CZECH UNIVERSTY OF LIFE SCIENCES PRAGUE

Faculty of Forestry and Wood Sciences

Diploma thesis

Theoretical Assessment of the Pedigree Reconstruction Methods in Forest Tree Breeding

Prague, 2020

Supervisor

Dr. Jiri Korecky

Author

Kevin Aculey

CZECH UNIVERSITY OF LIFE SCIENCES PRAGUE

Faculty of Forestry and Wood Sciences

DIPLOMA THESIS ASSIGNMENT

B.Sc. Kevin Aculey, BSc

Forestry Engineering
Forest Engineering

Thesis title

Theoretical assessment of the pedigree reconstruction methods in forest tree breeding

---

Objectives of thesis

Pedigree reconstruction methods have been developed to accomodate conventional breeding programs with powerful tool: analysis of wild, unimproved populations with either partical or full pedigree assembly on the basis of neutral genetic markers. In the current study, we plan to investigate existing data set on European larch in Austria, consisting of 20 stands in Wienna forest. Within this set, a retrospective study will be performed to investigate the joint effect of sample size, polymorphic information content, number of marker loci, and knowledge of maternal gametic contributions. The interplay among these factors is crutial to reaching declared accuracy of the estimated pedigree.

Methodology

1. Pedigree reconstruction using existing data (SSR markers)

2. Iterative reduction of the sample size and its effect on the accuracy of the resulting pedigree (part I. of the retrospective study)

3. Iterative reduction of the marker loci and its effect on the accuracy of the resulting pedigree (part II. of the retrospective study)

4. Results will be statistically evaluated and presented in tables and graphs.

The proposed extent of the thesis

40p.

Keywords

DNA markers, pedigree reconstruction, forestry

Recommended information sources

El-Kassaby, Y. A. & Lstibůrek, M. Breeding without breeding. Genet. Res. 91, 111-120 (2009).

Kalinowski, S. T. et al. Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. Mol. Ecol. 16, 1099-1106 (2007).

Marshall, T. C. et al. Statistical confidence for likelihood-based paternity inference in natural populations. Mol. Ecol. 7, 639-655 (1998)

Wagner, S. et al. Two highly informative dinucleotide SSR multiplexes for the conifer Larix decidua (European larch). Mol. Ecol. Resour. 12, 717-725 (2012).

White, T. L. et al. Forest Genetics (CABI, 2007).

Expected date of thesis defence

2017/18 SS – FFWS

The Diploma Thesis Supervisor

Ing. Jiří Korecký, Ph.D.

Supervising department

Department of Genetics and Physiology of Forest Trees

Electronic approval: 31. 1. 2018

prof. Ing. Milan Lstibůrek, MSc, Ph.D.

Head of department

Electronic approval: 5. 2. 2018

prof. Ing. Marek Turčáni, PhD.

Dean

Prague on 21. 05. 2020

Declaration

I declare that I wrote my graduation dissertation (master’s/graduation) independently, and that I have stated all the information sources and literature I used. Neither this thesis nor any substantial part of it have been submitted for the acquisition of another or the same academic degree. I consent to the lending of my dissertation for study purposes. By affixing his or her signature the user confirms using this dissertation for study purposes and declares that he or she has listed it among the sources used.

In Prague, June 15, 2020

Kevin Aculey

………………………...

Acknowledgement

My first appreciation goes to the almighty God for his love and protection over my life. I will also thank my family, especially Elizabeth, my wife for their support and encouragement.

My heart felt gratitude to my supervisor, Dr. Jiri Korecky for his time and guidance in making this work a success. My last but not the least of appreciation goes all the lecturers of Czech University of Life Sciences who through one way or the other, have influenced my life.

Abstract

Pedigree reconstruction using molecular markers can be used as an effective tool to manage and control gametic distribution in open pollinated breeding systems. The incorporation of parentage analysis into forest breeding plans has the tendency to improve pedigree records and increase the accuracy of selection in forest trees.

The effectiveness and the accuracy of a pedigree to a larger extent, will depend on the informativeness of the microsatellite markers. The informativeness of microsatellite markers is measured by parameters such as Polymorphic Information Content (PIC), Expected Heterozygosity, Null allele frequency and its conformation to the Hardy-Weinberg Equilibrium.

In this study, a genetic data set on European larch from Vienna forest was used in pedigree reconstruction. The effect of sample size and the number of loci on the pedigree was also investigated. The informativeness of the microsatellite markers used in the pedigree reconstruction was also evaluated.

An increase in sample size resulted in an increase in parent-pair assignment. Increase in the number of loci also resulted in an increase in simulation and parentage analysis assignments at a strict confidence of 95%. No assignment was observed when less than six loci was used in the simulation and parental analysis. All but one of the thirteen microsatellite markers used in the pedigree reconstruction had their polymorphic Information Content (PIC) values higher than 0.50.

Key words: Pedigree reconstruction, Polymorphic Information Content, European larch.

Klíčová slova: rekonstrukce rodokmene, polymorfní informační obsah, modřín opadavý.

Contents

List of Tables

List of Figures

LIST OF SYMBOLS AND ABBREVIATIONS USED

AFLD: Amplified Fragment Length Polymorphism

BwB: Breeding without Breeding

cpSSRs: Chloroplast simple sequence repeats

DNA: Deoxyribonucleic acid

FAO: Food and Agriculture Organization

F(null): Null allele frequency

HExp: Expected Heterozygosity

HObs: Observed Heterozygosity

HW: Hardy-Weinberg

K: Number of alleles

LOD: Natural logarithm of a combined likelihood

N: Number of individuals

NE-1P: Average non-exclusion probability for one candidate parent

NE-2P: Average non-exclusion probability for one candidate parent given the genotype of a known parent of the opposite sex

NE-PP: Average non-exclusion probability for candidate parent pair

NE-I: Average non-exclusion probability of two unrelated individuals

NE-SI: Average non-exclusion probability of two siblings

PCR: Polymerase Chain Reaction

PIC: Polymorphic Information Content

QTL: Quantitative Trait Loci

RAPD: Random Amplified Polymorphic DNA

SNP: Single Nucleotide Polymorphism

SSR: Simple Sequence Repeat

1. Introduction

The use of molecular techniques to determine parentage has been on the rise ever since hypervariable microsatellite markers were developed. These techniques have been used in various fields to answer questions relating to ecology, evolution and quantitative genetics. The highly polymorphic attributes of microsatellite markers also known as Simple Sequence Repeat (SSR), Short Tandem Repeat (STR) or Simple Sequence Length Polymorphisms (SSLP) has been one of the main reasons why it has been at the center of most genetic analysis. Polymorphism in SSR markers are as a result of differences in the number of repeats of the motif caused by either polymerase strand-slippage in DNA replication or recombination errors(Lucia, Vieira, Santini, Diniz, & Munhoz, 2016).

Microsatellite markers show many alleles at a locus which differ from each other in the number of repeats and are codominant with high rate of mutation which makes them suitable for estimation of within and between-breed genetic diversity and genetic mixtures among breeds. The reproductive success in forest populations is determined by the breadth of genetic diversity of future generations and their resilience to unpredictable environmental eventualities (El-Kassaby, Funda, & Lai, 2010). The most common parameters for assessing within-breed diversity include the number of alleles per population, expected and observed heterozygosity.

Marker assisted selection, genetic linkage analysis, kinship analysis and fingerprinting are some of the notable applications of microsatellite markers. SSR markers have been successfully used to assess the genetic relationship between populations and individuals by estimating their genetic distance (Beja-Pereira et al., 2003). They are usually the markers of choice when it comes to small-scale genetic studies with limited budget, and when there is a chance of detecting large genetic information and physiological parameters of the genome (Abdurakhmonov, 2016). When these genetic markers are combined with the appropriate statistical analysis, it is possible to make recommendations on the conservation on forest genetic resources, infer the origin of forest plants and woods and conduct molecular tree improvement (Garcia & Garcia, 2016). Species specificity of the SSR loci in plants is one of the main demerits of microsatellite markers.

2. Goals and aims

Most of the new pedigree reconstruction methods have been developed such that, they can be easily employed in conventional breeding programs. These pedigree reconstruction methods are equipped with powerful tools for the analysis of wild and unimproved populations with either partial or full pedigree assembly by using neutral genetic markers.

In the current study, I plan to investigate existing data set on European larch in Austria, consisting of 20 stands in Vienna forest. Within this set, a retrospective study was performed to investigate the joint effect of sample size, polymorphic information content, number of marker loci, and knowledge of maternal gametic contributions. The interplay among these factors is crucial to reaching declared accuracy of the estimated pedigree.

3 Literature Review

3.1. Tree Breeding

The utilization of forest products will continue to increase while land base used for wood production is expected to decline but the indication of models are that, the potential productivity in many regions can be much higher than the current situation (Schmidtling et al., 2002). Some 129 million hectares of forest has been lost since 1990 (*Global Forest Resources Assessment 2015 Desk reference*, 2015). Some of the consequences of the rapid decline in forest lands include habitat lost, decrease in biodiversity, soil degradation, and pollution of water bodies.

In response, efforts around the world are being developed to achieving sustainable forest management, an approach that seek to mitigate against the negative effects that arise due to the exploitation of the forest. The maintenance or restoration of ecosystem functions, protection of biological diversity, making money, improving the welfare of rural people, and the preservation of opportunities for research and recreation are the main driving force behind sustainable forest management (Putz & Redford, 2010). Of the various measures being developed to achieving sustainable forest management across the world, tree breeding has been identified as one of the most effective and environmentally friendly means to increase sustainable biomass production (Haappanen, Jansson, Bräuner Nielsen, Steffenrem, & Stener, 2015).

Breeding combines the science and art, seated in the ability of a breeder to identify differences in the traits of economic importance among plants and to improve these traits with available scientific knowledge (Farooq & Azam, 2002). It is an important component of tree improvement which involves the application of genetic principles for the mass production of seedlings with desired traits in order to achieve higher productivity, better adaptability of the environment and vigorous growth rate (Thakur & Schmerbeck, 2014). The success attained in the use of molecular markers in plant breeding necessitated the incorporation of this technology into the breeding of trees. That is, the domestication of trees through artificial selection and mating to meet the demand of a society. There exist a wide variation in traits such as tree growth rate, wood quality, stem straightness, resistance to pest and diseases and general adaptation to a given climatic condition. Tree breeders try to package desired traits in an individual with the aim of producing trees that are genetically and phenotypically superior than those in the wild.

Molecular markers have been exclusively used in breeding programs for the determination of genetic variation within, between and among populations, verification and characterization of genotypes as well as marker assisted selection (Gudeta, 2018). Breeding programs are directed at increasing genetic gain as well as decreasing the breeding cycle. This can be achieved through the integration of molecular breeding techniques with the conventional breeding methods, and the use of marker assisted selection and double haploid development respectively (Xu et al., 2017). Gains achieved in tree breeding are fastest, largest and cheapest when the right species and seed sources within the species are used (Thakur & Schmerbeck, 2014). Breeding programs should however, maintain sufficient genetic variation to allow for continued genetic gains over multiple generations (Johnson, Clair, & Lipow, 2001). Genomic selection which make use of large number of genome-wide markers in the prediction of complex phenotypes increase the potentials of accelerating breeding cycles, increasing selection intensity and improve the accuracy of breeding values (Grattapaglia, 2018).

The primary objective of a tree breeding program is to identify the traits of interest for improvement and estimate the genetic variation in these traits and how they are inherited (Eriksson, Ekberg, & Clapham, 2013). The objective of a breeding program should also depend on the genetic state of the population and its intended usage (Wellmann & Bennewitz, 2019).

Unlike most breeding techniques employed in agricultural plants, breeders of forest trees will have to consider the several unique ecological, population, and quantitative genetic issues in recurrent selection programs, and the deployment of strategies with forest trees. Since tree breeders work largely with wild populations, factors such as geographic patterns of genetic variation, seed transfer within environmentally similar zones of adaptation, special field and progeny test designs must address and meet the special biological features of forest trees (Yanchuk, 2009). Genetic linkage analysis in forest trees has been comparatively slow and factors such as the large genome size of trees and the difficulty involved in developing segregating F_2 population being some of the reasons behind the slow genetic linkage analysis (Tingey & Del Tufo, 1993).

3.2 European Larch

The two most important Larix genus for Western European forestry are the European larch (*Larix decidua*) and the Japanese larch (*Larix kaempferi*) (Pâques, Philippe, & Prat, 2006). European larch, *Larix decidua* an upland conifer is one of the important tree species in the central and eastern mountains of Europe. Large plantations of European larch have been established throughout Europe and northeastern and North America (Gilmore & David, 2002).

Being a deciduous conifer, European larch sheds its needles during the cold winter, an adaptation that helps it to escape foliage desiccation. It has adequate genetic diversity, grows rapidly and hybridizes readily (Einspahr, Wyckoff, & Fiscus, 1984). Their interspecific hybrids have also been identified to be of high value for lowland reforestation ever since it was first observed in Scotland at the beginning of the 20th century, and hybrids were also found to be more vigorous in height and diameter at the end of the first and the second growing season (Pâques et al., 2006).

3.2.1 Distribution

European larch can survive on different growing sites, differing with respect to both climate and soil, and it is adapted to 2500 m as well as 150 m elevations (Lewandowski & Mejnartowicz, 1991). European larch does not have a continuous range over the European continent but have four distinct, closed natural distribution regions, namely: Alps, Sudeten, Tatra and central part of Poland plus several outliers in the east and south Carpathian Mountains in Romania (Lewandowski & Mejnartowicz, 1991). Ten species of larch with numerous varieties and hybrids have been identified and all of them are found in the Northern Hemisphere with cold winters and even at the northern and altitudinal limits for tree growth, larches are widespread and often dominate a woodland zone north of evergreen dominated boreal forests or above subalpine forests (Gower & Richards, 1990).

3.2.2 Wood of European Larch

Larix species combine good form and rapid juvenile growth with moderately high wood density and good fiber characteristics (Keith & Chauret, 1988). European larch reaches the culmination of its increment in diameter early and maintains it for a long time, forming stands with a large abundance of valuable high-class timber (Nawrot, Pazdrowski, & Szymański, 2008). Wood from European larch is much appreciated for its good mechanical properties and also for the high durability of its heart wood (Pâques, 2001).

Stem production values for natural and plantation larch forests growing in relatively favorable environment range were found to be higher in plantation forest and this low production rate of stem wood in natural larch forest were partly attributed to the adverse effects of cold temperature, low nutrient and water availability on leaf area which positively correlates with forest production (Gower & Richards, 1990).

3.2.3. Pest and Diseases of European Larch

The pests of European larch include: *Ips typographus* and other species of Ips genus such as cembrae, the back beetle which is associated with fungal pathogens, the larch case-bearer (*Coleophora laricella*), larch bud moth (*Zeiraphora diniana*), and the large pine weevil (*Hylobius abietis*), a serious pest affecting young coniferous forests in Europe (Da Ronch, Caudullo, Tinner, & de Rigo, 2016). The caterpillars of the case-bearer moth, *Cleophora sibiricella* are defoliators whiles tortrix moth, *Cydia illutana* use the scales of the cones of European larch as a source of nourishment.

European larch is prone to several fungus diseases such as larch canker, root rot and but rot. The larch canker disease is caused by the fungus, *Larchnellula willkommii* whilst root rot and but rot diseases is caused by the fungi, *Heterobasidion annosum* and *Phaeolusis schweinitzii* respectively (Da Ronch et al., 2016). The cast fungus, *Meria laricis* feeds on the needles of European larch, causing significant defoliation. The larch canker disease is the most disastrous of all the diseases affecting larch species and resistance to this disease was found to be highest in European larch populations from the eastern Alps than those from the southern Alps of Europe (Matras & Paques, 2008).

