Revisiting the decoded genomes to promptly reveal their genomic perspectives

Shouvik Das[1], Deepak Bajaj[1], S. Gopala Krishnan[2], Ashok K. Singh[2] and Swarup K. Parida[1,]*

[1]National Institute of Plant Genome Research, Aruna Asaf Ali Marg, New Delhi 110 067, India
[2]Division of Genetics, Rice Section, Indian Agricultural Research Institute, New Delhi 110 012, India

Post *Arabidopsis thaliana*, 55 genomes comprising 49 different plant species have been decoded by use of clone-by-clone, whole genome shotgun and next-generation sequencing approaches. The structural outcomes of these sequenced genomes shed light on their genomic constitution, particularly the way genes, transposable elements and genetic markers are organized within the genomes. The functional outcomes provide a brief account of specific phenotypic trait characteristics of crop genomes by digging deep into the genetic make-up of transcription factors, regulatory elements and gene families governing multiple agronomic traits in these crop plants. The comparative and evolutionary outcomes deduce the genetic basis of biological diversity and basic process of genome evolution by analysing the syntenic relationships among genes and genomes/chromosomes of the sequenced crop plants. Therefore, a revisit to published genome sequence landmarks in 30 major cultivated food crops constituting major groups (cereals, legumes, vegetables, fruits, oilseeds and fibres) would significantly assist us to gain a detailed insight into their genome organization and dissect the structural, functional, comparative and evolutionary intricacies for identifying species- and lineage-specific genes controlling multiple characteristics in crop plants. The essential inputs obtained will be helpful in devising efficient strategies to develop high-yielding climate-ready crop varieties through translational genomics.

Keywords: Decoding, food crops, plant genome, translational genomics.

GENOMICS-ASSISTED breeding and transgenics are currently the most sought after strategies as far as genetic improvement pertaining to crop plants is concerned. To propel them further, the identification and implementation of innovative molecular diagnostic tools like sequence-based DNA markers as well as structurally and functionally well-characterized genes/QTLs (quantitative trait loci) and regulatory sequences (transcription factors) associated with specific plant characteristics, including disease resistance, stress tolerance, improved productivity and quality traits seem quite relevant. This could be effectively achieved by decoding all the indispensable structural, functional, comparative and evolutionary information encoded by the DNA through whole genome sequencing of crop plants. *Arabidopsis* remains the first plant and *indica*/*japonica* rice (*Oryza sativa*) the first crop species to be sequenced using the first-generation Sanger sequencing (FGS)-based clone-by-clone (CBC) and whole genome shotgun (WGS) approaches during the year 2000 and 2005 respectively[1,2]. The recent advancement of genome analysis driven by the development of high-throughput next-generation sequencing (NGS) platforms and innovative computational genomics tools has accelerated the whole genome sequencing programmes for diverse crop species with small diploid and large polypoid genomes. The potential of available NGS platforms such as long sequence read (600–700 bp)-based Roche 454 Pyrosequencer, and short sequence read (35–150 bp)-based Applied Biosystems SOLiD and Illumina Solexa Genome Analyzer in complete and draft genome sequencing of multiple crop plants has been well-understood. Among these NGS platforms, the 454 Pyrosequencer and Illumina Solexa Genome Analyzer are most widely used for complete as well as draft genome sequencing of plant species. Remarkably, the hybrid sequencing approaches that combined the traditional Sanger sequencing with NGS technologies are gaining popularity for sequencing both small diploid and large polyploid plant genomes. All sequencing strategies exploited until now for whole genome sequencing of crop plants can be broadly classified into four major approaches, including CBC–FGS, WGS–FGS, WGS–NGS and hybrid sequencing (CBC–FGS/WGS–FGS with WGS–NGS).

Utilizing the above, 55 plant genomes comprising 49 different plant species have been sequenced to date[3,4]. This includes 30 major food crops representing 28 cultivated plant species having higher worldwide average productivity[5]. Based on their global productivity and economic importance, these 28 sequenced plant genomes can be easily categorized into five different major food crop groups, namely cereals, legumes, vegetables, fruits and others (oilseeds, fibres and millets). The plant genome sequencing projects have generated colossal genomic and transcriptomic sequence resources, structurally and

*For correspondence. (e-mail: swarup@nipgr.ac.in) functionally annotated protein-coding and non-protein-coding genes, transcription factors and sequence-based molecular markers like simple sequence repeats (SSRs) and single nucleotide polymorphisms (SNPs). Moreover, the sequenced plant genomes put forth novel and significant biological insight into their genomic constitution as well as implications towards structural, functional, comparative genomic analyses and phylogenetics during their domestication. With sequencing of a diverse array of plant genomes, the current trend is inclined more towards exploration of novel structural, functional, comparative and evolutionary aspects of genome biology and genomic features for understanding the genome structure, domestication and complexity of individual crop plants. In this context, the integration/comparison and annotation of structural, functional, comparative and evolutionary information generated from the sequencing data of various plant genomes will provide relevant biological insight which will assist us to rapidly derive conclusive hypothesis. This in turn will greatly benefit plant genomics researchers, biologist and molecular breeders in their quest for better strategies aimed at crop genetic improvement.

To the best of our knowledge, no comprehensive comparative study involving structural, functional, comparative and evolutionary aspects has been undertaken that covers 30 major cultivated food crops culminating into 5 different major groups, namely cereals (rice, wheat, maize, sorghum and barley), legumes (*Lotus*, soybean, *Medicago*, pigeon pea, chickpea and common bean), vegetables (cucumber, potato, tomato, watermelon, melon, hot pepper and radish), fruits (grape, papaya, apple, strawberry, banana and sweet orange) and others (fibres: cotton and flax; millet: foxtail millet and oilseeds/ vegetables: sesame, *Brassica rapa* and *B. oleracea*). Therefore, revisiting all complete/draft genome sequence landmarks published hitherto in cultivated food crops constituting the aforesaid five major groups would significantly assist us to gain deeper insight into their genome organization and dissect the structural, functional, comparative and evolutionary intricacies for identifying species- and lineage-specific genes controlling multiple characteristics in crop plants. For instance, the sequencing of hexaploid wheat genome gave clues regarding the polyploidization and expansion of gene families comprising 200 genes involved in energy harvesting, metabolism and growth associated with high carbohydrate content in grain and crop productivity[6]. Likewise, the sorghum genome sequencing revealed retrotransposon accumulation in its recombinationally recalcitrant heterochromatic sequenced regions which possibly contributes to a larger genome size of sorghum (~75%) in contrast to rice[7]. Higher drought tolerance in sorghum than other cereal species is probably due to recent gene and microRNA duplications across its genome. The maize genome sequencing inferred the abundance (~85% of genome) of hundreds of transposable element families contributing to complexity and diversity of its genome[8]. The genome sequence of *Medicago* unravelled evolution of endosymbiotic–rhizobial nitrogen fixation by sub- and/or neofunctionalization of genes having specialized role in nodulation during its ancient whole genome duplication (WGD) event about 58 million years (Myr) ago[9], which led to the identification of a lineage-specific nodulin gene controlling nodulation in legumes. The tomato genome sequencing provided proper understanding of fleshy fruit evolution due to its genome triplications causing neo-functionalization of genes regulating fruit characteristics like colour and fleshiness, which are otherwise absent in other sequenced Solanaceous crop species like potato[10]. The autotetraploid potato genome hinted at gene family expansion, tissue-specific expression and recruitment of genes by novel pathways that facilitate the evolution of its tuber development[11]. The much needed inputs, including functionally relevant molecular tags (markers, genes, QTLs and alleles) regulating traits of agricultural importance collectively acquired by correlating all the aforementioned studies can prove useful for genetic enhancement of major food crops coupled with higher yield and stress tolerance through translational genomics approaches.

Therefore, the present study made an effort to revisit and compare all the genome sequence landmarks successfully accomplished in cultivated food crops encompassing five major groups to delve deeper into the structural, functional, comparative and evolutionary make-up of plant genomes. Further, the implications of species- and lineage-specific genes governing useful agronomic traits deciphered from the aforementioned comprehensive studies on sequenced genomes of food crops have been discussed briefly with an ultimate objective of genetic enhancement of the diverse crop plants through translational genomics.

Structural perspectives of decoded crop plant genomes

The genome sequences of five major cereal crops, namely rice, wheat, maize, sorghum and barley have been deciphered by CBC–FGS, WGS–FGS and WGS–FGS–NGS approaches. In rice, the genomes of its four species/ subspecies, *Oryza sativa* L. ssp. *indica* (cv. 93-11), *O. sativa* L. ssp. *japonica* (cv. Nipponbare), *O. sativa* L. ssp. *aus* (cv. Kasalath) and *O. brachyantha* have been sequenced till date using CBC–FGS, WGS–FGS and WGS–FGS–NGS approaches respectively[2,12–14]. In wheat, the genomes of two of its cultivated species, *Triticum aestivum* (cv. CS 42) and *T. urartu* (cv. G1812/PL428108) as well as one of its wild relative *Aegilops tauschii* (cv. AL8178) have been sequenced by WGS–NGS approaches[6,15–17]. The genomes of maize (*Zea mays* L. cv. B73), sorghum (*Sorghum bicolor* cv. Moench BI623) and

Table 1. Structural and functional outcomes of whole genome sequencing of five cereal crop plants

