Genetics and Molecular Biology, 24 (1-4), 161-167 (2001)

Sequence polymorphism from EST data in sugarcane: a fine analysis of 6-phosphogluconate dehydrogenase genes

L. Grivet[1,2], J.C. Glaszmann[2] and P. Arruda[1,3]*

Abstract

This paper presents preliminary results demonstrating the use of the sugarcane expressed sequence tag (EST) database (SUCEST) to detect single nucleotide polymorphisms (SNPs) inside 6-phosphogluconate dehydrogenase genes (Pgds). Sixty-four Pgd-related EST sequences were identified and partitioned into two clear-cut sets of 14 and 50 ESTs, probably corresponding to two genes, A and B, respectively. Alignment of A sequences allowed the detection of a single SNP while alignment of B sequences permitted the detection of 39 reliable SNPs, 27 of which in the coding sequence of the gene. Thirty-eight SNPs were binucleotidic and a single one was trinucleotidic. Nine insertions/deletions from one to 72 base pairs long were also detected in the noncoding 3' and 5' sequences. The soundness and the consequences of those preliminary observations on sequence polymorphism in sugarcane are discussed.

INTRODUCTION

The sugarcane genome is characterized by a high level of polyploidy. Current cultivars are derived from interspecific hybridization between a domesticated species, *Saccharum officinarum*, and a wild relative, *Saccharum spontaneum* (Daniels and Roach, 1987). *S. officinarum* has 2n = 80 chromosomes and is octoploid, while various ploidy levels from 5x to 14x have been reported for *S. spontaneum* where the basic chromosome number is x = 8 (Panje and Babu, 1960; D'Hont *et al.*, 1998). Cultivars are often aneuploids, their number of chromosomes generally being between 100 and 130, with 10% to 25% being contributed by *S. spontaneum* (D'Hont *et al.*, 1996; A. D'Hont, personal communication). It can thus be estimated that a gene with a single locus in the genome will be present in approximately 10 copies, each potentially corresponding to a specific haplotype, of which roughly eight or nine may be inherited from *S. officinarum* and one or two from *S. spontaneum*.

Precise knowledge about DNA polymorphism, especially single nucleotide polymorphism (SNP), is the first step toward access to the emerging high throughput genotyping technologies (Laken *et al.*, 1998; Lipshutz *et al.*, 1999), and expressed sequence tags (ESTs) constitute a useful raw material with which to detect SNP, which has already been explored on a large scale in human genomics (Buetow *et al.*, 1999; Marth *et al.*, 1999; Picoult-Newberg *et al.*, 1999).

The Sugarcane EST Genome Project (SUCEST) has generated about 260,000 ESTs derived from the sequencing of the 5 end, or both ends, of around 230,000 randomly cloned cDNAs (http://sucest.lbi.dcc.unicamp.br/en/). These cDNAs were recovered from 37 libraries constructed using different plant tissues from a limited number of cultivars, one of which contributed a little more than one half of the total number of the cDNA clones.

Although only a few distinct cultivars were used to construct the libraries, the polyploid nature of sugarcane should ensure haplotype diversity in the transcriptome if heterozygosity is high at the DNA level, a reasonable assumption based on marker data (Lu *et al.*, 1994a; Jannoo *et al.*, 1999), and if possible regulation of transcription does not excessively bias this pattern of diversity at the RNA level. From this perspective, genes with high expression levels should be the most interesting, since rare alleles will have a chance to be tagged and sequence redundancy will ensure a high level of confidence for any variation detected.

The work presented in this paper demonstrates the use of the SUCEST database as a source of sequence polymorphism in cultivated sugarcane. The 6-phosphogluconate dehydrogenase genes (Pgd) were chosen for detailed analysis because genetic determinism of the Pgd gene family is known in other grasses such as maize, rice and sorghum, and the number of Pgd-related sequences is high in the SUCEST database.

MATERIALS AND METHODS

Sugarcane ESTs, putatively orthologous to the cytosolic 6-phosphogluconate dehydrogenase *Pgd1* and *Pgd2* genes of maize (GenBank accession numbers AF061837 and AF061838, respectively), were identified in the

[1]*Centro de Biologia Molecular e Engenharia Genética, Universidade Estadual de Campinas, C.P. 6109, 13083-970 Campinas, SP, Brazil.*
[2]*CIRAD, UMR1096, TA 40/03, Avenue Agropolis, 34398 Montpellier cedex 5, France.*
[3]*Departamento de Genética e Evolução, Instituto de Biologia, Universidade Estadual de Campinas, C.P. 6109, 13083-970 Campinas, SP, Brazil.*
Send correspondence to Grivet L. E-mail: laurent@unicamp.br.

