New Zealand forestry enters the genomics era – towards a genome sequence of radiata pine

Phillip L. Wilcox and Lucy J. Macdonald

Abstract

The number of species for which an entire genetic code or genome has been sequenced is rising steeply. The first sequence of the human genome was announced in June 2000, 10 years after the project started. Since then genome sequencing has become both increasingly affordable, with the advent of cheap high-throughput DNA sequencing, and more technically feasible with advances in data analyses. Significant efforts have been made worldwide to sequence the genomes of an ever-increasing number of organisms including the mouse, chimpanzee, rice, maize and many others (Ellegren, 2014).

Once the genetic code of an organism has been deciphered, the way the knowledge is used can have profound effects. The human genome sequence has accelerated the development of a multitude of medical diagnostics and treatments, as well as providing new insights into human biology, disease, evolution and culture (Koboldt et al., 2013; Hood & Rowen, 2013). Black cottonwood (*Populus trichocarpa* Torr. and Gray) was the first forest tree species to be sequenced (Tuskan et al., 2006). More recently, the first genome sequences have been published of the following conifers: Norway spruce (*Picea abies* (L.) H. Karst), white spruce (*Picea glauca* Moench) (Birol et al., 2013; Nystedt et al., 2013) and loblolly pine (*Pinus taeda* L.) (Neale et al., 2014). The *Eucalytus grandis* W. Hill genome, another non-conifer species, has also been recently released (Myburg et al., 2014).

Scion, with assistance from other organisations, has begun sequencing the huge megagenome of *Pinus radiata* D. Don, New Zealand's most commercially important tree species. The first working draft assembly will be completed next year. As with all other large genome assemblies, this first working draft will only be a partially complete genome and is the start of a longer endeavour. Due to the nature and complexity of large genomes in general, and conifer genomes in particular, to complete and fully understand the genome will require input from a larger international community of researchers. Nonetheless, a partially completed genome, when combined with other resources, will be a powerful tool to inform other research. This paper discusses the basic concepts of genome sequencing, considers how the radiata pine genome sequence could be used, and the possible consequences for the New Zealand forest industry.

