Mol Breeding (2010) 25:501–516
DOI 10.1007/s11032-009-9348-x

Detection of somaclonal variation during cocoa somatic embryogenesis characterised using cleaved amplified polymorphic sequence and the new freeware Artbio

Carlos M. Rodríguez López · Hector Sicilia Bravo · Andrew C. Wetten · Michael J. Wilkinson

Received: 13 March 2009 / Accepted: 24 October 2009 / Published online: 12 November 2009

Abstract The scarcity and stochastic nature of genetic mutations presents a significant challenge for scientists seeking to characterise de novo mutation frequency at specific loci. Such mutations can be particularly numerous during regeneration of plants from in vitro culture and can undermine the value of germplasm conservation efforts. We used cleaved amplified polymorphic sequence (CAPS) analysis to characterise new mutations amongst a clonal population of cocoa plants regenerated via a somatic embryogenesis protocol used previously for cocoa cryopreservation. Efficacy of the CAPS system for mutation detection was greatly improved after an 'a priori' in silico screen of reference target sequences for actual and potential restriction enzyme recognition sites using a new freely available software called Artbio. Artbio surveys known sequences for existing restriction enzyme recognition sites but also identifies all single nucleotide polymorphism (SNP) deviations from such motifs. Using this software, we performed an in silico screen of seven loci for restriction sites and their potential mutant SNP variants that were possible from 21 restriction enzymes. The four most informative locus-enzyme combinations were then used to survey the regenerant populations for de novo mutants. We characterised the pattern of point mutations and, using the outputs of Artbio, calculated the ratio of base substitution in 114 somatic embryo-derived cocoa regenerants originating from two explant genotypes. We found 49 polymorphisms, comprising 26.3% of the samples screened, with an inferred rate of 2.8×10^{-3} substitutions/screened base. This elevated rate is of a similar order of magnitude to previous reports of de novo microsatellite length mutations arising in the crop and suggests caution should be exercised when applying somatic embryogenesis for the conservation of plant germplasm.

Keywords Artbio · Somaclonal variation · SNPs · CAPS · PCR–RFLP · *Theobroma cacao*

Electronic supplementary material The online version of this article (doi:10.1007/s11032-009-9348-x) contains supplementary material, which is available to authorized users.

A. C. Wetten
School of Biological Sciences, Reading University, Whiteknights, Reading RG6 6AS, UK

C. M. Rodríguez López (✉) · M. J. Wilkinson
Institute of Biological Environmental and Rural Sciences, Aberystwyth University, Penglais, Ceredigion SY23 3DA, UK
e-mail: ccr@aber.ac.uk

H. S. Bravo
Fundación de Estudios Portuarios, 38006 S/C de Tenerife, Spain

Introduction

Cocoa (*Theobroma cacao*) is a tropical tree crop that is susceptible to many pests and diseases (Fang et al. 2004). The crop is typically cultivated in small plantations comprising of few genotypes that are generally established from clonal cuttings. Difficulty in generating commercial quantities of cocoa ramets via cuttings has led to an increasing interest in the application of somatic embryogenesis for clonal multiplication (Traore et al. 2003; Li et al. 1998; Lopez-Baez et al. 1993). These methods have great economic and practical potential for the industry but equally have the capacity to cause the appearance of unwanted spontaneous mutants. Rodríguez López et al. (2004) reported a remarkable 31% of 233 cocoa somatic embryogenesis regenerants were at least chimeric for novel microsatellite length mutations. This figure implies that a low level of genetic fidelity would be maintained by commercial systems based on somatic embryogenesis. However, microsatellites are extremely prone to length slippage (e.g. Whittaker et al. 2003) and so these mutation rates do not necessarily relate to substitution-based mutations. Thus, there is a need to characterise the presence, nature and abundance of single nucleotide polymorphisms (SNPs) among cocoa regenerants via somatic embryogenesis.

Point mutations lead to the formation of SNPs between cells, cell lines or individuals. SNPs are becoming increasingly popular as a source of genetic markers for genetic linkage mapping (e.g. Yu and Buckler 2006), population genetics (González-Martínez et al. 2007), biogeography (Vettori et al. 2004), phylogenetics (Bala et al. 2003; Devos et al. 2005) and studies of allelic diversity (Zakeri et al. 2006). The study of rare 'de novo' mutations amongst clonal populations nevertheless provides the greatest challenge for researchers aiming to characterise genetic variability (Oh et al. 2007; Rodríguez López et al. 2004). There is little doubt that plant tissue culture is a highly mutagenic environment, with mutation rates typically greatly exceeding those observed in seed-derived plants. The enhanced genetic infidelity amongst regenerants derived from in vitro culture is known as 'somaclonal variation' and includes an increased incidence of all forms of mutation (Phillips et al. 1994). Kaeppler et al. (1998) noted that most tissue culture-induced mutations are attributable to point mutations rather than to gross chromosomal abnormalities. In most instances, cost implications mean that extensive sequencing is still an impractical option to characterise de novo point mutations arising from culture. For this reason point mutations are generally detected indirectly, usually in the form of altered mobility of a PCR amplicon or protein when subjected to electrophoresis (e.g. Jin et al. 2008; Rodríguez López et al. 2004; Albani and Wilkinson 1998). Thus, the exact nature of the DNA base sequence alteration is frequently unknown.

Endogenously generated DNA damage is often termed 'spontaneous' to distinguish it from that caused by exogenous sources. Single nucleotide changes to the code have many possible causes, including polymerase infidelity during DNA replication, misinsertion of a mutagenic nucleotide, ineffective mismatch repair systems, DNA damage by light or oxidizing molecules, or deamination of methylated cytosine residues (Maki 2002). These different causes are likely to produce distinct mutation profiles. For instance, Graur and Li (2000) argue that replication or repair problems are likely to result in any kind of base substitution, whereas Britt (1999) noted that light damage (e.g. UV light) will be most reactive on thymine dimers, and deamination of methylated cytosines results only in C to T transitions, and Wang et al. (1998) observed that the two dominant base substitutions induced by oxidative stress are G:C to A:T transitions and G:C to T:A transversions. Whilst several hypotheses have been proposed to try to explain the mechanisms giving rise to point mutations, there remains a paucity of sequence data relating to tissue culture-induced sequence changes, and the interrelations of the possible mechanisms remains obscure. Molecular evolutionists have long been interested in describing the pattern of spontaneous mutations, since point mutations are among the most important factors driving the evolution of genomes (Graur and Li 2000). The study of the dynamics of nucleotide substitution has to take into account the probability of substitution of one nucleotide by another. Numerous mathematical schemes

have been proposed to address this problem. The two simplest and most frequently used are: Jukes and Cantor's one parameter model, and Kimura's two-parameter model (Graur and Li 2000). Jukes and Cantor's model makes the unrealistic assumption that substitutions occur with equal probability among the four nucleotide types, i.e. no bias in the direction of change. Many studies have shown that these processes are not random, with transitions, and in particular $C \rightarrow T$ and $G \rightarrow A$, occurring more frequently than transversions (Wang et al. 1998). Indeed, the observed proportion of transitions (67%) is approximately twice the expected value (33%) if the mutational events were random (Kimura 1980). There is also a tendency for C and G to be more mutable (59.2% of all mutations) and for these bases to become A or T (56.4% of all mutations). Kimura (1980) accommodated for these tendencies when he proposed a two-parameter model in which the rate of transitions is different from the rate of transversional substitutions.

