VAGNER KATSUMI OKURA

Sintenia genomica entre sorgo e cana-de-açúcar inferida a partir do sequenciamento de um pool de BACs

Genomic synteny between sorghum and sugarcane inferred from a BAC pool sequencing

CAMPINAS
2015

VAGNER KATSUMI OKURA

SINTENIA GENOMICA ENTRE SORGO E CANA-DE-AÇÚCAR INFERIDA A PARTIR DO SEQUENCIAMENTO DE UM POOL DE BACS

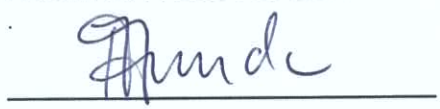
GENOMIC SYNTENY BETWEEN SORGHUM AND SUGARCANE INFERRED FROM A BAC POOL SEQUENCING

Tese apresentada ào Instituto de Biologia da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Doutor em Genética e Biologia Molecular, na área de Bioinformática.

Thesis presented to the Institute of Biology, State University of Campinas in partial fulfillment of the requirements for the degree of Doctor in Genetics and Molecular Biology, on the area of Bioinformatics.

ORIENTADOR/SUPERVISOR: PROF. DR. PAULO ARRUDA

ESTE EXEMPLAR CORRESPONDE À VERSÃO
FINAL DA TESE DEFENDIDA PELO
ALUNO VAGNER KATSUMI OKURA, E ORIENTADA PELO
PROF. DR. PAULO ARRUDA



CAMPINAS

2015

Agência(s) de fomento e nº(s) de processo(s): Não se aplica.

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Biologia
Mara Janaina de Oliveira - CRB 8/6972

Ok7s

Okura, Vagner Katsumi, 1973-
Sintenia genomica entre sorgo e cana-de-açúcar inferida a partir do sequenciamento de um pool de BACs / Vagner Katsumi Okura. – Campinas, SP : [s.n.], 2015.

Orientador: Paulo Arruda.
Tese (doutorado) – Universidade Estadual de Campinas, Instituto de Biologia.

1. Cromossomos artificiais bacterianos. 2. Cana-de-açúcar. 3. Sorgo. 4. Sequenciamento de DNA. 5. Genoma de planta. I. Arruda, Paulo,1952-. II. Universidade Estadual de Campinas. Instituto de Biologia. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: Genomic synteny between sorghum and sugarcane inferred from a BAC pool sequencing
Palavras-chave em inglês:
Chromosomes, Artificial, Bacterial
Sugarcane
Sorghum
DNA sequencing
Genome, Plant
Área de concentração: Bioinformática
Titulação: Doutor em Genética e Biologia Molecular
Banca examinadora:
Paulo Arruda [Orientador]
Claudia Teixeira Guimarães
João Paulo Fumio Whitaker Kitajima
Guilherme Pimentel Teles
Michel Eduardo Beleza Yamagishi
Data de defesa: 28-08-2015
Programa de Pós-Graduação: Genética e Biologia Molecular

Campinas, **28** de **agosto** de **2015**

BANCA EXAMINADORA

Prof. Dr. Paulo Arruda (orientador) _____
Assinatura

Dra. Claudia Teixeira Guimarães _____
Assinatura

Dr. João Paulo Fumio Whitaker Kitajima _____
Assinatura

Prof. Dr. Guilherme Pimentel Telles _____
Assinatura

Dr. Michel Eduardo Beleza Yamagishi _____
Assinatura

Dr. Francisco Pereira Lobo _____
Assinatura

Dr. Renato Vicentini Dos Santos _____
Assinatura

Dr. Plinio Tadeu Cristofoletti Junior _____
Assinatura

RESUMO

O sequenciamento genômico de plantas tem se acelerado nos últimos anos principalmente devido ao avanço das tecnologias de sequenciamento de nova geração, capazes de gerar um grande volume de dados com custo cada vez menor. No entanto, o sequenciamento e a montagem de genomas de plantas ainda continua sendo um grande desafio em função da alta complexidade desses genomas que na sua grande maioria possuem alto grau de ploidia e grande proporção de sequências repetitivas. O sequenciamento de bibliotecas produzidas com DNA genômico de plantas clonados em vetores BACs (*bacterial artificial chromosomes*) pode ser uma estratégia efetiva para sequenciamento de genomas complexos, por dividir a tarefa de montagem em problemas menores. No geral, bibliotecas de BACs contém fragmentos de DNA de 100 a 200 kilobases, cujo conjunto cobre o genoma clonado várias vezes. Entretanto, mesmo com as novas tecnologias de sequenciamento, o custo de sequenciar bibliotecas de BACs ainda é alto, pois na maioria das vezes o sequenciamento é realizado a partir do DNA isolado de cada BAC individualmente. Uma alternativa seria sequenciar pools contendo centenas de BACs amostrados randomicamente, que dessa forma diminuiria o custo proporcionalmente ao número de BACs do pool. Neste trabalho, desenvolvemos um modelo para sequenciamento e montagem de pools de BACs de uma biblioteca preparada a partir de uma variedade comercial de cana-de-açúcar. Como resultado, um pool com 178 BACs de cana-de-açúcar da variedade SP80-3280 foi sequenciado utilizando-se as tecnologias HighSeq2000 da Illumina e PacBio, e montados utilizando diferentes conjuntos de softwares. Por ser uma amostra de BACs selecionados randomicamente foi possível montar 2.451 scaffolds correspondentes a 88,2% do tamanho estimado total do conjunto de BACs do pool. A completeza da montagem foi verificada de várias maneiras incluindo a análise do número de BACs montados com tamanho esperado, a comparação com BACs depositados no NCBI e pela colinearidade e ordem de genes presentes entre scaffolds de cana e os cromossomos de sorgo. Os scaffolds com tamanho superior a 2 kb foram alinhados contra o genoma de sorgo, e no geral os alinhamentos mostraram uma distribuição uniforme ao longo dos 10 cromossomos do sorgo indicando a aleatoriedade da

amostragem. Pela análise sintênica entre os scaffolds de cana e os cromossomos de sorgo, observamos que o genoma monoploide da cana parece ser mais contraído em relação ao genoma do sorgo. No geral o trabalho mostrou que é possível sequenciar pool de BACs de genomas de plantas de alta complexidade como o genoma de cana-de-açúcar com altos níveis de ploidia.

ABSTRACT

The genomic sequencing of plants has accelerated in recent years mainly due to the advances of next generation sequencing technologies capable of generating a high volume of data with ever lower cost. However, the sequencing and assembly of plant genomes remains a major challenge due to the high complexity of these genomes that mostly have a high degree of ploidy and large proportion of repetitive sequences. The sequencing of libraries produced with genomic DNA of plants cloned into BAC (bacterial artificial chromosome) vectors can be an effective strategy for sequencing complex genomes, by breaking down the assembly task into smaller problems. Typical BAC libraries contain DNA fragments of 100 to 200 kilobases which together cover the genome cloned several times. However, even with the new sequencing technologies, the cost of sequencing BACs libraries is still high because most of the times the sequencing is individually performed from the isolated DNA of each BAC. An alternative would be the sequencing of pools containing hundreds of randomly sampled BACs, which thereby would decrease the cost in proportion to the number of BACs pooled. In this work we developed a model for sequencing and assembly BAC pools of a library prepared from a commercial sugarcane variety. As a result, a pool of 178 BACs from sugarcane variety SP80-3280 was sequenced using the technologies of the Illumina HighSeq2000 and PacBio and was assembled using different sets of softwares. Being a sample of randomly selected BACs was possible to assemble 2,451 scaffolds corresponding to 88.2% of the estimated total pool size set of BACs. The completeness of the assembly was verified in many ways including the analysis of the number of BACs assembled with expected size, comparison to sugarcane BACs deposited in NCBI and by the collinearity and gene order presented between sugarcane scaffolds and sorghum chromosomes. Scaffolds larger than 2 kb were aligned to the sorghum genome, and in general, alignments showed a uniform distribution over the 10 sorghum chromosomes indicating the randomness of sampling. By syntenic analysis between sugarcane scaffolds and sorghum chromosomes, we found that the monoploid sugarcane genome seems to be more contracted compared to the genome of sorghum. Overall the study showed that it is possible to sequence BAC pools from plant genomes with high complexity like the sugarcane genome with high level of ploidy.

AGRADECIMENTOS

Primeiramente, agradeço ao Prof. Paulo Arruda, pela oportunidade de desenvolver meu projeto de doutorado, pela confiança, pelo apoio e pela visão sempre otimista.

Agradeço o pessoal do laboratório, Ana Camargo, Eduardo Kiyota, Izabella Pena, Jader Armanhi, Laura Migliorini, Lucas Adjafre, Natalia Damasceno, Pedro Barreto, Renato Maia, Sandra Queiroz, Thais Figueira, Vinicius Almeida, pela convivência, pelas discussões nas reuniões e colaborações.

Ao LaCTAD, onde foram gerados os dados de sequenciamento Illumina, em particular a Suzely e Leonardo pela ajuda com a preparação dos BACs.

Ao Rafael Souza, pela sugestões e discussão sobre o meu projeto, e por ajudar na preparação do material para sequenciamento PacBio.

A FAPESP, pelo custeio dos dados de sequenciamento.

Aos amigos Adhemar, Chico, Dante, Felipe, Isabel, Juliana, Natalia, pela companhia durante os almoços, e pelas dicas e sugestões.

Aos membros da banca, pela disponibilidade e sugestões para enriquecimento do meu trabalho.

SUMÁRIO

INTRODUÇÃO

Sequenciamento de genoma de plantas

O sequenciamento genômico das plantas tem se tornado cada vez mais uma ferramenta essencial tanto para a pesquisa básica como para a pesquisa aplicada voltada para o desenvolvimento de variedades mais produtivas e resistentes a estresses bióticos e abióticos. Com o genoma sequenciado, é possível acessar a estrutura física dos cromossomos, identificar os genes tanto na sua porção codificadora como na regulatória, estudar os níveis de expressão a nível transcriptômico de plantas submetidas a estresses e assim entender melhor os mecanismos genéticos responsáveis pela fisiologia e desenvolvimento das plantas submetidas as mais diversas condições bióticas e abióticas. Arabidopsis thaliana [1] foi a primeira planta a ter seu genoma sequenciado devido ao tamanho reduzido do seu genoma e da baixa quantidade de DNA repetitivo. O sequenciamento do genoma da Arabidopsis, feito ainda com a tecnologia Sanger de baixo rendimento, representou um grande avanço para o conhecimento da estrutura do genoma de plantas para o desenvolvimento de estratégias de sequenciamento e de ferramentas de bioinformática utilizadas na montagem do genoma. A partir do sequenciamento da Arabidopsis as tecnologias desenvolvidas foram aplicadas para o sequenciamento de genomas de plantas economicamente importantes como o arroz [2], uva [3], sorgo [4], milho [5] e soja [6]. Em todos esses casos foram utilizadas as estratégias baseadas na primeira tecnologia de sequenciamento Sanger.

Sequenciamento genômico utilizando sequenciamento de próxima geração (NGS - Next Generation Sequencing)

Nos últimos 10 anos foram desenvolvidas várias tecnologias de sequenciamento massivo de genomas incluindo as tecnologias 454, Illumina e SOLiD [7]. Essas tecnologias possibilitaram a geração de grande volume de sequencias a um custo bastante reduzido. Além do grande volume de dados gerados, essas tecnologias dispensam as tarefas dispendiosas de clonagem. As estratégias de sequenciamento são baseadas essencialmente na preparação de bibliotecas obtidas por fragmentação randômica do DNA genômico, gerando fragmentos para sequenciamento "*paired-end* ou *mate pair"*, com tamanhos variando entre 170 e

20.000 pares de bases (bp). Porém, o tamanho dos reads gerados nestas novas plataformas é pequeno. Enquanto que a tecnologia Sanger gera reads de até 1.000 bp, os reads 454 tem até 700 bp, os reads Illumina tem até 300 bp e os reads SOLiD tem até 75 bp. Com o sequenciamento das duas pontas ods fragmentos de tamanhos variados, é feita a montagem de scaffolds que, pelo grande número, acabam cobrindo o genoma original até 1.000 vezes, dependendo do tamanho do genoma a ser sequenciado. Assim, essas novas tecnologias de sequenciamento, associadas a estratégia de fragmentação randômica do DNA (shotgun) permitiram o sequenciamento de genomas de eucariotos, no que se convencionou denominar "*whole genome sequencing*". Essas abordagens tornaram viáveis o sequenciamento genômico de eucariotos em termos de custo, rapidez e facilidade de preparação das bibliotecas de sequenciamento. Com isso, o número de organismos sequenciados tem aumentado continuamente, em particular genomas de plantas. Basicamente, na estratégia "shotgun", bibliotecas com diferentes tamanhos de fragmentos do genoma são preparadas e sequenciadas, e os reads gerados são montados para reconstruir a sequência original do genoma. Com maior capacidade de sequenciamento e menor custo por base, a plataforma Illumina foi utilizada para sequenciar os genomas da laranja [8] e do pimenteiro [9]. No entanto, devido a complexidade, a maior parte das plantas cujos genomas já foram sequenciados utilizaram mais de uma plataforma de sequenciamento. O genoma do algodão [10], por exemplo, foi sequenciado utilizando-se as plataformas Illumina e Sanger enquanto o genoma da seringueira [11] foi sequenciado utilizando-se as plataformas 454, Illumina e SOLiD. Os genomas do melão [12], da beterraba [13] e da cevada [14] foram sequenciados com 454, Illumina e Sanger. A utilização de múltiplas plataformas de sequenciamento pode ser explicada pelas dificuldades decorrentes do tamanho menor dos reads Illumina e SOLiD, e pelo tamanho limitado dos fragmentos das bibliotecas *paired end*, fatores importantes para lidar com longas regiões repetitivas na montagem. Mais recentemente foi lançada uma nova plataforma, chamada de terceira geração das tecnologias de sequenciamento, a plataforma PacBio, desenvolvida para produzir reads muito longos variando entre 1 kb e 20 kb de comprimento. Porém, os reads produzidos pela tecnologia PacBio apresentam maior taxa de erros. Além disso, a tecnologia PacBio tem menor capacidade de sequenciamento, pois produzem uma quantidade menor de bases por corrida.

Reads longos tendem a facilitar bastante as montagens de genomas com alta taxa de conteúdo repetitivo, característica comum em plantas. Por enquanto, não há nenhuma planta com genoma sequenciado e publicado com esta útlima tecnologia de sequenciamento.

Plantas possuem genomas complexos

Mesmo com a evolução das tecnologias de sequenciamento, é relativamente pequeno o número de plantas sequenciadas até o momento (http://www.ncbi.nlm.nih.govs/genome). A principal razão para isso está relacionada a estrutura e organização do genoma das plantas. Estas apresentam tamanho maior em relação genomas de animais, tem alta proporção de sequências repetitivas além de apresentarem alto nível de ploidia (número de cromossomos homólogos em uma célula). Estas características das plantas tornam o processo de sequenciamento e montagem do genoma bem mais desafiador. Isso pode ser observado com as plantas sequenciadas recentemente [8-14], que apesar de terem genomas menores e serem diplóides, as montagens resultantes foram em geral bem mais fragmentadas, com número maior de contigs/scaffods de tamanho menor. Associada a isso, está a limitação da estratégia de sequenciamento shotgun. Mesmo empregando as novas tecnologias de sequenciamento para lidar com genomas de tamanho maior e com certo grau de sequências repetitivas, o uso da estratégia shotgun para genomas de plantas poliplóides é um desafio. Devido a arquitetura genômica, com múltiplas cópias de cromossomos compartilhando regiões similares, a montagem usando sequenciamento shotgun resultará em contigs colapsados destas regiões, dificultando a montagem de sequências consenso mais longas. Um alternativa para trabalhar com genomas poliploides é usar a estratégia de sequenciamento baseada em BACs (do inglês *Bacterial Artificial Chromosomes*), que visa diminuir a complexidade da montagem em problemas menores, produzindo montagens com maior qualidade. O genoma de Arabidopsis [1], considerado uma boa referência de montagem, e os genomas de arroz [2] e de milho [5] foram sequenciados usando BACs.

Sequenciamento de BACs

A estratégia de sequenciamento baseada em BACs, também conhecida como shotgun hierárquico, é baseada em três etapas (Figura 1). A primeira etapa consiste na preparação de uma biblioteca de BACs a partir do DNA genômico da planta. Fragmentos de DNA genômico com tamanhos variando entre 100-200kb são gerados randomicamente a partir de digestão parcial utilizando-se uma enzima de restrição. Em seguida, fragmentos na faixa de tamanho desejada, são selecionados e clonados em vetores BACs. Na segunda etapa, há uma seleção dos BACs para sequenciamento, com base no mapa físico do genoma, que estabelece uma ordem para BACs. Mapas físicos são geralmente construídos usando enzimas de restrição (mapeamento por restrição), mas há também as técnicas de FISH (*Fluorescent in situ hybridization*) e STS (*Sequence tagged site*) [15]. No mapeamento por restrição, BACs são digeridos com diferentes enzimas de restrição e os fragmentos gerados são separados por eletroforese em gel de acrilamida de alta resolução. Em seguida os padrões dos tamanhos dos fragmentos de restrição são analisados para estabelecer a sobreposição entre os BACs e assim determinar a ordem entre os BACs ao longo dos cromossomos. Por fim, é feita uma seleção de um conjunto mínimo de clones que cobrem o genoma (*minimum tiling path).* Na terceira etapa, os BACs selecionados são individualmente sequenciados. No geral são preparadas bibliotecas shotgun a partir do DNA isolado de cada BAC individualmente. Os reads shotgun são então montados utilizando-se ferramentas de bioinformática (montadores ou *assemblers*), determinando-se a sequência de bases de cada BAC. As sequências dos BACs são então sobrepostas para gerar a montagem do genoma.

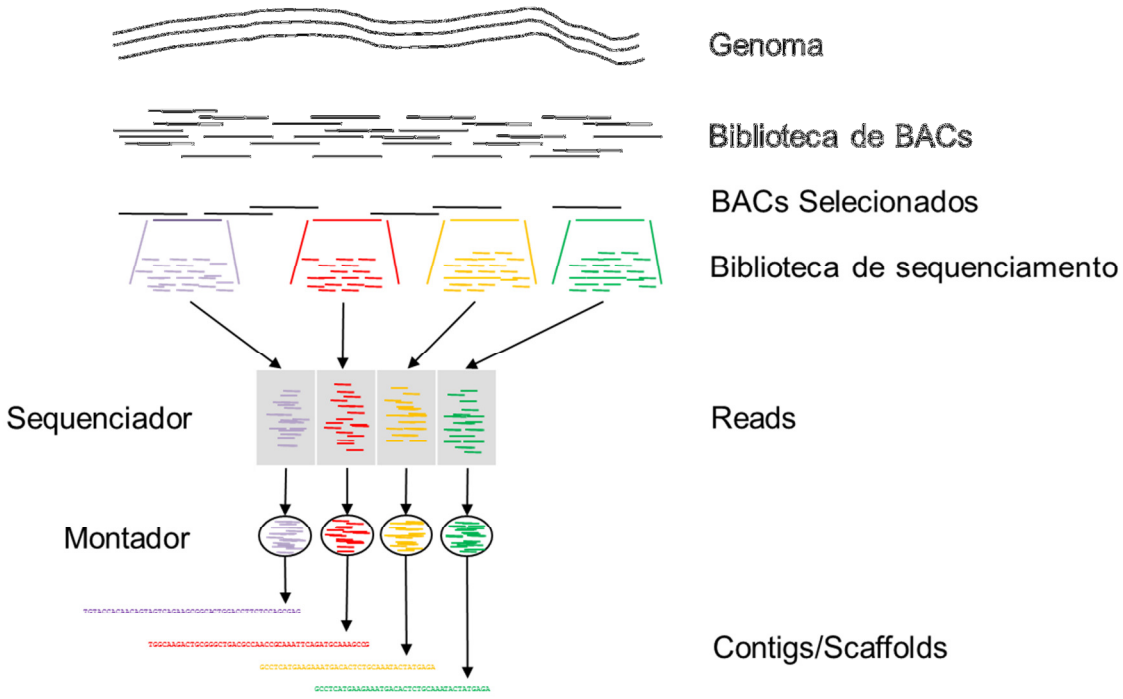


Figura 1: Esquema da estratégia de sequenciamento baseada em BACs.

Sequenciamento de pool de BACs

Embora a estratégia de sequenciamento de BACs facilite a montagem de genomas mais complexos, possui grande desvantagem de ter um custo elevado de esforço e tempo, associado a dispendiosas tarefas laboratoriais envolvidas na construção dos mapas físicos e na preparação dos BACs para sequenciamento. Mesmo com as novas tecnologias de sequenciamento, o custo de sequenciar BACs ainda é alto, devido a grande quantidade de BACs que devem ser sequenciados para compor um genoma de planta com um mínimo confiável de cobertura. Isso tem limitado o número de genoma de plantas sequenciados usando esta estratégia. Uma alternativa é sequenciar pool de BACs (Figura 2), onde um número de clones é misturado, sem uso de *tags* ou *barcodes* de individualização de amostras, para compor uma biblioteca de sequenciamento, não havendo neste caso a relação entre os reads gerados e os BACs. Neste sentido, haveria uma diminuição na quantidade de bibliotecas de sequenciamento e um aumento no número BACs por corrida, otimizando o capacidade dos sequenciadores, gerando uma diminuição do custo. Porém, o aumento do número de BACs por pool eleva-se a complexidade da montagem. Esta abordagem usando pool de BACs já foi testada para alguns organismos. Em arroz [16], seis pools, cada um contendo 28 BACs, foram

sequenciados usando a plataforma 454. Cada pool correspondia a um trecho de 3 Mb de uma região de 18 Mb, selecionada a partir do mapa físico. Para o genoma do salmão [17], foi sequenciado com a plataforma 454 um pool contendo oito BACs, correpondendo a um trecho de 1 Mb. Com o genoma do melão [18], um pool com 23 BACs e um pool com 35 BACs foram sequenciados com a platforma 454. Do total de 57 BACs, 50 tiveram a sequência completa montada.

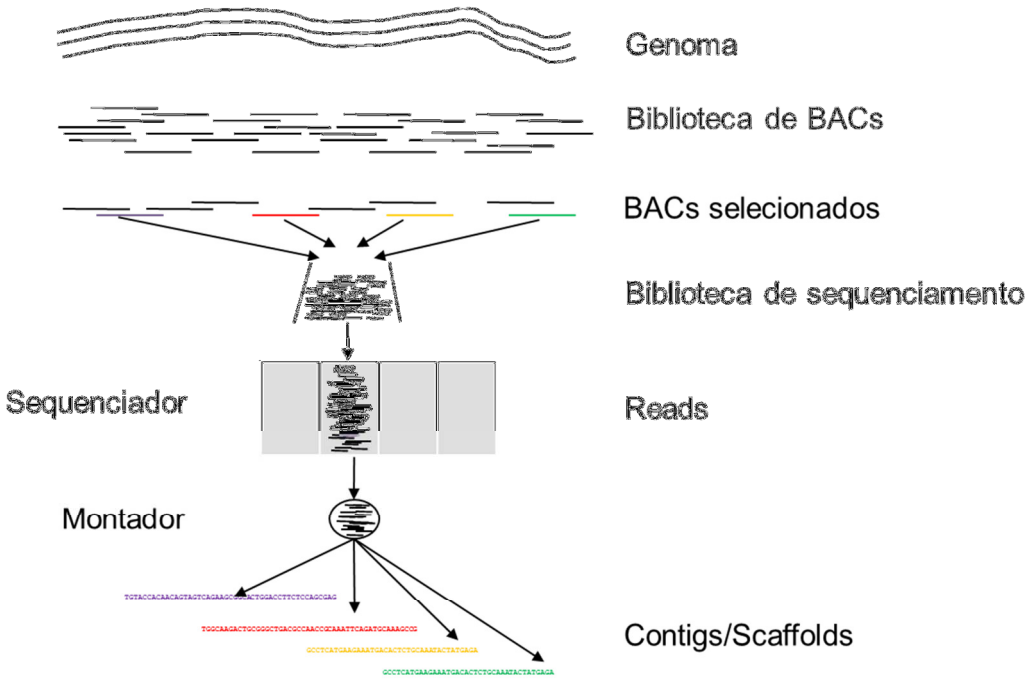


Figura 2: Esquema da abordagem de sequenciamento de pool de BACs.

Objetivos

Os objetivos deste trabalho foram:

- Desenvolver um modelo para sequenciamento e montagem de pools de BACs.
- Testar o modelo sequenciando um pool de constituído de BACs amostrados aleatoriamente de uma biblioteca de cana-de-açúcar.
- Validar os contigs e scaffolds gerados utilizando ferramentas de bioinformática, alinhamento sintênico com o genoma do sorgo e completude de sequências oriundas de regiões codificadoras.

Resultados

Os resultados da tese estão divididos em 3 capítulos.

O Capítulo 1 apresenta o manuscrito referente a construção e análises preliminares da bliblioteca de BACs de cana-de-açúcar da variedade comercial SP80-3280.

O Capítulo 2 descreve as simulações realizadas com número crescente de BACs por pool, para analisar o balanço entre o ganho em custo obtido com aumento to tamanho do pool e a perda causada por uma maior complexidade das montagens,

O Capítulo 3 apresenta o manuscrito referente ao sequenciamento e montagem de um pool com 178 BACS de cana-de-açúcar da variedade SP80-3280, e análise comparativa com sorgo mostrando informações interessantes sobre a sintenia entre os genomas destas duas plantas.

CAPITULO 1 - A BAC library of the SP80-3280 sugarcane variety (Saccharum sp.) and its inferred microsynteny with the sorghum genome

Este trabalho foi publicado na revista BMC Research Notes, 2012, 5:185.

A BAC library of the SP80-3280 sugarcane variety (*Saccharum* sp.) and its inferred microsynteny with the sorghum genome

Thais Rezende e Silva Figueira[1], Vagner Okura[1], Felipe Rodrigues da Silva[2], Marcio Jose da Silva[1], Dave Kudrna[3], Jetty SS Ammiraju[3], Jayson Talag[3], Rod Wing[3] and Paulo Arruda[1,4,*]

[1]Centro de Biologia Molecular e Engenharia Genética, Universidade Estadual de Campinas (UNICAMP), 13083-875, Campinas, SP, Brazil

[2]EMBRAPA Informática na Agricultura, São Paulo, Brazil

[3]Arizona Genomics Institute, School of Plant Sciences, BIO5 Institute, University of Arizona, Tucson, AZ 85721, USA

[4]Departamento de Genética e Evolução, Instituto de Biologia, Universidade Estadual de Campinas (UNICAMP), 13083-875, Campinas, SP, Brazil

* Corresponding author

Email addresses:

TRSF: trsfigueira@cbmeg.unicamp.br
VO: vagner@cbmeg.unicamp.br
FRS: felipes@cnptia.embrapa.br
MJS: marciojs@unicamp.br
DK: dkudrna@ag.arizona.edu
JSSA: jettyr@ag.arizona.edu
JT: jtalag@ag.arizona.edu
RW: rwing@ag.arizona.edu
PA: parruda@unicamp.br

Abstract

Background

Sugarcane breeding has significantly progressed in the last 30 years, but achieving additional yield gains has been difficult because of the constraints imposed by the complex ploidy of this crop. Sugarcane cultivars are interspecific hybrids between *Saccharum officinarum* and *Saccharum spontaneum*. *S. officinarum* is an octoploid with 2n=80 chromosomes while *S. spontaneum* has 2n=40 to 128 chromosomes and ploidy varying from 5 to 16. The hybrid genome is composed of 70-80% *S. officinaram* and 5-20% S*. spontaneum* chromosomes and a small proportion of recombinants. Sequencing the genome of this complex crop may help identify useful genes, either per se or through comparative genomics using closely related grasses. The construction and sequencing of a bacterial artificial chromosome (BAC) library of an elite commercial variety of sugarcane could help assembly the sugarcane genome.

Results

A BAC library designated SS_SBa was constructed with DNA isolated from the commercial sugarcane variety SP80-3280. The library contains 36,864 clones with an average insert size of 125 Kb, 88% of which has inserts larger than 90 Kb. Based on the estimated genome size of 760-930 Mb, the library exhibits 5-6 times coverage the monoploid sugarcane genome. Bidirectional BAC end sequencing (BESs) from a random sample of 192 BAC clones sampled genes and repetitive elements of the sugarcane genome. Forty-five per cent of the total BES nucleotides represents repetitive elements, 83% of which belonging to LTR retrotransposons. Alignment of BESs corresponding to 42 BACs to the genome sequence of the 10 sorghum chromosomes revealed regions of microsynteny, with expansions and contractions of sorghum genome regions relative to the sugarcane BAC clones. In general, the sampled sorghum genome regions presented an average 29% expansion in relation to the sugarcane syntenic BACs.

Conclusion

The SS_SBa BAC library represents a new resource for sugarcane genome sequencing. An analysis of insert size, genome coverage and orthologous alignment

with the sorghum genome revealed that the library presents whole genome coverage. The comparison of syntenic regions of the sorghum genome to 42 SS_SBa BES pairs revealed that the sorghum genome is expanded in relation to the sugarcane genome.

Keywords: Sugarcane genomics, BAC library, genome organization, microsynteny, sorghum

Background

Sugarcane is a C4 plant that stores 1/3 of its fixed carbon as sucrose in the parenchyma cells of mature stalks. The other 2/3 is stored in the leaves (1/3) and, the stalks (1/3) in the form of complex carbohydrates [1]. Sugarcane has been grown as a sugar source for a century, but in recent years, extensive industrial plantations have demonstrated this crop's value for the production of sustainable energy [2]. In industrial plantations, when sugarcane is harvested, its leaves are left in the field, contributing to the improvement of soil conservation and fertility. The stalks are transported to sugarcane mills and crushed. After crushing the juice enters a first-pass sucrose crystallisation, and the sugar remaining in the molasses goes to fermenters to produce fuel ethanol [3]. Currently, the dried bagasse resulting from the stalk crushing is used to produce bioelectricity, but it could also be used for the production of cellulosic ethanol [1]. Sugarcane juice has also been used as a carbon source by the synthetic biology industry to produce other fuels and high value molecules [3]. However, the worldwide use of sugarcane for sustainable energy production depends, on the development of superior varieties that are able to grow in less fertile soils, in stress-inducing biotic and abiotic conditions in a range of tropical and sub-tropical environments.

The cultivated sugarcane varieties derive from crosses performed at the beginning of the last century between *S. officinarum*, a species with a high sugar content in the stalk and *S. spontaneum*, a disease-resistant and vigorous wild relative [4,5]. After few backcrosses of the interspecific hybrid to *S. officinarum*, the breeders were able to select varieties less sensitive to biotic and abiotic stress and with a high sugar content in their stalks [5,6]. These early interspecific hybrids constitute the basic germplasm used in breeding programs around the world. However, breeding sugarcane is a complex task because of the high degree of ploidy of the ancestor

species [7,8]. *S. officinarum* is octoploid with a basic chromosome number of x=10 and 2n=80 chromosomes, while *S. spontaneum* has a basic chromosome number of x=8 and 2n= 40 to 128, and a ploidy varying from 5 to 16 [9,10]. The interspecific hybrid genome is a mixture of the genomes of both species with a ploidy varying between 2n=100 and 2n=130 chromosomes [11]. Intact chromosomes from both parents coexist in the interspecific hybrid in proportion of 5-20% from *S. spontaneum* and 70-80% from *S. officinarum*, along with a variable proportion of recombinants between the parental homoeologous chromosomes [12]. This genome architecture imposes constraints for the breeding process and prevents the use of seeds for progeny propagation because of the complex allelic segregation from the polyploidy hybrid [2]. This has limited the achievement of genetic gains in breeding programs, despite the use of crosses between numerous selected parental varieties and evaluation of hundreds of thousands or even millions of progenies in the large-scale field trials.

Because of its complexity, the complete sugarcane genome has not yet been sequenced, mainly due to the difficulty of assigning gene-containing fragments to a specific homologous/homeologous chromosome. However, a reference genome sequence could be assembled from fragments of different homologous and homeologous chromosomes and, even though this reference sequence would be chimeric, it could be useful for comparative genome analysis with close relatives, such as sorghum [13].

The estimated monoploid genome size of sugarcane is approximately 760-930 Mb [7], which is close to the 730 Mb size observed for sorghum [14]. A reference sugarcane genome sequence can be obtained by sequencing a representative bacterial artificial chromosome (BAC) library. Few sequenced BAC clones from the commercial Reunion Island R570 sugarcane variety has already demonstrated the viability of comparative genomics between sugarcane and sorghum [15-17].

This report describes the construction and initial analysis of a BAC library from the Brazilian sugarcane variety SP80-3280, which has been extensively cultivated during the past 18 years [2]. This library will be made available for the scientific community, and would be useful for the establishment of a reference genome sequence for sugarcane. The library was characterised in terms of insert size and

genome coverage based on the alignment of a random sample of BAC end sequences (BESs) into the sorghum genome. Gene annotation of these BESs provided an early glimpse into the sequence composition of the sugarcane genome compared to the sorghum genome.

Results

Construction and characterisation of the SP80-3280 BAC library

The sugarcane variety SP80-3280 was chosen to construct the BAC library because it has been widely cultivated in Brazil. Around 300 thousand Ha has been cultivated with SP80-3280 along the past, recent years in different regions of the country. The superior agronomic performance in such a vast area implies that breeders have selected adaptability traits responsible for yield stability. Thus, sequencing a BAC library from this variety may reveal allelic composition involved in crop performance, and by comparing with genome sequence from other sugarcane BAC libraries may reveal genomic regions responsible for crop adaptation to different environments. The SP80-3280 has also contributed to the cDNA libraries used for EST sequencing carried out by the sugarcane EST project (SUCEST) [18]. SUCEST sequences targeted over 70% of the expressed sugarcane genes [19] and have demonstrated its usefulness for genome annotation of sugarcane BAC sequences [17].

