# Asymétrie dans la division cellulaire

## un exemple de traitement statistique de données issues de la biologie

Benoîte de Saporta IMAG, Montpellier




# Plan

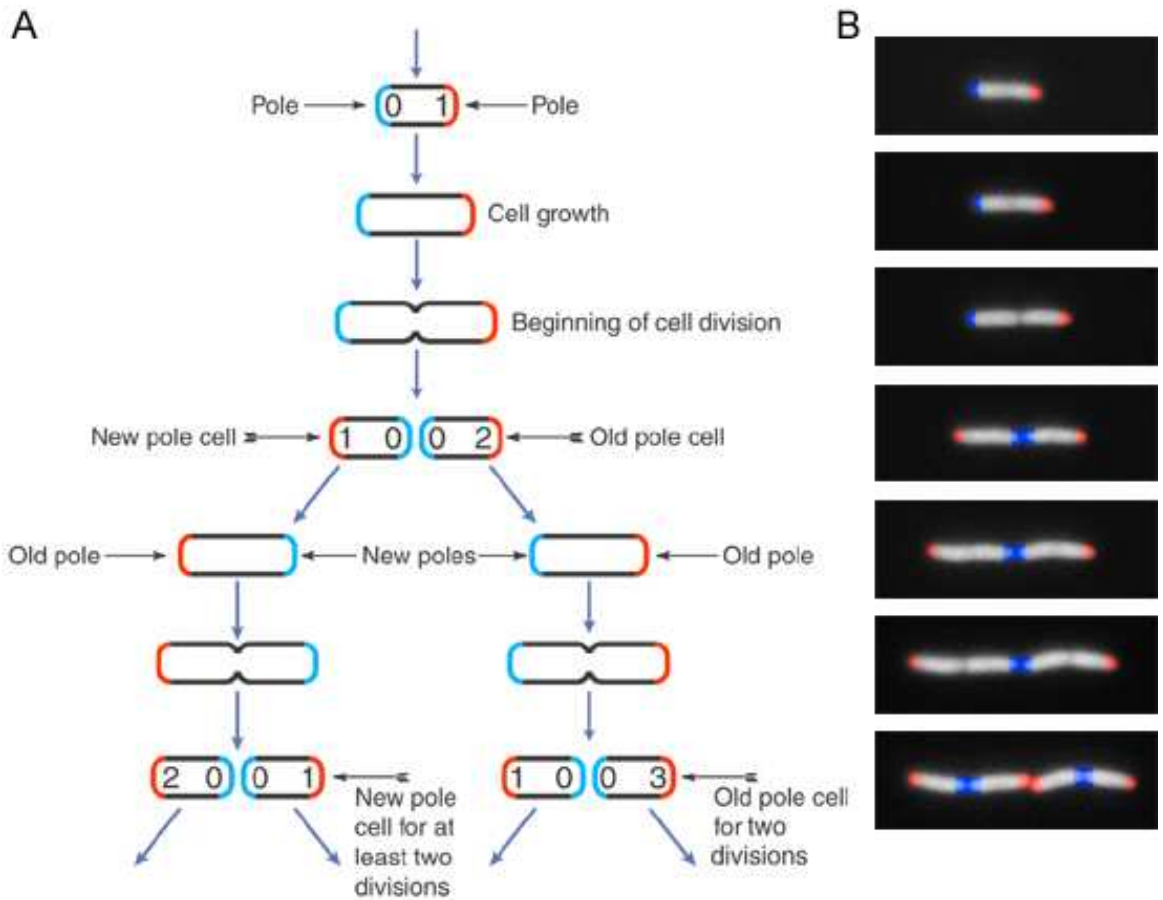
# Asymétrie de la division cellulaire

Est-ce que les organismes unicellulaires vieillissent ?
Est-ce que la division cellulaire est symétrique ?

- pas de signe visible de vieillissement des organismes unicellulaires
- deux cellules filles sont génétiquement identiques à leur mère
- la partage du matériel cellulaire est-il identique entre deux cellules filles ?

# Une mesure d'âge pour Escherichia coli

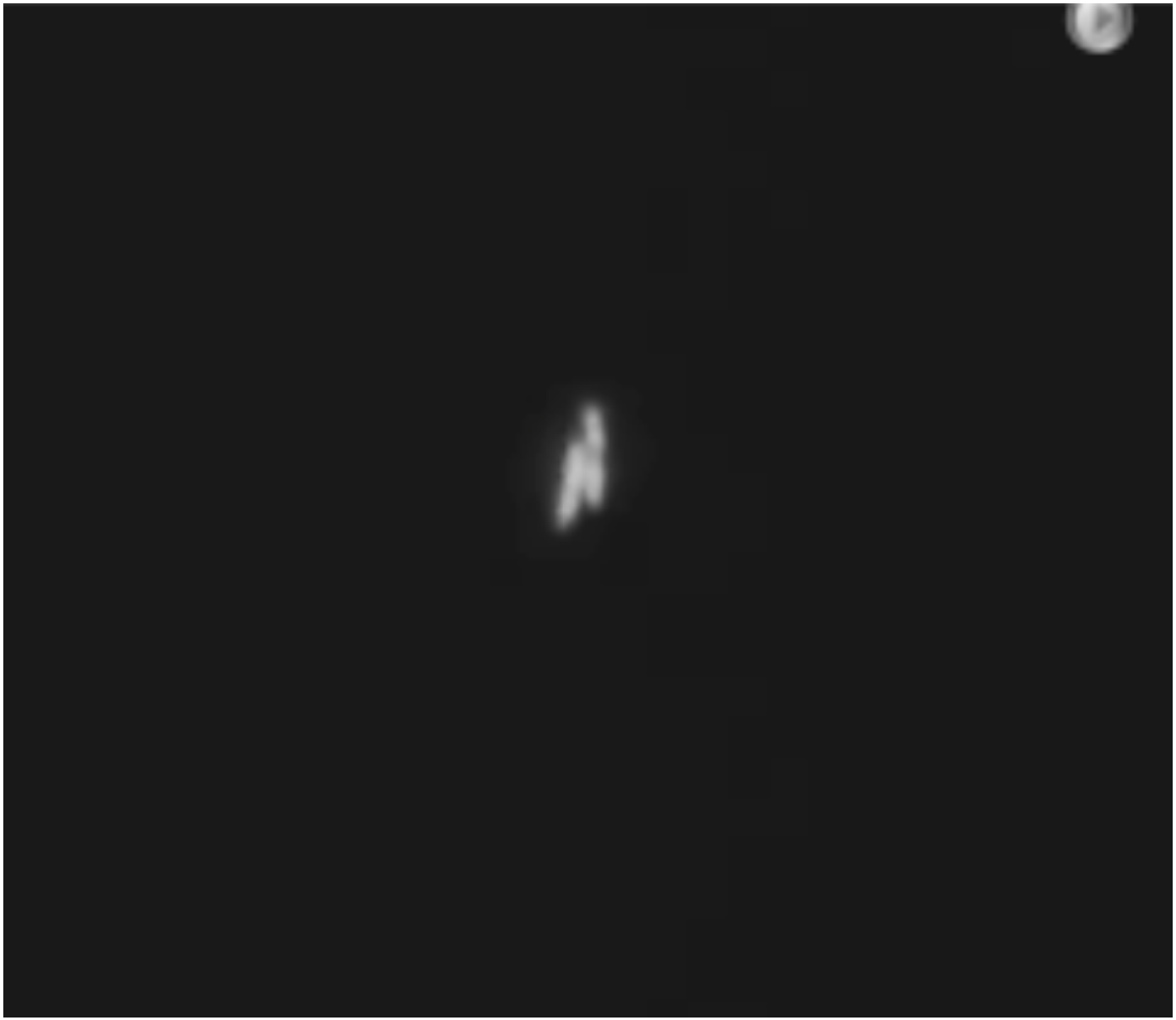
[Stewart & al. 2005]



# Expériences de Stewart et al.

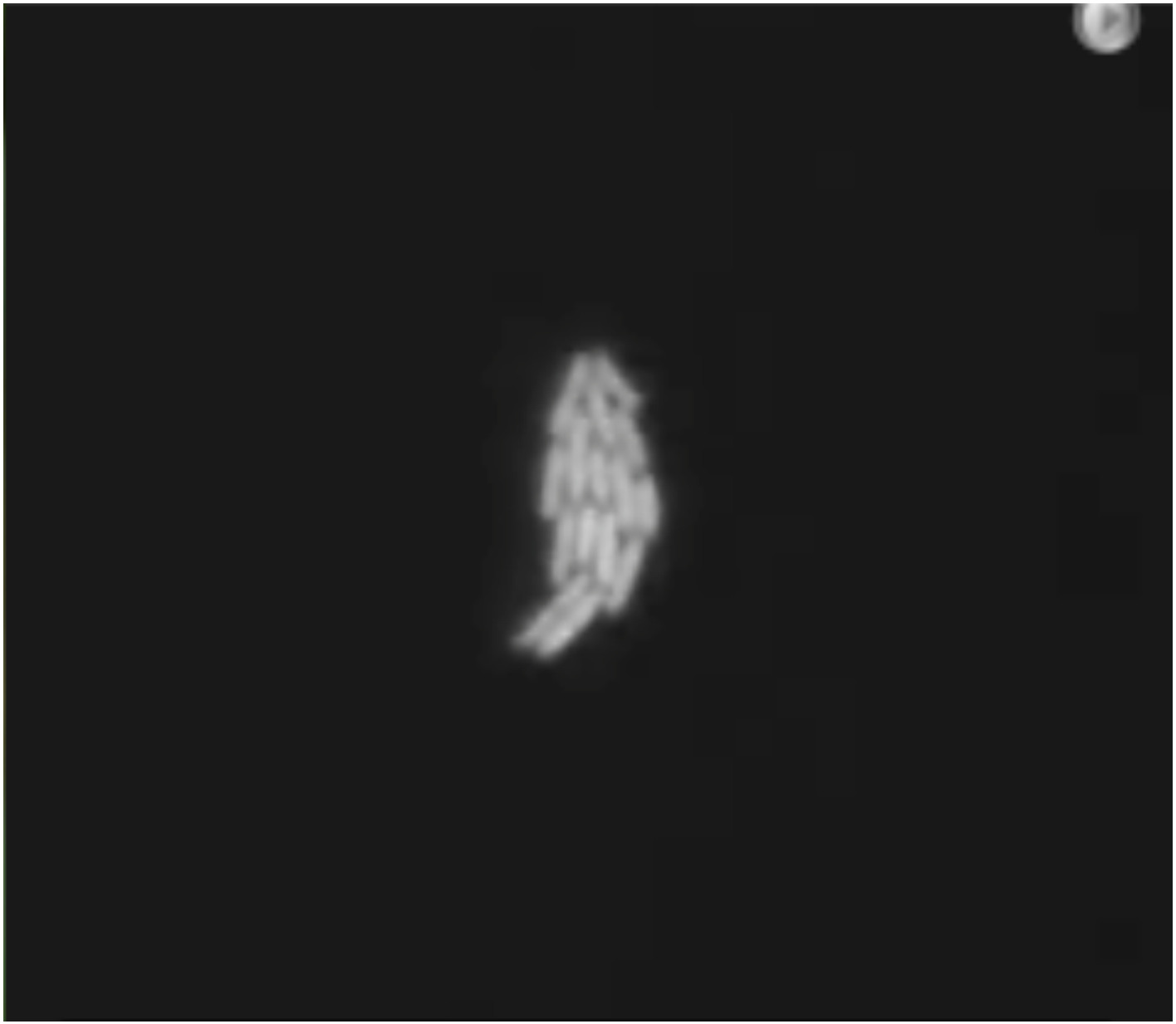
# Expériences de Stewart et al.

# Expériences de Stewart et al.



# Expériences de Stewart et al.

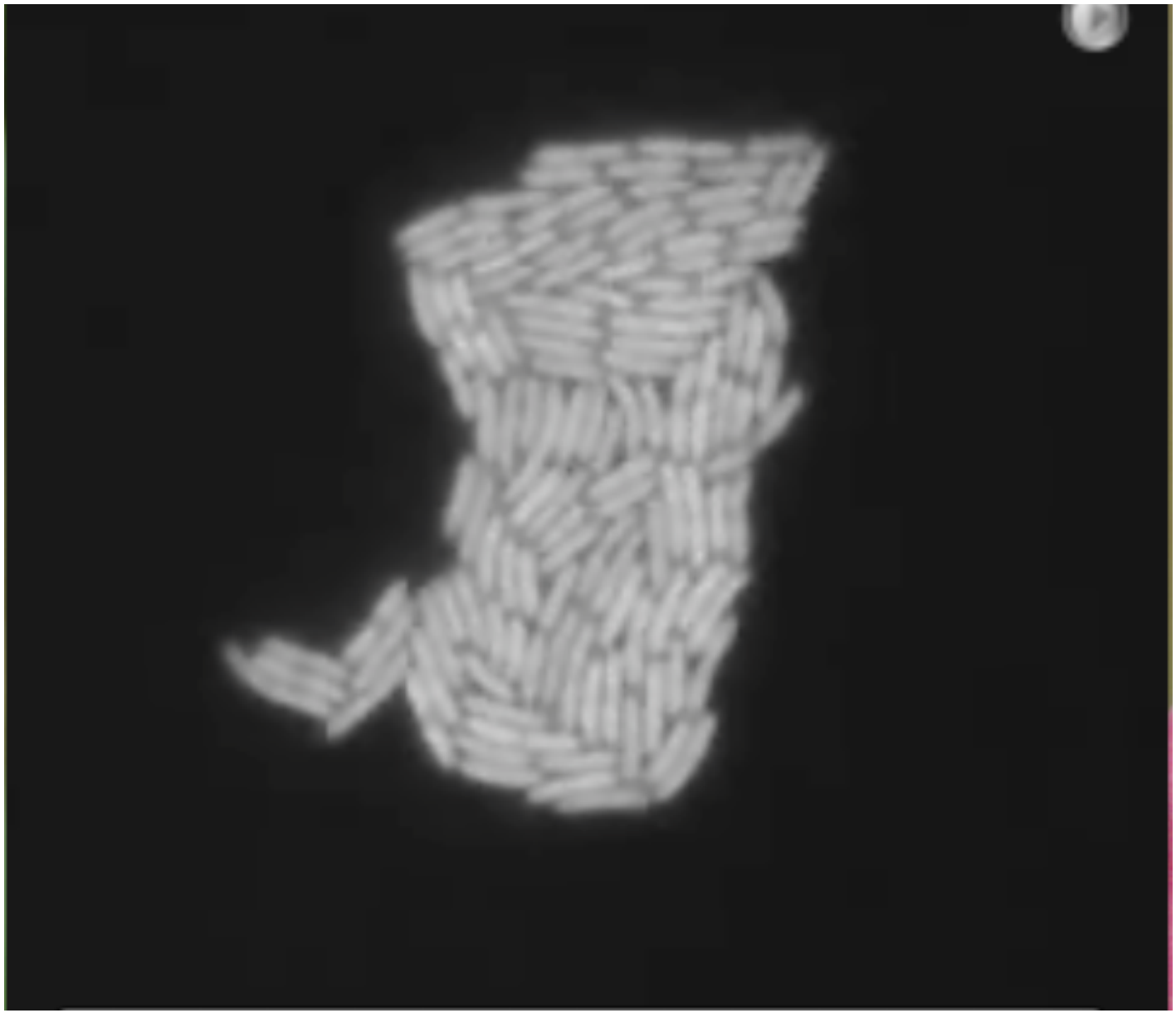
# Expériences de Stewart et al.



