in vitro iCLIP-based modeling uncovers how the splicing factor U2AF2 relies on regulation by co-factors

SUPPLEMENTAL MATERIAL

FX Reymond Sutandy*[1], Stefanie Ebersberger*[1], Lu Huang*[1], Anke Busch[1], Maximilian Bach[1], Hyun Seo Kang[2,3], Jörg Fallmann[4], Daniel Maticzka[5], Rolf Backofen[5,6], Peter F. Stadler[4], Kathi Zarnack[7], Michael Sattler[2,3], Stefan Legewie[#1], Julian König[#1]

[1]Institute of Molecular Biology (IMB) gGmbH, Ackermannweg 4, 55128 Mainz. [2]Institute of Structural Biology, Helmholtz Center Munich, Ingolstädter Landstr. 1, 85764 Neuherberg, Germany; [3]Biomolecular NMR and Center for Integrated Protein Science Munich at Department of Chemistry, Technical University of Munich, Lichtenbergstr. 4, 85747 Garching, Germany. [4]Bioinformatics Group, Department of Computer Science and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18,04107 Leipzig. [5]Bioinformatics Group, Department of Computer Science, University of Freiburg, Freiburg, Germany. [6]Centre for Biological Signalling Studies (BIOSS), University of Freiburg, Freiburg, Germany. [7]Buchmann Institute for Molecular Life Sciences (BMLS), Goethe University Frankfurt, Max-von-Laue-Str. 15, 60438 Frankfurt a.M., Germany

* shared first authors; [#] correspondence: s.legewie@imb-mainz.de, j.koenig@imb-mainz.de

Content:

Description of mathematical model

We developed a mathematical model of U2AF2 binding to all binding sites in the *in vitro* transcripts. In the following, it will be described how this model was derived, calibrated by fitting to *in vitro* iCLIP titration data with eight different concentrations of recombinant U2AF2^{RRM12} and then applied to better understand the *in vivo* landscapes of U2AF2 binding.

I. Modeling *in vitro* U2AF2 binding to its binding sites

1. Derivation of the binding model

The binding sites of U2AF2 were defined as peaks of U2AF2^{RRM12} iCLIP signal above background in a 9-nt window (see Supplemental Methods), corresponding to the commonly observed width of U2AF2 binding sites (Agrawal et al. 2016). The binding of U2AF2 to these sites was modeled using the following reversible and monomeric binding model:

$$
\begin{aligned}
U2AF2 + Site_1 &\underset{k_{off1}}{\overset{k_{on1}}{\longleftrightarrow}} [U2AF2{:}\,Site_1] \\
U2AF2 + Site_2 &\underset{k_{off2}}{\overset{k_{on2}}{\longleftrightarrow}} [U2AF2{:}\,Site_2] \\
&\cdots\ \cdots \\
U2AF2 + Site_i &\underset{k_{offi}}{\overset{k_{oni}}{\longleftrightarrow}} [U2AF2{:}\,Site_i] \\
&\cdots\ \cdots \\
U2AF2 + Site_n &\underset{k_{offn}}{\overset{k_{on\,n}}{\longleftrightarrow}} [U2AF2{:}\,Site_n]
\end{aligned}
$$

The temporal changes of the system's components were described by sets of ordinary differential equations (ODEs) and algebraic equations describing conservation relations.

The temporal changes of U2AF2-RNA complexes are described by ODEs:

$$
\begin{aligned}
\frac{d([U2AF2{:}\,Site_1])}{dt} &= k_{on1} \cdot [U2AF2] \cdot [Site_1] - k_{off1} \cdot [U2AF2{:}\,Site_1] \\
\frac{d([U2AF2{:}\,Site_2])}{dt} &= k_{on2} \cdot [U2AF2] \cdot [Site_2] - k_{off2} \cdot [U2AF2{:}\,Site_2] \\
&\cdots\ \cdots \\
\frac{d([U2AF2{:}\,Site_i])}{dt} &= k_{oni} \cdot [U2AF2] \cdot [Site_i] - k_{offi} \cdot [U2AF2{:}\,Site_i] \\
&\cdots\ \cdots \\
\frac{d([U2AF2{:}\,Site_n])}{dt} &= k_{onn} \cdot [U2AF2] \cdot [Site_n] - k_{offn} \cdot [U2AF2{:}\,Site_n]
\end{aligned}
\qquad (1)
$$

Each binding site can either be free or occupied by U2AF2, giving rise to the following conservation relations:

$$[Site_1]_{total} = [Site_1] + [U2AF2{:}\,Site_1\,]$$

… …

$$[Site_n]_{total} = [Site_n] + [U2AF2{:}\,Site_n\,]$$

(2)

U2AF2 is distributed among all binding sites and thus follows a conservation relation given by

$$[U2AF2]_{total} = [U2AF2] + [U2AF2{:}\,Site_1\,] + [U2AF2{:}\,Site_2\,] + \cdots + [U2AF2{:}\,Site_i\,] + \cdots + [U2AF2{:}\,Site_n\,]$$

(3)

Our modeling approach builds on data of *in vitro* iCLIP experiments. In these experiments, *in vitro* transcribed RNAs were incubated with recombinant U2AF2^{RRM12} for 10 min, suggesting that the U2AF2^{RRM12}-RNA complexes reached equilibrium. We therefore neglected the temporal changes of the complexes using an equilibrium assumption for each binding site *i*:

$$\frac{d([U2AF2{:}\,Site_i])}{dt} = 0$$

(4)

This yields the equilibrium concentration of bound complex on binding site *i*:

$$[U2AF2{:}\,Site_i] = \frac{[Site_i]_{total} \cdot [U2AF2]}{\frac{k_{offi}}{k_{oni}} + [U2AF2]} = \frac{[Site_i]_{total} \cdot [U2AF2]}{k_{di} + [U2AF2]}$$

(5)

Eq. 5 contains the dissociation constant of each RNA-protein complex $k_{di} = \frac{k_{offi}}{k_{oni}}$.

2. Linking the model to experimental *in vitro* iCLIP measurements

The dissociation constant of each binding site was estimated by fitting the binding model (Eq. 5) to the *in vitro* iCLIP measurements (see Section 3). Model and experimental data were compared using the following relationship

$$Signal_i = SF_i \cdot N \cdot [U2AF2{:}Site_i] \cdot e^{\sigma Z_i} = SF_i \cdot N \cdot \frac{[Site_i]_{total} \cdot [U2AF2]}{k_{di} + [U2AF2]} \cdot e^{\sigma Z_i} \quad (6)$$

The iCLIP signal is assumed to be proportional to the concentration of the complex. A proportional experimental error ($e^{\sigma Z_i}$) was considered (Z_i being an independent normal random variable), because the standard deviation of biological quadruplicates scaled with the intensity of the *in vitro* iCLIP signal (**Supplemental Fig. S7A**). In further support for a proportional error, we observed that the logarithm of the iCLIP signal was normally distributed, with a standard deviation that was constant across U2AF2 concentrations (**Supplemental Fig. S7B**).

The signal is also proportional to a "scaling factor" (SF_i) that is specific for each binding site (i), but the same across all experimental runs. The scaling factor can be interpreted as the sum of binding site-specific biases, such as a UV crosslinking or PCR amplification efficiency.

Finally, we assumed a normalization factor (N) that scales all binding sites for a given experimental replicate. This normalization factor reflects the batch effect of different sequencing depths. Our data contains such a batch effect, since the *in vitro* iCLIP signals of two replicates are highly correlated, but shifted by a global factor between replicates (**Supplemental Fig. S6A**). Importantly, these differences can be efficiently compensated for when normalizing the data to total library size (**Supplemental Fig. S6B**). However, the latter normalization also abolishes overall signal differences between U2AF2 concentrations, implying that the use of a fitted normalization factor is superior, as it allows for U2AF2 concentration-dependent normalization. The use of a fitted normalization factor yielded comparable results to normalization of each sample by a spike-in (a U2AF2-bound RNA of known concentration) which we added to each sample before applying the *in vitro* iCLIP procedure (**Supplemental Fig. S6C**).

3. Model calibration by maximum likelihood fitting

The simulated iCLIP signal (Eq. 6) was fitted to the *in vitro* iCLIP titration experiment, in which known, constant concentrations of eleven *in vitro* transcripts were incubated with eight different concentrations of recombinant U2AF2^{RRM12} (four replicates) in order to estimate the unknown parameters: SF_i, N, k_{di}, σ.

For all experiments, we assumed the same dissociation constant k_{di}, the same scaling factor SF_i and the same binding site concentration $[Site_i]_{total}$. In contrast, each experimental run (replicates or experiments with different U2AF2^{RRM12}

concentrations) was characterized by a specific normalization factor (N_j) and a specific relative (log-constant) error ($e^{\sigma Z_{ij}}$), as indicated by the subscript *j*:

$$Signal_{ij} = SF_i \cdot N_j \cdot \frac{[Site_i]_{total} \cdot [U2AF2_j]}{k_{di} + [U2AF2_j]} \cdot e^{\sigma Z_{ij}} \tag{7}$$

Our model assumes that the signal is normally distributed at log-scale (see above), giving rise to the following likelihood function:

$$-2\ln\left(L(\theta|Signal_{ij})\right) = \sum_{i=1}^{I}\sum_{j=1}^{J}\ln(2\pi\sigma^2) + \frac{1}{\sigma^2}\sum_{i=1}^{I}\sum_{j=1}^{J}\left[\ln(Signal_{ij}) - \ln\left(SF_i \cdot N_j \cdot \frac{[Site_i]_{total} \cdot [U2AF2_j]}{k_{di} + [U2AF2_j]}\right)\right]^2 \tag{8}$$

Here, θ denotes the parameter set { SF_i, N_j, k_{di}, σ }, *I* the total number of binding sites (*I*=795), and *J* the total number of experimental runs (*J*=31, different U2AF2^{RRM12} concentrations and/or replicates).

