

# Chapitre 3 - Annexe B :

## Modélisation de l'hétérogénéité pour les prélèvements issus de dons de moelle osseuse

Gurvan Hermange

### Contenu de l'annexe

Le chapitre 3 se concentrait sur l'étude des temps de première division des cellules souches et progénitrices (HSCP) dans le cas d'échantillons issus de sang de cordon.

Ici, nous nous intéresserons à des prélèvements de dons de moelle osseuse de différents individus, pour lesquels il y a une forte hétérogénéité. Nous explorons alors l'utilisation d'une méthode d'estimation Bayésienne hiérarchique. Nous simplifions cette étude en utilisant un modèle simplifié pour la modélisation des temps de première division des HSPC, à savoir un modèle log-normal dont le paramètre  $\sigma$  quantifiant la variabilité sera fixé égal pour les trois types cellulaires.

### Table des matières

<b>1</b>	<b>Données expérimentales</b>	<b>2</b>
<b>2</b>	<b>Données poolées</b>	<b>2</b>
2.1	Modèle simplifié . . . . .	2
2.2	Simplification additionnelle du modèle . . . . .	2
2.3	Estimation de l'incertitude sur les paramètres . . . . .	4
<b>3</b>	<b>Prise en compte de l'hétérogénéité entre individus</b>	<b>5</b>
3.1	Représentation des données à l'échelle de l'individu . . . . .	5
3.2	Individus considérés indépendants . . . . .	6
3.3	Effet populationnel et estimation hiérarchique . . . . .	6
3.4	Comparaison entre les approches . . . . .	11

# 1 Données expérimentales

Les données utilisées dans cette annexe proviennent d'échantillons de sang issus de dons de moelle osseuse. A chaque expérimentation correspond un individu. Contrairement aux prélèvements de sang de cordon étudiés dans le chapitre 3, ici les individus ont des âges différents, justifiant la plus forte hétérogénéité observée entre expérimentations. Les données sont représentées sur la figure 1.

En plus de la plus forte hétérogénéité entre individus, notons également des temps de première division en moyenne plus élevés.

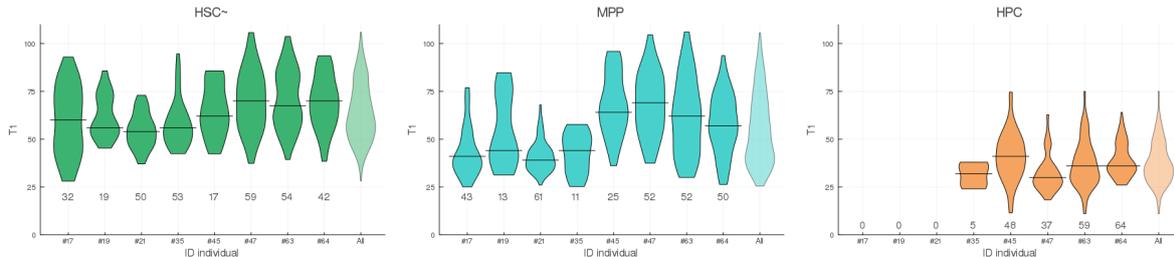


FIGURE 1 – Temps de première division des HSC\* (gauche), MPPs (centre) et HPCs (droite) en fonction des individus. La ligne verticale correspond à la médiane. Les valeurs en grises sous chacune des distributions indiquent le nombre de données disponibles pour ces individus. Les distributions plus claires à droite de chacun des graphes correspondent à celles obtenues en regroupant tous les individus (pool).

Le dataset présenté dans cette annexe comporte 921 observations. Parmi celles-ci, 326 correspondent à des temps d'observations de HSC, 307 à des MPPs et 213 à des HPCs. Les 75 données restantes sont labellisées "NA", ce qui signifie que la première division n'a pas été observée sur les 96 heures qu'a duré l'expérience. Il est possible que la première division n'ait pas été observée pour différentes raisons (cellules qui sortent du champ de l'appareil par exemple) mais que les divisions suivantes aient été observées (dans 3 cas). On exclut ces données.

Dans 30 cas, nous n'avons pas non plus d'indication sur des temps de divisions ultérieurs (deuxième ou troisième division), mais l'information comme quoi la taille de la colonie à 96 heures est strictement supérieure à un, ce qui signifie que la cellule s'est divisée, sans qu'on sache à quel moment. Ces données sont exclues.

Finalement, on se retrouve avec 42 observations labellisées "NA" (25 pour des HSC\*, 16 pour des MPPs et 1 pour des HPCs) que l'on considère correspondre à des divisions survenant après 96 heures.

## 2 Données poolées

### 2.1 Modèle simplifié

Nous considérons ici que les temps de première division, pour chacune des populations de cellules (HSC\*, MPP, HPC), sont distribuées suivant une loi log-normale de paramètres  $\mu$  et  $\sigma$  (ces derniers paramètres étant fonction de la population de cellules considérée).

Nous considérons les données de patients poolées ensemble (voir Fig. 2).

Nous pouvons alors calculer numériquement (en utilisant l'algorithme CMA-ES) la vraisemblance pour en trouver le maximum, et nous obtenons finalement les résultats suivants :

- Pour les HSC\* :  $\hat{\mu} = 4.16$ ,  $\hat{\sigma} = 0.279$  et un maximum de log-vraisemblance valant -1421.44
- Pour les MPPs :  $\hat{\mu} = 3.98$ ,  $\hat{\sigma} = 0.374$  et un maximum de log-vraisemblance valant -1363.99
- Pour les HPCs :  $\hat{\mu} = 3.57$ ,  $\hat{\sigma} = 0.319$  et un maximum de log-vraisemblance valant -820.57

### 2.2 Simplification supplémentaire du modèle

On se demande alors si on peut simplifier ce modèle, en diminuant le nombre de paramètres à estimer, par exemple en considérant une valeur  $\sigma$  commune aux trois catégories de cellules. Dans