	Cereal species/genotypes-sequenced									
	Rice				Wheat			Maize	Sorghum	Barley
Characteristics	*Oryza sativa* (*indica* cv. 93-11)	*O. sativa* (*japonica* cv. Nipponbare)	*O. brachyantha*	*O. sativa* (cv. Kasalath)	*Triticum aestivum* (cv. CS 42)	*Aegilops tauschii* (cv. AL8178)	*T. urartu* (cv. G1812/ PL428108)	*Zea mays* (cv. B73)	*Sorghum bicolor* (cv. Moench BI623)	*Hordeum vulgare* (cv. Morex)
Genome size sequenced	~362 Mb	~370 Mb	~261 Mb	~330.5 Mb	~3.8 Gb	~4.23 Gb	~4.66 Gb	~2.1 Gb	~700 Mb	~4.98 Gb
Estimated genome size	~466 Mb	~389 Mb	~300 Mb	~362 Mb	~17 Gb	~4.36 Gb	~4.94 Gb	~2.3 Gb	~739 Mb	~5.1 Gb
Approaches used for sequencing	WGS–FGS	CBC–FGS	WGS–FGS–NGS	NGS	WGS–NGS	WGS–NGS	WGS–NGS	CBC–FGS	WGS–FGS	WGS–FGS–NGS
Chromosomes sequenced	12	12	12	12	7	7	7	10	10	7
Number of protein-coding genes	~51,000	37,544	32,038	NA	94,000–96,000	43,150	34,879	32,000	27,640	26,159
Gene density (number of genes/Mb)	140.88	101.47	122.75	NA	5.58	10.20	7.48	15.23	39.48	5.25
Transposable elements (% of genome)	24.9	35	29.2	NA	NA	65.9	66.88	85	55	84
Number of miRNA	NA	158	NA	NA	NA	159	24	150	144	NA
Number of genes controlling traits of agronomic importance	600 RGAs and 1306 TFs	535 RGAs, 455 *CytP450* and >1000 TFs	>1000 RGAs		>200 energy-harvesting genes	878 RGAs, 216 cold-related genes, 485 *CytP450* and 1489 TFs	593 RGAs and >100 *CytP450*	129 RGAs and 261 *CytP450*	211 RGAs and 365 *CytP450*	191 RGAs
Molecular markers	48,351 SSRs	18,828 SSRs and 80,127 SNPs	NA	2,787,250 SNPs and 7393 InDels between Kasalath and Nipponbare; 2,216,251 SNPs and 3780 InDels between Kasalath and 93–11	>132,000 SNPs	860,126 SSRs and 711,907 SNPs	166,309 SSRs and 2,989,540 SNPs	>3.3 million SNPs and InDels	71,000 SSRs and 90,000 SNPs	>15 million SNPs

NA, Not available; TFs, Transcription factors; RGAs, Resistance gene analogues; SNPs, Single nucleotide polymorphisms; SSRs, Simple sequence repeats; InDel, Insertions/deletions; CBC, Clone-by-clone; FGS, First generation sequencing; WGS, Whole genome shotgun and NGS, Next-generation sequencing.

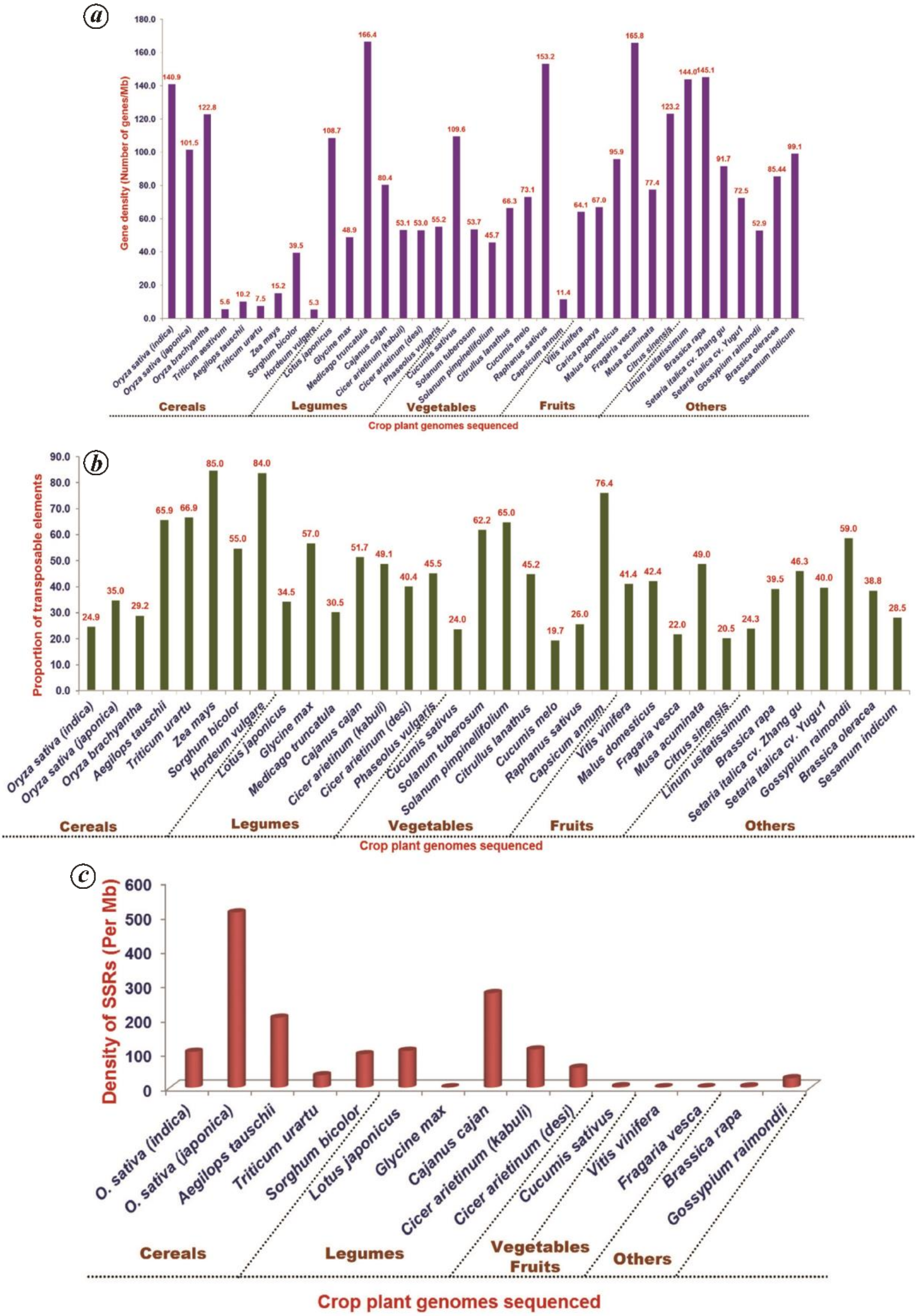

Figure 1. A brief overview of structural outcomes of sequenced crop plant genomes. ***a***, Gene density (number of genes/Mb); ***b***, Proportionate genomic distribution of transposable elements; ***c***, SSR (simple sequence repeat) density (number of SSRs/Mb) estimates of the sequenced plant genomes.

barley (*Hordeum vulgare* cv. Morex) have also been decoded using CBC–FGS, WGS–FGS and WGS–FGS–NGS approaches[7,8,18]. Table 1 and Figure 1 provided detailed information, including genome size sequenced and relative density of protein coding genes, transcription factors, disease resistance-related genes, transposable elements and sequence-based robust genetic markers which have been deciphered from five decoded cereal genomes. A comparative study of these outcomes was performed to unveil the structural perspectives of five sequenced cereal genomes.

Among the four species/subspecies of rice sequenced, *O. sativa* (*indica* cv. 93-11) has the highest gene density (140.88/Mb (mega base)) with a larger genome size (~466 Mb) and relatively less amount of transposable elements (24.9% of genome) in its genome (Figure 1 *a* and Table 1). Two copies of *Tos17*, an endogenous *Copia*-like retrotransposon insertion sites, are found to be abundant (11487) in *japonica* genome[2]. In contrast, the mutator-like retrotransposon elements are most frequent in *indica* genome[13]. The small genome (~300 Mb) of *O. brachyantha* has comparatively higher gene density (122.75/Mb) with relatively lower proportion of transposable elements (29.2% of genome). This implies that more compact genome of *O. brachyantha* is due to the low activity of long-terminal repeat (LTR) retrotransposons and massive internal deletion of ancient LTR elements.

A more recently sequenced bread wheat genome contains a total of 124,201 protein coding genes, which is higher compared to previously reported genome of *T. aestivum* (94,000–96,000 genes)[6,17]. The B genome of recently sequenced bread wheat genome contains the highest number of protein coding genes (44,523) followed by A (40,253) and D (39,425) genomes. In contrast, the A genome of previously reported genome sequence of *T. aestivum* contains lower number of genes (28,000) compared to B (38,000) and D (36,000) genomes. Both earlier and currently sequenced genomes of *T. aestivum* contain more than 75% transposable elements of the total genome size sequenced. The class I retrotransposon DNA elements are found to be more abundant in A genome chromosomes relative to B or D genome chromosomes (A > B > D), whereas class II retrotransposons are found to occur in a reverse manner (D > B > A).

Maize genome contains the highest percentage of transposable elements (85% of the genome) among all the five cereal genomes sequenced, resulting in a relatively lower gene density (15.23/Mb) in its genome. LTR retrotransposons which compose 75% of the maize genome exhibit family-specific and non-uniform distribution along the five chromosomes. For instance, *Copia*-like LTR elements are overrepresented in gene-rich euchromatic region, whereas *Gypsy*-like elements are abundant in gene-poor heterochromatic regions. DNA transposable elements make up 8.6% of the maize genome. The most complex of these superfamilies are mutator-like elements carrying fragments of 226 nuclear genes. Except *CACTA*, most of the maize DNA transposable elements are enriched in gene-rich recombinationally active chromosome ends. Sorghum genome contains lesser proportion of transposable elements (55% of genome) and relatively more gene density (39.48/Mb) compared to maize genome (Figure 1 *a*, *b* and Table 1). In barley genome, LTR retrotransposons and *Gypsy*-like elements are 1.5-fold more abundant than *Copia* superfamily compared to those documented in rice.

