



## Database tool

# YQFC: a web tool to compare quantitative biological features between two yeast gene lists

Wei-Sheng Wu<sup>1,\*</sup>, Lai-Ji Wang<sup>1</sup>, Han-Chen Yen<sup>1</sup> and Yan-Yuan Tseng<sup>2</sup>

<sup>1</sup>Department of Electrical Engineering, National Cheng Kung University, No.1, University Road, Tainan city, 70101 Taiwan and <sup>2</sup>Center for Molecular Medicine and Genetics, Wayne State University, School of Medicine, 3127 Scott Hall, 540 East Canfield, Detroit, MI 48201, USA

\*Corresponding author: Tel: +886-6-2757575 Ext: 62426; Fax: +886-6-2345482; Email: [wessonwu@mail.ncku.edu.tw](mailto:wessonwu@mail.ncku.edu.tw)

Citation details: Wu, W.-S., Wang, L.-J., Yen, H.-C., *et al.* YQFC: a web tool to compare quantitative biological features between two yeast gene lists. *Database* (2020) Vol. 00: article ID baaa076; doi:10.1093/database/baaa076

Received 1 June 2020; Revised 27 July 2020; Accepted 11 August 2020

## Abstract

Nowadays high-throughput omics technologies are routinely used in biological research. From the omics data, researchers can easily get two gene lists (e.g. stress-induced genes vs. stress-repressed genes) related to their biological question. The next step would be to apply enrichment analysis tools to identify distinct functional/regulatory features between these two gene lists for further investigation. Although various enrichment analysis tools are already available, two challenges remain to be addressed. First, most existing tools are designed to analyze only one gene list, so they cannot directly compare two gene lists. Second, almost all existing tools focus on identifying the enriched qualitative features (e.g. gene ontology [GO] terms, pathways, domains, etc.). Many quantitative features (e.g. number of mRNA isoforms of a gene, mRNA half-life, protein half-life, transcriptional plasticity, translational efficiency, etc.) are available in the yeast, but no existing tools provide analyses on these quantitative features. To address these two challenges, here we present Yeast Quantitative Features Comparator (YQFC) that can directly compare various quantitative features between two yeast gene lists. In YQFC, we comprehensively collected and processed 85 quantitative features from the yeast literature and yeast databases. For each quantitative feature, YQFC provides three statistical tests (t-test, U test and KS test) to test whether this quantitative feature is statistically different between the two input yeast gene lists. The distinct quantitative features identified by YQFC may help researchers to study the underlying molecular mechanisms that differentiate the two input yeast gene lists. We believe that YQFC is a useful tool to expedite the biological research that uses high-throughput omics technologies.

**Database URL:** <http://cosbi2.ee.ncku.edu.tw/YQFC/>

## Introduction

Nowadays high-throughput omics technologies (e.g. genomics, transcriptomics, proteomics, etc.) are routinely

used in biological research (1). From the omics data, researchers can easily get two gene lists (e.g. stress-induced genes vs. stress-repressed genes, drug-induced genes vs.

drug-repressed genes, hub genes vs. nonhub genes, genes with long mRNA half-life vs. genes with short mRNA half-life, etc.) related to their biological question. The ultimate goal is to understand the underlying molecular mechanisms that differentiate these two gene lists. To achieve this goal, researchers need to know which biological features are good candidates for further investigation. Since enrichment analysis tools can help researchers to identify possible biological features, which are worthy of pursue, the enrichment analysis tool development has become a popular research topic in the bioinformatics field (2).

Many enrichment analysis tools are available for analyzing yeast gene lists. First, DAVID (3) provides gene ontology enrichment, pathway enrichment, protein domain enrichment and chromosome enrichment analyses. Second, YeastMine (4) provides gene ontology enrichment, pathway enrichment and publication enrichment analysis. Third, modPhEA (5) provides phenotype enrichment analysis. Fourth, YGA (6) provides physical/genetic interaction enrichment and functional group enrichment analyses. Fifth, YHMI (7) provides histone modification enrichment analysis. Sixth, YARG (8) provides arsenic-related genes enrichment analysis. Seventh, YEASTRACT (9) provides single transcription factor regulation enrichment analysis. Eighth, YCRD (10) provides cooperative transcription factor pair regulation enrichment analysis. Descriptions of many other enrichment analysis tools could be found in a review paper (2).

Although many enrichment analysis tools are already available for testing various biological features, two challenges remain to be addressed. First, most existing tools are designed to analyze one given gene list, so they (e.g. YeastMine, YHMI, YARG, YEASTRACT, YCRD, etc.) cannot directly compare two given gene lists. Second, almost all existing tools (e.g. YeastMine, DAVID, modPhEA, YHMI, etc.) focus on finding the enriched qualitative features (e.g. gene ontology [GO] terms, pathways, domains, phenotypes, histone modifications, etc.). Many quantitative features (e.g. mRNA expression level, mRNA half-life, transcriptional plasticity, translational efficiency, protein abundance, protein half-life, UTR length, number of mRNA isoforms of a gene, etc.) are available in the yeast, but no existing tools provide analyses on these quantitative features. To address these two challenges, here we develop Yeast Quantitative Features Comparator (YQFC) to directly compare various quantitative features between two yeast gene lists.

In YQFC, we comprehensively collected and processed 85 quantitative features from the yeast literature and yeast databases. For each quantitative feature, YQFC provides three statistical tests (t-test, U test and KS test) to test

whether this quantitative feature is statistically different between the two given yeast gene lists. Both tables and figures (box plots and cumulative distribution function [CDF] plots) are provided to visualize the testing results.

## Construction and contents

### Collection of 85 quantitative features from 7 yeast publications and 6 yeast databases

We comprehensively collected and processed 85 quantitative features from 7 yeast publications (11–17) and 6 yeast databases (YeastMine, YeastNet (18), YAGM (19), SGD (20), BioGRID (21), SPELL (22)). We classified these 85 quantitative features into 4 categories: 12 gene features, 4 mRNA features, 52 protein features, and 17 network features (Table 1). The detailed source information of all quantitative features in YQFC is given in Supplementary Table 1.

