Size distribution of function-based human gene sets and the split–merge model

Wentian Li¹, Oscar Fontanelli² and Pedro Miramontes²,³

¹The Robert S. Boas Center for Genomics and Human Genetics, The Feinstein Institute for Medical Research, Northwell Health, Manhasset, NY, USA
²Departamento de Matemáticas, Facultad de Ciencias, Universidad Nacional Autónoma de México, Circuito Exterior, Ciudad Universitaria, México 04510 DF, México
³Bioinformatics Group and Interdisciplinary Center for Bioinformatics, University of Leipzig, Haertelstrasse 16–18, 04107 Leipzig, Germany

The sizes of paralogues—gene families produced by ancestral duplication—are known to follow a power-law distribution. We examine the size distribution of gene sets or gene families where genes are grouped by a similar function or share a common property. The size distribution of Human Gene Nomenclature Committee (HGNC) gene sets deviate from the power-law, and can be fitted much better by a beta rank function. We propose a simple mechanism to break a power-law size distribution by a combination of splitting and merging operations. The largest gene sets are split into two to account for the subfunctional categories, and a small proportion of other gene sets are merged into larger sets as new common themes might be realized. These operations are not uncommon for a curator of gene sets. A simulation shows that iteration of these operations changes the size distribution of Ensembl paralogues and could lead to a distribution fitted by a rank beta function. We further illustrate application of beta rank function by the example of distribution of transcription factors and drug target genes among HGNC gene families.

1. Introduction

A narrowly defined gene family within a genome [1,2] refers to all working genes, thus not pseudo-genes, that were produced by duplication events [3,4]. Genes in such a gene family share sequence similarity owing to their common origin, and are called paralogues [5,6]. The role of such a defined gene family in functional biology, however, is less clear as paralogues may diverge in functions [7–9]. Alternatively, a group of genes can be
connected in a variety of other ways, such as similar expression patterns [10], participating in the same biology pathway, carrying out similar biological processes, contributing to the same phenotype, chemical interaction, biological function and phenotype or medical condition [14].

To clearly distinguish gene families created by gene duplication and those by grouping using any other possible criteria, here we use the term ‘paralogues’ and ‘gene sets’ to refer to the two situations, and ‘gene family’ when such a distinction is not important. The size of paralogues in a genome can be quite uneven, with some large ones but more small ones. If $S$ denotes the number of genes (size) in a paralogue, and $f(S)$ the number of paralogues with size $S$, $f(S)$ as a function of $S$ has a power-law-like trend with negative exponents [15]. This statistical trend for paralogue sizes has been confirmed by other publications [16,17]. Similar trends have been extended to the protein universe for folds, domains, clusters, families and superfamilies [18–23].

Many theoretical and mathematical models have been proposed to explain the observed paralogue size distribution, with gene duplication [3] as the essential mechanism [24–34]. Meanwhile, biological details on paralogue evolution have been accumulated [35–44], making a realistic modelling possible. The robust and ubiquitous power-law gene and protein family size distribution is hailed as one of the laws of genome evolution [45,46].

There are three goals in this paper: (i) although the long-tailed size distribution of paralogues is well established with sufficient understanding, most analyses, with some exceptions [47], were focused on bacterial genomes. Here, we examine the human paralogues in more detail. (ii) For gene families not constructed by homology, which we call gene sets, we show that power-law rank function is not a good fitting model. A more flexible function called beta rank function can be used [48,49]. This greatly expands our toolbox in a quantitative analysis of gene families. (iii) Just as duplication operation is essential for the observed power-law distribution of paralogue sizes, we examine the effect of a combined operation of splitting and merging of gene sets. We show that these operations lead to a deviation from the power-law distribution, thus justifying the introduction of new types of fitting function such as the beta rank function.

2. Methods and data

2.1. Homology-based human gene family (paralogues) data

We use the Ensembl family (in human) which has a stable ID beginning with ENSFM [50]. The number of genes within a family (paralogue) is counted by the number of distinct Ensembl Gene IDs beginning with ENSG that share the same family ID. The data are obtained from the Ensembl Biomart (www.ensembl.org/biomart) [51] by selecting the following boxes: (i) Ensembl genes 82 in ‘choose database’; (ii) Homo sapiens genes (GRCh38.p3) in ‘choose dataset’; (iii) filtering ‘region (chromosome)’: 1–22, X, Y; (iv) filtering ‘gene (gene type)’: protein_coding and (v) filtering ‘gene (multi species comparisons)’: paralogous human genes.

2.2. Biological function-based human gene sets data

We use the Human Gene Nomenclature Committee (HGNC) gene families (www.genenames.org/cgi-bin/genefamilies/) [52,53]. The broad guideline for gene families (which we call gene sets here) is that they ‘signify groups of genes related by function, by sequence or by phenotype caused’ [54]. Although the hierarchical organization of gene sets is advocated [54], many high-level concepts for gene (super)families do not have an HGNC family ID. The number of genes in a gene set is counted by the number of gene names with the same gene family name or ID. Non-coding genes, RNAs, microRNAs and genes without an approved name are not counted. We examine the effect of including or excluding pseudo-genes.

