Pergamon

Bioorganic & Medicinal Chemistry 9 (2001) 2533–2541

BIOORGANIC & MEDICINAL CHEMISTRY

Selection of Genomic Target RNAs by Iterative Screening

Christine Brunel,[a,b,*] Bernard Ehresmann,[a] Chantal Ehresmann[a] and Michael McKeown[b,†]

[a]*UPR 9002 du CNRS, Institut de Biologie Cellulaire et Moléculaire, 15 rue Descartes, 67084 Strasbourg Cedex, France*
[b]*Molecular Biology and Virology Laboratory, Salk Institute, PO Box 85800, San Diego, CA 92186, USA*

Received 22 September 2000; revised 15 December 2000; accepted 2 January 2001

Abstract—A growing number of proteins are known to exert their regulatory or biological functions via RNA binding. In some cases genetic interactions allow us to infer candidate targets for RNA directed regulation, but in many other cases identification of potential regulatory targets is problematic. We have developed an in vitro biochemical screen, SETIS (SElection of << genomic >> Target RNAs by Iterative Screening) that allows screening of a major portion of the genome for identification of potential targets for RNA binding proteins.

Introduction

Post-transcriptional regulation of gene expression can involve the on/off regulation, modulation or even expression of alternative forms of particular gene products in a temporally and spatially regulated manner. This allows cells of different types or at different developmental stages to fine tune their patterns of gene expression. In fact, many important events in development such as pattern formation and terminal differentiation, are regulated by an array of post-transcriptional mechanisms controlling differential splicing, mRNA stability, localization and translation. Clearly, an understanding of the molecular basis of interactions between RNA and proteins is a prerequisite for understanding many physiological processes.

In vitro selection has been a powerful technique for identifying potential binding sequences for known nucleic acids binding proteins and for identifying nucleic acids sequences which can interact in novel ways with specific proteins.[1–4] Techniques of this type, such as SELEX,[2] are based on repeated selection cycles for protein binding from an initially large, often random pool (up to 10^{15} variants) of small nucleic acids (aptamers). In each round of selection, a population of RNA sequences is screened for protein binding. Protein-bound survivors are saved and amplified.[5] One can exert control over almost every aspect of the in vitro selection process from pool design (sequences can be either partially or completely randomized) to stringency of the selection.

Such techniques, based on synthetic generation of initial substrates, do not detect biological targets for RNA binding proteins, such as RNAs encoded in genomic DNA. To overcome this difficulty, we have developed the SETIS strategy (SElection of << genomic >> Target RNAs by Iterative Screening). SETIS couples the power of in vitro iterative selection–amplification techniques like SELEX with scanning the full potential capacity of the genome (Fig. 1). The use of genomic fragments to create the initial pool used as a starting population of the selection is remisniscent with the concept of genomic selex described by Gold and collaborators.[4,6] However the means used to create the library are differents, (restriction enzymes versus PCR).[6] Gold and collaborators applied a strict restriction on the size of the fragments used during the selection (the length was set at 50 nucleotides).[6] Moreover the choice of the organism was different: *D. melanogaster* in this work versus *E. coli*, *S. cerevisiae*, and human.[6]

As a first test of this technique with a known biological model, we have used the Drosophila Transformer 2 (Tra2)protein. Tra2 is an RNA-binding protein of the super family of Serine-Arginine like proteins, constitutive components of the splicing machinery.[7,8] Tra2 was initially identified from genetic screens as being

*Corresponding author. Tel.: +33-03-88-41-70-40; fax: +33-03-88-60-18-22; e-mail: c.brunel@ibmc.u-strasbg.fr
†Current address: Molecular Biology, Cell Biology, Biochemistry, Box G-J363, and Brown University, Providence, RI 02912, USA

0968-0896/01/$ - see front matter
PII: S0968-0896(01)00029-3

necessary for female sexual differentiation and male fertility.[9] Further studies of sexual differentiation identified two direct targets for the action of Tra 2, *doublesex* (*dsx*) and *fruitless* (*fru*) mRNAs.[10–12] Indeed, the *dsx* regulatory element, located in the female specific exon contains six 13-nucleotides repeated sequences (with consensus of UCUUCAAUCAACA), referred to as the *dsx* repeat element (*dsx*RE) and a purine-rich element (PRE).[13,14] Tra2 binds in vitro to both types of *dsx* regulatory elements, the *dsx*RE and the PRE.[13,14] Finally, Tra2 is implicated in regulation of expression and processing of three genes in the testes: *tra2* itself, *exuperantia* (*exu*) and *alternative testis target* (*att*).[15,16] Thus, Tra2 presents us with a chance to test the SETIS strategy for identification of known targets and for the potential identification of high affinity genomic targets for the action of Tra2, perhaps separate from the action of Tra.

