

## Survey sequencing of soybean elucidates the genome structure, composition and identifies novel repeats

Andrew Nunberg<sup>A</sup>, Joseph A. Bedell<sup>A</sup>, Mohammad A. Budiman<sup>A</sup>, Robert W. Citek<sup>A</sup>, Sandra W. Clifton<sup>B</sup>, Lucinda Fulton<sup>B</sup>, Deana Pape<sup>B</sup>, Zheng Cai<sup>C</sup>, Trupti Joshi<sup>C,D</sup>, Henry Nguyen<sup>E,F</sup>, Dong Xu<sup>C,D,E</sup> and Gary Stacey<sup>D,E,F,G,H</sup>

<sup>A</sup>Orion Genomics, LLC, 4041 Forest Park Ave, St Louis, MO 63108, USA.

<sup>B</sup>Genome Sequencing Center, School of Medicine, Washington University, St Louis, MO 63130, USA.

<sup>C</sup>Computer Science Department, University of Missouri, Columbia, MO 65211, USA.

<sup>D</sup>Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO 65211, USA.

<sup>E</sup>National Center for Soybean Biotechnology, University of Missouri, Columbia, MO 65211, USA.

<sup>F</sup>Division of Plant Science, University of Missouri, Columbia, MO 65211, USA.

<sup>G</sup>Division of Biochemistry, Department of Molecular Microbiology and Immunology, University of Missouri, Columbia, MO 65211, USA.

<sup>H</sup>Corresponding author. Email: staceyg@missouri.edu

*This paper originates from a presentation at the Third International Conference on Legume Genomics and Genetics, Brisbane, Queensland, Australia, April 2006.*

**Abstract.** In order to expand our knowledge of the soybean genome and to create a useful DNA repeat sequence database, over 24 000 DNA fragments from a soybean [*Glycine max* (L.) Merr.] cv. Williams 82 genomic shotgun library were sequenced. Additional sequences came from over 29 000 bacterial artificial chromosome (BAC) end sequences derived from a BstI library of the cv. Williams 82 genome. Analysis of these sequences identified 348 different DNA repeats, many of which appear to be novel. To extend the utility of the work, a pilot study was also conducted using methylation filtration to estimate the hypomethylated, soybean gene space. A comparison between 8366 sequences obtained from a filtered library and 23 788 from an unfiltered library indicate a gene-enrichment of ~3.2-fold in the hypomethylated sequences. Given the 1.1-Gb soybean genome, our analysis predicts a ~343-Mb hypomethylated, gene-rich space.

### Introduction

Soybean (*Glycine max*) is the most valuable legume crop, with numerous nutritional and industrial uses because of its unique seed chemical position. Over 85 million metric tonnes of soybeans were produced in the US on >30 million ha in 2004, with an estimated annual economic value exceeding US \$17 billion, second only to maize, and approximately twice that of wheat and ten times that of rice (<http://www.nass.usda.gov/index.asp>; verified 13 June 2006). While genomics is having a profound effect on plant biology, its direct impact on major crop species such as soybean remains limited. A primary difficulty is that the genomes of major crop species are large and complex, being as much as 50 times larger than that of the model plant

*Arabidopsis thaliana*. For example, the soybean genome at 1.12 billion bp (Arumuganathan and Earle 1991) remains a formidable challenge for genome sequencing using current technologies. Nonetheless, the scientific community clearly sees the need for sequencing of crop genomes. Indeed, the legume community recently chose soybean as the reference species for the phaseoloid legumes, which comprise many of the major legume crops, and recommended sequencing the soybean genome (Gepts *et al.* 2005).

Soybean is a paleopolyploid with  $2n = 40$  (Goldblatt 1981). Therefore, it is expected that any given gene will be present approximately four times in the genome. Indeed, Shoemaker *et al.* (1996) found that 60% of the time, hybridisation of random clones to soybean genomic

Abbreviations used: BAC, bacterial artificial chromosome; EST, expressed sequence tag; FISH, fluorescent *in situ* hybridisation; FP, filter power; gDNA, genomic DNA; Mb, megabases; MF, methylation filtration, methylation filtering; miRNA, microRNA; SSR, simple sequence repeats; UF, unfiltered; WGS, whole-genome shotgun.

DNA resulted in three or more detectable bands, with over 90% detecting more than two bands. The soybean genome also exhibits evidence of two duplication events, one ~15 million years ago and another ~40 million years ago (Schlueter *et al.* 2004; Shoemaker *et al.* 2006). The latter likely predates the division of the galeoid and phaseolid lineages.

DNA–DNA renaturation studies suggested that ~40–60% of the soybean genome sequence is repetitive (Goldberg 1978; Gurley *et al.* 1979). Recently, Lin *et al.* (2005) used fluorescent *in situ* hybridisation (FISH) to explore the distribution of repetitive sequences in the soybean genome. These data supported the notion that the soybean repeats are largely localised to the pericentromeric regions resulting in largely euchromatic chromosome arms. Digestion of repetitive sequences with a methylation-sensitive enzyme (*HpaII*) suggested the centromeric repeats were methylated. These results are consistent with earlier studies. For example, one study, analysing the nature of BAC-end sequences, suggested that selected clones were either repeat-rich or gene-rich but some interspersions of such sequences was also found (Marek *et al.* 2001). A complete karyotype of soybean has been reported based upon pachytene analysis (Singh and Hymowitz 1988). This analysis indicated that the soybean genome was ~35% heterochromatic but several of the chromosomes had largely euchromatic arms.

Relatively few specific soybean repetitive sequences have been reported. A 120-bp soybean repeat (STR120) was identified by Morgante *et al.* (1997) and estimated to exist in 5000–10 000 copies. Lin *et al.* (2005) identified the STR102 repeat (102 bp) with 82.6% sequence similarity to STR120. Vahedian *et al.* (1995) used FISH to localise the 92-bp SB92 repeat to four or five locations within the genome, including two associated with the centromere. Estimates suggested that this one repeat represented ~0.9% of the genome (~1 × 10<sup>5</sup> bp). This high copy number but limited distribution suggests that the SB92 repeat must be localised in megabase-sized regions. Lin *et al.* (2005) found the STR102 repeat in clusters up to ~435.6 kb in length.

