Volume 10 Number 15 1982 Nucleic Acids Research

Sequence variation and methylation of the flax 5S RNA genes

P.B.Goldsbrough*, T.H.N.Ellis and G.P.Lomonossoff

John Innes Institute, Colney Lane, Norwich NR4 7UH, UK

Received 27 May 1982; Accepted 29 June 1982

SUMMARY

The complete sequence of the flax 5S DNA repeat is presented. Length heterogeneity is the consequence of the presence or absence of a single direct repeat and the majority of single base changes are transition mutations. No sequence variation has been found in the coding sequence. The extent of methylation of cytosines has been measured at one location in the gene and one in the spacer. The relationship between the observed sequence heterogeneity and the level of methylation is discussed in the context of the operation of a correction mechanism.

INTRODUCTION

In flax the DNA which encodes the 5S RNA varies between 1.5% and 3% of the genome in different lines [1]. This tandemly repetitious DNA comprises two major length classes (Ca 340 or Ca 360 bp) each of which is heterogenous in length, with the shorter length class predominating (Ca 80%). These 5S RNA gene repeats (5S DNA) contain frequent 5 methyl-cytosine (5 me C) residues [1, 2].

In this paper we present the base sequence of repeat units from each major length class. Comparisons between these DNA sequences and also comparisons with other plant 5S RNA sequences [3, 4] show the major mutational pathways in this gene cluster. We conclude that most point mutations are likely to be a consequence of 5 me C residues in this DNA. This finding complements both the observation of mutational hotspots in the *E. coli lacI* gene at 5 me C [5] and the correlation between CpG deficiency and the extent of methylation in animal genomes [6].

MATERIALS AND METHODS

Genomic DNA extraction procedures, electrophoresis transfer and hybridisation methods have been previously described in detail [1, 2] and are not reiterated.

0305-1048/82/1015-4501$2.00/0

Enzymes

BamHI BstNI RsaI and TaqI were purchased from Bethesda Research Laboratories. DNA polymerase I was purchased from Boehringer Corporation Ltd. HpaII and MspI were purchased from New England Biolabs. Digestion conditions were in accordance with the suppliers recommendations.

Plasmids

Both pBG6 and pBG13 are pAT153 [7] derivatives which contain BamHI fragments of flax 5S DNA. pBG6 contains one repeat unit (0.34 Kb) and pBG13 contains ten tandem repeats of the 5S DNA, one repeat being slightly ($\sim$20 bp) longer than each of the other nine. All of these repeats contain a cleavable BamHI site, and all the repeats have the same orientation, as judged by digests with HpaII, TaqI and BstNI. Furthermore all ten repeats in pBG13 carry TaqI sites, but the single BamHI repeat in pBG6 does not. In genomic flax DNA digests the 5S DNA uncut by BamHI is cut by TaqI and vice versa [1]. Thus pBG13 is typical of those molecules uncut by BamHI in genomic DNA while the single insert in pBG6 is typical of those molecules of genomic flax DNA which are completely susceptible to BamHI. No evidence is available to suggest that pBG13 has been derived from multiple ligation events, though this does remain a remote possibility.

DNA sequencing

The recombinant plasmids pBG6 and pBG13 were digested with BamHI, the fragments ligated into the BamHI site of M13 mp2 Bam [8] and the DNA used to transform the E. coli strain JM101. Recombinant plaques were identified and DNA isolated from these phage as described elsewhere [9]. Clones were screened for the length and orientation of inserts by running dideoxythymidine tracks [10]. Sequencing of appropriate clones was carried out as described elsewhere [9] using a 17-nucleotide long synthetic oligodeoxyribonucleotide as primer [11]. DNA sequences from the opposite ends of M13 clones containing short repeats from pBG13 were determined as described [12].

Microdensitometry

This was performed using a Joyce Lobel microdensitometer as previously described [7]. For each band estimations of peak areas were obtained from four positions across each band. Linearity of the relationship between peak area and DNA amount was checked from a dilution series of a standard sample. Peak ares were in the range 400-4000 mm^2.