3.3 Genetic Markers

Traits that are inherited successfully and assigned unambiguously to the phenotype or to a set of one or more loci can be referred to as a genetic marker (Farooq & Azam, 2002). A genetic marker may be a short DNA sequence such as single nucleotide polymorphism (SNP) or a long DNA sequence such as minisatellites. Heritable genetic markers that are associated with economically important traits can be utilized by plant and tree breeders as a tool to increase the efficiency of selection by reducing the time period as well as the population size used during selection (Staub & Serquen, 1996). Genetic markers have come to play a significant role in the study of genetics of organisms, including trees at the level of single genes (White, Adams, & Neale, 2007). These markers are used to identify different features in DNA sequence and these differences can be used to determine if a specific region was inherited from the mother or the father. Their development and application have been swiftly incorporated into the field of tree breeding with remarkable results. Genetic markers are classified broadly into two categories, classical markers and DNA or molecular markers. Morphological, cytological and biochemical markers are classified under classical markers while markers such as restriction fragment length polymorphism, amplified fragment length polymorphism, simple sequence repeats, single nucleotide polymorphism and diversity arrays technology markers are classified as DNA markers. Each marker type has its own strength as well as limitations with respect to their development and application.

3.3.1 Morphological Markers

Morphological markers are related to the variation in observable and measurable features such as size, shape, colour and surfaces of plant parts. Any morphological trait which is controlled by a single locus can be considered as a genetic marker if only it can be replicated over a range of environments (Staub & Serquen, 1996). Morphological characters in forest trees suitable as genetic markers are few and most of them are observed mutation in seedlings such as albino needles and dwarfing and they have been successfully utilized in estimating self-pollination rates in conifers. However, morphological mutants are rare and most often have a deleterious effect, limiting their utilization (White et al., 2007). Morphological markers exhibit dominance and are influenced by environmental conditions. Some of these morphological markers can only be identified at specific growth stage of organisms and as such, their detection is dependent on the developmental stage of the organism. These factors have also limited the development and application of morphological markers.

3.3.2 Biochemical Markers

Biochemical markers were developed base on variation in protein and amino acid binding pattern. Biochemical markers such as terpenes and allozymes were developed in the 70s for trees. The best available genetic markers in the 60s and 70s for forest trees were monoterpene genetic markers with their utilization extended to taxonomic and evolutionary studies. Monoterpene marker loci are few and as a result, they express some form of dominance in their phenotype and require specialized and expensive equipment for assay. These factors contributed to the limited utilization of monoterpene markers after the development of allozyme genetic marker (White et al., 2007).

Allozymes are allelic variants of enzymes encoded by structural genes and isozyme, the general term for allozymes is defined as structurally different molecular forms of enzymes with qualitatively the same catalytic function (Kumar, Gupta, Misra, Modi, & Pandey, 2009). Isozymes have been extensively used in forestry to study genetic variation between and within populations, population structure, phylogeny, and to elucidate mating pattern in natural populations as well as experimental populations (Namkoong & Koshy, 2001). They generally provide ample genetic information and are relatively inexpensive, rapid and technically easy to apply (Neale et al. 1992). Isozyme markers can be genetically mapped onto chromosomes and then used as genetic markers to map other genes (Jiang, 2013). Determination of Mendelian inheritance of Isozyme is required before it can be used for genetic studies and this is achieved through similar crosses between trees as Mendelian crosses in peas. Isozymes, however, have a very limited number of possible markers and they are not distributed evenly on the chromosome (Yang, Kang, Nahm, & Kang, 2015). Isozymes exhibit a low level of polymorphism and may also be affected by environmental conditions (Kumar et al. 2009).

3.3.3 Cytological Markers

Cytological markers are related to the variations in the number, bidding patterns, shape, size and position of chromosome. These variations reveal differences in the distributions of euchromatin and heterochromatin. These chromosome landmarks can be used in the differentiation of normal and mutated chromosomes. They can also be used in the identification of linkage groups and in physical mapping (Nadeem et al. 2017). Physical maps based on morphological and cytological markers lay the foundation for genetic linkage mapping with the aid of molecular techniques but its direct use in plant breeding and genetic mapping has become limited in recent times (Jiang, 2013).

3.4 Molecular Markers/DNA Markers

Molecular markers are nucleotide sequences and can be investigated through the polymorphism present between the nucleotide sequences in different individuals. Insertion, deletion, point mutation duplication and translocation are the basis of this polymorphism. Different types of DNA molecular markers have been developed and successfully applied in genetics and breeding activities in various Agriculture crops (Nadeem et al., 2017). The ability of restriction enzymes to cut DNA molecules at occurrences of a recognition sequence throughout the DNA strand forms the basis of the technique used in developing molecular markers. Several hundred of these enzymes, each with its own specificity in terms of recognition sequence, are known (Kloch, Hawliczek-strulak, & Sekrecka-bielak, 2015).

Molecular markers for plant breeding are based on genetic differences among individuals in alleles at a certain locus (Yang et al., 2015). Molecular markers are categorized into hybridization-based markers and polymerase chain reaction (PCR) based markers.

3.4.1 Hybridization Based Markers

Genetic markers based on DNA- hybridization were developed in the 1970s and the first to be developed was the Restriction Fragment Length Polymorphism (RFLP) marker. The hybridization method involves the tagging of a probe, a short DNA fragment that is homologous to the target DNA with a radioisotope which then hybridizes with the DNA being analyzed. The revelation of polymorphism in the target DNA is based on the DNA-probe hybridization or the size of the hybridized DNA fragment. Insertions and deletions of small segments of DNA or the gain or loss of a restriction site are two types of restriction fragment length polymorphisms (RFLPs) which are easily detected (Neale et al., 1992). The genetic interpretation of Restriction Fragment length Polymorphism (RFLP) banding pattern can be difficult especially in conifers whose large genomes often lead to large numbers of fragments revealed by a single probe (White et al., 2007). The visualization of DNA fragments is made possible with southern blotting and probe hybridization technique.

3.4.2 Polymerase Chain Reaction Based (PCR) Markers

The development of polymerase chain reaction (PCR) technique paved the way for a new and sophisticated means of developing genetic markers. The polymerase chain reaction technique involves the in vitro enzymatic amplification of a specific DNA sequence by using suitable primers and DNA polymerase. Primers used in the PCR technique are of two kinds, namely: specific primer and random primer. The specific primers are used for the amplification of DNA fragments of at least, partially known sequences whiles the random primers are used for DNA probes of unknown sequence (Kloch et al., 2015). Polymerase chain reaction-based markers are categorized into locus non-specific markers and locus specific markers.

Locus specific polymerase chain reaction markers include, Random Amplified Polymorphic DNA (RAPD) and Amplified Fragment Length Polymorphism (AFLP) with the Random Amplified Polymorphic DNA marker as the most used. The sources of polymorphism for this type of markers are point mutation, deletions, insertions and chromosomal rearrangement that change the DNA-primer complementation pattern. Random amplified polymorphic DNA markers are dominant genetic markers but in conifers, the problem of dominance can be overcome if RAPD markers are assayed from the megagametophyte tissue (Neale et al., 1992). Species specific RAPD markers have been sought in *Quercus robur* and *Quercus petraea* and in order to detect natural hybrids between the species, many mapping projects have been conducted using RAPDs (Kloch et al., 2015).

Amplified Fragment Length Polymorphism (AFLP) combines the techniques of both the restriction fragment length polymorphism (RFLP) and the random amplified polymorphic DNA (RAPD). It in cooperates the positive aspects of polymerase chain reaction and the restriction digestion techniques. The development of these marker types is based on the generation of DNA fragments using restriction enzymes, oligonucleotide linkers and its amplification by polymerase chain reaction. AFLP markers were used in trees for the first time to genetically map a disease resistance gene in *Populus* (Cervera et al., 1996). Genetic linkage maps based on AFLPs have also be constructed in *Eucalyptus globulus* and *Eucalyptus tereticornis* and in *Pinus tidea* (White et al., 2007). AFLP markers are very sensitive and reproducible and it does not require prior knowledge of the sequence information. They are, however, tedious to analyze and requires skilled technicians. However, the development of commercial kits for AFLP analysis has alleviated these problems associated with AFLP markers.

Single nucleotide polymorphism (SNP) and simple sequence repeat (SSR) are examples of markers categorized under locus specific markers. Simple sequence repeat (SSR) markers also called microsatellites were first developed for used in genetic mapping in humans (Neale et al., 1992). Jarne & Lagoda, (1996) defined microsatellite as short tandemly repeated sequences whose unit of repetition is between 1 to 5 base pairs and may be classified into pure, compound and interrupted. They are co-dominant markers and are more informative for genotyping individuals and for linkage mapping than dominant markers such as RAPDS (Kloch et al., 2015). Microsatellites can be mononucleotide, dinucleotide, trinucleotide, tetranucleotide, pentanucleotide and hexanucleotide (Nadeem et al., 2017). According to Jarne & Lagoda, (1996) di, tri and tetra nucleotide repeats are the mostly used types. Microsatellites are distributed in the genome but are also found in the mitochondria and the chloroplast. Microsatellites represent the lesser repetition per locus with higher polymorphism level, with this high level of polymorphism attributed to the occurrence of various numbers of repeats in microsatellite regions which can be detected easily by the use of polymerase chain reaction (PCR) (Nadeem et al., 2017). Polymorphism in microsatellite can be detected by southern hybridization or polymerase chain reaction (PCR). The number of microsatellite repeats in plants varies from among individual and species. Flaking sequences of microsatellite repeats are used as PCR primers in the development of microsatellite markers. Formerly, probes containing repeat sequences were used to identify homolog repeats in genes from DNA libraries when developing microsatellite markers. Currently, they can be developed from gene bank data. Microsatellite markers developed from genomic libraries may belong to the transcribed region or the non-transcribed region of the genome, and rarely is there information available regarding their function (Kumar et al., 2009). Since the development of the first microsatellite markers in the forest tree, *Pinus radiate*, it has become an important tool in individual genotyping and studies of gene flow in forest trees. Microsatellite markers from the nuclear genome have been developed for temperate and tropical trees such as Eucalyptus, Quercus, Picea and *Melaleuca altenifolia* (Kloch et al., 2015). .

Chloroplast microsatellites (cpSSRs) have become a popular tool for the study of population genetics as a result of their higher polymorphism with easy genotyping. They are inherited uniparentally and the chloroplast chromosome is a non-recombinant molecule due to which all chloroplast microsatellites loci are linked (Nadeem et al., 2017). In most gymnosperms, the chloroplast genome is inherited paternally, and this creates the opportunity for paternity testing in forest trees.

Single nucleotide polymorphisms (SNPs) are termed the new generation of molecular markers since they were recently discovered. Single nucleotide polymorphisms (SNPs) are single base changes in the DNA sequences, the most abundant type of mutation. NSPs may be transitions or transversions based on the nucleotide substitution (Nadeem et al., 2017). There are numerous sites within a genome where a short stretch of DNA in a pair of homologous chromosomes differ by a single nucleotide. A single nucleotide base is the smallest unit of inheritance and SPN can provide the simplest and maximum number of markers (Nadeem et al., 2017). SNPs must be present in at least 1% of individuals in a population to be considered as polymorphic (Khlestkina & Salina, 2006). The abundance the ease to which they can be measured, makes its genetic variations significant. An SNP close to a gene acts as a marker for that gene. SNPs can be in both coding and the non-coding regions and those found within the coding regions have the potentials of altering the protein structure made by that coding region. The preference of SNP markers over SSR markers is attributed to their higher map resolution, higher through-put, lower cost and lower error rate

(Jones et al., 2009). Single nucleotide polymorphisms (SNPs) have become extremely popular in plant molecular genetics due to their genome-wide abundance and amenability for high- to ultra-high throughput detection platforms. Unlike earlier marker systems, single nucleotide polymorphisms (SNPs) has made it possible to create saturated, if not supersaturated genetic maps, thereby enabling genome-wide tracking, fine mapping of target regions, rapid association of markers with a trait and accelerated cloning of gene/QTL of interest. According to Butler, Coble, & Vallone (2007) SNPs are more stable in terms of inheritance and could aid parentage testing in some cases for kinship analysis. However, the polymorphic information content is lower than microsatellite markers since it is bi-allelic.

3.5 Pedigree Reconstruction

Pedigree reconstruction methods have been developed to accommodate conventional breeding programs with powerful tool: analysis of wild, unimproved populations with either partial or full pedigree assembly based on neutral genetic markers. It provides a solid foundation for studies in population evolutionary dynamics in the wild. Pedigree reconstruction does not only include parenting assignment, but when sampling of candidate parents is incomplete, also clustering of (half)-siblings sharing the same, non-genotyped parents (Huisman, 2017). There are several means by which relatedness can be measured, but the best depends on the number of alleles that a pair of individuals share by descent at random locus (Day-williams, Blangero, Dyer, Lange, & Ã, 2011).

The minimum number of loci required to accurately assign parentage depends on several factors that affects the informativeness, including allelic richness and diversity, linkage disequilibrium among marker loci due to physical linkage and other sources, number of parental pairs, mating design, frequency of null alleles and genotype errors, and unequal numbers of offspring per family (Matson, Camara, Eichert, & Banks, 2008).

3.5.1 Methods of Pedigree Reconstruction and Parental Analysis

The development and the introduction of microsatellite markers paved the way for the current status of parentage analysis (Jones, Small, Paczolt, & Ratterman, 2010). Parentage analysis is simply a means of testing assignment and this technique has been employed as a tool for the detection of ecological and evolutionary patterns in systems that are characterized with high levels of flow of genes (Christie, 2010). The ability to infer genealogical relationship among individuals has become an effective approach to investigate a wide variety of evolutionary, ecological, and behavioral questions (Harrison, Saenz-Agudelo, Planes, Jones, & Berumen, 2013). Many statistical approaches have been developed to accommodate these techniques with many biologists being oblivious of some of these new approaches and how pedigree information data could easily be extracted by using these techniques (Blouin, 2003). Exclusion, category allocation, fractional allocation, full probability analysis, parent reconstruction and sub-ship reconstruction are some of the approaches being applied in parental analysis. Most of these new approaches aim at improving upon earlier methods that were coupled with some limitations as regards to cost, accuracy and the ease to work with.

3.5.2 Exclusion Approach

The more genetically similar two individuals are, the more likely they are to share alleles for the genes involved in kin recognition (Städele & Vigilant, 2016). The development of the exclusion method of parentage analysis follows the Mendelian inheritance rules. According to Mendelian inheritance in diploid organisms, a parent and its offspring should at least share an allele per locus for a co-dominant marker. The exclusion approach for pedigree reconstruction use parent-offspring incompatibilities as the basis for rejecting a particular parent for being the true parent of an offspring. A candidate parent can therefore be rejected as the true parent of the offspring of interest if it does not meet this Mendelian inheritance condition of sharing an allele per locus for a co-dominant marker. The exclusion approach is an appealing method because, exclusion of all but one parent pair from a complete sample of all possible parents for each offspring in a population could be considered as a paragon of parentage analysis but only few studies have achieved this idea (Jones & Ardren, 2003). Under right conditions, the exclusion approach could be utilized as a powerful technique to detect parent-offspring pairing in a large open population (Harrison et al. 2012).

The exclusion approach is simple, but it has been identified with some drawbacks which can affect results in parental analysis. Any marker characteristics that prevents inheritance from appearing strictly Mendelian to the observer could result in false exclusion of a true parent and microsatellite markers have been found to be vulnerable to this type of phenomenon (Jones, Small, Paczolt, & Ratterman, 2010).

3.5.3 Categorical Allocation Approach

The inability to achieve a complete exclusion in some instances demanded the development of new methods that can accommodate such scenarios and this led to the development of the categorical and fractional allocation or likelihood methods which assigns an offspring to non-excluded parents based on likelihood scores derived from their genotype. The categorical allocation or likelihood approach assign the entire offspring to a parent with the highest likelihood or posterior probability of being the true parent. The categorical allocation approach easily accommodates scoring errors, missing data or null alleles that are frequently associated with microsatellite data sets (Harrison et al., 2013).

The updating and refining of the categorical allocation method has really transformed it into a very useful parentage analysis technique (Jones et al., 2010).

