Improvement of a high throughput method for identification of T-DNA insertion events in an activation-tagged population of *Populus tremula* x *Populus alba*.

By Jonas Lafave

A thesis submitted to the Graduate Program in Biology
In conformity with the requirements for the degree of Master of Science.

Queen’s University
Kingston, Ontario, Canada
May, 2021

COPYWRITE © JONAS M. LAFAVE, 2021
Abstract:

Mutant populations are a cornerstone of modern studies in genetics; their usefulness in gene discovery, and in understanding the various forms of gene regulation, can not be overstated. Since the introduction of mutation in plants nearly a century ago (Stadler, 1928), mutant populations have been generated in every major crop as well as many additional species (Sikora et al., 2011). While mutant populations have become commonplace in the study of plant biology, the usefulness of each population is limited by the ability of researchers to characterise the mutations present. This limitation has led me to search for an improved method to characterise the random T-DNA insertions sites of the activation tagged *Populus tremula x Populus alba* population generated by the Regan lab. In this thesis, I explore the potential of using a capture-based sequencing method to construct a high-throughput sequencing method that can be used to identify the genomic flanking regions that border T-DNA insertions. This theoretical process is then used to evaluate experimental data generated in 2 sequencing runs, examining the output for quality, consistency, and accuracy. Finally, I use my method to positively confirm T-DNA insertion sites in 4 selected lines of activation tagged *P. tremula x P. alba*, demonstrating the viability of my method to identify random T-DNA insert locations.
Acknowledgements

I’d like to acknowledge my supervisor, Sharon Regan, for allowing me to run with this project over the others that she’d no doubt like to get finished. And for displaying as much patience as she did during my writing of this thesis, since writing is not my strength.

I’d also like to acknowledge Brian Boyle, our collaborator at the University of Laval, and the huge role he played in my project. His experience from an earlier project involving identifying T-DNA insert locations in *Arabidopsis* was invaluable for this project, and it couldn’t have proceeded without him.

I’d like to thank our post-doctoral researcher, Sagar Datir, for his assistance in the lab and in reviewing my results. An extra set of hands or eyes is always a help in the lab.

Finally, I’d like to acknowledge Professor Rob Colautti and members of his lab (Almira and Graeme) for their assistance. Professor Colautti was willing to let me bounce ideas off of him and take the time to explore how different approaches to this project would play out. His staff, particularly Almira, were extremely helpful in getting the project started.
Abstract: ........................................................................................................................................... i
Acknowledgements ..........................................................................................................................ii
Table of Contents ............................................................................................................................ iii
List of Figures and Illustrations ........................................................................................................ v
List of Tables ................................................................................................................................... vi
List of Abbreviations ...................................................................................................................... vii
Chapter 1. Introduction .............................................................................................................. 1
  1.1 Mutant populations: their uses and limitations .............................................................. 2
  1.2 Identifying mutation locations in activation-tagged population ................................. 5
  1.3 High-throughput sequencing technologies reviewed ................................................ 9
  1.4 Analysing sequencing data ............................................................................................. 13
    1.4.1 Multiplexing samples .............................................................................................. 13
    1.4.2 Sequence assembly: de novo assembly vs. reference-guided assembly .......... 14
  1.5 Sequencing of Populus species, and the establishment of the activation-tagged P. tremula x P. alba population ..................................................................................... 15
  1.6 Developing an improved method to locate inserts from flanking sequences .............. 17
Chapter 2. Materials and Methods ........................................................................................... 19
  2.1 Sample collection and line selection ............................................................................. 19
  2.2 Sample preparation and genomic DNA extraction ......................................................... 21
  2.3 Command line alignment, de novo assembly, and use of BLAST to map reads to the P. trichocarpa genome ...................................................................................................... 26
  2.4 Geneious Prime 2019 sequence assembly ................................................................... 28
  2.5 PCR for the confirmation of left border locations .......................................................... 30
  2.6 Identifying genes bordering T-DNA inserts from selected lines .................................. 32
Chapter 3. Results ..................................................................................................................... 33
  3.1 Sequencing and assembling the first run ..................................................................... 33
  3.2 Sequencing and assembling the second run ................................................................. 38
  3.3 Confirming left border inserts through PCR and Sanger sequencing ......................... 45
List of Figures and Illustrations

Figure 1. Overview of plasmid rescue technique. ................................................................. 7
Figure 2. Overview of TAIL-PCR.......................................................................................... 8
Figure 3. Overview of Illumina, PacBio SMRT, and Oxford Nanopore sequencing technologies. 12
Figure 4. Simplified process for reference-guided and de novo assembly of sequence data. ... 14
Figure 5. The mechanism of enrichment of the genomic DNA with a capture probe. ............ 18
Figure 6. Steps of DNA library preparation for first sequencing run.................................. 24
Figure 7. Capture to enrich for desired sequences before final PCR.................................... 25
Figure 8. Bioinformatics workflow for first sequencing run................................................. 27
Figure 9. Bioinformatics workflow for second sequencing run.......................................... 28
Figure 10. Location of key features for confirming left border insert locations.................... 30
Figure 11. Heat maps of the first sequencing run................................................................. 35
Figure 12. Contigs have uneven distribution of reads......................................................... 35
Figure 13. Example of contigs mapping to multiple genome locations and displaying different sizes and structures. ..................................................................................................................... 36
Figure 14. Example of mapped locations displaying gaps between T-DNA and genomic ends of sequenced fragments. ........................................................................................................... 37
Figure 15. Comparison of read numbers from first and second sequencing runs. .............. 39
Figure 16. All read pairs in the second sequencing run were found to contain the T-DNA. .... 41
Figure 17. The second sequencing reads assembled into far fewer contigs than the first run. 42
Figure 18. Captured T-DNA demonstrates improved specificity for flanking genomic material and enables identification of T-DNA insert locations. .......................................................... 43
Figure 19. Contigs from second sequencing map to single genome locations ..................... 43
Figure 20. Sequences that map to multiple locations may indicate insertion in duplicated regions........................................................................................................................................... 44
Figure 21. Sequencing of some T-DNA sites reveals similarity to Agrobacterium plasmid DNA. 45
Figure 22. PCR results of left border confirmation............................................................. 46
Figure 23. Location of T-DNA activation-tagging insert in line #148.................................... 47
Figure 24. Location of T-DNA activation-tagging insert in line #316................................. 47
Figure 25. Location of T-DNA activation-tagging insert in line #638................................. 48
Figure 26. Location of T-DNA activation-tagging insert in line #903................................. 48
List of Tables
Table 1. Lines selected for sequencing from heavy metal screen................................. 19
Table 2. Lines selected from previously characterised phenotypes.............................. 20
Table 3. Lines selected from phenotyping of the field site.............................................. 20
Table 4. Lines selected from secondary cell wall phenotypes........................................ 21
Table 5. PCR conditions for TAIL-PCR of T-DNA flanking regions in activation-tagged mutants 23
Table 6. Primers for confirming left border genomic regions of activation-tagged lines...... 31
Table 7. T-DNA insert locations in 4 selected lines......................................................... 45
Table 8. Comparison of sequencing platforms................................................................. 51
List of Abbreviations

AGL9 – AGAMOUS-like 9
ApeKI – Aeropyrum pernix K1
bHLH13 – basic helix–loop–helix
BLAST – Basic Local Alignment Search Tool
bp – base pairs
BWA – Burrows-Wheeler Aligner
CaMV – Cauliflower Mosaic Virus
CaMV 35S – Cauliflower Mosaic Virus 35S promoter
Contig – contiguous DNA sequence
CSHL – cold Spring Harbour Labs
CTAB – cetyl-trimethyl ammonium bromide
DNA – deoxyribonucleic acid
DOE – Department of Energy
EDTA – Ethylenediaminetetraacetic acid
EJ2 – ENHANCER-OF-JOINTLESS-2
EMS – ethyl methanesulfonate
ExOTIC/JIC – Exon Trapping Insert Consortium/ John Innes Centre
gDNA – genomic DNA.
GenBank – Genetic Sequence Database
HAK1 – probable leucine-rich repeat receptor-like serine/threonine-protein kinase At1g06840
HCl – hydrochloric acid
INRA – Institut National de la Recherche Agronomique
IPTG – isopropyl β- d-1-thiogalactopyranoside
JGI – Joint Genome Institute
JIC – John Innes Centre
LB – left border

IncRNA – long, non-coding RNA

LRR – leucine-rich repeat

MADS-box – MCM1, AGAMOUS, DEFICIENS, SRF

PacBio – Pacific Biosciences

PAG – Platforme Analyse Genomique

PCR – polymerase chain reaction

PVP – polyvinylpyrrolidone

RB – right border

RNA – ribonucleic acid

SAIL – Syngenta Arabidopsis Insertion Library

SMRT – Single Molecule Real Time

T-DNA – transfer DNA

TAIR – The Arabidopsis Information Resource

TCAG – The Centre for Applied Genomics

TAIL PCR – Thermal Asymmetric Interlaced Polymerase Chain Reaction

TILLING – Targeting Induced Local Lesions IN Genomes

Tris – tris(hydroxymethyl)aminomethane

x-gal – 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside
Chapter 1. Introduction

Tree species are an important ecological resource and are used in industry, landscaping, and food production. The poplars are a diverse group of temperate dioecious trees in the genus *Populus* (family *Salicaceae*) that include the cottonwoods, aspens and balsam poplars. These fast growing trees are responsible for contributing numerous valuable ecosystem services, such as phytoremediation (Giachetti and Sebastiani, 2006), water purification (Khurelbaatar *et al*., 2017; Paniagua *et al*., 2016), erosion control (Hathaway and Penny, 1975), and carbon sequestration (Fortier *et al*., 2016). Compared to other angiosperm trees, the fast growth of *Populus* species also make them useful for habitat restoration and stabilization, home construction materials, pulp and paper, and as a biomass source for bioenergy (Geyer and Walawender, 2000). Fast growth, economic importance, and its comparatively small genome of 550 million bases (Tuskan *et al*., 2006) have contributed to several *Populus* species being considered the model systems for molecular biology and genetics of forest trees.

Considerable research has focused on improving the commercial value of poplar through both breeding programs and genetic modification. For instance, the analysis for genes of interest through genome-wide association studies and quantitative trait loci mapping (Bdeir *et al*., 2017) programs have yielded lines with faster growth (Guerra *et al*., 2019), altered lignin properties and wood traits (Porth *et al*., 2013; Ralph *et al*., 2012), better disease resistance (Jiang *et al*., 2019), changes in budding characteristics (Frewen *et al*., 2000) and improved resistance to biotic and abiotic stress (Chhetri *et al*., 2019; Guy *et al*., 2014; Monclus *et al*., 2012). Molecular studies building on this and other work have pushed even further, developing genetically modified lines
with herbicide resistance (Ault et al., 2016), altered cellulose and lignin biosynthesis pathways (Nabuqi et al., 2020) improved phytoremediation traits (Castro-Rodríguez et al., 2016; Rugh et al., 1998), and in general improving the understanding of the genes involved in regulating several other aspects of tree biology (Ye et al., 2011).

Understanding the function of genes relies on two processes, forward and reverse genetics. The process of forward genetics identifies the function of a gene by connecting an observed phenotype with the responsible gene (Alonso and Ecker, 2006). For this reason, forward genetics usually involves screening of large mutant populations for a phenotype of interest and then identifying the location of the mutation and ultimately discovering the responsible genes (Schreiber et al., 2012; Wu et al., 2015). Forward genetic screens have identified genes responsible for hundreds of functions (Ding et al., 2015; Dobritsa et al., 2011; Huang et al., 2008; Schreiber et al., 2012; Stamatiou et al., 2013; Zwiewka and Friml, 2012). Reverse genetics first identifies a gene, and then alters the gene through transgenic approaches to identify the impact on the plant (phenotype). Like forward genetics, reverse genetics studies have also been used to identify a diverse range of genes in mutant populations (Beris et al., 2016; Shkryl et al., 2021).

1.1 Mutant populations: their uses and limitations

In plants, there are several methods used to generate mutants for forward genetics studies. Populations can be generated with chemical mutagens such as ethyl methanesulfonate (EMS), physical damage such as exposure to radiological agents, or biological factors including transfer
of DNA (T-DNA) and transposons (Østergaard and Yanofsky, 2004). Chemical and physical mutagens have the advantage of being capable of generating many thousands of lines in a short period of time, however, they are known to generate multiple mutations in the genome that require time-consuming back crossing to generate lines with only a single mutation for study (Weigel and Glazebrook, 2008). Numerous populations have been constructed in model species Arabidopsis thaliana using these methods (Appendix H), but transgenic poplars lagged because of the difficulty in backcrossing them.

Mutant populations created by way of biological factors such as transposons (May et al., 2003) or transfer DNA (Krysan et al., 1999), offer more stable mutations (Alonso and Ecker, 2006; Xie and Yang, 2013; Sanson et al., 2018). Transfer DNA, or T-DNA, uses a plasmid naturally derived from strains of Agrobacterium tumefaciens known to infect plants and cause tumor-like growths called galls (Smith and Townsend, 1907). These growths are the result of plant cells producing two hormones (auxin and cytokinin) causing undifferentiated cell growth, similar to cancer cells in the human body (Akiyoshi et al., 1984). The instructions are transmitted to the cells by way of this special plasmid, called the Tumor-inducing (T) plasmid, which integrates itself into the plant’s genome through the DNA repair mechanism of non-homologous end joining (Akiyoshi et al., 1984; Jia et al., 2012). The plasmid feature which enables this integration are the flanking regions referred to the Left Border (LB) and Right Border (RB) (Jen and Chilton, 1986). Biologists have made use of this property to insert desired genes into plants, such as entirely new metabolic pathways (Beyer et al., 2002). The use of T-DNA to introduce large DNA fragments and mutation sites into a genome also offers fast characterisation of lines using PCR-based methods since the sequence of the insert is already known, though they insert randomly in the genome.
and may also lead to other genomic rearrangements as well (Alonso and Ecker, 2006; Gang et al., 2019).

There are difficulties in establishing large tree breeding populations for forward genetics studies, including the long time it takes trees to reproduce. Delving into the genetics of tree species such as *Populus* (poplar) is presented with difficulties. Tree species usually take more than a decade to progress from sapling to the reproductive stage of their life cycle. Therefore, it was necessary to look to reverse genetics approaches that may be applied to trees.