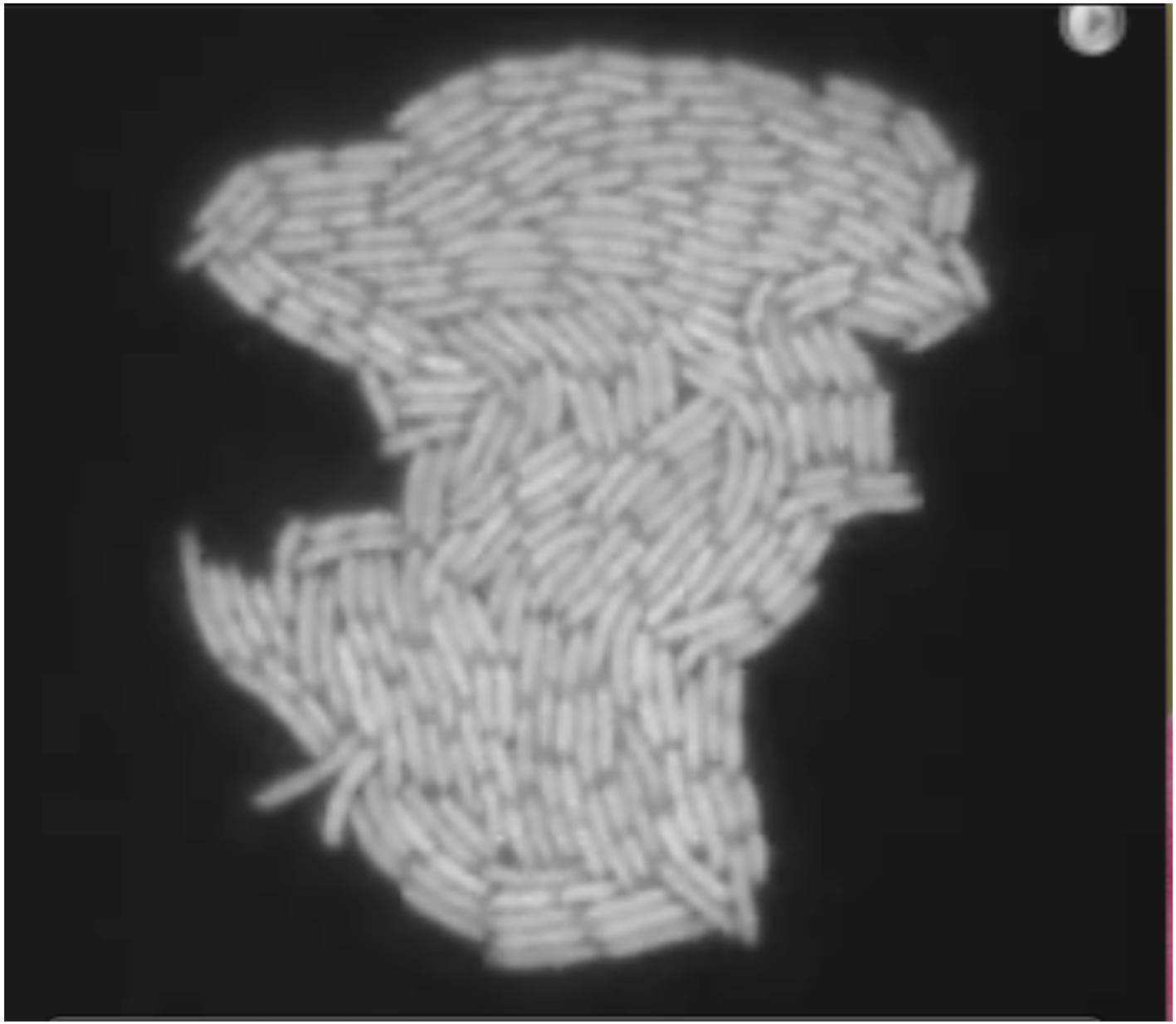
# Expériences de Stewart et al.

# Expériences de Stewart et al.

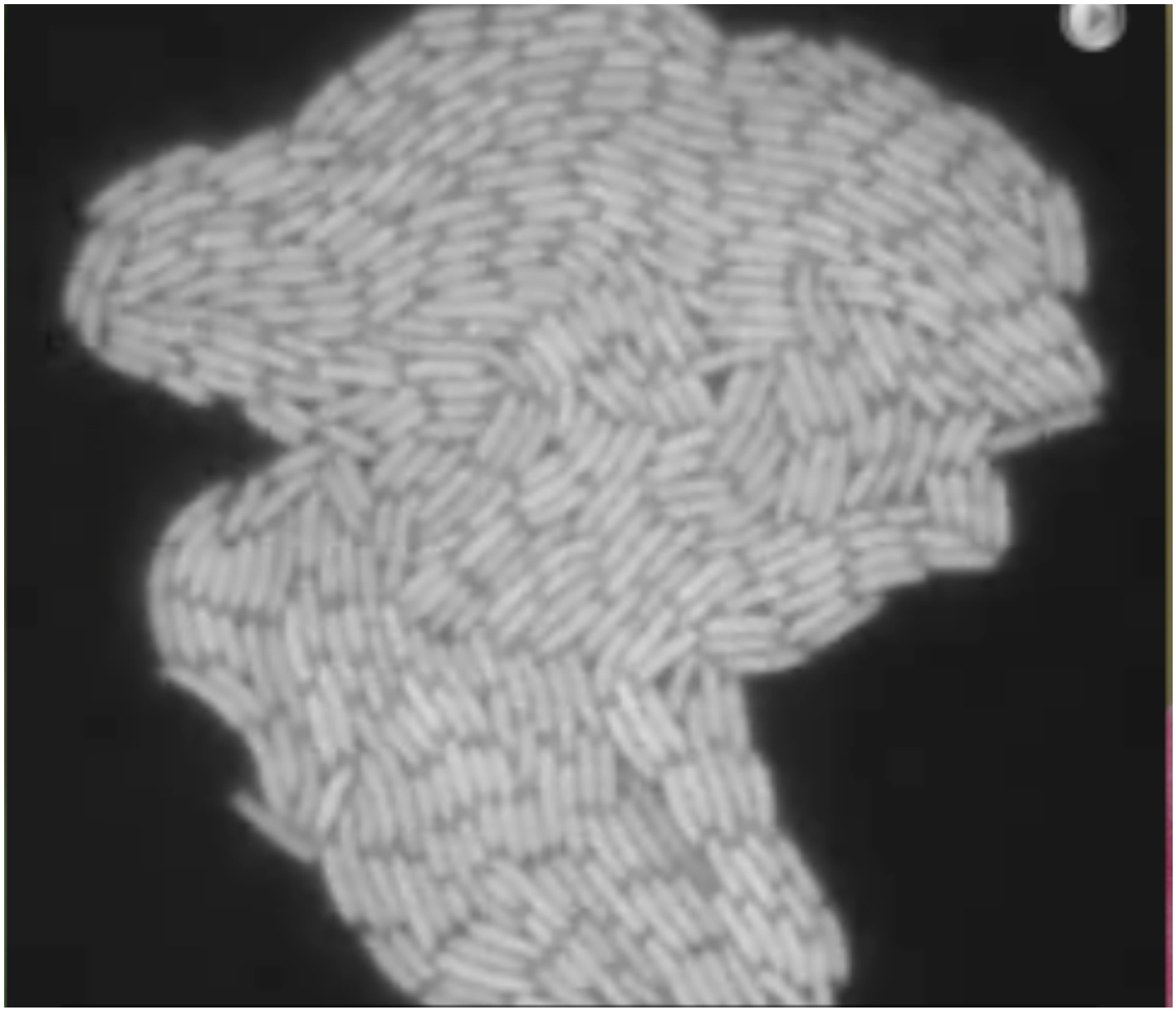
# Expériences de Stewart et al.



# Expériences de Stewart et al.

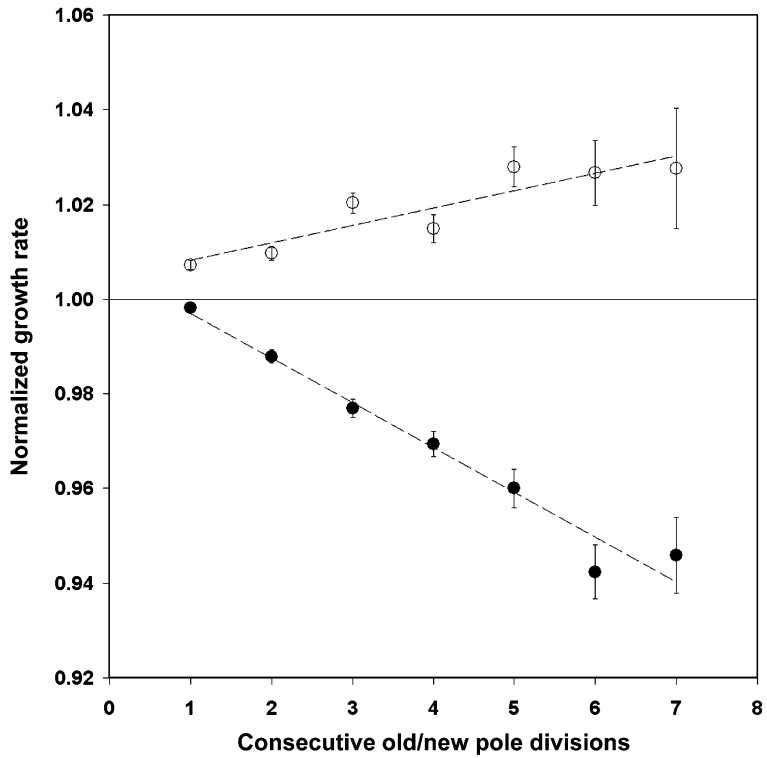


# Expériences de Stewart et al.



# Conclusions de Stewart et al.

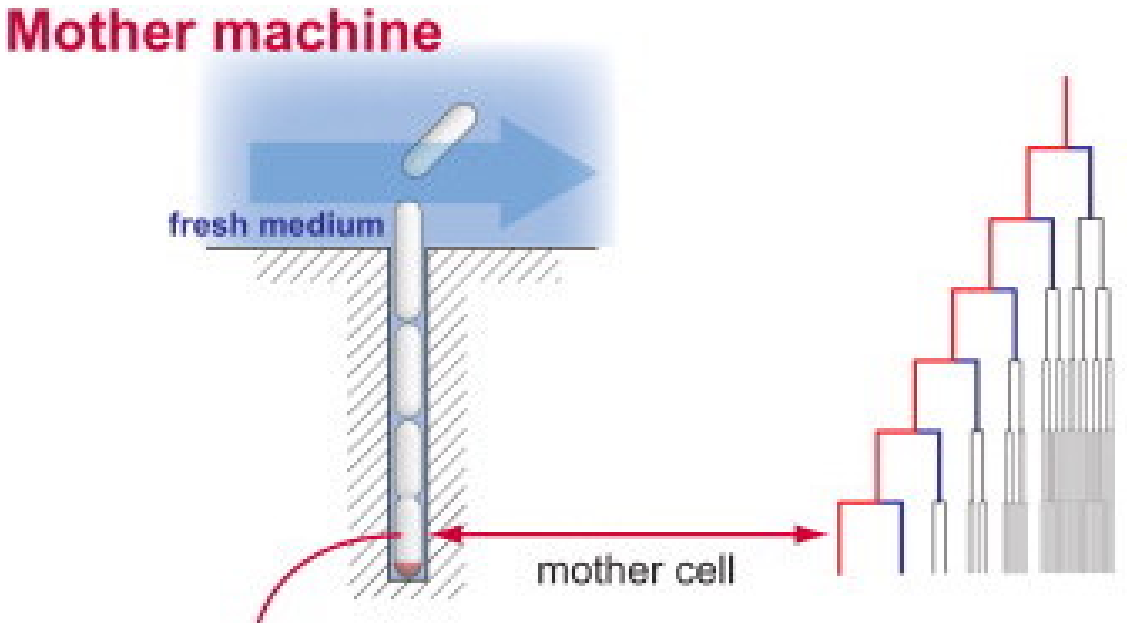
[Stewart & al. 2005]




*"(...) the cell that inherits the old pole exhibits a diminished growth rate, decreased offspring production, and an increased incidence of death."*

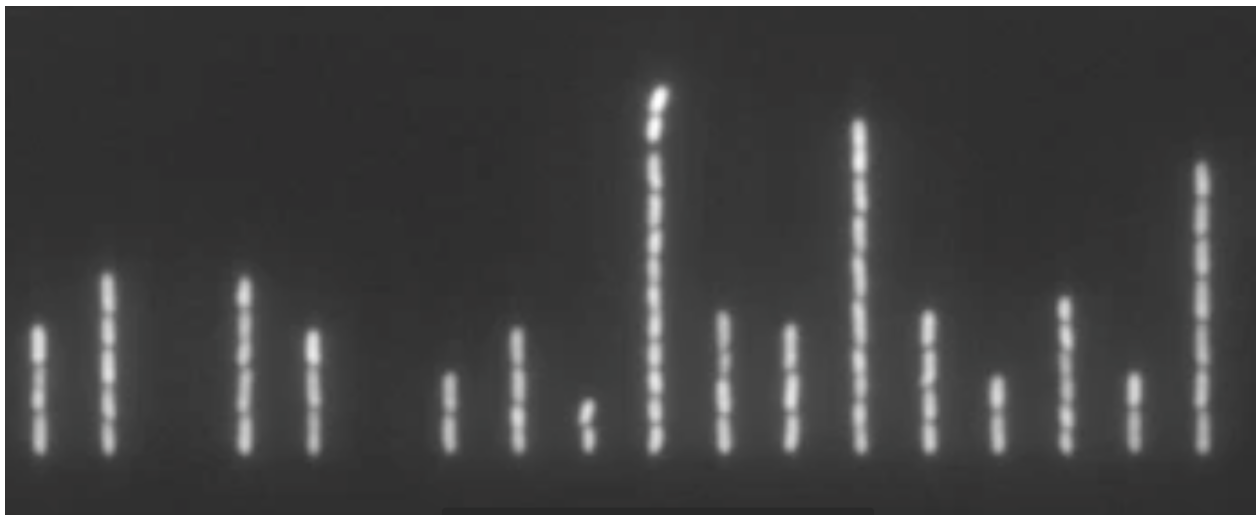
# Taux de croissance avec beaucoup plus de divisions ?

[Wang & al. 2012]



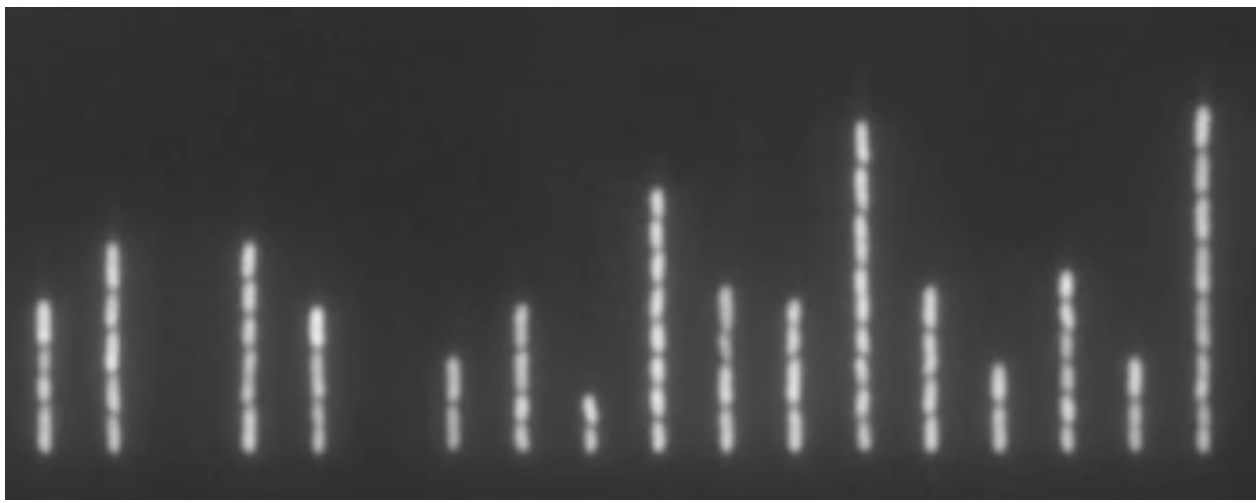
# Expériences de Wang et al.

[Wang & al. 2012]



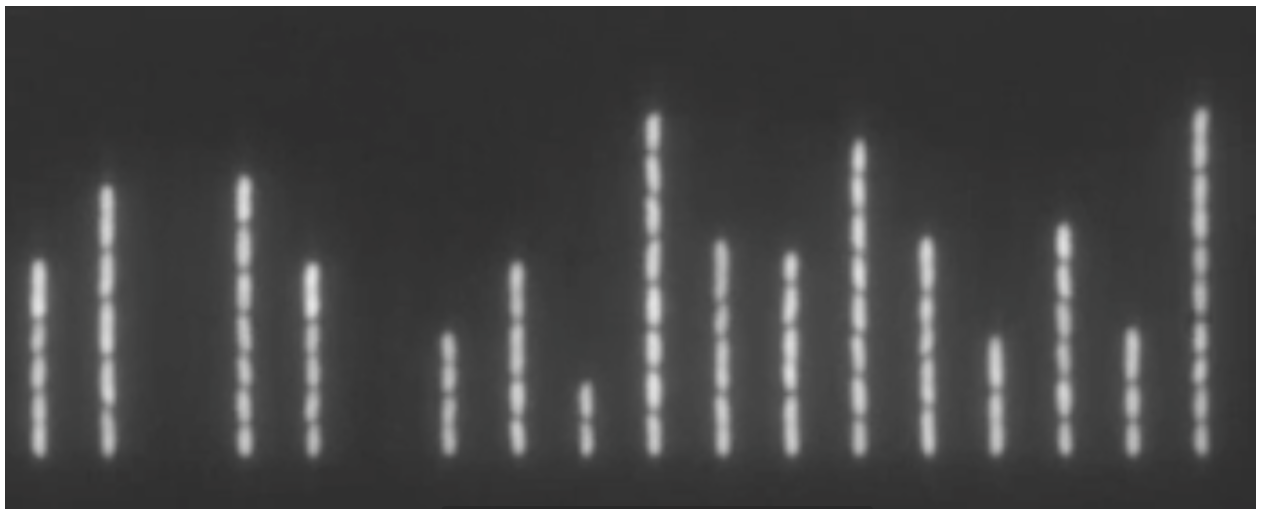
# Expériences de Wang et al.