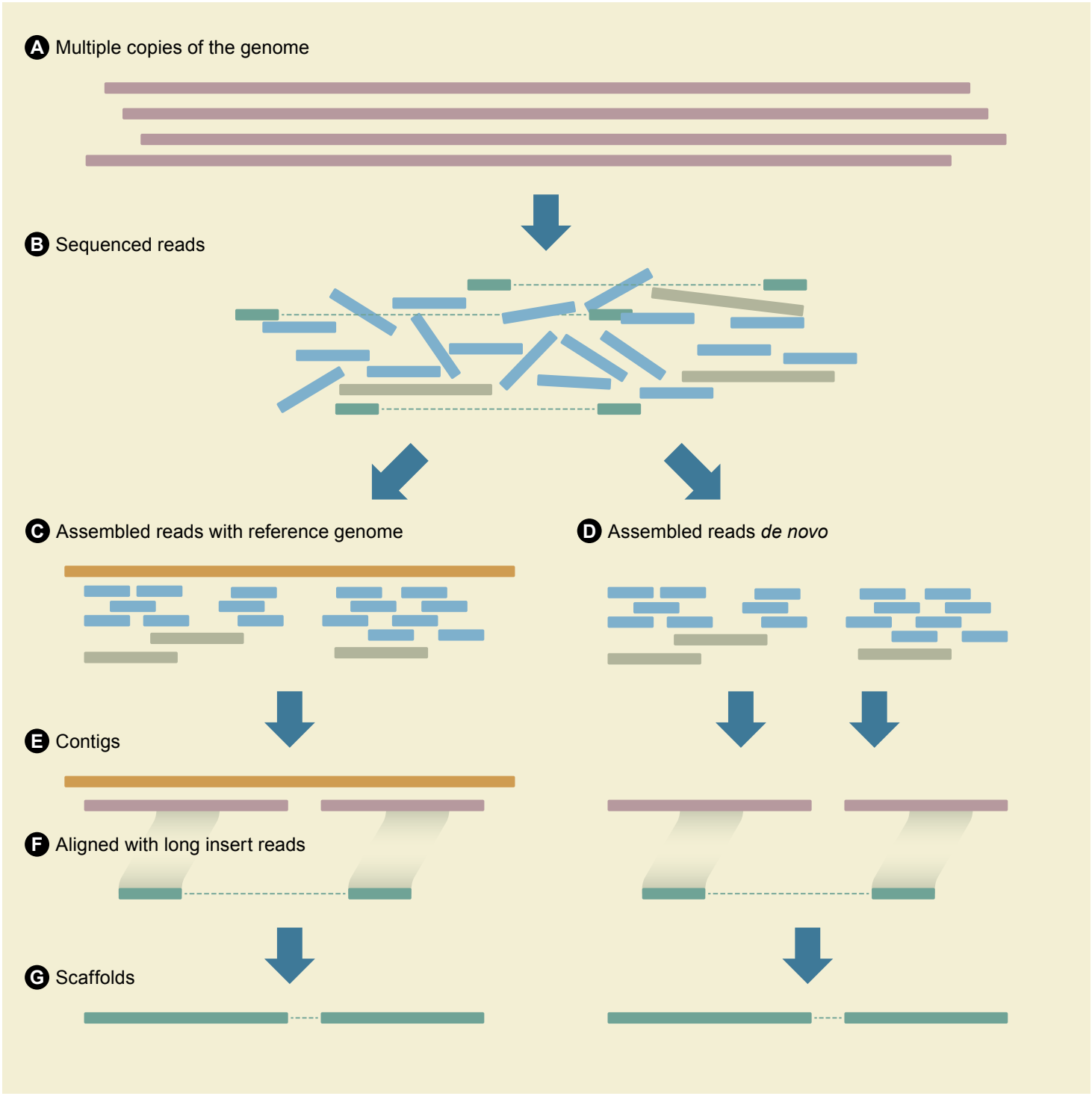
Generating whole genome assemblies

The first step is the extraction of DNA from the target organism (Box 1A). Practically any part of the organism can be used as a source since the genome sequence of an individual is virtually identical in every living tissue. Because there are no technologies that sequence whole chromosomes from end to end, DNA needs to be snipped into a multitude of short fragments that are simultaneously read by a DNA sequencing machine (Box 1B). Current high-throughput DNA sequencing technologies can generate millions of short sequences of DNA referred to as 'reads'. The reads range in size from a few tens of base pairs (bp) – letters of DNA code – to a few thousand bp. Different technologies vary in the length of fragment read and the number of reads performed per sequencing cycle. Each short read is a tiny fraction of the entire genome; a 100 bp read is only about a four billionth of the radiata pine genome.

The millions of short reads produced are then assembled into the genome using either a reference-based assembly (Box 1C), where the reads are aligned to a pre-existing reference genome from the same or a closely-related species, or a *de novo* assembly (Box 1D), where the reads are overlapped with each other in the most likely order. Assemblies of large genomes using *de novo* methods require considerable computing power and typically utilise super-computing resources with large numbers of CPUs and terabytes of RAM. The recent *de novo* assembly of the loblolly pine genome sequence, for example, used a 64-core super-computer with one terabyte of RAM and required three months for assembly (Zimin et al., 2014).

The typical results of both the reference-based and *de novo* assembly methods are a series of contiguous sequences, or contigs (Box 1E), which range from a few hundred to tens of thousands of bp. Longer read lengths produce longer contigs with more confidence since more of the reads will overlap. Contigs can then be further clustered into scaffolds, often using additional long insert reads that bridge gaps between contigs (Box 1F-G). However, the short reads are much cheaper to sequence and the adverse effects of such short length reads can be mitigated by repeatedly sequencing the genome. Researchers often combine the benefits of both short and long reads by sequencing genomes using multiple technologies.

The assembly process requires iterations of data generation to develop a near complete sequence. The sequence is near complete as some gaps in the DNA sequence always occur, which may be due to the stochastic nature of sequencing resulting in some regions being under-represented by chance. The chemical composition of some regions also renders them difficult to sequence with existing technologies. The most



thoroughly sequenced eukaryote genome – the human genome – is still incomplete (Chaisson et al., 2015) because the complex chemical structures of some regions hinder access to existing sequencing methods and/or they create difficulties for assembly. The random under-representation of specific regions can be compensated for by generating 40 to 100 times the total DNA content of the genome to ensure as much coverage as possible, and by using multiple sequencing technologies involving a combination of short and longer fragments. Even so, assemblies are usually imperfect, although they typically improve as more data become available.