The ease with which a novel genotype can be identified amongst regenerants depends on the nature of the mutation and the methodology used for genotype assignation itself. The advent of parallel sequencing capacity offers the greatest power for the extensive characterisation of low frequency mutation events and is becoming increasingly important in the detection and characterisation of mutations in genes implicated in the onset of cancer (e.g. Thomas et al. 2006, 2007; Dahl et al. 2007). In these instances there is a clear functional rationale behind selecting the genomic targets (i.e. cancer-associated gene regions) and a mechanistic reasoning for the expectation of raised mutation frequencies (cancer cell lines). This reasoning does not necessarily extend to studies aiming to characterise somaclonal variation of a species with an (as yet) unsequenced genome. Cost and resource considerations also come into play for micropropagation or cryopreservation initiatives where the priority is the regeneration of large quantities of clonal plant material. There may be a need to sequence large numbers of samples (to ensure mutation detection) or else to sequence a smaller number of genotypes but compare much larger lengths of DNA. Furthermore, somaclonal mutations are normally heterozygous, and so the sequencing effort must be sensitive enough to detect the heterozygous condition (Cotton 1997) quite possibly in chimeric tissues (Rodríguez López et al. 2004).

Methods have been developed to detect mismatched bases by electrophoresis of PCR amplicons that include the single base mutation (Sawa et al. 1997), and by sequencing using chip hybridisation (Cotton 1997). The last strategy has the advantage of allowing large regions of the genome to be compared across several genotypes but requires considerable establishment costs and complex data analysis (Paton et al. 2000). Cleaved amplified polymorphic sequence (CAPS), otherwise known as polymerase chain reaction restriction fragment length polymorphism (PCR–RFLP), generates a simplified restriction pattern relating to a single locus. When the amplicon sequence is known, restriction enzymes targeting interstitial regions of the amplicon can be deployed to identify somaclonal variants where a point mutation has destroyed the recognition site. Empirical screening of large numbers of regenerants using a randomly selected sequence may equally identify regenerants where mutation has given rise to a new restriction site and so uncovered new haplotypes. The problem lies in not knowing which loci are most likely to give rise to novel haplotypes and so most suited for such large-scale blind screens. It is also useful to know where new restriction sites are most likely to occur within the amplicon because some sites may be difficult to detect after electrophoresis (specifically, those occurring close to the termini) and because such information may allow the experimenter to infer the nature of the mutation without resorting to a posteriori sequencing. Thus, there is a need to be able to screen target sequences both for existing restriction sites and also for sites that could become restriction sites after a single mutation. Here, we describe and apply a new algorithm-based software tool (Artbio) for the computational prediction of SNP variants that can be recognised by DNA restriction digest. Whilst we expect the greatest usage of the software will derive from its use to select loci with the highest capacity to reveal CAPS-based polymorphisms in wild or cultivated germplasm, we elected to evaluate its utility for the more demanding search of contemporary point mutations among clonal regenerants from somatic embryo cultures of two cocoa genotypes.

Materials and methods

Artbio software development and programming

Task definition and design of algorithm

The program goal was defined as 'to determine the frequency and location of DNA code strings (considered here as 'words') representing the various restriction recognition sites and those of one-letter modifications of these 'words' within a given template sequence'. For example, when screening the string **GGATCGCTCTATC**, the algorithm should indicate that **GATC** appears in position 2 and subsequences **GCTC** and **TATC** occupy positions 6 and 10.

The reasoning used to attain the program goal recognised that basic problems encountered in classical text analysis have established solutions and that these could be applied to address equivalent problems in biological texts. Moreover, we chose to adapt an algorithm designed for text analysis to the specificities of biological sequences. To do this, we defined the following terms:

DNA String (**DNAS**): DNA sequence where the search is to be conducted.

Search String (**SS**): Searched DNA subsequence.

Search Differences: Number of differences allowed between the **Search String** and the **DNA String** subsequences.

IsEqual: If it is *True* indicates that the number of differences allowed are exactly those indicated by **Search Differences**. Conversely, if it is *False* the number of differences allowed can be the same or lower than those indicated by **SearchDifferences**.

The variables used by the algorithm were set as:

SizeDNAS, **SizeSS**: Number of bases that form the **DNAS** and the **SS**, respectively.

DNAS[**X**], **SS**[**X**]: Text character on position **X** of **DNAS** and the **SS**, respectively. The first base of any sequence is character **DNAS**[1] and **SS**[1] while the last base of any sequence is **DNAS**[**SizeDNAS**] and **SS**[**SizeSS**].

Sequence Index or **SI**: Position of the initial base of the **DNAS** subsequence where the search is going to be performed.

Differences: Detected number of differences between the **DNAS** subsequence and **SS** in every algorithm cycle.

Given these premises the design algorithm pseudocode was:

```
SizeDNAS = Length(DNAS);

SizeSS := Length(SS);

for SI = 1 to SizeDNA – SizeSS + 1 do
 {
 Differences = 0;
 for Count = 1 to SizeSS do
  {
  if SS[Count] <> DNAS[SI + Count - 1] then
   {
   Differences = Differences + 1;
     }
    }
   if ((   IsEqual) and (Differences  = SearchDifferences)) or
    ((not IsEqual) and (Differences <= SearchDifferences)) then
    {
    AddResult(SI);
    }
  }
```

The algorithm contains two nested cycles. In the main loop variable **SI** increases until string **DNAS** is completely searched for positions that fulfil the given conditions. The initial position of the last subchain that can be compared to **SS** is **SizeDNAS** – **SizeSS** + **1**, since this is the last subchain that presents **SizeSS** characters. In the sample given before **SizeDNAS** is 11 and **SizeSS** is 3, therefore the last comparable bases are on positions 9, 10 and 11.

The internal loop compares **SS** characters to those forming the subchain **DNSA** and every time it finds differences between both increases the value of

variable **Differences**. This variable starts as zero at the beginning of every run of the main loop. At the end of the loop, variable **Differences** is examined to determine if given conditions are fulfilled. If these are fulfilled, position **SI** is added to the results.

Implementation phase

The new algorithm has been implemented in a program called Artbio. The program reads sequence databases and searches for SNP variants of restriction enzyme recognition sequences (SNP-RSs). Artbio was written using Delphi 6.0 under a multi document interface (MDI) working environment. It allows editing and storage of DNA sequences in different formats (such as TXT, FASTA) and enables multiple searches to be performed simultaneously.

Sequences can be searched according to various factors including the location of existing or potential restriction recognition sequences (after one or more mutations), and for the number and location of SNP mutations from a reference sequence. The results can be visualized in the program window or saved on a text file. Artbio can be run on the following supports: Windows 98, Windows 2000 Server, Windows 2000 Pro and Windows XP and is easily transformed to be used under a Linux environment.

It is available free of charge at http://www.artbio.net.

Use of Artbio to characterise somaclonal variation

We sought to test the utility of Artbio to select loci for use in CAPS-based screens of culture-derived cocoa regenerants for somaclonal mutations.

Plant material

Explant material was taken from cocoa trees held in Reading University's Intermediate Quarantine Facility. Closed immature flower buds (4–5 mm) were collected from genotypes LCT EEN 162/S-1010, and SIAL 93 between 08.00 and 12.00 a.m. Staminodes isolated from these buds were used to induce and regenerate somatic embryos according to Li et al. (1998). A random selection of individuals from regenerant populations of genotypes LCT EEN 162/S-1010 (96 samples) and SIAL 93 (18 samples) were assessed for somaclonal variation.

DNA isolation

DNA was extracted from in vitro plantlets, and from recently expanded leaves of trees used as the source for explant material using the DNeasy 96 Plant Kit (Qiagen, Valencia, CA) and the Mixer Mill MM 300 (Retsch, Germany). Isolated DNA was diluted in nanopure water to produce working stocks of 5 ng μl^{-1}. There were 20 DNA replicate extractions performed from each ortet tree and two replicates for each regenerant plantlet.

