

Editorial Manager(tm) for Genomics  
Manuscript Draft

Manuscript Number:

Title: DNA tracts composed of only two bases concentrate in gene promoters.

Article Type: Regular Article

Section/Category:

Keywords: promoter; transcription initiation; purine; pyrimidine; DNA unwinding; cerevisiae; elegans; melanogaster; human chromosome 14; human chromosome 22; jannaschii; AT; GC; R.Y; W; S; K.M; single stranded; gene activation; binary DNA

Corresponding Author: Prof. Gad Yagil, PhD

Corresponding Author's Institution: Weizmann Institute of science

First Author: Gad Yagil, PhD

Order of Authors: Gad Yagil, PhD

Manuscript Region of Origin:

Abstract: We have previously shown that long DNA tracts composed of only two of the bases ("binary DNA") are highly overrepresented in sequenced eukaryotic genomes. Here we examine gene promoter regions, by superposing all genes in a chromosome at their transcription start sites (TSS). We find that of the four motifs made of two bases, three are concentrated in gene promoters: Purine/pyrimidine tracts are highly overrepresented in the promoters of yeast chromosome IV, of *C. elegans* Chromosome I, of *A. thaliana* Chr. 2 and in human chromosomes 14, 21 and 22 (a subset). A,T rich tracts (W tracts) are enriched in the same chromosomes, as well as in *D. melanogaster* chromosome 2R and in an archeon, *M. Jannaschii*. A third motif, K.M tracts, shows some concentration in *D. melanogaster* promoters. A propensity of binary DNA to unwind is proposed as explaining the high presence of the two base motifs in gene promoters.

Dear Editor,

I am enclosing a manuscript entitled: "DNA tracts composed of only two base pairs are overrepresented in gene promoters". I hope your referees will agree that the effect demonstrated is quite unexpected, and presents a challenge to those who try to decipher how promoters determine gene activity. The technique employed in this paper (not in our previous papers) is a computational one, a legitimate and efficient tool these days.

Sincerely yours,

G. Yagil

p.s. I requested R. Mural as editor; Ross Hardisson and Russ Altman can be alternatives. May I suggest Craig Benham, of University of California, Davis, as a referee? (cjbendam@ucdavis.edu)

---

Gad Yagil, Ph. D.  
L.G. Lawrence Professor of Cancer Research  
Dept. of Molecular Cell Biology  
The Weizmann Institute of Science  
Rehovot, Israel, 76100;  
Tel. 972-89-460-918 (home); Fax 089-344-125  
e-mail: gad.yagil@weizmann.ac.il  
web site: <http://www.weizmann.ac.il/~lcyagil>

---



We have previously studied two yeast genes which contain long pyrimidine tracts in their promoters, *ded1* and *cyc1*. We found these long tracts to be sensitive to single strand specific nuclease P1 when in negatively superhelical plasmids [16], implying that the pyrimidine containing sequences tend to be unwound, at least intermittently. Similar studies have been carried out with promoters containing W tracts [11,19,20]. Also, An "initiator" promoter element with a number of consecutive pyrimidines has been experimentally identified in a series of promoters [17-18]. Since only a limited number of genes can be studied by direct probing, studies that analyze the distribution of the various DNA binary tracts on the complete genome level are necessary. A systematic survey of promoters on the complete chromosome level are however not yet available.

In this communication the occurrence of DNA tracts of all four binary motifs in gene promoters of five eukaryotic and one archeon species is presented. An unexpected excessive concentration of R.Y, K.M and/or W tracts is found in the promoter regions of yeast, *Arabidopsis*, *C. elegans*, *D. melanogaster* as well as in two human chromosomes. The high concentrations observed in promoter regions strongly support a specific role for the binary DNA tracts in promoter function, and raise the possibility that one or more unwinding events may participate in gene activation processes.

## **Results**

Chromosomes of six species were examined. Their characteristics and data sources are shown in Table 1. All the genes annotated in each chromosome were superposed at their transcription start sites (TSS) or ATG's. For each chromosome, lists of all binary tracts with 13 or more bases ( $\geq 13$ nt) were prepared (see legend to Figs. 1 and 2) and the number of bases in these tracts was plotted against their distance from the TSS. These plots are shown in Figs. 1-3. The height of each column shows the cumulative number of bases present in tracts  $\geq 13$  nt – e.g., the brown color is for tracts of length of 13 to 15 bases, and so on (see box).

**R.Y tracts** The distribution of the R.Y (purine.pyrimidine) tracts is shown in panels A of Figs. 1 and 2. A distinct peaking in the immediate 5' region of yeast, *Arabidopsis*, *elegans* and human chromosomes (a subset, see next paragraph) is clearly evident. The highest concentration is either directly at the TSS (*Arabidopsis* and human), or 80-160 nt 5' upstream of the ATG (yeast and *C. elegans*). As the exact position of the TSS is not yet known for most yeast and *elegans* genes, some of the tracts plotted e.g. 40-80 bases from the "TSS" may still reside in the transcription initiation region. The transcribed regions of these four chromosomes have significantly fewer R.Y tracts, while in the intergenic regions which lie upstream of the promoters, a certain excessive level is maintained, but significantly less than in the promoter region. No distinct peak is evident in *Drosophila* chromosome, see nevertheless the next section. In yeast (Fig 1a-A), the possibility exists that this decrease in binary tracts is due to the coding region of the next gene upstream. Supplementary Fig. 1 describes the subset of yeast genes where the next gene ends at least 800 bases upstream; the data are more scattered in this subset of only 99 genes, but a distinct promoter preference is evident.

In human chromosome 14 the concentration of R.Y tracts in the promoter region is clearly evident only when R.Y tracts containing more than three consecutive A's or T's or more are excluded from the plot (Fig. 2b-A). This subset of R.Y tracts comprises 45% of all R.Y tracts  $\geq 10$ nt. The main contig of human chromosome 22 also shows a distinct peak near the TSS when the same subset is excluded (supplementary Fig. 2). Human chromosomes thus seem to fit the trend manifested by the other eukaryotic chromosomes, namely that homopurine.homopyrimidine tracts are a significant feature of gene promoters. When all R.Y tracts in chromosomes 14 or 22 are mapped, a peak only on the border of significance is observed (supplementary Fig. 3).