2.3. Inverse power-law rank function and Zipf’s law

Suppose there are $n_F$ gene families (paralogues or gene sets); these gene families are ranked from large (with many genes in a family) to small (few genes in a family), with the rank $r = 1, 2, \ldots n_F$. The gene family size $S_r$ as a function of rank $r$ can be called a rank-size distribution. The Zipf’s law or Pareto’s law uses an inverse power-law function: $\log(S_r) = c - a \log(r)$ to fit the data, in log-log scale [55,56].
In the linear regression framework, the relation can be written as \( \log(S_r) \sim \log(r) \) with the purpose of minimizing the error between expected (from the regression line) and observed in \( \log(S_r) \) scale. Without the log–log scale, that fitting \( S_r = C/r^a \) is to minimize the error between expected and observed in the original \( S_r \) scale. The two regressions may result in different fitting parameters. It is well known that the power-law Zipf’s law \( S_r \propto 1/r^{a+1} \) corresponds to a power-law probability density function \( f(S) \propto 1/S^{1+a+1} \) [57], as the two are connected by a switching of \( x \) and \( y \) axes, and a derivative [58].

2.4. Beta rank function

The beta rank function [48,49,58–60] is the following functional form for ranked data:

\[
S_r = C \frac{(n_F + 1 - r)^b}{r^a}
\]

with two free parameters (\( a \) and \( b \), where \( c \) is constrained by the normalization condition). In the regression framework, there are again two versions. One is a multiple linear regression in log–log scale: \( \log(S_r) = c + b \log(n_F + 1 - r) - a \log(r) \) or \( \log(S_r) \sim \log(n_F + 1 - r) + \log(r) \). The second is a nonlinear regression \( S_r \sim (n_F + 1 - r)^b/r^a \). The multiple linear regression aims at minimizing predicted–observed error in the \( \log(S_r) \) scale, whereas the nonlinear regression minimizes that in the \( S_r \) scale. The beta rank function becomes a power-law rank function when \( b = 0 \).

2.5. Measuring fitting performance of a rank function by mean-squared error and Pearson’s correlation

The fitting performance of a function/regression can be measured in different ways. We use two measures here: (i) mean-squared error (MSE):

\[
MSE = \frac{1}{n_F} \sum_{r=1}^{n_F} (O_r - E_r)^2,
\]

where ‘observed’ \( O_r \) can either be \( S_r \) or \( \log(S_r) \) and ‘expected’ \( E_r \) can either be regression-predicted \( S_r \) or \( \log(S_r) \). The quantity without the normalization factor \( (1/n_F) \) is called sum of squared errors (SSE), or residual sum of squares (RSS), or sum of squared residuals (SSR). (ii) Pearson’s correlation coefficient between \( O_r \) and \( E_r \),

\[
r_{OE} = \frac{\sum_{r=1}^{n_F} (E_r - \bar{E}_r)(O_r - \bar{O}_r)}{\sqrt{\sum_{r=1}^{n_F} (E_r - \bar{E}_r)^2 \sum_{r=1}^{n_F} (O_r - \bar{O}_r)^2}}
\]

again in either the \( S_r \) scale or \( \log(S_r) \) scale. The coefficient of determination \( R^2 \), used in, e.g. Martinez-Mekler et al. [49], is the square of \( r_{OE} \) [61, p. 29]:

\[
R^2 = r_{OE}^2 = \frac{\sum_{r=1}^{n_F} (E_r - \bar{E}_r)(O_r - \bar{O}_r)^2}{\sum_{r=1}^{n_F} (E_r - \bar{E}_r)^2 \sum_{r=1}^{n_F} (O_r - \bar{O}_r)^2}.
\]

(iii) Kolmogorov–Smirnov (KS) distance between the cumulative distribution function (cdf) from observed data and cdf from the fitted values. KS distance in cdf is defined as the maximum distance between the two cdfs in the cumulative probability axis. That axis is equivalent to the rank axis in a rank–frequency plot, normalized by the maximum rank.

