Science in China Series C: Life Sciences

Genome size and sequence composition of moso bamboo: A comparative study

GUI YiJie[1], WANG Sheng[1], QUAN LiYan[1], ZHOU ChangPing[2], LONG ShiBao[2], ZHENG HuaJun[3], JIN Liang[1], ZHANG XianYin[1], MA NaiXun[4] & FAN LongJiang[1†]

[1] Institute of Crop Science/Institute of Bioinformatics, Zhejiang University, Hangzhou 310029, China;
[2] Zhejiang Anji Bamboo Exposition Garden, Anji 313300, China;
[3] Chinese National Human Genome Center at Shanghai, Shanghai 201203, China;
[4] The Research Institute of Subtropical Forestry, Chinese Academy of Forestry, Fuyang 311400, China

Moso bamboo (*Phyllostachys pubescens*) is one of the world's most important bamboo species. It has the largest area of all planted bamboo——over two-thirds of the total bamboo forest area——and the highest economic value in China. Moso bamboo is a tetraploid (4*x*=48) and a special member of the grasses family. Although several genomes have been sequenced or are being sequenced in the grasses family, we know little about the genome of the bambusoids (bamboos). In this study, the moso bamboo genome size was estimated to be about 2034 Mb by flow cytometry (FCM), using maize (cv. B73) and rice (cv. Nipponbare) as internal references. The rice genome has been sequenced and the maize genome is being sequenced. We found that the size of the moso bamboo genome is similar to that of maize but significantly larger than that of rice. To determine whether the bamboo genome has a high proportion of repeat elements, similar to that of the maize genome, approximately 1000 genome survey sequences (GSS) were generated. Sequence analysis showed that the proportion of repeat elements was 23.3% for the bamboo genome, which is significantly lower than that of the maize genome (65.7%). The bamboo repeat elements were mainly Gypsy/DIRS1 and Ty1/Copia LTR retrotransposons (14.7%), with a few DNA transposons. However, more genomic sequences are needed to confirm the above results due to several factors, such as the limitation of our GSS data. This study is the first to investigate sequence composition of the bamboo genome. Our results are valuable for future genome research of moso and other bamboos.

flow cytometry (FCM), genome size, GSS, *Phyllostachys pubescens*, repeat elements

Bamboo is one of the most important forest resources and has over 70 genera and 1200 species. It is mainly distributed in tropical and subtropical regions, with a few species found in temperate and frigid regions. In China, 48 genera and nearly 500 species are distributed in subtropical regions, or south of 40° northern latitude, particularly south of the Yangtze River. This area includes Zhejiang, Fujian, Jiangxi and Hunan provinces, covering an area of 7.2 million hectares. Moso bamboo (*Phyllostachys pubescens*) is the most important bamboo species in China. It is the most widely distributed with the largest planting area (over two-thirds of the total planted bamboo area) and has the highest economic value in China[1]. Bamboo is a monocot, classified in the subfamily Bambusoideae within the family Poaceae. The family Poaceae also includes rice, maize, wheat and other cereals. Previously, rice was classified in the subfamily Bambusoideae, but recently it was classified in a new subfamily, Oryzoideae[2]. Moso bamboo is monopodial bamboo, belonging to the genus *Phyllostachys* and the tribe Shibataeeae. Its somatic chromosome

Received April 30, 2006; accepted July 4, 2007
doi: 10.1007/s11427-007-0081-6
†Corresponding author (email: fanlj@zju.edu.cn)
Supported by the Program for New Century Excellent Talents in University of China

number is $2n=48$[3]. The basic chromosome number of bamboo is considered to be $x=12$ and two kinds of polyploids are found: tetraploid ($2n=4x=48$) for the monopodial bamboo and hexaploid ($2n=6x=72$) for the caespitose bamboo. Until now, most molecular biology studies on bamboo have focused on DNA polymorphisms for phylogenetic analyses[4, 5]. Its DNA sequences in public nucleotide databases are limited; for example, only 100 sequences from moso bamboo were found in GenBank at the time of writing (2006-08-17).

