AN INTERNATIONAL PLAN FOR SEQUENCING AND ANNOTATION OF ONION

Michael J. Havey, USDA-ARS and Univ. of Wisconsin, USA (mjhavey@wisc.edu)

and

John McCallum, Plant and Food Research, New Zealand (john.mccallum@plantandfood.co.nz)

Abstract

As large-scale DNA sequencing technologies become more efficient and cheaper, the genomic DNAs of more and more plants are being sequenced, assembled, and annotated. These complete sequences are extremely valuable for the identification of specific genes associated with important phenotypes. The purpose of this document is to propose an efficient strategy for an international effort to sequence the onion nuclear DNA and provide the sequences and annotations on a freely accessible website. A single doubled-haploid (DH) line of onion should be chosen and its DNA made available to serve as the common reference genome for independent sequencing runs by different labs. This DH line should be used for development of segregating families and their DNAs freely distributed. These resources will enable translational genomics of onion by providing researchers world-wide tools to more efficiently select for important traits in onion improvement. Eventually the bioinformatics resources can be expanded to include other major Alliums, such as bunching onion, garlic, leek, and chive.

Introduction

Sequencing and annotation of the onion (*Allium cepa* L.) genome presents a huge challenge. At approximately 16.4 giga (billion) bases per 1C, the onion nuclear genome is one of the largest nuclear genomes among all diploid and is over six times more than maize or humans (Arumuganathan and Earle 1991). Onion is diploid with 16 chromosomes and there is no evidence of a recent polyploidization event. Biochemical analyses have provided insights about the structure of the nuclear genome of onion. The GC content of onion DNA is 32%, the lowest known among angiosperms (Kirk et al. 1970). Although gene-rich islands exist in some grass species with larger genome sizes (Keller and Feuillet 2000), Cot reassociation kinetics revealed that the onion genome consists of middle-repetitive sequences occurring in short-period interspersions among single-copy regions (Stack and Comings 1979). Fluorescent in-situ hybridizations (FISH) of random genomic fragments cloned into bacterial artificial chromosomes (BACs) supported significant amounts of repetitive DNAs in the onion genome; Suzuki et al. (2001) reported that 80% of random BACs 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. These results suggest that much of the onion genome is composed of repetitive elements; however unique sequences exist at specific regions of the genome.

Using next-generation sequencing technologies and large computational abilities, sequencing, assembly, and annotation of large genomes has become almost routine. Current technologies include the Roche 454, Illumina, and ABI Solid platforms, which are constantly being refined and updated. Third-generation technologies such as Pacific Biosciences offer the possibility of greater read lengths, facilitating the assembly of large and complex genomes such as onion. Paralleling these newer technologies, the declining cost of computational power and growing availability of on-demand cloud-based computing is making assembly and annotation more efficient. Standardized web-based resources allow sharing and browsing of integrated data from annotated genomes. As a result, large-scale and cost-effective sequencing, assembly, and annotation of onion DNA are becoming feasible.

There are three main approaches to sequencing relatively large genomes, some of which are applicable to sequencing gene-rich regions of the onion genome. The first is shot-gun sequencing, in which the DNA is fragmented into relatively small pieces which are then randomly sequenced. Although shot-gun sequencing works well for relatively small genomes with few repetitive regions, it is inefficient as the sole strategy for larger and more complex genomes, since a large proportion of sequence fragments will

correspond to repeated sequences. As the size of a genome increases, the numbers of random reads required to cover the genome also increases. Normally, enough shot-gun sequences are generated to attain at least 5x coverage of the genome. Approximately 123 million random reads of an average of 400 bp would be required for 1x coverage of the nuclear DNA of onion. Until recently sequencing at such depth was cost prohibitive, but current standard sequencing technologies, such as the Illumina HiSeq platform, could generate sequence equivalent to a haploid onion genome (~20 Giga bases) in a single sequencing lane for under US $2000.