Making sense of a genome

When an assembly is judged to be sufficiently complete, usually referred to as a working draft, the genome is annotated. There are two types of annotation: physical and functional. Physical annotation includes locating features such as repeat sequences, for example, long terminal repeat (LTR) retrotransposons, gene copy number variants, small tandemly repeated sequences (also known as microsatellites), and methylated regions (where they have been identified) within the entire assembled sequence. Regulatory elements, which influence the manner in which genes are expressed, are also identified.

Box 1: Assembling a genome

Key steps in genome assembly:

A. DNA is extracted from the target organism. The sample will contain multiple copies of the genome from the individual being sequenced.

B. The genomes are fragmented into millions of reads which are sequenced simultaneously. The genomes can be sequenced in different ways depending on the technology used:
   1. For short reads (blue) the genome is fragmented into uniform sized pieces called reads that can range in size from 25–300 bp.
   2. Long reads (grey) are typically between 3,000–20,000 bp. This is new technology and not routinely used yet.
   3. With long insert reads (green) fragments between 250–20,000 bp are produced but only the 100 or so bp at each end is sequenced.

C. For reference-based assemblies the reads are aligned to the reference genome (brown).

D. For *de novo* assemblies, where there is no reference genome, the reads are aligned against each other and overlapped. This is the most computationally intensive step. The more copies of the genome that are sequenced the more likely there will be overlapping reads.

E. These overlapping reads are the consensus sequences known as 'contigs' (contiguous sequence).

F. Reads from long insert DNA fragments (B.3) are used to concatenate contigs; if one end aligns to a contig and the other end aligns to another contig we know these contigs are adjacent to each other and the distance between them, even if we do not have all of the DNA sequence between the contigs.

G. Contigs joined in this way become known as 'scaffolds'. There will still be gaps in the genome that are not covered by the scaffolds, so there will be multiple scaffolds per chromosome. For example, the current working draft of the loblolly pine genome has 14 million scaffolds even though there are only 12 chromosomes.

Functional annotation is the assignment of presumed biochemical function(s) to specific genes, including their role in biological processes, as well as in the regulation of gene expression. Genes can be found either by searching the entire genome for genes or gene-like elements using a publicly available database of genes (*ab initio*), or by using our own data from prior research where previously identified gene sequences can be located on the genome. Functional information often comes from other species for which the biological functions of homologous, i.e. the same, genes have been determined previously and subsequently extrapolated on the basis of sequence similarity. Gene sequences are generally more similar between closely-related species. In conifers, the order of genes on the genome is also highly conserved. Moreover, genes often occur in families of the same or similar sequences (Morgante & De Paoli, 2011).

The final assembly containing the working draft of the genome sequence and its associated annotations is usually visualised in a genome browser. Visualisation tools are typically web-based: examples include the various genome viewers at http://genome.ucsc.edu (see Box 2), the human HapMap project http://hapmap.ncbi.nlm.nih.gov/, which includes all known variations in DNA sequence in humans, and the recently released Norway spruce and loblolly pine genome viewers (see http://congenie.org/gbrowse). As continual improvements are made to genomes each new working draft assembly, along with enhanced annotations, is released via these genome browsers.

Characteristics of conifer genomes

Radiata pine and other conifers have huge genomes with extensive repeat content that make whole genome sequencing a formidable task. With approximately 25 billion base pairs (Murray, 1998) the nuclear genome of radiata pine is over eight times the size of the human genome. It is one of the largest genomes in the Pinaceae, and amongst the largest of all gymnosperm genomes. In contrast, the genomes of hardwoods such as eucalypts and poplars, which contain fewer than one billion base pairs, are just 4% of the size of the radiata pine genome.

The nuclear DNA of radiata pine is packaged into 12 chromosomes, each with virtually identical physical appearance, such that they are very difficult to differentiate one from another via conventional microscopy. Genes in the nuclear genome are essentially small islands scattered amongst large tracts of repeat sequences. Over 80% of pine nuclear DNA is repeat sequence, consisting of material ranging from long strings of the same base pair to serially repeated elements, each consisting of thousands of base pairs. The longer repeats are particularly challenging to analytically assemble in their correct order due to the difficulty of assigning short fragments with near identical sequence to their appropriate copy with certainty. The nuclear genome is complemented by a small chloroplast genome (approximately 120,000 bp) and a mitochondria genome (approximately 1.3 million bp).

