Supplementary material for: "Mixture models and wavelet transforms reveal high confidence RNA-protein interaction sites in MOV10 PAR-CLIP data"

Cem Sievers[1], Tommy Schlumpf[1], Ritwick Sawarkar[1], Federico Comoglio[1] and Renato Paro[1,2]

[1]Department of Biosystems Science and Engineering, Swiss Federal Institute of Technology Zurich, Mattenstrasse 26, 4058 Basel, Switzerland

[2]Faculty of Science, University of Basel, Klingelbergstrasse 50, 4056 Basel, Switzerland

Supplementary Information

Text S1: Mixture models and Bayesian inference Mixture models represent linear combinations of distributions used to represent complex probability density functions (pdfs). These kind of models are particularly useful when the data generating distribution exhibits a hidden structure, which cannot be captured by a single pdf. This applies to examples where data instances are derived from two or more different distributions. In such a case the identity of the distribution responsible for generating a particular data instance can be represented as a hidden variable giving rise to a hierarchical model. The application of a mixture model seems natural in the analysis of PAR-CLIP data, because it can be assumed that all data instances (all observable T to C transitions) are derived from one out of two possible components. The first component encompasses all non-experimental causes, i.e. all transitions which are not derived by experimental induction and the second component accounts for experimentally induced observations. Hence, the entirety of all T to C transitions is the result of a data generating process in which a component is randomly chosen to generate a given data instance. Since we cannot observe which component actually generated the data instance, the challenge is to estimate the probability that either of the two components was responsible. For this reason the two pdfs are estimated using a Bayesian inference framework. In the Bayesian setting parameters that govern a model are considered to be random variables rather than constants fixed to a specific value. Hence, the parameters are distributed according to a pdf. One main goal in Bayesian inference is the estimation of this distribution by integrating the information of the observed data instances. For this purpose a prior distribution is chosen in order to reflect the prior belief about the parameters. In the method described in this work a uniform prior was chosen since there was no reason to assume that specific parameter values are more likely than others. In order to obtain the posterior pdf of a given parameter the prior is multiplied with the likelihood function, which accounts for the observed data instances and thereby changes the belief of the parameter values initially set to be uniform. Hence, the extent to which the belief in the parameters changes depends on the observations that were made. In cases of highly informative observations the posterior pdf will be closely centered around a specific value reflecting the extent to which the parameters are believed to take certain values. This is a major difference with respect to the maximum likelihood estimation, which returns a single value only. Since genomic positions used for the parameter estimation are not of equal coverage and hence convey different amount of information, a Bayesian approach was chosen in this work. Regions of higher coverage comprise more information regarding the position-dependent parameters, resulting in pdfs that are less spread around the given values. The Bayesian framework naturally accounts for the heterogenous information content in the data. For further information on the subject see [1, 2].

Text S2: Short introduction to wavelets analysis In many applications the representation of a signal that evolves in time (or space) in terms of its frequency components is desirable. The first mathematical framework applicable to study signals in either time or frequency was developed to become what is known as Fourier analysis. The use of the Fourier transform or its inverse allows the signal to be transformed between the different domains. However, the local frequency content of a signal cannot be captured in this way. The wavelet analysis provides the means to represent signals in both time and frequency (also called scale) domain thereby allowing to study local frequency properties. The wavelet transform of a signal can be understood as the inner product of the signal with a family of wavelets that are parameterized by shift and scale parameter. The shift parameter determines the location of the wavelet in time whereas the scale determines the length of its support and hence is equivalent to the frequency. The different wavelet coefficients (representing the inner product of the signal with a wavelet) therefore describe the different frequency components of the signal at different time points. Using the time-frequency representation of the signal, useful properties such as local signal-to-noise ratios can be computed. One important application is peak-calling where the goal is the detection of peaks corresponding to large amplitudes of the signal, possibly located within regions of high noise. Computing local signal-to-noise ratios can be achieved since, for a fixed time, the signal is represented by wavelet coefficients, which correspond to different scales and therefore provide more or less local approximations of the signal. Large scale wavelet coefficients correspond to rather global approximations and can be used to represent the local noise in the signal, whereas small scales result in local approximations. The comparison of the magnitudes of wavelet coefficients corresponding to similar time but different scales therefore results in local signal-to-noise ratio estimates. One major advantage of this approach is the consideration of local noise approximation. This strategy is favorable in cases where the noise is not constant and hence global noise estimates may lead to poor local approximations, resulting in suboptimal peak detection. For further information on the subject see [3].

