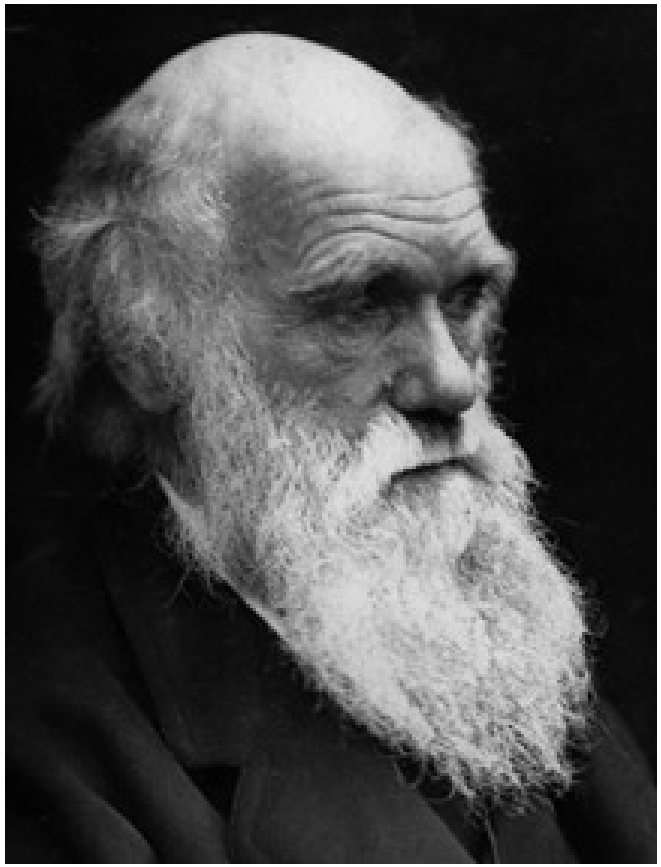
Modélisation mathématique de l'évolution de la diversité génétique

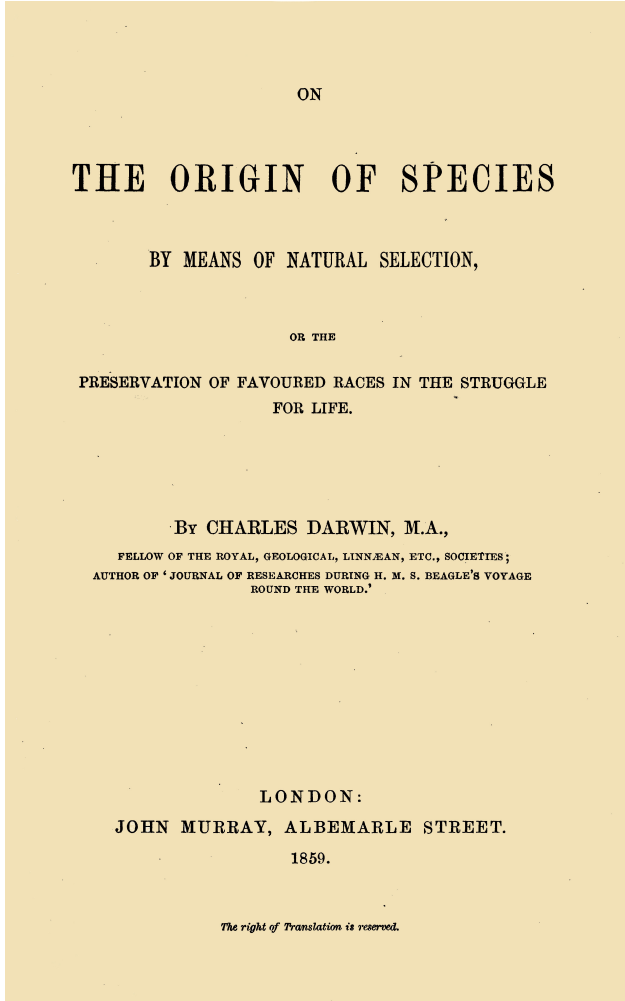
Raphaël Forien

INRA - BioSP - Avignon

Master class - CIRM 2019

Un peu d'histoire...





ON

THE ORIGIN OF SPECIES

BY MEANS OF NATURAL SELECTION,

OR THE

PRESERVATION OF FAVOURED RACES IN THE STRUGGLE FOR LIFE.

By CHARLES DARWIN, M.A.,

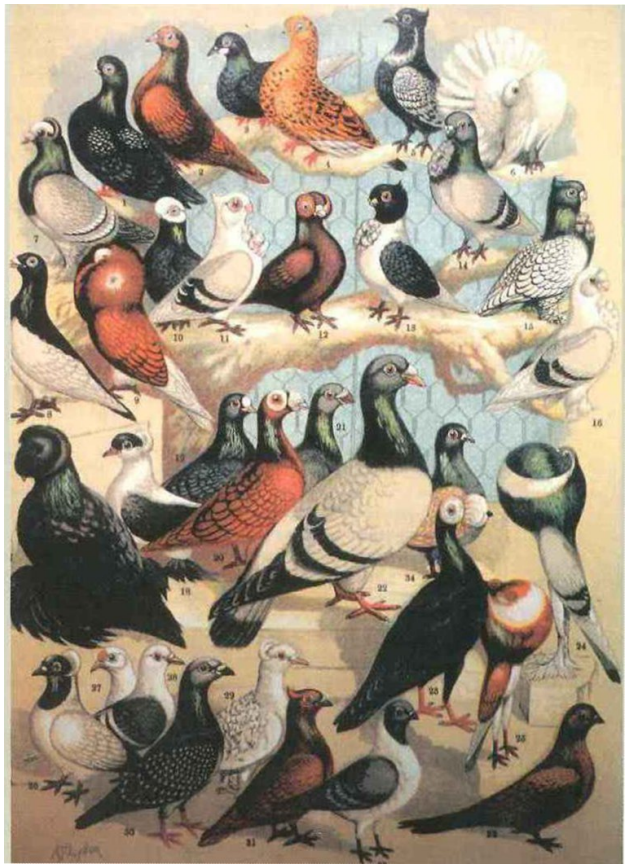
FELLOW OF THE ROYAL, GEOLOGICAL, LINNÆAN, ETC., SOCIETIES;
AUTHOR OF 'JOURNAL OF RESEARCHES DURING H. M. S. BEAGLE'S VOYAGE ROUND THE WORLD.'

LONDON:
JOHN MURRAY, ALBEMARLE STREET.
1859.

The right of Translation is reserved.

Figure – Charles Darwin publie *On the origin of species by means of natural selection* en 1859, où il expose pour la première fois sa théorie de l'évolution.

La sélection naturelle



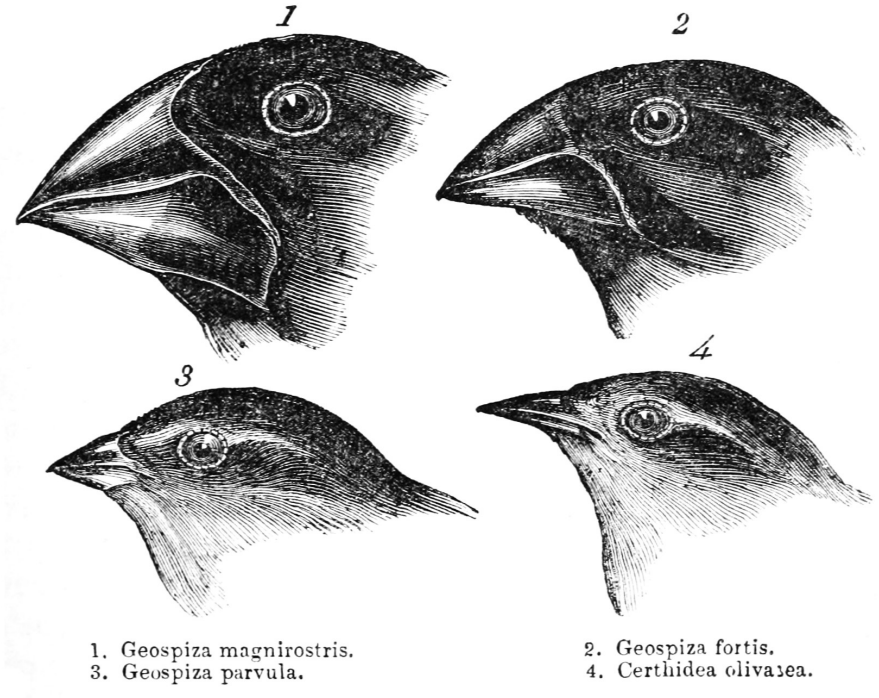


Figure – Sélection artificielle et sélection naturelle

La théorie de l'évolution

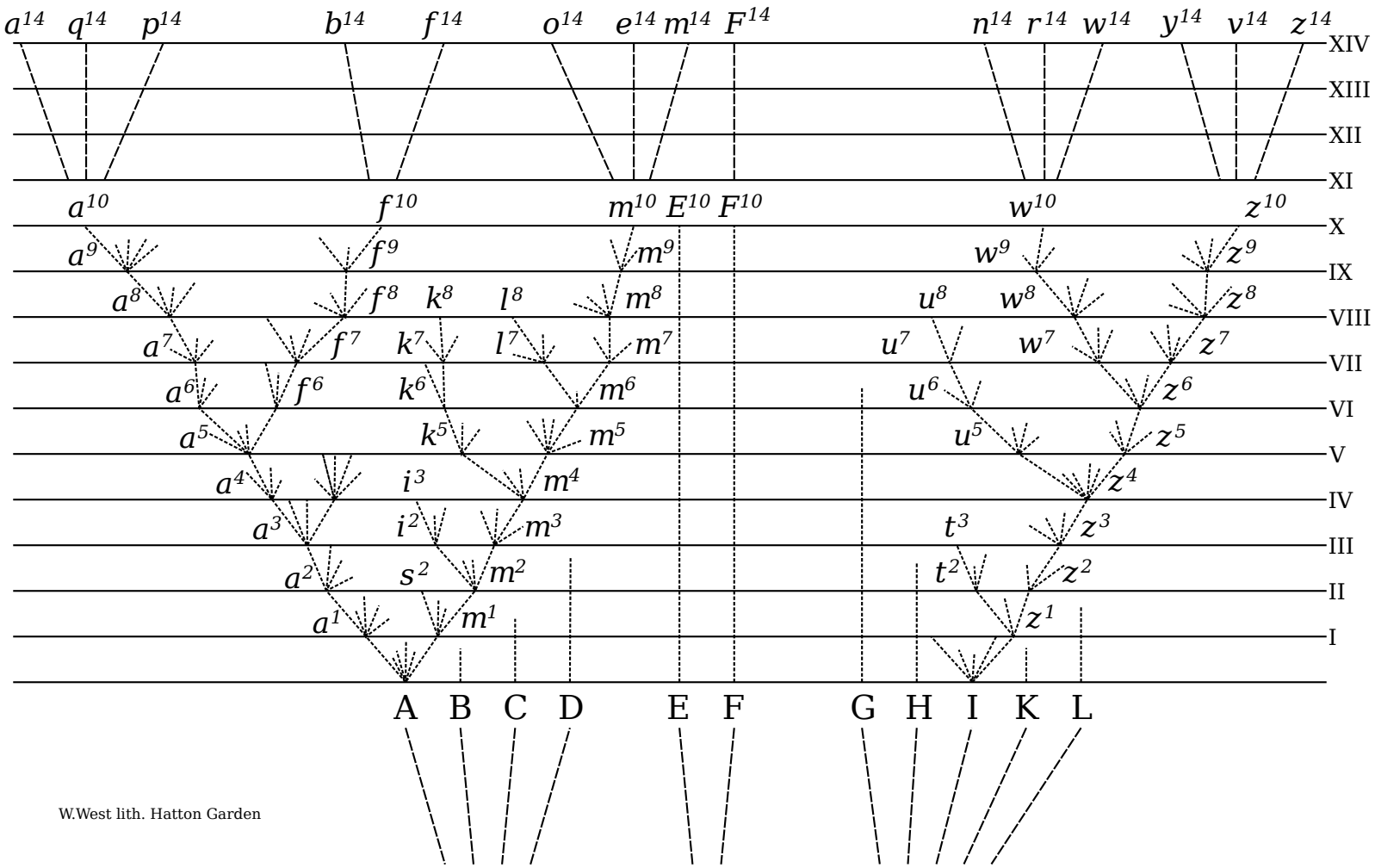


Figure – La diversification progressive associée à la sélection naturelle est à l'origine de la diversité du vivant. Cette théorie repose sur l'héritabilité des caractéristiques des êtres vivants, accompagnée d'une dose de variabilité.

Les lois de l'héritabilité

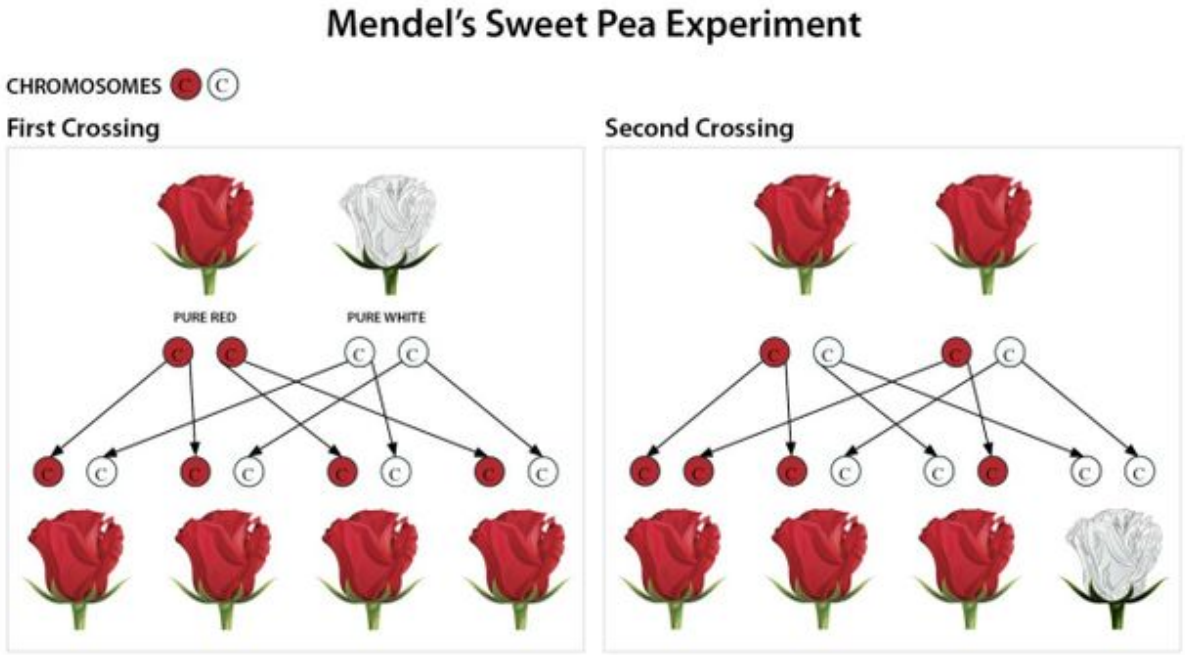


Figure – Gregor Mendel publie les résultats de ses expériences sur les croisements de plantes en 1866. Ils seront largement ignorés par la communauté scientifique jusque dans les années 1900. Le mot *gène* est inventé en 1909 par W. Johannsen pour désigner l'unité opérationnelle qui obéit aux lois de Mendel.