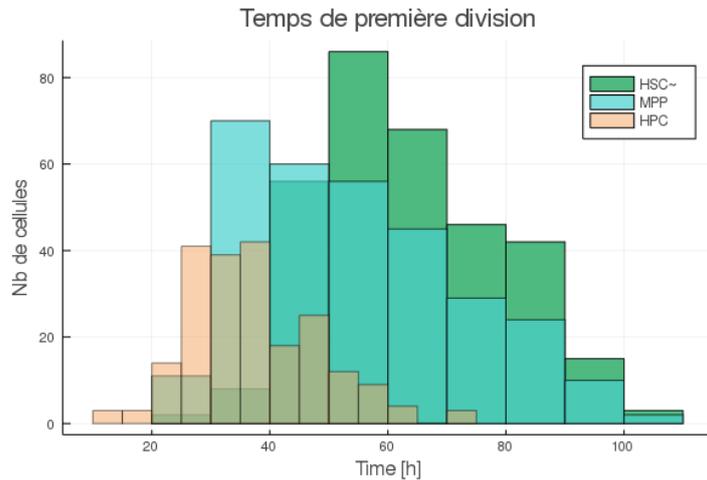


FIGURE 2 – Distribution des temps de premières divisions pour les HSC\*, MPP et HPC. A noter que certaines divisions survenant après 96 heures ne sont pas mesurées, et donc ne sont pas représentées sur ces histogrammes, qui doivent donc être considérés comme potentiellement censurés.

ce cas là, l'estimation est conjointe. En utilisant ici encore l'algorithme CMA-ES, on trouve, au maximum de la log-vraisemblance (valant -3619),  $\hat{\sigma} = 0.326$  et une valeur de  $\hat{\mu}$  égale à 4.17, 3.98 et 3.57 pour les HSC\*, MPPs, et HPCs respectivement. Ces valeurs sont quasiment les mêmes que celles obtenues précédemment. Les résultats des ajustements aux données sont présentées sur la figure 3, où l'on compare le modèle simplifié avec un paramètre  $\sigma$  commun aux trois catégories de cellules *vs* le modèle plus complet.

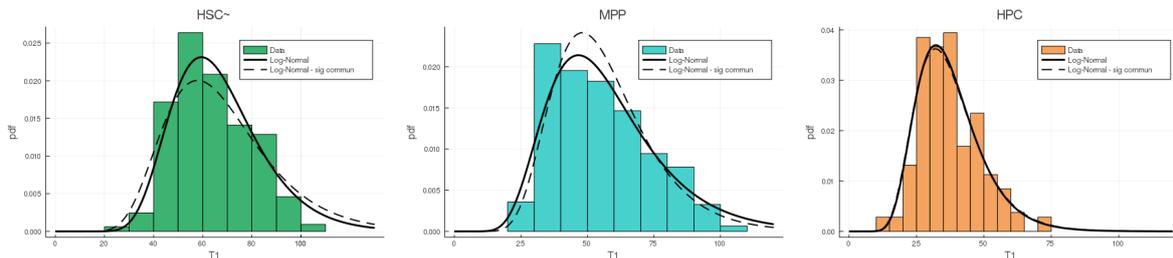


FIGURE 3 – Ajustement aux données du modèle log-normal. De gauche à droite, données des HSC\*, MPP et HPC. On compare le modèle à 6 degrés de liberté (courbes noires), avec une valeur de  $\sigma$  propre à chaque catégorie de cellules, *vs* le modèle simplifié (courbes en tirets) où le paramètre  $\sigma$  est commun aux trois catégories de cellules.

La log-vraisemblance du modèle à 6 paramètres est la somme des trois log-vraisemblances calculées de façon indépendante, et vaut -3606. Elle est bien sûr supérieure à celle que l'on obtient dans le modèle à 4 paramètres, mais ce dernier est aussi plus parcimonieux. Pour comparer ces deux modèles, on peut appliquer des critères classiques de sélection de modèle, par exemple le critère d'information d'Akaike (AIC) défini par [1, 2] :

$$AIC = -2\mathcal{L}(\hat{\theta}) + 2k \quad (1)$$

avec  $k$  le nombre de paramètres à estimer, ou encore le critère d'information Bayésien (BIC), défini par [6] :

$$BIC = -2\mathcal{L}(\hat{\theta}) + k \cdot \log(N_{\mathcal{D}}) \quad (2)$$

avec  $N_{\mathcal{D}}$  le nombre d'observations.

L'application de ces deux critères nous amène à sélectionner le modèle à 6 paramètres. Celui-ci serait plus précis pour décrire la dynamique de division des cellules. Néanmoins, dans l'optique de développer un modèle plus complet de prolifération et différenciation, il reste préférable d'avoir moins de paramètres. En plus du compromis entre parcimonie et qualité de l'ajustement aux

données (tel qu'évalué par les critères BIC ou AIC), les aspects computationnels devraient également être pris en compte. Un modèle avec plus de paramètres convergera plus difficilement, en un temps plus long et prendra plus de place en mémoire.

Vu que les estimations restent satisfaisantes avec le modèle à quatre paramètres, c'est ce dernier que nous utiliserons dans la suite. Notamment en vue de prendre en compte l'hétérogénéité entre individus, et donc de subdiviser notre dataset actuel (donc de réduire le nombre de données disponibles pour chaque estimation).

### 2.3 Estimation de l'incertitude sur les paramètres

Étudions maintenant un peu plus le modèle sélectionné précédemment (à quatre degrés de liberté), en essayant de quantifier l'incertitude que nous avons sur les paramètres. Pour cela, nous nous plaçons dans un cadre Bayésien. Le vecteur de paramètres  $\theta$  est maintenant considéré comme un vecteur aléatoire dont nous aimerions déterminer la loi *a posteriori*, c'est-à-dire sachant les données :

$$\mathbb{P}[\theta|\mathcal{D}, \mathcal{M}] \sim \mathbb{P}[\mathcal{D}|\theta, \mathcal{M}] \cdot \mathbb{P}[\theta] \quad (3)$$

Le terme de droite fait apparaître la vraisemblance  $\mathbb{P}[\mathcal{D}|\theta, \mathcal{M}]$  ainsi que  $\mathbb{P}[\theta]$  qui correspond à la distribution *a priori* des paramètres (prior). Ici, nous choisissons cette dernière uniforme.

Pour estimer la loi *a posteriori*, nous échantillonnons dans cette loi en utilisant une méthode Markov Chain Monte Carlo (MCMC), à savoir l'algorithme Metropolis-Hasting [7].