RESULTS

Digestion of pBG6 (see Plasmids) with BamHI and ligation of the fragments into M13 mp2 Bam resulted in a large number of recombinant clones, twelve

of which were examined by 'T-tracking' [11]. All were identical in size six being in one orientation and six in the other. The same procedure with pBG13 (see Plasmids) also resulted in large number of recombinants, 36 of which were examined by 'T-tracking'. Three of the clones contained the longer repeat unit, two in one orientation and one in the other. The remaining thirty three recombinants contained inserts identical in size to that contained in pBG6. No clones containing pAT153 cut at the BamHI site were obtained with either plasmid. One clone containing the long repeat of pBG13, and one clone of the pBG6 repeat in each orientation were sequenced giving the complete sequence of these repeats. Eight recombinant phages containing short repeats derived from pBG13 were sequenced from both ends [12] because of difficulties identifying recombinants of the same repeat cloned in opposite orientations. From these eight clones three repeat types were found. One of these repeat types represented by three clones contains no BstNI site, and restriction enzyme digests of pBG13 suggest that only one short repeat lacks a BstNI site. One sequence type represented by four clones is identical to the long repeat of pBG13 lacking one copy of the 21 base pair repeat.

The sequences are presented in Figure 1. Since the BamHI site lies in the middle of the putative 5S RNA gene the sequences have been written, for convenience, as if the repeats had been cut at one of the two RsaI sites.

In Figure 1 the sequence of the long repeat is shown in full (a) and differences between this and the short repeat sequences (b-e) are indicated The long repeat is 362 BP in length and the short repeat is 341 BP in length, this difference is due to a 21 BP direct repeat in the long repeat which is present in only one copy in the short repeat (duplication underlined). The position of the presumed coding sequence (marked by a wavy line above the sequence) was inferred by comparing the sequence of the repeat with 5S RNA sequences of several species [3, 4]. The length of the coding sequence is 120 bp if it is assumed that the pyrimidine-purine initiation rule [13] is obeyed and that transcription proceeds no further than in wheat or dwarf bean which have the longest known pyrimidine tracts at the 3' end of any plant 5S RNA.

The base composition of the long and short repeats are similar (49.4% GC for long, 48.7% GC for short) and both are GC rich when compared to flax genomic DNA (41% GC) [14]. It has not been possible to determine directly whether the long and short repeats are interspersed, but the existence of pBG13 suggests that the sequences may be interspersed.