High molecular weight (HMW) genomic DNA prepared from the isolated nuclei of young sugarcane leaves was partially digested with *Hind*III, and the fragments were fractionated by pulsed-field agarose gel electrophoresis [20]. After elution from the gel, the fragments were ligated into the *Hind*III cloning site of the pAGIBAC1 vector, and the ligations were transformed into the DH10B T1 *E. coli* strain to generate the SS_SBa BAC library comprised of 36,864 BAC clones (Table 1). Based on a genome size of 760-930 Mb for the monoploid chromosome set [7], we estimated that this library corresponds to approximately 5-6-monoploid sugarcane genome equivalents. However, as has recently been suggested based on the sequences of 19 BACs from the R570 sugarcane variety [17], the sugarcane genome could be 20% smaller than that of sorghum; therefore, the SS_SBa BAC library could represent 8-fold coverage of the monoploid sugarcane genome. The library was picked into 96 x 384-well plates, and 192 BAC clones, two for each 384-well plate,

were randomly selected for insert size estimation and BAC end sequencing. *Not*I restriction enzyme digestion showed that the library was composed of large insert clones (Fig. 1A) with an average estimated insert size of 125 Kb (ranging from 29 to 293 Kb), 87.5% of which contained inserts larger than 90 Kb (Fig. 1B). Restriction analysis of this 192 BAC clone sample revealed an absence of empty vectors among the 36,864 clones of the SS_SBa BAC library. The 36,864 SS_SBa BAC library clones were printed onto hybridisation screening filters for future experiments.

BES of a clone sample of the SS_SBa BAC library

The quality of the library and its potential genome coverage were examined by bidirectional end sequencing of the randomly selected 192 BAC clones for insert size estimation and its alignment to the genome sequence of the 10 sorghum chromosomes (Supplemental Table 1). After trimming the BES sequence reads for low quality and vector bases, 378 sequences, with an average read length of 944 nucleotides and a minimum length of 312 bases, were recovered (Table 1).

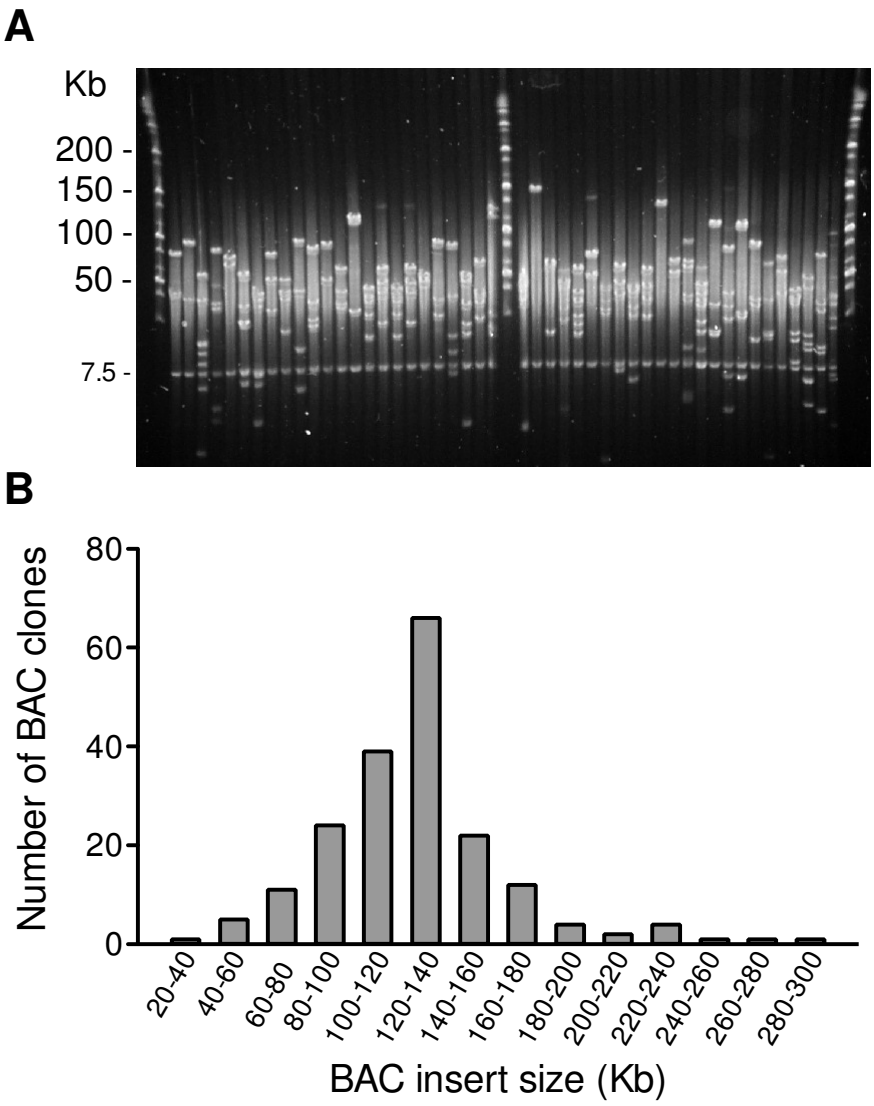


Figure 1- Insert size distribution in a random sample of 192 BAC clones of the SP80-3280 sugarcane BAC library. (A) Example of pulsed-field gel electrophoresis (PGE) of 48 BAC clones DNA digested with *Not*I. Lanes 1, 26 and 51 are Lambda Ladder PFG (New England Biolabs) molecular weight DNA markers. The 7.5-Kb band marks the position of the *Not*I-released cloning vector. (B) Insert size distribution of the 192 BAC clones as estimated by *Not*I digestion and PGE.

The sugarcane BESs were compared to the sugarcane chloroplast genome [21] and the rice mitochondria genome [22]. No significant similarity to mitochondrial genome was found in the library while 1 BAC, out of 192 (0.5%), showed similarities

with chloroplast genome (Table 1). Among the 378 BES sequences, 113 produced no hits with sorghum, either at the nucleotide or protein sequence level. Of these 113 sequences, 67 produced no significant hit against any nucleotide or protein sequence in GenBank, and 36 produced significant hits exclusively with sugarcane (Fig. 2). These 103 BES with no hit with the sorghum genome may represent sugarcane-exclusive sequences. This result is in keeping with those observed by the analysis of 19 sugarcane BAC sequences from the R570 sugarcane variety BAC library [17] and analysis of the sugarcane ESTs [19]. Among the remaining 10 BES with no hit against sorghum, 4 produced significant hits exclusively with maize, 4 with maize and sugarcane, 1 with maize and rice and 1 with maize, rice and sugarcane (Fig. 2). These BESs may represent conserved sequences from the Andropogoneae ancestor that gave rise to grasses but, may have been lost by the sorghum genome after the divergence of Saccharum/sorghum that occurred approximately 8 million years ago (MYA) [7,16,17].

Table 1 - Summary of the SS_SBa Sugarcane BAC library

Germplasm	Sugarcane variety SP80-3280
Cloning vector	pAGIBAC1
Partial digest enzyme	*HindIII*
Number of clones	36,864
Number of 384-well plates	96
Number of analyzed clones	192
Average insert size (kb)	125
Minimum insert size (kb)	29
Maximum insert size (kb)	293
Number of high quality BES	378
Average BES read length (bp)*	944
Chloroplast contamination (%)	0.5
Mitochondrial contamination	None
Number of monoploid genome equivalents**	5-6 X

* Reads were trimmed using parameters established by Telles and da Silva, 2001 [30] to maximize the number of nucleotides with useful sequence information
**Number of genome equivalent was estimated based on a non-redundant chromosome set of the sugarcane genome

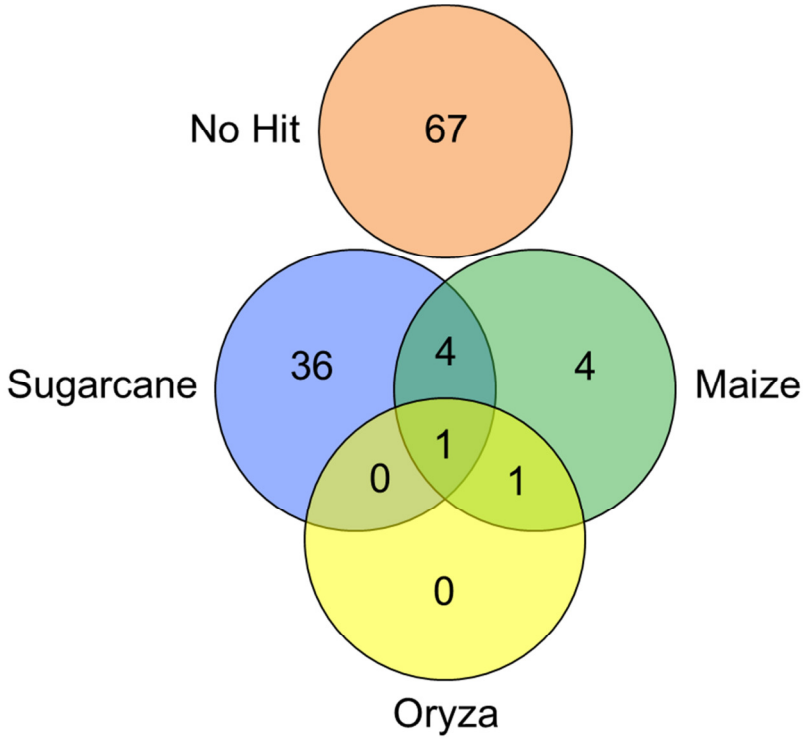


Figure 2- Distribution of BlastN hits among maize, rice and sugarcane of the 118 BES for which no significant hits against the sorghum genome were obtained.

Synteny and micro-collinearity with sorghum

The 378 BESs were aligned with the 10 sorghum chromosomes in search for synteny and micro-collinearity. From the 265 positive alignments, 84 BESs, corresponding to the end sequence pairs of 42 BACs (Table 2, Class1), aligned in a concordant manner with the genome sequence of at least one of the 10 sorghum chromosomes, indicating conformity to sugarcane/sorghum syntenic genome regions. This BES category was assigned as Class 1 and comprises all concordant alignments. Another set of 88 BESs, corresponding to 44 BACs, had both BES end aligned to genome sequences of the same sorghum chromosomes (Table 2, Classes 2 to 5). However, their BES sequence pairs aligned in a discordant manner - in the same orientation or at positions smaller than 20 Kb or larger than 300 Kb. These sequences may correspond to sugarcane genome regions that were inverted, expanded or contracted after the divergence of sugarcane/sorghum. A set of 18 BES, corresponding to 9 BACs, presented end sequence pairs aligned with different sorghum chromosomes (Table 2, Class 6). These sequences may represent

sugarcane regions that were rearranged by translocation after the sugarcane/sorghum divergence. Seventy five BES, corresponding to 75 BACs, aligned individually to sorghum chromosomes, 10 of which having a single match amongst the sorghum chromosomes (Table 2, Class 8) while 65 had more than one possible assigned position (Table 2, Class 9).

Table 2 - Classification of SP 80-3280 BAC end sequences as related to the alignments into the sorghum chromosomes

BES Class	BES per BAC	Aligned BES	Sorghum Chromosome	BES Orientation	Distance Between BES (Kb)	Type	BAC Count
1	2	2	Same	Opposite in (> <)	20 - 300	Concordant	42
2	2	2	Same	Same (< < or > >)	20 - 300	Discordant	1
3	2	2	Same	Opposite in (> <)	> 300	Discordant	16
4	2	2	Same	Same (<< or >>)	> 300	Discordant	12
5	2	2	Same	Opposite out (< >)	> 300	Discordant	15
6	2	2	Different	N/A	N/A	Discordant	9
7	2	1	N/A	N/A	N/A	1	10
8	2	1	N/A	N/A	N/A	> 1	65
9	2	0	N/A	N/A	N/A	N/A	22

Class 1, BAC end pairs that matched the same Sorghum chromosome at positions 20 to 300 Kb apart in opposite orientation. Class 2, BAC end pairs that matched the same Sorghum chromosome at positions 20 to 300 Kb apart in the same orientation. Class 3, BAC end pairs that matched the same Sorghum chromosome within a distance larger than 300Kb in the opposite in orientation. Class 4, BAC end pairs that matched the same Sorghum chromosome within a distance larger than 300Kb in the same orientation. Class 5, BAC end pairs that matched the same Sorghum chromosome within a distance larger than 300Kb in the opposite out orientation. Class 6, BAC end pairs that matched different Sorghum chromosome. Class 7, BAC end pairs for which only one sequence matched a sorghum chromosome at a single position. Class 8, BAC end pairs for which only one sequence matched sorghum chromosomes in more than one position. Class 9, BAC end pairs that didn't match Sorghum chromosome.

Distribution of BES into the sorghum chromosomes

A total of 112 BES, corresponding to one or both ends of 61 BACs, aligned into the 10 sorghum chromosomes (Fig. 3). Eighty four BES corresponding to paired ends of 42 BACs aligned in a concordant manner. Ten BACs had only one BES aligned in a single position into a sorghum chromosome (Table2, Class7). The other 18 BES from 9 BACs aligned in a discordant manner (Table 2 Class 2, 3, 4 and 5). The 61 BACs had their BES randomly aligned along the 10 sorghum chromosomes (Fig. 3). However, chromosomes 5 and 6 presented long regions without aligned BES

sequences. This could be attributed to several different factors, including bias in the constructed BAC library and regions of chromosome 5 and 6 without representation in the sugarcane genome due to sequence loss after the sugarcane/sorghum divergence. Another likely reason for the smaller number of aligned BESs on chromosomes 5 and 6 is that both of these chromosomes are richer in repetitive elements (Table 3). Since we did not align BES ends representing repetitive elements, this has introduced a bias in the BES distribution analysis. Of the 112 BESs analysed (Table 2, Class 1 to 8) only 84 (Table 2, Class 1) aligned in a concordant syntenic manner. The other 28 BESs (Class 2 to 8) aligned in a discordant manner, or each end aligned at different chromosome. This result suggests that the sugarcane genome has undergone extensive rearrangement, including inversions and translocations, relative to the sorghum genome. A sample of the concordant syntenic BACs for which insert size was estimated by restriction enzyme digestion was used to illustrate the expansions and contractions of the sugarcane genome relative to the sorghum genome (Supplemental Table 2). Of the 42 concordant BAC end sequence pairs, 22 aligned to syntenic regions of the sorghum genome that were larger than the estimated size of the sugarcane BAC. Other syntenic regions of the sorghum genome showed contractions compared to the corresponding sugarcane BAC (Supplemental Table 2). However, the sum of the nucleotides of the expanded and contracted syntenic regions showed a positive expansion of the sorghum genome compared to the corresponding sugarcane BACs. The sorghum syntenic regions were 29% expanded relative to the same region represented by the sugarcane BACs. This result is in keeping with the suggestion that the sugarcane genome may be 20% to 30% smaller than the sorghum genome [17]. We also investigated the nature of the genic sorghum region of conserved concordant syntenic regions relative to the sugarcane BACs (Table 3). Sorghum chromosomes (1, 2 and 3) with higher gene density and lower repetitive element content were associated with a higher proportion of sugarcane syntenic BACs. Sorghum chromosomes (5, 6 and 7) with higher repetitive content and lower gene density exhibited the lowest proportion of syntenic BACs found. These findings further suggest that the most syntenic chromosome is chromosome 2, which also shows the most expanded sequence compared to sugarcane. Additionally, the genes present in the syntenic regions of sorghum chromosomes were classified according

to Gene Ontology (GO) functional categories (Supplemental Table 3). Most of the GO terms (55%) associated with the expanded sorghum regions were related to binding metabolic processes. Genes in the biosynthetic process (28%) and nitrogen compound metabolic process (24%) categories, which fall under the biological process category, were the most represented in the expanded regions. GO terms related to transferase activity (48%) were the most widely observed in the regions that were contracted in sorghum in relation to sugarcane. The most (50%) represented biological process category found in these regions was the cellular metabolic process category. Neither the contracted nor the expanded regions appeared to be significantly discrepant, in terms of GO functional categories as compared to the complete genome.

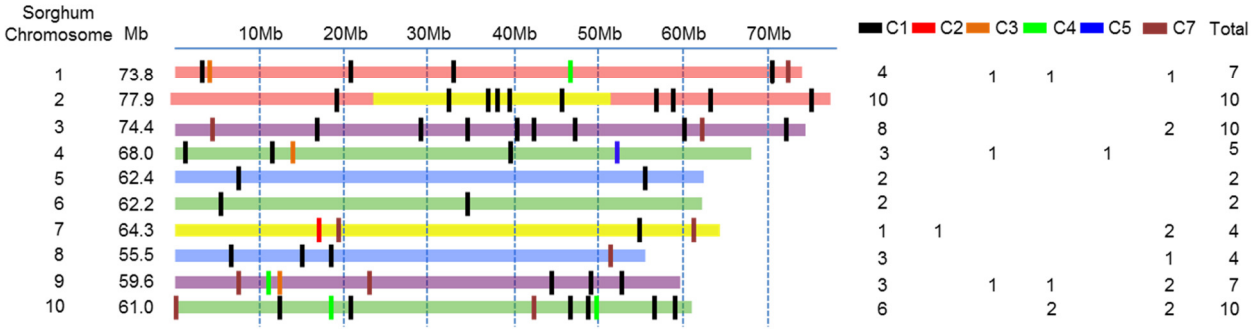


Figure 3- Orthologous alignment of the BES of a random sample of 61 clones of the SP80-3280 sugarcane BAC library on the 10 sorghum chromosomes. Sequences from the both ends of the sugarcane BAC clones were searched against the sorghum genome using BlastN, and significant hits were annotated and positioned on the corresponding sorghum chromosome. For non-repetitive sequences, positioning was based on the BAC insert size, concordance of the opposite end sequence direction and best hit. For repetitive sequences, alignment utilised only the best hit. The coloured solid lines represent the sorghum chromosomes with their predicted segmental duplication [1]. The coloured rectangles represent the sugarcane BES classes. C1, C2, C3, C4, C5, C6 and C7 refer to BESs classes as defined in Table 2. C8 and C9 classes are not represented in the figure.

Table 3 - Difference of expanded and contracted sorghum regions syntenic to sugarcane BACs and gene and repetitive elements content of the expanded/contracted region of sorghum

Chromosome	Number of syntenic regions	Sum of sugarcane BAC nucleotides (bp)	Sum of nucleotides of syntenic sorghum regions (bp)	Nucleotide difference between sorghum and sugarcane syntenic regions (bp)	Gene density of the sorghum chromosomes (%)	Repetitive elements of the sorghum chromosomes (%)	Distribution of sugarcane BACs among sorghum chromosomes (%)
Chromosome 1					21.3	43.4	11.5
Expanded regions	2	228,757	438,099	209,342			
Contracted regions	2	246,930	170,629	-76,301			
Chromosome 2					15.1	60.1	16.4
Expanded regions	7	843,489	1,530,484	686,995			
Contracted regions	3	390,747	307,693	-83,054			
Chromosome 3					16.8	58.2	16.4
Expanded regions	3	300,722	649,695	348,973	2.3	12.7	
Contracted regions	5	623,310	501,139	-122,171	16.9	6.5	
Chromosome 4					14.9	56.2	8.2
Expanded regions	2	228,311	542,706	314,395	5.5	15.6	
Contracted regions	1	129,840	60,597	-69,243	28.0	20.0	
Chromosome 5					8.1	65.9	3.3
Expanded regions	1	105,911	192,480	86,569	19.6	59.2	
Contracted regions	1	188,640	93,686	-94,954	13.6	47.8	
Chromosome 6					12.8	66.3	3.3
Expanded regions	0	na	na	na	na	na	
Contracted regions	2	247,660	68,939	-178,721	0.0	35.9	
Chromosome 7					9.2	66.2	6.6
Expanded regions	1	44,210	296,520	252,310	6.0	57.8	
Contracted regions	0	na	na	na	na	na	
Chromosome 8					9.0	65.6	6.6
Expanded regions	2	266,020	341,742	75,722	0.2	33.2	
Contracted regions	1	125,996	124,114	-1,882	13.1	52.1	
Chromosome 9					11.8	61.6	11.5
Expanded regions	2	221,580	423,004	201,424	27.2	23.5	
Contracted regions	1	114,955	91,453	-23,502	23.5	25.2	
Chromosome 10					12.0	60.9	16.4
Expanded regions	2	184,310	374,313	190,003	8.0	13.9	
Contracted regions	4	502,882	283,520	-219,362	12.2	5.7	
Total		4994270	6490813	1496543			

Number of regions: concordant syntenic regions with either increased or decreased size in sorghum compared to sugarcane.

Sum of sugarcane BAC nucleotides: size of sugarcane BAC in nucleotides as determined by *Not*I restriction digestion analysis.

Sum of nucleotides corresponding to syntenic sorghum regions: sum of nucleotides of the sorghum region corresponding to syntenic sugarcane BACs.

Nucleotide difference between sorghum and sugarcane syntenic regions: positive values indicate regions that are expanded and negative values indicated regions that are contracted in sorghum as related to sugarcane BACs.

Gene density of the sorghum chromosomes: percentage of gene encoding sequences (bp) in each sorghum chromosome.

Repetitive elements of the sorghum chromosomes: percentage of nucleotides corresponding to repetitive elements in each sorghum chromosome. Distribution of sugarcane BACs among sorghum chromosomes: percentage of concordant syntenic sugarcane BACs positioned in each one of the sorghum chromosome.

Repetitive elements content

Among the total number of nucleotides of the 378 BESs analysed, 45.2% produced significant hits with sequences in repetitive elements databases (Table 4). This is in keeping with the proportion of repetitive elements observed in the sample of 19 BACs sequenced from the R570 BAC library [17]. However, repetitive elements are highly lineage-specific, and because the limited sugarcane entries in repbase the data based on BESs may be underestimated. Nevertheless, this preliminary estimation suggests that the repetitive element counterpart of the sugarcane genome, may be smaller than that of the sorghum genome, which contains 61% repetitive sequences, most of which are located in centromeric and pericentromeric regions [14]. Most (98%) of the repetitive nucleotides found in the BES reads corresponded to transposable elements; 85.2% were LTR retrotransposons, of which 48.1% were assigned to the Copia family and 51.6% to the Gypsy family. Non-LTR retrotransposons of the L1, RTE, SINE and SINE/tRNA families corresponded to 3.5% of the total repetitive element nucleotides (Table 4). DNA transposable elements belonging to the EnSpm, Harbinger, Helitron, MuDr and hAT families represented 10.9% of the total BES repetitive nucleotides. Few sequences were found to correspond to integrated viruses (0.8%) or simple repeats (1.2%).

Table 4 - Summary of repetitive sequences among the sugarcane BESs

Repeat Element	Number of elements	Length (bp)	% of Total Bases
Transposable Element	293	160624	44.29
RNA transposon	234	142693	39.35
LTR Retrotransposon	221	136899	37.76
Copia	96	65873	18.17
Gypsy	123	70697	19.50
Non-LTR Retrotransposon	13	5794	1.59
L1	7	2626	0.72
RTE	4	2943	0.81
SINE	2	225	0.06
SINE2/tRNA	2	225	0.06
DNA transposon	59	17931	4.94
EnSpm	13	5358	1.48
Harbinger	11	2728	0.75
Helitron	2	1197	0.33
MuDr	6	2515	0.69
hAT	9	3111	0.86
Integrated Virus	2	1231	0.34
Caulimoviridae	2	1231	0.34
Simple Repeat	3	1923	0.53
Satellite	3	1923	0.53
Total	298	163778	45.16

Discussion

Two BAC libraries from the Reunion Island sugarcane cultivar R570, one constructed with DNA isolated from the commercial variety [23] and, the other constructed with DNA isolated from selfed progenies of R570 [24] are current available. These libraries have contributed with BAC sequencing for various purposes. Here, we described the construction and initial analyses of a new sugarcane BAC library prepared with genomic DNA from a Brazilian elite commercial sugarcane variety. This BAC library exhibits genome coverage of 5-6 times the monoploid chromosome set of sugarcane. The genome coverage was estimated based on a size of 760-930 Mb for the monoploid sugarcane genome [7]. However, in a previous study, syntenic alignment of 19 sugarcane BAC sequences from the R570 BAC library into the 20 sorghum chromosome arms revealed predominant local DNA sequence expansion of the sorghum genome in the regions syntenic with the sugarcane BAC sequences [17]. These results suggested that the monoploid sugarcane genome could be 20% smaller than the 730 Mb sorghum genome. The alignment of the 42 BES pairs into concordant syntenic regions of the sorghum genome revealed 29% expansion of

sorghum in relation to the sugarcane genome. This result is in keeping with the results observed for the R570 BAC library and suggests that the size of the monoploid sugarcane genome could be on the order of 580 Mb. If this is correct, the coverage of the SS_SBa BAC library could be on the order of 8 times the sugarcane monoploid genome.

The use of the sorghum genome sequence as a template to assemble the sugarcane genome has been proposed based on the close similarity between the two species [25,26]. The sequence of BAC clones from the R570 BAC library and comparison of its gene and repetitive element content to that of sorghum improved confidentiality with respect to these assumptions [16,17]. Sequence analysis of 19 BAC from the R570 BAC library revealed that almost 85% of its gene-encoding sequences are syntenic with sorghum orthologs [17]. We analysed the sorghum chromosomes for gene density as related to the distribution of the SP80-3280 BES. Sorghum chromosomes 1, 2 and 3 showed the highest gene density and had increased number of aligned sugarcane BESs (Table 3). Chromosomes 5 and 6 has reduced gene density were richer in repetitive elements and showed fewer aligned sugarcane BESs (Table 3).

The library described in this report is from an elite commercial sugarcane variety that has been cultivated on hundreds of thousands of hectares in a range of different environments, including regions of less favourable soils in terms of water and nutrient availability. This library would be useful in providing additional information regarding the allelic composition selected by breeders. The overlapping BACs in this library may represent different homeologous chromosomes from both *S. officinarum* and *S. spontaneum* parents. Since *S. officinarum* contributes mainly with yield and sugar alleles and, *S. spontaneum* contributes mainly with stress tolerance genes, the sequences of overlapping BACs representing both species could be identified by high stringency filter hybridisation with DNA from the two parents [16]. Furthermore, their gene and allele content could be identified, and the contribution of each of the parental genes to disease resistance and sugar content could be assigned. Additionally, expression patterns obtained using next generation platforms could provide additional useful information regarding this valuable genetic resource.

Conclusions

Sugarcane is a main crop for both sugar and bioenergy generation. To address the projections for sugarcane production, breeding and biotechnology approaches must be developed in the next few years, to assist the selection of high sugar yield varieties adapted to tropical and sub-tropical regions. Sequencing the genome of this complex crop may help to identify agronomically useful genes, either per se or through comparative genomics, and could also assist in the development of biotechnology tools for sugarcane improvement. This report describes the construction and preliminary analyses of a sugarcane BAC library from DNA isolated from a Brazilian elite sugarcane variety. The library comprises large insert clones and possesses 5-6 times coverage of the monoploid sugarcane genome. Sequencing and alignment of BAC end sequences from a sample of this library into orthologous regions of the sorghum genome revealed that the library presents sound genome coverage. In addition, comparison of the syntenic regions of the sorghum genome with respect to BAC end sequence pairs confirmed that the sugarcane genome might be between 20% and 30% smaller than the sorghum genome. This library represents a new resource for the community interested in sugarcane breeding and biotechnology coupled with sustainable bioenergy generation.

Methods

Germplasm and plant tissue processing

Twenty 10-week-old, field-grown sugarcane plants of the SP80-3280 variety were generously provided by the Cosan company (www.cosan.com.br). The plants were harvested at Usina Santa Helena in Fazenda Santo Antonio (GPS coordinates -22.735657, -47.305069), Piracicaba, State of São Paulo, Brazil. The plants were subjected to a 30-hour dark treatment, after which the healthy young leaves were collected, quickly washed to remove debris and immediately frozen by submersion in liquid nitrogen. The frozen leaves were stored at -80◦C until use.

Preparation of high molecular weight (HMW) sugarcane DNA in agarose plugs

The sugarcane SP-803280 BAC library was constructed in the Arizona Genomics Institute (AGI) using standard protocols [27,28]. Fifty grams of frozen tissue were ground under liquid nitrogen with a mortar and pestle. The ground tissue was transferred to a 1-L Erlenmeyer flask containing 500 mL of pre-chilled extraction buffer (10 mM Tris-HCL, pH 8.0, 10 mM EDTA, pH 8.0, 100 mM KCl, 0.5 M sucrose, 4 mM spermidine, 1 mM spermine, 2.0% w/v PVP-40, 0.13% w/v sodium diethyldithiocarbamate trihydrate and 800 µl β-mercaptoethanol). The suspension was gently shaken for 15 min, and the homogenate was filtered into an Erlenmeyer flask containing 500 mL of pre-chilled extraction buffer with 1.7% Triton X-100. The suspension was kept on ice for 15 min and then centrifuged for 15 min at 3,250 rpm at 4˚C. The resulting pellet was resuspended in pre-chilled extraction buffer, incubated for 5 min in a water bath at 45˚C and gently mixed with 1/3 v/v of 1.0% low melting temperature agarose that was previously prepared in extraction buffer and held at 45˚C. The mixture was transferred to plug moulds and allowed to solidify. Forty-six plugs were transferred into a 50-mL Falcon tube containing 40 mL of proteinase K solution (0.5 M EDTA pH 9.2, 1.0%N-lauroylsarcosine, 40 mg proteinase K and 2% PVP), and the tube was incubated in a hybridisation oven at 50˚C with gentle rotation for 24 h. The plugs were then washed with fresh proteinase K solution for an additional 24 h. Subsequently, the plugs were washed five times for 1 h at room temperature using 40 mL T10E10 containing phenylmethylsulfonyl fluoride (PMSF; 10 mM Tris-HCL, 10 mM EDTA, 1 mM PMSF, pH 8.0) and five times for 1 h with T10E1 plus PMSF (10 mM Tris-HCL, 1 mM EDTA, 1 mM PMSF, pH 8.00). The plugs were stored in TE at 4˚C.

Restriction digestion of HMW DNA and isolation of size-selected fragments

Eight DNA plugs were partially digested for 20 minutes with 0.6 U of the *Hind*III restriction enzyme for each half plug. The digested samples were loaded into a 1.0% agarose gel and subjected to pulsed-field gel electrophoresis (PFGE). DNA was visualised using a UV transilluminator, and fragments containing DNA ranging from 90 to 450 Kb were cut from the gel slabs. The fragments were subsequently purified through second and third PFGE runs to remove small trapped DNA fragments [27].

The gel fractions containing sized fragments were recovered from the gel slabs and stored at 4˚C.

Ligation of sized DNA fragments

High-molecular-weight genomic DNA fragments (120-200 ng) were ligated into a HindIII- linearized and dephosphorylated pAGIBAC1 plasmid vector [27]. The ligation reactions were incubated in a water bath at 16◦C for 19 h, transferred to 0.1 M glucose/1.0% agarose and allowed to desalt for 1.5 h on ice. The ligations were transferred into new microcentrifuge tubes and stored at 4◦C. The ligation samples were tested to determine the transformation efficiency and cloned insert quality. For the final transformations, 2.0 µl of ligation mixture was used to electroporate 20 µl of DH10B T1 phage-resistant E. coli cells (Invitrogen). The transformed cells were transferred into 3 mL of SOC media and incubated at 37◦C for 1 h in a shaker at 250 rpm, followed by the addition of an equal volume of sterile glycerol and gentle shaking for 3 min, after which the mixtures were immediately frozen by submersion into liquid nitrogen and stored at -80◦C. Subsequently, the cells were thawed and plated on 22.5 x 22.5 cm plates containing solid LB medium with 12.5 µg/mL chloramphenicol, 80 µg/mL X-gal and 100 µg/mL IPTG. The plates were incubated at 37◦C overnight. White recombinant colonies were transferred into liquid LB medium containing 12.5 mg/mL chloramphenicol and incubated overnight at 37◦C. The transformed E. coli from ligations that contained large inserts were arrayed into 96 x 384-well plates to constitute the SS_SBa BAC library.

Quality control and BES sequencing and analysis

Two 96-wells plates were set up using two clones from each 384-well plate of the SS_SBa BAC library. BAC DNA was isolated from these two 96-well plates, digested with NotI and separated by PFGE for fragment sizing. DNA from the same 192 BAC clones was used for BAC end sequencing with an ABI 3730 sequencer at the AGI facility. The BESs were trimmed for vector and low quality sequences using the SUCEST project trimming procedure [29]. The trimmed sequences were compared to the NCBI GenBank non-redundant protein database using BlastX (E-value cutoff of 1e-5), to NCBI GenBank nucleotide database, to sorghum, maize and rice genome sequences, sugarcane ESTs and BAC sequences and to the sugarcane chloroplast

genome [21] and rice mitochondria genome [22] using BlastN. For all BlastN searches, an E-value cutoff of 1e-20 was used. Additionally, for chloroplast and mitochondria BlastN searches a cutoff of 80% coverage was used. Repeats in the sugarcane BES were masked [30] and identified through searches for similarity to grass sequences in the RepBase [31] with Censor [32]. The BES sequences have been submitted GenBank/NCBI under ID: 1495713.

Comparative analysis and alignment of BESs into the sorghum genome

Regions of microsynteny between sorghum and sugarcane were mapped by the alignment of BESs onto sorghum genome sequences using BlastN alignments with an E-value cutoff of 1e-20. A BES was considered microsyntenic if both ends mapped within 20 Kb and 300 Kb in the opposite orientation. When the two ends were opposite oriented one to another, the region was considered collinear [33,34]. Otherwise, the region was considered to be rearranged between the two species. The best score sum of two ends was used to select among multiple mapping possibilities. Gene density and Gene Ontology analyses of the sorghum chromosomes and syntenic regions were based on Phytozome (V7.0) and the JGI sorghum genome annotation. Repetitive elements in the sorghum chromosomes and syntenic regions were identified with Censor [32] using RepBase [31].

Acknowledgements

TRSF was supported by CAPES. PA is a recipient of a CNPq productivity fellowship.