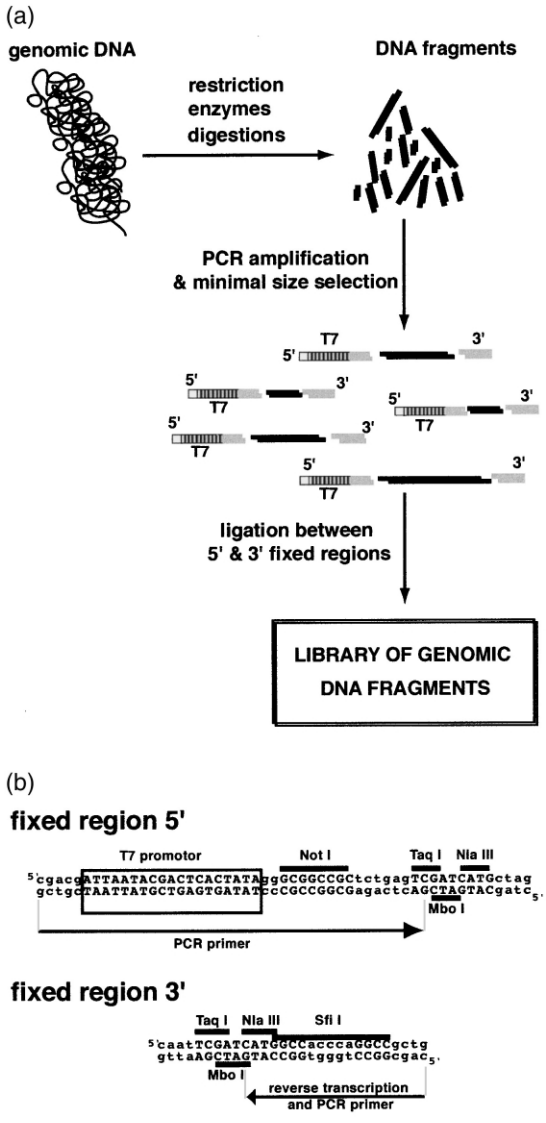

Figure 1. (A) General strategy of SETIS initial library construction. (B) Sequence and structure of the fixed regions flanking the sequence of interest and used to pursue the selection. 5′ and 3′ design respectively the 5′ and 3′ fixed region, and T7 the phage T7 RNA polymerase promotor.

In the present paper, we present the SETIS strategy. We describe the method of library constuction from *Drosophila melanogaster* genomic DNA, then the selection for binding to Tra2. Preliminary data indicates that the *dsx*RE has been selected providing a positive control to our selection procedure. We also present first preliminary evidence that new targets of Tra2 may have been selected.

Results

Library construction

SETIS is a technique designed to identify potential protein binding sites within primary transcripts, whether the binding sites are in exons, introns or the region distal to the poly (A) addition site. For SETIS to work it is necessary to produce a set of RNAs representing the euchromatic portion of the genome, subject those RNAs to selection for protein binding and amplify copies of those RNAs in a form that allows production of more copies of the RNAs for the next round of selection. As with SELEX, the production of RNAs and the post-selection amplification needs DNA libraries containing defined sequences flanking the genomic fragments. These two fixed regions were added by ligation (Fig. 1). The fixed 5′ region contains the phage T7 RNA polymerase promotor. These 5′ and 3′ fixed regions also allow RT–PCR reamplification of the RNAs which are bound in the selection process (Fig. 1). The efficiency of the phage polymerase and the RT–PCR reaction define the nature of the genomic DNA inserts to be used as the templates for RNA production. Specifically, for efficient RNA synthesis and amplification the inserts should be $\leqslant 1$ kb, with optimal sizes, for our purposes, in the range of 50–400 nt.

In order to produce an appropriate array of fragments while at the same time favoring efficient ligation to our defined-sequence 5′ and 3′ oligonucleotides, we chose to generate fragments via restriction enzyme digestion. The enzymes Taq I (T′CGA) and Dpn II (′GATC) were chosen because they have four base recognition sequences with equal GC and AT content, they generate sticky ends which favor ligation, and the sticky ends are compatible with sites generated by additional enzymes which cut less often such as Cla I (AT′CGAT), Bam H1 (G′GATCC) and Bgl II (A′GATCT). We reasoned that using appropriate combinations of single and double digests with these enzymes we could generate an almost uniform size distribution for sequences throughout the genome.

Prior to library construction, we electronically tested our predictions about the ability of these enzymes to generate the appropriate array of fragment sizes. Starting with a collection of a few hundred thousand bases of nucleotide sequence data we predicted the size profiles that would be generated with different combinations of enzymes. Single digestion with Taq I or Dpn II generates fragments with an average size of 250 and 318 nucleotides respectively. However a significant fraction

(14% and 39% respectively) of the DNA is left in fragments greater than 1 kb. Double digestion with both Taq I and Dpn II reduces the number of fragments with a length over 1000 nucleotides but creates numerous very small fragments (from 1 to 50 nucleotides). To generate intermediate arrays of fragment sizes between those seen with the four base cutter single digests and double digests, we did theoretical digests with one enzyme having a four base recognition sequence and one recognizing a six base sequence with a different core sequence (for example Taq I with Bam H1 or Bgl II; Dpn II with Cla I). Our calculations show that this kind of double digestion optimizes the number of medium size fragments (between 100 and 400 nucleotides). If appropriate libraries are generated from each of the single or double digest pairs plus Cla I-Bgl II and Cla I-Bam H1 digests, each nucleotide sequence is potentially present in eight different sequence contexts. We calculate that this pool of eight different libraries will have a theoretical complexity of 5×10^6 fragments. Almost 95% of the genome should be represented in fragments $\leqslant 1$ kb in this collection of libraries as the calculated fraction of fragments bigger than 1kb is only 4.3% (Fig. 2).

Based on the above calculations, we proceeded with the construction of the libraries. Chromosomal DNA was prepared from late third instar wild type (Canton-S) *Drosophila melanogaster* and digested using the different sets of restriction enzymes as described above. Aliquots of digested DNA were fractionated and visualized on agarose gels containing Ethidium Bromide. Observed size distributions on agarose gels are in good agreement with theoretical calculations (data not shown). Portions of each restriction digest, without size fractionation, were ligated to appropriately digested 5′ and 3′ defined-sequence oligonucleotides as described in Experimental.

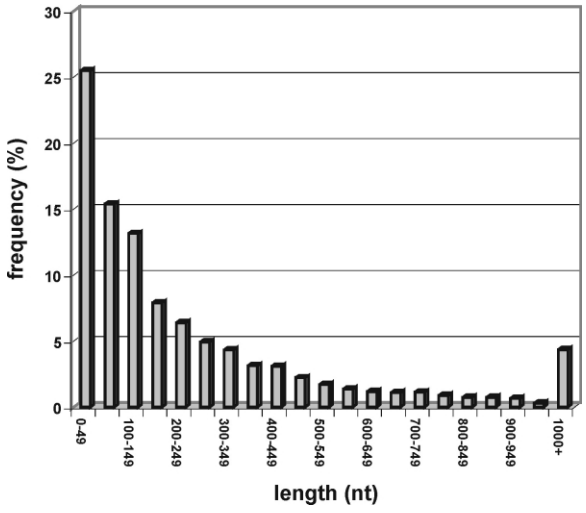

Figure 2. Theoretical prediction of genomic fragments size distribution. Random sequences covering over 200,000 pairs of bases were theoreitically digested with each set of restriction enzymes: Taq I, Dpn II, Taq I and Dpn II, Taq I and Bgl II, Taq I and Bam HI, Dpn II and Cla I, Cla I and Bam HI, or Cla I and Bgl II. The 5599 generated fragments were classified according to their length into 11 classes defined as 50 nucleotides steps and ranging from 0 to 49 nuleotides to over 1000 nucleotides. The frequency of each size class was calculated as the ratio of the number of fragments from a determinated class over the total number of fragments.

