Supplementary Information

1 Methods

1.1 Analysing and selecting mathematical models using Bayesian Approach.

We attempted to describe the kinetics of host and donor cells in busulfan chimeric mice using an array of mathematical models (described in section 1.4). Due to variation in the degree of depletion of host HSCs by busulfan treatment, the level at which the fraction of donor cells stabilize in each subset varies across mice. We address this issue by normalising the donor fractions in any given B cell subsets to the donor fractions in the upstream precursor populations. We refer to this as the normalized chimerism:

$f_{\mathrm{d}} = \dfrac{\text{Donor counts}}{\text{Total counts} \times \chi}$, where χ is the fraction of cells in the precursor population that are donor-derived.

This allows us to fit a single model to data from multiple mice, that may have responded differently to busulfan treatment. Each model was fitted simultaneously to the total cell counts (N, the sum of host and donor cells), the normalised donor fraction (f_d) and the proportions of Ki67$^{\mathrm{hi}}$cells (κ) in the host and donor compartments. Cell counts were log-transformed while f_d and κ were transformed using the *logit* function to ensure that measurement errors were normally distributed.

We then used the Bayesian approach to estimate the model parameters, the errors associated with the measurements in each dataset, and a measure of the support for the model. The inputs to this procedure are (i) an expression for the joint likelihood of the data, in terms of the model parameters and the (unknown) errors associated with each dataset; and (ii) a set of distributions on both the model parameters and the magnitude of the measurement errors in each dataset, representing any insights we may have regarding their values and referred to as *priors*. The Bayesian procedure updates these priors with the likelihood, to generate posterior distributions that reflect our knowledge of these parameters in the light of the data. Strong (narrow) priors help to *regularise* a model's behaviour and prevent it from learning too much from the data – and hence guard against over-fitting.

The *posterior* distribution of parameters (θ_{post}) for a given model is estimated using Bayes' rule,

$$p(\theta_{\mathrm{post}}|y) = \frac{p(y|\theta_{\mathrm{post}}) \,.\, p(\theta_{\mathrm{prior}})}{p(y)} \tag{1}$$

where $p(y)$ is the *likelihood* of the data (averaged over the *priors*) that normalises the *posterior* such that it integrates to 1. In our analysis, *priors* are considered to be tools that improve model's ability to learn from the data and are subjected to similar standards of evaluation and revaluation like any other component of the model.

Our each model is represented as a system of ordinary differential equations (ODEs, described in details in section 1.4), which is solved numerically using the '*integrate_ode_rk45*' solver available in *Stan* programming language. During an individual iteration of the fitting process, a set of parameter values are drawn from θ_{prior} repeatedly using a MCMC based *NUTS* sampler, in order to provide probable predictions of the combined data and calculate the joint *likelihood*. The generalised strategy that we use for Bayesian analysis is depicted

below. Model specific details of *prior* distributions of parameters are described in their respective paragraphs in section 1.4.

We assumed that the suitably transformed measurements of total cell counts,normalised chimerism and proportions of Ki67^{hi} cells in host and donor compartments were all normally distributed with unknown but constant errors at each time point,

$$\begin{aligned} N &\sim \text{Normal}(\mu_N, \sigma_N), \\ f_d &\sim \text{Normal}(\mu_{fd}, \sigma_{fd}), \\ \kappa_{\text{host}} &\sim \text{Normal}(\mu_{\kappa_{\text{host}}}, \sigma_{\kappa_{\text{host}}}), \\ \kappa_{\text{donor}} &\sim \text{Normal}(\mu_{\kappa_{\text{donor}}}, \sigma_{\kappa_{\text{donor}}}). \end{aligned} \tag{2}$$

Here, μ is the mean of model predictions estimated by averaging the model predictions across the whole *posterior* distribution weighted by the *likelihood* of data at a given time-point.
The standard deviation σ indicates the uncertainty in μ, which is estimated from the model fits. The *R-stan* package in *R* software is used to interface and compile the code written for model definitions, sampling and fitting procedures in *stan* programming language.