Authors' contributions

TRSF participated in sample collection, DNA preparation, data analysis and help drafting of the manuscript. TRSF and JT constructed the sugarcane BAC library under the supervision of DK, JSSA and RW. MJS was involved in sampling and coordination. VO and FRS performed the bioinformatics analysis, and PA directed the strategy of the work toward sequencing the sugarcane genome using a BAC library and drafted the manuscript. All of the authors have read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

References

1. Goldemberg J, Coelho ST, Guardabassi P: **The sustainability of ethanol production from sugarcane.** *Energy Policy* 2008, **36:** 2086-2097.
2. Matsuoka S, Ferro J, Arruda P: **The Brazilian experience of sugarcane ethanol industry**. *In Vitro Cell Dev Biol Plant* 2009, **45:** 372–381.
3. Arruda P: **Perspective of the Sugarcane Industry in Brazil.** *Tropical Plant Biol* 2011, **4:**3-8.
4. Arcenaux G: **Cultivated sugarcanes of the world and their botanical derivation.** *Proc Int Soc Sugarcane Technol* 1967, **12:**844-85.
5. Berding N, Roach BT: **Germplasm collection, maintenance, and use. In: Heinz DJ (ed.) Sugarcane improvement through breeding.** Elsevier, New York; 1987:143-210.
6. Roach BT: **Nobilisation of sugarcane.** *Proc Int Soc Sugar Cane Technol* 1972, **14:** 206-216.
7. D'Hont A, Glaszmann JC: **Sugarcane genome analysis with molecular markers, a first decade of research.** *Proceedings of the International Society of Sugarcane Technology* 2001, **24:** 556-559.
8. Lu YH, D'Hont A, Paulet F, Grivet L, Arnaud M, Glaszmann JC: **Molecular diversity and genome structure in modern sugarcane varieties.** *Euphytica* 1994, **78:** 217-226.
9. D'Hont A, Ison D, Alix K, Roux C, Glaszmann JC: **Determination of basic chromosome numbers in the genus Saccharum by physical mapping of ribosomal RNA genes.** *Genome* 1998, **41:** 221-225.
10. Ha S, Moore PH, Heinz D, Kato S, Ohmido N, Fukui K: **Quantitative chromosome map of the polyploid Saccharum spontaneum by multicolor fluorescence in situ hybridization and imaging methods.** *Plant Molecular Biology* 1999, **39:** 1165-1173.
11. Grivet L, Arruda P: **Sugarcane genomics: depicting the complex genome of an important tropical crop.** ***Current Opinion in Plant Biology*** 2001, **5:**122-127.
12. D'Hont A, Grivet L, Feldmann P, Rao S, Berding N, Glaszmann JC: **Characterisation of the double genome structure of modern sugarcane**

cultivars (Saccharum spp.) by molecular cytogenetics. *Molecular and General Genetics* 1996, **250:** 405-413.

13. Abrouk M, Murat F, Pont C, Messing J, Jackson S, Faraut T, Tannier E, Plomion C, Cooke R, Feuillet C, Salse J: **Palaeogenomics of plants: synteny-based modelling of extinct ancestors.** *Trends Plant Sci* 2010, **15:**479-87.
14. Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, Schmutz J, Spannagl M, Tang H, Wang X, Wicker T, Bharti AK, Chapman J, Feltus FA, Gowik U, Grigoriev IV, Lyons E, Maher CA, Martis M, Narechania A, Otillar RP, Penning BW, Salamov AA, Wang Y, Zhang L, Carpita NC, *et al*: **The sorghum bicolor genome and the diversification of grasses.** *Nature* 2009, **457:**551-556.
15. Garsmeur O, Charron C, Bocs S, Jouffe V, Samain S, Couloux A, Droc G, Zini C, Glaszmann JC, Van Sluys MA, D'Hont A: **High homologous gene conservation despite extreme autopolyploid redundancy in sugarcane.** *New Phytologist* 2011, **189:** 629–642
16. Jannoo N, Grivet L, Chantret N, Garsmeur O, Glaszmann JC, Arruda P, D'Hont A: **Orthologous comparison in a gene-rich region among grasses reveals stability in the sugarcane polyploid genome.** *The Plant Journal* 2007, **50:**574–585.
17. Wang J, Roe B, Macmil S, Yu Q, Murray JE, TangH, Chen C, Najar F, Wiley G, Bowers J, Van Sluys MA, Rokhsar DS, Hudson ME, Moose SP, Paterson AH, Ming R: **Microcollinearity between autopolyploid sugarcane and diploid sorghum genomes.** *BMC Genomics* 2010, **11:**261.
18. Vettore AL, da Silva FR, Kemper EL, Arruda P: **The libraries that made SUCEST**. *Genetics Molecular Biology* 2001, **24:**1-4.
19. Vettore AL, da Silva FR, Kemper EL, Souza GM, da Silva AM, Ferro MI, Henrique-Silva F, Giglioti EA, Lemos MV, Coutinho LL, Nobrega MP, Carrer H, Franca SC, Bacci Junior M, Goldman MH, Gomes SL, Nunes LR, Camargo LE, Siqueira WJ, Van Sluys MA, Thiemann OH, Kuramae EE, Santelli RV, Marino CL, Targon ML, Ferro JA, Silveira HC, Marini DC, Lemos EG, Monteiro-Vitorello CB, *et al*: **Analysis and functional annotation of an expressed sequence tag collection for tropical crop sugarcane**. *Genome Res* 2003, **13:**2725-2735.

20. Luo M, and Wing RA: **An improved method for plant BAC library construction.** *Plant Functional Genomics: Methods and Protocols*, E. Grotewold, Ed., Humana Press, Totowa, NJ, USA, 2003, 3-19.
21. Junior TC, Carraro DM, Benatti MR, Barbosa AC, Kitajima JP, Carrer H: **Structural features and transcript-editing analysis of sugarcane (Saccharum officinarum L.) chloroplast genome.** *Current Genetics* 2004, **46:**366-373.
22. Notsu Y, Masood S, Nishikawa T, Kubo N, Akiduki G, Nakazono M, Hirai A, Kadowaki K: **The complete sequence of the rice (Oryza sativa L.) mitochondrial genome: frequent DNA sequence acquisition and loss during the evolution of flowering plants.** *Molecular Genetics and Genomics* 2002, **268:**434-445.
23. Tomkins JP, Yu Y, Miller-Smith H, Frisch DA, Woo SS, Wing R: **A bacterial artificial chromosome library for sugarcane.** *Theoretical and Applied Genetics* 1999, **99:** 419-424.
24. Le Cunff L, Garsmeur O, Raboin LM, Pauquet J, Telismart H, Selvi A, Grivet L, Philippe R, Begum D, Deu M, Costet L, Wing R, Glaszmann JC, D'Hont A: **Diploid / polyploid syntenic shuttle mapping and haplotype-specific chromosome walking toward a rust resistance gene (Bru1) in highly polyploid sugarcane (2n - 12x - 115).** *Genetics* 2008 **180:** 649-660.
25. Bowers JE, Arias MA, Asher R, Avise JA, Ball RT, Brewer GA, Buss RW, Chen AH, Edwards TM, Estill JC, Exum HE, Goff VH, Herrick KL, James Steele CL, Karunakaran S, Lafayette GK, Lemke C, Marler BS, Masters SL, McMillan JM, Nelson LK, Newsome GA, Nwakanma CC, Odeh RN, Phelps CA, Rarick EA, Rogers CJ, Ryan SP, Slaughter KA, Soderlund CA *et al*: **Comparative physical mapping links conservation of microsynteny to chromosome structure and recombination in grasses.** *Proc Natl Acad Sci USA* 2005, **102:**13206-13211
26. Ming R, Liua SC, Lina YR, da Silva J Wilson W, Braga D, van Deynze A, Wenslaff TF, Wud KK, Mooree PH, Burnquist W, Sorrells ME, Irvine JE, Paterson AH: **Detailed alignment of saccharum and sorghum chromosomes: comparative organization of closely related diploid and polyploid genomes.** *Genetics* 1998, **150:**1663-168.
27. Lin J, Kudrna D, Wing RA: **Construction, characterization, and preliminary BAC-end sequence analysis of a bacterial artificial chromosome library of**

the tea plant (Camellia sinensis). *Journal of Biomedicine and Biotechnology* 2011, doi:10.1155/2011/476723.

28. Peterson DG, Tomkins JP, Frisch DA, Wing RA, PatersonAH: **Construction of plant bacterial artificial chromosome (BAC) libraries: an illustrated guide.** *Journal of Agricultural Genomics* 2000, **5:**1-100.
29. Telles GP, da Silva FR: **Trimming and clustering sugarcane ESTs**. *Genet Mol Biol* 2001, **24:** 17-23.
30. Tarailo-Graovac M, Chen N: **Using RepeatMasker to identify repetitive elements in genomic sequences**. *Curr Protoc Bioinformatics* 2009, **4:**4-10.
31. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J: **Repbase Update: a database of eukaryotic repetitive elements.** *Cytogenet Genome Res* 2005, **110:**462-7.
32. Kohany O, Gentles AJ, Hankus L, Jurka J.: **Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor.** *BMC Bioinformatics* 2006 **7:**474.
33. Kim H, Hurwitz B, Yu Y, Collura K, Gill N, SanMiguel P, Mullikin JC, Maher C, Nelson W, Wissotski M, Braidotti M, Kudrna D, Goicoechea JL, Stein L, Ware D, Jackson SA, Soderlund C, Wing RA: **Construction, alignment and analysis of twelve framework physical maps that represent the ten genome types of the genus Oryza.** *Genome Biol.* 2008, **9:**R45.
34. Ammiraju JSS, Luo M, Goicoechea JL, Wang W, Kudrna D, Mueller C, Talag J, Kim H, Sisneros NB, Blackmon B, Fang E, Tomkins JB, Brar D, MacKill D, McCouch S, Kurata N, Lambert G, Galbraith DW, Arumuganathan K, Rao K, Walling JG, Gill N, Yu Y, SanMiguel P, Soderlund C, Jackson S, Wing RA: **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 Research* 2006, **16:**140-147.

Additional files

Additional file 1: Table S1. Sequence ID, annotation and orthologous positioning of 384 BES from 192 clones of the Sugarcane SP 80-3280 BAC Library into sorghum chromosomes. Orthologous sequence present in the maize and rice genomes are also displayed.

Additional file 2: Table S2. Position coordinates of 42 BES pairs into the 10 sorghum chromosomes.

Additional file 3: Table S3. Gene ontology (GO) categories for genes in the expanded and contracted regions of the color genome compared with the complete genome.

Supplemental Table 2 - Position coordinates of 42 BES pairs into the 10 sorghum chromosomes.

Sugarcane BAC	Sorghum chromosome	Forward chromosome position (.b)	E-value	Reverse chromosome position (.g)	E-value	Estimated sorghum syntenc size (bp)	Sugarcane BAC size (bp)
SCSBa0003A24	2	74288971	0.0	74.486.173	4,00E-74	197.203	117.770
SCSBa0004C24	3	71.753.270	0.0	71.855.132	0.0	101.863	124.980
SCSBa0010C24	3	40.249.741	3,00E-65	40.178.428	1,00E-54	71.314	140.230
SCSBa0015A24	10	58.773.027	2,00E-116	58.853.259	6,00E-92	80.233	121.670
SCSBa0021C24	1	70.662.536	7,00E-35	70.379.692	5,00E-32	282.845	138.670
SCSBa0025A24	9	44281097	7,00E-121	44.404.512	7,00E-136	123.416	101.880
SCSBa0027A24	3	60.120.673	0.0	60.009.820	0.0	110.854	130.830
SCSBa0030C24	6	34491427	0.0	34.533.867	1,00E-151	42.441	126.480
SCSBa0031A24	10	20658817	4,00E-32	20.839.680	0.0	180.864	131.280
SCSBa0031C24	2	58171855	8,00E-24	58.335.967	3,00E-174	164.113	130.970
SCSBa0038A24	2	45184544	0.0	45.370.791	2,00E-120	186.248	78.600
SCSBa0039C24	3	16.889.392	0.0	16.769.011	2,00E-135	120.382	120.630
SCSBa0040C24	2	31802041	5,00E-61	32.047.384	1,00E-141	245.344	173.489
SCSBa0041C24	1	20507363	0.0	20.600.195	2,00E-45	92.833	137.210
SCSBa0042A24	2	37162464	2,00E-80	37.458.027	1,00E-82	295.564	127.200
SCSBa0046C24	5	7.494.523	0.0	7.400.838	1,00E-82	93.686	188.640
SCSBa0053C24	10	12.247.867	0.0	12.268.743	2,00E-112	20.877	124.548
SCSBa0054A24	2	36592889	6,00E-30	36.736.744	0.0	143.856	102.562
SCSBa0057C24	2	39.260.076	5,00E-167	39.171.718	3,00E-24	88.359	130.940
SCSBa0058A24	4	11.037.453	2,00E-159	11.298.892	4,00E-67	261.440	96.541
SCSBa0058C24	5	55.509.375	0.0	55.316.896	4,00E-47	192.480	105.911

SCSBa0062A24	8	14964531	0.0	15.145.853	2,00E-70	181.323	140.230
SCSBa0065C24	3	34240595	9,00E-49	34.143.870	1,00E-131	96.726	106.640
SCSBa0066A24	8	6.420.909	0.0	6.545.022	0.0	124.114	125.996
SCSBa0069C24	10	56342892	5,00E-23	56.536.340	1,00E-97	193.449	53.030
SCSBa0070A24	2	62.684.716	0.0	62.773.764	2,00E-86	89.049	92.127
SCSBa0072C24	7	54.224.489	7,00E-164	54.521.008	1,00E-41	296.520	44.210
SCSBa0074C24	1	31409470	1,00E-82	31.564.723	2,00E-136	155.254	90.087
SCSBa0076C24	10	48669699	0.0	48.597.659	1,00E-102	72.041	139.904
SCSBa0078C24	2	56096647	3,00E-95	56.394.802	2,00E-145	298.156	112.898
SCSBa0079A24	3	47406230	0.0	47.280.582	1,00E-52	125.649	111.460
SCSBa0080A24	9	48.831.048	0.0	49.130.635	0.0	299.588	119.700
SCSBa0080C24	3	29.113.999	2,00E-129	28.886.358	4,00E-52	227.642	125.512
SCSBa0081A24	3	42.095.365	2,00E-68	42.391.768	1,00E-67	296.404	63.750
SCSBa0083A24	9	52.719.464	0.0	52.810.916	3,00E-55	91.453	114.955
SCSBa0083C24	4	1.159.346	5,00E-96	1.219.942	2,00E-34	60.597	129.840
SCSBa0086A24	1	3.364.892	0.0	3.287.097	5,00E-167	77.796	109.720
SCSBa0092A24	4	39253285	3,00E-158	39.534.550	4,00E-82	281.266	131.770
SCSBa0094A24	10	46219935	3,00E-64	46.109.567	3,00E-105	110.369	116.760
SCSBa0094C24	6	5114815	8,00E-94	5.141.312	0.0	26.498	121.180
SCSBa0095C24	8	18265203	3,00E-35	18.104.785	0.0	160.419	125.790
SCSBa0096C24	2	19034783	6,00E-30	18.904.499	2,00E-130	130.285	167.680
Total:						6.490.813	4.994.270
				Sorghum genome expansion compared to Sugarcane:		29,97%	

Supplemental Table 2 - Position coordinates of 112 BES (61 BACs) into the 10 sorghum chromosomes.

Sugarcane BAC	Sorghum chromosome	Forward chromosome position (.b)	E-value	Reverse chromosome position (.g)	E-value	Estimated sorghum syntenic size (bp)	Class
SCSBa0002A24	8	51.247.493	0.0				Class07
SCSBa0003A24	2	74.288.971	0.0	74.486.173	3,00E-74	197.203	Class01
SCSBa0004C24	3	71.753.270	0.0	71.855.132	0.0	101.863	Class01
SCSBa0010C24	3	40.249.741	3,00E-65	40.178.428	9,00E-55	71.314	Class01
SCSBa0012C24	4	10.061.516	6,00E-45	13.899.451	3,00E-153	3.837.936	Class03
SCSBa0013A24	1			72.123.994	0.0		Class07
SCSBa0015A24	10	58.773.027	2,00E-116	58.853.259	6,00E-92	80.233	Class01
SCSBa0021C24	1	70.662.536	6,00E-35	70.379.692	4,00E-32	282.845	Class01
SCSBa0025A24	9	44.281.097	6,00E-121	44.404.512	7,00E-136	123.416	Class01
SCSBa0027A24	3	60.120.673	0.0	60.009.820	0.0	110.854	Class01
SCSBa0028C24	7	16.656.178	5,00E-52	16.770.660	4,00E-98	114.483	Class02
SCSBa0030C24	6	34.491.427	0.0	34.533.867	1,00E-151	42.441	Class01
SCSBa0031A24	10	20.658.817	4,00E-32	20.839.680	0.0	180.864	Class01
SCSBa0031C24	2	58.171.855	7,00E-24	58.335.967	3,00E-174	164.113	Class01
SCSBa0036C24	7			60.868.951	0.0		Class07
SCSBa0038A24	2	45.184.544	0.0	45.370.791	2,00E-120	186.248	Class01
SCSBa0039A24	9			22.922.845	6,00E-25		Class07
SCSBa0039C24	3	16.889.392	0.0	16.769.011	2,00E-135	120.382	Class01
SCSBa0040C24	2	31.802.041	5,00E-61	32.047.384	1,00E-141	245.344	Class01
SCSBa0041C24	1	20.507.363	0.0	20.600.195	2,00E-45	92.833	Class01
SCSBa0042A24	2	37.162.464	2,00E-80	37.458.027	1,00E-82	295.564	Class01
SCSBa0046C24	5	7.494.523	0.0	7.400.838	1,00E-82	93.686	Class01

SCSBa0049C24	3	4.336.800	9,00E-35				Class07
SCSBa0050C24	9	7.595.636	0.0				Class07
SCSBa0052A24	4	55.545.378	0.0	52.168.921	2,00E-51	3.376.458	Class05
SCSBa0053C24	10	12.247.867	0.0	12.268.743	2,00E-112	20.877	Class01
SCSBa0054A24	2	36.592.889	5,00E-30	36.736.744	0.0	143.856	Class01
SCSBa0057A24	3	61.848.303	0.0				Class07
SCSBa0057C24	2	39.260.076	5,00E-167	39.171.718	2,00E-24	88.359	Class01
SCSBa0058A24	4	11.037.453	2,00E-159	11.298.892	4,00E-67	261.440	Class01
SCSBa0058C24	5	55.509.375	0.0	55.316.896	4,00E-47	192.480	Class01
SCSBa0060A24	10	42.077.226	5,00E-170				Class07
SCSBa0062A24	8	14.964.531	0.0	15.145.853	2,00E-70	181.323	Class01
SCSBa0063A24	7			19.076.107	2,00E-49		Class07
SCSBa0063C24	10	49.786.791	3,00E-104	47.714.346	7,00E-76	2.072.446	Class04
SCSBa0065C24	3	34.240.595	8,00E-49	34.143.870	1,00E-131	96.726	Class01
SCSBa0066A24	8	6.420.909	0.0	6.545.022	0.0	124.114	Class01
SCSBa0069C24	10	56.342.892	5,00E-23	56.536.340	1,00E-97	193.449	Class01
SCSBa0070A24	2	62.684.716	0.0	62.773.764	2,00E-86	89.049	Class01
SCSBa0072C24	7	54.224.489	6,00E-164	54.521.008	1,00E-41	296.520	Class01
SCSBa0074C24	1	31.409.470	9,00E-83	31.564.723	1,00E-136	155.254	Class01
SCSBa0076C24	10	48.669.699	0.0	48.597.659	1,00E-102	72.041	Class01
SCSBa0078C24	2	56.096.647	3,00E-95	56.394.802	2,00E-145	298.156	Class01
SCSBa0079A24	3	47.406.230	0.0	47.280.582	1,00E-52	125.649	Class01
SCSBa0080A24	9	48.831.048	0.0	49.130.635	0.0	299.588	Class01
SCSBa0080C24	3	29.113.999	2,00E-129	28.886.358	4,00E-52	227.642	Class01
SCSBa0081A24	3	42.095.365	2,00E-68	42.391.768	1,00E-67	296.404	Class01
SCSBa0083A24	9	52.719.464	0.0	52.810.916	2,00E-55	91.453	Class01
SCSBa0083C24	4	1.159.346	4,00E-96	1.219.942	2,00E-34	60.597	Class01
SCSBa0084A24	10	61.667	1,00E-116				Class07
SCSBa0085C24	10	18.542.195	8,00E-109	16.294.920	0.0	2.247.276	Class04

SCSBa0086A24	1	3.364.892	0.0	3.287.097	5,00E-167	77.796	Class01
SCSBa0091A24	1	7.472.870	2,00E-160	4.052.722	0.0	3.420.149	Class03
SCSBa0091C24	9	12.270.280	0.0	12.997.862	0.0	727.583	Class03
SCSBa0092A24	4	39.253.285	2,00E-158	39.534.550	3,00E-82	281.266	Class01
SCSBa0093C24	1	46.651.602	0.0	50.906.083	1,00E-151	4.254.482	Class04
SCSBa0094A24	10	46.219.935	3,00E-64	46.109.567	3,00E-105	110.369	Class01
SCSBa0094C24	6	5.114.815	7,00E-94	5.141.312	0.0	26.498	Class01
SCSBa0095A24	9	10.976.049	3,00E-72	6.772.313	4,00E-126	4.203.737	Class04
SCSBa0095C24	8	18.265.203	2,00E-35	18.104.785	0.0	160.419	Class01
SCSBa0096C24	2	19.034.783	6,00E-30	18.904.499	1,00E-130	130.285	Class01

Supplementary Table 3 - Gene ontology (GO) categories for genes in the expanded and contracted regions of the Sorghum bicolor genome compared with the complete genome

GO Category		Expanded Regions		Contracted Regions		Sorghum Genome	
		Number	%*	Number	%**	Number	%***
Biological Process	biological adhesion					20	0,08
	biological regulation	26	18,18	13	14,61	4875	18,57
	carbon utilization		0,00		0,00	13	0,05
	cell killing		0,00		0,00	5	0,02
	cell proliferation	1	0,70		0,00	83	0,32
	cellular component organization or biogenesis	11	7,69	3	3,37	1946	7,41
	cellular process	67	46,85	49	55,06	13802	52,56
	death	1	0,70	2	2,25	513	1,95
	developmental process	19	13,29	8	8,99	3059	11,65

	establishment of localization	10	6,99	6	6,74	2510	9,56
	growth		0,00		0,00	557	2,12
	immune system process	3	2,10	6	6,74	504	1,92
	localization	10	6,99	6	6,74	2577	9,81
	locomotion		0,00		0,00	25	0,10
	metabolic process	74	51,75	53	59,55	14434	54,97
	multi-organism process	3	2,10	7	7,87	1271	4,84
	multicellular organismal process	15	10,49	8	8,99	2918	11,11
	negative regulation of biological process	3	2,10	1	1,12	619	2,36
	nitrogen utilization		0,00		0,00	6	0,02
	pigmentation		0,00		0,00	7	0,03
	positive regulation of biological process	3	2,10	3	3,37	534	2,03
	regulation of biological process	23	16,08	13	14,61	4592	17,49
	reproduction	8	5,59	3	3,37	1497	5,70
	reproductive process	8	5,59	3	3,37	1482	5,64
	response to stimulus	26	18,18	18	20,22	6259	23,84
	rhythmic process		0,00		0,00	126	0,48
	signaling	10	6,99	6	6,74	1354	5,16
	sulfur utilization		0,00		0,00	1	0,00
	viral reproduction		0,00		0,00	34	0,13
Cellular Component	cell	86	60,14	60	67,42	16432	62,58
	cell junction	5	3,50	3	3,37	1217	4,63
	cell part	86	60,14	60	67,42	16432	62,58
	extracellular region	5	3,50	2	2,25	572	2,18
	extracellular region part		0,00		0,00	26	0,10
	macromolecular complex	7	4,90	2	2,25	1781	6,78

	membrane-enclosed lumen	6	4,20		0,00	711	2,71
	organelle	50	34,97	24	26,97	9294	35,39
	organelle part	24	16,78	6	6,74	3533	13,45
	symplast	5	3,50	3	3,37	1215	4,63
Molecular Function	antioxidant activity		0,00	1	1,12	213	0,81
	binding	57	39,86	41	46,07	10834	41,26
	catalytic activity	60	41,96	41	46,07	10123	38,55
	channel regulator activity		0,00		0,00	1	0,00
	electron carrier activity	7	4,90	3	3,37	453	1,73
	enzyme regulator activity		0,00	1	1,12	346	1,32
	metallochaperone activity		0,00		0,00	6	0,02
	molecular transducer activity	1	0,70		0,00	290	1,10
	nucleic acid binding transcription factor activity	2	1,40	8	8,99	1836	6,99
	nutrient reservoir activity		0,00		0,00	71	0,27
	protein binding transcription factor activity		0,00		0,00	61	0,23
	protein tag		0,00		0,00	7	0,03
	receptor activity	1	0,70		0,00	163	0,62
	structural molecule activity	2	1,40	1	1,12	506	1,93
	translation regulator activity		0,00		0,00	7	0,03
	transporter activity	8	5,59	3	3,37	1562	5,95
*	Total number of unique Sorghum Genes in Expanded Regions (143)						
**	Total number of unique Sorghum Genes in Contracted Regions (89)						
***	Total number of unique Sorghum Genes in Sorghum Genome (26.259)						

CAPÍTULO 2 - Montagem de pools de BACs de sorgo usando dados simulados

O sequenciamento de pool de BACs utilizando as novas tecnologias de sequenciamento possibilita uma redução nos custos para preparar e sequenciar clones de BACs [16,17,18]. Embora o número de clones por pool aplicados nestes trabalhos não impactem na montagem, o aumento do número de clones por pool deve elevar a complexidade da montagem, resultando na diminuição do tamanho dos scaffolds montados. Assim, deve haver um balanço ideal entre custo e complexidade para sequenciar pool de BACs. Para isso, realizamos um estudo baseado em simulações com diferentes números de BACs por pool. As análises foram feitas usando BACs simulados de sorgo, cujo genoma é bem próximo ao genoma de cana-de-açúcar, que é foco de estudo deste trabalho. Devido ao maior rendimento e menor custo entre os sequenciadores da nova geração, foram simulados reads da plataforma Illumina. Um ponto importante destas simulações é maneira com que os pools de BACs são montados. Nos trabalhos anteriores [16,17,18], os pools foram montados a partir de BACs selecionados com base em mapas físicos e genéticos. Neste trabalho, a proposta é misturar BACs tomados aleatoriamente, sem utilizar mapas físicos para seleção dos clones, visto que na prática a construção destes mapas exige grande esforço e impacta o custo. Além disso, a seleção randômica pode favorecer uma menor complexidade das montagens. O racional para isso é que ao tomar clones de BACs ao acaso, a amostragem obtida contemple regiões distintas e diferentes do genoma, e portanto regiões que sejam diferentes a nível de sequência, fator que contribui para reduzir a complexidade no processo de montagem de pools.

Construção de uma biblioteca virtual de BACs de sorgo

Foi criada uma biblioteca virtual de BACs de sorgo com 26.500 clones, compreendendo uma cobertura de 5X do genoma. A amostragem contempla clones variando de 80 kb a 140 kb, com tamanho médio de 120 kb.

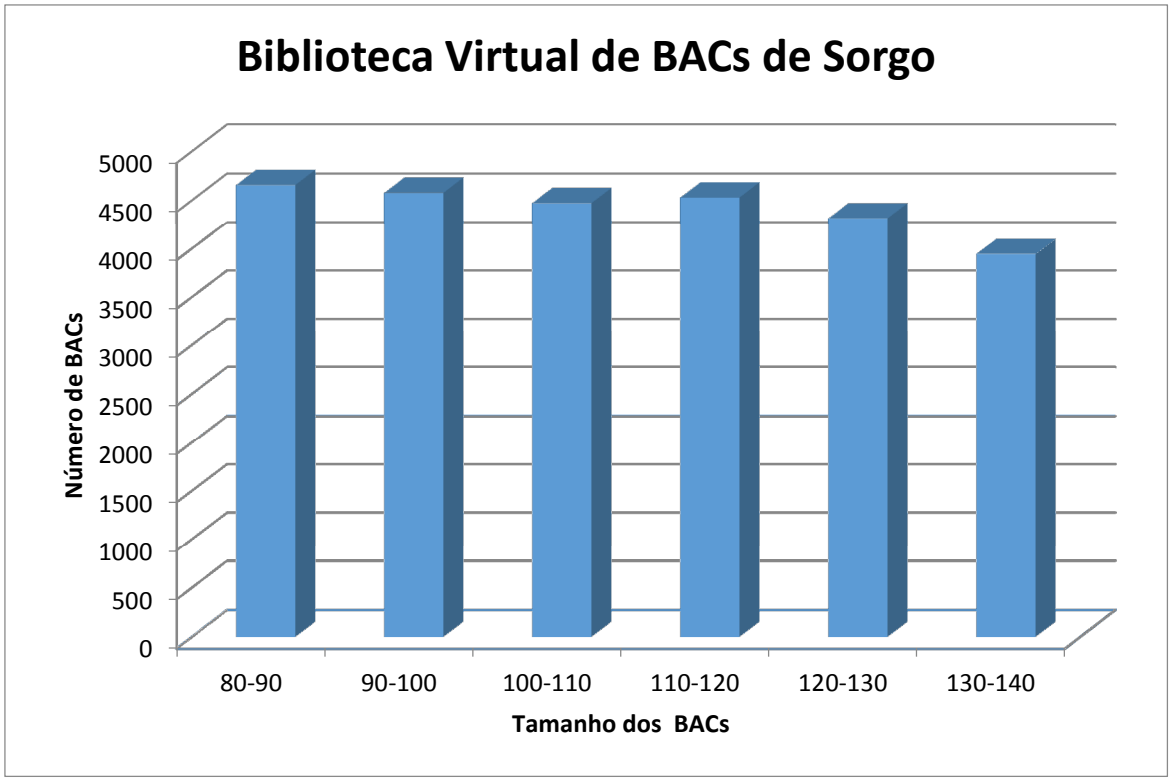


Figura 3: Distribuição do tamanho dos BACs da biblioteca simulada do genoma de sorgo.

Simulação dos reads

Para fazer a simulação de reads Illumina foi usado o software SimSeq [19], desenvolvido e aplicado em uma importante competição para avaliar softwares para montagem *de novo* [20]. O software SimSeq simula reads de bibliotecas de sequenciamento Illumina paired-end (reads de pontas de fragmentos até 800 bp) e mate pair (reads de pontas de fragmentos mais longos). Além disso, modela erros e valores de qualidade para as bases, e simula reads quiméricos das bibliotecas mate pair. Para as bibliotecas paired-end foram gerados reads de 100 bp e para as bibliotecas mate pair foram gerados reads de 50 bp. Foram simuladas para cada pool, bilbiotecas com os seguintes tamanhos de fragmentos: 150 bp, 400 bp, 800 bp, 3 kb, 5 kb e 10 kb. Foi permitida uma variação de 10% no tamanho dos fragmentos. O número de reads foi amostrado uniformemente por biblioteca, correspondendo a uma cobertura de 80X para as bibliotecas paired end e 40X para as bibliotecas mate pair.

Simulação de pools de BACs

Para avaliar o efeito do aumento do número de BACs, foram avaliados pools com 25, 50, 100, 200 e 400 BACs. Para cada tamanho de pool foram geradas 5 réplicas. A seleção dos BACs para cada pool foi feita aleatoriamente, sem informação prévia de mapas físicos ou *minimal tilling path*. Conforme descrito acima, para cada pool foram simulados reads para as bibliotecas paired-end (150 bp, 400 bp, 800 bp) e mate pair (3 kb, 5 kb e 10 kb), totalizando uma cobertura de 360X por BAC. As montagens dos pools de BACs foram realizadas com o software SOAPdenovo [21]. Primeiramente os reads paired-end (bibliotecas de inserto menor) foram montados em contigs. A seguir, os contigs foram juntados em scaffods usando os reads mate pair. Em cada passo é usado um valor diferente de K. No primeiro passo, envolvendo a montagem dos reads paired-end, o valor de K é usado para construir o grafo *de Bruijn* e no segundo passo, o valor de K é usado para mapear os reads mate pair nos contigs. Foram adotados dois pares de valores de K: (75,45) e (85,43). Estes valores de K apresentaram os mellhores montagens em uma simulação com pools de 10 BACs.

Avaliação das montagens

As montagens foram avaliadas em termos do número de scaffolds e da métrica de montagem N50. O valor de N50 corresponde ao tamanho do contig ou scaffold de uma montagem, onde 50% da bases estão em contigs/scaffolds que tenham pelo menos este tamanho. Além disso, as montagens foram avaliadas segundo o valor de cobertura entre as sequências conhecidas do BACs e os scaffolds montados. Para isso, o scaffold com maior similaridade foi mapeado para cada BAC. A seguir, cada par de BAC e scaffold foi alinhado com o cross_match [22], e os alinhamentos foram ordenados para identificar o maior segmento contíguo. O valor da cobertura corresponde a divisão entre o tamanho do maior segmento contíguo e o tamanho da sequência do BAC.

Efeito com o aumento do número de BACs por pool na montagem

As simulações com os diferentes tamanhos de pool estão sumarizadas na Tabela 1. Com o aumento do número de BACs há uma leve diminuição no valor médio de N50,

e um pequeno aumento na proporção entre número médio de scaffolds e número de BACs por pool.

Tabela 1: Estatísticas das montagens dos pools de BACs

Tamanho do Pool	Réplicas	K	Número de Scaffolds	N50 Scaffold
25 BACs	pool1	75,45	35	109.189
		85,43	27	110.066
	pool2	75,45	28	103.458
		85,43	29	99.789
	pool3	75,45	25	108.345
		85,43	26	99.771
	pool4	75,45	39	95.196
		85,43	33	98.225
	pool5	75,45	34	109.930
		85,43	36	93.268
	Média	75,45	32	105.224
		85,43	30	100.224
50 BACs	pool1	75,45	70	108.156
		85,43	57	111.121
	pool2	75,45	76	95.692
		85,43	64	99.087
	pool3	75,45	65	99.834
		85,43	59	99.771
	pool4	75,45	70	110.925
		85,43	63	111.940
	pool5	75,45	75	99.080
		85,43	67	103.516
	Média	75,45	71	102.737
		85,43	62	105.087
100 BACs	pool1	75,45	130	102.370
		85,43	122	105.310
	pool2	75,45	156	96.229
		85,43	132	103.116
	pool3	75,45	135	96.650
		85,43	124	100.636
	pool4	75,45	150	103.643
		85,43	131	106.187
	pool5	75,45	143	101.654
		85,43	129	103.516
	Média	75,45	143	100.109
		85,43	128	103.753
200 BACs	pool1	75,45	296	96.610
		85,43	252	105.166

	pool2	75,45	327	94.050
		85,43	271	103.047
	pool3	75,45	276	98.496
		85,43	252	98.816
	pool4	75,45	286	97.568
		85,43	254	100.476
	pool5	75,45	302	98.859
		85,43	260	101.052
	Média	75,45	297	97.117
		85,43	258	101.711
400 BACs	pool1	75,45	604	95.782
		85,43	558	95.863
	pool2	75,45	616	95.108
		85,43	527	101.676
	pool3	75,45	575	95.930
		85,43	528	96.806
	pool4	75,45	587	98.145
		85,43	518	101.643
	pool5	75,45	616	94.509
		85,43	543	102.647
	Média	75,45	600	95.895
		85,43	535	99.727

A análise da cobertura por pool (Figura 4) confirma a tendência de queda na qualidade da montagem. Com o aumento de BACs por pool, a proporção scaffolds com cobertura acima de 80% diminui, aumentando a proporção de scaffolds com coberturas menores, indicando montagens mais fragmentadas. O número de scaffolds bem montados também é influenciado pela composição dos pools, indicado pela variação entre as réplicas com mesmo número de BACs por pool.

