Tree Genetics & Genomes
DOI 10.1007/s11295-011-0385-3

ORIGINAL PAPER

Genome-wide BAC-end sequencing of *Musa acuminata* DH Pahang reveals further insights into the genome organization of banana

Rafael E. Arango · Roberto C. Togawa · Sebastien C. Carpentier · Nicolas Roux · Bas L. Hekkert · Gert H. J. Kema · Manoel T. Souza Jr.

Received: 26 February 2010 / Revised: 2 February 2011 / Accepted: 7 March 2011

Abstract Banana and plantain (*Musa* spp.) are grown in more than 120 countries in tropical and subtropical regions and constitute an important staple food for millions of people. A *Musa acuminata* ssp. *malaccencis* DH Pahang bacterial artificial chromosome (BAC) library (MAMB) was submitted for BAC-end sequencing. MAMB consists of 23,040 clones, with a 140-kbp average insert size, accounting for a five times coverage of the banana genome. A total of 46,080 reads were generated, and 42,750 (92.8%) high-quality sequences were obtained after trimming for vector and quality. Analysis of these data shows a GC content of 41.39%, whereas interspersed repeats comprise 32.3%. The most common repeated sequences found show homology to ribosomal RNA genes, particularly 18S rRNA, while the Ty3/gypsy type monkey retrotransposon is the most common retro element. The sequence data were used to generate a banana-specific repeat library containing 54 new repetitive elements which accounted for 11.86% of the total nucleotides. Simple sequence repeats represent 0.7% of the sequence data and allowed the identification of 2,455 potentially useful marker sites. Functional annotation identified 2,705 sequences that could code for proteins of known function. Microsynteny analysis shows a higher number of co-linear matches to *Oryza sativa*, in contrast to *Arabidopsis thaliana*. This database of BAC-end sequences is useful for the assembly of the complete banana genome

Communicated by F. Gmitter

Electronic supplementary material The online version of this article (doi:10.1007/s11295-011-0385-3) contains supplementary material, which is available to authorized users.

R. E. Arango
Unidad de Biotecnología Vegetal UNALMED-CIB,
Corporación para Investigaciones Biológicas (CIB),
Carrera 72 A No. 78 B-141,
Medellín, Colombia

R. E. Arango
Escuela de Biociencias, Facultad de Ciencias,
Universidad Nacional,
Carrera 64 Calle 65,
Medellín, Colombia

R. C. Togawa
Embrapa Genetic Resources & Biotechnology,
CP 02372, CEP 70770-900, Brasília, Federal District, Brazil

S. C. Carpentier
Division of Crop Biosystems, K.U.Leuven,
3001 Leuven, Belgium

N. Roux
Global Musa Genomics Consortium, Bioversity International,
Montpellier, France

B. L. Hekkert · G. H. J. Kema
Plant Research International,
6708 PB, Wageningen, The Netherlands

M. T. Souza Jr. (✉)
Embrapa LABEX Europe,
6708 PB, Wageningen, The Netherlands
e-mail: manoel.souza@embrapa.br

Published online: 12 April 2011

sequence and is important for identification in functional genomics experiments.

Keywords BAC library · *Musa acuminata* ssp. *malaccencis* · GMGC · Genomics

Introduction

Banana and plantain (*Musa* spp.) belong to the family *Musaceae*. Most of the edible *Musa* spp. cultivated and highly appreciated worldwide nowadays are derived from the several subspecies of *Musa acuminata* and from hybrids with *Musa balbisiana*. These edible bananas have a basic chromosome number set of 11 chromosomes and are mostly triploid, although some diploid and tetraploid varieties are also known and used in several countries (Osuji et al. 1997). The A and B genomes of *Musa* differ in size, according to estimates carried out by flow cytometry of nuclei stained by propidium iodide (Lysak et al. 1999); the average haploid genome size of *M. balbisiana* is 537 Mbp, while *M. acuminata* showed statistically significant variants among subspecies and clones, ranging from 591 to 615 Mbp. Thus, the B genome is only 15% larger than *Oryza sativa* L. ssp. *indica* (Yu et al. 2002), while the A genome is about 30% larger.