2.6. Comparing the fitting performance of two functions with different number of fitting parameters by Akaike information criterion and Bayesian information criterion

This part is essential identical to page 185 of Venables & Ripley [62] and the appendix of Li & Miramontes [59]. If we assume the difference between observed and fitted values follow a normal distribution at all ranks, the likelihood of the fitting function is

\[
L = \prod_{i=1}^{n_F} \frac{e^{-(O_i - E_i)^2/2\sigma^2}}{\sqrt{2\pi \sigma^2}} = \frac{e^{-\sum_{i=1}^{n_F} (O_i - E_i)^2/2\sigma^2}}{(2\pi \sigma^2)^{n_F/2}},
\]

where the variance of the noise can be estimated from the data: \( \hat{\sigma}^2 = MSE \).
The Akaike information criterion (AIC) \[63\] is defined as: 

$$\text{AIC} = -2\log \hat{L} + 2p,$$

where $\hat{L}$ is maximized likelihood, $p$ is the number of parameters in the statistic model and Bayesian information criterion (BIC) \[64\] defined as 

$$\text{BIC} = -2\log \hat{L} + p \log n.$$

Replacing $\sigma$ by the estimated $\hat{\sigma}$, we obtained the maximized likelihood, which after log is

$$\log(\hat{L}) = C - \frac{n_F}{2} \log(\hat{\sigma}^2) = C - \frac{n_F}{2} \log(\text{MSE})$$  \hspace{1cm} (2.6)

then 

$$\text{AIC} = n_F \log(\text{MSE}) + 2 \cdot p + \text{const.}$$ \hspace{1cm} (2.7)

and 

$$\text{BIC} = n_F \log(\text{MSE}) + \log(n_F) \cdot p + \text{const.}$$ \hspace{1cm} (2.8)

The fitting function with the lowest AIC/BIC is a better function \[65\].

2.7. Measuring fitting performance of a rank function by empirical $p$-value through simulation

Following Clauset et al. \[57\], we compare the distance between the observed data and fitted function with the distances generated from re-samplings (or bootstraps \[66\]). The re-sampling uses the true function to generate data, and the produced distances are supposed to be noise. When the observed distance tends to be larger than those from the re-sampling, we may question the validity of the fitted function. The proportion of re-sampling distances larger than the observed one is called empirical $p$-value. Then, a smaller empirical $p$-value can be used to reject the fitted function being a good function for this data \[57\] (see also, e.g. the appendix of Moreno-Sánchez et al. \[67\]).

We carry out these steps: (i) fitting a ranked data by a ranking function (e.g. Zipf’s function or beta rank function equation (2.1)); (ii) use the fitted rank function to construct an empirical cdf, noting that rank plot can be directly converted to an empirical cumulative plot \[58\]. We may write $y = \text{cdf}(x)$ with $y \in (0, 1)$; (iii) randomly sampling $y$s from the uniform distribution in $(0, 1)$, and through the empirical cdf, obtain a simulated dataset which is generated by the fitted rank function ($x = \text{cdf}^{-1}(y)$). The $x$s are rounded off to integers, and integer $x = 1$ is discarded as we require gene families to have at least one member; (iv) the simulated values $x$s are ranked and fitted by a fitted function (same functional form as the one used in the observed data, e.g. Zipf’s law or beta rank function). We use the KS distances as the distance measure between the data and the fitted function. KS distance is the maximum difference between the observed and fitted points in the $y$-direction. The KS distances between the simulated set and the corresponding fitted function are calculated; (v) repeat steps ii–iv. The proportion of KS distances from re-sampling replicates larger than the KS distance from the original observed data is the empirical $p$-value.

2.8. Computer program used

We use the R statistical package (www.r-project.org), where the function $\text{lm}$ is used for linear regression, $\text{cor.test}$ for correlation coefficient, $\text{ks.test}$ for KS distance, $\text{approxfun}$ for empirical cdf, etc.

2.9. List of transcription factor genes

A list of 1987 transcription factor (TF) genes is obtained from the supplementary table 2 of Vaquerizas et al. \[68\], with the URL: http://www.nature.com/nrg/journal/v10/n4/extref/nrg2538-s3.txt.

2.10. List of drug target genes

A list of 2829 drug target genes is obtained from the IUPHAR/BPS (International Union of Basic and Clinical Pharmacology/ British Pharmacological Society) Guide to PHARMACOLOGY site: http://www.guidetopharmacology.org/DATA/targets_and_families.csv \[69\].
3. Results

3.1. Human parologue sizes are characterized by the power-law distribution reasonably well

By the procedure described in the Method section, 3586 gene families (paralogues) are produced from Ensembl. By definition, a gene family should contain at least two genes. There are 2168 (60.5%) two-gene families, 810 (22.6%) three-gene families and 318 (8.9%) four-gene families. These ‘nuclear families’ already make up more than 90% of all families, as shown in the cumulative distribution (figure 1a). The fall off of the number of size-$S$ gene families as a function of $S$, when shown in double logarithmic scales in figure 1b, is roughly a straight line, indicating that the histogram is a power-law function. Similarly, the histogram can be obtained over the log($S$), as in figure 1c, with $S$ in log scale, which is again a straight line.

The distribution of human paralogues can be viewed in the rank–frequency plot (figure 1d in log–log scale). Besides the largest gene family, ENSFM00250000000002, a zinc finger family with 225 genes, other points in figure 1d can be reasonably approximated by a power-law function (i.e. Zipf’s law). Note that the second, third and fourth largest gene families, ENSFM00730001521153 (36 genes), ENSFM00730001521252 (33 genes) and ENSFM00760001714593 (31 genes), are also zinc fingers.

