Computational and Structural Biotechnology Journal 23 (2024) 2949–2962

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Review article

Systematic data analysis pipeline for quantitative morphological cell phenotyping

Farzan Ghanegolmohammadi [a,b,*], Mohammad Eslami [c], Yoshikazu Ohya [b,**]

[a] Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[b] Department of Integrated Biosciences, Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa, Chiba, Japan
[c] Harvard Ophthalmology AI Lab, Schepen's Eye Research Institute of Massachusetts Eye and Ear Infirmary, Harvard Medical School, Boston, USA

ARTICLE INFO

Keywords:
Morphological profile
Image-based cell profiling
Cell morphology
High-content imaging

ABSTRACT

Quantitative morphological phenotyping (QMP) is an image-based method used to capture morphological features at both the cellular and population level. Its interdisciplinary nature, spanning from data collection to result analysis and interpretation, can lead to uncertainties, particularly among those new to this actively growing field. High analytical specificity for a typical QMP is achieved through sophisticated approaches that can leverage subtle cellular morphological changes. Here, we outline a systematic workflow to refine the QMP methodology. For a practical review, we describe the main steps of a typical QMP; in each step, we discuss the available methods, their applications, advantages, and disadvantages, along with the R functions and packages for easy implementation. This review does not cover theoretical backgrounds, but provides several references for interested researchers. It aims to broaden the horizons for future phenome studies and demonstrate how to exploit years of endeavors to achieve more with less.

1. Introduction

Phenotypic diversity involves variations in molecular aspects (*e.g.*, proteome, transcriptome, and metabolome), cellular characteristics (*e.g.*, cell morphology), fitness metrics (*e.g.*, growth rate, colony size, and yield of biomass), and visible features (*e.g.*, colony morphology and invasive growth) of a cell population. Morphology defines the basic phenotypic characteristic of both unicellular and multicellular organisms. The effects of genetic or environmental perturbations result in diverse morphological phenotypes through natural selection [110,54,56]. Investigating this diversity enables direct observation of the cellular processes to answer various biological questions. In multicellular organisms, cell morphology reflects cell behaviors and intracellular communications [96,155]. In unicellular organisms, like *Saccharomyces cerevisiae*, morphology dynamically changes in response to life-cycle events (*e.g.*, cell cycle progression), stressors, genotype, and genetic networks and has been used to gain a global understanding of cell systems [154].

Imaging platforms are employed for morphological studies, including 1) phenotypic screening (a targeted method for quantifying a single process or cellular function), such as studying variations in cell size due to genetic or environmental perturbations, and 2) phenotypic profiling (an unbiased approach that quantifies as many features as possible). The latter is also known as quantitative morphological phenotyping (QMP). Due to its inclusive nature, QMP has played a significant role in opening new frontiers in biology (Supplementary Table 1). The amount of biological information captured by a typical QMP experiment is comparable with that of other high-throughput methods [78]. However, mapping variations in morphology related to specific phenotypes is not a straightforward task due to their complexity [71], and the intricate interactions among different biological domains [79,165]. Consequently, there is a demand for efficient data collection and analytical strategies. It is essential to define a proper methodological pipeline to complement the available conventional choices. This includes preprocessing, statistical modeling, and follow-up analysis (*e.g.*, clustering and classification) to obtain accurate results. A clear pipeline can also eliminate unnecessary duplication of efforts across different groups and organizations.

To support biologists in constructing a typical QMP pipeline, we illustrate a comprehensive QMP workflow to transform quantified morphological data into biologically meaningful insights. The workflow includes: 1) image analysis, 2) data modeling, 3) knowledge extraction,

* Corresponding author at: Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.
** Corresponding author.
E-mail addresses: farzang@mit.edu (F. Ghanegolmohammadi), ohya@edu.k.u-tokyo.ac.jp (Y. Ohya).

https://doi.org/10.1016/j.csbj.2024.07.012
Received 6 May 2024; Received in revised form 9 July 2024; Accepted 10 July 2024
Available online 14 July 2024

4) sharing, and 5) transformative knowledge exploration (Fig. 1). For each step, we documented the importance and applicability, various computational methods, often overlooked challenges, and relevant R functions and packages. Each step included a set of examples to further demonstrate QMP applications for addressing various biological questions, and readers are encouraged to review the cited publication(s) and references, vignettes in Supplementary Tables, and the R book [25] for further information and practical examples. It is important to note that before QMP initiation, well-designed experiments followed by proper sample preparation and image acquisition are required [73]. While this review does not cover them, these upstream steps are crucial for laying the foundation of a successful QMP.

2. Step 1: Image analysis

To generate reproducible results, high-quality assays and appropriate imaging techniques are vital as the initial steps of a QMP experiment. These conditions minimize artifacts and save time for further analysis, but are not always met. Thus, a typical QMP begins with checking the quality of the captured images (Step 1–1), followed by extracting morphological information from high-quality images using computer vision techniques (Step 1–2). This data lays the foundation for successful data analysis subsequently.

2.1. Step 1-1: Image quality assessment

Errors during sample preparation (*e.g.*, human errors, uncalibrated instruments, *etc.*) and image acquisition (*e.g.*, improper focusing) can introduce artificial variations into the data. Therefore, image analysis must include an objective image assessment step to filter out unwanted data and low-quality images [178].

Manual verification of image quality in a large number of images from high-throughput experiments is not feasible. Thus, a systematic approach is required to objectively flag or remove artifacts and noise, such as non-uniform light source or shady edges. Several methods have been proposed for image preprocessing (Supplementary Table 2). However, as a general rule, it is highly recommended to apply various methods to increase the likelihood of artifact identification. In our experience, inevitable situations, such as changing fluorescence filters, which might affect data modality (*Step 2–3: Modality*), can be added as confounders to the statistical model to avoid possible misleading inferences.

2.2. Step 1-2: Quantifying cell morphology

Quantifying cell morphology involves two sub-steps. First, segmentation involves partitioning cell boundaries and subcellular structures (Fig. 2a and Supplementary Table 3). This is usually performed by manually optimized algorithms or trained classifiers. The former typically requires human intervention and has limited application, particularly in high-throughput experiments [100]. A trained classifier is mainly preferable in large-scale experiments; however, classifiers are as good as their training set and their applications are limited in the absence of a standard training set (Supporting Text). A global segmentation algorithm [18] or foundation models could be used to address this problem, but are beyond the scope of this review.

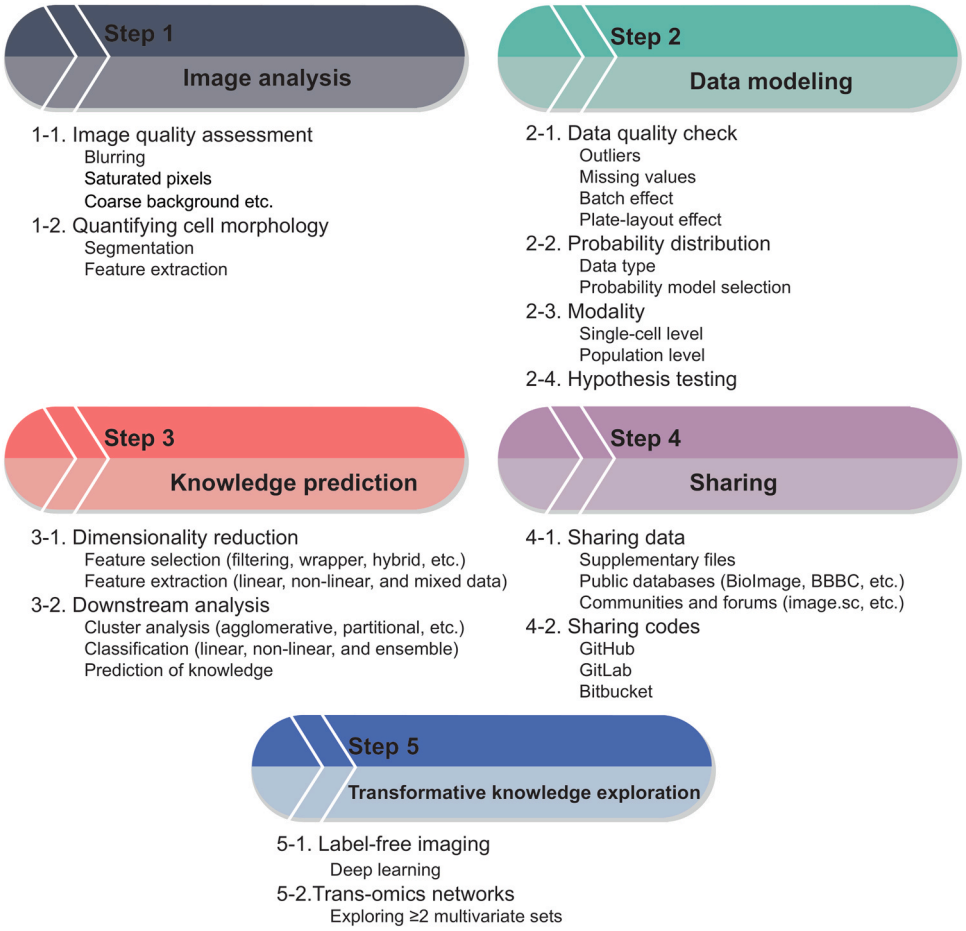

Fig. 1. Quantitative morphological phenotyping (QMP) workflow. This figure illustrates the main steps of a typical QMP process. As depicted, each step represents a broad spectrum of analytical methods and diverse approaches designed to accommodate the varying conditions in each study. Additional details are provided in Supplementary Tables 2–15.

