Mol Genet Genomics (2008) 280:287–292
DOI 10.1007/s00438-008-0364-z

ORIGINAL PAPER

# Pilot sequencing of onion genomic DNA reveals fragments of transposable elements, low gene densities, and significant gene enrichment after methyl filtration

**Jernej Jakše · Jenelle D. F. Meyer · Go Suzuki · John McCallum · Foo Cheung · Christopher D. Town · Michael J. Havey**

Received: 16 April 2008 / Accepted: 22 June 2008 / Published online: 10 July 2008


**Abstract** Sequencing of the onion (*Allium cepa*) genome is challenging because it has one of the largest nuclear genomes among cultivated plants. We undertook pilot sequencing of onion genomic DNA to estimate gene densities and investigate the nature and distribution of repetitive DNAs. Complete sequences from two onion BACs were AT rich (64.8%) and revealed long tracts of degenerated retroviral elements and transposons, similar to other larger plant genomes. Random BACs were end sequenced and only 3 of 460 ends showed significant ($e < -25$) non-organellar hits to the protein databases. The BAC-end sequences were AT rich (63.4%), similar to the completely sequenced BACs. A total of 499,997 bp of onion genomic DNA yielded an estimated mean density of one gene per 168 kb, among the lowest reported to date. Methyl filtration was highly effective relative to random shotgun reads in reducing frequencies of anonymous sequences from 82 to 55% and increasing non-organellar protein hits from 4 to 42%. Our results revealed no evidence for gene-dense regions and indicated that sequencing of methyl-filtered genomic fragments should be an efficient approach to reveal genic sequences in the onion genome.

**Keywords** Bacterial artificial chromosome · Reduced representation · Retrovirus · Transposon

## Introduction

Onion (*Allium cepa* L.) is a diploid ($2n = 2x = 16$) plant and possesses one of the largest nuclear genomes among all cultivated species with over 16 gigabasepairs of DNA per 1C, similar to that of hexaploid wheat (*Triticum aestivum*) and six times bigger than maize (*Zea mays*) (Arumuganathan and Earle 1991). The enormous genome of onion has slowed the development of genomic resources for this economically and phylogenetically important plant. In addition

Communicated by A. Tyagi.

Names are necessary to report factually on available data; however, the US Department of Agriculture (USDA) neither guarantees nor warrants the standard of the product, and the use of the name by USDA implies no approval of the product to the exclusion of others that may also be suitable.

J. Jakše · J. D. F. Meyer
Department of Horticulture, University of Wisconsin, 1575 Linden Drive, Madison, WI 53706, USA

J. Jakše
Agronomy Department, Biotechnical Faculty, University of Ljubljana, Jamnikarjeva 101, Ljubljana 1000, Slovenia

G. Suzuki
Laboratory of Plant Molecular Genetics, Division of Natural Science, Osaka Kyoiku University, 4-698-1 Asahigaoka, Kashiwara, Osaka 582-8582, Japan

J. McCallum
Crop and Food Research, Private Bag 4704, Christchurch, New Zealand

F. Cheung · C. D. Town
The Institute for Genomic Research, 9712 Medical Center Dr., Rockville, MD 20850, USA

M. J. Havey (✉)
Agricultural Research Service, USDA, Department of Horticulture, University of Wisconsin, 1575 Linden Drive, Madison, WI 53706, USA
e-mail: mjhavey@wisc.edu

to its huge size, the nuclear genome of onion possesses other distinguishing characteristics, including a unique GC-rich telomeric repeat (Adams et al. 2001, Fajkus et al. 2005) and GC content (32%) among the lowest known for angiosperms (Kirk et al. 1970). Codon biases and GC content of onion ESTs were more similar to *Arabidopsis thaliana* and the eudicots than to the grasses (Kuhl et al. 2004). Reassociation kinetics (Cot) of onion DNA have revealed a significant component of middle-repetitive sequences occurring in short-period interspersions among low-copy regions (Stack and Comings 1979). This structure of the onion DNA was supported by FISH analyses of random bacterial artificial chromosomes (BACs). An onion BAC library of 48,000 clones (0.3× coverage of the onion nuclear genome) has been synthesized and FISH analysis of random BACs revealed that 80% carried common repetitive DNAs and hybridized to entire chromosomes, 15% hybridized to centromeric or telomeric regions, and only 5% of BACs hybridized to specific regions on chromosomes (Suzuki et al. 2001). These results indicate that much of the onion genome is likely composed of many repetitive elements. In order to determine efficient approaches for sequencing of onion, we completed sequence analyses of onion BACs and BAC ends to estimate gene density and reveal types of repetitive elements. We also assessed the efficacy of methyl filtration to increase the proportion of genic hits after shotgun sequencing.

