Sequencing the maize genome: Rationale, current status and future prospects

Parvez Sofi[1,*], A. G. Rather[1], Abdul Mateen[2] and Amjad Husaini[3]

[1]Division of Plant Breeding and Genetics, SKUAST, Srinagar 191 121, India
[2]School of Life Sciences, Chicago University, Chicago, Illinois, USA
[3]Department of Biotechnology, Hamdard University, New Delhi 110 062, India

Maize is one of the important food crops and possesses one of the well-studied and most tractable genetic systems. Even though rice has been regarded as a model reference plant for genomic studies, the enormous amount of local rearrangements have distorted the local microcolinearity in maize that make rice a too distant model for map-based cloning in maize. The maize genome-sequencing project was launched on 20 September 2002 by National Science Foundation. Various sequencing techniques like methyl filtration and high CoT analysis have been standardized, both of which are based on differential methylation of gene-rich and gene-poor regions. The sequencing studies have reported that maize genome contains about 42,000–59,000 genes with average gene size of 3000–3200 bp and gene density of 1 per 40 to 1 per 53 bp. The information generated will help in gene identification, expression and regulation across grass genomes and also unravel the evolution of complex genomes.

Keywords: Filtration, high CoT analysis, maize sequencing.

THE grass family Gramineae (alternative name: Poaceae) is one of the largest among monocots, consisting of more than 10,000 species covering almost one-third of arable land and a major source of food, feed and fuel for a large population in both developed and developing countries. The family evolved some 65 million years ago and has diversified into a variety of crops like rice, maize, sugarcane, oats, sorghum, wheat, etc. Maize (*Zea mays* L.) belongs to tribe Maydeae which consists of two groups, namely Luxuriantes comprising *Z. luxurians*, *Z. diploperennis* and *Z. perennis*, and Zea comprising *Z. mays*, *Z. mexicana* and *Z. parviglumis*. It is a recently domesticated form of tropical grass teosinte[1]. Owing to its superb genetic variability and enormous biological diversity is referred to as the '*Drosophilla* of the plant kingdom' and provides significant clues in understanding several genetic principles that are of practical interest to plant breeders[2]. Besides its importance in global agricultural economy, maize also possess one of the most well studied and most tractable genetic systems among cereals. Even today it serves as a model genetic system for the study of transposons, mieosis[3] and limits of fine mapping of phenotypes[4]. The transposon *Mu* has been a particularly useful mutagen in maize that allows for subsequent tagging and cloning by a variety of methods[5]. However, lack of comprehensive understanding of gene content and organization is a serious limitation to the speedy advancement in basic and applied aspects of research on the maize genome[6]. In fact, rice, another member of the grass family has been a model crop for genomic research mostly due to its genomic features such as small genome size (430 Mbp), relatively higher total recombination distance, lower proportion of kilobases per recombination unit, lower amount of repetitive DNA, etc.[7].

The maize genome

Plant genomes differ from mammalian genomes in the enormity of size range. In fact, within the grass family itself there is wide variation in genome size[8]. The genome sizes of rice, maize, wheat, sorghum and barley are approximately 430, 2500, 16,000, 750 and 4900 Mbp respectively. Genera like *Saccharum* and *Festuca* are even more complicated with large complex genomes, displaying wide variation in ploidy level, with over 100 chromosomes in some species[9]. Such greater diversity is mainly due to the long history of evolution of Poaceae. Evolutionary divergence of the grass family occurred some 50–60 million years ago. Rice and maize diverged some 50 million years ago, whereas maize and sorghum diverged some 20 million years ago.

The maize ($N = 10$) genome size is about 2500 Mbp with recombination distance of 1500 cM, and proportion of kilobases per recombination unit at 1700 kb/cM. The percentage of repetitive DNA in maize[10] is about 66. One of the important features of plant genomes is the lack of correlation between the number of chromosomes, genome size and number of genes. Rice with $N = 12$ has a genome size of 430 Mbp, whereas maize ($N = 10$) has a genome size of 2500 Mbp, and barley ($N = 7$) has 5000 Mbp. The larger genomic size of maize is due to the fact that it has undergone whole genome duplication. Most of the maize genome is repetitive. The repetitive sequences consist of two broad groups: tandem arrays and interspersed repeats[11].