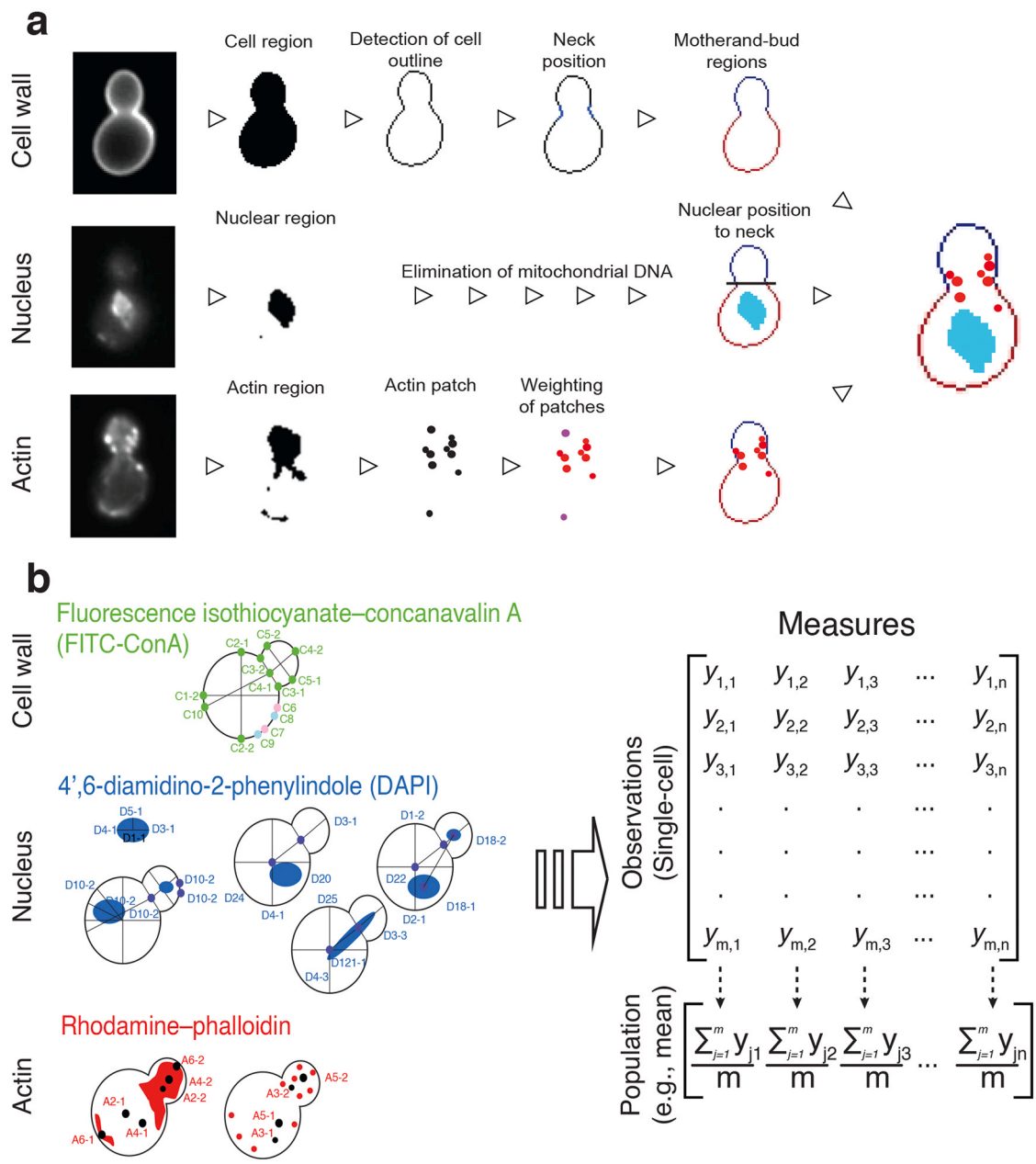

Fig. 2. Segmentation and feature extraction steps of a budding yeast cell. **a.** The segmentation steps in the CalMorph image analysis tool [114] include three stages of processing of triple-fluorescent stained cells: the cell region (top), the nucleus (middle), and the actin cytoskeleton (bottom). Different versions of CalMorph (http://www.yeast.ib.k.u-tokyo.ac.jp/CalMorph/index.html) can extract morphological features at either single-cell level or population level (*i.e.*, arithmetic mean of single cell data). **b.** CalMorph extracts 501 morphological features at the cell population level. Further information is provided in the "Image Processing in CalMorph" section in [35].
The figure has been modified from [35].

Second, feature extraction involves using an image processing tool/algorithm to extract as many features as possible (Supplementary Table 4). The extracted morphological measures cover a wide range of functionality and usability [29,168] particularly if the tool is designed for a specific goal [*e.g.*, CalMorph for *S. cerevisiae* [114], Fig. 2b] or as a multipurpose tool [*e.g.*, CellProfiler for human, fruit fly, worm, or yeast [149]]. When choosing an image analysis tool, we recommend considering the aims of the study (*i.e.*, the specific morphological measures one wants to acquire, such as single-cell data), accessibility, and ease of use.

Tools of any kind typically extract the morphological features of individual cells, but the output can be at either the single-cell or population-level (also referred to as image-level or well-level). Single-cell resolution is commonly employed to capture cell variations and rare phenotypes. Population-level measures, on the other hand, are ensemble averages of single-cell measurements that summarize the typical population features. The aggregation procedure varies depending on the aim and properties of the data. Previously applied strategies include: 1) *mean* profiles for normally distributed data [66,114]; however, these profiles are susceptible to outliers and 2) for non-normally distributed data, alternatives such as the *median* (or median absolute deviation), the *Kolmogorov–Smirnov* statistic [122], or the *Anderson–Darling* statistic, could be employed [17].

Another issue with population profiles is the nonlinear dependency between the morphological variations (coefficient of variation; CV) and mean measures in a heterogeneous population, introducing bias into the analysis. Locally estimated scatterplot smoothing (LOESS) regression

has been proposed to discriminate changes between variance and mean phenotypes [81]. As Levy and Siegal [81] suggested, after fitting a LOESS curve, the residual distance of each point from the regression line (*i.e.*, observed value − predicted value) represents a measure of variance controlled for the mean [6,81]. We suggest incorporating these residual values beside other morphological parameters, such as mean and ratio, to capture a global view of cell morphology. However, selecting the optimal smooth span (*f*) can be challenging. We recommend exploring a range of spans (*f*-values: 0.10–0.99) and determining the best-fit model using a model selection method, such as the Akaike Information Criterion (AIC, Supplementary Table 8).

3. Step 2: data modeling

After data collection, a methodology that balances specificity and sensitivity must be employed to detect subtle morphological changes. Additionally, before proceeding to hypothesis testing [51], data wrangling and verification of statistical assumptions are necessary (Supplementary Table 5). We will first detail the specific techniques and methods used for data wrangling, including approaches to handle outliers, missing data, batch effect, and plate-layout effect. Then, we will explore methods for verifying statistical assumptions. These preparatory steps are crucial for ensuring the integrity of our analyses and will be thoroughly discussed in the following sections to elucidate their impact on achieving robust and defensible research outcomes. Readers interested in practical examples of statistical learning and predictive models are encouraged to consult "An introduction to statistical learning" [34,63] and "Applied Predictive Modeling" [74]. These books provided numerous practical examples in R and Python.

3.1. Step 2-1: data quality check

The obtained morphological measures are based on empirical data. The accuracy of measures and consequently the data quality, significantly affect the final results in both enhancing and corrupting the final conclusion [118,128,3]. Quality control is a critical step that involves addressing misrecognized and unrecognized morphological features, as well as correcting for technical and experimental variations or confounds (Supplementary Table 6).

Outliers are extraordinary measures acquired from either unusual phenotypes or errors (Fig. 3a). These extreme values should be managed with careful techniques and protocols to avoid discarding

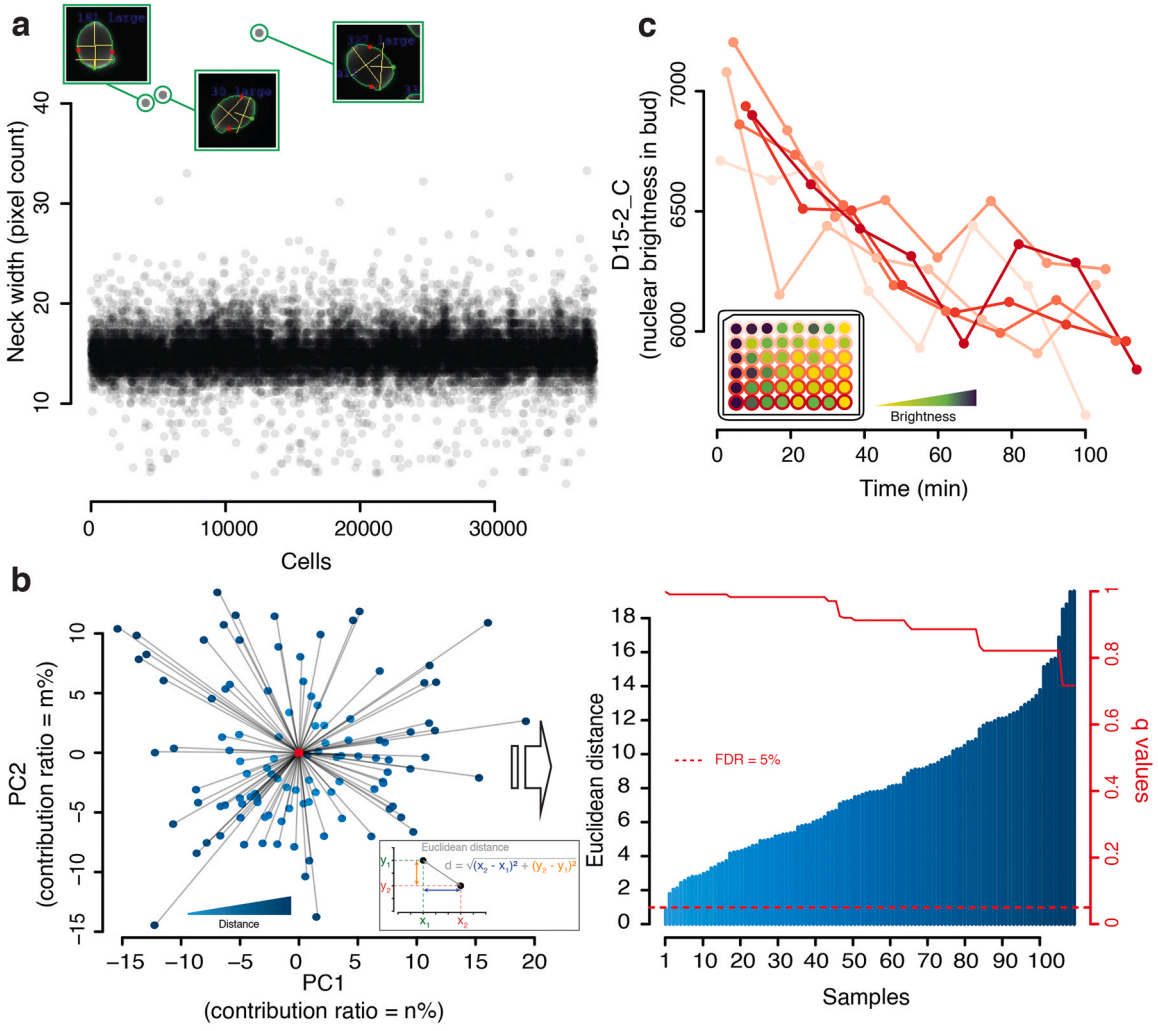

Fig. 3. Assessing the quality of extracted morphological features. **a.** Example of how outliers can be generated due to misrecognition of neck width in budding yeast cells by CalMorph. **b.** Example of using Euclidian distances to detect outliers, as previously described by [152]. Briefly, following Z-transformation, the data were subjected to PCA [scatter plot: circles represent 109 replicates of haploid wild-type yeast strain, and red square shows means of PC1 and PC2]. Euclidian distances (see inset formula) were calculated using the first two principal component (PC) scores (bar plot on the right). *P*-values were estimated by fitting a Gamma distribution to the Euclidean distances. The false discovery rate was estimated by converting *P*-values to q-values. **c.** Schematic illustration depicting the edge effect in a 48-well plate, where unfavorable edge conditions resulted in lower cell counts at the plate's edges.

subpopulations exhibiting interesting phenotypes. Otherwise, outliers might lead to misinterpretation of the results or undermine the analytical power. To detect outliers, we first recommend using simple strategies including visual presentations (*e.g.*, box plots) and univariate methods (*e.g.*, 3- or 5-standard deviation rules, Winsorization, and residual plots). Although these methods are not sufficiently robust, they can provide a general overview of the data. Alternatively, we assessed data variability using principal component analysis (PCA) to project the morphological profile onto lower dimensions [152]. We then use PC scores to calculate Euclidian distances of each entry (*i.e.*, a replicate in a population dataset or a cell in a single-cell dataset) from the center of the PC spaces. Significant deviations from the center of these spaces are considered outliers (Fig. 3b).

Other approaches include building multivariate statistical models, such as Hotelling's T-squared test, distance-based measures [132,134, 84], or training a classifier [24,55]. These methods have been proven efficient but require greater effort and experience. Perez and Tah [121] employed a multi-step approach to detect outliers, initially using a dimensionality reduction (DR) method (t-distributed stochastic neighbor embedding; t-SNE) to reduce high-dimensional features into a lower probability density distribution. Subsequently, they applied the interquartile range method to identify outliers from the density distribution of these features [121].

