Stefan Jansson · Rishikesh P. Bhalerao ·
Andrew T. Groover
Editors

Genetics and Genomics of *Populus*

Springer

Why and How *Populus* Became a "Model Tree"

Brian Ellis, Stefan Jansson, Steven H. Strauss, and Gerald A. Tuskan

Abstract Although *Populus* was not a favored experimental system for very many plant biologists in 2000, *P. trichocarpa* ultimately became only the third plant species to have its genome fully sequenced. Here we examine the many different factors that came into play when this species was abruptly elevated to the status of a new "model organism".

1 Model Systems Within Biological Research

The diversity and complexity of life-forms presents an enormous challenge to biologists. However, the common evolutionary origin of all organisms implies that what is learned about one organism can provide useful insight into its relatives. This concept has led to the selection of a wide array of "model or reference organisms" over the past 50 years, ranging from the early adoption of *Escherichia coli* as the model prokaryotic microbe to a recent focus on the mouse as a model for mammalian biology. There are few fixed criteria for selection of the ideal model organism, but the choice is typically strongly driven by the nature of the biological question(s) to be addressed and the availability of suitable tools or approaches to address the questions (Abzhanov et al., 2008). Thus, many aspects of prokaryotic biology can be profitably explored in *E. coli,* but if the question of interest involves bacterial spore formation, *Bacillus subtilis* is a better model. Similarly, the adoption of *Arabidopsis thaliana* as a model plant has allowed extraordinary progress to be made in understanding the fundamental features of plant biology over the last 20 years. The intense concentration of research on this single species fostered the development of powerful research tools and resources, including the first complete sequence of a plant genome. These resources, paired with genetic, genomic and other approaches, have revealed insights into fundamental plant biology, including induction and organogenesis of flowering, regulation of primary meristems and leaf

B. Ellis (✉)
Michael Smith Laboratories, University of British Columbia, Vancouver BC V6T 1Z4, Canada
e-mail: bee@interchange.ubc.ca; bee@msl.ubc.ca

S. Jansson et al. (eds.), *Genetics and Genomics of Populus*, Plant Genetics and Genomics: Crops and Models 8, DOI 10.1007/978-1-4419-1541-2_1,

development, and genes responsible for disease resistance. The list of accomplishments is long and impressive but *Arabidopsis* differs in many important respects from plant species of economic value such as legumes or cereals. *Medicago truncatula* and *Oryza sativa* have therefore also been chosen for intensive study, in order to explore symbiotic nitrogen fixation and monocot biology, respectively. In addition, to understand the evolutionary origins and mechanisms underlying developmental processes in seed plants necessitates the examination of similar developmental processes in diverse taxa.

Perennial woody plants are often closely related to annual herbaceous plants, yet woody species possess structural and lifestyle characteristics that differ dramatically from herbaceous annuals. *Arabidopsis* is thus not necessarily a good model for the study of arboreal traits, despite close taxonomic relationships with woody relatives. In light of the ecological and commercial importance of trees across the terrestrial landscape it has been clear for some time that, in order to address traits characteristic of woody plants and forest trees, a suitable "model tree" should be identified, around which key genetic and genomics resources could be developed (Fig. 1).

The primary contenders for this designation fell into two obvious classes – gymnosperms and angiosperms. From a commercial perspective, there was no question that conifers (*Pinus*, *Picea*, *Abies*, and *Larix*) dominated both the marketplace and much of the temperate landscape. Significant conifer genetic resources have therefore – for commercial purposes – been built that would be useful also for genomic research However, countering gymnosperms obvious utility were some serious experimental disadvantages, such as massive genome sizes, long generation times, inefficient transformation procedures and a relatively underdeveloped biological knowledge base. Among the angiosperm tree species, those in contention

Fig. 1 Which tree is the best model species? Photo from October 21 2007, by "Ragesoss" from Wikimedia commons (http://commons.wikimedia.org/wiki/Image:Autumn_leaves,_Talcott_Mountain_State_Park.jpg)

included poplar/aspen (*Populus*), willow (*Salix*), birch (*Betula*) and *Eucalyptus*, but at the time when active debate over selection of a model tree was underway, *Populus* already possessed two major advantages. One was a combination of several desirable biological traits, such as a modest genome size, facile genetic transformation, ease of vegetative propagation, rapid growth response after experimental treatments and a short generation time compared to most other forest trees. The other was the considerable body of baseline research and development already being conducted with *Populus* hybrids in Europe and North America (see below), driven largely by their exceptional vigor and commercial potential for short-rotation forestry.