According to the FAO (http://faostat.fao.org/), in 2008, bananas and plantains were produced in 130 and 49 countries, respectively. The total area harvested and the total production quantity were, respectively, 4,817,551 ha and 90,705,922 tonnes for bananas and 5,390,731 ha and 34,343,343 tonnes for plantains. Only 20% of the banana and 10% of the plantains, produced worldwide in 2006, entered the international market, generating about US $6 billion in export value (97% banana and 3% plantain). The remainder was traded on the internal markets of the producing countries, supporting the notion that besides being an important commodity for the economy of many (developing) countries, bananas and plantains have a major social and economic role in the tropics and sub-tropics, being a staple food for millions of people in Africa.

The global export banana production relies on closely genetically related clones of the Cavendish sub-group (AAA), which are known as sterile triploids. This subgroup of cultivars is extremely vulnerable to pests and disease and extremely difficult to improve by classical plant breeding. Diseases are a major constraint to banana production in most of the growing regions. Among the different diseases occurring in banana and plantain, Black leaf streak, caused by *Mycosphaerella fijiensis*, and Fusarium wilt, caused by *Fusarium oxysporum* f. sp. *cubense*, stand out for their economic importance and its destructiveness, respectively (Marin et al. 2003).

The construction of a bacterial artificial chromosome (BAC) library, together with the sequencing of the BAC ends (BAC end sequencing, BES), constitute two of the first steps in the process to perform a whole-genome project (WGP). Although the sequences sampled are not truly random, due to the requirement for a specific restriction enzyme site for the construction of a BAC library, a BES project can provide significant clues about the genome of a species. The use of data from such a project in whole-genome characterization was first proposed as a strategy for identifying overlapping clones during whole-genome sequencing projects (Venter et al. 1996) since paired end sequencing greatly facilitates the assembly of contigs into scaffolds (Goff et al. 2002). BAC-end sequences provide a sampling of the genome that can be used to estimate not only gene content but also the content of repetitive sequences, simple sequence repeats (SSR), and other classes of repeats (Hong et al. 2006; Lai et al. 2006; Paux et al. 2006).

In order to gain insight into the genome of the double haploid of the *M. acuminata* ssp. *malaccensis* var. Pahang (DH Pahang), which is the genotype being used in the WGP of banana, as well as to help in the assembly of this whole genome, we have performed a complete BES project of the DH Pahang BAC library. The data generated in this BES project and the insights obtained are here described.

Material and methods

Plant material, BAC library production, and BAC-end sequencing

A double haploid of the *M. acuminata* ssp. *malaccensis* var. Pahang was developed by CIRAD (www.cirad.fr). This DH Pahang genotype was then made available to the Global *Musa* Genomics Consortium (www.musagenomics.org) for the production of a BAC library at Amplicon Express (http://www.genomex.com/). This BAC library, denominated MAMB, was then made available to PRI-WUR and Embrapa for the BES project here reported. DH Pahang is the genotype being used in the banana Whole Genome Project (www.genoscope.cns.fr/spip/September-8th-2009-Banana-genome.html) recently approved by the French National Research Agency (L'Agence nationale de la recherche) to be sequenced by CIRAD and Genoscope (www.genoscope.cns.fr). For additional information on MAMB, please see at http://mgrc.musagenomics.org/display.php?page=baclibrary. BAC-end sequencing was performed at the Genome Sequencing Center at the HudsonAlpha Institute for Biotechnology, Huntsville, AL, USA (http://www.hudsonalpha.org/).

Sequence assembling and masking

The raw data were analyzed using the EGassembler web server (http://egassembler.hgc.jp/), which provides an automated as well as a user-customized analysis tool for cleaning, repeat masking, vector trimming, organelle masking, clustering, and assembling of genomic fragments (Masoudi-Nejad et al. 2006). The consensus sequences of contigs and singletons obtained as a result of the assembly were used for identification of repeat content, SSR analysis, functional annotation, and protein identification. For microsynteny, non-assembled BAC-end sequences were used, excluding those having known repeats.

Repeat content

Repeats were identified through similarity searches using RepeatMasker Version open-3.2.8 (http://www.repeatmasker.org/) with default parameters. Several databases were used for comparison including RepBase Update 20090604, TIGR_Oryza_GSS_Repeats, the EGrep repeats library (http://egassembler.hgc.jp/), and a custom-made repeat library containing several published banana-specific repeats. The repeat density was then calculated as the percentage of nucleotides in the BAC-end sequences that had one or more hits to the repeat database. Classification of repeat families was derived from the annotation in the RepBase database. In order to find new banana, specific repetitive elements BES were analyzed with RepeatModeler (http://www.repeatmasker.org/RepeatModeler.html) using default parameters. The repetitive elements generated by RepeatModeler were further filtered to remove published banana repeats. The type of repeat was further classified using TEclass (Abrusan et al. 2009).