## Materials and methods

### Onion BAC and BAC-end sequencing

Onion BACs from a 0.3×-coverage library synthesized from the cultivar 'Cheonjudaego' (Suzuki et al. 2001) were randomly selected, isolated from overnight cultures using a minipreparation (Marra et al. 1997), and end sequenced using 10 pmol of primer and an initial denaturation of 95°C for 5 min., followed by 50 cycles of 95°C for 30 s, 50°C for 20 s, and 60°C for 4 min. For comparisons with other plants, ten random samples of BAC-end sequences were randomly chosen from datasets of *Arabidopsis thaliana* (47,788 bp from http://www.tigr.org/tdb/at/atgenome/bac_end_search/bac_end_search.html.ld), *Zea mays* (370,117 bp from Genbank), *Medicago truncatula* (177,786 bp from Genbank), *Oryza sativa* (127,459 bp from ftp://ftp.genome.clemson.edu/pub), and *Sorghum bicolor* (17,283 bp from ftp://ftp.genome.arizona.edu/pub), and were compared to the random onion BAC-end sequences. The onion BAC ends were used in BLASTX comparisons with the Non-Redundant Protein Database at TIGR using a minimum threshold of $e < -25$. Any BAC ends containing a top protein hit similar retro- or DNA-transposons were removed. A second round of filtering was carried out using BLASTN against the TIGR Plant Repeat Database using a $e < -5$ threshold.

We screened DNA pools (Suzuki et al. 2002) from the onion BAC library using oligonucleotide primers from onion cDNAs and isolated one BAC (1G-12-89) carrying a region similar to sulfite reductase (McCallum et al. 2002). A second onion BAC (S1-D12) was sequenced that showed discrete FISH signals to onion chromosomes (Fig. 2c of Suzuki et al. 2001)]. BAC 1G-12-89 was nick translated and hybridized as previously described (Bark and Havey 1995) to *Eco*RI digests of DNAs of bunching onion (*A. fistulosum*), chive (*A. schoenoprasum*), Chinese chive (*A. tuberosum*), garlic (*A. sativum*), and onion (*A. cepa*). The two onion BACs were sequenced to GenBank HTGS phase 2 quality sequence so that the structure, relative orientation, and position of all genes in the contig were revealed. Random small (2–3 kb) and large (10–12 kb) insert libraries were constructed from hydrodynamically sheared, size-selected DNA in a medium copy vector. These libraries were sequenced at ~3:1 ratio (small:large) up to a total sequence coverage of at least 8x using 1/32 volume Big Dye reactions in 384-well format and analyzed on ABI 3730xl sequencers. Base-calling was performed by Paracel TraceTuner, specifically trained for machine type and polymer. Sequence quality trimming and elimination of vector and *E. coli* sequences were conducted using in-house software (Chou and Holmes 2001). Sequences were assembled using TIGR assembler and the assemblies were ordered and oriented with respect to one another using Bambus software (Pop et al. 2004). Assembled sequences were submitted to the HTGS Division of GenBank. Individual shotgun reads and contigs of assembled sequences were compared using nucleotide and translated searches with the databases to search for genic-like regions. Similarities among repetitive sequences on the onion BACs were revealed using online versions of PipMaker and MultiPipmaker (Schwartz et al. 2000) after masking with RepeatMasker (Smit et al. 1996) using the PANICOID/RICE repeats. Basic program parameters were used in all analyses.

### Methyl filtration and sequencing of onion DNA

DNA was isolated by CsCl-banding (Bark and Havey 1995) from etiolated seedlings of the onion double haploid (DH) 15197 (gift of Seminis Seed Company, Woodland CA). DNA was sheared by nebulization, the 1 to 2 kb fraction was size-selected, end polished, ligated with *Bst*XI adapters and cloned into the *Bst*XI site of TIGR vector pHOS2 (a pBR322 derivative). Clones were propagated in *E. coli* DH10B to recover all sequences and in *E. coli* DH5alpha for methyl filtration. Sequencing was carried out as described above.

