Project No. *283496*

transPLANT

Trans-national Infrastructure for Plant Genomic Science

Instrument: **Combination of Collaborative Project and Coordination and Support Action**

Thematic Priority: FP7-INFRASTRUCTURES-2011-2

MS27
Implementation of reference-free methods for transcriptome variation analysis

Due date of milestone: August 2012
Actual submission date: August 2012

Start date of project: 1.9.2011 Duration: 48 months

Organisation name of lead contractor for this milestone: TGAC

Project co-funded by the European Commission within the Seventh Framework Programme (2011-2014)		
Dissemination Level		
PU	Public	
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Contributor

The Genome Analysis Centre (TGAC)

Introduction

Milestone reference number: MS27

The advent of inexpensive high-throughput DNA sequencing technologies opened up a variety of novel applications that have enabled more sophisticated and innovative studies. One example is the analysis of genetic variants across relatively large populations. This approach is particularly suited for species for which good quality reference genomes are available. There are, however, many examples of plant species for which we don't yet have reference genome sequences. This is particularly true for crop plants where, mainly due to their intrinsically complex genome structures, the availability of complete draft genomes is not envisaged for several years. An alternative for species lacking reference genomes is to work directly with the transcriptome. This represents a smaller and less repetitive portion of the genome and therefore is easier to work with. There are, however, specific challenges to consider such as the biases introduced by the dynamic range of expression levels as well as the ambiguities in gene families or in some cases polyploidy. The demands for denser marker panels for important downstream activities required by breeding programmes justify the investment in developing analysis pipelines to deal directly with reference-free transcriptome data. In summary the main challenges these datasets present are:

- The multiple "references" defined by genes with alternative splicing;
- bias in sequencing coverage and the effect of different expression levels; and
- the confounding effect introduced by gene families (one example in plants are the R genes associated with disease resistance).

In this milestone we summarise the state-of-the-art in the technologies and analysis tools for reference-free variant analysis in transcriptome data.

Methods

The implementation of the general method for the analysis of transcriptome variants is illustrated in Figure 1. The method is organised in two stages:

1. The initial step is to generate a consensus sequence that can be used as a "reference" for the analysis downstream.
2. In the second step the sequencing reads are mapped against the transcriptome reference to call for variants.

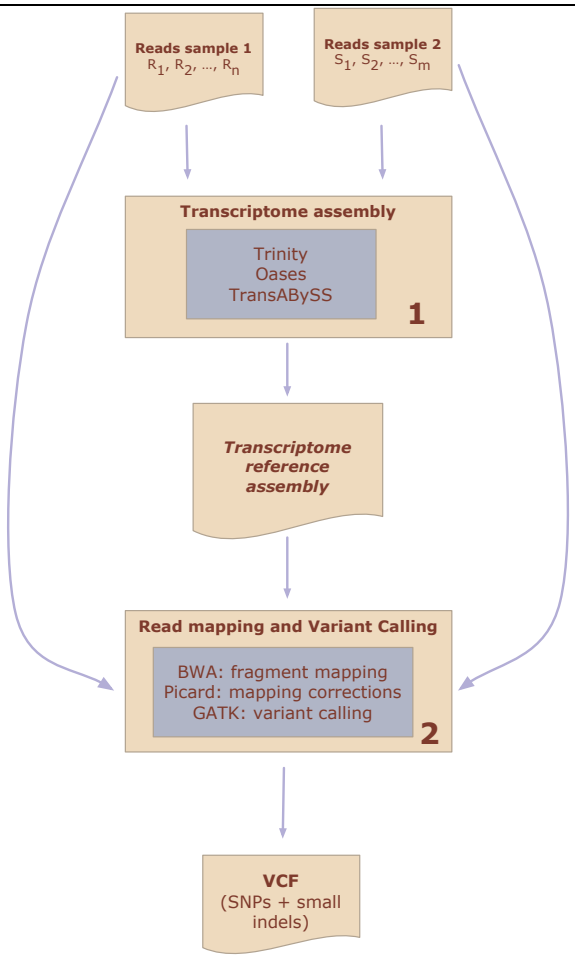

Figure 1: Strategy Overview

Evaluation samples and dataset

We have conducted the evaluation of the workflow with a transcriptome sample from two parental *Miscanthus* ecotypes. This is an example of a species which has a complex, polyploidy genome with large repeat content, and for which no reference genome is currently available. We have generated sequences for independent RNA-seq samples from stem, rhizome, pink tip and leaf for both individuals, sequenced using a standard Illumina paired end reads protocol. This has enabled us to conduct studies using different combinations of samples and tissues to produce the transcriptome reference, and perform cross-sample variant calling, to compare and validate results.

