AN APPROACH TO TRANSCRIPTOME ANALYSIS OF NON-MODEL ORGANISMS USING SHORT-READ SEQUENCES

LESLEY J COLLINS[1,2]
L.J.Collins@massey.ac.nz

PATRICK J BIGGS[1,2]
P.Biggs@massey.ac.nz

CLAUDIA VOELCKEL[1]
C.Voelckel@massey.ac.nz

SIMON JOLY[1,3]
S.Joly@massey.ac.nz

[1] *Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, Palmerston North, New Zealand*
[2] *Institute of Molecular BioSciences, Massey University, Palmerston North, New Zealand*
[3] *Current address: Department of Biology, McGill University, Montreal, Quebec, Canada*

Transcriptome analysis using high-throughput short-read sequencing technology is straightforward when the sequenced genome is the same species or extremely similar to the reference genome. We present an analysis approach for when the sequenced organism does not have an already sequenced genome that can be used for a reference, as will be the case of many non-model organisms. As proof of concept, data from Solexa sequencing of the polyploid plant *Pachycladon enysii* was analysed using our approach with its nearest model reference genome being the diploid plant *Arabidopsis thaliana.* By using a combination of mapping and *de novo* assembly tools we could determine duplicate genes belonging to one or other of the genome copies. Our approach demonstrates that transcriptome analysis using high-throughput short-read sequencing need not be restricted to the genomes of model organisms.

Keywords: short-read sequencing; next generation sequencing *Pachycladon*; transcriptome analysis.

1. Introduction

High-throughput short-read sequencing is one of the latest sequencing technologies to be released to the genomics community. For example, on average a single run on the Illumina Genome Analyser can result in over 30 to 40 million single-end (~35 nt) sequences. However, the resulting output can easily overwhelm genomic analysis systems designed for the length of traditional Sanger sequencing, or even the smaller volumes of data resulting from 454 (Roche) sequencing technology.

Typically, the initial use of short-read sequencing was confined to matching data from genomes that were nearly identical to the reference genome. This enabled easy comparisons between genomes in order to investigate differences either in the genomic sequence itself (SNPs - single nucleotide polymorphisms, and other mutations), gene expression (transcriptomics), small RNAs, methylation or chromatin mapping (ChIP-sequencing) (examples [1; 2]). However, researchers are now pushing the boundaries of this technology to sequence more distantly related genomes. Our study presents an approach to transcriptome analysis of a non-model genome.

Transcriptome analysis on a global gene expression level is an ideal application of short-read sequencing. Traditionally such analysis involved complementary DNA (cDNA) library construction, Sanger sequencing of ESTs, and microarray analysis. Next generation sequencing has become a feasible method for increasing sequencing depth and coverage while reducing time and cost compared to the traditional Sanger method. A method for non-model organisms using 454 pyrosequencing data was recently published [3], highlighting how next-generation sequencing enables transcriptome analysis from any species. Short-read sequencing produces a far greater coverage even though the sequences produced are shorter than those produced by pyrosequencing. Genome projects are now looking not only to produce sequence counts of individual ESTs obtained using short-read sequencing, but to produce the EST sequences in the first place to investigate EST characteristics prior to counting. Our study introduces an approach enabling the latter, demonstrating its usefulness on data obtained from the *Pachycladon* transcriptome project.

The genus *Pachycladon* is an emerging non-model system in the study of plant speciation. The whole genus (2n=4x=20) is of allopolyploid origin from distant parents in the Brassicaceae family (S. Joly, P. Heenan and P. Lockhart, unpublished data), meaning that we expect (most) genes to be duplicated. Both genome copies present in *Pachycladon* diverged from the model species *Arabidopsis thaliana*, a functional diploid (2n=2x=10), relatively recently (ca. 7-10 Mya). The small number of species, its young age and its close relationship with *A. thaliana*, suit *Pachycladon* for evolutionary studies investigating the ecological drivers and the molecular basis of species diversification. Multiple approaches can be used to address these questions, including gene expression profiling, QTL mapping and candidate gene studies, all of which require molecular resources such as an EST library. These applications also require prior characterization of duplicate gene copies. Short-read sequences of amplified cDNA from roots and shoots of *Pachycladon enysii* obtained with the Illumina Genome Analyzer provided an opportunity to explore an efficient, inexpensive and reliable approach to EST sequencing that can be readily adopted by researchers studying non-model organisms. Our analysis resulted in the identification of duplicate gene candidates from *Pachycladon* ESTs, some of which could be matched to *A. thaliana* ESTs showing that analysis of short-read sequences is feasible when the reference genome is distantly related.