## Results

### Sequence characteristics of onion BACs and BAC ends

Onion BAC-end sequencing yielded 460 unique sequences (Genbank accessions ET222737 to ET223030) from 316 BACs totaling 297,703 bp with an average AT content of 63.4%. These end sequences were searched against the non-redundant protein database at TIGR and 5% (16 sequences covering 16,144 bp) showed significant similarities ($e < -25$) to the protein or EST databases and 54% (250 of 460 ends) to retroelements or transposons at $e < -6$. Of the 16 BAC-end sequences with high-quality matches to the protein database, 13 ends from 10 BACs were highly similar to parts of organellar genes, yielding an estimate of 3.2% organellar BACs in the onion library. Three ends from three different BACs were putative nuclear genes (glutamine synthase, subtilase, and hypothetical protein AT2g13670) that gave high-quality alignments of open reading frames with well-defined exon boundaries, suggesting that they are parts of intact genes rather than degenerated pseudogenes. None of three putative nuclear genes was present in the onion gene index at TIGR (August 2007), although a paralog of glutamine synthase was present. We generated 10 random samples of 460 BAC-end sequences from *A. thaliana*, rice, *M. truncatula*, sorghum, and maize and revealed higher numbers of hits to protein databases

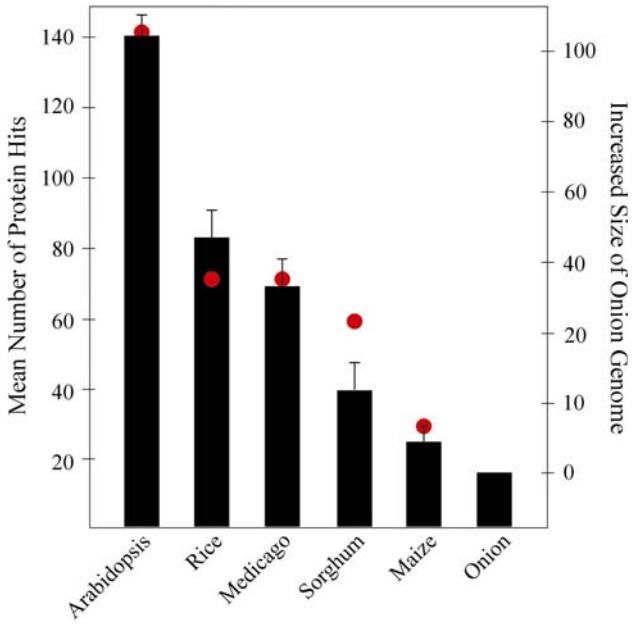


**Fig. 1** Histogram showing mean number of significant ($e < -25$) hits to protein databases (*left axis*) for 460 unique onion BAC-end sequences and for an equal number of randomly selected BAC-end sequences from *Arabidopsis*, rice, *Medicago*, sorghum, and maize. *Lines* show standard deviations among samples. *Dots* show the increased size of the onion nuclear genome relative to these plants (*right axis*)

($141.7 \pm 6.2$, $84.0 \pm 7.6$, $69.2 \pm 7.9$, $39.2 \pm 7.2$, and $24.5 \pm 5.2$, respectively) which closely paralleled differences in genome sizes (Fig. 1). Analyses of BAC-end sequences revealed high frequencies (>50%) of transposable elements, consistent with large genome sizes in other plants (SanMiguel et al. 1996).

We used onion cDNAs to screen a 0.3×-coverage BAC library of onion and identified one BAC (1G-12-89) carrying sequences similar to sulfite reductase (onion EST ACAAJ79). This BAC was 108,232 bp in size, 65.8% AT, and contained no intact genes (Table 1). The region between 49.5 and 53.7 kb carried several degenerated exons with similarity to the target sulfite reductase and the region between 54.9 and 60.1 had similarity to a MYB-related DNA-binding protein. The degrees of similarity between the genomic and EST sequences for sulfite reductase were low, indicating that this genomic region did not produce onion EST ACAAJ79. Over 50% of onion BAC 1G-12-89 showed similarities to transposon-like sequences (Fig. 2a). Hybridization of the entire 1G-12-89 BAC to a gel blot carrying DNA from various Alliums revealed the strongest cross-hybridizing repetitive DNAs in onion and bunching onion (Fig. 3), both members of *Allium* section *Cepa* (Hanelt 1990; Havey 1992; Raamsdonk et al. 1992).