SUCEST database through Basic local alignment tool (BlastN) comparison (Altschul *et al.*, 1990) and assembled with the software Phrap (Green, 1996). A stringent assembly criterion was used (Telles and Silva, 2001) with *minscore* = 100, mismatch *penalty* = 15, and the *shatter_greedy* option was turned on. With these parameters, only highly identical reads will be grouped in a same cluster, differences being mostly due to sequencing error, and sequences of a same haplotype (allele) will cluster if they overlap sufficiently.

Phrap software produces a consensus sequence for each cluster where each nucleotide corresponds to the base with the highest Phred (Ewing *et al.*, 1998) quality value among all aligned reads. The number of cluster consensus sequences is lower than the number of original ESTs, their quality is improved and their length is increased, making this new data-set more amenable to further analysis.

The sequence identity for each pair of consensus sequences was established with the Blast program over the specific overlapping region of the pair. As the distinct sequences were known to be related (either homologous or homoeologous) the penalty for a mismatch was reduced to one in order to permit sequence alignment over the longest possible region and filtering for low complexity regions was not used. As the overlapping region is not the same for each pair of sequences, we assumed that the variation of sequence divergence along the gene is negligible compared to the global sequence divergence between paralogous gene sequences. Only identity values established over 100 bp or more were considered valid.

Consensus sequences were further grouped in order to tentatively identify those tagging a same gene. Two sequences were assigned to a same group if the identity over their overlapping region was higher than 98%. Based on the principle of transitivity, other sequences already grouped with either one of the two sequences were also included in the same group.

The number of genes that was deduced from the EST grouping pattern was tested against other independent sources of information, like isozyme and EST data, in maize, sorghum and rice. For that purpose, Pgd gene and EST sequences were recovered from GenBank for those species.

Finally, consensus sequences from a same gene were aligned and corresponding clusters were fused into a super-cluster with the software Consed (Gordon *et al.*, 1998). Sequence polymorphism was investigated in super-clusters. The strategy was inspired from Picoult-Newbert *et al.* (1999) in order to identify reliable variation. Only the high quality part of each read, as determined by Phrap, was considered for the detection of polymorphism at a given site. An SNP was declared as true when adjacent bases were aligned inside a window of 10 bp and the least frequent variant occurred at least twice with a Phred base quality value ≥ 20. The same method was used for INDEL (insertion/deletion) of a single base pair (bp). For INDEL of larger size, a single event was retained as true if the base quality was ≥ 20 for at least two adjacent bases.

RESULTS

Grouping homologous sequences

Maize *Pgd1* and *Pgd2* sequences were recovered from the GenBank database and were compared with BlastN to sequences in the SUCEST database in order to identify related sugarcane ESTs. Seventy sugarcane ESTs were obtained with a Blast score ≥ 100. These ESTs were first assembled with the Phrap program, resulting in 13 clusters containing between 1 and 14 ESTs each. Three sequences with very poor mean quality were discarded. Cluster consensus sequences were compared pair-wise with the BlastN program. Out of the 78 two-by-two possible pairs an identity value could be established for 54 (Table I). Consensus sequences were then assembled into two groups based on rules defined previously, one consisting of a single cluster of 14 reads (group A) and another made up of 10 clusters totaling 50 reads (group B). The most simple and straightforward interpretation of the whole set of data is that group A and group B represent two Pgd genes in sugarcane. Clusters included in group B were fused giving rise to a single super-cluster and a single consensus sequence was deduced for group B. Open reading frames of 1443 and 1446 bp were detected for the consensus sequence of groups A and B respectively.

Crosschecking gene-related groups with information from related species

The evaluation of sequence polymorphism is strictly dependent on the correct estimation of the number of paralogous genes and the correct assignment of each EST sequence to each gene. We thus compared our `two sugarcane Pgd genes' hypothesis with independent information on Pgd genes from maize, rice and sorghum.

In maize, isozyme and gene expression data along with gene sequences point to the existence of two genes, *Pgd1* and *Pgd2*, mapped on chromosomes 6 and 3 respectively (Goodman and Stuber, 1983; Bailey-Serres *et al.*, 1992; Redinbaugh and Campbell, 1998). These data are compatible with the diversity among maize EST present in the GenBank database (produced by the laboratory of V. Walbot, Stanford University). Among the 23 ESTs that were recovered (Blast score ≥ 100 with either gene *Pgd1* or *Pgd2*), 9 could be unambiguously assigned to the *Pgd1* gene and 14 to the *Pgd2* gene base on sequence identity (Figure 1A).