By maximizing *L*, the maximum likelihood estimates of the unknown parameters can be obtained. It is more common, equivalent and numerically more efficient to minimize the negative logarithm of the likelihood function -2ln(L) instead (Eq. 8). Since the relative error σ in Eq. 8 is a constant, the minimization of -2ln(L) reduces to the minimization of the part *f*:

$$\begin{aligned} f &= \sum_{i=1}^{I}\sum_{j=1}^{J}\left[\ln(Signal_{ij}) - \ln\left(SF_i \cdot N_j \cdot \frac{[Site_i]_{total} \cdot [U2AF2_j]}{k_{di} + [U2AF2_j]}\right)\right]^2 \\ &= \sum_{i=1}^{I}\sum_{j=1}^{J}\left[\ln(Signal_{ij}) - \ln(SF_i) - \ln(N_j) - \ln\left(\frac{[Site_i]_{total} \cdot [U2AF2_j]}{k_{di} + [U2AF2_j]}\right)\right]^2 \end{aligned} \tag{9}$$

f is our cost function for maximum likelihood estimation of parameters, which is essentially a simple nonlinear least squares fit to the logarithm of the signal.

***Strategies for parameter estimation*:** In many cases, the parameters of a model cannot be unequivocally determined based on the available experimental data ("non-identifiability"), because the model contains too many independent parameters. An obvious non-identifiability problem can be observed when minimizing our objective function (Eq. 9): The normalization factor N_j and scaling factor SF_i both enter the formula as a proportional factor, implying that the fitting result would be unchanged by simultaneously scaling all N_j 2-fold up, and all SF_i 2-fold down. To circumvent this problem and to allow for more efficient parameter estimation, we ensured identifiability of the scaling and normalization factors by considering the following arbitrary "sum-to-zero" constraint during the optimization:

$$\sum_{j=1}^{J}\ln(N_j) = 0$$

(10)

In Eq. 9, there is a dissociation constant parameter (k_{di}) and a scaling factor parameter (SF_i) for each of the 795 binding sites, and a normalization factor (N_j) for each experimental run, giving rise to a total of 795*2+31=1621 parameters that need to be estimated from fitting to 24,645 data points (795 binding sites; 4 replicates, each with eight or nine U2AF2^{RRM12} concentrations). We introduced two simplification steps, described as I and II below, in order to reduce the complexity of the parameter space to 795 parameters (I) and to speed up the computation (II).

Simplification I: In Eq. 9, we let y_{ij} denote $\ln(Signal_{ij})$, sf_i denote $\ln(SF_i)$, n_j denote $\ln(N_j)$, c_{ij} denote the log-transformed complex concentration of binding site *i* at protein concentration *j*: $\ln(\frac{[Site_i]_{total}\cdot[U2AF2_j]}{k_{di}+[U2AF2_j]})$. Then, our objective function can be written as:

$$f = \sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij} - sf_i - n_j - c_{ij})^2$$

and the constraint is $g = 0$, where

$$g = \sum_{j=1}^{J} n_j.$$

(11)

We want to find the maximum likelihood estimates $\widehat{k_{di}}, \widehat{sf_i}, \widehat{n_j}$. In the following, we will show that only the binding affinities of the 795 binding sites need to be estimated by fitting to data, because $\widehat{sf_i}$ and $\widehat{n_j}$ can be directly calculated from $\widehat{k_{di}}$,

Let $\widehat{c_{ij}} = \ln(\frac{[Site_i]_{total}\cdot[U2AF2_j]}{\widehat{k_{di}}+[U2AF2_j]})$. Since

$$\frac{\partial f}{\partial(sf_i)} = -2\sum_{j=1}^{J}\left(y_{ij} - sf_i - n_j - c_{ij}\right),$$

the maximum likelihood estimates $\widehat{sf_i}$, $\widehat{n_j}$, $\widehat{c_{ij}}$ must satisfy

$$\sum_{j=1}^{J}(y_{ij} - \widehat{sf_i} - \widehat{n_j} - \widehat{c_{ij}}) = 0.$$

(12)

Since $\widehat{sf_i}$ does not depend on *j*, we can write:

$$\sum_{j=1}^{J}(y_{ij} - \widehat{n_j} - \widehat{c_{ij}}) - J\cdot\widehat{sf_i} = 0$$

and solve it for $\widehat{sf_i}$:

$$\widehat{sf_i} = \frac{1}{J}\sum_{j=1}^{J}(y_{ij} - \hat{n}_j - \widehat{c_{ij}})$$

From the above constraint, we have:

$$\sum_{j=1}^{J}\hat{n}_j = 0 \tag{13}$$

then $\widehat{sf_i}$ can be written as:

$$\widehat{sf_i} = \frac{1}{J}\sum_{j=1}^{J}(y_{ij} - \widehat{c_{ij}}) \tag{14}$$

This demonstrates that $\widehat{sf_i}$ can be directly determined from $\widehat{k_{di}}$.

Similar calculations can be done for $\hat{n}_j$. The method of Lagrange multipliers is employed in this case as a strategy for finding the local minima of a function subject to equality constraints (Eq. 11). Introducing a Lagrange multiplier λ for the constraint (Eq. 13), we seek solutions of

$$\frac{\partial f}{\partial(n_j)} + \lambda \cdot \frac{\partial g}{\partial(n_j)} = 0.$$

Thus, the maximum likelihood estimates must satisfy

$$-2\sum_{i=1}^{I}\left(y_{ij} - \widehat{sf_i} - \hat{n}_j - \widehat{c_{ij}}\right) + \lambda \cdot 1 = 0. \tag{15}$$

Since $\hat{n}_j$ does not depend on *i*, we can write:

$$\sum_{i=1}^{I}(y_{ij} - \widehat{sf_i} - \widehat{c_{ij}}) - I \cdot \hat{n}_j - \frac{\lambda}{2} = 0$$

and solve it for $\hat{n}_j$:

$$\hat{n}_j = \frac{1}{I}\left[\sum_{i=1}^{I}(y_{ij} - \widehat{sf_i} - \widehat{c_{ij}}) - \frac{\lambda}{2}\right] \tag{16}$$

We can solve for λ in the above equation by summing both sides of Eq. 15 with respect to all *j*:

$$-2\sum_{j=1}^{J}\sum_{i=1}^{I}\left(y_{ij}-\widehat{sf_i}-\widehat{n_j}-\widehat{c_{ij}}\right)+J\cdot\lambda=0$$

and hence

$$-2\sum_{i=1}^{I}\left[\sum_{j=1}^{J}\left(y_{ij}-\widehat{sf_i}-\widehat{n_j}-\widehat{c_{ij}}\right)\right]=-J\cdot\lambda$$

Together with Eq. 12, we get

$$\lambda=0$$

By replacing the λ in Eq. 16, we get

$$\widehat{n_j}=\frac{1}{I}\sum_{i=1}^{I}(y_{ij}-\widehat{sf_i}-\widehat{c_{ij}}) \tag{17}$$

This demonstrates that $\widehat{n_j}$ can be directly determined from $\widehat{sf_i}$ and $\widehat{k_{di}}$. Thus, using Eqs. 14 and 17, the set of parameters to be estimated by fitting can be simplified from {SF_i, N_j, k_{di}} to {k_{di}}.

Simplification II: Eq. 5 describes the U2AF2^{RRM12}-RNA complex concentration(s) as a function of free U2AF2^{RRM12}, with the additional constraint that all U2AF2^{RRM12}-containing model species need to sum up to the total U2AF2^{RRM12} concentration (Eq. 3). The latter condition reflects the possible competition of U2AF2 binding sites for a limiting U2AF2^{RRM12} concentration. The optimization procedure can be greatly speeded up when neglecting this competition and assuming that the free U2AF2^{RRM12} concentration at experimental run *j* is given by the total protein concentration, i.e., $\left[U2AF2_j\right]_{total}=[U2AF2_j]$. This assumption is justified for the *in vitro* titration experiments, since the RNA is present at a very low concentration of 0.2 nM, implying that only 0.1598 µM U2AF2^{RRM12} will be bound even if all 795 binding sites are fully occupied. Most of the U2AF2^{RRM12} concentrations used in the titration experiments (0.15 to 25 µM) largely exceed this level. For the potentially limiting lowest protein concentration (0.15 µM), we find little binding to RNA, and therefore apply this simplification during the optimization procedure.

***Treatment of low read counts*:** Our model assumes that the *in vitro* iCLIP signal is proportional to the concentration, scaling factor and normalization factors, but this basic proportionality assumption no longer applies if the signals are too low. In this case, the read count from sequencing exhibits a probabilistic component, as exemplified by the fact that most of the measured signals are 0 at very low U2AF2^{RRM12} concentrations (0.05 µM and below) which were therefore excluded from this study. More suitable error models like the Poisson model or negative-binomial model could be explored to accurately describe this scenario (Bullard et

al. 2010). In our fitting procedure, we added a pseudocount of 1 to all signals which is negligibly small compared to binding signals at high U2AF2^{RRM12} concentrations. This renders zero signals to be manageable by our logarithmic model.

***Parameter ranges*:** All model parameters are by definition non-negative. Thus, a log-scale is used in the parameter estimation in order to ensure that parameters being potentially different by orders of magnitude are handled with equal efficiency by numerical computations. The allowed search space for k_d parameters is [0.1 μM, 1000 μM], i.e., [-1,3] on a $\log_{10}$-scale. In search for the minimum of the objective function (Eq. 9), we used the local optimization algorithm *lsqnonlin* in MATLAB, and tried 100 multi-start optimization with latin-hypercube sampling from logarithmic space of the parameters, to ensure even sampling across all orders of magnitude. The Pearson correlation coefficient between model fit and data is 0.78 (p-value < 0.001; **Supplemental Fig. S6D**).