Six control runs, with random DNA of the length of yeast chr. 4, are shown in supplementary Figs. 4 and 5. No peaks are evident near the TSS. Also, no tracts longer than 21nt were detected in the control runs, while many longer tracts were found in the analyzed chromosomes, see the color coding of tract lengths in text Figs. 1-3. It should be noted that even in the transcribed regions, where significantly less tracts are evident, R.Y tracts are still in excess of the number expected in randomized DNA, as indicated by the red horizontal lines in text Figs 1-3.

The concentration of R.Y tracts is still enhanced in the upstream intergenic regions, beyond the first 400-600 core promoter bases, but less distinct than in the core promoter regions (compared to random DNA, red lines in Figs. 1 and 2). This enhanced presence further upstream may be due to several effects: 1)

Binary tracts may be associated with enhancers or other upstream promoter elements. 2) due to far upstream exons which have yet to be identified [21]; to examine this point we have examined an available set of 434 and 869 new “first exons” [21] in chromosomes 21 and 22 respectively, but no significant difference from the GenBank listed TSS set was apparent (unpublished). 3) The upstream regions may contain promoters for the many micro-RNA species recently identified [22]. 4) Binary tracts may also have a role in other DNA directed processes, such as recombination and repair, as discussed later. In summary, the clear excess of R.Y tracts in promoter regions alludes to a functional significance of these tracts.

**K.M tracts.** Panels B of Figs. 1a and 1b show clear peaks in yeast and *Arabidopsis*. In *C.elegans* only a minor excess is seen, while in *Drosophila* a shallow yet clear peak is found (Fig. 2a-B). In the human chromosomes (Fig 2b-B) a minimum around the TSS is observed. It should be noted that in *Drosophila*, the values on the vertical scale are higher for K.M tracts than for R.Y tracts. This is in accord with a previous finding that in the *Drosophila* chromosome K.M tracts are the most overrepresented binary motif [7].

**W tracts.** Panels C of Figs. 1 and 2 show highly significant peaks in the promoters for these “A,T rich” tracts, in all five organisms studied; the peak is least evident in *C. elegans* (Fig.1c-C). In human chromosome 14 the peak occurs further upstream than those for R.Y and K.M tracts and becomes evident only when bins up to -4000 nt are plotted (supplementary Fig 6). A clear peak is seen also in Chromosome 22 (supplementary fig. 2). The peak for the W tracts in *Arabidopsis* also occur further upstream and is also more broadly distributed, which may indicate separate functions for R.Y and W tracts. W tracts are well known to melt early, which is a necessary step in a DNA unwinding process. A clear peak for W tracts is also found in the promoters of an archeon, *M. jannaschii* (Fig. 3) as has been observed also in Bacteria [12] (neither R.Y nor K.M tracts are particularly overrepresented, as in bacteria). R.Y overrepresentation is so far limited to eukaryotic organisms.

**S tracts** (panels D): The peaks seen here for humans and *Arabidopsis* are mostly the well studied CpG islands of these two organisms [2]. *M. jannaschii*, Yeast and *C. elegans*, which lack CpG islands, show very few S tracts, while *Drosophila* may have some CpG Islands. The procedures used here can thus serve as an alternate way to detect CpG islands.

**Tract abundance:** More than 80% of the promoters have at least one binary DNA tract  $\geq 13$  nt in their promoter region (Table 2, column 3). Thus, in human chromosome 22, 458/476 of the genes have at least one binary tract  $\geq 13$  nt within the first 600 bases upstream (over 79% have an R.Y tract, 49% a K.M tract and 56% a W tract; many promoters carry all three motifs). The figures are of similar magnitude in the other chromosomes surveyed (Table 2), except for yeast, where only the first 300 nt upstream were counted. Only 7.3% of the promoters are expected to have a tract  $\geq 13$  nt in 600 bases of randomized DNA (equation 3 of ref. [7]). This again suggests that each of the three motifs has a role in promoter function. The preponderance in *Drosophila* promoters of K.M tracts over R.Y tracts (see abscissa) is again to be noted. Promoters with W tracts are least abundant in human chromosome 14, and most abundant in *Arabidopsis*, *elegans* and *Drosophila* (Table 2, column 6). Human chromosome 14 contains two special gene groups: Olfactory receptors genes (45 genes),

close to the centromere and immunoglobulin heavy chain genes, at the telomere end (111 V<sub>h</sub> genes, 27 D and 8 J segments, plus one constant chain). Superposing the 111 V<sub>h</sub> genes alone shows a clear R.Y peak around the ATG of these genes. At least one R.Y tract is present within 200 bases of most V<sub>h</sub> genes, occasionally within the first intron. Their functional significance is now being studied.

## **Discussion**

Each of the five organisms examined has two or more binary motifs highly overrepresented in their promoter, except for the *Drosophila*, where only W tracts are clearly overrepresented. The early melting property of W tracts strongly suggests that a DNA unwinding event may be involved. An unwinding role for W tracts has been previously suggested by several authors [19-20,23-24], based mainly on susceptibility of the promoter regions to single strand specific DNAses or to permanganate probing (reviewed in [25]). Promoter regions susceptible to single strand specific nucleases were termed DNA Unwinding Elements, or DUE's, by Kowalski and co-workers [14,19].

It is somewhat harder to assign a DNA unwinding role to R.Y tracts. There is nevertheless a considerable amount of evidence, starting with the seminal experiments of Larsen and Weintraub [13] that purine.pyrimidine rich tracts tend to be in the unwound state. Studies covering a wide range of gene promoters [15, 25-26] indicate that under suitable conditions gene promoters containing R.Y (or K.M tracts, [27]) may assume a strand separated state, at least intermittently. We have studied two selected yeast genes which contain long Y tracts in their promoters (*cyc1* and *ded1*) [16]. Under negative superhelical stress, the Y rich sequences became susceptible to single strand specific nuclease P1; topological analysis has indicated that six primary turns were unwound. This implies that under normal superhelical stress, these promoters assume an intermittently unwound state. Whether the unwinding event occurs prior to, or after the assembly of the transcription machinery, could not be assessed at that stage.

