The gene distribution in the genomes of pea, tomato and date palm

Abdelali Barakat[a,b], David Tran Han[a], Abdel-Ali Benslimane[b], André Rode[c], Giorgio Bernardi[a,d,*]

[a] *Laboratoire de Génétique Moléculaire, Institut Jacques Monod, 2, Place Jusssieu, F-75005 Paris, France*
[b] *Laboratoire des Sciences des Aliments, Faculté des Sciences Semlalia, P.O. Box S15, Marrakech, Morocco*
[c] *Institut de Biotechnologie des Plantes, Bâtiment 630, Université Paris Sud, F-91405 Orsay, France*
[d] *Laboratorio di Evoluzione Molecolare, Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Naples, Italy*

Received 13 September 1999; received in revised form 5 November 1999

Edited by Takashi Gojobori

Abstract The vast majority of genes of maize, rice, barley and wheat are contained in long gene-rich regions (collectively called the 'gene space') separated by long gene-empty regions. The gene space covers a narrow, 0.8–1.6%, GC range, possibly because of the presence of abundant transposons. Here we report that the gene space is not an exclusive property of Gramineae, because it also exists in the large genome of pea (5000 Mb). Moreover, the gene space is not just dependent upon genome size, since a gene space is found in rice (415 Mb), but not in *Arabidopsis* (120 Mb), nor in two other plants investigated in the present work, date palm (250 Mb) and tomato (1000 Mb).

Key words: Isochore; Gene; Genome; Plant

1. Introduction

Previous investigations showed that the nuclear genomes of maize, rice, barley and wheat exhibit an unexpected distribution of genes, the vast majority of them being contained in long gene-rich regions (collectively called the 'gene space') separated by long gene-empty regions [1,2]. In the cereals investigated so far, the gene space only covers a 0.8–1.6% GC range (GC is the molar fraction of guanine+cytosine in DNA), and only represents 12–24% of the genome. The concept of gene space refers to the fact that genes are not distributed over all or, at least, over most of the genome, but are restricted to some genome regions, which are characterized by a narrow compositional distribution.

In contrast, in the very small (ca. 120 Mb) genome of *Arabidopsis*, (i) genes cover a broad, 8%, GC range and are distributed over about 85% of the CsCl main band of DNA; (ii) open reading frames (ORFs) are fairly evenly distributed in the long sequences (>50 kb) available; moreover, (iii) the GC levels of coding sequences (and of their third codon positions, GC_3) are correlated with the GC levels of their flanking sequences [3].

The different pattern of gene distribution of Gramineae (cereals) compared to *Arabidopsis* is understood [3]. Indeed, in Gramineae (i) the genome comprises many large gene-empty regions (made up of repeated sequences) separating long gene-rich regions (the gene space), as already mentioned; and (ii) the base composition of the intergenic sequences of the long gene clusters, possibly comprising abundant transposons [4], is responsible for the narrow compositional distribution of the gene space (which contains coding sequences covering a very broad compositional spectrum). In contrast, in the genome of *Arabidopsis*, repeated sequences and transposons are absent or very scarce, leading to a continuity of gene-rich regions and to an absence of the influence of transposons on the buoyant density of these regions. A scheme of the gene distributions just described is presented in Fig. 4 in Barakat et al. [3].

An interesting question raised by these observations is whether gene distributions similar to those found in Gramineae and *Arabidopsis* are specific to given families of plants. Another question concerns the possible correlation of the existence of a gene space with genome size and with the abundance of repeated sequences. In order to answer these questions, we have investigated the gene distribution in the genomes of tomato (1000 Mb) and pea (5000 Mb), in which repeated sequences represent about 20% [5,6] and 70% [7,8] of the genome, respectively, as well as in the small genome (250 Mb; B. Benslimane, personal communication) of a non-graminaceous monocot, date palm.

2. Materials and methods

2.1. DNA preparation

Seeds from pea (*Pisum sativum* var. *finale*) and tomato (*Lycopersicon esculentum* var. *monolabo*) were obtained from André Blondeau, Bersée, France, and from the Station d'Amélioration des Plantes Maraichères, INRA, Montfavet, France, respectively. Nuclear DNA was prepared from etiolated seedlings using the method of Jofuku and Goldberg [9] with minor changes. Date palm (*Phoenix dactyliphera* L.) was from the Institut de Biotechnologie des Plantes (IBP), Université Paris Sud, Orsay, France. Nuclear and total cellular DNAs were prepared from date palm embryos and etiolated leaves using the methods of Jofuku and Goldberg [9] with minor changes, and of Dellaporta et al. [10] with several modifications [11], respectively.

