

Review

The impact of telomere-to-telomere genome assembly in the plant pan-genomics era

Yuta Aoyagi Blue, Hideaki Iimura, Mitsuhiro P. Sato and Kenta Shirasawa*

Department of Frontier Research and Development, Kazusa DNA Research Institute, 2-6-7 Kazusa-Kamatari, Kisarazu, Chiba 292-0818, Japan

Advances in sequencing technologies have enabled the determination of genome sequences of multiple lines within a single species. Comparative analysis of multiple genome sequences reveals all genes present within a species, providing insight into the genetic mechanisms that lead to the establishment of species. Highly accurate pan-genome analysis requires telomere-to-telomere gapless genome assembly, providing an ultimate genome sequence that covers all chromosomal regions without any undetermined nucleotide sequences. This review describes the genome sequencing technologies and sophisticated bioinformatics required for telomere-to-telomere gapless genome assembly, as well as a genetic mapping that can evaluate the accuracy of telomere-to-telomere genome assembly. Pan-genome analyses may contribute to the understanding of genetic mechanisms not only within a single species but also across species.

Key Words: genome project, long-read sequencing technology, next-generation sequencing technology, telomere-to-telomere sequence.

Introduction

Pan-genome analysis is an approach to understanding genetic components in a species by sequencing the genomes of multiple individuals (or cultivars) that cover the genetic diversity of that species (Bayer *et al.* 2020, Schreiber *et al.* 2024). Significant advances in genome sequencing technology over the past 20 years have enabled *de novo* genome assembly of multiple lines (Kong *et al.* 2023, Nurk *et al.* 2022). The completeness of genome assembly can be categorized on three levels: 1) contig-level sequences generated by assembling raw reads from a sequencer; 2) scaffold-level sequences, in which contigs are connected using information about reads that bridge two contigs; and 3) chromosome-level sequences, in which scaffold sequences are aligned on chromosomes of targets (Rhie *et al.* 2021).

In pan-genome analysis, the chromosome-level sequences of multiple lines have been required for analyses of comparative genome structures. This has enabled the detection of variations in chromosome structure (Bayer *et al.* 2020, Schreiber *et al.* 2024), such as translocations and inversions, as well as the presence or absence of gene variations and variations in gene copy number. This approach has

resulted in the identification of new genetic components that could not be detected by analysis of a single genome. Pan-genome analysis has enabled the classification of genes into two groups: core and dispensable genes, the latter of which are also known as accessory genes. Core genes are those detected in all individuals analyzed, suggesting that they may be essential for the species, whereas dispensable genes are present in some (not all) individuals of a species.

As of 2021, chromosome-level genome sequences had been reported in more than 100 plant species (Shirasawa *et al.* 2021). However, completeness of the genome sequences varied among plant species because of gaps in which nucleotide sequences were undetermined and/or because of the existence of sequences that could not be assigned to any chromosome. This situation would therefore prevent the definitive determination of the presence or absence of variations, as it is difficult to prove the non-existence of genes.

Most plant genomes are complex, being large in size, with high repeat content and heterozygosity, as well as variations in ploidy. Differences in haplotype are frequently ignored during chromosome-level genome assembly that involves the generation of consensus sequences from two haploid sequences without separating haplotype alleles. Haplotype-phased assembly plays a crucial role in detecting sequence and structure variants and in analyzing allele-specific functions (Zhou *et al.* 2020).

Recent advances in high-quality long-read sequencing methods, together with mate-pair sequencing, linked-read sequencing, Hi-C, optical mapping, and ultra-long reads,

Communicated by Sachiko Isobe

Received September 3, 2024. Accepted October 19, 2024.

First Published Online in J-STAGE on February 21, 2025.

*Corresponding author (e-mail: shirasaw@kazusa.or.jp)



are being used more frequently in analyzing plant genomes (Xie *et al.* 2024). The integration of these technologies has enabled the construction of telomere-to-telomere (T2T) haplotype-phased assembly (Nurk *et al.* 2022). High-fidelity (HiFi) long reads (PacBio, Menlo Park, CA, USA), due to their high accuracy and long read lengths, are mostly used for determining the nucleotide sequences of entire genomes, from a telomere on one end to a telomere on the other end of a chromosome without any gaps. Although these HiFi reads can resolve repetitive sequences, repetitive regions longer than HiFi reads remain unresolved. ONT ultra-long reads (Oxford Nanopore Technologies, Oxford, UK) can help resolve the repetitive sequences in plant genomes with high repetitive contents, although at relatively lower accuracy. This review describes recent advances in plant pan-genome analysis, including the introduction of sequencing and bioinformatic technologies that enable T2T-level genome assembly in plant species.

Classical genome sequencing in plants

Genome sequencing in higher plants started with Arabidopsis (The Arabidopsis Genome Initiative 2000), a member of the Brassicaceae. Arabidopsis has been recognized as a model for plant research because of its small-sized genome, rapid life cycle, and compact size, making it suitable for growth under laboratory conditions. The genome of rice, a model for monocots as well as an important cereal for staple food supply, was also one of the first to be analyzed (International Rice Genome Sequencing Project 2005). Since then, the genome sequences of several agronomically important crops, including vegetables, fruiting trees, and ornamental flowers, were analyzed, mostly for breeding purposes (Marks *et al.* 2021). Subsequently, the genomes of several non-crop wild plant species, including weeds, have been sequenced (Marks *et al.* 2021). Although this genome information has been used as “references” to understand the genetic and genomic bases of these species, these references were found to be insufficient to understand the genetics and genomics of these species. This issue has been partially addressed by resequencing analyses of the genomes of multiple individuals or cultivars of several species, including the Arabidopsis (The 1001 Genomes Consortium 2016) and rice (Wang *et al.* 2018). However, single nucleotide polymorphism was a major target since it was difficult to detect other types of sequence variations, such as insertions and deletions, by the short-read sequencing technology.

Plant genome sequences determined by next-generation sequencing technology

Advances in plant genomics have been associated with advances in sequencing technology. The genomes of Arabidopsis and rice were initially sequenced using the Sanger method (International Rice Genome Sequencing Project

2005, The Arabidopsis Genome Initiative 2000). In the genome projects, a clone-by-clone strategy was employed to determine genome sequences at the chromosome level. Due to high costs and limitations in techniques, however, the resultant genome sequences possess gap regions, in which nucleotide sequences were not determined. Next-generation sequencing (NGS) methods (Heather and Chain 2016) have introduced a paradigm shift in both genomics and biology in the mid to late 2000s. NGS technologies are based on massive parallel reactions of sequencing-by-synthesis (Illumina, San Diego, CA, USA) or sequencing-by-ligation (ThermoFisher Scientific, Waltham, MA, USA), yielding whole-genome shotgun sequence reads at an Mb- or Gb-scale during one experiment and at low cost. NGS enabled the sequencing of various non-model plants, including vegetables, fruiting trees, and ornamental flowers (Marks *et al.* 2021). Although the resultant genome assemblies were highly fragmented because of the short reads from the NGS technologies, this information was sufficient to list gene sets in the genomes. These sequences were subsequently expanded to chromosome levels by the development of scaffolding technologies, including optical mapping (Bionano Genomics, San Diego, CA, USA) and Hi-C (Dudchenko *et al.* 2017). These technologies enabled the chromosome-level or chromosome-scale assembly of the genomes of more than 100 plant species by 2021, although the quality of these chromosome-level assemblies varied in genome coverage and gap contents (Shirasawa *et al.* 2021). It might be difficult for the NGS and Sanger methods to assemble centromeric and peri-centromeric regions comprised of repetitive sequences like transposable elements and clusters of repeated ribosomal DNA genes. However, in case of tomato, for which the first chromosome-level genome assembly is established with the short-read sequencing technologies (The Tomato Genome Consortium 2012), a ribosomal DNA arrays spanning a 15 Mb stretch at the end of a chromosome has been finally assembled with a long-read sequencing technology as mentioned below (Shirasawa and Ariizumi 2024).