The genomes of six legume species, namely *Lotus japonicas* (cv. Miyakojima MG-20), *Glycine max* (cv. William82), *Medicago truncatula* (cv. A17), *Cajanus cajan* (cv. ICPL87119/Asha), *Cicer arietinum* (*kabuli* cv. CDC Frontier and *desi* cv. ICC4958), and *Phaseolus vulgaris* (cv. G19833) have been sequenced using WGS–FGS, WGS–FGS–NGS, WGS–NGS and CBC–FGS–WGS–NGS approaches[9,19–25]. Table 2 and Figure 1 show the significant outcomes, specifically the relative distribution frequency of protein-coding genes, transcription factors and transposable elements obtained from these decoded legume genome sequences. A comprehensive study on characteristic features of genomic constitution among these sequenced legume genomes led to uncover their structural perspectives. The draft genome sequence assembly of six leguminous crops reveals that *Medicago* genome contains the highest number of protein coding genes (62,388) with a very high gene density (166.36/Mb). Due to experiencing recent WGD (13 Myr), the soybean genome has the largest size (950 Mb) with more abundance of transposable elements (57% of genome) and relatively low gene density (48.87/Mb) (Figure 1 *a*, *b* and Table 2). The transposable elements present in soybean genome include Tc1/Mariner, haT, Mutator, PIF/Harbinger, pong, *CACTA* and *Helitrons*. Also, 2668 LTR retrotransposons distributed among 165 families, including 65 *Ty-copia* and 78 *Ty3-gypsy* elements have been identified primarily in common bean genome.

Among vegetables, the genomes of cucumber (*Cucumis sativus* cv. Chinese long-9930), potato (*Solanum tuberosum* cv. Phureja/DM1-3516 R44), tomato (*Solanum pimpinellifolium* cv. LA1589 and *S. lycopersicum* cv. Heinz1706), watermelon (*Citrullus lanathus* cv. 97103), melon (*Cucumis melo* cv. DHL92), radish (*Raphanus sativus* cv. Aokubi) and hot pepper (*Capsicum annum* cv. CM334) have been sequenced using WGS–NGS, WGS–FGS–NGS and CBC–FGS–WGS–NGS approaches[10,26–31]. Certain relevant aspects, including genome size sequenced and relative density of protein-coding genes, transcription factors, disease resistance-related genes, transposable elements and molecular markers inferred from these sequenced genomes have been summarized in Table S1 (see Supplementary Material online) and Figure 1. A comparative study on characteristic genomic features among these sequenced vegetable

Table 2. Structural and functional outcomes of whole genome sequencing of six legume crop plants

	Legume species/genotypes sequenced						
	Lotus	Soybean	*Medicago*	Pigeon pea	Chickpea		Common bean
Characteristics	*Lotus japonicas* (cv. Miyakojima MG-20)	*Glycine max* (cv. William 82)	*Medicago truncatula* (cv. A17)	*Cajanus cajan* (cv. ICPL87119/ Asha)	*Cicer arietinum* (cv. CDC Frontier)	*C. arietinum* (cv.ICC 4958)	*Phaseolus vulgaris* (cv. G19833)
Genome size sequenced	~315.1 Mb	~950 Mb	~375 Mb	~605.78 Mb	~532 Mb	~520 Mb	~472.5 Mb
Estimated genome size	~472 Mb	~1.1 Gb	~465 Mb	~833 Mb	~738 Mb	~740 Mb	~587 Mb
Approaches used for sequencing	WGS–FGS	WGS–FGS	CBC–FGS–WGS–NGS	WGS–FGS–NGS	WGS–NGS	WGS–NGS	WGS–NGS
Chromosomes sequenced	6	20	8	11	8	8	11
Number of protein-coding genes	34,245	46,430	62,388	48,680	28,269	27,571	27,197
Gene density (number of genes/Mb)	108.67	48.87	166.36	80.359	53.13	53.02	55.17
Transposable elements (% of genome)	34.5	57	30.5	51.67	49.1	40.4	45.52
Number of miRNA	1312	85	395	862	420	60	NA
Number of genes controlling traits of agronomic importance	229 RGAs, 1481 TFs, 1267 protein kinases, 1310 transporters and 313 *CytP450*	506 RGAs, 5671 TFs, 28 nodulin genes and 109 drought-responsive genes	764 RGAs, 3692 TFs and 593 nodule cysteine-rich peptide	406 RGAs and 111 drought-responsive genes	187 RGAs	119 RGAs,1680 TFs and 89 nodulin genes	376 disease resistance-related genes and 15 candidate genes associated with seed weight
Molecular markers	33,730 SSRs	874 SSRs and 4991 SNPs	>3 million SNPs	166,309 SSRs and 2,989,540 SNPs	81,845 SSRs and 76,084 SNPs	30,000 SSRs and 60,000 SNPs	8,890,318 SNPs–Mesoamerican subpopulation and 139,405 SNPs – Andean subpopulation

genomes reveals diverse salient attributes regarding structural perspectives of these genomes. From the draft genome sequences of vegetables (Solanaceae and Cucurbitaceae) it can be inferred that among the Solanaceae family, potato genome contains a large number of protein coding genes (39,031) with gene density of 53.68/Mb, which is higher compared to that of tomato (Figure 1 *a* and Table S1 (see Supplementary Material online)). Among the retrotransposons, LTRs are abundant in tomato and potato genomes. Though these genomes have the same ploidy level and comparable size, the gene density in tomato genome is much less compared to potato, implying the possibility of amplification of transposable elements in former genome. Within the Cucurbitaceae family, cucumber has the smallest genome size (243.5 Mb) with a very high gene density (109.57/Mb) and relatively less abundance of transposable elements (10.4% of genome; Figure 1 *a*, *b* and Table S1 (see Supplementary Material online)). Class I transposable elements, including *Copia* and *Gypsy* types are found to be the most frequent repetitive sequences present in radish and watermelon genomes. Transposon families, including *CACTA*, MULE and PIF/harbinger have amplified significantly in melon lineage.

Among fruits, the genomes of grapevine (*Vitis vinifera* cv. PN40024), papaya (*Carica papaya* cv. SunUP), apple (*Malus domesticus* cv. Golden delicious), strawberry (*Fragaria vesca* cv. Hawaii 4), banana (*Musa acuminate* cv. DH-Pahang) and sweet orange (*Citrus sinensis* cv. Valencia) have been sequenced using WGS–FGS, WGS–NGS and WGS–FGS–NGS approaches[32–37]. The significant outcomes, including genome size sequenced and relative density of protein-coding genes, transcription factors, disease resistance-related genes, transposable elements and molecular markers inferred from these decoded fruit genome sequences are briefly summarized in the Table S2 (see Supplementary Material online) and Figure 1. The structural components from these fruit genomes sequenced so far are unravelled through a comprehensive study on vital genomic constitution among these sequenced fruit genomes. The draft genome sequences of fruit crops reveal that within the Rosaceae family, strawberry genome contains the highest number of protein coding genes (34,809) with a very high gene density (165.75/Mb) and relatively less abundance of transposable elements (22% of genome) (Figure 1 *a* and Table S2 (see Supplementary Material online)). Class-I transposons are over-retained in grapevine genome compared to that of class-II transposons. Due to chromosomal duplication, rearrangements and translocation, apple has a larger genome size with massive amplification of transposable elements resulting in lower gene density (95.91/Mb) compared to other members of the same family, whose genomes have been sequenced till date (Figure 1 *a* and Table S2 (see Supplementary Material online)). The papaya and sweet orange genomes experienced no recent WGD leading to relatively smaller genome size of these two fruit crops. Nearly half of the banana genome is rich in transposable elements (45–50% of the genome), which accumulated into its genome during three rounds of WGD that occurred in the *Musa* lineage. LTR retrotransposons represent the largest part of transposable elements, with *Copia* elements (25.7%) being much more abundant than *Gypsy*-type elements (11.6%) in banana genome. A new type of MITEs (Miniature inverted-repeat transposable elements), i.e. MiM (MITE inserted in microsatellite) is identified in the *Citrus* genome.

The genome sequences of fibre crops, flax (*Linum* cv. CDC Bethune) and cotton (*Gossypium raimondii*); vegetable/oilseed crops – chinese cabbage (*Brassica rapa* cv. Chiffu-401-42) and wild mustard (*Brassica oleracea*); millet crop – foxtail millet (*Setaria italica* cv. Zhang gu and *S. italica* cv. Yugu1), and oilseed crop – sesame (*Sesamum indicum* cv. Zhongzhi No. 13) have been sequenced using WGS–FGS, WGS–NGS, WGS–FGS–NGS and CBC–WGS–FGS–NGS approaches[38–43]. A brief overview on the significant outcomes, including genome size sequenced and relative density of protein-coding genes, transcription factors, disease resistance-related genes, transposable elements and molecular markers inferred from these sequenced genomes is provided in Table S3 (see Supplementary Material online) and Figure 1. A comprehensive study on genomic constitution among these sequenced genomes is undertaken to infer structural perspectives of these genomes. The draft genome sequences of oilseed, fibre and bioenergy crops reveal that Chinese cabbage has comparatively smaller genome size (283.8 Mb) with a very high gene density (145.08/Mb), reflecting the role of whole genome triplication occurring in its genome around 13–17 Myr ago. A WGD event in flax genome occurred around 5–9 Myr ago, resulting in a large number of protein coding genes in its genome with a relatively high gene density (143.98/Mb) (Figure 1 *a* and Table S3 (see Supplementary Material online)). WGD is thought to have occurred in the cotton genome leading to massive amplification of transposable elements in the genome resulting in a comparatively lower gene density (52.85/Mb) (Figure 1 *a* and Table S3 (see Supplementary Material online)). Among the LTRs, *Copia*-like elements are the most abundant compared to *Gypsy*-like elements in flax and sesame genomes. In contrast, *Gypsy*-like elements are the most abundant compared to *Copia*-like elements in foxtail millet and cotton genomes.

