Theor Appl Genet (2004) 109:129–139
DOI 10.1007/s00122-004-1603-2

ORIGINAL PAPER

R. Aert · L. Sági · G. Volckaert

Gene content and density in banana (*Musa acuminata*) as revealed by genomic sequencing of BAC clones

Received: 4 December 2003 / Accepted: 20 January 2004 / Published online: 18 February 2004

Abstract The complete sequence of *Musa acuminata* bacterial artificial chromosome (BAC) clones is presented and, consequently, the first analysis of the banana genome organization. One clone (MuH9) is 82,723 bp long with an overall G+C content of 38.2%. Twelve putative protein-coding sequences were identified, representing a gene density of one per 6.9 kb, which is slightly less than that previously reported for Arabidopsis but similar to rice. One coding sequence was identified as a partial *M. acuminata* malate synthase, while the remaining sequences showed a similarity to predicted or hypothetical proteins identified in genome sequence data. A second BAC clone (MuG9) is 73,268 bp long with an overall G+C content of 38.5%. Only seven putative coding regions were discovered, representing a gene density of only one gene per 10.5 kb, which is strikingly lower than that of the first BAC. One coding sequence showed significant homology to the soybean ribonucleotide reductase (large subunit). A transition point between coding regions and repeated sequences was found at approximately 45 kb, separating the coding upstream BAC end from its downstream end that mainly contained transposon-like sequences and regions similar to known repetitive sequences of *M. acuminata*. This gene organization resembles Gramineae genome sequences, where genes are clustered in gene-rich regions separated by gene-poor DNA containing abundant transposons.

Communicated by J.S. Heslop-Harrison

R. Aert · G. Volckaert
Laboratory of Gene Technology,
Katholieke Universiteit Leuven,
Kasteelpark Arenberg 21, 3001 Leuven, Belgium

Present address:
R. Aert (✉) · L. Sági, Laboratory of Tropical Crop Improvement,
Katholieke Universiteit Leuven,
Kasteelpark Arenberg 13, 3001 Leuven, Belgium
e-mail: Rita.Aert@agr.kuleuven.ac.be
Tel.: +32-16-321683
Fax: +32-16-321993

Introduction

The continuous advance in high-throughput DNA automation and sequencing technologies has resulted in important breakthroughs in plant science. An example of this progress is the recent release of the high-accuracy complete sequence of rice chromosomes 1 (Sasaki et al. 2002) and 4 (Feng et al. 2002). High-quality sequencing becomes more desirable in view of current data, which demonstrate that the degree of collinearity between rice and Arabidopsis (currently the only other plant species with a draft genome sequence) is rather low at the level of the whole genome (Liu et al. 2001; Salse et al. 2002). An additional consequence is that ongoing genome projects in other plant species might not be able to rely on the information obtained from genomes already sequenced to the extent that had been expected. Therefore, despite the development of ever more intelligent algorithms for comparative genome analysis, the importance of individual genome projects in higher organisms cannot be underestimated.

With an annual production of about 100 million tons, banana and plantain (*Musa* sp.) are the most important fruit crop on worldwide and a staple food for about 400 million people in more than 120, mainly less-developed countries, in which it accounts for up to 90% of the carbohydrates consumed. However, this species has not been extensively used in genetic studies as triploidy and sterility is prevalent in most cultivars, which also results in low genetic variability. On the other hand, with a haploid genome size of 500–600 Mbp (Lysák et al. 1999), the banana genome is among the smaller ones found within non-graminaceous monocotyledons. Natural hybridizations between the wild diploid species, *Musa acuminata* Colla (A genome, $2n=2x=22$) and *Musa balbisiana* Colla (B genome, $2n=2x=22$) have given rise to various genome combinations at three ploidy levels.

This characteristic turns banana into an interesting candidate for comparative genomics. Being a monocotyledon but distantly related to rice, banana could represent a useful comparison point between dicotyle-

donous and monocotyledonous genomes. In addition, a number of important traits, not present in model plants, can be functionally analysed in banana. These unique features include (1) banana fruit physiology, which has been a model for a wide range of other fruit crops, and (2) the breeding system, both sexual and vegetative, characterized by different forms of sterility in combination with parthenocarpy that is rare in monocots. This breeding system, as well as the not widely known fact that banana was among the first crops domesticated (Simmonds 1966; Denham et al. 2003), enables the identification of variation related to natural mutations and polyploidy, as many sterile clones have been fixed for thousands of years by vegetative propagation in the same environment (Gowen 1995; Robinson 1996). In parallel, partially and highly fertile wild diploids have also been adapted to the same environment, making banana a fascinating model by which to study both plant evolution and plant-pathogen co-evolution at a genomic level. An attractive example for the latter is the integration of the banana streak badnavirus in the plant genome (Harper et al. 1999; Ndowora et al. 1999), which can be reactivated after recombination. Similarly, the widespread presence of *gypsy*-like long terminal region (LTR) retroelements (200–500 copies per haploid genome, Balint-Kurti et al. 2000) and Ty1-*copia*-like retrotransposons (Teo et al. 2002; Baurens et al. 1997) makes a challenge for genome studies in banana.

According to the current understanding, most plant genomes are organized into long clusters of genes occupying 12–25% of the genome ('gene space') that are separated by long stretches of gene-empty regions consisting mainly of repetitive sequences (Walbot and Petrov 2001). This explains why the observed gene density is similar between largely different plant genomes (Keller and Feuillet 2000; Bevan et al. 2001), whereas the expected gene density (based on random gene distribution) should be variable because of the variation in genome size. It was expected and has been confirmed that genes are more densely packed in large plant genomes than would be expected. Characteristic examples of such gene-dense regions are the *bronze* locus in maize (Fu et al. 2001) and the R-gene clusters in a wide range of plant species. In tomato, for example, five members of the *Cf* gene family span a cluster of about 35 kb (Parniske et al. 1997), and the *Fusarium I2* gene cluster contains seven genes over 90 kb (Simons et al. 1998). In maize, Webb et al. (2002) identified five rust resistance gene (*rp3*) paralogues in a region of 140 kb, and comparable gene densities were observed at the *rp1* locus in maize and sorghum (Ramakrishna et al. 2002) and in the *Lr21* region of *Triticum* (Brooks et al. 2002).

The nuclear genome organization of *Arabidopsis thaliana*, however, is drastically different from that of the large genomes of the Gramineae (Barakat et al. 1998). *Arabidopsis* genes are fairly evenly distributed over regions amounting to about 85% of the genome, whereas gene-empty regions are greatly reduced. In the rice chromosome 4 sequence (Feng et al. 2002), repeat sequences also seem to be dispersed along the whole chromosome with no obvious clustering. In rice chromosome 10, however, enrichment of repetitive elements on the short arm and enrichment of expressed genes on the long arm has been described (The Rice Chromosome 10 Sequencing Consortium 2003).

In the light of this global and local genome heterogeneity it would be interesting to study how the genome of other, previously not characterized plants is related to model plant genomes. As a first step in this direction, we present here a sequence analysis of two randomly chosen bacterial artificial chromosomes (BAC) clones from a wild diploid banana (*Musa acuminata*) with a combined length of 156 kb. The identification of genes and various repetitive as well as mobile elements provides a first insight in the genome organization of a non-graminaceous monocot plant.

Materials and methods

Plant material and BAC DNA preparation

A *Musa acuminata Bam*HI library from the wild diploid Calcutta 4 (International Transit Centre accession no. ITC.0249) was generated by a research consortium (INCO-DC project IC18-CT97–0192; James et al. unpublished). This genotype has been selected as a standard by the Global *Musa* Genomics Consortium (Gewolb 2001), which currently comprises 27 public institutions from 13 countries.