Analysis of simple sequence repeats

Microsatellites were detected using a modified version of the Sputnik software (http://wheat.pw.usda.gov/ITMI/EST-SSR/LaRota/). Running parameters were set to return all SSRs spanning at least 10 nucleotides, with a motif length between 1 and 5 (i.e., mono-, di-, tri-, tetra-, and penta-nucleotide repeats) and a minimum score of 5. Microsatellites identified in this manner were divided into two classes: class I, which has 10 or more motif repeats, and class II, which has fewer than 10 motif repeats (Shultz et al. 2007). For comparison, the BAC-end sequences of Calcutta 4 (Cheung and Town 2007) were directly downloaded and screened for SSRs. Additionally, BAC-end sequences were submitted to analysis in a bioinformatics pipeline for identification of potential primers useful for SSR detection. This pipeline consisted of a primer development module which chopped genomic sequences into 1,500-bp size fragments with a 100-bp overlap, then detects SSR using a Perl script allowing imperfect repeat SSR detection by using the scan-for-matches program (Dsouza et al. 1997). Finally, it generates a maximum of five primer sets by using Primer3 software (Rozen and Skaletsky 2000) that could amplify a SSR by PCR followed by a screening for unique primer pairs.

Functional annotation

After the assembly process, the obtained contigs and singletons were collected into one file for further processing. Sequence comparison using Blastx (Altschul et al. 1997) was performed locally. The cutoff *E* value of $<10^{-5}$ was used to define the similar orthologs, and the sequences set that did not meet this requirement were annotated as unknown. Several databases were used in the annotation step: GenBank nr (Benson et al. 2002), MIPS *Arabidopsis thaliana* (Schoof et al. 2004), and SwissProt (Gasteiger et al. 2001). Predicted protein sequences were aligned with Blastx against eukaryotic orthologous groups (KOG; Tatusov et al. 2003) and Gene Ontology (Ashburner et al. 2000).

Protein identification based on proteomics data

In addition to the nucleotide-based sequence annotation, the identification of protein coding regions based on homology to a mass spectrometry derived-database of banana peptide sequences was also done. The procedure from protein isolation to protein identification was essentially performed as described previously (Carpentier et al. 2007). Briefly, a sequence query was performed with the following parameters: enzyme, trypsin; fixed modifications, carbamidomethyl (*C*); variable modifications, deamidated (NQ), oxidation (*M*), mass values, monoisotopic, protein mass, unrestricted; peptide mass tolerance, ±40 ppm; fragment mass tolerance, ±0.2 Da; max missed cleavages, 0; and instrument type, matrix-assisted laser desorption/ionization tandem time of flight. Only protein scores greater than 58 are significant ($p<0.05$). Protein scores are derived from ions scores as a non-probabilistic basis for ranking protein hits.

Genome comparison

Unassembled high-quality sequences not containing known repetitive sequences were used to determine potential areas of microsynteny between the banana genome and the genomes of *Arabidopsis thaliana*, *Oryza sativa*, *Vitis vinifera*, and *Brachypodium distachyon* using BLASTN alignments. For each sequence, only the best match to the respective genome sequence with an *E* value lower than 10^{-10} was evaluated. A BAC clone was considered to have microsynteny to the target genome if both ends mapped within a distance of between 50 and 500 kb of one another.

Results and discussion

BAC-end sequencing and assembly

A total of 46,080 reads were generated. After trimming for vector and quality, 42,516 high-quality sequences were obtained. The average length of the high-quality sequences was 700 nucleotides, with a GC content of 41.39 %, giving a total base count of about 30 Mb, or about 5% equivalent of the whole genome (Table 1). These data have approximately seven times more coverage in terms of number of high-quality sequences used and total base count, in comparison with the only BES study done in banana so far Cheung and Town (2007). In addition, the present study has used the same genotype that is being used in the whole-genome sequencing project which is underway. Sequence assembly was done after vector and quality trimming and repeat and organelle masking, giving a total of 5,784 contigs and 18,573 singletons, which were submitted to GenBank dbGSS (dbGSS _Id: from 28987538 to 29011891).

