjugation dynamics under filter mating conditions.

Future studies should aim to refine and expand upon the methodologies applied in this research. First, MCMC algorithm can further be improved. For example, by running MCMC for a greater number of steps and employing more advanced algorithms, we can minimize the risk of becoming trapped in local maxima[44]. Additionally, integrating prior distributions from existing literature or new molecular studies could improve the calculation of posterior distributions[38]. A weighted error function could also be utilized to prioritize data. For example, short term conjugation may contain more information than long-term dynamics where subpopulations barely change after the first few days and the algorithms merely try to fit a constant lines. Second, there should be an in-depth exploration of the relationship between the training dataset and the MCMC's ability to determine posterior distribution of parameters. For example, one could investigate how the number of repeats, standard deviation across repeats, and the numbers and duration of time points influence the quality of posterior distribution estimation and prediction capabilities. Third, we should identify the hidden mechanisms causing discrepancies between experimental results and model predictions. For example, we currently assume that donor remain active all the time and conjugative transfer remain constant. We know that this is not always the case. Growth phase and microenvionment around cells can affect conjugative transfer[38,42,45]. The question is to what extend this effect has impact on overall plasmid population dynamics. Fourth, we should attempt to bridge the macroscopic population dynamic model with a microscopic model at the level of gene regulatory networks. The exploration of parameter correlations could be expanded to understand how plasmid population dynamics can be linked to the dynamic behaviors of genetic elements and gene networks. We could attempt to estimate the posterior parameter distribution of gene expression and then apply this to fit the macroscopic experimental data on cell population dynamics. Such insights would not only enrich our comprehension of multi-scale phenomena—from genetics to the evolution and ecology of microbes—but also inform our experimental design and data collection strategies[10,46,47]. This would enable the development of more robust predictive models for applications such as microbiome engineering and strategies to combat antibiotic resistance related to horizontal gene transfer (HGT).

Conclusion

We have introduced and applied a novel approach for the extraction and analysis of parameters governing plasmid spread dynamics within cell populations. Utilizing the Markov Chain Monte Carlo (MCMC) method, we were able to simultaneously derive a set of parameters from experimental observations. This innovative approach enabled us to evaluate the precision of our parameter estimations and the reliability of our predictions. Our findings underscore the necessity of long-term experiments for adequately constraining parameters, thus enabling accurate predictions concerning long-term dynamics. This underscores the importance of conducting mating experiments over varied timescales. This study also represents the first to document the short and long-term dynamics of the mini-RK2 plasmid in a simple *E. coli* population across varied initial donor-recipient ratios and timescales. The adoption of this new parameter estimation and analysis methodology has provided deeper insight into the certainties and limitations inherent in our current experimental setups and analytical techniques. Future research will necessitate broader experimental setups to sufficiently constrain the model for enhanced explanation and prediction capabilities. Furthermore, employing a simplified and standardized conjugation system like mini-RK2 could facilitate the exploration of the function of each genetic element within the system and its relationship to the overall observed plasmid population dynamics. This approach holds potential for application to other conjugation systems, mobile genetic elements (MGEs), or infection models, offering a promising avenue for advancing our understanding of microbial dynamics and antibiotic resistance spread.

Material & methods

Bacteria, plasmids and growth media

E. coli DH5α and plasmids used in this study are listed in Table 1. Plasmid X61 and X13 were transferred to *E. coli* DH5α via CCMB80 chemical transformation to be used as donor and recipient host cells, respectively. Selection was carried out on Luria-Bertani (LB) agar supplemented with the appropriate antibiotics: kanamycin (Km^R: 50 mg/mL) and chloramphenicol (Cm^R: 25 mg/mL).

Strains/Plasmids	Relevant characteristics	Source/Reference
E. coli strains		
DH5α (Donor and recipient cells)	hsdR17(rK – mK þ) F– mcr1 Δ(mrr-hsrRMS-mcrBC) 80(lacZΔM15) ΔlaX74 recA1 endA1 araD139 Δ(ara, leu)7697 galU galK rpsL nupG	Thermo Fisher Scientific
Plasmid		
X013	p1008, Ori ColE1, Cm^r	provided by Dr. Drew Endy, Endy lab
X061	pMATINGα-msfGFP, PEM7 → msfGFP, Ori RK2, Tra1 and Tra2 gene, Km^r	[31]

Table 1. Bacterial strains and plasmids used for the conjugation in this study.

