Università degli Studi di Udine
Dipartimento di Scienze Agrarie e Ambientali
Dottorato di Ricerca in Scienze e Biotecnologie Agrarie

Ph.D. Thesis

Next-generation sequencing in *Populus nigra*: *de novo* assembly, genome-wide SNP map and comparative genomic analysis

Candidate:
Stefania Giacomello

Supervisor:
Michele Morgante

March 14, 2012

Author's e-mail: giacomello@appliedgenomics.org

Author's address:

Dipartimento di Scienze Agrarie e Ambientali
Università degli Studi di Udine
Via delle Scienze, 208
33100 Udine
Italia

Abstract

The PhD thesis is divided in two parts. The first part proposes a whole genome Illumina resequencing approach to detect the genetic variation within the *Populus nigra* genome using the *Populus trichocarpa* genome sequence as reference. The SNP information obtained was applied to SNP frequency and population genetic studies at a genome scale to investigate the potential of next-generation sequencing to detect sequence variation within tree populations. This study was compared to a similar study conducted at a gene scale level on 18 *P. nigra* candidate genes for phenology traits resequenced using the traditional Sanger method. Moreover, it describes the application of the whole genome resequencing approach in 47 *P. nigra* genotypes to detect highly informative SNPs to be used in a 12k bead-chip, which was designed considering SNPs in candidate regions. The second part focuses on the exploitation of the Illumina next-generation sequencing to obtain the *P. nigra* genomic sequence using a *de novo* assembly approach and introduce the pan genome concept in poplar. In the first part of the thesis, 48 *P. nigra* (Sanger DP) genotypes were resequenced within 18 candidate genes using the Sanger method, whereas 4 *P. nigra* genotypes (4-clone DP) were resequenced at high coverage (> 20 X) across their whole genome using the Illumina technology. To evaluate the feasibility of using the Illumina technology in population genetic studies, we estimated the false discovery rate of the Illumina platform considering the ABI Sanger method as "gold standard" in terms of sequencing accuracy. It resulted to be 7.6%. The nucleotide diversity was estimated in different genic compartments. The Illumina 4 clone DP showed higher values for each class compared to the Sanger DP values. The higher diversity found at genome-wide level compared to that found in a small set of candidate genes may reflect weaker selective constraints on pseudogenes and non-functional gene copies. We also studied the nucleotide diversity distribution in the 19 chromosomes of the *P. nigra* genome within 100 kb windows: regions with higher gene content and lower repetitive sequences present lower nucleotide diversity estimates, and vice versa. The Tajima's D computation was performed within the 18 candidate genes resequenced by using the Sanger method and at a genome scale level in the 4 clones resequenced at high coverage. Gene D values were sligthly different between the 18 candidate genes (D = -0.679) and across the whole *P. nigra* genome (D = -0.373) probably due to the fact that in the first case we have a biased set of genes since they are candidates for the phenology traits and belong to the same metabolic pathway. Low-frequency alleles were more common than expected under a standard neutral model, which was reflected in negative distributions of Tajima's D for both sliding 100 kb windows and mean values in genes, as well as in putative transcripts (D = -0.356). In the DP of the 4 high coverage clones the interspecific divergence between the *P. nigra* and *P. trichocarpa* species was estimated at a genome level. Its value was on average almost an order of magnitude higher than the nucleotide diversity across the whole genome. As the nucleotide diversity, the higher values of the total divergence are localised in regions with higher gene content and lower repetitive sequences. Finally, for this part of the thesis, 47 *P. nigra* individuals were resequenced at low coverage to detect highly informative SNPs to be used in a high density bead chip which is under production. In this perspective, 42 clones out of the 47 were studied to characterise and evaluate the achievable SNP information using a stringent MAF parameter, which is required to provide markers to be used in marker assisted selection studies. In the second part of the thesis, the *de novo* sequencing of the *P. nigra* genome is presented. Two assemblers, CLC and ABySS, were used with a different combination of input data to have an alternative in terms of assembly quality. The input data consisted of standard paired-end reads, overlapping and mate-pair ones, for a total raw coverage of 123 X. The most accurate assembly, 339,551,115 bp long, was obtained using the CLC software using solely standard paired-end data. The assembly was validated in terms of capacity in reconstructing gene sequences, low percentage of chimeric reads and capacity in representing the heterozygous regions of the *P. nigra* genome. On this assembly, the scaffolding was performed using the SSPACE software and resulted to be reliable after its validation. To investigate the genomic differences between the *P. nigra* and the *P. trichocarpa* sequences, we designed a comparative genomic analysis pipeline that allowed us to introduce the pan genome concept in poplar composed by shared sequence portion between the two species and two species-specific portions. From the pipeline result, we concluded that the shared sequence portion is constituted by 200,586,680 bp and a putative shared sequence portion between the not-assembbled sequences of the two species. The *P. nigra*-specific portion is composed by 138,964,435 bp and by the *P. nigra* not-assembled sequence portion excluding those sequences that might be shared with the *P. trichocarpa* not-assembled sequence and, on the other hand, the *P. trichocarpa*-specific portion is composed by 216,551,264 bp and the *P. trichocarpa* not-assembled sequence portion excluding those sequences that might be shared with the *P. nigra* not-assembled portion. We also characterised the *P. nigra*-specific and the shared sequence portion and we saw that the shared portion contains a higher number of genes compared to the *P. nigra*-specific portion which is composed by a higher level of repetitive sequences.

Acknowledgments

I am grateful to Prof. Michele Morgante for encouraging and facilitating this work; Giusi Zaina, my mentor, for generous help and advice; Francesco Vezzi, Emanuele De Paoli, Simone Scalabrin and Fabio Marroni for helping me with some bioinformatic analyses; Nicoletta Felice, Irena Jurman and Federica Cattonaro for technical support; for INRA-Orleans for the plant material; INRA-Evry for the resequencing of 14 *Populus nigra* genotypes; UMEA University for useful discussions and technical exchange; Martin Lascoux, Stefan Jansson and Nathaniel for reviewing the thesis. This work has been financed by NoE EVOLTREE and the EU project NOVELTREE.

Contents

List of Figures

List of Tables

Introduction

Populus as a model system for plant biology

In plants, *Arabidopsis* has been adopted as the prime model system and an impressive number of tools and techniques are now available to understand gene function in this herbaceous species [1]. *Arabidopsis* was chosen as a model species for obvious reasons: small physical size, rapid generation time, straight-forward genetics, high fecundity, and small genome size. However, *Arabidopsis* is, in many respects, an unusual plant. As an almost obligate inbreeder, heterozygosity has been reduced to a minimum. Although this greatly facilitates functional studies, most plant species have different reproductive strategies, making *Arabidopsis* a genetic extreme. The very accelerated life cycle of *Arabidopsis* also makes many traits that are essential in many (or most) plants unimportant in *Arabidopsis.* Two obvious examples are wood formation and seasonality of growth.

The development of wood, or secondary xylem, from the vascular cambium is essential for tree growth and development and in providing support for a tall structure [3]. It is difficult to imagine that the formation of secondary xylem can ever be studied in anything other than a true tree. Surprisingly, however, *Arabidopsis* has proved useful in this role: many xylem-forming genes from *Arabidopsis* have been shown to be present in wood-forming tissues of pine [4].

In addition to secondary xylem, trees also exhibit complex patterns of activity and dormancy and the control of, for exemple, bud-burst must involve a complex of interactions between environmental signals (including day-length and temperature) and plant signal transduction pathways. It is hard to imagine that *Arabidopsis* could be adequate for the study of such processes in their entirety, since control may not be at the level of gene expression. Similarly, the age at which plants flower varies in woody species, occurring approximately 1 (*Salix*), 6 (*Populus*) and 60 (*Quercus*) years after germination. Despite this, many of the genes that control flowering in *Arabidopsis* are present in trees but they are used in different ways [3, 1]. This is illustrated by recent studies on the CONSTANS(CO)/FLOWERING LOCUS T (FT) regulatory module. In *Arabidopsis*, this pathway regulates the photoperiod-dependent induction of flowering [5]. In *Populus*, however, this module regulates not only flowering but also bud set in the late season [6], a process absent in *Arabidopsis.* Obviously, in the *Populus* lineage, the signal provided by the photoperiod-dependent oscillations of CO and FT mRNAs has been co-opted to regulate two different developmental pathways, whereas *Arabidopsis* only uses one of these. This example illustrates the power of comparative biology to provide important insights into the evolution of signal transduction pathways, insights that would be less obvious given a single model [1].

Physiologically and genetically, in many respects trees represent an opposite extreme to *Arabidopsis* in the spectrum of land plants, with long life spans and generation times, and woody perennial growth habits. However, trees do not form a monophylogenetic group but are found among many higher plant genera and families, and have arisen multiple times during land plant evolution. *Populus* is found in the angiosperm Euroside I clade together with *Arabidopsis*. Thus, *Arabidopsis* is more related to *Populus* than to the vast majority of other dicot taxa including those with trees, not to mention monocots like rice or gymnosperm trees such as conifers, lineages that separated from the eudicots long before the radiation of eudicot families 100-120 million years ago (mya). Thus, *Populus*, although relatively closely related to *Arabidopsis*, offers a new model system to study an expanded repertoire of biological processes that better represent the breadth of plant biology. The development of *Populus* as a model system for tree and woody perennial plant biology has been largely driven by the rapid development of genomic and molecular biology resources for this genus, as discussed below, culminating in the completion of a draft sequence of the *Populus trichocarpa* (black cottonwood) genome [2].

Populus nigra and its importance

The European black poplar (*Populus nigra* L.) is dioecious and wind-pollinated (Figure 1).

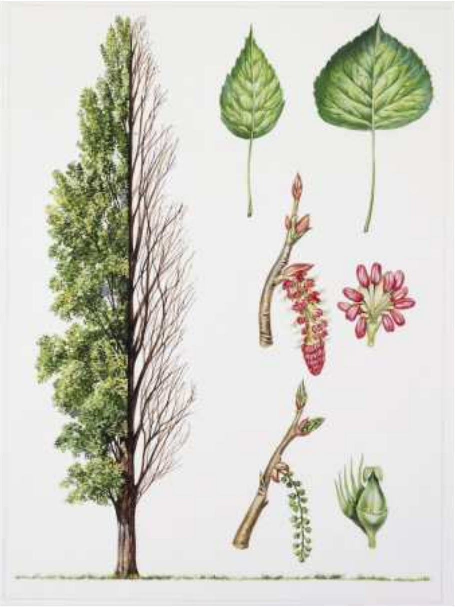


Figure 1: As a dioecious species, black poplar trees are either male or female. They reach the reproductive stage when they are 10-15 years old. Approximately 1-3 weeks prior to leaf initiation in the early spring (March-April), during the flood peak period along rivers in temperate Europe, male and female trees produce flowers clustered in pendulous catkins [23]. Red catkins belong to male trees, green catkins belong to female trees.

It is a keystone species for softwood floodplain forest ecosystems. As a pioneer species, it regenerates through colonisation of bare soil along the riverbank created by heavy flooding events. The wind- and water-dispersed seeds have a short viability period. They need highly specific water/soil conditions for germination [7, 8]. Successful regeneration therefore only occurs when the moisture of sediment remains high enough for seedling roots to establish [9]. Due to inappropriate conditions in many years, successful regeneration is absent. Accordingly, the history of flooding is reflected in a strong age structure that frequently exists in naturally occurring stands [10]. The natural regeneration of *P. nigra*, which is already patchy and sporadic, is additionally restricted by human activities. The drainage of rivers or management of riverbanks prevents natural flooding dynamics, which causes a lack of suitable areas for seedling establishment. Furthermore, today, many native populations of *P. nigra* have been replaced or fragmented by the widespread cultivation of commercially exploited hybrid poplars. From an economic point of view the most important hybrid combination is *P. x canadensis* (Dode) Guinier produced by crossings of *P. nigra* and the North American species *P. deltoides* Bartr. [11]. Cultivated *P. x canadensis* is a potential mating partner for *P. nigra.* Several studies have already reported on the production of viable progeny (e.g. [12, 13, 14]).

Forest fragmentation leads to a breakup of pollen- and seed-mediated gene flow [15]. Limitations of pollen and seed dispersal result in spatial aggregation of related individuals that is called isolation by distance [16], [17]. However, genetic diversity is important to allow a population to survive and reproduce under changing environmental conditions [18]. Due to the reasons outlined above, the European black poplar is one of Europe's rarest native trees ([19], [20]) although it has a wide distribution (Figure 2), ranging from Central and Southern Europe to Central Asia and North Africa [21]. In an attempt to conserve the genetic diversity that remains within this endangered species several European countries have independently set up *ex situ* genebanks in which cuttings of native black poplars from within each country are grown.

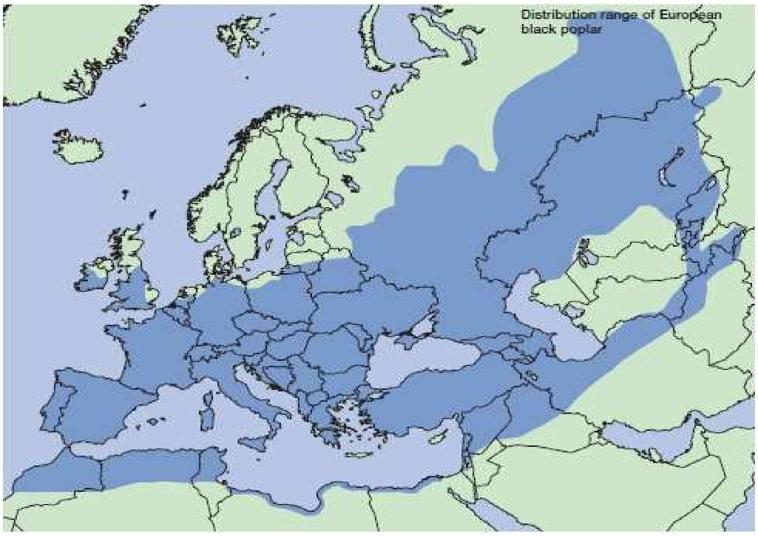


Figure 2: Black poplar has a large distribution area throughout Europe and is also found in northern Africa and central and west Asia. The distribution area extends from the Mediterranean in the south to approximately 64 latitude in the north and from the British Isles in the west to Kazakstan and China in the east. The distribution area also includes the Caucasus and large parts of the Middle East [23].

Populus trichocarpa: a close relative of *Populus nigra*

In 2006 the draft genome sequence of *Populus trichocarpa*, tree species native to western North America, was published [2]. The availability of such genomic sequence allows us to investigate deeper in the black poplar genome as the two species are closely related. I will describe some *P. trichocarpa* genomic features useful to better understand *P. nigra* genomic structure.

The *Populus* genome size was estimated to be 485 +/- 10 Mb (+/- SD), in rough agreement with previous cytogenetic estimates of about 550 Mb [42]. The near completeness of the shotgun assembly in protein-coding regions is supported by the identification of more than 95% of known *Populus* cDNA in the assembly. The ~75 Mb of unassembled genomic sequence is consistent with cytogenetic evidence that ~30% of the genome is heterochromatic [2]. The draft genome sequence consists of 2447 major scaffolds containing an estimated 410 megabases (Mb) of genomic DNA representing the 19 *Populus* chromosomes.

Populus and *Arabidopsis* lineages diverged about 100 to 120 million years ago (Ma). Analysis of the *Populus* genome provided evidence of a more recent duplication event that affected roughly 92% of the *Populus* genome. Nearly 8000 pairs of paralogous genes of similar age (excluding tandem or local duplications) were identified. Comparison of 1825 *Populus* and *Salix* orthologous genes derived from *Salix* EST suggests that both genera share this whole-genome duplication event. Moreover, the parallel karyotypes and collinear genetic maps [25] of *Salix* and *Populus* also support the conclusion that both lineages share the same large-scale genome history.

If we naively calibrated the molecular clock using synonymous rates observed in the *Brassicaceae* [26] or derived from the *Arabidopsis-Oryza* divergence [27], we would conclude that the genome duplication in *Populus* is very recent (8 to 13 Ma, as reported by Sterk [28]). Yet the fossil record shows that the *Populus* and *Salix* lineages diverged 60 to 65Ma [29, 30, 31].

The *Populus* gene duplications occurred as a single genome-wide event. We refer to this duplication event as the salicoid duplication event. Nearly every mapped segment of the *Populus* genome had a parallel paralogous segment elsewhere in the genome as a result of the salicoid event [2]. The colinearity of genetic maps among multiple *Populus* species suggests that the genome reorganization occurred before the evolution of the modern taxa of *Populus* [2].

Genomic diversity in forest trees

Today, many commercially important species have large natural distribution areas. Important phenotypic traits, such as timing of growth and cold tolerance, have clinal variation across environmental gradients in many widely distributed tree species [32, 33]. In contrast to many crop species, most commercially important forest species are essentially found in natural populations, with few traces of domestication and artificial selection.

In forestry, genomic discovery will support genetic improvement of tree varieties for solid wood, pulp and paper, biofuels, and biomaterials through integration into traditional breeding approaches in domesticated tree populations [34]. However, forest trees are also found throughout the world in an undomesticated state and are fundamentally important for noncommodity values such as plant and animal biodiversity, carbon sequestration, clean air and water, and human recreation [35].

Tree breeding in general is slow, as most species have long generation times. Thus, it is clearly desirable to find genetic and genomic tools that could speed up the detection of functionally important regions in the genome [36]. The genetic structure of populations defines how the natural variation can be used to detect and map genomic areas of functional importance. For instance, the scope for association studies needs to be assessed against population phenotypic variation and the extent of *linkage disequilibrium* (LD) [37], [38].

Natural genetic variation in forest trees has traditionally been investigated using two approaches: either quantitative genetics using common gardens or population genetics using markers. The common garden approach estimates genetic variation based on the measurement of phenotypic variation. Common gardens can be replicated over many different environments using either clonal or family-based testing [39]. This allows estimation of genotype by environment interactions. Quantitative genetic parameters such as heritability and additive and dominance variance components are used to characterize genetic variation. This approach is extremely useful for characterizing broad patterns of adaptive genetic variation and has been used in a practical manner to define seed or breeding zones in reforestation programs [39]. However, the individual genes underlying complex adaptive traits are not known, so single-locus population genetic theory cannot be applied. The alternative has been to use genetic markers for study of natural genetic variation. Markers such as isozymes, RFLPs, RAPDs, AFLPs, SSRs, and ESTPs have all been used but for the most part all reveal neutral genetic variation. Such markers are useful for characterizing demographic patterns of variation (migration and drift) but are not instructive of adaptive patterns of genetic variation. So before population genetic theory can be applied to genes controlling complex adaptive traits, they must first be discovered by complex trait dissection experiments. The QTL approach was first used, but because of low map resolution of QTLs, the underlying genes could not be determined [37]. The association genetics approach does provide much higher-level map resolution and it potentially can reveal individual genes underlying complex traits. Once the genes are identified that underly adaptive traits, then it is just one more step to discover the naturally occurring allelic variation in populations and test for the presence of selection using modern population genetic methods [40]. There are four aspects in forest trees which make advantageous the application of association genetic approach for dissecting complex traits [39]:

1. large and random-mating populations with minimal population structure;
2. adequate levels for nucleotide diversity for single nucleotide polymorphism (SNP) markers;
3. rapid decay of linkage disequilibrium (LD);
4. access to large clonal or family-based genetic tests for precise evaluation of phenotypes.

Effects of evolution on tree genomes

At nucleotide sites not influenced by selection (neutral sites), the rate of evolution is governed by the mutation rate. The nucleotide substitution rate at synonymous sites provides a direct estimate of the mutation rate. If we know the level of divergence of two species at the DNA level and have an estimate of their time of divergence, we can obtain an estimate of the rate of nucleotide substitution. In *Arabidopsis*, the rate of synonymous substitution has been estimated to be 1.5 x 10^{-8} per site per year. In *Populus*, the evolutionary rate has been estimated to be only one-sixth of *Arabidopsis* rate [2].

The standard neutral model provides a counterintuitive explanation of species diversity patterns. This model considers finite populations, where mating is random and there is no population subdivision. All individuals are equally likely to survive and reproduce. Synonymous variants at nucleotide sites can be assumed to evolve according to this model. The level of nucleotide diversity (calculated as π, the mean number of differences between all pairs of alleles in the sample [41]) in this neutral model is governed by the mutation rate and the size of the population. Nucleotide diversity is expected to be higher in large populations than in small ones, because random sampling (genetic drift) reduces variation in small populations more effectively. The diversity estimates of trees are highly variable, but not higher than those for annual plants (Figure 3).

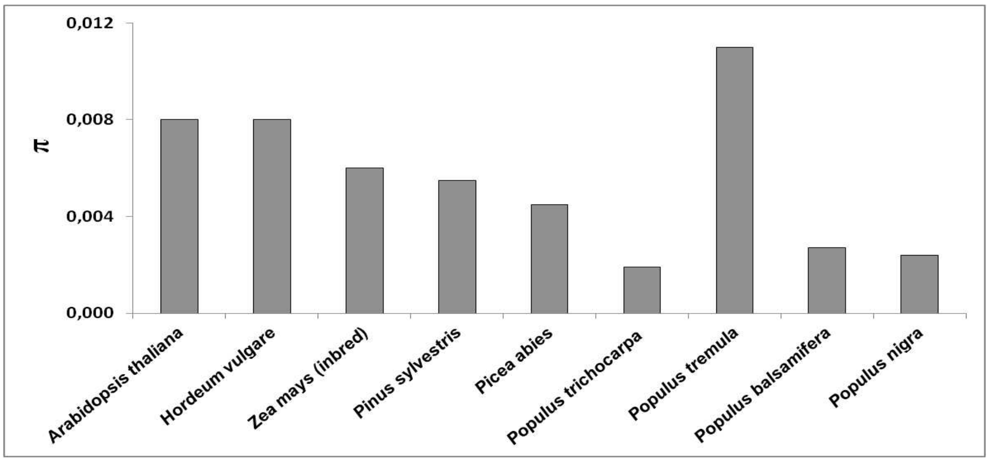


Figure 3: The histogram represents the nucleotide diversity (π) in the following species: *Arabidopsis thaliana* [42], *Hordeum vulgare* [43], *Zea mays* [44], *Pinus sylvestris* [45], *Picea abies* [46], *Populus trichocarpa* [47], *Populus tremula* [48], *Populus balsamifera* [49], *Populus nigra* [50].

This suggests that many tree populations might not be at an equilibrium situation, and that factors other than current population sizes, such as historical changes in population size, could have reduced the genome-wide diversity [36].

The theory above also predicts that there will be variation in the level of polymorphism across different loci, depending on the mutation rate. Loci that are much more or much less variable than predicted on the basis of divergence might be influenced by other evolutionary factors.

Closely linked nucleotide sites, such as sites within a single gene, are not expected to evolve independently. The statistical correlation between variant frequencies at two nucleotide sites (i.e. *linkage disequilibrium*, LD) is greater between closely linked sites than between those that are further apart. A pattern of low LD and rapid decay (R^2<0.05 within less than 500 bp) has been found in *Populus tremula* [48]; whereas a pattern of medium LD (R^2=0.37 over 600 bp) has been found in *P. trichocarpa* [47]. This is in strong contrast to the situation in highly inbred species such as *Arabidopsis thaliana* [51], where LD extends over distances of more than 20 kb in worldwide samples [52], but resembles the situation found in maize [53]. Tree populations have large distribution ranges and, to date, many species have been found to have rather uniform frequencies of different alleles across the range [36]. Within Europe, most nucleotide diversity resides within populations, and less than 5% is found between populations for *Populus tremula* [48]. For comparison, in a group of Central European populations of *Arabidopsis lyrata*, 17% of the variation was between the populations, indicating a much higher divergence [54].

The very slow neutral processes in the large long-lived tree populations are an important feature of tree populations [55, 56]. After population size changes, populations reach the new equilibria between mutation and genetic drift at a rate governed by 4N (where N is the population size). Before the new equilibrium is reached, the population genetic makeup differs from the equilibrium expectation. A population-size bottleneck first eliminates rare alleles, whereas a population expansion phase results in an increased number of nucleotide sites that have low frequency variants. Such consequences of population size changes have genome-wide effects.

Detecting selection in tree genomes

Natural selection leaves many traces in the genome. It is not easy to detect these signals, however, because many demographic events can result in similar patterns of polymorphism. For instance, the spread of a new favorable allele (selective sweep) is accompanied by a reduction of variability in areas surrounding the selected site [52]. However, such reduced diversity could also be due to a demographic departure from the standard neutral model. An excess of rare alleles can reflect the spread of a new favorable allele, but it can also follow from an expansion phase of the population. A high frequency of nucleotide sites that have intermediate frequencies of variants could be due to selection that maintains many alleles, such as at self-incompatibility loci, but it could also be due to a recent bottleneck.

Natural selection can also cause an increase in the local level of disequilibrium, as has been found in *A. thaliana* [51], [52], but LD also varies because of variation in recombination rates or due to population structure. Differential selection between populations in their local conditions can give rise to high differentiation, but this can also result from the isolation of populations. Currently, the best way to detect the effects of selection on a locus is to demonstrate that its pattern of polymorphism differs significantly from the genome-wide polymorphism [57]. Attempts to do this are being made using the very extensive datasets on populations of humans [58], maize [59] and *A. thaliana* [60]. These analyses require multi-locus datasets that are rarely available in trees.

A list of reports of selection on individual loci in trees can be found in [36]. Of the studied loci, 15% have been reported to be under some kind of selection, whereas in *A. thaliana*, selection was reported at 38% of the genes [57]. Among the many possible causes for such a difference, apparent patterns of selection due to demography may be less likely in the large random mating populations of forest trees than in other species. The mechanisms of selection in general are not known.

While the molecular traces of selection are still poorly understood in trees, studies on quantitative trait variation have demonstrated that selection can be highly efficient in large tree populations. For instance, many northern European populations of trees (e.g. willows, birches, sprucesand pines) have evolved genotypes whose growth is arrested at very long day lengths [61]. Such locally adapted ecotypes must have become frequent during the post glacial recolonization of the northern areas, just a few thousand years ago. The evolution in response to natural selection is very rapid compared to the slow rate of neutral evolution.

Even more can be learnt about selection when the nucleotide polymorphism can be correlated with phenotypic variation. The large, random-mating, low LD populations of trees are very well suited for association genetic studies [37], [38].

Revolution in Genomics: the next-generation sequencing

Over the past four years, there has been a fundamental shift away from the application of automated Sanger sequencing for genome analysis. Prior to this departure, the automated Sanger method had dominated the industry for almost two decades and led to a number of monumental accomplishments, including the completion of the only finished-grade human genome sequence [62]. Despite many technical improvements during this era, the limitations of automated Sanger sequencing showed a need for new and improved technologies for sequencing large numbers of human genomes. The automated Sanger method is considered as a first-generation technology, and newer methods are referred to as next-generation sequencing (NGS) [63].

Sanger sequencing: first generation

Since the early 1990s, DNA sequence production has almost exclusively been carried out with capillary-based, semi-automated implementations of the Sanger biochemistry [64, 65, 66] (Figure 4a). In high-throughput production pipelines, DNA to be sequenced is prepared by one of two approaches: first, for shotgun *de novo* sequencing, randomly fragmented DNA is cloned into a high-copy-number plasmid, which is then used to transform *Escherichia coli*; second, for targeted resequencing, PCR amplification is carried out with primers that flank the target [67]. The output of both approaches is an amplified template, either as many clonal copies of a single plasmid insert present within a spatially isolated bacterial colony that can be picked, or as many PCR amplicons present within a single reaction volume. The sequencing biochemistry takes place in a cycle sequencing reaction, in which cycles of template denaturation, primer annealing and primer extension are performed. The primer is complementary to known sequence immediately flanking the region of interest. Each round of primer extension is stochastically terminated by the incorporation of fluorescently labeled dideoxynucleotides (ddNTPs). In the resulting mixture of end-labeled extension products, the label on the terminating ddNTP of any given fragment corresponds to the nucleotide identity of its terminal position. Sequence is determined by high-resolution electrophoretic separation of the single-stranded, end-labeled extension products in a capillary based polymer gel. Laser excitation of fluorescent labels as fragments of discreet lengths exit the capillary, coupled to four-color detection of emission spectra, provides the readout that is represented in a Sanger sequencing trace. Software translates these traces into DNA sequence, while also generating error probabilities for each base-call [27], [69]. The approach that is taken for subsequent analysis for example, genome assembly or variant identification, depends on precisely what is being sequenced and why. Simultaneous electrophoresis in 96 or 384 independent capillaries provides a limited level of parallelization. After three decades of gradual improvement, the Sanger biochemistry can be applied to achieve read-lengths of up to ~1,000 bp, and per-base raw accuracies as high as 99.999%. In the context of high-throughput shotgun genomic sequencing, Sanger sequencing costs on the order of $0.50 per kilobase.

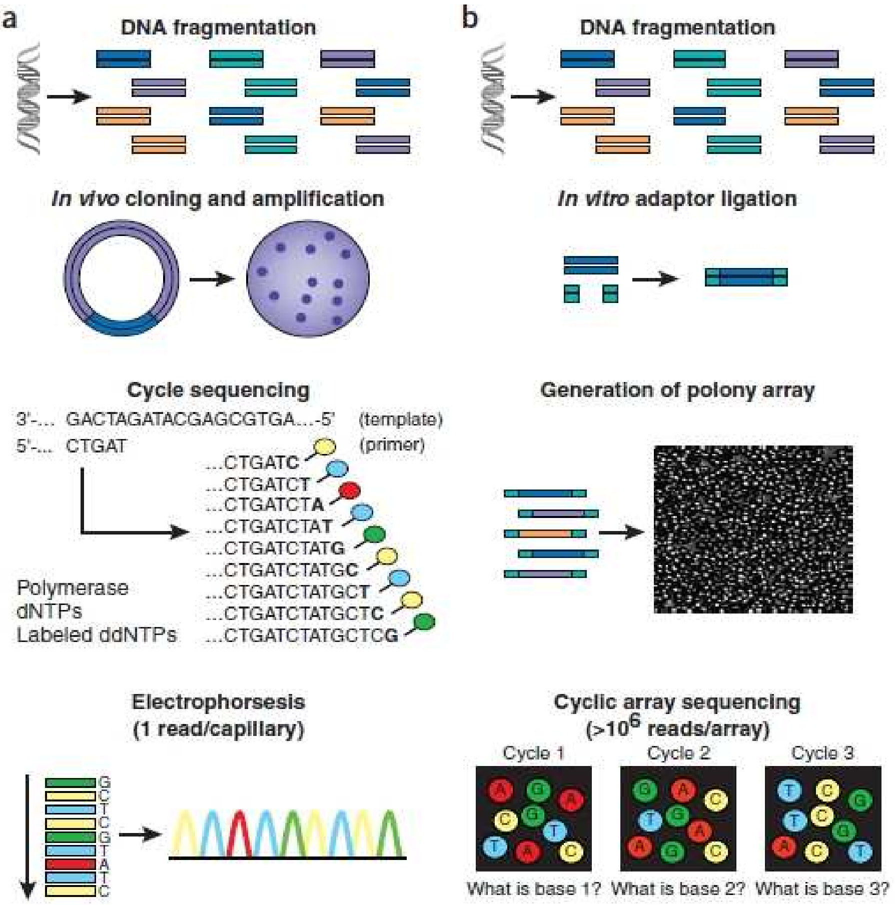


Figure 4: Work flow of conventional versus second-generation sequencing. (a) With high-throughput shotgun Sanger sequencing, genomic DNA is fragmented, then cloned to a plasmid vector and used to transform *E. coli.* For each sequencing reaction, a single bacterial colony is picked and plasmid DNA isolated. Each cycle sequencing reaction takes place within a microliter-scale volume, generating a ladder of ddNTP-terminated, dye-labeled products, which are subjected to high-resolution electrophoretic separation within one of 96 or 384 capillaries in one run of a sequencing instrument. As fluorescently labeled fragments of discrete sizes pass a detector, the four-channel emission spectrum is used to generate a sequencing trace. (b) In shotgun sequencing with cyclic-array methods, common adaptors are ligated to fragmented genomic DNA, which is then subjected to one of several protocols that results in an array of millions of spatially immobilized PCR colonies or polonies [70]. Each polony consists of many copies of a single shotgun library fragment. As all polonies are tethered to a planar array, a single microliter-scale reagent volume (e.g., for primer hybridization and then for enzymatic extension reactions) can be applied to manipulate all array features in parallel. Similarly, imaging-based detection of fluorescent labels incorporated with each extension can be used to acquire sequencing data on all features in parallel. Successive iterations of enzymatic interrogation and imaging are used to build up a contiguous sequencing read for each array feature.

Impact of next-generation sequencing technology on genomics

The impact of next-generation sequencing technology on genomics is in turn causing a revolution in genetics that will fundamentally change the nature of genetic experimentation due to a variety of factors [71]. When coupled with the appropriate computational algorithms, our ability to answer questions about the mutational spectrum of an organism, from single base to large copy number polymorphisms, on a genome-wide scale, is likely to radically alter our understanding of model organisms and ultimately of ourselves.

At the moment, there are five commercially available technologies: Roche/454, Illumina, Life/APG-Helicos BioSciences, the Polonator instrument and the near-term technology of Pacific Biosciences. These newer technologies constitute various strategies that rely on a combination of template preparation, sequencing and imaging, and genome alignment and assembly methods [63].

These platforms belong to the "cyclic-array sequencing". The concept of cyclic-array sequencing can be summarized as the sequencing of a dense array of DNA features by iterative cycles of enzymatic manipulation and imaging-based data collection [70], [72]. Although these platforms are quite diverse in sequencing biochemistry as well as in how the array is generated, their work flows are conceptually similar (Figure 4b). Library preparation is accomplished by random fragmentation of DNA, followed by *in vitro* ligation of common adaptor sequences. Alternative protocols can be used to generate jumping libraries of mate-paired tags with controllable distance distributions. The generation of clonally clustered amplicons to serve as sequencing features can be achieved by several approaches, including in situ polonies [70], emulsion PCR [73] or bridge PCR [74], [75]. What is common to these methods is that PCR amplicons derived from any given single library molecule end up spatially clustered, either to a single location on a planar substrate (in situ polonies, bridge PCR), or to the surface of micron-scale beads, which can be recovered and arrayed (emulsion PCR). The sequencing process itself consists of alternating cycles of enzyme-driven bio- chemistry and imaging-based data acquisition.