2. Approach Overview

Our overall approach to non-model organism transcriptome analysis (as shown in Figure 1) is to use high-throughput short-read sequences, optimize assembly and mapping parameters using partial data, then process the total data using these optimized mapping and *de novo* parameters. Assembled contigs are compared to themselves and also to the nominated reference genome using BLAST, leading to the extraction of candidate duplicate genes. Results are visualized at different stages for validation purposes.

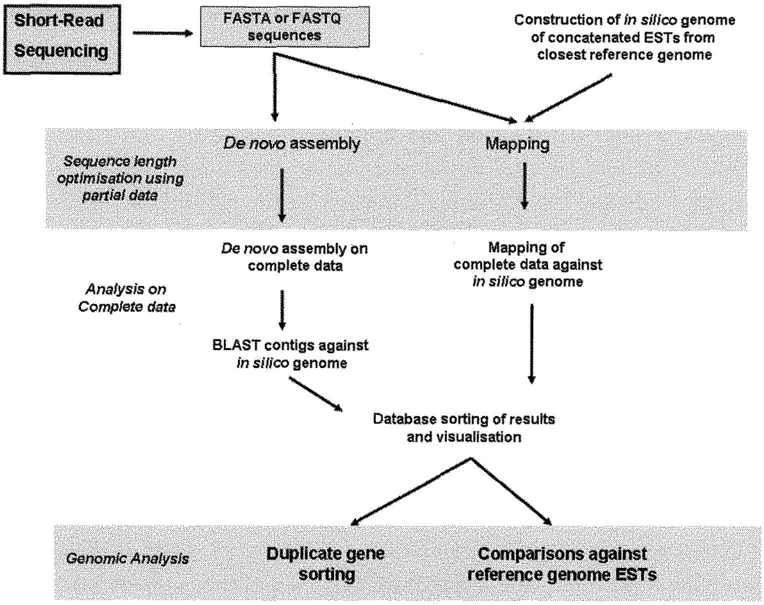
Short-Read Sequencing
FASTA or FASTQ sequences
Construction of *in silico* genome of concatenated ESTs from closest reference genome
De novo assembly
Mapping
Sequence length optimisation using partial data
De novo assembly on complete data
Mapping of complete data against *in silico* genome
Analysis on Complete data
BLAST contigs against *in silico* genome
Database sorting of results and visualisation
Genomic Analysis
Duplicate gene sorting
Comparisons against reference genome ESTs

Figure 1. Overview of short-read based transcriptome analysis approach for non-model organisms. Mapping is an option only if a suitable genome is available, otherwise FASTA (or FASTQ) sequences can only proceed down the *de novo* track or into other project-specific analysis such as sequence counting (not shown). However mapping against even a distant genome can provide valuable information about genome conservation so should be done where possible.

Since the output is large, all data is managed and curated with a MySQL database from which genome areas of interest can be extracted. Reformatting and data extraction is handled through the use of Perl and MySQL scripts. Details about each stage of this approach are given below.

2.1 Dataset volumes and data management

Data output from short-read sequencing is large, consisting of millions of sequences and preliminary mapping output. To manage these data volumes as well as sequence and result curation we used the MySQL database system (version 5.0.45, running under Windows XP-Pro). This database was also used to store BLAST results, EST location and other relevant information. The MySQL database was also linked to the Gbrowse genome browser [4], to enable viewing of data subsets. We see no problems to other databases being used so long as they are robust enough to handle these data volumes, data types and genome viewer integration. We used data from the Illumina Genome Analyser (also known as Solexa Sequencing), but this approach is applicable to data produced from other platforms (such as the SOLiD platform from Applied Biosystems), so long as the sequence output has already been converted from any internal and/or proprietary forms (such as the SOLiD 'colour space') to the more standard FASTA or FASTQ format.