2 Key Events That Led to Adoption of *Populus* as the Prime Tree Model System

The experience of the plant biology community with the power of gene manipulation in *Arabidopsis* for revealing gene function made the ability to efficiently transform any model tree a particularly high priority. Transformation capability in *Populus* had been examined soon after general methods for plant transformation and regeneration were first established in 1984–85, and the first publication on regenerated transgenic poplar occurred in 1987 (Filatti et al. 1987). Leading researchers in poplar transformation included M. Gordon and E. Nester at the University of Washington (USA), where many of the pioneering advances in *Agrobacterium* biology were made, and B. McCown in Wisconsin (USA), who had long worked on in vitro systems for poplar regeneration. M. DeBlock and W. Boerjan (Belgium), L. Jouanin and G. Pilate (France), A. Seguin (Quebec), and R. Meilan and S. Strauss (Oregon) had all produced and field tested transgenic poplars in the late 1980s and early 1990s. These studies demonstrated convincingly that phenotypically stable traits could be readily produced in *Populus* spp. using transgenic methods. The rate of somaclonal variation has been reported to be very low, and the stability in transgene expression very high, in transgenic *Populus*.

Clonal propagation of select genotypes is another important trait amongst non-domesticated forest tree species. Many forest trees are difficult to vegetatively propagate, or show substantial "maturation effects" after propagation that confound genetic differences. For example, many *Eucalyptus* species root poorly and in many other species, rooted cuttings become increasingly difficult with age of the parent tree. The derived plants also often show variable degrees of maturation effects, such as slow growth and modified wood properties. The success and uniformity in response to micropropagation and other tissue culture methods also declines with age. The maturation effects tend to be much smaller for *Populus* than for most other taxa of forest trees, which means that trees of a variety of ages, and varying tissue sources, can be used to establish clonal populations whose primary differences are genetic rather than physiological in origin.

As genomics technologies became more broadly accessible to the life science community and sequencing of more complex eukaryotic genomes gained momentum in the late 1990s it became clear that producing a genome sequence for a model

plant could launch a new era in plant biology research. Publication of the 125 Mb *A. thaliana* genome sequence in 2000 was indeed a landmark event in plant science, but interestingly, part of the justification for this first plant genomics effort was its potential impact on improvement of agricultural crops and forest productivity. Given the many unique characteristics of trees, the power of having an *Arabidopsis* genome sequence available only whetted the appetite of the tree biology community for similar "global biology" resources devoted to a "model tree", and the lobbying began in earnest.

In the early 1990s, three different efforts began to converge on the choice of *Populus* as the "model tree". In Sweden, researchers at the Swedish University of Agricultural Sciences in Umeå were successful in genetically transforming *Populus* and in 1997, the Swedish *Populus* genome program was launched as a collaboration between researchers in Umeå (both at Umeå University and the Swedish University of Agricultural Sciences) and Stockholm (the Royal Institute of Technology, KTH). ESTs were sequenced from wood-forming tissues and other sources. Spotted DNA microarrays were produced and, again, first used to study wood-formation but later many other processes, and the third generation array contained 25 k features. The primary motivation for choosing *Populus* as the model tree for the Swedish effort was purely scientific – transformability allowing for functional analysis. The genotype that was used for transformation (T89) is a hybrid between *P. tremula* and *P. tremuloides*, but EST sequencing was also performed on tissue from *P. tremula* growing naturally in the area, and also from *P. trichocarpa*. In total, over 100,000 ESTs were sequenced. *Populus* proteomics and metabolomics were also being developed at the time, and a gene knockout project and the first public *Populus* EST database were launched. Although full genome sequencing was discussed in Sweden, such an enterprise appeared beyond reach, considering the estimated cost of the project.