While figure 1 shows qualitatively that the paralogues follow a power-law-like distribution, we would like to examine whether it is good quantitatively. The MSE of the three fitting lines in figure 1b–d is 0.107, 0.255, 0.010; and the $R^2$ and 95% CIs are 0.973 (0.922–0.991), 0.943 (0.804–0.984), 0.936 (0.931–0.940). All these measures are calculated in the scale in where the regression was carried out (i.e. log–log scale in figure 1b,d, linear–log in figure 1c). If the largest paralogue ENSFM00250000000002 is included in a fitting, then all fitting performances go down: MSEs are 1.08, 0.79, 0.011, $R^2$s and 95% confidence intervals are 0.767 (0.552–0.888), 0.843 (0.992–0.952), 0.933 (0.929–0.937).

The fitting of points with non-zero values in figure 1b leads to an estimation of the exponent $\alpha$ in $\log(p) \sim -\alpha \log(S)$ to be 3.1. The estimation of $\alpha$ from figure 1d leads to $1/0.389 + 1 \approx 3.6$. The difference between the two estimations is perhaps due to the slightly different scalings at the two ends: in figure 1b, smaller paralogues contribute more to the fitting, whereas in figure 1d, individual large paralogues contribute more to the parameter fitting.

---

**Figure 1.** Distributions of human paralogue sizes from Ensembl. (a) Cumulative distribution. (b) Histogram (in log–log scale). (c) Histogram of log-transformed gene family sizes (y-axis in log scale). (d) Rank–frequency distribution (ranked individual paralogue sizes). A possible error bar is drawn in grey colour.
We use the re-sampling approach described in the Method section to assess the fitting of Zipf’s law. Of the 10 000 replicates, 9692 have larger KS distances with their own fitted function than the KS distance from the observed data, and 46 have equal KS distances. This leads to an empirical p-value of 0.97, indicating that Zipf’s law is a very good fitting function of the observed data. Another indirect confirmation is by drawing an error bar ($\pm \sqrt{\hat{y}}$) [70] for the data point in figure 1d. The shaded area enveloped by error bars maintains the straight line trend.

3.2. Curated human gene family sizes are better characterized by the beta rank function

We compile a list of 749 gene families from HGNC, with 717 contain at least two genes. ‘Not approved’ genes and RNAs (including microRNA, tRNA, piRNA, etc.) are excluded. A gene might be present in multiple families, either because the gene can have multiple functions or a gene can be viewed from different perspectives in the construction of the gene families. Including the size-1 families, there are 19 461 entries in the gene-genefamily file, with 15 181 genes appearing once, and 1931 genes appearing in multiple families. The largest five families are C2H2-type zinc fingers (ZNF; 714 without counting pseudo-genes and 720 if pseudo-genes are included), Solute carriers (SLC) (390/396), cluster of differentiation molecules (CD) (389/394), ring finger proteins (RNF) (263/275) and WD (or WD40 for 40 tryptophan–aspartic amino acids)-repeat-domain containing (WDR) (261/262).

Various distributions for the HGNC gene families are shown in figure 2. The empirical cdf of gene family sizes in figure 2a shows that unlike the Ensembl paralogues, most HGNC families are not ‘nuclear families’. It is reasonable, because curating a gene family with only two or three genes is less useful for practical purposes. The histogram of gene family sizes in figure 2b may seem to point to a power-law distribution. However, fitting all points in figure 2b with $y > 0$ (dashed line, slope = −1.02) and fitting only points with $y \geq 2$ (solid line, slope = −1.23) are not consistent with each other. On the other hand, a coarse-graining of the large families by using the log size before fitting with power-law function seems to work better (figure 2c, slope = −1.79). The discrepancy with the Zipf’s power-law becomes
more pronounced if we plot the ranked individual gene family sizes (figure 2d). Including pseudo-genes (circles) versus excluding pseudo-genes (crosses) exhibits some difference in the middle rank ranges.

The power-law or Zipf’s law’s fitting performances of figure 2b–d are as follows (for figure 2b, points with \( y \geq 2 \) are used): MSE is 0.158, 0.217, 0.079; \( R^2 \) and 95% CI is: 0.856 (0.752–0.919), 0.876 (0.728–0.946), 0.933 (0.925–0.944); KS distance is: 0.156, 0.130, 0.173. These fitting performances are worse than those in figure 1b–d. In figure 2b, if points with \( y = 1 \) are also used, MSE = 0.378, \( R^2 = 0.765 \) (0.667–0.838), then the fitting is much worse. In figure 2d, if pseudo-genes are considered, MSE = 0.070, \( R^2 = 0.935 \) (0.925–0.944), then fitting is very similar to the fitting performance when pseudo-genes are excluded.