Missing values arise when the image-processing tool fails to extract morphological features (Step 1–1), resulting in gaps in the data. These incomplete values are generally indicated by specific symbols [*e.g.*, NaN; Not a Number] or a numerical value (*e.g.*, −1). Although missing data are relatively common in the data collection step, they pose serious challenges during data analysis. The approach to handling missing values depends on available logistics, experimental conditions, and the nature of extracted data. Ideally, the best course of action is to repeat the experiment, although this is often not feasible. For a small proportion of observations with missing values, general options include removing affected cells or using imputation techniques [57,77]. It is important to note that artificially replacing missing values can reduce data dispersion. Therefore, as Caicedo et al. [17] previously recommended, features with a large proportion of observations having missing values should be omitted.

Batch effects refer to unexpected technical variations (*e.g.*, different laboratory conditions or equipment calibrations) that can lead to false conclusions and misinterpretation of the results. Thus, correcting these misleading signals is an important step. At the experimental level, potential sources of variations should be identified and removed. However, achieving this desired uniformity is often challenging.

There are several approaches to mitigate the adverse effects of batch effects, such as standardization and quantile normalization at the plate level rather than to the entire screen [11], or canonical correlation analysis to transform data and maximize the similarity between technical replicates across experiments [161]. In our labs, we address batch effects by including sources of variation as confounding factors in the statistical model [111]. Subsequently, a model selection analysis (such as likelihood-ratio test) between the null and confounding factor models determines the goodness of fit. This process is straightforward when batch effects are known (*e.g.*, using different microscopes for imaging). However, detecting unknown variability is also essential. Recommended methods to identify possible batch effects include estimating the distance from the means of all replicates (Fig. 3b), analyzing class distinctions [5], and performing correlation analysis among profiles [17, 108]. The latter involves creating heatmaps that illustrate the correlations between all well pairs in an experiment, sorted by repeated experiments. Batch effects can then be detected as patterns of high correlation that indicate technical artifacts.

Plate-layout effect: Edge effects result from slight environmental differences between peripheral and inner wells of a multi-well plate in high-throughput assays (Fig. 3c), which affect experimental consistency and demand correction during both sample preparation and data analysis steps. For the former, a common correction strategy is to distribute samples randomly across the plates given the experimental conditions [89]. At the data analysis level, we primarily suggest a visual check of the measured variable as a heatmap in the same spatial format as the plate (Fig. 3c, inset). Objective approaches include two-way median polish for correcting positional effects [14], 2D polynomial regression (using the LOESS function) and employing corresponding residuals to correct spatial biases [129], and building a distance matrix in a reduced multi-dimensional space [42].

3.2. Step 2-2: probability distribution

A common practice in biology involves comparing a phenotype of interest with a baseline condition; thus, accurately estimating the true values from noisy biological measurements is central to quantitative biological studies. Here, we review core concepts and the most recent approaches for maximizing insights from data.

Data type. Each data type has specific properties that can be best described by certain probability distributions (Supplementary Table 7). Defining the data type is a crucial preliminary step in planning further analytical approaches such as normalization, statistical inference, *etc*. We previously demonstrated [36,113] that morphometric measures generally fall into five categories (Fig. 4): 1) Basic morphometric features of a cell population, such as shapes, intensities, and contexts, are continuous semi-infinite measures with non-negative values (*i.e.*, $0 \leq y$); 2) Ratios of two related morphological features (*e.g.*, cell axis ratio) are continuous bounded measures ($0 \leq y \leq 1$); 3) Residual or noise variables are continuous positive and negative values ($-\infty < y < +\infty$). As mentioned earlier, these values are derived from decoupling variance and mean phenotypes using methods such as LOESS regression; 4) Proportions of specific features in the population, such as proportion of budded cells to all cells, are discrete finite measures ($0 \leq y \leq 1$); and 5) Single-cell features (*e.g.*, number of actin patches) are discrete infinite measures with positive values ($0 \leq y$). Additionally, Rohban and coworkers [133] illustrated how fusions of different data types can add new dimensions to the morphological profile. In their study, incorporating dispersion and covariance estimates to the population averages improved the predictive performance in linking mechanism of action of a compound to gene pathways.

Understanding specific data types is not just a technical detail; it is fundamental to achieving broader research objectives [8,32]. It also facilitates the development of more robust models that effectively capture the complex interactions within biological systems [115]. Rigby and colleague' book [131] offers both a theoretical background and practical examples in R for various data types, detailing appropriate probability models for each. These can be implemented using the gamlss package [148].

Probability model selection: Natural variables do not always follow a normal distribution [113,174,8]. However, the apparent simplicity of applying normal distribution models often makes them the method of choice. It is essential to first understand the main characteristics of data by conducting simple exploratory data analysis, such as plotting (*e.g.*, histograms, cumulative distribution curves, and quantile–quantile plots) or performing normality tests (*e.g.*, Shapiro and Kolmogorov-Smirnov tests).

Non-normal data are often transformed to achieve approximately normal distributions through various methods, including square-root transformation, logarithmic transformation [31], Box-Cox transformation [6], and centering and scaling by the mean and standard deviation of negative controls [94]. It is necessary to employ and compare several normalization methods because their performances can vary significantly. For example, square-root transformations generally perform better with discrete infinite measures, while arcsine transformations are more suitable for discrete finite measures [9104].

Although normalization is an accepted approach for achieving fairly accurate repeatability from non-normal data that have been transformed

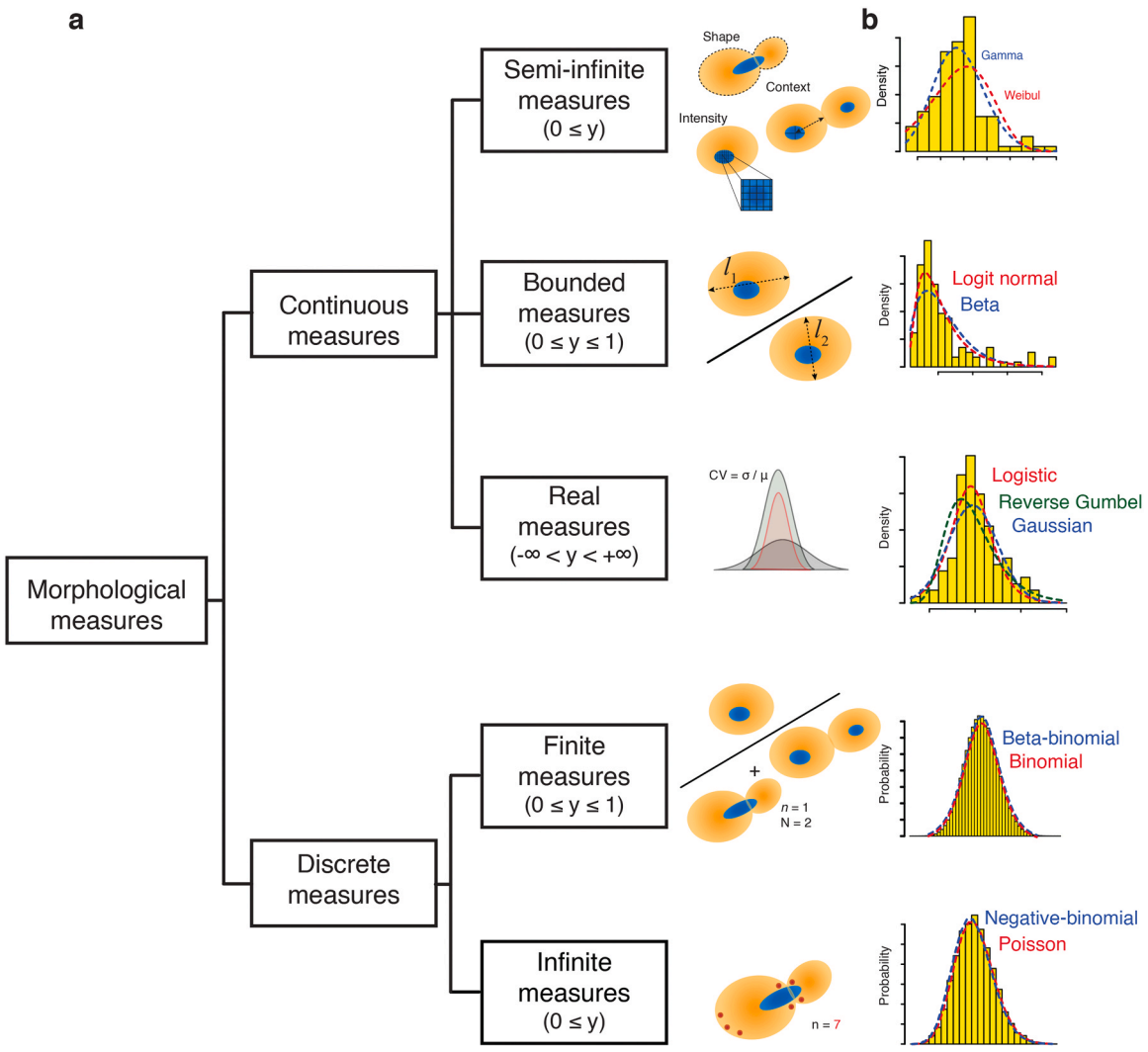

Fig. 4. Morphological data at a glimpse. **a.** The five types of morphological measurements, each with a clear definition and characteristics. Schematic examples of features defined in yeast cell morphology, along with their associated biological measurements for each data type, are depicted. **b.** Example histograms for each data type illustrating the distributions. To increase statistical specificity, models of the probability distributions for each morphological parameter must be determined to accommodate the statistical model used in a generalized linear model analysis (Supplementary Table 7). The best fit can be determined using a model selection metric (Supplementary Table 8).
The figure has been modified from [36].

to approximate normal distributions, it may limit the ability to detect subtle changes. Indeed, true values are best described by appropriate probability distributions defined according to the data type and its distribution (Supplementary Table 7).

We recently examined various probability distributions for morphological measures of budding yeast generated by CalMorph [36]. After testing 33 probability distributions, we determined the best fit using AIC, considering model complexity, available computation power, and runtime. We ultimately selected nine distributions that best matched the experimental error of the morphological features. We then compared the results with our previous study [114]. The impact of selecting the appropriate probability models on detecting subtle morphologies was significant, with approximately 1.5 times greater detection power achieved (Fig. 5a). Further pathway enrichment analysis revealed that the additional data provided useful biological information that was previously masked.

3.3. Step 2-3: Modality

In biology, the definition of modality is context-dependent. It sometimes refers to mixed populations, while in other instances, it can indicate various datasets representing different aspects of cellular biology (*e.g.*, transcriptome, proteome, *etc.*). In this review, we focused on the former, where profiling of subpopulations helps identify the full range of phenotypes within a population (*e.g.*, majority and over- and/or under-represented). Generally, biological modality enhances our understanding of dynamic transitions in cellular phenomena [47].

In statistics, modality refers to the number of peaks or modes in a probability distribution. It reflects the complexity of distribution, where the possible asymmetry of multimodal distributions can violate assumptions regarding the mean and dispersion [1160]. Therefore, to estimate true values and achieve more accurate predictions and inferences, simpler statistical approaches are more effective with unimodal distributions.

