Supplementary computational methods

PREPARATION AND PRE-PROCESSING FOR SEQUENCING DATA

Genomic reference

We used the hg18 reference genome throughout this project. Chromosome files were downloaded from the UCSC genome browser website [1].

Preparation of the transcriptome

We prepared a modified transcriptome from RefSeq [2]. The RefSeq exon annotations were acquired from the UCSC genome browser website, as were lookup tables linking each RefSeq identifier to an official gene symbol. After associating each transcript to a gene symbol, we removed any transcript that had exons on multiple chromosomes, or on both positive and negative strands. We then constructed our reference by collapsing all isoforms for a given gene into one super-transcript. We defined the super-transcript boundaries as the 5' end of the 5'-most exon and the 3' end of the 3'-most.

Preparation of regions

We defined the regions of 3' UTR, 5' UTR, coding sequence and introns by collapsing all such labelled regions associated with RefSeq identifiers obtained from the UCSC genome browser. In cases where there is ambiguity, we use the following order of preference: 3' UTR, 5' UTR, coding, and intron. We consider any region of the genome not falling on of these four categories as intergenic.

STATISTICAL ANALYSES

Corrections for multiple hypothesis testing

Unless otherwise stated, all *p*-values reported in this manuscript have been corrected for multiple hypothesis testing using the method of Benjamini and Hochberg [3].

READ MAPPING AND PRELIMINARY PROCESSING FOR SEQUENCING DATA

Mapping of data

We constructed two masked versions of the hg18 chromosomes. In the first, we masked all regions that are non-exonic (i.e. not 5'UTR, 3'UTR or coding) with Ns; we call this the *exonic-masked genome*. In the second, we do the same, but allow intronic regions; we call this the *transcript-masked genome*. We also construct a junction database from all of the exon-exon junctions for each super-transcript. We used RMAP [4] to map the iCLIP data to the transcript-masked genome and the junction database, while we mapped the RNA-Seq data to the exon-masked genome.

Assignment of reads to regions, exons and genes

In the case of the RNA-Seq data, we count the number of reads mapping with their first mapped position within each exon of our super-transcript reference. The count of reads within a gene is then simply the sum of all read counts for the exons in its super-transcript.

IDENTIFICATION AND ANALYSIS OF MSI1 ICLIP SITES AND TARGETS

Peak calling in iCLIP data

We call peaks in iCLIP data using Piranha [5], using a bin size of 1nt. We consider significant peaks to be those that have a corrected p-value less than 0.05.

Target identification from iCLIP data

Target genes are defined to be those with at least a single site in 3' UTR or 5' UTR in at least two of the three iCLIP replicates.

Analysis of motif enrichment in iCLIP data

There is a bias in iCLIP data towards cross-linking at triple-uracil sequences [6]. We observed a strong enrichment for these tri-nucleotides around our identified iCLIP sites. To ameliorate this when trying to determine enriched sequences around Msi1 binding sites, we computed an expected number of occurrences for each tri-nucleotide. To do this, we identified the top 1000 most enriched locations from a set of public iCLIP datasets using Piranha [5] – dataset details below – and counted the number of times each possible tri-nucleotide occurs within +/- 2nt of the cross-link location as follows:

$$e_t = \frac{\sum_{j=1}^{n}\sum_{k=1}^{1000}\sum_{l=-2}^{2} I(S_{j,k,l},t)}{5 \times 1000 \times n},$$

where e_i is the normalized count for tri-nucleotide t, computed over n total datasets and $I(S_{j,k,l},t)$ is the indicator function that returns 1 if the tri-nucleotide starting at position l of sequence k from dataset j is equal to t, and 0 otherwise. We compute the observed counts from our three Msi1 iCLIP replicates analogously:

$$o_t = \frac{\sum_{j=1}^{3}\sum_{k=1}^{m_j}\sum_{l=-2}^{2} I(S_{j,k,l},t)}{5 \times 1000 \times n},$$

where m_j is the number of significant iCLIP sites reported for replicate j. We use the following public iCLIP datasets to compute the expected values:

RBP	Citation
HuR	[7]
TIAL	[8]
TIA1	[8]
hnRNPC	[9]
TDP43	[10]

Calculating paired and unpaired UAGs around Msi1 iCLIP sites

We extracted the 500bp region centred on each iCLIP site, and the 500bp region flanking upstream and downstream. We split each of these into 20bp bins and within each bin we count the number of times GUAG occurs, and the number of times [C/T/A]UAG occurs. We define the number of pairs as two times the minimum of these two counts, and the number of orphans as the difference between these two counts. By overlaying the flanking bins with those centred on the iCLIP sites, we construct the following table for each pair of bins:

	iCLIP	Flanking
Paired	T11	T12
Orphaned	T21	T22

Where

- T11 is the number of pairs in the 'iCLIP bin'
- T12 is the number of pairs in the 'flanking bin'
- T21 is the number of orphans in the iCLIP bin
- T22 is the number of orphans in the flanking bin

We then perform a Fisher's exact test on each bin to determine an odds ratio and p-value.

