Supplementary Materials

1 Supplementary Methods

1.1 CLIP-seq Data preparation

HITS-CLIP experiments were performed as follows. *C. elegans* transgenic strains carrying a single copy of a modified lin-28 gene, encoding a fusion GFP, flag, HAHA at the C-terminus, were generated by bombardment. The expression of the transgene at the proper time and place was verified by RT-PCR, western blot and by its ability to fully rescue the phenotype of the *lin-28* n719 mutant strain. Liquid cultures of L1 larvae (five millions) were harvested by centrifugation and treated with UV in a stratalinker (3.6 mJ/cm^2). Subsequently, worms were lysated with zirconia beads in a MP Fastprep 24 in buffer A (20 mM Hepes pH 7.4, 0.1% SDS, 0.5% deoxycholate, 0.5% NP40, 20 mM EDTA and 20 mM EGTA), and lysate was cleared by ultracentrifugation (100,000 X g, 30 minutes). Subsequent steps were performed as described previously (Jensen and Darnell, 2008; Ule, et al., 2005). LIN-28/RNA complexes were purified with a commercial antibody anti-HA (HA-7, Sigma H3663) conjugated with Dynabeads (Life Technologies 112-01D). During the subsequent washing steps, the complexes were treated with micrococcal nuclease to achieve an average RNA size of about seventy nuceotides, as estimated by gel electrophoresis and by Alkaline Phosphatase. A 5' end adapter was ligated overnight. Following SDS-PAGE purification and proteinase K treatment, a 3' end adapter was ligated, and Reverse Transcription/PCR was performed. Libraries thus prepared were sequenced in an Illumina HighSeq 2000 machine. RNA-seq libraries were performed from total RNA purified from L1 larvae reared the same way, following oligo(dT) selection, according to standard Illumina protocol.

1.2 Background estimation

We used the Poisson, Negative Binomial (NB), Beta-Binomial, Poisson regression, and NB regression models to estimate background distribution of CLIP-seq with or without a RNA-seq control. Parameters of all the models were estimated using the maximum likelihood estimation (MLE) unless otherwise stated. Let k be the number of windows and n be the total number of reads in regions of interest.

1.2.1 Poisson model.

For the Poisson model, we have the probability mass function as

$$P(x \mid \lambda_{bg}) = \frac{\lambda_{bg}^{x}}{x!} e^{-\lambda_{bg}}.$$

Hence, we have the log likelihood function as

$$L = \sum_{i=1}^{k} [x_i \log(\lambda_{bg}) - \lambda_{bg} - \log(x_i!)]$$

and the parameter λ_{bg} is estimated to be

$$\hat{\lambda}_{bg} = \frac{\sum x_i}{k}$$

1.2.2 Gamma Poisson/ negative binomial model.

We considered the Gamma-Poisson model:

$$\begin{aligned} &\lambda \sim Gamma(a,b), \\ &X \mid \lambda \sim Poisson(\lambda). \end{aligned}$$

Then, the marginal distribution is obtained as

$$\begin{aligned} p(x|a,b) &= \int p(x \mid \lambda) p(\lambda \mid a,b) d\lambda \\ &= \int Poisson(x \mid \lambda) Gamma(\lambda \mid a,b) d\lambda \\ &= \int \frac{\lambda^x e^{-\lambda}}{x!} \frac{1}{\Gamma(a) b^a} \lambda^{a-1} e^{-\lambda/b} d\lambda \\ &= \frac{1}{x!\,\Gamma(a) b^a} \int \lambda^{x+a-1} e^{-(b^{-1}+1)\lambda} d\lambda \\ &= \frac{\Gamma(x+a)}{x!\,\Gamma(a) b^a (b^{-1}+1)^{x+a}} \\ &= \frac{\Gamma(x+a)}{x!\,\Gamma(a)} \left(\frac{b}{b+1}\right)^x \left(\frac{1}{b+1}\right)^a . \end{aligned}$$