Onion BAC S1-D12 showed discrete FISH signals at the ends of two onion chromosomes (Fig. 2c in Suzuki et al. 2001) and was selected for sequencing to determine if this BAC was gene-rich or possessed fewer repetitive DNAs. We produced two contigs (84,316 and 9,746 bp) from S1-D12 totaling 94,062 bp, which showed no significant hits to any of the protein or EST databases. This BAC possessed no genes and a plethora of AT-rich regions and short regions similar to parts of retroviruses or transposons (Fig. 2b). This result indicates that discrete FISH signals are not necessarily indicative gene-rich regions.

### Methyl filtration of onion genomic DNA

Reduced representation sequencing using methyl-filtered libraries is an effective tool to reduce the frequencies of repetitive DNAs and increase the proportion of random shot-gun reads showing significant similarities to expressed sequences (Rabinowicz et al. 1999; Palmer et al. 2003; Whitelaw et al. 2003). We completed pilot sequencing from whole genome shot-gun (WGS) and methyl-filtered genomic libraries created from a doubled-haploid population of onion. Sequences from the WGS and methyl-filtered libraries had average phred 20 read lengths of 883 and 781 bases yielding 95% and 88% sequencing efficiencies, respectively. Out of 6,590 random unfiltered sequences (Genbank accessions ET642110 through ET648699), 14% matched transposons and 2.6% matched other nuclear-encoded proteins. Out of 2,712 methyl-filtered sequences (Genbank

**Table 1** Sequence characteristics and estimated gene densities of onion bacterial artificial chromosomes (BACs) and BAC ends

| BAC | Genbank accessions | Sizes (bp) | AT (%) | Estimated gene density |
|---|---|---|---|---|
| Ends | ET222737 to ET223030 and ET437813 to ET437978 | 297,703 | 63.4 | 1 gene/99 kb |
| 1G-12-89 | DQ273270 | 108,232 | 65.8 | 1 gene/108 kb |
| S1-D12 | DQ273272 | 94,062 | 63.8 | No genes |

