Compendium of Plant Genomes
Series Editor: Chittaranjan Kole

Frank M. You
Bourlaye Fofana *Editors*

The Flax Genome

Springer

Compendium of Plant Genomes

Series Editor

Chittaranjan Kole, President, International Climate Resilient Crop Genomics Consortium (ICRCGC), President, International Phytomedomics and Nutriomics Consortium (IPNC) and President, Genome India International (GII), Kolkata, India

Whole-genome sequencing is at the cutting edge of life sciences in the new millennium. Since the first genome sequencing of the model plant *Arabidopsis thaliana* in 2000, whole genomes of about 100 plant species have been sequenced and genome sequences of several other plants are in the pipeline. Research publications on these genome initiatives are scattered on dedicated web sites and in journals with all too brief descriptions. The individual volumes elucidate the background history of the national and international genome initiatives; public and private partners involved; strategies and genomic resources and tools utilized; enumeration on the sequences and their assembly; repetitive sequences; gene annotation and genome duplication. In addition, synteny with other sequences, comparison of gene families and most importantly potential of the genome sequence information for gene pool characterization and genetic improvement of crop plants are described.

Frank M. You • Bourlaye Fofana
Editors

The Flax Genome

Springer

Editors
Frank M. You
Ottawa Research
and Development Centre
Agriculture and Agri-Food Canada
Ottawa, ON, Canada

Bourlaye Fofana
Charlottetown Research
and Development Centre
Agriculture and Agri-Food Canada
Charlottetown, PE, Canada

ISSN 2199-4781 ISSN 2199-479X (electronic)
Compendium of Plant Genomes
ISBN 978-3-031-16060-8 ISBN 978-3-031-16061-5 (eBook)
https://doi.org/10.1007/978-3-031-16061-5

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

This book series is dedicated to my wife Phullara and our children Sourav and Devleena

Chittaranjan Kole

Preface to the Series

Genome sequencing has emerged as the leading discipline in the plant sciences coinciding with the start of the new century. For much of the twentieth century, plant geneticists were only successful in delineating putative chromosomal location, function, and changes in genes indirectly through the use of a number of "markers" physically linked to them. These included visible or morphological, cytological, protein, and molecular or DNA markers. Among them, the first DNA marker, the RFLPs, introduced a revolutionary change in plant genetics and breeding in the mid-1980s, mainly because of their infinite number and thus potential to cover maximum chromosomal regions, phenotypic neutrality, absence of epistasis, and codominant nature. An array of other hybridization-based markers, PCR-based markers, and markers based on both facilitated construction of genetic linkage maps, mapping of genes controlling simply inherited traits, and even gene clusters (QTLs) controlling polygenic traits in a large number of model and crop plants. During this period, a number of new mapping populations beyond F_2 were utilized and a number of computer programs were developed for map construction, mapping of genes, and mapping of polygenic clusters or QTLs. Molecular markers were also used in the studies of evolution and phylogenetic relationship, genetic diversity, DNA fingerprinting, and map-based cloning. Markers tightly linked to the genes were used in crop improvement by employing the so-called marker-assisted selection. These strategies of molecular genetic mapping and molecular breeding made a spectacular impact during the last one and a half decades of the twentieth century. But still, they remained "indirect" approaches for elucidation and utilization of plant genomes since much of the chromosomes remained unknown and the complete chemical depiction of them was yet to be unraveled.

Physical mapping of genomes was the obvious consequence that facilitated the development of the "genomic resources" including BAC and YAC libraries to develop physical maps in some plant genomes. Subsequently, integrated genetic–physical maps were also developed in many plants. This led to the concept of structural genomics. Later on, emphasis was laid on EST and transcriptome analysis to decipher the function of the active gene sequences leading to another concept defined as functional genomics. The advent of techniques of bacteriophage gene and DNA sequencing in the 1970s was extended to facilitate the sequencing of these genomic resources in the last decade of the twentieth century.

As expected, sequencing of chromosomal regions would have led to too much data to store, characterize, and utilize with the-then available computer software could handle. But the development of information technology made the life of biologists easier by leading to a swift and sweet marriage of biology and informatics, and a new subject was born—bioinformatics.

Thus, the evolution of the concepts, strategies, and tools of sequencing and bioinformatics reinforced the subject of genomics—structural and functional. Today, genome sequencing has traveled much beyond biology and involves biophysics, biochemistry, and bioinformatics!

Thanks to the efforts of both public and private agencies, genome sequencing strategies are evolving very fast, leading to cheaper, quicker, and automated techniques right from clone-by-clone and whole-genome shotgun approaches to a succession of second-generation sequencing methods. The development of software for different generations facilitated this genome sequencing. At the same time, newer concepts and strategies were emerging to handle sequencing of the complex genomes, particularly the polyploids.

It became a reality to chemically—and so directly—define plant genomes, popularly called whole-genome sequencing or simply genome sequencing.

The history of plant genome sequencing will always cite the sequencing of the genome of the model plant Arabidopsis thaliana in 2000 that was followed by sequencing the genome of the crop and model plant rice in 2002. Since then, the number of sequenced genomes of higher plants has been increasing exponentially, mainly due to the development of cheaper and quicker genomic techniques and, most importantly, the development of collaborative platforms such as national and international consortia involving partners from public and/or private agencies.

As I write this preface for the first volume of the new series “Compendium of Plant Genomes”, a net search tells me that complete or nearly complete whole-genome sequencing of 45 crop plants, eight crop and model plants, eight model plants, 15 crop progenitors and relatives, and three basal plants is accomplished, the majority of which are in the public domain. This means that we nowadays know many of our model and crop plants chemically, i.e., directly, and we may depict them and utilize them precisely better than ever. Genome sequencing has covered all groups of crop plants. Hence, information on the precise depiction of plant genomes and the scope of their utilization are growing rapidly every day. However, the information is scattered in research articles and review papers in journals and dedicated web pages of the consortia and databases. There is no compilation of plant genomes and the opportunity of using the information in sequence-assisted breeding or further genomic studies. This is the underlying rationale for starting this book series, with each volume dedicated to a particular plant.

Plant genome science has emerged as an important subject in academia, and the present compendium of plant genomes will be highly useful to both students and teaching faculties. Most importantly, research scientists involved in genomics research will have access to systematic deliberations on the plant genomes of their interest. Elucidation of plant genomes is of interest not only for the geneticists and breeders but also for practitioners of an array of plant science disciplines, such as taxonomy, evolution, cytology, physiology, pathology, entomology, nematology, crop production, biochemistry, and obviously bioinformatics. It must be mentioned that information regarding each plant genome is ever-growing. The contents of the volumes of this compendium are, therefore, focusing on the basic aspects of the genomes and their utility. They include information on the academic and/or economic importance of the plants, a description of their genomes from a molecular genetic and cytogenetic point of view, and the genomic resources developed. Detailed deliberations focus on the background history of the national and international genome initiatives, public and private partners involved, strategies and genomic resources and tools utilized, enumeration of the sequences and their assembly, repetitive sequences, gene annotation, and genome duplication. In addition, synteny with other sequences, comparison of gene families, and, most importantly, the potential of the genome sequence information for gene pool characterization through genotyping by sequencing (GBS) and genetic improvement of crop plants have been described. As expected, there is a lot of variation of these topics in the volumes based on the information available on the crop, model, or reference plants.

I must confess that as the series editor, it has been a daunting task for me to work on such a huge and broad knowledge base that spans so many diverse plant species. However, pioneering scientists with lifetime experience and expertise in the particular crops did excellent jobs editing the respective volumes. I myself have been a small science worker on plant genomes since the mid-1980s and that provided me the opportunity to personally know several stalwarts of plant genomics from all over the globe. Most, if not all, of the volume editors, are my longtime friends and colleagues. It has been highly comfortable and enriching for me to work with them on this book series. To be honest, while working on this series, I have been and will remain a student first, a science worker second, and a series editor last. And I must express my gratitude to the volume editors and the chapter authors for providing me the opportunity to work with them on this compendium.

I also wish to mention here my thanks and gratitude to the Springer staff, particularly Dr. Christina Eckey and Dr. Jutta Lindenborn for the earlier set of volumes and presently Ing. Zuzana Bernhart for all their timely help and support.

I always had to set aside additional hours to edit books beside my professional and personal commitments—hours I could and should have given to my wife, Phullara, and our kids, Sourav and Devleena. I must mention that they not only allowed me the freedom to take away those hours from them but also offered their support in the editing job itself. I am really not sure whether my dedication to this compendium to them will suffice to do justice to their sacrifices for the interest of science and the science community.

New Delhi, India Chittaranjan Kole

Contents

Reference Genome Sequence of Flax

1

Frank M. You, Ismael Moumen, Nadeem Khan, and Sylvie Cloutier

1.1 Introduction

Flax (*Linum usitatissimum* L., 2*n* = 2*x* = 30), also called common flax or linseed, is a self-pollinating crop belonging to the Linaceae family (Singh et al. 2011). Its domestication by humans started around 8000 to 10,000 years ago in the Near-Middle East during the Neolithic period. It then propagated to the Nile Valley, Europe, and finally to the rest of the world (Fu 2011). To meet the growing industry demand, flax is one of the few crops that is cultivated as two main morphotypes: fibre and linseed (Liu et al. 2011). The linseed-type flax is the oilseed type also known as flaxseed. These two morphotypes have different morphology and agronomic traits. The fibre-type accessions are generally taller and have few branches, greater straw strength, and fewer and smaller seeds than the linseed-type accessions, which are comparatively shorter, more branched, have greater seed weight, oil content, and seed yield (Diederichsen and Ulrich 2009; You et al. 2017).

