

# The Absence of Universally Conserved Protein-coding Genes<sup>1</sup>

Change Laura Tan, Division of Biological Sciences, University of Missouri, Columbia, MO 65211.

## Abstract

Public access to thousands of completely sequenced and annotated genomes provides a great opportunity to address the relationships of different organisms, at the molecular level and on a genome-wide scale. Via comparing the phylogenetic profiles of all protein-coding genes in 317 model species described in the OrthoInspector3.0 database, we found that approximately 29.8% of the total protein-coding genes were orphan genes (genes unique to a specific species) while <0.01% were universal genes (genes with homologs in each of the 317 species analyzed). When weighted by potential birth event (i.e., assuming that the conventional wisdom that a gene and all its homologs are derived from a common ancestral gene is true), the orphan genes comprised 82% of the total, while the universal genes accounted for less than 0.00008%. Strikingly, as the analyzed genomes increased, the sum total of universal and nearly-universal genes plateaued while that of orphan and nearly-orphan genes grew continuously. When the compared species increased to the inclusion of 3863 bacteria, 711 eukaryotes, and 179 archaea, not one of the universal genes remained universal. In other words, all genes are taxonomically-restricted genes. The results speak to a previously unappreciated degree of genetic diversity and challenge the popular view of genes diversification by gene duplication, mutation, and natural selection.

**Keywords:** conserved genes, universal genes, orphan genes, taxonomically restricted genes, origin of genes, origin of life, biodiversity

## Introduction

The rapid advances of whole genome sequencing technologies have facilitated comparative genomic analyses, at fine molecular detail, mostly based on comparisons of similarities of DNA, RNA, or protein sequences, for example, homologous genes or gene contents.

Many methods have been developed to identify homologous genes. By definition, any gene and its homologs, including paralogs (a particular class of homologous genes presumably resulted from gene duplication within an organism) and xenologs (a type of ortholog where the homologous sequences are found in different species presumably because of horizontal gene transfer), are derived from a common ancestral gene. Since the history of genes are generally not known, homology-identification is challenging. Consequently, different methods that are based on different assumptions and algorithms may differ in homology classifications. OrthoInspector is one of the three most balanced methods (the other two are InParanoid and Hieranoid) of orthology inference in specificity and sensitivity (Altenhoff et al. 2016; Liebeskind, McWhite, and Marcotte 2016; Linard et al. 2011).

OrthoInspector identifies orthologs by dividing genes into inparalog groups based on all-to-all proteome BLAST comparisons and then searching

for a reciprocal-best-hit relationship between inparalog groups (Linard et al. 2011). Since it does not require a reciprocal-best-hit between individual genes it is more sensitive than many other methods, including Inparanoid and OrthoMCL (Altenhoff et al. 2016; Li, Stoeckert, and Roos 2003; O'Brien, Remm, and Sonnhammer 2005; Linard et al. 2011). The OrthoInspector algorithm has been used to determine orthologs in 4,753 organisms (3863 bacteria, 711 eukaryotes, and 179 archaea), generating an orthology resource with the broadest species coverage (except viruses) (Nevers et al. 2019). Of the 4,753 organisms, 317 (144 eukaryotes, 142 bacteria, and 31 archaea) are deemed model species (referred to as NMS (Nevers's Model Species) hereafter) either due to their importance in the biological field or due to a consideration of taxonomic coverage. Orthologous genes across three domains of life are available in the OrthoInspector website for these species. For a non-model species, only orthologous genes within its own domain of life are available.

To investigate the relationship of gene contents of different species, we compared the phylogenetic profiles of all NMS. A phylogenetic profile of a protein describes the presence or absence of its homologs across a given set of organisms (Pellegrini et al. 1999). Two proteins with the same phylogenetic profiles tend to function in the same biological process, though the

<sup>1</sup> An earlier version of this article was deposited at bioRxiv preprint: <https://doi.org/10.1101/842633> in 2019 under a CC0 license.

accuracy of the functional-linkage prediction depends on the criteria of defining homology and selection of reference species (Kim and Subramaniam 2006; Pellegrini et al. 1999; Ranea et al. 2007; Sun, Li, and Zhao 2007; Weißenborn and Walther 2017). We focused our attention to proteins with two extreme distributions: universal or orphan. To accommodate the potential bias of the method used to identify homologous genes and the consequence of occasional gene loss or genome sequencing or annotation errors, we have also analyzed nearly-universal genes and nearly-orphan genes.

We discovered an unexpected pattern of the distribution of universal genes and orphan genes. We found that every species has a large number of orphan and nearly-orphan genes, but none, or only a few, universal and nearly-universal genes. Contrary to the common expectation that homologs would be found for orphan genes so that the number of orphan genes would decrease as more species are analyzed, the number of orphan genes grows continuously; each addition of species brings new orphan genes, though often resulting in a decrease of universal genes. Strikingly, not only the homologs of a universal gene are generally not universal genes, but also not a single universal gene maintained its status as a universal gene when enough species were sampled. In other words, all genes are taxonomically restricted, though at different levels of restriction.

### Materials and Methods

Phylogenetic profiles of all the 317 NMS were provided by Nevers and Lecompte in CSV (delimited by tabs). The phylogenetic profile for each species contains information about the presence (indicated with a 1) or absence (indicated with a 0) of homologs in the 317 NMS (columns) for all its protein-coding genes (rows). (The phylogenetic profiles can be obtained from Nevers and Lecompte [Nevers et al. 2019]). For the column that corresponds to the species itself, all cells are 0. When the CSV files were imported into Excel, all species names were shifted one cell to the left. After that was corrected, the phylogenetic profiles were saved as Excel files. Number of species in which there are orthologous genes for a NMS protein was calculated based on the phylogenetic profiles using Microsoft Excel and/or a script written for this project by Andrew Jones. Identity of proteins and that of their orthologs were manually curated from the OrthoInspector website (<https://lbgf.fr/orthoinspectorv3/>). Gene function annotations were mostly from the UniProt Knowledgebase (UniProtKB, <https://www.uniprot.org/uniprot/>), occasionally from organism-specific databases, for example, the *Drosophila*

Genome Database (<https://flybase.org/>) and the *Saccharomyces* Genome Database (<https://www.yeastgenome.org/>). All figures were generated using Microsoft Excel and PowerPoint.

### Categorizing Genes According to the Numbers of Species Having Their Homologs

The number of species having homologs for a specific gene is the sum total for the row to which that gene belongs, since the presence or absence of its homologs in a species (except its own host or home species, also referred to as the reference species) is indicated with 1 or 0, respectively. Note that the presence of multiple homologs in a species does not increase the count beyond 1. A gene was called an orphan, orphan +1 (o+1, “o” means orphan), orphan +2 (o+2)...or a universal gene if the sum total is 0, 1, 2...or 316. Thus, an orphan gene of an organism is unique to that organism, while an “o+1” gene has homologs in one additional organism, except the reference organism.

### Identification of Orphan, Nearly-orphan, Universal, and Nearly-universal Genes

If the number of species having homologs for a specific gene is 316, then that gene is a universal gene, since it has homolog(s) in every organism analyzed. If the number is 0, then that gene is an orphan gene, a gene unique to the reference species; no homolog exists in any of the other species analyzed. Nearly-universal genes are genes conserved in all but five or fewer species analyzed, that is, a sum total of 315, 314, 313, 312, or 311 (corresponding to u-1, u-2, u-3, u-4, u-5 genes, “u” means universal). Nearly-orphan genes, by contrast, are genes that are shared by no more than five of the species analyzed, that is, a sum total of 1, 2, 3, 4, or 5 (corresponding to o+1, o+2, o+3, o+4, o+5 genes).

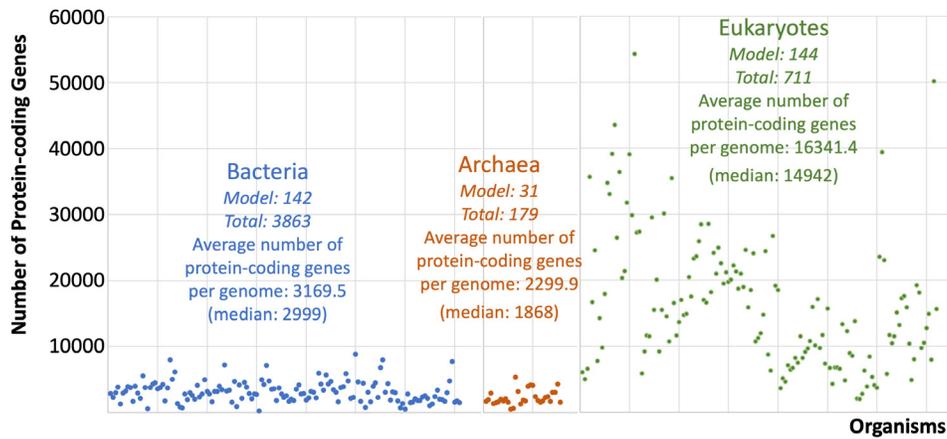
### Weighted Counts of Genes by Potential Birth Event

The weighted value of a gene is the inverse of the number of species, including the gene’s home species (that is., the reference species), that have homologs for that gene. Thus, an orphan gene was counted as one (=1/1) gene, while a universal gene was counted as 0.003155 (=1/317) gene. An o+x gene was counted as 1/(1+x) gene, x is any integer between 1 and 316.