Mating assay

Overnight cultures of donor and recipient strains were diluted 1:100 in fresh media and re-grown back to exponential phase for 2–3 h in a 200 rpm shaking incubator at 37 °C. Following incubation, the cultures were washed three times with 1 mL of phosphate buffer saline (PBS) to eliminate residual antibiotics. After removing the supernatant, the cellular pellet was resuspended and adjusted to an OD600 of 0.3 using PBS, approximately half of OD600 at an early exponential phase. Donors and recipients were combined at ratios of 1:1 or 1:1000 as indicated. Ten microliters of the mixture were applied to a 3 × 3 mm nitrocellulose membrane placed on LB agar plates. Subsequently, the plates were incubated at 37 °C for the specified durations. Following filter mating, the nitrocellulose membrane was resuspended in 1 ml of PBS through gentle pipetting or vortexing. For long term conjugation experimenting lasting for multiple days, we refreshed media every 24 h. Specifically, once a day, mating samples were resuspended from nitrocellulose membrane in 1 ml PBS. Then, ten microliters of resuspended sample were dropped on a new nitrocellulose membrane on fresh LB agar.

Cell quantification

For each experiment, all mixtures were serially diluted, and subsequently, ten microliters of the mixtures was dropped on selective agar plates to quantify the numbers of donors (D), recipients (R), and transconjugants (TC). Plasmid transfer frequency (f) was determined by counting colonies and calculated using the formula $f = TC/R$.

Computational model and parameter estimation

Mass action kinetic model of plasmid conjugation and Metropolis Monte Carlo algorithm were implemented in python on google colab platform (see detail and codes in supplementary). Data visualisation and analysis were performed using Matplotlib and Seaborn package in python. Generative AI (ChatGPT) was used for guiding python programming and revising manuscript.

We used ordinary differential equations (ODEs) to model the population dynamics of four cell types: DH5α (DH5a), DH5α with X13 (DH5a-X13), DH5α with X61 (DH5a-X61), and DH5α with both X13 and X61 (DH5a-X13-X61). The model includes eight kinetic parameters: the transfer rate of X61 (η), the growth rate of DH5α (μ), the growth rate of DH5α-X13 (μ_{13}), the growth rate of DH5α-X61 (μ_{61}), the growth rate of DH5α-X13-X61 (μ_{1361}), the loss rate of X13 (κ_{13}), the loss rate of X61 (κ_{61}), and the combined rate of cell death and dilution (D). The complete set of equations and the simulation workflow are shown in Figures S1 and S2. These ODEs were used to generate synthetic data (points in Fig. 2A), simulate population time courses (curves in Fig. 2A A), and calculate the posterior probability of parameters (Figs S3 and S4).

The default parameter values used to generate the synthetic dataset were: η = 1E-10 / cell / hr, μ = 1.4 / hr, $\mu_{13} = \mu_{61} = 1.3$ / hr, $\mu_{1361} = 1.2$ / hr, and D = 0.1 / hr. The carrying capacity of the environment (the maximum total cell count that allows continued population growth) was set at 1E + 9 cells. These values were manually selected to produce population time courses resembling those observed experimentally in short-term conjugation (Fig. 3A-i) while remaining within an order of magnitude of previously reported values[18,38]. To generate synthetic data points, log-normal noise with a standard deviation of 1 was added to the simulated time course values. This noise distribution and standard deviation were chosen to ensure that the synthetic data points closely matched the experimentally observed data points in the short-term conjugation setup (compare data points in Fig. 2A-i with Fig. 3A-i).

For a given training dataset, whether synthetic or experimental, we used MCMC to obtain an ensemble of parameters. The process involves performing random walks across the parameter space while calculating the posterior probability of each parameter set. This posterior probability guides the direction of the random walk. The resulting parameter ensemble is a collection of parameter sets visited during the random walk, with sets that have higher posterior probabilities being visited more frequently and thus appearing more often in the ensemble (Fig. S5). The posterior probability was calculated based on the likelihood (i.e., how well a parameter set explains the training dataset, indicated by the similarity between the simulated and training data time courses, Fig. S3) and the prior (existing knowledge about parameter values).