Transposable DNAs comprise a significant proportion of the repetitive DNA found in eukaryotic genomes. Vodkin *et al.* (1983) identified the first transposable element in soybean (termed *Tgm*). Seven different classes of the *Tgm* transposon were identified ranging in size from 1.6 to >12 kb (Rhodes and Vodkin 1988); however, the copy number of this transposon was not determined. A mariner-like transposon (*Soymar1*) was estimated to be present up to ~10 000 copies per haploid genome, with the largest member of this class being 3.5 kb (Jarvik and Lark 1998). Soybean retrotransposons include the *gypsy*/*Ty3*-like retroelement, *Calypso* (Wright and Voytas 2002), similar to the *Arabidopsis* element *Athila4* found in the centromere, and a *copial*/*Ty1*-like retroelement (SIRE-1, Laten and Morris 1993). These retroelements vary in size from 11 to 14 kb and are duplicated

a few hundred times in the soybean genome. Lin *et al.* (2005) found both SIRE1 and *Calypso*-like elements in soybean BAC clones mapping to centromeres.

Compared with crop plants, the *Arabidopsis* genome is relatively gene-rich with an average gene density of 20–25 genes per 100 kb. These genes are relatively evenly distributed across the genome (Barakat *et al.* 1998). In contrast, estimates for large cereal genomes suggest that the gene-rich portion may only comprise 10–20% of the genome. Fewer studies have been done with soybean. DNA methylation in plants is associated with silent, heterochromatic DNA that in large cereal genomes is rich in transposable elements (Bennetzen 1996). Marek *et al.* (2001) compared 2000 soybean BAC-end sequences either selected based on RFLPs from methylation-sensitive restriction enzymes or based on the presence of SSRs (simple sequence repeats). The conclusions of this study were that the RFLP-selected BACs had 50% less repetitive sequences with a similar enrichment in genic sequences. Other data also suggest that hypomethylated, gene-rich regions exist in the soybean genome. For example, using a similar strategy, Mudge *et al.* (2004) selected BAC clones using *PstI*-generated RFLP probes. *PstI* is a methylation-sensitive enzyme and restriction sites would be expected in regions of hypomethylation. The probes used in this study identified BAC clones from only 24% of the genome, giving an estimate of the gene space of ~264 Mbp (i.e. 1.1 billion bp × 0.24). Only a few soybean BAC clones have been completely sequenced. Analysis of one such region (330 kb) revealed sequence containing few repetitive sequences and a gene density of ~1 gene per 5 kb (Foster-Hartnett *et al.* 2002). It is clear that more needs to be known about the soybean genome before an effective strategy for genomic sequencing can be developed. Collectively, the current data suggest a soybean genome containing hypomethylated, gene-rich segments with the hypermethylated DNA regions largely confined to islands of repetitive sequence found in pericentromeric regions.

In order to expand our knowledge of the soybean genome and to develop a useful DNA repeat sequence database, we sequenced over 24 000 DNA fragments from a shotgun genomic library of soybean cv. Williams 82. Also included in our analysis were over 29 000 BAC-end sequences from a *Bst*I library of Williams 82 DNA (J Tomkins unpubl. data). This cultivar has been chosen as the primary model by the soybean community (Stacey *et al.* 2004). In order to derive more information from this effort and to explore a possible means for determining the genic sequence of the soybean genome, we also conducted a pilot study using methylation filtration (MF). This is a very simple and robust method for enriching for hypomethylated, gene-rich sequences in complex plant genomes (Rabinowicz *et al.* 1999). Briefly, genomic DNA is used to transform an *E. coli* strain that preferentially cleaves methylated

DNA sequences. Consequently, only hypomethylated DNA inserts 'survive' the cloning process. The MF strategy greatly reduces the time and cost of gene identification in plants by filtering out methylated repetitive elements while retaining hypomethylated gene fragments. MF has been successfully applied to genomes of more than a dozen plant species across the plant kingdom including monocots, dicots, gymnosperms, and even moss, a non-vascular plant (Rabinowicz *et al.* 2005). The data from soybean suggests that the MF method results in a ~3.2-fold gene-enrichment of an estimated ~343-Mb hypomethylated gene space.

## Materials and methods

### *GeneThresher*<sup>®</sup> library construction

Nuclear DNA was obtained from 4–6-week-old greenhouse-grown soybean seedlings from<sup>®</sup> *Glycine max* (L.) Merr. cv. Williams 82. Shearing of nuclear DNA was performed using either a nebuliser (Cis-U.S., Inc., Bedford, MA) or Hydroshear (GeneMachines, San Carlos, CA). Sheared fragments were end-repaired with a variety of enzymes including Mungbean Nuclease, T4 DNA Polymerase, Klenow fragment, and T4 Polynucleotide kinase. End-repaired fragments were size-selected on an agarose gel and DNA fragments ranging from 0.7 to 1.5 kb were extracted and ligated to dephosphorylated, *Bst*XI-digested pOT2 vector which was used to construct both methylation filtered (MF, GeneThresher<sup>TM</sup>) and unfiltered (UF) libraries. Ligation reactions were transformed into McrBC+ and McrBC– strains of *Escherichia coli* for generation of filtered and unfiltered libraries, respectively. Recombinant clones were picked using a Genetix Q-bot robot (Research Genetics, Carlsbad, CA) and stored individually in 384-well microtitre plates.

### DNA sequencing

The subclone libraries were quality tested, and, once passed, they entered the production-sequencing queue at Genome Sequencing Center, Washington University. Libraries were plated, and colonies resulting after overnight growth were harvested by robotic pickers (Q-Pix) that array subclones in 384-well microtitre trays. These trays hold glycerol-containing media and provide both a source of DNA for sequencing as well as a subclone archive. Subclone DNAs were purified using a robotic assembly line and paramagnetic particle-based separation technology (CCS Packard, Inc., Torrance, CA; Hawkins *et al.* 1987; Clifton *et al.* 2004). ABI DyeTerminator (Applied Biosystems, Foster City, CA) reactions were used for sequencing. The reactions were performed in a 384-well format and were assembled using a BiomekFX robot. Thermocycling reaction times were as previously reported (Lander *et al.* 2001). The reaction products were loaded onto 3730xl DNA sequencers, and as sequence runs were completed, data were automatically processed and recorded in an Oracle database. For all of the above activities, sample processing and tracking were facilitated by a bar-code system that also is linked to the Oracle database. The reads were base called using the ABI KB software. Traces were submitted to the Trace Archive division of GenBank, and the reads were sent to the GSS section of GenBank; accession numbers are noted below.