Single-cell level: Cells, even in monoclonal populations, respond to perturbations in a variety of ways; thus, cell populations are usually multimodal or heterogeneous [1145,146,84]. Recognizing this fact can reveal important biological insights [87,119], such as observing how different groups of cells respond to a drug treatment. This can lead to more personalized medical approaches [53] or aid in identifying

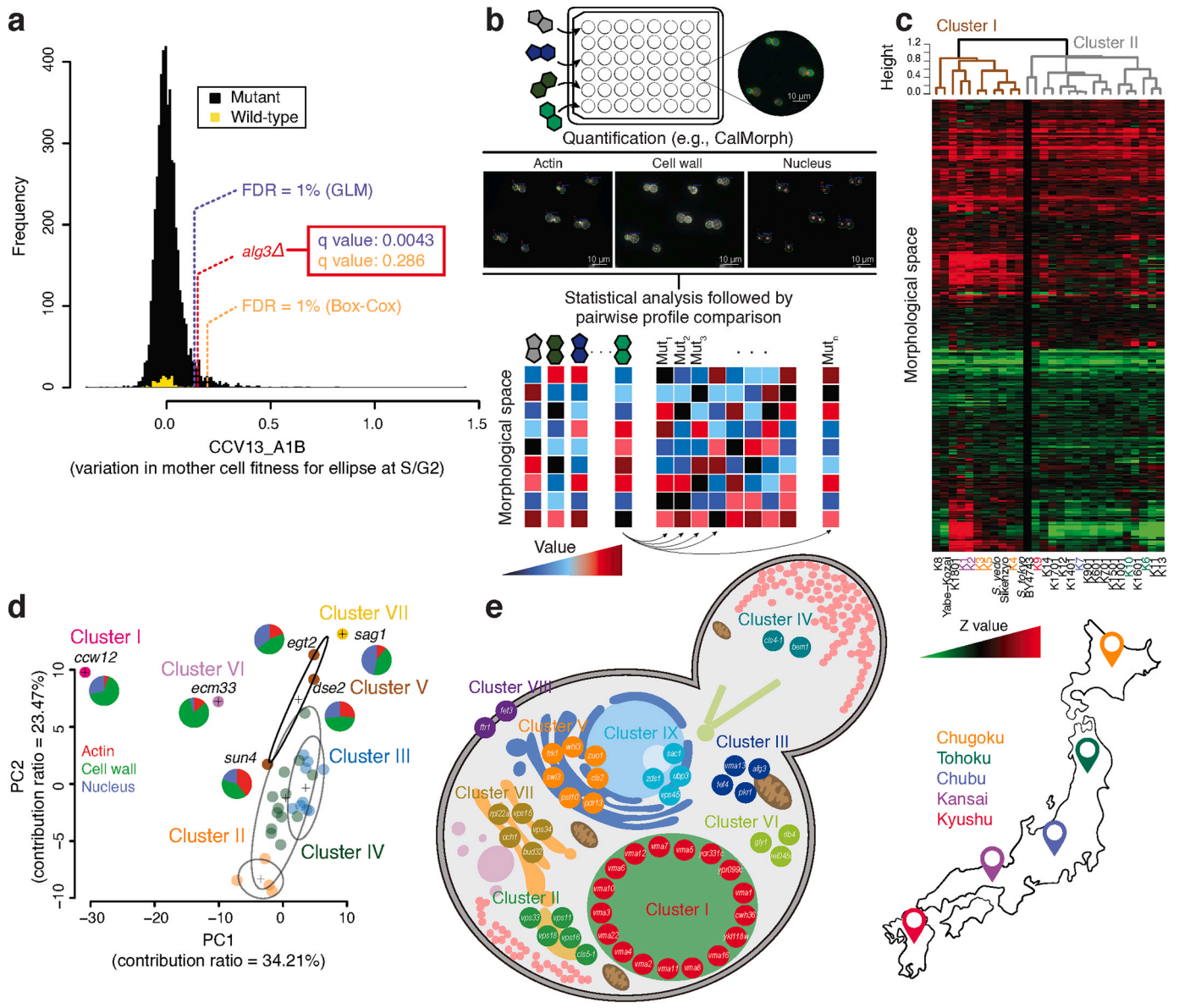

Fig. 5. QMP Applications. **a.** Detection of subtle morphological differences. An example of undetected morphological defect in *alg3*Δ (*ybl082*) cells by Box-Cox method and a histogram illustrating variation (noise values after LOESS smoothing) in mother cell fitness for the ellipse at the S/G2 stage (CCV13_A1B) are presented. For further details, review the CalMorph user manual. Non-essential mutants (4708) and wild-type replicates (109) are shown in black and yellow, respectively. Dashed lines indicate significant thresholds for the GLM (purple) and Box–Cox (orange) methods at FDR = 1 %. Data were obtained from [36,114] **b.** Morphological profiling to predict intracellular targets of compounds. Morphological defects in a mutant caused by loss of gene function can mimic those in cells treated with chemical compounds. An abstract methodology of this process is shown, where morphological profiles of drug-treated cells are compared to a library of mutants. Mutants with significantly high positive correlation values (*e.g.*, Pearson correlation coefficient) are identified as the most probable target. Various computational approaches for generating mechanism of action hypotheses have been reported by [158]. **c.** Morphological phenotyping for phenotypic diagnosis of lineage. Examples of 27 sake yeast strains and BY4743 strain (control). Morphological analysis revealed two clusters (brown and dark gray) at AU (Approximately Unbiased) p-values > 0.95. Strains are color-coded based on their origin, and are shown on the map as light-gray, indicating an unknown origin. **d.** A biplot illustrating the clustering of mannoprotein mutants according to their morphological defects. Morphological space was subjected to Gaussian mixture model clustering with no prior assumptions, resulting in mutants exhibiting similar defects grouped together. Pie charts show the proportion of significant morphological parameters. **e.** Clustering of Ca^{2+}-sensitive mutants based on morphology. An example of chemical-genetic morphological profiles illustrates that mutants with similar morphological defects exhibited similar functions.
(a) The figure has been modified from [112]. (b) The figure has been modified from [37]. (c) The figure has been modified from [38].

subpopulations within tumor cell populations [44,93].

A powerful approach to detect modality at the single-cell level involves grouping cells based on a previously known cellular state. In our laboratories, we use CalMorph to categorize cells into three groups according to their spatial (*i.e.*, mother/bud cells) and temporal (*i.e.*, cell cycle stages) attributes (Fig. 6a). Others applicable approaches include using the cell shape [137,175,2] and cell type [144,145]. Cell heterogeneity can also be dissected prior to imaging using techniques, such as microfluidics [153] and morphology-based cell sorting, which offer several advantages, including reduced cell damage and stress. Further information on these methods is available from published sources [109,116,139,159].

Population level: Morphological measures at the population level reflect dominant biological characteristics influenced by the

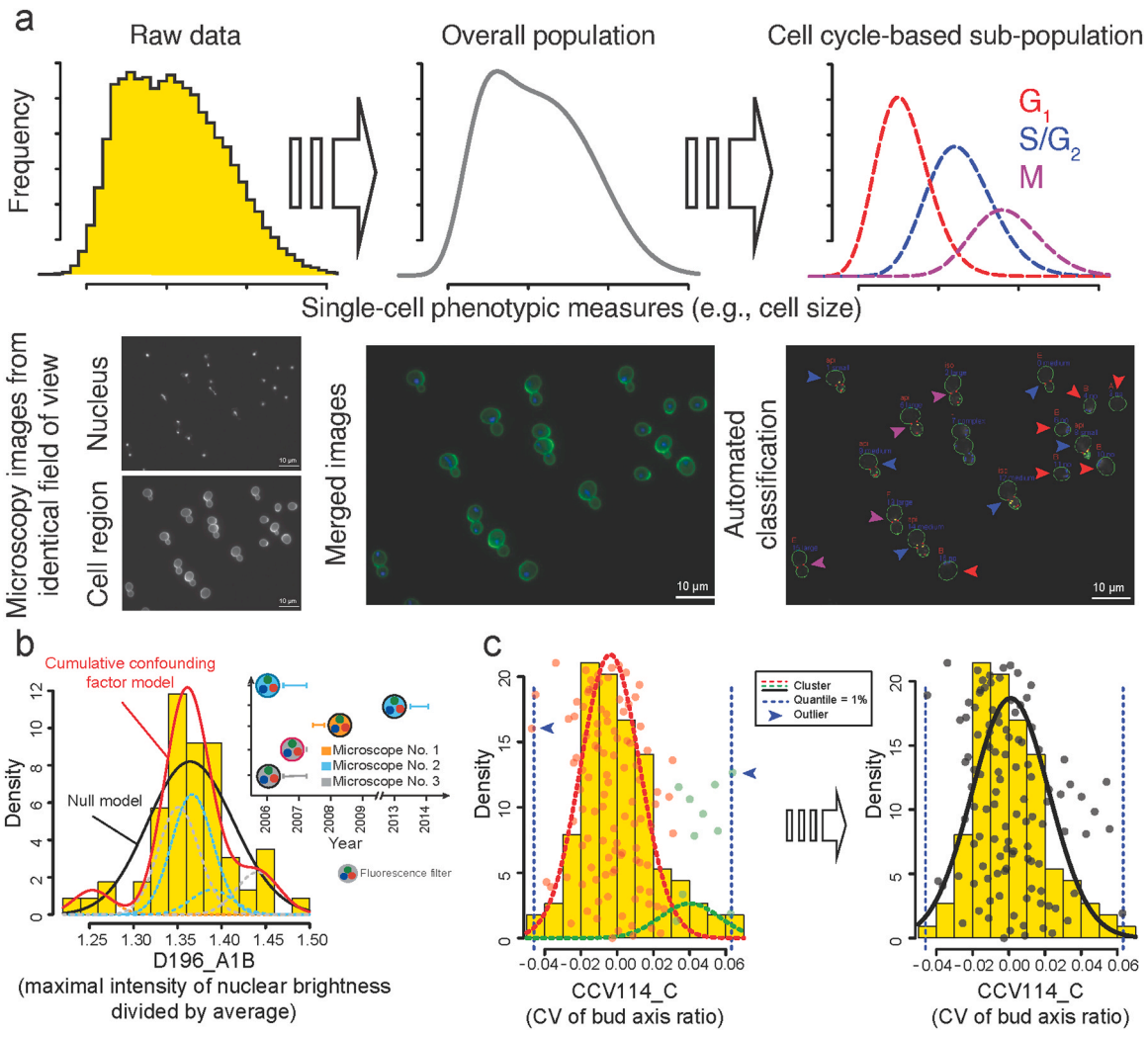

Fig. 6. Data modality concepts. **a.** Histogram depicting the whole-cell size of haploid wild-type (WT) yeast cells (n = 109). The built-in classification algorithm in CalMorph uses both cell shape and nuclear DNA images to identify the nuclear cell cycle phase, categorizing cells into G1, S/G2, and M phases. **b.** An illustration of the effect of confounding factors on data modality. Dashed lines are various experimental conditions; for further detailed information see [111]. The solid black and red lines show the null and overall distributions, respectively, influenced by confounding factors. **c.** Impact of outliers on modality. **Left:** Gaussian mixture modeling of diploid WT yeast cells (n = 114). Each circle is color-coded according to the detected cluster. CCV114_C represents a noise parameter of the "bud axis ratio" (*i.e.*, normally distributed). **Right:** Unimodal distribution after outlier removal.
These figures have been adapted from [36].

experimental conditions. However, it remains unclear whether the population data are truly unimodal. Thus, investigating the possibility of subpopulations using either supervised or unsupervised methods is necessary (Supporting Text). Since reference phenotypes are not always available, unsupervised approaches, such as *k*-means [137,162] and mixture model clustering [133] are more commonly used. These methods organize the population into clusters based on feature similarity, potentially revealing hidden patterns that correspond to different sub-populations.