La synthèse moderne

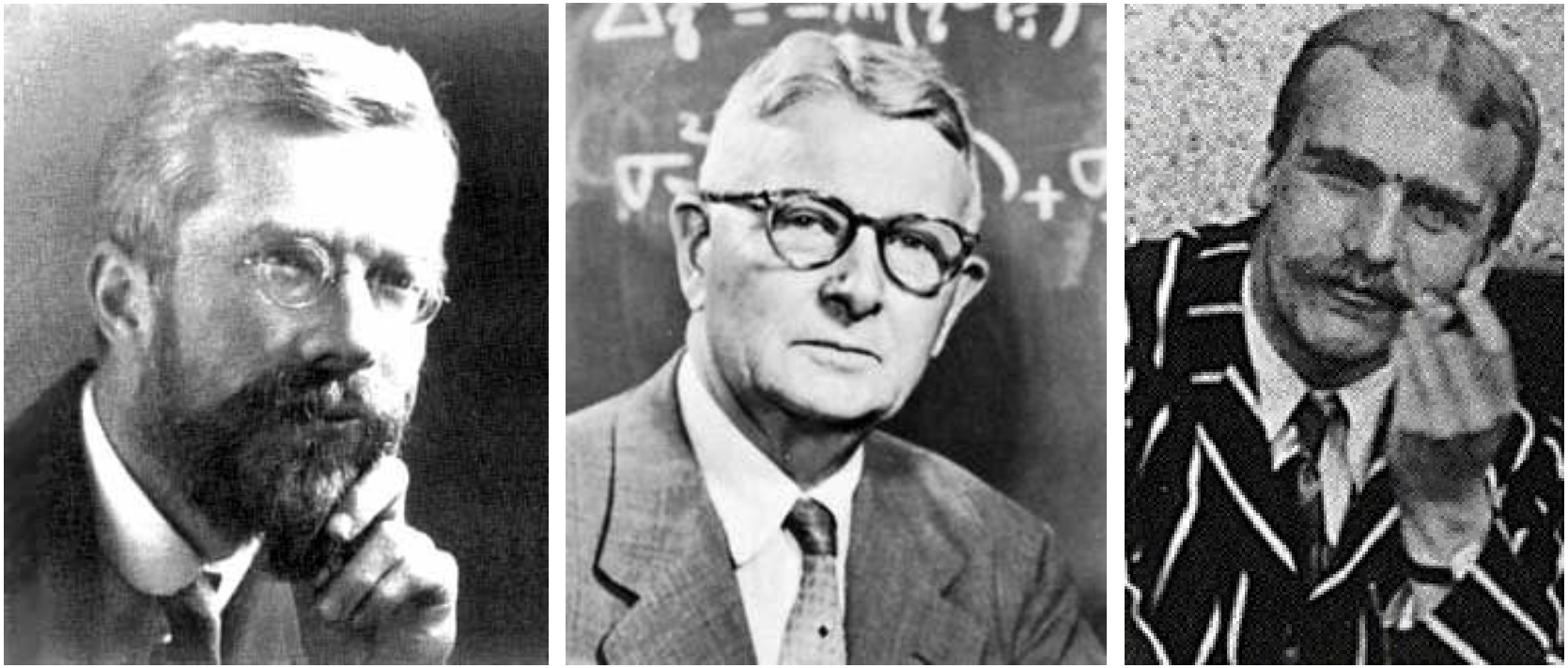


Figure – Dans les années 1920, Ronald Fisher, Sewall Wright et John B. S. Haldane mettent au point les premiers modèles mathématiques de la génétique des populations et permettent de réconcilier les travaux de Mendel avec la théorie de l'évolution de Darwin.

La séquence génétique

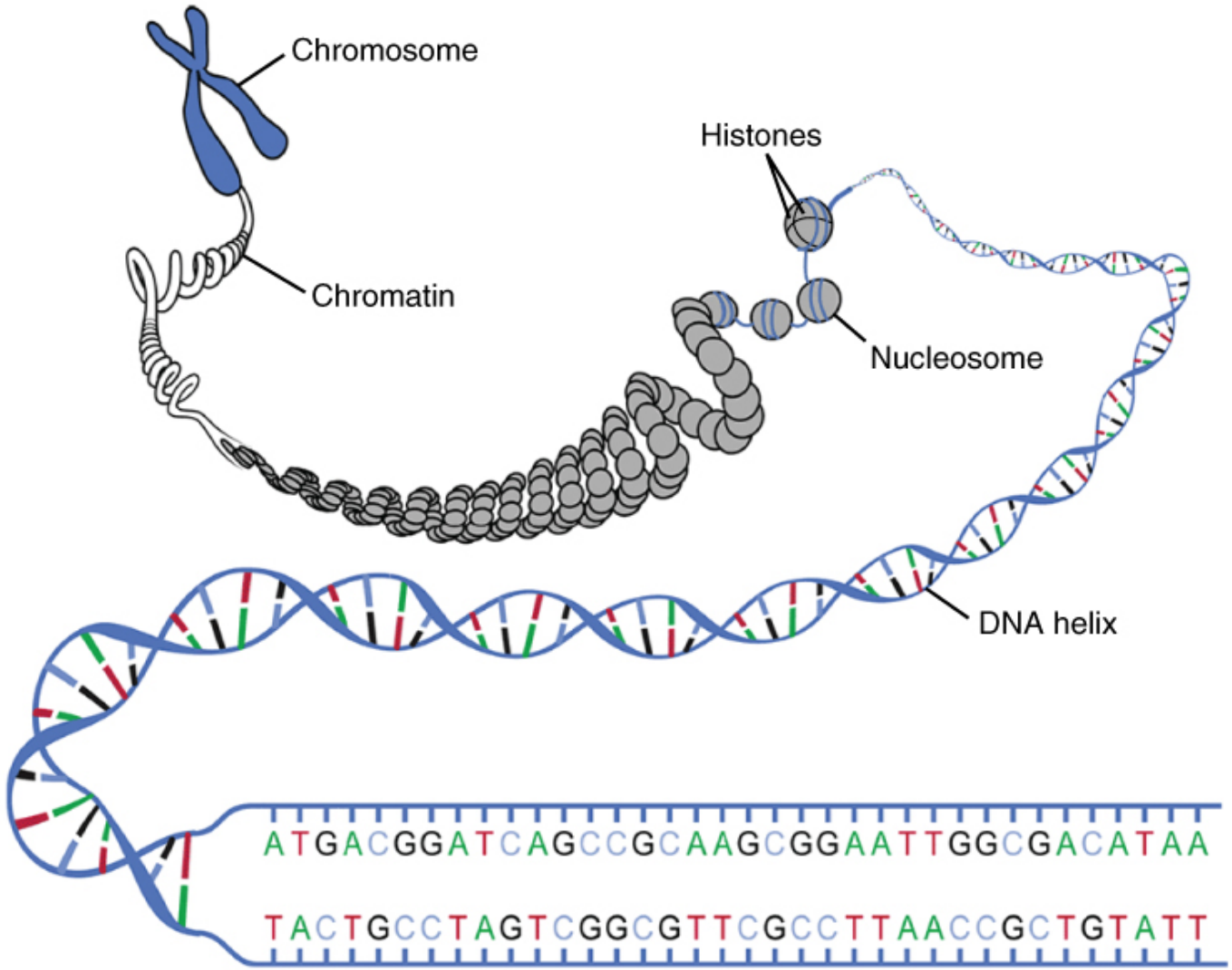


Figure – Les chromosomes sont le support de la séquence génétique

OpenStax via Wikimedia Commons CC-BY-4.0

Le support des gènes

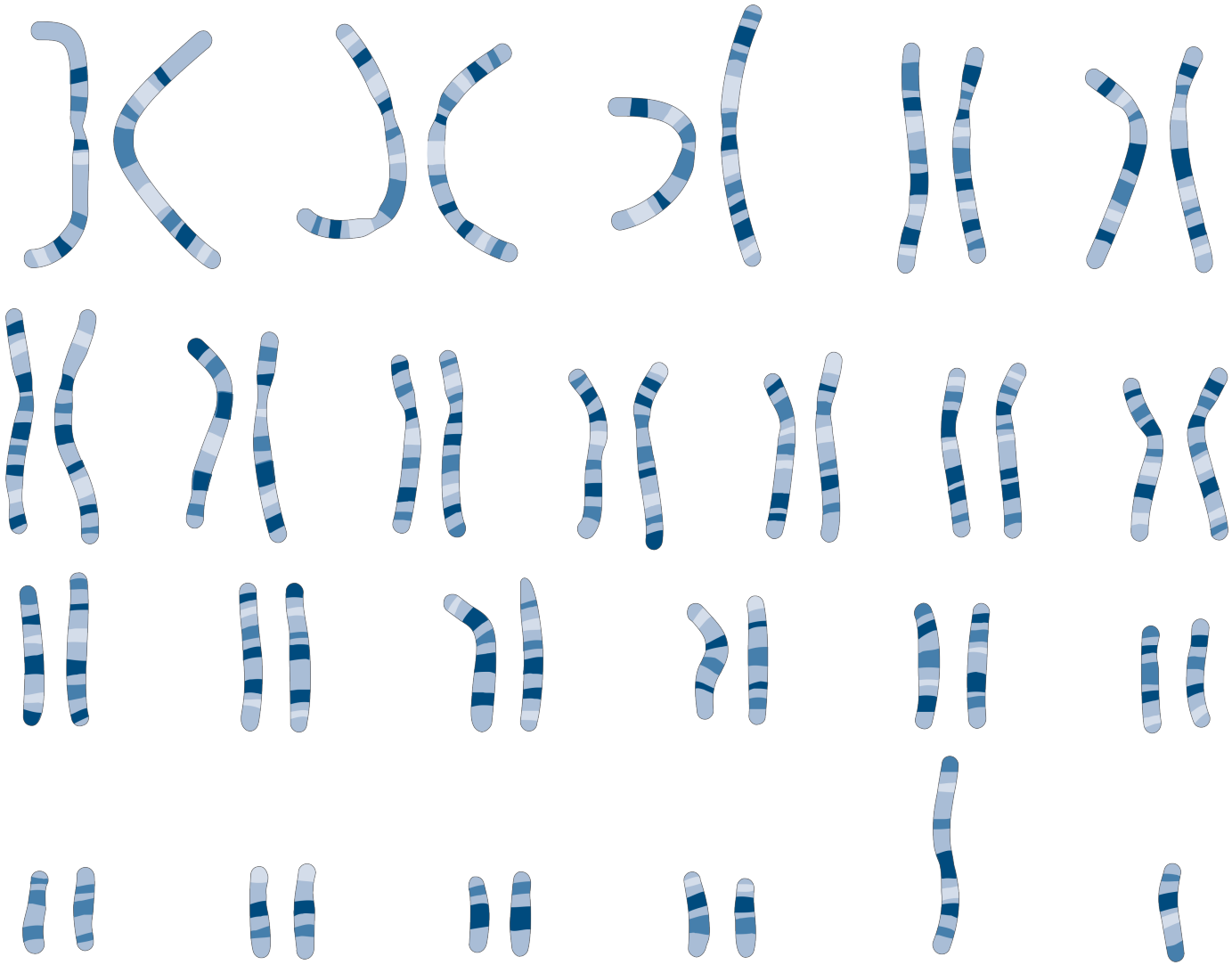


Figure – Caryotype humain

Serviermedicalart CC-BY-2.0

La Méiose

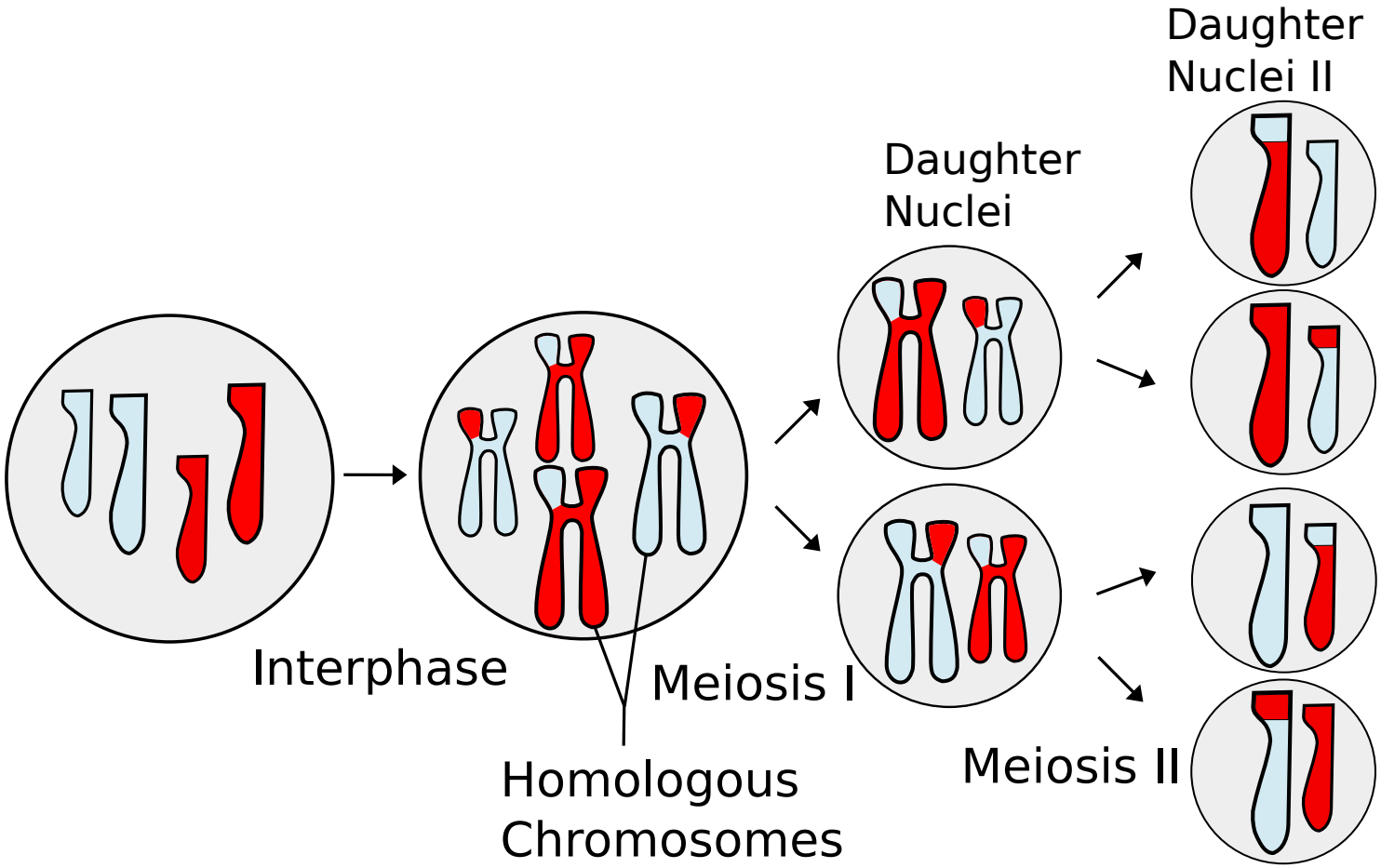


Figure – Les chromosomes sont répartis au hasard lors de la production des gamètes.

Transmission des chromosomes

Figure – Chaque individu reçoit un chromosome de chaque parent. Chaque chromosome est le résultat d'une éventuelle recombinaison entre les deux chromosomes du parent en question.

adapté de Gklambauer [CC-BY-SA 3.0] via Wikimedia Commons

1. Premiers modèles

Le modèle de Wright-Fisher

N individus se reproduisent aléatoirement. Chaque individu de la génération $n+1$ choisit son parent uniformément au hasard dans les individus de la génération n.

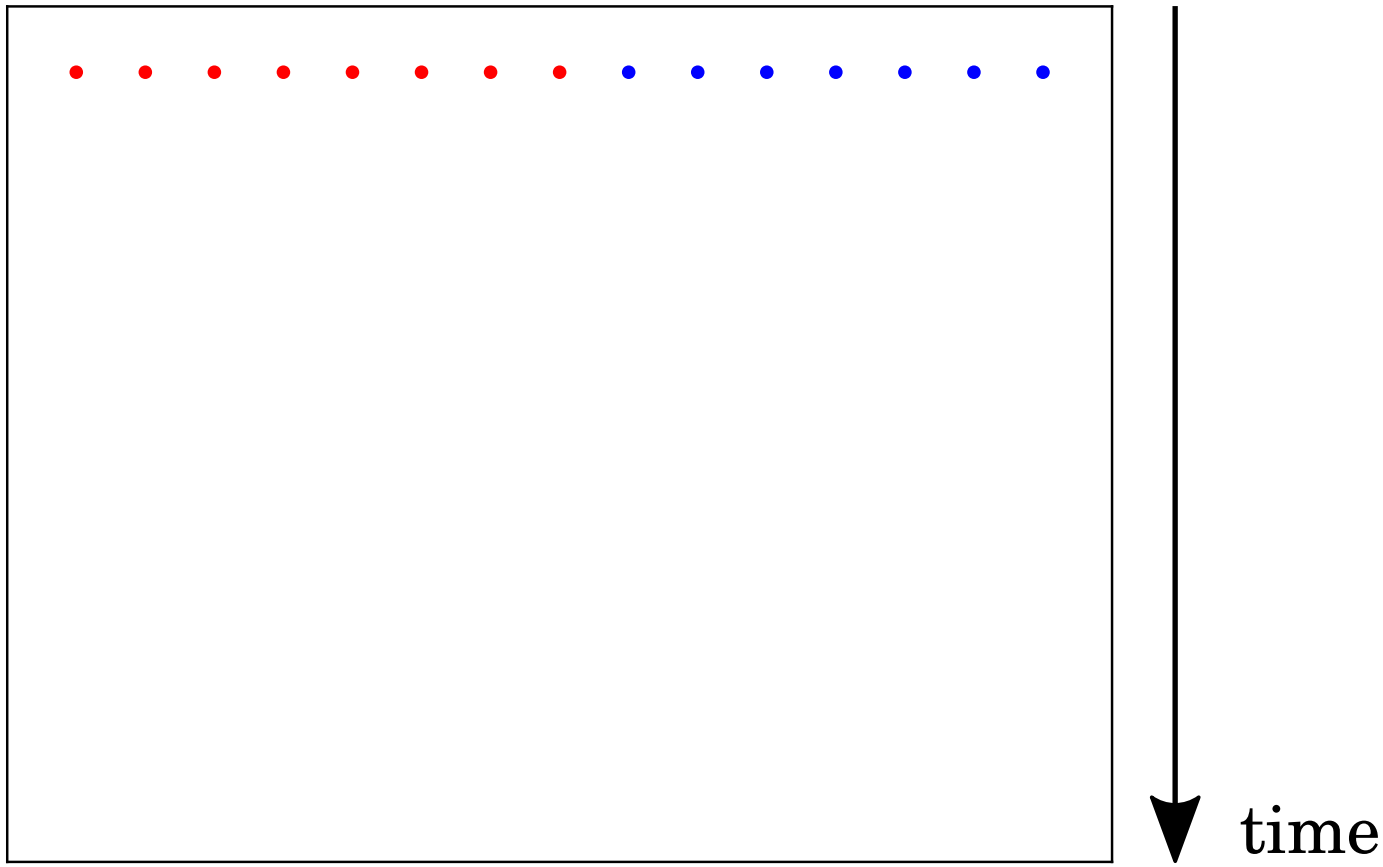


Figure – modèle de Wright-Fisher à $N = 15$ individus

Le modèle de Wright-Fisher

N individus se reproduisent aléatoirement. Chaque individu de la génération $n+1$ choisit son parent uniformément au hasard dans les individus de la génération n.