For our Bayesian inference, we used non-negative uniform prior distributions (Fig. S4). This means that the prior contributes no additional information to the calculation of the posterior, other than the hard constraint that all parameters must be greater than or equal to zero. The rationale behind this choice is that, although the RK2 plasmid and its derivatives have been extensively studied, there is no well-established information about the distribution of these specific parameters. Even for parameters such as plasmid transfer rates and cell growth rates, which have been previously measured, those measurements were conducted under conditions quite different from ours—such as using different host cells or media conditions. Therefore, it was more appropriate to use an unbiased prior distribution and allow the likelihood to primarily determine the posterior values.

We selected starting points for the random walk from a log-uniform distribution centered around the default parameter values. The minimum and maximum values for these distributions were as follows: (1E-13, 1E-13) / cell / hr for η, (1E-1, 1E+1) / hr for all four growth parameters, (1E-7, 1E+1) for both plasmid loss parameters, and (1E-2, 1E+1) for D. These ranges were chosen arbitrarily to encompass at least a few orders of magnitude around the default values. We initiated the random walks from 10 randomly chosen starting points and continued them for 20,000 steps until the parameter and posterior values appeared to converge. From steps 15,000 to 20,000, 200 parameter sets were randomly selected to form the parameter ensemble. This ensemble was then used for simulating population dynamics (as shown in Fig. 2A, 3A) and for parameter analyses (Fig. 2B,C, 3B,C).

Data availability

The authors confirm that the data supporting the findings of this study are available within the article [and/or] its supplementary materials.

