*Entropy* **2011**, *13*, 1465-1480; doi:10.3390/e13081465

OPEN ACCESS

*entropy*

**ISSN 1099-4300**

www.mdpi.com/journal/entropy

*Article*

# Size of the Whole *versus* Number of Parts in Genomes

**Antoni Hernández-Fernández [1,4], Jaume Baixeries [2], Núria Forns [3] and Ramon Ferrer-i-Cancho [4,*]**

[1] Departament de Lingüística General, Universitat de Barcelona, Gran Via de les Corts Catalanes 585, 08007 Barcelona (Catalonia), Spain; E-Mail: antonio.hernandez@upc.edu

[2] LARCA Research Group, Complexity and Quantitative Linguistics Lab, Departament de Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya, Campus Nord, Edifici Omega, Jordi Girona Salgado 1-3, 08034 Barcelona (Catalonia), Spain; E-Mail: jbaixer@lsi.upc.edu

[3] Departament de Microbiologia, Facultat de Biologia, Universitat de Barcelona, Av. Diagonal 645, 08028 Barcelona (Catalonia), Spain; E-Mail: nuforns@ub.edu

[4] Complexity and Quantitative Linguistics Lab, Departament de Llenguatges i Sistemes Informàtics, TALP Research Center, Universitat Politècnica de Catalunya, Campus Nord, Edifici Omega, Jordi Girona Salgado 1-3, 08034 Barcelona (Catalonia), Spain

* Author to whom correspondence should be addressed; E-Mail: rferrericancho@lsi.upc.edu.

*Received: 5 July 2011; in revised form: 22 July 2011 / Accepted: 28 July 2011 / Published: 5 August 2011*

---

**Abstract:** It is known that chromosome number tends to decrease as genome size increases in angiosperm plants. Here the relationship between number of parts (the chromosomes) and size of the whole (the genome) is studied for other groups of organisms from different kingdoms. Two major results are obtained. First, the finding of relationships of the kind "the more parts the smaller the whole" as in angiosperms, but also relationships of the kind "the more parts the larger the whole". Second, these dependencies are not linear in general. The implications of the dependencies between genome size and chromosome number are two-fold. First, they indicate that arguments against the relevance of the finding of negative correlations consistent with Menzerath-Altmann law (a linguistic law that relates the size of the parts with the size of the whole) in genomes are seriously flawed. Second, they unravel the weakness of a recent model of chromosome lengths based upon random breakage that assumes that chromosome number and genome size are independent.

**Keywords:** Menzerath-Altmann law; genome size; chromosomes

**PACS Codes:** 87.18.Wd Genomics; 89.75.Da Systems Obeying Scaling Laws; 87.15.A-, Theory, modeling, and computer simulation; 87.16.Sr Chromosomes, histones; 87.14.gk DNA

## 1. Introduction

Various studies have reported a negative correlation between genome size and number of chromosomes or B chromosomes in angiosperm plants [1,2]. Interestingly, Vinogradov argues that this negative correlation could be explained as a trade-off between different recombination mechanisms [1]. In contrast, it has been argued recently that theoretical models of chromosome length evolution [3,4] "and the current knowledge on the fluid nature of chromosomal rearrangements through time rule **against any special multiscale link between genome-level and chromosome-level patterns.** (boldface is ours)" [5]. Here it will be shown that dependencies between chromosome number and genome size are not a peculiarity of flowering plants, as it may be concluded from the pioneering work of Vinogradov [1], by examining various groups of organisms from different kingdoms: fungi, plants, and animals. As the size of the genomes increases, it will be shown that the number of chromosomes increases in some groups while in others it decreases. Evidence that these dependencies are not simply linear will be provided.

## 2. Results

$N$ is defined as the number of organisms of a group that is being analyzed. $G$ and $L_g$ are defined, respectively, as length of a genome in million base pairs (Mb) and the size of the genome in chromosomes.

### *2.1. Correlations between Genome Size and Chromosome Number*

Figures 1 and 2 show the relationship between $G$ and $L_g$ for the major groups of organisms analyzed in [6]. It can be seen that certain groups of organisms such as reptiles, birds and fungi, cluster in different regions of the space defined by $G$ and $L_g$. For certain groups of organisms (e.g., reptiles), a dependency between $G$ and $L_g$ can be seen. However, a rigorous statistical correlation test is necessary. Separate plots of the relationship between $G$ and $L_g$ for each group are provided in Appendix A. Table 1 shows a significant correlation between $G$ and $L_g$ is found in 9 out of 11 groups of organisms at a significance level of 0.05. The only groups where no significant correlation is found are birds and cartilaginous fishes. Therefore, $G$ is not indeed a constant function of $L_g$ for the majority of groups.

### *2.2. Non-Linearity*

Some light on the kind of functional dependency between $G$ and $L_g$ can be shed. If the relationship was purely linear, the point estimation of the slope should not show any dependency with either $G$ or $L_g$. Table 2 shows that this linearity test (see Methods for further details) rejects the null hypothesis that $G$ is a purely linear function of $L_g$ for all groups (p-value $< 10^{-7}$). Non-linearity is consistent with

the plots in Figures 1 and 2 and in the Appendix A where it can easily be seen that the slope of a linear approximation in double logarithmic scale deviates, in many cases, clearly from one, the expected slope if the relationship was linear. However, our test cannot exclude that linearity is present in some part of the series despite the fact that pure linearity has been rejected for the whole series.

**Figure 1.** Genome size $G$ (in Mb) *versus* the number of chromosomes $L_g$ (in $1n$) for all the major groups of organisms analyzed in [6] excluding plants, which were plotted separately (Figure 2) due to the high dispersion of angiosperms.