[Wang & al. 2012]



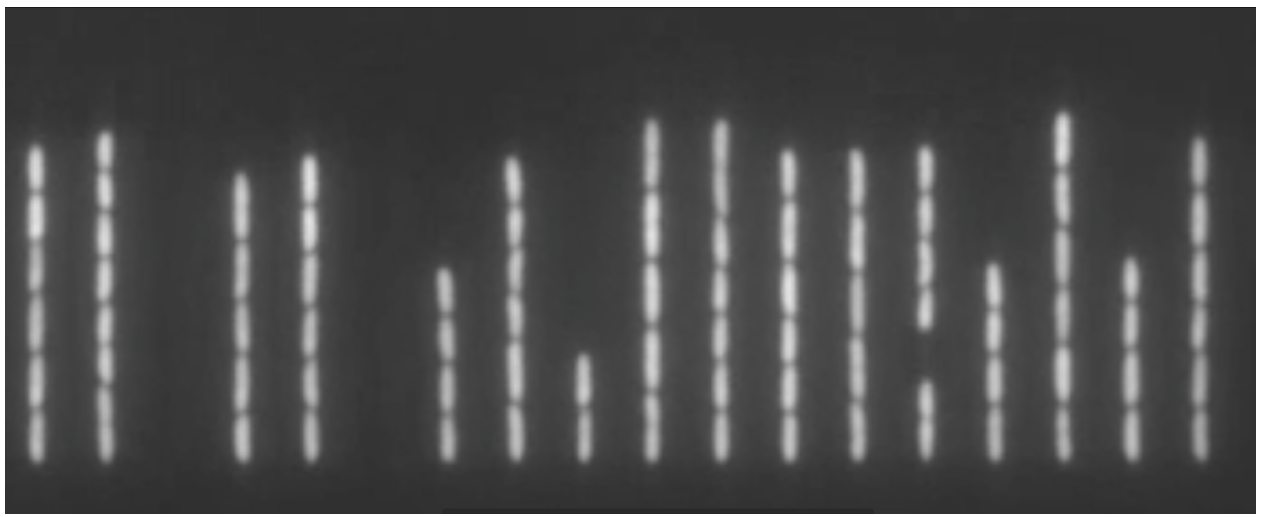
# Expériences de Wang et al.

[Wang & al. 2012]



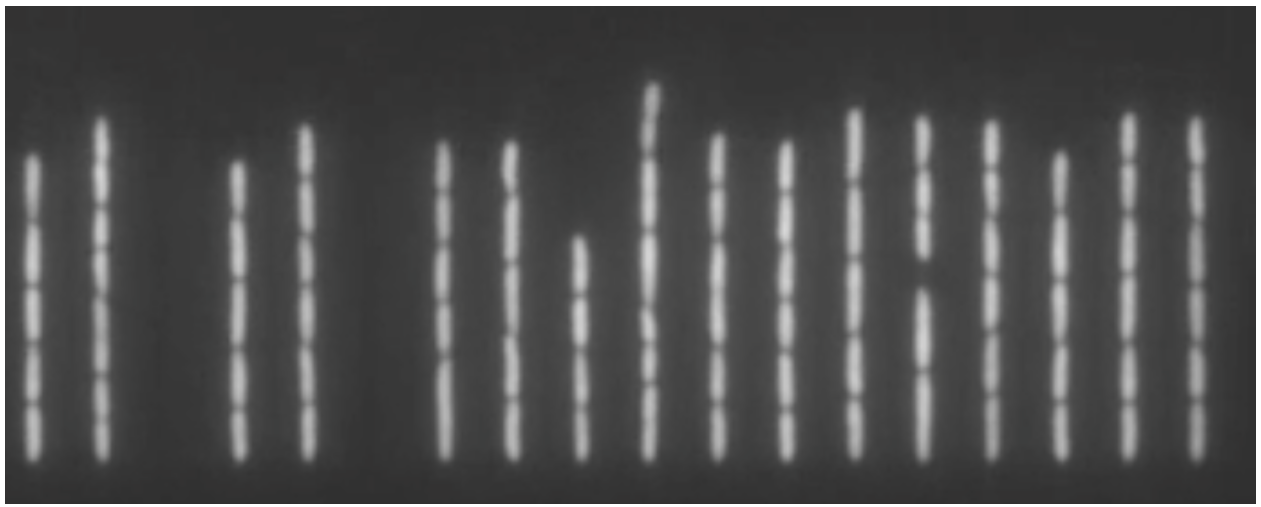
# Expériences de Wang et al.

[Wang & al. 2012]



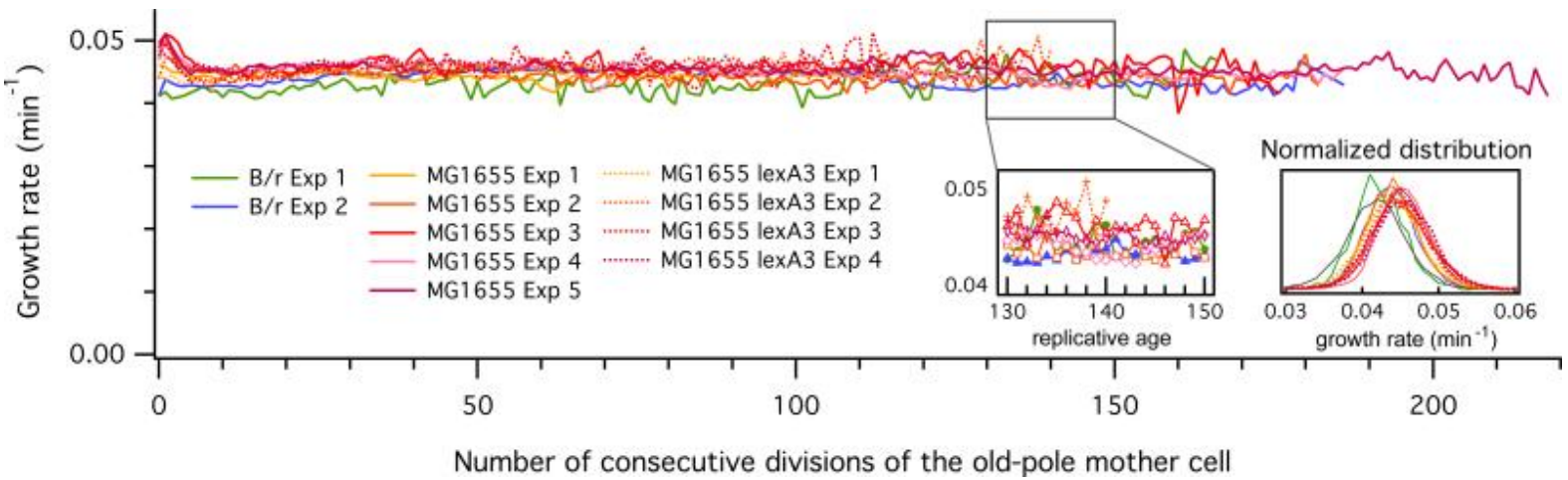
# Expériences de Wang et al.

[Wang & al. 2012]



# Conclusions de Wang et al.

[Wang & al. 2012]




*"Our analysis (...) reveals a remarkable stability of growth whereby the mother cell inherits the same pole for hundreds of generations."*

## Questions

- Les deux expériences sont-elles contradictoires ?
- La division d'E. coli est-elle asymétrique ?

$\Longrightarrow$ Nouvelle analyse statistique des deux jeux de données

# Plan

# Démarche statistique

- recueil des données [Stewart & al. 2005], [Wang & al. 2012]

# Démarche statistique

- recueil des données [Stewart & al. 2005], [Wang & al. 2012]
- statistique exploratoire ou descriptive : synthétiser l'information, la représenter graphiquement

# Démarche statistique

- recueil des données [Stewart & al. 2005], [Wang & al. 2012]

  $\Downarrow$ mise en forme des données, prétraitement $\Downarrow$

- statistique exploratoire ou descriptive : synthétiser l'information, la représenter graphiquement

# Démarche statistique

- recueil des données [Stewart & al. 2005], [Wang & al. 2012]

  $\Downarrow$ mise en forme des données, prétraitement $\Downarrow$

- statistique exploratoire ou descriptive : synthétiser l'information, la représenter graphiquement
- statistique inférentielle : estimer des paramètres, valider ou infirmer des hypothèses au vu des données

# Démarche statistique

- recueil des données [Stewart & al. 2005], [Wang & al. 2012]

  $\Downarrow$ mise en forme des données, prétraitement $\Downarrow$

- statistique exploratoire ou descriptive : synthétiser l'information, la représenter graphiquement
- statistique inférentielle : estimer des paramètres, valider ou infirmer des hypothèses au vu des données
  $\Rightarrow$ modèle probabiliste

# Démarche statistique

- recueil des données [Stewart & al. 2005], [Wang & al. 2012]

  $\Downarrow$ mise en forme des données, prétraitement $\Downarrow$

- statistique exploratoire ou descriptive : synthétiser l'information, la représenter graphiquement
- statistique inférentielle : estimer des paramètres, valider ou infirmer des hypothèses au vu des données
  $\Rightarrow$ modèle probabiliste
- modélisation statistique : chercher des liens (approximatifs) entre les variables

# Statistique inférentielle

But : étant donné un modèle et des réalisations de ce modèle, on veut estimer ses paramètres

# Statistique inférentielle

But : étant donné un modèle et des réalisations de ce modèle, on veut estimer ses paramètres

Exemple

- Modèle : loi normale d'espérance $m$ et de variance $\sigma^2$
- $X_1, X_2, \ldots, X_n$ variables iid suivant ce modèle
- paramètre d'intérêt $\theta = (m, \sigma^2)$

# Statistique inférentielle

But : étant donné un modèle et des réalisations de ce modèle, on veut estimer ses paramètres

Exemple

- Modèle : loi normale d'espérance $m$ et de variance $\sigma^2$
- $X_1, X_2, \ldots, X_n$ variables iid suivant ce modèle
- paramètre d'intérêt $\theta = (m, \sigma^2)$

Un estimateur est une quantité $\widehat{\theta}_n$ calculable à partir des données $X_1, X_2, \ldots, X_n$ et qui *approche* $\theta$.

# Comment estimer les paramètres d'un modèle ? (1/2)

- méthode empirique basée sur la loi des grands nombres
  Exemple estimer $\theta = \mathbb{E}[X]$ par

$$\widehat{\theta}_n = \frac{1}{n}\sum_{k=1}^{n} X_k$$

# Comment estimer les paramètres d'un modèle ? (1/2)

- méthode empirique basée sur la loi des grands nombres
  Exemple estimer $\theta = \mathbb{E}[X]$ par

$$\widehat{\theta}_n = \frac{1}{n}\sum_{k=1}^{n} X_k$$

- méthode des moindres carrés minimise une erreur quadratique
  Exemple estimer $\theta = (a, b)$ dans un modèle $Y = a + bX + \epsilon$ par

$$\widehat{\theta}_n = \arg\min_{(a,b)} \sum_{k=1}^{n} (Y_k - a - bX_k)^2$$

# Comment estimer les paramètres d'un modèle ? (2/2)

- méthode des moments basée sur la résolution d'un système faisant intervenir les moments
  Exemple estimer $\theta = (m, \sigma^2) = (\mathbb{E}[X], \mathbb{V}(X))$ en résolvant

$$\begin{cases} \widehat{m}_n &= \frac{1}{n}\sum_{k=1}^{n} X_k \\ \widehat{\sigma}_n^2 &= \frac{1}{n}\sum_{k=1}^{n}(X_k - \widehat{m}_n)^2 \end{cases}$$

# Comment estimer les paramètres d'un modèle ? (2/2)

- méthode des moments basée sur la résolution d'un système faisant intervenir les moments
  Exemple estimer $\theta = (m, \sigma^2) = (\mathbb{E}[X], \mathbb{V}(X))$ en résolvant

$$\left\{ \begin{array}{rcl} \widehat{m}_n & = & \frac{1}{n}\sum_{k=1}^{n} X_k \\ \widehat{\sigma}_n^2 & = & \frac{1}{n}\sum_{k=1}^{n}(X_k - \widehat{m}_n)^2 \end{array} \right.$$

- méthode du maximum de vraisemblance maximise la fonction de vraisemblance ou son logarithme Exemple pour $X$ de loi $\mathcal{N}(m, \sigma^2)$ et $\theta = m$

$$\widehat{\theta}_n = \arg\max_m \log\left(\prod_{k=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(X_k-m)^2}{2\sigma^2}}\right) = \arg\max_m -\sum_{k=1}^{n}(X_k-m)^2$$

  impose de connaître la loi du modèle

# Comment mesurer la qualité d'un estimateur ?

- estimateur sans biais si $\mathbb{E}[\widehat{\theta}_n] = \theta$
- estimateur asymptotiquement sans biais si $\mathbb{E}[\widehat{\theta}_n] \to \theta$
- risque quadratique $\mathbb{E}[(\widehat{\theta}_n - \theta)^2] = \mathbb{V}(\widehat{\theta}_n) + (\mathbb{E}[\widehat{\theta}_n] - \theta)^2$

Propriétés souhaitables

- estimateur convergent $\widehat{\theta}_n \to \theta$ ps
- normalité asymptotique $\sqrt{n}(\widehat{\theta}_n - \theta) \to \mathcal{N}(0, v^2)$ en loi

Connaître la loi limite permet d'obtenir des intervalles de confiance et de faire des tests

# Intervalles de confiance

Un estimateur $\widehat{\theta}_n$ donne une approximation de la valeur du paramètre $\theta$.

Un intervalle de confiance donne en plus une indication sur la précision de cette estimation. Si $\sqrt{n}(\widehat{\theta}_n - \theta) \to \mathcal{N}(0, v^2)$ en loi alors

$$\mathbb{P}(\widehat{\theta}_n - 1.96\sqrt{\frac{v^2}{n}} \leq \theta \leq \widehat{\theta}_n + 1.96\sqrt{\frac{v^2}{n}}) \to 95\%$$

Si la valeur de $v^2$ est inconnue, on la remplace par un estimateur

# Définition d'un test statistique

Test : règle de décision permettant de choisir entre deux hypothèses au vu d'une statistique de test $T_n$ dépendant des données.

## Exemple

- Modèle : loi normale d'espérance $m$ et de variance $\sigma^2$
- $X_1, X_2, \ldots, X_n$ variables iid suivant ce modèle
- Tester si $m = 0$ ou $m \neq 0$.

# Hypothèses de test

- Hypothèse nulle H0 : hypothèse de travail qui sera retenue à défaut d"information
- Hypothèse alternative H1

## Décision

- si $T_n <$ seuil, on rejette H0, le test est significatif
- si $T_n >$ seuil, on ne rejette pas H0, le test est non significatif, les données ne contredisent pas H0

# Risques d'erreur dans un test

| | Réalité (inconnue) $H_0$ | Réalité (inconnue) $H_1$ |
|---|---|---|
| Décision : $H_0$ | bonne décision | erreur $\beta$ |
| Décision : $H_1$ | erreur $\alpha$ | bonne décision |

- risque de première espèce : probabilité de rejeter $H_0$ à tort

$$\alpha = \mathbb{P}(\text{rejet de } H_0 | H_0 \text{ vraie}).$$

- risque de deuxième espèce : probabilité de accepter $H_0$ à tort

$$\beta = \mathbb{P}(\text{accepter } H_0 | H_0 \text{ fausse}).$$

puissance d'un test $\pi = 1 - \beta = \mathbb{P}(\text{rejeter } H_0 | H_0 \text{ fausse}).$

# Comment construire un test ?

- Choisir le risque de première espèce $\alpha$ par exemple 5%
- Trouver une statistique de test $T_n$ de loi (asymptotique) connue sous H0 avec $T_n \to \infty$ sous H1
  Exemple Si $\sqrt{n}(\widehat{\theta}_n - \theta) \to \mathcal{N}(0, v^2)$ en loi sous H0, alors $T_n = \frac{n}{v^2}(\widehat{\theta}_n - \theta)^2$ suit une loi $\chi^2$ à 1 degré de liberté
- Le seuil de la règle de décision correspond au quantile $1 - \alpha$ de la loi sous H0
  Exemple si $T_n > 3.841$ rejet, si $T_n < 3.841$ non rejet

En général il est difficile de calculer la puissance du test

# Division cellulaire

## Questions

- Les deux expériences sont-elles contradictoires ?
- La division d'E. coli est-elle asymétrique ?