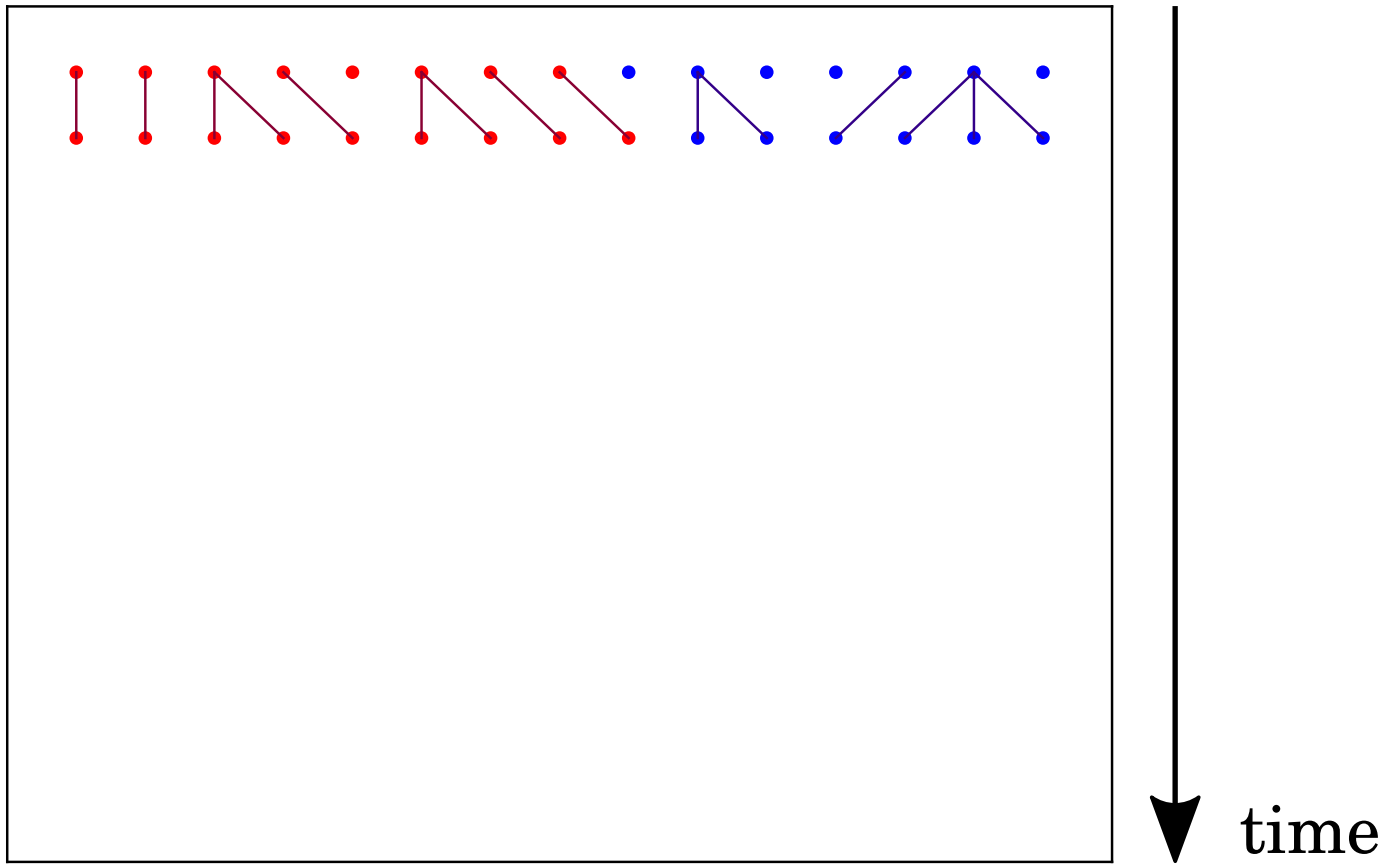


Figure – modèle de Wright-Fisher à $N = 15$ individus

Le modèle de Wright-Fisher

N individus se reproduisent aléatoirement. Chaque individu de la génération $n+1$ choisit son parent uniformément au hasard dans les individus de la génération n.

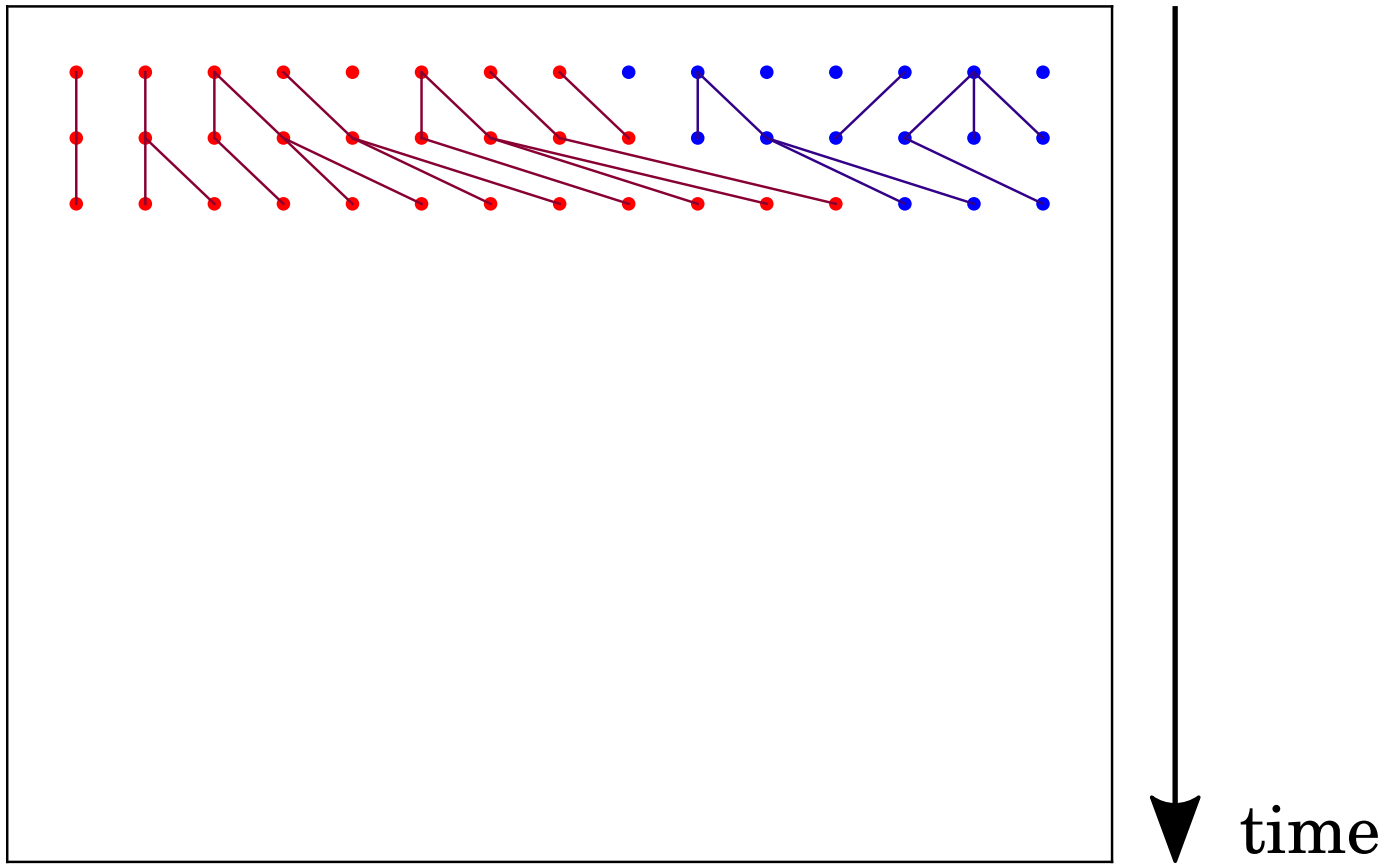


Figure – modèle de Wright-Fisher à $N = 15$ individus

Le modèle de Wright-Fisher

N individus se reproduisent aléatoirement. Chaque individu de la génération $n+1$ choisit son parent uniformément au hasard dans les individus de la génération n.

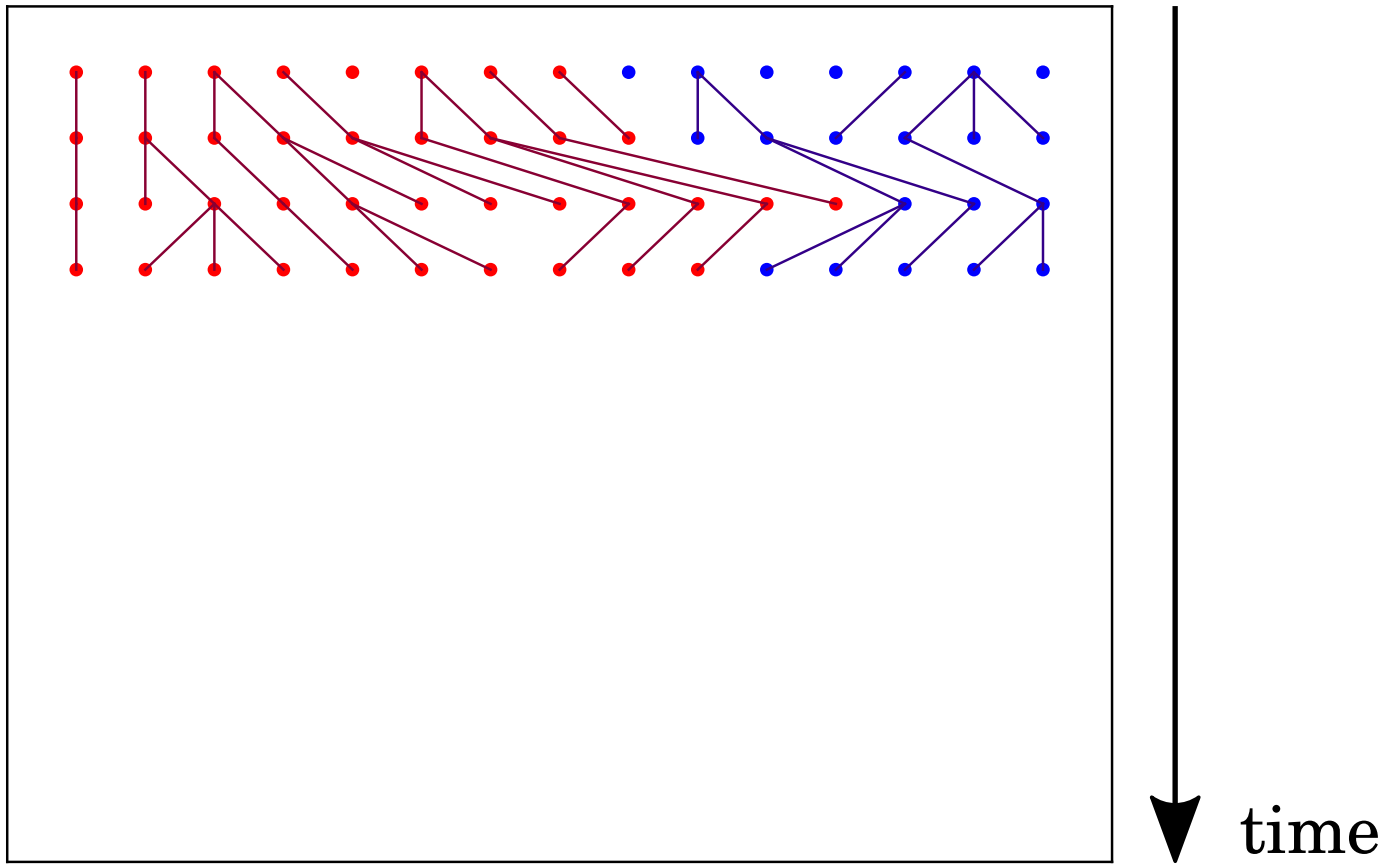


Figure – modèle de Wright-Fisher à $N = 15$ individus

Le modèle de Wright-Fisher

N individus se reproduisent aléatoirement. Chaque individu de la génération $n+1$ choisit son parent uniformément au hasard dans les individus de la génération n.

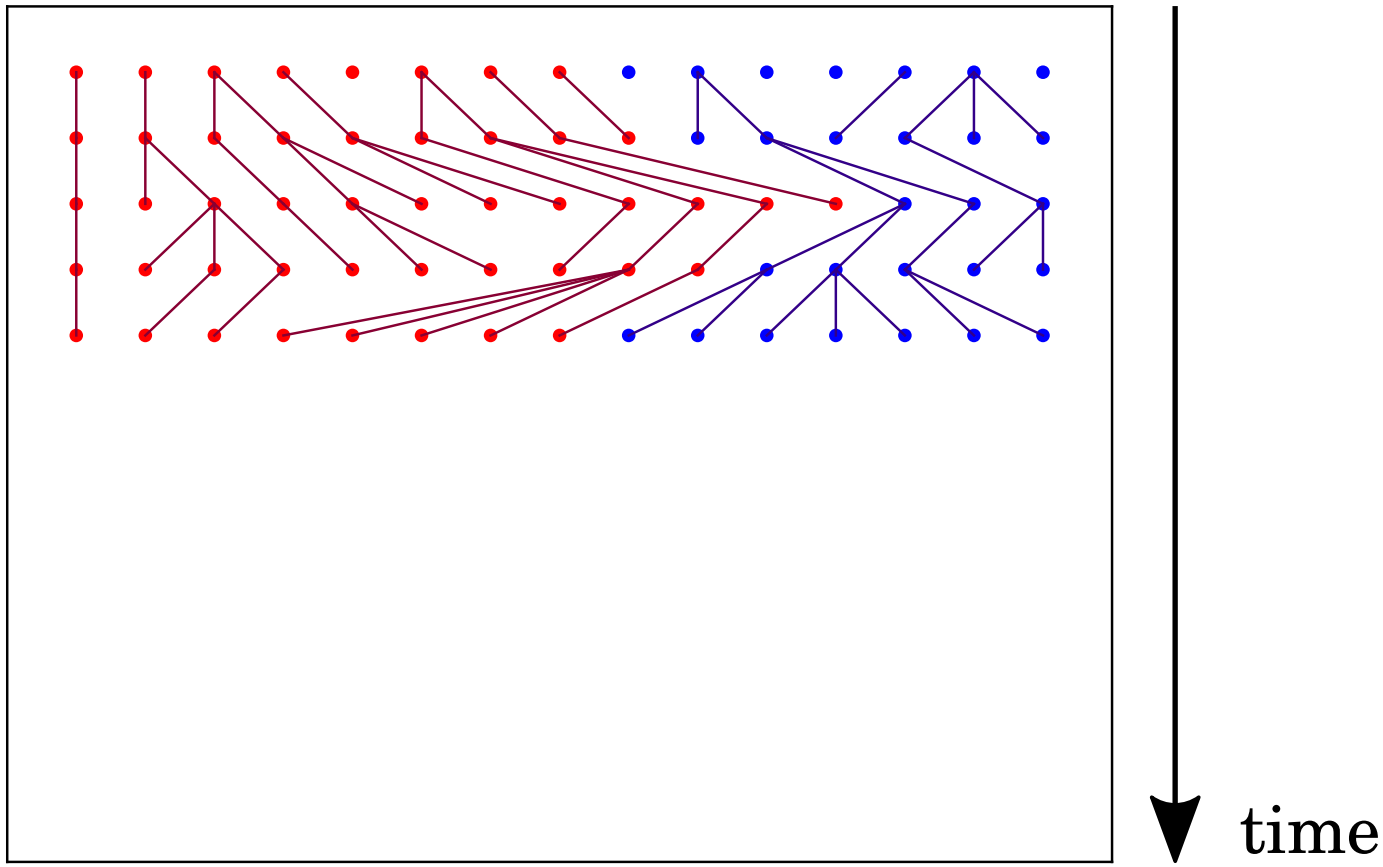


Figure – modèle de Wright-Fisher à $N = 15$ individus

Le modèle de Wright-Fisher

N individus se reproduisent aléatoirement. Chaque individu de la génération $n+1$ choisit son parent uniformément au hasard dans les individus de la génération n.