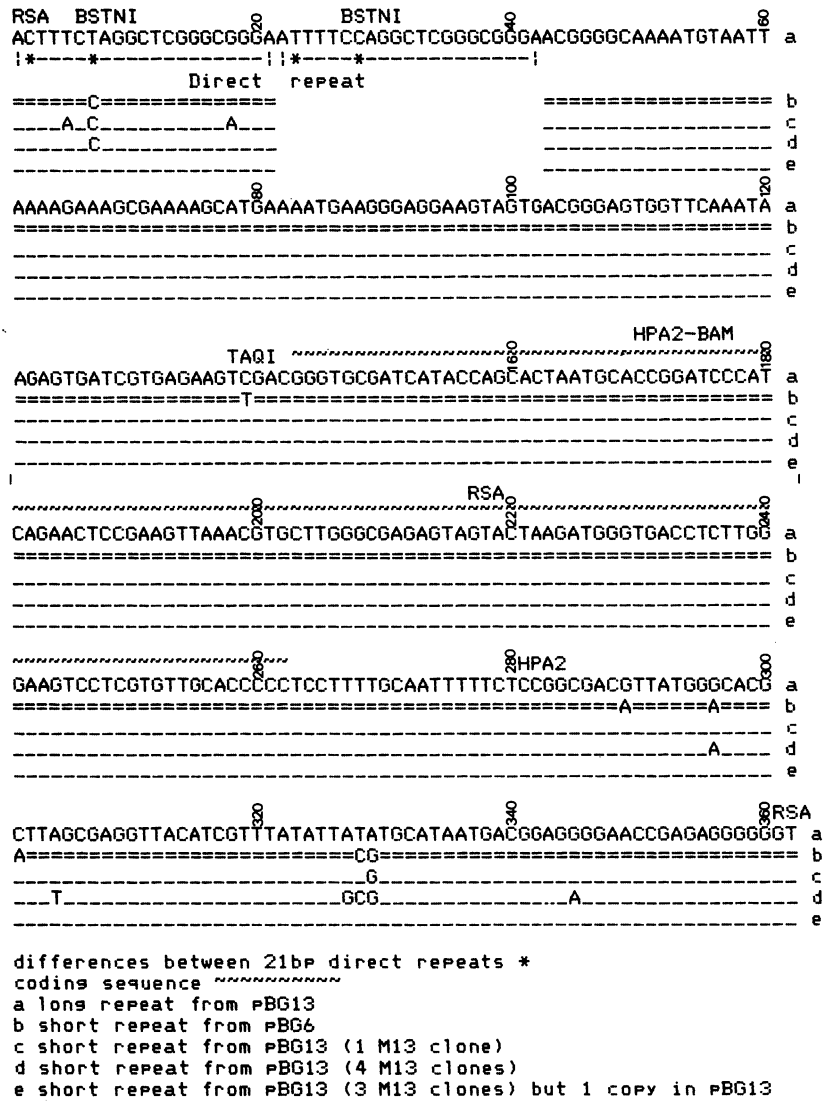


Fig. 1. The complete sequence of one long repeat of the flax 5S gene cluster is shown (a), directly beneath this the differences between this and the short repeats are indicated (b-e). The two copies of the 21 BP duplication in the long repeat are underlined (one copy is absent in the short repeat). Note that there are two single base differences between the copies of the duplicated 21 BP sequences (marked *). As a consequence of these differences only one BstNI site is present in both the long and short repeats. Other differences between the long and short repeats are shown by indicating the base present at that position in the short repeat. Restriction enzyme sites discussed in the text are indicated. A wavy line above the sequence marks the presumed limits of the coding sequence.

3' end

The 5S DNA shows a pronounced strand assymetry, with 63% purines in the non-coding strand of the short repeat. However this strand assymetry is reversed in the region following the 3' end of the coding sequence where only 3 or 20 consecutive bases are purines in the non-coding strand. This situation is reminiscent of the sequence of 14 T residues in the non-coding strand following the 3' end of the wheat 5S RNA gene [15], and also the composition of the region immediately following the 3' end of the Drosophila and Xenopus 5S RNA genes [16, 13].

5' end

If the region flanking the 5' end of the flax 5S RNA gene is compared with the equivalent region of the wheat 5S DNA repeat then the pentanucleotide sequence ATAAG is found in the non-coding strand of both DNAs. The sequence is separated from the 5' end of the wheat 5S RNA gene by 24 base pairs, but the distance is only 20 bases in flax. A related sequence (AGAAG) is found in a similar position in the Xenopus leavis and X. borealis oocyte 5S RNA gene repeat; an identical sequence is found near the 5' end of the Drosophila 5S RNA gene [13].

If the wheat and flax sequences are aligned at this pentanucleotide sequence [Fig. 2] then regions of homology with about 10 BP spacing are evident. Five base pairs from the 5' end of the flax 5S RNA gene the hexanucleotide sequence GAGAAG is found which aligns with the purine tract GGGG in wheat, and is in a similar position to the sequence AAAAG in Xenopus and ATAAG in Drosophila [13].

Thirty three bases from the 5' end of the flax 5S RNA gene the pentanucleotide GGGAG can be found in the non-coding strand. An identical sequence is aligned in the wheat sequence [Fig. 2]. This pentanucleotide sequence is in the same position as the GAC trinucleotide in Yeast and Xenopus [13].

Both the spacing, which presents these sequences to the same side of the DNA molecule, and conservation of these sequences suggests that they may be protein recognition sequences involved with the initiation of 5S RNA transcription.

In animals directly repeating sequences have been found near the 5' end of the 5S RNA gene [13],but no such sequences have been found in wheat [15]. In flax two candidate direct repeats can be found [Fig. 2b] but neither are perfect with 2 or 3 out of 9 bases mismatched. The direct repeats can be seen to have a weak relationship to each other [2c]. These