Model comparisons

We evaluate the mathematical models studied here in two separate ways. First by running posterior-predictive checks on the same data that was used for fitting procedures and second by estimating the out-of-sample prediction error i.e the ability of the model to accurately predict new data.

Posterior-predictive checks involve replicating data (y^{rep}) by simulating the fitted model using the posterior distribution of parameters and comparing that to the observed data (y^{obs}). Since, we use the whole *posterior*, these checks include the uncertainty associated with parameter estimations, which is an advantage over non-bayesian methods. Posterior-predictive check helps assess the validity of the model by displaying systematic discrepancies between y^{rep} and y^{obs} and therefore works as a diagnostic tool in rebuilding the model. Its just a systematic way to ask whether your model makes sense in the light of observed data.

We then estimate the out-of-sample prediction error which is slightly more complicated and computationally expensive. First the log predictive density $\log\big(p(y|\theta)\big)$ (also called as the log-likelihood) of the given observation is estimated by simulating 'D' number draws from the posterior distribution for the fitted model. The whole process is repeated for each data point $(y_1, ..., y_n)$ to calculate log point-wise predictive density (*lppd*) as described in eq.3. Then the out-of-sample prediction error is estimated using the method of cross-validation, where data is divided into (i) a training sample (y_{train}), which is used to fit the model and (ii) a test sample (y_{test}), which is used to evaluate the predictive accuracy of the model.

$$\text{lppd} = \sum_{i=1}^{n} \log\left(\frac{1}{D}\sum_{d=1}^{D} p(y_i|\theta^d_{\text{post}})\right) \tag{3}$$

The advantage of using *lppd* over deviance is that it measures the difficulty in predicting observations at each unique point in the data and thus provides more accurate descriptions of the uncertainty in model predictions. Additionally we use a special case of cross-validation called 'leave-one-out' (*loo*) method to compute the out-of-sample prediction error, which substantially minimises the problem of over-fitting. (i.e. model learning too much from the data). In this method, a dataset of size 'n' $(y_1, ..., y_n)$ is partitioned into 'n' number of samples such that each test sample contains a single observation. The model is fitted on the training set

containing (n-1) observations and is tested for the prediction accuracy of the 'left-out' observation. This process is repeated for all the elements in the test sample and the out-of-sample prediction error is estimated by simulating 'D' number draws from the posterior distribution. The *loo* estimate of lppd is therefore the summed value of the log-predictive densities calculated for each element in test sample from the posterior draws fitted on the observations excluding the test sample. (eq.4).

$$\begin{aligned} \text{lppd}^{loo} &= \sum_{i=1}^{n} \log\, p(y_i|\theta_{\text{post}}^{(\text{-}i)}) \\ &= \sum_{i=1}^{n} \log\, \left(\frac{1}{D} \sum_{d=1}^{D} p(y_i|\theta_{\text{post}}^{(\text{-}i)d}) \right) \end{aligned} \tag{4}$$

Where $\theta_{\text{post}}^{(\text{-}i)d}$ characterizes D posterior simulations fitted on (n-1) observations when ith observation is left out.

We use the *loo-2.0* package in *R* to calculate the lppd^{loo} estimates for each model. The *loo-2.0* package uses Pareto-smoothed importance sampling (PSIS), which is an approximation of *loo* cross-validation that uses existing posterior draws from the model fits to estimate the lppd^{loo} values (Vehtari et al, Statistics and Computing, Aug 2016).

