Theor Appl Genet (2003) 106:190–197
DOI 10.1007/s00122-002-1075-1

L. Grivet · J.-C. Glaszmann · M. Vincentz
F. da Silva · P. Arruda

ESTs as a source for sequence polymorphism discovery in sugarcane: example of the *Adh* genes

Received: 14 March 2002 / Accepted: 6 May 2002 / Published online: 1 October 2002

Abstract Expressed sequence tags (ESTs) have proven to be a valuable tool to discover single nucleotide polymorphism (SNP) in human genes but their use for this purpose is still limited in higher plants. Using a database of approximately 250,000 sugarcane ESTs we have recovered 219 sequences encoding alcohol dehydrogenases (*Adh*), which tagged 178 distinct cDNAs from 27 libraries, constructed from at least four different cultivars. The partitioning of these ESTs into paralogous genes revealed three *Adh* genes expressed in sugarcane, one *Adh2* and two *Adh1*. The soundness of the partition was carefully checked by comparison to external data, especially from the closely related sorghum. Analysis of polymorphism in the alignments of EST sequences revealed a total of 37 highly reliable SNPs in the coding and untranslated regions of the three *Adh* genes. In the coding regions, the mean occurrence of SNPs was one for every 122 base pair. A total of eight insertion-deletions was observed, their occurrence being limited to untranslated regions. These results show that EST data constitute an invaluable source of sequence polymorphism for sugarcane that is worth carefully collecting for the future development of new marker tools.

Keywords *Adh* · EST · INDEL · SNP · Sugarcane

Communicated by C. Möllers

L. Grivet (✉) · M. Vincentz · F. da Silva · P. Arruda
Centro de Biologia Molecular e Engenharia Genética,
Universidade Estadual de Campinas, PO Box 6010, 13083-970,
Campinas, SP, Brazil
e-mail: laurent@unicamp.br
Tel.: +55-19-3788-11-40, Fax: +55-19-3788-10-89

L. Grivet · J.-C. Glaszmann
CIRAD, TA 40/03, UMR 1096, Avenue Agropolis,
34398 Montpellier cedex 5, France

M. Vincentz · P. Arruda
Departmento de Genética e Evolução, Instituto de Biologia,
Universidade Estadual de Campinas, PO Box. 6109, 13083-970,
Campinas, SP, Brazil

Introduction

Sugarcane is an important cash crop cultivated in the tropics for its sugar-rich stalks. Cultivars are clones propagated by stem cuttings. They are polyploid and derive from interspecific hybridization between *Saccharum officinarum*, a domesticated species, and *Saccharum spontaneum*, a wild and vigorous relative. They are often aneuploids with chromosome numbers varying between 100 and 130, 15% to 25% being contributed by *S. spontaneum* (D'Hont et al. 1996). The chromosome pairing scheme at meiosis is complex, but disomy (i.e. allopolyploidy) is very unlikely (Grivet et al. 1996; Hoarau et al. 2001) indicating that all homologous chromosomes are susceptible to recombination. This genome organization implies that each single-copy gene is represented by around ten alleles, each of which potentially corresponds to a distinct sequence haplotype. Among the ten alleles, roughly eight or nine should be inherited from *S. officinarum* and one or two from *S. spontaneum*.

Sugarcane breeding was initiated more than 100 years ago. Interspecific crosses have been performed a limited number of times and, since then, breeding relies on intercrossing of current elite cultivars. As cultivars are clones and as the generation time is 10 to 15 years, few meioses had the opportunity to recombine founder chromosomes and a high level of linkage disequilibrium is therefore expected. This has been verified on a sample of Mauritian cultivars genotyped with mapped RFLP markers (Jannoo et al. 1999b). The configuration is thus favorable to the tracking and use of Quantitative Trait Alleles (QTA) in breeding programs without developing specific crosses dedicated to this task. It requires, however, a powerful high-throughput genotyping technology. Marker techniques such as RAPD (Al-Janabi et al. 1993), RFLP (Da Silva et al. 1993; Lu et al. 1994; Grivet et al. 1996) and AFLP (Hoarau et al. 2001; Lima et al. 2002) have already been used in sugarcane for genotyping or genetic mapping. However, RAPD lacks repeatability, RFLP and AFLP are not likely to be easily automated and the three techniques rely on gel electrophoresis, which does not

allow easy and unambiguous identification of alleles, especially when bench work is disconnected in time and space.

High-throughput genotyping technologies based on Single Nucleotide Polymorphism (SNP) or small-scale insertion/deletion (INDEL) could become efficient alternative tools in the future (Syvänen 2001). They require first identifying and localizing SNP and other local variations on the genome. An efficient strategy, which is extensively tested in humans, relies on the use of Expressed Sequenced Tags (ESTs) generated from various chromosome haplotypes (Buetow et al. 1999; Garg et al. 1999; Picoult-Newberg et al. 1999; Deutsch et al. 2001). In sugarcane, the Brazilian Sugarcane EST Project (SUCEST) recently produced about 250,000 ESTs derived from the sequencing of the 5′ end or both ends of approximately 230,000 randomly cloned cDNAs from 27 libraries (http://sucest.lad.ic.unicamp.br/en/). Four cultivars contributed to almost all cDNAs and a single one contributed for more than a half. Although the number of distinct genotypes is low, polyploidy should ensure allelic diversity, as heterozygozity is high (Lu et al. 1994; Jannoo et al. 1999a).