2.2 Data subset extraction and optimal parameter evaluation for mapping

At the end of the sequencing run the short-read sequences were converted to FASTA or FASTQ output and mapped against a nominated reference genome as part of the Illumina Genome Analyser Pipeline. However, this preliminary analysis can use parameters that may not produce optimal results. For example, the maximum number of mismatches allowed between reads and the reference sequence by the pipeline software (ELAND) is two, which could be a too restrictive value when one is using a more distant genome as a reference. A related parameter is the sequence length for ELAND mapping because longer reads mean more potential mismatches between the two genomes, and thus resulting in more non-mapped sequences.

One way to choose optimal parameters for analysis is by running simulations, but it is not possible to simulate data from a genome not already sequenced; this can only be done after the sequencing run. The primary parameter that required determination for this application was the sequence length used for both ELAND mapping and *de novo* assembly. Because of the large volumes of output from a single short-read run it is not efficient to determine experimental parameters on the entire dataset. Instead we use as standard the data from one lane (approximately 4 million sequences from a lower titration).

The Illumina pipeline software ELAND was run on this data subset with the sequence length parameter set for 17 initially, and then increased by one until the maximum of 32 was reached. This means the first 17 bases of the sequences are used for mapping to the reference genome. If the sequence length is set too short then we can expect to see a steep increase in the number of repeat-matches as the 'specificity' of the match lowers. However, if the sequence length is set too long then we run the risk of generating more non-matches as the number of differences between the sequenced genome and the reference genome will push the match beyond ELAND's limit of two mismatches. By rerunning a subset of data over a wide range of sequence lengths, an optimal length can thus be selected.

Another popular mapping software Maq (Release 0.5.0 [5]) was briefly compared. Maq offers the advantage of allowing a higher number of mismatches (three opposed to the two offered with ELAND) but is much slower when this is permitted. Maq uses FASTQ input incorporating quality information as well as sequence information. Users of Solexa produced FASTQ data should be aware that the scores are calculated differently from Sanger-type sequencing FASTQ and can include calibration from the initial mapping to a reference genome. When working with distantly related reference genomes, potential users of this software should specify 'uncalibrated quality scores' from a Solexa sequencing service. There are also some functions in Maq that have been specifically written for the SOLiD platform.

The third piece of software we compared was SOAP [6]. SOAP is similar to ELAND in that it uses hash look-up table algorithms to speed up analysis and runs comparably [6]. It also has a limit of two mismatches. Although we used ELAND for the proof of concept of our approach, any of these other software packages could theoretically be substituted for short-read sequences from any platform.

For analysis, an *in silico* reference genome must also be prepared from the many discontiguous sequences within EST sequence libraries. Mapping to each EST separately is possible so long as the conditions for running ELAND are met (ELAND documentation from Illumina). To construct the *in silico* EST 'genome' we concatenated the EST sequences leaving 50 'N's between each EST sequence. Co-ordinates for each sequence are retained during this process so that mappings against each EST can be determined separately.

2.3 De novo assembly

Out of the *de novo* assemblers available for handling short-read output (including Velvet [8], SSAKE [7] VCAKE [8]and SHARCGS [9]), we chose to primarily use Velvet (version 0.5) [10] as it was found to produce consistent and sizable contigs. Velvet was developed specifically for manipulating short-read sequences and uses de Bruijn graphs for sequence assembly. However, a downside is that it runs only in a 64 bit Linux environment. As with the mapping, the optimal 'k-mer' size (a Velvet parameter comparable to 'word' size used for BLAST searches) is determined using a subset of the entire data, although it is feasible for entire datasets to be assembled with a variety of k-mer lengths and the results compared. Assembled contigs are then BLASTed against themselves to find exact copies or against any other similar genomes using BLAST [11].

The results of the BLAST analyses are loaded into a MySQL database for further processing. The sequences of the contigs and coordinates of the hits to the *A. thaliana* EST genome were output so they could be viewed with a combination of Gbrowse and MySQL. The combination of reference genome mapping and BLASTing of contigs from *de novo* assembly then allows us to pull out regions corresponding to duplicate genes.

3. *Pachycladon* Transcriptome short-read analysis

The *Pachycladon enysii* Transcriptome project presented two genomic challenges. The first relates to the fact that the closest reference genome that could be used was the plant *A. thaliana*, a species that diverged 7-10 million years from both genomic copies present in *Pachycladon* (S. Joly, P. Heenan and P. Lockhart, unpublished data). Prior to this work, there were no published studies on whether a genome of a different species could be used as a reference genome in short-read sequencing. Thus our approach was used to both study this effect and to aid the construction of *Pachycladon* ESTs for further analyses. The second type of genomic challenge is that *P. enysii* is a polyploid organism with two genome copies whereas *A. thaliana* is a diploid organism with one genome copy. Polyploidy on any level creates issues for genome analysis. Our aim was to map *Pachycladon* orthologues to specific *A. thaliana* loci and find putative duplicate *Pachycladon* genes.

