Growing confidence in forestry's future Research Programme

SCION FORESTS PRODUCTS INNOVATION

Date: March 2018
Reference: GCFF TN020

Technical Note

Reconstructing parentage in *Pinus radiata* using exome capture genotyping – proof of concept in FR10/0

Author/s: Natalie Graham, Ahmed Ismael & Toby Stovold

Corresponding author: natalie.graham@scionresearch.com

Summary: We have successfully recreated the pedigree by assigning parentage to progeny from two of the seedlots used to establish a silvicultural research trial, FR10/0. Trees were of rotation age and, due to their height and inaccessibility of the foliage, cambial tissue from bark windows was used as a source of DNA.

We have also demonstrated the recreation of a deceased parental genotypic profile using megagametophyte tissue (tissue that surrounds the embryo in a seed), a useful approach where candidate parents are no longer available. Genotypic profiles were generated for a total of 311 individuals, including 31 candidate parents and 280 progeny, using an exome capture probe panel for radiata pine. This panel delivered 93,160 SNP (single nucleotide polymorphism) markers; ten random subsets of 1000 SNP markers each were used to assign parentage using a statistical likelihood approach.

Introduction

One of the goals of workstream 2.1b of the GCFF programme is pursuing a phenotyping platform, created through the combination of genetics and remote sensing. Part of this vision is the recreation of pedigrees through DNA-based parentage assignments. This would then enable the evaluation of which genotypes have contributed towards final stand composition, and thus indicate which genotypes might be the best for certain sites. In many sites, only GF Plus ratings are available and only at the seedlot level, with individual genetic or parentage information unavailable, as is the case with FR10/0. We aimed to demonstrate that the pedigree in this silvicultural trial could be reconstructed using DNA marker-based methods.

What is fingerprinting?

Within a diploid forestry species, such as radiata pine, an individual carries two copies of the genome in every cell, and thus two copies of every DNA marker - one inherited from each of the parents. Different types of markers, such as simple sequence repeats (SSRs) or single nucleotide polymorphisms (SNPs) can be used to generate a profile (pattern) of markers in an individual, which is called a DNA fingerprint. The method only requires that there are sufficient markers to ensure these profiles are unique between individuals. In order to perform pedigree reconstruction in a population, unique DNA fingerprints must be generated for candidate parents and progeny. Thereafter, software algorithms look for matches between the profiles of the parents and progeny to determine who the most likely parents are for a particular individual.

This experiment

The trial - FR10/0

This silvicultural research trial, planted in 1987 at Glengarry to compare improved *Pinus radiata* breeds, was established from 4 main seedlots (Fig. 1). The first seedlot comprised known female parents fertilised with mixed pollen from known male parents (A), and the second was control-pollinated with known female and male parents (B). The third was open-pollinated, with known female parents but male parents unknown (C), and the fourth was climbing select with no parentage information available (D). In terms of demonstrating a proof of concept parentage reconstruction in FR10/0, the decision was made to focus on seedlots A and B for which we have a defined set of potential parents from which to choose,

supported by forestgrowers commodity levy

FOREST OWNERS ASSOCIATION

This programme is also funded by Scion and the Ministry of Business, Innovation and Employment, New Zealand.

and most of which already had foliage in storage at Scion for DNA.

A: GF21 Amberley '268' Seedlot (6/3/86/46) 9 cone (female) parents 1984 pollen mix (equal contributions from 21 pollen (male) parents) Total 26 unique parents – DNA for 25, recreated one missing parent from megagametophytes 143 progeny with cambium collected	**C: GF14 Gwavas '850' Seedlot (3/3/85/1)** 24 cone (female) parents, OP DNA for 6 female parents, 18 female parents not available. 279 progeny with cambium collected
B: 870 seedlot (9/3/86/166) 4 cone (female) parents, 4 pollen (male) parents Total 5 unique parents – DNA for all 137 progeny with cambium collected	**D: GF7 - climbing select** No parental information 140 progeny with cambium collected

Fig. 1: Seedlot information for FR10/0

Tissue collection

The trees comprising the stands of FR10/0 trees were 28 years old, making collection of needle tissue for DNA extraction logistically impractical. As such, bark windows (5 cm diameter), as shown in Fig. 2, were collected from these trees in March 2015, as the underlying cambial tissue provided an alternative tissue from which to extract DNA.

Bark windows were collected from 143 (seedlot A progeny) and 137 (seedlot B progeny) trees, frozen and stored at -20 °C at Scion.

Fig. 2: Example of a bark window

Needle tissue was used for obtaining DNA for the parents; for the one deceased parent with no material in storage, we attempted to recreate this genotype using seed from Scion's seed stores, from which maternal haploid megagametophyte tissue was sourced. Previous estimates (data not shown) have indicated that the combination of 12 mega-gametophytes should enable representation of all the alleles present in the original diploid maternal genome.