The lppd^{loo} estimate by itself does not provide any information about model's suitability to the data. However, it can be used to rank different models, such that the model with lowest lppd^{loo} value has the highest support. It is also used to calculate *Akaike weight*, which is the probability that a given model will explain new data better as compared to all the other models considered in the analysis. We estimate *Akaike weights* by exponentiating the negative half of Δlppd^{loo} (i.e the difference between each lppd^{loo} and the lowest lppd^{loo}) value and then standardise it by dividing by the total value for 'M' number of models. (eq. 5).

$$Aw_i = \frac{\exp(-\frac{1}{2}\,\Delta\text{lppd}_i^{loo})}{\sum_{m=1}^{M} \exp(-\frac{1}{2}\,\Delta\text{lppd}_m^{loo})} \tag{5}$$

1.2 Modelling the dynamics of chimerism in the precursor population

We describe the changes in chimerism in the source populations over time using a phenomenological function (equation 6) that reasonably depicts the shape of the timecourse of fraction of donor cells across all animals.

$$\chi(t) = \chi_{stable}\,(1 - e^{-q\,t}) \tag{6}$$

Parameters 'χ_{stable}' and 'q' are estimated by fitting the spline in equation 6 to the observed donor fractions in the source compartment.

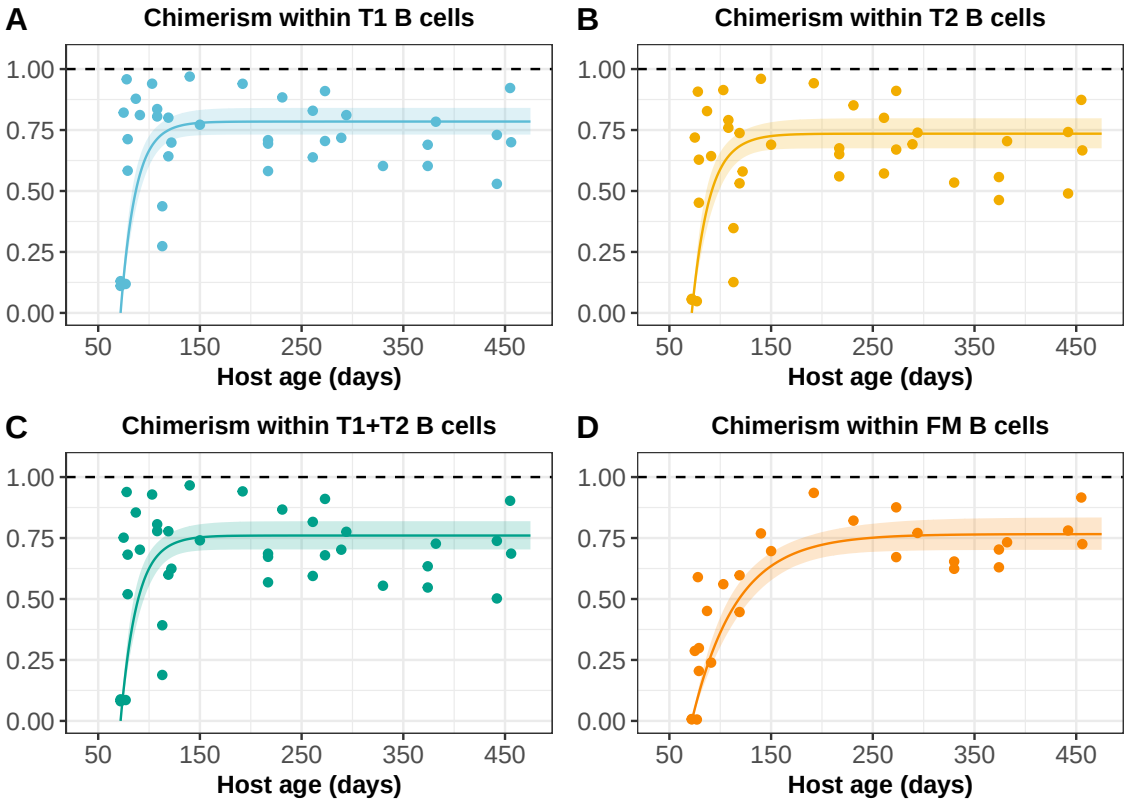

Figure 1: **The time-course of absolute chimerism within precursor compartments.** Solid lines indicate the most probable description of changes in the chimerism with host age, modelled according to equation 6. Shaded envelopes denote the uncertainty in the most probable prediction and are drawn by taking 4.5 and 95.5 percentiles intervals of model predictions made from the posterior distribution of parameters.