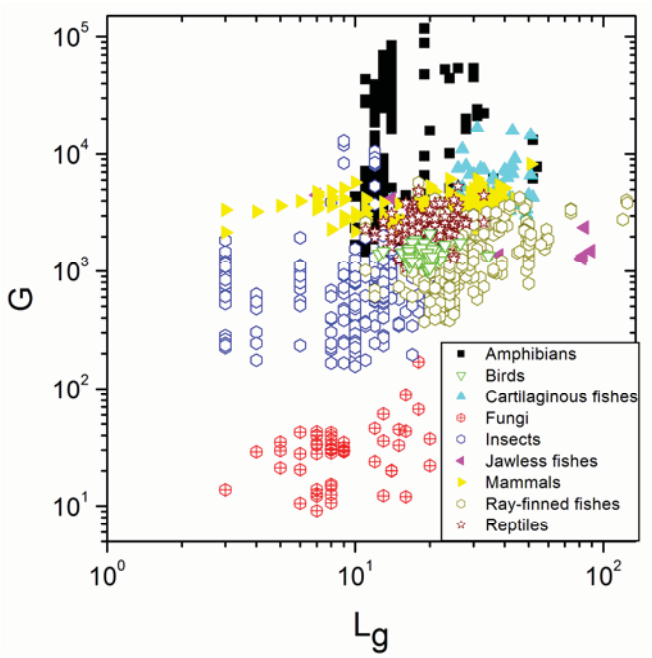



**Figure 2.** Genome size $G$ (in Mb) *versus* the number of chromosomes $L_g$ (in $1n$) for the major groups of plants analyzed in [6].

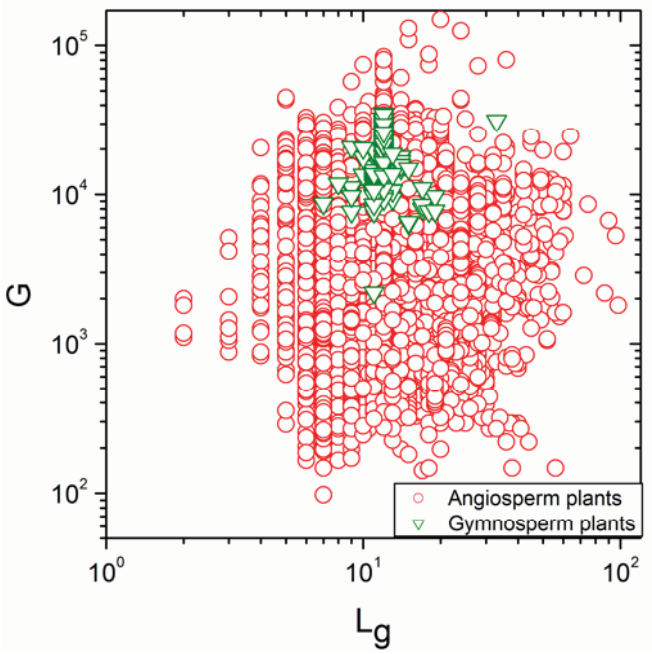


**Table 1.** Summary of the correlation analysis between genome size $G$ (in Mb) and genome size $L_g$ in number of chromosomes. $N$, $\rho$, and $p$ are defined, respectively, as the number of different organisms, the value of Spearman's rank correlation statistic for $G$ *versus* $L_g$, and the $p$-value of $\rho$ within a group of organisms. The values of $\rho$ were rounded to leave only three decimals and the p-values were rounded to leave only one significant digit.

| **Group** | ***N*** | ***ρ*** | ***p*** |
|---|---|---|---|
| Fungi | 56 | 0.280 | 0.04 |
| Angiosperm plants | 4706 | −0.38 | 0.008 |
| Gymnosperm plants | 170 | 0.315 | $3 \times 10^{-5}$ |
| Insects | 269 | 0.220 | 0.0003 |
| Reptiles | 170 | 0.243 | 0.001 |
| Birds | 99 | 0.008 | 0.9 |
| Mammals | 371 | 0.297 | $5 \times 10^{-9}$ |
| Cartilaginous fishes | 52 | −0.129 | 0.4 |
| Jawless fishes | 13 | −0.744 | 0.004 |
| Ray-finned fishes | 647 | 0.487 | $<10^{-17}$ |
| Amphibians | 315 | 0.446 | $9 \times 10^{-17}$ |

**Table 2.** Summary of the correlation analysis between genome size $G$ (in million base pairs) and $a = (G - c)/L_g$, where $L_g$ is the genome size in number of chromosomes and $c$ is the intercept of a linear approximation of the dependency between $G$ and $L_g$ by a non-parametric linear regression method. $N$, $\rho$, and $p$ are defined, respectively, as the number of different organisms, the value of Spearman's rank correlation statistic for $G$ *versus* $a$, and the $p$-value of $\rho$ within a group of organisms. The values of $\rho$ were rounded to leave only three decimals and the $p$-values were rounded to leave only one significant digit.

| **Group** | ***N*** | ***ρ*** | ***p*** |
|---|---|---|---|
| Fungi | 56 | 0.666 | $2 \times 10^{-8}$ |
| Angiosperm plants | 4706 | 0.925 | $<10^{-17}$ |
| Gymnosperm plants | 170 | 0.992 | $<10^{-17}$ |
| Insects | 269 | 0.802 | $<10^{-17}$ |
| Reptiles | 170 | 0.791 | $<10^{-17}$ |
| Birds | 99 | 0.771 | $<10^{-17}$ |
| Mammals | 371 | 0.278 | $5 \times 10^{-8}$ |
| Cartilaginous fishes | 52 | 0.886 | $<10^{-17}$ |
| Jawless fishes | 13 | 0.951 | $<10^{-17}$ |
| Ray-finned fishes | 647 | 0.812930 | $<10^{-17}$ |
| Amphibians | 315 | 0.983 | $<10^{-17}$ |

## 3. Discussion

According to Table 1, the dependencies between $G$ and $L_g$ can be classified into three qualitative types:

- "The more parts, the larger the whole"

This is the case of fungi, gymnosperm plants, insects, reptiles, mammals, ray-finned fishes and amphibians.

