OPEN ACCESS Freely available online

PLoS one

A Genome-Wide Survey of Switchgrass Genome Structure and Organization

Manoj K. Sharma[1,2], Rita Sharma[1,2], Peijian Cao[3], Jerry Jenkins[4,5], Laura E. Bartley[1,2¤], Morgan Qualls[4,5], Jane Grimwood[4,5], Jeremy Schmutz[4,5], Daniel Rokhsar[5,6], Pamela C. Ronald[1,2]*

1 Department of Plant Pathology, University of California Davis, Davis, California, United States of America, **2** Joint BioEnergy Institute, Emeryville, California, United States of America, **3** China Tobacco Gene Research Center, Zhengzhou Tobacco Research Institute, Zhengzhou, China, **4** HudsonAlpha Institute of Biotechnology, Huntsville, Alabama, United States of America, **5** United States Department of Energy Joint Genome Institute, Walnut Creek, California, United States of America, **6** University of California, Berkeley, California, United States of America

Abstract

The perennial grass, switchgrass (*Panicum virgatum* L.), is a promising bioenergy crop and the target of whole genome sequencing. We constructed two bacterial artificial chromosome (BAC) libraries from the AP13 clone of switchgrass to gain insight into the genome structure and organization, initiate functional and comparative genomic studies, and assist with genome assembly. Together representing 16 haploid genome equivalents of switchgrass, each library comprises 101,376 clones with average insert sizes of 144 (*Hin*dIII-generated) and 110 kb (*Bst*YI-generated). A total of 330,297 high quality BAC-end sequences (BES) were generated, accounting for 263.2 Mbp (16.4%) of the switchgrass genome. Analysis of the BES identified 279,099 known repetitive elements, >50,000 SSRs, and 2,528 novel repeat elements, named switchgrass repetitive elements (SREs). Comparative mapping of 47 full-length BAC sequences and 330K BES revealed high levels of synteny with the grass genomes sorghum, rice, maize, and *Brachypodium*. Our data indicate that the sorghum genome has retained larger microsyntenous regions with switchgrass besides high gene order conservation with rice. The resources generated in this effort will be useful for a broad range of applications.

Citation: Sharma MK, Sharma R, Cao P, Jenkins J, Bartley LE, et al. (2012) A Genome-Wide Survey of Switchgrass Genome Structure and Organization. PLoS ONE 7(4): e33892. doi:10.1371/journal.pone.0033892

Editor: Frank G. Harmon, USDA-ARS, United States of America

Received November 17, 2011; **Accepted** February 19, 2012; **Published** April 12, 2012

Funding: This work was supported by the Office of Science of the United States Department of Energy under Contract No. DE-AC02-05CH11231 to the United States Department of Energy Joint Genome Institute, the Office of Biological and Environmental Research of the United States DOE contract No. DE-AC02-05CH11231 to the Joint BioEnergy Institute, and the United States Department of Agriculture National Institute of Food and Agriculture agreement No. 2011-67009-30153 to PCR. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: pcronald@ucdavis.edu

¤ Current address: Department of Botany and Microbiology, The University of Oklahoma, Norman, Oklahoma, United States of America

Introduction

The C4 perennial grass, switchgrass (*Panicum virgatum* L.), a member of Paniceae tribe of the Panicoideae subfamily of the Poaceae is a promising bioenergy crop [1,2]. Striking features include its high productivity, adaptability to growth on marginal lands, low nutrient and water requirements, and ability to sequester carbon and recycle nutrients [3,4,5,6,7].

The work reported here is part of an effort directed towards generating the genetic and genomic resources for switchgrass needed for gene discovery and breeding efforts [8,9]. Considering the highly outcrossing and tetraploid features of lowland switchgrass with two heterozygous genomes [10], major challenges will be independently assembling the subgenomes into a reference and reaching chromosome-scale contiguity. An accurate estimate of genome structure and composition prior to full genome sequencing is needed. Generation and sequencing of BAC libraries is an efficient strategy to obtain this information and support assembly of the large and complex underlying genomes [11,12,13,14,15,16]. Recently, an *Eco*RI-generated BAC library was reported from the SL93 2001-1 genotype of Alamo switchgrass [17]. Based on the analysis of homoeologous genomic regions harboring orthologs of the rice *Brassinosteroid insensitive 1* (*OsBRI1*), those authors made an attempt to provide a glimpse of switchgrass genome structure and complexity. However, the analysis was limited to a single locus and only one restriction enzyme (*Eco*RI) was used. Additional libraries are required to achieve unbiased and near-complete representation for genome-wide studies.

Here, we describe the generation and characterization of two high-quality BAC libraries using two different restriction endonucleases (*Bst*YI and *Hin*dIII) prepared from the switchgrass genotype Alamo clone AP13. Because this clone was the parent of the first mapping population described for switchgrass and has been further used in defined crosses [18], it was chosen as the consensus target for sequencing by the switchgrass community. Collection of 330,297 high-quality BAC-end sequences (BES) were generated from both the libraries that provided the basis for a genome-wide survey of switchgrass genome structure and organization. Comparative mapping of full-length BACs and BES onto four other grass genomes reveals high levels of synteny and micro-collinearity. Gene annotations and analysis of BES provide an estimate of protein signatures, GC content, repeat elements and SSRs in switchgrass genome.

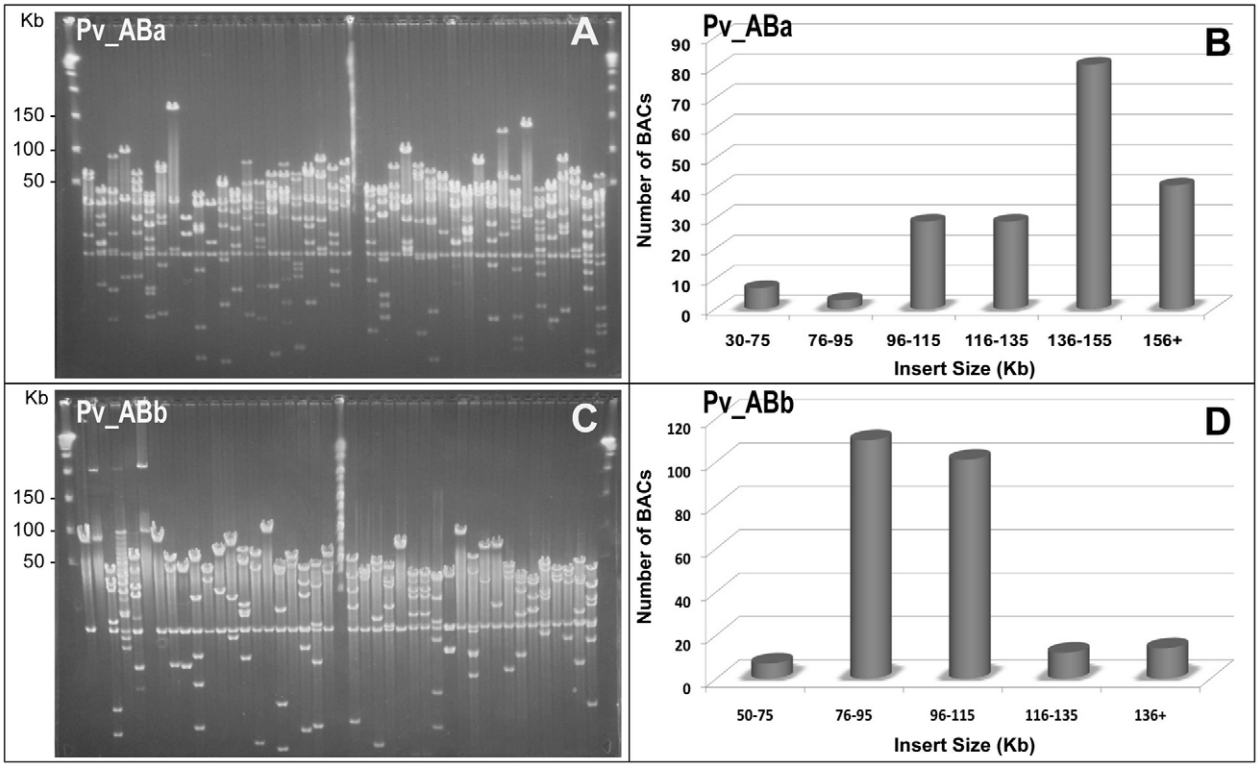

Figure 1. Determination and distribution of switchgrass BAC clone insert sizes. DNA was digested from >180 randomly selected BAC clones from Pv_ABa and Pv_ABb libraries and analyzed by Pulsed Field Gel Electrophoresis. A, C), Representative gel pictures of *Not*I digested BAC DNA from Pv_ABa and Pv_ABb libraries, respectively. B, D) Estimated BAC insert sizes with their relative frequencies.
doi:10.1371/journal.pone.0033892.g001

Results

Construction and Characterization of BAC Libraries

We constructed two BAC libraries, Pv_ABa and Pv_ABb, from AP13 clone of switchgrass using *Hin*dIII and *Bst*YI, respectively. Each library consists of 101,376 clones. To estimate insert size, >180 clones were randomly picked from each library. *Not*I digestion of these clones generated 7.8 kb vector band and various-sized insert fragments (Figure 1A, C). The inserts in Pv_ABa ranged from 30 to 280 kb, with the majority of fragments in the 136–155 kb size range (Figure 1B) and an average size of 144 kb. For Pv_ABb, insert sizes ranged from 50 to 200 kb, with the majority of fragments in the 76–115 kb size range and an average size of 110 kb (Fig. 1D). More than 80% of tested clones from both libraries had an insert size larger than 100 kb. A very low percentage (<1%) of empty clones were detected in both the libraries. The detailed characteristics of both the libraries are summarized in Table 1.

To assess the quality of the BAC libraries, high-density colony filters were hybridized with chloroplast/mitochondria-specific probes spanning the whole genome of respective organelle. Using a pool of chloroplast-specific genes, viz., *rbc*L, *ndh*A, *rpo*B and *trn*L, 209 and 62 clones among 36,864 clones from Pv_ABa and Pv_ABb BAC libraries, respectively, produced hybridization signal. We, therefore, estimate that 0.57 and 0.17% clones in Pv_ABa (Figure 2A) and Pv_ABb (Figure 2D), respectively, carry chloroplast-originated DNA sequences. Similarly, hybridizations with the mitochondrial DNA probe containing mixture of *atp*6, *atp*9, *cob* and *cox*1 gene-specific amplicons identified 79 and 23 mitochondrial clones in Pv_ABa (Figure 2B) and Pv_ABb (Figure 2E), respectively. This amounts to 0.21 and 0.06% contamination from mitochondrial clones in Pv_ABa and Pv_ABb library, respectively. The overall contamination of organellar DNA in Pv_ABa and Pv_ABb is, therefore, estimated to be 0.78 and 0.23%, respectively.