3.5.4 Fractional Allocation Approach

The fractional allocation approach is similar to the categorical approach but instead of assigning the entire offspring to the most likely male as in the categorical approach, offspring are assigned partially to each non-excluded candidate parent based on their relative likelihoods or posterior probabilities (Jones & Wang, 2010). At a statistical point of view, the fractional allocation approach provides exact estimates of important mating system parameters but it does not represent the biological truth, because an offspring can have only one mother and one father and there is no such thing as fractional parentage (Jones & Ardren, 2003). The fractional allocation approach may possess better statistical properties for problems that involve the estimation of population-level variables such as the relative fitness of genotypic classes or variances in reproductive success (Jones & Wang, 2010). This approach also allows population level patterns of paternity to be assessed even when the discriminatory power of marker is low (Marshall; Slate; Kruuk; Pemberton, 1998).

3.5.5 Full Probability Analysis

This approach of parental analysis estimates the population-level parameters of interest simultaneously with the parent-offspring relationships in a single modeling framework that interfaces very naturally with the fractional allocation techniques. The full probability approach put individuals into clustered family groups and the likelihood of different clusters is evaluated to identify the parsimonious configuration (Harrison et al., 2013). The accuracy of assignment for this approach is very high as accounting for the various family groups provides valuable information (Harrison et al., 2013). The relationships between some of these variables and the probability of parentage could be known with certainty and this knowledge would affect prior probabilities of parentage for certain individuals in the population. Other variables could have unknown relationships with parentage and the estimation of these relationships would be part of the analysis of the model (Jones et al., 2010). Difficulty in the specification of models is one of the drawbacks of this approach.

3.5.6 Sib-ship Reconstruction

The algorithms of this approach of pedigree reconstruction have been improving and now, can provide reconstructed parental genotypes or use candidate genotypes to guide sibship reconstruction procedure and the technique comes into play when a group of offspring can be collected from the population, but family groups cannot be identified a priori even though the sample is known to contain some full- and half-sibs model (Jones et al., 2010). In species for which the population can be expected to mainly contain groups of full and/ or half siblings, sibship reconstruction is a powerful tool for identifying the related individuals. This approach is more accurate than evaluating dyads because it considers the relationships among all genotypes simultaneously (Sheikh, Berger-Wolf, Khokhar, & Dasgupta, 2008).

Sibship reconstruction techniques can be classified into the likelihood-based method and the combinatorial method. In the likelihood-based method, the algorithm attempts to partition the sampled individuals into sibling groups in way that maximizes the probability of the data whereas the combinatorial method takes advantage of a strong focus on Mendelian segregation, and essentially, all implementations of these methods are sufficiently computationally challenging that , stochastic optimization techniques are required to obtain a solution in timeframe relevant to a typical human lifespan model (Jones et al., 2010). According to Städele & Vigilant (2016) the success of sibship reconstruction generally improves with increase in the number of individuals per full- or half-sib family, although full sibship may be determined with accuracy for siblings' groups as small as four but may decrease with increasing number of families. Sibship reconstruction can be useful in the context of parentage analysis when a large group of offspring can be collected, but they are not associated with any particular parent and not in family groups. If a pool of candidate parents is available, then an assignment technique can be used, with sibship reconstruction serving as a complementary approach and if candidate parents are not available, then sibship reconstruction could allow some inference of patterns of parentage through the comparison of reconstruction genotype model (Jones et al., 2010).

3.5.7. Parentage Analysis Program, CERVUS

CERVUS has been effectively used in the field of genetics for parentage analysis in both plant and animal population. It is the most used likelihood-based paternity inference program (Christie, 2010). The development of the software was based on the assumption that, the species is diploid with autosomal genetic markers that are inherited independently of each other. CERVUS make use of genetic data from co-dominant markers such as microsatellites and single nucleotide polymorphism (SNP) to perform analysis such as allele frequency analysis, parentage analysis simulation, parentage analysis and identity analysis. Even though CERVUS accommodates incomplete parental sampling, the software is most powerful when all parents are sampled. It calculates a likelihood score for each possible parent-offspring pairing and uses this value to assign parentage across a group of offspring (Jones et al., 2010). The advantage of this software over the exclusionary-based method is that, multiple non-excluded males can be statistically distinguished, it makes room for laboratory typing error and confidence is statistically determined for assigned paternities through simulation (Slate, Marshall, & Pemberton, 2000).

One of the major problems that researchers performing genotype analysis had to deal with was genotyping errors. However, by using CERVUS 1.0 and 2.0 versions, researchers take the option for allowing for genotyping errors (Kalinowski, Taper, & Marshall, 2007). If genotyping errors are not accommodated during data analysis, increasing the number of loci scored will probably increase the probability of a false exclusion (Kalinowski et al., 2007) . CERVUS has no formal framework for handling null alleles, but it does detect their presence and estimate their frequencies by examining deviations from Hardy-Weinberg equilibrium (Jones et al., 2010). The user is then left to decide whether to exclude affected loci in an analysis or not.

4. Materials and methodology

4.1. Materials

Two highly polymorphic microsatellite multiplexes, a 7-plex and a 6-plex had been developed for European larch (*Larix decidua*). These markers were tested on 413 individuals from 18 populations and their 13 loci were found to have allele numbers between 9 and 36 (Wagner, Gerber, & Petit, 2012). To enhance their polymorphism, only microsatellite motif with high number of repeats were selected for their development.

The existing genetic data set on European larch was used for the study. The data was obtained from twenty forest stands in Vienna, Austria. The data set is made up of parental genotype file and offspring genotype file. The genotype files consist of all typed individuals that will be involved in the pedigree reconstruction. The parental genotype file consists of typed genotypes of all individuals contesting the parentage of the offspring whiles the offspring genotype file consists of all typed genotypes of individuals who are to be assigned parentage. The data consisted of 53 parents and 1417 offspring. The number of typed loci was 13.

The statistical tool excel was used to prepare the genotype files for use by CERVUS 3.0 in the allele frequency analysis, simulation, and parentage analysis.

4.2 Methodology

A. Pedigree reconstruction using existing data (SSR markers)

B. Iterative reduction of the sample size and its effect on the accuracy of the resulting pedigree (part I. of the retrospective study)

C. Iterative reduction of the marker loci and its effect on the accuracy of the resulting pedigree (part II. of the retrospective study)

A. Pedigree Reconstruction Using Existing Data (SSR Markers)

4.3 Allele Frequency Analysis

The allele frequency analysis was performed to generate the allele frequencies at each locus and help in the generation of the various statistical information on each locus which, will eventually be used in the simulation and parental analysis. The genotype file, a combined genotype of the parents and the offspring was used in allele frequency analysis. The following inputs were made into the allele frequency analysis set up: the IDs and first allele of the genotype file were in the 1st and the 2nd columns respectively. The number of loci was thirteen while the minimum expected frequency was set at five. Estimation of null allele frequency was performed as well as the testing of Hardy-Weinberg. Bonferroni correction was used to evaluate significance. The data generated in this analysis was saved and used in the simulation and parental analysis respectively.

4.4 Simulation Analysis

The total offspring and potential parent's population used in the CERVUS 3.0 simulation setup were 50000 and 53 respectively. A proportion sample of 0.9, typed loci of 0.986100, and a minimum typed locus of 6 were inputted into the CERVUS simulation setup. Confidence was calculated using Delta and the simulation output was saved and used in the parental analysis. The above information is summarized in table 1 below.

Table 1. Summary of parameters used in simulation

Parameters	Used in simulation
Number of parents	53
Number of offspring	50000
Proportion sampled	0.9
Typed loci	0.9861
Mistyped loci	0.0139
Relaxed confidence level	80%
Strict confidence level	95%

4.5 Pedigree Reconstruction

In the CERVUS 3.0. analysis module, the parental and offspring population was set at 53 and 50000 respectively. The total number of maker loci from the genotype data was thirteen. Candidate parent sampled was set at 0.9 and the proportion of typed loci was set at 0.9861. The minimum typed locus was set at six. The presence of null alleles as well Hardy-Weinberg equilibrium were tested. Delta was used to determine confidence with strict and relaxed confidence set at 95% and 80% respectively. Bonferroni correction was used to evaluate significance.

The previous information on the upstream analysis, that is, allele frequency analysis and simulation analysis were required for the pedigree reconstruction. The parentage analysis module was fed with the names of the offspring and names of potential parents. CERVUS automatically calculates the LOD scores for every potential parent or parent pair and also, evaluating the confidence of the LOD or Delta score of the most likely parent or parent pair by using the appropriate LOD or Delta criteria by the simulation. The output data generated from the parentage analysis were saved and the trio confidence for each output data was evaluated and presented in tables and graphs.

B. Iterative reduction of the sample Size and its effect on the resulting pedigree (part I. of the retrospective study).

4.6 Reduced Sample size

The parent population of 53 was systematically reduced by 10%. Each reduced parent population formed a sample size. The sample size was therefore between 1 and 0.1. In the allele frequency analysis, ID and first allele of offspring in the column were set as one and two respectively in the CERVUS 3.0 module. The number of loci was set at thirteen. Testing for Hardy-Weinberg equilibrium, Bonferroni correction and estimation of null alleles were selected in the CERVUS 3.0 module. The selected minimum expected frequency was set at five. Each reduced sample size of the parent population was used in simulation and parentage analysis. The output data generated for each parent sample size analyzed was evaluated against pedigree constructed using total offspring population and the result presented in tables and graphs.

C. Iterative reduction of the number of loci and its effect on the accuracy of the resulting pedigree (part II. of the retrospective study).

4.7 Reduced Loci

The total candidate parents and offspring population of 53 and 50000 respectively, were inputted into the CERVUS analysis module. Sampled candidate parents was set at 0.9 and proportion of typed loci set at 0.9861. Strict and relaxed confidence was set at 95% and 80% respectively. Delta was used in the determination of the confidence.

The thirteen loci marker were systematically reduced by one locus and each reduced number of loci was used in allele frequency analysis, simulation and parental analysis. The output data from each reduced locus were evaluated, and their results presented in tables and graphs.

5. Results

5.1 Evaluation of Individual Microsatellite Markers

The evaluation of the microsatellite markers was done to ascertain their suitability in inferring parentage based on allele frequencies, Observed and Expected Heterozygosity (H_E and H_O), Polymorphic Information Content (PIC), exclusion probability as well as null allele frequency. Results of the parameters evaluated from the genetic data of European larch are presented in tables 1, 2, 3 and 4.

5.1.1 Allele Frequency

A text-based file of genotypes is read at one or more loci by the allele frequency module and the number of times that each allele at each locus occurs is counted. The data generated is used to calculate statistical parameters such as Expected heterozygosity, Polymorphic Information Content (PIC), Average Exclusion Probability. The estimation of Hardy-Weinberg equilibrium and null allele frequency is however optional.

The number of alleles per locus for the thirteen microsatellite markers ranged between seven and thirty-three. Two of the markers, bcKL263_55 and bcLK211_FA were found to have 33 alleles each while Ld101_56 have only seven alleles. The allele frequencies of the thirteen microsatellite markers tested ranged between 0.0003 and 0.8490. Only two of the microsatellite markers (Ld101_56 and Ld42_FA) have their allele frequencies greater than 0.50.

The most common allele for marker bcKL263_55 was found to be allele 217 with a 15.25% frequency while marker bcLK211_FA with allele 195 as most common allele has a 27.55% frequency. Marker Ld101_56 having only seven alleles was also found to have allele 198 as its most common allele and it formed 84.90% of the total alleles for this marker. The proportion of each allele for each marker is illustrated on pie charts in the various appendixes.

5.1.2 Estimated Heterozygotes and Homozygotes

The markers, bcLK253_53 and Ld101_56 were found to contain the highest and the least numbers of heterozygote individuals respectively. bcLK253_53 is made up of 86.8% of heterozygote individuals whilst Ld101_56 is made up of 26.1% heterozygotes. Ld101_56 is the only marker having more homozygote individuals than heterozygotes. The marker with the least homozygote individuals corresponded with bcLK253_53. This information is presented in Table 2.

Table 2. A summary of the estimated Heterozygotes and Homozygotes

Locus	Typed individuals	Percentage of Heterozygotes (%)	Percentage of Homozygotes (%)
bcLK189_FA	1456	78.1	21.9
Ld56_53	1463	75.4	24.6
Ld30_55	1451	76.9	23.1
bcLK263_55	1456	84.2	15.8
bcLK228_56	1459	81.9	18.1
bcLK211_FA	1452	78.7	21.3
bcLK253_53	1455	86.8	13.2
Ld50_55	1444	68.8	31.2
Ld31_56	1443	78.3	21.7
Ld42_FA	1453	50.4	49.6
Ld45_53	1433	65.2	34.8
Ld58_55	1451	82.4	17.6
Ld101_56	1454	26.1	73.9

5.1.3 Expected Heterozygosity

The Expected Heterozygosity is used to estimate the fraction of all the individuals in a population that are heterozygous at any random chosen locus in the population. It is a measure of the informativeness of a locus. Loci of expected heterozygosity of 0.5 or less is generally not desirable for large-scale parentage analysis.

The expected heterozygosity for the thirteen microsatellite markers ranged between 0.271 and 0.912. Apart from the marker, Ld101_56, all the microsatellite markers have their expected heterozygosity greater than 0.5. All markers have expected heterozygosity values greater than their respective observed heterozygosity. Four markers have their expected heterozygosity values less than the mean heterozygosity which have a value of 0.774. The marker with the largest expected heterozygosity was bcKL263_55 (0.912).

Table 3. A summary of allele frequency analysis

Locus	K	N	HObs	HExp	PIC	NE-1P	NE-2P	NE-PP	NE-I	NE-SI	HW	F (Null)
bcLK189_FA	15	1456	0.781	0.88	0.867	0.396	0.245	0.093	0.027	0.317	***	**0.0589**
Ld56_53	16	1463	0.754	0.792	0.767	0.567	0.388	0.197	0.067	0.371	NS	0.0248
Ld30_55	19	1451	0.769	0.803	0.778	0.554	0.378	0.193	0.064	0.364	***	0.0204
bcLK263_55	33	1456	0.842	0.912	0.905	0.301	0.177	0.051	0.014	0.298	***	0.04
bcLK228_56	18	1459	0.819	0.898	0.888	0.346	0.208	0.069	0.02	0.306	***	0.0453
bcLK211_FA	33	1452	0.787	0.866	0.854	0.412	0.259	0.096	0.03	0.324	***	**0.0501**
bcLK253_53	18	1455	0.868	0.876	0.863	0.4	0.249	0.093	0.028	0.319	NS	0.0038
Ld50_55	22	1444	0.688	0.762	0.727	0.625	0.447	0.258	0.091	0.392	***	**0.0525**
Ld31_56	14	1443	0.783	0.828	0.809	0.499	0.329	0.148	0.048	0.348	***	0.0259
Ld42_FA	9	1453	0.504	0.552	0.503	0.835	0.681	0.51	0.25	0.537	***	0.0423
Ld45_53	12	1433	0.652	0.772	0.736	0.619	0.442	0.258	0.088	0.386	***	**0.084**
Ld58_55	22	1451	0.824	0.854	0.84	0.44	0.281	0.111	0.035	0.332	NS	0.0179
Ld101_56	7	1454	0.261	0.271	0.259	0.962	0.853	0.74	0.544	0.751	NS	0.017

5.1.4 Polymorphic information content

Polymorphic information content (PIC) and expected heterozygosity values are both used to measure the genetic variability among breeding populations. Polymorphism in genetic markers is thus used to determine genetic diversity in populations.

During allele frequency analysis, the expected heterozygosity values were found to be between 0.271 (Ld101_56) and 0.912 (bcLK263_55). The observed heterozygosity values were also found to range between 0.261 (Ld101_56) and 0.868 (bcLK253_53). All the expected heterozygosity values were greater than their respective observed heterozygosity values.

The polymorphic information content values for the markers were found to be between 0.259 (Ld101_56) and 0.905 (bcLK263_55). The polymorphic information values were all found to be less than their respective expected heterozygosity values. This information is presented in Table 3.

5.1.5 Null Allele Frequency

The presence of null alleles in microsatellite markers is normally an indication of more homozygous individuals in the population for specific markers. This normally leads to mismatches in the genotype of the offspring and the parents at the locus. The null allele frequency which is denoted as F (null) values as presented in Table 3 ranged from 0.0038 (bcLK253_53) to 0.0840 (Ld45_53). The markers, bcLK189_FA, Ld50_55 and Ld45_53 all had their null allele frequencies greater than 0.05. Null allele frequencies greater than 0.05 are printed in bold in Table 3.