$\Longrightarrow$ Nouvelle analyse statistique des deux jeux de données

# Démarche statistique

- recueil des données [Stewart & al. 2005], [Wang & al. 2012]

  $\Downarrow$ mise en forme des données, prétraitement $\Downarrow$

- statistique exploratoire ou descriptive : synthétiser l'information, la représenter graphiquement
- statistique inférentielle : estimer des paramètres, valider ou infirmer des hypothèses au vu des données
  $\Rightarrow$ modèle probabiliste
- modélisation statistique : chercher des liens (approximatifs) entre les variables

# Plan

# Format original des données

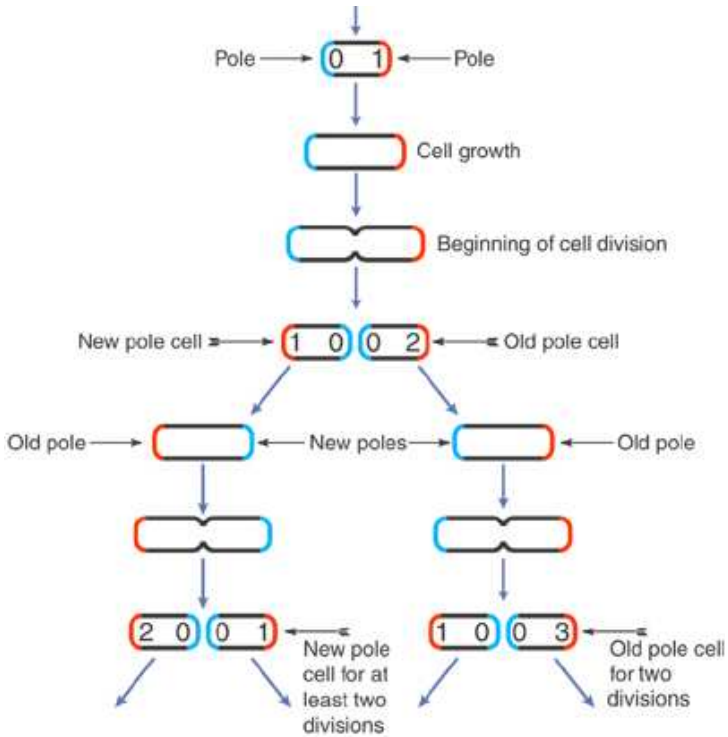
- 94 fichiers .dat correspondant aux 94 films, identifiés par la date et le numéro de l'expérience

2002-10-02-1.dat

| Simple Name | Cell Length | LSF Log2 length rate | Border Distance |
|---|---|---|---|
| 0T | 66.006881713867187 | 0.021434691117359853 | 3.7708301544189453 |
| 0H | 70.550323486328125 | 0.024786635298190567 | 5.0358486175537109 |
| 0HH | 64.124320983886719 | 0.032522424132207212 | 3.2866756916046143 |
| 0HT | 69.74560546875 | 0.028880125030165918 | 4.7025094032287598 |
| 0TH | 75.189002990722656 | 0.030998329618137591 | 3.7274696826934814 |
| 0TT | 75.981613159179688 | 0.029723926616769265 | 2.3916752338409424 |
| 0HTT | 64.626998901367188 | 0.03311138136990853 | 3.1578500270843506 |
| 0HTH | 65.085319519042969 | 0.035072581550857233 | 7.2902283668518066 |
| 0HHT | 81.493141174316406 | 0.036126451673668375 | 2.8658080101013184 |

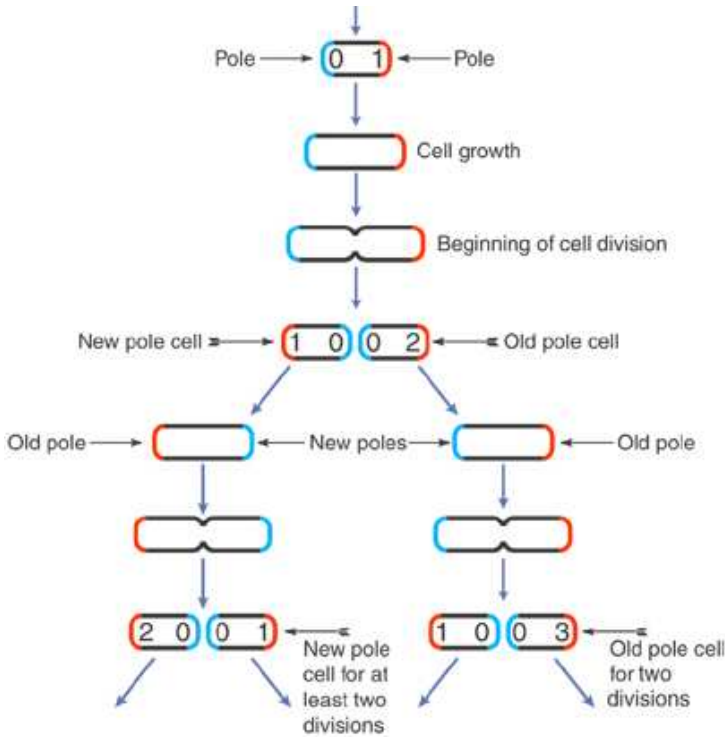
- garder uniquement le taux de croissance
- numéroter les cellules pour reconstruire la généalogie

# Numérotation des cellules



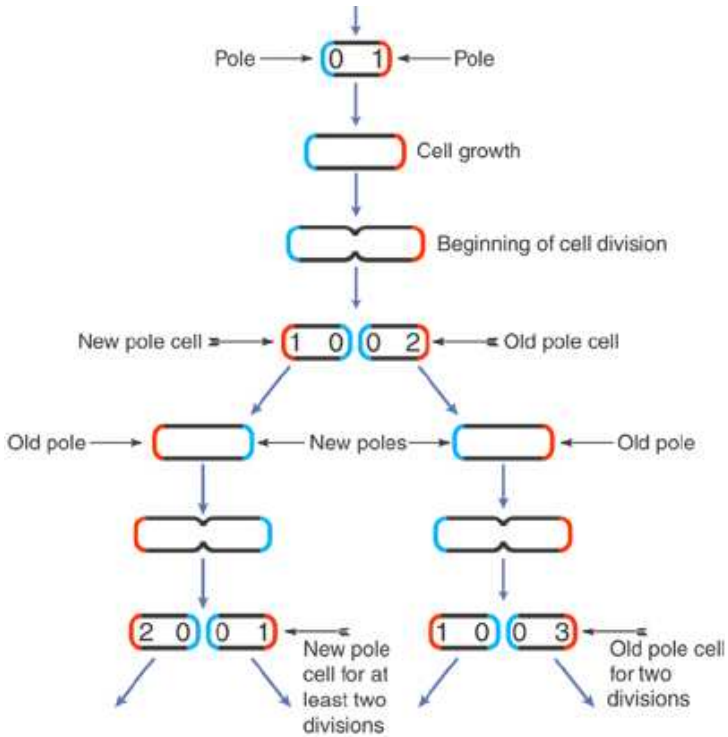

0T indéterminé

# Numérotation des cellules



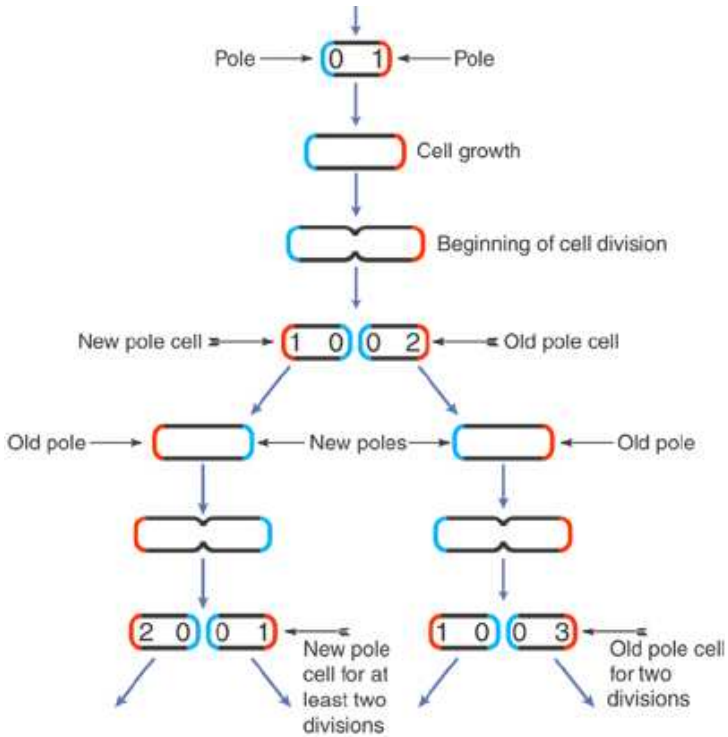

| | |
|---|---|
| 0T | indéterminé |
| 0H | indéterminé |

# Numérotation des cellules



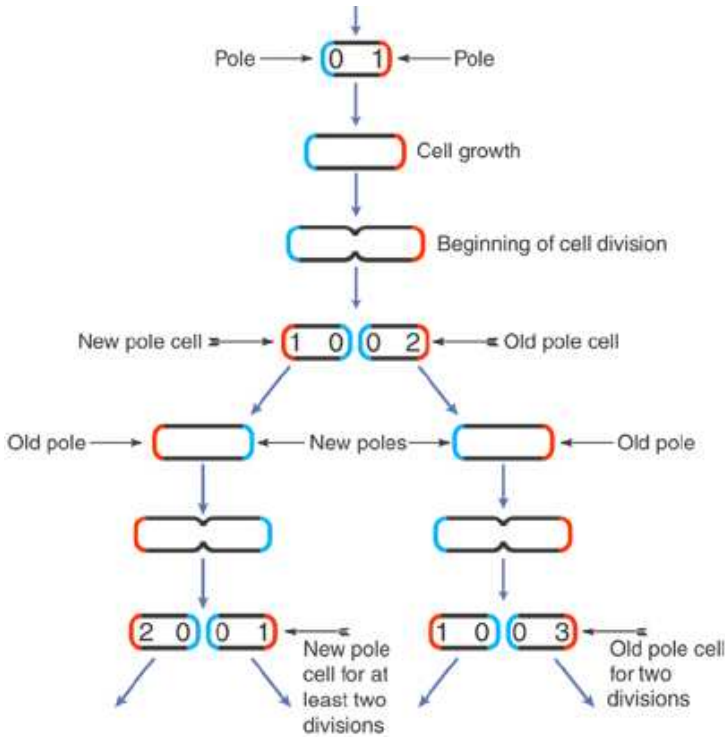

| | |
|---|---|
| 0T | indéterminé |
| 0H | indéterminé |
| 0HH | vieux pôle |

# Numérotation des cellules



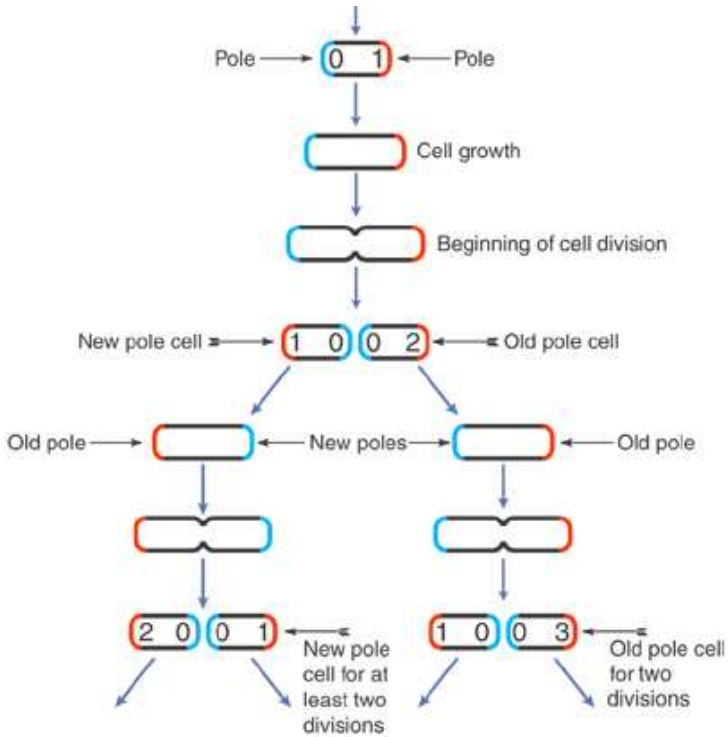

| | |
|---|---|
| 0T | indéterminé |
| 0H | indéterminé |
| 0HH | vieux pôle |
| 0HT | nouveau pôle |

# Numérotation des cellules



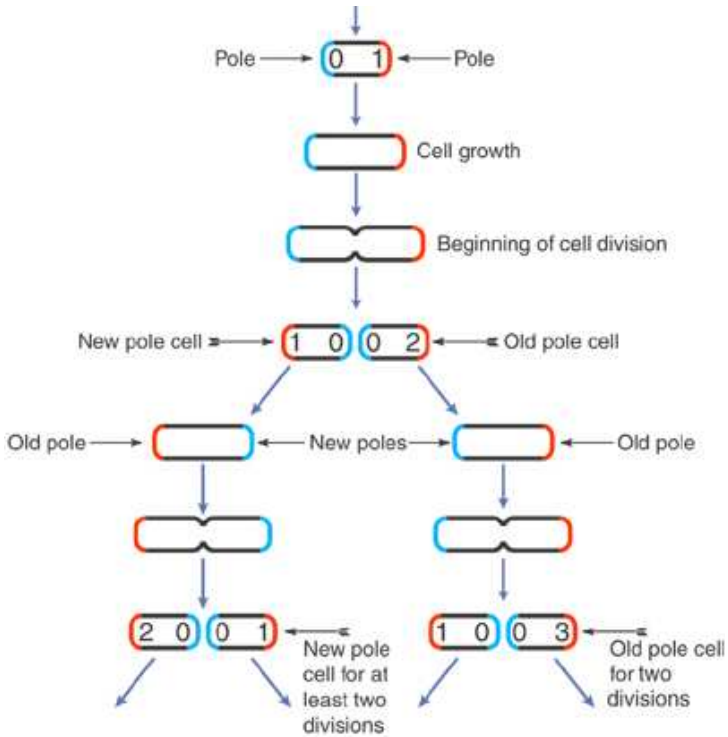

| | |
|---|---|
| 0T | indéterminé |
| 0H | indéterminé |
| 0HH | vieux pôle |
| 0HT | nouveau pôle |
| 0TH | nouveau pôle |

# Numérotation des cellules



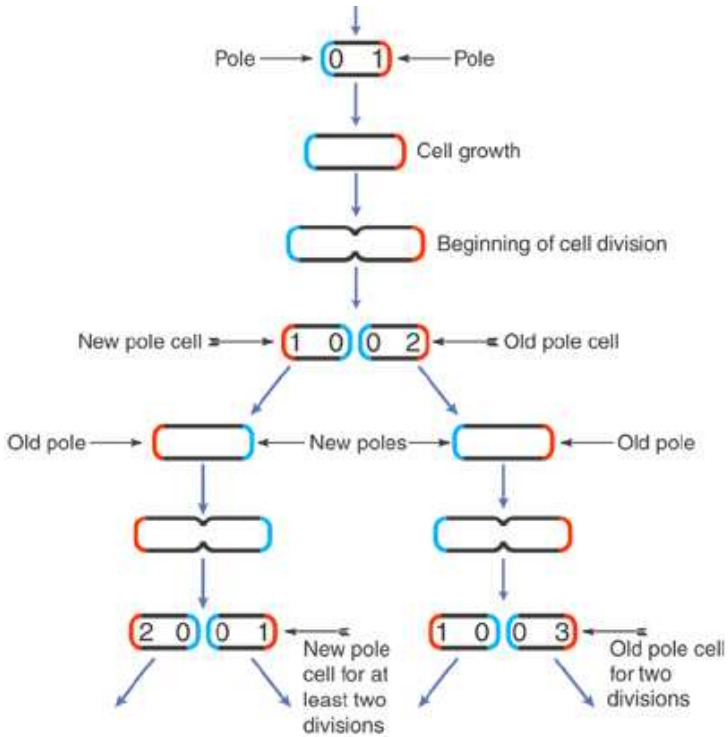

| | |
|---|---|
| 0T | indéterminé |
| 0H | indéterminé |
| 0HH | vieux pôle |
| 0HT | nouveau pôle |
| 0TH | nouveau pôle |
| 0TT | vieux pôle |

# Numérotation des cellules



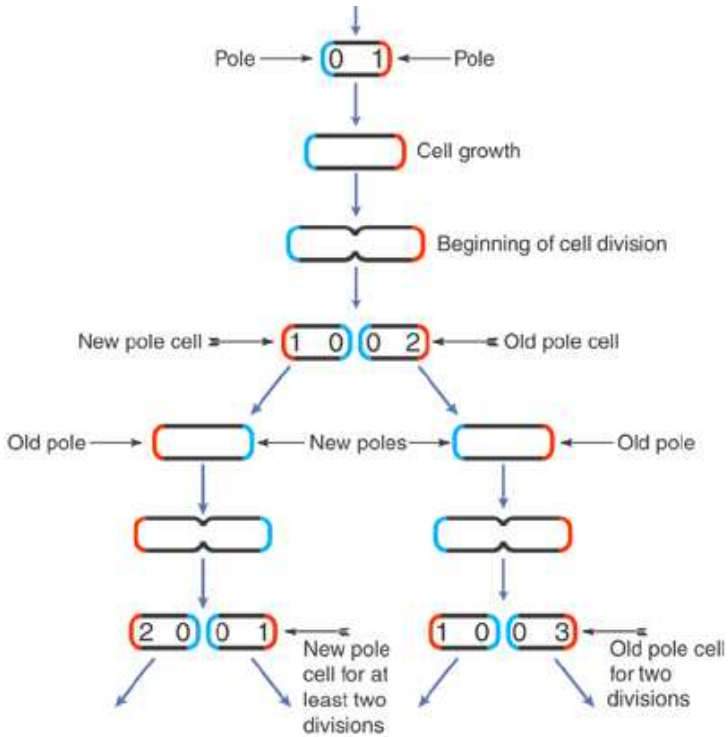

| | |
|---|---|
| 0T | indéterminé |
| 0H | indéterminé |
| 0HH | vieux pôle |
| 0HT | nouveau pôle |
| 0TH | nouveau pôle |
| 0TT | vieux pôle |
| 0HTT | nouveau puis |