*For correspondence. (e-mail: phdpbg@yahoo.com)

The former group comprises telomeric, sub-telomeric and centromeric repeats, ribosomal RNA genes and satellite DNA. These likely arise due to replication slippage (especially in case of SSR's) or unequal recombination events. In contrast, interspersed repeats are usually small translocated segments, mostly active or defective transposable elements[12]. Such elements are characterized by their ability to catalyse their own movements, resulting in an increase in their copy number. Out of all such elements, retrotransposons constitute the bulk of repetitive sequences in larger genomes, as in the case of maize. Functional genes represent only 15% genome, whereas the bulk comprises retrotransposons[13], which are dispersed throughout the genome. The rest of the genome consists of additional repeats, including the yet uncharacterized sequences or sequences that have diverged too far from known repeats[10]. In fact, retrotransposons are far more frequent in the maize genome than DNA transposons, whereas in the case of rice it is the opposite. Various kinds of repetitive sequences in maize are shown in Table 1. The maize genome as such can be regarded as genes interspersed with large islands of nested retrotransposons inserted between genes, and the genes are often associated with inverted repeat transposons[14]. The active genes are relatively under-methylated in comparison to the repetitive sequences. Thus the genome may have regions of varied gene density. The BACs cloned so far have been found to house some 2–16 genes[6].

The maize genome is a striking example of genome evolution in terms of polyploidization and transposon expansion[15]. Due to these processes, the retrotransposons constitute nearly 80% of the maize genome. The sequencing of a 225 kb region near the *Adh-1* region revealed nine genes and 74% retrotransposons, whereas the same orthologous region in sorghum also contains 14 genes and is only 78 kb in size. Such large amount of insertions have taken place in about 6 million years, which have drastically affected gene spacing in the maize genome. These retrotransposon explosions have dispersed thousands of copies periodically over the entire length of the maize genome without specificity to any particular sequences. However, natural selection cannot be ruled out in this process, as such selection favoured or disfavoured fixation of such insertions, with the result that gene distances do not correlate with the proposition that all genomic regions have been subjected to similar retrotransposition process. The general evidences are that such a process has been more rapid in gene-poor regions than gene-rich regions. Besides, the process is more effective in expanding the larger non-coding sequences compared to smaller ones, thus resulting in extreme differences in gene distances. Haberer *et al.*[10] estimated the (G + C) content of coding (exons) and non-coding (introns) sequences and found-marked differences in the (G + C) content. In the case of exons it was about 55.4% (40–75%) and in the case of introns it was 42.3% (30–60%). The average size of the intron was 407 bp and that of the exon was 259 bp. The (G + C) content was almost similar to that of rice, but greater than that of *Arabidopsis*. Similarly, the exon size in maize and rice was comparable, but intron size in maize was much higher than rice and *Arabidopsis*.

Table 1. Various classes of maize repetitive sequences (V4.0)

Class	Number
Transposable elements	
Retrotransposons	18,510
Transposons	617
MITES	816
Centromere-related sequences	353
Telomere-related sequences	26
Ribosomal genes	273
Other repetitive sequences	
Known repetitive sequences	1041
Unknown repetitive sequences	102
RECON-predicted repeats	5063

Sequencing the maize genome: the rationale

The Gramineae family comprises some of the important crop plants which provide food, feed and industrial raw materials and has thus a significant position in global agricultural economy and food security. Growth of agricultural productivity at the national and international level is undoubtly a function of pace of improvement in the productivity of cereal crops. The economic and scientific importance of cereals has motivated a rich history of research on genetics, development and evolution. Thus, understanding the genetic make-up of cereals will have a great impact for such an endeavour, in addition to the fact that it provides an insight into some of the interesting biological phenomena which are of practical significance to geneticists and plant breeders. Rice has been identified as a model crop for genomic studies and has been sequenced. Now one of the important questions that can be raised is that, given the amount of evidences coming up about genome conservation in grasses, would it be suitable to sequence the maize genome? Comparative analysis of cereal genomes, including rice, maize, wheat, sorghum, barley, etc. has revealed a high level of gene order conservation at macro level[16,17]. Rice genome sequencing, in this context, was viewed as a major breakthrough for use as a reference plant to cut the expenditure and labour involved in individual crop sequencing[18]. Besides, most cereals other than rice have large complex genomes with high proportion of repetitive sequences. This is a major obstacle in speedy progress during systematic sequencing[19]. Moreover, it needs to be validated whether genomics will bring about perceived enhancements in levels of productivity, given the great deal of information accumulating regarding gene discoveries, gene localization and characterization.