### Database curation and filter power calculation

A first pass definition-line curation of publicly available sequence databases was done to eliminate obvious transposon sequences that would hamper subsequent analyses by virtue of inflating the true 'gene' content of the given database. The *Arabidopsis* protein set, which was used for the gene enrichment calculations and assessment

of cross-genome annotation potential, was downloaded from the NCBI ([ftp://ftp.ncbi.nih.gov/genomes/Arabidopsis\\_thaliana/CHR\\_\\*/\\*.faa](ftp://ftp.ncbi.nih.gov/genomes/Arabidopsis_thaliana/CHR_*/*.faa)). The files were dated 23 May 2003 and contained 28 581 sequences (12 112 846 total letters). Repeats were removed from this dataset if the definition line met both of the following two criteria:

- (1) matched the case-insensitive regular expression `/retro|mutator|transpos|reverse transcriptase|polyprotein|bgag\b|BARE-1|athila/`
- (2) did not match `^\[.*retro.*\]|leucine|WD-repeat|WD repeat|WD40|WD-40|ankyrin|telomere|arm repeat|PPR-repeat|armadillo|tetratricopeptide|TPR-repeat|TPR repeat|Kelch|pentapeptide|C-repeat/`.

This second step was used to replace falsely identified non-repetitive elements. Removing repeats reduced the database size by 640 sequences to 27 941, which included 4412 sequences identified as hypothetical by matching the definition line to the case-insensitive regular expression `/hypothetical/`.

Gene enrichment was calculated by comparing the rate of gene discovery between methylation-filtered sequences and unfiltered sequences. To ensure high quality, unique sampling events, reads were chosen that contained at least 100 contiguous Phred Q20 bases and only one read per clone. Detection of genes was accomplished by an NCBI-BLASTX (Parameters: `-e 0.01 -b 5 -v 5`) search of the curated *Arabidopsis* protein database (Methods). Aside from the curation of the *Arabidopsis* database to remove repetitive elements, matches to proteins annotated as hypothetical were not counted. Hypothetical genes are often false gene predictions or unknown repetitive elements. In order to calculate a gene enrichment factor, or Filter Power (FP), the proportion of matches from MF sequences are compared to the proportion of matches in UF sequences over a range of Expectation values (E-values) from  $1e-5$  to  $1e-20$ , such that all matches better than the given E-value are tabulated (Table 1). For soybean, the genome size is estimated at 1.1 Mb (Arumuganathan and Earle 1991). Dividing the genome size by the median 3.2 FP provides an estimate of a 343-Mb sampled space.

### Collapsing read pairs

Read pairs from each nuclear clone were assembled using Phrap (`-minmatch 17 -minscore 40 -forcelevel 1`). The resulting contigs were trimmed for quality and vector. Read pairs that did not collapse were trimmed for sequence and length. These sequence sets are hereto referred to as the nuclear collapsed set.

### Repeat analysis with RepeatMasker

Repetitive elements were identified in the nuclear collapsed set for filtered and unfiltered sequences using RepeatMasker (AFA Smit, R Hubley, P Green RepeatMasker at <http://repeatmasker.org>; verified 6 June 2006) with the MaskerAid speed enhancement (Bedell *et al.* 2000) and a collection of plant repeat databases from TIGR. Version 2 of the Brassicaceae, Fabaceae, Solanaceae repeat libraries were downloaded from TIGR. In addition, the TIGR cereal repeat database, dated 11 July 2003 was downloaded from [ftp://ftp.tigr.org/pub/data/TIGR\\_Plant\\_Repeats/](ftp://ftp.tigr.org/pub/data/TIGR_Plant_Repeats/) and contained 11 043 repeat entries. This collection of plant repeats was supplemented with the soybean sequences of the retrotransposon Calypso (AF186182–AF186186, AF378062–AF378073), repeat SB92(U11026), and repeat STR120 (U26697, U26698, U26700, U26701). RepeatMasker was run with the following parameters: `-s -w -no-is -nolow`.

### De novo repetitive sequence identification

De novo repetitive sequence identification was performed on the RepeatMasker masked nuclear collapsed sets for filtered and unfiltered sequences (see above). The masked sequences were split on 20 or more masked nucleotides (N's) and each segment was given

**Table 1. Distribution of soybean sequence repeats (%)**

Genomic sequences for the various organisms were analysed for repeat content with RepeatMasker. The fraction for each repeat class is shown. The 'mixed' classes are those repeat features that are made up of overlapping sequences of two or more sub-classes. WGS, soybean whole-genome shotgun sequences; BAC ends, soybean sequences derived from BAC ends

Class	Sub-class	<i>Glycine max</i> WGS	<i>Glycine max</i> BAC ends	<i>Lotus japonicus</i>	<i>Medicago truncatula</i>	<i>Arabidopsis thaliana</i>
Retrotransposons						
	Ty1-copia	3.44	4.86	1.67	0.78	0.43
	Ty3-gypsy	0.49	0.39	0.44	0.25	0.20
	Calypso	8.87	4.89	0.17	0.23	0.07
	Unclassified	2.15	3.00	2.09	4.19	2.23
	LINE	0.06	0.06	0.05	0.06	0.11
Transposons						
		0.61	0.63	0.30	0.70	0.60
Mites						
		0.10	0.07	0.04	0.38	0.02
Ribosomal						
		2.95	1.06	1.85	0.14	0.11
Centromeric						
		0.14	0.07	0.05	0.02	0.50
Telomeric						
		0.04	0.01	0.03	0.00	0.05
Unclassified						
		–	–	–	–	0.00
	SB92	0.25	0.03	0.00	0.00	0.00
	STR120	0.98	0.30	0.00	0.00	0.00
	Unclassified	0.96	0.38	0.45	0.17	0.43
	Soy novel	16.78	14.82	0.08	0.41	0.00
Mixed						
		4.43	3.85	1.05	1.71	1.07
Total						
		42.26	34.42	8.29	9.04	5.81

a unique identifier. The sequences were analysed with RECON (Bao and Eddy 2002a, b) for *de novo* repetitive sequence identification. The computation of RECON involved the following steps.

