Volume 16 Number 8 1988 Nucleic Acids Research

The YYRR box: a conserved dipyrimidine–dipurine sequence element in *Drosophila* and other eukaryotes

Douglas Cavener, Yue Feng, Brian Foster, Philip Krasney, Michael Murtha, Christopher Schonbaum and Xiao Xiao

Department of Molecular Biology, Vanderbilt University, Nashville, TN 37235, USA

Received December 2, 1987; Revised and Accepted March 9, 1988

ABSTRACT

We have discovered a novel DNA sequence element in Drosophila which is based upon a CTGA tandem repeat. This element has been named the YYRR box to emphasize its dipyrimidine-dipurine nature which is predicted to have unusual structural features. Southern hybridization analysis of genomic DNA indicates the presence of 25-30 copies of the YYRR box in each of three Drosophila species (melanogaster, pseudoobscura, and virilis) and conservation of genomic location within species. Similar analysis of human and rat DNA indicates the presence of YYRR related sequences in mammals as well. YYRR boxes have been localized to two genetic loci in Drosophila: Gld and a gene tentative identified as ted. These two genes exhibit correlated patterns of developmental expression and an identical mutant phenotype. Sequence analysis of the Gld YYRR box in three Drosophila species revealed a high degree of conservation despite its intronic location.

INTRODUCTION

Recent refinements in models of DNA structure and results of X-ray diffraction studies have stimulated a number of investigations into the structural and functional consequences of sequence dependent variation in the conformation of DNA (1-9). Calladine (1) proposed that large distortions in the archetypical B-DNA conformation are generated as a consequence of avoiding steric hindrance of nearest-neighbor purines on complementary DNA strands. Furthermore, he suggested four alterations in the DNA duplex structure which could potentially alleviate the purine-purine clash. Based upon the assumption that these alterations would be compensated by nearest-neighbor nucleotides, a simple quantitative model (denoted Calladine's rules) was proposed to describe the relative changes in these helix parameters (2). Dickerson and coworkers have shown that this model is able to predict such alterations in short oligonucleotide crystals (2). Tung and Harvey (8) have recently developed a more elaborate theoretical model to estimate absolute values for sequence dependent conformational parameters. Whether these two theoretical models are directly applicable to DNA in vivo is certainly not

clear. However, considerable evidence exists that alterations in the conformation of B-DNA strongly influence the recognition sites of various endonucleases (e.g. DNAase I)
(6,10-13)

We have discovered a conserved multicopy DNA sequence element in Drosophila which has the highest possible variance in the helix twist angle allowed by Calladine's rules. This sequence is 50-72 bp in length and is based upon a tetranucleotide dipyrimidine-dipurine motif. We have named this sequence the YYRR (pronounced wire) box to emphasize its dipyrimidine-dipurine nature. Two of the YYRR boxes have been localized within or near the developmentally related Drosophila genes, glucose dehydrogenase (Gld) and a transcription unit tentatively identified as trapped (ted). Gld and ted exhibit an identical mutant phenotype and are tightly linked (14,15). Furthermore, the expression of the mRNAs from Gld and the putative ted gene exhibit highly correlated developmental patterns. Their correlated pattern of expression may be related to the presence of the YYRR boxes.

MATERIALS AND METHODS

Drosophila DNA clones

Genomic DNA clones containing the Gld and ted genes for D. melanogaster were previously described (14). These clones were isolated from an EMBL-4 lambda library of the Oregon R. strain. An additional ted clone was isolated in this study from a Charon 4 lambda library of the Canton S strain. The D. pseudoobscura and D. virilis Gld clones were isolated from an EMBL lambda library and a Charon 30 library, respectively, via standard plaque hybridization techniques using D melanogaster Gld clones as the hybridization probes. The YYRR related sequences found in the stranded-at-second (sas) gene and the Antennapedia Complex (ANT-C) were identified from a series of genomic clones from two chromosome walks in the 84BC region (14,16) using the 16-mer GACT oligonucleotide probe desribed below.

DNA hybridization

Two synthetic oligonucleotides were used to screen Southern blots of various genomic clones and genomic DNAs for the YYRR sequences: a 16 mer, $d(GACT)_4$ and a 32-mer, $d(GACT)_8$. These will be referred to as GACT-16 and GACT-32, respectively. Hybridization experiments with ^{32}P end-labelled GACT-16 used the following conditions: 5X SSC, 0.1% SDS, 0.01% heparin, and 20% formamide at 25° C. Blots were washed in 2X SCC-0.2% SDS at 25° C. Hybridization experiments with GACT-32 were conducted at 42° C using one of

the following hybridization buffers: (Buffer A) 5X SSC, 0.1% SDS, 0.01% heparin and 50% formamide: (Buffer B) 0.25 M sodium phosphate, pH 7, 0.3 M NaCl, 0.01 M EDTA, 7% SDS, and 50% formamide. Blots were washed in 2X SSC-0.2% SSC-0.2%SDS first at 25° C and then at 60° C.