Interspersed repeats

Similarity-based searches on existing repeat databases were able to find a repeat content in MAMB of 20.41%. Sequences similar to 18 rRNA gene accounted for 7.41%, being the most common repeat found in this *Musa* accession. In contrast, analysis of the Calcutta 4 BAC-end sequences (Cheung and Town 2007) downloaded from the NCBI database showed very little sequences with similarity to ribosomal RNA genes. Other ribosomal RNA-related sequences, such as Radka2 5S banana rRNA and LSU_rRNA_Ath (Bartos et al. 2005), were more abundant in DH Pahang and almost absent in Calcutta 4.

Apart from the ribosomal RNA repeats, the most common element found in both DH Pahang and Calcutta 4 was the Ty3/gypsy type monkey retrotransposon (Balint-Kurti et al. 2000), which accounted for 2.5% of total sequence length in DH Pahang (Fig. 1). This result is consistent with previous reports showing the high abundance of the monkey retrotransposon in banana (Balint-Kurti et al. 2000; Cheung and Town 2007) and supports the notion that class I retrotransposons are contributing to a large portion of the genome in several plant species (Cheung and Town 2007; Lai et al. 2006; SanMiguel et al. 1998). Very recently, in a study of Calcutta 4 repetitive elements using 454 sequencing technology, Hribova et al. (2010) described that the most abundant retrotransposon found was belonged to Ty1/copia type instead of Ty3/gypsy as described in the present study as well as in previous reports (Balint-Kurti et al. 2000; Cheung and Town 2007). As discussed by Hribova et al., the difference in these results might be due to different chromosomal distributions of the repetitive elements (Hribova et al. 2010).

Table 1 Description of BAC-end sequences used in this study

	Calcutta 4[a]	DH Pahang
Total # sequences	6,252	42,516
Total base count (bp)	4,420,944	29,756,005
Average length (bp)	707	700
GC content (%)	39.25	41.39
# of contigs	370	5,784
# of singletons	4,394	18,573

[a] Cheung and Town (2007)

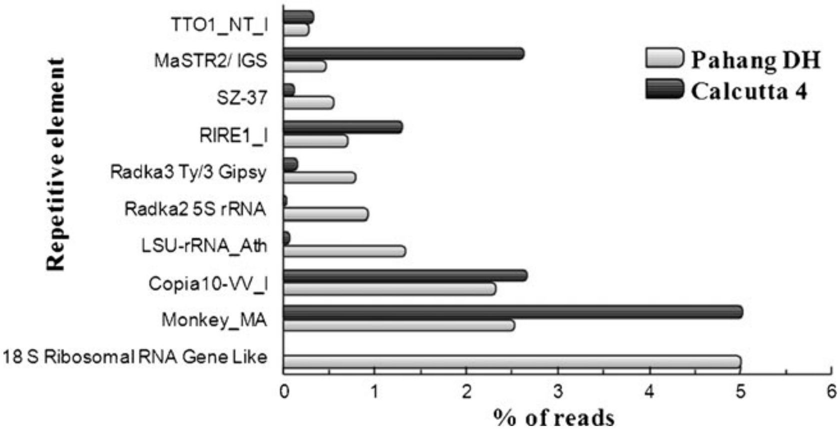
In spite of the fact that in the above-mentioned analysis we used a custom-made repeat database that included all published banana-specific repeats (Balint-Kurti et al. 2000; Bartos et al. 2005; Hribova et al. 2007; Valarik et al. 2002), the estimated size of the repetitive fraction of the banana genome was most likely underestimated. As a base for comparison, rice (*O. sativa*) has a smaller genome than banana, 1C=490 Mbp (Bennett and Smith 1991), and about 35% corresponds to repeated sequences (Vij et al. 2006). Assuming that banana has approximately the same number of genes in a 600-Mb genome, it could have up to 40% of repetitive sequences. In order to prove this hypothesis, we searched for de novo repeats using RepeatModeler (http://www.repeatmasker.org/RepeatModeler.html). This analysis allowed us to generate a library of 54 new banana-specific repetitive elements. Of the 54 new repetitive elements found, 10 were classified as DNA transposons, 25 as long terminal repeats, 5 as long interspersed DNA elements, and 14 as unknown. The newly found banana repeats accounted to 11.86% of the BES elevating the total percentage of repetitive elements to 32.3%.

The repeat library available as a result of this BES project can be the starting point for the design of a banana-specific repeat database. Such database will be very helpful in the process of assembling the entire genome sequence of banana. The importance of a profound analysis of banana-specific repeats is further emphasized by the fact that the known differences in genome sizes between the A and B genomes (Lysak et al. 1999) can be most probably attributed to repetitive sequences (Valarik et al. 2002).