Ninety-six BAC clones from this library were obtained from the Musa Genome Resources Centre (http://www.inibap.org/mgrc/). DNA preparation of BAC clones was carried out on a BioRobot 9600 (Qiagen, Hilden, Germany) using the R.E.A.L. BioRobot kit (Qiagen).

BAC clone sequencing

The *M. acuminata* BAC clones were selected after restriction enzyme analysis and sequenced by standard methods using a shotgun approach (Bodenteich et al. 1993). DNA purified by Qiagen Q-Tip100 was sheared by nebulization (Roe et al. 1996) to an average size of 1.5 kb. After end-filling, DNA fragments were size-fractionated on HPEC (Applied Biosystems, Foster City, Calif.) and cloned into the *Sma*I site of pUC18 (Amersham Biosciences, Piscataway, N.J.). The resulting plasmids were electroporated into XL1-blue cells (Stratagene, La Jolla, Calif.). Clones were picked into 96-well trays filled with LB culture medium, grown for 2 h and frozen until needed. The clones were sequenced with the ABI PRISM Dye Terminator Cycle Sequencing Ready Reaction kit (Applied Biosystems) on ABI 377 (Applied Biosystems) sequencing gels. The sequence data were assembled using Sequencher software (Gene Codes, Ann Arbor, Mich.).

Analysis of sequence data

The complete BAC sequences and the predicted coding sequences (CDSs) were subjected to BLAST (Altschul et al. 1997) or FASTA analysis against all GenBank+RefSeq Nucleotides+EMBL+DDBJ+PDB sequences. In addition, the whole sequences were searched by WU-BLAST 2.0 (Gish and States 1993; http://blast.wustl.edu) against the TIGR (The Institute for Genomic Research) Rice Gene Index (http://tigrblast.tigr.org/tgi/), the TIGR Rice Repeat and Cereal Repeat Databases (http://www.tigr.org/tdb/e2k1/osa1/blastsearch.shtml) and Repbase (Jurka 2000) with plant or rice

settings using CENSOR (Jurka et al. 1996). Gene predictions were perfor- med using DIGIT version 1.0 (Yada et al. 2003; http://digit.gsc.riken.go.jp), FGENESH version 1.1 (Salamov and Solovyev 2000; http://www.softberry.com), GENEMARK.HMM version 2.2a (Lukashin and Borodovsky 1998; http://opal.biology.gatech.edu/GeneMark/), and GENSCAN version 1.0 (Burge and Karlin 1997; http://genes.mit.edu/GENSCAN.html). For FGENESH, monocot settings were used; for GENEMARK.HMM, the *O. sativa* settings; for GENSCAN, the maize settings were taken into account. A gene with significant homology [E less than (e−20)] to a known protein is classified according to the protein name as 'putative' or 'like protein'. A gene without significant homology to any protein but with substantial expressed sequence tag (EST) homology or with homology to a protein with EST homology is classified as 'expressed' or 'unknown' protein. A gene identified with a gene prediction programme (or homologous to a gene identified with a gene prediction programme) only is classified as a 'hypothetical' one.

Sequences were analysed and masked by REPEATMASKER (Smit and Green, unpublished; http://ftp.genome.washington.edu/RM/RepeatMasker.html) prior to use for gene prediction. The identified genes were checked by the domain search software INTERPRO (Mulder et al. 2003; http://www.ebi.ac.uk/interpro/). ESTs were located on the sequences by searching a *Musa* database donated by Syngenta to the Global *Musa* Genomics Consortium and the RGP EST database (http://riceblast.dna.affrc.go.jp/). The search for tRNA-encoding genes was performed by TRNASCAN-SE version 1.21 (Lowe and Eddy 1997; http://www.genetics.wustl.edu/eddy/tRNAscan-SE/). Repeats and LTRs were also analysed using PALINDROME (http://bioweb.pasteur.fr/seqanal/interfaces/palindrome.html).

Results

Selection of BAC clones

In order to estimate the insert size of the BACs, all 96 clones were analysed using three restriction enzymes (*Apa*LI, *Eco*RV and *Dra*III). We randomly chose two clones from those BACs showing a clear restriction pattern and containing inserts of adequate length. For MuH9, an insert length of 80 kb was estimated, and for MuG9, the insert size was approximately 70 kb. The shotgun sequencing of 314 (MuH9) and 470 (MuG9) subclones provided 204.1 kb and 282 kb, respectively, of high-quality bases. Finishing was carried out with 256 (MuH9) and 284 (MuG9) sequencing reactions. A final sequence coverage of 3.5-fold for MuH9 and 3.8-fold for MuG9 was obtained. Sequences from the two BACs were deposited in Genbank (accession nos.: MuG9=AY484588, MuH9=AY484589).

BAC clone MuH9

For MuH9 (82,723 bp), an overall G+C content of 38.2% was calculated (Table 1). BLASTN analysis against all GenBank+RefSeq Nucleotides+EMBL+DDBJ+PDB sequences was performed, and one short (335 bp) fragment in a 3-kb intron had significant sequence identity (E=4e−76) with the 23S rRNA gene of plant chloroplast DNA.

Four gene prediction programmes (DIGIT, FGENESH, GENEMARK.HMM, GENSCAN) were used for modeling exon-intron structure, and these detected a total of 17 candidates (Table 2). FGENESH and GENEMARK.HMM concurrently predicted the presence of 10 and 11 CDSs from the 12 positively identified ones (numbered from H9-1 to H9-12 in Fig. 1 and Table 3). The exon-intron structure of the putative genes differed between the two programmes. GENEMARK.HMM separated H9-7 (*Lotus japonicus* Ca-binding protein, E=2.6e−64) and H9-8, whereas FGENESH detected a single CDS in that position. GENSCAN identified more or less correctly only H9-1, H9-2, H9-6, H9-11 and H9-12, whereas DIGIT appeared to detect the least number of CDSs. Of the four prediction programmes, GENEMARK.HMM clearly found the highest number of best fitting hits in this BAC clone.

BLASTP analysis identified H9-1 (E=0) as a truncated malate synthase gene that coded for only 483 out of the total 556 amino acids of the *M. acuminata* malate synthase (Pua et al. 2003). H9-6 (distinguished by all programmes) showed extensive homology (E=0) to rice OSCR4, a maize crinkly 4 orthologue. The remaining CDSs were identified by BLASTP analysis as putative proteins with similarities to different hypothetical proteins present in the databases. The presence of functional domains in the predicted proteins was explored using INTERPRO, and characterized domains were found for 10 of the 12 putative genes (Table 3).