1.3 Modelling Source influx

We assumed that the rate of influx from source ψ, stays constant and used the size of the source compartment as a proxy to estimate the number of cells maturing into the target pool. Equation 7, is a spline that is used to describes the changes in the source over time. Parameters S_0 and ν are estimated by fitting the spline in equation 7 to the observed cell counts in the source compartment.

$$S(t) = S_0\,e^{-\nu\,t} \tag{7}$$

The proportion of source population entering the target pool (ψ) is assumed to remains constant over time and is estimated from the model fits. Daily source influx into FM and GC populations is therefore given by equation 8,

$$\begin{aligned} \phi(t) &= \psi\, S(t), \\ \phi_{\text{donor}}(t) &= \psi\, S(t)\, \chi(t), \\ \phi_{\text{host}}(t) &= \phi(t) - \phi_{\text{donor}}(t). \end{aligned} \tag{8}$$

We also explored variant of our models in which ψ varied with time ($\psi(t) = \psi_0\, e^{p\,t}$), which failed to improve on the quality of fits that we received with constant ψ. Parameters ψ_0 and 'p' are estimated from model fits to the time-course of total counts, normalised chimerism and Ki67$^{\text{hi}}$fractions of the target population.

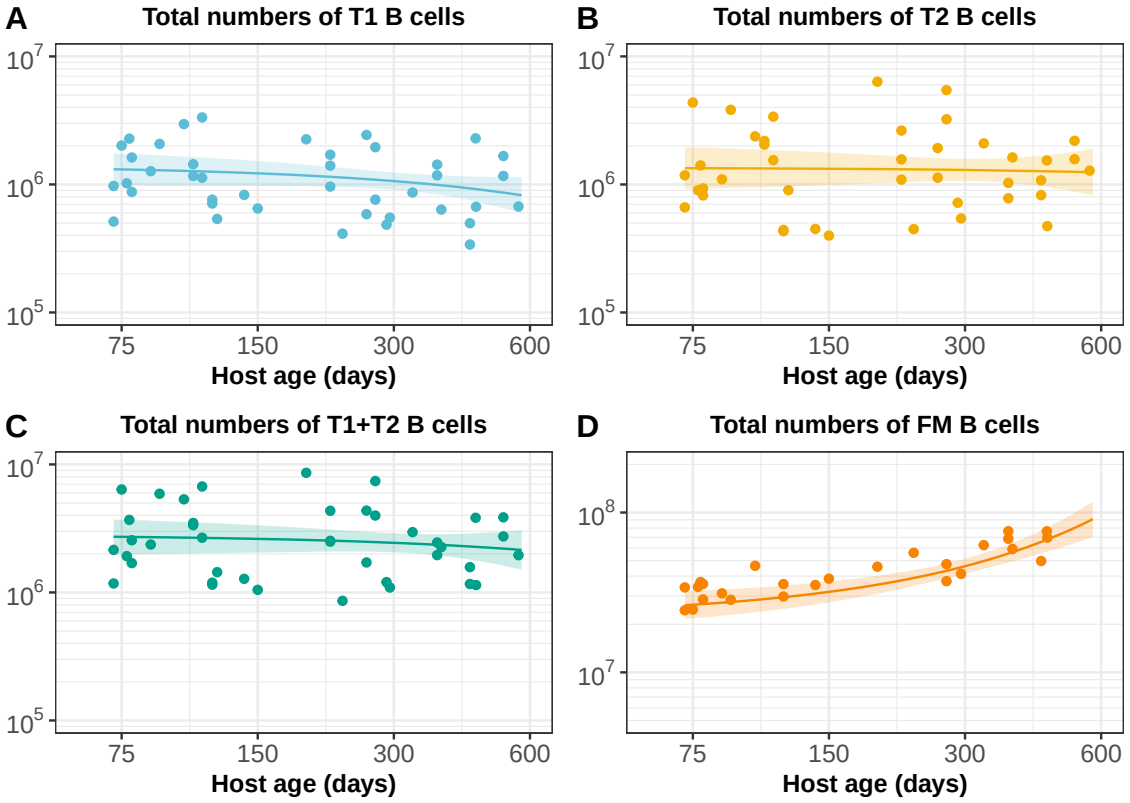

Figure 2: **The time-course of total counts of precursor compartments.** Solid lines indicate the most probable description of changes in the size of the precursor population with host age, modelled according to equation 7. Shaded envelopes denote the uncertainty in the most probable prediction and are drawn by taking 4.5 and 95.5 percentiles intervals of model predictions made from the posterior distribution of parameters.

1.4 Mathematical models used to describe B cell population dynamics