No data are yet available for sequence polymorphism in sugarcane genes. We are thus in an exploratory phase, and an important risk to consider is the possible confusion between orthologous alleles and paralogous genes. For that reason we concentrated our efforts here on one family of genes, the Alcohol dehydrogenase (*Adh*), for which the evolution pattern has been well studied in the Poaceae (Gaut et al. 1999), and for which expression level is high, so that many ESTs are available. In the grasses, the *Adh* gene family is usually composed of two or three genes: *Adh1* and *Adh2* probably diverged from a common ancestor before the radiation of grass species around 65 million years ago (Gaut et al. 1999). In some species one of the two genes is duplicated, as for example the case for *Adh2* in barley (Trick et al. 1988).

In this article we demonstrate, on the particular example of three genes of the *Adh* family, that a large collection of ESTs is a highly valuable source to discover sequence polymorphism, i.e. SNPs and INDELs, in sugarcane.

Materials and methods

Sugarcane ESTs encoding *Adh* genes were retrieved from the SUCEST database (http://sucest.lad.ic.unicamp.br/en/) with BLASTN (Altshul et al. 1990) using maize *Adh1* (GenBank accession number AF123535) and *Adh2* (X01965), and sorghum *Adh1* (AF124045), as query sequences. Histogram files were available for all ESTs. They were base-called with *Phred* (Ewing et al. 1998) to obtain the quality value associated with each residue. The quality value is directly related to the estimated error-probability for the base-call (Ewing and Green 1998). ESTs were then assembled with the software *phrap* (Green et al. 1994) using a stringent criterion: the *Minscore* parameter was given a value of 100, mismatch *penalty* was given a value of 15 and option *shatter_greedy* was turned on. With those criteria, a cluster assembles reads that overlap over a large portion and that tag a more-or-less unique haplotype. When they exist, sequence differences are usually not supported by high base-called quality values. *Phrap* establishes a consensus sequence for each cluster by choosing, at each site, the base with the highest quality inside the alignment. Consensus sequences established by *phrap* for each cluster are lower in number than the original crude EST sequences, their quality is improved and their length is increased, making them more amenable for further analysis. Consensus sequences were then compared pairwise with the program BLASTN in order to construct a matrix of sequence identity. As sequences are known to be either homo- or homo-eologous, the penalty for a mismatch was reduced to one and the filter for low complexity was not used to permit alignment over the largest possible region. As the overlapping region is not the same for all pairs, we assumed that the variation of sequence identity along the gene is negligible compared to the global sequence identity between paralogous gene sequences. The minimum overlapping length for an identity to be computed between two sequences was 100 bp. When smaller, the identity value was discarded and replaced by missing data. The most pre-eminent information present in the overall identity matrix was extracted through a Factorial Analysis (FA) performed with Statistica (1997).

Groups of sequences revealed by FA were used to establish the number of genes in the family, assuming that alleles of the same gene are more identical than alleles of paralogous genes. The reliability of the number of genes established in this way was then tested by comparing them to estimations established through independent sources of information, not only in sugarcane but also in other cereals, especially sorghum. For that purpose we retrieved various published sequences from GenBank, including rice (X16296) and pearlmillet (M59082) *Adh1* genes, the rice *Adh2* gene (AF172282) and various sorghum ESTs (http://dogwood.botany.uga.edu/~prattlab/). Base-calling error probabilities were not available for sorghum ESTs.

Clusters that assemble reads corresponding to a same gene were aligned with the software *consed* (Gordon et al. 1998) and were fused to give a super-cluster. Sequence polymorphism was investigated in super-clusters. For SNP, polymorphic sites were identified with a two-step procedure. The first step permitted the identification of sites that are polymorphic in a well-aligned sequence stretch. It was based on criteria inspired by Picoult-Newbert et al. (1999). A site was retained when surrounding adjacent bases were perfectly matched for all ESTs, inside a window of 10 base pairs (bp), and the least-frequent variant at the site occurred at least twice with a *phred* base quality value ≥20. The second step is intended to establish the level of reliability of each candidate SNP site retained in step one, based on statistics comparable across sites. This was performed with the software Polybayes (Marth et al. 1999), by computing for each candidate SNP site the probability P_{SNP} that an SNP is present at the site, depending on base composition, depth of alignment and *Phred* base-call qualities.

For INDEL of a single bp, the procedure was the same as for SNP. For an INDEL of larger size, an event was retained as true if the base quality inside the INDEL was ≥20 for at least two adjacent bases in a single read.

Results

Resolution of orthologous vs paralogous sequences

ESTs related to the *Adh* gene family were retrieved from the SUCEST database using as query sequences maize *Adh1* and *Adh2*, and sorghum *Adh1*. The threshold was a BLASTN score ≥100 with at least one of the three query sequences. A total of 219 EST sequences were obtained that tagged 178 independent cDNAs from 27 libraries: 178 were first-pass 5′ sequences, 15 were control reads, that is second-pass 5′ end sequences, and 26 were 3′ end