The exon-intron structures of malate synthase and the crinkly4 candidate were compared with all known genomic sequences from databases. The malate syn-

Table 1 Characteristics of the sequenced banana BAC clones MuH9 and MuG9

BAC clone (length in bp)	MuH9 (82,723)	MuG9 (73,268)	Average values
GenBank no.	AY484589	AY484588	
Base composition (%)			
A/T−T/A=	61.8	61.4	61.6
C/G−G/C=	38.2	38.6	38.4
Predicted genes[a]	12	7	9.5
Gene density[a] (kb)	6.9	10.5	8.7
Gene regions[a] (%)	43.8	27.3	35.6
tRNAs (%)	0	0	0
LTR retrotransposons (%)	0	10.5	4.9
Other repetitive elements (%)	3.4	4.1	3.7
Non-annotated sequences (%)	52.8	58.1	55.8

[a] As determined by GENEMARK.HMM. For MUG9, the polyprotein-like sequences (*G9pp1–G9pp4*) are not taken into account

Table 2 Comparison of gene prediction programs in the sequenced *Musa* BAC MuH9

CDS[a]	DGT[b] (masked or unmasked)	FSH[c] (masked)	GMK	GSN[c]
1	**59–1,929 (–), 4 ex[c]**	65–1,929 (–), 4 ex[c]	65–1,929 (–), 4 ex[c]	29–1,929 (–), 4 ex[c]
2			**4,765–5,238 (+), 1 ex**	4,811–5,148 (–), 2 ex
3		8,223–14,668 (–), 12 ex	**11,060–14,668 (–), 6 ex**	11,060–12,469 (–), 3 ex
4				14,667–16,209 (+), 2 ex
5		16,123–18,805 (–), 6 ex	**16,789–18,916 (–), 3 ex**	
6	22,865–34,662 (+), 7 ex	22,865–36,888 (+), 19 ex	**22,864–32,851 (+), 14 ex**	17,798–32,852 (+), 5 ex
7		**38,972–41,656 (–), 1 ex**	**38,972–41,656 (–), 1 ex**	38,941–41,564 (–), 4 ex
8	41,929–46,530 (+), 2 ex or 45,396–46,530 (+), 2 ex		41,929–42,252 (+), 1 ex	
9			43,380–44,036 (–), 3 ex	43,860–46,530 (+), 2 ex
10			**45,395–46,529 (+), 2 ex**	
11	53,688–57,588 (+), 4 ex	46,078–57,588 (+), 11 ex	**47,752–57,864 (+), 11 ex**	
12				53,101–59,263 (-), 2 ex
13			**59,398–59,865 (+), 1 ex**	
14		63,279–65,890 (–), 3 ex	**63,352–64,420 (–), 2 ex**	
15				70,799–71,291 (+), 3 ex
16	68,512–73,443 (–), 6 ex	72,051–73,443 (–), 4 ex	72,050–73,442 (–), 4 ex	**72,051–73,443 (–), 4 ex**
17	79,646–81,297 (+), 5 ex	79,646–81,405 (+), 6 ex	**79,646–81,877 (+), 7 ex**	79,646–80,517 (+), 2 ex[c]
Total	6	9	14	11
Best:	**1**	**-**	**10**	**1**

[a] Coding sequences (CDS) are numbered starting from the 5′ end of the DNA sequence as deposited in GenBank
[b] Localization of predicted coding sequences on plus or minus strand is indicated in brackets followed by the number of exons predicted; the most correct coding sequences, i.e. the ones with the highest homology to hits in the database, are indicated in bold. Prediction programmes: DGT, DIGIT; FSH, FGENESH; GMK, GENEMARK.HMM; GSN, GENSCAN
[c] Truncated genes

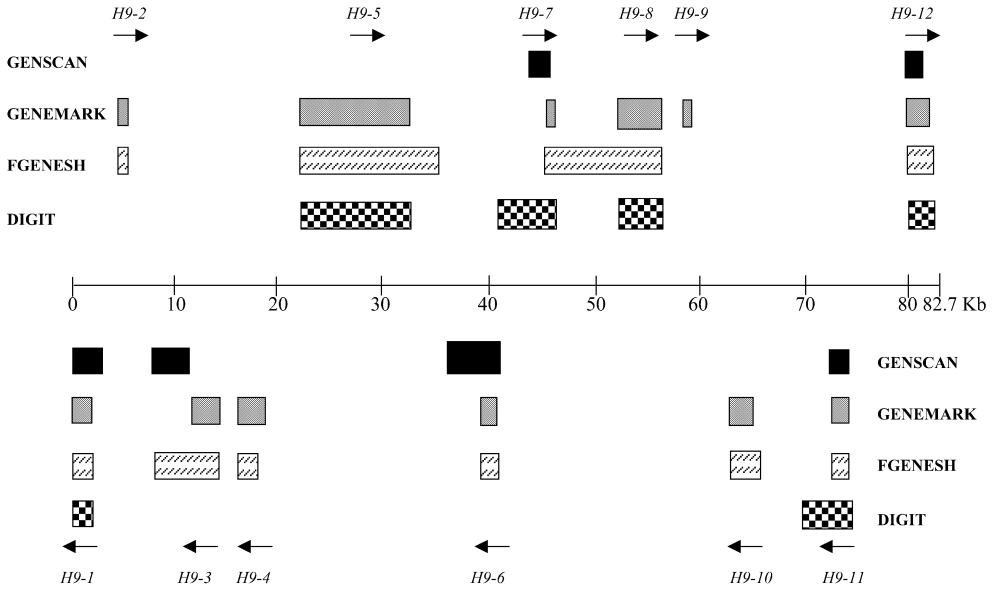

Fig. 1 Schematic overview of MuH9 (AY484589) coding sequences for GENSCAN, GENEMARK, FGENESH and DIGIT predictions. The genes' start and end are depicted without detailed indication of exons

thase genes from rice (accession nos. AL662946 and AL627350), *Arabidopsis* (AL162873) and cucumber (X15425) as well as the banana orthologue all contained four exons with a similar length distribution. The *M. acuminata* malate synthase gene is truncated at its 5′ end in MuH9 and, consequently, its first exon is shorter. As expected, a higher length and sequence variation was observed among the introns of malate synthase genes. The distribution and average length (142 bp) of the introns in banana were more similar to those of rice (145 bp), whereas the average length of the introns in the malate synthase genes of *Arabidopsis* and cucumber was markedly longer (190 bp and 292 bp, respectively). In terms of overall sequence homology, the banana gene was again the closest to the two rice sequences (BLASTN 346, E=7e–92). The 2,685-bp single exon of the banana crinkly 4 orthologue was also more homologous to rice (AB057787, BLASTN 546, E=e–152) or the maize mRNA sequence (U67422, BLASTN 619, E=e–174) than to the better fitting of two *Arabidopsis* genes (AL356014, BLASTN 392, E=e–105).

