

L'énigme de la valeur C

J.R. Lobry & A. Necşulea & É. Vautrin

Quelques exercices de manipulation et de représentation graphique des données. La *valeur C* représente la taille du génome. Elle est mesurée le plus souvent en pico-grammes (pg), ou encore en nombre de paires de bases nucléotidiques. On pourrait lui supposer un lien *a priori* avec la complexité morphologique des organismes. Cette fiche propose de voir ce qu'il en est exactement.

Table des matières

1	La valeur C : un invariant biologique	2
1.1	Haploïde et diploïde	2
1.2	Unités de masse	3
1.3	Unités de longueur	3
1.3.1	Nombre de paires de bases	3
1.3.2	Longueur physique	4
2	Le paradoxe de la valeur C	4
3	La taille du génome et le nombre de gènes	4
4	La variabilité de la taille du génome chez les vertébrés	7
5	Valeur C et taille des globules rouges	7
	Références	8

Le contenu en acide désoxyribonucléique des tissus animaux a déjà fait l'objet de nombreuses recherches¹. Les auteurs ont exprimé leurs résultats en pourcentage du poids frais ou sec des noyaux et cela leur a permis de relever une corrélation, au moins grossière, entre la teneur des tissus en acide désoxyribonucléique et leur richesse en noyaux, telle que la révèlent les images histologiques. Il nous a semblé qu'il serait d'un réel intérêt de préciser la quantité d'acide désoxyribonucléique *par noyau* et de voir quelles variations cette quantité est susceptible de subir de noyau à noyau chez un même individu, d'individu à individu dans une même espèce et d'espèce animale à espèce animale. Quelques premiers coups de sonde dans cette direction, effectués en collaboration avec notre maître, le Prof. A. BOIVIN, nous ont laissé entrevoir une remarquable constance dans la teneur en acide désoxyribonucléique du noyau de toutes les cellules et chez tous les individus dans une même espèce animale². Nous voudrions apporter ici de nouvelles données d'expérience à l'appui de cette thèse et décrire sommairement les techniques que nous avons mises en œuvre.

FIG. 1 – Copie d'écran du paragraphe introductif de l'article de Vendrely & Vendrely de 1948 [8] qui est à l'origine du terme de *valeur C* pour désigner la taille d'un génome.

1 La valeur C : un invariant biologique

La variabilité étant le substrat de l'évolution, il est très rare de rencontrer des constantes dans le domaine du vivant. Les premiers biologistes qui mesurèrent la masse d'ADN dans les noyaux des cellules des eucarya (*cf* figure 1) trouvèrent donc très remarquable que cette valeur soit constante :

1. entre les cellules des différents tissus d'un même organisme.
2. entre les cellules des différents individus d'une même espèce.

On a pris depuis l'habitude de désigner par *valeur C*, pour valeur constante, la masse de ce que l'on appelle aujourd'hui le génome. L'ADN étant un hétéropolymère linéaire sans ramifications, il y a une relation de proportionalité entre sa masse et sa longueur, aussi la valeur C est également utilisée pour désigner la taille du génome.

1.1 Haploïde et diploïde

Il n'est pas tout à fait exact de dire que la masse d'ADN de toutes vos cellules est constante (*cf* figure 2). Par convention, la valeur C désigne la taille du

Teneurs comparées en acide désoxyribonucléique des cellules diploïdes et des cellules haploïdes chez l'espèce boeuf (en γ)					
	Cellules diploïdes				Cellules haploïdes
	Foie	Thymus	Rein	Pancréas	Spermatozoïdes
Acide désoxyribonucléique. . . .	$6,4 \cdot 10^{-6}$	$6,4 \cdot 10^{-6}$	$5,9 \cdot 10^{-6}$	$6,9 \cdot 10^{-6}$	$3,3 \cdot 10^{-6}$

FIG. 2 – Copie d'écran du tableau III de l'article de Vendrely & Vendrely de 1948 [8]. La notation γ est une notation obsolète pour μg . Notez que la masse d'ADN des cellules haploïdes (lignée germinale) est la moitié de celle des cellules diploïdes (lignée somatique).

génomme haploïde. Ceci permet de faciliter les comparaisons avec les espèces qui n'ont pas ou peu de cellules diploïdes. Mais notez bien que l'immense majorité des cellules qui forment votre corps (*soma*) sont diploïdes. La valeur C n'est par convention que celle de vos spermatozoïdes ou ovules.