Consulting the intermediate output lists of TRACTS, it is observed that individual R tracts and Y tracts are equally overrepresented in the promoters analyzed – in some promoters the R tracts, in others the Y tract are on the coding strand. This is the rationale behind the grouping of the R and Y tracts into one class. The function of the R.Y or W tracts is thus orientation independent; it does not matter whether the purines or the pyrimidines are on the coding strand, in line with a DNA unwinding role. The fact that binary tracts do not occur at a very definite distance from a TSS is also more in favor of an unwinding role than in favor of binding particular factors, like TBP, the TATA binding protein, to the TATA box. To serve as a DNA unwinding center, dynamic stabilization of the separated state by suitable proteins is likely to be required. A number of proteins which preferentially bind to single stranded oligopurines [28] or to single stranded A,T rich (W) tracts [23,29] have been identified. These single strand binding proteins are essential partners in many DNA directed processes, including replication [30-32], transcription [15,26], recombination [27], repair [33], and even translation [34]. A requirement for binary motifs in processes other than transcription can thus explain the overrepresentations observed in intergenic regions, beyond the specific promoter sequence.

In summary, the very high presence of long W, R.Y and K.M tracts in the promoters of a range of eukaryotic species indicates a special function of these tracts. A DNA unwinding event can explain most observations reported here, including early melting, independence of orientation, non-fixed position, and the presence of at least one long binary tract in almost every promoter. The omnipresence of binary DNA tracts in gene promoter and adjacent regions seems thus to be a topic that warrants further investigation.

## **Methods**

The Data shown in Figs 1 and 2 were created by the following pipeline of computer programs: First, program ANEX (in FORTRAN, on a unix platform) parses a flat GenBank file of a of a chromosome or a contig, to produce a file with one line for each gene, listing the designation, orientation, start and end sites of that gene, as well as a brief functional description. The first mRNA listed, generally the longest one, is parsed. Next, based on the one line per gene file, program TRAPAR (in perl) creates a complete set of the promoter sequences, of length and range specified during input (e.g. from -2000 to +500, i.e. 2500 bases). A second input to TRAPAR is the GenBank sequence. The GenBank files utilized are shown in Table 1. Then, program TRACTS [35], in FORTRAN, is applied to the promoter set produced by TRAPAR, to produce lists of all binary tracts  $\geq 10$ nt present in the promoter set, one list for each binary motif (R.Y; K.M; W and S; also R and Y separately). Finally program BINS (in perl) determines, for each contig or chromosome, the distance of every tract from the GenBank listed start site (ATG's or true TSS, as available) and allocates each tract into a bin of specified distance from the TSS (40 to 100nt). BINS also produces a matrix giving the number of bases in each tract-length interval as function of its distance to the TSS. These matrices are transferred to EXCEL (Microsoft Inc.) to create the plots shown in the Figures. Programs are available upon request.

**Acknowledgements:** The author thanks Dr. Shifra Ben-Dor for critical remarks to the manuscript, to many members of the Weizmann Institute Computing Center and of the Biological Computing Unit for continuous and friendly support and advice.

## **References:**

1. S. Karlin, G. Ghandour, The use of multiple alphabets in kappa-gene immunoglobulin DNA sequence comparisons. *EMBO J.* 4 (1985) 1217-1223.
2. A. Bird, DNA methylation patterns and epigenetic memory. *Genes and Develop.* 16 (2002) 6-21.
3. C. Tamm, H.S. Shapiro, R. Lipshitz, E. Chargaff, Distribution density of nucleotides within a deoxyribonucleic acid chain. *J. Biol. Chem.* 203 (1952) 673-698.
4. E. Chargaff, *Essays in Nucleic Acids*. Elsevier, Amsterdam, 1963, pp.126ff, 146ff.
5. W. Szybalski, H. Kubinski, P. Sheldrick, Pyrimidine clusters on the transcribing strands of DNA and their possible role in the initiation of RNA synthesis. *Cold Spring Harbor Symp. Quant. Biol.* 31 (1966) 123-127.
6. P. Bucher, G. Yagil, The occurrence of oligopurine.oligopyrimidine tracts in eukaryotic and prokaryotic genes. *DNA Sequence*, 1 (1991) 27-43.
7. G. Yagil, The over-representation of binary DNA tracts in seven sequenced chromosomes. *BMC Genomics* 5 (2004) 19-37.
- 8 J. Moreau, L. Matash-Smirniaguina, and K. Scherrer, Systematic punctuation of Eukaryotic DNA by A+T sequences. *Proc. Natl. Acad. Sci. USA* 78 (1981) 1341-1345.
9. R. Nussinov, A. Sara , G.W. Smythers, R.L. Jernigan, Sequence context of oligomer tracts in eukaryotic DNA. Biological and conformational implications. *J. Bioch. Struct. Dynamics* 6 (1988) 543-562.
10. S. Lisser, H. Margalit, Compilation of mRNA promoter sequences. *Nucl. Acids Res.* 21 (1993) 1507-1516.
11. W. Ross, E. Alexander, L. Gourse, Fine structure of E. coli RNA polymerase-promoter interactions: A subunit binding to the UP element minor groove. *Genes and Dev.* 15 (2000) 491-506.
12. Shomer, G. Yagil, Long W tracts are over-represented in the *E. coli* and *H. Influenzae* genomes. *Nucl. Acids Res.* 27 (1999) 4491-4480.
13. A. Larsen, H. Weintraub, An altered DNA conformation detected by S1 nuclease occurs at specific regions in active chick globin chromatin. *Cell*, 29 (1982) 609-616.