Nous exécutons l'algorithme Metropolis-Hasting sur 300,000 itérations. Nous monitorons la bonne exécution de l'algorithme en vérifiant que le taux d'acceptation est satisfaisant et que les chaînes MCMC explorent suffisamment l'espace des paramètres (Fig. 4 gauche et centre). La convergence de l'algorithme pourrait également être évaluée en calculant la moyenne ergodique des paramètres au cours des itérations. Ici, nous avons initialisé l'algorithme à des valeurs proches du maximum de vraisemblance, à partir des résultats obtenus précédemment, ce qui permet d'obtenir une convergence rapidement. Nous obtenons alors les distributions *a posteriori* des paramètres (Fig. 4 droite et Fig. 5). Finalement, en échantillonnant dans ces distributions, nous pouvons propager les incertitudes des paramètres vers la sortie du modèle (ici les densités de probabilités des temps de première division des HSC\*, MPP et HPC) et nous obtenons les résultats de la figure 6.

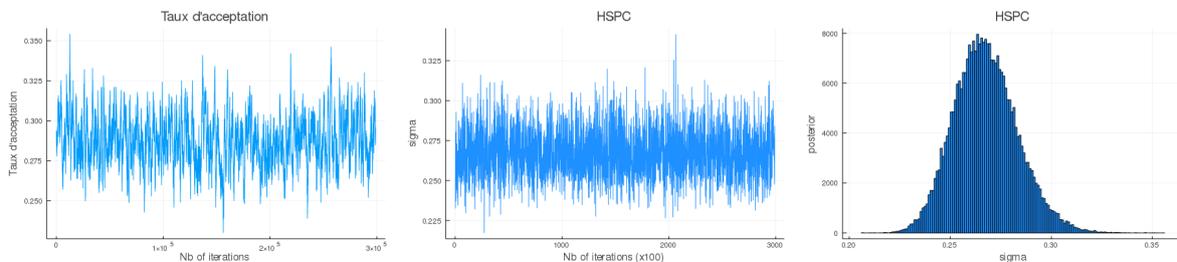


FIGURE 4 – Monitoring de l'exécution de l'algorithme Metropolis-Hasting. A gauche, taux d'acceptation au cours des itérations. Au centre, chaîne MCMC générée pour le paramètre  $\sigma$  et à droite, distribution *a posteriori* qui en résulte.

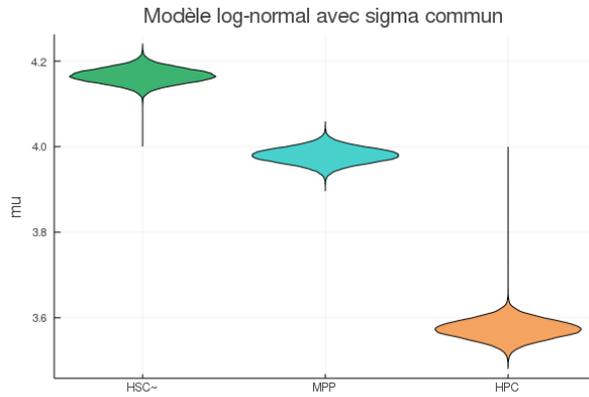


FIGURE 5 – Estimation de la distribution *a posteriori* du paramètre  $\mu$  pour les HSC\*, MPP et HPC (lorsque les données sont regroupées (poolées) ensemble, pour l’ensemble des individus). On observe clairement que les divisions sont plus rapides pour des cellules moins immatures.

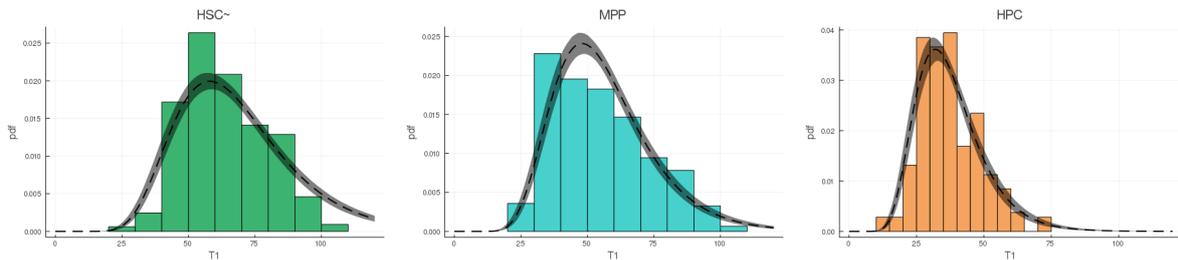


FIGURE 6 – Confrontation du modèle Log-normal avec  $\sigma$  commun (4 degrés de liberté) aux données des HSC\*, MPPs et HPCs. On représente un intervalle de crédibilité à 95%. La ligne en tirets correspond à la courbe médiane.

### 3 Prise en compte de l’hétérogénéité entre individus

#### 3.1 Représentation des données à l’échelle de l’individu

Jusqu’à présent, nous avons considéré les observations des temps de première division des HSC\* (puis des MPPs et HPCs) issues des prélèvements de moelle osseuse, sans prendre en compte que ces prélèvements provenaient de différents individus. Nous avons donc regroupé les données ensemble, négligeant l’hétérogénéité entre les individus. Nous remettons en cause cette hypothèse ici, et cherchons à étudier plus précisément cette hétérogénéité. En séparant les observations selon les individus, nous obtenons les distributions de la figure 1. Visuellement, les distributions des temps de première division sont assez proches dans le cas des HSC\*, et semblent plus hétérogènes dans le cas des MPPs. Si on effectue un test de Kruskal-Wallis[9, 12], on rejette (avec une très forte significativité quelle que soit la catégorie de cellules considérée) l’hypothèse selon laquelle il n’y aurait pas un individu dont la distribution dominerait celle des autres.

Si on effectue des tests de Mann-Whitney [13, 11], comparant les individus deux à deux, les tests nous conduisent également dans plusieurs cas à rejeter l’hypothèse selon laquelle les distributions des temps de première division de deux individus sont proches.

