cells

Article

A Trans-Omics Comparison Reveals Common Gene Expression Strategies in Four Model Organisms and Exposes Similarities and Differences between Them

Jaume Forés-Martos [1], Anabel Forte [2], José García-Martínez [1,*] and José E. Pérez-Ortín [1,*]

1. Instituto de Biotecnología y Biomedicina (Biotecmed), Universitat de València, C/Dr. Moliner 50, E46100 Burjassot, Spain; fores.martos.jaume@gmail.com
2. Departamento de Estadística e Investigación Operativa, Facultad de Matemáticas, Universitat de València, C/Dr. Moliner 50, E46100 Burjassot, Spain; anabel.forte@uv.es
* Correspondence: jose.garcia-martinez@uv.es (J.G.-M.); jose.e.perez@uv.es (J.E.P.-O.); Tel.: +34-963543402 (J.G.-M.); +34-963543467 (J.E.P.-O.)

Abstract: The ultimate goal of gene expression regulation is on the protein level. However, because the amounts of mRNAs and proteins are controlled by their synthesis and degradation rates, the cellular amount of a given protein can be attained by following different strategies. By studying omics data for six expression variables (mRNA and protein amounts, plus their synthesis and decay rates), we previously demonstrated the existence of common expression strategies (CESs) for functionally related genes in the yeast *Saccharomyces cerevisiae*. Here we extend that study to two other eukaryotes: the yeast *Schizosaccharomyces pombe* and cultured human HeLa cells. We also use genomic data from the model prokaryote *Escherichia coli* as an external reference. We show that six-variable profiles (6VPs) can be constructed for every gene and that these 6VPs are similar for genes with similar functions in all the studied organisms. The differences in 6VPs between organisms can be used to establish their phylogenetic relationships. The analysis of the correlations among the six variables supports the hypothesis that most gene expression control occurs in actively growing organisms at the transcription rate level, and that translation plays a minor role. We propose that living organisms use CESs for the genes acting on the same physiological pathways, especially for those belonging to stable macromolecular complexes, but CESs have been modeled by evolution to adapt to the specific life circumstances of each organism.

Keywords: transcription rate; mRNA stability; translation rate; protein stability; phenogram

1. Introduction

The central dogma of molecular biology states that information flows from DNA to protein [1]. The information flux for protein-coding genes has an obligate intermediary: mRNA. The regulation of the expression for these genes ultimately addresses the control of protein levels in the cell because the final goal is readily available protein. In fact, protein abundance (PA) seems to correlate much more between different organisms than mRNA abundance (RA; [2]). However, given the central position of mRNA, and because both RA and PA are controlled by synthesis and degradation rates (Figure 1), the desired PA can be obtained through different strategies [3] that balance the contribution of productive and destructive steps, as well as the relative importance of transcriptional and translational regulation [4]. Recent evidence shows that transcription, and not translation, determines PA under steady-state conditions from yeast [5,6] to mammals [7,8]. Moreover, several studies have suggested that changes in mRNA levels in dynamic scenarios strongly determine protein dynamics (discussed in [9]). This topic, however, is open to discussion [4,10–12]. In fact, it has been argued that under severe pleiotropic stress conditions, the contributions of protein-level regulation, translation rate (TLR) and protein stability (PS) are more important.

Hence, the relative contribution of mRNA-level and protein-level regulation can be context-dependent [4]. It is clear that PA depends on a dynamic balance among these processes, but how this balance is achieved and to what extent all these processes contribute to the regulation of cellular PA are still open questions.
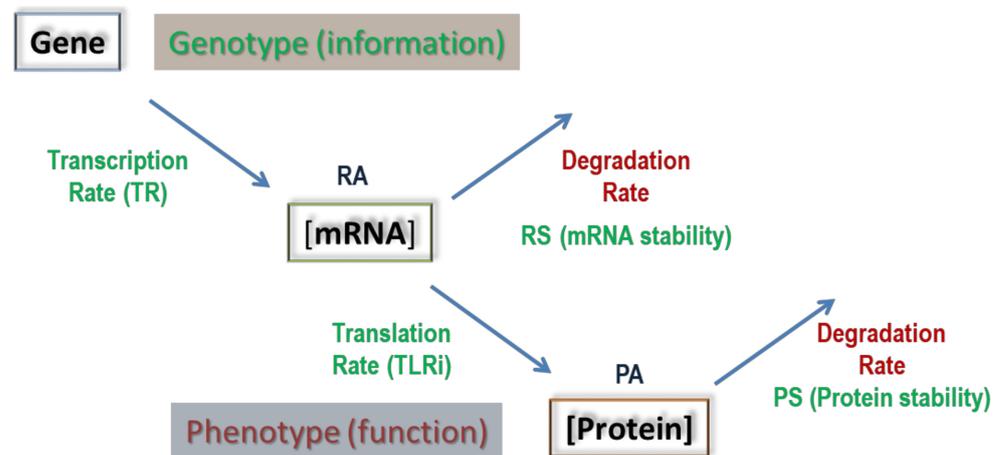

Figure 1. The six variables of the gene expression flux. The genetic information (genes = genotype) is transcribed to mRNAs, of which the concentration (RA), under steady-state conditions, depends on the equilibrium between synthesis (TR) and the degradation rates (herein expressed as the reverse parameter, mRNA stability: RS). The phenotype is, however, dependent mainly on the protein concentration (PA), which, in turn, depends on their synthesis (TLRi) and degradation rates (or stabilities, PS). See the main text for additional explanations.

The evolution of extant cells should have taken into account the energy costs of each step of gene expression. Proteins are thousands of times more abundant than mRNAs and have a larger dynamic range [3,10], which makes their synthesis and regulation much more costly processes. This may be the reason for selecting gene expression mechanisms at the mRNA level [13]. Other variables that influence the selection of specific strategies of gene expression are appropriate speediness and the required level and gradation of the response to potential changes in the environment [14], the feasibility of post-transcriptional and/or post-translational regulatory mechanisms [15] and the optimal biological noise associated with each step [15–17]. In this sense the respective influence of transcription rates and mRNA stabilities on biological noise has been studied in yeast, showing that the two different mRNA decay machineries, Xrn1 and the exosome, contribute differently to the extrinsic noise of gene expression [18].

In a previous study conducted with the model yeast *Saccharomyces cerevisiae*, we addressed these questions by comparing omics data for the abundances of mRNAs and proteins, and their synthesis and degradation rates (studied herein as their reverse variable: stabilities) [3]. We found that yeast cells use common expression strategies (CESs) for genes belonging to the same physiological pathways. Thus we defined a six-variable profile (6VP) for each functional group of genes to illustrate the particular average expression strategy it followed. Our results also showed that synthesis rates and molecule amounts tend to have higher correlations between one another than with stabilities, which suggests a more important role for synthesis rates in the regulation of gene expression.

In the present study, we checked if the results obtained in budding yeast were general to other organisms. To answer this question, we selected three additional model cells for which omics data were available for the six variables: transcription rate (TR), RNA stability (RS), RA, TLR, PS and PA. Thus, we chose another single cell eukaryote, the fission yeast *Schizosaccharomyces pombe*, because it is distantly related to budding yeast, but has converged with it in lifestyle [19–22]. As a higher eukaryote model, we selected HeLa cells, because this is the human cell line with the best and most extensive omics data [23]. Finally, we chose the model bacterium *Escherichia coli* to be representative of prokaryotes for similar

reasons. Our results show that all the investigated organisms have common expression strategies for functionally related gene groups and that CESs are especially robust for genes coding for subunits of stable macromolecular complexes. 6VPs have similarities and differences between organisms that allow phenetic dendograms (phenograms) based on them to be constructed, which, recapitulate the evolutionary tree to a great extent. There are, however, interesting features which indicate that functional divergence is not always strictly proportional to sequence divergence. Finally, our results support the notion that most gene regulation takes place at the mRNA synthesis level, whereas translation plays a minor role, serving to potentiate the effect of transcription.

2. Materials and Methods

2.1. Selection and Features of the Original Data

We used data from several publications that have followed, in many cases, different methodologies or different experimental setups. In all cases, data were obtained from the standard reference strains of the four organisms. For some organisms, the data available came from a single study. However, for others, some of the six variables were measured in more than one study (see Supplementary Table S1 for a summary of datasets used). In order to check the compatibility of the datasets we used the Pearson correlation. If the Pearson correlation was above 0.3, we used them all (see below for specific details). If not, we used the dataset having the highest number of gene data. When having more than one valid dataset, in order to manage missing data for some genes, the k-nearest neighbor method (KNN), available in the Bioconductor library impute [24], was used with all the parameters set to their default values. Then, the medians of all datasets were matched. Finally, the average was calculated between the values represented for each gene in the different datasets to obtain the final values of the amounts, synthesis rates or stabilities. Once the data to be used were selected, we created 4 arrays, one for each organism. The actual data employed for the comparisons are shown in Supplementary Table S2.

In particular, for *S. cerevisiae* (S288c background), we updated the data used in our previous study [3]. These data contained two sources for RS [25,26] obtained through dynamic transcriptome analysis (DTA) and the RNA approach to equilibrium sequencing (RATE-seq), respectively. The correlation for RS in both sources was R = 0.31, just above the established limit. Moreover, despite not having a large correlation, they were combined for convenience, since although RATE-seq technology is the most reliable in methodological terms, analysis with DTA included the RA and TR data from the experiment, which improves correspondence (decreases noise) compared to other parameters. For the RA variable, we used the data from [25,27]. These were combined to give a file with 4985 analyzed genes. The PS data were obtained from [28]. For PA, we used the data from [29]. The TR data came from averaging the results of [25,30]. Finally, we employed individual translation rate data (TLRi) collected from [31]. We used the gene identifiers described in Saccharomyces Genome Database (SGD) (http://www.yeastgenome.org/ (accessed on 1 February 2021)).

For the *S. pombe* data, we utilized the following datasets. For the RS variable, the data came from [32–34] (DNA microarrays). Data were combined in a single dataset. For RA, data were obtained from [34,35] (RNA-seq), as well as [36]. Data were combined into a single dataset. For TR, data were obtained from [34–36]. The TLRi data were collected from [36]. The PA data were obtained from [35,37]. Finally, the PS data were from [28]. We used the gene identifiers described in pomBase (https://www.pombase.org/ (accessed on 1 February 2021)).

For HeLa cells, we employed the following data. For RS, the data were obtained from [38]. With RA, data were available from three sources [39–41], which were combined to create a final dataset. The PS data came from analyzing [41]. For PA, data came from three sources that correlated well (R > 0.3). As what actually matters is the relative expression of each protein compared to the others, although two of them presented expression data as number of protein copies per cell (PCN) [42,43] and the other as intensity-Based

Absolute Quantification (iBAQ) [39], the latter was converted into PCN with a conversion factor obtained from the plot between them. As a result, a single dataset was obtained. Finally, there is no published HeLa TR and TLRi data; therefore, these were calculated mathematically from the quantity and synthesis data using Equations (1) and (2):

$$TR = \frac{RA}{RS} \tag{1}$$

$$TLR = \frac{PA}{PS}$$
$$TLR = RA \times TLRi$$

Hence:

$$TLRi = \frac{PA}{RA} \times PS \tag{2}$$

The gene identifiers were converted into the HUGO nomenclature (http://www.genenames.org/ (accessed on 1 February 2021)) because it is the most widely used one among different sources.