Coverage of the Switchgrass Genome

Prior analyses suggest that switchgrass is an allotetraploid with an effective genome size of 2x = 1n = 1600 Mbp [19]. Considering the ~1600 Mbp effective genome size of *Panicum virgatum* L. var.

Table 1. Characteristics of the switchgrass BAC libraries.

Characteristic	Pv_ABa	PV_ABb
Cloning vector used	plindigoBAC536	plindigoBAC536
Restriction enzyme	*Hin*d III	*Bst*Y I
Total number of clones	101,376	101,376
Percent empty clones	<1%	<1%
Maximum insert size	280 Kb	200 Kb
Minimum insert size	30 Kb	50 Kb
Average insert size	144 Kb	110 Kb
Chloroplast DNA contamination	0.57%	0.17%
Mitochondrial DNA contamination	0.79%	0.29%
Number of genome equivalents	9X	7X

doi:10.1371/journal.pone.0033892.t001

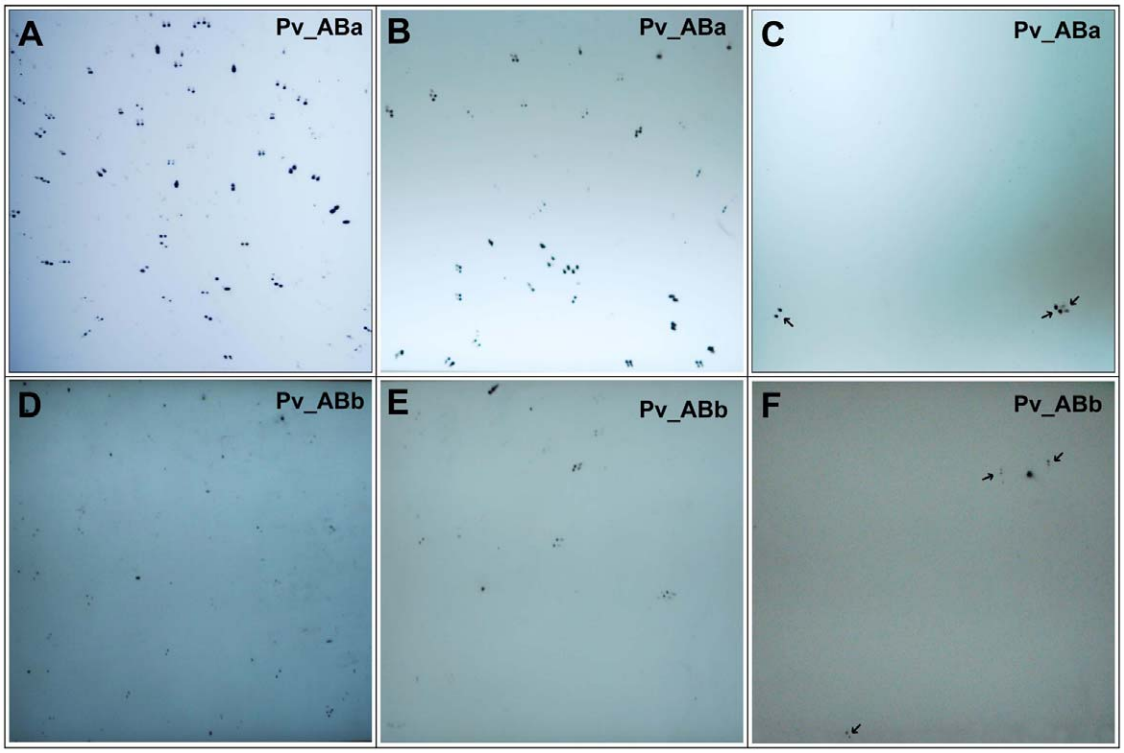

Figure 2. Estimation of organellar DNA contamination and representation of low-copy genes. Switchgrass BAC libraries (Pv_ABa and Pv_ABb) were screened by high-density filter hybridizations to estimate chloroplast- or mitochondrial-specific DNA and representation of single/low copy genes. Representative filter hybridization data used to estimate chloroplast (A, D) and mitochondrial (B, E) contaminants, and the library coverage based on the presence of a single copy gene, *brittle culm 10* (C, F). Black arrows in C and F identify the signal from *BC10* probes.
doi:10.1371/journal.pone.0033892.g002

Alamo and removing estimated organellar DNA-specific (0.78 and 0.23%) as well as empty clones (1%), each library represents ~9 and 7 haploid genome equivalents. Therefore, the theoretical probability of finding a sequence of interest in these library resources is more than 99.9%.

We empirically validated the coverage using filter hybridizations with single/low copy genes (Figure 2C, F). The copy number of six genes, including *brittle culm 10* (*BC10*), *xyloglucan endotransglucosylase/hydrolase* (*OsXTH*) and, *Teosinte branched 1* (*TB1*) of rice and *Tubulin-4*, *Opaque*, and *Starch branching enzyme 1 (SBE1)* of maize, was determined using Southern hybridizations. In switchgrass, *OsXTH* and *TB1* appear to have several copies or exhibit variability among homoeologous regions, whereas, *BC10*, *Tubulin-4*, *Opaque* and *SBE1* have single or low copy number (Figure S1). Using a *BC10* gene-specific probe, three clones were identified among 18,432 clones of each library (Figure 2C, F). Similarly 3, 2 and 2 clones specific to *Tubulin-4*, *SBE1* and *Opaque*, respectively, were identified among 18,432 clones of Pv_ABa library (data not shown). Conversely, 2, 1 and 3 clones were identified for *Tubulin-4*, *Opaque* and *SBE1*, respectively, from the second library. Therefore, an average of two clones were obtained per single/low copy gene in the 18% of the clones represented on the filters, corresponding to about 11 hits in each library and consistent with the high coverage of each BAC library.

BAC-end Sequencing and Analysis

Because BES data represent a random snapshot of a genome, it can be used to perform a genome-wide survey of structural features. We sequenced paired ends of 101,376 and 84,480 clones from Pv_ABa and Pv_ABb, respectively. After removing *E. coli*-specific sequences, vector sequences, short/failed sequences, and organelle-specific DNA, a data set of 330,297 (~263 Mbp) high quality sequences (> = 400 HQ bases) was generated. These represent ~16.4% of the switchgrass genome. 95.9% BES were paired. The length of BAC-end sequences varied from ~100 to 1000 bp with an average length of 761 bp (Figure 3). More than 73% clones of each library had a read length longer than 700 bp. Based upon homology with coding sequences from other grass genomes and the presence of protein domains, approximately, 15.4% (40 Mbp) of BES had a protein signature. A protein signature refers to the contiguous pattern of amino acids associated with a particular structure or function of proteins [20]. Based on the BES analyzed, the GC content of switchgrass is estimated to be ~45.5%. Further, GC content in the sequences with a protein signature is 57.8%, which is significantly higher that the GC content (43.3%) of non-coding region in the BES (222 Mbp).

Analysis of Simple Sequence Repeats. We identified a total of 50,206 SSRs from BES that includes 1–3 nt repeats (at least 12 nt in length) and 4–6 nt repeats (having at least four tandem repeat units) adding up to 870,808 bases. The density of SSRs is therefore, estimated to be one SSR per 5.2 kb of sequence. The most abundant of these were trimeric SSRs (55%,), followed by dimers (20.4%) and monomers (16.6%; Figure 4a). However, tetramers, pentamers and hexamers were much lower in abundance and all together add up to less than 10% of total microsatellites. Furthermore, GC-rich trimers constitute 63% of total trimers with GCC/GGC and CGC/GCG being most abundant (Figure 4b). ACT/AGT trinucleotides were least in number (Figure 4b). About 14% of the SSRs (6812 in number) were longer than 20 nucleotides. Details of SSRs and their frequencies are given in File S1.

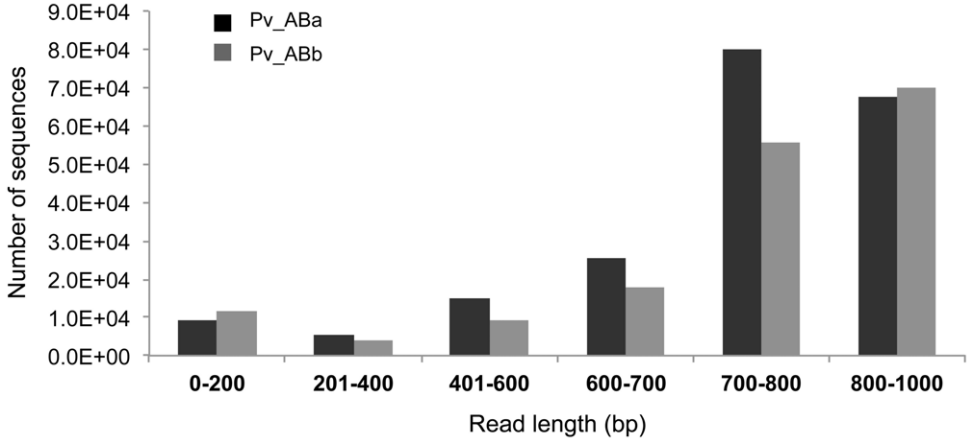

Figure 3. Size distribution of BAC-end sequences. The x-axis represents read length with the numbers of sequences indicated on y-axis.
doi:10.1371/journal.pone.0033892.g003

Analysis of Repetitive Elements. Based upon homology with known plant repeat elements, 279,099 repeat elements were identified from the switchgrass BES (Table 2). Such repeats correspond to 30.97% of the total sequence analyzed. Class I and class II transposons account for 73.7 and 26.3% of total transposons, respectively, thereby suggesting an approximate