## Results

### Distribution of Total Genes

To gain a broad view of the species analyzed, we compared the sizes of their proteomes (fig. 1). Not surprisingly, on average, eukaryotes have much larger proteome sizes than bacteria and archaea. Eukaryote proteome sizes also vary the most.



**Fig. 1.** Number of protein-coding genes of NMS. Each data point represents a species. Species are arranged according to the order of (Nevers et al. 2019). The first two numbers underneath the name of each domain of life are the number of the model species (indicated with “Model”) and that of the total species (indicated with “Total”) analyzed in the corresponding domain.

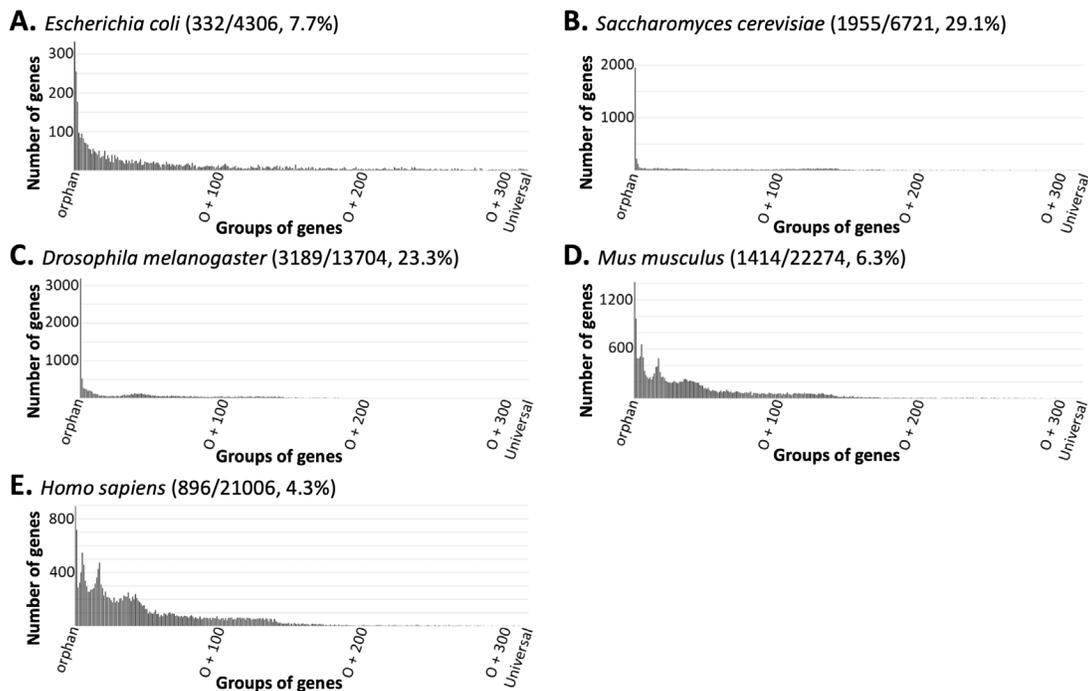
**Grouping Genes According to the Occurrences of Their Homologs**

Next, we placed every protein-coding gene encoded in the genomes of all NMS into orphan, orphan +1, orphan +2... or universal groups (o, o+1, o+2... universal) based on whether the number of species, except the reference species, in which its homologs exist is 0, 1, 2... or 316 (S2 table). Surprisingly, for every species analyzed, the group with most members was the orphan group, and the number of genes in a group quickly dropped into a handful or

even zero with the increase of species containing the corresponding homologs (fig. 2 and S1 fig.).

**Distribution of Universal, Nearly-Universal, Orphan, and Nearly-orphan Genes**

In order to comprehend the above data, we focused our attention on orphan, nearly-orphan, universal, and nearly-universal genes. A nearly-universal gene is conserved in all but five or fewer species analyzed, while a nearly-orphan gene is shared by no more than five of the species analyzed.



**Fig. 2.** The number of genes in a group decreases rapidly as the number of species sharing the corresponding homologs increases. Note that the first vertical line in each panel is not its Y-axis but the orphan gene number in the corresponding organism. The numbers in the parenthesis next to the name of a species in each panel are the number of its orphan genes, all protein-coding genes, and the percent of orphan genes. Only five species are shown here. More examples can be found in S1 fig. and S2 table.

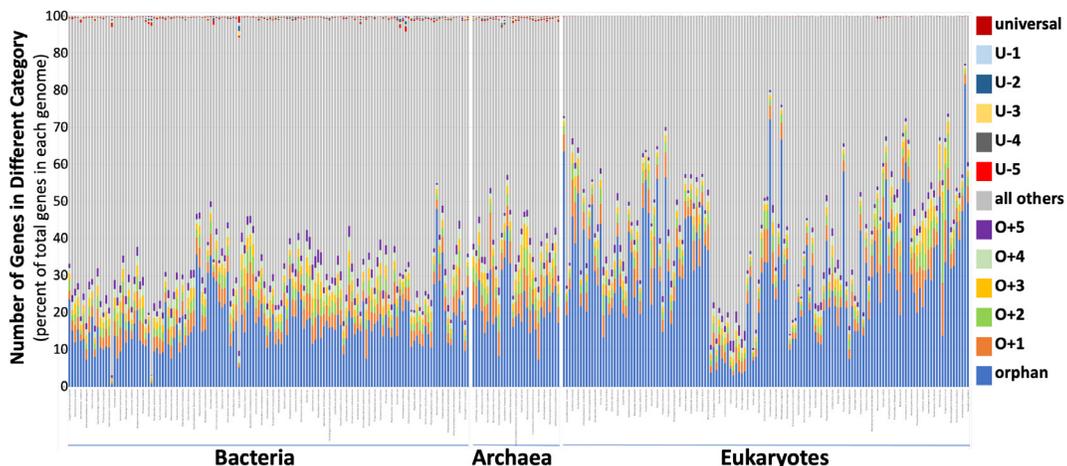
**Table 1.** Numbers of orphan, nearly-orphan, universal, and nearly-universal genes in 17 chosen organisms.