Fibre flax was one of the top three fibre crops used in the textile industry, whereas linseed flax ranked fifth oilseed crop in the world (Ottai et al. 2011). Flax has been widely cultivated in broad geographical regions (Fig. 1.1). In the last 25 years, the main fibre production regions were Europe (73.8%) and Asia (24.5%), while the Americas (41.2%), Asia (35.4%), and Europe (18.5%) were the leading producers of linseed (Fig. 1.2). Fibre flax is mainly grown in Western Europe, Russia, and China, while linseed flax is primarily cultivated in Canada, USA, China, India, Western Europe, Russia, and Kazakhstan (Foulk et al. 2004; You et al. 2016; Soni 2021). The fluctuations of linseed and fibre production by the main world producing regions from 1994–2019 are presented in Fig. 1.3. In recent years, France has led fibre production, while Kazakhstan has become the top flax seed producer.

Recent advancements in flax research have improved our level of knowledge regarding this crop. Specifically, genomic studies have produced large amounts of genomic data, providing the required resources to enhance flax genetic improvement using genomics-based technologies and strategies. One of the major achievements in the past decade was the release of the first flax reference genome sequence of the Canadian cultivar CDC Bethune (Wang et al. 2012), followed by its first version of chromosome-scale pseudomolecules (You et al. 2018). Recently, five more flax genotypes have been sequenced (Dmitriev et al. 2020; Zhang et al. 2020; Sa et al. 2021), providing additional genome sequences for the flax research community. These include

F. M. You (✉) · I. Moumen · N. Khan · S. Cloutier
Ottawa Research and Development Centre,
Agriculture and Agri-Food Canada, 960 Carling
Avenue, Ottawa, ON K1A 0C6, Canada
e-mail: frank.you@agr.gc.ca

F. M. You and B. Fofana (eds.), *The Flax Genome*, Compendium of Plant Genomes,
https://doi.org/10.1007/978-3-031-16061-5_1

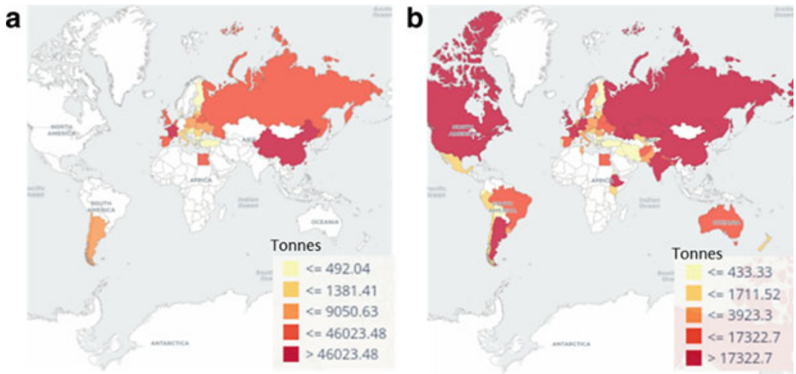

Fig. 1.1 Average production of flax fibre and tow (**a**) and linseed (**b**) in the world from 1994–2020. *Source* FAOSTAT

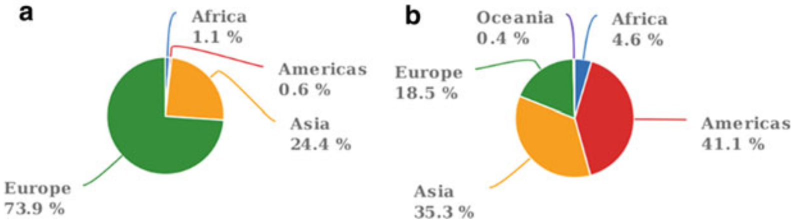

Fig. 1.2 Average production of flax fibre and tow (**a**) and linseed (**b**) by region from 1994–2020. *Source* FAOSTAT

the Chinese linseed cultivar Longya-10, the Chinese fibre cultivars Heiya-14 and Yiya-5, and the Russian fibre cultivar Atlant, as well as one accession of pale flax (*L. bienne*), which is the closest wild relative of cultivated flax. This chapter briefly reviews the major advances in flax genome sequencing, assembly, annotation, and comparative analysis between these genome sequences.

1.2 Flax Genome Assemblies

A reference genome developed from a genotype of a species represents a standard of genome sequence that delivers both the nucleotide sequence of its chromosomes and its associated structural information for genomics studies, providing a basis for comparison with other genotypes within or between species. Since the first plant genome assembly, i.e., *Arabidopsis thaliana,* was published in 2000 (Arabidopsis Genome Initiative 2000) and the first draft human reference genome released in 2001 (Lander et al. 2001), hundreds of plant species have been sequenced (Marks et al. 2021). The rapid evolution of sequencing technologies, the development of new genome scaffolding techniques, and the improvement of genome assembly algorithms and software tools (Ghurye and Pop 2019; Wee et al. 2019) have led to comprehensive and high-quality genome assemblies covering long repeat sequence gaps (Zimin et al. 2017a; Nurk et al. 2022). The third generation of sequencing technologies includes PacBio single-molecule high fidelity (HiFi) sequencing (Hon et al. 2020) and ultra-long-read Oxford Nanopore technology (ONT) (Jain et al. 2016; Bocklandt et al. 2019). The new genome scaffolding techniques have optical mapping such as BioNano genome mapping (Lam et al. 2012) and Hi-C sequencing (Belton et al. 2012; Burton et al. 2013).

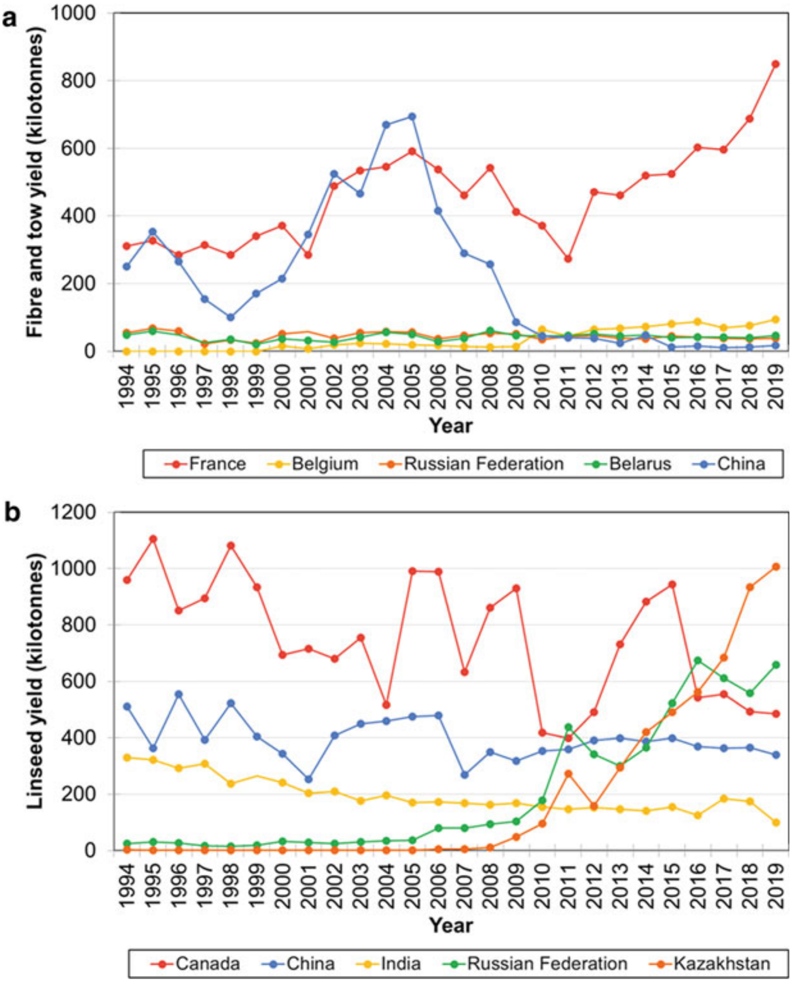

Fig. 1.3 Fibre and tow (**a**) and linseed (**b**) production by the main world producing countries from 1994–2019. *Source* FAOSTAT

1.2.1 The First Version of the Flax Genome Assembly for CDC Bethune

CDC Bethune (Rowland et al. 2002), the Canadian high-yielding and medium-late-maturing linseed flax cultivar, developed by the Crop Development Centre at the University of Saskatchewan, was selected for the development of the first flax reference genome proposed in the Genome Canada research project entitled "Total Utilization Flax Genomics (TUFGEN)" which ran between 2009–2014 (https://genomecanada.ca/project/total-utilization-flax-genomics/). The genome size of CDC Bethune was estimated at 368 Mb based on the bacterial artificial chromosome (BAC)-based physical map (Ragupathy et al. 2011) and at nearly 373 Mb based on flow cytometry (Wang et al. 2012).

This pioneering CDC Bethune sequencing project used a whole-genome sequencing (WGS) strategy based on the Illumina sequencing platform. A total of 25.88 Gb Illumina short reads, corresponding to 94X genome coverage from seven paired-end and mate-pair libraries, with an insert size of 300 bp to 10 Kb, were generated. De novo assembly was performed using SOAPdenovo (Li et al. 2009). This led to an assembly consisting of 116,602 contigs (302 Mb) or 88,384 supercontigs (scaffolds) (318 Mb), covering approximately 81% of the flax genome, estimated at 370 Mb (Wang et al.

Table 1.1 Statistics of the first version (Wang et al. 2012) of the flax genome assembly and its annotation as deposited and summarized in Phytozome

Item	Statistics
Annotation version	v1.0
Total scaffold length (bp)	318,250,901
Number of scaffolds	88,420
Scaffold L50	132
Scaffold N50 (bp)	693,492
Total contig length (bp)	302,186,967
Number of contigs (bp)	116,824
Contig L50	4427
Contig N50 (bp)	20,125
Number of protein-coding transcripts	43,484
Number of protein-coding genes	43,471
Percentage of eukaryote BUSCO genes	97.7
Percentage of embryophyte BUSCO genes	92.1

L50: the minimum number of scaffolds or contigs containing half of the assembly; N50: the length of the shortest scaffold or contig from the L50 set
Source https://phytozome-next.jgi.doe.gov/

2012). This assembly was the first flax reference genome sequence and opened a new era to flax genomic studies despite its large number of short scaffolds, with a scaffold N50 of only ~ 693 Kb. The sequences and genome annotation information are now available to download from Phytozome (Table 1.1).