YQFC collected and processed data of 85 quantitative features in yeast. For each yeast gene, the values of all these 85 quantitative features are stored. Users can download our unique and useful feature data from the Download page of YQFC.

### Three statistical tests (t-test, U test and KS test) used to compare a quantitative feature between two yeast gene lists

For each quantitative feature, YQFC provides three statistical tests (t-test, U test and KS test) to test whether this quantitative feature is statistically different between the two given yeast gene lists (denoted as  $L1$  and  $L2$ ). For example, assume that a user wants to compare the 5'UTR length between  $L1$  and  $L2$ . YQFC provides the following three statistical tests to do this task. First, the t-test (or called Student's t-test) is used to determine if the mean 5'UTR length of the genes in  $L1$  is statistically longer/shorter than the mean 5'UTR length of the genes in  $L2$  (23). Note that the t-test is a parametric test based on the assumption that the observed data come from normal distributions. Second, the U test (or called Mann-Whitney U test) is used to determine if the median 5'UTR length of the genes in  $L1$  is statistically longer/shorter than the median 5'UTR length of the genes in  $L2$  under the location shift assumption (24). Note that the U test is a nonparametric test that does not assume anything about the distribution of the underlying populations. Third, the KS test (or called Kolmogorov-Smirnov test) is used to determine if the cumulative distribution of the 5'UTR length of the genes in  $L1$  is statistically larger/smaller than the cumulative distribution of the 5'UTR length of the genes in  $L2$  (24). Note that the KS test is a nonparametric test that does not

**Table 1.** The 85 collected quantitative features could be divided into 4 categories: 12 gene features, 4 mRNA features, 52 protein features and 17 network features

12 Gene features	CDS length, 5'UTR length, 3'UTR length, Number of publications, number of GO terms, number of GO slim terms, number of pathways, number of mutant phenotypes, number of mRNA isoforms, number of transcriptional regulators, number of fungal homologs, number of nonfungal and <i>Saccharomyces cerevisiae</i> homologs
4 mRNA features	mRNA level (3 datasets <sup>a</sup> ), mRNA half-life, transcriptional plasticity, translational efficiency (5 datasets <sup>b</sup> )
52 protein features	Number of domains, number of PTMs, protein half-life, amino acid composition (20 kinds <sup>c</sup> ), atomic composition (5 kinds <sup>d</sup> ), extinction coefficient at 280 nm (2 kinds <sup>e</sup> ), protein abundance (in the normal growth condition, 23 datasets <sup>f</sup> ), protein abundance (in various stress conditions, 11 kinds <sup>g</sup> ), protein physical details (protein length, molecular weight, pI, aliphatic index, instability index), coding region translation calculation (codon bias, codon adaptation index, frequency of optimal codons, hydrophaticity of protein, aromaticity score)
17 network features	Number of interactors in the following networks: PI (3 datasets <sup>h</sup> ) network GI network CC network CX network DC network GN (similar genomic context) network GT (similar profiles of GI partners) network PG (similar phylogenetic profiles) network TS (3-D protein structure of interacting orthologous proteins) network EPA network FAA network GIA network LEA network MPA network PIA network TFBA network TFRA network

CC, co-citation; CDS, coding sequence; CX, co-expression; DC, domain co-occurrence; EPA, expression profile association; FAA, functional annotation association; GI, genetic interaction; GIA, genetic interaction association; LEA, literature evidence association; MPA, mutant phenotype association; pI, isoelectric point; PI, physical interaction; PIA, physical interaction association; PTM, posttranslational modification; TFBA, transcription factor-binding association; TFRA, transcription factor regulation association; UTR, untranslated region.

<sup>a</sup>We collected three datasets of mRNA levels generated from microarray (11) and RNA-seq (12).

<sup>b</sup>We downloaded five datasets of translational efficiency from Csárdi *et al.* (16) who collected them from the literature.

<sup>c</sup>Amino acid composition contains 20 kinds of amino acids.

<sup>d</sup>Atomic composition contains five kinds of atoms: carbon, hydrogen, nitrogen, oxygen, and sulfur.

<sup>e</sup>Extinction coefficient at 280 nm contains two kinds: (i) all cys residues appear as half cystines and (ii) no cys residues appear as half cystines.

<sup>f</sup>We downloaded 23 datasets of protein abundance measured in the normal growth condition from Ho *et al.* (17) who collected them from the literature.

<sup>g</sup>We downloaded the datasets of protein abundance measured in 11 different stress conditions (dithiothreitol, H<sub>2</sub>O<sub>2</sub>, hydroxyurea, methyl methanesulfonate, etc.) from Ho *et al.* (17) who collected them from the literature.

<sup>h</sup>We collected three datasets of protein-protein physical interaction from BioGRID (21) and YeastNet (18).

PI (physical interaction, 3 datasets<sup>h</sup>) network

GI (genetic interaction) network

CC (co-citation) network

CX (co-expression) network

DC (domain co-occurrence) network

GN (similar genomic context) network

GT (similar profiles of genetic interaction partners) network

PG (similar phylogenetic profiles) network

TS (3-D protein structure of interacting orthologous proteins) network

EPA (expression profile association) network

FAA (functional annotation association) network

GIA (genetic interaction association) network

LEA (literature evidence association) network

MPA (mutant phenotype association) network

PIA (physical interaction association) network

TFBA (transcription factor binding association) network

TFRA (transcription factor regulation association) network

assume anything about the distribution of the underlying populations.

### Implementation of the web interface of YQFC

Figure 1 illustrates the configuration of YQFC. The web interface of YQFC was developed in Python using the Django MTV framework. The 85 processed quantitative

feature data were deposited in MySQL. All tables, box plots, and CDF plots were produced by the JavaScript and feature-rich JavaScript libraries (jQuery, DataTables and Plotly.js) to visualize data on the webpage. Except for the main website (<http://cosbi2.ee.ncku.edu.tw/YQFC/>), we also have two backup sites (<http://cosbi4.ee.ncku.edu.tw/YQFC/> and <http://cosbi6.ee.ncku.edu.tw/YQFC/>).