In rice, isozyme data indicate the presence of two genes, *Pgd-1* and *Pgd-2,* which have been mapped on chromosomes 11 and 6, respectively (Morishima and Glaszmann, 1990). A single rice Pgd sequence, likely to

Table I - Sequence identity between cluster consensus sequences (CCS) produced by Phrap for SUCEST ESTs related to *Pgds*. The number (n) of EST sequences in each cluster is given. For each pair of sequences the identity (upper number) and the length in base pairs (lower number) on which it was established is given. The grouping, performed according to the rules described in the text, is given in the last column.

CCS	n	CCS													grp
		1	2	3	4	5	6	7	8	9	10	11	12	13	
1	1	100% 325													-
2	2	-	100% 284												-
3	2	-	-	100% 479											B
4	3	-	92% 275	99% 479	100% 1165										B
5	3	-	93% 275	99% 479	99% 852	100% 852									B
6	5	-	92% 284	98% 218	95% 611	97% 540	100% 719								B
7	5	-	-	-	100% 101	-	-	100% 809							B
8	5	-	92% 284	99% 479	98% 810	99% 810	98% 527	-	100% 805						B
9	6	-	93% 284	99% 307	97% 637	98% 626	98% 546	-	99% 613	100% 632					B
10	6	73% 130	-	100% 104	99% 403	99% 159	-	97% 721	99% 125	-	100% 1220				B
11	7	63% 218	-	99% 142	97% 441	99% 197	-	97% 721	99% 163	-	98% 1146	100% 1242			B
12	8	63% 218	-	97% 288	97% 587	99% 343	-	98% 719	99% 309	96% 116	98% 1146	98% 1240	100% 1384		B
13	14	95% 314	82% 270	88% 476	89% 978	88% 733	88% 418	90% 591	88% 699	88% 507	89% 893	90% 930	89% 1077	100% 1910	A

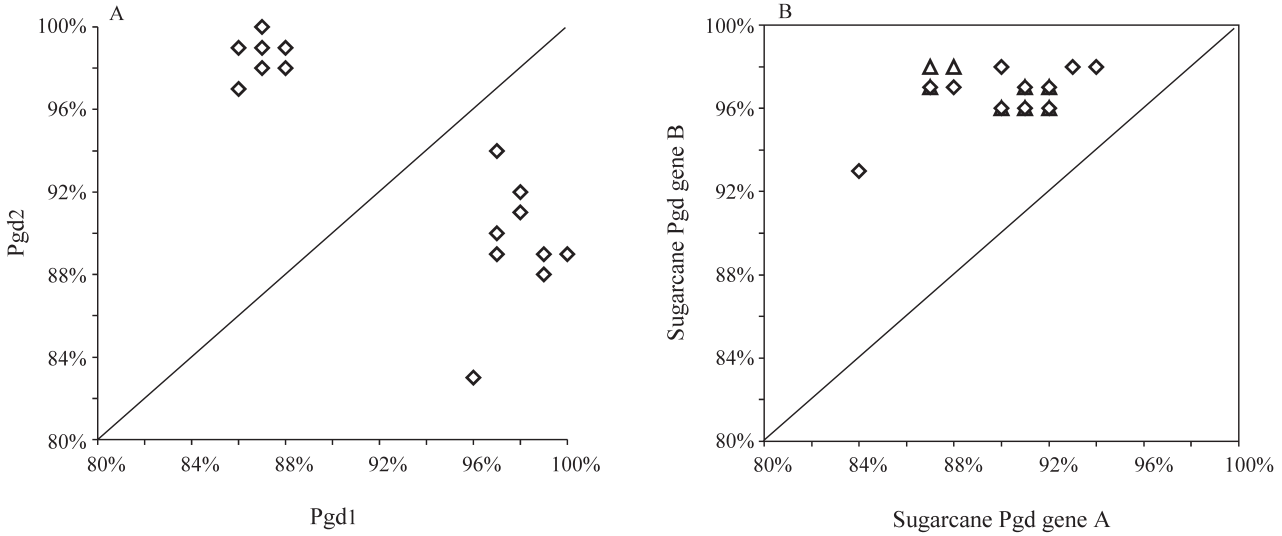

Figure 1 - Sorting of orthologous *vs.* paralogous ESTs in maize and sorghum. Sequence alignments was established between ESTs and known genes over the longest local alignment using Blast analysis without the filter option. ESTs were recovered from the GenBank database as explained in the text. **A** = Sequence identity between 23 maize ESTs and the maize *Pgd*1 and *Pgd*2 genes. **B** = Sequence identity between 21 *Sorghum bicolor* (triangles) and 25 *Sorghum propimquum* (lozenges) ESTs and the consensus sequences of sugarcane clusters A and B.