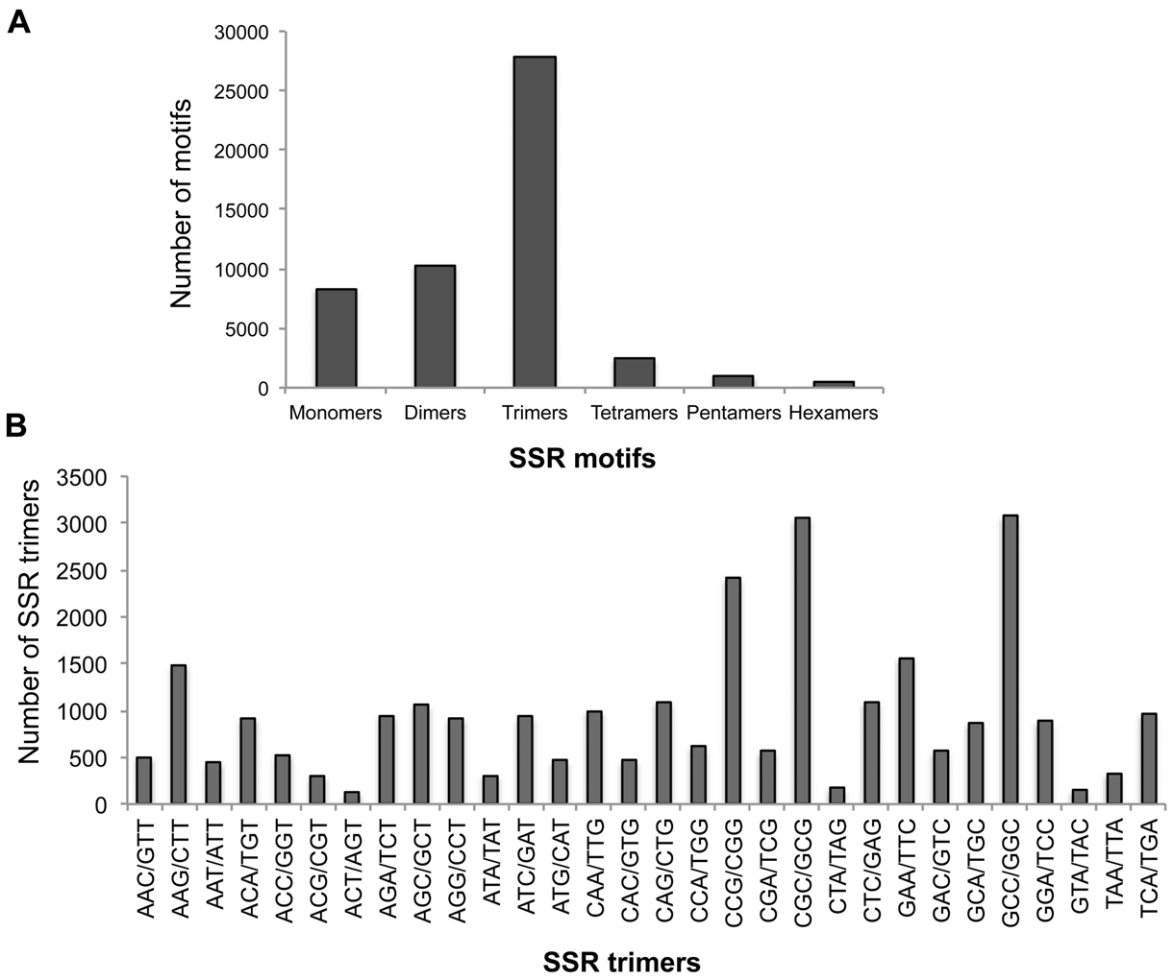

Figure 4. Analysis of Simple Sequence Repeats (SSRs) from BAC-end sequences. a) Distribution of total number of repeat loci. The x-axis represents the length of SSRs and the y-axis indicates total number of motifs observed. **b**) Distribution of SSR trimers. X-axis represents the various trinucleotide SSRs and the y-axis represents number of SSRs.
doi:10.1371/journal.pone.0033892.g004

ratio of 3:1 in the switchgrass genome. Class I transposons include Long Terminal Repeats (LTR) elements, Short Interspersed Elements (SINEs), and Long Interspersed Elements (LINEs). LTR-elements were most abundant and comprise 90.4% of total retrotransposons identified; however, SINEs (1.2%) and LINEs (8.3%) were very low in number (Table 2). LTR-elements are further classified into five major groups including BEL, Ty1/Copia, Ty3/gypsy, DIRS1 and vertebrate retroviruses. We did not find any BEL or retrovirus type elements in switchgrass. Ty3/Gypsy and DIRS1 together comprise 67% of the LTRs in switchgrass. Similarly, Class II transposons include 35% En-spm, 13.5% Tourist/Harbinger, 17.8% MuDR-IS905, 12.4% Hobo-Activator, 9.06% Tc1-IS630-Pogo and others (8.7%). Based on these results, we estimate that ~31% of the switchgrass genome corresponds to known repeat sequences. Several retro-element subfamilies including Penelope, CRE/SLACS, L2/CR1/Rex, R1/LOA/Jocky, R2/R4/NeSL, BEL/Pao, Rolling-circles and, DNA transposons viz., PiggyBac, Mirage, P-element and Tarnsib were not found in the sequence analyzed.

Identification of Novel Repeats. Similarity-based repeat detection is generally limited by the size and diversity of the available databases. To identify switchgrass-specific novel repeat elements, we carried out a self-comparison of the BES. Even with the stringent threshold requirement that each 100 bp window matches another BES with at least 90% identity, 61.2% (202,280) of the switchgrass BES matched at least one other BES (Figure 5). We identified 2,948 repeat sequences among those BES with at least six matches with other switchgrass BES. When these sequences were queried against the RepBase repeat database, MSU Plant Repeat Databases, Triticeae repetitive sequence database (TREP), NCBI GenBank non-redundant nucleic acid sequence database and Swissprot database (release 2011_08), 420 repeat sequences matched at least one record in the mentioned databases and were therefore, removed from the list of putative switchgrass repetitive elements (SREs). The remaining 2,528 SREs were present in 7 to 548 copies in the BES database and their sizes ranged from 80 to 300 bp (File S2). Overall, these SREs matched 83,289 BES, covering a ~6 Mbp region that accounts for ~2.3% of the total BES length. Extrapolating to the level of the switchgrass genome, there could be as many as 3,341 copies of the most frequent SREs.

Functional Annotation and Gene Ontology Analysis. To better characterize this valuable resource and provide an overview of the expanse of biological functions encoded by the switchgrass genome, we performed functional annotation and GO analysis of protein-coding signatures obtained from the BES with regard to the three major gene ontology terms viz., molecular function, biological process and cellular locations. Out of the 330,297 BES, 5052 could be associated with at least one GO term (File S3). In total, 716 terms were associated with 5052 reads. 4507 reads were assigned at least one of the 377 molecular function categories, 3244 reads were annotated with at least one of 259 biological function categories and 1144 reads were associated with at least one of the 80 cellular location categories. Figure 6 presents the distribution of GO terms identified from the switchgrass BES. The top most terms highlighted in the cellular location category included membrane (37%) and those comprising protein complexes (21%). Equal representation (11%) of those associated with nucleotide binding, metal ion binding, nucleic acid binding and hydrolase activity were found in the molecular function

Table 2. Distributions of repeat elements identified from switchgrass BAC-end sequences.

Type of Element	Total number of elements*	Total length occupied (bp)	%age of total sequence analyzed
Retroelements	178318	64563658	24.53%
SINEs	2306	348493	0.13%
Penelope	27	3174	0.00%
LINEs	14816	6018230	2.29%
R2/R4/NeSL	2	106	0.00%
RTE/Bov-B	3780	1911841	0.73%
L1/CIN4	11001	4102823	1.56%
LTR elements	161196	58196935	22.11%
Ty1/Copia	50870	18995037	7.22%
Ty3/Gypsy/DIRS1	108785	38941535	14.79%
DNA transposons	63616	14149503	5.38%
hobo-Activator	7917	1872711	0.71%
Tc1-IS630-Pogo	5764	867375	0.33%
En-Spm	22287	6169405	2.34%
MuDR-IS905	11351	2572080	0.98%
Tourist/Harbinger	8620	1441495	0.55%
Others	-	1226437	0.47%
Unclassified	4424	804772	0.31%
Total interspersed repeats:		79517933	30.21%
Small RNA:	2382	568637	0.22%
Satellites:	1634	244518	0.09%
Low complexity:	28725	1199167	0.46%

*most repeats fragmented by insertions or deletions have been counted as one element.

doi:10.1371/journal.pone.0033892.t002

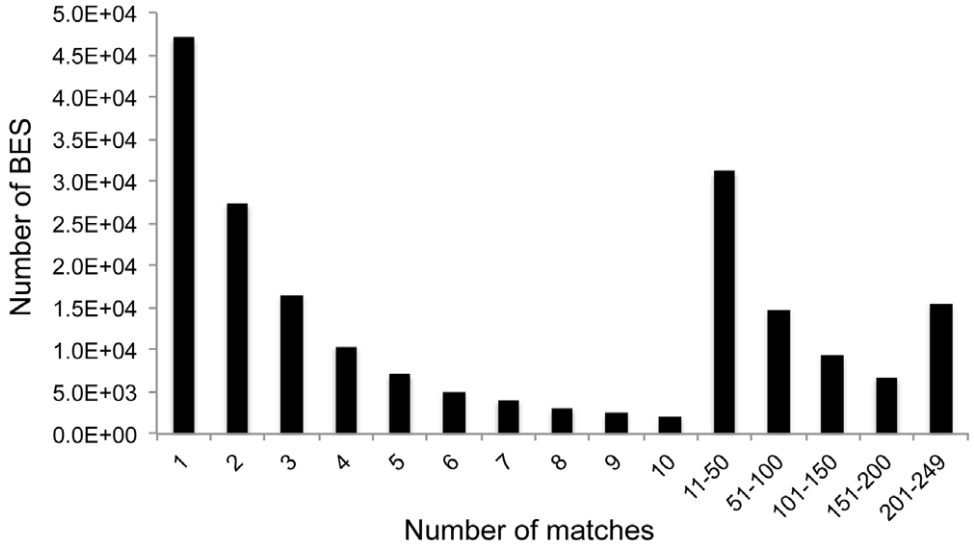

Figure 5. Distribution of BAC end sequences that show significant homology to other BES. The x-axis represents the number of matches and y-axis contains total number of BAC-end sequences.
doi:10.1371/journal.pone.0033892.g005