Table 3 Predicted genes in the sequenced *Musa* BAC MuH9

Gene ID[a]	Number of exon(s)	Number of amino acids	InterPro domains	Putative function[b]; supporting evidence[c]
H9-1 (−)[d]; 59–1,929	4	483	Malate synthase (IPR001465) (EC 4.1.3.2), malate synthase A (IPR006252)	*Musa acuminata* malate synthase, AF321286 (980, E=0); DGT, FSH, GMK, GSN
H9–2 (+); 4,765–5,238	1	158	Zinc finger, AN1-like (IPR000058)	*Arabidopsis thaliana* expressed (PRP3-like) protein, At3g12630 (105, E=2e−22); GMK, FSH, GSN
H9-3 (−); 11,060–14,668	6	148	-	*A. thaliana* 4-nitrophenylphosphatase-like protein, BAA98057 (242, E=6e−64); GMK, FSH, GSN
H9-4 (−); 16,789–18,916	3	639	BED-finger (IPR003656), DUF659 (IPR007021)	*Oryza sativa* unknown (transposase-like) protein, BAA94530 (332, E=9e-90); *Musa* EST 600017447T1 (585, E=e−165); GMK, FSH
H9-5 (+); 22,864–32,851	14	492	Zn-finger, RING (IPR001841) Zn-finger, C6HC (IPR002867), 0001876	*A. thaliana* ARIADNE-like (ARI7, ARI8) protein, At2g31510 (724, E=0); GMK, FSH, DGT, GSN
H9-6 (−); 38,972–41,656	1	894	Protein kinase (IPR000719)	*O. sativa* crinkly4 (CR4), BAB68389 (1344, E=0); GMK, FSH, GSN
H9-7 (+); 45,395–46,529	2	210	Ca-binding EF-hand (IPR002048)	*A. thaliana* Ca-binding (calmodulin-like) protein, At3g07490 (248, E=6e-65); *Musa* EST 600105768T1 (799, E=0); GMK, DGT, GSN
H9-8 (+); 47,752–57,864	11	1,382	Plant regulator RWP-RK (IPR003035), octicosapeptide/Phox/Bem1P (IPR000270), Zn-finger, TRAF-type (IPR001293)	*O. sativa japonica* CAD41257 (679, E=0); *Musa* ESTs 600164505T1 (416, E=e−114) and 600084815T1 (206, E=8e−51), rice ESTs AU197502 (200, 8e−49) and AU108333 (184, 5e−44); GMK, FSH, DGT
H9-9 (+); 59,398–59,865	1	155	-	*A. thaliana* expressed protein, At3g59680 (136, E=e−31); GMK
H9-10 (−); 63,352–64,420	2	313	Cyclin-like F-box (IPR001810), leucine-rich repeat, cysteine subtype (IPR006553)	*A. thaliana* F-box-like protein, At3g07550 (278, E=3.0e−77); GMK, FSH
H9-11 (−); 72,646–81877	4	295	Short-chain dehydrogenase/reductase SDR (IPR002198)	*O. sativa japonica* short-chain hydrogenase/reductase, CAD41255 (368, E=e−101); GSN, GMK, FSH, DGT
H9-12 (+); 79,646–81,877	7	308	RNA-binding region RNP-1 (RNA recognition motif, RRM, IPR000504)	*O. sativa japonica* expressed (RNA-binding) protein, CAD41253 (276, E=4e−73); GMK, FSH, GSN, DGT

[a] Genes are numbered starting from the 5′ end of the DNA sequence as deposited in GenBank. Localization of a predicted gene on plus or minus strand is indicated in brackets
[b] Protein function is based on BLASTP results with the score and E-value indicated in brackets
[c] Evidence supporting the existence of gene is as follows: EST, similar gene transcripts with GeneBank accession number and BLASTN score and E-value in brackets. DGT, DIGIT; FSH, FGENESH; GMK, GENEMARK.HMM; GSN, GENSCAN (ordered according to the precision of prediction for each entry
[d] Truncated CDS

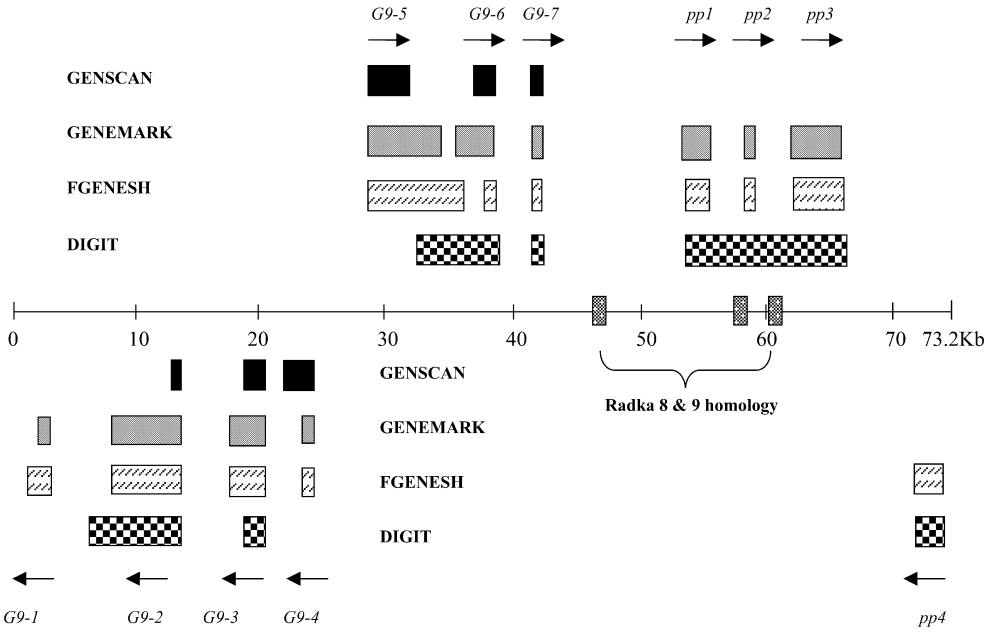

Fig. 2 Schematic overview of MuG9 (AY484588) coding sequences for GENSCAN, GENEMARK, FGENESH and DIGIT predictions. The genes' start and end are depicted without detailed indication of exons. *pp* Polyprotein-like

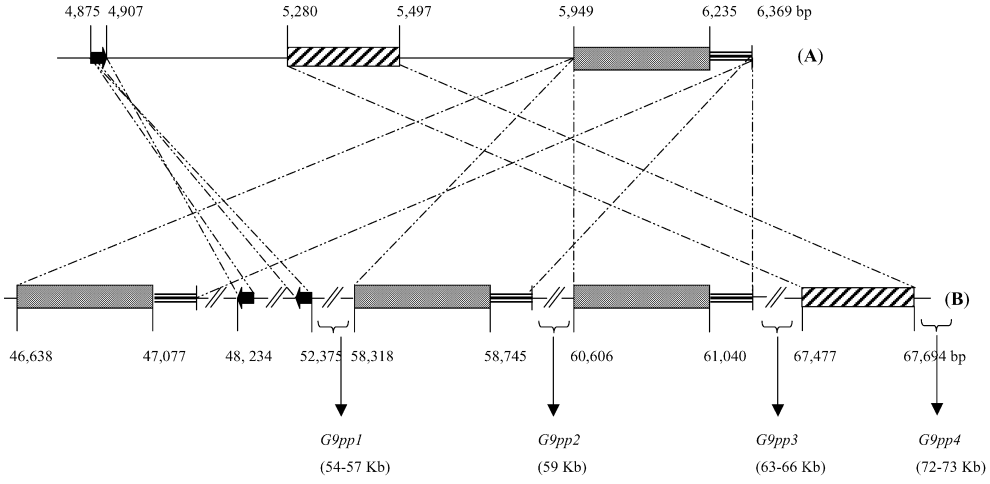

Fig. 3 Schematic structure and comparison of the *monkey* region from triploid banana (**A**) with the 3′end in MuG9 (**B**). *G9pp1–G9pp4* Polyprotein-like sequences in MuG9 (see also Table 5). *Shaded box* Radka8, Radka9 sequences (Valárik et al. 2002), *diagonally striped box monkey* retrotransposon sequence (Balint-Kurti et al. 2000), *large black arrow* 133-bp *monkey* flanking sequence

BAC clone MuG9

For MuG9 (73,268 bp), an overall G+C content of 38.5% was calculated. An overview of the main characteristics of this BAC clone is given in Table 1. BLASTN analysis against all GenBank+RefSeq Nucleotides+EMBL+DDBJ+PDB sequences was performed and showed high homology in three locations to the short repetitive sequences *Radka*8 (AF39948, E=e–142) and *Radka*9 (AF39938, E=e–116) (Valárik et al. 2002) (Fig. 2) as well as in one position to a 662-bp *M. acuminata* sequence encoding a partial putative protein homologous to a *pol* gene product (Y10860, 85% homology, E=e–127). Both *Radka*8 and *Radka*9 are homologous to the 3′ non-coding part of a 4.0-kb *Eco*RI fragment of the *Ty3/gypsy* type *monkey* retrotransposon (Balint-Kurti et al. 2000). Indeed, highly homologous fragments to *monkey* were found in the same three regions where the *Radka* repetitive sequences were localized (Fig. 2). A closer analysis of the *monkey* sequence revealed that the region homologous to *Radka* was in fact banana sequence downstream from the actual *monkey* retrotransposon. In addition, another 133-bp sequence from the *monkey* flanking sequence was found to occur twice between the first two repeat-containing regions in MuG9 and constituted part of a 449-bp LTR (Fig. 3). The similarity with the LTR was 93%. Adjacent to the LTR were a putative primer binding site and a polypurine tract in the 5′ end and the 3′ end, respectively, which showed the structural features of a LTR retrotransposon. However, target site duplication flanking the LTR was absent, and the 3.7-kb sequence between the LTR contained open reading frames (ORFs) with no homology to known DNA or proteins. In contrast, retrotransposon-like sequences with