Comparative genomics deals with cross genome organization and function to estimate the synteny of biological organization[20]. Organismal evolution often provides threads of continuity allowing for analysis of the biological systems to link genes, protein expression, and growth and developmental traits across species and genera. Such a comparative analysis determines the relatedness patterns leading to new elucidations. Cereals exhibit almost 35-fold variation in genome size despite the fact that they share a common set of genes. This invites questions into the evolution of genome structure and size in grasses[21]. Moreover, most of the members share whole genome duplication before divergence into isolated lineages. In this context, the genomes that underwent duplication (maize, sugarcane) or polyploidization (wheat) can be potential genetic systems for exploring the repeating patterns of genes or gene families. An important aim of comparative genomic analysis is to transfer information from model plants to related organisms either due to the complexity of their genomes or due to the fact that their economic contribution does not warrant such genome analysis at individual crop level (orphan crops). Ahn and Tanksley[16] first demonstrated genome conservation between rice and maize. However, comparative genomic studies in grasses have shown that collinearity is more profound at megabase level (macrocollinearity) and not so much at the level of small genomic regions (microcollinearity). There are two compelling situations for studying the maize genome in finer details. One, that the process of genome diversification has resulted in about 15,000 rearrangements which distort the microsysteny providing for differentiations of grass genome[22–25]. Such duplications, inversions and translocations over a long period of evolution make rice too distant a model crop for efficient map-based cloning[26]. Therefore, a broader understanding of genes present in maize in terms of their organization, expression and interaction is imperative to improve the agronomic characteristics of this valuable food crop. In this regard, development of high-quality integrated, physical and genetic maps is a prerequisite, and serves as a foundation for such studies. Once our understanding of the maize genome is better in relation to rice, inferences drawn from rice genomic analysis can be more efficiently used to make predictions on the basis of comparisons. More importantly, sequencing the maize genome will also help in functional characterization of rice genes on an equally reciprocal basis, similar to comparative genomic analysis of mouse and human genomes[27].

Secondly, the application of synteny due to macrocollinearity in cereals has serious limitations due to the absence of comprehensive genomic information in crops other than rice. Besides such collinearity at recombinational map level does not always correspond to local genome structure level[28]. Therefore, if all the maize genes are characterized in terms of their location on the genome and ordered on a physical–genetic map and compared to the map of rice which already stands sequenced, the inferences drawn from such comparisons will provide more useful information regarding the structure, function and evolution of genes and the missing links between related genomes. This in turn would broadly empower the maize community, leading to more rapid understanding of maize genes[6]. Thus sequencing the maize genome and the functional analysis of its genes will help unravel the molecular basis of agronomically important traits. The fact that maize holds an important position in the global agricultural economy and food security should be an appropriate rationale for directing our efforts towards an initiative for sequencing its genome. However, an even stronger justification comes from the fact that complete sequencing of its genome will surely have significant implications for plant biology[29]. Therefore, it is high time that we develop efficient strategies for maize genome sequencing to get a better insight into the genetic systems underlying morphological, physiological and developmental phases of plant growth.

The maize genome sequencing project

Maize offers one of the best genetic systems that provides good opportunity to open up new vistas on complex organization, evolution and dynamic behaviour of what is arguably the most interesting genome. Maize is one of the highly diversified crop species owing to its long history of evolution and high outer-crossing, which have resulted in greater genetic variability and biological diversity. This was one of the strongest temptations for geneticists to consider the maize genome to be sequenced after *Arbidopsis* and rice, even though there were other candidates from the plant kingdom. This will have stronger implications on understanding of the structure, function and evolution of plant genomes, as well as a challenge given the complexity of the maize genome, because a greater proportion of it comprises of retrotransposons and other repetitive elements.