Calculation of base-pair probability

To calculate the base-pair probability, we selected a subset of the significant iCLIP sites such that each was at least 200 nucleotides from the closest other site. We then computed the base-pair probabilities for a window of 100 nucleotides around each selected iCLIP site using a modified version of the RNA Vienna package [11]. Then we took the average base pair probability for each location in the window over all sequences. Since the folding algorithm favours the ends of the sequences to be single stranded to obtain more stable structures, the base pairing probabilities of the ends of sequences are biased towards zero. In order to fix this problem, although we folded a window of 100 nucleotides around the peaks, we only took the base pairing probability of the middle 80 nucleotides into account, and dropped 10 nucleotides from each end of all sequences.

Calculation of secondary structure context

We folded a window of 100 nucleotides around each identified iCLIP site using Mfold [12] to obtain the minimum free energy structure for these sequences. Then we parsed the resulting secondary structure using covariance models introduced in (Eddy and Durbin, 1994) to assign a context to each nucleotide in the 100nt region: bulge-loop, internal loop, hairpin loop, multi-branch loop or dangling ends. We then counted the number of times each specific type of loop is observed. For each loop type then we have four counts: the number of times that loop type is seen at the iCLIP site (including +/- 1nt of the iCLIP site), the number of times it is seen in flanking regions, the number of times a different loop type is seen at iCLIP sites and the number of times a different loop type is seen in flanking regions. We then construct a contingency table; for example, the following for hairpin loops:

	Within binding site (+/- 1nt)	Outside binding site
Hairpin	T11	T12
Not hairpin	T21	T22

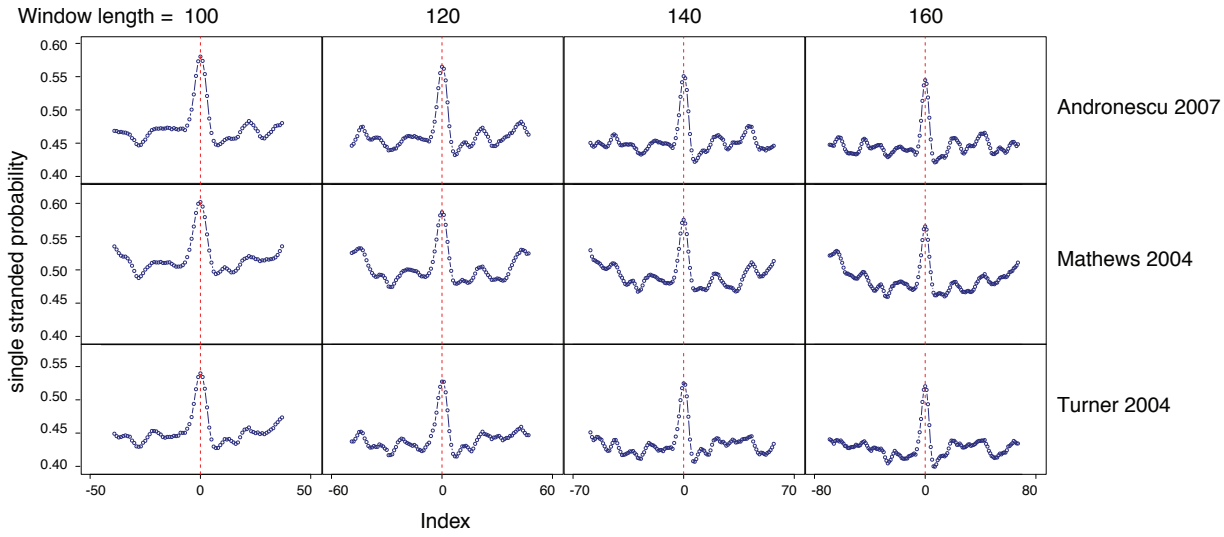
where:

- T11 is the number of times that peaks appear in a Hairpin loop region,
- T12 is the total number of observed hairpin loop regions minus T11,
- T21 is the total number of peaks minus T11 and
- T22 is the total number of single stranded regions that occur in any type of loop other than hairpin loop minus T21.