4. Parameter uncertainty analysis

The titration experiments were performed at a limited set of U2AF2^{RRM12} concentrations, and even the highest of these concentrations will not be sufficient to completely saturate all low-affinity binding sites. Owing to these limitations, it is likely that the model fitting to the data may not allow us to precisely estimate all binding affinities, scaling factors and normalization factors ("non-identifiability"). As described in the following, we applied uncertainty analysis to better understand these uncertainties and to assign a confidence interval to each of the parameter estimates.

Parameter uncertainties were assessed using the profile likelihood approach (Raue et al. 2009). These likelihood profiles break down the uncertainty contained in the high-dimensional likelihood to a footprint in one dimension: Each parameter is systematically perturbed around its best-fit value, and fixed to this perturbed value, while allowing all remaining parameters to change when refitting the model to the data. Using this approach, one obtains a two-dimensional profile for each parameter, the profile likelihood, in which the goodness-of-fit (here, it is the -2log(likelihood)) is shown as a function of the fixed parameter value. Profile likelihood-based confidence interval for each parameter could be calculated using the likelihood ratio test at a 95% confidence level (α=0.05, degrees of freedom=1) (Raue et al. 2009).

Three outcomes are feasible when calculating the profile likelihood for a given parameter: (i) A profile that is completely flat with no unique minimum implies that the parameter confidence interval is infinitely extended in both increasing and decreasing directions of the (logarithmic) parameter space. This indicates that the parameter is structurally non-identifiable, i.e., that it determines the goodness-of-fit only in combination with functionally related parameters. (ii) There is a unique minimum on the likelihood profile, but the likelihood-based confidence region is infinitely extended in one or both directions of parameter space of the fixed parameter. This indicates that the parameter is practically non-

identifiable, implying that there is not enough experimental data or that the experimental noise is too large. (iii) A convex profile, in which the goodness-of-fit drastically decreases when the parameter deviates from the best-fit value, indicates that the parameter can be well estimated based on the available experimental data. In this case, a finite upper and lower bound of the confidence interval can be identified.

We obtained finite confidence intervals for 633 out of 795 parameters, indicating that most of the parameters are identifiable based on the available experimental data. The confidence interval of the estimated binding affinities (k_d) gets larger for larger best-fit values of the k_d (**Fig. 2D**). This reflects the expectation that k_d can no longer be distinguished in Eq. 6 if the binding affinity exceeds the applied $U2AF2^{RRM12}$ concentration. The k_d values of the high-affinity and intermediate binding sites up to k_d < 18 µM were well constrained by the data, as the upper and lower boundaries of the confidence intervals on average differed by a factor of 5.16. The other k_d parameters are practically non-identifiable (according to the above definition), because they have only the lower boundary of the confidence interval well identified. The unclosed upper boundary is because of lack of experimental data at even higher $U2AF2^{RRM12}$ concentrations which would allow these binding sites to go into saturation.

5. Absence of cooperativity in U2AF2 binding

Nucleic acid-binding proteins frequently bind nearby target sequences with (anti-) cooperativity, i.e., the binding of a protein to the first site strongly enhances (reduces) the binding of a second protein to the neighboring binding site. In our *in vitro* assays, we characterized the binding of the truncated $U2AF2^{RRM12}$ protein to RNA. Given that the truncation of the $U2AF2^{RRM12}$ protein eliminates the RS domain as prominent protein-protein interaction interface, we considered cooperation of two molecules as unlikely. To support this notion, we compared the fitting results of our simple model to the fits of a more complex (anti-)cooperative binding model.

In mathematical terms, (anti-)cooperativity was implemented by fitting a Hill equation which represents an extension of Eq. 5 to the binding data:

$$[U2AF2{:}\,Site_i] = \frac{[Site_i]_{total} \cdot [U2AF2]^n}{k_{di} + [U2AF2]^n}$$

For n>1, this equation gives rise to a steep, sigmoidal dose-response curve, reflecting cooperativity. During fitting, it was assumed that the Hill coefficient *n* is a fitted parameter that is estimated separately for each binding site (within the range of n = [0.5;5]).

A direct comparison shows that the goodness-of-fit is similar for the two binding models (**Supplemental Fig. S7C**), with the exception of stronger deviations for few binding sites (off-diagonal points) for which the model predicts (anti-)cooperative regulation. The total cost function over all binding sites, i.e., the *-2log(likelihood)*, is $-2L_n$=37816 and $-2L_1$=39347 for n≠1 and n=1, respectively.

This is consistent with a better goodness-of-fit for the more complex model incorporating cooperativity.

We calculated the Akaike Information Criterion to estimate whether the addition of the Hill coefficient as a parameter resulted in a significant improvement, or merely reflected an overfitting of the data (model selection problem). In support of the one-step model, we find that $(AIC)_n > (AIC)_1$ (or specifically $\Delta(AIC) = -2L_n-(-2L_1)+2*799=66.6$), i.e., there is no obvious improvement by introducing new parameters *n*. Taken together, this suggests that a simple one-step model is sufficient to describe the data with an accuracy close to measurement noise, whereas the cooperative model yields no significant further improvement, except for very few binding sites.

II. Model-based analysis of *in vivo* binding landscapes

The intrinsic binding behavior of recombinant U2AF2^{RRM12} in '*in vitro* iCLIP' showed no specific enrichment of high-affinity binding at 3' splice sites (**Fig. 2G**). In contrast, we found U2AF2 to be enriched at these sites when performing *in vivo* iCLIP measurements from living cells (**Fig. 1E**). This suggests that U2AF2 binding is modulated *in vivo* by auxiliary RNA-binding proteins (RBPs) recognizing sequence elements nearby the 3' splice site.

We employed our *in vitro* binding model to systematically identify differences between the *in vitro* and *in vivo* binding landscapes. We restricted these analyses to the nine *in vitro* transcripts that are derived from protein-coding genes and display well-defined splicing patterns *in vivo*, corresponding to 571 binding sites. Specifically, we asked for which binding sites we can assume the binding affinity to be unchanged, and where we have to assume additional regulation *in vivo*. To this end, we searched for the best overlap by fitting the *in vitro* model to the *in vivo* iCLIP landscape under the assumption that the *in vitro* binding affinities continue to hold, whereas the RNA and protein concentrations can be different *in vivo*. This *in vivo* modeling approach is described in more detail below.

1. Model extension and fitting to *in vivo* iCLIP landscapes

The model describing *in vivo* U2AF2 binding essentially corresponds to the *in vitro* model, as U2AF2-RNA binding and the iCLIP signals are still described by Eqs. 5 and 6, respectively. Some biophysical parameters such as the intrinsic U2AF2 binding affinity for RNA and the crosslinking efficiency at each binding site are assumed to be the same *in vitro* and *in vivo*. Furthermore, the same set of transcripts (and thus binding sites) is analyzed.

However, there are also several differences between the *in vitro* and *in vivo* situations:

a) The U2AF2 protein and transcript concentrations in living cells are not known. Furthermore, the transcripts are spliced and degraded in living cells, and each intron may be turned over with a different half-life. Thus,

different parts of the transcripts (i.e., sets of neighboring binding sites) may have different concentrations in living cells.

b) Living cells do not only contain the eleven tested transcripts, but a large number of additional binding sites in the transcriptome that may sequester U2AF2. Analytical calculations demonstrated that the presence of such additional binding sites only affects the free U2AF2 concentration that diffuses in the cell, but does not affect the general structure of the simple binding model in Eq. 5 (not shown). We therefore continued to use this model *in vivo*.

Considering all the above points, the *in vivo* binding signal is a modification of Eq. 6 and given by:

$$Signal_{i,invivo} = SF_i \cdot N_{invivo} \cdot \frac{\left[Site_{i,intron}\right]_{total} \cdot [U2AF2_{invivo}]}{k_{di} + [U2AF2_{invivo}]} \cdot e^{\sigma Z_{i,invivo}} \tag{18}$$

Each of the three *in vivo* iCLIP replicates is normalized by the median over all signals in the respective replicate. Therefore, the three replicates will get the same *in vivo* normalization factor (N_{invivo}) as in Eq. 18. Since this normalization factor enters Eq. 18 as an overall proportionality factor, it also introduces "structural non-identifiability" (a change in N_{invivo} could be compensated by a change in all $[Site_{i,intron}]_{total}$). To address this problem, we lumped the product of N_{invivo} and $[Site_{i,intron}]_{total}$ into one identifiable parameter ($Site_{i,lump,intron}$), and obtain

$$Signal_{i,invivo} = SF_i \cdot \frac{Site_{i,lump,intron} \cdot [U2AF2_{invivo}]}{k_{di} + [U2AF2_{invivo}]} \cdot e^{\sigma Z_{i,invivo}} \tag{19}$$

The parameters common to the *in vitro* and *in vivo* situation (SF_i, k_{di}) were estimated from the previously described fitting to the *in vitro* iCLIP titration experiments (Section I.3), and were fixed to these values. The remaining parameter values were determined by fitting Eq. 19 to the *in vivo* iCLIP signals using a maximum likelihood approach and a local multi-start optimization strategy (see above). In contrast to the *in vitro* model fitting, we did not assume the free pool of U2AF2 to be present in excess over the transcripts, and hence allowed for protein sequestration effects between the binding sites.

We restricted the following analyses to the nine *in vitro* transcripts (excluding *MALAT1* and *MIRLET7A2*) that are derived from protein-coding genes and display well-defined splicing patterns *in vivo*. The nine *in vitro* transcripts contain 571 U2AF2 binding sites on 29 introns that were used for fitting. The remaining free parameters are the free U2AF2 protein concentration ($[U2AF2_{invivo}]$), 29 lumped binding site concentrations ($Site_{i,lump,intron}$) - each corresponding to one intron, and the *in vivo* noise factor (σ). During fitting, the free U2AF2 protein concentration was allowed to vary on a range of 0.01 – 1000 µM, and the lumped binding site concentration between 10^{-2} - 10^{6} µM.