For *E. coli*, we used the following data obtained from the K12 strain exponentially grown at 37 °C in M9 or LB media. For RS, we used the data from [44,45]. They were combined, creating a single dataset. For RA, we utilized the datasets from [44–47]. They were combined to give a single dataset. For TLRi, we used the data from [46]. For PA, we employed the data from [45,47,48]. They were combined to give a single dataset. The TR and PS data were obtained mathematically from other data using Equations (1) and (2). We used the gene identifiers described in the EcoCyc Database [49].

2.2. Correlation Analyses

To test the global correlation among all the pairwise combinations of the six variables (obtained as explained in the previous section), Pearson's correlation coefficients (R) and their associated *p*-values were calculated using the data from the genes for which complete information was available (Spearman rank correlations yielded equivalent results, not shown).

However, in a complex process like this, Pearson's correlations are not enough to capture how the 5 considered variables are related to PA, taking into account the fact that they are correlated among them. Hence it is more appropriate to consider a multiple regression analysis in order to capture this interdependence. Moreover, it is important to keep in mind that there exist many different linear combinations of these 5 variables and some of them may be more suitable than others. That is, not all variables may be directly related with PA in the presence of the rest and these relations may change among organisms. Hence, we decided to incorporate the uncertainty about the suitability of the model by using a Bayesian model averaging (BMA) procedure [50].

BMA uses the strength of Bayesian analysis by averaging estimations and predictions from different models weighted by the posterior probabilities of each model. In particular, the analysis was performed using the BayesVarSel R package [51], allowing us to make a robust estimation of the coefficients of each variable. It worth mentioning that this type of procedure cannot be performed using classical statistics since we need to determine the posterior probability of each possible combination of covariates, which is only available when using the Bayesian approach. Moreover, BMA estimation of a coefficient takes into account the potential correlations among the independent variables included in the linear regression. Note that in order to make all the coefficients comparable, data were transformed using Z-scores (since Z-scores maintain the internal structure of the data, still accounting for their dispersion within and among organisms). Accordingly, the posterior mean and 95% credible interval for each coefficient and organism were calculated. Certain variables were not included in the analysis for some organisms given their direct mathematical calculation from the response variable (PA), which was the case for TLRi in HeLa and PS in *E. coli*.

2.3. Six-Variable Profiles (6VPs)

For 6VPs we followed the protocol previously described in [3]: we analyzed the behavior of the genes from all four organisms using rank data values (0 was the lowest value, and 1 was the highest). Ranking, similarly to Z-scores, serves to avoid the wide dispersion in the unit ranges seen when comparing the different datasets for the six variables obtained with very distinct experimental techniques. In addition, for the comparison of 6VPs between different organisms, the use of ranking offers an advantage over Z-scores that use the same scale (0–1) for all groups and organisms. In this way, although some information was lost, the results were much more robust. Matrices were obtained with 3613, 4139, 3350 and 1643 genes represented for HeLa, *S. cerevisiae*, *S. pombe* and *E. coli*, respectively (Supplementary Table S3).

To show the performance of the 6VP protocol we selected some Gene Ontology (GO) terms with enough genes (>50) in all three eukaryotes or groups of functionally related genes that were obtained from a previous selection, described in [3]. The definition of groups was based on *S. cerevisiae* genes. The orthologous genes from the other eukaryotes were obtained from the YeastMine database (http://yeastmine.yeastgenome.org/ (accessed on 1 February 2021)) for HeLa and from pomBase (https://www.pombase.org/ (accessed on 1 February 2021)) for *S. pombe*.

We calculated the average rank value and represented these values for the six variables in this order—TR, RS, RA, TLRi, PS and PA, to yield a 6VP for each studied group. We also calculated the standard error (SE) associated with each average and represented it in the profile as error bars. A control test was done by averaging values of 1000 random samplings with the same sample size as the analyzed functional group. In all cases they appeared as a flat line at the 0.5 score. They have been omitted in the graphs for clarity.

2.4. Cluster Analyses

In order to identify the groups of genes with similar expression profiles, gene clustering was done according to their 6VP (Figure 3). In this way, we settled characteristic expression profiles with the data of the six variables for the genes with at least four represented variables (of which at least two had to be of mRNA and the other two of protein).

For the clustering analysis, to allow a better comparison with the previous study we used the same algorithm that was used in that study [3]: the *sota*() function of the R clValid package [52]. This function performs clustering with the self-organizing tree algorithm (SOTA) [53] using the linear correlation coefficient among the six variable vectors as the distance between genes, following a splitter scheme that allows the algorithm to be stopped at any point to gain the desired number of clusters. This algorithm does not allow any variable with missing data, so the Bioconductor *impute* package [24] was used to impute unrepresented data by means of the k-nearest neighbors method (KNN). The algorithm was settled in order to avoid clusters with less than five genes.

Trees were allowed to grow until 10, 15, 20 or 30 clusters were produced. Then, clusters were manually selected from any clustering level by considering the *p*-value of the enrichment for GO terms by looking for the best ones, and in such a way that clusters do not overlap. The whole set of clusters and its GO searches can be seen in Supplementary Files S1–S4.

2.5. Gene Ontology Category Searches

To test potential enrichment in GO terms in the different clusters obtained by SOTA, and as previously explained, we used the GOstats R package. For this purpose, a test based on hypergeometric distribution was run for the terms or divisions of ontologies Cellular Component (CC) and Biological Process (BP). Only GO terms were considered significant when applying the false discovery rate (FDR) method [54] for multiple comparisons, and they yielded an adjusted *p*-value of < 0.001. These terms were filtered by a semantic comparison process with the help of the GOSemSim package [55] to avoid redundant GO

terms. Using this package, a function was designed to select the GO group with the lowest adjusted *p*-value of all those with a semantic similarity greater than 70%.

2.6. Comparison of the Proximity in 6VPs among the Genes Belonging to Macromolecular Complexes and Those Belonging to GO Categories Not Forming Macromolecular Complexes

We selected the 18 stable protein complexes in *S. cerevisiae* that had more than five and less than 150 genes from the MIPS database described in the previous study [3]. We also selected the 15 GO categories that included less than 350 genes and were known not to include complexes. Supplementary Table S4 indicates the selected GO terms and the genes included in those for which complete data were available.

2.7. Phenogram Tree Construction

We performed both neighbor-joining (NJ) and hierarchical clustering by Unweighted Pair Group Method with Arithmetic mean (UPGMA) using the information deriving from our GO term-level 6VPs as input. We removed excessively broad GO terms and the GO terms containing very few genes. Thus, in order to keep a functional category for the downstream analysis, it had to contain a number of genes between 5 and 275 in all four species. We only included those functional categories for which we had data about all six variables.

We employed only Biological Process (BP) and Cellular Component (CC) ontologies for searches, removing very closely related GO terms, which would probably present large gene overlaps. First, we used the GOSemSim package [55], a package designed for the semantic comparisons of Gene Ontology (GO) annotations. All the analyses were carried out by taking the human GO database as a reference. From our list dataset of the GO terms, we selected those found in the human GO database. Then for the list of retrieved GO terms, we computed a matrix of pairwise similarity values using the Rel method. In short, the Rel method combines Resnik's and Lin's methods [56,57] to compute semantic similarity (SimREL) between any given pair of GO terms. SimREL values range from 0 to 1, and the higher the value is, the greater the similarity between GO terms. The pairwise matrix of the SimREL values were transformed into a distance matrix (SimREL-dist) by computing 1-SimREL. For each functional category, a parameter called uniqueness was computed as 1 minus the average of the SimREL values of each GO term to all the other terms. This parameter indicates how different a specific GO term is compared to all the others.

The distance matrix, SimREL-dist, was then employed to perform hierarchical clustering using the average (UPGMA) method. The mean silhouette information was extracted for any possible divisions from 1 to the number of the included functional categories −1. The number of clusters yielding the highest average silhouette value was selected. Then for each cluster of the GO terms, the term with the highest uniqueness value was selected as the most representative element of each cluster. For those clusters with only one element, this one element was selected. The set of representative elements of each cluster included the GO terms used in the downstream analysis. Lists of the GO terms employed in each tree are shown in Supplementary Table S5.

Using the whole set of GO terms yielded by the previous procedure, we carried out a clustering analysis using two different methods: hierarchical average (UPGMA) and neighbor-joining employing the NJ and hclust methods implemented in phangorn [58] and stats packages. We also visually inspected the GO terms and created groups of GO terms linked with the following cellular processes: cytoplasmic translation, mitochondrion, transcription and replication to make the clustering analysis of selected terms shown in Figure 5B and those of cytoplasmic Golgi, vacuole, ribosome biogenesis, nuclear pore, proteasome and spliceosome in Supplementary Figure S2.

3. Results

3.1. Datasets for the Variables Selected in This Study

In this work, we studied the six variables that control gene expression (see Figure 1)—transcription rate (TR), mRNA amount (RA), mRNA stability (RS), translation rate per mRNA (TLRi), protein amount (PA) and protein stability (PS)—using omics datasets under a standard growth condition for each studied organism. When the published work studied different growth conditions, we selected that with the highest growth rate because in the rest of studies this condition was used. Although the right parameter to be used is concentration rather than amount, we can assume that the cell volume for each organism is constant for all their datasets obtained under the same culture conditions and, thus, variations in the number of molecules and their concentrations are equivalent. As there were large differences within the ranges of actual values among the six variables, we used ranks and Z-scores instead of absolute values to make the results more robust.

In a previous study into *S. cerevisiae* [3], we used the datasets available for the omics data at that time. We have updated some of the datasets from *S. cerevisiae* by taking special care with the mRNA half-lives dataset (RS) that produces very different results from the previous one (see below). For *S. pombe*, HeLa and *E. coli* we selected from the available datasets. When two datasets or more were available, we took the average of the well-correlated ones to be the final value (see Materials and Methods for a detailed protocol). For human cells, we selected the HeLa datasets because this cultured cell line covers more information about omics data. Although other human cells lines in culture may evidently have different data for specific genes, we assume that the global behavior in HeLa cells is the best available possibility and is a representative of human cells in culture.

The actual synthesis rates of mRNAs and proteins, TR and TLR are in fact the product of individual rates (or rate constants), namely TRi and TLRi, multiplied by the number of genes or mRNA copies, respectively. For mRNA synthesis rates, in practice TR and TRi are equivalents for single-cell organisms because most genes have one (in haploids) or two (diploids) copies. Hence, we use the acronym TR throughout this paper. Although HeLa cells are considered mostly diploid, they have a high aneuploidy level and numerous large chromosome structural variants [59]. Therefore, the TR in HeLa is not equivalent to TRi, but is the right value to be used because is the actual determinant of mRNA levels. For protein synthesis, TLR and TLRi are, however, essentially different. Given its dependence on RA (see Materials and Methods), TLR is mathematically linked with it. Yet TLRi reflects an intrinsic property of mRNA and has been calculated experimentally, usually as ribosome density, by ribosome profiling [31]. For this reason, we employed TLRi values for our analyses.

