Minireview

Maize DNA-sequencing strategies and genome organization

Ron J Okagaki and Ronald L Phillips

Address: Department of Agronomy and Plant Genetics, and Center for Plant and Microbial Genomics, The University of Minnesota, St. Paul, MN 55108, USA.

Correspondence: Ron J Okagaki. E-mail: okaga002@tc.umn.edu

Published: 16 April 2004

Genome ***Biology*** 2004, **5:**223

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2004/5/5/223

Abstract

A large amount of repetitive DNA complicates the assembly of the maize genome sequence. Genome-filtration techniques, such as methylation-filtration and high-CoT separation, enrich gene sequences in genomic libraries. These methods may provide a low-cost alternative to whole-genome sequencing for maize and other complex genomes.

The maize and human genomes have similar sizes (2,500 and 3,200 megabases, respectively) and contain large amounts of repetitive sequence [1,2]. But differences between the two genomes create unique challenges. The available data suggest that most maize repetitive sequences accumulated in the past six million years [3]. This means that they should be more conserved than human repetitive sequences, most of which are over 25 million years old [2]. Plant genes, including maize genes, tend to be small; *Arabidopsis* and rice genes average between 2.4 and 5 kilobases [4-6], whereas human genes average about 27 kilobases [2]. Identifying genes may therefore be easier in maize; but whole-genome sequence assembly may prove more difficult because of the degree of conservation of its repetitive sequences.

Completion of a draft rice genome sequence [5,7] stimulated discussion on how to proceed with similar efforts for other crops. This discussion is tempered by an awareness of the difficulties to be faced with most crops. Plant genomes are usually large, composed largely of repetitive sequences, and are often polyploid. The costs of whole-genome sequencing will be substantial. In 2001, the National Science Foundation (NSF) sponsored a workshop to discuss sequencing the maize genome in light of these realities [1]. Out of these discussions came a strategy for using genome filtration as a low-cost alternative to fully sequencing the maize genome, so as to sequence clones from libraries enriched for genes, and then place these sequences on genetic or physical maps.

Two genome-filtration techniques were proposed for enriching gene sequences in genomic libraries. The first technique uses 'high-CoT' libraries; in this approach renaturation kinetics (represented by the product of DNA concentration (Co) and time (T), CoT, at which renaturation occurs) are used to separate repetitive sequences from low-copy sequences. The low-copy DNA renatures more slowly than repetitive sequences, and this fraction is enriched for genes [8]. The second technique, methylation filtration, is based on the tendency for repetitive sequences to be hyper-methylated in higher plants. Genomic libraries are constructed in *Escherichia coli* strains that have a functional *McrBC* restriction-modification system, which does not permit the propagation of heavily methylated DNA, thus excluding most repetitive sequences and enriching the library for gene-rich sequences [9]. Among major cereal crops, maize has an intermediate-sized genome, whereas the genomes of wheat, barley and oat are much larger. Decisions made with maize will thus help determine how to proceed with sequencing other crop genomes. Two recent papers by Palmer *et al.* [9] and Whitelaw *et al.* [10] describe the application of genome filtration to maize.

Genome filtration works

The Whitelaw *et al.* paper [10] compared genome filtration with random genomic shotgun sequencing. From the random library, 73% of 34,074 sequences were identified as repetitive. In contrast, 35% of the 95,233 methylation-filtered and 21% of 100,000 high-CoT sequences were repetitive. Over 900,000 sequence reads of the latter two libraries have now been completed and deposited in a public database [11]. The high-CoT and methylation-filtered clone sequences were found to be enriched for sequences related to known plant genes. For example, 13% of methylation-filtered and 11% of high-CoT sequences were similar to known plant expressed sequence tags (ESTs), whereas only 4% of sequences from random libraries were similar. Palmer and coworkers [9] developed an independent set of approximately 100,000 methylation-filtered sequences, and found that 8.6% of these exhibited sequence similarity to their gene database, while 24% of them matched a known repetitive sequence. They additionally showed that rates of new gene discovery per sequence read were similar for EST and methylation-filtration libraries [9].