1.2 Unités de masse

On utilise une sous-unité du système international, en général le picogramme (pg), soit :

$$1\text{pg} = 10^{-12}\text{g} = 10^{-15}\text{kg}$$

Exercice Sachant que la valeur C de l'homme est de l'ordre de 3 pg et votre corps contient environ 10^{14} cellules, quelle masse d'ADN, exprimée en kg, contient votre organisme ?

[1] 0.6

1.3 Unités de longueur

1.3.1 Nombre de paires de bases

Le monomère de base de l'ADN est la paire de base, on utilise donc souvent le nombre de paire de bases comme unité de longueur. Les multiples suivants sont les plus utilisés :

- 1 kb = 10^3 bp
- 1 Mb = 10^6 bp
- 1 Gb = 10^9 bp

On a la relation de conversion [2] suivante : nombre de paires de bases = masse en pg $\times 0.978 \cdot 10^9$, soit :

- 1 kb $\approx 10^{-6}$ pg
- 1 Mb $\approx 10^{-3}$ pg
- 1 Gb ≈ 1 pg

Exercice Donner le nombre de paire de bases présentes dans une de vos cellule (diploïde).

[1] 5.868e+09

1.3.2 Longueur physique

On a la relation [1] suivante : 1 paire de base ≈ 0.33 nm, soit :

- 1 kb ≈ 0.33 μ m
- 1 Mb ≈ 0.33 mm
- 1 Gb ≈ 0.33 m

Exercice Quelle serait la longueur, exprimée en mètre, si l'on mettait bout à bout toutes les molécules d'ADN de toutes les cellules de votre corps ?

[1] 1.93644e+14

Exprimer ce résultat en unités astronomiques, ua, correspondant à la distance moyenne entre la terre et le soleil (1 ua $\approx 1.5 \cdot 10^{11}$ m).


[1] 1290.96

Les données que nous allons étudier pour la suite se trouvent à l'adresse suivante :

<http://pbil.univ-lyon1.fr/R/donnees/>

2 Le paradoxe de la valeur C

La valeur C présente une très grande variabilité entre espèces. Comme cette mesure représente la quantité de matériel génétique d'un organisme, il est *a priori* raisonnable d'imaginer un lien entre la valeur C et la complexité morphologique. Nous vous proposons dans un premier temps d'analyser de manière qualitative la relation entre taille des génomes et la complexité des organismes.

Récupérez le fichier `cvalcomplex.txt` et importez ces données dans . Ce fichier contient pour 53 espèces le domaine taxonomique auquel elles appartiennent (virus, procaryote ou eucaryote), la nomenclature de l'espèce, le nom usuel (s'il est connu) et la valeur C, exprimée en Kb (milliers de paires de bases). Les données proviennent en grande partie de [3].

Représentez graphiquement la valeur C selon l'espèce. Donnez des couleurs différentes pour les trois domaines taxonomiques (virus, procaryote et eucaryote). Un exemple de représentation possible est donnée dans la figure 3.

Selon ce graphique, quelles sont les espèces qui ont la plus grande taille de génome ? Est-ce que la taille du génome dépend du domaine taxonomique ? Quelle est la position de l'homme sur ce graphique ?