Structural prospects of decoded crop plant genomes

A structural comparison of genome sequences of different species and subspecies of crop plants reveals that no interrelationship persists between the density of protein-coding genes and non-coding transposable elements underlying overall size differentiation of their genomes sequenced. A significant direct correlation of transposable elements-density with contraction/expansion of genome size, while inverse correlation between density of protein-coding genes and genome size variation is apparent, as evident from the decoded sequences of *O. brachyantha*, maize, sorghum and soybean genomes. In *T. aestivum*, the expansion of both transposable elements and protein-coding genes, whereas in *O. sativa* (*indica*), the expansion of genes contribute more towards their large genome size. On the contrary, in a large genome of *T. urartu*, both the transposable elements (66.9%) and protein-coding genes (7.484/Mb) are less abundant. This collectively infers intricacies in establishing transposable element- and gene-density estimation in the decoded crop plants with their varying genome size. A comprehensive understanding of genomic constitution and complete decoding of coding as well as non-coding sequence components especially of the draft plant genomes is essential to derive their possible impact on size variation of large and small genome species. The available complete and draft reference genome sequences of plant species sequenced so far have expedited the genome resequencing and global transcriptome sequencing of diverse accessions by utilizing multiple NGS approaches. These genomic and genic sequence resources often have the potential to develop numerous SSR, SNP and insertion/deletion (InDel) markers at a genome-wide scale, which are structurally and functionally annotated in different coding and non-coding sequence components of genes/genomes (chromosomes) of crop plants. For instance, ~20 million genome-wide SNPs have been discovered by an international initiative on 'The 3000 rice genome sequence project' through genome resequencing of 3000 rice accessions[44]. In chickpea, the whole genome resequencing of 90 cultivated and wild *Cicer* accessions discovered 4.4 million sequence variants (SNPs and InDels) at a genome-wide scale[24]. These informative genome- and gene-derived markers have been genotyped in phenotypically well-characterized natural germplasm lines and mapping populations using various high-throughput marker genotyping assays for their effective deployment in genomics-assisted crop improvement.

Functional perspectives of decoded crop plant genomes

The *O. brachyantha* wild rice genome contains a large number of disease resistance-related genes (>1000 resistance gene analogues (RGAs)) reflecting their role in adaptation to various adverse environments (Figure 2 and Table 1). The low abundance of protein-coding genes in *O. brachyantha* wild genome compared to *O. sativa* cultivated genome suggests their massive amplification in presently domesticated rice genome. This amplification is possibly caused due to tandem gene duplication and gene transposition. Genes encoding protein kinase and disease resistance-related protein are overrepresented in the Kasalath genome. Notably, a functionally characterized phosphorus uptake 1 (*Pup1*) gene known to be involved in phosphorus-deficiency tolerance is absent in *japonica* Nipponbare genome, but present on chromosome 11 of *aus*-type Kasalath genome, reflecting the adaptive evolution of this gene during rice domestication. A high predominance of RGAs in wheat *Ae. tauschii* genome is observed compared to the other two species of wheat and barley genomes sequenced so far. The expansion of energy-harvesting genes in *T. aestivum* genome leads to more nutrient content in its grain. Moreover, several gene families underlying components of photosystem II, storage proteins, NB-ARC domain-containing protein, and growth and metabolism-related protein have expanded in bread wheat genome, reflecting their role in the accumulation of more nutrient content in its grain. A large number of genes encoding cytrochrome P450 family (*CytP450* genes; 485), cold-responsive genes (216) and myeloblastoma (*MYB*)-like transcription factors (103) are found to be present in *Ae. tauschii* genome, reflecting their role in abiotic stress responses, especially in biosynthetic and detoxification pathways and cold acclimatization. Several grain quality-related genes, including high-molecular weight glutenin subunits (*HMW-GS*), low-molecular weight glutenin subunits (*LMW-GS*), gibberellin-regulated GASA/GAST/Snakin protein family (*GASR7*), *Puroindolines a* (*PINa*), *Puroindolines b* (*PINb*), grain texture proteins (*GSP*) and storage protein activator (*Spa*) are present in *Ae. tauschii* ancestral genome making this a vital source for many grain-quality genes in presently cultivated hexaploid wheat. A large number of abiotic stress-tolerant genes encoding for *CytP450* (261 genes) are identified in maize genome, resulting in its greater adaptation towards abiotic stress. Due to the redirection of C3 progenitor genes as well as recruitment and functional divergence of both ancient and recent gene duplicates, C4 photosynthetic pathway is evolved in sorghum lineages. The sole sorghum C4 pyruvate orthophosphate dikinase (*ppdk*) and phosphoenol pyruvate carboxylase kinase (*ppck*) genes and their two isoforms have only single orthologs in rice. Recent gene and micro-RNA duplication contribute majorly towards drought tolerance in sorghum. The number of genes encoding expansion enzymes is abundant in sorghum (82) compared to rice (58), *Arabidopsis* (40) and poplar (40), which could be linked to the durability of sorghum. Certain gene families, including genes encoding (1,3)-*β*-glucan synthase, protease inhibitors, sugar binding proteins and sugar transporters are expanded in the barley genome.

The *Medicago* genome is rich in disease resistance- (764) and nodulin (593)-related genes (Figure 2 and Table 2). The expansion of these genes in the *Medicago*

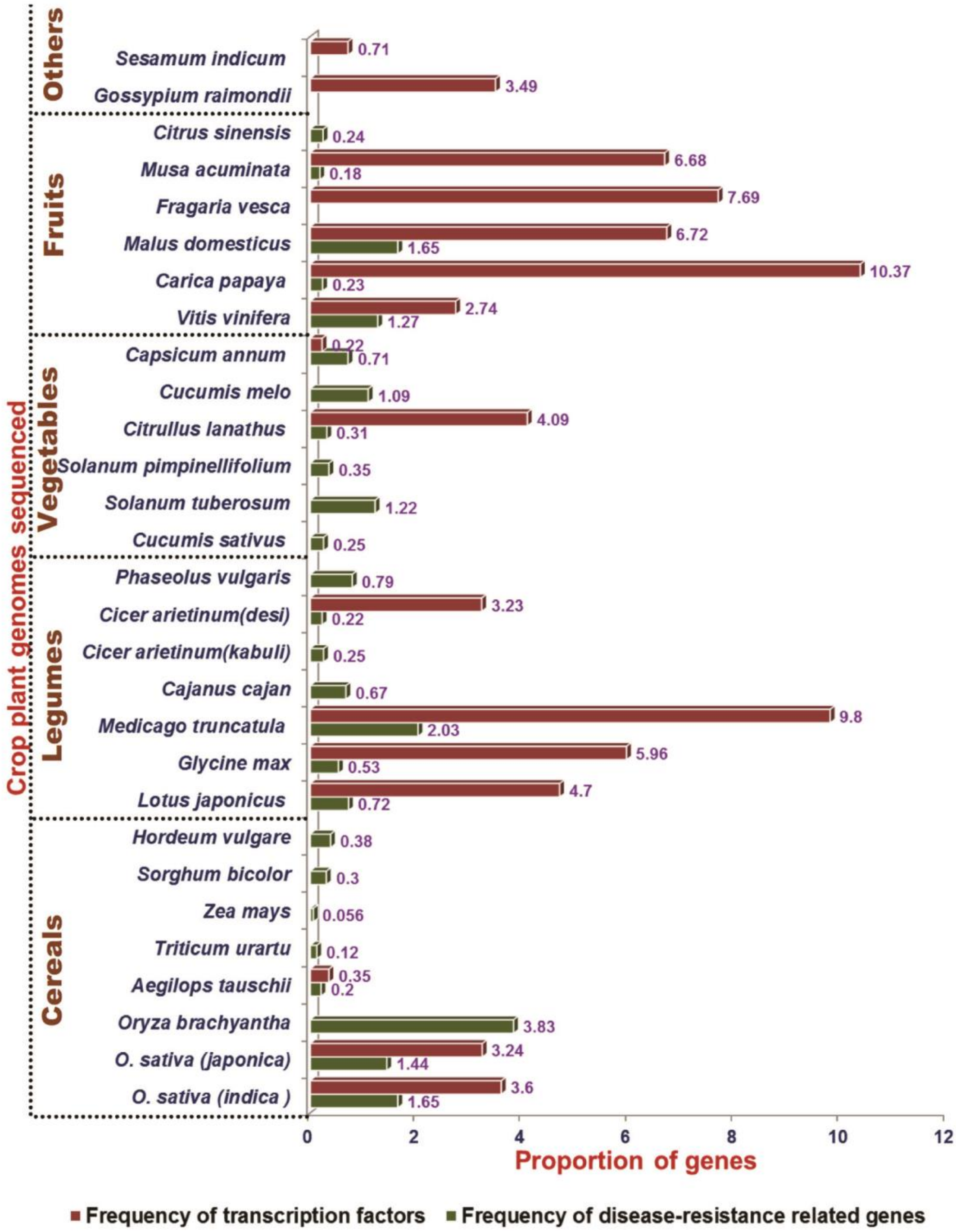

Figure 2. A brief overview of functional outcomes of sequenced crop plant genomes. The proportionate genomic distribution of transcription factors and disease resistance-related genes measured in the sequenced plant genomes is also depicted.

genome occurred during divergence of *Medicago* from the papilionoid family around 58 Myr ago after experiencing a WGD. The functional domains and gene families for several transcription factors (1481), transporters (1310) and receptor protein kinases (1267) (including calmodulin binding transcription activator (*CAMTA*), squamosa promoter binding protein (*SBP*) and lysosome membrane protein 2 (*LIMP2*)-encoding genes) have expanded in the *Lotus* genome, reflecting its role in symbiotic nitrogen fixation. The genome of *Lotus* contains a large number (1312) of functional micro-RNAs. Due to segmental and whole genome-wide duplication, a large number of drought tolerance-related genes (111 in pigeon pea and 109 in soybean) are evolved in pigeon pea and soybean, implying an insight regarding genetic architecture of the pigeon pea and soybean genomes for drought tolerance. A large number of nodulin-related genes like *GRAS* (gibberellic acid insensitive (*GAI*), repressor of

GAI (*RGA*), and scarecrow (SCR)), ethylene responsive transcription factor (*ERF*) and basic leucine zipper (*bZIP*) are identified in *desi* genome, reflecting their role in nodulation and symbiotic nitrogen fixation. Candidate genes for domestication involved in nitrogen metabolism, flowering and plant size regulation have been identified in common bean genome. These genes include orthologs of coat protein1 (*COP1*), cullin4 (*CUL4*) and *AGAMOUS* like 42 (*AGL42*).

