APPLYING SECOND-GENERATION SEQUENCING TO NON-MODEL SPECIES

Carsten Külheim*
Research School of Biology, Australian National University, Canberra, ACT 0200
*Corresponding author: carsten.kulheim@anu.edu.au

Introduction

Early whole-genome sequencing focused on model species such as *Arabidopsis* and mice (*Mus musculus*) (1,2) because they had small genomes and were widely used in biomedicine and plant science. Consequently, there are many useful genomic resources for these species, including genetic databases, annotated genomes and maps of metabolic pathways. However, for non-model species, the equivalent resources are scarce. A significant focus of Australian science is on studies of our unique native flora and fauna, including monotremes such as the platypus (*Ornithorhynchus anatinus*) (3), marsupials such as the Common Brushtail possum (*Trichosurus vulpecula*) (4), and the dominant plant radiations of the eucalypt (5) and *Acacia* (6) groups. Second-generation sequencing provides a cost-effective tool for genetic research on non-model organisms, but it remains under-utilised despite the large number of sequencing platforms being installed in universities and research institutes around the country. There are numerous platforms available, with the three most common in Australia being the 454 GS-FLX (Roche Life Science), GA*IIx* (Illumina) and SOLiD 4 (Applied Biosystems; ABI). Each system has its advantages and disadvantages (reviewed in (7)). One of the major differences is that the amount of data produced by the GS-FLX is approximately an order of magnitude smaller than that of the other two platforms. Whilst this may seem like a disadvantage, it actually holds many advantages for research on non-model organisms. This is due to the fact that downstream analysis can be performed on any average personal computer, whereas the data handling with the other platforms requires more advanced computer technology. As data handling and analysis can represent the majority of any second-generation sequencing experiment, this is a major consideration.

This article provides an overview of how these novel technologies can be used on non-model organisms and in ways that are scalable to most laboratories' interests. These applications include transcriptome sequencing, targeted sequencing, genotyping and population genetics.

Transcriptome Sequencing

Extracting and sequencing expressed genes can be simple and informative on many levels. First, it can provide novel DNA sequences from a previously unknown genome. Second, allelic variants can be identified when sequencing from pools of cDNA or from multiple samples. Third, insertions, deletions and data on alternate splicing sites can be found. Finally, transcript levels can be quantified from multiple samples, giving information on relative gene expression within and between each sample.

During the past two decades, transcript quantification has gone through several revolutionising developments. The method of choice up to the late '90s was based on blotting RNA to a membrane after separation on a gel, with subsequent detection of a single or a few transcripts by specific hybridisation with a labelled probe (Northern blot). This approach was laborious, costly, often involved radioactively labelled probes, and had a low throughput. More importantly, it also required prior knowledge of the gene of interest. At the end of the '90s, microarrays became the method of choice. This method was based on hybridising labelled cDNA to a small chip with either complementary DNA derived from expressed sequence tags (ESTs) or, later, whole genome oligonucleotides. This method was scalable to some extent and, because it was not necessary to know the function of each probe, provided a less biased approach to investigating the transcriptome. For non-model species, arrays could be designed from a small EST library. This meant that there were high initial costs for the discovery of the EST library, plus the preparation of the microarray slide. Another method that has been used for over a decade is quantitative real-time PCR (qRT-PCR) (8). Although this method can be used with a large number of samples, the number of transcripts that can be investigated is low and dependent on previous knowledge of the sequence of the transcript.