As a model of the grass family, rice has the smallest genome among cereals, and other members are usually several folds greater than rice in genome size. A clear trend, along with the increase in genome size, is an increase in the proportion of repeat elements, particularly retrotransposons[6]. The comparative genomics also suggest that insertions of repeat elements have contributed to the significant difference in genome size[7]. According to previous studies, the DNA contents of bamboo were 2.45—5.3 pg DNA/2C, and 4.17—5.3 pg for the temperate bamboo (*Phyllostachys*)[8]. This implies that bamboo genomes are relatively large within the grass family. Bamboo is a special node in plant genomics, particularly in poaceous genomics. Although massive genomic sequences have been determined in the Poaceae family, some of its key nodes are still blank, such as bamboo. Most poaceous genes are identifiable through large-scale sequencing. However, there remain many gaps in knowledge of intragenic variation across different species of the family, in which the bamboos represent an important branch[9].

In this study, genome size of moso bamboo, the most important bamboo species in China, was estimated using flow cytometry (FCM) with maize and rice as references. Nearly 1000 genomic sequences (Genome Survey Sequences, GSS) were determined. The moso bamboo genome was over 2000 Mb, clearly larger than that of rice and slightly smaller than that of maize. The proportion of repeat elements in our GSS sample was significantly lower than that of maize.

1 Materials and methods

1.1 Materials

Leaves of moso bamboo (*P. pubescens*) were collected from Anji Bamboo Exposition Garden, Zhejiang Province, China. Diploid rice (*Oryza sativa* ssp. *japonica*) (cv. Nipponbare) and wild tetraploid rice (*Oryza latifolia*) were provided by the China National Rice Research Institute. Maize (*Zea mays*) (cv. B73) was provided by the Institute of Crop Science, Zhejiang University, Hangzhou, China.

1.2 Methods

1.2.1 Genome size estimation. A FACSCalibur flow cytometer (Becton Dickinson, USA) was used to compare the genomic sizes of bamboo, maize and rice.

1.2.2 Sample preparation. The Otto (1990) method was adopted for sample preparation[10]: (1) midrib tissue excised from young leaves (20 mg) was chopped with a sharp scalpel in 1 mL of ice-cold Otto buffer I in a petri dish; (2) the suspension was filtered through a 40-μm mesh nylon filter and transferred to a 1.5 mL microtube; (3) the filtrate was centrifuged at 15700 *g* for 30 s; (4) the supernatant was gently discarded to 0.1 mL, and then the suspension was re-suspended by adding 200 μL Otto buffer I. Before determination, 600 μL Otto buffer II and 100 μL staining solution PI (Becton Dickinson, USA) were added.

1.2.3 Sample estimation. The flow cytometer was operated at 488 nm. The staining solution used was PI/RNase (Becton Dickinson, USA, Cat. No:550825). Before estimation, the system was preheated for 5 min. The equipment conditions, such as sampling rate, FSC/SSC detector voltage and FSC threshold, were optimized according to the sample. Maize and rice were used as internal references. All experiments were carried out in triplicate.

1.2.4 Genomic DNA extraction and sequencing. DNA was extracted from fresh bamboo leaves using the CTAB method[11]. For sequencing, the DNA fragments generated by sonification were cloned into pUC18 and transformed into *Escherichia coli* DH10B. Random clones were chosen to perform the sequencing reaction on an ABI 3730 DNA sequencer (Applied Biosystems). A total of 1008 reads were generated. After removing vector sequences by VecScreen (http://www.ncbi.nlm.nih.gov), the sequences longer than 150 bp were deposited into GenBank under accession numbers ED017875-ED018870. A total of 996 sequences were deposited.

1.2.5 Genome sequence analysis. Aside from the 996 GSS from moso bamboo, *Arabidopsis* (ATH1 Version 5.0) and rice (OSA1 Version 4.0) genome sequences were downloaded from TIGR (http://ww.tigr.org).

Non-random (gene-rich) maize BAC sequences were also downloaded from TIGR, and random BAC end sequences (BESs) were downloaded from ftp://ftp.genome.arizona.edu/pub/stc/maize/ [12]. Repeat elements were identified using RepeatMasker and RepeatProteinMask (http://www. repeatmasker.org) based on protein similarity with rice as the reference species. Besides RepBase[13], the plant repeat element databases in MIPS (http://mips.gsf.de) and TIGR[14] were also used. Potential coding sequences in the 996 GSSs were searched against the SWISS-PROT protein database using BLASTX (E<1e-7)[15].