Table 4 Comparison of gene prediction programs in the sequenced *Musa* BAC MuG9

CDS[a]	DGT[b]	FSH[b]	GMK[b]	GSN[b]
1		**1,479–3,077 (–), 4 ex**	2,052–3,077 (–), 2 ex	
2	6,241–14,693 (–), 13 ex	6,180–6,608 (–), 2 ex	6,113–6,608 (–), 2 ex	3,449–6,608 (–), 2 ex
3		9,499–14,693 (–), 17 ex	**9,499–14,693 (–), 17 ex**	13,158–14,693 (–), 2 ex
4		15,688–16,060 (+), 2 ex	15,688–16,060 (+), 2 ex	
5	20,135–20,747 (–), 2 ex	**17,266–20,747 (–), 7 ex**	**17,266–20,747 (–), 7 ex**	20,011–20,747 (–), 4 ex
6		**24,902–25,748 (–), 3 ex**	**24,902–25,748 (–), 3 ex**	22,604–25,777 (–), 7 ex
7	33,667–36,891 (+), 2 ex 33,667–39,056 (+), 5 ex	**29,329–37,222 (+), 11 ex**	29,176–35,696 (+), 11 ex	29,329–32,387 (+), 2 ex
8	37,949–39,056 (+), 3 ex	**38,113–39,055 (+), 3 ex**	36,364–39,056 (+), 5/7 ex	37,949–39,056 (+), 3 ex
9				39,569–40,788 (–), 2 ex
10				41,590–42,016 (+), 2 ex
11	42,530–43,606 (+), 2 ex	**42,530–43,606 (+), 2 ex**	41,590–43,606 (+), 4 ex	42,737–43,606 (+), 2 ex
12	44,783–46,034 (–), 2 ex	44,783–46,034 (–), 2 ex		
13			48,669–49,979 (–), 6 ex	
14			50,006–50,519 (–), 3 ex	
15			50,806–51,641 (–), 3 ex	
16				52,353–53,794 (–), 3 ex
17	54,424–66,224 (+), 4 ex	54,688–57,380 (+), 4 ex	**54,424–57,695 (+), 8 ex**	
18		59,515–59,970 (+), 1 ex	**59,578–59,970 (+), 1 ex**	
19		62,730–66,224 (+), 5 ex	**62,798–66,224 (+), 6 ex**	62,798–63,072 (+), 2 ex
20				65,986–67,260 (+), 2 ex
21	71,915–73,002 (–), 2 ex	**71,915–73,002 (–), 3 ex**		
Total	7 or 8	14	15	12
Best:	**-**	**7**	**6**	**-**

[a] Coding sequences (CDS) are numbered starting from the 5′ end of the DNA sequence as deposited in GenBank
[b] Localization of predicted coding sequences on plus or minus strand is indicated in brackets followed by the number of exons predicted; the most correct coding sequences, i.e. the ones with the highest homology to hits in the database, are indicated in bold. Prediction programmes: DGT, DIGIT; FSH, FGENESH; GMK, GENEMARK.HMM; GSN, GENSCAN

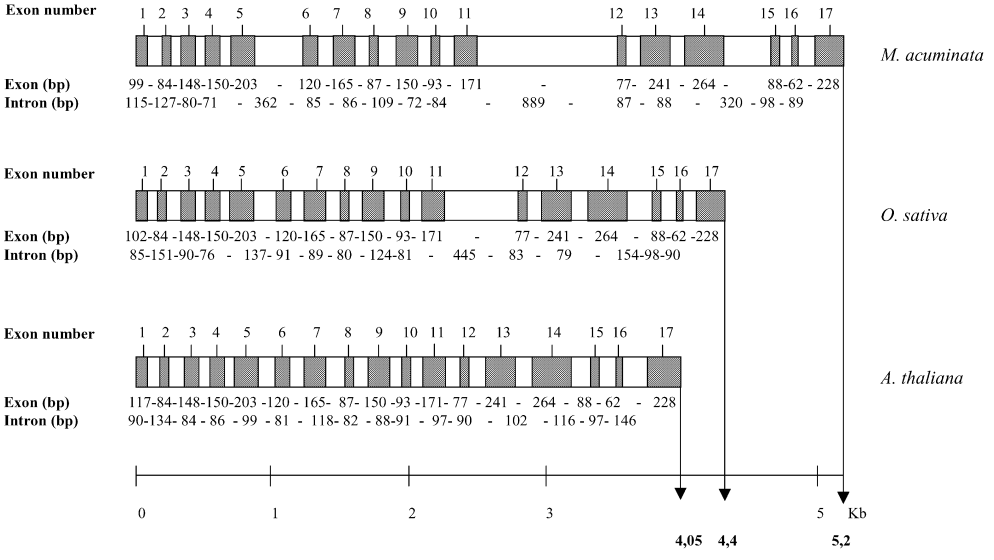

Fig. 4 Comparison of exon-intron structure for ribonucleotide reductase genes in *Musa acuminata* (AY484588, 810 amino acids), *Oryza sativa* (AB023482, 811 amino acids) and *Arabidopsis thaliana* (AC007019, 816 amino acids). *Shaded box* Exon, *white box* intron

no apparent LTR (*G9pp1–-G9pp4*, Table 4) were also found inserted around the three repeated regions (Fig. 3).

When DIGIT, FGENESH, GENEMARK.HMM and GENSCAN were used for modeling exon structure, a total of 21 candidates were identified by the four programmes (Table 4). FGENESH and GENEMARK.HMM predicted the presence of seven and six putative proteins, respectively, but the exon-intron structure frequently differed between the two programmes. GENEMARK also calculated the presence of three polyprotein-like sequences (*G9pp1–G9pp3*), whereas *G9pp4* was discovered by FGENESH and DIGIT only. Using GENSCAN with maize settings, we were only able to identify three CDSs (Table 4, Fig. 2). FGENESH and GENEMARK.HMM were the best programmes for predicting novel genes in BAC MuG9.