Haplotype-resolved genome assembly

Most practical reference genome sequences are haploid genome sequence. Pure lines of autogamous diploid plants have highly homozygous genomes. Thus, in assembling the genomes of these homozygotes, it is unnecessary to distinguish between pairs of homologous chromosomes containing the same base sequences. In allogamous and/or polyploid plants, however, the genome assembly of highly heterozygous individuals complicate assembly graphs, increasing the numbers of misassemblies and redundant contigs (Mochizuki *et al.* 2023). Additionally, typical haploid genome assembly can generate mosaic sequences that are mixtures of two haplotypes with heterozygous regions in a single chromosome (Delorean *et al.* 2023). The assembly of genomes that does not accurately represent the true genome

may be misleading in downstream analysis. Advances in sequencing and bioinformatics methods may allow the generation of haplotype-resolved genome sequences, even in allogamous plants, polyploid species, and interspecific hybrids.

Plant genome sequences at the T2T level

Haploids have been shown effective for improving the accuracy of genome assembly. Because the Human Genome Project used diploid cells of multiple individuals, the assembled human genomes contained euchromatic gaps caused by differences in segmental duplications (International Human Genome Sequencing Consortium 2004). The T2T Consortium used a complete hydatidiform mole (CHM) derived from a specific individual (Nurk *et al.* 2022). CHM is a tumor that develops from an abnormal fertilized egg in which the nucleus of the egg disappears and the haploid genome of the sperm doubles. The homozygous CHM genome was sequenced using various methods, such as HiFi long-read technology (PacBio), ultralong-read sequencing (Oxford Nanopore Technologies), PCR-free sequencing (Illumina), high-throughput chromosome conformation (Arima Genomics, Carlsbad, CA, USA), optical maps (Bionano Genomics), and single-cell DNA template strand sequencing (10X Genomics, Pleasanton, CA, USA). The first T2T genome assembly in plants involved the assembly of two of the ten chromosomes in maize (Liu *et al.* 2020).

Through a search with a term of “telomere-to-telomere” to NCBI PubMed database (<https://pubmed.ncbi.nlm.nih.gov>), T2T genome assemblies have been found to date in as many as 121 assemblies of 65 plant species (Table 1). T2T genome assembly in plants has shed light on the “dark matter” not detected in the genome sequences determined by classical sequencing and bioinformatics technologies, e.g., repetitive sequences associated with centromeric and pericentromeric regions (Sato *et al.* 2023) and clusters consisting of highly repeated sequences like ribosomal DNA genes (Shirasawa and Ariizumi 2024). These findings suggest that sophisticated bioinformatics technology is essential to accelerate T2T genome assembly in plant species.

Sequencing technologies and bioinformatics tools for T2T genome assembly

Telomere-to-Telomere genome assembly

Scaffolding (grouping, ordering, and orienting contigs) has been shown necessary to improve continuity and generate telomere-to-telomere assembly. Various types of long-distance linkage information, including optical maps and Hi-C/Omni-C contact maps, are used to construct chromosome-level scaffolds. These are used in genome assembly projects that aim to generate high-quality reference genome assemblies for diverse species, such as the Darwin Tree of Life Project (The Darwin Tree of Life Project

Consortium 2022) (<https://www.darwintreeoflife.org/>) and the Vertebrate Genomes Project (Rhie *et al.* 2021) (<https://vertebrategenomesproject.org>).

Optical mapping involves the use of a light microscope-based technique to physically locate nicking enzyme recognition sites in the genome to produce DNA sequence fingerprints (Schwartz *et al.* 1993, Yuan *et al.* 2020). Optical maps generated using optical mapping technologies, such as Bionano (Bionano Genomics), provide information on the physical location and relative separation of labeling sites and have been used to improve contiguity and validate the genome assembly (Yuan *et al.* 2020).

Hi-C/Omni-C reads and information on long-range interactions are also valuable for scaffolding (Burton *et al.* 2013) and for the phasing of haplotypes. This technology has an advantage over optical mapping, in that the latter requires a large amount of high-molecular-weight genomic DNA, whereas Hi-C/Omni-C requires only chromatin from a single individual and short-read sequencing. Hi-C/Omni-C scaffolding has therefore been adopted as a major strategy in many studies of T2T genome assembly. For example, Hi-C/Omni-C scaffolding has been performed using tools such as YaHS (Zhou *et al.* 2023) and SALSA (Ghurye *et al.* 2017, 2019).

Telomeres in assembled genome sequences are identified by detecting telomeric repeat sequences using tools such as a Telomere Identification toolKit (tidk) (Brown *et al.* 2023). A telomeric repeat (TTTAGGG)_n, first detected in *Arabidopsis thaliana* (Richards and Ausubel 1988), has been widely identified in plant genomes (Fuchs *et al.* 1995). Telomeric repeats in many plant species have been collected in comprehensive databases such as TeloBase (Lyčka *et al.* 2024) (<http://cfb.ceitec.muni.cz/telobase/>). In addition to the telomere detection, identifying centromeric regions could also validate the completeness of the genome assemblies, for which CentroMiner implemented in quarTeT (Lin *et al.* 2023) and TRASH (Włodzimierz *et al.* 2023) are available. These tools and databases are valuable for identifying telomeres and centromeres in assembled genome sequences.

Haplotype-phased genome assembly

Most plant genomes are complex structures, being large in size, having a high repeat content, and having high heterozygosity and polyploidy. Haplotype phasing in plant genomes could be achieved with trio data, sequencing reads obtained from both parents of a target individual, and/or high-throughput chromosome conformation capture (Hi-C/Omni-C), the latter of which provide information on long-range chromatin interactions. Trio binning partitions the target long sequence reads from both parents, followed by two separate assemblies of the partitioned reads (Koren *et al.* 2018). On the other hand, because Hi-C interactions occur more frequently within the same chromosome rather than between different chromosomes, with interaction frequency being inversely related to genome distance