Advantages of the next-generation sequencing

What is it that sets next-generation sequencers apart from conventional capillary-based sequencing? Many features answer this question [71, 67]:

- the ability to process millions of sequence reads in parallel rather than 96 at a time. This massively parallel throughput may require only one or two instrument runs to complete an experiment;

- next generation sequence reads are produced from fragment libraries that have not been subject to the conventional vector-based cloning and *Escherichia coli* based amplification stages used in capillary sequencing. As such, some of the cloning bias issues that impact genome representation in sequencing projects may be avoided, although each sequencing platform may have its own associated biases. The workflow to produce next-generation sequence-ready libraries is straightforward; DNA fragments that may originate from a variety of front-end processes (described below) are prepared for sequencing by ligating specific adaptor oligos to both ends of each DNA fragment;

- relatively little input DNA (a few micrograms at most) is needed to produce a library;

- the array features are immobilized to a planar surface, so that they can be enzymatically manipulated by a single reagent volume. Although microliter-scale reagent volumes are used in practice, these are essentially amortized over the full set of sequencing features on the array, dropping the effective reagent volume per feature to the scale of picoliters or femtoliters.

Collectively, these differences translate into dramatically lower costs for DNA sequence production.

Looking in depth at the Illumina sequencing platform

Introduced in 2006, the Illumina sequencing platform is based on the concept of “sequencing by synthesis” (SBS) to produce sequence reads of 100-150 bp from tens of millions of surface-amplified DNA fragments simultaneously [71]. Starting from a mixture of single-stranded, adaptor oligo-ligated DNA fragments, the Illumina process uses a microfluidic cluster station to add these fragments to the surface of a glass flowcell. Each flowcell is divided into eight separate lanes, and the interior surfaces have covalently attached oligos complementary to the specific adapters that are ligated onto the library fragments. Hybridization of these DNAs to the oligos on the flowcell occurs by an active heating and cooling step, followed by a subsequent incubation with reactants and an isothermal polymerase that amplifies the fragments in a discrete area or cluster on the flow cell surfaces. In particular, amplified sequencing features are generated by bridge PCR [74], [75] (Figure 5).

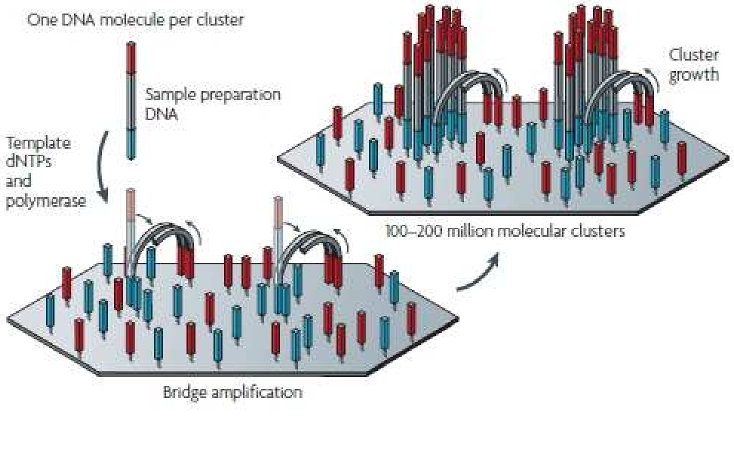

One DNA molecule per cluster
Sample preparation DNA
Template dNTPs and polymerase
Cluster growth
100–200 million molecular clusters
Bridge amplification

Figure 5: Illumina technology relies on bridge PCR [74],[75] to amplify clonal sequencing features. In brief, an *in vitro*constructed adaptor-flanked shotgun library is PCR amplified, but both primers densely coat the surface of a solid substrate, attached at their 5’ ends by a flexible linker. As a consequence, amplification products originating from any given member of the template library remain locally tethered near the point of origin. At the conclusion of the PCR, each clonal cluster contains ~1,000 copies of a single member of the template library. Accurate measurement of the concentration of the template library is critical to maximize the cluster density while simultaneously avoiding overcrowding [67].

In this approach, both forward and reverse PCR primers are tethered to a solid substrate by a flexible linker, such that all amplicons arising from any single template molecule during the amplification remain immobilized and clustered to a single physical location on an array. The resulting clusters each consist of ~1,000 clonal amplicons. Several million clusters can be amplified to distinguishable locations within each of eight independent lanes that are on a single flow-cell. After cluster generation, the amplicons are single stranded (linearization) and a sequencing primer is hybridized to a universal sequence flanking the region of interest.

Subsequently, the flowcell is placed into a fluidics cassette within the sequencer to start the **Cyclic Reversible Termination** (CRT). The CRT uses reversible terminators in a cyclic method that comprises nucleotide incorporation, fluorescence imaging and cleavage [76]. In the first step, a DNA polymerase, bound to the primed template, adds or incorporates just one fluorescently modified nucleotide (Figure 6), which represents the complement of the template base.

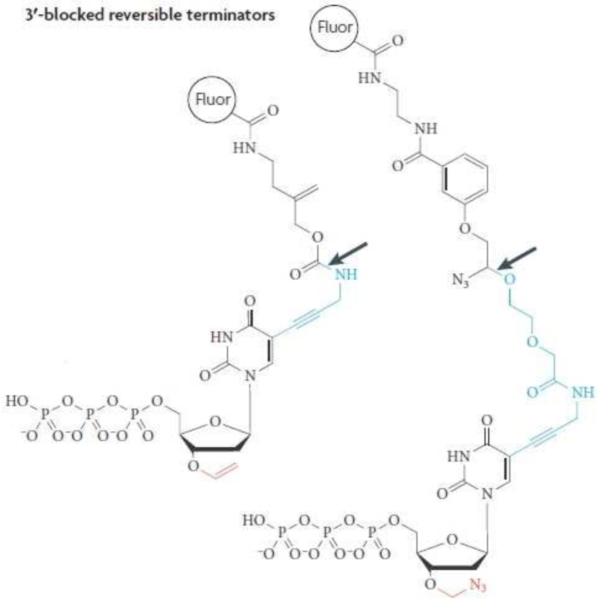


Figure 6: In the figure, red chemical structures denote terminating functional groups. Arrows indicate the site of cleavage separating the fluorophore from the nucleotide, and the blue chemical structures denote residual linker structures or molecular scars that are attached to the base and accumulate with subsequent cycles. DNA synthesis is terminated by reversible terminators following the incorporation of one modified nucleotide by DNA polymerase. 3'-blocked terminators contain a cleavable group attached to the 3'-oxygen of the 2'-deoxyribose sugar.

The termination of DNA synthesis after the addition of a single nucleotide is an important feature of CRT. Following incorporation, the remaining unincorporated nucleotides are washed away. Imaging (Figure 7) is then performed to determine the identity of the incorporated nucleotide. This is followed by a cleavage step, which removes the terminating/inhibiting group and the fluorescent dye. Additional washing is performed before starting the next incorporation step.

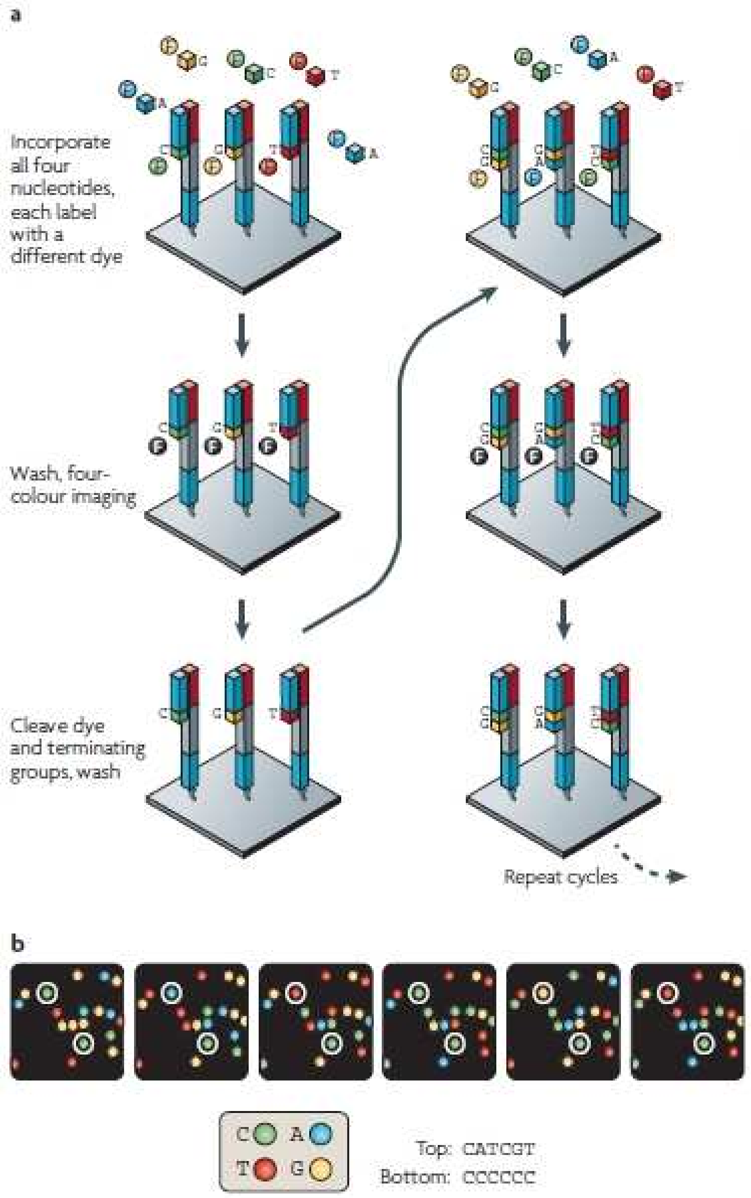

a
Incorporate all four nucleotides, each label with a different dye
Wash, four-colour imaging
Cleave dye and terminating groups, wash
Repeat cycles
b
C A
T G
Top: CATCGT
Bottom: CCCCCC

Figure 7: a The four-colour cyclic reversible termination (CRT) method uses Illumina 3'-O-azidomethyl reversible terminator chemistry [77], [78] using solid-phase-amplified template clusters. Following imaging, a cleavage step removes the fluorescent dyes and regenerates the 3'-OH group using the reducing agent tris 2-carboxyethylphosphine (TCEP) [77].

Currently, the Illumina HiSeq 2000 dominates the NGS market. It uses the clonally amplified template method illustrated in Figure 5, coupled with the four-colour CRT method illustrated in Figure 7a. The four colours are detected by total internal reflection fluorescence (TIRF) imaging using two lasers, the output of which is depicted in Figure 7b. At the end of the sequencing run (the run usually lasts ~11 days), the sequence of each cluster is computed and subjected to quality filtering to eliminate low-quality reads. Sequence data obtained per run are usually 540-600 Gb.

Read-lengths are limited by multiple factors that cause signal decay and dephasing, such as incomplete cleavage of fluorescent labels or terminating moieties. The dominant error type is substitution rather than insertions or deletions (and homopolymers are certainly less of an issue than with other platforms such as 454). The substitution error has a higher occurring rate when the previous incorporated nucleotide is a G base [99]. Genome analysis of Illumina data has revealed an underrepresentation of AT-rich [99], [25] and GC-rich regions [25], which is probably due to amplification bias during template preparation.

Next-generation sequencing applications

The production of large numbers of low-cost reads makes the NGS platforms described above useful for many applications. These include:

- variant discovery by resequencing targeted regions of interest or whole genomes,
- *de novo* assemblies of bacterial and lower eukaryotic genomes,
- cataloguing the transcriptomes of cells, tissues and organisms (RNAseq) [81],
- genome-wide profiling of epigenetic marks and chromatin structure using other seq-based methods (ChIPseq, methylseq and DNaseseq) [82],
- species classification and/or gene discovery by metagenomics studies [83].

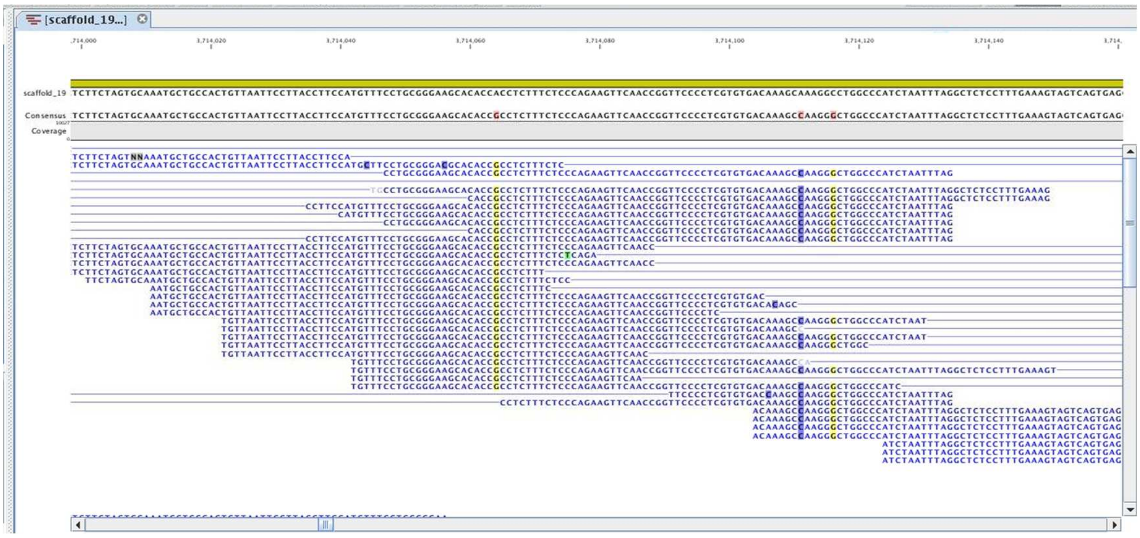
In this introduction we will focus on resequencing of a whole genome and *de novo* assembly.

The advantage of resequencing whole genomes

In the post-genome era, one of the central goals of evolutionary biologists is to understand the evolutionary histories of genetic and regulatory mechanisms that underlie organismal diversity [86]. Much of our information about genome structure, function and evolution depends on sequence data [87]. Using genomic technologies, such as high-throughput Sanger sequencing and DNA microarrays, comparative studies of natural variation at the molecular level have yielded important insights into population histories, as well as into the genetic mechanisms underlying adaptation and speciation. For example, studies of the patterns of nucleotide diversity (namely, studies of variation at the genomic level) have been used to infer the nature of selective forces acting on specific loci [88], as well as to perform genome-wide scans for the signatures of natural selection [74], [90].

At the genomic level, most studies of natural variation focused on populations of species for which a sequenced reference genome was available, because it was nearly inconceivable for single laboratories to sequence entire genomes. Moreover, owing to the cost of traditional sequencing methods, most large-scale studies of genetic diversity (especially in species with large genomes) chose to genotype previously identified polymorphisms rather than discover new mutations by direct sequencing. The genotype approach typically relies on the assumption that for many questions it is sufficient to genotype only a subset of variants (e.g. because un-typed variation is in linkage disequilibrium with typed variation [91]). However, even in cases where this assumption is mostly valid, these studies inevitably result in a low-resolution description of genetic variation (e.g. in the case of mapping the genetic basis for a trait, such as susceptibility to a disease, genotyping studies usually result in the identification of a genomic region associated with the trait, not a specific functional variant). By contrast, with the development of next-generation sequencing technologies, complete surveys of all genetic variation in a large number of individuals have become feasible.

The availability of an increasing range of high-quality reference genome sequences for different species provides a new opportunity to study genetic variation on an unprecedented scale [87]. Whole genomes, regions, or genes can be re-sequenced in multiple individuals. The sequence data from each individual are aligned to the appropriate reference, and the genetic variants between the different samples can be detected as high-confidence sequence differences [84] (Figure 8).



```
scaffold_19 TCTTCTAGTGCAAATGCTGCCACTGTTAATTCCTTACCTTCCATGTTTCCTGCGGGAAGCACACCACCTCTTTCTCCCAGAAGTTCAACCGGTTCCCCTCGTGTGACAAAGCAAAGGCCTGGCCCATCTAATTTAGGCTCTCCTTTGAAAGTAGTCAGTGAG
Consensus   TCTTCTAGTGCAAATGCTGCCACTGTTAATTCCTTACCTTCCATGTTTCCTGCGGGAAGCACACCGCCTCTTTCTCCCAGAAGTTCAACCGGTTCCCCTCGTGTGACAAAGCCAAGGGCTGGCCCATCTAATTTAGGCTCTCCTTTGAAAGTAGTCAGTGAG
```

Figure 8: CLC Genomics Workbench screenshot.

The variants between genotypes can be identified either on a genome-wide scale or by comparison to the reference genotype [85]. Thus, next-generation sequencing approaches have the potential to allow one to work on any species, collect genome-wide natural variation data at unprecedented resolution, and provide considerable additional insight into the mechanisms of regulatory evolution [86].

De novo assembly

The development of automated sequencing technologies has revolutionized biological research by allowing scientists to decode the genomes of many organisms. NGS technologies can accelerate the pace at which we explore the natural world, yet pose new challenges to the software tools used to reconstruct genetic information from the raw data produced by sequencing machines [92]. Despite a dramatic increase in the number of complete genome sequences available in public databases, the vast majority of the biological diversity in our world remains unexplored. Combining NGS technologies and *de novo* assembly approach, the scientific community could potentially investigate the structural complexity of all species genomes. Nowadays *de novo* assembly of NGS data require the development of new software tools that can overcome the technical limitations of these technologies. Indeed, the main limitation is a rapid deterioration in assembly quality as the read length decreases.

De novo assembly problem

De novo genome assembly is often likened to solving a large jigsaw puzzle without knowing the picture we are trying to reconstruct. Repetitive DNA segments correspond to similarly colored pieces in this puzzle (e.g. sky) that further complicate the reconstruction. Mathematically, the *de novo* assembly problem is difficult irrespective of the sequencing technology used, falling in the class of NP-hard problems [93], computational problems for which no exact solution is known. Repeats are the primary source of this complexity, specifically repetitive segments longer than the length of a read. An assembler must either guess (more often incorrectly) the correct genome from among a large number of alternatives (a number that grows exponentially with the number of repeats in the genome) or restrict itself to assembling only the nonrepetitive segments of the genome, thereby producing a fragmented assembly.

As genome sequencing technology has evolved, methods for assembling genomes have changed with it. Genome sequencers have never been able to read more than a relatively short stretch of DNA at once, with read lengths gradually increasing over time. Reconstructing a complete genome from a set of reads requires an assembly program, and a variety of genome assemblers have been used for this task. Redundant coverage, in which on average every nucleotide is sequenced many times over, is required to produce a high-quality assembly. Another benefit of redundancy is greatly increased accuracy compared with a single read: where a single read might have an error rate of 1%, eight fold coverage has an error rate as low as 10^{-16} when eight high-quality reads agree with one another. High coverage is also necessary to sequence polymorphic alleles within diploid or polyploid genomes [94]. Next-generation sequencing (NGS) technologies produce read lengths ranging from 35 to 400 bp, at far greater speed and much lower cost than Sanger sequencing. However, as reads get shorter, coverage needs to increase to compensate for the decreased connectivity and produce a comparable assembly. Certain problems cannot be overcome by deeper coverage: if a repetitive sequence is longer than a read, then coverage alone will never compensate, and all copies of that sequence will produce gaps in the assembly. These gaps can be spanned by paired reads consisting of two reads generated from a single fragment of DNA and separated by a known distance as long as the pair separation distance is longer than the repeat.

Today, thanks to changes in sequencing technology, a major question confronting genome projects is: can we sequence a large genome (>100Mbp) using short reads? If so, what are the limitations on read length, coverage, and error rates? How much paired-end sequencing is necessary? And what will the assembly look like?

Assembly method

Current genome sequencing technology can only sequence a tiny portion of a genome in a contiguous read. Nevertheless, just as a jigsaw puzzle can be assembled from small puzzle pieces, a complete genome sequence can be assembled from short reads. Unlike jigsaw puzzle pieces that precisely lock together, DNA sequence reads may fit together in more than one way because of repetitive sequences within the genome. Assembly methods aim to create the most complete reconstruction possible without introducing errors. The central challenge of genome assembly is resolving repetitive sequences. The magnitude of the challenge depends on the sequencing technology, because the fraction of repetitive reads depends on the length of reads themselves. At one extreme, if the reads were just one base long, every read would be repetitive; at the other extreme, if we could simply read an entire chromosome from one end to the other, repeats would pose no problem at all. In between these extremes, the fraction of unique sequences increases as the read length increases, until eventually every sequence in the genome is unique. If DNA sequences were random (which they are not), then the expected number of occurrences of any sequence would decrease exponentially as the length of the sequence increases, and a modest increase in read length could dramatically reduce the number of repeats in the genome. However, real genomes have complicated repeat structures making some sequences nearly impossible to assemble correctly.

Early genome assemblers used a simple greedy algorithm, in which all pairs of reads are compared with each other, and the ones that overlap most are merged first. To allow for sequencing errors, assemblers compute these overlaps with a variant of the Smith-Waterman algorithm [95], which allows for a small number of differences in the overlapping sequence, typically 1%-10%. Once all overlaps are computed, the reads with the longest overlap are concatenated to form a contig (contiguous sequence). The process then repeats, each time merging the sequences with the longest overlap until all overlaps are used. This simple merging process will accurately reconstruct the simplest genomes, but fails for repetitive sequences longer than the read length. The greedy algorithm will assemble all copies of a repeat into a single instance, because all reads with the repetitive sequence overlap equally well. The problem is that the greedy algorithm cannot tell how to connect the unique sequences on either end of a repeat, and it can easily assemble together distant portions of the genome into misassembled, chimeric contigs. Beginning in the 1990s, assembly of bacterial genomes required development of more sophisticated methods to handle repetitive sequences. Assembly of large eukaryotic genomes required further innovations, not only in the handling of repeats, but also in the computational requirements for memory and processing time [94].

Scaffolding

The scaffolding phase of assembly focuses on resolving repeats by linking the initial contigs into scaffolds, guided by mate-pair data. Mate pairs constrain the separation distance and the orientation of contigs containing mated reads. A scaffold is a collection of contigs linked by mate pairs, in which the gaps between contigs may represent either repeats, in which case the gap can in theory be filled with one or more copies of the repeat,or true gaps in which the original sequencing project did not capture the sequence needed to fill the gap. If the mate pair distances are long enough, they permit the assembler to link contigs across almost all repeats. Assemblers vary in their strategies for calling a contig repetitive, but most of them rely on some combination of the length of the contig and the number of reads it contains. If a contig contains too many reads, then it is flagged as a repeat. High copy-number repeats are easy to identify, because the coverage statistics make it obvious that they are repetitive; in contrast, two-copy repeats are the most difficult to identify using statistical methods. After flagging repeats, an assembler can build scaffolds by connecting unique contigs using mate-pair links. If the contigs in a scaffold overlap, the assembler can merge them at this point. Otherwise, the assembler will record a gap of approximately known size within the scaffold. Assemblers can also include repetitive contigs in these scaffolds, as long as the repeats are connected by mate pairs to unique contigs [94].

Short read assembly

In principle, assemblers created for long reads should also function for short reads. The principles of detecting overlap and building contigs are no different. In practice, initial attempts to use existing assemblers with very short reads either failed or performed very poorly, for a variety of reasons. Some of these failures were mundane: for example, assemblers impose a minimum read length, or they require a minimum amount of overlap that is too long for a short-read sequencing project. Other failures are caused by more fundamental problems [94].

The computation of overlaps is one of the most critical steps in any assembly algorithm. Short-read sequencing projects require that this step be redesigned to make it computationally feasible, especially since many more short reads than long reads are needed to achieve the same level of coverage. As such, the number of overlaps to compute will increase, and any per-read or per-overlap overhead will be greatly magnified. This problem is exacerbated by the fact that short-read projects compensate for read length by obtaining deeper coverage, and it is not unusual to see NGS projects at 30X, 40X, or 50X coverage rather than the 8X coverage that is typical of Sanger sequencing projects.

A new generation of genome assemblers has been developed specifically to address the challenges of assembling very short reads. These assemblers include for exemple ABySS [23] and CLC Genomics Workbench (www.clcbio.com/genomics). Rather than using an overlap graph, all of these assemblers use a de Bruijn graph algorithm [24]. In this approach, the reads are decomposed into k-mers that in turn become the nodes of a de Bruijn graph. A directed edge between nodes indicates that the k-mers on those nodes occur consecutively in one or more reads. These k-mers take the place of the seeds used for overlap computation in other assemblers (Fig. 2).

Parameters to be considered in a *de novo* assembly

There is a direct and dramatic tradeoff among read length, coverage, and expected contig length in a genome assembly. Further complicating any modeling strategy, next-generation sequencing methods have sequence-dependent coverage biases and nonuniform error rates [99]. These sequencing irregularities will cause unexpectedly low coverage regions (e.g., Illumina sequencers have lower coverage in low-GC regions) and consequently more gaps in an assembly. Fortunately, many of these limitations can be overcome by additional oversampling of the genome to boost the low coverage regions.

An important parameter to estimate an assembly is the N50. An N50 contig size of N means that 50% of the assembled bases are contained in contigs of length N or larger. N50 sizes are often used as a measure of assembly quality because they capture how much of the genome is covered by relatively large contigs.

An important side note here is that coverage cannot be computed precisely based on the number of reads generated, because all NGS technologies have a non negligible failure rate. This is best illustrated by resequencing projects, in which it is typical to find only 70%-75% of the reads mapping onto the genome. The remaining 25%-30% of the reads fail to map primarily due to low quality.

In the ideal case, the quality of an assembly will be determined by the read lengths, mate-pair distances, and by the repeat structure of the genome. In general, longer reads make better assemblies because they span more repeats. Similarly, longer insert sizes (mate-pair distances) will increase scaffold sizes, but longer inserts will not always improve contig sizes. For an assembler to close a gap within a scaffold, it must find a set of reads that form an unambiguous path between the flanking contigs. With large gaps, multiple alternative paths through the overlap or de Bruijn graph are much more likely. For this and other reasons, using a mixture of insert sizes can be very effective. The shortest inserts are used to resolve the small repeats, and longer inserts can resolve progressively longer repeats. In practice, long inserts tend to be less reliable, with a much higher variance in their length distribution [94].

The keys to good assembly results include deep coverage by reads with lengths longer than common repeats, and paired-end reads from short (0.53 kb) and long (>3 kb) DNA fragments. To obtain large scaffolds and fill in repeat-induced gaps, a sequencing project should also generate a large set of reliable paired-end reads. As long as both ends of a pair map uniquely to contigs, the pair can be used for scaffolding, and, to fill in scaffold gaps, we need paired reads in which one read is anchored in a contig and its mate falls in the gap.

More important than the read length of paired reads, however, is the number of distinct, nonchimeric pairs produced. Protocols to generate paired reads are still being refined, and we have seen sequencing runs that suffered from having very few distinct pairs in them, from having numerous redundant pairs (the same pairs occurred repeatedly), and from having chimeric pairs (the paired sequences were not at the expected separation and orientation in the genome).

Introducing the pan-genome concept in plants

With the advent of the NGS re-sequencing technologies, we have enteredan exciting era in which we can finally learn what differences are found among individuals within a species at the DNA sequence level [100]. Transposable element movement is largely responsible for variation both in intergenic region sequence content and in local genic content. Both class I (long terminal repeat [LTR]-retrotransposons) and class II (DNA transposons of different superfamilies) transposons contribute to sometimes dramatic differences in local sequence content among individuals belonging to the same species. A comparison of four randomly chosen genomic regions between the maize inbred lines B73 and Mo17 revealed that, on average, only 50% of the sequences are shared. Approximately 25% of the sequences were observed in a homologous location in one of the inbred lines but not in the other [101].

Similar, but less dramatic, differences have been observed in rice [102] and barley [103]. These observations have prompted us to borrow the concept of the pan-genome, which has been proposed for bacterial species [104], to describe the developing view of genomic variation within plant species.

Indeed, the comparison of the genomic sequences of eight strains of the bacterial species *Streptococcus agalactiae* [104] revealed that a bacterial species can be best described by its "pan-genome". The pan-genome includes a core genome containing genes that are present in all strains and a dispensable genome composed of partially shared and strain-specific DNA sequence elements. Unique genes were detected in each of the eight sequenced genomes, and mathematical modelling indicates that new genes will still be found after sequencing many more strains. Thus, the genomes of multiple, independent isolates are required to understand the global complexity of bacterial species.

In the two previously mentioned inbred lines of maize, taking the estimates from the cited paper [101], the pan-genome would comprise a core genome representing the 50% of the genome that is shared between the two lines (corresponding to a size of 1.67 Gb, if we assume an approximate total genome size for each of the lines of 2.50 Gb) and a dispensable genome of the same total size that is equally distributed among the two lines. The core genome comprises both single-copy sequences (including most if not all genes) and transposable elements that are found among all individuals in a certain genomic location.The dispensable genome is made up mostly of transposable elements of different types that, although present in multiple copies in each individual, can be found in a specific location only in some of them. A gene-like fraction can also be found in the dispensable maize genome [100].

Considering the previous works, we would like to introduce the pan-genome concept between the two closely-related species of *Populus*, *nigra* and *trichocarpa.* We require to find a core genome composed of shared genes between the two species and a dispensable genome for both species composed of partially shared and species-specific DNA sequence elements.

Bibliography

[1] Jansson S. and Douglas C. *Populus*: a model system for plant biology. *Annual review of plant biology* **58**, 435-58 (2007).

[2] Tuskan, G.A., Difazio, S., Jansson, S. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313** 1596-604 (2006).

[3] Taylor, G. *Populus: Arabidopsis* for Forestry. Do We Need a Model Tree?, *Annals of Botany* **90**, 681-689 (2002).

[4] Allona, I. *et al.* Analysis of xylem formation in pine by cDNA sequencing. *Proceedings of the National Academy of Sciences of the USA **95**, 9693-9698 (1998).*

[5] Koornneef, M., Alonso-Blanco, C., Peeters A.J., Soppe, W. Genetic control of flowering time in *Arabidopsis. Annu. Rev. Plant Physiol. Plant Mol. Biol.*, **49**, 34570 (1998).

[6] Bohlenius, H., Huang, T., Charbonnel-Campaa, L., Brunner, A.M., Jansson S., *et al.* CO/FT regulatory module controls timing of flowering and seasonal growth cessation in trees. *Science* **312**, 104043 (2006).

[7] Braatne, J.H., Rood, S.B., Heilman, P.E. Life history, ecology, and conservation of riparian cottonwoods in North America. In: Stettler, R.F., Bradshaw, H.D., Heiman, P.E., Hinckley, T.M. (eds) *Biology of Populus.* NRC Research Press, Ottawa, 5780 (1996).

[8] Guilloy-Froget, H., Muller, E., Barsoum, N., Hughes, F.M.R. Dispersal, germination, and survival of *Populus nigra* L. (*Salicaceae*) in changing hydrologic conditions. *Wetlands* **22**, 478488 (2002).

[9] Legionnet, A., Muranty, H., Lefevre, F. Genetic variation of the riparian pioneer tree species *Populus nigra.* II. Variation in susceptibility to the foliar rust Melampsora larici-populina. *Heredity* **82**, 318327 (1999).

[10] Rathmacher, G., Niggemann, M., Khnen, M., Ziegenhagen, B., Bialozyt, R. Short-distance gene flow in *Populus nigra* L. accounts for small-scale spatial genetic structures: implications for in situ conservation measures. *Conservation Genetics* **11**, 1327-1338 (2009).

[11] FAO. Poplars and willows in wood production and land use. *FAO Forest Series* No. 10, (1979).

[12] Vanden Broeck, A., Storme, V., Cottrell, J.E., Boerjan, W., Van Bockstaele, E., Quataert, P., Van Slycken, J. Gene flow between cultivated poplars and native black poplar (*Populus nigra* L.): a case study along the river Meuse on the Dutch-Belgian border. *For Ecol Manag* **197**, 307310 (2004).

[13] Smulders, M.J.M., Beringen, R., Volosyanchuk. R., Broeck, A.V., van der Schoot. J., Arens, P., Vosman, B. Natural hybridisation between *Populus nigra* L. and *P. x canadensis* Moench. Hybrid offspring competes for niches along the Rhine River in the Netherlands. *Tree Genet Genomes* **4**, 663675 (2008).

[14] Ziegenhagen, B., Gneuss, S., Rathmacher, G., Leyer, I., Bialozyt, R., Heinze, B., Liepelt, S. A fast and simple genetic survey reveals the spread of poplar hybrids at a natural Elbe river site. *Conserv Genet* **9**, 373379 (2008).

[15] Jump, A.S., Penuelas, J. Genetic effects of chronic habitat fragmentation in a wind-pollinated tree. *Proc Natl Acad Sci USA* **103**, 80968100 (2006).

[16] Hardy, O.J., Vekemans, X. Isolation by distance in a continuous population: reconciliation between spatial autocorrelation analysis and population genetics models. *Heredity* **83**, 145154 (1999).

[17] Born, C., Hardy, O.J., Chevallier, M.H., Ossari, S., Atteke, C., Wickings, J., Hossaert-Mckey, M. Small-scale spatial genetic structure in the Central African rainforest tree species *Aucoumea klaineana*: a stepwise approach to infer the impact of limited gene dispersal, population history and habitat fragmentation. *Mol Ecol* **17**, 20412050 (2008).

[18] Primack, R.B. A primer of conservation biology. 2nd edn. Sinauer associates, Inc., Sunderland (2000).

[19] Tabbener, H.E., Cottrell, J.E. The use of PCR based DNA markers to study the paternity of poplar seedlings. *For Ecol Manag* **179**, 363376 (2003).

[20] Pospiskova, M., Salkova, I. Population structure and parentage analysis of black poplar along the Morava River. *Can J For Res* **36**, 10671076 (2006).

[21] Zsuffa, L. The genetics of *Populus nigra. L. Ann. Forestales (Zagreb)* **6/2**, 2953 (1974).

[22] Cottrell, J.E., Krystufek, V., Tabbener, H.E., *et al.* Postglacial migration of *Populus nigra* L.: lessons learnt from chloroplast DNA. *Forest Ecology and Management* **206**, 71-90 (2005).

[23] Vanden Broeck, A. Technical guidelines for genetic conservation and use for European black poplar (*Populus nigra*). *EUFORGEN* 6 pages (2003).

[24] Bradshaw, H.D., Stettler, R.F. *Theor. Appl. Genet.* **86**, 301 (1993).

[25] Hanley, S.J., Mallott, M.D., Karp, A. Alignment of a *Salix* linkage map to the *Populus* genomic sequence reveals macrosynteny between willow and poplar genomes. *Tree Genet. Genomes* **3**, 35-48 (2006).

[26] Koch, M.A., Haubold, B., Mitchell-Olds, T. *Mol. Biol. Evol.* **17**, 1483 (2000).

[27] Lynch, M., Conery, J.S. *Science* **290**, 1151 (2000).

