Published online 22 October 2012 *Nucleic Acids Research, 2013, Vol. 41, No. 1* ***e31***
doi:10.1093/nar/gks968

X-inactivation: quantitative predictions of protein interactions in the *Xist* network

Federico Agostini[1,2], Davide Cirillo[1,2], Benedetta Bolognesi[1,2] and Gian Gaetano Tartaglia[1,2,*]

[1]Centre for Genomic Regulation (CRG), Dr. Aiguader 88, 08003 Barcelona and [2]Universidat Pompeu Fabra (UPF), 08003 Barcelona, Spain

Received May 30, 2012; Revised September 21, 2012; Accepted September 25, 2012

ABSTRACT

The transcriptional silencing of one of the female X-chromosomes is a finely regulated process that requires accumulation in *cis* of the long non-coding RNA X-inactive-specific transcript (*Xist*) followed by a series of epigenetic modifications. Little is known about the molecular machinery regulating initiation and maintenance of chromosomal silencing. Here, we introduce a new version of our algorithm catRAPID to investigate *Xist* associations with a number of proteins involved in epigenetic regulation, nuclear scaffolding, transcription and splicing processes. Our method correctly identifies binding regions and affinities of protein interactions, providing a powerful theoretical framework for the study of X-chromosome inactivation and other events mediated by ribonucleoprotein associations.

INTRODUCTION

X-chromosome inactivation (XCI) is a highly regulated process that involves the transcriptional silencing of one of the female X-chromosomes (1). The silencing process is mainly attributable to the long non-coding RNA X-inactive-specific transcript (*Xist*) transcribed from the *Xist* gene located on the XCI inactivation centre (1). *Xist*-mediated X-inactivation involves two distinct phases: initiation and maintenance. First, *Xist* transcript coats *in cis* the entire X-chromosome triggering transcriptional silencing (2). Subsequently, stabilization of the repressed state is facilitated by a number of epigenetic processes, such as DNA methylation and chromatin modifications mediated by the Polycomb group (PcG) proteins (3). Notably, *Xist* is regulated in *cis* by its antisense partner *Tsix* (4), which also interacts with PcG proteins (5).

Using an inducible expression system in mouse embryonic stem cells, Wutz *et al.* (6) identified a number of *Xist* domains associated with chromatin localization. Interestingly, these domains do not contain sequence or structural motifs and could be low-affinity protein-binding sites (6). In contrast to the poorly defined sequence properties associated with RNA localization, the 5′-repeat region A (RepA) represents a structured domain involved in X-chromosome silencing (6). Secondary structure predictions indicate that RepA folds in two stem loops of ∼200 nt containing a number of repeats (6,7).

To date, the precise mechanisms underlying localization and confinement of *Xist* onto the X-chromosome as well as the molecular details of the silencing process remain poorly understood. Recent experiments suggest that: (i) alternative splicing factor SFRS1 regulates *Xist* processing (8); (ii) transcriptional repressor Ying and Yang (YY1) tethers *Xist* onto the X-chromosome (9); (iii) the RNA-binding domains of scaffold attachment factor SAF-A bind to *Xist*-inducing chromatin reorganization (10) and (iv) the special AT-rich sequence-binding protein SATB1 co-localizes with *Xist* in the nucleus (11). Yet, due to the limited amount of experimental evidence, the challenge of identifying protein–RNA interactions associated with XCI still stands (11).

Here, we use our theoretical framework, catRAPID, to investigate *Xist* interactions with a number of epigenetic modifiers as well as transcription and splicing factors (12). Our approach exploits physicochemical properties of nucleotide and amino acid chains such as secondary structure, hydrogen bonding and van der Waals' propensities to predict protein–RNA associations with a confidence of 78% or higher (12). In the original implementation of the method, we calculated interactions with transcripts <3 kb ('Materials and Methods' section) (12). In order to investigate *Xist*, which is 16–19 kb long and represents the largest non-coding transcript with known function, we developed an extension of the algorithm. In addition to the fine calculation of protein–RNA interactions (interaction propensity), we present here an algorithm to estimate the specificity of associations (interaction strength) and a method to identify binding regions in transcripts (interaction fragments). These new developments are introduced to facilitate the characterization of protein interactions with long non-coding RNA and guide future experimental design. Notably, the new versions of the method do not require introduction of fitting

[*]To whom correspondence should be addressed. Tel: +34 93 316 01 16; Fax: +34 93 396 99 83; Email: gian.tartaglia@cgr.es

parameters and represent a conceptual and methodological advance to study ribonucleoprotein associations. A new version of our web servers is released at http://tartaglialab.crg.cat/.

MATERIALS AND METHODS

Interaction propensity

We use the catRAPID method to predict protein–RNA interactions (12). In catRAPID, the contributions of secondary structure, hydrogen bonding and van der Waals' are combined together into the 'interaction profile':

$$|\Phi_x\rangle = \alpha_S|S_x\rangle + \alpha_H|H_x\rangle + \alpha_W|W_x\rangle \tag{1}$$

In Equation (1), $|Y\rangle$ indicates the physicochemical profile of a property Y calculated for each amino acid (nucleotide) starting from the N-terminus (5′). For example, the hydrogen bonding profile, denoted by $|H\rangle$, is the hydrogen bonding ability of each amino acid (nucleotide) in the sequence:

$$|H\rangle = H_1, H_2, \ldots, H_L \tag{2}$$

Similarly, $|S\rangle$ represents the secondary structure occupancy profile and $|W\rangle$ the van der Waals' profile. The variable x indicates RNA ($x = r$) or protein ($x = p$) profiles. Secondary structure, hydrogen bonding and van der Waals contributions are calculated as described in the original articles (12). In particular, the RNA secondary structure is predicted from sequence using the Vienna package including the algorithms RNAfold, RNAsubopt and RNAplot (13). Model structures, ranked by energy, are used as input for *cat*RAPID. For each model structure, the RNAplot algorithm is used to generate secondary structure coordinates. Using the coordinates, we define the 'secondary structure occupancy' by counting the number of contacts within the nucleotide chain. High values of secondary structure occupancy indicate that base pairing occurs in regions with high propensity to form stems, while low values are associated with junctions or multi-loops.