Due to two consecutive triplication events experienced by the genome of *Solanum* lineage, several new genes are evolved in the lineage leading to fleshy fruit evolution in tomato and tuberization in potato. More than 100 xyloglucanendotransglucosylase/hydrolases (XTHs) genes, many transcription factors and enzymes necessary for ethylene biosynthesis (Ripening Inhibitor (*RIN*), Colourless non ripening (*CNR*) and Acyl CoA synthetase (*ACS*)), red light photoreceptors-Phytochrome B1/Phytochrome B2 (*PHYB1/PHYB2*) as well asethylene and light regulated genes mediating lycopene biosynthesis-Phytoene synthase 1/Phytoene synthase 2 (*PSY1/PSY2*) are evolved in tomato, influencing its fruit quality. A set of 28 Kunitz protease inhibitors and a large number of RGAs are evolved in potato genome, implying its role in resistance to biotic stress. Several gene families underlying *Lox* (lipoxygenase) pathway, cell-wall biogenesis and citrulline biosynthesis pathway are evolved in watermelon genome. Like watermelon, genes underlying *Lox* pathway, cadmium sensitivity and tendril formation (*EXLA*) have expanded in cucumber genome, whereas genes for phytochelatin synthase and mandelonitrilelyase are also expanded in melon genome. Cucumber has less number of disease resistance-related genes (61), indicating that the *Lox* pathway may provide a complementary pathway to cope up with biotic stress. Capsaicin synthase[45] appears in hot pepper genome due to unequal tandem gene duplication that occurred after speciation, thus resulting in pungency of hot pepper[45]. Several genes, including capsanthin-capsorubin synthase (CCS)[46], *GDP-*L*-galactose* phosphorylase (*GGP1*) and dehydroascorbate reductase (*DHAR*) are found to be highly expressed during fruit ripening. Gene families involved in disease resistance and cellular function, such as *cytP450* and heat shock protein 70 are found to be significantly expanded in the hot pepper genome.

The recent genome-wide duplication in apple genome has resulted in the expansion of several gene families underlying disease resistance, lignin biosynthesis, terpene synthase, sorbitol synthesis and fruit development, explaining the formation of pome, a pyreae-specific false fruit type. Several gene families underlying tannin and terpene biosynthesis pathways have also expanded in the grapevine genome, reflecting their role in aroma production in this particular fruit crop. Unlike other angiosperms sequenced so far, papaya genome lacks recent WGD. However, expansion of several gene families, including genes for expansion B, starch synthase and ethylene responsive binding factor has occurred in the papaya genome providing several new properties such as tree-like habit, deposition and remobilization of starch reserves and attraction of seed-dispersal agents in papaya plant. Genes underlying vitamin C biosynthesis pathway, galactouranic acid reductase and oxidoreductase have expanded in sweet orange genome. Genes overrepresented in fruits of strawberry are enriched for several categories of biological processes and molecular functions associated with fruit development and carbohydrate metabolic activity. Transcription factors like *R2R3MYB* and transparent testa 2 (*TT2*)-like *MYB* have expanded in the strawberry genome, suggesting a key function in strawberry pro-anthocyanin synthesis. Transcription factor families are strikingly expanded in banana compared with other plant genomes, implying their most effective contribution towards specific aspects of banana development, including cell-wall modification and ripening process. The banana genome is rich in deeply conserved non-coding sequences (116) within commelinid monocotyledons, and between monocotyledons and eudicotyledons[47], representing their role in detecting novel motifs with a gene regulation function.

Foxtail millet evolved from its common ancestor sorghum and maize around 27 Myr ago through chromosomal rearrangement and local gene duplication, which resulted in the evolution of foxtail millet-specific gene families like 586 stress response genes, reflecting its role in adaptation of the crop to semi-arid environments. It is hypothesized that the evolution of C4 photosynthesis occurred in foxtail millet and sorghum lineage, which suggests that C4 isoform of malic enzyme is recruited from a different C3 paralog in foxtail millet than maize and sorghum. The flax genome is relatively less abundant in leucine-rich repeat (LRR) domain-containing protein-coding genes. Genes underlying hormone biosynthesis (auxin, cytokinin, brassinosteroid) and *TCP* (teosinte branched1 (*tb1*), cycloidea (*CYC*) and proliferating cell factor (*PCF*)) transcription factors (39) are overrepresented in the Chinese cabbage genome. Genes responsive to important environmental factors, including cold, salt and osmotic stress, and plant hormone biosynthetic genes are overrepresented in the Chinese cabbage genome. Except *Cocoa*, cotton is the only plant species that contains an authentic *CDN1* gene family for gossypol biosynthesis. Cotton genome contains a large number of micro-RNA (348). Whole genome triplication event has led to the expansion of genes involved in auxin functioning (such as auxinindole-acetic acid (*AUX-IAA*), Gretchen Hagen (*GH*$_3$), pin formed 1 (*PIN*), small auxin upregulated RNA (*SAUR*), tryptophan aminotransferase of *Arabidopsis* (*TAA*), transport inhibitor 1 (*TIR*), topless (*TPL*) and YUCCA), morphological specification (*TCP*) and flowering time control (flowering locus C (*FLC*), constans (*CO*), vernalization1 (*VRN1*), leafy (*LFY*), apetala 1

(*AP1*) and gigantia (*GI*)) in *B. oleracea* genome. The gene families encoding lipid transfer protein 1 (*LTP1*)[48], midchain alkane hydroxylase, FAD4-like desaturase and alcohol forming fatty acyl-CoA reductase have been found to be expanded in sesame genome, leading to high oil accumulation by strengthening transport of the fatty acid and other lipid molecules in sesame. In addition, the two cytosolic lipoxygenase and lipid acyl hydrolase-like families related to degradation of lipids are both contracted in sesame. Two genes encoding dirigent protein and sesamin synthase involved in sesamin biosynthesis are found to be present in the sesame genome, implying their role in genetic foundation for sesame-specific product.

Functional prospects of decoded crop plant genomes

The functional annotation (protein sequence homology searches (BLASTX), gene ontology (GO) and genome-scale/interactive pathway analysis) of genes discovered by decoding the plant genomes has led to the identification of multiple species-specific genes contributing towards contrasting characteristics in the crop plants. For instance, a large number of energy-harvesting genes have been identified in a hexaploid genome of wheat, making its grain rich in protein content. Species-specific genes, including nodule-related genes (nodulin) in *Medicago*, capsaicin synthase genes in hot pepper, terpene and tannin biosynthesis genes in grape, vitamin C biosynthesis genes in sweet orange, gossypol biosynthesis genes in cotton, and glucosinolate biosynthesis/catabolism genes in *B. oleracea* providing distinct features to these crop plants have been identified. The assignment of putative function to annotated protein-coding genes will assist in user-specific selection of candidate genes and informative markers from different coding and non-coding sequence components of genes on a priority basis for their further utilization in genomics-assisted breeding applications of crop plants. In addition, this will be useful in rapidly establishing marker-trait linkages and identification/mapping of genes/QTLs governing important agronomic traits in crop plants. The much needed functional viewpoints obtained from the decoded plant genomes will provide the necessary clues to accelerate functional validation and molecular characterization of genes for trait genetic enhancement studies in plant species. Significant progress has been made in this regard to clone and characterize more than 2000 genes using map-based isolation and diverse functional genomics approaches in a completely sequenced rice genome vis-à-vis other draft cereal genomes sequenced hitherto[49]. However, with the complete whole genome sequencing of draft crop plant genomes and subsequent progress in genome resequencing and transcriptome sequencing of numerous crop accessions, more number of genes (those encoding hypothetical and unknown expressed proteins) can be functionally annotated and characterized in the near future in an efficient manner to expedite the functional genomic analysis in crop plants.

Comparative and phylogenetic perspectives of decoded crop plant genomes

All the cereals evolved from a common ancestral genome $n = 5$ near about 90 Myr ago. Rice and wheat evolved from its ancestral genome $n = 12$ after breakage and fusion of chromosomes, chromosomal translocation and segmental duplications. Like maize and sorghum, expansion in rice genome occurred due to two massive recent amplifications of LTR retrotransposon (<0.5 and 1–2 Myr ago). However, specifically in *O. brachyantha*, massive removal of ancient gene families through unequal homologous recombination and illegitimate recombination has led to its smaller genome size. Though wheat and rice share a common ancestor, a significant difference in ploidy level ($n = 12$ for rice and $n = 7$ for wheat) is clearly observed in their genomes. During evolution, a large number of transposable elements accumulated in the wheat genome resulting in its lower gene density compared to rice. The chromosomal fusion in wheat genome is likely to have occurred in *Triticeae* tribe (wheat and barley), leading to their similar ploidy level and relatively similar gene density. As a result of polyploidization, a large number of gene families have been lost from the bread wheat genome. However, the scale of gene loss in hexaploid wheat compared with maize and Chinese cabbage is significantly smaller. This is possibly because of relatively recent origin and absence of intergenome recombination in wheat[50]. Maize and sorghum evolved independently from a common ancestral genome ($n = 10$). Maize genome underwent a WGD following several chromosomal fusions, leading to the accumulation of transposable elements in its genome resulting a lower gene density. Barley being a member of the Triticeae tribe shared the same ploidy level with wheat ($n = 7$). It is hypothesized that the five chromosomal fusions in wheat that occurred during evolution are also likely to have occurred in the barley genome giving rise to same ploidy level and comparable genome size[51]. Also, 8443 gene families are shared by 5 grass genomes (rice, *Ae. tauschii*, *Brachypodium*, barley and sorghum). On the basis of orthology analysis, 24,339 gene families are identified in 5 grass species (rice, *T. urartu*, maize, *Brachypodium* and sorghum), of which 9836 gene families are common to all 5 grasses. Among 11,892 maize gene families, 11,088 are shared by maize and sorghum, which is highest in number compared to that of maize-rice (10,898) and maize–*Arabidopsis* (8715), suggesting that maize is phylogenetically closely related to sorghum.