14. D. Kowalski, M.J. Eddy, The DNA unwinding element A novel, cis acting component that facilitates the opening of the *E. Coli* replication origin. *EMBO J.* 8 (1989) 4335-4339.
- 15 C. Giardina, J. Lis, Dynamic protein-DNA architecture of a yeast heat shock protein. *Mol. Cell. Biol.* 15 (1995) 2737-2744.
16. G. Yagil, F. Shimron, M. Tal, DNA unwinding in the CYC1 and DED1 yeast promoters. *Gene* 225 (1998) 153–162.
17. S.T. Smale, J.T. Kadonaga, The RNA polymerase core promoter. *Ann. Rev. Biochem.* 72 (2003) 449-479.
18. R.P. Perry, The architecture of mammalian ribosomal protein promoters. *BMC Evolut. Biol.* 5 (2005) 15-36.
19. R.M. Umek, D. Kowalski, The DNA unwinding element in a yeast replication origin functions independently of easily unwound sequences present elsewhere on a plasmid. *Nucl. Acids Res.* 18 (1990) 6601-6605.
20. H. Wang, M. Noordewier, C.J. Benham, Stress induced DNA duplex destabilization (SIDD) in the *E. coli* genome: SIDD sites are closely associated with promoters *Genome Res.* 14 (2004) 1575-1584.
21. R.V. Davuluri, I.V. Grosse, M.Q. Zhang, Computational identification of promoters and first exons in the human genome. *Nature Genetics* 29 (2001) 412-417.
22. P. Kapranov, S.E. Cawley, J. Drenkow, S. Bekiranov, R.L. Strausberg, S.P.A. Fodor, T.R. Gingeras, Large scale transcriptional activity in chromosomes 21 and 22. *Science* 296 (2002) 916-919.
23. D. Bramhill, A. Komberg, A model for initiation at origins of DNA initiation. *Cell* 54 (1988) 915-917.
24. M. Tal, F. Shimron, G. Yagil, Unwound regions in yeast centromere DNA. *J. Mol. Biol.* 243 (1994) 179-189.
25. G. Yagil, Paranemic structures of DNA and their role in DNA unwinding. *Crit. Revs. Biochem. Mol. Biol.* 26 (1991) 475-559.
26. G.A. Michelotti, E.F. Michelotti, A. Pullner, R.C. Duncan, Dirk E, Levens A, Multiple single-stranded cis elements are associated with activated chromatin of the human c-myc gene in vivo. *Mol. Cell Biol.* 16 (1996) 2656-2669.

27. R.B. Tracy, J.K. Baumohl, S.C. Kowalczykowski, The preference for GT-rich DNA by the yeast Rad51 proteins defines a set of universal pairing sequences. *Genes and Develop.* 11 (1997) 3423-3431.
28. A.D. Bergemann, E.M. Johnson, The HeLa pur factor binds single-stranded DNA at a specific element conserved in gene flanking regions and origins of replication. *Mol. Cell. Biol.* 12 (1992) 1257-1265.
29. A.M. Makhov, P. Boehmer, I.R. Lehman, J.D. Griffith, The herpes simplex virus type 1 origin-binding protein carries out origin specific DNA and forms unwound stem-loop structures. *EMBO J.* 15 (1996) 1742-1750.
30. M.S. Wold, J.J. Li, D.H. Weinberg, D.N. Virshup, D.L. Sherley, E. Verheyen, T. Kelly, Cellular proteins are required for SV40 DNA replication in vitro. *Cancer cells* 1988 133-141.
31. M.K. Kenny, S-K. Lee, J. Hurvitz, Multiple functions of a human single-stranded protein in SV40 DNA replication: Single strand stabilization and stimulation of DNA polymerases alpha and delta. *Proc. Natl. Acad. Sci. USA* 86 (1989) 5757-9761.
32. X. He, I.R. Lehman, Unwinding of a herpes simplex virus type 1 origin of replication oris by a complex of the viral origin binding protein and a single stranded DNA binding protein. *J. Virol.* 74 (2000) 5726-5728.
33. X. Wang, J.E. Haber, Role of *Saccharomyces* single-stranded DNA-Binding protein RPA in the strand invasion step of double-strand break repair. *PLoS Biology* 2 (2004) 104-112.
34. O. Meyuhas, Synthesis of the translational apparatus is regulated at the translational level. *Europ. J. Biochem.* 267 (2000) 6321-6330.
35. M. Gal, T. Katz, A. Ovadia, G. Yagil, TRACTS: A program to map oligopurine.oligopyrimidine and other binary DNA tracts. *Nucl. Acids Res.* 31 (2003) 3679-3681.

## Tables.

**Table 1. The chromosomes analyzed.**

	Chromosome-	Build ,Date	GenBank accession number:	Length	No. of genes <sup>b</sup>
<i>S. cerevisiae</i>	4	16/6/02	NC_001136.2	1,531,929	856
<i>A. thaliana</i>	2	21/12/99	AE002093	19,647,091	4116
<i>C. elegans</i>	1	23/4/99	"chr_I"	16,183,83 <sup>a</sup>	2516
<i>D. melanogaster</i>	2R (right arm)	7/11/02	NT_033778.1	20,302,755	2805
<i>H. sapiens</i>	14, contig "87"	10/4/03, Build 34.3	NT_026437.10	87,191,216	1005
<i>H. sapiens</i>	22, contig "23"	19/4/04, Build 34.3	NT_011520.9	23,178,213	476
<i>M. jannaschii</i>	Main	10/9/01	NC_000909	1,664,970	1712

- a- 1,441,828 "N" bases excluded.  
b- As detected by ANEX in the GenBank annotation file.  
References for the sequences utilized are quoted in ref. 7

**Table 2. Percentage of genes which have a DNA binary tract  $\geq 13$  nt in their promoter (-600 - 1 upstream: -300 - 1 for yeast)**

	Chromosome	All three Motifs <sup>a</sup>	R.Y	K.M	W
<i>S. cerevisiae</i>	4	79	35	25	56
<i>A. thaliana</i>	2	98	80	74	80
<i>C. elegans</i>	1	94	68	56	91
<i>D. melanogaster</i>	2R (Right arm)	92	34	50	77
<i>H. Sapiens</i>	22, contig 23	96	79	49	56
<i>H. sapiens</i>	14, contig "87"	92	76 <sup>b</sup>	51	39

Tracts were counted as follows: Program BINS produces a file which lists for each promoter (bases -600 and 0 (TSS)) every binary tract longer than 10 nt. These files were sorted so that all tracts less than 13 nt were deleted, and a single line was left for each promoter. The number of lines was counted, and divided by the number of genes in the contig (Table 1). The expected number of tracts  $\geq 13$ , in randomized 600 nt, was calculated by eq. (3) of Yagil (2004), with 50% A, and the result is 0.073, i.e. 10 times lower than for most promoter regions in Table2. The number of promoters containing at least one tract of one of the 3 motifs was calculated by concatenating the 3 lists of tracts  $\geq 13$ nt, deleting by script GENECOUNT all promoters appearing more than once, and counting the number of lines (promoters) left.

a - Unary tracts like polyA may be common to all 3 motifs.

b - Full set of R.Y tracts.