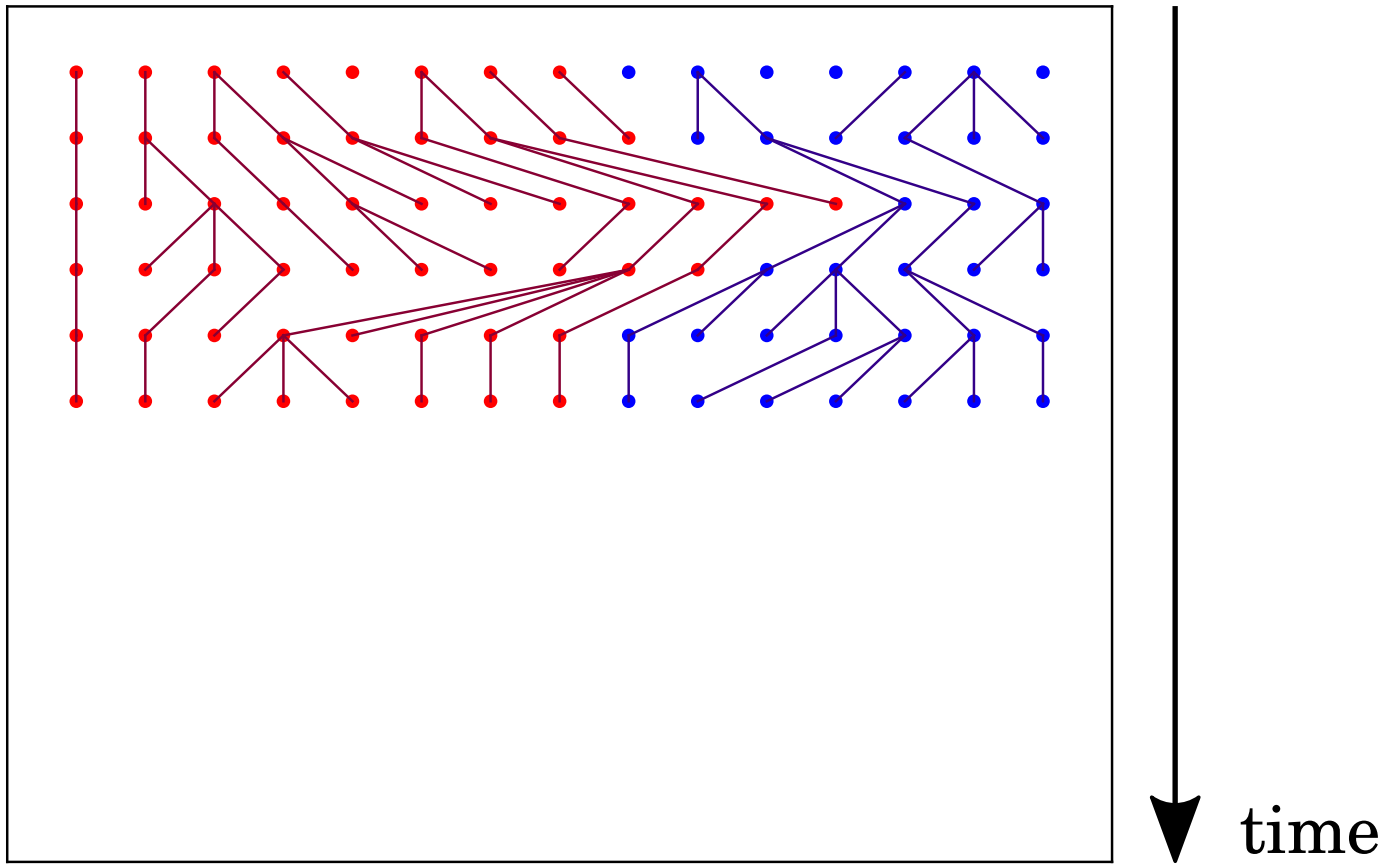


Figure – modèle de Wright-Fisher à $N = 15$ individus

Le modèle de Wright-Fisher

N individus se reproduisent aléatoirement. Chaque individu de la génération $n+1$ choisit son parent uniformément au hasard dans les individus de la génération n.

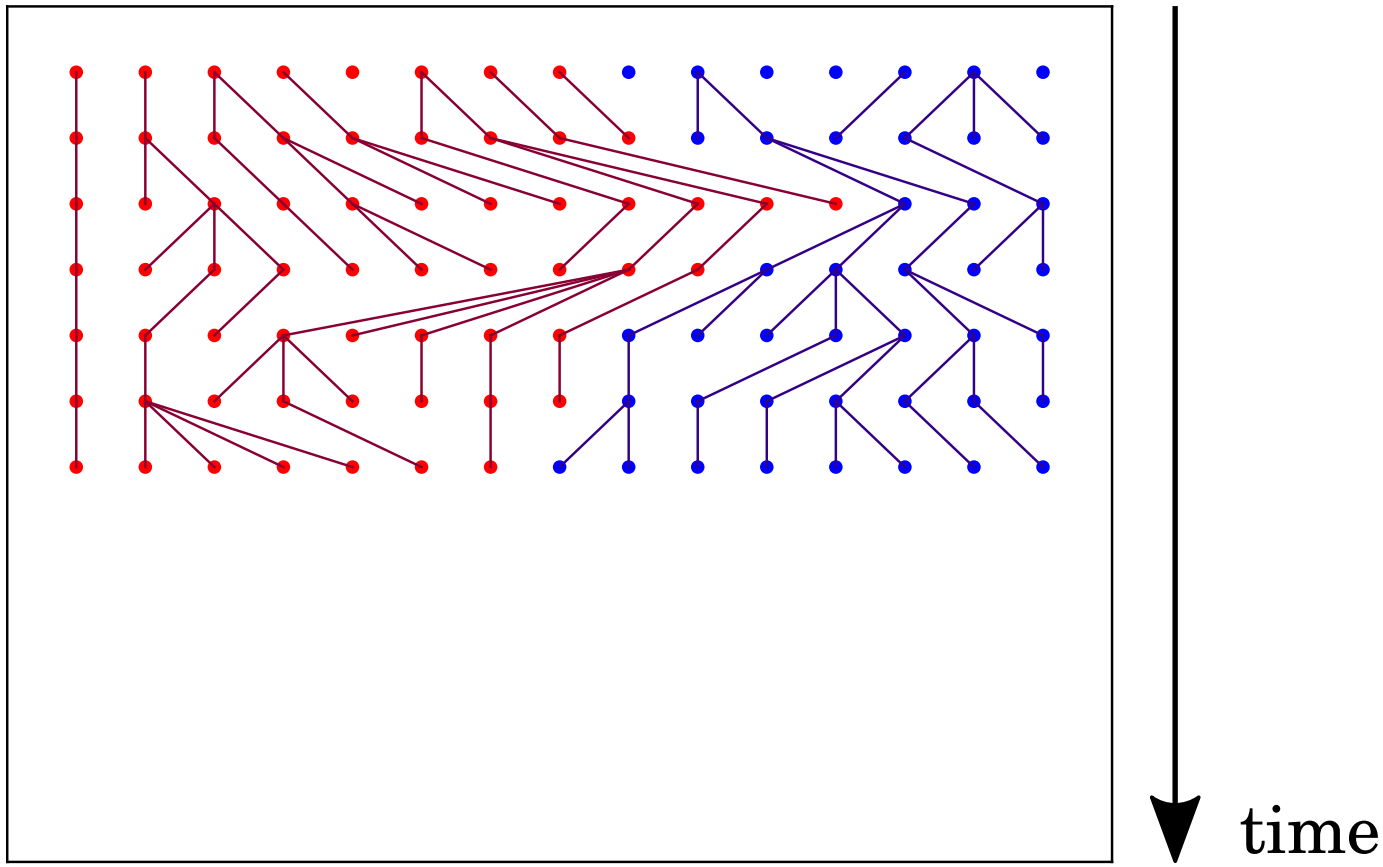


Figure – modèle de Wright-Fisher à $N = 15$ individus

Le modèle de Wright-Fisher

N individus se reproduisent aléatoirement. Chaque individu de la génération $n+1$ choisit son parent uniformément au hasard dans les individus de la génération n.

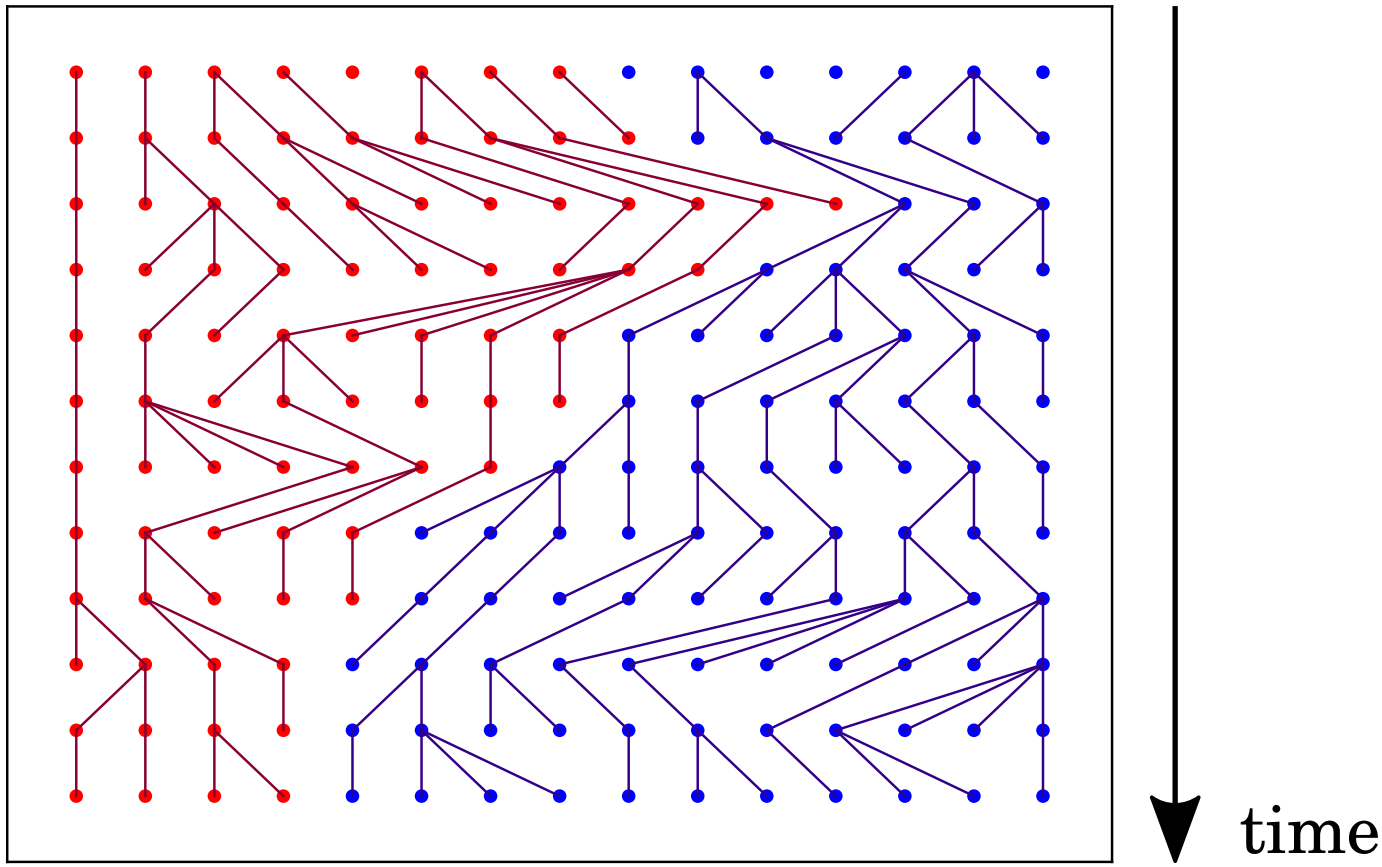


Figure – modèle de Wright-Fisher à $N = 15$ individus

Le modèle de Wright-Fisher

On note $X(t)$ le nombre d'individus portant le gène rouge à la génération t. Alors, conditionnellement à $X(t)$,

$$X(t+1) \sim Bin\left(N, \frac{X(t)}{N}\right).$$

C'est équivalent à

$$\mathbb{P}\left(X(t+1) = k \mid X(t) = x\right) = \binom{n}{k} \left(\frac{x}{N}\right)^k \left(1 - \frac{x}{N}\right)^{N-k}.$$

On remarque que dès que $X(t) = 0$ ou $X(t) = N$, alors $X(t+1) = X(t+2) = \ldots = X(t)$. On dit que 0 et N sont des états absorbants.
On peut noter T_{fix} la première génération pour laquelle $X(t) \in \{0, N\}$. On appelle T_{fix} le temps de fixation.

Le point de vue généalogique

On peut reconstruire la composition de la population à partir de ses ancêtres.

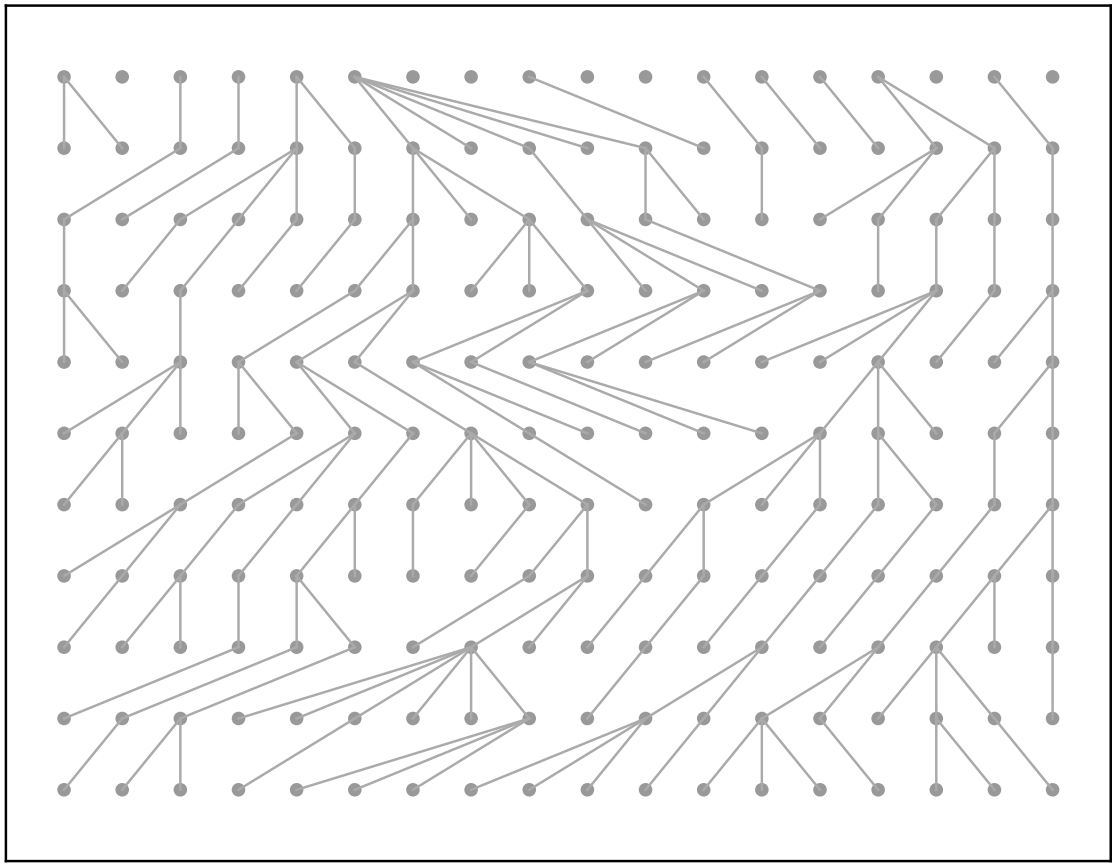


Figure – modèle de Wright-Fisher à $N = 18$ individus

On parle de relation de dualité.

