Compendium of Plant Genomes
Series Editor: Chittaranjan Kole

Ilga M. Porth
Amanda R. De la Torre *Editors*

The Spruce Genome

MOREMEDIA
Springer

Compendium of Plant Genomes

Series Editor

Chittaranjan Kole, Raja Ramanna Fellow, Government of India, ICAR-National Research Center on Plant Biotechnology, Pusa, New Delhi, 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.

Interested in editing a volume on a crop or model plant? Please contact Prof. C. Kole, Series Editor, at ckoleorg@gmail.com

More information about this series at http://www.springer.com/series/11805

Ilga M. Porth • Amanda R. De la Torre
Editors

The Spruce Genome

Springer

Editors
Ilga M. Porth
Faculté de foresterie, de géographie
et de géomatique
Université Laval
Quebec, QC, Canada

Amanda R. De la Torre
School of Forestry
Northern Arizona University
Flagstaff, AZ, USA

ISSN 2199-4781 ISSN 2199-479X (electronic)
Compendium of Plant Genomes
ISBN 978-3-030-21000-7 ISBN 978-3-030-21001-4 (eBook)
https://doi.org/10.1007/978-3-030-21001-4

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

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 for 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 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 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 of 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, 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 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, 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 on 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 of 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

Preface

The Spruce Genome, an Important Resource to Fundamental Biological Research and Selective Tree Breeding

Main Text

Spruces (*Picea spp*.) are naturally abundant and widely distributed conifer tree species in the Northern hemisphere. Due to their enormous ecological and economic value, management of this important forest genetic resource has focused on conservation and tree improvement. Recently, with the aid of improved sequencing technologies and bioinformatics advances, a draft genome sequence of the 20 Gigabases Norway spruce (*P. abies*) genome was published (Nature 497:581 (2013)). Canadian white spruce hybrid (*P. glauca* × *engelmannii* × *sitchensis*) genome assembly followed in the same year (Bioinformatics 29:1492 (2013)), establishing spruce as a model species in gymnosperm genomics. Continuous efforts to improve the spruce genome assembly are underway, but are challenged by the inherent characteristics of conifer genomes: high amounts of repetitive sequences (introns and transposable elements) and large gene family expansions related to abiotic stress responses, secondary metabolism, and their defense responses against pathogens and herbivory. Because the assembly is still highly fragmented with millions of scaffolds, the generation of ultra-dense genetic maps allows anchoring these scaffolds onto the 12 haploid spruce chromosomes represented by the 12 linkage groups in a spruce genetic map. The generation of RNA-seq data further aids to improve scaffolds. Such data are also particularly valuable in comparative genomics and can highlight the functional divergence between species. Bacterial artificial chromosomes (BACs) sequencing has also served the spruce genomics research community greatly, by (a) unraveling the substantial presence of pseudogenes, (b) supporting the isolation of entire metabolite-biosynthetic genes, (c) facilitating conifer genome comparisons for microsynteny, and last but not least (d) proving indispensable for spruce genome assembly. Some of these BAC sequencing efforts predated the spruce genome sequencing project.

The post-genomic era has seen a surge in genomic applications not only for species amenable to population genomics using whole genome data. In fact, genomics applications using a reduced representation of the massive spruce genome have become very popular (e.g., exome capture sequencing, genotyping-by-sequencing, restriction site-associated DNA sequencing). Throughout this book, we highlight all areas that have been impacted by the acquisition of a high-quality reference genome for spruce. In brief, this volume aims to provide the latest information on (1) status of the genome assembly, (2) detailed insights into whole genome and gene family structure, (3) comprehensive genomic resources available for research, (4) emerging genomics tools for tree improvement programs, (5) genomics related to genetic conservation programs, and (6) functional genomics to improve gene function annotations.

Chapters 1 and 2 focus on the current state of the nuclear and organelle genome assemblies since the first publication of draft genomes and the newest attempts to use whole genome re-sequencing (WGS) data for variant calling. WGS is unprecedented for conifers' complex genomes, where reduced-representation-sequencing-based genotyping has been the state-of-the-art genomic method. By contrast, confident WGS-based variant calling in a population of 1000 individuals for poplar, an angiosperm forest tree species with a 45× smaller genome size, constitutes no major obstacle nowadays. For spruce, however, current challenges regarding such an approach remain. These challenges are highlighted in the respective book chapter on *Picea abies* and potential solutions are extensively discussed. The following two chapters are on repetitive elements, which represent an important 70% fraction of the genome, and retrotransposons are suspected to actually drive spruce genome expansion. The significant differences overall with angiosperm transposable elements dynamics are highlighted; an example also illustrates how BAC sequencing conclusively helps characterize features of a retroelement family important in explaining spruce genome evolutionary dynamics. The epigenomics chapter focuses on the current state of knowledge about epigenetic variation in spruce. One of the chapters devoted to comparative genomics looks at the comparison of nuclear and organelle genomes among spruces, and among spruces and other gymnosperms, focusing on aspects of comparative mapping, and rates of sequence evolution. Another comparative genomics chapter focuses on the sequencing and annotation of a few randomly selected BACs and provides further insights into whole genome evolution comparisons and genome structural features among conifers (spruce versus pine) with regard to genes versus transposons. Spruce genomic resources also have important implications for modern tree selective breeding. A separate chapter is therefore dedicated to genomic selection in white spruce and the increased ability to capture genetic gain by more accurate phenotype prediction models obtained from improved genomic resources. This became possible with the implementation of the genomic pairwise kinship relationship matrix among individuals. This relationship captures the traditional contemporary pedigree (i.e., half- and full-sib family relationships) as well as the historical pedigree through the identification of common DNA variants (SNPs) passed through generations. The following two chapters deal with