2.2. DNA fractionation

Nuclear DNA fractionation was performed using equilibrium centrifugation in Cs_2SO_4/BAMD gradients [12]; BAMD is 3,6-bis(acetatomercurimethyl)-1,4-dioxane. The buoyant density of the fractions in which genes were localized was determined by analytical ultracentrifugation in CsCl gradients. In the case of date palm, total cellular DNA was used with the shallow CsCl gradient technique [13].

2.3. Probes

Probes used for pea and tomato were *Arabidopsis thaliana* ESTs

*Corresponding author. Fax: (39)-81-2455807.
E-mail: bernardi@alpha.szn.it

Abbreviations: BAMD, 3,6-bis(acetatomercurimethyl)-1,4-dioxane; EST(s), expressed sequence tag(s); GC, molar fraction of guanine+cytosine in DNA; GC_3, average GC level of third codon positions of genes; ORF, open reading frame; PCR, polymerase chain reaction

PII: S0014-5793(99)01587-2

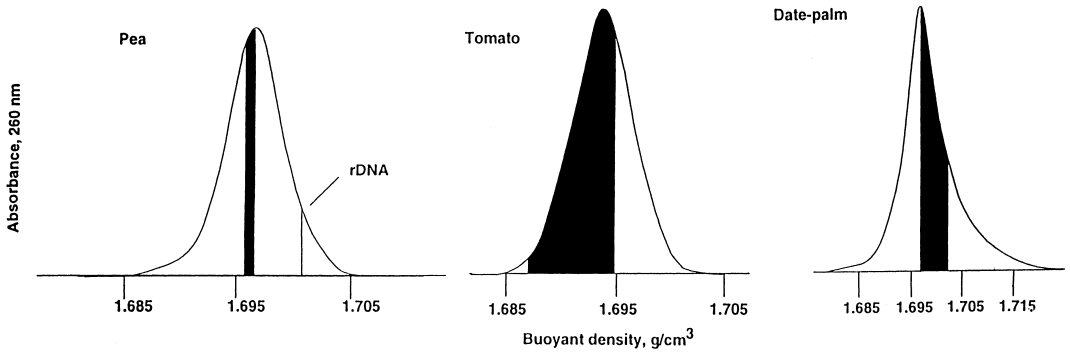

Fig. 1. CsCl profile of nuclear DNAs from pea, tomato and date palm as obtained by centrifugation in an analytical density gradient. The black areas corresponds to the gene space.

(expressed sequence tags), obtained from the Arabidopsis Biological Center, Columbus, OH, USA, and from Abdellah Hamel (IBP, Orsay). Some of the ESTs used corresponded to unknown genes (accession numbers are available upon request). Probes for date palm were maize ESTs obtained from T. Musket (Columbia, MO, USA). Sequences encoding pea legumins (leg J) were amplified according to [14].

2.4. Gene localization

Gene localization was performed by hybridization of probes on DNA fractions, as already described [2].

3. Results

3.1. Compositional distributions of DNAs from pea, tomato and date palm

Fig. 1 displays the CsCl profiles of nuclear DNAs from pea and tomato. Pea DNA ranges from 1.686 to 1.705 g/cm^3 (corresponding to ca. 40–49% GC, after correction for methylation; see [15]) with a maximum at 1.696 g/cm^3 (about 44% GC). Tomato DNA ranges from 1.685 to 1.702 g/cm^3 with a maximum at 1.693 g/cm^3, corresponding to ca. 35–42% GC (if methylation is ignored; real values are likely to be 2–4% higher). The CsCl profile from date palm (Fig. 1) ranges from 1.685 to 1.715 g/cm^3, approximately corresponding to 35–55% GC (not corrected for methylation) with a maximum at 1.697 g/cm^3. Table 1 summarizes these values.

3.2. Gene distribution in the genome of pea

Twenty-five ESTs corresponding to *Arabidopsis* genes whose GC_3 values covered a 29–69% range [3] were hybridized on pea DNA fractions from Cs_2SO_4/BAMD gradients digested by *Eco*RI. Because of the similarity of the gene distributions of pea, tomato and *Arabidopsis* and of the correlation between the GC_3 of homologous genes from dicots [16], one can consider that the probe sample used in this study covered the whole range of the gene distribution of pea and tomato. The results obtained (see Fig. 2, for an example) show that the hybridization maxima corresponded to fractions 4 and 5, having modal buoyant densities of 1.6961 and 1.6966 g/cm^3 (corresponding to 50–51% GC). These fractions represent 20% of the pea genome. The results obtained with the ESTs mentioned above were confirmed by localizing 30 additional *Arabidopsis* ESTs corresponding to unknown genes. rDNA genes were localized in a GC-rich fraction having a buoyant density of 1.705 g/cm^3. Genes encoding legumins were localized in fractions 4 and 5, except for one gene which was found in fraction 8.