Desenvolvimento de scripts para simulação

Para lidar com a grande quantidade de dados de sequenciamento e montagem, foram desenvolvidos alguns programas/scripts em PERL. Um dos programas foi desenvolvido para gerar sequências de fragmentos *in silico*. O programa recebe como entrada uma sequência de nucleotídeos (S), menor tamanho de uma sequência (m), maior tamanho de uma sequência (M), número de sequências (N), e produz aleatoriamente um número N de sequências de nucleotídeos (subsequências), originadas da sequência S, com tamanhos entre m e M. O

programa simula também a geração de reads *paired end*. Para isso, deve ser informado o tamanho do fragmento.

O principal programa de computador desenvolvido, realiza grande parte do processamento. Dado o número de BACs por pool, o número de réplicas, a especificação das bibliotecas para os diferentes tamanhos de fragmentos e número de reads, o script seleciona aleatoriamente os BACs para cada pool, gera os reads simulados para cada biblioteca de sequenciamento, e faz a montagem dos reads. Outros scripts foram criados para comparar os scaffolds montados com as sequências conhecidas dos BACs e para gerar alguns relatórios das montagens.

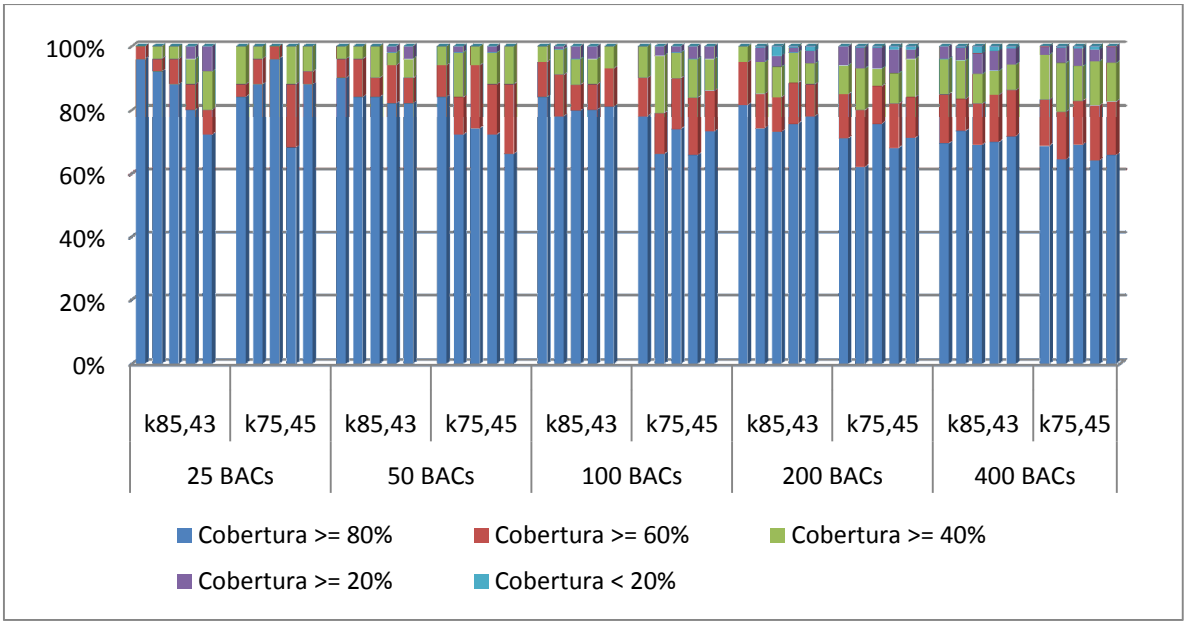


Figura 4: Distribuição das coberturas entre os pools com número crescente de BACs.

CAPÍTULO 3 - BAC-pool sequencing strategy for assembly the complex sugarcane genome

BAC-pool sequencing strategy for assembly the complex sugarcane genome

Vagner Okura[1,2], Rafael Soares Correa de Souza[1], Susely Ferraz de Siqueira Tada[2] and Paulo Arruda[1,3,*]

[1]Centro de Biologia Molecular e Engenharia Genética, Universidade Estadual de Campinas (UNICAMP), 13083-875, Campinas, SP, Brazil

[2]Laboratório Central de Tecnologias de Alto Desempenho em Ciências da Vida (LaCTAD), Universidade Estadual de Campinas (UNICAMP), 13083-875, Campinas, SP, Brazil=

[3]Departamento de Genética e Evolução, Instituto de Biologia, Universidade Estadual de Campinas (UNICAMP), 13083-875, Campinas, SP, Brazil

[*]Corresponding author

Email addresses:
VO: vagnerko@unicamp.br
PA: parruda@unicamp.br
RSCS: scs.rafael@gmail.com
SFST: susy.tada@reitoria.unicamp.br

Abstract

Sequencing plant genomes is often challenging because of its complex genomic architecture and high content of repetitive sequences. Sugarcane has one of the most complex genomes. It is highly polyploid, preserves intact homeologous chromosomes from its parental species and contains over 55% repetitive sequences. Although bacterial artificial chromosome (BAC) libraries have emerged as an alternative for accessing sugarcane genome, sequencing individual clones is laborious and expensive. Here we present a strategy to sequence and assembly reads produced from pooled BAC clones DNA. A set of 178 BAC clones was randomly sampled from the SP80-3280 sugarcane BAC library, pooled and sequenced using Illumina HighSeq2000 and PacBio platforms. A hybrid assembly strategy (AHA) generated 2,451 scaffolds comprising 19.2MB of assembled genome sequence. Scaffolds ≥ 20Kb corresponded to 80% of the assembled sequences and the full sequences of forty BACs were recovered in one or two contigs. Alignment of BAC scaffolds with sorghum chromosome sequences showed high degree of collinearity and gene order. The gene containing scaffolds aligned homogeneously along the sorghum chromosome confirming the randomness of BAC sampling. The alignment of the BAC scaffolds to the 10 sorghum chromosomes indicates that the genome of the SP80-3280 sugarcane variety is ~19% contracted in relation to the sorghum genome. In conclusion our data shows that sequencing pools composed of high numbers of BAC clones may help to construct a reference scaffold mapping of the sugarcane genome.

Keywords: Sugarcane, BAC pool sequencing, synteny, sorghum, sugarcane genome

Introduction

Grasses have evolved by complete duplication of their chromosome sets (REF). Some grasses species show variable degree of ploidy and high content of repetitive sequences (Wang et al., 2010; Wang et al., 2011). This is the case of sugarcane genome. The modern sugarcane varieties are hybrids derived from crosses between *Saccharum officinarum* which has a chromosome constitution of 2n = 80, and *Saccharum spontaneum* which has a chromosome constitution of 2n = 40-128 (Cheavegatti-Gianotto et al., 2011). The commercial varieties grown worldwide have been selected from populations produced by few backcross cycles between the interspecific hybrid and the high sugar content parent *S. officinarum*. This crossing and selection scheme resulted in varieties with chromosome constitutions varying between 2n=100-130 with 5-20% of the chromosomes inherited from *S. spontaneum*, 70-80% inherited from *S. officinarum* and recombinant chromosomes formed between homeologous of both species (Grivet and Arruda, 2001). This complex genomic architecture with multiple homo/homeoalleles at each locus (Daugrois et al., 1996; Guimaraes et al., 1999) makes assembling very difficult using shotgun sequencing, as reads arising from homeoalleles would collapse, making it difficult to recover large consensus sequences or contigs. As a consequence, the complete sugarcane genome sequence has not yet been assembled and it could be envisaged that at some extent a sugarcane consensus genome sequence may comprises mosaic sequence arrangements with impaired biological meaning. However, partial alignment of overlapping regions of large contigs would help understanding the genome organization as different homologous/homeologous chromosomes could be represented in the alignments. Such a reference map could be created by sequencing bacterial artificial chromosome (BAC) libraries and aligning the sequences using the sorghum genome sequence as a syntenic template [Paterson et al. 2009].

Efforts in sequencing sugarcane BAC clones have already been reported (de Setta et al., 2014). In this case the sequences were generated by individually sequencing and assembling each BAC clone. However, this strategy is time consuming and costly as sequencing libraries must be generated from the DNA individually isolated from each BAC clone. An alternative is sequencing polls of BAC

clones, with or without previous mapping, covering the entire genome. BAC pool sequencing has being used to generate megabases (MB) of genome sequence for several species. For example, 3 MB of rice sequences were generated from 6 pools composed of 28 BACs clones each using the 454 sequencing platform (Rounsley et al., 2009). In another example, a pool composed of 8 BACs was used to generate 1 MB of sequences from salmon genome using the 454 platform (Quinn et al., 2008). In these two cases the number of BACs per pool was very small and authors used minimum tiling path to fingerprint the pooled BACs. In a third example, two pools of 35 and 23 BACs from a BAC library constructed from a melon line were sequenced using the 454 platform (Gonzalez et al., 2010).

In this report we describe the sequencing of a sugarcane BAC pool composed of a large number BACs as a cost-effective way of generating large contigs of non-overlapping BAC clones. By random sampling BAC clones from a sugarcane BAC library (Figueira et al., 2012), we were able to generate ~20 MB of sequences assembled into 2,451 scaffolds with a minimum sequence size of 2Kb. By syntenic aligning of scaffolds to the sorghum genome we could access scaffold completeness, the randomness distribution of the scaffolds along sorghum chromosomes, the sugarcane/sorghum synteny and the gene and repetitive sequence content of a sample of the sugarcane genome.

Material and methods

BAC library

The SS_SBa BAC library comprises 36,864 clones prepared with genomic DNA isolated from the sugarcane variety SP80-3280, by HindIII partial digestion and ligation into the pAGIBAC1 vector (Figueira et al., 2012). The library represents approximately 6 genomic equivalents of the monoploid sugarcane genome.

BAC library size determination and pooling

A total of 192 BAC clones were randomly selected from 96 X 384-well plates, two for each plate, and replated into two 96-wells plates. Clones were grown overnight and the cultures used to prepare three additional replicates for the two 96 well plates that was stored at -80^{o}C in Circle Grow medium containing 20% glycerol. The clone inserts size was estimated by NotI restriction enzyme digestion (Figueira et al.,

2012). Each one of 192 clones were individually grown overnight in 50 ml falcon tubes containing 10mL of Circle Grow medium containing 12,5ug/mL chloramphenicol at 37℃ and 300 rpm. A total of 178 clones cultures that growth at ODs ranging from 0.6-1.0 (Supplemental File 1) were pooled, pelleted and the DNA extracted using the QIAGEN Large-Construct Kit.

Illumina sequencing

One µg of DNA prepared from the BAC pool was used to prepare small-insert (150, 400 and 800bp) libraries. For this the DNA was randomly fragmented by sonication using Bioruptor (Diagenode, Denville, NJ, USA) and the desired fragments size-selected by gel electrophoresis. Illumina paired-end sequencing libraries were prepared using TruSeq DNA Sample Preparation Kit V2 and sequenced in a HiSeq2000 platform. Sonication, library preparation and sequencing were carried out at Central Laboratory of High Performance Technologies (LaCTAD) from Universidade Estadual de Campinas (www.lactad.unicamp.br).

PacBio sequencing

A total of 23 µg BAC pool DNA was submitted to Duke University Genome Sequencing & Analysis Core Resource (www.genome.duke.edu) sequencing using the PacBio platform. One large insert library (4kb-10kb) was sequenced in one SMRT cell using the XL-C2 chemistry.

Sequence assembly

Illumina reads were pre-filtered by quality criteria (90% of bases with phred quality 30) and primer/adaptor contamination removal using NGS QC Toolkit (Patel and Jain, 2012). Vector contamination (pBeloBAC11) and E. coli (DH10B) reads were identified and removed using Bowtie (Langmead et al., 2009). Assembly of Illumina reads was performed by Edena (Hernandez et al., 2008). PacBio sequence data was loaded to SMRT Analysis Software v2.1.1 (http://www.pacb.com/devnet), and by applying RS_CeleraAssembler protocol, reads were error corrected with 400X coverage Illumina reads using PacBioToCa (Koren et al., 2012). The corrected reads were assembled with Celera Assembler (Meyers et al., 2000) (PacBio contigs). Illumina Contigs and PacBio contigs were assembled with CeleraAssembler (wgs8.0). Hybrid scaffolding of Illumina contigs using PacBio reads were performed by SSPACE-LongRead (Boetzer and Pirovano, 2014) and by AHA (A Hybrid

Assembler) (Bashir et al., 2012), a module of SMRT Analysis Software. In addition to standard assembly metrics (number of contigs/scaffolds, largest sequence length, N50), sugarcane BAC end sequences (BES) (Figueira et al., 2012) positioning in the assembled contigs/scaffolds was used to validate assemblies. The number of BES uniquely anchored at the end of a contig/scaffold (less than 1000 from sequence end) was considered a parameter to verify the consistency of an assembly (number of correctly anchored BES). BES position in the contigs/scaffolds was determined using BlastN (E-value cutoff of 1e-10). BES uniquely positioned at middle of contigs/scaffolds contributes negatively to the assembly. Complete BAC sequence ("One Contig") was determined as the contig/scaffold that had its corresponding BES pair mapped at the end of their sequence and their length presented a size similar to the expected BAC length.

Sequence analysis

Repeat elements identification and masking were performed by Censor (Kohany et al., 2006) software using grass sequences of Repbase (Jurka et al. 2005). Repeat masked version of scaffold sequences was submitted to gene prediction processing. Genes were predicted using EVidenceModeler (EVM) (Haas et al., 2008) annotation tool by combining predictions from Augustus (Stanke et al., 2008), GlimmerHMM (Majoros et al., 2004) and GeneMark (Lomsadze et al., 2005), and EST alignments processed by PASA (Haas et al., 2003) using SUCEST EST sequences (Vettore et al., 2003). Predicted genes were searched against Swissprot, Uniref90, NCBI non-redundant protein database using BlastX (evalue cutoff of $1e^{-5}$) and searched against SUCEST EST and sorghum CDS using BlastN (evalue cutoff of $1e^{-10}$). Blast2GO software was used to determine GO term and protein code. Masked scaffold sequences ≥ 2,000 bp were mapped to sorghum chromosomes using BlastN (evalue cutoff of $1e^{-10}$) and Perl and shell scripts. High-scoring segment pairs (HSP) were sorted by scaffold positioning and an 'expanded alignment' was determined by joining non overlapping HSPs. Scaffolds with a minimum of 1,000 bp expanded alignment length were considered mapped to sorghum chromosomes. Synteny analysis between sugarcane and sorghum was performed based on the expanded alignment.

Results and discussion

Rational for sugarcane BAC pool sequencing

Taken into account the closest phylogenetic relationship of sugarcane and sorghum, we based our sugarcane BAC pool sequencing rational on the sorghum genome size. The monoploid sorghum genome is comprised of 10 chromosomes with sizes ranging from ~55 to 78 MB (Paterson et al., 2009). The SP80-3280 BAC library used in this work comprises ~37,000 clones with an average size of ~120 Kb (Figueira et al., 2012). Thus, we estimated that a pool of 200 BACs would account for ~24MB of sequences. If the sugarcane genome have a size similar to the ~780MB sorghum genome, a BAC pool of 200 clones would approximately correspond to ~3% of the nucleotide sequence of the sugarcane genome. Thus if these clones are randomly sampled there is 97% chance that they do not overlap. Absence of overlapping would facilitate the assembling process, excepted for the repetitive sequences, as the reads produced from each of the individual BAC clones in the pool will be recovered in an isolated contig. If this rational works it would not be necessary to have any additional information from the individual BACs in the pool. To test the rational, we made a single pool containing DNA individually isolated from 178 BACs from the SP-80-3280 library (Figueira et al., 2012). The DNA from each BAC was pooled at equimolar amounts and sequenced using the Illumina HiSeq2000 paired end reads and the PacBio RSII long reads.

Sequencing output and assembly

For sequencing in the Illumina platform we prepared paired end libraries with insert sizes of 170, 400 and 800 bp using the DNA pool from the 178 BAC clones. Libraries were sequenced in a single lane of the HiSeq2000 resulting in 24.6 Gb of usable reads (Supplemental Table 1). We have previously estimated the size of each BAC clone (Figueira et al., 2012) used to construct the pool (Supplemental Table 2). The sum of the sizes of the 178 BACs was estimated in 21.7 Mb. Thus, the sequence reads produced by the HiSeq2000 platform was in excess of 1,000-fold coverage of the estimated sum of BAC clone sequences. The same DNA pool was sequenced using the PacBio SMRT sequencing platform. Using a single Smart Cell we produced 101,841 reads with an average length of 3,637 bp totaling 370.4 Mb of sequence

corresponding to 17-fold coverage of the estimated sum of BAC clone sequences (Supplemental Table 3).

To assembly the BAC pool sequence reads produced by the two sequencing platforms we tested three hybrid-assembling strategies. The PacBio sequencing platform produces long sequence reads but these reads possess 15-20% base errors while the Illumina sequencing platform produces shorter sequence reads but with higher base accuracy. Thus, in the first strategy, the Illumina reads were used for error correction of PacBio long reads and then the long corrected PacBio reads were assembled using the Celera Assembler (Myers et al., 2000). In the second strategy, the Illumina reads were first assembled using Edena (Hernandez et al., 2008) and than a hybrid assembly was performed using the Illumina assembled contigs and the PacBio contigs assembled in the first strategy. The hybrid assembly was performed using Celera Assembler. In the third strategy hybrid scaffolding was performed in which the PacBio corrected reads were used to anchoring the Illumina assembled contigs. The assembly results from the three different strategies were examined according to standard assembly metrics (number of contigs/scaffolds, largest contig/scaffold length, N50 value) and two additional criteria: anchoring BAC end sequences (BES) (Figueira et al., 2012) to the assembled scaffolds and the number of large contigs corresponding to the estimated BAC size (Supplemental Table 4). Scaffolds generated by the hybrid assembler (AHA) (Bashir et al., 2012) tool produced the best assembly and was chosen as the reference assembly of the BAC pool sequences. Although not having the best N50 value, AHA assembly resulted in the lower number of scaffolds, scaffolds with the largest sizes, the highest number of BES correctly anchored at the scaffolds ends and the highest number of contigs corresponding to complete BAC sequences. The AHA assembly generated 2,451 scaffolds corresponding to a total of 19.2 MB sequences (Table 1), which accounted for 88.2% of the 21.7 MB estimated sum of bases of the 178 BACs in the pool. The difference between the 19.2 MB assembled sequences and the 21,7 MB of estimated sum of BAC sizes could be due the inaccuracy of the BAC size estimation by partial restriction digestion and gel electrophoresis fractionation.

Table 1 - Size distribution of assembled scaffolds

Scaffold length	N° scafolds	Total bases	Bases (% total)
< 2.000	1,758	743,310	3.88
2,000 - 10,000	321	1,700,748	12.76
10,000 - 20,000	104	1,480,189	20.49
20,000 - 40,000	110	3,309,119	37.76
40,000 - 60,000	63	3,087,796	53.88
60,000 - 80,000	36	2,464,222	66.74
80,000 - 100,000	31	2,715,881	80.92
100,000 - 120,000	7	797,899	85.09
120,000 - 140,000	16	2,062,825	95.86
140,000 - 160,000	4	590,630	98.94
> 160,000	1	203,132	100.00
Total	2,451	19,155,751	
Estimated total bases		21,717,887	

Scaffolds larger than 20 Kb accounted for ~80% of the assembled sequences. A total of 8 BACs was recovered as one contig as compared with the estimated BAC size. The one contig scaffolds were considered complete assembled BACs as its BES exactly anchored to the termini sequence of the scaffold (Supplemental Table 5). Furthermore, scaffolds with one unique correctly anchored BES were analysed and 32 additional BACs represented by two scaffolds have sum equivalent of the estimated BAC size (Figueira et al., 2012) (Supplemental Table 6). Collinearity analysis of sugarcane scaffolds along sorghum chromosomes showed 133 scaffolds sharing two or more collinear genes with sorghum chromosomes indicating preserved gene order and correctness of the assembly (Supplemental Table 7). The recovery of these additional BAC clones with complete insert sequence along with the syntenic gene orders along sorghum chromosomes represents additional validation of the correctness of AHA assembled scaffolds. Finally, we retrieved from NCBI the nucleotide sequences of two sugarcane BACs (GI:530278086, GI:530279041) that matched to four sugarcane scaffolds assembled from our BAC pool sequencing. The alignments showed high level of sequence identity indicating a high accuracy of assembled nucleotides sequence of our scaffolds (Supplemental Figure 1). Thus we concluded that the sequencing strategies used in this work that involved generating short high accuracy reads from Illumina platform and long reads

from PacBio platform and the use of AHA assembling process resulted in a cost effective manner to generate high accurate long contigs of sugarcane pools composed of high number of BAC clones.

Content and nature of repetitive sequences

Sequence analysis of the 19.1MB assembled nucleotides revealed a content of 54.6% of repetitive sequences among which transposable elements are the predominant group comprising 53.3% of total repetitive sequence bases. Among the group of transposable elements, the long terminal repeat (LTR) category is the most abundant comprising 43.3% of total bases, followed by DNA transposons with 7.7% and Non-LTR retrotransposon with 2.25% (Table 2). Among the LTR group, the *Gypsy* and *Copia* elements accounted respectively for 30.3% and 12.9% of assembled nucleotides. Simple repeats, integrated viruses and unclassified repeat sequences accounted for 1.08, 0.23 and 0.02% of de total bases, respectively. These data are in accordance of previous repetitive elements found in a total of 317 sequenced sugarcane BACs (de Setta et al., 2014). We have previous estimated a slightly smaller proportion of repetitive regions (45.6%) based on BAC end sequences (Figueira et al., 2012). Our new estimates are more accurate as they are based on a large sequence dataset. The ratio of *Gypsy* and *Copia* LTR elements was 2.3:1 that is higher than that observed in the 317 sequenced sugarcane BACs (de Setta et al., 2014). We believe the differences reflects the fact that our data is based on sequences evenly distributed across the genome as reveled by scaffold anchoring along the sorghum chromosomes.

Table 2 - Summary of repetitive sequences among the sugarcane BACs

Repeat element	Number of Elements	Length (bp)	% of Total Bases
Transposable Element	1,279	10,209,529	53.30
DNA transposon	407	1,479,344	7.72
EnSpm/CACTA	102	545,515	2.85
Harbinger	77	282,933	1.48
Helitron	28	74,200	0.39
Mariner/Tc1	11	7,401	0.04
MuDR	44	218,853	1.14
hAT	87	162,922	0.85
Other	58	187,520	0.98
LTR Retrotransposon	732	8,297,946	43.32
Copia	291	2,473,755	12.91
Gypsy	426	5,795,891	30.26
Other	15	28,300	0.15
Non-LTR Retrotransposon	138	431,477	2.25
Other	2	762	0.004
Simple Repeat	9	206,466	1.08
Satellite	9	206,466	1.08
Integrated Virus	3	43,599	0.23
Caulimoviridae	3	43,599	0.23
Unclassified	5	3,965	0.02
Total of repeat elements	1,296	10,463,559	54.62
Total of assembled bases		19,155,751	100.00

Syntenic mapping of scaffolds into sorghum chromosomes

A total of 292 scaffolds corresponding to 12.4 Mb (67.8% of assembled sequences) with a minimum size of 2Kb were mapped by syntenic sequence alignment to the nucleotide sequences of sorghum chromosomes (Table 3). To avoid miss alignment of scaffolds at several locations within and among the sorghum chromosomes, the repetitive sequences were masked. In general scaffolds aligned with high accuracy and were homogeneous distributed along the 10 sorghum chromosomes, except for the chromosomes 6, 8 and 10 that had smaller number of mapped scaffolds. Scaffolds are slightly higher represented in chromosomes 1, 3 and 5. No mapped scaffolds correspond to sequence with high repeat sequence composition. In terms of localization, the 292 scaffolds aligned homogeneously over sorghum chromosomes (Figure 1). This alignment uniformity must be directly related to the random selection of BACs.

Table 3 - Scaffolds longer than 2,000 bp mapped to Sorghum chromosomes

Chromosome	Total Bases	Bases (% Total)	N° Scaffolds	Scaffold size range
1	1,821,039	9,89	42	3,008 - 203,132
2	1,607,121	8,73	36	3,106 - 152,524
3	1,589,753	8,63	44	2,882 - 138,830
4	1,716,141	9,32	37	3,734 - 141,339
5	1,121,959	6,09	40	2,086 - 125,065
6	689,413	3,74	15	6,387 - 97,701
7	1,268,000	6,89	22	7,190 - 151,964
8	722,184	3,92	14	10,282 - 123,118
9	1,270,664	6,90	24	4,562 - 137,971
10	675,823	3,67	18	2,688 - 135,690
No Mapped	5,930,344	32,21	401	2,003 - 129,783
Total	18,412,441		693	

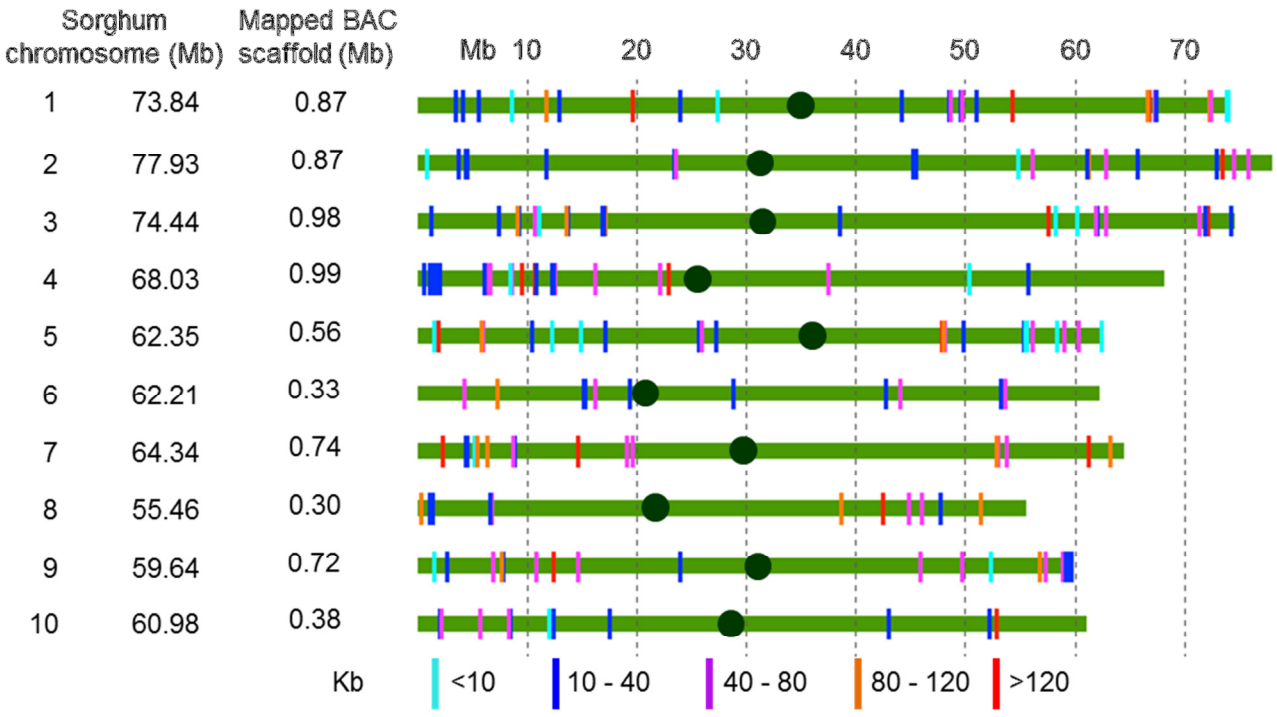


Figure 1 - Orthologous alignment of assembled BAC scaffolds on the 10 sorghum chromosomes. Scaffold sequences were aligned along the sorghum chromosome sequences. Repetitive sequences were masked to avoid misalingment. The coloured solid lines represent the sorghum chromosomes. The coloured vertical bars represent the sugarcane scaffolds.

Gene content and distribution among scaffolds

The annotation pipeline based on *ab initio* gene predictions combined with spliced alignments of transcripts generated a set of 1,338 gene models. Predicted genes were distributed in 431 scaffolds, which correspond to 15.4 Mb (80.57%) of total assembled sequences (Supplemental Table 8). Among scaffolds containing predicted genes, 245 sequences have two or more genes and 16 sequences have ten or mores genes. Gene density was estimated to be 3.1 genes per scaffold with a coding average size of 713 bp, exon average size of 246 bp and intron average size of 647 bp. A total of 884 genes (66.1%) presented similarity to protein databases with 565 (63.9%) of them being supported by sugarcane EST sequences (SUCEST) (Vettore et al., 2001). Genes were classified using the Gene Ontology (GO) functional categories (Supplemental Figure 2). An amount of 2,330 GO terms were assigned to 558 genes. Biological Process GO category comprised 41.9% of the identified terms, with the most representative classes are involved in metabolic, cellular and single-organism process. Catalytic activity and binding are the two most representative classes in Molecular Function category (33.4% of terms). Most of terms were assigned to cell, organelle and membrane classes for Cellular Component category (24.7% of terms). Collinearity of genes between sugarcane and sorghum was found in 133 scaffolds (>= 2 genes) containing 431 genes (Supplemental Table 7).

Sugarcane and sorghum genome comparison

A customized BLAST pipeline was applied to map the sugarcane scaffolds onto the sorghum chromosomes and determine the syntenic regions. Our results shows expanded and contracted regions between sugarcane and sorghum (Supplemental Figure 3). A summary of expanded and contracted regions shows a positive rate of sorghum syntenic regions in relation to sugarcane to all sorghum chromosomes (1.04 – 1.41) (Table 4). Taking all the regions into account, a total of 6,550,682 bp of sugarcane syntenic regions were aligned to 7,809,102 bp of sorghum chromosomes, showing an expansion of sorghum genome of 19% in comparison to sugarcane BAC scaffolds. This result is in keeping with previous studies where sorghum genome may be 20% to 30% longer than sugarcane genome (Figueira et al., 2012; Wang et al., 2010).

Table 4 - Expanded and Contracted Regions between Sorghum and Sugarcane

Chromosome	Number of Scaffolds	Scaffold Mapped size	Chr Mapped size	Syntenic regions rate between Sorghum and Sugarcane	Sum of Sugarcane expanded regions (bp)	Sum of Sugarcane contracted regions (bp)	Sum of Sorghum expanded regions (bp)	Sum of Sorghum contracted regions (bp)
1	42	866,716	905,270	1.04	556,830	309,886	608,863	296,407
2	36	857,217	1,124,624	1.31	547,673	309,544	830,799	293,825
3	44	977,446	1,036,227	1.06	627,837	349,609	601,346	434,881
4	37	994,271	1,115,670	1.12	708,032	286,239	653,976	461,694
5	40	555,615	721,043	1.30	327,292	228,323	568,417	152,626
6	15	328,515	455,977	1.39	163,097	165,418	354,613	101,364
7	22	705,244	856,835	1.21	418,892	286,352	635,808	221,027
8	14	296,242	417,336	1.41	140,887	155,355	320,335	97,001
9	24	590,229	699,434	1.19	384,530	205,699	449,331	250,103
10	18	379,187	476,686	1.26	281,509	97,678	264,014	212,672
Total	292	6,550,682	7,809,102	1.19	4,156,579	2,394,103	5,287,502	2,521,600

Conclusion

Sequencing complex genomes such as the one of sugarcane is challenging due to the interspecific hybrid nature of the crop, high degree of ploidy and high proportion of repetitive DNA sequences. Furthermore, the presence of variable sequence size along noncoding and repetitive and regions among multiple homologous and homeologous chromosomes makes difficult to use shotgun approaches from NGS platforms such as Illumina that generates short reads.. To avoid such immense difficulty it has been suggest and already taken by several research groups the strategy to sequence BAC libraries prepared from sugarcane genomic DNA. However, sequencing individuals BAC is costly and time consuming. In this work, we have tested a cost effective strategy to sequence BAC libraries in a pool arrangement. To test this strategy, Illumina and PacBio platforms were used to sequence 178 BAC clones randomly sampled from a sugarcane BAC library. The completeness of scaffolds as verified by several criterions and most important the alignment of the scaffolds into the sorghum chromosomes strongly support the idea that pooling high number of sugarcane BAC clones randomly chosen from libraries is a very cost effective way to produce a sugarcane genome sequence map. The genome information produce from this work is highly valuable in terms of unraveling the structure and sequence composition of sugarcane genome. Such information allowed us to concluded, for example, that the sequenced sugarcane scaffolds aligned to sorghum chromosomes is ~19% contracted in relation to the sorghum syntenic regions. This information raises the question if this is because the assembled sequenced produced represents only 3% of the sugarcane genome or if this is a particularity of the genome of the SP80-3280 sugarcane variety that may be smaller than the sorghum genome, while other BAC sequences produced from the R570 sugarcane variety have indicated a genome size higher that the sorghum genome.

Accession

This BAC pool Whole-Genome Shotgun reads has been deposited at NCBI GenBank under the accession PRJNA299804.

Authors contribution

Vagner Okura and Paulo Arruda conceived the project; Rafael S. C. de Souza and Susely F. S. Tada prepared the BAC clone cultures, isolated DNA, prepared the sequencing libraries and executed the Illumina sequencing. Vagner Okura assembled the BAC sequences and performed the sequence analysis; Vagner Okura, Rafael S. C. de Souza and Paulo Arruda wrote the manuscript.

Conflict of Interest Statements

The authors declare that the research has been conducted in the absence of potential conflict of interest.

Acknowledgments

This study was funded by FAPESP – 10/50114-4. Paulo Arruda is a CNPq productivity research fellow. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

References

Bashir, A., Klammer, A. A., Robins, W. P., Chin, C. S., Webster, D., Paxinos, E., et al. (2012) A hybrid approach for the automated finishing of bacterial genomes. *Nat. Biotechnol*. 30, 701-707.

Boetzer, M., and Pirovano, W. (2014) SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. *BMC Bioinformatics* 15, p. 211.