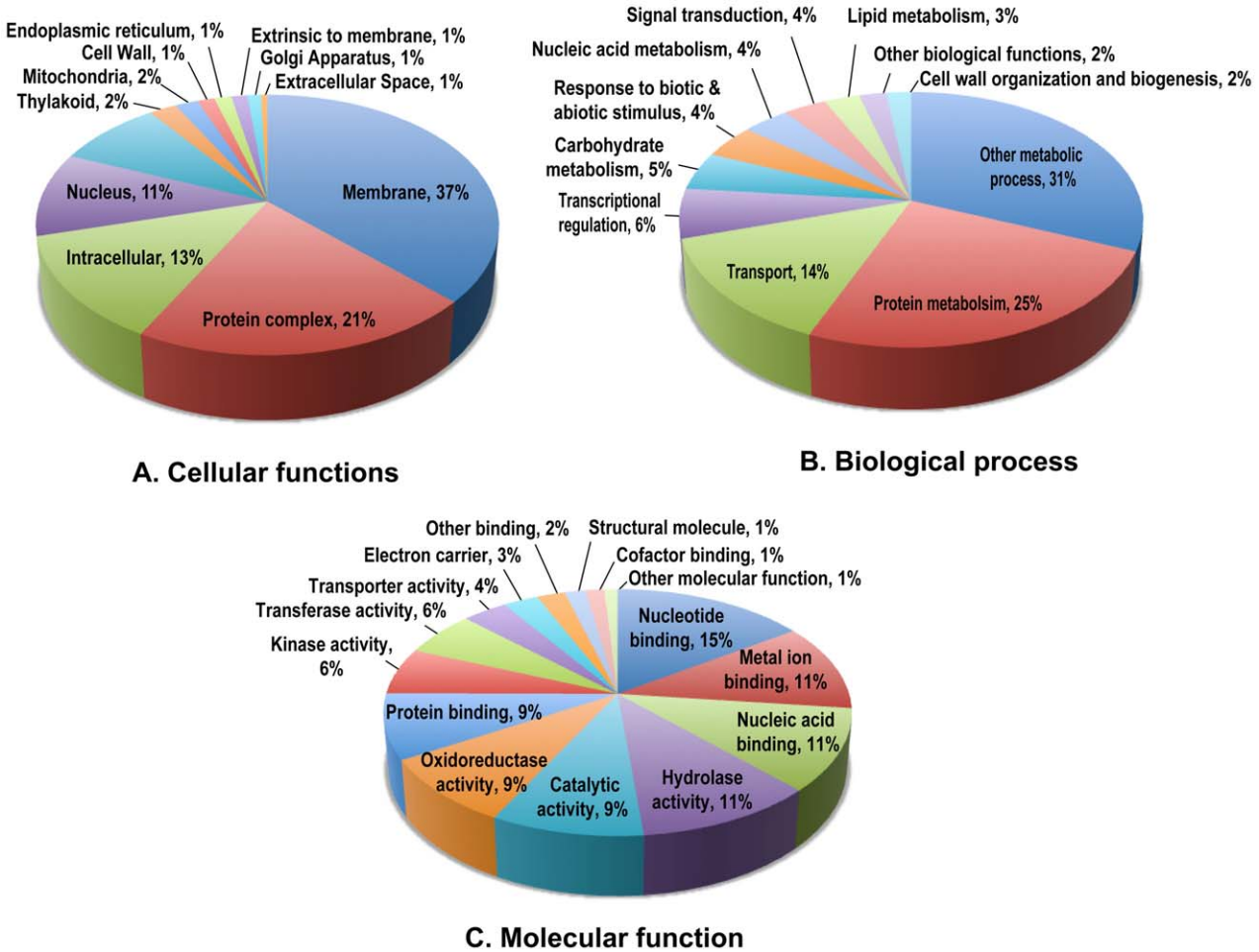

Figure 6. Distribution of GO-slim annotations of putative gene products predicted from switchgrass BAC-end sequences. A, Cellular locations −12 groups of gene ontology; B, Biological processes −11 groups of gene ontology; C, Molecular functions −15 groups of gene ontology terms.
doi:10.1371/journal.pone.0033892.g006

category followed closely by catalytic activity (9%), oxidoreductase activity (9%) and protein binding (9%) terms (Figure 6). With regard to biological functions, terms associated with metabolic processes were most abundant, followed by transporters (14%) and transcriptional regulators (6%). Overall, genes annotated to encode kinases, transcription factors, metal ion binding proteins and oxidoreductases comprise a large proportion of the coding regions of switchgrass genome.

Comparative Mapping of Switchgrass BES. For comparative mapping, we initially mapped switchgrass BES to rice peptides, which were subsequently mapped onto sorghum and *Brachypodium* genomes. A GBrowse-based synteny browser, GBrowse-syn [21], was used to display the synteny between the rice, sorghum and *Brachypodium* genomes. Approximately 8% of the BES mapped to sorghum, 7% to rice, and 5.5% to the *Brachypodium* genome. In total, 4522 (1%) paired end reads mapped to sorghum; whereas, 24,758 (~7%) reads mapped as high scoring singlets. Mapping onto the rice genome placed 2400 (0.7%) paired ends and 22,158 (6.4%) high scoring singlets. Similarly, 1568 (0.5%) paired ends and 17,517 (5%) high scoring singlets mapped onto the *Brachypodium* genome. Figure 7 displays a snapshot of a 2.0 Mbp region of rice with mapping results from corresponding regions of sorghum, *Brachypodium* and switchgrass BAC-end sequences. In the region, 332 BAC-ends mapped to sorghum, 298 to rice and 275 to *Brachypodium* genome. Forty-six BAC-end sequences that mapped to sorghum had both ends placed within 500 kb of one another. Similarly, 24 paired-BES were mapped to orthologous region in rice and 22 to *Brachypodium* genome. Based on the paired placements in the region shown in Figure 7; 74.7, 89.45 and 43.29% BES mapped to coding sequence in sorghum, *Brachypodium* and rice, respectively. The regions with both ends mapped within 500 kb represent microsyntenous regions in these genomes.

Analysis of Microcollinearity using Full-length BAC Sequences. Forty-seven randomly selected BACs from Pv_ABa were sequenced to essentially full-length using Sanger's method. The average size of these BACs was 153.6 kb. The distribution of SSRs and repeat elements in the full-length BAC sequences (File S4) is very similar to their distribution among the BES. A total of 439 gene loci (451 gene models; File S5) were annotated from ~7.2 Mbp of switchgrass genomic sequence, obtained from full-length BAC sequences. The gene density is therefore estimated to be one per 16.4 kb of genomic sequence. Predicted cDNA, protein and genomic sequences of these loci are given in File S6. The genes predicted from these sequences were mapped onto other grass genomes. Corresponding orthologs for 370 (84%), 363 (83%), 357 (82%) and 336 (77%) gene loci could be identified from rice, maize, sorghum and *Brachypodium*, respectively (File S7).

We compared the order of switchgrass genes and their transcriptional orientations with orthologous regions in sorghum, maize, rice and *Brachypodium*. Figure 8 shows the pictorial representation of micro-collinearity among five BAC clones of switchgrass and the corresponding regions in other sequenced grasses. Generally, the length of corresponding regions is longer in maize and smaller in *Brachypodium*, in agreement with the whole genome size rankings. Despite various local rearrangements in these regions including inversions, translocations, deletions and insertions, we generally observed a high level of micro-collinearity in terms of gene content. A few genes have undergone tandem duplication in switchgrass resulting in paralogs. The list of genes from rice, sorghum and *Brachypodium*, not represented in

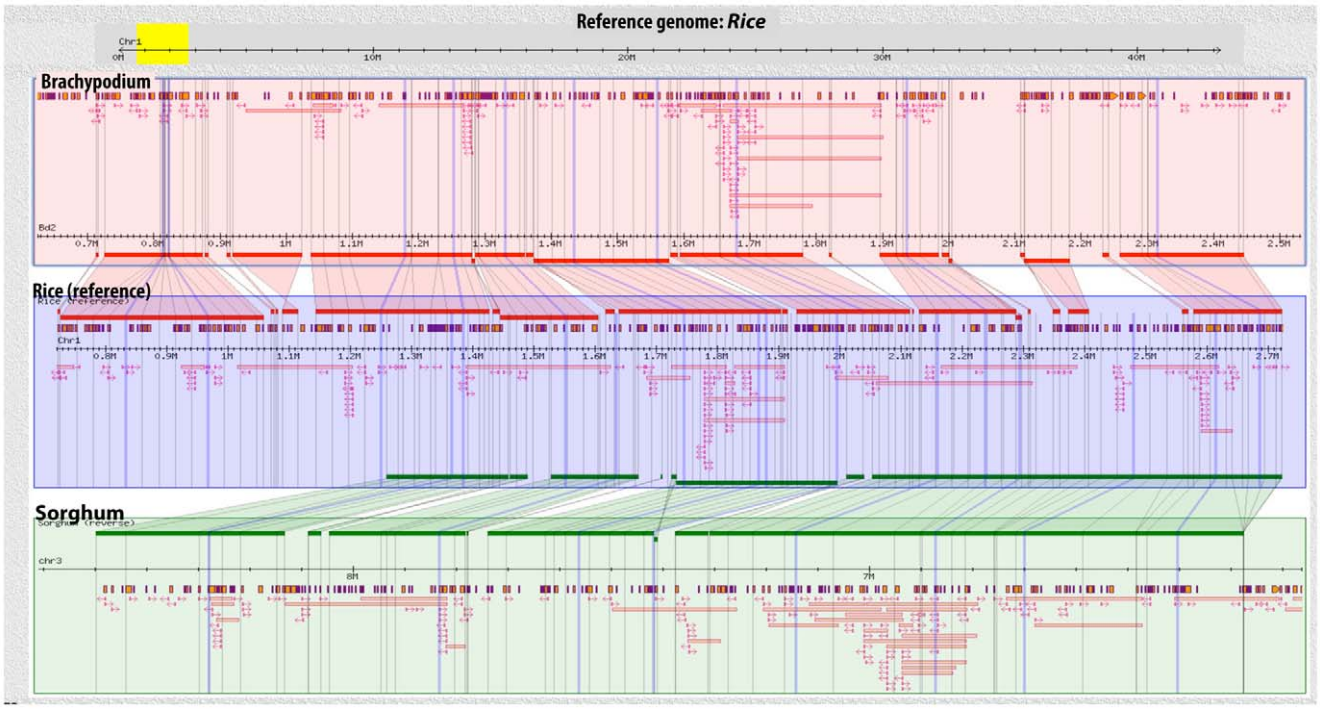

Figure 7. Mapping results of switchgrass BAC-end sequences to a 2 Mbp region of rice with orthologous regions from sorghum and *Brachypodium*. BES having base pair identity >75% with e value <1e-20 and coverage of >50% were placed on to rice peptides. The equivalent regions in sorghum and *Brachpyodium* were identified and used for mapping BES. BES pairs that were placed within 500 kb are represented by red bars. Red blocks show high-quality singlets of switchgrass with arrows indicating orientation. Synteny between grass genomes is marked by mauve alignment.
doi:10.1371/journal.pone.0033892.g007

switchgrass, is given in File S8. Out of 47 BACs analyzed, half are significantly collinear with other grass genomes; whereas, the rest show varying rearrangements (File S7). Reduced collinearity in some of the BACs seems to be due to low representation of coding sequences in these BACs. Overall, order, transcriptional orientations and gene structures of switchgrass genes seem more conserved with those of rice and sorghum, than those of maize and *Brachypodium* (Figure 8).