L’hétérogénéité entre les différentes expériences (à un individu correspond une expérience) n’est pas surprenante, et la multiplication du nombre d’expériences augmente le risque d’hétérogénéité. Elle peut s’expliquer par des différences dans les expériences elles-mêmes, mais aussi pas la variabilité inter-individuelle, ou encore par le fait que les prélèvements sont issus d’individus d’âges différents, sachant que l’hématopoïèse s’altère avec l’âge [3].

Malgré l’hétérogénéité entre expériences, en biologie, il est souvent nécessaire de quand-même regrouper les données ensemble pour pouvoir les exploiter, comme nous l’avons fait jusqu’à présent. Il est cependant important d’être conscient des limites éventuelles de cette approche.

Ici, nous allons étudier deux autres approches. L’une vise à considérer les individus indépendants les uns des autres, et à faire des analyses séparées pour chacun. Bien sûr, cela augmente significativement le nombre de calculs à effectuer, tout en augmentant l’incertitude sur les résultats

car diminuant la quantité de données disponibles pour chacun. La seconde approche est intermédiaire entre regrouper les individus ensemble et les considérer indépendants : elle introduit un effet populationnel. Plus robuste, elle est aussi plus compliquée à mettre en place.

### 3.2 Individus considérés indépendants

Jusqu'à maintenant, on avait regroupé les données des individus ensemble (pool) pour estimer les paramètres du modèle log-normal (à quatre degrés de liberté, avec  $\sigma$  commun aux HSC\*, MPPs et HPCs). On peut adopter l'approche inverse, et cette fois-ci considérer tous les individus indépendants les uns des autres. On estimera alors des paramètres propres à chaque individu. On multiplie donc le nombre de paramètres à estimer par le nombre d'individus, et chaque estimation se fait maintenant sur un jeu de données plus réduit. Cette approche augmente donc le risque d'overfitting (sur-ajustement aux données) et l'incertitude sur les estimations.

Après exécution de l'algorithme Metropolis-Hasting (idem que pour l'estimation poolée), on obtient les résultats de la figure 7. On peut notamment observer une incertitude plus forte lorsqu'il y a moins de données (individu #45 par exemple).

Sur les figures 10, 11 et 12, on montre comment le modèle s'adapte aux données des HSC\*, MPPs et HPCs respectivement.

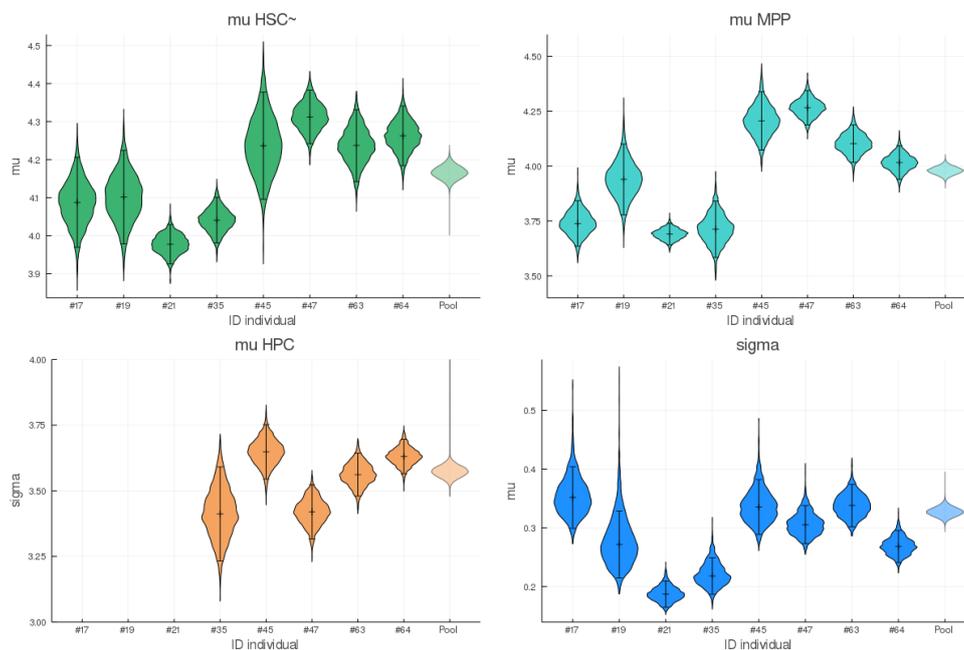


FIGURE 7 – Distributions *a posteriori* des paramètres  $\mu_{\text{HSC}^*}$  (en haut à gauche),  $\mu_{\text{MPP}}$  (en haut à droite),  $\mu_{\text{HPC}}$  (en bas à gauche) et  $\sigma$  (en bas à droite) dans le cas d'estimations indépendantes pour chacun des individus. La ligne verticale sur chacune des distributions représente la médiane et les quantiles à 2.5% et 97.5% (elle sera utile pour comparer avec l'approche hiérarchique, voir Fig. 8). Les distributions plus claires à droite de chaque graphe représentent celles obtenues en regroupant (poolant) les données des individus (elles correspondent aux distribution de la figure 4 droite et 5).

### 3.3 Effet populationnel et estimation hiérarchique

Les limites de l'approche précédente - où chaque individu est considéré séparément - sont claires : risque d'overfitting et augmentation de l'incertitude. L'approche opposée, où les données sont poolées ensemble, a aussi ses limites : elle fait l'hypothèse d'une homogénéité entre individus, hypothèse qu'on peut remettre en cause en observant les données, et sous-estime l'incertitude. La méthode d'estimation Bayésienne hiérarchique permet de pallier aux problématiques rencontrées ci-dessus : on continue à estimer des paramètres propres à chaque individu, tout en faisant l'hypothèse que ces paramètres proviennent *a priori* de distributions de populations, dont on veut estimer la moyenne et la variance (qu'on appellera hyper-paramètres), et dont on suppose la

variance la plus faible possible (en un sens qu'on va définir dans la suite). Cette méthode permet ainsi d'augmenter la robustesse des résultats en réduisant la variance entre individus [4, 10].