The *Pachycladon* RNA was extracted separately from the roots and leaves of one rosette-stage *P. enysii* specimen originating from Avalanche Peak, South Island, New Zealand using the Qiagen RNeasy kit (Biolab Ltd.). An equal amount of root and leaf RNA (12.5μg) was pooled and reverse-transcribed using the SuperScript™ Double-Stranded cDNA Synthesis Kit (Invitrogen) and oligo(dT) primers (Invitrogen). Double-stranded cDNA (3μl) was subsequently amplified using the Qiagen REPLI-g Mini Kit (Biolab Ltd.). Five μg of the REPLI-g-amplified *P. enysii* cDNA was then used as a template for Solexa Genomic DNA preparation. Solexa sequencing used the Genome DNA Sample Preparation kit (FC-102-1001, Illumina) over 36 cycles.

Solexa sequencing produced a total of 40 million single-end short-reads of 36 nucleotides (nt). Seven lanes were used and contained different numbers of sequences due to a titration of DNA concentrations being used to generate clusters on the flowcell. This data was analyzed using our approach and the results are described below.

3.1. *Mapping against Arabidopsis ESTs*

An *in silico* genome was constructed from *A. thaliana* ESTs by concatenating the TAIR7 EST dataset (TAIR7_cDNA_20070425) [12]. Each EST was separated by 50 'N's (to prevent short-read sequences mapping to more than one EST), and all coordinates recorded for later mapping. Because of the sequence distance between the *Pachycladon* genome and the *A. thaliana* EST reference genome, we recognized that using the full length of the sequence for the match may potentially exclude many sequences due to the mapping software (ELAND) only allowing up to two mismatches per sequence. However, even with a low percentage of unique matches expected, mapping to a nearby reference genome enables us to examine the conserved portion of the *Pachycladon* transcriptome.

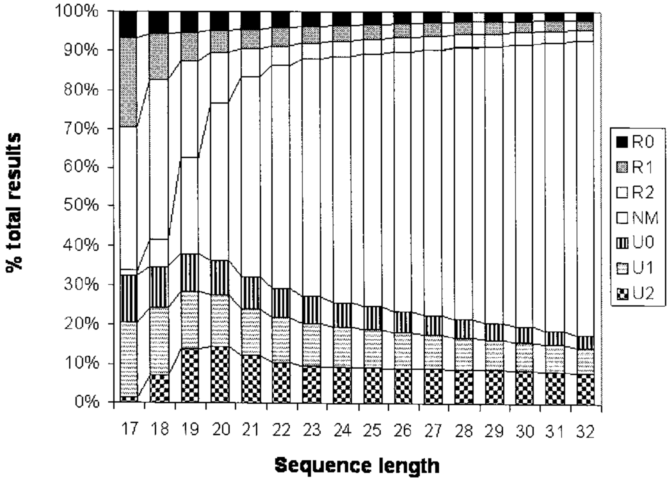

Figure 2. Graph of ELAND performed at different sequence lengths on one lane of sequences (4 million). Results are scored as a percentage of the total number of sequences. We get the greatest percentage of unique hits (37%) using a sequence length of 19.Key: U - Unique match U0 (no mismatches) U1 (1 mismatch) U2 (2 mismatches); R - Repeat match R0, R1 and R2 as for the unique hits; QC -Quality filter fail; NM - no match to reference genome. (QC results are omitted as they are too small to be seen on this graph.)

Table 1. Results from ELAND analysis of *Pachycladon* sequences against *A. thaliana* ESTs. Key: U - Unique match U0 (no mismatches) U1 (1 mismatch) U2 (2 mismatches); R - Repeat match R0, R1 and R2 as for the unique hits; QC -Quality filter fail; NM - no match to reference genome. The percentages obtained from a single lane of data were comparable with that from the entire 7 lanes.