Simple sequence repeats

SSRs occupied 225,195 (0.75%) nucleotides of de DH Pahang BAC-end sequences. Approximately 26% in Pahang DH and 24% in Calcutta 4 of the SSRs found belonged to class I which are considered to be hypervariable and thus more useful for designing primers for

Fig. 1 Distribution of the most abundant interspersed repeats in the DH Pahang and Calcutta 4 BAC-end sequences. The *Y*-axis shows the type of repeat and the *X*-axis shows the percentage of nucleotides occupied by that repeat



molecular markers. The most abundant SSR was (AT)n followed by (TC)n and (AAG)n. No significant differences were found in Calcutta 4 (Cheung and Town 2007; Fig. 2). The relative amounts of SSRs found in banana are similar to some other plant species such as rice (Huo et al. 2008). However, the most abundant type of SSRs seems to vary depending on the plant species analyzed. For example, AT dinucleotide repeats are more common in rice and banana, whereas AT-rich trinucleotides are much more abundant in soybean, papaya, and wheat (Huo et al. 2008).

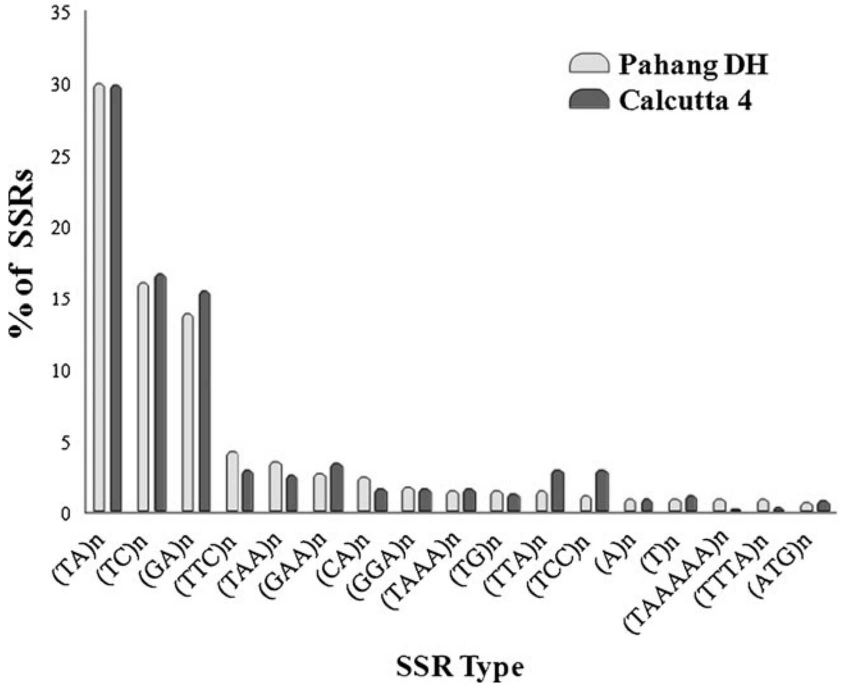
In order to scan for potentially useful SSR markers of size between 24 and 60 nucleotides, a perl script based on Scan_for_matches (Ross Overbeek and Mark D'Souza Argonne National Laboratory, USA) and Primer3 (Rozen and Skaletsky 2000) was used. A total of 2,455 sites were identified, and a list of primers was generated and is available as supplementary information.

Functional annotation

In order to predict functions for the proteins encoded by the BAC-end sequences, a search with all the masked contigs was carried out in the KOG database (Tatusov et al. 2003), which contain groups of orthologous proteins of seven eukaryotic genomes. This comparison resulted in 21,652 sequences with no evident homologies with sequences in the database and 2,705 homologous to proteins with known functions. These latter were categorized into 22 functional classes, the most common being signal transduction mechanisms with 13.01%, general function prediction with 10.35%, non-conclusive with 10.24%, and posttranslational modification, protein turnover, chaperones with 9.83% (Fig. 3).