| Species                              | Orphan Genes | Nearly-orphan Genes |      |      |     |     | All Others | Nearly-universal Genes |      |      |      |      | Universal Genes | Total | Orphan percent of total |
|--------------------------------------|--------------|---------------------|------|------|-----|-----|------------|------------------------|------|------|------|------|-----------------|-------|-------------------------|
|                                      | o            | o+1                 | o+2  | o+3  | o+4 | o+5 |            | u -5                   | u -4 | u -3 | u -2 | u -1 |                 |       |                         |
| <i>Escherichia coli</i>              | 332          | 255                 | 177  | 97   | 85  | 95  | 3248       | 4                      | 3    | 4    | 2    | 3    | 1               | 4306  | 7.7                     |
| <i>Caulobacter crescentus</i>        | 635          | 130                 | 99   | 86   | 81  | 74  | 2596       | 5                      | 1    | 2    | 0    | 6    | 0               | 3715  | 17.1                    |
| <i>Mycoplasma genitalium</i>         | 114          | 16                  | 11   | 4    | 5   | 3   | 310        | 5                      | 2    | 2    | 4    | 5    | 2               | 483   | 23.6                    |
| <i>Sulfolobus solfataricus</i>       | 843          | 169                 | 108  | 102  | 69  | 51  | 1567       | 3                      | 3    | 2    | 2    | 4    | 1               | 2924  | 28.8                    |
| <i>Methanocaldococcus jannaschii</i> | 353          | 94                  | 60   | 59   | 36  | 34  | 1127       | 4                      | 4    | 4    | 5    | 3    | 4               | 1787  | 19.8                    |
| <i>Haloferox volcanii</i>            | 700          | 311                 | 265  | 248  | 116 | 60  | 2260       | 5                      | 4    | 3    | 2    | 4    | 4               | 3982  | 17.6                    |
| <i>Zea mays</i>                      | 10719        | 1869                | 1933 | 1033 | 863 | 761 | 21973      | 3                      | 0    | 0    | 0    | 0    | 0               | 39154 | 27.4                    |
| <i>Oryza sativa</i>                  | 19132        | 2380                | 1847 | 958  | 740 | 627 | 17862      | 2                      | 2    | 0    | 0    | 0    | 0               | 43550 | 43.9                    |
| <i>Arabidopsis thaliana</i>          | 5057         | 893                 | 558  | 544  | 546 | 545 | 19108      | 0                      | 1    | 0    | 0    | 0    | 0               | 27252 | 18.6                    |
| <i>Saccharomyces cerevisiae</i>      | 1955         | 225                 | 125  | 61   | 45  | 44  | 4257       | 3                      | 2    | 0    | 0    | 1    | 0               | 6721  | 29.1                    |
| <i>Caenorhabditis elegans</i>        | 9418         | 1285                | 994  | 355  | 201 | 180 | 7634       | 2                      | 1    | 0    | 0    | 1    | 0               | 20071 | 46.9                    |
| <i>Drosophila melanogaster</i>       | 3189         | 525                 | 269  | 250  | 238 | 195 | 9033       | 1                      | 4    | 0    | 0    | 0    | 0               | 13704 | 23.3                    |
| <i>Danio rerio</i>                   | 2808         | 744                 | 496  | 468  | 449 | 339 | 19603      | 0                      | 2    | 0    | 0    | 0    | 0               | 24929 | 11.3                    |
| <i>Canis lupus</i>                   | 973          | 171                 | 188  | 283  | 311 | 534 | 17293      | 2                      | 1    | 0    | 0    | 0    | 0               | 19756 | 4.9                     |
| <i>Mus musculus</i>                  | 1414         | 972                 | 489  | 488  | 510 | 658 | 17739      | 2                      | 2    | 0    | 0    | 0    | 0               | 22274 | 6.3                     |
| <i>Homo sapiens</i>                  | 896          | 719                 | 288  | 325  | 402 | 548 | 17825      | 1                      | 2    | 0    | 0    | 0    | 0               | 21006 | 4.3                     |
| <i>Gallus gallus</i>                 | 1136         | 143                 | 150  | 79   | 87  | 77  | 14236      | 1                      | 2    | 0    | 1    | 1    | 0               | 15913 | 7.1                     |

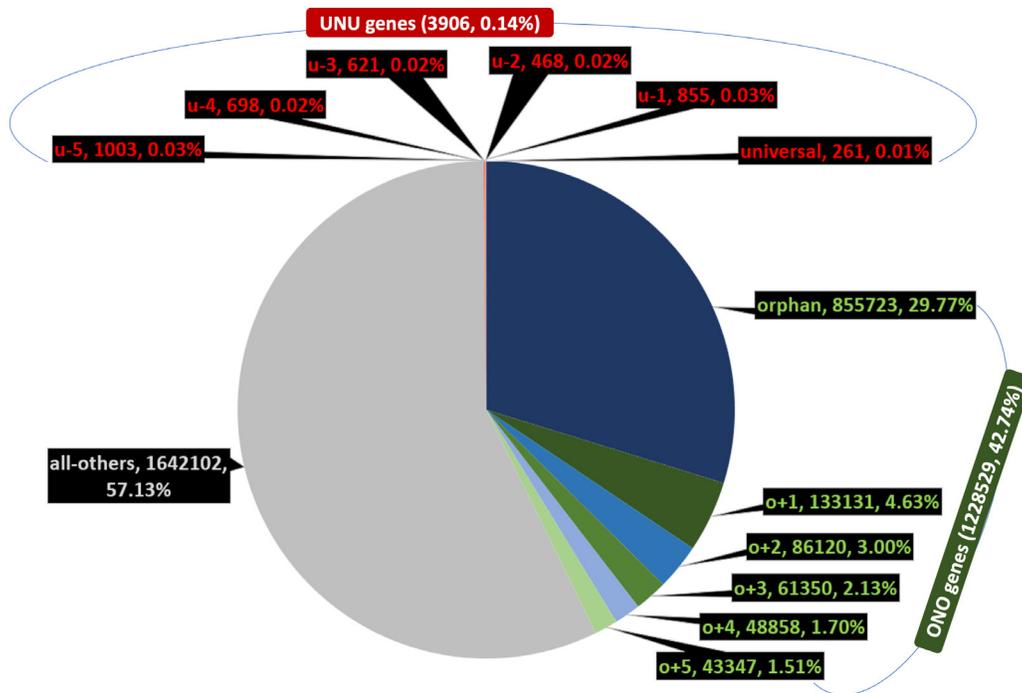
As one would expect from a quick glance at fig. 2, S1 fig., and S2 table, the number of orphan and nearly-orphan (ONO) genes greatly exceeded that of universal and nearly-universal (UNU) genes. Table 1 lists the ONO, UNU, and the proteome of our 17 chosen organisms, including three bacteria, three archaea, three plants, one fungus, and seven animals. S3 table lists the ONO, UNU, and the proteome of all NMS. Fig. 3 shows the percentages of ONO genes (colored sections at the bottom), UNU genes (colored sections at the top), and all other

genes (gray) in each of the model species. Note the great portion of the ONO genes. In contrast, the portion of the UNU genes are barely visible, especially for eukaryotes.

Fig. 4 is a representation of the grouping of genes of all NMS together. Amazingly, the ONO genes represent 42.7% (1,228,529) of the total (2,874,537), with the orphan group itself occupies about 29.8% (855,723) of the total. The UNU groups account for less than 0.14% (3,906) of the total, with the universal group about 0.01% (261) of the total.



**Fig. 3.** Distribution of ONO and UNU genes of all NMS. Each column represents a species whose order of appearance is according to that of (Nevers et al. 2019). Not all organisms are labeled, and the species names are too small to read in this figure due to the limitation of space. For a higher resolution figure with all organisms legibly labeled, see the big graph located at the bottom of S3 table.



**Fig. 4.** Distribution of orphan, nearly-orphan, universal, and nearly-universal genes of NMS as a whole.

Strangely, the number of universal genes, 261, is even smaller than the number of model species, 317. This creates a contradiction with the definition of universal genes, since, by definition, if one true universal gene existed in the model species, then we should have 317 universal genes, because a universal gene should (again, by definition) possess a homolog in each of the other model species. To make the situation worse, more than half of the model species have none, while some of them have several universal genes (table 2). For example, *Candidatus caldiarchaeum* has seven universal genes, almost twice as many as the ten species with the second largest number (four) of universal genes. Although eukaryotic proteomes are generally much larger than bacterial and archaeal proteomes, they have the least number of universal genes. Of the 144 model

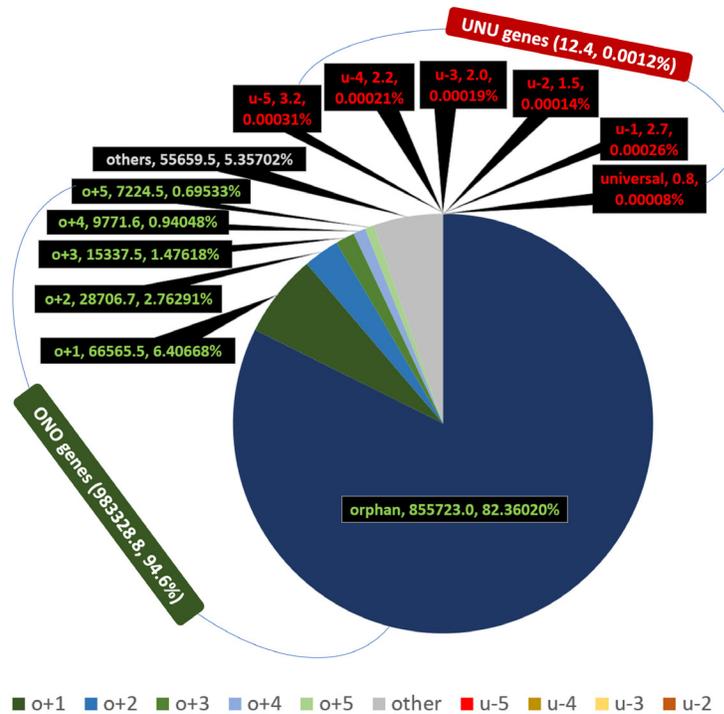
eukaryotes, only ten have universal genes and none has more than one universal genes.