We use discrete Fourier transform to compare interaction profiles of different length:

$$\Psi_{k,x} = \sqrt{\frac{2}{\text{length}}}\sum_{n=0}^{\text{length}} \Phi_{n,x}\cos\left[\frac{\pi}{\text{length}}\left(n+\frac{1}{2}\right)\left(k+\frac{1}{2}\right)\right]$$

$$k = 0,1,\ldots\ell$$

where the number of coefficients is $\ell = 50$. **(3)**

The 'interaction propensity' π is defined as the inner product between the protein propensity profile $|\Psi_\mathrm{p}\rangle$ and the RNA propensity profile $|\Psi_\mathrm{r}\rangle$ weighted by the 'interaction matrix' I:

$$\pi = \langle\Psi_\mathrm{p}|\mathrm{I}|\Psi_\mathrm{r}\rangle \tag{4}$$

To calculate the interaction propensity π, we exploit that the squared norm of π is conserved under Fourier transform:

$$\sum_{\mathrm{i,j}}^{\text{length}_\mathrm{p},\text{length}_\mathrm{r}} \left|\langle\psi_\mathrm{p}|\mathrm{I}|\psi_\mathrm{r}\rangle\right|^2 \approx \sum_{\mathrm{i,j}}^{\ell_\mathrm{r},\ell_\mathrm{p}} \left|\langle\psi_\mathrm{p}|\mathrm{I}|\psi_\mathrm{r}\rangle\right|^2 \tag{5}$$

The interaction matrix I as well as the parameters α_S, α_H and α_W are derived under the condition that interaction propensities π take maximal values for associations present in the positive training set (and minimal values for those in the negative training set):

$$\mathrm{I}: \begin{cases} \max\ \langle\Psi_\mathrm{p}|\mathrm{I}|\Psi_\mathrm{r}\rangle\ \forall\{\mathrm{r,p}\} \in \{\text{positive training set}\} \\ \min\ \langle\Psi_\mathrm{p}|\mathrm{I}|\Psi_\mathrm{r}\rangle\ \forall\{\mathrm{r,p}\} \in \{\text{negative training set}\} \end{cases} \tag{6}$$

In the training and test phases, we used protein and RNA sequences in the range of 50–750 amino acids and 50–3000 nt, respectively (12). We note that prediction of RNA secondary structures results in intense CPU usage when sequences are >1500 nt and simulations cannot be completed on standard processors (2.5 GHz; 4–8 GB memory).

The server to compute the interaction propensity with respect to the negative training set (discriminative power) is available at: http://tartaglialab.crg.cat/catrapid.html.

Interaction strength

Computational models indicate that RNA sequence length and secondary structure free energies are correlated (Supplementary Figure S1a) (14). Hence, one would expect that long RNAs are more stable and prone to bind to proteins than short RNAs (see also section 'interaction fragments'). Indeed, we observe a weak correlation between secondary structure energy and protein–RNA interaction propensity in our algorithm (Pearson's correlation = 20%; $P = 0.07$) (Supplementary Figure S2b). Nevertheless, as no experimental evidence indicates that long transcripts interact more than small RNAs, we eliminated the length dependence introducing a 'reference set' composed by protein and RNA sequences that have exactly the same lengths as the molecules under investigation. In our calculations, we use random associations between polypeptide and nucleotide sequences. Since little interaction propensities are expected from random associations, the reference set represents a 'negative control'.

For each protein–RNA pair under investigation, we use a reference set of 10^2 protein and 10^2 RNA molecules (the number of sequences is chosen to guarantee sufficient statistical sampling). To assess the strength of a particular association, we compute the interaction propensity π and compare it with the interaction propensities $\tilde{\pi}$ of the reference set (total of 10^4 protein–RNA pairs). Using the interaction propensity distribution of the reference set, we generate the 'interaction score':

$$\text{Interaction score} = \frac{\pi - \mu}{\sigma}$$

$$\begin{cases} \mu = \frac{1}{\Lambda}\sum_{i=1}^{\Lambda}\tilde{\pi}_i \\ \sigma^2 = \frac{1}{\Lambda}\sum_{i=1}^{\Lambda}(\tilde{\pi}_i - \mu)^2 \end{cases} \tag{7}$$

The number of interactions is $\Lambda = 10^4$. From the distribution of interaction propensities, we compute the 'interaction strength':

$$\text{Interaction strength} = P(\tilde{\pi} \leq \pi) = \text{cumulative distribution function (cdf)} \tag{8}$$

Reference sequences have the same lengths as the pair of interest to guarantee that the interaction strength is independent of protein and RNA lengths. The interaction strength ranges from 0 (non-interacting) to 100% (interacting). Interaction strengths >50% indicate propensity to bind. The 'RNA interaction strength' and the 'protein interaction strength' are special cases of the interaction strength in which only a reference set is generated using RNA or protein sequences. The RNA interaction strengths used for the analysis of RepA, 4R and 2R represent the RNA-binding abilities of SUZ12 and EZH2 with respect to the polynucleotide reference set (Figure 1). Similarly, the protein interaction strengths used for SFRS1, SAF-A and SATB1 are the protein-binding abilities of the experimental RNA fragments with respect to the polypeptide reference set (Figures 2 and 4). The interaction strength is also used to compare YY1- and green fluorescent protein (GFP)-binding propensities (proteins are RNA fragments are of different lengths; Figure 3). It should be noted that in the case of *Xist* fragment BC (nt 1898–4940), the RNA sequence is >3 kb. In order to calculate the abilities of fragment BC to interact with YY1 and GFP, we analyzed all *Xist* fragments of size 1500 nt contained in the region 1898–4940 nt, computed the corresponding interaction strengths and averaged the scores.

The server to compute the interaction strength is available at: http://tartaglialab.crg.cat/catrapid.strength.html.