It is also the ability for T-DNA to insert into the genome that enabled the creation of a forward genetics tool that can be used in tree species – activation tagging. The cauliflower mosaic virus (CaMV) promoter is often used to over express a gene in plants (Odell et al., 1985). Activation tagging makes use of a portion of the CaMV promoter, called the 35S enhancer, to enhance expression of genes; plasmids that include multiple copies of the 35S enhancer such as pSKI074 (Appendix A) can increase gene expression several times over it’s normal rate (Weigel et al., 2000). Activation tagging results in a gain-of-function phenotype because the gene is overexpressed and will have a phenotype indicative of that overproduction of the protein. Most importantly, activation tagging using the 35S enhancers results in an increase in gene expression but in its natural expression pattern (ex: leaf-specific gene expression will still only be expressed in the leaf, but the level of expression will be higher). This is in stark contrast to experiments where a transgenic plant is created with the gene of interest is inserted under the direction of the CaMV promoter. The CaMV promoter drives high levels of expression in almost all tissues of the plant, also known as ectopic expression (meaning “out-of-place” expression). This ectopic
expression can generate ectopic phenotypes such as flower organs on leaves (Rutledge et al., 1998).

Activation tagging was first established in Arabidopsis (Weigel et al., 2000; Robinson et al., 2009; Taheri et al., 2014) and has since been developed in other plants including Populus (Busov et al., 2003; Harrison et al., 2007; Busov et al., 2011). The Regan lab created the largest population of activation-tagged Populus in the world with more than 2000 independent lines (Harrison et al., 2007) and the population has resulted in more than 75 distinct phenotypic mutants. Several of these mutants have been fully characterized and the genes responsible for that phenotype have been identified (Harrison et al., 2007; Plett et al., 2010; Yordanov et al., 2010; Trupiano et al., 2013; Plett et al., 2014). While activation tagging is an excellent tool for forward genetics in trees, it still requires that insert locations be identified. Lines in some species may need to have extensive studies done to characterise the mutations (Oladosu et al., 2016), since the crossing practices used for plant species such as A. thaliana runs into the same difficulties as establishing a large forward genetics population.

1.2 Identifying mutation locations in activation-tagged population

One of the challenges of using a forward genetics approach is identifying the location of the mutation. Random mutagenesis through chemical mutagens, for example, is a well established method of generating multiple mutations at once (Wu et al., 2005), but it is difficult to identify the location of the mutations unless you are using a genetically tractable system such as Arabidopsis and can use map-based cloning to identify the mutation (Jander et al., 2002).
Identifying all the changes induced by chemical mutagen exposure can only truly be performed by whole genome sequencing on each line (Hartwig et al., 2012) and comparison to a high quality reference genome; while sequencing costs have come down and accelerated methods developed (Schneeberger, 2014), they have not yet reached the point where this is a viable option for large collections. Trees are not well suited to this type of mutagenesis since their flowers are not produced for several years even decades (Fladung et al., 2004), and they are less-well characterized making map-based cloning impossible.

For populations created by T-DNA insertions or transposons, the DNA sequence that is inserted into the genome that can be used to identify the site of the insertion. The two most common methods for identifying these insertions are plasmid rescue and TAIL-PCR (or Genome Walking™) that are used. Plasmid rescue incorporates bacterial origin of replication sites and bacterial antibiotic resistance genes to construct an insertable fragment of DNA (either T-DNA or transposon) that can later be recovered. For recovery or “rescue”, the genome is cut using a restriction enzyme not present inside the insert and compatible overhangs ligated together. When ligated, a new plasmid is produced which contains the flanking genomic material (Behringer and Medford, 1992; Radhamony et al., 2005; Sangwan et al., 2012), as well as the bacterial origin of replication and antibiotic resistance genes, making it a self-replicating construct that can be grown up as any other plasmid (Figure 1). When the plasmid is sequenced, the flanking genomic regions provide the location of a T-DNA insertion. Plasmid rescue has successfully been used to identify T-DNA insertions in Arabidopsis and poplar, however only a few mutants are investigated at a time and the process does not adapt well into a high-throughput method (Weigel et al., 2000; Johansson et al., 2003).
Figure 1. Overview of plasmid rescue technique.

A DNA insert is designed to have high copy number *Escherichia coli* (*E. coli*) bacterial origin of replication and both bacterial and plant antibiotic selection markers. When the T-DNA inserts into a genome, it can later be cut out with a restriction digest and ligated onto itself. Once ligated, a self-replicating construct containing high copy number bacterial origin of replication and antibiotic resistance gene, as well as genomic flanking material, will be able to be cloned as any other plasmid and sequenced (Sangwan *et al.*, 2012).

An alternative method to identify T-DNA location is through Thermal Asymmetric Interlaced Polymerase Chain Reaction (TAIL PCR) (Singer and Burke, 2003). Commercially available kits that make use of TAIL-PCR, such as from TakaraBio, for example, refer to it as “genome walking”. One set of nested PCR primers are designed in the T-DNA, as well as degenerate primers that anneal throughout the genome (Figure 2). These nested PCRs use
Figure 2. Overview of TAIL-PCR.

TAIL-PCR can be used to identify flanking regions of DNA inserts by using multiple cycles of nested PCR. In the first stage, a primer specific to the insert (LB1) and random primers in the genome (AD) generate both specific and non-specific amplicons. Following PCR stages reduce the amplification of non-specific sequences, leaving only the specific amplicon for sequencing. (Singer and Burke, 2003)

T-DNA primers adjacent to but not overlapping the previous primer (Figure 2, Step 1). Sorting the fragments still represents a major challenge, requiring each mutant line in a population to have a library of fragments made using different degenerate primers, and multiple fragments sequenced to identify the insert location. Scaling this approach to identify the T-DNA locations
for an entire population of activation-tagged lines would require too many PCR reactions to be financially viable (Singer and Burke, 2003). Previously, all lines in the Regan lab have used this method to identify insert locations (Harrison et al., 2007; Plett et al., 2010; Plett et al., 2014), however this was to identify insertions in single lines. To identify insertions in a population of 2000 mutant lines such as the activation-tagged lines in the Regan lab, finding the T-DNA location for one border in all lines would require approximately 4,000 PCR reactions.

1.3 High-throughput sequencing technologies reviewed

With the recent advances in sequencing technologies (Heather and Chain, 2016) and the simultaneous decrease in sequencing costs (Marshall et al., 2017), it is becoming increasingly feasible to use genome sequencing to identify mutations, including the T-DNA location in activation-tagged lines (Boyle et al., 2013). With these advances, it is now possible to sequence T-DNA locations directly, without the need for intermediate steps such as Genome Walking™ or plasmid rescue and may also be able to sequence far enough to overcome the genome disruptions common to T-DNA insertions (Boyle et al., 2013). Many sequencing methods exist, and here I focus on large-scale sequencing platforms compatible with high-throughput sequencing that could be adapted to identify T-DNA locations in mutant populations.

Short-read sequencing is the oldest of the sequencing methods, and includes Sanger sequencing, more modern Illumina sequencing, 454 pyrosequencing, and Ion Torrent platforms (Ansorge, 2009; Morozova and Marra, 2008). Starting in the mid-1960’s with the addition of radiolabelled nucleic acids and refined over the following decade, sequencing by synthesis
dominated the short-read sequencing market. Beginning with the early work of Fred Sanger (Sanger et al., 1973), sequencing by synthesis relies on the ability of DNA polymerases to work from a template and assemble a new copy of DNA, labelled with radioactive $^{35}S$ or $^{32}P$ dideoxynucleosides (Heather and Chain, 2016). While there were competing technologies such as Maxam and Gilbert (Maxam and Gilbert, 1977), the Sanger method eventually won out as the most used technique and continues to be used extensively today. The usefulness of Sanger sequencing was enhanced in 1986 with the development of fluorescently labelled dideoxy nucleotides by Applied Biosystems that propelled Sanger sequencing into the leading position for sequencing (Kurowska et al., 2011). Improvements over the following decades, led to increases in efficiency that enabled the development of automated sequencing systems. The most prominent of these automated systems were created by Illumina. Illumina sequencing platforms use flow cells with covalently bound oligos which are complimentary to adapters that are ligated onto the ends of the nucleic acid strand to be sequenced (Bharti and Grimm, 2019). Continued innovation by Illumina have enabled them to push their short-read technology from 100 base fragments to sequence longer and longer pieces, but at much higher costs, and they still fall short of true long-read sequencing platforms which are capable of reading fragments thousands of bases long. Numerous factors contribute to the prevalence of Illumina over its competitors, Ion Torrent and 454 pyrosequencing, but the ability to be used multiple types of specialised sequencing projects, coupled with its low cost, have certainly contributed to its market dominance.

Currently there are 3 major platforms that offer true long read sequencing, Pacific Biosciences (PacBio) and Oxford Nanopore, and 10X Genomics (Dijk et al., 2018). Given the
uncertain future of 10X Genomics due to multiple patent infringement suits, they will not be reviewed. Like the Illumina platform, PacBio’s Single Molecule, Real Time (SMRT) system uses sequencing by synthesis, with ligated adapters functioning as primers; while Illumina uses those adapters to immobilise oligos to a flow cell, the SMRT system instead forms single-stranded circular DNA constructs made up of hairpin adapters and oligo of interest (Figure 3). The circular constructs are then isolated into waveguides that channel photons to detectors, and as the polymerase uses labelled nucleotides to synthesise a new strand of DNA the photons released are used to determine what base was added. The SMRT system can operate in 2 modes depending on the length of the oligonucleotide of interest and has demonstrated the capability to sequence fragments of over 20 kb in length. Unlike sequencing by synthesis and sequencing by degradation methods, Oxford Nanopore’s MinION system does not rely on building new copies of or breaking down DNA to gather information about its structure; instead, it is able to read the molecule by passing it through a pore in the surface of a flow cell and reading a change in voltage as each base passes though a current gate without making any changes to its structure. This may be useful when the synthesis process itself may be generating errors in reads, however the physical limitation of the flow cell size reduces the number of reads that can be simultaneously processed, and current flow cells require specialised buffers and storage conditions due to the use of proteins to pull DNA into the pores which also increases the handling and storage complexity.
In Illumina “sequencing by synthesis” systems, adapters are ligated onto 5’ and 3’ ends of DNA strands and bind to their complementary sequences in the flow cell. Paired-end sequencing involves sequencing one end of the fragment, the DNA folds over so that the adapter on the opposite end can bind to the cell, and then sequencing is carried out starting with the opposite end. PacBio long read sequencing ligates hairpin adapters to a double-stranded DNA sequence, and an immobilized polymerase adds nucleotides in a “sequencing by synthesis” method similar to rolling circle amplification. Finally, Nanopore sequencing by Oxford uses ligated adapters to feed a DNA strand into a motor protein and through a tiny hole that has a current passing through it. As each nucleotide passes through, it causes a detectable change in voltage, providing real-time sequencing of long DNA fragments (Bharti and Grimm, 2019).
1.4 Analysing sequencing data

One of the challenges of DNA sequencing is that the results come out as short fragments (called reads) which must be assembled together into contiguous sequences (contigs), and eventually into larger scaffolds (Miller et al., 2010). As well, different sequencing methods output their data in formats that are not easily read by standard programs such as Word, and often include supplementary data such as chromatograms and quality reports. Along with the advancements in sequencing technologies has been the complementary design of suitable software to analyse and assemble the DNA reads, such as Burrows-Wheeler Aligner (BWA), Bowtie, and NovoAlign (Li and Durbin, 2009; Thankaswamy-Kosalai et al., 2017).

1.4.1 Multiplexing samples

A single sequencing run can output large amounts of data and is capable of providing the entire genome for simple organisms, with extensive coverage of all bases. That high coverage (the number of times a single base is represented in an assembly) is often wasted on smaller projects. Since high coverage is not required for simple projects, and cost is often an issue when sequencing is involved, methods for increasing the number of samples that can be sequenced in a single run have been developed, at the cost of lower coverage.

The usual method for increasing the number of samples that can be sequenced in a single run is to add barcodes to individual samples. Barcodes are short DNA sequences added to the ends of a library by ligation or modified PCR primers (Wong et al., 2013). These short DNA sequences allow a researcher to tie a DNA sequence to a specific sample. By doing so, samples can then be pooled for library preparation, regardless of the sequencing platform to be used. The
addition of these barcodes will add an extra step to the bioinformatics process, since the reads will need to be sorted out in order to gather all the reads for a single sample in one group.

1.4.2 Sequence assembly: *de novo* assembly vs. reference-guided assembly

Once reads have been sorted into those belonging to individual samples, then it becomes possible to perform the actual assembly of reads into larger fragments. This can be performed using two methods – reference-guided assembly, in which an existing sequence is used as a guide for reads to be aligned against (Figure 4), and *de novo* assembly, which aligns the reads against one another and merges them together into larger fragments called contigs, short for contiguous sequences.

![Figure 4. Simplified process for reference-guided and de novo assembly of sequence data.](image)

Both reference and *de novo* assembly can be used to generate contiguous sequences (contigs) representing a stretch of DNA longer than the individual reads but use different steps to arrive at the same final product.

Reference assembly of DNA sequences make use of an existing stretch of sequence to line up reads against, as the name suggests. This is particularly useful where there are existing high quality genome assemblies (Lischer and Shimizu, 2017; Schneeberger *et al.*, 2011), but may not be an option for researchers who do not have access to such tool. Researchers working with
seldom studied species may have to resort to \textit{de novo} assembly or combination of \textit{de novo} and reference-based assembly (Whittall \textit{et al.}, 2010).

\textit{De novo} assembly was the initial method of assembly used for sequencing, and has been used successfully to assemble bacterial (Dohm \textit{et al.}, 2007; Studholme \textit{et al.}, 2009; Whiteford \textit{et al.}, 2005) as well as plastid genomes (Whittall \textit{et al.}, 2010) from short-read sequence data. Unfortunately, \textit{de novo} assembly with short reads suffers from an inability to distinguish complex repetitive regions from one another (Paszkiewicz and Studholme, 2010). While using long reads would seem to be the solution to this problem, the lower accuracy and reduced coverage can often result in errors in the assembly (Koren \textit{et al.}, 2012; Miller \textit{et al.}, 2010). This can be alleviated by using a combination of both long reads and highly accurate short reads in what is called “hybrid assembly”, where the long reads help to bridge complex repeats and serve to anchor the short reads (Kamada \textit{et al.}, 2014; Koren \textit{et al.}, 2012).