Strategy and in vitro selection of Tra2 binding RNAs

The candidate protein used in the selection was Tra2 overexpressed from a baculovirus construct in insect cells.[17] Nitrocellulose filter binding assays gave a dissociation constant (K_d) of 700 nM for Tra2 bound to its target site in the *dsx* female exon RNA. The K_d for Tra2 bound to the anti-sense of the same RNA was over 5×10^{-4} M, similar to the K_d for Tra2 and the starting RNA. Non-specific retention of female exon RNA on nitrocellulose filters in the absence of any Tra2 protein is rather high, about 40% of the input RNA.

The pool of RNAs generated from the initial genomic library was carried through 10 rounds of selection–amplification. Tra2-bound RNAs were selected by retention of the ribonucleo–protein complex on nitrocellulose. Selected RNAs were eluted from the nitrocellulose filter, copied into DNA using reverse-transcription then amplified using PCR. This new population was then ready for another cycle of selection. Enrichment of RNAs with high affinity for Tra2 was evaluate at each cycle by measuring the fraction of the RNA bound to the Tra2 (Fig. 3). The non-specific retention level was determined in the absence of protein.

The fraction of RNAs retained specifically in complexes with Tra2 increased from 0% measured at the first cycle to almost 30% measured at cycle 8 (Fig. 3), but non-specific binding also increased. To increase the specificity and quality of binding, prior to cycle 9 RNAs were subjected to counter selection against nitrocellulose binding in the absence of protein, and the stringency of the binding assays was increased in cycles 9 and 10. As a result, the fractions of RNA specifically binding to Tra2 increased drastically after the 9th round of selection

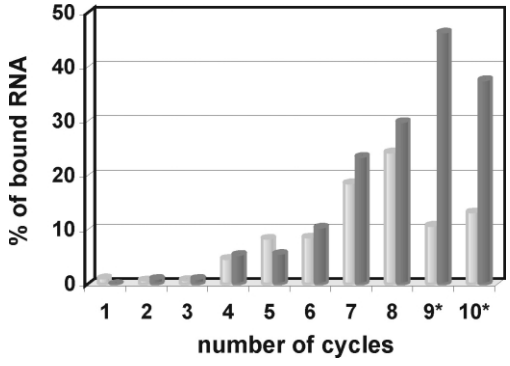

Figure 3. Progress of the selection of Tra2-binding RNAs. The graph shows the percentage of RNA retained on the nitro-cellulose filter (calculated as the ratio: RNA retained*100/total RNA used in the essay) as a function of the number of selection cycles. For each cycle, the light grey columns show the RNA retention level measured in absence of protein and the dark grey columns the RNA retention level measured in presence of Tra2. The latter were measured at a concentration of Tra2 protein of 10^{-5} M during cycles 1 to 4, 5×10^{-6} M during cycles 5 to 8, 1.7×10^{-6} M during cycle 9 and 6.7×10^{-7} M during cycle 10. The stars indicate the introduction of a counter-selection against non-specific nitrocellulose binding during selection cycles 9 and 10.

(Fig. 3). The fraction of RNA complexed to Tra2 at rounds 9 and 10 reached the same level seen with *dsx* female exon mRNA. A slight decrease in specific retention at the 10th round indicates likely completion of the selection procedure. cDNAs generated from the pool of RNAs from the 10th round were amplified by PCR, cloned into a bacterial vector and sequenced.

Selected RNA sequences

Over 100 clones were sequenced. Some cloned sequences were found in several copies but the complexity of the selected pool was quite high (Table 1). Based on sequence analysis, the sequenced clones issued from the selection can be grouped in four different families (Table 1).

Members of the family I, which represents 40% of clones from the final pool, contain of a track of uninterrupted adenines followed by a degenerate *dsx*RE motif: 8 out of 13 nucleotides are conserved, including the CAAAA sequence at the 3′ terminus of the degenerate repeat. Only minor sequence differences are observed between members of this family. These may represent PCR and reverse-transcription errors. This lead us to postulate that all clones are derivated from the same original sequence. In agreement with this hypothesis, this sequence, including the A rich sequence, corresponds to a single copy sequence from the sequenced Drosophila genome as shown by a BLAST search througth the drosophila genome (http://www.ncbi.nlm.nih.gov/blast/blast.cgi).

Sequences from family II, representing 10% of the selected population, share the same general organization as members of the first one: a track of adenines starting with a CAAAA sequence, followed by a different conserved sequence motif [GGGUAGCUGAUCAAUU$(A)_{5\ to\ 18}$CACUGAAUCGA]. This conserved sequence representes a single copy sequence within the sequenced Drosophila genome according to the result of the BLAST search (Table 1). However, because of sequence divergence between the family II clones. It os not certain that all clones are derivatives from a single sequence.

Family III (30% of the final population), is a collection of unrelated orphan clones. Two-thirds, but not all of them contain adenine tracks. Strikingly, each contains at least one copy of the motif CAAAA (Table 1). No other consensus sequence motifs are conserved among members of this group.