Received: 7 July 2024; Accepted: 9 December 2024
Published online: 03 March 2025

References

1. Frost, L. S., Leplae, R., Summers, A. O. & Toussaint, A. Mobile genetic elements: the agents of open source evolution. *Nat. Rev. Microbiol.* **3**, 722–732 (2005).
2. Ghaly, T. M. & Gillings, M. R. Mobile DNAs as ecologically and evolutionarily independent units of life. *Trends Microbiol.* **26**, 904–912 (2018).
3. Haudiquet, M., De Sousa, J. M., Touchon, M. & Rocha, E. P. C. Selfish, promiscuous and sometimes useful: how mobile genetic elements drive horizontal gene transfer in microbial populations. *Philos. Trans. R Soc. B Biol. Sci.* **377**, (2022).
4. Zhu, S., Hong, J. & Wang, T. Horizontal gene transfer is predicted to overcome the diversity limit of competing microbial species. *Nat. Commun.* **15**, 1–9 (2024).
5. Partridge, S. R., Kwong, S. M., Firth, N. & Jensen, S. O. Mobile genetic elements associated with antimicrobial resistance. *Clin. Microbiol. Rev.* **31**, (2018).
6. Leclerc, Q. J., Lindsay, J. A. & Knight, G. M. Mathematical modelling to study the horizontal transfer of antimicrobial resistance genes in bacteria: current state of the field and recommendations. *J. R Soc. Interface* **16**, (2019).
7. Sheth, R. U., Cabral, V., Chen, S. P. & Wang, H. H. Manipulating bacterial communities by in situ Microbiome Engineering. *Trends Genet.* **32**, 189–200 (2016).
8. Bober, J. R., Beisei, C. L. & Nair, N. U. Synthetic Biology approaches to engineer probiotics and members of the human microbiota for Biomedical Applications. *Annu. Rev. Biomed. Eng.* https://doi.org/10.1016/j.physbeh.2017.03.040 (2018).
9. Marsh, J. W., Kirk, C. & Ley, R. E. Toward Microbiome Engineering: expanding the repertoire of genetically tractable members of the human gut Microbiome. *Annu. Rev. Microbiol.* **77**, 427–449 (2023).
10. Hernández-Beltrán, J. C. R., San Millán, A. & Fuentes-Hernández, A. Peña-Miller, R. Mathematical Models of Plasmid Population Dynamics. *Front. Microbiol.* **12**, 1–18 (2021).
11. Stewart, F. M. & Levin, B. R. The Population Biology of bacterial plasmids: a PRIORI conditions for the existence of Conjugationally transmitted factors. *Genetics* **87**, 209–228 (1977).
12. Levin, B. R., Stewart, F. M. & Rice, V. A. The kinetics of conjugative plasmid transmission: fit of a simple mass action model. *Plasmid* **2**, 247–260 (1979).
13. Levin, B. R. & Stewart, F. M. The population biology of bacterial plasmids: a priori conditions for the existence of mobilizable nonconjugative factors. *Genetics* (1980).
14. Simonsen, L., Gordon, D. M., Art^, F. M. S. & In^, B. R. L. estimating the rate of plasmid transfer: an end-point method. *J. Gen. Microbiol.* **136**, 2319–2325 (1990).
15. Krone, S. M., Lu, R., Fox, R., Suzuki, H. & Top, E. M. Modelling the spatial dynamics of plasmid transfer and persistence. *Microbiology* **153**, 2803–2816 (2007).
16. Seoane, J. et al. An individual-based approach to explain plasmid invasion in bacterial populations. *FEMS Microbiol. Ecol.* **75**, 17–27 (2011).
17. Zhong, X., Droesch, J., Fox, R., Top, E. M. & Krone, S. M. On the meaning and estimation of plasmid transfer rates for surface-associated and well-mixed bacterial populations. *J. Theor. Biol.* https://doi.org/10.1016/j.jtbi.2011.10.034 (2012).
18. Lopatkin, A. J. et al. Persistence and reversal of plasmid-mediated antibiotic resistance. *Nat. Commun.* **8**, (2017).
19. Malwade, A., Nguyen, A., Sadat-Mousavi, P. & Ingalls, B. P. Predictive modeling of a batch filter mating process. *Front. Microbiol.* **8**, 1–11 (2017).
20. Wang, T. & You, L. The persistence potential of transferable plasmids. *Nat. Commun.* **11**, 1–10 (2020).
21. Huisman, J. S. et al. Estimating plasmid conjugation rates: a new computational tool and a critical comparison of methods. *Plasmid* **121**, 102627 (2022).
22. Kosterlitz, O. & Huisman, J. S. Guidelines for the estimation and reporting of plasmid conjugation rates. *Plasmid* **126**, 102685 (2023).
23. Linden, N. J., Kramer, B. & Rangamani, P. *Bayesian parameter estimation for dynamical models in systems biology*. *PLoS Comput. Biol.* **18**, (2022).
24. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. Equation of state calculations by fast Computing machines. *J. Chem. Phys.* **21**, 1087–1092 (1953).
25. Hastings, W. K. Monte carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109 (1970).
26. Tierney, L. Markov Chains for exploring posterior distributions. *Annu. Stat.* **22**, 1701–1762 (1994).
27. Mathews, J. D., McCaw, C. T., McVernon, J., McBryda, E. S. & McCaw, J. M. A biological model for influenza transmission: pandemic planning implications of asymptomatic infection and immunity. *PLoS One* **2**, (2007).
28. Keersmaekers, N., Ogunjimi, B., Van Damme, P., Beutels, P. & Hens, N. An ODE-based mixed modelling approach for B- and T-cell dynamics induced by varicella-zoster virus vaccines in adults shows higher T-cell proliferation with Shingrix than with Varilrix. *Vaccine* **37**, 2537–2553 (2019).
29. Valderrama-Bahamóndez, G. I. & Fröhlich, H. MCMC techniques for parameter estimation of ODE based models in Systems Biology. *Front. Appl. Math. Stat.* **5**, 1–10 (2019).
30. Rossini, L., Bruzzone, O. A., Speranza, S. & Delfino, I. Estimation and analysis of insect population dynamics parameters via physiologically based models and hybrid genetic algorithm MCMC methods. *Ecol. Inf.* **77**, 102232 (2023).
31. Aparicio, T., Silbert, J., Cepeda, S. & de Lorenzo, V. Propagation of recombinant genes through complex microbiomes with synthetic mini-RP4 plasmid vectors. *BioDesign Res.* **1–15**, 9850305. https://doi.org/10.34133/2022/9850305 (2022).
32. Simon, R., Priefer, U. & Puhler, A. A. Broad Host Range Mobilization System for in Vivo Genetic Engineering: Transposon Mutagenesis in Gram negative Bacteria. *Nat. Biotechnol.* **1**, 784–791 (1983).
33. Silbert, J., Lorenzo, V. & Aparicio, T. Refactoring the conjugation machinery of promiscuous plasmid RP4 into a device for conversion of Gram-negative isolates to hfr strains. *ACS Synth. Biol.* **10**, 690–697 (2021).
34. Klümper, U. et al. Broad host range plasmids can invade an unexpectedly diverse fraction of a soil bacterial community. *ISME J.* **9**, 934–945 (2015).
35. Fu, J. et al. Aquatic animals promote antibiotic resistance gene dissemination in water via conjugation: role of different regions within the zebra fish intestinal tract, and impact on fish intestinal microbiota. *Mol. Ecol.* **26**, 5318–5333 (2017).
36. Fan, X. T. et al. Fate of antibiotic resistant pseudomonas putida and broad host range plasmid in natural soil microcosms. *Front. Microbiol.* **10**, 1–10 (2019).
37. Ronda, C., Chen, S. P., Cabral, V., Yaung, S. J. & Wang, H. H. Metagenomic engineering of the mammalian gut microbiome in situ. *Nat. Methods.* **16**, 167–170 (2019).
38. Sheppard, R. J., Beddis, A. E. & Barraclough, T. G. The role of hosts, plasmids and environment in determining plasmid transfer rates: a meta-analysis. *Plasmid* **108**, 102489 (2020).
39. Pansegrau, W. & Lanka, E. Enzymology of DNA transfer by conjugative mechanisms. *Prog Nucleic Acid Res. Mol. Biol.* **54**, 197–251 (1996).
40. Wang, T. et al. Horizontal gene transfer enables programmable gene stability in synthetic microbiota. *Nat. Chem. Biol.* https://doi.org/10.1038/s41589-022-01114-3 (2022).