DNA cloning and sequencing

Nested deletions of several DNA clones described below were constructed using the Henikoff ExoIII method (17). D melanogaster Gld YYRR box: A series of ExoIII deletions were constructed in pEG25, a pEMBL18 subclone of a 7 kb Sal-Xba fragment from intron-I. D. melanogaster ted YYRR box: 340bp AccI fragments (pCT40) from genomic clones of Oregon-R and Canton-S were subcloned into the Bluescript vector (Stratagene, Inc.) D. melanogaster sas YYRR sequence: A series of ExoIII deletions were constructed in pCS58, a Bluescript subclone of a 2.5 kb EcoRI fragment from intron-I of the sas gene. D. melanogaster ANT-C YYRR sequence: a series of ExoIII deletions were constructed in pCX62, a Bluescript subclone of a 2.4 kb EcoRI fragment located between the divergent promoters of the ANT-C genes Sex combs reduced (Scr) and fushi tarazu (ftz) (Dr. T. C. Kaufman, personal communication). D. pseudoobscura Gld YYRR box: A series of ExoIII deletions were constructed in pBy2.1, a Bluescribe (Stratagene, Inc.) subclone of a 2.1 kb Sst I fragment from intron I. D virilis Gld YYRR box: a 350 bp PvuII-SstI fragment was subcloned into Bluescript (pCG79Ω). After verifying the location of YYRR sequences within the subclones and deletion mutants using Southern analysis, their DNA sequences were determined using the Sanger dideoxy sequencing method (18). Analysis of the DNA sequences was performed on a Compaq 286 computer using the IBI Pustell DNA sequence analysis programs.

RESULTS

D. melanogaster Gld YYRR box

DNA sequence analysis of the 8 kb intron separating the start site of transcription from the start codon of Gld revealed an unusual 72 bp tetranucleotide tandem repeat element (Fig.1). This repeat is based upon the sequence CTGA. Single nucleotide differences in the CTGA motif exist in seven of the eighteen repeats. Interestingly, only two of these differences disrupt the dipyrimidine-dipurine (YYRR) nature of the repeat. Although these two repeats (YYYR and RRRY) also destroy the simple dyad symmetry (YYRR/RRYY), they give rise to a higher order palindromic pattern (YY R YYY RR / YY RRR Y RR). Because YYRR sequences have a high density of YR and RY base steps, high levels of variation were found for all four helix parameters (helix twist

D. melanogaster Gld YYRR BOX

CTGA CTGG CTGA CTGG CTGA CTGG CTGA CTGA CTGC CTGA CCGA GTGA CTGA CTGA CTGA CTGA CTAA CTGA

$(YYRR)_8$ YY R YYY RR | YY RRR Y RR $(YYRR)_6$

2 1 3 2 2 3 1 2

D. pseudoobscura Gld YYRR BOX

CTGA CTGA CTGA CTGA CTGG CTGA CTGA CTGA CTCA CTAA CTGA CTGA TTGA CTGA CTGG CAGA CTGA

$(YYRR)_8$ YYY R (YY | $RR)_6$ Y RRR YYRR

3 1 2 2 1 3

D. virilis Gld YYRR BOX

AA CTGA CTCA CTGA CTCA CTGA CTGA CTGA CTGA CGGA CTGA CAGA CTGA

RR YYRR YYY R YYRR YYY R (YY | $RR)_4$ Y RRR YYRR Y RRR YYRR

3 1 2 2 3 1 2 2 1 3 2 2 1 3

D. melanogaster ted YYRR BOX

AA CTGA CTGA CTGT CTCA CTGC CTCA CTGA CTGA CTGA CTGA CTGC CTGA ATGA CTGA CTGA CTAA CTGA

$(RRYY)_3$ R Y_4 R YY R Y_4 R $(YYRR)_4$ Y_2 R Y_3 | R_3 Y $(RRYY)_4$ RR

2 1 4 1 2 1 4 1 2 2 1 3 3 1 2

Figure 1. Primary DNA sequences and derived symmetries of Gld and ted YYRR boxes. All of the sequence symmetries observed for the Gld gene are palindromic. The axis of the dyad summetry is indicated by " "in the second line for each sequence. A direct repeat is observed in the ted YSSR box. The ted YYRR box was sequenced in both the Canton-S and Oregon-R strains of D. melanogaster and found to be identical.

angle, base-plane roll angle, main-chain torsion angle, and propeller twist angle) as calculated by the sum rules of Dickerson (2). According to these rules, alterations in the helix twist angle and base-plane roll angle are compensated for by opposite changes in the same angles in adjacent base pairs. The YYRR sequence is unique in this regard because it contains the highest possible packing of YR and RY steps which generate the clash and YY and RR steps which compensate for the clash. As a consequence, YYRR tandem repeats generate the highest possible variance in the helix twist angle as computed by Calladine's rules. The expected deviations and variance in the helix twist angle for the Gld YYRR box and flanking sequence are depicted in Figure 2. The central symmetrical pattern in Figure 2a is due to the higher order