Hence, the marginal distribution of the Gamma-Poisson model is equivalent to Negative Binomial distribution. Alternatively, if we let $\mu = ab, \alpha = 1/a$, we have

$$p(x \mid \mu, \alpha) = \frac{\Gamma(x+\alpha^{-1})}{x!\,\Gamma(\alpha^{-1})} \left(\frac{\mu}{\mu+\alpha^{-1}}\right)^x \left(\frac{\alpha^{-1}}{\mu+\alpha^{-1}}\right)^{1/\alpha},$$

denoted as

$$X \sim NegBin(\mu, \alpha).$$

In order to obtain MLE for parameters, we first let $b = \frac{p}{1-p}$. Then, we have the log likelihood function as

$$l(a,p) = \sum_i \log\Gamma(x_i + a) - \sum_i \log(x_i!) - k\log\Gamma(a) + ka\log(1-p) + \sum_i x_i \log(p)$$

By differentiating the log likelihood function with respect to p and a, we have

$$\begin{cases} \frac{\partial l(a,p)}{\partial p} = -\frac{ka}{1-p} + \sum_i x_i \frac{1}{p} = 0 \\ \frac{\partial l(a,p)}{\partial a} = \sum_i \psi(x_i + a) - k\psi(a) + k\log(1-p) = 0 \end{cases} \qquad (1)$$

where digamma function $\psi(x) = \Gamma'(x)/\Gamma(x)$. From the equation (1), we have

$$p = \frac{\sum_i x_i}{\sum_i x_i + ka} \qquad (2)$$

We substitute p in the log likelihood function $l(a,p)$ by the equation (2) and obtain

$$\begin{aligned} l(a,p) = &\sum_i \log(\Gamma(x_i + a)) - \sum_i \log(x_i!) - k\log(\Gamma(a)) \\ &+ ka\log(\frac{ka}{\sum_i x_i + ka}) + \sum_i x_i \log(\frac{\sum_i x_i}{\sum_i x_i + ka}) \end{aligned} \qquad (3)$$

Finally, we obtain MLE for a as a maximizer of equation (3) using Newton's method and then obtain MLE for p using equation (2).

1.2.3 Beta-binomial model.

We considered the Beta-Binomial model:

$$\begin{aligned} &\theta \sim Beta(a,b), \\ &X \mid \theta \sim binomial(n,\theta). \end{aligned}$$

Then, the marginal distribution is obtained as

$$
\begin{aligned}
p(x|a,b) &= \int p(x \mid \theta)p(\theta \mid a,b)d\theta \\
&= \int \binom{n}{x} \theta^{x}(1-\theta)^{n-x} \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)} d\theta \\
&= \binom{n}{x} \frac{1}{B(a,b)} \int \theta^{x+a-1}(1-\theta)^{n-x+b-1} d\theta \\
&= \binom{n}{x} \frac{B(x+a,n-x+b)}{B(a,b)},
\end{aligned}
$$

denoted as

$$X \sim Bb(a,b,n).$$

Because we do not have explicit solutions for MLE of the Beta-Binomial model, we use numerical optimization routine to estimate the parameters. In order to obtain more reliable numerical solutions, we used the method of moments (MOM) to estimate initial values of parameters described as follows. For the Beta-Binomial model, we have the first and second moments as

$$
\begin{aligned}
E(X) &= \frac{na}{a+b} \\
E(X^2) &= \frac{na(na+n+b)}{(a+b)(a+b+1)}
\end{aligned}
$$

By solving the equations above in terms of a and b and substituting $E(X)$ and $E(X^2)$ by their corresponding sample moments m_1 and m_2, we have

$$
\begin{aligned}
\hat{a} &= \frac{m_1 m_2 - n m_1^2}{(n-1)m_1^2 + n m_1 - n m_2}, \\
\hat{b} &= \frac{n m_2 - n^2 m_1 - m_1 m_2 + n m_1^2}{(n-1)m_1^2 + n m_1 - n m_2}
\end{aligned}
$$

where

$$m_1 = \frac{\sum_i x_i}{k}, m_2 = \frac{\sum_i x_i^2}{k},$$

Then, we use these $\hat{a}$ and $\hat{b}$ obtained from MOM as initial values for MLE.