ELAND result type	Length = 19 (1 lane)	% Total (1 lane)	Number (7 lanes)	Av % Total (7 lanes)	Std Dev (7 lanes)
U0	341804	8.99	3577678	8.86	1.2
U1	543248	14.29	5697913	14.23	0.34
U2	520442	13.69	5547612	13.85	0.29
R0	960119	25.26	1943961	24.84	0.93
R1	191895	5.05	2914570	4.88	0.22
R2	275036	7.24	10455270	7.27	0.16
QC	14	0.00	121	0.00	0.00
NM	968842	25.49	9916563	26.07	0.79
Total	**3801400**	100	**40053567**	100	-

ELAND was run using a wide range of input sequence lengths (17-32) and the number of matches, repeat matches and non-matched sequences noted. This was graphed (Figure 2) indicating that using a sequence length of 19 was optimal for further analysis. All data was then mapped using ELAND with a sequence length input of 19. The results of this mapping are summarized in Table 1.

Analysis of four duplicate genes (*NIA* (106 nt), *CHS* (1135 nt), *PRK* (476 nt), and *MS* (394 nt)) prior to the short-read sequencing (data not shown) gave an average distance (per nucleotide) between *A. thaliana* and *Pachycladon* of 0.064 ± 0.021 substitutions per site, and an average distance between the two *Pachycladon* copies of 0.058 ± 0.023 substitutions per site. Despite this distance there were a surprising number of unique matches to the *A. thaliana* EST library (37.0% U0, U1 and U2 combined for all ELAND-19 data). These mapped short-reads can be viewed to determine coverage but they are even more useful when assembled into longer contigs. We can then use these *A. thaliana*-mapped contigs to search for potential duplicate copies within the *Pachycladon* transcriptome. As expected, the number of repeat matches was higher in the shorter length mappings and we suspect that most of these mappings were later designated as non-matches as the mapping length increased. However, these repeat matches may be useful in future analysis of repeat regions in *Pachycladon.*

3.2. *de novo Assembly*

A single FASTA file of 40 million 35-mers was used as input for the assembly with Velvet [10]. Assemblies were performed independently with k-mers having the values 15, 17, 19, 21, 23 and 25 using default parameters and all assembled contigs being returned (Figure 3), thereby covering the range of the ELAND analyses (for computational reasons Velvet only allows odd numbered k-mers). Different k-mer lengths were tested as we were unsure as to how duplicate copies of the *Pachycladon* ESTs would affect *de novo* assembly.

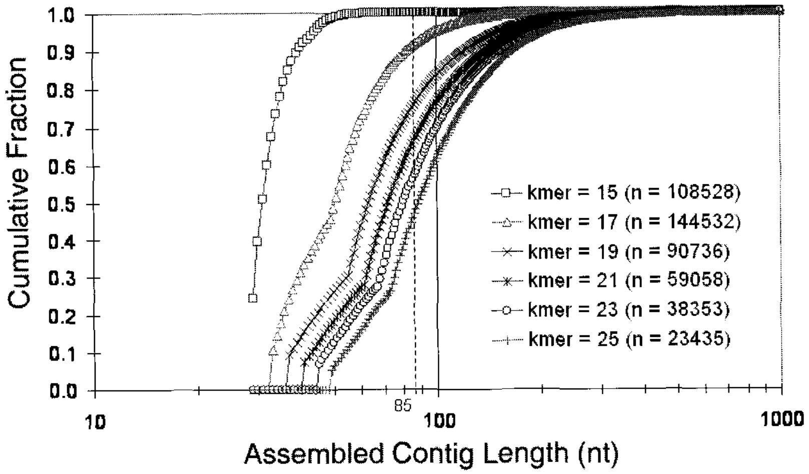

Figure 3 - Graph of *de novo* assembly results at different k-mer sizes. The length of contigs assembled under different k-mers is plotted as a cumulative fraction (the number of contigs generated is shown in the key). The dashed line at 85 nt shows the contig length cutoff that was used for further analyses (see text for details).