[28] Sterck, L. *et al. New Phytol.* **167**, 165 (2005).

[29] Dode, L.A. *Bull. Soc. Hist. Nat. Autun* **18**, 161 (1905).

[30] Collinson, M.E., *Proc. R. Soc. Edinburgh B Bio. Sci.* **98**, 155 (1992).

[31] Eckenwalder, J.E. in *Biology of Populus and Its Implications for Management and Conservation* Stettler, R.F., Bradshaw H.D.Jr., Heilman, P.E., Hinckley, T.M.Eds. (NRC Research Press, Ottawa, 1996), chap. 1.

[32] Garcia-Gil, M.R., Mikkonen, M., Savolainen, O. Nucleotide diversity at two phytochrome loci along a latitudinal cline in Pinus sylvestris. *Mol Ecol* **12**, 1195-1206 (2003).

[33] Howe, G.T., Aitken, S.N., Neale, D.B., Jermstad, K.D., Wheeler, N., Chen, T.H.H. From genotype to phenotype: unraveling the complexities of cold adaptation in forest trees. *Can J Bot* **81**, 1247-1266 (2003).

[34] White, T.L., Adams, W.T., Neale. D.B. Forest Genetics. *CABI Publishing* (2007).

[35] Neale, D.B. Genomics to tree breeding and forest health. *Current Opinion in Genetics & Development* **17**, 539-544 (2007).

[36] Savolainen, O., Pyha, T. Genomic diversity in forest trees. *Current Opinion in Plant Biology* **10**, 162167 (2007).

[37] Neale, D.B., Savolainen, O. Association genetics of complex traits in conifers. *Trends Plant Sci* **9**, 325-330 (2004).

[38] Gonzalez-Martinez, S.C., Krutovsky, K.V., Neale, D.B. Forest tree genomics and adaptive evolution. *New Phytol* **170**, 227-238 (2006).

[39] Neale, D.B., Ingvarsson, P.K. Population , quantitative and comparative genomics of adaptation in forest trees. *Current Opinion in Plant Biology* **11**, 149-155 (2008).

[40] Nielsen, R. Statistical tests of selective neutrality in the age of genomics. *Heredity* **86**, 641-647 (2001).

[41] Nei, M. Molecular evolutionary genetics. *New York: Columbia University Press* (1987).

[42] Schmid, K.J., Ramos-Onsins, S., Ringys-Beckstein, H., Weisshaar, B., Mitchell-Olds, T. A multilocus sequence survey in *Arabidopsis thaliana* reveals a genome-wide departure from a neutral model of DNA sequence polymorphism. *Genetics* **169**, 1601-1615 (2005).

[43] Morrell, P.L., Lundy, K.E., Clegg, M.T. Distinct geographic patterns of genetic diversity are maintained in wild barley (*Hordeum vulgare* ssp. *spontaneum*) despite migration. *Proc Natl Acad Sci USA* **100**, 10812-10817 (2003).

[44] Ching, A., Caldwell, K.S., *et al.* SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. *BMC Genetics* **3**, 19 (2002).

[45] Dvornyk, V., Sirvio, A., Mikkonen, M., Savolainen, O. Low nucleotide diversity at the *pal1* locus in the widely distributed *Pinus sylvestris. Molecular Biology and Evolution* **19**, 179-188 (2002).

[46] Heuertz, M., De Paoli, E., Kllman, T., Larsson, H., Jurman, I., Morgante, M., Lascoux, M., Gyllenstrand, N. Multilocus patterns of nucleotide diversity, *linkage disequilibrium* and demographic history of Norway spruce (*Picea abies* (L.) Karst). *Genetics* **174**, 2095-2105 (2006).

[47] Gilchrist, E.J., Haughn, G.W. *et al.* Use of Ecotilling as an efficient SNP discovery tool to survey genetic variation in wild populations of *Populus trichocarpa. Molecular Ecology* **15**, 13671378 (2006).

[48] Ingvarsson, P.K. Nucleotide polymorphism and *linkage disequilibrium* within and among natural populations of European aspen (*Populus tremula* L., Salicaceae). *Genetics* **169**, 945-953 (2005).

[49] Olson, S., Robertson, A.L., *et al.* Nucleotide diversity and linkage disequilibrium in balsam poplar (*Populus balsamifera*). *New Phytologist* **186**, 526-536 (2010).

[50] Zaina, G., Giacomello, S., Morgante, M. Naturally occuring polymorphisms in *Populus nigra* candidate genes for phenology. Conference Proceedings, p. 38, of Forest Ecosystem Genomics and Adaptation Conference (San Lorenzo de El Escorial - Madrid, 9-11 June 2010).

[51] Nordborg, M., Borewitz, J.O., Bergelson, J., Berry, C., Chory, J., Hagenblad, J., Kreitman, M., Maloof, J., Noyes, T., Oefner, P.J. *et al.* The extent of *linkage disequilibrium* in *Arabidopsis thaliana. Nat Genet* **30**, 190-193 (2002).

[52] Nordborg, M., Hu, T.T., Ishino, Y., Juahveri, J., Toomajian, C., Zheng, H., Bakker, E., Calabrese, P., Galdstone, J., Goyal, R. *et al.* The pattern of polymorphism in *Arabidopsis thaliana. PLoS Biol* **3**, 1289-1299 (2005).

[53] Remington, D.L., Thornsberry, J.M., Matsuoka, Y., Wilson, L.M., Whitt, S.R., Doebley, J., Kresovich, S., Goodman, M.M., Buckler, E.S. Structure of *linkage disequilibrium* and phenotypic associations in the maize genome. *Proc Natl Acad Sci USA* **98**, 11479-11484 (2001).

[54] Clauss, M.J., Mitchell-Olds, T. Population genetic structure of *Arabidopsis lyrata* in Europe. *Mol Ecol* **15**, 2753-2766 (2006).

[55] Brown, G.R., Gill, G.P., Kuntz, R.K., Langley, C.H., Neale, D.B.: Nucleotide diversity and linkage disequilibrium in loblolly pine. *Proc Natl Acad Sci USA* **101**, 15255-15260 (2004).

[56] Bouille, M., Bousquet, J. Trans-species shared polymorphisms at orthologous nuclear gene loci among distant species in the conifer *Picea* (Pinaceae): implications for the long-term maintenance of genetic diversity in trees. *Am J Bot* **92**, 63-73 (2005).

[57] Wright, S.I., Gaut, B.S. Molecular population genetics and the search for adaptive evolution in plants. *Mol Biol Evol* **22**, 506-519 (2005).

[58] Biswas, S., Akey, J.M. Genomic insights into positive selection. *Trends Genet* **22**, 437-445 (2006).

[59] Wright, S.I., Bi, I.V., Schroeder, S.G., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S. The effects of artificial selection on the maize genome. *Science* **308**, 1310-1314 (2005).

[60] Toomajian, C., Hu, T.T., Aranzana, M.J., Lister, C., Tang, C., Zheng, H., Zhao, K., Calabrese, P., Dean, C., Nordborg, M. A nonparametric test reveals selection for rapid flowering in the *Arabidopsis genome. PLoS Biology* **4**, e137 (2006).

[61] Li, C., Vihera-Aarnio, A., Puhakainen, T., Junttila, O., Heino, P., Palva, E.T. Ecotype-dependent control of growth, dormancy and freezing tolerance under seasonal changes in *Betula pendula* Roth. *Trees* **17**, 127-132 (2003).

[62] International Human Genome Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931945 (2004).

[63] Metzker, M.L. Sequencing technologies the next generation. *Nature Reviews Genetics* **11**, 31-46 (2010).

[64] Sanger, F. *et al.* Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* **265**, 687695 (1977).

[65] Swerdlow, H., Wu, S.L., Harke, H., Dovichi, N.J. Capillary gel electrophoresis for DNA sequencing. Laser-induced fluorescence detection with the sheath flow cuvette. *J. Chromatogr.* **516**, 6167 (1990).

[66] Hunkapiller, T., Kaiser, R.J., Koop, B.F., Hood, L. Large-scale and automated DNA sequence determination. *Science* **254**, 5967 (1991).

[67] Shendure, J., Ji, H. Next-generation DNA sequencing. *Nature biotechnology* **26**, 1135-45 (2008).

[68] Ewing, B., Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8**, 186194 (1998).

[69] Ewing, B., Hillier, L., Wendl, M.C., Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8**, 175185 (1998).

[70] Mitra, R.D., Church, G.M. In situ localized amplification and contact replication of many individual DNA molecules. *Nucleic Acids Res.* **27**, e34 (1999).

[71] Mardis, E. The impact of next-generation sequencing technology on genetics. *Trends in genetics : TIG* **24**, 133-41 (2008).

[72] Mitra, R.D., Shendure, J., Olejnik, J., Edyta Krzymanska, O., Church, G.M. Fluorescent in situ sequencing on polymerase colonies. *Anal. Biochem. 320*, 5565 (2003).

[73] Dressman, D., Yan, H., Traverso, G., Kinzler, K.W., Vogelstein, B. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proc. Natl. Acad. Sci. USA* **100**, 88178822 (2003).

[74] Adessi, C. *et al.* Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. *Nucleic Acids Res.* **28**, e87 (2000).

[75] Fedurco, M., Romieu, A., Williams, S., Lawrence, I., Turcatti, G. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic Acids Res.* **34**, e22 (2006).

[76] Metzker, M.L. Emerging technologies in DNA sequencing. *Genome Res.* **15**, 17671776 (2005).

[77] Bentley, D.R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 5359 (2008).

[78] Barnes,C., Balasubramanian, S., Liu, X., Swerdlow, H., Milton, J. Labelled nucleotides. US Patent 7,057,026 (2002).

[79] Dohm, J.C., Lottaz, C., Borodina, T., Himmelbauer, H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.* **36**, e105 (2008).

[80] Harismendy, O. *et al.* Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol.* **10**, R32 (2009).

[81] Wang, Z., Gerstein, M., Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Rev. Genet.* **10**, 5763 (2009).

[82] Wold, B., Myers, R.M. Sequence census methods for functional genomics. *Nature Methods* **5**, 1921 (2008).

[83] Petrosino, J. F., Highlander, S., Luna, R.A., Gibbs, R.A., Versalovic, J. Metagenomic pyrosequencing and microbial identification. *Clin. Chem.* **55**, 856866 (2009).

[84] The International SNP Map Working Group. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928-933 (2001).

[85] Varshney, R.K., Nayak, S.N., May, G. D., Jackson, S.A. Next-generation sequencing technologies and their implications for crop genetics and breeding. *Trends in biotechnology* **27**, 522-30 (2009).

[86] Gilad, Y., Pritchard, J.K., Thornton, K. Characterizing natural variation using next-generation sequencing technologies. *Trends in genetics*, 463-471 (2009).

[87] Bentley, D.R. Whole-genome re-sequencing. *Current Opinion in Genetics & Development* **16**, 545-552 (2006).

[88] Kelley, J.L., Swanson, W.J. Positive selection in the human genome: from genome scans to biological significance. *Annu. Rev. Genomics Hum. Genet.* **9**, 143160 (2008).

[89] Clark, A.G. *et al.* Positive selection in the human genome inferred from human-chimp-mouse orthologous gene alignments. *Cold Spring Harb. Symp. Quant. Biol.* **68**, 471477 (2003).

[90] Voight, B.F. *et al.* A map of recent positive selection in the human genome. *PLoS Biol.* **4**, e72 (2006).

[91] McCarthy, M.I. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.* **9**, 356369 (2008).

[92] Pop, M., Salzberg, S.L. Bioinformatics challenges of new sequencing technology. *Trends in genetics* **24**, 142-149 (2008).

[93] Medvedev, P. *et al.* Computability and equivalence of models for sequence assembly. *Lecture Notes Comput. Sci.* **4645**, 289301 (2007).

[94] Schatz, M.C., Delcher, A.L., Salzberg, S.L. Assembly of large genomes using second-generation sequencing. *Genome research* **20**, 1165-73 (2010).

[95] Smith, T.F., Waterman, M,S. Identification of common molecular subsequences. *J Mol Biol* **147**, 195197 (1981).

[96] Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J., Birol, I. ABySS: A parallel assembler for short read sequence data. *Genome Res* **19**, 1117 1123 (2009).

[97] Pevzner, P.A., Tang, H., Waterman, M.S. An Eulerian path approach to DNA fragment assembly. *Proc Natl Acad Sci* **98**, 97489753 (2001).

[98] Myers, E.W. The fragment assembly string graph. *Bioinformatics* **21**, ii79ii85 (2005).

[99] Dohm ,J.C., Lottaz, C., Borodina, T., Himmelbauer, H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res* **36**, e105. doi: 10.1093/nar/gkn425 (2008).

[100] Morgante, M., De Paoli, E., Radovic, S. Transposable elements and the plant pan-genomes. *Current opinion in plant biology* **10**, 149-155 (2007).

[101] Brunner, S., Fengler, K., Morgante, M., Tingey, S., Rafalski, A. Evolution of DNA sequence nonhomologies among maize inbreds. *Plant Cell* **17**, 343-360 (2005).

[102] Yu, J., Wang, J., Lin, W., Li, S., Li, H., Zhou, J., Ni, P., Dong, W., Hu, S., Zeng, C. *et al.* The genomes of *Oryza sativa*: a history of duplications. *PLoS Biol* 3:e38 (2005).

[103] Scherrer, B., Isidore, E., Klein, P., Kim, J.S., Bellec, A., Chalhoub, B., Keller, B., Feuillet, C. Large intraspecific haplotype variability at the Rph7 locus results from rapid and recent divergence in the barley genome. *Plant Cell* **17**, 361-374 (2005).

[104] Tettelin, H., Masignani, V., Cieslewicz, M.J., *et al.* Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial “pan-genome”. *Proc Natl Acad Sci USA* **102**, 13950-13955 (2005).

I

Exploiting next-generation sequencing for genetic variation and population genetic studies across the whole genome of Populus nigra

1
Introduction

The justification for genomics research in forest trees follows paradigms from both agriculture and human medicine [1]. In forestry, genomic discovery will support genetic improvement of tree varieties for solid wood, pulp and paper, biofuels, and biomaterials through integration into traditional breeding approaches in domesticated tree populations [2]. However, forest trees are also found throughout the world in an undomesticated state and are fundamentally important for noncommodity values such as plant and animal biodiversity, carbon sequestration, clean air and water, and human recreation [1].

Molecular markers are widely employed in plant research and plant breeding. During plant breeding, markers are being used for the acceleration of plant selection gains through marker-assisted selection (MAS) on the basis of individual genes or on a genome level through the selection of chromosomal segments [3]. In plant genetic research, molecular markers are also being used for the analysis of population structure, the study of evolutionary relationships, and in sequenced model systems such as *Arabidopsis* for studies of the genetic structure of individuals at the whole-genome level [4]. In recent years, SNP markers have gained much interest in the scientific and breeding community [5]. They occur in virtually unlimited numbers as differences of single nucleotides between individuals and every SNP in single copy DNA is a potentially useful marker. The potential of SNP markers is clearly demonstrated in human genome analysis. On the basis of massive research efforts and the full sequence of the human genome, several million SNP markers [6] have been identified and technologies to analyze large numbers of SNP markers simultaneously (currently up to 1 million) have been developed [3]. With such large marker numbers it has become possible to scan the entire genome at extremely high marker densities for associations of individual markers with specific quantitatively inherited traits which is called whole-genome scanning (WGS), genome-wide association studies (GWAS), or association genetics [7].

Although thousands of SNP markers are widely used in animal and human genome analysis, their use in plants is still in its infancy [3]. There are several SNP discovery techniques that are used for the identification of large numbers of SNPs in a given plant. For example, the identification of large numbers of SNPs can be based on EST sequence data. The problem with this approach is the sequence quality of the ESTs which is usually not very high so that several ESTs from each of the compared lines need to be available for the same gene in order to reliably identify a SNP [3]. As a result of these limitations, the number of identified SNPs is relatively low in many species with validation rates usually between 50 and 85%. Such results have been published for example in maize [8], barley [9], tomato [10], or pine [11]. Another approach toward the identification of SNPs that is also based on available ESTs involves the use of arrays containing oligonucleotides derived from large numbers of genes [3]. Furthermore, such arrays can be used not only for the species from which the ESTs have been derived but also for closely related species with limited sequence diversity in genes as it has been demonstrated through the use of a soybean genome array for the identification of SNPs in cowpea [12]. However, the EST clustering approach has a high false discovery rate of 25-50% since array hybridizations with organisms containing large genomes can create spurious results [3]. Another alternative can be offered by resequencing amplicons. In summary, it involves the development of primers for the amplification of DNA fragments derived from genes, ESTs or other single copy genomic sequences. The amplification products from a number of representative lines are fully sequenced and the corresponding sequences are subsequently compared with one another using sequence alignment tools and bioinformatic pipelines [3]. The advantage of this approach is that the sequence from each investigated individual is determined through double-strand sequencing and SNPs can be identified in a very reliable way with a false discovery rate usually significantly below 5%. Instead, the main disadvantage of this approach is that it requires an enormous effort for the analysis of many genes since for each gene, specific primers have to be developed and usually a larger number of individuals need to be amplified and sequenced. Moreover, heterozygous positions may be difficult to score from direct sequencing of PCR products. Indeed, if the primer is designed where a SNP is present, then only the allele not containing that polymorphism will be sequenced causing the non-detection of the polymorphic site. Probably the best-investigated crop plant in that respect is maize [13, 14]. Recently, high expectations toward the fast identification of many SNPs between individuals or lines have been placed onto the next-generation high-throughput genomic sequencing technologies (454, Illumina, SOLiD). With a throughput in the area of hundreds of millions to several billions of bases per run, these methods should permit the identification of many SNPs in a species at much lower cost [15]. Next-generation sequencing (NGS) data can suffer from high error rates that are due to multiple factors, including base-calling and alignment errors [16]. Moreover, many NGS studies rely on low-coverage sequencing (<5X per site per individual, on average), for which there is high probability that only one of the two chromosomes of a diploid individual has been sampled at a specified site. Under such circumstances, accurate SNP calling and genotype calling are difficult, and there is often considerable uncertainty associated with the results. It is crucial to quantify and account for this uncertainty, as it influences downstream analyses based on the inferred SNPs, such as the identification of rare mutations, the estimation of allele frequencies and association mapping. One method for reducing uncertainty associated with genotype and SNP calling is to sequence target regions deeply (at >20X coverage) [16]. Using currently available sequencing technology, the most cost-effective way to obtain an high sequence coverage, required for reliable SNP identification, is to use paired-end Illumina reads [17].

Forest trees are found in an undomesticated state. This provides extensive experimental opportunity for the study of the relationship between naturally occurring genotypic and phenotypic variation. Having access to populations with little or no human disturbance means that extant populations are the result of natural evolutionary forces and questions pertaining to speciation, adaptation, and demography will not be anthropomorphically confounded. This contrasts with the large number of agricultural species that cannot be found in natural populations and have often been through large domestication bottlenecks [18]. DNA re-sequencing of candidate genes in small panels of individuals to discover SNPs in trees began just a few years ago [19, 70]. A number of studies in conifers and poplars have been published and have been recently reviewed [1, 21]. Two general types of inference can be made from resequencing data. First, are the estimates of the neutral mutation rate based on the amount of synonymous nucleotide diversity. The second type of inference made from resequencing data is departures from neutrality [18]. The availability of the sequence genome of *Populus trichocarpa* [22] provides a new opportunity to study the genetic variation on an unprecedented scale in *Populus nigra*. Indeed, whole genomes, regions, or genes can be resequenced in multiple poplar individuals and the sequence data from each individual can be aligned to the *P. trichocarpa* reference genome sequence, and the genetic variants between the different samples can be detected as high-confidence sequence differences, as investigated in humans [23, 24]. Anyway, exploiting the genome sequence of a species to align sequences of another species it is not always possible.

In this part of the PhD thesis we demonstrated that this approach can be exploited between *P. trichocarpa* and *P. nigra* genome sequences as they are two closely related species. However, using the *P. trichocarpa* genome sequence as reference does not allow to reconstruct the *P. nigra*-specific genomic regions as *P. nigra* reads do not align to any *P. trichocarpa* genome portion preventing from detecting SNPs in those regions. We exploited the Illumina sequencing technology to entirely resequence fiftyone *P. nigra* genotypes and detect the sequence polymorphisms within the *nigra* species and between the two species *trichocarpa* and *nigra*. The genotypes were selected in order to maximize the genetic variation, indeed they cover a vast geographical distribution, ranging from the Netherlands to the South Italy in latitude, and from Spain to Hungary in longitude. Four clones, out of the fiftyone, were resequenced at high coverage (>20). Their sequence information was exploited to demonstrate the feasibility of applying the Illumina sequencing to generate accurate sequence data to infer demographic and evolutionary statistics. First, we performed a comparative analysis between the Illumina and the Sanger base call methods. We chose the Sanger method as it is considered the “gold standard” in terms of sequencing accuracy [25]. Then, we performed some population genetic analyses in the *P. nigra* species (i.e. nucleotide diversity estimation and Tajima’s D) and between the *P. nigra* and the *P. trichocarpa* species (divergence level). The remaining 47 clones were resequenced at low coverage (< 20). In 42 genotypes we studied the distribution of both homozygous and heterozygous SNPs within and outside the gene regions, and the SNP frequency in different genic compartments. Finally, the sequence data, belonging both to the four high coverage clones and to the 47 low coverage clones, were aligned *versus* a *P. nigra consensus* sequence to maximize the sequence alignments and detect highly informative *P. nigra* SNPs. The *consensus* sequence was obtained from the sequence alignment of one high coverage clone against the *P. trichocarpa* genome sequence. An Illumina Infinium iSelect HD Custom BeadChip (12000 beadtypes) was designed exploiting the collected SNP information. The SNP chip will provide markers to be used in marker assisted selection (MAS) studies to improve the breeding programs in black poplar in the frame of three large European projects, NovelTree and EnergyPoplar.

2
Materials and Methods

2.1 Sanger resequencing

Plant material

Fourty-eight naturally occurring *Populus nigra* plants were included in the study. The black poplar (*P. nigra*) plants (Table 2.1) were obtained from the Unité Amélioration Génétique et Physiologie Forestiéres - I.N.R.A. (Orleans, France) and the DiSAFRi - University of Tuscia (Viterbo, Italy). All the plants are different genotypes and originate from some different and geographically distant populations. Bud or leaf tissue from field-grown clones was collected for DNA extraction.

DNA extraction

Leaf material (fresh or frozen at -80 ℃) was ground into a fine powder using mortar and pestle in the presence of liquid nitrogen. The genomic DNA was then extracted using the DNeasy PLANT KIT Qiagen (Inc. Valencia, CA).

Gene sequences and primer design

A total of 18 genes, specific for the phenology pathway, were considered in the present study. Each gene was reconstructed by the amplification of overlapping fragments. Sequence-specific primer pairs for PCR (Table 2.2) were designed on *Populus trichocarpa* genome sequence [22] using PRIMER3 software (http://frodo.wi.mit.edu/primer3/input.htm). The expected size of the amplification fragments was chosen to range from 400 to 800 bp. In order to promote the amplification of the selected regions, primer design was targeted to exon sequences when possible.

Table 2.1: Discovery panel of *P. nigra* genotypes screened for sequence variation in eighteen candidate genes resequenced using the Sanger method.

Genotype ID	Country	River system	Latitude (N°)	Longitude (E°)
NL1217	The Netherlands	Rhine (IJssel)	52°33'	5°55'
NL1238	The Netherlands	Rhine (IJssel)	51°54'	6°15'
NL1328	The Netherlands	Rhine (IJssel)	52°31'	6°14'
NL1329	The Netherlands	Rhine (IJssel)	52°31'	6°14'
NL1421	The Netherlands	Rhine (IJssel)	51°58'	5°42'
NL1797	The Netherlands	Rhine (Waal/Maas)	51°48'	5°12'
NL1513	The Netherlands	Rhine (Waal/Maas)	51°51'	5°14'
NL1684	The Netherlands	Rhine (Waal/Maas)	51°49'	5°23'
NL1737	The Netherlands	Rhine (Waal/Maas)	51°52'	5°00'
NL2051	The Netherlands	Unknown	Unknown	Unknown
Ginsheim1	Germany	Rhine	Unknown	Unknown
Ginsheim3	Germany	Rhine	Unknown	Unknown
NVHOF2/19	Germany	Rhine	49°49'	8°30'
NVHOF3/11	Germany	Rhine	49°49'	8°30'
NVHOF3/17	Germany	Rhine	49°49'	8°30'
NVHOF3/5	Germany	Rhine	49°49'	8°30'
GG 404A/1	Germany	Rhine	49°49'	8°25'
GG 501/9	Germany	Rhine	49°49'	8°30'
GG 22D/1	Germany	Rhine	49°49'	8°30'
71041-3-402	France	Arc	45°13'	6°28'
71073-305	France	Unknown	Unknown	Unknown
71077-308	France	Ain	45°54'	5°12'
92510-1	France	Loire EstOrl	47°28'	2°54'
BDG	France	Garonne	44°04'	0°54'
SRZ	France	Lot	44°28'	1°26'
1-A05	France	Drôme	44°41'	5°24'
1-A10	France	Drôme	44°41'	5°24'
6-A06	France	Drôme	44°45'	4°55'
6-A23	France	Drôme	44°45'	4°55'
POLI	Italy	Sinni	40°00'	16°00'
58-861	Italy	Dora	45°00'	7°00'
SN11	Italy	Ticino (right side)	45°12'	9°04'
SN21	Italy	Ticino (right side)	45°12'	9°04'
SN26	Italy	Ticino (right side)	45°12'	9°04'
SN40	Italy	Ticino (right side)	45°12'	9°04'
N11	Italy	Ticino (left side)	45°16'	8°59'
N17	Italy	Ticino (left side)	45°16'	8°59'
N47	Italy	Ticino (left side)	45°16'	8°59'
N62	Italy	Ticino (left side)	45°16'	8°59'
C1	Spain	Ebro	41°55'	1°21'
C12	Spain	Ebro	41°55'	1°21'
C2	Spain	Ebro	41°55'	1°21'
C6	Spain	Ebro	41°55'	1°21'
ag1	Spain	Ebro	41°35'	0°44'
BEN3	Spain	Ebro	41°35'	0°44'
CART2	Spain	Ebro	41°35'	0°44'
FR7	Spain	Ebro	41°35'	0°44'
MEJS4	Spain	Ebro	41°35'	0°44'

Gene	Database ID	Sense primer (5′ to 3′)	Database ID	Antisense primer (5′ to 3′)	Tm (℃)
Phytocrome A	PN700	TCATAACTGGGTTCCGCTTC	PN701	GCTAGCAAGCATCCAAAAGG	57
	PN702	TGGATATGATAGGGCGATGG	PN703	TCCAAGGCAAACTCCTTGTC	57
	PN704	ATGAACCTGGGGAGAAGGAT	PN705	CCGAGCAAAGAAACCAAAAG	57
	PN706	GAGTTTACCCTGCGTGAGGT	PN707	TCAATTGGCAAATGCGTAAA	57
Phytocrome B1	PN720	CGCAATTTTGCATAGGATTGA	PN721	CTGTCCTTGGTAATAAAGAGCTG	58
	PN722	TCAGAGAAACATGTCTTGAGGA	PN723	CATGTAAAACTTATGAGTGAGCCAAT	59
	PN724	TGACCATTGAGATGAGTCCA	PN725	AACATCCTGACCGACAAAGC	57
	PN726	CTGATGCAGGTGAGGAGGAT	PN727	GGGAATGAAACTGCATGAGAA	59
	PN728	GAGAGCATCGAAAATGGGTTA	PN729	TCCATATTGATCCTTAAATTTTCTCA	57
	PN730	AAAACGCTGGTGGTTTATGG	PN731	AATCTTCCTGCACATGCTGA	59
	PN732	CCTGAATTAGTTCAAGACATGTTCC	PN733	TTACGCTAGTTGTACACGAAGAAA	57
	PN734	TTTCCTTCCCCCTTGTCTCT	PN735	TGTGAAATCGAACGAACTGC	57
	PN923	CCAAGAATTCCGGGAAGC	PN924	TTTTATTTGGGCTTCGGTTG	58
Phytocrome B2	PN712	TAGCAGGGGCTGTTCAGTCT	PN713	AAAGACTTCCCCATGGCTTC	57
	PN711	CACAAGCCGTCCGTATACCT			
	PN714	TGGTTCACGATCTTGTTTATAAGG	PN715	TTGGTGTTCAGAGCAAAAGG	57
	PN716	TGGTGGTAAATGCTTGTTCG	PN717	AAAGAATGGTACCAAAGCAGAAA	57
	PN718	TGCTATTGCTGAGGGAAAGAA	PN719	CGCCGATAAAGAAGCAAAAA	57
	PN824	CCCAAAGGCGACGATATAAA	PN825	TTGCACCAAATGCTTTTTCA	58
	PN955	GCAACCAATTCTAAGGCTGT	PN956	GTATGTTTTCCAAGGTTTATTT	57
	PN957	GATAGATGCCTCACCACAAG	PN958	GACGACTAAAAACAGAGCAC	54
Gigantea I	PN774	TCTGAGAGATGGATTGATGGTC	PN775	GCATGGTTGTTTTTCCCAAA	59
	PN776	TCATGTACTGTTGCCTCAGATTG	PN777	AAACTTCATGAAAAGTAAAAGTCTCA	59
	PN778	CCTCATCTGCAGGCACAAAT	PN779	TATCGGTGATCCAGGGAGAC	59
	PN780	AACTCATTACAATCGCCCAAT	PN781	GTCATTGGATCCACACATGC	59
	PN782	TTCAGCATCCCATGTAGGAA	PN783	TCTCCCGCAGCATATTTACC	59
	PN786	TGGGTCTGGAAAACATCCTC	PN787	CAATGCAGGCTGTGAGAGAA	59
	PN788	CTTCCCAGGAACTGGATGC	PN789	ATTCAGTTTTGCAGGCTTGG	59
	PN790	TCTGGCTGCTGTATGTGCTC	PN791	CCTCATGCTTCCAAGTCACA	59
	PN792	AATAGCATCACCCGAAACTCA	PN793	CAACAAGCATCCCATCTGTG	59
	PN886	CCATGTTCTTGCAGAGCAGTA	PN887	CAAGAATTGTTAAAAAGACAAAGGAA	53
	PN888	GGCAACTGTTGGTTCTGCTA	PN889	CTGTTGTTGGAGGAGGGAGT	58
	PN890	TAGAGCTTGCATCGCGACTA	PN891	AGAAACATGAGCAGCAACCA	55
	PN892	CATCACCCGAAACTCAACCT	PN893	CTGCAACTGCAAATCCAGAC	59
	PN894	GCCTTTCTCATCCAAGTGCT	PN895	TTTGGGTATCAGAATTTATTTCCAT	58
Gigantea II	PN796	GGGCAGATAACTTTTTAATGGCTA	PN797	CCAGGAAGGGAAAAACCAAT	59
	PN798	TGACGTGATCAGGAAATTTGA	PN799	TCATTTCCCCACTACCCATC	59
	PN800	TGGAAATGGAGGAGAAAATTG	PN801	GCATTGGCAAAGAATCGTTA	59
	PN802	TGGTTCATGTTTCATTTCACAAG	PN803	CAAGGAGTTCGCAGTTTTTGT	59
	PN806	GCACCTAAACTTGGATGAAGAGA	PN807	GCTGAGGATGCTTTCCTGAC	59
	PN812	CAACATGAACCTTTGGTTGG	PN813	AAAAAGTGCCTCCAAAATAGC	59
	PN814	AGCCAAGCCTGCAAAACTAA	PN815	CGTATCATGCATCCCAGTCA	59
	PN896	CAGCAAAGAAAGGTCAGTTGAA	PN897	GCCCACAACTTTTACTTGCTC	55
	PN898	CGCATGCAATTTCTGTATGG	PN899	GACACCGATCATGTCTTCCA	55
	PN900	AGCTGGGAGCTATTATGAGGA	PN901	AGCTCGAAGGAGTTCCACAA	58
	PN902	TCTTGAAGCACCGCCATC	PN903	CATGGACATTGCAGTCCCTA	58
	PN904	GCATCGCATCAGAACAGAGA	PN905	ATTTCTTGCGTGCAATAGGG	58
	PN906	CAGCATCTCCAACAAAAGCA	PN907	TGCCATGTGGCATTGAGTAT	58
	PN947	CTTTGGTCAATCTATGGTTATC	PN948	GTTCCAATAGAAGATGGTTTC	55
Cryptochrome 1	PN736	CACCCACTAAATTTGGAGCTGTA	PN737	AATCCACACGAACCTCAAGC	59
	PN738	ACTAGCCTTTCATCCCGACA	PN739	TGTAGCGCTATCAGCCTTCC	59
	PN740	TCAAGTCAATTGGTCTTAGGGAAT	PN741	CCGGCGTACATATTCTCCAT	59
	PN742	GATCAGATGGTCCCAAGCAT	PN743	TTTGGCAAAAGAAGCAGTCA	59
Cryptocrome 2	PN744	CGTGATCTTGATCAACAAAGAA	PN745	ATTGTCCCAAGTGAGCAAGG	59
	PN746	AGGAGCATGGCTGATGAAAT	PN747	CAATGCCAGTTAAAGACTGAAATG	59
	PN748	TGTTTTGCAATTTAGTCTTTCATTC	PN749	TGCTCAGAGCCCAGAAAAAT	59
	PN750	GGTGGAAGCTTGTGTTTCCT	PN751	AAGTTTCTCAACAACGATCTCTCA	59
	PN752	CATGGAGATGGGGAATGAAA	PN753	ACGCATCAAAACCAGTTTCC	59
	PN754	ACCCAGAGGGTGAATACGTG	PN755	CACAGAGGCTATTCACTGTTTATTT	59