**Weighted Distribution of UNU and ONO Genes**

Homologs of a gene are generally thought to have shared a common gene ancestor. In other words, they shared one “birth event.” Therefore, it is only logical that genes should not have been counted equally; they should be weighted according to their birth events. Consequently, each orphan gene shares its birth right with no other sequence, and should be counted once, while each o+1 gene shares its birth right with an ortholog, and should be counted as 0.5, and so on. When thus weighted, the ONO genes represent about 95% of the total, while the UNU groups represent only about 0.0012% (fig. 5 and S4 table). The orphan group itself makes up more than

**Table 2.** Distribution of species with zero to seven universal genes

| Domain of Life                    | Bacteria | Archaea | Eukaryotes | All  |     |
|-----------------------------------|----------|---------|------------|------|-----|
| <b>Number of Model Species</b>    | 142      | 31      | 144        | 317  |     |
| <b>Number of Universal Genes</b>  | 0        | 33      | 1          | 134  | 168 |
|                                   | 1        | 59      | 4          | 10   | 73  |
|                                   | 2        | 45      | 9          | 0    | 54  |
|                                   | 3        | 3       | 8          | 0    | 11  |
|                                   | 4        | 2       | 8          | 0    | 10  |
|                                   | 5        | 0       | 0          | 0    | 0   |
|                                   | 6        | 0       | 0          | 0    | 0   |
|                                   | 7        | 0       | 1          | 0    | 1   |
| <b>Universal Genes per Genome</b> | 1.17     | 2.74    | 0.07       | 0.82 |     |



**Fig. 5.** Distribution of weighted orphan, nearly-orphan, universal, and nearly-universal genes of NMS as a whole.

82% of the total. In comparison, without weighting, when every gene is counted equally in each species, the ONO genes represent about 43% of the total, while the UNU genes account for about 0.14%.

### Change of the Total Numbers of UNU and ONO Genes with the Species Coverage

To determine the accumulated number of UNU, ONO, and all the protein-coding genes as more species were added, we simply summed the UNU, the ONO, and the proteome of each of the model species, one-by-one, in the sequence of the Nevers's original species order (fig. 6). Thus, each dot in the fig. 6 represents a species and its coordinates (X, Y) were obtained via the following equations:

$$X = \sum_{n=1}^n P_n$$

$$Y_{\text{UNU}} = \sum_{n=1}^n (\text{UNU})_n$$

$$Y_{\text{ONO}} = \sum_{n=1}^n (\text{ONO})_n$$

$P_n$  is the number of protein-coding genes in species  $n$ , whose coordinates are (X, Y). X is the sum of the proteins encoded by organisms 1 to  $n$ .  $Y_{\text{UNU}}$  and  $Y_{\text{ONO}}$  are the sum of the ONU or UNU, respectively, of these organisms.

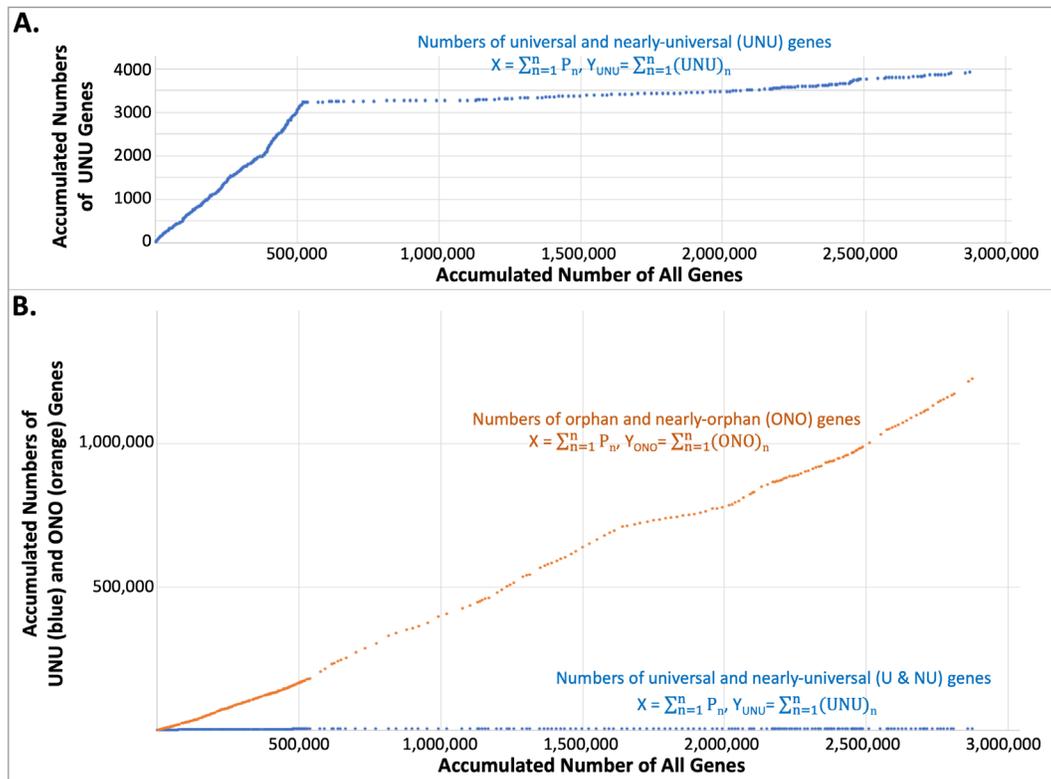
Note that the number of UNU grew quickly initially with the addition of species, but the growth slowed down soon, and almost plateaued at around 3,300 genes (fig. 6A). Strikingly, the ONO number increased continuously, at a much greater speed than

the initial, fastest, growth rate of the UNU number (fig. 6B, orange). Viewed at the same scale, the number of UNU genes appears to show a trend, or slope, of zero (that is, unchanging along the vertical axis) (fig. 6B, blue).

### Identity of the Universal Genes

The numbers of different groups of genes are interesting and important to know but their identity is even more informative for our understanding of life. Here we will describe the identity of the universal genes, while that of the orphan genes will be described in other publications.

We first analyzed the homologs of the universal genes of our seventeen chosen organisms. We observed three phenomena. First, the vast majority of the homologs of a gene encode proteins performing the same function in different organisms, when functional data are available (S5 table). This indicates the specificity of homolog inference of OrthoInspector. The occasional out-of-place homologs, for example, the *Escherichia coli* lysine-tRNA ligase amongst the homologs of asparagine-tRNA ligase and the *Caulobacter crescentus* peptide chain release factor 3 in the mist of elongation factor G homologs, indicates the ability of OrthoInspector to recognize even very low similarity. For example, the *E. coli* lysine-tRNA ligase, though not an asparagine-tRNA ligase, belongs to the same subgroup (class IIb) of tRNA ligases as the asparagine-tRNA ligase, with the sole other member of this group being aspartate-tRNA



**Fig. 6.** Change of the number of UNU, ONO, and all the protein-coding genes as new organisms were added. Each data point represents a species whose coordinates are (X, Y) and whose encoded protein number is  $P_n$ . X is the sum of the proteins encoded by all the organisms up to and including that species. Y is the sum of the UNU or ONO of these organisms.

ligase (Giege and Springer 2016). These three tRNA ligases recognize related anticodons: aspartate GUC, asparagine GUU, and lysine UUU triplets. The *C. crescentus* peptide chain release factor 3 belongs to the same subfamily of GTPases and probably has similar structure and binds to overlapping regions of the bacterial ribosomes as the elongation factor G (Burroughs and Aravind 2019; Graf et al. 2018; Leipe et al. 2002). The fact that most homologs of a gene encode proteins with the same characteristics and/or functions demonstrates that OrthoInspector is a reliable method to identify homologs. Second, the homolog of a universal gene is commonly not a universal gene. This is surprising. However, it explains why the number of universal genes is not an integer multiplication of the number of model species (317), though not why it is smaller than 317. Third, none of the universal genes in our seventeen chosen organisms kept their status as universal genes when checked against their in-domain non-model species in OrthoInspector (S5 table).

Next, we expanded our analysis to all the model species. Consistent with the observation of our seventeen chosen organisms, all universal genes lost their status as a universal gene when checked against their in-domain non-model species in OrthoInspector.

The protein characteristics of five universal genes (Q74MY3\_NANEQ, R1E424\_9ARCH, E4WXB9\_OIKDI, C4V6P4\_NOSCE, A2ER26\_TRIVA) were not clear. To solve this problem, we examined their homologs in our seventeen chosen organisms (S6 table). Assuming a gene's homologs share its identity, we called Q74MY3\_NANEQ and R1E424\_9ARCH Obg-like ATPase 1, A2ER26\_TRIVA and C4V6P4\_NOSCE elongation factor 2, and E4WXB9\_OIKDI isoleucine-tRNA ligase.