- "The more parts, the smaller the whole"

This is only the case of angiosperm plants and jawless fishes. A negative correlation between genome size and number of chromosomes in angiosperm plants has previously been reported [1].

- "Other"

Birds and cartilaginous fishes fall into this category, which includes the possibility that the number of parts and the size of the whole are independent. However, independence is not necessarily the only explanation (recall that absence of correlation does not imply independence [7]). We just mention a couple of possibilities. First, the dependency is not monotonic (rank correlation tests of the kind that we have used are more appropriate for strictly monotonically increasing or decreasing functional dependencies). Second, the dataset is not large enough to allow one to unravel the underlying trend for that particular group since only a very small fraction of all the species that actually belong to the groups has been explored (e.g., Table 1.1 of [8]). In sum, absences of correlations are not the rule but the exception in these major groups.

The class "The more parts, the larger the whole" could have simple explanations if $G$ was an increasing linear function of $L_g$, *i.e.*, $G = aL_g + c$ with $a > 0$. First, imagine that all chromosomes are of about the same size $a$ (and that $a$ does not depend on the number of chromosomes). Then genomes size $G$ would be proportional to $L_g$, *i.e.*, $G = aL_g$. Second, consider the case of genome duplication. Imagine that a new species is produced by adding $k$ copies copy of the genome of an origin species (with $k = 1$ for genome duplication). The genomes that would be generated by this mechanism would satisfy the relationship $G = aL_g$, where $a = G^0/L_g^0$ would be the ratio between $G^0$ and $Lg^0$, respectively, the genome size and the chromosome number of the origin species. Here it has been shown that a linear relationship between G and $L_g$ is not supported for any group. In sum, a purely increasing linear dependency between $G$ and $L_g$ is not supported for any group in our dataset. This has an important biological implication: Simple genome duplication is unlikely to be the only force shaping the class of organisms where "the more parts, the larger the whole".

We have presented a classification into three classes of growth of the whole with regard to its parts at a given taxonomic scale of analysis which does not need to be preserved at lower taxonomic scales. For instance, although angiosperm plants fall into the class "the more parts, the larger the whole", at the level of families, only seven families show this behavior, 22 families show the opposite pattern ("the more parts, the smaller the whole") but an overwhelming number of families, *i.e.*, 194, show no significant part-whole correlation (see the Appendix B for further information on group subdivision). This and other results discussed in the Appendix B mean that these three classes must be interpreted as only valid a priori at their taxonomic scale. The Appendix B also shows that subdividing does not help to unravel a trend in the only two groups where no correlations were found: Birds and cartilaginous fishes.

Our empirical analysis has implications for the debate about the relevance of a connection between human language and genomes through a common pattern: the tendency of the mean size of the parts (syllables or chromosomes) to decrease as the number of parts of the whole (a word or a genome) increases [6]. This pattern is known as Menzerath-Altmann law in quantitative linguistics [9] and is

found not only in language at many levels of description but also in music (see [10] and references therein). According to [5], the finding of this negative correlation between the mean size of the parts and the number of parts in genomes is a trivial consequence of the definition of the size of the parts, $L_c$ as a mean, *i.e.*, $L_c = G/L_g$, which leads to $L_c = a/L_g$ where $a$ is a constant. However, $L_c = a/L_g$ holds if and only if $G$ is a constant function of $L_g$. In other words, the relationship between the mean size of the parts and the number of parts is trivial if and only if $G$ is constant. In contrast, here it has been shown that $G$ and $L_g$ are significantly correlated in many groups of organisms. The classes "The more parts, the larger the whole" and the classes "The more parts, the smaller the whole" violate the constancy assumption of [5]. Furthermore, it has been shown that, when such significant correlation is not found, the possibility that this is due to the small size of the group sample cannot be denied. Notice that [5] evaluates the goodness of the fit of $L_c = a/L_g$ to actual data with a flawed test, which consists of fitting $L_c = a/L_g^b$ to actual data. If $b = -1$ is obtained this implies that the hypothesis $L_c = a/L_g$ is correct, according to [5]. However, obtaining $b = -1$ from data is a necessary but not a sufficient condition for $L_c = a/L_g$. In contrast, here we have investigated a sufficient condition for $L_c \neq a/L_g$: if $G$ is not a constant function of $L_g$ then $L_c = a/L_g$ cannot be true, at least in some region.

Similarly, our findings unravel the weakness of a random breakage model of chromosome lengths that has been proposed recently [5]. In this model, the information about a certain organism is generated in the following way:

- $G$ is chosen uniformly at random within the interval $(G^m, G^M)$.
- $L_g$ (the number of chromosomes of the organism) is chosen uniformly at random within the interval $(L_g^m, L_g^M)$.
- Chromosome lengths are produced from $G$ and $L_g$ following a random breakage procedure [11,12].