- (1) Run BLAST using each sequence against the database, with the expectation value threshold  $1e^{-30}$ .
- (2) Apply the MSPCollect tool of RECON, which converts each BLAST output into an MSP file. In the MSP file, each pairwise MSP result reported by BLAST is turned into a one-line summary in a certain format, containing variables of 'score', 'identical percentage', 'sequence start position of query', 'sequence end position of query', 'query name', 'sequence start position of subject', 'sequence end position of subject', and 'subject name'.
- (3) Filter MSP files through three steps: (a) remove repeat sequences that are less than 50 bp; (b) remove those sequences with frequency of less than 10; (c) remove more lines of self-hits (i.e. query sequence is the same as the subject name) in the MSP file.
- (4) Run RECON using the remaining MSP file.

RECON outputs all the repetitive sequences, but it does not provide a representative sequence for a group of sequences that are repetitive from each other. We then developed and ran a C-shell script that incorporates ClustalW (Thompson *et al.* 1994) for all the elements in a repeat family to obtain the aligned sequences. A consensus sequence was built for the family, where for each column in the alignment, we took the letter with largest occurring frequency in a column as its representative. This consensus sequence is considered to be the representative repeat sequence for this whole family.

#### Novel repeat copy number

Novel repeats were compared to the original nuclear collapsed dataset from unfiltered sequences using BLAST. Matches below 95% identity and under 50 bp were discarded. The total base pairs matched was tallied for each repeat and used to determine copy number.

## Results

### *Reducing the soybean genome to the hypomethylated, gene-rich space*

Methylation filtered libraries were constructed from soybean nuclear DNA in host strains of bacteria that restrict methylated DNA (Rabinowicz *et al.* 1999). Of 10 751 attempts from the MF library, 8632 were successful, 8366 of which were considered of nuclear origin based on comparison with chloroplast, mitochondrial, viral and bacterial databases. As a control and a tool for assessment of the whole genome composition, the same DNA ligations were propagated in host strains that do not restrict methylated DNA (unfiltered libraries, UF). The soybean whole genome shotgun, or unfiltered (UF), sequences comprise 26 108 attempts, of which 24 224 were considered successful and 23 788 were nuclear.

To calculate the genome space sampled by GeneThresher<sup>®</sup> technology, a method that relies on gene enrichment was used. The gene-enrichment method works by assuming that genes are enriched in the MF libraries proportional to the reduction in genome size. For example, if the genome is reduced by 3-fold, then gene discovery should occur 3-fold faster in MF *v.* whole-genome shotgun libraries.

The gene-enrichment factor is called Filter Power (FP) and FP can be used to derive the sampled genome space by dividing it into the size of the whole genome (G). We calculated the soybean FP using a subset of our filtered and unfiltered sequences compared with a curated database

of known genes over a range of BLAST E-values ( $1e^{-5}$  to  $1e^{-20}$ ). The FP is between 2.7 and 3.5 with a median value of 3.2. By dividing this range of FP values into the 1.1-Gb soybean genome, the sampled genome is estimated to be between 314 and 407 Mb with a median of 343 Mb (Fig. 1). The MF dataset consists of a nuclear coverage, after collapsing read pairs, of 3.66 Mb which is approximately a  $0.01 \times$  coverage of the sampled space.

#### Gene ontology analysis

To determine whether there is an apparent bias in the enrichment for genes using MF, an analysis of the gene ontology terms for sequences with significant similarity (with E-value less than  $1e^{-8}$ ) to an *Arabidopsis* protein (see Materials and methods) was performed for both the filtered and unfiltered sequences. As seen in Fig. 2, there was no significant difference between the genes enriched through methylation filtration (FGM) and through whole-genome shotgun (UGM).

#### Simple sequence repeats

Simple sequence repeats are stretches of DNA with simple sequence pattern repetitions, usually in the form of di-, tri-, or tetra-nucleotide expansions such as  $(CA)_n$ ,  $(CAG)_n$ , or  $(GATA)_n$ . These stretches of DNA are useful for genetic marker analysis because they are unstable and often are polymorphic between closely related individuals (Cordeiro *et al.* 2001; Klein *et al.* 2003). Overall, there was a higher

density of SSRs in methylation-filtered soybean sequences with one SSR per 10.0 kb in MF, compared with one SSR per 15.3 kb in UF. Additionally, the SSRs obtained from MF are more likely to be gene-associated.

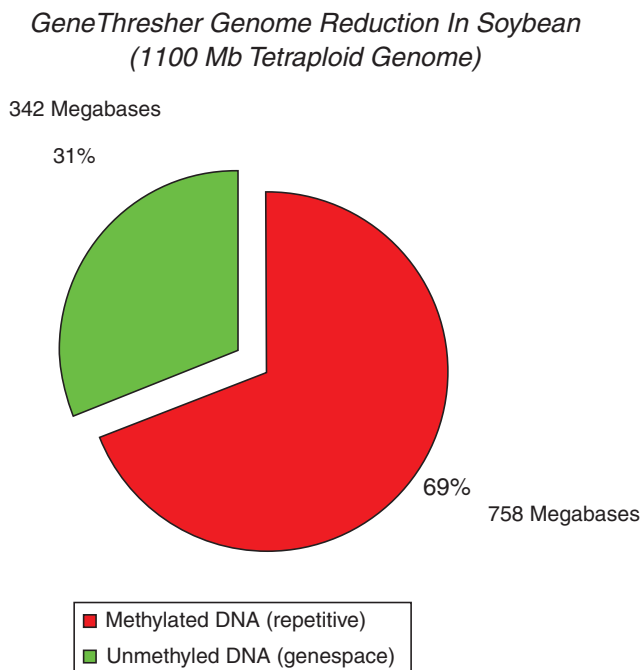
GC-rich trinucleotide SSRs (TNR) in monocots have been shown to be preferentially associated with coding regions (McCouch *et al.* 2002; Morgante *et al.* 2002). For soybean, there were 37 out of 739 (5%) GC-rich SSRs in UF v. 29 of 378 (7.7%) in MF, which is not a statistically significant difference.