3.3. Gene distribution in the tomato genome

Hybridization of the 25 *Arabidopsis* ESTs mentioned above showed that tomato genes are localized in Cs_2SO_4/BAMD fractions 2–7 (see Fig. 3 for an example) which cover a 5% GC range (between 1.688 and 1.693 g/cm^3) and represent 60% of the genome.

3.4. Gene distribution in the date palm genome

Twenty-one maize ESTs corresponding to the genes listed in Table 2 and ranging in GC_3 values from 37% to 95% [1] were hybridized on DNA fractions obtained by centrifugation in shallow CsCl gradients [13]. The results showed that the genes analyzed were localized in DNA fractions having buoyant densities between 1.698 and 1.7035 g/cm^3 covering a broad, 6% GC, compositional range (Fig. 1). The mean of the distribution is 1.698 g/cm^3, a value close to the maximum of the DNA fragment distribution. rDNA genes were localized in a GC-rich fraction having a buoyant density of 1.705 g/cm^3.

4. Discussion

4.1. The plant DNAs

The modal buoyant density of pea nuclear DNA (1.6956 g/cm^3) is close to that (1.6961 g/cm^3) previously reported [15] and slightly higher than that (1.6945 g/cm^3) reported in [17]. In the case of tomato (1.693 g/cm^3) and date palm nuclear DNAs (1.697 g/cm^3), buoyant densities were not previously reported.

It should be noted that the average GC levels of coding sequences of pea, 44.0%, and tomato, 44.9% (see [16]), are very close to those of the corresponding CsCl peaks, about

Table 1
CsCl analysis of nuclear DNA

	ρ, modal (g/cm^3)	ρ, range (g/cm^3)	GC (%)	GC range (%)
Pea	1.696	1.685–1.705	44	40–49
Tomato	1.639	1.685–1.702		35–42[a]
Date palm	1.697	1.685–1.715		35–55[a]

[a]Not corrected for methylation; actual values will be slightly higher.

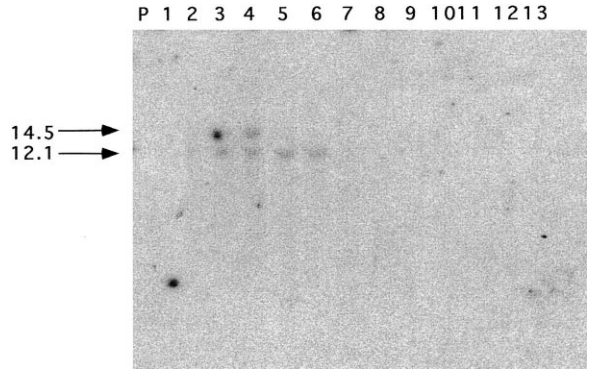

Fig. 2. Localization of a cDNA sequence corresponding to heat shock protein (HSP 76) on pea DNA fractions. Arrows (Kb) indicate hybridization bands. Lanes: P, pellet; 1–13, fractions.

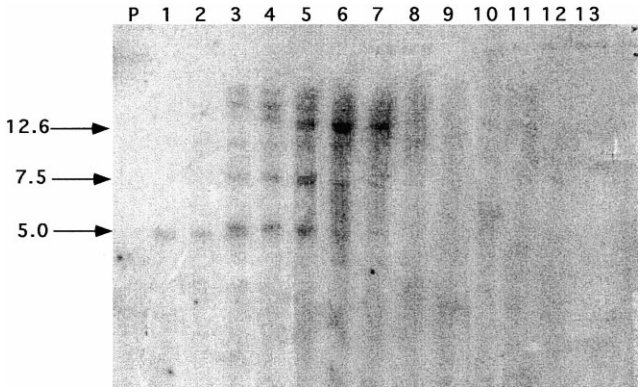

Fig. 3. Localization of three ESTs on tomato DNA fractions. Arrows indicate hybridization bands. Lanes: P, pellet; 1–13, fractions.

44%, and about 42% (the latter being an estimated value; see Section 3). These levels are much lower than those found (59–61% GC) in Gramineae [16,18,19].