5.1.6 Hardy-Weinberg Test

The 13 microsatellite markers were tested to determine if the population was in Hardy-Weinberg equilibrium. 9 out 13 of the microsatellite markers were found to conform to Hardy-Weinberg equilibrium while 4 of the markers were not. The above information is summarized in table 4.

Table 4. A summary of Hardy-Weinberg Equilibrium

Locus	Degree of freedom	Chi-square	P-Value	Significance
bcLK189_FA	36	202.23	1.03E-23	***
Ld56_53	21	34.64	0.0309	NS
Ld30_55	21	59.99	0.00001282	***
bcLK263_55	28	94.33	4.12E-09	***
bcLK228_56	28	97.52	1.27E-09	***
bcLK211_FA	15	102.22	4.95E-15	***
bcLK253_53	36	48	8.71E-02	NS
Ld50_55	15	93.06	2.66E-13	***
Ld31_56	15	49.21	1.62E-05	***
Ld42_FA	6	33.21	0.00000957	***
Ld45_53	10	149.73	4.24E-27	***
Ld58_55	15	17.99	0.2633	NS
Ld101_5	3	8.32	0.0398	NS

5.2 A. Pedigree reconstruction using existing data (SSR markers)

When 50000 offspring and 53 putative parents were used in the simulation analysis, assignments were achieved at both confidence levels. At a strict confidence of 95%, 42,667 offspring were successfully assigned to their respective parents and this accounted for 85% assignment rate at this confidence level. At a relaxed confidence level 80%, 45076 of the offspring were successfully assigned to their parents and this corresponds to a 90% assignment at this confidence level. 4924 of the offspring were however not assigned to any of the putative parents and this number represents approximately 10% of the offspring population. This information is presented in table 5.

Table 5. Pedigree reconstruction for simulation analysis

Level	Confidence (%)	Critical Delta	Assignment	Assignment Rate (%)
Strict	95	2.47	42667	85
Relaxed	80	0	45076	90
Unassigned			4924	10
Total			50000	100

In real data analysis or parentage analysis, 1124 out of the 1417 offspring population were successfully assigned to their respective parents at a 95% confidence level and this figure represents a 79% of assignment. At a relax confidence level of 80%, 1180 offspring were assigned successfully. This number corresponds with 83% of successful assignment. 234 offspring which represents about 17% of the offspring population were not assigned to any parent at this confidence level. This information is presented in table 6.

Table 6. Pedigree reconstruction for real data

			Assignment		Assignment Rate (%)	
Level	**Confidence (%)**	**Critical Delta**	**Observed**	**Expected**	**Observed**	**Expected**
Strict	95	2.47	1124	1207	79	85
Relaxed	80	0	1180	1275	83	90
Unassigned			234	139	17	10
Total			1414	1414	100	100

5.3 B. Iterative reduction of the sample Size and its effect on the resulting pedigree (part I. of the retrospective study).

The parent population size was systematically reduced by 10% in order to verify the effect that the reduced population sizes will have on the pedigree constructed. During simulation where 50000 offspring were used in the analysis, fluctuations in assignments were observed as the parent sample sizes got smaller. At a 95% confidence level, 42667 assignment was observed when the total parent population was used in the simulation analysis. A sample size of 0.9 resulted in a reduced assignment of 42396 assignments. An increase in the number of assignments were recorded from sample sizes 0.9 to 0.7, 0.6 to 0.5 and 0.4 to 0.3. The highest number of assignments was observed when the parent population was 0.3 of the total parent population.

At 80% confidence, the number of assignments declined as the parent sample size was systematically reduced by 10%. When the total parent population of 53 was used in simulation analysis, 45076 number of assignments were observed whiles 41446 number of assignments were recorded with a parent population of 0.1. The number of unassigned offspring increased with declining sample size. 4924 unassigned offspring were observed when the total population was used for simulation and by the time the sample size was reduced to 0.1, the number of unassigned offspring had increased to 8554.

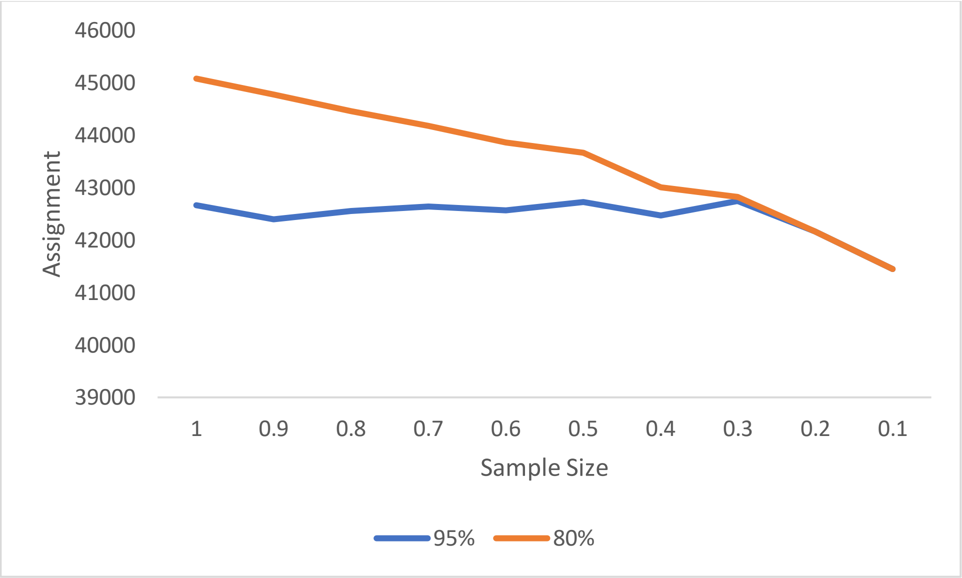


Figure 1. A chart showing parent pair assignment for Simulation analysis

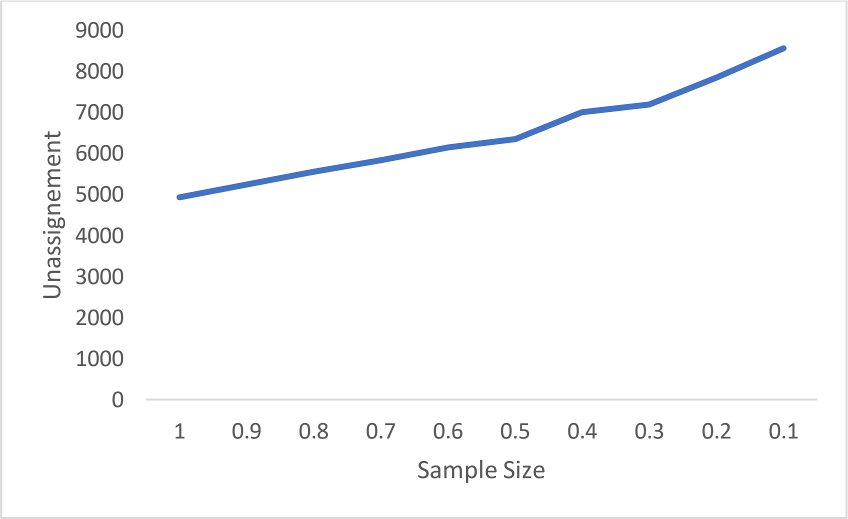


Figure 2. A chart showing unassigned parentage for Simulation analysis

The number of parent pair assignment for the simulation analysis was found to decline as the size of the parental population decrease in size. 1431 and 15 number of parent pair assignments were observed for sample sizes of 1 and 0.1 respectively. This information is illustrated with figure 3.

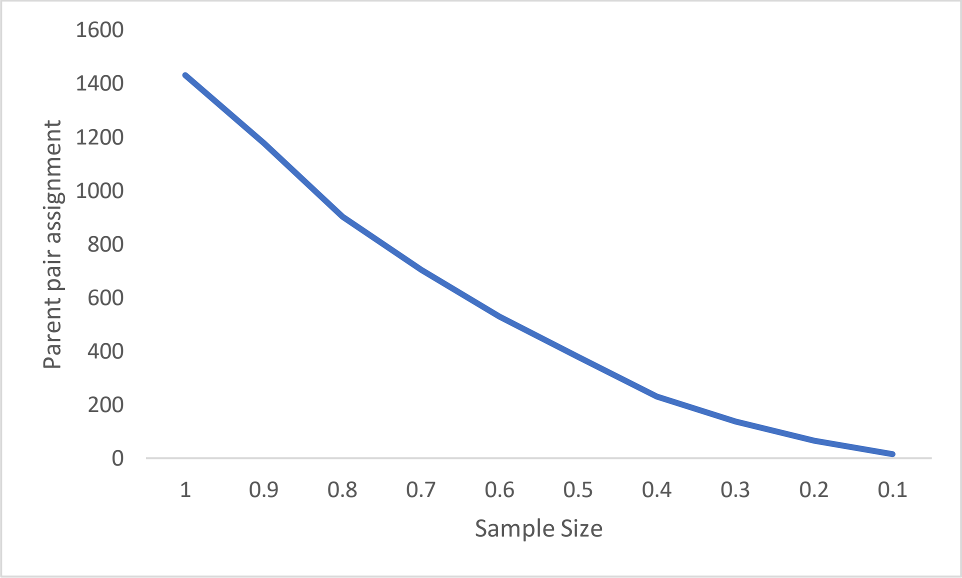


Figure 3. A chart showing parent pair assignment for simulation analysis

For real data analysis, the number of assignments increased as the sample sizes declined. At a 90% confidence level, 1124 number of assignments were recorded when all parent sample size was used for parentage analysis. The highest number of assignments at this confidence level was observed with the least sample size of 0.1.

A constant assignment of 1180 was realized for all sample sizes at a confidence level of 80%. The number of offspring not assigned at this confidence level was 234 for all sample sizes and this number represents 17% of the progeny population. This information is represented by figure 4.

The parent pair assignments for parentage analysis were found to decline as the parent population size was systematically reduced by 10%. The parent pair assignment for real data analysis was the same as the results obtained for simulation analysis.

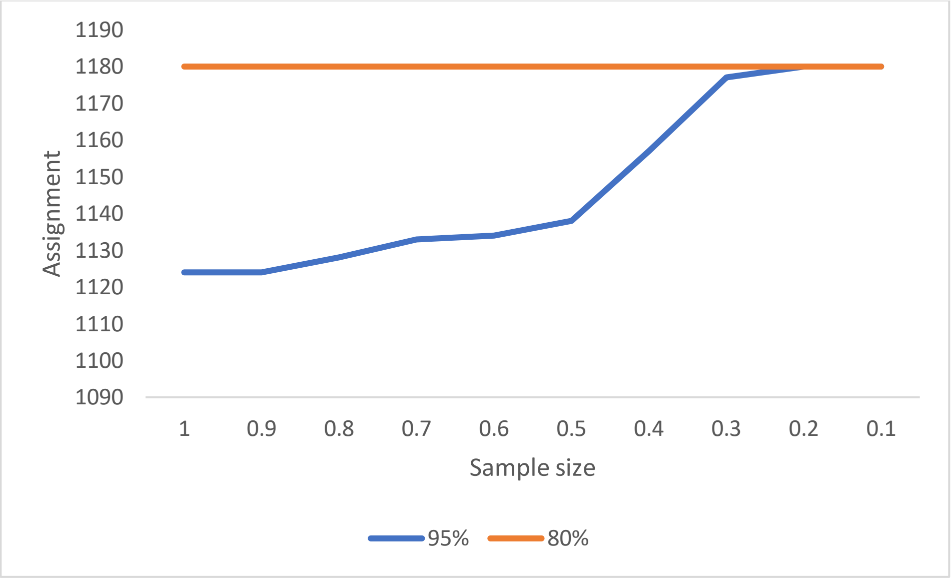


Figure 4. Assignment for real data analysis

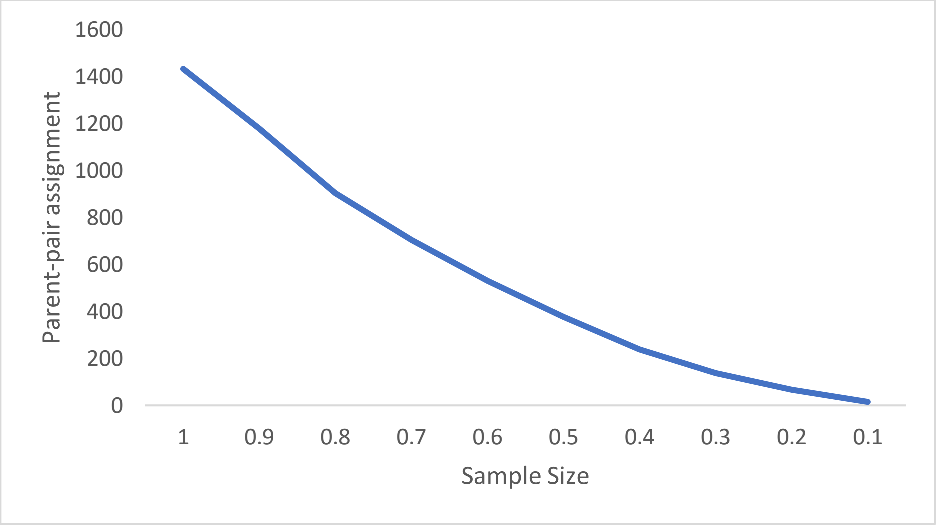


Figure 5. A chart showing parent pair assignment for real data analysis

5.3 Trio Confidence Level

The confidence level was set at strict and relaxed confidence of 95% and 80% respectively and this was to give an indication on the tolerance of false positive assignments. During the evaluation of the trio confidence, it was realized that majority of the assignments were observed at 95% confidence and the assignments at this confidence level were also found to decline with decreasing sample size. (*) and (+) is used to denote assignment at strict 95% and relaxed 80% confidence respectively while (–) is used to denote the confidence of parent pair assignment in which the parent is the most likely candidate parent but could be assigned at either confidence level. Candidates not considered as most likely parents were not assigned any confidence level. A constant figure of 237 individuals were not assigned at the trio confidence level at all the sample sizes and this figure represents 16.7% of the progeny population. The above information is presented in table 7.

Table 7. A table showing the rate of trio confidence of assignment at different sample sizes

	1	**0.9**	**0.8**	**0.7**	**0.6**	**0.5**	**0.4**	**0.3**	**0.2**	**0.1**
(*)	79.3%	79.3%	79.6%	79.9%	80.0%	80.4%	81.6%	83.1%	83.3%	83.3%
(+)	4.0%	4%	3.7%	3.3%	3.2%	2.8%	1.6%	0.2%	0%	0%
(-)	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%

5.4 Reduced Loci Number

In the simulation analysis, no assignment was realized when the number of loci were less than six. Assignments were however realized when six or more loci were used in the simulation analysis. With six number of loci, 50000 offspring population, 53 prospective parents, and a 0.9 candidate parents sampled, 29106 parent pair assignments were realized at a strict 95% confidence. The assignment at this confidence represents 56% of the offspring population. The number of assignments increased up to twelve loci and a decline in assignment was observed when thirteen loci was used in the simulation analysis. This information is represented in figure 6.