We previously attempted to characterize modality within a large set of wild-type yeast replicates using probabilistic mixture models with no prior assumption [36]. Our analysis revealed that the predominance of the observed modality could be attributed to confounding variables (Fig. 6b) and outliers (Fig. 6c). Subsequently, by incorporating confounding variables into our statistical model and excluding outliers using strategies, such as the one percentile rule, we enhanced the statistical robustness of subsequent analytical steps.

3.4. *Step 2-4: hypothesis testing*

QMP experiments have a high level of information, but appropriate data analysis is essential to effectively use this information [91]. Morphological data are not always normally distributed (*Step 2–2*). For non-normal data, non-parametric methods, such as Mann-Whitney U test [47,64] (Supplementary Table 9a) are applicable. However, parametric approaches (Supplementary Table 9b) offer improved differentiation and are, therefore, preferable [104,108]. This necessitates the application of generalized linear models (GLMs) that cover additional probability distributions besides the Gaussian distributions, accommodating the data type (*Step 2–2*) and answering biological questions more precisely.

GLM analysis is widely utilized in morphological analysis, as demonstrated in several studies [167,67,85,95]. We primarily employed GLM to answer various questions, such as to identify relevant patterns among morphological profiles and define the intracellular target of a compound (Fig. 5b), study genetic diversity and ecological origins (Fig. 5c), and to investigate risk management in industrial settings [35]. However, the field has used diverse approaches, including supervised

machine learning algorithms (see below) for prediction [10] and capturing non-linear morphological variabilities [2,64], factor analysis [86], and studying morphological diversity on principal component space [6,31]. Additional examples are provided in Supplementary Table 9. Finally, deep learning algorithms (DL; see *Step 5*), which require minimal data transformation, can be more effective in capturing biological information contained within the images [72] and have recently gained widespread acceptance for morphological phenotyping [20].

4. Step 3: knowledge extraction

After applying a statistical modeling method, various ML approaches can be employed to recognize, understand, and obtain new biological knowledge by comparing, categorizing, and predicting (dis)similar morphologies. This section briefly covers the main approaches. Readers are encouraged to review the cited references for additional information.

4.1. Step 3-1: dimensionality reduction

QMP experiments are inherently high-dimensional, generating a detailed morphological spectrum for the phenotype of interest. This complexity poses a significant challenge in identifying the features carrying the most valuable information to address the biological questions. Additionally, a large feature space contributes to higher data sparsity, variance, and multicollinearity due to redundant features [4]. These challenges are typically addressed by DR approaches, which provide a better representation of informative features. DR improves the identification of hidden structures, enhances visualization, facilitates accurate model building, reduces the risk of overfitting, and decreases computational cost (CPU time and memory).

Selecting pertinent features based on prior biological knowledge is the most straightforward but subjective DR method. However, our research [37] has demonstrated that selecting a limited subset of features may not comprehensively represent the entire morphological spectrum. This is evident from the observations that mutations in cell wall components (*e.g.*, *DSE2*, *EGT2*, and *SUN4*) can concurrently affect the morphology of the nucleus and actin structures (Fig. 5d).

There are two main techniques to reduce dimensionality: *feature selection*, which involves reducing dimensionality by selecting a subset of original features to remove redundant or irrelevant features. Maintaining the original features helps preserve data interpretability in downstream models [45,136]. As reviewed previously [17], filtering based on replicate correlation is the most commonly used feature selection approach.

Feature extraction projects the features onto a lower dimensional subspace while maintaining the information content. This transformation, however, can render the resulting feature set less interpretable. PCA is the most commonly choice in biology demonstrating superior performance over alternative DR techniques [123,130]. Nonetheless, PCA primarily focuses on capturing variance (*i.e.*, intergroup variability) and identifies linear relationships. This limitation may restrict its applicability in scenarios that require the detection of complex, non-linear patterns within the data. The applications of various feature selection and feature extraction methods are discussed in the Supporting Text and Supplementary Table 10.

4.2. Step 3-2: downstream analysis

Complementary to data analysis, downstream analysis is the process of discovering, interpreting, and validating biological patterns in morphological profiles using a ML approach, such as clustering and classification. These techniques should not be applied arbitrarily, but selecting the optimal method can sometimes be challenging [41,103]. The choice largely hinges on the access to supplementary information sources and the confidence level in the predictive power [50,83]. In our laboratories, collaboration between computational and experimental biologists consistently enrich our methodologies, facilitating more nuanced and accurate data interpretations.

Morphology-based clustering is a practical tool for investigating cellular and molecular phenomena, and has effectively been employed to study gene functions [107]. Improvements in QMP enable objective clustering of mutants. For instance, hierarchical clustering (HC) is broadly used to identify patterns in morphological data [137,172,26]. We have also previously [38] applied HC to categorize morphological responses of Ca^{2+}-sensitive mutants (Fig. 5e). However, HC requires: 1) a predefined distance metric to estimate (dis)similarity; 2) a linkage criterion to determine how distances between clusters are defined and how clusters are merged at each step; and 3) a cut-off height for the tree. These choices significantly affect the final clustering results. In our experience, probabilistic methods, such as Gaussian mixture model (GMM) clustering, present a preferable alternative for several reasons. Firstly, morphological data often undergoes Z value transformation and/or PCA projection prior to clustering, ensuring data normality. Secondly, GMM uses the Expectation-Maximization algorithm, which aids in optimal determination of the number of clusters and maintaining objectivity [140,61,88]. Clustering methods and their applications are discussed in Supporting Text, Supplementary Table 11, and Fig. S1. Additionally, Giordani and coworkers' book [39] on clustering with R is an application-oriented textbook that covers the most common clustering methods from a theoretical perspective, complemented by various examples.

Classification serves as an essential tool for labeling new samples using pre-existing data [17]. The application of classification in morphological phenotyping is vast, encompassing developmental biology [157], functional genomics [12], decoding disease mechanisms [68,141], and guiding drug discovery strategies [20]. Advanced imaging technologies and computational analysis have significantly enhanced the accuracy and scalability of these classifications [101], for further information, see Supporting Text, Supplementary Table 12 and Fig. S2.

4.3. Step 3-3: knowledge prediction

As discussed previously, cell morphology can be affected by both external and internal stimuli. Thus, predicting cellular networks based on phenomic variations is an important goal in biological research. The primary tools for predicting these relationships include regression, classification [21,69,90], and DL models (see *Label-free imaging* section). However, the predictability of the model and associations among variables, whether linear, nonlinear, or (co)variational, are critical factors in choosing the appropriate tool [82]. No singular model can universally address all challenges, emphasizing the need for adopting a multifaceted strategy. In our approach, building multiple models and rigorously evaluating their predictive capabilities (Supplementary Table 13) was deemed necessary.

Integrating predictive modeling with morphological data has proven potential for advancing our understanding across a broad spectrum of biological and medical fields. This includes studying drug efficacy [143], immune responses [126], cell morphogenesis and differentiation [16,92], underlying causes of diseases [28], and metabolic activities [62]. Advancements in imaging techniques, combined with the advent of more sophisticated analytical algorithms, has accelerated the pace of discovery and opened new avenues for exploring cellular mechanisms with unprecedented depth and scope.

5. Step 4: sharing

The final step in any study is sharing, which offers undeniable benefits for transparency [179], standardization [40,170], archiving [46], and reproducibility [120,60,7]. Scientists greatly benefit from sharing codes and data, known as open science, through public databases or, at a minimum, *via* institutional website or journal supplementary files. Open

science enhances data aggregation methods, bolsters confidence in reported results, provides opportunities for deeper data analysis, and increases the impact of research [135,142,171,40].

Sharing code and scripts is an integral scientific practice that ensures a broader audience for replicating studies, validates findings, and builds further upon existing work. Platforms, such as GitHub and Bitbucket, are commonly used for this purpose, ensuring that the code is accessible, well-documented, and well-maintained. Furthermore, these platforms help educate students and early-career researchers about best practices in coding and research methodologies [19].

In addition to the new data management and sharing policy of the National Institutes of Health (effective since January, 2023), the morphological phenotyping community has increasingly embraced sharing and collaboration (Supplementary Table 14). However, there is still a lack of a unified reference database for consolidating available and future data and preventing redundancies. One solution is to integrate data into community databases (*e.g.*, the *Saccharomyces* Genome Database for budding yeast) or well-known databases (*e.g.*, NCBI), because of their popularity and logistics. Such databases could serve as a benchmark for developing novel methods. Meanwhile, researchers are encouraged to adhere to the FAIR (Findable, Accessible, Interoperable, and Reusable) principles [169] and deposit their data into an appropriate database (Supplementary Table 14a).

Similar principles should apply to image analysis tools to support open research and avoid duplicative efforts. One solution could be to develop a plugin or extension for existing platforms (Supplementary Table 4), if possible. However, developing new tools may be the best option for meeting unprecedented needs. When developing a new image analysis tool, the true costs of development and maintenance, interoperability, and a long-term sustainability plan should be considered [80].

Eventually, scientific image forums (Supplementary Table 14b) are ideal for discussing bioimage analysis-related questions, sharing new bioimage tools or libraries, and fostering collaborations. These forums are typically sponsored by experienced field experts and facilitate learning, problem-solving, and networking.

6. Step 5: transformative knowledge exploration

6.1. Step 5-1: label-free imaging

High-throughput screening *via* fluorescent labeling provides high contrast and specificity. However, challenges, such as labor intensity and potential artifacts or perturbations caused by labeling, limit its applications [10,69]. In contrast, high-throughput label-free imaging overcomes these barriers. In label-free imaging, features are extracted based on variations in the optical properties of cellular and subcellular structures, such as light absorption, scattering, and phase. Although this approach yields lower contrast and resolution compared to labeled images [29], advanced DL methods can extract abundant of data from label-free images with high sensitivity [43,64].

A common method to extract data from bright field images is training neural network classifiers [117,147,41,99], such as convolutional neural networks (CNNs) [151,164,90,96], U-Net [30,150], and region-based CNN (R-CNN) [33,59]. These techniques have been widely adopted for morphological phenotyping and actively contribute to advancements in biomedical problem-solving [23], identification of disease mechanisms [102,138,58], and drug discovery [20]. For further information readers are encouraged to refer to previously published sources [125,163]. Further enhanced capabilities may be achieved by combining DL and high-throughput cell microscopy [109], ghost cytometry [116] and imaging flow cytometry [10], which may eventually replace classical image processing.

While this integration offers abundant opportunities, it also introduces new challenges. Two major limitations can be noted with DL. First, these methods perform reliably when the classifier is trained by a large set, increasing the computational cost and requiring high-quality ground truth data [75,127]. This underscores the need for a well-annotated global database (see *Sharing* section). Second, the training sets for classifier learning are restricted to our knowledge boundaries, which increases the risk of overfitting. Consequently, caution must be exercised while interpreting DL-based results. Students new to DL as well as more experienced individuals looking to broaden their understanding can benefit from the abundant practical examples provided in “Deep learning with Python” [22].

6.2. Step 5-2: towards trans-omics networks

Linking morphological changes to a single omics data type does not fully address the intricate biological questions that require a global understanding of the interplay among various biological domains. A prime example of this complexity is observed in cancer development and progression, where cellular processes are orchestrated by networks spanning multiple omics layers [13,48,65,70]. Such phenomena can be better understood through the integration of multi-omics data into a cohesive framework [52].

