Rationalizing investment and effort in whole genome sequencing for harvesting applied benefits

*S. R. Bhat**

Whole genome sequencing of higher organisms was considered a major challenge about two decades ago. Thanks to the rapid developments in sequencing technologies, genome sequencing has now become faster, cheaper and technically less demanding. India participated in international genome sequencing projects of rice and tomato, and has independently initiated work on whole genome sequencing of Mesorhizobium ciceri and buffalo. With whole genome sequence data of more than 1500 organisms already available in public databases, and more added on a weekly basis, the excitement is waning. Considering that structural genomics is only the starting point for a detailed analysis of function, the investment in whole genome sequencing needs to be balanced with its intended downstream applications. In this article, we discuss the relevance of whole genome sequencing for agricultural applications and emphasize the need for urgent investment in development of appropriate tools, biological resources and human capacity in biotechnology and bioinformatics to reap full benefits.

Keywords: Microarray, molecular markers, transgenics, whole genome sequencing.

THE Human Genome Project[1] (HGP) launched in 1990 could be regarded as the starting point of the 'genomics era'. HGP was a watershed in the history of biology and had some unique features. For example, it was the first mega network project in biology involving several countries and laboratories around the globe. The project stirred debates about strategies and spurred innovation of alternative technologies, including equipment design, computation and data-handling. This accelerated the work and led to considerable contraction of time and cost. In fact, innovations in sequencing technologies have continued and are currently following the so-called Moore's Law relating to transistors[2]. These new technologies and data in turn have opened fresh opportunities and novel applications which hitherto were considered not feasible due to cost or/and time consideration. The draft human genome sequence was published in 2000 and within the next five years besides many prokaryotes, whole genome sequence data of most of the model organisms such as *Saccharomyces cerevisiae* (1996), *Caenorhabditis elegans* (1998), *Arabidopsis thaliana* (2000), fruitfly (2000), rice (2002), mouse (2002), *Neurospora crassa* (2003), rat (2004), silkworm (2004), chicken (2004), dog (2005) and chimpanzee (2005) became available. So far, nearly 1500 species have been sequenced (http://www.genomes-online.org/Large_scale_projects.htm; http://www.genome-newsnetwork.org/resources/sequenced_genomes/genome_guide_index.shtml; http://en.wikipedia.org/wiki/List_of_sequenced_eukaryotic_genomes), and a new species is added to the list almost on a weekly basis.

Following the development of the next generation sequencing (NGS) technologies[3], the whole genome sequencing projects have quickly shifted from multi-country, multi-institutional collaborative efforts into individual, laboratory-based mechanistic exercises. There are now projects aimed at sequencing individuals of a species (representing ethnic groups, varieties/strains), or comprehensive collection of biological diversity (the 10K genome project)[4], metagenomics, etc. In fact, some whole genome sequencing projects appear to have been motivated by the desire to demonstrate technological prowess. In India too, desire for such a demonstration is expressed in informal scientific discussions. Although no public announcements have been made so far, whole genome sequencing of buffalo genome is underway in India. The genome sequence information undoubtedly provides an opportunity for incisive scientific study as well as for genetic manipulation. Notwithstanding the fact that the complete gamut of applications ranging from rapid and cheap disease prognosis and diagnosis to personalized medicine that was envisaged from the availability of complete human genome sequence information has not

S. R. Bhat is in the National Research Centre on Plant Biotechnology, Pusa Campus, New Delhi 110 012, India.
*e-mail: srbhat22@rediffmail.com

Box 1. Explanation of some terms/phrases used in this article.

1. *Next generation sequencing technologies:* Traditional automated sequencing technology used Sanger's di-deoxy chain termination chemistry. In recent years, several novel sequencing technologies based on alternative chemistry and employing automated micro-devises have been developed. These technologies which need less preparatory work have radically transformed genome sequencing, making it faster and cheaper.

2. *Molecular markers/marker-assisted selection:* Molecular markers refer to DNA, protein, isozyme or any other biomolecules which show Mendelian inheritance. In recent years DNA-based markers have become popular as they are the true units of inheritance and are not affected by the environment. By studying co-segregation of markers and traits over generations, traits could be tagged (linked) with specific molecular markers. Thus, based on the presence or absence of specific markers one could predict the phenotype of the individual. In this way one could make selections for a desired trait at the seedling stage. Further, by following several markers simultaneously, one can identify rare desirable recombinants from a segregating population.