Around 54 Myr ago, the papilionoideae subfamily diverged into two major subgroups, the millettioid (soybean and pigeon pea) and galetoid (*Medicago*, *Lotus* and chickpea). Within the millettioid clade, pigeon pea diverged from soybean around 10–20 Myr ago. After the divergence, soybean genome underwent recent WGD around 13 Myr ago, resulting in the expansion of its genome size and accumulation of transposable elements in its genome. This recent WGD resulted in local gene duplication and gene rearrangements in soybean genome, resulting in the expansion of oil biosynthesis genes and nodulin-related genes in its genome. The recent WGD which occurred in soybean genome around 13 Myr ago is missing in pigeon pea, resulting in its smaller genome size and lower abundance of transposable elements compared to soybean. Within the galetoids clade, chickpea diverged from *Lotus* and *Medicago* around 20–30 and 10–20 Myr ago respectively. A WGD (58 Myr ago) and local gene rearrangements occurred in *Medicago* and *Lotus* genomes, resulting in sub- or neo-functionalization of signalling components and regulators showing specialized role in nodulation. Synteny analysis reveals that among legumes, soybean has the highest number of syntenic blocks with chickpea, reflecting its recent polyploidy ancestry. The divergence of two wild subpopulations of common bean (Mesoamerican and Andean) is thought to have occurred around 165,000 years ago. The Andean subpopulation underwent an exponential growth phase which began around 90,000 years ago. The predomestication bottleneck is found to be present in the Andean population but absent in the Mesoamerican population. Comparative analyses between millettioid and galegoid subgroups reveal that 16,380 gene families are shared by these two subgroups, whereas a total of 2951 and 5331 gene families are specific to the millettioid and galegoid subgroups respectively. Comparative analysis between chickpea, *Lotus* and *Medicago* shows that of 10,869 gene families are shared by all the three species, whereas *Medicago* shares a large number of gene families (3237) with chickpea compared to that of *Lotus* (1437), suggesting that chickpea is more closely related to *Medicago*. A total of 4311 gene families containing 72,193 genes are shared by *Medicago*, soybean, *Lotus*, grapevine and pigeon pea. About 1024 gene families are identified to be specific to the soybean and pigeon pea genomes, suggesting that soybean is more closely related to pigeon pea.

Solanum lineage is thought to have experienced two consecutive genome triplications (one is ancient shared with rosid and another is recent that occurred around 71 Myr ago). Potato diverged from tomato around 7.3 Myr ago. Both of these two triplication events resulted in neo-functionalization of genes controlling fruit characteristics in tomato and hot pepper. This also led to the expansion of gene families (*SP6A* (Self Pruning 6A), *SP5G* (Self Pruning 6A) and *GGP1*), tissue-specific expression and recruitment of genes towards a new pathway (flowering control pathway) in potato genome causing tuberization to evolve exclusively in *Solanum* section *Petota*. Comparative analysis of resistance genes of hot pepper and other *Solanaceae* plants implies that expansion and diversification of resistance genes have been involved in lineage-specific parallel evolution through unequal gene duplication events, resulting in different gene repertoires even in closely related species. Cucurbitaceae genome speciation is thought to have occurred around 15–23 Myr ago. Cucumber diverged from melon around 10.1 Myr ago. Though melon and cucumber belong to the same genus, there is a significant difference in ploidy level (*C. melo*: $2n = 2x = 24$ and *C. sativus*: $2n = 2x = 14$) and gene density (73.13/Mb and 109.57/Mb) (Figure 1 *a*, Table S1), implying that during divergence a large number of gene rearrangements, duplication and deletion occurred in the cucumber genome. No recent WGD event is found to prevail in melon and cucumber genomes. However, segmental duplication occurred in melon genome, resulting in expansion of several genes, including defence response and apoptosis functional process in its genome. A total of 18,320 distinct orthologous tomato–potato gene-pairs are identified through comparative analysis between tomato and potato genomes. Also 18,809 gene orthologous groups are identified among *S. lycopersicum*, *S. pimpinellifolium* and *S. tuberosum*. Within the Cucurbitaceae family (cucumber, watermelon and melon), a total of 3543 orthologous relationships covering 60% of the watermelon genome are identified. Syntenic analysis reveals that 5473, 6525, 9842, 8439 and 3992 cucumber genes are collinear to *Arabidopsis*, papaya, poplar, grapevine and rice respectively. Synteny is also observed between cucumber chromosome 6 and melon chromosome 3, indicating that interchromosome rearrangement occurred in one of the two genomes after speciation. Cucumber chromosome 4 is syntenous with melon chromosomes 7 and 8, indicating that the rearrangements most likely occurred before the divergence of cucumber and melon. It is observed that cucumber chromosomes 1, 2, 3, 5 and 6 are collinear to 10 melon chromosomes (1 = 2 and 12), (2 = 3 and 5), (3 = 4 and 6), (5 = 9 and 10) and (6 = 8 and 11), indicating that chromosomal fusion events occurred in cucumber genome during speciation.

All members of the Rosaceae family whose genomes have been sequenced so far are evolved from a common ancestor $n = 9$. A relatively recent (>50 Myr ago) WGD has resulted in transition from 9 ancestral chromosomes to 17 in apple leading to massive amplification of transposable elements causing its larger genome size compared to other crops of the Rosaceae family whose genomes have been sequenced so far. The genome of strawberry is the only one sequenced to date with no evidence of large-scale within genome duplication resulting in a small genome size with a very high gene density and less accumulation of transposable elements in its genome. Banana genome has experienced three genome-wide duplications (two consecutive recent α and β, and one ancient γ). No recent WGD is reported to occur in sweet orange, grapevine and strawberry genomes. Grapevine genome is thought to have originated from the contribution of three ancestral genomes. The palaeo-hexaploidy observed in grapevine genome is also present in its common ancestral genomes, poplar and *Arabidopsis*[52]. The ancient duplication events in grapevine genome resulted in the accumulation of transposable elements in it. Unlike other angiosperm genomes sequenced so far, the papaya genome lacks the recent WGD[53], resulting in a smaller genome with relatively less gene numbers compared to that of other angiosperm. However, striking amplification of genes occurred within a particular functional group of papaya genome, resulting in the evolution of tree-like habit, deposition and remobilization of starch reserves, attraction of seed-dispersal agents and its adaptation to tropical environment. Comparative analysis among banana, rice, sorghum, *Brachypodium*, date palm and *Arabidopsis* reveals that 7674 gene families are shared by these plant species, of which 759 are found to be specific to the banana genome. Comparative genome analysis between papaya and *Arabidopsis* reveals that papaya segments show collinearity with 2–4 *Arabidopsis* segments, indicating that either one or two genome duplications have affected the *Arabidopsis* lineage, since its divergence from papaya. In contrast, individual *Arabidopsis* segments are collinear to only one papaya segment, suggesting that no genome duplication has occurred in the papaya genome since its divergence from *Arabidopsis* around 72 Myr ago. Orthologous regions between *Arabidopsis* and grapevine are found to be different for each of the three grapevine chromosomes, indicating that *Arabidopsis*/grapevine ancestor had the same palaeo-hexaploid content. Conversely, orthologous regions in rice are the same for the three paralogous chromosomes of grapevine, suggesting that the triplication event is absent in common ancestor of monocotyledons and dicotyledons. Syntenic analysis among orange, cacao, *Arabidopsis*, apple, strawberry and grape reveals that no recent WGD occurred in orange genome, except the shared ancient triplication. Comparative analysis among rice, *Arabidopsis*, grape and strawberry reveals that 663 gene families are unique to *Arabidopsis* and strawberry, whereas 262 and 6233 gene families are shared between strawberry and rice as well as among all these four plant species respectively.

A whole genome triplication event is thought to have occurred in Chinese cabbage genome leading to its mesohexaploidization, resulting in gene loss along with substantial gene conversion. The whole genome triplication event in *B. oleracea* genome has led to gene loss, sub- or neo-functionalization of duplicated genes and over-retention of some genes underlying metabolic pathways, including oxidative phosphorylation, carbon fixation, photosynthesis and circadian rhythm. Cotton diverged from its common ancestor around 33.7 Myr ago and later experienced a WGD, reflecting its role in larger genome size of this crop. The evolution of flax genome involved chromosome doubling followed by loss of one or more chromosome. A recent WGD event (5–9 Myr ago) in the flax genome led to the evolution of some *Linum* lineage-specific genes (*Agglutinin* gene). The divergence of foxtail millet from its common ancestor (sorghum and maize) is thought to have occurred around 27 Myr ago. During the divergence, several chromosomal fusion and gene rearrangements occurred in its genome leading to evolution of C4 photosynthetic pathway independently in the *Setaria* lineage. Comparative genome analysis among foxtail millet, *Brachypodium*, rice, sorghum and maize reveals that about 71.8, 72.1, 61.5 and 86.7 proportion of foxtail millet genome is collinear with rice, sorghum, *Brachypodium* and maize respectively. Foxtail millet chromosomes 2, 3 and 9 are found to be collinear with 6 rice chromosomes (2 = 7 and 9), (3 = 5 and 12) and (9 = 3 and 10) respectively, suggesting that three pairs of these chromosomes separately fused to form three chromosomes in foxtail millet after divergence from common ancestor. Comparative analysis among cotton, cacao, *Arabidopsis* and maize reveals that 9525 gene families are shared by all these four plant species. Likewise, 9909 gene families are shared by all four plant species – papaya, *Arabidopsis*, Chinese cabbage and grapevine. Notably, 108.6 Mb genomic region covering 90.1% of *Arabidopsis* genome is found to be collinear with 259.6 Mb (covering 91.1%) of Chinese cabbage genome. Sesame genome experienced a WGD event resulting in the expansion of genes encoding lipid transfer protein1, mid-chain alkane hydroxylase, FAD4-like desaturase and alcohol forming fatty acyl-CoA reductase, implying the complex genetic architecture of high oil content in sesame.