Interestingly, $G$ and $L_g$ are chosen independently in this model. Such independence is totally unrealistic as our analyses and previous research [1] have revealed. Notice that the independence between $G$ and $L_g$ needs (if genomes with chromosomes of length zero are considered as not allowed or totally unrealistic) that the condition $L_g^M \leq G^m + 2$ is satisfied so that all chromosomes can have length greater or equal than one. This condition follows from $L_g \leq L_g^M - 1$, $G^m + 1 \leq G$ and the condition for non-empty chromosomes, *i.e.*, $L_g \leq G$.

Our study is just one among many evidences of the "multiscale link between genome-level and chromosome-level" that the random breakage model above and accompanying arguments deny [5]. Laboratory experiments indicate that "upper and lower tolerance limits for chromosome size are apparently determined by the genome size, chromosome number and karyotype structure of a given species" (see [13] and references therein). Along these lines, a recent statistical study shows that it is possible to predict, for a given species, chromosome sizes by chromosome number, and furthermore, given either genome size or average chromosome length it is possible to predict the size range of all chromosomes of that species [14].

Future work should address the question of the precise mathematical form of the dependency between chromosome number and genome size. By having shown its statistical significance and excluded that it is trivially linear for all groups, the foundations for further research have been established and the actual scope of multiscale links between the genome and the chromosome level has

been clarified. Our selection of groups of organisms was motivated by [5,6] but the same analysis should be extended to other groups of organisms in the future.

## 4. Methods

### *4.1. Data*

For consistency with [6], the same major groups of organisms (listed on Table 1) were used. The information about each organism was retrieved in June 2011 from the same databases of [6]. The same methods of [6] for filtering incorrect data were applied.

### *4.2. A Test of Pure Linearity between G and $L_g$*

$G$ is a purely linear function of $L_g$, if and only if $G = aL_g + c$, where $a$ and $c$ are constants. If $G$ was a purely linear function of $L_g$, one would have that $a = (G - c)/L_g$ is a constant function of $G$ with $c$ obtained from least squares linear regression. A two-sided Spearman rank correlation test was used to determine if there is a correlation between $(G - c)/L_g$ and $G$. Notice that here the term 'pure' or 'purely' is not used to mean that the relationship between $G$ and $L_g$ is deterministically linear but to mean indeed that $E[G|L_g]$, the expectation of $G$ given $L_g$ is exactly linear, *i.e.*, $E[G|L_g] = aL_g + c$. The general assumption of regression (and also ours) is that $G = E[G|L_g] + \varepsilon$, where $\varepsilon$ is an error that is typically assumed to be normally distributed with mean zero and constant standard deviation [15]. However, a non-parametric linear regression method, Theil's incomplete method [16], was used to estimate $a$. This method has the following advantages over a simple parametric least squares linear regression [16]:

- It does not assume that all the errors are only in the $y$-direction.
- It does not assume that either the $x$- or $y$-direction errors are normally distributed.
- It is robust in the sense that it is not affected by the presence of outliers.

## Acknowledgments

We are grateful to J. Perarnau and X. Messeguer for helpful discussions. This work was supported by the project SESAAME-BAR (TIN2008-06582-C03-01) of the Spanish Ministry of Science and Innovation.

## References and Notes

1. Vinogradov, A.E. Mirrored genome size distributions in monocot and dicot plants. *Acta Biotheoretica* **2001**, *49*, 43–51.
2. Trivers, R.; Burt, A; Palestis, B.G. B chromosomes and genome size in flowering plants. *Genome* **2004**, *47*, 1–8.
3. Sankoff, D.; Ferretti, V. Karyotype distributions in a stochastic model of reciprocal translocation. *Genome Res.* **1996**, *6*, 1–9.

4. De, A.; Ferguson, M.; Sindi, S.; Durrett, R. The equilibrium distribution for a generalized Sankoff-Ferretti model accurately predicts chromosome size distribution in a wide variety of species. *J. Appl. Probab.* **2001**, *38*, 324–334.
5. Solé, R.V. Genome size, self-organization and DNA's dark matter. *Complexity* **2010**, *16*, 20–23.
6. Ferrer-i-Cancho, R.; Forns, N. The self-organization of genomes. *Complexity* **2009**, *15*, 34–36.
7. DeGroot, M.H. *Probability and Statistics*, 2nd ed.; Addison-Wesley: Reading, MA, USA, 1989; p. 215.
8. Gregory, T.R. Genome size evolution in animals. In *The Evolution of the Genome*; Gregory, T.R., Ed.; Elsevier: San Diego, CA, USA, 2005; pp. 4–71.
9. Altmann, G. Prolegomena to Menzerath's law. *Glottometrika* **1980**, *2*, 1–10.
10. Boroda, M.G.; Altmann, G. Menzerath's law in musical texts. *Musikometrika* **1991**, *3*, 1–13.
11. Fuquan, K.; Kui, Z.; Yong, Z.; Tianguang, C.; Meinan, N.; Li, S.; Minghui, C.; Yizhong, Z. Analysis of length distribution of short DNA fragments induced by $^{7}$Li ions using the random-breakage model. *Chin. Sci. Bull.* **2005**, *50*, 841–844.
12. Becker, T.S.; Lenhard, B. The random *versus* fragile breakage models of chromosome evolution: A matter of resolution. *Mol. Genet. Genomics* **2007**, *278*, 487–491.
13. Schubert, I. Chromosome evolution. *Curr. Opin. Plant Biol.* **2007**, *10*, 109–115.
14. Li, X.; Zhu, C.; Lin, Z.; Wu, Y.; Zhang, D.; Bai, G.; Song, W.; Ma, J.; Muehlbauer, G.J.; Scalon, M.J.; *et al.* Chromosome size in diploid eukaryotic species centers on the average length with a conserved boundary. *Mol. Biol. Evol.* **2011**, doi:10.1093/nar/gkl828.
15. Ritz, C.; Streibig, J.C. *Nonlinear Regression with R*; Springer: New York, NY, USA, 2008.
16. Miller, J.C; Miller, J.N. *Statistics for Analytical Chemistry*, 3rd ed.; Prentice Hall: London, UK, 1993; pp. 159–161.
17. Pearl, J. *Causality: Models, Reasoning, and Inference*; Cambridge University Press: Cambridge, UK, 2000.