Family IV (about 20% of the final population), consist of repeated sequences derived from heterochromatic satellite DNA (Table 1). About 21% of the haploid genome of *Drosophila melanogaster* correspond to highly repeated sequences DNA located mainly in the centromeric heterochromatin.[18] Most of this family of Tra2 binding clones contains the sequence motif (AAUAACAUAG)n, corresponding to the known satellite DNA with the density 1686.[18] The remaining 15% sequences of this fourth family consist of (UG)n repeats. The (UG)n repeats have not yet been described in the literature but their sequences are in agreement with the general structure of the satellite DNAs.

From comparison of the sequenced clones, two general properties can be drawn. Nearly all the sequences share a very high adenine content, and, strikingly, the small motif CAAAA appears in all clones belonging to families I to III, representing 80% of the selected population.

Selected population may contain sequences from Tra2 regulated genes

In the course of sequencing clones from the selected population, no clones with the exact sequence of either the *dsx* female specific exon or *dsx*RE were found. To test if *dsx* was selected but not cloned, we performed DNA blots on cDNAs derived from each round of selection, using a 13-mer oligonucleotide containing the consensus sequence of the *dsx*RE. Figure 4A shows that a faint signal was detecTable after five rounds of selection–amplification for a population of fragments between 100 and 200 nucleotides. Its intensity increased as the selection proceeded showing a strong signal for two populations: one between 100 to 200 and the other between 300 and 450 nucleotides. The same blot was also probed with the plasmid containing the full-length *dsx* female-specific exon (Fig. 4B). This probe was used to ensure a higher specificity. A similar pattern was observed as a signal appears from the 5th round population and its intensity increased as the selection proceeds.

In a reverse experiment to verify futher the presence of the *dsx*RE or female exon in the selected pool, we made a DNA blot of the plasmid carrying the *dsx* female-specific exon and probed it with the PCR product from the 10th round of selection (Fig. 4C). A clear signal was observed with the *dsx* DNA further demonstrating the presence of the *dsx*RE sequence in the 10th round population (Fig. 4C).

These results show that the *dsx*RE sequences are present in the selected population. This provides an important positive control validating the SETIS strategy.

Discussion

Genomic libraries

When searching for biological targets of RNA binding proteins, genomic DNA libraries offer key advantages: (1) all sequences, including introns and transcribed regions downstream of poly A sites are included; (2) not only are all genes included, nearly all genes are represented at equal molar levels. We have developed a strategy for production and use of high representation genomic libraries for interative RNA selection. This strategy relies on digestion with appropriate combinations of restriction enzymes followed by ligation to oligonucleotides creating 5′ and 3′ fixed regions that flank each fragment and allow RNA expression, selection, and reamplification through multi-round selection pro-

Table 1. Selected RNAs sequences

Clone ID	Selected sequences[a]
	Family I
#A26	p5AGGCUAUAACUAGUAUAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAGUA- CAAGCAAAAAGGCCCAAGCCAAUUGCAGUAAACUUp3
	Family II
#122	p5CUGNGUCGUCNNUUAAAAACGUUAAAAACGUUNNCGCAAAAAAAAAAAAAGGCCNCNGNNGNGNNUCGGUGGNN- GAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAGGGUAGCUGAUCAAUUAAAAAAAAAAAAAACACUGAAU
#219	AUAAAUGUCAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAGGGUAGCUGAUCAAUUAAAAAAAAAAAAA CACUGAAUCGA
	Family III
#164	p5UGAGUCGAAAAAAAAAAAAAACAAAAAAAAAAAAAAAAAAAAAAAAAAAAACAAUAAUGCCUUUAAAAAUAACAAAAAGAGCCUAAAAAA- CUGCAAAAUGGAUUCCAAAAACCCCAAAAUCAAGCAAAUAAAAGUGCNNNGACNNUGCGUCAUGG
#218	p5CUGUGCGUCNNUGCCUUCGNGNUGCGUGNNGNNGNCUGCUUGAAAAUCAAAAAAAAAAAAAAAAAAAAAAAA- CAAAAAAAAAAAAAAAAAAAAAAAACAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAACUUCUUAUUUGUGUCACCUUAUCCCUU
#191	p5UNGNGUCGNNNCNNUGCNNNUUNCNNCNNCNNCAGGNNUNUGNUCNNCU- GAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAACUGUCAAUGAAAAGACAGCAAAAAAAAAAAAAAAAAAAAAAGU
#197	p5UGNUGCGUCNCUGCCUCGAAAAAAAUNUU- NUUAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAACAGAAGAAAAA- GAAUAGCUAAAAAACAAGUACCAUAAAACGCUGGCAAACAGAUAG
#199	p5UGNGUCGAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA- CAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAACACAAUAAAAUAAGAACGACAUUAAUUGAAUU- CUAUAGCAUACCAGCUAUGCAAGGUUUUCUCCAAUUGCUGAAU
#201	P5UGAGUCGAUAAAUCAAUUAAACAAAUUAAAUCAAUUCCACAUUUCUUNCUCGUAUCCUAUAAAAUAAUAAACCUGAUUAp3
#292	AAUAGCAGUGAGUAUACCCGNNNNNUGUCACUGCAAAAAAAAAAAAAAAAAAAAACAGACGGCCACAAAAACAGAAACACACAAGA- GUAAAUGCAAAAAAAAAAAAAAAAAAAAAAAAGUAAACAAAGGCAAAUAAACAAAAUACAAAUAAAAAUAACCGAAAUAAAGAAAAUAGAAAUAAAAA
	Family IV (satellite DNA)
#26	p5CUGAGUCGAUGUUAUUCUAUGUUAUUCUAUGUUAUUCUAUGUUAUUCUAUAUUAUUCUAUGUUAUUCUAUGUUAUUCUAUAUUAUUCUAU- GUUAUUCUAUCAUGGCCACCCANNGCUGNAUCAGUAC
#268	ANCNAACNCCAAAACCANNNNCNNAUGUGUGUGUGUGUGUGUGUGUGUGUGUGUGUGUGUGUGUGUGUGUGUGUGUGUGUGUGUGCCGAANGUU

[a]Only one clone from family I is shown as the minor sequence differences observed between members of this family were postulated coming from reverse-transcription and PCR. Few representative sequences are shown as typical example of sequences belonging to family II and III. Sequence from two clones containing either satellite sequence were arbitrarily choosen as an illustration of family IV sequences. CAAAA sequences are written in bold. Consensus sequences of family II, degenerate *dsx*RE of family I as well as the repeated motifs of the satellite DNA are underlined. P5 and p3 indicated that fixed sequence flanking the region of interest have been sequenced.

cedures. Because this strategy does not rely on organism-specific characteristics of the genomic DNA such as Alu repeats, CpG islands or methylation, it should be general to a wide range of organisms.