Cheavegatti-Gianotto, A., de Abreu, H. M. C., Arruda, P., Bespalhok Filho, J. C., Burnquist, W. L., Creste, S., et al. (2011) Sugarcane (Saccharum X officinarum): A Reference study for the regulation of genetically modified cultivars in Brazil. *Tropical Plant Biol*. 4, 62-89.

Daugrois, J. H., Grivet, L., Roques, D., Hoarau, J. Y., Lombard, H., Glaszmann, J. C., D'Hont, A. (1996) A putative major gene for rust resistance linked with a RFLP marker in sugarcane cultivar R570. *Theor. Appl. Gen*et. 92, 1059-1064.

de Setta, N., Monteiro-Vitorello, C. B., Metcalfe, C. J., Cruz, G. M. Q., Del Bem, L. E., Vicentini, R., et al. (2014) Building the sugarcane genome for biotechnology and identifying evolutionary trends. *BMC Genomics* 15, p. 540.

Figueira, T. R., Okura, V., Rodrigues da Silva, F., da Silva, M. J., Kudrna, D., Ammiraju, J. S., et al. (2012). A BAC library of the SP80-3280 sugarcane variety (saccharum sp.) and its inferred microsynteny with the sorghum genome. *BMC Res. Notes* 5, p. 185.

González, V. M., Benjak, A., Hénaff, E. M., Mir, G., Casacuberta, J. M., Garcia-Mas, J., et al. (2010) Sequencing of 6.7 Mb of the melon genome using a BAC pooling strategy. *BMC Plant Biol.* 10, 246.

Grivet, L., and Arruda, P., (2001) Sugarcane genomics: depicting the complex genome of an important tropical crop. Current Opinion in Plant Biology 5, 122-127.

Guimaraes, C. T., Honeycutt, R. J., Sills, G. R., and Sobral, B. W. S. (1999) Genetic maps of *Saccharum officinarum* L. and Saccharum robustum. Genet. Mol. Biol. 22, 125-132.

Haas, B. J., Delcher, A. L., Mount S. M., Wortman, J. R., Smith, R. K., Hannick, L. I., (2003) Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. 31, 5654-5666.

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., et al. (2008) Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol*. 9, p. R7.

Hernandez, D., François, P., Farinelli, L., Osteras, M., and Schrenzel, J. (2008) De novo bacterial genome sequencing: Millions of very short reads assembled on a desktop computer. *Genome Res*. 18, 802-809.

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res*. 110, 462-467.

Kohany, O., Gentles, A. J., Hankus, L., and Jurka, J. (2006) Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinformatics* 7, p. 474.

Koren, S., Schatz, M. C., Walenz, B. P., Martin, J., Howard, J. T., Ganapathy, G., et al. (2012) Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat. Biotechnol*. 30, 693-700.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L., (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*. 10, p. R25.

Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y. O., and Borodovsky, M. (2005) Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res*. 33, 6494-6506.

Majoros, W. H., Pertea, M., and Salzberg, S. L., (2004) TigrScan and GlimmerHMM: Two open source ab initio eukaryotic gene-finders. *Bioinformatics* 20, 2878-2879.

Myers, E. W., Sutton, G. G., Delcher, A. L., Dew, I. M., Fasulo, D. P., Flanigan, M. J, et al. (2000) A whole-genome assembly of Drosophila. *Science* 287, 2196-2205.

Patel, R.K., and Jain, M., (2012) NGS QC toolkit: A toolkit for quality control of next generation sequencing data. *PLoS One*, 7, no. 2.

Paterson, A. H., Bowers, J. E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., et al. (2009) The Sorghum bicolor genome and the diversification of grasses. *Nature* 457, 551-556.

Quinn, N. L., Levenkova, N., Chow, W., Bouffard, P., Boroevich, K. A., Knight, J. R., et al. (2008) Assessing the feasibility of GS FLX pyrosequencing for sequencing the Atlantic salmon genome. *BMC Genomics* 9, 404.

Rounsley, S., Marri, P. R., Yu, Y., He, R., Sisneros, N., Goicoechea, J. L., et al. (2009) De novo next generation sequencing of plant genomes. *Rice* 2, 35-43.

Stanke, M., Diekhans, M., Baertsch, R., and Haussler, D. (2008) Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* 24, 637-644.

Vettore, A. L., da Silva, F. R., Kemper, E. L., Souza, G. M., da Silva, A. M., Ferro, M. T., et al. (2003) Analysis and functional annotation of an expressed sequence tag collection for tropical crop sugarcane. *Genome Res*. 13, 2725-2735.

Wang, J., Roe, B., Macmil, S., Yu, Q., Murray, J. E., Tang, H., et al. (2010) Microcollinearity between autopolyploid sugarcane and diploid sorghum genomes. *BMC Genomics* 11, p. 261.

Wang, X., Tang, H., and Paterson, A. H. (2011) Seventy million years of concerted evolution of a homoeologous chromosome pair, in parallel, in major poaceae lineages. Plant Cell 23, 27-37.

Figure Legends

Figure 1 - Orthologous alignment of assembled BAC scaffolds on the 10 sorghum chromosomes. Scaffold sequences were aligned along the sorghum chromosome sequences. Repetitive sequences were masked to avoid misalignment. The coloured solid lines represent the sorghum chromosomes. The coloured vertical bars represent the sugarcane scaffolds.

Supplemental Figure 1 - Alignment of assembled scaffolds nucleotide sequences to nucleotide sequence regions of sequenced BACs deposited in NCBI. In the two cases the scaffolds aligned to contiguous regions of the BACs indicating the correctness of the assembly.

Supplemental Figure 2 - Gene Ontology categorization of sugarcane genes annotated in the 19.2MB of scaffold sequences.

Supplemental Figure 3 - Examples of contracted (A and B) and expanded (C and D) regions of sugarcane scaffolds in relation to overlapping regions of sorghum chromosomes.

Supplemental Tables

Supplemental Table 1: Illumina HiSeq2000 sequencing of Sugarcane BAC pool

Supplemental Table 2: List of BACs pooled with their size in nucleotides as determined by NotI restriction digestion analysis

Supplemental Table 3: PacBio Sequencing of Sugarcane BAC pool

Supplemental Table 4: BAC pool assembly numbers and metrics

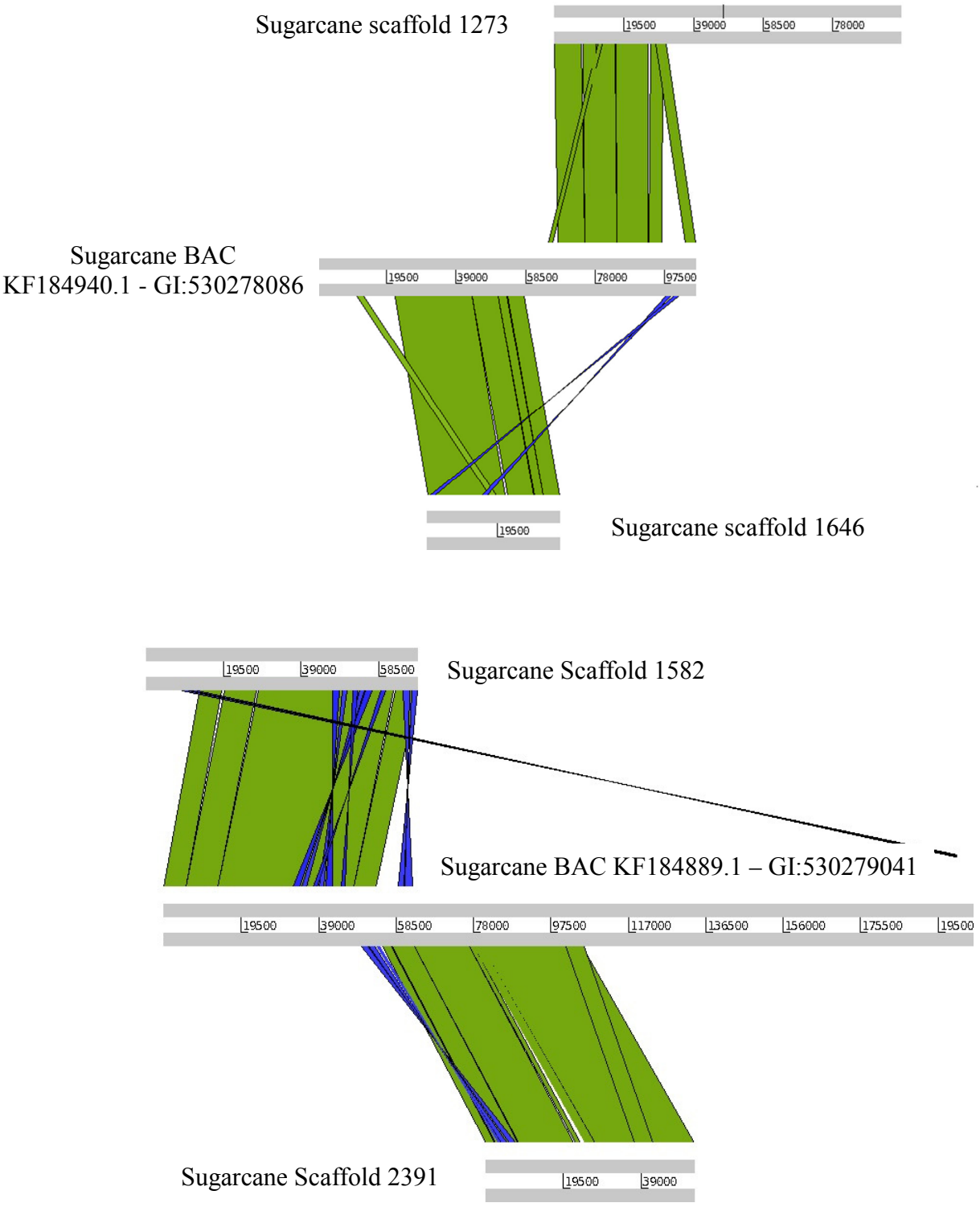
Supplemental Table 5: BACs assembled in “One Contig” – BES match consistently at end of scaffolds and scaffold length is similar to expected BAC length

Supplemental Table 6: BACs assembled in two scaffolds - BES match consistently at end of scaffolds and sum of scaffolds length is similar to expected BAC length

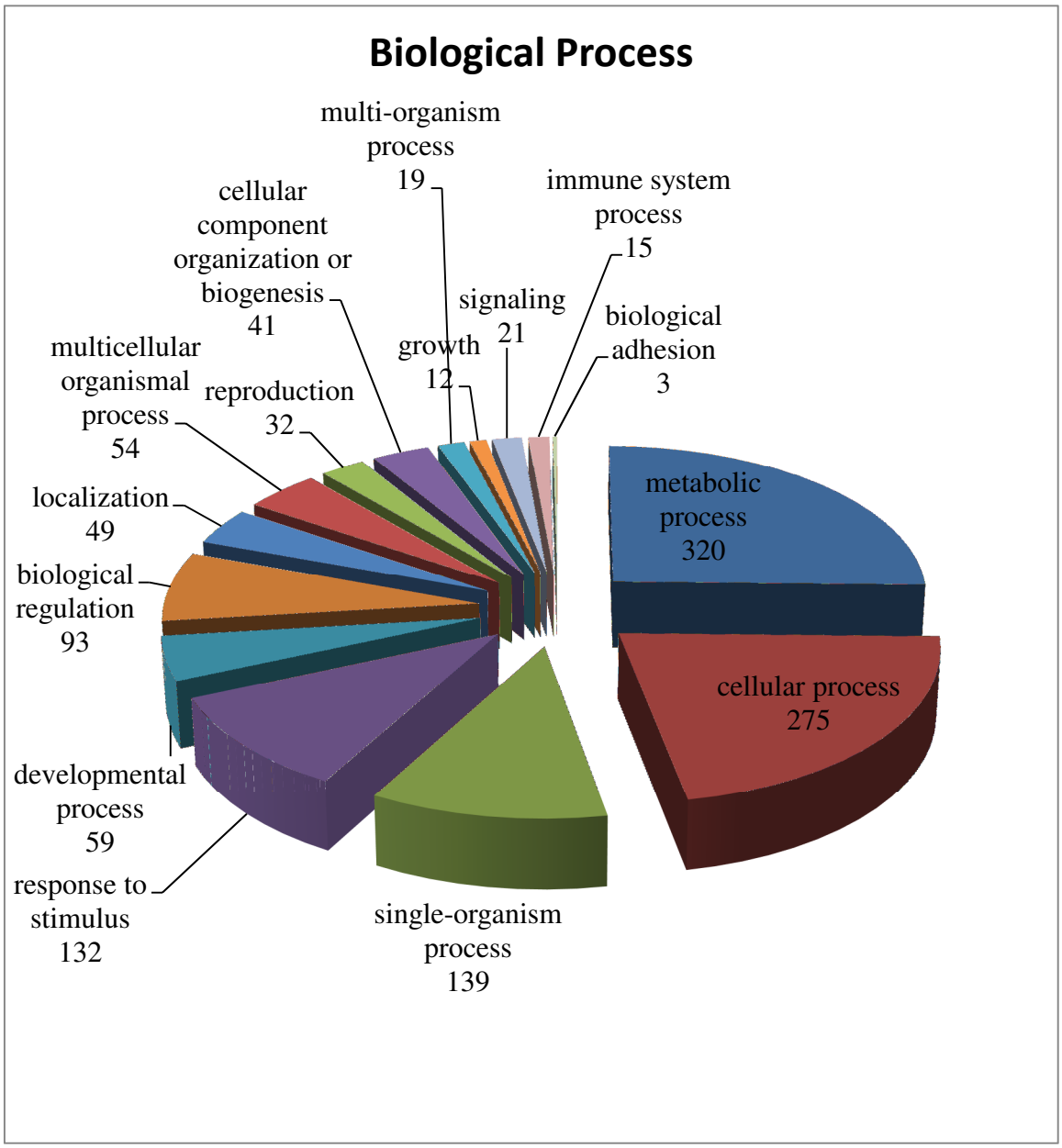
Supplemental Table 7: Collinearity between Sugarcane genes and Sorghum genes. 133 Sugarcane scaffolds present gene collinearity with Sorghum.

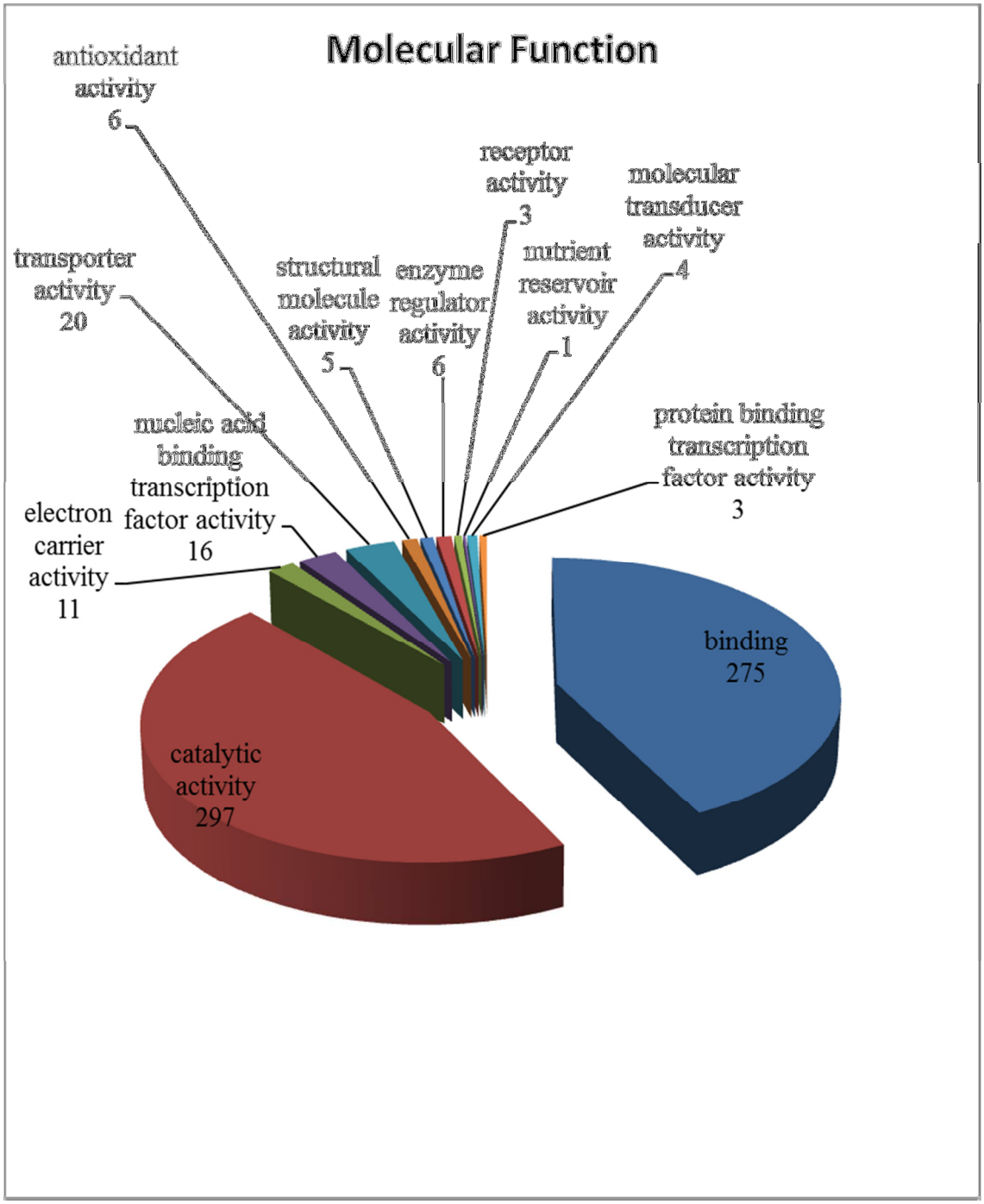
Supplementary Table 8: Distribution of gene models among scaffolds

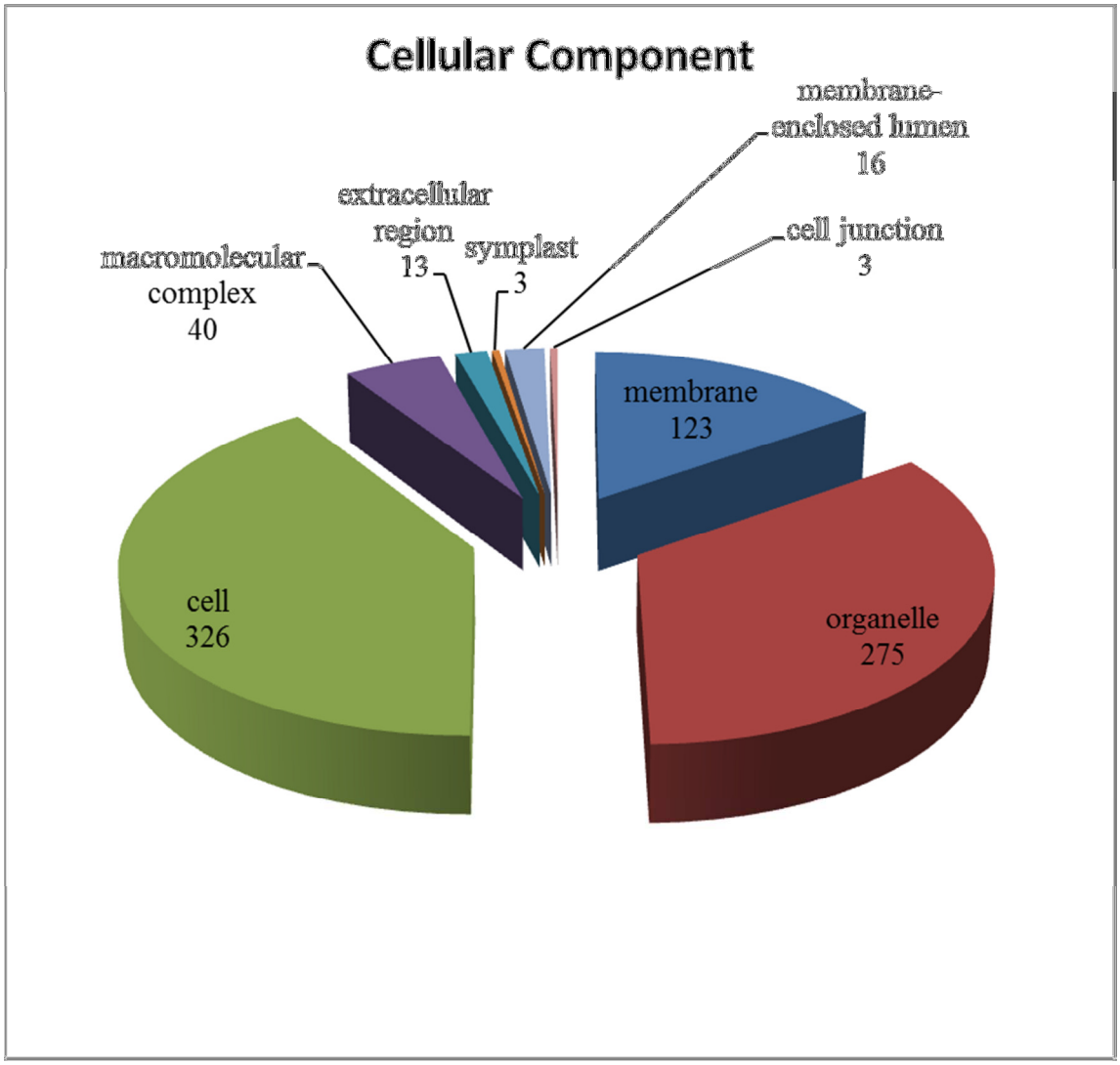
Supplemental Figure 1 - Alignment of assembled scaffolds nucleotide sequences to nucleotide sequence regions of sequenced BACs deposited in NCBI. In the two cases the scaffolds aligned to contiguous regions of the BACs indicating the correctness of the assembly.



Supplemental Figure 2

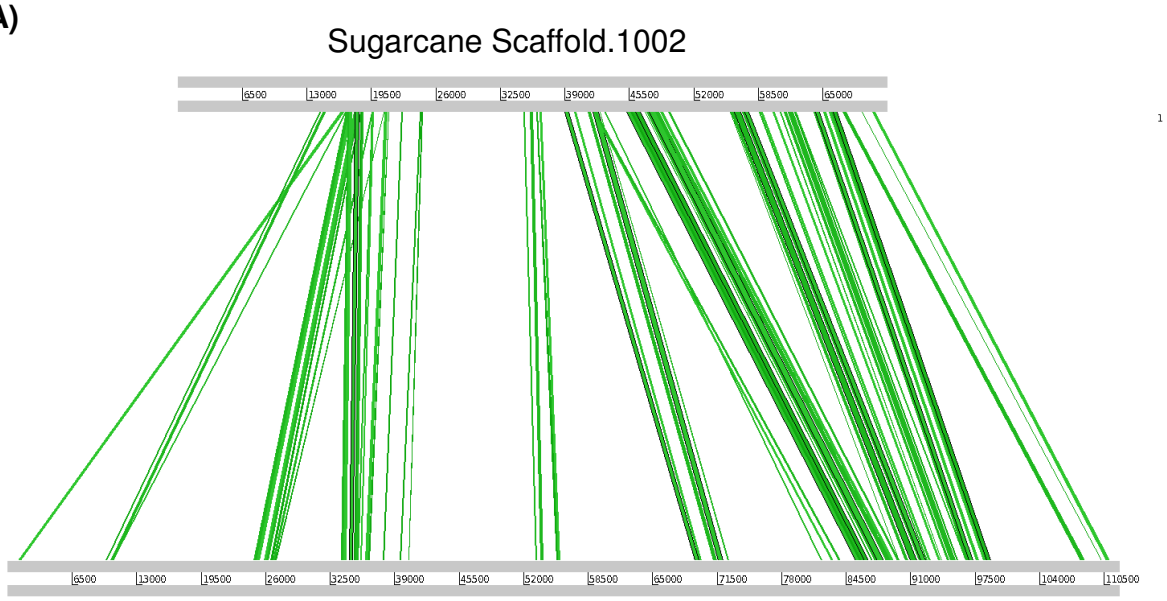



Molecular Function
antioxidant activity 6
receptor activity 3
molecular transducer activity 4
transporter activity 20
structural molecule activity 5
enzyme regulator activity 6
nutrient reservoir activity 1
protein binding transcription factor activity 3
nucleic acid binding transcription factor activity 16
electron carrier activity 11
binding 275
catalytic activity 297


Cellular Component
membrane-enclosed lumen 16
extracellular region 13
symplast 3
cell junction 3
macromolecular complex 40
membrane 123
cell 326
organelle 275

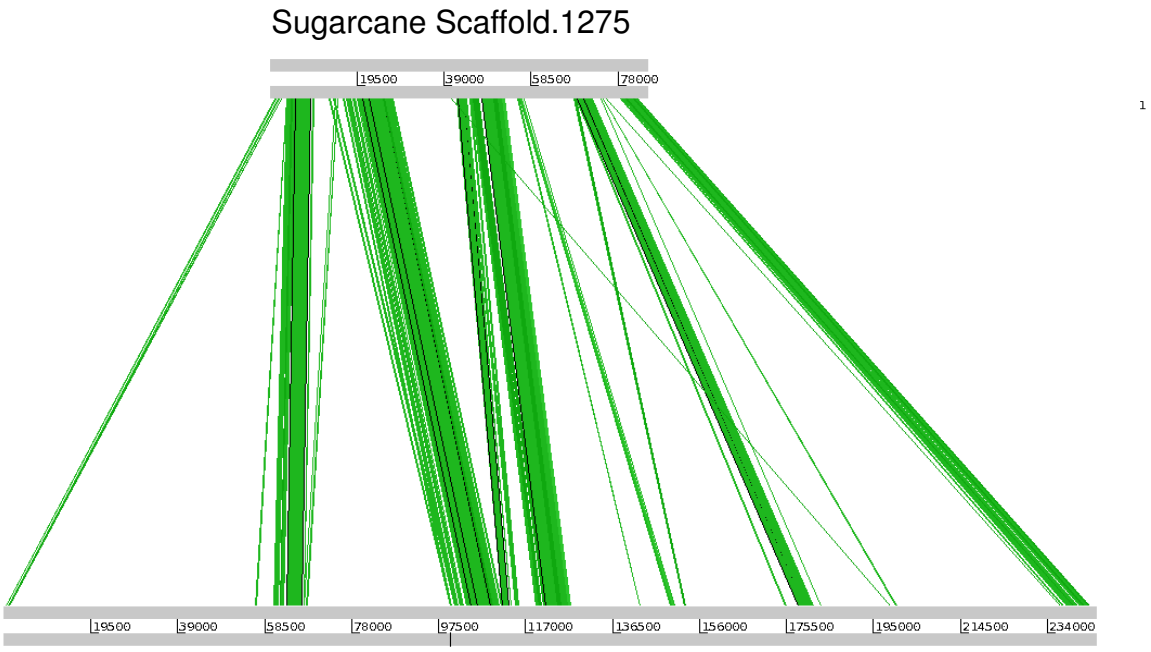
Supplemental Figure 3 - Examples of contracted (A and B) and expanded (C and D) regions of sugarcane scaffolds in relation to overlapping regions of sorghum chromosomes.

(A)



Sorghum chromosome 4 (8,406,000-8,517,000)

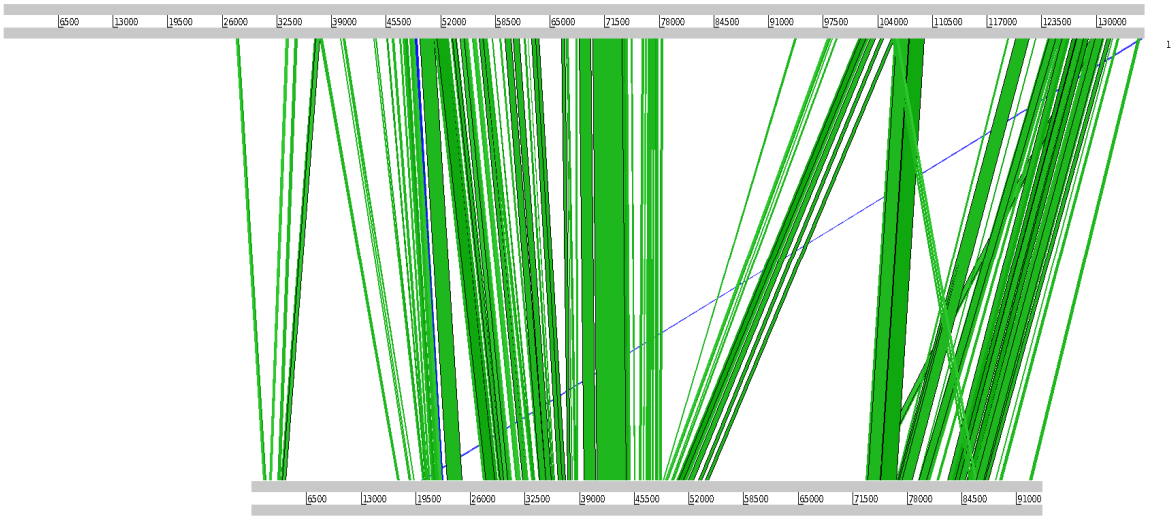
(B)



Sorghum chromosome 6 (4,116,000-4,360,000)

(C)

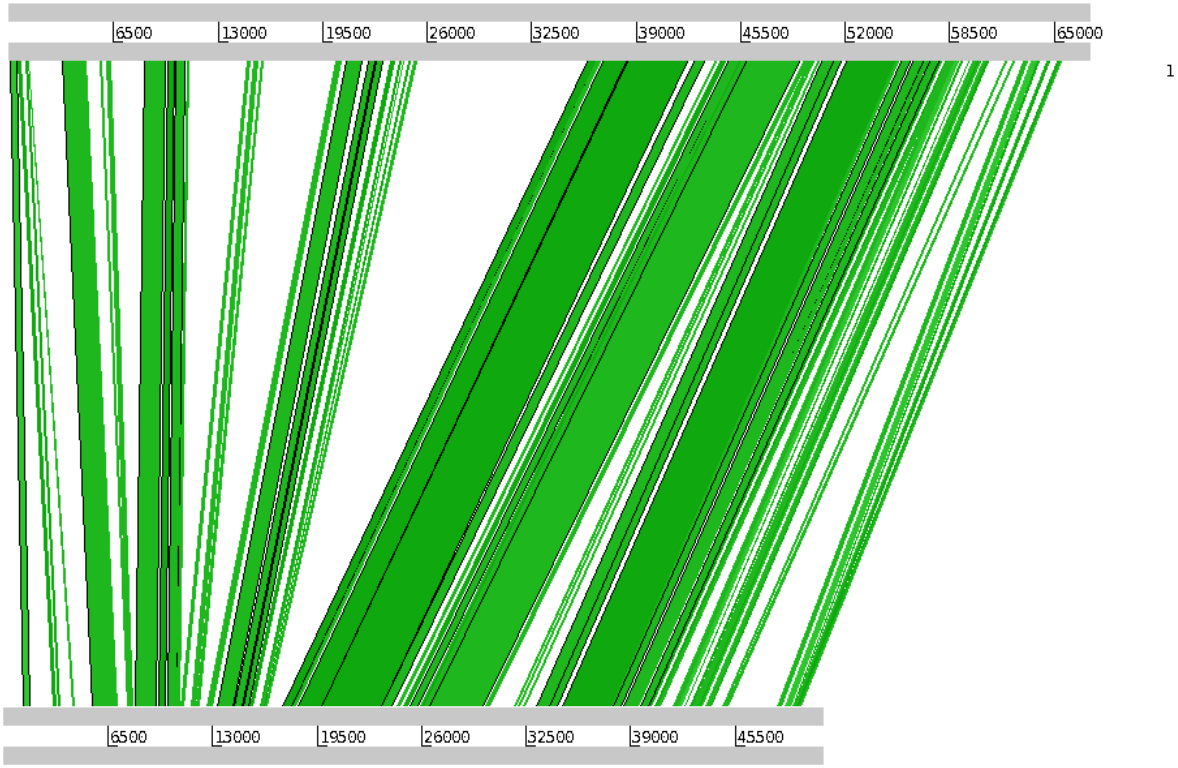
Sugarcane Scaffold.1071



Sorghum chromosome 10 (52,659,000-52,750,000)

(D)

Sugarcane Scaffold.1229



Sorghum chromosome 2 (62,725,000-62,774,000)

Supplemental Table 1 - Illumina HiSeq2000 sequencing of Sugarcane BAC pool

Library	Number Of Reads	Number of Reads after Quality Filtering	Number Of Reads After E. coli Filtering	Number of Reads after pbelo vector Filtering	Number of Bases after pbelo vector Filtering (bp)	Coverage per BAC
BAC170	123,610,182	96,092,842	94,999,141	86,788,234	8,678,823,400	400
BAC400	173,447,366	121,857,784	120,665,875	108,149,370	10,814,937,000	498
BAC800	123,313,864	60,511,394	59,824,685	51,776,460	5,177,646,000	238
Total	420,371,412	278,462,020	275,489,701	246,714,064	24,671,406,400	1,136

Supplemental Table 2 - List of BACs pooled with their size in nucleotides as determined by NotI restriction digestion analysis

BAC Name	96 well pl.address	Copy 2 Pl#	Well Picked	Fragment length 1	Fragment length 2	Fragment length 3	Fragment length 4	Fragment length 5	Fragment length 6	Fragment length 7	Fragment length 8		Total BAC size (Kb)
	SS_SBa	Pl#1											
SCSBa0001A24	a1	1	A24	73,83	18,59	17,46	15,66						125,54
SCSBa0001C24	a2	1	C24	61,9	28,89	22,48							113,27
SCSBa0002A24	a3	2	A24	80,09	23,62	11,48							115,19
SCSBa0002C24	a4	2	C24	73,83	36,86	13,97							124,66
SCSBa0003A24	a5	3	A24	65,92	38,21	13,64							117,77
SCSBa0003C24	a6	3	C24	42,93	31,91	27,55	18,25	10,21					130,85
SCSBa0004A24	a7	4	A24	39,12	28,36	25,95	14,99	14,42					122,84
SCSBa0004C24	a8	4	C24	73,23	29,27	22,48							124,98
SCSBa0005A24	a9	5	A24	22,71	20,88	18,03	16,9	14,87	11,48				104,87
SCSBa0005C24	a10	5	C24	29,65	26,75	11,48							67,88
SCSBa0006C24	a12	6	C24	105,9	23,62	18,25							147,77
SCSBa0007A24	b1	7	A24	37,31	32,81	16,23							86,35
SCSBa0007C24	b2	7	C24	38,21	17,69	11,87							67,77
SCSBa0008A24	b3	8	A24	62,57	57,21								119,78
SCSBa0008C24	b4	8	C24	51,96	42,93	24,35	12,06	10,56					141,86
SCSBa0009A24	b5	9	A24	103,4	99,67	17,35	12,74						233,16
SCSBa0009C24	b6	9	C24	44,85	30,56	24,88	21,11						121,4
SCSBa0010A24	b7	10	A24	54,59	36,41								91
SCSBa0011C24	b10	11	C24	96,5	70,81	50,61	34,16	30,1	10,56				292,74
SCSBa0012A24	b11	12	A24	63,24	49,33	28,89							141,46
SCSBa0012C24	b12	12	C24	65,92	46,77	14,99							127,68
SCSBa0013A24	c1	13	A24	83,6	35,51	11,77							130,88