Amplification of target sequences with PCR

One SSR primer pair (mTcCIR15) and two sequence characterized amplified region (SCAR) markers primer pairs (BA and B4) were tested for the production of markers in 20 DNA extractions from cocoa genotypes LCT EEN 162/S-1010 and SIAL 93. The primer sequences, map positions and conditions for amplification are described by Lanaud et al. (1999), Risterucci et al. (2000), and Charters (2000). PCR amplifications were performed in 20 µl reaction mixtures in a Phoenix thermal cycler (Helena BioScience, UK).

Using primers BE and mTcCIR 15, PCR products were generated from replicated DNA extractions of randomly selected regenerated individuals from genotypes LCT EEN 162/S-1010 (96 samples) and SIAL 93 (18 samples). The amplicons were subsequently digested using endonucleases *Mse*I or α-*Taq*I.

Cloning and cycle sequencing reactions

Cloning of selected fragments

PCR products obtained from wild-type ortet plants of genotypes LCT EEN 162/S-1010 and SIAL 93 were cloned using a TOPO TA Cloning® kit (Invitrogen™ life technologies). Cloning was used to further amplify individual PCR products and to ensure that only single alleles were used for subsequent DNA sequencing. In this way we avoided problems associated with mutation chimeras (Rodríguez López et al. 2004).

Cycle sequencing of cloned fragments

Cloned products were subjected to cycle sequencing using the ABI PRISM™ Big Dye™ Terminator Cycle Sequencing Ready Reaction Kit according to

the manufacturer's instructions. Twenty single stranded sequencing reactions were performed per sample (ten for each, M13 forward and reverse primers) and each included 3 μl of PCR product, 0.16 μl primer, 4 μl terminator ready reaction mix and 2.8 μl sterile nanopure water. The thermal profile for all reactions was: 35 cycles of 30 s at 96°C, 15 s at 50°C, and 4 min at 60°C.

Products from the sequencing reactions were cleaned using Edge Biosystems AGTC Gel Filtration Cartridges. Samples were then fractionated on an ABI 373XL automated sequencer according to the manufacturer's protocols (Applied Biosystems, Inc.).

Direct cycle sequencing of PCR amplified fragments

The only deviation from the above protocol involved the sequencing of PCR amplicons generated using specific primer pairs BA and BE. Here, primer volume was unchanged when sequencing cloned fragments although PCR products were further diluted to 1.5 μM for direct sequencing reactions.

Sequence analysis and alignment

The sequences obtained from each primer (forward and reverse) were manually edited and a consensus sequence was obtained using the Bioedit Sequence Alignment Editor Software version 5.0.3 (North Carolina State University, USA). Subsequently, replicate sequences from the same genotype were aligned and compared in order to ratify the consensus sequences.

Statistical screen of template DNA sequences for palindromic and quasi-palindromic sites

DNA sequences obtained above for the SCAR markers BA and BE, and for SSR marker mTcCIR 15 (two alleles), were added to a further three sequences taken from publicly available databases (*T. cacao* csv gene for seed vicilin NCBI accession X62625; *T. cacao* class I chitinase gene NCBI accession U30324; *T. cacao* Resistant protein-like protein gene NCBI accession AF402719). All sequences were screened for the presence of 26 (Supplementary Table 1) restriction recognition sequences and single base deviations from these sequences (known hereafter as quasi-palindromic sequences, with the 'deviant' nucleotide position termed as Single Nucleotide Polymorphic Recognition Sequences or SNP-RS's) using Artbio software (Table 1). Collectively, these sequences comprised 6,686 base pairs. Counts were performed for: (1) the relative frequency of every possible oligonucleotide, (2) the frequency of every possible base substitution, and (3) the position of the SNP-RSs within the quasi-palindromic sequences detected. Such positions were divided into SNP-RSs present on the 1st and 4th bases and 2nd and 3rd bases for 4-bp sequences and SNP-RSs present on the 1st and 6th bases, 2nd and 5th bases, and 3rd and 4th bases for the 6-bp sequences.

Karling et al. (1992) produced a model that predicted counts of each screened 'word' based on described mononucleotide frequencies and under the assumption that nucleotides are both independently and randomly distributed. For a string of N letters sampled independently, the expected probability of observing a certain word w (e.g. AGCT), at any specified location is the product of the frequencies of the component letters:

$$f_w = (f_A X f_G X f_C X f_T)$$

This prediction was tested by comparison with observed frequencies of a given word. To achieve this, all target sequences used in the screen above were amalgamated into a single sequence comprising 6,686 bp. A search for each of the four DNA bases was performed using the 'Find and replace' function and so the number of possible substitutions obtained

Table 1 Comparison of expected restriction sites and SNP-RSs and restriction sites and SNP-RSs detected in-silico using Artbio

	Restriction sites		SNP-RSs		χ^2
	Expected	Counts	Expected	Counts	
Four bases words	195.5	167	2,376	2,041	3.8×10^{-20}
Six bases words	35.5	34	608	561	0.025

Counts of palindromes and quasi-palindromic sequences using Artbio software in seven *T. cacao* sequences. Expected counts for any specific word was obtained by multiplying the expected frequency of observing the word (given by the frequency of the letters forming such word) by the total number of bases screened (6,689)

was deemed to be the number of each base present in the sequence.

Frequent and rare 'words' were statistically defined comparing the expected number of counts based on the base composition of the studied sequences, and the number of counts obtained with Artbio using a Chi-square test.

Enzymatic restriction

We then performed an empirical test of the utility of Artbio for predicting and characterising mutant CAPS alleles amongst a regenerant plant population. PCR products were amplified using primer pairs BE and mTcCIR 15 from parental and regenerant genotypes LCT EEN 162/S-1010 and SIAL 93, and restricted using enzymes *Mse*I or α-*Taq*I (Promega) in 40 µl reaction mixtures containing 20.8 µl of nanopure water, 0.2 µl of restriction enzyme, 15 µl of PCR product and 4 µl of 10× buffer React1 (Promega). The restriction reaction was performed in a Phoenix Thermal cycler for 30 min at 37°C (this was sufficient to ensure partial rather than complete digestion so that labelled unrestricted template and a labelled restriction product were both visible after capillary electrophoresis). The restriction enzyme was subsequently denatured by incubation for 4 min at 94°C.

Fluorescently labelled PCR amplicons and their restriction products were diluted 1/10 using nanopure sterile water and 1 µl was combined with 1 µl of ROX/HiDi mix (50 µl ROX plus 1 ml of HiDi formamide). Samples were denatured at 95°C for 3–5 min and snap-cooled on ice for 2 min and then fractionated on an ABI PRISM 3100 (PE Applied Biosystems, Foster City, CA) at 3 kV for 22 s and at 15 kV for 45 min at 60°C.

Data analysis

Scoring of polymorphic bands

Following restriction, polymorphic bands were identified through the use of Genotyper 3.7NT software (Applera Corporation) for the comparison of unrestricted profiles from ortets and regenerants with their respective restricted profiles. Since restriction time was too short to allow complete digestion of the PCR amplicons, each restriction profile should carry the uncut labelled fragment plus restricted labelled fragments (both true-to-type or polymorphic). Polymorphic fragment sizes were then compared to the database of expected restriction sites and possible SNP-RS's obtained using Artbio.

Two kinds of polymorphisms were detected (1) the loss of an expected restriction fragment indicating possible base substitutions or small in/dels and (2) the appearance of a polymorphic restriction fragment, indicating the creation of a new restriction site for the enzyme used. The type of base substitution was obtained using the restriction site/SNP-RS maps for each sequence-restriction enzyme combination using Artbio.

Determination of the ratio of base substitution

Mutation frequency was calculated by dividing the number of detected novel polymorphisms by the total number of bases screened. The number of screened bases in instances where restriction sites have been lost is the number of screened samples (114) multiplied by the size of the recognition sequence (four bases for both enzymes used) and by the number of restriction sites detected. In the case where mutations produce the gain of a restriction site two-thirds of the possible mutations at the target site are undetectable by restriction analysis and so this criterion was accommodated in calculating inferred mutation frequency. Thus, the number of bases analysed is the number of screened samples multiplied by the number of SNP-RSs detected and multiplied by three (assuming all bases are equally mutable).

