Mapping and Analysis of Illumina Reads for Transcriptome of *Medicago truncatula* During the Early Organogenesis of the Nodule

Alexandre Boscari[1*], Alberto Ferrarini[2], Jennifer del Giudice[2], Luca Venturini[2], Massimo Delledone[2], Alain Puppo[2*]

[1]Institut National de la Recherche Agronomique, Unité Mixte de Recherche 1355, Centre National de la Recherche Scientifique, Unité Mixte de Recherche 7254, and Université de Nice-Sophia Antipolis, Unité Mixte de Recherche Institut Sophia Agrobiotech, Sophia Antipolis, France;
[2]Università degli Studi di Verona, Dipartimento di Biotecnologie, Centro di Genomica Funzionale, Verona, Italy
*For correspondence: alexandre.boscari@sophia.inra.fr; puppo@unice.fr

[Abstract] *Medicago truncatula* serves as a model plant for legume genetics and genomics. We used RNA-Seq to characterize the transcriptome during the early organogenesis of the nodule and during its functioning. We generated approximately 135.5 million high-quality 36-bp reads, which were then aligned with the *M. truncatula* genome sequence (Mt3.0 version) and with sequences from a custom splice-junction database, for the detection of transcribed regions and splicing sites. Mapping and analysis of the reads conducted to the detection of 37,333 expressed transcription units (TUs), 1,670 had never been described before and were functionally annotated. We identified 7,595 new transcribed regions, mostly corresponding to 5' and 3' UTR extensions and new exons associated with 5,264 previously annotated genes. We also assessed the complexity in the nodulation transcriptome by performing a Cufflinks analysis to determine the frequency of the various alternatively spliced forms. Thus, we identified 23,164 different transcripts derived from 6,587 genes. Finally, we carried out a differential expression analysis, which provided a comprehensive view of transcriptional reprogramming during nodulation.

All Illumina sequence data have been deposited in the NCBI short-read archive, and Sanger-sequenced PCR products have been deposited in GenBank (SRA048731). Assembled contigs longer than 200 bp have been deposited at TSA (JR366937-JR375780). Coverage data are available at http://ddlab.sci.univr.it/cgi-bin/gbrowse/medicago/.

Materials and Reagents

1. mRNA-Seq 8 sample prep kit (Illumina, catalog number: RS-100-0801)
2. QIAquick Gel Extraction Kit (QIAGEN, catalog number: 28704)
3. RNeasy Plant Mini Kit (QIAGEN, catalog number: 74903)
4. Standart agarose for electrophoresis (Sigma-Aldrich, catalog number: A9539)

Equipment

1. Bioanalyzer Chip DNA 1000 series II (Agilent)
2. Gel electrophoresis apparatus
3. Illumina Genome Analyzer II (Illumina, model: SY-301-1201)

Software

1. Bowtie (Langmead *et al*., 2009)
2. BEDTools suite (Quinlan and Hall, 2010)
3. TopHat (Trapnell *et al*., 2009)
4. Cufflinks (Roberts *et al*., 2011)
5. Velvet (Zerbino and Birney, 2008)
6. CAP3 (Huang and Madan, 1999)
7. GMAP (version 2012-04-21) (Wu and Watanabe, 2005)
8. Cistematic 2.5 (http://cistematic.caltech.edu/)
9. ERANGE software (3.1) (Mortazavi *et al*., 2008)
10. Medicago3.py: script to import Medicago annotation into cystematic (Boscari *et al*., 2012)
11. Gff2knowngene.pl: script to convert from General Feature Format (GFF) format to UCSC knowngene format (Zenoni *et al*., 2010)
12. CASAVA (Illumina)
13. GenomeStudio (Illumina)

Procedure

1. Poly(A) mRNA was isolated from the extracted RNA to prepare a nondirectional Illumina RNA-Seq library with mRNA-Seq 8 sample prep kit. We modified the gel extraction step by dissolving excised gel slices in QG buffer of QIAquick Gel Extraction Kit at room temperature to avoid under representation of AT-rich sequences.
2. Quality control and quantification of each library of 200 bp was performed with a Bioanalyzer Chip DNA 1000 series II.
3. 36 to 44 bp sequences were generated on an Illumina genome analyzer II. A total of 135 million of reads were obtained for the different conditions.
4. *M. truncatula* Genome and the Splice Database Sequence alignments were generated with Bowtie (http://bowtie-bio.sourceforge.net).
5. Alignment of the reads was made on the Mt3.0 version of the *M. truncatula* genome sequence (www.medicago.org). For our analysis we allowed up to 2 mismatches, and

sequences that matched with more than 10 different loci were discarded. Genome index is built with command “bowtie-build –f medicago.fasta genome” where medicago.fasta is the complete Mt3.0 fasta file. Reads are then aligned with command “bowtie -v 2 –m 10 –k 11 -S genome reads.fastq output.sam”, where reads.fastq are the raw sequencing reads of a sample, and the output.sam was processed using the software suite BedTools (http://code.google.com/p/bedtools/) in order to assign each read to an exon, intron, UTR, or intergenic region. Reads mapped onto external exons fell within a 3-kb catchment from both ends of a gene, promoting the investigation of putative undiscovered exons. Intergenic reads represented those sequence reads that fell outside this catchment. The program ERANGE defined potentially novel clusters of expression on the basis of their alignment; they were categorized as novel sections (exons/UTRs) of a known gene if they fell inside a radius of 3,000 bps from them (average gene density/2). The remaining expressed clusters were marked as potential new genes.

