Bayesian inference elucidates the varying dynamics of alternative end joining mechanisms.

M. Woods and C.P. Barnes

SUPPLEMENTARY APPENDIX

1 Biochemical kinetic models, initial conditions and priors

Our model consists of two reactions for each repair mechanism, giving a maximum total of six reactions for each data set:

$$x_1^d + E_1^d \xrightarrow{K_1^d} x_2^d \tag{1}$$

$$x_2^d \xrightarrow{K_1^d} \emptyset \tag{2}$$

$$x_1^d + E_2^d \xrightarrow{K_2^d} x_3^d \tag{3}$$

$$x_3^d \xrightarrow{K_2^d} \emptyset \tag{4}$$

$$x_1^d + E_3^d \xrightarrow{K_3^d} x_4^d \tag{5}$$

$$x_4^d \xrightarrow{K_3^d} \emptyset \tag{6}$$

where x_i^d, for $i \in \{1,2,3,4\}$ is the state of a double strand break (DSB) for dataset d. K_i^d and E_i^d for $i \in \{1,2,3\}$ are the parameter and recruitment protein for dataset d and repair mechanism i. Each model d has three conservation equations

$$E_1^d + x_2^d = C_1^d \tag{7}$$

$$E_2^d + x_3^d = C_2^d \tag{8}$$

$$E_3^d + x_4^d = C_3^d, \tag{9}$$

where C_i^d is the total amount of recruitment protein for dataset d and repair mechanism i. If for any mechanism, the first protein to bind is repressed, then we remove all reactions corresponding to that mechanism. If a protein downstream of the first protein to bind is repressed, then we remove the reaction corresponding to end ligation for that mechanism. When the reactions are taken to be deterministic, the wild type system can be described by the following set of nonlinear ordinary differential equations.

$$\frac{dx_1}{dt} = V - x_1(K_1E_1 + K_2E_2 + K_3E_3) \tag{10}$$

$$\frac{dx_2}{dt} = K_1(E_1x_1 - 1) \tag{11}$$

$$\frac{dx_3}{dt} = K_2(E_2x_1 - 1) \tag{12}$$

$$\frac{dx_4}{dt} = K_3(E_3x_1 - 1) \tag{13}$$

This system of equations can be solved numerically, however we are interested in making predictions for single cell data and so we solve the stochastic model. To determine the most probable set of parameters to give rise to the experimental data - termed the posterior distribution - we apply approximate Bayesian computation sequential Monte Carlo.

2 Approximate Bayesian computation

In the Bayesian framework, we are interested in the posterior distribution $\pi_\epsilon(\theta, x|y)$, where θ is a vector of parameters and $x|y$ is the simulated data conditioned on the experimental data. To obtain samples from the posterior distribution we must condition on the data y and this is done via an indicator function $I_{\mathscr{A}_{y,\epsilon}}(x)$. We then have

$$\pi_\epsilon(\theta, x|y) = \frac{\pi(\theta)f(x|\theta)I_{\mathscr{A}_{y,\epsilon}}(x)}{\int_{\mathscr{A}_{y,\epsilon}\times\Theta}\pi(\theta)f(x|\theta)dxd\theta},$$

where $\mathscr{A}_{y,\epsilon} = \{x \in \mathscr{D} : \rho(x,y) \leq \epsilon\}$, $\rho : \mathscr{D}\times\mathscr{D} \rightarrow R^+$ is a distance function comparing the simulated data to the observed data and π_ϵ is an approximation to the true posterior distribution. This approximation is obtained via an algorithm that repetitively samples from the parameter space until ϵ is small such that the resulting approximate posterior, π_ϵ, is close to the true posterior. There are different algorithms that can be applied to obtain this approximation. We use the method of sequential importance sampling or more specifically ABC SMC, which is implemented in the software ABC-SysBio. For further details on the algorithms available in ABC-SysBio see [1–4].