## Appendix A

The relationship between genome size and chromosome number is shown in Figure 3 for fungi, Figure 4 for angiosperm plants, Figure 5 for gymnosperm plants, Figure 6 for insects, Figure 7 for reptiles, Figure 8 for birds, Figure 9 for mammals, Figure 10 for cartilaginous fishes, Figure 11 for jawless fishes, Figure 12 for ray-finned fishes and Figure 13 for amphibians.

**Figure 3.** Genome size $G$ (in Mb) *versus* the number of chromosomes $L_g$ (in 1$n$) for fungi.

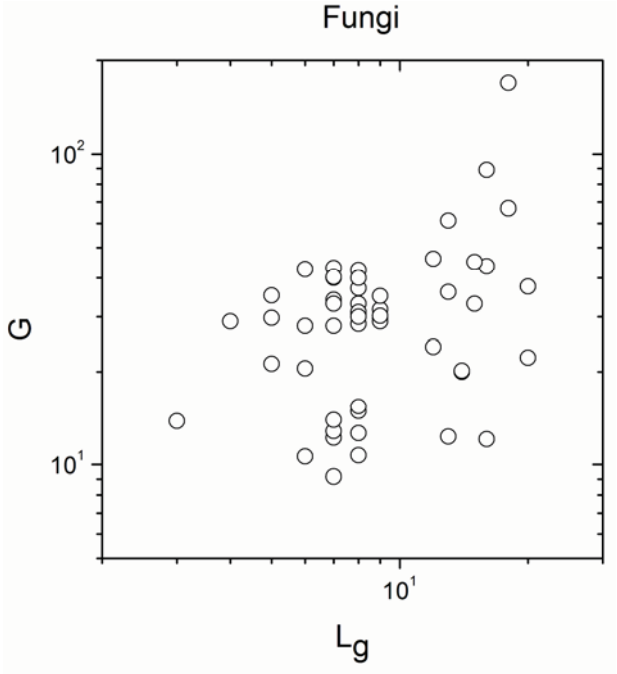



**Figure 4.** Genome size $G$ (in Mb) *versus* the number of chromosomes $L_g$ (in 1$n$) for angiosperm plants.

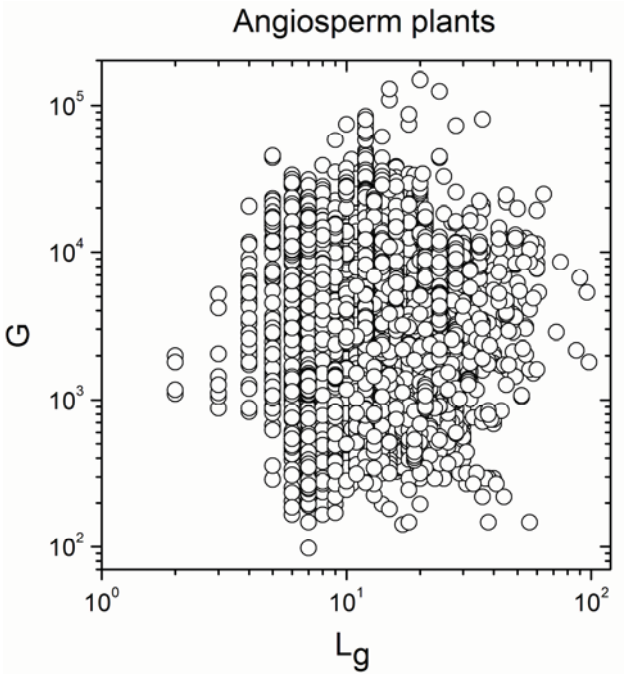