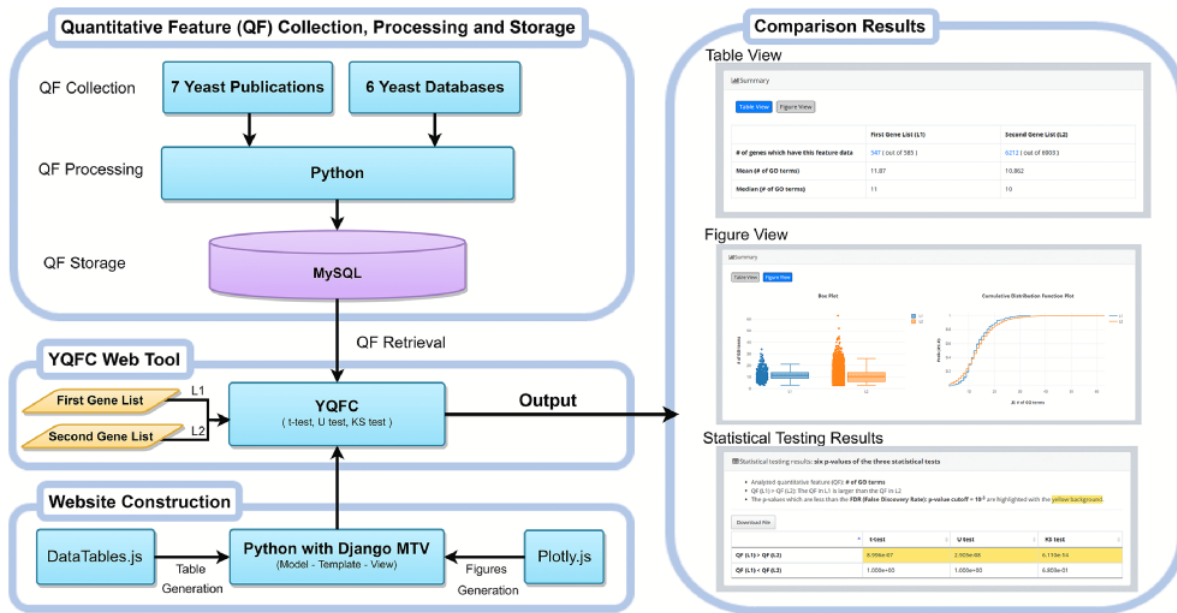


Figure 1. The configuration of YQFC.

### Compare Quantitative Features Between Two Yeast Gene Lists

Figure 2. The input page. To use YQFC, users have to go through a three-step process.

## Utility and discussion

### The usage of YQFC

YQFC is a web tool for identifying the distinct quantitative features between two input yeast gene lists. To use YQFC, users have to go through a three-step process

(Figure 2). [Step 1] Users need to input two gene lists to be compared. Standard names, systematic names, or aliases are all acceptable. If users only have one input gene list, they can use our precompiled gene lists (e.g. 6604 ORF genes, 299 tRNA genes, 27 rRNA genes, etc.) to serve as the second input gene list. In other words, YQFC also

## Comparison Results

**User's Specification**

# of genes in L1	585	# of genes in L2	6903
Multiple hypotheses testing	FDR (False Discovery Rate): p-value cutoff = $10^{-3}$		

**See the testing result of a chosen quantitative feature**

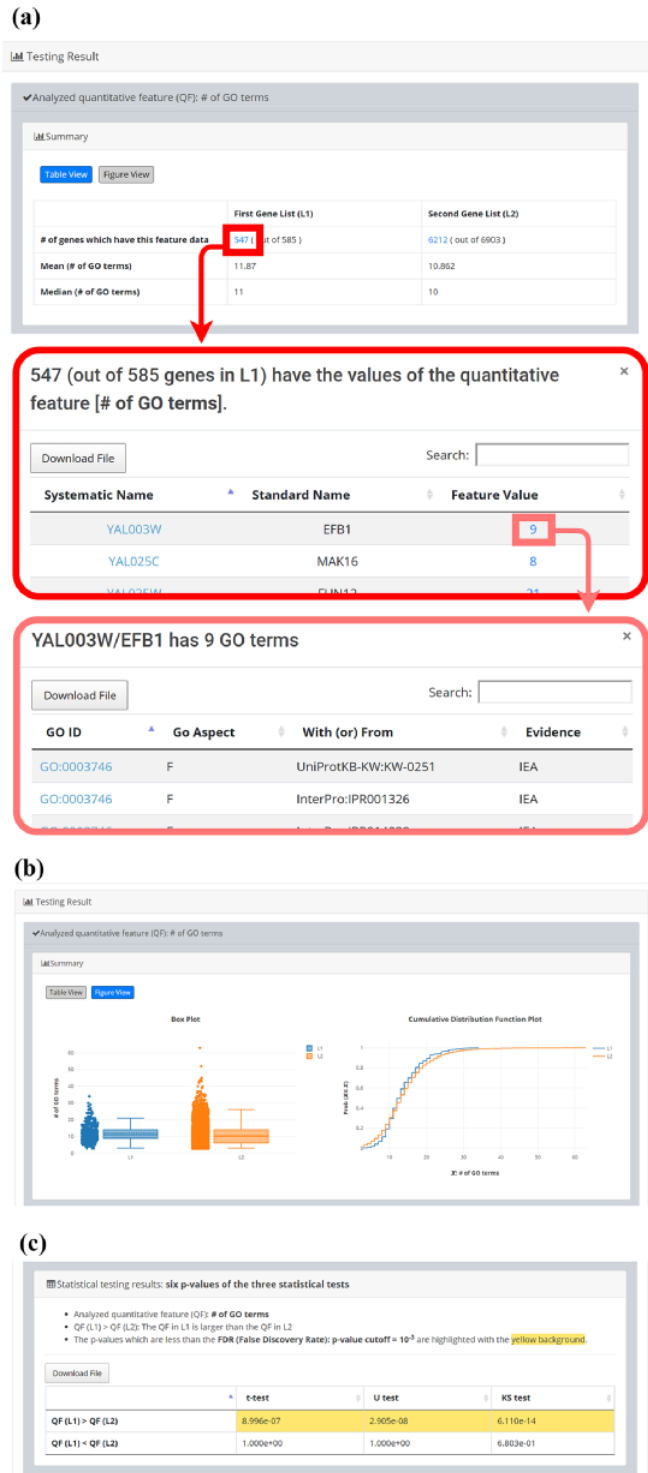
Gene Features	<input type="checkbox"/> CDS length <input type="checkbox"/> 5'UTR length <input type="checkbox"/> 3'UTR length <input type="checkbox"/> # of publications <input checked="" type="checkbox"/> # of GO terms <input type="checkbox"/> # of GO slim terms <input type="checkbox"/> # of pathways <input type="checkbox"/> # of mutant phenotypes <input type="checkbox"/> # of mRNA isoforms <input type="checkbox"/> # of transcriptional regulators <input type="checkbox"/> # of fungal homologs <input type="checkbox"/> # of non-fungal and S. cerevisiae homologs
mRNA Features	<input type="checkbox"/> mRNA level <input type="checkbox"/> mRNA half-life <input type="checkbox"/> Transcriptional plasticity <input type="checkbox"/> Translational efficiency
Protein Features	<input type="checkbox"/> Amino acid composition <input type="checkbox"/> Atomic composition <input type="checkbox"/> Protein abundance (in the normal growth condition) <input type="checkbox"/> Protein abundance (in various stress conditions) <input type="checkbox"/> Extinction coefficient at 280nm <input type="checkbox"/> # of protein domains <input type="checkbox"/> # of PTMs <input type="checkbox"/> Protein half-life <b>Protein Physical Details:</b> <input type="checkbox"/> Protein length <input type="checkbox"/> Molecular weight <input type="checkbox"/> Isoelectric point <input type="checkbox"/> Aliphatic index <input type="checkbox"/> Instability index <b>Coding Region Translation Calculations:</b> <input type="checkbox"/> Codon bias <input type="checkbox"/> Codon adaptation index <input type="checkbox"/> Frequency of optimal codons <input type="checkbox"/> Hydropathicity of protein <input type="checkbox"/> Aromaticity score
Network Features	<b># of interactors in the</b> <input type="checkbox"/> PI network <input type="checkbox"/> GI network <input type="checkbox"/> CC network <input type="checkbox"/> CX network <input type="checkbox"/> DC network <input type="checkbox"/> GN network <input type="checkbox"/> GT network <input type="checkbox"/> PG network <input type="checkbox"/> TS network <input type="checkbox"/> PIA network <input type="checkbox"/> GIA network <input type="checkbox"/> TFBA network <input type="checkbox"/> TFRA network <input type="checkbox"/> EPA network <input type="checkbox"/> FAA network <input type="checkbox"/> LEA network <input type="checkbox"/> MPA network