Long-read second-generation sequencing has several advantages over these early methods of transcript quantification (**Fig. 1**). The number of transcripts that can be assayed simultaneously is very high, and the method is cheap, fast and easy compared to qRT-PCR or development of a microarray. Transcript discovery and quantification can be done in a single step. Nonetheless, some sequencing platforms are more useful than others for non-model organisms for which there are few genomic resources. The first step after transcriptome sequencing of an unknown genome is to perform a *de novo* assembly of the reads, as compared to a reference assembly on organisms with a known genome sequence, which reduces the advantage of the most popular short-read platforms (Illumina's GA*IIx* and ABI's SOLiD). The *de novo* assembly can be done more easily with read lengths around 400bp as compared to 100bp (Illumina) or 50bp (ABI). Using the GS-FLX (Roche) reduces the number of reads per run from 100s of millions to a mere 1 million, but also reduces the dynamic range of the transcriptome experiment such that only highly expressed genes are discovered and assayed. That said, the number of transcripts assayed is still very high and may range from 7,500 transcripts (about 25% of

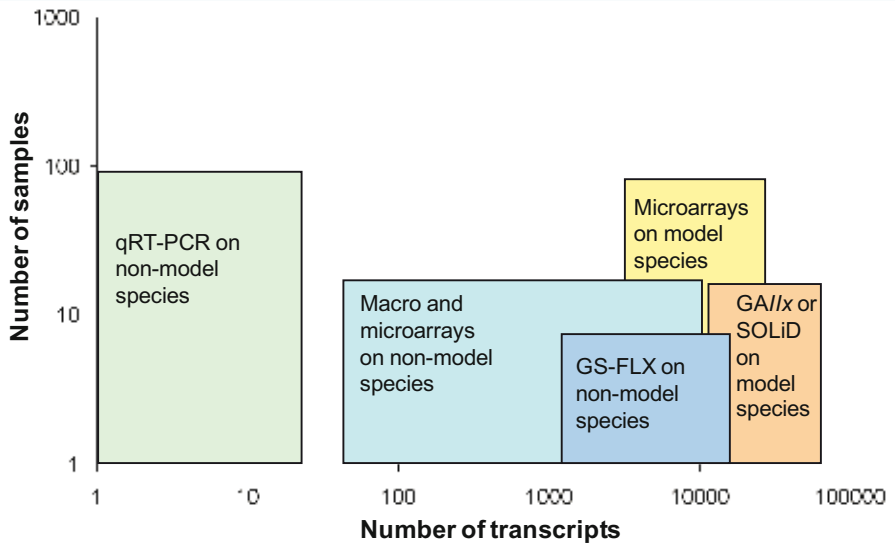

Fig. 1. Different methods of transcript quantification currently in use.

Depending on the experimental requirements, different quantification platforms are available. Custom-made macro- and microarrays for non-model species are compared to microarrays from model species, such as Affymetrix tiling arrays. Short-read (SOLiD and GAIIx) and long-read (GS-FLX) second-generation sequencing are compared.

the transcriptome) in *Eucalyptus melliodora* (Potts, Külheim, Keszei and Foley, unpublished data) to 11,800 transcripts (about 65% of the transcriptome) in zebra finch (9).

Our lab at ANU is concerned with the genetic control of secondary chemistry related to herbivore defence in Australian plants. Recently, we analysed two cDNA samples on a quarter of a GS-FLX picotitre plate (PTP) each. The samples were from two different branches of a mosaic *Eucalyptus melliodora* tree (10), which have vastly different terpene and phenolic chemistries within branches of the same tree. The two samples were combined prior to a *de novo* assembly in order to increase the number of transcripts that could be discovered from the data. The 283,439 reads were assembled into 7,469 contigs. The contigs could then be used as a reference in new assemblies from each of the two samples and the number of reads per transcript could be compared. To quantify the expression levels, reads per kilobase of exons values were calculated for each sample and contig (11). We were able to make approximately 3,200 meaningful comparisons and found that about 100 transcripts had a strong differential expression between the two branches of the same tree (**Fig. 2**). These transcripts of interest were then compared to public databases and biologically meaningful conclusions about the biosynthesis of terpenes and phenolics in the different branches could be drawn (Potts, Külheim, Keszei and Foley, unpublished data). We were also able to identify a number of alternative splicing sites, which is of significant interest in recent literature (12, 13).