represent *Pgd-2* based on sequence identity with the mapped probe R2869 (Rice Genome Research Program database; http:// rgp.dna.affrc.go.jp/), is available in the GenBank database (accession number AP001552). Using the maize and rice Pgd sequences we recovered 12 rice Pgd-related ESTs from the GenBank database, of which nine had an identity ≥ 98% with the available rice Pgd sequence. The identity values for the three other ESTs were lower (89%, 94% and 96%) but this likely was because of the low quality of the sequences (data not shown). Based on the Southern hybridization pattern of probe R2869 (Rice Genome Research Program database; http:// rgp.dna.affrc.go.jp/), the *Pgd-1* sequence seems to have an identity < 90% with the *Pgd*-2 sequence. It is thus likely that all EST sequences available for rice are *Pgd-2* tags.

In cultivated sorghum, *Sorghum bicolor*, isozyme data indicate the presence of two *Pgd* loci (Morden *et al.*, 1989) as in maize and rice, but none has yet been cloned. We searched for ESTs in the GenBank database using maize, rice and sugarcane Pgd sequences as query sequences for Blast comparison. We recovered 21 ESTs for *S. bicolor* and 25 for a closely related wild species *S. propinquum*, which have all been produced by the Pratt laboratory from the University of Georgia. Their identities with the consensus sequences of sugarcane groups A and B indicate that all the ESTs of the two sorghum species are closer to consensus sequence of group B (Figure 1B). The ESTs were assembled for each sorghum species separately and for both species two non-overlapping clusters were obtained that were aligned to the two extremities of the sugarcane and maize Pgd genes.

A neighbor-joining tree was constructed including all relevant sequences for sugarcane, maize, rice and the two sorghum species, over the longest possible complete alignment inside the coding sequence (Figure 2). This showed that the consensus sequences of sugarcane groups A and B are probably orthologous of maize *Pgd2* and *Pgd1*, respectively, and that the sole Pgd gene detected for the two sorghum species is probably orthologous to *Pgd2* and sugarcane group B consensus. Rice *Pgd-2* may also be part of this lineage.

The absence of EST for the Pgd gene orthologue to sugarcane group A in rice, *S. bicolor* and *S. propinquum* may indicate that this gene is generally less expressed in all grass species, at least in the tissues used to produce cDNA libraries. This observation is in line with the highly unbalanced number of ESTs detected in sugarcane between group B and A (50 *vs.* 14, respectively) and to a leather extend with the slightly unbalanced number of ESTs observed between maize *Pgd2* and *Pgd1* (14 *vs.* 9).

Although incomplete, available data are compatible with a `two-gene' hypothesis for Pgds in maize, rice and sorghum. Those two genes are likely orthologous to those deduced from the partition of sugarcane ESTs alone. Therefore, the partition of sugarcane ESTs in genes A and B seems reasonable.

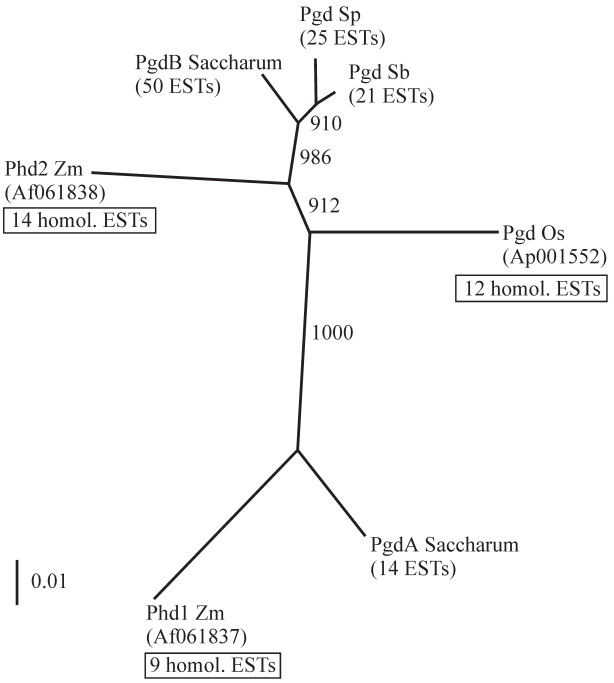

Figure 2 - Neighbor-joining tree of *Pgds* in grasses. The tree was constructed over the alignment of the 759 bp coding region, cumulating 395 bp from the 5end and 364 bp from the 3 end. The sequences of maize *Pgd1* and *Pgd2* and of rice *Pgd* were recovered from the GenBank database. For each of these three genes the number of homologous ESTs detected in the GenBank database is indicated, although these were not used to construct the tree. Sequences for sugarcane, *Sorghum bicolor* and *Sorghum propimquum* are consensus sequences of EST assemblies. The number of constitutive ESTs is indicated for each. Bootstrap values are indicated on internal branches, the total number of iterations being 1000. Zm = maize, Os = rice, Sb = *S. bicolor* and Sp = *S. propimquum*.