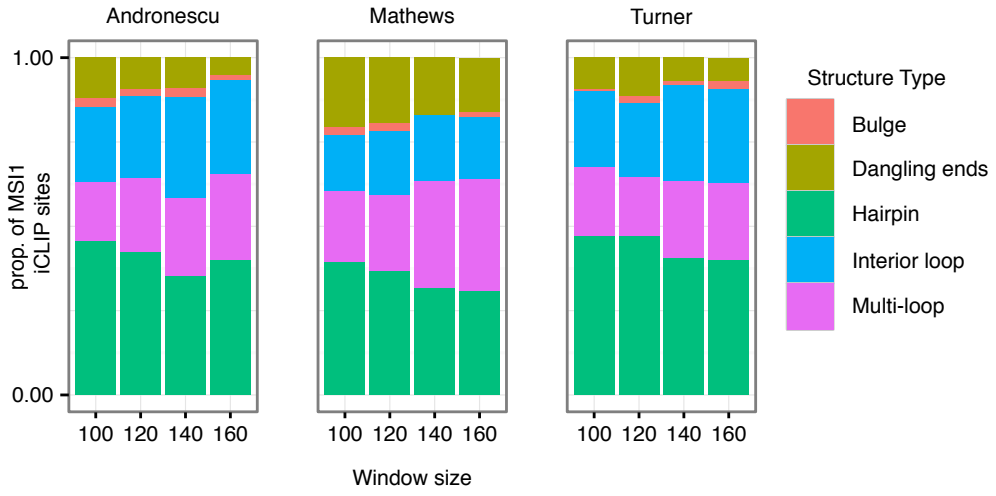
The p-value and odds-ratio resulting from performing a Fisher's exact test illustrates the significance of preference for Musashi to bind to a specific type of single stranded loop.

Discussion on different energy models and parameters for RNA structure prediction

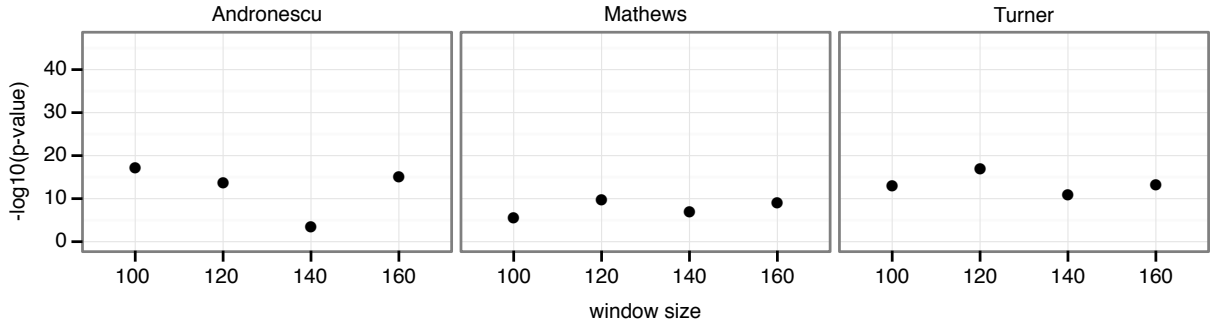
To ensure our prediction of RNA secondary structure was not heavily dependent upon the energy model or parameters used, we explored a number of alternatives. Firstly, we examined larger window sizes and found that the peak in ssRNA probability is observed in all of those examined, regardless of window size or energy models used:



We did the same for the proportion of MSI1 sites that are contained within particular secondary structures, and found these also to be relatively invariant in the face of energy model or window size used:



Notably, the tendency for MSI1 binding sites to be enriched in hairpin loops remains significant under all of these models and window sizes:



Other recent work by Fukunaka et al. has also found that considering long-range interactions increases computational cost, but does not largely impact the assignment of locations within the RNA to secondary structure types, such as we have done here [13].

Although the choice of model and window size can impact the predictions for individual targets, our analysis involves many targets, and our aim is to describe the general properties of MSI1 binding, rather than the structure of any individual target. Given the above observations, the conclusions drawn (namely that MSI1 binding sites exist mainly in single-stranded RNA that folds into hairpin loops), are robust to changes in parameters for structure predictions.