# Numérotation des cellules



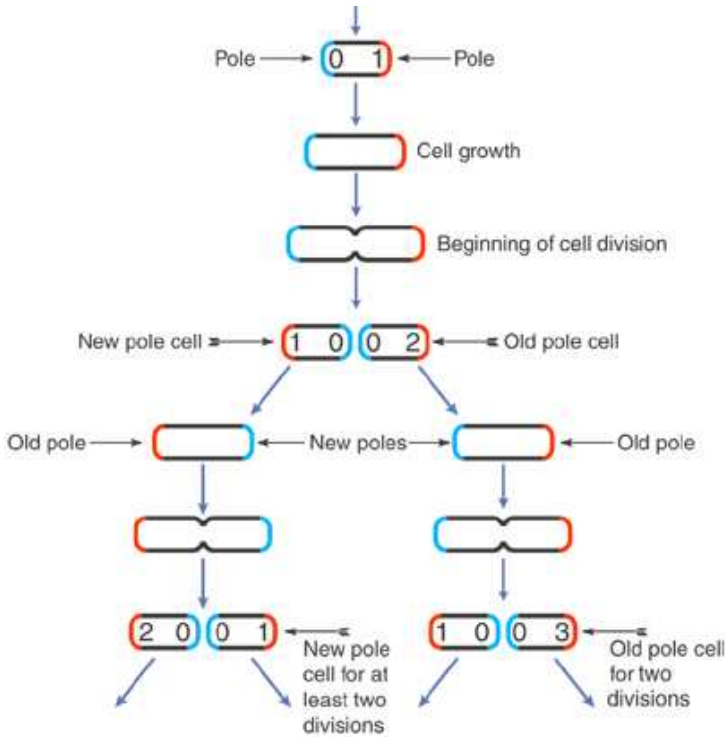

| | |
|---|---|
| 0T | indéterminé |
| 0H | indéterminé |
| 0HH | vieux pôle |
| 0HT | nouveau pôle |
| 0TH | nouveau pôle |
| 0TT | vieux pôle |
| 0HTT | nouveau puis vieux pôle |

# Numérotation des cellules



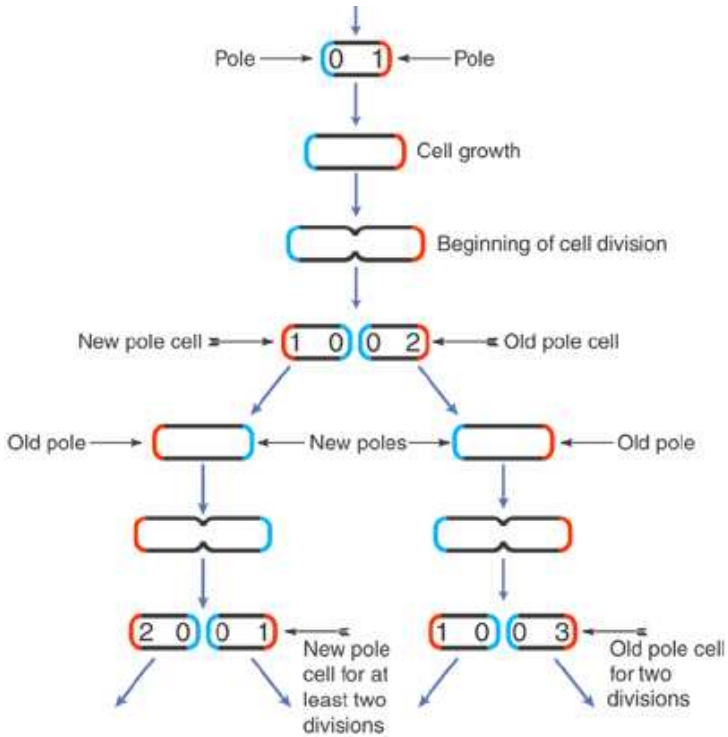

| | |
|---|---|
| 0T | indéterminé |
| 0H | indéterminé |
| 0HH | vieux pôle |
| 0HT | nouveau pôle |
| 0TH | nouveau pôle |
| 0TT | vieux pôle |
| 0HTT | nouveau puis vieux pôle |
| 0HTH | nouveau pôle |

# Numérotation des cellules



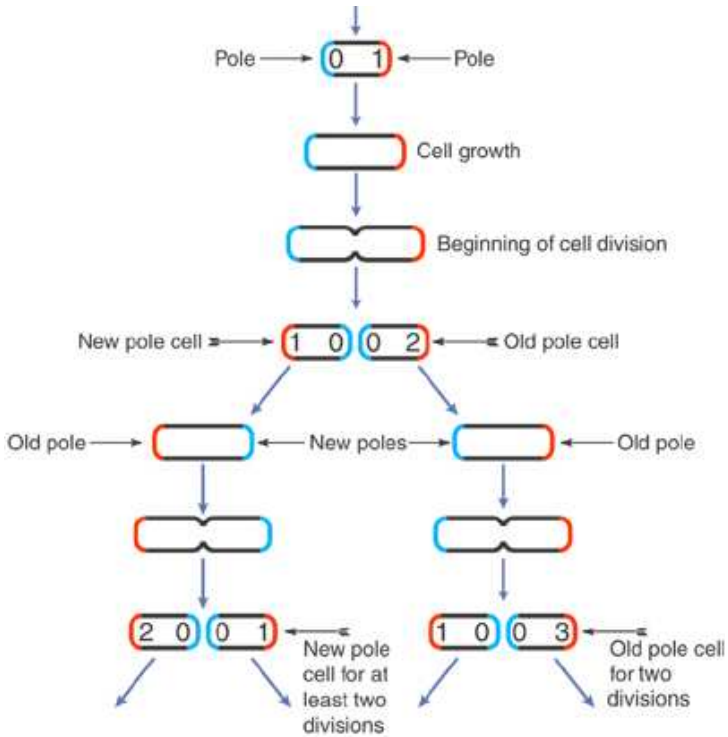

| | |
|---|---|
| 0T | indéterminé |
| 0H | indéterminé |
| 0HH | vieux pôle |
| 0HT | nouveau pôle |
| 0TH | nouveau pôle |
| 0TT | vieux pôle |
| 0HTT | nouveau puis vieux pôle |
| 0HTH | 2 nouveaux pôles |

# Numérotation des cellules




| | |
|---|---|
| 0T | indéterminé |
| 0H | indéterminé |
| 0HH | vieux pôle |
| 0HT | nouveau pôle |
| 0TH | nouveau pôle |
| 0TT | vieux pôle |
| 0HTT | nouveau puis vieux pôle |
| 0HTH | 2 nouveaux pôles |
| 0HHT | vieux puis |

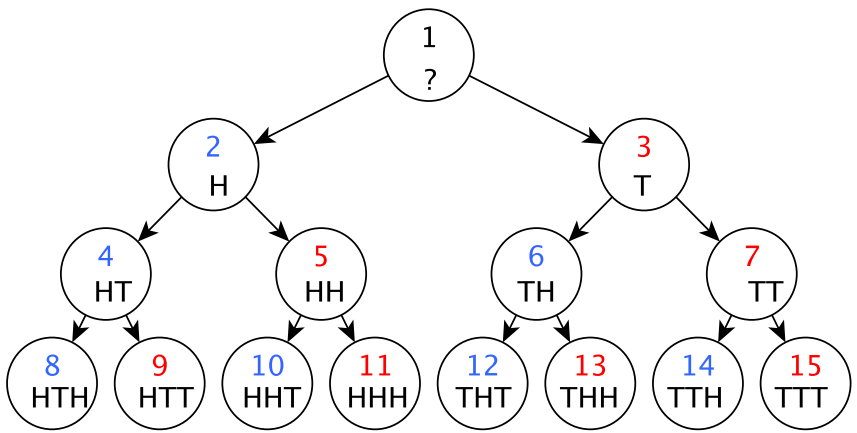
# Numérotation des cellules




| | |
|---|---|
| 0T | indéterminé |
| 0H | indéterminé |
| 0HH | vieux pôle |
| 0HT | nouveau pôle |
| 0TH | nouveau pôle |
| 0TT | vieux pôle |
| 0HTT | nouveau puis vieux pôle |
| 0HTH | 2 nouveaux pôles |
| 0HHT | vieux puis nouveau pôle |

# Codage des cellules

- numérotation correspondant de façon unique à une position dans l'arbre généalogique
- pair : nouveau pôle
- impair : vieux pôle



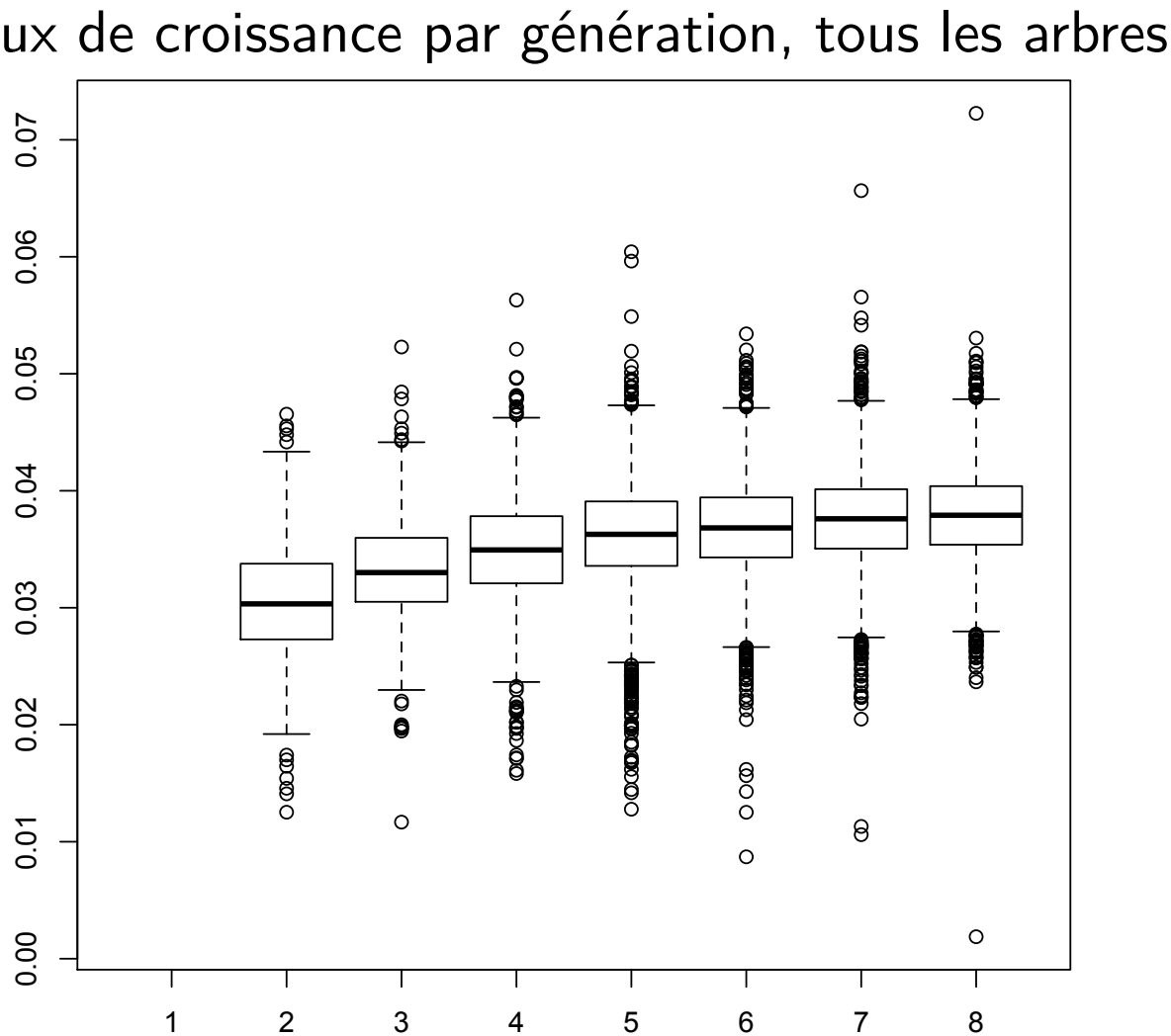

| | |
|---|---|
| 0T | 3 |
| 0H | 2 |
| 0HH | 5 |
| 0HT | 4 |
| 0TH | 6 |
| 0TT | 7 |
| 0HTT | 9 |
| 0HTH | 8 |
| 0HHT | 10 |

# Résumé des données

- 94 films $\rightarrow$ 101 arbres généalogiques
- 4 à 9 générations observées
- 22732 cellules
- nombreuses données manquantes

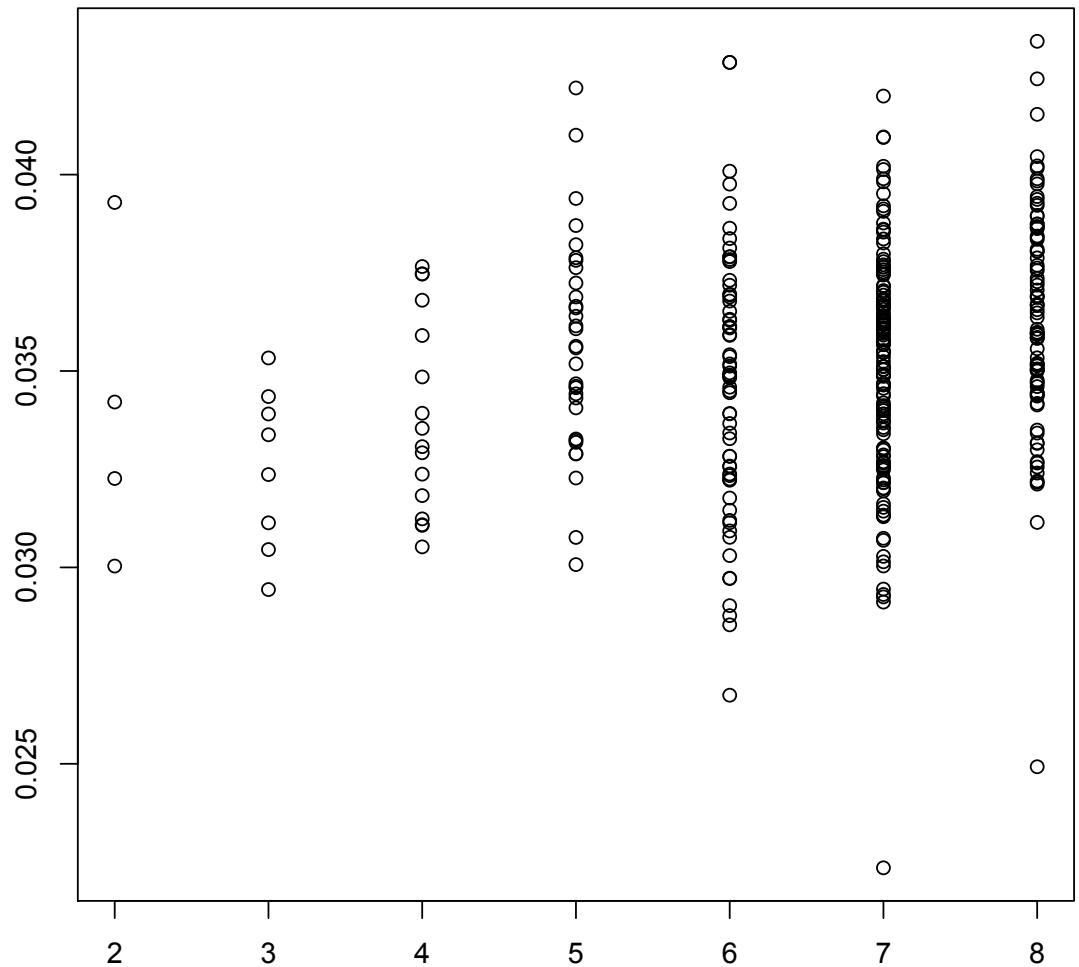
- mesure des taux de croissance
- mesure des types :
  ancien / nouveau pôle à partir de la génération 2
- lignées connues via la numérotation

# Analyse exploratoire



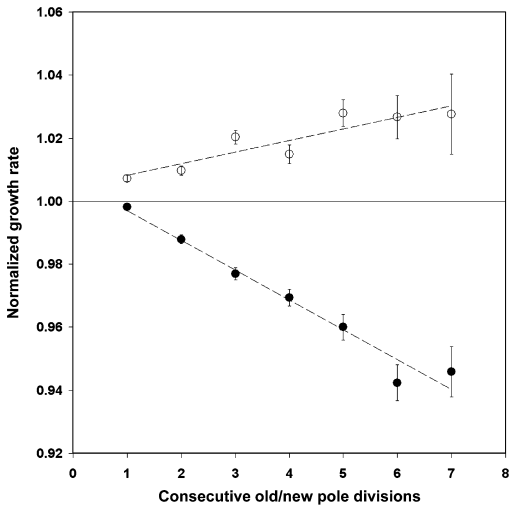
# Analyse exploratoire

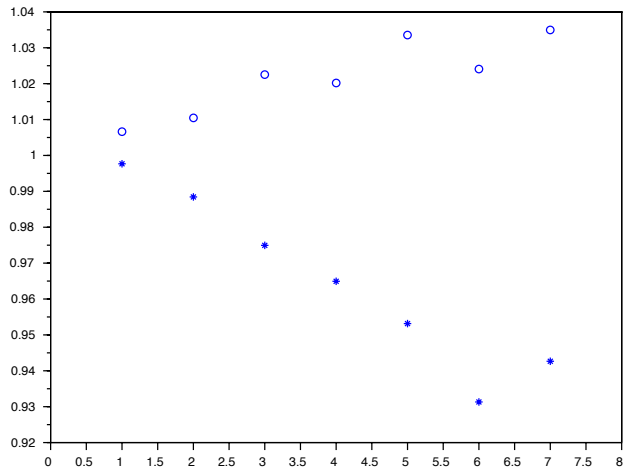
Taux de croissance par génération, arbre 1