1.2.2 The Second Version of the Flax Genome Assembly for CDC Bethune: Chromosome-Level Pseudomolecules

A pseudomolecule refers to the DNA sequence assembly representing a biological chromosome or full genome. To achieve that the assembled scaffolds or contigs are sorted and assigned to individual chromosomes with the aid of consensus genetic maps (Cloutier et al. 2012), BAC-based physical maps (Luo et al. 2010), and more recent scaffolding technologies, including optical mapping, such as BioNano genome mapping (Hastie et al. 2013; Stankova et al. 2016) and Hi-C sequencing (Belton et al. 2012).

The first chromosome-level pseudomolecules of CDC Bethune (v2.0) (You et al. 2018) were constructed by integrating information from the BAC-based physical map (Ragupathy et al. 2011), the simple sequence repeat (SSR) marker-based consensus genetic map (Cloutier et al. 2012a), and the BioNano genome optical maps (You et al. 2018). The long scaffolds in the assembly v1.0 (Wang et al. 2012) were sorted and assigned to the 15 chromosomes of flax. This new 316.2 Mb assembly represented the 15 chromosomes of flax with sizes ranging from 15.6 Mb for chromosome (Chr) 15 to 29.4 Mb for Chr 1 (You et al. 2018) (Table 1.2, Fig. 1.4). The pseudomolecules contain ~ 47 Mb of gaps within original scaffolds generated by de novo assembly and scaffolding with mate-pair sequences and between sorted scaffolds estimated with BioNano genome maps. The 15 chromosome sequences were deposited in the NCBI database (CP027619–CP027633). This chromosome-scale reference sequence represents a significant improvement over the first version of the draft flax genome reference sequence (Wang et al. 2012), benefiting genome-wide SNP discovery, QTL identification, genome-wide association studies, and comparative genome analyses.

Table 1.2 Chromosome-scale pseudomolecules of the 15 chromosomes of the linseed cultivars CDC Bethune (You et al. 2018) and Longya-10 (Zhang et al. 2020)) and the fibre cultivar Yiya-5 (Sa et al. 2021)

Chr	CDC Bethune			Longya-10			Yiya-5[a]	
	NCBI accessions	Size (Mb)	Gap (Mb)	NCBI accessions	Size (Mb)	Gap (Mb)	Size (Mb)	Gap (Kb)
1	CP027619.1	29.43	5.46	CM036262.1	22.66	0.36	31.79	0.82
2	CP027626.1	25.73	5.09	CM036263.1	22.13	0.26	29.85	0.60
3	CP027627.1	26.64	3.81	CM036264.1	21.91	0.36	26.36	0.40
4	CP027628.1	19.93	2.66	CM036265.1	21.42	0.31	24.86	0.40
5	CP027629.1	17.70	1.95	CM036266.1	20.60	0.35	24.52	0.60
6	CP027630.1	18.08	1.99	CM036267.1	19.56	0.33	28.49	0.70
7	CP027631.1	18.30	2.63	CM036268.1	18.79	0.38	17.72	0.60
8	CP027632.1	23.79	3.77	CM036269.1	18.16	0.38	31.05	0.40
9	CP027633.1	22.09	3.85	CM036270.1	17.74	0.21	32.04	0.90
10	CP027620.1	18.20	1.79	CM036271.1	17.30	0.29	33.28	0.50
11	CP027621.1	19.89	2.42	CM036272.1	16.85	0.36	31.41	0.20
12	CP027622.1	20.89	3.66	CM036273.1	17.15	0.24	19.88	0.40
13	CP027623.1	20.48	2.14	CM036274.1	16.35	0.24	21.42	0.20
14	CP027624.1	19.39	2.86	CM036275.1	16.86	0.26	39.91	0.40
15	CP027625.1	15.64	2.50	CM036276.1	14.75	0.25	30.52	0.40
Total		316.17	46.58		282.23	4.57	423.10	7.52

[a] NCBI accession numbers are not available for Yiya-5 because its assembly and annotation files are deposited in Zenodo (https://doi.org/10.5281/zenodo.4872893)

1.2.3 Recent Assemblies Expanding the Representation to Both Morphotypes and to the Closest Wild Relative of Cultivated Flax

In recent years, five other flax genotypes, including one linseed and three fibre cultivars, as well as one wild relative of flax (pale flax, *L. bienne*) have also been sequenced and assembled. Zhang et al. (2020) performed a WGS of three flax genotypes: the linseed-type cultivar Longya-10, the fibre-type cultivar Heiya-14, and a pale flax accession (Fig. 1.5). Illumina paired-end reads of 68.2, 73.5, and 49.1 high-quality Gbp corresponding to 133, 142, and 93X genome coverage were generated for the three genotypes, respectively. De novo assemblies were performed using ALLPATH-LG (Gnerre et al. 2011), scaffolding with mate-pair information was conducted using SSPACE (Boetzer et al. 2011), and gap-filling was performed using GapCloser from the SOAPdenovo2 package (Luo et al. 2012). As a result, assemblies of 306.0, 303.7, and 293.5 Mb genome sequences with the scaffold N50 of 1,235 Kb, 700 Kb, and 384 Kb were obtained for Longya-10, Heiya-14, and pale flax, respectively. Gaps in the assemblies were estimated at 5.8 Mb for Longya-10, 2.8 Mb for Heiya-14, and 5.6 Mb for the pale flax genome. Hi-C data and a genetic map were used to enhance the Longya-10 genome assembly, leading to 434 scaffolds totaling 295.7 Mb in length for the chromosomal-level assembly. The longest scaffolds corresponding to 15 chromosomes have a total length of 282.23 Mb (Table 1.2).

Around the same time, Dmitriev et al. (2020) released the genome sequence of the Russian fibre cultivar Atlant using both ONT and Illumina platforms. A total of 8.4 Gb ONT long reads with an N50 of 12 Kb corresponding to

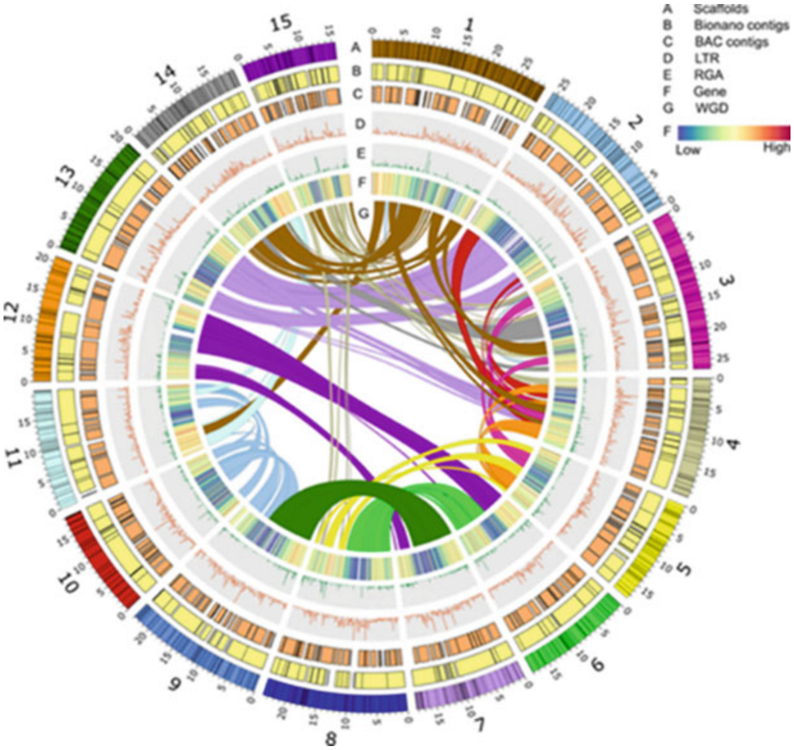

Fig. 1.4 Circos map illustrating the 15 chromosome of CDC Bethune with Track A, scaffolds integrated in the pseudomolecules; Track B, BioNano contigs mapped to scaffolds; Track C, bacterial artificial chromosome (BAC)-based contigs mapped to scaffolds and BioNano contigs; Track D, frequency distribution of long terminal repeats (LTRs) on chromosomes with bin sizes of 0.1 Mb; Track E, frequency distribution of resistance gene analogues (RGAs) on chromosomes with bin sizes of 0.1 Mb; Track F, heat map of genes with bin sizes of 0.1 Mb; and Track G, the central region showing whole-genome duplication (WGD). *Source* You et al. (2018)

23X flax genome coverage and 22.6 million 250 bp paired-end reads corresponding to 30X genome coverage were generated. The ONT reads were assembled separately by several assemblers, including Canu 2.0 (Koren et al. 2017), Flye 2.7 (Kolmogorov et al. 2019), Shasta 0.5.0 (Shafin et al. 2020), and wtdbg2 2.5 (Ruan and Li 2020). Contigs were polished using Illumina reads by Racon (Vaser et al. 2017), Medaka (https://github.com/nanoporetech/medaka), and POLCA in the assembler MaSuRCA (Zimin et al. 2017b) to improve the accuracy. A comparison of the assemblies generated by the different assemblers and polishing tools indicated that the most complete and accurate assembly was obtained using Canu combined with the polishing tools Racon + Medaka + POLCA. This assembly was 361.7 Mb in length, but the N50 was only 350 Kb (Table 1.3).