In Canada, intense lobbying by the biomedical research community, spearheaded by M. Smith, had finally convinced the federal government to commit some major research funding specifically to large-scale genomics projects that would be relevant to Canada. The vehicle for this activity was a new foundation (Genome Canada), where funds would be awarded on a competitive basis. In the first Genome Canada competition, two multi-million dollar forest tree genomics projects were funded – *Treenomix* (based in the University of British Columbia, Vancouver) and *Arborea* (based in Laval University, Quebec), both of which incorporated some component of *Populus* genomics research, as well as work on conifers. The focus on conifers in these projects reflected the reality of Canadian forestry, which relies almost exclusively on harvesting coniferous species. Interestingly, however, the inclusion of *Populus* was justified on the grounds that it had already become the de facto "model tree" from a genomics perspective, as attested by the recent development of the *Populus* EST resource in Sweden.

In the USA in 1990 *Populus* was selected as a U.S. Department of Energy's model woody crop. Funding through this program was managed by G. Tuskan and included research at many universities and government laboratories. At a Poplar Genome Steering Committee meeting in Portland, Oregon, on November 14, 2001, a sequencing strategy was presented, and bioinformatics and genetic resources

and other issues were discussed, not only by US but also Canadian and Swedish researchers. During this period, intensive work to lay down the strategy took place, and in 2002, as the Human Genome efforts at DOE were winding down, there was a petition within DOE to utilize the high-throughput sequencing capacity at the Joint Genome Institute (JGI) to address DOE-relevant missions. *Populus* was nominated, reviewed by an external committee and accepted in 2002. At that time, *Populus* represented the largest, most complicated genome to be sequenced, assembled and annotated by a single facility.

3 The *Populus* Genome Sequencing

The *Populus* clone chosen for sequencing was the female clone *Nisqually-1*, originally collected by R. Stettler along the Nisqually River south of Seattle. This clone had been used in control crosses, and a 10X BAC library had been created as part of a QTL cloning effort. After being dubbed the extra-ordinary *Populus* genotype, scions of this genotype are now growing replicated in several places around the world (Fig. 2).

Sequencing began in earnest in 2003 and was met with a number of serious challenges. *Populus*, like other perennial plants, is comprised of multiple genomes, i.e., the nuclear genome, the mitochondrial genome, the chloroplast genome and the genomes of multiple endophytes. Shotgun sequencing such a "mega-genome" had never been attempted before. First, although DNA is found in all living tissues within a plant, the quality and quantity of high-molecular weight DNA that can be

Fig. 2 Nisqually-1 growing in a greenhouse in Umeå Sweden (flanked by two of the editors of this volume)

extracted varies with tissue type. Young, partially expanded leaves provide the highest quantity of DNA per unit of volume of tissue. However, these tissues also contain very large amounts of mitochondrial and chloroplast DNA. The first libraries prepared for the shotgun sequencing were prepared using DNA from young leaves, and although efforts were made to reduce the amount of organellar DNA in the preparations, these libraries contained too much organelle DNA (≈10% chloroplast DNA) to allow for cost-efficient sequencing. To reduce the amount of organellar DNA, root tips were selected and DNA template was isolated using a sucrose gradient to separate the nuclei from organelles, followed by cesium chloride gradient centrifugation and pulsed-field gel electrophoresis. This approach successfully eliminated the majority of organellar DNA, although over 40 partial putative endophyte genomes remained in the template pool. Of these endophytes, six genomes have subsequently been fully sequenced, assembled and annotated – *M. populi* BJ001, *S. maltophilia*, *S. proteamaculans*, *Enterobacter* sp. 638, *Burkholderia cepacia* Bu72 and *P. putida* W619 [http://www.jgi.doe.gov/]

Once high-molecular weight DNA template was obtained it was used to create three cloning libraries with 3 , 8 and 40 kb inserts that were characterized using 700 bp end-reads from a bank of capillary sequencing machines. The first two billion high-quality bases were subjected to several rounds of assembly, each giving more complete coverage of the genome. The first draft assembly was completed in November 2003 and represented 384 Mb of captured sequence, with the 100 largest scaffolds containing ca 50% of the sequence. During the first half of 2003 2.2 billion additional bases were sequenced and added to the assembly. In early 2004, the final draft assembly was completed and represented 429 Mb contained in 2447 scaffolds, with N50 scaffold size of 1.9 Mb and N50 scaffold number of 58. With the aid of newly created physical and genetic maps, these sequence scaffolds were used to create a linear combination of 19 chromosomal units.