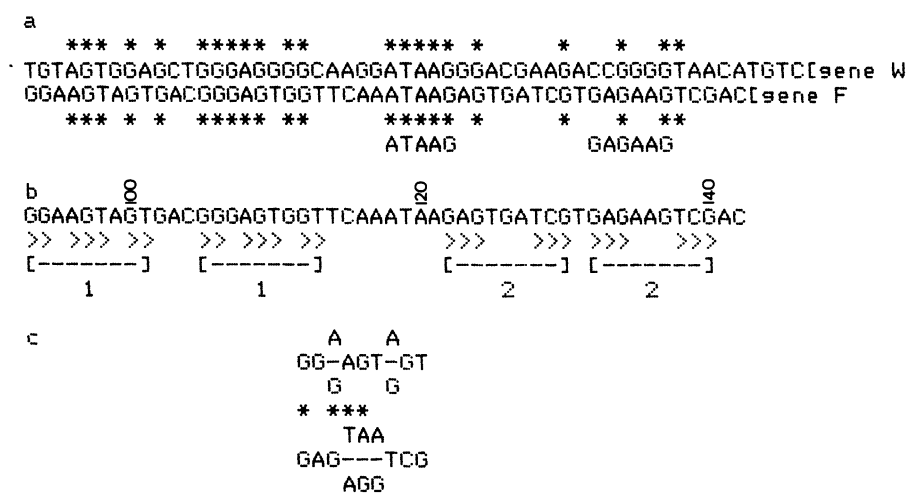


Fig. 2. Comparisons of the region flanking the 5' end of wheat and flax 5S RNA genes.
a) These two sequences are aligned so that the sequence ATAAG in wheat (W) and flax (F) are together. Identical bases are marked (*) over a region of 45 bases.
b) Candidate sequences for a direct repeat in this region are indicated and compared in (C). Note that if this sequence is regarded as a purine tract the probability of finding such direct repeats by chance is quite large.

repeats may either be analagous to the direct repeats of animal 5S DNAs or they may be artifacts of the purine richness of this region of DNA.

Differences between the long and short repeats

As can be seen from Figure 1 sequence differences between the long and short repeat are confined to the spacer region. The length variation between these two repeats is due to a 21 bp tandem duplication in the spacer of the long repeat type [Fig. 1], only one copy of this 21 BP sequence is present in the short repeat. Both of these observations contrast with the situation in the wheat 5S DNA where the length variation is a consequence of a tendem duplication in the coding sequence, and both the coding sequence and spacer exhibit sequence heterogeneity [15]

Of the fourteen single base differences observed between these cloned sequences two derive from differences between the copies of the 21 BP tandem duplication in the long repeat. Eleven of these fourteen differences are transition mutations which could have arisen as a consequence of deamination of 5 me C. Five of the eleven transition mutations involve the symmetrical dinucleotide CG (one of which inactivates a TaqI site in the short repeat).

and two involve the symmetrical trinucleotide $C^{A}_{T}G$ in a BstNI site. The inactivation of these sites has been confirmed by digestion of the plasmids pBG6 and pBG13 with TaqI and BstNI (data not shown, see methods). The digestion experiments also showed that all nine of the short repeats in pBG13 contain cleavable TaqI sites and one of the short repeats in pBG13 lacks a cleavable BstNI site. The overall sequence heterogeneity can be estimated from the sequences presented in Figure 1 as approximately 0.7%. Consequently the frequency of mutant restriction enzyme sites should be about 1% for tetranucleotide recognition sites and 0.1% for hexanucleotide recognition sites. Because the sequence variation is restricted to a few locations and not randomly distributed these estimates of restriction enzyme site mutation frequencies must be underestimates for those sites known to be affected, and overestimates for those sites not seen to vary.

Interspecific differences in the coding region

The coding sequence of the flax 5S RNA gene has been inferred by comparing the DNA sequence with available 5S RNA sequences, [3, 4] and the wheat 5S DNA sequence [15]. Sequence differences between flax and other plant 5S RNA genes are indicated in Figure 3. There are four positions at which deletions can be identified and three of these arise from differences in the location of the 5' and 3' end of the RNA. Only one deletion occurs within the coding sequence when it is compared with 5S RNA sequences, but this deletion is not observed when the sequence is compared with the wheat 5S DNA sequence. The wheat 5S RNA reportedly lacks this C residue [15]. There are seventeen positions at which differences between plant 5S RNAs can be found and of these fourteen show exclusively transition mutations and one position shows both transition and transversion mutations.

Near the 5' end of the flax 5S RNA the sequence CATAC is found, which is replaced by the sequence ACCAT in other higher plants [Fig. 3 and refs 3, 4]. It is noteworthy that this flax sequence is common to all known vertebrate 5S RNA sequences [3, 4].

Differences at the 5' flanking region

The alignment of the 5' flanking regions shown in Figure 2 gives 50% homology between this region in wheat and flax. Of the twenty single base changes identified within the region of 'homology' nine are transition mutations.

It can therefore be said that the abundance of transition mutations is limited to regions of good homology. This will be discussed later.