1.5 Sequencing of \textit{Populus} species, and the establishment of the activation-tagged \textit{P. tremula} x \textit{P. alba} population

In 2002, the United States Department of Energy (DOE) Joint Genome Institute (JGI) announced it’s intent to sequence a poplar tree (genus \textit{Populus}), and female specimen “Nisqually 1” from \textit{Populus trichocarpa} was selected for this purpose (Tuskan \textit{et al.}, 2004). The genome of \textit{P. trichocarpa} was publicly released in 2006 (Tuskan \textit{et al.}, 2006). While \textit{P. trichocarpa} was an excellent choice for initial sequencing because it was an important tree for the forestry industry, other species of poplar are more amenable to transformation. Of the numerous species of poplar, the \textit{Populus tremula} x \textit{Populus alba} INRA 717-1B4 line is a workhorse of poplar genetics.
and molecular biology work. A vegetatively-propagated hybrid of *P. tremula* and *P. alba* was generated as a part of a breeding program developed by Dr. Michel Lemoine in the 1960’s. Both parent trees were lost to storms in 1999, resulting in vegetative copies of parent lines being unavailable (Kersten *et al.*, 2016). INRA 717-1B4 is a female clone, and used by researchers all over the world for it’s ease of transformation due to a susceptibility to *A. tumefaciens* (Leple *et al.*, 1992). The amenability to transformation that INRA 717-1B4 possesses makes it ideal to generate mutants for identifying gene function in *Populus*. INRA 717-1B4 *P. tremula x P. alba* line was subsequently sequenced because of its long standing superiority for genetic transformation and molecular biology research (Mader *et al.*, 2016). The INRA 717-1B4 line underwent genome sequencing and was published in 2016 (Mader *et al.*, 2016), resulting in sequence for both *P. tremula* and *P. alba* genomes becoming available to the public.

Finding that gene discovery in poplar was limited by the number of activation-tagged lines available (Busov *et al.*, 2003), in 2005 the Regan lab developed the largest activation-tagged population in *Populus* species using INRA 717-1B4 (Harrison *et al.*, 2007). That population has been used to discover genes that affect key aspects of tree physiology, such as root development (Trupiano *et al.*, 2013), leaf morphology (Plett *et al.*, 2010; Williams *et al.*, 2015), and stem architecture (Harrison *et al.*, 2007), to name a few. Since moving the population from the greenhouse to a field site, the number of phenotypes presented is nearly seven times (6.5%) higher than greenhouse-grown *Arabidopsis* populations generated with the same transformation vector. Most of the phenotypes observed were only apparent after the first year in the field, and it is believed that the long lifespan and interaction with biotic and abiotic factors that are responsible for the population being so rich in phenotypes.
1.6 Developing an improved method to locate inserts from flanking sequences

Because T-DNA inserts in random locations in the genome, locating inserts previously has been limited to time consuming processes such as plasmid rescue or TAIL PCR (Genome Walking TM). With the lower cost of sequencing and the development of several high throughput methods, it is now feasible to sequence the mutants to identify the location of the T-DNA. The most direct method to identify the insert locations in the activation-tagged P. tremula x P. alba population would still be to perform with whole genome sequencing approach on each line or mutant. However, this would be an expensive undertaking, with approximately 2000 lines in the population (Harrison et al., 2007) the cost would range from $800 CAD to $2800 CAD per tree depending on the level of coverage requested (The Centre for Applied Genomics - Next Generation Sequencing Facility, 2020). Researchers at the University of Montreal and the Platforme Analyse Genomique (PAG) at the University of Laval developed a targeted genome sequencing approach to identify T-DNA inserts in A. thaliana lines (Boyle et al., 2013). Unlike plasmid rescue, the method can be used in a high-throughput manner as it relies on a combination of sequence capture (Figure 5) and PCR rather than the more difficult and time-consuming process of generating a library of bacterial lines for each tree. Rather than carrying out 100 separate sequencing reactions, samples will be barcoded to combine into a single pool and PCRs for only T-DNA flanking regions used to focus the sequencing to the insert border. The literature reviewed supports the use of short-read sequencing to assemble reads into usable
contigs to provide insert locations, and this also allows us to take advantage of the low cost and well documented Illumina MiSeq platform for sequencing.

**Figure 5. The mechanism of enrichment of the genomic DNA with a capture probe.**

Top: Magnetic beads with streptavidin bound to them are attracted to the biotin-tagged capture probe. Bottom: When genomic DNA with a complementary sequence bind to the capture probe, they can be extracted by applying a magnet to the sample.

In this thesis, I explore the possibility of using both a TAIL-PCR method and a sequence capture method to directly identify flanking locations of T-DNA inserts using high-throughput Illumina sequencing to identify T-DNA insert locations in an activation-tagged population of *P. tremula x P. alba* in a cost and time efficient manner. Here I provide the results of two attempts to use high-throughput sequencing to identify where the T-DNAs are located in activation-tagged poplar mutants in a cost-effective manner.
Chapter 2. Materials and Methods

2.1 Sample collection and line selection

Data from the field site was reviewed for phenotypes of interest. Forty-eight activation-tagged lines of *P. tremula* x *P. alba* were initially selected based on a range of traits including bioremediation potential, previously characterised phenotypes, unique leaf phenotypes, secondary cell wall properties, and (Table 1-4). Leaves were selected from each line from field grown trees, flash frozen in liquid nitrogen, and subsequently stored at -80°C.

Table 1. Lines selected for sequencing from heavy metal screen.

<table>
<thead>
<tr>
<th>Line number</th>
<th>Transformation ID</th>
<th>Heavy metal screen results</th>
</tr>
</thead>
<tbody>
<tr>
<td>882</td>
<td>E4-26</td>
<td>Extreme signs of stress including death</td>
</tr>
<tr>
<td>1026</td>
<td>E20-5</td>
<td>Extreme signs of stress including death</td>
</tr>
<tr>
<td>290</td>
<td>E13-37</td>
<td>Less than average biomass</td>
</tr>
<tr>
<td>907</td>
<td>D4-41</td>
<td>Extreme signs of stress including death</td>
</tr>
<tr>
<td>240</td>
<td>NA</td>
<td>Less than average biomass</td>
</tr>
<tr>
<td>141</td>
<td>WG4-23</td>
<td>More than average biomass</td>
</tr>
<tr>
<td>1493</td>
<td>Z-55</td>
<td>Reduced signs of stress</td>
</tr>
<tr>
<td>1415</td>
<td>NA</td>
<td>Oxidative stress</td>
</tr>
<tr>
<td>542</td>
<td>NA</td>
<td>Reduced signs of stress</td>
</tr>
<tr>
<td>952</td>
<td>DG1-14</td>
<td>Reduced signs of stress</td>
</tr>
<tr>
<td>133</td>
<td>NA</td>
<td>More than average biomass</td>
</tr>
<tr>
<td>638</td>
<td>E9-20</td>
<td>Extreme signs of stress including death, necrotic patches</td>
</tr>
<tr>
<td>1636</td>
<td>E13-34</td>
<td>Extreme signs of stress including death</td>
</tr>
<tr>
<td>508</td>
<td>NA</td>
<td>Reduced signs of stress</td>
</tr>
<tr>
<td>225</td>
<td>NA</td>
<td>Less than average biomass</td>
</tr>
</tbody>
</table>

*Fifteen lines were selected from phenotypes displayed an arsenic and cadmium survivability screen.*
Table 2. Lines selected from previously characterised phenotypes.

<table>
<thead>
<tr>
<th>Line number</th>
<th>Transformation ID</th>
<th>Characterised control name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1562</td>
<td>E19-12</td>
<td>rippled leaf</td>
</tr>
<tr>
<td>1289</td>
<td>E13-43</td>
<td>fuzzy</td>
</tr>
<tr>
<td>1522</td>
<td>E11-25</td>
<td>shrivelled leaf</td>
</tr>
<tr>
<td>630</td>
<td>D23-28</td>
<td>adventitious root</td>
</tr>
<tr>
<td>1290</td>
<td>E1-5</td>
<td>corky</td>
</tr>
<tr>
<td>956</td>
<td>DGQ-22</td>
<td>bent leaf</td>
</tr>
<tr>
<td>351</td>
<td>E7-1</td>
<td>bald leaf</td>
</tr>
<tr>
<td>WT</td>
<td>None</td>
<td>Wildtype</td>
</tr>
</tbody>
</table>

*Seven lines that were previously characterised were selected as controls, as well as a presumed wild-type line.

Table 3. Lines selected from phenotyping of the field site.

<table>
<thead>
<tr>
<th>Line number</th>
<th>Transformation ID</th>
<th>Reason selected</th>
</tr>
</thead>
<tbody>
<tr>
<td>1023</td>
<td>E11-22</td>
<td>yellow leaf</td>
</tr>
<tr>
<td>964</td>
<td>E11-21-9</td>
<td>thin leaf</td>
</tr>
<tr>
<td>875</td>
<td>WG1-11</td>
<td>Leaf phenotype</td>
</tr>
<tr>
<td>380.5</td>
<td>E13-39</td>
<td>Curled down leaf margin</td>
</tr>
<tr>
<td>1508</td>
<td>E19-20</td>
<td>Leaf phenotype</td>
</tr>
<tr>
<td>1739</td>
<td>E1-5-3</td>
<td>Bark phenotype</td>
</tr>
<tr>
<td>1504</td>
<td>R2-6</td>
<td>Leaf phenotype</td>
</tr>
<tr>
<td>301</td>
<td>Z-15</td>
<td>Leaf phenotype</td>
</tr>
<tr>
<td>342</td>
<td>NA</td>
<td>Leaf phenotype</td>
</tr>
<tr>
<td>1088</td>
<td>NA</td>
<td>Leaf phenotype</td>
</tr>
<tr>
<td>1549</td>
<td>E10-24</td>
<td>sensitivity to caterpillars</td>
</tr>
<tr>
<td>321</td>
<td>B7-28</td>
<td>Early leaf emergence</td>
</tr>
<tr>
<td>163</td>
<td>NA</td>
<td>Leaf phenotype</td>
</tr>
<tr>
<td>325</td>
<td>E19-44</td>
<td>rosewood</td>
</tr>
<tr>
<td>903</td>
<td>B3-9</td>
<td>Leaf shape phenotype</td>
</tr>
<tr>
<td>1631</td>
<td>E13-11</td>
<td>random tree</td>
</tr>
<tr>
<td>1054</td>
<td>E3-10</td>
<td>Leaf phenotype</td>
</tr>
</tbody>
</table>

*Seventeen lines were chosen for their leaf and stem phenotypes from previous studies of the field site.*
Table 4. Lines selected from secondary cell wall phenotypes.

<table>
<thead>
<tr>
<th>Line number</th>
<th>Transformation ID</th>
<th>Lignin digest screen results</th>
</tr>
</thead>
<tbody>
<tr>
<td>316</td>
<td>E26-14B</td>
<td>High glucose/xylose</td>
</tr>
<tr>
<td>864</td>
<td>DGQ23-6</td>
<td>High xylose release</td>
</tr>
<tr>
<td>1366</td>
<td>E10-1</td>
<td>Low glucose/xylose release</td>
</tr>
<tr>
<td>1539</td>
<td>B7-17</td>
<td>High xylose release</td>
</tr>
<tr>
<td>1109</td>
<td>B3-9</td>
<td>High glucose release</td>
</tr>
<tr>
<td>903</td>
<td>B3-9</td>
<td>High glucose/high xylose release</td>
</tr>
<tr>
<td>657</td>
<td>E26-16</td>
<td>High lignin</td>
</tr>
<tr>
<td>593</td>
<td>E18-23</td>
<td>High lignin</td>
</tr>
<tr>
<td>142</td>
<td>R8-39</td>
<td>Low S:G ratio</td>
</tr>
<tr>
<td>1109</td>
<td>DGQ-9-3</td>
<td>High S:G ratio</td>
</tr>
<tr>
<td>1366</td>
<td>E10-1</td>
<td>Low glucose/xylose release</td>
</tr>
<tr>
<td>1539</td>
<td>B7-17</td>
<td>High xylose release</td>
</tr>
<tr>
<td>934</td>
<td>R15-23</td>
<td>High glucose release</td>
</tr>
<tr>
<td>148</td>
<td>E10-35</td>
<td>Low xylose release</td>
</tr>
<tr>
<td>1631</td>
<td>E13-11</td>
<td>random tree</td>
</tr>
</tbody>
</table>

*Fifteen lines were chosen from a secondary cell wall study. Lines were selected based on potentially valuable commercial traits.

2.2 Sample preparation and genomic DNA extraction

Genomic DNA (gDNA) from lines of interest was extracted using a modified version of Cold Spring Harbour cetrimonium bromide (CTAB) DNA extraction protocol (Hsia et al., 2010). CTAB buffer (2% CTAB, 100 mM Tris, and 20 mM EDTA at a pH of 8.0, with 40 mg/mL PVP) was heated to 65°C in a water bath and approximately 30 mg of frozen ground leaf tissue was added to 0.5 mL of CTAB extraction buffer, vortexed briefly to mix, and subsequently incubated at 65°C for 30 minutes with mixing at 10-minute intervals. Samples were centrifuged at 13,000 xg for 2 minutes, and approximately 400 µL of supernatant was transferred to a clean Eppendorf tube. To each sample, 2.5 µL of RNase (10 mg/mL, Molecular Solutions) was added and incubated at 37°C for
10 minutes. Following this, 7.5 uL of proteinase K (20 mg/mL, Molecular Solutions) was added and incubated at 56°C for 10 minutes. After the proteinase K incubation 0.5x volume (200 uL) of guanidine -HCl [2M] and 0.5x volume (200 uL) of chloroform were added to the aqueous solution and vortexed for 5 seconds to mix thoroughly. Samples were centrifuged at max speed (21,000 xg) for 2 minutes. Aqueous layers were transferred to a new Eppendorf tube (~300 uL) without disturbing the interface between the layers. To each sample 0.7x (200uL) of 100% isopropyl alcohol was added. Samples were shaken to mix and left to incubate at room temperature for 10 minutes. Following incubation, samples were centrifuged at 21,000 xg for 15 minutes and supernatant was removed and discarded. Pellets were washed twice with 1 mL of 70% ethanol before centrifuging at 21000 xg for 2 minutes. DNA pellets were air dried for 5 minutes at room temperature and resuspended in 50 uL of nuclease free H2O. The DNA was electrophoresed on 1% agarose gel to confirm there was minimal shearing of the gDNA. DNA quality and quantity was assessed by use of a DX-11 FX+ nanodrop (Denovix, DE) device comparing absorbance at 260 nm (A260), the 260nm/230nm, and 260nm/280nm values. Samples were then measured using the Denovix Broad Range Assay Kit for DNA concentration (Denovix, DE). Using the data from the DeNovix assay, all samples were normalised to a concentration of 10 ng/uL before next steps. Individual samples were assigned locations on 96 well plates.