ANALYSIS OF CHANGES IN MRNA LEVELS AND EXON-INCLUSION RATES

Identification of exons with changed inclusion ratio from RNA-Seq data

To identify those exons with changes in inclusion ratios between the control and the knockdown condition, we paired each control replicate with its corresponding KD replicate and for each gene we constructed a two-by-two contingency table as follows:

	Total RNA	Msi1 KD
Within target exon	T11	T12
Within other exons (same gene)	T21	T22

where:

- T11 is the number of reads mapping into the exon of interest in the total RNA sample
- T12 is the number of reads mapping into the exon of interest in the Msi1 KD sample
- T21 is the number of reads mapping into the gene, but outside the exon of interest in the total RNA sample
- T22 is the number of reads mapping into the gene, but outside the exon of interest in the Msi1 KD sample

We then performed Fisher's exact test on this table to compute an odds-ratio and a two-tailed p-value for the significance of the change from an odds-ratio of 1. We perform this comparison for the 3 pairs of control/KD samples. We considered significant changes to be those with a corrected *p*-value less than 0.01 and an odds ratio greater than 1.5 (increase) or less than 0.66 (decrease) in three replicates.

Identification of differentially expressed genes from RNA-Seq data

We used EdgeR to identify differentially expressed genes (ref). Three replicates of control total RNA and three replicates of Msi1 knockdown RNA-Seq were used to construct a 6xN matrix of gene-read counts (where N is the number of genes in our reference) that was provided to EdgeR. We considered those genes reported as having a corrected *p*-value less than 0.05 as showing significant changes in expression. We divided this set into two lists: those genes that were up-regulated on Msi1 KD (i.e. had a greater normalized read-count in the Msi1 KD condition) and those genes that were down-regulated on Msi1 KD (i.e. had a lesser normalised read-count in the Msi1 KD condition)

GENE ONTOLOGY AND PATHWAY ANALYSIS

Enrichment of biological processes and KEGG pathways was performed with DAVID [14]. For all analyses we used a background gene set constructed by finding those genes that had at least a single read in all replicates of our total-RNA RNA-Seq samples. Protein-protein interactions were extracted from iRefIndex [15].

REFERENCES

1. Dreszer, T.R., et al., *The UCSC Genome Browser database: extensions and updates 2011.* Nucleic Acids Research, 2012. **40**(D1): p. D918-D923.
2. Pruitt, K.D., et al., *NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy.* Nucleic Acids Research, 2012. **40**(D1): p. D130-D135.
3. Hochberg, Y.B.a.Y., *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.* Journal of the Royal Statistical Society. Series B (Methodological), 1995. **57**(1): p. 289-300.
4. Smith, A.D., et al., *Updates to the RMAP short-read mapping software.* Bioinformatics, 2009. **25**(21): p. 2841-2842.
5. Uren, P.J., et al., *Site identification in high-throughput RNA–protein interaction data.* Bioinformatics, 2012. **28**(23): p. 3013-3020.
6. Sugimoto, Y., et al., *Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions.* Genome Biology, 2012. **13**(8): p. R67.
7. Uren, P.J., et al., *Genomic Analyses of the RNA-binding Protein Hu Antigen R (HuR) Identify a Complex Network of Target Genes and Novel Characteristics of Its Binding Sites.* Journal of Biological Chemistry, 2011. **286**(43): p. 37063-37066.
8. Wang, Z., et al., *iCLIP Predicts the Dual Splicing Effects of TIA-RNA Interactions.* PLoS Biol, 2010. **8**(10): p. e1000530.
9. Konig, J., et al., *iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution.* Nat Struct Mol Biol, 2010. **17**(7): p. 909-915.
10. Tollervey, J.R., et al., *Characterizing the RNA targets and position-dependent splicing regulation by TDP-43.* Nat Neurosci, 2011. **14**(4): p. 452-458.
11. Hofacker, I.L., *Vienna RNA secondary structure server.* Nucleic Acids Research, 2003. **31**(13): p. 3429-3431.
12. Zuker, M., *Mfold web server for nucleic acid folding and hybridization prediction.* Nucleic Acids Research, 2003. **31**(13): p. 3406-3415.
13. Fukunaga, T., et al., *CapR: revealing structural specificities of RNA-binding protein target recognition using CLIP-seq data.* Genome Biology, 2014. **15**(1): p. R16.
14. Huang, D.W., B.T. Sherman, and R.A. Lempicki, *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.* Nat. Protocols, 2008. **4**(1): p. 44-57.
15. Razick, S., G. Magklaras, and I. Donaldson, *iRefIndex: A consolidated protein interaction database with provenance.* BMC Bioinformatics, 2008. **9**(1): p. 405.