The Chinese fibre-type genotype Yiya-5, a high fibre-yielding cultivar bred by the Xinjiang Yili Institute of Agricultural Sciences, China, has also been sequenced using the PacBio HiFi sequencing technology. A total of 21.80 Gb of circular consensus sequence (CCS) reads were generated with an N50 of 12,191 bp. The reads were assembled using Hifiasm (v0.13-r308) (Chen et al. 2020), generating a draft assembly of 1,632 contigs totalling 537.51 Mb. Removal of the redundant haplotigs yielded a refined assembly (v1.0) of 336 contigs with an N50 of 9.61 Mb totalling 454.95 Mb. Using 58.61 Gb high-quality Hi-C data, the contigs in the assembly v1.0 were further scaffolded, resulting

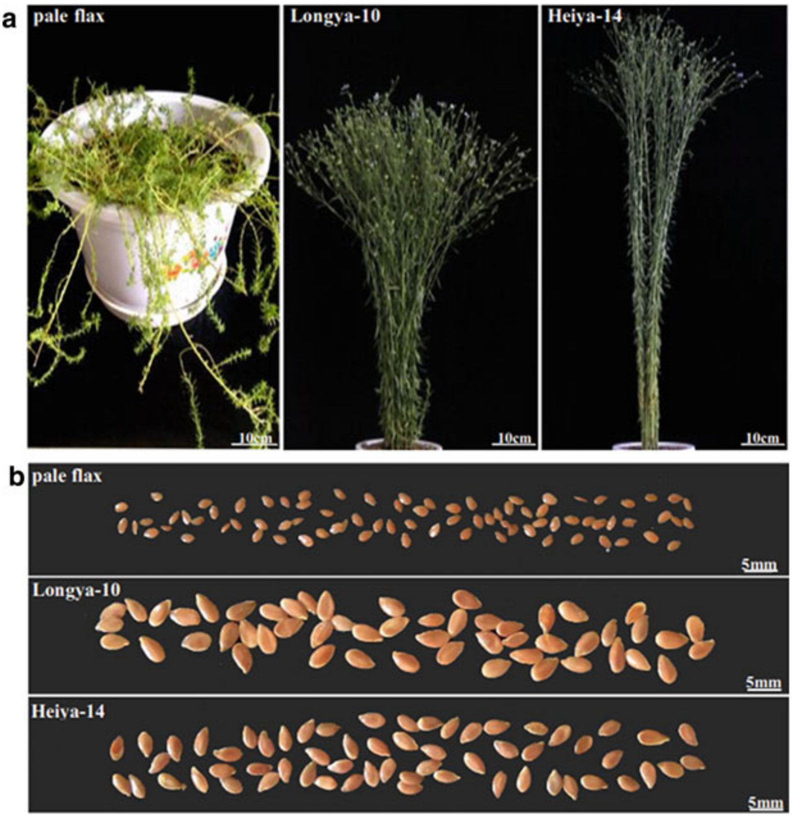

Fig. 1.5 Plant (**a**) and seed (**b**) morphology of pale flax, Longya-10, and Heiya-14. *Source* Zhang et al. (2020)

in 15 chromosome-length scaffolds totalling 423 Mb (v2.0) (Table 1.2) and covering 93.0% of the sequences in the assembly v1.0.

For the chromosome-scale pseudomolecules, Longya-10 has a smaller genome assembly size with gaps (282.23 Mb) than CDC Bethune (316.17 Mb), but similar genome assembly size without gaps (277.7 Mb) compared to CDC Bethune (269.6 Mb) (Table 1.2). The similarity in pseudomolecule sizes between CDC Bethune and Longya-10 could be because a similar Illumina-based sequencing technology strategy was used in both projects.

At 423 Mb, Yiya-5 v2.0 pseudomolecules yielded a larger genome assembly size than CDC Bethune and Longya-10, with chromosomes ranging from 17.72 Mb (Chr 7) to 39.91 Mb (Chr 14) and only 7.5 Kb of gaps. The Yiya-5 v2.0 genome assembly is highly collinear with the CDC Bethune v2.0 genome assembly except for the central regions of chromosomes that likely contain the centromeric repeat sequences resolved in the Yiya-5 assembly but missing in the CDC Bethune v2.0 assembly (Sa et al. 2021) (Fig. 1.6), suggesting that PacBio HiFi sequencing scaffolded with Hi-C data has significantly improved the genome assembly by providing a more complete assembly of the repeat sequences.

1.2.4 Quality Examination of Flax Genome Assemblies

The Benchmarking Universal Single-Copy Orthologue (BUSCO) is a widely adopted assessment tool for genome assembly quality that uses a predefined and expected set of single-copy marker genes as a proxy for genome-wide completeness (Manni et al. 2021). To compare the assembly quality of all seven flax assemblies, we performed BUSCO analyses of the released

Table 1.3 Description of flax genome assemblies released between 2012 and 2021

Genotype	Morphotype	Sequence technology	Assembler	Scaffolds	Total size (Mbp)	N50 (Mbp)	Longest scaffold (Mbp)	References
CDC Bethune	Linseed	Illumina	SOAPdenovo	88,384 (≥ 100 bp)	318.25	0.69	3.09	Wang et al. (2012)
Longya-10	Linseed	Illumina, Hi-C	ALLPATH-LG SSPACE SOAPdenovo2	1865 (≥ 911 bp)	306.00	1.24	4.61	Zhang et al. (2020)
Heiya-14	Fibre			2748 (≥ 899 bp)	303.68	0.70	3.04	
Pale flax				2609 (≥ 883 bp)	293.58	0.38	3.51	
Atlant	Fibre	Oxford Nanopore, Illumina	Canu 2.0 Flye 2.7 Shasta 0.5.0 wtdbg2 2.5	2458 (≥ 1012 bp)	361.70	0.35	5.00	Dmitriev et al. (2020)
Yiya-5	Fibre	PacBio HiFi, Hi-C	Hifiasm v0.13-r308 minimap2 2.17-r941	262 (≥ 9885 bp)	454.96	9.61	39.91	Sa et al. (2021)

genome assemblies (Table 1.3) using the same BUSCO eudicots_odb10 dataset (Fig. 1.7). Of the protein-coding genes, a high percentage of complete single-copy and duplicated genes (~95%) with approximately 4% missing genes was observed for the Atlant, Heiya-14, and Longya-10 assemblies, indicative of high assembly quality. Approximately 6–8% of the genes were missing in the assembly of Yiya-5 and pale flax, and ~ 9.5% were absent from the CDC Bethune assembly.

1.3 Repeat Sequence

Genome annotation of assembled genome sequences aims to assign biological information to sequences. Genome annotation primarily consists of two steps: identification of non-protein-coding elements, such as repetitive DNA sequences, including the major transposable element (TE) classes, and gene annotation that involves the prediction of protein-coding genes and their functional annotation.

Based on the mechanism of transposition, TEs can be grouped into two major classes: Class I retrotransposons, transposing via a copy-and-paste mechanism involving RNA intermediates, and Class II DNA transposons, transposing through a simple cut-and-paste mechanism without an RNA intermediate (Wicker et al. 2007). TEs are extremely diverse, and the thousands of distinct TE families in plants (Feschotte et al. 2002; Morgante 2006) account for a large portion of the genome in many plant species. In eukaryotic genomes, TEs contribute substantially to the genome size and they play important roles in structural and functional genomics (Tollis and Boissinot 2012). These sequences are known to cause significant changes in genomes, reflecting evolutionary differences across species (Mehrotra and Goyal 2014). In the genus *Linum*, there are more than 200 diploid species, characterized by karyotype variabilities observed as size, number, and structure of chromosomes (Goldblatt 2007; Rice et al. 2014). Such variability is mostly determined by the amount and composition of repeated sequences. Using high-throughput

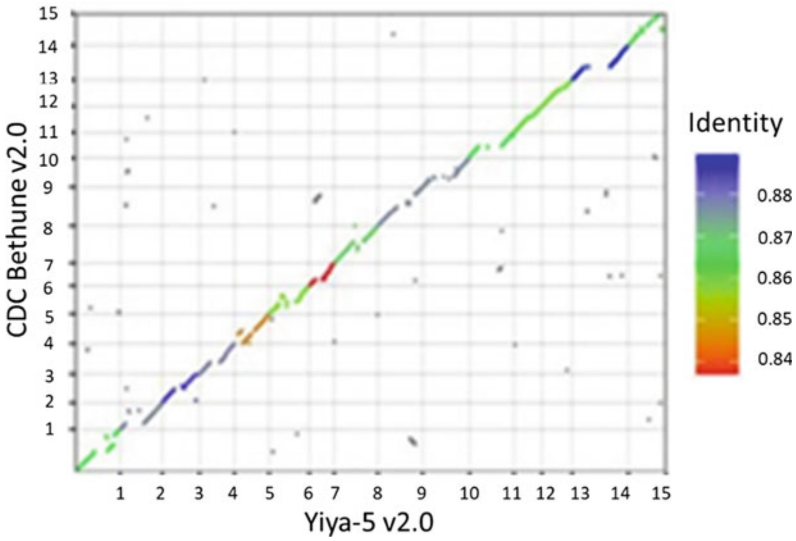

Fig. 1.6 Genomic synteny similarity between Yiya-5 v2.0 and CDC Bethune v2.0. *Source* modified from Sa et al. (2021)

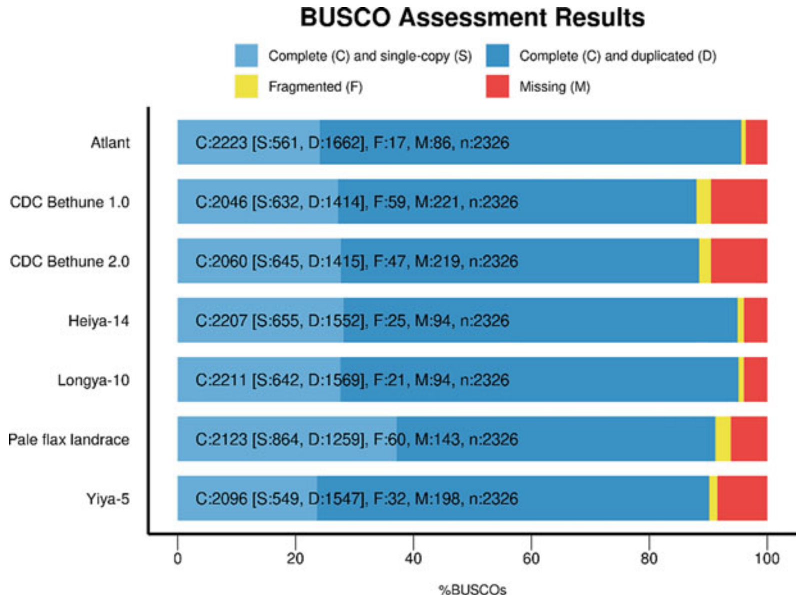

Fig. 1.7 BUSCO assessment results of seven flax assemblies including five cultivated flax cultivars (two linseed and three fibre) and one pale flax accession

genome sequencing, Bolsheva et al. (2019) performed a comparative study on repeat sequences in 12 blue-flowered flax species, including cultivated and wild flax. These *Linum* species were found to largely differ in their satellite DNA families and relative content in genomes. Their evolution was accompanied by waves of amplification of satellite DNAs and long terminal repeat (LTR) retrotransposons (Bolsheva et al. 2019).