The proportion of base changes (P_{ij}) from ith type to the jth type of nucleotide (i, j = A, T, C or G, and $i \neq j$) is calculated as

$$P_{ij} = n_{ij}/n_i$$

where n_{ij} is the number of substitutions from i to j, and n_i is the number of the i nucleotides in the ancestral sequence. To be able to compare the patterns of nucleotide substitution when the four bases are not equally frequent it is necessary to define the relative substitution frequency from nucleotide i to j (f_{ij}):

$$f_{ij} = \left(P_{ij} \Big/ \sum_i \sum_{j \neq i} P_{ij} \right) \times 100$$

Results

Artbio

A user's manual was created for the Artbio software and is available as supplementary online material and at http://www.artbio.net. Source code and precompiled executables are also freely available upon request.

SCAR marker amplification and sequencing

Clean sequences were obtained for all seven markers (SCAR markers BE, B4, and SSR marker mTcCIR 15, etc.) using template DNA from leaves of both explant genotypes (LCT EEN 162/S-1010 and SIAL 93). When these sequences were analyzed with Artbio, the output revealed that the seven cocoa DNA sequences collectively contained 200 known restriction sites (166 four base recognition sites and 34 six base sites), and 2,602 SNP-RSs (2,012 single base deviations from four base restriction sites and 557 from six base restriction sites; e.g. Fig. 1). Curiously, both restriction sites and SNP-RSs were slightly but significantly less abundant than predicted purely on the basis of chance (Table 1). Furthermore, not all types of restriction enzyme recognition sequences and associated SNP-RSs were detected within the template sequences; out of the possible 26 restriction sites and 420 SNP-RSs, four restriction sites and 90 SNP-RSs (all of them 6 bp sequences) were not detected in any of the DNA template sequences analysed (Supplementary Table 1).

Relative frequency of transitions and transversions to create new restriction sites

Provided that all search 'words' of the same length are equally abundant, it would be expected that the ratio of restriction sites/SNP-RSs would be 1/12 for four-bases restriction enzymes and 1/18 for six-bases enzymes. Our data for both types of restriction sites and their single-base deviants fitted with this hypothesis, with the detected ratios failing to differ significantly from the expected ratios (Table 1).

For any given base there is one type of possible transition and two types of transversions. In the absence of selection or of a mechanistic bias favouring one type of mutation over another, the chances of finding a SNP-RS that would produce a restriction site after a transversional mutation should be approximately twice the chance of finding one that would produce it after a transition. Comparison of both possible types of changes amongst the restriction products screened concurred with this premise (874 possible transitions and 1,728 possible transversions; Supplementary Table 1).

Nucleotides A and T formed a significantly higher relative number of observed SNP-RSs (59%) (Table 2) than did G and C while 54% of all predicted SNP-RSs would arise after a base substitution to C or G (Table 2).

Positional effects of SNP-RSs

The relative abundance of alternative SNP-RSs was then compared for each position in each restriction enzyme recognition site. All possibilities were divided into SNP-RSs where the mutation occurred in the 1st and 4th bases, or 2nd and 3rd bases for 4-bp sequences and SNP-RSs were present on the 1st and 6th bases, 2nd and 5th bases, and 3rd and 4th bases for the 6-bp sequences. Curiously there was an excess of SNP-RSs observed in central base positions (1,160 for 2nd and 3rd bases for 4-bp sequences and 225 for 3rd and 4th bases for the 6-bp sequences) when compared with terminal bases (852 for 1st and 4th bases for 4-bp sequences and 186 for 1st and 6th bases and 146 for 2nd and 5th bases for the 6-bp sequences; Chi squared $= 8.13 \times 10^{-14}$, $P = 0.05$).

CAPS analysis

Restriction enzymes *Mse*I and α-*Taq*I yielded the most restriction recognition sites and also the greatest predicted frequency of SNP-RS among the loci screened. Accordingly, these loci were selected for the screen of genetic mutations among somatic embryo-derived regenerants. PCR products amplified using primer pairs BE and mTcCIR15 from both explant sources and from 114 regenerant samples included two kinds of polymorphisms with *Mse*I and α-*Taq*I: (1) the loss of an expected restriction fragment, and (2) the appearance of a novel restriction fragment (Fig. 2).

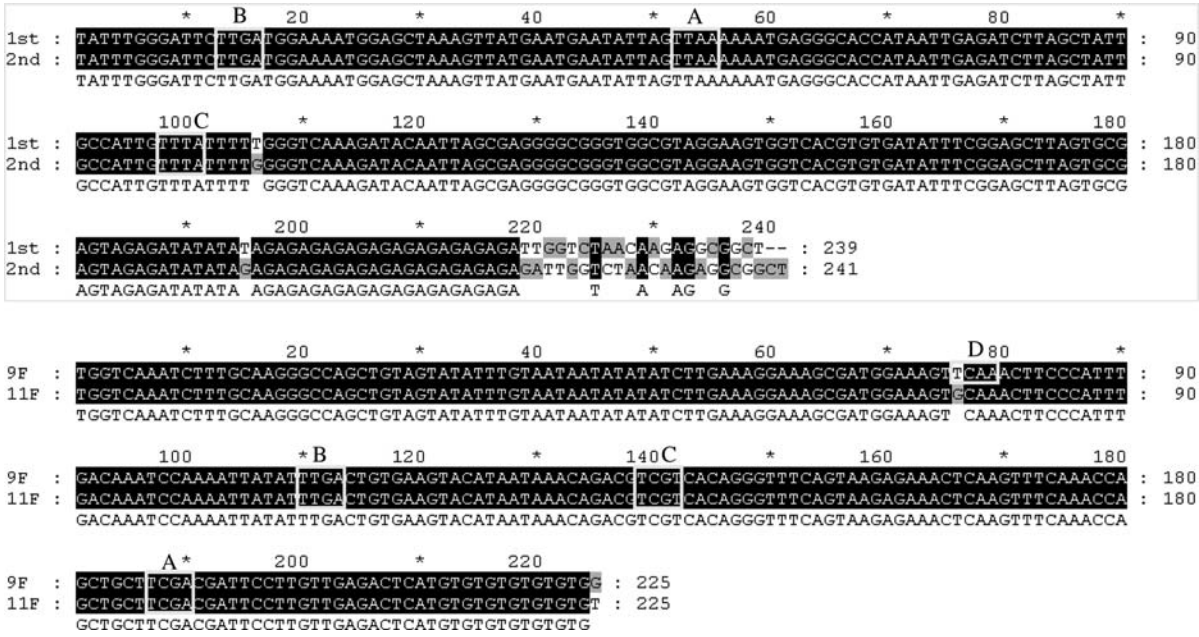

Fig. 1 Restriction mapping of SSR markers (mTcCIR 15) and BE SCAR marker: 1st and 2nd sequences were obtained from the two mTcCIR 15 alleles present in genotype SIAL 93. Sequences 9F and 11F were obtained from the SCAR marker BE from genotypes LCT EEN 162/S-1010 and SIAL 93, respectively after direct sequencing. *Boxes* show examples of (1) restrictions sites (*A*) for *Mse*I on 1st and 2nd sequences and for *α-Taq*I for sequences 9F and 11F (2) SNP-RS carrying a possible transition (*B*) G → A in all four sequences to generate a *Mse*I restriction site (3) SNP-RS carrying a possible transversion (*C*) T → A in sequences 1st and 2nd to generate a *Mse*I restriction site and T → A in sequences 9F and 11F to generate an *α-Taq*I restriction site (4) unexpected *Mse*I restriction site (*D*) not present on the obtained LCT EEN 162/S-1010 BE consensus sequence but present in all LCT EEN 162/S-1010 samples digested with this enzyme. In this case, the appearance on the non-expected restriction fragment was not considered an artefact

