R scripts for checking the best-fit distribution and modality of the budding yeast morphological data, obtained by CalMorph.

These scripts are used to perform the UNImodal MOrphology (UNIMO) statistical pipeline described in Ghanegolmohammadi et al. (2021; Bioinformatics). For detailed descriptions read “Supporting Text”.

**Descriptions of the R scripts:**

* **01 Noise parameters (114 WT):** Calculating the best smooth span value (f = {0.10, 0.11, 0.12, 0.13, ..., 0.99}) for LOESS regression using lo() function of 'gamlss' package (Stasinopoulos & Rigby, 2007; R. J. Stat. Softw., 23(7): 1-46) and Akaike information criterion (AIC).
* **02 Data preparation (114 WT):** Transformation of the 501 CalMorph parameters (Ohya et al, 2005; PNAS 102(52): 19015-19020) including:
  + Uncoupling the dependency between the coefficient of variation (CV) and the mean values (Levy and Siegal, 2008; PLoS Biol. 6(11): e264; Yvert et al. 2013; BMC Syst. Biol., 7(1): 54).
  + Handling inversion of seven ratio parameters.
* **03 Reference distribution check (114 WT):** Check the equality of each 501 CalMorph parameters (Ohya et al, 2005; PNAS 102(52): 19015-19020) with the presumed reference distribution using:
  + Kolmogorov-Smirnov test for
    - Non-negative continuous values [0, +Inf) includes 183 CalMorph parameters 🡪 Gamma (GA) distribution.
    - Bounded continuous values; i.e., ratio [0,1] includes 37 CalMorph parameters 🡪 Beta (BE) distribution.
    - Finite discrete values; i.e., proportion [0,1) includes 61 CalMorph parameters 🡪 Binomial (BI) distribution.
  + Shapiro-Wilk test of normality for
    - Real continuous values; i.e., noise (-Inf, +Inf) includes 220 CalMorph parameters 🡪 Gaussian (NO) distribution.
* **04 Probability distribution function (114 WT):** Rather than reference distribution for each data type, we tested various probability distributions to survey changes of skewness, kurtosis, and specific cases where the data are a mixture of discrete and continuous values:
  + Non-negative continuous values [0, +Inf) includes 183 CalMorph parameters:
    - Reference distribution: Gamma (GA)
    - Alternative distributions:
      * Exponential distribution (EXP)
      * Generalized Pareto distribution (GP)
      * Inverse gamma distribution (IGAMMA)
      * Inverse Gaussian distribution (IG)
      * Log normal distribution (LOGNO)
      * Log-normal 2 distribution (LOGNO2)
      * Pareto type 1 distribution for y > 0 (PARETO1)
      * Pareto type 2 distribution for y > 0 (PARETO2)
      * Pareto type 2 distribution for y > 0 (PARETO2o)
      * Weibull distribution (WEI)
      * Weibull distribution for PH parameterization (WEI2)
      * Weibull distribution where mu is the mean of the distribution (WEI3)
      * Zero-adjusted Gamma distribution (ZAGA)
      * Zero-adjusted inverse Gaussian distribution (ZAIG)
  + Bounded continuous values; i.e., ratio values [0,1] includes 37 CalMorph parameters:
    - Reference distribution: Beta (BE) distribution
    - Alternative distributions:
      * Beta zero-inflated distribution (BEINF0)
      * Beta one-inflated distribution (BEINF1)
      * Beta zero- and one-inflated distribution (BEINF)
      * One-inflated beta distribution (BEOI)
      * Zero-inflated beta distribution (BEZI)
      * Logit-normal distribution (LOGITNO)
  + Real continuous values; i.e., noise values (-Inf, +Inf) includes 220 CalMorph parameters:
    - Reference distribution: Gaussian (NO) distribution
    - Alternative distributions:
      * Gumbel distribution (GU)
      * Logistic distribution (LO)
      * Normal distribution 2 with variance as sigma parameter (NO2)
      * Reverse Gumbel distribution (RG)
  + Finite discrete values; i.e., proportion values [0,1) includes 61 CalMorph parameters:
    - Reference distribution: Binomial (BI) distribution
    - Alternative distributions:
      * Beta-binomial distribution (BB)
      * Zero-adjusted beta-binomial distribution (ZABB)
      * Zero-adjusted binomial distribution (ZABI)
      * Zero-inflated beta-binomial distribution (ZIBB)
      * Zero-inflated binomial distribution (ZIBI)
* **05 Confounding factors (114 WT):** To minimize the effects of experimental error among replicates, a group of five confounding factors was considered, including a combination of different fluorescence filters for microscopy and the period of image acquisition (Ohnuki and Ohya, 2018; PLoS Biol., 16(5): e2005130). A generalized linear model was introduced by constructing a linear model (ANOVA) of the confounding factors:
  + y: The fitted value.
  + f: An appropriate link function.
  + Beta: Fixed effect of each confounding factor (0,1).
  + MS1: Microscope #1 before replacement of the fluorescence filter.
  + MS2: Microscope #2 before replacement of the fluorescence filter.
  + MS2a: Microscope #2 after replacement of the fluorescence filter over time.
  + MS2b: Microscope #2 after replacement of the fluorescence filter.
  + MS3: Microscope #3.