**Figure 3.** The result page (the first part). The first part contains the information of the user's settings (i.e. the number of genes in  $L1$ , the number of genes in  $L2$ , the selected quantitative features, the selected multiple hypotheses correction method, and the selected  $p$ -value threshold).

supports analysis of a single gene list. [Step 2] Users need to select the quantitative features to be analyzed. [Step 3] Since YQFC tests many quantitative features (i.e. multiple hypotheses testing), users have to select a statistical method (Bonferroni correction or false discovery rate (FDR)) for multiple hypotheses correction and set the  $P$ -value threshold. Note that using different correction method will give you different corrected  $P$ -value. Bonferroni correction is more conservative than FDR. That is, Bonferroni correction has a smaller type I error rate, resulting in a smaller power, than FDR does. The  $P$ -value threshold determines the statistical significance of how different of a quantitative feature is between the two input gene lists. The more stringent the  $P$ -value threshold, the higher the statistical significance of the identified distinct quantitative feature.

After submission, YQFC will perform three statistical tests (t-test, U test and KS test) to test whether a selected quantitative feature is statistically different between the two input gene lists (denoted as  $L1$  and  $L2$ ). This process will go through all the selected quantitative features. Once the analysis process is complete, YQFC will return the results with two parts. The first part (Figure 3) contains the information of the user's settings (i.e. the number of genes in  $L1$ , the number of genes in  $L2$ , the selected quantitative features, the selected multiple hypotheses correction method, and the selected  $P$ -value threshold). The second part contains the result of each selected quantitative feature shown as two sections: (i) Summary and (ii) Statistical testing results. In the 'Summary' section, users can choose

'Table View' or 'Figure View'. 'Table View' (Figure 4a) provides a table containing two kinds of information. First, the numbers of genes (in  $L1$  and  $L2$ , respectively), which have the feature values are given. If users click on the number, they will see the names and the feature values of these genes. By clicking on a feature value, users will see the original sources of the feature value. Second, the mean and median feature values of the genes (having feature values) in  $L1$  and  $L2$ , respectively, are given. 'Figure View' (Figure 4b) provides two kinds of plots for visualization. Box plots are used to display variation in the feature values in  $L1$  and  $L2$ , respectively. CDF plots are used to show the probability that the feature value  $X$  is less than or equal to a specific value  $x$  (i.e.  $Prob(X \leq x)$ ). In the 'Statistical testing results' section (Figure 4c), users can see a table with six  $P$ -values. Three  $P$ -values (calculated by t-test, U test and KS test) represent the statistical significance of claiming the quantitative feature of  $L1$  is larger than that of  $L2$  (denoted as  $QF(L1) > QF(L2)$ ). The other three  $P$ -values (calculated by t-test, U test and KS test) represent the statistical significance of claiming the quantitative feature of  $L1$  is smaller than that of  $L2$  (denoted as  $QF(L1) < QF(L2)$ ). To draw the users' attention, the  $P$ -values which are less than the  $P$ -value threshold are highlighted with the yellow background. To ensure the robustness of the testing results, we suggest the users to investigate the quantitative features with statistical significance in all three statistical tests (t-test, U test and KS test). As for the  $P$ -value threshold, we suggest the users to use 0.01 or less. The more stringent the  $P$ -value threshold,