Interaction fragments

The use of RNA fragments is introduced to identify RNA regions involved in protein binding. The RNALfold algorithm from the Vienna package (www.tbi.univie.ac.at/RNA/) is used to select RNA fragments in the range of 100–200 nt with predicted stable secondary structure. Secondary structure stabilities are estimated by calculating the RNA free energy predicted by RNALfold (15). As long RNA segments have lower free energy for the higher number of bases that can be paired (Supplementary Figure S1a) (14), the choice of segments in the range of 100–200 nt is optimal because it allows simultaneously: (i) selection of secondary structures with comparable free energy (Supplementary Figure S1b) and (ii) high sequence coverage (>90%) for long transcripts such as *Xist* (Supplementary Figure S1c). Once the RNA fragments are selected, catRAPID is used to predict their ability to bind to polypeptide chains. Conceptually, the interaction fragments algorithm is a variant of the RNA interaction strength algorithm that allows identification of putative binding areas in long sequences. If the exact protein and/or RNA domains are known, we recommend the use of the interaction strength method to predict the binding specificity (Figure 3).

The server to compute fragment interactions is available at: http://tartaglialab.crg.cat/catrapid.fragments.html.

RESULTS

Xist-mediated X-chromosome silencing implies a complex network of macromolecular associations orchestrated by epigenetic modifiers as well as splicing and transcription factors. *Xist* function at the initiation of X-inactivation has been extensively studied in mouse embryonic stem cells. The mouse system is more accessible to experimental investigation than the human one and is here investigated

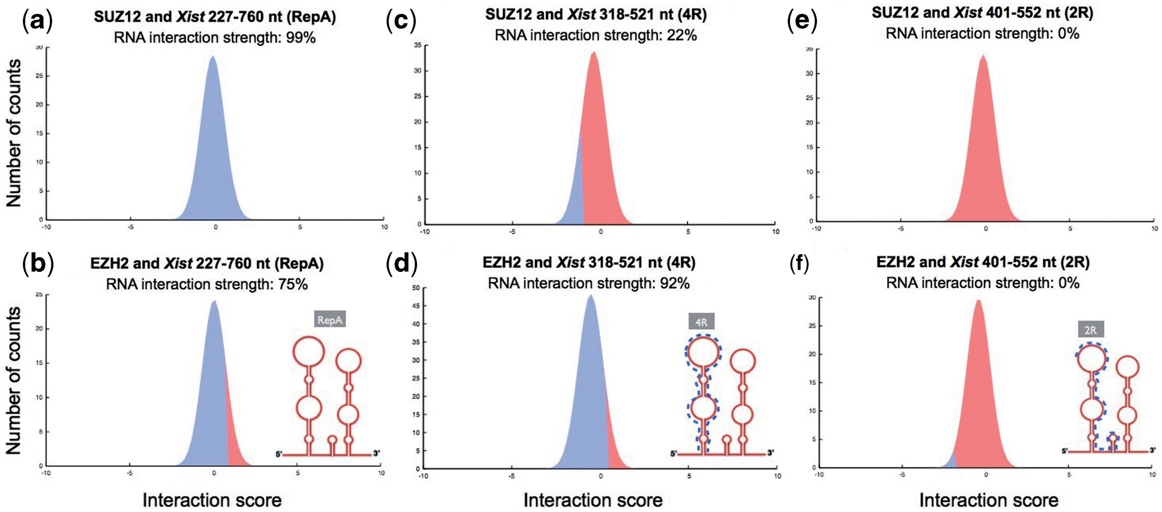

Figure 1. *Xist* RepA, 4R, 2R and PcG proteins. We predict that *Xist* RepA (227–760 nt) binds strongly to (**a**) SUZ12 (RNA interaction strength = 99%), and (**b**) EZH2 (RNA interaction strength = 75%), in agreement with experimental evidence; (**c**) SUZ12 does not bind to repeat 4R (318–521 nt; RNA interaction strength = 22%), while (**d**) EZH2 shows high interaction propensity (RNA interaction strength = 92%). Neither (**e**) SUZ12 nor (**f**) EZH2 are in contact with repeat 2R (401–552 nt; RNA interaction strengths = 0; Supplementary Table S1c) (7). Insets (b, d and f) are secondary structures of RepA (red line), 4R and 2R (blue dots) proposed by Maenner *et al.* (7).