3. *Synteny/crop circle:* The linear order of genes on chromosomes (also called linkage groups) is unique for each species. During species evolution, this gene order is largely conserved. Such a conserved order of genes on chromosomes is called synteny. The study of gene order in closely related species has revealed that species evolution has occurred through duplication, inversion and translocation of parts of chromosomes. Despite large variation in genome size and chromosome number, members of Gramineae display striking synteny. The synteny relationship among chromosomes of different cereal/millet crop species depicted as circles is termed crop circles.

4. *GWA analysis/linkage disequilibrium mapping:* In GWA analysis, a large number of individuals from diverse populations is genotyped for a number of molecular markers spread throughout the genome using high-throughput techniques. These individuals are also phenotyped for specific traits of interest. In the next step, using statistical tools, markers showing strong association with traits are identified. The markers that are closely linked to the trait will show linkage disequilibrium among populations. This is particularly relevant to crops/species where standard genetic analysis is not feasible due to various reasons.

5. *QTLs:* A majority of traits of economic significance (flowering time, height, grain number, yield, etc.) are controlled by polygenes, each with a small effect. Such genes cannot be accurately identified through conventional genetic analysis. The advent of molecular marker techniques has made their identification possible. By following the segregation of markers and the trait, one can identify marker(s) whose presence or absence results in a significant change in the phenotypic value of the individual. This indicates that the gene regulating/influencing the trait is located in the vicinity of the marker and the marker is called QTL.

6. *Microarray/gene chips:* Gene chip refers to glass, silicon or any other solid support onto which short, defined DNA sequences are covalently attached. Multiple copies of specific DNA sequences are spotted (immobilized) on the support and several thousand spots each carrying different species of DNA are arranged in an orderly manner to create an array. Such DNA-chips can be used to query for the presence of transcripts or closely matching DNA sequences in a given sample through Southern hybridization. Gene chips thus allow large-scale assessment of gene presence/expression in a single reaction.

7. *c-DNA library:* A collection of complementary DNA clones obtained following amplification through RT–PCR using mRNA as template. They represent the genes that are transcribed in that tissue.

yet been fully realized, the broad proof-of-principle has been well demonstrated. Hence, there is general consensus that whole genome sequence information of organisms is invaluable for dealing with them appropriately (depending on whether the organism is beneficial or harmful).

Given the limited public funds for research and considering the fact that a large number of institutions and nations have embarked on various whole genome sequencing projects, should we undertake a similar exercise? If so, what should be our priority? Are there alternative ways to reach our ultimate goals? In this article I shall attempt to answer these questions. Owing to familiarity, I will use examples from agricultural biotechnology to illustrate my points. Box 1 provides an explanation of some of the terms/phrases used in this article.

Deciding priority

In any discussion about the choice of an organism for whole genome sequencing, people come forward with

different names with their own justification for the choice. Usually the justifications fall under one or more of the following categories.

- Economic importance: The crop/animal is economically most valuable as a major contributor to national food, feed or nutritional security, export, and livelihood of the poor and marginal farmer.
- Nativity: The species is native and relevant to India, and is not likely to be chosen by anyone else. For instance, mango is a major fruit crop native to India, or large-scale rearing of lac insect is found only in India.
- Unique features of economic significance: Cotton is the only crop plant that produces fibre on seed, or curcuma is an important medicinal plant, or mangroves thrive under special conditions.
- Academic curiosity: The organism is a representative of a special biological category.

India joined the international effort on rice genome sequencing primarily for the first reason stated above. Subsequently, India also participated in international tomato genome sequencing project. Buffalo is one species which meets the above criteria and has been rightly chosen for whole genome sequencing. However, other candidates are not as easily identifiable. Although economic reason is easily understood, this alone is insufficient and need not override other considerations. In particular, with the availability of complete genome sequence of several organisms, there is diminishing return on investments in sequencing new organisms. Instead of seeking justification on the above criteria, it is more pertinent to answer the questions: 'What difference would the sequence information make towards management or improvement of the crop/animal or its product?' and 'Are we adequately equipped to take benefit of the genome data?'. In case of agriculturally important organisms, genome sequence information is primarily sought for breeding improved varieties through diversity studies, association analysis, fingerprinting, etc. The relevance of whole genome sequence data for agricultural applications is depicted in Figure 1.