Table 1. Statistics of plant genomes sequenced at the telomere-to-telomere level

Plant species name	Accession name ^a	Basic chr. no. ^b	Zygosity of the accession	Ploidy level of the accession	Total length of assembled sequences ^c	Reference DOI
<i>Actinidia chinensis</i>	Donghong	29	Heterozygous	Diploid	608 Mb	10.1016/j.molp.2022.12.022
<i>Actinidia chinensis</i>	Hongyang	29	Heterozygous	Diploid	606 Mb and 600 Mb	10.1093/hr/uhac264
<i>Actinidia eriantha</i>	Midao 31	29	Heterozygous	Diploid	619 Mb and 612 Mb	10.1186/s43897-023-00052-5
<i>Actinidia latifolia</i>	Kuoye	29	Heterozygous	Diploid	641 Mb	10.1016/j.molp.2022.12.022
<i>Arabidopsis thaliana</i>	Col-0	5	Homozygous	Diploid	134 Mb	10.1016/j.gpb.2021.08.003
<i>Armoracia rusticana</i>	HD15	16	Heterozygous	Allotetraploid	610 Mb	10.1038/s41467-023-39800-y
<i>Brassica napus</i>	Xiang5A	19	Homozygous	Allotetraploid	1.0 Gb	10.1093/hr/uhad171
<i>Brassica rapa</i>	Chiifu-401-42	10	Homozygous	Diploid	425 Mb	10.1111/pbi.14015
<i>Capsicum annuum</i>	G1-36576	12	Homozygous	Doubled haploid	3.1 Gb	10.1038/s41467-024-48643-0
<i>Capsicum rhomboideum</i>	PI 645680	13	n.a.	Diploid	1.7 Gb	10.1038/s41467-024-48643-0
<i>Chaenomeles speciosa</i>	n.a.	17	Heterozygous	Diploid	632 Mb	10.1093/hr/uhad183
<i>Citrullus amarus</i>	PI 189225	11	Homozygous	Diploid	378 Mb	10.1038/s41588-024-01823-6
<i>Citrullus amarus</i>	PI 271769	11	Homozygous	Diploid	378 Mb	10.1038/s41588-024-01823-6
<i>Citrullus amarus</i>	PI 296341-FR	11	Homozygous	Diploid	381 Mb	10.1038/s41588-024-01823-6
<i>Citrullus amarus</i>	PI 482276	11	Homozygous	Diploid	382 Mb	10.1038/s41588-024-01823-6
<i>Citrullus amarus</i>	RCAT 055816	11	Homozygous	Diploid	380 Mb	10.1038/s41588-024-01823-6
<i>Citrullus colocynthis</i>	PI 525081	11	Homozygous	Diploid	379 Mb	10.1038/s41588-024-01823-6
<i>Citrullus colocynthis</i>	PI 537300	11	Homozygous	Diploid	380 Mb	10.1038/s41588-024-01823-6
<i>Citrullus colocynthis</i>	PI 632755	11	Homozygous	Diploid	379 Mb	10.1038/s41588-024-01823-6
<i>Citrullus colocynthis</i>	PI 652554	11	Homozygous	Diploid	361 Mb	10.1038/s41588-024-01823-6
<i>Citrullus ecirrhosus</i>	PI 673135	11	Homozygous	Diploid	402 Mb	10.1038/s41588-024-01823-6
<i>Citrullus lanatus</i>	AllSugar	11	Homozygous	Diploid	369 Mb	10.1038/s41588-024-01823-6
<i>Citrullus lanatus</i>	Calhoun Gray	11	Homozygous	Diploid	370 Mb	10.1038/s41588-024-01823-6
<i>Citrullus lanatus</i>	Charleston Gray	11	Homozygous	Diploid	369 Mb	10.1038/s41588-024-01823-6
<i>Citrullus lanatus</i>	DBHZGua	11	Homozygous	Diploid	370 Mb	10.1038/s41588-024-01823-6
<i>Citrullus lanatus</i>	G42	11	Homozygous	Diploid	369 Mb	10.1016/j.molp.2022.06.010
<i>Citrullus lanatus</i>	G42	11	Homozygous	Diploid	369 Mb	10.1038/s41588-024-01823-6
<i>Citrullus lanatus</i>	HeiShanRen	11	Homozygous	Diploid	371 Mb	10.1038/s41588-024-01823-6
<i>Citrullus lanatus</i>	PI 254622	11	Homozygous	Diploid	374 Mb	10.1038/s41588-024-01823-6
<i>Citrullus lanatus</i>	PI 288522	11	Homozygous	Diploid	368 Mb	10.1038/s41588-024-01823-6
<i>Citrullus lanatus</i>	PI 381740	11	Homozygous	Diploid	370 Mb	10.1038/s41588-024-01823-6
<i>Citrullus lanatus</i>	PKR6	11	Homozygous	Diploid	371 Mb	10.1038/s41588-024-01823-6
<i>Citrullus lanatus</i>	SanBaiGua	11	Homozygous	Diploid	370 Mb	10.1038/s41588-024-01823-6
<i>Citrullus lanatus</i>	ShiHong No. 2	11	Homozygous	Diploid	369 Mb	10.1038/s41588-024-01823-6
<i>Citrullus lanatus</i>	Sugarlee	11	Homozygous	Diploid	369 Mb	10.1038/s41588-024-01823-6
<i>Citrullus mucosospermus</i>	PI 532732	11	Homozygous	Diploid	371 Mb	10.1038/s41588-024-01823-6
<i>Citrullus mucosospermus</i>	PI 595203	11	Homozygous	Diploid	371 Mb	10.1038/s41588-024-01823-6
<i>Citrullus naudinianus</i>	PI 596694	11	Homozygous	Diploid	365 Mb	10.1038/s41588-024-01823-6
<i>Citrullus rehmii</i>	PI 670011	11	Homozygous	Diploid	414 Mb	10.1038/s41588-024-01823-6
<i>Cucumis melo</i>	Kuizilikjiz	12	Homozygous	Diploid	379 Mb	10.1093/hr/uhad189
<i>Cucumis melo</i>	PI511890	12	Homozygous	Diploid	375 Mb	10.1111/tpj.16705
<i>Daucus carota</i>	Kurodagosun	9	Homozygous	Diploid	430 Mb	10.1093/hr/uhad103
<i>Dianthus caryophyllus</i>	Baltico	15	Heterozygous	Diploid	564 Mb and 568 Mb	10.1093/hr/uhad244
<i>Echinocloa phyllopogon</i>	R511	18	Homozygous	Allotetraploid	1.0 Gb	10.1093/dnares/dsad023
<i>Eleocharis dulcis</i>	n.a.	111	Heterozygous	Diploid	493 Mb	10.1038/s41597-024-03717-y
<i>Ficus hispida</i>	n.a.	14	Heterozygous	Diploid	372 Mb	10.1093/hr/uhad257
<i>Fragaria vesca</i>	Hawaii 4	7	Homozygous	Diploid	221 Mb	10.1093/hr/uhad027
<i>Glycine max</i>	Lee	20	Homozygous	Diploid	1.0 Gb	10.1002/tpg2.20382
<i>Glycine max</i>	Williams 82	20	Homozygous	Diploid	1.0 Gb	10.1002/tpg2.20382
<i>Glycine max</i>	Wm82-NJAU	20	Homozygous	Diploid	1.0 Gb	10.1016/j.molp.2023.08.012
<i>Glycine max</i>	Yundou1	20	Homozygous	Diploid	1.0 Gb	10.1016/j.xplc.2024.100919
<i>Glycine soja</i>	Yesheng71	20	Homozygous	Diploid	1.0 Gb	10.1016/j.xplc.2024.100919
<i>Gossypium raimondii</i>	D5-3	13	Homozygous	Diploid	776 Mb	10.1038/s41588-024-01877-6
<i>Gynostemma pentaphyllum</i>	n.a.	11	n.a.	Diploid	599 Mb	10.1016/j.xplc.2024.100932
<i>Ipomoea cairica</i>	n.a.	15	Heterozygous	Diploid	733 Mb	10.1093/g3journal/jkac187
<i>Jasminum sambac</i>	n.a.	13	Heterozygous	Diploid	495 Mb	10.1093/jxb/erac464
<i>Lactuca sativa</i>	PKU06	9	Homozygous	Diploid	2.6 Gb	10.1016/j.xplc.2024.101011
<i>Lolium perenne</i>	Kyuss	7	Homozygous	Doubled haploid	2.3 Gb	10.1093/gbe/evab159
<i>Mangifera indica</i>	Irwin	20	Heterozygous	Diploid	365 Mb	10.1002/tpg2.20441
<i>Manihot esculenta</i>	Xinxuan 048	18	Heterozygous	Diploid	665 Mb	10.1093/hr/uhad200
<i>Momordica charantia</i>	Jin ling zi or Lai pu tao	11	Homozygous	Diploid	296 Mb	10.1093/hr/uhac228
<i>Morella rubra</i>	Zaojia	8	Heterozygous	Diploid	293 Mb	10.1093/hr/uhac033