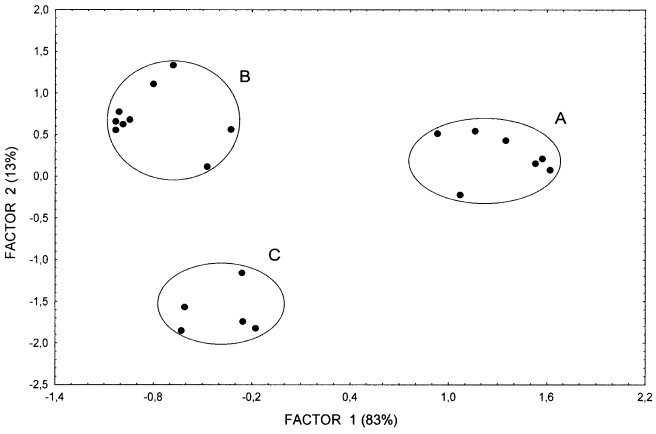

Fig. 1 Identity between 21 consensus sequences of *Phrap* EST clusters as revealed on the first plane of a Factorial Analysis. Each point represents a cluster consensus sequence. The first factorial axis is horizontal and is individualized sequence group A. The second factorial axis is vertical and separates sequence groups B and C

sequences of cDNA for which the 5′ end sequence already existed. The 219 ESTs were first assembled with *phrap*, giving 21 clusters that contained between 1 and 64 reads each. More than 70% of bases inside individual reads that presented a discrepancy with the cluster consensus sequence had a quality ≤20. Three reads were excluded because of their poor quality. Cluster consensus sequences given by *phrap* were trimmed from their terminal low quality regions. Then, the 210 (= 21 × 20/2) possible sequence pairs were compared pairwise with BLASTN over the longest local alignment. A sequence identity value was computed over more than 100 bp for 170 of them. These values were used to construct a 21 × 21 diagonal and symmetric square-matrix of identities. Missing data were scattered inside the matrix. A FA was performed over the smallest sub-matrix, complete for all rows (21 × 7 matrix). The first factor of the FA explained 83% of the variability, and was allowed to discriminate one group of seven cluster consensus sequences from the others (group A in Fig. 1). The second axis explained 13% and permitted the discrimination of two other groups (B with nine consensus sequences and C with five consensus sequences in Fig. 1). The cumulated variance explained by the first plane was 96%, while the third axis explained only 3% of the total variance. The most straightforward interpretation of this picture is the existence of three *Adh* genes in sugarcane, one, A, being more divergent from the two others, B and C. Clusters of each group were fused into a super-cluster with *consed*, resulting in the alignment of all ESTs tagging the same putative *Adh* gene. The total numbers of aligned ESTs tagging putative genes A, B and C, were 67, 123 and 26, respectively. In the three cases, the consensus sequence established for the putative gene revealed an open reading frame (ORF) of 1,140 bp. This length is the most-frequent and the largest observed for *Adh* coding sequences in other grasses (maize, sorghum, rice, pearl-millet), indicating that EST alignments obtained here very probably cover the complete coding sequence of the three putative *Adh* sugarcane genes. Due to the high number of reads, the sequences of ORFs were of high quality with a number of errors per kb estimated to be ≤10^{-3} with *consed*.

Polyadenylation was suspected for the three putative genes, as several positions of initiation for poly-(A) tails were observed among the ESTs. The putative gene B was the most expressed of the three genes as it presented the highest number of ESTs. It could be detected in all organs, except seeds. Putative gene C seemed to have a similar expression profile, also it was globally less expressed, and putative gene A seemed to be relatively more expressed in roots and less in aerial parts, as compared to B and C (data not shown).

Validation of genes with external data

The evaluation of sequence polymorphism critically depends on the correct estimation of the number of genes in the *Adh* family and on the correct assignment of each EST to each gene. We thus crosschecked the gene number evaluated from the sole sugarcane EST data with independent sources of information.

We first examined the genetic mapping information. In sugarcane, mapping of *Adh* genes through Southern hybridization of the maize *Adh1* gene revealed two loci in cultivar ‘R570’ (Grivet et al. 1996). In sorghum, two loci orthologous to the two loci detected in sugarcane have been reported (Whitkus et al. 1992; Dufour et al. 1997). In maize, map location of the two genes *Adh1* and *Adh2* has been reported on many occasions on chromosomes 1 and 4, respectively (MaizeDB; http://www.agron.missouri.edu/). A number of loci deduced from genetic mapping information must however be considered a lower limit, because (1) lack of polymorphism may prevent detection of some loci, and (2) the resolution of recombinational mapping is not high enough to distinguish possible duplications in tandem.

We then tried to investigate further the number of *Adh* loci in sorghum from available sequence data. Sorghum is a particularly attractive reference for sugarcane because it is a closely related diploid, inbred lines are available, chromosomes are highly collinear with sugarcane (Dufour et al. 1997; Ming et al. 1998) and C-value per monoploid genome is close to sugarcane (Butterfield et al. 2001). Publicly available sequences for sorghum *Adhs* included one completely sequenced *Adh1* gene (AF124045) and 34 ESTs retrieved from GenBank using *Adh2* of maize and *Adh1* of sorghum as query sequences. ESTs are derived from a single inbred line, BT × 623 (http://dogwood.botany.uga.edu/~prattlab/). It is thus a favorable case to resolve paralogous vs orthologous sequences, because no or very limited intragenic polymorphism is expected and any variant, supported by a fair-enough sequence quality, should introduce suspicion of an eventual supplementary member in the *Adh* family.