Gene	Database ID	Sense primer (5′ to 3′)	Database ID	Antisense primer (5′ to 3′)	Tm (℃)
	PN959	CATTCAATTACACTTTCTCA	PN960	AGAGCCCAGAAAAATAACAA	52
Cryptocrome 3	PN756	GAGGTCCCATAATGAAGTCGAG	PN757	AAGAAAACAAAGCCCAATTCAA	59
	PN760	TTGCAAGTGCCCTATTTTCA	PN761	TTGCAATGACTGATTGTGATGA	59
	PN939	CTTTTGGCAACTGGTTGGAT	PN940	CCATTCAGTTGGCATTCTTG	58
	PN961	GTATGTTTTCTTTGTTTCCT	PN962	CCCCATTCCACCACAGCAAC	52
	PN963	GTAAAAGTTCTTCTTCTTCCA	PN964	CCCCATTCCACCACAGCAAC	54
Cryptocrome 4	PN766	CACCCAATACATTTGGAGCA	PN767	CCGGTAGATTTGACAACCTCA	59
	PN770	GAAAATTTTGTGTCAGCGTGTT	PN771	TTTGATCCCAGCTCAATTCC	61
	PN772	AGAGCAGAGGTGCCAAGAGA	PN773	TCAAGCAAGAAAGCTCAAACC	61
Frigida	PN820	TCCATGGCCATTACTCTCAA	PN821	GCTGCCGGAACTTCTTCTC	58
	PN929	GATCAGCCAGATCTCAAGAAA	PN930	GCGTTGAAGGTGATTGTTGT	58
	PN931	GGGTTTGGGATACCGAAGTT	PN932	AGTTCATGCATGGTGTTGGA	58
	PN933	AAAAGCAAATAATTCACCTGCAA	PN934	TTTGATCAGGCAACCTCACA	58
SOC1	PN828	ACTTTTAAACAAATTGCGTTCC	PN829	TGCTTGTGGCGTTCTCTATG	58
	PN830	AAACCTTTTCTGCTGGGTGT	PN831	TTATCATGTCTACAAGTACACCAAAA	58
	PN834	TGAAATTAGAAATGTTACACACACAA	PN835	GTCCAGCGAGCATGATAGAA	56
	PN840	AAAAACTTGGGCCAACAAAG	PN841	AACAAAAGAGGCCAAAAGCA	56
	PN842	TCCCGATAGACTTGGACTTGA	PN843	GACAATTTTCACGTGGGAGA	58
	PN844	CGTCGTCAACAGTTGCATGT	PN845	CCCAATTTAACCCCCAAGTT	58
	PN908	AAACCTTTTCTGCTGGGTGT	PN909	ACCCTAAGGTGCAAGAACCA	58
	PN910	TTTTTCTCTCCGGGTTCTGA	PN911	TCCAAGTTGGTAAGAAAATACAACA	55
	PN914	GCGCTTTGTCAAATAAGCAA	PN915	TTGTGTGAAACCCAAAACTCTG	58
	PN918	GACTGAACAATGTAGTATCCTTGAA	PN919	GAAAAGTCTAACGATACGTCAAATA	58
	PN953	GCCAACAAAGCACCTATGAC	PN919	GGATTTTCTCTCTCAAACTC	55
Constans 1	PN848	AGCCTGCGGGATACTGTCTA	PN849	CTGTCACCCCCACAACTCTT	59
	PN852	AGGCCTATGCAGAGACCAGA	PN853	TTCACGGCAGAGCCTTAGTT	61
	PN943	TGAAAACCCCACCAAATGTT	PN944	ATACGCGTGCCCAATTGTTA	58
Constans 2	PN854	TGCAATGTGGAATTTGAAACC	PN855	TCAAGCAAGAAAGCTCAAACC	61
	PN920	CGGGTTTGACAGCTATGTTG	PN921	CACAGGCAGTGCAGAGAGAC	58
	PN922	CAACTGCGACCTACCTCAAA	PN921	CACAGGCAGTGCAGAGAGAC	58
PAT1a	PN856	TGCAAGCATCGAAACAACTC	PN857	CAGAACCTCACAGCCAGACA	58
	PN858	AGGTGGTCACCCTTGTTGAG	PN859	AGCCAGTCACCACTGTCCTT	58
PAT1LGI	PN860	TGCAAACATCCCAGAAGAAA	PN861	CTCAGCACCCTCACATGCTA	58
	PN862	GAGGTGTAAGGAGCCTGCTG	PN863	CAGTGGAAGTGAAGGGGATG	58
PAT1LGVI	PN870	AATGCATTGCCGAGGATAAG	PN871	TTGAACCTCACAACCATCCA	58
	PN925	TCGACTTCCAGATTGCACAG	PN926	GGAGCTGAAGCTATTCGGTTT	58
PAT1LGXVI	PN864	GATGGGTTTTGTCTGAAATTGA	PN865	AAGGGGAAAATTAATTCCCTGTA	58
	PN866	TTTGCTGTTTCCACCAATGA	PN867	TCGGCTCTCCAGAGACTGAC	58
	PN868	GGCAGAAATTAAGGGAGTTGG	PN869	TTGTTTCATTCATGGTGGACA	58
	PN876	AGGTATGGAATCGCACCAGT	PN877	GGCTGCCATGTAACCAAACT	58
	PN878	TTCCAAATTGCTCAGGGAAC	PN879	GCACCATTCCCACCTATGAA	58

Table from previous page

PCR amplification

DNA amplifications were performed in a 25 μl volume. The reactions contained 10 ng of genomic DNA, 10 μM of each primer, 2.5 mM of each dNTP, 2% DMSO, 1.25 units AMPLITAQ GOLD (Applied Biosystems, Foster City, CA) and 1X AMPLITAQ GOLD buffer. The reactions were performed in the GENEAMP 9700 PCR system (Applied Biosystems, Foster City, CA), under the following conditions: 95 ℃for 10 min., 40 cycles of 20 sec. at 94 ℃, 30 sec. at Tm ℃and 1 min. and 30 sec. at 72 ℃, followed by a final extension of 10 min. at 72 ℃. PCR products were analysed on agarose gel and purified using Agencourt Ampure magnetic beads (Beckmann Coulter, Fullerton, CA) using a Biomek FX robot (Beckmann Coulter, Fullerton, CA).

DNA sequencing and analysis

Purified PCR products were sequenced directly on both strands using the fragment specific primers and the ABI PRISM Dye Terminator Cycle Sequencing Ready Reaction kit v3.1 (Applied Biosystems, Foster City, CA), then separated on an ABI3730 sequencer (Applied Biosystems, Foster City, CA). The output sequences of each region for all genotypes were trimmed using Lucy software [26] with its default settings in order to mask bad and/or confusing tracts, on the basis of base quality scores (Q phred) calculated by the base caller Phred [27, 28]. The trimmed sequences were aligned and visualised using Phrap and Consed programs [29]. Base changes (SNP) and insertion/deletion polymorphisms (DIP) were identified directly by visual inspection of sequence alignments after the PolyPhred SNP detection [30]. Potentially polymorphic sites were called SNPs whenever Q phred was greater than 40 for each nucleotide for homozygous positions. In the case of heterozygous ones, the reliability of findings was assessed using the PolyPhred ranking system: only the best ranks (1, 2 or 3 out of 6) were considered. Frequencies of polymorphic sites (per 100 bp) were calculated by dividing the total number of polymorphic sites by the length of the DNA sequence examined. Synonymous/replacement changes were estimated using the program DNAsp v5 [31].

The following population genetic analyses were carried out with DNAsp 3.50: nucleotide diversity (calculated as π, the mean number of differences between all pairs of alleles in the sample) and Tajima's D [32].

2.2 Illumina resequencing

Plant material

46 naturally occurring *Populus nigra* plants were included in the study. The black poplar plants (Table 2.3) were obtained from the Unit Amlioration Gntique et Physiologie Forestires - I.N.R.A. (Orleans, France) and the DiSAFRi - University of Tuscia (Viterbo, Italy) and the University of Southampton (Southampton, UK). All the plants represent different genotypes and originate from different and geographically distant populations. Bud or leaf tissue from field-grown clones was collected for DNA extraction. Some of these 46 clones are present in the discovery panel of the 48 *P. nigra* clones listed in Table 2.1 and are indicated with *.

Table 2.3: Discovery panel of *P. nigra* genotypes screened for sequence variation across their whole genome. All the individuals, except 71077-308, BDG, POLI and BEN3, were resequenced at a low coverage (<20). Genotypes indicated with * were used in the SNP frequency analysis.

Genotype ID	Country	River system	Latitude (N°)	Longitude (E°)
NL2051*	The Netherlands	Unknown	Unknown	Unknown
NL1217*	The Netherlands	Rhine (IJssel)	52°33'	5°55'
NL1238*	The Netherlands	Rhine (IJssel)	51°54'	6°15'
NL1329*	The Netherlands	Rhine (IJssel)	52°31'	6°14'
NL1682*	The Netherlands	Rhine (Waal/Maas)	Unknown	Unknown
NL1797	The Netherlands	Rhine (Waal/Maas)	51°48'	5°12'
Ginsheim1	Germany	Rhine	Unknown	Unknown
Ginsheim3	Germany	Rhine	Unknown	Unknown
NVHOF2/19*	Germany	Rhine	49°49'	8°30'
NVHOF3/17	Germany	Rhine	49°49'	8°30'
NVHOF3/5*	Germany	Rhine	49°49'	8°30'
FTNY18*	Hungary	Tisa	46°24'	19°20'
FTNY19*	Hungary	Tisa	46°24'	19°20'
98568-1*	France	Rhine	48°45'	7°58'
99582-1*	France	Loire	47°16'	0°05'
92510-3*	France	Loire	47°28'	2°54'
92520-6*	France	Loire	47°15'	2°58'
92525-25*	France	Loire	46°50'	3°23'
92538*	France	Unknown	46°15'	1°54'
71077-308*	France	Unknown	45°54'	5°12'
71072-501*	France	Rhone	45°35'	5°36'
SRZ*	France	Lot	44°28'	1°26'
1-A10*	France	Drôme	44°41'	5°24'
6-A06*	France	Drôme	44°45'	4°55'
6-A23*	France	Drôme	44°45'	4°55'
6-A31*	France	Drôme	44°45'	4°55'
BDG*	France	Garonne	44°04'	0°54'
VGN*	France	Garonne	Unknown	Unknown
72145-7*	France	Gard	43°59'	4°13'
73193-25*	France	Le Gave de Pau	42°57'	0°04'
BDX-06*	France	Unknown	Unknown	Unknown
CZB-25*	France	Unknown	43°59'	3°27'
N11*	Italy	Ticino (left side)	45°16'	8°59'
N47	Italy	Ticino (left side)	45°16'	8°59'
N38*	Italy	Ticino (right side)	45°12'	9°04'
SN11*	Italy	Ticino (right side)	45°12'	9°04'
SN21*	Italy	Ticino (right side)	45°12'	9°04'
SN26*	Italy	Ticino (right side)	45°12'	9°04'
SN40*	Italy	Ticino (right side)	45°12'	9°04'
58-861*	Italy	Dora	45°00'	7°00'
PG-5*	Italy	Paglia	42°50'	11°45'
PG-13*	Italy	Paglia	42°49'	11°46'
PG-22*	Italy	Paglia	42°47'	11°49'
POLI*	Italy	Sinni	40°00'	16°00'
C1*	Spain	Ebro	41°55'	1°21'
C12*	Spain	Ebro	41°55'	1°21'
C2*	Spain	Ebro	41°55'	1°21'
C6*	Spain	Ebro	41°55'	1°21'
BEN3*	Spain	Ebro	41°35'	0°44'
CART5*	Spain	Ebro	Unknown	Unknown
RIN4*	Spain	Ebro	Unknown	Unknown

Nuclei and DNA extraction

Leaf material (fresh or frozen at -80 ℃) from young leaves was ground into a fine powder using mortar and pestle in the presence of liquid nitrogen. Nuclei were extracted from 5 g of grounded material per preparation, according to [35]. The genomic DNA was then extracted from the nuclei following a modified Doyle&Doyle protocol [36].

Illumina library preparation

Paired-end sequencing libraries were constructed with insert size of about 300-600 bp. Library preparation followed “Illumina Paired-End Sample Preparation” modified protocol.

Illumina sequencing

The sequencing process followed Illumina instruction for Genome Analyzer IIx and HiSeq2000. Runs were performed for 75, 100, 110 and 114 cycles. The genotypes POLI, BEN3, 71077-308 and BDG (see Table 2.3) were resequenced at high coverage (>20X), the remaining 42 clones were resequenced at low coverage (<20X). The fluorescent images were processed to sequences using the Illumina data processing pipeline (v1.5 and v1.7).

Read mapping

Paired-end data were mapped to the reference genome sequence of *P. trichocarpa* v2.0 [22] using the software CLC Genomics Workbench (v4) (http://www.clcbio.com/). Parameters used were set as follow: Length fraction=0.9, Similarity=0.9, Minimum paired-end distance=120 or 250 (depending on the library insert size), Maximum paired-end distance=300 or 800 (depending on the library insert size), Multiple hits ignored.

SNP detection

High coverage individuals

SNP detections were performed using the software CLC Genomics Workbench (v4). Parameters used were set as follow: Window length=11, Maximum number of gaps and mismatches=2, Minimum average quality of surrounding bases=15, Minimum quality of central base=20, Minimum coverage=50% of the average coverage of the alignment, Maximum coverage=200% the average coverage of the alignment , Minimum variant frequency (%)=20, Required minimum variant count=1, Sufficient minimum variant count=the maximum coverage, Non-specific and low-quality matches were ignored during SNP detection, Maximum expected variations=2.

Low coverage individuals

SNP detections were performed using the software CLC Genomics Workbench (v4). Parameters used were set as follow: Window length=11, Maximum number of gaps and mismatches=2, Minimum average quality of surrounding bases=15, Minimum quality of central base=20, Minimum coverage=114 (50% of the average coverage of the alignment), Maximum coverage=454 (200% the average coverage of the alignment), Minimum variant frequency (%)=15, Required minimum variant count=1, Sufficient minimum variant count=454 (the maximum coverage), Non-specific and low-quality matches were ignored during SNP detection, Maximum expected variations=2.

SNP chip

SNP detections were performed using the software CLC Genomics Workbench (v4). Parameters used were set as follow: Window length=11, Maximum number of gaps and mismatches=2, Minimum average quality of surrounding bases=15, Minimum quality of central base=20, Minimum coverage=spanning from 10 to 50% the average coverage of the alignment, Minimum paired-end coverage=10% of the average coverage of the alignment, Maximum coverage=150% of the average coverage of the alignment, Minimum variant frequency (%)=35% for the high coverage clones and 15% for the low coverage clones, Non-specific and low-quality matches were ignored during SNP detection, Maximum expected variations=2.

2.3 SNP validation

Experimental design

To evaluate the Illumina sequencing technology accuracy, we compared it with the ABI3730 Sanger sequencing method. We considered the ABI Sanger method the “gold standard” [37]: its base call was assumed to be correct. The SNP validation was conducted in the four clones resequenced at high coverage using the Illumina technology (71077-308, BDG, BEN3 and POLI) (Table 2.3). These clones were also present in the discovery panel resequenced using the Sanger method (Table 2.1). The validation focused on the 18 candidate genes resequenced using both methods. Specifically, for each genotype and for each gene, only the sites covered by both resequencing methods were considered for the comparison. Moreover, the base call of each genotype covered site was examinated in the *consensus* sequence of the 48 genotype discovery panel resequenced using only the Sanger method. In order to assess the performance of the Illumina sequencing technology, several metrics were defined.

Site classification

The covered sites considered for the comparative analysis were classified in 4 different categories:

- **TRUE POSITIVE** (TP): site called as polymorphic in the Illumina data and confirmed by the Sanger information;
- **TRUE NEGATIVE** (TN): site NOT called as polymorphic in the Illumina data and confirmed by the Sanger information;
- **FALSE POSITIVE** (FP): site called as polymorphic in the Illumina data and NOT confirmed by the Sanger information. Two subcategories of the FP sites were defined:
  - **TRUE-FALSE POSITIVE** (T-FP): the Illumina base call was even not confirmed by the site call in the Sanger DP48 sequence;

  - **FALSE-FALSE POSITIVE** (F-FP): the Illumina base call was confirmed by the site call in the Sanger DP48 sequence;

- **FALSE NEGATIVE** (FN): site not called as polymorphic in the Illumina data but polymorphic in the Sanger information.

Performance metrics

Sensitivity, specificity, accuracy and false discovery rate are variables that are usually used to estimate the effectiveness of a test procedure [38].

Sensitivity

Sensitivity relates to the ability of the method to call known polymorphic sites as such.
Sensitivity = TP / (TP + FN)

Specificity

Specificity relates to the ability of the method to make a correct call at not polymorphic positions.
Specificity = TN / (TN + FP)

False Discovery Rate (FDR)

The False Discovery Rate measures the expected proportion of sites called as polymorphic when they are not polymorphic.
FDR = FP / (FP + TP)

Accuracy

Accuracy measures the proportion of concordant calls between ABI Sanger and Illumina.
Accuracy = (TP + TN) / (TP + TN + FP + FN)

2.4 Population genetic analysis

The following population genetic analyses were carried out: nucleotide diversity across the whole genome (calculated as π, the mean number of differences between all pairs of alleles), Tajima's D [32] within genes and across the genome, total divergence between the *P. nigra* and *P. trichocarpa* genome sequences.

The four genotypes (71077-308, BDG, BEN3 and POLI) (Table 2.3) resequenced at high coverage were considered for these analyses.

SNP detection, for each of the four clones aligned to the *P. trichocarpa* reference sequence, was performed individually. A combined table, containing the overall shared positions (polymorphic and non) among the 4 clones respecting the individual SNP detection coverage parameters, was created.

Nucleotide diversity (π), defined as the average number of nucleotide differences per site in pairwise comparisons among DNA sequences, was computed according to the formula reported in Figure 2.1 [33]:

$$\pi = \sum_{ij} x_i x_j \pi_{ij} = 2 * \sum_{i=1}^{n} \sum_{j=1}^{i-1} x_i x_j \pi_{ij}$$

Figure 2.1: Nucleotide diversity computation.

where xi and xj are the respective frequencies of the ith and jth sequences, π is the number of nucleotide differences per nucleotide site between the ith and jth sequences, and n is the number of sequences in the sample. π values were obtained for each nucleotide locus that showed polymorphism among the eight chromosomes while π values for non-polymorphic sites were set to 0 by definition. Then, to obtain π values across genes or their compartments (i.e. exons [mRNA], introns, 5'-UTR, 3'-UTR, CDS), π values of all the polymorphic sites mapped within those regions were summed up and the sum was divided by the total number of covered bases in the gene or compartment taken into account. When multiple transcript models had been predicted within the same gene locus, only the first model (.1) was taken into account to calculate π for gene compartments.

Departures from neutral molecular evolution [34] were tested using the method of Tajima [32]. The D statistic was computed for all the genes, mRNAs and coding sequences (CDS) separately. When multiple transcript models had been predicted within the same gene locus, only the first model (.1) was taken into account. The D statistic was also computed in chromosome windows of 100,000 bp covered at least for their 30%. Tajima's D of all the polymorphic sites mapped within each window were summed up and the sum was divided by the total number of covered bases in the window. The D statistic was then computed for the genome as the average value within the total window number.

Total divergence between the *P. nigra* and *P. trichocarpa* genome sequences was computed for genomic windows of 100,000 bp covered at least for their 30%. In each window, all the polymorphic positions between the two species, resulting from the comparisons between the overall shared positions (polymorphic and non) among the 4 clones and the corresponding *P. trichocarpa* positions, were summed up and the sum was divided by the total number of covered bases in the window. The *ratio* between nucleotide diversity and corresponding total divergence was computed for each window covered at least for its 30%.

2.5 SNP chip pipeline

An Illumina Infinium iSelect HD Custom BeadChip (attempted beadtype 11999) was designed. A pipeline to extract the suitable SNPs for the chip was developped using the whole set of polymorphisms detected on those *P. nigra* genotypes having paired-end read data.

A targeted region approach was chosen to select SNPs within candidate genomic intervals putatively responsible for a set of traits (for example: phenology, biomass, rust resitance and water stress resistance) on the basis of previous QTL studies. In particular, 15 large QTL/genomic regions (152,315 Mb) and 2,916 gene models (66,534 Mb). A number of 1000 control loci belonging to regions outside the candidate ones were also included in the chip.

The input data for the pipeline were produced as follow:

1. reference assembly of 71077-308 clone *versus* the *P. trichocarpa* genome sequence v2.0 annotation 156 (performed using the CLC software);

2. extraction of a *P. nigra consensus* sequence with gaps (shown as N) from the alignment obtained at the previous step (only first 40 scaffolds were considered);

3. masking the *P. nigra consensus* sequence from step 2 for duplications and repetitions (performed using RepeatScout [34]);

4. alignment of all the low-coverage genotype reads (Table 2.3)(without trimming) to the *P. nigra* consensus got in 2. using a Similarity of 0.95. The same was done for the four high-coverage genotypes separately (performed using CLC);

5. SNP detection on the alignments obtained at the previous step: Maximum coverage=1.5 the average coverage of the alignment, Minimum coverage=spanning from 0.1 to 0.5 the average coverage, Minimum Allele Frequence (MAF)=35% for the high-coverage genotypes and 15% for the low-coverage genotypes, Minimum paired-end coverage=0.1 the average coverage (performed by CLC);

6. using the SNP detected in the whole set of genotypes (51) to create a *P. nigra* consensus with the IUPAC codes representing all the polymorphic positions;

7. DIP detection on the alignments obtained at the step 4: Maximum coverage=1.5 the average coverage, Minimum coverage=0.5 the average coverage, MAF=10%, Minimum paired-end coverage=0.1 the average coverage (by CLC);

8. creation a DIP map of the *P. nigra consensus*.

Then, the pipeline was run on the input data as follows:

1. extract SNPs within the candidate regions (select all SNPs within and outside gene space);

2. select only the SNPs with a coverage $>$ 0.5 the average coverage of the alignment and with a MAF ≥ 0.15 (15%) by using the 47 low-coverage individuals (discovery panel, DP). A SNP was considered if present in the DP and confirmed by at least one high-coverage clone;

3. extract SNPs with 60-bp flanking sequences (provide within each flanking region all the polymorphic sites, if present, in IUPAC code and all the DIPs, if present, as N);

4. select only the SNPs whose flanking sequences do not contain any SNP (if it is not possible, limit the number of polymorphisms in the flanking sequences). No SNPs and/or DIP are allowed within ± 10 bp the target SNP;

5. discard the SNPs within duplicated or repetitive regions (use the 71077-308 *P. nigra consensus*-masked fasta file);

6. discard redundant SNPs (nucleotide loci with the same position) due to putative gene model duplicates in the input list of candidates.

3

Results

Sanger resequencing

Good full-length sequences (UTRs included with two exceptions) were obtained for 18 genes (Table 3.1) resequenced in 48 *P. nigra* genotypes by bidirectional PCR-based Sanger resequencing. The *consensus* sequence of each gene was obtained by the PhredPhrap-based alignment of chromatograms and, in the worst case, was the result of at least 54 aligned chromatograms. The full-length genes successfully resequenced were selected as putative candidates for phenology traits and they were identified based on literature searches with respect to the flowering time pathway of *Arabidopsis thaliana* [40]. The candidate genes analysed were first blasted (BLASTn analysis) through their *consensus* sequence to the NCBI nr database confirming they all belong to the photoreceptors and circadian clock factors responsible for the photoperiod control in *Arabidopsis.* In the hypothesis of a co-linear behaviour of poplar genomes, the candidate genes were then blasted to the *P. trichocarpa* genome sequence at the Phytozome 7.0 website [41] (genome version 2.0) confirming they are single-copy loci, whose coding and non coding portions were assigned on the bases of the annotation version 156. In *P. nigra* a total of 1034 single nucleotide sequence variants, over 87194 bp analysed, were identified (Table 3.1). Annotation of the polymorphisms for each full-length gene is detailed in Table 3.1.

Table 3.1: SNP statistics in the 18 candidate genes.

Gene (ID)	Total bp	Total SNPs	Synonymous SNPs	Non synonymous SNPs
Phytochrome A (PHYA)	5728	48	12	6
Phytochrome B1 (PHYB1)	7993	51	6	12
Phytochrome B2 (PHYB2)	6366	67	14	4
Cryptochrome 1 (CRY1)	3692	30	6	4
Cryptochrome 2 (CRY2)	6553	95	4	5
Cryptochrome 3 (CRY3)	3678	74	17	24
Cryptochrome 4 (CRY4)	4202	39	6	9
Constans-like 1 (CoL1)	3105	57	7	6
Constans-like 2 (CoL2)	2152	22	3	7
Suppressor of Constans 1 (SOC1)	7291	115	2	2
Frigida (FRI)	3463	32	2	4
Gigantea 1 (GI1)	10852	109	14	7
Gigantea 2 (GI2)	9165	179	16	23
PhyA signal transduction factor (PAT1a)	3114	32	6	2
PhyA signal transduction factor-like LGI (PAT1LGI)	1960	19	10	5
PhyA signal transduction factor-like LGVI (PAT1LGVI)	1786	16	9	3
PhyA signal transduction factor-like LGXVI (PAT1LGXVI)	3801	39	5	5
PhyA signal transduction factor (PAT1b)	2293	10	3	6

SNP frequency and nucleotide diversity analysis in 18 candidate genes

SNP frequency was estimated for the 18 candidate genes in different genic domains as reported in Table 3.2. The SNP frequency in the candidate genes resulted to be of 1 SNP every 84 bp. Among the different gene compartments, the highest frequency was found in introns, and the lowest in CDSs, as expected. Intermediate values were found for the UTR regions.

Nucleotide diversity was estimated in each candidate gene sequenced (Figure 3.1), with values usually close to the mean (π=0.0016) with few exceptions of highly variable genes: Gigantea 2 (GI2), Suppressor of Constans (SOC1) and Constans-like 1 (COL1). Nucleotide diversity for synonymous and replacement sites was also computed, as well as for the different genic domains (Table 3.2). π for non coding domains was on average higher than that observed in coding ones, with the highest π value in introns, as expected. Within coding sites, π for the synonymous sites was more than four times higher respect to the nucleotide diversity for replacement ones (Table 3.2).

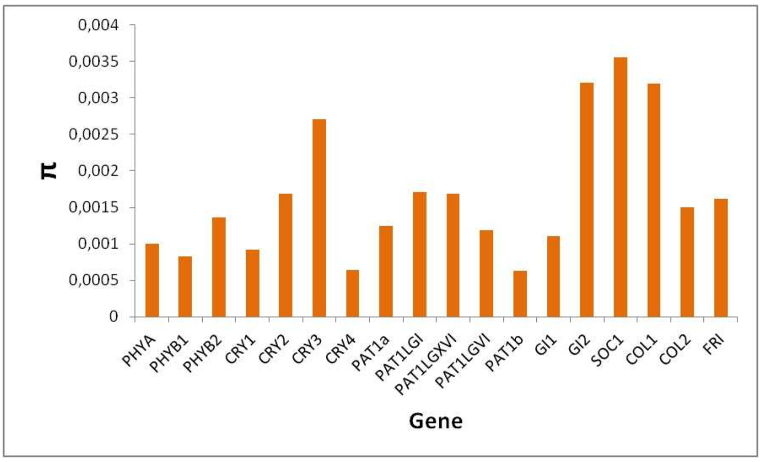


Figure 3.1: Nucleotide diversity (π) distribution in the 18 candidate genes.

Table 3.2: SNP frequency and nucleotide diversity (π) estimated in different compartments of 18 candidate genes.

Compartment	Covered bases	Total bases (%)	Polymorphic sites	Freq.: 1 SNP/... bp	Mean π
Genes	87,194	-	1,034	84	0.0016
mRNAs	52,051	59.70	484	108	0.0013
CDSs	35,030	40.17	276	127	0.0010
Synonymous	-	-	142	-	0.0029
Replacement	-	-	134	-	0.0006
UTRs	17,021	19.52	208	82	0.0019
UTR 5’	8,291	9.51	105	79	0.0023
UTR 3’	8,730	10.01	103	85	0.0016
Introns	35,143	40.30	550	64	0.0022

The minor allele frequency (MAF) for each polymorphic site, of the total 1034, was considered. Its spectrum is presented in Figure 3.2. A clear abundance, about 40%, of rare alleles (MAF<0.05) was detected.

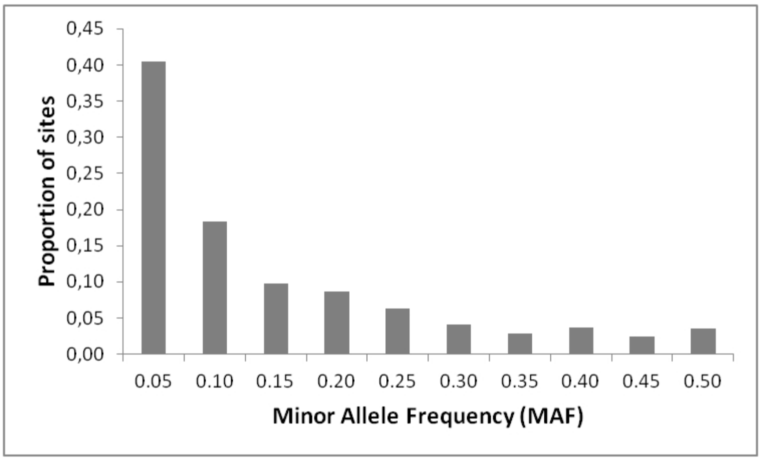


Figure 3.2: MAF distribution for 1034 polymorphic sites in the 18 candidate genes.

Tajima's D analysis

To evaluate the allele frequency distribution of the loci and its possible deviation from neutral expectations, Tajima's D statistics [32] were computed on the 18 *P.nigra* gene sequences. Owing to the relatively small sample size (48 genotypes, N=96) only major deviations from neutrality would be detectable as significant. Tajima's D resulted on average negative (-0.679) (Figure 3.3) and statistically significant only for the Cryptochorme 4 (CRY4). Only one extreme positive value was detected in SOC1 gene, even if still not statistically significant (Figure 3.3). Computation of D within the genic domains (data not shown) resulted in few negative and significant values: the replacement sites of Phytochrome B1, all coding and replacement sites of both the Cryptochrome 3 and 4, all coding and synonymous sites of Frigida.

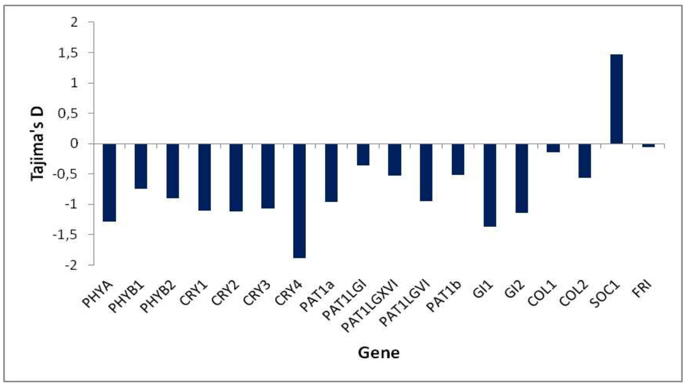


Figure 3.3: Tajima's D distribution in the 18 candidate genes.

Illumina resequencing

We used Illumina next generation DNA sequencing technology to resequence 4 *P. nigra* genotypes (71077-308, BDG, BEN3 and POLI) at a high coverage (>20X) and 47 others at low coverage (< 20X) to detect SNPs across the whole black poplar genome. Read data and relative raw coverage obtained for each genotype are reported in Table 3.3. As shown in Figure 3.4, the majority of the low coverage individuals had a raw coverage spanning between 1 and 10X.

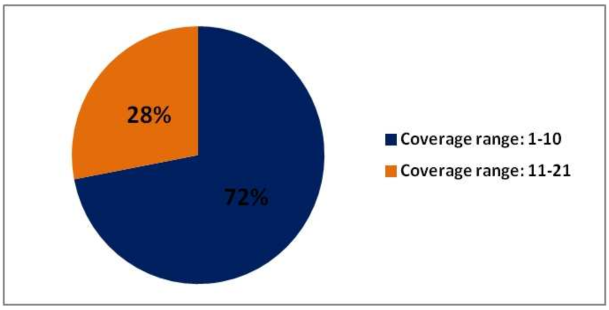


Figure 3.4: Coverage distribution for the 47 low coverage clones.

The sequence data of the *P. nigra* genotypes listed in Table 3.3 were exploited to conduct population genetic analysis at a whole-genome level, calculate the SNP frequency in different genic domains and identify highly informative *P. nigra* SNPs. For the population genetic analysis, 71077-308, BDG, BEN3 and POLI sequences were aligned *versus* the *P. trichocarpa* genome sequence. The same was done to calculate the SNP frequency in these 4 high coverage clones and in 42 low coverage clones (indicated with * in Table 3.3) selected among the 47 for informatic contraints.

The feasibility of a *P. trichocarpa*-*P. nigra* alignment was demonstrated aligning BEN3 reads *versus* two *P. nigra* BAC sequences and *versus* the *P. trichocarpa* BAC corresponding sequence portions. It was shown that in the intraspecific alignment, the BAC sequences were covered for their 98% length, as expected, and in the interspecific alignment, the *P. trichocarpa* BAC corrisponding sequence portions were covered for their 80%. We considered the latter alignment percentage being adequate for our purpose.

Instead, for the identification of highly informative SNPs to design a 12k SNP-chip, within the *P. nigra* species, the sequences of all 51 clones listed in Table 3.3 were aligned *versus* a *P. nigra consensus* sequence to maximise the sequence alignment. In particular, the 4 high coverage clone reads were mapped singularly, whereas the 47 low coverage clone reads were mapped in pool.

High coverage resequencing in 4 *P. nigra* clones to conduct population genetic analysis

SNP frequency, nucleotide diversity and Tajima's D were estimated within the 4 high coverage clones (71077-308, BDG, BEN3 and POLI) on a genome-wide scale.

Paired-end reads of the 4 clones were aligned to the reference genome of *P. trichocarpa* singularly. Statistics concerning the number of mapped base pairs, the percentage of

Table 3.3: *P. nigra* genotypes resequenced exploiting the Illumina technology. Genotypes indicated with * were resequenced at low coverage and used to calculate SNP frequency in different genic domains.