It was in response to a strong mandate from maize geneticists that a workshop sponsored by the National Science Foundation (NSF) was held on 2 July 2001 at St. Louis, USA to deliberate upon the technical feasibility of sequencing the maize genome. The participants were unanimous that sequencing all maize genes and placing them on a cross-referenced physical–genetic map was an extremely worthwhile exercise, which was a feasible and timely goal achievable at a reasonable cost[6]. As a follow-up to the understanding arrived at the St. Louis workshop, the NSF launched the maize genome sequencing project on 20 September 2002. The fact that such an endeavour aims at concerted efforts directed towards this goal, was driven by advances in DNA sequence technology and high resolution and high throughput DNA fingerprinting methods. Besides, implications of using the information

for comparative studies of grass genomes for possible gene synteny were immense, because *Arabidopsis* and rice genome sequences will not provide sufficient understanding of the genetic system of maize[26].

A number of research groups are currently working towards the sequencing of the maize genome under the aegies of the Maize Genomics Consortium (MGC). The goal was to finish and map out the entire gene islands within the maize genome by the end of 2006. A number of other agencies, especially the National Corn Growers Association and several industries are making valuable political and financial contribution towards the realization of the set goals, which is a clear indication of the collaborative interaction between academia and private sector that has been the characteristic of maize research[30].

The MGC consists of four member institutions, namely Donald Danforth Plant Sciences Center, The Institute for Genomic Research (TIGR), Purdue University and Orion Genomics. MGC was awarded a two-year plant genomics grant of US$ 6 million in September 2002 by NSF, to develop strategies for isolation and sequencing of the maize genome. Other institutions working in collaboration with MGC are University of Georgia, University of Arizona, Cold Spring Harbor Laboratory, Iowa State University, The Waksman Institute, Munich Information Centre for Protein Sequences, MIT Centre for Genomic Research and University of Missouri. Table 2 elaborates the various functions of members of the MGC.

Strategies for sequencing the maize genome

With the set objectives, the members and collaborators of the MGC had a challenging task to devise strategies for sequencing the maize genome. The clone-by-clone approach and whole genome shotgun approaches used in sequencing the human and rice genomes will not be effective in the case of maize due to the large size and complexity of the maize genome and the large proportion of repetitive sequences. Therefore, alternate sequencing strategies were developed and standardized, namely methyl filtration method and high CoT method. To meet its goals, MGC intends to generate about 800,000 total sequence reads, out of which the analysis of 200,000 reads has been presented in the 2003 edition of *Science*[31]. The aim is to develop gene-rich libraries as a fast and cost-effective method of genome sequencing. However, the usefulness of this approach will depend on successful assembly and annotation of the obtained sequence[30]. According to the agreement reached at the NSF-sponsored workshop in 2001, the maize inbred line B73 will be the primary focus of the sequencing project. This inbred line is the source of BAC libraries presently used to develop framework physical map and the many public ESTs besides being a good representative of the maize germplasm[6]. This is in line with the international rice genomic sequencing project rationale of using a single Japonica variety 'Nipponbare' for sequencing the rice genome[32].

Table 2. Responsibilities of MGC members

Institution	Responsibilities
Donald Danforth Plant Sciences Center	Evaluate gene hit rates of high CoT and methyl filtrated assemblies
	Estimate gene coverage
	Estimate maize genome coverage
	Compare methyl filtration vs high CoT
	Correlate assembled sequence islands to maize genetic/physical map
TIGR	Sequencing methyl filtration and high CoT libraries
	Sequences assembly and integration
	Sequence annonation
Purdue University	Construction of high CoT libraries
Orion Genomics	Construction of methyl filtration libraries.

The methyl filtration method was developed by Robert Martienssen and Richard McComble at Cold Spring Harbor Laboratory and is exclusively licenced to Orion Genomics. It is based on the premise that repetitive or non-coding sequences are hypermethylated and the genes are relatively under methylated[33]. In this technique the genomic libraries are constructed in *Escherichia coli* strains having functional *Mcr BC* restriction-modification system, which does not permit propagation of hypermethylated DNA, thus excluding repetitive sequences and enriching the library with gene-rich sequences by 5–7 times[31]. Methyl filtration has the potential to detect 95% maize genes[34,35]. However, its efficiency to detect small proteins, RNAs or tandem duplications, which are common in maize genome, has been questioned. Besides, it is being proposed that such a strategy could miss a number of maize genes[33,36]. However, this strategy has worked well in maize as the genes contain little repetitive DNA which might interfere with selection strategies, except MITES (miniature inverted-repeat transposable elements) which are characterized by poor conservation[37].