Variation in DNA sequences among trees also creates challenges for assembly. Pines are diploid organisms, inheriting one copy of each chromosome from the pollen and cone parents. As with humans, the sequence of the DNA inherited from the male parent slightly differs from that of the DNA inherited from the female parent, leading to 'polymorphism' within an individual tree. Therefore, within the same tree, there are many differences in DNA sequence for the same chromosome, creating further

Box 2: An example of a genome browser

Genome browsers are used to display all information related to areas of the genome sequences. This is a snapshot of the latest human genome (Feb. 2009 GRCh37/hg19) assembly from the University of California, Santa Cruz website (http://genome.ucsc.edu/, Kent et al., 2002):

A. The diagram represents human chromosome 21. The red line depicts the specific region being viewed below which is in a region known as q22.11.

B. The top line represents the scale of 2,000 bp of sequence and the figures below indicate the position on the chromosome, i.e., 33,033,000–33,042,000 bp from the start of the chromosome (full length of region is approx. 9,900 bp).

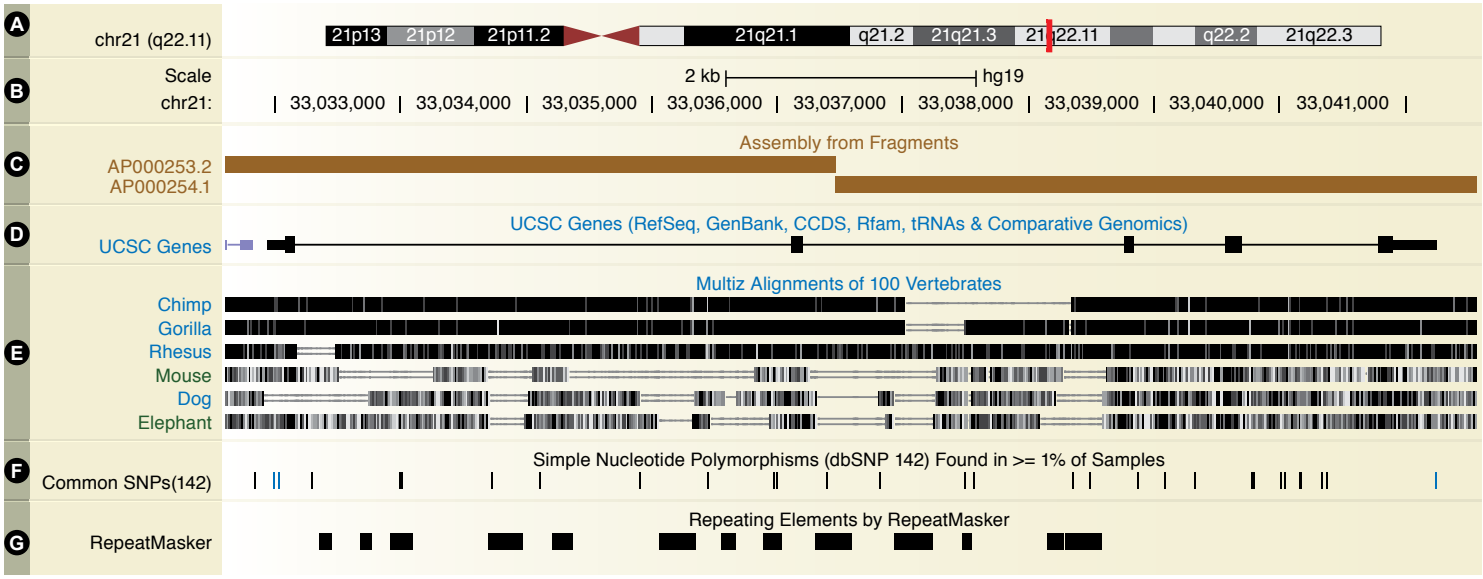
C. The two brown bars correspond to two scaffolds: AP000253.2, AP000254.1.

D. Each box shows the position of predicted genes within this region of the chromosome. These genes have been identified in various public databases.

E. Similarity between human DNA sequence in this region and other mammalian genomes (chimp, gorilla, rhesus monkey, mouse, dog and elephant). Regions in black show high levels of similarity with the human genome, regions in grey are moderately similar, and horizontal lines correspond to regions that are present only in the human genome.