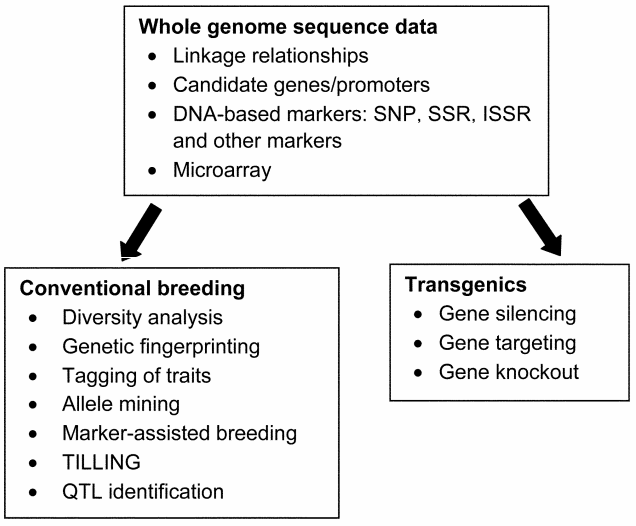

Figure 1. Applications of genome sequence data in agriculture.

Applications of sequence data

The complete genome sequence of an organism, say rice, does not mean that all accessions of rice contain the same genetic constitution (or nucleotide sequence). Instead, each accession will contain nearly the same set of genes but with varying changes in nucleotide sequence (including substitutions, insertions, deletions and structural variations). Indeed, sequencing of multiple human individuals has revealed significant additional genomic sequences not present in the draft genome sequence[5,6]. Genome sequence data (structural genomics) does not automatically yield functions of all genes contained therein. In fact, complete functional characterization of all genes of even simple organisms is still incomplete. Given the complex interactions among genes (or their products), which ultimately result in a phenotype, assigning function to each of the genes will be a major challenge. Thus genome sequence information is the starting point for incisive study.

The two options of biotechnological approaches for breeding improvement are: (i) molecular marker-based selection and (ii) transgenic manipulation. Complete genome sequence information is helpful but not essential for the above applications.

Molecular markers

A variety of DNA-based molecular markers have been devised during the past 2–3 decades. These markers are physical units of inheritance and show typical Mendelian segregation. Hence, these markers are used to construct detailed linkage maps to tag traits of interest, assess genetic diversity and obtain genetic fingerprints of varieties, strains or accessions. Further, molecular markers have made possible the identification of genes governing complex traits (quantitative trait loci, QTLs). By following the markers linked to traits of interest (for example, disease resistance, stress tolerance or seed yield), one can make indirect selection for the trait without phenotyping. Such marker-assisted selection not only reduces the time required for breeding new varieties, but also makes the process more efficient and accurate. Therefore, one of the main applications of genome sequence information is in the identification of markers. Among the molecular markers, microsatellite markers are popular due to their

reliability and ease of use. Availability of complete genome sequence information is undoubtedly useful in identifying densely distributed simple sequence repeat (SSR) markers on all linkage groups. Complete genome sequence also provides an opportunity for identifying another type of markers called the single nucleotide polymorphism (SNP) markers. These SNP markers are particularly useful in unearthing functionally relevant genetic variants (allele mining) for use in breeding improved varieties. New high-throughput method of SNP detection based on microarrays has now been developed and complete sequence data help in devising such chips for large-scale genotyping and related applications.

Genome sequence information is also critical for isolation of specific gene mutations. Using primers for amplification of the desired gene sequences, one can rapidly screen a large number of individuals subjected to mutagenic treatment to isolate rare mutations in the concerned gene. This technology called TILLING (Targeting Induced Local Lesions in Genomes) is not only used in functional genomics, but also in breeding improved varieties.