2 Results

2.1 Comparative study on genome size

Flow cytometry is mainly used to give quantitative estimates and to generate specific multi-parameter data for cells or biological particles in rapid flow. The nuclei are first treated with specific fluorescent dyes, then rapidly passed through the flow chamber one by one. As these particles pass through the laser light source they emit a fluorescent signal. The signal strength reflects the concentration of the material within the nucleus, that is, the nuclear DNA content. The signals are then converted into electrical pulses by optical detectors and digital signals which can be recognized by a computer. FCM is widely applied to the estimation of ploidy in plant cells, including detection of chimeras and aneuploids. It has also been used to estimate genome size and to show changes during the cell cycle. For example, it has been used to show that DNA content changes cyclically during the cell cycle (G0, G1, S, G2 and M). Coefficients of variation in FCM analysis reflect the resolution or precision during the experiments, and a value less than 5% is considered to be reliable. In this study, the coefficient of variation was less than 3.5%.

DNA content of moso bamboo was estimated by FCM (Figure 1). The results indicated that the genome size of the tetraploid bamboo was slightly smaller than that of the maize genome. According to the comparison of G0/G1 peak values using maize cv. B73 (4.85 ± 0.18 pg $2C^{-1}$ [16] as an internal reference, the bamboo genome size was 86.92% of the maize genome (Table 1). It could then be estimated that the relative 2C DNA content was 4.22 pg, corresponding to a genome size of 2034 Mb for moso bamboo (1 pg DNA=0.965 × 10^9 bp[17]). In the same way using rice as internal reference, it indicated that the tetraploid bamboo genome (peak value 215) was clearly bigger than the diploid cultivated rice, and bigger than tetraploid wild rice (peak values 40 and 116, respectively). According to their peak values, the bamboo genome size is approximately 5.4 times that of the diploid cultivated rice, and 1.9 times that of the tetraploid wild rice. Previous studies suggested that the genome size of cv. Nipponbare was 389 Mb[18] and 2C DNA content of the tetraploid wild rice was 2.32 pg[19]. Taken together, genome size of the moso bamboo could also be estimated, which was consistent with that obtained by using maize as reference.

Two peaks were observed in maize during FCM analysis (Figure 1). Apparently, the first peak represented the 2C DNA content in G0/G1 phase, while the second peak represented the 4C DNA content in G2/M phase. The double peaks may be due to the maize samples used in this study, which may have been in the state of cell division.

Table 1 Peak values of moso bamboo and the internal reference (maize) in G0/G1 phase using flow cytometry

Replicate No.	Bamboo	Maize	Bamboo/Maize (%)
1	41.78	48.06	86.93
2	40.49	46.64	86.81
3	41.72	47.95	87.01
Average	41.33	47.55	86.92 ± 0.06

2.2 Analysis of genome sequence composition

Rice has the smallest size genome among cereals, and has the lowest proportion of repeat elements[20, 21]. Moso bamboo has a genome size similar to maize, which suggests that like maize, it may be rich in repeat elements. To test this hypothesis, 1008 GSS reads were generated. After removing vector sequences, 996 GSS greater than 150 bp in length, or 0.92 Mb in total length, were deposited in GenBank (accession numbers: ED017875-ED018870) and were used in this study. Most of them (85%) were over 900 bp in length, with an average length of 926.2 bp.

Repeat element analysis was performed based on the 996 GSS (Table 2). As shown in Table 2, moso bamboo and other cereals had relatively high proportions of repeat elements compared with the dicot *Arabidopsis*. However, the repeat element proportion of moso bamboo was significantly lower than that of maize. The proportion was 23.23% for the moso bamboo genome, which was similar to that of rice and 42% lower than that of maize. The composition of repeat elements in