Simple homogeneous model: In this model we assume that cells follow random birth-death processes to form a kinetically homogeneous population that self renews through homoeostatic division (ρ) and decays by death and maturation. We define the rate of cellular turnover 'δ', which accounts for combined loss of cells by death and by differentiation events. Influx of cells from the source compartment is given by ϕ and is modelled as shown in section 1.3. We model the dynamics of Ki67$^{\text{hi}}$(Y) and Ki67$^{\text{lo}}$(X) cells using the coupled system of ODEs depicted below,

$$\begin{aligned}\dot{Y}(t) &= \phi(t)\epsilon + \rho\,(2\,X(t) + Y(t)) - (\beta + \delta)\,Y(t),\\ \dot{X}(t) &= \phi(t)(1-\epsilon) + \beta\,Y(t) - (\rho + \delta)\,X(t).\end{aligned} \tag{9}$$

Here, β is the rate of loss of Ki67 expression after mitosis, which is estimated form the model fits and ϵ denotes the proportion of the cells entering from source that are Ki67$^{\text{hi}}$, which is calculated from observations and used as an input in the model fitting process.

We define time 't_0' as the age at BMT of the youngest recipient. The proportions of Ki67$^{\text{hi}}$cells in the host compartment at t_0 $\left(\kappa_{\text{host}}(0) = \kappa_0\right)$, are estimated from the model fits. At the time of BMT, the size of donor compartment is zero *i.e.*$N_{\text{donor}} = 0$. Therefore, the proportions of Ki67$^{\text{hi}}$cells in the donor compartment at t_0 would reflect the proportions of Ki67$^{\text{hi}}$cells in the source influx *i.e.* $\kappa_{\text{donor}}(0) = \epsilon$. The initial sizes of Ki67$^{\text{hi}}$and Ki67$^{\text{lo}}$subsets within the host compartment are therefore given as $N_0\,\kappa_0$ and $N_0\,(1-\kappa_0)$, respectively.