We use the same re-sampling procedure to test the performance of Zipf’s law for the whole range (for gene set sizes excluding pseudo-genes). In one round of re-sampling with 10,000 replicates, two of the KS distances between the randomly generated data and their fitting function are larger than the observed one. This 0.0002 empirical \( p \)-value indicates that Zipf’s law is a poor fitting function. The shaded area in figure 2d enveloped by error bars also shows that straight lines do not fit the data well.

It is a common practice to dismiss the tail that does not conform to a power-law as finite-size effects, based on the assumption that we do not have enough samples to observe the true value [71]. Although we may gradually improve the fitting performance of Zipf’s law by removing the smaller families (empirical \( p \)-values are 0.01, 0.16, 0.20, 0.48 if we only fit family sizes larger than or equal to 8, 9, 10, 11), we actually remove common events by doing so, not rare events as implied in the idea of finite-size effects.

Not only could the finite-size-effect argument not justify smaller gene sets being rare events, we also have a general belief that all points in the rank–frequency plot such as figure 2d should be fitted, and not to use that argument of finite-size effects arbitrarily ([72]). Previous studies illustrate great potential of the beta rank function in fitting observational data [48,49]. The fitted parameter values in equation (2.1) are \( a = 0.85 \) and \( b = 0.34 \) by linear regression of log size over log-rank (\( a = 0.82 \) and \( b = 0.32 \) with pseudo-genes). Because \( a > b \), we may consider our fitting function a modification of the Zipf’s law.

Figure 2d shows a much better fitting of the data by the beta function (blue) than by the power-law function (red). The improvement of the fitting performance can be further quantified by MSE (0.0179 without pseudo-genes, 0.0155 with pseudo-genes) and \( R^2 \) (0.985 without pseudo-genes, 0.986 with pseudo-genes). Because it is not fair to compare the fitting performance between two functions of different numbers of fitting parameters, we require that the increase of likelihood (or decrease of the fitting error) is more than compensated by the cost of using one more parameter

\[
\Delta AIC = AIC_{\text{beta}} - AIC_{\text{zipf}} = n_F \log \frac{\text{MSE}_{\text{beta}}}{\text{MSE}_{\text{zipf}}} + 2 < 0 \tag{3.1}
\]

or

\[
\Delta BIC = BIC_{\text{beta}} - BIC_{\text{zipf}} = n_F \log \frac{\text{MSE}_{\text{beta}}}{\text{MSE}_{\text{zipf}}} + n_F < 0 \tag{3.2}
\]

The first term in equation (3.1) and equation (3.2) is \(-1065.9\) (without pseudo-genes) and \(-1086.4\) (with pseudo-genes), proving that the rank beta function is better than the power-law function.

### 3.3. Simulating a beta rank function by a split–merge model

We propose a simple model to explain rank functional form such as equation (2.1) which describes the sizes of curated gene sets. We first assume gene set sizes to initially follow a power-law distribution. A curator of the gene sets may fine-tune the collection by (i) splitting the larger (and the second largest) gene sets into two smaller ones and (ii) merging 6% of randomly chosen sets into larger ones by a two-to-one operation. The reason for not choosing the smaller gene sets, but any gene sets, is that once we chose a threshold gene set size to decide which one is small, a discontinuity is introduced to the distribution.

The motivation for the split operation is that curators of gene family databases may find subtle functional difference between genes in the largest families. The motivation for merging gene families could be due to a realization of a common function between two families. These operations are distinctly different from duplication which is an intrinsic mechanism to grow gene family sizes. Our two operations can be considered to be extrinsic ones imposed by curators.

We use the paralogues from Ensembl as the starting gene sets. The ranked gene set sizes after every 10 rounds of operation according to our dynamic model are plotted in figure 3. At the 100th round, the ranked gene set sizes are fitted by a beta rank function (solid line in figure 3). Visual impression indicates that beta rank function fits well the data points at the 100th step. Simulation with other initial distributions has also been carried out, such as uniform distribution, absolute normal distribution, lognormal distribution, chi-square distribution, etc. The resulting distributions after 100 or so iterations
Figure 3. Dynamics of gene set sizes. The initial gene set sizes are taken from the Ensembl paralogues. At each round, the largest gene set is split randomly into two sets, and 6% of all gene sets are merged by a two-to-one operation. The ranked gene set sizes are plotted after every 10 rounds, until reaching the 100th round. The fitting of the 100th round gene set sizes by the rank beta function leads to $a = 0.16$ and $b = 0.84$.

are all similar, although there are subtle differences. We also observe that the $b > a$ in equation (2.1) holds true in our simulation of this particular model.

Because the initial gene set sizes, which are the sizes of paralogues, should be a consequence of duplication and deletion [15] or duplication and mutation [24], the sizes of gene sets produced by our dynamic model is a consequence of all these operations: duplication, deletion, mutation, splitting and merging, though their relative contribution to the final distribution is unclear. Besides modelling of gene family sizes, there has been theoretical modelling of power-law distribution on various other entities in genomes [73–85]. Although these works do not model gene family sizes directly, the connection between a mechanism (duplication) and a feature of the distribution (power-law) is unmistakable. We expect the extra mechanisms discussed here could be a cause of deviating the power-law distribution in these entities also.