Finally, it should be noted that the quality of the datasets is not the same for the four organisms. These kinds of omics studies have been conducted more frequently in yeasts. This is because the gene coverage in both yeasts for most variables is much better than for HeLa and *E. coli*. The existence of mRNA and protein isoforms also adds a complication to the interpretation of the study. The number of this isoforms is higher in HeLa. In this study we have grouped all isoforms present in datasets under a single gene name. The number of datasets for both yeast species is much larger, and we could select several and make an average dataset. Moreover, the yeast datasets for all variables consist of direct experimental data. For HeLa and *E. coli*, however, some variable datasets were not available, and we had to estimate them mathematically from others (see Materials and Methods). Therefore, we consider that our conclusions are more robust when taken from budding and fission yeast variables.

3.2. Comparisons and Correlations between Variables

It is believed that protein amounts depend mainly on mRNA amounts under steady-state conditions [4,8], although the correlation value is still a matter of discussion [6,60,61]. Thus, a positive correlation between RA and PA datasets is expected. In fact, this has been previously observed for the three eukaryotes studied herein [3,35,39,62]. Other correlations

have been much less studied [3,6,9], but may illustrate the strategy adopted by cells to determine a given PA.

In this study we obtained pairwise correlations (Pearson) among the six variables considered for each organism (Figure 2). It is worth noting that the actual correlations between variables may be underestimates of the true correlations due to measurement errors in the employed datasets. For all the organisms, we found positive and statistically significant correlations (red backgrounds in Figure 2A) between RS and TR with RA, which means that both synthesis and degradation rates (i.e., mRNA stabilities) control mRNA levels. The higher TR/RA correlations indicate a stronger influence of TR on mRNA levels. RAs also positively correlate with their stabilities, and take lower values, in the three eukaryotes, but not in *E. coli*. This indicates that stabilities are not as important as synthesis rates for determining mRNA levels. In our previous work [3], we found a non-correlation or even a negative correlation between RS and all the other variables in *S. cerevisiae*. This time we used new more reliable RS datasets, and this result changed. We think that, together with other authors [26,34,63–70], the RS datasets obtained by transcription shutoff methods are affected by marked biases that invert the observed correlation [63]. Thus, we consider that the current results reflect that mRNA stability has a positive role on mRNA levels. In *E. coli*, we found no correlation or a negative correlation between RS and RA or TR, probably due to the RS dataset having been obtained by transcription shutoff [44].
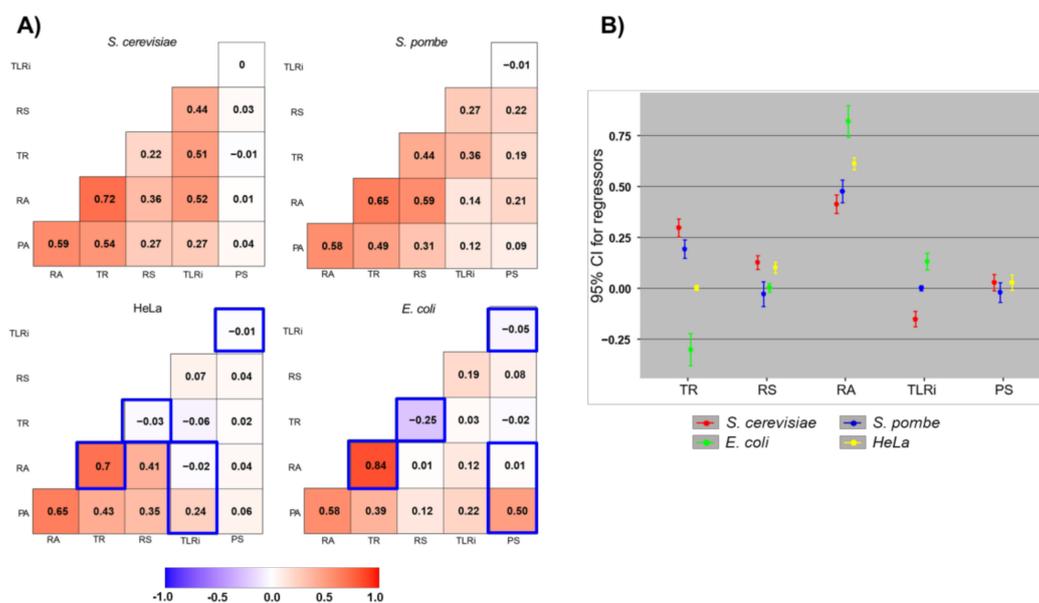

Figure 2. Correlations between the six variables of the gene expression in four model organisms. (**A**) The pairwise Pearson's correlations among the six variables in each organism. Exact values are indicated in frames. The background color scale indicates positive (red) or negative (blue) correlations. Some correlations are less reliable because they are between the variables calculated from some others in HeLa and *E. coli* (see the main text) and are marked with blue boxes. (**B**) Estimation of the coefficients of a multiple regression model per organism using Bayesian model averaging (posterior mean and 95% credible interval). These coefficients show the contribution of each variable across organisms to the final protein amount (PA) using a Z-score scale to make them comparable. The TLRi for HeLa and the PS for *E. coli* are not shown because they were mathematically calculated from PA values (see Materials and Methods).

The large correlations between RA and PA and the lesser correlations between PA and TLRi suggest that, as recently suggested [5–8], mRNA levels are much more important for determining protein levels than specific translation rates, although TLRi can be very different between mRNAs and explain part of the actual PA level [6]. However, it should be taken into account that, for HeLa, the TLRi values were derived as a mathematical calculation involving PA, RA and PS (see Materials and Methods) and, therefore, the corresponding correlations (marked with a blue box in Figure 2A) may be influenced

by this fact. In any case, the message of the low positive, but statistically significant, TLRi-PA correlation in the other three organisms supports the idea that abundant mRNAs tend to be better translatable. This is probably because they are also enriched in optimal codons [71], which has been called potentiation or the amplification exponent [5,8]. The positive correlations of TLRi seen in the yeasts with RA, TR and RS also support the idea that a potentiation effect occurs during translation. Following the same argument, we can conclude from the TLRi-RS positive correlation that a direct proportionality appears in free living cells (at least in eukaryotes) between the stability of an mRNA and its capacity to be translated, which confirms the results from J. Coller's [72] and other laboratories [63] by showing that in *S. cerevisiae* mRNA enriched in optimal codons (better translatable) are more stable than those depleted in such codons. Given that Coller's results were based on RS calculated by means of the transcription shutoff method, these results seem contradictory with our previous results, showing that these kinds of methods provoke unreliable RS datasets. However, the study of Carneiro et al. [63] and our own unpublished studies (García-Martínez, Arnau, Choder and Perez-Ortín, in preparation) have shown that, for the specific case of correlation between codon optimality and RS, some studies (including Coller's one) using transcription shutoff studies show a positive correlation and appear reliable for this purpose. Similar results on the variability of RS datasets have been obtained for *S. pombe* [73] in *E. coli*, zebrafish and mammalian cells (reviewed in [71]). Our analysis cannot confirm this for *E. coli* and, especially HeLa, which may be due to a genuine lack of correlation or to an artefactual bias due to the direct mathematical calculation of some variables, as mentioned above.

The existence of much better correlations between TLRi and PA than between PS and PA argues that for protein amount, and even more strongly than for the mRNA amount, the total synthesis rate (TLR, e.g., multiplying RA by TLRi) is the main determinant. In fact, it would seem that protein degradation is not used by any organism as a way to determine the most steady-state protein levels. This is not surprising for fast living single cells in which protein half-lives are usually much longer than generation times, which implies that dilution by cell division is a much more important factor for protein disappearance than protein degradation [13]. Moreover, for all organisms, including HeLa cells, where the generation time is longer, the high energy cost of regulating abundant proteins by degradation does not seem to be a suitable strategy [13].

In order to compare the differences in the contribution of each covariate among organisms we performed a multiple regression analysis with the variables transformed using a Z-score scale. However, given our reduced knowledge about the optimal combination of variables with this purpose, we opted to consider them all by incorporating a Bayesian model averaging approach [51]. This approach estimates the coefficient associated to each variable by using a weighted average of its estimation under each possible model. In this context, the weights are the posterior probability for each given model. Figure 2B then shows the posterior weighted estimation for each parameter, together with a 95% credible interval. This analysis confirms that for all four organisms, RA is the most important contributor to PA. In the two yeasts, TR also makes a significant contribution. In *E. coli* and HeLa, as previously explained, TR was mathematically calculated and is less reliable. The other three parameters, RS, TLRi and especially PS contribute much less to PA.

These results support the conclusion that the main determinant of the protein level in a cell is the corresponding mRNA level and that RA, in turn, depends mostly on the transcription rate of the gene.

3.3. Clustering of Genes According to the Six Variables of Gene Expression into Four Different Organisms

Our previous results obtained with *S. cerevisiae* demonstrated that functionally related genes tend to be grouped according to their gene expression variables [3]. In this work, we repeated the clustering for that yeast with new datasets, and for the other three organisms studied using the previously described omics datasets. Here, however, we used ranked values instead of the Z-scores used there [3] (see Supplementary Table S3) because the