The TE content of the flax genome assemblies ranged from 23 to 55% (Wang et al. 2012; Zhang et al. 2020; Sa et al. 2021) (Table 1.4). LTRs are the most prominent TE type, accounting for 36–80% depending on the genotypes (Table 1.4). For example, LTRs accounted for 75.35%, 36.45%, 36.34%, 80.49%, and 36.47% of all TEs identified in the assemblies of flax genotypes CDC Bethune, Longya-10, Heiya-14, Yiya-5,

Table 1.4 Transposable elements (TEs) in the genome assemblies of five flax genotypes showing the percentages of each TE types per genome and as a proportion of all identified TEs (in parentheses)

Type	Sequence percentage (%) of genome and all TEs (in parentheses)					
	CDC Bethune v1.0[a]	CDC Bethune v1.0[b]	Longya-10	Heiya-14	Pale flax	Yiya-5 v2.0
Class I: Retrotransposon						
LTR/*Copia*	9.79 (40.09)	9.30 (40.33)	7.93 (20.50)	7.66 (20.76)	7.55 (20.56)	5.67 (10.24)
LTR/*Gypsy*	8.31 (34.03)	7.89 (34.22)	6.12 (15.82)	5.53 (14.99)	5.79 (15.77)	14.70 (26.55)
LTR/Unknown	0.30 (1.23)	0.28 (1.21)	0.05 (0.13)	0.22 (0.60)	0.05 (0.14)	3.82 (6.90)
Other	2.22 (9.09)	2.11 (9.15)	9.68 (25.03)	8.88 (24.07)	10.84 (29.52)	2.11 (7.55)
Class I total	20.62 (84.44)	19.58 (84.91)	23.78 (64.48)	22.29 (64.01)	24.23 (65.99)	26.29 (47.49)
Class II: Transposon						
Class II total	3.80 (15.56)	3.62 (15.70)	4.84 (12.51)	4.79 (12.98)	4.13 (11.25)	9.98 (18.03)
Other repeats	NA	NA	10.06 (26.01)	9.82 (26.61)	8.36 (22.77)	19.09 (34.47)
Overall total	24.42 (100)	23.06 (100)	38.68 (100)	36.90 (100)	36.72 (100)	55.36 (100)

NA not available
Sources Wang et al. (2012) for CDC Bethune[a]; Gonzalez and Deyholos (2012) for CDC Bethune[b]; Zhang et al. (2020) for Longya-10, Heiya-14, and Pale flax; Sa et al. (2021) for Yiya-5

and a pale flax accession, respectively (Table 1.4).

Besides the inherent genome features of flax genotypes, differences in TE proportions of the assemblies may be a consequence of the sequencing technologies used. The third-generation sequencing platforms improved the completeness of genome assemblies because long reads can span entire repeat sequences. For example, the fibre cultivar Yiya-5 was sequenced using PacBio HiFi, generating an assembly of 454 Mb in size, of which 55% were repeat sequences. Both the size of the genome and the proportion of TEs in the Yiya-5 assembly are significantly greater than those in the other four flax genotypes (Tables 1.2 and 1.4). Another factor affecting TE identification could be the software tools and repeat libraries which differed across the flax assemblies (Agrios 2005; Gonzalez and Deyholos 2012; Wang et al. 2012; You et al. 2018; Zhang et al. 2020; Sa et al. 2021). Therefore, a standardized TE identification procedure with a common set of software tools is required to provide a better comparative TE analysis of flax genomes (You et al. 2015).

1.4 Gene Annotation

The gene annotation of a draft assembly includes gene prediction and their functional annotation. There are three major strategies for protein-coding gene prediction of assembled genome sequences: ab initio-based, evidence-based, and/or combination thereof. Ab initio gene prediction methods use statistical models to identify intrinsic gene content and signals and predict potential protein-coding genes strictly based on the genome sequence. As such, ab initio gene prediction can identify putative genes even if they share no similarity to known gene sequences or protein domains. On the other hand, some may also be erroneous calls. The evidence-based methods predict genes based on evidence for their transcription obtained from cDNAs, RNA-seq data, or other gene expression data such as PacBio IsoSeq for example. The accuracy of gene prediction relies on their expression in the sample(s) sequenced and the integrity of the sample(s). Some real genes may potentially be missed because they are not represented in the sample(s) or because their expression and/or the quality of the sample(s) were too low. These may be captured based on statistical models; hence, a strategy that combines evidence-based and ab initio approaches by mapping proteins, expressed sequence tags (ESTs), and RNA-seq data to the target genome to validate predicted gene structures outperforms the individual strategies through complementarity (Holt and Yandell 2011; Hoff et al. 2016; Bruna et al. 2021).

In the last decade, some combined approach-based software tools have been implemented and continuously improved. These tools integrate ab initio gene prediction, evidence from transcripts, and homology-based gene prediction, which relies on gene models of related and well-annotated species into an automatic pipeline, increased accuracy of protein-coding gene prediction, and they provide an efficient way to solve some computational complexities (Holt and Yandell 2011; Hoff et al. 2016; Bruna et al. 2021). Maker2 (Holt and Yandell 2011) is a pipeline that integrates ab initio gene predictors, including SNAP (Korf 2004), Augustus (Stanke et al. 2006), GeneMark (Lomsadze et al. 2014), and RNA-seq data, whereas Braker2 is a more recent pipeline for unsupervised RNA-seq-based genome annotation that combines the advantages of GeneMark-EP+ (Lomsadze et al. 2014) and Augustus (Stanke et al. 2006).

Ab initio gene prediction for the assemblies of the six flax genotypes (CDC Bethune, Longya-10, Heiya-14, Atlant, Yiya-5, and pale flax accession) has been performed (Wang et al. 2012; Dmitriev et al. 2020; Sa et al. 2021; Zhang et al. 2021). In these annotations, Augustus was used for ab initio gene prediction followed by validation using ESTs (CDC Bethune) or RNA-seq data sequenced from different tissues, including stem and boll tissues (Longya-10, Heiya-14, and pale flax), and leaf, stem, flower, root, and bolls (Yiya-5). A total of 43,484 genes with an average gene size of 2307 bp were predicted from the first version of the CDC Bethune genome, with 89.5% of them aligned to one or more proteins in the NCBI nr protein database (Wang et al. 2012). A similar number of protein-coding genes, approximately 43,500, with similar gene lengths (2.3–2.5 Kb) were predicted for Longya-10 (linseed), Heiya-14 (fibre), and the pale flax accession. More protein-coding genes were predicted for the fibre genotypes Yiya-5 (49,616) and Atlant (77,522), which were sequenced using the third-generation sequencing technologies PacBio HiFi and ONT, respectively, and produced larger genome assemblies (454.96 Mb for Yiya-5 and 361.80 Mb for Atlant) than those obtained using Illumina short reads (Table 1.5). Of note, the Yiya-5 predicted genes are larger (3.7 Kb) than that of the other four flax genotypes (Table 1.5).

Functional annotation is the process of assigning biological information to the predicted genes. The homology-based sequence alignment tool BLAST and some other bioinformatics tools such as InterProScan program (Jones et al. 2014), eggNOG-mapper (Huerta-Cepas et al. 2017), and DIAMOND (Buchfink et al. 2015) are commonly used for functional annotation. They are based on dedicated databases, including eggNOG 5.0 (Huerta-Cepas et al. 2019), GO (Harris et al. 2004), KEGG (Kanehisa et al. 2002), Pfam (Mistry et al. 2021), Swiss-Prot (Bairoch and Apweiler 2000), and NCBI non-redundant protein database nr.

Table 1.6 lists the number of predicted genes that have been annotated using some common gene annotation databases (Wang et al. 2012; Dmitriev et al. 2020; Sa et al. 2021; Zhang et al. 2021). In the CDC Bethune assembly v1.0, 39,288 of 43,484 predicted genes aligned to one or more *Arabidopsis* proteins, resulting in 35,727 predicted genes with assigned functions from the protein annotation of *Arabidopsis* genes (Wang et al. 2012). In the assembly v2.0 of Yiya-5, of the 49,616 protein-coding genes, 34,938

Table 1.5 Genes predicted from the genome assemblies of six flax genotypes

Cultivar	Assembly	Gene prediction tools used	No. of protein-coding genes	Average gene length (bp)	Average exon length (bp)	References
CDC Bethune	v1.0	Augustus v. 2.5.5, GLIMMERHMM v. 3.0.1	43,484	2,307	237	Wang et al. (2012)
Longya-10	v1.0	Genscan v1.0, Augustus v2.5.5 GlimmerHMM v3.0.1, GeneID v1.3, SNAP	43,668	2,505	238	Zhang et al. (2020)
Heiya-14	v1.0		43,826	2,501	236	
Pale flax	v1.0		43,424	2,344	231	
Atlant	v1.0	PASA 2.4.1, Augustus 3.3.3, GlimmerHMM 3.0.4, SNAP v. 2006-07-28, GeneMark 4.61, CodingQuarry 2.0, EvidenceModeller 1.1.1	77,522	NA	NA	Dmitriev et al. (2020)
Yiya-5	v2.0	Braker2 v2.1.6 with HISAT2 v2.1.0, Augustus v3.4.0, GUSHR v1.0.0	49,616	3,702	215	Sa et al. (2021)