While DR techniques are commonly applied to single-omics datasets, methods that concurrently decompose and integrate multiple datasets have been less frequently explored. Meng et al. [97] listed several approaches for integrative analysis of multi-omics data [97]. Among them, we employed canonical correlation analysis (CCA; Supplementary Table 15) in our previous works [36,111] to link morphological variations to functional genomics characteristics (Fig. 7). CCA facilitates the exploration of the relationship between two sets of multidimensional data by identifying their maximally correlated linear combinations. CCA provided us with significant predictive power regarding gene functions based on morphological defects. Remarkably, the obtained precision was comparable with other omics data such as protein interaction and genetic interaction profiles [111].

To our knowledge, there are only a few examples of such system-level integrations in morphological phenotyping. Kurita et al. combined image-based screening of HeLa cells with untargeted metabolomics analysis to predict the identity and mode of action of natural products [76]. Nassiri and McCall [105], Hasle et al. [49], and Way et al. [166] used different approaches to associate morphological profiles with transcriptomic changes to infer and predict mechanisms of action [105,166,49]. The demand for integrating two or more profiles is expected to grow as image-based analysis continues to evolve rapidly using deep learning approaches [124,156,173].

System-level integration of multi-omics studies, *i.e.*, trans-omics models [176], can bring different types of knowledge together and allow models to learn from specific and/or common connections. Such integrative models can offer novel insights if the combined profiles carry complementary information [15,177,98] or facilitate data translation if they are largely redundant. Genomics data have been commonly integrated with epigenomics, transcriptomics, proteomics [106], and chromatography–mass spectrometry data. Despite the potential of these advanced methodologies to impact our understanding of cell morphology, comprehensive incorporation of morphological data into trans-omics models remains an underexplored frontier.

7. Limitations of computational models

Before the advent of computers, exploring complex biological processes through computational studies was largely impractical, primarily due to the large number of equations that needed to be constructed and solved by hand. Nowadays, a wide array of software tools is accessible, enabling modeling and analysis of cellular structures and molecular dynamics. These tools allow researchers to effectively modify model properties to test hypotheses, delve into the causes of specific outcomes, and distinguish differences between models. However, this process is not without its challenges [41].

Determining the most appropriate model is a critical step that hinges

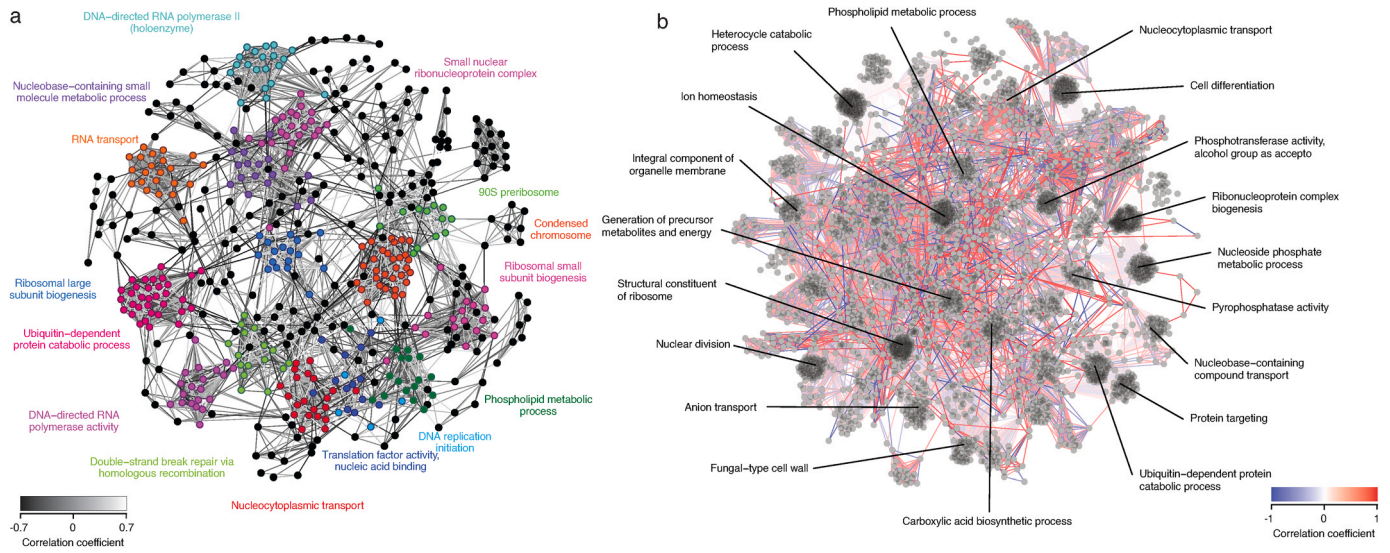

Fig. 7. Linking morphological defects to functional profiles of budding yeast mutants. **a.** A global representation of haplo-insufficient essential genes, illustrated based on the detected morphological abnormalities (Spring layout). Each circle represents a mutant, with edges in black and white indicating negative and positive phenotypic correlations, respectively. The haplo-insufficient gene with similar functions are color-coded. **b.** A graphical representation of 2915 nonessential genes exhibiting morphological defects and their associated functions, depicted using the Spring layout. Each circle represents a mutant, with red and blue lines representing positive and negative correlations, respectively.
(a) This figure has been adapted from [111]. (b) The figure has been adapted from [36].

on a thorough understanding of the biological system and adequate computational expertise to design a structured experiment. Key steps in model development include conceptual model formulation (*i.e.*, cataloging of all variables), counting for potential associations, identifying confounders, and minimizing errors. However, building an efficient model demands significant effort and time, and may necessitate further assumptions or revisions. Poor choices can produce misleading results, emphasizing the importance of diligence in these procedures. Once a computational model is finalized, it must undergo rigorous testing to ensure accuracy and reliability.

Another significant challenge is the vast heterogeneity across datasets. Major sources of variability are sample preparation and different microscopy processes (*e.g.*, varying resolutions, bit depths, magnifications, dimensions (2D or 3D), and multiple light wavelength channels). Adequate data annotation and labeling are essential for mitigating these challenges.

Ultimately, even though computational models offer varying perspectives, they cannot replace wet-lab experiments. Instead, they should be viewed as valuable, complementary tools that enrich our understanding [15,51].

8. Coding dilemma

Morphological phenotyping is an interdisciplinary field that encompasses biology, computer science, optics, microfluidics, and data science. Users with no or limited computational knowledge find it challenging to write or modify scripts in programming languages (mainly Python, R, or MATLAB). However, a general understanding of the principles underlying each method allows researchers without a strong background in data science to critically analyze their data using publicly available tools (Supplementary Table 16). This allows a broader range of users to build their own statistical and ML models, adding a new dimension to their research. As highlighted throughout this review, using these tools without understanding their output can cause misinterpretation of the results, and it is researchers' responsibility to be their own strongest critics.

9. Conclusions

With the wealth of available morphological data, it is challenging to extract the most relevant information without appropriate skills and background. This review aimed to bridge this gap by documenting available methods, providing the principles underlying each step, and their proper applications. The presented workflow offers an extensible approach to enhancing state-of-the-art QMP method.

While routine analytical methods, utilizing various open-source and commercial software (*e.g.*, SPSS), are valued for their simplicity and availability, they often overlook considerable statistical power. Proper use of this power could not only detect subtle changes, but also reduce bias induced by misestimations of true effects. Moreover, a higher statistical power directly impacts reproducibility, ensuring that statistically significant findings accurately reflect the actual effects and their magnitude.

The QMP workflow lays a foundation for further studies, but has its own limitations. Particularly, GLM implementation requires familiarity with a broad range of probability distributions beyond just the Gaussian distribution. Basic programming knowledge is also essential for a successful study. Therefore, the community would benefit from engaging researchers capable of conducting interdisciplinary studies, such as computational biologist, biostatisticians, and data scientists. However, recruiting and retaining experts, as well as providing sufficient time and computational resources to develop and maintain software, all come at a cost.

CRediT authorship contribution statement

Farzan Ghanegolmohammadi: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Methodology, Formal analysis, Data curation, Conceptualization. **Yoshikazu Ohya:** Funding acquisition, Conceptualization. **Mohammad Eslami:** Writing – original draft, Visualization, Validation, Formal analysis, Data curation.

Declaration of Competing Interest

There are no competing financial interests.

Data availability

Plots and figures presented in this paper were obtained by re-analysis of publicly available data from the *Saccharomyces cerevisiae* Morphological Database 2 (http://www.yeast.ib.k.u-tokyo.ac.jp/SCMD/datasheet.php). This data includes replicates of wild-type strains and mutant collections of budding yeast.

All analyses were performed using R or Python. Probability distributions were defined using the gamlss package [148], and the false discovery rate was estimated using the qvalue package [27]. All source code for analyzing and defining unimodal parameters is available at GitHub: (https://github.com/OhyaLab/UNIMO).

Acknowledgments

This work was supported by Grants-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science and Technology, Japan to Y.O. (19H03205 and 22H02216) and a MEXT scholarship to F.G. (160693).

Author contribution

All authors contributed to writing the manuscript and editing the text.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2024.07.012.