A profile likelihood analysis showed that the fitted parameters could be well identified from the available experimental data. Furthermore, the best-fit parameters are physiologically reasonable: For instance, the concentration of the free (unbound) U2AF2 level *in vivo* (~11 µM) is only slightly above the total U2AF2 concentration reported for NIH3T3 cells (~7 µM) (Schwanhäusser et al. 2011).

2. Identification of regulatory hotspots *in vivo*

The fitting procedure described above yielded the maximal overlap between *in vitro* and *in vivo* binding behaviors that can be obtained by arbitrarily choosing the RNA and protein concentrations, while fixing the U2AF2-RNA binding affinities. This overlap generates hypotheses at which binding sites U2AF2 binding is regulated *in vivo* beyond simple RNA sequence recognition. In order to identify these regulatory hotspots, we needed to quantify at which binding sites the "expected *in vivo* signal" given by the model fit differs from the *in vivo* measurement.

We quantified this difference for binding site *i* and normalized it to the experimental variation to obtain a z-score:

$$z_i = \frac{\ln(Signal_{i,invivo}) - \ln(Signal_{i,model})}{\sigma_{invivo}} \tag{20}$$

Here, σ_{invivo} is the relative (log-constant) error estimated as the standard deviation of the three *in vivo* iCLIP replicates. We called binding sites as regulatory hotspots if the difference between model fit and experiment was bigger than the experimental variation ($|z_i| > 1$). The sign of the z-score indicates whether a binding site showed higher or lower binding affinity *in vivo* when compared to the *in vitro* situation ($z > 1$ and $z < -1$, respectively).

Based on this strategy, we estimated 57% (324 out of 571) of the binding sites to be regulated *in vivo*, with 26% (151) and 30% (173) being stabilized and cleared, respectively. The distribution of z-scores is symmetric around $z = 0$. Most 3' splice sites show enhanced binding *in vivo*, whereas the z-scores are symmetrically distributed for intronic sites (**Supplemental Fig. S3A**). We validated the set of predicted regulatory hotspots using a step-wise fitting approach described in the following.

***Step-wise fitting approach*:** We asked whether the assumption that dissociation constants are identical between *in vitro* and *in vivo* may be too stringent, possibly introducing a bias in the identification of regulatory hotspots. Therefore, we implemented a more realistic fitting approach in which subsets of the k_{di} values were allowed to change during fitting the *in vivo* landscape. These changes reflect that auxiliary RBPs acting *in vivo* will affect the apparent U2AF2 binding affinity at some sites, thereby also influencing the fitting result.

To this end, we employed a step-wise, greedy hill-climbing approach, in which the model was repeatedly fitted to the *in vivo* binding landscape, while allowing an increasing number of k_{di} values to be distinct from their *in vitro*-derived estimate. The fitting sequence started with the above-mentioned fit, where all k_{di} values corresponded to the *in vitro* estimates. At each subsequent iteration, we additionally allowed the affinity of the binding site that had the greatest z-score in the previous iteration (i.e., was most different from the *in vivo* signal) to have a k_{di} that deviates from its *in vitro* value. In this way, additional degrees of freedom allowed the model fit to improve in a step-wise manner, and the distance measurement (z-score) was recalculated adaptively in each iteration. The procedure was terminated if the model exhibited too many degrees of freedom and overfitted the data according to the Bayesian Information Criterion (BIC). The list of *in vivo* regulated binding sites retrieved from this approach strongly overlapped with the set from the simpler approach (see above), especially for the 100 top-ranked binding sites (**Supplemental Fig. S6E**), suggesting that our predictions of *in vivo* regulated binding sites are robust.

Supplemental methods

Preparation of recombinant proteins

6xHis-tagged U2AF2^{RRM12}, full-length U2AF2 and hnRNPC1 recombinant constructs were overexpressed in a *Escherichia coli* BL21-CodonPlus(DE3)-RIL strain under IPTG induction for 3-4 hours (h). The recombinant proteins were then purified with affinity purification by using Ni Sepharose 6 Fast Flow (GE Healthcare). Eluted recombinant proteins were concentrated in binding buffer (10 mM HEPES pH 7.2, 100 mM KCl, 3 mM $MgCl_2$, 5% glycerol, 1 mM DTT) with Spin-X UF 500 5K MWCO columns (Corning). Additional purification with size selection chromatography was applied for the recombinant full-length U2AF2 and hnRNPC1 protein preparation to achieve higher purity. Recombinant FLAG-tagged PTBP1 expressed and purified from mammalian cells was obtained from Kelifa Arab (Heidelberg University). For the co-factor experiments, all GST-tagged recombinant RBPs except hnRNPC1 and PTBP1 (CELF6, ELAVL1, FUBP1, KHDRBS1, MBNL1, PCBP1, RBM24, RBM41 and SNRPA) were purchased from Abnova as *in vitro* translation products (**Supplemental Fig. S4**). We noted that the preparations for PTBP1, RBM41, RBM24, PCBP1 and SNRPA showed small amounts of additional minor bands which could potentially impact on the *in vitro* iCLIP co-factor assays.

Preparation of *in vitro* transcripts

In total, eleven different *in vitro* transcripts were used for the *in vitro* iCLIP experiments (**Supplemental Table S1**). The transcripts were chosen to harbor a diverse set of constitute and alternative exons as well as to show high coverage with U2AF2 *in vivo* iCLIP reads which facilitated comparative *in vitro – in vivo* analysis. Briefly, vectors harboring genes for the transcript set were *in vitro* transcribed by using Riboprobe System-T7 (Promega) according to the manufacturer's instructions. The *in vitro* transcripts were then treated with TURBO DNase (Ambion) and purified with the RNeasy MinElute Cleanup Kit (Qiagen). Concentrations of each purified transcript were determined with a NanoDrop 2000 system to estimate the required volume for making the stock of the equimolar *in vitro* transcript mix.

in vivo iCLIP library preparation and sequencing

in vivo iCLIP libraries were prepared from HeLa cells under *U2AF2* KD and wild-type conditions according to the previously published protocol (Huppertz et al. 2014; Sutandy et al. 2016). HeLa cells were obtained from ATCC (number: CCL-2). For immunoprecipitation, we used 7.5 µg monoclonal anti-U2AF2 antibody produced in mouse (Sigma cat. no. U4758) per sample. Each condition was done in triplicates. The libraries were sequenced as single-end reads on an Illumina HiSeq 2500 and an NextSeq 500 sequencing system. An overview of the *in vivo* iCLIP libraries is given in **Supplemental Table S7**.

In vitro iCLIP library preparation and sequencing

The *in vitro* iCLIP protocol was developed by modifying the early steps of the standard iCLIP protocol (Huppertz et al. 2014; Sutandy et al. 2016). Briefly, beads were prepared by twice washing 40 µl of protein-G Dynabeads per sample with dilution buffer (50 mM Tris-HCl pH 7.4, 100 mM NaCl, 1% Igepal CA-630, 0.1% SDS, 0.5% sodium deoxycholate; corresponding to the lysis buffer in the *in vivo* iCLIP protocol). After the second wash, 40 µl dilution buffer was added to resuspend the beads and followed by mixing with 3 µg anti-U2AF2 antibody. The beads were rotated at room temperature for 30-60 minutes (min). One-time high-salt buffer (50 mM Tris-HCl pH 7.4, 1 M NaCl, 1 mM EDTA, 1% Igepal CA-630, 0.1% SDS, 0.5% sodium deoxycholate) and twice dilution buffer washes were applied to wash the beads before proceeding with immunoprecipitation.

The *in vitro* transcripts were preheated for 5 min at 70°C to reduce large-scale RNA secondary structures. Titrated concentrations of $U2AF2^{RRM12}$ (150 nM, 250 nM, 450 nM, 750 nM, 1.5 µM, 3 µM, 5 µM, 15 µM) and 2.2 nM *in vitro* transcript mix (eleven transcripts) were used for the K_d measurements. For the initial hnRNPC1 titration experiment, 1 µM $U2AF2^{RRM12}$ was mixed with 6.75 nM *in vitro* transcript mix (nine transcripts; excluding *MALAT1* and *MIRLET7A2*) and different concentrations of recombinant hnRNPC1 (200 nM, 500 nM, and 1 µM) in binding buffer. For the co-factor experiments, 500 nM $U2AF2^{RRM12}$ was mixed with 6.75 nM *in vitro* transcript mix (nine transcripts) and different concentrations of eleven recombinant RBPs in binding buffer. In addition, 500 nM BSA was added to 500 nM $U2AF2^{RRM12}$ and 6.75 nM *in vitro* transcript mix as a control. Moreover, to test the linearity between input material and output of the *in vitro* iCLIP experiment, five different dilutions (1x, 2x, 4x, 8x and 16x) of a mixture of 2.5 µM $U2AF2^{RRM12}$ and 6.75 nM *in vitro* transcripts (nine transcripts) were prepared.

All *in vitro* mixtures were incubated for 10 min at 37°C. After the incubation, the mixtures were placed on a parafilm-coated plate on top of an ice plate and UV-irradiated with 5 mJ/cm^2 250 nm UV wavelength (Stratalinker 2400). Since only a minor fraction of the overall interactions (<5%) are expected to be crosslinked during this time, the irradiation should not dramatically shift the binding equilibrium. The irradiated *in vitro* mixtures were pooled back to the tubes, and dilution buffer was added to fill the samples to a volume of 1 ml. To normalize the final *in vitro* iCLIP libraries, 10 µl crosslinked mixture containing 250 nM $U2AF2^{RRM12}$ and 6 nM *NUP133 in vitro* transcript was spiked in to each sample. Partial RNase digestion was performed by adding 10 µl of 1:1500 diluted RNase I (Ambion) to each sample. In addition, 2 µl TURBO DNase was added to each sample to avoid DNA contamination. The sample mixtures were incubated for 3 min at 37°C, added to the prepared beads and incubated for 2 h at 4°C. Beads were washed twice with high-salt buffer and twice with wash buffer (20 mM Tris-HCl pH 7.4, 10 mM $MgCl_2$, 0.2% Tween-20).