Prior to production of the libraries we modeled the expected size distribution of fragments and genomic representation based on a pool of actual Drosophila DNA sequences. This modeling of the fragment sizes helped us choose the sets of restriction endonucleases used. Using single and appropriate combinations of enzymes digest, we generated a population of 5×10^6 fragments with sizes between 20 and 1000 nt, with a majority of 50–400 nucleotide fragments. The same nucleotide sequence is potentially present in eight different contexts, corresponding to the eight combinations of enzymes used. No fragment size restrictions were applied to the starting pool so as not to bias the selection in a situation where little is known about the effect of fragment length or conformation on protein-RNA binding. Such choice in the library construction strategy contrasts with the choices of Gold and collaborators.[4,6] One of our goals was to allow the possibility that the generated RNAs could adopt their own natural secondary and tertiairy folding, ensuring the recognition of complex binding site by their partners proteins.

Relevance of selected sequences

Our results clearly show that it is possible to select and amplify very specific sequences from genomic DNA using this procedure. Over 50% of all cloned fragments sequenced after 10 cycles of selection and amplification are probably derived from just two single copy portions of the genome. Given the size of the fragments, this is an amplification of approximately one millionfold. This clearly indicates the ability of the technique to select individual pieces of the genome from amid the noise of the rest of the genome.

Also included within the sequenced population of clones were approximately 15% of clones corresponding to a single heterochromatic satellite sequence.[18] Although we used a DNA population enriched for euchromatic DNA, a number of cells containing a full complement of heterochromatic DNA were present in the tissues from which we made DNA, so it is not surprising that such a heterochromatic sequence might appear in our

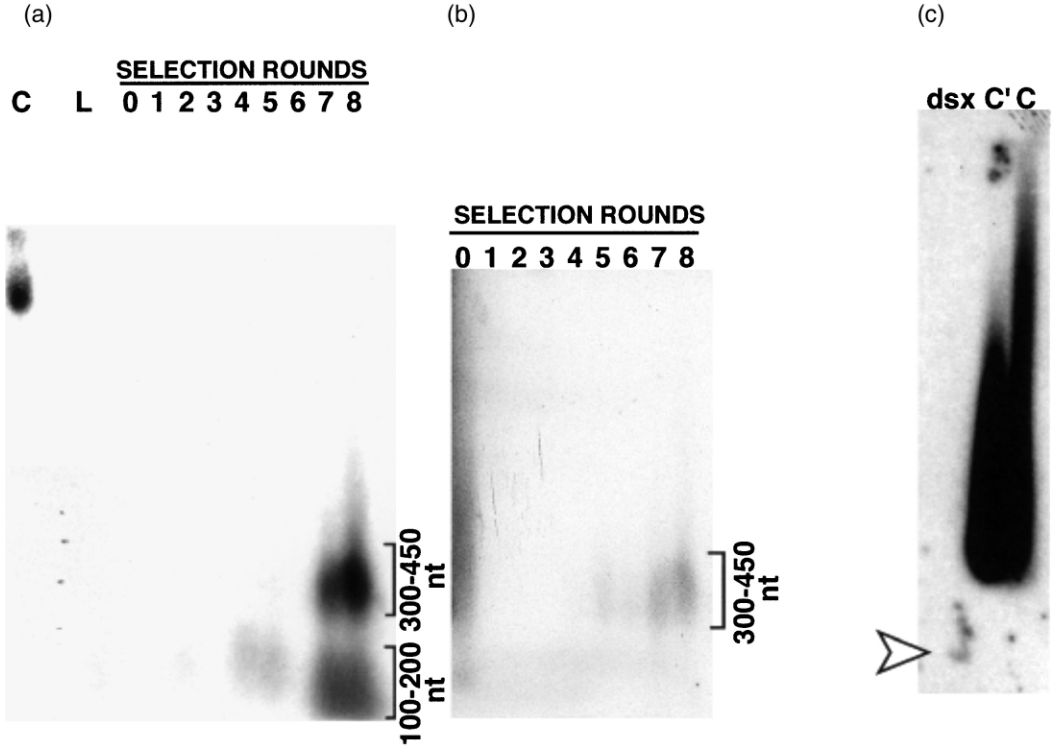

Figure 4. *dsx*RE may be present in the selected population. (A) Southern blot of the cDNA generated from each selection round (lane 0 to 8) probed with the *dsx*RE 13-mer oligonucleotide. From lane 4 onward, a faint signal appears for a cDNA population of an approximate size of 100 to 200 nucleotides (nt). From lane 7 onward, a clear signal is observed for the former population and for a population of cDNA ranging from 300 to 450 nt in length. A size ladder (lane L) and the plasmid pSK-BC (lane C) containing the *dsx* female specific exon sequence were run in parallel. The latter is meant as a positive control. (B) The same blot was probed with the plasmid pSK-BC labelled by random priming. The specific signal observed from cDNA population issued from the 5th to 8th cycles of selection (lane 5 to 8). This signal decreased at least 10 times compared to the signal observed when the blot was probed with the *dsx*RE 13-mer oligonucleotide. The lanes containing the size ladder and the plasmid pSK-BC are not shown. (C) Southern blot of the pSK-BC plasmid (lane dsx) probed with the cDNA population issued from the 10th round of selection. The cDNA generated from the 0th and 10th round selection (lane C and C′ respectively) were run on the same gel but were loaded latter because of the smaller size of these cDNAs. Lane C and C′ are meant as positive controls.

enriched population. It is not clear if isolation of this sequence after 10 rounds of selection indicates a relatively high affinity for Tra2, or if selection occurred based on only moderate affinity but a very high copy number. In any case, just as with the single copy sequences, only this defined satellite sequence was isolated while other major satellites did not contribute significantly to the final sequenced population, again indicating that the technique can select one sequence from among many.