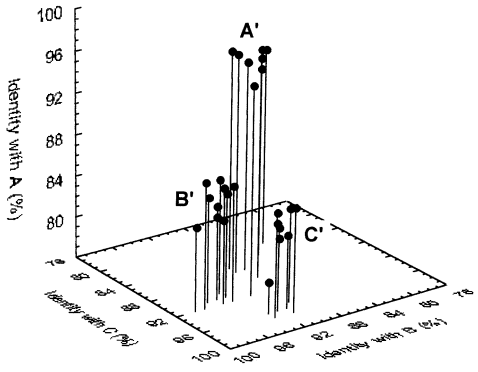

Fig. 2 Sequence identity between sorghum ESTs and three sugarcane putative *Adh* genes. Each point represents a sorghum EST. Identity with sugarcane putative genes A′, B′ and C′ (consensus sequence of ESTs) is reported on the three axes of the graph. It was established with BLASTN over the longest local alignment

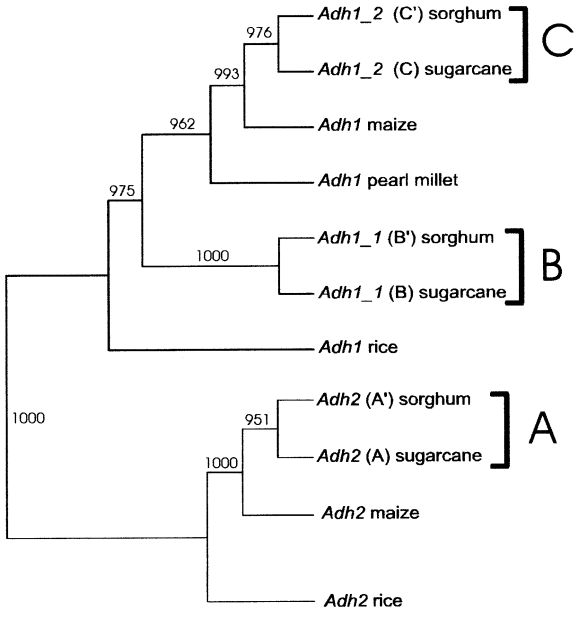

Fig. 3 The neighbor-joining tree performed on the coding sequence of the *Adh* genes of sugarcane, sorghum, maize, pearlmillet and rice. For sugarcane and sorghum, sequences are consensus-deduced from ESTs. For the three other species, Genbank accession numbers are given in the text. A first tree was constructed with the complete coding sequence of 1,140 bp, including all sequences except *Adh2* of sorghum. A second tree was then constructed with all sequences over the 600-bp region available for sorghum *Adh2*. Tree topologies were identical in both cases. The result of the first analysis is presented for the *Adh1* lineage, and the result of the second analysis is presented for the *Adh2* lineage. Bootstrap values are reported on the branches, the total number of trials performed being 1,000

Sorghum EST sequences were compared to sugarcane putative genes A, B and C consensus sequences with BLASTN, and identity was established over the longest local alignment. Three groups, A′, B′ and C′, containing 12, 9 and 13 ESTs, respectively, were observed (Fig. 2). In each group ESTs presented a higher identity with one of the three sugarcane genes, and respective genes were distinct for the three groups. This can simply be interpreted as three sorghum genes being orthologous to the three sugarcane genes. ESTs of each group were assembled with *phrap* and a consensus sequence was derived. No reliable polymorphism was detected inside alignments. As trace files were not available, this was established by searching for variants appearing at least twice. An ORF of 1,140 bp was detected for B′ and C′ and a shorter incomplete ORF of 600 bp was observed for A′. Sorghum *Adh1* sequence AF124045 was used as a control: it presented a single nucleotide difference with the consensus of C′, thus allowing us to be confident regarding the quality of sequences retrieved from sorghum ESTs.

Sugarcane and sorghum gene sequences derived from ESTs were used in conjunction with published sequences of maize, rice and pearlmillet *Adhs* to construct a Neighbor Joining tree (Fig. 3). This permitted us to show that: (1) the three sugarcane/sorghum pairs of putative orthologous genes were distinguished as three branches, each individualized with a high bootstrap value; (2) the position of branches C and A in the *Adh1* and *Adh2* lineages, respectively, were in good agreement with what is known of the grass phylogeny (Spangler et al. 1999; Kellogg 2001); and (3) branch B appeared as a duplication inside the *Adh1* lineage, that may have occurred before the radiation of the Andropogoneae. The global picture retrieved from the tree gives credit to the theory that three pairs of sugarcane/sorghum *Adh* genes are pairs of orthologs. EST groups A, B and C will thus further be referred to as genes *Adh2*, *Adh1_1* and *Adh1_2*, respectively.

Isozyme data for sorghum makes out a strong case for three *Adh* loci, one *Adh2* and two *Adh1* (Ellstrand et al. 1983), which is perfectly in line with the results established above.

Sequence variation inside sugarcane Adh genes

EST alignments for each of the three sugarcane genes were investigated for SNP and INDEL polymorphism. *Consed* alignments were occasionally modified manually near the ends, as large INDELs have sometimes not been well resolved by the program. Global results for three *Adh* genes are presented in Table 1. For *Adh2*, details of reliable variation are given in Fig. 4.