local adaptation in spruce, the genetic underpinnings of resistance to drought as well as of cold hardiness. This represents a highlight of the current knowledge in clinal genetic variation in spruces. The last two chapters describe genes and gene families implicated in the formation of terpenes and phenols, the most important secondary compounds in spruce defense. Some of these genes have anti-herbivory and pathogen resistance potential. The book will close with an outlook into emerging fields of research in spruce genomics.

Quebec, Canada — Ilga M. Porth
Flagstaff, USA — Amanda R. De la Torre
November 2019

Contents

1 Sequencing and Assembling the Nuclear and Organelle Genomes of North American Spruces

Inanc Birol and Amanda R. De la Torre

Abstract

Reference genomes provide valuable information to study the molecular biology and the genomic architecture of species, and constitute a baseline for applied sciences such as molecular breeding and gene editing. The sequencing of conifer genomes still lags behind other plant and animal species, with only a few available conifers having full sequence genomes to date. This chapter aims to describe details on the sequencing and bioinformatics analysis of the nuclear and organelle genome assemblies of the economically important white spruce (*Picea glauca*), and closely related *Picea* species *P. sitchensis* and *P. engelmannii*. The chapter finishes by providing some perspectives for future genome assemblies of North American species.

1.1 Introduction

We are at the dawn of an era for conifer genomics. Several groups around the world have developed reference resources for a variety of conifer species (Birol et al. 2013; Nystedt et al. 2013; Zimin et al. 2014; Warren et al. 2015)—valuable reference material to study the molecular biology of these species. Beyond advancing basic science, reference genome assemblies and their detailed annotations are keys to support the development of marker systems for applications in conifer breeding programs (De La Torre et al. 2014a).

The flurry of activity in the field is a manifestation of the wider availability and reducing costs of high throughput sequencing platforms. The Illumina sequencing technology (Illumina Inc., San Diego, CA), in its various forms, has dominated the field of de novo genome sequencing until very recently. Illumina, offers several instruments built on the concept of sequencing by synthesis, where clusters of amplified DNA fragments are interrogated by a series of fluorescently labeled reversible termination reactions (Bentley et al. 2008). On their high-throughput platform, the approach generates over a billion short and accurate sequences, typically achieving error rates under 1% with up to 250 base pair (bp) reads.

Recently, the proven performance and favourable per nucleotide cost of the Illumina

I. Birol
Department of Medical Genetics, Michael Smith Genome Sciences Centre, University of British Columbia, Vancouver, Canada
e-mail: ibirol@bcgsc.ca

A. R. De la Torre (✉)
School of Forestry, Northern Arizona University, 200 E. Pine Knoll, Flagstaff AZ86001, AZ, USA
e-mail: Amanda.de-la-torre@nau.edu

I. M. Porth and A. R. De la Torre (eds.), *The Spruce Genome*, Compendium of Plant Genomes,
https://doi.org/10.1007/978-3-030-21001-4_1

sequencers have been leveraged to provide co-localization information on short reads. The Chromium sequencing library preparation instrument from 10X Genomics (Pleasanton, CA) uses microfluidics to label fragments from large DNA molecules (reaching beyond 100,000 bp), expanding the utility of the short reads generated (Zheng et al. 2016).

In contrast to Illumina sequencing, instruments from Pacific Biosciences (PacBio, Menlo Park, CA), and Oxford Nanopore Technologies (ONT, Oxford, UK) generate long reads from single molecules, opening up new possibilities. These reads can be three orders of magnitude longer than the reads from Illumina platforms, albeit with higher error rates. The PacBio technology is built on real-time observation of DNA polymerase reactions on single-stranded DNA templates immobilized in zero-mode waveguide arrays (Eid et al. 2009). The ONT technology is the only sequencing approach on major commercial platforms that does not synthesise DNA, but reads electrical signals off single-strand molecules while they transition through nanopores.

These sequencing platforms continue to improve, and new technologies promise longer reads, better data quality, and higher throughput. With each wave of progress in sequencing technologies, new frontiers emerge for life sciences. However, this *emergence* is by no means spontaneous; enabling research on new frontiers invariably requires new analytics capabilities. The rapidly evolving field of conifer genomics is a prime example of how large-scale problems benefit from these developments, while also inspiring the development of cutting-edge bioinformatics technologies.