We also demonstrated that a natural regulatory target of Tra2 action, *dsx*, was present in the population of amplified sequences and that its representation increased through selection. This serves as a positive control for the technique. Interestingly, *dsx* was not among the major products isolated after selection. At first this may appear surprising, but further consideration suggests that it makes sense and can inform future SETIS screens. The biology of *dsx* and *tra2* shows that Tra2 only acts on *dsx* in the presence of the female-specific protein Tra. In vivo, in culture, and biochemical experiments show that Tra and Tra2 proteins work together in binding to and regulation of *dsx* RNA (refs 14,17,19 and see below). In our selection, on the other hand, we used only the region around the RNP motif of Tra2 without Tra. Thus, it is likely that under these conditions *dsx* sequences will be of only moderate affinity for Tra2. Indeed, our test results show that the Tra2-*dsx* RNA complex has a K_d of about 10^{-6} molar. This contrasts, for example, with the affinity of another sex-specific splicing regulator, Sex-lethal, for its targets such as *transformer* and *Sex-lethal* RNAs: the complexes exhibit K_d values of about 10^{-9} molar.[20–22] Thus, for an RNA binding protein with a potentially large number of target RNAs, some of which are of interest and limited to a particular process, it may be important to include multiple members of an appropriate RNA-binding complex.

In addition to the major classes of isolated DNAs, a number of DNAs represented only once or a few times were found. Sequences comparison between these and the other selected clones reveals useful information about the RNA recognition site of Tra2. These comparisons confirm and extend two previous studies on the binding specificity of the *D. melanogaster* Tra2 protein and of the human Tra2 proteins alpha and beta.[23]

Every selected sequence from our clone families I, II and III exhibits the motif CAAAA (Table 1) close to the motif CAACA found in the 3′ half of the *dsx*RE. We propose that the motif CAA(c/a)A defined a possible primary binding site for Tra2. This is is in full agreement with results obtained by Lynch and Maniatis.[14] Crosslinking data demonstrate that Drosophila Tra2 binds to the 3′ part of the *dsx*RE[14] while Tra binds to the formed Tra2-RNA complex. Tra2 and Tra recruit different members of the SR family of splicing factors to the repeats and the PRE. Thus, Tra2, which contains a single RNA binding domain, can recognize distinct sequences in the repeats and the PRE in conjunction with specific SR proteins. Thus, the specificity of the regulatory complex could be determined by a combination of protein–protein and protein–RNA interactions. This would account for the diversity of Tra2 regulatory mechanisms: the activation of the female specific 3′ splice site in the case of *dsx* regulation, the activation of the 5′ splice site in the case of the *fru* female specific splicing and a potential intron retention in the case of *tra2* autoregulation. Moreover, ours results on the CAA(c/a)A motifs emphasizes the role of the CAA(c/a)A motif as part of the splicing enhancer. Coulter and collaborators[24] suggest the possibility that the *dsx* enhancer is a member of the A/C-rich splicing enhancers (ACES) family. This new class of exonic splicing enhancers was selected by in vivo selection that uses transient tranfection in an iterative procedure to select for exon sequences that enhance exon inclusion.[24] Of note their results also suggest the interesting possibility that the ACEs selected are targets for human Tra2.[24]

Another common feature of all or many of our clones in families I, II and III is the presence of poly or oligo A regions. Interestingly, mammalian Tra2 proteins are also sequence-specific splicing activators that likely participate in the control of cell-specific splicing patterns.[23] All 20 sequences selected in a SELEX procedure by Human Tra2 alpha contained extended stretches of poly(A).[25] In one case, the entire selected sequence consisted of a continuous poly(A) sequence of 24 adenosine residues.[25] Sequences selected by Human Tra2 beta showed greater variety. Although many of them also contained stretches of poly(A), others contained purine rich motifs or ACEs motifs.[25] Those results are in complete agreement with the results of SETIS procedure with Drosophila Tra2 and emphasize the interesting possibility that RNA binding specificity of those two Tra2 homologs are similar. In this regard it is worth noting that the human and Drosophila Tra2 RBDs are 54% identical.[23]

Conclusions

We have described a new procedure, SETIS, to identify potential natural targets for RNA binding proteins and pilot tested it with the Drosophila Tra2 protein. Our results show that the SETIS strategy allows selection and amplification of Tra2 binding targets from a genomic library. As part of our procedure we selected sequences containing a natural target of Tra2 action, the *dsx*RE. This validates our selection–amplification procedure and the procedure we used to give quality control over the complexity of our initial genomic library. Thus, SETIS is a useful and versatile strategy to screen for parts of biological networks of RNA-proteins interactions. It provides a tool to investigate arrays of processing and translational regulation events. It is complementary to methods like yeast two- and three-hybrid screening or RNA–protein or protein–protein crosslinking. Interestingly, it may also prove useful as screening strategy. Although not tested, the methods should be easily extended to screening 5′ and 3′ EST libraries or even to looking for RNA–RNA interactions.