## Legend to Figures 1 and 2

Promoter regions of specified length of an entire chromosome (or contig) were superposed so that their GenBank annotated Transcription Start Sites (TSS; ATG in yeast and *elegans*) coincide. The number of bases in binary DNA tracts longer than 12 bases (9 for S tracts is plotted against the distance of each tract from the TSS (in bins of 40 -100 nt, the 5' end of a tract determines its bin). Tract lengths are color coded according to length, in intervals of 3nt, as specified in the legend box. The horizontal red lines show the number of bases in tracts longer than 12 nt expected in randomized DNA of the same base composition; the number is indicated in red on the ordinate. The expected base numbers are calculated as in ref. 7

The panels in horizontal direction show:

Fig. 1a: *S. cerevisiae*, chr. 4.

Fig. 1b: *A. thaliana*, chr. 2.

Fig. 1c: *C. elegans*, chr. 1.

Fig. 2a: *D.melanogaster*, chr. 2, right arm.

Fig. 2b: *H. sapiens*, contig 87 of chr. 14 (practically the entire chromosome).

The GenBank data files used are listed in Table 1.

The panels in vertical direction show:

A: R.Y tracts,

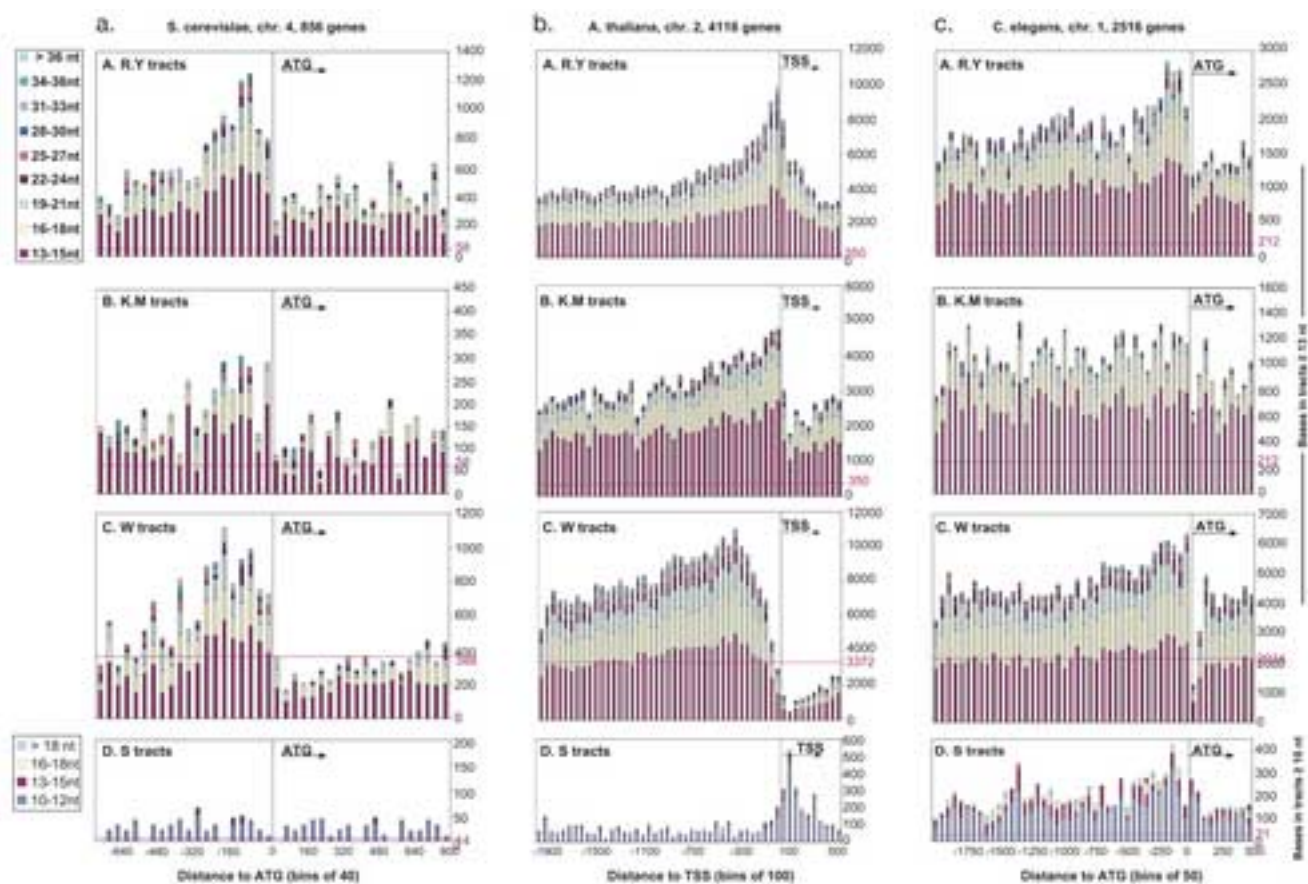
B: K.M tracts,

C: W tracts

D: S tracts.

## Fig.3

Distribution of binary DNA tracts around the Transcription Start Sites of an archeon, *M. Jannaschii*. For details see legend to Figs. 1 and 2. S tracts were not plotted, as there are only very few S tracts in this chromosome.