Table 2 Number of SNP-RSs detected in-silico using Artbio

From	To					
	A	T	C	G	Row totals	Relative totals (%)
A	–	260	252	*259*	771	30
T	252	–	*274*	230	756	29
C	158	*182*	–	180	520	20
G	*159*	189	207	–	555	21
Colum totals	569	631	733	669	2,602	
Relative totals (%)	22	24	28	26		

Counts of quasi-palindromic sequences using Artbio software in seven *T. cacao* sequences. In total 2,602 SNP-RSs were detected. The four pairs of elements from the upper right corner to the lower left corner (Italicised) are the values for transitions, while the other eight pairs of elements represent transversions

Loss of restriction sites

A survey of 19 restriction sites by CAPS analysis across the regenerant population for possible loss of restriction showed 18 instances in different individuals. The total number of screened bases considered for the loss of restriction sites was 8,664 [114 samples multiplied by the size of the recognition sequence (four bases for both enzymes used) and multiplied by the 19 restriction sites scored].

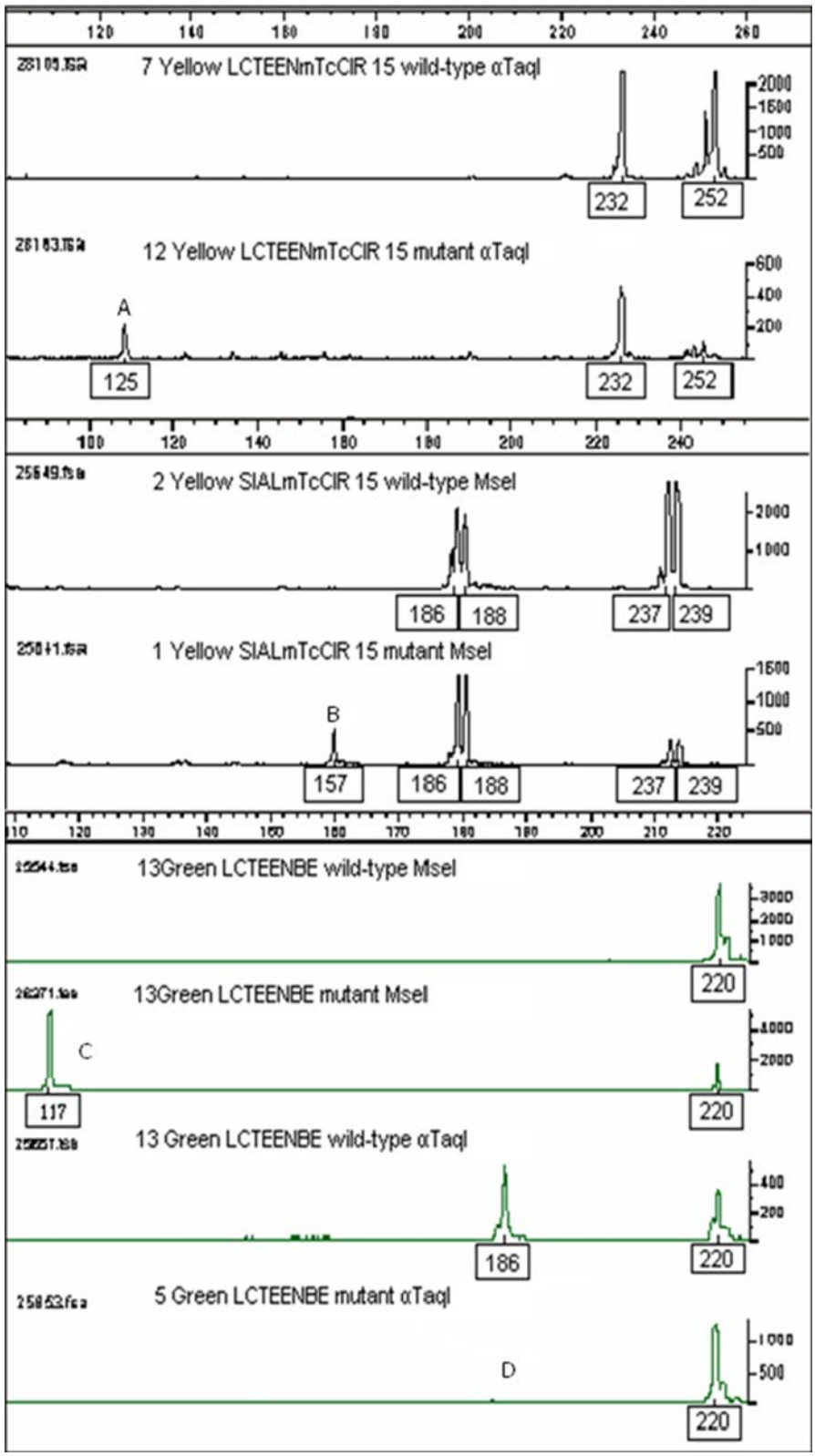

Fig. 2 Detection of SNP's using restriction enzymes in *Theobroma cacao* SSR marker mTcCIR 15 and SCAR marker BE: parental and polymorphic profiles from genotypes LCT EEN 162/S-1010 and SIAL 93 for SSR marker mTcCIR 15 and BE restricted with *Mse*I or α-*Taq*I. *Capital letters* (*A*, *B*, *C* and *D*) indicate polymorphic restriction fragments: *A* indicates creation of a new α-*Taq*I restriction site on the 252 bp mTcCIR 15 allele by a G → T base substitution on base 125. *B* indicates creation of a new *Mse*I restriction site on the 237 bp mTcCIR 15 allele by a G → A base substitution on base 85. *C* indicates creation of a new *Mse*I restriction site by a G → T base substitution on base 117. *D* indicates loss of a α-*Taq*I restriction site that should generate a 186 bp restriction fragment. *Horizontal axis* represents allele size in bp (measured allele size in bp is indicated by *figures in boxes*). *Vertical axis* represents peak intensity. Fragment sizes (in bp) are shown by *figures in boxes below each peak*

Therefore, the inferred mutation frequency arising in culture during somatic embryogenesis was estimated as 2.1×10^{-3} substitutions/base (18/8,664).

Gain of restriction sites

There were 224 potential SNP-RSs identified by Artbio for the restriction enzymes α-*Taq*I and *Mse*I when applied to the two target loci. Therefore, the total number of screened bases for gain of restriction sites was 25,536 (114 regenerant samples multiplied by the 224 SNP-RSs detected). Manual comparison of unexpected restriction fragment sizes to the original sequence traces for detection of sequencing errors showed two instances of unique or rare genotypes in which restriction products did not concur with expectations from sequence information. These were deemed as probable non-SNP mutations (e.g. indels) or artefacts (e.g. anomalous restriction), and were discarded from further analysis. Overall, we retained 30 deviant restriction sites that were present amongst the regenerant population but not in the parental ortet plants. These were all of a size expected from the output of Artbio. The frequency of base substitutions resulting in the gain of a restriction site was 1.2×10^{-3} substitutions/screened base. However, since only a third of all possible base substitutions would be detected by the system (i.e. the base change that creates a restriction recognition motif), a better estimate of mutation frequency was three times this value (3.5×10^{-3}). Thus, the mutation rate estimate arising from the frequency of gained restriction sites was of the same order of magnitude as that noted from the loss of restriction.

Pattern of base substitution

The nature of base substitution detected by a change in CAPS profile was inferred by reference to the restriction site/SNP-RS maps obtained for each sequence-restriction enzyme combination from Artbio. Using this approach we detected a higher level of SNP-RSs originating from template sites occupied by nucleotides A and T than from those containing G or C (59 and 41%, respectively; Table 2). It was also noted that 54% of all SNP-RSs were predicted to generate a new restriction site after a base substitution to either C or G (Table 2).