Plus précisément (comme décrit dans [8]), si on considère une population  $\mathcal{P} = \{1, \dots, N\}$  de  $N$  individus, dont les distributions des temps de première division des HSPC sont décrites par le modèle  $\mathcal{M}$ ,  $\boldsymbol{\theta} = \left\{ \boldsymbol{\theta}^{(i)} \right\}_{i \in \mathcal{P}}$  décrit l'ensemble des paramètres des individus avec :

$$\begin{aligned} \boldsymbol{\theta}^{(1)} &= \left( \theta_1^{(1)}, \dots, \theta_P^{(1)} \right) \\ &\vdots \\ \boldsymbol{\theta}^{(N)} &= \left( \theta_1^{(N)}, \dots, \theta_P^{(N)} \right) \end{aligned}$$

où  $P$  est le nombre de paramètres à estimer. Avec la méthode d'inférence hiérarchique, au lieu d'estimer chaque  $\boldsymbol{\theta}^{(i)}$  de façon indépendante, nous supposons que tous les vecteurs de paramètres des individus sont des réalisations de la même variable aléatoire d'une distribution inconnue dans un modèle statistique. Ainsi, le modèle hiérarchique (aussi connu sous le nom de modèle à effet aléatoire) peut prendre en compte la variabilité inter-individuelle mais aussi la similarité entre individus. En pratique, nous considérons ici :

$$\forall i \in \mathcal{P}, \forall k \in \{1, \dots, P\}, \theta_k^{(i)} \mid \tau_k, \zeta_k^2 \sim \mathcal{N}_{c,k}(\tau_k, \zeta_k^2) \quad (4)$$

où la distribution de population de chacun des paramètres du modèle est une loi gaussienne tronquée  $\mathcal{N}_{c,k}$  (sur un intervalle qui dépend du paramètre considéré  $k$ ), et  $\boldsymbol{\tau} = (\tau_1, \dots, \tau_P)$  et  $\boldsymbol{\zeta}^2 = (\zeta_1^2, \dots, \zeta_P^2)$  sont les hyper-paramètres (HP).

On peut alors estimer la loi jointe (a posteriori) de  $\boldsymbol{\theta}$  et des hyper-paramètres  $\boldsymbol{\tau}$  et  $\boldsymbol{\zeta}^2$  :

$$\begin{aligned} \mathbb{P}[\boldsymbol{\theta}, \boldsymbol{\tau}, \boldsymbol{\zeta}^2 \mid \mathcal{D}] &\propto \mathbb{P}[\mathcal{D} \mid \boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(N)}, \boldsymbol{\tau}, \boldsymbol{\zeta}^2] \mathbb{P}[\boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(N)}, \boldsymbol{\tau}, \boldsymbol{\zeta}^2] \\ &\propto \prod_{i \in \mathcal{P}} \left( \mathbb{P}[\mathcal{D}_i \mid \boldsymbol{\theta}^{(i)}] \mathbb{P}[\boldsymbol{\theta}^{(i)} \mid \boldsymbol{\tau}, \boldsymbol{\zeta}^2] \right) \mathbb{P}[\boldsymbol{\tau}] \mathbb{P}[\boldsymbol{\zeta}^2] \end{aligned} \quad (5)$$

On échantillonne alors dans la distribution a posteriori en utilisant l'algorithme Metropolis-Hastings avec Gibbs sampling [7, 5]. Conditionnellement aux valeurs des hyper-paramètres, les individus sont indépendants et leurs paramètres peuvent être échantillonnés suivant une méthode Metropolis-Hasting standard, comme au paragraphe 2.3. Pour échantillonner les valeurs pour les hyper-paramètres (composante par composante), on utilise la méthode de Gibbs qui consiste à choisir, pour le proposal dans l'algorithme Metropolis-Hasting, la loi a posteriori conditionnelle des hyper-paramètres. Avec un tel choix, les nouveaux échantillons sont toujours acceptés. Nous présentons dans le manuscrit (aux chapitre 6) le détail des calculs.

Les priors pour  $\boldsymbol{\tau}$  sont choisis uniformes, tandis que pour  $\boldsymbol{\zeta}^2$ , on choisit pour prior des lois gamma-inverse  $(\alpha_k, 0)$  (tronquées sur un intervalle). Ces lois sont impropres, mais conjuguées à la vraisemblances, elles conduisent à des lois conditionnelles a posteriori qui sont des lois inverse-gamma régulières (tronquées également).  $\alpha_k$  dépend de l'hyper-paramètre considéré. Plus il est élevé, plus on donne d'importance aux faibles valeurs, c'est-à-dire plus on suppose une variance faible pour la distribution de population. On illustre ce choix au paragraphe suivant.

La méthode ci-dessus est alors implémentée et exécutée sur 1 million d'itérations. On obtient alors les résultats de la figure 8. Par rapport à l'estimation où chaque individu est considéré séparément (ligne verticale), on observe ici que les distributions des paramètres des patients se rapprochent. C'est le cas par exemple de l'individu #35, pour qui il y avait assez peu de données et pour qui la distribution de tous ses paramètres prenait des valeurs assez faibles par rapport aux autres individus. De même pour l'individu #45 qui lui avait des valeurs estimées plus élevées que le reste des individus.

On peut alors regarder, pour chacun des individus, l'ajustement du modèle aux données. Ces résultats sont représentés sur les figures 10, 11 et 12 pour les HSC\*, MPPs et HPCs respectivement. On compare ici l'ajustement dans le cas hiérarchique et dans le cas où les individus sont

indépendants. Les fits sont bien sûr meilleurs dans ce dernier cas, mais avec un risque d'overfitting.

Pour les individus avec beaucoup de données (#21, 64 ou 65 par exemple), les résultats changent peu avec l'approche hiérarchique, contrairement aux individus avec peu de données.