1.2.4 Poisson regression model.

Suppose that we observe RNA-seq tag counts in i-th window, r_i, which corresponds to CLIP-seq tag counts in i-th window, x_i. In Poisson regression model, we assume that logarithm-transformed expectation of x_i is a linear function of log(r_i):

$$\log E(X_i \mid R_i = r_i) = \log \mu(r_i) = a + b\log(r_i)$$

Then, we have marginal distribution for CLIP-seq count in i-th window as

$$p(x_i \mid r_i, a, b) = \frac{\mu(r_i)^{x_i} e^{-\mu(r_i)}}{x_i!}.$$

We have the log likelihood function as

$$l(a,b \mid X,R) = \sum_i \{x_i(a + b\log(r_i)) - e^{a+b\log(r_i)} - \log(x_i!)\}$$

Because we do not have explicit solutions for this log likelihood function, we used numerical optimization routine to estimate the parameters.

1.2.5 Negative binomial regression model.

As in the case of Poisson regression model, we suppose that we observe RNA-seq tag counts in i-th window, r_i, which corresponds to CLIP-seq tag counts in i-th window, x_i. Again, we assume that logarithm-transformed expectation of x_i is a linear function of $\log(r_i)$:

$$\log E(X_i \mid R_i = r_i) = \log \mu(r_i) = a + b\log(r_i)$$

Then, we have marginal distribution for CLIP-seq count in i-th window as

$$p(x_i \mid r_i, a, b, \alpha) = \frac{\Gamma(x_i + \alpha^{-1})}{x_i!\Gamma(\alpha^{-1})} \left(\frac{\mu(r_i)}{\mu(r_i) + \alpha^{-1}} \right)^{x_i} \left(\frac{\alpha^{-1}}{\mu(r_i) + \alpha^{-1}}\right)^{1/\alpha}.$$

We have the log likelihood function as

$$l(a,b,\alpha \mid X,R) = \sum_i \left[\log\Gamma(x_i+\alpha^{-1}) - \log(x_i!) - \log\Gamma(\alpha^{-1}) + x_i \log\frac{e^{a+b\log(r_i)}}{e^{a+b\log(r_i)}+\alpha^{-1}} - \frac{1}{\alpha}\log(\alpha e^{a+b\log(r_i)}+1) \right]$$

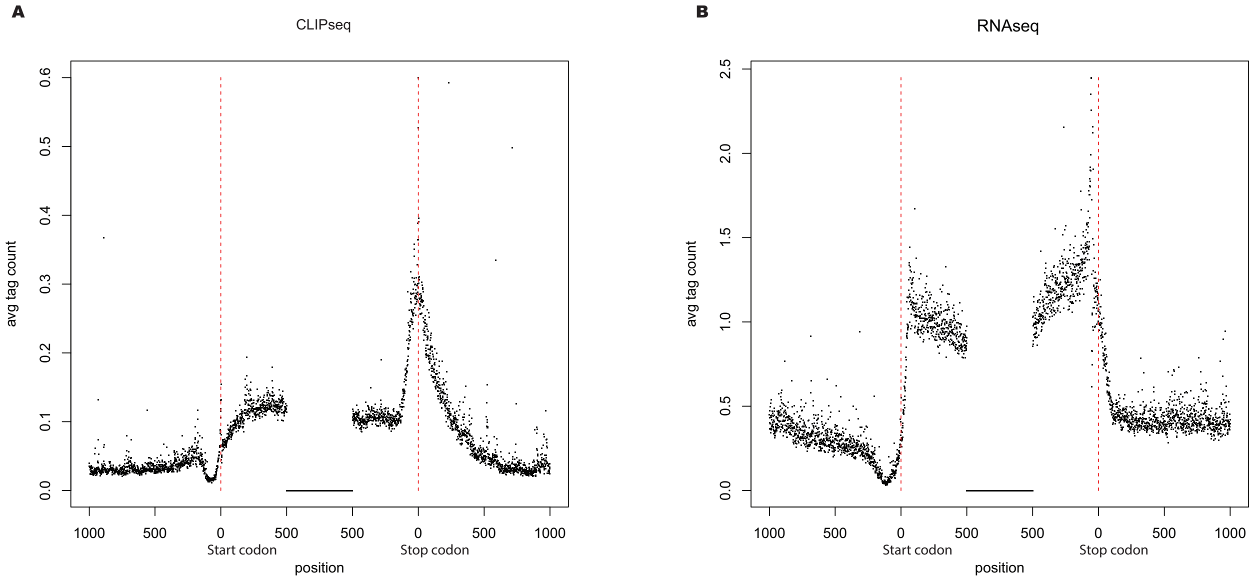
Because we do not have explicit solutions for this log likelihood function, we used numerical optimization routine to estimate the parameters. The regression coefficients were estimated by MLE and dispersion parameter was estimated through the residual degrees of freedom.