**Figure 4.** The result page (the second part). The second part contains the result of each selected quantitative feature shown as two sections: (i) Summary and (ii) Statistical testing results. In the “Summary” section, users can choose “Table View” or “Figure View.” (a) “Table View” provides a table containing two kinds of information. First, the numbers of genes (in  $L1$  and  $L2$ , respectively) which have the feature values are given. If users click on the number, they will see the names and the feature values of these genes. By clicking on a feature value, users will see the original sources of the feature value. Second, the mean and median feature values of the genes (having feature values) in  $L1$  and  $L2$ , respectively, are given. (b) “Figure View” provides two kinds of plots for visualization. Box plots are used to display variation in the feature values in  $L1$  and  $L2$ , respectively. Cumulative distribution function (CDF) plots are used to show the probability that the feature value  $X$  is less than or equal to a specific value  $x$  (i.e.  $\text{Prob}(X \leq x)$ ). (c) In the “Statistical testing results” section, users can see a table with six  $p$ -values. Three  $p$ -values (calculated by t-test, U test, and KS test) represent the statistical significance of claiming the quantitative feature of  $L1$  is larger than that of  $L2$  (denoted as  $\text{QF}(L1) > \text{QF}(L2)$ ). The other three  $p$ -values (calculated by t-test, U test, and KS test) represent the statistical significance of claiming the quantitative feature of  $L1$  is smaller than that of  $L2$  (denoted as  $\text{QF}(L1) < \text{QF}(L2)$ ). To draw the users’ attention, the  $p$ -values which are less than the  $p$ -value threshold are highlighted with the yellow background.

the less testing results that can be called statistical significance.

### A case study

Microarray experiments have identified approximately 900 environmental stress response (ESR) genes whose mRNA levels respond to various environmental stresses (e.g. heat shock, oxidative stress, osmotic stress, etc.) in a stereotypical manner (25). These ESR genes can be divided into 585 repressed ESR (rESR) genes and 281 induced ESR (iESR) genes. Previous studies have shown that many rESR genes are housekeeping genes (e.g. ribosomal genes and genes involved in growth-related processes), while iESR genes are usually involved in various stress defense mechanisms (6, 25). We are interested in knowing which quantitative features can differentiate rESR genes from iESR genes. Therefore, we input these two gene lists (first list: 585 rESR genes and second list: 281 iESR genes) into our YQFC as a case study (Supplementary Table 2).

Many interesting observations have revealed from the analysis results of our YQFC. To ensure the statistical significance and robustness of the observations, we only discuss those with FDR-corrected  $P$ -values  $< 0.001$  in all three statistical tests (t-test, U test and KS test). First, in the normal growth condition, rESR genes have higher mRNA level, larger protein abundance and larger translational efficiency than iESR genes do (Table 2). These observations are not surprising. Since many rESR genes are housekeeping genes, their proteins are highly needed in the normal growth condition. On the contrary, iESR genes are involved in the stress defense mechanisms, so their proteins are not highly needed in the normal growth condition. Second, rESR genes also have larger protein abundance than iESR genes do (Table 2) in various stress conditions (e.g. dithiothreitol,  $H_2O_2$ , hydroxyurea, and methyl methanesulfonate). This indicates that even in the stress conditions, the needed quantity of housekeeping proteins are still larger than that of the stress response proteins to maintain the cell physiology. Third, in the normal growth condition, rESR genes have shorter mRNA half-life than iESR genes do (Table 3), consistent with a previous study reporting that ribosomal genes (belonging to rESR genes) have short mRNA half-life (26). Fourth, iESR genes have larger transcriptional plasticity, more transcriptional regulators and longer 5'UTR length than rESR genes do (Table 3). Since iESR proteins are only highly needed in the stress conditions, their expression must be under complicated controls. Having larger transcriptional plasticity and more transcriptional regulators indicate that the transcription of iESR genes may be more tightly regulated than that of rESR genes. Our finding is supported by a previous study reporting that the stress-induced genes are under a more complicated transcriptional

regulation than the housekeeping genes (27). Since 5'UTR is known to contain translational regulation signals (28), having longer 5'UTR suggests that the translation of iESR genes may be under more complicated regulation than that of rESR genes. For example, uORFs are one kind of translational regulation signals (29). It is known that the proportion of genes that contain uORFs in the 5'UTR is much greater for translationally up-regulated genes than down-regulated genes under various stress conditions (29). Therefore, we suspect that this observation is also hold for iESR genes vs. rESR genes. Indeed, we found that the proportion of genes that contain uORFs in the 5'UTR is much greater for iESR genes than rESR genes (Supplementary Table 3). Fifth, rESR genes have more GO term annotations, higher mRNA isoforms, and more interactors in the physical/genetic networks than iESR genes do. The biological meaning of these observations still needs further investigation. Several other quantitative features that can differentiate rESR genes from iESR genes could be found in Tables 2 and 3.

### Quantitative comparison vs. qualitative enrichment analysis

Since quantitative comparison between a single input gene list ( $L1$ ) and the background gene list ( $L2$ ) is possible in YQFC, users may want to correlate the quantitative comparison results with the qualitative enrichment analysis results. The qualitative enrichment analysis is to test the enrichment of a qualitative feature (e.g. GO terms) in the single input gene list compared to the background. Here we provide an example to illustrate the differences between these two testing strategies. Assume that we want to compare the 5'UTR length of 281 iESR genes and the background (4418 genes which have 5'UTR length data). For quantitative comparison, YQFC found that the 5'UTR length of 281 iESR genes ( $L1$ ) is statistically significantly longer than that of the background ( $L2$ ) when using U test or KS test with  $P$ -value threshold 0.01 (Figure 5). Note that only 233 out of 281 iESR genes have 5'UTR length data.

For qualitative enrichment analysis, first we need to define a set of genes with long 5'UTR length from the background (4418 genes which have 5'UTR length data). Let us define a set of genes with long 5'UTR length by including all the genes whose 5'UTR lengths are within the largest  $X\%$  ( $X = 10, 20, 30, 40$ ) of the background. Then we can use the hypergeometric test to check whether the iESR genes are enriched with genes with long 5'UTR length.