Discussion

High Quality BAC Libraries Provide a Valuable Resource for Diverse Genetic and Genomic Studies in Switchgrass

While trying to assemble the tetraploid genome of switchgrass, a major challenge will be to discriminate between paralogous, orthologous and homoeologous regions. Further repetitive regions longer than the read length and similarity in homoeologous regions may lead to potential misassemblies, which could require a great deal of directed sequencing to accurately resolve [22]. An ordered clone sequencing [23] approach using large insert clones can assist in assembly of the shorter genome sequences generated by next generation sequencing technologies [24,25,26]. BAC libraries are preferred over fosmid, cosmids or yeast artificial chromosomes, for this purpose because of their ability to preserve larger DNA fragments and lower level of chimerism [27,28,29,30,31].

Here we report construction of two BAC libraries from switchgrass accounting for ~16 haploid genome equivalents of switchgrass with >99.9% probability of finding a particular sequence. The large insert size, high coverage and low organellar DNA contamination indicate that these libraries provide a useful resource for diverse genetic and genomic studies including genetic and physical mapping, exon trapping, isolation of closely-linked polymorphic markers, FISH analysis, as well as functional and comparative genomics studies [28,32,33,34,35]. The percentage of empty clones observed (~1%) is also comparable or significantly lower than other reports for maize (0.4%; [36]), *Panax ginseng* (2.7%; [37]), *Vitis vinifera* (2.2%; [38]) and *Brachypodium* (4.6 and 5.1%; [39]). As these libraries have been constructed from the same clone (AP13) that is being sequenced at JGI, the sequences generated will prove instrumental for assembly and gap filling of the genome sequence of switchgrass.

GC-rich Trinucleotides are the Most Abundant SSRs in Switchgrass

Microsatellites play an important role in genome evolution and gene regulation. They have been extensively used in several research areas including linkage mapping, comparative genomics and population genetics [40,41]. Monocot genomes are enriched in GC-rich SSRs [42] with trinucleotide SSRs being most abundant in sorghum, maize and rice genomes (File S9; [43]. We find that switchgrass also, trinucleotide SSRs predominate (55.3%), with 63% of them being GC-rich, reflecting the codon bias. These observations are similar to the results observed for rice (65%) and *Brachypodium* (67.4%). Distributions of SSRs in full-length BAC sequences also showed similar distribution patterns as identified with BES. In plants, a negative correlation exists among SSR density and genome size [42] and our data also conforms to this general trend (File S9). Out of >50,000 SSR sequences discovered here, 6,812 are longer than 20 nucleotide in length and will serve as a valuable resource to develop highly heterozygous and polymorphic markers for saturating existing linkage maps.

Repeat Content in Switchgrass is Estimated to be ~33%

Transposable elements are abundant in plant genomes and play an important role in determining the size of grass genomes and driving genome evolution in response to environmental cues [44,45]. Known repeat elements accounted for approximately 31% of the total BES analyzed, with transposable elements representing about 86.7% of the repetitive-DNA fractions. Therefore, the estimated transposon content in switchgrass is approximately 29.9%. The percentage of retroelements in switchgrass (24.53%) is more than double compared to *Arabidopsis* (10%; [46]), similar to that of rice (26%), half of sorghum (55%) but less than one third of maize (79%; [47]). Analysis of full-length BAC sequences also showed similar patterns (File S4). Similar to poplar, rice and sorghum [48], the Gypsy group of LTRs is the most abundant repetitive elements in switchgrass. The ratio of Gypsy to Copia elements in switchgrass is ~2:1, similar to the ratio reported for rice [49]. LTRs have not only been implicated in genome reorganization but are also involved regulating plant adaptation to biotic and abiotic stresses [50]. Therefore, these elements might have significant contribution in stress adaptation and shaping the switchgrass genome.

In addition to the repetitive DNA fraction identified by classical analysis (30.97%), novel SREs (~2.3%) bring the total repetitive DNA content of switchgrass to a minimum of ~33% which is similar to estimated repeat content in rice in spite of the much greater genome size of switchgrass (File S9).

GC Content in Switchgrass is Comparable to Other Grasses

GC content is an important feature of a genome as indicated in several studies of prokaryotes, vertebrates and plants [51,52,53,54]. Gene density, patterns of codon usage, distribution of repeat elements, methylation patterns and recombination rate are all associated with GC content [52,55,56]. GC content is correlated with codon bias specifically at the third position and is reported higher in monocot plant species (File S9; [56]). Based on BES data, the estimated GC content in switchgrass is 45.5%, which is comparable to other monocot species (File S9). However, GC content of coding regions (57.8%) is noticeably higher than that of non-coding regions (43.3%), which may be the result of GC-rich codon usage and will be important for gene annotations of this species [57].

Gene Density in Switchgrass is more similar to that of Rice

Due to its large genome size, the genes in switchgrass are expected to have longer intergenic regions as compared to rice and other shorter genomes. Based on BAC-end sequence analysis, the estimated gene density in switchgrass is one gene per 16.4 kb, which varies in gene-rich and gene poor or repetitive regions. The highest density observed among the full-length BAC sequences is one gene per 6.8 kb (AC243226) and lowest was one gene per 59.4 kb (AC243244). Conversely, gene density in rice, sorghum and *Brachypodium* is one gene per 13.4 Kb, 26.7 Kb and 10.6 Kb, respectively [58]. However, gene density in maize is estimated to be three times lower than that of rice [59]. Closer inspection of some BACs suggested that in the regions of high gene density, most of the genes are clustered within a short distance. Therefore, the gene arrangement in switchgrass is more similar to that of rice.

Synteny and Collinearity of Switchgrass with Evolutionarily Diverged Grass Species

Investigation of genomic organization and comparative mapping to other grasses using RFLP (restriction fragment length polymorphism) markers revealed several syntenic regions between the rice and switchgrass genomes. [18]. Similarly, ESTs and other marker-based studies have also revealed significant similarity of