**Library preparation for the first sequencing run**

To sequence the genomic DNA next to the T-DNA for 48 samples but also to decrease the cost of sequencing, unique barcodes were ligated onto each DNA sample. Adapter A was added to each well, and samples and adapters were digested together for 1 h at 75°C with ApeKI (New England Biolabs, MA) in 25 µL volumes containing 1x NEB Buffer 3.1 (New England Biolabs, MA),
and 4 U ApeKl. Following digestion, adapters were ligated to sticky ends by adding 25 µL of a solution containing 2x ligase buffer with ATP and T4 DNA ligase (New England Biolabs, MA) to each well. Samples were incubated at 22°C for 1 h and heated to 65°C for 30 min to inactivate the ligase. Samples were pooled (10 uL each) and purified using a QIAquick PCR Purification Kit (Qiagen, ON) according to the manufacturer’s instructions. Samples were eluted in a final volume of 50 µL.

Following barcoding, the first sequencing run pooled all samples and performed a pair of nested, non-overlapping PCRs using Illumina primers on the genomic end and primers of known location within the activation tag insert (Figure 6) to amplify the bordering region. Primers were diluted to a concentration of 10 pg to reduce non-specific binding. Table 5 details the PCR conditions, with stage 1 using the outer nested primer (CTCTTTCTTTTCCTCCATATTGACCATCATAC), and stage 2 the inner nested primer (AGATTTCGCCGACATGAAGCCTTACAATTGAATATAT). Following the PCR reactions above, sequencing was accomplished on an Illumina Mi-Seq 300 bp paired-end flow cell at the Platforme Analyse Genomique.

Table 5. PCR conditions for TAIL-PCR of T-DNA flanking regions in activation-tagged mutants.

<table>
<thead>
<tr>
<th>PCR stage</th>
<th>Denaturation</th>
<th>Amplification</th>
<th>Annealing</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>98°C, 45s</td>
<td>25 cycles - 98°C, 45 s // 55°C, 30s // 72°C, 1 min</td>
<td>72°C 2 min</td>
</tr>
<tr>
<td>2</td>
<td>98°C, 45s</td>
<td>35 cycles - 98°C, 45 s // 55°C, 30s // 72°C, 1 min</td>
<td>72°C 2 min</td>
</tr>
</tbody>
</table>

PCR conditions using the outer nested primer (Stage 1) and inner nested primer (Stage 2) for modified TAIL-PCR used the same conditions, but a different number of amplification cycles.
Figure 6. Steps of DNA library preparation for first sequencing run.

A: The first sequencing run used modified TAIL-PCR with a barcode/adapter ligated onto only the genomic end. The first round of PCR used an outer primer located in the T-DNA and an Illumina adapter primer. B: The second round of PCR used the same Illumina adapter primer with a T-DNA inner primer fused with the other Illumina sequencing adapter. C: The final product contains Illumina adapters at either end, a barcode adjacent to the genomic flanking sequence, and a small fragment of the T-DNA.

Library preparation for the second sequencing run

Following the results of the approach described above, the following modifications were performed for the second sequencing run to reduce amplification of non-specific product. The concentration of gDNA was increased to 50 ng/uL. Partial Illumina adapters A and C were ligated to both ends of the DNA fragments in the same manner as above. The genomic end adapter consisted of part of the Adapter A sequence and a short barcode. Capture enrichment was performed using the discontinued NimbleGen SeqCap EZ Hypercap (Roche, ON) workflow using the capture probe sequence TAATTACTCTTTTTTCCATATGACCA.
TCATACTCATTGCTGATCCATGTAGATTTCCCGGACATG) (Figure 7). Following the capture, PCR on the capture product was performed using Illumina primers referred to as Primer A and Primer C to complete the adapters. All other steps were performed identically to those of the first run. Lines with a high value such as control lines and those already partially characterized by phenotype were assigned to multiple wells to increase the possibility of retrieving locations and for comparison to one another.

Figure 7. Capture to enrich for desired sequences before final PCR.

A: Following ligation of partial adapters and barcode to the end of DNA fragments, the capture oligo was used to enrich for the T-DNA region immediately adjacent to the LB. B: After capture, the adapter sequences were completed by using proprietary Illumina Primers A and C. C: On completion of the final PCR, enriched samples contain complete Illumina adapters at both ends.

Following library preparation, the sample pool was sequenced using a MiSeq 300bp paired end flow cell. Once completed, reads were sorted by barcode and Illumina adapter sequences were trimmed. Reads were also sorted to remove any that did not have T-DNA present before continuing with bioinformatics to match reads to locations in the \( P. \) alba genome.
2.3 Command line alignment, *de novo* assembly, and use of BLAST to map reads to the *P. trichocarpa* genome

The initial workflow used for identifying T-DNA locations made use of BWA to align sequences to the *P. alba* genome. This workflow was taken from the textbook *Bioinformatics: A Practical Handbook of Next Generation Sequencing and Its Applications* (Low and Tammi, 2016), and tested on a tutorial dataset of 150bp paired-end Illumina reads from an *Escherichia coli* genome. The test performed perfect with no errors. When used on the sequence data from my first sequencing run, the workflow correctly mapped reads to the genome, however when it was converting from SAM file to BAM file types the output file size was 0, and it was unable to continue. The process was repeated with multiple samples, but always failed at the same step. Attempts were made to substitute NovoAlign, another alignment tool, for BWA using a similar previously tested workflow, however since both tools have a similar step requiring the conversion from one file type to another, it also failed. Subsequent to failure in the alignment method, Velvet, a *de novo* assembly tool, was used to assemble reads into contigs. Once completed a review of the data determined that multiple contigs were being generated in each well.

Finally, an attempt was made to use megaBLAST to place reads on the *P. trichocarpa* genome. Both forward and reverse reads were loaded into BLAST, and when complete reads were found to be scattered across multiple locations in the genome, with many reads being discarded as unable to match. I also attempted to use discontinuous BLAST, which has a lower threshold for specificity, to place reads, but it resulted in the same pattern of reads scattered across the genome and many reads discarded. These issues with command line sequencing led
to use of alternative sequencing tool Geneious Prime, and the second sequencing run made exclusive use of Geneious Prime. A summary of steps can be found below in Figures 8 and 9.

![Bioinformatics workflow for first sequencing run.](image)

**Figure 8. Bioinformatics workflow for first sequencing run.**

The bioinformatics workflow for the first sequencing run was complicated by unexpected errors. Highlighted in red are the command line tools intended for alignment, assembly, and BLAST. In green are the steps performed in Genious Prime. Both of the alignment tools (BWA and NovoAlign) stopped during the workflow and were unable to recover. The de novo assembly produced multiple contigs, and the BLAST method resulted in reads scattered throughout the genome. Failures in these methods led to the use of Geneious Prime for data visualisation and analysis.
Figure 9. Bioinformatics workflow for second sequencing run.

In green, the second sequencing run analysis made exclusive use of the Geneious Prime 2019 software. The workflow was essentially the same as that used for analysis of the first sequencing run, with the exception of step 6, in which contig structures were checked for anomalies prior to mapping.

2.4 Geneious Prime 2019 sequence assembly

Geneious Prime is a Windows-OS tool for viewing and analyzing DNA, RNA, and protein sequences. It includes not only its own assembly algorithm, but also allows several other alignment tools (such as Bowtie) and assembly tools (Velvet, SPADES, and others) to be used as plugins. The Geneious Prime 2019 version was used to assemble the first run of sequencing using the tutorial for de novo assembly, in which files of forward and reverse reads for a line are
dragged in and set to paired reads, with a maximum size of 350 bases. The reads are then trimmed with DDBuk module (BBDuk Adapter/Quality Trimming Version 38.37 by Brian Bushnell) to discard short reads of less than 100 bp and to remove low quality reads. De novo assembly is then selected from the dropdown of assembly options, and the Geneious assembly algorithm was used on medium sensitivity to construct contigs from the forward and reverse reads.

Following de novo assembly, contigs were mapped to the *P. alba* genome using the Genious mapper at high sensitivity. Contigs that mapped were examined for key components in a specific arrangement, specifically that T-DNA sequence should be present on a single end of the contig. Contigs that failed to map were put through BLAST (Basic Local Alignment Search Tool - https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastSearch&PROG_DEF=blastn &BLAST_PROG_DEF=megaBlast&BLAST_SPEC=OGP__3694__10770) for the *P. trichocarpa* genome using discontinuous megablast to identify possible matches. Contigs that did not generate matches in the *P. trichocarpa* genome and were made up of >100 reads were put through the same discontinuous megablast for all species.

Bioinformatics for the second run was done entirely using Geneious. The similar procedure for de novo assembly from the first sequencing run was performed on each well that generated reads. The bordering genomic area from each contig was mapped to the *P. alba* genome using the Geneious mapper at high sensitivity, and any samples that did not map to genomic areas were put through BLAST for the standard database of all organisms, with program selection for discontinuous megablast.
2.5 PCR for the confirmation of left border locations

Once T-DNA locations were identified for mutants, 4 lines were selected for confirmation by PCR. PCR reactions were set up using a primer within the insert that is 70 base pairs upstream of the capture oligo and a mutant-specific primer selected in the genome near the T-DNA insert but outside of the sequence produced by the Illumina plate sequencing (Figure 10). Amplicons were predicted to be of varying size for each line (Table 6). Annealing temperature was determined with a gradient PCR from 55°C to 69°C at 1°C intervals using FroggaBio 2x Taq Mastermix (FroggaBio, ON). Extension time was determined based on the length of each product according to manufacturer’s suggested extension rate of *taq* polymerase of 1000 bases/min.

![Figure 10: Location of key features for confirming left border insert locations.](image)

A primer to be used for confirming border inserts (in red) was selected approximately 70 bp from the outer nester primer (orange) used in sequencing to ensure that amplified products were generated independent of the same primers used to produce sequencing results.
Table 6. Primers for confirming left border genomic regions of activation-tagged lines.

<table>
<thead>
<tr>
<th>Primer name</th>
<th>Primer sequence (5’ to 3’)</th>
<th>Tm (°C)</th>
<th>Exp. Size (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>T-DNA insert</td>
<td>CGCCTATAAATACGACGGATG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Line 638</td>
<td>CACACGGATTGAAAGGAGAGG</td>
<td>66</td>
<td>817</td>
</tr>
<tr>
<td>Line 316</td>
<td>CAGTCTAGCATATGCTTAGGTAGCC</td>
<td>66</td>
<td>541</td>
</tr>
<tr>
<td>Line 903</td>
<td>GTGCAATATTGAAGGGACGAGG</td>
<td>66</td>
<td>1034</td>
</tr>
<tr>
<td>Line 148</td>
<td>GCATGGGTGGTGAAGTTATGAAAAAGGATC</td>
<td>67</td>
<td>1466</td>
</tr>
</tbody>
</table>

The T-DNA insert primer was used in conjunction with each line-specific primer to generate products of the given sizes at the Tm given. The expected size for each amplicon varied depending on the location of each primer in the genomic flanking region.

PCR products from above were ligated into pGEM-T using the pGEM-T Easy Vector kit (Promega, WI) following the manufacturer’s protocol and using 3 uL of PCR product. Plasmids were transformed into chemically competent *Escherichia coli* DH5α included with the pGEM-T Easy Vector kit using manufacturer’s instructions using 1 uL of plasmid. Transformed DH5α cells were grown overnight on selective LB agar plates supplemented with 100 μg/mL of ampicillin and spread with 100 μL of 100 mM isopropyl β- d-1-thiogalactopyranoside (IPTG) and 10 μL of 100mg/mL 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside (X-GAL). For each line, a white colony was selected and used to inoculate 10 mL of LB with ampicillin [10 mg/mg] and incubated overnight at 37°C. Plasmids were isolated with plasmid extraction kit Invtrogen Purelink Quick Plasmid Mini-prep Kit (and DNA quantity was determined with the Denovix DS-11 FX+ spectrophotometer and confirmed by separation on a 1% agarose gel. Four lines that mapped to single locations were selected for further study and the DNA was sent for Sanger sequencing.
at The Centre for Applied Genomics (TCAG). DNA from line 903 was also used to confirm the
sequence of the T-DNA outside of the nested PCRs used for sequencing.

Results from the TCAG facility were de novo assembled in Geneious against the Illumina
sequencing result and a copied section of genomic material from the mapping/BLAST locations
to confirm the border location sequence, and a BLAST for high similarity sequences in the genome
of *P. trichocarpa* was also used to identify nearby genes.

2.6 Identifying genes bordering T-DNA inserts from selected lines

Sequences from the 4 lines selected for further research had the T-DNA end trimmed and were
used in BLAST to identify locations in the *P. trichocarpa* genome. The genome map of each
location was used to identify genes upstream and downstream that may have been responsible
for observed phenotypes.
Chapter 3. Results

From the field-grown population of activation-tagged *P. tremula* x *P. alba*, 48 mutants were initially selected for sequencing using the Illumina MiSeq platform. These mutants were selected based on the results of a heavy metal (arsenic and cadmium) survivability screen, altered secondary cell wall properties, visible phenotypes. A summary of the mutants, as well as known phenotypes is shown in Tables 1-4. Also included in this study were 6 control lines that include mutants that have already been fully characterized including *fuzzy* (Plett et al., 2010), *rippled leaf* (Williams et al., 2015), *bent leaf* (Harrison et al., 2007) 3 others that remain unpublished, and a sample of wild type. Sequencing of T-DNA insert locations were carried out in two separate sequencing runs, with slightly different procedures for each.