#### Repetitive sequences in soybean

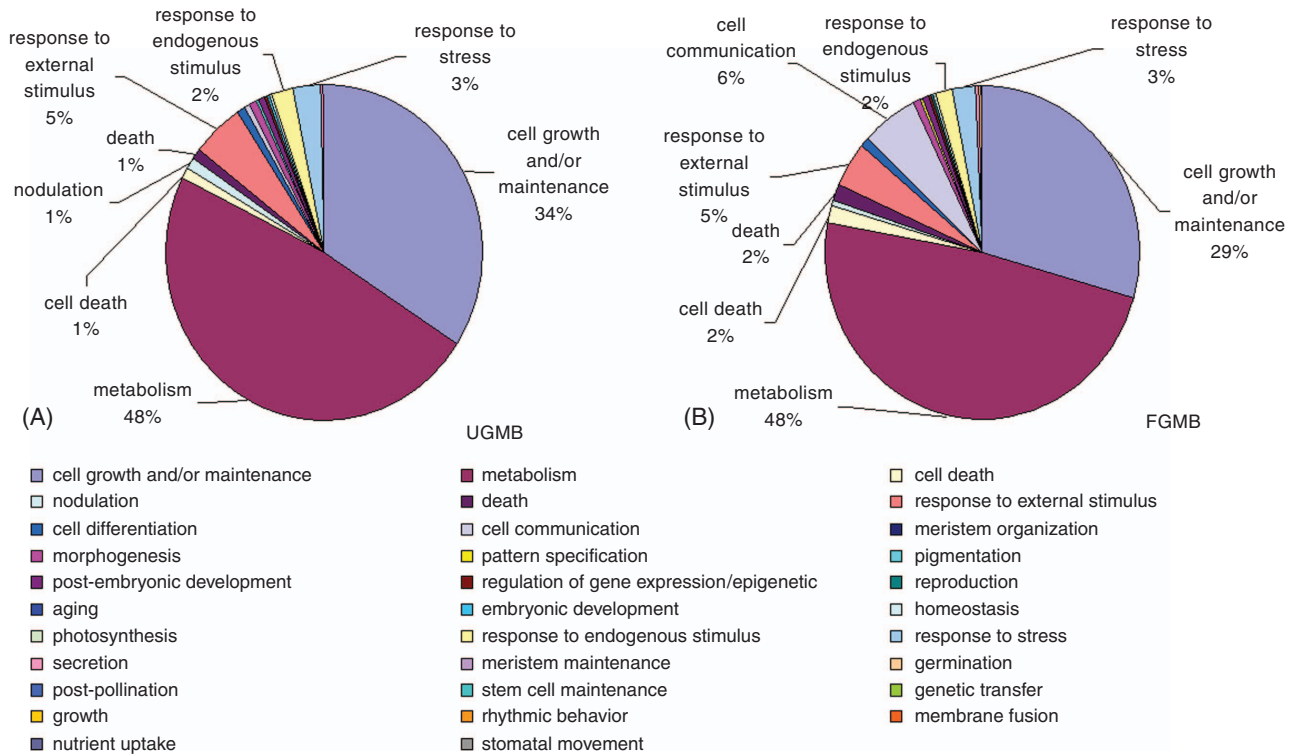
Estimates of the gene space by MF suggest that  $\sim 70\%$  of the soybean genome is composed of methylated DNA. To determine the extent to which the soybean genome is repetitive, the unfiltered soybean sequences were masked using a compilation of plant repeats from TIGR with RepeatMasker (Table 1) (see Materials and methods). As a basis for comparison, 20 000 GenBank GSS sequences of the legumes *Lotus japonicus* and *Medicago truncatula*, as well as 20 000 random reads from the *Arabidopsis thaliana* genome, were also masked after contaminating sequences were removed. The results are shown in Table 1. Excluding the novel soybean repeats (see below), the most abundant class of repeats for soybean were retrotransposons (14.97%) and this also appeared true for *M. truncatula* and *L. japonicus*. Counting all the known repeat sequences, at least 25% of the soybean genome is repetitive. In addition, we included 29 117 soybean BAC-end sequences (with GenBank accession numbers from CZ498303 to CZ527432) of soybean in our repeat analysis. The relative distribution of repeats found in the BAC-end sequences was roughly equivalent to that found in the whole-genome shotgun sequences (Table 1).

The reduction of repeat sequences was calculated by comparing the fraction of repeat-masked sequences from unfiltered and filtered libraries with respect to the whole genome. Thus, the calculation becomes: Fraction repeat-masked Unfiltered / (Fraction repeat-masked Filtered / Filter Power). MF significantly reduced the known repetitive fraction by  $\sim 12$ -fold (see Table 2). Reduction of various repeat classes was not uniform. For example, ribosomal repeats were reduced by more than 40-fold, while DNA transposons were reduced roughly 8-fold. Such differences may be, in part, due to differential methylation status of these repeat classes as was observed when MF was applied to maize (Whitelaw *et al.* 2003) and sorghum (Bedell *et al.* 2005). Differences in reduction may also be due to the GC content of different repeat classes.

In summary, 235 representative repeats were identified from the unfiltered and filtered libraries of soybean. After using them as a filter, we identified 113 additional representative repeats for the 29 117 BAC-end sequences of soybean, i.e. 348 representative repeats in total were identified for soybean. These repeats can be found at



**Fig. 1.** Genome reduction. Methylation filtration reduces the soybean genome by 69%, sampling a hypomethylated space of  $\sim 342$  Mb (green) and filtering out 758 Mb (red) of the 1100-Mb soybean genome.



**Fig. 2.** Gene ontology analysis. Gene ontology terms, with significant similarity ( $>1e-8$ ) to an *Arabidopsis* protein (see Materials and methods), derived from the predicted gene sequences were compared between the filtered and unfiltered sequences. There was no significant difference between the classes of genes enriched by methylation filtration (FGM) and through whole-genome shotgun (UGM).