3 La taille du génome et le nombre de gènes

Nous avons vu précédemment qu'il n'existe pas de relation évidente entre la taille des génomes et la complexité des organismes ; c'est ce qu'on appelle "le paradoxe de la valeur C". Ce paradoxe a été résolu avec la découverte de l'ADN non-codant. Les génomes eucaryotes sont (généralement) composés en grande proportion d'ADN qui ne code pas pour des protéines ou pour d'autres éléments fonctionnels. Il est donc possible qu'un génome de taille très importante ne contienne qu'une très faible proportion de matériel informatif, ce qui

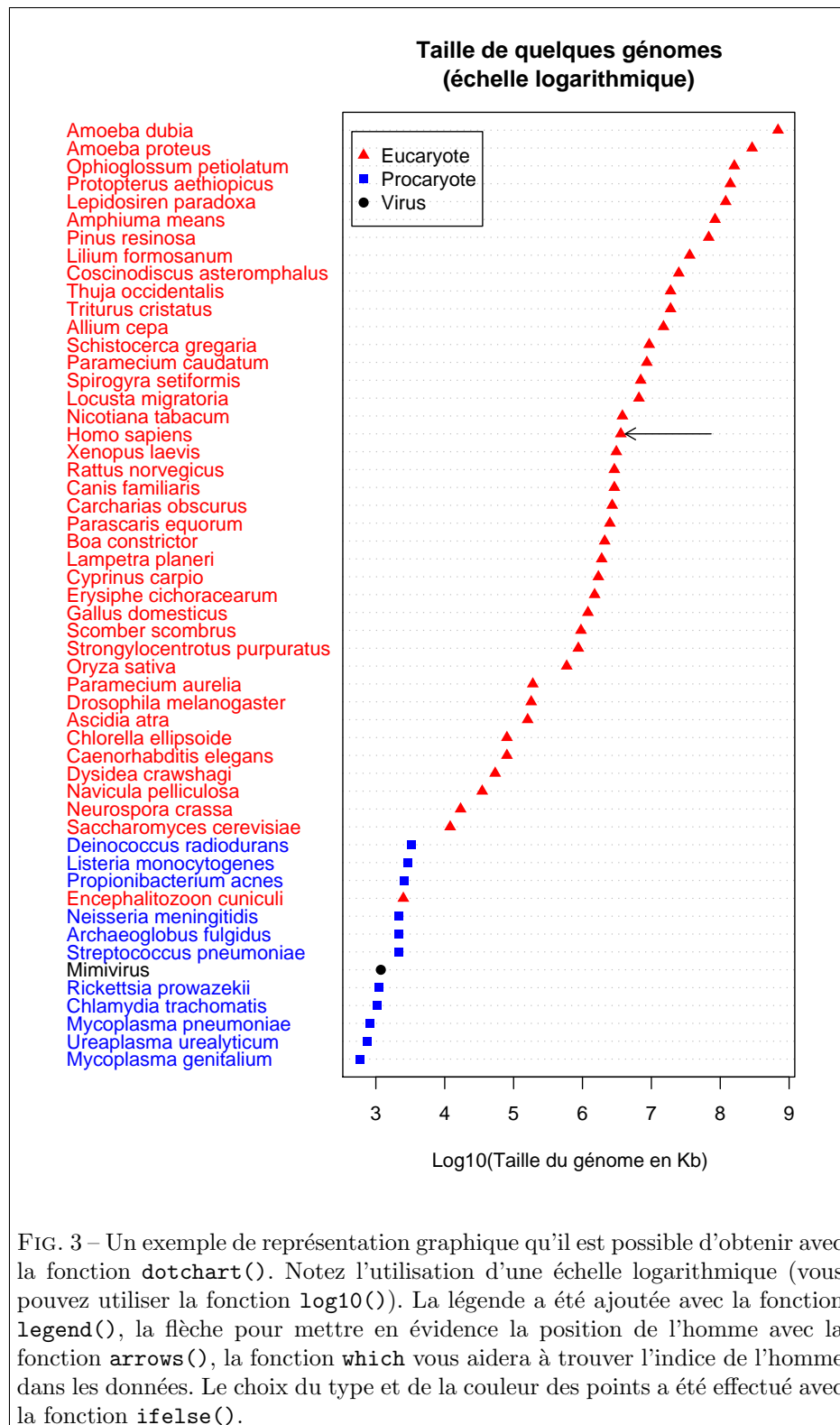
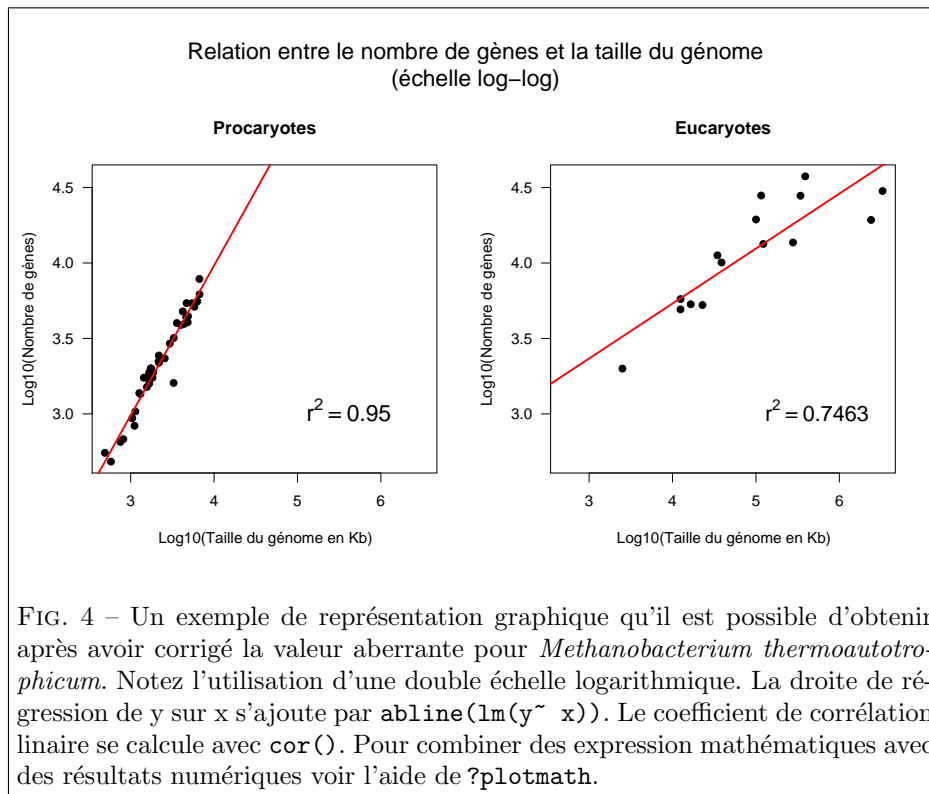