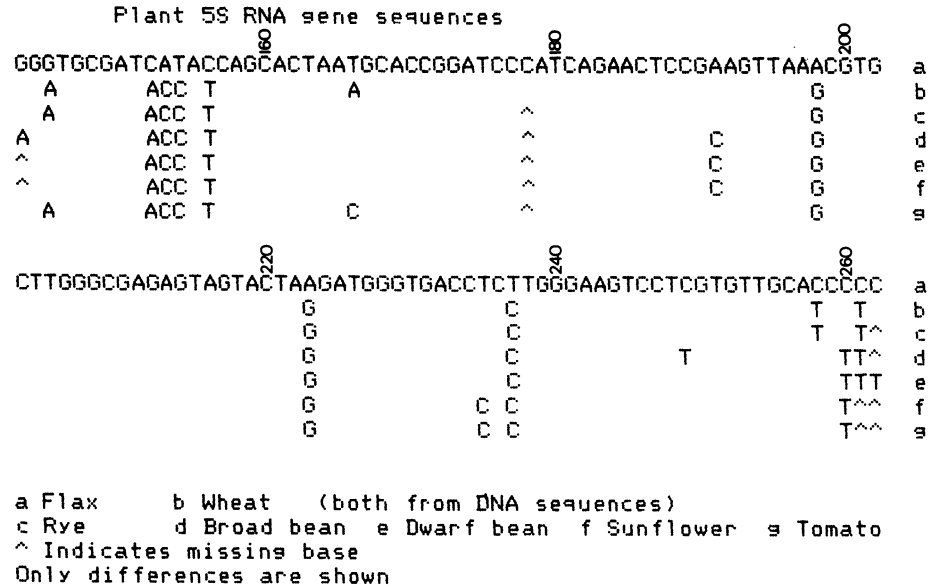
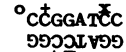


Fig. 3. Comparison of higher plant 5S genes.
The sequence of the flax 5S gene is compared with the sequence of other plant 5S DNA either as determined from DNA sequencing in the case of wheat [15] or from their RNA sequences [3, 4].

Pattern of Methylation

The predominance of transition mutations both in the gene and the spacer led us to determine whether the observed difference in the rate of accumulation of sequence heterogeneity between these regions could be a consequence of differential methylation of C residues. To this end it was decided to determine the cleavage frequency of BamHI, HpaII and MspI at their appropriate sites within the genomic 5S gene cluster. One HpaII/MspI site overlaps a BamHI site (a feature of all plant 5S RNA studied to date). At this site methylation of one C residue will block both the enzymes BamHI and MspI. This permits two estimations of the frequency of this methylation event in genomic DNA:

In the sequence  MspI is blocked by methylation at C residues marked O BamHI is blocked by methylation at C residues marked X and HpaII is blocked by methylation at C residues marked + [17]. Because of the low sequence heterogeneity of the 5S DNA ($\sim$ 0.7%) mutation as a source of inactivation of these restriction enzyme sites is considered unlikely.

As the HpaII/MspI sites map in different RsaI fragments [Fig.1, Fig.6], the fraction of each of these two smallest RsaI fragments remaining in Rsa/HpaII or Rsa/MspI double digests is the frequency at which these sites are blocked. [Fig. 4 fragments arrowed]. Similarly the fraction of the larger (200-220 BP) of these two RsaI fragments remaining in RsaI/BamHI double digests gives the frequency at which the BamHI site is blocked. The amount of the smaller RsaI (140 BP) fragment should be unaltered by the BamHI enzyme.

Microdensitometry of these bands in double digests compared with the same bands in a dilution series (shown in Figure 4) permits the estimation of the frequency of cleavage by the enzymes BamHI, HpaII and MspI. For the BamHI double digest the 140 BP fragment acts as an internal control, but this fragment is cleaved by MspI so the estimations of MspI cleavage can only be verified by comparing the predicted and observed ratios of the 300 BP fragments in RsaI/BamHI and RsaI/MspI double digests (Fragment arrowed in Fig. 4). In this experiment both these controls behaved as expected, within the errors of measurement. This method did not permit a determination of the frequency of HpaII cleavage and previous experiments [2] show that this cleavage frequency is small. The cleavage frequency at the HpaII sites could not be determined separately but if the frequencies are assumed to be equal it is possible to interpret these previous experiments to suggest an average HpaII cleavage frequency as 0.026 ± 0.002.

The double digest experiment shown in Figure 4 suggested that the cleavage frequency of MspI at the site which overlaps a BamHI site was 0.65 ± 0.10, and the MspI cleavage frequency at the other site was 0.55 ± 0.11. These data suggest that previous estimates of MspI cleavage frequency were too low [2]. This may be a consequence of the extensive (97-100% methylation of C at the CG sequence in these sites reducing the MspI reaction rate. In these experiments 100 fold excess of MspI was used whereas previously only a 36 fold excess was used.