SCSBa0013C24	c2	13	C24	65,25	50,61	41,94	8,984						166,784
SCSBa0014A24	c3	14	A24	72,02	64,58								136,6
SCSBa0014C24	c4	14	C24	56,54	50,61	22,25	16,23	15,1					160,73
SCSBa0015C24	c6	15	C24	78,19	42,93	31,46	19,73						172,31
SCSBa0016A24	c7	16	A24	67,26	41,45	32,36	23,17						164,24
SCSBa0016C24	c8	16	C24	73,23		16,79							90,02
SCSBa0017A24	c9	17	A24	90,78	31,91								122,69
SCSBa0018A24	c11	18	A24	79,46	55,87								135,33
SCSBa0019A24	d1	19	A24	74,44	37,76	27,02							139,22
SCSBa0019C24	d2	19	C24	55,23	41,94	37,31	32,36	20,19					187,03
SCSBa0020A24	d3	20	A24	95,87	64,58								160,45
SCSBa0020C24	d4	20	C24	137,6									137,6
SCSBa0022A24	d7	22	A24	92,05	45,49	32,36							169,9
SCSBa0023A24	d9	23	A24	53,31	43,57	29,65							126,53
SCSBa0023C24	d10	23	C24	70,81	60,56	33,71							165,08
SCSBa0024C24	d12	24	C24	33,26	27,02	19,28	15,55						95,11
SCSBa0025A24	e1	25	A24	73,37	28,51								101,88
SCSBa0025C24	e2	25	C24	84,98	25,99								110,97
SCSBa0026A24	e3	26	A24	48,91	25,56	11,24	10,01	8,465	6,74				110,925
SCSBa0026C24	e4	26	C24	77,07	35,29	23,46	18,83						154,65
SCSBa0027A24	e5	27	A24	69,67	61,16								130,83
SCSBa0027C24	e6	27	C24	50,35		20,94	13,94	13,4	6,969				105,599
SCSBa0028A24	e7	28	A24	30,19	25,56	20,94	6,912						83,602
SCSBa0029A24	e9	29	A24	43,28	31,56	25,56	12,86						113,26
SCSBa0029C24	e10	29	C24	86,2	32,63		10,47	6,682					135,982
SCSBa0030A24	e11	30	A24	78,9	23,88	14,4	14,32	13,55					145,05
SCSBa0030C24	e12	30	C24	82,55	43,93								126,48

SCSBa0031A24	f1	31	A24	57	31,56		23,04	19,68					131,28
SCSBa0031C24	f2	31	C24	113,4	17,57								130,97
SCSBa0032A24	f3	32	A24	34,75	28,09	23,88	15,89	12,09					114,7
SCSBa0032C24	f4	32	C24	59,17	51,07	42,63	31,56	27,25	13,94				225,62
SCSBa0033A24	f5	33	A24	34,75	24,3	16,14	12,24						87,43
SCSBa0033C24	f6	33	C24	58,5	40,02	30,61	14,02						143,15
SCSBa0034A24	f7	34	A24	51,07	43,28								94,35
SCSBa0036A24	f11	36	A24	66,49	28,93	24,3	14,56						134,28
SCSBa0036C24	f12	36	C24	124,3	117,2								241,5
SCSBa0037A24	g1	37	A24										0
SCSBa0037C24	g2	37	C24	143,9									143,9
SCSBa0038A24	g3	38	A24	65,82		12,78							78,6
SCSBa0038C24	g4	38	C24										0
SCSBa0039A24	g5	39	A24	58,5	50,35	32,1	25,56	15,64	13,78	12,63			208,56
SCSBa0039C24	g6	39	C24	70,28	50,35								120,63
SCSBa0040A24	g7	40	A24		33,69	24,72	19,68	16,65					94,74
SCSBa0040C24	g8	40	C24	61,16	41,33	34,22	28,93	7,849					173,489
SCSBa0041A24	g9	41	A24	31,56	20,94	15,89	7,17						75,56
SCSBa0041C24	g10	41	C24	57,75	37,41	26,41	15,64						137,21
SCSBa0042A24	g11	42	A24	127,2									127,2
SCSBa0042C24	g12	42	C24	66,49	52,52								119,01
SCSBa0043A24	h1	43	A24	86,2	65,82	59,17	28,93	23,46	10,86				274,44
SCSBa0043C24	h2	43	C24	57,75	39,37	29,77	25,56	20,94	14,09	12,86	10,47		210,81
SCSBa0044A24	h3	44	A24	108,1	22,62	13,01	12,55						156,28
SCSBa0044C24	h4	44	C24	78,9	75,22	30,61	20,94	14,48	7,457	5,679			233,286
SCSBa0045A24	h5	45	A24	106	16,9	15,14	7,285						145,325
SCSBa0045C24	h6	45	C24	83,77	25,99	21,36	11,78						142,9

SCSBa0046A24	h7	46	A24	62,49	12,47								74,96
SCSBa0046C24	h8	46	C24	69,05	52,52	36,88	30,19						188,64
SCSBa0047A24	h9	47	A24	33,16	26,83	23,04	12,63	11,7	8,851				116,211
SCSBa0047C24	h10	47	C24	44,59	33,69	12,4	11,32	10,47	7,113	5,965			125,548
SCSBa0048A24	h11	48	A24	71,52	23,46	18,41	10,78	10,16	5,679				140,009
SCSBa0048C24	h12	48	C24	93,64	28,51	22,2	14,88	13,4					172,63
	SS_SBa		**Pl#2**										
SCSBa0049A24	a1	49	A24	84,89	12,53								97,42
SCSBa0049C24	a2	49	C24	76,95	16,28	12,62	11,81	6,204					123,864
SCSBa0050A24	a3	50	A24	72,57	17,23	15,52							105,32
SCSBa0050C24	a4	50	C24	61,49	14,39	11,9	10,92						98,7
SCSBa0051A24	a5	51	A24	34,99	21,85	19	10,83	6,641					93,311
SCSBa0051C24	a6	51	C24	112,7	8,05								120,75
SCSBa0052A24	a7	52	A24	33,73	26,39	17,74	9,408	6,86					94,128
SCSBa0052C24	a8	52	C24	62,9	40,03	12,88							115,81
SCSBa0053A24	a9	53	A24	67,13	23,11	20,58							110,82
SCSBa0053C24	a10	53	C24	69,95	21,85	19,32	6,86	6,568					124,548
SCSBa0054A24	a11	54	A24	51,02	14,01	12,17	10,92	8,53	5,912				102,562
SCSBa0054C24	a12	54	C24	36,25	25,92	20,58	17,42	12,53					112,7
SCSBa0055A24	b1	55	A24	63,6	24,05	16,09							103,74
SCSBa0055C24	b2	55	C24	54,02	44,21	14,2							112,43
SCSBa0056A24	b3	56	A24	41,29	33,21	23,11							97,61
SCSBa0056C24	b4	56	C24	44,21	38,14	27,96							110,31
SCSBa0057A24	b5	57	A24	35,62	29,54	14,01							79,17
SCSBa0057C24	b6	57	C24	66,42	25,45	17,23	15,15	6,69					130,94
SCSBa0058A24	b7	58	A24	54,02	19,95	14,01	8,561						96,541
SCSBa0058C24	b8	58	C24	51,77	24,05	21,53	8,561						105,911

SCSBa0059A24	b9	59	A24	72,57	41,29								113,86
SCSBa0059C24	b10	59	C24	109	16,85								125,85
SCSBa0060A24	b11	60	A24	58,46	45,74								104,2
SCSBa0060C24	b12	60	C24	75,74	17,23	9,853	7,196	0					110,019
SCSBa0061A24	c1	61	A24	61,49	43,45								104,94
SCSBa0061C24	c2	61	C24	55,52	19,63	12,26							87,41
SCSBa0062A24	c3	62	A24	77,55	20,27	17,04	13,82	11,55					140,23
SCSBa0062C24	c4	62	C24	70,65	60,73								131,38
SCSBa0063A24	c5	63	A24	108,3	65,72	29,54	22,48						226,04
SCSBa0063C24	c6	63	C24	56,27	37,51	28,49							122,27
SCSBa0064A24	c7	64	A24	131,9									131,9
SCSBa0064C24	c8	64	C24	40,66	31,64	23,58	18,05	13,06					126,99
SCSBa0065A24	c9	65	A24	141,2	38,77								179,97
SCSBa0065C24	c10	65	C24	78,15	28,49								106,64
SCSBa0066A24	c11	66	A24	118,8	7,196								125,996
SCSBa0066C24	c12	66	C24	73,21	39,4	24,05	12,08						148,74
SCSBa0067A24	d1	67	A24	87,38									87,38
SCSBa0067C24	d2	67	C24	124,4	12,62	11,01							148,03
SCSBa0068A24	d3	68	A24	51,02	42,69	26,39	21,85	11,9					153,85
SCSBa0068C24	d4	68	C24	94,05	26,91								120,96
SCSBa0069A24	d5	69	A24	82,4	46,51								128,91
SCSBa0069C24	d6	69	C24	37,51	15,52								53,03
SCSBa0070A24	d7	70	A24	47,26	16,66	11,55	9,408	7,249					92,127
SCSBa0070C24	d8	70	C24	81,17	44,21	10,03							135,41
SCSBa0071A24	d9	71	A24	79,96	32,69								112,65
SCSBa0071C24	d10	71	C24	69,95	41,29	13,44							124,68
SCSBa0072A24	d11	72	A24	110,5	13,63								124,13

SCSBa0072C24	d12	72	C24	44,21									44,21
SCSBa0073A24	e1	73	A24	109,7									109,7
SCSBa0073C24	e2	73	C24	69,97	31,23								101,2
SCSBa0074A24	e3	74	A24	38,95	31,23	24,97							95,15
SCSBa0074C24	e4	74	C24	39,62	24,97	11,49	8,007	6					90,087
SCSBa0075A24	e5	75	A24	59,63	22,78	14,1							96,51
SCSBa0075C24	e6	75	C24	83,96	18,41	11,05	7,024	5,749					126,193
SCSBa0076A24	e7	76	A24	35,91	9,659	6,53							52,099
SCSBa0076C24	e8	76	C24	115,3	17,97	6,634							139,904
SCSBa0077A24	e9	77	A24	45,03	37,63	24,09	15,07						121,82
SCSBa0077C24	e10	77	C24	43,59	34,25								77,84
SCSBa0078A24	e11	78	A24	50,86	24,97								75,83
SCSBa0078C24	e12	78	C24	66,38	24,53	14,86	7,128						112,898
SCSBa0079A24	f1	79	A24	93,05	18,41								111,46
SCSBa0079C24	f2	79	C24	60,39	51,63	14,01							126,03
SCSBa0080A24	f3	80	A24	119,7									119,7
SCSBa0080C24	f4	80	C24	37,63	31,23	25,41	12,01	10,79	8,442				125,512
SCSBa0081A24	f5	81	A24	43,59	20,16								63,75
SCSBa0081C24	f6	81	C24	35,91	27,59	23,66	14,36	9,572	8,615				119,707
SCSBa0082A24	f7	82	A24	23,22	11,4	10,7	7,128						52,448
SCSBa0082C24	f8	82	C24	47,92	30,22	15,7	13,57	10,18					117,59
SCSBa0083A24	f9	83	A24	44,31	26,72	14,44	12,88	9,659	6,946				114,955
SCSBa0083C24	f10	83	C24	67,57	50,09	12,18							129,84
SCSBa0084A24	f11	84	A24	71,76	14,01	12,88	10,79	7,232	6,66				123,332
SCSBa0084C24	f12	84	C24	95,08	26,28								121,36
SCSBa0085A24	g1	85	A24										0
SCSBa0085C24	g2	85	C24	100,6	13,4								114

SCSBa0086A24	g3	86	A24	48,65	23,22	14,01	12,44	11,4					109,72
SCSBa0086C24	g4	86	C24	38,95	19,28	15,7							73,93
SCSBa0087A24	g5	87	A24	79	46,48	11,66							137,14
SCSBa0087C24	g6	87	C24	52,4	44,31								96,71
SCSBa0088A24	g7	88	A24	54,83	42,26	17,09							114,18
SCSBa0088C24	g8	88	C24	117,3	21,91								139,21
SCSBa0089A24	g9	89	A24	64,97	37,06	24,09	14,18	9,746	7				157,046
SCSBa0089C24	g10	89	C24	51,63	35,91	22,34	16,12	9,92					135,92
SCSBa0090A24	g11	90	A24	114	12,79	7,474							134,264
SCSBa0090C24	g12	90	C24	124,7									124,7
SCSBa0091A24	h1	91	A24	42,26	27,59	15,07							84,92
SCSBa0091C24	h2	91	C24	52,4	22,34								74,74
SCSBa0092A24	h3	92	A24	35,33	29,34	22,34	17,09	14,36	13,31				131,77
SCSBa0092C24	h4	92	C24	38,29	22,78	12,96	11,05						85,08
SCSBa0093A24	h5	93	A24	42,93	15,91								58,84
SCSBa0093C24	h6	93	C24	23,22	5,775								28,995
SCSBa0094A24	h7	94	A24	58,1	21,03	15,7	11,49	10,44					116,76
SCSBa0094C24	h8	94	C24	68,77	39,62	12,79							121,18
SCSBa0095A24	h9	95	A24	69,97	46,48								116,45
SCSBa0095C24	h10	95	C24	68,17	44,31	13,31							125,79
SCSBa0096A24	h11	96	A24	61,92	26,72	23,22	11,31	10,62					133,79
SCSBa0096C24	h12	96	C24	144,9	22,78								167,68
													21717,887

Supplemental Table 3 - PacBio Sequencing of Sugarcane BAC pool

Number of Raw Reads	Number of Filtered Reads *	Number of Bases (Filtered Reads)	Average read length	Coverage
150.292	101.841	370.446.255	3.637	17

* Minimum Read Length = 50bp

Number of reads	Length	Percentage
14.925	1.000	14,655%
25.288	2.000	24,831%
16.675	3.000	16,374%
11.132	4.000	10,931%
8.067	5.000	7,921%
6.314	6.000	6,200%
4.834	7.000	4,747%
4.033	8.000	3,960%
3.059	9.000	3,004%
2.238	1.000	2,198%
1.642	11.000	1,612%
1.126	12.000	1,106%
872	13.000	0,856%
600	14.000	0,589%
421	15.000	0,413%
262	16.000	0,257%
170	17.000	0,167%
84	18.000	0,082%
55	19.000	0,054%
27	20.000	0,027%
7	21.000	0,007%
5	22.000	0,005%
5	23.000	0,005%
101.841		

Supplemental Table 4 - BAC pool assembly numbers and metrics

Input	Reads Illumina	Reads Illumina + Reads PacBio	Contigs PacBio + Contigs Edena	Contigs Edena + Reads PacBio	Contigs Edena + Reads PacBio
Software	Edena	PacBioToCA + Celera Assembler	Celera Assembler	SSPACE-LongRead	AHA
Output	Contigs	Contigs	Contigs	Scaffolds	Scaffolds
Number of contigs/scaffolds				2.697	2.451
Number of bases in contigs/scaffolds (bp)				19.085.165	19.155.751
N50 (bp)				56.229	54.129
Largest contig/scaffold size (bp)				195.258	203.132
BES matching end of scaffolds		240	223	255	272
BES matching middle of scaffolds		31	50	51	33
BES matching more than one scaffold		35	46	19	20
BES not matching scaffolds		45	32	26	26
Number of “One Contig” scaffolds		0	2	7	8

Supplemental Table 5 - BACs assembled in "One Contig" – BES match consistently at end of scaffolds and scaffold length is similar to expected BAC length

BAC	Expected BAC length	Scaffold length	End Distance BES (.b)	End Distance BES (.g)	Alignment Coverage Percentage BES (.b)	Alignment Coverage Percentage BES (.g)	Number of Scaffolds matching BES (.b)	Number of Scaffolds matching BES (.g)	Scaffold ID
SCSBa0008C24	141,860	138,830	91	96	100.00	85.58	2	1	scaffold5/869
SCSBa0042A24	127,200	129,783	103	1	98.89	93.96	1	1	scaffold5/2173
SCSBa0045C24	142,900	126,859	93	95	100.00	100.00	1	1	scaffold5/770
SCSBa0055C24	112,430	141,339	119	95	97.28	100.00	1	1	scaffold5/2050
SCSBa0067A24	87,380	64,808	97	95	93.56	100.00	1	1	scaffold5/1460
SCSBa0084A24	123,332	119,761	93	97	99.46	100.00	1	1	scaffold5/1840
SCSBa0085C24	114,000	123,118	97	96	99.44	100.00	1	1	scaffold5/2075
SCSBa0093C24	28,995	26,337	96	97	100.00	99.88	1	1	scaffold5/1329

Supplemental Table 6 - BACs assembled in two scaffolds - BES match consistently at end of scaffolds and sum of scaffolds length is similar to expected BAC length

BAC	Expected BAC length	Scaffold1 length	Scaffold2 length	Sum of Scaffolds bases	End Distance BES (.b)	End Distance BES (.g)	Aligment Percentage BES (.b)	Aligment Percentage BES (.g)	Number of Scaffolds matching BES (.b)	Number of Scaffolds matching BES (.g)	Scaffold1 ID	Scaffold2 ID
SCSBa0001C24	113,270	34,483	85,032	119,515	98	95	87.48	100.00	1	1	scaffold5/1164	scaffold5/720
SCSBa0003A24	117,770	57,550	24,333	81,883	89	92	78.77	80.95	1	1	scaffold5/1694	scaffold5/659
SCSBa0005A24	104,870	18,716	63,273	81,989	104	95	99.04	100.00	1	1	scaffold5/190	scaffold5/688
SCSBa0006C24	147,770	86,420	22,698	109,118	93	78	90.54	97.41	1	1	scaffold5/1427	scaffold5/494
SCSBa0007C24	67,770	3,779	60,625	64,404	98	97	93.42	84.35	1	1	scaffold5/792	scaffold5/388
SCSBa0008A24	119,780	98,065	1,774	99,839	97	95	81.78	100.00	1	1	scaffold5/809	scaffold5/2029
SCSBa0012A24	141,460	122,587	3,705	126,292	98	97	100.00	100.00	1	1	scaffold5/264	scaffold5/760
SCSBa0013A24	130,880	54,738	93,774	148,512	95	96	99.88	99.48	1	1	scaffold5/2344	scaffold5/1483
SCSBa0014A24	136,600	108,756	8,515	117,271	97	86	100.31	85.39	1	1	scaffold5/257	scaffold5/1643
SCSBa0018A24	135,330	46,546	76,087	122,633	96	93	84.92	72.31	1	1	scaffold5/726	scaffold5/2369
SCSBa0018C24	148,850	69,356	51,720	121,076	95	97	51.40	96.32	1	1	scaffold5/169	scaffold5/1409
SCSBa0026C24	154,650	152,524	27,694	180,218	101	97	100.00	100.00	1	1	scaffold5/1048	scaffold5/1411
SCSBa0027C24	105,599	44,181	80,959	125,140	96	96	98.93	99.00	1	1	scaffold5/1839	scaffold5/1901
SCSBa0029C24	135,982	127,769	14,615	142,384	92	98	100.00	99.89	1	1	scaffold5/1608	scaffold5/1882
SCSBa0030C24	126,480	29,436	81,952	111,388	95	95	60.38	100.00	1	1	scaffold5/1376	scaffold5/2361
SCSBa0033C24	143,150	79,234	54,926	134,160	96	57	100.00	52.58	1	1	scaffold5/407	scaffold5/56
SCSBa0046A24	74,960	82,024	12,706	94,730	97	77	100.00	94.48	1	1	scaffold5/467	scaffold5/1968
SCSBa0054A24	102,562	77,911	5,732	83,643	100	96	54.43	99.49	1	1	scaffold5/321	scaffold5/1538
SCSBa0056A24	97,610	81,486	39,234	120,720	96	94	99.89	100.00	1	1	scaffold5/1098	scaffold5/301
SCSBa0056C24	110,310	43,442	34,662	78,104	96	77	89.66	77.09	1	1	scaffold5/1226	scaffold5/405
SCSBa0060A24	104,200	68,438	6,379	74,817	97	96	99.08	100.12	1	1	scaffold5/1582	scaffold5/1679

SCSBa0068A24	153,850	116,291	9,946	126,237	98	1	100.00	52.43	1	1	scaffold5/1010	scaffold5/1731
SCSBa0070C24	135,410	86,830	40,175	127,005	93	97	87.53	90.22	1	1	scaffold5/799	scaffold5/1423
SCSBa0071C24	124,680	151,964	7,579	159,543	93	94	100.00	88.21	1	1	scaffold5/377	scaffold5/2072
SCSBa0073C24	101,200	81,578	871	82,449	97	1	80.62	92.16	1	1	scaffold5/2305	scaffold5/708
SCSBa0075C24	126,193	48,563	47,432	95,995	101	0	99.10	88.96	1	1	scaffold5/1096	scaffold5/629
SCSBa0077A24	121,820	61,721	30,502	92,223	100	98	98.90	99.88	1	1	scaffold5/1931	scaffold5/1689
SCSBa0078A24	75,830	51,632	3,505	55,137	95	1	81.27	84.54	1	1	scaffold5/552	scaffold5/1204
SCSBa0080C24	125,512	62,075	74,360	136,435	96	92	99.65	100.00	1	1	scaffold5/1333	scaffold5/2232
SCSBa0088A24	114,180	63,407	33,110	96,517	91	96	90.29	100.00	1	1	scaffold5/810	scaffold5/679
SCSBa0092C24	85,080	17,620	87,231	104,851	94	93	100.00	95.44	1	1	scaffold5/1675	scaffold5/1410
SCSBa0095A24	116,450	58,321	58,703	117,024	94	96	100.00	100.00	1	1	scaffold5/1870	scaffold5/1093

Supplemental Table 7 - Collinearity between Sugarcane genes and Sorghum genes.
133 Sugarcane scaffolds present gene collinearity with Sorghum.
Different gray colors are used to group genes by scaffold.

Sugarcane Genes	Sorghum				
	Gene	Chr	start	end	strand
evm.model.scaffold_1037.1	Sb05g022900.1|PACid:1970764	5	55.256.680	55.258.337	+
evm.model.scaffold_1037.3	Sb05g022910.1|PACid:1970765	5	55.258.506	55.259.773	-
evm.model.scaffold_1048.3	Sb02g039140.1|PACid:1959381	2	73.279.014	73.281.452	-
evm.model.scaffold_1048.4	Sb02g039130.1|PACid:1959379	2	73.255.405	73.277.570	+
evm.model.scaffold_1048.5	Sb02g039130.1|PACid:1959379	2	73.255.405	73.277.570	+
evm.model.scaffold_1048.7	Sb10g021650.1|PACid:1984000	10	47.786.528	47.787.967	-
evm.model.scaffold_1071.4	Sb10g023950.1|PACid:1984309	10	52.722.451	52.723.722	+
evm.model.scaffold_1071.6	Sb10g023940.1|PACid:1984308	10	52.703.684	52.705.035	-
evm.model.scaffold_1071.8	Sb10g023930.2|PACid:1984307	10	52.700.092	52.701.651	+
evm.model.scaffold_1071.9	Sb10g023920.1|PACid:1984305	10	52.675.382	52.677.175	+
evm.model.scaffold_1071.10	Sb10g023920.1|PACid:1984305	10	52.675.382	52.677.175	+
evm.model.scaffold_1071.12	Sb10g023910.1|PACid:1984304	10	52.664.722	52.669.195	+
evm.model.scaffold_1071.13	Sb10g023910.1|PACid:1984304	10	52.664.722	52.669.195	+
evm.model.scaffold_1093.1	Sb09g005270.1|PACid:1979960	9	6.771.485	6.775.784	+
evm.model.scaffold_1093.4	Sb09g005260.1|PACid:1979959	9	6.732.398	6.733.860	-
evm.model.scaffold_1093.5	Sb09g005250.1|PACid:1979958	9	6.728.382	6.732.228	+
evm.model.scaffold_1094.4	Sb04g001400.1|PACid:1965020	4	1.189.374	1.192.740	+
evm.model.scaffold_1094.5	Sb04g003200.1|PACid:1965255	4	2.997.394	3.005.273	-
evm.model.scaffold_11.1	Sb03g034460.1|PACid:1963275	3	62.662.537	62.663.700	-
evm.model.scaffold_11.2	Sb03g034470.1|PACid:1963276	3	62.663.960	62.669.234	+
evm.model.scaffold_1103.1	Sb02g003915.2|PACid:1955728	2	4.360.603	4.363.528	-
evm.model.scaffold_1103.2	Sb02g003915.2|PACid:1955728	2	4.360.603	4.363.528	-
evm.model.scaffold_1115.2	Sb05g019130.1|PACid:1970336	5	46.662.285	46.665.108	+
evm.model.scaffold_1115.4	Sb05g019510.1|PACid:1970365	5	47.721.853	47.726.473	-
evm.model.scaffold_1115.6	Sb05g019520.1|PACid:1970366	5	47.748.224	47.756.510	-
evm.model.scaffold_1128.1	Sb03g013460.1|PACid:1961655	3	16.813.617	16.818.046	+
evm.model.scaffold_1128.2	Sb03g013470.1|PACid:1961656	3	16.818.148	16.824.079	-
evm.model.scaffold_1138.1	Sb02g018600.1|PACid:1956992	2	45.389.213	45.391.901	+
evm.model.scaffold_1138.2	Sb02g018590.1|PACid:1956990	2	45.381.698	45.383.806	+
evm.model.scaffold_1138.3	Sb02g018580.1|PACid:1956989	2	45.296.523	45.298.834	+
evm.model.scaffold_1148.2	Sb02g008065.1|PACid:1956315	2	10.435.272	10.440.088	+
evm.model.scaffold_1148.3	Sb02g008065.1|PACid:1956315	2	10.435.272	10.440.088	+
evm.model.scaffold_1172.1	Sb02g001040.1|PACid:1955355	2	847.022	849.208	-
evm.model.scaffold_1172.3	Sb02g001390.1|PACid:1955400	2	1.200.626	1.202.092	+
evm.model.scaffold_1172.5	Sb02g001390.1|PACid:1955400	2	1.200.626	1.202.092	+
evm.model.scaffold_1172.7	Sb02g001470.1|PACid:1955414	2	1.289.500	1.291.511	+

evm.model.scaffold_1172.8	Sb02g001380.1|PACid:1955399	2	1.195.779	1.197.298	+
evm.model.scaffold_1174.2	Sb05g006380.1|PACid:1969627	5	10.116.867	10.118.570	+
evm.model.scaffold_1174.3	Sb05g006420.1|PACid:1969637	5	10.349.434	10.351.941	+
evm.model.scaffold_1175.1	Sb09g007360.1|PACid:1980209	9	12.846.922	12.847.461	-
evm.model.scaffold_1175.5	Sb09g007330.1|PACid:1980204	9	12.628.569	12.633.952	-
evm.model.scaffold_1175.6	Sb09g007310.1|PACid:1980201	9	12.271.454	12.276.979	-
evm.model.scaffold_1175.7	Sb09g007300.1|PACid:1980200	9	12.267.354	12.270.694	-
evm.model.scaffold_1186.1	Sb07g004760.1|PACid:1975170	7	6.297.720	6.299.461	+
evm.model.scaffold_1186.3	Sb07g004750.1|PACid:1975169	7	6.290.776	6.296.000	+
evm.model.scaffold_1186.4	Sb07g004740.1|PACid:1975168	7	6.276.609	6.277.436	-
evm.model.scaffold_1186.5	Sb07g004730.1|PACid:1975167	7	6.270.751	6.274.903	+
evm.model.scaffold_1226.1	Sb04g006360.1|PACid:1965654	4	6.407.138	6.407.287	+
evm.model.scaffold_1226.2	Sb04g006370.1|PACid:1965655	4	6.412.424	6.415.352	+
evm.model.scaffold_1226.3	Sb04g006375.1|PACid:1965656	4	6.416.362	6.416.694	+
evm.model.scaffold_1226.4	Sb04g006380.1|PACid:1965657	4	6.418.055	6.426.683	-
evm.model.scaffold_1229.1	Sb02g027500.1|PACid:1957946	2	62.768.206	62.769.481	+
evm.model.scaffold_1229.2	Sb02g027490.1|PACid:1957945	2	62.764.446	62.766.566	+
evm.model.scaffold_1229.3	Sb02g027480.1|PACid:1957944	2	62.760.165	62.761.194	+
evm.model.scaffold_1229.5	Sb02g027475.1|PACid:1957943	2	62.752.418	62.757.575	-
evm.model.scaffold_1229.6	Sb02g027470.1|PACid:1957942	2	62.745.198	62.748.327	-
evm.model.scaffold_1229.7	Sb02g027460.1|PACid:1957941	2	62.734.582	62.741.600	+
evm.model.scaffold_1273.2	Sb04g009590.2|PACid:1966049	4	12.101.128	12.105.407	-
evm.model.scaffold_1273.4	Sb04g009600.1|PACid:1966050	4	12.109.948	12.110.367	+
evm.model.scaffold_1273.5	Sb04g009610.1|PACid:1966051	4	12.112.034	12.114.080	-
evm.model.scaffold_1273.7	Sb04g009620.2|PACid:1966053	4	12.119.970	12.124.658	-
evm.model.scaffold_1274.1	Sb01g043360.1|PACid:1954343	1	66.555.852	66.556.315	-
evm.model.scaffold_1274.3	Sb01g043350.1|PACid:1954342	1	66.549.192	66.552.200	-
evm.model.scaffold_1274.6	Sb01g043390.1|PACid:1954346	1	66.570.966	66.572.572	-
evm.model.scaffold_1274.7	Sb01g043390.1|PACid:1954346	1	66.570.966	66.572.572	-
evm.model.scaffold_1274.8	Sb01g043380.1|PACid:1954345	1	66.565.728	66.568.055	-
evm.model.scaffold_1274.9	Sb01g043370.1|PACid:1954344	1	66.560.139	66.565.073	+
evm.model.scaffold_1275.1	Sb06g002250.1|PACid:1971715	6	4.249.152	4.256.224	+
evm.model.scaffold_1275.3	Sb06g002240.1|PACid:1971714	6	4.233.147	4.240.507	+
evm.model.scaffold_1275.5	Sb06g002220.1|PACid:1971711	6	4.178.738	4.181.898	+
evm.model.scaffold_1276.1	Sb02g003870.1|PACid:1955721	2	4.288.179	4.295.402	-
evm.model.scaffold_1276.3	Sb02g003860.1|PACid:1955720	2	4.283.726	4.287.550	+
evm.model.scaffold_1276.4	Sb02g003880.1|PACid:1955722	2	4.296.212	4.306.762	-
evm.model.scaffold_1329.1	Sb01g029130.1|PACid:1952582	1	50.905.862	50.907.648	+
evm.model.scaffold_1329.2	Sb01g029120.1|PACid:1952581	1	50.880.180	50.890.807	+
evm.model.scaffold_1341.1	Sb06g024380.1|PACid:1973371	6	53.508.596	53.510.077	+
evm.model.scaffold_1341.3	Sb06g024390.1|PACid:1973372	6	53.521.619	53.522.489	+
evm.model.scaffold_1341.5	Sb06g024400.1|PACid:1973373	6	53.535.970	53.538.827	+
evm.model.scaffold_1405.2	Sb09g020320.1|PACid:1980868	9	49.589.167	49.592.182	-
evm.model.scaffold_1405.3	Sb09g020310.1|PACid:1980867	9	49.572.300	49.588.443	-
evm.model.scaffold_1405.5	Sb09g020310.1|PACid:1980867	9	49.572.300	49.588.443	-