Genotype	Total bp produced	Raw coverage (X)
71077-308	11,614,046,643	27.8
BDG	10,499,784,562	25.1
BEN3	21,882,737,550	52.5
POLI	34,031,232,782	81.6
N47	691,873,200	1.7
NVHOF3/17	878,908,000	2.1
NL1797	910,082,000	2.2
71072-501*	1,020,158,073	2.4
C12*	1,026,605,990	2.5
SN40*	1,195,698,229	2.9
RIN4*	1,224,325,600	2.9
BDX-06*	1,199,931,013	2.9
FTNY18*	1,336,413,883	3.2
58-861*	1,425,822,523	3.4
C6*	1,460,806,904	3.5
N38*	1,540,547,636	3.7
73193-25*	1,647,799,444	4.0
PG13*	1,665,449,401	4.0
N11*	1,676,606,505	4.0
NL2051*	1,826,967,332	4.4
CART5*	1,936,051,399	4.6
NL1682*	2,046,322,170	4.9
PG05*	2,055,865,151	4.9
NL1329*	2,067,806,626	5.0
C1*	2,116,880,335	5.0
C2*	2,160,560,966	5.2
SN26*	2,174,897,241	5.2
Ginsheim1	2,351,224,600	5.6
FTNY19*	2,419,647,905	5.8
NVHOF3/5*	2,475,035,580	5.9
NL1217*	2,543,452,219	6.1
SN11*	2,791,982,335	6.7
98568-1*	2,811,019,907	6.7
NL1238*	3,095,875,836	7.4
Ginsheim3	3,114,417,000	7.5
SN21*	3,183,780,277	7.6
PG22*	3,542,852,254	8.5
CZB-25*	3,885,764,113	9.3
99582-1*	4,749,535,204	11.4
6A31*	4,957,635,050	11.9
NVHOF2/19*	5,638,954,091	13.5
6A23*	5,733,143,633	13.7
VGN*	5,865,971,615	14.0
SRZ*	6,545,172,797	15.7
92510-3*	6,599,547,430	15.8
92520-6*	7,100,652,141	17.0
92525-25*	7,379,085,905	17.7
1A10*	7,616,642,138	18.3
6A06*	8,124,691,652	19.5
72145-7*	8,279,967,553	19.8
92538*	8,874,612,395	21.3

reference covered, the mean alignment coverage and the mean coverage excluding zero coverage regions of the alignment are indicated in Table 3.4 for each clone. Approximately 50-60% input reads, of each clone, were aligned to a unique position in the reference genome and were used for SNP calling. Despite the difference in the number of mapped reads among the individuals, the percentage of reference covered was similar in the 4 clones, spanning between 70 and 79%.

Table 3.4: Alignment statistics of the 4 clones resequenced at high coverage.

Clone	Total bp mapped	Reference covered (%)	Mean coverage	Mean cov. no 0 cov. reg.
POLI	21,388,890,649	79	50.83	64.21
BEN3	7,138,941,598	79	16.96	21.59
BDG	8,430,106,825	77	19.90	25.81
71077-308	5,978,752,683	70	14.20	20.24

SNP detection on the four alignments was performed with stringent parameters regarding minimum and maximum coverage. More than 7 million SNPs were detected both in 71077-308 and POLI (Table 3.5), whereas around 5 million ones were detected in the other two clones. In the total amount of SNPs detected, the majority was composed by homozygous SNPs (Figure 3.5, Table 3.5). Both in homozygous and heterozygous SNP classes, the SNPs were more aboundant in the intergenic regions (Figure 3.6), as expected.

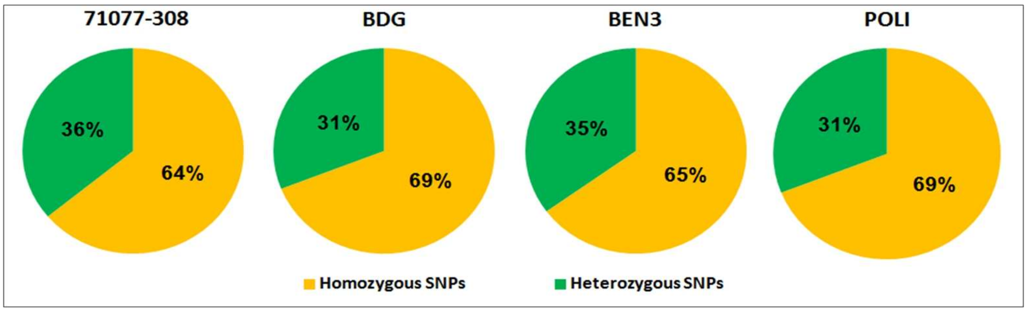


Figure 3.5: Proportion of total homozygous and heterozygous SNPs in the four clones.

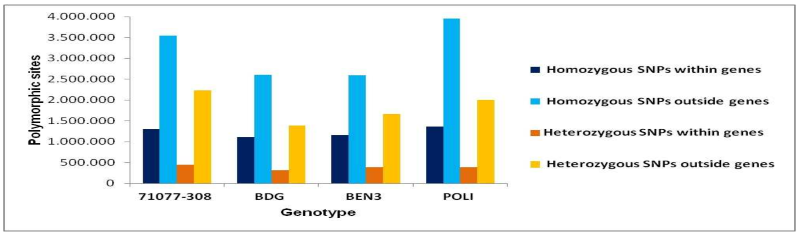


Figure 3.6: Distribution of SNPs within and outside genes for homozygous and heterozygous SNPs.

Table 3.5: SNP statistics of the 4 clones resequenced at high coverage.

Clone	Total SNPs	Homoz.-SNPs within genes	Homoz.-SNPs outside genes	Heteroz.-SNPs within genes	Heteroz.-SNPs outside genes
71077-308	7,512,400	1,296,262	3,545,897	443,572	2,226,669
BDG	5,407,484	1,111,258	2,602,894	312,136	1,381,196
BEN3	5,783,591	1,154,127	2,587,401	383,786	1,658,277
POLI	7,706,023	1,357,523	3,955,369	391,751	2,001,380

SNP validation

To validate the SNP detection based on the Illumina sequencing method, a SNP validation was conducted. The validation was performed on the 18 candidate genes resequenced in the clones 71077-308, BDG, BEN3 and POLI by the Sanger method (see paragraph "Sanger resequencing", pag. 19) using the ABI sequencer and procedure. The ABI Sanger method is considered the "gold standard" in terms of sequencing accuracy [37] and was used as such in this validation. The 18 candidate genes were considered for the validation since their sequences did result to be covered by the high coverage Illumina resequencing effort on the same clones. For each genotype and for each gene, only the sites covered by both resequencing methods were considered for the comparison. Moreover, the base call of each genotype covered site was examined in the *consensus* sequence of the 48-genotype discovery panel resequenced using the Sanger method.

A total of 96,164 sites were analysed. Among these sites, the number of positions correctly called by the Illumina sequencing *versus* the Sanger one were evaluated. In Table 3.6 the number of the different kinds of sites is reported (definitions of the different sites are in the paragraph "Site classification" at pag. 14). Among them, it's interesting the presence of the 43 sites evaluated as false-false positive which were not detected by Sanger in a single clone (thus classified as false positive in a first instance), but resulted to be a polymorphic site in the population data given by the 48-clone discovery panel (thus classified as false-false positive and eventually rescued as site correctly called by Illumina). An explanation for these sites might be the imbalanced amplification of the two alleles for some amplicons, produced by the PCR sample preparation in the bidirectional PCR-based Sanger resequencing method, which results in incorrect genotype calls at variant bases by specifically calling heterozygous sites as homozygous [42]. Imbalanced amplification is usually suspected to result from polymorphisms in or near the oligonucleotide priming sites that result in greater efficiency of amplification for one of the two alleles.

Table 3.6: SNP validation results.

Kind of site	Number
TP	1,186
TN	94,747
FP	141
T-FP	98
F-FP	43
FN	90

Four performance metrics were calculated for the Illumina sequencing technology: sensitivity, specificity, accuracy and false discovery rate (Table 3.7; definitions of the different metrics are in the paragraph "Performance metrics" at pag. 15). Illumina technology proved to be highly accurate (~100%), sensitive at 93% (sensitivity indicates the Illumina ability to make a correct call at polymorphic positions) and specific at 100% (specificity indicates the Illumina ability to make a correct call at not-polymorphic positions). The false discovery rate was evaluated to be 7.6%, when considering as false positive sites only the false-false positives. Possible explanations for the false positive and negative calls might be search in difficult sequence contexts such as repetitive elements, homopolymer stretches, simple repeats and indels.

Table 3.7: Performance metrics of the Illumina sequencing technology.

Performance metric	%
Sensitivity	92.9
Specificity	99.8
Specificity calculated with T-FPs	99.8
Accuracy	99.7
Accuracy calculated with T-FPs	99.8
False discovery rate	10.6
False discovery rate calculated with T-FPs	7.6

Population genetic analysis in the 4 high coverage clones

SNP frequency and nucleotide diversity analysis

SNP frequency and nucleotide diversity (Table 3.8) were estimated in the discovery panel composed by the four high coverage clones considering 2,683,139 polymorphic sites distributed in a resequenced genome sequence of 128,206,359 bp, which resulted from the common nucleotide sites, i.e. those covered by Illumina reads in all the clones. Both the SNP frequency and π were calculated in the same genic domains adopted for the 18 candidate genes plus intergenic regions. The SNP frequency of the intergenic domains (1 SNP every 31 bp) resulted to be higher than all compartments considered, as expected. The total SNP frequency (1 SNP every 48 bp) was estimated to be two times higher than the one within genes. The lowest SNP frequency was observed in the CDS compartment, as expected.

Total nucleotide diversity estimated in the 4 clones (π=0.0071) was observed to have same trends of the SNP frequencies among the different gene domains (Table 3.8), as expected. π for the synonymous sites resulted to be two-thirds of the total one, while π of the replacement sites was one-third of the total one. π at synonymous sites resulted to be double than that at replacement sites.

Table 3.8: SNP frequency and nucleotide diversity (π) in the discovery panel of the 4 high coverage clones.

Compartment	Covered bases	Total bases (%)	Polymorphic sites	Freq.: 1 SNP/... bp	Mean π
Total	128,206,359	-	2,683,139	48	0.0071
Genes	67,564,889	52.70	721,871	94	0.0036
mRNAs	34,018,046	26.53	309,624	110	0.0031
CDSs	27,842,480	21.72	232,704	120	0.0028
Synonymous	4,647,072	3.62	25,422	-	0.0052
Replacement	16,347,683	12.75	40,866	-	0.0024
UTRs	6,175,566	4.82	76,920	80	0.0042
UTR 5'	1,792,164	1.40	21,953	82	0.0041
UTR 3'	4,383,402	3.42	54,967	80	0.0042
Introns	33,608,910	26.21	412,648	81	0.0041
Intergenic	60,641,470	47.30	1,961,268	31	0.0110

The minor allele frequency (MAF) for each polymorphic site was considered. Its spectrum is presented in Figure 3.7. A clear abundance, about 42%, of single alleles was detected, whereas the less represented chromosome-based MAF class (~5%) was the 4/8 (0.50), as expected.

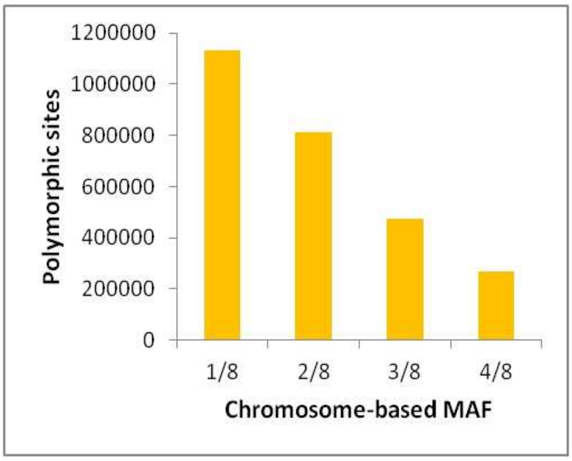


Figure 3.7: MAF distribution for 2,683,139 polymorphic sites in the 4 high coverage clones.

Tajima's D analysis

To evaluate the allelic frequency distribution of the loci and their possible deviation from neutral expectation, Tajima's D statistics were computed on the discovery panel composed by the four high coverage clones. Tajima's D was calculated across the genome and for gene, CDS and mRNA compartments, and resulted on average negative (Table 3.9). 4,295 genes, according to the *P. trichocarpa* annotation v156, were not covered. 10,493 genes had a D value >0.

Table 3.9: Tajima's D statistics for the 4 high coverage clones.

Compartment	Mean Tajima's D
Genes	-0.373
CDSs	-0.344
mRNAs	-0.356
Genome	-0.434

Genomic distribution of nucleotide diversity, Tajima's D and total divergence

A genome-scale study of the distribution of nucleotide diversity, Tajima's D and total divergence between the two species *P. nigra* and *P. trichocarpa* was conducted in the discovery panel of the 4 high coverage clones considering the polymorphic sites distributed in a total length of 128,206,359 bp across the genome, which resulted from the common nucleotide sequence covered by all the clones.

In Figure 3.8, the genomic distribution of polymorphic positions, nucleotide diversity and Tajima's D are shown together with the gene and genome k-mer distributions. K-mer level is a measure of the degree of repetitiveness of genomic regions. As overall trend in the circular plot, we can observe that the regions with higher SNP number, and with higher nucleotide diversity, correspond to regions with lower gene content and a higher level of repetitive elements, as expected. SNP and nucleotide diversity were estimated in 100 kb windows only when at least 30% of the windows was covered by aligned reads. The estimates are often missing in the highly repetitive regions because of insufficient coverage. The genomic distribution of Tajima's D is clearly negative, except for few positive spikes (74). Also Tajima's D was calculated within 100 kb windows covered at least for their 30%.

In Figure 3.9, the genomic distribution of nucleotide diversity, interspecific divergence and the *ratio* of nucleotide diversity *versus* interspecific divergence is shown together with the gene and genome k-mer distributions. The mentioned statistics were computed within 100 kb chromosome windows covered at least for their 30%. We can observed that the interspecific divergence between the two species is on average almost an order of magnitude higher than the nucleotide diversity within the *P. nigra* species. In fact, the average value of *ratio* between nucleotide diversity and interspecific divergence, across each chromosome, ranges from 0.14 to 0.18. In particular, 13 chromosomes have an average value of 0.15-0.16, 4 of 0.14, and only 2 have an average value corresponding to 0.17 and 0.18 (respectively, chromosome 19 and chromosome 17). Across each chromosome, the highest spikes of the total divergence are localised in highly repetitive regions which mostly correspond to the putative centromeric regions.

Low coverage resequencing in 42 *P. nigra* clones to identify informative SNPs

Paired-end reads of the 42 clones were aligned to the reference genome of *P. trichocarpa* as pool. Statistics concerning the number of mapped base pairs, the percentage of reference covered, the mean coverage alignment and the mean coverage excluding zero coverage regions of the alignment are indicated in Table 3.10. Approximately 54% input reads were aligned to a unique position in the reference genome and was used for SNP calling.

Table 3.10: Alignment statistics of the 42 clones resequenced at low coverage.

Clone	Total bp mapped	Refrence covered (%)	Mean coverage	Mean cov. no 0 cov. reg.
42 clones	80,727,480,833	84	191.89	227.17

SNP detection on the alignment of the 42 low coverage clones was performed with stringent parameters regarding minimum and maximum coverage. 7,920,987 SNPs were detected (Table 3.11). 55% of all SNPs detected was composed by homozygous SNPs, whereas 45% by heterozygous SNPs. Both in homozygous and heterozygous SNP class, the SNPs were more aboundant in the intergenic regions, as expected.

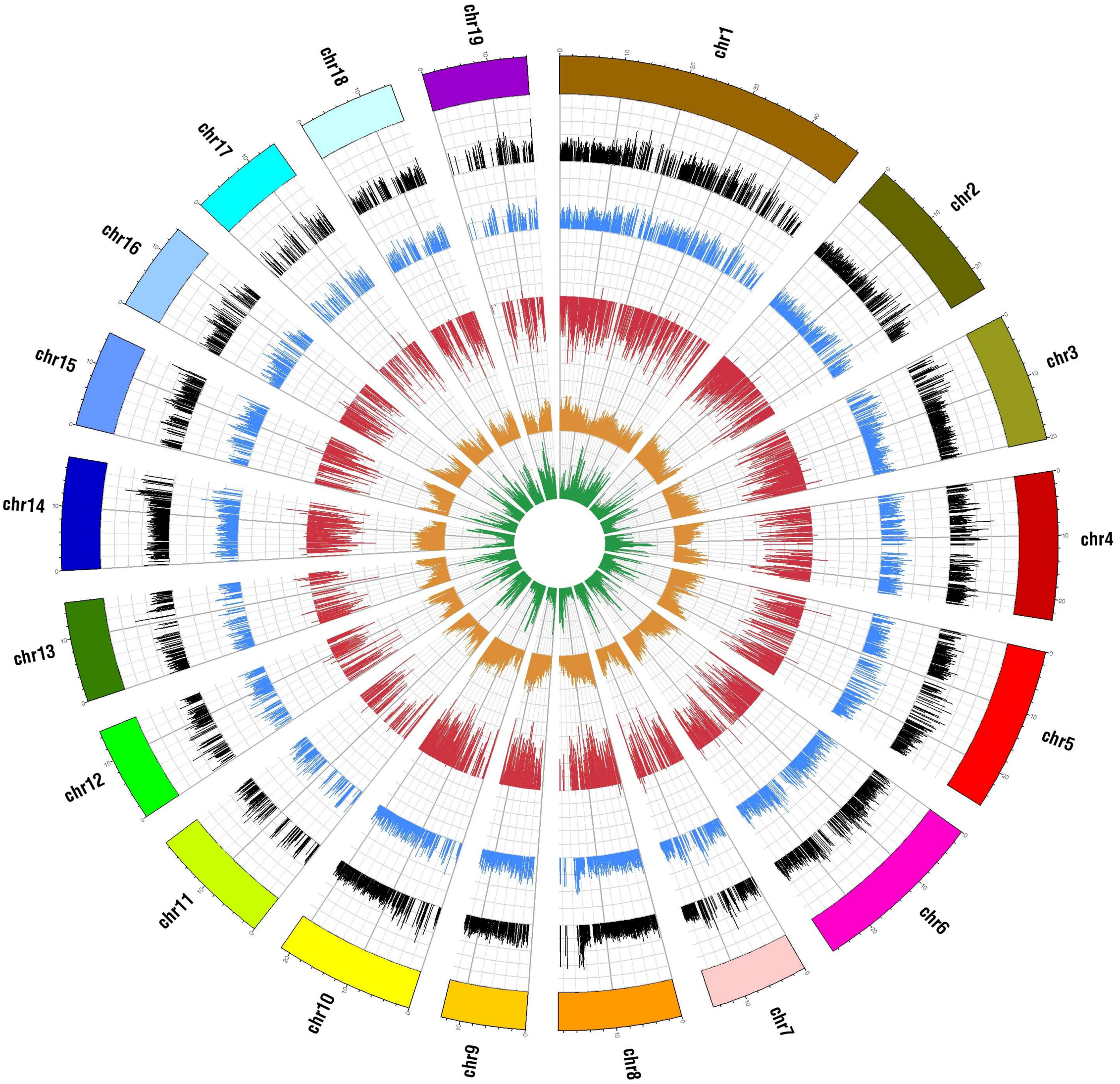


Figure 3.8: Genomic distribution of SNPs, nucleotide diversity, Tajima's D, genes and genome k-mers in the discovery panel represented by the 4 high coverage clone. SNP, nucleotide diversity and Tajima's D distributions were calculated in windows of 100 kb covered at least for their 30%. Windows not having such coverage parameters correspond to empty (blank) intervals. Colored bars=*Populus* 19 chromosomes (unit=2 Mb). In the plot: outer layer=SNP distribution (vertical unit of 500, spanning from 0 to 2500 -maximum number of SNPs that can be represented by a spike-); second layer=nucleotide diversity (x1000) (vertical unit=10, spanning from 0 to 40); third layer=Tajima's D (vertical unit=0.2, spanning from +1.0 to -1.0); forth layer=number of genes every 100 kb (vertical unit=10, spanning from 0 to 40); fifth layer=repetitiveness of the genome calculated with a k-mer analysis using the tool Tallymer [43] (vertical unit=2, spanning from 0 to 8).

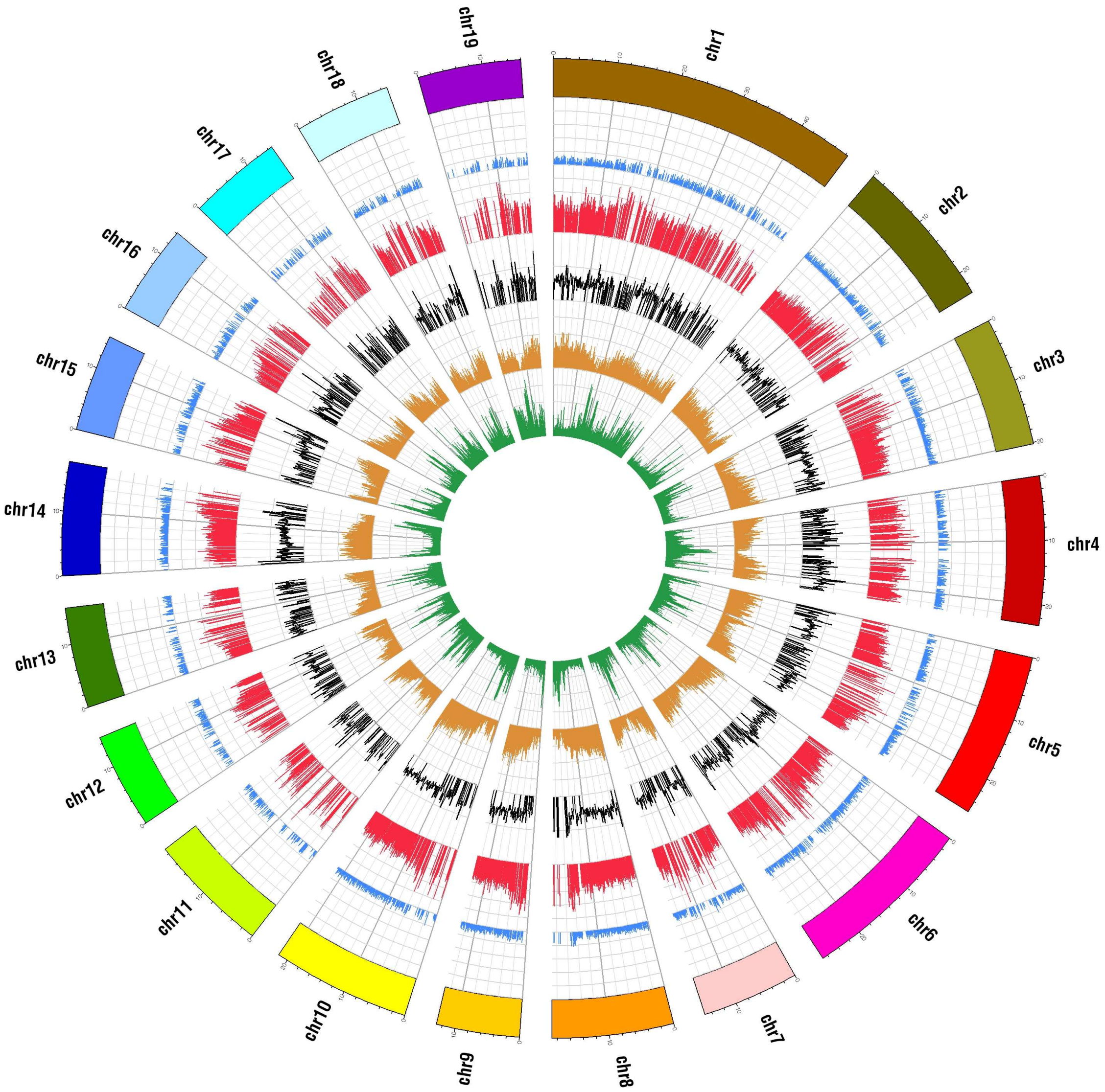


Figure 3.9: Genomic distribution of nucleotide diversity, total divergence, nucleotide diversity-total divergence *ratio*, genes and genome k-mers in the discovery panel represented by the 4 high coverage clones. Nucleotide diversity, total divergence and nucleotide diversity-total divergence *ratio* distributions were calculated in windows of 100 kb covered at least for their 30%. Windows not having such coverage parameters correspond to empty (blank) intervals. Colored bars=*Populus* 19 chromosomes (unit=2 Mb). In the plot: outer layer=nucleotide diversity (x1000) (vertical unit=20, spanning from 0 to 100); second layer=total divergence (x1000) (vertical unit=20, spanning from 0 to 100); third layer=nucleotide diversity-total divergence *ratio* (x100) (vertical unit=25, spanning from 0 to 50); forth layer=number of genes every 100 kb (vertical unit=10, spanning from 0 to 40); fifth layer=repetitiveness of the genome calculated with a k-mer analysis using the tool Tallymer [43] (vertical unit=2, spanning from 0 to 8).

Table 3.11: SNP statistics of the pool of 42 clones resequenced at low coverage.

Clone	Total SNPs	Homoz.-SNPs within/outside genes	Heteroz.-SNPs within/outside genes
42 clones	7,920,987	1,221,987/3,142,790	659,945/2,896,265

SNP frequency analysis in the 42 low coverage clones

SNP frequency was estimated in the discovery panel composed by the 42 low coverage clones considering 3,069,680 polymorphic sites distributed in a total length of 206,962,540 bp (these are the covered base pairs satisfying the parameters in the SNP detection) (Table 3.12). The SNP frequency of the intergenic domains (1 SNP every 46 bp) resulted to be the higher among all compartments considered, as expected. The total SNP frequency (1 SNP every 67 bp) was estimated to be two times higher than in gene domains, UTR and intronic regions. The two lower SNP frequencies were observed in mRNA and CDS compartments.

Table 3.12: SNP frequency in the discovery panel of the 42 low coverage clones.

Compartment	Covered bases	Total bases (%)	Polymorphic sites	Freq.: 1 SNP/... bp
Total	206,962,540	-	3,069,680	67
Genes	93,276,175	45.07	624,427	149
mRNAs	46,319,903	22.38	271,975	170
CDSs	37,258,729	18.00	204,139	183
UTRs	9,061,174	4.38	67,836	134
UTR 5'	2,974,098	1.44	22,144	134
UTR 3'	6,087,076	2.94	45,692	133
Introns	47,021,448	22.72	352,611	133
Intergenic	113,686,365	54.93	2,445,252	46

Highly informative SNP identification to design a SNP-chip

To design a 12k SNP-chip, we wanted to identify higly informative SNPs. To achieve this aim, we combined the SNP information obtained by the resequencing of both the 4 high coverage clones and the 47 low coverage clones. The sequences of the clones were aligned to a *P. nigra consensus* sequence to maximize the sequence alignment.

P. nigra consensus sequence resulted to be of 388,572,533 bp (including gaps). This length is relative to the first 40 scaffolds considered for the analyses. The 47 low coverage clones, and the 4 high coverage clones singolarly, were aligned to this *consensus*. SNPs detected in these five alignments (Input SNPs in Table 3.13) underwent several filtering step before being included in the SNP chip. The first step consisted in selecting only those SNPs which respected stringent coverage parameters (SNPs filtered, Table 3.13). The second one eliminated the SNPs in repetitive regions, such as repetitions, duplications and paralogue sequences. Left SNPs are indicated as "SNPs no repeats" in Table 3.13. In the third step, only SNPs present in the genomic intervals selected for the chip were kept (SNPs intervals, Table 3.13). The fourth step consisted in removing those polymorfic positions

Table 3.13: SNP statistics for the filtering steps adopted to extract the SNPs for the chip.

SNP description	47 clones	POLI	BEN3	BDG	71077-308
Input SNPs	758,043	937,790	282,299	491,850	460,047
SNPs filtered	548,544	629,431	204,401	319,251	285,299
SNPs no repeats	548,365	629,101	204,337	319,119	285,118
SNPs intervals	296,964	344,709	112,262	174,035	155,846
SNPs InDels	279,813	314,457	105,212	157,061	143,312
SNPs 5 accessions	278,330	not applied	not applied	not applied	not applied
Final SNPs	189,616	not applied	not applied	not applied	not applied

included in DIP and in their flanking regions ± 10 bp. Remaining SNPs from this step are reported as "SNPs InDels" in Table 3.13. Penultimate filtering step selected only those SNPs (SNPs 5 accessions), in the SNP detection performed on the 47 clone alignment, supported by at least 5 low coverage clones. The final step of the filtering pipeline selected only those SNPs, of the previous step, present in at least one high coverage clone SNP detection ("Final SNPs" in Table 3.13). The first filtering step decreased sensibly the number of SNPs considered, prooving the stringent parameters applied. Among those, a further selection was applied to get a total of 12,000 beads. This last selection was made considering the SNPs with higher score higher number of accessions supporting the locus.

Final SNPs were submitted to the Illumina chip scoring test and 113,821 SNPs had a score suitable for the chip.

SNP display and sharing

The huge amount of data detected in the high-throughput resequencing and the jointly feature of this resequencing effort enabled to set up a user-friendly tool in order to display and share among NovelTree, EvolTree and EnergyPoplar partners all the sequence data obtained. The tool selected for this purpose was the Genome Browser, which was populated with the *P. trichocarpa* genome information as reference and *P. nigra* sequence data obtained during the high-throughput resequencing. The *Populus* Genome Browser was made available at the web site http://services.appliedgenomics.org/gbrowse/populus/ (Figure 3.10). The access is restricted by providing a username and password common to all the members concerned.

The *Populus* Genome Browser hosts the sequence data by grouping them into tracks, which are classified in different sections. The sections included in the browser are as follow:

- DNA-seq Illumina profiles;
- DNA/GC Content;
- Genes. Gene annotation based on *P. trichocarpa* v156 Phytozome;
- SNPs;
- Kmers (20-mers), genome profile;
- Restriction site analysis.

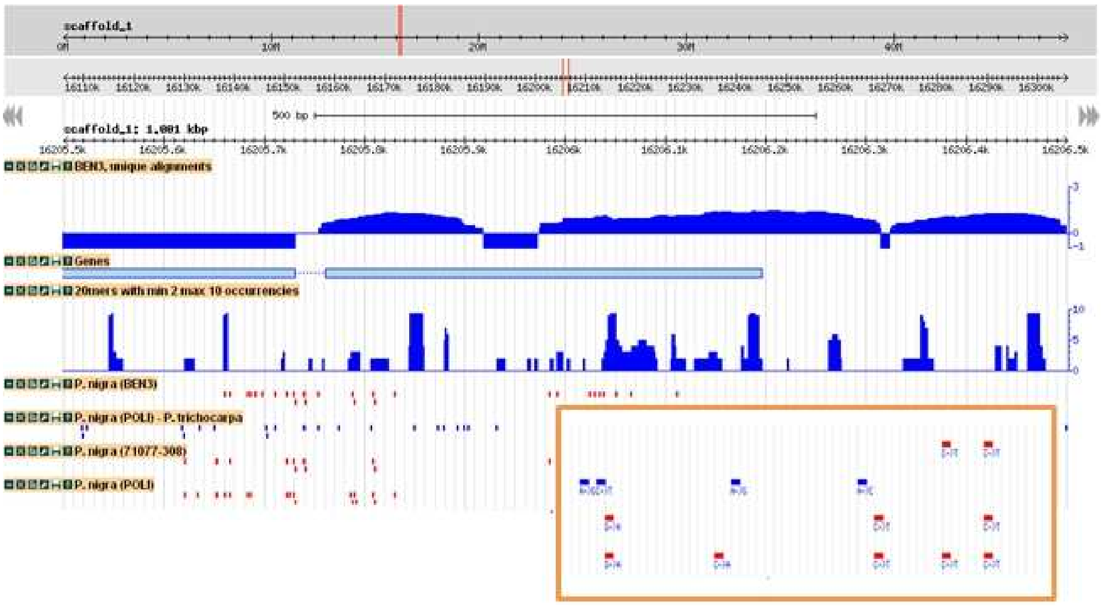

scaffold_1
0M
10M
20M
30M
40M
16110k 16120k 16130k 16140k 16150k 16160k 16170k 16180k 16190k 16200k 16210k 16220k 16230k 16240k 16250k 16260k 16270k 16280k 16290k 16300k
500 bp
scaffold_1: 1,001 kbp
16205.5k 16205.6k 16205.7k 16205.8k 16205.9k 16206k 16206.1k 16206.2k 16206.3k 16206.4k 16206.5k
BEN3, unique alignments
3
0
-1
Genes
20mers with min 2 max 10 occurrencies
10
5
0
P. nigra (BEN3)
P. nigra (POLI) - P. trichocarpa
P. nigra (71077-308)
P. nigra (POLI)

Figure 3.10: Gbrowse screenshot. SNP tracks are zoomed in.

4 Discussion

The work presented in this part of the PhD thesis proposes a whole genome Illumina resequencing approach to detect the genetic variation within the *P. nigra* genome using the *P. trichocarpa* genome sequence as reference. The SNP information obtained was applied to SNP frequency and population genetic studies at a genome scale to investigate the potential of next-generation sequencing to detect sequence variation within tree populations. This study was compared to a similar study conducted at a gene scale level on 18 *P. nigra* candidate genes for phenology traits resequenced using the Sanger method. Moreover, it describes the application of the whole genome resequencing approach in 47 *P. nigra* genotypes to detect highly informative SNPs to be used in a 12k bead-chip, which was designed considering SNPs in candidate regions.

Population genetics originated in the first half of the 20th century as a field driven by theoretical insights but with very limited empirical data, and for several decades theory remained well ahead of the data available to test its predictions. This situation began to change with the emergence of protein electrophoretic variation [44, 45, 46, 47]. Since the introduction of polymerase chain reaction (PCR) technology, the scale of data has grown exponentially, as restriction fragment length polymorphisms, microsatellites, and small-scale DNA sequencing [48] broadened the range of questions open to empirical investigation. With the recent flood of genome-wide single nucleotide polymorphism (SNP) data, and now the advent of fully sequenced population samples of genomes, population genetics has become a fundamentally data-driven discipline [49]. Population genetics is now at the core of analyses in molecular ecology and conservation biology, where it provides a framework for understanding the distribution of genetic variability among populations and for inferring the demographic histories of natural populations from molecular data. It is also central in studies of molecular evolution, providing a foundation for understanding the contributions of mutation, genetic drift, and natural selection in the evolution of genes and genomes [49]. Next-generation sequencing data will be the foundation of many future population genomic studies but it will carry new demands for computational infrastructure and statistical and bioinformatics training as the analysis of these data is currently in its infancy. The special nature of the data produced by next-generation sequencing platforms may entail a new set of challenges for unbiased estimation of population genetic parameters [49]. In contrast to traditional approaches, as the Sanger method, where a defined fragment is amplified by PCR and then sequenced, sequence reads from next-generation technologies stem from individual DNA molecules and are distributed across the genome in a largely random fashion (although regions with very high or very low GC content may be underrepresented) [50]. Data produced by these technologies, in our case the Illumina technology, suffer from three basic problems: sequence errors, assembly errors, and missing data. The severity of these problems will depend in part on the depth of sequencing, with higher coverage potentially minimizing many errors [51]. In our study we experienced the three cited problems and considered how they affect the SNP calling and consequently the nucleotide diversity estimation among our *P. nigra* genotypes.