The first step of the procedure for SNP detection was performed manually by visual inspection of alignments obtained with *consed*, using rules defined in Materials and methods. This permitted the identification of a total of 37 bi-nucleotide candidate SNPs along with the three *Adh* genes, in both coding and non-coding regions. In the second step of the procedure, P_{SNP} was computed for each candidate SNP with Polybases. All candidates were retained at this step, as they all gave P_{SNP} higher than

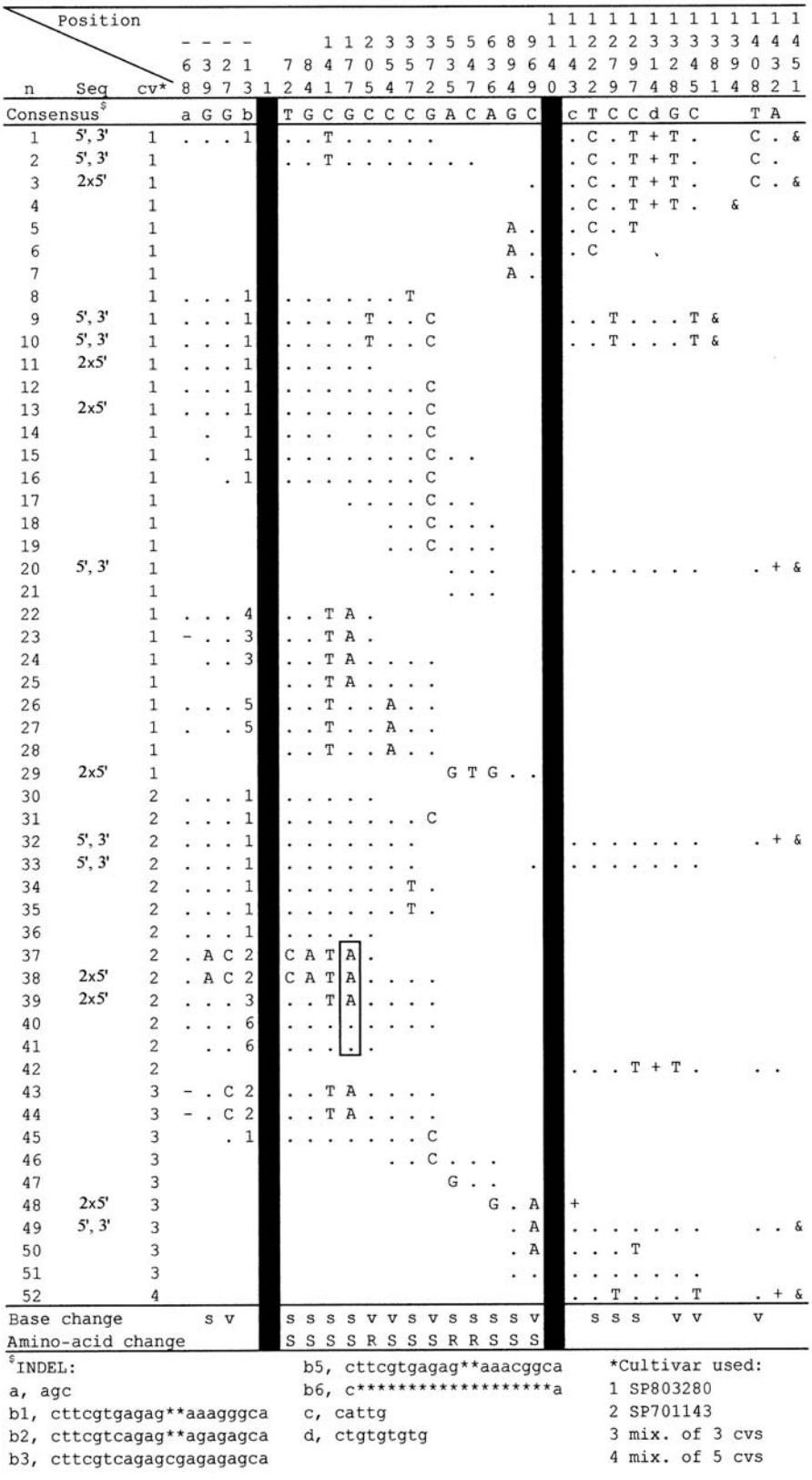

Fig. 4 Sequence polymorphism inside the sugarcane *Adh2* gene, as revealed with ESTs. Reliable sequence variation observed in the alignment of 67 ESTs from 52 independent cDNAs of *Adh2* gene is shown (see text). Each line represents a cDNA, identified in column one. Number, either single (blank) or double (2×), and direction, either 5′ or 3′, of sequencing is given in column two. Cultivar (*cv*) origin of each cDNA is given in column three. Each following column represents a variant site that passed the procedure described in the text. The position of each site is given relative to the initiation site of the translation (numbers should be read vertically). The *two black vertical lines* indicate the initiation and the end of the translation. The *consensus line* gives the majority variant at each site. A *point* indicates that the cDNA sequence is identical to the consensus and a blank indicates that data is missing. INDELs are shown by a '+' when present and by a '–' when absent. They are visualized by a small capital letter in the consensus sequence, which is detailed at the bottom of the figure. For a particularly complex region (shown as 'b' in the consensus sequence), haplotypes were numbered from 1 to 6 and are each explained at the bottom of the figure. When identified, the poly-(A) start was represented by a '&'. The type of nucleotide change, either transition (*s*) or transversion (*v*), and the type of amino-acid change in the coding region, either synonymous (*S*) or replacement (*R*), are given on the last two lines. Variants for cDNAs 37 to 41 at position 177 are *boxed*, indicating that corresponding trace files are shown in Fig. 5

Table 1 Summary of SNP and INDEL detected for three *Adh* sugarcane genes through EST data