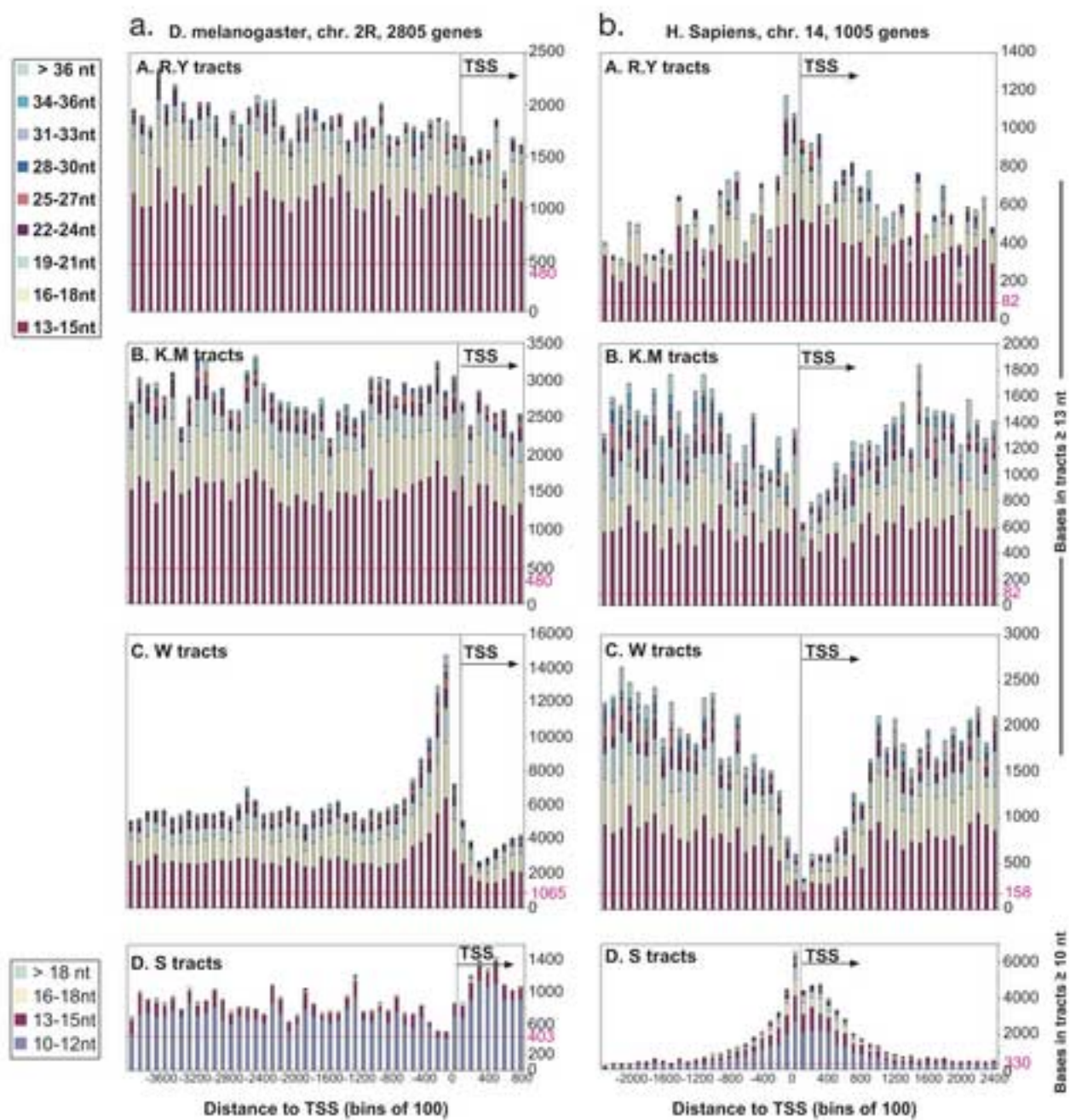
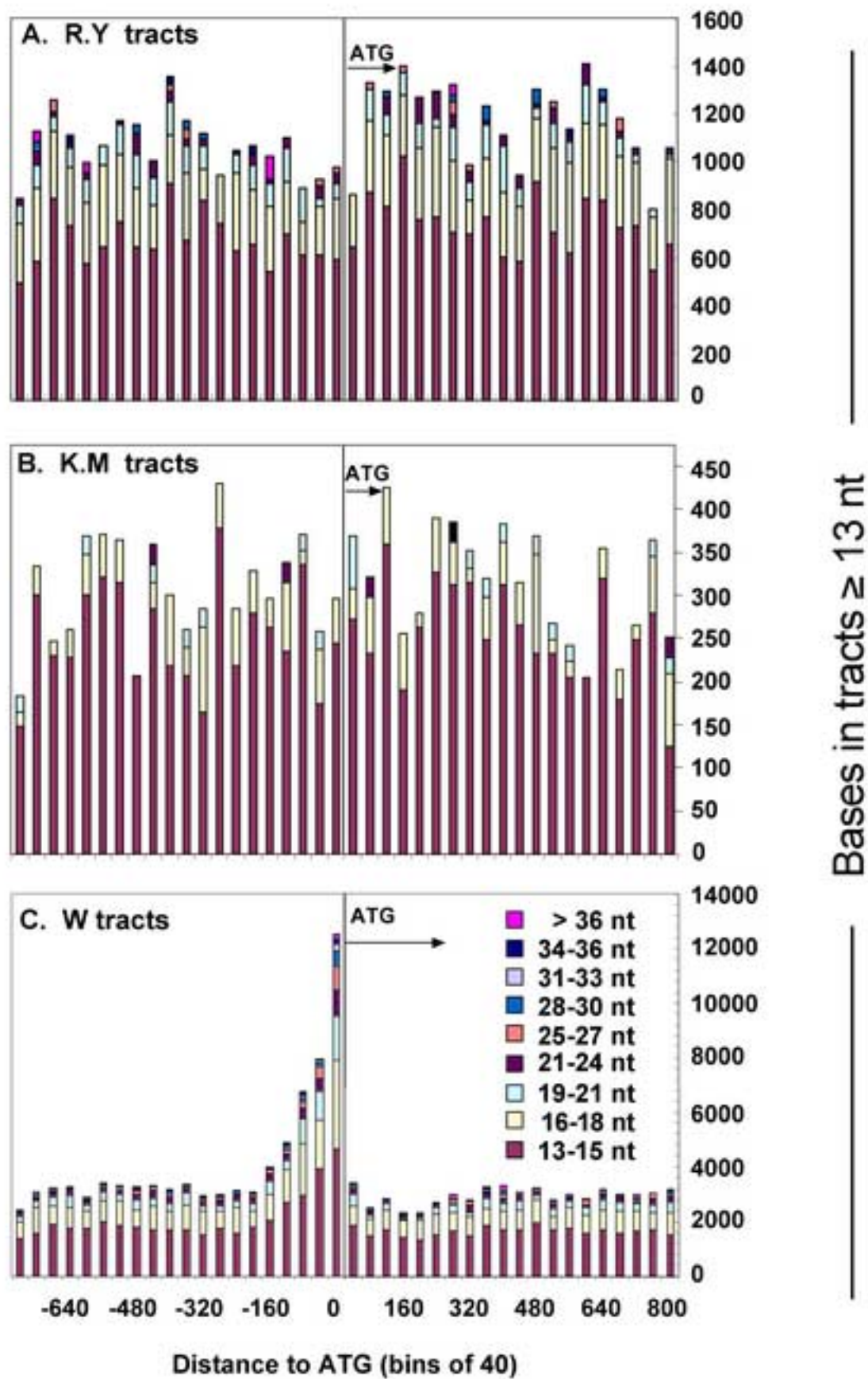
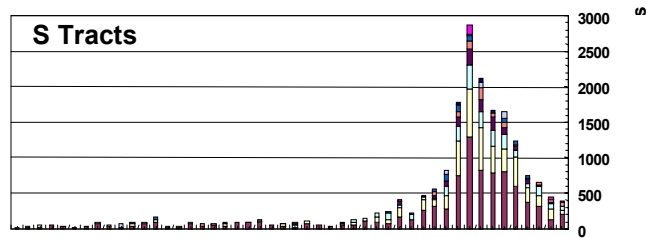
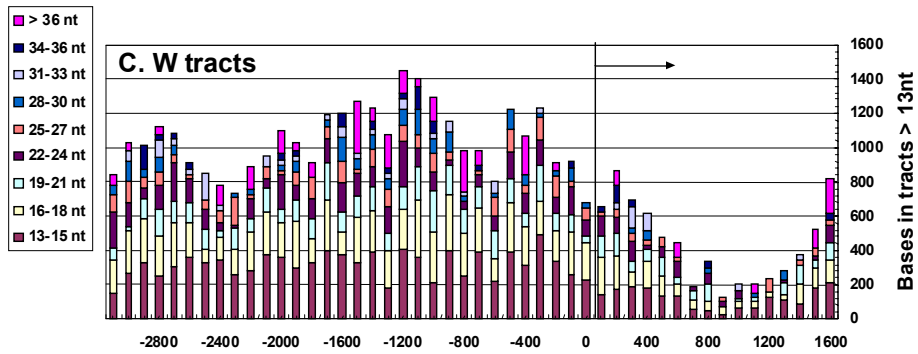
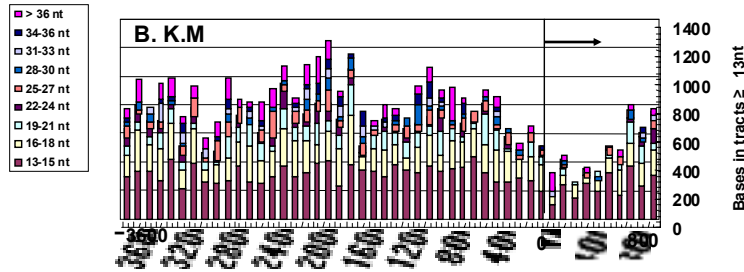
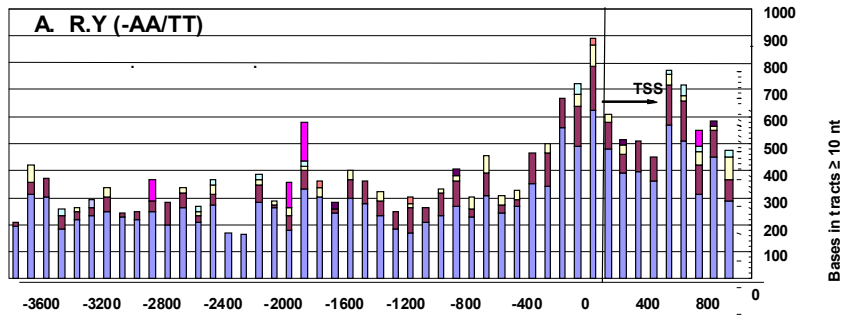