3 The hierarchical model

We have modified the software ABC-SysBio to include an option to perform ABC SMC on a hierarchical model structure. In our hierarchical framework, we wish to obtain K_i^d for three processes $i \in \{1,2,3\}$ and eight datasets $d \in \{1,...,8\}$. These parameters are in some sense nuisance variables, with “true parameters" (or parameters of interest) μ_{1-5}. The μ_{1-4} represent the means of four lognormal distributions and μ_5 the variance. The K_i^d are drawn from these population level distributions. In this case, the joint density can be written

$$p(y,K,\mu) = p(y|K,\mu)p(K|\mu)p(\mu) \tag{14}$$

where Bayes rule becomes

$$p(\mu|y) = \frac{p(\mu)\int p(y|K,\mu)p(K|\mu)}{p(y)}. \tag{15}$$

This is the posterior of the hyper parameters given the data, y. The integral indicates that we sum over (marginalise) the K values. We can include this into ABC by simulating data x^* using the following scheme:

$$\mu \sim U(\alpha,\beta)$$

$$K \sim LN(\mu,\sigma)$$

$$x^* \sim f(x|K)$$

In our study, $f(x|K)$ is the data generating model, and is the solution to the reaction systems presented in section 1. We perturb only the μ but the distance is calculated on the simulation using the sampled K values.

The model prior distributions for the hyper parameters were fixed across all datasets and had the following limits:

Hierarchical priors: $\mu_1 \sim U(1,4)$, $\mu_2 \sim U(-4,-1)$, $\mu_3 \sim U(-2,4)$, $\mu_4 \sim U(-4,-1)$, $\mu_5 \sim U(0.05,0.9)$.

The total amount of protein and initial conditions were set according to the data.

constant: $C_1 = 700$, $C_2 = 700$, $C_3 = 700$, $C_4 = 2800$, $C_5 = 2800$, $C_6 = 1906$, $C_7 = 1814$ and $C_8 = 1128.7$.

Recruitment protein initial conditions: $E_1^1(0) = 700$, $E_2^1(0) = 700$, $E_3^1(0) = 700$, $E_1^2(0) = 700$, $E_2^2(0) = 700$, $E_3^2(0) = 700$, $E_1^3(0) = 700$, $E_2^3(0) = 700$, $E_3^3(0) = 700$, $E_1^4(0) = 2800$, $E_2^4(0) = 2800$, $E_3^4(0) = 2800$, $E_2^5(0) = 2800$, $E_3^5(0) = 2800$, $E_2^6(0) = 1906$, $E_3^6(0) = 1906$, $E_2^7(0) = 1814$, $E_1^8(0) = 1128.7$, $E_2^8(0) = 1128.7$.

State vector initial conditions: $x_1^1(0) = 700$, $x_2^1(0) = 0$, $x_3^1(0) = 0$, $x_4^1(0) = 0$, $x_1^2(0) = 700$, $x_2^2(0) = 0$, $x_3^2(0) = 0$, $x_4^2(0) = 0$, $x_1^3(0) = 700$, $x_2^3(0) = 0$, $x_3^3(0) = 0$, $x_4^3(0) = 0$, $x_1^4(0) = 2800$, $x_2^4(0) = 0$, $x_3^4(0) = 0$, $x_4^4(0) = 0$, $x_1^5(0) = 2800$, $x_2^5(0) = 0$, $x_3^5(0) = 0$, $x_4^5(0) = 0$, $x_1^6(0) = 1906$, $x_2^6(0) = 0$, $x_3^6(0) = 0$, $x_4^6(0) = 0$, $x_1^7(0) = 1814$, $x_2^7(0) = 0$, $x_3^7(0) = 0$, $x_4^7(0) = 0$, $x_1^8(0) = 1128.7$, $x_2^8(0) = 0$, $x_3^8(0) = 0$, $x_4^8(0) = 0$.

4 The cumulative number of DSBs

Proportions of DSBs repaired by each mechanism are estimated by calculating the cumulative number of DSBs that enter each individual pathway with the density weighted integral,

$$G_j^d(t)\Big|_{t=T} = -\int_0^{t=T} \sigma_j^d(t) X^{d'}(t)\,dt, \tag{16}$$

$$\sigma_j^d(t) = \frac{x_{j+1}^d(t)}{\sum_{k\in\{2,3,4\}} x_k^d(t)}. \tag{17}$$

Equation 16 is the product of the change in total DSBs and density $\sigma_j(t)$, $j \in \{1,2,3\}$ of DSBs in repair mechanism j integrated over time. This contribution to the overall repair can be used to predict the proportion of DSBs $P_j \forall j \in R$ repaired by each mechanism:

$$P_j = G_j(T)/X(T). \tag{18}$$

Where $G_j(T)$ is the total amount of DSBs repaired by mechanism j at time T.