With the new assignment for these five universal genes, there are a total of 261 universal genes in the 317 model species. These genes encode the following eight universal proteins: aspartate-tRNA ligase, phenylalanine-tRNA ligase alpha subunit, valine-tRNA ligase, isoleucine-tRNA ligase, elongation factor G (name according to bacteria, corresponding to the archaeal and eukaryotic elongation factor 2), elongation factor Tu (name according to bacteria, corresponding to the archaeal and eukaryotic elongation factor 1), DNA-directed RNA polymerase subunit beta, and Obg-like ATPase 1 (table 3 and S7 table). Five of these eight belong to the up-to-date bacterial core gene set (the two elongation factors and valine-tRNA ligase do not) (Na et al. 2018). Note that the aspartate-tRNA ligase (or any of the other seven

**Table 3.** Identity of universal genes and their distribution in the three domains of life

| Bacteria  | Archaea  | Eukaryotes   |
|---|--|--|
| Aspartate-tRNA ligase (1)                       | Aspartate-tRNA(Asp/Asn) ligase (10)                |  |
| DNA-directed RNA polymerase subunit beta (3)    | DNA-directed RNA polymerase subunit beta (5)       | DNA-directed RNA polymerase subunit beta (4)         |
| Elongation factor G (Elongation factor 2) (7)   | Elongation factor 2 (31)                           | Elongation factor 2 (5) (A2ER26_TRIVA, C4V6P4_NOSCE) |
| Elongation factor Tu (Elongation factor 1) (57) | Elongation factor 1-alpha (11)                     |  |
|   | Isoleucine-tRNA ligase (19)                        | Isoleucine-tRNA ligase (1) (E4WXB9_OIKDI)            |
| Phenylalanine-tRNA ligase alpha subunit (97)    | Phenylalanine-tRNA ligase alpha subunit (7)        |  |
| Valine-tRNA ligase (1)                          |  |  |
|   | Obg-like ATPase 1 (2) (Q74MY3_NANEQ, R1E424_9ARCH) |  |

Note: The numbers in parentheses after the names of genes indicate the numbers of species in which the corresponding genes were recognized as universal genes. Genes recognized as universal genes in species across three domains are shaded in green, those between two domains in tan, and those unique to a single domain are not shaded. The original names for the five universal genes that we assigned new identities have also been included in the table.

of the eight universal proteins) of different species may differ in amino acid sequences and normally those differences are assumed to be the result of accumulated mutation in the history of life. However, the differences may be introduced by God when He created different organisms at the beginning.

Therefore, even though by definition, if there is one universal gene in the model species, we should have 317 such genes because it should have a homologous gene in each of the model species. And all of the 317 homologous genes should be the same, at least very similar. But we have only 261 universal genes and they are not all the same, although they all, except one (Obg-like ATPase 1), are involved in either gene transcription (the DNA-directed RNA polymerase subunit beta) or gene translation (the t-RNA ligases and elongation factors). Even Obg-like ATPase 1 may be involved in translation. For example, human Obg-like ATPase 1 can prevent eIF2 (eukaryotic initiation factor 2) ternary complex formation, leading to inhibition of protein synthesis and promotion of integrated stress response (Balasingam et al. 2019; Chen, Song et al. 2015).

## Discussion

Our comparisons of the genetic profiles of 317 proteomes revealed that ONO genes are a common occurrence, while the UNU genes are very rare. The more organisms are included in the analysis, the more ONO genes are detected and the smaller the percentage of UNU gene becomes. Lastly, not a single universal gene remains its status as universal when enough organisms are included.

### Universal vs Non-universal Genes

The continuous increase of the ONO numbers and

the leveling off of the UNU numbers is consistent with earlier observations (Charlebois and Doolittle 2004; Graham et al. 2000; Khalturin et al. 2009; Koonin 2003; Lagesen, Ussery, and Wassenaar 2010; Mushegian and Koonin 1996; Prabh et al. 2018; Tautz and Domazet-Lošo 2011; Wilson et al. 2005, 2007; Wissler et al. 2013). These analyses have led to a dramatic shrinking, or even vanishing, of the “universal,” or universally conserved, core set of genes and proteins—with a concomitant linear growth in the so-called “orphan” or “taxonomically restricted” sequences. “Accordingly,” notes Koonin, “the universal core of life has shrunk almost to the point of vanishing” (Koonin 2016). Indeed, after complete sequencing of the first two bacterial genomes, a comparison of the 1,727 protein-coding genes of *Haemophilus influenzae* and the 468 *Mycoplasma genitalium* genes identified 240 homologous genes between the two (Mushegian and Koonin 1996). When the number of included prokaryotic genomes increased to 100, the number of universally conserved homologous genes decreased to 63 (Koonin 2003). With the inclusion of 1,000 genomes, the number of universally conserved homologous genes became zero—not a single protein-coding gene was conserved across the 1,000 prokaryotes compared (Lagesen, Ussery, and Wassenaar 2010).

What is surprising, and counterintuitive, is that the homologs of universal genes normally are not universal genes. This results from how we define and detect homologs. Two genes are deemed homologous as long as part of their encoded proteins share some sequence similarity, normally an e-value of  $10^{-3}$  to  $10^{-5}$  in a BLASTp search. OrthoInspector uses an e-value cutoff of  $1e^{-9}$  (about  $1.2 \times 10^{-4}$ ) (Linard et al. 2011). At this condition, the *C. crescentus* peptide chain release factor 3 and elongation factor G are

detected as homologs because they share some sequence homology in their GTPase domains. With the normal e-value cutoffs, the well-known *Drosophila* orphan genes *Jingwei* and *Zeus* will not be recognized as orphan genes because their high sequence similarity with other widely distributed genes (Bai et al. 2007; Chen et al. 2012; Chen, Krinsky, and Long 2013; Long et al. 2003; Long and Langley 1993; Wang et al. 2000; Zhang et al. 2004, 2010). For example, *Drosophila melanogaster* *Zeus*, though without homologs in the bacterial and archaeal model species, has homologs in 100 of Nevers's 143 eukaryotic model species, including *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, and *Gallus gallus* (Tan, unpublished observation).

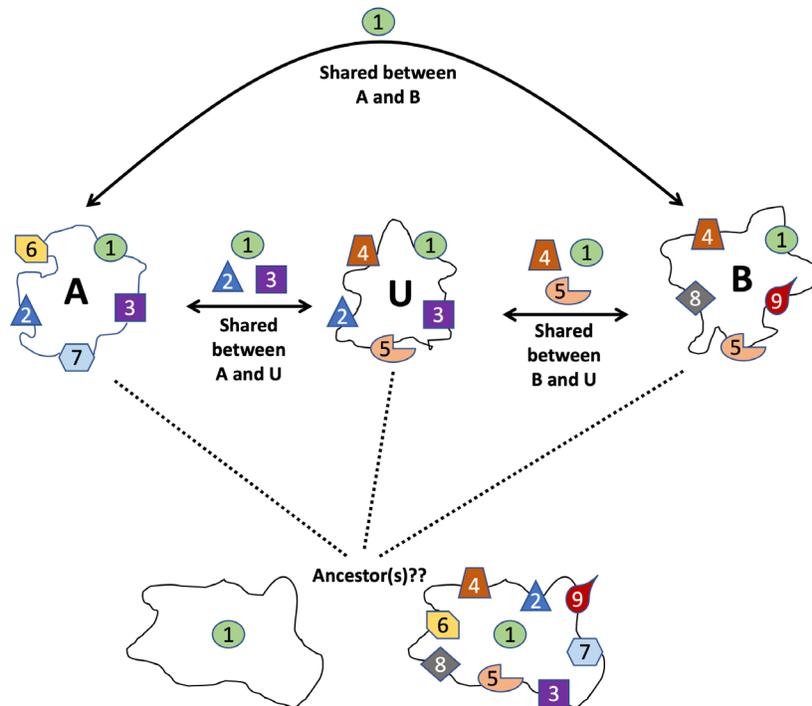
To illustrate how the homologs of a universal gene can be non-universal genes, opposite to what one would expect by the definition of a universal gene, we made up three hypothetical proteins, A, B, and U (fig. 7). U shares parts 1, 2, and 3 with A and parts 1, 4, and 5 with B. Each part can be one or a group of amino acids. Under the criterion that two proteins are homologous if they share three parts, U and A are homologs, so are U and B. However, A and B only share part 1, thus are not homologous. U is a universal gene in this scenario, while A and B are not. How would the ancestor(s) of A, B, and U look like, the one on the bottom left, the one on the bottom right, or something else? How can we know?