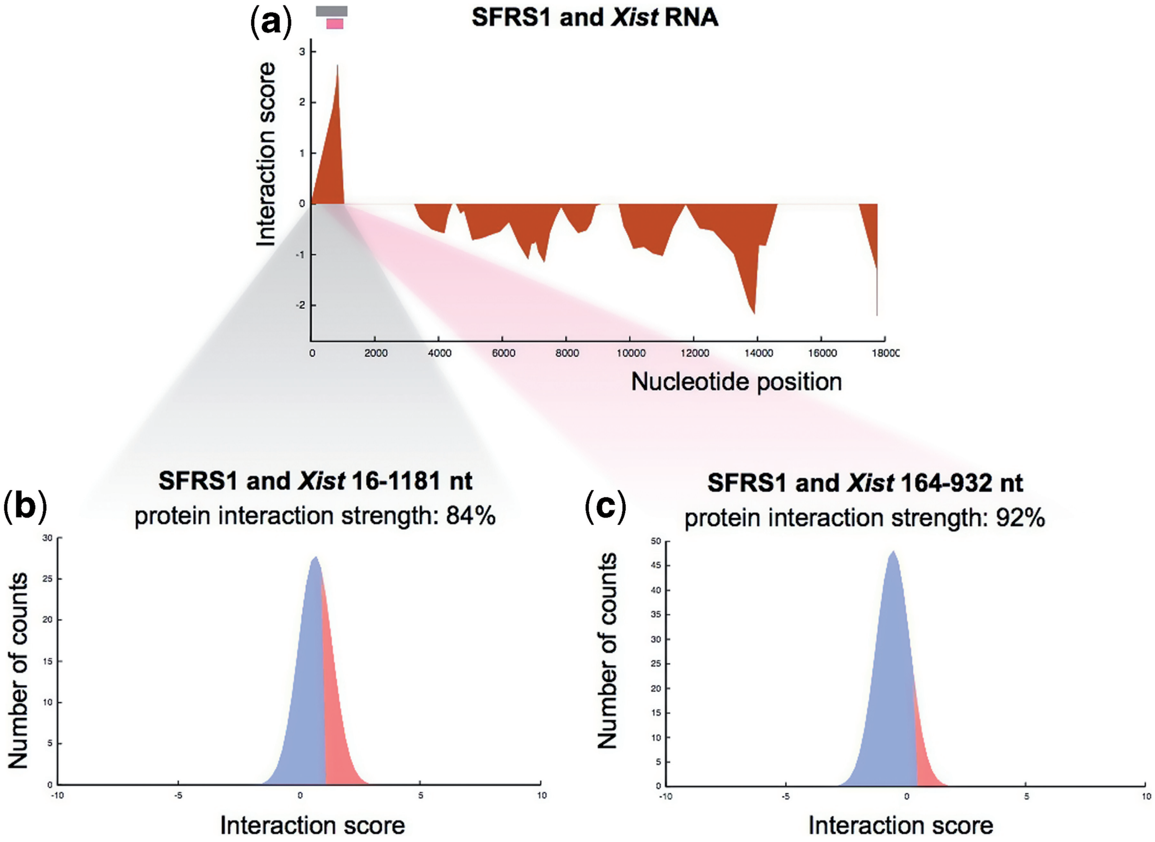

Figure 2. *Xist* and alternative splicing factor SFRS1. The interaction fragments algorithm is used to predict *Xist* ability to interact with SFRS1. (**a**) SFRS1 shows high propensity to contact *Xist* 5′. (**b**) The region studied by Royce-Tolland *et al.* (8) is marked in grey (nt 16–1181). In agreement with experimental evidence, strong interaction propensity is predicted between SFRS1 and nt 16–1181 (protein interaction strength = 84%); (**c**) nucleotides 164–932 nt (marked in red) correspond to an RNA region whose deletion abolishes *Xist* splicing (8). Strong interaction propensity is predicted between SFRS1 and nt 164–930 (protein interaction strength = 92%), as previously reported (8).

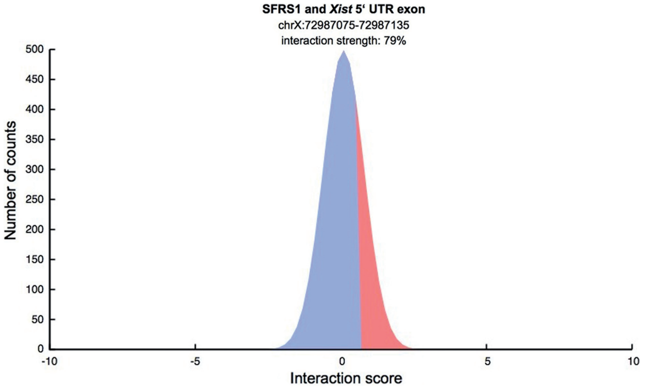

Figure 3. SFRS1 and *Xist* 5′-UTR. We predict that SFRS1 interacts with the 5′-UTR exon region of *Xist*, in agreement with CLIP-seq experiments (18).

using two novel algorithms: interaction strength and interaction fragments.

SUZ12 and EZH2 bind to RepA

The Polycomb repressive complex 2 (PRC2) is one of the two classes of PcG proteins and plays a major role in the epigenetic silencing of X-chromosome (7). More specifically, PRC2 is associated with histone modifications promoting tri-methylation of histone H3 lysine 27 along the X-chromosome, which is thought to generate a repressive compartment for silencing (16). In agreement with experimental evidence, we predict that *Xist* Repeat A region (RepA) interacts with PRC2 (7). More specifically, we find that Suppressor of Zeste 12 (SUZ12) protein homolog and Enhancer of Zeste homolog 2 (EZH2) have

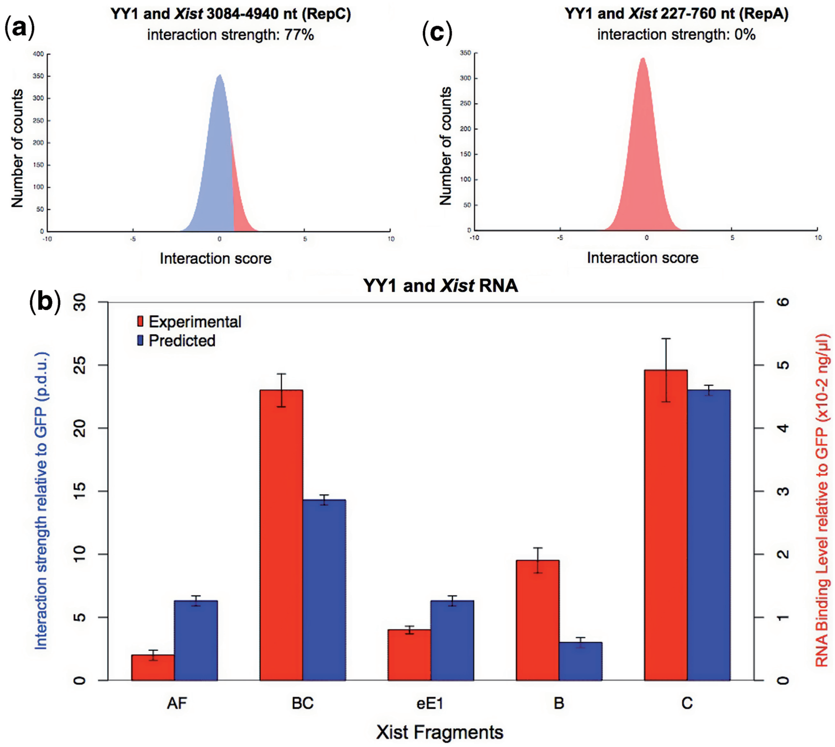

Figure 4. *Xist* and transcriptional repressor Ying and Yang (YY1). The interaction strength algorithm is used to predict YY1 ability to interact with *Xist*. (**a**) High interaction propensity is found between YY1 and *Xist* Repeat C region (RepC; interaction strength = 77%). (**b**) No interaction is predicted between YY1 and RepA (interaction strength = 0%), as previously reported (9). (**c**) Experimental binding levels of AF, B, C, BC and eE1 fragments (red bars) are reproduced by catRAPID (blue bars) with high accuracy (Pearson's correlation = 92%; $P = 0.04$ estimated with analysis of variance, two-tailed *t*-test (9) (Supplementary Table S1b). Interactions strengths and RNA-binding levels are normalized subtracting GFP signals (Supplementary Figure S2b). Errors on catRAPID predictions are evaluated using the second derivative of the cumulative distribution function associated with the interaction strength.

strong propensities to bind to RepA (region 227–760 nt; RNA interaction strengths >75%; Figures 1a and d and 5; 'Material and Methods' section). Hence, our results clearly indicate that *Xist* is able to contact PRC2 without mediation of other molecules (7).