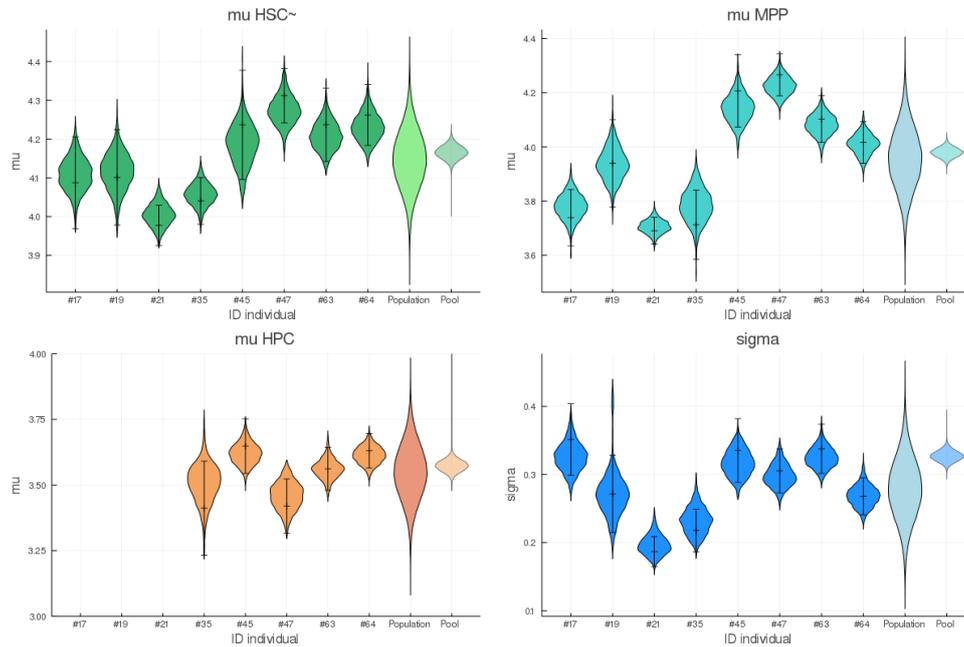


FIGURE 8 – Distributions *a posteriori* des paramètres  $\mu_{\text{HSC}^*}$  (en haut à gauche),  $\mu_{\text{MPP}}$  (en haut à droite),  $\mu_{\text{HPC}}$  (en bas à gauche) et  $\sigma$  (en bas à droite) dans le cas de l'estimation Bayésienne hiérarchique. Les lignes verticales correspondent aux estimations effectuées en considérant les patients indépendants (voir Fig. 7). Les distributions plus claires à droite de chaque graphe représentent celles obtenues en regroupant (poolant) les données des individus (elles correspondent aux distribution de la figure 4 droite et 5). La distribution de population correspond à une loi gaussienne (tronquée) de moyenne  $\mathbb{E}[\tau_k|\mathcal{D}]$  et de variance  $\mathbb{E}[\zeta_k|\mathcal{D}]$  (HP associés au paramètre  $k$ ). On peut observer qu'elles sont plus étendues que les distributions obtenues dans le cas "poolé".

De même qu'on avait étudié l'incertitude sur l'ajustement du modèle aux données poolées sur la figure 6, on peut ici échantillonner les paramètres du modèle suivant la distribution de population (loi gaussienne de paramètres les moyennes *a posteriori* des hyper-paramètres) et propager les incertitudes. Les résultats sont présentés sur la figure 9. En prenant en compte les individus, on a une loi de population qui a une variance bien plus élevée que dans le cas "poolé", ce qui conduit à des intervalles de crédibilités à 95% plus large que ceux qu'on avait jusqu'à présent. Il est intéressant de voir ici que les histogrammes sont maintenant inclus dans ces intervalles de crédibilité.

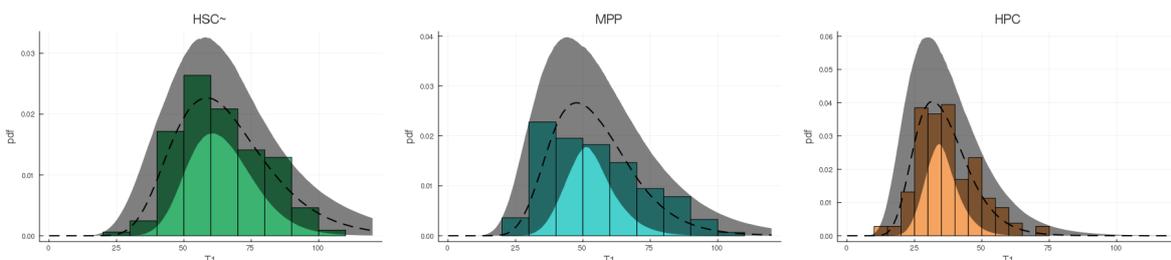


FIGURE 9 – Confrontation des données au modèle log-normal dont les paramètres sont échantillonnés suivant les distributions de population, c'est-à-dire des lois gaussiennes (tronquées) de paramètres les moyennes *a posteriori* des hyper-paramètres. On représente un intervalle de crédibilité à 95%.