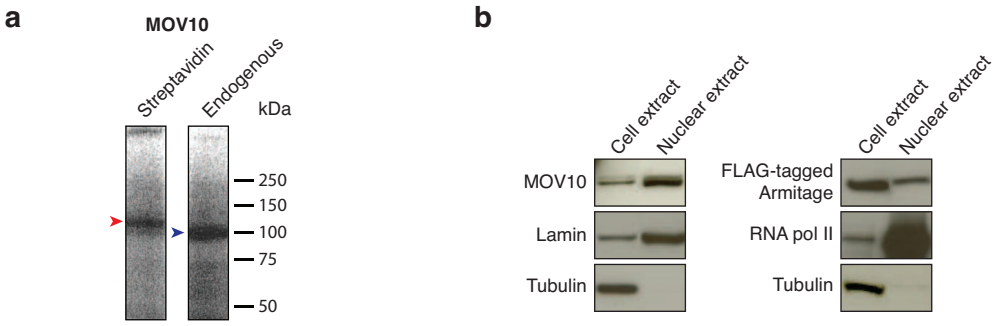

Figure S1 **a**) Autoradiograph obtained by UV-mediated crosslinking of HEK293 cells followed by MOV10 immunoprecipitation and radio labeling of the RNA molecules. Left lane: Streptavidin-HA-tagged MOV10, immunoprecipitated using Streptavidin-tag. Right lane: endogeneous MOV10, immunoprecipitated using MOV10 antibody. Radioactive band corresponding to MOV10 size (100 kDa), the size shift corresponds to the tag mass. Protein mass is indicated in kDa. **b**) Fractionation experiments using total cell and nuclear extract. Experiments were performed in HEK293 cells, expressing Streptavidin-HA-tagged MOV10 (left panel) and in *Drosophila* S2 DRSC cells expressing FLAG-tagged Armitage (homologue of MOV10, right panel). MOV10 immunoblotting was done using endogenous antibody, detecting endogenous and tagged MOV10 as indicated by the double band. Tagged Armitage was detected using FLAG-specific antibodies. Lamin and RNA polymerase II serve as marker for the nuclear compartment. Tubulin serves as marker for the cytosolic compartment.