The 30 new restriction sites created due to single nucleotide base substitutions (Table 3) included 63.9% that originated from template sites occupied by C or G. This represents a significant divergence from the expected one to one ratio (1 C or G: 1 A or T). Most changes (52.1%) resulted in these bases being converted to A or T (Table 3). Overall, we detected 15 transitions and 15 transversions. Thus, the relative frequencies for detected transitions and transversions (53.7 and 46.3%; Table 3) was substantially different from the expected one transition to every two transversions (Chi squared = $1.55\ 10^{-5}$, $P < 0.05$).

Discussion

Artbio

Algorithms for text analysis of biological sequences have been described previously (e.g. Sagot and Wakabayashi 2003) and implemented as tools for the analysis and alignment of biological sequences (Rognes and Seeberg 1998). However, these software tools are usually complex and do not easily allow the inclusion of errors or mutations in their searches. In this study the software package Artbio was developed to aid the study of molecular evolution and allelic diversity amongst uncharacterised populations. By allowing base substitutions and the analysis of multiple sequences at the same time, Artbio is able to identify SNPs between such sequences and to predict which possible base substitutions would generate specified sequences after any given mutation. In this way, Artbio facilitates the molecular description of point mutations from an evolutionary standpoint.

There is increasing interest in characterising mutations across the genome either arising spontaneously or in response to some form of mutagenic agent. One of the cheapest and least technically demanding approach for characterising such change is the use of sequence-specific restriction enzymes (e.g. by CAPS/PCR–RFLP analysis or RFLP). Given the highly variable cost of different restriction enzymes, the ability of Artbio not only to identify existing recognition sites but also to recognise and locate those sites that have the potential to become such sites through a single mutation provides

Table 3 Pattern of somatic embryogenesis-derived base substitutions in cocoa

From	To														
	A			T			C			G			n_i	n_{ij}	f_{ij}
	n_i	n_{ij}	f_{ij} (%)	n_i	n_{ij}	f_{ij} (%)	n_i	n_{ij}	f_{ij} (%)	n_i	n_{ij}	f_{ij} (%)			
A	–	–	–	1,612	1	1.9	113	1	26.5	*934*	*1*	*3.2*	2,659	3	13.1%
T	1,032	1	2.9	–	–	–	*1,258*	*4*	*9.5*	226	0	0	2,516	5	23.0%
C	678	2	8.8	*1,386*	*1*	*2.2*	–	–	–	324	1	9.2	2,388	4	19.4%
G	*2,094*	*9*	*12.9*	1,936	6	9.3	663	3	13.6	–	–	–	4,693	18	44.5%
	3,804	12	34.2	4,934	8	17.9	2,034	8	40.9	1,484	2	7.0	12,256	30	1.2×10^{-3}

The number of single nucleotide polymorphisms detected for each type of base substitution is compared to the number of SNP-RSs. The four pairs of elements from the upper right corner to the lower left corner (Italicised) are the values for transitions, while the other eight pairs of elements represent transversions. Relative frequencies were computed by applying the formula $f_{ij} = \left(P_{ij} \Big/ \sum_i \sum_{j \neq i} P_{ij}\right) \times 100$

researchers with a useful aide in the selection of appropriate enzymes for particular target sequences. One particularly attractive feature of the software is that allows post-hoc re-analysis of existing, possibly unpublished data.

Detection of restriction enzyme recognition sites and single-base deviant motifs in DNA sequences

Analysis of the 6,686 bases of cocoa DNA revealed a frequency of 4-bases restriction enzyme recognition sites and single base variants slightly lower than expectations based on the base composition of target sequences. Equally, the relative frequency of enzyme recognition motifs and single-base variants of these sequences was in agreement with expected ratios made under the assumption of random base-associations and distributions (1/12 and 1/18 for 4- and 6-bp sequences, respectively). These data therefore provide no strong evidence consistent with skewing (either avoidance or preference of 4-base recognition site motifs) within the small but random section of the cocoa genome sampled here.

As expected, the possible 874 transitions and 1,728 transversions leading to the formation of new restriction sites as predicted in-silico by Artbio did not deviate significantly from the one transition for every two transversions anticipated on the basis of random mutation. A deviation from the expected ratio would inevitably have led to a tendency either to favour (if transitions were more frequent than statistically expected) or to avoid (if transversions were more frequent than statistically expected) the creation of new restriction sites. The absence of random deviation reported here therefore indicates, as expected, an absence of bias towards either the creation or loss of restriction enzymes recognition motifs once base composition is taken into account. There were nevertheless indications of an A-T bias among the targeted sequence (59% AT vs. 41% GC). In consequence, Artbio predicted significantly more SNP-RSs originating from A and T nucleotides rather than G and C (59 and 41%, respectively). At the same time, we noted that 54% of all SNP-RSs were predicted to generate a new restriction site after a base substitution resulting in either C or G. In comparison, we found 52.1% of de novo mutations featured new A or Ts among the regenerated plants and 63.9% of all SNP mutations originated from G or C (compared with 41% in the explant material). This is indicative of a modest but significant bias in mutational tendency that favours A:T accumulation at the expense of G:C. Certainly, this tendency is in broad agreement with the AT bias in the starting explant material (59%) and also of that noted in other genomes [*Lycopersicon esculentum* 60%, *Petunia hybrida* 59%, *Pisum sativum* 49.5%, *Triticum aestivum* 56.3% (Marie and Brown 1993) *Medicago truncatula* 61.4–61.9% (Blondon et al. 1994), *Quercus* sp. 61.1% (Zoldos et al. 1998)]. These observations nevertheless contrast with those reported by Noro et al. (2007), who noted an extremely high frequency of transition mutations of A:T to G:C among in vitro cultured cell lines of *Oryza sativa*.

The purpose of the present molecular survey was to illustrate the utility of Artbio rather than to perform a large-scale characterisation of genetic change over the genome amongst cocoa somatic embryo regenerants and so care must be exercised before extrapolating from the limited data set presented here. However, if these results apply more generally then it may be pertinent to question whether AT accumulation is specifically favoured by in vitro regeneration or is systematic across the genome and is revealed as a symptom of the accentuated mutation rate associated with in vitro plant culture. Certainly, the AT bias among newly formed sites correlates well with the base composition bias within the (unmutated) templates within the parental ortet clones (A = 29.8%; T = 29.2%; G = 21.8%; and C = 19.2%).

Enzymatic restriction analysis

Gain versus loss of restriction sites

We generated data for the identification of two enzymes with greatest coverage to detect random point mutations by CAPS/PCR–RFLP analysis across two loci. With this very limited screen, we were nevertheless able to detect mutations leading to restriction site loss (2.1×10^{-3}). This rate is of the same order of magnitude as that observed for mutations that led to a gain in restriction site (3.5×10^{-3}), implying that no strong mutation bias occurs within restriction enzyme recognition sites. Overall, the mean frequency of point mutations observed in this study was 2.8×10^{-3} substitutions/screened base. This rate is similar to that reported previously for cocoa somatic embryogenesis (Rodríguez López et al. 2004) but far higher than has reported previously among chloroplast, mitochondrial, and nuclear genes from in vivo material of other species, where rates typically fall in the range ($0.9–11 \times 10^{-9}$) (Wolfe et al. 1987, 1989; GuhaMajumdar and Sears 2005). It should be noted, however, that rates of molecular change vary quite widely among plant families (Clegg et al. 1994) and even between genes, and can depend on variable functional constraints acting over different parts of the genome (i.e. coding and non-coding regions; Graur and Li 2000). We noted the relationship between time in culture and variation in SSR allele size was consistent with mutational events occurring throughout callus development (data not shown), leading to a progressive accumulation of mutations during the first weeks in culture. This observation was in accordance with the widespread belief that genetic somaclonal variants accumulate with exposure to callus culture (Peschke and Phillips 1992; Bouman and De Klerk 2001). Rather surprisingly, however, embryos initiated late in the lifespan of a callus apparently yielded regenerants with significantly fewer aberrant alleles.

