Synthetic Physical Interactions with the yeast centrosome - Supplementary Information

May 9, 2019

Supplementary Methods

Theory

Mixture models have been proposed as an alternative to calculating p-values based on the assumption that data is normally distributed Efron [2004] and have previously been used to analyse genome-wide datasets. The theory behind their use is that genome-wide screens are conducted in order to identify genes involved in a given process and that this divides the genome into two categories: those that are involved in this process (hits) and those that aren't. Typically, non-hits will have a normal distribution centred around 0, due to variation caused by inherent noise in the system. In contrast, measurement of each of the hits can be thought of as a sample of a normal distribution with mean (and potentially variance) determined by the individual hit. In combination, these hits will form a distribution with properties that will depend on the underlying biology of the screen. The aim of analysing genome-wide screen data is to distinguish these two categories. If there are few enough hits, they will simply form a tail at the edge of the distribution of non-hits and will not significantly effect the mean or standard deviation of the overall distribution. However, when there are significant numbers of hits, they will effect these summary statistics and a fitted normal distribution is unlikely to accurately reflect the real distribution of non-hits. This will render methods based on this approximation, such as the calculation of p-values and application of Z-transformations, inaccurate. The mixture model approach attempts to overcome this limitation by directly identifying the distribution of each of the two categories. Efron's original method [Efron, 2004] involved fitting a normal component to the central peak of the data, representing non-hits, based on the shape of this peak. He then estimated the distribution of the hit peak from the difference between the overall distribution and the fitted null distribution. A limitation of this approach is that the null model is fitted to a relatively small region of the distribution of non-hits and furthermore, it gives no information about the distribution of the hits. In this study, we fitted two normal modes to the data, using an Expectation-Maximization (ME) algorithm, which iteratively improves the fit of the model based on the likelihood of the generating the observed data from the given model. This means all of the data is used to fit the model and the end result is a parameterised model of the distribution of the hits which can be used to compare different genome-wide screens.

Fitting

We fit two-peak normal mixture models to the smoothed LGR data for each of the screens, using the Mclust package Scrucca et al. [2016], which uses an ME algorithm to fit the model. The model fitting process yields 6 parameters: $\rho_1, \rho_2, \mu_1, \mu_2, \sigma_1, \sigma_2$ which fully define the mixture model. A table of all parameters of fitted models is included in Supplementary Table S1.

Peak Identification

After fitting, we distinguished two types of fit: good fits that had two clearly defined distributions representing hits and non-hits; and poor fits where the distributions were not clearly defined. These poor fits were defined as those in which

$$\mu_2 < \mu_1 + 1.5\sigma_1,$$

these screens were excluded from further analysis with mixture models, in the supplementary data these are referred to as "failed" fits. In the remaining 20 cases where the fit was good, we identified the "hit peak" as the peak shifted furthest to the right and the distribution of non-hits, or "central peak" as the leftmost distribution. We refer to these two components of the distribution as C_1 for the central peak and C_2 for the hit peak. We can consider the genome-wide screen as a process for assigning LGRs to particular genes, the first step of this process is to decide whether the gene is a hit or not, which is a Bernouilli variable or weighted coin flip, where the probability of being a hit is given by ρ_2. Then a gene G_i has identity I_i given by:

$$\mathbb{P}(I_i = C_k) = \begin{cases} \rho_1, & k = 1 \\ \rho_2, & k = 2 \end{cases} .$$

Once the identity is determined, the measured LGR, LGR_i, is assigned as a normal variable distributed with mean and standard deviation μ_1, σ_1 or μ_2, σ_2 as determined by the category in which the gene was placed.

We wanted to define metrics to inform about the significance of results. In some cases we wish to draw a line that distinguishes LGRs from hits and non-hits and these metrics allow for such definitions. While cutoffs are a widely used tool and help to focus on significant results, they will always be to some extent arbitrary, as cases on the border may be placed either side by chance. On top of this, the strength of the interaction will vary depending on the particular genes, and depending on the application we may want only strong hits or we may want to include more subtle phenotypes. Therefore we propose different metrics to give a fuller picture of the data and so that a relevant metric can be chosen depending on context.

p-value and Adjustments

The central peak of the distribution provides a natural null model for the data and this can be used to calculate a p-value for a given LGR, x:

$$p(x) = \mathbb{P}(LGR_i > x | I_i = C_1) = \int_x^\infty f_{LGR_i|I_i=C_1}(z)dz,$$

where $f_X(x)$ represents the probability distribution function of the random variable X. This value gives a measure of the probability that a given LGR would have been measured if the identity of gene G_i was the central peak C_1. Genome-wide screens test multiple hypotheses so we may adjust the p-values to account for this, using for example either Bonferroni or FDR q-value adjustments [Benjamini and Hochberg, 1995]. A p-value of 0.05 is generally considered to be the cutoff for significance.