It can be seen from Figure 3 that the longer k-mer values resulted in longer assembled contigs but fewer of them, and that a k-mer of 15 gave very different results to all the other k-mers. Because of the previous ELAND results (Figure 2) a k-mer size of 19 was selected as optimal for the *de novo* assemblies. The resultant contigs were converted to a tab delimited form and loaded into a MySQL database in a way that kept the k-mer length as a searchable variable. The number of contigs generated and how they match to *A. thaliana* using BLAST is shown in Table 2. Contigs greater than 85 nt in length were then converted into a FASTA file. A cut-off of 85nt was chosen as it was a length where a reasonable fraction of all contigs made with a k-mer greater than 19 would be represented (34.4% of all contigs: Table 2) and analysis with all contigs becomes difficult to manage due to larger numbers of multiple lower scoring hits. From Table 2 it can be seen that the maximal number of contigs made is with a k-mer of 19, and that about 74% of all contigs have BLAST hits with a bit-score greater than 40.

Table 2 – *de novo* assembly results for contigs greater than 85 nt.

K-mer Size	**Number of Contigs**	**BLAST hits**	**Number of Contigs with BLAST hits**	**% Contigs with BLAST hits**	**Number of genes hit**		
					All alignments	**Alignment >40 nt**	**Alignment >85 nt**
15	2	3	1	50.00	3	2	0
17	12638	18844	8780	69.50	10579	9312	6786
19	22631	38535	16413	72.50	14906	13105	10304
21	20531	39554	15227	74.20	14594	12267	9504
23	16873	34486	12739	75.50	13209	10732	8222
25	12750	27360	9751	76.50	10871	8852	6595
Total	**85425**		**62911**	**73.65**			

To give an initial indication on the lengths of the resultant BLAST hits, the number of genes was calculated for all contigs irrespective of length, or where they were at least 40 or 85 nt long. Again the k-mer of 19 gives the highest number of genes hit (10304) with a contig length of at least 85 nt (Table 2). At this stage although we used *de novo* assembly, we were not attempting to completely assemble the *Pachycladon* EST transcriptome, but analyze sections of it for future experimentation.

3.3. *Gene Analysis*

The *Pachycladon* total contigs file as described in Table 2 (containing 85425 sequences), was BLASTed against the concatenated *A. thaliana* ESTs. The output was subsequently parsed with a filtering script to remove low bit-score values (less than 40) and to convert the remaining hits into a tab delimited format. To be conservative we used the set of *Pachycladon* contigs (assembled with a k-mer of 19) that mapped uniquely to a given *A. thaliana* EST, resulting in sequence alignment information for 4283 putative *Pachycladon* genes. The original *Pachycladon* contigs file (85425 sequences) was indexed using BLAST v2.16 [11] to convert it into a BLAST database and subsequently BLASTed against itself. These results were then used for duplicate gene analysis. Of the 4283 potential *Pachycladon* genes, 1155 showed evidence of overlap between contigs that mapped to a corresponding *A. thaliana* EST. The distribution of the amount of overlap for these *Pachycladon* genes is shown in Figure 4. We find 141 *Pachycladon* genes with a >100 nt overlap and 9 genes with a >300 nt overlap. These longer cases will be used in SNP and QTL analyses. One example is shown in Figure 5 where possible SNPs can be seen in the alignment. These SNPs are potentially useful for QTL analysis or estimates of genome divergence times.

We found another dataset of contigs that did not match to the *A. thaliana* ESTs but instead matched to other *Pachycladon* contigs. These contigs may represent gene copies which are different from *A. thaliana* and could be indicative of more recent duplications. Further analysis will be required to determine if this is the case. A sample of genes was viewed and analyzed in more detail for evidence that they are possibly duplicate genes from *Pachycladon*.

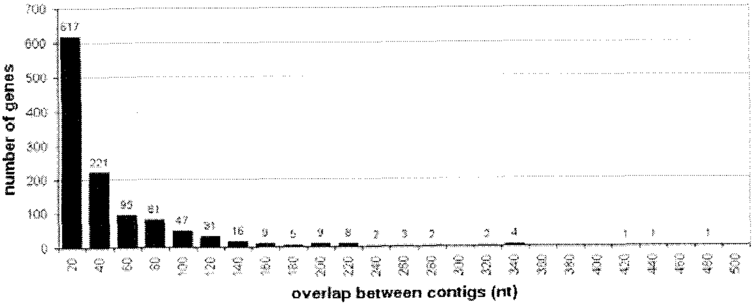

Figure 4 - Graph showing the number of *A. thaliana* ESTs to which *Pachycladon* contigs mapped with any overlap (bin size = 20 nt). For example, *Pachycladon* contigs were mapped 617 *A. thaliana* ESTs with an overlap of 1 to 20 nucleotides.