Low mutation rates noted within some species have been correlated with their "cultivability", and is taken by some as an indication of low levels of stress (Phillips et al. 1994). Indeed, high mutation rates among in vitro-regenerated plants can be explained by the fact that plants have developed responses to various environmental stresses (Joyce et al. 2003). Radman (1999) provided a mechanism for such increases by suggesting that under stress conditions, DNA polymerase IV acts on undamaged DNA and randomly introduces new nucleotide bases and, therefore, new mutations. Furthermore, during the cell cycle, the temporary termination of mitosis facilitates DNA repair (Chan et al. 1999; Piwnica-Worms 1999), although Joyce et al. (2003) suggested that this internal control may be disrupted by the in vitro culture conditions because of the presence of unbalanced growth regulators, sucrose, and minerals in the medium.

The use of Artbio in the present study has thus enhanced our capacity to interpret mutations from CAPS analysis and meant that we were able to obtain independent estimates of mutation rates resulting from gain or loss of restriction sites. In effect, this greatly increased the power of analysis without the need for high throughput sequencing.

Patterns of base substitution

This small study was able to provide a strong indication that some nucleotides are more mutable than others. The last column of Table 3 indicates the relative frequencies of mutations from A, T, C, and G. Clearly, if all the four nucleotides had been equally mutable, the expected value for each column element would be 25%. In practice, it was observed that G mutates at a relatively high frequency of 44.5% whereas A mutates at only 13.1%. This correlates with results obtained from human pseudogenes where A and G present ratios of substitution of 20.3 and 29.7, respectively (Graur and Li 2000). The

least frequent base substitution detected in this experiment (apart from the T to G changes of which none were found) were transitions from C to T (2.2%) while, the most common were transversions from A to C (26.5%). There are three major causes of DNA damage (Maki 2002) (1) erroneous action of the replicative apparatus during DNA replication with an intact DNA template and the usual dNTPs; (2) misinsertion of a mutagenic nucleotide, which has a loose specificity of base-pairing, during DNA replication; and (3) chemical reactions by endogenous mutagens, such as active oxygen species, and spontaneous decomposition of DNA bases. Whilst we report elevated mutation rates normally associated with induced mutation, there are some features of the results that differ from those derived using chemical mutagens. First, our results follow "the A rule" which suggests that adenine is preferentially misinserted in spontaneous mutations (Sahasrabudhe et al. 1991). Second, the only slightly higher frequencies of transitions than transversions (53.7 and 46.3%, respectively) detected in the present study closely mirrors the work by Kohler et al. (1991), who reported 56% transitions and 44% transversions among spontaneous mutations in *Mus musculus*. In contrast, the same work also noted a marked increase in transversions following exposure to the chemical mutagens *N*-ethyl-*N*-nitrosourea and benzo(a)pyrene (62 and 100%, respectively). Likewise Cheng et al. (1992)reported 100% transversions among de novo mutations induced by the chemical mutagen 8-hydroxyguanine. Thus, in the present study although we observed a mutation frequency more usually associated with the application of chemical mutagens, the ratio of transversions to transitions is more typical of that seen amongst spontaneous mutations. Given that our work implies that there is a tendency for C and G to be more mutable (63.9% of all mutations) and a slight tendency for mutations to produce A or T (52.1% rather than 50%), then this may help to explain the observation that studied sequences are AT-rich (59%). Other, more substantial studies have noted similar compositional bias in genomic DNA, and suggested that these result from a mutational bias (Gautier 2000). It is a matter worth noting that non-coding regions are generally recognised as being more AT-rich than coding regions (Graur and Li 2000). Given that we detected a similar de novo A–T mutation bias arising from culture, it is tempting to speculate that some aspect of the DNA mutation-repair mechanism that acts to ameliorate the A–T mutation bias but operates preferentially in coding regions. Clearly, more exhaustive studies would be required to test this hypothesis.

To conclude, the application of Artbio software has enabled the application of low-cost CAPS/PCR–RFLP analysis to characterise the nature of point mutations in more detail than would have been possible otherwise. A simple protocol has been developed to detect the ratio and patterns of nucleotide substitution based on the predicted creation of new restriction sites, allowing not only a quantitative analysis of the frequency of base substitution but also a qualitative study of the molecular changes occurring during tissue culture-induced point mutation. Three conclusions may be made from the survey. First, the frequency of tissue culture-induced base substitution is much higher than the expected spontaneous mutation rate. Second, the type of changes seen (i.e. some bases more mutable than others, transitions are more frequent than transversions, etc.) are generally of the same kind as those occurring in vivo. Third, the software we have developed in this paper may be of use in allowing efficient conversion of monomorphic SCAR markers into polymorphic CAPS/PCR RFLP markers.

Acknowledgments We thank Cocoa Research UK for funding this study.

References

Albani MC, Wilkinson MJ (1998) Inter simple sequence repeat polymerase chain reaction for the detection of somaclonal variation. Plant Breed 117:573–575

Bala A, Murphy P, Giller KE (2003) Distribution and diversity of rhizobia nodulating agroforestry legumes in soils from three continents in the tropics. Mol Ecol 12:917–930

Blondon F, Marie D, Brown S, Kondorosi A (1994) Genome size and base composition in *Medicago sativa* and *M. Truncatula* species. Plant Genet Breed 37(2):264–270

Bouman H, De Klerk GJ (2001) Measurement of the extent of somaclonal variation in begonia plants regenerated under various conditions. Comparison of three assays. Theor Appl Genet 102:111–117

Britt BA (1999) Molecular genetics of DNA repair in higher plants. Trends Plant Sci 4:20–22

Chan TA, Hermekin H, Lengauer C, Kinzler KW, Vogelstein B (1999) 14–3-3 is required to prevent mitotic catastrophe after DNA damage. Nature 401:616–620

Charters YM (2000) The potential of anchored microsatellite analysis for cocoa germplasm characterization. PhD thesis, School of Plant Sciences, Reading University

Cheng KC, Cahill SD, Kasai H, Nishimura S, Loeb LA (1992) 8-Hydroxyguanine, an abundant form of oxidative DNA damage, causes G-T and A-C substitutions. J Biol Chem 267(1):166–172

Clegg MT, Gaus BS, Learn GH, Morton BR (1994) Rates and patterns of chloroplast DNA evolution. Proc Natl Acad Sci USA 91:6795–6801

Cotton RGH (1997) Mutation detection. Oxford University Press, Oxford

Dahl F, Stenberg J, Fredriksson S, Welch K, Zhang M, Nilsson M, Bicknell D, Bodmer WF, Davis RW, Ji H (2007) Multigene amplification and massively parallel sequencing for cancer mutation discovery. Proc Natl Acad Sci USA 104(22):9387–9392

Devos KM, Beales J, Ogihara Y, Doust AN (2005) Comparative sequence analysis of the *Phytochrome C* Gene and its upstream region in allohexaploid wheat reveals new data on the evolution of its three constituent genomes. Plant Mol Biol 58(5):625–641

Fang JY, Wetten A, Hadley P (2004) Cryopreservation of cocoa (*Theobroma cacao* L.) somatic embryos for long-term germplasm storage. Plant Sci 166(3):669–675

Gautier C (2000) Compositional bias in DNA. Curr Opin Genet Dev 10:656–661

González-Martínez SC, Wheeler NC, Ersoz E, Nelson CD, Neale DB (2007) Association genetics in *Pinus taeda* L. I. Wood Property Traits. Genetics 175:399–409

