Supplementary Materials and Methods

Experimental data protocols

The protocol details for both technologies can be found in Manakov *et al.*, (2022) and Van Nostrand *et al.*, (2016), respectively. Briefly, miR-eCLIP is an extension of the standard eCLIP protocol using an anti-AGO2 antibody and modified to enable chimeric ligation of miRNA and target mRNA. RBFOX2 RBP-eCLIP (RBFOX2-eCLIP), on the other hand, was obtained by applying RBP-eCLIP against RBFOX2, an RNA-binding protein with a widespread role in several cellular mechanisms and tissue-specific effects (Arya *et al.*, 2014). RBP-eCLIP extends the single-end eCLIP protocol (Van Nostrand *et al.*, 2016) by optimizing several aspects of the original technique, like cell input requirement, UV crosslinking settings, and the lysis protocol.

The resulting sequencing data went through multiple steps of preprocessing and then clusters of aligned reads were found using CLIPper (v2.0.1). For the miR-eCLIP data sets, each cluster was annotated with the names of miRNAs responsible for that target. For the RBFOX2 eCLIP, the IP levels were normalized with input levels to calculate fold enrichments, and p-values were calculated using a Yates' Chi-Square test (or Fisher Exact Test if the observed or expected read number was below 5). In all datasets, peaks were annotated using transcript information from GENCODE release 41 (GRCh38.p13) (Frankish *et al.*, 2022) with the following priority hierarchy to define the final annotation of overlapping features: protein-coding transcript (CDS, UTRs, intron), followed by non-coding transcripts (e.g., exon, intron).

Reproducibility assessment methods: models and algorithms

Estimation of mixture copula model parameters

Suppose $\mathbf{X} = (x_{ij})_{i=1,\dots,n}^{j=1,\dots,r}$ is a matrix with positive values standing for significance or intensity of n omic features in r technical replicates. As in standard IDR, gIDR and mIDR, used a copula mixture model on $\mathbf{X}$. In the case of mIDR, the model fits in every pairwise combination of replicates. The derivation of the copula model and parameters estimation is as follows. First, feature values are transformed according to the probability integral transform (Savits, 1994) and re-scaled to avoid infinities:

$$u_{ij} \equiv \frac{n}{n+1}\hat{F}_j(x_{ij}). \quad (1)$$

Where $\hat{F}_j$ is the empirical cumulative probability distribution (ECDF) of omic feature values across replicate j.

Second, the vector of model parameter values is initialized $\boldsymbol{\theta} = \boldsymbol{\theta}^{(0)} = \left(\pi_1^{(0)},\ \mu_1^{(0)},\ \boldsymbol{\Sigma}_1^{(0)}\right)$. Where $\boldsymbol{\Sigma}_1 = \sigma_1^2(1-\rho_1)\left[\mathbf{I} + \frac{\rho_1}{1-\rho_1}\mathbf{J}\right]$, σ_1^2 is the reproducible class variance, ρ_1 is the reproducible class correlation, $\mathbf{I}$ is a rxr matrix with ones in the diagonal and zero otherwise and $\mathbf{J}$ is rxr matrix with all ones. In the case of the irreproducible class, $\boldsymbol{\Sigma}_0 = \mathbf{I}$.

Third, $\mathbf{X}$ is turned into "pseudo data" $\mathbf{Z} = (z_{ij})_{i=1,\dots,n}^{j=1,\dots,r}$, where $z_{ij} = G^{-1}(u_{ij}|\boldsymbol{\theta})$, with

$$u_{ij} \equiv G(z_{ij}) = \frac{\pi_1}{\sigma_1}\Phi\left(\frac{z_{ij}-\mu_1}{\sigma_1}\right) + \Phi(z_{ij}) \quad (2)$$

and Φ is the univariate standard normal cumulative density function. The value of G^{-1} is obtained by calculating G on regular grid of 1,000 u_{ij} values followed by linear interpolation as in Li et al. (2011).

Then, an EM algorithm on the augmented pseudo data $\boldsymbol{y}_i = (K_i, \boldsymbol{z}_i)$ is used to obtain $\boldsymbol{\theta}^{(t)}$ (i.e., the estimates of parameter values at iteration t) where $\boldsymbol{z}_i = [z_{i1}, \dots, z_{ir}]$. After setting $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$**,** the last two steps are repeated until convergence or a maximum value to t (t_{MAX}) are reached.

gIDR

The principal difference between IDR and gIDR is that the latter incorporates information from all replicates by using the following complete log-pseudolikelihood formula:

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \begin{bmatrix} (1-K_i)(\log(1-\pi) + \log(\phi_0(\boldsymbol{z}_i|\mu_0, \boldsymbol{\Sigma}_0))) + \\ K_i(\log(\pi) + \log(\phi_1(\boldsymbol{z}_i|\mu_1, \boldsymbol{\Sigma}_1))) \end{bmatrix}. \quad (3)$$

where ϕ_k is the probability density function of a "General" Multivariate Normal distribution, with dimension equal to the number of replicates r (see equation (2))