DNA extractions

Scion's standard DNA extraction method, using the NucleoSpin® Plant II (Machery-Nagel, Düren, GER) kit, had previously been demonstrated as suitable for cambium [1] and had been used to prepare DNA from FR10/0 as described in GCFF FN-01. However, any DNA that was of adequate concentration had been used up in testing the TruSeq Custom Amplicon (TSCA) (Illumina, Carlsbad, CA, USA) Parentage Panel, with the remaining samples being of poor yields. Due to ongoing technical challenges and delays associated with the TSCA platform, the decision was made to use the full exome capture probe panel developed and co-owned by Scion and the RPBC in the Genomic Selection programme. This required new DNA extractions to be performed for all samples.

DNA extractions were then performed as per the manufacturer's instructions, with the modifications as described in Telfer, Graham, Stanbra, Manley, and Wilcox (2013). Megagametophyte tissue was excised from each seed, 20 in total, and DNA extracted individually using the modified CTAB method [2]. The 12 samples with the best yields were selected for genotyping.

All DNA was frozen and stored at -20 °C until shipping in 96-well format, capped and vacuum-sealed, and on ice.

Genotyping

DNA samples were submitted to Rapid Genomics LL, in Gainesville, Florida, for genotyping by sequencing. Data were made available for download by Scion on 8 Feb 2018, and filtered according to standard pipelines developed in the Genomic Selection programme. For the parental genotype that was to be recreated from haploid megagametophytes, data were combined across the 12 individually genotyped samples. Any markers that were still showing as heterozygous in these individual haploid samples were removed from the entire dataset. If no alternative alleles were observed in any of the 12 samples for a particular marker, that marker was called as homozygous for the reference marker. Similarly, if no reference alleles were observed in any of the 12 samples, then that marker was called homozygous for the alternative allele. If there was at least one call for the alternative allele, a marker was called as heterozygous.

Parentage analysis

Parentage analysis was performed using CERVUS software (version 3.0.7) [3] which uses a statistical likelihood approach, and allows for the possibility of incomplete or mistyped genetic marker datasets and simulated sampling of the most likely parent.

Results

DNA was successfully isolated from the cambial tissue. Genotyping was successfully performed by Rapid Genomics using the exome capture panel, with marker profiles successfully generated for all 31 parents and 280 progeny. The filtered dataset comprised 93,160 SNP markers.

Parentage assignments using the "parent pair analysis" option in CERVUS were found to be unsatisfactory due to the low power of the analysis. The availability of some prior information as to which parents were used as females vs males, or if

controlled crosses were made, improved the confidence of the assignments. As such, parentage assignments for this study were performed using sequential maternity and paternity assignments, and are summarised in Tables 1 (seedlot A) and 2 (seedlot B). Details of the specific assignments per individual are available on request.

Table 1: Assignments per parent for seedlot A

Female parent	No. of assignments	Ave PLS
268_109	81	253.49
268_323	2	187.34
268_345	1	114.04
268_402	6	379.96
268_405	21	382.74
268_494	18	348.23
268_532	7	207.27
268_547	1	304.48
850_055	6	287.76
Male parent	**No. of assignments**	**Ave PLS**
268_007	1	46.11
268_041	3	51.67
268_107	4	14.32
268_118	0	29.98
268_131	4	-74.09
268_162	10	25.94
268_169	8	7.08
268_208	3	-109.6
268_232	17	32.73
268_249	3	-41.24
268_266	4	32.12
268_308	20	-74.31
268_345	12	-166.1
268_402	1	-191.79
268_455	12	27.92
268_494	1	46.27
268_514	4	35.53
268_530	13	-117.42
268_593	3	-41.93
268_609	9	9.60

For each assignment, CERVUS determined a Pair LOD (logarithm (base 10) of odds) Score (PLS), which is an indicator of how likely that the candidate selected by the software is the true parent. A positive LOD score suggests that the candidate parent is more likely to be the true parent, while a score of zero means that the candidate parent is equally likely to be the true parent or not the true parent. A negative LOD score means that the candidate parent is less likely to be the true parent than not the true parent. It should be noted, however, that negative LOD scores can also occur when the alleles shared between the candidate parent and offspring are very common in the population, which makes those alleles less useful at discriminating relationships. Another cause for low PLS values is due to mismatches between the candidate parent and the progeny at one or more loci, which can result from genotyping errors.

On average, maternal predictions were made with greater confidence (PLS = 287.28 in seedlot A, 4.40 in seedlot B) than paternal predictions (PLS = -27.09 in seedlot A, -92.15 in seedlot B).

Table 2: Assignments per parent for seedlot B

Female parent	No. of assignments	Ave PLS
870_533	43	6.09
870_580	64	5.11
870_609	12	-17.85
870_589	18	12.67
Male parent	**No. of assignments**	**Ave PLS**
870_580	18	-55.77
870_609	28	-199.39
870_589	61	-104.3
870_529	30	10.82

We also observed a number of predicted crosses in seedlot B that were not in alignment with the documented crosses (see Fig. 3). The PLS values for seedlot B were also lower than for seedlot A, even though we had a smaller number of parents from which to select. We noted that these parents are all long internode selections which represent around 10% of New Zealand's germplasm. It is possible that the lower likelihood scores reflect some hidden relatedness within these individuals which is compromising the ability of the software to make confident assignments. Employing a larger number of SNPs in the predictions could improve these scores. This was attempted with the full set of 93,160 SNPs, however, the memory requirements of CERVUS to process such a large dataset exceeded the computing capacity that was available for this study.