Comparative analysis of genome sequences of different organisms has revealed a high degree of conservation of sequence order (synteny) among phylogenetically related organisms. This is best illustrated in the so-called crop circles of members of the Gramineae. Hence, sequence information from one organism could be profitably applied to closely related organisms. Thus whole genome sequence of an organism is less important for the purpose of identifying molecular markers if a related organism is completely sequenced. For instance, a detailed molecular map of *Brassica juncea* could be constructed based on *Arabidopsis thaliana* sequence[7]. Cross-transferability of markers is also demonstrated among members of the Gramineae[8,9] and Asteraceae[10]. Further, alternative approaches are available to identify microsatellite and other markers for use in routine breeding applications[11]. Similarly, microarray chips have been found to work well in closely related organisms[12,13]. Therefore, if the organism chosen for sequencing is phylogenetically distinct, it could serve a wider purpose.

The second aspect that may be worth considering is the nature of the crop and its breeding behaviour. In annual crops amenable to controlled breeding, genetic maps could be readily constructed and various marker-based techniques could be readily applied. Also, a variety of markers could be developed and used. However, in many perennial horticultural crops, breeding is complicated and genetic information about traits is scanty. In such cases, meaningful information can be gathered only through large-scale genotyping by high-throughput methods using naturally available variation. Application of molecular markers would be crucial for breeding improvement of such crops. Genome sequencing will be particularly relevant for construction of linkage maps and for tagging of traits using genome-wide association (GWA) analysis. For example, mango is a crop that would qualify for consideration for whole genome sequencing.

Another point often stressed is the uniqueness of the organism. For instance, genome sequencing of cotton is advocated on the ground that it is the only cultivated species that produces fibre on the seed. Such unique traits are also cited for advocating sequencing of various medicinal plants. However, a unique trait need not necessarily justify complete genome sequencing. For instance, cotton plant is likely to share with other dicots many genes related to growth and metabolism. If the uniqueness relates to fibre development on seed, one could identify relevant genes through c-DNA library of developing ovules. This will be economical and quicker compared to whole genome sequencing, and will uncover the genes governing this unique trait. Considering the fact that most of the genes involved in basic plant metabolism and growth are conserved among plants, information about genes governing agronomic traits (like growth habit, flowering, fruit development, stress tolerance, etc.) could be obtained from a comparative study of related model plants following functional genomics approaches.

Transgenics

Direct genetic manipulation through transformation is a powerful breeding approach that has become feasible in recent years. By introducing genes and appropriate regulatory sequences (promoters) into the host, transformation alters host metabolism to confer new traits. Such transgenic crops bearing foreign genes (from bacteria, virus or other sources) and displaying novel traits such as herbicide tolerance, pest resistance or novel quality have been widely commercialized[14]. Transformation technology is also widely used for functional genomics. How does genome sequencing help in transgenic breeding? As stated above, knowledge of host metabolism, genes and promoters governing the pathway is critical for transgenic manipulation. Complete genome sequence will help gain such information. However, in most cases, the genes and promoters needed for manipulating the trait are sourced from other organisms (bacteria/fungi or model plants). Hence, genome sequence information of the host is not important for transgenic-based improvement. In recent years, gene replacement and gene-knockout technologies have been devised that permit accurate genetic manipulation. This is particularly significant because commonly used plant transformation methods insert transgenes at random locations in the host chromosome and thereby introduce event-associated variation in performance. Genome sequence information will facilitate targeted insertion of transgenes and thereby avoid event-based biosafety assessment. Once again, whole genome sequence information is not essential for this purpose. Thus

on the whole, transgenic technology is less likely to be constrained by the nonavailability of whole genome sequence information.

Sequencing technology

With the arrival of NGS technologies, whole genome sequencing has become cheaper and faster. There are different technology platforms with varying costs, efficiencies and accuracies[3]. This adds another dimension to the genome sequencing debate regarding the choice of technology platform for sequencing. While re-sequencing could be easily accomplished with different approaches, *de novo* assemblies could prove tricky. If the aim is to obtain a rough draft sequence for facilitating molecular breeding, one should opt for less stringent methods and save on cost. On the other hand, if the aim is to obtain a high quality, accurate sequence for basic studies, cost considerations should be secondary.

Preparedness to avail the benefits

Mere availability of sequence information is not a guarantee that benefits will flow. We will need to mine relevant alleles/QTLs from the germplasm to incorporate into appropriate varieties, or to assemble improved genotypes. Therefore, a critical examination of the availability of the following would be useful while choosing the species for sequencing.