Because Illumina reads originate from a single DNA molecule, errors in the sequences can be due to DNA damage, errors introduced during amplification, and sequencing errors [49]. Errors will inflate nucleotide diversity and skew the allele frequency spectrum (AFS) toward rare alleles, which will mainly be visible as an excess of singletons [52]. Thus far, the processing of sequence data and especially the calling of SNPs focused on minimizing the false-positive rate, by introducing stringent quality criteria to call SNPs [53]. On the other hand, Johnson and Slatkin [52] noted that stringent SNP-calling criteria will bias diversity estimates by excluding many true SNPs (especially rare alleles) from the data. We considered a compromise in parameter setting in the SNP calling to balance the false-positive rate and not-called positions.

Illumina sequence reads are still shorter than in traditional Sanger sequencing and this poses serious challenges for assembling reads [54, 55], as well as mapping reads to a reference genome [56, 57] even if it is of a closely related species as *P. trichocarpa.* Assembly was challenging in repetitive or highly polymorphic genomic regions, as demonstrated by the SNP distribution in Figure 3.8 (SNP information is missing in the chromosome regions which are highly repetitive). It is worthwhile to consider the potential biases that an imperfect assembly may introduce. For some mapping algorithms, sequence reads with more than one or two differences from a reference genome will not be placed [56]. This makes the mapping of alleles that are different from the reference genome less probable than for a reference-matching allele, causing a bias in allele frequency toward the allele found in the reference sequence. It may additionally reduce the number of SNPs discovered and bias estimates of nucleotide diversity toward smaller values. Indels present in the resequenced alleles may cause similar problems, since many mapping algorithms are unable to deal with them, especially when they are large. We tried to reduce this bias keeping in mind this problem when we set the alignment parameters. A trade-off between sequence specificty and mismatching bases was chosen.

Another challenge for the analysis of whole-genome sequence polymorphism is missing data. Due to the stochastic placement of sequence reads across the genome, the sampled chromosomes at any particular site may not include all individuals [49]. And unless all samples are sequenced at very high genomic coverage, it may not be clear whether both of a diploid individuals alleles have been sequenced. A study showed that accurate detection of variant loci necessitates a 20-fold read depth per base [58, 16]. We resequenced four *P. nigra* genotypes at high coverage ($\geq$20X) to perform the population genetic analyses aware of these problems.

Before the advent of next-generation sequencing, a common approach to detect SNPs was the resequencing of amplicons using the traditional Sanger method. This technique has been mainly used for the haplotype-based study of human genetic variation [3], but there are many examples of this technique application also in herbaceous plants [13, 14, 59, 60, 61, 62] and trees [63, 64, 65]. The advantage of this approach is that the sequence from each individual is determined through double-strand sequencing and SNPs can be identified in a reliable way with a false discovery rate usually significantly below 5% [3]. The main disadvantage of this approach is that it requires an enormous effort for the analysis of many genes since for each gene, specific primers have to be developed and usually a larger number of genotypes need to be amplified and sequenced. A concern about the Sanger resequencing regards the design of primers. Indeed, PCR primers sometimes overlapping unknown DNA variants lead to imbalanced amplification of the two alleles [42] and therefore to a miscall of a polymorphic position.

Illumina resequencing has two large advantages compared to the Sanger resequencing: it is cheaper and quicker. These characteristics allow to resequence many individuals on a genome scale to detect SNP variation and conduct population genetic analysis.

In this work we performed the SNP detections with stringent parameters as we wanted to be as much accurate as possibile, despite missing some rare alleles. To detect the SNPs in the Illumina data set, we were very stringent with minimum and maximum coverage parameters to avoid the SNP calling respectively in regions covered by a low number of reads and in regions covered by a high number of reads, likely representing repetitive regions or paralogos. On the other hand, to detect SNPs in the Sanger data set, we called the polymorphic positions when associated to high quality score values. In general, also the MAF (Minor Allele Frequency) of a SNP can be used as stringent parameter when set at too high values. Indeed, a high MAF excludes the rare alleles. To evaluate the feasibility of using the Illumina technology in population genetic studies, we estimated the false discovery rate (FDR) of the Illumina platform considering the ABI Sanger method as "gold standard" in terms of sequencing accuracy [37]. We found the Illumina sequencing is sensitive at 93%, accurated at ~100% and its FDR is 7.6%. We reckoned these statistics are acceptable to apply the Illumina technology to population genetic studies.

The polymorphism frequency (~1 SNP every 84 bp) detected in the 18 *P. nigra* candidate genes was lower than that detected in *P. tremula* (~1 SNP every 60 bp) [64] and higher than that detected in *P. balsamifera* (~1 SNP every 95 bp) [65] and *P. trichocarpa* (~1 SNP every 130 bp) [66].

The level of genetic variation in the genome under study can be expressed as nucleotide diversity (π). π is key to most phenotypic variation and can reflect evolutionary history. It was calculated exploiting the SNP information detected both in the Sanger 48 clone DP within 18 candidate genes specific for the phenology pathway and in the Illumina 4 clone DP across the whole genome. The SNP information obtained in the Illumina 42 clone DP was not considered for the population genetic analyses as the 42 clones were resequenced at low coverage and so not enough reliable for the detection of heterozygous positions. Considering the different nucleotide diversity classes reported in Table 3.8, the Illumina 4 clone DP showed higher values for each class. The higher diversity found at genome-wide level compared to that found in a small set of candidate genes, selected as putative responsible for a specific trait phenology-, may reflect weaker selective constraints on pseudogenes and non-functional gene copies. The overall π of the Sanger 48 clone DP estimated within the 18 candidate genes (π=0.0016) resulted not significantly different from that detected in other forest trees: *P. trichocarpa* (π = 0.0019, [66]), *Picea abies* (π = 0.0021, [69]) and *Populus balsamifera* (π = 0.0026, [65]). Such nucleotide diversity can be considered limited, being lower than the averages found in European aspen (π= 0.011, [64]), loblolly pine (π= 0.0039, [70]), maize (π = 0.0063, [71]) and sugar beet (π = 0.0076, [72]). Instead, overall π of the Illumina 4 clone DP estimated across the whole genome (π = 0.0071) resulted to be higher than the π estimated among all cited species, except for the European aspen and the sugar beet. We also studied the nucleotide diversity distribution in the 19 chromosomes of the *P. nigra* genome in 100 kb windows (Figure 3.8). We observed a clear correlation among the distribution of the nucleotide diversity, the genes and the repetitive regions across the genome. Regions with higher gene content and lower repetitive sequences present lower nucleotide diversity estimates, and vice versa. Nucleotide diversity and SNP frequency were significantly higher in intergenic regions than in genes and even than in the non coding portions of genes (introns, UTRs). Our validation analysis compared SNP identification in genic regions and thus we can not be sure that the same FDR applies to the intergenic ones. It is possible that misalignment of reads may occur more frequently in intergenic regions due to the presence of repeats and this may explain the overall rather high SNP frequency and nucleotide diversity estimates at the whole genome level. In the genomic windows not covered at least for the 30%, the nucleotide diversity was not computed. These windows are frequently localised in repetitive regions indicating problems in the read alignment. Moreover, we observed a negative correlation (which is a decreasing of nucleotide diversity) from centromeric to telomeric regions between nucleotide diversity and distance from the centromere. This kind of negative correlation was seen also in *Medicago truncatula* [73] and *A. thaliana* [74]. By contrast, nucleotide diversity increases with increasing distance from the centromeric regions in *Zea mays* [75], *Drosophila simulans* [76], and humans [77]. We supposed that not finding reduced diversity near the centromeres and telomeres may be related to heterochromatic regions that could be missing from the reference genome and/or the assembly and may not reflect fundamental differences in the forces shaping nucleotide diversity. The interspecific divergence between *P. nigra* and *P. trichocarpa* species was also estimated at a genome level. Its value is on average almost an order of magnitude higher than the nucleotide diversity across the whole genome. As the nucleotide diversity, the higher values of the total divergence are localised in regions with higher gene content and lower repetitive sequences.

In this work it is also presented one of the first studies of Tajima's D at genome scale (Figure 3.8). The Tajima's D test measures the allele frequency distribution of nucleotide sequence data. This statistic can be influenced by both population history and natural selection. A positive value indicates an excess of intermediate frequency (polymorphic) alleles, while a negative value indicates an excess of rare alleles. The null hypothesis of the Tajima's D test is neutral evolution in an equilibrium population. This implies that no selection is acting at the locus and that the population has not experienced any recent growth or contraction [32]. Only a recent publication in *M. truncatula* [73] presented a Tajima's D study at genome level. In other *Populus* species, Tajima's D was calculated on a restricted set of candidate genes [64, 65, 66]. Gene D values are sligthly different between the 18 candidate genes (D = -0.679) and across the whole *P. nigra* genome (D = -0.373) probably due to the fact that in the first case we have a biased set of genes since they are candidates for the phenology traits and belong to the same metabolic pathway. Low-frequency alleles were more common than expected under a standard neutral model, which was reflected in negative distributions of Tajima's D for both sliding 100 kb windows and mean values in genes, as well as in putative transcripts (D = -0.356). In the presence of strong selection shaping diversity, we would expect to find windows or regions exhibiting an excess of low-frequency variants (low D). Regions of low diversity or very negative Tajima's D suggest the presence of selective sweeps, which represent the reduction or elimination of variation among the nucleotides as the result of recent and strong positive purifying natural selection. Thus, these regions of low diversity or very negative D are obvious *a posteriori* regions to search for targets of recent species-wide selective sweeps. We detected 22 windows/regions having very negative D. We have to identify those genes with putative functions (e.g., pathogen defense) harboured by these regions that make them obvious targets of strong selection. On the other hand, we found 107 windows with a positive value for Tajima's D which may be yielded by balancing selection. In these regions the genes may be exposed to a more efficient selection because they represent crucial keys for important quantitative or adaptive traits.

To conclude, the presented work proved the feasibility of applying the Illumina sequencing to population genetic studies at a genome level. Moreover, exploiting the Illumina sequencing technology, 47 *P. nigra* individuals (42 clones used for the SNP frequency study plus 5 clones to increase the information) were resequenced at low coverage to detect highly informative SNPs to be used in a high density bead chip which is under production. In this perspective, 42 clones were studied to characterise and evaluate the achievable SNP information using a stringent MAF parameter, which is required to provide markers to be used in marker assisted selection (MAS) studies. The chip will be functional for association studies to improve breeding programs in black poplar. Four high coverage clones information has been used to populate a Genome Browser which is available for the members working in three European projects: NovelTree, EvolTree and EnergyPoplar.

Bibliography

[1] Neale, D.B. Genomics to tree breeding and forest health. *Current Opinion in Genetics & Development* **17**, 539-544 (2007).

[2] White, T.L., Adams, W.T., Neale. D.B. Forest Genetics. *CABI Publishing* (2007).

[3] Ganal, M.W., Altmann, T., Róder, M.S. SNP identification in crop plants. *Current Opinion in Plant Biology* **12**, 211-217 (2009).

[4] Bomblies, K., Weigel, D. *Arabidopsis*: a model genus for speciation. *Curr Opin Genet Dev* **17**, 500-504 (2007).

[5] Rafalski, A. Applications of single nucleotide polymorphisms in crop genetics. *Curr Opin Plant Biol* **5**, 94-100 (2002).

[6] International HapMap Consortium, Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851-861 (2007).

[7] McCarthy, M.I., Abecasis, G.R., Cardon, L.R., Goldstein, D.B., Little, J., Ioannidis, J.P., Hirschhorn, J.N. Genome wide association studies for complex traits: *consensus*, uncertainty and challenges. *Nat Rev Genet* **9**, 356-369 (2008).

[8] Batley, J., Barker, G., O'Sullivan, H., Edwards, K., Edwards, D. Mining for single nucleotide polymorphisms in insertions/deletions in maize expressed sequence tag data. *Plant Phys* **132**, 84-91 (2003).

[9] Kota, R., Rudd, S., Facius, A., Kolesov, G., Thiel, T., Zhang, H., Stein, N., Mayer, K., Graner, A. Snipping polymorphisms from large EST collections in barley (*Hordeum vulgare* L.). *Mol Genet Genomics* **270**, 24-33 (2003).

[10] Yamamoto, N., Tsugane, T., Watanabe, M., Yano, K., Maeda, F. *et al.* Expressed sequence tags from the laboratory-grown miniature tomato (*Lycopersicon esculentum*) cultivar Micro-Tom and mining for single nucleotide polymorphisms and insertions/deletions in tomato cultivars. *Gene* **356**, 127-134 (2005).

[11] Dantec, L.L., Chagné, D., Pot, D. *et al.* Automated SNP detection in expressed sequence tags: statistical considerations and applications to maritime pine sequences. *Plant Mol Biol* **54**, 461-470 (2004).

[12] Das, S., Bhat, P.R., Sudhakar, C. *et al.* Detection and validation of single feature polymorphisms in cowpea (*Vigna unguiculata* L. Walp) using a soybean genome array. *BMC Genomics* **9**, 107 (2008).

[13] Wright, S.I., Bi, I.V., Schroeder, S.G., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S. The effects of artificial selection on the maize genome. *Science* **308**, 1310-1314 (2005).

[14] Yamasaki, M., Tenaillon, M.I., *et al.* A large-scale screen for artificial selection in maize identifies candidate agronomic loci for domestication and crop improvement. *Plant Cell* **17**, 2859-2872 (2005).

[15] Mardis, E. The impact of next-generation sequencing technology on genetics. *Trends in genetics : TIG* **24**, 133-41 (2008).

[16] Nielsen, R., Paul, J.S., Albrechtsen, A., Song, Y.S. Genotype and SNP calling from next-generation sequencing data. *Nature reviews. Genetics* **12**, 443-451 (2011).

[17] Schatz, M.C., Delcher, A.L., Salzberg, S.L. Assembly of large genomes using second-generation sequencing. *Genome research* **20**, 1165-73 (2010).

[18] Neale, D.B., Ingvarsson, P.K. Population , quantitative and comparative genomics of adaptation in forest trees. *Current Opinion in Plant Biology* **11**, 149-155 (2008).

[19] Dvornyk, V., Sirvio, A., Mikkonen, M., Savolainen, O- Low nucleotide diversity at the pal1 locus in the widely distributed *Pinus sylvestris. Mol Biol Evol* **19**, 179-188 (2002).

[20] Brown, G.R., Gill, G.P., Kuntz, R.J., Langley, C.H., Neale, D.B. Nucleotide diversity and linkage disequilibrium in loblolly pine. *Proc Natl Acad Sci USA* **101**, 15255-15260 (2004).

[21] Savolainen, O., Pyha, T. Genomic diversity in forest trees. *Current Opinion in Plant Biology* **10**, 162167 (2007).

[22] Tuskan, G.A., Difazio, S., Jansson, S. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313** 1596-604 (2006).

[23] The International SNP Map Working Group. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928-933 (2001).

[24] Bentley, D.R. Whole-genome re-sequencing. *Current Opinion in Genetics & Development* **16**, 545-552 (2006).

[25] Harismendy, O., Ng, P.C., Strausberg, R.L., *et al.* Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biology* **10**, R32 (2009).

[26] Chou, H. and Holmes, M.H. DNA sequence quality trimming and vector removal. *Bioinformatics* **17**, 1093-1104 (2001).

[27] Ewing, B., Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research* **8**, 186-94 (1998).

[28] Ewing, B., Hillier, L., Wendl, M.C., Green, P. Base-Calling of Automated Sequencer Traces UsingPhred. I. Accuracy Assessment. *Genome Research* **8**, 175-185 (1998).

[29] Gordon, D., Abajian, C., Green, P. Consed: a graphical tool for sequence finishing. *Genome Research* **8**, 195-202 (1998).

[30] Nickerson, D.A., Tobe, V.O., Taylor, S.L. PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Research* **25**, 27452751 (1997).

[31] Librado, P. and Rozas, J. DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**, 1451-1452 (2009).

[32] Tajima, F. Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. *Genetics* **123**, 585-595 (1989).

[33] Nei, M. and Li, W.H. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc Natl Acad Sci USA* **76**, 5269-5273 (1979).

[34] Kimura, M. Rare variant alleles in the light of the neutral theory. *Mol Biol Evol* **1**, 84-93 (1983).

[35] Zhang, H., Zhao, X., Ding, X., Paterson, A.H., Wing, R.A. Preparation of megabase-size DNA from plant nuclei. *The Plant Journal* **7**, 175-184 (1995).

[36] Doyle, J.J., Doyle, J.L. A rapid DNA isolation procedure from small quantities of fresh leaf tissues. *Phytochem Bull* **19**, 11-15 (1987).

[37] Bonetta, L.: Genome sequencing in the fast lane. *Nature Methods* **3**, 141-147 (2006).

[38] Elston, R.C., Johnson, W.D. Essential of biostatistics. 2nd ed. Philadelphia: FA Davis (1994).

[39] Benjamini, Y. and Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society* **57**, 289-300 (1995).

[40] Simpson, G.G., Dean, C. *Arabidopsis*, the Rosetta Stone of Flowering Time? *Science* **296**, 285-289 (2002).

[41] Goodstein1, D.M., Shu., S., *et al.* Phytozome: a comparative platform for green plant genomics. *Nucl. Acids Res.* 1-9 (2011).

[42] Quinlan, A.R., Marth, G.T. Primer-site SNP mask mutations. *Nature Methods* **4**, 192 (2007).

[43] Kurtz, S. Narechania, A., Stein, J.C., Ware, D. A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. *BMC genomics* **9**, 517 (2008).

[44] Harris, H. Enzyme polymorphisms in man. *Proc R Soc Lond B Biol Sci* **164**, 298310 (1966).

[45] Hubby, J.L., Lewontin, R.C. A molecular approach to the study of genic heterozygosity in natural populations. I. The number of alleles at different loci in *Drosophila pseudoobscura. Genetics* **54**, 577594 (1966).

[46] Lewontin, R.C., Hubby, J.L. A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura. Genetics* **54**, 595609 (1966).

[47] Lewontin, R.C. The apportionment of human diversity. In Evolutionary biology (ed. TH Dobzhansky et al.), pp. 381398. Kluwer Academic Publishers, New York (1972).

[48] Kreitman, M. Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster. Nature* **304**, 412417 (1983).

[49] Pool, J.E., Hellmann, I., Jensen, J.D., Nielsen, R. Population genetic inference from genomic sequence variation. *Genome research* **20**, 291-300 (2010).

[50] Ossowski, S., Schneeberger, K., Clark, R.M., Lanz, C., Warthmann. N, Weigel, D. Sequencing of natural strains of *Arabidopsis thaliana* with short reads. *Genome Res* **18**, 20242033 (2008).

[51] Bentley, D.R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 5359 (2008).

[52] Johnson, P.L.F., Slatkin M. Accounting for bias from sequencing error in population genetic estimates. *Mol Biol Evol* **25**, 199206 (2008).

[53] Altshuler, D., Pollara, V.J., Cowles, C.R. *et al.* An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* **407**, 513516 (2000).

[54] Sundquist, A., Ronaghi, M., Tang, H., Pevzner, P., Batzoglou, S. Whole-genome sequencing and assembly with high-throughput, short-read technologies. *PLoS One* **2**, e484. doi: 10.1371/journal.pone.0000484 (2007).

[55] Chaisson, M.J., Pevzner, P.A. Short read fragment assembly of bacterial genomes. *Genome Research* **18**, 324330 (2008).

[56] Li, H., Ruan, J., Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research* **18**, 18511858 (2008).

[57] Li, R., Li, Y., Kristiansen, K.,Wang, J. SOAP: Short Oligonucleotide Alignment Program. *Bioinformatics* **24**, 713714 (2008).

[58] Craig, D.W., Pearson, J.V., *et al.* Identification of genetic variants using bar-coded multiplexed sequencing. *Nature Methods* **5**, 887-893 (2008).

[59] Choi, I.Y., Hyten, D.L., Matukumalli, L.K.*et al.* A soybean transcript map: gene distribution, haplotype and single-nucleotide polymorphism analysis. *Genetics* **176**, 686-696 (2007).

[60] Nordborg, M., Hu, T.T., Ishino, Y. *et al.* The pattern of polymorphism in *Arabidopsis thaliana. PLoS Biol* 3:e196 (2005).

[61] Schmid, K., Ramos-Onsins, S. *et al.* A multilocus sequence survey in *Arabidopsis thaliana* reveals a genome-wide departure from a neutral model of DNA sequence polymorphism. *Genetics* **169**, 1601-1615 (2005).

[62] Nasu, S., Suzuki, J., Ohta, R. *et al.* Search for and analysis of single nucleotide polymorphisms (SNPs) in rice (*Oryza sativa, Oryza rufipogon*) and establishment of SNP markers. *DNA Res* **9**, 163-171 (2002).

[63] Pavy, N., Pelgas, B., Beauseigle, S. *et al.* Enhancing genetic mapping of complex genomes through the design of highly-multiplexed SNP arrays: application to the large and unsequenced genomes of white and black spruce. *BMCGenomics* 9:21 (2008).

[64] Ingvarsson, P.K. Nucleotide polymorphism and linkage disequilibrium within and among natural populations of European aspen (*Populus tremula* L., *Salicaceae*). *Genetics* **169**, 945-953 (2005).

[65] Olson, S., Robertson, A.L., *et al.* Nucleotide diversity and linkage disequilibrium in balsam poplar (*Populus balsamifera*). *New Phytologist* **186**, 526-536 (2010).

[66] Gilchrist, E.J., Haughn, G.W. *et al.* Use of Ecotilling as an efficient SNP discovery tool to survey genetic variation in wild populations of *Populus trichocarpa. Molecular Ecology* **15**, 13671378 (2006).

[67] Imbert, E., Lefevre, F. Dispersal and gene flow of *Populus nigra* (Salicaceae) along a dynamic river system. *Jurnal of Ecology* **91**, 447-456 (2003).

[68] van Dam, B.C. EUROPOP: Genetic diversity in river populations of european black poplar for evaluation of biodiversity, conservation strategies, nature development and genetic improvement. In: van Dam, B.C., Bordacs, S., eds. Genetic diversity in river populatins of european black poplar, Budapest: Csiszar Nyomda,15-32.

[69] Heuertz, M., De Paoli, E., Kllman, T., Larsson, H., Jurman, I., Morgante, M., Lascoux, M., Gyllenstrand, N. Multilocus patterns of nucleotide diversity, *linkage disequilibrium* and demographic history of Norway spruce (*Picea abies* (L.) Karst). *Genetics* **174**, 2095-2105 (2006).

[70] Brown, G.R., Gill, G.P., Kuntz, R.K., Langley, C.H., Neale, D.B.: Nucleotide diversity and linkage disequilibrium in loblolly pine. *Proc Natl Acad Sci USA* **101**, 15255-15260 (2004).

[71] Ching, A., Caldwell, K.S., *et al.* SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. *BMC Genetics* **3**, 19 (2002).

[72] Schneider, K., Weisshaar, B., Borchardt, D.C., Salamini, F. SNP frequency and allelic haplotype structure of *Beta vulgaris* expressed genes. *Molecular breeding* **8**, 63-74 (2001).

[73] Branca, A., Paape, T.D., Zhou, P. *et al.* Whole-genome nucleotide diversity, recombination, and linkage disequilibrium in the model legume *Medicago truncatula. Proceedings of the National Academy of Sciences*, **108**, E864E870 (2011).

[74] Clark, R.M. *et al.* Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana. Science* **317**, 338-342 (2007).

[75] Gore, M.A. *et al.* A first-generation haplotype map of maize. *Science* **326**, 1115-1117 (2009).

[76] Begun, D.J. *et al.* Population genomics: Whole-genome analysis of polymorphism and divergence in *Drosophila simulans. PLoS Biol* 5:e310 (2007).

[77] Hellmann, I. *et al.* Population genetic analysis of shotgun assemblies of genomic sequences from multiple individuals. *Genome Res* **18**, 1020-1029 (2008).

II

De novo assembly and comparative genomic analysis in Populus nigra

5 Introduction

Forests cover 30% (about 3.8 billion ha) of Earth's terrestrial surface, harbor substantial biodiversity, and provide humanity with benefits such as clean air and water, lumber, fiber and fuels [1]. Worldwide, one-quarter of all industrial feedstocks have their origins in forest-based resources [2]. Large and long-lived forest trees grow in extensive wild populations across continents, and they have evolved under selective pressures unlike those of annual herbaceous plants.

Populus nigra L., the European black poplar, has a wide distribution area ranging from the Mediterranean border in the south to 64°latitude in the north, and from the British Isles to western Asia [3]. It is a pioneer tree species of the riparian ecosystem, strictly heliophilous, which forms metapopulations by colonising open areas through seeds and propagules (cuttings, root suckers) [3]. *P. nigra* is characterised by a great diversity of population types, from isolated trees to huge pure or mixed stands. The species is dioecious and anemogamous. *P. nigra* is of economic interest as a pure species. Due to its plasticity, it is also used as a pure species for soil protection, and afforestation in polluted industrial zones [4]. The necessity for model species in plants is well recognized and, in this role, *Arabidopsis thaliana* has gained a supreme acceptance amongst plant scientists [5]. It has often been considered, however, that *Arabidopsis* cannot be a useful model for trees as it does not undergo wood formation and seasonality of growth. The major sets of tools and resources that have propelled *Populus* into the set of model plants are the genomic resources [6] culminating in the completion of a draft sequence of the *Populus trichocarpa* genome [1]. The *Populus* genera typically has a haploid chromosome number of 19 and small and convenient genome size (~ 500 Mb). *Populus* has experienced at least two genome-wide duplication events, the salicoid event approximately 60-65 MYA and the older eudicot triplication event, as well as numerous segmental and tandem duplication events [1]. The assembled and assigned genomic sequence of the poplar nuclear genome is mostly euchromatic DNA including the majority of transcribed genes. A draft reference set of 45,555 protein-coding gene loci was identified [1]. Moreover, 40% of the whole poplar genome consists of repetitive elements, likely the majority of the unassembled genome sequence is heterochromatic DNA [7].

Considering the economic interest for *P. nigra* in Europe and the avaibility of next generation sequencing technologies, we propose a draft genomic sequence of *P. nigra* exploiting a *de novo* assembly approach.

After the successful assembly of the human [8, 9] and mouse [10] genomes by whole-genome shotgun sequencing, most large-scale genome projects quickly moved to adopt the WGS approach, which has subsequently been used for dozens of eukaryotic genomes. Today, thanks to changes in sequencing technology, a major issue confronting genome projects is whether we sequence a large genome (>100Mbp) using short reads [11]. Several large draft genomes have been published that used a combination of Sanger and short-read sequencing. The draft assembly of grapevine (*Vitis vinifera*, genome size ~ 500 Mb) reported in [12] combined Sanger and 454 sequencing. The draft genome sequence of cucumber, *Cucumis sativus*, genome size ~ 360 Mb, was obtained using a combination of Sanger and Illumina sequencing [13]. The first *de novo* assembly of a novel large genome, exclusively obtained through short-reads sequencing, belongs to the giant panda, *Ailuropoda melanoleura* (genome size ~ 2.4 Gb), which was published by the Beijing Genome Institute in 2010 [14]. This assembly used only Illumina reads averaging 52 bp in length and was done with the SOAPdenovo assembler. An assembly of the cod genome (*Gadus morhua*, genome size ~ 800 Mb) has recently been published [15]. The genome assembly was obtained exclusively by 454 sequencing of shotgun and paired-end libraries. Instead, the *de novo* assembly of the wild strawberry genome (*Fragaria vesca*, genome size ~ 240 Mb) was performed using a combination of 454, Illumina and SOLiD data [16].

The results achieved in the cited works make it clear that assemblies using NGS reads alone are substantially inferior to what can be accomplished using Sanger sequencing [11]. The two-to-three orders of magnitude cost advantage of NGS, however, will continue to make it much more appealing, and for many genomes it may be the only affordable option. The assembly results now being obtained with NGS are scientifically valuable: they cover most of the genome and they produce contigs and scaffolds long enough for comprehensive gene-annotation efforts [11]. These results will continue to improve as NGS read lengths grow, paired-end protocols improve, and assembly software innovations appear. The keys to good assembly results include deep coverage by reads with lengths longer than common repeats, and paired-end reads from short (0.53 kb) and long (>3 kb) DNA fragments [11]. To obtain large scaffolds and fill in repeat-induced gaps, a sequencing project should also generate a large set of reliable paired-end reads. As long as both ends of a pair map uniquely to contigs, the pair can be used for scaffolding, and, to fill in scaffold gaps, we need paired reads in which one read is anchored in a contig and its mate falls in the gap. For this reason, a mixture of several fragment sizes is necessary to resolve the short and long repeats in a genome. Longer (e.g., 454-based) reads are also advantageous in resolving the most complicated repeats, but (potentially modest) [11] improvements to assembly quality may not justify the higher costs of long reads. More important than the read length of paired reads, however, is the number of distinct, nonchimeric pairs produced. Protocols to generate paired reads are still being refined, and we have seen sequencing runs that suffered from having very few distinct pairs in them, from having numerous redundant pairs (the same pairs occurred repeatedly), and from having chimeric pairs (the paired sequences were not at the expected separation and orientation in the genome).

As we have briefly shown, the rapid and continuing development of next-generation sequencing technologies has made it feasible to contemplate sequencing the genomes of hundreds, if not thousands, of species of agronomic, evolutionary, and ecological importance, as well as biomedical interest [17]. Indeed, the Genome 10K Community has proposed to assemble a “virtual collection” of frozen or otherwise suitably preserved tissues or DNA samples representing on the order of 10,000 extant vertebrate species, including some recently extinct species that are amenable to genomic sequencing [17]. It is already well accepted that in few years we will obtain a huge collection of genomic sequences, representing the variability among the main living organisms. However, it is also well-known the biological analysis of the sequenced genomes is not developing as quickly as the sequencing technology. This will lead us to possess valuable genomic resources without a proper and complete meaning. The presence of sequenced genomes from different closely related species, will allow us to conduct several comparative genomic analyses to identify the biological correlations among species.

This part of the PhD project aims to produce the *de novo* assembly of the *P. nigra* genome. We propose an assembly composed solely of Illumina paired-end and mate-pair data in order to define a *de novo* assembly pipeline for Illumina data. We called "mate-pair" reads those read pairs having a reverse-forward orientation, contrary to the "paired-end" reads, and which are useful to order and orient the contigs. Furthermore, we propose a genomic analysis pipeline to discover and define the main genomic content differences between two closely related tree species: *P. nigra* and *P. trichocarpa*. In this way, we introduce the pan-genome concept in forest trees. The pan-genome concept was initially introduced in bacterial species [18] and subsequently in maize [19]. In particular, the pan-genome includes a core genome containing genes that are present in all individuals and a dispensable genome composed of partially shared and individual-specific DNA sequence elements [19].

6

Materials and Methods

Plant material

A *Populus nigra* heterozygous male genotype, POLI, was selected for the *de novo* assembly purpose. This genotype was provided by the DiSAFRi - University of Tuscia (Viterbo, Italy). POLI was originally collected in the South of Italy by Sinni river.

Nuclei and DNA extraction

Leaf material (fresh or frozen at -80 ℃) from young leaves was ground into a fine powder using mortar and pestle in the presence of liquid nitrogen. Nuclei were extracted from 5 g of grounded material per preparation, according to [20]. The genomic DNA was then extracted from the nuclei following a modified Doyle&Doyle protocol [21].

Illumina library preparation

Standard paired-end libraries

Paired-end sequencing libraries were constructed with insert sizes of about 500-600 bp. The library preparation followed the “Illumina Paired-End Sample Preparation” TruSeq protocol, according to the manufacturer guidelines (Illumina, Inc. San Diego, CA, USA).

Overlapping paired-end libraries

Paired-end sequencing libraries were constructed with insert sizes of 180 bp. The library preparation followed the “Illumina Paired-End Sample Preparation” protocol, according to the manufacturer guidelines (Illumina, Inc. San Diego, CA, USA). The read pairs were merged using the SHERA software [22].

Hybrid mate-pair libraries

Mate-pair sequencing libraries were constructed with insert sizes of about 3 kb. Library preparation followed the “Paired End Library Preparation Method Manual, 3kb Span” 454 protocol (provided by Roche Applied Science, Basel, Switzerland) up to the Library immobilization step, then the preparation followed the “Illumina Paired-End Sample Preparation” protocol, according to the manufacturer guidelines (Illumina, Inc. San Diego, CA, USA).

K-mer analysis

To estimate the quality of the libraries, a home-made script was developped to calculate the *k*-mers of the reads. Read *k*-mers were calculating in sliding windows of 16 bp.

Illumina sequencing

The sequencing process followed Illumina instruction for Genome Analyzer IIx and HiSeq2000. Runs were performed for 101 or 110 cycles. The fluorescent images were processed to sequences using the Illumina data processing pipeline (v1.7 and v1.8).

De novo assembly

Paired-end short reads were assembled using two softwares: CLC Genomics Workbench (http://www.clcbio.com) and ABySS [23]. Both assemblers are based on a de Bruijn-like data structure [24]. For ABySS, *k* parameter was set to 50. Before assembling the reads, they were trimmed for quality, using Mott's trimming algorithm (http://www.clcbio.com/manual/genomics/Quality_trimming.html) [25], and filtered for chloroplast and mitochondrial contaminations. Both operations were performed using the rNA software [26].

Assembly merging

The GAM software [27] was used to merge the two assemblies produced using both CLC and ABySS. Each assembly was set both as master and as slave. The most fundamental parameter of this tool is the number of reads that two segments must share in order to be merged. This parameter was set to 500 in order to be as conservative as possible to avoid creation of chimeric contigs.

De novo assembly validation

Masking of the *de novo* assembly onto *P. trichocarpa* genome sequence

To verify the portion of the *P. trichocarpa* genome sequence covered by the *P. nigra de novo* assembly, *P. nigra* contigs were mapped using Mummer3 [28] onto the 19 Linkage Groups sequences of *P. trichocarpa* [1]. RepeatMasker [29] was used to mask onto *P. trichocarpa* 19 Linkage Groups sequences a library of known *P. trichocarpa* transposable and repetitive elements [30].

Gene reconstruction in *de novo* assembly

A BLASTn analysis was done to investigate how many contigs on average are needed to reconstruct a gene. BLASTn query was the *de novo* contigs and the subject the *P. trichocarpa* CDS sequences (v156 Phytozome [31]), e-value=1E-5, perc_identity=90. The BLASTn output was the input for an home-made script which calculates how many contigs are necessary to reconstruct one gene.

Validation alignments

The validation alignments were carried out using the rNA software [26] with default values. 21,696,229 Mb of high quality POLI paired-end reads were used for the alignments. Paired-end insert size ranges between 500 and 600 bp. The count of chimeric reads was carried out exploiting the Picard command-line tool-CollectAlignmentSummaryMetrics.

Heterozygosity evaluation

Contig coverage was calculated using qaCompute, a package of the qaTools (https://github.com/CosteaPaul/qaTools).