The BamHI cleavage frequency estimated from the double digest experiment [Fig. 4] was 0.83 ± 0.04. A method for determining restriction enzyme site cleavage frequencies from one terminal and several partial digests has been previously described [2]. Such an experiment with BamHI is shown using hybridisation to blots of flax genomic DNA [Fig. 5]. The partial as well as the terminal cleavage frequencies are shown. This experiment suggests that the BamHI cleavage frequency of genomic flax 5S DNA is 0.84 ± 0.05 in good agreement with the estimate for double digests. These data

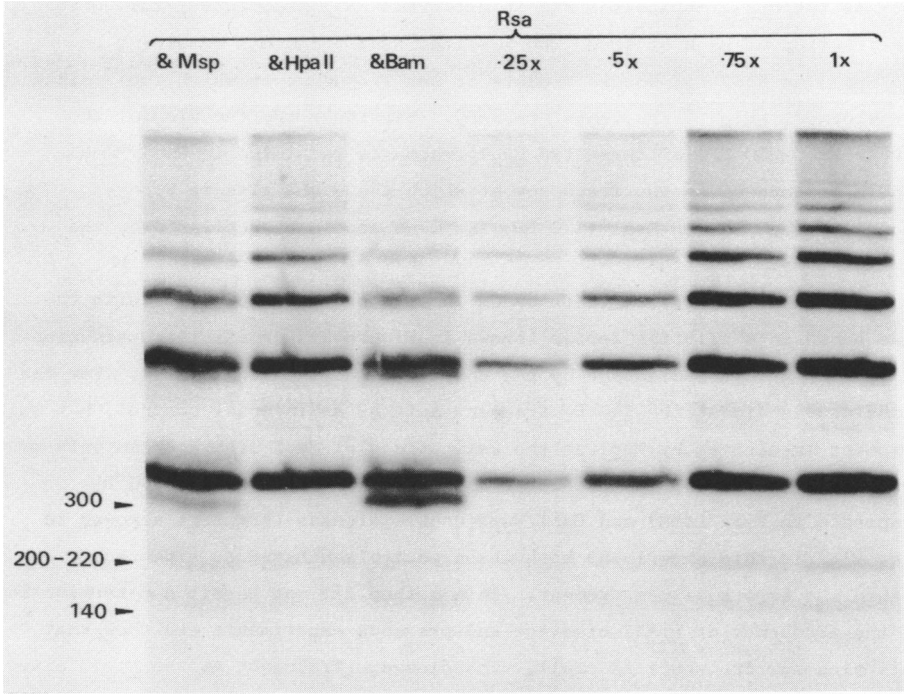


Fig. 4. Autoradiograph of Southern transfer of RsaI, RsaI and BamHI, RsaI and HpaII and RsaI and MspI digests of flax genomic DNA (line P1 see ref 1) hybridized with pBG13. The bands arrowed were used for microdensitometry. Other bands are from monomer, dimer, trimer etc. of the 5S DNA repeat in these digests. The calibration dilution series is shown. 1µg of DNA was loaded in the double digest tracks and in the RsaI single digest marked 1x. The other tracks were: .75x - 0.75µg; .5x - 0.5µg .25x - 0.25µg. The gel was 2% agarose, 2 cm sample wells were used to aid microdensitometry.

are summarized in Figure 6.

The frequency of C methylation at CpC in the HpaII/MspI sites similar in both the gene and spacer region. Approximately 20% of these C residues are methylated (if all site inactivation is a consequence of DNA methylation). This is approximately the same as the overall proportion of 5 me C in flax genomic DNA [14]. Because of the low cleavage frequency of HpaII it has not been possible to distinguish between the cleavage frequencies at the two sites; but because 97% of the sites are blocked the difference in C methylation at these two positions cannot be more than 3%. These experiments suggest there is no difference in the extent of C methylation between the 5S gene and its spacer.