**Table 2.** Frequency of repeat classes in filtered and unfiltered soybean libraries

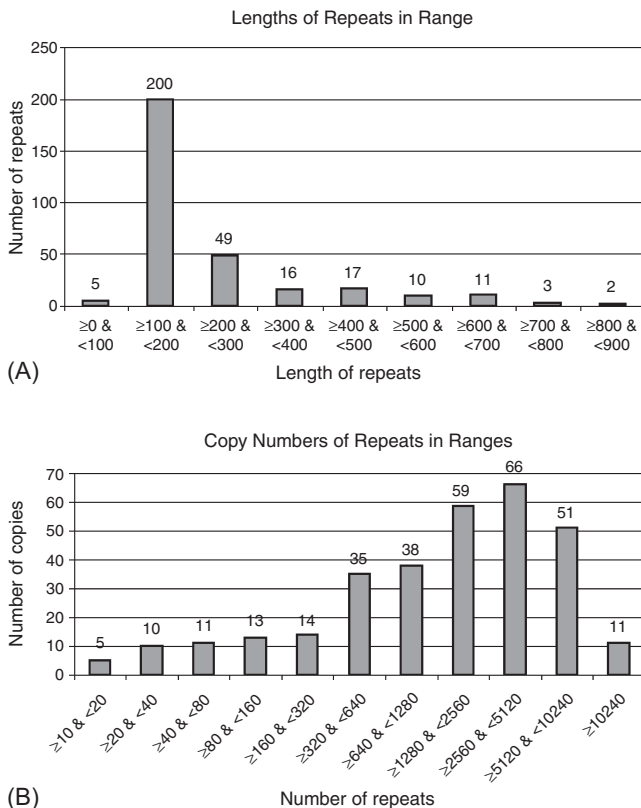
The reduction of repeats sequences was calculated by comparing the fraction of repeat-masked sequence from unfiltered and filtered libraries with respect to the whole genome

Class	Sub-class	Unfiltered soybean (%)	Filtered soybean (%)	Reduction factor
Retrotransposons	Ty1-copia	3.44	1.05	10.52
	Ty3-gypsy	0.49	0.36	4.3
	Calypso	8.87	1.23	23.16
	Unclassified	2.15	1.09	6.31
	LINE	0.06	0.02	8.49
Transposons		0.61	0.23	8.39
Mites		0.10	0.13	2.47
Ribosomal		2.95	0.21	44.71
Centromeric		0.14	0.13	3.32
Telomeric		0.04	0.08	1.61
Unclassified	SB92	0.25	0.02	40.76
	STR120	0.98	0.14	22.33
	unclassified	0.96	0.18	16.9
	Soy novel	16.78	5.18	10.36
Mixed		4.43	0.67	21.15
Total		42.26	10.73	12.6

<http://www.soybeanome.org/> (verified 6 June 2006). The distribution of the repeats v. their length is shown in Fig. 3A, while a histogram for the number of copies of repeats within the dataset is shown in Fig. 3B. It is worthwhile mentioning that some of the identified repeats are very long; 26 of them are longer than 500 bp. Assuming that our sampling was random, the relative copy number shown should be representative of the whole genome. This analysis estimated that repetitive sequences make up  $\sim 10\%$  of the genespace of soybean (data not shown).

#### Novel repeats in soybean

Results in Table 1 indicate that only approximately 5% of the *Arabidopsis* genome contains repetitive DNA. This value does not agree with other estimates of the repeat content of *Arabidopsis* (The Arabidopsis Genome Initiative 2000). This under-estimated value of repeat content is due to the lack of curated repeats. The TIGR Fabaceae repeat library contains 404 sequences of which only 14 are curated retrotransposons, the most abundant class of known repeats in soybean (see above). To further curate repeats in soybean, a *de novo* approach to repeat identification was taken using unfiltered, filtered, combined filtered and unfiltered soybean sequences, and BAC-end sequences (see Materials and



**Fig. 3.** Statistical analysis of the 383 identified repeats. (A) The distribution of the number of repeats v. repeat lengths. (B) A histogram for the number of repeats v. number of copies.

methods). Because it appears soybean has undergone at least two genome duplications (Schlueter *et al.* 2004), each gene may have four paralogs. To avoid such a trivial case, we set the cut off for repetitive sequences to a minimum of six members.

Putative repeats were screened for coding regions by comparison to an *Arabidopsis* protein database. They were also screened for known repeats with RepeatMasker. Of the 348 repeats derived from this analysis that were not coding for a known gene, two (<1%) can be classified as similar to a known repeat (with E-value less than  $1e^{-8}$ ), leaving 346 (>99%) as novel, unclassified repeats. Using these repeats with RepeatMasker, 16.78% of the unfiltered soybean sequence was determined to contain these repeats and 5.18% of the filtered soybean sequence. These novel sequences appear to be fairly specific for soybean. As seen in Table 1, only 0.08 and 0.41% of *L. japonicus* and *M. truncatula* sequences, respectively, were masked. Of the novel sequences, only 15 masked *L. japonicus* and 21 masked *M. truncatula*.

## Discussion

The soybean genome, like most important crop genomes, is large and complex with a high proportion of repetitive

elements (Stacey *et al.* 2004). Our results confirm this composition of the genome and also extend the analysis to the epigenetic component. The 1.1-Gb genome contains 40% identifiable repetitive elements with a hypomethylated gene space of ~340 Mb. The hypomethylated fraction contains relatively few repeats and is enriched for genes by more than 3-fold.

We have identified a total of 348 repetitive elements by analysis of whole-genome shotgun (WGS) and BAC-end sequences. This represents the only public repeat database available for use in masking repeats during analysis of soybean genomic sequence (Holmes 2002). Given that the TIGR Fabaceae repeat library is relatively small (404) and only a small fraction (14 of 404) represent retrotransposons, these sequences represent a significant contribution to our knowledge of legume DNA repeats. A significant fraction (~17%) appears to be soybean specific, as they are not found in *Medicago truncatula*, *Lotus japonicus*, or *Arabidopsis*. Methylation filtration is clearly an effective method for removing repetitive DNA. Analysis of the repeats found in the filtered libraries estimated that only around 10% of the hypomethylated, gene-rich segments of the genome are repetitive.