- Diverse germplasm collection for identifying new alleles necessary for breeding.
- Appropriate mapping populations to tag traits.
- Precise, high-throughput phenotyping facilities/ screening techniques.
- Trained manpower to undertake work.

Depending on the crop and specific breeding goals, the downstream requirements for exploiting genome information will vary. Although India has made good investment in collecting and conserving germplasm resources of species relevant to agriculture, including crops, animals and of late, microbes and insects, the same have not been well characterized. Accurate characterization of germplasm for various phenotypic traits such as tolerance to various biotic and abiotic stresses, yield, etc. would require sophisticated controlled conditions. Phenomic facilities for germplasm characterization are lacking, and only now some effort is being initiated towards addressing this deficiency. Similarly, development of mapping populations for specific traits in different crops is still in its infancy.

In India, our preparedness for utilizing genome information is debatable. Trained human resource is limited and laboratory facilities for large-scale genomics are lacking. Arguably, our weakest link in the whole process is our bioinformatics capabilities. Whole genome sequencing effort cannot fructify without strong bioinformatics support. Outsourcing such critical components could prove counter-productive. Hence, human capacity building in this area needs immediate attention.

In summary, while whole genome sequence information of crop/animal species is invaluable for application in crop and animal improvement, investment on genome sequencing project needs careful consideration of various intended downstream uses and our preparedness to harness the information. Therefore, both the aspects should be jointly considered for funding to realize the ultimate goals.

1. Watson, J. D., The human genome project: past, present, and future. *Science*, 1990, **248**, 44–49.
2. Moore, G. E., Cramming more components onto integrated circuits. *Electronics*, 1965, **38**, 114–117.
3. Metzke, M. L., Sequencing technologies – the next generation. *Nat. Rev. Genet.*, 2010, **11**, 31–46.
4. Genome 10K Community of Scientists, Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. *J. Hered.*, 2009, **100**, 659–674.
5. Wang *et al*., Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 2008, **456**, 60–65.
6. Kidd *et al*., Characterization of missing human genome sequences and copy-number polymorphic insertions. *Nature Methods*, 2010, **7**, 365–371.
7. Punjabi, P. *et al*., Comparative mapping of *Brassica juncea* and *Arabidopsis thaliana* using Intron Polymorphism (IP) markers: homoeologous relationships, diversification and evolution of the A, B and C *Brassica* genomes. *BMC Genomics*, 2008, **9**, 113.
8. Sim, S.-U., Yu, J.-K., Jo, Y.-K., Sorrells, M. E. and Jung, G., Transferability of cereal EST–SSR markers to ryegrass. *Genome*, 2009, **52**, 431–437.
9. Yadav, O. P., Mitchell, S. E., Fulton, T. M. and Kresovich, S., Transferring molecular markers from sorghum, rice and other cereals to pearl millet and identifying polymorphic markers. *J. SAT Agric. Res.*, 2008, **6**, 1–4.
10. García-Moreno, M. J., Velasco, L. and Pérez-Vich, B., Transferability of non-genic microsatellite and gene-based sunflower markers to safflower. *Euphytica*, 2010, **175**, 145–150.
11. Mondini, L., Noorani, A. and Pagnotta, N. A., Assessing plant genetic diversity by molecular tools. *Diversity*, 2009, **1**, 19–35.
12. Horvath, D. P., Schaffer, R., West, M. and Wisman, B., *Arabidopsis* microarrays identify conserved and differentially expressed genes involved in shoot growth and development from distantly related plant species. *Plant J.*, 2003, **34**, 125–134.
13. Bar-Or, C., Czosnek, H. and Koltai, H., Cross-species microarray hybridizations: a developing tool for studying species diversity. *Trends Genet.*, 2007, **23**, 200–207.
14. James, C., Global status of commercialized biotech/GM crops. *ISAAA Brief*, ISAAA, Ithaca, NY, 2009, 41.

ACKNOWLEDGEMENTS. I thank Prof. V. L. Chopra and Dr R. Srinivasan, National Research Centre on Plant Biotechnology, New Delhi for suggestions and comments on the manuscript.

Received 23 September 2010; accepted 24 March 2011