Our approach could easily be applied to a number of projects and experiments on non-model species with no known reference genome. Further, this approach would be scalable to some degree. Instead of 2 x one-quarter PTP, 2 x one-half PTP could be used for a two sample comparison or 4 x one-quarter PTP for four samples. A single experiment could be the basis of an expressed gene library and future experiments in that or a related species could be done on one of the short-read platforms, where many more samples could be studied simultaneously. Further, small barcodes (see below) could be ligated to the cDNA prior to sequencing and the number of samples per run could thus be increased. This seems only feasible if a genome or a reference transcriptome is present.

Reduced Genome and Targeted Sequencing

It is often neither feasible nor practical to attempt to sequence the whole genome of a species of interest. However, smaller amounts of genomic data may be obtained from specific genes, reduced genomes or genomic regions of interest from one or several species, and few or many individuals. When planning a second-generation sequencing experiment, the first consideration is the aim of the experiment. If the aim is to discover sequence variants within a species or a population, DNA from multiple individuals can be pooled together, the genes of interest amplified and then sequenced. This can also be done for individuals from multiple species and the species can either be physically separated on the PTP or barcoded by kits provided by several vendors. This can also be done for different populations or even different individuals of the same species. If there are more than 12 samples (or 2x12 or 4x12 for the different PTP conformations), custom designed and ligated barcodes can be used to later separate the sequences based on their origin (see below).

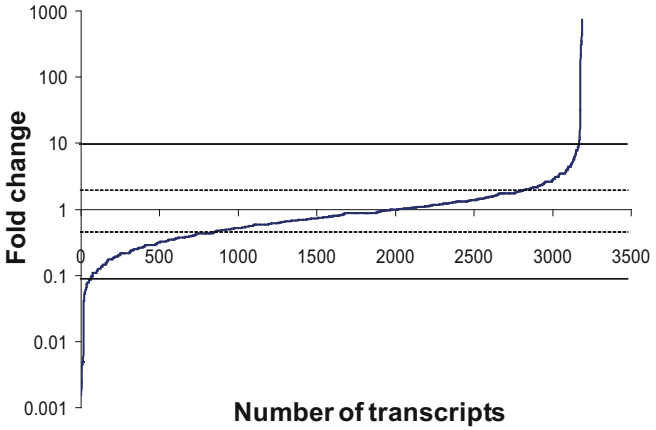

Fig. 2. Ratio of relative transcript levels between two samples of Eucalyptus melliodora.

Relative reads per kilobase of exons values are shown for approximately 3,200 transcripts. Two-fold (dashed line) and ten-fold (solid line) changes are indicated.

The availability of existing genomic resources for the species or family of interest is also critical in experimental design. If no genomic resources are available, degenerate primers can be designed for a selected number of genes of interest prior to amplification and sequencing. This is possible only for a small number of genes, because of time and cost constraints. Alternatively, when dealing with eukaryotic genomes, which are still too large for comparative whole genome sequencing, sequence enrichment can be achieved by restriction fractioning of the genome, separation, and selection by size prior to sequencing (14, 15). If some genetic resources are available, for example, small- to medium-size EST databases, then a search for homologues of a gene of interest can lead to the discovery of expressed genes for which specific primers can be designed and the genomic DNA can then be amplified and sequenced (16). When significant genetic resources are available, genes or genomic regions can be incorporated in a Roche NimbleGen capture array (17) or enriched with the Agilent SureSelect Target enrichment system (18) prior to second-generation sequencing. This removes the PCR amplification step and clean-up, thereby greatly reducing the amount of lab work and costs involved.