***De novo* Transcriptome assembly and its output**

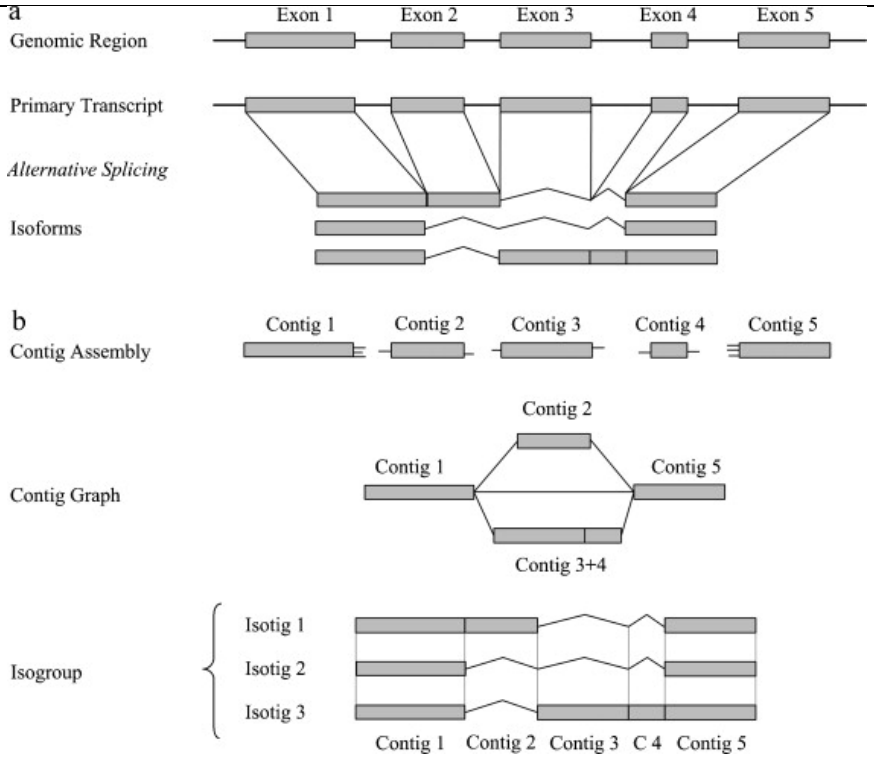

Figure 2 Transcriptome Assembly Overview

As described in the introduction the assembly of transcriptome data introduces a number of issues that are not present in the usual *de novo* assembly problem. A number of tools have been published recently which were designed to address some of these issues. For the implementation of this milestone we have evaluated the performance of three popular software packages: **Oases** [1], **TransABySS** [2] and **Trinity** [3]. The behaviour and output of these tools are substantially different, hinting to the fact that these are still early days in the development of robust algorithms and mature software. The three tools are *de Bruijn* graph-based, and they each take into account coverage variation due to expression levels and alternative splicing. The shorter nature of transcripts compared to genomes, coupled with the additional connections introduced due to alternative splicing, results in a more compact graph when compared to genomic samples. The highly linked graph structure limits the ability of the standard heuristics to detect long paths within the graph, and therefore shorter contigs are produced. Scaffolding, as performed on genomic samples, is not feasible with transcriptome data, so both Oases (based on Velvet) and TransABySS (based on ABySS) run a modified version of the contig construction step of their genomic counterparts and then continue the process from there. In both cases, a multi k-mer approach is performed by running the contig construction step on multiple k-mer sizes and then merging the results. This is a strategy to overcome the more highly connected nature and uneven coverage of these graphs, allowing transcript reconstruction at different expression levels, since transcripts with lower read depths are assembled more effectively with lower k values, and those with higher read depth are assembled more effectively with higher k values[4].
Trinity's first tool, Inchworm, deals with a single k-mer size *de Bruijn* graph construction, attempting to build the longest possible path using a coverage-based approach, aiming to assemble the most common isoform. The output is, none the less, separated into sections of sequence, roughly similar to the contigs generated by the other tools.