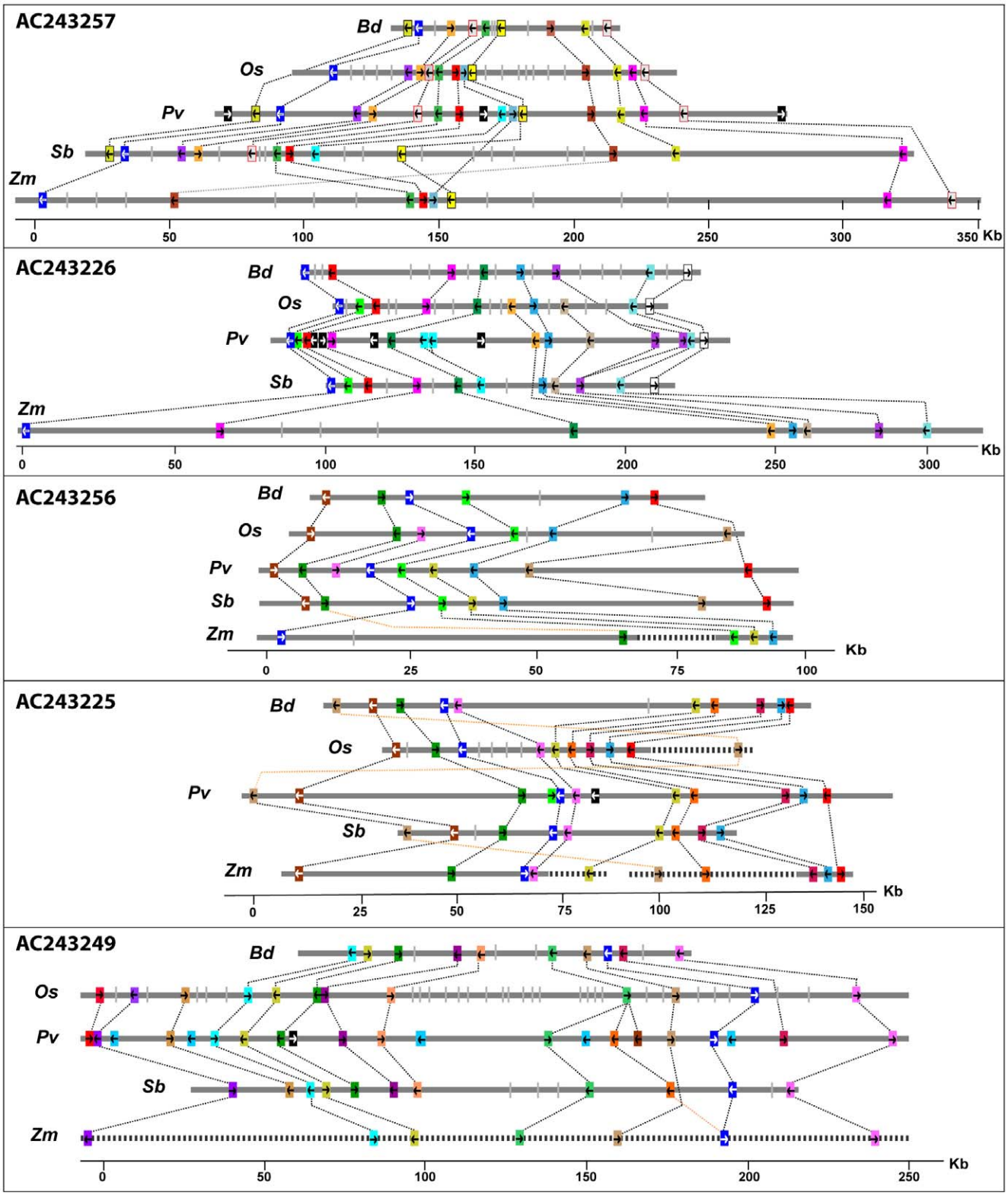

Figure 8. Micro-collinearity between switchgrass BAC clones and orthologous regions from *Brachypodium* (*Bd*), rice (*Os*), sorghum (*Sb*) and maize (*Zm*). Colored boxes along the physical location in the genome of each species represent genes and arrows in the colored boxes indicate the transcriptional orientation of each gene. Orthologous genes are given the same color and are connected by dotted lines. Grey bars represent genes from respective genomes lacking syntenic match in switchgrass. Dashed lines represent breaks in contiguity to allow larger genomic regions of the chromosomes to fit in the scale of the figures and the genes from these regions lacking syntenic match have not been plotted. The scale is shown at the bottom of each section. NCBI accession numbers for switchgrass BAC clones are given at top left of each section. Detailed information on accession numbers and gene names is given in File S7.
doi:10.1371/journal.pone.0033892.g008

switchgrass genome to sorghum, pearl millet and rice [10,60,61,62]. However, conservation of marker order at the level of a genetic map may not reflect the micro-collinearity at the genic level [63,64]. Sequence comparisons at various loci have shown that local rearrangements including deletions, insertions, duplications and translocations have occurred among related genomes at loci that otherwise seem collinear at in genetic mapping [65]. These results indicate that a closer look at gene-level collinearity is needed.

Single-pass BAC-end sequences are generally very specific and hence can be used as markers for comparative genomic studies. The BES reported here covers 16.4% of switchgrass genome and thus provides a reliable resource for anchoring switchgrass sequences to related grass model genomes. We picked four genomes with varying evolutionary distances viz., sorghum, maize, rice and *Brachypodium*, for genome-wide comparisons with switchgrass. Based on the BES mapping, we identified 3338, 2400 and 1568 putative microsyntenic regions with sorghum, rice and *Brachypodium*, respectively. Identification of orthologous segments in these regions may facilitate functional genomic studies in switchgrass.

Comparisons of full-length BAC sequences of switchgrass also revealed its higher similarity to sorghum followed closely by rice and then maize and *Brachypodium*. Sorghum and maize diverged from switchgrass about 28 million years ago [66]; whereas, rice and *Brachypodium* have diverged from switchgrass >50 and 60 mya, respectively. Reiterating the significance of genic-level sequence comparisons, the phylogenetic divergence between these genomes does not correspond to the pattern of collinearity we observed.

Due to difficulty of cloning and characterizing genes in polyploids like switchgrass; rice and *Brachypodium* have been promoted as surrogates for gene discovery and genomic analysis of other grasses [65]. Our results suggest that findings from the model genomes can be utilized for initiating functional genomic studies in switchgrass. However, due to widespread genome rearrangements, sorghum, along with soon to-be-completed foxtail millet genome will better serve as reference for assembling the genic region of the switchgrass genome.

It will be intriguing to investigate what makes switchgrass so different from these crops in terms of morphology, effective genome size (~1600 Mbp; four times than that of rice), ploidy level (polyploid vs diploid rice) and physiological processes (C4 vs C3 in rice and *Brachypodium*). Certainly, the substantial rearrangements observed in some of the BACs would contribute to these factors. The set of genes identified from switchgrass that lack syntenic matches with other genomes may represent lineage-specific loci with novel or divergent functions. Detailed analysis of switchgrass gene functions is needed to enlighten this area.

The results reported here represent an important milestone for advancement of functional and comparative genomic studies of switchgrass. The BAC library resources and comparative anchoring of BES will be useful for SSR marker development, saturating existing linkage maps, anchoring physical and genetic maps, and assembly of ongoing genome sequence of switchgrass.

Materials and Methods

Plant Material and HMW DNA Preparation

Leaf tissue from young plantlets of *Panicum virgatum* L. cv. Alamo clone AP13, provided by the group of Michael Udvardi at The Samuel Roberts Noble Foundation, was used for preparation of high molecular weight (HMW) DNA. Briefly, nodes from greenhouse-grown plants were sterilized with 20% commercial bleach containing 0.1% Tween 20 followed by *in vitro* culture. New shoots were cut and transferred to rooting medium. Leaf tissue was harvested from plantlets after 16 h of dark treatment and frozen in liquid nitrogen.

BAC Library Construction

BAC libraries were constructed at Clemson University Genomics Institute (CUGI) according to a published protocol [67] with minor modifications. Briefly, 100 g tissue was ground to powder in liquid nitrogen with pestle and mortar, and nuclei were isolated. To remove charged molecules as well as small and sheared gDNA, nuclei embedded into agarose plugs were exposed to pre-electrophoresis by loading onto a 1% TBE CHEF gel under the following conditions: 1 to 4 s switch times run at 4 V/cm for 3 h at 14°C. Genomic DNA was digested with *Hin*dIII and *Bst*Y1 restriction enzymes, separately, and large fragments were retrieved from gel fractions. *Hin*dIII and *Bst*YI digested fragments were used for DNA ligation into *Hin*dIII and *Bam*HI digested and dephosphorylated pIndigoBAC536 vectors [67], respectively.

Gene Copy Number Estimation

For Southern blot analysis, total genomic DNA was isolated from leaf tissue of *Panicum virgatum* L. var. Alamo clone AP13 as described [68]. Briefly, 1 g frozen leaf tissue was ground to fine powder using a pre-chilled mortar and pestle. Powdered leaf tissue was transferred to a 30 mL centrifuge tube containing 15 mL extraction buffer (100 mM Tris-HCl, pH 8.0; 50 mM EDTA, pH 8.0; 500 mM NaCl and 10 mM β-mercaptoethanol). After lysis with 20% sodium dodecyl sulphate (SDS), DNA was precipitated using isopropanol and treated with RNase A (30 µL of 10 mg/mL stock per sample) for 1–2 h at 37°C. The samples were extracted once with phenol:chloroform:isoamyl alcohol, followed by another extraction with chloroform:isoamyl alcohol only. DNA was precipitated with 0.1 volume of 3 M sodium acetate and 2.5 volumes of absolute ethanol for 1 h at −20°C.

Aliquots of genomic DNA (12 µg each) were digested with four different restriction enzymes (*Bam*HI, *Eco*RI, *Hin*dIII and *Sac*I) separately. Digested DNA samples were analyzed on 0.8% w/v agarose gel and blotted on nylon membrane (Hybond-N+^{TM}, Amersham Pharmacia Biotech Ltd.) by capillary transfer. To prepare probes, gene-specific primers were designed for known single copy genes from closely related genera (rice and maize) of the Poaceae family. The list of primers is given in File S10. DNA fragments, amplified using switchgrass DNA as a template, were labeled with alkaline phosphatase enzyme using Amersham Gene Images AlkPhos Direct Labeling and Detection System from GE Healthcare). Hybridizations and detection were performed according to manufacturer's instructions. In brief, approximately 5 ng probe was used per mL of hybridization buffer. Hybridizations of labeled DNA with membrane filters were performed overnight at 60°C in hybridization oven using hybridization bottles at 10 rpm. Primary washes were performed at 58°C for 20 min each. CDP-Star^{TM} chemiluminescent detection reagent was used for signal generation. Chemiluminescence was captured on an X-ray film, purchased from ISC-BioExpress USA and recorded using a document scanner.

Library Characterization

Approximately, 180 BAC clones were randomly selected from each library and inoculated to 2 mL overnight cultures of LB media containing 12.5 µg/mL chloramphenicol in 15 mL culture tubes. Cells were collected at 16,000 g for 10 min and BAC DNA was prepared using Qiagen's plasmid isolation kit. BAC DNA was digested with 10 U of *Not*I and analyzed on an agarose gel. Insert

size of BAC clones was estimated by comparing with the Lambda ladder PFG marker (New England Biolabs Inc.) as standard. High-density filter hybridizations were performed to check extra-nuclear DNA contamination and library coverage. Each filter contained 18,432 individual clones, arrayed in a 4×4 pattern in duplicate. Gene-specific DNA sequences (500–1000 bp in length) spanning through chloroplast (*trn*L, *rpo*B, *ndh*A and *rbc*L) and mitochondria (*atp*6, *atp*9, *cob* and *cox*1) genomes of rice/sorghum were used to design primers (File S10). The corresponding DNA sequences were amplified using switchgrass genomic DNA, labeled and used for filter hybridizations, as described earlier. The Clarke-Carbon equation [69], $N = \ln(1-P)/\ln(1-[I/GS])$, where N is the number of clones, GS is genome size and I is insert size, was used to calculate the theoretical probability of finding a sequence of interest among the BAC clones.

Full-length BAC Sequencing

Essentially full-length sequences for randomly selected BAC clones were obtained at the HudsonAlpha Institute of Biotechnology (www.hudsonalpha.org) by Sanger's method on ABI 3730XL DNA analyzers. The resulting trace data was base called using Phred V 0.020425. The Phred/Phrap/Consed suite of programs was used for assembling and editing the sequence [70,71,72]. After manual inspection of the assembled sequences, finishing was performed both by re-sequencing plasmid subclones and by primer walking on plasmid subclones or the BAC clone using custom primers. All finishing reactions were performed using dGTP BigDye Terminator Chemistry (Applied Biosystems). Hard-to-sequence gaps or small repeats were completed using small insert shatter libraries generated using Roche/454 sequencing technology or transposon libraries generated using Sanger technology.

BAC-End Sequencing (BES)

The BES reads were obtained by Sanger's method on ABI 3730XL capillary sequencing machines at the HudsonAlpha Institute of Biotechnology. The resulting trace data was base called using Phred V 0.020425 and vector sequences were masked using cross_match. Masked terminal vector sequences and BES less than 50 bp in length were removed. High quality sequences were then filtered for plant-organelle genomes-specific or *Escherichia coli*-specific sequences.