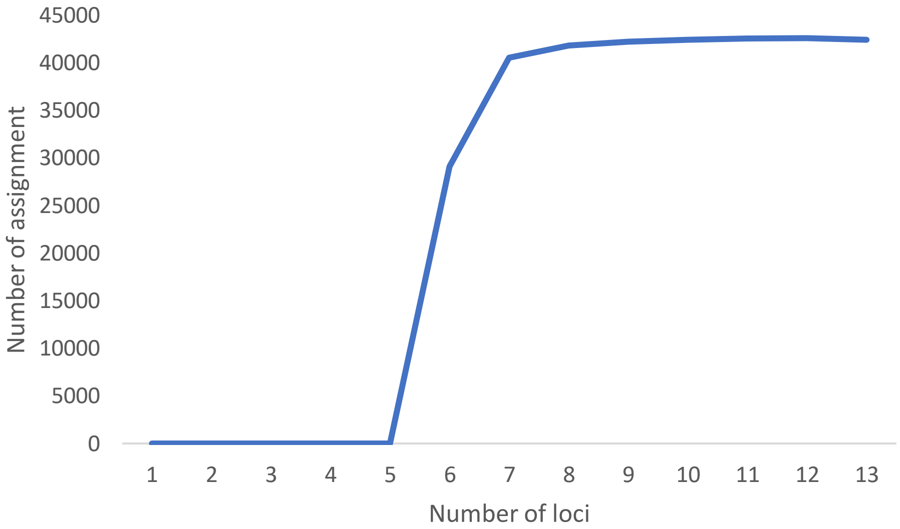


Figure 6. A graph showing simulation analysis assignments with different loci numbers at 95% confidence

At a relaxed confidence of 80%, assignment was possible only when more than five loci were used in the simulation analysis. When six loci were used in the simulation, 37896 of the offspring were assigned to the respective parents. From six loci to eight loci, an increase in assignments were observed. One possible reason for this increase in assignment is that, as the number of marker loci increase in number, more genetic information is made available which will help to facilitate the assignment of the offspring to their parents. A decline in assignment were observed from nine to thirteen loci. The least and the highest assignments corresponded with six and eight number of loci respectively. (see figure 7).

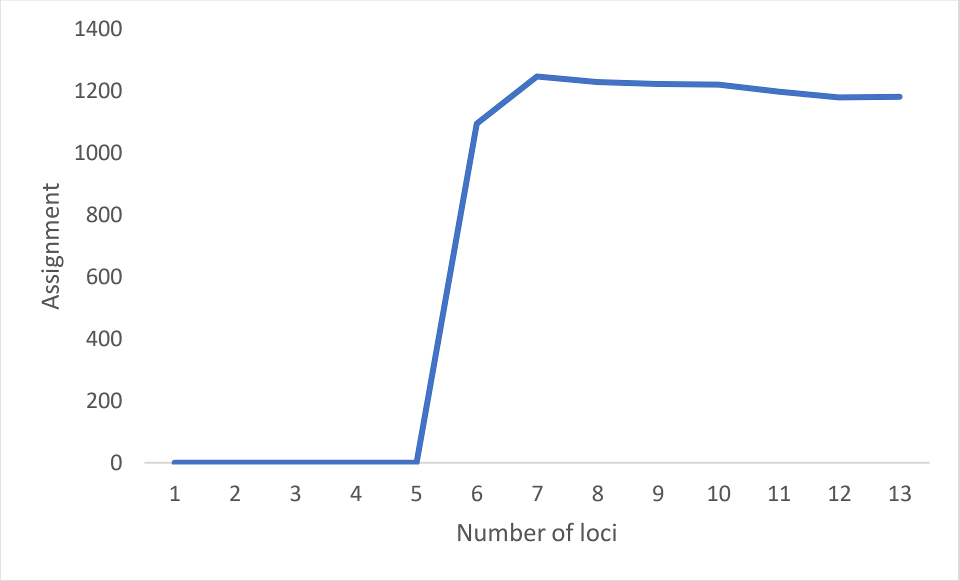


Figure 7. A chart showing parent pair assignments with different loci numbers at 80% confidence

In real data or parentage analysis, no assignment was observed with loci number less than six. Fewer loci mean less information and with more potential parents around, it was not obvious which parent is the most probable and no assignment was observed.

At strict 95% confidence, assignment increased from six loci to twelve loci with a decline in assignment at thirteen loci. All observed assignments were found to be slightly lower than their expected, except assignment at six loci, which was equal to the expected assignment. Highest number of assignments corresponded with twelve loci.

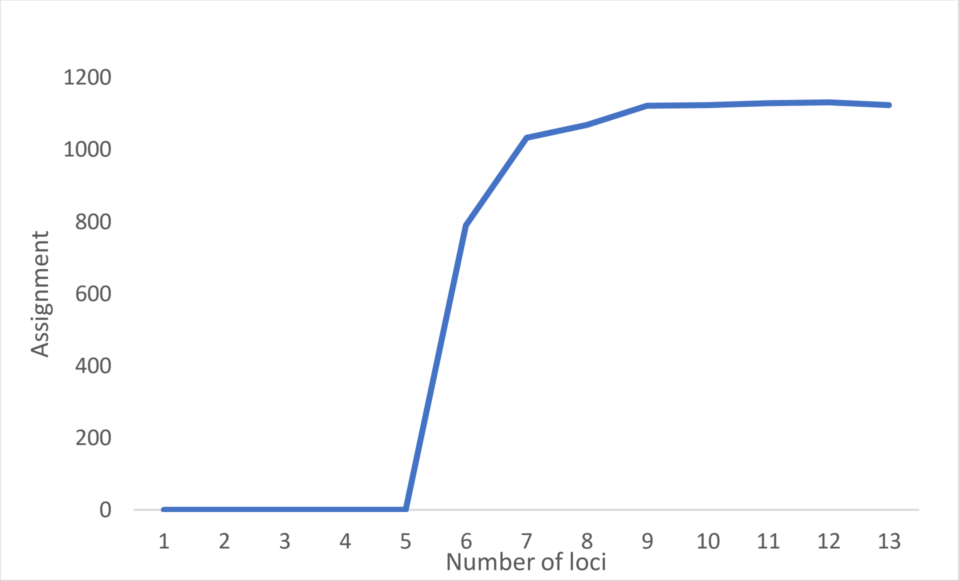


Figure 8. A chart showing real data assignments with different loci numbers at 95% confidence

At 80% confidence, assignment was not achieved when the number of loci was below six. Assignments for parentage analysis at this confidence level increased from six loci to eight loci after which a decline in assignment was recorded. The highest number of assignments was realized with eight loci. Unassigned parentage was at its peak when the loci number was six.

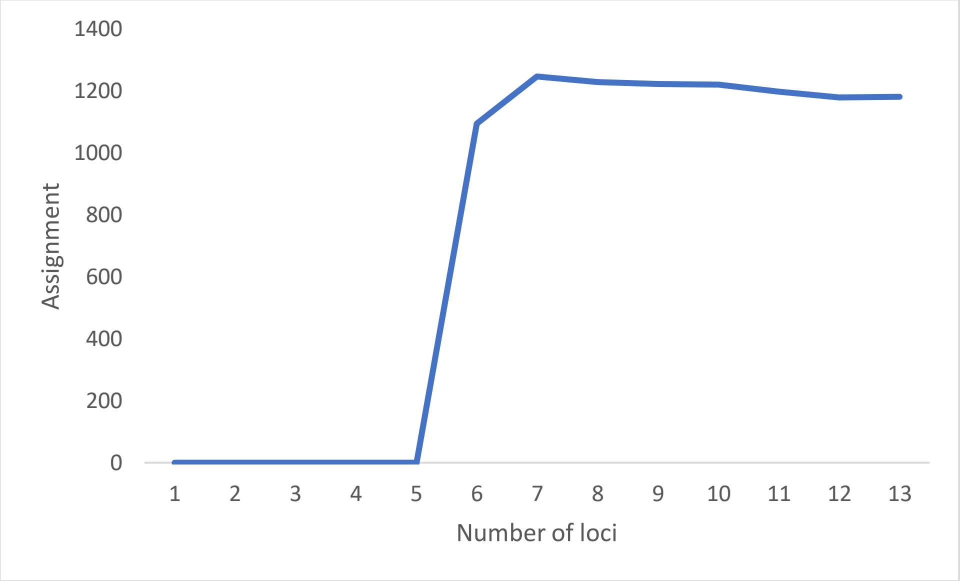


Figure 9. A chart showing parentage analysis assignments with different loci numbers at 80% confidence

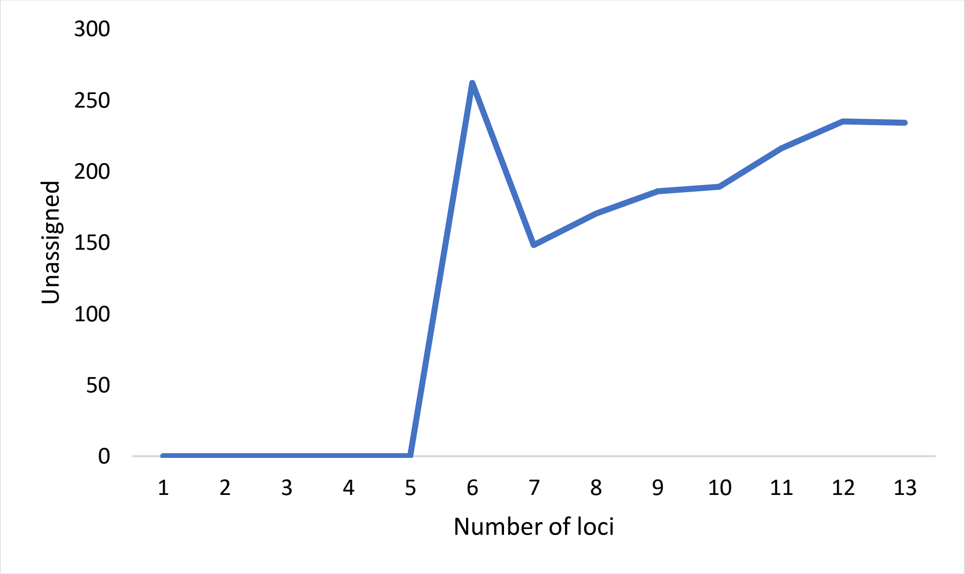


Figure 10. A chart showing unassigned parentage for different loci numbers for parentage analysis

5.5. Trio Confidence

Except for six number of locus, all the assignments were either made at a strict or relaxed confidence of 95% and 80% respectively. However, majority of the parent pair assignments were made at a strict confidence of 95%. The number of parent pair assignments observed with a strict confidence of 95% increased from locus number of six to eleven but from locus number of eleven to thirteen, there was a reduction in the number of assignments at this confidence level.

Generally, the number of parent pair assignments observed at a relaxed confidence of 80% declined with increased number of locus. From locus number of twelve to thirteen, there was a slight increase in the number of assignments made with a relaxed confidence of 80%. It was only in locus number of six that the most likely candidate parents were not able to be assigned to neither strict confidence of 95% nor relaxed confidence of 80%. (*), (+), and (-) were used to denote strict confidence of 95%, relaxed confidence of 80%, and the inability to make assignment at either confidence level respectively. This information is presented in figure 11.

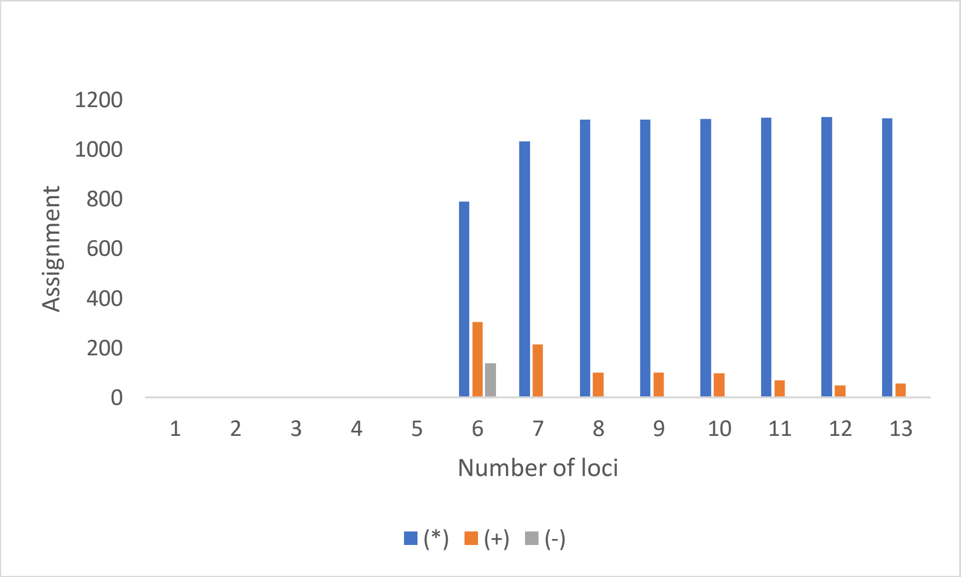


Figure 11. A chart showing trio confidence of assignment for different loci numbers for real data analysis

6. Discussion

6.1 (A). Real Data Evaluation

6.1.1. Allele Frequency

The term, allele frequency is used to describe how common an allele is in a population. It reflects the genetic variability of a population. A population may increase or decrease some of its alleles in the population as an adaptation to an evolutional force acting on it.

An allele may be common or rare in a population based on its frequency in the population. Joyce & Tavar (1995) described a rare allele as one that appear twice in every 100 sample size or 200 times in every 10,000 sample size. Based on this definition for a rare allele, all thirteen microsatellite markers used for the study can be described as having rare allele since they all have alleles with frequencies less than 0.02.

Frequent or infrequent alleles do not really influence fitness. Most alleles with low frequency are normally underrepresented in a population but they may be beneficial to the population, only that they may be new and will need time for it to be replicated in the population. Rare alleles have been identified to confer a fertility advantage in plants since pollens carrying this rare allele are not rejected by the incompatibility reaction of the recipient plant (Charlesworth & Guttman, 1997).

The locus bcLK263_55 and bcLK211_FA have the same number of alleles and they represent the markers with the greatest number of alleles with their most frequent alleles being 217 and 195 respectively. The major cause of large numbers of alleles in a population is attributed to mutation(Charlesworth & Guttman, 1997). Generally, the higher the number of alleles for a loci and the closer the PIC value to 1, the more desirable the loci is for parentage analysis (Botstein, White, Skolnick, & Davis, 1980). The number of alleles for the loci this study were found to be between seven and thirty-three but in a similar study conducted by Wagner et al., (2012) recorded alleles ranging between nine and thirty-six even though thirteen loci were used in both studies.

6.1.2. Estimated Heterozygosity and Homozygosity

From Table 4, all markers were made up of heterozygous and homozygous individuals. Except for the marker, Ld101_56, the number of heterozygous individuals were more than the homozygous individuals for the rest of the markers. Heterozygosity in a population is highest when all the allele frequencies in the population are equal. Higher than expected heterozygosity values normally occur when two isolated populations that were each homozygous for different alleles start to interbreed.

Higher homozygosity in a population can be attributed fragmentation of anthropogenic habitats which results in reduced fitness due to inbreeding depression (Pérez-Tris et al., 2019). More homozygous individuals indicate a lower genetic variation in the population while more heterozygous individuals denote a higher genetic variation in the population (Friedrich, Supervisor, & Visser, 2009). In this study, the marker, Ld101_56 was found to have excess homozygous individuals. This marker is expected to be low in genetic variability. High homozygosity in a population will usually lead to loss of heterozygosity and the fixation of deleterious alleles that can drive the population into extinction

Since reduced population size, strong founder effects and geographic isolation have been linked with higher homozygosity and its negative impact on a population, mixing of populations can be an ideal means of elevating high homozygosity and its effects on the population (Bertolini et al., 2018).

6.1.3. Expected Heterozygosity

Expected heterozygosity also known as gene diversity is usually one of the means by which variation within a population can be measured. The estimation of expected heterozygosity of related or inbred individuals usually results in a decline in accuracy and precision owing to the fact that all individuals in the sample will be sharing the same copies of the alleles present in the sample population (Harris & DeGiorgio, 2017).

Observed heterozygosity of markers are usually compared with the expected under Hardy-Weinberg equilibrium. Lower observed heterozygosity than the expected is usually linked with inbreeding while higher observed heterozygosity will depict the mixing of two populations that were previously isolated.

In this study, as shown in table 3, twelve of the microsatellite markers were found to have their expected heterozygosity values greater than their respective observed heterozygosity. The mean expected heterozygosity value for the thirteen SSR marker used in this study was 0.774. However, in a similar study conducted by Nardin et al., (2015), recorded a slight lower mean heterozygosity value of 0.761.

6.1.4. Polymorphic Information Content (PIC)

The polymorphic information content value is the probability of a marker genotype to allow for the deduction that, a marker allele was from its parent (Elston, 2005). It is used to infer the degree of informativeness of genetic markers. Polymorphic information content values greater than 0.5 are highly informative, when less than 0.5 but greater than 0.25, it is reasonably informative and when less than 0.25, it is slightly informative (Botstein et al. 1980).