The second approach is sequencing of random cDNAs, which is an efficient method to sample expressed regions and enable gene discovery (Rounsley et al. 1996). In spite of its enormous genome, there is no evidence that onion carries significantly more genes than any other diploid plant. By sequencing only the expressed regions of the onion genome, one avoids large numbers of sequencing reads from non-expressed regions of the genome. Messenger RNA can be efficiently isolated and converted to cDNAs. Sequencing random cDNAs reveals the collection of genes (the transcriptome) expressed in any given tissue, at a specific developmental stage, or after a treatment. Normalization of the cDNA population reduces the frequencies of highly expressed genes and increases the number of sequencing reads from rarer transcripts. However rare transcripts may not be revealed by cDNA sequencing, even after normalization.

The third approach involves enrichment for lower-copy regions by removing repetitive DNAs. Reduced-representation sequencing focuses on genomic DNA enriched for non-repetitive or hypo-methylated regions (Rabinowicz et al. 1999, Peterson et al. 2002, Springer et al. 2004, Shagina et al. 2010). For both of these approaches, the nuclear DNA is randomly sheared to relatively small (a few hundred basepairs) fragments. The first strategy exploits CoT reassociation kinetics (Peterson et al. 2002). The population of random fragments is heat-denatured to produce single-stranded DNAs. The temperature of the solution carrying single-stranded DNAs is slowly lowered. DNA fragments carrying repetitive motifs are much more likely to find a match among single-stranded molecules, as compared to less repetitive DNAs. By removing fragments with double-stranded regions, either by column exclusion or double-strand nuclease (DSN) treatments, repetitive DNAs are selected against and the frequencies of unique DNA fragments increase. After cycles of selection against repetitive DNAs, the remaining fragments are sequenced. A second approach for reduced representation sequencing involves selection against methylated DNA. Expressed regions of the genome tend to be hypo (lowly) methylated; repetitive or non-expressed regions tend to have more methyl groups attached to their DNA. After shearing the DNA, random fragments are cloned into a plasmid vector and transformed into a specific bacterial strain that degrades methylated DNAs. When the methylated DNA is degraded, antibiotic resistance carried on the plasmid is lost and the bacteria become susceptible to the antibiotic; only bacteria carrying a non-methylated fragment survive. The corresponding DNA fragments can be then be purified and sequenced.

Onion sequencing to date

Preliminary sequencing of onion DNA has exploited all three of these approaches and revealed important characteristics of the onion genome.

Sequencing of Random Genomic Fragments: Random genomic fragments of onion have been sequenced, assembled, and annotated in three different experiments. The first utilized large genomic fragments cloned into BACs (Suzuki et al. 2001). Sanger sequencing of two entire BAC clones, as well as the random ends of genomic fragments cloned into BACs, yielded 298 kilobases of random-end sequences and 202 kb from the entire BAC clones (Jakse et al. 2008). The AT content of this random genomic sequence was 35.7%, close to the estimate of 32% by Kirk et al. (1970). Only 5% of this random genomic sequences showed significant similarities to non-organellar proteins, 25% were highly similar to retrotransposons or transposons, and 70% showed no significant hits to the databases and were primarily degenerated retroelements (Jakse et al. 2008). Therefore Sanger sequencing of BAC clones or ends yielded an estimate of only one gene every 168 kb! This very low gene density has been supported by

two additional sequence surveys of the onion genome. DNA was isolated from etiolated seedlings of a doubled haploid (DH) line 15197 of onion (gift of Seminis Seed Company, Woodland CA). The DNA was sheared and random fragments sequenced using Sanger and 454 technologies. Sanger sequencing generated 6,590 reads (Genbank accessions ET642110 through ET648699), of which 82% had no significant hits in the databases, 14% matched transposons, and 4% matched nuclear-encoded proteins (organellar hits were removed). This same DNA sample was then sequenced using one-half of a Roche 454 plate. A total of 845,529 reads yielded 15,341,323 bp of sequence, of which 5% showed significant similarities to non-organellar proteins (Havey unpublished). Overall, these three survey sequencing projects revealed that the onion genome may have experienced numerous explosions of transposons without elimination of ancient elements, resulting in very low gene densities (Jakse et al. 2008). Therefore, widely used approaches such as BAC-to-BAC or whole-genome shotgun sequencing will be inefficient and expensive for onion.