An earlier study suggested that methylation-filtration can detect 95% of maize exons [12], and analyses in the two recent papers [9,10] suggest that most maize genes may be captured by filtration. These predictions are, however, based on detecting typical polypeptide-encoding genes. Will enrichment techniques capture genes encoding very small proteins or small RNAs? Tandem duplications, which are common in plant genomes, are another concern [4,6]. Will filtration be able to distinguish between copies, including those that have evolved distinct functions? It is possible that genome filtration could miss a number of genes.

There are, however, reasons for optimism. First, sequences for genes encoding small polypetides or RNAs could be among the uncharacterized sequences found in the filtered libraries. After sequencing reads were assembled into contigs, 63% of high-CoT assemblies and 39% of methylation-filtration assemblies had no significant matches to a gene or repeat sequence in the database at The Institute for Genomic Research [10,11]. Second, the methylation-filtration and high-CoT techniques sample from partially different fractions of the maize genome. It was estimated that of all the sequences sampled in the methylation-filtration and high-CoT libraries, approximately one-third were recovered by both approaches [10]. Using both techniques thus samples a greater fraction of the genome, and it seems possible that genes encoding microRNAs and small polypeptides will be captured by one or other technique.

The application of genome filtration for sequencing the maize genome would require the mapping of sequences onto physical or genetic maps, as noted at the NSF workshop [1]. How this mapping step is carried out will be a critical decision. As positional cloning is likely to be a major use of the mapped sequences, high-resolution map data are desirable. Placing sequences onto maps derived from bacterial artificial chromosome (BAC) contigs by hybridization or low-pass sequencing, would be appropriate. Genome filtration may prove to be most effective when a closely related species has already been sequenced, because synteny between species can then provide the positional information. Studies of cereal genomes suggest that rice is not sufficiently related to maize to adequately fill this gap in genome information [13,14]. In this light, synteny to important crops, in addition to genome size, may be an important criterion for selecting model species to sequence in the future.

When is genome filtration appropriate?

Enrichment may not be an appropriate approach for all species. Methylation filtration has worked well in maize because plant genes are largely unmethylated [12]. Furthermore, there is little repetitive sequence within plant genes themselves that could interfere with high-CoT selection, the exception being MITES (miniature inverted-repeat transposable elements), which are very small and usually poorly conserved [15]. Plant transcription units tend to be small [4-6], and their regulatory regions are compact. A wealth of experience with transgene constructs in plants demonstrates that in general only a few kilobases of flanking sequence are required for tissue and developmental regulation, although exceptions do exist. For instance, the maize *P1* gene promoter is unusually large, extending 5 kilobases upstream of the transcription start site [16]. Gene and genome organization must be considered before applying genome-filtration techniques to other organisms.

If funding becomes available, there are strong reasons for sequencing the entire maize genome. Access to the hundreds of mutations isolated over the past 75 years is one compelling reason. The agronomic importance of maize, in the United States and other countries, is another. A complete sequence of the maize genome would provide researchers with gene sequences, regulatory sequences, precise positional information, and markers for high-resolution mapping. These are the obvious reasons for whole-genome sequencing, but others may in fact prove more rewarding. We now know that different maize lines do not have identical complements of genes. In one region sequenced from two lines, four of the ten genes present in one line were absent from the other [17]. Tandem duplications provide an opportunity for gene number to increase or decrease within pedigrees [18,19], and duplication allows epigenetic regulation of gene expression [19,20]. Perhaps epigenetic interactions and variation in gene content underlie heterosis, whereby hybrids show increased vigor compared to their parents. This, together with the long breeding records and extraordinary genetic variation in maize, provides very special opportunities. Genome filtration coupled with mapping relatively inexpensively provides much of the same information that can be found in a complete genome sequence. But a full genome sequence provides a much broader foundation for exploring the complete genome.