References

[1] Altschuler SJ, Wu LF. Cellular heterogeneity: do differences make a difference? Cell 2010;141(4):559–63.
[2] Bakal C, et al. Quantitative morphological signatures define local signaling networks regulating cell morphology. Science 2007;316(5832):1753–6.
[3] Baker M. Reproducibility crisis? Nature 2016;533(26):353–66.
[4] Banerjee J, et al. Machine learning in rare disease. Nat Methods 2023:1–12.
[5] Baryshnikova A, et al. Quantitative analysis of fitness and genetic interactions in yeast on a genome scale. Nat Methods 2010;7(12):1017.
[6] Bauer CR, et al. Essential gene disruptions reveal complex relationships between phenotypic robustness, pleiotropy, and fitness. Mol Syst Biol 2015;11(1):773.
[7] Beaulieu-Jones BK, Greene CS. Reproducibility of computational workflows is automated using continuous analysis. Nat Biotechnol 2017;35(4):342–6.
[8] Bertolet A, et al. The complexity of DNA damage by radiation follows a Gamma distribution: insights from the Microdosimetric Gamma Model. Front Oncol 2023;13:1196502.
[9] Birmingham A, et al. Statistical methods for analysis of high-throughput RNA interference screens. Nat Methods 2009;6(8):569–75.
[10] Blasi T, et al. Label-free cell cycle analysis for high-throughput imaging flow cytometry. Nat Commun 2016;7(1):1–9.
[11] Bolstad BM, et al. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 2003;19(2):185–93.
[12] Bougen-Zhukov N, et al. Large-scale image-based screening and profiling of cellular phenotypes. Cytom Part A 2017;91(2):115–25.
[13] Boyle EA, et al. An expanded view of complex traits: from polygenic to omnigenic. Cell 2017;169(7):1177–86.
[14] Brideau C, et al. Improved statistical methods for hit selection in high-throughput screening. J Biomol Screen 2003;8(6):634–47.
[15] Brodland GW. How computational models can help unlock biological systems. Seminars in cell & developmental biology. Elsevier,; 2015.
[16] Buggenthin F, et al. Prospective identification of hematopoietic lineage choice by deep learning. Nat Methods 2017;14(4):403–6.
[17] Caicedo JC, et al. Data-analysis strategies for image-based cell profiling. Nat Methods 2017;14(9):849.
[18] Caicedo JC, et al. Nucleus segmentation across imaging experiments: the 2018 Data Science Bowl. Nat Methods 2019;16(12):1247–53.
[19] Chakraborty S, Aithal P. A Practical Approach to GIT Using Bitbucket, GitHub and SourceTree. Int J Appl Eng Manag Lett (IJAEML) 6(2 2022:254–63.
[20] Chandrasekaran SN, et al. Image-based profiling for drug discovery: due for a machine-learning upgrade? Nat Rev Drug Discov 2021;20(2):145–59.
[21] Ching T, et al. Opportunities and obstacles for deep learning in biology and medicine. J R Soc Interface 2018;15(141):20170387.
[22] Chollet F. Deep learning with Python. Simon and Schuster,; 2021.
[23] Christensen BC, et al. update in progress. Opportunities and obstacles for deep learning in biology and medicine. Manubot,; 2021.
[24] Cox MJ, et al. Tales of 1,008 Small Molecules: Phenomic Profiling through Live-cell Imaging in a Panel of Reporter Cell Lines. bioRxiv 2020.
[25] Crawley MJ. The R book. John Wiley & Sons,; 2012.
[26] Cutiongco MF, et al. Predicting gene expression using morphological cell responses to nanotopography. Nat Commun 2020;11(1):1384.
[27] Dabney A, et al. qvalue: Q-value estimation for false discovery rate control. R Package Version 2010;1(0).
[28] Eddy CZ, et al. Morphodynamics facilitate cancer cells to navigate 3D extracellular matrix. Sci Rep 2021;11(1):20434.
[29] Eliceiri KW, et al. Biological imaging software tools. Nat Methods 2012;9(7):697.
[30] Falk T, et al. U-Net: deep learning for cell counting, detection, and morphometry. Nat Methods 2019;16(1):67–70.
[31] Farkas Z, et al. Gene loss and compensatory evolution promotes the emergence of morphological novelties in budding yeast. Nat Ecol Evol 2022;6(6):763–73.
[32] Fidelis CR, et al. Reparametrized generalized gamma partially linear regression with application to breast cancer data. J Appl Stat 2024:1–18.
[33] Fujita S, Han X-H. Cell detection and segmentation in microscopy images with improved mask R-CNN. Proc Asian Conf Comput Vis 2020.
[34] Gareth J, et al. An introduction to statistical learning: with applications in R. Spinger; 2013.
[35] Ghanegolmohammadi F, et al. Single-Cell Phenomics in Budding Yeast: Technologies and Applications. Single-Cell Omics. Elsevier; 2019. p. 355–79.
[36] Ghanegolmohammadi F, et al. Assignment of unimodal probability distribution models for quantitative morphological phenotyping. BMC Biol 2022;20(1):1–13.
[37] Ghanegolmohammadi F, et al. Defining functions of mannoproteins in *Saccharomyces cerevisiae* by high-dimensional morphological phenotyping. J Fungi 2021;7(9):769.
[38] Ghanegolmohammadi F, et al. Systematic analysis of Ca^{2+} homeostasis in *Saccharomyces cerevisiae* based on chemical-genetic interaction profiles. Mol Biol Cell 2017;28(23):3415–27.
[39] Giordani P, et al. Introduction to clustering. Springer; 2020.
[40] Gonzalez-Beltran AN, et al. Community standards for open cell migration data. GigaScience 2020;9(5). giaa041.
[41] Greener JG, et al. A guide to machine learning for biologists. Nat Rev Mol Cell Biol 2022;23(1):40–55.
[42] Guo Q, et al. A novel edge effect detection method for real-time cellular analyzer using functional principal component analysis. IEEE/ACM Trans Comput Biol Bioinforma 2019.
[43] Gupta A, et al. Deep learning in image cytometry: a review. Cytom Part A 2019;95(4):366–80.
[44] Haffner MC, et al. Genomic and phenotypic heterogeneity in prostate cancer. Nat Rev Urol 2021;18(2):79–92.
[45] Hancer E, et al. A survey on feature selection approaches for clustering. Artif Intell Rev 2020:1–27.
[46] Hanson B, et al. Making data maximally available, American Association for the Advancement of. Am Assoc Adv Sci 2011.
[47] Hartmann J, et al. An image-based data-driven analysis of cellular architecture in a developing tissue. Elife 2020;9:e55913.
[48] Hasin Y, et al. Multi-omics approaches to disease. Genome Biol 2017;18(1):83.
[49] Hasle N, et al. High-throughput, microscope-based sorting to dissect cellular heterogeneity. Mol Syst Biol 2020;16(6):e9442.
[50] Hastie T, et al. The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media; 2009.
[51] Henley SS, et al. Statistical modeling methods: challenges and strategies. Biostat Epidemiol 2020;4(1):105–39.
[52] Heumos L, et al. Best practices for single-cell analysis across modalities. Nat Rev Genet 2023;24(8):550–72.
[53] Hiley C, et al. Deciphering intratumor heterogeneity and temporal acquisition of driver events to refine precision medicine. Genome Biol 2014;15:1–10.
[54] Ho W-C, Zhang J. Evolutionary adaptations to new environments generally reverse plastic phenotypic changes. Nat Commun 2018;9(1):1–11.
[55] Horvath P, et al. Machine learning improves the precision and robustness of high-content screens: using nonlinear multiparametric methods to analyze screening results. J Biomol Screen 2011;16(9):1059–67.
[56] Houle D, et al. Phenomics: the next challenge. Nat Rev Genet 2010;11(12):855–66.
[57] Hu M-x, Salvucci S. A study of imputation algorithms. Work Pap Ser 2001.
[58] Huang S, et al. Artificial intelligence in cancer diagnosis and prognosis: opportunities and challenges. Cancer Lett 2020;471:61–71.
[59] Hung J, et al. Keras R-CNN: library for cell detection in biological images using deep neural networks. BMC Bioinforma 2020;21(1):1–7.
[60] Ince DC, et al. The case for open computer programs. Nature 2012;482(7386):485–8.
[61] Inoue H, et al. Automatic quantitative segmentation of myotubes reveals single-cell dynamics of S6 kinase activation. Cell Struct Funct 2018;43(2):153–69.
[62] Itto-Nakama K, et al. AI-based forecasting of ethanol fermentation using yeast morphological data. Biosci, Biotechnol, Biochem 2022;86(1):125–34.
[63] James G, et al. An introduction to statistical learning: With applications in python. Springer Nature; 2023.
[64] Jang J, et al. A deep learning-based segmentation pipeline for profiling cellular morphodynamics using multiple types of live cell microscopy. Cell Rep Methods 2021;1(7).
[65] Jung GT, et al. How to interpret and integrate multi-omics data at systems level. Anim Cells Syst 2020;24(1):1–7.