Based on secondary structure predictions, it has been proposed that RepA contains two long stem-loop structures of ~200 nt, each containing four repeats (6,7). Nuclear magnetic resonance studies have given indication that the second loop has higher propensity to pair (17). This pairing propensity can lead to multiple interactions and complex folding. Such folding was indeed observed by structural probing of RepA and a large set of interactions have been observed with no direct evolutionary conservation or consistency with known mutations (7).

By using chemical and enzymatic probes as well as Förster resonance energy transfer experiments, EZH2 was shown to bind to RepA and repeat 4R located at position 318–521 nt within RepA (7) (Figure 5 and Supplementary Table S1a and b). By contrast, SUZ12 was found to interact with RepA and not 4R (7). Our predictions show that both EZH2 and SUZ12 contact RepA (RNA interaction strengths >75%) and that EZH2 binds to 4R (RNA interaction strength = 92%), whereas SUZ12 shows much lower binding propensity (RNA interaction strength = 22%). Moreover, we predict that neither EZH2 nor SUZ12 is able to interact with region 2R; (nt 401–552; RNA interactions strengths = 0%), as previously demonstrated by immunoprecipitation assays and western blot analysis (7) (Supplementary Table S1). In agreement with experimental evidence, we also predict that EZH2 binds to the reverse complement of RepA present in *Tsix* (2073–2239 nt; Supplementary Figure S2a) (5).

SFRS1 associates with RepA

Stochastic differences in *Xist* RNA levels influence the production of spliced RNA in the two X-chromosomes, thus leading to inactivation of one chromosome upon differentiation (8). Using HeLa cell nuclear extracts and ultraviolet cross-linking, Royce-Tolland *et al*. (8) showed that the splicing factor SFRS1 is able to associate with RepA. Here, we use the interaction fragments algorithm to predict the ability of SFRS1 to interact with *Xist*. In our analysis, the interaction propensities are calculated using RNA fragments with predicted stable secondary structure ('Materials and Methods' section). In agreement with *in vitro* and *in vivo* experiments (8), we find that SFRS1 interacts with RepA (nt 682–881, 707–826 and 726–907; Supplementary Table S1a). In particular, we predict that SFRS1 has strong propensity to bind to the domain investigated by Royce-Tolland *et al.* (nt 16–1181; protein interaction strength = 84%; Figure 2b; 'Material and Methods' section) and with a

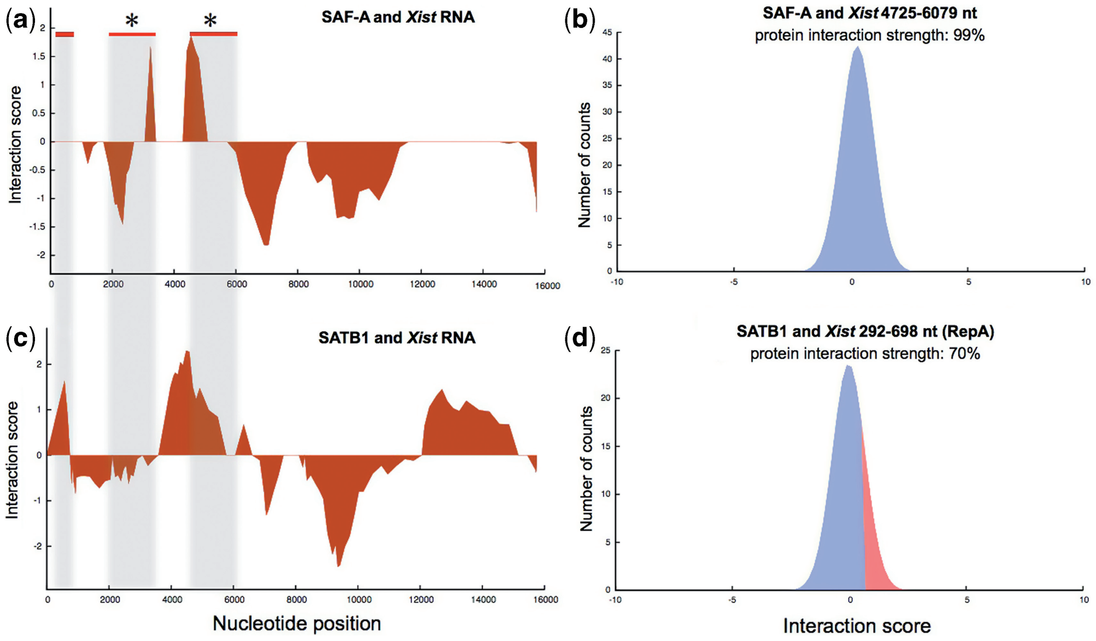

Figure 5. *Xist*, scaffold attachment factor SAF-A and special AT-rich sequence-binding protein SATB1. (**a**) In agreement with experimental evidence, SAF-A is predicted to contact *Xist* in more than one region (10). Red lines and grey boxes indicate experimentally validated regions involved in *Xist* localization (6). Stars mark primers of elements studied by Hasegawa *et al.* (10). (**b**) SAF-A shows strong propensity to bind to *Xist* region 4934–5056 nt (protein interaction strength = 99%). (**c**) Multiple binding sites are predicted between *Xist* and SATB1. (**d**) We predict that SATB1 binds strongly to nt 292–698 (RepA; protein interaction propensity = 70%), as previously suggested (6,11).

fragment whose deletion abrogates *Xist* splicing (nt 164–932; protein interaction strength = 92%; Figure 2c); (8). Thus, our results indicate that SFRS1 is directly recruited for selective inactivation of the X-chromosome (8).