Le point de vue généalogique

On peut reconstruire la composition de la population à partir de ses ancêtres.

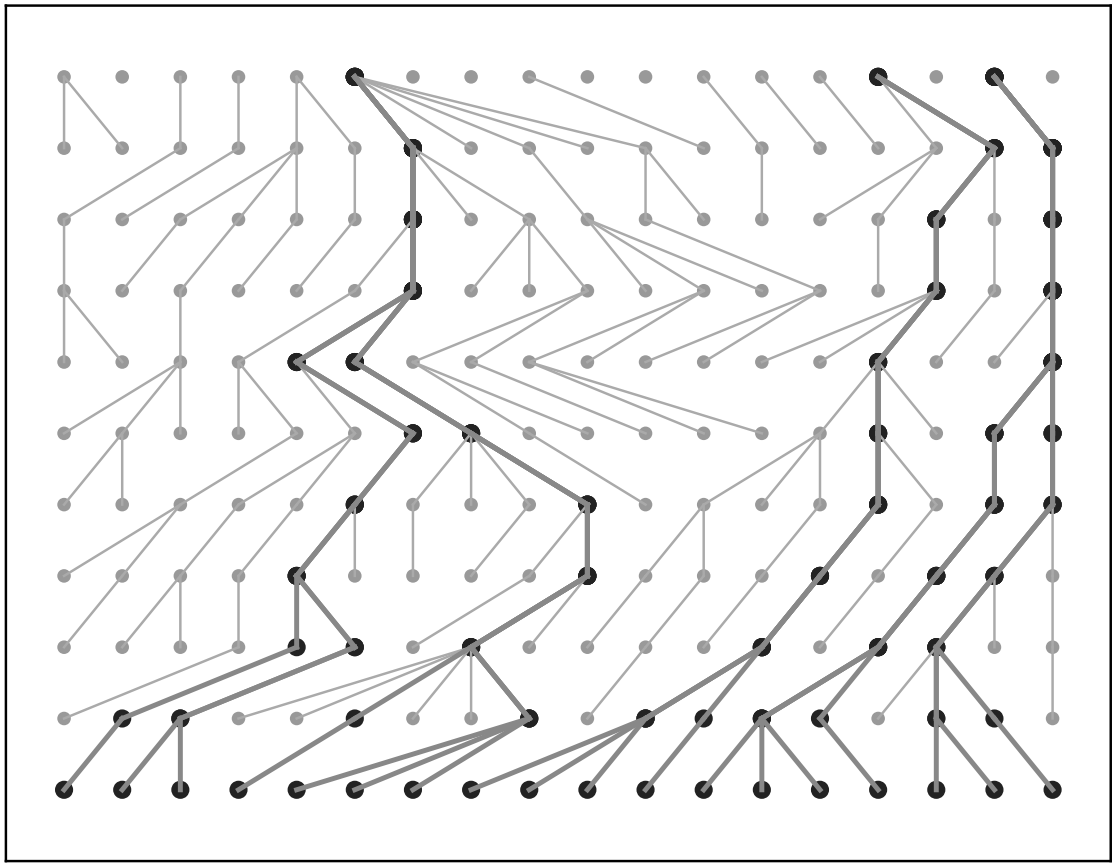


Figure – modèle de Wright-Fisher à $N = 18$ individus

On parle de relation de dualité.

Le point de vue généalogique

On peut reconstruire la composition de la population à partir de ses ancêtres.

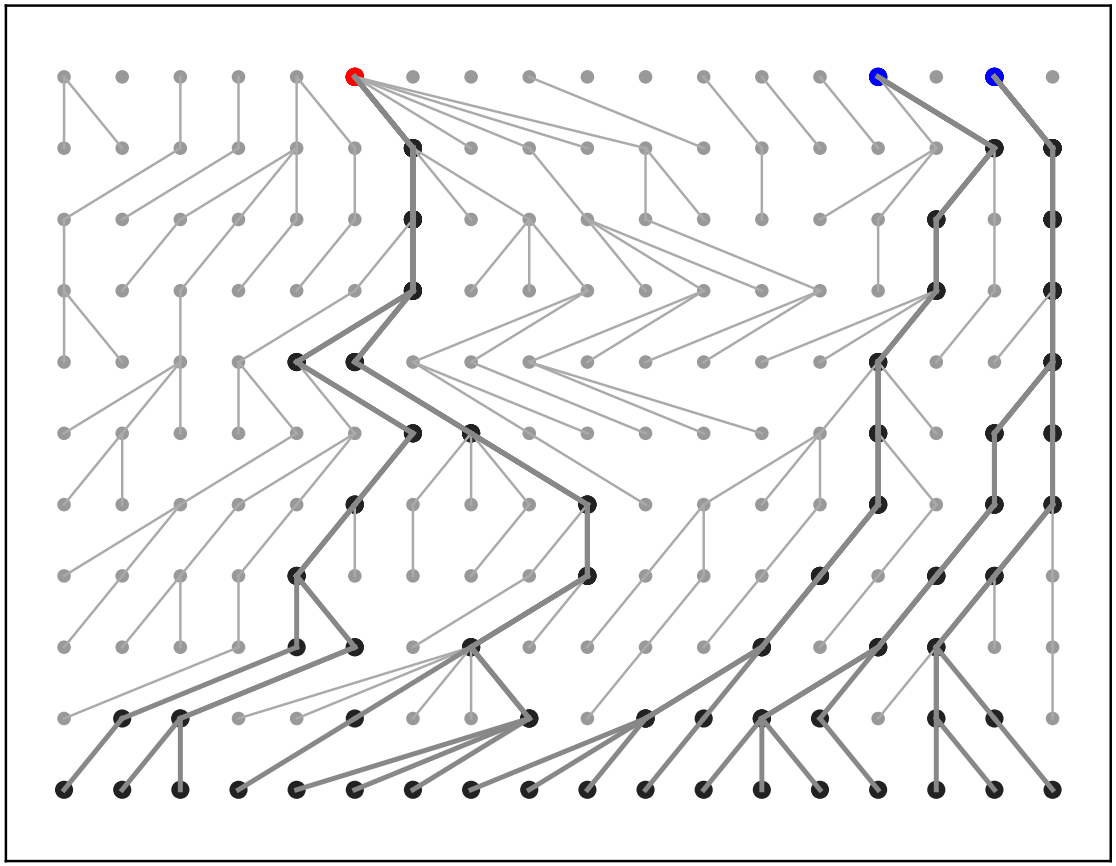


Figure – modèle de Wright-Fisher à $N = 18$ individus

On parle de relation de dualité.

Le point de vue généalogique

On peut reconstruire la composition de la population à partir de ses ancêtres.

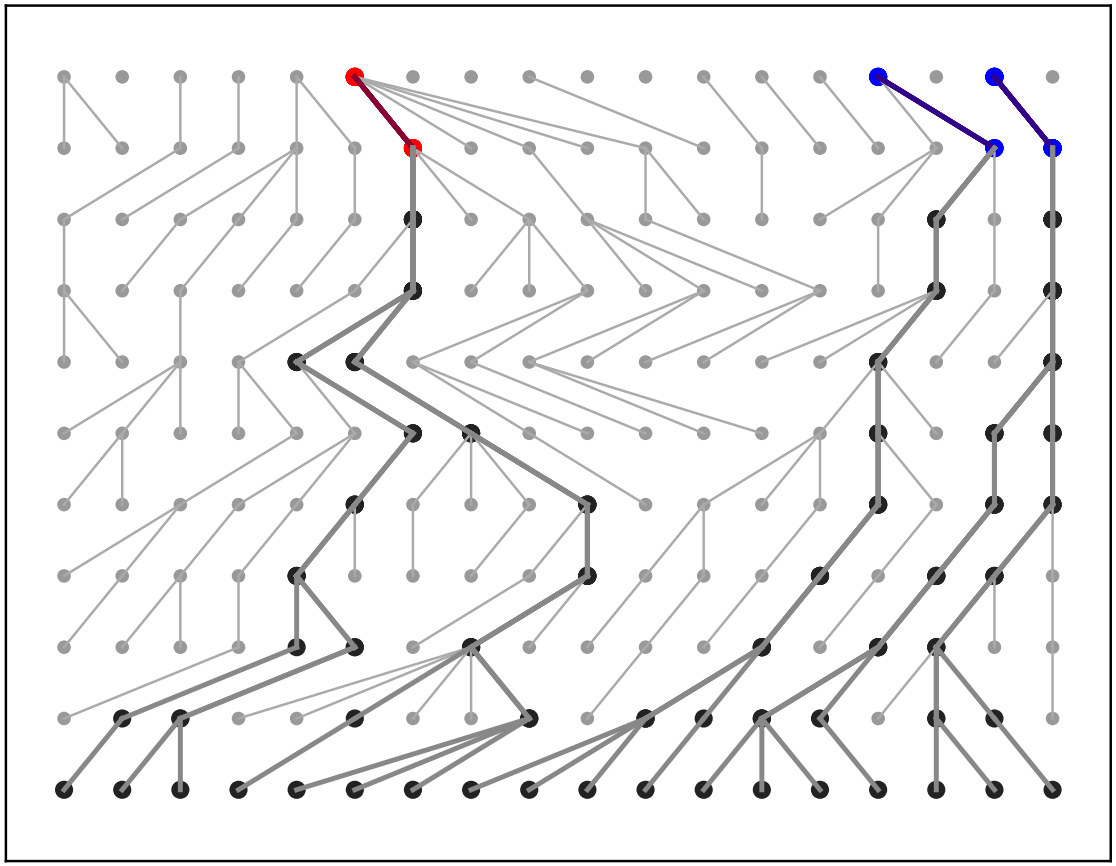


Figure – modèle de Wright-Fisher à $N = 18$ individus

On parle de relation de dualité.

Le point de vue généalogique

On peut reconstruire la composition de la population à partir de ses ancêtres.

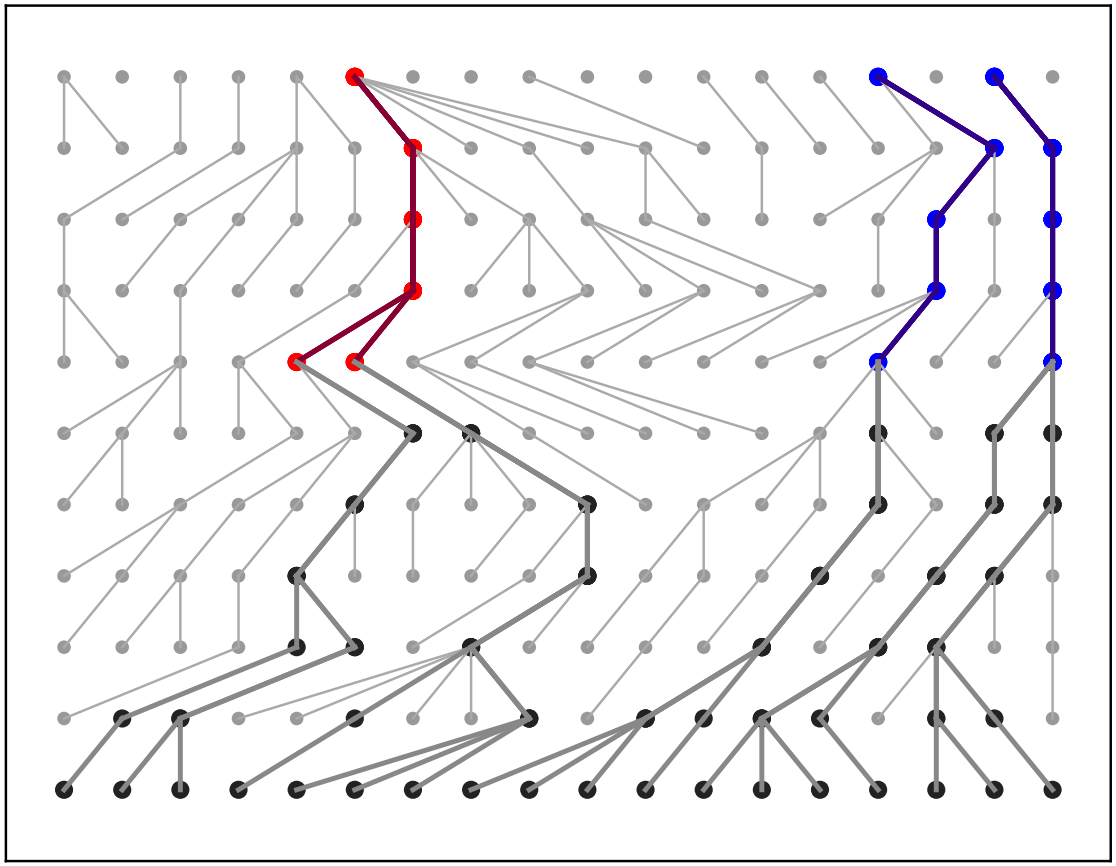


Figure – modèle de Wright-Fisher à $N = 18$ individus

On parle de relation de dualité.

Le point de vue généalogique

On peut reconstruire la composition de la population à partir de ses ancêtres.

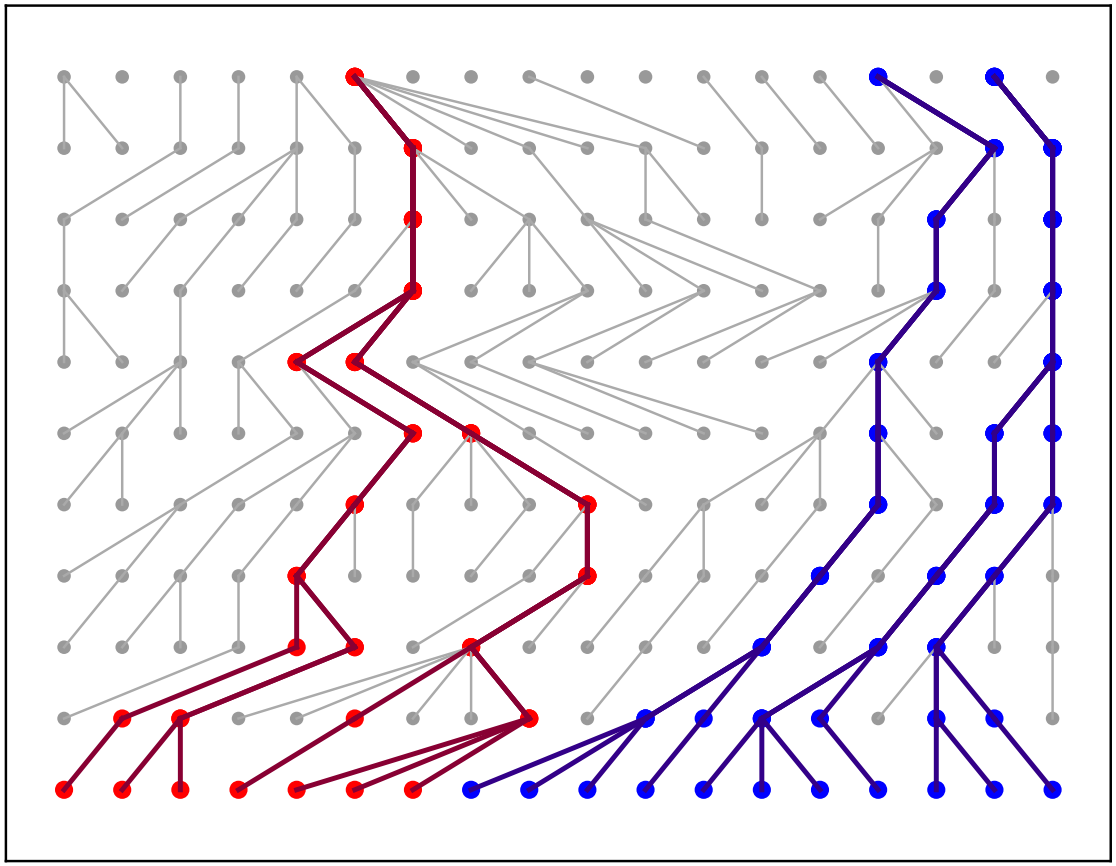


Figure – modèle de Wright-Fisher à $N = 18$ individus

On parle de relation de dualité.

