Growing confidence in forestry's future Research Programme

SCION FORESTS ▪ PRODUCTS ▪ INNOVATION

Date: June 2019
Reference: GCFF TN-024

Technical Note

DNA fingerprinting to reconstruct parentage from trees in the forest – a second proof of concept

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

Corresponding author: natalie.graham@scionresearch.com

Summary: The ability to recreate parentage (pedigree reconstruction) has long been a goal of the New Zealand forestry industry and more recently, the GCFF programme. It opens up the possibility of using operational forests as our experiments, rather than relying exclusively on well-characterised genetic trials. New genotyping technologies and analysis software are now making this a reality for radiata pine. In this study, we aimed to recreate the parentage from DNA isolated from high performing trees. These were selected using characteristics derived from remotely sensed data as proxies for DBH and total stem volume. We compared predicted and documented pedigrees in individuals selected from a genetics trial, and showed substantial improvements in the parentage predictions obtained by using a new software package called Apparent. Using this approach, we have now successfully predicted parentage in high performing trees similarly identified using remotely sensed data from an operational stand (Kang1229). We can now begin to understand which parents within a seedlot are the ones contributing to high performing progeny on a site-by-site basis, or whether certain combinations of parents might outperform others. These outcomes therefore support the development of the phenotyping platform which will enable the implementation of precision forestry for New Zealand.

Introduction

Previous work under workstream 2.1b of the GCFF programme, towards creating a phenotyping platform through the combination of genetics and remote sensing, demonstrated the successful recreation of parentage using DNA in a research trial, FR10/0. Many operational stands have incomplete seedlot records, which is one of the more important variables to include in productivity models. DNA testing provides a means by which this information could be retrospectively determined. Furthermore, the ability to recreate pedigrees through DNA-based parentage assignments would allow the performance of seedlots to be teased apart further, by identifying which genotypes have contributed towards final stand composition. By identifying the best (or worst) performers, this would indicate which genotypes might be the best (or least recommended) for certain sites. In the previous study, we did not have documented pedigree information against which to compare the results that we obtained.

Therefore, in this second proof of concept study, we aimed to compare the results of DNA fingerprinting-based parentage predictions with documented pedigree information. Thereafter, we applied this technique to determine the parentage in a selection of best performing trees identified in an operational stand.

Methods

Trial 1 – FR260/1

This genetics trial was one of four trials established in 1995 from controlled crosses between high density parents. The trial comprised 30 families, planted out

supported by forestgrowers commodity levy

FOREST GROWERS RESEARCH

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

using a single tree plots, sets in rep design, with 30 replications per site. FR260/1 was established at a 1.44 ha site in compartment 1334 in Kaingaroa Forest.

Trial 2 – Kang 1229

This much larger site (35.5 ha) represented an operational planting stand in Kaingaroa Forest (compartment 1229) that fulfilled the following requirements: age >20 years, thinned, known seedlot (91/294), and with a limited number of genetic combinations for parentage assignments ("top 16" crossed with 850055).

Lidar data and selection of candidates

In FR260/1, 921 trees were phenotyped (height, DBH, total stem volume) using LIDAR (operational forest inventory data acquired in 2014), and 64 potential candidates were selected. In Kang 1229, 10,726 trees were similarly phenotyped and 170 candidates selected.

Field collections

Candidate trees were assigned GPS coordinates on a geo-referenced map and an efficient path to navigate between them determined. Individual trees were located using a Garmin handheld GPS and a tablet with internal GPS loaded with a georeferenced map.
Due to tree ages (>20 years), collection of needle tissue for DNA extraction was logistically impractical due to the height of the canopy. As such, bark windows (5 cm diameter) were collected using a bark hammer, as the underlying cambial tissue provided an alternative tissue from which to extract DNA [1].

Bark windows were collected in November 2018, frozen and stored at -20 °C until DNA could be extracted.

Where existing DNA profiles were not available for parents, needle tissue already in storage in Scion's freezers was used for obtaining DNA.