Scanning for variation among EST sequences

Alignments for groups A and B of sugarcane Pgd sequences were investigated in relation to SNP and INDEL polymorphism. This was performed manually by observing slides of aligned bases with the defined threshold parameters. Sequence alignment was occasionally corrected manually, especially near the ends of alignments, because INDEL delineation may not have always been well managed through the sequential cycles of group alignment and fusion performed with the Phrap software. For group A, the alignment contained 14 sequences generated from 11 independent cDNAs. A single SNP was observed. It located inside the coding sequence and induced no change in the specified amino acid. Group B contained sequence data for 44 independent cDNAs. A single-pass 5' end sequence was available for 38 cDNA, a double-pass 5' end sequence was available for three cDNAs and a single-pass for both 5' and 3' end was available for three cDNAs (Figure 2). In all, 50 ESTs were available. A total of 39 SNP were observed, four

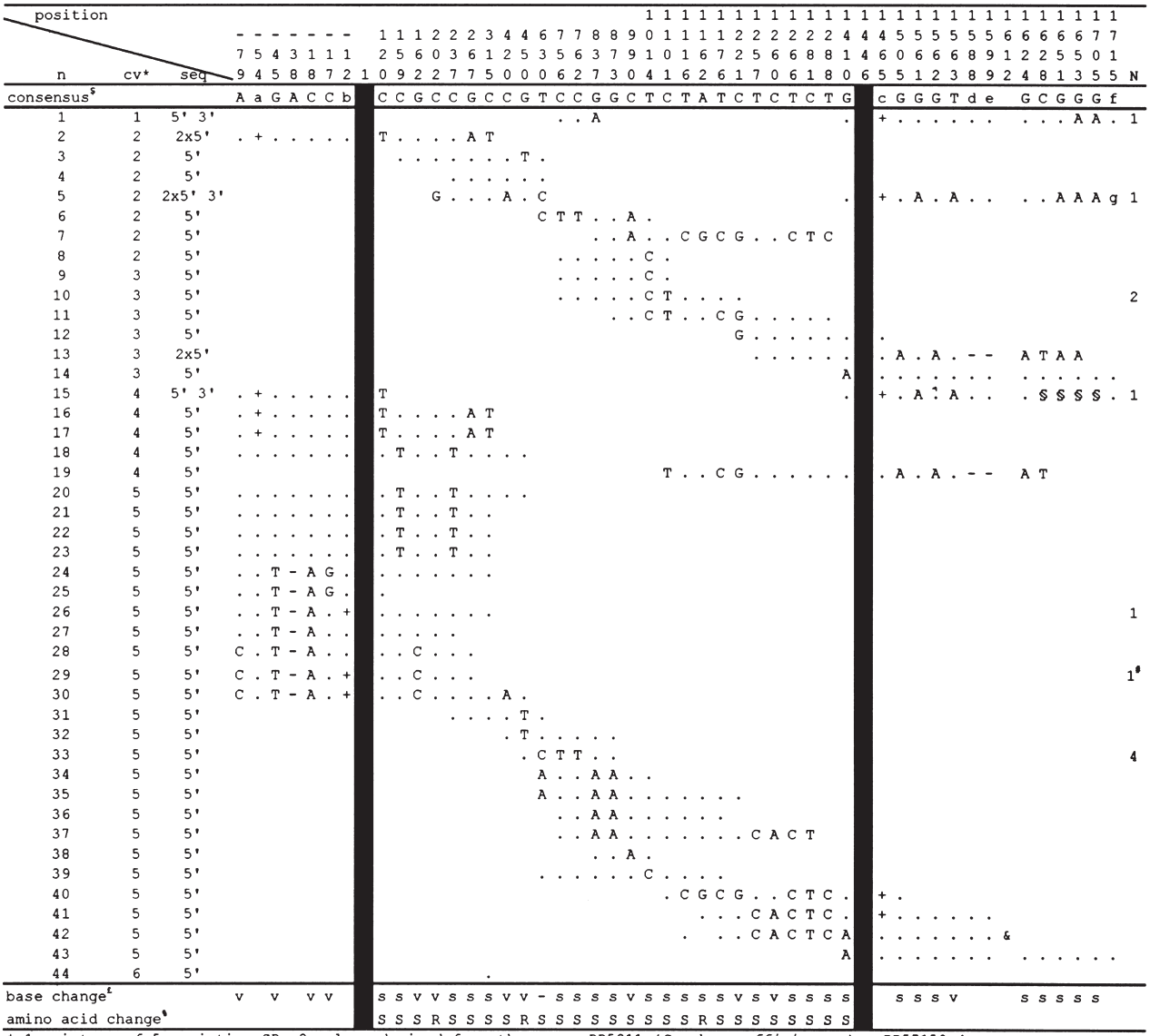

Figure 3 - Sequence variation in a sugarcane *Pgd* gene. Reliable sequence variation observed in the alignment of 50 ESTs generated from 44 independent cDNA probably representing a single 6-phosphogluconate dehydrogenase gene, named `B' (see text), is presented. Each line represents a cDNA sequence, identified in column one. Cultivar origin of each cDNA, cv, is given in column two. Sequencing number, (single or double) and direction (5 or 3) are given in column three. Each following column represents a reliable variant identified according to criterion defined in the text. The position of the variant is given relative to the translation initiation site. The two vertical bold lines represent translation initiation and termination. A point indicates that the cDNA sequence is identical to the consensus sequence (which gives the most frequent variant). A white space indicates missing data. INDELs are represented by a plus when they are present and by a minus when they are absent. In the last column is given the number, N, of variants appearing with a Phred quality higher than 20 in a single EST.

in the 5' leader, 27 in the coding region and nine in the 3' noncoding end. Thirty eight were binucleotidic and one was trinucleotidic (position 630). Sixty-nine percent of SNPs were transition *vs.* transversion, which is about the proportion observed in the nuclear genome of animals (Graur and Li, 1999). Twenty-four out of the 27 SNPs observed in the coding region were synonymous, while only three (11%) led to an amino acid replacement. The mean density of SNPs in the coding region was one every 54 bp. A total of nine INDELs were observed, four in the 5' leader