From Table 3, all but marker Ld101_56 have polymorphic information content values greater than 0.5 and this observation was also reported by (Gramazio et al., 2018). The PIC value of Ld101_56 was slightly above 0.25. Based on the assertion of Botstein et al., (1980) twelve out of the Thirteen microsatellite markers can be described as highly informative while one (Ld101_56) is reasonably informative. The average polymorphic information content of 0.754 observed for the thirteen SSR markers employed in this study is a bit higher than the average PIC value of 0.713 that was observed by Gramazio et al., (2018) in a similar study. The PIC value obtained for this study is higher than what was obtained for genic SSRs in *Larix kaempferi* (Chen, Xie, & Sun, 2015).

6.1.5. Null Allele Frequency

From Table 3, the null allele frequencies for the microsatellite markers ranged between 0.0038 and 0.084. For this study, only three of the thirteen microsatellite markers (bcLK189_FA, Ld50_55 and Ld45_53) were found to have null alleles. All the three markers including an extra two markers (Ld30_55 and Ld42_FA) were identified by Wagner et al., (2012) to have null alleles.

The null allele frequency values for the three markers were greater than 0.05. Markers with null allele frequencies greater than 0.05 tend to have many individuals with many homozygous alleles which results in low level of polymorphism in these markers. These markers are usually not desirable in pedigree analysis. Null alleles are also source of mismatch (Marshall et al., 1998) . The Three microsatellite markers, bcLK189_FA, Ld50_55 and Ld45_53 are characterized with low level of polymorphism and may not be ideal for parentage analysis on the basis of their null allele frequency values.

6.1.6. Hardy-Weinberg Equilibrium

Nine out of the thirteen microsatellite markers tested conformed to the Hardy-Weinberg equilibrium. The microsatellite markers which were in disequilibrium with Hardy-Weinberg expectations included locus Ld56_53, bcLK253_53, Ld58_55, and Ld101_56. A deviation at a single locus may be as a result of natural selection acting on nearby gene. However, Hardy-Weinberg deviation at several locus or all loci gives an indication of population substructure.

It is not common to have a natural population with all genotypes conforming to a Hardy-Weinberg equilibrium since the tendency of one or more of the Hardy-Weinberg assumptions being violated is very high.

6.2 A. Pedigree reconstruction using existing data (SSR markers)

With a parent population of 53 and a progeny population of 1417, a pedigree was successfully constructed using the SSR markers for European larch. 85% of the offspring population were assigned to their respective parents at a strict confidence of 95%. At a relaxed confidence of 80%, a rate of 90% assignment was achieved. 10% of the offspring population were not assigned to any parent.

6.3 B. Iterative reduction of sample Size

One of the basic quantities in statistics that influence many aspects of a study is the sample size. One of the aims of the study was to investigate the influence of the sample size on a given pedigree. In most statistical studies, the widely accepted practice is that, the larger the sample size you have, the more likely your results will also be better (Sánchez-montes, Ariño, Vizmanos, Wang, & Martínez-solano, 2017) . The results obtained for parent pair assignment for this study reflects this common practice in most statistical studies. The wisdom behind this practice is that, the more sample size that is acquired, the more it becomes representative of the whole population. This means that more individuals in the population will be included in your study. In genetic analysis, allele frequencies are often estimated from the sample being analyzed for relatedness. Smaller sample size in this case can affect estimated allele frequencies resulting in less precise estimates which has the potential of introducing bias due to relatedness between individuals that were not accounted (Wang, 2012).

This practice, however, does not hold for all statistical analysis. Studies that involve unites that depend on other units may not always conform to this rule (Raffa & Thompson 2016) . In studies involving genetic analysis, the sample size alone does not determine the precision and accuracy of the pedigree. The quality of the markers in the study can equally affect outcome of the study. As shown by the results of the study, the number of assignments were not always high with higher sample sizes in both simulation and parentage analysis.

6.4 C. Iterative reduction of the number of loci and its effect on the accuracy of the resulting pedigree (part II. of the retrospective study)

The second part of the retrospective study was set to investigate the influence of the number of marker loci on pedigree. Based on the studies conducted, the number of marker loci have been found to influence the output of a pedigree. The number of assignments observed in simulation and parentage analysis were greatly affected by the number of loci used in the allele frequency analysis.

One of the key factors to consider in genetic analysis is informativeness of the marker loci used (Wang, 2012). In this study, no assignments were observed for the simulation and parentage analysis when the number of loci used in the allele frequency analysis was less than six. One possible explanation to this unsigned parentage is that, with less than six loci, and many potential parents, less information is made available and as a result, identification of the true parent becomes difficult. Assignment at both confidence of 95% and 80% saw an increase in assignment from six loci to a point after which the increase in loci number resulted in a decline in assignment. This observation is in line with Slate et al., (2000) assertion that large numbers of loci can cause mismatches between parents and offspring at some loci due to mutation or typing errors. These mismatches can prevent an offspring from being assigned to its true parent. The highest number of assignments was realized when twelve loci was used in the allele frequency analysis.

The number of markers needed for a study will somehow depend on the kind and purpose of the study. Markers required for genetic linkage is less demanding as compared to markers required for counselling purposes (Botstein et al., 1980).

7. Conclusion

The evaluation of the microsatellite markers was done to authenticate their informativeness and suitability for genetic analysis. The following conclusions were arrived at after the evaluation of the markers.

None of the thirteen SSR markers have alleles less than four and as a result, they are all desirable for the evaluation of genetic analysis. The microsatellite Ld101_56 was found to have more homozygote individuals than heterozygotes. Its expected heterozygosity was also found to be less than 0.5. This excess homozygosity and low expected heterozygosity for the marker Ld101_56 is an indication of it being low in variability and as such, might not be ideal for parentage analysis. The observed heterozygosity values of the SSR markers were found to be higher than their respective expected heterozygosity values and this can be an indication of the force of inbreeding operating in the European larch population.

The marker, Ld101_56 have PIC value of 0.259 while the rest of the markers have their PIC values higher than 0.5. Ld101_56 is therefore considered to be reasonably informative while the rest are deemed to be highly informative. Three of the markers, bcLK189_FA, Ld50_55 and Ld45_53 have null alleles frequencies greater than 0.05 and this might be attributed to the presence of excess homozygous alleles. The markers Ld56_53, bcLK253_53, Ld58_55 and Ld101_56 do not conform to Hardy-Weinberg equilibrium. Since the deviation did not occur at a single locus but several, this might be an indication of inbreeding in the population or the presence of null alleles.

Parent-pair assignments in simulation and parentage analysis were found to increase with increasing parent sample size. But the results for assignments at 95% confidence level for simulation and parentage analysis was opposite to what was observed for parent pair assignment. The effect of sample size on pedigree constructed using SSR markers is not straight forward and may be dependent on factors such as quality of the marker used.

Increasing the number of loci will result in an increase in the number of assignments at a 95% confidence whiles at a confidence level of 80% the opposite result will be observed. At a given number of loci, a decline in the number of assignments should be expected. From this study, it has been established that, at least six microsatellite marker loci for European larch will be required for an assignment to be made in simulation and parentage analysis.

8. References

Abdurakhmonov, I. Y. (2016). Introduction to Microsatellites: Basics, Trends and Highlights. *Microsatellite Markers*. https://doi.org/10.5772/66446

Beja-Pereira, A., Alexandrino, P., Bessa, I., Carretero, Y., Dunner, S., Ferrand, N., … Cañon, J. (2003). Genetic characterization of Southwestern European bovine breeds: A historical and biogeographical reassessment with a set of 16 microsatellites. *Journal of Heredity*, *94*(3), 243–250. https://doi.org/10.1093/jhered/esg055

Bertolini, F., Cardoso, T. F., Marras, G., Nicolazzi, E. L., Rothschild, M. F., & Amills, M. (2018). Genome-wide patterns of homozygosity provide clues about the population history and adaptation of goats. *Genetics Selection Evolution*, *50*(1), 1–12. https://doi.org/10.1186/s12711-018-0424-8

Blouin, M. S. (2003). DNA-based methods for pedigree reconstruction and kinship analysis in natural populations, *18*(10), 503–511. https://doi.org/10.1016/S0169-5347(03)00225-8

Botstein, D., White, R. L., Skolnick, M., & Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics*. https://doi.org/10.17348/era.9.0.151-162

Butler, J. M., Coble, M. D., & Vallone, P. M. (2007). STRs vs. SNPs: Thoughts on the future of forensic DNA testing. *Forensic Science, Medicine, and Pathology*, *3*(3), 200–205. https://doi.org/10.1007/s12024-007-0018-1

Cervera, M. T., Gusmão, J., Steenackers, M., Peleman, J., Storme, V., Vanden Broeck, A., … Boerjan, W. (1996). Identification of AFLP molecular markers for resistance against Melampsora larici-populina in Populus. *Theoretical and Applied Genetics*. https://doi.org/10.1007/BF00224069

Charlesworth, D., & Guttman, D. S. (1997). Plant genetics: Seeing selection in S allele sequences. *Current Biology*, *7*(1), 34–37. https://doi.org/10.1016/s0960-9822(06)00015-7

Chen, X. Bin, Xie, Y. H., & Sun, X. M. (2015). Development and characterization of polymorphic genic-SSR markers in Larix kaempferi. *Molecules*, *20*(4), 6060–6067. https://doi.org/10.3390/molecules20046060

Christie, M. R. (2010). Parentage in natural populations: Novel methods to detect parent-offspring pairs in large data sets. *Molecular Ecology Resources*, *10*(1), 115–128. https://doi.org/10.1111/j.1755-0998.2009.02687.x

Da Ronch, F., Caudullo, G., Tinner, W., & de Rigo, D. (2016). Larix decidua and other larches in Europe : distribution , habitat , usage and threats. *San-Miguel- Ayanz, J., de Rigo, D., Caudullo, G., Houston Durrant, T., Mauri, A. (Eds.), European Atlas of Forest Tree Species.*, 108–110.

Day-williams, A. G., Blangero, J., Dyer, T. D., Lange, K., & Ã, E. M. S. (2011). Linkage Analysis Without Defined Pedigrees, *370*, 360–370. https://doi.org/10.1002/gepi.20584

Einspahr, D. W., Wyckoff, G. W., & Fiscus, M. H. (1984). Larch A Fast-Growing Fiber Source States and Northeast For the Lake Forests of the Lake contain enough hardwood, 104–106.

El-Kassaby, Y. A., Funda, T., & Lai, B. S. K. (2010). Female reproductive success variation in a pseudotsuga menziesii seed orchard as revealed by pedigree reconstruction from a bulk seed collection. *Journal of Heredity*, *101*(2), 164–168. https://doi.org/10.1093/jhered/esp126

Elston, R. C. (2005). Polymorphism Information Content. *Encyclopedia of Biostatistics*, 5078. https://doi.org/10.1002/0470011815.b2a05078

Eriksson, G., Ekberg, I., & Clapham, D. (2013). Genetics Applied to Forestry An Introduction Third edition, 208. Retrieved from http://vaxt2.vbsg.slu.se/forgen/Forestry_Genetics.pdf

Farooq, S., & Azam, F. (2002). Molecular Markers in Plant Breeding-I : Concepts and Characterization Molecular Markers in Plant Breeding-I : Concepts and Characterization, (March 2014). https://doi.org/10.3923/pjbs.2002.1135.1140

Friedrich, H., Supervisor, P., & Visser, M. C. (2009). *Evaluation of microsatellite markers* *for parentage verification in South African Angora goats*.

Garcia, M. V., & Garcia, M. V. (n.d.). We are IntechOpen , the world ’ s leading publisher of Open Access books Built by scientists , for scientists TOP 1 %.

Gilmore, D. W., & David, A. J. (2002). Current trends in management practices for European larch in North America. *Forestry Chronicle*. https://doi.org/10.5558/tfc78822-6

Global Forest Resources Assessment 2015 Desk reference. (2015).

Gower, S. T., & Richards, J. H. (1990). Larches: Deciduous Conifers in an Evergreen World. *BioScience*. https://doi.org/10.2307/1311484

Gramazio, P., Plesa, I. M., Truta, A. M., Sestras, A. F., Vilanova, S., Plazas, M., … Sestras, R. E. (2018). Highly informative SSR genotyping reveals large genetic diversity and limited differentiation in European larch (Larix decidua) populations from Romania. *Turkish Journal of Agriculture and Forestry*, *42*(3), 165–175. https://doi.org/10.3906/tar-1801-41

Grattapaglia, D. (2018). Quantitative Genetics and Genomics Converge to Accelerate Forest Tree, *9*(November), 1–10. https://doi.org/10.3389/fpls.2018.01693

Gudeta, T. B. (2018). Molecular marker based genetic diversity in forest tree populations, *2*(4), 176–182. https://doi.org/10.15406/freij.2018.02.00044

Haappanen, M., Jansson, G., Bräuner Nielsen, U., Steffenrem, A., & Stener, L.-G. S. (2015). *The status of tree breeding and its potential for improving biomass production - A review of breeding activities and genetic gains in Scandinavia and Finland*.

Harris, A. M., & DeGiorgio, M. (2017). An unbiased estimator of gene diversity with improved variance for samples containing related and inbred individuals of any ploidy. *G3: Genes, Genomes, Genetics*, *7*(2), 671–691. https://doi.org/10.1534/g3.116.037168

Harrison, H. B., Saenz-Agudelo, P., Planes, S., Jones, G. P., & Berumen, M. L. (2013). Relative accuracy of three common methods of parentage analysis in natural populations. *Molecular Ecology*, *22*(4), 1158–1170. https://doi.org/10.1111/mec.12138

Huisman, J. (2017). Pedigree reconstruction from SNP data: parentage assignment, sibship clustering and beyond. *Molecular Ecology Resources*, *17*(5), 1009–1024. https://doi.org/10.1111/1755-0998.12665

Jarne, P., & Lagoda, P. J. L. (1996). Microsatellites, from molecules to populations and back. *Trends in Ecology and Evolution*, *11*(10), 424–429. https://doi.org/10.1016/0169-5347(96)10049-5

Jiang, G.-L. (2013). Molecular Markers and Marker-Assisted Breeding in Plants. In *Plant Breeding from Laboratories to Fields*. https://doi.org/10.5772/52583

Johnson, R., Clair, B. S. T., & Lipow, S. (n.d.). Genetic Conservation in Applied Tree Breeding Programs, 215–230.

Jones, A. G., & Ardren, W. R. (2003). Methods of parentage analysis in natural populations. *Molecular Ecology*, *12*(10), 2511–2523. https://doi.org/10.1046/j.1365-294X.2003.01928.x

Jones, A. G., Small, C. M., Paczolt, K. A., & Ratterman, N. L. (2010). A practical guide to methods of parentage analysis. *Molecular Ecology Resources*, *10*(1), 6–30. https://doi.org/10.1111/j.1755-0998.2009.02778.x

Jones, E., Chu, W. C., Ayele, M., Ho, J., Bruggeman, E., Yourstone, K., … Smith, S. (2009). Development of single nucleotide polymorphism (SNP) markers for use in commercial maize (Zea mays L.) germplasm. *Molecular Breeding*, *24*(2), 165–176. https://doi.org/10.1007/s11032-009-9281-z

Jones, O. R., & Wang, J. (2010). COLONY: A program for parentage and sibship inference from multilocus genotype data. *Molecular Ecology Resources*, *10*(3), 551–555. https://doi.org/10.1111/j.1755-0998.2009.02787.x

Joyce, P., & Tavar, S. (1995). Mathematical Mology The distribution of rare alleles, 602–618.