Table 1. (continued)

Plant species name	Accession name ^a	Basic chr. no. ^b	Zygosity of the accession	Ploidy level of the accession	Total length of assembled sequences ^c	Reference DOI
<i>Morus notabilis</i>	n.a.	6	Heterozygous	Diploid	410 Mb	10.1093/hr/uhad111
<i>Musa acuminata</i>	Baxijiao	11	Heterozygous	Triploid	477 Mb, 477 Mb, and 470 Mb	10.1093/hr/uhad153
<i>Musa acuminata</i>	DH-Pahang	11	Homozygous	Doubled haploid	485 Mb	10.1038/s42003-021-02559-3
<i>Musa acuminata</i>	n.a.	11	Heterozygous	Diploid	470 Mb and 470 Mb	10.1038/s41597-023-02546-9
<i>Oldenlandia diffusa</i>	n.a.	16	Homozygous	Diploid	500 Mb	10.1093/dnares/dsae012
<i>Olea europaea</i>	Leccino	23	n.a.	Diploid	1.3 Gb	10.1093/hr/uhae168
<i>Oryza barthii</i>	NH278	12	Homozygous	Diploid	350 Mb	10.1111/jipb.13607
<i>Oryza barthii</i>	NH279	12	Homozygous	Diploid	349 Mb	10.1111/jipb.13607
<i>Oryza barthii</i>	NH280	12	Homozygous	Diploid	349 Mb	10.1111/jipb.13607
<i>Oryza barthii</i>	NH281	12	Homozygous	Diploid	351 Mb	10.1111/jipb.13607
<i>Oryza barthii</i>	NH283	12	Homozygous	Diploid	349 Mb	10.1111/jipb.13607
<i>Oryza barthii</i>	NH284	12	Homozygous	Diploid	348 Mb	10.1111/jipb.13607
<i>Oryza barthii</i>	NH285	12	Homozygous	Diploid	348 Mb	10.1111/jipb.13607
<i>Oryza glaberrima</i>	NH266	12	Homozygous	Diploid	344 Mb	10.1111/jipb.13607
<i>Oryza glaberrima</i>	NH267	12	Homozygous	Diploid	347 Mb	10.1111/jipb.13607
<i>Oryza glaberrima</i>	NH268	12	Homozygous	Diploid	346 Mb	10.1111/jipb.13607
<i>Oryza glaberrima</i>	NH269	12	Homozygous	Diploid	348 Mb	10.1111/jipb.13607
<i>Oryza glaberrima</i>	NH270	12	Homozygous	Diploid	347 Mb	10.1111/jipb.13607
<i>Oryza glaberrima</i>	NH271	12	Homozygous	Diploid	345 Mb	10.1111/jipb.13607
<i>Oryza glaberrima</i>	NH272	12	Homozygous	Diploid	346 Mb	10.1111/jipb.13607
<i>Oryza glaberrima</i>	NH273	12	Homozygous	Diploid	346 Mb	10.1111/jipb.13607
<i>Oryza glaberrima</i>	NH274	12	Homozygous	Diploid	344 Mb	10.1111/jipb.13607
<i>Oryza glaberrima</i>	NH275	12	Homozygous	Diploid	348 Mb	10.1111/jipb.13607
<i>Oryza sativa</i>	HN	12	Homozygous	Diploid	394 Mb	10.1111/pbi.13880
<i>Oryza sativa</i>	J4155S	12	Homozygous	Diploid	395 Mb	10.1111/pbi.13880
<i>Oryza sativa</i>	LK638S	12	Homozygous	Diploid	396 Mb	10.1111/pbi.13880
<i>Oryza sativa</i>	XL628S	12	Homozygous	Diploid	398 Mb	10.1111/pbi.13880
<i>Oryza</i> sp.	NH277	12	Homozygous	Diploid	347 Mb	10.1111/jipb.13607
<i>Oryza</i> sp.	NH282	12	Homozygous	Diploid	348 Mb	10.1111/jipb.13607
<i>Panax ginseng</i>	n.a.	24	Heterozygous	Allotetraploid	3.5 Gb	10.1093/hr/uhae107
<i>Panicum miliaceum</i>	AJ8	18	n.a.	Allotetraploid	835 Mb	10.1038/s41597-024-03489-5
<i>Penthorum chinense</i>	n.a.	9	Homozygous	Diploid	258 Mb	10.1093/hr/uhad274
<i>Persea americana</i>	West Indian	12	Heterozygous	Diploid	842 Mb	10.1093/hr/uhae119
<i>Peucedanum praeruptorum</i>	n.a.	11	Heterozygous	Diploid	1.8 Gb	10.1093/gigascience/giae025
<i>Phragmites australis</i>	CN	25	n.a.	Allotetraploid	920 Mb	10.1038/s42003-024-06660-1
<i>Populus alba</i> × <i>Populus tremula</i>	84K	19	Heterozygous	Interspecific hybrid	417 Mb and 400 Mb	10.1093/plphys/kiae078
<i>Populus tremula</i> × <i>Populus alba</i>	INRA 717-1B4	19	Heterozygous	Interspecific hybrid	394 Mb and 403 Mb	10.1111/tpj.16454
<i>Prunus salicina</i>	Fengtangli	8	Heterozygous	Diploid	251 Mb and 251 Mb	10.1093/hr/uhae109
<i>Pyrus pyrifolia</i>	Yunhong No. 1	17	Heterozygous	Diploid	501 Mb	10.1093/hr/uhad201
<i>Quercus variabilis</i>	n.a.	12	Heterozygous	Diploid	789 Mb and 768 Mb	10.3389/fpls.2023.1290913
<i>Scutellaria baicalensis</i>	n.a.	9	Heterozygous	Diploid	385 Mb	10.1093/hr/uhad235
<i>Sesbania cannabina</i>	LJ5	12	n.a.	Allotetraploid	2.1 Gb	10.1007/s11427-023-2463-y
<i>Sorghum bicolor</i>	BTx623	10	Homozygous	Diploid	720 Mb	10.1002/imt2.193
<i>Sorghum bicolor</i>	Cuohu Bazi	10	Homozygous	Diploid	725 Mb	10.1038/s41597-024-03664-8
<i>Sorghum bicolor</i>	Hongyingzi	10	Homozygous	Diploid	746 Mb	10.1016/j.xplc.2024.100933
<i>Sorghum bicolor</i>	Huandianuo	10	Homozygous	Diploid	739 Mb	10.1016/j.xplc.2024.100933
<i>Sorghum bicolor</i>	Ji2055	10	Homozygous	Diploid	723 Mb	10.1002/imt2.193
<i>Theobroma grandiflorum</i>	1074	10	Heterozygous	Diploid	423 Mb	10.1093/gigascience/giae027
<i>Vaccinium duculoxii</i>	SGLD20220023	12	Heterozygous	Diploid	574 Mb	10.1093/hr/uhad209
<i>Vigna unguiculata</i>	Fengchan 6	11	Homozygous	Diploid	521 Mb	10.1111/pbi.14142
<i>Vitis</i> sp.	Thompson Seedless	19	Heterozygous	Diploid	505 Mb	10.1093/hr/uhad260
<i>Vitis vinifera</i>	Chasselas	19	Heterozygous	Diploid	500 Mb	10.1073/pnas.2403750121
<i>Vitis vinifera</i>	PN40024	19	Homozygous	Diploid	495 Mb	10.1093/hr/uhad061
<i>Vitis vinifera</i>	Ugni Blanc	19	Heterozygous	Diploid	495 Mb	10.1073/pnas.2403750121
<i>Vitis vinifera</i>	Yan73	19	Heterozygous	Diploid	502 Mb and 493 Mb	10.1093/hr/uhad205
<i>Zea mays</i>	B73-Ab10	10	Homozygous	Diploid	2.2 Gb	10.1186/s13059-020-02029-9
<i>Zea mays</i>	Mo17	10	Homozygous	Diploid	2.2 Gb	10.1038/s41588-023-01419-6
<i>Ziziphus jujuba</i>	Junzao	12	Heterozygous	Diploid	386 Mb	10.1093/hr/uhae071
<i>Ziziphus jujuba</i>	SZ	12	Heterozygous	Diploid	375 Mb	10.1093/hr/uhae071