**Figure 5.** Genome size $G$ (in Mb) *versus* the number of chromosomes $L_g$ (in 1$n$) for gymnosperm plants.

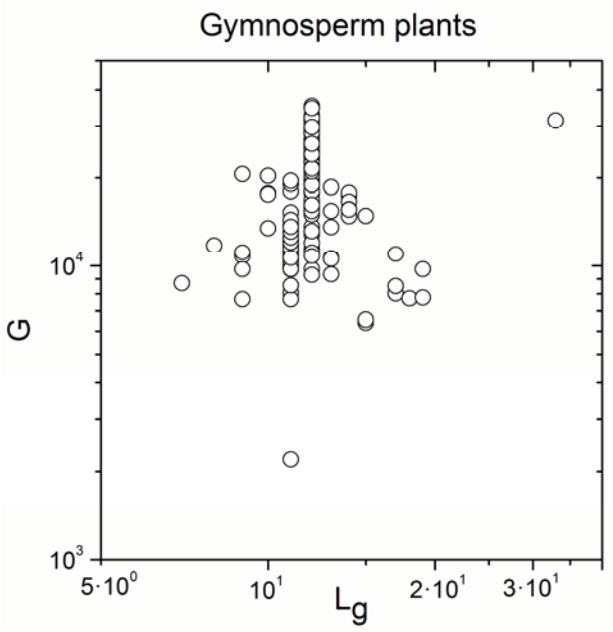



**Figure 6.** Genome size $G$ (in Mb) *versus* the number of chromosomes $L_g$ (in 1$n$) for insects.

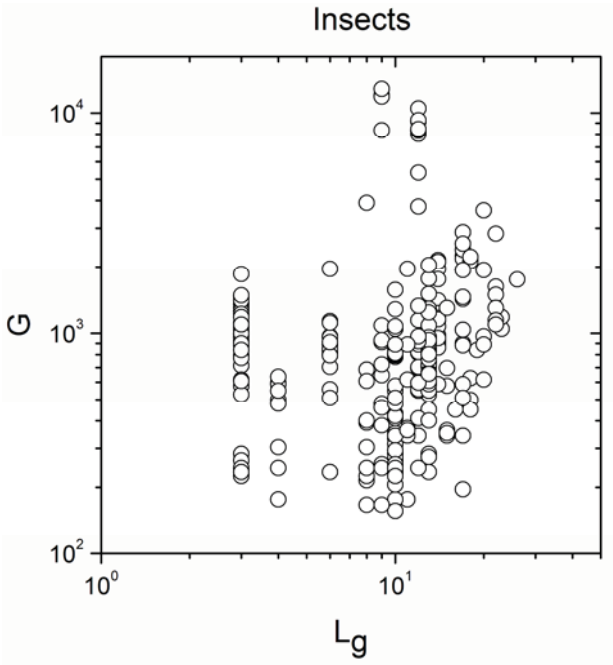


**Figure 7.** Genome size *G* (in Mb) *versus* the number of chromosomes $L_g$ (in 1*n*) for reptiles.

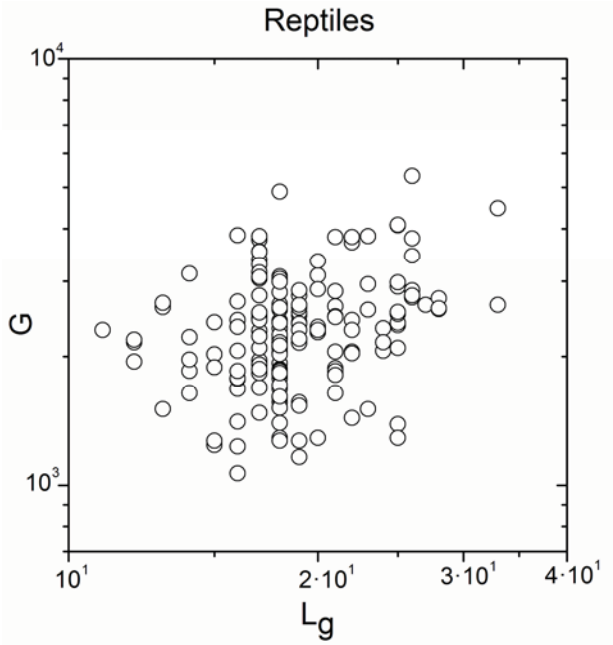



**Figure 8.** Genome size *G* (in Mb) *versus* the number of chromosomes $L_g$ (in 1*n*) for birds.

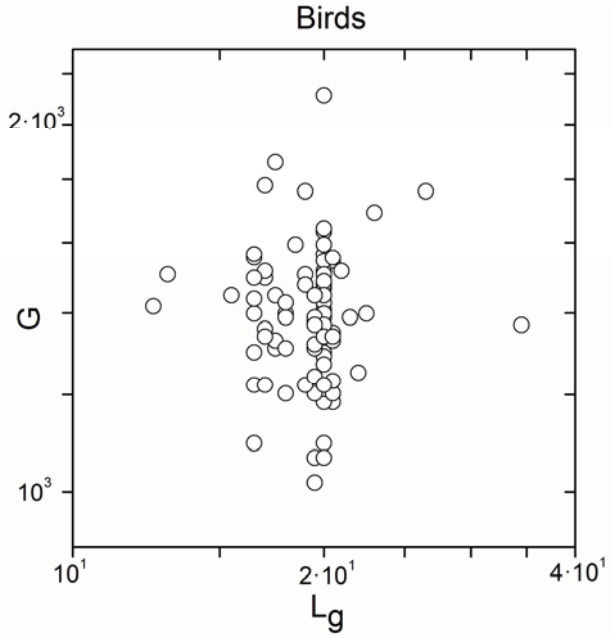