If we set the  $P$ -value threshold for calling enrichment as 0.01, then we can conclude that iESR genes are enriched with genes with long 5'UTR length only when  $X = 40$  but not 10, 20 or 30 (Table 4). It can be seen that the testing

**Table 2.** YQFC found that rESR genes are statistically larger than iESR genes in 23 QFs. To ensure the statistical significance and robustness of the findings, we only consider those with FDR-corrected  $P$ -values  $< 0.001$  in all three statistical tests (t-test, U test and KS test)

QF(rESR) > QF(iESR)	$P$ -value (t-test)	$P$ -value (U test)	$P$ -value (KS test)
Number of mRNA isoforms	2.393E-36	1.047E-33	1.421E-24
Number of GO terms	1.074E-04	2.751E-06	1.341E-06
mRNA level—mRNA expression level	1.925E-37	1.390E-53	3.282E-44
Translational efficiency—Dataset1 (Csárdi)	3.143E-17	1.955E-29	7.422E-27
Protein abundance (normal condition)—Dataset1 (Mean molecules per cell)	3.597E-11	2.328E-23	1.580E-18
Protein abundance (stress condition)—2 mM dithiothreitol, 2 h	8.394E-05	7.610E-07	9.150E-05
Protein abundance (stress-condition)—1mM H <sub>2</sub> O <sub>2</sub> , 1 h	3.487E-07	6.236E-13	5.831E-09
Protein abundance (stress-condition)—0.2M hydroxyurea, 160 min	1.193E-11	3.315E-16	1.748E-10
Protein abundance (stress-condition)—0.2M hydroxyurea, 2 h	3.708E-08	1.920E-10	1.715E-08
Protein abundance (stress-condition)—0.03% methyl methanesulfonate, 2 h	3.371E-06	2.133E-10	2.027E-09
Codon bias	9.216E-38	7.312E-21	2.745E-16
Codon adaptation index	1.313E-45	1.472E-28	6.607E-21
Frequency of optimal codons	3.524E-38	2.1344E-21	2.180E-16
Atomic composition—Hydrogen	3.4931E-16	1.449E-15	8.345E-12
Atomic composition—Nitrogen	1.134E-06	7.026E-07	1.582E-05
Amino acid composition—A	7.937E-07	2.9351E-07	5.746E-06
Amino acid composition—E	3.997E-06	7.183E-05	1.094E-05
Amino acid composition—K	2.466E-15	1.033E-11	8.844E-12
Amino acid composition—R	2.879E-16	6.514E-11	9.999E-10
Amino acid composition—V	3.859E-07	2.224E-06	1.899E-04
Number of interactors in the PI network—BioGRID	1.123E-14	2.408E-41	2.631E-34
Number of interactors in the GI network	3.556E-19	1.530E-18	1.226E-19
Number of interactors in the CX network	3.079E-07	1.900E-05	6.763E-06

QFs quantitative features

**Table 3.** YQFC found that iESR genes are statistically larger than rESR genes in 17 QFs. To ensure the statistical significance and robustness of the findings, we only consider those with FDR-corrected  $P$ -values  $< 0.001$  in all three statistical tests (t-test, U test and KS test)

QF(iESR) > QF(rESR)	$P$ -value (t test)	$P$ -value (U test)	$P$ -value (KS test)
5'UTR length	9.172E-07	1.932E-14	2.831E-12
Number of transcriptional regulators	1.544E-06	6.367E-06	7.281E-04
mRNA half-life	3.335E-32	6.538E-78	3.760E-71
Transcriptional plasticity	1.747E-07	4.046E-08	6.679E-07
Aromaticity score	8.413E-17	6.953E-21	6.270E-17
Extinction coefficient at 280 nm—All Cys residues appear as half cystines	1.701E-06	3.080E-08	4.875E-08
Extinction coefficient at 280 nm—No Cys residues appear as half cystines	1.864E-06	3.687E-08	5.844E-08
Atomic composition—Carbon	1.967E-25	5.450E-25	5.226E-18
Amino acid composition—F	1.485E-06	9.373E-12	1.123E-08
Amino acid composition—H	7.552E-06	3.296E-07	3.961E-06
Amino acid composition—N	1.161E-08	2.001E-08	4.775E-06
Amino acid composition—P	1.303E-10	5.362E-12	1.515E-08
Amino acid composition—W	7.552E-06	1.842E-06	1.128E-05
Amino acid composition—Y	5.226E-13	1.106E-14	5.057E-13
Number of interactors in the PIA network	7.839E-08	3.209E-06	8.477E-05
Number of interactors in the LEA network	5.942E-04	5.907E-05	3.889E-05
Number of interactors in the MPA network	4.532E-05	3.046E-06	1.371E-04