The success rate of functional high-throughput studies like transcriptomics and proteomics are highly dependent

Fig. 2 Distribution of the most abundant SSR motifs in the DH Pahang and Calcutta 4 BAC-end sequences. The values on the *Y*-axis represent the percentage of SSRs present in each library. The *X*-axis shows the 17 most abundant SSRs



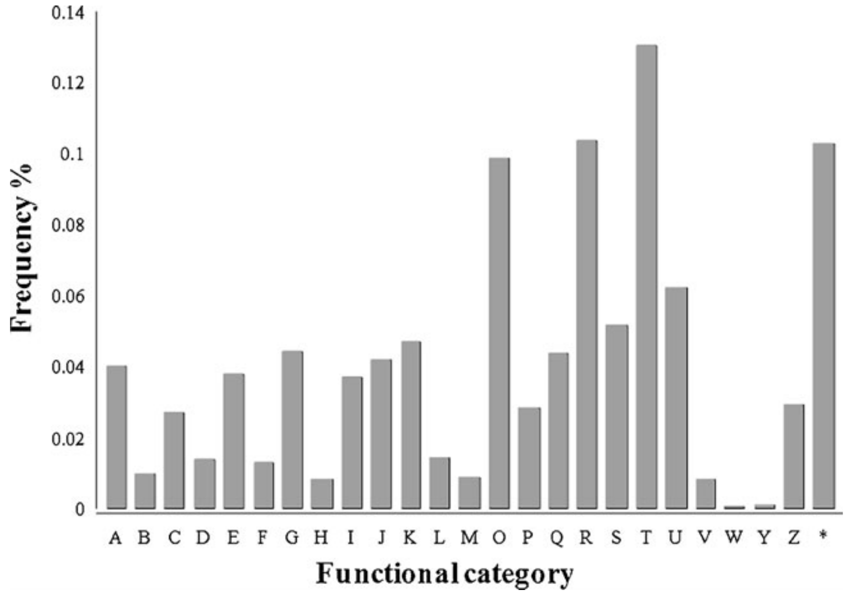


Fig. 3 Functional classification and class frequency of DH Pahang sequences according to eukaryotic clusters of orthologs (KOG). Designations of functional categories: *A* RNA processing and modification, *B* chromatin structure and dynamics, *C* energy production and conversion, *D* cell-cycle control and mitosis, *E* amino acid metabolism and transport, *F* nucleotide metabolism and transport, *G* carbohydrate metabolism and transport, *H* coenzyme metabolism, *I* lipid metabolism, *J* translation, *K* transcription, *L* replication and repair, *O* posttranslational modification, protein turnover, chaperone functions, *P* inorganic ion transport and metabolism, *Q* secondary metabolites biosynthesis, transport, and catabolism, *R* general functional prediction only, *S* function unknown, *T* signal transduction, *U* intracellular trafficking and secretion, *Z* cytoskeleton, *asterisk* non-conclusive

on the presence of databases (Carpentier et al. 2008a). The identification of mRNA fragments in a transcriptomic approach or the identification of peptides in a proteomics approach of a non-sequenced organism like banana is greatly dependent on EST databases. Unfortunately, EST libraries only contain a fraction of the mRNA and its corresponding protein, which complicates the identification. In order to test the use of the library for proteomics and check some of the BAC predicted proteins, a search was performed with existing proteome data. As a proof of principle, the spectra of 1,000 proteins were searched against the total of 5,784 contigs. In this fraction of the meristem proteome, 56 proteins originating from 33 different contigs could be confidently identified covering up to eight different peptides per protein (Supplementary Table 1A). A BLAST of those 5,784 contigs against NCBI nr resulted in 1,267 potential hits (*E* value lower than 10^{-5}). Nine proteins were identified with the proteomics approach but not via BLASTx (Supplementary Table 1B). From a proteomics viewpoint, the extensive protein coverage is a great advantage and opens possibilities to perform a shotgun approach for non-sequenced organisms (Carpentier et al. 2008b). The advantage of this additional proteomic annotation is that we can confirm the BLAST predicted proteins with real protein data and even detect proteins that have a too low homology with proteins present in the nr database (negative BLAST result). A proteomics confirmation also helps in the functional annotation of novel genes.