Gene	Group	Nb of ESTs	Nb of c DNA	3′ leader			CDS			5′ end			Poly A[d]
				Length (bp)	SNP	INDEL	Length (bp)	SNP[c]	INDEL	Length (bp)	SNP	INDEL	
Adh1_1	B	123	100	97	0	2	1,140	10 (8)	0	210	1	1	2
Adh1_2	C	26	23	52	0	0	1,140	5 (5)	0	247	0	0	2
Adh2	A	67	54	95	2[a]	3[b]	1,140	13 (10)	0	310	6	3	3

[a] The three SNPs included in motive 'b' of Fig. 4 are not included here
[b] Two of these INDELs are included in motive 'b' of Fig. 4
[c] Number of silent changes is given in parentheses
[d] Number of Poly-adenylation sites observed

0.99. Trace file availability was an important aspect of quality assessment for detected variants in our procedure (Fig. 5).

The numbers of SNPs detected were 11, 5 and 21 for *Adh1_1*, *Adh1_2* and *Adh2*, respectively, indicating a high discrepancy between genes (Table 1). A first explanation may be that the number of ESTs available was very different for the three genes. It is clear that the exploration of the sequence diversity and the reliability of SNP detection increase with the volume of sequences available. Moreover, many SNPs may result from sequence divergence between *S. officinarum* and *S. spontaneum*, and the more or less ten-to-one allele origin ratio in cultivars may prevent the detection of this source of polymorphism when the number of ESTs is low (i.e. <20). Indeed, the lowest number of SNPs was detected for *Adh1_2*, which is also represented by the smallest number of ESTs. A second explanation for the discrepancy in the number of SNPs detected among genes may be the specific evolutionary history of each gene, as has been reported for barley *Adh*s (Lin et al. 2001). This may, for example, explain why the number of SNPs detected was higher in *Adh2* than in *Adh1_1*, despite the lower number of ESTs available.

Among the 37 SNPs detected, 24 (64%) corresponded to transition vs transversion. This value is in line with the proportion reported for similar studies in humans (Garg et al. 1999; Deutsch et al. 2001). Among the 28 SNPs detected inside coding regions, 23 (82%) corresponded to synonymous amino-acid substitutions. This proportion is about the same as that observed in humans. Deutsch et al. (2001), for example, reported 72 synonymous changes out of 82 SNPs (88%) detected in the coding regions of several genes.

INDELs were detected in 5′ leaders and 3′ terminal ends, which presented a cumulated length of around 1,000 bp (only segments covered by at least two sequences were considered). However, none was detected in the 3,420 bp of the cumulated coding sequence for the three genes. The size of the INDEL varied between 1 bp and 22 bp. The diversity of the 5′ region preceding the coding sequence over 50 to 100 bp was particularly well explored because the large majority of ESTs used in our study were sequences produced from the 5′ end. For *Adh1_1*, all combinations of the presence/absence of two INDELs of 4 bp and 22 bp (four haplotypes), separated by 15 bp, were observed in the sample of ESTs examined (data not shown). For *Adh2*, a complex pattern of variation, combining three SNPs and two INDELs, allowed us to distinguish up to six distinct haplotypes over 19 bp in the sample of ESTs examined (Fig. 4). Note that the three SNPs were not detected with the procedure mentioned above because of the proximity between each other and the proximity with the INDELs. However, all three gave a $P_{SNP} > 0.99$. No INDEL was detected for *Adh1_2*, which is the gene represented by the smallest number of ESTs. The high level of variation detected for the 5′ leader over short sequences in *Adh1_1* and *Adh2* may be fortuitous. It cannot be excluded, however, that it may be a common feature of sugarcane genes. It was also observed in a sugarcane 6-phosphogluconate dehydrogenase gene (unpublished results).

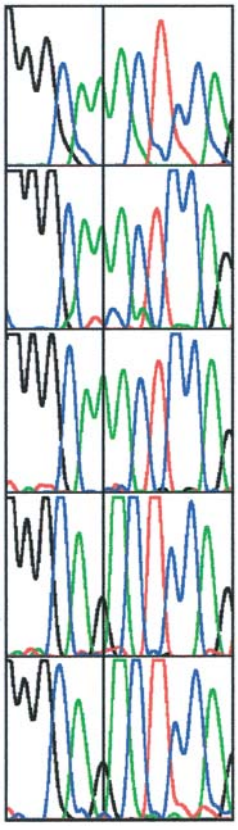

Fig. 5 Example of a reliable SNP as visualized from EST trace files. The '*green*', '*black*', '*blue*' and '*red*' lines correspond to the 'A', 'G', 'C' and 'T' residue detection, respectively. The five histograms are a zooming of EST trace files around position 177 (boxed on Fig. 4) of gene *Adh2*, where a G/A SNP was detected. Note that for cDNA 38 and 39 (Fig. 4) only one of the two available trace files is shown. A *vertical bar* visualizes position 177. The first three sequences show an 'A' and the last two, a 'G' variant. Base quality is > 30 for the five residues and P_{SNP} for this subsample is > 0.99

Discussion

Unlike human genetics, the use of EST data has been limited so far to access the sequence polymorphism in higher plants. The reason may be that, until recently, EST resources remain limited for most species, including the best studied, compared to the sequencing effort in humans. Moreover, the frequently inbred nature of plant material (for example, in Arabidopsis, rice, tomato, soybean, wheat and sorghum, etc.), and the generally limited diversity of genotypes used to construct cDNA libraries for sequencing, may have limited the interest of the approach. The situation is *a priori* more favorable in sugarcane as this species is highly polyploid and heterozygous, and as a large collection of ESTs is now available. The present work is intended to give credit to this hypothesis, by demonstrating the interest of sugarcane ESTs to reveal polymorphism in a few well-known genes.

