The maize gene space is compositionally compartimentalized

Nicolas Carels*

Laboratório de Bioinformática, Universidade Estadual de Santa Cruz, Rod. Ilhéus/Itabuna km. 16, 45650-000 Ilhéus Bahia, Brazil

Received 8 April 2005; accepted 13 May 2005

Available online 20 June 2005

Edited by Takashi Gojobori

Abstract Previous investigations by Southern hybridization of cDNA with compositional DNA fractions showed that the majority of maize genes are located in a narrow GC range of DNA fragments and that the corresponding gene space was GC-richer than the region of the genome where zein genes are found. Here, we revisited the maize gene space using new data from the maize genome sequencing initiative. We found that the maize gene space itself is formed of two compositional compartments, i.e., a GC-poor and a GC-rich, characterized by a different distribution of Opie and Huck retrotransposons. The GC-rich compartment tends to be richer in GC-rich genes than the GC-poor compartment. However, the gene space compartimentalization of maize is much simpler than that of human.

Keywords: Maize; Genome; Gene space; Isochores; Retrotransposon; Zein

1. Introduction

The compositional approach to the genome organization of angiosperms showed that they are made up of compositional compartments [1–3]. Considering DNA fragments in the size range ~100 kb, *Gramineae* are richer in GC (guanine + cytosine) than the other angiosperms studied so far [1]. In maize, DNA fragments vary between ~34% and ~62% GC [4]. The major parts of the intergenic sequences are composed of various retroelement families that cover the whole GC composition of the genome [5–7]. Non-storage protein genes were preferentially found in ~100 kb DNA fragments with GC levels above the average GC% (46%) of the whole genome [4], within a range of GC variation far smaller than that of warm-blooded vertebrates. In contrast, storage-protein genes were generally found in a compositional compartment of lower GC level. The mosaic distribution pattern of gene islands in the so-called gene space was later confirmed in *Gramineae* by other authors [7–10].

In the following, the genome organization of maize has been revisited analyzing the distribution of (i) protein genes, (ii) α-zeins and (iii) Opie and Huck retrotransposons on a dataset recently available from the Maize Sequencing Initiative [7]. It has been found that the maize gene space itself can be divided in two compositional compartments, i.e., GC-poor and GC-rich. GC-poor genes and Opie tend to be more frequent in the GC-poor compartment while GC-rich and Huck are preferentially associated to the GC-rich one.

2. Materials and methods

Maize sequences were from GenBank (release 144, 15 October 2004). They were extracted with the ACNUC/QUERY retrieving system [11] using the Infobiogen server (see: http://www.infobiogen.fr). Large sequences (LS) higher than 100 kb and corresponding coding sequences (CDS) were extracted with the options "modify" (settled to $L > 100000$) and "t = cds", respectively. GC level of LS (GCLS%) and of CDS in third codon positions (GC3%) were calculated using CODONW [12].

A total of 188 LS covering 30 Mbp and corresponding to 108 CDS were extracted. The CDS with the annotations "unknown", "hypothetical", and of retrotransposon (identified by their name or function) or tRNA origin were not taken into account.

The intergenic sequences of each annotated LS were extracted and assembled by concatenation to generate shorter sequences that were compared for GC% to the original ones. The null hypothesis of average GC% equality between the 24 gene containing LS and the product of their intergenic sequence concatenation has been checked with the student *t* test in Excel comparing the *t* variable to the theoretical one for 23 degrees of freedom and a probability level of 0.975. The average GC3% per LS was calculated by multiplying the GC3% of each CDS by its relative size to the complete CDS pool per LS. The weighted correlation was calculated by multiplying each *xy* couple (the average GC3% and the GC% of the corresponding LS) by the number of gene per LS.