FIG. 3 – Un exemple de représentation graphique qu'il est possible d'obtenir avec la fonction `dotchart()`. Notez l'utilisation d'une échelle logarithmique (vous pouvez utiliser la fonction `log10()`). La légende a été ajoutée avec la fonction `legend()`, la flèche pour mettre en évidence la position de l'homme avec la fonction `arrows()`, la fonction `which` vous aidera à trouver l'indice de l'homme dans les données. Le choix du type et de la couleur des points a été effectué avec la fonction `ifelse()`.



explique l'absence de lien avec la complexité.

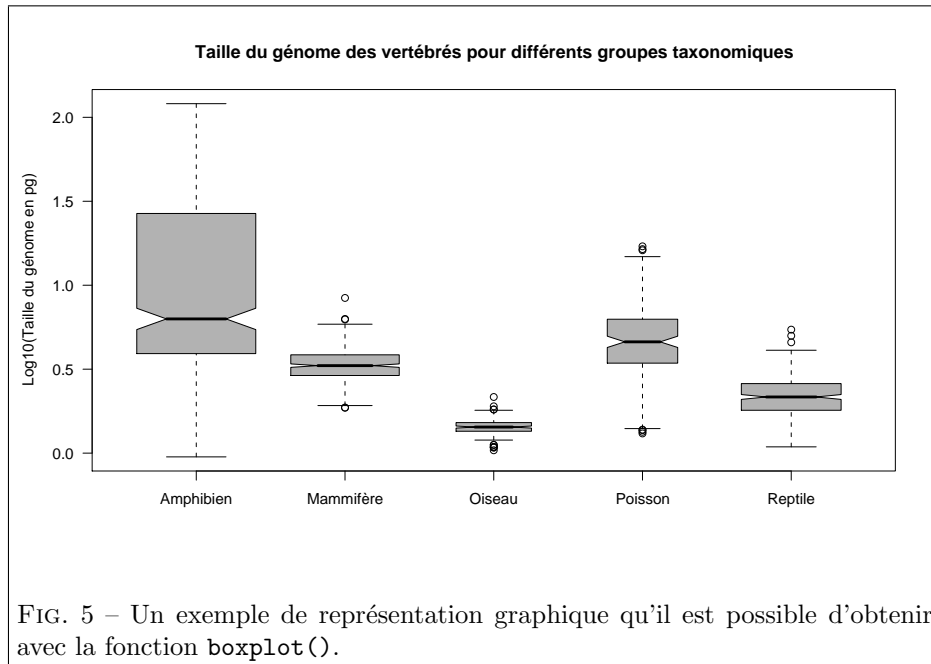
Pour mettre en évidence l'existence de l'ADN non-codant, nous vous proposons d'analyser la relation entre la taille des génomes et le nombre de gènes protéiques.

Récupérez le fichier suivant : `nbgenes.txt` et importez ces données dans . Ce fichier est similaire à celui étudié au-dessus ; il contient en plus le nombre de gènes protéiques pour chacune des espèces étudiées. **Attention** : ce jeu de données contient volontairement une valeur aberrante qu'il conviendra de corriger (cette valeur aberrante sera évidente sur le graphique des procaryotes).

Les données sont extraites du cours de J.W. Kimball : <http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/G/GenomeSizes.html>.

Représentez graphiquement le nombre de gènes en fonction de la taille du génome. Calculez le coefficient de corrélation entre ces deux variables et superposez le modèle linéaire sur le graphique précédent.

Représentez ensuite sur deux graphiques côte à côte le nombre de gènes en fonction de la valeur C, pour les procaryotes et les eucaryotes, en superposant les modèles linéaires correspondants. Calculez les deux coefficients de corrélation (*cf* figure 4). Quelle interprétation biologique donnez-vous aux résultats précédents ?



4 La variabilité de la taille du génome chez les vertébrés

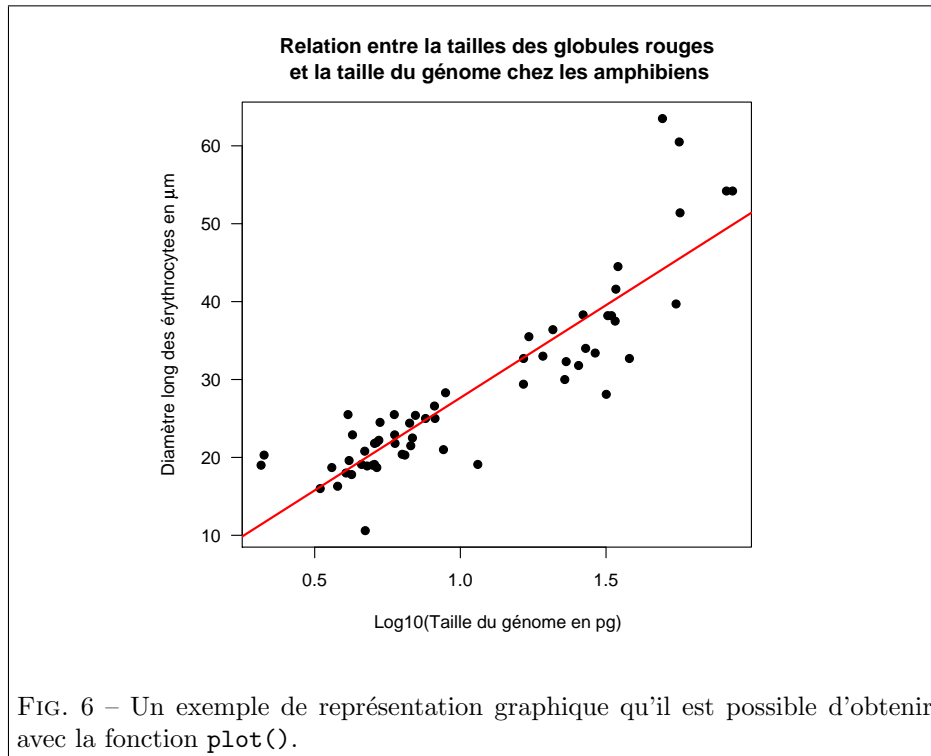
Au sein des vertébrés la taille du génome présente une très grande variabilité. Nous vous proposons pour la suite d'étudier la variation de la valeur C au sein des différentes classes d'animaux (mammifères, oiseaux, reptiles, amphibiens et poissons).