References

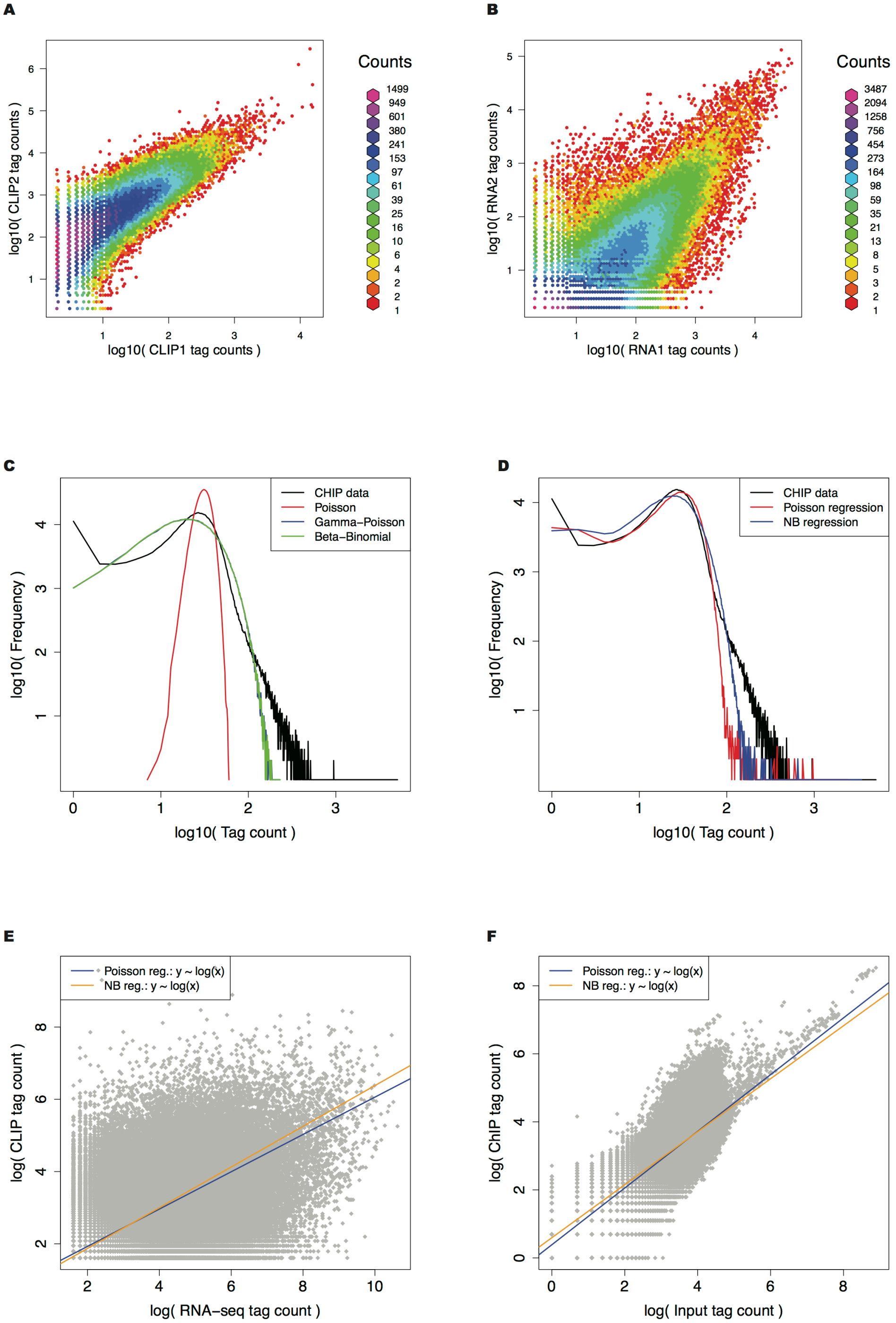
Jensen, K.B. and Darnell, R.B. (2008) CLIP: crosslinking and immunoprecipitation of in vivo RNA targets of RNA-binding proteins, *Methods in molecular biology*, **488**, 85-98.