where the alignment covered 130 bp and five in the 3' end where the alignment covered 301 bp. The length of the INDELs varied between one and 72 bp. Interestingly, no INDEL were observed in the coding region. The poly (A)-addition site in one cDNA sequence started at position 1612, indicating the possible occurrence of alternate polyadenylation for this gene. For cultivars PB5211xPB57150-4, SP701143 and SP803280, several cDNA sequences were available. This permitted to identify a minimum number of haplotypes of 4, 2 and 6 respectively. In cultivar SP803280, for which the highest number of ESTs was available (24), it seems that two populations of haplotypes may coexist based on the alignment over the first 450 bp. This may possibly correspond to haplotype populations inherited from the two ancestral species *S. officinarum* and *S. spontaneum*. In several cases, a same haplotype was present in more than one cultivar (Figure 2), probably reflecting the high linkage disequilibrium expected inside sugarcane elite germplasm (Jannoo *et al.*, 1999).

Knowing the organ from which each cDNA was extracted did not permit to delineate clear-cut pattern of expression for the two Pgd genes in sugarcane. Both seemed however somewhat more expressed in seeds and roots relative to other organs. Available sequences extended upstream from the initiation codon for 40% of the cDNAs, indicating that they represent full-length inserts.

DISCUSSION

This paper is the first report of sequence variation in sugarcane and to our knowledge, one of the first report on the use of EST data to access sequence polymorphism in higher plants. It is important to appreciate the level of confidence that can be attributed to such data. There are two types of risks that should be considered: firstly that of confusing genuine polymorphism with sequencing errors, and secondly of confounding polymorphism at a unique locus with fixed differences between paralogous loci.

The first risk was controlled by the procedure used for reliable variant identification, which was a compromise permitting the elimination of obvious sequencing errors and the retention of most of the genuine variants. As such, it was not very stringent. However, as the number of sequence was high, especially in gene B, the threshold configuration was largely overshoot in most cases, regarding both the frequency of the least frequent variant and the base-calling qualities. Moreover, indirect verification of the validity of this methodology comes from the proportion of synonymous and non-synonymous amino acid changes in the coding region. A proportion of around 25% synonymous changes would be expected if the variation detected corresponded to random artifacts. In contrast, we observed a proportion of 89%, which is highly significantly different, and which can be explained by the purifying selection operating at non-synonymous sites.