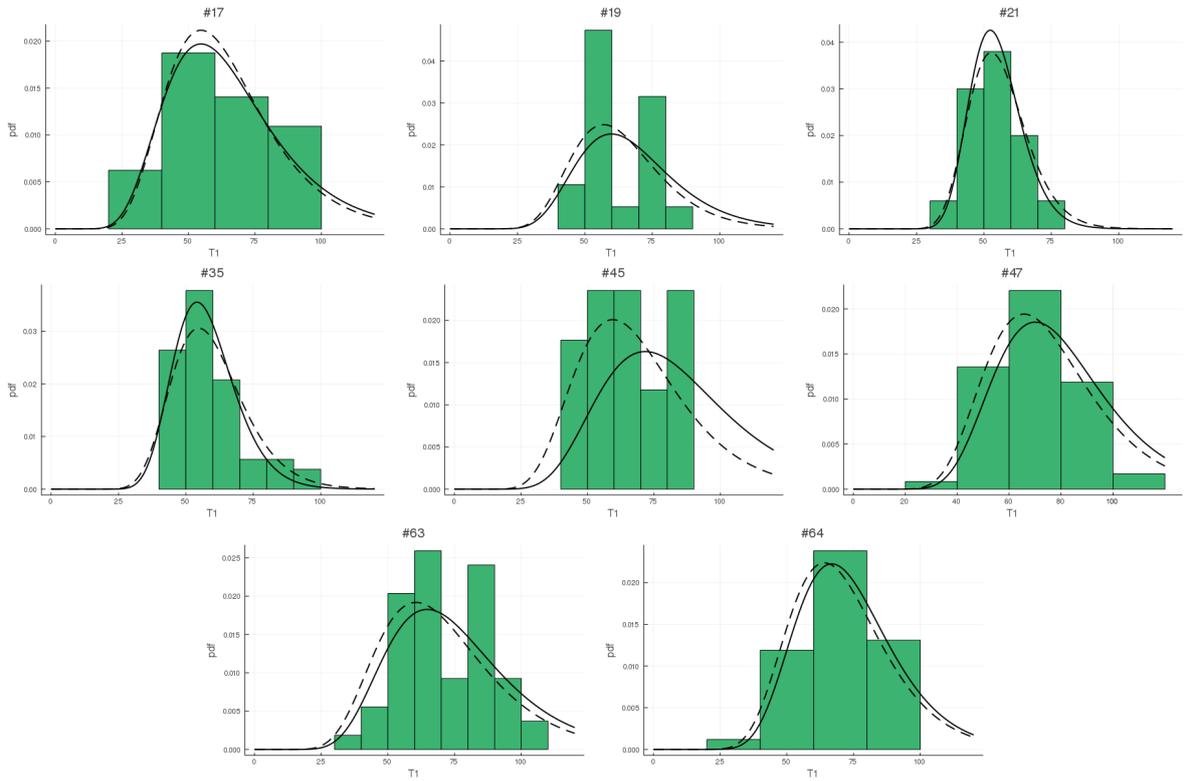


FIGURE 10 – Histogrammes des temps de première division des HSC\* pour chacun des individus, et confrontation au modèle log-normal dont les paramètres (moyennes a posteriori) sont soit estimés individu par individu (courbes noires) ou via la méthode hiérarchique (courbes en pointillés).

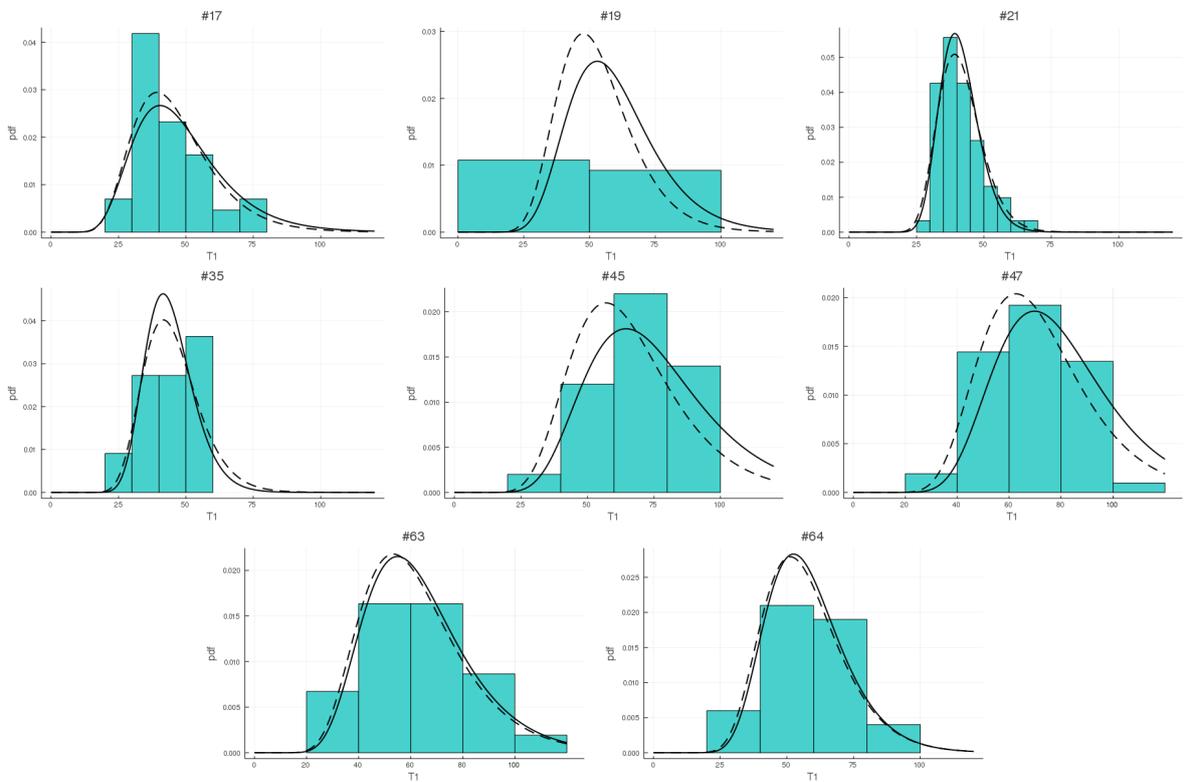


FIGURE 11 – Histogrammes des temps de première division des MPPs pour chacun des individus, et confrontation au modèle log-normal dont les paramètres (moyennes a posteriori) sont soit estimés individu par individu (courbes noires) ou via la méthode hiérarchique (courbes en pointillés).

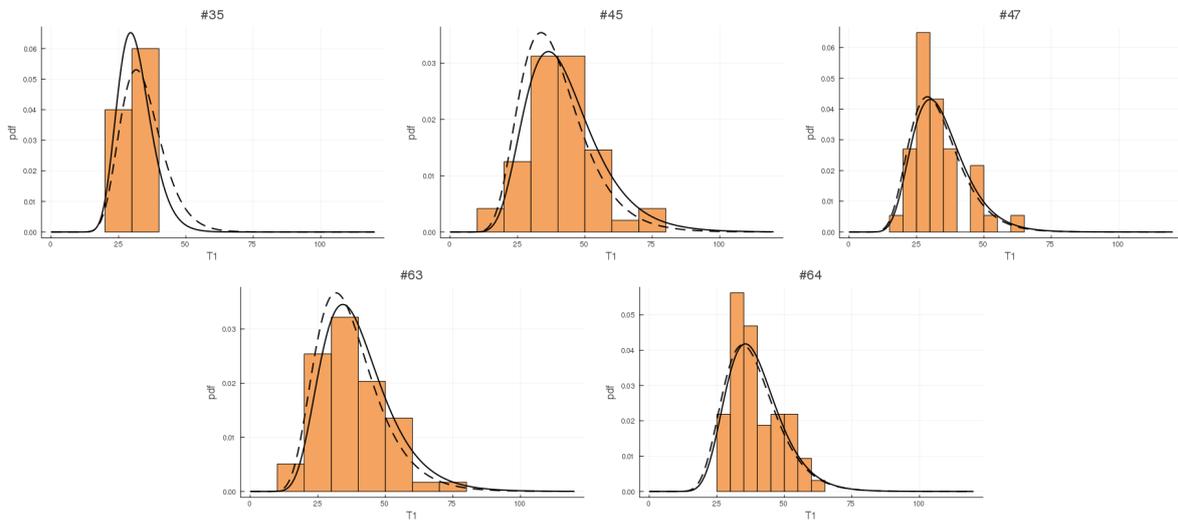


FIGURE 12 – Histogrammes des temps de première division des HPCs pour chacun des individus, et confrontation au modèle log-normal dont les paramètres (moyennes a posteriori) sont soit estimés individu par individu (courbes noires) ou via la méthode hiérarchique (courbes en pointillés).