4. Discussion and conclusions

4.1. Gene family size information is important for enrichment tests

Gene set is an indispensable part of a microarray expression analysis [13, 86–89]. A typical question is whether the overlapping genes between a gene set with a specific function and a list of differentially expressed genes (between two phenotypes) from microarray data are more than expected by chance. If such an overlap is unlikely by chance, measured by the $p$-value from a statistical test on enrichment, one may infer the relevance of that gene set to the phenotype under study. Because the test result depends on the size of the gene set, enrichment results of two different gene sets may not be comparable. In Subramanian et al. [13], gene sets with size $S$ smaller than 15 are not considered, and enrichment score is normalized by a permutation-obtained mean value which is gene-set-size-dependent. Thus, it is important to characterize the distribution of all gene sets and one should be aware of the gene set size before reaching conclusion on the relevance of a gene set.

4.2. Distribution of transcription factor genes among gene families

The beta rank function equation (2.1) used in figure 2d expands our ability to fit gene set size data. Here, we show two more examples to illustrate its flexibility. We examine how human TF genes are clustered in different HGNC gene sets. Towards this, we use the 1987 TFs listed in Vaquerizas et al. [68], and search them in the HGNC gene sets. Some HGNC gene sets are clearly exclusively TFs. For example, in the C2H2 zinc finger family (ZNF), 633 out of 720 genes are TFs; in the basic helix–loop–helix family (BHLH), 96 out of 110 are TFs; all 52 genes in HOXL subclass homeoboxes and 47 out of 49 nuclear hormone receptors (NR) are TFs, etc.
Figure 4 shows the ranked distribution of TFs in HGNC gene sets (families) when there is at least two TFs in a set. There are 1389 TFs satisfying the condition (in 56 families). Some gene sets may only contain a few TF genes, with the majority being non-TFs (the original gene set sizes are linked to the number of TFs in that gene set in dashed lines in figure 4). A beta rank function is used to fit the data. Re-sampling with 10,000 replicates, 1618 with KS distance larger than observed, and 2180 replicates with the same KS distance as observed, leads to an empirical p-value of 0.38. As a comparison, an empirical p-value using the power-law rank function is 0.12.

4.3. Distribution of drug target genes among gene families

A similar analysis is carried out for the pharmacological or drug target genes [69]. We use the target list obtained from the The International Union of Basic and Clinical Pharmacology/British Pharmacological Society (IUPHAR/BPS) Guide to PHARMACOLOGY site, which only includes those with a UniProtKB/Swiss-Prot ID, thus also having an HGNC name. The first groups of targets annotated in detail include G protein-coupled receptors, ion channels and nuclear hormone receptors [90]. Many proteins labelled as receptor, channel, transporter, kinase, etc. are drug targets or target candidates.

Figure 5 shows the distribution of these target genes in the HGNC gene families. The majority of genes in the solute carrier gene family (390/396) are IUPHAR/BPS targets. Solute carriers are transporters that transport a large variety of smaller molecules such as ions, amino acids, neurotransmitters, sugars, etc, and are known to be good candidates for drug targets [91]. G protein-coupled receptors (GPCRs) reside on the cell membrane (pass through it seven times), sense molecular signals from outside the cell and initiate signal transduction inside the cell. As a result, GPCRs are also candidates for drug targets [92]. Although CD molecules or cell surface molecules provide the second most numerous targets, 60% of CD proteins are still not listed as targets. As seen in figure 5, the beta rank function provides a qualitative trend of the distribution of drug target genes in the HGNC gene families. The systematic deviation from the fitting curve in figure 5 is very similar to that for TFs (figure 4), even though the gene families involved in the two examples are completely different.

4.4. The effect of including pseudo-genes on gene family sizes

There are more than 10,000 pseudo-genes in the human genome [93]. As most pseudo-genes do not have any known function, only a small proportion of them show up in the HGNC gene families/sets.
Excluding pseudo-genes reduces the total number of entries in the gene sets by 1396. In figure 6, we plot the increase of gene set size due to the inclusion of pseudo-genes, as a function of the gene set size. Pseudo-genes have the most dramatic impact on olfactory gene sets [94,95], as expected because the reduced importance of sense of smell to human survival leads to a larger amount of disabling mutations in these genes. The gene set vomeronasal receptors, which is related to the olfactory sense, is labelled individually: it contains three genes without pseudo-genes, whereas 130 genes with pseudo-genes are included. Pseudo-genes also cause large increase of size for immunoglobulin gene sets. On the other hand, pseudo-genes have either no or very little impact on zinc finger gene families.
In conclusion, we aim at bridging two aspects of biological fields, evolutionary biology and functional biology (as well as human biomedical research), by comparing two types of gene families: paralogues and gene sets. While the size distribution of paralogues is well studied and its understanding is through the dynamics of duplication/deletion/mutation, the size distribution of gene sets has rarely been studied and its functional form was unknown. We observed that the size distribution of functional gene sets does not need to follow a power-law distribution, and its deviation from the power-law can be fitted by beta rank function. We propose a mechanism to drive the size distribution away from a power-law function, by splitting the largest gene sets and by randomly merging a small proportion of other gene sets. These results are potentially useful in a gene enrichment analysis as gene set size affects a gene enrichment test result.