Analysis of Simple Sequence Repeats (SSRs) and Repeat Elements

We used mreps [73], a simple repeat identification software, to identify Simple Sequence Repeats (SSRs) from fasta-formatted unique BES. Parameters used were 1–3 nt repeats at least 12 nt in length and 4–6 nt repeats with at least 4 unit repetition. Other known repeat elements like TEs, rRNAs, centromere-/telomere-related sequences were identified with RepeatMasker 3.3.0 (http://www.repeatmasker.org/) [74,75] and AB-BLAST v3 (http://blast.advbiocomp.com/) using the Viridiplantae section of the RepBase repeat database (release 20110419) [76]. To identify novel repeat elements, switchgrass BES were masked with RepeatMasker 3.3.0 [75] and compared to themselves using MegaBlast (E-value = 10^{-50}). BES with at least six hits were analyzed using MEME V3.5.7 [77] to identify DNA motifs (E-value = 10^{-4}). Resulting putative switchgrass repeat elements (SREs) were queried in the RepBase repeat database (release April 2011) [76], MSU Plant Repeat Database release May 2009 [78], Triticeae repetitive sequence database (TREP) (release 10; http://wheat.pw.usda.gov/ITMI/Repeats/index.shtml), NCBI GenBank non-redundant nucleic acid sequence database (Release 184.0; http://www.ncbi.nlm.nih.gov/RefSeq/) and Swissprot database (release August 2011; http://www.ebi.ac.uk/uniprot/) with BLASTN and BLASTX under E-value cutoff of 10^{-4} to check for their uniqueness.

Functional Annotation and GO Analysis

Gene predictions from switchgrass BES was performed using Geneid v 1.4.4 [79] and PASA (http://pasa.sourceforge.net/). Predicted proteins were functionally annotated by comparison with Pfam database (version 25.0) using HMMER 3.0 [79]. GO terms were converted from Pfam domains using the mapping tool of the Gene Ontology project (http://www.geneontology.org/).

Comparative Mapping of BAC-end Sequences

To map BAC-end sequences onto grass genomes, the BES were first aligned to rice peptide sequences using BlastX. The equivalent regions in sorghum and *Brachypodium* were identified and used for mapping BES. All genome sequences were extracted from Phytozome (http://www.phytozome.net/). Best alignments were identified for each BES that placed above a base pair identity of 75% with e value <1e-20, and coverage of the BES >50%. Furthermore, a best placement for BES that aligned to multiple locations after applying the aforementioned screening criteria was determined by sorting the placements using the blast score. Pairs were identified with a maximum insert size of 500 KB. If only one side of the pair placed in coding sequences, then we performed a blast alignment of the mate on the nucleotide sequence of the whole rice region (equivalently in *Brachypodium* and sorghum) to find the mate. The syntenic relationship among genomes and mapping results are displayed using the Gbrowse-syn module [21].

Gene Annotations and Mapping of Full-length BAC Sequences

To produce high-quality non-redundant genomic sequences, repeat elements from full-length BAC sequences were masked using RepeatMasker 3.3.0 [75]. Gene models were identified using GenomeScan (http://genes.mit.edu/genomescan.html). Further PASA (http://pasa.sourceforge.net/) and NCBI EST sequences were used to update GenomeScan predictions. BLAST analysis to rice and *Arabidopsis* databases (http://rice.plantbiology.msu.edu/, http://www.arabidopsis.org/) and Pfam domain analysis (http://pfam.janelia.org/search) was performed to identify the conserved domains.

Genomic sequences of sorghum and *Brachypodium* were downloaded from Phytozome v6.0 (http://www.phytozome.net/), maize from MaizeSequence release 5b.60 (http://www.maizesequence.org) and of rice from MSU v6.1 (http://rice.plantbiology.msu.edu/). Discontinuous mega blast with a cutoff of $1e^{-20}$ was used to compare switchgrass gene models with other grass genomes. The microcolinearity among genomes was visually identified and displayed using Adobe Illustrator CS4. Direction of genes was determined using online databases (http://rice.plantbiology.msu.edu/; http://www.phytozome.net/).

The libraries and filters have been made available to the public through the Clemson University Genomics Institute (CUGI; www.genome.clemson.edu). Full-length BAC sequences for randomly selected 47 BAC clones have been submitted to GenBank under accession numbers AC243215–AC243261. GenBank accession numbers for BES are HR309496–HR503629 (Pv_ABa) and JM786703–JM972700 (Pv_ABb).

Supporting Information

Figure S1 Southern hybridizations for gene copy number estimations in switchgrass. We used Southern hybrid-

izations to determine the copy number, of single/low copy genes from closely related monocotyledonous plant species in switchgrass. The results of Southern hybridizations using four different restriction enzymes for each gene are presented.
(JPG)

File S1 Distribution of simple sequence repeats identified in switchgrass BAC-end sequences.
(XLS)

File S2 List of nucleotide sequence of novel switchgrass repetitive repeats (SREs).
(TXT)

File S3 Distribution of GO annotations with regard to A, Functional classes of gene products encoded from BAC end sequences; B, Biological processes associated with gene products and their C, cellular locations.
(XLS)

File S4 Distribution of simple sequence repeats and plant repeat elements identified from full-length BAC sequences.
(DOC)

File S5 List of 439 switchgrass gene loci (451 gene models) annotated from switchgrass full-length BAC sequences.
(XLS)

File S6 List of A, cDNA; B, genomic and C, protein sequences of switchgrass genes predicted from full-length BAC sequences.
(XLS)

File S7 List of switchgrass gene models with their corresponding orthologs from rice, sorghum, maize and *Brachypodium*.
(XLS)

File S8 List of genes from rice, sorghum, maize and *Brachypodium* that are not present in the corresponding regions in switchgrass in figure 8.
(XLSX)

File S9 Genome characteristics of various plant genomes based upon BAC-end or genome sequence data.
(XLS)

File S10 List of primers used for BAC library characterization and Southern hybridizations.
(DOC)

Acknowledgments

We would like to thank Uffe Hellsten and Kerry Berry (JGI) for sequencing and technical advice and Malay Saha (Nobel Foundation), Michael Udvardi, Jiyi Zhang and Yuhong Tang (BESC, Nobel Foundation) for providing AP13 material. We thank Christopher Saski at Clemson University for construction of the BAC libraries.

Author Contributions

Conceived and designed the experiments: MKS RS LEB PCR. Performed the experiments: MKS RS PC JG JS PC JJ MQ. Analyzed the data: MKS RS PC JG JS PC JJ MQ. Contributed reagents/materials/analysis tools: DR. Wrote the paper: MKS RS PCR.

References

1. Sanderson MA, Reed RL, McLaughlin SB, Wullschleger SD, Congcr BV, et al. (1996) Switchgrass as a sutainable bioenergy crop. Bioresource Technology 56: 83–93.
2. Schmer MR, Vogel KP, Mitchell RB, Perrin RK (2008) Net energy of cellulosic ethanol from switchgrass. Proc Natl Acad Sci U S A 105: 464–469.
3. McLaughlin SB, Kszos LA (2005) Development of switchgrass (*Panicum virgatum*) as a bioenergy feedstock in the United States. Biomass and Bioenergy 28: 515–535.
4. Parrish DJ, Fike JH (2005) The Biology and Agronomy of Switchgrass for Biofuels. Critical Rev Plant Sci 24: 423–459.
5. Bouton J (2008) Improvement of switchgrass as a bioenergy crop. W. Vermerris, ed. Genetic Improvement of Bioenergy Crops, Springer New york. pp 309–345.
6. Fike J, Parrish D, Wolf D, Balasko J, Green J, et al. (2006) Long-term yield potential of switchgrass-for-biofuel systems. Biomass Bioenergy 30: 198–206.
7. Sanderson MA, Adler PR, Boateng AA, Casler MD, Sarath G (2006) Switchgrass as a biofuels feedstock in the USA. CANADIAN JOURNAL OF PLANT SCIENCE 86: 1315–1325.
8. Rubin EM (2008) Genomics of cellulosic biofuels. Nature 454: 841–845.
9. Wright L Historical perspective on how and why switchgrass was selected as "Model" high-potential energy crop. http://wwwostigov/bridge.
10. Okada M, Lanzatella C, Saha MC, Bouton J, Wu R, et al. (2010) Complete Switchgrass Genetic Maps Reveal Subgenome Collinearity, Preferential Pairing, and Multilocus Interactions. Geneticsdoi:10.1534/genetics.110.113910.
11. Cheung F, Town CD (2007) A BAC end view of the Musa acuminata genome. BMC plant biology 7: 29.
12. Febrer M, Cheung F, Town CD, Cannon SB, Young ND, et al. (2007) Construction, characterization, and preliminary BAC-end sequencing analysis of a bacterial artificial chromosome library of white clover (Trifolium repens L.). Genome/National Research Council Canada = Genome/Conseil national de recherches Canada 50: 412–421.
13. Gonzalez VM, Rodriguez-Moreno L, Centeno E, Benjak A, Garcia-Mas J, et al. (2010) Genome-wide BAC-end sequencing of Cucumis melo using two BAC libraries. BMC Genomics 11: 618.
14. Han Y, Korban SS (2008) An overview of the apple genome through BAC end sequence analysis. Plant molecular biology 67: 581–588.
15. Faivre Rampant P, Lesur I, Boussardon C, Bitton F, Martin-Magniette ML, et al. (2011) Analysis of BAC end sequences in oak, a keystone forest tree species, providing insight into the composition of its genome. BMC Genomics 12: 292.
16. Jeukens J, Boyle B, Kukavica-Ibrulj I, St-Cyr J, Levesque RC, et al. (2011) BAC library construction, screening and clone sequencing of lake whitefish (Coregonus clupeaformis, Salmonidae) towards the elucidation of adaptive species divergence. Molecular Ecology Resources 11: 541–549.
17. Saski CA, Li Z, Feltus FA, Luo H (2011) New genomic resources for switchgrass: a BAC library and comparative analysis of homoeologous genomic regions harboring bioenergy traits. BMC Genomics 12: 369.
18. Missaoui AM, Paterson AH, Bouton JH (2005) Investigation of genomic organization in switchgrass (*Panicum virgatum* L.) using DNA markers. Theor Appl Genet 110: 1372–1383.
19. Tobias CM (2009) A genome may reduce your carbon footprint. Plant Genome 2: 5–8.
20. Sheridan RP, Venkataraghavan R (1992) A systematic search for protein signature sequences. Proteins 14: 16–28.
21. McKay SJ, Vergara IA, Stajich JE (2010) Using the Generic Synteny Browser (GBrowse_syn). Current protocols in bioinformatics/editoral board, Andreas D Baxevanis [et al] Chapter 9: Unit 9 12.
22. Schatz MC, Delcher AL, Salzberg SL (2010) Assembly of large genomes using second-generation sequencing. Genome research 20: 1165–1173.
23. *C elegans* Sequencing Consortium (1998) Genome sequence of the nematode C. elegans: a platform for investigating biology. Science 282: 2012–2018.
24. Huang S, Li R, Zhang Z, Li L, Gu X, et al. (2009) The genome of the cucumber, Cucumis sativus L. Nature genetics 41: 1275–1281.
25. Velasco R, Zharkikh A, Affourtit J, Dhingra A, Cestaro A, et al. (2010) The genome of the domesticated apple (Malus×domestica Borkh.). Nature genetics 42: 833–839.
26. Velasco R, Zharkikh A, Troggio M, Cartwright DA, Cestaro A, et al. (2007) A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. PloS one 2: e1326.
27. Shizuya H, Birren B, Kim UJ, Mancino V, Slepak T, et al. (1992) Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector. Proc Natl Acad Sci U S A 89: 8794–8797.
28. Tomkins JP, Yu Y, Miller-Smith H, Frisch DA, Woo SS, et al. (1999) A bacterial artificial chromosome library for sugarcane. Theor Appl Genet 99: 419–424.
29. Venter JC, Smith HO, Hood L (1996) A new strategy for genome sequencing. Nature 381: 364–366.
30. Wang GL, Holsten TE, Song WY, Wang HP, Ronald PC (1995) Construction of a rice bacterial artificial chromosome library and identification of clones linked to the Xa-21 disease resistance locus. Plant J 7: 525–533.
31. Woo SS, Jiang J, Gill BS, Paterson AH, Wing RA (1994) Construction and characterization of a bacterial artificial chromosome library of Sorghum bicolor. Nucleic Acids Res 22: 4922–4931.