Ancêtre commun le plus récent

Soient deux individus dans la population. On note T le nombre de générations qu'il faut remonter dans le passé pour trouver l'ancêtre commun le plus récent entre ces deux individus.

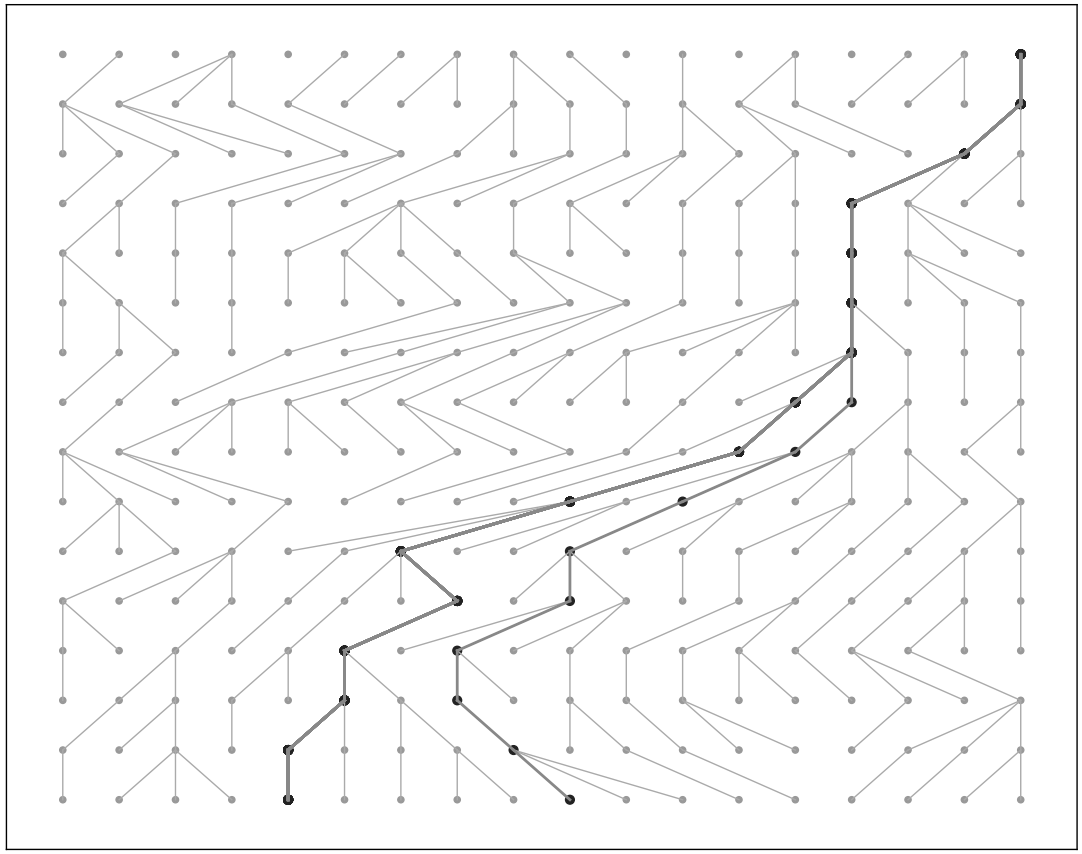


Figure – Ancêtre commun le plus récent de deux individus

Ancêtre commun le plus récent

Soient deux individus dans la population. On note T le nombre de générations qu'il faut remonter dans le passé pour trouver l'ancêtre commun le plus récent entre ces deux individus.
La variable aléatoire T suit une loi géométrique de paramètre $1/N$, *i.e.* pour $n \geq 0$,

$$\mathbb{P}(T = n) = \frac{1}{N}\left(1 - \frac{1}{N}\right)^{n-1}.$$

En particulier,

$$\mathbb{E}[T] = N.$$

De plus, lorsque $N \to \infty$,

$$\frac{1}{N} T \xrightarrow[N\to\infty]{loi} \mathcal{E}(1).$$

Le coalescent de Kingman

Cela suggère de considérer l'objet suivant :

On attache à chaque **paire** d'individus une variable exponentielle de paramètre $1/N$. Soit T_1 la plus petite de ces variables exponentielles, alors à $t = T_1$, les deux lignées corresondantes fusionnent, puis on retire de nouvelles variables exponentielles pour chaque paire d'individus restant, et ainsi de suite.

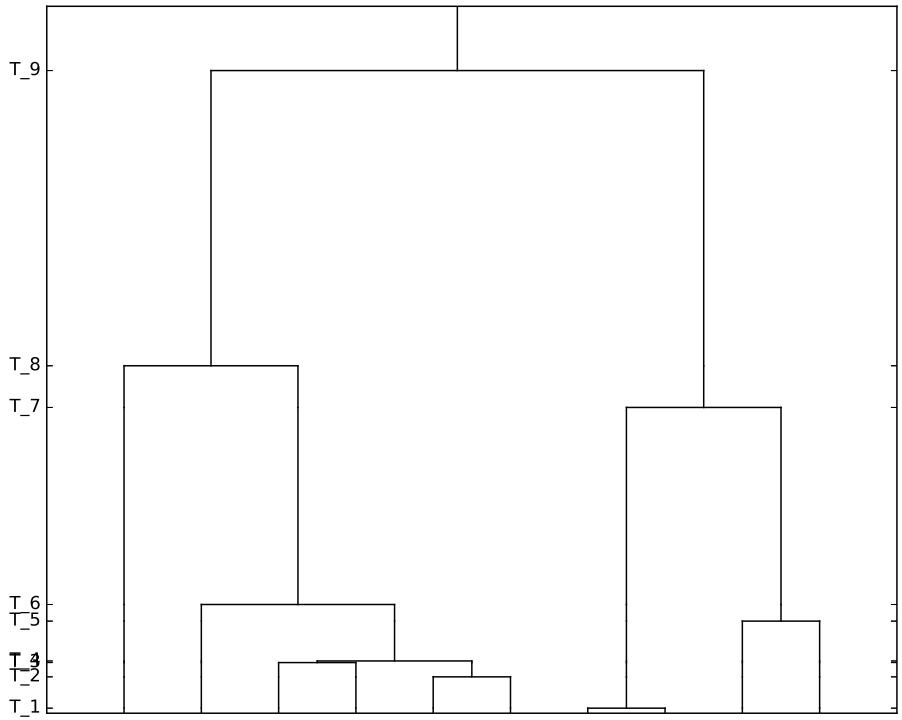


Figure – Coalescent de Kingman avec 10 lignées

Ancêtre commun le plus récent

On échantillone k individus dans une population qui en compte N au total et on note $T_{MRCA}(k)$ l'âge de leur ancêtre commun le plus récent. Alors

$$\mathbb{E}\left[T_{MRCA}(k)\right] = 2N\left(1 - \frac{1}{k}\right).$$

et

$$\mathbb{V}\left[T_{MRCA}(k)\right] = 4N^2 \times \left[2\sum_{i=1}^{k-1}\frac{1}{i^2} - 3 + \frac{2k+1}{k^2}\right].$$

Ancêtre commun le plus récent

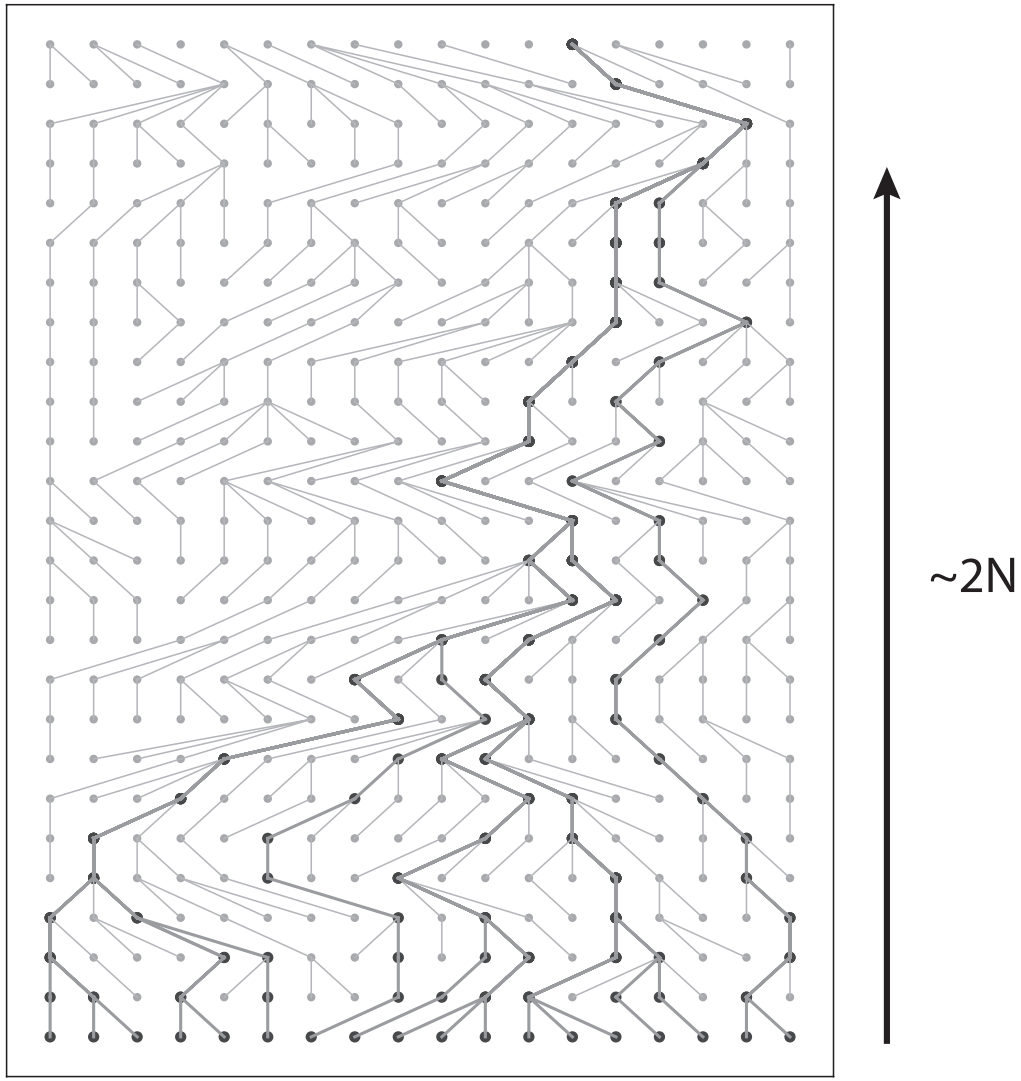


Figure – Temps de coalescence dans le modèle de Wright-Fisher

Et pour *homo sapiens* ?

Les cellules de tous les eucaryotes contiennent des mitochondries, qui produisent l'ATP, servant à la production d'énergie dans la cellule. Les mitochondries contiennent des petites séquences d'ADN, qu'on appelle ADN mitochondrial.

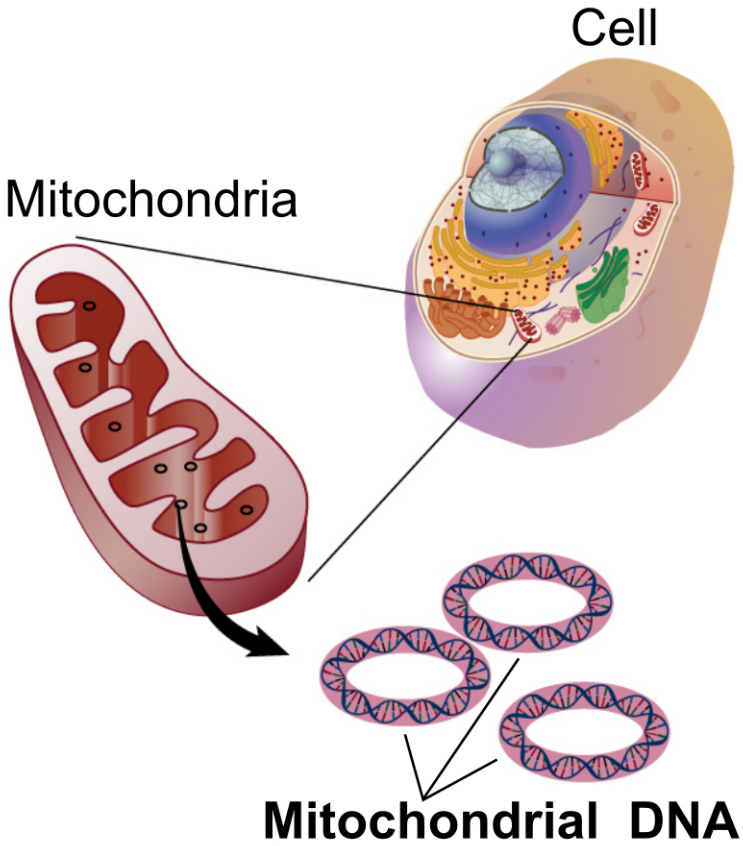


Figure – L'ADN mitochondrial

Et pour *homo sapiens* ?

Les cellules de tous les eucaryotes contiennent des mitochondries, qui produisent l'ATP, servant à la production d'énergie dans la cellule. Les mitochondries contiennent des petites séquences d'ADN, qu'on appelle ADN mitochondrial.

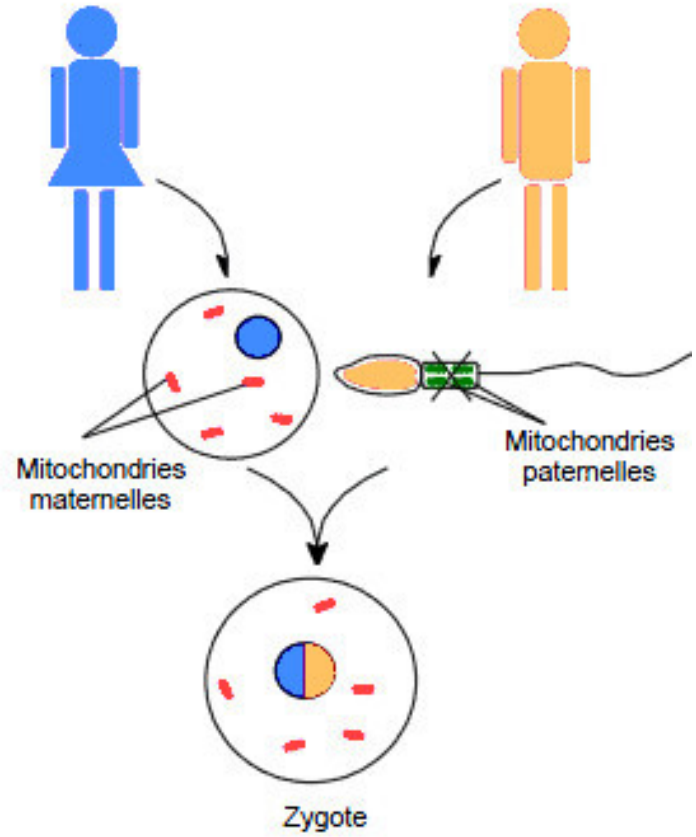


Figure – L'ADN mitochondrial est hérité uniquement de la mère

Et pour *homo sapiens* ?

Les cellules de tous les eucaryotes contiennent des mitochondries, qui produisent l'ATP, servant à la production d'énergie dans la cellule. Les mitochondries contiennent des petites séquences d'ADN, qu'on appelle ADN mitochondrial.

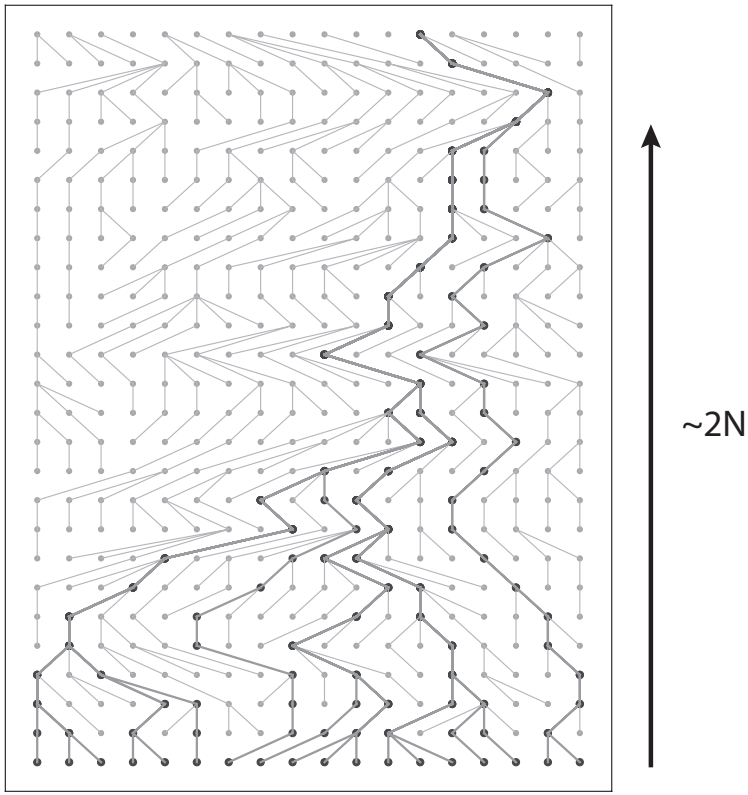


Figure – La généalogie de l'ADN mitochondrial correspond à un modèle de reproduction haploide

← Ève mitochondriale

Pour homo sapiens, on estime que cette Ève mitochondriale vivait en Afrique il y a entre 100 000 et 200 000 ans.