Henceforth, we followed the steps of the standard iCLIP protocol. Briefly, 3' end RNA dephosphorylation was performed by resuspending the beads in 20 µl of a

mixture containing 4 µl 5x PNK buffer (350 mM Tris-HCl pH 6.5, 50 mM $MgCl_2$, 5 mM DTT), 0.5 µl PNK (NEB), 0.5 µl RNasin Ribonuclease Inhibitor (Promega), and 15 µl water, followed by incubation for 20 min at 37°C. The beads were washed once with wash buffer, once with high-salt buffer and twice with wash buffer.

For the linker ligation, pre-adenylated L3 linker (5'-App-AGATCGGAAGAGCGGTTCAG-dideoxycytidine-3') was ligated by resuspending the beads in the ligation mixture containing 5 µl 4x ligation buffer (200 mM Tris-HCl pH 7.8, 40 mM $MgCl_2$, 4 mM DTT), 1 µl T4 RNA ligase (NEB), 0.5 µl RNasin, 1.5 µl pre-adenylated L3 linker (20 µM), 4 µl PEG400, and 8 µl water. The samples were incubated at 16°C overnight. The next day, the samples were washed twice with high-salt buffer and twice with wash buffer.

Interacting RNAs were radioactively labeled by resuspending the beads in hot PNK mix (0.2 µl PNK [NEB], 0.4 µl 10x PNK buffer [NEB], 0.4 µl ^{32}P-γ-ATP, and 3 µl water). The beads were incubated at 1,100 rpm for 5 min at 37°C. Supernatants were removed and the beads were boiled in 20 µl 1x NuPAGE loading buffer (Invitrogen) for 5 min at 70°C. Boiled beads were placed on a magnetic rack. The supernatants were then loaded into the 4-12% NuPAGE Bis-Tris gel (Invitrogen) and run in 1x MOPS buffer for 50 min at 180 V. Protein-RNA complexes from the gel were transferred to a nitrocellulose membrane for 1 h at 30 V.

To extract the interacting RNAs, the membrane was cut into pieces and digested with 10 µl proteinase K (Roche) in 200 µl PK buffer (100 mM Tris-HCl pH 7.4, 50 mM NaCl, 10 mM EDTA) for 20 min at 37°C. Another 200 µl PK buffer containing 7 M urea were added for further 20 min incubation at 37°C. The RNA-containing mixtures were transferred to Phase Lock Gel Heavy tubes and mixed with 400 µl phenol/chloroform by shaking with 1,100 rpm for 5 min at 30°C. RNAs were extracted by centrifugation for 5 min at 16,000 xg to separate the phases followed by transferring the top aqueous phase containing RNAs to new tubes. The samples were then mixed with 0.75 µl GlycoBlue (Ambion), 40 µl 3 M sodium acetate pH 5.5 and 1 ml ethanol absolute, and incubated overnight at -20°C. To precipitate the RNAs, the samples were centrifuged with 21,000 xg for 20 min at 4°C, washed with 80% ethanol and resuspended in 5 µl water.

cDNA synthesis was performed by adding 1 µl dNTP mix and 1 µl RT primers containing different barcode sequences to each sample (**Supplemental Table S7**), and incubating them for 5 min at 70°C. The reaction was started by adding RT mixture (4 µl 5x RT buffer [Invitrogen], 1 µl 0.1 M DTT, 0.5 µl RNasin, 0.5 µl Superscript III [Invitrogen], 7 µl water) to the samples and incubating them for 5 min at 25°C, 20 min at 42°C, 40 min at 50°C, 5 min at 80°C, and hold at 4°C. To hydrolyze hot RNA templates, 1.65 µl 1 M NaOH was added, followed by 20 min incubation at 98°C. After the incubation, 20 µl 1 M HEPES-NaOH was added to neutralize the samples' pH. The cDNA libraries were mixed with 0.75 µl GlycoBlue, 40 µl 3 M sodium acetate pH 5.5 and 1 ml ethanol absolute, and incubated overnight at -20°C. The next day, the cDNA libraries were precipitated

by spinning the samples with 21,000 xg for 20 min at 4°C, washing with 80% ethanol, and resuspension in 6 µl water.

cDNA libraries were mixed with 6 µl 2x TBE-urea loading buffer (Invitrogen), heated for 5 min at 80°C, and then loaded and run in a 6% TBE-urea gel for 40 min at 180 V. DNA low molecular weight size marker (NEB) was used as the ladder. The libraries were size-selected by cutting out the gel within the range of 80-100 nt based on the ladder. Each piece of the gel was then crushed into smaller pieces and mixed with 400 µl diffusion buffer (0.5 M ammonium acetate, 10 mM magnesium acetate, 1 mM EDTA, 0.1% SDS). The mixtures were incubated for 30 min at 50°C, and moved to a Costar SpinX column (Corning) prepared with two 1 cm glass pre-filters (Whatman). To extract the cDNA libraries, the mixtures were spun at 16,000 xg for 5 min, and the eluates were added together with 400 µl phenol/chloroform into a Phase Lock Gel Heavy tube. The samples were incubated for 5 min at 30°C, and spun at 16,000 xg for 5 min to separate the phases. The aqueous top layers containing the libraries were moved to new tubes, mixed with 1 µl GlycoBlue, 40 µl 3 M sodium acetate pH 5.5 and 1 ml ethanol absolute, and then stored at -20°C overnight.

For circularization, libraries were centrifuged with 21,000 xg for 20 min at 4°C, washed with 80% ethanol, and resuspended in 8 µl ligation mixture (0.8 µl 10x CircLigase buffer II [Epicentre], 0.4 µl 50 mM $MnCl_2$, 0.3 µl CircLigase II [Epicentre], 6.5 µl water). The libraries were transferred into PCR tubes and incubated for 1 h at 60°C. To re-linearize the libraries, 30 µl oligo annealing mix containing 3 µl FastDigest buffer (Thermo Fischer), 1 µl 10 µM cut_oligo (5'-GTTCAGGATCCACGACGACGACGCTCTTCaaaa-3'), and 26 µl water were added. The annealing program was performed by running the samples in successive cycles of 20 seconds from 95°C to 25°C with decreasing the temperature by 1°C in each cycle. After the end of the program, 2 µl of BamHI was added to each sample followed by incubation for 30 min at 37°C and heat inactivation for 5 min at 80°C. The samples were mixed with 350 µl TE, 0.75 µl GlycoBlue, 40 µl 3 M sodium acetate pH 5.5 and 1 ml ethanol absolute, and then precipitated overnight at -20°C. The next day, the libraries were extracted by spinning the samples with 21,000 xg for 20 min at 4°C, washing with 80% ethanol, and resuspension in 20 µl water.

The libraries were amplified by mixing the cDNA libraries in a PCR reaction containing 0.5 µM P3/P5 Solexa primers mix and 1x Accuprime Supermix 1 enzyme (Invitrogen). The PCR mixes were run with a program comprising a 2 min denaturation step at 94°C, 17-25 cycles of 15 seconds at 94°C, 30 seconds at 65°C and 30 seconds at 68°C, and a final elongation step for 3 min at 68°C. Several pre-PCR steps were performed to estimate the minimal number of cycles that is necessary to amplify the libraries. The amplified libraries were pooled together by purification with the MinElute PCR purification kit (Qiagen). The purified libraries were size-selected with LabChip XT DNA 300 kit (Perkin Elmer) to remove residual P3/P5 Solexa primers. The final libraries were quantified with the Qubit dsDNA HS assay kit (Invitrogen) and sequenced as

single-end reads on an Illumina MiSeq sequencing system. An overview of the *in vitro* iCLIP libraries is given in **Supplemental Table S7**.

Initial processing and genomic mapping of iCLIP sequencing reads

Basic quality checks were applied to all sequenced reads using FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Afterwards, iCLIP reads were filtered based on sequencing quality (Phred score) in the barcode region. Only reads with at most one position with a Phred score < 20 in the experimental barcode (positions 4 to 7) and without any position with a Phred score < 17 in the random barcode (positions 1 to 3 and 8 to 9) were kept for further analysis. The reads were then de-multiplexed based on the experimental barcode at positions 4 to 7 using Flexbar (version 2.4) without allowing mismatches (Dodt et al. 2012).

The following analysis steps were applied to all individual samples: Remaining adapter sequences were trimmed from the read ends using Flexbar (version 2.4) allowing one mismatch in 10 nt, requiring a minimal overlap of 1 nt between read and adapter as well as removing all reads with a remaining length of less than 24 nt (including the 9-nt barcode). After adapter trimming, quality checks were repeated on individual samples using FastQC. Afterwards, the first 9-nt of each read containing the barcode were trimmed off and added to the header of the read in the fastq file, such that the information is kept for downstream analysis.

Filtered and trimmed reads were mapped to the human genome (hg19/GRCh37) and its annotation (GENCODE release 19) (Harrow et al. 2012) using STAR (version 2.4.0h) (Dobin et al. 2013). Since the new genome version (GRCh38) was reported to only moderately improve the average mapping rate (0.0017%) (Guo et al. 2017), we do not anticipate an impact on the results. When running STAR, up to two mismatches were allowed, soft-clipping was prohibited and only uniquely mapped, unspliced reads were kept for further analysis. Unspliced reads were defined as reads mapping without N-stretches longer than 5 in the CIGAR string. The nucleotide position upstream of each aligned read was considered as the ‘crosslink nucleotide’, with each read counted as individual ‘crosslink event’. The total number of reads for all *in vitro* and *in vivo* iCLIP libraries can be found in **Supplemental Table S7**.