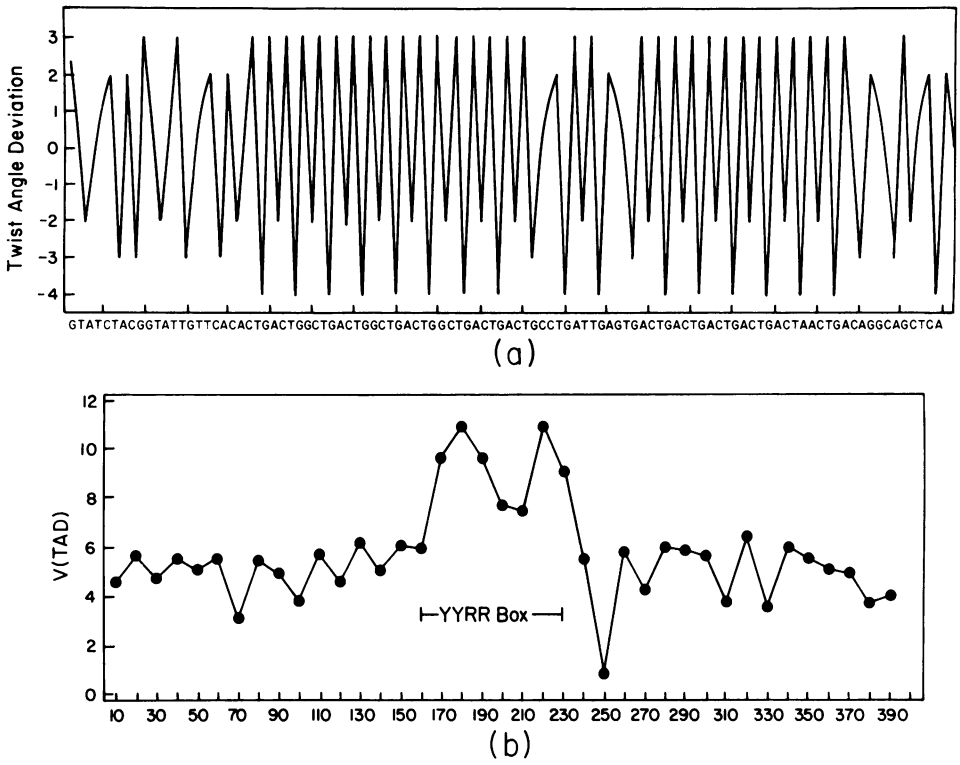
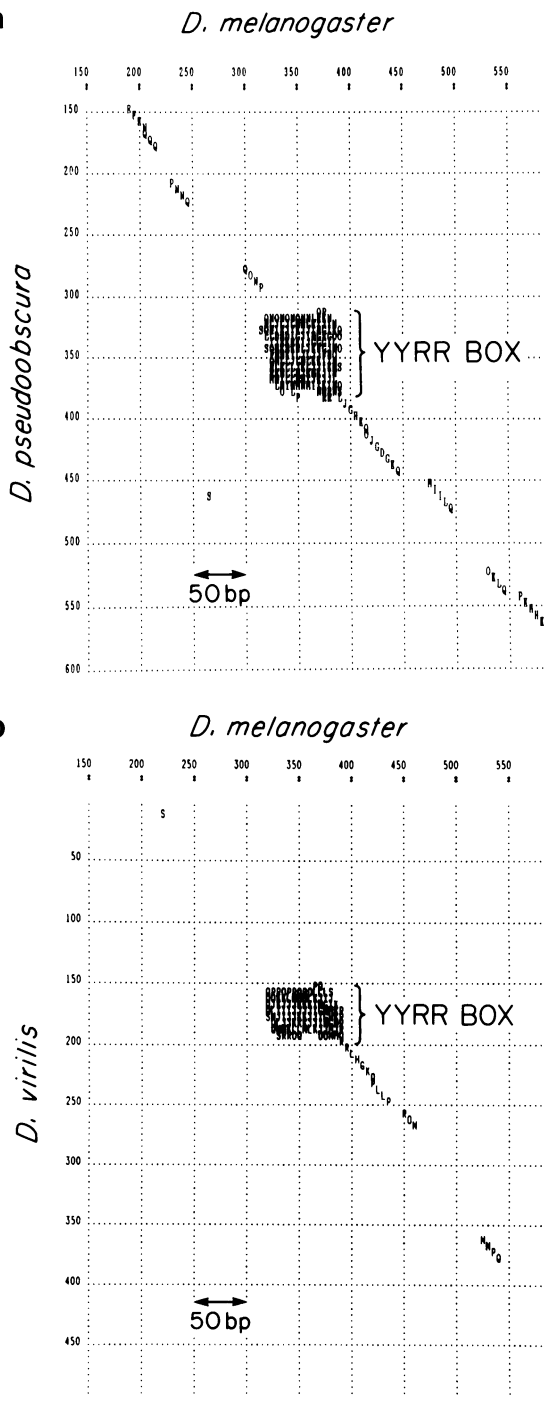


Figure 2. Helix twist plot for the D. melanogaster Gld YYRR box.
(a) The expected deviations in the helix twist angle were calculated using Caladine's rules. Zero is approximately equal to an angle of 36° (i.e. the mean angle for B-DNA). Each unit is approximately equal to 2° (Dickerson, 1983). Thus, the predicted range in the helix twist angle is 28° to 42°.
(b) The variances of the twist angle deviations, V(TAD), are plotted for the Gld YYRR box and flanking sequences. Each data point represents the variance calculated from ten adjacent angles. The two highest peaks in the variance represents the YYRR box (bases160-230). The intervening valley between these two peaks is caused by the palindromic sequence of the YYRR box which has a lower density of purine-purine clash than the simple dipyrimidine-dipurine motif.

palindromic sequence described above. The tight packing of helix twist deviations in the YYRR box leads to a two fold higher variance in this parameter than that observed in flanking sequences (Figure 2b). Using the methods of Tung and Harvey (8) the following series of helix twist angle is predicted for the basic CTGA tandem motif CT-34.9°, TG-27.7°, GA-36.6°, and AC-36.4°. Although both models predict a similar sharp reduction for the TG step twist angle, the Tung-Harvey model does not predict the large positive


a
D. melanogaster
D. pseudoobscura
YYRR BOX
50 bp
b
D. melanogaster
D. virilis
YYRR BOX
50 bp