Recently, Sanford *et al.* (18) used cross-linking immunoprecipitation coupled with high-throughput sequencing (CLIP-seq) to characterize SFRS1's interactome. Using HEK293T cells, the authors gathered a large amount of information on the RNA-binding sites targeted by SFRS1. In particular, CLIP-seq experiments indicate that SFRS1 binds to the 5′-UTR exon region of *Xist* (coordinates chrX:72987075–72987135 in the Human Genome Assembly 18) (18). In agreement with this finding, we predict high interaction propensity between SFRS1 and the 5′-UTR exon region (interaction strength: 79%; Figure 3).

We take the opportunity offered by CLIP-seq experiments to assess catRAPID's ability to predict SFSR1's interactions. In our analysis, we use RNA regions containing the highest number of CLIP-seq-binding sites (i.e. CLIP-seq 'clusters'). Using the interaction strength algorithm, we predict that 78 out of 100 large (>50 nt) clusters bind to SFSR1 with average interaction strength of 69% (Supplementary Figure S3a), which indicates strong agreement between observed and predicted interactions. Based on the analysis of SFRS1 CLIP-seq experiments, Wang *et al.* (19) developed the 'RNAMotifModeler' algorithm to predict RNA-binding sites using sequence features and secondary structures. RNAMotifModeler identifies binding motifs in 72 out of 100 large clusters (motifs AGAAGA, AAGAAG and GA AGAA; Supplementary Figure S3a), which is fully compatible with catRAPID's performances. We also analyse the interaction propensity of 100 small (<50 nt) clusters and their corresponding upstream and downstream regions (Supplementary Figure S3b). High interaction propensities are observed for regions containing SFSR1-binding sites (interactions predicted by catRAPID: 76; RNAMotifModeler motifs: 25; Supplementary Figure S3b), while lower interaction strengths and fewer binding motifs are predicted in the flanking regions (interactions predicted by catRAPID: 30; RNAMotifModeler motifs: 10; Supplementary Figure S3b).

YY1 contacts RepC

To study *Xist* RNA localization onto the X-chromosome, Jeon and Lee (9) introduced a doxycycline-inducible *Xist* transgene into female mouse embryonic fibroblasts. Multiple independent clones showed that *Xist* transgenes act on endogenous locus *in trans* and squelch *Xist* RNA clouds on the inactive X (9). The authors reported that RepA elimination does not abolish *Xist* RNA clouds squelching, which indicates that the region is not required for X-chromosome localization (9). By contrast, knocking down of transcriptional repressor YY1 can be correlated with 70% loss of *Xist* clouds. Importantly, pull-down assays showed that *Xist* RNA repeat C, a conserved C-rich element repeated 14 times in tandem (RepC; 3084–4940 nt; Figure 6), has a pronounced ability to bind to YY1 with respect to GFP.

Using the interaction strength approach, we are able to recapitulate all the *in vitro* assays performed by Jeon and Lee to probe YY1 affinity for *Xist* fragments (9).

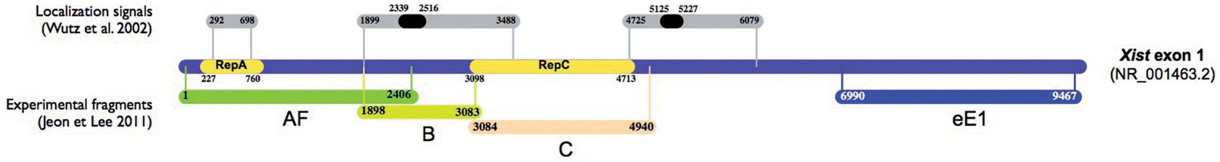

Figure 6. *Xist* first exon. RepA and RepC (yellow lines) encompass nt 227–760 and 3098–4713 (8,15). YY1 interactions investigated by Jeon and Lee (9) correspond to nt 1–2406 (AF), 1898–3083 (B), 3084–4940 (C) and 6990–9467 (eE1). The localization signals identified by Wutz *et al.* (6) are indicated by grey lines at nt 292–698, 1899–3488 and 4725–6079. The primers used by Hasegawa *et al.* correspond to nt 2339–2515 and 5125–5227 (10).

According to our calculations, *Xist* RepC shows very high propensity to interact with YY1 (Figure 4a and b), followed by one region containing an overlap between RepC and Repeat B region (RepB; Figure 4b).

In striking agreement with experimental evidence, we predict that YY1 interacts with *Xist* through RepC and RepB (Figure 4c; Pearson's correlation = 92%; $P = 0.04$) and does not associate directly with RepA (9,20).

SAF-A interacts with *Xist* 5′

Xist chromosomal localization is regulated by *cis*-elements in the 5′-half of the transcript located at nt 292–698 (RepA), 1899–3488 and 4725–6079 (Supplementary Table S1b) (6). Recently, the nuclear scaffold protein SAF-A has been linked with *Xist* localization (21). SAF-A contains three conserved domains: a SAF-box (22) binding to AT-rich DNA regions (23), Spla and Ryanodine receptor (SPRY) domain of unknown function (24) and an arginine–glycine glycine (RGG) RNA-binding domain. Deletion of the RGG-binding domain strongly reduces *Xist* chromosomal localization, suggesting direct interaction with *Xist* (10).