The three tools aim to reconstruct the transcripts and cluster them together into "isogroups" (Figure 2) . With Trinity this is performed in a two-phase approach. First Chrysalis groups together all of the contigs belonging to each isoform (isotig) and then creates a set with these contigs and all associated reads. This effectively creates a "partitioned problem" for the next stage to resolve. Then Butterfly processes each cluster independently and reports all the alternative full-length transcripts reconstructed from the set. Oases creates a contig graph for the whole set of contigs produced by the different k-mer Velvet runs and uses a specially designed scaffolding heuristic to connect contigs into "isogroups", starting with long contigs, then incorporating smaller contigs afterwards. Trans-ABySS, on the other hand, works by evaluating its contigs and the reads mapping to them, progressively generating fusions of contigs, annotating the splicing events each fusion represents. It should be noted that this approach is greatly improved by having some kind of reference.

Both Trans-ABySS and Oases are more sensitive, reconstructing more rare transcripts than Trinity[1]. The Trinity approach, however, also finds new transcripts not produced by the other two assemblers. If one, in fact, wants to reconstruct the largest possible set of transcripts, running the three different assemblers and creating a consolidated superset of their output is a reasonable approach, which also indicates the immaturity of this particular field.

Trans-ABySS is a lot more computationally expensive to run on deep-coverage samples, for one of our examples taking approximately 10 times the amount of processing time needed for Oases.

Each one of the assemblers reports both contigs and the reconstructed transcripts, and each one has a different output format to indicate how a transcript has been put together from the contigs. This opens the question of whether variant calling should be performed directly over the contigs or over the transcripts, and in each case some measure of expression levels should be taken into account. **It is important to acknowledge the fact that references generated by these tools are not of comparable quality to full-length cDNAs generated by previous technologies.**

Mapping of RNA-seq reads and variant calling

Once a reference set of contigs/transcripts is available the next step is to map the raw sequencing reads. There are a number of challenges, specific to trancriptome data, which must be considered:

- **Coverage and expression tags**
- **Splice-site junctions**
- **Effect of ploidy**
- **Multiple matches when using transcripts**

There are several pipelines developed to call for variants in genomics sequence that can be directly applied to the detection of variants from transcriptome data. A well-established strategy is to run BWA[5] for read mapping, followed by a duplication marking and realignment of the matches using Picard Tools (http://picard.sourceforge.net/, unpublished), and then a SNP and small indel variant call using GATK[6]. As future work we will evaluate the performance of this approach (mapping reads from one individual to the transcriptome assembly of the other), using the *Miscanthus* samples. We will then compare the different variant calls for the runs and produce a consolidated variant set, and further validate this approach using the *Brassica napus* dataset from Trick *et al.*[7] to compare the SNP predictions to the sets of validated SNPs.

References

1. Schulz, M. H., Zerbino, D. R., Vingron, M. & Birney, E. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics (Oxford, England)* **28**, 1086–92 (2012).

2. Robertson, G. *et al.* De novo assembly and analysis of RNA-seq data. *Nature methods* **7**, 909–12 (2010).

3. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology* **29**, 644–52 (2011).

4. Surget-Groba, Y. & Montoya-Burgos, J. I. Optimization of de novo transcriptome assembly from next-generation sequencing data. *Genome research* **20**, 1432–40 (2010).

5. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)* **25**, 1754–60 (2009).

6. DePristo, M. a *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics* **43**, 491–8 (2011).

7. Trick, M., Long, Y., Meng, J. & Bancroft, I. Single nucleotide polymorphism (SNP) discovery in the polyploid Brassica napus using Solexa transcriptome sequencing. *Plant biotechnology journal* **7**, 334–46 (2009).

Results (if applicable, interactions with other workpackages)

Publications