^a n.a. indicates not available.^b Basic chromosome number (n).^c Multiple values indicate haplotype-resolved assembly.

(Lieberman-Aiden *et al.* 2009, Xu and Dixon 2020), the information on the long-range interactions from Hi-C/Omni-C is also used to phase haplotypes (Kronenberg *et al.* 2021). A combination of trio data and Hi-C/Omni-C, together with HiFi and ultra-long reads, could be useful for haplotype-phased assembly in Hifiasm (Cheng *et al.* 2021) and Verkko (Rautiainen *et al.* 2023).

Repetitive sequence annotation

Continuous genome assemblies can be constructed and long repeats can be resolved using long-read sequencing and assembly technology. Repetitive sequences in assembled genomes can be detected by *de novo*, homology-based, and structure-based methods (Liao *et al.* 2023), for which the pipeline of RepeatModeler2 is available (Flynn *et al.* 2020). *De novo* detection methods identify repetitive sequences without relying on the structures of repeat elements or their similarity to known repetitive sequences, with one *de novo* detection strategy using high-frequency *k*-mers and space seed extension, RepeatScout (Price *et al.* 2005). Homology-based detection methods, RepeatMasker (<https://www.repeatmasker.org>), identify repetitive sequences based on their similarity with known repeats. Many repetitive sequences from a wide range of species have been stored in databases, such as Repbase (Bao *et al.* 2015, Jurka 1998) and Dfam (Storer *et al.* 2021). Most repeats, especially transposable elements, have specific structures, similar to those of proteins or non-coding domains (Liao *et al.* 2023), which could be annotated by LTRharvest (Ellinghaus *et al.* 2008).

Structural and functional annotation of genes

The structures of protein-encoding genes can be predicted by four methods: 1) *ab initio*, 2) transcript-based, 3) homology-based prediction, and 4) integration method.

1) The *ab initio* methods, e.g., AUGUSTUS (Stanke *et al.* 2006) and GeneMark-ES (Lomsadze *et al.* 2005), predict gene structures based on information contained in the genomic sequence and statistical models, such as Hidden Markov Models. Extrinsic evidence can also be used to train the model.

2) Alignment of information on RNA-seq reads, expressed sequence tags, full-length cDNA sequences, and/or isoform transcriptome sequences, so-called Iso-Seq (PacBio), from the target species can be used as extrinsic evidence, improving the accuracy of gene prediction. Transcript-based methods of Trinity (Grabherr *et al.* 2011), StringTie (Pertea *et al.* 2015), and TransDecoder (<https://github.com/TransDecoder/TransDecoder>) predict gene structures based on transcripts assembled using short-read RNA-seq and full-length RNA from the target genome. This method is highly reliable because it uses mRNA information as evidence of expression. Since not all genes are expressed in a single tissue/organ, highly comprehensive gene annotation requires the collection of RNA-seq data from multiple tissues and developmental stages.

3) Homology-based prediction methods implemented in Spaln (Gotoh 2008, 2024) and GeMoMa (Keilwagen *et al.* 2016, 2018) are based on spliced alignments of protein sequences from closely related species with the target genome. A large number of protein sequences from a wide range of species have been published and deposited in public databases, such as RefSeq (O'Leary *et al.* 2016) and UniProt Knowledgebase (UniProt Consortium 2023). This method, however, is less sensitive when the only protein sequences available are from species that are evolutionarily distant from the target species.

4) Integration methods, for which EVIDENCEModeler (Haas *et al.* 2008) and GINGER (Taniguchi *et al.* 2023) are available, integrate gene structures from other predictive methods to construct consensus gene structures, with these structures being more accurate and comprehensive than those based on other methods.

Because each of these methods has advantages and disadvantages, combining different methods is essential for complete gene annotation. The gene prediction pipelines of BRAKER3 (Gabriel *et al.* 2024) and MAKER2 (Holt and Yandell 2011) could simplify the complicated procedure. Gene annotation using extrinsic data depends on the quality and quantity of these data. Gene prediction tools using deep learning have also been described, with gene annotation based on DNA sequences alone using deep learning showing high-quality gene prediction in eukaryotic genomes (Stiehler *et al.* 2020).

In addition to protein-encoding genes, non-coding RNAs (ncRNAs), including transfer RNAs, ribosomal RNAs, small nucleolar RNAs, micro RNAs, and small-interfering RNAs, have also been annotated with tools, such as tRNAscan-SE2.0 (Chan *et al.* 2021) and RNAmmer (Lagesen *et al.* 2007). RNA genes are usually annotated based on sequence- and structure-based alignments with known RNAs. ncRNAs should be annotated using Rfam, a database of ncRNAs (Kalvari *et al.* 2021), and Infernal, a tool for sequence- and structure-based RNA alignments (Nawrocki and Eddy 2013).

Functional annotation in general relies on similarities between sequences from the target genome and those with characterized function from other genomes. The accumulation of gene function data from diverse species has enabled the annotation of functional information on unknown genes with the function of thousands of known genes using local alignment search tools such as BLAST.

UniProtKB is a high-quality and comprehensive resource on protein sequences and functions of various species (UniProt Consortium 2023). To date, UniProtKB contains over 245 million entries (release 2024_03 of UniProtKB; <https://www.uniprot.org/>). Pfam is a database of protein families and domains that has been used to analyze protein sequences from novel genomes (Mistry *et al.* 2021). To date, Pfam 37.0 contains 21,979 entries (<http://pfam.xfam.org/>).

The Gene Ontology (GO) knowledge base provides a

comprehensive summary of gene structure and function. Gene function in GO has been described using three types (aspects) of functional characteristics: molecular function (molecular-level activity of the gene product), cellular component (location of the gene product), and biological process (biological program accomplished by multiple molecular activities) (The Gene Ontology Consortium 2023).