This chapter offers an account of this interaction, within the specifics of building reference resources for spruce genomes. It primarily focuses on algorithms developed at Canada's Michael Smith Genome Sciences Centre (Vancouver, BC), and how they were used in assembling the nuclear genome of the Canadian white spruce (*Picea glauca*) (Birol et al. 2013; Warren et al. 2015), and the organellar genomes of *P. glauca* and the Sitka spruce (*P. sitchensis*) (Jackman et al. 2016; Coombe et al. 2016; Lin et al. 2019). While the story of the spruce genomes is still being written, where projects are underway to complete their assemblies, we expect the algorithms reported here and their future versions will continue to play a significant role in collating and refining these valuable resources.

1.1.1 De Novo Assembly

The first DNA genome of an organism, the 5,375 base pair (bp) sequence of the bacteriophage phi X174, was published in 1977 by Sanger et al. (1977), inspiring generations of researchers attempt larger and larger genomes. While the de novo assembly problem this pioneering work had tackled was tractable even for manual reconstruction of the genome, the problem quickly becomes intractable for larger targets.

De novo sequence assembly refers to the reconstruction of a sequence (usually DNA or RNA sequence, in the context of genomics research) using redundant random sampling of the underlying sequence (usually, genome or transcriptome), without consulting a similar reference sequence. The sampling redundancy of a genome is called the fold coverage, and represented by x. Although wrong, fold coverage is often simply referred to as coverage. More correctly, the latter refers to the fraction of the genome covered by reads, yet the ambiguity in terminology is compensated by the unit used. For example, 30× coverage indicates that a genome is on average covered by 30 reads, while 30% coverage indicates the percentage of the bases in the genome represented by at least one read.

When two sequences overlap partially and unambiguously (meaning, the overlapping sequence is unique in the underlying genome) they can be merged to obtain an extended contiguous sequence, or contig in short. As such, de novo assembly task relies heavily on identifying pairs of overlapping sequence reads in experimental data. When a sequencing experiment generates n reads, in its naïve conception, the problem of identifying pairs of overlapping sequence requires comparison of every sequence

to every other sequence, hence is an $O(n^2)$ problem, the notation indicating the order of magnitude of the number of operations needed to perform the task.

Fast-forwarding to date, where the number of reads in a sequencing experiment is in the billions for large genomes, it becomes apparent that performing $O(10^{18})$ operations, even on modern high performance computers, would not be feasible.

Practical de novo sequence assembly algorithms use a variety of approaches to simplify this problem. One common technique in such complex computer science problems is to balance algorithmic complexity with computational memory use. In this problem space, search for read-to-read overlaps can be reformulated as a table look up, and $O(n)$ operation, by limiting the sought after overlaps to be of $k - 1$ bp, and by using the fact that the genomic alphabet is composed of four letters (bases): A for adenine, C for cytosine, G for guanine, and T for thymine.

To accommodate, every read is *shredded* into sub-sequences of length k bp, assuming the reads are at least of this length, and redundant observations of the same k bp sequences, or k-mers for short, are collapsed into single representations. The result of this simplification is a table of k-mers, and all other k-mers overlapping a given k-mer by $k - 1$ bp can be found by consulting this table to interrogate $k - 1$ bp sequences appending one of the four bases. Accounting for overlaps on both ends, this totals to eight lookups for each k-mer. When one conceptualizes k-mers as nodes on graph, and overlaps between k-mers as directed edges, the result is called a de Bruijn graph (Pevzner and Tang 2001).

Although this basic component of the genome assembly is well described, the subsequent stages of assembly pipelines vary from algorithm to algorithm, and are less defined. The most modern sequence assembly algorithms, can use paired end sequencing data to (1) disambiguate contig extensions and (2) build scaffolds. The latter happens when flanking sequences around a difficult-to-assemble sequence is unambiguous. When multiple sequencing data types are available, often bespoke pipelines are implemented to best leverage their information.

1.1.2 The Nuclear Genome

The engine behind reconstructing the 20 billion bp (Gbp) genome of the Canadian white spruce (*Picea glauca*) (Birol et al. 2013; Warren et al. 2015) was several algorithms implemented in the ABySS package (Simpson et al. 2009), including ABySS-Bloom and ABySS release v1.5.2, which made use of memory-efficient Bloom filters for analyzing the sequence of large genomes (Birol et al. 2013). The original draft assembly of the genome in 2013 (Birol et al. 2013) reported a contiguity of NG50 = 22,967 bp, reconstructing 20.8 Gbp in 4.9 million scaffolds. Only two years later, this was surpassed by an NG50 of 83,000 bp (Warren et al. 2015), resulting in the most contiguous spruce assembly to date (Table 1.1). This significant increase in contiguity was achieved by the use of transcriptome re-scaffolding and large fragment mate pair sequences. The coding gene space in white spruce only represents 0.11–0.37% of the genome, therefore the increase in contiguity obtained with transcriptome re-scaffolding was low in comparison to the high increase obtained with the use of large fragment mate pair sequences (Birol et al. 2013). The assembly strategy included three main steps. In the first one, the genome assembly PG29 v2 (Birol et al. 2013) was re-scaffolded using RNA-seq libraries and large-fragment mate pair data. In the second step, a second genome assembly (WS77111, Warren et al. 2015) was created. The draft assembly of genotype WS77111 was sequenced after genomic analyses of PG29 suggested the presence of introgression from other spruces (*P. engelmannii*, *P. sitchensis*) (De La Torre et al. 2014b). The third step included the use of the WS77111 draft assembly for second-stage long-range re-scaffolding of PG29 v3 informed by scaffold alignments to the assembly WS77111 (Birol et al. 2013). Besides increases in contiguity, all spruce (and conifer) reference genomes to date