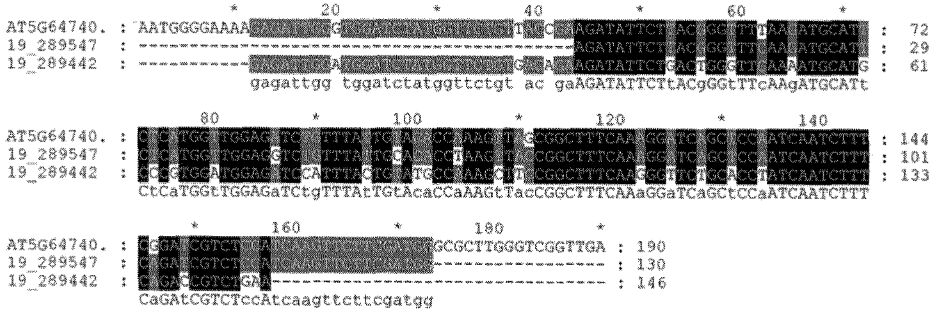
AT5G64740.
19_289547
19_289442
20
40
60
72
29
61
gagattgg tggatctatggttctgt ac gaAGATATTCTtACgGGtTTcAAgATGCATt
80
100
120
140
144
101
133
CtCaTGGtTGGAGaTCtgTTTAtTGtAcaCCaAAGtTacCGGCTTTCAAaGGaTCaGCtCCaATCAATCTTT
160
180
190
130
146
CaGAtCGTCTccAtcaagttcttcgatgg

Figure 5 - Alignment of contigs from one of the duplicate *Pachycladon* genes against the equivalent *A. thaliana* gene (AT5G64740.1 also known as CESA6 or Cellulase Synthase 6). The darker the shading, the higher the conservation between sequences and the consensus sequence is given below the alignment (absolutely conserved positions in upper case and variable positions in lower case). SNPs between the *Pachycladon* copies as well as between *Pachycladon* and *A thaliana* can be seen.

To summarize; using an optimal ELAND mapping parameter of 19 nt, 37% of the *Pachycladon* short-reads mapped uniquely to the distant reference genome of 33122 *A. thaliana* ESTs. The 33122 *A. thaliana* ESTs correlate to 28152 gene loci, of which 24292 have only a single transcript (singletons). Our results uniquely mapped to 22438 of these singleton genes (92.4%). From a total of 40 million *Pachycladon* short-read sequences, 85425 contigs were assembled *de novo* under a variety of assembly conditions. 22631 *Pachycladon* contigs were assembled under the optimal assembly parameter k-mer = 19. BLAST results with the assembled contigs identified 4283 potential *Pachycladon* genes that matched *A. thaliana* ESTs. 1155 of these *Pachycladon* genes (27%) indicated some measure of *de novo* contig overlap which will enable future duplicate gene SNP and QTL analysis.

4. Discussion

Next-generation sequencing is an emerging technology that produces millions of short-read sequences and opens the way to rapid genome analysis of non-model organisms. Although the molecular biology and mechanics of this type of sequencing are well commercialized, the bioinformatics and especially practical downstream genomic approaches are not. Researchers receiving short-read output do have software tools available for mapping and *de novo* assembly but little guidance on how to apply them. Given the high data volumes of short-read sequencing, methods which in the past worked well for traditional Sanger sequencing, may fail especially for non-model genomes. Our approach was successful in showing that a distantly related reference genome could be used for mapping and for duplicate gene analysis. Other duplicate genes, not mapped to an *A. thaliana* equivalent were found after *de novo* assembly of contigs and comparison to the contig dataset. Although this was preliminary analysis of the *Pachycladon* short-read data, we gained valuable information leading to SNP detection between the duplicate copies (and *A. thaliana* where appropriate).

The *Pachycladon* transcriptome project posed problems not only due to the non-model nature of the genome, but results had the potential to be complicated by the polyploid nature of the genome. Having multiple copies of a gene in a genome (i.e. paralogy/gene families) is common, as polyploidy (having multiple copies of a genome) is extremely common in plants. Although our approach was ultimately targeted for the finding of near exact gene copies we cannot rule out that some copies may be from recent paralogous events. This again requires further research.

