Challenges associated with adapting genomics technologies to forestry

E. J. Telfer [1], N. J. Graham[1] and J. Klápště[1]

[1] Scion, Forest Genetics, 49 Sala St, Rotorua 3010, New Zealand
emily.telfer@scionresearch.com (Corresponding Author)

Summary

The emergence of the polymerase chain reaction, Sanger DNA sequencing, and microarrays, to the latest in massively parallel sequencing of RNA and DNA and single nucleotide polymorphism (SNP) arrays, have revolutionised the field of Genetics. However these technologies have in almost all instances been developed and optimised for Human and mammalian models. Despite commercial providers assurance that new technologies are species (or Kingdom!) neutral, frequently the adaption of these technologies for plants has taken considerable optimisation. In the case of large, pre-reference genome conifer species, our journey from cutting edge technology to practical application has been a challenge.

Keywords: Megagenome, Next Generation Genomics, Forestry, Genotyping

Introduction

Molecular biotechnology has changed our understanding of life on earth. At a sub-cellular level, we can begin to reveal the processes by which the genetic DNA code is translated into the phenotypes we observe. Once we began reading the sequence of DNA and targeting the amplification of very specific regions, the progress from unlocking molecular details of retroviruses, bacteria, and simple eukaryotes, moved incredibly fast. Understandably, the species we find most fascinating to explore the workings of is our own and, in 2004, the first entire Human genome was revealed to the world. Similarly, many of the genotyping technology platforms available today have been developed and optimised with human medical discovery in mind. Although the first tree genome (*Populus trichocarpa*) appeared just two years later (Tuskan, et al., 2006), of the 26,390 species genomes listed in the last 21 years on the National Center for Biotechnology Information website ("Database Resources of the National Center for Biotechnology Information," 2017), less than 1% (238) are those of land plants. In particular, the large complex genomes of conifers (De La Torre, et al., 2014) have proven a challenge to assemble. As a consequence, many tools and resources available to agricultural species in the post-genome world remain out of reach for forest tree breeding programmes. This remains a frustration given the considerable impact that genomic selection could have by shortening the delivery time for improved genetics through earlier predictions of genetic potential in forest trees (Grattapaglia, et al., 2011; Isik, 2014).

Extracting suitable quantities of high quality DNA from conifers and several other forest trees remains a challenge, and finding a “one size fits all” extraction method for any species or tissue type, that can also be utilized in any genotyping platform, is unlikely. Successful methods for one species or tissue type or genotyping platform don’t necessarily transfer to other species or tissue types or genotyping platforms. Some platforms require extremely pure DNA, and fragment size is less of an issue, e.g. MALDI-TOF (Jurinke, et al., 2004). Polymerase chain reaction (PCR)-based protocols with single locus targets can be relatively forgiving in terms of DNA quality. This tolerance has allowed for the development of many field-diagnostic assays (Schaad, et al., 2002; Tomlinson, et al., 2005), where relatively crude DNA extractions are still

usable for single-target PCR or loop mediated isothermal amplifications (LAMP) (Kiddle, et al., 2012). Illumina or Affymetrix SNP arrays are surprisingly robust with respect to both DNA purity and fragment size. Other methods require large quantities of DNA, and may have additional purity or molecular weight constraints, such as PacBio sequencing (Korlach, et al., 2010) and restriction genotype-by-sequencing (GBS) (Elshire, et al., 2011). Furthermore, the impact of cost per extraction both in terms of consumables and labour should not be dismissed, particularly where large population-wide experiments are required and small additional costs per sample rapidly compound. We present three examples of genomic technologies that have been adapted for application in forestry.

Case studies in technology adoption for forestry

Mass-array genotyping

With the pursuit of medium to high-throughput genotyping, which enabled the interrogation of multiple targets per reaction, the impact of trace quantities of inhibitors have become an important consideration (Bayés, et al., 2011). Many plants including conifers are known to contain high levels of secondary metabolites, and the complete removal of these during DNA extraction can be difficult. CTAB-based methods (adapted from Cato, et al. (1996) have delivered relatively good DNA yields for *Pinus radiata*, however, high yielding methods can be potentially less discriminating in terms of co-precipitating unwanted co-extractives. Our first attempts at using the Sequenom (now Agena) iPLEX® Gold MassARRAY® gave poor call rates when using DNA extracted with this method. More than 20% of samples failed to deliver a genotype, and 37% of the markers departed from Hardy-Weinberg equilibrium due to incomplete detection of heterozygous alleles. Multiple extraction and purification combinations were subsequently trialed (Telfer, et al., 2013) and quickly demonstrated that a simple additional ethanol purification could reduce the fail rate to 6.46%. Once sufficient improvements were achieved to reduce genotyping failure rates, we were able to perform successful association studies (Li, et al. 2016). Taking scalability, cost and ease of use into consideration for future studies, we have now adapted the Nucleospin® Plant II kit (Machery-Nagel, Düren, GER) for most species and tissues extracted in our facility.