Transcriptome Sequencing: Sequencing of random onion cDNAs has been completed and supported the efficiency of this approach for onion. In a pilot study, McCallum et al. (2001) completed Sanger sequencing of random cDNAs from three non-normalized libraries of onion. In another study, Kuhl et al. (2004) completed 20,000 Sanger sequencing reactions from a normalized cDNA library of onion (Genbank accessions CF434396 to CF452784). Codon biases and GC content of the onion cDNAs were more similar to the eudicots than to the grasses, and the GC content of coding regions was higher than that of the onion genome as a whole (Kuhl et al. 2004). The cDNAs have been assembled, annotated, and used to create the onion gene index (http://compbio.dfci.harvard.edu/tgi/cgi-bin/tgi/gimain.pl?gudb=onion). More recently, sequencing of normalized cDNA libraries has been conducted in two sets of parental lines used for mapping (Havey unpublished; Baldwin et al 2012 *submitted*, NCBI BioProject 60277 http://www.ncbi.nlm.nih.gov/bioproject/60277). These projects are providing a wealth of expressed sequences for gene discovery and marker development, and will aid future genome annotation.

Reduced-Representation Sequencing of Genomic Regions: Pilot sequencing of methyl-filtered DNA fragments has been completed from onion DH 15197. Out of 2,712 methyl-filtered fragments (Genbank accessions ET639398 through ET642109), 55% were anonymous, 3% matched transposons, and 42% were similar to non-organellar proteins. These results indicate that methyl-filtration of onion DNA was effective in reducing the proportion of both transposons (from 14% to 3% for random reads) and anonymous sequences (from 82% to 55% for random reads), as well as increasing non-organellar protein hits over 10-fold (Jakse et al. 2008). These reductions in transposon-like and anonymous sequences are the highest so far reported for any plant and indicate that sequencing of methyl-filtered DNA fragments from onion is an efficient approach to enrich for genic regions in the enormous onion genome (Jakse et al. 2008). To date, there is no report of random sequencing of onion DNA after Cot re-association. In pilot studies (McCallum et al unpublished), we constructed DSN-normalized genomic libraries from DH 2150 from Cornell University (Alan et al. 2007). These libraries show significant reductions in abundance of high-copy sequences and increase in frequency of single-copy markers as assessed by qPCR analyses. Pilot sequencing (Illumina HiSeq platform) has confirmed reduction of repeat family abundance similar to that achieved by methyl filtration (Jakse et al. 2008). More extensive sequencing is now being conducted to characterize frequency of repetitive and genic regions in these libraries, and assess the practicality of these methods for the very large *Allium* genomes.

An International Effort for Sequencing of the Onion Genome

We propose to undertake sequencing of the nuclear genome of onion focusing on the transcriptome and reduced representation of genomic DNA. All sequencing efforts should focus on a single DH population available for distribution to private and public sector labs world-wide. We propose that long-day DH lines from Cornell University, such as 2107 or 2150 (Alan et al. 2007), be used because they are relatively vigorous and produce adequate amounts of seed. Over the longer term, additional DH populations, such as short or intermediate day lengths, should also be used. DNAs from these DH lines and segregating

families, as well as mapping data, should be freely available to the international community. The development of BIN mapping families from crosses among diverse DH populations will avoid the costly maintenance of onion seed and would allow researchers to assign new markers or candidate genes to specific recombinational units (BINs). Data assigning phenotypes, either as qualitative or quantitative traits, to specific regions should be available so locations of candidate genes can be compared to previously evaluated phenotypes.

For the transcriptome, cDNA sequencing to date has used a variety of normalized and non-normalized libraries. Future sequencing efforts should focus on cDNAs from a single DH population subjected to different growing conditions, treatments, etc. Libraries should be normalized in order to maximize the number unique sequences produced. Use of a single DH population will also help to reveal expressed paralogs in the onion genome. We recommend that all cDNA sequencing projects share their sequences with the international research community by placing assembled and annotated sequences into publically curated database(s).

Complementing transcriptome sequencing, the second major focus should be reduced representation sequencing of genomic DNAs. These efforts should use the same DH line as transcriptome sequencing. Methyl- or Cot-filtered or DSN-treated libraries should be synthesized from one DH line and made freely available to the research community. It is desirable to use different sequencing technologies in order to benefit from longer reads versus greater confidence (such as across homopolymers) offered by specific platforms. The random sequence reads should be assembled against the transcriptome in order to build contigs covering introns and promoter regions. We expect that a large number of the genomic sequences will not assemble because they show no significant similarities to the cDNAs, due simply to the size of the onion genome.