**Figure 9.** Genome size $G$ (in Mb) *versus* the number of chromosomes $L_g$ (in 1$n$) for mammals.

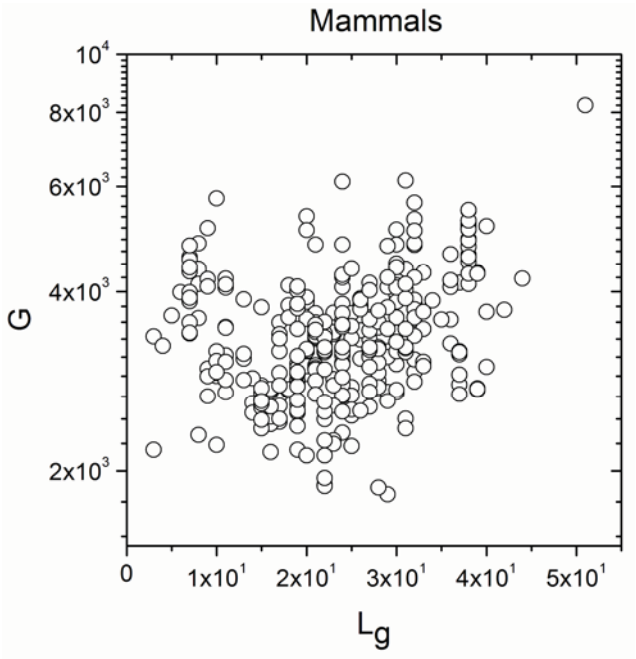



**Figure 10.** Genome size $G$ (in Mb) *versus* the number of chromosomes $L_g$ (in 1$n$) for cartilaginous fishes.

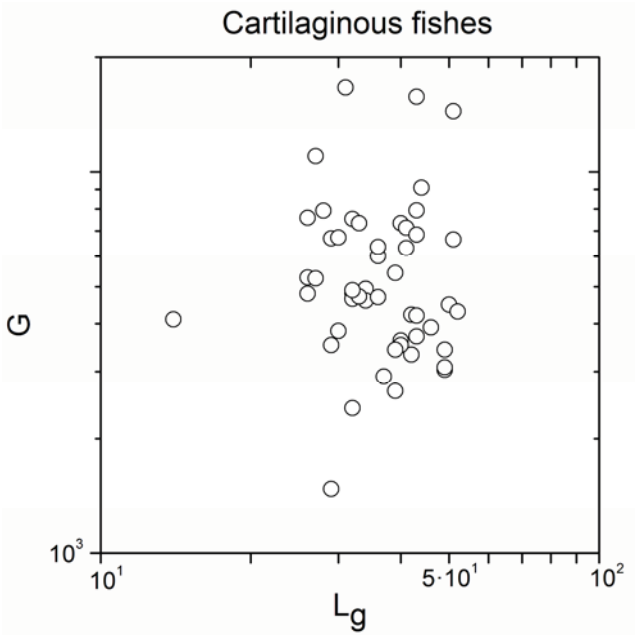


**Figure 11.** Genome size $G$ (in Mb) *versus* the number of chromosomes $L_g$ (in 1$n$) for jawless fishes.

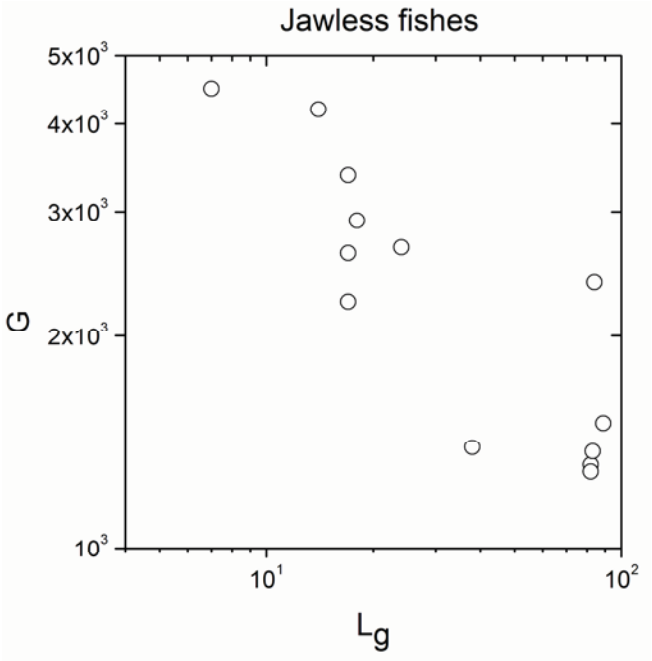



**Figure 12.** Genome size $G$ (in Mb) *versus* the number of chromosomes $L_g$ (in 1$n$) for ray-finned fishes.

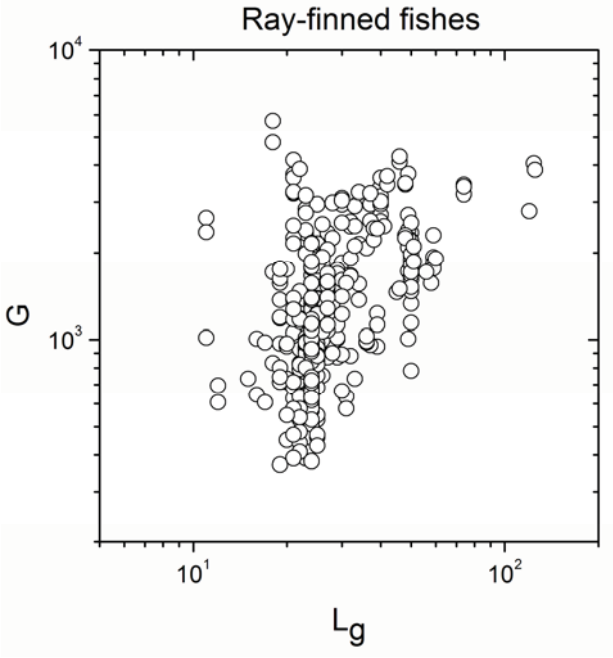
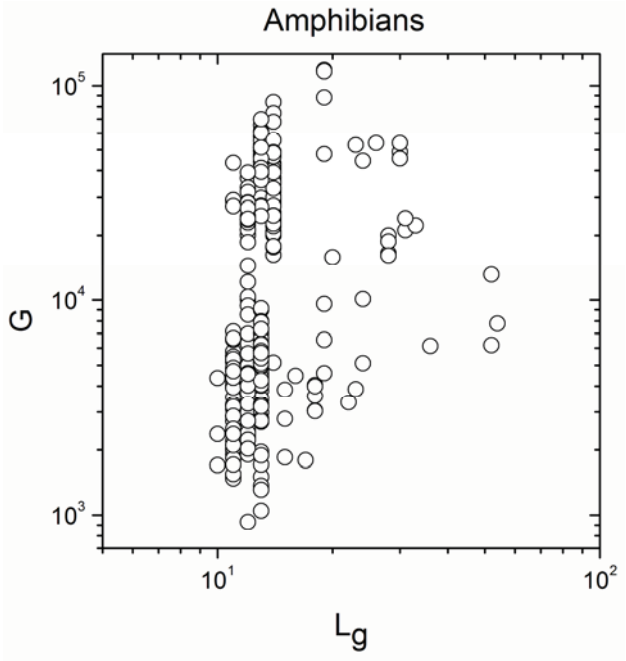