Priors on the model parameters for the Simple homogeneous model

Unknowns		**FM**	**Spleen GC**	**LN GC**
Parameters	$\psi\ (\geq 0)\ (\leq 1)$	Normal(0.5, 0.25)	Normal(0.001, 0.5)	Normal(0.001, 0.5)
	$\rho\ (\geq 0)$	Normal(0.01, 0.5)	Normal(0.01, 0.5)	Normal(0.01, 0.5)
	$\delta\ (\geq 0)$	Normal(0.01, 0.5)	Normal(0.01, 0.5)	Normal(0.01, 0.5)
	$\beta\ (\geq 0)$	Normal(3.5, 1)	Normal(3.5, 1)	Normal(3.5, 1)
Initial conditions	$\log(N_0)$	Normal(17, 1)	Normal(11, 1)	Normal(11, 0.4)
	$\kappa_0\ (\geq 0)\ (\leq 1)$	Normal(0.2, 0.15)	Normal(0.9, 0.05)	Normal(0.9, 0.05)
Variance	$\sigma_N\ (\geq 0)$	Cauchy(1, 1)	Cauchy(0.7, 1)	Cauchy(0.3, 0.5)
	$\sigma_{fd}\ (\geq 0)$	Cauchy(1.5, 1)	Cauchy(0.2, 1)	Cauchy(0.4, 0.5)
	$\sigma_{\kappa_{donor}}\ (\geq 0)$	Cauchy(1, 1)	Cauchy(0.05, 1)	Cauchy(0.9, 0.5)
	$\sigma_{\kappa_{host}}\ (\geq 0)$	Cauchy(0.5, 1)	Cauchy(0.05, 1)	Cauchy(0.5, 0.5)

Time-dependent model: This is a homogeneous model where the fitness of a population, defined as its ability to grow either by enhanced cellular division (ρ) or decreased turnover (δ), varies with time. However, at any given instant of time all cells in the population behave identically. Variation in the fitness may be a result of changes in the cellular environment manifested as animals age. We explored different forms of either rates of cell division (ρ) or turnover (δ) changing with time and the best-fit was obtained when turnover declines

exponentially with host age ($\delta_0\, e^{-r\, t}$) while the rate of division remains unaltered.

$$\begin{aligned}\dot{Y}(t) &= \phi(t)\epsilon\, + \rho(t)\,(2\,X(t) + Y(t)) - (\beta + \delta(t))\,Y(t),\\ \dot{X}(t) &= \phi(t)(1-\epsilon)\, + \beta\,Y(t) - (\rho(t) + \delta(t))\,X(t).\end{aligned} \tag{10}$$

Priors on the model parameters for the Time-dependent model

Unknowns		**FM**	**Spleen GC**	**LN GC**
Parameters	$\psi\ (\geq 0)\ (\leq 1)$	Normal(0.5, 0.25)	Normal(0.001, 0.5)	Normal(0.001, 0.5)
	$\rho\ (\geq 0)$	Normal(0.01, 0.5)	Normal(0.01, 0.5)	Normal(0.01, 0.5)
	$\delta_0\ (\geq 0)$	Normal(0.01, 0.5)	Normal(0.01, 0.5)	Normal(0.01, 0.5)
	r	Normal(0, 0.3)	Normal(0, 0.3)	Normal(0, 0.3)
	$\beta\ (\geq 0)$	Normal(3.5, 1)	Normal(3.5, 1)	Normal(3.5, 1)
Initial conditions	$\log(N_0)$	Normal(17, 1)	Normal(11, 1)	Normal(11, 0.4)
	$\kappa_0\ (\geq 0)\ (\leq 1)$	Normal(0.2, 0.15)	Normal(0.9, 0.05)	Normal(0.9, 0.05)
Variance	$\sigma_N\ (\geq 0)$	Cauchy(1, 1)	Cauchy(0.7, 1)	Cauchy(0.3, 0.5)
	$\sigma_{fd}\ (\geq 0)$	Cauchy(1.5, 1)	Cauchy(0.2, 1)	Cauchy(0.4, 0.5)
	$\sigma_{\kappa_{donor}}\ (\geq 0)$	Cauchy(1, 1)	Cauchy(0.05, 1)	Cauchy(0.9, 0.5)
	$\sigma_{\kappa_{host}}\ (\geq 0)$	Cauchy(0.5, 1)	Cauchy(0.05, 1)	Cauchy(0.5, 0.5)

Kinetic-heterogeneity model: This model assumes that the given population is heterogeneous, composed of kinetically distinct subsets that turnover or divide at different rates. The subset with faster kinetics of net loss would get replaced quickly new cells coming form the source, while the one with slower kinetics would resist the replacement. As a result, we expect a sharp increase in the donor fraction initially which then stabilises slowly to the level of donor faction in the source compartment.