32. Ammiraju JS, Luo M, Goicoechea JL, Wang W, Kudrna D, et al. (2006) The Oryza bacterial artificial chromosome library resource: construction and analysis of 12 deep-coverage large-insert BAC libraries that represent the 10 genome types of the genus Oryza. Genome Res 16: 140–147.
33. Liang H, Fang EG, Tomkins JP, Luo M, Kudrna D, et al. (2007) Development of a BAC library for yellow-poplar (*Liriodendron tulipifera*) and the identification of genes associated with flower development and lignin biosynthesis. Tree Genetics Genomes 3: 215–225.
34. Marek LF, Shoemaker RC (1997) BAC contig development by fingerprint analysis in soybean. Genome 40: 420–427.
35. Buckler AJ, Chang DD, Graw SL, Brook JD, Haber DA, et al. (1991) Exon amplification: a strategy to isolate mammalian genes based on RNA splicing. Proc Natl Acad Sci U S A 88: 4005–4009.
36. Tomkins JP, Davis G, Main D, Yim Y, Duru N, et al. (2002) Construction and characterization of a deep-coverage bacterial artiWcial chromosome library for maize. Crop Sci 42: 928–933.
37. Hong CP, Lee SJ, Park JY, Plaha P, Park YS, et al. (2004) Construction of a BAC library of Korean ginseng and initial analysis of BAC-end sequences. Mol Genet Genomics 271: 709–716.
38. Adam-Blondon A-F, Bernole A, Faes G, Lamoureux D, Pateyron S, et al. (2005) Construction and characterization of BAC libraries from major grapevine cultivars. Theor Appl Genet 110: 1363–1371.
39. Huo N, Gu YQ, Lazo GR, Vogel JP, Coleman-Derr D, et al. (2006) Construction and characterization of two BAC libraries from Brachypodium distachyon, a new model for grass genomics. Genome/National Research Council Canada 49: 1099–1108.
40. Agarwal M, Shrivastava N, Padh H (2008) Advances in molecular marker techniques and their applications in plant sciences. Plant Cell Rep 27: 617–631.
41. Cavagnaro PF, Senalik DA, Yang L, Simon PW, Harkins TT, et al. (2010) Genome-wide characterization of simple sequence repeats in cucumber (Cucumis sativus L.). BMC Genomics 11: 569.
42. Morgante M, Hanafey M, Powell W (2002) Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nat Genet 30: 194–200.
43. Jayashree B, Punna R, Prasad P, Bantte K, Hash CT, et al. (2006) A database of simple sequence repeats from cereal and legume expressed sequence tags mined in silico: survey and evaluation. In Silico Biol 6: 607–620.
44. Bennetzen JL (2000) Transposable element contributions to plant gene and genome evolution. Plant molecular biology 42: 251–269.
45. Flowers JM, Purugganan MD (2008) The evolution of plant genomes: scaling up from a population perspective. Current opinion in genetics & development 18: 565–570.
46. The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408: 796–815.
47. Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, et al. (2009) The Sorghum bicolor genome and the diversification of grasses. Nature 457: 551–556.
48. Wang J, Roe B, Macmil S, Yu Q, Murray JE, et al. (2010) Microcollinearity between autopolyploid sugarcane and diploid sorghum genomes. BMC Genomics 11: 261.
49. Paterson AH, Lan TH, Reischmann KP, Chang C, Lin YR, et al. (1996) Toward a unified genetic map of higher plants, transcending the monocot-dicot divergence. Nature genetics 14: 380–382.
50. Grandbastien MA (2004) [Stress activation and genomic impact of plant retrotransposons]. J Soc Biol 198: 425–432.
51. Barow M, Meister A (2002) Lack of correlation between AT frequency and genome size in higher plants and the effect of nonrandomness of base sequences on dye binding. Cytometry 47: 1–7.
52. Fullerton SM, Bernardo Carvalho A, Clark AG (2001) Local rates of recombination are positively correlated with GC content in the human genome. Molecular biology and evolution 18: 1139–1142.
53. Musto H, Naya H, Zavala A, Romero H, Alvarez-Valin F, et al. (2004) Correlations between genomic GC levels and optimal growth temperatures in prokaryotes. FEBS letters 573: 73–77.
54. Smarda P, Bures P, Horova L, Foggi B, Rossi G (2008) Genome size and GC content evolution of Festuca: ancestral expansion and subsequent reduction. Annals of botany 101: 421–433.
55. Galtier N, Piganeau G, Mouchiroud D, Duret L (2001) GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. Genetics 159: 907–911.
56. Kawabe A, Miyashita NT (2003) Patterns of codon usage bias in three dicot and four monocot plant species. Genes & genetic systems 78: 343–352.
57. Rabinowicz PD, Sachidanandam R (2002) Genomics: more than the sum of the parts. Genome research 12: 1015–1016.
58. Huo N, Vogel JP, Lazo GR, You FM, Ma Y, et al. (2009) Structural characterization of Brachypodium genome and its syntenic relationship with rice and wheat. Plant molecular biology 70: 47–61.
59. Lai J, Ma J, Swigonova Z, Ramakrishna W, Linton E, et al. (2004) Gene loss and movement in the maize genome. Genome research 14: 1924–1931.
60. Devos KM (2005) Updating the 'crop circle'. Curr Opin Plant Biol 8: 155–162.
61. Devos KM, Gale MD (2000) Genome relationships: the grass model in current research. Plant Cell 12: 637–646.
62. Tobias CM, Sarath G, Twigg P, Lindquist E, Pangilinan J, et al. (2008) Comparative Genomics in Switchgrass Using 61,585 High-Quality Expressed Sequence Tags. The Plant Genome 1: 111–124.
63. Paterson AH, Bowers JE, Burow MD, Draye X, Elsik CG, et al. (2000) Comparative genomics of plant chromosomes. Plant Cell 12: 1523–1540.
64. Bennetzen JL (2000) Comparative sequence analysis of plant nuclear genomes:m microcolinearity and its many exceptions. The Plant cell 12: 1021–1029.
65. Feuillet C, Keller B (2002) Comparative genomics in the grass family: molecular characterization of grass genome structure and evolution. Annals of botany 89: 3–10.
66. Kellogg EA (2001) Evolutionary history of the grasses. Plant Physiol 125: 1198–1205.
67. Luo M, Wing RA (2003) An improved method for plant BAC library construction. Methods Mol Biol 236: 3–20.
68. Dellaporta SL, Wood J, Hicks JB (1983) A plant DNA minipreparation: version I. Plant Mol Biol Rep 1: 19–21.
69. Clark L, Carbon J (1976) A colony bank containing synthetic Col E1 hybrids representative of the entire *E. coli* genome. Cell 9: 91–99.
70. Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome research 8: 186–194.
71. Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome research 8: 175–185.
72. Gordon D, Abajian C, Green P (1998) Consed: a graphical tool for sequence finishing. Genome research 8: 195–202.
73. Kolpakov R, Bana G, Kucherov G (2003) mreps: Efficient and flexible detection of tandem repeats in DNA. Nucleic acids research 31: 3672–3678.
74. Smit A, Hubley R, Green P RepeatMasker Open-3.0. 1996–2010.
75. Tarailo-Graovac M, Chen N (2009) Using RepeatMasker to identify repetitive elements in genomic sequences. Current protocols in bioinformatics/editoral board, Andreas D Baxevanis [et al] Chapter 4: Unit 4 10.
76. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, et al. (2005) Repbase Update, a database of eukaryotic repetitive elements. Cytogenetic and genome research 110: 462–467.
77. Bailey TL, Boden M, Whitington T, Machanick P (2010) The value of position-specific priors in motif discovery using MEME. BMC bioinformatics 11: 179.
78. Ouyang S, Buell CR (2004) The TIGR Plant Repeat Databases: a collective resource for the identification of repetitive sequences in plants. Nucleic acids research 32: D360–363.
79. Blanco E, Parra G, Guigó R (2003) Using geneid to identify genes. Current protocols in bioinformatics. pp 4.3.1–4.3.26.