Graur D, Li WH (2000) Fundamentals of molecular evolution. Sinauer, Sunderland

GuhaMajumdar M, Sears BB (2005) Chloroplast DNA base substitutions: an experimental assessment. Mol Gen Genomics 273:177–183

Jin N, Chow CY, Liu L, Zolov SN, Bronson R, Davisson M, Petersen JL, Zhang Y, Park S, Duex JE, Goldowitz D, Meisler MH, Weisman LS (2008) VAC14 nucleates a protein complex essential for the acute interconversion of PI3P and PI(3, 5)P2 in yeast and mouse. EMBO J 27(24):3221–3234

Joyce SM, Cassells AC, Jain SM (2003) Stress and aberrant phenotypes in in vitro culture. Plant Cell Tissue Organ Cult 74:103–121

Kaeppler SM, Phillips RL, Olhoft P (1998) Molecular basis of heritable tissue culture-induced variation in plants. Current plant science and biotechnology in agriculture. Kluwer, Dordrecht

Karling S, Burge C, Campbell AM (1992) Statistical analyses of counts and distributions of restriction sites in DNA sequences. Nucl Acids Res 20:1363–1370

Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16(2):111–120

Kohler SW, Provost GS, Fieck A, Kretz PL, Bullock WO, Sorge JA, Putmant DL, Short JM (1991) Spectra of spontaneous and mutagen-induced mutations in the lacI-gene in transgenic mice. Proc Natl Acad Sci USA 88:7958–7962

Lanaud C, Risterucci AM, Pieretti I, Falque M, Bouet A, Lagoda PJL (1999) Isolation and characterization of microsatellites in *Theobroma cacao* L. Mol Ecol 8:2141–2143

Li Z, Traore A, Maximova SN, Gupta PK (1998) Somatic embryogenesis and plant regeneration from floral explants of cocoa (*Theobroma cacao* L.) using thidiazuron. In Vitro Cell Dev Biol Plant 34:293–299

Lopez-Baez O, Bollon H, Eskes A (1993) Embryogenèse somatique de cacaoyer *Theobroma cacao* L. à partir de pièces florales. C R Acad Sci Paris 316:579–584

Maki H (2002) Origins of spontaneous mutations: specificity and directionality of base-substitution, frameshift, and sequence-substitution mutageneses. Annu Rev Genet 36:279–303

Marie D, Brown SC (1993) A cytometric exercise in plant DNA histograms, with 2C values for 70 species. Biol Cell 78:41–51

Noro Y, Takano-Shimizu T, Syono K, Kishima Y, Sano Y (2007) Genetic variations in rice in vitro cultures at the EPSPs-RPS20 region. Theor Appl Genet 114(4):705–711

Oh TJ, Cullis MA, Kunert K, Engelborghs I, Swennen R, Cullis CA (2007) Genomic changes associated with somaclonal variation in banana (*Musa* spp.). Physiol Plant 129(4): 766–774

Paton NW, Khan SA, Hayes A, Moussouni F, Brass A, Eilbeck K, Goble CA, Hubbard SJ (2000) Conceptual modelling of genomic information. Bioinformatics 16(6):548–557

Peschke VM, Phillips RL (1992) Genetic implications of somaclonal variation in plants. Adv Genet 13:41–76

Phillips RL, Kaeppler SM, Olhoft P (1994) Genetic instability of plant tissue cultures: breakdown of normal controls. Proc Natl Acad Sci USA 91:5222–5226

Piwnica-Worms H (1999) Fools in rush in. Nature 401:535–537

Radman M (1999) Enzymes of evolutionary change. Nature 401:866–869

Risterucci AM, Grivet L, N'Goran J, Pieretti I, Flament MH, Lanaud C (2000) A high density linkage map of *Theobroma cacao* L. Theor Appl Genet 101:948–955

Rodríguez López CM, Wetten AC, Wilkinson MJ (2004) Detection and quantification of in vitro-culture induced chimerism using simple sequence repeats (SSR) analysis in *Theobroma cacao* (L.). Theor Appl Genet 110:157–166

Rognes T, Seeberg E (1998) SALSA: improved protein database searching by a new algorithm for assembly of sequence fragments into gapped alignments. Bioinformatics 14:839–845

Sagot M-F, Wakabayashi Y (2003) Pattern inference under many guises. In: Reed BA, Sales CL (eds) Recent advances in algorithms and combinatorics. Springer, New York, pp 245–287

Sahasrabudhe SR, Luo X, Humayun MZ (1991) Specificity of base substitutions induced by the acridine mutagen ICR-191: mispairing by guanine N7 adducts as a mutagenic mechanism. Genetics 119:981–989

Sawa S, Ito T, Okada K (1997) A rapid method for detection of single base changes in *Arabidopsis thaliana* using the polymerase chain reaction. Plant Mol Biol Reporter 15:179–185

Thomas RK, Nickerson E, Simons JF, Jänne PA, Tengs T, Yuza Y, Garraway LA, LaFramboise T, Lee JC, Shah K, O'Neill K, Sasaki H, Lindeman N, Wong K-K, Borras AM, Gutmann EJ, Dragnev KH, DeBiasi R, Chen T-H, Glatt KA, Greulich H, Desany B, Lubeski CK, Brockman W, Alvarez P, Hutchison SK, Leamon JH, Ronan MT, Turenchalk GS, Egholm M, Sellers WR, Rothberg JM, Meyerson M (2006) Sensitive mutation detection in heterogeneous cancer specimens by massively parallel picoliter reactor sequencing. Nature Med 12:852–855

Thomas RK, Baker AC, Debiasi RM, Winckler W, Laframboise T, Lin WM, Wang M, Feng W, Zander T, Macconnaill LE, Lee JC, Nicoletti R, Hatton C, Goyette M, Girard L, Majmudar K, Ziaugra L, Wong KK, Gabriel S, Beroukhim R, Peyton M, Barretina J, Dutt A, Emery C (2007) High-throughput oncogene mutation profiling in human cancer. Nature Genet 39:347–351

Traore A, Maximova SN, Guiltinan MJ (2003) Micropropagation of *Theobroma cacao* L. using somatic embryo-derived plants. In vitro Cell Dev Biol Plant 39:332–337

Vettori C, Vendramin GG, Anzidei M, Pastorelli R, Paffetti D, Glannini R (2004) Geographic distribution of chloroplast variation in Italian populations of beech (*Fagus sylvatica* L.). Theor Appl Genet 109:1–9

Wang D, Kreutzer DA, Essigmann JM (1998) Mutagenicity and repair of oxidative DNA damage: insights from studies using defined lesions. Mutat Res 400(1–2):99–115

Whittaker JC, Harbord RM, Boxall N, Mackay I, Dawson G, Sibly RM (2003) Likelihood-based estimation of microsatellite mutation rates. Genetics 164:781–787

Wolfe KH, Sharp PM, Wen-Hsiung L (1987) Rates of nucleotide substitutions vary greatly among plant mitochondrial, chloroplast and nuclear DNAs. Proc Natl Acad Sci USA 84:9054–9058

Wolfe KH, Sharp PM, Wen-Hsiung L (1989) Rates of synonymous substitutions in plant nuclear genes. Mol Evol 29:208–211

Yu J, Buckler ES (2006) Genetic association mapping and genome organization of maize. Curr Opin Biotech 17(2): 155–160

Zakeri S, Mehrizi AA, Djadid ND, Snounou G (2006) Circumsporozoite protein gene diversity among temperate and tropical *Plasmodium vivax* isolates from Iran. Trop Med Int Health 11(5):729–737

Zoldos V, Papes D, Brown SC, Panaud O, Siljak-Yakovlev S (1998) Genome size and base composition of seven *Quercus* species: inter- and intra-population. Genome 41:162–168