References

[1] Toni, T., Welch, D., Strelkowa, N., Ipsen, A., and Stumpf, M. P. H. (2009) Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems.. *Journal of the Royal Society, Interface / the Royal Society,* **6**(31), 187–202.

[2] Liepe, J., Barnes, C., Cule, E., Erguler, K., Kirk, P., Toni, T., and Stumpf, M. P. (2010) ABC-SysBio–approximate Bayesian computation in Python with GPU support. *Bioinformatics,* **26**(14), 1797–1799.

[3] Toni, T. and Stumpf, M. P. H. (2010) Simulation-based model selection for dynamical systems in systems and population biology.. *Bioinformatics,* **26**(1), 104–110.

[4] Liepe, J., Kirk, P., Filippi, S., Toni, T., Barnes, C. P., and Stumpf, M. P. H. (2014) A framework for parameter estimation and model selection from experimental data in systems biology using approximate Bayesian computation.. *Nature protocols,* **9**(2), 439–456.

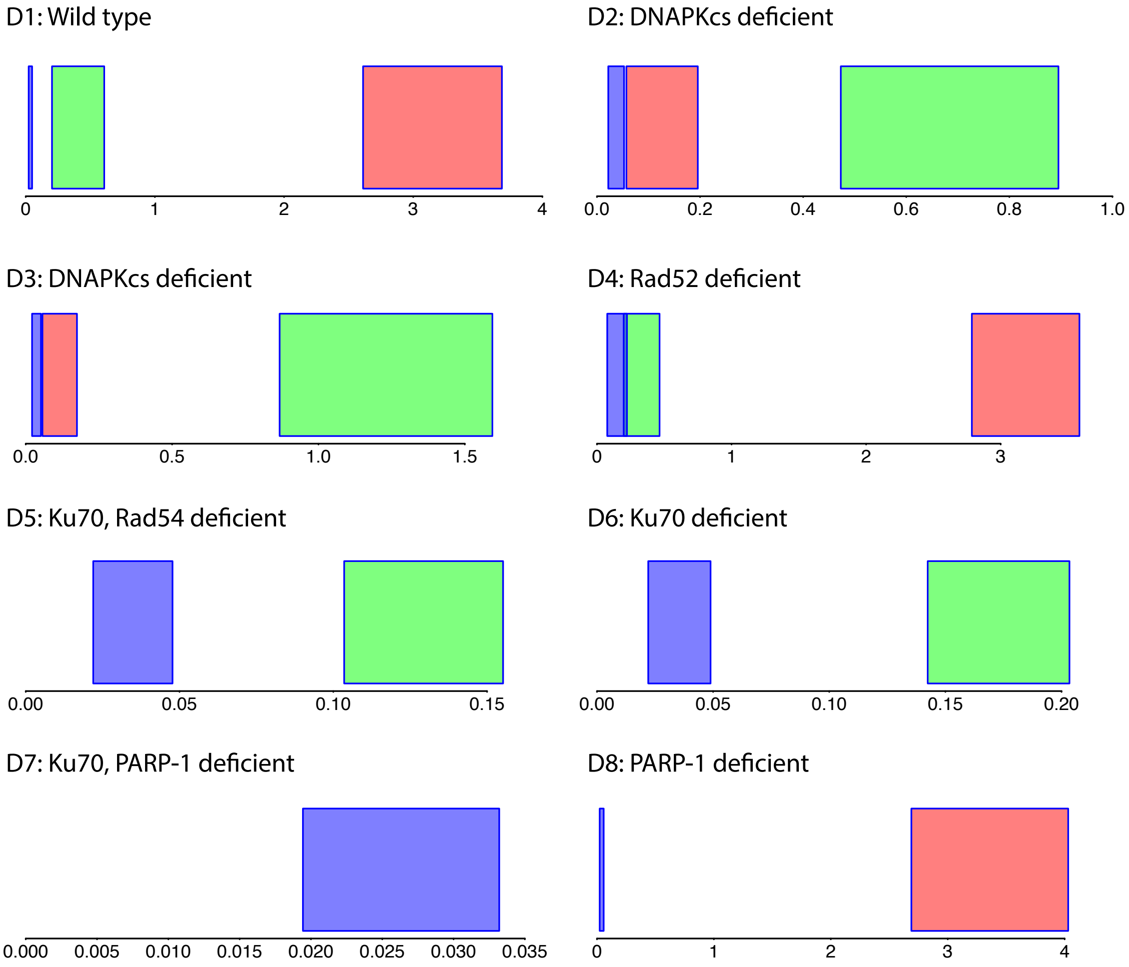

Figure S1: The interquartile range for all the parameters in each dataset. Red (fast repair), blue (slow repair) and green (intermediate repair).