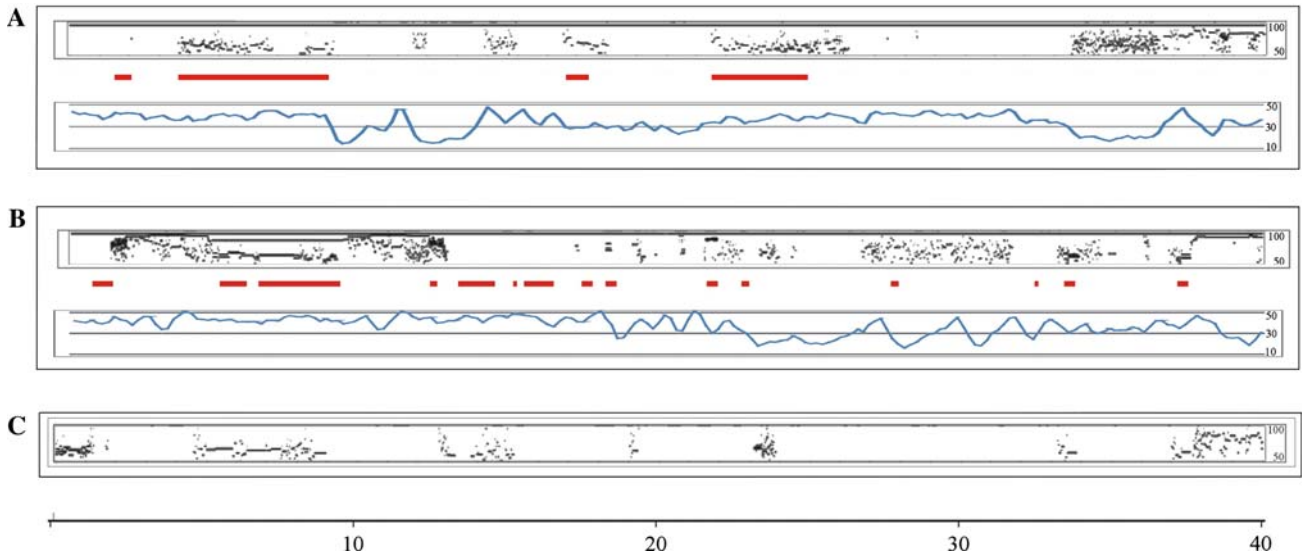


**Fig. 2** Percent identity plots (PIPs) showing similar (50–100%) repetitive DNAs across the first 40 kilobases of BAC 1G-12-89 (**a**), BAC S1-D12 (**b**), and between BACs 1G-12-89 and S1-D12 (**c**) of onion. Below PIPs are regions (*boxes*) showing significant ($e < -6$) similarities to retroviral or transposon-like sequences and percent GC plots (30–50%). Positions in kb are shown on scale at *bottom*

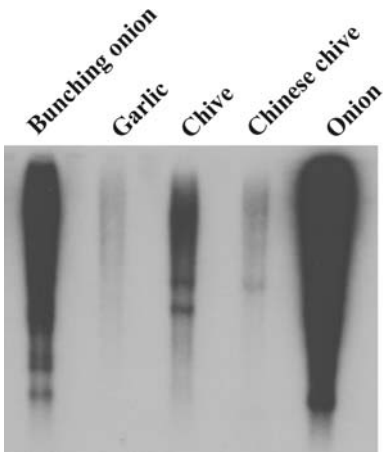


**Fig. 3** Autoradiogram from hybridization of onion BAC 1G-12-89 to DNA of bunching onion (*A. fistulosum*), garlic (*A. sativum*), chive (*A. schoenoprasum*), Chinese chive (*A. tuberosum*), and onion (*A. cepa*)

accessions ET639398 through ET642109), 3% matched transposons and 42% were similar to other non-organellar proteins. These results indicate that methyl filtration of onion DNA was very effective in reducing the proportion of both identifiable transposons (from 14 to 3%) and anonymous sequences (from 82 to 55%), and increasing non-organellar protein hits (Fig. 4).

## Discussion

Our sequence analyses of onion BACs revealed AT-rich regions and low gene densities, as expected given enormous genome of onion. In total, we produced 499,997 bp of onion genomic sequence and identified only three putatively nuclear-encoded genes, yielding an average gene density of one per 168 kb. A similar low gene density was previously reported for an onion BAC (GeneBank accession AB111058) that carried only the target alliinase gene (Do et al. 2003). The average AT content of the BACs was 64.8%, comparable to the BAC ends (63.4%) and higher than the mean of 55% for onion ESTs (Kuhl et al. 2004). Multi-PIP analysis revealed that the two onion BACs shared repetitive elements (Fig. 2). Over 50% of the onion BAC sequences were similar to transposons, many of which were degenerated. In the grasses, transposable elements contributed significantly to genome-size increases; approximately 14% of the rice genome (Jiang and Wessler

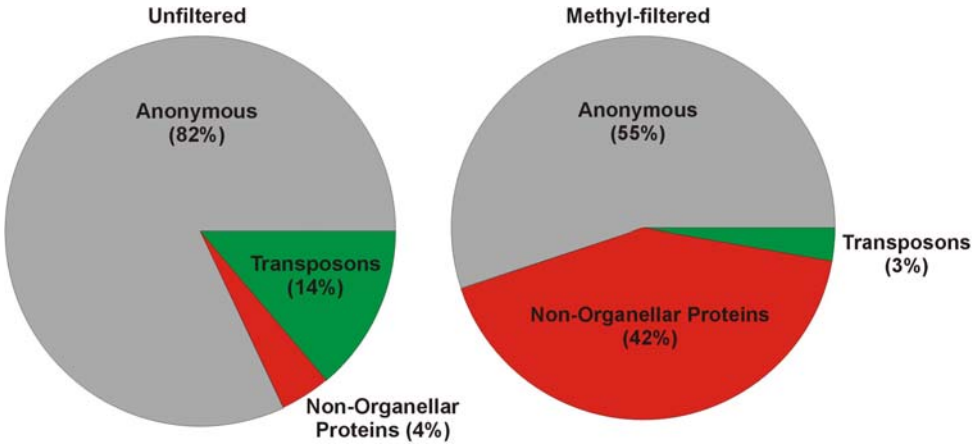


**Fig. 4** Relative percentages of randomly sequenced fragments from whole genome shot-gun (*left*) and methyl-filtered (*right*) libraries of onion showing no significant ($e < -9$) similarities to databases (anonymous) or were similar to transposons or non-organellar proteins