4.2. The gene distribution in the pea genome

Hybridization experiments showed that almost all the pea genes detected by heterologous hybridization with *Arabidopsis* ESTs are located in DNA fractions covering a very narrow, 1%, GC range. Incidentally, these results are in agreement with previous very limited data [20], which showed that all genes tested (glutamine synthase, ribulose 1,5-bisphosphate carboxylase, alcohol dehydrogenase and legumin A, with the exception of the sequence encoding chlorophyll *a/b* binding protein) were located within a 1% GC range. These fractions correspond to 20% of the genome and represent the 'gene space' of pea genome. Seed storage protein genes (legumins), unlike those of Gramineae, were found in the gene space, except for one gene which was localized in a fraction higher in buoyant density than that of the gene space.

Table 2
Genes localized in compositional fractions of date palm DNA

Name	Gene GC_3	DNA ρ (g/cm^3)
Kinase	37	1.700
GTP1	41	1.698
Clx1	46	1.699
mdh4	46.5	1.699
RPL19	50	1.699
Grx1	51	1.698
rnp1	53	1.700
ATPase 1	56.5	1.699
NDME	59	1.699
Enolase	60	1.702
POD	69	1.7015
GAPDH	72	1.703
Sucrose P S	78	1.702
Knotted	80	1.702
Cab	94	1.7035
LHCB	95	1.7025
T12669	–	1.702
T12730	–	1.700
T12654	–	1.703
T12735	–	1.702
T12718	–	1.701

GTP1, GTP-binding protein; Clx1, calnexin; mdh4, malate dehydrogenase 4; RPL19, ribosomal protein L19; Grx1, glutaredoxin; rnp1, RNA binding protein 1; NDME, NADP-dependent malic enzyme; Enolase, enolase 1; POD, pyruvate orthophosphate synthase; GAPDH, glyceraldehyde 3-phosphate dehydrogenase 1; Sucrose P S, sucrose phosphate synthase; Cab, light-harvesting chlorophyll *a/b* binding protein; LHCB, light-harvesting protein. T12669, T12730, T12654, T12735, T12718, unknown functions (C. Baysdorfer, 1993, personal communication to the Maize Genetic Database).

4.3. The gene distribution in the tomato genome

Tomato presents a different gene distribution compared to pea in that genes were localized in fractions covering a broad, 5%, GC range and representing 60% of the genome (see Fig. 3 for an example). This broad compositional distribution is associated with a lower level (20% of the genome under high stringency) of repeated sequences [5,6] in its genome compared to pea, and with a high level of non-transcribed single-copy sequences [6]. Sequences encoding ribosomal genes were localized in a fraction which was GC-richer (fraction 9) than the gene space. This fraction also contains satellite DNA that, together with ribosomal genes, represents 5% of the genome [21]. This colocalization is in agreement with earlier results [22].

4.4. The gene distribution in the date palm genome

Date palm presents a widespread gene distribution relative to other monocots, like Gramineae. Indeed, genes detected by intergeneric hybridization with maize ESTs are located in DNA fractions again covering a broad, 6%, GC range and representing 41% of the genome.

5. Conclusions

The present results provide answers to the questions raised in Section 1 and add to the picture previously reported [3] of genome organization and gene distribution in angiosperms. Indeed, the gene distribution found in the genome of pea is similar to those previously reported for Gramineae, showing that the existence of a gene space is not confined to Gramineae. A first reason for this situation is that genes occupy only 20% of the large (5000 Mb) pea genome. This does not account, however, for the narrow compositional distribution of the gene space. A possible explanation for this might be the narrow compositional distribution of coding sequences of pea [16]. This cannot, however, be the explanation because the compositional distribution of coding sequences from pea is similar to that of tomato [16], which does not show the narrow gene space of pea. An alternative explanation is that, as in the case of Gramineae, abundant transposons in the intergenic sequences of the gene space are responsible for it. The

giant *Cyclops* transposons, present in 5000 copies in the pea genome [23], are likely to be the explanation, since their base composition is 42% GC.

In the case of tomato, the situation is different in that genes occupy a larger fraction of the relatively small (1000 Mb) genome, and repeated sequences are relatively scarce, non-transcribed single-copy sequences are abundant, and transposons do not appear to contribute to the compositional homogenization of this fraction. In contrast, this seems to be the case for the smaller (415 Mb) genome of rice. The gene distribution of other dicots might be similar to, or intermediate between, those found in the genomes of pea and tomato, because of the homology between their coding sequences, the colinearity of their genomes [24–26] and the similarities of their isochore patterns [19].