F. Vertical black lines correspond to positions where DNA sequence variations (single nucleotide polymorphisms or SNPs) have been found within this region.

G. Black boxes show the positions of repeated DNA sequence. These will also be present elsewhere in the genome.



challenges for assembly. These features require novel and innovative approaches to data acquisition and analyses. To avoid these issues, researchers sequencing conifers sometimes use DNA from megagametophytes, the tissue surrounding the embryo in a pine seed, which is haploid and identical to the DNA inherited from the female parent (Neale et al., 2014).

The radiata pine genome sequencing project

Scion is assembling the first working draft of a whole genome of radiata pine. The individual tree being sequenced, known as 268.345, is an important parent in the Radiata Pine Breeding Company (RPBC) breeding population. The actual sequencing has been carried out by New Zealand Genomics Ltd, which has generated approximately eight billion short reads, equivalent to 36.3 times the total DNA content of radiata pine.

A reference-based assembly of the nuclear portion of the radiata pine genome is being used due to the comparative high cost of alternatively generating sufficient data for a *de novo* sequence and the availability of a closely-related genome, loblolly pine. We are therefore using the loblolly pine genome sequence as a template. Previous research by Scion scientists has indicated a very high degree of sequence homology between genes of the two species. Scion is receiving advice on assembling the genome from Professor David Neale at the University of California, Davis and Professor Steven Salzberg at Johns Hopkins University who, between them, have led the United States-funded loblolly pine genome assembly. The huge size of the radiata pine genome means that this is the largest reference-based assembly ever attempted to date. Only the *de novo* assembly of the 38 billion base pair sugar pine (*P. lambertiana* Douglas) genome currently underway is bigger (Murray, 1998). This project is being carried out by the Pine Reference Sequence consortium funded by the US Department of Agriculture.

We anticipate that the first radiata pine assembly, with limited annotation and a genome browser, will be

completed by early 2016. It is likely that the first working draft will be very similar to the loblolly pine assembly, which consists of over 14 million scaffolds (Zimin et al., 2014). Owing to the sheer scale of this endeavour, Scion intends to release the sequence publicly so that it can be improved by others. To put this in context the human genome, approximately 12% of the size of the radiata pine genome, cost US$1 billion to sequence and annotate and involved a large international consortium (International Human Genome Sequence Consortium, 2004). Even though genome sequencing costs are now only a fraction of when the human genome was first sequenced, contributions by other researchers will accelerate improvements in the radiata pine genome sequence at a much faster rate and with more detail than can be achieved by a single research group. We envision continual improvements in the assembly over time as more DNA sequence data are generated and re-assemblies are undertaken. Scion is currently exploring options for an international collaboration to undertake a more complete *de novo* assembly.

Impacts of the whole genome sequence on NZ forestry

The high cost of sequencing large genomes like radiata pine currently prevents widespread application of whole genome sequencing in operations such as breeding, seed production and propagation. Although sequencing costs have dropped over the last decade – 10,000-fold in the case of the human genome – further reductions are needed for whole genome sequencing to be cost-effective for everyday use in forestry.

Over the next few years, researchers are most likely to use the radiata pine genome sequence to accelerate the development of cheaper DNA-based diagnostic tools such as for DNA fingerprinting, parentage assignment, breeding population management and selective breeding, as well as for the discovery of genes of interest for further manipulation, or selection using new breeding techniques. The sequence will therefore inform research in the RPBC-led Genomic Selection in Radiata Pine Partnership Programme (see paper by Li et al. in this issue), and thus indirectly contributes to the future selection of superior genotypes in breeding populations based on the DNA markers. We anticipate that marker panels will require ongoing improvement and optimisation for specific applications, which will require the latest and most accurate information from the genome sequence for marker panel design.

For researchers, in the short term the whole genome sequence will accelerate the discovery of useful genes by providing an underpinning resource that already contains the DNA sequence of genes of interest. This contrasts with the present pre-genome situation where the development of DNA-based tools and the discovery of useful genes require project-by-project laboratory-based research, which can be slow and significantly more expensive. Identifying the sequence of specific genes will accelerate unique modifications using transgenic or targeted mutagenesis such as new gene editing technologies (Kim & Kim, 2014), improving the opportunity for developing special purpose germplasm and products. Similar outcomes are forecast from whole genome sequencing in other agronomically important species such as wheat (Mayer et al., 2014).