Since the model is in essence a mixture of two normal distributions on the pseudo data, standard EM formulas apply (Xu *et al.*, 2016). Hence, for the Expectation step (*E-step*) we have

$$K_i^{(t+1)} = E\left(K_i|z_{i1}, \dots, z_{ir}, \boldsymbol{\theta}^{(t)}\right) = \frac{\pi_1^{(t)}\phi_1\left(z_{i1},\dots,z_{ir}|\mu_1^{(t)},\boldsymbol{\Sigma}_1{}^{(t)}\right)}{\sum_{k=0}^{1}\pi_k^{(t)}\phi_k\left(z_{i1},\dots,z_{ir}|\mu_k^{(t)},\boldsymbol{\Sigma}_k{}^{(t)}\right)}, \quad (7)$$

For the Maximization step (*M-step*), we have:

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \pi_1} = \sum_{i=1}^{n} \frac{(1 - K_i)}{1 - \pi_1} - \frac{K_i}{\pi_1} = 0, \tag{8}$$

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \mu_1} = \sum_{i=1}^{n} K_i \, \boldsymbol{z}_i^T \boldsymbol{\Sigma}_1^{-1} - \mu_1 \mathbf{1}^T \boldsymbol{\Sigma}_1^{-1} \sum_{i=1}^{n} K_i = 0, \tag{9}$$

and

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\Sigma}_1^{-1}} = \boldsymbol{\Sigma}_1 \sum_{i=1}^{n} K_i - \sum_{i=1}^{n} K_i \, (\boldsymbol{z}_i - \mu_k \mathbf{1})(\boldsymbol{z}_i - \mu_k \mathbf{1})^T = 0 \tag{4}$$

Where in equation (10) we used a) the derivatives of the logarithm of a matrix's determinant with respect to $\boldsymbol{\Sigma}_1^{-1}$ and b) the derivative of the trace of the product between a matrix and $\boldsymbol{\Sigma}_1^{-1}$ with respect to $\boldsymbol{\Sigma}_1^{-1}$ (Dwyer, 1967).

The maximum likelihood estimates of each parameter in the $(t+1)$-th iteration are given by the following expressions:

$$\pi_1^{(t+1)} = \sum_{i=1}^{n} \frac{K_i^{(t+1)}}{n}, \tag{11}$$

$$\mu_1^{(t+1)} = \frac{\sum_{i=1}^{n} K_i^{(t+1)} \sum_{j=1}^{r} z_{ij}}{r \sum_{i=1}^{n} K_i^{(t+1)}} \tag{12}$$

and

$$\boldsymbol{\Sigma}_1^{(t+1)} = \frac{\sum_{i=1}^{n} K_i^{(t+1)} \left(\boldsymbol{z}_i - \mu_1^{(t+1)}\mathbf{1}\right)\left(\boldsymbol{z}_i - \mu_1^{(t+1)}\mathbf{1}\right)^T}{\sum_{i=1}^{n} K_i^{(t+1)}}. \quad (13)$$

Finally, since $\boldsymbol{\Sigma}_1 = \sigma_1^2(1-\rho_1)\left[\mathbf{I} + \frac{\rho_1}{1-\rho_1}\mathbf{J}\right]$, we have:

$$(\sigma_1^2)^{(t+1)} = \frac{\mathrm{tr}\left(\boldsymbol{\Sigma}_1^{(t+1)}\right)}{r} = \frac{\sum_{i=1}^{n} K_i^{(t+1)} \sum_{j=1}^{r}\left(z_{ij} - \mu_1^{(t+1)}\right)^2}{r\sum_{i=1}^{n} K_i^{(t+1)}} \quad (5)$$

and

$$\rho_1^{(t+1)} = \frac{\mathrm{trU}\left(\boldsymbol{\Sigma}_1^{(t+1)}\right)}{(\sigma_1^2)^{(t+1)}\binom{r}{2}} = \frac{2\sum_{i=1}^{n} K_i^{(t+1)} \sum_{j\neq l}\left(z_{ij} - \mu_1^{(t+1)}\right)\left(z_{il} - \mu_1^{(t+1)}\right)}{(\sigma_1^2)^{(t+1)} r(r-1)\sum_{i=1}^{n} K_i^{(t+1)}}, \quad (6)$$

where tr and trU are the trace and upper-triangular trace of a matrix, respectively, and $\binom{r}{2}$ is the number r replicates combinations in two. Finally, the local IDR is approximated as $\Pr(K_i = 0|\boldsymbol{\theta}, \mathbf{X}) \approx 1 - K_i^{(t_{MAX})}$, where t_{MAX} is the number of iterations at convergence. Notice that when $r = 2$, the *E-step* and *M-step* formulas above reduce to ones derived for standard IDR (see Li *et al*, 2011 and supplementary materials there).