6. In order to identify potential new isoforms of known genes, we remapped all reads against the *M. truncatula* genome using TopHat (http://tophat.cbcb.umd.edu/) with a segment length of 16 due to the short length of our reads, and defined the new isoforms of known genes performing a Cufflinks (http://cufflinks.cbcb.umd.edu/) analysis on each sample, with standard parameters, followed by an analysis with Cuffcompare to merge transcripts identified on different samples. We used the latest genome sequence and annotations provided by the Medicago research community (Mt3.5, http://www.medicago.org/).
7. To identify novel transcribed regions, we used the reads which had not been mapped against the Mt3.0 sequence from every sample to assemble separately our contigs, using the Velvet program (http://www.ebi.ac.uk/~zerbino/velvet/), using a sensitive hash length of 29 for the reads with a length of 44 bps and of 21 for the rest. The contigs were subsequently clustered together using the software CAP3 (http://bioinformatics.ca/links_directory/tool/9319/cap3-sequence-assembly-program), with a minimum overlap of 90%, requiring an overlap identity of 80%. Contigs mapping against the reference genome with identity ≥ 90% and coverage ≥ 90% after BLAT alignment were discarded from further analysis. All the contigs were also mapped against the accompanying RNA-Seq data of the Mt3.5 version with GMAP (version 2012-04-21). The contigs, with an alignment coverage on the sequence length >= 90% and on the identity >= 90%, were merged together using the program mergeBed from the BEDTools suite (http://code.google.com/p/bedtools/).
8. The evaluation of gene expression was performed with the ERANGE software (3.1), available at http://woldlab.caltech.edu/RNA-Seq. ERANGE requires Cistematic 2.5 to

execute RunStandardanalysis.sh. Therefore, a Python script (medicago3.py) was developed to import *M. truncatula* reference sequence (Mt3.0) and annotation in General Feature Format (GFF) into Cistematics Genomes sqlite database, and a Perl script (gff2knowngene.pl) was used to convert the GFF annotation file to the knowngene.txt file used by RunStandardanalysis.sh. ERANGE reports the number of mapped reads per kilobase of exon per million mapped reads, measuring the transcriptional activity for each gene. To obtain an accurate measure of gene expression not biased by reads mapping to splice junctions in genes with many introns, ERANGE considers both reads mapping to genome or to the custom splice junctions database. ERANGE was preferred over commercial packages such as CASAVA and GenomeStudio platform from Illumina because of its open nature. This allowed us to adapt and reuse code for our own analysis with greater flexibility than a comparable closed source commercial package.

9. Differential Gene Expression Statistic for RNA-Seq. ERANGE software computes the normalized gene locus expression level (named RPKM) by assigning reads to their site of origin and counting them. In the case of reads that match equally well to several sites, ERANGE assigns them proportionally to their most likely site(s) of origin (Mortazavi *et al*., 2008). The RPKM value for a given gene locus can be estimated as follows:

   $RPKM = N/(L * N_{Tot}) * 10^9$

   Where N = number of mapping reads at a given gene locus, L = estimated length (bp) of the gene locus, N_{Tot} = number of total mapping reads, and 10^9 is correspond to 1,000 bp transcript multiple 1 million reads. The null hypothesis of no differential gene expression for each gene was tested using the R package qvalue (Storey, 2002; Storey and Tibshirani, 2003; Dabney *et al*., 2010) on the R working environment. False Discovery Rates were calculated based on p-values obtained running a t test on the raw read counts using the basic R package stats.

   Variance = (1/RPKM1) + (1/RPKM2)

   $Stat = (\log(RPKM1/N_{Tot1}) - \log(RPKM2/N_{Tot2}))/\sqrt{(Variance)}$

   p.value = (1-pnorm(abs(Stat))) * 2

   The threshold value for the FDR was 0.001 and genes were first selected using this filter.

   Differentially expressed genes were then filtered again based on a Fold Change (FC) > 2.

Acknowledgments

This protocol is adapted from Boscari *et al*. (2013).

References

1. Boscari, A., Del Giudice, J., Ferrarini, A., Venturini, L., Zaffini, A. L., Delledonne, M. and Puppo, A. (2013). Expression dynamics of the *Medicago truncatula* transcriptome during the symbiotic interaction with *Sinorhizobium meliloti*: which role for nitric oxide? *Plant Physiol* 161(1): 425-439.
2. Huang, X. and Madan, A. (1999). CAP3: A DNA sequence assembly program. *Genome Res* 9(9): 868-877.
3. Langmead, B., Trapnell, C., Pop, M. and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10(3): R25.
4. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 5(7): 621-628.
5. Quinlan, A. R. and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26(6): 841-842.
6. Roberts, A., Pimentel, H., Trapnell, C. and Pachter, L. (2011). Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics* 27(17): 2325-2329.
7. Trapnell, C., Pachter, L. and Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25(9): 1105-1111.
8. Wu, T. D. and Watanabe, C. K. (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* 21(9): 1859-1875.
9. Zenoni, S., Ferrarini, A., Giacomelli, E., Xumerle, L., Fasoli, M., Malerba, G., Bellin, D., Pezzotti, M. and Delledonne, M. (2010). Characterization of transcriptional complexity during berry development in *Vitis vinifera* using RNA-Seq. *Plant Physiol* 152(4): 1787-1795.
10. Zerbino, D. R. and Birney, E. (2008). Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res* 18(5): 821-829.