After mapping and filtering, duplicate reads were marked in selected samples using the dedup function from bamUtil (version 1.0.7), which defines duplicates as reads whose 5' ends map to the same position in the genome (https://github.com/statgen/bamUtil). We then removed all marked duplicates with an identical random barcode representing technical duplicates, while biological duplicates with different random barcodes were kept.

Samtools (Li et al. 2009) was used to sort and index the resulting bam files. Based on the bam files, bedgraph files were created, considering only the position upstream of the 5' mapping position of the read. bedgraph files were then transformed to bigWig format using bedGraphToBigWig of the UCSC tool suite (Kent et al. 2010).

iCLIP peak calling on *in vitro* transcripts

In order to comprehensively identify all U2AF2 binding sites within the eleven *in vitro* transcripts, peaks were called on combined *in vitro* and *in vivo* iCLIP data: From the *in vitro* iCLIP data, the experiment with 3000 nM U2AF2^{RRM12} was taken as representative *in vitro* condition (libraries no. 39-42; **Supplemental Table S7**). For *in vivo* iCLIP, we used two previously published datasets from untreated and mock transformed HeLa cells (Zarnack et al. 2013) plus a newly produced dataset (libraries no. 1-3; **Supplemental Table S7**; replicates were summed up). Peak calling was performed on iCLIP counts, i.e., crosslink events per crosslink nucleotide (corresponding to the position upstream of each read start as decribed above). In order to restrict the analyses to pre-mRNAs, only reads mapping in a unspliced manner were taken into account. Due to the high sequencing depth of the *in vitro* iCLIP experiments, we omitted duplicate removal since the number of possible random barcodes runs into exhaustion.

For merging *in vivo* and *in vitro* iCLIP data, data were first normalized for each replicate/dataset, then separately combined within *in vitro* and *in vivo* and finally summed up as follows: In order to account for differences in sequencing depth, the data within each sample were normalized per intron and transcript by the 85% quantile of the nucleotide-wise signal. For introns shorter than 100 nt, the median of 85% quantiles of all other introns in the same transcript was taken. The nucleotide-wise median was taken as summary statistics within the four *in vitro* replicates as well as within the three *in vivo* datasets, and subsequently summed up to combine *in vivo* and *in vitro* data. Peaks were then called on the summed normalized *in vitro* and *in vivo* iCLIP counts in a sequential manner using a window-based approach: We iteratively identified the 9-nt peak window with the highest cumulative signal. Peaks were separated by at least 1 nt and called exhaustively, until no further peak of width 9 nt could be placed.

In order to retain only relevant peaks, we assessed their iCLIP count enrichment over a uniform background distribution per intron and transcript. The background signal was calculated by randomly distributing the same number of crosslink events across the respective region. Only those peaks were kept which exceeded the uniform background distribution per intron and transcript in at least 3 out of 4 *in vitro* and all 3 *in vivo* experiments. This procedure yielded a total of 795 binding sites (**Supplemental Table S2**).

These and all following computations were done with R version 3.3.1 (R Core Team 2016).

Peak calling on *in vivo* iCLIP data

For the 4-mer comparison (**Fig. 1D**), we identified all U2AF2 binding sites in the *in vivo* iCLIP dataset with the highest sequencing depth (library no. 3, **Supplemental Table S7**). Peak calling was done on merged replicates using ASPeak (Kucukural et al. 2013), taking only non-spliced reads after duplicate removal and restricted to annotated introns. Gencode v19 (Harrow et al. 2012) was used as annotation, considering only entries with ‘support level 3’ and gene-

type/transcript-type ‘protein-coding’. The predicted peaks were centered on the position at which half of the total peak signal is reached and then extended to 9 nt. Overlapping peaks were merged and newly centered on the position with the half maximum signal. This procedure yielded a total of 406,671 *in vivo* U2AF2 binding sites. For the GraphProt analysis, we repeated the ASPeak procedure on the combination of all three *in vivo* U2AF2 iCLIP datasets yielding a total 458,942 U2AF2 binding sites.

Using the same approach, we identified a total of 82,127 hnRNPC binding sites from merged replicates of previously published *in vivo* iCLIP data from HeLa cells (Harrow et al. 2012). Within the region of the *in vitro* transcripts, 126 U2AF2^{RRM12} binding sites overlap with *in vivo* hnRNPC binding sites (**Fig. 4C**).

RNA sequence composition at U2AF2 binding sites

In order to compare the RNA sequence composition at U2AF2 binding sites, we counted all 4-mers in the 9-nt peak region. In order to not provide advantage for simple motifs such as homopolymer runs, identical 4-mers within the same binding site were counted twice at most. We considered all 795 binding sites within the region of the eleven *in vitro* transcripts and compared these to the top 100,000 peaks with highest iCLIP count in the *in vivo* iCLIP dataset with the highest sequencing depth (library no. 3; **Supplemental Table S7**). Relative 4-mer frequencies are depicted in **Fig. 1D**.

For **Fig. 2F**, the occurrence of pyrimidine-rich motifs within the 9-nt peak region of the 795 U2AF2 binding sites within the eleven *in vitro* transcripts was screened in the following decreasing hierarchy: TTTT-TTTT, TTTT-YYYY, YYYY-YYYY, TTTT-YYYR, YYYY-YYYR, TTTT-NNNN, YYYY-NNNN and NNNN-NNNN. The order of the two half-sites within each motif is commutative, i.e. TTTT-YYYY and YYYY-TTTT are assigned to the same motif class. Y refers to T or C, YYYR to exactly one A or G at any position among otherwise just Y. NNNN refers to at least two A or G.

Calculation of binding site accessibility

RNAplfold (Bernhart et al. 2006) was used to compute local sequence accessibility considering potential mid-range interactions (W = 240, L = 160). The underlying idea is that the most likely secondary structure formed by a stretch of RNA is of less importance than its accessibility derived from the ensemble of structures which this RNA stretch can form. In order to determine accessibility, RNAplfold splits a sequence into windows of defined size (W) and searches for base pairs within this window with a maximum span between the bases (L). This allows to efficiently scan large sequences for their base-pair probabilities which yields the nucleotide-wise accessibility profile of the scanned sequence. The probability value for each binding site to be unpaired was then calculated as the mean of nucleotide-wise probabilities across the 9-nt peak region. We observe that the majority of U2AF2 binding sites have a low probability of being unpaired. Binding sites with a probability of being unpaired greater/smaller than the

75% quantile of the probability distribution (0.15) were classified as accessible/unaccessible. According to this classification, 596 U2AF2 binding sites are unaccessible, and 199 accessible (**Fig. 2F**).

Assignment of U2AF2 binding sites to transcript regions

Transcript regions were defined as follows: (i) The 5' splice site comprises the first 40 nt at the 5' end of each intron. (ii) For defining the extent of the 3' splice site, we scanned for the boundaries of the polyprimidine tract (Py-tract). To this end, a 39-nt region upstream of the AG dinucleotide at the 3' splice site was screened by means of sliding windows (width 5-30 nt), to identify the window with the highest Py-tract strength. The Py-tract strength of each window was calculated as the X^2 test statistic with 1 degree of freedom, comparing the observed number of pyrimidines with the expected number based on the assumption of uniform nucleotide distribution. Additionally, candidate Py-tracts were required to end within 10 nt upstream of the AG dinucleotide. Using this approach, the median length of identified Py-tracts is 17 nt. Together, the maximum achievable length of the 3' splice site region was thus 41 nt, consisting of 2 nt (AG) + 9 nt (max. allowed distance to AG) + 30 nt (max. allowed width of Py-tract). The median length of identified 3' splice site regions (start of Py-tract to end of intron) was 21 nt. This is in agreement with a recent report showing that 90% of all human branch points occur within 39 nt upstream of the 3' splice site (Mercer et al. 2015). (iii) The remainder of the intron body was considered as 'intronic'.

For calculating the distribution of *in vivo* and *in vitro* iCLIP signal across transcript regions in the eleven *in vitro* transcripts (**Fig. 1E**), only introns longer than 85 nt were considered (40 nt and 21 nt 5' splice site and average 3' splice site length, respective, leaving 24 nt for the intervening intron). In addition, we only used 'complete' introns of which both the 5' and the 3' splice site were contained within the *in vitro* transcript boundaries, leaving 17 introns in total. The iCLIP signal distribution was calculated as the number of normalized iCLIP reads per intron and transcript scaled by the width of the underlying transcript region.

In order to investigate the genome-wide distribution of *in vivo* U2AF2 iCLIP signal upon partial *U2AF2* knockdown (**Fig. 3D**), we used Gencode v19 (Harrow et al. 2012) annotation as described above. For this analysis, introns overlapping with exonic or UTR sequences as well as introns shorter than 100 nt were removed, resulting in a final number of 169,872 introns. Replicates from control and partial *U2AF2* knockdown conditions were summed up, and signal distribution was calculated as for **Fig. 1E**. Read numbers can be found in **Supplemental Table S7**.

Prediction of branch point motifs

For the analysis of branch point locations, only introns were considered which had their 3' splice site within the boundaries of the nine protein-coding *in vitro* transcripts (26 introns harboring 496 of 571 U2AF2 binding sites). Prediction of

branch points was done using SVM-BP finder requiring a support vector machine score > 0 (Corvelo et al. 2010). 25 out of 26 introns were predicted to host at least one predicted branch point. Intronic U2AF2 binding sites were considered to be associated with a branch point if the distance to the next upstream branch point is smaller than the 75% quantile (24 nt) of distances of U2AF2 binding sites at 3' splice sites to their closest upstream branch point. Based on this definition, **Fig. 3F** includes 55 U2AF2 binding sites at 3' splice sites and 230/191 intronic U2AF2 binding sites with\without upstream branch point, respectively.