The Kyoto Encyclopedia of Genes and Genomes (KEGG) is a resource that integrates various biological objects categorized into systems, genomic, chemical, and health information (Kanehisa *et al.* 2023). The KEGG provides 16 manually curated databases in these four categories: molecular interaction, reaction, and relation networks in the systems information category; genes and proteins in the genomic information category; chemical compounds and reactions in the chemical information category; and human diseases and drugs in the health information category (Kanehisa *et al.* 2023).

The continued accumulation of sequence data and evidence from molecular, biochemical, and biophysical experiments, coupled with advances in informatics analysis technologies, is expected to facilitate comprehensive and accurate annotation of sequences and genes.

Validation of the accuracy of genome assembly

Following the completion of assembly using advanced sequencing and bioinformatics technologies, the accuracy of the sequences should be validated to confirm that these sequences represent the actual genome structure. The N50 value is an indicator of continuity in draft genomes but is influenced by chromosome length. Therefore, the N50 values become meaningless in the chromosome-level assemblies.

One frequent validation method is bioinformatics metrics of BUSCO to assess genome assembly and annotation completeness with single-copy orthologs (Simão *et al.* 2015), Merqury to evaluate assembly quality with *k*-mer spectrum of whole genome sequencing reads (Rhie *et al.* 2020), and LAI to indicate the assembly quality of the intergenic and repetitive sequence space by the amount of identifiable intact long-terminal repeat elements (Ou *et al.* 2018). QUAST is also a quality assessment tool for evaluating and comparing genome assemblies providing reports, summary tables and plots on contig sizes, misassemblies and structural variations, genome representation and its functional elements, and variations of the N50 based on aligned blocks (Gurevich *et al.* 2013).

Alternative validation method is genetic mapping. Specifically, a genetic map is constructed based on the frequency of recombination between markers using a mapping population, thereby showing the relative positions of markers on chromosomes. Based on comparison between the positions of markers on genetic maps and on the assembled contigs, contigs can be ordered correctly and oriented to

reconstruct chromosome-level scaffolds (Fierst 2015, Tang *et al.* 2015).

One importance drawback of genetic mapping is the requirement for mapping populations, especially in plants. It is not always possible to generate mapping population in plants. Woody plants require a long time to generate progenitors, and self-incompatible species require multiple individuals. These drawbacks may be resolved by using gametophytes, i.e., pollens in plants. Because each pollen possesses a haploid genome, consisting of recombinant chromosomes from the parent, each pollen could be recognized as an individual in a mapping population. Mapping based on haploid genomes in pollen has a great advantage over mapping based on diploid and polyploid genomes. Genotyping of a haploid requires only 1× coverage because of the absence of heterozygosity, although genotyping by sequencing generally requires 30× coverage. A tool, Scaffold Extender with Low Depth Linkage Analysis (SELDLA), has been developed to make linkage maps from low coverage data. A hybrid with distinguishable parent genomes can be regarded as the equivalent of two haploid organisms. SELDLA has been used to construct a linkage map of hybrids of two fish species, *Takifugu rubripes* and *Takifugu stictonotus*, from an average of 1.8× coverage data (Yoshitake *et al.* 2018). However, it may not be possible to prepare a sufficient population of organisms that are difficult to breed or cross, and more individuals may be required to construct a high-density linkage map. To solve these problems, a linkage map was constructed with sperm from *Gasterosteus nipponicus* (Yoshitake *et al.* 2022). A single-cell DNA library was constructed using Chromium (10x Genomics, Pleasanton, CA, USA) and a high-density linkage map was constructed from an average of 0.13× coverage data by SELDLA. Multiple displacement amplification, which efficiently amplifies a small amount of DNA, may also amplify the small amount of genomic DNA in a single pollen, such that it is sufficient for genotyping. The resultant high-density linkage maps would be useful for validation of assembled sequences by determining the positions and/or directions of contigs.

Future perspectives

Pan-genome analysis with T2T genome assembly can enhance understanding of the genetic and genomic components of target species. To date, 4,731 genome sequences are available for 1,875 plant species (by a search with a term of “Magnoliopsida” to NCBI Genome database; <https://www.ncbi.nlm.nih.gov/datasets/genome/>), and T2T assemblies have been completed for 121 genomes across 65 species (Table 1). However, while 111 of them have simple genomes, e.g., diploid and doubled haploid, only 10 plants possess complex genomes of allopolyploid and interspecific hybrid. No T2T assemblies have been so far reported for autopolyploidy. Moreover, >99% species in the plant kingdom have not yet been sequenced (Vallée *et al.* 2016).

Furthermore, as Marks *et al.* (2021) mentioned, there are substantial taxonomic gaps in the current plant genome research and are numerous disconnects between the native range of focal species and the national affiliation of the researchers studying them. Genomic information on plants that have not yet been employed for breeding and industrial purposes is required to understand and dissect the genetics and genomics of the species and utilize this information to enhance plant breeding. We propose that, by focusing on plants, animals, and microorganisms endemic to Japan, for instance, researchers in Japan shall describe the importance of the genetic diversity (Washoku BioGenome Consortium launched by Prof. S. Kuraku of National Institute of Genetics and his colleagues). This regional attempt should be expanded over the world under the concept of the Earth BioGenome Project to record all genetic and genomic information on the planet toward the future (Lewin *et al.* 2022). The challenge would open up neo pan-genomics that covers not only within a single species but also bridges plants, animals, and microorganisms. In addition to sequencing, data management, including curation, distribution, and maintenance, should also be considered, with artificial intelligence possibly being a new interface to handle massive amounts of genomic data in the era of pan-genomics.

Author Contribution Statement

KS designed the structure of the manuscript. All authors wrote the manuscript and approved the final version.

Acknowledgments

The authors thank Y. Kishida, S. Nakayama, and A. Watanabe (Kazusa DNA Research Institute) for technical assistance. This work was supported by JSPS KAKENHI (22H05172 and 22H05181) and the Kazusa DNA Research Institute Foundation.