The high CoT selection strategy was developed at Purdue University by Joff Bennettzen. This exploits the fact that gene sequences are in relatively low copy abundances compared to large copy number repetitive non-coding sequences. It makes use of 'high CoT libraries'[38], and is based on the premise that genes which are unmethylated could be separated from repetitive sequences on the basis of renaturation kinetics.

$$\text{Renaturation kinetics} = \text{DNA concentration} \times \text{time at which the renaturation occurs.}$$

The low copy number gene sequences usually renature slowly compared with the repetitive sequences and this can be enriched for genes. Thus, higher CoT values are found in case of genes where reassociation after denaturation is slower. High CoT sequencing procedure may exclude

gene families. However, both strategies used together can cover almost 95% of the maize genes. Whitelaw *et al.*[35] found that of all the sequences sampled in methylation filtration and high CoT libraries, about one third were recorded in both strategies. Thus using both strategies, even a sizeable proportion of the gene encoding small proteins and small RNAs can be recovered. They also gave comparative data regarding the efficacy of sequencing strategies. It was reported that out of 34,074, 95,233 and 100,000 sequences from random genomic shotgun, methyl filtration and high CoT libraries respectively, the percentage of repetitive sequences was 73, 32 and 21 respectively. Besides, the percentage of sequences with similarity to known plant ESTs was 13, 11 and 4 respectively. Thus high CoT and methyl filtration clone sequences were highly enriched for genes compared to random shotgun library. Yuan *et al.*[39] also reported that sequencing of unrenatured DNA enriches the gene sequences by more than fourfold, from 5% in case of random library to more than 20% for high CoT library, and enhances the prediction of gene discovery to more than 95%. Consequently, the sequence reads necessary to sequence the full gene space would be lowered by fourfold compared to shotgun library. Springer *et al.*[40] used the methyl filtration and high CoT method and reported 7–8-fold increase in the rate of gene discovery.

The results pertaining to differential efficacy of various sequencing strategies have been presented by Whitelaw *et al.*[35]. They used both methyl filtration and high CoT methods individually and simultaneously, as well as unfiltered sequences. The study revealed that greater proportion of unique sequences was recovered using methyl filtration and high CoT methods simultaneously (93 Mbp) followed by high CoT method (54 Mbp), methyl filtration method (42 Mbp) and unfiltered sequences (24 Mbp). Consequently, the proportion of repetitive sequences was highest in unfiltered sequences (68%) followed by methyl filtration (33%), methyl filtration + high CoT (23%) and high CoT (14%).

Current status of sequencing the maize genome

The preliminary findings of the maize genome strategies were presented in *Science*[31,35]. The maize inbred line B73 was used for sequencing as a standard representative of the maize germplasm. Palmer *et al.*[31] constructed methyl filtrated libraries from immature ear nuclear DNA and generated 96,576 methyl filtration reads. Comparison with known plant ESTs from GenBank revealed that 8.6% of the reads were gene-enriched sequences, whereas 24% reads matched repetitive sequences. The rest of the reads were presumably unknown repeats, intergenic regions and a minor number of promoters and introns. The analysis removed 93% of repeats. They also analysed 5679 unfiltered or shotgun sequences, but were able to detect only 1.4% sequences matching genes, whereas 57% were repeats. Assuming a genome size of 25,000 Mbp, they estimated that 17% of the genome amounting to 425 Mbp, is undermethylated or gene space. This is almost equal to the rice genome size and can be conveniently sequenced. Similarly, Whitelaw *et al.*[35] analysed about 100,000 high CoT libraries and reported that one-third of sequences were recovered similar to those recovered by the methylation filtration method of Palmer *et al.*[31].

Messing *et al.*[8] analysed three BAC libraries of B73 using *Hin*dIII, *Eco*RI and *Mbo*I. The study found 7.5% of the genome to be gene-rich, amounting to 178 Mbp assuming a genome size of 2365 Mbp. The average size of the gene was about 3000 bp, which is comparable to that of rice genes, suggesting that the possible number of genes in maize is about 59,000, slightly higher than the estimated number of 37,500 in rice[41], but with a slightly lesser gene density of 1 per 40 kb. The study revealed that 58% of the genome is repetitive, but did not characterize the remaining 34.5% amounting to 816 Mbp (Table 3).