Following BLASTP analysis, one CDS (G9-1) showed extensive similarity (E=0) to a gene encoding the large subunit of the soybean ribonucleoside-diphosphate reductase (AF118784). The six remaining CDSs showed similarity only to hypothetical proteins in genome se-

Table 5 Predicted genes in the sequenced *Musa* BAC MuG9

Gene ID[a]	Number of exon(s)	Number of amino acids	InterPro domains	Putative function[b]; supporting evidence[c]
G9-1 (−); 1,479–3,077	4	112	DUF502 (IPR007462)	*Oryza sativa* hypothetical (membrane?) protein, AP003286 (116, 8e−26); rice EST C97538 (109, 8e-24); FSH, GMK
G9-2 (−); 9,499–14,693	17	810	Ribonucleotide reductase large subunit (IPR000788) (EC 1.17.4.1) ATP cone domain (IPR005144)	*Glycine max* ribonucleoside-diphosphate reductase large subunit, AF118784 (1506, E=0); rice EST C27938 (139, 2e-30); GMK, FSH, DGT
G9-3 (−); 17,266–20,747	7	382	Phosphatidylinositol-specific phospholipase C, X-domain (IPR000909), Serpin (IPR000215), Carbohydrate kinase (IPR 000577)	*Arabidopsis thaliana* hypothetical (MAP3K-like) protein kinase, AC011807, At1g49740 (409, E=e−113); FSH, GMK, DGT
G9-4 (−); 24,902–25,748	3	225	As above	*A. thaliana* hypothetical (MAP3K-like) protein kinase, At4g36950 (75, E=6e−13); FSH, GMK (GSN)
G9-5 (+); 29,329–37,222	11	617	Protein kinase (IPR000719), serine/threonine protein kinase (IPR002290) and active site (IPR008271), tyrosine protein kinase (IPR001245)	*O. sativa* hypothetical (serine/threonine-like) protein kinase, CAB53482 (503, E=e−141); FSH, GMK, DGT
G9-6 (+); 38,114–39,056	3	267	As above	*O. sativa* hypothetical (serine/threonine-like) protein kinase, CAB53482 (323, E=2e−87); FSH, GMK, DGT
G9-7 (+); 42,530–43,606	2	202	RNA-binding region RNP-1 (RNA recognition motif, RRM, IPR000504)	*O. sativa* hypothetical protein, CAB53480 (211, E=6e−54); *Musa* EST 600124625T1 (553, E=e−155); FSH, DGT, GMK
G9pp1 (+); 54,424–57,695	8	512	-	*O. australiensis* RIRE1 retrotransposon, BAA22288 (258, E=2e−67); *Musa* EST 600179507T1 (216, E=7e−54); GMK, FSH
G9pp2 (+); 59,578–59,970	1	130	-	*O. australiensis* RIRE1 retrotransposon, BAA22288 (196, E=7e−50); GMK, FSH
G9pp3 (+); 62,798–66,224	6	521	Zn-finger, CCHC type (IPR001878)	*O. sativa* retrotransposon-like protein, BAB63501 (232, E=2e−59); GMK, FSH, DGT
G9pp4 (−); 71,915–73,002	3	215	Zn-finger, CCHC type (IPR001878)	*O. sativa* putative Opie-2 retrotransposon, AC107224 (92, E=4e−18); FSH, DGT

[a] Genes are numbered starting from the 5′ end of the DNA sequence as deposited in GenBank. Localization of predicted gene on the plus or minus strand is indicated in brackets
[b] Protein function is based on BLASTP results with the score and E-value indicated in brackets
[c] Evidence supporting the existence of gene is as follows: EST, similar gene transcripts with GeneBank accession number and the BLASTN score and E-value in brackets; DGT, DIGIT; FSH, FGENESH; GMK, GENEMARK.HMM; GSN, GENSCAN (ordered according to the precision of prediction for each entry)

quence data. Functional domains of the predicted proteins were searched using INTERPRO and could be identified for the six putative genes (Table 5).

The exon-intron structures of ribonucleoside-diphosphate reductase genes from banana (G9-2), rice (AB023482) and *Arabidopsis* (AC007019) were compared as depicted in Fig. 4. Exon lengths were perfectly conserved in all three species, except for slight differences in the very last exon (banana: 99 bp; rice: 102 bp; *Arabidopsis*: 117 bp). The predicted protein displayed the highest homology to rice with 86% identity and 92% similarity. Positions and phases of all 16 introns in the gene were conserved among the three species, indicating that ribonucleoside-diphosphate reductase genes are highly conserved. However, individual introns were, on average, almost twice as long in banana as in *Arabidopsis*, with the difference less pronounced in comparison with rice: banana average = 173 bp; *Arabidopsis* average = 100 bp; rice average = 122 bp. The overall sequence homology of the banana gene was also higher to rice (BLASTN 281, E=9e−72) than to *Arabidopsis* (BLASTN 264, 2e−66).

Discussion

In the *M. acuminata* BAC MuH9 (82,723 bp), 12 putative CDSs were identified, which represents a gene density of one gene per 6.9 kb, similar to that of rice chromosome 4 (Feng et al. 2002), and, as expected, lower than that of *Arabidopsis* (The *Arabidopsis* Genome Initiative 2000). The function of several predicted proteins could be positively deduced: malate synthase (H9-1) and crinkly 4 (H9-6). Malate synthase (EC 4.1.3.2) is one of the key enzymes of the glyoxylate shunt (Huang et al. 1983) and catalyses the aldol condensation of glyoxylate with acetyl-CoA to form malate. The shunt enables the cell to generate increased levels of tricarboxylic acid cycle intermediates for biosynthetic pathways such as gluconeogenesis. The organic acids generated in this cycle are implicated in aluminium tolerance mechanisms for a range of plant species (Ma et al. 2001). Since tropical soils are characterized by high residual levels of aluminium, this banana gene could be an interesting candidate for further analysis.

The maize crinkly 4 (*cr4*) gene encodes a receptor protein kinase (RPK) that is critical for normal cell differentiation. RPKs are components of signal transduction pathways that elicit cellular responses to extracellular information. RPKs are essential for a variety of plant processes, including development (Becraft 1998), self-incompatibility and disease-resistance (Walker-Simmons 1998). Comparison of the *Musa* putative *cr4* gene with other genomic *cr4* sequences in the databases indicated no significant distinction between monocot and dicot genes. Both *Arabidopsis* and rice possess putative *cr4* genes that are similar to the maize *cr4*.

Analysis of *M. acuminata* BAC MuG9 (73,268 bp) revealed that this BAC contains seven CDSs that show homology to (hypothetical) proteins found in the databases (E less than e−20), resulting in a gene density of one gene per 10.5 kb. G9-2 showed highly significant homology to the rice ribonucleoside-diphosphate reductase large subunit. Ribonucleoside-diphosphate reductase (EC 1.17.4.1) is an essential enzyme in the DNA synthetic pathway (Reichard 1988). This single enzyme reduces four ribonucleotides to their corresponding deoxyribonucleotides. Searching the databases, we could only find mRNA sequences for the large subunit of *Glycine max* (AF118784) and *Nicotiana tabacum* (CAA71815). For *Arabidopsis*, both a cDNA (Sauge-Merle et al. 1999) and a genomic (AC007019) sequence were present. Searching the EMBL database, we found mRNA sequences for the large subunit of *Glycine max* (AF118784) and *Nicotiana tabacum* (CAA71815). For *A. thaliana*, both a cDNA (Sauge-Merle et al. 1999) and a genomic (AC007019) sequence were present. Gene prediction analysis of the genomic sequence showed that this sequence also consists of 17 exons. The lengths of the exons are similar to that of the *Musa* sequence (Fig. 4).