Kalinowski, S. T., Taper, M. L., & Marshall, T. C. (2007). Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. *Molecular Ecology*, *16*(5), 1099–1106. https://doi.org/10.1111/j.1365-294X.2007.03089.x

Keith, C. T., & Chauret, G. (1988). Basic wood properties of European larch from fast-growth plantations in eastern Canada. *Canadian Journal of Forest Research*. https://doi.org/10.1139/x88-204

Khlestkina, E. K., & Salina, E. A. (2006). SNP markers: Methods of analysis, ways of development, and comparison on an example of common wheat. *Russian Journal of Genetics*, *42*(6), 585–594. https://doi.org/10.1134/S1022795406060019

Kloch, M., Hawliczek-strulak, A., & Sekrecka-bielak, A. (2015). Molecular Markers In Forest Management And Tree Breeding : A review. *Annals of Warsaw University of Life Sciences*, *199*, 193–199.

Kumar, P., Gupta, V. K., Misra, a K., Modi, D. R., & Pandey, B. K. (2009). Southern Cross Journals © 2009 Potential of Molecular Markers in Plant Biotechnology. *Plant Biotechnology*, *2*(4), 141–162. Retrieved from http://www.pomics.com/Pradeep_2_4_2009_141_162.pdf

LEWANDOWSKI, A., & MEJNARTOWICZ, L. (1991). Linkage analysis of allozyme loci in Polish larch (Larix decidua subsp. polonica (Racib.) (Domin). *Hereditas*. https://doi.org/10.1111/j.1601-5223.1991.tb00560.x

Lucia, M., Vieira, C., Santini, L., Diniz, A. L., & Munhoz, C. D. F. (2016). Microsatellite markers : what they mean and why they are so useful, *328*, 312–328.

Matras J, L. P. (2008). *Technical Guidelines for genetic conservation of European larch (Larix decidua)*. google books.

Matson, S. E., Camara, M. D., Eichert, W., & Banks, M. A. (2008). P-LOCI: A computer program for choosing the most efficient set of loci for parentage assignment. *Molecular Ecology Resources*. https://doi.org/10.1111/j.1755-0998.2008.02128.x

Nadeem, M. A., Nawaz, M. A., Shahid, M. Q., Doğan, Y., Comertpay, G., Yıldız, M., …

Baloch, F. S. (2017). DNA molecular markers in plant breeding: current status and recent advancements in genomic selection and genome editing. *Biotechnology & Biotechnological Equipment*, *2818*(November), 1–25. https://doi.org/10.1080/13102818.2017.1400401

Namkoong, G., & Koshy, M. P. (2001). *Application of Genetic Markers to Forest tree species Draft report to IPGRI of the project "Developing Decision-making Strategies on Priorities for Conservation and Use of Forest Genetic Resources."*

Nardin, M., Musch, B., Rousselle, Y., Guérin, V., Sanchez, L., Rossi, J., … Rozenberg, P. (2015). Genetic differentiation of European larch along an altitudinal gradient in the French Alps, 517–527. https://doi.org/10.1007/s13595-015-0483-8

Nawrot, M., Pazdrowski, W., & Szymański, M. (2008). Dynamics of heartwood formation and axial and radial distribution of sapwood and heartwood in stems of European larch (Larix decidua Mill.). *Journal of Forest Science*. https://doi.org/10.17221/30/2008-jfs

Neale, D. B., Devey, M. E., Jermstad, I. D., Ahuja, M. R., Alosi, M. C., & Marshall, K. A. (1992). Use of DNA markers in forest tree improvement research. *New Forests*, *6*, 391–407. https://doi.org/10.1007/BF00120654

Pâques, L. E. (2001). Genetic control of heartwood content in larch. *Silvae Genetica*.

Pâques, L. E., Philippe, G., & Prat, D. (2006). Identification of European and Japanese larch and their interspecific hybrid with morphological markers: Application to young seedlings. *Silvae Genetica*. https://doi.org/10.1515/sg-2006-0018

Pérez-Tris, J., Llanos-Garrido, A., Bloor, P., Carbonell, R., Tellería, J. L., Santos, T., & Díaz, J. A. (2019). Increased individual homozygosity is correlated with low fitness in a fragmented lizard population. *Biological Journal of the Linnean Society*, *128*(4), 952–962. https://doi.org/10.1093/biolinnean/blz144

Putz, F. E., & Redford, K. H. (2010). The importance of defining "Forest": Tropical forest degradation, deforestation, long-term phase shifts, and further transitions. *Biotropica*, *42*(1), 10–20. https://doi.org/10.1111/j.1744-7429.2009.00567.x

Raffa, J. D., & Thompson, E. A. (2016). Power and Effective Study Size in Heritability

Studies. *Statistics in Biosciences*, (April 2015). https://doi.org/10.1007/s12561-016-9143-2

Sánchez-montes, G., Ariño, A. H., Vizmanos, J. L., Wang, J., & Martínez-solano, Í. (2017). Effects of Sample Size and Full Sibs on Genetic Diversity Characterization : A Case Study of Three Syntopic Iberian Pond-Breeding Amphibians, 1–9. https://doi.org/10.1093/jhered/esx038

Schmidtling, R. C., Robison, T. L., Mckeand, S. E., Rousseau, R. J., Allen, H. L., & Goldfarb, B. (2002). ree Improvement.

Sheikh, S. I., Berger-Wolf, T. Y., Khokhar, A. A., & Dasgupta, B. (2008). Consensus Methods for Reconstruction of Sibling Relationships from Genetic Data. *Proceedings of the 4th Multidisciplinary Workshop on Advances in Preference Handling*, 97–102.

Slate, J., Marshall, T., & Pemberton, J. (2000). A retrospective assessment of the accuracy of the paternity inference program CERVUS. *Molecular Ecology*. https://doi.org/10.1046/j.1365-294X.2000.00930.x

Städele, V., & Vigilant, L. (2016). Strategies for determining kinship in wild populations using genetic data. *Ecology and Evolution*, *6*(17), 6107–6120. https://doi.org/10.1002/ece3.2346

Staub, J. E., & Serquen, F. C. (1996). Genetic Markers , Map Construction , and Their Application in Plant Breeding, *31*(5), 729–741.

T. C. MARSHALL; J. SLATE; L. E. B. KRUUK; J. M. PEMBERTON. (1998). Statistical confidence for likelihood-based paternity, 639–655.

Thakur, R. B., & Schmerbeck, J. (2014). Role of Tree Breeding in Timber and Wood Supply in World and India: Status and Outlook. *The Initiation*. https://doi.org/10.3126/init.v5i0.10266

Tingey, S. V., & Del Tufo, J. P. (1993). Genetic analysis with random amplified polymorphic DNA markers. *Plant Physiology*. https://doi.org/10.1104/pp.101.2.349

Wagner, S., Gerber, S., & Petit, R. J. (2012). Two highly informative dinucleotide SSR multiplexes for the conifer Larix decidua (European larch). *Molecular Ecology Resources*, *12*(4), 717–725. https://doi.org/10.1111/j.1755-0998.2012.03139.x

Wang, J. (2012). Computationally efficient sibship and parentage assignment from multilocus marker data. *Genetics*. https://doi.org/10.1534/genetics.111.138149

Wellmann, R., & Bennewitz, J. (2019). Key Genetic Parameters for Population Management, *10*(August), 1–20. https://doi.org/10.3389/fgene.2019.00667

White, T L, Adams, W. T., & Neale, D. B. (2007). Genetic markers - morphological, biochemical and molecular markers. *Forest Genetics*, (5), 53–74. https://doi.org/10.1079/9781845932855.0053

White, Timothy L., Adams, W. T., & Neale, D. B. (2007). *Forest Genetics*. *CABI Publishing*. https://doi.org/10.1038/167764a0

Xu, Y., Li, P., Zou, C., Lu, Y., Xie, C., Zhang, X., & Prasanna, B. M. (2017). Enhancing genetic gain in the era of molecular breeding, *68*(11), 2641–2666. https://doi.org/10.1093/jxb/erx135

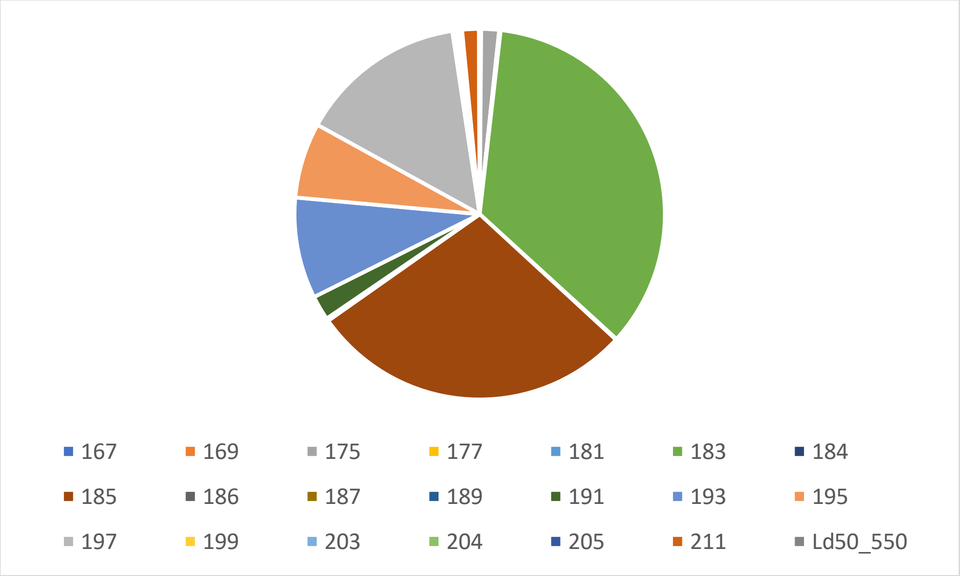
Yanchuk, A. D. (2009). Techniques in forest tree breeding. *Forests and Forest Plants*, *III*, 40–47.

Yang, H., Kang, W., Nahm, S., & Kang, B. (2015). *Current Technologies in Plant Molecular Breeding*. https://doi.org/10.1007/978-94-017-9996-6

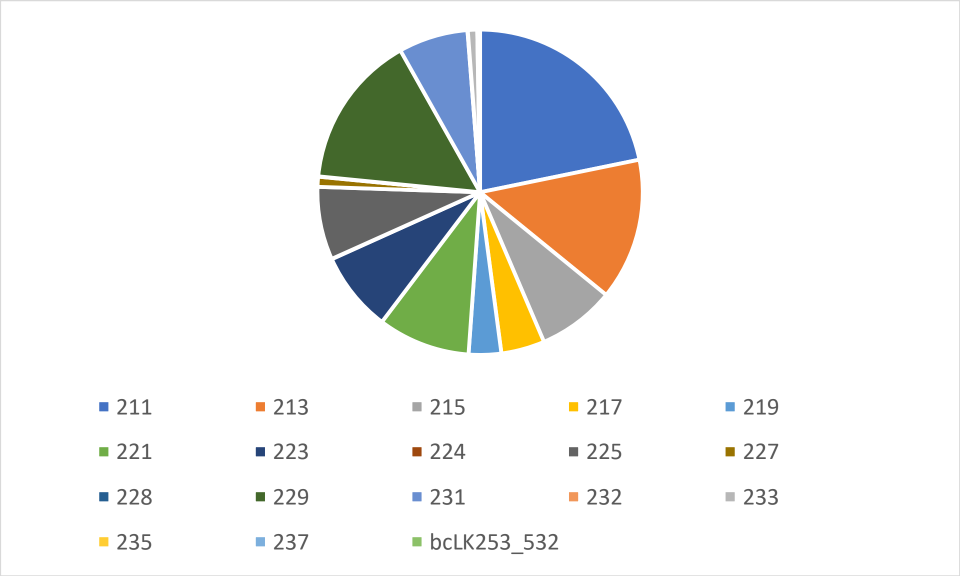
(http://www.tiem.utk.edu/~gross/bioed/bealsmodules/hardy-weinberg.html).

Appendices

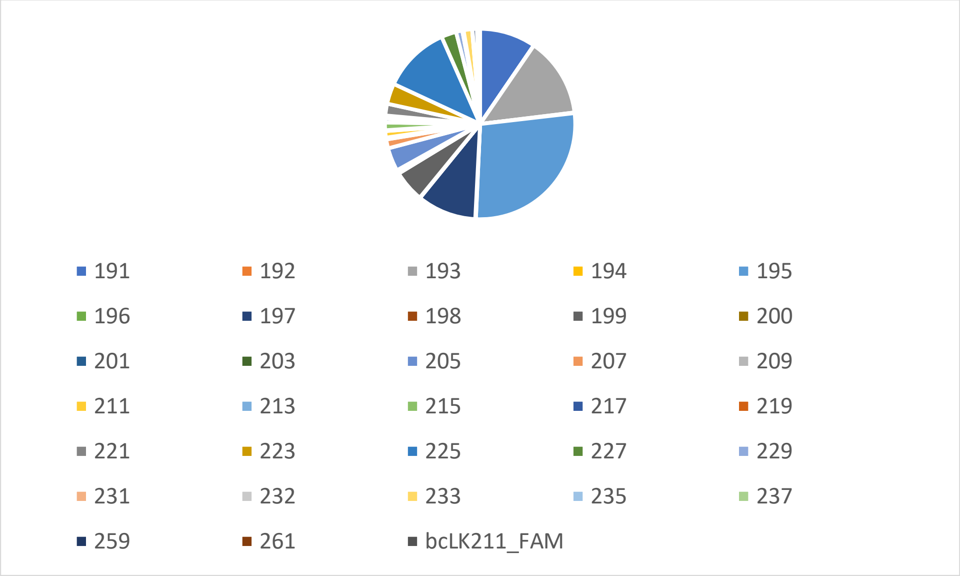
1. A chart showing the proportion of alleles for marker Ld50_55



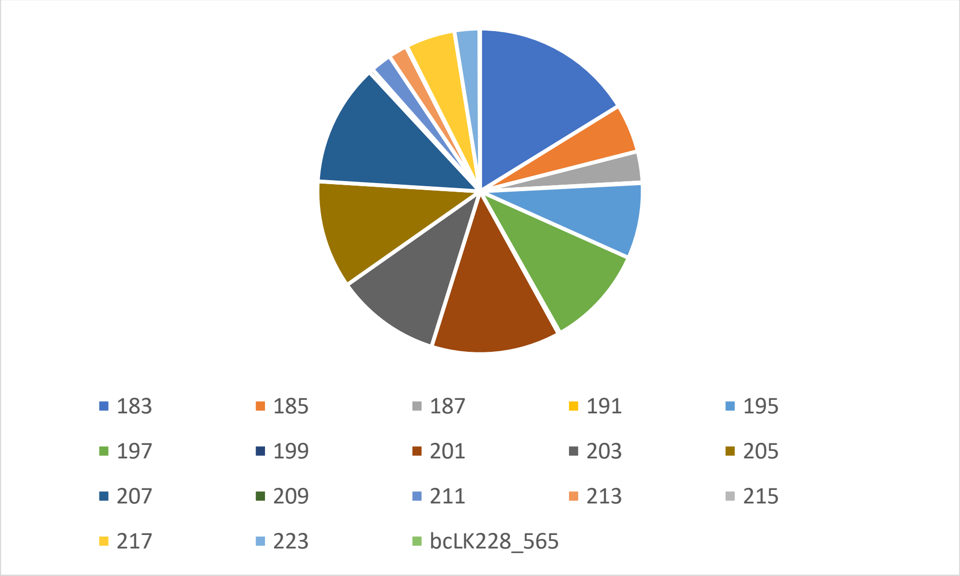
2. A chart showing the proportion of alleles for marker bcLK253_53