Table 1.1 Comparison among sequence assemblies published for North American spruce genomes

Species	Individual sequenced	Assembly	Genome size (Gb)	Scaffold NG50 (kb)	References
white spruce	WS77111	v1	22.4	19.9	Warren et al. (2015)
white x Engelmann x Sitka spruce	PG29	v1	20.8	22.9	Birol et al. (2013)
	PG29	v2	20.8	41.9	Birol et al. (2013)
	PG29	v3	20.8	71.5	Warren et al. (2015)
	PG29	v4	20.8	83	Warren et al. (2015)

are still highly fragmented due to the challenges associated with sequencing and assembling highly repetitive, complex, and very large genome sizes (De La Torre et al. 2014a).

1.1.3 Organelle Genomes

Following the sequencing of the nuclear genomes, efforts were dedicated into sequencing the organelle genomes of white spruce and other North American spruces. Chloroplast and mitochondrial genomes were sequenced from clone PG29, the individual used in the PG29 v1-4 nuclear genome assemblies of white x Engelmann x Sitka spruce (Jackman et al. 2015). The 123-kb chloroplast and 5.9 mitochondrial genomes were sequenced with Illumina MiSeq and HiSeq at 80X and 30X of coverage, respectively. Pair-end reads were assembled using ABySS, gaps were closed with additional Illumina sequencing or PacBio long reads, and coding regions were annotated with MAKER (Campbell et al. 2014). Transcript abundance of the annotated mitochondrial genome was quantified using RNA-Seq data from three developmental stages and five tissues. The mitochondrial genome had an N50 of 369 kb and contained 106 protein-coding genes (51 distinct genes), 29 tRNAs, and 8rRNAs, whereas the chloroplast genome was composed by 74 protein-coding genes, 36 tRNAs, and 4rRNAs. In contrast to the highly variable angiosperm chloroplast genomes, the chloroplast genome of white spruce had high levels of synteny and co-linearity in relation to Norway spruce even though the species diverged more than 10 Mya (Jackman et al. 2015). The chloroplast of non-admixed white spruce genotype WS77111 was also sequenced a few years later (Lin et al. 2019). In comparison with the PG29 assembly, the WS77111 genotype produced a higher-quality assembly with a N50 lengths of 3692, 1313, and 949 bp (43X, 88X, and 172X) after assembly with ABySS (Lin et al. 2019), even though having a similar size (123,421 bp vs. 123,266 bp).

Sitka spruce (*Picea sitchensis*) and Engelmann spruce (*Picea engelmannii*) are other economically important species growing in western North America. The chloroplast genome of Sitka spruce was sequenced earlier than white spruce, using Solexa sequencing-by-synthesis technology (Cronn et al. 2008). A more recent assembly was sequenced using long-read technologies (Coombe et al. 2016). Sequencing libraries were prepared with 10X Genomics platform that allows the incorporation of long DNA fragments, and libraries were later sequenced using Illumina HiSeq. Assembly followed with ABySS and gaps were filled with Sealer (Paulino et al. 2015). The final assembly had a size of 124,049 bp, about 3873 bp larger than the previous assembly, and a 99% sequence identity with other spruce chloroplast genomes such as white and Norway

spruce (Coombe et al. 2016). The Engelmann spruce chloroplast genome was sequenced using whole-genome shotgun sequencing, and later assembled with ABySS to result in a 123,542 bp assembly (Lin et al. 2019). Finally, the mitochondrial genome of Sitka spruce was sequenced using the Oxford Nanopore MinION technology. Reads were assembled using Unicycler (Wick et al. 2017), resulting in a 5.5-Mbp mitochondrial genome assembly (Jackman et al. 2019).

1.2 Nucleotide Sequence Alignment

Bioinformatics technologies have always been in a race with genomics technologies, especially in the era of high-throughput sequencing, to deliver timely results. For example, as the sequencing throughput on the Illumina platforms increased from millions to tens and hundreds of millions of reads per lane—recently pushing the billion mark on their NovaSeq instrument—sequence alignment methods shifted through several paradigms to offer analytical capability to projects that use these platforms.

At the dawn of the NGS era, short-read instruments were initially marketed as re-sequencing platforms. Their reads ($\sim$30 base pairs (bp) at the time) were meant to be interpreted only with respect to a reference genome. To fill the void of bioinformatics tools capable of handling large volumes of data, Illumina provided Eland (Cox, unpublished), a sequence alignment algorithm based on the pigeon-holing principle—where a sequence is partitioned into n non-overlapping sections, requiring at least one of them to have a perfect match to the reference, allowing guaranteed performance for $n - 1$ mismatches.