DNA extractions

DNA extractions were performed using ~100 mg of tissue and the NucleoSpin® Plant II (Machery-Nagel, Düren, GER) kit, as per the manufacturer's instructions, with the modifications as described in [2]. All DNA was frozen and stored at -20 °C until shipping to Rapid Genomics in Gainesville, Florida, on 10 Dec 2018 in 96-well plates, capped and vacuum-sealed, and on ice.

Genotyping

Rapid Genomics performed the exome capture genotyping by sequencing. Data were made available for download by Scion on 20 Feb 2019, and filtered according to standard pipelines developed in the Genomic Selection programme.

Parentage analysis

Parentage analysis was performed using both CERVUS software (version 3.0.7) [3] and a newly published package called Apparent [4]. This package examines all possible combinations of parents for each progeny, using markers that are homozygous in both parents to calculate a pairwise genetic distance (Gower's Dissimilarity coefficient (GD)) for each trio. A score of 0 indicates perfect identity, and 1 indicates perfect dissimilarity, therefore the lower the GD values, the more likely that the assigned parents are the true parents of an individual.

For CERVUS, the software determines a Pair LOD (logarithm (base 10) of odds) Score (PLS) for each assignment, 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.

Results

Tree sampling

When locating trees in the forest, some degree of error was noted when using GPS technology, particularly when under the canopy and when the collection crew were stationary. In such instances, other features such as canopy shape and proximity to canopy gaps were used to confirm that the correct tree had been located. Of the original selection candidates, 28 were sampled from FR260/1 and 160 were sampled from Kang 1229, with all sampled trees were processed for DNA extraction and genotyping. However, it was noted that many of these trees were not considered acceptable breeding candidates in terms of form issues (e.g., forking, sweep, multi-leader), with only 18 of 28 trees from FR260/1 and 65 of 160 trees from Kang 1229 deemed to have sufficiently good form.

Genotyping

Genotyping was successfully performed by Rapid Genomics on all candidates, including 4 parent samples for which profiles were not previously available. Profiles for most other candidate parents were sourced from previous genotyping experiments, but two were not available. The filtered dataset comprised 105,954 SNP markers, however, due to some of the parental genotypes originating from different genotyping experiments, this set had to be further refined to those that overlapped between all datasets (31,405 SNPs) and were thus informative in this study.

Comparison of CERVUS and Apparent

When investigating the documented pedigree information for the FR260/1 individuals, it was discovered that 12 individuals were either not in the trial or were controls and were thus excluded from further analysis. Of the remaining 16 individuals, 6 were affected by the unavailability of documented father genotype files. To compare the performance of CERVUS with Apparent, the refined set of SNPs was randomly reduced to 4000 to accommodate the computational limitations of CERVUS. Using these SNPs and the 16 remaining individuals, CERVUS was able to assign mother parents for 13 individuals (11 matched the documented pedigree) and fathers for 8 of the 10 individuals (7 matched the documented pedigree) that were expected to have an available father genotype file (see Appendix 1). Confidence scores ranged from -424.89 (highly unlikely yet it matched the documented pedigree) through to 227.94 (also matched the documented pedigree).

For Apparent, an assignment is always made for the trio with the best (lowest) GD score, however, these GD scores can vary (0.08 – 0.22 in this analysis). For the female parents, 12 of the 16 assignments matched the documented pedigree. Interestingly, 2 of the mismatches were a match for the CERVUS assignment, suggesting that perhaps the documented pedigree is not correct. The other 2 mismatches had no calls with CERVUS and had much worse GD scores, suggesting that perhaps the true parent is not among the candidates. It is worth noting that while we have examined the trios with the best GD scores, these are not generally reaching the recommended level of significance. This indicates that some further refinements are required, both in terms of genotyping accuracy (sequencing based methods are prone to missing heterozygotes, i.e., there are likely more false homozygote calls) combined with the fact that this package specifically targets SNPs that are allegedly homozygous in the parents. Furthermore, many of the parent genotype files have been sourced from earlier genotyping projects that were known to have a higher level of data quality issues and missing heterozygotes. In spite of this, Apparent still performed better than CERVUS at predicting parentage, with the added advantage of being able to use all available SNPs, and not just the 4,000 that we used for FR260/1 to enable a fair comparison between the packages.