Genome comparison

Comparison of banana BAC-end sequences with whole-genome sequence of *A. thaliana*, *O. sativa*, *V. vinifera*, and *B. distachyon* was done in order to find regions of microsynteny with these genomes. For this purpose, only 21,580 BAC-end sequences that did not contain known repetitive regions were used. Blast hits were divided into three categories as is shown in Table 2. The unpaired category was composed of BAC-end sequences pairs from which only one of the two sequences had a match to the target genome. Gapped matches contained pairs that matched to the same chromosome with a distance between the paired matches that was either smaller than 50 kb or larger than 500 kb. Collinear matches contained pairs that matched to the same chromosome with a distance between 50 and 500 kb and where considered to represent regions of

Table 2 Number and type of BLASTN matches of paired BAC-end sequences with four different plant genomes

	Collinear matches	Gapped matches	Unpaired matches
Oryza sativa	156	376	6,977
Vitis vinifera	75	0	2,971
Arabidopsis thaliana	0	0	351
Brachypodium distachyon	0	0	571

microsynteny. Rice was by far the species with the biggest number of matches in all categories with 156 collinear matches. In contrast, *A. thaliana* or *B. distachyon* showed no collinear matches and even very few (351 and 571) unpaired matches.

The microsynteny results obtained in the present study are in agreement with the study done by Lescot et al. (2008), in which 11 *M. acuminata* BAC clones containing genes with a high degree of sequence conservation among monocots were used to study the relationships of *Musa* with rice and *Arabidopsis*. Comparison of the 11 *Musa* BAC sequences with the complete genome sequence of rice led to the identification of several syntenic regions between these two species. However, comparison between *Musa* and the more distantly related dicot *A. thaliana* revealed only one case of micro-collinearity (Lescot et al. 2008).

Conclusion

The data and analysis generated in this study give a first glimpse of the genome constitution of the DH Pahang accession which is the accession chosen for the whole-genome sequencing of banana. Additionally, it constitutes a resource for the *Musa* genomics consortium providing important information for understanding of the genome organization and the biology of the banana plant. DNA fingerprinting of the BAC library is being done at the moment and has already been used for scaffold assembly of the DH Pahang Genome (http://gnpannot.musagenomics.org/cgi-bin/gbrowse/musa_fpc/?name=ctg0:50639000..50798000).

Acknowledgments This project was funded through a grant of the Stichting Het Groene Woudt and was partially supported by the Dioraphte Foundation. We thank Dr. Jane Grimwood (HudsonAlpha Institute for Bioinformatics, Huntsville, AL, USA) for the BAC-end sequencing and Stefaan Vandamme and Dr. Kris Laukens (CEPROMA, Antwerp, Belgium) for technical assistance with the protein searching.

References

Abrusan G, Grundmann N, DeMester L, Makalowski W (2009) TEclass—a tool for automated classification of unknown eukaryotic transposable elements. Bioinformatics 25:1329–1330

Altschul SF, Madden TL, Schaffer 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

Ashburner M, Ball C, Blake J et al (2000) Gene Ontology: tool for the unification of biology. Nat Genet 25:25–29

Balint-Kurti P, Clendennen S, Dolezelova M, Valarik M, Dolezel J, Beetham P, May G (2000) Identification and chromosomal localization of the monkey retrotransposon in *Musa* sp. Mol Gen Genet 263:908–915

Bartos J, Alkhimova O, Dolezelova M, De Langhe E, Dolezel J (2005) Nuclear genome size and genomic distribution of ribosomal DNA in *Musa* and *Ensete* (Musaceae): taxonomic implications. Cytogenet Genome Res 109:50–57. doi:10.1159/000082381

Bennett M, Smith J (1991) Nuclear-DNA amounts in angiosperms. Philos Trans The R Soc Lond Ser 334:309–345

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Rapp BA, Wheeler DL (2002) GenBank. Nucleic Acids Res 30:17–20

Carpentier SC, Witters E, Laukens K, Onckelen HV, Swennen R, Panis B (2007) Banana (*Musa* spp.) as a model to study the meristem proteome: acclimation to osmotic stress. Proteomics 7:92–105. doi:10.1002/pmic.200600533

Carpentier SC, Coemans B, Podevin N, Laukens K, Witters E, Matsumura H, Terauchi R, Swennen R, Panis B (2008a) Functional genomics in a non-model crop: transcriptomics or proteomics? Physiol Plant 133:117–130. doi:10.1111/j.1399-3054.2008.01069.x

Carpentier SC, Panis B, Vertommen A, Swennen R, Sergeant K, Renaut J, Laukens K, Witters E, Samyn B, Devreese B (2008b) Proteome analysis of non-model plants: a challenging but powerful approach. Mass Spectrom Rev 27:354–377. doi:10.1002/mas.20170

Cheung F, Town CD (2007) A BAC end view of the *Musa acuminata* genome. BMC Plant Biol 7:29