3.1 Sequencing and assembling the first run

For the first sequencing run, 48 lines were barcoded and sequenced in a single pool. Following sequencing, Illumina output files were de-multiplexed to separate the 48 individual DNA samples by barcode. A total of 614,372 pairs (1,228,744 individual reads) were obtained and a summary of the results are shown in Figure 11. Approximately 80% of the sequenced pairs (506,572 pairs) were confirmed to contain the T-DNA sequence, dropping 107,800 pairs from further bioinformatics processes. This indicated that approximately 20% of the total fragments sequenced were due to non-specific binding of the modified TAIL-PCR primers. The number of pairs with barcodes varied from well to well (Figure 11). Two wells (A2 and C3) were found to have no data due to the plate wells breaking.
Early attempts to use command line software hit unexpected problems. Due to failures in the alignment process, it was impossible to continue with them. The expected result of *de novo* assembly would be the generation of a single or double contig, however, when the program Velvet was used it produced far more (varied by sample). An attempt to use BLAST confirmed the *de novo* assembly results, indicating that there were many inserts scattered across the entire genome, but unable to better visualise why I was not getting a single location instead of many. The bioinformatics tool Geneious Prime (version 2019) offered multiple sequence assembly options, as well as the ability to visualise data at each step of the assembly process. Using the built in *de novo* assembly process, reads for each well were assembled into contigs. Wells varied in the number of contigs generated for each, ranging from a single contig to 981 (Figure 11). When examined as a whole, over 1000 unique contigs were generated from the reads in all wells. The number of reads used to assemble each contig was uneven, with some contigs made up of large numbers of reads, and others made up of only 2 (Figure 12).
Figure 11. Heat maps of the first sequencing run.

A: The number of paired reads per well displayed high variability from well to well, where lighter colours indicate fewer reads and dark colours indicate higher number of reads. Two wells, A2 and C3, were broken during shipment and had no data. B: The number of reads possessing the T-DNA border also varied well to well, with some wells discarding up to 90% of their reads. C: Reads from all wells assembled into multiple contigs. Green numbers in all maps represent placement of control lines.

Figure 12. Contigs have uneven distribution of reads.

The distribution of reads is uneven between contigs. A: Some of the contigs are made up of large numbers of reads, with most of the reads from the well being formed into 3-5 contigs. B: a much larger number of contigs are made up of only individual read pairs.
When mapped to the *P. alba* genome, contigs from each well mapped to multiple locations (Figure 13). Several of those locations were recurring throughout the samples, suggesting non-specific binding of the PCR primers. As an example, well A4, containing line 1366, had contigs mapping to 9 different chromosomes, of which 3 chromosomes had contigs mapping to more than one location in the chromosome, and 2 additional contigs mapping to scaffolds not yet associated with a chromosome (Figure 13-A). One of those locations, on chromosome 1, was found to be present in 15 different samples. There was also some variation observed in the structure of contigs when mapped (Figure 13-B and C).

**Figure 13.** Example of contigs mapping to multiple genome locations and displaying different sizes and structures.

A: Contigs map to multiple locations within each sample suggesting non-specific amplification. B: Contigs are expected to have a mismatch on one end as shown in Contig 21 (A4). C: While one end is expected to have a low level of similarity to the genome, many contigs possess several such mismatches as seen in contigs 69 and 119 (well B1).
Despite the presence of multiple contigs in samples mapping to the *P. alba*, a large number of discarded contigs, and some variation in the structure of contigs, largely the structure of the data matched the expected results. Contigs that mapped to the genome of *P. alba* had high levels of similarity to the genome, except for one end where the remainder of the insert was left behind. Some mapped locations had contigs separated by short gaps (Figure 14), consistent with longer fragments with both ends sequenced. Unfortunately, while the data had some good properties, the variation in read and contig numbers due to non-specific amplification and samples having contigs that mapped to multiple locations left a need to address these issues to determine if this would be a viable method for identifying T-DNA insert locations in a high-throughput manner.

**Figure 14.** Example of mapped locations displaying gaps between T-DNA and genomic ends of sequenced fragments.

In this example, contig 21 (top) mapped in a forward orientation, and contig 23 (bottom) mapped in a reverse direction. Note the small partial mismatch at the leftmost edge of contig 23, which is where the small fragment of left border region from the insert is present.
3.2 Sequencing and assembling the second run

The low success of the first sequencing run prompted the change of a few steps in the protocol as described in the methods (Section 2.2), adding the capture step to before the PCR. After refinements to the methods were completed, I increased the number of samples from 48 to 96 and loaded control samples into multiple wells. A second plate containing 96 samples was sequenced using the same platform and parameters and analysed. Similar to the first sequencing run, Illumina output files were de-multiplexed to sort wells by barcode. Once de-multiplexed, the output files were still found to vary in size, though not as widely as the first sequencing run (Figure 11). A total of 866 pairs (1732 individual reads) were obtained. All pairs were confirmed to contain capture sequence. There were no broken wells in the second run, and while the wells did not possess the high level of variation seen in the first run, an additional problem made itself known - many of the wells had no sequence data when demultiplexed, as if there was no DNA in them to be sequenced.

In second seq run, the reads per well did not display as much variability in the number of reads present when compared to the first run (Figure 15), with most wells having fewer than 100 reads. All read pairs were found to contain the T-DNA border region (Figure 16), with no reads discarded. The lower number of reads in each well translated into a reduction in the number of contigs generated once de novo assembly was completed (Figure 17), and in contrast to the first sequencing run, where numerous of contigs were generated from de novo assembly, most wells generated a single contig from their reads. The lower number of reads were able to be de novo assembled much quicker than in the first run, where reads frequently numbered in the
thousands. When compared to the first sequencing run, the greatly reduced number of contigs indicate that putting the capture step before PCR during the library construction had the desired effect of increasing the specificity fragments of DNA sequenced.

**Figure 15. Comparison of read numbers from first and second sequencing runs.**

Reads from the second run (capture sequencing) displayed better normalisation than the first sequencing run (TAIL-PCR). Wells with green numbers represent control lines, and those with a B in them indicate blanks (wells with no DNA included). White wells represent samples that had no reads.
Examination of those contigs that mapped to single locations demonstrated expected T-DNA mismatches to the genome on single ends (Figure 18) and did not show T-DNA in multiple spots as seen in the first run (Figure 14). As well, several contigs mapped to a single location (Figure 19), in comparison to the first run where wells often had contigs mapping to multiple locations in the *P. alba* genome. Samples that mapped to multiple locations indicate possible insertion in highly conserved sequences that were replicated during the *Populus* genome duplication events (Figure 20). Contigs from 4 lines failed to map to any genomic location, and a BLAST search identified the possibility that, in addition to the insert, DNA from the *A. tumefaciens* plasmid At may have also integrated genomically at the time of transformation (Figure 21), disrupting the identification of flanking genomic regions.

While the second run did not perform as efficiently as expected (far fewer wells had data to be analysed than expected), the quality of the results was much higher. Three indications of the improved quality were the reads assembling into fewer (in most cases single) contigs, Not discarding contigs when mapping to the genome, and being able to use those contigs in either mapping or BLAST processes to identify single likely points of T-DNA insertion.
Figure 16. All read pairs in the second sequencing run were found to contain the T-DNA.

In contrast to the fist sequencing run, where wells discarded up to 99% of the read pairs if they didn’t contain the T-DNA border sequence, the second sequencing run kept all read pairs, demonstrating the improvement in specificity that the capture process has added.
**Figure 17.** The second sequencing reads assembled into far fewer contigs than the first run.

Read pairs from the second sequencing run assembled into far fewer contigs than the first run, where some wells had over 100 contigs generated. Wells in the second run formed single contigs, indicating increased specificity from the incorporation of the capture oligo.
Figure 18. Captured T-DNA demonstrates improved specificity for flanking genomic material and enables identification of T-DNA insert locations.

The clean assembly into a single contig, and the mapping to a single location in the genome of *P. alba* demonstrate that the improvement in read quality allow this method to be used to identify the location of T-DNA inserts in poplar.

Figure 19. Contigs from second sequencing map to single genome locations.

A lower number of contigs map to locations in the *P. alba* genome, however the mapping generated single locations rather than the multiple locations seen in the first sequencing run, indicating improved specificity in the capture process and improving the overall methodology for use as a high-throughput technique for identifying T-DNA inserts.
Figure 20. Sequences that map to multiple locations may indicate insertion in duplicated regions.

The most recent poplar genome duplication event (the salicoid event) resulted in genomic rearrangement (Tuskan et al., 2006), after which chromosomes shared homologous gene blocks. Chromosomes are indicated on the left by number (I to IXX). The same pattern and linkage are indicated to the right, with common colour blocks indicating homologous DNA fragments. T-DNA inserts for lines 1694 and 1290 were located in chromosomes 12 (XII) and 4 (XIV) respectively (inset). Adapted from Tuskan et al., 2006.
Figure 21. Sequencing of some T-DNA sites reveals similarity to Agrobacterium plasmid DNA.

An example of a flanking sequence that aligned with the DNA of cryptic plasmid At present in A. tumefaciens. GenBank display of DNA from line 351 (Query) compared to A. tumefaciens (Subject).

3.3 Confirming left border inserts through PCR and Sanger sequencing

Once sequence assembly and mapping were completed, it was necessary to confirm T-DNA insert locations. Using a forward primer located the insert (common to all 4 lines selected), and primers located in nearby flanking regions unique to each line, PCR amplicons were generated that offered presumptive confirmation of insert location based on amplicon size (Figure 22). Sanger sequencing subsequently confirmed the flanking genome sequence for all 4 lines.

Table 7. T-DNA insert locations in 4 selected lines.

<table>
<thead>
<tr>
<th>Line #</th>
<th>Chromosome</th>
<th>Location (bp)</th>
<th>Phenotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>148</td>
<td>19</td>
<td>15,738,161</td>
<td>Low xylose</td>
</tr>
<tr>
<td>316</td>
<td>1</td>
<td>4,417,839</td>
<td>High glucose/xylose</td>
</tr>
<tr>
<td>638</td>
<td>14</td>
<td>7,765,422</td>
<td>Extreme signs of stress from heavy metals</td>
</tr>
<tr>
<td>903</td>
<td>3</td>
<td>16,287,262</td>
<td>Leaf phenotype</td>
</tr>
</tbody>
</table>

*Insert locations for selected lines confirmed by Sanger sequencing of flanking genomic DNA.
Figure 22. PCR results of left border confirmation.

PCR conducted on 4 lines (top) to confirm the left border insertion used a common T-DNA primer and generated amplicons of unique size for each line.

3.4 Identifying nearby genes for future study in lines with confirmed inserts

Once locations were confirmed, DNA sequences were put through BLAST and genome locations reviewed. Gain-of-function phenotypes have been reported as far away as 11 kb (Hsing et al., 2007), so multiple genes in both downstream and upstream directions were reviewed. In line 148, insertion was found to have occurred into the 5’ UTR of the HAK1 gene (Figure 23). Figures 23-26 summarise the locations of the T-DNA insert and the genes found nearby.
**Figure 23. Location of T-DNA activation-tagging insert in line #148.**

The activation-tagging T-DNA insert LB in line 148 is located in chromosome 19, at 15,738,161, inside the 5’ UTR of the *HAK1* gene (POPTR_019G131800v3). Immediately flanking it upstream is ROOT PHOTOTROPISM PROTEIN 2 (POPTR_019G131600v3), and immediately downstream is 60S RIBOSOMAL PROTEIN L10 (POPTR_019G131900v3).

**Figure 24. Location of T-DNA activation-tagging insert in line #316.**

The T-DNA insert in line 316 is located at in chromosome 1 at 4,417,740 with 2 genes flanking it within approximately 10kb. The flanking gene upstream (POPTR_001G058100v3) is an uncharacterised gene with kinase domains, and the one downstream (POPTR_001G058200v3) is MADS-box protein EJ2.
**Figure 25. Location of T-DNA activation-tagging insert in line #638.**

Line 638 was selected for its phenotype during a heavy metal screen with arsenic and cadmium. The T-DNA insert is located in chromosome 14 at position 7,765,422. Located about 10 kb upstream is CYCLIC PYRANOPTERIN MONOPHOSPHATE SYNTHASE (POPTR_014G099500v3). Downstream approximately 8 kb is transcription factor BASIC HELIX-LOOP-HELIX PROTEIN 13 (POPTR_014G099700v3).

**Figure 26. Location of T-DNA activation-tagging insert in line #903.**

The T-DNA was located in chromosome 3, position 16,287,262. Upstream approximately 5 kb away lysophospholipase BODYGUARD 3 (POPTR_003G147400v3) is located. Downstream almost 10 kb away is ENDOGLUCANASE 1 (POPTR_003G147600v3).
To summarise, the first sequencing run made use of a modified TAIL-PCR which attempted to directly identify T-DNA flanking regions in a manner similar to Singer and Burke (2003), but without the use of expensive Genome Walking kits. After discovering non-specific products in numerous wells, a capture probe as per Lepage et al. (2013) was included in the method to increase specificity in the second sequencing run. The improvement in quality from sequence capture was apparent in the second run, with all reads possessing the T-DNA border, a reduction in the number of contigs per well, and with 10 wells giving single locations when contigs were mapped to the *P. alba* genome. The improvement in quality also provided insight into lines that mapped to multiple locations indicating T-DNA may have inserted into areas duplicated in the *Populus* genome, as well as genomic rearrangement that incorporated *Agrobacterium* At plasmid DNA into the genome of some activation-tagged *Populus* lines.
Mutant populations in plants are incredibly valuable tools, both for identifying genes that are involved in a wide range of cellular processes. The Regan lab’s population of *P. tremula x P. alba* activation-tagged mutants has revealed a large number of important phenotypes ranging from trichome development (Plett *et al.*, 2010), leaf morphology (Plett *et al.*, 2014), stem structure, tree form and root architecture (Harrison *et al.*, 2007) and lead to the discovery of many new genes. Unlike similar populations in model systems like *Arabidopsis*, the frequency of visible phenotypes is much higher (more than 7X) in this poplar population, likely due to the long life of the plant and growth in natural field conditions for many years that expose the trees to many stresses that can reveal phenotypes (Busov *et al.*, 2011).