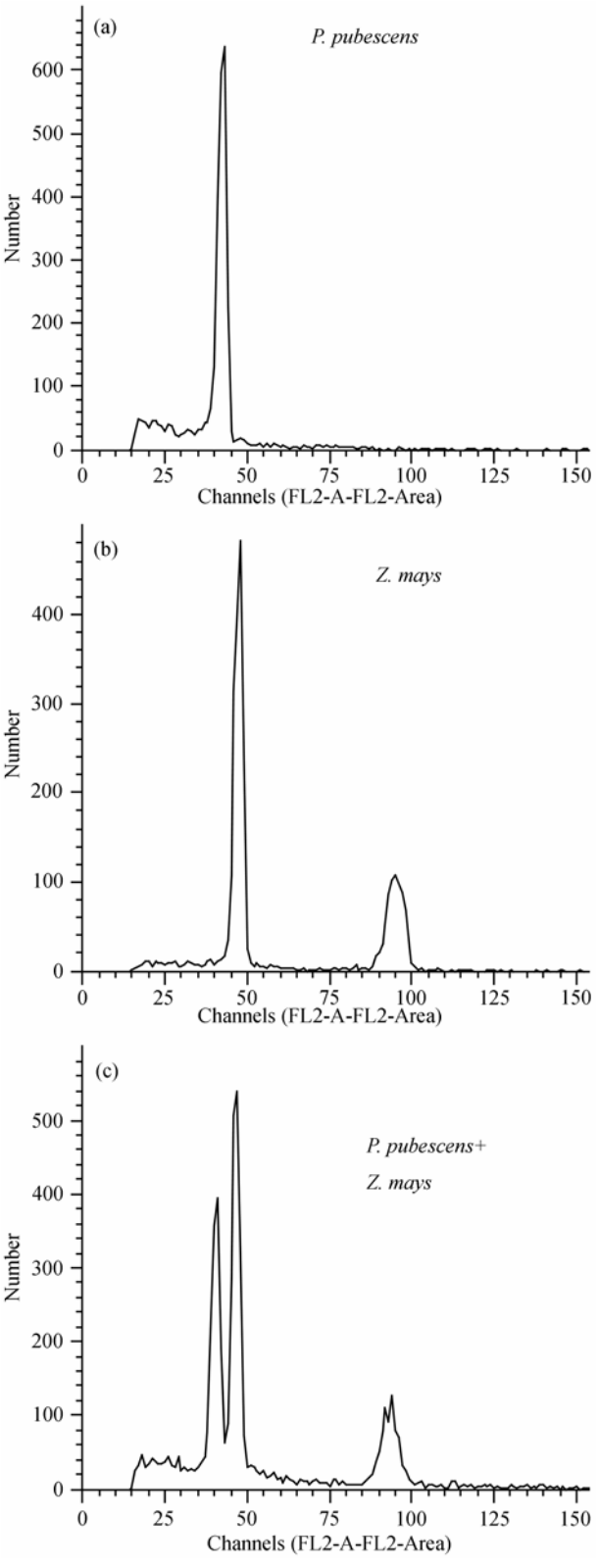

Figure 1 Comparison of fluorescence intensity between bamboo and maize. Top and middle: moso bamboo (*P. pubescens*) and cultivated maize (*Z. mays*), respectively. Bottom: moso bamboo with maize as an internal reference

moso bamboo was similar to maize; the majority was LTR retrotransposons (18.89%), while the proportion of DNA transposons was low (2.53%). The LTR retrotransposons were mainly Gypsy/DIRS1 type, and some were Ty1/Copia type. In rice, however, the proportions of LTR and DNA transposons were nearly equal, and most LTR transposons were the Gypsy/DIRS1 type. The proportions of other kinds of repeat elements (e.g. simple repeat sequences) did not show significant differences between the moso bamboo and other species. There was also no significant difference in base composition (e.g. GC content) between the three poaceous species. Our results need further confirmation by larger-scale genome sequencing in respect to our limited GSS data (less than 0.1% of the bamboo genome).

Potential coding sequences in the bamboo GSS were further determined. A total of 323 (32.4%) sequences had hits with SWISS-PROT protein sequences (<1e-7), and 104 (10.4%) sequences had hits with non-repetitive protein sequences, such as the NBS-LRR resistance gene and the S-adenosylmethionine decarboxylase gene. The best hits of the GSS were sequences from rice.