Haberer *et al.*[10] also constructed BAC libraries of inbred B73 using *Hin*dIII, *Eco*RI and *Mbo*I. They constructed 100 BAC clones out of which 89 yielded ordered and oriented sequence assemblies and 11 clones were not fully ordered. Their study revealed that 7% of genome amounting to 167 Mbp, assuming a genome size of 2365 Mbp, was represented by coding sequences. Gene number and gene density were 42,000–56,000 and 1 gene per 43 kbp respectively. In fact, Lai *et al.*[42] proposed that even though maize lost half of its genes after the whole genome duplication event, still the gene number is expected to be higher than the rice genome. Analysis revealed that 66% of the genome is repetitive. However, it may be a lower estimate, since there are additional repeats which are yet uncharacterized. An interesting finding reported is that gene density had a wide range of 0.5–10.7 genes per 100 kbp over a relatively even distribution and thus contradicts the earlier suggestions that a large proportion of maize genes are tightly clustered in islands.

Recently, TIGR published the results of its sequencing strategies[43]. The results are also available at www.maize.tigr.org. Using methylation filtration and high CoT sequencing strategies, they sequenced 895,731 gene-enriched sequence reads and assembled them based on sequence similarity, to generate assembled *Zea mays* (AZM) sequences to reconstruct genic regions. The TIGR database revealed that average gene size in maize is 3200 bp and the average gene density is 1 gene per 53 basis both slightly higher than those reported in previous studies[8,10,31]. The proportion of repetitive DNA is about 60–80%, comprising mainly retrotransposons. The TIGR repeat database is available at www.tigr.org/tdb/e2ki/plant repeats[44]. In the current repeat database, out of 28,249 sequences (22 Mbp), 74% constitutes transposable elements, 1% centromeric and telomeric repeats, 3% ribosomal repeats and 22% unclassified repeats. The TIGR database

Table 3. Distribution of repetitive DNA in genomic survey sequences of maize (modified from Messing *et al.*[8])

	Total BES	*Hin*dIII BES	*Eco*RI BES	*Mbo*I BES	Unfiltered sequences	Methyl filtration method	High CoT method
Total sequences	474,604	309,560	78,313	86,731	50,876	30,000	30,000
No. of base pairs	307,169,410	206,221,247	46,673,217	54,274,946	37,621,118	21,649,324	21,649,324
Per cent of genome	12.99	8.72	1.97	2.29	1.59	0.92	0.91
Class I retroelements (%)	55.60	58.39	46.77	52.58	57.73	15.70	6.55
Class II DNA transposons (%)	0.98	0.94	1.16	0.98	0.92	1.14	1.58
Simple repeats (%)	0.40	0.27	0.47	0.83	1.66	1.27	0.19
High copy number genes (%)	0.82	0.12	1.13	3.17	1.95	0.17	0.19
Other repeats (%)	0.12	0.11	0.12	0.16	0.30	0.09	0.06
Total repeats (%)	57.91	59.82	49.65	57.72	62.55	18.38	8.57

BES = Back end sequences.

reported an assembly of 298 and 52 Mbp of AZM and BAC sequences representing much of the maize gene space and amounting to 14% of the whole genome.

Future perspective

Sequencing, of the whole maize genome, even though a more challenging task than earlier plant genome sequencing initiatives, will surely be one of the worthy achievements of biologists in terms of its implications in the understanding of not only the maize genetic system, but also the whole grass family. The results will be of more interest to complex genomes like wheat, barley and sugarcane, which are currently less ideal candidates for genome sequencing. The sequencing process is progressing using methyl filtration, high CoT and other strategies. Accurate gene annonation of these sequences is a major challenge, because presence of transposable elements causes overprediction of genes[10]. One possible solution is to remove all repetitive sequences from the gene set. However, this may lead us to miss out some gene families which are mistaken for repetitive sequences, thus causing underprediction of genes. Presently, 397,000 ESTs have been clustered to 49,991 unigenes, even though these may also contain paralogous sequences[42].

One important area of focus will be the comparative analysis of various grass genomes based on synteny or conservation of gene order, besides getting an insight into genome evolution of this highly diversified group of plants. In fact Messing *et al.*[8] proposed that collinearity of maize genome to rice and sorghum may be up to 86%. Therefore, cross-referencing of various cereal genomes may help identify the positions of the genes. As far as the understanding of genome evolution is concerned, it is significant to point out that maize lies in the middle with regard to size and complexity of grass genomes is concerned. Due to common descent, the effects of various processes such as gene duplications, insertions, transpositions and other genome modifications in maize can be generalized for other cereals with greater degree of precision.