One of the time consuming and labour-intensive steps with gene discovery using mutant populations is identifying the site of the mutation that is causing the phenotypic effect. Two technologies have traditionally been used, plasmid rescue (Behringer and Medford, 1992; Kemppainen *et al.*, 2008) and TAIL-PCR (Singer and Burke, 2003) but neither of these are suited for high-throughput applications. The recent advances in large-scale sequencing, and my approach to use barcodes, pooling of samples, and sequence capture before sequencing holds promise to make large-scale sequencing affordable.

**Deciding on the best sequencing technology for this project**

In considering large-scale sequencing to identify the location of the T-DNAs in this population, the read size, coverage, accuracy, cost, and availability of several major sequencing platforms were considered (Table 8). Read size is an important factor to consider when
identifying T-DNA flanking locations, since enough genome sequence to locate insertion sites is needed (Singer and Burke, 2003). Coverage and accuracy come into effect when looking at the reads themselves – Illumina platforms make up for the possible errors in sequencing by having very high coverage, something other technologies have still have difficulty with (Li and Harkess, 2018; Loman et al., 2015). Cost per run, which ultimately becomes cost per mutant line, is a primary concern in developing a cost-effective method. Availability of the platforms, and therefore the demand on them and the wait time for data, had to be considered as well; for example, the PacBio SMRT platform is only listed as being available at one location in Canada, so demand and wait times can be anticipated as high. Other considerations that were difficult to quantify, and not included in the table, are the availability of support and the difficulty in troubleshooting problems.

Table 8. Comparison of sequencing platforms.

<table>
<thead>
<tr>
<th>Platform</th>
<th>Read size</th>
<th>Coverage</th>
<th>Accuracy</th>
<th>Cost</th>
<th>Availability</th>
<th>Overall Suitability</th>
</tr>
</thead>
<tbody>
<tr>
<td>WGS on HiSeqX</td>
<td>NA</td>
<td>Medium</td>
<td>High</td>
<td>Medium</td>
<td>Medium</td>
<td>***</td>
</tr>
<tr>
<td>Illumina MiSeq</td>
<td>2x300</td>
<td>Medium</td>
<td>High</td>
<td>Low</td>
<td>High</td>
<td>*****</td>
</tr>
<tr>
<td>Illumina NextSeq</td>
<td>2x150</td>
<td>High</td>
<td>High</td>
<td>Medium</td>
<td>Low</td>
<td>***</td>
</tr>
<tr>
<td>Illumina Novaseq</td>
<td>2x250</td>
<td>Very High</td>
<td>High</td>
<td>M. to High</td>
<td>Low</td>
<td>**</td>
</tr>
<tr>
<td>PacBio SMRT</td>
<td>&gt;1kb</td>
<td>Low</td>
<td>Medium</td>
<td>High</td>
<td>Low</td>
<td>*</td>
</tr>
<tr>
<td>Oxofrd Nanopore</td>
<td>&gt;1kb</td>
<td>Low</td>
<td>Medium</td>
<td>High</td>
<td>NA</td>
<td>*</td>
</tr>
</tbody>
</table>

Sequencing platforms were evaluated by several factors to determine their suitability for this project. NA refers to Not Applicable.

The following considerations were made in selecting the best sequencing technology for the P. tremula x P. alba activation-tagged population.
1) The population is large

The population of activation-tagged mutant poplar in the Regan lab field site numbers approximately 2000 lines. Using TCAG pricing, WGS for each line ($800/line with 15x coverage, including library preparation) would cost $1.6 million - a price far outside the budget for all but the best funded projects. Continuing with TCAG pricing for other Illumina sequencing options, prices range greatly depending on read size, but quickly reach costs too great for all lines to be sequenced individually. Oxford Nanopore was initially considered to be quite low, with only $1000 for a device, however flow cells are expensive ($1000 each) and can be easily damaged by improper handling. In addition, this is also a relatively new technology and both accuracy and coverage are low. Of all the options examined, the PacBio SMRT sequencing was the most expensive, with cost per line potentially reaching thousands of dollars. When discussing the cost with Brian Boyle, Director of the PAG at University of Laval, the use of barcodes was introduced, and with them the ability to reduce cost. The cost breakdown for 96 wells individually tagged is approximately $1992.00, or $20.75 per line using a MiSeq 300bp paired end flow cell. By barcoding lines, the same 96 samples could be sequenced for $1080, or $11.25/line, bringing the cost for sequencing the entire population to $22,500.

2) Short read length

Previous studies of T-DNA have shown insertion sites can be identified with stretches of flanking DNA ranging from 100 to 1500 bp in length (Boyle et al., 2013; Singer and Burke, 2003). Therefore, whole genome sequencing is not necessary, and high coverage from NextSeq and NovaSeq platforms are wasted (Wong et al., 2013). Both the use of TAIL-PCR and capture
enrichment can focus the sequencing onto only the region of interest. This has the potential to maximise the amount of genomic flanking material sequenced, improving the ability to confirm insertion sites.

3) Accuracy and coverage

Looking at the sequencing technologies, while long read sequencing has its appeal, it is also known to be more error prone (Koren et al., 2012) than established second generation technologies such as the Illumina platforms. *Populus* have undergone multiple genome duplication and genome rearrangement events (Hou et al., 2016), making accuracy a key concern. Part of that accuracy is achieved in Illumina sequencing by having a region sequenced multiple times. A review of papers with a focus on sequencing suggest that 25X coverage will be sufficient to get highly accurate output. In addition to using TAIL-PCR to amplify a region of interest (Singer and Burke, 2003), a recent paper on locating T-DNA inserts in *Arabidopsis* (Boyle et al., 2013), suggested the possibility of using an oligonucleotide complementary to the region of interest (the insert borders) to capture it for sequencing, allowing a high coverage on a very short piece of DNA. This would resolve the need for reasonable 25x coverage and allow reduction in cost without sacrificing read length.

All of these details were brought together in choosing the MiSeq platform. Like other Illumina technologies, the MiSeq system is known for its high accuracy and coverage. Being a well-developed 2nd generation platform, the availability is high, and many institutions possess systems. It offers a range of read sizes, maxing out at 300 bp paired-end flow cells, and with the difference in sizes comes scalability in cost. In addition to all these considerations, being well
established also makes troubleshooting problems and additional support (both from the manufacturer and other researchers) available. Taken together, all these factors make MiSeq sequencing the best option for this project.

4.1 Sequencing and assembly of first sequencing run

The first sequencing run made use of modified TAIL-PCR to amplify the region around the LB insertion site. The principle is similar to the TAIL-PCR employed by Singer and Burke (2003), which was subsequently developed into Genome Walking™ kits by Takara Bio. The incorporation of barcodes to the sequencing adapters allowed samples to be multiplexed and sequenced together, making better use of the flow cell’s capacity.

The first sequencing run generated a very large dataset to analyse. With the exception of the 2 broken wells, every well produced some reads. De-multiplexing and trimming of the reads was completed successfully. The initial analysis of the output in Figure 11-B showed that, with very few exceptions, at least 30% of the reads in every well contained the probe sequence, indicating that the T-DNA border would be present and identifying insert locations should be possible. There were inconsistent numbers of reads from each well. Despite the fact that DNA library preparation included a normalisation step in which the same concentration and volume (10 ng/ul) of DNA was added to each well, there were significant difference in the number of sequencing reads from each well ranging from 2 (in a broken well) to 410,482 with the average being 12,822 (Figure 11-A). Shown in Figure 11-B, the variation is not only in the number of reads,
but in how many of them contain the T-DNA border - there are sequences sequenced that do not have the capture probe present. This suggests that the TAIL-PCR, while amplifying some the targeted T-DNA, is not as specific as it should be and is generating a large number of non-specific products.

In addition to the problem with normalisation, the initial attempt to analyse the data using standard command line software was unable to function. An unexpected error during alignment forced me to abandon that method of identifying insert locations, as I was unable to find any specific error that would cause a failure in converting from one file type to another. The completion of the first half of the alignment was completed without any issue, and the cause of failure remains unknown. Subsequent to the failure of alignment, I moved to the assembly tool Velvet, which was able to convert reads into contigs. The large number of contigs produced when used caused uncertainty as to whether it was functioning correctly (or more likely if an incorrect operator had been used). The expected result was a single or double contig, indicating single or double insertion events, as have been previously documented in the population; the generation of dozens of contigs and multiple matches to the genome was unexpected. Using BLAST to try and align reads with the genome produced the same result as Velvet – insertions were numerous and scattered throughout the genome. That left the question of whether the T-DNA left border was inserting multiple times, but because each event was unique previous use of Southern blots were unable to detect the. I needed a better way to visualise the output at each step in order to analyse the data and understand if this was accurate new information about the mutant lines, or if it was being caused by some other factor.
Geneious Prime is a bioinformatics tool developed by Biomatters Ltd. It offers the ability to plug in a number of commonly used assembly and alignment tools, such as Bowtie and ClustalW, as well as having its own built-in algorithms. When Geneious Prime was used to de novo assemble reads into contigs, reads displayed lengths ranging from 55 to 300, with the N50 varying from well to well and dependant on the length of the reads (Appendix H). The second problem was in the number contigs generated. As seen in Figure 13 -C, the number of contigs in a well ranged from single digits to nearly a thousand. The contigs from all of the wells can be clustered into 2 groups. The first group consists of a small number of contigs, each of which had a large number of reads; overall they made up over 50% of the reads in a well. The second group is made up of many more contigs, sometimes numbering over 200, and these contigs were only made up of a few reads, usually 2 - 1 forward and 1 reverse – which indicated that such contigs were the result of single fragments of anomalous DNA picked up in the sample.

The high variation in the number of contigs generated for each well does not appear to be correlate to the number of reads; that is, wells with more reads did not necessarily generate more contigs. For example, well C2, which contained 1066 pairs assembled into 35 contigs; well A1, with only 205 pairs, formed 56 contigs (Figure 11-A and 11-C). In addition, the same contigs presented themselves in multiple wells, suggesting that there was either a form of cross contamination or some other effect. The most convincing point for some form of contamination was the presence of T-DNA in well H3, which was supposed to contain wild-type DNA, and therefore should not have presented any T-DNA. Reviewing the DNA extractions, samples were examined for the presence of the common contigs found in multiple wells, but no correlation was found. This means that some contigs commonly found in wells were not the result of
contamination during DNA extraction, but from some other source further in the process. Given the damage sustained to 2 of the wells on the plate during shipping, the sequences were examined for cross-contamination at a plate level, but there was no pattern indicating such, and no other damage was reported by the sequencing company. The factor that was most convincing that cross contamination was a not the cause of issues in the assembly of contigs was in the number of contigs generated. Assuming for a moment that each sample contaminated each other sample, I would expect to see at most 48 contigs present in a well. Instead, wells often displayed hundreds of contigs, with a large amount of variation, as seen in Figures 11-C.

The most concerning development was found in the contig structure. Contigs often displayed unusual internal structures with regards to the location of the capture oligo and left border, which was expected to be at one end (Figure 13). Contigs that failed to map to locations in the genome of *P. alba* often had more of these disruptions than those that did map. This is a not a problem with the capture oligo – the capture sequence had to be present to make it past the sorting step early on, and in fact was often present in more than one location on these contigs. While the rearrangement could have occurred during T-DNA insertion years ago when the mutant lines were generated, if that were the case the representation of these contigs would be expected to be higher.

Taken together, these points reinforced my belief that the TAIL-PCR generates non-specific products which, because of the arrangement of the Illumina adapters, are then carried through to sequencing. The result is fragments of DNA that contain properties that we are looking for (the T-DNA), but in arrangements that are highly unlikely to be present in the samples. While the read quality was acceptable, and some locations appear valid, these issues present problems to
be overcome before this new method of locating inserts can be deemed viable for widespread use. If the PCR is introducing non-specific sequences into the capture process, the most effective way to counter it would be to perform a step to enrich the fragment of interest, such as sequence capture, first then use PCR to amplify the enriched product from capture.

4.2 Sequencing and assembly of the second run

Prior to the second sequencing run, careful planning went into correcting the biggest issues with the trial run – that when assembled, there were more contigs than samples, and that when mapped to the *P. alba* genome multiple locations per sample were revealed. To address the mapping to multiple locations, triplicate samples of control lines were loaded into a second plate. If contamination were occurring as an artifact of the pooling process, then different well locations would have contigs that map to different places in the genome based on the samples they mixed with; it would then be possible to identify the correct location by looking for mapped locations that are present in all 3 wells. Based on the results from the first sequencing run, and in consultation with our collaborator at Universite Laval, Brian Boyle, it was decided to perform the sequence capture step before any PCR steps, in the belief that removing the excess genomic material prior to the final PCR would reduce any non-specific amplification that might be occurring as observed in the first run.

In the second sequencing run wells had far fewer reads than the first run, and while there was some variation it was far less than with the first sequencing run (Figure 15). The number of reads overall was reduced, with the highest number of reads being about 100 compared to
hundreds of thousands in the first sequencing run. All reads contained the T-DNA sequence and assembled into single contigs almost without exception (Figures 16 and 17). Unlike the first sequencing run, which had T-DNA sequences appear in the WT sample (Figure 11), blank wells in the second run did not have T-DNA present, indicating that the library preparation and transport of the samples was not a likely a factor in the high number of reads or contigs in different wells mapping to the same location. Approximately 60% of the contigs from the second sequencing run mapped to single locations in the *P. alba* genome. Indicating that the addition of the capture step before the final amplification resulted in higher quality reads and contigs.