A single bioinformatic resource should be developed that continually scans public databases for new onion sequences, assembles and aligns these new sequences with established records, and automatically provides useful information such as annotations, intron-exon borders, and potentially paralogous regions. As sequencing information from other onion populations becomes available, molecular polymorphisms (such as single nucleotide polymorphisms and simple sequence repeats) should be revealed across specific genes or regions. AlliumMap is a web-based resource established at http://alliumgenetics.org to hold curated data concerning genetic maps constructed in onion and allied Alliums using transcriptome data (McCallum et al. 2012). Although this site contains links from markers to corresponding sequence data and primers, it is desirable that the site should contain more sequence data and associated annotations, to enable integration with larger volumes of expressed and genomic sequences at they become available. We propose to migrate current data in AlliumMap to a more advanced web-based database using the Tripal database construction toolkit (Ficklin et al. 2011).

One challenge over the long term will be continued support of bioinformatic resources for the Alliums. As public-sector funding becomes more restricted and sporadic, long-term maintenance of an Allium-specific bioinformatic resources is a serious challenge. The Allium research community must identify sources of public and private funds to maintain up-to-date bioinformatics and sequence information for tagging of important phenotypes.

References

Alan, A., Lim, M., Mutschler, M., and Earle, E.D. (2007) Complementary strategies for ploidy manipulations in gynogenic onion (*Allium cepa* L.). Plant Sci. 173:25-31.

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

Feuillet, C., and Keller, B. (1999) High gene density is conserved at syntenic loci of small and large grass genomes. Proc. Natl. Acad. Sci. (USA) 96:8265-8270.

Ficklin, S.P., Sanderson, L., Cheng, Staton, M.E., Lee, T. , Cho, I., Jung, S., Bett, K.E., and Main, D. (2011) Tripal: a construction toolkit for online genome databases. Database (Oxford).

Jakše, J., Meyer, J.D.F., Suzuki, G., McCallum, J., Cheung, F., Town, C.D., and Havey, M.J. (2008) Pilot sequencing of onion genomic DNA reveals fragmented transposable elements, low gene densities, and significant gene enrichment after methyl filtration. Mol. Genet. Genomics 280:287-292.

Kirk, J.T.O., Rees, H., and Evans, G. (1970) Base composition of nuclear DNA with the genus *Allium*. Heredity 25:507-512.

Kuhl, J.C., Cheung, F., Yuan, Q., Martin, W., Zewdie, Y., McCallum, J., Catanach, A., Rutherford, P., Sink, K.C., Jenderek, M., Prince, J.P., Town, C.D., and Havey, M.J. (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.

McCallum, J., Leite, D., Pither-Joyce, M., and Havey, M.J. (2001) Expressed sequence markers for genetic analysis of bulb onion (*Allium cepa)*. Theor. Appl. Genet. 103:979-991.

McCallum, J.A., Baldwin, S., Shigyo, M., Deng, Y., van Heusden, S., Pither-Joyce, M., Kenel, F. (2012) AlliumMap-a comparative genomics resource for cultivated Allium vegetables. BMC Genomics 13:168.

Peterson, D., Schulze, S., Sciara, E., Lee, S., Bowers, J., Nagel, A., Jiang, N., Tibbitts, D., Wessler, S., Paterson, A. (2002) Integration of Cot analysis, DNA cloning, and high-throughput sequencing facilitates genome characterization and gene discovery. Genome Res. 12:795-807.

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

Shagina, I., Bogdanova, E., Mamedov, I.Z., Lebedev, Y., Lukyanov, S., and Shagin, D. (2010) Normalization of genomic DNA using duplex-specific nuclease. Biotechniques 48:455-459.

Springer, N., Xu, X., and Barbazuk, W.B. (2004) Utility of different gene enrichment approaches toward identifying and sequencing the maize gene space. Plant Physiol. 136:3023-3033.

Stack, S.M., and Comings, D.E. (1979) The chromosomes and DNA of *Allium cepa.* Chromosoma 70, 161-181.

Suzuki, G., Ura A., Saito N., Do G., So B., Yamamoto M., and Mukai Y. (2001) BAC FISH analysis in *Allium cepa*. Genes Genetic Systems 76:251-255.