Female parent \ Male parent	870_580	870_609	870_589	870_529
870_533	**10**	4	18	11
870_580		**17**	32	15
870_609			**11**	1
870_589	8	7		**3**

Fig. 3: No. of progeny per predicted cross in seedlot B (bold = expected control-pollinated crosses as per seedlot records)

Lessons learned

This "proof of concept" study has shown that is possible to reconstruct the pedigree within a seedlot of known parents or known crosses, however, there are a number of observations and future suggestions that have arisen from this study.

Needle tissue remains the best material from which to extract DNA for radiata pine, both in terms of ease-of-use and DNA yields. Bark windows require a lot

more hands-on manipulation to prepare the cambial tissue for extraction, and trace fibres can clog the filtration columns used in the extraction process. This adds extra cost to the DNA extraction process, however, where no alternative is available, cambial DNA remains a viable option. One recommendation is that, where possible, bark windows are collected in the early growing season and preferably after good rainfalls, as this improves the thickness and softness of the cambial layer, facilitating the removal of the cambial layer and improving DNA yields.

For parents with no surviving ramets and needle tissue is no longer available, we showed that it was possible to recreate the genotype for such parents from megagametophyte tissue from their seeds (as maternal parent). This remains a last resort, however, as genotyping this parent costs 12 times as much as any other parent due to the number of samples that required testing. We would recommend that a small amount of tissue is banked for DNA extraction for as many parental genotypes as possible to mitigate this issue for future studies.

We had initially hoped to reconstruct parentage using an alternative platform, called Truseq Custom Amplicon Sequencing (TSCA), which interrogated 110 radiata pine SNPs. However, due to technical challenges, we were unable to successfully genotype FR10/0 progeny for this study using the TSCA platform. Exome capture was a more expensive approach but it was a proven platform, and delivered SNP data well in excess of the requirements for parentage assignments.

The assignment of parentage does rely on the accuracy of the candidate parental genotypes. In this study, we sampled from existing collections available in Scion's freezers, but there remains the possibility that the actual ramet used to produce the crosses contained in seedlots A and B are genetically different. Ideally, multiple copies of these parents would be genotyped to confirm their genetic identities, in combination with careful tracking of which exact ramet was used in a particular cross.

There are multiple approaches used for assigning parentage – these can employ either an exclusion approach, where potential parents are eliminated on the basis of mismatches with progeny and the "last parent standing" is assigned, or a likelihood approach, where the parent with the most similar profile is assigned based on likelihood ratios. Both have advantages and disadvantages that are not discussed further in this report. In this study, we have used the second approach of likelihood by applying the software package called CERVUS. Even within CERVUS, there are several approaches to assign parentage, depending on what level of prior information is available to the user. Including information such as known control-pollinated crosses, or even just which parents were used as females and which as males, can improve the robustness of the analysis and increased confidence in the parentage assignments.

Next steps

We propose that additional collections are made for some of the key parental genotypes to confirm the genetic identities of these trees. Furthermore, we recommend banking tissue from any parent that has been included as a female or male parent in deployed seedlots to ensure that DNA remains available for future testing. Some individuals will be permanently unavailable, and will require reconstruction from megagametophyte tissue. We also recommend the establishment of a parentage database in which genotypes for as many potential parents as possible are stored. While performing parentage assignments in germplasm with no known parentage such as climbing select, or with limited information such as open-pollinated crosses, is challenging due to the limited power of this approach, such a database could allow for the identification of relatives using a relationship matrix approach.

One of the questions originally motivating this study was the tracking of the GF rating of a stand from establishment to harvest. The rating for a seedlot is traditionally based on the performance and relative contribution of the parents used in the crosses. However, the impact of various natural and silvicultural factors over time could shift the relative representation of those parents over the life of the trial. Now that we have recreated the pedigree for the individuals sampled in this study, we will be exploring if there is any impact on GF ratings.

Acknowledgements

Funding for this research came from the "Growing Confidence in Forestry's Future" research programme (C04X1306), which is jointly funded by the Ministry of Business Information and Employment (MBIE) and the Forest Growers Levy Trust, with the support of the NZ Forest Owners Association (FOA) and the NZ Farm Forestry Association (FFA). Development of the exome capture probe panel was funded by the RPBC's MBIE Genomic Selection Partnership programme (RPBC1301).

References

1. Murray, M., *Extraction of genomic DNA from cambial bark windows*. 2015, University of Waikato student placement report.
2. Telfer, E.J., et al., *Extraction of high purity genomic DNA from pine for use in a high-throughput Genotyping Platform.* New Zealand Journal of Forestry Science, 2013. **43**(3).
3. Kalinowski, S.T., M.L. Taper, and T.C. Marshall, *Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment.* Molecular Ecology, 2007. **16**(5): p. 1099-1106.