Experimental

Transformer2 protein purification

Recombinant baculovirus expressing Tra2 (as the isoform containing 264 amino acids) (a gift from K. Lynch and T. Maniatis) and the purification of Tra2 were as described in Tian and Maniatis.[17] Purity of Tra2 was verified on SDS–PAGE gels. The protein concentration was measured via the dye binding procedure with Coomassie Brilliant Blue in phosphoric acid.[26]

Primers and plasmids

The sequences of the DNA primers used for the library construction are: 5′-CGACGATTAATACGACTCACTATAGGGCGGCCGCTCTGAGTCGATCATGCTAG-3′5, 5′-CTAGCATGATCGACTCAGAGCGGCCGCCCTATAGTGAGTCGTATTAATCGTCG-3′, 5′-GAATTCGATCATGGCCACCCAGGCCGCTG-3′, 5′-CAGCGGCCTGGGTGGCCATGATCGAATTC-3′. The sequences of reverse transcription and PCR DNA primers are 5′-CGACGATTAATACGACTCACTATAGGGCGGCCGCTCTGAGTCGA-3′, 5′-CAGCGGCCTGGGTGGCCAT-3′.

The pSK-BC plasmid contained the Bgl II- Cla I fragment of the female *doublesex* exon subcloned in the pSK bluescript derivative.[11]

Library construction

The SETIS technique is dependent on the production of RNA populations representing the single copy DNA complexity of the genome. Our strategy to produce such RNA populations is to create ''libraries'' of euchromatin-enriched genomic DNA fragments bounded by oligonucleotides of known sequence containing transcription initiation sites for phage RNA polymerases.

In order to obtain a DNA population enriched for gene rich regions, genomic DNA was prepared from third instar larvae. In such larvae, most of the tissue is composed of polyploid cells in which euchromatic DNA has been amplified many times with very little amplification of the highly repetitive DNA of the centric heterochromatin. Late third instar Canton S larvae were collected in the 'wandering' stage which immediately precedes pupariation. One thousand larvae were ground in 4 mL of grind buffer (24 mM Tris–HCl pH8, 10 mM Tris-base pH 9.2, 80 mM NaCl, 58 mM EDTA, 0.5% SDS and 160 mM sucrose) and 130 μL of a 1/5 dilution, in ethanol, of DEPC were added. The mixture was incubated for 30 min at 65 °C. After addition of potassium acetate to 1 M final concentration, the mixture was incubated 15 min on ice then centrifuged at 4 °C and 10,000 rpm for 15 min. The supernatant liquid was removed, extracted with phenol/chloroform (1:1) and ethanol precipitated twice. The pellet was resuspended in TE buffer (Tris–HCl 10 mM, EDTA 1 mM) and incubated with RNAse A (0.4 μg/μL). RNase A was removed by phenol-chloroform extraction and the aqueous phase was ethanol precipitated. The final DNA pellet was resuspended in 500 μL of TE at a concentration of 3.4 μg/μL.

Eight pools of sticky-ended genomic DNA fragments were generated by digestion with restriction enzymes. 17 μg samples of genomic DNA were digested with Taq I (20 units/μL), Dpn II (10 units/μL), Taq I and Dpn II, Taq I and Bgl II (10 units/μL), Taq I and Bam HI (20 units/μL), Dpn II and Cla I (5 units/μL), Cla I and Bam HI, or Cla I and Bgl II. The completion of the digestion was visualized on 0.6 and 1.6% agarose gel containing ethidium bromide.

Equimolar mixtures of the two oligonucleotides forming the 5′ or 3′ fixed regions were heated for 5 min at 80 °C, allowed to annealed at room temperature for 5 min and digested with Dpn II or Taq I restriction enzymes. Digested double-stranded fixed regions were purified after electrophoresis on a non-denaturing 15% acryl amide (1/30 acrylamide/bis) gel by elution into TE buffer.

Restriction enzyme digested 5′ and 3′ fixed regions were ligated to appropriately digested DNA sample pools in reaction mixtures containing 0.1 nmol of digested and purified 5′ and 3′ fixed regions and 1.7 μg of the compatibly digested genomic sample pools. Ligation mixtures were passed through TE Midi SELECT®-D, G50 columns (5 prime -3 prime Inc., Boulder) to remove the non-ligated fixed regions and ligation products smaller than 72 nucleotides.

The primary libraries generated as above were then amplified in PCR reactions containing one-fifth of the library, 2 pmol/μL of each PCR primer, 0.2 mM dNTPs, 50 mM KCl, 10 mM Tris–HCl, pH9 at 25 °C, 0,1% TritonX-100, 3 mM $MgCl_2$, 0.01U/μL Taq DNA polymerase. After treatment at 94 °C for 2 min, samples were subjected to 19 cycles of 94 °C for 1 min, 52 °C for 30 seconds and 72 °C for 3 min. Double stranded DNA was phenol/chloroform extracted, ethanol precipitated and further purified on a TE Midi SELECT®-D, G50 column.

These singly amplified libraries were then used to prime synthesis of the RNAs used in the first round of the selection process, with 20 to 50 μg of DNA used in a 100 μL T7 transcription reaction. These reactions were run in parallel with reactions in which 2 μg of DNA served as template in a T7 transcription reaction containing α^{32}P CTP (400 Ci/mmol) 50 μCi, in a 10 μL reaction. Transcription reactions were digested with DNase I to eliminate template DNA, transcripts separated from unincorporated ribonucleotides using a TE Midi SELECT®-D, G50 column, phenol extracted and ethanol precipitated. RNAs for subsequent rounds of selection were generated in reactions containing 10 to 30 μg of DNA in a 50 μL reaction, then purified as described above.

Selection

RNAs (a mixture of unlabeled and uniformly ^{32}P-labeled T7 transcribed RNAs as described above) were denatured for 3 min at 65 °C then renatured for 10 min at room temperature in 50 mM Tris HCl (pH 8), 145 mM KCl, 3.2 mM $MgCl_2$, 1 mM ATP with 50 μg/mL BSA

and 0.5 mg/mL yeast tRNA. Selections were done at a Tra2 protein concentration varying between 10^{-5} M (first round) and 6.7×10^{-7} M (10th round) in 60 μL; the RNA concentration was 10^{-7} M. The RNA–protein mixes were incubated 15 min at 30 °C and RNA–protein complexes separated from free RNA by filtration through a nitrocellulose membrane (Millipore, 0.45 μ, HAWP) presoaked in binding buffer. Unbound RNAs were rinsed off with 2 mL of 50 mM Tris HCl (pH 8), 145 mM KCl, 3.2 mM $MgCl_2$. Retention of RNAs on filters was measured by counting the << Cerenkov effect >>. RNA was recovered after elution from the filter as described by Tuerk and Gold.[2] Controls without Tra2 were performed at each round.

cDNA synthesis and subsequent rounds of selection

RNA recovered from the filter was subjected to reverse transcription (AMV reverse transcriptase, Life Sciences) using 'primer 3' in a 50 μL reaction mix for 30 min at 48 °C and the cDNA product was amplified with DNA polymerase and PCR as described above. The dsDNA amplified product was then transcribed as described above to produce an RNA pool for the next round of selection.