Data accessibility. The data files used in figures 1, 2, 4 and 5 are available at http://dx.doi.org/10.5061/dryad.rclkv.

Authors’ contributions. W.L. and P.M. discussed the initial idea of the paper; W.L. carried out most of the analyses presented in the paper; O.F. carried out part of the simulation of the split–merge model; W.L. wrote the initial draft of the paper; W.L., P.M. and O.F. read and improved the draft and approved the final version of the paper.

Competing interests. We have no competing interests.

Funding. W.L. is supported by the Robert S. Boas Center for Genomics and Human Genetics. P.M. is supported by PAPIIT/UNAM IN107414 and PASPA/UNAM. O.F. is supported by CONACYT Mexico.

Acknowledgements. We thank Jan Freudenberg for discussion and recommending references. We also thank the two reviewers for their feedback and comments.

References

1. Walsh JB, Stephan W. 2001 Multigene families: evolution. eLS. (doi:10.1002/0470843381.a00003510001702)
2. Ohta T. 2008 Gene families: multigene families and superfamilies. eLS. (doi:10.1002/9780470015902.a00015160001702)
3. Ohno S. 1970 Evolution by gene duplication. New York, NY: Springer.
4. Taylor JS, Rats E. 2004 Duplication and divergence: the evolution of new genes and old ideas. Annu. Rev. Genet. 38, 605–643. (doi:10.1146/annurev.genet.38.070202.092831)
5. Fitch WM. 2000 Homology: a personal view on some of the problems. Trends Genet. 16, 227–231. (doi:10.1016/S0168-9525(00)02009-9)
6. Koonin EV. 2005 Orthologs, paralogs, and evolutionary genomics. Annu. Rev. Genet. 39, 309–338. (doi:10.1146/annurev.genet.39.070303.114725)
7. Van de Peer Y, Taylor JS, Braasch I, Meyer A. 2001 The ghost of selection past: rates of evolution and functional divergence of anciently duplicated genes. J. Mol. Evol. 53, 456–446. (doi:10.1007/s002390010032)
8. Blanc G, Wolfe KH. 2004 Functional divergence of duplicated genes formed by polyploidization during Arabidopsis evolution. Plant Cell 16, 1679–1691. (doi:10.1105/tpc.024340)
9. Szklarczyk D, Huynen MA, Snel B. 2008 Complex fate of paralogs. BMC Biol. 6, 337. (doi:10.1186/1741-7007-6-337)
10. Haoze T, Ribohsarn R, Eben MA, Alizadeh A, Levy R, Staudt L, Chan WC, Botstein D, Brown P. 2000 Gene shaving as a method for identifying distinct sets of genes with similar expression patterns. Genome Biol. 1, 003. (doi:10.1186/gb-2000-1-2-research0005)
11. Benitez GF, King GD, Bryant B, Sander C, Roth FP. 2003 Characterizing gene sets with FuncAssociate. Bioinformatics 19, 2502–2504. (doi:10.1093/bioinformatics/btg363)
12. Zhang B, Kirov S, Snoddy J. 2005 WebGestalt: an integrated system for exploring gene sets in various biological contexts. Nucl. Acids Res. 33, W741–W748. (doi:10.1093/nar/gqi475)
13. Subramanian A et al. 2005 Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl Acad. Sci. USA 102, 15545–15550. (doi:10.1073/pnas.0506580102)
14. Li W, Freudenberg J, Oswald M. 2015 Principles for the organization of gene-sets. Comp. Biol. Chem. 59, 139–149. (doi:10.1016/j.compbioc hem.2015.04.005)
15. Huyten MA, van Nimwegen E. 1998 The frequency distribution of gene family sizes in complete genomes. Mol. Biol. Evol. 15, 583–589. (doi:10.1093/oxfordjournals.molev.a02559)
16. Molina N, van Nimwegen E. 2009 Scaling laws in functional genome content across prokaryotic clades and lifestyles. Trends Genet. 25, 243–247. (doi:10.1016/j.tig.2009.04.004)
17. Collins RE, Merz H, Higgins PG. 2011 Origin and evolution of gene families in bacteria and archaea. BMC Bioinformatics 12, S14. (doi:10.1186/1471-2105-12-S3-S14)
18. Qian J, Luscombe NM, Gerstein M. 2001 Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model. J. Mol. Biol. 313, 673–681. (doi:10.1006/jmbi.2001.5079)
19. Dolhoyan JV. 2002 Expanding protein universe and its origin from the biological Big Bang. Proc. Natl. Acad. Sci. USA 99, 14132–14136. (doi:10.1073/pnas.202499799)
20. Koonin EV, Wolf YI, Karev GP. 2002 The structure of the protein universe and genome evolution. Nature 420, 216–223. (doi:10.1038/nature01256)
21. Unger R, Ulель S, Havlin S. 2003 Scaling law in sizes of protein sequence families: from super-families to orphan genes. Proteins: Struct. Func. Bioinformatics 51, 549–576. (doi:10.1002/pro.10347)
22. Lee D, Grant A, Marsden RL, Orengo C. 2005 Identification and distribution of protein families in 120 completed genomes using Gene3D. Proteins: Struct. Func. Bioinformatics 59, 603–615. (doi:10.1002/prot.20409)
23. Zueck BE, Huang H, McGarevy P, Mazumder R, Wu CH. 2007 UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics 23, 1282–1288. (doi:10.1093/bioinformatics/btm098)
24. Yanai I, Camacho CJ, DeLisi C. 2000 Predictions of gene family distributions in microbial genomes: evolution by gene duplication and modification. Phys. Rev. Lett. 85, 2641–2644. (doi:10.1103/PhysRevLett.85.2641)
25. Reed WJ, Hughes BD. 2002 A model explaining the size distribution of gene and protein families. Math. Biosci. 189, 97–102. (doi:10.1016/j.mbs.2003.11.002)
26. Karev GP, Wolf YI, Zhretskyi AV, Berezovskaya FS, Koonin EV. 2002 Birth and death of protein domains: a simple model of evolution explains power-law behavior. BMC Evol. Biol. 2, 18. (doi:10.1186/1471-2148-2-18)
27. Karev GP, Wolf YI, Berezovskaya FS, Koonin EV. 2004 Gene family evolution: an in-depth theoretical and simulation analysis of nonlinear birth–death–innovation models. BMC Evol. Biol. 4, 32. (doi:10.1186/1471-2148-4-32)
28. Hahn MW, de Bie T, Stajich JE, Nguyen C, Christianini N. 2005 Estimating the tempo and mode of gene family evolution from comparative genomic data. Genome Biol. 15, 1153–1160. (doi:10.1002/gf.6795)
29. Csuros M, Miklos I. 2006 A probabilistic model for gene content evolution with duplication, loss and horizontal transfer. In Research in Computational Molecular Biology, 10th Annu. Int. Conf., RECOMB 2006, Venice, Italy, 2–5 April 2006, Springer LNBI 3909 (eds AApostolico, CGuerra, SIstrail, PA) 225–251.
30. Zeldovich KB, Chen P, Shakhnovich BE, Shakhnovich EL. 2007 A first-principles model of early evolution: emergence of gene families, species, and preferred
87. Irizarry RA, Wang C, Zhou Y, Speed TP. 2011 Gene set enrichment analysis made simple. Stat. Methods Med. Res. 10, 565–575. (doi:10.1177/0962280211403602)