evm.model.scaffold_1407.1	Sb04g001900.1|PACid:1965085	4	1.628.407	1.630.029	+
evm.model.scaffold_1407.2	Sb04g001913.1|PACid:1965087	4	1.635.912	1.639.088	-
evm.model.scaffold_1407.3	Sb04g001916.1|PACid:1965088	4	1.641.482	1.646.286	-
evm.model.scaffold_1407.4	Sb04g001916.1|PACid:1965088	4	1.641.482	1.646.286	-
evm.model.scaffold_1408.1	Sb05g020240.1|PACid:1970445	5	49.632.483	49.634.711	-
evm.model.scaffold_1408.3	Sb05g020250.1|PACid:1970446	5	49.637.709	49.638.670	-
evm.model.scaffold_1408.4	Sb05g020240.1|PACid:1970445	5	49.632.483	49.634.711	-
evm.model.scaffold_1409.1	Sb09g030110.1|PACid:1982045	9	58.756.716	58.759.154	+
evm.model.scaffold_1409.5	Sb09g030080.1|PACid:1982042	9	58.732.078	58.734.419	+
evm.model.scaffold_1409.7	Sb09g030125.1|PACid:1982047	9	58.762.375	58.763.460	+
evm.model.scaffold_1410.1	Sb01g012660.1|PACid:1950858	1	11.645.237	11.646.322	-
evm.model.scaffold_1410.2	Sb01g012650.1|PACid:1950857	1	11.635.128	11.640.213	-
evm.model.scaffold_1410.3	Sb01g012640.1|PACid:1950856	1	11.631.883	11.634.199	+
evm.model.scaffold_1410.4	Sb01g012630.1|PACid:1950855	1	11.627.412	11.628.237	-
evm.model.scaffold_1410.5	Sb01g012620.1|PACid:1950854	1	11.623.001	11.623.904	-
evm.model.scaffold_1410.6	Sb01g012610.1|PACid:1950853	1	11.618.856	11.622.800	+
evm.model.scaffold_1415.1	Sb09g030820.1|PACid:1982134	9	59.434.853	59.435.662	-
evm.model.scaffold_1415.2	Sb09g030810.1|PACid:1982132	9	59.431.871	59.434.202	+
evm.model.scaffold_1427.1	Sb09g030510.1|PACid:1982090	9	59.082.694	59.086.551	+
evm.model.scaffold_1427.2	Sb09g030520.1|PACid:1982091	9	59.097.658	59.100.427	+
evm.model.scaffold_1427.3	Sb09g030530.5|PACid:1982096	9	59.100.907	59.105.416	-
evm.model.scaffold_1427.4	Sb09g030540.1|PACid:1982097	9	59.109.618	59.110.761	+
evm.model.scaffold_145.1	Sb03g010150.1|PACid:1961249	3	10.986.309	10.987.758	-
evm.model.scaffold_145.2	Sb03g010150.1|PACid:1961249	3	10.986.309	10.987.758	-
evm.model.scaffold_145.5	Sb03g010140.1|PACid:1961248	3	10.980.530	10.984.962	+
evm.model.scaffold_145.6	Sb03g010130.1|PACid:1961246	3	10.942.412	10.948.914	-
evm.model.scaffold_145.7	Sb03g010120.1|PACid:1961245	3	10.935.750	10.939.816	+
evm.model.scaffold_1483.1	Sb01g049140.1|PACid:1955051	1	72.125.370	72.129.092	+
evm.model.scaffold_1483.4	Sb01g049150.1|PACid:1955052	1	72.144.875	72.146.368	-
evm.model.scaffold_1483.6	Sb01g049160.1|PACid:1955053	1	72.170.465	72.172.197	+
evm.model.scaffold_1483.7	Sb01g049180.1|PACid:1955056	1	72.192.758	72.194.092	+
evm.model.scaffold_1492.1	Sb08g001240.1|PACid:1977245	8	1.213.938	1.216.326	+
evm.model.scaffold_1492.2	Sb08g001240.1|PACid:1977245	8	1.213.938	1.216.326	+
evm.model.scaffold_1575.1	Sb03g044485.1|PACid:1964461	3	71.843.543	71.847.704	+
evm.model.scaffold_1575.2	Sb03g044480.1|PACid:1964460	3	71.834.871	71.837.933	-
evm.model.scaffold_1575.3	Sb03g044470.1|PACid:1964459	3	71.807.529	71.811.781	-
evm.model.scaffold_1575.4	Sb03g044460.1|PACid:1964458	3	71.806.009	71.806.845	+
evm.model.scaffold_1575.5	Sb03g044450.1|PACid:1964457	3	71.799.425	71.805.341	-
evm.model.scaffold_1575.7	Sb03g044440.1|PACid:1964456	3	71.785.070	71.786.496	+
evm.model.scaffold_1575.8	Sb03g044430.1|PACid:1964455	3	71.779.051	71.781.483	+
evm.model.scaffold_1575.9	Sb03g044420.1|PACid:1964454	3	71.771.395	71.774.897	-
evm.model.scaffold_1575.10	Sb03g044410.1|PACid:1964453	3	71.769.461	71.771.085	+
evm.model.scaffold_158.1	Sb06g005410.1|PACid:1971886	6	13.580.446	13.581.583	+
evm.model.scaffold_158.2	Sb06g005410.1|PACid:1971886	6	13.580.446	13.581.583	+
evm.model.scaffold_1581.1	Sb01g028370.1|PACid:1952483	1	49.479.727	49.483.109	-

evm.model.scaffold_1581.2	Sb01g028360.4|PACid:1952482	1	49.475.358	49.478.716	-
evm.model.scaffold_1581.3	Sb01g028350.1|PACid:1952478	1	49.473.863	49.474.708	+
evm.model.scaffold_1581.4	Sb01g028340.1|PACid:1952477	1	49.471.186	49.473.008	+
evm.model.scaffold_1609.1	Sb01g044070.1|PACid:1954422	1	67.226.697	67.231.065	-
evm.model.scaffold_1609.4	Sb01g044050.1|PACid:1954420	1	67.203.376	67.204.311	+
evm.model.scaffold_1620.1	Sb08g005125.1|PACid:1977749	8	6.557.130	6.559.784	+
evm.model.scaffold_1620.2	Sb08g005110.1|PACid:1977746	8	6.531.235	6.532.218	+
evm.model.scaffold_1689.1	Sb02g022780.1|PACid:1957336	2	56.039.429	56.044.860	+
evm.model.scaffold_1689.2	Sb02g022780.1|PACid:1957336	2	56.039.429	56.044.860	+
evm.model.scaffold_1689.3	Sb02g022770.1|PACid:1957335	2	56.020.096	56.024.150	-
evm.model.scaffold_169.1	Sb09g030110.1|PACid:1982045	9	58.756.716	58.759.154	+
evm.model.scaffold_169.3	Sb09g027550.1|PACid:1981741	9	56.646.159	56.648.853	+
evm.model.scaffold_169.4	Sb09g027560.1|PACid:1981742	9	56.649.198	56.651.080	-
evm.model.scaffold_169.6	Sb09g027720.1|PACid:1981760	9	56.779.590	56.782.217	+
evm.model.scaffold_1715.1	Sb03g043960.1|PACid:1964402	3	71.204.119	71.206.483	-
evm.model.scaffold_1715.2	Sb03g043970.1|PACid:1964403	3	71.210.660	71.215.483	-
evm.model.scaffold_1715.3	Sb03g043980.1|PACid:1964404	3	71.219.341	71.224.191	-
evm.model.scaffold_172.2	Sb01g009700.1|PACid:1950501	1	8.482.372	8.483.178	+
evm.model.scaffold_172.3	Sb01g009710.1|PACid:1950502	1	8.483.890	8.488.828	-
evm.model.scaffold_1738.1	Sb09g030800.1|PACid:1982131	9	59.425.403	59.431.498	+
evm.model.scaffold_1738.2	Sb09g030790.1|PACid:1982130	9	59.420.443	59.425.222	-
evm.model.scaffold_1741.2	Sb05g001700.1|PACid:1968998	5	1.841.352	1.845.079	+
evm.model.scaffold_1741.3	Sb05g001700.1|PACid:1968998	5	1.841.352	1.845.079	+
evm.model.scaffold_1741.4	Sb05g001700.1|PACid:1968998	5	1.841.352	1.845.079	+
evm.model.scaffold_1741.6	Sb05g001710.1|PACid:1968999	5	1.858.647	1.860.913	+
evm.model.scaffold_1744.2	Sb10g002260.1|PACid:1982410	10	1.956.003	1.958.393	-
evm.model.scaffold_1744.3	Sb10g002250.1|PACid:1982409	10	1.950.327	1.955.286	-
evm.model.scaffold_1744.5	Sb10g002240.1|PACid:1982408	10	1.947.290	1.949.502	-
evm.model.scaffold_1744.6	Sb10g002230.1|PACid:1982407	10	1.942.572	1.946.659	+
evm.model.scaffold_1840.1	Sb03g013540.2|PACid:1961665	3	17.006.847	17.017.981	+
evm.model.scaffold_1840.2	Sb03g013540.2|PACid:1961665	3	17.006.847	17.017.981	+
evm.model.scaffold_1846.1	Sb09g006630.1|PACid:1980153	9	10.750.243	10.755.595	-
evm.model.scaffold_1846.2	Sb09g006640.1|PACid:1980154	9	10.759.109	10.761.601	+
evm.model.scaffold_1846.3	Sb09g006650.1|PACid:1980155	9	10.761.911	10.769.582	-
evm.model.scaffold_1846.4	Sb09g006670.1|PACid:1980158	9	10.779.285	10.781.813	-
evm.model.scaffold_1868.2	Sb02g025966.1|PACid:1957745	2	61.017.466	61.019.287	-
evm.model.scaffold_1868.8	Sb02g025970.2|PACid:1957747	2	61.020.478	61.023.193	-
evm.model.scaffold_1868.9	Sb02g025970.1|PACid:1957746	2	61.020.478	61.023.193	-
evm.model.scaffold_1868.10	Sb02g025980.1|PACid:1957748	2	61.026.436	61.028.668	+
evm.model.scaffold_1868.11	Sb02g025980.1|PACid:1957748	2	61.026.436	61.028.668	+
evm.model.scaffold_1895.1	Sb08g020340.1|PACid:1978911	8	51.392.301	51.395.362	+
evm.model.scaffold_1895.3	Sb08g020320.1|PACid:1978909	8	51.369.057	51.372.606	+
evm.model.scaffold_1901.2	Sb05g019530.1|PACid:1970369	5	47.867.151	47.870.659	+
evm.model.scaffold_1901.3	Sb05g019540.2|PACid:1970371	5	47.874.380	47.882.162	-
evm.model.scaffold_1909.1	Sb04g006390.1|PACid:1965658	4	6.427.801	6.431.435	+

evm.model.scaffold_1909.2	Sb04g006400.1|PACid:1965659	4	6.431.878	6.434.573	-
evm.model.scaffold_1909.4	Sb04g006410.1|PACid:1965660	4	6.437.936	6.442.557	+
evm.model.scaffold_1909.5	Sb04g006420.1|PACid:1965661	4	6.442.793	6.447.458	-
evm.model.scaffold_1909.6	Sb04g006430.1|PACid:1965662	4	6.447.652	6.450.831	+
evm.model.scaffold_1909.7	Sb04g006440.1|PACid:1965663	4	6.451.450	6.456.805	-
evm.model.scaffold_1909.8	Sb04g006450.1|PACid:1965664	4	6.459.118	6.461.711	+
evm.model.scaffold_1909.9	Sb04g006460.1|PACid:1965665	4	6.463.268	6.466.720	-
evm.model.scaffold_1921.4	Sb04g025800.1|PACid:1967223	4	55.567.645	55.571.544	-
evm.model.scaffold_1921.6	Sb04g025780.1|PACid:1967221	4	55.555.555	55.558.749	+
evm.model.scaffold_1941.3	Sb03g044000.1|PACid:1964407	3	71.239.079	71.245.071	+
evm.model.scaffold_1941.4	Sb03g043995.1|PACid:1964406	3	71.236.601	71.238.540	-
evm.model.scaffold_1941.5	Sb03g043990.1|PACid:1964405	3	71.233.079	71.235.839	+
evm.model.scaffold_1942.1	Sb03g031670.1|PACid:1962925	3	60.080.358	60.084.458	+
evm.model.scaffold_1942.2	Sb03g031670.1|PACid:1962925	3	60.080.358	60.084.458	+
evm.model.scaffold_1944.4	Sb03g047220.1|PACid:1964793	3	74.145.011	74.148.340	-
evm.model.scaffold_1944.5	Sb03g047210.1|PACid:1964792	3	74.141.132	74.143.901	+
evm.model.scaffold_1956.1	Sb01g004090.1|PACid:1949782	1	3.310.786	3.312.790	+
evm.model.scaffold_1956.2	Sb01g004100.1|PACid:1949783	1	3.316.366	3.318.406	+
evm.model.scaffold_1956.4	Sb01g004110.1|PACid:1949784	1	3.319.006	3.320.520	+
evm.model.scaffold_1956.6	Sb01g004120.1|PACid:1949785	1	3.326.638	3.327.186	+
evm.model.scaffold_1956.7	Sb01g004130.1|PACid:1949786	1	3.327.321	3.332.332	-
evm.model.scaffold_1956.10	Sb01g004150.1|PACid:1949788	1	3.344.432	3.348.438	-
evm.model.scaffold_1969.1	Sb06g024355.1|PACid:1973368	6	53.448.316	53.449.768	-
evm.model.scaffold_1969.2	Sb06g024370.1|PACid:1973370	6	53.491.945	53.492.949	+
evm.model.scaffold_2050.5	Sb04g008300.1|PACid:1965904	4	9.420.645	9.421.922	+
evm.model.scaffold_2050.7	Sb04g008280.1|PACid:1965902	4	9.386.737	9.388.263	-
evm.model.scaffold_2050.9	Sb04g008270.1|PACid:1965901	4	9.381.944	9.383.000	+
evm.model.scaffold_2050.10	Sb04g008310.1|PACid:1965905	4	9.423.788	9.429.225	-
evm.model.scaffold_2093.3	Sb03g044630.1|PACid:1964481	3	71.975.563	71.980.937	+
evm.model.scaffold_2093.4	Sb03g044640.1|PACid:1964482	3	71.985.660	71.992.831	-
evm.model.scaffold_2093.5	Sb03g044650.1|PACid:1964483	3	71.993.909	71.998.666	-
evm.model.scaffold_2093.6	Sb03g044660.1|PACid:1964484	3	72.005.974	72.008.320	-
evm.model.scaffold_2096.1	Sb03g010160.1|PACid:1961251	3	11.002.937	11.007.215	+
evm.model.scaffold_2096.2	Sb03g010170.1|PACid:1961252	3	11.009.212	11.010.604	+
evm.model.scaffold_2133.1	Sb05g020240.1|PACid:1970445	5	49.632.483	49.634.711	-
evm.model.scaffold_2133.2	Sb05g020240.1|PACid:1970445	5	49.632.483	49.634.711	-
evm.model.scaffold_2135.1	Sb04g001756.1|PACid:1965064	4	1.494.155	1.498.653	+
evm.model.scaffold_2135.2	Sb04g001753.1|PACid:1965063	4	1.481.100	1.485.994	-
evm.model.scaffold_2194.2	Sb03g010190.1|PACid:1961254	3	11.034.168	11.035.612	+
evm.model.scaffold_2194.4	Sb03g010180.1|PACid:1961253	3	11.025.544	11.027.075	+
evm.model.scaffold_2201.1	Sb04g001210.1|PACid:1964997	4	1.026.974	1.030.045	+
evm.model.scaffold_2201.3	Sb04g001220.1|PACid:1964998	4	1.048.656	1.051.236	-
evm.model.scaffold_2201.4	Sb04g001620.1|PACid:1965045	4	1.372.821	1.374.648	+
evm.model.scaffold_2209.1	Sb07g002080.1|PACid:1974815	7	2.223.319	2.227.173	+
evm.model.scaffold_2209.2	Sb07g002080.1|PACid:1974815	7	2.223.319	2.227.173	+

evm.model.scaffold_2209.4	Sb07g002070.2|PACid:1974814	7	2.208.945	2.212.826	-
evm.model.scaffold_2209.7	Sb07g002050.1|PACid:1974811	7	2.197.663	2.202.438	-
evm.model.scaffold_2209.9	Sb07g002040.1|PACid:1974810	7	2.189.593	2.190.568	-
evm.model.scaffold_2209.11	Sb07g002030.1|PACid:1974809	7	2.181.443	2.182.918	-
evm.model.scaffold_2209.13	Sb07g002030.1|PACid:1974809	7	2.181.443	2.182.918	-
evm.model.scaffold_2209.15	Sb07g002020.2|PACid:1974807	7	2.176.461	2.177.735	+
evm.model.scaffold_2209.16	Sb07g002010.1|PACid:1974805	7	2.171.218	2.173.923	+
evm.model.scaffold_2234.1	Sb03g017600.1|PACid:1961837	3	28.877.458	28.878.912	-
evm.model.scaffold_2234.2	Sb03g039445.1|PACid:1963858	3	67.130.423	67.130.569	+
evm.model.scaffold_2234.4	Sb03g017600.1|PACid:1961837	3	28.877.458	28.878.912	-
evm.model.scaffold_2234.8	Sb03g020182.1|PACid:1961890	3	38.389.958	38.390.431	+
evm.model.scaffold_2236.3	Sb10g008220.1|PACid:1983166	10	8.330.795	8.331.640	+
evm.model.scaffold_2236.4	Sb10g008230.1|PACid:1983167	10	8.349.157	8.351.340	+
evm.model.scaffold_2236.5	Sb10g008240.1|PACid:1983169	10	8.351.999	8.354.688	-
evm.model.scaffold_2236.6	Sb10g008250.1|PACid:1983170	10	8.357.631	8.361.511	-
evm.model.scaffold_2303.2	Sb02g022230.1|PACid:1957263	2	54.789.662	54.791.283	-
evm.model.scaffold_2303.4	Sb02g022225.1|PACid:1957262	2	54.699.239	54.701.232	-
evm.model.scaffold_2303.6	Sb02g022220.1|PACid:1957261	2	54.694.244	54.696.319	+
evm.model.scaffold_2303.7	Sb02g022210.1|PACid:1957260	2	54.689.030	54.689.704	+
evm.model.scaffold_2305.1	Sb02g003900.1|PACid:1955725	2	4.328.360	4.338.839	-
evm.model.scaffold_2305.2	Sb02g003890.1|PACid:1955724	2	4.321.584	4.328.217	+
evm.model.scaffold_2305.6	Sb02g003880.2|PACid:1955723	2	4.296.212	4.306.762	-
evm.model.scaffold_2394.1	Sb06g024380.1|PACid:1973371	6	53.508.596	53.510.077	+
evm.model.scaffold_2394.5	Sb06g024370.1|PACid:1973370	6	53.491.945	53.492.949	+
evm.model.scaffold_2394.6	Sb06g024360.1|PACid:1973369	6	53.478.706	53.479.806	+
evm.model.scaffold_2407.1	Sb02g040460.1|PACid:1959546	2	74.386.506	74.389.007	-
evm.model.scaffold_2407.2	Sb02g040450.1|PACid:1959545	2	74.382.202	74.384.095	-
evm.model.scaffold_257.2	Sb07g028040.1|PACid:1976958	7	63.002.738	63.004.800	+
evm.model.scaffold_257.4	Sb07g028050.1|PACid:1976959	7	63.007.131	63.012.078	-
evm.model.scaffold_257.5	Sb07g028065.1|PACid:1976962	7	63.027.930	63.028.675	+
evm.model.scaffold_257.6	Sb07g028070.1|PACid:1976963	7	63.029.524	63.033.671	-
evm.model.scaffold_264.8	Sb04g008920.1|PACid:1965987	4	10.628.717	10.629.429	-
evm.model.scaffold_264.9	Sb04g008930.1|PACid:1965988	4	10.636.413	10.637.242	-
evm.model.scaffold_264.10	Sb04g008940.1|PACid:1965989	4	10.645.399	10.646.097	-
evm.model.scaffold_264.13	Sb04g008950.1|PACid:1965990	4	10.671.311	10.671.829	-
evm.model.scaffold_264.14	Sb04g008960.1|PACid:1965991	4	10.701.214	10.701.703	-
evm.model.scaffold_264.15	Sb04g008970.1|PACid:1965992	4	10.702.946	10.703.617	-
evm.model.scaffold_265.2	Sb09g030830.1|PACid:1982135	9	59.468.049	59.469.774	+
evm.model.scaffold_265.3	Sb09g030840.1|PACid:1982136	9	59.470.162	59.474.046	-
evm.model.scaffold_265.7	Sb09g030850.1|PACid:1982137	9	59.479.642	59.483.425	-
evm.model.scaffold_269.1	Sb03g031630.1|PACid:1962920	3	60.053.930	60.057.542	+
evm.model.scaffold_269.2	Sb03g031640.1|PACid:1962921	3	60.057.640	60.063.820	-
evm.model.scaffold_270.1	Sb08g005110.1|PACid:1977746	8	6.531.235	6.532.218	+
evm.model.scaffold_270.2	Sb08g005060.1|PACid:1977739	8	6.418.158	6.421.690	-
evm.model.scaffold_270.3	Sb08g005090.1|PACid:1977744	8	6.497.715	6.501.035	-

evm.model.scaffold_271.3	Sb02g041980.1|PACid:1959729	2	75.701.312	75.702.846	-
evm.model.scaffold_271.4	Sb02g041970.2|PACid:1959728	2	75.694.339	75.700.466	+
evm.model.scaffold_274.1	Sb09g027610.1|PACid:1981747	9	56.669.479	56.675.110	+
evm.model.scaffold_274.2	Sb09g027610.1|PACid:1981747	9	56.669.479	56.675.110	+
evm.model.scaffold_274.4	Sb09g027630.1|PACid:1981749	9	56.692.165	56.695.716	-
evm.model.scaffold_274.5	Sb09g027620.1|PACid:1981748	9	56.686.018	56.690.207	+
evm.model.scaffold_275.1	Sb01g004190.1|PACid:1949792	1	3.363.318	3.364.904	-
evm.model.scaffold_275.2	Sb01g004180.1|PACid:1949791	1	3.360.775	3.362.611	+
evm.model.scaffold_275.3	Sb01g004160.1|PACid:1949789	1	3.356.299	3.357.587	+
evm.model.scaffold_286.1	Sb09g020340.1|PACid:1980872	9	49.616.534	49.617.892	+
evm.model.scaffold_286.5	Sb09g020330.1|PACid:1980869	9	49.600.234	49.604.287	+
evm.model.scaffold_299.3	Sb02g027450.1|PACid:1957940	2	62.703.169	62.707.399	-
evm.model.scaffold_299.4	Sb02g027440.1|PACid:1957939	2	62.698.863	62.699.970	+
evm.model.scaffold_299.6	Sb02g027430.1|PACid:1957938	2	62.690.316	62.692.694	+
evm.model.scaffold_3.1	Sb05g026330.1|PACid:1971199	5	60.184.870	60.188.044	-
evm.model.scaffold_3.2	Sb05g026330.1|PACid:1971199	5	60.184.870	60.188.044	-
evm.model.scaffold_3.3	Sb05g026330.1|PACid:1971199	5	60.184.870	60.188.044	-
evm.model.scaffold_3.4	Sb06g029060.1|PACid:1973931	6	57.762.682	57.763.123	+
evm.model.scaffold_308.1	Sb04g006100.1|PACid:1965622	4	5.996.166	6.004.566	-
evm.model.scaffold_308.2	Sb04g006100.1|PACid:1965622	4	5.996.166	6.004.566	-
evm.model.scaffold_345.1	Sb01g044090.1|PACid:1954424	1	67.251.110	67.255.570	-
evm.model.scaffold_345.2	Sb01g044080.1|PACid:1954423	1	67.244.974	67.245.378	+
evm.model.scaffold_377.6	Sb07g026270.1|PACid:1976747	7	61.481.521	61.484.475	-
evm.model.scaffold_377.7	Sb07g026280.1|PACid:1976748	7	61.486.878	61.488.048	+
evm.model.scaffold_377.9	Sb07g026290.1|PACid:1976749	7	61.488.389	61.491.915	-
evm.model.scaffold_377.10	Sb07g026300.1|PACid:1976750	7	61.494.736	61.496.535	-
evm.model.scaffold_377.11	Sb07g026300.1|PACid:1976750	7	61.494.736	61.496.535	-
evm.model.scaffold_388.1	Sb04g001760.1|PACid:1965065	4	1.503.538	1.506.346	+
evm.model.scaffold_388.3	Sb04g001760.1|PACid:1965065	4	1.503.538	1.506.346	+
evm.model.scaffold_388.5	Sb04g001760.1|PACid:1965065	4	1.503.538	1.506.346	+
evm.model.scaffold_405.2	Sb04g006470.1|PACid:1965667	4	6.477.568	6.482.182	-
evm.model.scaffold_405.4	Sb04g006470.1|PACid:1965667	4	6.477.568	6.482.182	-
evm.model.scaffold_427.1	Sb03g033590.1|PACid:1963165	3	61.845.919	61.850.880	+
evm.model.scaffold_427.2	Sb03g033585.1|PACid:1963164	3	61.840.092	61.841.466	-
evm.model.scaffold_448.1	Sb04g009750.1|PACid:1966069	4	12.355.789	12.359.949	+
evm.model.scaffold_448.2	Sb04g009750.1|PACid:1966069	4	12.355.789	12.359.949	+
evm.model.scaffold_467.3	Sb07g005850.1|PACid:1975318	7	8.311.066	8.313.249	-
evm.model.scaffold_467.6	Sb07g004720.1|PACid:1975166	7	6.261.121	6.262.757	-
evm.model.scaffold_469.2	Sb03g033570.1|PACid:1963162	3	61.832.915	61.833.329	+
evm.model.scaffold_469.3	Sb03g033560.1|PACid:1963161	3	61.827.664	61.832.436	+
evm.model.scaffold_469.6	Sb03g033550.1|PACid:1963160	3	61.820.363	61.822.645	-
evm.model.scaffold_494.2	Sb09g030500.1|PACid:1982088	9	59.057.081	59.059.977	-
evm.model.scaffold_494.3	Sb09g030505.1|PACid:1982089	9	59.066.774	59.067.136	+
evm.model.scaffold_54.1	Sb05g026315.1|PACid:1971197	5	60.176.072	60.177.292	+
evm.model.scaffold_54.2	Sb05g026330.1|PACid:1971199	5	60.184.870	60.188.044	-

evm.model.scaffold_55.1	Sb02g040430.1|PACid:1959543	2	74.369.213	74.370.619	+
evm.model.scaffold_55.2	Sb02g040420.1|PACid:1959542	2	74.360.493	74.362.043	-
evm.model.scaffold_55.3	Sb02g025930.2|PACid:1957739	2	60.978.501	60.981.390	+
evm.model.scaffold_552.1	Sb05g026630.1|PACid:1971239	5	60.626.278	60.629.989	+
evm.model.scaffold_552.2	Sb05g025570.1|PACid:1971089	5	58.861.750	58.865.331	+
evm.model.scaffold_552.3	Sb05g025570.1|PACid:1971089	5	58.861.750	58.865.331	+
evm.model.scaffold_552.4	Sb05g025590.1|PACid:1971091	5	58.875.263	58.876.930	-
evm.model.scaffold_557.1	Sb01g050610.1|PACid:1955233	1	73.690.442	73.693.499	-
evm.model.scaffold_557.2	Sb01g050600.1|PACid:1955232	1	73.686.461	73.690.254	-
evm.model.scaffold_56.1	Sb09g007920.2|PACid:1980235	9	14.526.212	14.529.007	+
evm.model.scaffold_56.2	Sb09g007930.1|PACid:1980236	9	14.529.316	14.531.555	-
evm.model.scaffold_564.1	Sb03g031720.1|PACid:1962933	3	60.119.631	60.119.924	-
evm.model.scaffold_564.3	Sb03g031710.1|PACid:1962930	3	60.115.828	60.117.031	+
evm.model.scaffold_57.2	Sb01g027810.1|PACid:1952412	1	48.417.157	48.423.798	+
evm.model.scaffold_57.3	Sb01g027820.1|PACid:1952413	1	48.449.904	48.456.508	+
evm.model.scaffold_593.7	Sb01g018700.1|PACid:1951569	1	19.545.193	19.552.496	+
evm.model.scaffold_593.10	Sb01g018690.1|PACid:1951567	1	19.522.762	19.523.178	-
evm.model.scaffold_601.1	Sb03g044500.2|PACid:1964464	3	71.853.994	71.856.810	-
evm.model.scaffold_601.2	Sb03g044485.1|PACid:1964461	3	71.843.543	71.847.704	+
evm.model.scaffold_614.1	Sb02g003940.1|PACid:1955731	2	4.376.629	4.377.866	+
evm.model.scaffold_614.2	Sb02g003950.1|PACid:1955732	2	4.378.871	4.380.601	-
evm.model.scaffold_614.6	Sb02g003970.1|PACid:1955734	2	4.383.018	4.387.609	-
evm.model.scaffold_614.7	Sb02g003980.1|PACid:1955736	2	4.390.494	4.392.601	+
evm.model.scaffold_614.8	Sb02g003990.1|PACid:1955737	2	4.392.808	4.396.800	-
evm.model.scaffold_649.2	Sb06g005410.1|PACid:1971886	6	13.580.446	13.581.583	+
evm.model.scaffold_649.3	Sb06g005410.1|PACid:1971886	6	13.580.446	13.581.583	+
evm.model.scaffold_659.2	Sb04g001190.1|PACid:1964994	4	1.012.924	1.016.515	+
evm.model.scaffold_659.3	Sb04g001200.1|PACid:1964995	4	1.016.937	1.019.794	-
evm.model.scaffold_66.4	Sb07g020340.1|PACid:1975990	7	52.819.468	52.819.737	+
evm.model.scaffold_66.5	Sb07g020335.1|PACid:1975989	7	52.754.411	52.758.032	-
evm.model.scaffold_66.6	Sb07g020310.1|PACid:1975986	7	52.723.574	52.727.545	-
evm.model.scaffold_66.7	Sb07g020300.1|PACid:1975982	7	52.685.122	52.689.766	-
evm.model.scaffold_683.1	Sb04g000650.1|PACid:1964914	4	465.415	469.383	+
evm.model.scaffold_683.2	Sb04g000660.1|PACid:1964915	4	470.601	473.118	-
evm.model.scaffold_683.6	Sb04g000670.1|PACid:1964917	4	483.669	491.585	+
evm.model.scaffold_683.7	Sb04g000670.1|PACid:1964917	4	483.669	491.585	+
evm.model.scaffold_683.8	Sb04g000800.1|PACid:1964935	4	619.923	621.568	-
evm.model.scaffold_683.9	Sb04g000690.1|PACid:1964919	4	495.706	497.775	-
evm.model.scaffold_688.2	Sb06g006720.1|PACid:1971911	6	16.139.824	16.161.531	+
evm.model.scaffold_688.4	Sb06g006720.1|PACid:1971911	6	16.139.824	16.161.531	+
evm.model.scaffold_688.6	Sb06g005900.1|PACid:1971902	6	15.077.564	15.099.135	-
evm.model.scaffold_698.5	Sb03g034540.1|PACid:1963283	3	62.722.950	62.724.361	-
evm.model.scaffold_698.7	Sb03g034530.1|PACid:1963282	3	62.718.341	62.719.099	-
evm.model.scaffold_698.9	Sb03g034510.1|PACid:1963280	3	62.698.868	62.708.365	-
evm.model.scaffold_722.1	Sb01g004090.1|PACid:1949782	1	3.310.786	3.312.790	+

evm.model.scaffold_722.2	Sb01g004090.1|PACid:1949782	1	3.310.786	3.312.790	+
evm.model.scaffold_722.3	Sb01g004080.1|PACid:1949781	1	3.304.490	3.307.002	-
evm.model.scaffold_722.4	Sb01g004070.1|PACid:1949780	1	3.302.963	3.304.103	+
evm.model.scaffold_722.6	Sb01g003050.1|PACid:1949654	1	2.466.904	2.468.916	+
evm.model.scaffold_723.1	Sb02g022760.1|PACid:1957334	2	56.018.219	56.019.628	+
evm.model.scaffold_723.2	Sb02g022750.1|PACid:1957333	2	56.011.526	56.012.048	+
evm.model.scaffold_723.3	Sb02g022750.1|PACid:1957333	2	56.011.526	56.012.048	+
evm.model.scaffold_723.5	Sb02g022660.1|PACid:1957321	2	55.810.778	55.811.482	-
evm.model.scaffold_729.1	Sb03g031650.1|PACid:1962922	3	60.064.828	60.066.130	-
evm.model.scaffold_729.2	Sb03g031680.1|PACid:1962926	3	60.085.127	60.088.181	+
evm.model.scaffold_729.3	Sb03g031670.1|PACid:1962925	3	60.080.358	60.084.458	+
evm.model.scaffold_729.4	Sb03g031655.1|PACid:1962923	3	60.075.964	60.076.713	-
evm.model.scaffold_732.1	Sb03g009820.1|PACid:1961212	3	10.583.678	10.588.407	-
evm.model.scaffold_732.3	Sb03g009830.1|PACid:1961213	3	10.601.371	10.604.803	+
evm.model.scaffold_732.5	Sb03g009840.1|PACid:1961214	3	10.606.733	10.608.811	-
evm.model.scaffold_732.6	Sb03g009840.1|PACid:1961214	3	10.606.733	10.608.811	-
evm.model.scaffold_735.1	Sb04g002063.1|PACid:1965109	4	1.880.652	1.890.831	+
evm.model.scaffold_735.3	Sb04g002060.1|PACid:1965108	4	1.863.229	1.876.993	+
evm.model.scaffold_743.1	Sb10g010040.1|PACid:1983409	10	12.267.089	12.269.181	-
evm.model.scaffold_743.3	Sb10g010030.1|PACid:1983408	10	12.261.245	12.262.235	+
evm.model.scaffold_770.3	Sb07g008810.1|PACid:1975550	7	14.772.904	14.773.601	+
evm.model.scaffold_770.8	Sb07g008580.1|PACid:1975543	7	14.515.561	14.516.832	-
evm.model.scaffold_770.9	Sb07g008570.1|PACid:1975541	7	14.506.211	14.511.902	+
evm.model.scaffold_808.3	Sb09g028180.1|PACid:1981812	9	57.142.354	57.144.361	-
evm.model.scaffold_808.4	Sb09g028200.1|PACid:1981814	9	57.156.820	57.157.911	-
evm.model.scaffold_808.6	Sb09g028210.1|PACid:1981815	9	57.163.124	57.167.527	-
evm.model.scaffold_809.2	Sb03g008480.1|PACid:1961056	3	9.036.759	9.039.997	+
evm.model.scaffold_809.4	Sb03g008490.1|PACid:1961057	3	9.048.157	9.057.522	+
evm.model.scaffold_809.5	Sb03g008500.1|PACid:1961058	3	9.061.875	9.067.760	-
evm.model.scaffold_809.7	Sb03g008510.1|PACid:1961059	3	9.097.287	9.099.902	-
evm.model.scaffold_810.1	Sb08g017310.1|PACid:1978508	8	45.897.314	45.898.876	-
evm.model.scaffold_810.2	Sb08g017300.1|PACid:1978507	8	45.863.449	45.881.012	-
evm.model.scaffold_810.3	Sb08g017290.1|PACid:1978506	8	45.856.092	45.860.678	-
evm.model.scaffold_812.1	Sb09g002420.1|PACid:1979597	9	2.641.848	2.643.249	-
evm.model.scaffold_812.2	Sb09g002420.1|PACid:1979597	9	2.641.848	2.643.249	-
evm.model.scaffold_814.1	Sb01g027890.2|PACid:1952424	1	48.536.375	48.539.226	-
evm.model.scaffold_814.2	Sb01g027880.1|PACid:1952422	1	48.529.772	48.533.838	-
evm.model.scaffold_814.3	Sb01g027870.1|PACid:1952421	1	48.525.230	48.527.120	+
evm.model.scaffold_851.1	Sb10g010030.1|PACid:1983408	10	12.261.245	12.262.235	+
evm.model.scaffold_851.2	Sb10g010020.1|PACid:1983407	10	12.253.035	12.254.424	+
evm.model.scaffold_853.1	Sb10g002310.1|PACid:1982416	10	1.994.362	1.997.869	-
evm.model.scaffold_853.2	Sb10g002300.1|PACid:1982415	10	1.986.832	1.990.238	+
evm.model.scaffold_854.1	Sb01g004960.1|PACid:1949895	1	4.050.457	4.054.657	-
evm.model.scaffold_854.2	Sb01g004970.1|PACid:1949896	1	4.055.998	4.057.549	-
evm.model.scaffold_854.3	Sb01g004980.1|PACid:1949897	1	4.061.511	4.061.840	+

evm.model.scaffold_859.4	Sb01g018790.1|PACid:1951578	1	19.707.212	19.717.615	-
evm.model.scaffold_859.5	Sb01g018790.1|PACid:1951578	1	19.707.212	19.717.615	-
evm.model.scaffold_859.6	Sb01g027850.2|PACid:1952417	1	48.510.652	48.513.447	+
evm.model.scaffold_859.7	Sb01g027830.1|PACid:1952414	1	48.459.131	48.462.694	+
evm.model.scaffold_859.9	Sb01g027820.1|PACid:1952413	1	48.449.904	48.456.508	+
evm.model.scaffold_860.1	Sb08g020220.1|PACid:1978898	8	51.245.972	51.249.106	+
evm.model.scaffold_860.2	Sb08g020230.1|PACid:1978899	8	51.249.500	51.252.088	-
evm.model.scaffold_860.3	Sb08g020240.1|PACid:1978900	8	51.253.721	51.254.884	-
evm.model.scaffold_860.4	Sb08g020250.1|PACid:1978901	8	51.258.182	51.264.057	-
evm.model.scaffold_860.5	Sb08g020260.1|PACid:1978902	8	51.264.909	51.270.727	+
evm.model.scaffold_860.7	Sb08g020260.1|PACid:1978902	8	51.264.909	51.270.727	+
evm.model.scaffold_860.8	Sb08g020270.1|PACid:1978903	8	51.276.577	51.278.999	-
evm.model.scaffold_860.9	Sb08g020280.1|PACid:1978905	8	51.307.733	51.309.202	+
evm.model.scaffold_861.2	Sb05g026335.1|PACid:1971200	5	60.193.990	60.197.805	+
evm.model.scaffold_861.8	Sb05g026335.1|PACid:1971200	5	60.193.990	60.197.805	+
evm.model.scaffold_869.1	Sb03g029340.1|PACid:1962636	3	57.485.238	57.490.269	-
evm.model.scaffold_869.3	Sb03g029330.1|PACid:1962635	3	57.465.294	57.469.736	-
evm.model.scaffold_869.5	Sb03g029320.1|PACid:1962634	3	57.449.784	57.453.744	+
evm.model.scaffold_869.6	Sb03g029310.1|PACid:1962633	3	57.439.748	57.442.504	-
evm.model.scaffold_869.7	Sb03g029300.1|PACid:1962632	3	57.434.330	57.435.217	-
evm.model.scaffold_899.2	Sb09g005720.1|PACid:1980021	9	7.620.684	7.625.880	-
evm.model.scaffold_899.3	Sb09g005700.1|PACid:1980019	9	7.601.936	7.603.335	-
evm.model.scaffold_899.4	Sb09g005695.1|PACid:1980018	9	7.595.722	7.599.150	-
evm.model.scaffold_921.1	Sb07g020770.1|PACid:1976045	7	53.641.687	53.645.584	-
evm.model.scaffold_921.2	Sb07g020760.1|PACid:1976044	7	53.640.854	53.641.924	+
evm.model.scaffold_940.4	Sb02g026010.1|PACid:1957752	2	61.045.346	61.051.060	-
evm.model.scaffold_940.6	Sb02g026020.1|PACid:1957753	2	61.058.535	61.063.537	+
evm.model.scaffold_940.7	Sb02g026030.1|PACid:1957754	2	61.065.808	61.066.747	-
evm.model.scaffold_961.1	Sb02g025980.1|PACid:1957748	2	61.026.436	61.028.668	+
evm.model.scaffold_961.2	Sb02g025985.1|PACid:1957749	2	61.029.252	61.030.784	-
evm.model.scaffold_961.3	Sb02g025990.1|PACid:1957750	2	61.037.031	61.038.134	+
evm.model.scaffold_980.1	Sb05g013170.1|PACid:1970096	5	27.166.690	27.211.521	+
evm.model.scaffold_980.2	Sb05g012300.1|PACid:1970081	5	25.491.813	25.493.458	-
evm.model.scaffold_982.1	Sb03g013500.1|PACid:1961659	3	16.877.573	16.879.769	-
evm.model.scaffold_982.2	Sb03g013510.1|PACid:1961660	3	16.887.767	16.890.526	-