References

1. Bennetzen JL, Chandler VL, Schnable P: **National Science Foundation-sponsored workshop report. Maize genome sequencing project.** *Plant Physiol* 2001, **127:**1572-1578.
2. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, *et al.*: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:**860-921.
3. SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL: **The paleontology of intergene retrotransposons of maize.** *Nat Genet* 1998, **20:**43-45.
4. The Arabidopsis Genome Initiative: **Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*.** *Nature* 2000, **408:**796-815.
5. Yu J, Hu S, Wang J, Wong GK-S, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X, *et al.*: **A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*).** *Science* 2002, **296:**79-92.
6. Sasaki T, Matsumoto T, Yamamoto K, Sakata K, Baba T, Katayose Y, Wu J, Niimura Y, Cheng Z, Nagamura S, *et al.*: **The genome sequence and structure of rice chromosome 1.** *Nature* 2002, **420:**312-316.
7. Goff SA, Ricke D, Lan T-H, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H, *et al.*: **A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*).** *Science* 2002, **296:**92-100.
8. Peterson DG, Schulze SR, Sciara EB, Lee SA, Bowers JE, Nagel A, Jiang N, Tibbitts DC, Wessler SR, Paterson AH: **Integration of Cot analysis, DNA cloning, and high-throughput sequencing facilitates genome characterization and gene discovery.** *Genome Res* 2002, **12:**795-807.
9. Palmer LE, Rabinowicz PD, O'Shaughnessy AL, Balija VS, Nascimento LU, Dike S, de la Bastide M, Martienssen RA, McCombie WR: **Maize genome sequencing by methylation filtration.** *Science* 2003, **302:**2115-2117.
10. Whitelaw CA, Barbazuk WB, Pertea G, Chan AP, Cheung F, Lee Y, Zheng L, van Heeringen S, Karamycheva S, Bennetzen JL, *et al.*: **Enrichment of gene-coding sequences in maize by genome filtration.** *Science* 2003, **302:**2118-2120.
11. **The TIGR Maize Database** [http://www.tigr.org/tdb/tgi/maize/]
12. Rabinowicz PD, Palmer LE, May BP, Hemann MT, Lowe SW, McCombie WR, Martienssen RA: **Genes and transposons are differentially methylated in plants, but not in mammals.** *Genome Res* 2003, **13:**2658-2664.
13. Song R, Llaca V, Messing J: **Mosaic organization of orthologous sequences in grass genomes.** *Genome Res* 2002, **12:**1549-1555.
14. Bennetzen JL, Ma J: **The genetic colinearity of rice and other cereals on the basis of genomic sequence analysis.** *Curr Opin Plant Biol* 2003, **6:**128-133.
15. Bureau TE, Wessler SR: **Mobile inverted-repeat elements of the *Tourist* family are associated with the genes of many cereal grasses.** *Proc Natl Acad Sci USA* 1994, **91:**1411-1415.
16. Sidorenko LV, Li X, Cocciolone SM, Chopra S, Tagliani L, Bowen B, Daniels M, Peterson T: **Complex structure of a maize *Myb* gene promoter: functional analysis in transgenic plants.** *Plant J* 2000, **22:**471-482.
17. Fu H, Dooner HK: **Intraspecific violation of genetic colinearity and its implications in maize.** *Proc Natl Acad Sci USA* 2002, **99:**9573-9578.
18. Lynch M, Conery JS: **The evolutionary fate and consequences of duplicate genes.** *Science* 2000, **290:**1151-1155.
19. Kermicle JL, Eggleston WB, Alleman A: **Organization of paramutagenicity in *R-stippled* maize.** *Genetics* 1995, **141:**361-372.
20. Assaad FF, Tucker KL, Signer ER: **Epigenetic repeat-induced gene silencing (RIGS) in *Arabidopsis*.** *Plant Mol Biol* 1993, **22:**1067-1085.

comment
reviews
reports
deposited research
refereed research
interactions
information