Early DNA–DNA renaturation studies suggested that 40–60% of the soybean genome sequence is repetitive (Goldberg 1978; Gurley *et al.* 1979). It has also been shown that more than 35% of the genome is made up of heterochromatin (Singh and Hymowitz 1988). Our estimate that the soybean gene space is ~342 Mb (31%) is roughly consistent with these earlier estimates. Additionally, analysis of eight gene-rich soybean BACs deposited in GenBank indicates that they are remarkably free of identifiable repeats. Using the most current plant repeat database from TIGR, supplemented with our current set of newly identified soybean-specific repeats, these eight BACs (i.e. GenBank Accessions AC166090, AC152056, AC166330, AC166092, AC166091, AC144537, AY262686, AF541963) average less than 5% repetitive elements with a range of 1.8–8.7%. This suggests that ~95% of the euchromatic BAC DNA will be accessible to MF clones. Therefore, the MF clones should supplement euchromatic, gene-rich BAC sequencing to a similar extent as WGS, allowing very extensive assemblies across the length of the BAC, similar to what was done to finish the rat genome (Gibbs *et al.* 2004). In maize, skim sequencing from BAC clones at less than 1× coverage combined with a deep coverage through gene enrichment is predicted to generate a high quality sequence map for a fraction of the cost of whole-genome sequencing (Martienssen *et al.* 2004). This will be even more efficient in soybean, given the paucity of repetitive elements in the euchromatic, gene-rich portion.

For the method development, we improved the procedure of RECON for repeat identification. In particular, we applied

ClustalW to systematically retrieve consensus sequences from the multiple alignment of potential repetitive sequences. Such an approach allowed us to more accurately identify repeat representatives. Nevertheless, it should be noted that our procedure based on RECON cannot guarantee to identify all possible repeats. Some repeats with a weak pattern may be missed.

### Accession numbers

The soybean GeneThresher<sup>®</sup> sequences are deposited in the Genome Survey Sequence (GSS) division of GenBank under accessions: CL868625–CL874567, CL874569–CL876819, CL876829–CL877015, CL884614. The unfiltered sequences are deposited under accessions CL876820–CL876828, CL877016–CL884613, CL884615–CL900625. Soybean BAC-end sequences are deposited in the Genome Survey Sequence (GSS) division of GenBank under accessions: CZ498303–CZ527432.

### Acknowledgments

Research was supported by a grant (to GS and SWC) from the National Science Foundation (grant DBI-0417357) and by the Missouri Soybean Merchandising Council.

### References

- Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815. doi: 10.1038/35048692
- Arumuganathan K, Earle ED (1991) Estimation of nuclear DNA content of plants by flow cytometry. *Plant Molecular Biology Reporter* **9**, 229–241.
- Bao Z, Eddy SR (2002a) Automated *de novo* identification of repeat sequence families in sequenced genomes. *Genome Research* **12**, 1269–1276. doi: 10.1101/gr.88502
- Bao Z, Eddy SR (2002b) RECON 00README — finding repeat families from biological sequences Version 1.03.
- Barakat A, Matassi G, Bernardi G (1998) Distribution of genes in the genome of *Arabidopsis thaliana* and its implications for the genome organization of plants. *Proceedings of the National Academy of Sciences USA* **95**, 10 044–10 049. doi: 10.1073/pnas.95.17.10044
- Bedell JA, Korf I, Gish W (2000) MaskerAid: a performance enhancement to RepeatMasker. *Bioinformatics* **16**, 1040–1041. doi: 10.1093/bioinformatics/16.11.1040
- Bedell JA, Budiman MA, Nunberg A, Citek RW, Robbins D, *et al.* (2005) Sorghum genome sequencing by methylation filtration. *PLoS Biology* **3**, e13. doi: 10.1371/journal.pbio.0030013
- Bennetzen JL (1996) The contributions of retroelements to plant genome organization, function and evolution. *Trends in Microbiology* **4**, 347–353. doi: 10.1016/0966-842X(96)10042-1
- Clifton SW, Minx P, Fauron CMR, Gibson M, Allen JO, *et al.* (2004) Sequence and comparative analysis of the maize NB mitochondrial genome. *Plant Physiology* **136**, 3486–3503. doi: 10.1104/pp.104.044602
- Cordeiro GM, Casu R, McIntyre CL, Manners JM, Henry RJ (2001) Microsatellite markers from sugarcane (*Saccharum* spp.) ESTs cross transferable to erianthus and sorghum. *Plant Science* **160**, 1115–1123. doi: 10.1016/S0168-9452(01)00365-X
- Foster-Hartnett D, Mudge J, Larsen D, Danesh D, Yan H, Denny R, Penuela S, Young ND (2002) Comparative genomic analysis of sequences sampled from a small region on soybean (*Glycine max*) molecular linkage group G. *Genome* **45**, 634–645. doi: 10.1139/g02-027
- Gepts P, Beavis WD, Brummer EC, Shoemaker RC, Stalker HT, Weeden NF, Young ND (2005) Legumes as a model plant family. Genomics for food and feed report of the cross-legume advances through genomics conference. *Plant Physiology* **137**, 1228–1235. doi: 10.1104/pp.105.060871
- Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, *et al.* (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**, 493–521. doi: 10.1038/nature02426
- Goldberg RB (1978) DNA sequence organization in the soybean plant. *Biochemical Genetics* **16**, 45–51. doi: 10.1007/BF00484384
- Goldblatt P (1981) Cytology and phylogeny of Leguminosae. In 'Advances in legume systematics. Part 2'. (Eds RM Polhill, PH Raven) pp. 427–463. (Royal Botanic Gardens: Kew)
- Gurley WB, Hepburn AG, Key JL (1979) Sequence organization of the soybean genome. *Biochimica et Biophysica Acta* **561**, 167–183.
- Hawkins JW, Van Keuren ML, Piatigorsky J, Law ML, Patterson D, Kao FT (1987) Confirmation of assignment of the human  $\alpha$  1-crystallin gene (*CRYA1*) to chromosome 21 with regional localization to q22.3. *Human Genetics* **76**, 375–380. doi: 10.1007/BF00272448
- Holmes I (2002) Transcendent elements: whole-genome transposon screens and open evolutionary questions. *Genome Research* **12**, 1152–1155. doi: 10.1101/gr.453102
- Jarvik T, Lark KG (1998) Characterization of *Soymar1*, a *Mariner* element in soybean. *Genetics* **149**, 1569–1574.
- Klein PE, Klein RR, Vrebalov J, Mullet JE (2003) Sequence-based alignment of sorghum chromosome 3 and rice chromosome 1 reveals extensive conservation of gene order and one major chromosomal rearrangement. *The Plant Journal* **34**, 605–621. doi: 10.1046/j.1365-313X.2003.01751.x
- Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921. doi: 10.1038/35057062
- Laten HM, Morris RO (1993) *SIRE-1*, a long interspersed repetitive DNA element from soybean with weak sequence similarity to retrotransposons: initial characterization and partial sequence. *Gene* **134**, 153–159. doi: 10.1016/0378-1119(93)90089-L
- Lin J-Y, Jacobus BH, SanMiguel P, Walling JG, Yuan Y, Shoemaker RC, Young ND, Jackson SA (2005) Pericentromeric regions of soybean (*Glycine max* L. Merr.) chromosomes consist of retroelements and tandemly repeated DNA and are structurally related and evolutionarily labile. *Genetics* **170**, 1221–1230. doi: 10.1534/genetics.105.041616
- Marek LF, Mudge J, Darnielle L, Grant D, Hanson N, *et al.* (2001) Soybean genomic survey: BAC-end sequences near RFLP and SSR markers. *Genome* **44**, 572–581. doi: 10.1139/gen-44-4-572
- Martienssen RA, Rabinowicz PD, O'Shaughnessy A, McCombie WR (2004) Sequencing the maize genome. *Current Opinion in Plant Biology* **7**, 102–107. doi: 10.1016/j.pbi.2004.01.010
- McCouch SR, Teytelman L, Xu YB, Lobos KB, Clare K, *et al.* (2002) Development and mapping of 2240 new SSR markers for rice (*Oryza sativa* L.). *DNA Research* **9**, 199–207. doi: 10.1093/dnares/9.6.199
- Morgante M, Jurman I, Shi L, Zhu T, Keim P, Rafalski JA (1997) The STR120 satellite DNA of soybean: organization, evolution and chromosomal specificity. *Chromosome Research* **5**, 363–373. doi: 10.1023/A:1018492208247