For faster searches through the reference genome, MAQ algorithm (Li et al. 2008) introduced the concept of hash indexing with multiple spaced seeds (called noncontiguous seed templates in the paper), essentially allowing for missed hits, as long as there is enough statistical evidence from matching seeds. When the hash indexing paradigm also buckled under the increasing throughputs, the community moved on to using concepts from data compression algorithms, such as FM indexing (Ferragina et al. 2000). Algorithms that implemented this paradigm, such as BWA (Li et al. 2008) and Bowtie (Langmead and Salzberg 2012), offered an order of magnitude faster run times (Hatem et al. 2013), briefly catching up with the increase in sequencing throughputs.

While precise sequence alignment may be needed in many established sequence analysis pipelines, for instance, for variant analysis (Robertson et al. 2010; Van der Auwera et al. 2013), in certain other applications alignment-free methods may be more suitable, such as in quantifying gene expression levels. Methods like Salmon and Kallisto (Patro et al. 2014; Bray et al. 2016) employ a light mapping strategy based on databases of k-mers, sequences of uniform length k. To distinguish the two paradigms, we will call the conventional methods as sequence *alignment* methods, and methods like the ones implemented in Salmon and Kallisto as sequence *mapping* methods.

1.3 Gene Annotation

Information is not knowledge. While the former is about the question "what," the latter is about exploring the "how" and "why." While the former is about foraging and collating data, the latter is about interpreting the data, building hypotheses, and models. The turn of this century witnessed a rapid transformation in biology, which arguably was mostly an observational and descriptive (i.e., information) science, to a more interactive and predictive (i.e., knowledge) science. This transformation has been the starkest in genomics.

The DNA molecule is essentially an information storage medium. It harbours all the information needed for life, and enables the transfer of that information across generations. Since the discovery of the double helix structure of DNA, and the realization that it was the *sequence* of the constituting nucleotides that encode that information, researchers looked for

ways to read those sequences. As discussed in previous sections, there are widely available commercial platforms for DNA sequencing, using various experimental approaches, and a range of bioinformatics methods tightly coupled with their strengths and weaknesses.

An assembled genome, in and of itself, is just raw information, even when it is highly contiguous and correct. Even though that information harbours all the essential biology about the species it describes, inferring that link is what makes the assembled genome valuable. The exercise of cataloguing that link is called genome annotation.

Though the wide selection of methods we see in the de novo assembly field is not paralleled in the annotation of assembled genomes, several tools have served the community well. The genome annotation engines, such as Maker and its derivatives (Holt and Yandell 2011; Campbell et al. 2014; Liu et al. 2014), ab initio gene finders, such as SNAP (Korf et al. 2004) and Augustus (Stanke et al. 2006), and specialized tools, such as DOGMA for organellar genomes (Wyman et al. 2004), Prokka for prokaryotic genomes (Seemann 2014), and MaGe for microbial genomes (Vallenet et al. 2006) use a range of different paradigms. But, broadly speaking, they are based on the concept of sequence homology, carrying over what is known in previously annotated genomes of evolutionarily related species to the genome under study. Some annotation tools, such as Maker and Augustus, do this indirectly using machine learning approaches, such as hidden Markov models (HMMs). As recent exciting results from fundamental research on machine learning are fueling various genomics applications (Libbrecht and Noble 2015), it is worth revisiting the problem of automated genome annotation and interpretation.

In the annotation of the spruce genomes, we used one of the most widely used and sophisticated annotation tools, the Maker pipeline (Holt and Yandell 2011). Now in its second version, Maker2, the pipeline uses evidence from data sources, such as RNA-seq, for de novo gene annotations, and integrates predictions from established ab initio gene finders, including SNAP (Korf et al. 2004), Augustus (Stanke et al. 2006), and GeneMark (Besemer and Borodovsky 2005). We have successfully used Maker in a number of genome assembly projects (Quevillon et al. 2005; Diguistini et al. 2009, 2011; Chan et al. 2011; Birol et al. 2013; Feau et al. 2016; Haridas et al. 2013; Keeling et al. 2013; Jackman et al. 2016; Warren et al. 2015; Coombe et al. 2016; Hammond et al. 2017; Jones et al. 2017a, b). To our experience, the pipeline works best for contiguous assemblies, but genes that are split across multiple contigs or fall across scaffold gaps are partially annotated or missed. Also, while the pipeline includes a workflow that integrates InterProScan (Quevillon et al. 2005) to predict protein family (PFam) domains (Finn et al. 2006) for annotated coding transcripts, it does not provide suggested biological functions for its predicted genes.

1.4 Glossary

Contig—It is a set of overlapping reads (DNA segments) that after multiple sequence alignment result in a consensus sequence representing a region of the draft genome being sequenced.

Unitig—It is a high-confidence contig. Contigs consist of one or more unitigs.

Scaffold—It is an ordered and oriented set of one or more contigs. The scaffold representing a single sequence may also contain gaps.