2001), 50–60% of the maize (Meyers et al. 2001), and more than 70% of the barley (Vicient et al. 1999) genomes are comprised of retrotransposons. Our survey sequencing indicates that the onion genome likely has low gene density, carries a plethora of degenerated retroviral and transposable elements, and may have experienced numerous increases of transposable elements without elimination of more ancient elements. These repetitive sequences were shared between onion and closely related bunching onion, and showed lower hybridization intensities to more distantly related chive, garlic, and Chinese chive (Fig. 3) (Klaas 1998). These results indicate that map-based cloning of onion genes will likely be difficult due to very low gene densities and common repetitive elements. However, construction of BAC contigs by fingerprinting (Marra et al. 1997) may be productive because tracts of unique sequences and degenerated transposable elements should yield unique restriction patterns.

Reduced representation sequencing has been proposed as a valid approach to increase the proportion of genic sequences from shot-gun sequencing of large genomes (Whitelaw et al. 2003). Methyl filtration of onion DNA reduced the numbers of anonymous (82–55%) and transposon-like (14–3%) sequences, as well as increased non-organellar protein hits over tenfold (Fig. 4). These reductions in transposon-like and anonymous sequences were greater than those reported for methyl filtration of maize DNA (Rabinowicz et al. 1999; Palmer et al. 2003; Whitelaw et al. 2003). The relatively low frequencies of organellar DNAs among the methyl-filtered fragments indicate that purification of nuclei prior to cloning was successful in reducing hypo-methylated organellar DNAs (McCullough et al. 1992). Therefore, sequencing of methyl-filtered genomic clones should complement EST sequencing (Kuhl et al. 2004) as an efficient approach to enrich for genic regions of the onion genome.

**Acknowledgments** This work was completed in compliance with the current laws governing genetic experimentation in Japan, New Zealand, and USA and was supported by the Initiative for Future Agriculture and Food Systems Grant no. 2001-52100-11344 from the USDA Cooperative State Research, Education, and Extension Service and a Fulbright-Hayes Post-doctoral Fellowship to JJ.

## References

Adams SP, Hartman TPV, Lim YK, Chase MW, Leitch AR (2001) Loss and recovery of *Arabidopsis*-type telomere repeat sequences 5′-(TTTAGGG)n-3′ in the evolution of a major radiation of flowering plants. Proc R Soc Lond Ser B 268:1541–1546

Arumuganathan K, Earle ED (1991) Nuclear DNA content of some important plant species. Plant Mol Biol Rep 9:208–218

Bark OH, Havey MJ (1995) Similarities and relationships among open-pollinated populations of the bulb onion as estimated by nuclear RFLPs. Theor Appl Genet 90:607–614

Chou HH, Holmes MH (2001) DNA sequence quality trimming and vector removal. Bioinformatics 17:1093–104

Do S, Suzuki G, Mukai Y (2003) Genomic organization of a novel root alliinase gene, ALL1, in onion. Gene 325:17–24

Fajkus J, Sykorova E, Leitch AR (2005) Telomeres in evolution and evolution of telomeres. Chrom Res 13:469–479

Hanelt P (1990) Taxonomy, evolution, and history. In: Brewster J, Rabinowitch H (eds) Onions and allied crops, vol 1. CRC Press, Boca Raton, pp 1–26

Havey M (1992) Restriction enzyme analysis of the chloroplast and nuclear 45s ribosomal DNA of *Allium* sections *Cepa* and *Phyllodolon*. Plant Syst Evol 183:17–31

Jiang N, Wessler SR (2001) Insertion preference of maize and rice miniature inverted repeat transposable elements as revealed by the analysis of nested elements. Plant Cell 13:2553–2564