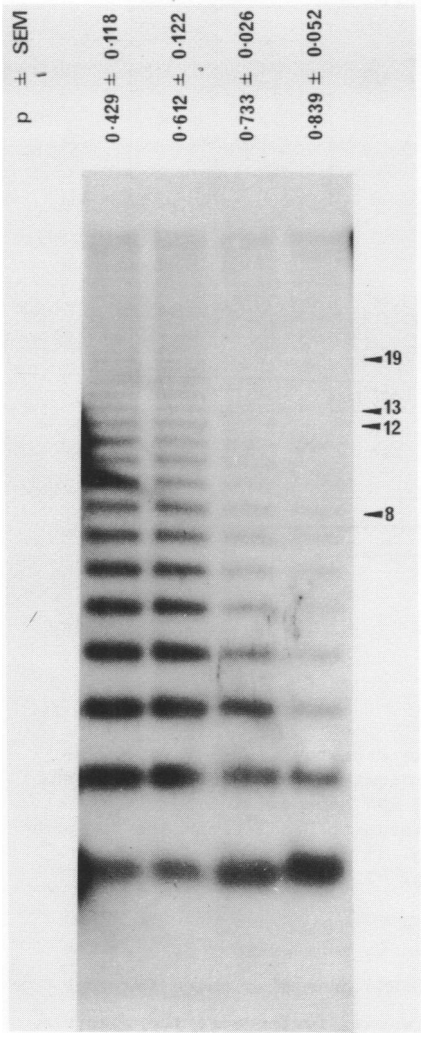


Fig. 5. Autoradiograph of a series of BamHI digests of flax genomic DNA (line P1 see ref 1, Southern transfer from 1% agarose gel). These digests are tending towards completion in the final track. Extents of digestion (p) were determined as previously described [2] and are indicated for each track. The position of expected single copy fragments are indicated for each extent of digestion; the presence of larger fragments in all tracks is indicative of non-randomness in the distribution of inactivated BamHI sites.

DISCUSSION

The flax 5S gene repeat shares many features with other known 5S genes, both in plants and in animals. The putative gene sequence appears to be more different from known plant 5S RNA sequences than these 5S RNA sequences are from each other [3, 4], but because only a small number of differences are known neither the phylogenetic or statistical significance of this can be assessed. Furthermore most of this 'excessive' difference of the flax 5S RNA gene is due to the sequence CATAC near the 5' end of the non-coding strand which is replaced by ACCAT in all other known plant 5S RNA genes

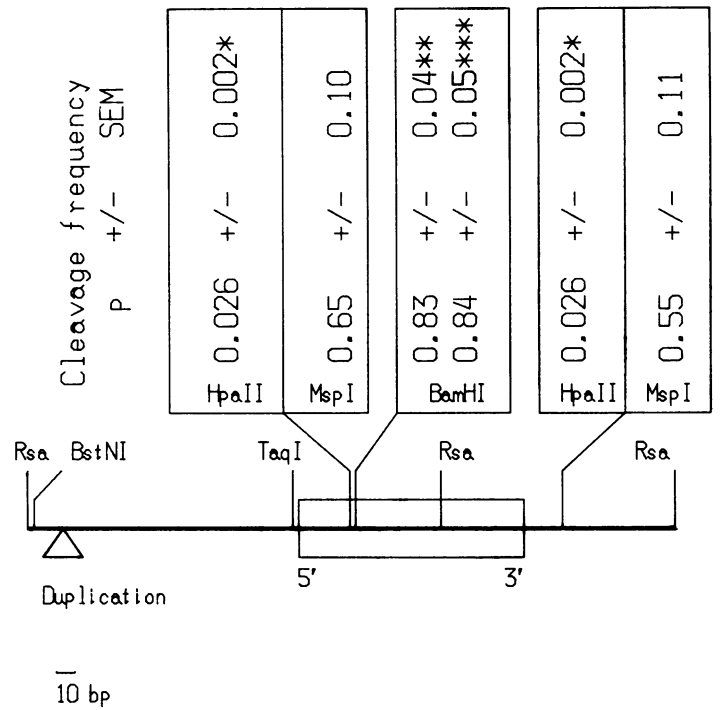


Fig. 6. Summary map of the flax 5S RNA gene repeat. Restriction enzyme si tes are shown and cleavage frequencies indicated for the HpaII, MspI and BamHI enzymes. The 5' and 3' ends of the 5S RNA gene are indicated.
* estimated from HpaII/MspI comparison [2] assuming HpaII sites are the same.
** estimated from Rsa/Bam double digest (Fig. 3)
*** estimated from Bam digestion series (Fig. 4).

[Fig. 1], however the flax sequence is common in vertebrates [3, 4]. The regions abutting the 5' and 3' ends of the 5S RNA gene also share features with other 5S RNA genes both in animals and wheat, furthermore the possible direct repeats near the 5' end of the gene may be another way in which this flax gene is more akin to animal 5S RNA genes than is wheat.