Two types of risks should be considered in this exercise, first the risk of confusing genuine polymorphism with sequencing errors, and second the risk of confusing polymorphism at a unique locus with differences between recently diverged paralogous loci.

The first risk was controlled by a two-step procedure. The first step was a compromise to permit eliminating obvious artifacts and retaining most genuine variant sites. The second step consisted in computing a statistic, P_{SNP}, for sites that passed the first step, which allowed us to compare the level of significance for the presence of an SNP whatever the number of EST sequences available, the balance between variant frequencies and the base-call qualities. The variant sites that passed the first step were all conserved in the second step with a high level of probability.

The proportion of synonymous vs non-synonymous variation in coding regions permitted a supplementary validation of detected SNPs. In a random DNA sequence, 25% of the base changes are synonymous. A similar proportion would be expected in case the variation detected corresponded to artifacts. In contrast, we observed 82% synonymous changes. This value is significantly higher than 25% ($P \approx 1$), leaving little chance for the global set of SNPs to result from artifacts. This high proportion of synonymous variation illustrates the heavy selection operating against amino-acid-altering mutations. It is in line with observations performed on similar types of data in humans (Deutsch et al. 2001).

The second risk may result from a bad delineation of individual genes inside the family. The three *Adh* genes detected with sugarcane ESTs are supported by a similar gene partition in sorghum with clear orthological relationships between the two species. Sorghum is closely related to sugarcane with an estimated divergence time of about 9 million years (unpublished results). It is diploid and inbred, making interpretation of sequence variation much easier and more straightforward. Moreover, the three genes detected through nucleotide sequence differences are supported by isozyme data, which demonstrate the expression of three distinct *Adh* proteins in sorghum. No supplementary *Adh* lineage, or a potential indication of an eventual supplementary gene in the family, could be traced in maize or pearlmillet, two species of the Panicoideae. All those elements give a fair support to the partition of *Adh* ESTs into three genes for sugarcane.

The variation detected in sugarcane *Adh* genes was huge. The mean density of SNP in the coding sequences was one every 122 bp, which is much higher than that observed in human EST data, where for example Deutsch et al. (2001) reported one every 1,500 bp and Garg et al. (1999) one every 3,330 bp. This difference may be related to the specific evolutionary history of sugarcane, particularly, the double *S. officinarum–S. spontaneum* origin of current cultivars. This may also be due to the method of detection that we used here, which was intended to perform an exhaustive sorting of orthologous vs paralogous EST sequences. In this manner, all ESTs contributed to the detection of sequence variants in a specific gene, and none was withdrawn according to a particular threshold.

The data reported here are limited to three genes, but they are consistent and sound enough to suggest that the EST collection now available for sugarcane is very rich in sequence polymorphism information. It will be worthwhile in making an inventory of this variation, as it could become the raw material for future high-throughput genotyping technologies. Similar information is already being collected in human and model organisms such as mouse, and a new GenBank database, dbSNP, has been specially created to stock this information. Polymorphism revealed from ESTs is especially amenable for detection in genes that are highly expressed; that is, for which several tens of ESTs or more are available. A corollary is that polymorphism may not always be accessible for specific genes of interest that are poorly expressed and thus for which few ESTs are accessible from the database.

The specific application of sequence polymorphism to sugarcane breeding will be challenging. On one hand, linkage disequilibrium is likely to be very strong. But on the other hand, the high ploidy level may hamper a straightforward valorization of emerging technologies developed in Man and model diploid organisms. Based on a sample of Mauritian cultivars, it has been tentatively estimated that founding chromosome segments as long as 10 cM may still be segregating as unrecombined units in current material (Jannoo et al. 1999b). A tentative estimate of the global genetical to physical distance ratio (A. D'Hont, personal communication) suggests that 10 cM may correspond to a physical distance of at least several hundred kb and possibly up to several Megabases. Regarding polymorphism detected in the present study, this is probably enough for SNP and INDEL variants adjacent to a given QTA to combine into a specific haplotype signature. The challenge will be to read such signatures in a context where up to ten different ones coexist in a particular individual. In this perspective, the 5′ leader regions may offer interesting possibilities if the

particularly high diversity observed here appear as a more general feature of sugarcane genes.

Acknowledgements This work was supported by Sugarcane Genome Project from Fundação de Amparo à Pesquisa do Estado de São Paulo/FAPESP. The Sugarcane program of CIRAD funded the stay of L. G. as a visiting scientist at the Universidade Estadual de Campinas. P.A. received a research fellowship from CNPq. We thank Drs. Gabor T. Marth, Mark D. Yandell, Ian F. Korf and Warren R. Gish who gave us access to PolyBayes software. The present work has been carried out in compliance with the current laws governing genetic experimentation in Brazil.