- Morgante M, Hanafey M, Powell W (2002) Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. *Nature Genetics* **30**, 194–200. doi: 10.1038/ng822
- Mudge J, Yan HH, Denny RL, Howe DK, Danesh D, Marek LF, Retzel E, Shoemaker RC, Young ND (2004) Soybean bacterial artificial chromosome contigs anchored with RFLPs: insights into genome duplication and gene clustering. *Genome* **47**, 361–372.
- Rabinowicz PD, Schutz K, Dedhia N, Yordan C, Parnell LD, Stein L, McCombie WR, Martienssen RA (1999) Differential methylation of genes and retrotransposons facilitates shotgun sequencing of the maize genome. *Nature Genetics* **23**, 305–308. doi: 10.1038/15479
- Rabinowicz PD, Citek R, Budiman MA, Nunberg A, Bedell JA, Lakey N, O'Shaughnessy AL, Nascimento LU, McCombie WR, Martienssen RA (2005) Differential methylation of genes and repeats in land plants. *Genome Research* **15**, 1431–1440. doi: 10.1101/gr.4100405
- Rhodes PR, Vodkin LO (1988) Organization of the Tgm family of transposable elements in soybean. *Genetics* **120**, 597–604.
- Schlueter J, Dixon P, Granger C, Grant D, Clark L, Doyle JJ, Shoemaker RC (2004) Mining EST databases to resolve evolutionary events in major plant species. *Genome* **47**, 868–876. doi: 10.1139/g04-047
- Shoemaker RC, Polzin K, Labate J, Specht J, Brummer EC, *et al.* (1996) Genome duplication in soybean (*Glycine* subgenus *soja*). *Genetics* **144**, 329–338.
- Shoemaker RC, Schlueter J, Doyle JJ (2006) Paleopolyploidy and gene duplication in soybean and other legumes. *Current Opinion in Plant Biology* **9**, 104–109. doi: 10.1016/j.pbi.2006.01.007
- Singh RJ, Hymowitz T (1988) The genomic relationship between *Glycine max* (L.) Merr. and *G. soja* Sieb. and Zucc. as revealed by pachytene chromosome analysis. *Theoretical and Applied Genetics* **76**, 705–711. doi: 10.1007/BF00303516
- Stacey G, Vodkin L, Parrott WA, Shoemaker RC (2004) National science foundation-sponsored workshop report: draft plan for soybean genomics. *Plant Physiology* **135**, 59–70. doi: 10.1104/pp.103.037903
- Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* **22**, 4673–4680.
- Vahedian M, Shi L, Zhu T, Okimoto R, Danna K, Keim P (1995) Genomic organization and evolution of the soybean SB92 satellite sequence. *Plant Molecular Biology* **29**, 857–862. doi: 10.1007/BF00041174
- Vodkin LO, Rhodes PR, Goldberg RB (1983) A lectin gene insertion has the structural features of a transposable element. *Cell* **34**, 1023–1031. doi: 10.1016/0092-8674(83)90560-3
- Whitelaw CA, Barbazuk WB, Pertea G, Chan AP, Cheung F, *et al.* (2003) Enrichment of gene-coding sequences in maize by genome filtration. *Science* **302**, 2118–2120. doi: 10.1126/science.1090047
- Wright DA, Voytas DF (2002) *Athila4* of *Arabidopsis* and *Calypso* of soybean define a lineage of endogenous plant retroviruses. *Genome Research* **12**, 122–131. doi: 10.1101/gr.196001

Manuscript received 27 April 2006, accepted 24 May 2006