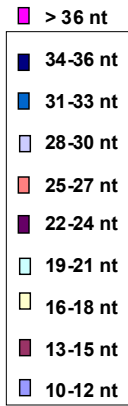
Fig 3. *M. Jannaschii*, 1712 genes



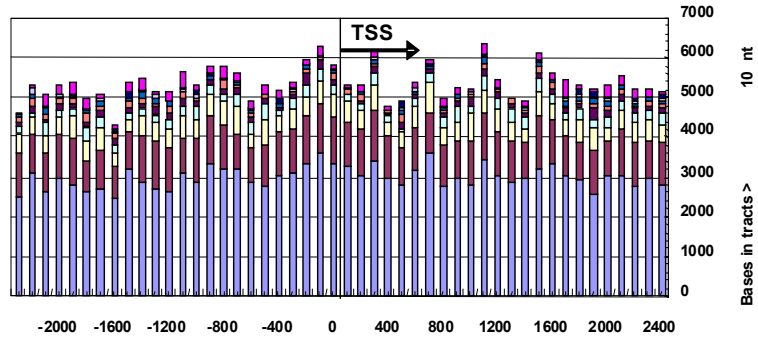
Suppl. Fig. 2 contigs 23 of chromosome 22



Distance from TSS (nt)

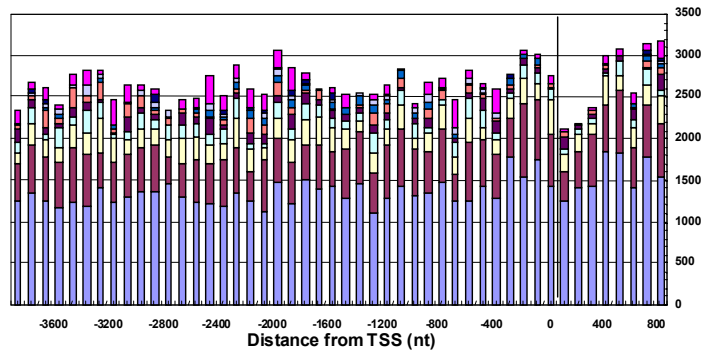


Supl. Fig. 3a. *H. Sap* chr.14, contig 87: R.Y tracts  $\geq 10$ nt, All A|T ctg. tracts