Comparative and phylogenetics prospects of decoded crop plant genomes

The effective deployment of cereal crop genome sequences in comparative mapping at macro (chromosome)- and micro (gene)-syntenic level reveals evolutionary and domestication patterns as well as changing ploidy level of these crop species. However, the origin and domestication progression of crop plants is controversial and has long been debated in many studies. However, genome resequencing of numerous crop accessions representing diverse cultivated (landraces, breeding lines and varieties) and wild genetic backgrounds is underway for a more comprehensive understanding of evolution, phylogenetics and genetic structure of their genomes. These efforts led to the identification of several potential genomic (selective sweep) regions that have undergone selection during domestication of a particular agronomic

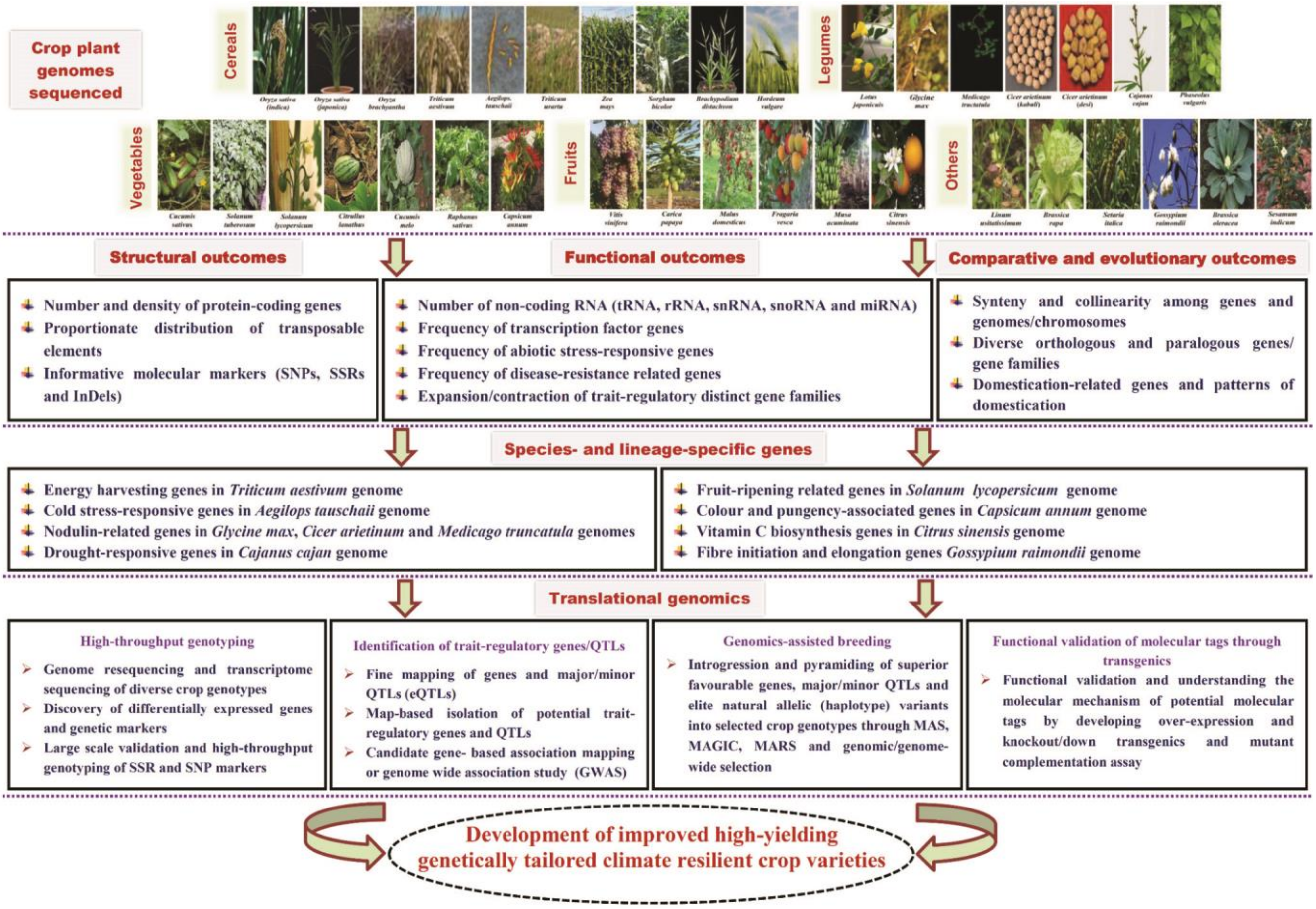
Crop plant genomes sequenced
Cereals
Oryza sativa (indica)
Oryza sativa (japonica)
Oryza brachyantha
Triticum aestivum
Aegilops. tauschaii
Triticum urartu
Zea mays
Sorghum bicolor
Brachypodium distachyon
Hordeum vulgare
Legumes
Lotus japonicuis
Glycine max
Medicago tructatula
Cicer arietinum (kabuli)
Cicer arietinum (desi)
Cajanus cajan
Phaseolus vulgaris
Vegetables
Cucumis sativus
Solanum tuberosum
Solanum lycopersicum
Citrullus lanathus
Cucumis melo
Raphanus sativus
Capsicum annum
Fruits
Vitis vinifera
Carica papaya
Malus domesticus
Fragaria vesca
Musa acuminata
Citrus sinensis
Others
Linum usitatissimum
Brassica rapa
Setaria italica
Gossypium raimondii
Brassica oleracea
Sesamum indicum
Structural outcomes
Number and density of protein-coding genes
Proportionate distribution of transposable elements
Informative molecular markers (SNPs, SSRs and InDels)
Functional outcomes
Number of non-coding RNA (tRNA, rRNA, snRNA, snoRNA and miRNA)
Frequency of transcription factor genes
Frequency of abiotic stress-responsive genes
Frequency of disease-resistance related genes
Expansion/contraction of trait-regulatory distinct gene families
Comparative and evolutionary outcomes
Synteny and collinearity among genes and genomes/chromosomes
Diverse orthologous and paralogous genes/ gene families
Domestication-related genes and patterns of domestication
Species- and lineage-specific genes
Energy harvesting genes in Triticum aestivum genome
Cold stress-responsive genes in Aegilops tauschaii genome
Nodulin-related genes in Glycine max, Cicer arietinum and Medicago truncatula genomes
Drought-responsive genes in Cajanus cajan genome
Fruit-ripening related genes in Solanum lycopersicum genome
Colour and pungency-associated genes in Capsicum annum genome
Vitamin C biosynthesis genes in Citrus sinensis genome
Fibre initiation and elongation genes Gossypium raimondii genome
Translational genomics
High-throughput genotyping
Genome resequencing and transcriptome sequencing of diverse crop genotypes
Discovery of differentially expressed genes and genetic markers
Large scale validation and high-throughput genotyping of SSR and SNP markers
Identification of trait-regulatory genes/QTLs
Fine mapping of genes and major/minor QTLs (eQTLs)
Map-based isolation of potential trait-regulatory genes and QTLs
Candidate gene- based association mapping or genome wide association study (GWAS)
Genomics-assisted breeding
Introgression and pyramiding of superior favourable genes, major/minor QTLs and elite natural allelic (haplotype) variants into selected crop genotypes through MAS, MAGIC, MARS and genomic/genome-wide selection
Functional validation of molecular tags through transgenics
Functional validation and understanding the molecular mechanism of potential molecular tags by developing over-expression and knockout/down transgenics and mutant complementation assay
Development of improved high-yielding genetically tailored climate resilient crop varieties

Figure 3. A brief overview demonstrating the broad practical applicability of structural, functional, comparative and evolutionary outcomes of sequenced crop plant genomes in translational genomics for crop genetic enhancement. Snapshots of crop plants illustrated have been adapted from http://en.wikipedia.org.

trait and uncover the trait domestication pattern within and/or among plant species. For instance, a comprehensive map of rice genome variation detected 55 selective sweeps-targeted genomic regions that have occurred during domestication by genome resequencing of diverse 1083 cultivated *indica* and *japonica* varieties and 446 accessions of wild rice species (*O. rufipogon*)[54]. In soybean, 302 wild and cultivated accessions have been resequenced leading to the identification of genes underlying domestication and improvement traits (including oil content, plant height, seed-coat colour and pubescence form) at a genome-wide scale[55]. In cucumber, the resequencing of 115 accessions has uncovered 112 domestication sweep regions containing a gene responsible for loss of bitterness in this vegetable crop[56]. Several accessions of sweet orange have been resequenced at whole genome level, thereby documenting their ancestral origin from pummelo and mandarin, and deciphering the domestication pattern of diverse traits undergoing selection in this fruit crop[57]. The whole genome resquencing of 48 flax accessions provides insight into the impact and processes of flax domestication, especially for the trait of winter hardiness[58]. The signatures of selection, in-depth analyses of the domestication sweeps and identification of natural allelic variants associated with domestication traits (like pericarp colour, seed size/weight and seed shattering) provide a vital resource/genomics strategy for research on selective breeding and domestication leading genetic enhancement of crop plants.

Conclusion and future prospects

This review presents a comprehensive picture regarding the structural, functional, comparative and evolutionary studies that have been undertaken so far by revisiting and precisely comparing all the 28 genome sequence landmark reports on cultivated food crops encompassing five major groups. However, the dependence of these comparative outcomes upon diverse methods/experimental strategies, computational genomics, bioinformatics software/pipelines and algorithms adopted in the sequencing and assembling of different complete and draft crop plant genomes cannot be overruled. The structural outcomes equating all these sequenced crop genomes overall indicate their genomic constitution, particularly the way protein-coding genes, transposable elements and molecular markers (SSRs and SNPs) are organized within the genomes. The functional outcomes provide a brief account of specific phenotypic trait characteristics of crop genomes by understanding the genetic make-up of transcription factor genes, abiotic and biotic stress-responsive genes and expansion/contraction of distinct gene families governing various agronomic traits in these plant species. The comparative and evolutionary outcomes deduce the genetic basis of biological diversity and basic process of genome evolution by analysing the syntenic relationships and collinearity among genes and genomes/chromosomes of the sequenced crop plants. Meanwhile, diverse orthologous and paralogous gene families/genes, domestication-related genes, and species- and lineage-specific genes controlling traits of agricultural importance have been identified from these crop genomes. The cues, including trait-regulating genes, alleles and markers obtained by correlating and integrating structural, functional, comparative and evolutionary outcomes from the above sequenced 30 major food crop genomes can be essentially utilized for genetic improvement of crop plants via translational genomics (Figure 3).