Another important aspect will be the functional characterization of repetitive sequences. At present nothing is known about the effect of complex arrays of multicopy sequences within which genes are interspersed on the gene expression. The non-random distribution of such sequences in plant genomes with ample evidences for location specificities, point to the fact, the mechanisms underlying such spatial variations are important to discover their roles in gene regulation, expression and genome evolution[29]. There is a need to characterize the unclassified portion of repetitive sequences as it might have some important clues regarding genome evolution.

1. Doebley, J., The genetics of maize evolution. *Annu. Rev. Genet.*, 2004, **38**, 37–59.
2. Sarkar, K. R., Recent developments in maize genetics. In *Maize Genetics Perspectives* (eds Sarkar, K. R., Singh, N. N. and Sachan, J. K. S.), ICAR, New Delhi, 1991, pp. 1–15.
3. Harper, L., Golubovskaya, I. and Cande, W., A bouquet of chromosomes. *J. Cell Sci.*, 2004, **117**, 4025–4032.
4. Wright, S., Bi, I., Shroeder, S., Yamasaki, M., Doebley, J., McMullen, M. and Gaut, B., The effects of artificial selection of maize genome. *Science*, 2005, **308**, 1310–1314.
5. Lisch, D., Mutator transposons. *Trends Plant Sci.*, 2002, **7**, 498–504.
6. Bennetzen, J., Chandler, V. and Schnable, P., National Science Foundation – sponsored workshop report: Maize genome sequencing project. *Plant Physiol.*, 2001, **127**, 1572–1578.
7. Parvez Sofi and Trag, A. R., Genomics in rice improvement. *Asian J. Biochem.*, 2006, **1**, 194–210.
8. Messing, J. *et al.*, Sequence composition and genome organisation of maize. *Proc. Natl. Acad. Sci. USA*, 2004, **101**, 14349–14354.
9. Gaut, B., Ennequin, M. T., Peek, A. and Sawkins, M., Maize as a model for evolution of plant nuclear genomes. *Proc. Natl. Acad. Sci. USA*, 2000, **97**, 7008–7015.
10. Haberer, G. *et al.*, Structure and architecture of maize genome. *Plant Physiol.*, 2005, **139**, 1612–1624.
11. Lapiten, N. L. V., Organisation and evolution of higher plant nuclear genomes. *Genome*, 1992, **35**, 171–181.
12. SanMiguel, P. and Bennetzen, J., Evidence that increase in maize genome size was caused by massive amplification of integenic retrotransposons. *Ann. Bot.*, 1998, **82**, 37–44.
13. Bennetzen, J. L., The contribution of retro elements to plant genome organisation, function and evolution. *Trends Microbiol.*, 1996, **4**, 347–353.