$N = 4000$? (une génération $\approx$ 25 ans)

Un modèle diploïde

On reprend le modèle de Wright-Fisher, mais à chaque génération, chaque individu choisit **deux** parents au hasard parmi ceux de la génération précédente.

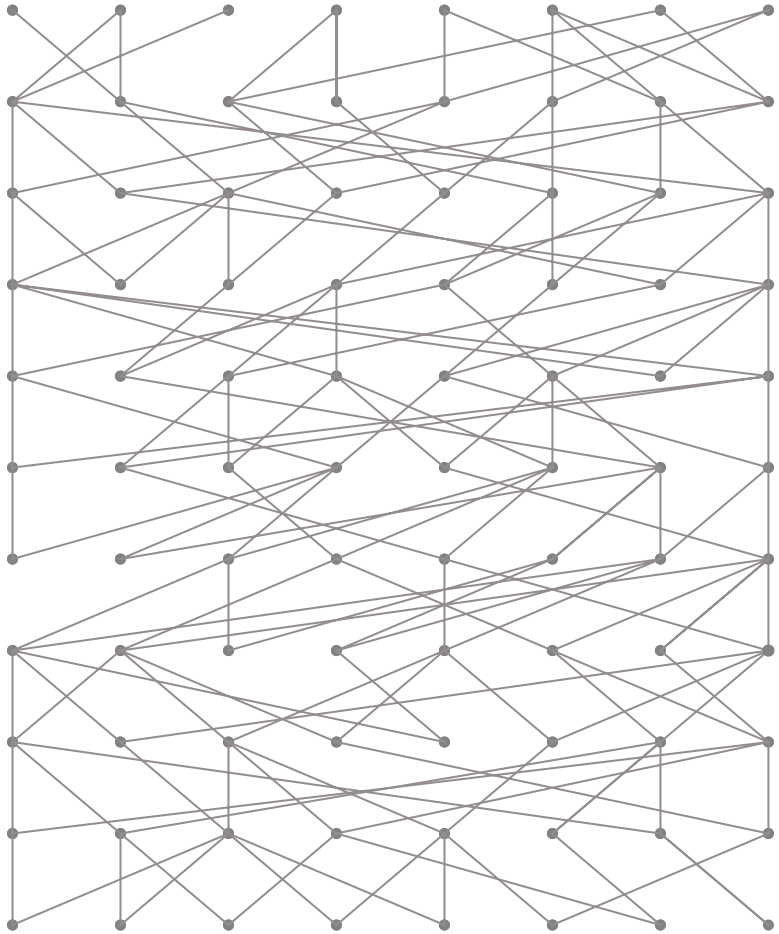


Figure – Modèle de Wright-Fisher diploide

Un modèle diploïde

On reprend le modèle de Wright-Fisher, mais à chaque génération, chaque individu choisit **deux** parents au hasard parmi ceux de la génération précédente.

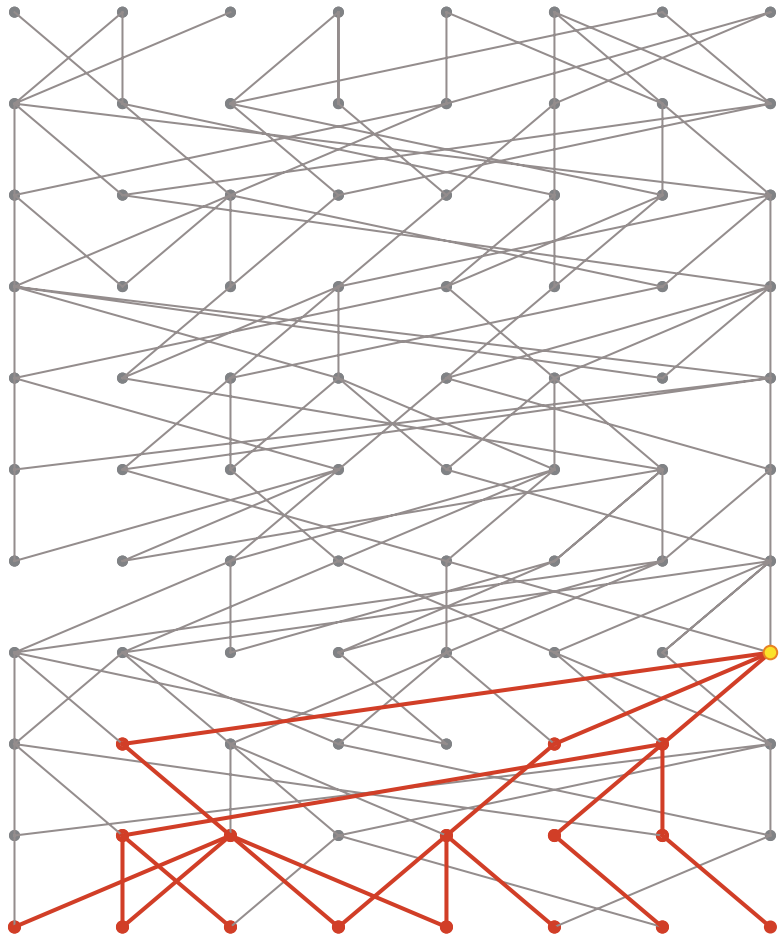


Figure – Modèle de Wright-Fisher diploide

Ancêtres communs en population diploide

On note $\mathcal{T}_N$ le nombre de générations qu'il faut remonter pour trouver un individu qui est un ancêtre de **tous** les individus de la génération présente.

Lorsque N tend vers l'infini, $\mathcal{T}_N$ est du même ordre que

1. N ?
2. $N/2$?
3. $\sqrt{N}$?
4. 1 ?

Ancêtres communs en population diploide

On note $\mathcal{T}_N$ le nombre de générations qu'il faut remonter pour trouver un individu qui est un ancêtre de **tous** les individus de la génération présente.

Lorsque N tend vers l'infini, $\mathcal{T}_N$ est du même ordre que

1. N ?
2. $N/2$?
3. $\sqrt{N}$?
4. 1 ?

$$\mathcal{T}_N \approx \log_2(N)$$

Ancêtres communs en population diploide

On note $\mathcal{T}_N$ le nombre de générations qu'il faut remonter pour trouver un individu qui est un ancêtre de **tous** les individus de la génération présente.

Theorem (Chang 1999)

Pour tout $\varepsilon > 0$, lorsque N tend vers l'infini,

$$\mathbb{P}\left(\left|\frac{\mathcal{T}_N}{\log_2(N)} - 1\right| > \varepsilon\right) \to 0.$$

Ancêtres communs en population diploide

Après un nombre suffisant de générations, arrive un moment où **tous** les individus de la génération n sont **soit** des ancêtres communs de toute la population, **soit** n'ont laissé aucun descendant dans la population.
Soit $\mathcal{U}_N$ la première génération pour laquelle c'est le cas.

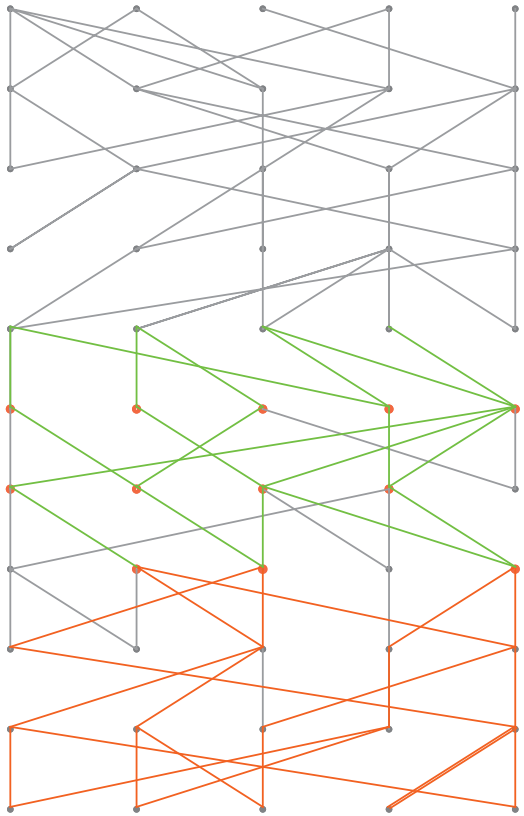


Figure – Illustration avec $N = 5$.

Ancêtres communs en population diploide

Après un nombre suffisant de générations, arrive un moment où **tous** les individus de la génération n sont **soit** des ancêtres communs de toute la population, **soit** n'ont laissé aucun descendant dans la population.
Soit $\mathcal{U}_N$ la première génération pour laquelle c'est le cas.

Theorem (Chang 1999)

Il existe $\gamma > 1$ *tel que, pour tout* $\varepsilon > 0$,

$$\mathbb{P}\left(\left|\frac{\mathcal{U}_N}{\log_2(N)} - \gamma\right| > \varepsilon\right) \underset{N\to\infty}{\longrightarrow} 0.$$

La valeur de γ est ≈ 1.7698.

Arbres de gènes

Si l'on suit l'histoire d'un gène de chaque individu, on obtient un sous-arbre de l'arbre diploide correspondant à un modèle de Wright-Fisher haploide.

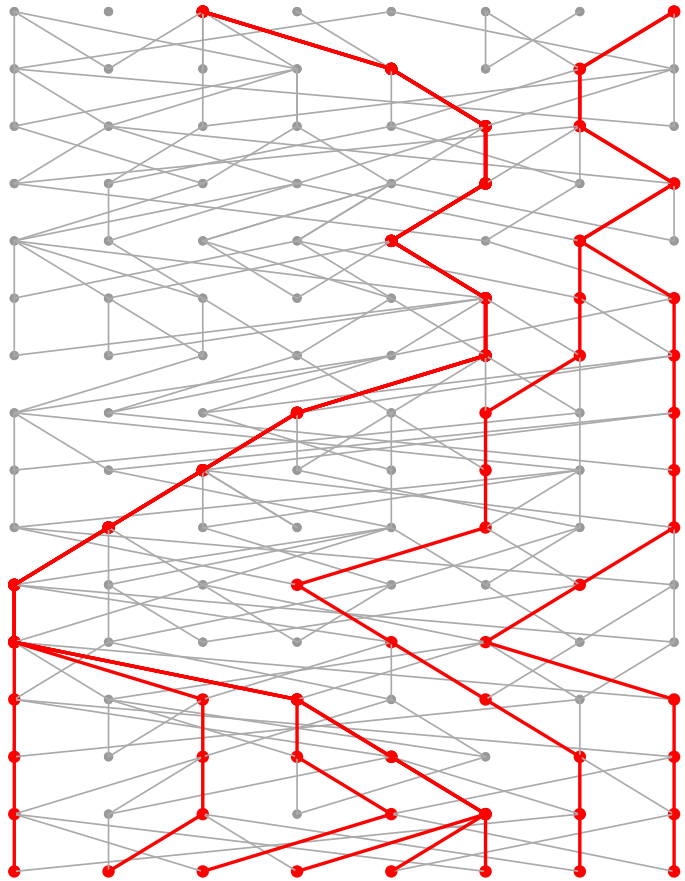


Figure – Chaque gène possède sa propre généalogie

Arbres de gènes

Si l'on suit l'histoire d'un gène de chaque individu, on obtient un sous-arbre de l'arbre diploide correspondant à un modèle de Wright-Fisher haploide.

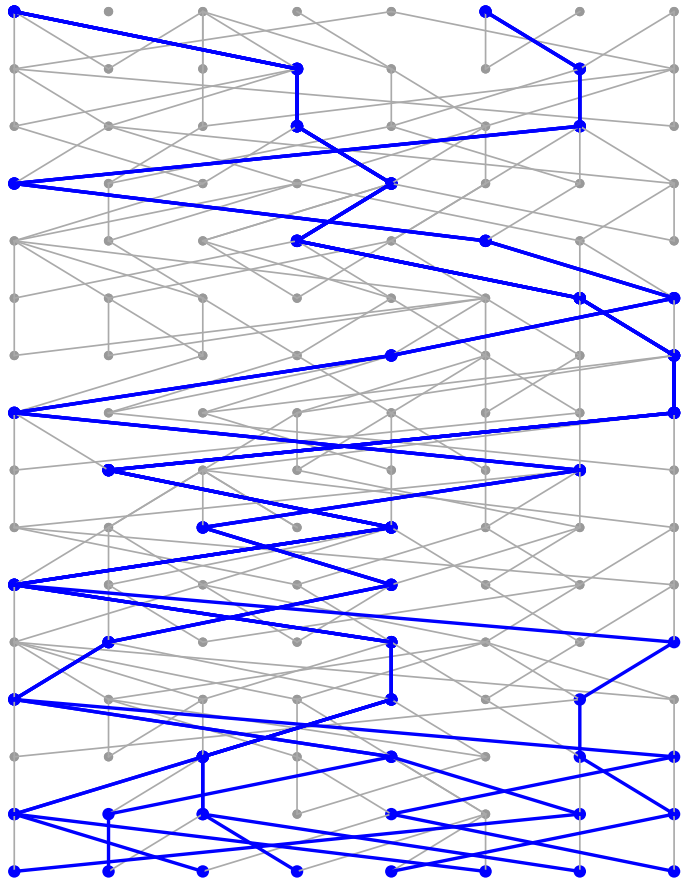


Figure – Chaque gène possède sa propre généalogie

2. Inférence démographique

Le coalescent de Kingman avec mutations

On suppose que des mutations se produisent chez les individus à un taux constant μ au cours du temps. On suppose que chaque mutation se produit à un site différent du génome, un individu présente donc toutes les mutations qui se sont produites chez ses ancêtres. On définit alors le coalescent de Kingman avec mutations.

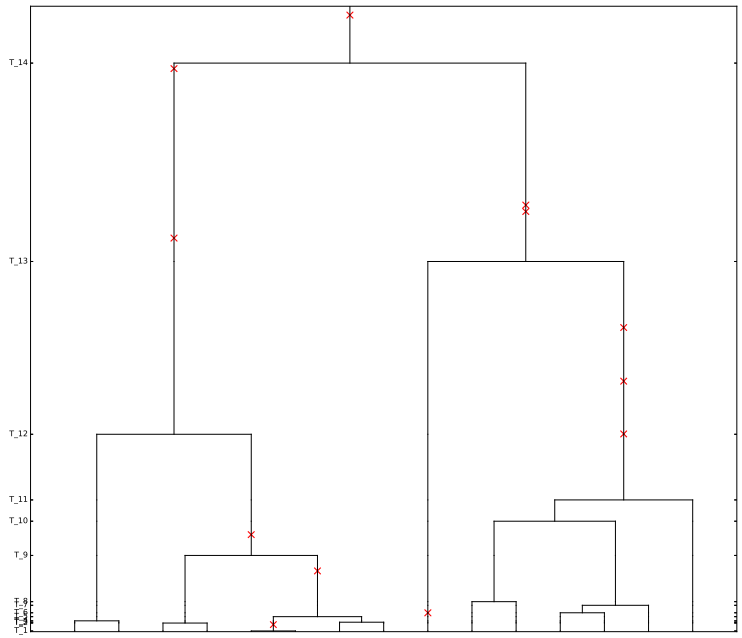


Figure – Coalescent de Kingman avec mutations

Chaque branche présente un nombre de mutations suivant une loi de Poisson de paramètre μl, où l est la longueur de la branche en question.