Kirk JTO, Rees H, Evans G (1970) Base composition of nuclear DNA with the genus *Allium*. Heredity 25:507–512

Klaas M (1998) Applications and impact of molecular markers on evolutionary and diversity studies in the genus *Allium*. Plant Breed 117:297–308

Kuhl JC, Cheung F, Yuan Q, Martin W, Zewdie Y, McCallum J, Catanach A, Rutherford P, Sink KC, Jenderek M, Prince JP, Town CD, Havey MJ (2004) A unique set of 11,008 onion (*Allium cepa*) ESTs reveals expressed sequence and genomic differences between monocot orders Asparagales and Poales. Plant Cell 16:114–125

Marra MA, Kucaba TA, Dietrich NL, Green ED, Brownstein B, Wilson RK, McDonald KM, Hillier LW, McPherson JD, Waterston RH (1997) High throughput fingerprint analysis of large-insert clones. Genome Res 11:1072–1084

McCallum JA, Pither-Joyce M, Shaw M (2002) Sulfur deprivation and genotype affect gene expression and metabolism of onion roots. J Am Soc Hort Sci 127:583–589

McCullough AJ, Kangasjarvi J, Gengenbach BG, Jones RJ (1992) Plastid DNA in develoing maize endosperm: genome structure, methylation, and transcript accumulation patterns. Plant Physiol 100:958–964

Meyers BC, Tingey SV, Morgante M (2001) Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. Genome Res 110:1660–1676

Palmer LE, Rabinowicz PD, O'Shaughnessy AL, Balija VS, Nascimento LU, Dike S, de la Bastide M, Martienssen RA, McCombie WR (2003) Maize genome sequencing by methylation filtration. Science 302:2115–2117

Pop M, Kosack D, Salzberg S (2004) Hierarchical scaffolding with *Bambus*. Genome Res 14:149–159

van Raamsdonk L, Wietsma W, de Vries J (1992) Crossing experiments in *Allium* L. section *Cepa*. Bot J Linn Soc 109:293–303

Rabinowicz P, Schutz K, Dedhia N, Yordan C, Parnell L, Stein L, McCombie W, Martienssen R (1999) Differential methylation of genes and retrotransposons facilitates shotgun sequencing of the maize genome. Nature Genet 23:305–308

SanMiguel P, Tikhonov A, Jin YK, Motchoulskaia N, Zakharov D, Melake-Berhan A, Springer PS, Edwards KJ, Lee M, Avramova Z, Bennetzen JL (1996) Nested retrotransposons in the intergenic regions of the maize genome. Science 274:765–768

Schwartz S, Zhang Z, Frazer K, Smit A, Riemer C, Bouck J, Gibbs R, Hardison R, Miller W (2000) PipMakerA web server for aligning two genomic DNA sequences. Genome Res 10:577–586

Smit AFA, Hubley R, Green P (1996) RepeatMasker Open-3.0. http://www.repeatmasker.org, verified February 2008

Stack SM, Comings DE (1979) The chromosomes and DNA of *Allium cepa*. Chromosoma 70:161–181

Suzuki G, Do G, Mukai Y (2002) Efficient storage and screening system for onion BAC clones. Breed Sci (Japan) 52:157–159

Suzuki G, Ura A, Saito N, Do G, So B, Yamamoto M, Mukai Y (2001) BAC FISH analysis in *Allium cepa*. Genes Genet Syst (Japan) 76:251–255

Vicient CM, Suoniemi A, Anamthawat-Jonsson K, Tanskanen J, Beharav A, Nevo E, Schulman AH (1999) Retrotransposon BARE-1 and its role in genome evolution in the genus *Hordeum*. Plant Cell 11:1769–1784

Whitelaw CA, Barbazuk WB, Pertea G, Chan AP, Cheung F, Lee Y, Zheng L, van Heeringen S, Karamycheva S, Bennetzen JL, SanMiguel P, Lakey N, Bedell J, Yuan Y, Budiman MA, Resnick A, Van Aken S, Utterback T, Riedmuller S, Williams M, Feldblyum T, Schubert K, Beachy R, Fraser CM, Quackenbush J (2003) Enrichment of gene-coding sequences in maize by genome filtration. Science 302:2118–2120