Parentage in Kang1229

For Kang1229, we only used Apparent predictions for parentage and used the full filtered SNP set (~84k SNPs for this trial). This operational stand was established from a single seedlot that comprised 850055 (as both male and female parent) crossed with 16 other parents, therefore we would expect 850055 to be called as a parent for all samples from this trial. For all of the 160 progeny sampled, 231 different trio combinations were examined and the GD scores calculated, with the trio with the lowest GD score used to assign parents. Tree 850055 was predicted as the parent in 130 cases, with average GD scores of 0.049 (SD 0.034). In the 30 other instances where 850055 was not considered a parent, the parents assigned by Apparent were combinations of the other 16 candidate parents. In one example, progeny 7048 was assigned parents 268494 and 268109, with a GD score of 0.0176, which was the lowest (best) score obtained for the entire trial, and in fact the only trio that met the statistical significance requirements for true parentage. In general, however, the average GD scores in the cases where 850055 was not assigned as a parent were generally higher at 0.09 (SD 0.02) and were found to be statistically significantly different (*P-value* 2.31 x 10^{-13}, single factor ANOVA). While these GD scores were still quite low compared to what was observed for FR260/1, this could suggest that the true parent(s) were in fact not among the candidates. Performing parentage assignments with a wider pool of potential parental candidates could allow for the true parent(s) to be identified, however, this would rely on the availability of true parent profile(s) within an extensive database. We recommend the development of such a resource, in parallel with further investigation into determining the thresholds below which parentage can be confidently assigned. As significance levels are not being reached in most instances, it is important to know the GD threshold above which we can be sure that the true parent is not in fact among the candidates.

Within the assignments made for Kang1229, it was interesting to note the high representation of 268054 (Fig 1), both in the full progeny set and within the subset of progeny that possessed suitable form characteristics to warrant selection.

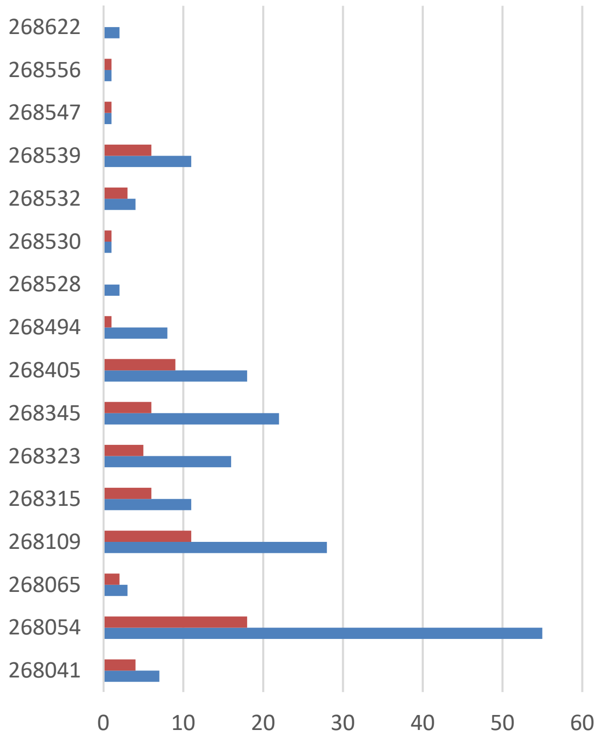

Figure 1: Number of progeny assigned to each of the candidate parents (excluding 850055) in Kang1229 for the full set of 160 progeny (blue bars) and the subset of 65 progeny that had suitable form (red bars).

Conversely, parents 268622, 268556, 268547, 268530, 268528 and 268065 were underrepresented, with 268622 and 268528 not represented at all in the 65 progeny that had sufficiently good form. Some of these trends could be explained by the relative contribution of these parents to the seedlot mix in terms of number of cones sampled. However, some parents do appear to deviate from their expected contribution. For example, progeny from the cross between 850055x268323 were below what might be expected, and progeny from 850055268345 were above what might be expected, based on the number of cones sampled. However, there are several other factors that could explain these variances which should be explored before any solid conclusions can be drawn.