Cytosines in both the gene and spacer regions are methylated to a similar extent. Transition mutations are responsible for most of the observed sequence differences in the spacer region of the flax 5S gene repeat. Most single base changes found by comparing the flax 5S gene with the sequence of the 5S gene of other species are also transition mutations. Similarly the sequence heterogeneity observed in the wheat 5S RNA gene is largely that generated by transition mutations [15]. The opportunities for the operation of this major mutational pathway are similar in both the gene

and spacer regions. Because the gene sequence has not been found to vary between species to the same extent as the spacer sequence then, as the opportunities for transition mutations are similar in both regions, it must be concluded that the gene sequence is under selective restraint. This selective restraint must operate where a two fold reduction in 5S gene copy number has no observable phenotypic effect, and where the total 5S gene copy number is about 80,000 per 2C nucleus [1].

This apparent paradox can be resolved if it is accepted that a 'correction mechanism' [18, 19] operates to drive the gene cluster to homogeneity. One consequence of the operation of this mechanism can be copy number variation. If any newly arisen variant has a high probability of reaching fixation within the gene cluster then the selective pressures on that variant can be large. Thus interspecific differences can accumulate rapidly in the spacer but not the gene. Furthermore if transition mutations are the most frequent single mutational event then the observed sequence heterogeneity within the flax 5S spacer will be those mutations which have not yet had time to reach fixation and are therefore most likely to be transition mutations. Similarly the 5S RNA gene sequence variants observed will be either those tolerated by selection, or those which have not yet experienced the full force of selection exerted as a consequence of the fixation process. Therefore gene sequence variants are also most likely to be those which have arisen by transition mutation because this is the most frequent mutational event.

Acknowledgements

We thank Dr. M.J. Gait for providing the synthetic oligodeoxynucleotides used in the sequencing and C.A. Cullis and D.R. Davis for reading and commenting on the manuscript.

*Present address: Department of Botany and Plant Pathology, Purdue University, West Lafayette, IN 47907, USA

REFERENCES

1. Goldsbrough, P.B.; Ellis. T.H.N. and Cullis, C.A. (1981) Nucl. Acid. Res. 9, 5895-5904.
2. Ellis, T.H.N. and Goldsbrough, P.B. (1981) Nucl. Acid. Res. 7, 1551-1558.
3. Erdmann, V.A. (1981) Nucl. Acid. Res. 8, r31-r47.
4. Erdmann, V.A. (1981) Nucl. Acid. Res. 9, r25-r42.
5. Coulondre, C.; Miller, J.H.; Farabaugh, P.J. and Gilbert, W. (1978) Nature 274, 775-780.
6. Bird, A.P. (1980) Nucl. Acid. Res, 8, 1499-1504.
7. Twigg, A.J. and Sherratt, D. (1980) Nature 283, 216-218.

8. Rothstein, R.S.; Lau, L.F.; Bahl, C.P.; Narang, S.A. and Wu, R. (1979) Methods in Enzymology 68, 98-109.
9. Sanger, F.; Coulson, A.R.; Barrell,B.G.; Smith, A.J.H. and Roe, B.A. (1980) J. Molec. Biol. 143, 161-178.
10. Fields, S.; Winter, G. and Brownlee, G.G. (1981) Nature 290, 213-217.
11. Duckworth, M.L.; Gait, M.J.; Goelet, P.; Hong, G.F.; Singh, M. and Titmas, R.C. (1981) Nucl. Acid. Res. 9, 1691-1706.
12. Hong, G.F. (1981) Bioscience reports 1, 243-252.
13. Korn, L.J. and Brown, D.D. (1978) Cell 15, 1145-1156.
14. Handbook of Biochemistry selected for Molecular Biology (1968). Ed: Sober, H.A. p. 1139. publ: Chemical Rubber Corporation 18901 Granwood Parkway Cleveland Ohio 44128.
15. Gerlach, W.L. and Dyer, T.A. (1980) Nucl. Acid. Res. 8, 4851-4865.
16. Tschudi, C. and Pirotta, V. (1980) Nucl. Acid. Res. 8, 441-451.
17. McClelland, M. (1981) Nucl. Acid. Res. 9, 5859-5866.
18. Brown, D.D.; Wensink, P.C. and Jordan, E. (1972) J. Molec. Biol. 63, 57-73.
19. Miller, J.R. and Brownlee, G.G. (1978) Nature 275, 556-558.