When reviewed in BLAST with the *P. trichocarpa* genome, some insert locations were found to be highly conserved, indicating T-DNA insertion into regions that were doubled during the last genome duplication event in the evolution of poplars (Figure 20) (Tuskan et al., 2006). The difficulty in identifying which location is correct is compounded by the 717-1B4 line being a hybrid of *P. tremula* and *P. alba*, and the genome in BLAST being that of *P. trichocarpa*. Those that did not map or have BLAST results in poplar species were found to contain DNA from the cryptic plasmid At of *A. tumefaciens*, which is a known occurrence when using *Agrobacterium*-mediated transformation; fragments of plasmid DNA can become incorporated into the genome of the transformed organism (Gelvin, 2003). Mutant lines that belong to either of these groups may require the use of long read sequencing for identifying insert locations, to move past fragments of bacterial DNA or genome rearrangements and into unique portions of the poplar genome. One option for long read sequencing would be to use this method to identify candidates, use Covaris shearing and barcoded library prep with the same capture step, but finish with sequencing being done on a long-read platform such as Oxford Nanopore.
Unfortunately, 53% the wells did not display any data at all. Wells that did not have DNA did not display any recognizable pattern. This may represent a problem with normalisation, in that wells did not contain the same quantity of DNA despite the use of the Denovix Broad Range assay kit. This would result in some wells in the second run having few or no reads and may also be part of the reason for such high variation in read numbers from the first sequencing run. While the normalisation may continue to present difficulties as a high throughput method, there were still 30 wells that displayed sufficient improvement to proceed with mapping or BLAST for identifying insert locations. Of those 30 wells, 10 unique lines mapped to single locations (Figure 19), 7 wells had BLAST results that identified *A. tumefaciens* plasmid DNA that inserted at the time of transformation (Figure 21), 3 lines did not produce high quality matches to either poplar or bacterial genomes, 2 lines produced BLAST results to regions with known genome duplication events (Figure 20), and 1 resulted in matches to unplaced scaffolds. Some of the samples were replicates, and those produced the same contig sequence and BLAST or mapping results, confirming that the capture method is able to generate reproducible results, and giving a high degree of confidence that the sequenced locations are accurate representations of T-DNA insertion.

### 4.3 Confirming left border inserts through PCR and Sanger sequencing

T-DNA insertions were further confirmed through PCR and Sanger sequencing. Given that T-DNA can insert in a range of configurations (Robinson *et al.*, 2009), and cause rearrangements like the insertion of *A. tumefaciens* DNA, it’s necessary to check outside of the sequence generated from the capture method. Based on the locations predicted by sequencing, and using
the *P. alba* genome sequence, primers were designed to confirm the location of the DNA for 6 activation-tagged lines. Of those 6, 4 lines produced single amplicons at their designed Tm. One of those lines (line 1290) was located in a region that may have undergone genome duplication and rearrangement and generated multiple amplicons, indicating the need to move further into the genome to find a unique location in which to base emplification.

4.4 Identifying candidate genes for future study in lines with confirmed inserts

Having confirmed the location of T-DNA inserts, the final step in characterising the lines would have been to identify what genes are overexpressed as a result of activation tagging. This would also have confirmed that the insert remained intact and that the CaMV 35s repeats were in that location in the genome. It should also be noted that while genes are immediately surrounding the T-DNA insert, they need to be actively expressed for the enhancer subunit to cause overexpression (Weigel *et al.*, 2000); there is no guarantee that they will be upregulated in all circumstances, or at all. Activation tagging has been known to upregulate genes from as close as 10kb or as far away as 50 kb, depending on unknown factors (Regan lab, unpublished).

**Line 148 (Figure 23)**

Line 148 (Figure 23) was selected for its low xylose content in the wood. Xylans are a hemicellulose made up of xylose molecules that are responsible for increasing the strength of plant cell walls and increasing recalcitrance to enzymatic digestion (DeMartini *et al.*, 2013). The receptor kinase predicted from the *A. thaliana* gene At5g58300 is inactive, and no expression has
been detected in poplar leaf cells according to the bioinformatics tool UniProt (https://www.uniprot.org/). ROOT PHOTOTROPISM PROTEIN 2 has been well studied (Haga et al., 2015), and is unlikely to play a role in this phenotype. The same can be said for the ubiquitin-conjugating enzyme and the ribosomal protein L10. While some K homology proteins have been implicated in flower development (Mockler et al., 2004), there is no evidence to support that such a gene would be responsible to secondary cell wall biosynthesis.

The insertion into a leucine-rich repeat (LRR) receptor-like serine/threonine-protein kinase may be responsible for the low xylose phenotype. Serine/threonine receptor kinases are a diverse family of genes that have been implicated in the response to plant pathogens (Afzal et al., 2008) such as fungi, and some members of the family have previously been implicated in secondary cell wall modification (Draeger et al., 2015). Since the insertion was in the 5’ UTR could still produce a viable protein, the translocation may be affected for a single copy of the gene, however with a single copy functional it is possible that the CaMV35S enhancer sequence could overexpress the gene to the point where it would be able to compensate.

**Line 316 (Figure 24)**

Mutant #316 was selected for its high glucose levels, in Figure 24 the insertion occurred in an uncharacterised lncRNA (LOC112328619). Upstream, 3 uncharacterised proteins of varying lengths were located, potentially representing new genes linked to secondary cell wall properties. The nearest gene (POPTR_001G058100v3) has also been identified in a screen to determine possible genes related to forest tent caterpillar herbivory (Ralph et al., 2008). This
gene could have numerous effects on the response of the tree to herbivory, ranging from production of glucosinolate compounds for defense to improved immune response to deal with opportunistic pathogens, however it needs further investigation.

Following the T-DNA are two transcription factors, MADS-box protein EJ2 and a homolog of agamous-like MADS-box protein AGL9. The AGL9 gene is unlikely to play a role in high glucose, as it has been well studied in its role in regulating flower development (Mandel and Yanofsky, 1998). MADS-box genes have been identified as important factors in abiotic stress response in a range of species (Schilling et al., 2018), and may be of interest in further work on improving poplar tolerance to heavy metals or for other bioremediation efforts. It has also been implicated as playing an important role in crop domestication, having roles in plant development (Castelán-Muñoz et al., 2019), and in particular flowering traits. Because transcription factors can affect many genes simultaneously, either gene may play a key role in the high glucose levels detected (Davies et al., 1996), and both of these genes may be of interest for crop species, such as hazelnut or apples.

Line 638 (Figure 25)

Line 638 was selected for its phenotype during a Regan lab heavy metal screen with arsenic and cadmium. While there are several genes nearby, neither UV-stimulated scaffold protein A homolog or aspartyl protease family protein 2 stand out as genes that would have potential for bioremediation. Upstream is CYCLIC PYRANOPTERIN MONOPHOSPHATE SYNTHASE, an enzyme pathway that interacts with other metals, such as molybdenum (Kruse et al., 2018). There have been occasions in which enzymes that react to one metal have been found to also interact with
others (Wu et al., 2011). However it is far more likely that transcription factor BHLH13, which interacts with the jasmonate stress response pathway (Huang et al., 2018) or ETHYLENE-RESPONSIVE TRANSCRIPTION FACTOR 23 are involved, given the “extreme stress” response described in the results of the heavy metal screen.

Ethylene response factors are a very large family of transcription factors that have been found to regulate plant growth and response to abiotic stresses (Dey and Corina Vlot, 2015). To date, cadmium uptake studies conducted in A. thaliana and Oryza sativa indicate an increase in ethylene biosynthesis during exposure to heavy metals (Abiri et al., 2017), leading to a change in osmotic potential in cell membranes and the packaging of heavy metal ions into transport vesicles (Abiri et al., 2017). The gene BHLH13 has been implicated in the regulation of the jasmonic acid pathway (Huang et al., 2018). The BHLH subgroup IIId transcription factors (BHLH3, BHLH13, BHLH14 AND BHLH17), function as transcription repressors to negatively regulate jasmonic acid responses. Studies of jasmonic acid effects in response to cadmium treatment in rapeseed indicate that higher levels of jasmonic acid will reduce the effects of cadmium exposure (Ali et al., 2018). In the case of an activation-tagged mutant, over expression of the BHLH13 would reduce the biosynthesis of jasmonic acid, limiting the tree’s ability to respond to the increased stress of cadmium exposure. The “extreme stress“ phenotype could be explained by the reduction in jasmonic acid as the overexpressed bHLH13 represses synthesis.

**Line 903 (Figure 26)**

Finally, line 903 was selected for a leaf phenotype, in which the leaves had a waxier appearance than those of trees nearby. The T-DNA was located in chromosome 3, at position
16,287,263, surrounded by genes that are related to plant stress response. Upstream are the MEDIATOR 10b (MED10b) genes and the transcription factor MYC2. Downstream the BODYGUARD3 gene (BDG3), and downstream almost 10kb is endoglucanase 1. While it is possible that the endoglucanase 1 gene downstream could influence leaf structure through manipulation of secondary cell wall biosynthesis (Glass et al., 2015), it is far more likely caused by BODYGUARD 3, a controller of leaf cuticle biosynthesis (Kurdyukov et al., 2006). Leaf cuticle is responsible for a range functions such as defense against pathogens and abiotic stresses (Kurdyukov et al., 2006), and further investigation could lead to useful discoveries in improving plant response to disease and environmental stresses. The presence of other stress-response regulatory genes could also play a role, and whose upregulation may mean that line 903 is more sensitive and responsive to environmental signals. Transcription factor MYC2 is known to be involved in regulating crosstalk between several important plant hormone pathways, and regulation of the jasmonic acid-mediated metabolites (Kazan and Manners, 2013). A subunit of the mediator complex (MED10b), an important regulator of transcription, is also nearby; while the exact function is unknown (Kidd et al., 2011), it is known to form an important role in the formation of the central structure of the complex (Maji et al., 2019).
Chapter 5. Conclusions

In the present study, I have used a combination of DNA barcoding and sequence capture with high-throughput sequencing to construct a method for the identification of T-DNA inserts in an activation-tagged population of *P. tremula x P. alba*. Using this method, I was able to identify T-DNA insert locations in 10 unique lines and further confirmed 4 of those locations through the use of PCR and Sanger sequencing. In addition, I have identified lines that have t-DNA inserts within areas that have undergone genome duplication and lines with T-DNA inserts that contained *A. tumefaciens* DNA, in which it may be necessary to switch to a longer sequencing method to move past the bacterial DNA and into the genomic flanking regions. Unfortunately, while the method has been successful in identifying insert locations, the efficiency is still quite low. Although further work is needed before the method can be applied as a high-efficiency means of identifying T-DNA inserts, advances to date indicate it is a matter of refining the existing process, and the successful location of flanking region sequences presents a useful tool for improved use of the Regan lab activation-tagged *P. tremula x P. alba* population.
References


ecophysiology, and chemical and metabolomic composition of wood of *Populus trichocarpa*. BMC Genomics 20.


novo assembly of single-molecule sequencing reads. Nature Biotechnology 30, 693–700. https://doi.org/10.1038/nbt.2280


Marshall, D.A., MacDonald, K. V., Robinson, J.O., Barcellos, L.F., Gianfrancesco, M., Helm, M.,


Ye, X., Busov, V., Zhao, N., Meilan, R., Mcdonnell, L.M., Coleman, H.D., Mansfield, S.D., Chen, F., Li, Y., Cheng, Z., 2011. Transgenic Populus trees for forest products, bioenergy, and


Appendix B Left border flanking region of pSKI074 plasmid
### Appendix C Oligonucleotide sequences for left border confirmation PCR

<table>
<thead>
<tr>
<th>Oligo name</th>
<th>Sequence (5’ to 3’)</th>
<th>Tm (°C)</th>
<th>Exp. Size (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Capture oligo</td>
<td>CGGACATCTACATTTTTGAAATTGAAAAAAATTGGTAATTACTC TTTTTTCTCCATATTGACCATACTCATT</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>T-DNA insert</td>
<td>CGCCTATAAATACGACGGATG</td>
<td>66</td>
<td>817</td>
</tr>
<tr>
<td>Line 638</td>
<td>CACACGGATTGAAGGGAGAGG</td>
<td>66</td>
<td>817</td>
</tr>
<tr>
<td>Line 316</td>
<td>CAGTCTAGCATATGCTTAGGTAGCC</td>
<td>66</td>
<td>541</td>
</tr>
<tr>
<td>Line 903</td>
<td>GTGCAATATTGAAGGGACGAGG</td>
<td>66</td>
<td>1034</td>
</tr>
<tr>
<td>Line 148</td>
<td>GCATGCTGGTAAGTTATGAAAAGGATC</td>
<td>67</td>
<td>1466</td>
</tr>
</tbody>
</table>
Appendix D command line pipeline for initial assembly attempt

Bowtie 2 pipeline for trial run (command line only)

$create reference sequence index

samtools faidx GCF_000002775.4_Pop_tri_v3_genomic.fna

$create BWA reference index

bwa index GCF_000002775.4_Pop_tri_v3_genomic.fna

$align reads to reference

$bwa mem (reference.fna) (forward reads.fastq) (reverse reads.fastq) > (output) 2> (log for problems)

bwa mem GCF_000002775.4_Pop_tri_v3_genomic.fna B2.fastq B2r.fastq > B2.bwa.sam 2> bwa.log

$view the aligned reads

view B2.bwa.sam less -S B2.bwa.sam

$convert SAM to BAM – problem spot, files output with size of 0

samtools view -uS -o B2.bwa.bam B2.bwa.sam

$sort alignment with reference coordinate order
samtools sort B2.bwa.bam B2.bwa_sorted.bam

$index alignment in samtools

samtools index B2.bwa_sorted.bam

$mark/remove duplicates
$remember to check the log for issues

samtools rmdup B2.bwa_sorted.bam B2.bwa_rmdup.bam 2> samtools_rmdup.log

$second round of indexing on the BAM output

samtools index B2.bwa_rmdup.bam

bedtools bamtofastq -i B2.bwa_rmdup.bam \
    -fq aln.end1.fq \
    -fq2 aln.end2.fq

$convert FASTQ format to FASTA format

paste - - - - < B2-aln.end1.fq | cut -f 1,2 | sed 's/^@/>/' | tr "\t" "\n" > B2.fa
Appendix E DNA Barcodes, read counts, and technical specifications for first sequencing run