# Comparaison des taux de croissance

Reconstitution du graphique de [Stewart & al. 2005]



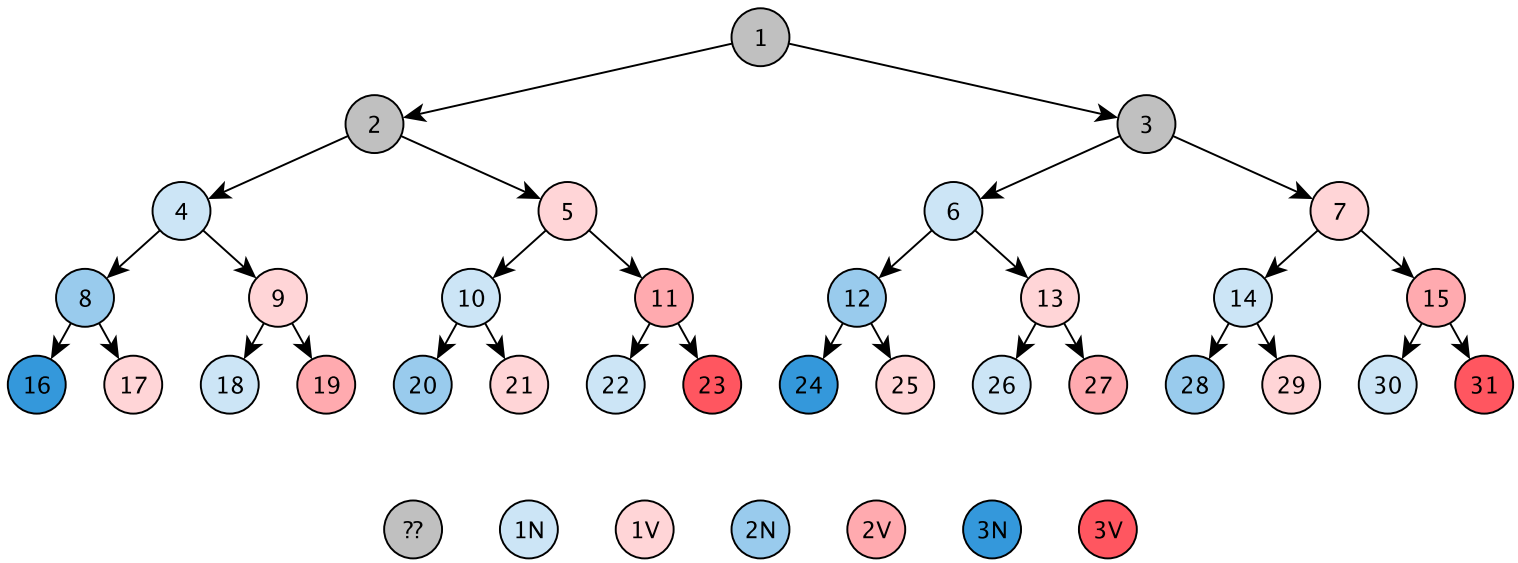



- normalisation par le taux de croissance moyen par génération par arbre
- moyenne entre cellules ayant hérité du même nombre de vieux / nouveaux pôles dans toutes les données

# Questions statistiques sur la comparaison

Problèmes liés à la structure en arbre des données

- chaque point correspond à un nombre de cellules très différent



# Questions statistiques sur la comparaison

Problèmes liés à la structure en arbre des données

- chaque point correspond à un nombre de cellules très différent
- moyenne entre cellules dans des arbres différents mais aussi dans des générations différentes du même arbre $\Rightarrow$ pas d'indépendance

# Questions statistiques sur la comparaison

Problèmes liés à la structure en arbre des données

- chaque point correspond à un nombre de cellules très différent
- moyenne entre cellules dans des arbres différents mais aussi dans des générations différentes du même arbre $\Rightarrow$ pas d'indépendance

On ne peut pas tirer de conclusion rigoureuse d'un tel graphique
$\Rightarrow$ Nouvelle procédure de statistique inférentielle et de test de symétrie

# Objectif

- déterminer si au vu des données il y a une asymétrie significative dans la division d'E. coli
- statistique inférentielle $\Rightarrow$ besoin d'un modèle adapté à la structure d'arbre généalogique des données
  - estimer les paramètres
  - tester la symétrie

# Modélisation

- le taux de croissance d'une cellule fille dépend de celui de sa mère
- deux cellules soeurs peuvent être corrélées
- deux cellules soeurs n'ont pas forcément le même taux de croissance

$\Rightarrow$ Modèle auto-régressif adapté à la structure d'arbre

# Modèle BAR

Modèle auto-régressif de bifurcation
[Cowan & Staudte 1986] [Guyon 2007]

$$\begin{cases} X_{2k} &= a + bX_k + \epsilon_{2k} \\ X_{2k+1} &= c + dX_k + \epsilon_{2k+1} \end{cases}$$

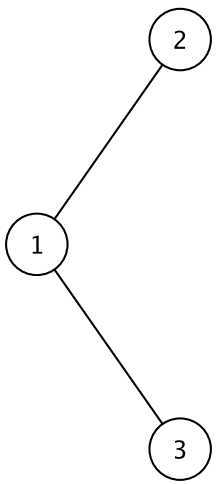
1

Estimer les paramètres pour tester l'asymétrie

- $(a, b) = (c, d)$

# Modèle BAR

Modèle auto-régressif de bifurcation
[Cowan & Staudte 1986] [Guyon 2007]



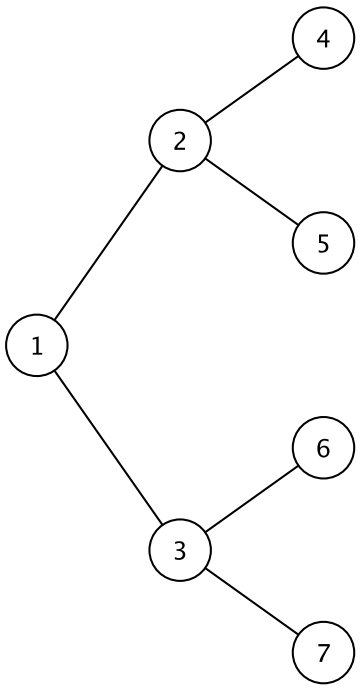

$$\begin{cases} X_{2k} &= a + bX_k + \epsilon_{2k} \\ X_{2k+1} &= c + dX_k + \epsilon_{2k+1} \end{cases}$$

Estimer les paramètres pour tester l'asymétrie

- $(a, b) = (c, d)$

# Modèle BAR

Modèle auto-régressif de bifurcation
[Cowan & Staudte 1986] [Guyon 2007]



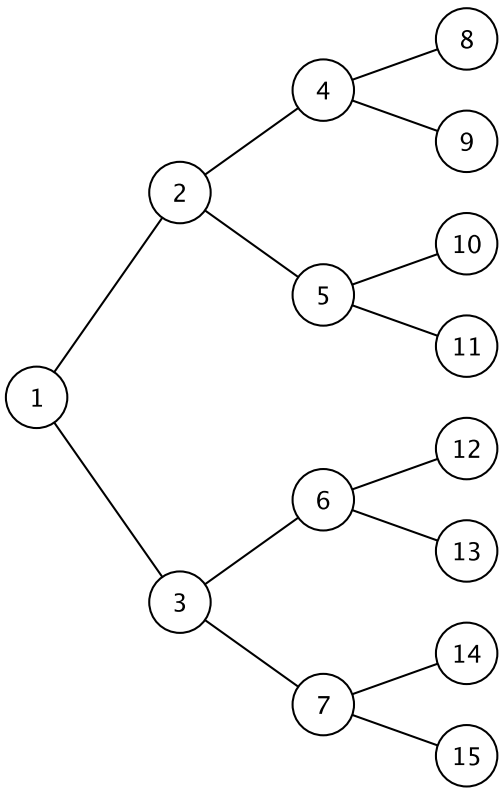

$$\begin{cases} X_{2k} &= a + bX_k + \epsilon_{2k} \\ X_{2k+1} &= c + dX_k + \epsilon_{2k+1} \end{cases}$$

Estimer les paramètres pour tester l'asymétrie

- $(a, b) = (c, d)$

# Modèle BAR

Modèle auto-régressif de bifurcation
[Cowan & Staudte 1986] [Guyon 2007]




$$\begin{cases} X_{2k} &= a + bX_k + \epsilon_{2k} \\ X_{2k+1} &= c + dX_k + \epsilon_{2k+1} \end{cases}$$

Estimer les paramètres pour tester l'asymétrie

- $(a, b) = (c, d)$

# Modélisation : données manquantes

- la probabilité d'être observé dépend du type de la fille et de celui de sa mère
- une cellule non observée n'a pas de descendance observée
- indépendance des observations des enfants de chaque cellule mère d'une même génération
- indépendance entre le processus d'observation et le processus BAR

$\Rightarrow$ Modèle de Galton-Watson à deux types

# Modèle d'observation : Galton-Watson à deux types

[Delmas & Marsalle 2010] [dS, Gégout-Petit, Marsalle 2011]

- $\delta_k = 1$ si la cellule $k$ est observée, 0 sinon
- probabilité $p^{(i)}(j_0, j_1)$ pour une cellule mère de type $i$ d'avoir $j_0$ fille de type 0 et $j_1$ fille de type 1

# Modèle d'observation : Galton-Watson à deux types

[Delmas & Marsalle 2010] [dS, Gégout-Petit, Marsalle 2011]

- $\delta_k = 1$ si la cellule $k$ est observée, 0 sinon
- probabilité $p^{(i)}(j_0, j_1)$ pour une cellule mère de type $i$ d'avoir $j_0$ fille de type 0 et $j_1$ fille de type 1

Matrice de descendance

$$P = \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix}$$

$p_{i0} = p^{(i)}(1,0) + p^{(i)}(1,1)$: nombre moyen de filles de type 0
$p_{i1} = p^{(i)}(0,1) + p^{(i)}(1,1)$: nombre moyen de filles de type 1

# Modèle d'observation : Galton-Watson à deux types

[Delmas & Marsalle 2010] [dS, Gégout-Petit, Marsalle 2011]

- $\delta_k = 1$ si la cellule $k$ est observée, 0 sinon
- probabilité $p^{(i)}(j_0, j_1)$ pour une cellule mère de type $i$ d'avoir $j_0$ fille de type 0 et $j_1$ fille de type 1

Matrice de descendance

$$P = \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix}$$

$p_{i0} = p^{(i)}(1,0) + p^{(i)}(1,1)$: nombre moyen de filles de type 0
$p_{i1} = p^{(i)}(0,1) + p^{(i)}(1,1)$: nombre moyen de filles de type 1

## Critère d'extinction

$\pi$ rayon spectral de $P$

- si $\pi \leq 1$, extinction presque sure
- si $\pi > 1$, extinction avec probabilité $< 1$

# Paramètres à estimer

## Paramètres d'intérêt

- coefficients de l'auto-régression $\theta = (a, b, c, d)$ pour tester la symétrie

# Paramètres à estimer

## Paramètres d'intérêt

- coefficients de l'auto-régression $\boldsymbol{\theta} = (a, b, c, d)$ pour tester la symétrie

## Paramètres annexes nécessaires

- variance $\sigma^2$ et covariance $\rho$ du bruit BAR, nécessaires pour construire la statistique du test de symétrie

# Paramètres à estimer

## Paramètres d'intérêt

- coefficients de l'auto-régression $\boldsymbol{\theta} = (a, b, c, d)$ pour tester la symétrie

## Paramètres annexes nécessaires

- variance $\sigma^2$ et covariance $\rho$ du bruit BAR, nécessaires pour construire la statistique du test de symétrie

## Paramètres annexes non nécessaires

- moments d'ordre supérieur du bruit BAR
- coefficients du processus d'observation

# Méthodes d'estimation

- méthode des moindres carrés pour $\boldsymbol{\theta}$ : minimiser

$$\Delta_n(\boldsymbol{\theta}) = \frac{1}{2} \sum_{k \in \mathbb{T}_{n-1}} \delta_{2k}(X_{2k} - a - bX_k)^2 + \delta_{2k+1}(X_{2k+1} - c - dX_k)^2.$$

- méthode empirique pour tous les autres paramètres

Génération $n$

$$\mathbb{G}_n = \{2^n, 2^n + 1, \ldots, 2^{n+1} - 1\}$$

Arbre jusqu'à la génération $n$

$$\mathbb{T}_n = \{1, 2, \ldots, 2^{n+1} - 1\} = \cup_{\ell=0}^{n} \mathbb{G}_\ell$$

# Méthodes d'estimation

- méthode des moindres carrés pour $\boldsymbol{\theta}$ : minimiser

$$\Delta_n(\boldsymbol{\theta}) = \frac{1}{2} \sum_{k \in \mathbb{T}_{n-1}} \delta_{2k}(X_{2k} - a - bX_k)^2 + \delta_{2k+1}(X_{2k+1} - c - dX_k)^2.$$

- méthode empirique pour tous les autres paramètres

Génération $n$ observée

$$\mathbb{G}_n^* = \{k \in \mathbb{G}_n \ ; \ \delta_k = 1\}$$

Arbre observé jusqu'à la génération $n$

$$\mathbb{T}_n^* = \{k \in \mathbb{T}_n \ ; \ \delta_k = 1\} = \cup_{\ell=0}^{n} \mathbb{G}_\ell^*$$

# Estimateur de $\theta$

[dS, Gégout-Petit, Marsalle 2011]

## Estimateur des moindres carrés pour $\theta$

$$\widehat{\theta}_n = \begin{pmatrix} \widehat{a}_n \\ \widehat{b}_n \\ \widehat{c}_n \\ \widehat{d}_n \end{pmatrix} = \boldsymbol{S}_{n-1}^{-1} \sum_{k \in \mathbb{T}_{n-1}} \begin{pmatrix} \delta_{2k} X_{2k} \\ \delta_{2k} X_k X_{2k} \\ \delta_{2k+1} X_{2k+1} \\ \delta_{2k+1} X_k X_{2k+1} \end{pmatrix}$$

avec

$$\boldsymbol{S}_n = \begin{pmatrix} \boldsymbol{S}_n^0 & 0 \\ 0 & \boldsymbol{S}_n^1 \end{pmatrix}$$

$$\boldsymbol{S}_n^0 = \sum_{k \in \mathbb{T}_n} \delta_{2k} \begin{pmatrix} 1 & X_k \\ X_k & X_k^2 \end{pmatrix} \qquad \boldsymbol{S}_n^1 = \sum_{k \in \mathbb{T}_n} \delta_{2k+1} \begin{pmatrix} 1 & X_k \\ X_k & X_k^2 \end{pmatrix}$$

# Convergence de l'estimateur

## Théorème

$$\mathbb{1}_{\{|\mathbb{G}_n^*|>0\}} \parallel \widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta} \parallel^2 = \mathbb{1}_{\{|\mathbb{G}_n^*|>0\}} \mathcal{O}\left(\frac{\log |\mathbb{T}_{n-1}^*|}{|\mathbb{T}_{n-1}^*|}\right)$$

## Théorème

Conditionnellement à la non extinction

$$\sqrt{|\mathbb{T}_{n-1}^*|}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \boldsymbol{\Omega})$$

$\boldsymbol{\Omega}$ : matrices dont on sait estimer empiriquement les coefficients

# Estimation sur un arbre

Plus grand arbre 2002-10-04-4
663 cellules jusqu'à la génération 9

| | | |
|---|---|---|
| $a$ | 0.03627 | [0.03276; 0.03979] |
| $b$ | 0.02662 | [-0.06866; 0.12191] |
| $c$ | 0.03058 | [0.02696; 0.03420] |
| $d$ | 0.17055 | [0.07247; 0.26863] |

# Statistiques de test

[dS, Gégout-Petit, Marsalle 2012]

- H0 : $(a, b) = (c, d)$
- H1 : $(a, b) \neq (c, d)$

## Statistique de test

$$T_n = (\widehat{a}_n - \widehat{c}_n, \widehat{b}_n - \widehat{d}_n)\mathbf{\Delta}_n^{-1}(\widehat{a}_n - \widehat{c}_n, \widehat{b}_n - \widehat{d}_n)'$$

avec

$$\mathbf{\Delta}_n = \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{pmatrix} \widehat{\mathbf{\Omega}}_n \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -1 & 0 \\ 0 & -1 \end{pmatrix}$$

**Théorème**

Sous H0 $T_n \xrightarrow{\mathcal{L}} \chi^2(2)$, sous H1 $\|T_n\|^2 \to +\infty$

# Test sur un arbre

Plus grand arbre 2002-10-04-4
663 cellules jusqu'à la génération 9
$\alpha = 5\%$

$$T_9 > 5.591 \Rightarrow \text{H0 rejetée}$$

# Test sur plusieurs arbres

p-valeur d'un test : plus petit niveau de risque $\alpha$ pour lequel $H0$ est rejetée