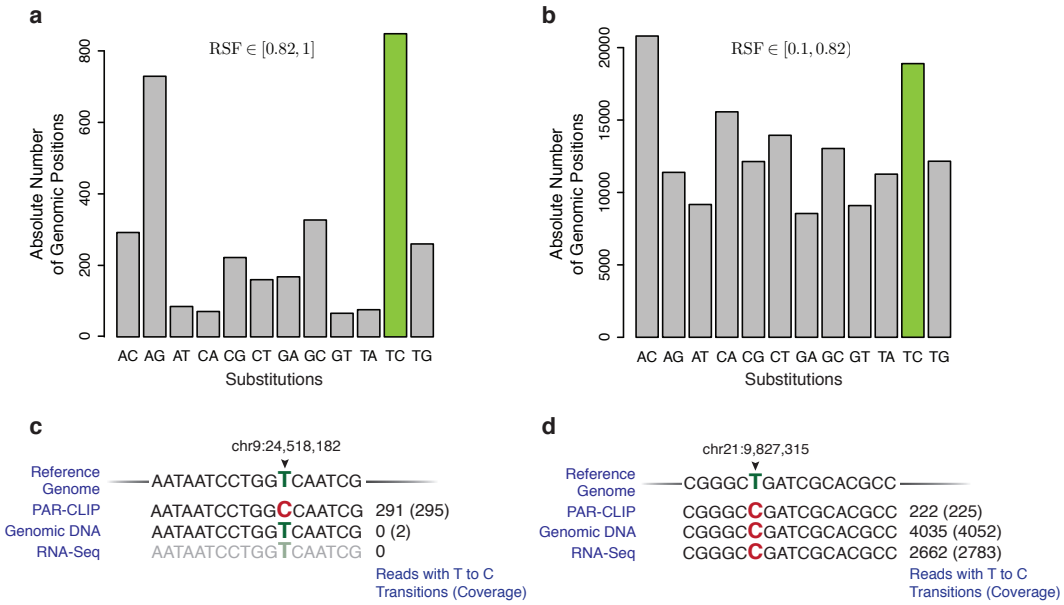

Figure S2: **a**,**b**) Absolute number of genomic sites exhibiting specified substitutions within the RSF intervals specified in the figure. Reads were obtained from nuclear RNA-seq (+4-SU) experiments, no UV-mediated crosslink was performed, only genomic positions of minimum coverage 20 were considered. **c**) Example high-TC site likely to be the result of external RNA contamination. Genomic position is indicated on top. Total Numbers of all aligned reads (in brackets) and observed T to C transitions are shown on the right. Experimental read sources are indicated on the left. PAR-CLIP: reads obtained from MOV10 PAR-CLIP experiments. Genomic DNA: reads obtained from multiple RNA-Seq and ChIP-Seq experiments performed in the same cell line (not published). RNA-Seq: reads obtained from nuclear RNA-Seq control experiments (Methods). **d**) Example high-TC site likely to be the result of a cell-type specific SNP. Annotation same as in Figure S1c.

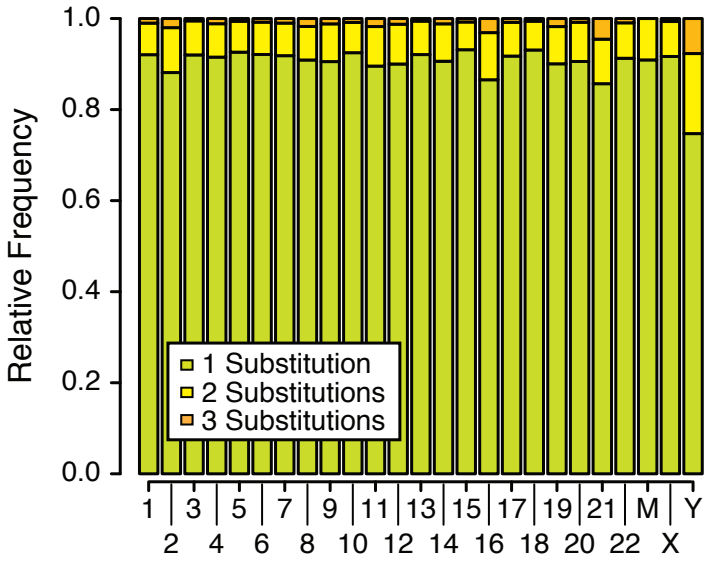

Figure S3: The number of distinct substitutions at chromosomal positions with at least one substitution. Shown are the relative occurrences of chromosomal positions with either 1, 2 or 3 distinct substitutions. Genomic positions without substitutions were not considered, a minimum coverage of 20 was required. Labels on horizontal axis indicate the chromosome names.