Dsouza M, Larsen N, Overbeek, R (1997) Searching for patterns in genomic data. Trends Genet 13:497–498

Gasteiger E, Jung E, Bairoch A (2001) SWISS-PROT: connecting biomolecular knowledge via a protein database. Curr Issues Mol Biol 3:47–56

Goff SA, Ricke D, Lan T-H, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H et al (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. japonica). Science 296:92–100

Hong CP, Plaha P, Koo DH, Yang TJ, Choi SR, Lee YK, Uhm T, Bang JW, Edwards D, Bancroft I, Park BS, Lee J, Lim YP (2006) A survey of the *Brassica rapa* genome by BAC-end sequence analysis and comparison with *Arabidopsis thaliana*. Mol Cells 22:300–307

Hribova E, Dolezelova M, Town CD, Macas J, Dolezel J (2007) Isolation and characterization of the highly repeated fraction of the banana genome. Cytogenet Genome Res 119:268–274

Hribova E, Neumann P, Matsumoto T, Roux N, Macas J, Dolezel J (2010) Repetitive part of the banana (*Musa acuminata*) genome investigated by low-depth 454 sequencing. BMC Plant Biol 10 (1):204

Huo N, Lazo G, Vogel J et al (2008) The nuclear genome of *Brachypodium distachyon*: analysis of BAC end sequences. Funct Integr Genomics 8:135–147. doi:10.1007/s10142-007-0062-7

Lai C, Yu Q, Hou S et al (2006) Analysis of papaya BAC end sequences reveals first insights into the organization of a fruit tree genome. Mol Genet Genomics 276:1–12. doi:10.1007/s00438-006-0122-z

Lescot M, Piffanelli P, Ciampi AY et al (2008) Insights into the *Musa* genome: syntenic relationships to rice and between *Musa* species. BMC Genomics 9:58

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

Marin DH, Romero RA, Guzman M, Sutton TB (2003) Black Sigatoka: an increasing threat to banana cultivation. Plant Dis 87:208–222

Masoudi-Nejad A, Tonomura K, Kawashima S, Moriya Y, Suzuki M, Itoh M, Kanehisa M, Endo T, Goto S (2006) EGassembler: online bioinformatics service for large-scale processing, clustering and assembling ESTs and genomic DNA fragments. Nucleic Acids Res 34:459–462

Osuji J, Harrison G, Crouch J, Heslop-Harrison J (1997) Identification of the genomic constitution of *Musa* L. lines (bananas, plantains and hybrids) using molecular cytogenetics. Ann Bot 80:787–793

Paux E, Roger D, Badaeva E, Gay G, Bernard M, Sourdille P, Feuillet C (2006) Characterizing the composition and evolution of homoeologous genomes in hexaploid wheat through BAC-end sequencing on chromosome 3B. Plant J 48:463–474

Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. Meth Mol Biol 132:365–386

SanMiguel P, Gaut B, Tikhonov A, Nakajima Y, Bennetzen J (1998) The paleontology of intergene retrotransposons of maize. Nat Genet 20:43–45

Schoof H, Ernst R, Nazarov V, Pfeifer L, Mewes H, Mayer K (2004) MIPS *Arabidopsis thaliana* Database (MAtDB): an integrated biological knowledge resource for plant genomics. Nucleic Acid Res 32:D373–D376. doi:10.1093/nar/gkh068

Shultz J, Kazi S, Bashir R, Afzal J, Lightfoot D (2007) The development of BAC-end sequence-based microsatellite markers and placement in the physical and genetic maps of soybean. Theor Appl Genet 114:1081–1090

Tatusov RL, Fedorova ND, Jackson JD et al (2003) The COG database: an updated version includes eukaryotes. BMC Bioinform 4:41. doi:10.1186/1471-2105-4-41

Valarik M, Simkova H, Hribova E, Safar J, Dolezelova M, Dolezel J (2002) Isolation, characterization and chromosome localization of repetitive DNA sequences in bananas (*Musa* spp.). Chromosome Res 10:89–100

Venter JC, Smith HO, Hood L (1996) A new strategy for genome sequencing. Nature 381:364–366

Vij S, Gupta V, Kumar D, Vydianathan R, Raghuvanshi S, Khurana P, Khurana J, Tyagi A (2006) Decoding the rice genome. Bioessays 28:421–432. doi:10.1002/bies.20399

Yu J, Hu S, Wang J et al (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). Science 296:79–92