[66] Kamentsky L, et al. Improved structure, function and compatibility for CellProfiler: modular high-throughput image analysis software. Bioinformatics 2011;27(8):1179–80.
[67] Kardum Hjort C, et al. Morphological variation in bumblebees (Bombus terrestris) (Hymenoptera: Apidae) after three decades of an Island invasion. J Insect Sci 2023;23(1):10.
[68] Kelley ME, et al. High-content microscopy reveals a morphological signature of bortezomib resistance. Elife 2023;12:e91362.
[69] Kobayashi-Kirschvink KJ, et al. Linear regression links transcriptomic data and cellular raman spectra. Cell Syst 2018;7(1):e104. 104-117.
[70] Koseki J, et al. Comput Anal Cancer Biol Based Exhaust Exp Backgr 2019.
[71] Kosmicki JA, et al. Discovery of rare variants for complex phenotypes. Hum Genet 2016;135(6):625–34.
[72] Kraus OZ, et al. Automated analysis of high-content microscopy data with deep learning. Mol Syst Biol 2017;13(4).
[73] Krzywinski M, Altman N. Points of significance: power and sample size. Nat Publ Group 2013.
[74] Kuhn M, Johnson K. Applied predictive modeling. Springer; 2013.
[75] Kulikov V, et al. DoGNet: a deep architecture for synapse detection in multiplexed fluorescence images. PLoS Comput Biol 2019;15(5):e1007012.
[76] Kurita KL, et al. Integration of high-content screening and untargeted metabolomics for comprehensive functional annotation of natural product libraries. Proc Natl Acad Sci 2015;112(39):11999–2004.
[77] Kwak SK, Kim JH. Statistical data preparation: management of missing values and outliers. Korean J Anesthesiol 2017;70(4):407.
[78] Lapins M, Spjuth O. Evaluation of gene expression and phenotypic profiling data as quantitative descriptors for predicting drug targets and mechanisms of action. bioRxiv 2019:580654.
[79] Lee I. Probabilistic functional gene societies. Prog Biophys Mol Biol 2011;106(2):435–42.
[80] Levet F, et al. Developing open-source software for bioimage analysis: opportunities and challenges. F1000Research 2021;10.
[81] Levy SF, Siegal ML. Network hubs buffer environmental variation in *Saccharomyces cerevisiae*. PLoS Biol 2008;6(11).
[82] Li Y, et al. Cell morphology-based machine learning models for human cell state classification. npj Syst Biol Appl 2021;7(1):23.
[83] Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nat Rev Genet 2015;16(6):321–32.
[84] Liberali P, et al. Single-cell and multivariate approaches in genetic perturbation screens. Nat Rev Genet 2015;16(1):18–32.
[85] Lin Y, et al. Cucurbitacin B exerts antiaging effects in yeast by regulating autophagy and oxidative stress. Oxid Med Cell Longev 2019;2019.
[86] Ljosa V, et al. Comparison of methods for image-based profiling of cellular morphological responses to small-molecule treatment. J Biomol Screen 2013;18(10):1321–9.
[87] Loo L-H, et al. An approach for extensibly profiling the molecular states of cellular subpopulations. Nat Methods 2009;6(10):759.
[88] Luengo-Sanchez S, et al. A univocal definition of the neuronal soma morphology using Gaussian mixture models. Front Neuroanat 2015;9:137.
[89] Lundholt BK, et al. A simple technique for reducing edge effect in cell-based assays. J Biomol Screen 2003;8(5):566–70.
[90] Ma J, et al. Using deep learning to model the hierarchical structure and function of a cell. Nat Methods 2018;15(4):290.
[91] Makin TR, de Xivry J-JO. Ten common statistical mistakes to watch out for when writing or reviewing a manuscript. Elife 2019;8.
[92] Mao Y, Green JB. Systems morphodynamics: understanding the development of tissue hardware. R Soc 2017;372:20160505.
[93] Marusyk A, et al. Intra-tumour heterogeneity: a looking glass for cancer? Nat Rev Cancer 2012;12(5):323–34.
[94] Mattiazzi Usaj M, et al. Systematic genetics and single-cell imaging reveal widespread morphological pleiotropy and cell-to-cell variability. Mol Syst Biol 2020;16(2):e9243.
[95] May-Tec AL, et al. Differential blood cells associated with parasitism in the wild puffer fish Lagocephalus laevigatus (Tetraodontiformes) of the Campeche Coast, southern Mexico. Parasitol Res 2024;123(1):24.
[96] Mencattini A, et al. Discovering the hidden messages within cell trajectories using a deep learning approach for in vitro evaluation of cancer drug treatments. Sci Rep 2020;10(1):1–11.
[97] Meng C, et al. Dimension reduction techniques for the integrative analysis of multi-omics data. Brief Bioinforma 2016;17(4):628–41.
[98] Misra BB, et al. Integrated omics: tools, advances and future approaches. J Mol Endocrinol 2019;62(1):R21–45.
[99] Moen E, et al. Deep learning for cellular image analysis. Nat Methods 2019;16(12):1233–46.
[100] Molnar C, et al. Accurate morphology preserving segmentation of overlapping cells based on active contours. Sci Rep 2016;6:32412.
[101] Moshkov N, et al. Learning representations for image-based profiling of perturbations. Nat Commun 2024;15(1):1594.
[102] Munir K, et al. Cancer diagnosis using deep learning: a bibliographic review. Cancers 2019;11(9):1235.
[103] Murphy RF. An active role for machine learning in drug development. Nat Chem Biol 2011;7(6):327.
[104] Nakagawa S, Schielzeth H. Repeatability for Gaussian and non-Gaussian data: a practical guide for biologists. Biol Rev 2010;85(4):935–56.
[105] Nassiri I, McCall MN. Systematic exploration of cell morphological phenotypes associated with a transcriptomic query. Nucleic Acids Res 2018;46(19):e116. e116.
[106] Nesvizhskii AI. Proteogenomics: concepts, applications and computational strategies. Nat Methods 2014;11(11):1114–25.
[107] Neumann B, et al. Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes. Nature 2010;464(7289):721–7.
[108] Nicholson G, Holmes C. A note on statistical repeatability and study design for high-throughput assays. Stat Med 2017;36(5):790–8.
[109] Nitta N, et al. Intelligent image-activated cell sorting. Cell 2018;175(1):e213. 266-276.
[110] Ohairwe ME, et al. A fitness landscape instability governs the morphological diversity of tip-growing cells. Cell Rep 2024;43(4).
[111] Ohnuki S, Ohya Y. High-dimensional single-cell phenotyping reveals extensive haploinsufficiency. PLoS Biol 2018;16(5).
[112] Ohnuki S, et al. Phenotypic diagnosis of lineage and differentiation during sake yeast breeding. G3: Genes, Genomes, Genet 2017;7(8):2807–20.
[113] Ohya Y, et al. Application of unimodal probability distribution models for morphological phenotyping of budding yeast. FEMS yeast Res: foad056 2024.
[114] Ohya Y, et al. High-dimensional and large-scale phenotyping of yeast mutants. Proc Natl Acad Sci 2005;102(52):19015–20.
[115] Oshio K, et al. Interpretation of diffusion MR imaging data using a gamma distribution model. Magn Reson Med Sci 2014;13(3):191–5.
[116] Ota S, et al. Ghost cytometry. Science 2018;360(6394):1246–51.
[117] Ounkomol C, et al. Label-free prediction of three-dimensional fluorescence images from transmitted-light microscopy. Nat Methods 2018;15(11):917–20.
[118] Park S, Khan S. arXiv preprint. GSSMD: N Metr Robust Interpret Assay Qual Assess Hit Sel 2020. arXiv:2001.06384.
[119] Pelkmans L. Using cell-to-cell variability—a new era in molecular biology. Science 2012;336(6080):425–6.
[120] Peng RD. Reproducible research in computational science. Science 2011;334(6060):1226–7.
[121] Perez H, Tah JH. Improving the accuracy of convolutional neural networks by identifying and removing outlier images in datasets using t-SNE. Mathematics 2020;8(5):662.
[122] Perlman ZE, et al. Multidimensional drug profiling by automated microscopy. Science 2004;306(5699):1194–8.
[123] Pincus Z, Theriot J. Comparison of quantitative methods for cell-shape analysis. J Microsc 2007;227(2):140–56.
[124] Piran Z, et al. Disentanglement of single-cell data with biolord. Nat Biotechnol 2024:1–6.
[125] Pratapa A, et al. Image-based cell phenotyping with deep learning. Curr Opin Chem Biol 2021;65:9–17.
[126] Qu C, et al. Mitochondria in the biology, pathogenesis, and treatment of hepatitis virus infections. Rev Med Virol 2019;29(5):e2075.
[127] Ramezani M, et al. A genome-wide atlas of human cell morphology. bioRxiv 2023.
[128] Raunig DL, et al. Quantitative imaging biomarkers: a review of statistical methods for technical performance assessment. Stat Methods Med Res 2015;24(1):27–67.
[129] Reisen F, et al. Linking phenotypes and modes of action through high-content screen fingerprints. Assay Drug Dev Technol 2015;13(7):415–27.
[130] Reisen F, et al. Benchmarking of multivariate similarity measures for high-content screening fingerprints in phenotypic drug discovery. J Biomol Screen 2013;18(10):1284–97.
[131] Rigby RA, et al. Distributions for modeling location, scale, and shape: Using GAMLSS in R. Chapman and Hall/CRC; 2019.
[132] Rofatto VF, et al. A Monte Carlo-Based Outlier Diagnosis Method for Sensitivity Analysis. Remote Sens 2020;12(5):860.
[133] Rohban MH, et al. Capturing single-cell heterogeneity via data fusion improves image-based profiling. Nat Commun 2019;10(1):1–6.
[134] Rousseeuw PJ, Leroy AM. Robust regression and outlier detection. John wiley & sons; 2005.
[135] Rueden CT, et al. Scientific community image forum: a discussion forum for scientific image software. PLoS Biol 2019;17(6):e3000340.
[136] Saeys Y, et al. A review of feature selection techniques in bioinformatics. Bioinformatics 2007;23(19):2507–17.
[137] Sailem H, et al. Cross-talk between Rho and Rac GTPases drives deterministic exploration of cellular shape space and morphological heterogeneity. Open Biol 2014;4(1):130132.
[138] Schneider L, et al. Integration of deep learning-based image analysis and genomic data in cancer pathology: A systematic review. Eur J Cancer 2022;160:80–91.
[139] Schraivogel D, et al. High-speed fluorescence image–enabled cell sorting. Science 2022;375(6578):315–20.
[140] Scrucca L, et al. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. R J 2016;8(1):289.
[141] Seal S, et al. Insights into drug cardiotoxicity from biological and chemical data: the first public classifiers for FDA drug-induced cardiotoxicity rank. J Chem Inf Model 2024.
[142] Shen H. Interactive notebooks: Sharing the code. Nature 2014;515(7525):151–2.
[143] Simm J, et al. Repurposing high-throughput image assays enables biological activity prediction for drug discovery. Cell Chem Biol 2018;25(5):e613. 611-618.
[144] Singh DK, et al. Patterns of basal signaling heterogeneity can distinguish cellular populations with different drug sensitivities. Mol Syst Biol 2010;6(1).
[145] Slack MD, et al. Characterizing heterogeneous cellular responses to perturbations. Proc Natl Acad Sci 2008;105(49):19306–11.

[146] Snijder B, Pelkmans L. Origins of regulated cell-to-cell variability. Nat Rev Mol Cell Biol 2011;12(2):119–25.
[147] Srinidhi CL, et al. Deep neural network models for computational histopathology: A survey. Med Image Anal 2021;67:101813.
[148] Stasinopoulos DM, Rigby RA. Generalized additive models for location scale and shape (GAMLSS) in R. J Stat Softw 2007;23(7):1–46.
[149] Stirling DR, et al. CellProfiler 4: improvements in speed, utility and usability. BMC Bioinforma 2021;22(1):1–11.
[150] Stringer C, et al. Cellpose: a generalist algorithm for cellular segmentation. Nat Methods 2021;18(1):100–6.
[151] Sun J, et al. Deep learning-based single-cell optical image studies. Cytom Part A 2020;97(3):226–40.
[152] Suzuki G, et al. Global study of holistic morphological effectors in the budding yeast Saccharomyces cerevisiae. BMC Genom 2018;19(1):149.
[153] Széles E, et al. Microfluidic platforms designed for morphological and photosynthetic investigations of Chlamydomonas reinhardtii on a Single-Cell Level. Cells 2022;11(2):285.
[154] Taylor MB, Ehrenreich IM. Genetic interactions involving five or more genes contribute to a complex trait in yeast. PLoS Genet 2014;10(5):e1004324.
[155] Tegtmeyer M, et al. High-dimensional phenotyping to define the genetic basis of cellular morphology. Nat Commun 2024;15(1):347.
[156] Ternes L, et al. A multi-encoder variational autoencoder controls multiple transformational features in single-cell image analysis. Commun Biol 2022;5(1):255.
[157] Thomas OO, et al. Automated morphological phenotyping using learned shape descriptors and functional maps: A novel approach to geometric morphometrics. PLoS Comput Biol 2023;19(1):e1009061.
[158] Trapotsi M-A, et al. Computational analyses of mechanism of action (MoA): data, methods and integration. RSC Chem Biol 2022;3(2):170–200.
[159] Ugawa M, et al. In silico-labeled ghost cytometry. Elife 2021;10:e67660.
[160] Usaj MM, et al. High-content screening for quantitative cell biology. Trends Cell Biol 2016;26(8):598–611.
[161] Vaisipour S. Detect, correcting, Prev batch Eff multi-site data, a Focus gene Expr Micro 2014.
[162] Volz HC, et al. Single-cell phenotyping of human induced pluripotent stem cells by high-throughput imaging. bioRxiv: 026955 2015.
[163] von Chamier L, et al. Artificial intelligence for microscopy: what you should know. Biochem Soc Trans 2019;47(4):1029–40.
[164] Warchal SJ, et al. Evaluation of machine learning classifiers to predict compound mechanism of action when transferred across distinct cell lines. SLAS DISCOVERY: Adv Life Sci RD 2019;24(3):224–33.
[165] Watt WB, Dean AM. Molecular-functional studies of adaptive genetic variation in prokaryotes and eukaryotes. Annu Rev Genet 2000;34(1):593–622.
[166] Way GP, et al. Morphology and gene expression profiling provide complementary information for mapping cell state. Cell Syst 2022;13(11):911–23. e919.
[167] Whittaker BA, et al. Zebra finches have style: nest morphology is repeatable and associated with experience. iScience 2023;26(11).
[168] Wiesmann V, et al. Review of free software tools for image analysis of fluorescence cell micrographs. J Microsc 2015;257(1):39–53.
[169] Wilkinson MD, et al. The FAIR guiding principles for scientific data management and stewardship. Sci data 2016;3(1):1–9.
[170] Williams E, et al. Image data resource: a bioimage data integration and publication platform. Nat Methods 2017;14(8):775.
[171] Wilson SL, et al. Sharing biological data: why, when, and how. FEBS Lett 2021;595(7):847–63.
[172] Wu P-H, et al. Single-cell morphology encodes metastatic potential. Sci Adv 2020;6(4). eaaw6938.
[173] Yang KD, et al. Multi-domain translation between single-cell imaging and sequencing data using autoencoders. Nat Commun 2021;12(1):31.
[174] Yang M, et al. Unveiling nonessential gene deletions that confer significant morphological phenotypes beyond natural yeast strains. BMC Genom 2014;15(1):932.
[175] Yin Z, et al. A screen for morphological complexity identifies regulators of switch-like transitions between discrete cell shapes. Nat Cell Biol 2013;15(7):860–71.
[176] Yugi K, et al. Trans-omics: how to reconstruct biochemical networks across multiple 'omic'layers. Trends Biotechnol 2016;34(4):276–90.
[177] Zeng ISL, Lumley T. Review of statistical learning methods in integrated omics studies (an integrated information science). Bioinforma Biol Insights 2018;12.1177932218759292.
[178] Zhai G, Min X. Perceptual image quality assessment: a survey. Sci China Inf Sci 2020;63(11):1–52.
[179] Zhang Y, et al. Crop phenomics: current status and perspectives. Front Plant Sci 2019;10:714.