41. Herman, D., Thomas, C. M. & Stekel, D. J. Global transcription regulation of RK2 plasmids: a case study in the combined use of dynamical mathematical models and statistical inference for integration of experimental data and hypothesis exploration. *BMC Syst. Biol.* **5**, 119 (2011).
42. Sysoeva, T. A., Kim, Y., Rodriguez, J., Lopatkin, A. J. & You, L. Growth-stage-dependent regulation of conjugation. *AIChE J.* **66**, 1–10 (2020).
43. Simonsen, L. Dynamics of plasmid transfer on surfaces. *J. Gen. Microbiol.* **136**, 0–1 (1990).
44. Ballnus, B. et al. Comprehensive benchmarking of Markov chain Monte Carlo methods for dynamical systems. *BMC Syst. Biol.* **11**, 1–18 (2017).
45. Hunter, P. R., Wilkinson, D. C., Catling, L. A. & Barker, G. C. Meta-analysis of experimental data concerning antimicrobial resistance gene transfer rates during conjugation. *Appl. Environ. Microbiol.* **74**, 6085–6090 (2008).
46. Sheppard, R. J., Barraclough, T. G. & Jansen, V. A. A. The evolution of plasmid transfer rate in bacteria and its effect on plasmid persistence. *Am. Nat.* **198**, 473–488 (2021).
47. Bethke, J. H. et al. Vertical and horizontal gene transfer tradeoffs direct plasmid fitness. *Mol. Syst. Biol.* **19**, 1–10 (2023).

Acknowledgements

This study is financially supported by the Air Force Office of Scientific Research, USA, under award number FA2386-23-1-4017 and Ministry of Higher Education, Science, Research and Innovation, Thailand, under award number RGNS 63-131. We would like to thank Dr. Sudarat Chadsuthi from the Department of Physics, Faculty of Science, Naresuan University, for her valuable comment on the manuscript. We would also like to thank Faculty of Medical Science, Naresuan University for supporting all facilities.

Author contributions

S. K., C. J., C. J. conducted experimental work; S. K. and C. J. (Jaichuen) analyzed data. C. J. (Jaichuen) and P. S. wrote the manuscript. All authors reviewed the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-82799-5.

Correspondence and requests for materials should be addressed to P.S.

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

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