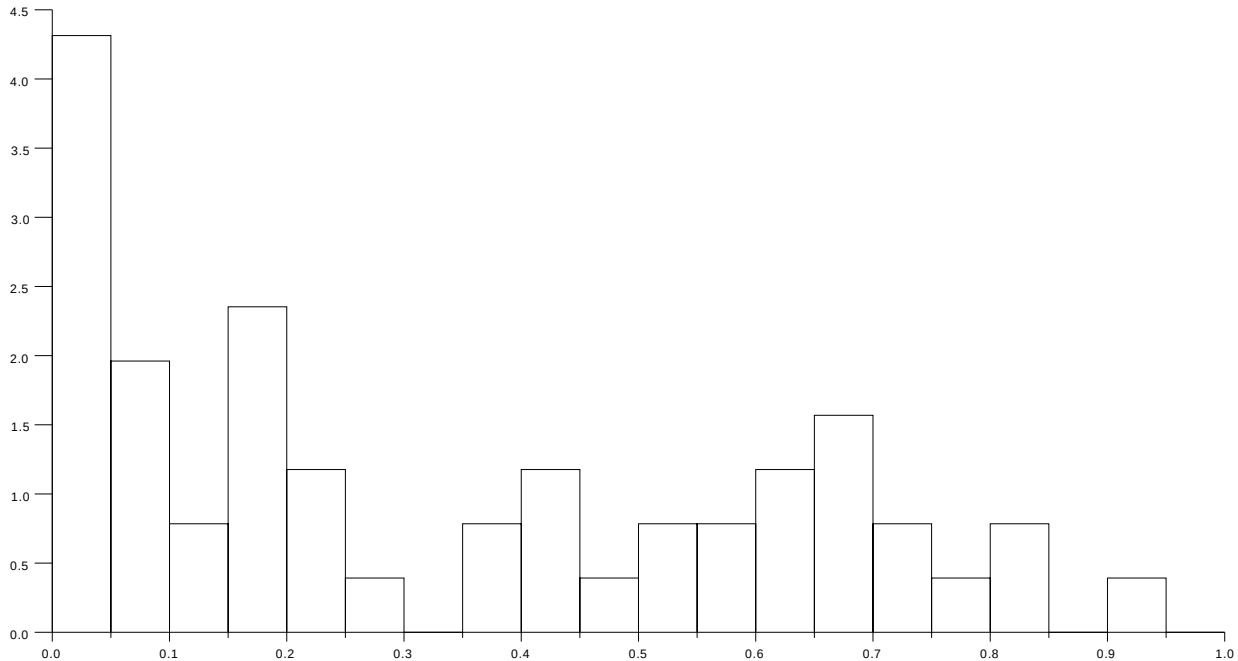
$$\text{p-valeur} < 5\% \Leftrightarrow \text{H0 est rejetée au niveau de risque 5\%}$$

Sous H0, la statistique de test suit une loi connue, on peut obtenir toutes les valeurs entre 0 et 1 comme $p$-valeur

Sous H1, on ne devrait avoir que des $p$-valeurs petites

# Test de symétrie

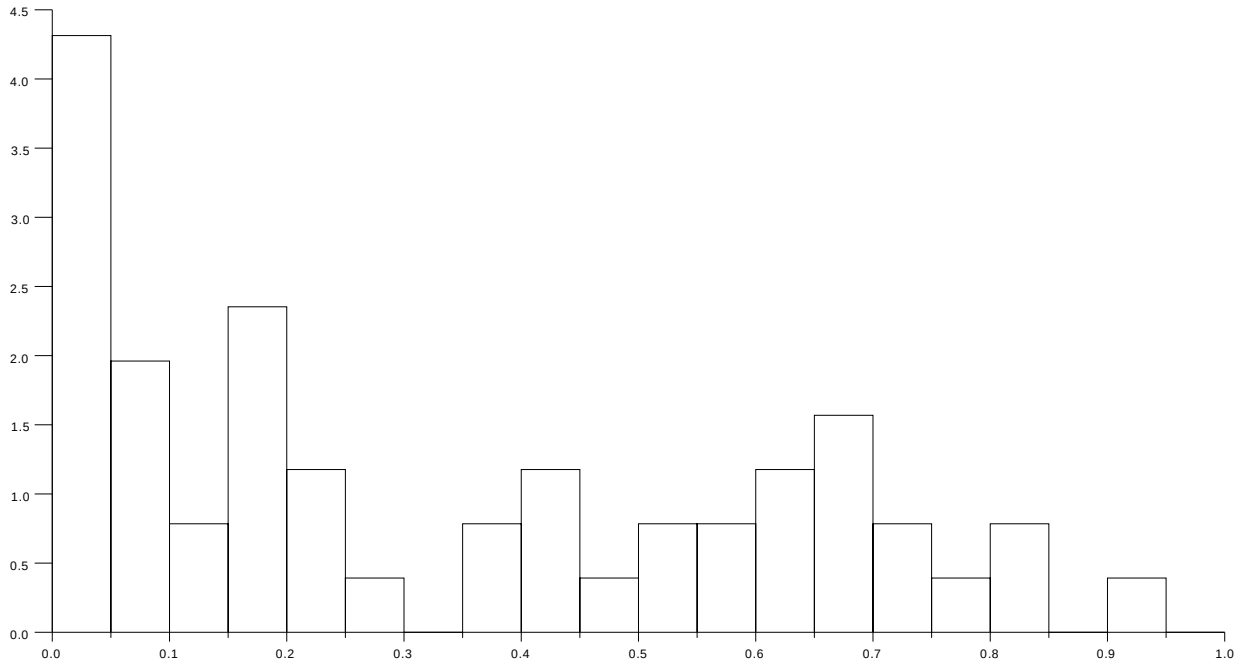
p-valeurs pour les 51 arbres comportant 8 ou 9 générations




Test $(a, b) = (c, d)$

# Test de symétrie

p-valeurs pour les 51 arbres comportant 8 ou 9 générations




Test $(a, b) = (c, d)$

On ne peut pas rejeter H0

# Etude empirique de la puissance du test

[dS, Gégout-Petit, Marsalle 2012]

| Gen | sous H0 | sous H1 |
|---|---|---|
| | $p < 0.05$ | $p < 0.05$ |
| 7 | 6.6 | 37.4 |
| 8 | 5.5 | 53.6 |
| 9 | 5.5 | 71.1 |
| 10 | 6.3 | 86.8 |
| 11 | 5.9 | 95.7 |

Proportions de p-valeurs sous le seuils 0.05 pour le test de symétrie (1000 replications) $a = b = 0.5$ (sous H1, $c = 0.5; d = 0.4$)

Faible puissance du test pour les petits nombres de générations

# Estimateur multi-arbres

[dS, Gégout-Petit, Marsalle 2014]

## Estimateur des moindres carrés pour $\theta$

$$\widehat{\theta}_n = \boldsymbol{S}_{n-1}^{-1} \sum_{j=1}^{m} \sum_{k \in \mathbb{T}_{n-1}} \begin{pmatrix} \delta_{j,2k} X_{j,2k} \\ \delta_{j,2k} X_{j,k} X_{j,2k} \\ \delta_{j,2k+1} X_{j,2k+1} \\ \delta_{j,2k+1} X_{j,k} X_{j,2k+1} \end{pmatrix}$$

avec

$$\boldsymbol{S}_n = \begin{pmatrix} \boldsymbol{S}_n^0 & 0 \\ 0 & \boldsymbol{S}_n^1 \end{pmatrix}$$

$$\boldsymbol{S}_n^i = \sum_{j=1}^{m} \sum_{k \in \mathbb{T}_n} \delta_{j,2k+i} \begin{pmatrix} 1 & X_{j,k} \\ X_{j,k} & X_{j,k}^2 \end{pmatrix}$$

# Analyse multi-arbres

Estimation de $\theta \Longrightarrow$ hypothèse $\max\{|b|, |d|\} < 1$ vraie

| | | | |
|---|---|---|---|
| $a$ | $0.0203\ [0.0197; 0.0210]$ | $c$ | $0.0195\ [0.0188; 0.0201]$ |
| $b$ | $0.4615\ [0.4437; 0.4792]$ | $d$ | $0.4782\ [0.4605; 0.4959]$ |

Estimation des moments du bruit

| | |
|---|---|
| $\sigma^2$ | $1.81 \cdot 10^{-5}\ [1.12 \cdot 10^{-5}; 2.50 \cdot 10^{-5}]$ |
| $\rho$ | $0.48 \cdot 10^{-5}\ [0.44 \cdot 10^{-5}; 0.52 \cdot 10^{-5}]$ |

Tests : hypothèse $(a, b) = (c, d)$ rejetée (p-valeur$= 10^{-5}$)

# Analyse multi-arbres des données E. coli : Galton-Watson

Estimation des lois de reproduction

| | | | |
|---|---|---|---|
| $p^{(0)}(0,0)$ | 0.35579 [0.35574; 0.35583] | $p^{(1)}(0,0)$ | 0.35611 [0.35606; 0.35616] |
| $p^{(0)}(1,0)$ | 0.03621 [0.03620; 0.03622] | $p^{(1)}(1,0)$ | 0.04707 [0.04706; 0.04708] |
| $p^{(0)}(0,1)$ | 0.04740 [0.04739; 0.04741] | $p^{(1)}(0,1)$ | 0.03755 [0.03754; 0.03756] |
| $p^{(0)}(1,1)$ | 0.56060 [0.56055; 0.56065] | $p^{(1)}(1,1)$ | 0.55928 [0.55923; 0.55933] |

Estimation de $\pi$ : $1.204\ [1.191; 1.217] \Longrightarrow$ hypothèse $\pi > 1$ vraie

Tests : hypothèse d'égalité des moyennes des deux lois non rejetée (p-valeur$= 0.9$), hypothèse d'égalité des deux vecteurs rejetée (p-valeur$= 2 \cdot 10^{-5}$)

# Plan

# Format original des données

- 448 fichiers .dat correspondant à 224 canaux, un ficheri pour les cellules vieux pôle et une pour les nouveaux pôles

`xy01c1/ch0_cell0.dat`

| index | division | length | width | area | intensity | CMx | CMy |
|---|---|---|---|---|---|---|---|
| 0 | 0 | 51.113 | 14.2234 | 727 | 168.531 | 632.084 | 333.744 |
| 1 | 0 | 54.1365 | 13.743 | 744 | 173.957 | 631.977 | 335.464 |
| 2 | 0 | 54.0628 | 14.4277 | 780 | 175.956 | 632.63 | 336.681 |
| 3 | 0 | 58.2108 | 14.0696 | 819 | 170.506 | 632.399 | 338.183 |
| 4 | 0 | 60.0982 | 14.077 | 846 | 172.301 | 631.405 | 339.792 |
| 5 | 0 | 61.0536 | 14.5282 | 887 | 173.015 | 630.493 | 340.86 |
| 6 | 0 | 64.0388 | 14.4131 | 923 | 174.315 | 630.088 | 342.173 |
| 7 | 0 | 66.1035 | 14.4924 | 958 | 175.429 | 629.842 | 343.641 |
| 8 | 0 | 68.0593 | 14.7665 | 1005 | 171.961 | 630.698 | 345.902 |
| 9 | 0 | 72.0969 | 14.5915 | 1052 | 169.44 | 630.175 | 347.913 |

- calculer les taux de croissance
- numéroter les cellules pour reconstruire la généalogie

# Calcul des taux de croissance

Régression de la forme $\texttt{length} = \exp(\tau t)$

- pas de calcul pour les cellules dont toute la vie n'est pas observée
- données aberrantes (longueurs / taux négatifs)

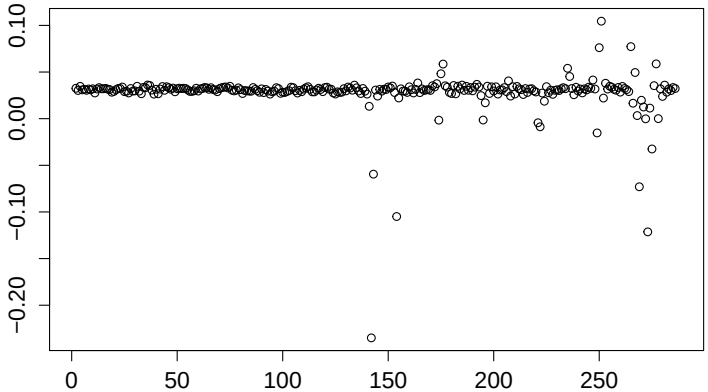
$\Rightarrow$ Eliminer les arbres

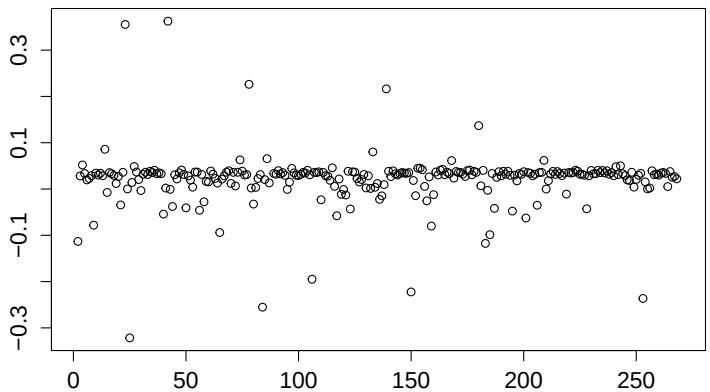
- de moins de 20 générations
- dont les taux de croissance moyens sont trop éloignés de la moyenne globale

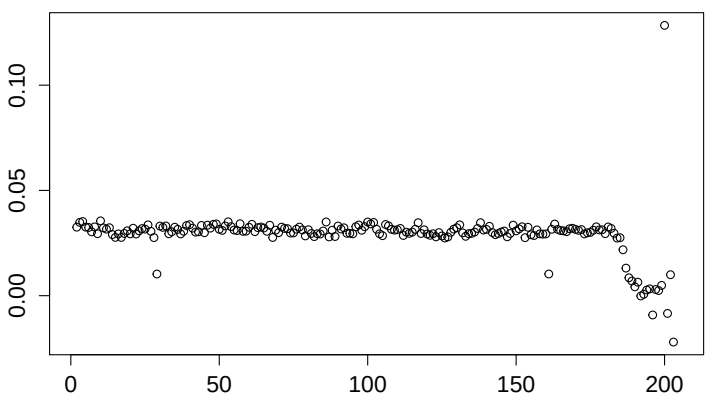
Dans les arbres restants

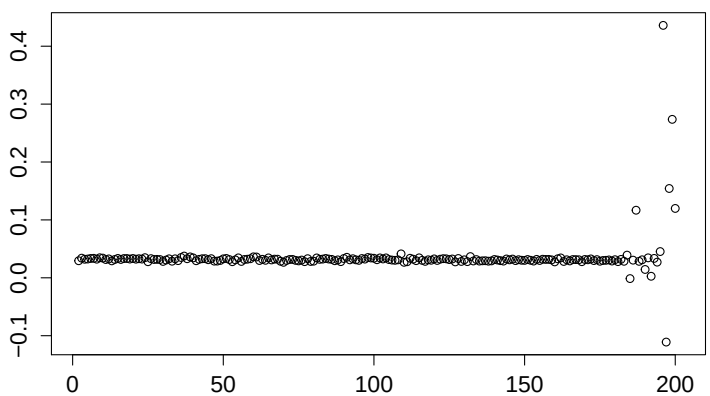
- éliminer les taux de croissances trop éloignés de la médiane