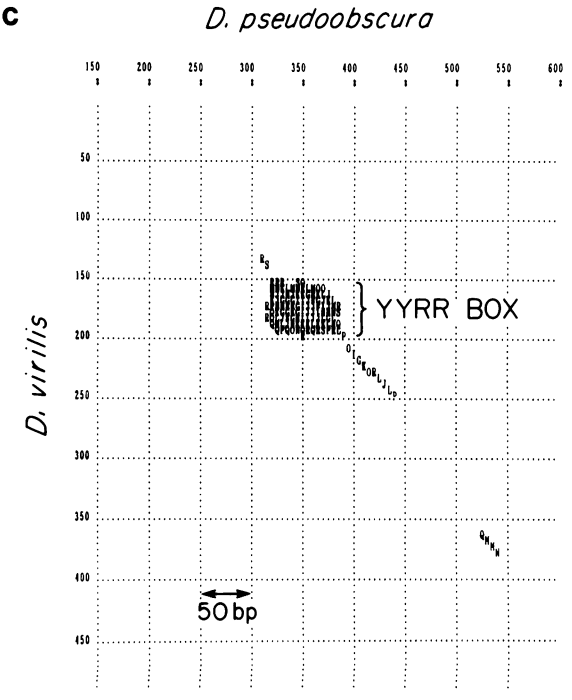


Figure 3. Homology matrices for Gld YYRR boxes in three Drosophila species. The Pustell - IBI DNA Sequence Analysis program was used to compare the nucleotide sequence of D. melanogaster. D. pseudoobscura. and D. virilis. (a) melanogaster & pseudoobscura. (b) melanogaster & virilis, (c) pseudoobscura & virilis. Homology between sequences is indicated by letters which represent the degree of sequence similarity within a centrally-weighted region of 31 nucleotides. Sequence similarity ranges from 100% (letter A) to a lower cutoff of 65% (letter S). The three sequences containing the YYRR boxes are all located within the large 5' intron of the Gld gene. The orientation (top to bottom and left to right) corresponds to the 5' to 3' orientation of the Gld transcription unit.

deviations in the CT and step angles. Consequently the helix twist variance within the YYRR box is not greater than variance in the sequences flanking the YYRR box as calculated by the Tung-Harvey method (data not shown) in contrast to Figure 2b (Calladine helix twist variance).

Gld YYRR box in D. pseudoobscura and D. virilis

Genomic DNA clones of the D. melanogaster Gld gene were used as hybridization probes to isolate the homologous clones in D. virilis and D. pseudoobscura from genomic DNA libraries (Krasney and Cavener, unpublished data). YYRR boxes were found to exist in the homologous regions of the Gld

genes of these two species as well. The sizes of the Gld YYRR boxes are similar in the three species (D. melanogaster = 72 bp. D. pseudoobscura = 68 bp. and D. virilis = 50 bp). A total of eight mismatches to the dipyrimidine-dipurine motif is observed among the three Gld YYRR boxes, and all of these are part of higher order palindromic sequences. However, each one of the YYRR boxes contains a different palindromic pattern. One common feature of the palindromes is that tripyrimidines are found to the left of the axis of symmetry whereas tripurines are found to the right. Also, a tripyrimidine is present in the same position of all three relative to the 3' boundary of the YYRR box.

The Pustell-IBI DNA Sequence Analysis program was used to examine the conservation of the sequences flanking the Gld YYRR box in the three Drosophila species. The sequences immediately downstream of the Gld YYRR boxes are highly conserved among all three species (Fig. 3). Interestingly, this region is responsible for the low variance in the helix twist angle immediately to the left of the YYRR box in Figure 2b. A more extensive region of homology is observed in the 300 bp region surrounding the YYRR box between D. melanogaster and D. pseudoobscura. It is important to note that we have examined this itronic region in D. melanogaster and find no evidence for protein coding regions.

YYRR sequences in other regions

Two oligonucleotide probes, GACT-16 and GACT-32 were used to detect YYRR sequences among several previously cloned D. melanogaster genes. GACT-16 and GACT-32 are four and eight tandem repeats of GACT, respectively. We chose hybridization conditions which would allow for approximately two mismatches using GACT-16 and five mismatches using GACT-32. All DNA sequences examined were probed in parallel with the two oligonucleotides. We have chosen to use the general term YYRR sequence to denote sequences which hybridize with at least the GACT-16 probe and the specific term YYRR box to denote those sequences which also hybridize with the GACT-32 probe.

The following regions of the genome were examined for YYRR sequencs by probing Southern blots containing genomic DNA clones: (a) 84C region containing nine genes including Gld, 160 kb; (b) 84AB region which contains the Antennapedia Complex, 320 kb; (16), (c) 89E region which contains the Bithorax Complex, 400 kb (19.20), (d) 25F-26A region which contains the α-glycerophosphate dehydrogenase and β-galactosidase genes as well as several other genes, 195 kb (D. Knipple & R. MacIntyre, personal communication): and (e) 74E region which contains the E74 gene, 100 kb (C. Thummel & D. Hogness,