Genomic selection in a pre-genome world

Genomic selection as proposed by Meuwissen, et al. (2001) promises to deliver early selection of elite individuals using DNA markers to predict traits. Application requires a sufficiently large set of genome-wide markers at a density that ensures at least some of the markers will be in linkage disequilibrium with genes controlling the trait of interest (Grattapaglia, et al., 2011). SNP markers are widely abundant throughout genomes and can be discovered via DNA resequencing. Ideally, this is done with whole genome resequencing of an appropriate reference population (Faivre-Rampant, et al., 2016; Silva-Junior, et al., 2015), and mapping against a detailed reference genome (Myburg, et al., 2014; Tuskan, et al., 2006) to detect and order SNPs across the genome and assess their frequency within the population. However, whole genome sequencing and resequencing is often cost-prohibitive in large conifer megagenomes (Birol, et al., 2013; Nystedt, et al., 2013; Stevens, et al., 2016; Zimin, et al., 2014).

To overcome the barriers of large genomes, SNP discovery can be carried out within the expressed portion of the genome only, specifically the exome. However, this presents an additional challenge in conifers. It is estimated that between 200 and 300 million years ago, several whole genome duplication events occurred in the gymnosperm lineage (Li, et al., 2015), resulting in large families of paralogous genes. Fifty thousand genes models were predicted in *P. taeda*, yet they represent only 20,000 gene families (Wegrzyn, et al., 2014). Using a *P. radiata*

transcriptome, we developed a set of 44,366 exome capture probes using the method previously described for *P. taeda* (Neves, et al., 2013). Resulting genotypic data were filtered to eliminate probes that appeared to capture a composite of multiple paralogous genes, and not just a single locus. However, this filtering initially removed over a quarter of the capture probes as paralogous genes. Reordering the filters to account for data quality first, rather than elimination of probes with low quality "pseudo-heterozygous" genotypes in a haploid control samples, vastly improved the SNP discovery rate (Table 1).

Table 1. Impact of filtering order on probe retention

Original filter		Revised filter	
Filter	Number of probes remaining	Filter	Number of probes remaining
None	44,378	None	44,378
Polymorphic haploids	33,833	Read depth per SNP	44,367
Quality Score		Biallelic	44,363
Read depth per SNP	33,817	Read depth/ individual	44,363
Biallelic		Allele ratio	44,363
Allele ratio	33,783	Polymorphic haploids	41,768

Restriction enzyme-mediated genotype-by-sequencing

An alternative method for reduced representation GBS utilizes restriction enzyme cleavage sites (Elshire, et al., 2011). This method requires intact high molecular weight DNA so that markers detected are a result of the underlying genomic sequence, rather than due to random shearing during extraction. Low molecular weight DNA can result in inconsistent genotypes with lower reproducibility. The utilisation of GBS as a cost-effective alternative to SNP arrays or exome capture requires a high degree of fidelity between parent and progeny genotypes for analysis such as pedigree reconstruction or genomic selection to be possible. We compared the reproducibility of technical replicates of *Eucalyptus nitens* samples across multiple marker panels: microsatellites (simple sequence repeats or SSRs), the EUChip60K (Telfer, et al., 2015) and restriction GBS (Table 2). We have identified that GBS is suitable for one-off SNP discovery or diversity studies, but not for applications that require comparisons to be made over time.

Table 2. Reproducibility in technical replicates across genotyping platforms

Genotyping assay	Mistyping rate between replicates
Scion 13-plex SSRs	10%
EMBRAPA 15-plex SSRs	3%
EUChip60K	0.007%
Restriction GBS	34%

Conclusions

A clear understanding of the DNA requirements and performance metrics of each platform and the biological parameters of the species of interest has been key to adapting these genotyping technologies for application in forestry. Those limitations need to be understood by both the researcher and more importantly, the service provider performing the genotyping. In our

experience, most platforms tested to date have required considerable iterative optimisations to adapt these platforms to forestry species, an essential step to deliver cost-effective services for operational implementation. However, with tenacity, most genomics methods can be adapted for plants including conifers.

List of References

Bayés, M., & Gut, I. G. (2011). Overview of genotyping. Molecular analysis and genome discovery: John Wiley & Sons, Ltd, 1-23.

Birol, I., Raymond, A., Jackman, S. D., Pleasance, S., Coope, R., *et al*. (2013). Assembling the 20 Gb white spruce (Picea glauca) genome from whole-genome shotgun sequencing data. Bioinformatics, 29(12): 1492-7.

Cato, S. A., & Richardson, T. E. (1996). Inter- and intra-specific polymorphism at chloroplast SSR loci and the inheritance of plastids in *Pinus radiata* D.Don. Theoretical and Applied Genetics, 93: 587-592.