# Données aberrantes



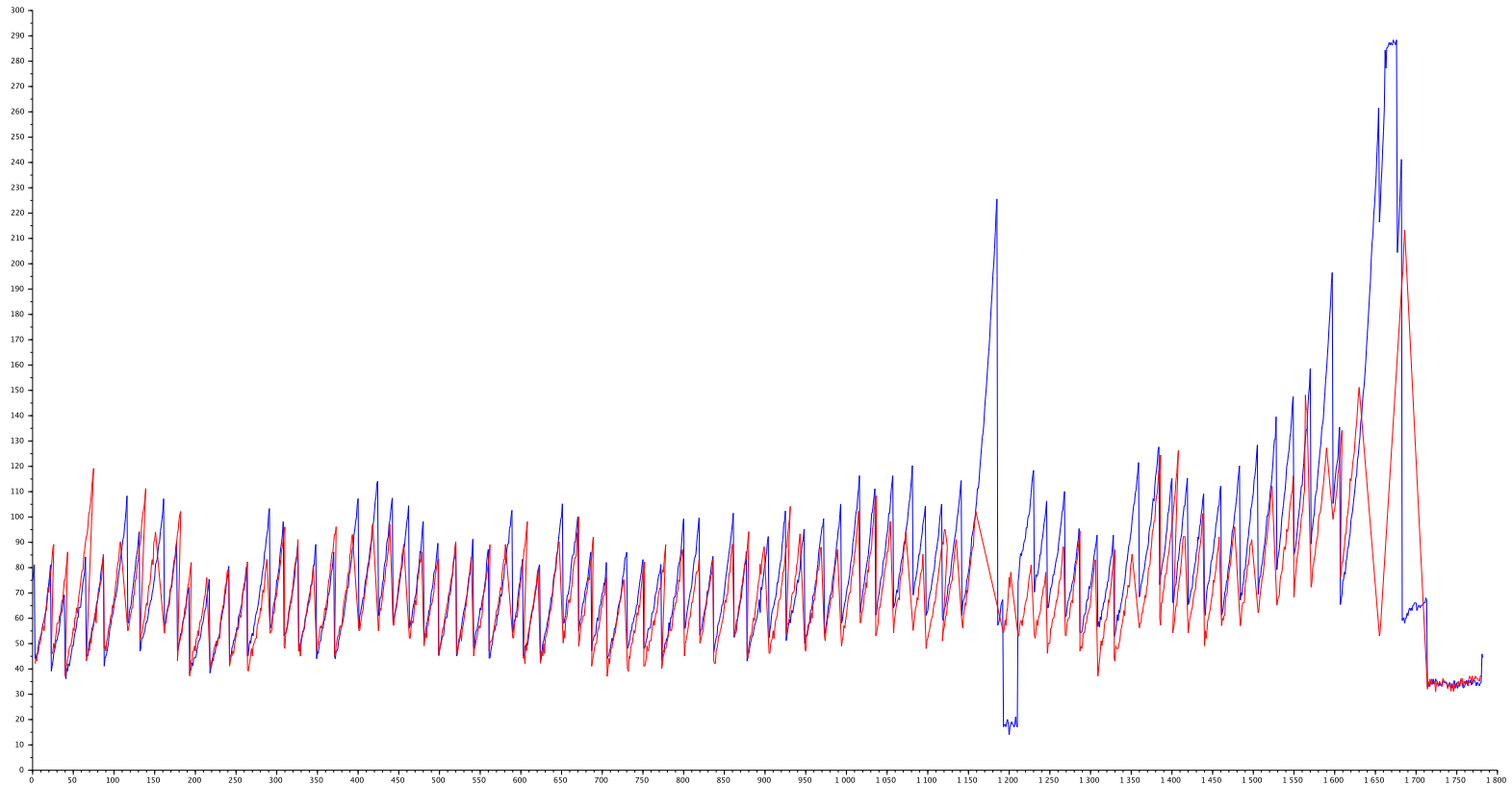



# Reconstitution des généalogies

Correspondance entre les dates de naissance des cellules



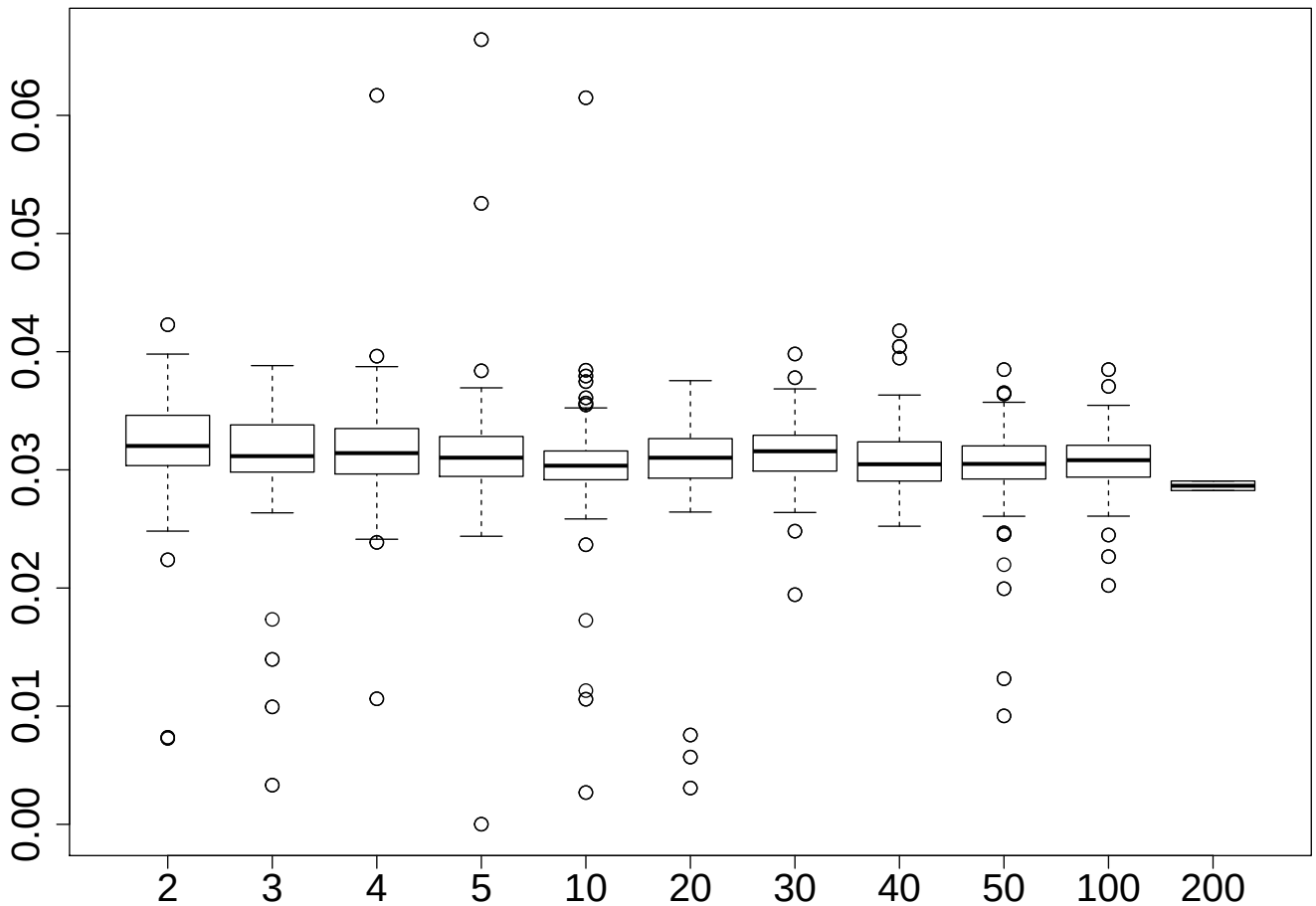
# Résumé des données

- 224 canaux
- 6 à 302 générations observées
- 45255 cellules
- quelques données manquantes, quelques données aberrantes

- calcul des taux de croissance des cellules
- types connus entre les deux filles
- lignées connues via la numérotation par génération

# Analyse exploratoire

Taux de croissance par génération, tous les arbres



# Modèle BAR

[Delyon, dS, Krell, 2016]

Estimation des paramètres

| | | |
|---|---|---|
| $a$ | 0.0304 | $[0.0200; 0.0410]$ |
| $b$ | 0.0664 | $[-0.4652; 0.5980]$ |
| $c$ | 0.0281 | $[0.0178; 0.0385]$ |
| $d$ | 0.0994 | $[-0.3194; 0.5182]$ |

- les intervalles de confiance pour b et d contiennent 0
- résultats non significatifs, données trop bruitées, impossible de valider le modèle

# Modèle de régression

[Delyon, dS, Krell, 2016]

$$X_{n+1} = \beta_m X_n + \beta_g X_{n-1} + \beta_0 + e_n$$

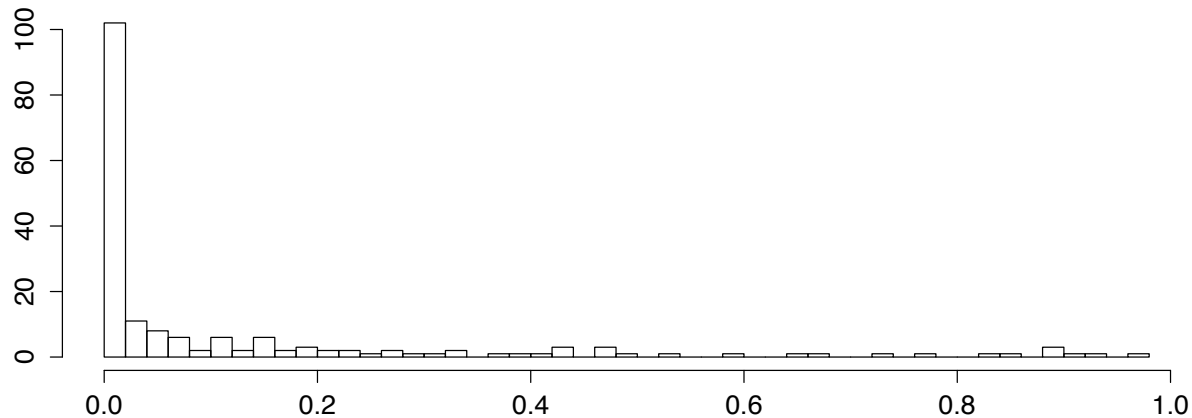
- $X_{n+1}$ taux de croissance d'une cellule de la génération $n+1$
- $X_n$ taux de croissance de sa mère
- $X_{n-1}$ taux de croissance de sa grand-mère

# Modèle de régression

[Delyon, dS, Krell, 2016]

$$X_{n+1} = \beta_m X_n + \beta_g X_{n-1} + \beta_0 + e_n$$

Histogramme des p-valeurs pour le test de significativité de la mère

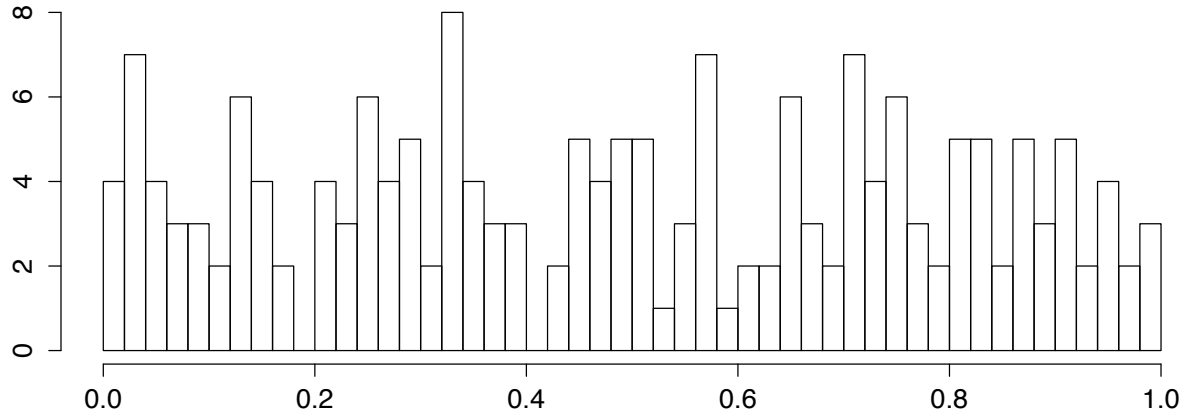


# Modèle de régression

[Delyon, dS, Krell, 2016]

$$X_{n+1} = \beta_m X_n + \beta_g X_{n-1} + \beta_0 + e_n$$

Histogramme des p-valeurs pour le test de significativité de la grand-mère

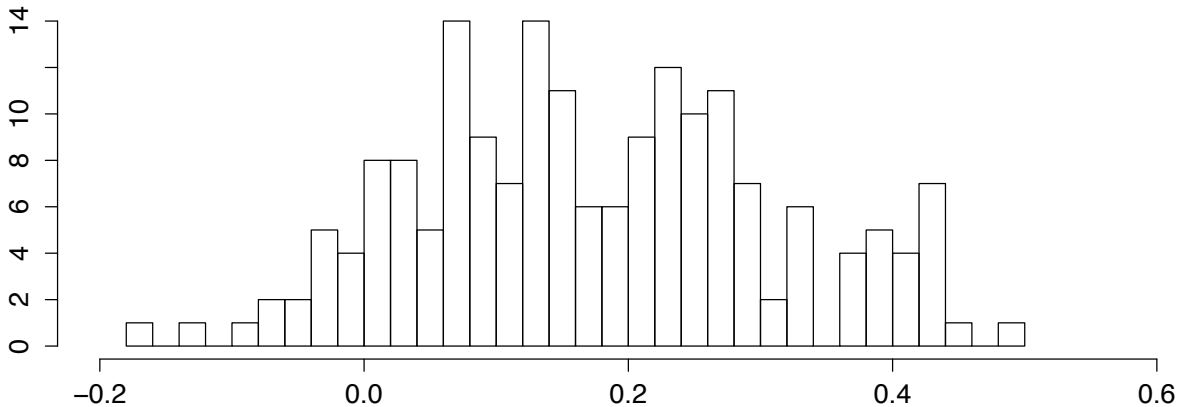


# Modèle de régression

[Delyon, dS, Krell, Robert 2016]

$$X_{n+1} = \beta_m X_n + \beta_0 + e_n$$

Histogramme des $\beta_m$ pour les cellules vieux pôle



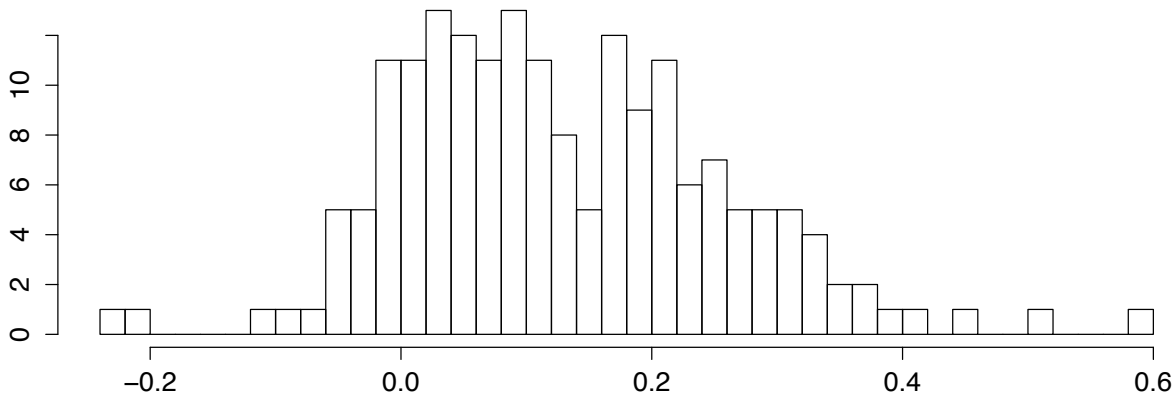

Différence significative des taux de croissance moyens, des coefficients de corrélation mère-fille.

# Modèle de régression

[Delyon, dS, Krell, Robert 2016]

$$X_{n+1} = \beta_m X_n + \beta_0 + e_n$$

Histogramme des $\beta_m$ pour les cellules nouveau pôle




Différence significative des taux de croissance moyens, des coefficients de corrélation mère-fille.

# Plan de l'exposé

# Problèmes rencontrés

- structure complexe des données en arbre
- données manquantes
- données aberrantes
- données trop bruitées

# Réponses

- Les deux expériences sont-elles contradictoires ?
- La division d'E. coli est-elle asymétrique ?

# Réponses

- Les deux expériences sont-elles contradictoires ?
- La division d'E. coli est-elle asymétrique ?

- La taux de croissance d'une cellule fille dépend de celui de sa mère et pas de sa grand-mère et ancêtres d'ordre plus élévé
- Les deux jeux de données font apparaître une asymétrie significative entre les cellules nouveau / vieux pôle
- Les données de [Stewart & al. 2005] sont en régime transitoire
- Les données de [Wang & al. 2012] sont en régime stationnaire

# Perspectives

- Encore peu de données disponibles à l'échelle de la cellule
- Concevoir de nouveaux modèles pour expliquer l'origine du bruit : mémoire ?
  - prendre en compte la dynamique individuelle
  - prendre en compte le temps continu
  - prendre en compte la fabrication de protéines marquées, les marqueurs de fluorescence
- Concevoir de nouveaux outils statistiques
  - utilisables sur des données de population
  - permettant de discriminer les modèles

# Rérérences


[Cowan & Staudte 1986] Cowan and Staudte The bifurcating autoregressive model in cell lineage studies. *Biometrics* (1986).

[Stewart & al. 2005] Stewart, Madden, Paul, and Taddei Aging and death in an organism that reproduces by morphologically symmetric division. *PLoS Biol.* (2005)

[Guyon 2007] Guyon Limit theorems for bifurcating Markov chains. Application to the detection of cellular aging. *Ann. Appl. Probab.* (2007)

[Bercu, dS, Gégout-Petit 2009] Bercu, de Saporta and Gégout-Petit Asymptotic analysis for bifurcating autoregressive processes via a martingale approach. *Electron. J. Probab.* (2009)

[Delmas & Marsalle 2010] Delmas and Marsalle Detection of cellular aging in a Galton-Watson process. *Stoch. Process. and Appl.* (2010)

[dS, Gégout-Petit, Marsalle 2011] de Saporta, Gégout-Petit and Marsalle Parameters estimation for asymmetric bifurcating autoregressive processes with missing data. *Electron. J. Statist.* (2011)

[Wang & al. 2012] Wang, Robertn Pelletier, Dang, Taggei, Wright and Jun Robust growth of escherichia coli. *Current Biology* (2012)

[dS, Gégout-Petit, Marsalle 2012] de Saporta, Gégout-Petit and Marsalle Symmetry tests for bifurcating autoregressive processes with missing data. *Statistics & Probability Letters* (2012)

[dS, Gégout-Petit, Marsalle 2014] de Saporta, Gégout-Petit and Marsalle Random coefficients bifurcating autoregressive processes. *ESAIM proba. stat.* (2014)

[dS, Gégout-Petit, Marsalle 2014] de Saporta, Gégout-Petit and Marsalle Statistical study of asymmetry in cell lineage data. *Comp. Stat. Data Anal.* (2014)

[Delyon, dS, Krell, Robert 2016] Delyon, de Saporta, Krell and Robert Investigation of asymmetry in E. coli growth rate *CSBIGS* (to appear)