Superscaffold—Represents a group of scaffolds, usually obtained when the contiguity of an existing draft genome is improved by the use of long reads.

HMM—Stands for Hidden Markov Model, a statistical model used to build sequence analysis algorithms in computational molecular biology.

De novo *assembly*—Refers to the assembly of a novel genome of a species that does not present any previous reference sequences available for alignment.

References

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J et al (2008) *Accurate whole human genome sequencing using reversible terminator chemistry*. Nature 456(7218):53–59

Besemer J, Borodovsky M (2005) *GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses*. Nucleic Acids Res 33:451–454

Birol I, Raymond A, Jackman SD, Pleasance S, Coope R et al (2013) *Assembling the 20 Gb white spruce (Picea glauca) genome from whole-genome shotgun sequencing data*. Bioinformatics 29(12):1492–1497

Bray NL, Pimentel H, Melsted P, Pachter L (2016) *Near-optimal probabilistic RNA-seq quantification*. Nat Biotechnol 34(5):525–527

Campbell MS, Holt C, Moore B, Yandell M (2014) *Genome Annotation and Curation Using MAKER and MAKER-P*. Curr Protoc Bioinformatics **48**: 4 11 1–39

Chan QW, Cornman RS, Birol I, Liao NY, Chan SK et al (2011) *Updated genome assembly and annotation of Paenibacillus larvae, the agent of American foulbrood disease of honey bees*. BMC Genom 12:450

Coombe L, Warren RL, Jackman SD, Yang C, Vandervalk BP et al (2016) *Assembly of the Complete Sitka Spruce Chloroplast Genome Using 10X Genomics' GemCode Sequencing Data*. PLoS ONE 11(9): e0163059

Cronn R, Liston A, Parks M, Gernandt DS, Shen R, Mockler T (2008) Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. Nucleic Acids Res 36:e122. https://doi.org/10.1093/nar/gkn502 PMID: 18753151

De La Torre AR, Birol I, Bousquet J, Ingvarsson PK, Jansson S, Jones SJM, Keeling CI, MacKay J, Nilsson O, Ritland K, Street N, Yanchuk A, Zerbe P, Bohlmann J (2014a) Insights into Conifer Giga-genomes. Plant Physiol 166:1–9

De La Torre AR, Roberts D, Aitken SN (2014b) Genome-wide admixture and ecological niche modeling reveal the maintenance of species boundaries despite long history of interspecific gene flow. Mol Ecol 23:2046–2059

Diguistini S, Liao NY, Platt D, Robertson G, Seidel M, Chan SK et al (2009) *De novo genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data*. Genome Biol 10(9):R94

Diguistini S, Wang Y, Liao NY, Taylor G, Tanguay P, Feau N et al (2011) *Genome and transcriptome analyses of the mountain pine beetle-fungal symbiont Grosmannia clavigera, a lodgepole pine pathogen*. Proc Natl Acad Sci U S A 108(6):2504–2509

Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G et al (2009) *Real-time DNA sequencing from single polymerase molecules*. Science 323(5910):133–138

Feau N, Taylor G, Dale AL, Dhillon B, Bilodeau GJ, Birol I et al (2016) *Genome sequences of six Phytophthora species threatening forest ecosystems*. Genomics Data 10:85–88

Ferragina P, Manzini G (2000) *Opportunistic data structures with applications*, in *41st Annual Symposium on Foundations of Computer Science, Proceedings* 390–398

Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T et al (2006) *Pfam: clans, web tools and services*. Nucleic Acids Res 34:D247–D251

Haridas S, Wang Y, Lim L, Alamouti SM, Jackman S, Docking R et al (2013) *The genome and transcriptome of the pine saprophyte Ophiostoma piceae, and a comparison with the bark beetle-associated pine pathogen Grosmannia clavigera*. BMC Genom 14:373

Hammond SA, Warren RL, Vandervalk BP, Kucuk E, Khan H, Gibb EA et al (2017) *The North American bullfrog draft genome provides insight into hormonal regulation of long noncoding RNA*. Nature Communications 8(1):1433

Hatem A, Bozdag D, Toland AE, Catalyurek UV (2013) *Benchmarking short sequence mapping tools*. BMC Bioinformatics 14:184

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

Jackman SD, Warren RL, Gibb EA, Vandervalk BP, Mohamadi H, Chu J et al (2016) *Organellar Genomes of White Spruce (Picea glauca): Assembly and Annotation*. Genome Biology and Evolution 8(1): 29–41

Jones SJ, Haulena M, Taylor GA, Chan S, Bilobram S, Warren RL (2017) *The Genome of the Northern Sea Otter (Enhydra lutris kenyoni)*. Genes (Basel) **8**(12)

Jones SJM, Taylor GA, Chan S, Warren RL, Hammond SA, Bilobram S, Mordecai G et al (2017) *The Genome of the Beluga Whale (Delphinapterus leucas)*. Genes (Basel) **8**(12)