Probability of Inclusion

As the intention of a genome-wide screen is to distinguish hits from non-hits, rather than considering the p-value we can consider the probability of inclusion in a given category. For a given LGR, x, the probability of inclusion in Component 2 is:

$$q(x) = \mathbb{P}(I_i = C_2 | LGR_i = x).$$

By Bayes' theorem

$$q(x) = \frac{f_{LGR_i|I_i=C_2}(x)\mathbb{P}(I_i = C_2)}{f_{LGR_i}(x)},$$

where $f_{LGR_i|I_i=C_2}(x)$ and $f_{LGR_i}(x)$ can be calculated from the fitted distributions. A sensible cutoff according to this approach is the point where a given gene is more likely to belong to Component 2 than Component 1, in other words $q(x) = 0.5$. We refer to this cutoff as $L_{q,0.5}$.

Validation prediction

We validated our SPI screens against GFP-free controls, however this can be a time-consuming activity and so we developed analytical methods to predict the probability of validation. A strain is considered to be a validated hit if its retested LGR exceeds the mean plus two standard deviations of the LGRs of GFP-free controls on the plate. Note this is different to the methodology of Berry et al. [2016], in which the maximum LGR of the GFP-free controls was used as a cutoff. We define the probability of validation for a given LGR, x to be :

$$p_V(x) = \mathbb{P}(LGR_i^V > K|LGR_i = x).$$

Using the law of total probability and conditioning on which of the categories gene G_i belongs to,

$$\begin{aligned} p_V(x) = {} & \mathbb{P}(LGR_i^V > K|I_i = C_1)\mathbb{P}(I_i = C_1|LGR_i = x) \\ & + \mathbb{P}(LGR_i^V > K|I_i = C_2, LGR_i = x)\mathbb{P}(I_i = C_2|LGR_i = x). \end{aligned}$$

These values may all be simply calculated from the fitted mixture model, with the exception of $\mathbb{P}(LGR_i^V > K|I_i = C_2, LGR_i = x)$. We assume that

$$\mathbb{P}(LGR_i^V > K|I_i = C_2, LGR_i = x) \sim \text{Normal}\left(\mu = x, \sigma^2 = \frac{\alpha(\sigma_2)^2}{4}\right),$$

where α is a tunable parameter. We chose to centre the distribution on the original measurement of the LGR based on our observation that generally validation LGRs are similar to the genome-wide screen values. The variance of this distribution is not trivial to describe as it represents both noise in the system and batch effects. We chose to use $\frac{\alpha(\sigma_2)^2}{4}$, where the factor of four is derived from the higher density of colonies (16 rather than 4) used in the retest, and α is a tunable parameter representing batch effects. We found good accuracy using $\alpha = 4$ and used this in all analysis.

We found that $p_V(x)$ performed well at predicting validation rate and FPR, with some exceptions (see main text). We propose that the curve $p_V(x)$ could be used as a tool when making decisions about how many results to validate in a genome-wide screen.

Code accessibility

R scripts for data formatting and analysis are freely available at `https://github.com/RowanHowell/data-analysis`.

References

Y. Benjamini and Y. Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B Methodological*, 57(1):289–300, Oct. 1995.

L. K. Berry, G. Ólafsson, E. Ledesma-Fernandez, and P. H. Thorpe. Synthetic protein interactions reveal a functional map of the cell. *eLife*, 5:e13053–17, Apr. 2016.

B. Efron. Large-Scale Simultaneous Hypothesis Testing. *Journal of the American Statistical Association*, 99(465):96–104, Mar. 2004.

L. Scrucca, M. Fop, T. B. Murphy, and A. E. Raftery. mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *The R journal*, 8(1):289–317, Aug. 2016.