The gene (or CDS) density was estimated through the ratio of the sequence size (bp) resulting from the gene (or CDS) concatenation per LS to the size of the corresponding LS. The average gene (or CDS) size was weighted with the number of gene per LS. Weighted correlation and weighted average gene (or CDS) density were calculated without taking storage-protein genes into consideration. The orthogonal regression line of the plot of GC3% according to GCLS% was calculated as described by Jolicoeur [13].

To identify compositional domains in the gene space, we studied the effect of DNA sequence size on the compositional distribution of the retroelements Opie (46% GC) and Huck (63% GC). These retroelements were chosen because they are among the most numerous in the maize gene space and their GC level may explain its compositional organization [5,6]. To do this, the Opie and Huck location in the 188 LS were identified by BLASTN homology search using their *gag* and *pol* CDS as queries. Homologies that were at least 600 bp with Expected <0.0001 and identity level >80% were only considered. The homologous regions were retrieved by their coordinates and extracted together with their flanking sequences on the 5′ and 3′ sides. By this way, sequences of 10, 20, 40, 60, 80, 100, 120, 140 and 160 kb were automatically extracted using a Perl script. The 10 kb sequences are in the range of the retroelement size [6] and were a control to ensure consistency of their GC level with that of Opie (46%) and Huck (63%). This in silico methodology corresponds to a Southern hybridization of gene space DNA fragments of various class size with Opie or Huck probes. The statistical significancy of the difference between Opie and Huck proportions in GC-poor and GC-rich compartments for LS = 120 kb was established using the χ^2 test. The X^2 variable was compared to the theoretical one for 1 degree of freedom and a probability level of 0.95.

*Fax: +55 73 680 5226.
E-mail address: carels@uesc.br (N. Carels).

0014-5793/$30.00 © 2005 Federation of European Biochemical Societies. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.febslet.2005.05.063

3. Results

The maize genome heterogeneity covers an ~28% GC interval between ~34% and ~62% GC (Fig. 1A). LS distribution of our sample ranged between 36% and 52% GC with a mean value of 47.5% GC (Fig. 1B). We found that the genes containing LS cover a GC interval of ~10% (between ~44% and ~54%) with the highest concentration in an interval of ~4% (between ~46% and ~50% GC). Zein genes were exclusively found in a GC% interval of 2% between 44% and 46%. The correlation between GC3% of non-storage protein genes and GCLS% is rather high ($r = 0.61$) with a steep orthogonal regression line ($y = 14 * x - 605.59$) at 85.9° (data not shown). The weighted correlation between average GC3% and GCLS% was even higher: 0.70, with an orthogonal regression line $y = 8.5 * x - 344.64$ at lower slope and an angle of 83.3° (Fig. 1C). Both correlation coefficients were significant since $P < 0.001$. The average weighted gene and CDS densities were 12% and 6%, respectively. In addition, a difference of CDS densities between GC-poor (7%) and GC-rich (4.5%) compartments was detected.