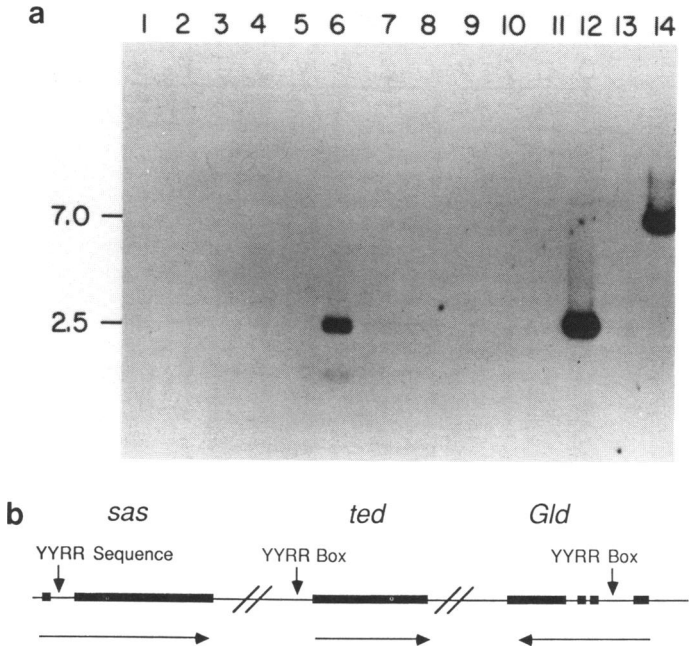


Figure 4. YYRR sequences in the 84C region of D. melanogaster. (a) Overlapping genomic DNA clones (152 kb) containing the entire 84C region of the right arm of the third chromosome (Cavener, et al, 1986a) were hybridized with the GACT-16 probe. Clone 6 contains the YYRR sequence within an intron of the sas gene. Clones 12 and 14 contain the ted and Gld YYRR boxes, respectively. (b) The location of the YYRR sequences/boxes relative to the transcript maps of sas, ted and Gld. The thick lines represent the approximate location of exonic sequences. The precise boundaries for the exonic sequences have only been determined for exons I, II, and III for the Gld gene.

personal communication). Using the GACT-16 probe, a total of 15 YYRR sequences were detected among the 1.2 megabases of DNA examined. The results for the 84C region are shown in Figure 4. Only two of the 15 YYRR sequences were also detected by the GACT-32 probe. Besides the YYRR box at the Gld locus the other YYRR box detected by GACT-32 maps either within or near the trapped (ted) locus. ted and Gld are tightly linked and reside within a 100 kb region. (14,15). The location of the ted YYRR box relative to the ted transcription unit has not been precisely determined. Preliminary evidence indicates that this YYRR box is most likely within a 5' intron of the ted gene.

The sequence of the ted YYRR box is shown in Figure 1. Its length is essentially the same as that of the D. melanogaster Gld YYRR box. The ted

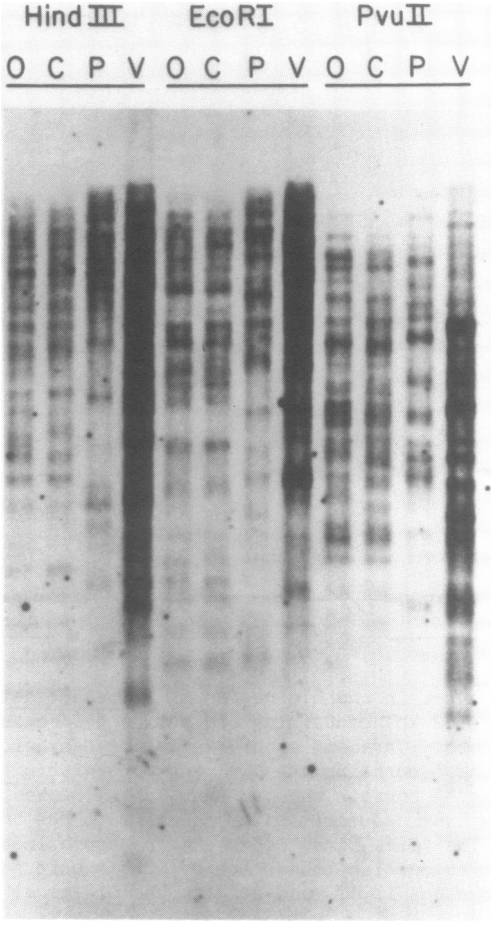


Figure 5. YYRR boxes in the Drosophila genome. Genomic Southern analysis was performed on total genomic DNAs of D. melanogaster, D. pseudoobscura, and D. virilis using the GACT-32 probe. O = D. melanogaster Oregon-R strain, C = D. melanogaster Canton-S strain. P = D. pseudoobscura and V = D. virilis. High stringency hybridization and washing conditions were used to eliminate the detection of YYRR sequences with less than a 25 of 32 match with GACT-32. The filters were exposed to XAR-5 film for 15 hours at -70 C°.

YYRR box contains six mismatches to the YYRR motif. Four of these mismatches are within a tandem repeat (YY R YYYY R, YY R YYYY R) while the other two mismatches impose a higher order palindromic sequence much like the palindromes observed for the Gld YYRR boxes of D. melanogaster, D. pseudoobscura, and D. virilis. Another striking feature of all four YYRR boxes which have been sequenced is that their four base repeat element is

uninterrupted by any insertions or deletions which would disturb the four base frame.

Given the heterogeneity of the symmetries observed among the four sequenced YYRR boxes, we decided to determine if the ted YYRR box sequence differed in another D. melangoaster strain. We sequenced the ted YYRR box from the Canton-S strain and found that it was identical to that of the Oregon-R strain.

In order to compare the YYRR sequences which only are detected by the GACT-16 probe, we sequenced the YYRR sequences detected in the sas gene (formerly denoted 1(3)84Cd) (14,15) and in the ANT-C between the Scr and ftz genes. These YYRR sequences contain four or five tandem repeats of the GACT motif with matches of 15 (sas) and 14 (ANT-C) bases to the GACT-16 probe (data not shown). The significance of these short YYRR sequences is questionable, inasmuch as their probable occurence in the Drosophila genome approaches random expectation. Nonetheless, the clones containing them have proven to be useful negative controls for experiments conducted with GACT-32. The best match for each of these two YYRR sequences to the GACT-32 oligomer is 21 of 32 (ANT-C) and 20 of 32 (sas).