14. Bennetzen, J., SanMiguel, P., Chan, M., Tikhonov, A. and Avramova, Z., Grass genomes. *Proc. Natl. Acad. Sci. USA*, 1998, **95**, 1975–1978.
15. Walbot, V. and Petrov, D., Gene galaxies in the maize genome. *Proc. Natl. Acad. Sci. USA*, 2001, **98**, 8163–8164.
16. Ahn, S. and Tanksley, S. D., Comparative linkage maps of rice and maize genomes. *Proc. Natl. Acad. Sci. USA*, 1993, **90**, 7980–7984.
17. Gale, M. D. and Devos, K. M., Comparative genetics in grasses. *Proc. Natl. Acad. Sci. USA*, 1998, **95**, 1971–1974.
18. Freeling, M. F., Grasses as a single genetic system: A reassessment. *Plant Physiol.*, 2001, **15**, 191–197.
19. Leung, H. and An, G., Rice functional genomics. Large-scale gene discovery and application in crop improvement. *Adv. Agron.*, 2004, **82**, 55–111.
20. Sorrels, M., Rota, M. L., Bermidez, C., Kandianis, E., Greene, R. E. and Kantely, R., Comparative DNA sequences of wheat and rice genomes. *Genome Res.*, 2003, **13**, 1818–1827.
21. Paterson, A. H., Freeling, M. and Sasaki, T., Grains of knowledge: Genomics of model cereals. *Genome Res.*, 2005, **15**, 1643–1650.
22. Tikhonov, A., SanMiguel, P., Nakajima, Y., Gorentein, N., Bennetzen, J. and Avramova, Z., Collinearity and its exceptions in orthologous adh region of maize and sorghum. *Proc. Natl. Acad. Sci. USA*, 1999, **96**, 7409–7414.
23. Keller, B. and Feuillet, C., Collinearity and gene density in grass genome. *Trends Plant Sci.*, 2002, **5**, 246–251.
24. Fu, J. and Dooner, H. K., Intraspecific violation of genetic collinearity and its implications in maize. *Proc. Natl. Acad. Sci. USA*, 2002, **99**, 9573–9578.
25. Song, R., Liaca, V. and Messing, J., Mosaic organisation of orthologous sequences in grass genome. *Genome Res.*, 2002, **12**, 1549–1555.
26. Chandler, V. L. and Brendel, V., The maize genome sequencing project. *Plant Physiol.*, 2002, **130**, 1594–1597.
27. Gregory, S. *et al.*, A physical map of mouse genome. *Nature*, 2002, **418**, 743–750.
28. Feuillet, C. and Keller, B., High gene density is conserved at syntenic loci of small and large grass genomes. *Proc. Natl. Acad. Sci. USA*, 2002, **96**, 8265–8270.
29. Jorgensen, R. A., Sequencing maize: Just sample a salsa or go for the whole enchilada. *Plant Cell*, 2004, **16**, 787–788.
30. Timmermans, M. C., Brutnell, T. and Becraft, P. W., The 46th Annual Maize Genetics conference: Unlocking the secrets of maize genome. *Plant Physiol.*, 2004, **136**, 2633–2640.
31. Palmer, L. *et al.*, Maize genome sequencing by methylation filtration. *Science*, 2003, **302**, 2115–2117.
32. Sasaki, T. and Burr, B., International rice genome sequencing project. The effect to completely sequence rice genome. *Curr. Opin. Plant Biol.*, 2000, **3**, 138–141.
33. Okagaki, R. and Phillips, R. L., Maize DNA sequence strategies and genome organisation. *Genome Biol.*, 2004, **5**, 223–226.
34. Rabinowicz, P., Palmer, L., May, B., Hemann, M., Lowe, S., McComble, W. and Martienssen, R., Genes and transposons are differentially methylated in plants but not in mammals. *Genome Res.*, 2003, **13**, 2658–2664.
35. Whitelaw, C. *et al.*, Enrichment of gene coding sequences in maize by genome filtration. *Science*, 2003, **302**, 2118–2120.
36. Sasaki, T. M. *et al.*, The genome sequences and structure of rice chromosome 1. *Nature*, 2002, **420**, 312–316.
37. Bureau, T. E. and Wessler, S. R., Mobile inverted repeat elements of the tourise family are associated with genes of may cereal grasses. *Proc. Natl. Acad. Sci. USA*, 1994, **91**, 1411–1415.
38. Peterson, D. *et al.*, Integration of CoT analysis, DNA cloning and high throughput sequencing facilitates genome characterisation and gene discovery. *Genome Res.*, 2002, **12**, 795–807.
39. Yuan, Y., SanMiguel, P. and Bennetzen, J., High-CoT sequence analysis of the maize genome. *Plant J.*, 2003, **34**, 249–255.
40. Springer, N., Xu, X. and Barbazuk, W., Utility of different gene enrichment approaches toward identifying and sequencing the maize genome. *Plant Physiol.*, 2004, **136**, 3023–3033.
41. International Rice Genome Sequencing Project, 2005.
42. Lai, J. *et al.*, Characterisation of maize endosperm transcriptome and its comparison to rice genome. *Geneme Res.*, 2004, **34**, 1932–1937.
43. Chan, A. P. *et al.*, The TIGR maize database. *Nucleic Acids Res.*, 2006, **34**, 771–776.
44. Ouyang, S. and Buell, C., The TIGR repeat data bases: A collective resource for identification of repetitive sequences in plants. *Nucleic Acids Res.*, 2004, **32**, 360–363.

Received 12 September 2006; revised 23 January 2007