In our laboratory's research, we wished to discover all or most nucleotide variants from 36 genes in four different *Eucalyptus* species. At that time, there were modest EST databases, but no publicly available genome. Therefore, we searched public databases for homologues of genes in *Eucalyptus* from four secondary metabolite biosynthetic pathways and designed specific primers for one species, *Eucalyptus globulus*. The primers were designed in conserved regions of the exons, so that we were able to use the same primers on multiple species from the same genus. We pooled DNA from 381 to 511 individuals from the four species, covering each species' geographic range (in some cases, this approximates about 50% of the Australian landmass). The 36 loci from the four pools of DNA were amplified, equimolar amounts of amplicons were pooled together for each species, then barcoded using Roche's multiplex identifier (MID) adaptors and sequenced on a 454 GS-FLX pyrosequencer. Data were obtained for 23 loci, covering 50kbp. The remainder of the loci came from multi-gene families, where our primers amplified more than one locus at a time, making the discovery of discrete allelic variants impossible, which is a disadvantage of using primers in conserved regions. The reads were sorted by their MID identifier and assembled to a reference sequence. We discovered 8,631 high-confidence single nucleotide polymorphisms (SNP) from the four species, with average densities between one SNP in every 33bp in *E. nitens* and one SNP in every 16bp in *E. camaldulensis* (16). We used these data to genotype approximately 400 common SNPs in 480 individuals of *E. globulus* and then used association mapping to relate these variants to phenotypic variation within the species (Külheim *et al.*., unpublished data).

Barcoded Amplicons Sequencing

Most second-generation sequencing platforms offer the possibility of identifying individual samples using short oligonucleotides that can be ligated to the DNA of interest. For example, 454 Life Science offers twelve MIDs of 10 nucleotide (nt) length, while Illumina offers a set of twelve oligonucleotide tags of 6nt length. These additional bases reduce the effective read length to some degree, but they can greatly enhance the potential applications of second-generation sequencing. If the expected coverage per position greatly exceeds the experimental needs, it becomes feasible to sequence multiple samples at the same time, for example, when sequencing short amplicons or mitochondrial genomes. Coupled with the physical conformations of the sequencing platform, 96 samples (Illumina) or 24/48/96/192 samples (depending on the setup of the PTP) (454 Life Science) can be sequenced in a single sequencing experiment. Using conformations with more regions on the 454 platform has the disadvantage of greatly decreasing the total number of reads generated per sequencing run. It can therefore be a good idea to apply custom-made barcodes, such as those developed by Meyer and colleagues (19), who developed sets of 6, 7 and 8nt long barcodes giving up to 475 different identifiers with at least two substitutions between them. This barcode set had to be reduced to decrease runs of homopolymers within the barcode and a subset of 226 fulfils these requirements. Combined with the PTP setup, 452, 904, 1808 or 3616 samples can be sequenced in parallel. Galan and colleagues (20) developed a set of short barcode tags that are added to both the forward (36 tags) and reverse (24 tags) primer at the PCR amplification step, leading to a maximum of 864 different combinations. The advantage of this method is the high number of combinations, while the disadvantage lies in the need to design a large number of different long primers (20).

The number of different applications with this method is very large. For example, Stiller and co-workers used this method to sequence full mitochondrial genomes from 31 extinct cave bears (*Ursus spelaeus*) (21), while Williams and co-workers sequenced a reduced genome of killifish (*Fundulus heteroclitus*) in eight individuals, each from ten different collection sites. This enabled them to discover and genotype allelic variants for population genetic estimates (15). While the discipline of meta-genomics sequences environmental samples and then tries to reconstruct which species were present in that sample, barcoding can be applied to phylogenetics, which can be done by sequencing informative amplicons (e.g., *matK, ITS, ndhF*) from a large number of species that are pooled together after barcoding. This is especially applicable for some of the very large and diverse plant families present in Australia, such as the Myrtaceae *(Eucalyptus, Melaleuca, Syzygium),* Proteaceae *(Banksia, Grevillea, Hakea)* and Fabaceae *(Acacia)* families.