Copy number and conservation of location

Whole genomic Southern hybridization indicated that the YYRR box copy number ranges from 25-30 for D. melanogaster, D. pseudoobscura, and D. virilis (Figure 5). Furthermore, the locations of YYRR boxes within D. melanogaster appear to be highly conserved between the Oregon-R and Canton-S strains (which were isolated from different populations more than fifty years ago). This result is in contrast to the lack of intraspecific conservation in location of several other middle repetitive sequences in D. melanogaster (21).

Southern hybridization experiments were also conducted with genomic DNA from rats and humans (Figure 6). Humans have four distinct PvuII fragments (8.7 kb, 2.8 kb, 1,2 kb, and 0.7 kb) which hybridize with GACT-32. This restriction pattern is conserved among eight unrelated human DNA samples. The YYRR copy number in rats is difficult to estimate because one of the restriction fragments containing a YYRR sequence exhibits a very high signal (thus obscuring the signal from other fragments). Analysis of five diverse rat tissues uterus, intestine, liver, heart, and spleen) indicates that the 6.5 kb PvuII fragment has very high signal in all tissues. It is important to note that under these hybridization conditions, high concentrations of restriction fragments containing the short sas and ANT-C YYRR sequences are not detected at all with GACT-32. Thus the genomic restriction fragments

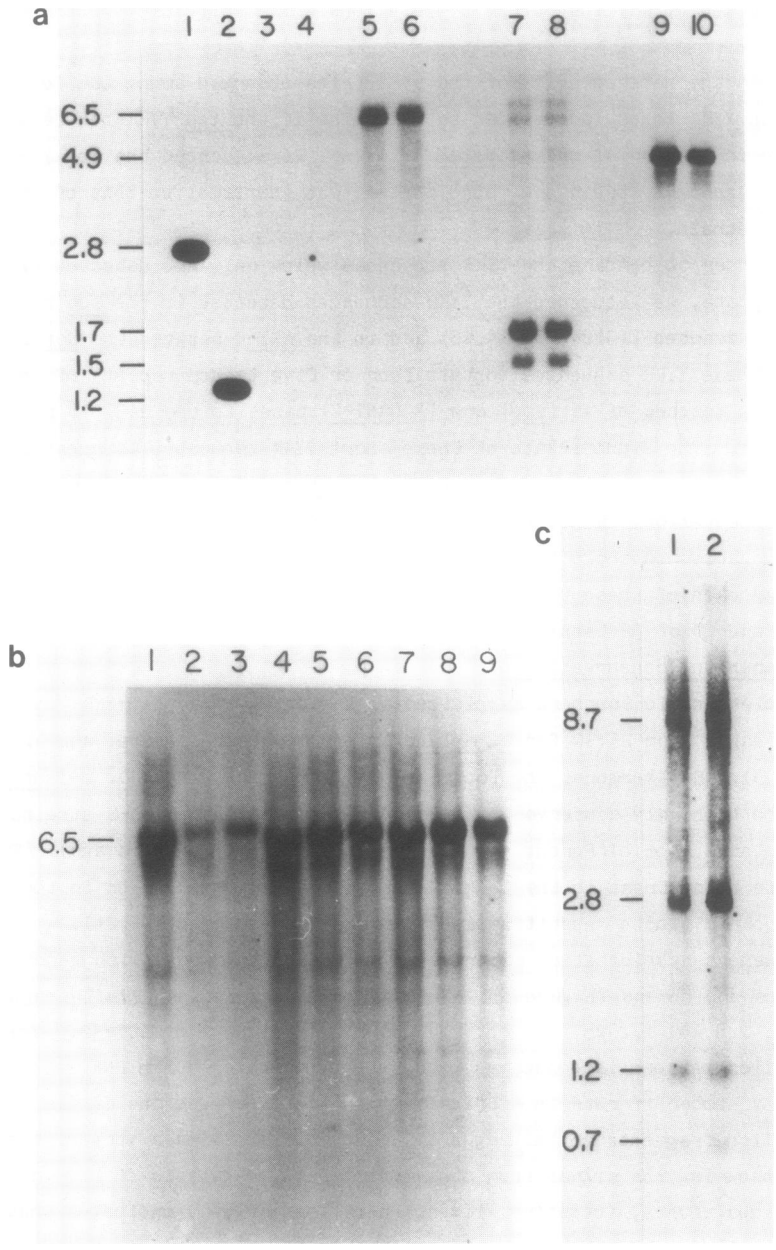


Figure 6. Human and rat YYRR box related elements. Total genomic DNAs were subjected to Southern analysis using the GACT^{32} probe and high stringency hybridization conditions. (a) Lane 1, cloned ted YYRR box; Lane 2, cloned Gld YYRR box; Lane 3, cloned sas YYRR sequence; Lane 4, cloned ANT^{3C} YYRR

sequence; Lanes 5, 7, & 9 rat uterus^{0} DNA (20 µg) digested with PvuII, HindIII, & PstI, respectively: and Lanes 6, 8, & 10 rat uterus^{24} DNA (20 µg) digested with PvuII, HindIII, and PstI, respectively. Rat DNA was isolated from Sprague Dawley rat uteri either untreated, uterus^{0}, or treated for 24 hours with estradiol, uterus^{24}. Lanes 1 & 2 contain the two positive controls (ted and Gld YYRR boxes) with the predicted 2.8 kb and 1.2 kb fragments, respectively, hybridizing with the probe. Lanes 3 & 4 contain the two negative controls (sas) and ANT-C YYRR sequences) which fail to hybridize to GACT-32 as predicted under these hybridization conditions. The restriction fragments containing the control YYRR sequences (Lanes 1-4) contain between 20-60 ng of DNA. Exposure time was for 6 hours. (b) Rat genomic DNA from various tissues digested with PvuII. Lane 1, spleen (15 µg); lanes 2 & 3, liver (7 µg); lanes 4 & 5 heart (15 µg); lanes 6 & 7 intestine (15 µg); and lanes 8 & 9 uterus (15 µg). (c) Lanes 1 & 2 contain DNA samples (20 µg) isolated from peripheral lymphocytes from two different human individuals both digested with PvuII. Exposure time was for 40 hours. The four PvuII fragments detected have also been seen in six other individual samples (data not shown).