Preparatory work for gene modeling and annotation in *Populus* was occurring simultaneously with the assembly. Modeling gene structure (i.e. intron and exons, and transcribed but untranslated regions) in a new organism requires both robust algorithms and high-quality training sets, typically ESTs and/or full-length cDNA sequences. In this stage of the project, the entire *Populus* community rallied to provide relevant EST sequences. About ten groups throughout the world that had been creating small or large scale EST data sets provided their sequence data to create the training set for gene-calling algorithms. With three bioinformatics groups working on gene modeling, the strategy was to initially let each group independently perform individual gene calls on the assembled sequence. Three ab initio gene prediction algorithms, EUGENE, GRAIL, and FgenesH, were trained based on over 5,000 true and in silico full-length cDNAs and a pool of around 500,000 EST sequences. Even though the EST dataset was rather extensive and the average *Populus* gene shares high similarity with orthologous *Arabidopsis* genes, the different algorithms – or even the same algorithm but with different settings – produced results that were quite variable. Not surprisingly, genes with good EST support were often identically predicted while in the absence of ESTs, results could be confusing (Fig. 3).

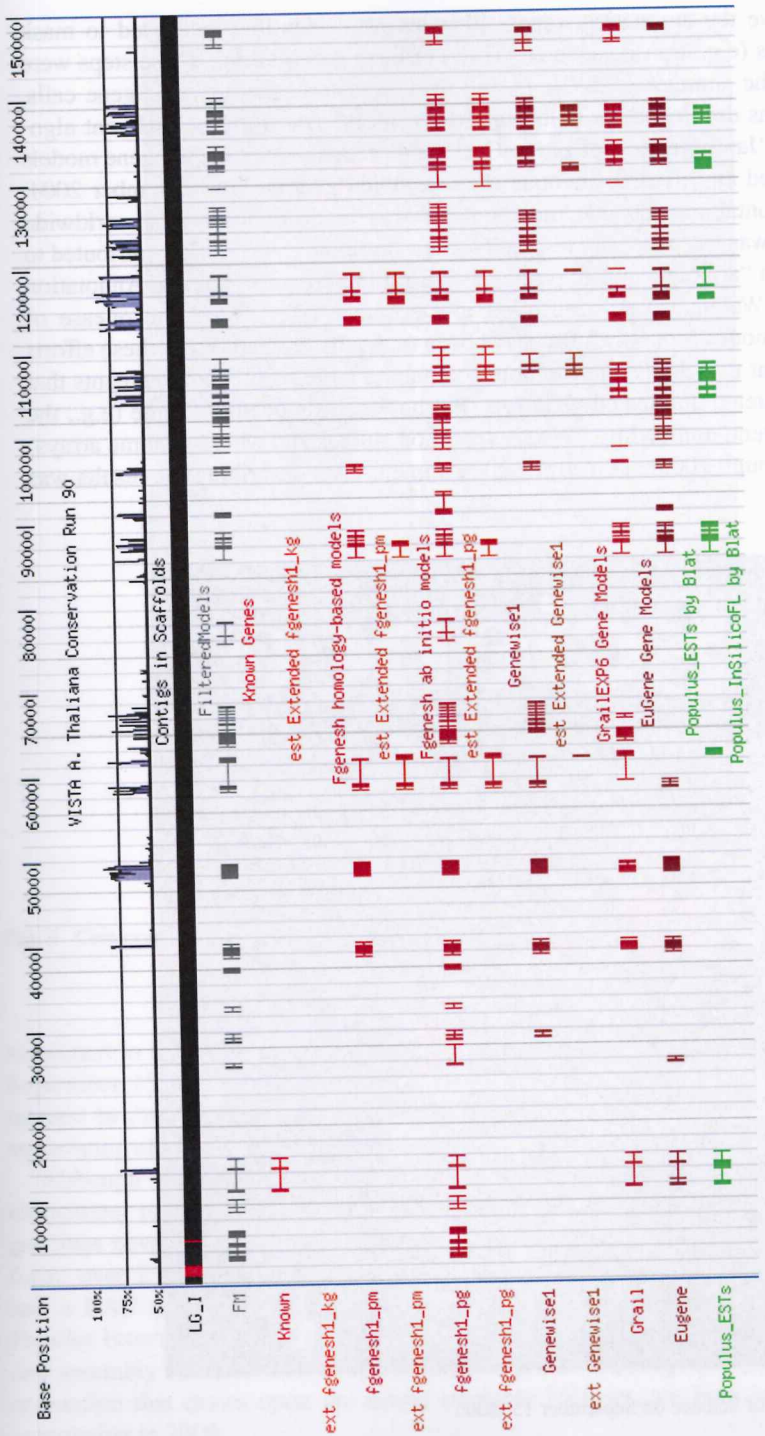

Fig. 3 Which gene model to choose? Genes predicted by FgenesH, Genewise, Grail and Eugene and EST coverage at the first 150 kb of LG I – a difficult region – in the first version of the Populus Genome Browser. The track at the top (FM, filtered models) represents the models that went on into the "Jamboree set"