Literature Cited

- Bao, W., K.K. Kojima and O. Kohany (2015) Repbase update, a database of repetitive elements in eukaryotic genomes. *Mob DNA* 6: 11.
- Bayer, P.E., A.A. Golicz, A. Scheben, J. Batley and D. Edwards (2020) Plant pan-genomes are the new reference. *Nat Plants* 6: 914–920.
- Brown, M., P.M. Gonzalez de la Rosa and M. Blaxter (2023) A Telomere Identification Toolkit. *Zunodo* doi: 10.5281/zenodo.10091385.
- Burton, J.N., A. Adey, R.P. Patwardhan, R. Qiu, J.O. Kitzman and J. Shendure (2013) Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. *Nat Biotechnol* 31: 1119–1125.
- Chan, P.P., B.Y. Lin, A.J. Mak and T.M. Lowe (2021) tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Res* 49: 9077–9096.
- Cheng, H., G.T. Concepcion, X. Feng, H. Zhang and H. Li (2021) Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. *Nat Methods* 18: 170–175.
- Delorean, E.E., R.C. Youngblood, S.A. Simpson, A.N. Schoonmaker, B.E. Scheffler, W.B. Rutter and A.M. Hulse-Kemp (2023) Representing true plant genomes: haplotype-resolved hybrid pepper genome with trio-binning. *Front Plant Sci* 14: 1184112.
- Dudchenko, O., S.S. Batra, A.D. Omer, S.K. Nyquist, M. Hoeger, N.C. Durand, M.S. Shamim, I. Machol, E.S. Lander, A.P. Aiden *et al.* (2017) *De novo* assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* 356: 92–95.
- Ellinghaus, D., S. Kurtz and U. Willhoeft (2008) *LTRharvest*, an efficient and flexible software for *de novo* detection of LTR retrotransposons. *BMC Bioinformatics* 9: 18.
- Fierst, J.L. (2015) Using linkage maps to correct and scaffold *de novo* genome assemblies: methods, challenges, and computational tools. *Front Genet* 6: 220.
- Flynn, J.M., R. Hubley, C. Goubert, J. Rosen, A.G. Clark, C. Feschotte and A.F. Smit (2020) RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci USA* 117: 9451–9457.
- Fuchs, J., A. Brandes and I. Schubert (1995) Telomere sequence localization and karyotype evolution in higher plants. *Plant Syst Evol* 196: 227–241.
- Gabriel, L., T. Bruna, K.J., Hoff, M. Ebel, A. Lomsadze, M. Borodovsky and M. Stanke (2024) BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. *Genome Res* 34: 769–777.
- Ghurye, J., M. Pop, S. Koren, D. Bickhart and C.S. Chin (2017) Scaffolding of long read assemblies using long range contact information. *BMC Genomics* 18: 527.
- Ghurye, J., A. Rhie, B.P. Walenz, A. Schmitt, S. Selvaraj, M. Pop, A.M. Phillippy and S. Koren (2019) Integrating Hi-C links with assembly graphs for chromosome-scale assembly. *PLoS Comput Biol* 15: e1007273.
- Gotoh, O. (2008) A space-efficient and accurate method for mapping and aligning cDNA sequences onto genomic sequence. *Nucleic Acids Res* 36: 2630–2638.
- Gotoh, O. (2024) Spaln3: improvement in speed and accuracy of genome mapping and spliced alignment of protein query sequences. *Bioinformatics* 40: btae517.
- Grabherr, M.G., B.J. Haas, M. Yassour, J.Z. Levin, D.A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* 29: 644–652.
- Gurevich, A., V. Saveliev, N. Vyahhi and G. Tesler (2013) QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 29: 1072–1075.
- Haas, B.J., S.L. Salzberg, W. Zhu, M. Pertea, J.E. Allen, J. Orvis, O. White, C.R. Buell and J.R. Wortman (2008) Automated eukaryotic gene structure annotation using EVIDENCEModeler and the Program to Assemble Spliced Alignments. *Genome Biol* 9: R7.
- Heather, J.M. and B. Chain (2016) The sequence of sequencers: The history of sequencing DNA. *Genomics* 107: 1–8.
- Holt, C. and M. Yandell (2011) MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* 12: 491.
- International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. *Nature* 431: 931–945.

- International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. *Nature* 436: 793–800.
- Jurka, J. (1998) Repeats in genomic DNA: mining and meaning. *Curr Opin Struct Biol* 8: 333–337.
- Kalvari, I., E.P. Nawrocki, N. Ontiveros-Palacios, J. Argasinska, K. Lamkiewicz, M. Marz, S. Griffiths-Jones, C. Toffano-Nioche, D. Gautheret, Z. Weingberg *et al.* (2021) Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Res* 49: D192–D200.
- Kanehisa, M., M. Furumichi, Y. Sato, M. Kawashima and M. Ishiguro-Watanabe (2023) KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res* 51: D587–D592.
- Keilwagen, J., M. Wenk, J.L. Erickson, M.H. Schattat, J. Grau and F. Hartung (2016) Using intron position conservation for homology-based gene prediction. *Nucleic Acids Res* 44: e89.
- Keilwagen, J., F. Hartung, M. Paulini, S.O. Twardziok and J. Grau (2018) Combining RNA-seq data and homology-based gene prediction for plants, animals and fungi. *BMC Bioinformatics* 19: 189.
- Kong, W., Y. Wang, S. Zhang, J. Yu and X. Zhang (2023) Recent advances in assembly of complex plant genomes. *Genomics Proteomics Bioinformatics* 21: 427–439.
- Koren, S., A. Rhie, B.P. Walenz, A.T. Dilthey, D.M. Bickhart, S.B. Kingan, S. Hiendleder, J.L. Williams, T.P.L. Smith and A.M. Phillippy (2018) *De novo* assembly of haplotype-resolved genomes with trio binning. *Nat Biotechnol* 36: 1174–1182.
- Kronenberg, Z.N., A. Rhie, S. Koren, G.T. Concepcion, P. Peluso, K.M. Munson, D. Porubsky, K. Kuhn, K.A. Mueller, W.Y. Low *et al.* (2021) Extended haplotype-phasing of long-read *de novo* genome assemblies using Hi-C. *Nat Commun* 12: 1935.
- Lagesen, K., P. Hallin, E.A. Rødland, H.H. Stærfeldt, T. Rognes and D.W. Ussery (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res* 35: 3100–3108.
- Lewin, H.A., S. Richards, E.L. Aiden, M.L. Allende, J.M. Archibald, M. Bálint, K.B. Barker, B. Baumgartner, K. Belov, G. Bertorelle *et al.* (2022) The Earth BioGenome Project 2020: starting the clock. *Proc Natl Acad Sci USA* 119: e2115635118.
- Liao, X., W. Zhu, J. Zhou, H. Li, X. Xu, B. Zhang and X. Gao (2023) Repetitive DNA sequence detection and its role in the human genome. *Commun Biol* 6: 954.
- Lieberman-Aiden, E., N.L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B.R. Lajoie, P.J. Sabo, M.O. Dorschner *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326: 289–293.
- Lin, Y., C. Ye, X. Li, Q. Chen, Y. Wu, F. Zhang, R. Pan, S. Zhang, S. Chen, X. Wang *et al.* (2023) quarTeT: a telomere-to-telomere toolkit for gap-free genome assembly and centromeric repeat identification. *Hortic Res* 10: uhad127.
- Liu, J., A.S. Seetharam, K. Chougule, S. Ou, K.W. Swentowsky, J.I. Gent, V. Llaca, M.R. Woodhouse, N. Manchanda, G.G. Presting *et al.* (2020) Gapless assembly of maize chromosomes using long-read technologies. *Genome Biol* 21: 121.
- Lomsadze, A., V. Ter-Hovhannisyan, Y.O. Chernoff and M. Borodovsky (2005) Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res* 33: 6494–6506.
- Lyčka, M., M. Bubeník, M. Závodník, V. Peska, P. Fajkus, M. Demko, J. Fajkus and M. Fojtová (2024) TeloBase: a community-curated database of telomere sequences across the tree of life. *Nucleic Acids Res* 52: D311–D321.
- Marks, R.A., S. Hotaling, P.B. Frandsen and R. VanBuren (2021) Representation and participation across 20 years of plant genome sequencing. *Nat Plants* 7: 1571–1578.
- Mistry, J., S. Chuguransky, L. Williams, M. Qureshi, G.A. Salazar, E.L.L. Sonnhammer, S.C.E. Tosatto, L. Paladin, S. Raj, L.J. Richardson *et al.* (2021) Pfam: The protein families database in 2021. *Nucleic Acids Res* 49: D412–D419.
- Mochizuki, T., M. Sakamoto, Y. Tanizawa, T. Nakayama, G. Tanifuji, R. Kamikawa and Y. Nakamura (2023) A practical assembly guideline for genomes with various levels of heterozygosity. *Brief Bioinform* 24: bbad337.
- Nawrocki, E.P. and S.R. Eddy (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* 29: 2933–2935.
- Nurk, S., S. Koren, A. Rhie, M. Rautiainen, A.V. Bzikadze, A. Mikheenko, M.R. Vollger, N. Altomose, L. Uralsky, A. Gershman *et al.* (2022) The complete sequence of a human genome. *Science* 376: 44–53.
- O’Leary, N.A., M.W. Wright, J.R. Brister, S. Ciufu, D. Haddad, R. McVeigh, B. Rajput, B. Robbertse, B. Smith-White, D. Ako-Adjei *et al.* (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* 44: D733–D745.
- Ou, S., J. Chen and N. Jiang (2018) Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res* 46: e126.
- Pertea, M., G.M. Pertea, C.M. Antonescu, T.C. Chang, J.T. Mendell and S.L. Salzberg (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* 33: 290–295.
- Price, A.L., N.C. Jones and P.A. Pevzner (2005) *De novo* identification of repeat families in large genomes. *Bioinformatics* 21: i351–i358.
- Rautiainen, M., S. Nurk, B.P. Walenz, G.A. Logsdon, D. Porubsky, A. Rhie, E.E. Eichler, A.M. Phillippy and S. Koren (2023) Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nat Biotechnol* 41: 1474–1482.
- Rhie, A., B.P. Walenz, S. Koren and A.M. Phillippy (2020) Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol* 21: 245.
- Rhie, A., S.A. McCarthy, O. Fedrigo, J. Damas, G. Formenti, S. Koren, M. Uliano-Silva, W. Chow, A. Fungtammasan, J. Kim *et al.* (2021) Towards complete and error-free genome assemblies of all vertebrate species. *Nature* 592: 737–746.
- Richards, E.J. and F.M. Ausubel (1988) Isolation of a higher eukaryotic telomere from *Arabidopsis thaliana*. *Cell* 53: 127–136.
- Sato, M.P., S. Iwakami, K. Fukunishi, K. Sugiura, K. Yasuda, S. Isobe and K. Shirasawa (2023) Telomere-to-telomere genome assembly of an allotetraploid pernicious weed, *Echinochloa phyllopogon*. *DNA Res* 30: dsad023.
- Schreiber, M., M. Jayakodi, N. Stein and M. Mascher (2024) Plant pangenomes for crop improvement, biodiversity and evolution. *Nat Rev Genet* 25: 563–577.
- Schwartz, D.C., X. Li, L.I. Hernandez, S.P. Ramnarain, E.J. Huff and Y.K. Wang (1993) Ordered restriction maps of *Saccharomyces cerevisiae* chromosomes constructed by optical mapping. *Science* 262: 110–114.
- Shirasawa, K., D. Harada, H. Hirakawa, S. Isobe and C. Kole (2021) Chromosome-level *de novo* genome assemblies of over 100 plant species. *Breed Sci* 71: 117–124.
- Shirasawa, K. and T. Ariizumi (2024) Near-complete genome assembly of tomato (*Solanum lycopersicum*) cultivar Micro-Tom. *Plant*