K_d measurements by MST and ITC

For the microscale thermophoresis (MST) experiment (**Figs. 2E, S2D**), RNA oligonucleotides were selected based on the *in vitro* iCLIP binding landscape. Each selected RNA oligonucleotides contained a U2AF2 binding site plus a few nucleotides upstream and downstream of the corresponding site (sequences in **Supplemental Table S3**). 5'-Cy5-labeled RNA oligonucleotides were chemically synthesized from IDT. Briefly, 5'-Cy5-labeled RNA oligonucleotides were mixed to obtain a final reaction containing 150 nM RNA and titrated concentrations of recombinant $U2AF2^{RRM12}$ in MST buffer (50 mM Tris-HCl, 150 mM NaCl, 10 mM $MgCl_2$, 0,05% Tween-20). Each mixture was loaded into an MST capillary. The K_d measurements were then performed with Monolith NT.115 (NanoTemper Technologies) at room temperature according to manufacturer's instructions and fitted with a Hill equation (Goutelle et al. 2008). For each RNA oligonucleotide, the measurements were done in triplicate.

Isothermal titration calorimetry (ITC; **Supplemental Fig. S2C,D**) was performed using MicroCal PEAQ-ITC (Microcal) at 25°C. Briefly, 300 µl 20 µM $U2AF2^{RRM12}$ protein sample (20 mM sodium phosphate, pH 6.5, 50 mM NaCl) in the ITC cell was titrated with 50 µl of 100/100/150/200 µM OR1, 200/200/200 µM OR2, 200/400 µM OR4, 250/300 µM OR5, 200/250 µM OR6, 150/200 µM OR7, and 115/115/130 µM OR8 RNA (IBA) in the same buffer (**Supplemental Table S3**). The data was further analyzed using Origin v5.0 from Microcal.

Random Forests analysis

Random Forests (RF) machine learning (Breiman 2001) was used as a classification tool to learn whether a binding site is cleared *in vivo* (z-score < -1) or stabilized *in vivo* (z-score > 1). Each binding site was characterized by a collection of features (*k*-mers, position-specific scoring matrices [PSSMs] and positional information; see below) which is used by RF to predict the direction of regulation. RF grow many classification trees, and to be classified each new binding site is put down each tree of the forest. Each tree returns a classification, and the majority vote of all trees is the final classification of the forest.

Feature Selection: Three types of features were considered: *k*-mers, PSSMs and positional information. *k*-mers and PSSMs were evaluated at three regions: an extended binding site region (9 nt peak + 5 nt up-/downstream) as well as two adjacent regions (40-nt windows flanking the extended binding site region).

Counts were normalized by the width of the underlying transcript region. In order to avoid an advantage for simple motifs such as homopolymer runs, only non-overlapping hits were counted. 6 nt was chosen as reasonable *k*-mer size to neither detect too degenerate nor too complex motifs. PSSMs for 120 unique RBPs were extracted from CISBP-RNA (Ray et al. 2013) and scored requiring at least 90% identity. If multiple PSSMs were available for a certain RBP, the highest scoring PSSM was taken. Additionally, the following features describing positional information were considered: 3' and 5' splice site score calculated by the Maximum Entropy method (Yeo and Burge 2004), the Py-tract score (as described above), as well as the length of Py-tract, AG exclusion zone and the distance to the next downstream AG, 5' splice site and 3' splice site.

Out-of-bag (OOB) error estimate: The training set for each tree is selected by sampling with replacement from the input data leaving about one third of the data unused. This out-of-bag data is used to get an unbiased estimate of the classification error. Each binding site left out in the construction of the k-th tree is put down the *k*-th tree to get a classification. Take *j* to be the class that got most of the votes every time that binding site *n* was OOB. The proportion of times that *j* is not equal to the true class of *n* averaged over all binding sites is the OOB error estimate. We achieve a misclassification rate of ~12% (133/151 and 152/173 binding sites being classified correctly as stabilized and cleared, respectively), indicating a high classification accuracy of our machine learning approach.

Feature importance: Importance informs about the relevance of each feature in discriminating between binding sites that are stabilized *in vivo* vs. cleared *in vivo*. Importance of individual features is assessed via random permutation of the values of each feature. The difference in the OOB when using the original and the permuted feature is averaged over all trees, resulting in a raw importance score for each variable.

RF parameters: Features exceeding a correlation cutoff of 0.85 were merged. Merging was done separately for each feature class. Two runs of RF with each 20,000 trees were done. The top 30% features of the first run were input to the second run. The resulting mean OOB rate was 12.6%, i.e. 21/18 out of 172/151 cleared/stabilized binding sites were misclassified, respectively. Moreover, control runs (10 repetitions) comparing two sets of randomly picked non-regulated U2AF2 binding sites (|z-score| < 0.5; 70 binding sites each) resulted in an average overlap of 3/100 for the top 100 features. Computation was done with the R package randomForest (v4.6-12) for R 3.3.1 (Liaw and Wiener 2002).

Identification of relevant regulatory groups

In order to identify putative regulators of U2AF2 binding *in vivo*, we considered the top 100 features ranked by importance (**Supplemental Table S4**). To increase their interpretability, *k*-mers were mapped to RBPs as follows: Starting from the available PSSMs (length 6-9 nt), each *k*-mer was positioned in all possible registers (e.g. position 1-6, 2-7, ...) and checked for an PSSM match

with at least 80% of the maximum possible score. This results in a loose assignment of *k*-mers to RBPs. No PSSM information is so far available for FUBP1, yet there is evidence for a TG-rich motif to be recognized by FUBP1 (Miro et al. 2015). Thus, all *k*-mers containing only T and/or G were considered as candidate FUBP1 motifs.

RBPs were combined into regulatory groups to simplify analysis (**Supplemental Table S5**). In total, we defined 13 regulatory groups with the following number of members: *CELF* (4), *ELAV* (2), *FUBP* (1), *HNRNPC* (3), '*other HNRNP*' (16), *MBNL* (3), *PCBP* (4), *PTB* (3), *RBFOX* (3), *RBM* (18), *SF* (other splice factors, 6), '*SR proteins*' (11) and '*other*' (47). For each RBP in a regulatory group, the highest observed importance was taken as representative value. Importance scores of all RBPs in each regulatory group were summarized by the 75% quantile ('majority vote'), and scaled to the maximum observed importance of any feature in the RF analysis. In order to capture the specificity to bind at either stabilized or cleared U2AF2 binding sites, we further assigned a 'purity' score for each RBP. To this end, we counted the number of stabilized and cleared U2AF2 binding sites that harbor a feature of the RBP (either a PSSM match or an associated *k*-mer) within the extended 99-nt window (see above), and calculated the ratio of the difference of binding sites in both groups over the number of binding sites in the larger group. Purity thus reflects the percentage of binding sites unique to one direction of regulation, such that a purity of 1 is achieved when an RBP binds only at either cleared or stabilized binding sites, while a purity of 0.5 indicates that the RBP binds at 2 times more stabilized than cleared binding sites or vice versa (**Fig. 4B**). Purity was first calculated for each RBP, and then summarized for RBPs within a regulatory group by taking the 75% quantile.

For **Fig. 4B**, the number of regulated U2AF2 binding sites associated with a given regulatory group was calculated as follows: For each RBP, the total number of predicted binding sites within the extended 99-nt window was summed over all U2AF2 binding sites that are stabilized\cleared *in vivo*, and then scaled by the size of the respective group. The 75% quantile over all RBPs was taken as summary statistics for each regulatory group. Since a total of 64 *k*-mers are considered as possible FUBP1 motifs, the median rather than the 75% quantile was taken as a summary statistics for the *FUBP* regulatory group. The same approach was followed for calculating the summarized importance and purity scores for this group.

Analysis of *in vitro* iCLIP co-factor assays

In order to facilitate direct comparisons, reads from each *in vitro* iCLIP co-factor replicate were downsampled to 100,000 reads, followed by removal of PCR duplicates by means of random barcodes and spike-in normalization to account for differences in sequencing depth. iCLIP counts were summed up per binding site and represented as 'signal-over-background' (SOB). Background was defined as the 75% quantile of signal on all nucleotides within introns (minus binding sites +/- 2 nt) per *in vitro* transcript. Binding sites with a ratio of

K_d confidence interval boundaries greater than 10 were removed, leaving 420 binding sites. SOB values from replicate experiments were averaged. Only binding sites with a SOB greater than the 10% quantile of the SOB distribution (combined $U2AF2^{RRM12}$ and $U2AF2^{RRM12}$+co-factor samples) were taken into consideration.

For **Fig. 5B**, binding sites were assigned to regulatory categories as follows: (i) Based on the model-based comparison of *in vivo* and *in vitro* U2AF2 binding landscapes (see 'Model-based analysis of *in vivo* regulatory hotspots'), U2AF2 binding sites were classified as not regulated (|z-score| < 0.5), stabilized *in vivo* (z-score > 1) or cleared *in vivo* (z-score < -1). (ii) *in silico* prediction of associated RBP binding sites was done as described for the Random Forests analysis, considering the 99-nt extended binding site region. For each category, the data was centered such that the control group of U2AF2 binding sites with no overlapping RBP motifs had a median $\log_2$ fold change ($\log_2$FC) of 0. Each set was tested against the control group using a two-sided Student's t-test, followed by multiple testing correction (Benjamini-Hochberg). For **Fig. 5A**, binding sites were required to show an SOB greater than the 25% of the SOB distribution as well as an absolute $\log_2$FC > 2 in at least one co-factor experiment. $\log_2$FC values were centered on zero for each co-factor to make data comparable. For **Fig. 7A**, we used all U2AF2 binding sites within 600 nt upstream of the 3' splice site. Note that the preceding introns of *MYL6* exon 6 and *PCBP2* exon 9 are only 304 nt and 510 nt in length. Binding sites that are lowly covered (SOB smaller than the 25% quantile of the SOB distribution) are not shown for the respective condition (indicated by dark gray color) or completely removed if present in less than half of the eleven KD experiments. $\log_2$FC values were centered on zero for each co-factor to make data more easily comparable.