To improve the annotation, repeat libraries were identified and used to mask repeat regions (e.g., transposable elements) prior to gene calling. These steps were finalized in the summer of 2004. In an effort to restrain the ab initio gene calls, a protocol was developed for collapsing the gene models from the different algorithms into a "Jamboree set" of genes. Of 55,054 predicted loci, 45,500 gene models were promoted and used to annotate the assembled genome. In September 2004, a database containing the genome sequence was made public and a worldwide press release was issued. Many research groups throughout the world contributed to this step, both "at home" but in particular during the *Populus* Genome Annotation Jamboree in Walnut Creek, California in December 2004. Since the release of 45,500 gene models, roughly 5,000 have been manually curated. From these efforts it was apparent that the *Populus* genome contained large paralogous segments that contained syntenic duplicated gene sets. Further analysis of the genome (e.g., the duplication event, non-coding RNAs, expression studies and whole-genome arrays) continued through 2005 and in April 2006 a manuscript describing the results was

Fig. 4 The cover of Science on September 15, 2006

Fig. 5 Ceremonial planting of Nisqually-1 at JGI by Jerry Tuskan and Dan Rokhsar

submitted to Science. The manuscript was accepted on August 9 and published on September 15, 2006 (Tuskan et al. 2006, Figs. 4 and 5). The increased scientific interest in *Populus* (Fig. 6) has of course been much influenced by the genome sequencing effort.

Although the publication marked the formal end of the *Populus* genome sequencing project, the work did not stop. For example, two additional *Populus* genomes have recently been resequenced by JGI using the Solexa short-read platform, over 2 million EST/cDNAs have been sequenced using the 454 platform, and a BAC minimum tiling path and QTL tracks have been added to the JGI *Populus* browser [http://genome.jgi-psf.org/Poptr1_1/Poptr1_1.home.html]. A second assembly based on subcloning BACs and primer walking, as well as a second annotation that draws upon the newly available EST set, are both scheduled for completion in 2009.