- Biotechnol 41: 367–374.
- Simão, F.A., R.M. Waterhouse, P. Ioannidis, E.V. Kriventseva and E.M. Zdobnov (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31: 3210–3212.
- Stanke, M., O. Keller, I. Gunduz, A. Hayes, S. Waack and B. Morgenstern (2006) AUGUSTUS: *ab initio* prediction of alternative transcripts. *Nucleic Acids Res* 34: W435–W439.
- Stiehler, F., M. Steinborn, S. Scholz, D. Dey, A.P.M. Weber and A.K. Denton (2020) Helixer: cross-species gene annotation of large eukaryotic genomes using deep learning. *Bioinformatics* 36: 5291–5298.
- Storer, J., R. Hubley, J. Rosen, T.J. Wheeler and A.F. Smit (2021) The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mob DNA* 12: 2.
- Tang, H., X. Zhang, C. Miao, J. Zhang, R. Ming, J.C. Schnable, P.S. Schnable, E. Lyons and J. Lu (2015) ALLMAPS: robust scaffold ordering based on multiple maps. *Genome Biol* 16: 3.
- Taniguchi, T., M. Okuno, T. Shinoda, F. Kobayashi, K. Takahashi, H. Yuasa, Y. Nakamura, H. Tanaka, R. Kajitani and T. Itoh (2023) GINGER: an integrated method for high-accuracy prediction of gene structure in higher eukaryotes at the gene and exon level. *DNA Res* 30: dsad017.
- The 1001 Genomes Consortium (2016) 1,135 Genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell* 166: 481–491.
- The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408: 796–815.
- The Darwin Tree of Life Project Consortium (2022) Sequence locally, think globally: The Darwin Tree of Life Project. *Proc Natl Acad Sci USA* 119: e2115642118.
- The Gene Ontology Consortium (2023) The Gene ontology knowledgebase in 2023. *Genetics* 224: iyad031.
- The Tomato Genome Consortium (2012) The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* 485: 635–641.
- UniProt Consortium (2023) UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res* 51: D523–D531.
- Vallée, G.C., D.S. Muñoz and D. Sankoff (2016) Economic importance, taxonomic representation and scientific priority as drivers of genome sequencing projects. *BMC Genomics* 17: 782.
- Wang, W., R. Mauleon, Z. Hu, D. Chebotarov, S. Tai, Z. Wu, M. Li, T. Zheng, R.R. Fuentes, F. Zhang *et al.* (2018) Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature* 557: 43–49.
- Wlodzimierz, P., M. Hong and I.R. Henderson (2023) TRASH: tandem repeat annotation and structural hierarchy. *Bioinformatics* 39: btad308.
- Xie, L., X. Gong, K. Yang, Y. Huang, S. Zhang, L. Shen, Y. Sun, D. Wu, C. Ye, Q.H. Zhu *et al.* (2024) Technology-enabled great leap in deciphering plant genomes. *Nat Plants* 10: 551–566.
- Xu, Z. and J.R. Dixon (2020) Genome reconstruction and haplotype phasing using chromosome conformation capture methodologies. *Brief Funct Genomics* 19: 139–150.
- Yoshitake, K., Y. Igarashi, M. Mizukoshi, S. Kinoshita, S. Mitsuyama, Y. Suzuki, K. Saito, S. Watabe and S. Asakawa (2018) Artificially designed hybrids facilitate efficient generation of high-resolution linkage maps. *Sci Rep* 8: 16104.
- Yoshitake, K., A. Ishikawa, R. Yonezawa, S. Kinoshita, J. Kitano and S. Asakawa (2022) Construction of a chromosome-level Japanese stickleback species genome using ultra-dense linkage analysis with single-cell sperm sequencing. *NAR Genom Bioinform* 4: lqac026.
- Yuan, Y., C.Y.L. Chung and T.F. Chan (2020) Advances in optical mapping for genomic research. *Comput Struct Biotechnol J* 18: 2051–2062.
- Zhou, C., S.A. McCarthy and R. Durbin (2023) YaHS: yet another Hi-C scaffolding tool. *Bioinformatics* 39: btac808.
- Zhou, Q., D. Tang, W. Huang, Z. Yang, Y. Zhang, J.P. Hamilton, R.G.F. Visser, C.W.B. Bachem, C.R. Buell, Z. Zhang *et al.* (2020) Haplotype-resolved genome analyses of a heterozygous diploid potato. *Nat Genet* 52: 1018–1023.