**Figure 13.** Genome size $G$ (in Mb) *versus* the number of chromosomes $L_g$ (in 1$n$) for amphibians.




## Appendix B

Simpson's paradox [7,17] suggests that the conclusions about the correlations between $G$ and $L_g$ for a certain groups of organisms (Table 1) could change when these groups are subdivided using taxonomic information. Subdividing could yield paradoxical results such as (a) that a group of organisms shows no significant dependency but its subgroups do show a significant correlation or the opposite, that the significant correlation of the group is lost in the subgroups [7] or (b) that the sign of the significant correlation of the original group is the opposite of that of its subgroups [17].

When attempting to study how that correlation changes when taxonomic subgroups are considered, various serious problems were encountered. First, the necessary taxonomic information is not available for all species in public genome size databases. This is especially worrying for fungi, where the amount of missing information is massive. Second, due to the very limited coverage of the genome size databases, taxonomic subdivisions may contain only one subgroup or a few unless the taxonomic subgroup is low enough. Thirdly, at low taxonomic levels, subgroups turn out to have so little members that no significant correlations can be detected in the majority of them. The few significant correlations may not be representative of that scale of analysis due to the very limited coverage of genome size databases. Table 3 summarizes the results of the analysis of the dependency between the size of whole and the size of the parts at lower taxonomic levels within each original group. For simplicity, for each taxonomic sublevel, only those sublevels for which the group yielded more than one subgroup are considered.

**Table 3**. Summary of the correlations between genome size ($G$) and chromosome number ($L_g$) at different taxonomic levels. Boldface is used to indicate the taxonomic groups that are the target of our main analysis. +, −, ? are attached to the name of each target group to indicate, respectively, that the correlation between $G$ and $L_g$ was significant and positive, significant and negative, and none of them (at a significance level of 0.05). Below each target group of organisms, the total number of organisms in our dataset is shown. In each cell for which taxonomic data is available, a triple of numbers is shown above and a pair of numbers is shown below. The triple follows the format $x$,$y$,$z$, where $x$, $y$ are respectively, the number of subgroups with significant positive and significant negative correlations, and $z$ is the total number of subgroups. The pair follows the format $x'$,$y'$, where $x'$ and $y'$ are the number of organisms involved in significant positive and significant negative correlations, respectively.

| **Kingdom** | **Phylum/Division** | **Class** | **Order** | **Family** | **Genus** |
|---|---|---|---|---|---|
| **Fungi** + 56 | 0,3,5<br>0,55 | | 0,4,5<br>0,34 | | 0,1,40<br>0,5 |
| Plants | **Angiosperm** − 4706 | | | 22,7,194<br>2374,965 | 66,8,1114<br>1608,186 |
| | **Gymnosperm** + 170 | | | 0,4,14<br>0,122 | 0,2,52<br>0,13 |
| Animals | Arthropoda | **Insects** + 269 | 3,1,7<br>189,56 | 0,1,26<br>0,13 | |
| | Chordata | **Reptiles** + 170 | 0,0,4<br>0,0 | 1,1,34<br>14,18 | |
| | | **Birds ?** 99 | 0,0,17<br>0,0 | 0,0,33<br>0,0 | |
| | | **Mammals** + 371 | 2,1,17<br>162,54 | 5,0,63<br>89,0 | |
| | | **Cartilaginous fishes ?** 52 | 0,1,9<br>0,24 | 1,0,20<br>7,0 | |
| | | **Jawless fishes** − 13 | 0,0,2<br>0,0 | 0,0,2<br>0,0 | 0,0,2<br>0,0 |
| | | **Ray finned fishes** + 647 | 4,0,30<br>262,0 | 3,0,115<br>214,0 | |
| | | **Amphibians** + 315 | 1,0,3<br>185,0 | 3,1,26<br>42,72 | |

To scrutinize the results of Table 3, we consider two definitions of Simpson's paradox: (a) the reversing of the sign of significant correlation between $G$ and $L_g$ when splitting a group into subgroups (b) the emergence or the loss of significant correlations between $G$ and $L_g$ when splitting a group into subgroups. Table 3 shows that, after splitting,

- The sign of the significant correlations was totally reversed, in full agreement with definition (a) of Simpson's paradox, only in fungi and gymnosperm plants.
- The sign of the significant correlation was totally maintained only in ray-finned fishes.

- The significant correlation was lost in jawless fishes, in agreement with definition (b) of the paradox.
- Significant correlations became a mixture of positive and negative correlations in angiosperm plants, insects, reptiles, mammals and amphibians.
- Non-significant correlations remained totally for birds.
- Significant correlations emerged only exceptionally in cartilaginous fishes (the number of significant correlations was very small with regard to the total number of subgroups), consistently with definition (b) of the paradox, but the sign of the correlation was not coherent.

This suggests that, with the currently available data, Simpson's paradox is only supported in some groups: Fungi, gymnosperm plants, jawless fishes and cartilaginous fishes. The limited coverage of genome sizes databases cannot exclude that the paradox appears in more groups when more organisms are added but also, the opposite effect could be found, namely, that the paradox disappears when more species are included.