Database Resources of the National Center for Biotechnology Information. (2017). Nucleic Acids Res, 45(D1): D12-d17.

De La Torre, A. R., Birol, I., Bousquet, J., Ingvarsson, P. K., Jansson, S., *et al*. (2014). Insights into Conifer Giga-Genomes. Plant Physiology, 166(4): 1724–1732.

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., *et al*. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE, 6(5): e19379.

Faivre-Rampant, P., Zaina, G., Jorge, V., Giacomello, S., Segura, V., *et al*. (2016). New resources for genetic studies in Populus nigra: genome-wide SNP discovery and development of a 12k Infinium array. Molecular Ecology Resources, 16(4): 1023-1036.

Grattapaglia, D., & Resende, M. D. V. (2011). Genomic selection in forest tree breeding. Tree Genetics and Genomes, 7(2): 241-255.

Isik, F. (2014). Genomic selection in forest tree breeding: The concept and an outlook to the future. New Forests, 45(3): 379-401.

Jurinke, C., Oeth, P., & van den Boom, D. (2004). MALDI-TOF Mass Spectrometry. Molecular Biotechnology, 26: 147 - 163.

Kiddle, G., Hardinge, P., Buttigieg, N., Gandelman, O., Pereira, C., *et al*. (2012). GMO detection using a bioluminescent real time reporter (BART) of loop mediated isothermal amplification (LAMP) suitable for field use. BMC Biotechnology, 12(1): 15.

Li, Y., Wilcox, P., Telfer, E., Graham, N., Stanbra, L., 2016. Association of single nucleotide polymorphisms with form traits in three New Zealand populations of radiata pine in the presence of genotype by environment interactions. Tree Genetics & Genomes 12:63.

Li, Z., Baniaga, A. E., Sessa, E. B., Scascitelli, M., Graham, S. W., *et al*. (2015). Early genome duplications in conifers and other seed plants. Science Advances, 1(10): e1501084.

Meuwissen, T. H. E., Hayes, B. J., & Goddard, M. E. (2001). Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics, 157(4): 1819-1829.

Myburg, A. A., Grattapaglia, D., Tuskan, G. A., Hellsten, U., Hayes, R. D., et al. (2014). The genome of Eucalyptus grandis. Nature, 510: 356-362.

Neves, L. G., Davis, J. M., Barbazuk, W. B., & Kirst, M. (2013). Whole-exome targeted sequencing of the uncharacterized pine genome. Plant Journal, 75(1): 146-156.

Nystedt, B., Street, N. R., Wetterbom, A., Zuccolo, A., Lin, Y.-C., *et al*. (2013). The Norway spruce genome sequence and conifer genome evolution. Nature, 497: 579-584.

Schaad, N. W., Opgenorth, D., & Gaush, P. (2002). Real-Time Polymerase Chain Reaction for One-Hour On-Site Diagnosis of Pierce's Disease of Grape in Early Season Asymptomatic

Vines. Phytopathology, 92(7): 721-728.

Silva-Junior, O. B., Faria, D. A., & Grattapaglia, D. (2015). A flexible multi-species genome-wide 60K SNP chip developed from pooled resequencing of 240 Eucalyptus tree genomes across 12 species. New Phytologist, 206: 1527-1540.

Stevens, K. A., Wegrzyn, J. L., Zimin, A., Puiu, D., Crepeau, M., *et al*. (2016). Sequence of the Sugar Pine Megagenome. Genetics, 204(4): 1613-1626.

Telfer, E. J., Graham, N., Stanbra, L. K., Manley, T., & Wilcox, P. L. (2013). Extraction of high purity genomic DNA from pine for use in a high-throughput Genotyping Platform. New Zealand Journal of Forestry Science, 43(3).

Telfer, E. J., Stovold, G. T., Li, Y., Silva-Junior, O. B., Grattapaglia, D. G., & Dungey, H. S. (2015). Parentage reconstruction in Eucalyptus nitens using SNPs and microsatellite markers: a comparative analysis of marker data power and robustness. PLoS ONE, 10(7): e0130601.

Tomlinson, J., Boonham, N., Hughes, K., Griffin, R., & Barker, I. (2005). On-site DNA extraction and real-time PCR for detection of Phytophthora ramorum in the field. Applied and Environmental Microbiology, 71(11): 6702-6710.Tuskan, G. A., DiFazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., *et al*. (2006). The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). Science, 313(5793): 1596-1604.

Wegrzyn, J. L., Liechty, J. D., Stevens, K. A., Wu, L. S., Loopstra, C. A., *et al*. (2014). Unique features of the loblolly pine (Pinus taeda L.) megagenome revealed through sequence annotation. Genetics, 196(3): 891-909.

Zimin, A., Stevens, K. A., Crepeau, M. W., Holtz-Morris, A., Koriabine, M., *et al*. (2014). Sequencing and assembly of the 22-Gb loblolly pine genome. Genetics, 196(3): 875-890.