BLASTP analysis further revealed the presence of polyprotein-like sequences in MuG9. The CCHC Zn-finger-like domain of *G9pp3* and *G9pp4* is mainly found in the nucleocapsid protein of retroviruses (Jentoft and Katz 1989). It is also found in eukaryotic proteins involved in RNA-binding or single-stranded DNA-bind-

Table 6 Characteristics of gene regions in the sequenced banana BAC clones MuH9 and MuG9 and comparison with rice and *Arabidopsis* genome data (*ND* not determined)

BAC clone (length in bp)	MuH9 (82,723)	MuG9 (73,268)[a]	Average	Rice[b]	*Arabidopsis*[c]
Average gene length (bp)	3,251	3,301	3,275	2,414	1,968
Average protein length (no. of amino acids)	485	445	466	ND	426
Average no. of exons	5.0	7.8	6.3	4.2	5.2
Exon size: average/minimum/maximum (bp)	286/39/2,685	176/18/552	234/29/1,682	296 (average)	247/1/7,713
G+C content (%)	48.9	48.7	48.8	54.9	44.1
Intron size: average/minimum/maximum (bp)	473/74/3,191	277/68/1,875	381/71/2,572	371/49 (average/minimum)	164/1/6,442
G+C content (%)	38.5	39.3	38.9	38.9	32.7

[a] For MuG9, the polyprotein-like sequences (*G9pp1-G9pp4*) are not taken into account
[b] From the TIGR Rice Genome database (http://www.tigr.org/tdb/e2k1/osa1/riceInfo/info.shtml)
[c] From the MIPS website (http://www.mips.biochem.mpg.de/proj/thal/db/tables/chrall_tables/)

ing. The whole *monkey* sequence (Balint-Kurti et al. 2000) is derived from the triploid banana Grand Nain (AAA), whereas *Radka* was originally isolated from the diploid *M. acuminata* Pisang Mas (AA). Consequently, the homology of the common *Radka-monkey* region with MuG9 from another wild diploid banana (Calcutta 4, AA) points to a relationship between the A genomes of these three cultivars.

In both BACs there were distinct differences in the GC% content between exons and introns (Table 6) as evaluated on the basis of 103 exons and 85 introns: 48.9% versus 38.5% for MuH9 and 48.7% versus 39.3% for MuG9. Similar results were found for *Arabidopsis* (44.1% versus 32.7%; The *Arabidopsis* Genome Initiative 2000) and rice (55.5% versus 36.8%, Yu et al. 2002 or 54.9% versus 38.9%, the TIGR Rice Genome database: http://www.tigr.org). These data also suggest that in terms of GC content the banana genome resembles rice more than *Arabidopsis*.

The gene density of MuG9 (one gene per 10.5 kb) is strikingly lower than that calculated for MuH9 (one gene per 6.9 kb). However, a transition point between coding regions and repeated sequences was found at approximately 45 kb that separated the coding upstream BAC end from its downstream end that contained mainly transposon-like sequences and regions similar to known repetitive sequences of *M. acuminata*. Re-calculation of gene density taking the 45 kb into account results in a gene density of one gene per 6.4 kb, which is comparable with that for MuH9. These findings indicate that MuG9 might be located in a gene-empty region, which separates gene-rich areas, as has been described for the Gramineae (Barakat et al. 1997) but not for *Arabidopsis* (Barakat et al. 1998) and rice chromosome 4 (Feng 2002). In *Arabidopsis*, genes are fairly even distributed over regions, gene-empty regions are greatly reduced and repeat sequences seem to be dispersed, with no obvious clustering. Gramineae genomes, however, appear to comprise many large gene-empty regions (containing abundant transposons) separating gene clusters.

Acknowledgements The authors are grateful to Jaroslav Dolezel (Institute of Experimental Botany, Olomouc, Czech Republic) for providing a set of 96 BAC clones for this study and to Phillip SanMiguel (Purdue University, West Lafayette, USA) for initial help with retroelement analysis. Access to the Syngenta *Musa* EST database maintained at MIPS (Munich, Germany) is acknowledged.

References

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Balint-Kurti PJ, Clendennen SK, Doleželová M, Valárik M, Doležel J, Beetham PR, May GD (2000) Identification and chromosomal localization of the *monkey* retrotransposon in *Musa* sp. Mol Gen Genet 263:908–915

Barakat A, Carels N, Bernardi G (1997) The distribution of genes in the genomes of Gramineae. Proc Natl Acad Sci USA 94:6857–6861

Barakat A, Matassi G, Bernardi G (1998) Distribution of genes in the genome of *Arabidopsis thaliana* and its implications for the genome organization in plants. Proc Natl Acad Sci USA 95:10044–10049

Baurens FC, Noyer JL, Lanaud C, Lagoda P (1997) Copia-like elements in banana. J Genet Breed 51:135–142

Becraft PW (1998) Receptor kinases in plant development. Trends Plant Sci 3:384–388

Bevan M, Mayer K, White O, Eisen J, Preuss D, Bureau T, Salzberg S, Mewes H-M (2001) Sequence and analysis of the *Arabidopsis* genome. Curr Opin Plant Biol 4:105–110

Bodenteich A, Chissoe S, Wang YF, Roe BA (1993) Shot-gun cloning as the strategy of choice to generate templates for high-throughput dideoxynucleotide sequencing. In: Venter JC (ed) Automated DNA sequencing and analysis techniques, Academic Press, London, pp 42–50

Brooks SA, Huang L, Gill BS, Fellers JP (2002) Analysis of 106 kb of contiguous DNA sequence from the D genome of wheat reveals high gene density and a complex arrangement of genes related to disease resistance. Genome 45:963–972

Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA. J Mol Biol 268:78–94

Denham TP, Haberle SG, Lentfer C, Fullagar R, Field J, Therin M, Porch N, Winsborough B (2003) Origins of agriculture at Kuk Swamp in the highlands of New Guinea. Science 301:189–193

Feng Q, Zhang Y, Hao P, Wang S, Fu G, Huang Y, Li Y, Zhu J, Liu Y, Hu X, Jia P, Zhang Y, Zhao Q, Ying K, Yu S, Tang Y, Weng Q, Zhang L, Lu Y, Mu J, Lu Y, Zhang LS, Yu Z, Fan D, Liu X, Lu T, Li C, Wu Y, Sun T, Lei H, Li T, Yin HH, Cai Z, Ren S, Lu G, Gu W, Zhu G, Tu Y, Jia J, Zhang Y, Chen J, Kang H, Chen X, Shao C, Sun Y, Hu Q, Zhang X, Zhang W, Wang L, Ding C, Sheng H, Gu J, Chen S, Ni L, Zhu F, Chen W, Lan L, Lai Y, Cheng Z, Gu M, Jiang J, Li J, Hong G, Xue Y, Han B (2002) Sequence and analysis of rice chromosome 4. Nature 420:316–319

Fu HH, Park WK, Yan XH, Zheng ZW, Shen BZ, Dooner HK (2001) The highly recombinogenic *bz* locus lies in an unusually gene-rich region of the maize genome. Proc Natl Acad Sci USA 98:8903–8908

Gewolb J (2001) DNA sequencers to go Banana's? Science 293:585–586

Gish W, States DJ (1993) Identification of protein coding regions by database similarity search. Nat Genet 3:266–272

Gowen S (1995) Bananas and Plantains. Chapman and Hall, London

Harper G, Osuji JO, Heslop-Harrison JS, Hull R (1999) Integration of banana streak badnavirus into the *Musa* genome: molecular and cytogenetic evidence. Virology 255:207–213

Huang AHC, Trelease RN, Moore TS (1983) Plant peroxisomes. Academic Press, New York

Jentoft JE, Katz RA (1989) What is the role of the cys-his motif in retroviral nucleocapsid (NC) proteins? Bioessays 11:176–181

Jurka J (2000) Repbase update: a database and an electronic journal of repetitive elements. Trends Genet 9:418–420

Jurka J, Klonowski P, Dagman V, Pelton P (1996) CENSOR— a program for identification and elimination of repetitive elements from DNA sequences. Comput Chem 20:119–122

Keller B, Feuillet C (2000) Colinearity and gene density in grasses. Trends Plant Sci 5:246–251