**Different Levels of Homology**

The puzzle that homologs of a universal gene

are normally non-universal genes necessitates distinguishing different levels of homology. The concept of gene homology and the identification of homologous genes among different species are foundational to our study and numerous other comparative genomic studies. We think that it will be fruitful to define and distinguish different levels of homology. Here we propose the following levels of homology: 1) protein-domains (or partial sequence homology, or regional homology), 2) full length proteins, 3) full length genes, including the 5'- and 3'- untranslated regions and introns, 4) signal transduction pathways, 5) tissues, 6) organs, 7) body parts. The first three levels are at the molecular level and concern protein-coding genes. The higher the level of homology two genes share, the more likely they will perform the same function. Each level has its own value, even the lowest level, the level one. For instance, if two proteins both have a kinase domain, then they will be a kinase of some sort. Their other non-homologous regions may determine their substrates and their cellular locations of functioning. If these two proteins share homology throughout their entire length, then we can predict that they possess the same substrate specificity and function in the same cell compartment.

Currently, it is a general practice to call two proteins homologous as long as they share level one homology and be put into the same protein family. This can cause unnecessary challenges for functional annotation of genes because thus assigned family members may perform opposite or unrelated



**Fig. 7.** A possible scenario of how a homolog of a universal gene could be a non-universal gene.

functions. This can be confusing, even misleading, especially when thus-identified “homologous” genes are given the same or similar names. For example, Frizzled, a seven-transmembrane protein, and FrzB (also known as soluble or secreted frizzled-related proteins), a protein that is similar to the amino-terminal cysteine rich domain of Frizzled but has no transmembrane segments, are included in the same protein family (InterPro: <https://www.ebi.ac.uk/interpro/entry/InterPro/IPR015526/>, PANTHER: <http://www.pantherdb.org/panther/family.do?clsAccession=PTHR11309>, UniProtKB: <https://www.uniprot.org/uniprot/Q92765>). However, the former is a Wnt receptor necessary for Wnt signaling, while the latter inhibits Wnt-signaling. A similar example is the *C. crescentus* peptide chain release factor 3 and the elongation factor G discussed earlier. They are put into the same subfamily of GTPases but perform different functions. Many bacterial and archaeal genes are called globins because they share partial sequence similarity with hemoglobins, but instead of carrying oxygen around like a hemoglobin, a bacterial globin does not bind oxygen, instead may function as a nitrogen monoxide detoxifier (Vinogradov et al. 2006).

We propose to provide the following information when declaring two genes homologous: 1) level of homology, or homology coverage, 2) degree of homology, 3) e-value cutoff, and 4) a visual representation of the homology. Homology coverage should indicate whether the homology level is of protein-domain, full-length protein, or full-length DNA sequence. For example, we may divide every protein into four quarters, 1 to 4 from the N-terminal to the C-terminal. The homology between FrzB and Frizzled can be described as FrzB-p1-4/Frizzled-p1, “p” for protein. That is, the full length FrzB protein is homologous to the first quarter of the Frizzled protein. Homology degree can be indicated with the percent identity of a BLASTp search. A visual representation of the homology should include both the regions that can be aligned and those that cannot

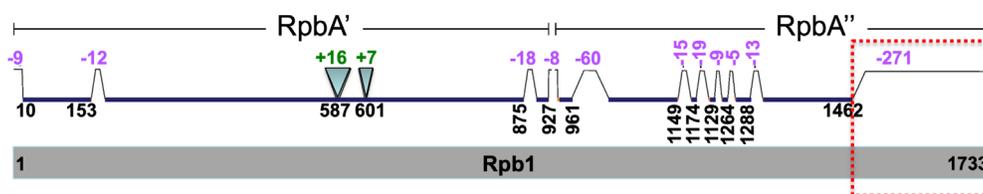
be aligned. Fig. 8 depicts a visual representation of homology between two subunits, Rpb A’ and A”, of archaea *Pyrococcus furiosus* RNA polymerase and the largest subunit Rpb1 of *S. cerevisiae* RNA polymerase II. A’ aligns with the first two quarters of Rpb1. A” aligns with the third quarter of Rpb1. The C-terminal quarter of Rpb1 is unique to eukaryotes and is essential for eukaryotic gene transcription initiation, elongation, termination and intron splicing (Tan 2017). Specific information about the levels at which proteins are homologous will not only avoid making incorrect connection of protein functions but also facilitate understanding of true relationship of genes.

### A Novel Index of Gene Diversity

Our birth-event-weighted gene distribution method can be used as a reasonable indication of the diversity of genes, or different types of genes. Since homologous genes (correctly defined as discussed above) tend to be similar, a non-discriminating counting would inflate the total gene number count encoded by all species on earth. The weighting approach corrects this, and thus can be used as an index of the number of gene types or gene diversity. Moreover, weighting is logically required if indeed homologous genes share a common ancestral gene. But this makes the explanation of the origin of ONO genes more acutely mysterious and makes it more important to study their functions, which we will address in the future. We propose to use the birth-event-weighted distribution of genes as an indication of gene diversity, even though the weighting makes it more difficult to explain the origin of genes by the common belief that genes were generated via duplication and diversification because of the greatly enlarged portion of the ONO genes.

### Limitation of This Study

The number and identity of genes in each of our orphan, o+1, o+2... and universal homologous groups may be different using a different method and/or a



**Fig. 8.** A comparison of *S. cerevisiae* Rpb1 (gray, bottom) with *P. Furiosus* Rpb A’ and A” (green, top). Segment locations are based on amino acid positions of Rpb1 protein. Segments present in Rpb1 but not in Rpb A’ and A” are indicated with trapezoids and negative numbers, while those segments absent in Rpb1 but present in Rpb A’ and A” are indicated with inverted triangles and positive numbers. The values of the numbers, which correlate with the sizes of the triangles and trapezoids, represent the numbers of amino acids of the corresponding regions that are present in only one of the proteins compared. The C-terminal tail of Rpb1 that is missing in *P. furiosus* RNA polymerase is highlighted with a red dash-lined box. The comparison is from Fig. 11 of (Tan 2017).

different e-value cutoff for homology calling or when different species are included, though the trend of the differential growth of ONO and UNU genes will not change. It will be interesting to study how the phylogenetic profiles of proteins will change with a change of the criteria of homolog calling, such as the percentage of gene length covered, percentage of identity, alignment gap penalty, e-value cutoff, and calculation models (alignment algorithms) for these parameters. Furthermore, how the phylogenetic profiles would change when different species are included.

Recent years have witnessed a growth of interests in orphan genes (Arendsee, Li, and Wurtele 2014; Bianchi et al. 1999; Bowling, Schultheis, and Strome 2016; Johnson 2018; Johnson and Tsutsui 2011; Khalturin et al. 2009; McLysaght and Hurst 2016; Palmieri, Kosiol, and Schlotterer 2014; Prabh and Rodelsperger 2016, 2019; Schmitz and Bornberg-Bauer 2017; Tan 2015; Tautz and Domazet-Loso 2011; Toll-Riera et al. 2009; Van Oss and Carvunis 2019; Verster et al. 2017; Wilson et al. 2005; 2007; Wissler et al. 2013; Wu and Knudson 2018; Zhang et al. 2019). However, the increase of interests is incomparable to the increase of orphan gene number. Most of the orphans have unknown functions and will be a rich soil for discovery of new enzymes and/or unknown substrates of known enzymes (Hanson et al. 2009) or new genotype-phenotype connections. The broad existence of orphan genes calls for a greater attention to them from the biological community.

## Conclusions

Our in-depth analysis of phylogenetic profiles of 317 proteomes across the three domains of life shows that ONO genes are a common occurrence in the sense that each organism has a significant number of them, while UNU genes are very rare. Most organisms, especially eukaryotes, do not have any UNU genes. Furthermore, the sum total of UNU genes almost plateaued while the number of ONO genes grew continuously when the number of organisms being analyzed increased. More importantly, every universal gene lost its status as a universal gene when the sampled number of organisms is increased. These results revealed a great challenge to explain not only the origin of genes but also the origin of life and the origin of biodiversity. Finally, we have introduced the birth-event-weighted distribution of genes and propose to use it as an indication of gene diversity, even though the weighting makes it more difficult to explain the origin of genes by the common belief that genes were generated via duplication and diversification because of the greatly enlarged portion of the ONO genes.

## Acknowledgments

The author thanks Andrew Jones for help with calculating the numbers of homologs, Rob Sadler and Paul Nelson for comments on the manuscript, Alan Marshall and Nicholas Valentine for help with computers.