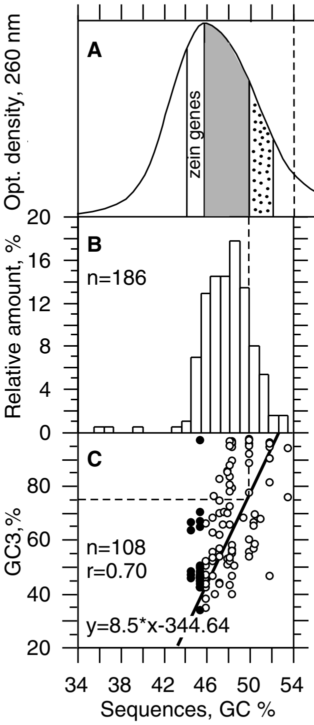

Fig. 1. Relationship between coding sequences and intergenic sequences in maize: (A) Adapted from Carels et al. [4]. It shows the sedimentation profile of DNA fragments (150 kb on the average) in CsCl. The relative amount is given in optical density at 260 nm and plotted according to GC% measured by HPLC. The three compositional compartments of the genome are shown with, from the left to the right, (i) the zein space, (ii) the gene space with the GC-poor compartment (gray area) and the GC-rich compartment (dotted area). The dotted line is at 54% GC. Above this line, protein genes do not seem to be present. It also corresponds to the rDNA location. (B) Distribution of the relative amount (%) of large sequences (>100 kb) according to their GC%. The dashed line is in continuation of that of panel C. (C) Plot of independent GC3% of coding sequences according to the GC% of the large sequences containing them. Black circle are for zein genes and open circle are for other protein genes. The correlation coefficient r is significant since $P < 0.001$. It corresponds to the weighted orthogonal line, i.e., each xy couple corresponding to an average GC3% value has been multiplied by the corresponding number of coding sequences. n is the number of coding sequences analyzed. y is the linear function for the orthogonal regression line. The angle of that line is 86°. The dashed line represents the separation between GC-poor and GC-rich genes. It is also used to identify GC-poor and GC-rich compartments.

Because of the high level and the statistical significance of the weighted correlation between GC3% and GCLS%, we identified 2 compositional compartments within the gene space (Fig. 1A). We determined their separation limit (50% GC) by reference to the distribution of GC-poor and GC-rich genes whose separation (dotted line in Fig. 1B and C) has been estimated at GC3 = 75% [14,15].

The vast majority (85%) of our gene sample was found in the GCLS% interval between 46% and 50% corresponding to the GC-poor compartment of the gene space accounting for 34.2% of the genome. The other remaining 15% were found in the GC-rich compartment that accounts for 10.6% of the genome. The GC-poor compartment is therefore ~3 times larger than the GC-rich one.

The GC% of the product of the intergenic sequences concatenation of each LS was not significantly different from that of the original LS containing genes ($n = 24$) since the t variable (0.084) from the one tail student test was below the threshold of the theoretical value (2.07). The orthogonal regression line corresponding to these variables was $y = 1.02 * x - 0.97$ and the correlation coefficient: 0.98. Neither the gene density (12% on the average) nor the CDS density (6% on the average) was high enough to dramatically influence the GC% of intergenic sequences.

By BLASTN homology search, it was found that the difference of GC% of the Opie and/or Huck containing sequences decreases with their increase in size (Fig. 2). At 10 kb size, each sequence distribution peaked at a GC% value corresponding to that of their corresponding retroelement, i.e., 46% GC for Opie and 63% GC for Huck (Fig. 2). The GC difference of each sequence distribution at the peak reached a plateau for sequence size >100 kb corresponding to 47% GC for Opie and 50% GC

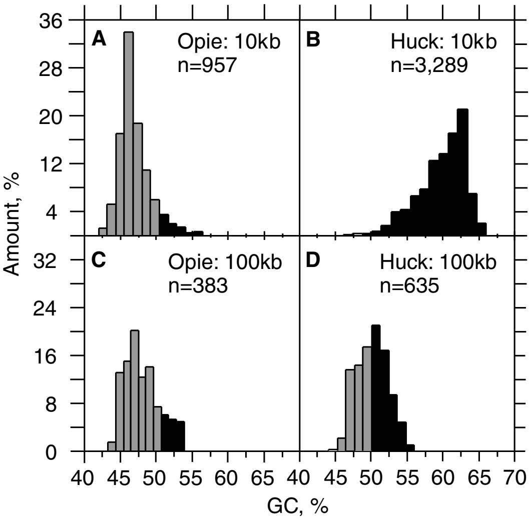

Fig. 2. Relationship between GC% and size of Opie and Huck containing sequences. Distribution of the relative amount of 10 kb sequences having a significant homologous region to Opie (A) and Huck (B) >600 bp according to their GC%. C and D are similar to A and B except that they concern 100 kb sequences. Gray and black bars represent GC-poor and GC-rich compartment of Fig. 1A, respectively.