ranges for the six variables were quite different between the four organisms and variables (see Materials and Methods). We employed the six values for clustering in this order: TR, RS, RA, TLRi, PS and PA (6VP; see Figure 3A). We performed a clustering analysis of the genes for which the data on at least four of the six variables were available (Supplementary Table S3). In this way, we analyzed 4139 genes for *S. cerevisiae*, 3350 genes for *S. pombe*, 1653 genes for *E. coli* and 3613 genes for HeLa cells.
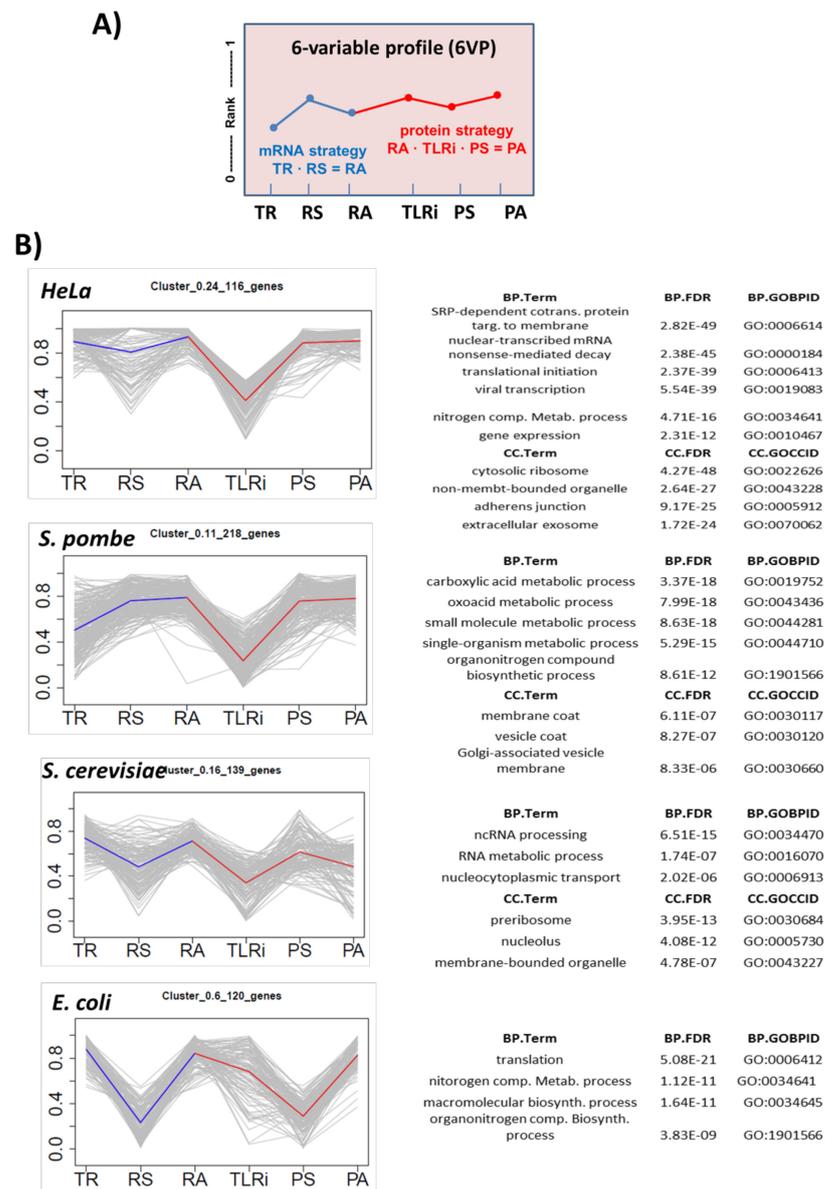

Figure 3. The 6-variable profile (6VP) allows researchers to cluster all genes according to their expression strategy. (**A**) Genes were ranked from lowest (0) to highest (1) in each variable (TR, RS, RA, TLRi, PS, PA). The shape of the line linking the six points is characteristic of an expression strategy: a 6VP. The strategy to gain a defined RA corresponds to the first part of the line (in blue): in the steady state the mRNA level depends on both its transcription rate and stability. For the protein strategy (in red), the final level depends on both the total translation rate (TLRi x RA) and protein stability. (**B**) Many clusters selected from the four studied organisms show defined profiles and statistically enriched GO terms. The individual gene profiles are shown in gray and the average profile of the cluster on the colored line. Some of the GOs with the highest *p*-values are shown on the right for the Biological Process (BP) and Cellular Component (CC) ontologies. Other examples are found in Supplementary Figure S1 and the whole lists are given as Appendices.

Thus we obtained a 6VP for each gene. This allowed us to compare all the genes for common profiles by means of standard clustering methods. Through this procedure, we found that clusters had genes with similar profiles that were statistically enriched in the Gene Ontology (GO) categories (terms) in all four organisms (Figure 3B, Supplementary Figure S1 and Supplementary Files). This result extends our previous results obtained with *S. cerevisiae* [3] and demonstrates that common expression strategies (CES) for the genes with a related biological function are a common feature of living beings. It should be noted that the quality of the datasets for the four organisms (see above) can influence clustering quality. We consider that the 6VP is more robust for the two yeasts, and in HeLa at a lower level. For *E. coli*, the poorer quality of the datasets and the lower quality of the information in the GO annotations for this organism [74] precluded the finding of strongly enriched clusters.

3.4. Detailed Analysis of the Selected Functional Groups in Eukaryotes

If all the analyzed organisms have CES for at least part of their gene groups, we may wonder if the particular CES (depicted as a particular 6VP) followed by a given gene group is similar or different in distinct organisms. Given the low quality of the *E. coli* annotations, we compared only the three eukaryotes. Figure 4 depicts some examples of these comparisons. We studied either selected GOs (Figure 4A) or some of the manually curated groups (Figure 4B) analyzed in a previous work on *S. cerevisiae* [3] to look for the orthologous genes in *S. pombe* and HeLa.

It can be seen that the average 6VPs for the groups generally show a similar ranking for all the variables in the three organisms. For instance, protein folding (GO:0006457) shows that all the six variables rank 0.6–0.8 in all three organisms, whereas cytosolic ribosome (GO:0022625) ranks higher than 0.8 for most variables and the nuclear pore (GO:0005643) ranks mostly between 0.4 and 0.6. This demonstrates that the particular levels of mRNAs and proteins for a given group of genes tend to be similar in all eukaryotes and, to a lesser extent, that the strategies followed to this end are also similar. It is interesting to note that some other groups differ. For instance, the spliceosomal complex (GO:0005681) ranks higher in HeLa than in the two yeasts. This result is logical given the much more marked importance of splicing for human genes [75].

Regarding the shape of profiles, we found similarities and differences. V-shaped profiles (meaning lower stabilities than synthesis rates) are common in stable macromolecular complexes in *S. cerevisiae*, especially for the mRNA part (see Figure 3A,B). This has been previously noted [3]. This feature has also been noted for human THP-1 and C2C12 cells [76] but is not so common in *S. pombe* and HeLa cells, where only some stable complexes behave as such. For instance, cytosolic and mitochondrial ribosomes have V-shaped profiles in budding yeast, but are not so marked in fission yeast and HeLa for the protein part. The spliceosomal complex is more clearly V-shaped in HeLa than in both yeasts but proteasome is conversely V-shaped in both yeasts, but not in HeLa. To conclude, we can state that the existence of cases with similar and different strategies in the three model eukaryotes makes 6VPs suitable for comparing the expression strategies for the whole gene sets between different organisms.
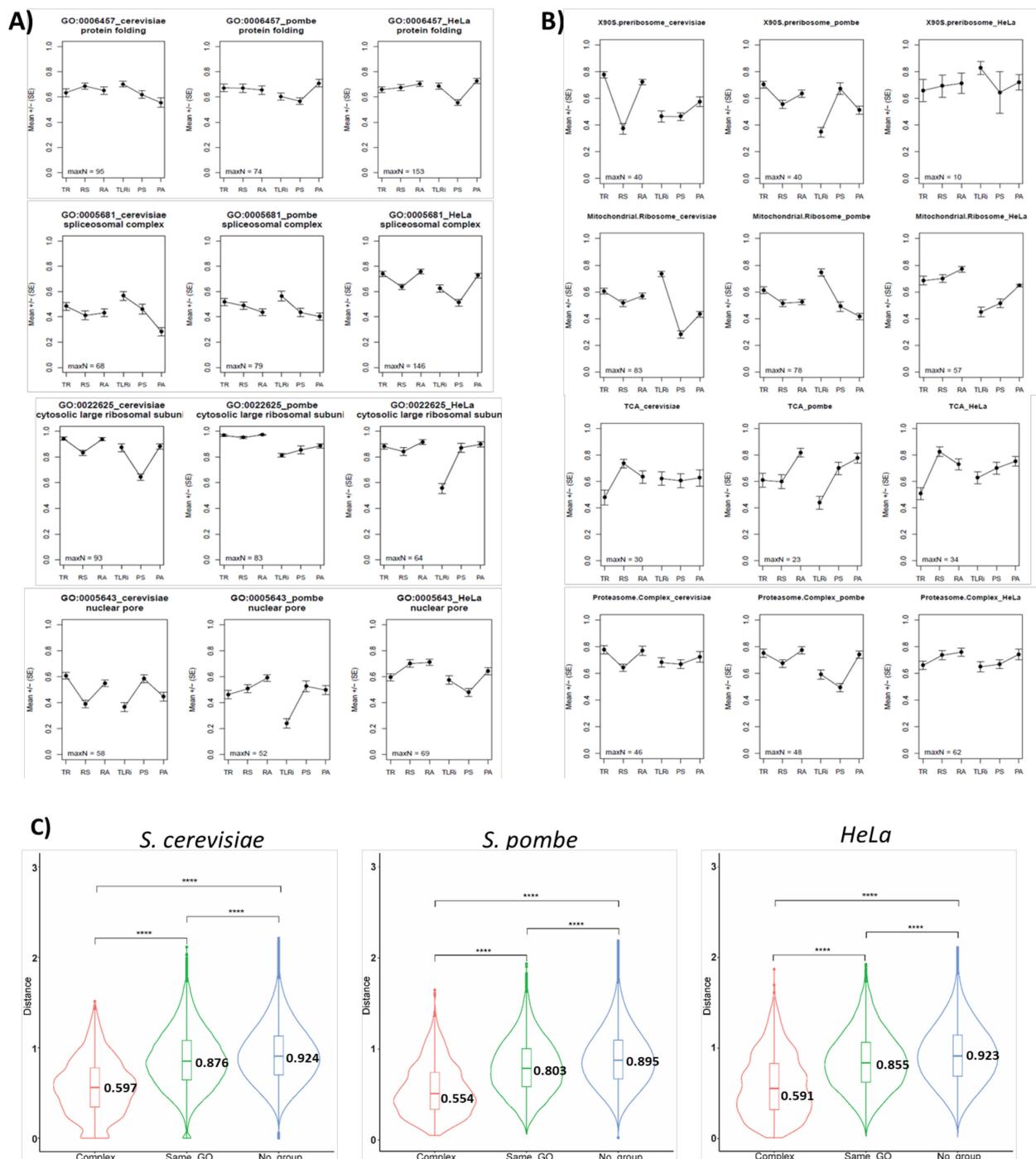
Figure 4. 6VPs for the selected GOs and manually curated groups in *S. cerevisiae*, *S. pombe* and HeLa cells. (**A**) We selected four GO terms corresponding to well-known groups of functionally related genes/proteins. The ID of the GO and its name are shown for *S. cerevisiae* (left), *S. pombe* (center) and HeLa (right). The maximum number of genes used (maxN) to determine the average value and standard error is shown. In some variables, this number may be lower due to the lack of information for some particular genes in some datasets. In this figure, unlike Figure 3, we independently represent mRNA and protein parts without a connecting line to better display the differences between them. (**B**) A similar analysis to that in (**A**), but done with the manually-curated categories that were used in a previous study on *S. cerevisiae* [3]. (**C**) Violin plots showing the mean of proximity (values shown on the plots) between the genes coding for the proteins acting in macromolecular complexes (Complex, red) belonging to the same GO, but do not form complexes (Same_GO, green) and other gene pairs that do not belong to the previous groups (No_group, blue). A *t*-test was applied for the statistical significance of the difference. **** *p*-value < 2.2×10^{-16} (lowest numerical value displayed in R). An ANOVA also detected a significant difference (*p*-value < 2.2×10^{-16}) among the three gene groups in all the organisms the first time they were cited.