### 3.4 Comparaison entre les approches

L'effet populationnel a été introduit par l'ajout d'hyper-paramètres, et notamment l'a priori selon lequel la variance, pour les distributions de population, doit être faible. Mathématiquement, cela se traduit par le choix d'une loi inverse-gamma  $(\alpha, 0)$ (impropre) avec des choix de  $\alpha$  ajustés en fonction des paramètres. Suivant la valeur choisie pour ce paramètre, il est possible de donner plus d'importance aux individus ou alors à la population. Nous illustrons l'influence de ce paramètre sur la figure 13. Nous avons ici fait varier  $\alpha$  pour l'hyper-paramètre associé à  $\mu_{\text{HPC}}$ . Pour  $\alpha = 0$ , on se retrouve avec des résultats proches de ceux qu'on avait lorsque les patients étaient considérés indépendants, et en augmentant  $\alpha$  jusqu'à  $\alpha = 3$ , on observe qu'on se retrouve dans la situation où les individus étaient poolés ensemble. Ainsi, la méthode d'estimation Bayésienne hiérarchique généralise les deux approches précédentes, et permet par le choix des priors populationnels de donner plus ou moins d'importance à l'effet population.

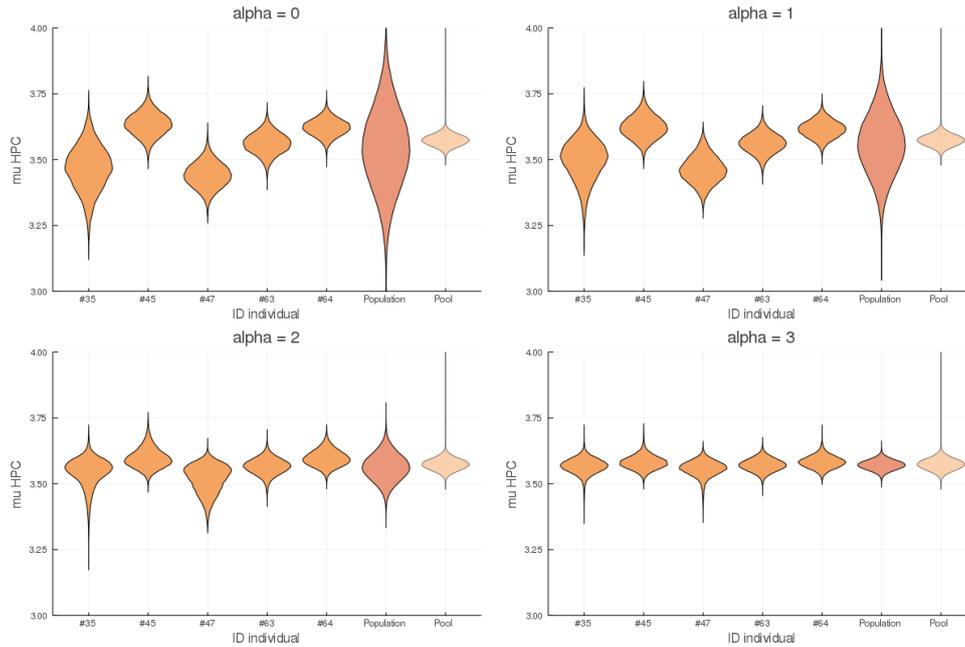


FIGURE 13 – Estimations des paramètres  $\mu_{\text{HPC}}$  pour chaque individu ainsi que l'effet population suivant différents choix de prior sur la variance de la distribution de population (lois gamma-inverse  $(\alpha, 0)$ ). Plus  $\alpha$  est grand, plus on fait l'hypothèse a priori que la variance dans la population doit être faible et plus on diminue la variabilité inter-individuelle.

## Références

- [1] Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6) :716–723, 1974.
- [2] Kenneth P Burnham and David R Anderson. A practical information-theoretic approach. *Model selection and multimodel inference*, 2, 2002.
- [3] M Carolina Florian, Markus Klose, Mehmet Sacma, Jelena Jablanovic, Luke Knudson, Kalpana J Nattamai, Gina Marka, Angelika Vollmer, Karin Soller, Vadim Sakk, et al. Aging alters the epigenetic asymmetry of hsc division. *PLoS biology*, 16(9) :e2003389, 2018.
- [4] Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. Bayesian data analysis chapman & hall. *CRC Texts in Statistical Science*, 2004.
- [5] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6) :721–741, 1984.
- [6] Schwarz Gideon et al. Estimating the dimension of a model. *The annals of statistics*, 6(2) :461–464, 1978.
- [7] W Keith Hastings. Monte carlo sampling methods using markov chains and their applications. 1970.
- [8] William H Kruskal and W Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260) :583–621, 1952.
- [9] Artémis Llamosi, Andres M Gonzalez-Vargas, Cristian Versari, Eugenio Cinquemani, Giancarlo Ferrari-Trecate, Pascal Hersen, and Gregory Batt. What population reveals about individual cell identity : single-cell parameter estimation of models of gene expression in yeast. *PLoS computational biology*, 12(2) :e1004706, 2016.
- [10] Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60, 1947.
- [11] J Patrick Meyer and Michael A Seaman. Expanded tables of critical values for the kruskal-wallis h statistic. In *annual meeting of the American Educational Research Association, San Francisco*, 2006.
- [12] Frank Wilcoxon. Individual comparisons by ranking methods. In *Breakthroughs in statistics*, pages 196–202. Springer, 1992.