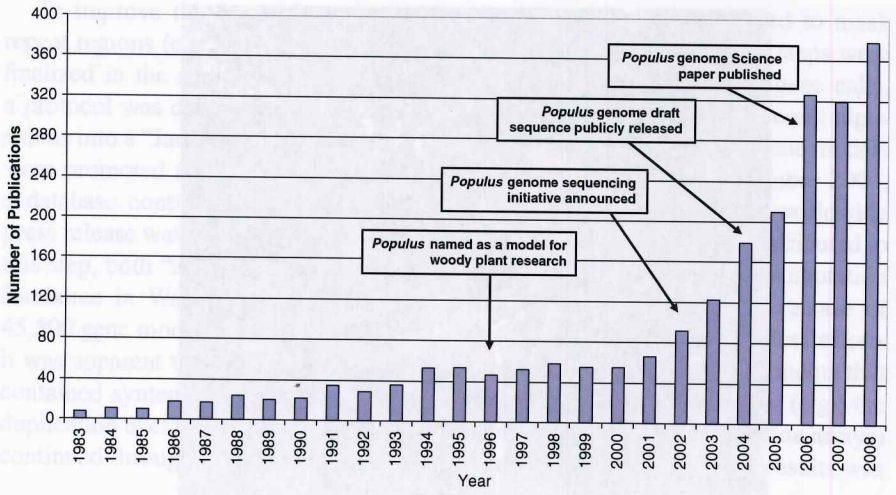

Fig. 6 Populus publomics

4 *Populus* Biotechnology and Breeding – Past and Future Visions

The ability of genomics methods to generate new DNA sequence data, and thus new possibilities for research and breeding, is growing rapidly. This trend is likely to continue for many years. However, the extent to which this knowledge is translated into benefits for society depends on social and economic factors that are difficult to predict. For example, the impact of transformation is extremely powerful for trees, in contrast to annual crops, since transformation allows for the introduction of new traits directly into elite germplasm without rounds of sexual propagation. This is especially important in trees, where genetic gains of clones (specific combining ability) are lost during outcrosssing. However, the ability to use transgenic traits is at present highly restricted, even for field research. Commercial applications are largely restricted to China, as a result of regulation and marketplace factors, and thus investments in applied research by government and industry sources outside of China are limited. This restriction may even grow greater if the pressures for living transgenic trees being incorporated into negotiations under the Cartagena Protocol on Biodiversity continue to grow (Strauss et al. 2009). In addition, in contrast to food crops, simply inherited and quality traits are rarely of major importance in forest tree breeding. The complex traits that are important, such as yield, adaptability, and wood quality are far more difficult to link to major genes. It is therefore unclear to what extent the limited numbers of molecular markers that are robustly identified in QTL and association genetics studies will be useful for marker-aided breeding. High levels of linkage equilibrium require that very large numbers of markers are employed to enable whole-genome marker selection, thus challenging current genotyping platforms and economics. This is especially true for

hybrid poplar programs, which are likely to have a complex QTL structure, and for which there is extensive genetic variation already present that can be captured by short-term trials and cloning without the use of markers. Thus, the economic driver for translational poplar genomic research is uncertain and reflects in part costs associated with sequencing and informatics. Whether translational research will take place with the expected growth of lignocellulosic bioenergy crops remains to be seen; it is likely that this will be highly influenced by costs and efficiencies of new sequencing and informatics technologies.

It is difficult to predict what scientists working with tree genetics and genomics will have achieved by 2020, and where future research emphasis may grow. It is, however, safe to assume that the present stage of *Populus* research will look rather primitive in 2020. The enormous advances in sequencing and profiling techniques will allow for full genome sequencing not only for additional species but also of a vast number of individuals of each species. When combined with thorough characterizations of the transcriptome, proteome, metabolome, lipidome, phosphorylome, etc. new systems-based approaches will be enabled that will more accurately model complex, multigenic traits such as maturation and wood formation. In the past, forest tree research has been limited by technical challenges. In the future, our understanding will increasingly be limited by our ability to pose relevant biological questions, accurately measure the phenotype of each genotype, dissect the relevant biological processes down to the subcellular level, and understand the complexity of genetic networks and signaling pathways in a scientific context that is related to the natural environment, where trees and other organisms constantly interact.

As discussed above, the selection of *Populus* as a model forest trees was highly influenced by practical issues. However, these considerations do not typically coincide with evolutionary or taxonomic realities for trees. For example, *Populus* is much more closely related to Arabidopsis and annual crop species than to coniferous tree species. Woodiness is possibly the ancestral state for angiosperms, and secondary growth and wood formation may even have homologous origins in angiosperms and gymnosperms. Increasingly, it will be vital to consider evolution of woody growth and taxonomic relationships in the study of trees. Perhaps the greatest contribution of *Populus* genomics research will be to identify in detail the basal mechanisms underlying secondary growth, wood development, maturation and perennial habit, ultimately providing a view of the evolution and development of perennial seed plants.