Both this analysis and the previous one in *S. cerevisiae* [3] suggest that the genes coding for proteins that are subunits of stoichiometric stable complexes tend to have better defined 6VPs than other functionally related gene groups not forming complexes. To test this hypothesis, we performed a comparative analysis of the profile distances between the gene pairs belonging to protein macromolecular complexes and the genes belonging to the GO categories not including macromolecular complexes (Figure 4C). The *S. cerevisiae* distance matrix included information about 2592 gene pairs that participate in the same protein complexes (Complex), 33558 gene pairs placed in the same selected GO categories with no protein complexes (Same_GO) and 3537651 gene pairs which were not included in either group (No_group). The mean pairwise distances for the genes included in protein complexes was 0.597, whereas these distances were respectively 0.876 and 0.924 for the gene pairs included in the same GO category and the gene pairs not included in either group. Given that the differences between Same_GO and No_group were small, we used two statistical tests to check for the significance. ANOVA showed significant differences between groups and the post hoc *t*-test determined that all the pairwise tests were significant. With *S. pombe*, after the orthologous conversion, the distance matrix included information on 2396 gene pairs that participate in the same protein complexes, 15,279 gene pairs placed in the same selected GO categories and 1,339,453 gene pairs not included in either group. The mean pairwise distances for the genes included in protein complexes was 0.554, whereas they were respectively 0.803 and 0.895 for the gene pairs included in the same GO category and the gene pairs not included in either group. The numbers for Hela were as follows: 2660 for the gene pairs in complexes, 8512 for the gene pair distances in the same GO categories, and 793,374 gene pairs for the No Group class. The mean pairwise distances for the genes included in protein complexes was 0.591, whereas they were 0.855 and 0.923, respectively, for the gene pairs included in the same GO category and the gene pairs not included in either group.

Therefore, it is clear in all three eukaryotes that 6VP distances are much lower between the genes belonging to complexes, although the genes belonging to the same GO category that do not form complexes are still closer than random gene pairs.

3.5. Trans-Organism Clusters Comparison: 6VP Phenograms

As we previously observed cases in which 6VPs for GO terms were similar between some organisms, but different in others, we wondered if the whole similarity of the 6VPs among the four studied organisms could be used to make a phenetic tree based on the similarities of the expression strategies for the same functional groups among the four organisms. In order to do this, we selected the GO terms with a number of genes between 5 and 275 to avoid excessively small groups, which can bias clustering, and the excessively broad ones containing genes that were not true functionally related genes. In this set of GO terms (from Biological Process, BP, and Cellular Component, CC, ontologies), we reduced the redundancy of similar GO terms by applying a procedure inspired by the REVIGO pipeline [77] (see Materials and Methods). The goal of the redundancy reduction step was to avoid the information about genes present in highly redundant GO terms to excessively influence the following cluster and tree construction.

As we can see in Figure 5A, the topology of the global neighbor joining (NJ) and UPGMA trees is identical to the known topology of the DNA sequence-based tree [20,78]. Given the poor quality of the GO annotation in *E. coli*, the global tree can only use 53 common GO terms for the four organisms (51 BP + 2 CC), and we consider that the branching of this prokaryote is less robust. However, *E. coli* can be considered an outgroup for the eukaryote tree. We repeated the tree using only the three eukaryotes, which extended the set of common GO terms to 437 (353 BP + 84 CC). The topology of this NJ tree is identical and the relative branch lengths are similar to the previous one. The lengths of the branches between the two yeasts and HeLa indicate that *S. cerevisiae* comes slightly closer to human cells in gene expression strategies than *S. pombe*.
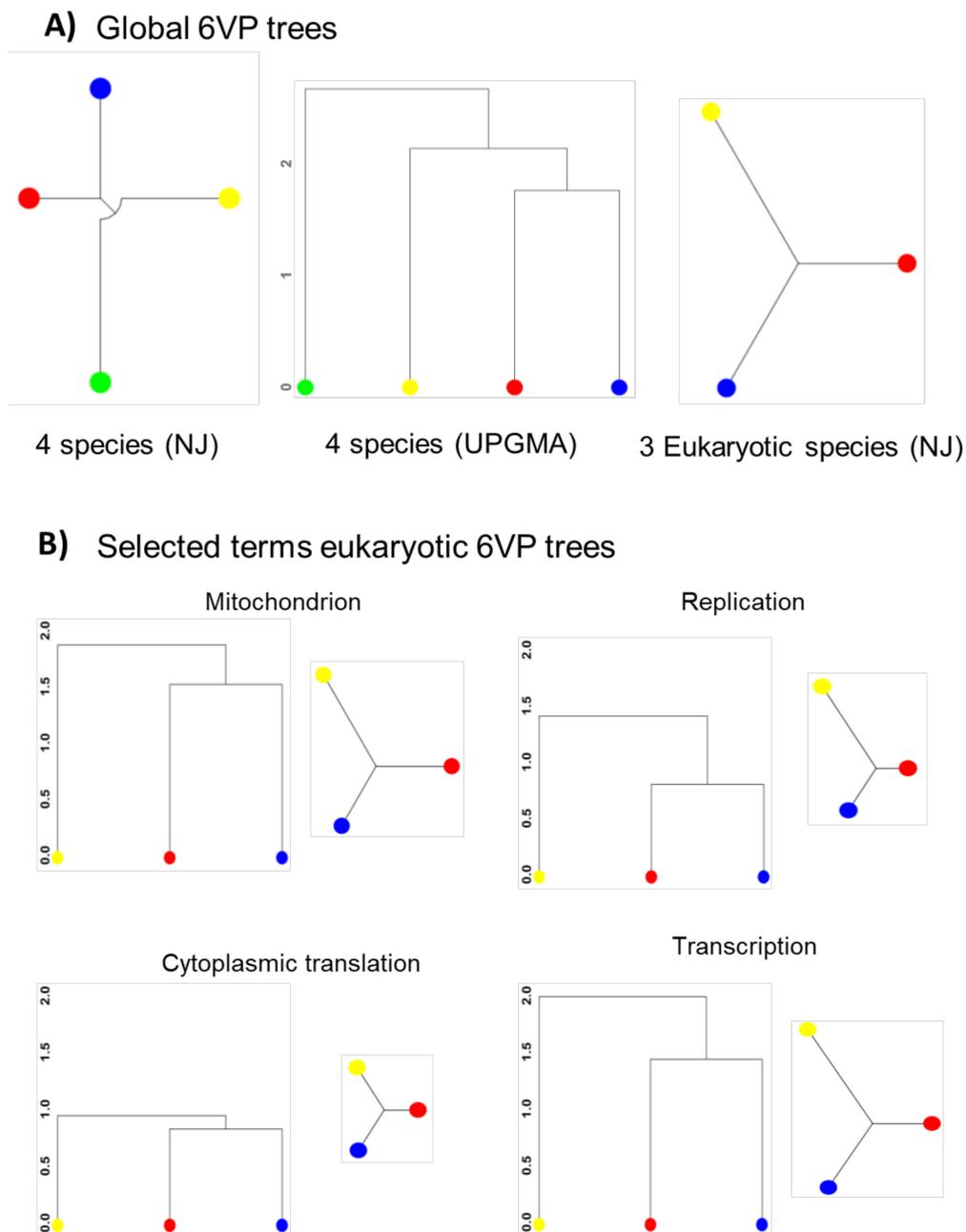
Figure 5. Phenograms of the four organisms based on their 6VPs. (**A**) Global neighbor-joining (NJ) and UPGMA trees based on the 6VPs derived from 53 GO Biological Process (BP) and 2 GO Cellular Component (CC) terms for *S. cerevisiae* (red dots), *S. pombe* (blue dots), *E. coli* (green dots) and H. sapiens (HeLa cell line, yellow dots) (left and center), or only the NJ tree for the three eukaryotes with 353 GO BP, plus 84 GO CC terms (right). (**B**) NJ and UPGMA trees based on the available GO terms linked (manually curated) with cytoplasmic translation, transcription, mitochondrion and replication. All the NJ and UPGMA unrooted trees are, respectively, on the same scale, but NJ and UPGMA are represented with a different kind of branching to show complementary information. The lists of the GO terms used in each tree are provided in Supplementary Table S5. Additional phenograms for other GO terms are shown in Supplementary Figure S2.

We wondered if the topology of the tree and the lengths of branches were the same for the different functional categories of genes. We repeated the clustering and tree reconstruction by separately using groups of the GO terms belonging to broad eukaryotic cellular functions, such as those related to macromolecule synthesis processes (replication, transcription, cytoplasmic translation) or to the mitochondrion (Figure 5B), because this is a large set of genes known to be coordinately regulated [3,79]. We also performed

this analysis (Supplementary Figure S2) for other general GOs, comprising large protein complexes (proteasome, spliceosome, nuclear pore), organelles (Golgi, vacuole) and wide co-regulated groups (ribosome biogenesis, RiBi). Figure 5B and Supplementary Figure S2 illustrate that the UPGMA trees tend to repeat the topology of the global one. There were, however, some exceptions in some groups (Supplementary Figure S2) that show almost identical branch lengths for the three organisms (Golgi, proteasome, RiBi, vacuole). Distances in NJ trees tend to be shorter between budding yeast and human cells, except for the mitochondrion group (Figure 5B) and some of the cases noted in Supplementary Figure S2 (spliceosome, Golgi), where the fission yeast comes closer to HeLa. Thus, we can conclude that in terms of cell functions, the two yeasts are much more similar to one another than to human cells, and budding yeast comes slightly closer to human cells than the fission yeast. Finally, the lengths of the branches in both the UPGMA and NJ trees differ for each broad group, which suggests that distinct cell functions and components are more conserved (cytoplasmic translation, vacuole, proteasome, nuclear pore) than others (i.e., mitochondrion, transcription, spliceosome).

4. Discussion

A capital question in biology is how genetic information is converted into function—how the variable copy number of a given protein is obtained from the constant copy number of its gene. Cells have multiple steps in which the gene expression flux can be regulated. In a series of relevant papers, M. Biggin's group [8,9] and others [5,6] have shown that protein levels under steady-state conditions are explained mostly by mRNA levels in both yeasts and mammals, and that these levels depend mainly on synthesis rates [9]. In a previous study in the model organism *S. cerevisiae*, we developed a pipeline to compare the respective influences of mRNA and protein synthesis and degradation rates to the final level of most of the proteins of this yeast. We found that the genes belonging to functionally related groups followed a similar gene expression strategy, which can be defined by a 'six-variable profile (6VP)' followed by the genes included in it [3]. In the present study, we extended this pipeline to three other model cells to check the extensibility of our previous conclusions.

Our results support the idea that the main pathway used for gene expression is based on synthesis rates for all organisms. The transcription rate is the main determinant in all four organisms of the mRNA level, which in turn is the main determinant of the protein level (Supplementary Figure S3), as confirmed by the MBA analysis presented in Figure 2B, which showed that for all four organisms studied, the mRNA level (RA) had the highest coefficient. Obviously, this result does not exclude the existence of cases in which other variables also have an effect or are even the main determinants of protein levels. By using tRNA modification enzyme mutants in *S. cerevisiae*, Chou et al., [80] have demonstrated that the number of translationally regulated genes is quite small (57 cases). In another recent study, the careful analysis of RA and TLRi correlations with PA in the same yeast found fewer than 200 proteins, apart from the general tendency of strict correlation between RA and PA [81]. All these results suggest that translational regulation is statistically uncommon and is therefore, globally, a minor participant in eukaryotic (and possibly prokaryotic) gene expression regulation. However, it is worth noting that in the four organisms studied herein, the data were obtained from actively growing cells. Nonetheless, it is rather possible that the respective importance of the synthesis rates and stabilities vary under no active or very slow growth conditions.