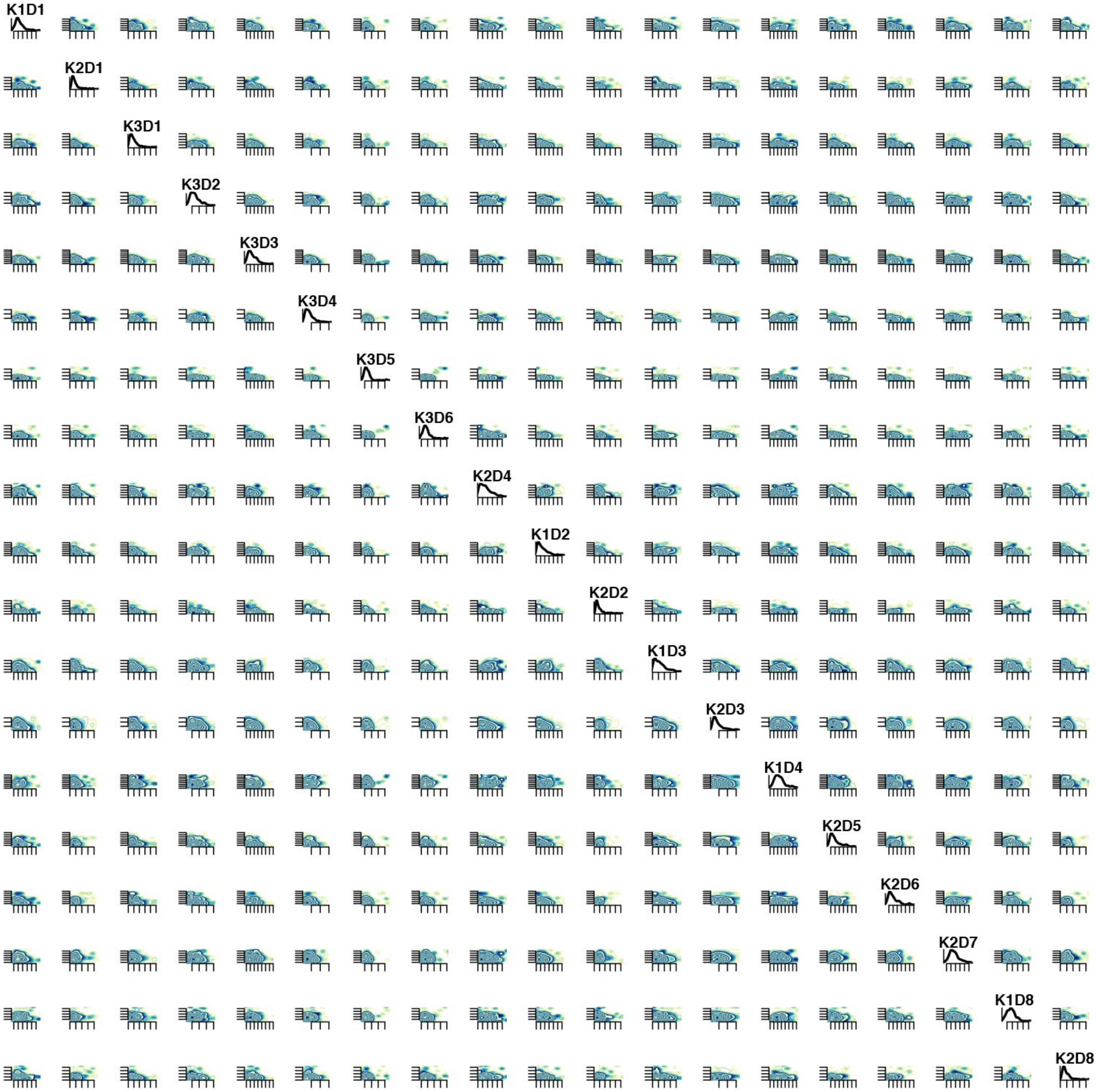

Figure S2: Posterior of the latent parameters K_i^d.