Acknowledgements: We thank the Arabidopsis Biological Resource Center, our colleague Abdellah Hamel and Alain Lecharny (Institut de Biotechnologie des Plantes, Université Paris Sud, Orsay, France) for the gift of *Arabidopsis* ESTs. We also thank E. Vierling (University of Arizona, Tucson, AZ, USA), T. Musket (Columbia, OH, USA) and D. de Vienne (Université Paris Sud, Orsay, France) for giving us the cDNA of the heat shock gene from pea and ESTs from maize. We also thank Nicolas Carels for his help in analytical ultracentrifugation and Bouchra Benslimane for communicating the size of the date palm genome to us.

References

[1] Carels, N., Barakat, A. and Bernardi, G. (1995) Proc. Natl. Acad. Sci. USA 92, 11057–11060.
[2] Barakat, A., Carels, N. and Bernardi, G. (1997) Proc. Natl. Acad. Sci. USA 94, 6857–6861.
[3] Barakat, A., Matassi, G. and Bernardi, G. (1998) Proc. Natl. Acad. Sci. USA 95, 10044–10049.
[4] SanMiguel, P., Tikhonov, A., Jin, Y., Moutchouslskaia, N., Zakharov, D., Melake-Berhan, A., Springer, P.S., Edwards, K.J., Lee, M., Avramova, Z. and Bennetzen, J.L. (1996) Science 274, 765–768.
[5] Zamir, D. and Tanksley, S.D. (1988) Mol. Gen. Genet. 213, 254–261.
[6] Peterson, D.G., Pearson, W.R. and Stack, S.M. (1998) Genome 41, 346–356.
[7] Murray, M.G., Cuellar, R.E. and Thompson, W.F. (1978) Biochemistry 17, 5781–5790.
[8] Murray, M.G., Peters, D.L. and Thompson, W.F. (1981) J. Mol. Evol. 17, 31–42.
[9] Jofuku, K.O. and Goldberg, R.B. (1988) in: Plant Molecular Biology – A Practical Approach (Shaw, C.H., Ed.), pp. 37–66, IRL Press, Oxford.
[10] Dellaporta, S.I., Wood, J. and Hicks, J.B. (1983) Plant Mol. Biol. Rep. 1, 19–21.
[11] Rode, A., Hartmann, C., Benslimane, A.A., Picard, E. and Quétier, F. (1987) Theor. Appl. Genet. 74, 31–37.
[12] Cortadas, J., Macaya, G. and Bernardi, G. (1977) Eur. J. Biochem. 76, 13–19.
[13] De Sario, A., Geigl, E.M. and Bernardi, G. (1995) Nucleic Acids Res. 23, 4013–4014.
[14] Turner, L., Hellens, R.P., Lee, D. and Ellis, T.H.N. (1993) Plant Mol. Biol. 22, 101–112.
[15] Matassi, G., Melis, R., Kuo, K.C., Macaya, G., Gehrke, C.W. and Bernardi, G. (1992) Gene 122, 239–245.
[16] Carels, N., Hatey, A., Jabbari, K. and Bernardi, G. (1998) J. Mol. Evol. 45, 45–53.
[17] Pivec, L., Horska, K., Vitek, A. and Doskocil, L. (1974) Biochim. Biophys. Acta 340, 199–206.
[18] Salinas, J., Matassi, G., Montero, L.M. and Bernardi, G. (1988) Nucleic Acids Res. 19, 5561–5567.
[19] Matassi, G., Montero, L.M., Salinas, J. and Bernardi, G. (1989) Nucleic Acids Res. 17, 5273–5290.
[20] Montero, L.M.J., Matassi, G. and Bernardi, G. (1990) Nucleic Acids Res. 18, 1859–1867.
[21] Ganal, M.W., Lapitan, N.L.V. and Tanksley, S.D. (1988) Mol. Gen. Genet. 213, 262–268.
[22] Chilton, M.D. (1975) Genetics 81, 469–483.
[23] Chavanne, F., Zhang, D.X., Liaud, M.F. and Cerff, R. (1998) Plant Mol. Biol. 37, 363–375.
[24] Tanksley, S.D., Ganal, M.W., Prince, J.P., de Vicente, M.C., Bonierbale, M.W., Broun, P., Fulton, T.M., Giovannomi, J.J., Grandillo, S. and Martin, G.B. et al. (1992) Genetics 132, 1141–1160.
[25] Cavell, A.C., Lydiate, D.J., Parkin, I.A., Dean, C. and Trick, M. (1998) Genome 41, 62–69.
[26] Lagercrantz, U., Putterill, J., Coupland, G. and Lydiate, D. (1996) Plant J. 9, 13–20.