which are detected in these experiments must have a match of greater than 21 of 32 with GACT^{4}32. Given the high stringency conditions used, we estimate that the percent match is greater than 80%.

DISCUSSION

Tandem repeats of short oligonucleotides are common to a variety of eukaryotic taxa and have been studied extensively in D. melanogaster and D. virilis (22). Although the simple sequence complexitiy of the YYRR box is similar to these previously reported tandem repeats, there are some important differences germane to their size, genomic location, and evolutionary conservation. A large majority of the previously characterized tandem repeats occur in large blocks (10^{5}-10^{7} copies per clusters), and are typically located in heterochromatin. As a consequence of their length and GC content bias, they are often isolated as satellite DNA. Phylogentic studies indicate that these sequence elements evolve rapidly such that satellites in closely related species are quite dissimilar and in some cases may have independent origins. In contrast, the YYRR box repeats sequenced to date are relatively small (50-72 bp), are found within euchromatic genes, and exhibit a much greater degree of evolutionary conservation of size, location, and copy number both within and between species.

Three other tandem repeats are more similar to the YYRR box than the satellite DNAs mentioned above. The Opa sequence is a trinucleotide repeat found in the coding region of a number of Drosphila genes (23). Given its location in protein coding regions, its functional consequence is more obvious than that of the YYRR box which apparently does not occur in any mRNA (Cavener and Cox, unpublished data). The other two tandem repeats, found in

immunoglobulin heavy chain genes and chorion genes, are found near or within their transcription units but not within coding regions. Immunoglobulin heavy chain genes contain pentanucleotide tandem repeats of GAGCT upstream of the alternative C exons (24). This repeat is thought to be essential for the DNA rearrangements which give rise to immunoglubulin class switching. The Drosophila chorion gene cluster also contains a pentameric DNA tandem repeat. This repeat has been implicated in the tissue specific amplicification of the chorion gene clusters on both the X and third chromosomes (25). Presently, we have no evidence that a DNA rearrangement or amplification occurs in the region of the Drosophila YYRR boxes. However, the small size of Drosophila makes it technically difficult to rule out tissue-specific DNA rearrangements.

The conservation of the Gld YYRR box in D. melanogaster, D. pseudoobscura, and D. virilis strongly suggests that the YYRR box has a function. Given the estimated time of divergence between the Sophophora subgenus (containing D. melanogaster and D. pseudoobscura) and the Drosophila subgenus (containing D. virilis) (26, 27) unconstrained DNA sequences in intronic and flanking regions surrounding genes are expected to be completely diverged between species from the two main subgenera. Indeed, comparisons of intronic sequences between D. melanogaster, D. pseudoobscura, and D virilis for the Hsp¬82 gene reveal complete divergence using the identical program and parameters values we have used herein (28). Thus, the conservation of the Gld YYRR box and the first 50 bp immediately downstream of it (Figure 3) implies a conserved function. This hypothesis is further supported by the presence of a YYRR box associated with the ted gene which exhibits an identical mutant phenotype and has a very similar pattern of mRNA expression to Gld (Cox & Cavener, unpublished data). The Gld gene in D. pseudoobscura and D. virilis are not abundantly expressed in adult males as is the case for the Gld gene in D. melanogaster, so it is unlikely that the YYRR box is involved in adult expression. The most compelling hypothesis at this time is that the YYRR box serves as an enhancer or DNA amplification element which affects the developmental expression of Gld, ted, and other related genes during preadult development. If this hypothesis proves to be correct, then it is likely that the 25-30 copies of the YYRR box in Drosophila are contained within a large family of developmentally related genes. We have begun in vitro mutagenesis - gene transformation experiments to test the hypothesis that the YYRR serves a regulatory function in the Gld gene.

Although it is tempting to speculate about the functional relationship of the YYRR box related sequences found in fruit flies, humans and rats, there

are some important differences observed. Firstly, the copy number of the YYRR boxes in humans is an order of magnitude smaller than that of Drosphila. However, like Drosophila, intraspecific variation in the restriction fragments detected by GACT-32 is virtually absent suggesting that the YYRR box is a non-mobile element and evolutionarily stable. Secondly, rats contain a highly amplified YYRR sequence. But this amplification is not tissue specific, and therefore, it is unlikely that the amplification is involved in developmental regulation.

In addition to functional and evolutionary questions, the potential structure of the YYRR box is of considerable interest because it represents an extreme case of perturbation in the helix twist angle as estimated by Calladine's rules. Although the estimates of the helix twist angles in the YYRR box by the method of Tung and Harvey (8) predicts less extreme average deviations, it is interesting to note that the predict four angle series for the YYRR box (34.9°, 27.7°, 36.6°, and 36.4°) is very similar to the consensus five angle series (36.2°, 37.8°, 26.4°, 37.8° and 36.2°) found in E coli promoters (29). Both the YYRR box and the E coli consensus structure contain a YR step leading to a large decrease in the twist angle.