NA not available

Table 1.6 Number of annotated protein-coding genes for six flax genotypes

Database	CDC Bethune v1.0	Longya-10	Heiya-14	Pale flax	Atlant	Yiya-5 v2.0
KOG	17,319	25,055	15,775	21,540	NA	NA
GO	23,571	24,919	25,798	22,268	NA	22,600
KEGG	5999	9450	9677	13,978	NA	21,611
Pfam	32,166	NA	NA	NA	18,946	34,938
eggNOG	NA	NA	NA	NA	19,741	42,697
UniProt	NA	NA	NA	NA	3725	NA
Swiss-Prot	NA	33,005	34,147	27,472	NA	34,654

NA not available
Source Wang et al. (2012) for CDC Bethune v1.0; Zhang et al. (2020) for Longya-10, Heiya-14, and Pale flax; Dmitriev et al. (2020) for Atlant; and Sa et al. (2021) for Yiya-5

(70.42%), 42,697 (86.05%), 22,600 (45.55%), 21,611 (43.56%), 34, 654 (69.84%), and 41,847 (84.34%) genes had significant hits with Pfam, eggNOG, GO, KEGG, Swiss-Prot, and nr databases, respectively. Overall, 43,364 (87.40%) genes were successfully annotated with at least one database (Sa et al. 2021). In the assembly of Atlant, 18,946 predicted genes were successfully annotated using the Pfam database, 19,741 using eggNOG, and 3,725 using UniProt (Dmitriev et al. 2020). In the assemblies of Longya-10, Heiya-14, and pale flax, even though similar numbers of the predicted genes were obtained from three genotypes (Table 1.5), large differences in hits to several annotation databases were observed (Table 1.6) (Zhang et al. 2020). Overall, the annotation results of the four genome sequencing studies differed substantially (Wang et al. 2012; Dmitriev et al. 2020; Sa et al. 2021; Zhang et al. 2021).

1.5 Non-coding RNAs

Non-coding RNAs (ncRNAs) comprise the majority of cellular RNAs and are a part of the transcriptome without having protein-coding roles. They play important roles in diverse biological processes, including translation (tRNA and rRNA), synthesis of the translational apparatus (snRNA), and gene regulation (miRNA). The genome sequence of four flax genotypes has been annotated for functional ncRNAs (Wang et al. 2012; Zhang et al. 2021). More than 700 copies of both tRNAs and rRNAs and 115–297 copies of putative miRNA precursor loci were identified from the four assemblies. More ncRNA copies were detected in CDC Bethune than in the flax cultivars Longya-10 and Heiya-14 and the pale flax accession (Table 1.7).

1.6 Chloroplast Genome

The chloroplast (cp) is the photosynthetic double membrane-bound organelle that converts light energy to carbohydrates in plants and algae. The size of the chloroplast genome (plastome) of autotrophic angiosperms is generally conserved (Guo et al. 2021). The average cp genome size of land plants is 151 Kb, with most ranging from 130 to 170 Kb (Guo et al. 2021). The majority of cp genomes are circular DNA molecules with two copies of inverted repeats (IRs) of approximately 20–28 Kb, one large single-copy region (LSC) of 80–90 Kb, and one small single-copy region (SCR) of 16–27 Kb (Jansen et al. 2005; Li and Zheng 2018). The large IR might help to protect the cp genome from major structural changes (Wu et al. 2011; Wu and Chaw 2014). The cp genomes were found to have highly conserved gene content and order and have been widely used for plant species identification, taxonomy, and phylogenetic analysis (Raubeson and Jansen 2005). Thus, cp genome sequencing has revealed significant sequence and structural variation within and between plant species, such as SNPs, indels, small inversions, and inverted repeats, which have proven to be valuable resources in the study of plant genome evolution (Borsch and Quandt 2009). This knowledge has also been useful in improving our understanding of the climatic adaptation of economically important crops, as well as the identification and study of important traits in closely related species (Wambugu et al. 2015; Brozynska et al. 2016). However, some angiosperms, such as the Campanulaceae (Knox 2014; Hong et al. 2017), Geraniaceae (Guisinger et al. 2008; Marcussen and Meseguer 2017), and some legume family species (Schwarz et al. 2015) are prone to large-scale rearrangements.

The rapid advances in chloroplast genetics and genomics have been greatly facilitated by the advent of high-throughput sequencing technologies. Since the first cp genome from tobacco (*Nicotiana tabacum*) was sequenced 35 years ago (Shinozaki et al. 1986), the National Center for Biotechnology Information (NCBI) organelle genome database has grown substantially and

Table 1.7 Copy number of non-coding RNAs in four flax genotypes

Accession	CDC Bethune v1.0	Longya-10	Heiya-14	Pale flax
rRNA	1100	955	722	866
tRNA	1100	965	986	969
miRNA	297	126	115	128
snRNA	462	207	202	184
snoRNA	NA	555	543	534
Total	2959	2808	2568	2681

rRNA ribosomal ribonucleic acid; *tRNA* transfer RNA; *miRNA* microRNA; *snRNA* small nuclear RNA; *snoRNA* small nucleolar RNA; *NA* not available

Source Wang et al. (2012) for CDC Bethune v1.0; and Zhang et al. (2020) for Longya-10, Heiya-14 and Pale flax

Table 1.8 Summary of chloroplast (cp) genomes in some plant species

Species	CPG (bp)	IRs (bp)	LSC (bp)	SSC (bp)	References
Flax	156,721	31,990	81,767	10,974	de Santana Lopes et al. (2018)
Tomato	155,443–155,561	25,612–25,639	85,857–85,911	18,362–18,387	Wu (2016)
Soybean	52,218	25,574	83,175	17,895	Saski et al. (2005)
Wheat	135,766	21,487–21,485	80,003	12,791	Fu (2021)
Aegilops	135,502	21,483–21,481	79,766	12,772	Fu (2021)
Grape	160,928	26,358	89,147	19,065	Jansen et al. (2006)
Black mustard	153,633	26,193	83,552	17,695	Seol et al. (2017)
Cabbage	153,366	26,197	83,137	17,835	Seol et al. (2017)

CPG chloroplast genome; *IRs* inverted repeats; *LSC* large single copy; *SSC* small single copy

now has more than 4200 entries of cp genome sequences of land plants (Guo et al. 2021). Several plant cp genomes are listed in Table 1.8. The cp genome of flax was completely sequenced by de Santana Lopes et al. (2018). Flax has a circular cp genome of 156,721 bp with IRs of 31,990 bp separating the LSC of 81,767 bp and the SSC of 10,974 bp and containing 109 unique genes and two pseudogenes (*rpl23* and *ndhF*) (de Santana Lopes et al. 2018). In addition, 176 SSRs, 20 tandem repeats, and 39 dispersed repeats were also identified (de Santana Lopes et al. 2018).

1.7 Concluding Remarks

The availability of high-quality reference genomes has a significant impact on the understanding of genome structure and function, species evolution, as well as applications in genetics and breeding. In the last decade, the first reference genome (v1.0) and its subsequent chromosome-scale pseudomolecule iteration (v2.0) of the flax cultivar CDC Bethune have been widely used as a reference in genomic studies and breeding applications. An additional five flax genotypes, including linseed, fibre flax, and the closely related wild flax (pale flax), have also been sequenced using different sequencing platforms. Their genome assemblies and annotations constitute precious genomic resources for genome-wide comparative analyses. It is noteworthy that all these assemblies and their annotations have revealed large variations in genome assembly size (304–455 Mb), predicted protein-coding genes (43,424–77,522 with an average gene size of 2.3–3.7 Kb), and repeat content (23–55% of the genome). However, to date, insufficient evidence exists to conclude that these variations are due to inherent genome features of the flax genotypes because they were generated by different laboratories using different sequencing technologies, computational tools, and combinations thereof. Therefore, additional genome size information of the sequenced genotypes (such as estimate by flow cytometry) and genome annotation using consistent software tools and criteria are warranted to improve the comparability across genotypes. However, such comparisons will remain hindered by the limits imposed by the sequencing and assembly strategies employed.

Acknowledgements The authors thank Dr. Bourlaye Fofana for reviewing and editing and Tara Edwards for English editing.

References

Agrios GN (2005) Plant Pathology. Elsevier Academic Press, Amsterdam

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796–815

Bairoch A, Apweiler R (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res 28:45–48

Belton JM, McCord RP, Gibcus JH, Naumova N, Zhan Y et al (2012) Hi-C: a comprehensive technique to capture the conformation of genomes. Methods 58:268–276

Bocklandt S, Hastie A, Cao H (2019) Bionano genome mapping: high-throughput, ultra-long molecule genome analysis system for precision genome assembly and haploid-resolved structural variation discovery. Adv Exp Med Biol 1129:97–118

Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W (2011) Scaffolding pre-assembled contigs using SSPACE. Bioinformatics 27:578–579

Bolsheva NL, Melnikova NV, Kirov IV, Dmitriev AA, Krasnov GS et al (2019) Characterization of repeated DNA sequences in genomes of blue-flowered flax. BMC Evol Biol 19:49

Borsch T, Quandt D (2009) Mutational dynamics and phylogenetic utility of noncoding chloroplast DNA. Plant Syst Evol 282:169–199

Brozynska M, Furtado A, Henry RJ (2016) Genomics of crop wild relatives: expanding the gene pool for crop improvement. Plant Biotechnol J 14:1070–1085

Bruna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M (2021) BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. NAR Genom Bioinform 3:lqaa108

Buchfink B, Xie C, Huson DH (2015) Fast and sensitive protein alignment using DIAMOND. Nat Methods 12:59–60

Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO et al (2013) Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. Nat Biotechnol 31:1119–1125

Chen H, Zeng Y, Yang Y, Huang L, Tang B et al (2020) Allele-aware chromosome-level genome assembly and efficient transgene-free genome editing for the autotetraploid cultivated alfalfa. Nat Commun 11:2494

Cloutier S, Ragupathy R, Miranda E, Radovanovic N, Reimer E et al (2012) Integrated consensus genetic and physical maps of flax (*Linum usitatissimum* L.). Theor Appl Genet 125:1783–1795

de Santana LA, Pacheco TG, Santos KGD, Vieira LDN, Guerra MP et al (2018) The *Linum usitatissimum* L. plastome reveals a typical structural evolution, new editing sites, and the phylogenetic position of Linaceae within Malpighiales. Plant Cell Rep 37:307–328