Using co-immunoprecipitation assays, Hasegawa *et al.* (10) reported that SAF-A contacts nt 1899–3488 and 4725–6079 (Figure 6). Employing the interaction fragments method, we find that these regions are highly prone to interact with SAF-A (Figure 5a and Supplementary Table S1a). In particular, we predict that nt 4725–6079 have strong propensity to bind to SAF-A (protein interaction strength = 77%; Figure 4b). In our analysis, we used a protein region spanning residues 50–800, which contain the uncharacterized SPRY region and the RNA-binding domain RGG (Supplementary Table S1c). By sliding a window of 750 amino acids from the N- to the C-terminus of SAF-A, we observe that the interaction fragments profiles correlate significantly (mean Pearson's correlation = 90%; $P = 0.01$; Supplementary Figure S4a). Intriguingly, when the SAF-box is included in the analysis (residues 9–759), we predict an increased ability to bind to RepA (Supplementary Figure S4b). The binding region present in RepA (Supplementary Figure S4b and Table S1b) was not investigated by Hasegawa *et al.* (10), but is consistent with the observations made by Wutz *et al.* (6) and the fact that deletion of SAF-box abolishes *Xist* chromosomal localization (10).

In agreement with experimental data, we expect that direct interaction between *Xist* and SAF-A could have an effect on *Xist* localization in the nuclear matrix, thus facilitating association with chromosomal DNA (6,10).

Does SATB1 binds to multiple *Xist* sites?

In a thymic lymphoma model, the nuclear protein SATB1 was identified as a critical component for gene silencing (25). In fact, it has been shown that viral expression of SATB1 in fibroblasts—in which *Xist* does not induce gene repression—could establish *Xist* silencing (3,25). As SATB1 co-localizes with *Xist* at the initiation of X-inactivation (25), it has been proposed that it could act as an anchor promoting RepA-mediated chromosomal reorganization (26). Nevertheless, it should be noted that SATB1 binds and regulates chromatin domains containing genes, whereas *Xist* overlaps chromosomal regions that are enriched for genomic repeats and deprived of genes. This aspect could lead to the idea that SATB1 makes genes susceptible to *Xist* by positioning gene-rich chromatin, without direct interaction (3).

In our calculations, we use SATB1 residues 23–764 (Supplementary Table S1c), which contain all the functional domains with exclusion of protein localization signals. Employing the interaction fragments method, we predict interactions for two regions identified by Wutz *et al.* (nt 292–698 and 4725–6079; Figure 6) (6). In particular, we find that SATB1 has strong propensity to bind to RepA (region 292–698 nt; interaction strength: 85%; Figure 5d), as suggested by Arthold *et al.* (11). Intriguingly, we observe previously uncharacterized binding sites in correspondence of the 3′-region (Figure 5c), in agreement with the fact that more than one *Xist* region could be involved in low-affinity cooperative binding of protein factors (3,6).

DISCUSSION

XCI is a complex process that requires several regulated events such as the *Xist* localization onto the X-chromosome and its spatial confinement. These steps are controlled by transcriptional factors and nuclear scaffold proteins, which play a role in the selection of chromosome and recruitment of silencing machinery. One of the first processes during XCI is the random selection of the X-chromosome to be silenced. The choice has been suggested to be stochastically determined by levels of spliced *Xist* RNA accumulated on the X-chromosome (8). We find that the splicing factor SFRS1 binds to the 5′-UTR exon (Figure 3) and RepA (Figure 2b and c), which suggests direct involvement of this protein in the production of mature *Xist* (8). Although RepA is fundamental for PCR2 recruitment and chromosomal silencing (7), we predict that it is unlikely to be involved in the interaction with YY1 (9) (Figure 4b). By contrast, we find that RepC has high interaction propensity for YY1 (Figure 4a and b). Hence, our predictions support the current hypothesis that PRC2 is co-transcriptionally recruited by RepA, while YY1 tethers RepC on the X-inactivation centre (9).

How the *Xist*–PRC2 complex translocates *in cis* along the X-chromosome is an open and tantalizing question. It has been reported that the nuclear scaffold factor SAF-A facilitates the association of *Xist* with nuclear matrix (10). Indeed, the nuclear matrix could provide a highly dynamic structure (27,28) to control *Xist* movements. We observe that the interaction profile of SAF-A correlates (Figure 5a) with that of the nuclear matrix protein SATB1 (Figure 5c) at the 5′, suggesting a possible synergistic mechanism of action to organize *Xist* translocation along the X-chromosome. The involvement of matrix-associated factors in the X-chromosome coating represents an intriguing scenario to be further investigated experimentally.

Our calculations suggest that localization and confinement of *Xist* are finely regulated by multiple factors acting at the interface between chromosome X and the nuclear matrix. Our results are compatible with a model in which following X-chromosome docking mediated by YY1 (9), matrix-associated proteins SAF-A and SATB1 recruit the 5′-half of *Xist* and drive the translocation *in cis* of the *Xist*–PRC2 complex.

In this work, we presented a new version of the catRAPID method to study *Xist* associations with a number of proteins, including SUZ12, EZH2, YY1, SAF-A, SFRS1 and SATB1. In striking agreement with experimental evidence, we demonstrated that our algorithms predict RNA-binding sites and affinities for a number of epigenetic, splicing and transcription factors. In particular, we investigated the association with transcription repressor YY1, which favours *Xist* tethering onto the X-chromosome, and nuclear matrix proteins SAF-A and SATB1, which guide its translocation. We also applied our method to SFRS1's interactome, showing that catRAPID predicts CLIP-seq-binding sites with great accuracy (18). Most importantly, we showed that computational approaches can provide a solid basis for the investigation of protein interactions with long non-coding transcripts (20).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Table 1 and Supplementary Figures 1–4.

ACKNOWLEDGEMENTS

The authors thank Domenica Marchese; Prof. R. Guigo'; Dr B. Keyes and Dr L. di Croce for stimulating discussions and Dr J.R. Sanford for having provided CLIP-seq data on SFSR1.

FUNDING

Funding for open access charge: Spanish Ministry of Economy and Competitiveness [SAF2011-26211].

Conflict of interest statement. None declared.