Tra2 filter binding assay

Filter binding reactions for determination of binding affinity were done in 50 mM Tris HCl (pH 8), 145 mM KCl, 3.2 mM $MgCl_2$, 1 mM ATP with 50 μg/mL BSA and 0.5 mg/mL yeast tRNA in a total volume of 10 μL. Binding reactions were incubated at 30 °C for 20 min, and then returned to ice. Reactions were diluted before loading with 90 μL of dilution buffer (50 mM Tris HCl (pH 8) 145 mM KCl and 3.2 mM $MgCl_2$) at 4 °C, and then filtered immediately through a nitrocellulose filter (Millipore, 0.45 μ, HAWP) on a filtration minifold. Filters were washed with 500 μL of dilution buffer. Retention of labeled RNA on filters was analyzed by liquid scintillation.

Cloning methods and molecular biology

Individual clones obtained after selection were sequenced after cloning into pGEM-T vector according to Promega's protocol. Sequencing was by the dideoxy chain termination method using the Sequenase Kit (United States Biochemical, manufacturer's protocol). Sequence analysis and assembly made use the DNA Star and DNA Strider packages.[27]

Drosophila culture, strains

Drosophila were grown on standard cornmeal, molasses, yeast, agar medium at 25 °C or room temperature. Mutations and chromosomes are as described in Lindsley and Zimm.[28]

Acknowledgements

The authors thank Phil Edeen and Kim Finley for helpful discussions and constant support, Steve Lodmell and Pascale Romby for critical reading of the manuscript, and Martin Broome for helping with insects cells culture. This work was supported by funding from Centre National de la Recherche Scientifique (CNRS) and by grants from the United States National Institutes of Health to M.McK. M.McK was a member of the U.S. N.I.H. Cancer Center at the Salk Institute.

References

1. Ellington, A. D.; Szostak, J. W. *Nature* **1990**, *346*, 818.
2. Tuerk, C.; Gold, L. *Science* **1990**, *249*, 505.
3. Gold, L. *J. Biol. Chem.* **1995**, *270*, 13581.
4. Gold, L.; Brown, D.; He, Y.-Y.; Shtatland, T.; Singer, B.; Wu, Y. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 59.
5. Marshall, K. A.; Ellington, A. D. In *Methods in Enzymology*, Vol. 318; Celander, D. W., Abelson, J. N.; Academic Press, San Diego, 2000, pp 193.
6. Singer, B.; Shtatland, T.; Brown, D.; Gold, L. *Nucleic Acids Res.* **1997**, *25*, 781.
7. Valcarcel, J.; Green, M. R. *Trends Biochem. Sci.* **1996**, *21*, 296.
8. Manley, J. L.; Tacke, R. *Genes Dev.* **1996**, *10*, 1569.
9. McKeown, M. *Dev. Genet.* **1994**, *15*, 201.
10. Nagoshi, R. N.; McKeown, M.; Burtis, K. C.; Belote, J. M.; Baker, B. S. *Cell* **1988**, *53*, 229.
11. McKeown, M.; Belote, J. M.; Boggs, R. T. *Cell* **1988**, *53*, 887.
12. Heinrichs, V.; Ryner, L. C.; Baker, B. S. *Mol. Cell. Biol.* **1998**, *18*, 450.
13. Lynch, K.; Maniatis, T. *Genes Dev.* **1995**, *9*, 284.
14. Lynch, K.; Maniatis, T. *Genes Dev.* **1996**, *10*, 2089.
15. MacDougall, C.; Harbison, D.; Bownes, M. *Dev. Biol.* **1995**, *172*, 353.
16. Madigan, S.; Edeen, P.; Esnayra, J.; McKeown, M. *Mol. Cell. Biol.* **1996**, *16*, 4222.
17. Tian, M.; Maniatis, T. *Science* **1992**, *256*, 237.
18. Ashburner, M. *Drosophila, A Laboratory Manual*; Cold Spring Harbor, New York: 1989.
19. Tian, M.; Maniatis, T. *Genes Dev.* **1994**, *8*, 1703.
20. Wang, J.; Bell, L. R. *Genes Dev.* **1994**, *8*, 2072.
21. Sosnowski, B. A.; Davis, D. D.; Boggs, R. T.; Madigan, S. J.; McKeown, M. *Dev. Biol.* **1994**, *161*, 302.
22. Singh, R.; Valcarcel, J.; Green, M. R. *Science* **1995**, *268*, 1173.
23. Dauwalder, B.; Amaya-Manzanares, F.; Mattox, W. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 9004.
24. Coulter, L. R.; Landree, M. A.; Cooper, T. A. *Mol. Cell. Biol.* **1997**, *17*, 2143.
25. Tacke, R.; Tohyama, M.; Ogawa, S.; Manley, J. L. *Cell* **1998**, *93*, 139.
26. Dunn, M. J. In *Protein Purification Methods. A Practical Approach*; Harris, E. L. V., Angal, S.), IRL PRESS, Oxford, 1989, pp 10.
27. Marck, C. *Nucleic Acids Res.* **1988**, *16*, 1829.
28. Lindsley, D. L., Zimm, G. G., In *The Genome of Drosophila Melanogaster*; Academic Press, San Diego, 1992.