Diederichsen A, Ulrich A (2009) Variability in stem fibre content and its association with other characteristics in 1177 flax (*Linum usitatissimum* L.) genebank accessions. Ind Crops Prod 30:33–39

Dmitriev AA, Pushkova EN, Novakovskiy RO, Beniaminov AD, Rozhmina TA et al (2020) Genome sequencing of fiber flax cultivar Atlant using Oxford Nanopore and Illumina platforms. Front Genet 11:590282

Feschotte C, Jiang N, Wessler SR (2002) Plant transposable elements: where genetics meets genomics. Nat Rev Genet 3:329–341

Foulk JA, Akin DE, Dodd RB, Frederick JR (2004) Optimising flax production in the South Atlantic region of the USA. J Sci Food Agri 84:870–876

Fu Y-B (2021) Characterizing chloroplast genomes and inferring maternal divergence of the *Triticum-Aegilops* complex. Sci Rep 11:15363

Fu YB (2011) Genetic evidence for early flax domestication with capsular dehiscence. Genet Resour Crop Evol 58:1119–1128

Ghurye J, Pop M (2019) Modern technologies and algorithms for scaffolding assembled genomes. PLoS Comput Biol 15:e1006994

Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN et al (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci USA 108:1513–1518

Goldblatt P (2007) The index to plant chromosome numbers: past and future. Taxon 56:984–986

Gonzalez LG, Deyholos MK (2012) Identification, characterization and distribution of transposable elements in the flax (*Linum usitatissimum* L.) genome. BMC Genomics 13:644

Guisinger MM, Kuehl JV, Boore JL, Jansen RK (2008) Genome-wide analyses of Geraniaceae plastid DNA reveal unprecedented patterns of increased nucleotide substitutions. Proc Natl Acad Sci USA 105:18424–18429

Guo YY, Yang JX, Li HK, Zhao HS (2021) Chloroplast genomes of two species of *Cypripedium*: expanded genome size and proliferation of AT-biased repeat sequences. Front Plant Sci 12:609729

Harris MA, Clark J, Ireland A, Lomax J, Ashburner M et al (2004) The gene ontology (GO) database and informatics resource. Nucleic Acids Res 32:D258-261

Hastie AR, Dong L, Smith A, Finklestein J, Lam ET et al (2013) Rapid genome mapping in nanochannel arrays for highly complete and accurate *de novo* sequence assembly of the complex *Aegilops tauschii* genome. PLoS ONE 8:e55864

Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M (2016) BRAKER1: unsupervised RNA-Seq-Based genome annotation with GeneMark-ET and AUGUSTUS. Bioinformatics 32:767–769

Holt C, Yandell M (2011) MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics 12:491

Hon T, Mars K, Young G, Tsai YC, Karalius JW et al (2020) Highly accurate long-read HiFi sequencing data for five complex genomes. Sci Data 7:399

Hong CP, Park J, Lee Y, Lee M, Park SG et al (2017) accD nuclear transfer of *Platycodon grandiflorum* and the plastid of early Campanulaceae. BMC Genomics 18:607

Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ et al (2017) Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. Mol Biol Evol 34:2115–2122

Huerta-Cepas J, Szklarczyk D, Heller D, Hernandez-Plaza A, Forslund SK et al (2019) eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res 47:D309–D314

Jain M, Olsen HE, Paten B, Akeson M (2016) The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biol 17:239

Jansen RK, Kaittanis C, Saski C, Lee S-B, Tomkins J et al (2006) Phylogenetic analyses of *Vitis* (Vitaceae) based on complete chloroplast genome sequences: effects of taxon sampling and phylogenetic methods on resolving relationships among rosids. BMC Evol Biol 6:32

Jansen RK, Raubeson LA, Boore JL, dePamphilis CW, Chumley TW et al (2005) Methods for obtaining and analyzing whole chloroplast genome sequences. Methods Enzymol 395:348–384

Jones P, Binns D, Chang HY, Fraser M, Li W et al (2014) InterProScan 5: genome-scale protein function classification. Bioinformatics 30:1236–1240

Kanehisa M, Goto S, Kawashima S, Nakaya A (2002) The KEGG databases at GenomeNet. Nucleic Acids Res 30:42–46

Knox EB (2014) The dynamic history of plastid genomes in the Campanulaceae *sensu lato* is unique among angiosperms. Proc Natl Acad Sci USA 111:11097–11102

Kolmogorov M, Yuan J, Lin Y, Pevzner PA (2019) Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol 37:540–546

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH et al (2017) Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res 27:722–736

Korf I (2004) Gene finding in novel genomes. BMC Bioinformatics 5:59

Lam ET, Hastie A, Lin C, Ehrlich D, Das SK et al (2012) Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. Nat Biotechnol 30:771–776

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC et al (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921

Li B, Zheng Y (2018) Dynamic evolution and phylogenomic analysis of the chloroplast genome in Schisandraceae. Sci Rep 8:9285

Li R, Yu C, Li Y, Lam TW, Yiu SM et al (2009) SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25:1966–1967

Liu F-H, Chen X, Long B, Shuai R-Y, Long C-L (2011) Historical and botanical evidence of distribution, cultivation and utilization of *Linum usitatissimum* L. (flax) in China. Veget Hist Archaeobot 20:561–566

Lomsadze A, Burns PD, Borodovsky M (2014) Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. Nucleic Acids Res 42:e119

Luo MC, Ma Y, You FM, Anderson OD, Kopecky D et al (2010) Feasibility of physical map construction from fingerprinted bacterial artificial chromosome libraries of polyploid plant species. BMC Genomics 11:122

Luo R, Liu B, Xie Y, Li Z, Huang W et al (2012) SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. Gigascience 1:18

Manni M, Berkeley MR, Seppey M, Zdobnov EM (2021) BUSCO: assessing genomic data quality and beyond. Curr Protoc 1:e323

Marcussen T, Meseguer AS (2017) Species-level phylogeny, fruit evolution and diversification history of Geranium (Geraniaceae). Mol Phylogenet Evol 110:134–149

Marks RA, Hotaling S, Frandsen PB, VanBuren R (2021) Representation and participation across 20 years of plant genome sequencing. Nat Plants 7:1571–1578

Mehrotra S, Goyal V (2014) Repetitive sequences in plant nuclear DNA: types, distribution, evolution and function. Genom Proteom Bioinform 12:164–171

Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA et al (2021) Pfam: the protein families database in 2021. Nucleic Acids Res 49:D412–D419

Morgante M (2006) Plant genome organisation and diversity: the year of the junk! Curr Opin Biotechnol 17:168–173

Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV et al (2022) The complete sequence of a human genome. Science 376:44–53

Ottai MES, Al-Kordy MAA, Afiah SA (2011) Evaluation, correlation and path coefficient analysis among seed yield and its attributes of oil flax (*Linum usitatissimum*) genotypes. Aust J Basic Appl Sci 5:252–258

Ragupathy R, Rathinavelu R, Cloutier S (2011) Physical mapping and BAC-end sequence analysis provide initial insights into the flax (*Linum usitatissimum* L.) genome. BMC Genomics 12:217

Raubeson LA, Jansen RK (2005) Plant diversity and evolution: genotypic and phenotypic variation in higher plants. In: Henry RJ (ed) Chloroplast genomes of plants. CABI Publishing, Wallingford, pp 45–68

Rice A, Glick L, Abadi S, Einhorn M, Kopelman NM et al (2014) The chromosome counts database (CCDB)—a community resource of plant chromosome numbers. New Phytol 206:19–26

Rowland GG, Hormis YA, Rashid KY (2002) CDC Bethune flax. Can J Plant Sci 82:101–102

Ruan J, Li H (2020) Fast and accurate long-read assembly with wtdbg2. Nat Methods 17:155–158

Sa R, Yi L, Siqin B, An M, Bao H et al (2021) Chromosome-level genome assembly and annotation of the fiber flax (*Linum usitatissimum*) genome. Front Genet 12:735690

Saski C, Lee SB, Daniell H, Wood TC, Tomkins J et al (2005) Complete chloroplast genome sequence of *Gycine max* and comparative analyses with other legume genomes. Plant Mol Biol 59:309–322

Schwarz EN, Ruhlman TA, Sabir JSM, Hajrah NH, Alharbi NS et al (2015) Plastid genome sequences of legumes reveal parallel inversions and multiple losses of rps16 in papilionoids. J Syst Evol 53:458–468

Seol Y-J, Kim K, Kang S-H, Perumal S, Lee J et al (2017) The complete chloroplast genome of two *Brassica* species, *Brassica nigra* and *B. Oleracea*. Mitochondrial DNA Part A 28:167–168

Shafin K, Pesout T, Lorig-Roach R, Haukness M, Olsen HE et al (2020) Nanopore sequencing and the Shasta toolkit enable efficient *de novo* assembly of eleven human genomes. Nat Biotechnol 38:1044–1053

Shinozaki K, Ohme M, Tanaka M, Wakasugi T, Hayashida N et al (1986) The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. EMBO J 5:2043–2049

Singh KK, Mridula D, Rehal J, Barnwal P (2011) Flaxseed: a potential source of food, feed and fiber. Crit Rev Food Sci Nutr 51:210–222

Soni S (2021) A complete guide on flaxseed cultivation. https://krishijagran.com/agripedia/a-complete-guide-on-flaxseed-cultivation/

Stanke M, Keller O, Gunduz I, Hayes A, Waack S et al (2006) AUGUSTUS: *ab initio* prediction of alternative transcripts. Nucleic Acids Res 34:W435-439

Stankova H, Hastie AR, Chan S, Vrana J, Tulpova Z et al (2016) BioNano genome mapping of individual chromosomes supports physical mapping and sequence assembly in complex plant genomes. Plant Biotechnol J 14:1523–1531