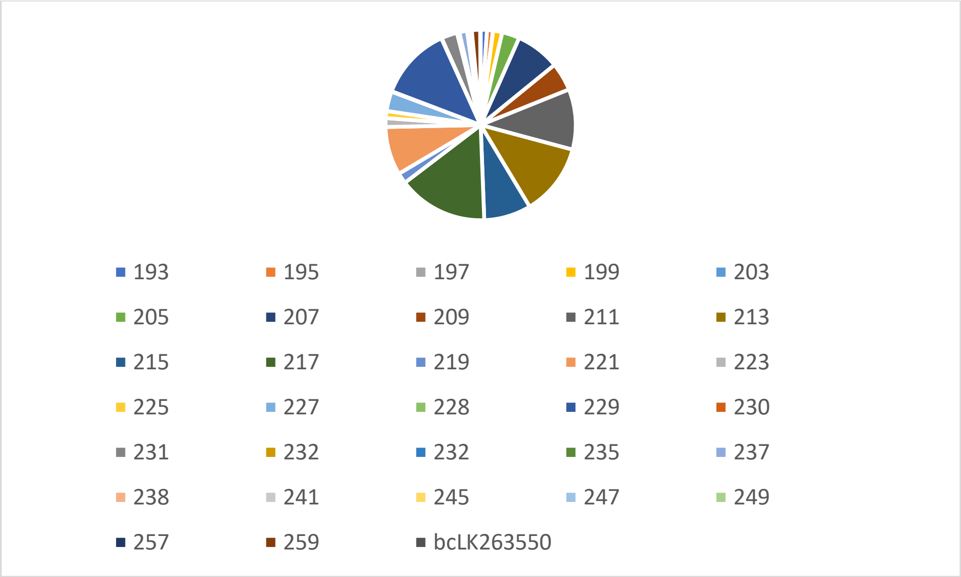
3. A chart showing the proportion of alleles for marker bcLK211_FA



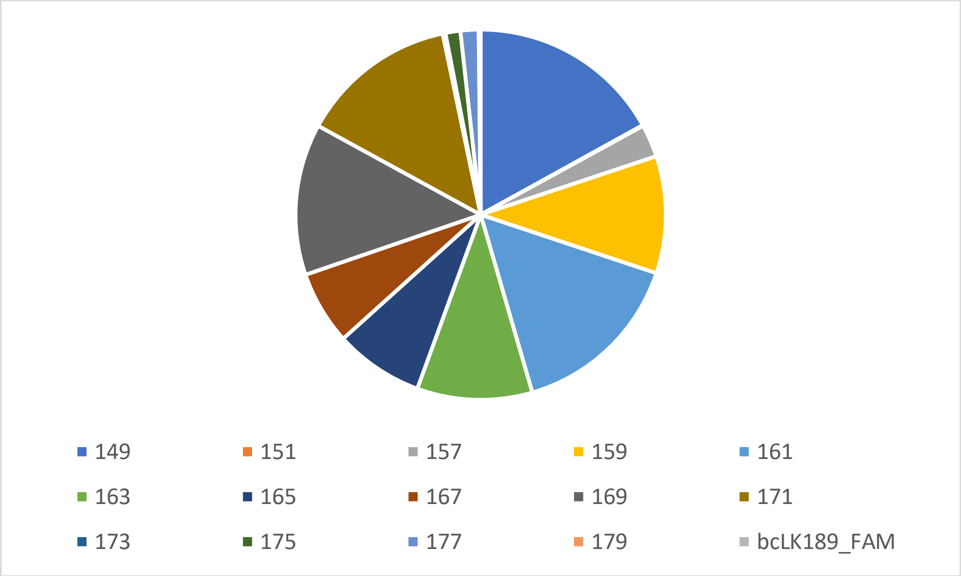
4. A chart showing the proportion of alleles for marker bcLK228_56



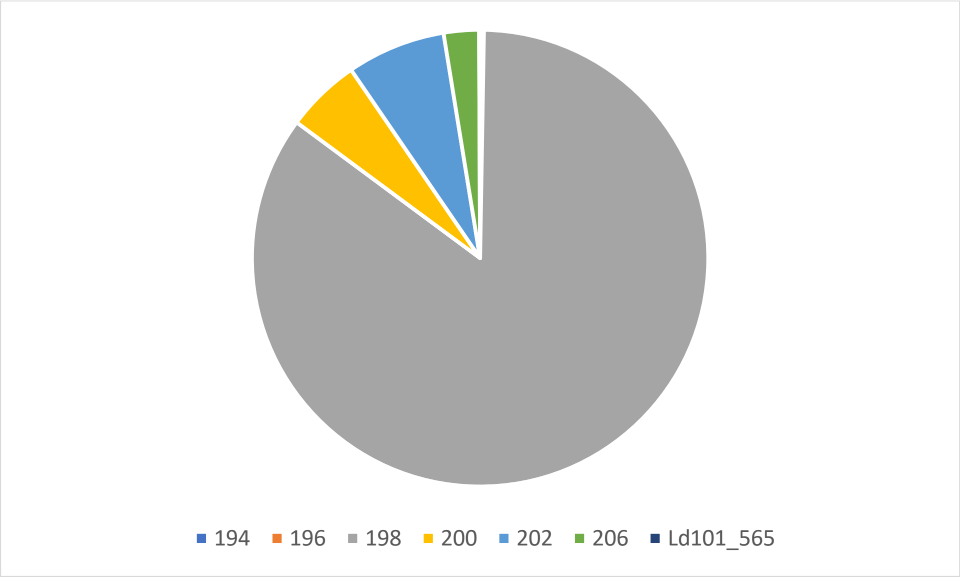
5. A chart showing the proportion of alleles for marker bcLK263_55



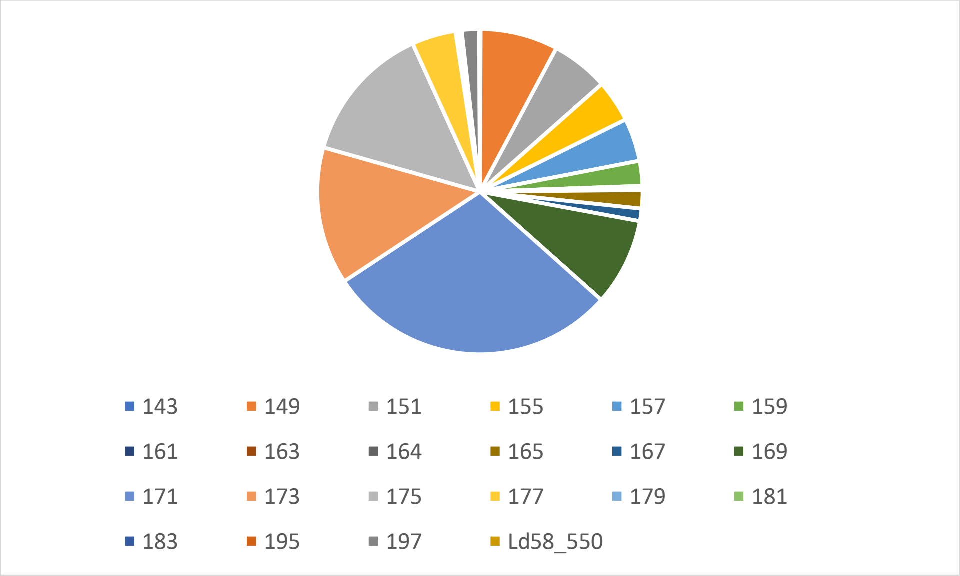
6. A chart showing the proportion of alleles for marker bcLK189_FA



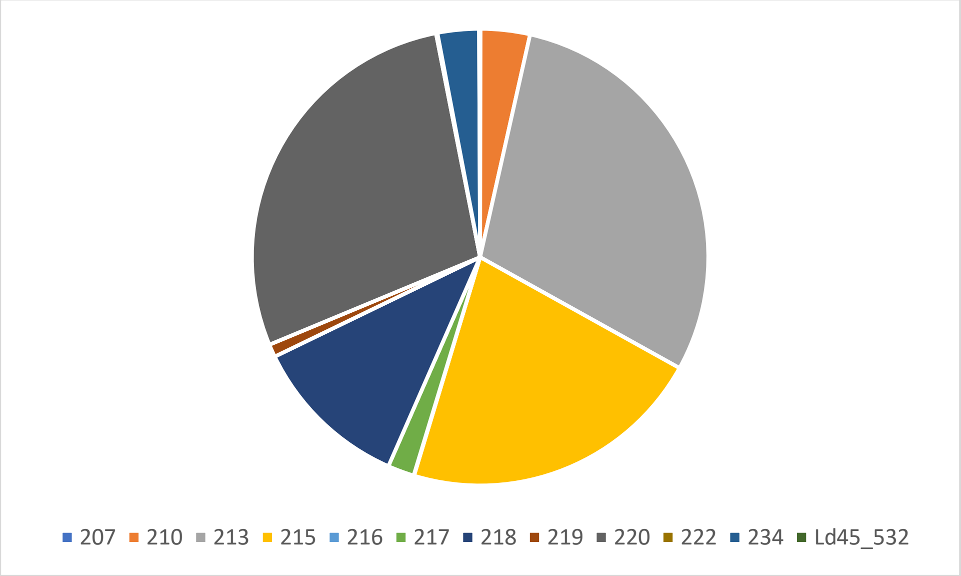
7. A chart showing the proportion of alleles for marker Ld101_56



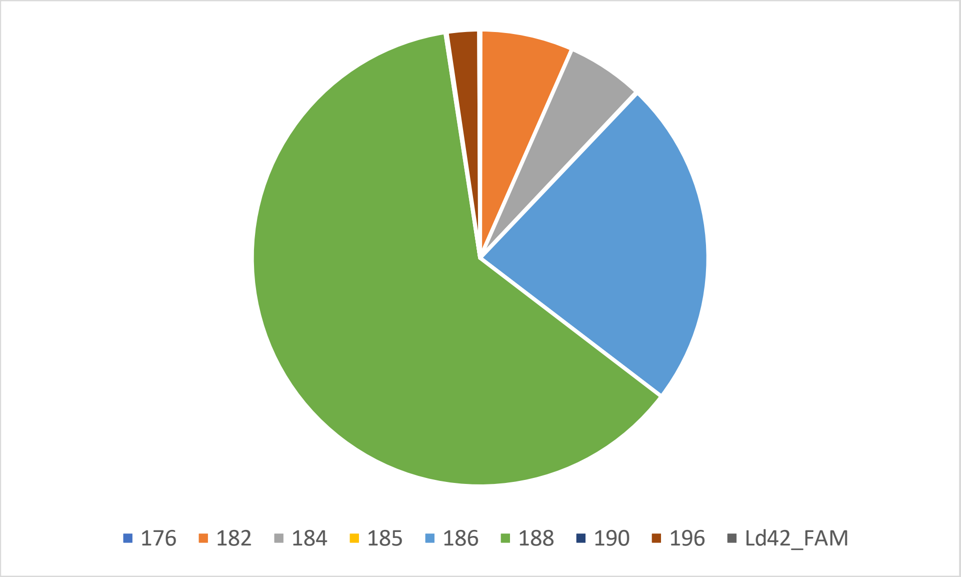
8. A chart showing the proportion of alleles for marker Ld58_55



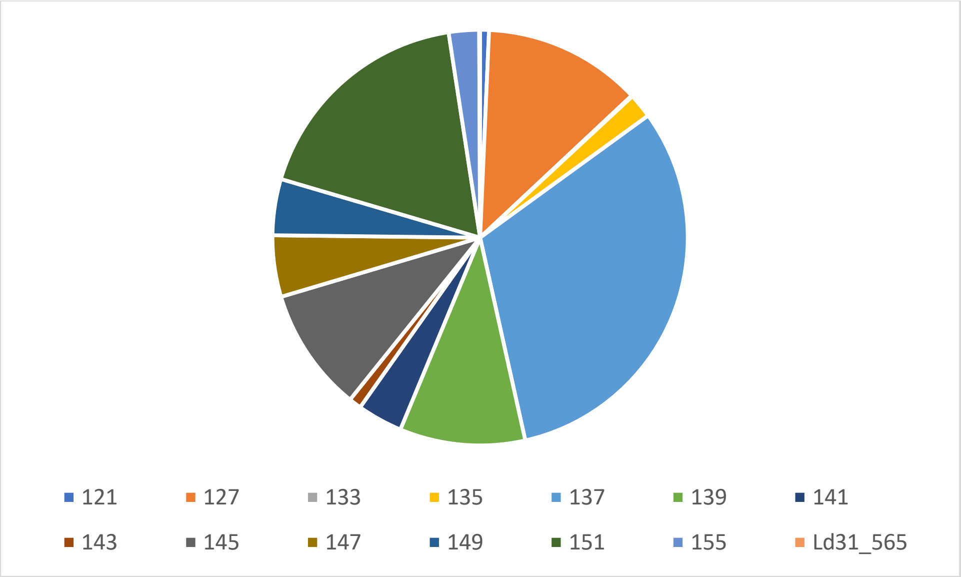
9. A chart showing the proportion of alleles for marker Ld45_53



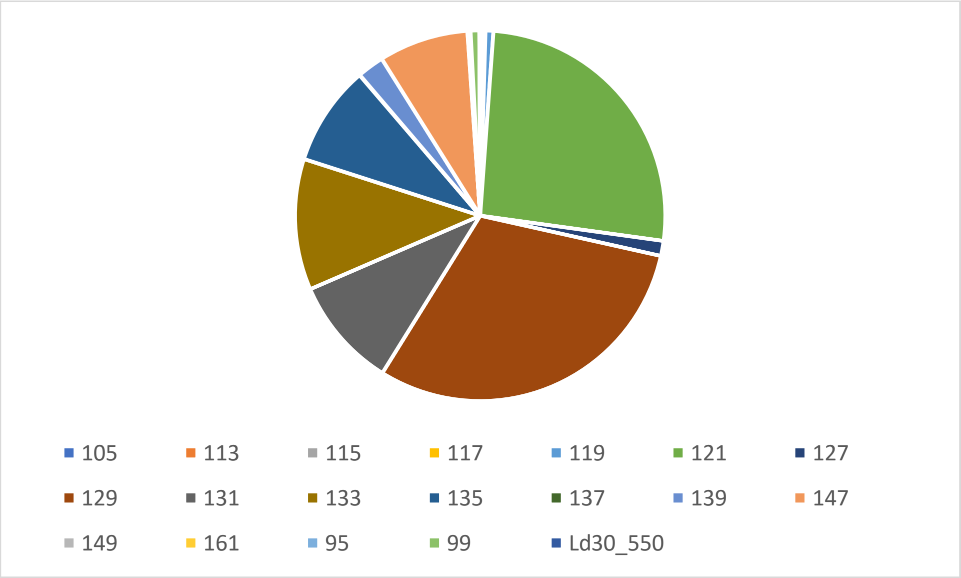
10. A chart showing the proportion of alleles for marker Ld42_FA



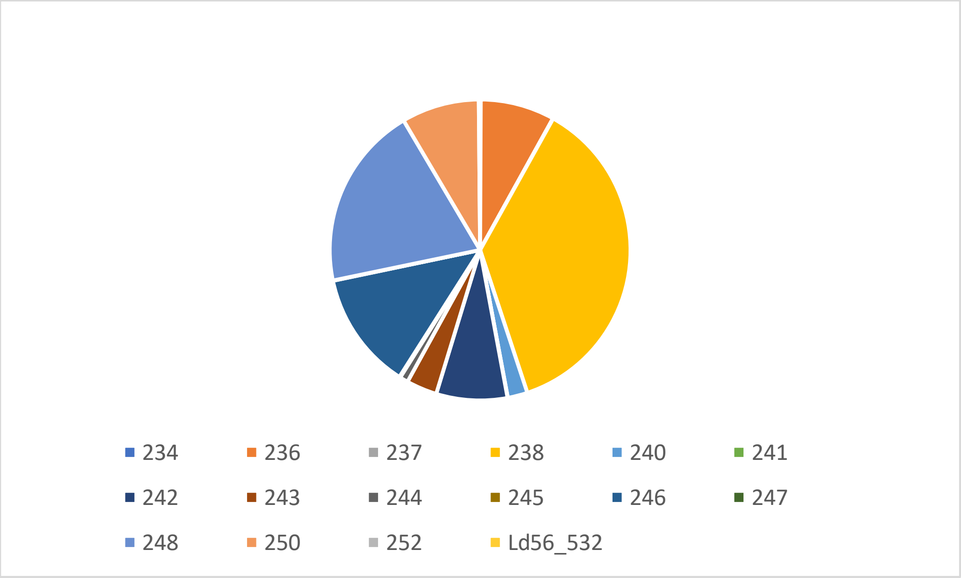
11. A chart showing the proportion of alleles for marker Ld31_56



12. A chart showing the proportion of alleles for marker Ld30_55



13. A chart showing the proportion of alleles for marker Ld56_53


Ld56_532