We found during the course of the *Pachycladon* analysis that the viewing of data was essential to understanding the genomic issues we faced. Using Gbrowse we were able to view the potentially duplicated genes as they mapped to *A. thaliana* and evaluate the consistency in nucleotide differences seen in each gene copy. It can also be used to connect data from other sources such as prior experiments. The use of longer reads from for example, the FLX-454 sequencing platform (Roche) can only enhance both the mapping and *de novo* aspects of our approach and this is planned for future work. Many *de novo* assemblers can now use a mixture of short and longer sequences.

A key part of our approach consists of testing parameters on a single lane of data prior to complete analysis. This is essential in situations where simulations cannot be done prior to the sequencing run. The basic idea of testing subsets of data to determine mapping and *de novo* assembly parameters can be applied to other applications using short-read output especially when even the analysis of a single lane of data takes an extraordinary length of time. Researchers at present only have a limited amount of software that can reliably handle large short-read datasets. As more software becomes available the same principle of testing parameters in these cases should of course apply.

In conclusion, we show that even though these are in fact early days in the use of high-throughput short-read sequencing technology, we can move beyond the analysis of the few model or well-sequenced genomes and into the larger world of biological organisms and systems.

Acknowledgments

The authors would like to thank Peter Lockhart and the *Pachycladon* transcriptome Project team for the use of the Solexa-output, and the Genome Sequencing Facility at the Allan Wilson Centre, especially Lorraine Berry, Tim White and Maurice Collins. This work was funded by the Allan Wilson Centre and the New Zealand Marsden Fund. Claudia Voelckel holds a Feodor-Lynen Fellowship from the Alexander von Humboldt Foundation, and Simon Joly holds a post-doctoral fellowship from the National Sciences and Engineering Research Council of Canada. The authors would also like to thank Peter Lockhart and David Penny for valuable reading of this manuscript.

References

1. G. Robertson, M. Hirst, M. Bainbridge, M. Bilenky, Y. Zhao, T. Zeng, G. Euskirchen, B. Bernier, R. Varhol, A. Delaney, N. Thiessen, O.L. Griffith, A. He, M. Marra, M. Snyder, and S. Jones, Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods*,4:651-7,2007.
2. N.L. Hiller, B. Janto, J.S. Hogg, R. Boissy, S. Yu, E. Powell, R. Keefe, N.E. Ehrlich, K. Shen, J. Hayes, K. Barbadora, W. Klimke, D. Dernovoy, T. Tatusova, J. Parkhill, S.D. Bentley, J.C. Post, G.D. Ehrlich, and F.Z. Hu, Comparative genomic analyses of seventeen Streptococcus pneumoniae strains: insights into the pneumococcal supragenome. *J Bacteriol*,189:8186-95,2007.
3. J.C. Vera, C.W. Wheat, H.W. Fescemyer, M.J. Frilander, D.L. Crawford, I. Hanski, and J.H. Marden, Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Mol Ecol*,17:1636-47,2008.
4. L.D. Stein, C. Mungall, S. Shu, M. Caudy, M. Mangone, A. Day, E. Nickerson, J.E. Stajich, T.W. Harris, A. Arva, and S. Lewis, The generic genome browser: a building block for a model organism system database. *Genome Res*,12:1599-610,2002.
5. http://maq.sourceforge.net
6. R. Li, Y. Li, K. Kristiansen, and J. Wang, SOAP: short oligonucleotide alignment program. *Bioinformatics*,24:713-4,2008.
7. R.L. Warren, G.G. Sutton, S.J. Jones, and R.A. Holt, Assembling millions of short DNA sequences using SSAKE. *Bioinformatics*,23:500-1,2007.
8. W.R. Jeck, J.A. Reinhardt, D.A. Baltrus, M.T. Hickenbotham, V. Magrini, E.R. Mardis, J.L. Dangl, and C.D. Jones, Extending assembly of short DNA sequences to handle error. *Bioinformatics*,2007.
9. J.C. Dohm, C. Lottaz, T. Borodina, and H. Himmelbauer, SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing. *Genome Res*,17:1697-706,2007.
10. D. Zerbino, and E. Birney, Velvet: Algorithms for De Novo Short Read Assembly Using De Bruijn Graphs. *Genome Res*,2008.
11. S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*,25:3389-402,1997.
12. http://www.arabidopsis.org