$$\begin{aligned}\dot{Y}_f(t) &= \phi(t)\, f\, \epsilon\, + \rho_f\,(2\,X_f(t) + Y_f(t)) - (\beta + \delta_f)\,Y_f(t),\\ \dot{X}_f(t) &= \phi(t)\, f\,(1-\epsilon)\, + \beta\,Y_f(t) - (\rho_f + \delta_f)\,X_f(t),\\ \dot{Y}_s(t) &= \phi(t)\,(1-f)\,\epsilon\, + \rho_s\,(2\,X_s(t) + Y_s(t)) - (\beta + \delta_s)\,Y_s(t),\\ \dot{X}_s(t) &= \phi(t)\,(1-f)\,(1-\epsilon)\, + \beta\,Y_s(t) - (\rho_s + \delta_s)\,X_s(t).\end{aligned} \tag{11}$$

We assume that fast and slow subsets exist in both host and donor sub-populations and therefore solve the ODEs depicted in eq. 11 separately for each compartment. f and α are the proportions of fast subset in the source and the target population, respectively. Initial sizes of the fast and slow subsets within the host compartment are given as $N_0\,\alpha$ and $N_0\,(1-\alpha)$, respectively. The initial fraction of Ki67$^{\text{hi}}$cells (κ_0) in fast and slow subsets were allowed to be different and were estimated

Incumbent model: The incumbent model (Hogan et al PNAS 2015 and Rane et al PLoS Bio 2018) is the variant of kinetic-heterogeneity model which assumes that heterogeneity is exhibited only in the host com-

Priors on the model parameters for the Kinetic heterogeneity model

Unknowns		FM	Spleen GC	LN GC
Parameters	$\psi\ (\geq 0)\ (\leq 1)$	Normal(0.5, 0.25)	Normal(0.001, 0.5)	Normal(0.001, 0.5)
	$\rho_{\text{fast}}\ (\geq 0)$	Normal(0.01, 0.5)	Normal(0.01, 0.5)	Normal(0.01, 0.5)
	$\delta_{\text{fast}}\ (\geq 0)$	Normal(0.01, 0.5)	Normal(0.01, 0.5)	Normal(0.01, 0.5)
	$\rho_{\text{slow}}\ (\geq 0)$	Normal(0.01, 0.5)	Normal(0.01, 0.5)	Normal(0.01, 0.5)
	$\delta_{\text{slow}}\ (\geq 0)$	Normal(0.01, 0.5)	Normal(0.01, 0.5)	Normal(0.01, 0.5)
	$\beta\ (\geq 0)$	Normal(3.5, 1)	Normal(3.5, 1)	Normal(3.5, 1)
	$f\ (\geq 0)\ (\leq 1)$	Normal(0.05, 0.25)	Normal(0.05, 0.5)	Normal(0.05, 0.5)
Initial conditions	$\log(N_0)$	Normal(17, 1)	Normal(11, 1)	Normal(11, 0.4)
	$\alpha\ (\geq 0)\ (\leq 1)$	Normal(0.5, 0.25)	Normal(0.5, 0.25)	Normal(0.5, 0.25)
	$\kappa_{\text{Fast}}(0)\ (\geq 0)\ (\leq 1)$	Normal(0.2, 0.15)	Normal(0.9, 0.05)	Normal(0.9, 0.05)
	$\kappa_{\text{slow}}(0)\ (\geq 0)\ (\leq 1)$	Normal(0.2, 0.15)	Normal(0.9, 0.05)	Normal(0.9, 0.05)
Variance	$\sigma_N\ (\geq 0)$	Cauchy(1, 1)	Cauchy(0.7, 1)	Cauchy(0.3, 0.5)
	$\sigma_{fd}\ (\geq 0)$	Cauchy(1.5, 1)	Cauchy(0.2, 1)	Cauchy(0.4, 0.5)
	$\sigma_{\kappa_{donor}}\ (\geq 0)$	Cauchy(1, 1)	Cauchy(0.05, 1)	Cauchy(0.9, 0.5)
	$\sigma_{\kappa_{host}}\ (\geq 0)$	Cauchy(0.5, 1)	Cauchy(0.05, 1)	Cauchy(0.5, 0.5)