The second risk may result from a bad estimation of the number of Pgd genes. The analysis of sugarcane EST data allowed us to divide the sequences into two populations, most probably corresponding to two genes. This bipartition is supported by data on the three closely related grass species, maize, rice and sorghum, for which two Pgd genes are the most likely hypothesis. Moreover, the contrasting number of ESTs observed for the two sugarcane genes are comforted by similar expression differences between the two orthologue genes, in sorghum, rice, and to a leather extend in maize.

The comparison of sequence polymorphism in *Sorghum* and *Saccharum* again provides indirect support for good partitioning of EST between genes. The EST sequences available for sugarcane come from two species (*S. officinarum* and *S. spontaneum*) that can easily intercross but which clearly represent divergent gene pools as assessed using restriction fragment length polymorphisms (Lu *et al.*, 1994b). EST sequences from two *Sorghum* species, *S. bicolor* and *S. propinquum*, were also available, that are close enough to intercross but that present contrasting RFLP patterns for most loci (Chittenden *et al.* 1994). The Pgd sequence of the two sorghum species differed for 11 nucleotides along the 759 bp of available coding sequence, while the number of SNPs detected in the orthologous region of sugarcane was 18. Since the species divergence seems roughly equivalent in both genera and more variants are expected to be detected in sugarcane because of the higher number of alleles observed, we can conclude that the variation detected in group B reasonably accounts for the polymorphism at a single gene. The low level of polymorphism detected in group A relative to group B, is likely due to the much lower number of EST sequences available.

Although the present data are limited to the analysis of a small family of two genes, they seem consistent and sound enough to suggest that the SUCEST database is a gold mine for sequence polymorphism detection in sugarcane. This arises from the combination of a biological property of the plant, which is its high level of ploidy, and a property of the database, which is huge in size and high in quality. Polymorphism will be especially easy to detect in highly expressed genes for which several tens of EST sequences, or more, are available. A corollary is that polymorphism may not be accessible for genes of interest that are poorly expressed.

It would be worthwhile inventorying SNP and INDEL variation, as it could become the raw material for future high throughput genotyping technologies. Such information is presently being collected in humans and a division of GenBank, dbSNP, has been specially devoted to the storage of such data. In sugarcane, the use of SNP and INDEL as marker tools will be challenging because the high ploidy level may prevent the straightforward application of emerging technologies developed in diploid model organisms. However, it is hoped that with the rapid evolu-

tion of technology and the increasing knowledge of sugarcane molecular variation patterns, specific tools will emerge which will possibly impact on future sugarcane breeding programs.

RESUMO

O presente estudo apresenta resultados preliminares demonstrando a utilização da base de dados de ESTs de cana-de-açúcar para detectar polimorfismo de base única (SNP para Single Nucleotide Polymorphism). Sessenta e quatro ESTs relacionados aos genes da 6-phosphogluconate deshydrogenases (Pgds) foram identificados e divididos em dois conjuntos bem delimitados, de 14 e 50 ESTs, correspondendo a dois genes, A e B. O alinhamento das seqüências do grupo A permitiu a detecção de um único SNP e o alinhamento das seqüências do grupo B permitiu a detecção de 39 SNP, incluindo 27 na região codificante do gene. Trinta e oito SNP foram bi-nucleotídicos e um único tri-nucleotídico. Nove inserções/supressões de um até 72 pares de base foram detectados nas regiões não-codificantes 3' ou 5'. A robustez e as conseqüências dessas observações preliminares são discutidas.

REFERENCES

Altshul, S.F., Gish, W., Miller, W., Myers, E.W. and **Lipman, J.** (1990). Basic local alignment tool. *J. Mol. Biol. 215*: 1651-1656.

Bailey-Serres, J., Tom, J. and **Freeling, M.** (1992). Expression and distribution of cytosolic 6-phosphogluconate deshydrogenase isozymes in maize. *Biochemical Genetics 30*: 233-246.

Buetow, K.H., Edmonson, M.N. and **Cassidy, A.B.** (1999). Reliable identification of large numbers of candidate SNPs from public EST data. *Nature Genet. 21*: 323-325.

Chittenden, L.M., Schertz, K.F., Lin, Y.R., Wing, R.A. and **Paterson, A.H.** (1994). A detailed RFLP map of *Sorghum bicolor* x *S. propinquum*, suitable for high density mapping, suggests ancestral duplication of sorghum chromosomes or chromosomal segments. *Theor. Appl. Genet. 87*: 925-933.