Nous allons analyser les données contenues dans le fichier suivant : `genome-size.txt`. Importez les données dans \mathbb{R} . Ce fichier contient pour 1416 espèces la classe taxonomique à laquelle elles appartiennent, la nomenclature officielle des espèces et leurs noms usuels, ainsi que la valeur C , exprimée en pg. Les données proviennent de la base de données de tailles de génome de T.R. Gregory (www.genomesize.com, [5]).

Représentez graphiquement la valeur C en fonction de la classe taxonomique (ou groupe). Quelles sont les groupes qui présentent la plus grande variabilité? Pouvez-vous comparer les valeurs C médianes des différents groupes? Un exemple de représentation graphique possible est donné dans la figure 5.

5 Valeur C et taille des globules rouges

La taille des génomes peut être liée à d'autres paramètres morphologiques et physiologiques des organismes. Dès les premières études sur la quantité d'ADN des cellules animales, une corrélation entre la taille du génome et la taille des cellules a été mise en évidence ([7], voir aussi [4]).



Nous vous proposons ici d’analyser la relation entre la valeur C et la taille des globules rouges (erythrocytes) chez les amphibiens. Vous disposez du fichier suivant : `cellgenome.amphi.txt`. Il contient les nomenclatures officielles et les noms usuels (en anglais) des espèces, la surface (mesurée en μm^2), le diamètre long et le diamètre court des globules rouges (mesurés en μm), ainsi que les valeurs C (exprimées en pg). Les données concernant la taille du génome proviennent de [5], les données concernant les tailles des globules rouges sont extraites de [6].

Analysez ce jeu de données. Est-ce que vous confirmez la relation mise en évidence par [7]? Un exemple de représentation possible est donné dans la figure 6.

Références

- [1] R.E. Dickerson, H.R. Drew, B.N. Conner, R.M. Wing, A.V. Fratini, and M.L. Kopka. The anatomy of A-, B-, and Z-DNA. *Science*, 216:475–485, 1982.
- [2] J. Doležel, J. Bartoš, H. Voglmayr, and J. Greilhuber. Nuclear DNA content and genome size of trout and human. *Cytometry*, 51A:127–128, 2003.
- [3] D. Graur and W.-H. Li. *Fundamentals of Molecular evolution (second edition)*. Sinauer associates, inc., Sunderland, Massachusetts, USA, 2000. ISBN 0-87893-266-6.
- [4] T.R. Gregory. Coincidence, coevolution or causation ? DNA content, cell size, and the C-value enigma. *Biological Reviews*, 76:65–101, 2001.
- [5] T.R. Gregory. Animal genome size database (<http://www.genomesize.com>). 2005.

- [6] T.R. Gregory. Cell size database (<http://www.genomesize.com/cellsizes>). 2005.
- [7] A. E. Mirsky and H. Ris. The desoxyribonucleic acid content of animal cells and its evolutionary significance. *Journal of General Physiology*, 34(4):451–462, 1951.
- [8] R. Vendrely and C. Vendrely. La teneur du noyau cellulaire en acide désoxyribonucléique à travers les organes, les individus et les espèces animales. *Experientia*, 4:434–436, 1948.