Consequently, the available gold standard reference whole/draft genome sequences have propelled the whole genome resequencing and transcriptome sequencing of diverse crop genotypes in recent years by use of NGS. This in turn led to the development of enormous resources in the form of differentially expressed known/candidate genes, regulatory elements and genomic (genic) SSR and SNP markers at a genome-wide scale in crop plants. The large-scale validation and high-throughput genotyping of these functionally relevant molecular tags employing various high-throughput array-based NGS and marker genotyping technologies in diverse natural germplasm (core and mini-core) collections, advanced generation mapping and mutant populations of crop plants are underway. These efforts ensued many promising outcomes, including scanning of novel functional allelic variants from diverse crop genetic resources and understanding the molecular diversity, population genetic structure and domestication patterns, particularly among natural germplasm lines. Consequently, these exertions assisted in the construction of ultra-high density genetic linkage maps, QTL/eQTL (expression QTL) mapping, fine mapping, candidate gene-based/genome-wide association study (GWAS) and map-based cloning for identifying potential genomic loci (major/minor QTLs, genes and alleles) associated with many qualitative and complex quantitative traits of agronomic importance in crop plants. The introgression of these validated superior genes, QTLs and natural elite allelic variants into diverse crop genotypes is now possible with the use of marker-assisted selection, multi-parent advanced generation inter-cross (MAGIC) and genomic/haplotype selection and developing over-expression and knockdown/out transgenics (Figure 3). This will eventually help us develop genetically tailored, high-yielding, stress-tolerant crop varieties for sustaining global food security amidst acute climate change scenario in the near future.

1. Arabidopsis genome initiative, Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, 2000, **6814**, 796–815.
2. International Rice Genome Sequencing Project, The map-based sequence of the rice genome. *Nature*, 2005, **436**, 793–800.
3. Michael, T. P. and Jackson, S., The first 50 plant genomes. *Plant Genome*, 2013, **6**, 1–7.
4. Bolger, M. E. *et al.*, Plant genome sequencing – applications for crop improvement. *Curr. Opin. Biotechnol.*, 2014, **26**, 31–37.
5. www.faostat.fao.org
6. Brenchley, R. *et al.*, Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature*, 2012, **491**, 705–710.
7. Paterson, A. H. *et al.*, The *Sorghum bicolor* genome and the diversification of grasses. *Nature*, 2009, **457**, 551–556.
8. Schanable, P. S., *et al.*, The B73 maize genome; complexity, diversity, and dynamics. *Science*, 2009, **326**, 1112–1117.
9. Young, N. D., *et al.*, The *Medicago* genome provides insight into the evolution of rhizobial symbioses. *Nature*, 2011, **480**, 520–524.
10. The tomato genome consortium, The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, 2012, **485**, 635–641.
11. Xu, X. *et al.*, Genome sequence and analysis of the tuber crop potato. *Nature*, 2011, **475**, 189–195.
12. Chen, J. *et al.*, Whole-genome sequencing of *Oryza brachyantha* reveals mechanisms underlying *Oryza* genome evolution. *Nature Commun.*, 2012, **4**, 1595–1601.
13. Jun, Y. *et al.*, A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science*, 2002, **296**, 79–87.
14. Sakai, H. *et al.*, Construction of pseudomolecule sequences of the *aus* rice cultivar Kasalath for comparative genomics of Asian cultivated rice. *DNA Res.*, 2014, **21**, 397–405.
15. Jia, J. *et al.*, *Aegilops tauschii* draft genome sequence reveals a gene repertoire for wheat adaptation. *Nature*, 2013, **496**, 91–95.
16. Ling, H. Q. *et al.*, Draft genome of the wheat A – genome progenitor *Triticum urartu*. *Nature*, 2013, **496**, 87–90.
17. International Wheat Genome Sequencing Consortium, A chromosome-based draft sequence of the hexaploid bread wheat (*Triticumaestivum*) genome. *Science*, 2014, **345**, 1251788.
18. Mayer, K. F. *et al.*, A physical, genetic and functional sequence assembly of the barley genome. *Nature*, 2012, **491**, 711–716.
19. Sato, S. *et al.*, Genome structure of the legume, *Lotus japonicus*. *DNA Res.*, 2008, **15**, 227–239.
20. Schmutz, J. *et al.*, Genome sequence of the palaeopolyploid soybean. *Nature*, 2010, **463**, 178–183.
21. Varshney, R. K. *et al.*, Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. *Nature Biotechnol.*, 2011, **30**, 83–89.
22. Singh, N. K. *et al.*, The first draft of the pigeon pea genome sequence. *J. Plant Biochem. Biotechnol.*, 2012, **21**, 98–112.
23. Varshney, R. K. *et al.*, Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement. *Nature Biotechnol.*, 2013, **31**, 240–246.
24. Jain, M. *et al.*, A draft genome sequence of the pulse crop chickpea (*Cicer arietinum* L.). *Plant J.*, 2013, **74**, 715–729.
25. Schmutz, J. *et al.*, A reference genome for common bean and genome-wide analysis of dual domestications. *Nature Genet.*, 2014, **46**, 707–713.
26. Huang, S. *et al.*, The genome of the cucumber, *Cucumis sativus* L. *Nature Genet.*, 2009, **41**, 1275–1281.

27. Xu, X. *et al.*, Genome sequence and analysis of the tuber crop potato. *Nature*, 2011, **475**, 189–195.
28. Guo, S. *et al.*, The draft genome of watermelon (*Citrullus lanatus*) and resequencing of 20 diverse accessions. *Nature Genet*., 2013, **45**, 51–58.
29. Mas-Garcia, J. *et al.*, The genome of melon (*Cucumis melo* L.). *Proc. Natl. Acad. Sci. USA*, 2012, **109**, 11872–11877.
30. Kitashiba, H. *et al.*, Draft sequences of the radish (*Raphanus sativus* L.) genome. *DNA Res*., 2014, **21**, 481–490.
31. Kim, S., *et al.*, Genome sequence of the hot pepper provides insights into the evolution of pungency in *Capsicum* species. *Nature Genet*., 2014, **46**, 270–278.
32. Jaillon, O. *et al.*, French–Italian Public Consortium for Grapevine Genome Characterization. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, 2007, **449**, 463–467.
33. Ming, R. *et al.*, The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature*, 2008, **452**, 991–996.
34. Velasco, R. *et al.*, The genome of the domesticated apple (*Malus* × *domestica* Borkh.). *Nature Genet*., 2010, **42**, 833–839.
35. Shulaev, V. *et al.*, The genome of woodland strawberry (*Fragaria vesca*). *Nature Genet*., 2011, **43**, 109–116.
36. D'Hont, A. *et al.*, The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. *Nature*, 2012, **488**, 213–217.
37. Xu, Q. *et al.*, The draft genome of sweet orange (*Citrus sinensis*). *Nature Genet*., 2012, **45**, 59–66.
38. Wang, Z. *et al.*, The genome of flax (*Linum usitatissimum*) assembled *de novo* from short shotgun sequence reads. *Plant J.*, 2012, **72**, 461–473.
39. Wang, K. *et al.*, The draft genome of a diploid cotton *Gossypium raimondii*. *Nature Genet*., 2012, **44**, 1098–1103.
40. Wang, X. *et al.*, The genome of the mesopolyploid crop species *Brassica rapa*. *Nature Genet*., 2011, **43**, 1035–1039.
41. Liu, S. *et al.*, The *Brassica oleracea* genome reveals the asymmetrical evolution of polyploid genomes. *Nature Commun*., 2014, **5**, 3930–3935.
42. Zhang, G. *et al.*, Genome sequence of foxtail millet (*Setaria italica*) provides insights into grass evolution and biofuel potential. *Nature Biotechnol*., 2012, **30**, 549–554.
43. Bennetzen, J. L. *et al.*, Reference genome sequence of the model plant *Setaria*. *Nature Biotechnol*., 2012, **30**, 555–561.
44. Alexandrov, N. *et al.*, SNP-seek database of SNPs derived from 3000 rice genomes. *Nucl. Acids Res*., 2015, **43**, D1023–D1027.
45. Bennett, D. J. and Kirby, G. W., Constitution and biosynthesis of capsaicin. *J. Chem. Soc. C*, 1968, 442–446.
46. Hugueney, P. *et al.*, Metabolism of cyclic carotenoids: a model for the alteration of this biosynthetic pathway in *Capsicum annuum* chromoplasts. *Plant J*., 1995, **8**, 417–424.
47. Freeling, M. and Subramaniam, S., Conserved noncoding sequences (CNSs) in higher plants. *Curr. Opin. Plant Biol*., 2009, **12**, 126–132.
48. Kader, J. C., Lipid-transfer proteins in plants. *Annu. Rev. Plant Physiol. Plant Mol. Biol*., 1996, **47**, 627–654.
49. http://venyao.github.io/RICENCODE
50. Akhunov, E. D. *et al.*, Comparative analysis of syntenic genes in grass genomes reveals accelerated rates of gene structure and coding sequence evolution in polyploid wheat. *Plant Physiol*., 2013, **161**, 252–265.
51. Salse, J. *et al.*, Identification and characterization of shared duplications between rice and wheat provide new insight into grass genome evolution. *Plant Cell*, 2008, **20**, 11–24.
52. Bodt, D. S., Maere, S. and Van de Peer, Y., Genome duplication and the origin of angiosperms. *Trends Ecol. Evol*., 2005, **20**, 591–597.
53. Bowers, J. E. *et al.*, Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature*, 2003, **422**, 433–438.
54. Huang, X. *et al.*, A map of rice genome variation reveals the origin of cultivated rice. *Nature*, 2012, **490**, 497–501.
55. Zhou, Z. *et al.*, Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nature Biotechnol*., 2015, **33**, 408–414.
56. Qi, J. *et al.*, A genomic variation map provides insights into the genetic basis of cucumber domestication and diversity. *Nature Genet*., 2013, **45**, 1510–1515.
57. Wu, G. A. *et al.*, Sequencing of diverse mandarin, pummelo and orange genomes reveals complex history of admixture during citrus domestication. *Nature Biotechnol*., 2014, **32**, 656–662.
58. Fu, Y. B., Population-based resequencing revealed an ancestral winter group of cultivated flax: implication for flax domestication processes. *Ecol. Evol*., 2012, **2**, 622–635.

ACKNOWLEDGEMENTS. We thank the Department of Biotechnology (DBT), Government of India, for financial support through a research grant (102/IFD/SAN/2161/2013-14). S.D. thanks DBT for award of the Senior Research Fellowship.

Received 16 October 2015; revised accepted 18 August 2016

doi: 10.18520/cs/v112/i02/279-294