Tollis M, Boissinot S (2012) The evolutionary dynamics of transposable elements in eukaryote genomes. Genome Dyn 7:68–91

Vaser R, Sovic I, Nagarajan N, Sikic M (2017) Fast and accurate *de novo* genome assembly from long uncorrected reads. Genome Res 27:737–746

Wambugu PW, Brozynska M, Furtado A, Waters DL, Henry RJ (2015) Relationships of wild and domesticated rices (*Oryza* AA genome species) based upon whole chloroplast genome sequences. Sci Rep 5:13957

Wang Z, Hobson N, Galindo L, Zhu S, Shi D et al (2012) The genome of flax (*Linum usitatissimum*) assembled *de novo* from short shotgun sequence reads. Plant J 72:461–473

Wee Y, Bhyan SB, Liu Y, Lu J, Li X et al (2019) The bioinformatics tools for the genome assembly and analysis based on third-generation sequencing. Brief Funct Genomics 18:1–12

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P et al (2007) A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8:973–982

Wu CS, Chaw SM (2014) Highly rearranged and size-variable chloroplast genomes in conifers II clade (cupressophytes): evolution towards shorter intergenic spacers. Plant Biotechnol J 12:344–353

Wu CS, Wang YN, Hsu CY, Lin CP, Chaw SM (2011) Loss of different inverted repeat copies from the chloroplast genomes of Pinaceae and cupressophytes and influence of heterotachy on the evaluation of gymnosperm phylogeny. Genome Biol Evol 3:1284–1295

Wu Z (2016) The completed eight chloroplast genomes of tomato from Solanum genus. Mitochondrial DNA A DNA Mapp Seq Anal 27:4155–4157

You FM, Cloutier S, Shan Y, Ragupathy R (2015) LTR Annotator: automated identification and annotation of LTR retrotransposons in plant genomes. Int J Biosci Biochem Bioinforma 5:165–174

You FM, Duguid SD, Lam I, Cloutier S, Rashid KY et al (2016) Pedigrees and genetic base of the flax varieties registered in Canada. Can J Plant Sci 96:837–852

You FM, Jia G, Xiao J, Duguid SD, Rashid KY et al (2017) Genetic variability of 27 traits in a core collection of flax (*Linum usitatissimum* L.). Front Plant Sci 8:1636

You FM, Xiao J, Li P, Yao Z, Gao J et al (2018) Chromosome-scale pseudomolecules refined by optical, physical, and genetic maps in flax. Plant J 95:371–384

Zhang J, Qi Y, Wang L, Wang L, Yan X et al (2020) Genomic comparison and population diversity analysis provide onsights into the domestication and improvement of flax. iScience 23:100967

Zhang Y, Edwards D, Batley J (2021) Comparison and evolutionary analysis of Brassica nucleotide binding site leucine rich repeat (NLR) genes and importance for disease resistance breeding. Plant Genome 14: e20060

Zimin AV, Puiu D, Hall R, Kingan S, Clavijo BJ et al (2017a) The first near-complete assembly of the hexaploid bread wheat genome, *Triticum aestivum*. Gigascience 6:1–7

Zimin AV, Puiu D, Luo MC, Zhu T, Koren S et al (2017b) Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. Genome Res 27:787–792

2 Repeat DNA Sequences in Flax Genomes

Nadeem Khan, Hamna Shazadee,
Frank M. You, and Sylvie Cloutier

2.1 Introduction

Humans have been growing flax (*Linum usitatissimum* L.) for its seeds and fiber since ancient times (Vaisey-Genser et al. 2003). Fiber flax is taller and has fewer branches toward the top of the stem than linseed. Linseed branches, on the other hand, develop from the center of the stem and yield large quantities of seeds (Diederichsen et al. 2003). Flax seeds are a rich source of omega-3 fatty acids and contain the essential alpha-linolenic and linoleic acids. Its health benefits have been proven in several studies (Mazza et al. 1989; Caligiuri et al. 2014; Goyal et al. 2014; Kezimana et al. 2018; Parikh et al. 2019). Also, flax seed contains lignans which are associated with reducing certain types of cancer (Goyal et al. 2014). In recent years, flax fiber has been used as a component of composite materials, with some fibers holding considerable promise for automotive, aerospace, and packaging industries where the length of the fiber is not as important as its other physico-chemical properties (Zhu et al. 2013; Mokhothu et al. 2015; Wu et al. 2016; Dhakal et al. 2019; Fombuena et al. 2019; Zhang et al. 2020a). Therefore, a greater understanding of the genes that influence the quality and yield, especially for seed and fiber, is expected to positively contribute to flax improvement.

The first draft of the genome sequence of the Canadian flax cultivar CDC Bethune, published in 2012, was obtained using Illumina short reads, which resulted in a contig assembly of 302 Mb of non-redundant sequences, representing a genome coverage of ~ 81% (Wang et al. 2012). In 2018, employing a BioNano optical map, a BAC-based physical map and genetic maps, the long scaffold sequences of the assembly were further validated, orientated, ordered, and assigned to chromosomes (You et al. 2018). This chromosome-scale pseudomolecule assembly contains a total of 316 Mb (including ~ 50 Mb gaps), with individual chromosome lengths of 15.60–29.40 Mb, covering 97% of the annotated genes in the original scaffolds-based assembly. Based on Illumina sequencing, Hi-C technology, and genetic mapping, scaffold-level genome assemblies of the Chinese linseed cultivar Longya-10, the fiber cultivar Heiya-14, and a pale flax landrace (*Linum bienne*) were released in 2020 (Zhang et al. 2020b). These three assemblies have 306.0, 303.7, and 293.5 Mb in

N. Khan · H. Shazadee · F. M. You (✉) ·
S. Cloutier (✉)
Ottawa Research and Development Centre,
Agriculture and Agri-Food Canada, 960 Carling
Avenue, Ottawa, ON K1A 0C6, Canada
e-mail: frank.you@agr.gc.ca

S. Cloutier
e-mail: sylviej.cloutier@agr.gc.ca

N. Khan · H. Shazadee
Department of Biology, University of Ottawa,
30 Marie Curie, Ottawa, ON K1N 6N5, Canada

F. M. You and B. Fofana (eds.), *The Flax Genome*, Compendium of Plant Genomes,
https://doi.org/10.1007/978-3-031-16061-5_2

total length, with the scaffold N50 lengths of 1235 kb, 700 kb, and 384 kb for Longya-10, Heiya-14, and the pale flax landrace, respectively. More recently, utilizing Oxford Nanopore Technologies (ONT) and Illumina platforms, the Russian flax fiber cultivar Atlant was sequenced and assembled, having a total sequence length of 361.7 Mb and an N50 of 350 kb (Dmitriev et al. 2021). Lastly, the Pacific Biosciences (PacBio) Hifi combined with Hi-C scaffolding was used to sequence the Chinese fiber flax cultivar Yiya-5 (Sa et al. 2021). The Hifi sequences for Yiya-5 were assembled with an N50 of 9.61 Mb and 336 contigs totaling 454.95 Mb. Hi-C scaffolding produced 15 chromosome-length pseudomolecules that covered 93% of the total length.

The reference genome sequence resources are extremely important for future research progress in functional genomics and evolutionary studies of flax, such as in the discovery of transposable elements (TEs). The flax genome has recently undergone whole-genome duplication (WGD) and is 55.36% covered with repeat elements in the fiber cultivar Yiya-5 (You et al. 2018; Sa et al. 2021). For example, using short reads, it is very easy to collapse during the assembly process due to homologous or repeat sequences (Dmitriev et al. 2021). As per reassociation kinetics studies, approximately half of the flax genome is low-copy-number sequences, while 35% is highly repetitive, and the remaining 15% belongs to the middle-repetitive sequence types, which often encompass transposable elements (Cullis 1981). Repetitive sequences, or repeats, therefore account for a significant fraction of the flax genome. Tandem and interspersed repeats, as well as copy number of variants, are key structural polymorphisms that can occur in either of these types of repetitive sequences (Hannan 2018). Generally, repetitive elements constitute the major proportion of genomes, indicating that they serve for vital biological purposes and should not be termed ‘junk DNA’ as they were referred to during the last century (Sperling et al. 2013). Current evidence points to their significant roles in evolution and human disease (Madireddy et al. 2017; Paulson 2018). The ubiquitous presence of repetitive DNA sequences in genomes causes difficulties during sequence assembly and automated annotation not only in flax, but also in other species. Thus, their identification and annotation is vital to understand functions and to aid in improving genome assemblies (Ragupathy et al. 2013). Hence, the major focus of this review will be on summarizing our current knowledge of the distribution of repeat sequences in cultivated and wild flax and on describing an automated pipeline for their discovery. The latter also holds potential for repeat discovery and characterization in other species.

2.2 Types and Distribution of Repetitive DNA Sequences

The efficiency of genomic sequencing has grown by a factor of ten in the last decade, and next-generation sequencing, PacBio, and ONT platforms can now sequence the whole human genome in a matter of days. This potential has sparked numerous studies targeted at sequencing the genomes of tens of thousands of individuals from both animal and plant species (Jain et al. 2018; Hunt et al. 2020; Dmitriev et al. 2021; Sa et al. 2021). For instance, PacBio and ONT sequencing platforms are revolutionizing our ability to capture an accurate picture of the molecular processes within the cell, leading to a deeper knowledge of the complex structural variants in a genome. However, some of the most difficult technical challenges associated with these new methods are caused by repetitive DNA: sequences that are similar or identical to other sequences in the genome. The majority of large genomes contain an abundance of repetitive sequences. For instance, repeats cover approximately half or more of the flax, wheat, and maize genomes (Cullis 1981; Garbus et al. 2015; Haberer et al. 2020). Repetitive DNA found in all domains of life—bacteria, archaea, and eukaryota—is classified into two types: interspersed repeats, which include TEs that occur in multiple loci across the genome, and tandem repeats (TRs) that occur at a single locus (Tørresen et al. 2019). TEs are typically several thousand base pairs (kbp) in length, but in eukaryotes, their size ranges from 100 bp to 20 kbp (Kidwell 2002).