References

Abzhanov A, Extavour CG, Groover A, Hodges SA, Hoekstra HE, Kramer EM, Monteiro A (2008) Are we there yet? Tracking the development of new model systems. Trends Plant Sci 24: 353–360.

Filatti JJ, Sellmer J, McCown B, Haissig B, Comai L (1987) Agrobacterium-mediated transformation and regeneration of Populus. Mol Gen Genet 206:192–199.

Strauss SH, Tan H, Boerjan W, Sedjo. R (2009) Strangled at birth? Forest biotech and the convention on biological diversity. Nat Biotechnol 27:519–527.

Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, Schein J, Sterck L, Aerts A, Bhalerao RR, Bhalerao RP, Blaudez D, Boerjan W, Brun A, Brunner A, Busov V, Campbell M, Carlson J, Chalot M, Chapman J, Chen GL, Cooper D, Coutinho PM, Couturier J, Covert S, Cronk Q, Cunningham R, Davis J, Degroeve S, Dejardin A, Depamphilis C, Detter J, Dirks B, Dubchak I, Duplessis S, Ehlting J, Ellis B, Gendler K, Goodstein D, Gribskov M, Grimwood J, Groover A, Gunter L, Hamberger B, Heinze B, Helariutta Y, Henrissat B, Holligan D, Holt R, Huang W, Islam-Faridi N, Jones S, Jones-Rhoades M, Jorgensen R, Joshi C, Kangasjarvi J, Karlsson J, Kelleher C, Kirkpatrick R, Kirst M, Kohler A, Kalluri U, Larimer F, Leebens-Mack J, Leple JC, Locascio P, Lou Y, Lucas S, Martin F, Montanini B, Napoli C, Nelson DR, Nelson C, Nieminen K, Nilsson O, Pereda V, Peter G, Philippe R, Pilate G, Poliakov A, Razumovskaya J, Richardson P, Rinaldi C, Ritland K, Rouze P, Ryaboy D, Schmutz J, Schrader J, Segerman B, Shin H, Siddiqui A, Sterky F, Terry A, Tsai CJ, Uberbacher E, Unneberg P, Vahala J, Wall K, Wessler S, Yang G, Yin T, Douglas C, Marra M, Sandberg G, Van de Peer Y, Rokhsar D (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science 313:1596–1604.

Salient Biological Features, Systematics, and Genetic Variation of *Populus*

Gancho T. Slavov and Peter Zhelev

Abstract The genus *Populus* includes morphologically diverse species of deciduous, relatively short-lived, and fast-growing trees. Most species have wide ranges of distribution but tend to occur primarily in riparian or mountainous habitats. Trees from this genus are typically dioecious, flower before leaf emergence, and produce large amounts of wind-dispersed pollen or seeds. Seedlings are drought- and shade-intolerant, and their establishment depends on disturbance and high soil moisture. Asexual reproduction is common and occurs via root sprouting and/or rooting of shoots. Fossil records suggest that the genus appeared in the late Paleocene or early Eocene (i.e., 50–60 million years BP). According to one commonly used classification, the genus is comprised of 29 species divided into six sections, but a number of phylogenetic inconsistencies remain. Natural hybridization both within and among sections is extensive and is believed to have played a major role in the evolution of extant species of *Populus*. Both neutral molecular markers and adaptive traits reveal high levels of genetic variation within populations. Deviations from Hardy–Weinberg equilibrium are commonly detected in molecular marker studies. These deviations typically have small to moderate magnitudes and tend to be caused by heterozygote deficiency, indicating the possible existence of population substructure. Genetic differentiation among populations is much stronger for adaptive traits than for neutral markers, which suggests that divergent selection has played a dominant role in shaping patterns of adaptive genetic variation. Molecular and bioinformatic resources are actively being developed for multiple species of *Populus*, which makes this genus an excellent system for studying tree genetics and genomics.

G.T. Slavov (✉)
Department of Biology, West Virginia University, Morgantown, WV 26506, USA
e-mail: gancho.slavov@mail.wvu.edu

S. Jansson et al. (eds.), *Genetics and Genomics of Populus*, Plant Genetics and Genomics: Crops and Models 8, DOI 10.1007/978-1-4419-1541-2_2,