Supplementary Table 8 - Distribution of gene models among scaffolds

Number of scaffolds	Number of Gene Models	Scaffold Size Range
1	17	144,803 - 144,803
1	15	122,587 - 122,587
2	14	135,690 - 203,132
1	13	120,035 - 120,035
1	12	151,964 - 151,964
4	11	51,028 - 134,888
6	10	50,115 - 141,339
9	9	49,010 - 131,796
11	8	39,928 - 138,830
24	7	37,498 - 152,524
22	6	39,569 - 127,716
24	5	12,278 - 128,730
28	4	16,384 - 90,365
40	3	5,459 - 80,959
71	2	1,695 - 119,761
186	1	233 - 129,783
Total 431	1,338	**Total Bases:** 15,435,186

DISCUSSÃO GERAL

A cana-de-açúcar é uma cultura de grande importância econômica para o país e de interesse do nosso laboratório no estudo da evolução das gramíneas. Derivado do cruzamento entre *Saccharum officinarum* and *Saccharum spontaneum*, o genoma de cana apresenta alto nível de ploidia e tem em torno de 55% de sequências repetitivas. Devido a complexidade do genoma, uma alternativa adotada para acessar a informação genética é sequenciar clones de BACs. Assim, foi construída uma biblioteca de BACs da variedade comercial SP80-3280 de cana-de-açúcar. Uma amostra aleatória de 192 clones foi selecionada para análises preliminares, a partir de 384 sequências Sanger oriundas das pontas destes BACs (*BAC end sequences* – BES). O mapeamento no genoma de sorgo mostrou uma boa distribuição dos BES de cana-de-açúcar ao longo dos cromossomos de sorgo. Além disso, o alinhamento de 42 pares de BES em regiões sintênicas concordantes do genoma de sorgo mostraram uma expansão de 29% do genoma de sorgo em relação ao genoma monoploide de cana. Este resultado está de acordo com as análises realizadas com o alinhamento de regiões sintênicas de 20 BACs de cana da variedade R570 com o genoma de sorgo [23], sugerindo uma expansão de 20.7% do genoma sorgo quando comparado ao genoma de cana-de-açúcar.

Devido ao alto custo e à dificuldade do sequenciamento individual BAC a BAC, foi proposto neste trabalho o sequenciamento de BACs em pool, anteriormente realizado para alguns organismos usando a plataforma de sequenciamento 454 [16,17,18]. Devido ao maior rendimento e menor custo entre os sequenciadores da nova geração, nossa proposta foi utilizar a plataforma de sequenciamento Illumina, onde uma quantidade maior de BACs em pools pode ser sequenciada em uma única corrida. Para avaliar o balanço entre quantidade de BACs por pool e a complexidade da montagem derivada do aumento do tamanho dos pools, foi realizada uma análise com dados simulados de sequenciamento Illumina, usando o genoma sequenciado de sorgo. Embora a simulação contemple alguns aspectos que impactam na montagem, como qualidade das bases, taxas de erros e reads derivados de bibliotecas com diferentes tamanhos de fragmentos, o cenário proposto não considera diferenças de amostragem no número de reads entre bibliotecas, e possíveis regiões não cobertas pelo sequenciamento. Mesmo assim, com um

cenário mais ideal, foi possível avaliar o efeito do aumento no número de BACs por pool, e verificar que a maior parte dos BACs (60%-70%) esta bem montada, mesmo quando há algumas centenas de BACs por pool.

Foi feito então um teste para sequenciar um pool com 178 BACs de cana-de-açúcar, amostrados aleatoriamente. O pool de BACs foi sequenciado com a plataforma Illumina e a plataforma PacBio. Os reads longos PacBio foram usados para suprir a falta das bibliotecas Illumina de fragmentos maiores, que apresentam baixo rendimento e geram reads quiméricos [24]. Embora não previsto anteriormente, a utilização dos reads longos PacBio traz a vantagem da sequência completa de ponta a ponta, frente a pares de reads curtos ligados a uma determinada distância. Por se tratar de uma abordagem nova, com um número muito reduzido de softwares e algoritmos, e com uma metodologia ainda não muito bem estabelecida, foi necessário testar diferentes estratégias de montagem híbrida para combinar reads curtos Illumina com reads longos PacBio. Três estratégias de montagem híbrida Illumina-PacBio foram utilizadas: montagem de reads PacBio corrigidos com reads Illumina [25], montagem de contigs Illumina e contigs PacBio, scaffolding dos contigs Illumina com reads PacBio [26, 27]. Alguns softwares de montagem (phrap, cap3, MIRA) foram usados para montar diretamente os reads curtos Illumina e reads longos PacBio, mas falharam. Além das métricas de montagem, o tamanho estimado dos BACs e as sequências das pontas dos BACs (BES) foram utilizados na avaliação da melhor estratégia para montagem híbrida. A estratégia de scaffolding dos contigs Illumina com reads PacBio usando o software AHA [27] produziu a melhor montagem para o pool de 178 BACs, com 19.2 Mb distribuídos em 2.451 scaffolds, dos quais ~80% têm no mínimo 20Kb. O alinhamento dos scaffolds de cana-de-açúcar com os cromossomos de sorgo mostrou alto grau de colineridade e ordem dos genes, indicando a corretude da montagem. Além disso, os scaffolds apresentam uma distribuição uniforme ao longo dos cromossomos de sorgo, confirmando a aleatoriedade da amostra de BACs selecionada. As sequências alinhadas sintenicamente revelaram regiões expandidas e contraídas entre cana e sorgo, que somadas mostram uma expansão de 19% do genoma de sorgo em relação ao genoma monoploide de cana-de-açúcar. Este resultado concorda com o trabalho anterior realizado com os BES de cana-de-açúcar da variedade SP80-3280, e com os resultados da análise de 20 BACs de cana da variedade R570 [23]. Porém,

um trabalho recente envolvendo 317 BACs cana da variedade R570 [28] diverge deste resultado, mostrando que o genoma de cana é expandido em relação ao genoma de sorgo. Como a expansão pode variar entre as regiões sintenicas analisadas, uma amostra mais representativa tende a indicar analises mais corretas. Apesar do número de 317 BACs ser maior e possivelmente mais representativo, a seleção dos BACs foi direcionada para regiões de interesse [28]. Diferentemente, os 178 BACs deste trabalho foram selecionados aleatoriamente, resultando em scaffolds bem distribuídos ao longo dos cromossomos de sorgo, e que alinham em algumas regiões do genoma de sorgo não mapeadas com os 317 BACs. Outro indicativo que nossa amostra de BACs é mais representativa deriva da análise entre os elementos Gypsy e Copia da família de retrotransposons LTR. Em sorgo [4], os elementos Gypsy são mais abundantes do que elementos Copia na razão 3.7 para 1, e estão mais concentrados em regiões próximas aos centromeros. Supondo que cana-de-açúcar tenha com composição parecida de elementos retrotransposons LTR, nossa amostra de BACs apresenta uma razão de 2.3 para 1 que é mais similar do que a razão de 1.3 para 1 identificada com os 317 BACs. Estes números demonstram a natureza randômica do pool de BACs apresentado neste trabalho, que amostra regiões distintas do genoma, inclusive regiões próximas ao centromero.

CONCLUSÕES

Os resultados obtidos neste trabalho mostram a viabilidade de sequenciar pools de BACs de genomas complexos, como cana-de-açúcar, cujo genoma apresenta altos níveis de ploidia. Com um custo baixo, foi sequenciado e montado um pool com 178 BACs selecionados diretamente da biblioteca de cana-de-açúcar da variedade comercial SP80-3280, usando as plataformas de sequenciamento de nova geração Illumina HiSeq2000 e PacBio. Além disso, seleção randômica dos clones de BACs para montar o pool, possibilitou uma amostragem representativa do genoma, visto a distribuição homogênea dos scaffolds de cana quando alinhados contra os cromossomos de sorgo. A estratégia de sequenciar genomas complexos usando pools de BACs selecionados diretamente e randomicamente da biblioteca de BACs, pode ser uma alternativa mais custo-efetiva se comparada a estratégia de *whole genome shotgun*.

REFERÊNCIAS

[1] T. Arabidopsis, T. Arabidopsis, G. Initiative, G. Initiative, A. G. Initiative, and A. G. Initiative, “Analysis of the genome sequence of the flowering plant Arabidopsis thaliana.,” *Nature*, vol. 408, no. 6814, pp. 796–815, 2000.

[2] I. Rice and G. Sequencing, “The map-based sequence of the rice genome.,” *Nature*, vol. 436, no. 7052, pp. 793–800, 2005.

[3] O. Jaillon, J.-M. Aury, B. Noel, A. Policriti, C. Clepet, A. Casagrande, N. Choisne, S. Aubourg, N. Vitulo, C. Jubin, A. Vezzi, F. Legeai, P. Hugueney, C. Dasilva, D. Horner, E. Mica, D. Jublot, J. Poulain, C. Bruyère, A. Billault, B. Segurens, M. Gouyvenoux, E. Ugarte, F. Cattonaro, V. Anthouard, V. Vico, C. Del Fabbro, M. Alaux, G. Di Gaspero, V. Dumas, N. Felice, S. Paillard, I. Juman, M. Moroldo, S. Scalabrin, A. Canaguier, I. Le Clainche, G. Malacrida, E. Durand, G. Pesole, V. Laucou, P. Chatelet, D. Merdinoglu, M. Delledonne, M. Pezzotti, A. Lecharny, C. Scarpelli, F. Artiguenave, M. E. Pè, G. Valle, M. Morgante, M. Caboche, A.-F. Adam-Blondon, J. Weissenbach, F. Quétier, and P. Wincker, “The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla.,” *Nature*, vol. 449, no. 7161, pp. 463–467, 2007.

[4] A. H. Paterson, J. E. Bowers, R. Bruggmann, I. Dubchak, J. Grimwood, H. Gundlach, G. Haberer, U. Hellsten, T. Mitros, A. Poliakov, J. Schmutz, M. Spannagl, H. Tang, X. Wang, T. Wicker, A. K. Bharti, J. Chapman, F. A. Feltus, U. Gowik, I. V Grigoriev, E. Lyons, C. a Maher, M. Martis, A. Narechania, R. P. Otillar, B. W. Penning, A. a Salamov, Y. Wang, L. Zhang, N. C. Carpita, M. Freeling, A. R. Gingle, C. T. Hash, B. Keller, P. Klein, S. Kresovich, M. C. McCann, R. Ming, D. G. Peterson, Mehboob-ur-Rahman, D. Ware, P. Westhoff, K. F. X. Mayer, J. Messing, and D. S. Rokhsar, “The Sorghum bicolor genome and the diversification of grasses.,” *Nature*, vol. 457, no. 7229, pp. 551–556, 2009.

[5] C. June, S. E. E. L. Page, S. Pasternak, C. Liang, J. Zhang, L. Fulton, T. A. Graves, P. Minx, A. D. Reily, L. Courtney, S. S. Kruchowski, C. Tomlinson, C. Strong, K. Delehaunty, C. Fronick, B. Courtney, S. M. Rock, E. Belter, F. Du, K. Kim, R. M. Abbott, M. Cotton, A. Levy, P. Marchetto, K. Ochoa, S. M. Jackson, B. Gillam, W. Chen, L. Yan, J. Higginbotham, M. Cardenas, J. Waligorski, E. Applebaum, L. Phelps, J. Falcone, K. Kanchi, T. Thane, A. Scimone, N. Thane, J. Henke, T. Wang, J. Ruppert, N. Shah, K. Rotter, J. Hodges, E. Ingenthron, M. Cordes, S. Kohlberg, J. Sgro, B. Delgado, K. Mead, A. Chinwalla, S. Leonard, K. Crouse, K. Collura, D. Kudrna, J. Currie, R. He, A. Angelova, S. Rajasekar, T. Mueller, R. Lomeli, G. Scara, A. Ko, K. Delaney, M. Wissotski, G. Lopez, D. Campos, M. Braidotti, E. Ashley, W. Golser, H. Kim, S. Lee, J. Lin, Z. Dujmic, W. Kim, J. Talag, A. Zuccolo, C. Fan, A. Sebastian, M. Kramer, L. Spiegel, L. Nascimento, T. Zutavern, B. Miller, C. Ambroise, S. Muller, W.

Spooner, A. Narechania, L. Ren, S. Wei, and S. Kumari, “The B73 Maize Genome: Complexity, Diversity, and Dynamics,” no. June, 2012.

[6] J. Schmutz, S. B. Cannon, J. Schlueter, J. Ma, T. Mitros, W. Nelson, D. L. Hyten, Q. Song, J. J. Thelen, J. Cheng, D. Xu, U. Hellsten, G. D. May, Y. Yu, T. Sakurai, T. Umezawa, M. K. Bhattacharyya, D. Sandhu, B. Valliyodan, E. Lindquist, M. Peto, D. Grant, S. Shu, D. Goodstein, K. Barry, M. Futrell-Griggs, B. Abernathy, J. Du, Z. Tian, L. Zhu, N. Gill, T. Joshi, M. Libault, A. Sethuraman, X.-C. Zhang, K. Shinozaki, H. T. Nguyen, R. a Wing, P. Cregan, J. Specht, J. Grimwood, D. Rokhsar, G. Stacey, R. C. Shoemaker, and S. a Jackson, “Genome sequence of the palaeopolyploid soybean.,” *Nature*, vol. 463, no. 7278, pp. 178–183, 2010.

[7] M. L. Metzker, “Sequencing technologies - the next generation.,” *Nat. Rev. Genet.*, vol. 11, no. 1, pp. 31–46, 2010.

[8] Q. Xu, L.-L. Chen, X. Ruan, D. Chen, A. Zhu, C. Chen, D. Bertrand, W.-B. Jiao, B.-H. Hao, M. P. Lyon, J. Chen, S. Gao, F. Xing, H. Lan, J.-W. Chang, X. Ge, Y. Lei, Q. Hu, Y. Miao, L. Wang, S. Xiao, M. K. Biswas, W. Zeng, F. Guo, H. Cao, X. Yang, X.-W. Xu, Y.-J. Cheng, J. Xu, J.-H. Liu, O. J. Luo, Z. Tang, W.-W. Guo, H. Kuang, H.-Y. Zhang, M. L. Roose, N. Nagarajan, X.-X. Deng, and Y. Ruan, “The draft genome of sweet orange (Citrus sinensis).,” *Nat. Genet.*, vol. 45, no. 1, pp. 59–66, 2012.

[9] S. Kim, M. Park, S.-I. Yeom, Y.-M. Kim, J. M. Lee, H.-A. Lee, E. Seo, J. Choi, K. Cheong, K.-T. Kim, K. Jung, G.-W. Lee, S.-K. Oh, C. Bae, S.-B. Kim, H.-Y. Lee, S.-Y. Kim, M.-S. Kim, B.-C. Kang, Y. D. Jo, H.-B. Yang, H.-J. Jeong, W.-H. Kang, J.-K. Kwon, C. Shin, J. Y. Lim, J. H. Park, J. H. Huh, J.-S. Kim, B.-D. Kim, O. Cohen, I. Paran, M. C. Suh, S. B. Lee, Y.-K. Kim, Y. Shin, S.-J. Noh, J. Park, Y. S. Seo, S.-Y. Kwon, H. a Kim, J. M. Park, H.-J. Kim, S.-B. Choi, P. W. Bosland, G. Reeves, S.-H. Jo, B.-W. Lee, H.-T. Cho, H.-S. Choi, M.-S. Lee, Y. Yu, Y. Do Choi, B.-S. Park, A. van Deynze, H. Ashrafi, T. Hill, W. T. Kim, H.-S. Pai, H. K. Ahn, I. Yeam, J. J. Giovannoni, J. K. C. Rose, I. Sørensen, S.-J. Lee, R. W. Kim, I.-Y. Choi, B.-S. Choi, J.-S. Lim, Y.-H. Lee, and D. Choi, “Genome sequence of the hot pepper provides insights into the evolution of pungency in Capsicum species.,” *Nat. Genet.*, vol. 46, no. 3, pp. 270–8, 2014.

[10] T. Zhang, Y. Hu, W. Jiang, L. Fang, X. Guan, J. Chen, J. Zhang, C. a Saski, B. E. Scheffler, D. M. Stelly, A. M. Hulse-Kemp, Q. Wan, B. Liu, C. Liu, S. Wang, M. Pan, Y. Wang, D. Wang, W. Ye, L. Chang, W. Zhang, Q. Song, R. C. Kirkbride, X. Chen, E. Dennis, D. J. Llewellyn, D. G. Peterson, P. Thaxton, D. C. Jones, Q. Wang, X. Xu, H. Zhang, H. Wu, L. Zhou, G. Mei, S. Chen, Y. Tian, D. Xiang, X. Li, J. Ding, Q. Zuo, L. Tao, Y. Liu, J. Li, Y. Lin, Y. Hui, Z. Cao, C. Cai, X. Zhu, Z. Jiang, B. Zhou, W. Guo, R. Li, and Z. J. Chen, “Sequencing of allotetraploid cotton (Gossypium hirsutum L. acc. TM-1) provides a resource for fiber improvement,” *Nat. Biotechnol.*, vol. 33, no. 5, 2015.

[11] A. Y. A. Rahman, A. O. Usharraj, B. B. Misra, G. P. Thottathil, K. Jayasekaran, Y. Feng, S. Hou, S. Y. Ong, F. L. Ng, L. S. Lee, H. S. Tan, M. K. L. M. Sakaff,

B. S. Teh, B. F. Khoo, S. S. Badai, N. A. Aziz, A. Yuryev, B. Knudsen, A. Dionne-Laporte, N. P. Mchunu, Q. Yu, B. J. Langston, T. A. K. Freitas, A. G. Young, R. Chen, L. Wang, N. Najimudin, J. a Saito, and M. Alam, “Draft genome sequence of the rubber tree Hevea brasiliensis.,” *BMC Genomics*, vol. 14, p. 75, 2013.

[12] J. Garcia-Mas, a. Benjak, W. Sanseverino, M. Bourgeois, G. Mir, V. M. Gonzalez, E. Henaff, F. Camara, L. Cozzuto, E. Lowy, T. Alioto, S. Capella-Gutierrez, J. Blanca, J. Canizares, P. Ziarsolo, D. Gonzalez-Ibeas, L. Rodriguez-Moreno, M. Droege, L. Du, M. Alvarez-Tejado, B. Lorente-Galdos, M. Mele, L. Yang, Y. Weng, a. Navarro, T. Marques-Bonet, M. a. Aranda, F. Nuez, B. Pico, T. Gabaldon, G. Roma, R. Guigo, J. M. Casacuberta, P. Arus, and P. Puigdomenech, “The genome of melon (Cucumis melo L.),” *Proc. Natl. Acad. Sci.*, vol. 109, no. 29, pp. 11872–11877, 2012.

[13] J. C. Dohm, A. E. Minoche, D. Holtgräwe, S. Capella-Gutiérrez, F. Zakrzewski, H. Tafer, O. Rupp, T. R. Sörensen, R. Stracke, R. Reinhardt, A. Goesmann, T. Kraft, B. Schulz, P. F. Stadler, T. Schmidt, T. Gabaldón, H. Lehrach, B. Weisshaar, and H. Himmelbauer, “The genome of the recently domesticated crop plant sugar beet (Beta vulgaris).,” *Nature*, vol. 505, no. 7484, pp. 546–9, 2014.

[14] K. F. X. Mayer, R. Waugh, P. Langridge, T. J. Close, R. P. Wise, A. Graner, T. Matsumoto, K. Sato, A. Schulman, G. J. Muehlbauer, N. Stein, R. Ariyadasa, D. Schulte, N. Poursarebani, R. Zhou, B. Steuernagel, M. Mascher, U. Scholz, B. Shi, P. Langridge, K. Madishetty, J. T. Svensson, P. Bhat, M. Moscou, J. Resnik, T. J. Close, G. J. Muehlbauer, P. Hedley, H. Liu, J. Morris, R. Waugh, Z. Frenkel, A. Korol, H. Bergès, A. Graner, N. Stein, B. Steuernagel, U. Scholz, S. Taudien, M. Felder, M. Groth, M. Platzer, N. Stein, B. Steuernagel, U. Scholz, A. Himmelbach, S. Taudien, M. Felder, M. Platzer, S. Lonardi, D. Duma, M. Alpert, F. Cordero, M. Beccuti, G. Ciardo, Y. Ma, S. Wanamaker, T. J. Close, N. Stein, F. Cattonaro, V. Vendramin, S. Scalabrin, S. Radovic, R. Wing, D. Schulte, B. Steuernagel, M. Morgante, N. Stein, R. Waugh, T. Nussbaumer, H. Gundlach, M. Martis, R. Ariyadasa, N. Poursarebani, B. Steuernagel, U. Scholz, R. P. Wise, J. Poland, N. Stein, K. F. X. Mayer, M. Spannagl, M. Pfeifer, H. Gundlach, K. F. X. Mayer, H. Gundlach, C. Moisy, J. Tanskanen, S. Scalabrin, A. Zuccolo, V. Vendramin, M. Morgante, K. F. X. Mayer, A. Schulman, M. Pfeifer, M. Spannagl, P. Hedley, J. Morris, J. Russell, A. Druka, D. Marshall, M. Bayer, D. Swarbreck, D. Sampath, S. Ayling, M. Febrer, M. Caccamo, T. Matsumoto, T. Tanaka, K. Sato, R. P. Wise, T. J. Close, S. Wannamaker, G. J. Muehlbauer, N. Stein, K. F. X. Mayer, R. Waugh, B. Steuernagel, T. Schmutzer, M. Mascher, U. Scholz, S. Taudien, M. Platzer, K. Sato, D. Marshall, M. Bayer, R. Waugh, N. Stein, K. F. X. Mayer, R. Waugh, J. W. S. Brown, A. Schulman, P. Langridge, M. Platzer, G. B. Fincher, G. J. Muehlbauer, K. Sato, T. J. Close, R. P. Wise, and N. Stein, “A physical, genetic and functional sequence assembly of the barley genome,” *Nature*, vol. 491, no. 7426, pp. 711–716, 2012.

[15] B. TA, “Genomes,” in *Genomes*, Second Edi., Oxford: Wiley-Liss, 2002.

[16] S. Rounsley, P. R. Marri, Y. Yu, R. He, N. Sisneros, J. L. Goicoechea, S. J. Lee, A. Angelova, D. Kudrna, M. Luo, J. Affourtit, B. Desany, J. Knight, F. Niazi, M. Egholm, and R. a. Wing, “De novo next generation sequencing of plant genomes,” *Rice*, vol. 2, no. 1, pp. 35–43, 2009.

[17] N. L. Quinn, N. Levenkova, W. Chow, P. Bouffard, K. a Boroevich, J. R. Knight, T. P. Jarvie, K. P. Lubieniecki, B. a Desany, B. F. Koop, T. T. Harkins, and W. S. Davidson, “Assessing the feasibility of GS FLX Pyrosequencing for sequencing the Atlantic salmon genome.,” *BMC Genomics*, vol. 9, p. 404, 2008.

[18] V. M. González, A. Benjak, E. M. Hénaff, G. Mir, J. M. Casacuberta, J. Garcia-Mas, and P. Puigdomènech, “Sequencing of 6.7 Mb of the melon genome using a BAC pooling strategy.,” *BMC Plant Biol.*, vol. 10, no. 1, p. 246, 2010.

[19] “SimSeq.” [Online]. Available: https://github.com/jstjohn/SimSeq.

[20] D. Earl, K. Bradnam, J. St. John, A. Darling, D. Lin, J. Fass, H. O. K. Yu, V. Buffalo, D. R. Zerbino, M. Diekhans, N. Nguyen, P. N. Ariyaratne, W. K. Sung, Z. Ning, M. Haimel, J. T. Simpson, N. a. Fonseca, I. Birol, T. R. Docking, I. Y. Ho, D. S. Rokhsar, R. Chikhi, D. Lavenier, G. Chapuis, D. Naquin, N. Maillet, M. C. Schatz, D. R. Kelley, A. M. Phillippy, S. Koren, S. P. Yang, W. Wu, W. C. Chou, A. Srivastava, T. I. Shaw, J. G. Ruby, P. Skewes-Cox, M. Betegon, M. T. Dimon, V. Solovyev, I. Seledtsov, P. Kosarev, D. Vorobyev, R. Ramirez-Gonzalez, R. Leggett, D. MacLean, F. Xia, R. Luo, Z. Li, Y. Xie, B. Liu, S. Gnerre, I. MacCallum, D. Przybylski, F. J. Ribeiro, T. Sharpe, G. Hall, P. J. Kersey, R. Durbin, S. D. Jackman, J. a. Chapman, X. Huang, J. L. DeRisi, M. Caccamo, Y. Li, D. B. Jaffe, R. E. Green, D. Haussler, I. Korf, and B. Paten, “Assemblathon 1: A competitive assessment of de novo short read assembly methods,” *Genome Res.*, vol. 21, no. 12, pp. 2224–2241, 2011.

[21] R. Li, H. Zhu, J. Ruan, W. Qian, X. Fang, Z. Shi, Y. Li, S. Li, G. Shan, K. Kristiansen, S. Li, H. Yang, J. Wang, and J. Wang, “De novo assembly of human genomes with massively parallel short read sequencing,” *Genome Res.*, vol. 20, no. 2, pp. 265–272, 2010.

[22] P. Green, “cross_match.” [Online]. Available: http://www.phrap.org.

[23] J. Wang, B. Roe, S. Macmil, Q. Yu, J. E. Murray, H. Tang, C. Chen, F. Najar, G. Wiley, J. Bowers, M.-A. Van Sluys, D. S. Rokhsar, M. E. Hudson, S. P. Moose, A. H. Paterson, and R. Ming, “Microcollinearity between autopolyploid sugarcane and diploid sorghum genomes.,” *BMC Genomics*, vol. 11, p. 261, 2010.

[24] F. Van Nieuwerburgh, R. C. Thompson, J. Ledesma, D. Deforce, T. Gaasterland, P. Ordoukhanian, and S. R. Head, “Illumina mate-paired DNA sequencing-library preparation using Cre-Lox recombination,” *Nucleic Acids Res.*, vol. 40, no. 3, pp. 1–8, 2012.

[25] S. Koren, M. C. Schatz, B. P. Walenz, J. Martin, J. T. Howard, G. Ganapathy, Z. Wang, D. a Rasko, W. R. McCombie, E. D. Jarvis, and A. M. Phillippy, “Hybrid error correction and de novo assembly of single-molecule sequencing reads,” *Nat. Biotechnol.*, vol. 30, no. 7, pp. 693–700, 2012.

[26] M. Boetzer and W. Pirovano, “SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information.,” *BMC Bioinformatics*, vol. 15, no. 1, p. 211, 2014.

[27] A. Bashir, A. a Klammer, W. P. Robins, C.-S. Chin, D. Webster, E. Paxinos, D. Hsu, M. Ashby, S. Wang, P. Peluso, R. Sebra, J. Sorenson, J. Bullard, J. Yen, M. Valdovino, E. Mollova, K. Luong, S. Lin, B. LaMay, A. Joshi, L. Rowe, M. Frace, C. L. Tarr, M. Turnsek, B. M. Davis, A. Kasarskis, J. J. Mekalanos, M. K. Waldor, and E. E. Schadt, “A hybrid approach for the automated finishing of bacterial genomes,” *Nat. Biotechnol.*, vol. 30, no. 7, pp. 701–707, 2012.

[28] N. de Setta, C. B. Monteiro-Vitorello, C. J. Metcalfe, G. M. Q. Cruz, L. E. Del Bem, R. Vicentini, F. T. S. Nogueira, R. A. Campos, S. L. Nunes, P. C. G. Turrini, A. P. Vieira, E. A. Ochoa Cruz, T. C. S. Corrêa, C. T. Hotta, A. de Mello Varani, S. Vautrin, A. S. da Trindade, M. de Mendonça Vilela, C. G. Lembke, P. M. Sato, R. F. de Andrade, M. Y. Nishiyama, C. B. Cardoso-Silva, K. C. Scortecci, A. A. F. Garcia, M. S. Carneiro, C. Kim, A. H. Paterson, H. Bergès, A. D’Hont, A. P. de Souza, G. M. Souza, M. Vincentz, J. P. Kitajima, and M.-A. Van Sluys, “Building the sugarcane genome for biotechnology and identifying evolutionary trends.,” *BMC Genomics*, vol. 15, no. 1, p. 540, 2014.

ANEXOS

COORDENADORIA DE PÓS-GRADUAÇÃO
INSTITUTO DE BIOLOGIA
Universidade Estadual de Campinas
Caixa Postal 6109. 13083-970, Campinas, SP, Brasil
Fone (19) 3521-6378. email: cpgib@unicamp.br

UNICAMP

IB

DECLARAÇÃO

Em observância ao **§4º do Artigo 1º da Informação CCPG-UNICAMP/002/13**, de 14/08/2013, referente a Bioética e Biossegurança, declaro que o conteúdo de minha Tese de Doutorado, intitulada "***Sintenia genomica entre sorgo e cana-de-açúcar inferida a partir do sequenciamento de um pool de BACs***", desenvolvida no Programa de Pós-Graduação em Genética e Biologia Molecular do Instituto de Biologia da Unicamp, não versa sobre pesquisa envolvendo seres humanos, animais ou temas afetos a Biossegurança.

Assinatura: _____
Nome do(a) aluno(a): Vagner Katsumi Okura

Assinatura: _____
Nome do(a) orientador(a): Paulo Arruda

Data: 20/07/2015

Profa. Dra. Rachel Meneguello
Presidente
Comissão Central de Pós-Graduação
Declaração

As cópias de artigos de minha autoria ou de minha co-autoria, já publicados ou submetidos para publicação em revistas científicas ou anais de congressos sujeitos a arbitragem, que constam da minha Dissertação/Tese de Mestrado/Doutorado, intitulada **Sintenia genomica entre sorgo e cana-de-açúcar inferida a partir do sequenciamento de um pool de BACs**, não infringem os dispositivos da Lei n.° 9.610/98, nem o direito autoral de qualquer editora.

Campinas, 20 de Julho de 2015.

Assinatura : _____
Nome do(a) autor(a): **Vagner Katsumi Okura**
RG n.° 4598958-5

Assinatura : _____
Nome do(a) orientador(a): **Paulo Arruda**
RG n.° 6642720