partment. All donor cells follow the same rules of turnover and division in this model. Whereas the host compartment is composed of - (i) an incumbent subset of older, self-renewing cells that are resistant to displacement by new cells and (ii) a dis-placeable subset that turns over with a constant net loss rate 'λ' and is replaced continuously by cohorts of new cells entering the pool.

$$\begin{aligned}
\dot{Y}(t) &= \phi(t)\epsilon\ + \rho\,(2\,X(t) + Y(t)) - (\beta + \delta)\,Y(t),\\
\dot{X}(t) &= \phi(t)(1-\epsilon)\ + \beta\,Y(t) - (\rho + \delta)\,X(t),\\
\dot{Y}_{\text{inc}}(t) &= \rho_{\text{inc}}\,(2\,X_{\text{inc}}(t) + Y_{\text{inc}}(t)) - (\beta + \delta_{\text{inc}})\,Y_{\text{inc}}(t),\\
\dot{X}_{\text{inc}}(t) &= \beta\,Y_{\text{inc}}(t) - (\rho_{\text{inc}} + \delta_{\text{inc}})\,X_{\text{inc}}(t).
\end{aligned} \tag{12}$$

We assume that the incumbent subset is established very early in the mouse life and hence could only be present in the host compartment, since the minimum age of BMT in chimeric animals was $\sim$7 weeks.

Priors on the model parameters for the Incumbent model

Unknowns		FM	Spleen GC	LN GC
Parameters	ψ (≥ 0) (≤ 1)	Normal(0.5, 0.25)	Normal(0.001, 0.5)	Normal(0.001, 0.5)
	ρ (≥ 0)	Normal(0.01, 0.5)	Normal(0.01, 0.5)	Normal(0.01, 0.5)
	δ (≥ 0)	Normal(0.01, 0.5)	Normal(0.01, 0.5)	Normal(0.01, 0.5)
	ρ_{inc} (≥ 0)	Normal(0.01, 0.5)	Normal(0.01, 0.5)	Normal(0.01, 0.5)
	δ_{inc} (≥ 0)	Normal(0.01, 0.5)	Normal(0.01, 0.5)	Normal(0.01, 0.5)
	β (≥ 0)	Normal(3.5, 1)	Normal(3.5, 1)	Normal(3.5, 1)
Initial conditions	log(N_0)	Normal(17, 1)	Normal(11, 1)	Normal(11, 0.4)
	α (≥ 0) (≤ 1)	Normal(0, 0.25)	Normal(0, 0.25)	Normal(0, 0.25)
	κ_0 (≥ 0) (≤ 1)	Normal(0.2, 0.15)	Normal(0.9, 0.05)	Normal(0.9, 0.05)
	$\kappa_{\text{inc}}(0)$ (≥ 0) (≤ 1)	Normal(0.2, 0.15)	Normal(0.9, 0.05)	Normal(0.9, 0.05)
Variance	σ_N (≥ 0)	Cauchy(1, 1)	Cauchy(0.7, 1)	Cauchy(0.3, 0.5)
	σ_{fd} (≥ 0)	Cauchy(1.5, 1)	Cauchy(0.2, 1)	Cauchy(0.4, 0.5)
	$\sigma_{\kappa_{donor}}$ (≥ 0)	Cauchy(1, 1)	Cauchy(0.05, 1)	Cauchy(0.9, 0.5)
	$\sigma_{\kappa_{host}}$ (≥ 0)	Cauchy(0.5, 1)	Cauchy(0.05, 1)	Cauchy(0.5, 0.5)