QFs quantitative features

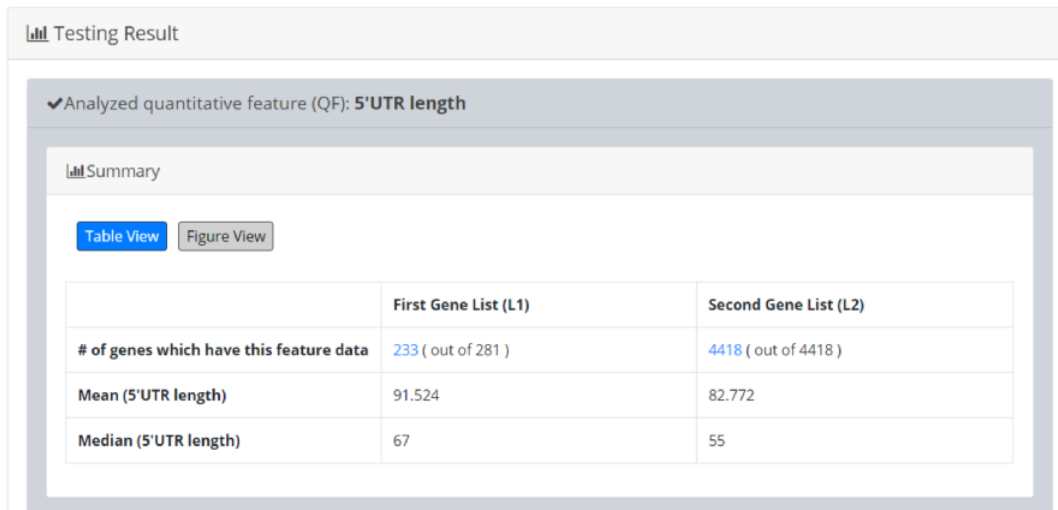
PIA physical interaction association

LEA literature evidence association

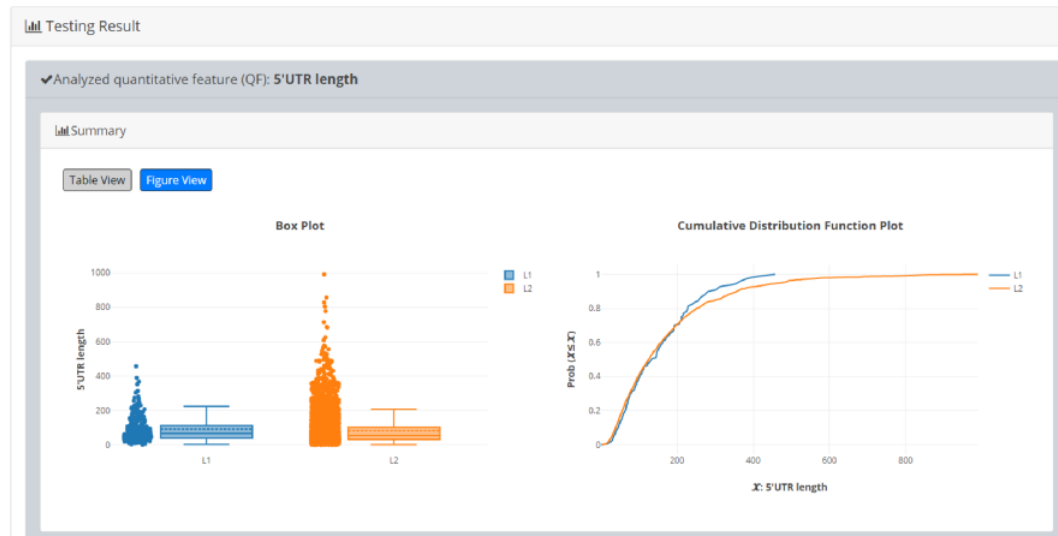
MPA mutant phenotype association



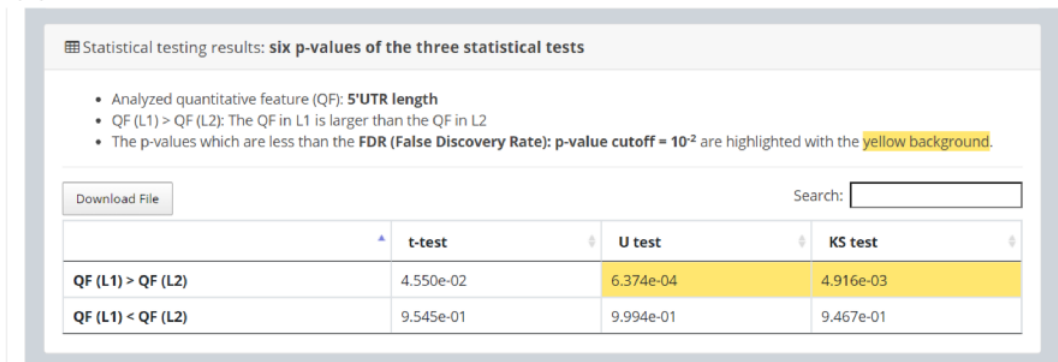
(a)



(b)



(c)



**Figure 5.** Quantitative comparison of 5'UTR length between 281 iESR genes and the background (4418 genes which have 5'UTR length data). (a) Table View (b) Figure View (c) Statistical testing results. It can be seen that the 5'UTR length of 281 iESR genes (*L1*) is statistically significantly longer than that of the background (*L2*) when using U test or KS test with *p*-value threshold 0.01. Note that only 233 out of 281 iESR genes have 5'UTR length data.

**Table 4.** For qualitative enrichment analysis, first we need to define a set of genes with long 5'UTR length from the background (4418 genes, which have 5'UTR length data). Let us define a set of genes with long 5'UTR length by including all the genes whose 5'UTR lengths are within the largest X% (X = 10, 20, 30, 40) of the background. Then we can use the hypergeometric test to check whether the iESR genes are enriched with genes with long 5'UTR length. If we set the *P*-value threshold for calling enrichment as 0.01, then we can conclude that iESR genes are enriched with genes with long 5'UTR length only when X = 40 but not 10, 20 or 30

X	Expected ratio in background	Observed ratio in iESR genes	<i>P</i> -value for enrichment calculated by hypergeometric test
10	10% (441/4418)	12.88% (30/233)	0.08381
20	20% (883/4418)	24.46% (57/233)	0.04971
30	30% (1325/4418)	36.05% (84/233)	0.02398
40	40% (1767/4418)	48.07% (112/233)	0.006232

results of the qualitative enrichment analysis are sensitive to how we define a set of genes with long 5'UTR length. Therefore, we suggest researchers to use quantitative comparison rather than qualitative enrichment analysis when quantitative feature data are available.

## Conclusion

In this study, we have developed a web tool called YQFC. We have comprehensively collected 85 quantitative features (including 12 gene features, 4 mRNA features, 52 protein features and 17 network features) in YQFC. For each quantitative feature, YQFC provides three statistical tests (t-test, U test and KS test) to test whether this quantitative feature is statistically different between the two input yeast gene lists. In the case study, we demonstrated the functionality and usefulness of YQFC by identifying 40 distinct quantitative features between rESR genes and iESR genes (Tables 2 and 3). The distinct quantitative features (e.g. 5'UTR length, mRNA half-life, transcriptional plasticity, translational efficiency, etc.) identified by YQFC may help researchers to study the underlying molecular mechanisms that differentiate rESR genes from iESR genes. We believe that YQFC is a useful tool to expedite the biological research that uses high-throughput omics technologies. In the future, we will keep updating YQFC once new quantitative features become available in the yeast literature or yeast databases. Now is YQFC Version 1.0. When YQFC has a minor update, we will change the version number to 1.1, 1.2, etc. In the future, if YQFC has a major update, we will change the version number to 2.0.

## Acknowledgements

We thank Dr Jagat Rathod for helpful comments.