Keeling CI, Yuen MMS, Liao NY, Docking TR, Chan SK, Taylor GA et al (2013) *Draft genome of the mountain pine beetle, Dendroctonus ponderosae Hopkins, a major forest pest*. Genome Biol 14(3):R27

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

Langmead B, Salzberg SL (2012) *Fast gapped-read alignment with Bowtie 2*. Nat Methods 9(4):357–359

Li H, Ruan J, Durbin R (2008) *Mapping short DNA sequencing reads and calling variants using mapping quality scores*. Genome Res 18(11):1851–1858

Libbrecht MW, Noble WS (2015) *Machine learning applications in genetics and genomics*. Nat Rev Genet 16(6):321–332

Lin D, Coombe L, Jackman SD, Gagalova KK, Warren RL, Hammond SA, Kirk H et al (2019) *Complete Chloroplast Genome Sequence of a White Spruce (Picea glauca, Genotype WS77111) from Eastern Canada*. Microbiology Resource Announcements **8** (23)

Liu J, Xiao H, Huang S, Li F (2014) *OMIGA: Optimized Maker-Based Insect Genome Annotation*. Mol Genet Genomics 289(4):567–573

Nystedt B, Street NR, Wetterbom A, Zuccolo A, LIn YC, Scofield DG (2013) *The Norway spruce genome sequence and conifer genome evolution.* Nature **497** (7451): 579–84

Patro R, Mount SM, Kingsford C (2014) *Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms*. Nat Biotechnol 32 (5):462–464

Paulino D, Warren RL, Vandervalk BP, Raymond A, Jackman SD, Birol I (2015) Sealer: a scalable gap-closing application for finishing draft genomes. BMC Bioinformatics 16:230. https://doi.org/10.1186/s12859-015-0663-4 PMID: 26209068

Pevzner PA, Tang H (2001) *Fragment assembly with double-barreled data*. Bioinformatics 17(Suppl 1): S225–S233

Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R (2005) *InterProScan: protein domains identifier*. Nucleic Acids Res 33:W116–W120

Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD et al (2010) *De novo assembly and analysis of RNA-seq data*. Nat Methods 7(11):909–912

Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes JC (1977) *Nucleotide sequence of bacteriophage phi X174 DNA*. Nature 265(5596):687–695

Seemann T (2014) *Prokka: rapid prokaryotic genome annotation*. Bioinformatics 30(14):2068–2069

Simpson JT, Wong K, Jackman SD, Schein JD, Jones SJM, Birol I (2009) *ABySS: a parallel assembler for short read sequence data*. Genome Res 19 (6):1117–1123

Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B (2006) *AUGUSTUS: ab initio prediction of alternative transcripts.* Nucleic Acids Research **34**(Web Server issue): W435–9

Vallenet D, Labarre L, Rouy Z, Barbe V, Bocs S, Cruvellier S et al (2006) *MaGe: a microbial genome annotation system supported by synteny results*. Nucleic Acids Res 34(1):53–65

Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A et al (2013) *From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline.* Curr Protoc Bioinformatics **43**: 11 10 1–33

Warren RL, Keeling CI, Yuen MMS, Raymond A, Taylor GA, Vandervalk BP et al (2015) *Improved white spruce (Picea glauca) genome assemblies and annotation of large gene families of conifer terpenoid and phenolic defense metabolism*. Plant J 83(2):189–212

Wick RR, Judd LM, Gorrie CL, Holt KE (2017) Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads Phillippy, AM, editor. PLoS Comput Biol 13:e1005595. https://doi.org/10.1371/journal.pcbi.1005595

Wyman SK, Jansen RK, Boore JL (2004) *Automatic annotation of organellar genomes with DOGMA*. Bioinformatics 20(17):3252–3255

Zheng GX, Lau BT, Schnall-Levin M, Jarosz M, Bell JM, Hindson CM et al (2016) *Haplotyping germline and cancer genomes with high-throughput linked-read sequencing*. Nat Biotechnol 34(3):303–311

Zimin A, Stevens KA, Crepeau MW, Holtz-Morris A, Koriabine M, Marcais G et al (2014) *Sequencing and assembly of the 22-gb loblolly pine genome*. Genetics 196(3):875–890

2 Variant Calling Using Whole Genome Resequencing and Sequence Capture for Population and Evolutionary Genomic Inferences in Norway Spruce (*Picea Abies*)

Carolina Bernhardsson, Xi Wang, Helena Eklöf, and Pär K. Ingvarsson

Abstract

Advances in next-generation sequencing methods and the development of new statistical and computational methods have opened up possibilities for large-scale, high-quality genotyping in most organisms. Conifer genomes are large and are known to contain a high fraction of repetitive elements and this complex genome structure has bearings for approaches that aim to use next-generation sequencing methods for genotyping. In this chapter, we provide a detailed description of a workflow for variant calling using next-generation sequencing in Norway spruce (*Picea abies*). The workflow starts with raw sequencing reads and proceeds through read mapping to variant calling and variant filtering. We illustrate the pipeline using data derived from both whole-genome resequencing data and reduced representation sequencing. We highlight possible problems and pitfalls of using next-generation sequencing data for genotyping stemming from the complex genome structure of conifers and how those issues can be mitigated or eliminated.