Scaffolding

The SSPACE software [32] was used to scaffold the assembly contigs. Mate-pair reads were used for this step. Before using the mate-pair reads for the scaffolding, 454 Circularization Adaptor sequence (5'-TCGTATAACTTCGTATAATGTATGCTATACGAAGTTATTACG-3'; 3'-AGCATATTGAAGCATATTACATACGATATGCTTCAATAATGC-5') was removed from the reads. This was executed applying the following pipeline:

1. remove 454 Circularization Adaptor sequence using Cutadapt [33] and requiring a minimum match of 12 bp;

2. trim 12 bp from reads ends when the 454 Circularization Adaptor sequence is not found in the reads;

3. remove short reads (<15bp) and exclude the corresponding orphaned reads;

4. quality trimming (using Mott's trimming algorithm [25]) and contaminant removal (chloroplast and mitochondrial sequences) on the remaining reads using the rNA software [26];

5. align the reads *versus P. trichocarpa* with rNA, sort the alignment file for coordinate position using the Picard command-line tool-SortSam (http://picard.sourceforge.net/command-line-overview.shtml), remove PCR duplicates using the Picard command-line tool-MarkDuplicates (default values).

To evaluate the contamination level of forward-reverse reads in the hybrid 454-Illumina mate-pair library, the Picard command-line tool-CollectInsertSizeMetrics (default values).

Scaffolding validation

To determine the correctness of the scaffolding performed by SSPACE, a BLASTn analysis was carried out exploiting two *P. nigra* BAC sequences of the Ghoy genotype (BAC1=130,840 bp, BAC1 coordinates on the *P. trichocarpa* genome [1]: LG_I:5706420-5826004; BAC3=101,368 bp, BAC3 coordinates on the *P. trichocarpa* genome: LG_VI: 16793284-16895000). The BAC clones belong to a genomic BAC library of *P. nigra* x *P. deltoides* (Department of Soil & Crop Sciences and Institute for Plant Genomics & Biotechnology, Texas A&M University). The insert sequence of *P. nigra*-specific clones, representing random single-copy regions in the *Populus* genome, were provided by the Department of Agricultural Sciences, University of Udine.

In the BLASTn analysis (parameters used: e-value=1E-5, perc_identity=90), the scaffold sequences were the query and the *P. nigra* BAC sequences were the subject. Subsequently, the contig sequences composing the matching scaffolds were aligned singularly to the relative *P. nigra* BAC sequence using the bl2seq-BLASTn tool. The correctness of the contig order and orientation in the scaffold was determined comparing its coordinates and orientation on the BAC sequence.

In the cases in which a contig did not align to the BAC sequence, it was verified if its sequence aligns to the same genomic region of the BAC sequence on the *P. trichocarpa* genome or to a different genomic locus. The contig sequence was aligned to the *P. trichocarpa* genome sequence [1] on the Phytozome website [31] and the alignment coordinates were compared to those of the BAC on the *P. trichocarpa* genome sequence.

Comparative genomic analysis

A pipeline was designed to identify the genomic differences between *P. nigra de novo* assembly contigs and *P. trichocarpa* reference genome sequence. This pipeline works on a BLASTn output: the *de novo* assembly contigs are the query and the reference genome sequence is the subject. From the pipeline point of view, a contig is composed by one or more hits aligning to the reference, or by any hit. The pipeline classifies a *de novo* assembly contig into one of three different sets. The sets are defined considering the concept of *ratio* (contig alignment length/contig length) as follows:

- U_C: set of the *de novo* assembly contigs (C) which do not produce any significant hit against the reference genome sequence (R) in the BLASTn analysis, or produce one or more non contiguous, but relatively close hits which reconstruct the contig for a small portion ($ratio<0.3$). Two or more hits are defined as relatively close if they align within either 20,000 bp or five times the contig length if the contig is longer than 20,000 bp;

- $U_{CRtotal}$: set of the *de novo* assembly contigs which produce either a unique significant hit against the reference genome sequence with a $ratio \geq 0.95$, or non contiguous, but relatively close hits which reconstructs the contig almost for its whole length ($0.8 \leq ratio < 0.95$);

- $U_{Cpartial}$: set of the *de novo* assembly contigs which produce a set of non contiguous, but relatively close hits which partially reconstruct the contig ($0.3 \leq ratio < 0.8$).

The pipeline considers also the *P. trichocarpa* portions not covered between the contigs. These portions represent the *P. trichocarpa*-specific sequences and are classified into a set called U_{BH}.

For each pipeline set, except U_{BH}, the total length of the hits aligned to the reference (Total_hit_length) and the total length of the contig not-aligned portions (Total_not-aligned_contig_length) were calculated. The sum of Total_hit_length among the set U_C, $U_{CRtotal}$ and $U_{Cpartial}$ represents the total length of the shared sequence portions between the *P. nigra de novo* assembly and the *P. trichocarpa* genome, whereas the sum of Total_not-aligned_contig_length among the three sets represents the total length of the *P. nigra*-specific portions.

Comparative genomic analysis pipeline

The comparative genomic analysis pipeline works as follows:

1. Filter the *de novo* assembly contigs to remove mitochondrial and chloroplastic sequences;

2. BLASTn of the *de novo* assembly contigs *versus* the reference genome sequence;

3. anchor, if possible, each *de novo* assembly contig only to one scaffold of the reference genome sequence (described in the following subsection);

4. classify each anchored contig into a specific set described above. If a contig is classified into U_C, then remove it from the scaffold of the reference genome sequence where it was anchored;

5. classify every portion between two hits of the contig into the set U_{BH}.

Anchoring of a contig

C is the contig to be anchored and R the reference genome sequence subdivided into more chromosomes indicated as R_1, R_2,..., R_n.

h is a hit of C and it is defined as a quadruple h=(x_r,y_r; x_c,y_c) where x_r and y_r are the coordinates of the position of h on the reference genome sequence ($x_r > y_r$), and x_c and y_c are the coordinate of the position of h on the contig C ($x_c > y_c$).

Two hits h_1 and h_2 belonging to the same contig C and anchored to a chromosome R_i are ordered if:

- the sequences of the two hits h1 and h2 on R_i ($R_i[h_1.x_r , h_1.y_r]$ and $R_i[h_2.x_r , h_2.y_r]$) respect the contig order.

Two hits h_1 and h_2 belonging to the same contig C and anchored to a chromosome R_i are separated if:

- $h_1.y_r > h_2.x_r$;

- $h_1.y_c > h_2.y_c$.

The max-span of a contig C is the maximum portion of the reference genome sequence that C can cover without nullifying the alignment accuracy. It is defined as max (20.000, 5*|C|).

The contig-span of an anchored contig C is the reference genome sequence portion included between the first hit start and the last hit end.

Considering the BLASTn output, to anchor a contig C:

1. discard all the hits with a percentage of identity minor to a fixed threshold and length <100 bp;

2. calculate the max-span;

3. for every R_i: run R_i using a window of length max-span and calculate for every point the total length of the hits (hits-length) belonging to that interval;

4. anchor the contig at the chromosome position with maximum hits-length. If this position is called X, then the reference genome portion covered by C is (X-X+contig-span).

Characterisation of the pipeline sets

Repetitive sequence (RS) content

The software RepeatMasker [29] was used to establish the RS content of the pipeline sets U_C, $U_{CRtotal}$ and $U_{Cpartial}$. Parameters used: -no_is (skips bacterial insertion element check), -nolow (does not mask low complexity DNA or simple repeats), -qq (about 10% less sensitive than default). The RS library was specific for transposable and repetitive elements of different plant species including the sequences of RepBase-plants (v 16.02), Triticeae Repeat Sequence Database, Plant Repeat Databases of the TIGR Institute at Michigan State University, known *P. trichocarpa* transposable and repetitive elements [30], predicted *Populus* repetitive elements by RepeatScout [34], and some *Vitis* and *Prunus* repetitive sequences.

Protein content

A protein sequence database was created to characterise the protein content of the pipeline sets U_C, $U_{CRtotal}$ and $U_{Cpartial}$. The database was constituted by the protein sequences [31] of five different species: *Arabidopsis thaliana*, *Oryza sativa*, *Populus trichocarpa*, *Prunus persica* and *Vitis vinifera*. A BLASTx analysis was performed for each pipeline set (e-value=1E-05). To calculate the percentage of matching contigs, only the contigs with a “Positives” value $\geq$ 90% were considered in the BLASTx output.

7

Results

De novo assembly

Input data for the assembly

A single male genotype, "POLI", was selected and used in Illumina paired-end sequencing and *de novo* assembly strategy. Three different kinds of sequencing libraries were produced (see BOX 1 at page 62 for terminology description): standard paired-end, overlapping paired-end (producing a single merged read) and mate-pair. Insert size, read length, number of reads and total read length for each library are reported in Table 7.1. The table contains also the statistics for the overlapping paired-end library after the two overlapping reads were merged, obtaining a single long read.

The quality of the standard paired-end libraries was estimated by exploiting the k-mer analysis. An example of k-mer plot of high quality libraries is shown in Figure 7.1. As *Populus* genus underwent ancient genome duplication [1] and POLI is an heterozygous genotype, in the k-mer plot three peaks were expected: one at expected coverage indicating the homozygous sequences, one at 0.5X expected coverage indicating the heterozygous sequences and one at 2X expected coverage indicating the anciently duplicated sequences. In the k-mer plot of the low quality standard paired-end libraries only the peak at the expected coverage was detectable (data not shown). The best performing standard paired-end library, in terms of quality, was selected as input data for the *de novo* assembly.

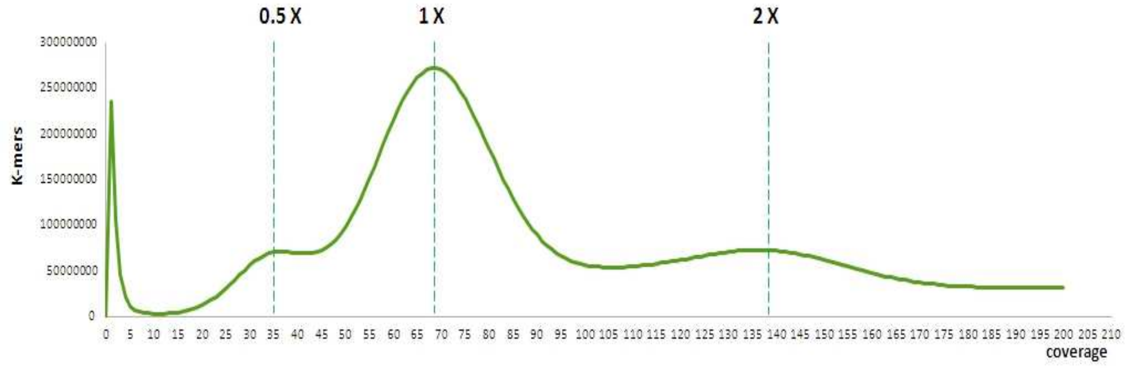


Figure 7.1: K-mer (16 bp) plot of a high quality standard paired-end library.

Before using 454-Illumina hybrid mate-pair reads for the assembly, they underwent the removal of both the 454 Circularization Adaptor sequence and the PCR duplicates. The two applied filters decreased the initial number of mate-pair reads as reported in Table 7.2. Moreover, the forward-reverse paired-end contamination was estimated (Figure 7.2). It resulted to be 14%.

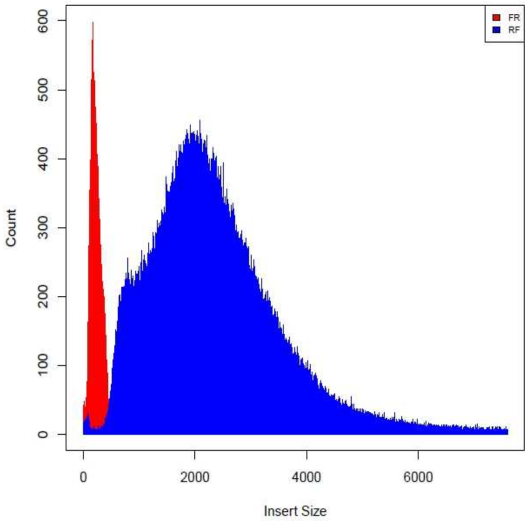
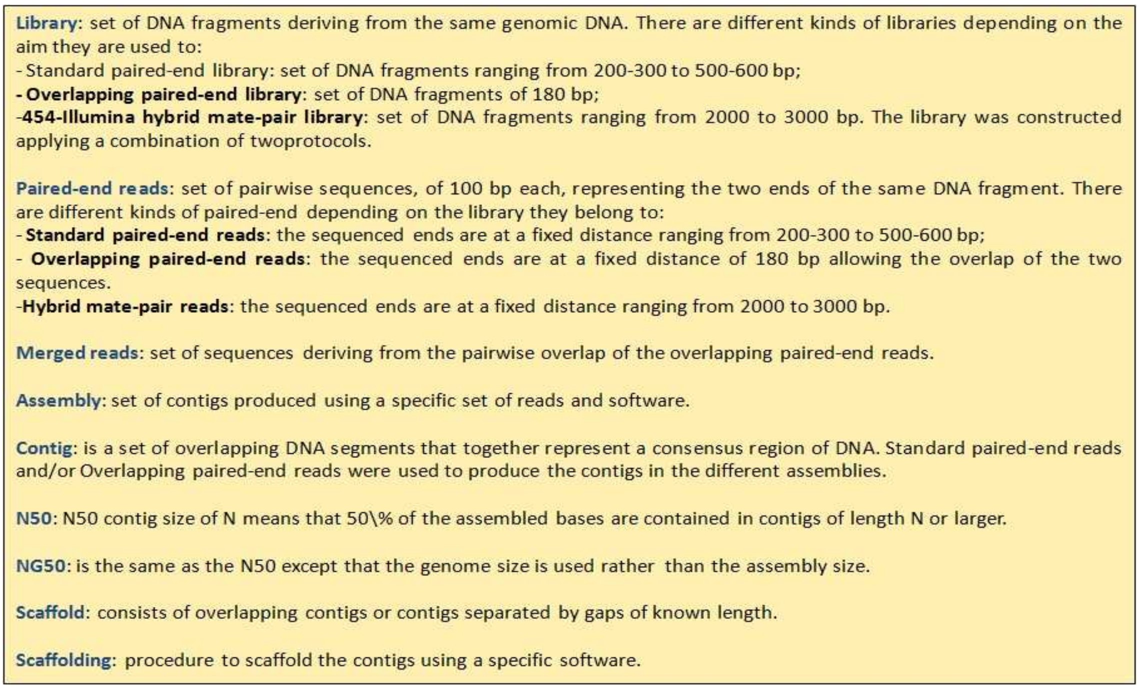


Figure 7.2: Contamination level of forward-reverse reads in a hybrid 454-Illumina mate-pair library.



Library: set of DNA fragments deriving from the same genomic DNA. There are different kinds of libraries depending on the aim they are used to:
- Standard paired-end library: set of DNA fragments ranging from 200-300 to 500-600 bp;
- **Overlapping paired-end library**: set of DNA fragments of 180 bp;
-**454-Illumina hybrid mate-pair library**: set of DNA fragments ranging from 2000 to 3000 bp. The library was constructed applying a combination of twoprotocols.

Paired-end reads: set of pairwise sequences, of 100 bp each, representing the two ends of the same DNA fragment. There are different kinds of paired-end depending on the library they belong to:
- **Standard paired-end reads**: the sequenced ends are at a fixed distance ranging from 200-300 to 500-600 bp;
- **Overlapping paired-end reads**: the sequenced ends are at a fixed distance of 180 bp allowing the overlap of the two sequences.
-**Hybrid mate-pair reads**: the sequenced ends are at a fixed distance ranging from 2000 to 3000 bp.

Merged reads: set of sequences deriving from the pairwise overlap of the overlapping paired-end reads.

Assembly: set of contigs produced using a specific set of reads and software.

Contig: is a set of overlapping DNA segments that together represent a consensus region of DNA. Standard paired-end reads and/or Overlapping paired-end reads were used to produce the contigs in the different assemblies.

N50: N50 contig size of N means that 50\% of the assembled bases are contained in contigs of length N or larger.

NG50: is the same as the N50 except that the genome size is used rather than the assembly size.

Scaffold: consists of overlapping contigs or contigs separated by gaps of known length.

Scaffolding: procedure to scaffold the contigs using a specific software.

Figure 7.3: BOX 1-Terminology summary.

Table 7.1: Paired-end sequencing libraries produced for the *de novo* assembly of POLI.

Library type	AVG insert size (bp)	Read length (bp)	Number of reads	Total read length (bp)	Raw coverage (X)
Standard paired-end	500	101	378,736,750	38,252,411,750	91.7
Overlapping paired-end	180	110	173,442,946	19,078,724,060	45.7
Overlapping paired-end after the merge	180	180	62,706,893	11,202,286,729	26.9
454-Illumina hybrid mate-pair	2200	101	16,281,308	1,644,412,108	3.9

Table 7.2: 454-Illumina hybrid mate-pair library statistics.

Library type	Number of reads	Total read length	Raw coverage (X)	% of duplicated reads
454-Illumina hybrid mate-pair	16,281,308	1,644,412,108	3.9	21
After adaptor removal	12,909,224	~1,303,831,624	~3.1	21
After adaptor and duplicated reads removal	10,616,968	~1,072,313,768	~2.6	9

Assemblies performed

Two *de novo* assemblers, CLC Genomics Workbench and ABySS, were tested combining different sequencing libraries in order to evaluate their performances related to different insert size, read orientation and length. Performed assembly combinations are described in Table 7.3.

Table 7.3: Assemblies performed with different combinations of sequencing libraries.

Input data	CLC	ABySS
Standard paired-end (Std pe)	X	X
Standard paired-end + merged reads (Std pe + mr)	X	X
Standard paired-end + hybrid mate-pair (Std pe + hmp)	X	X

Statistics were produced for each assembly combination as reported in Table 7.4. A comparison of "NG50 size" and "Number of contigs" in the different assemblies is reported in Figure 7.4. For each combination of input data, the assembler ABySS resulted the best performing in terms of NG50 values, number of contigs, maximum and mean contig length. Assembly length cannot be directly compared. Although the assemblies obtained by the combination of Standard paired-end and Hybrid mate-pair reads gave the best results, they were not considered for subsequent analyses as we decided to use the Hybrid mate-pair reads only in the scaffolding phase.

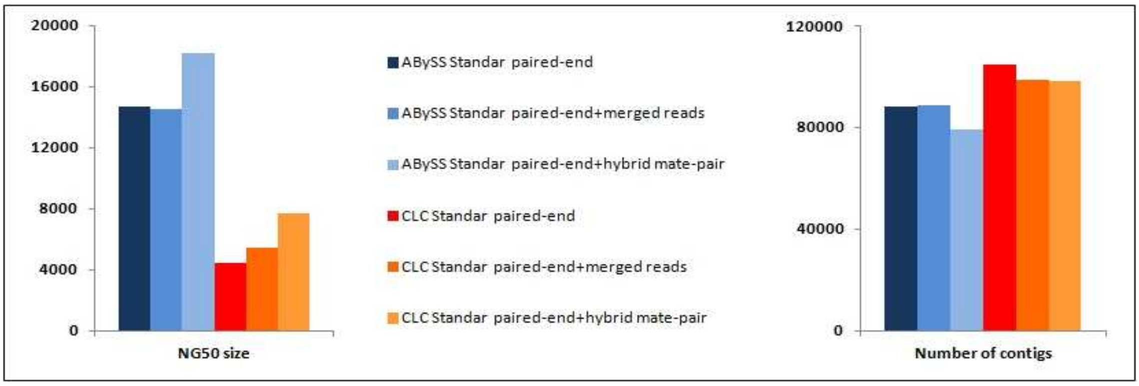


Figure 7.4: Comparison of "NG50 size" and "Number of contigs" in the different assemblies.

A third assembler, GAM [27], underwent preliminary tests. As this assembler merges two assemblies, two experimental tests were performed. The first one considered as master the CLC *de novo* assembly obtained from the combination of Standard paired-end reads and merged reads, and as slave the ABySS *de novo* assembly, obtained only using the Standard paired-end reads. The second experimental test was performed inverting the *de novo* assemblies used as master and slave. The two assemblies were selected as resulting the most performing assemblies obtained with CLC and ABySS. Statistics of GAM assemblies are shown in Table 7.5. GAM resulted to be effective in improving the two master assemblies, as demonstrated by the improvements in all parameters used to describe the assembly.

Table 7.4: Statistics concerning each *de novo* assembly combination considering input data and assembler used.

Parameters	CLC_Std pe	ABySS_Std pe	CLC_Std pe+mr	ABySS_Std pe+mr	CLC_Std pe+hmp	ABySS_Std pe+hmp
N50 size (bp)	6,130	11,768	7,034	11,620	8,825	14,339
NG50 size (bp)	4,434	14,677	5,443	14,482	7,711	18,218
Number of contigs	104,431	88,192	98,438	88,865	98,305	79,103
Assembly length (bp)	339,551,115	526,633,186	352,379,839	527,215,361	383,905,113	527,487,677
Maximum contig length (bp)	97,096	102,955	120,870	103,018	183,822	228,444
Mean contig length (bp)	3,251	5,971	3,579	5,932	3,905	6,668

Table 7.5: Statistics concerning GAM assemblies. In brackets the statistics for the assembly used as master are reported.

Parameters	CLC as master	ABySS as master
N50 size (bp)	13,569 (7,034)	14,438 (11,768)
NG50 size (bp)	14,242 (5,443)	17,493 (14,677)
Number of contigs	83,397 (98,438)	73,953 (88,192)
Assembly length (bp)	433,027,504 (352,379,839)	508,199,447 (526,633,186)
Maximum contig length (bp)	146,628 (120,870)	124,008 (102,955)
Mean contig length (bp)	5,192 (3,579)	6,871 (5,971)

Assembly validation

Gene reconstruction

In order to verify the completeness and integrity of the assemblies reported in Table 7.6 we determined for each of them how many genes have been reconstructed in a single contig. The 45,033 *P. trichocarpa* genes (CDS sequence, Phytozome156) were used for this analysis. A contig, to be considered for the gene reconstruction, had to cover at least the 80% of the gene length. The same figures are shown as percentages in Figure 7.5.

Table 7.6: Gene reconstruction (more than 80%) using only one contig.

Assembly	Total gene number rec.	Total gene length rec. (bp)
ABySS Standard paired-end	5,849	5,545,476
CLC Standard paired-end	10,471	9,903,999
CLC Standard paired-end+merged reads	5,901	6,662,201
GAM ABySS master-CLC slave	6,413	6,223,190
GAM CLC master-ABySS slave	8,321	8,141,319

The CLC_Standard paired-end assembly resulted being definitely more effective in reconstructing *P. trichocarpa* genes in a single contig.

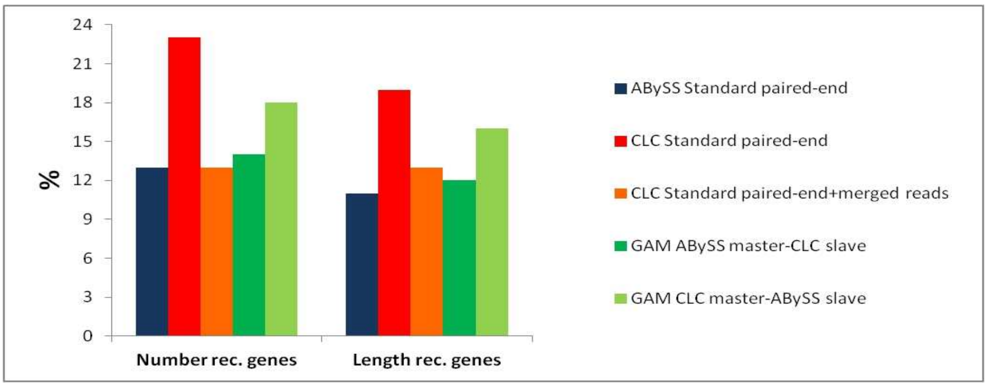


Figure 7.5: Percentages of genes and of gene length reconstructed by only single contig in each assembly.

Alignments

21,696,229 POLI high quality paired-end reads were selected to be mapped against the *P. trichocarpa* genome sequence and three *de novo* assemblies in order to evaluate how correct the assemblies are. The most accurate assembly of each assembler, considering the experimental test on gene reconstruction, was chosen as reference sequence. Percentages of paired-end and chimeric reads aligned both to the reference sequence and to the three selected assemblies are reported in Figure 7.6. The *P. trichocarpa* genome sequence was used as reference for alignment comparison. Although the percentage of aligned paired-end reads was similar in the three de novo assemblies, the performance of the CLC_Standard paired-end assembly was superior to the others as it had the lowest percentage of chimeric reads mapped, indicating that the CLC assembly allowed a more accurate alignment.

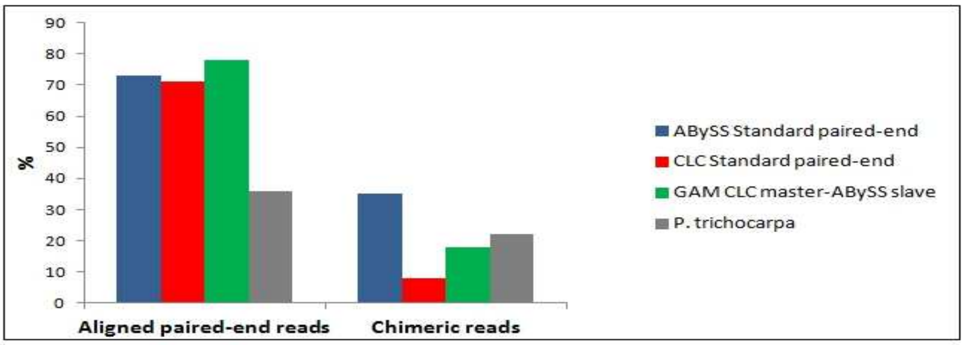


Figure 7.6: Percentages of both paired-end reads aligned to the four selected reference sequences and chimeric reads.

Heterozygosity evaluation

The same standard paired-end reads, used to produce both the CLC_Standard paired-end assembly and the ABySS_Standard paired-end one, were mapped against the two assemblies to verify the assembly contig coverage. The plots shown in Figure 7.7 represent the contig coverage distribution of the two assemblies. As POLI is an heterozygous genotype, two peaks were expected in the plot if the assemblers had identified distinct contigs for each haplotype: one at the assembly average coverage (~75X) and one at 0.5X average coverage (~37X). The peak at the assembly average coverage indicates homozygous regions and heterozygous regions that were assembled into a single contig. The peak at 0.5X assembly average coverage indicates heterozygous regions that were assembled separately and hemizygous regions. The peak at 2X assembly average coverage, indicating the duplicated regions, is not visible as multiple alignments of the reads were not allowed. In the plot of the CLC_Standard paired-end assembly, the two peaks are distinct; the peak at 0-4X indicates the misassembled contigs. A different situation is shown in the plot of the ABySS_Standard paired-end assembly, in which the peak at 0-4X indicates a very high number of misassembled contigs. Moreover, the number of contigs at the assembly mean coverage is smaller than that of the CLC_Standard paired-end assembly as a large number of reads were spread onto the misassembled contigs.

For the contigs with coverages between 0-4X, 25-40X and 65-85X, the statistics reported in the Table 7.7 and 7.8 were calculated. Very discordant values were observed

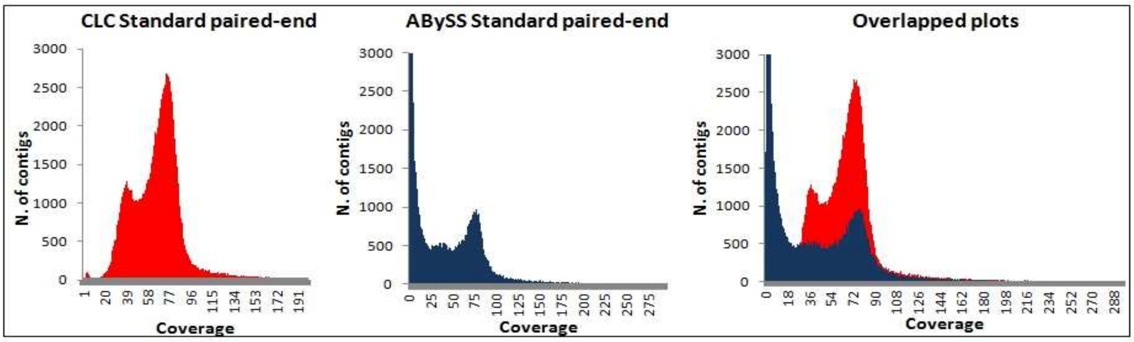


Figure 7.7: Contig coverage distribution. In each plot, the peak on the left indicates the number of misassembled contigs, the peak on the right is the peak at the assembly average coverage, whereas the peak between the two is the peak at 0.5X assembly average coverage.

Table 7.7: Statistics concerning the misassembled contigs in the CLC_Standard paired-end and ABySS_Standard paired-end assembly.

Parameters	CLC_0-4X	ABySS_0-4X
Number of contigs	134	22,542
Total contig length (bp)	113,974	235,189,229
Mean contig length (bp)	856	10,433

Table 7.8: Statistics concerning the contigs representing the heterozygous and hemizygous regions, and those representing homozygous regions and similar regions uniquely assembled in the CLC_Standard paired-end and ABySS_Standard paired-end assembly.

Parameters	CLC_25-40X	ABySS_25-40X	CLC_65-85X	ABySS_65-85X
Number of contigs	12,914	8,018	46,103	16,792
Total contig length (bp)	16,355,386	26,829,972	222,300,563	121,351,351
Mean contig length (bp)	1,266	3,346	4,821	7,227

comparing the two assemblies. ABySS misassembled contigs were 22,542 (total length of 235,189,229 bp) with mean length of 10,433 bp, whereas the CLC ones were only 134 (total length of 113,974 bp) with mean length of 856 bp. Discordance was also observed comparing the statistics of the contigs representing the homozygous regions and the similar regions uniquely assembled. These contigs, in the ABySS assembly, were 16,792 (total length of 121,351,351 bp) with mean length of 7,227 bp, instead in the CLC assembly were 46,103 (total length of 222,300,563 bp) with mean length of 4,821. The statistics of the contigs representing the heterozygous and hemizygous were also discordant between the two assemblies. It is interesting to notice the different proportion of contigs representing the "homozygous" and "heterozygous" regions in the two assemblies. In the ABySS assembly, the number of contigs representing "heterozygous" regions are half of those representing "homozygous" regions, whereas in the CLC assembly they are less than one-third.

From the analyses of the number of genes reconstructed by one contig, the read alignments against the four reference sequences and the plots in Figure 7.7 and the statistics reported in Table 7.7 and 7.8, it is clearly evident the CLC_Standard paired-end assembly is more reliable than the ABySS_Standard paired-end. Given these results, the CLC_Standard paired-end assembly was selected for the comparative genomic analysis between *P. nigra* and *P. trichocarpa*.

Scaffolding

Scaffolding was performed in the CLC_Standard paired-end assembly. Standard paired-end reads and 454-Illumina hybrid mate-pair reads, after the removal of the 454 Circularization Adaptor sequences and PCR duplicates (see Table 7.2), were used to scaffold the contigs. Results are presented in Table 7.9.

Table 7.9: Statistics concerning the scaffolding of the CLC_Standard paired-end assembly.

Parameters	CLC
N50 size	22,648
NG50 size	16,544
Number of scaffolds	59,569
Assembly length	349,935,450
Maximum scaffold length	318,552
Mean scaffold length	5,874

Scaffolding validation

Two *P. nigra* BAC sequences, belonging to the GHOY genotype, were used to verify order and orientation of the contigs scaffolded by the SSPACE software in the CLC_Standard paired-end assembly. 10 scaffolds, constituted by 57 contigs in total, were considered for the validation since they could reconstruct the BAC insert sequences. Among the 57 contigs, 37 contigs related to 9 scaffolds matched with the BAC sequences confirming their exact order and orientation within their scaffolds. Other 12 contigs, related to the same 9 scaffolds, represented sequences outside the BAC borders. However, we deduced they were in the right order within their scaffolds on the basis of their coordinates on the *P. trichocarpa* genome sequence given its similarity to the *P. nigra* genome.

1 scaffold out of 10 (10%) could be considered wrongly built as only one out of its eigth contigs matched to the BAC sequence, whereas all the others aligned consecutively to a different genomic region when blasted on the *P. trichocarpa* genome sequence.

Comparative genomic analysis

Comparative genomic analysis pipeline

A pipeline was designed to identify the genomic differences between *P. nigra* CLC_Standard paired-end *de novo* assembly contigs and *P. trichocarpa* genome sequence. CLC_Standard paired-end assembly was selected for this analysis as it resulted to be the most accurate among the assemblies produced. The pipeline runs on a BLASTn output where the selected assembly was aligned against the *P. trichocarpa* genome sequence. To establish the appropiate percentage of identity to consider as correct the alignment of a contig hit to the *P. trichocarpa* genome, different identity percentage thresholds were tested in the pipeline. From the pipeline output and for each identity percentage threshold, we calculated the shared sequence portion between the two genomes (sum of Total_hit_length among the set U_C, $\mathrm{U}_{CRtotal}$ and $\mathrm{U}_{Cpartial}$), the *P. nigra*-specific portion (sum of Total_not-aligned_contig_length among the three sets) and the *P. trichocarpa*-specific sequence portion (U_{BH}). In Table 7.10 the Total_hit_length of the pipeline sets and the length of the set U_{BH} are reported for the different identity percentage thresholds; the Total_not-aligned_contig_length is reported in Table 7.11.

Table 7.10: Statistics concerning the Total_hit_length of each pipeline set relatively to CLC_Standard paired-end assembly.

Identity percentage threshold	$\mathrm{U}_{CRtotal}$ (bp)	$\mathrm{U}_{Cpartial}$ (bp)	U_C (bp)	U_{BH} (bp)
80%	51,338,187	39,974,774	1,330,146	51,338,187
82.5%	252,629,667	40,796,181	1,544,761	53,503,725
85%	240,244,578	43,540,747	1,946,338	58,466,125
87.5%	220,716,241	49,769,208	2,747,270	69,192,340
90%	182,173,910	60,929,665	4,294,481	91,494,546
92.5%	126,067,695	67,915,760	6,603,225	138,966,183
95%	69,949,481	54,455,180	8,093,802	222,752,088
97.5%	16,174,681	15,003,495	5,605,518	345,559,988

Table 7.11: Statistics concerning the Total_not-aligned_contig_length of each pipeline set relatively to CLC_Standard paired-end assembly.