Liu H, Sachidanandam R, Stein L (2001) Comparative genomics between rice and Arabidopsis shows scant co linearity in gene order. Genome Res 11:2020–2026

Lowe TM, Eddy SR (1997) TRNASCAN-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25:955–964

Lukashin AV, Borodovsky M (1998) GENEMARK.HMM: new solutions for gene finding. Nucleic Acids Res 26:1107–1115

Lysák M, Doleželova M, Horry JP, Swennen R, Doležel J (1999) Flow cytometric analysis of nuclear DNA content in *Musa*. Theor Appl Genet 98:1344–1350

Ma JF, Ryan PR, Delhaize E (2001) Aluminium tolerance in plants and the complexing role of organic acids. Trends Plant Sci 6:273–278

Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Barrell D, Bateman A, Binns D, Biswas M, Bradley P, Bork P, Bucher P, Copley RR, Courcelle E, Das U, Durbin R, Falquet L, Fleischmann W, Griffiths-Jones S, Haft D, Harte N, Hulo N, Kahn D, Kanapin A, Krestyaninova M, Lopez R, Letunic I, Lonsdale D, Silventoinen V, Orchard SE, Pagni M, Peyruc D, Ponting CP, Selengut JD, Servant F, Sigrist CJA, Vaughan R, Zdobnov EM (2003) The INTERPRO database 2003 brings increased coverage and new features. Nucleic Acids Res 31:315–318

Ndowora T, Dahal G, LaFleur D, Harper G, Hull R, Olszewski NE, Lockhart B (1999) Evidence that badnavirus infection in *Musa* can originate from integrated pararetroviral sequences. Virology 255:214–220

Parniske M, Hammond-Kosack KE, Golstein C, Thomas CM, Jones DA, Harrison K, Wulff BBH, Jones JDG (1997) Novel disease resistance specificities result from sequence exchange between tandemly repeated genes at the *Cf-4/9* locus of tomato. Cell 91:821–832

Pua EC, Chandramouli S, Han P, Liu P (2003) Malate synthase gene expression during fruit ripening of Cavendish banana (*Musa acuminata* cv. Williams). J Exp Bot 54:309–316

Ramakrishna W, Emberton J, SanMiguel P, Ogden M, Llaca V, Messing J, Bennetzen JL (2002) Comparative sequence analysis of the sorghum *rph* region and the maize *rp1* resistance gene complex. Plant Physiol 130:1728–1738

Reichard P (1988) Interactions between deoxyribonucleotide and DNA synthesis. Annu Rev Biochem 57:349–374

Robinson JC (1996) Bananas and plantains. CAB Int, Wallingford

Roe BA, Crabtree JS, Khan AS (1996) DNA isolation and sequencing. John Wiley and Sons, New York

Salamov A, Solovyev V (2000) Ab initio gene finding in Drosophila genomic DNA. Genome Res10:516–522

Salse J, Piegu B, Cooke R, Delseny M (2002) Synteny between *Arabidopsis thaliana* and rice at the genome level: a tool to identify conservation in the ongoing rice genome sequencing project. Nucleic Acids Res 30:2316–2328

Sasaki T, Matsumoto T, Yamamoto K, Sakata K, Baba T, Katayose Y, Wu J, Niimura Y, Cheng Z, Nagamura Y, Antonio BA, Kanamori H, Hosokawa S, Masukawa M, Arikawa K, Chiden Y, Hayashi M, Okamoto M, Ando T, Aoki H, Arita K, Hamada M, Harada C, Hijishita S, Honda M, Ichikawa Y, Idonuma A, Iijima M, Ikeda M, Ikeno M, Ito S, Ito T, Ito Y, Ito Y, Iwabuchi A, Kamiya K, Karasawa W, Katagiri S, Kikuta A, Kobayashi N, Kono I, Machita K, Maehara T, Mizuno H, Mizubayashi T, Mukai Y, Nagasaki H, Nakashima M, Nakama Y, Nakamichi Y, Nakamura M, Namiki N, Negishi M, Ohta I, Ono N, Saji S, Sakai K, Shibata M, Shimokawa T, Shomura A, Song J, Takazaki Y, Terasawa K, Tsuji K, Waki K, Yamagata H, Yamane H, Yoshiki S, Yoshihara R, Yukawa K, Zhong H, Iwama H, Endo T, Ito H, Ho Hahn J, Kim HI, Eun MY, Yano M, Jiang J, Gojobori T (2002) The genome sequence and structure of rice chromosome 1. Nature 420:312–315

Sauge-Merle S, Falcone S, Fontecave M (1999) An active ribonucleotide reductase from *Arabidopsis thaliana*. Eur J Biochem 266:62–69

Simmonds NW (1966) Bananas, 2nd edn. Longmans, London

Simons G, Groenendijk J, Wijbrandi J, Reijans M, Groenen J, Diergaarde P, Van der Lee T, Bleeker M, Onstenk J, de Both M, Haring M, Mes J, Cornelissen B, Zabeau M, Vos P (1998) Dissection of the *Fusarium I2* gene cluster in tomato reveals six homologs and one active gene copy. Plant Cell 10:1055–1068

Teo CH, Tan SH, Othman YR, Schwarzacher T (2002) The cloning of Ty1-*copia*-like retrotransposons from 10 varieties of banana (*Musa* sp.). J Biochem Mol Biol Biophys 6:193–201

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796–815

The Rice Chromosome 10 Sequencing Consortium (2003) In-depth view of structure, activity and evolution of rice chromosome 10. Science 300:1566–1569

Valárik M, Simková H, Hribová E, Safár J, Doleželová M, Doležel J (2002) Isolation, characterization and chromosome localization of repetitive DNA sequences in banana (*Musa* spp.). Chromos Res 10:89–100

Walbot V, Petrov DA (2001) Gene galaxies in the maize genome. Proc Natl Acad Sci USA 98:8163–8164

Walker-Simmons MK (1998) Protein kinases in seeds. Seed Sci Res 8:193–200

Webb CA, Richter TE, Collins NC, Nicolas M, Trick HN, Pryor T, Hulbert SH (2002) Genetic and molecular characterization of the maize *rp3* rust resistance locus. Genetics 162:381–394

Yada T, Takagi T, Totoki Y, Sakaki Y, Takaeda Y (2003) DIGIT: a novel gene finding program by combining gene-finders. Pac Symp Biocomput 2003:375–387

Yu J, Hu S, Wang J, Wong G, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X, Cao M, Liu J, Sun J, Tang J, Chen Y, Huang X, Lin W, Ye C, Tong W, Cong L, Geng J, Han Y, Li L, Li W, Hu G, Huang X, Li W, Li J, Liu Z, Li L, Liu J, Qi Q, Liu J, Li L, Li T, Wang X, Lu H, Wu T, Zhu M, Ni P, Han H, Dong W, Ren X, Feng X, Cui P, Li X, Wang H, Xu X, Zhai W, Xu Z, Zhang J, He S, Zhang J, Xu J, Zhang K, Zheng X, Dong J, Zeng W, Tao L, Ye J, Tan J, Ren X, Chen X, He J, Liu D, Tian W, Tian C, Xia H, Bao Q, Li G, Gao H, Cao T, Wang J, Zhao W, Li P, Chen W, Wang X, Zhang Y, Hu J, Wang J, Liu S, Yang J, Zhang G, Xiong Y, Li Z, Mao L, Zhou C, Zhu Z, Chen R, Hao B, Zheng W, Chen S, Guo W, Li G, Liu S, Tao M, Wang J, Zhu L, Yuan L, Yang H (2002) A draft sequence of the Rice genome (*Oryza sativa* L. ssp *indica*). Science 296:79–92