Table 2 Comparative analysis of repeat elements in the moso bamboo genome. RepeatMasker and MIPS plant repeat element database (http://mips.gsf.de) were used

Repeat elements (%)	Bamboo	Maize	Rice	Arabidopsis
Sequence number	996	996	996	996
GC content (%)	44.66	42.34	43.75	35.98
Total length (Mb)	0.92	0.89	0.92	0.92
Total repeats (%)	23.28	65.68	38.87	14.74
Retroelements	18.89	58.45	20.77	6.37
SINEs	0.01	0	0.53	0.03
LINEs	0.34	0.56	0.88	0.5
LTR elements	18.54	57.88	19.35	5.84
BEL/Pao	0	0	0	0
Ty1/Copia	7.12	18.49	2.42	0.98
Gypsy/DIRS1	9.47	29.18	11.87	1.61
Retroviral	0	0	0	0
DNA transposons	2.53	1.37	13.90	3.53
hobo-Activator	0.44	0.08	0.48	0.49
Tc1-IS630-Pogo	0	0	0	0.07
En-Spm	0.75	0.61	3.23	0.79
MuDR-IS905	0.75	0.16	1.71	1.84
Tourist/Harbinger	0	0.07	0.02	0.06
Unclassified	0.22	0.36	0.99	0.48
Total interspersed repeats	21.64	60.18	35.65	10.39
Small RNA	0.33	0.10	0.02	0.25
Satellites	0	0	0.11	0.48
Simple repeats	0.36	0.15	0.87	0.34
Low complexity	0.73	0.36	0.64	1.24

3 Discussion

In this study, we estimated the relative size of the moso bamboo genome using rice and maize as reference species, the genomes of which had been or are currently being sequenced. The sequence composition, including the amount and types of repeat elements, was analyzed by random genomic survey sequencing. The genome size of moso bamboo was similar to maize, but significantly bigger than the rice genome. The moso bamboo genome was not rich in repeat elements compared to rice and maize. About 10% of the GSSs had high similarities to known genes. Our results are helpful to genomic and genetic research on moso bamboo and also other bamboo species. Besides moso bamboo, the genus *Phyllostachys* contains many species with high economic values for China and other Asian countries, such as timber bamboo (*Phyllostachys bambusoides*), zaozhu (*Phyllostachys praecox*), and woody bamboo (*Phyllostachys acuta*). This study is the first to investigate genomic sequence composition of a bamboo species.

The relative 2C DNA content was 4.22 pg or 2034 Mb for moso bamboo (1 pg DNA=0.965 × 10^9 bp). The result was similar to 4.19 pg or 2021 Mb reported by Gielis et al. (1997)[8]. In early studies, chicken red blood cell (CBRC; 2.33 pg DNA/2C) was usually used as the reference. The chicken 2C DNA contents usually vary within a certain range (2.33—2.5 pg), and some studies have suggested alternative reference material[22]. Many model organisms have been sequenced and provide more accurate and reliable references for FCM. Johnston et al. (1999)[23] believed that the best references for FCM estimation of plant DNA content wer from plants, and DNA content of reference species should be similar, but not entirely equal to that of the sample. Diploid rice cv. Nipponbare and wild tetraploid rice were used as internal references in our first attempt and we found that their genome size greatly differed from that of moso bamboo. Using the genome size of Nipponbare as a reference, the bamboo genome size was estimated to be over 2000 Mb. We therefore chose maize as a more accurate reference.

In this study, 1000 random sequences from the bamboo genome were used to investigate sequence composition. To determine a reasonable sequence number, we performed computational simulations in rice, maize and *Arabidopsis*, and found that 1000 random sequences were an appropriate sample size for an entire genome according to our results (Table 2). Our results for proportion and composition of repeat elements of rice, maize and *Arabidopsis* were essentially the same based on their genome analysis[12,18,21,24]. Meanwhile, the variation among different samples (1000 sequences as one sampling event) was very low. For example, in the rice genome the proportion of total repeat elements was 38.81% ± 1.29% in our simulation experiments. The results indicated that the 1000 random genomic sequences used in this study were reasonable.

Our study indicated that there were fewer repeat elements than expected in moso bamboo. We carried out several measures to improve our estimations. First, we used all available repeat element databases and searching methods, including the method based on protein similarity. Second, we used the same method of genome sequence sampling across bamboo and other species. The sequences sampled from three reference species were 926 bp, the average length of the 996 bamboo GSS. Finally, we used rice as a reference species in repeat database searching. However, it should be noted that some factors will lead to underestimation of repeat proportion in the moso bamboo genome in this study. For example, the moso bamboo genome may contain novel repeat elements that have not yet been identified. In addition, the cloning method used in the present study may be biased to clones with non-repeat elements. Therefore the estimation of repeat elements of the moso bamboo genome needs further investigation in the future, based on more genomic sequences.