Identity percentage threshold	$\mathrm{U}_{CRtotal}$ (bp)	$\mathrm{U}_{Cpartial}$ (bp)	U_C (bp)
80%	13,949,335	23,596,985	8,070,208
82.5%	13,903,888	24,164,321	10,000,325
85%	13,775,898	26,138,624	13,904,930
87.5%	13,365,356	30,564,188	22,388,852
90%	11,743,800	38,646,101	41,763,158
92.5%	8,243,420	47,414,199	83,306,816
95%	4,465,302	42,574,176	160,013,174
97.5%	764,422	14,121,579	287,881,420

To establish the correct identity percentage threshold to identify the real shared sequence portion between our *P. nigra de novo* assembly and the *P. trichocarpa* genome, we concentrated on the *P. trichocarpa* genome reconstruction as described in the following paragraph.

P. trichocarpa genome reconstruction

The *P. trichocarpa* genome length is ~500 Mb [1] but the assembled genome length is 417,137,944 bp. Comparing the length of the *P. trichocarpa* assembled genome to our *P. nigra de novo* assembly length (339,551,115 bp), our *P. nigra de novo* assembly is

77,586,829 bp shorter. From this evidence, the *P. trichocarpa* assembled genome is calculated as the sum of the shared sequence portion between the genomes of the two species, the *P. trichocarpa*-specific sequence portion and the 77,586,829 bp. Depending on the identity percentage threshold applied to the pipeline, the shared sequence portion between the genomes and the *P. trichocarpa*-specific sequence portion resulted variable. This is shown in the plot of Figure 7.8 representig the length of *P. trichocarpa* assembled genome at the different identity percentage thresholds. The black arrow in the plot indicates the identity percentage threshold to be considered for the more precise *P. trichocarpa* assembled genome reconstruction. At the identity percentage threshold of 92.5%, the length of the reconstructed *P. trichocarpa* assembled genome is 417,139,692 bp.

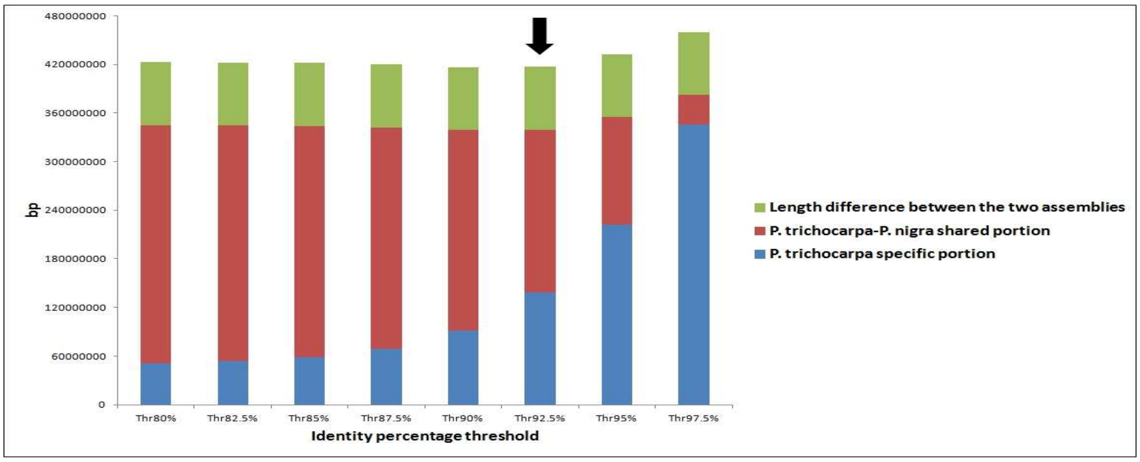


Figure 7.8: Size trend of the *P. trichocarpa* assembled genome depending on the length of the shared sequence between the *P. trichocarpa* assembled genome and the *P. nigra de novo* assembly.

The pipeline error rate, considering an error when the pipeline assigned a contig to the wrong set, was calculated for two identity percentage thresholds: 90 and 95%. The pipeline validation was conducted by blasting, in the Phytozome website, 270 *de novo* assembly contigs, distributed among the different pipeline sets. When the Phytozome *ratio* was not in accordance with the pipeline *ratio* interval, then the contig was considered wrongly assigned to the pipeline set. The pipeline error rate using a identity percentage thresholds of 90% was 10%, whereas the error rate using a identity percentage thresholds of 95% was 1%. Considering the results of the pipeline tests at the different identity percentage thresholds, the length of the reconstructed *P. trichocarpa* assembled genome and the error rate, we decided to run the pipeline using an identity percentage threshold of 92.5% for the pipeline applications in poplar. At this threshold, the amount of shared portion between the two genomes, resulted to be equal to 200,586,680 bp (Table 7.10), while the *P. nigra*-specific portion resulted to be equal to 138,964,435 bp (Table 7.11).

Characterisation of the pipeline set content

The pipeline sets U_C, $U_{Cpartial}$ and $U_{CRtotal}$, containing *P. nigra*-specific sequences in different proportions, were investigated for transposable and repetitive elements (repetitive sequences, RS), and protein-coding potential. RepeatMasker was used to mask the three pipeline sets with a library containing plant transposons and repetitive elements, and a BLASTx analysis, with a plant protein database, was carried out to determine the protein-coding potential.

In Figure 7.9, the different distribution of RSs and protein-coding potential is reported. It is clearly evident that the U_C set had a larger RS content (51% of its total sequence information) compared to the other two sets. This result was expected as the U_C set, for definition, collects those contigs sharing a maximum of 30% of their sequence with the *P. trichocarpa* genome sequence. This means the contigs likely represent *P. nigra*-specific repetitive sequences. This hypothesis is also supported by the low protein-coding potential (7%) among the U_C contigs.

$\text{U}_{Cpartial}$ and $\text{U}_{CRtotal}$ had a comparable RS content. Indeed $\text{U}_{Cpartial}$ contains the contigs which share from 30 to 80% of their sequence with the *P. trichocarpa* genome, while $\text{U}_{CRtotal}$ contains the contigs sharing the sequence from their 80 to 100%. However, although the RS content is comparable between the two sets, $\text{U}_{CRtotal}$ has a smaller RS content. This is due to its larger protein-coding potential content which justifies the higher similarity between its contigs and the *P. trichocarpa* genome.

Alignment of the *P. nigra de novo* assembly to the *P. trichocarpa* genome

The trend of the BLASTn alignment of the *P. nigra de novo* assembly to the *P. trichocarpa* genome sequence (Figure 7.10) was compared to the trend of the *P. trichocarpa* gene distribution and repetitiveness of its genome. In Figure 7.10, each black spike of the plot represents a *P. trichocarpa* genomic region of 100 kb covered from 0 to 100% of its length by the *de novo* assembly sequence. On average, the 100 kb regions were covered for the 50%. However, chromosome regions with high gene content were covered, in general, for more than 50%, whereas those highly repetitive were covered for less than 20%. This trend confirms that assembling repetitive sequences is a problematic issue.

The first 2 million base pairs in the chromosome 8 might represent an interesting region to be analysed as it has high gene content and low repetitiveness in but it is only partially covered by the *P. nigra* contigs.

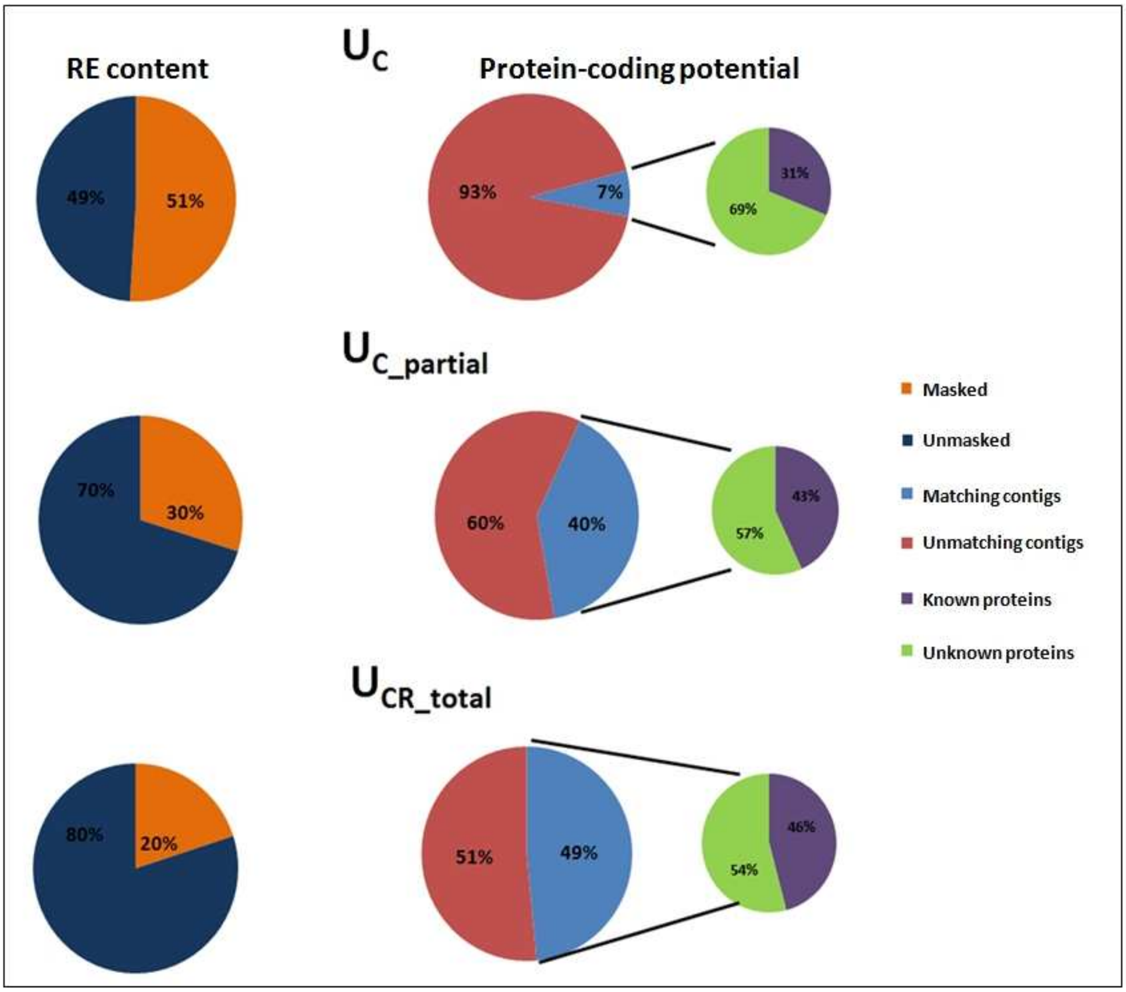


Figure 7.9: Characterisation of the pipeline set content in terms of repetitive sequences and protein-coding potential. The two analyses were performed using respectively RepeatMasker and the BLASTx tool.

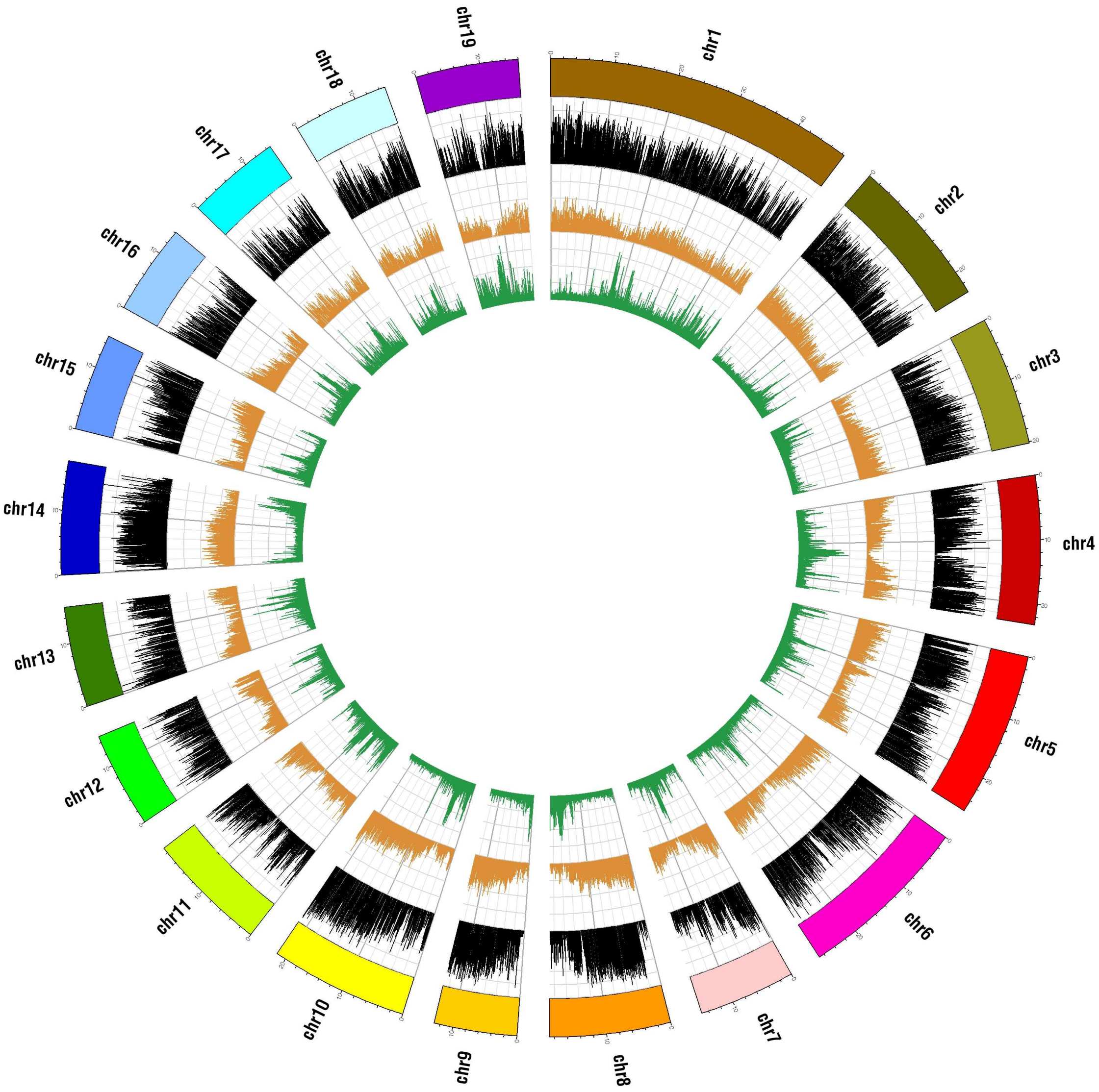


Figure 7.10: Alignment of the *P. nigra de novo* assembly to the *P. trichocarpa* genome sequence. Colored bars=*Populus* 19 chromosomes (unit=2 Mb). In the plot: outer layer=black spikes represent *P. trichocarpa* genomic regions of 100 kb covered from 0 to 100% of its length by the *de novo* assembly sequence (vertical unit=0.2, spanning from 0 to 1); second layer=number of genes every 100 kb (vertical unit=10, spanning from 0 to 40; third layer=repetitiveness of the genome calculated with a k-mer analysis using the tool Tallymer [43] (vertical unit=2, spanning from 0 to 8).

8
Discussion

The work presented in this part of the PhD thesis defines an accurate *de novo* assembly approach to obtain the *Populus nigra* genome sequence exploiting solely the Illumina sequencing technology. Moreover, it also provides a useful pipeline to compare two closely related genomic sequences in order to classify the *de novo* assembly contig content in terms of novelty or similarity between the two genomes.

Input data for the assembly

In literature there are several examples of *de novo* assemblies performed with a mixture of input data [12, 13, 16]. It is quite common to combine short-read data with 454 and Sanger data to fill in repeat-induced gaps and scaffold the contigs. The first *de novo* assembly exclusively obtained through short-reads sequencing, belongs to the giant panda, *Ailuropoda melanoleura* [14]. Redundant coverage, in which on average every nucleotide is sequenced many times over, is required to produce a high-quality assembly. Another benefit of redundancy is greatly increased accuracy compared with a single read: where a single read might have an error rate of 1%, eightfold coverage has an error rate as low as 10^{-16} when eight high-quality reads agree with one another. High coverage is also necessary to sequence polymorphic alleles within diploid or polyploid genomes [11]. To assemble *de novo* the *P. nigra* heterozygous genome, we produced three different kinds of Illumina data for a total raw coverage of ~ 123 X. Three-quarters of the raw coverage was constitued by standard paired-end reads to provide a high representation of the genome, while almost one quarter was constitued by overlapping single reads. These reads were produced from paired-end reads with a small insert size which allowed the overlap of the pair reads [22] with subsequent merger. The aim of the merged single reads was to solve small genome repetitions. In general, longer reads make better assemblies because they span more repeats [11]. The third kind of data was represented by long-insert paired-end reads (mate-pair reads). Mate-pair reads have a reverse-forward orientation, contrary to the standard paired-end reads, and in general are used in the scaffolding process which consists in ordering and orienting the contigs. In our assembly the mate-pair reads were exclusively used for scaffolding purpose. We obtained the mate-pair reads combining two different protocols to increase the library performance reducing both the contamination by forward-reverse paired-end reads and the mate-pair read redundancy (same pairs occurred repeatedly). Indeed, comparing the performance of the hybrid 454-Illumina protocol to the Illumina one (data not shown), the first one resulted being better for both aspects. In general, although the redundancy can be reduced *a posteriori* using, for example, Picard command-line tools, it is important to start from a low redundant library, otherwise, the input data will decrease sensibly. An example was given by the panda genome project [14]: the redundancy of the 10-kb library was 77%, implying that the actual coverage was just under one-fourth of that indicated by the total sequence length.

Others important aspects to keep in mind when performing a *de novo* assembly concern the use of distinct, nonchimeric pair-end reads [11] and high quality input data. Chimeric pairs are paired sequences not at the expected separation and orientation in the genome. Obviously, high percentages of chimeric reads in the input data cause wrong contig construction joining sequence portions distant on the genome sequence. An analogous scenario is expected when the chimeric reads are part of the read set used in the scaffolding phase. Here, the chimeric reads will closely join contigs which are separated by ten of thousands base pair or will join contigs in wrong orientation. Quality of the input data is also important for the contig reconstruction: low quality reads, containing sequencing errors, can lead to a not proper assembly of the contig.

Assemblies performed

Nowadays many *de novo* assemblers are available. Some of the most widely used assemblers for Illumina data are: CLC Genomics Workbench (http://www.clcbio.com), ABySS [23], ALLPATHS-LG [35], SOAPdenovo [36], CABOG [37]. We assembled the *P. nigra* genome using the softwares CLC and ABySS to have an alternative in terms of assembly quality. We tried different combinations of input data for both assemblers (Table 7.3) to select the best performing combination for our genome sequence. The initial parameters considered to measure the assembly quality were the NG50 size, the number of contigs obtained and their mean length. For each combination of input data, ABySS demonstrated a better performance for all the three parameters (Figure 7.4).

A third assembler, GAM [27], was partially tested. This software merges two assemblies obtained with different softwares. The two assemblies selected to be merged by GAM, CLC_Standar paired-end+merged reads and ABySS_Standar paired-end, were chosen as more performing on the basis of a first quality evaluation. Both selected assemblies were used alternatively as master in GAM. Between the two GAM assemblies, and among the others assemblies, the combination using ABySS_Standar paired-end as master, presented a better NG50 size and fewer number of contigs. A premature conclusion, from the results discussed so far, could be that ABySS and GAM had better performances in assembling the *P. nigra* genome. This was denied by our assembly validation analysis.

Assembly validation

Assembly validation is a critical step. Most of the traditional metrics used to evaluate assemblies (N50, NG50, mean contig size, etc.) emphasize only size, while nothing (or almost nothing) is said about how correct the assemblies are [38]. Being aware of this problem, we went further in validating the assemblies reported in Table 7.6 analysing the number of genes reconstructed for at least their 80% length by one contig, on the bases of the *P. trichocarpa* CDS sequence. Moreover, we calculated the total reconstructed gene length using, again, only one contig per gene. We surprisingly noticed that single contigs of CLC_Standard paired-end assembly reconstruct a higher number of genes compared to the other four assemblies. The same happened considering the total reconstructed gene length. That meant CLC_Standard paired-end assembly contigs were more correctly constructed.

Although the merged reads should improve the quality assembly, the CLC_Standar paired-end+merged reads assembly had larger contigs but low quality. This fact might be due to a lower quality of the standard paired-end reads used as overlapping standard paired-end than those used simply as standard paired-end. Moreover, it must be noticed that the GAM_CLC-master assembly showed a better result than the GAM_ABySS-master (in contrast with what demonstrated by the parameters NG50 and number of contig). A possible explanation to this observation is that when GAM considers ABySS_Standard paired-end assembly as master, it merges the contigs in a wrong order imitating what has been assembled in ABySS_Standard paired-end.

A second evidence, confirming CLC_Standard paired-end as the most correct assembly, was obtained mapping ~21 Mb of POLI good quality paired-end reads against the CLC_Standard paired-end assembly, the two more performing assemblies obtained with ABySS and GAM (ABySS_Standard paired-end and GAM_CLC-master) and the *P. trichocarpa* genome sequence. Although the GAM_CLC-master assembly and the ABySS_Standard paired-end had the two higher percentage of paired-end reads mapped against them, they presented the two higher percentages of mapped chimeric reads comparing. Again, this result confirmed that ABySS wrongly constructed the *P. nigra* contigs. As it is demonstrated in a recent published work of Salzberg *et al.* [39], when two contigs are erroneously concatenated, the resulting assembly has larger contigs, but the assembly is worse.

In this work we also considered the assembly problem caused by the heterozygousity of our genotype. Because assembly critically relies on finding perfect (or nearly perfect) overlaps between individual reads, the *de novo* assembly approach is more efficient for species with haploid or nearly homozygous diploid genomes with relatively few repetitive sequences. However, the *de novo* assembly approach is also applied to genomes with large natural effective population sizes and considerable genetic diversity. In particular, genomes that are a mosaic of homozygous and heterozygous regions are particularly challenging, as paralogous genes and alleles can be easily confused in the short reads. POLI is an heterozygous genotype and the genus *Populus* underwent ancient genome duplication [1]. The first characteristic was observed in our heterozygousity analysis (Figure 7.7). The results allowed us to conclude the CLC assembler had a better performance in reconstructing the heterozygous regions of the genome compared to ABySS (the causes were not investigated as it was not the aim of this study) and produced a good quality assembly as the number of misassembled contigs was very small (134 contigs for a total length of 113,974 bp). The observed misassembled contigs might be due to sequencing errors.

Scaffolding

The scaffolding phase of assembly focuses on resolving repeats by linking the initial contigs into scaffolds, guided by mate-pair data. A scaffold is a collection of contigs linked by mate pairs, in which the gaps between contigs may represent either repeats, in which case the gap can in theory be filled with one or more copies of the repeat,or true gaps in which the original sequencing project did not capture the sequence needed to fill the gap. If the mate pair distances are long enough, they permit the assembler to link contigs across almost all repeats [11]. During the scaffolding phase "misjoin" errors can occur. A misjoin occurs when the assembler incorrectly joins two distant loci of the genome, which most often occurs within a repeat sequence, causing a significant structural error [39]. Three kind of misjoins can be in general observed: (1) inversions, where a contig is reversed with respect to the true genome; (2) relocations, or rearrangements that move a contig within a chromosome; and (3) translocations, or rearrangements between chromosomes. Having a high quality and reliably scaffolded *de novo* assembly is a powerful instrument for functional annotation of the genome [40] and for the identification of structural variants, in terms of copy, content and structure [41]. The scaffolding we performed on the CLC_Standar paired-end assembly showed to be effective in joining the contigs as the error rate was 10%. The misjoing error was ascrivable to a putative translocation. However, although the observed translocation was quite unlikely to be real, we could not be totally sure about the misjoing given that the scaffolding validation was conducted exploiting BAC sequences of a different *P. nigra* genotype. In this sense, the misjoing error could be ascrivable to a structural variation between the two genomes reducing to 0 the error rate.

Comparative genomic analysis

The number of sequenced genomes of both related and closely related species is increasing. Many comparative genomic analyses could be done on the available genomic sequences. We designed a comparative genomic analysis pipeline that was successfully applied to the genomes of the two closely related species, *P. nigra* and *P. trichocarpa*. We investigated the two species sequences in terms of their pan genome. The pan-genome concept was initially introduced in bacterial species [18] and subsequently in maize [19]. In particular, the pan-genome includes a core genome containing genes that are present in all individuals and a dispensable genome composed of partially shared and individual-specific DNA sequence elements [19]. Our pipeline allowed us to introduce the pan genome concept in poplar (Figure 8.1) establishing the shared sequence portion between the two species and the *P. nigra*- and *P. trichocarpa*-specific portions.

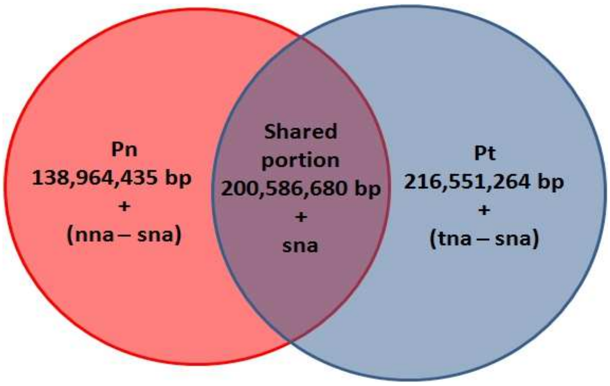


Figure 8.1: Pan genome in poplar. Pn=*P. nigra*; Pt=P. trichocarpa; nna=not-assembled sequence portion in *P. nigra*; tna=not-assembled sequence portion in *P. trichocarpa*; sna=shared sequence portion between the not-assembled sequences of the two genome.

Considering that the *Populus* genome size is ~500 Mb and both the *P. nigra de novo* assembly and the *P. trichocarpa* assembled genome sequence are shorter than 500 Mb, the core genome is composed by the shared sequence portion between the assembled *P. nigra* and *P. trichocarpa* sequences, that we found being 200,586,680 bp long, and by a putative shared sequence portion between the not-assembbled sequences of the two species. The *P.*

nigra-specific portion is constitued by 138,964,435 bp (pipeline result) and by the *P. nigra* not-assembled sequence portion excluding those sequences that might be shared with the *P. trichocarpa* not-assembled sequence. On the other hand, the *P. trichocarpa*-specific portion is composed by 216,551,264 bp [=417,137,944 bp (assembled genome sequence) - 200,586,680 bp (shared sequence portion between the two species)] and the *P. trichocarpa* not-assembled sequence portion excluding those sequences that might be shared with the *P. nigra* not-assembled portion.

To calculate the size of the shared portion between the two genomes, the pipeline assigned the *de novo* contigs to three different set classified in terms of shared sequence length with the reference genome. Characterising the content of each contig set for repetitive sequences and protein-coding potential (Figure 7.9), we observed that the set containing mainly *P. nigra*-specific contigs had a higher repetitive sequence content (51%) and a much lower protein-coding potential (7%) compared to the other two set that were mostly composed by contigs sharing a longer part of their sequence with the *P. trichocarpa* genome. On the other hand, the shared *P. nigra* sequence portion contained a higher number of genes compared to the *P. nigra*-specific portion which was composed by a higher level of repetitive sequences. This is in accordance to the pan genome concept.

Finally, we observed the BLASTn alignment trend of the *P. nigra de novo* assembly contigs against the *P. trichocarpa* reference genome (Figure 7.10). The contigs mainly cover the *P. trichocarpa* regions with a high gene number and a low level of repetitiveness. The fact that the highly repetitive *P. trichocarpa* regions were not extensively covered as the gene regions, might be explained by difficulty in assembling the repetitive regions (assembled in single copy) or as the presence of *P. trichocarpa*-specific repetitive sequences according to the pan genome concept.

Bibliography

[1] Tuskan, G.A., Difazio, S., Jansson, S. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313** 1596-604 (2006).

[2] Food and Agricultural Organization of the United Nations, State of the World's Forests 2003 (FAO, Rome, 2003).

[3] Lefevre, F., Vries, S.D., Turok, J. Strategies for the conservation of a pioneer tree species, *Populus nigra* L ., in Europe. *Genet. Selv. Evol.* **30**, S181-S196 (1998).

[4] Popivshchy, I.I., Prokazin, A.E. *et al. Populus nigra* Network. Report of the third meeting, Sarvar, Hungary, 5-7 October 1996, IPGRI, Rome, 1997, pp. 46-52.

[5] Taylor, G. *Populus: Arabidopsis* for Forestry. Do We Need a Model Tree?, *Annals of Botany* **90**, 681-689 (2002).

[6] Jansson S. and Douglas C. *Populus*: a model system for plant biology. *Annual review of plant biology* **58**, 435-58 (2007).

[7] Zhou, F., Xu, Y. RepPop: a database for repetitive elements in *Populus trichocarpa. BMC Genomics* **10**, 14 doi:10.1186/1471-2164-10-14

[8] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860921 (2001).

[9] Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A. *et al.* The sequence of the human genome. *Science* **291**, 13041351 (2001).

[10] Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520562 (2002).

[11] Schatz, M.C., Delcher, A.L., Salzberg, S.L. Assembly of large genomes using second-generation sequencing. *Genome research* **20**, 1165-73 (2010).

[12] Velasco, R., Zharkikh, A., Troggio, M. *et al.* A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PLoS One* **2**, e1326. doi: 10.1371/journal.pone.0001326 (2007).

[13] Huang, S., Li, R., Zhang, Z., Li, L. *et al.* The genome of the cucumber, *Cucumis sativus* L. *Nat Genet* **41**, 12751281 (2009).

[14] Ruiqiang, L., Wei, F., Geng, T., Hongmei, Zhu., Lin, H. *et al.* The sequence and *de novo* assembly of the giant panda genome. *Nature* **463**, 311-317 (2010).

[15] Star, B., Nederbragt, A.J., Jentoft, S., Grimholt, U. The genome sequence of Atlantic cod reveals a unique immune system. *Nature* **477** 207210 (2011).

[16] Shulaev, V., Sargent, D.J., Crowhurst, R.N., Mockler, T.C. The genome of woodland strawberry (*Fragaria vesca*). *Nature genetics* **43**, 109-16 (2001).

[17] Genome 10K Community of Scientists. Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. *J Hered* **100**, 659674 (2009).

[18] Tettelin, H., Masignani, V., Cieslewicz, M.J., *et al.* Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proc Natl Acad Sci USA* **102**, 13950-13955 (2005).

[19] Morgante, M., De Paoli, E., Radovic, S. Transposable elements and the plant pan-genomes. *Current opinion in plant biology* **10**, 149-155 (2007).

[20] Zhang, H., Zhao, X., Ding, X., Paterson, A.H., Wing, R.A. Preparation of megabase-size DNA from plant nuclei. *The Plant Journal* **7**, 175-184 (1995).

[21] Doyle, J.J., Doyle, J.L. A rapid DNA isolation procedure from small quantities of fresh leaf tissues. *Phytochem Bull* **19**, 11-15 (1987).

[22] Rodrigue, S., Materna, A.C., *et al.* Unlocking Short Read Sequencing for Metagenomics. *Publ Lib of Sci One* **5**, e11840 (2010).

[23] Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J.M., Birol, I. ABySS: a parallel assembler for short read sequence data. *Genome research* **19**, 1117-23 (2009).

[24] Pevzner, P.A., Tang, H. & Waterman, M.S. An Eulerian path approach to DNA fragment assembly. *Proc. Natl Acad. Sci. USA* **98**, 97489753 (2001).

[25] Mott, R. Personal communication.

[26] Vezzi1, F., Del Fabbro, C., Tomescu, A.I., Policriti, A. rNA: a Fast and Accurate Short Reads Numerical Aligner. *Bioinformatics* (2011).

[27] Casagrande, A., Del Fabbro, C., Scalabrin, S., Policriti, A. GAM: Genomic Assemblies Merger: A Graph Based Method to Integrate Different Assemblies. Bioinformatics and Biomedicine, 2009. BIBM '09. IEEE International Conference. 1-4 Nov. 2009, Washington, DC. 321 - 326

[28] Kurtz, S., Phillippy, A., Delcher, A.L. *et al.* Versatile and open software for comparing large genomes. *Genome Biology* 5:R12 (2004).

[29] Smit, A.F.A., Hubley, R., Green, P. RepeatMasker Open-3.0. 1996-2010 http://www.repeatmasker.org.

[30] Cossu, R.M., Buti, M., Giordani, T., Natali, L., Cavallini, A. A computational study of the dynamics of LTR retrotransposons in the *Populus trichocarpa* genome. *Tree Genetics & Genomes* DOI 10.1007/s11295-011-0421-3 (2011).

[31] Goodstein1, D.M., Shu., S., *et al.* Phytozome: a comparative platform for green plant genomics. *Nucl. Acids Res.* 1-9 (2011).

[32] Boetzer, M., Henkel, C., Jansen, H.J., Butler, D., Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* (2010).

[33] Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17** 10-12.

[34] Price, A.L., Jones, N.C., Pevzner, P.A. *De novo* identification of repeat families in large genomes. *Proc 13 Annual Intern conference on Intelligent Systems for Molecular Biology* Detroit, Michigan (2005).

[35] Gnerre, S., MacCallum, I., Przybylski, D. *et al.* High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Nat. Acad. of Sci. of USA* doi:10.1073/pnas.1017351108 (2011).

[36] Li, R., Zhu, H., Ruan, J. *et al. De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Research* doi: 10.1101/gr.097261.109 (2009).

[37] Miller, J.R., Delcher, A. *et al.* Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics* **24**, 2818-2824 (2008).

[38] Vezzi, F., Narzisi, G., Mishra, B. Feature-by-feature - Evaluating *De novo* Sequence Assembly. *PlosOne* **7**, e31002 (2011).

[39] Salzberg, S.L., Phillippy, A.M., Zimin, A.V. *et al.* GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Research* doi: 10.1101/gr.131383.111 (2011).

[40] Barriére, A., Yang, S.P., *et al.* Detecting heterozygosity in shotgun genome assemblies: Lessons from obligately outcrossing nematodes. *Genome research* **19**, 470-480 (2009).

[41] Alkan, C.C., Bradley, P., Eichler, E. Genome structural variation discovery and genotyping. *Nature reviews. Genetics* **12**, 363-376 (2011).

[42] Bradshaw, H.D., Stettler, R.F. Molecular genetics of growth and development in *Populus.* I. Triploidy in hybrid poplars. *Theor. Appl. Genet.* **86**, 301-307 (1993).

[43] Kurtz, S. Narechania, A., Stein, J.C., Ware, D. A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. *BMC genomics* **9**, 517 (2008).