For every 501 CalMorph parameter (Ohya et al, 2005; PNAS 102(52): 19015-19020) the best-fitted model between the confounding factor model (CFM) and the null model was selected using AIC given the predefined probability distribution functions (PDFs).

* **06 Modality (114 WT):** Mixture-model-based clustering to check the modality of the 501 CalMorph parameters (Ohya et al, 2005; PNAS 102(52): 19015-19020) according to their pre-defined distributions. Bayesian information criterion (BIC) is employed to compare the primary probability models that differed in the number of components:
  + n: Number of observations (i.e., sample size).
  + k: Number of hyper-parameters in the model multiplied by the number of the clusters (i.e., 1 <= c <= 10).
  + L: Estimated maximum value of the likelihood function for the model.

In the case of edge peak distributions, a model-free outlier-detection method (one-percentile deviation rule) was used to remove expected sets of outliers from the multimodal parameters.

* **07 Noise parameters (4718 mutants):** Re-analyzed morphological variations of the 4718 yeast nonessential gene mutants by comparing them to a data set of haploid wild-type yeast strains (his3; 109 replicates). Calculating the best smooth span value (f = {0.10, 0.11, 0.12, 0.13, ..., 0.99}) for LOESS regression using lo() function of 'gamlss' package (Stasinopoulos & Rigby, 2007; R. J. Stat. Softw., 23(7): 1-46) and Akaike information criterion (AIC).
* **08 Data preparation (4718 mutants):** Re-analyzed morphological variations of the 4718 yeast nonessential gene mutants by comparing them to a data set of haploid wild-type yeast strains (his3; 109 replicates). Transformation of the 501 CalMorph parameters (Ohya et al, 2005; PNAS 102(52): 19015-19020) including:
  + Uncoupling the dependency between the coefficient of variation (CV) and the mean values (Levy and Siegal, 2008; PLoS Biol. 6(11): e264; Yvert et al. 2013; BMC Syst. Biol., 7(1): 54).
  + Handling inversion of seven ratio parameters.
* **09 Data analysis (4718 mutants):** Re-analyzed morphological variations of the 4718 yeast nonessential gene mutants by comparing them to a data set of haploid wild-type yeast strains (his3; 109 replicates). Models of the probability distributions for the 490 unimodal CalMorph parameters were determined (see '04 Probability distribution function (114 WT).r' and '06 Modality (114 WT).r' files) to accommodate the pre-defined statistical models. Then, P values (two-sided test) as the deviation of each mutant from the wild-type (WT; null model) using functions in the 'gamlss' package (Stasinopoulos & Rigby, 2007; R. J. Stat. Softw., 23(7): 1-46) are calculated. False discovery rate (FDR), the rate of type I errors in the rejected null hypothesis due to multiple comparisons, was estimated by the 'qvalue' package (Storey, 2002; J. R. Stat. Soc. B, 64(3): 479-498).

**Descriptions of the additional data folder:**

* **501-Info.csv:** Final information of best fit probability model, related link functions, reference distribution, modality, and description of each CalMorph parameter (Ohya et al, 2005; PNAS 102(52): 19015-19020).
* Confounding\_factors.rdata: Information of experimental errors among replicates, a group of five confounding factors was considered, including a combination of different fluorescence filters for microscopy and the period of image acquisition (Ohnuki and Ohya, 2018; PLoS Biol., 16(5): e2005130).

**Descriptions of the functions:**

* **loess.fit.r:** LOESS (locally estimated scatterplot smoothing) regression function to uncouple non-linear dependency between the coefficient of variation (CV) of their related mean (220 CalMorph parameters). Finally, instead of CV parameters, noise parameters as the residuals (i.e., observed value - predicted value) are calculated. For a detailed description see Levy and Siegal (2008; PLoS Biol. 6(11):e264) and Yvert et al. (2013; BMC Syst. Biol. 7(1):54).
* **pBB2.r:** pBB() function of gamlss package (Stasinopoulos & Rigby, 2007; R. J. Stat. Softw., 23(7): 1-46) defines probability mass function of the beta-binomial distribution at any given sample between zero and one (0 < y < 1). In case of very large quantiles (numerator equals denominator) very small p values (less than 1E-13) for the upper-tail test fail to be estimated. This function (pBB2) solves this problem.