2.1 Introduction

Conifers were one of the last plant groups lacking genome assemblies; but recently, several draft genomes have become available for a number of conifers such as Norway spruce (*Picea abies*, Nystedt et al. 2013), Loblolly pine (*Pinus taeda*, Zimin et al. 2014; 2017), Sugar pine (*Pinus lambertiana*, Stevens et al. 2016), and Douglas fir (*Pseudotsuga menziesii*, Neale et al. 2017). This has opened up new possibilities to assess genome-wide levels of genetic diversity in conifers. Earlier studies of genetic diversity in Norway spruce has either been limited to coding regions (e.g., Heuertz et al. 2006; Chen et al.. 2012) or have used various complexity reduction methods, such as genotyping-by-sequencing,

Carolina Bernhardsson and Xi Wang—These authors contributed equally.

C. Bernhardsson · X. Wang · H. Eklöf · P. K. Ingvarsson (✉)
Department of Plant Biology, Linnean Centre for Plant Biology, Swedish University of Agricultural Sciences, 750 05 Uppsala, Sweden
e-mail: par.ingvarsson@slu.se

C. Bernhardsson
e-mail: carolina.bernhardsson@gmail.com

X. Wang
e-mail: xi.wang@umu.se

H. Eklöf
e-mail: helena.dahlberg@umu.se

C. Bernhardsson · X. Wang
Department of Ecology and Environmental Science, Umeå Plant Science Centre, Umeå University, 901 87 Umeå, Sweden

I. M. Porth and A. R. De la Torre (eds.), *The Spruce Genome*, Compendium of Plant Genomes,
https://doi.org/10.1007/978-3-030-21001-4_2

restriction site associated sequencing, or targeted capture sequencing (Baison et al. 2019) to estimate levels of genetic diversity within species. While we have learned a lot about levels of genetic diversity in Norway spruce from such studies, we still lack detailed information on, for instance, levels of nucleotide polymorphism and linkage disequilibrium in non-genic regions. However, with the availability of a reference genome sequence (Nystedt et al. 2013), whole genome resequencing is now also possible in conifers such as Norway spruce.

Conifer genomes are large (20–40 Gb) and have high repetitive content and current draft genome assemblies in conifers are therefore often fragmented into many, relatively short scaffolds. In addition, large fractions of the predicted genome sizes are also missing from reference genomes. The fragmented nature of conifer reference genome assemblies, combined with the high repetitive content make variant calling in conifers difficult. This is true regardless of which techniques have been used to generate sequencing data but perhaps more so for whole-genome resequencing data that can be expected to provide a relatively unbiased coverage of the target genome. In this chapter, we review methods available for variant calling using NGS data and outline some of the issues one may face when performing analyses of data from whole-genome resequencing (WGS) in Norway spruce. In particular, we discuss the performance of variant calling across different genomic contexts, such as coding and non-coding regions and regions known to be composed of repetitive elements. We also compare variant calling using WGS data with data derived from sequence capture probes, designed to target non-repetitive sequences in the *P. abies* genome and discuss how collapsed genomic regions in the assembly complicates the task of filtering for good, reliable variant- and genotype calls. Having access to robust variant calls is important for downstream analyses, such as population genomic analyses or inferences of the demographic history of individuals, populations or the species as a whole. To highlight these issues, we end by assessing how different approaches to variant calling alter the site frequency spectrum of variants and hence possible evolutionary inferences drawn from the data.

2.2 Sample Collection

We sampled 35 individuals of Norway spruce (*Picea abies*) spanning their natural distributions, mainly from Russia, Finland, Sweden, Norway, Belarus, Poland, and Romania for use in whole genome resequencing. Individuals Pab001–Pab015 were all derived from unique populations and no specific measurements were taken when they were collected. Samples were taken from newly emerged needles or dormant buds for each individual and stored at −80 °C until DNA extraction. In contrast, individuals Pab016–Pab035 were sampled from two different areas, one in the eastern and one in the western part of Västerbotten province in northern Sweden. Two different populations were sampled in each area, one old and untouched forest (>100 years old) and one young planted population (<20 years old). For each population, a transect was made and five trees were sampled from each population along the transect. From each tree, a number of fresh shoots were broken off and put into prelabeled ziplock bags before being taken back to the lab for DNA extractions.

Genomic DNA was extracted using Qiagen plant mini kit following manufacturer's instructions. All sequencing was performed at the National Genomics Initiative platform (NGI) at SciLifeLab facilities, Stockholm, Sweden; paired-end libraries with an insert size of 500 bp were run on different Illumina HiSeq platforms (Pab001–Pab006 on HiSeq 2000, while the remaining individuals on HiSeq X). The original location, estimated coverage from raw sequencing reads and coverage of mapped reads for each individual are given in Table 2.1.

Samples analyzed using sequence capturing methods were obtained from Bernhardsson et al. (2019; 1,997 haploid samples) and Baison et al. (2019; 526 diploid samples). For further information on sample collection, DNA extraction