Analysis of *in vivo* U2AF2 binding upon *HNRNPC* knockdown

In order to validate the hnRNPC-mediated regulation in our *in vitro* iCLIP co-factor assay, we compared the results to changes in *in vivo* U2AF2 binding upon *HNRNPC* knockdown (**Fig. 4C,E**). To this end, we used our previously published *in vivo* U2AF2 iCLIP data from control (lujh23a) and *HNRNPC* knockdown (lujh21a) HeLa cells (Zarnack et al. 2013) to calculate SOB values for both conditions. Binding sites which showed a $\log_2$ fold change ($\log_2$FC) in SOB < -1 (control over knockdown) and harbored at least one SOB value in control or knockdown condition that exceeds the mean SOB of all binding sites were defined as 'downregulated upon *HNRNPC* knockdown *in vivo*' (**Fig. 4C**). For **Fig. 4E**, we restricted the analysis to 126 U2AF2 binding sites which are located within 40 nt from an *in vivo* hnRNPC binding site in the regions of the nine *in vitro* transcripts. In addition, binding sites must exceed the 10% quantile of the SOB distribution in all conditions.

GraphProt analysis

A GraphProt sequence model (version 1.1.2, default parameters) (Maticzka et al. 2014) was trained on 20,000 U2AF2 binding sites that were randomly selected

from 458,942 intronic *in vivo* U2AF2 binding sites and 20,000 unbound sites. The 9-nt binding sites were set as viewpoint regions and flanked by 10 nt of non-viewpoint context on either side. The model was then used to score the 438,942 *in vivo* U2AF2 binding sites that were not used for training, a corresponding number of unbound sites, and the 571 U2AF2^{RRM12} binding sites measured by *in vitro* iCLIP. **Supplemental Fig. S1D** compares the distribution of scores for the top 50 *in vivo* U2AF2 binding sites with most crosslink events and the top 50 *in vitro* U2AF2^{RRM12} binding sites with highest affinity within the nine protein-coding *in vitro* transcripts.

Knockdown of RBPs

HeLa cells were grown in 6-well plate until they reached about 25% confluence. For all RBP knockdowns, siRNAs were transfected with Lipofectamine RNAiMax reagent according to manufacturer's instructions. All siRNAs are listed in **Supplemental Table S8**. The cells were grown for 48 h post-transfection and then harvested by scrapping and centrifugation. The cell pellets were stored at -80°C for subsequent experiments.

For the partial *U2AF2* knockdown, confirmation of the knockdown efficiency was done with Western blot. For the detection, we used a monoclonal mouse anti-U2AF2 antibody (Sigma cat. no. U4758) and a monoclonal mouse anti-Actin beta (ACTB) antibody (Sigma cat. no. A5316) as primary antibodies, and anti-mouse IgG HRP-linked antibody (NEB cat no. #7076) as secondary antibody.

For the knockdown of all other RBPs in the context of the *in vivo* alternative splicing quantifications, confirmations were done by measuring mRNA levels with Luminaris HiGreen Low ROX qPCR Master Mix (Thermo Fisher) in a ViiA 7 Real-time PCR system (Thermo Fisher) according to manufacturer's instructions. All primers that were used for the measurements are listed in **Supplemental Table S9**.

Minigene reporter assays

All minigene reporters were constructed by using pCDNA5 backbone via ligation of a 2,727 bp insert containing exons 9-11 of *PTBP2* (Chr1, 96804170 - 96806896 nt). Mutations introduced to different constructs are listed in **Supplemental Table S6**. Mutant constructs were generated by using the Q5® Site-Directed Mutagenesis Kit (NEB) according to manufacturer's instructions. All primers used for the minigene construction are listed on **Supplemental Table S9**. The *FUBP1* KD and the *PTBP1/2* double-KD were performed for 48 h as described above. The media were discarded and the cells were further transfected with 2 µg of different minigene constructs (wild type and mutant variants). Cells were harvested on the next day and total RNA was extracted with RNeasy Plus Mini Kit. cDNAs were synthesized with Revert Aid First Strand cDNA Synthesis by using oligo(dT)$_{18}$ primer. The resulting cDNAs were amplified with up to 25 cycles with One Taq polymerase (NEB), and the PCR products were visualized in 2200 Tape station system with D1000 DNA screen tape kit

(Agilent) to obtain the molar ratio of each splicing product. All primers used in these experiments are listed in **Supplemental Table S9**. The relative inclusion ('percent spliced in', PSI) in each sample was calculated with the following formula:

$$PSI = \frac{molar\ conc.of\ inclusion\ product}{molar\ conc.of\ inclusion\ product + \ molar\ conc.of\ skipping\ product}$$

In vivo splicing assays

Splicing assays were done by monitoring inclusion of four different alternative exons from *PTBP2*, *MYL6, CD55,* and *PCBP2* via RT-PCR under control conditions and knockdowns of twelve different RBPs (CELF6, ELAVL1, FUBP1, HNRNPC, KHDRBS1, MBNL1, PCBP1, PTBP1/2, RBM24, RBM41 and SNRPA). Total RNA was extracted 48 h post-transfection with RNeasy Plus Mini Kit (Qiagen), and cDNAs were synthesized with Revert Aid First Strand cDNA Synthesis by using oligo$(dT)_{18}$ primer. The resulting cDNAs were amplified with up to 35 cycles with One Taq polymerase (NEB), and the PCR products were visualized in 2200 Tape station system. PSI values for each sample were calculated as described above. All primers used in these experiments are listed in **Supplemental Table S9**.

Supplemental references

Agrawal AA, Salsi E, Chatrikhi R, Henderson S, Jenkins JL, Green MR, Ermolenko DN, Kielkopf CL. 2016. An extended U2AF65‐RNA-binding domain recognizes the 3' splice site signal. *Nat Commun* **7**: 10950.

Bernhart SH, Hofacker IL, Stadler PF. 2006. Local RNA base pairing probabilities in large sequences. *Bioinformatics* **22**: 614-615.

Breiman L. 2001. Random Forests. *Machine Learning* **45**: 5-32.

Bullard JH, Purdom E, Hansen KD, Dudoit S. 2010. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* **11**: 94.

Corvelo A, Hallegger M, Smith CWJ, Eyras E. 2010. Genome-Wide Association between Branch Point Properties and Alternative Splicing. *PLoS Comput Biol* **6**: e1001016.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15-21.

Dodt M, Roehr J, Ahmed R, Dieterich C. 2012. FLEXBAR—Flexible Barcode and Adapter Processing for Next-Generation Sequencing Platforms. *Biology* **1**: 895.

Goutelle S, Maurin M, Rougier F, Barbaut X, Bourguignon L, Ducher M, Maire P. 2008. The Hill equation: a review of its capabilities in pharmacological modelling. *Fundam Clin Pharmacol* **22**: 633-648.

Guo Y, Dai Y, Yu H, Zhao S, Samuels DC, Shyr Y. 2017. Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis. *Genomics* **109**: 83-90.

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S et al. 2012. GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res* **22**: 1760-1774.

Huppertz I, Attig J, D'Ambrogio A, Easton LE, Sibley CR, Sugimoto Y, Tajnik M, König J, Ule J. 2014. iCLIP: Protein–RNA interactions at nucleotide resolution. *Methods* **65**: 274-287.

Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. 2010. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* **26**: 2204-2207.

Kucukural A, Özadam H, Singh G, Moore MJ, Cenik C. 2013. ASPeak: an abundance sensitive peak detection algorithm for RIP-Seq. *Bioinformatics* **29**: 2485-2486.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078-2079.

Liaw A, Wiener M. 2002. Classification and Regression by Random Forest. *R News* **2**: 18-22.

Maticzka D, Lange SJ, Costa F, Backofen R. 2014. GraphProt: modeling binding preferences of RNA-binding proteins. *Genome Biol* **15**: R17.

Mercer TR, Clark MB, Andersen SB, Brunck ME, Haerty W, Crawford J, Taft RJ, Nielsen LK, Dinger ME, Mattick JS. 2015. Genome-wide discovery of human splicing branchpoints. *Genome Res* **25**: 290-303.

Miro J, Laaref AM, Rofidal V, Lagrafeuille R, Hem S, Thorel D, Méchin D, Mamchaoui K, Mouly V, Claustres M et al. 2015. FUBP1: a new protagonist in splicing regulation of the DMD gene. *Nucleic Acids Res* **43**: 2378-2389.

R Core Team. 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria.

Raue A, Kreutz C, Maiwald T, Bachmann J, Schilling M, Klingmüller U, Timmer J. 2009. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics* **25**: 1923-1929.

Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, Li X, Gueroussov S, Albu M, Zheng H, Yang A et al. 2013. A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **499**: 172-177.

Schwanhäusser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M. 2011. Global quantification of mammalian gene expression control. *Nature* **473**: 337-342.

Sutandy FXR, Hildebrandt A, König J. 2016. Profiling the Binding Sites of RNA-Binding Proteins with Nucleotide Resolution Using iCLIP. In *Post-Transcriptional Gene Regulation*, doi:10.1007/978-1-4939-3067-8_11 (ed. E Dassi), pp. 175-195. Springer New York, New York, NY.

Yeo G, Burge CB. 2004. Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals. *J Computat Biol* **11**: 377-394.

Zarnack K, König J, Tajnik M, Martincorena I, Eustermann S, Stévant I, Reyes A, Anders S, Luscombe NM, Ule J. 2013. Direct Competition between hnRNP C and U2AF65 Protects the Transcriptome from the Exonization of *Alu* Elements. *Cell* **152**: 453-466.