## References

- Altenhoff, Adrian M., Brigitte Boeckmann, Salvador Capella-Gutierrez, Daniel A. Dalquen, Todd DeLuca, Kristoffer Forslund, Jaime Huerta-Cepas, et al. 2016. "Standardized Benchmarking in the Quest for Orthologs." *Nature Methods* 13, no. 5 (4 April): 425–430.
- Arendsee, Zebulun W., Ling Li, and Eve Syrkin Wurtele. 2014. "Coming of Age: Orphan Genes in Plants." *Trends in Plant Science* 19, no. 11 (November): 698–708.
- Bai, Yongsheng, Claudio Casola, Cédric Feschotte, and Esther Betran. 2007. "Comparative Genomics Reveals a Constant Rate of Origination and Convergent Acquisition of Functional Retrogenes in *Drosophila*." *Genome Biology* 8, no. 1 (18 January): R11. <https://doi.org/10.1186/gb-2007-8-1-r11>. <https://www.ncbi.nlm.nih.gov/pubmed/17233920>.
- Balasingam, Nirujah, Harland E. Brandon, Joseph A. Ross, Hans-Joachim Wieden, and Nehal Thakor. 2019. "Cellular Roles of the Human Obg-like ATPase 1 (hOLA1) and its YchF Homologs." *Biochemistry and Cell Biology* 98, no. 1 (February): 1–11.
- Bianchi, Michele M., Geppo Sartori, Micheline Vandenbol, Aneta Kaniak, Daniela Uccelletti, Cristina Mazzoni, Jean-Paul Di Rago, Giovanna Carignani, Piotr P. Slonimski, and Laura Frontali. 1999. "How to Bring Orphan Genes into Functional Families." *Yeast* 15, no. 6 (April): 513–526.
- Bowling, Bethany V., Patrick J. Schultheis, and Erin D. Strome. 2016. "Implementation and Assessment of a Yeast Orphan Gene Research Project: Involving Undergraduates in Authentic Research Experiences and Progressing our Understanding of Uncharacterized Open Reading Frames." *Yeast* 33, no. 2 (13 October): 43–53.
- Burroughs, A. Maxwell, and L. Aravind. 2019. "The Origin and Evolution of Release Factors: Implications for Translation Termination, Ribosome Rescue, and Quality Control Pathways." *International Journal of Molecular Sciences* 20, no. 8 (April).
- Charlebois, Robert L., and W. Ford Doolittle. 2004. "Computing Prokaryotic Gene Ubiquity: Rescuing the Core from Extinction." *Genome Reseach* 14, no. 12 (December): 2469–2477.
- Chen, Huarong, Renduo Song, Guohui Wang, Zonghui Ding, Chunying Yang, Jiawei Zhang, Zihau Zeng, et al. 2015. "OLA1 Regulates Protein Synthesis and Integrated Stress Response by Inhibiting eIF2 Ternary Complex Formation." *Scientific Reports* 5 (18 August): 13241. <https://doi.org/10.1038/srep13241>.
- Chen, Sidi, Benjamin H. Krinsky, and Manyuan Long. 2013. "New Genes as Drivers of Phenotypic Evolution." *Nature Reviews Genetics* 14, no. 9 (16 August): 645–660.
- Chen, Sidi, Xiaochun Ni, Benjamin H. Krinsky, Yong E. Zhang, Maria D. Vibranovski, Kevin P. White, and Manyuan Long. 2012. "Reshaping of Global Gene Expression Networks and Sex-biased Gene Expression by Integration of a Young Gene." *The EMBO Journal* 31, no. 12 (13 June): 2798–2809.

- Giegé, Richard, and Mathias Springer. 2016. "Aminoacyl-tRNA Synthetases in the Bacterial World." *EcoSal Plus* 7, no.1 (May). <https://doi.org/10.1128/ecosalplus.ESP-0002-2016>.
- Graf, Michael, Paul Huter, Cristina Maracci, Mirosław Peterek, Marina V. Rodnina, and Daniel N. Wilson. 2018. "Visualization of Translation Termination Intermediates Trapped by the Apidaecin 137 Peptide During RF3-mediated Recycling of RF1." *Nature Communications* 9, no.1 (August 3): 3053. <https://doi.org/10.1038/s41467-018-05465-1>.
- Graham, David E., Ross Overbeek, Gary J. Olsen, and Carl R. Woese. 2000. "An Archaeal Genomic Signature." *Proceedings of the National Academy of Sciences USA* 97, no.7 (March 28): 3304–3308.
- Hanson, Andrew D. , Anne Pribat, Jeffrey C. Waller, and Valérie de Crécy-Lagard. 2009. "Unknown' Proteins and 'Orphan' Enzymes: The Missing Half of the Engineering Parts List—And How to Find it." *Biochemical Journal* 425, no.1 (December 14): 1–11.
- Johnson, Brian R. 2018. "Taxonomically Restricted Genes Are Fundamental to Biology and Evolution." *Frontiers in Genetics* 9: 407. <https://doi.org/10.3389/fgene.2018.00407>.
- Johnson, Brian R., and Neil D. Tsutsui. 2011. "Taxonomically Restricted Genes are Associated with the Evolution of Sociality in the Honey Bee." *BMC Genomics* 12 (29 March): 164. <https://doi.org/10.1186/1471-2164-12-164>.
- Khalturin, Konstantin, Georg Hemmrich, Sebastian Fraune, René Augustin, and Thomas C.G. Bosch. 2009. "More Than just Orphans: Are Taxonomically-Restricted Genes Important in Evolution?" *Trends in Genetics* 25, no.9 (September): 404–413.
- Kim, Yohan, and Shankar Subramaniam. 2006. "Locally Defined Protein Phylogenetic Profiles Reveal Previously Missed Protein Interactions and Functional Relationships." *Proteins: Structure, Function, and Bioinformatics* 62, no. 4 (1 March): 1115–1124.
- Koonin, Eugene V. 2003. "Comparative Genomics, Minimal Gene-sets and the Last Universal Common Ancestor." *Nature Reviews, Microbiology* 1, no.2 (November): 127–136.
- Koonin, Eugene V. 2016. "Evolution of the Genomic Universe." In *Genetics, Evolution and Radiation*, edited by Victoria L. Korogodina, Carmel E. Mothersill, Sergey G. Inge-Vechtomov, and Colin B. Seymour, 413–440. New York, New York: Springer International.
- Lagesen, Karin, Dave W. Ussery, and Trudy M. Wassenaar. 2010. "Genome Update: The 1000th Genome—A Cautionary Tale." *Microbiology* 156, no.3 (March): 603–608. <https://doi.org/10.1099/mic.0.038257-0>.
- Leipe, Detlef D., Yuri I. Wolf, Eugene V. Koonin, and L. Aravind. 2002. "Classification and Evolution of P-loop GTPases and Related ATPases." *Journal of Molecular Biology* 317, no.1 (15 March): 41–72.
- Li, Li, Christian J. Stoeckert Jr., and David S. Roos. 2003. "OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes." *Genome Research* 13, no.9 (September 13): 2178–2189.
- Liebeskind, Benjamin J., Claire D. McWhite, and Edward M. Marcotte. 2016. "Towards Consensus Gene Ages." *Genome Biology and Evolution* 8, no.6 (3 June): 1812–1823.
- Linard, Benjamin, Julie D. Thompson, Olivier Poch, and Odile Lecompte. 2011. "OrthoInspector: Comprehensive Orthology Analysis and Visual Exploration." *BMC Bioinformatics* 12: 11. <https://doi.org/10.1186/1471-2105-12-11>.
- Long, Manyuan, Esther Betrán, Kevin Thornton, and Wen Wang. 2003. "The Origin of New Genes: Glimpses From the Young and Old." *Nature Reviews: Genetics* 4, no.11 (November): 865–875.
- Long, Manyuan, and Charles H. Langley. 1993. "Natural Selection and the Origin of *jingwei*, a Chimeric Processed Functional Gene in *Drosophila*." *Science* 260. no.5104 (2 April): 91–95.
- McLysaght, Aoife, and Laurence D. Hurst. 2016. "Open Questions in the Study of *de novo* Genes: What, How and Why." *Nature Reviews Genetics* 17, no.9 (25 July): 567–578.
- Mushegian, Arcady R., and Eugene V. Koonin. 1996. "A Minimal Gene Set for Cellular Life Derived by Comparison of Complete Bacterial Genomes." *Proceedings of the National Academy of Sciences USA* 93, no.19 (September 17): 10268–10273.
- Na, Seong-In, Yeong Ouk Kim, Seok-Hwan Yoon, Sung-min Ha, Inwoo Baek, and Jongsik Chun. 2018. "UBCG: Up-to-date Bacterial Core Gene Set and Pipeline for Phylogenomic Tree Reconstruction." *Journal of Microbiology* 56, no. 4 (28 February): 280–285.
- Nevers, Yannis, Arnaud Kress, Audrey Defosset, Raymond Ripp, Benjamin Linard, Julie D. Thompson, Olivier Poch, and Odile Lecompte. 2019. "OrthoInspector 3.0: Open Portal for Comparative Genomics." *Nucleic Acids Research* 47, database issue (8 January): D411–D418.
- O'Brien, Kevin P., Maida Remm, and Erik L. L. Sonnhammer. 2005. "Inparanoid: A Comprehensive Database of Eukaryotic Orthologs." *Nucleic Acids Research* 33, database issue (1 January): D476–480.
- Palmieri, Nicola, Carolin Kosiol, and Christian Schlötterer. 2014. "The Life Cycle of *Drosophila* Orphan Genes." *Elife* 3 (February 19): e01311. <https://doi.org/10.7554/eLife.01311>.
- Pellegrini, Matteo, Edward M. Marcotte, Michael J. Thompson, David Eisenberg, and Todd O. Yeates. 1999. "Assigning Protein Functions by Comparative Genome Analysis: Protein Phylogenetic Profiles." *Proceedings of the National Academy of Sciences USA* 96, no. 8 (April 13): 4285–4288.
- Prabh, Neel, and Christian Rödelsperger. 2016. "Are Orphan Genes Protein-Coding, Prediction Artifacts, or Non-Coding RNAs?" *BMC Bioinformatics* 17, no. 1 (31 May): 226. <https://doi.org/10.1186/s12859-016-1102-x>.
- Prabh, Neel, and Christian Rödelsperger. 2019. "De Novo, Divergence, and Mixed Origin Contribute to the Emergence of Orphan Genes in *Pristionchus* Nematodes." *G3 Genes, Genomes, Genetics* 9, no. 7 (July 9): 2277–2286.
- Prabh, Neel, Waltraud Roeseler, Hanh Witte, Gabi Eberhardt, Ralf J. Sommer, and Christian Rödelsperger. 2018. "Deep Taxon Sampling Reveals the Evolutionary Dynamics of Novel Gene Families in *Pristionchus* Nematodes." *Genome Research* 28, no.11 (November): 1664–1674.
- Ranea, Juan A. G., Corin Yeats, Alastair Grant, and Christine A. Orengo. 2007. "Predicting Protein Function with Hierarchical Phylogenetic Profiles: The Gene3D Phylo-Tuner Method Applied to Eukaryotic Genomes." *PLoS Computational Biology* 3, no.11 (November): e237. <https://doi.org/10.1371/journal.pcbi.0030237>.
- Schmitz, Jonathan F., and Erich Bornberg-Bauer. 2017. "Fact or Fiction: Updates on How Protein-coding Genes