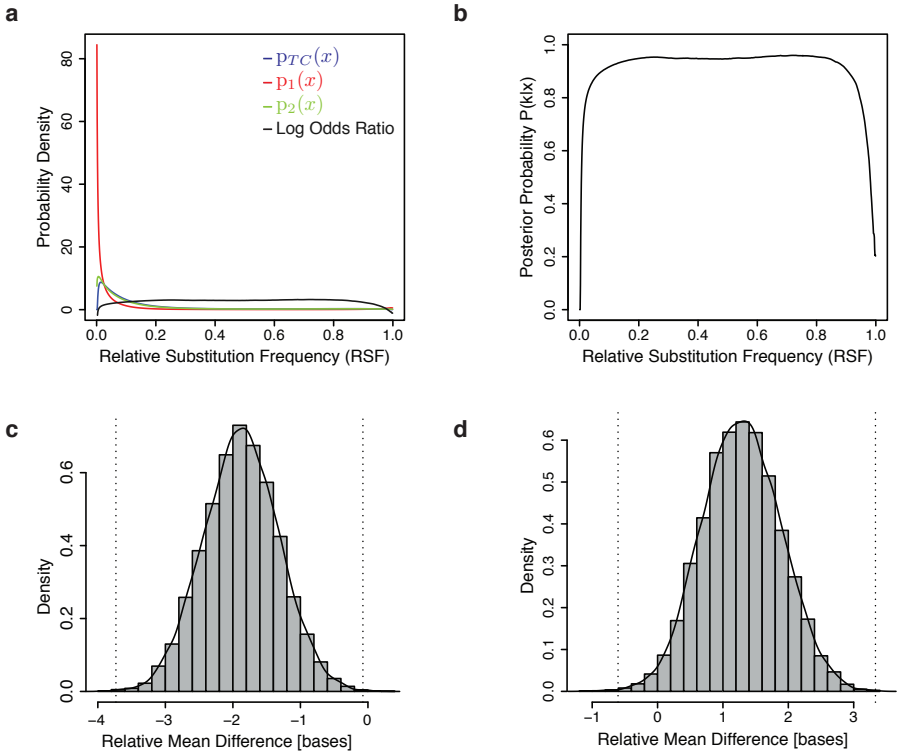

Figure S4: **a**) Probability densities obtained from pooled replicate AGO2 PAR-CLIP data sets. **b**) Posterior probability of an observation being experimentally induced. The interval determining high significance interaction sites was chosen to be $[0.2, 0.8]$. **c**) Bootstrap distribution of mean start difference of overlapping AGO2 and MOV10 wavCluster. 10000 bootstrap samples were computed. Dashed vertical lines (-3.73, -0.07) indicate bootstrap confidence intervals considering a significance level of 0.001. Solid line represents the kernel density estimate. **d**) Bootstrap distribution of mean end difference of overlapping AGO2 and MOV10 wavCluster. 10000 bootstrap samples were computed. Dashed vertical lines (-0.61, 3.33) indicate bootstrap confidence intervals considering a significance level of 0.001. Solid line represents the kernel density estimate.

Table S1: GO enrichment analysis for process and binding terms. Different gene sets bound by AGO2, MOV10 and AGO2 and MOV10 were considered for the analysis. P-values are indicated in parenthesis.

Gene set bound by:	**AGO2**	**MOV10**	**AGO2 and MOV10**
Process Enrichment	Regulation of transcription (e-44)	RNA splicing (e-26)	Regulation of transcription (e-23)
Process Enrichment	Metabolic control (e-24)	processing and metabolism (e-26)	RNA Metabolism (e-18)
Process Enrichment	Cell cycle (e-22)	Cell cycle (e-22)	Cell cycle (e-17)
Binding Enrichment	DNA binding (e-22)	RNA binding (e-40)	RNA binding (e-27)

Table S2: Most abundantly expressed miRNAs in HEK293 cells. Second and third column indicate presence of the according wavCluster. The arm was of the miRNA gene was ignored. (+) indicates cases exhibiting only one wavCluster in different loci giving raise to the same mature miRNA. See [4] for a description of nomenclature.

miRNA	**AGO2 wavCluster**	**MOV10 wavCluster**
hsa-miR-106b	+	−
hsa-miR-10a	+	−
hsa-miR-1307	+	−
hsa-miR-140	+	−
hsa-miR-15a	+	−
hsa-miR-15b	+	−
hsa-miR-16	+	−
hsa-miR-17	+	+
hsa-miR-185	+	−
hsa-miR-186	−	−
hsa-miR-18a	−	−
hsa-miR-191	+	+
hsa-miR-196a	+	−
hsa-miR-196b	+	−
hsa-miR-19a	+	+
hsa-miR-19b	+	−
hsa-miR-20a	+	−
hsa-miR-221	+	−
hsa-miR-222	+	−
hsa-miR-24	+	−
hsa-miR-29b	−	−
hsa-miR-30a	+	−
hsa-miR-30d	+	−
hsa-miR-30e	+	−
hsa-miR-32	+	−
hsa-miR-324	+	−
hsa-miR-339	+	−
hsa-miR-33b	+	−
hsa-miR-423	+	−
hsa-miR-615	+	−
hsa-miR-92a	+	(+) -2 only
hsa-miR-93	+	−
hsa-miR-99a	+	−
total	30 out of 33	4 out of 33

References

[1] Bishop,C. (2007) *Pattern Recognition and Machine Learning*. Springer, 233 Spring Street, New York, NY 10013 USA

[2] Hastie, T., Tibshirani, R. & Friedman, J., (2007) *The Elements of Statistical Learning*. Springer, 233 Spring Street, New York, NY 10013 USA

[3] Daubechies,I. (1992) *Ten Lectures on Wavelets*. SIAM, 3600 Market Street, 6th Floor Philadelphia, PA 19104-2688 USA

[4] Griffiths-Jones,S., Grocock,R.J., van Dongen,S., Bateman,A. & Enright,A.J. (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res*, **34**, D140D144