1 Jiang Z H. World Bamboo and Rattan (in Chinese). Shengyang: Liaoning Science and Technology Publishing House, 2002

2 GPWG (Grass Phylogeny Working Group). Phylogeny and subfamilial classification of the grasses (Poaceae). Ann Missouri Bot Garden, 2001, 88(3): 373—457

3 Li X L, Liu S, Song W Q, et al. Chromosome number of forty species of scattered bamboos. Acta Phytotaxonomica Sin (in Chinese), 1999, 37(6): 541—544

4 Mathews S, Tsai R C, Kellogg E A. Phylogenetic structure in the grass family (Poaceae): Evidence from the nuclear gene phytochrome B. Am J Bot, 2000, 87(1): 96—107

5 Guo Z H, Li D Z. Phylogenetics of the Thamnocalamus group and its allies (Gramineae: Bambusoideae): Inference from the sequences of GBSSI gene and ITS spacer. Mol Phylogenet Evol, 2004, 30(1): 1—12

6 SanMiguel P, Tikhonov A, Jin Y K, et al. Nested retrotransposons in the intergenic regions of the maize genome. Science, 1996, 274(5288): 765—768

7 Song R, Llaca V, Messing J. Mosaic organization of orthologous sequences in grass genomes. Genome Res, 2002, 12(10): 1549—1555
8 Gielis J, Valente P, Bridts C, et al. Estimation of DNA content of bamboos using flow cytometry and confocal laser scanning microscopy. The Bamboos. London: Academic Press, 1997. 215—223
9 Paterson A H, Freeling M, Sasaki T. Grains of knowledge: Genomics of model cereals. Genome Res, 2005, 15(12): 1643—1650
10 Otto F J. DAPI staining of fixed cells for high-resolution flow cytometry of nuclear DNA. In: Darzynkiewickz Z, Crissman H A, eds. Methods in Cell Biology. Vol 33. San Diego: Academic Press, 1990. 105—110
11 Wang G L, Fang H J. Plant Gene Engineering (in Chinese). 2nd ed. Beijing: Science Press, 2002. 744
12 Messing J, Bharti A K, Karlowski W M, et al. Sequence composition and genome organization of maize. Proc Natl Acad Sci USA, 2004, 101(40): 14349—14354
13 Jurka J. Repbase update: A database and an electronic journal of repetitive elements. Trends Genet, 2000, 16(9): 418—420
14 Ouyang S, Buell C R. The TIGR Plant Repeat Databases:collective resource for the identification of repetitive sequences in plants. Nucleic Acids Res, 2004, 32 (Database issue): 360—363
15 Altschul S F, Madden T L, Schaffer A A, et al. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res, 1997, 25(17): 3389—3402
16 Rayburn A L, Biradar D P, Bullock D G, et al. Nuclear DNA content in F sub(1) hybrids of maize. Heredity, 1993, 70(3): 294—300
17 Bennett M D, Smith J B. Nuclear DNA amounts in angiosperms. Phils Trans R Soc Lond B Biol Sci, 1976, 274(933): 227—274
18 International Rice Genome Sequencing Project. The map-based sequence of the rice genome. Nature, 2005, 436(7052): 793—800
19 Martinez C P, Arumuganathan K, Kikuchi H, et al. Estimation of nuclear DNA content in Oryza by flow cytometry. Rice Genet Newslett, 1993, 10: 116—119
20 Haberer G, Young S, Bharti A K,et al. Structure and architecture of the maize genome. Plant Physiol, 2005, 139(4): 1612—1624
21 Bedell J A, Budiman M A, Nunberg A, et al. Sorghum genome sequencing by methylation filtration. PLoS Biol, 2005, 3(1): e13
22 Tiersch T R, Chandler R W, Wachtel S S, et al. Reference standards for flow cytometry and application in comparative studies of nuclear DNA content. Cytometry, 1989, 10: 706—710
23 Johnston J P, Bennett M D, Rayburn A L, et al. Reference standards for determination of DNA content of plant nuclei. Am J Bot, 1999, 86(5): 609—613
24 Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature, 2000, 408(6814): 796—815