- Might Emerge *de novo* From Previously Non-coding DNA.” *F1000Res* 6 (19 January): 57. <https://doi.org/10.12688/f1000research.10079.1>.
- Sun, Jingchun, Yixue Li, and Zhongming Zhao. 2007. “Phylogenetic Profiles for the Prediction of Protein-Protein Interactions: How to Select Reference Organisms?” *Biochemical and Biophysical Research Communications* 353, no. 4 (23 February): 985–991.
- Tan, Change Laura. 2015. “Using Taxonomically Restricted Essential Genes to Determine Whether Two Organisms Can Belong to the Same Family Tree.” *Answers Research Journal* 8 (November 4): 413-435. [https://assets.answersingenes.org/doc/articles/pdf-versions/arj/v8/taxonomically\\_restricted\\_genes\\_family\\_tree.pdf](https://assets.answersingenes.org/doc/articles/pdf-versions/arj/v8/taxonomically_restricted_genes_family_tree.pdf).
- Tan, Change Laura. 2017. “Holistic Study of Whole Genomes.” *Journal of Genome* 1, no.1 (December 14): 1000e102. <https://www.omicsonline.org/open-access/holistic-study-of-whole-genomes-96453.html>.
- Tautz, Diethard, and Tomislav Domazet-Lošo. 2011. “The Evolutionary Origin of Orphan Genes.” *Nature Reviews Genetics* 12, no.10 (31 August): 692–702.
- Toll-Riera, Macarena, Nina Bosch, Nicolás Bellora, Robert Castelo, Lluís Armengol, Xavier Estivill, and M. Mar Albà. 2009. “Origin of Primate Orphan Genes: A Comparative Genomics Approach.” *Molecular Biology and Evolution* 26, no. 3 (March): 603–612.
- Van Oss, Stephen Branden, and Anne-Ruxandra Carvunis. 2019. “*De novo* Gene Birth.” *PLoS Genetics* 15, no.5 (May 23): e1008160. <https://doi.org/10.1371/journal.pgen.1008160>.
- Verster, Adrian J., Erin B. Styles, Abigail Mateo, W. Brent Derry, Brenda J. Andrews, and Andrew G. Fraser. 2017. “Taxonomically Restricted Genes with Essential Functions Frequently Play Roles in Chromosome Segregation in *Caenorhabditis elegans* and *Saccharomyces cerevisiae*.” *G3 Genes, Genomes, Genetics* 7, no. 10 (1 October): 3337–3347.
- Vinogradov, Serge N., David Hoogewijs, Xavier Bailly, Raúl Arredondo-Peter, Julian Gough, Sylvia Dewilde, Luc Moens, and Jacques R. Vanfleteren. 2006. “A Phylogenomic Profile of Globins.” *BMC Evolutionary Biology* 6: 31. <https://doi.org/10.1186/1471-2148-6-31>.
- Wang, Wen, Jianming Zhang, Carlos Alvarez, Ana Llopart, and Manyuan Long. 2000. “The Origin of the *Jingwei* Gene and the Complex Modular Structure of Its Parental Gene, *Yellow Emperor*, in *Drosophila melanogaster*.” *Molecular Biology and Evolution* 17, no. 9 (September): 1294–1301.
- Weißborn, Sandra, and Dirk Walther. 2017. “Metabolic Pathway Assignment of Plant Genes Based on Phylogenetic Profiling—A Feasibility Study.” *Frontiers in Plant Science* 8 (27 October): 1831. <https://doi.org/10.3389/fpls.2017.01831>.
- Wilson, G.A., N. Bertrand, Y. Patel, J.B. Hughes, E.J. Feil, and D. Field. 2005. “Orphans as Taxonomically Restricted and Ecologically Important Genes.” *Microbiology* 151, no.8 (1 August): 2499–2501. <https://doi.org/10.1099/mic.0.28146-0>.
- Wilson, Gareth A., Edward J. Feil, Andrew K. Lilley, and Dawn Field. 2007. “Large-Scale Comparative Genomic Ranking of Taxonomically Restricted Genes (TRGs) in Bacterial and Archaeal Genomes.” *PLoS One* 2, no.3 (March 28): e324. <https://doi.org/10.1371/journal.pone.0000324>.
- Wissler, Lothar, Jürgen Gadau, Daniel F. Simola, Martin Helmkampf, and Erich Bornberg-Bauer. 2013. “Mechanisms and Dynamics of Orphan Gene Emergence in Insect Genomes.” *Genome Biology and Evolution* 5, no.2 (February): 439–455.
- Wu, Baojun, and Alicia Knudson. 2018. “Tracing the *De Novo* Origin of Protein-Coding Genes in Yeast.” *mBio* 9, no.4 (July–August). <https://doi.org/10.1128/mBio.01024-18>.
- Zhang, Jianming, Antony M. Dean, Frédéric Brunet, and Manyuan Long. 2004. “Evolving Protein Functional Diversity in New Genes of *Drosophila*.” *Proceedings of the National Academy of Sciences USA* 101, no.46 (November 16): 16246–16250. <https://doi.org/10.1073/pnas.0407066101>.
- Zhang, Jianming, Huyuan Yang, Manyuan Long, Liming Li, and Antony M. Dean. 2010. “Evolution of Enzymatic Activities of Testis-Specific Short-Chain Dehydrogenase/Reductase in *Drosophila*.” *Journal of Molecular Evolution* 71, no.4 (31 August): 241–249.
- Zhang, Li, Yan Ren, Tao Yang, Guangwei Li, Jianhai Chen, Andrea R. Gschwend, Yeisoo Yu, et al. 2019. “Rapid Evolution of Protein Diversity by *De Novo* Origination in *Oryza*.” *Nature Ecology and Evolution* 3, no. 4 (11 March): 679–690.

## Supplementary Materials

- S1 fig. Change of numbers of genes in different gene groups in 12 additional species
- S2 table. Number of genes in different gene groups for all NMS
- S3 table. Distribution of orphan, nearly-orphan, universal, and nearly-universal genes of all NMS. The 17 chosen species were highlighted with red font.
- S4 table. Weighted distribution of orphan, nearly-orphan, universal, and nearly-universal genes of all NMS
- S5 table. Analysis of universal genes of selected organisms. Reference genes are highlighted with red font. The seemingly out-of-place homologs are shaded with yellow.
- S6 table. Analysis of universal genes of unknown identity. Reference genes are highlighted with red font.
- S7 table. Analysis of the identity of universal genes of all NMS.