We adapted the method developed by Meyer *et al.* (19) in our studies to our experimental design and amplified seven loci (approximately 11kbp) from 188 individuals

of medicinal teatree (*Melaleuca alternifolia*). The seven loci were pooled, an 8nt barcode was ligated to the amplicons of each individual and the combined sample was sequenced on one-quarter of a GS-FLX PTP. All reads were assembled to the reference sequence of the seven loci and a CLC Genomics Workbench (CLC bio, Aarhus, Denmark) was used to discover allelic variants. Approximately 100 common SNPs were selected to be genotyped from each individual. Neither the Roche GS Reference Mapper software nor CLC Genomics Workbench were capable of separating the reads by their barcode, which necessitated a custom Biopython script that sorted each read into a file corresponding to their barcode. After separation by barcode, each individual could be genotyped for the SNPs of interest (Webb, Külheim, Moran and Foley, unpublished data). This approach was also used for eleven loci (26kbp) from 110 individuals of *E. globulus* (Yeoh, Külheim, Moran and Foley, unpublished data).

Conclusions

Second-generation sequencing provides powerful tools to many investigators around Australia working with non-model organisms. The applications are vast, ranging from whole genome sequencing (for small genomes), reduced genome sequencing, transcriptome sequencing for the discovery of transcript sequences as well as quantification of transcripts and splicing variants, targeted sequencing (amplicons) and barcoded targeted sequencing for phylogenetics or population genetics. Second-generation sequencing is a very recent development and the technology is changing rapidly. Initial application faced high sequencing costs and had low throughput, and thus, few publications on non-model species were produced. In the past couple of years, this has changed and laboratories around the world are applying second-generation sequencing more and more commonly to non-model organisms.

References

1. The Arabidopsis Genome Initiative. (2000) *Nature* **408**, 796-815
2. Waterston, R.H., Lindblad-Toh, K., Birney, E., *et al.* (2002) *Nature* **420**, 520-562
3. Warren, W.C., Hillier, L.W., Graves, J.A.M., *et al.* (2008) *Nature* **453**, 175-183
4. Degabriel, J.L., Moore, B.D., Foley, W.J., and Johnson, C.N. (2009) *Ecology* **90**, 711-719
5. Ladiges, P.Y., Udovicic, F., and Nelson, G. (2003) *J Biogeogr.* **30**, 989-998
6. Miller, J.T., Andrew, R., and Bayer, R.J. (2003) *Aust. J. Bot.* **51**, 167-177
7. Shendure, J. and Ji, H.L. (2008) *Nat. Biotechnol.* **26**, 1135-1145
8. Gibson, U.E.M., Heid, C.A., and Williams, P.M. (1996) *Genome Res.* **6**, 995-1001
9. Ekblom, R., Balakrishnan, C.N., Burke, T., and Slate, J. (2010) *BMC Genomics* **11**, 219
10. Edwards, P.B., Wanjura, W.J., Brown, W.V., and Dearn, J.M. (1990) *Nature* **347**, 434
11. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008) *Nat. Methods* **5**, 621-628
12. Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010) *Nat. Biotechnol.* **28**, 511-515
13. Guttman, M., Garber, M., Levin, J.Z., *et al.* (2010) *Nat. Biotechnol.* **28**, 503-510
14. Castano Sanchez, C., Smith, T., Wiedmann, R., Vallejo, R., Salem, M., Yao, J., and Rexroad III, C. (2010) *BMC Genomics* **10**, 559
15. Williams, L.M., Ma, X., Boyko, A.R., Bustamante, C.D., and Oleksiak, M.F. (2010) *BMC Genet.* **11**, 32
16. Külheim, C., Yeoh, S.H., Maintz, J., Foley, W.J., and Moran, G.F. (2009) *BMC Genomics* **10**, 11
17. Fu, Y., Springer, N.M., Gerhardt, D.J., *et al.* (2010) *Plant J.* **62**, 898-909
18. Gnirke, A., Melnikov, A., Maguire, J., *et al.* (2009) *Nat. Biotechnol.* **27**, 182-189
19. Meyer, M., Stenzel, U., and Hofreiter, M. (2008) *Nat. Protoc.* **3**, 267-278
20. Galan, M., Guivier, E., Caraux, G., Charbonnel, N., and Cosson, J.-F. (2010) *BMC Genomics* **11**, 296
21. Stiller, M., Knapp, M., Stenzel, U., Hofreiter, M., and Meyer, M. (2009) *Genome Res.* **19**, 1843-1848