References

Al-Janabi SM, Honeycutt RJ, McClelland M, Sobral BWS (1993) A genetic linkage map *of Saccharum spontaneum* L. 'SES 208'. Genetics 134:1249–1260

Altshul SF, Gish W, Miller W, Myers EW, Lipman J (1990) Basic local alignment tool. J Mol Biol 215:1651–1656

Buetow KH, Edmonson MN, Cassidy AB (1999) Reliable identification of large numbers of candidate SNPs from public EST data. Nature Genet 21:323–325

Butterfield MK, D'Hont A, Berding N (2001) The sugarcane genome: a synthesis of current understanding, and lessons for breeding and biotechnology. Proc S Afr Sugar Technol Assoc 75:1–5

Da Silva JAG, Burnquist WL, Tanksley SD (1993) RFLP linkage map and genome analysis of *Saccharum spontaneum*. Genome 36:782–791

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

Deutsch S, Iseli C, Bucher P, Antonarakis SE, Scott HS (2001) A cSNP map and database for human chromosome 21. Genome Res 11:300–307

Dufour P, Deu M, Grivet L, D'Hont A, Paulet F, Bouet A, Lanaud C, Glaszmann JC, Hamon P (1997) Construction of a composite sorghum genome map and comparison with sugarcane a related complex polyploid. Theor Appl Genet 94:409–418

Ellstrand NC, Lee JM, Foster KW (1983) Alcohol dehydrogenase isozymes in grain sorghum (*Sorghum bicolor*): evidence for a gene duplication. Biochem Genet 21:147–154

Ewing B, Green P (1998) Base-calling of automated sequencer traces using Phrep. II. Error probabilities. Genome Res 8:186–194

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

Garg K, Green P, Nickerson DA (1999) Identification of candidate coding region single nucleotide polymorphisms in 165 human genes using assembled expressed sequence tags. Genome Res 9:1087–1092

Gaut BS, Peek AS, Morton BR, Clegg MT (1999) Patterns of genetic diversification within the *Adh* gene family in the grasses (Poaceae). Mol Biol Evol 16:1086–1097

Gordon D, Abajian C, Green P (1998) *consed*: a graphical tool for sequence finishing. Genome Res 8:197–202

Green P (1994) *phrap*: http://www.genome.washington.edu/

Grivet L, D'Hont A, Roques D, Feldmann P, Lanaud C, Glaszmann JC (1996) RFLP mapping in a highly polyploid and aneuploid interspecific hybrid. Genetics 142:987–1000

Hoarau JY, Offmann B, D'Hont A, Risterucci AM, Roques D, Glaszmann JC, Grivet L (2001) Genetic dissection of a modern sugarcane cultivar (*Saccharum* spp).I. Genome mapping with AFLP markers. Theor Appl Genet 103:84–97

Jannoo N, Grivet L, Seguin M, Paulet F, Domaingue R, Rao PS, Dookun A, D'Hont A, Glaszmann JC (1999a) Molecular investigation of the genetic base of sugarcane cultivars. Theor Appl Genet 99:171–184

Jannoo N, Grivet L, Dookun A, D'Hont A, Glaszmann JC (1999b) Linkage disequilibrium among sugarcane cultivars. Theor Appl Genet 99:1053–1060

Kellogg E (2001) Evolutionary history of the grasses Plant Physiol 125:1198–1205

Lima de L AM, Garcia AAF, Oliveira KM, Matsuoka S, de Souza Jr CL, Souza de AP (2002) Comparative analysis of genetic similarity detected by AFLP and coefficient of parentage among cultivars of sugarcane. Theor Appl Genet 104:30–38

Lin JZ, Brown AHD, Clegg MT (2001) Heterogeneous geographic patterns of nucleotide sequence diversity between two alcohol dehydrogenase genes in wild barley (*Hordeum vulgare* subspecies *spontaneum*). Proc Natl Acad Sci USA 98:531–536

Lu YH, D'Hont A, Paulet F, Grivet L, Arnaud M, Glaszmann JC (1994) Molecular diversity and genome structure in modern sugarcane varieties. Euphytica 78:217–226

Marth GT, Korf I, Yandell MD, Yeh RT, Gu Z, Zakeri H, Stiziel NO, Hillier L, Kwok PY, Gish WR (1999) A general approach to single-nucleotide polymorphism discovery. Nature Genet 23:452–456

Ming R, Liu SC, Lin YR, Silva da J, Wilson W, Braga D, van Deynze A, Wenslaff TF, Wu KK, Moore PH, Burnquist W, Sorrells ME, Irvine JE, Paterson AH (1998) Detailed alignment of Saccharum and Sorghum chromosomes: comparative organization of closely related diploid and polyploid genomes. Genetics 150:1663–1682

Picoult-Newberg L, Ideker TE, Pohl MG, Taylor SL, Donaldson MA, Nickerson DA, Boyce-Jacino M (1999) Mining SNPs from EST databases. Genome Res 9:167–174

Spangler R, Zaitchik B, Russo E, Kellogg E (1999) Andropogoneae evolution and generic limits in Sorghum (Poaceae) using ndhF sequences. Syst Bot 24:267–281

Statistica (1997) V 5.1, Statsoft Inc, Tulsa, Oklahoma, USA

Syvänen A-C (2001) Accessing genetic variation: genotyping single nucleotide polymorphisms. Nature Rev 12:930–942

Trick M, Dennis ES, Edwards KJR Peacock WJ (1988) Molecular analysis of the alcohol dehydrogenase gene family of barley. Plant Mol Biol 11:147–160

Whitkus R, Doebley J, Lee M (1992) Comparative genome mapping of sorghum and maize. Genetics 132:1119–1130