Disease resistance, growth and wood quality are a few of the traits that could be more rapidly improved using these combined approaches. For disease and insect resistance, identifying portions of radiata pine DNA sequence that are similar to genes involved in resistances to pathogens or insects in other species could provide a rapid means of developing new germplasm for enhanced resistance. Similarly, knowledge of genes influencing traits such as growth and wood quality in other tree species could also be applied by utilising variation in the corresponding DNA sequences of radiata pine.

The future

The first draft of the radiata pine genome is just the beginning. Scion-led efforts and ongoing international contributions will improve the genome by increasing the genome coverage and creating a more accurate assembly over time. The gene annotations are also expected to be updated continually as knowledge increases regarding the role(s) of specific genes and how they are regulated. DNA sequence variation will be added to the genome sequence and integrated with other information resources, including genetic maps.

These continuous improvements in DNA-based methods will spark new research. It is expected that identifying regions of the genome that have been subjected to natural and anthropogenic selection will point the way to the discovery of promising new genes for subsequent investigation. Similar methods have been used for humans (Fu & Akey, 2013) and agronomically important species to identify genes of interest and to reveal insights into evolutionary processes (e.g. Gu et al., 2009). Ultimately, these genomic resources will enable approaches to breeding in which selections are routinely made on the basis of DNA sequence variation. We also anticipate that new products will be generated, both in radiata pine and other tree species, using whole genome sequence information. These products could include gene modifications leading to the creation of new trees that are essentially factories making new products and speciality chemicals, thereby expanding the product range of radiata pine and other commercial forest tree species.

A new era of genomics-based breeding of radiata pine is also anticipated. Together with technical achievements in the Genomic Selection Partnership Programme, we anticipate that a combination of the genome sequence, knowledge of DNA sequence variants known to contribute to trait variation in specific environments, and analytical tools and techniques developed for other species will be routinely applied in radiata pine breeding. Such tools are already in use or are being developed for human medical genetics and livestock improvement (e.g. Cadzow et al., 2014; Varshney et al., 2014). We therefore

foresee that the genome sequence is likely to contribute substantively to the New Zealand forest industry, albeit slowly and incrementally at first, but ultimately becoming a key information resource that will underpin a new generation of trees, products and processes.

Acknowledgements

The authors would like to thank: the Scion Board of Directors for allocating Core Funding for the genome sequence project; Drs John Butcher and John Hay from the RPBC, members of the Forest Genetics team at Scion for their support and advocacy; Associate Professor Mik Black and Dr Beckie Laurie from the University of Otago and New Zealand Genomics Ltd for sequence data generation and provision of super-computer resources; Dr Shane Sturrock from BioMatters Ltd and Dr Gregory Gimenez from the University of Otago and New Zealand Genomics Ltd for assistance with assemblies; Professors David Neale and Charles Langley from the University of California, Davis and Professors Steven Salzberg and Daniela Puiu from Johns Hopkins University for advice and assistance; Dr Emily Telfer and Natalie Graham from Scion for generation of DNA material; and three reviewers for comments on earlier versions of the manuscript.

References

Birol I., et al. 2013. Assembling the 20 Gb White Spruce (*Picea Glauca*) Genome from Whole-Genome Shotgun Sequencing Data. *Bioinformatics*, 29(12): 1492–1497. doi:10.1093/bioinformatics/btt178/.

Cadzow, M., Boocock, J., Nguyen, H.T., Wilcox, P., Merriman, T.R. and Black, M.A. 2014. A Bioinformatics Workflow for Detecting Signatures of Selection in Genomic Data. *Frontiers in Genetics*, 5: 293. doi:10.3389/fgene.2014.00293.

Chaisson, M.J.P., et al. 2015. Resolving the Complexity of the Human Genome Using Single-Molecule Sequencing. *Nature,* 517(7536): 608–611. doi:10.1038/nature13907. Available from: www.nature.com/nature/journal/v517/n7536/abs/nature13907.html#supplementary-information.

Ellegren, H. 2014. Genome Sequencing and Population Genomics in Non-Model Organisms. *Trends Ecol Evol*, 29(1): 51–63. Available from: http://dx.doi.org/10.1016/j.tree.2013.09.008.