REFERENCES

1. Navarro,P. and Avner,P. (2010) An embryonic story: analysis of the gene regulative network controlling *Xist* expression in mouse embryonic stem cells. *Bioessays*, **32**, 581–588.
2. Tattermusch,A. and Brockdorff,N. (2011) A scaffold for X chromosome inactivation. *Hum. Genet.*, **130**, 247–253.
3. Wutz,A. (2011) Gene silencing in X-chromosome inactivation: advances in understanding facultative heterochromatin formation. *Nat. Rev. Genet.*, **12**, 542–553.
4. Lee,J.T., Davidow,L.S. and Warshawsky,D. (1999) Tsix, a gene antisense to *Xist* at the X-inactivation centre. *Nat. Genet.*, **21**, 400–404.
5. Zhao,J., Sun,B.K., Erwin,J.A., Song,J.-J. and Lee,J.T. (2008) Polycomb proteins targeted by a short repeat RNA to the mouse X-chromosome. *Science*, **322**, 750–756.
6. Wutz,A., Rasmussen,T.P. and Jaenisch,R. (2002) Chromosomal silencing and localization are mediated by different domains of *Xist* RNA. *Nat. Genet.*, **30**, 167–174.
7. Maenner,S., Blaud,M., Fouillen,L., Savoye,A., Marchand,V., Dubois,A., Sanglier-Cianférani,S., Van Dorsselaer,A., Clerc,P., Avner,P. *et al.* (2010) 2-D structure of the A region of *Xist* RNA and its implication for PRC2 association. *PLoS Biol.*, **8**, e1000276.
8. Royce-Tolland,M.E., Andersen,A.A., Koyfman,H.R., Talbot,D.J., Wutz,A., Tonks,I.D., Kay,G.F. and Panning,B. (2010) The A-repeat links ASF/SF2-dependent *Xist* RNA processing with random choice during X inactivation. *Nat. Struct. Mol. Biol.*, **17**, 948–954.
9. Jeon,Y. and Lee,J.T. (2011) YY1 tethers *Xist* RNA to the inactive X nucleation center. *Cell*, **146**, 119–133.
10. Hasegawa,Y., Brockdorff,N., Kawano,S., Tsutui,K., Tsutui,K. and Nakagawa,S. (2010) The matrix protein hnRNP U is required for chromosomal localization of *Xist* RNA. *Dev. Cell*, **19**, 469–476.
11. Arthold,S., Kurowski,A. and Wutz,A. (2011) Mechanistic insights into chromosome-wide silencing in X inactivation. *Hum. Genet.*, **130**, 295–305.
12. Bellucci,M., Agostini,F., Masin,M. and Tartaglia,G.G. (2011) Predicting protein associations with long noncoding RNAs. *Nat. Methods*, **8**, 444–445.
13. Gruber,A.R., Lorenz,R., Bernhart,S.H., Neubock,R. and Hofacker,I.L. (2008) The Vienna RNA Websuite. *Nucleic Acids Res.*, **36**, W70–W74.
14. Pervouchine,D.D., Graber,J.H. and Kasif,S. (2003) On the normalization of RNA equilibrium free energy to the length of the sequence. *Nucleic Acids Res.*, **31**, e49–e49.
15. Hofacker,I.L., Priwitzer,B. and Stadler,P.F. (2004) Prediction of locally stable RNA secondary structures for genome-wide surveys. *Bioinformatics*, **20**, 186–190.
16. Fang,J., Chen,T., Chadwick,B., Li,E. and Zhang,Y. (2004) Ring1b-mediated H2A ubiquitination associates with inactive X chromosomes and is involved in initiation of X inactivation. *J. Biol. Chem.*, **279**, 52812–52815.
17. Duszczyk,M.M., Zanier,K. and Sattler,M. (2008) A NMR strategy to unambiguously distinguish nucleic acid hairpin and duplex conformations applied to a *Xist* RNA A-repeat. *Nucleic Acids Res.*, **36**, 7068–7077.
18. Sanford,J.R., Wang,X., Mort,M., Vanduyn,N., Cooper,D.N., Mooney,S.D., Edenberg,H.J. and Liu,Y. (2009) Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts. *Genome Res.*, **19**, 381–394.
19. Wang,X., Juan,L., Lv,J., Wang,K., Sanford,J.R. and Liu,Y. (2011) Predicting sequence and structural specificities of RNA binding regions recognized by splicing factor SRSF1. *BMC Genomics*, **12(Suppl. 5)**, S8.
20. Thorvaldsen,J.L., Weaver,J.R. and Bartolomei,M.S. (2011) A YY1 Bridge for X Inactivation. *Cell*, **146**, 11–13.

21. Fackelmayer,F.O. (2005) A stable proteinaceous structure in the territory of inactive X chromosomes. *J. Biol. Chem.*, **280**, 1720–1723.
22. Kipp,M., Göhring,F., Ostendorp,T., van Drunen,C.M., van Driel,R., Przybylski,M. and Fackelmayer,F.O. (2000) SAF-Box, a conserved protein domain that specifically recognizes scaffold attachment region DNA. *Mol. Cell. Biol.*, **20**, 7480–7489.
23. Mirkovitch,J., Gasser,S.M. and Laemmli,U.K. (1987) Relation of chromosome structure and gene expression. *Phil. Trans. R. Soc. Lond. B, Biol. Sci.*, **317**, 563–574.
24. Ponting,C., Schultz,J. and Bork,P. (1997) SPRY domains in ryanodine receptors (Ca(2+)-release channels). *Trends Biochem. Sci.*, **22**, 193–194.
25. Agrelo,R., Souabni,A., Novatchkova,M., Haslinger,C., Leeb,M., Komnenovic,V., Kishimoto,H., Gresh,L., Kohwi-Shigematsu,T., Kenner,L. *et al.* (2009) SATB1 defines the developmental context for gene silencing by *Xist* in lymphoma and embryonic cells. *Dev. Cell*, **16**, 507–516.
26. Chaumeil,J., Le Baccon,P., Wutz,A. and Heard,E. (2006) A novel role for *Xist* RNA in the formation of a repressive nuclear compartment into which genes are recruited when silenced. *Genes Dev.*, **20**, 2223–2237.
27. Albrethsen,J., Knol,J.C. and Jimenez,C.R. (2009) Unravelling the nuclear matrix proteome. *J. Proteomics*, **72**, 71–81.
28. Simon,D.N. and Wilson,K.L. (2011) The nucleoskeleton as a genome-associated dynamic ‘network of networks’. *Nat. Rev. Mol. Cell Biol.*, **12**, 695–708.