Another interesting topic is the positive correlation of the individual translation rate of each mRNA molecule (TLRi) and the protein level (PA). If TLRi were the same for all the mRNAs, the correlation between RA and PA would still be observed. Therefore, the additional positive correlation of TLRi with RA and PA in all four organisms means that more abundant mRNAs tend to be specifically better translated. This can be explained using the known enrichment of abundant mRNAs in optimal codons, which makes them more stable and translatable [71,72]. Another possibility would be that transcription

imprints mRNAs at a level that depends on the actual TR, in which case mRNAs could be more or less translatable (TLRi) according to their TR. This has been shown to occur in mammalian cells, where methylation of adenines in position 6 (me6A) is greater for less transcribed mRNAs and this feature is detrimental for their translation [82]. This cannot be a reason for the potentiation effect on microorganisms because in those organisms, me6A is not present in *S. pombe* and *E. coli* and occurs only during meiosis in *S. cerevisiae* [83]. Whatever the molecular reason, this can be considered a "potentiation" of the effect of RA on PA because it provokes the exponential amplification of mRNA abundance [8]. Finally, our study indicates that stabilities in mRNAs and especially proteins seem to play minor roles in gene regulation in general (as previously stated by [9,62]). Nevertheless, they may play important roles for some genes [12] or during dynamic processes such as cell differentiation (discussed in [4,76]).

As the four studied organisms showed 6VP gene clusters that are statistically enriched in genes by behaving as specific functional categories, we reaffirm the conclusion that we drew in our previous study on *S. cerevisiae*: all kinds of cells use common expression strategies for the genes acting in the same physiological pathways. A seminal work from P. Brown's laboratory [84] found that genes belonging to same pathways, especially those belonging to stable stoichiometric protein complexes, tend to have similar RS values. Our current study extends this idea to other organisms and to the rest of the gene expression variables. However, CES are not always identical between organisms, which suggests that evolution has adapted the particular expression strategies to the life styles and cell organizations of different species.

These differences in CESs between species prompted us to employ a quantitative parameter to classify organisms and to make 6VP phenograms. We are aware that a phenogram based on four organisms provides limited information because it has relatively few possible topologies. In the future, it would be necessary to use more organisms to draw more in-depth conclusions. In any case, the topology of the 6VP phenogram is identical to the phylogenetic tree obtained by the sequence comparison of 16/18S rRNA genes or from the whole genome content [85,86]. This demonstrates the phylogenetic consistency of our functional clustering, although the relative distance between the prokaryotes and eukaryotes (compared to the internal distances between eukaryotes) is clearly shorter than in sequence-based phylogenetic trees. This can reflect the fact that DNA sequences have diverged much more than gene functions and gene regulatory mechanisms [87]. A previous study [88] constructed a phenetic dendogram based on antibiotic resistance and concluded that it should be seen as the outward manifestation of the evolutionary history of the translational apparatus. In the same line, we conclude here that our 6VP dendogram is a global meter of the evolutionary history of whole cell functions. However, we recognize that the results from the bacterium *E. coli* are not as relevant as those obtained from the three eukaryotes, but we think that this is due mainly the poor-quality GO annotation, which lowered the number of analyzable GO terms. In any case, we used *E. coli* as an outgroup for the comparative analyses of eukaryotes.

When comparing the three eukaryotes, namely two free-living yeasts separated by about 330–600 My of evolution [19,21,22] and a multicellular higher eukaryote separated from yeasts by about 1000–1100 My [19,21], the 6VP dendograms illustrate the respective influence of gene sequence evolution (nucleotide or amino acid conservation) and gene expression evolution (similarity of 6VPs). Once again, the distance between HeLa and yeasts is relatively shorter than that predicted upon the evolutionary distance basis. In fact, the high success rate in the systematic functional replacement of essential yeast genes with their human counterparts [87] has already suggested that gene functions and regulatory mechanisms are more conserved than gene sequences. As for human cells, it is clear that HeLa cells represent a special kind of cells, which are useful for performing omics studies but which are not representative of all kinds of human cells, especially of those with a much lower or even zero growth rate. Our study thus reflects the similarities between eukaryotic cells growing at their fastest capability.

Regarding the similarity of the three eukaryotes in different cell components and major physiological functions, it is interesting to see how our 6VP phenogram analysis (Figure 5B, Supplementary Figure S2) suggests that some components and functions have closer gene expression strategies than others. As we showed the identical scaling for both UPGMA and NJ trees, it became evident how some broad GO groups show globally less 6VP differences in the three organisms than others. Thus, in the major macromolecular synthesis related to the central dogma, it would seem that translation (cytoplasmic translation and ribosome biogenesis) has more regulatory strategy similarity than replication or transcription (including splicing). It is necessary to point out, however, that GO categories have been established arbitrarily by human curators [89], and it is not easy to be sure about the homogeneity of them all being identical. For instance, our transcription broad group includes many GO terms related to the transcription process, including transcription factors, RNA polymerases, chromatin modifiers and others, which are perhaps broader than the GO terms included in our "cytoplasmic translation" group. In any case, we think that this kind of analysis may open a new window to investigate the evolution of cellular functions.

The two yeasts were the most closely related organisms based on 6VP similarity (Figure 5). It has been argued that they are almost as different from one another as from animals [19]. Yet in spite of this statement, the evolutionary time-distance between them is about half of that which both have in regard to animals [18,20]. Moreover, as they have very similar lifestyles [20], a convergence in gene regulatory strategies would seem logical. In spite of being separated from the common ancestor with humans at the same time, globally, gene expression strategies are closer to HeLa in *S. cerevisiae* than in *S. pombe*. This was unexpected, because it has often been stated that *S. pombe* is more similar to higher eukaryotes as the cell cycle is more similar, and it has many more introns than budding yeast and an RNAi mechanism, although it has no well-developed peroxisomes and proliferates mainly in the haploid state, whereas *S. cerevisiae* has peroxisomes and in the wild it proliferates as diploid [22]. However, it is necessary to point out that our study deals with a different matter—gene regulatory mechanisms—which could have evolved to converge between budding yeast and human cells. This suggestion can be supported by the hypothesis that *S. pombe* is a more "ancient" yeast than *S. cerevisiae* based on its biological features because it appears to have undergone fewer evolutionary changes since diverging from their common ancestor [22].

We conclude that the comparative analysis of all the variables affecting the flow of gene expression is a useful strategy for investigating the regulatory strategies used by cells, at least by eukaryotes. We also conclude with our study that the transcription rate is the main determinant of the amount of the corresponding protein, and that all kinds of organisms, either prokaryotic, single-cell or higher eukaryotic cells, use CESs for genes acting on the same physiological pathways. This feature can be reflected as a 6VP that defines the average behavior of a given gene group. CESs are more clearly seen for genes coding for large and stable protein complexes, such as the ribosome or the spliceosome, but can be seen even in groups of genes that do not form stable complexes, but which are functionally related, like those involved in different metabolic pathways. Some of the 6VPs are similar between different organisms, which reflects either a common evolutionary origin or a convergent evolution due to similar functional constraints or horizontal gene transfers. We propose that comparing 6VPs for a series of organisms is another way to draw phenograms to reveal how the environment of cells influences gene expression strategies. The use of omics data for phenetic classifications is not new [78,88,90], but our analysis extends the available tools for phenetic classifications by allowing researchers to study global expression strategies adapted to lifestyle.

Abbreviations

6VP	six variable profiles
ANOV	Aanalysis of variance
BMA	bayesian model averaging
BP	biological process
CC	cellular component
CES	common expressions strategies
CHX	cycloheximide
DTA	dynamic transcriptome analysis
FDR	false discovery rate
GFP	green fluorescent protein
GO	gene ontology
GRO	genomic run on
HUGO	human genome organization
iBAQ	intensity-based absolute quantification
KNN	k-nearest neighbor method
LB	Luria-Bertani medium
LC-MS/MS	liquid chromatography-tandem mass spectrometry
MIPS	Munich information center for protein sequences
NJ	neighbor-joining
PA	proteín abundance
PCN	protein copies per cell
PS	proteín stability
RA	mRNA abundance

RATE-seq	RNA approach to equilibrium sequencing
REVIGO	reduce + visualize gene ontology
RiBi	ribosome biogenesis
RNA-seq	RNA sequencing
RS	mRNA stability
SGD	Saccharomyces Genome Database
SILAC	stable isotope labeling by/with amino acids in cell culture
SOTA	self-organizing tree algorithm
TLR	translation rate
TLRi	individual translation rate
TR	transcription rate
UPGMA	unweighted pair group method with arithmetic mean