D'Hont, A., Ison, D., Alix, K., Roux, C. and **Glaszmann, J.C.** (1998). Determination of basic chromosome numbers in the genus Saccharum by physical mapping of ribosomal RNA genes. *Genome 41*: 221-225.

D'Hont A., Grivet, L., Feldmann, P., Rao, S., Berding, N. and **Glaszmann, J.C.** (1996). Characterisation of the double genome structure of modern sugarcane cultivars (*Saccharum* spp) by molecular cytogenetics. *Mol. Gen. Genet. 250*: 405-413.

Daniels, J. and **Roach, B.T.** (1987). Taxonomy and evolution. In: *Sugarcane Improvement through breeding* (Heinz, D.J., ed.) Elsevier, Amsterdam, pp. 7-84.

Ewing, B., Hillier, L., Wendl, M.C. and **Green, P.** (1998). Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. *Genome Res. 8*: 175-185.

Goodman, M.M. and **Stuber, C.W.** (1983). Maize. In: *Isozymes in plant genetics and breeding, part B*. (Tanksley, S.D. and Orton, T.J., eds.) Elsevier, Amsterdam, pp. 1-33.

Gordon, D., Abajian, C. and **Green, P.** (1998). Consed: A graphical tool for sequence finishing. *Genome Res. 8*: 197-202.

Graur, D. and **Li, W.H.** (1999). *Fundamentals of molecular evolution*. Sinauer Associates Inc., Sunderland, Massachussetts.

Green, P. (1996) PHRAP software: http://www.genome.washington.edu/.

Jannoo, N., Grivet, L., Seguin, M., Paulet, F., Domaingue, R., Rao, P.S., Dookun, A., D'Hont, A. and **Glaszmann, J.C.** (1999). Molecular investigation of the genetic base of sugarcane cultivars. *Theor. Appl. Genet. 99*: 171-184.

Laken, S.J., Jackson, P.E., Kinzler, K.W., Vogelstein, B., Strickland, P.T., Groopman, J.D. and **Friesen, M.D.** (1998). Genotyping by mass spectrometric analysis of short DNA fragments. *Nature biotechnol. 16*: 1352-1356.

Lipshutz, R.J., Fodor, S.P.A., Gingeras, T.R. and **Lockhart, D.J.** (1999). High density synthetic oligonucleotide arrays. *Nature Genet. Suppl. 21*: 20-24.

Lu, Y.H., D'Hont, A., Paulet, F., Grivet, L., Arnaud, M. and **Glaszmann, J.C.** (1994a). Molecular diversity and genome structure in modern sugarcane varieties. *Euphytica 78*: 217-226.

Lu, Y.H., D'Hont, A., Walker, D.I.T., Rao, P.S., Feldmann, P. and **Glaszmann, J.C.** (1994b). Relationships among ancestral species of sugarcane revealed with RFLP using single-copy maize nuclear probes. *Euphytica 78*: 7-18.

Marth, G.T., Korf, I., Yandell, M.D., Yeh, R.T., Gu, Z., Zakeri, H., Stiziel, N.O., Hillier, L., Kwok, P.Y. and **Gish, W.R.** (1999). A general approach to single-nucleotide polymorphism discovery. *Nature Genet. 23*: 452-456.

Morden, C.M., Doebley, J.F. and **Schertz, K.F.** (1989). Allozyme variation in old world races of *Sorghum bicolor* (Poaceae). *Amer. J. Bot. 76*: 247-255.

Morishima, H. and **Glaszmann, J.C.** (1990). Current status of isozyme gene symbols. *Rice Genet. Newsl. 7*: 50-57.

Panje, R.R. and **Babu, C.N.** (1960). Studies in *S. spontaneum*. Distribution and geographical association of chromosome numbers. *Cytologia* 25: 152-172.

Picoult-Newberg, L., Ideker, T.E., Pohl, M.G., Taylor, S.L., Donaldson, M.A., Nickerson, D.A. and **Boyce-Jacino, M.** (1999). Mining SNPs from EST databases. *Genome Res. 9*: 167-174.

Redinbaugh, M.G. and **Campbell, W.H.** (1998). Nitrate regulation of the oxidative pentose phosphate pathway in maize (*Zea mays* L.) root plastids: indication of 6-phosphogluconate deshydrogenase activity, protein and transcript levels. *Plant Science 134*: 129-140.

Walbot, V. (1999). Maize ESTs from various cDNA libraries sequenced at Stanford University (unpublished).

Yamamoto, K. and **Sasaki, T.** (1997). Large-scale EST Sequencing in Rice. *Plant. Mol. Biol. 35*: 135-144.