mIDR

Method mIDR is derived using a simple heuristics method that resembles a meta-analysis score (Shelby and Vaske, 2008). First, we perform standard IDR in each one of every pairwise combination of replicates. Then, a meta score is calculated to get the probability of a feature being irreproducible in at least one of the $\binom{r}{2}$ comparisons:

$$\Pr(K_i = 0|\boldsymbol{\theta}, \mathbf{X}) \approx 1 - \prod\nolimits_{j \neq l} \Pr\left(K_i = 1 \middle| \boldsymbol{\theta}, x_{ij}, x_{il}\right) \quad (7)$$

eCV

The eCV method assumes omic feature values across replicates $\boldsymbol{y}_i = [y_{i1}, y_{i2}, \ldots, y_{ir}]$ are positive and approximately Normal. If x_{ij} stands for the number of aligned sequencings reads or another measurement of the feature's intensity, with distribution approximately log-Normal, taking $y_{ij} = \log(x_{ij} + \in)$, where $\in$ is an added positive value that prevents infinite values, yields the same distribution as in (1).

The estimated eCV is calculated as:

$$\widehat{eCV}_i = \frac{\left|\hat{\sigma}_i^2 - \hat{\mu}_i^2\right|}{\hat{\mu}_i}, \quad (8)$$

where $\hat{\sigma}_i^2 = \frac{1}{(r-1)} \sum_{j=1}^{r} \left(y_{ij} - \hat{\mu}_i\right)^2$ and $\hat{\mu}_i = \frac{1}{r} \sum_{j=1}^{r} y_{ij}$. Smaller values of $\widehat{eCV}_i$ would show reproducibility and vice versa.

To make inferences on the reproducibility of feature i, we first compute the following expression

$$\Pr\left(eCV_i \leq \widehat{eCV}_i | \boldsymbol{y}_i, \boldsymbol{\theta}, K_i = 1\right) \cong \hat{F}(\widehat{eCV}_i | \boldsymbol{y}_i, \boldsymbol{\theta}, K_i = 1) \quad (9)$$

With $\hat{F}$ being the ECDF of eCV_i. Equation (18) is analogous to a p-value. The larger the value of $\widehat{eCV}_i$, the larger the value of (18).

Second, we apply the Bayes theorem and the chain rule of probability to connect (18) with $\Pr(K_i|\boldsymbol{\theta}, \mathbf{X})$:

$$\Pr(K_i = 1|\boldsymbol{\theta}, \mathbf{X}) \propto$$

$$F\left(\widehat{eCV}_i \middle| \boldsymbol{y}_i, \boldsymbol{\theta}, K_i = 1\right) \Pr(\boldsymbol{y}_i|\boldsymbol{\theta}, K_i = 1) \Pr(\boldsymbol{\theta}|K_i = 1) \Pr(K_i = 1) \quad (19)$$

In a fully Bayesian treatment of (19), prior distributions for $\boldsymbol{\theta}$,K_i and their hyperparameters (e.g., π) might be used. For eCV, however, we assume a somewhat radical "Empirical Bayes" (EB) (Morris, 1983) method, where we set $\left(\widehat{\boldsymbol{\theta}}\middle|K_i = 1\right) \approx \left(\hat{\pi} = 0.5, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}\right)$. The quantities $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\Sigma}}$ are, respectively, the vector of means and the variance-covariance matrix across features. The values of $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\Sigma}}$ are assumed as good proxies for the parameters of prior distribution of reproducible features.

Third, EB estimates are plugged into equation (1) to randomly sample 10,000 values of $\boldsymbol{y}_i$ ($\tilde{\boldsymbol{y}}_i$). eCV is calculated in every $\tilde{\boldsymbol{y}}_i$ to obtain $\widetilde{eCV}_i$ values. The ECDF is used to obtain $\hat{F}$. Lastly, $\widehat{eCV}_i$ is plugged in $\hat{F}$ to obtain an estimate of (19).

References

Arya,A.D. *et al.* (2014) RBFOX2 protein domains and cellular activities. *Biochem. Soc. Trans.*, **42**, 1180–1183.

Dwyer,P.S. (1967) Some Applications of Matrix Derivatives in Multivariate Analysis. *J. Am. Stat. Assoc.*, **62**, 607–625.

Frankish,A. *et al.* (2022) GENCODE: reference annotation for the human and mouse genomes in 2023. *Nucleic Acids Res.*, **51**, D942–D949.

Manakov,S.A. *et al.* (2022) Scalable and deep profiling of mRNA targets for individual microRNAs with chimeric eCLIP. 2022.02.13.480296.

Morris,C.N. (1983) Parametric Empirical Bayes Inference: Theory and Applications. *J. Am. Stat. Assoc.*

Savits,T.H. (1994) On integration, substitution and the probability integral transform. *Stat. Probab. Lett.*, **21**, 173–179.

Shelby,L.B. and Vaske,J.J. (2008) Understanding Meta-Analysis: A Review of the Methodological Literature. *Leis. Sci.*, **30**, 96–110.

Van Nostrand,E.L. *et al.* (2016) Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat. Methods*, **13**, 508–514.

Xu,J. *et al.* (2016) Global Analysis of Expectation Maximization for Mixtures of Two Gaussians. In, *Advances in Neural Information Processing Systems*. Curran Associates, Inc.