References

1. Crick, F.H.C. On protein synthesis. *Symp. Soc. Exp. Biol.* **1958**, *13*, 138–163.
2. Laurent, J.M.; Vogel, C.; Kwon, T.; Craig, S.A.; Boutz, D.R.; Huse, H.K.; Nozue, K.; Walia, H.; Whiteley, M.; Ronald, P.C.; et al. Protein abundances are more conserved than mRNA abundances across diverse taxa. *Proteomics* **2010**, *10*, 4209–4212. [CrossRef]
3. García-Martínez, J.; González-Candelas, F.; Pérez-Ortín, J.E. Common gene expression strategies revealed by genome-wide analysis in yeast. *Genome Biol.* **2007**, *8*. [CrossRef] [PubMed]
4. Liu, Y.; Beyer, A.; Aebersold, R. On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell* **2016**, *165*, 535–550. [CrossRef] [PubMed]
5. Csárdi, G.; Franks, A.; Choi, D.S.; Airoldi, E.; Drummond, D.A. Accounting for Experimental Noise Reveals That mRNA Levels, Amplified by Post-Transcriptional Processes, Largely Determine Steady-State Protein Levels in Yeast. *PLoS Genet.* **2015**, *11*, e1005206. [CrossRef]
6. Lahtvee, P.-J.; Sánchez, B.J.; Smialowska, A.; Kasvandik, S.; Elsemman, I.E.; Gatto, F.; Nielsen, J. Absolute Quantification of Protein and mRNA Abundances Demonstrate Variability in Gene-Specific Translation Efficiency in Yeast. *Cell Syst.* **2017**, *4*, 495–504.e5. [CrossRef]
7. Li, J.J.; Bickel, P.J.; Biggin, M.D. System wide analyses have underestimated protein abundances and the importance of transcription in mammals. *PeerJ* **2014**, *2*, e270. [CrossRef]
8. Li, J.J.; Chew, G.-L.; Biggin, M.D. Quantitating translational control: mRNA abundance-dependent and independent contributions and the mRNA sequences that specify them. *Nucleic Acids Res.* **2017**, *45*, 11821–11836. [CrossRef]
9. Li, J.J.; Biggin, M.D. Statistics requantitates the central dogma. *Science* **2015**, *347*, 1066–1067. [CrossRef]
10. McManus, J.; Cheng, Z.; Vogel, C. Next-generation analysis of gene expression regulation–comparing the roles of synthesis and degradation. *Mol. BioSyst.* **2015**, *11*, 2680–2689. [CrossRef] [PubMed]
11. Vogel, C.; Marcotte, E.M. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat. Rev. Genet.* **2012**, *13*, 227–232. [CrossRef]
12. Belle, A.; Tanay, A.; Bitincka, L.; Shamir, R.; O'Shea, E.K. Quantification of protein half-lives in the budding yeast proteome. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 13004–13009. [CrossRef]
13. Pérez-Ortín, J.E.; Tordera, V.; Chávez, S. Homeostasis in the Central Dogma of molecular biology: The importance of mRNA instability. *RNA Biol.* **2019**, *16*, 1659–1666. [CrossRef] [PubMed]
14. Pérez-Ortín, J.E.; Alepuz, P.M.; Moreno, J. Genomics and gene transcription kinetics in yeast. *Trends Genet.* **2007**, *23*, 250–257. [CrossRef] [PubMed]
15. Newman, J.R.S.; Ghaemmaghami, S.; Ihmels, J.; Breslow, D.K.; Noble, M. Single-cell proteomic analysis of S. cerevisiae re-veals the architecture of biological noise. *Nature* **2006**, *441*, 840–846. [CrossRef] [PubMed]
16. Fraser, H.B.; Hirsh, A.E.; Giaever, G.; Kumm, J.; Eisen, M.B. Noise Minimization in Eukaryotic Gene Expression. *PLoS Biol.* **2004**, *2*, e137. [CrossRef] [PubMed]
17. Bar-Even, A.; Paulsson, J.; Maheshri, N.; Carmi, M.; Oshea, E.K.; Pilpel, Y.; Barkai, N. Noise in protein expression scales with natural protein abundance. *Nat. Genet.* **2006**, *38*, 636–643. [CrossRef]
18. Baudrimont, A.; Jaquet, V.; Wallerich, S.; Voegeli, S.; Becskei, A. Contribution of RNA Degradation to Intrinsic and Extrinsic Noise in Gene Expression. *Cell Rep.* **2019**, *26*, 3752–3761.e5. [CrossRef]
19. Sipiczki, M. Where does fission yeast sit on the tree of life? *Genome Biol.* **2000**, *1*, 101. [CrossRef]
20. Dujon, B. Yeast evolutionary genomics. *Nat. Rev. Genet.* **2010**, *11*, 512–524. [CrossRef]
21. Douzery, E.J.; Snell, E.A.; Bapteste, E.; Delsuc, F.; Philippe, H. The timing of eukaryotic evolution: Does a relaxed molecular clock reconcile proteins and fossils? *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 15386–15391. [CrossRef] [PubMed]
22. Hoffman, C.S.; Wood, V.; Fantes, P.A. An ancient yeast for young geneticists: A Primer on the Schizosaccharomyces pombe model system. *Genetics* **2016**, *201*, 403–423, Erratum in **2016**, *202*, 1241. [CrossRef] [PubMed]
23. Kim, H.S.; Sung, Y.-J.; Paik, S. Cancer Cell Line Panels Empower Genomics-Based Discovery of Precision Cancer Medicine. *Yonsei Med. J.* **2015**, *56*, 1186–1198. [CrossRef]

24. Hastie, T.; Tibshirani, R.; Narasimhan, B.; Chu, G. Impute: Imputation for Microarray Data. R package version 1.64.0. 2020. Available online: https://www.bioconductor.org/packages/release/bioc/html/impute.html (accessed on 1 February 2021).

25. Miller, C.; Schwalb, B.; Maier, K.; Schulz, D.; Dümcke, S.; Zacher, B.; Mayer, A.; Sydow, J.; Marcinowski, L.; Dölken, L.; et al. Dynamic transcriptome analysis measures rates of mRNA synthesis and decay in yeast. *Mol. Syst. Biol.* **2011**, *7*, 458. [CrossRef] [PubMed]

26. Neymotin, B.; Athanasiadou, R.; Gresham, D. Determination of in vivo RNA kinetics using RATE-seq. *RNA* **2014**, *20*, 1645–1652. [CrossRef]

27. Siwiak, M.; Zielenkiewicz, P. A Comprehensive, Quantitative, and Genome-Wide Model of Translation. *PLoS Comput. Biol.* **2010**, *6*, e1000865. [CrossRef] [PubMed]

28. Christiano, R.; Nagaraj, N.; Fröhlich, F.; Walther, T.C. Global Proteome Turnover Analyses of the Yeasts S. cerevisiae and S. pombe. *Cell Rep.* **2014**, *9*, 1959–1965. [CrossRef] [PubMed]

29. Chong, Y.T.; Koh, J.L.; Friesen, H.; Duffy, S.K.; Cox, M.J.; Moses, A.M.; Moffat, J.; Boone, C.; Andrews, B.J. Yeast Proteome Dynamics from Single Cell Imaging and Automated Analysis. *Cell* **2015**, *161*, 1413–1424. [CrossRef] [PubMed]

30. Pelechano, V.; Ch&vez, S.; Pérez-Ortín, J.E. A complete set of nascent transcription rates for yeast genes. *PLoS ONE* **2010**, *5*, e15442. [CrossRef] [PubMed]

31. Ingolia, N.T.; Ghaemmaghami, S.; Newman, J.R.S.; Weissman, J.S. Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. *Science* **2009**, *324*, 218–223. [CrossRef] [PubMed]

32. Amorim, M.J.; Cotobal, C.; Duncan, C.; Mata, J. Global coordination of transcriptional control and mRNA decay during cellular differentiation. *Mol. Syst. Biol.* **2010**, *6*, 380. [CrossRef]

33. Eser, P.; Wachutka, L.; Maier, K.C.; Demel, C.; Boroni, M.; Iyer, S.; Cramer, P.; Gagneur, J. Determinants of RNA metabolism in the Schizosaccharomyces pombe genome. *Mol. Syst. Biol.* **2016**, *12*, 857. [CrossRef]

34. Sun, M.; Schwalb, B.; Schulz, D.; Pirkl, N.; Etzold, S.; Larivière, L.; Maier, K.C.; Seizl, M.; Tresch, A.; Cramer, P. Comparative dynamic transcriptome analysis (cDTA) reveals mutual feedback between mRNA synthesis and degradation. *Genome Res.* **2012**, *22*, 1350–1359. [CrossRef] [PubMed]

35. Marguerat, S.; Schmidt, A.; Codlin, S.; Chen, W.; Aebersold, R.; Bähler, J. Quantitative Analysis of Fission Yeast Transcriptomes and Proteomes in Proliferating and Quiescent Cells. *Cell* **2012**, *151*, 671–683. [CrossRef] [PubMed]

36. Lackner, D.H.; Beilharz, T.H.; Marguerat, S.; Mata, J.; Watt, S.; Schubert, F.; Preiss, T.; Bähler, J. A Network of Multiple Regulatory Layers Shapes Gene Expression in Fission Yeast. *Mol. Cell* **2007**, *26*, 145–155. [CrossRef] [PubMed]

37. Gunaratne, J.; Schmidt, A.; Quandt, A.; Neo, S.P.; Saraç, O.S.; Gracia, T.; Loguercio, S.; Ahrné, E.; Xia, R.L.; Tan, K.H.; et al. Extensive mass spectrometry-based analysis of the fission yeast proteome: The Schizosaccharomyces pombe PeptideAtlas. *Mol Cell Proteom.* **2013**, *12*, 1741–1751. [CrossRef] [PubMed]

38. Tani, H.; Mizutani, R.; Salam, K.A.; Tano, K.; Ijiri, K.; Wakamatsu, A.; Isogai, T.; Suzuki, Y.; Akimitsu, N. Genome-wide determination of RNA stability reveals hundreds of short-lived noncoding transcripts in mammals. *Genome Res.* **2012**, *22*, 947–956. [CrossRef]

39. Nagaraj, N.; Wisniewski, J.R.; Geiger, T.; Cox, J.; Kircher, M.; Kelso, J.; Pääbo, S.; Mann, M. Deep proteome and transcriptome mapping of a human cancer cell line. *Mol. Syst. Biol.* **2011**, *7*, 548. [CrossRef] [PubMed]

40. Uhlén, M.; Hallström, B.M.; Lindskog, C.; Mardinoglu, A.; Pontén, F.; Nielsen, J. Transcriptomics resources of human tissues and organs. *Mol. Syst. Biol.* **2016**, *12*, 862.

41. Cambridge, S.B.; Gnad, F.; Nguyen, C.; Bermejo, J.L.; Krüger, M.; Mann, M. Systems-wide Proteomic Analysis in Mammalian Cells Reveals Conserved, Functional Protein Turnover. *J. Proteome Res.* **2011**, *10*, 5275–5284. [CrossRef] [PubMed]

42. Wiśniewski, J.R.; Ostasiewicz, P.; Duś, K.; Zielińska, D.F.; Gnad, F.; Mann, M. Extensive quantitative remodeling of the proteome between normal colon tissue and adenocarcinoma. *Mol. Syst. Biol.* **2012**, *8*, 611. [CrossRef] [PubMed]

43. Kulak, N.A.; Pichler, G.; Paron, I.; Nagaraj, N.; Mann, M. Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nat. Methods* **2014**, *11*, 319–324. [CrossRef]

44. Esquerré, T.; Laguerre, S.; Turlan, C.; Carpousis, A.J.; Girbal, L.; Cocaign-Bousquet, M. Dual role of transcription and transcript stability in the regulation of gene expression in Escherichia coli cells cultured on glucose at different growth rates. *Nucleic Acids Res.* **2013**, *42*, 2460–2472. [CrossRef] [PubMed]

45. Esquerré, T.; Moisan, A.; Chiapello, H.; Arike, L.; Vilu, R.; Gaspin, C.; Cocaign-Bousquet, M.; Girbal, L. Genome-wide investigation of mRNA lifetime determinants in Escherichia coli cells cultured at different growth rates. *BMC Genom.* **2015**, *16*, 275. [CrossRef] [PubMed]

46. Li, G.-W.; Burkhardt, D.; Gross, C.; Weissman, J.S. Quantifying Absolute Protein Synthesis Rates Reveals Principles Underlying Allocation of Cellular Resources. *Cell* **2014**, *157*, 624–635. [CrossRef]

47. Valgepea, K.; Adamberg, K.; Seiman, A.; Vilu, R. Escherichia coli achieves faster growth by increasing catalytic and translation rates of proteins. *Mol. Biosyst.* **2013**, *9*, 2344–2358. [CrossRef] [PubMed]