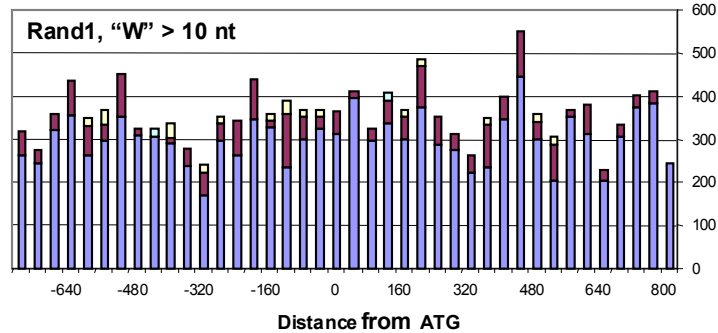
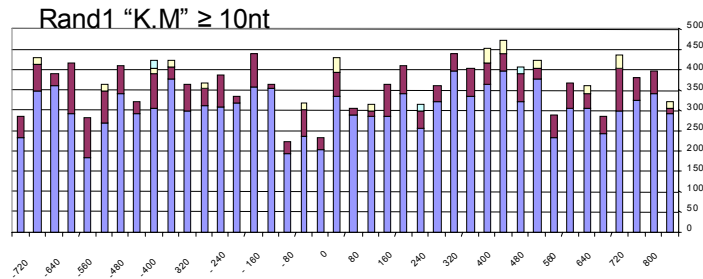
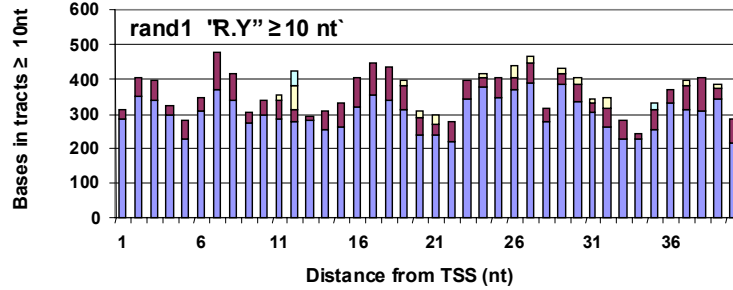
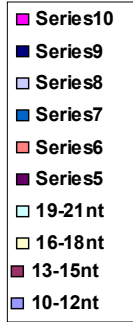
Sup. Fig. 3b. *H. Sap* chr.22, contig 23: R.Y tracts  $\geq 10$ nt, All A|T ctg. tracts

476genes



**Supl. Fig 4. Three randomized yeast promoter sequences  
(845 “gene promoters”)**

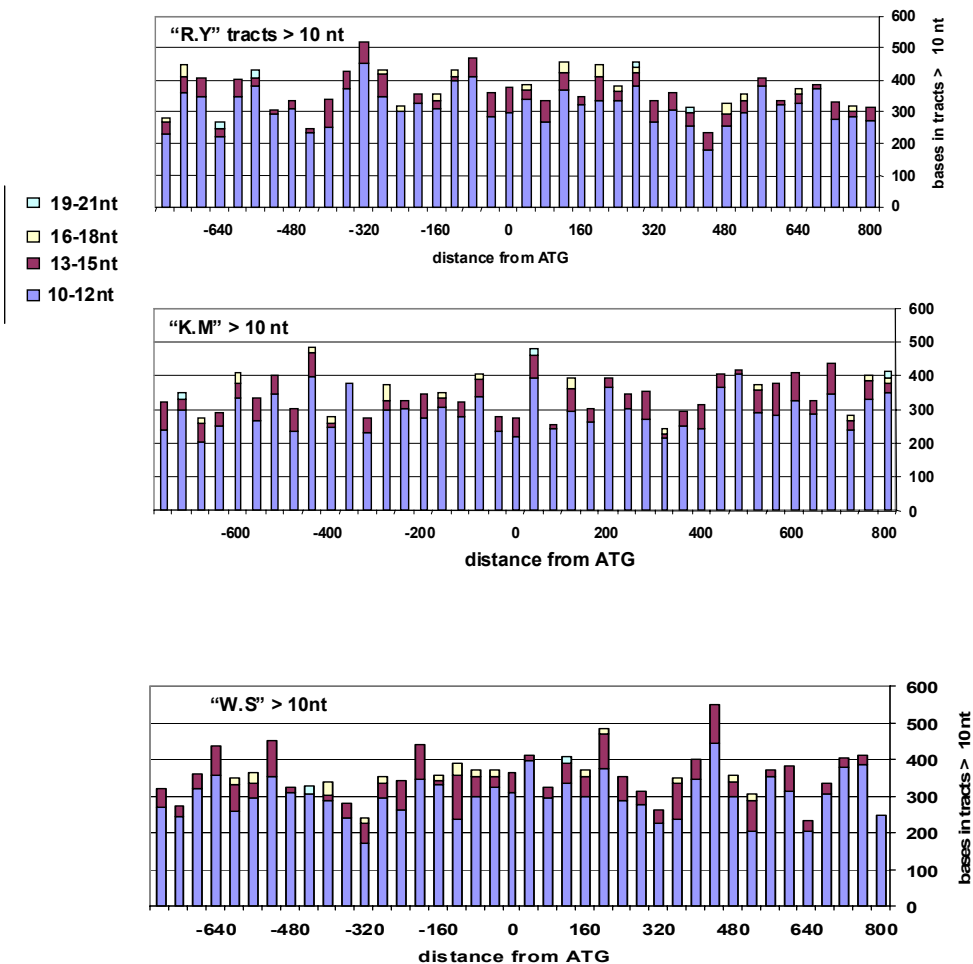
Please note that: 1) The vertical scale is 5-10 times less that in the real plots.  
2) No tracts longer than 21 nt are found in the randomized set and only few > 20 nt  
3) No. of bases can change only in steps > 10 nt.



Rand 1: 1,352,000 bases were randomly generated, with 50% A,G (“ry”), 50% A,C (“km”) Or 62.1% A,T (“ws”), as in yeast chromosome 4s, and analyzed by TRACTS and BINS.pl, as explained in text Fig. 1. 1,352,000 means 1600 promoter bases of 845 “genes” .

**Supl. fig. 5. Rand2: three additional randomized yeast chr 4 “promoters”**

Please note that: 1) The vertical scale is 5-10 times less that in the real plots.  
 2) No tracts longer than 21 nt are found in the randomized set and only few > 20  
 3) No. of bases can change only in steps > 10 nt.



Rand 2: 1,352,000 bases were randomly generated, with 50% A,G (“ry”), 50% A,C (“km”) Or 62.1% A,T (“w.s”, as in yeast chromosome 4), and analyzed by TRACTS and BINS.pl, as explained in text Figs. 1 and 2. 1,352,000 means 1600 promoter bases of 845 “genes” .

Supl. Fig. 6: *H. sap* Chr 14, contig 87, from base -4000 to +800