## Funding

This work was supported by Ministry of Science and Technology of Taiwan [107-2221-E-006-225-MY3 and 108-2628-E-006-004-MY3]. Dr Y.-Y.T. is supported by NCI RO1 CA204962. Funding

for open access charge: Ministry of Science and Technology of Taiwan.

## Supplementary data

Supplementary data are available at *Database* Online.

*Conflict of interest.* None declared.

## References

- Horgan, R. and Kenny, L. (2011) 'Omic' technologies: genomics, transcriptomics, proteomics and metabolomics. *Obstet. Gynaecol.*, **13**, 189–195.
- Huang, D.W., Sherman, B.T. and Lempicki, R.A. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, **37**, 1–13.
- Jiao, X., Sherman, B.T., Stephens, R. *et al.* (2012) DAVID-WS: a stateful web service to facilitate gene/protein list analysis. *Bioinformatics*, **28**, 1805–1806.
- Balakrishnan, R., Park, J., Karra, K. *et al.* (2012) YeastMine— an integrated data warehouse for *Saccharomyces cerevisiae* data as a multipurpose tool-kit. *Database (Oxford)*, **2012**, bar062.
- Weng, M.P. and Liao, B.Y. (2017) modPhEA: model organism phenotype enrichment analysis on eukaryotic gene sets. *Bioinformatics*, **33**, 3505–3507.
- Chang, D.T., Li, W.S., Bai, Y.H. *et al.* (2013) YGA: identifying distinct biological features between yeast gene sets. *Gene*, **518**, 26–34.
- Wu, W.S., Tu, H.P., Chu, Y.H. *et al.* (2018) YHMI: a web tool to identify histone modifications and histone/chromatin regulators from a gene list in yeast. *Database (Oxford)*, **2018**, bay116.
- Rathod, J., Tu, H.P., Chang, Y.I. *et al.* (2018) YARG: a repository for arsenic-related genes in yeast. *PLoS One*, **13**, e0201204.
- Teixeira, M.C., Monteiro, P.T., Palma, M. *et al.* (2018) YEASTRACT, an upgraded database for the analysis of transcription regulatory networks in *Saccharomyces cerevisiae*. *Nucleic Acids Res.*, **46**, D348–D353.
- Wu, W.S., Hsieh, Y.C. and Lai, F.J. (2016) YCRD: yeast combinatorial regulation database. *PLoS One*, **11**, e0159213.

11. Holstege, F.C., Jennings, E.G., Wyrick, J.J. *et al.* (1998) Dissecting the regulatory circuitry of a eukaryotic genome. *Cell*, **95**, 717–728.
12. Nagalakshmi, U., Wang, Z., Waern, K. *et al.* (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, **320**, 1344–1349.
13. Lin, Z., Wu, W.S., Liang, H. *et al.* (2010) The spatial distribution of cis regulatory elements in yeast promoters and its implications for transcriptional regulation. *BMC Genomics*, **11**, 581.
14. Pelechano, V., Wei, W. and Steinmetz, L.M. (2013) Extensive transcriptional heterogeneity revealed by isoform profiling. *Nature*, **497**, 127–131.
15. Christiano, R., Nagaraj, N., Fröhlich, F. *et al.* (2014) Global proteome turnover analyses of the Yeasts *S. cerevisiae* and *S. pombe*. *Cell Rep.*, **9**, 1959–1965.
16. Csárdi, G., Franks, A., Choi, D.S. *et al.* (2015) Accounting for experimental noise reveals that mRNA levels, amplified by post-transcriptional processes, largely determine steady-state protein levels in yeast. *PLoS Genet.*, **11**, e1005206.
17. Ho, B., Baryshnikova, A. and Brown, G.W. (2018) Unification of protein abundance datasets yields a quantitative *Saccharomyces cerevisiae* proteome. *Cell Syst.*, **6**, 192–205.
18. Kim, H., Shin, J., Kim, E. *et al.* (2014) YeastNet v3: a public database of data-specific and integrated functional gene networks for *Saccharomyces cerevisiae*. *Nucleic Acids Res.*, **42**, D731–D736.
19. Wu, W.S., Wang, C.C., Jhou, M.J. *et al.* (2015) YAGM: a web tool for mining associated genes in yeast based on diverse biological associations. *BMC Syst. Biol.*, **9**, S1.
20. Lang, O.W., Nash, R.S., Hellerstedt, S.T. *et al.* (2018) An introduction to the Saccharomyces Genome Database (SGD). *Methods Mol. Biol.*, **1757**, 21–30.
21. Oughtred, R., Stark, C., Breitkreutz, B.J. *et al.* (2019) The BioGRID interaction database: 2019 update. *Nucleic Acids Res.*, **47**, D529–D541.
22. Hibbs, M.A., Hess, D.C., Myers, C.L. *et al.* (2007) Exploring the functional landscape of gene expression: directed search of large microarray compendia. *Bioinformatics*, **23**, 2692–2699.
23. Hogg, R.V., McKean, J.W. and Craig, A.T. (2018) *Introduction to Mathematical Statistics*. Pearson, London.
24. Corder, G.W. and Foreman, D.I. (2014) *Nonparametric Statistics: A Step-by-Step Approach*. Wiley, Hoboken, NJ.
25. Gasch, A.P., Spellman, P.T., Kao, C.M. *et al.* (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell.*, **11**, 4241–4257.
26. Moore, J., Jacobs, H.T. and Kaiser, K. (1995) The relationship between mRNA half-life and gene function in the yeast *Saccharomyces cerevisiae*. *Gene*, **166**, 145–149.
27. Basehoar, A.D., Zanton, S.J. and Pugh, B.F. (2004) Identification and distinct regulation of yeast TATA box-containing genes. *Cell*, **116**, 699–709.
28. Gray, N.K. (1998) Translational control by repressor proteins binding to the 5'UTR of mRNAs. *Methods Mol. Biol.*, **77**, 379–397.
29. Lawless, C., Pearson, R.D., Selley, J.N. *et al.* (2009) Upstream sequence elements direct post-transcriptional regulation of gene expression under stress conditions in yeast. *BMC Genomics*, **10**, 7.