48. Schmidt, A.; Kochanowski, K.; Vedelaar, S.; Ahrné, E.; Volkmer, B.; Callipo, L.; Knoops, K.; Bauer, M.; Aebersold, R.; Heinemann, M. The quantitative and condition-dependent Escherichia coli proteome. *Nat. Biotechnol.* **2016**, *34*, 104–110. [CrossRef]

49. Keseler, I.M.; Mackie, A.; Santos-Zavaleta, A.; Billington, R.; Bonavides-Martínez, C.; Caspi, R.; Fulcher, C.; Gama-Castro, S.; Kothari, A.; Krummenacker, M.; et al. The EcoCyc database: Reflecting new knowledge aboutEscherichia coliK-12. *Nucleic Acids Res.* **2017**, *45*, D543–D550. [CrossRef] [PubMed]

50. Steel, M.F.J. Model Averaging and Its Use in Economics. *J. Econ. Lit.* **2020**, *58*, 644–719. [CrossRef]

51. Garcia-Donato, G.; Forte-Deltell, A. Bayesian testing, variable selection and model averaging in linear models using R with BayesVarSel. *R J.* **2018**, *10*, 155–174. [CrossRef]

52. Brock, G.; Pihur, V.; Datta, S.; Datta, S. clValid: An R package for cluster validation. *J. Statist. Softw.* **2008**, *25*, 4. [CrossRef]

53. Herrero, J.; Valencia, A.; Dopazo, J. A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics* **2001**, *17*, 126–136. [CrossRef] [PubMed]

54. Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate—A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B-Methodol.* **1995**, *57*, 289–300. [CrossRef]

55. Yu, G. Gene Ontology Semantic Similarity Analysis Using GOSemSim. *Toxicity Assess.* **2020**, *2117*, 207–215. [CrossRef]

56. Resnik, P. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *J. Artif. Intell. Res.* **1999**, *11*, 95–130. [CrossRef]

57. Lin, D. An information-theoretic definition of similarity. In Proceedings of the 15th International Conference on Machine Learning, Madison, WI, USA, 24–27 July 1998; pp. 296–304.

58. Schliep, K.P. phangorn: Phylogenetic analysis in R. *Bioinformstics* **2010**, *27*, 592–593. [CrossRef] [PubMed]

59. Landry, J.J.M.; Pyl, P.T.; Rausch, T.; Zichner, T.; Tekkedil, M.M.; Stütz, A.M.; Jauch, A.; Aiyar, R.S.; Pau, G.; Delhomme, N.; et al. The Genomic and Transcriptomic Landscape of a HeLa Cell Line. *G3 Genes Genomes Genet.* **2013**, *3*, 1213–1224. [CrossRef]

60. Fortelny, N.; Overall, C.M.; Pavlidis, P.; Freue, G.V.C. Can we predict protein from mRNA levels? *Nat. Cell Biol.* **2017**, *547*, E19–E20. [CrossRef]

61. Wilhelm, M.; Hahne, H.; Savitski, M.; Marx, H.; Lemeer, S.; Bantscheff, M.; Kuster, B. Wilhelm et al. reply. *Nat. Cell Biol.* **2017**, *547*, E23. [CrossRef]

62. Schwanhäusser, B.; Busse, D.; Li, N.; Dittmar, G.; Schuchhardt, J.; Wolf, J.; Chen, W.; Selbach, M. Global quantification of mammalian gene expression control. *Nat. Cell Biol.* **2011**, *473*, 337–342. [CrossRef]

63. Carneiro, R.L.; Requião, R.D.; Rossetto, S.; Domitrovic, T.; Palhano, F.L. Codon stabilization coefficient as a metric to gain insights into mRNA stability and codon bias and their relationships with translation. *Nucleic Acids Res.* **2019**, *47*, 2216–2228. [CrossRef]

64. Chan, L.Y.; Mugler, C.F.; Heinrich, S.; Vallotton, P.; Weis, K. Non-invasive measurement of mRNA decay reveals translation initiation as the major determinant of mRNA stability. *eLife* **2018**, *7*. [CrossRef] [PubMed]

65. Grigull, J.; Mnaimneh, S.; Pootoolal, J.; Robinson, M.D.; Hughes, T.R. Genome-Wide Analysis of mRNA Stability Using Transcription Inhibitors and Microarrays Reveals Posttranscriptional Control of Ribosome Biogenesis Factors. *Mol. Cell. Biol.* **2004**, *24*, 5534–5547. [CrossRef]

66. Pérez-Ortín, J.E.; Alepuz, P.; Chávez, S.; Choder, M. Eukaryotic mRNA Decay: Methodologies, Pathways, and Links to Other Stages of Gene Expression. *J. Mol. Biol.* **2013**, *425*, 3750–3775. [CrossRef] [PubMed]

67. Eshleman, N.; Luo, X.; Capaldi, A.; Buchan, J.R. Alterations of signaling pathways in response to chemical perturbations used to measure mRNA decay rates in yeast. *RNA* **2020**, *26*, 10–18. [CrossRef] [PubMed]

68. Russo, J.; Heck, A.M.; Wilusz, J.; Wilusz, C.J. Metabolic labeling and recovery of nascent RNA to accurately quantify mRNA stability. *Methods* **2017**, *120*, 39–48. [CrossRef] [PubMed]

69. Baudrimont, A.; Voegeli, S.; Viloria, E.C.; Stritt, F.; Lenon, M.; Wada, T.; Jaquet, V.; Becskei, A. Multiplexed gene control reveals rapid mRNA turnover. *Sci. Adv.* **2017**, *3*, e1700006. [CrossRef]

70. Furlan, M.; De Pretis, S.; Pelizzola, M. Dynamics of transcriptional and post-transcriptional regulation. *Brief. Bioinform.* **2020**, bbaa389. [CrossRef] [PubMed]

71. Hanson, G.; Coller, J. Codon optimality, bias and usage in translation and mRNA decay. *Nat. Rev. Mol. Cell Biol.* **2018**, *19*, 20–30. [CrossRef]

72. Presnyak, V.; Alhusaini, N.; Chen, Y.-H.; Martin, S.; Morris, N.; Kline, N.; Olson, S.; Weinberg, D.; Baker, K.E.; Graveley, B.R.; et al. Codon Optimality Is a Major Determinant of mRNA Stability. *Cell* **2015**, *160*, 1111–1124. [CrossRef] [PubMed]

73. Harigaya, Y.; Parker, R. Analysis of the association between codon optimality and mRNA stability in Schizosaccharomyces pombe. *BMC Genom.* **2016**, *17*, 895. [CrossRef]

74. Hu, J.C.; Karp, P.D.; Keseler, I.M.; Krummenacker, M.; Siegele, D.A. What we can learn about Escherichia coli through application of Gene Ontology. *Trends Microbiol.* **2009**, *17*, 269–278. [CrossRef]

75. Krebs, J.E.; Goldstein, E.S.; Kilpatrick, S.T. *Lewin's Genes XII*, 12th ed.; Jones & Burlett Learning: Burlington, MA, USA, 2018.

76. Kristensen, A.R.; Gsponer, J.; Foster, L.J. Protein synthesis rate is the predominant regulator of protein expression during differentiation. *Mol. Syst. Biol.* **2013**, *9*, 689. [CrossRef] [PubMed]

77. Supek, F.; Bošnjak, M.; Škunca, N.; Šmuc, T. REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. *PLoS ONE* **2011**, *6*, e21800. [CrossRef] [PubMed]

78. House, C.H. The Tree of Life Viewed Through the Contents of Genomes. *Toxic. Assess.* **2009**, *532*, 141–161. [CrossRef]

79. García-Martínez, J.; Aranda, A.; Pérez-Ortín, J.E. Genomic Run-On Evaluates Transcription Rates for All Yeast Genes and Identifies Gene Regulatory Mechanisms. *Mol. Cell* **2004**, *15*, 303–313. [CrossRef] [PubMed]

80. Chou, H.-J.; Donnard, E.; Gustafsson, H.T.; Garber, M.; Rando, O.J. Transcriptome-wide Analysis of Roles for tRNA Modifications in Translational Regulation. *Mol. Cell* **2017**, *68*, 978–992.e4. [CrossRef]

81. Ho, B.; Baryshnikova, A.; Brown, G.W. Unification of Protein Abundance Datasets Yields a Quantitative Saccharomyces cerevisiae Proteome. *Cell Syst.* **2018**, *6*, 192–205.e3. [CrossRef] [PubMed]

82. Slobodin, B.; Han, R.; Calderone, V.; Vrielink, J.; Joachim, A.O.; Loayza-Puch, F.; Elkon, R.; Agami, R. Transcription Impacts the Efficiency of mRNA Translation via Co-transcriptional N6-adenosine Methylation. *Cell* **2017**, *169*, 326–337.e12. [CrossRef] [PubMed]

83. Schwartz, S.; Agarwala, S.D.; Mumbach, M.R.; Jovanovic, M.; Mertins, P.; Shishkin, A.; Tabach, Y.; Mikkelsen, T.S.; Satija, R.; Ruvkun, G.; et al. High-Resolution Mapping Reveals a Conserved, Widespread, Dynamic mRNA Methylation Program in Yeast Meiosis. *Cell* **2013**, *155*, 1409–1421. [CrossRef]

84. Wang, Y.; Liu, C.L.; Storey, J.D.; Tibshirani, R.J.; Herschlag, D.; Brown, P.O. Precision and functional specificity in mRNA decay. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 5860–5865. [CrossRef] [PubMed]

85. House, C.H.; Fitz-Gibbon, S.T. Using Homolog Groups to Create a Whole-Genomic Tree of Free-Living Organisms: An Update. *J. Mol. Evol.* **2002**, *54*, 539–547. [CrossRef]

86. Tekaia, F.; Lazcano, A.; Dujon, B. The Genomic Tree as Revealed from Whole Proteome Comparisons. *Genome Res.* **1999**, *9*, 550–557.

87. Kachroo, A.H.; Laurent, J.M.; Yellman, C.M.; Meyer, A.G.; Wilke, C.O.; Marcotte, E.M. Systematic humanization of yeast genes reveals conserved functions and genetic modularity. *Sci.* **2015**, *348*, 921–925. [CrossRef] [PubMed]

88. Briones, C.; Manrubia, S.C.; Lázaro, E.; Lazcano, A.; Amils, R. Reconstructing evolutionary relationships from functional data: A consistent classification of organisms based on translation inhibition response. *Mol. Phylogenet. Evol.* **2005**, *34*, 371–381. [CrossRef] [PubMed]

89. Gaudet, P.; Dessimoz, C. Gene Ontology: Pitfalls, Biases, and Remedies. In *The Gene Ontology Handbook*; Methods in Molecular Biology; Dessimoz, C., Škunca, N., Eds.; Humana Press: New York, NY, USA, 2017; Volume 1446.

90. Gerstein, M.B.; Rozowsky, J.; Yan, K.-K.; Wang, D.; Cheng, C.; Brown, J.B.; Davis, C.A.; Hillier, L.W.; Sisu, C.; Li, J.J.; et al. Comparative analysis of the transcriptome across distant species. *Nat. Cell Biol.* **2014**, *512*, 445–448. [CrossRef] [PubMed]