<table>
<thead>
<tr>
<th>Well #</th>
<th>Reads</th>
<th>Pairs</th>
<th>Pairs w/ probe</th>
<th>% w/ probe</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>410</td>
<td>205</td>
<td>164</td>
<td>80%</td>
</tr>
<tr>
<td>A2</td>
<td>16</td>
<td>8</td>
<td>8</td>
<td>100%</td>
</tr>
<tr>
<td>A3</td>
<td>3046</td>
<td>1523</td>
<td>228</td>
<td>15%</td>
</tr>
<tr>
<td>A4</td>
<td>15460</td>
<td>7730</td>
<td>5823</td>
<td>75%</td>
</tr>
<tr>
<td>A5</td>
<td>226</td>
<td>113</td>
<td>73</td>
<td>65%</td>
</tr>
<tr>
<td>A6</td>
<td>2620</td>
<td>1310</td>
<td>523</td>
<td>40%</td>
</tr>
<tr>
<td>B1</td>
<td>14094</td>
<td>7047</td>
<td>4897</td>
<td>69%</td>
</tr>
<tr>
<td>B2</td>
<td>502</td>
<td>251</td>
<td>89</td>
<td>35%</td>
</tr>
<tr>
<td>B3</td>
<td>1064</td>
<td>532</td>
<td>459</td>
<td>86%</td>
</tr>
<tr>
<td>B4</td>
<td>1538</td>
<td>769</td>
<td>417</td>
<td>54%</td>
</tr>
<tr>
<td>B5</td>
<td>2908</td>
<td>1454</td>
<td>568</td>
<td>39%</td>
</tr>
<tr>
<td>B6</td>
<td>46112</td>
<td>23056</td>
<td>22416</td>
<td>97%</td>
</tr>
<tr>
<td>C1</td>
<td>55164</td>
<td>27582</td>
<td>26614</td>
<td>96%</td>
</tr>
<tr>
<td>C2</td>
<td>2132</td>
<td>1066</td>
<td>74</td>
<td>7%</td>
</tr>
<tr>
<td>C3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>100%</td>
</tr>
<tr>
<td>C4</td>
<td>21916</td>
<td>10958</td>
<td>10335</td>
<td>94%</td>
</tr>
<tr>
<td>C5</td>
<td>32314</td>
<td>16157</td>
<td>714</td>
<td>4%</td>
</tr>
<tr>
<td>C6</td>
<td>42100</td>
<td>21050</td>
<td>18011</td>
<td>86%</td>
</tr>
<tr>
<td>D1</td>
<td>38028</td>
<td>19014</td>
<td>13458</td>
<td>71%</td>
</tr>
<tr>
<td>D2</td>
<td>3000</td>
<td>1500</td>
<td>1198</td>
<td>80%</td>
</tr>
<tr>
<td>D3</td>
<td>2900</td>
<td>1450</td>
<td>558</td>
<td>38%</td>
</tr>
<tr>
<td>D4</td>
<td>41432</td>
<td>20716</td>
<td>14914</td>
<td>72%</td>
</tr>
<tr>
<td>D5</td>
<td>410482</td>
<td>205241</td>
<td>143012</td>
<td>70%</td>
</tr>
<tr>
<td>D6</td>
<td>61674</td>
<td>30837</td>
<td>906</td>
<td>3%</td>
</tr>
<tr>
<td>E1</td>
<td>23940</td>
<td>11970</td>
<td>150</td>
<td>1%</td>
</tr>
<tr>
<td>E2</td>
<td>1176</td>
<td>588</td>
<td>518</td>
<td>88%</td>
</tr>
<tr>
<td>E3</td>
<td>11610</td>
<td>5805</td>
<td>4190</td>
<td>72%</td>
</tr>
<tr>
<td>E4</td>
<td>408</td>
<td>204</td>
<td>148</td>
<td>73%</td>
</tr>
<tr>
<td>E5</td>
<td>1436</td>
<td>718</td>
<td>523</td>
<td>73%</td>
</tr>
<tr>
<td>E6</td>
<td>3662</td>
<td>1831</td>
<td>709</td>
<td>39%</td>
</tr>
<tr>
<td>F1</td>
<td>4448</td>
<td>2224</td>
<td>1919</td>
<td>86%</td>
</tr>
<tr>
<td>F2</td>
<td>2158</td>
<td>1079</td>
<td>1012</td>
<td>94%</td>
</tr>
<tr>
<td>F3</td>
<td>3348</td>
<td>1674</td>
<td>805</td>
<td>48%</td>
</tr>
<tr>
<td>F4</td>
<td>45248</td>
<td>22624</td>
<td>21181</td>
<td>94%</td>
</tr>
<tr>
<td>F5</td>
<td>11402</td>
<td>5701</td>
<td>1903</td>
<td>33%</td>
</tr>
<tr>
<td>F6</td>
<td>7964</td>
<td>3982</td>
<td>2615</td>
<td>66%</td>
</tr>
<tr>
<td>G1</td>
<td>366</td>
<td>183</td>
<td>116</td>
<td>63%</td>
</tr>
<tr>
<td>G2</td>
<td>620</td>
<td>310</td>
<td>253</td>
<td>82%</td>
</tr>
<tr>
<td>G3</td>
<td>2270</td>
<td>1135</td>
<td>683</td>
<td>60%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>G4</td>
<td>33970</td>
<td>16985</td>
<td>1049</td>
<td>6%</td>
</tr>
<tr>
<td>G5</td>
<td>5574</td>
<td>2787</td>
<td>1540</td>
<td>55%</td>
</tr>
<tr>
<td>G6</td>
<td>35590</td>
<td>17795</td>
<td>593</td>
<td>3%</td>
</tr>
<tr>
<td>H1</td>
<td>410</td>
<td>205</td>
<td>129</td>
<td>63%</td>
</tr>
<tr>
<td>H2</td>
<td>9916</td>
<td>4958</td>
<td>4561</td>
<td>92%</td>
</tr>
<tr>
<td>H3</td>
<td>1064</td>
<td>532</td>
<td>457</td>
<td>86%</td>
</tr>
<tr>
<td>H4</td>
<td>1966</td>
<td>983</td>
<td>677</td>
<td>69%</td>
</tr>
<tr>
<td>H5</td>
<td>2876</td>
<td>1438</td>
<td>1314</td>
<td>91%</td>
</tr>
<tr>
<td>H6</td>
<td>2582</td>
<td>1291</td>
<td>885</td>
<td>69%</td>
</tr>
</tbody>
</table>

Number of mismatches allowed  0

Barcoded pairs  506572

FastQ records with no barcode match  215600  107800

Total pairs  614372
### Appendix F DNA Barcodes for second sequencing run

<table>
<thead>
<tr>
<th>Barcode</th>
<th>Well #</th>
<th>Barcode</th>
<th>Well #</th>
<th>Barcode</th>
<th>Well #</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTCC</td>
<td>A1</td>
<td>CGTGTGGT</td>
<td>C12</td>
<td>AACCGAGA</td>
<td>F11</td>
</tr>
<tr>
<td>AGGC</td>
<td>A2</td>
<td>CAGA</td>
<td>D1</td>
<td>GTGAGGGA</td>
<td>F12</td>
</tr>
<tr>
<td>TTCT</td>
<td>A3</td>
<td>TGCAGA</td>
<td>D2</td>
<td>CGAT</td>
<td>G1</td>
</tr>
<tr>
<td>ATTGA</td>
<td>A4</td>
<td>CTGTA</td>
<td>D3</td>
<td>CTAGC</td>
<td>G2</td>
</tr>
<tr>
<td>TCGTT</td>
<td>A5</td>
<td>GAGGA</td>
<td>D4</td>
<td>GGTGT</td>
<td>G3</td>
</tr>
<tr>
<td>GAGATA</td>
<td>A6</td>
<td>TTCAGA</td>
<td>D5</td>
<td>TAATA</td>
<td>G4</td>
</tr>
<tr>
<td>CTATTA</td>
<td>A7</td>
<td>ACCTAA</td>
<td>D6</td>
<td>CCACAA</td>
<td>G5</td>
</tr>
<tr>
<td>CTTGCTT</td>
<td>A8</td>
<td>GTACTT</td>
<td>D7</td>
<td>CATCGT</td>
<td>G6</td>
</tr>
<tr>
<td>AATATGC</td>
<td>A9</td>
<td>GAATTCA</td>
<td>D8</td>
<td>TGGCTA</td>
<td>G7</td>
</tr>
<tr>
<td>GCGGAAT</td>
<td>A10</td>
<td>ATTGGAT</td>
<td>D9</td>
<td>GTCGATT</td>
<td>G8</td>
</tr>
<tr>
<td>TGCAAGGA</td>
<td>A11</td>
<td>TCTGTGA</td>
<td>D10</td>
<td>CGGTAGA</td>
<td>G9</td>
</tr>
<tr>
<td>CACATGAGT</td>
<td>A12</td>
<td>CCGGATAT</td>
<td>D11</td>
<td>TAGCATGC</td>
<td>G11</td>
</tr>
<tr>
<td>TGCA</td>
<td>B1</td>
<td>GCTGTGGA</td>
<td>D12</td>
<td>ACAGGGA</td>
<td>G11</td>
</tr>
<tr>
<td>GATC</td>
<td>B2</td>
<td>AACT</td>
<td>E1</td>
<td>TATCGGA</td>
<td>G12</td>
</tr>
<tr>
<td>AGCCCC</td>
<td>B3</td>
<td>CGCTT</td>
<td>E2</td>
<td>GTAA</td>
<td>H1</td>
</tr>
<tr>
<td>CATCT</td>
<td>B4</td>
<td>ACCGT</td>
<td>E3</td>
<td>ACAAA</td>
<td>H2</td>
</tr>
<tr>
<td>GGTGTT</td>
<td>B5</td>
<td>GAAC</td>
<td>E4</td>
<td>AGGAT</td>
<td>H3</td>
</tr>
<tr>
<td>ATGCCT</td>
<td>B6</td>
<td>TAGGAA</td>
<td>E5</td>
<td>TACAT</td>
<td>H4</td>
</tr>
<tr>
<td>GCCAGT</td>
<td>B7</td>
<td>ATATGT</td>
<td>E6</td>
<td>CTTCA</td>
<td>H5</td>
</tr>
<tr>
<td>ATGAAAC</td>
<td>B8</td>
<td>GTTGAA</td>
<td>E7</td>
<td>CGCGGT</td>
<td>H6</td>
</tr>
<tr>
<td>ACGTGT</td>
<td>B8</td>
<td>GAACCTC</td>
<td>E8</td>
<td>AACGCC</td>
<td>H8</td>
</tr>
<tr>
<td>TAGCGGA</td>
<td>B10</td>
<td>CATAACT</td>
<td>E9</td>
<td>ACTACCGA</td>
<td>H9</td>
</tr>
<tr>
<td>TGGTACGT</td>
<td>B11</td>
<td>TGCTGGA</td>
<td>E10</td>
<td>TATGGCCAT</td>
<td>H10</td>
</tr>
<tr>
<td>CGCCGAGA</td>
<td>B12</td>
<td>CGGCTTAT</td>
<td>E11</td>
<td>ACCTGGTA</td>
<td>H11</td>
</tr>
<tr>
<td>ACTA</td>
<td>C1</td>
<td>GGATGGT</td>
<td>E12</td>
<td>TTCCTGGA</td>
<td>H12</td>
</tr>
<tr>
<td>TCAC</td>
<td>C2</td>
<td>GCCT</td>
<td>F1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GTATT</td>
<td>C3</td>
<td>TCACC</td>
<td>F2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CCTAC</td>
<td>C4</td>
<td>GCTTA</td>
<td>F3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CCACGT</td>
<td>C5</td>
<td>GTCAA</td>
<td>F4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGTGGGA</td>
<td>C6</td>
<td>GCTCTA</td>
<td>F5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GGAAGA</td>
<td>C7</td>
<td>ATCGTA</td>
<td>F6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AAAAATT</td>
<td>C8</td>
<td>TAAGCAGA</td>
<td>F7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATTAATT</td>
<td>C9</td>
<td>GGATCTA</td>
<td>F8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TCGAAGA</td>
<td>C10</td>
<td>CGCTGAT</td>
<td>F9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TCTCAGTC</td>
<td>C11</td>
<td>ACGACTAC</td>
<td>F10</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Appendix G Geneious settings for *de novo* assembly and mapping to the *P. alba* genome

*De novo* assembly parameters for assembling reads into contigs.

Reads for each well were assembled into contigs using Geneious’ own assembly algorithm, using medium sensitivity. Reads were not end-trimmed and lists of both used and unused reads, as well as assembly reports, were saved. Any contigs generated were also saved separately for use downstream.
Genome mapping parameters to map contigs to *P. alba* genome.

The genome mapping was conducted using the Geneious algorithm at high sensitivity due to multiple genome duplication events in the poplar genome. Contigs generated by *de novo* assembly were not re-assembled, and reports were saved for contigs used and unused, as well as an assembly report detailing locations and technical data.
**Appendix H  *Arabidopsis* mutant population resources**

<table>
<thead>
<tr>
<th>Resource</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSHL</td>
<td><a href="http://genetrap.cshl.org/">http://genetrap.cshl.org/</a></td>
</tr>
<tr>
<td>GABI-Kat</td>
<td><a href="http://www.gabi-kat.de/">http://www.gabi-kat.de/</a></td>
</tr>
<tr>
<td>IMA</td>
<td><a href="https://www.arabidopsis.org/abrc/ima.jsp">https://www.arabidopsis.org/abrc/ima.jsp</a></td>
</tr>
<tr>
<td>JIC Activate</td>
<td><a href="http://arabidopsis.info/CollectionInfo?id=29">http://arabidopsis.info/CollectionInfo?id=29</a></td>
</tr>
<tr>
<td>SAIL (formerly GARLIC/TMRI)</td>
<td><a href="https://www.arabidopsis.org/abrc/sail.jsp">https://www.arabidopsis.org/abrc/sail.jsp</a></td>
</tr>
<tr>
<td>SALK</td>
<td><a href="http://signal.salk.edu/tabout.html">http://signal.salk.edu/tabout.html</a></td>
</tr>
<tr>
<td>TAIR</td>
<td><a href="https://www.arabidopsis.org/portals/mutants/index.jsp">https://www.arabidopsis.org/portals/mutants/index.jsp</a></td>
</tr>
<tr>
<td>TILLING (Col)</td>
<td><a href="http://tilling.ucdavis.edu/index.php/Main_Page">http://tilling.ucdavis.edu/index.php/Main_Page</a></td>
</tr>
</tbody>
</table>

*Arabidopsis* mutant population resources.

Numerous mutant populations have been created using a range of mutagenic methods. Many of the lines have been partially characterised and are available for researchers from their respective seed banks, which can be located in the TAIR website or at the resource center itself.