Ule, J., *et al.* (2005) CLIP: a method for identifying protein-RNA interaction sites in living cells, *Methods*, **37**, 376-386.

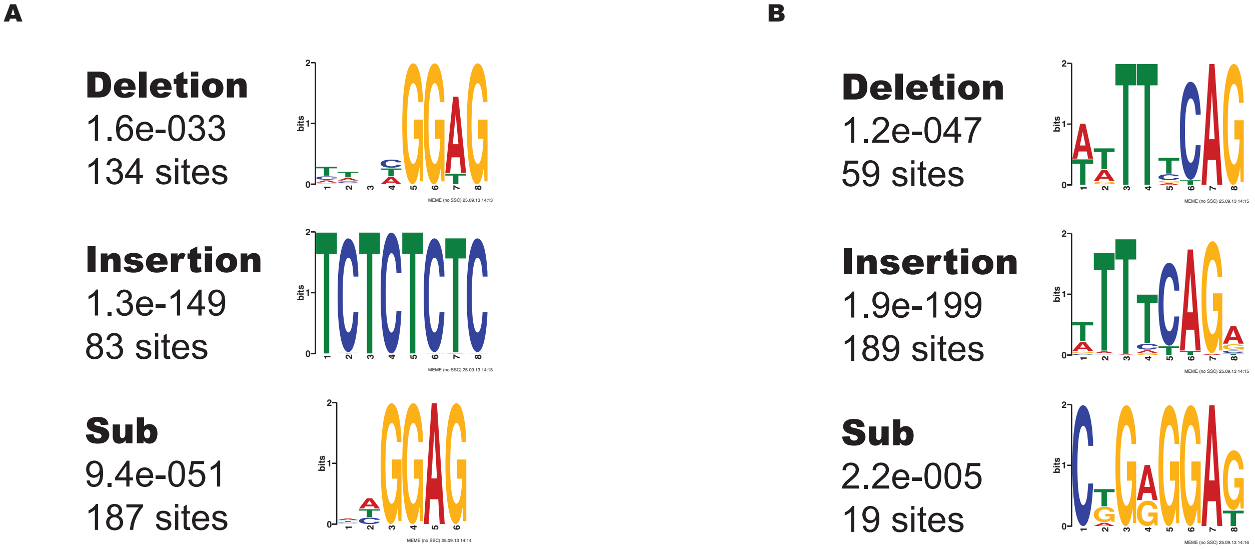
2 Supplementary Figures and Tables



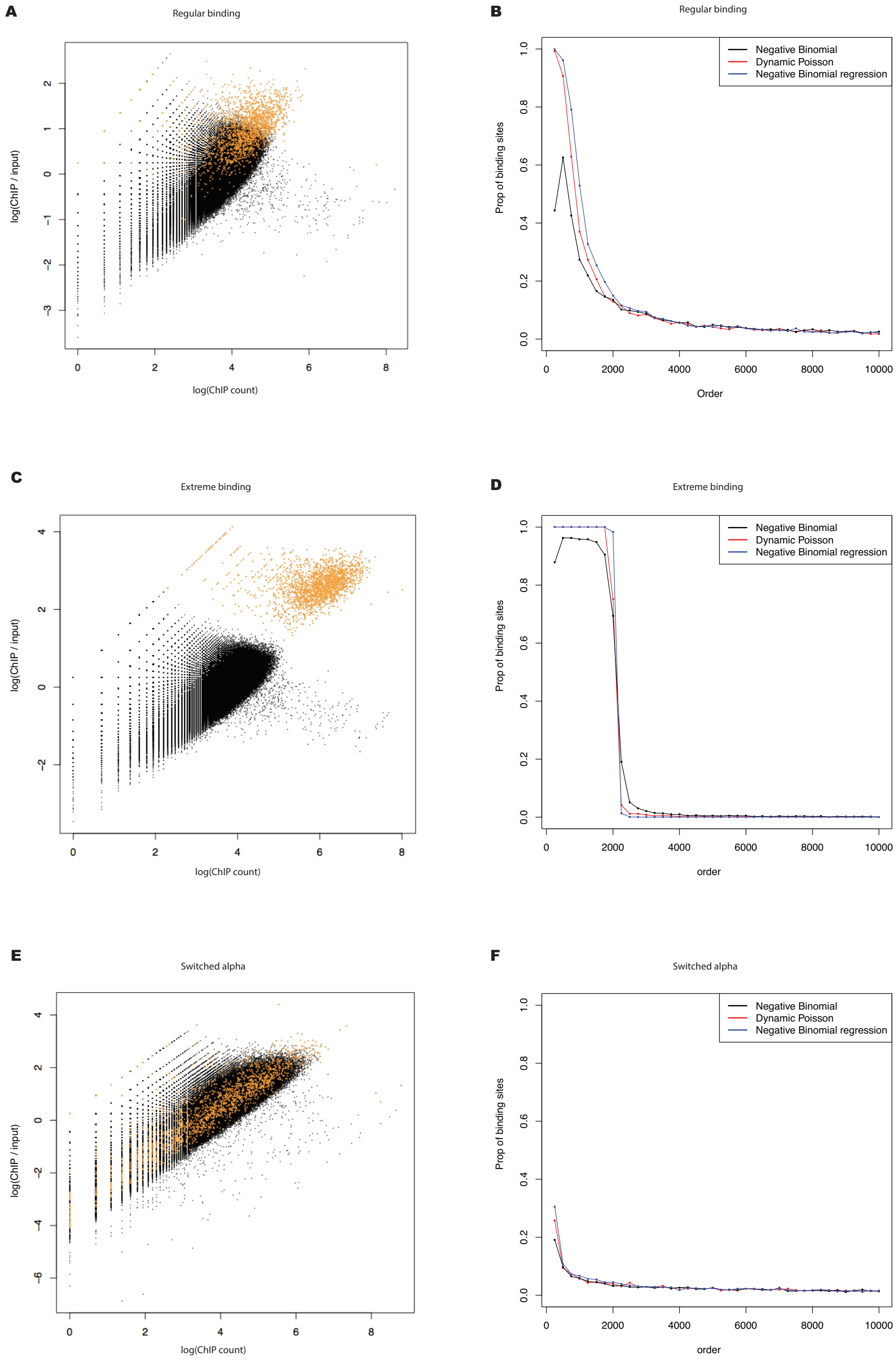
Suppl Fig 1. Read mapping around genes in CLIP-seq (A) and RNA-seq (B). The X axis is the position between 1000bp upstream of start codons and 1000bp downstream of stop codons. The Y axis is the average mapped tag count on each position. Both figures show that not many reads are mapped to 5'end of genes and reads are uniformly distributed over exon regions except for the regions close to stop codon. CLIP-seq shows longer mapping on 3'UTRs of genes. Considering that majority of 5'UTRs and 3'UTRs have lengths less than 200bp and 450bp, respectively, in Wormbase, we extended 200bp for 5'end and 750bp for 3'end to include most of potential UTRs in our study.