L'estimateur de Watterson

Soit $S^{(k)}$ le nombre de mutations présentes dans un échantillon de k individus. On pose

$$\hat{N}_k = \frac{S^{(k)}}{2\mu \sum_{i=1}^{k-1} \frac{1}{i}}.$$

Theorem

Lorsque k tend vers l'infini,

$$\sqrt{\frac{2\mu}{N} \log(k)} \left(\hat{N}_k - N \right) \xrightarrow[k \to \infty]{loi} \mathcal{N}(0,1).$$

L'estimateur de Watterson

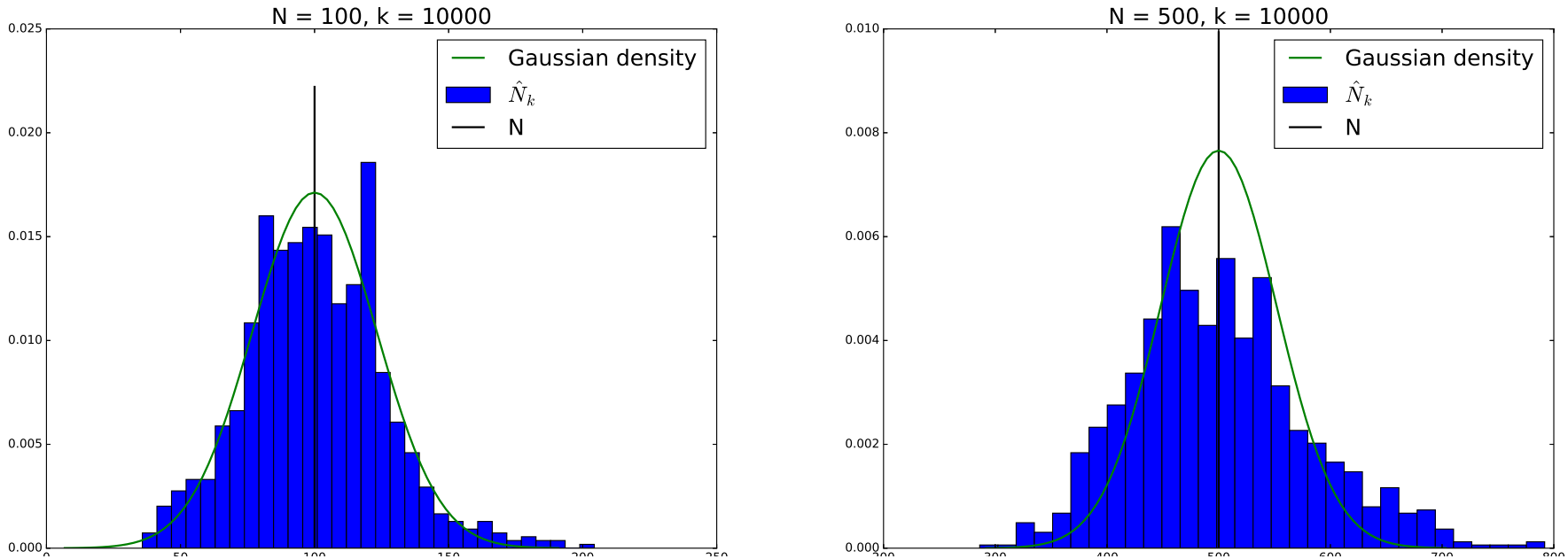


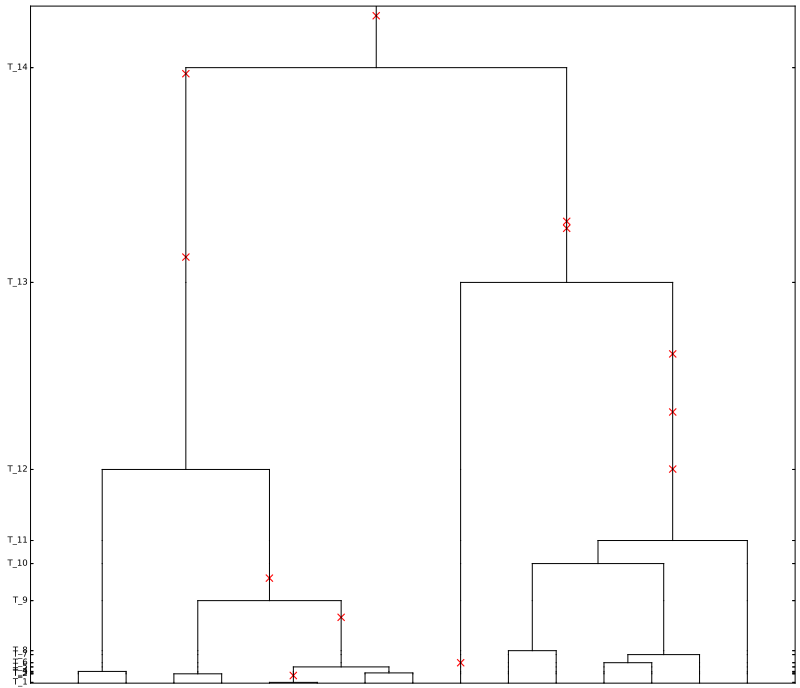
Figure – Histogramme de l'estimateur de Watterson pour $k = 10000$, $N = 100$ et $N = 500$.

La valeur de cet estimateur correspond très rarement au nombre réel d'individus dans la population, mais représente plutôt la taille "idéale" de population qui correspond à la diversité génétique observée dans la population. On parle de *taille efficace* de population.

Loi du nombre d'allèles distincts

On associe entre eux les individus qui portent les mêmes mutations. On obtient ainsi un certain nombre de classes d'individus qui ont la même séquence génétique (au locus que l'on étudie). On dit qu'ils portent le même allèle.

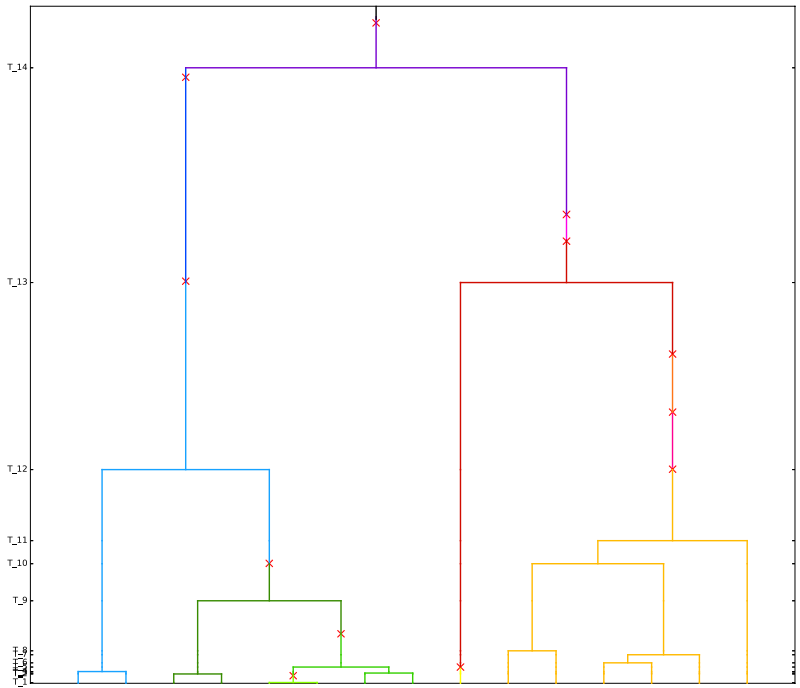
Soit A_k le nombre d'allèles distincts (= le nombre de classes) lorsque l'on séquence k individus.



Loi du nombre d'allèles distincts

On associe entre eux les individus qui portent les mêmes mutations. On obtient ainsi un certain nombre de classes d'individus qui ont la même séquence génétique (au locus que l'on étudie). On dit qu'ils portent le même allèle.

Soit A_k le nombre d'allèles distincts (= le nombre de classes) lorsque l'on séquence k individus.



L'urne de Hoppe

On considère une urne dans laquelle on place initialement θ boules noires. On itère ensuite la procédure suivante

- On tire une boule de l'urne uniformément au hasard.
- Si c'est une boule noire, on met dans l'urne une boule d'une nouvelle couleur (pas déjà présente dans l'urne), et on replace la boule noire dans l'urne.
- Sinon on prend une deuxième boule de la même couleur que celle qu'on vient de tirer et on replace les deux dans l'urne.

On note $\tilde{A}_k$ le nombre de couleurs différentes (hors boules noires) après k tirages. Alors si $\theta = 2N\mu$,

$$\tilde{A}_k \overset{d}{=} A_k.$$

La formule d'Ewens

On en déduit

$$\mathbb{E}[A_k] \sim \theta \ln k.$$

Si de plus on note N_j le nombre d'allèles portés par j individus, alors on a forcément

$$\sum_{j=1}^{k} jN_j = k.$$

Theorem (Formule d'Ewens)

Soient $k \geq 1$ et $n_1, \ldots, n_k$ tels que $\sum_{j=1}^{k} jn_j = k$, alors

$$\mathbb{P}(N_1 = n_1, \ldots, N_k = n_k) = \frac{1}{\binom{\theta+k-1}{k-1}} \prod_{j=1}^{k} \frac{\left(\frac{\theta}{j}\right)^{n_j}}{n_j!}.$$

3. La sélection naturelle

Le modèle de Wright-Fisher avec sélection

N individus qui se reproduisent et qui peuvent être de type A ou bien a. On note

$$X(t) = \text{ le nombre d'individus de type } A \text{ à la génération } t.$$

Lorsqu'un individu de la génération $t+1$ choisit son parent dans la génération t, la probabilité qu'il choisisse un individu de type A est

$$\frac{(1+s)X(t)}{(1+s)X(t)+N-X(t)},$$

tandis que la probabilité qu'il choisisse un individu de type a est

$$\frac{N-X(t)}{(1+s)X(t)+N-X(t)}.$$

Si $s>0$, l'allèle A est sélectionné positivement, si $s<0$, il est sélectionné négativement.

Le modèle de Wright-Fisher avec sélection

Si $s = \frac{s_0}{N}$, alors $X^N(t) = \frac{X(\lfloor Nt \rfloor)}{N}$ converge vers X_t tel que

$$X_{t+dt} = X_t + s_0 X_t(1 - X_t)dt + \mathcal{N}(0, X_t(1 - X_t)dt).$$

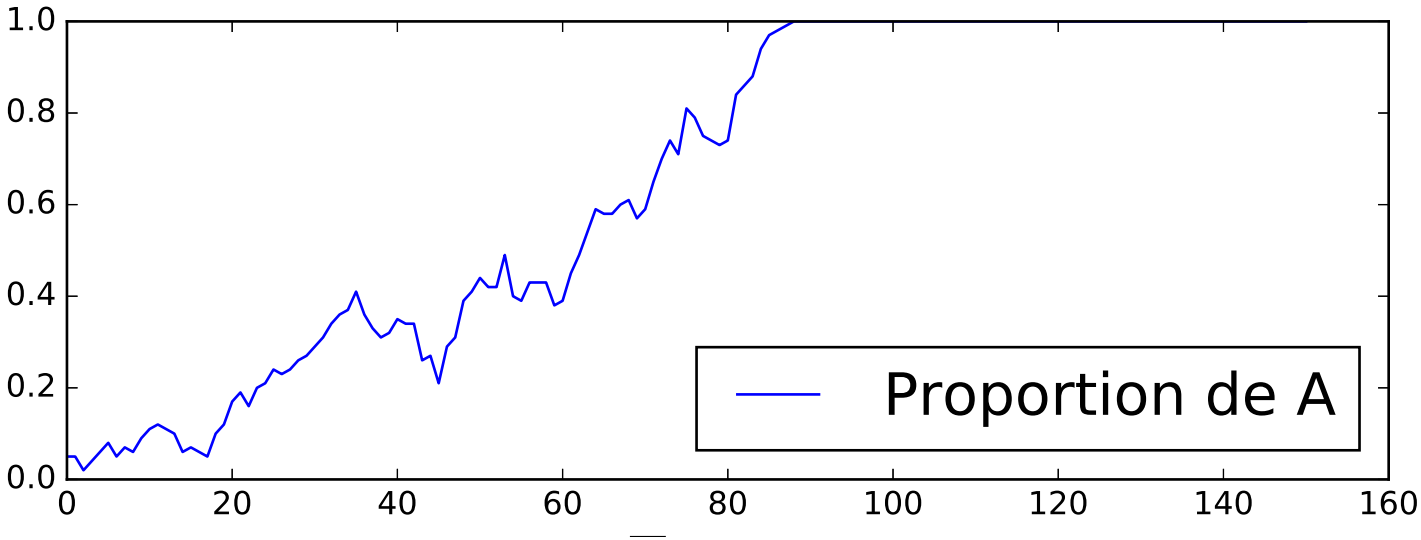
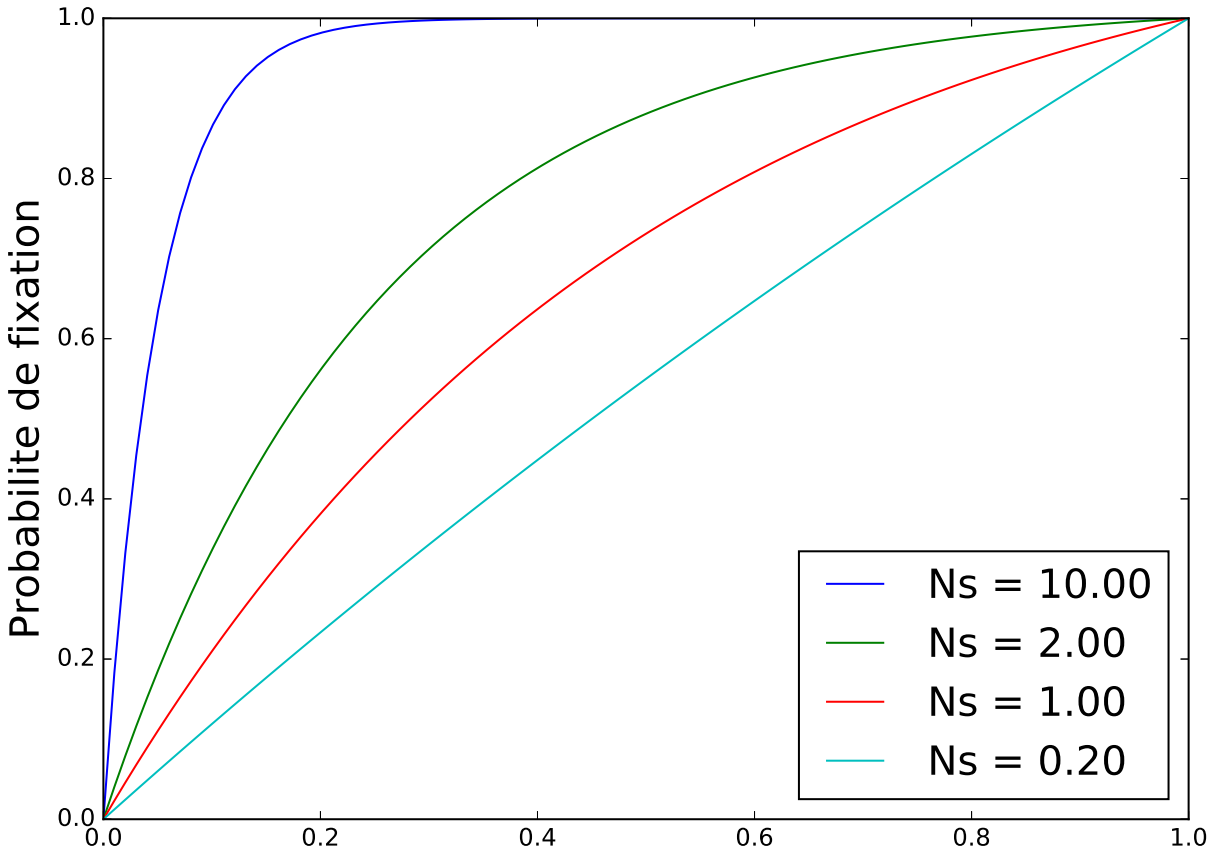


Figure – Modèle de Wright-Fisher avec sélection

Le modèle de Wright-Fisher avec sélection

Alors

$$\mathbb{P}\left(X_\infty = 1 \mid X_0 = x_0\right) = \frac{1 - e^{-2s_0 x_0}}{1 - e^{-2s_0}}.$$



Références

- Sylvie Méléard (2008). « Modèles Aléatoires En Ecologie et Evolution ». In : *Cours de 3ème année à l'Ecole Polytechnique*
- Alison Etheridge (2011). *Some Mathematical Models from Population Genetics*. T. 2012. Lecture Notes in Mathematics. Lectures from the 39th Probability Summer School held in Saint-Flour, 2009, Ecole d'Eté de Probabilités de Saint-Flour. Springer, Heidelberg. isbn : 978-3-642-16631-0
- Joseph T. Chang (1999). « Recent Common Ancestors of All Present-Day Individuals ». In : *Advances in Applied Probability* 31.4, p. 1002–1026
- Peter Ralph et Graham Coop (2013). « The Geography of Recent Genetic Ancestry across Europe ». In : *PLoS Biol* 11.5, e1001555