Conclusions and recommendations

In this study we have shown an improved performance in parentage predictions using a new package called Apparent, as demonstrated in a genetics trial (FR260/1), and subsequently applied to an operational stand (Kang1229). This package can use all available SNPs and outperforms the previously used CERVUS, which was limited to 4000 SNPs or less. Any parentage prediction remains a statistical probability, based on the available data. Currently, we are limited to assessing the performance of these packages relative to the documented "gold standard" pedigrees, which are also highly likely to contain errors. In many instances, candidate parent profiles used in this study were generated in earlier genotyping experiments, known to have a higher level of error; several improvements have since been made to this platform and newer datasets have reduced amounts of genotyping error. This discrepancy in error rates, and the approach of Apparent of specifically using SNPS that are homozygous in the parents could negatively impact our results. Therefore, we expect further improvements in parentage assignments that correspond with expected improvements data quality as we move to using the new radiata pine SNP array, and strongly recommend the development of a database of potential parents using this more robust and more affordable genotyping technology. Apparent will also select a parent from the pool of candidates that have been supplied, although the likelihood of these being true can be gauged by the GD scores. Understanding how these GD scores are impacted by genotyping error, and the threshold above which we can confidently assume that the true parent is not within the pool of candidates, is a further recommendation of this study.

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. Graham, N., A. Ismael, and T. Stovold, *Reconstructing parentage in Pinus radiata using exome capture genotyping proof of concept in FR10/0*. 2018.
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.
4. Melo, A.T.O. and I. Hale, *'apparent': a simple and flexible R package for accurate SNP-based parentage analysis in the absence of guiding information.* BMC Bioinformatics, 2019. **20**(1): p. 108.

Appendix 1: Comparing documented and assigned pedigrees in FR260/1 using CERVUS and Apparent software

	Documented		**CERVUS**				**Apparent**			
Tree ID	**female parent**	**male parent**	**female parent**	**PLS**	**male parent**	**PLS**	**female parent**	**male parent**	**No. SNPs**	**GD**
1006	880770	880730	880770	-424.89	880730	93.43	880770	880730	1475	0.08
1008	875293	268609	875293	-30.8	268609	2.72	875293	268609	743	0.08
1014	268556	875257	268556	132.55	875257	227.94	268556	875257	1465	0.04
1015	875954	268494	🟥		268494	146.36	875954	268494	1019	0.17
1026	875293	268609	880730	100.99	880770	-322.22	880730	880770	1490	0.08
1035	880733	880732	875255	-67.4	🟥		875255	880732	1572	0.11
1044	268288	875255	268288	-89.95	🟥		268288	875257	1683	0.17
1047	880770	880730	880770	-362.15	880730	79.52	880770	880730	1493	0.09
1057	875954	268494	🟥		268494	167.41	875293	268494	1053	0.18
1061	880770	880730	880770	-321.85	880730	129.43	880770	880730	1495	0.08
1004	268041	268429	268041	166.34	🟦		268041	880730	1506	0.14
1007	268041	268429	268041	151.11	🟦		268041	268556	1443	0.14
1028	268041	268429	268041	166.35	🟦		268041	875293	1458	0.14
1058	268041	268429	268041	171.75	🟦		268041	268556	1457	0.14
1062	268041	268429	268041	121.99	🟦		268041	880730	1500	0.14
1063	268228	875294	🟥		🟦		875293	880730	1529	0.22

PLS – probability score (>0 = more likely)

GD – genetic dissimilarity (closer to zero = more likely)

Colour code:

- Green – match to documented pedigree
- Blue – genotype file for documented parent not available, noting that Apparent will select the next best match available
- Orange – mismatch to documented pedigree but match between Cervus and Apparent
- Red – no assignment or mismatched to documented pedigree