88. Hung JH, Yang TH, Hu Z, Weng Z, DeLisi C. 2012 Gene set enrichment analysis: performance evaluation and usage guidelines. Brief. Bioinformatics 13, 281–291. (doi:10.1093/bib/bbr049)

89. Tamayo P, Steinhardt G, Liberzon A, Mesirov JP. 2016 The limitations of simple gene set enrichment analysis assuming gene independence. Stat. Methods Med. Res. 25, 472–487. (doi:10.1177/0962280216640441)

90. Pawson AJ et al. 2014 The IUPHAR/BPS Guide to PHARMACOLOGY: an expert-driven knowledgebase of drug targets and their ligands. Nucl. Acids Res. 42, D1098–D1106. (doi:10.1093/nar/gkt1143)

91. Lin L, Yee SW, Kim RB, Giacomini KM. 2015 SLC transporters as therapeutic targets: emerging opportunities. Nat. Rev. Drug Dis. 14, 543–560. (doi:10.1038/nrd4626)

92. Lappano R, Maggiolino M. 2011 G protein-coupled receptors: novel targets for drug discovery in cancer. Nat. Rev. Drug Dis. 10, 47–60. (doi:10.1038/nd320)

93. Harrow J et al. 2010 GENCODE: the reference human genome annotation for the ENCODE Project. Genome Res. 22, 1760–1774. (doi:10.1101/gr.135350.111)

94. Mombaert P. 2001 The human repertoire of odorant receptor genes and pseudogenes. Annu. Rev. Genomics Hum. Genet. 2, 493–510. (doi:10.1146/annurev.genom.2.1.493)

95. Gilad Y, Man O, Paabo S, Lancet D. 2003 Human specific loss of olfactory receptor genes. Proc. Natl Acad. Sci. USA 100, 3324–3327. (doi:10.1073/pnas.0535697100)