Fu, W. and Akey, J.M. 2013. Selection and Adaptation in the Human Genome. *Annual Review of Genomics and Human Genetics*, 14: 467–489.

Gu, J., Orr, N., Park, S., Katz, L., Sulimova, G., MacHugh, D. and Hill, E. 2009. A Genome Scan for Positive Selection in Thoroughbred Horses. *PLoS ONE* 4, (6): e5767.

Hood, L. and Rowen, L. 2013. The Human Genome Project: Big Science Transforms Biology and Medicine. *Genome Med*, 5(9): 79.

International Human Genome Sequencing Consortium. 2004. Finishing the Euchromatic Sequence of the Human Genome. *Nature*, 431(7011): 931–945. Available from: www.nature.com/nature/journal/v431/n7011/suppinfo/nature03001_S1.html.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M. and Haussler, D. 2002. The Human Genome Browser at UCSC. *Genome Research*, 12(6): 996–1006.

Kim, H. and Kim, J-S. 2014. A Guide To Genome Engineering with Programmable Nucleases. *Nat Rev Genet*, 15(5): 321–334. doi:10.1038/nrg3686.

Koboldt, D.C., Steinberg, K.M., Larson, D.E., Wilson, R.K. and Mardis, E.R. 2013. The Next-Generation Sequencing Revolution and its Impact on Genomics. *Cell*, 155(1): 27–38.

Mayer K.F., Rogers, J., Doležel, J., Pozniak, C., Eversole, K., Feuillet, C., Gill, B., Friebe, B., Lukaszewski, A.J. and Sourdille, P. 2014. A Chromosome-Based Draft Sequence of the Hexaploid Bread Wheat (Triticum Aestivum) Genome. *Science*, 345(6194): 1251788.

Morgante, M. and De Paoli, E. 2011. Toward the Conifer Genome Sequence. In Plomion, C., Bousquet, J. and Kole, C. (Eds). *Genetics, Genomics and Breeding of Conifers Trees* (1st edn). Science Publishers, Edenbridge Ltd, St Helier, Jersey, British Channel Islands, 389–403.

Murray B.G. 1998. Nuclear DNA Amounts in Gymnosperms. *Annals of Botany*, 82 (SUPPL. A): 3–15.

Myburg, A.A., Grattapaglia, D., Tuskan. G.A., Hellsten, U., Hayes, R.D., Grimwood, J., Jenkins, J., Lindquist, E., Tice, H. and Bauer, D. 2014. *The Genome of Eucalyptus grandis*. *Nature*, 510(7505): 356–362.

Neale, D.B., Wegrzyn, J.L., Stevens, K.A., Zimin, A.V., Puiu, D., Crepeau, M.W., Cardeno, C., Koriabine, M, Holtz-Morris, A.E. and Liechty, J.D. 2014. Decoding the Massive Genome of Loblolly Pine Using Haploid DNA and Novel Assembly Strategies. *Genome Biology*, 15(3): R59.

Nystedt, B., Street, N.R., Wetterbom, A., Zuccolo, A., Lin, Y-C., Scofield, D.G., Vezzi, F., Delhomme, N., Giacomello, S. and Alexeyenko, A. 2013. The Norway Spruce Genome Sequence and Conifer Genome Evolution. *Nature*, 497(7451): 579–584.

Tuskan, G.A., et al. 2006. The Genome of Black Cottonwood, *Populus trichocarpa* (Torr. & Gray). Science, 313(5793): 1596–1604.

Varshney, R.K., Terauchi, R. and McCouch, S.R. 2014. Harvesting the Promising Fruits of Genomics: Applying Genome Sequencing Technologies to Crop Breeding. *PLoS Biol*, 12(6): e1001883.

Zimin A., Stevens, K.A., Crepeau, M.W., Holtz-Morris, A., Koriabine, M., Marçais, G., Puiu, D,, Roberts, M., Wegrzyn, J.L. and de Jong, P.J. 2014. Sequencing and Assembly of the 22-Gb Loblolly Pine Genome. *Genetics*, 196(3): 875–890.

Phillip Wilcox is a Senior Scientist at Scion and a part-time Senior Research Fellow at the University of Otago. Lucy Macdonald is a Bioinformatician at Scion based in Rotorua.