The notion that conformational effects of the dipyrimidine - dipurine nature of the YYRR box are more important than the primary sequence is supported by the pattern of mismatches to the CTGA motif. None of the mismatches to the CTGA motif give rise to symmetries observable in the primary sequence, whereas among the four sequenced YYRR boxes all fourteen of the mismatches to the YYRR motif are organized into palindromic symmetries (10 mismatches) or direct repeats (4 mismatches in the ted YYRR box). Although this striking observation implies that the palindromic sequences serve a function, it is curious that the palindromes are not precisely conserved for the Gld YYRR box among the three Drosophila species. One feature which is conserved among the three is the presence of tripyrimidines and tripurines to the left and right, respectively, of the axis of symmetry. This is also a characteristic of the ted YYRR box palindrome.

A number of investigators have suggested that helix structure variants may serve as recognition sites for specific regulatory proteins or proteins involved in DNA packaging (2,5-9,30). We have obtained preliminary results which indicate that Drosophila embryo extracts contain a protease sensitive factor(s) which strongly binds to a 217 bp restriction fragment containing the Gld YYRR box. Experiments are underway in our laboratory to characterize this factor and determine whether the basis of its recognition is dependent upon the primary sequence and/or the conformation of the YYRR box.

ACKNOWLEDGEMENTS

We thank Thomas Kaufman, Welcome Bender, Carl Thummel, and Ross MacIntyre for providing us with genomic clones or Southern blots of clones for four region of the D melanogaster genome. We are grateful to Charles Singleton for pointing out the possible structural significance of the YYRR box and to Fritz Parl for providing us the rat and human DNAs.

REFERENCES

1. Calladine, C.R. (1982). J. Mol. Biol. 161, 343-352..
2. Dickerson, R.E. (1983). J. Mol. Biol. 166, 419-441.
3. Kabsch, W. Sander, C. & Trifonov, E.N. (1982). Nucl. Acids Res. 10,1097-1104.
4. Fratini, A.V., Kopka, M.L., Drew, H.R. & Dickerson, R.E. (1982). J. Biol. Chem. 257, 14686-14707.
5. Nussinov, R., Shapiro, B., Lipkin, L.E. & Maizal, J.B. (1984a). Bioch. Bioph. Acts 783, 246-257.
6. Nussinov, R., Shapiro, B., Lipkin, L.E. & Maizal, J.B. (1984b), J. Mol. Biol. 177, 591-607.
7. Tung, C.S., & Harvey, S.C. (1984). Nucl. Acids Res. 12, 3343-3356.
8. Tung, C.S., & Harvey, S.C. (1986). J. Biol. Chem. 261, 3700-3709.
9. Tung, C.S., & Harvey, S.C. (1986). Nucl. Acids Res. 14, 381-388.
10. Drew, H.R., & Travers, A.A. (1984). Cell 37, 491-502.
11. Drew, H.R. (1984). J. Mol. Biol. 176, 535-557.
12. Lomonosoff, G.P., Butler, P.J.G., Klug, A. (1981), J. Mol. Biol. 149, 745-760.
13. Keene, M.A. & Elgin, S.C.R. (1984), Cell 36, 121-129.
14. Cavener, D., Corbett, G., Cox, D. & Whetten, R. (1986), EMBO J. 5, 2939-2948.
15. Cavener, D.R., Otteson, D., & Kaufman, T.C. (1986b). Genetics, 114, 111-123.
16. Scott, M.P., Weiner, A.J., Hazelrigg, T., Polisky, B.A., Pirrota, V., Scalenghe., F., & Kaufman, T.C. (1983). Cell 35, 763-776.
17. Henikoff, S. (1984). Gene 28, 351-359.
18. Sanger, F., Nicklen, S., & Coulson, A.R. (1977). Proc. Natl. Acad. Sci. USA 74, 5643-5467.
19. Bender, W. Akam. M. Karch, F., Beachy, P., Peifer, M. Soierer, P., Lewis, E. B., & Hogness, D.S. (1983). Science 221,23-29.
20. Karch, F., Weiffenbach, B., Peifer, M., Bender, W., Duncan, I., Celniker, S., Crosby, M., & Lewis, E.B. (1985). Cell 43, 81-96.
21. Young, M.W. (1979). Proc. Natl. Acad. Sci. USA 76, 6274-6278.
22. Lewin, B. (1980). Gene Expression 2: Eukaryotic Chromosomes, 2nd ed. John Wiley & Sons, NY, pp. 540-569.
23. Wharton, K. A., Yedvobnick, B., Finnerty, V. G., & Artavanis-Tsakonas, S. (1985). Cell 40, 55-62.
24. Shimizu, A. & Honjo, T. (1984). Cell 36, 801-803.
25. Spradling, A. C., de Cicco, D. V., Wakimoto, B.T., Levine, J.F., Kalfayan, L. J., & Cooley, L. (1987) EMBO J. 6, 1045-1053.
26. Beverly, S. M., & Wilson, A.C. (1984). J. Mol. Evol. 21, 1-13.
27. Throckmorton, L.H. (1975). In Handbook of Gentics (King, R. C., ed.), vol. 3, pp. 421-469, Plenum Press, New York.
28. Blackman, R.K. & Meselson, M. (1986). J. Mol. Biol. 188, 499-515.
29. Tung, C.S. & Harvey, S.C. (1987). Nucl. Acids Res. 15, 4973-4984.
30. Nussinov, R. & Lennon, G.G. (1984). J. Mol. Evol. 20, 106-110.