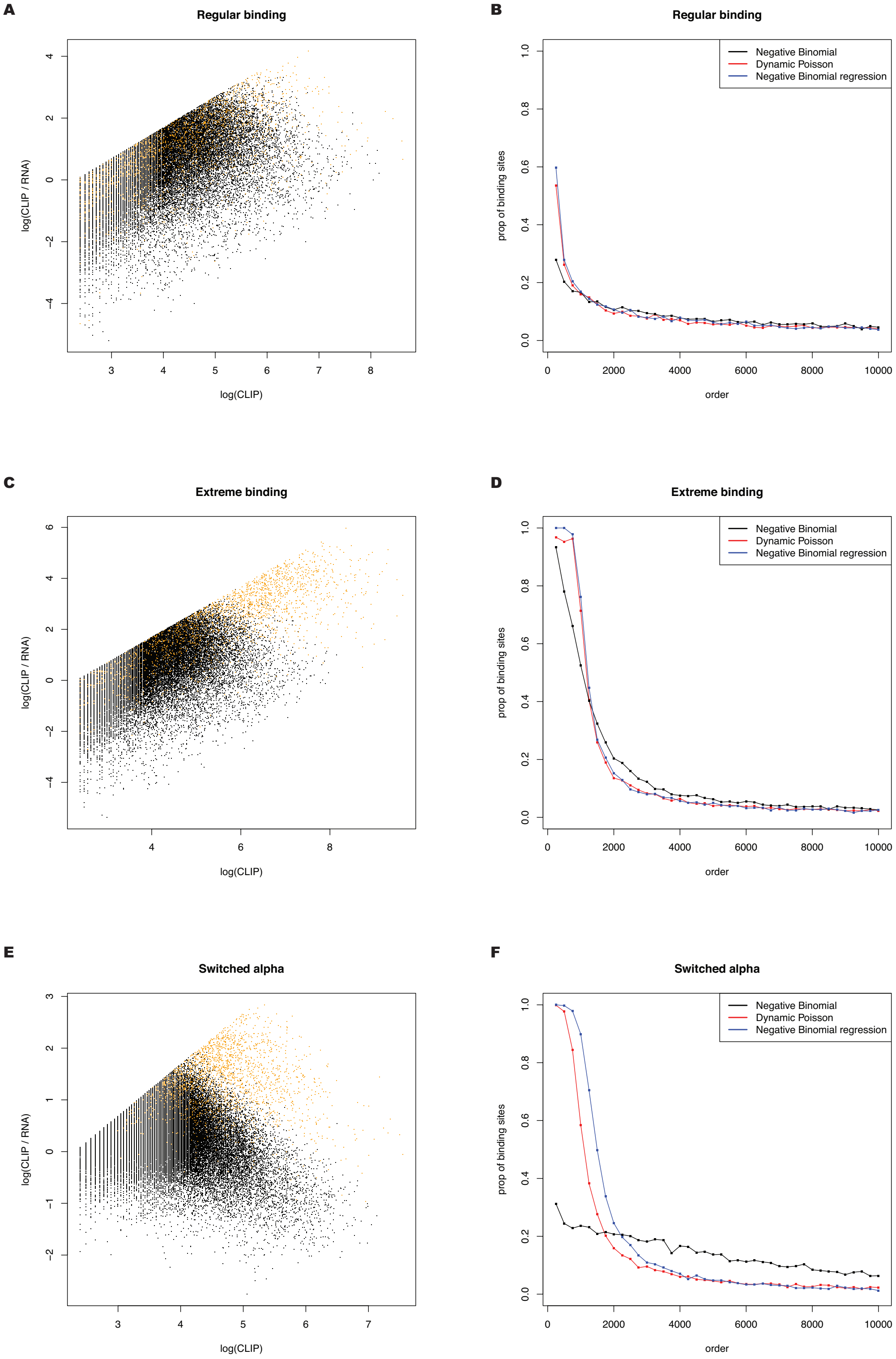
Suppl Fig 2. Background estimation and model fitting. (A) Scatter plot of two CLIP-seq replicates. (B) Scatter plot of two RNA-seq replicates. (C) One-sample model fitting on ChIP-seq. (D) Two-sample model fitting on ChIP-seq with input control. (E) Scatter plot of CLIP-seq VS RNA-seq counts (for windows with count >3), with linear regression fits of CLIP-seq on RNA-seq counts. (F) Scatter plot of ChIP-seq VS input counts with linear regression fits on ChIP-seq on input counts.



Suppl Fig 3. Motif identified for top 500 mutations in CLIP-seq2 (A) and RNA-seq (B) by MEME.



Suppl Fig 4. ChIP-seq simulation. (A, C, E) True binding sites (yellow ones) on scatter plot of log(ChIP-seq count / input count) VS log(ChIP-seq count). (B, D, F) Proportion of true binding windows among the windows identified to be binding sites by each method, where windows are ordered by their p-values.



Suppl Fig 5. CLIP-seq simulation. (A, C, E) True binding sites (yellow ones) on scatter plot of log(CLIP-seq count / RNA-seq count) VS log(CLIP-seq count). (B, D, F) Proportion of true binding windows among the windows identified to be binding sites by each method, where windows are ordered by their p-values.

Table S1. Bayesian information criterion (BIC) values for models in background estimation. Lower values indicate better model fits.

Models	CLIP-seq	ChIP-seq
Poisson	17,526,877	9,081,887
Negative binomial	2,289,642	4,279,481
Beta-binomial	2,289,658	4,279,489
Poisson regression	11,390,532	5,424,135
Negative binomial regression	2,040,052	3,772,524