Transcriptome analysis in the post-genomic era

Faccioli P., Ciceri G.P., Provero P., Stanca A.M., Morcia C., Terzi V.

in

Molina-Cano J.L. (ed.), Christou P. (ed.), Graner A. (ed.), Hammer K. (ed.), Jouve N. (ed.), Keller B. (ed.), Lasa J.M. (ed.), Powell W. (ed.), Royo C. (ed.), Shewry P. (ed.), Stanca A.M. (ed.).
Cereal science and technology for feeding ten billion people: genomics era and beyond

Zaragoza : CIHEAM / IRTA
Options Méditerranéennes : Série A. Séminaires Méditerranéens; n. 81

2008
pages 421-426

Transcriptome analysis in the post-genomic era

P. Faccioli*[1], G.P. Ciceri*, P. Provero, A.M. Stanca*, C. Morcia* and V. Terzi***
*CRA, Experimental Institute for Cereal Research, Via S. Protaso 302,
I-29017 Fiorenzuola d'Arda (PC), Italy
**Molecular Biotechnology Center and Dept. of Genetics, Biology and Biochemistry,
University of Turin, Via Nizza 52, I-10100 Torino, Italy

SUMMARY – The advent of high-throughput sequencing tools and bioinformatics has allowed a whole-genome analysis approach to gene expression, shifting the focus from single genes to genomes. Expressed Sequence Tags (ESTs) databases have thus been created for several plant species and species-specific Gene Indices have been developed with the final aim to cluster the raw EST sequences into groups of related transcripts. In such a scenario, the integration of *in silico*-wet methods plays a fundamental role in the process that goes from data to information. Here we reported a recently published example of such a working strategy applied on advance gene expression analysis.

Transcriptome sequencing: ESTs and Gene Indices

The advent of high-throughput sequencing tools and bioinformatics has allowed a whole-genome analysis approach to gene expression, shifting the focus from single genes to genomes. In particular, Expressed Sequence Tags (ESTs) databases are being created for several plant species because they add information on the expressed part of the genome, thus representing a valuable tool in a wide range of applications, from the theoretical aspects of plant biology to the breeding process (Faccioli *et al.*, 2001). Redundancy is a general property of EST dataset (Marra *et al.*, 1998): for this reason Gene Indices have been developed with the final aim to cluster the raw EST sequences into groups of related transcripts, thus providing a more queryable and biologically meaningful dataset (Yuan *et al.*, 2001). The TIGR gene index (www.tigr.org) is an example of such EST-based species-specific database (Table 1) and it is constructed by first clustering then assembling EST and annotated sequences (Quackenbush *et al.*, 2001). This process gives a set of unique, high fidelity virtual transcripts (TC, Tentative Consensus). TC sequences thus represent a fundamental resource for plant functional genomics and they have been previously used to provide information on the abundance of gene transcripts in cDNA libraries (Stekel *et al.*, 2000), for the identification of groups of potentially related genes (Faccioli *et al.*, 2005) and recently for candidate housekeeping identification (Faccioli *et al.*, 2007).

Table 1. Gene Indices available for the cereal research community

Gene index	www address	Cereal species
TIGR Gene Indices	www.tigr.org	Barley, maize, sorghum, rice, rye, wheat
NCBI Unigene	www.ncbi.nlm.nih.gov	Barley, maize, sorghum, rice, wheat
Plant GDB	www.plantgdb.org	Barley, maize, oat, rice, rye, sorghum, wheat

Transcriptome analysis: making sense of gene-expression data

Advanced gene expression analysis methods, such as microarray and RT-Real Time PCR, as well as more traditional ones, such as Northern blot, require efficient normalization to be informative. Normalization requires adjustment of expression data to permit comparisons among different samples.

Traditionally housekeeping genes, so called because they encode proteins mediating basic cellular functions and are thus synthesized in all cell types, have been employed as reference genes for

[1] Corresponding author: primetta.faccioli@entecra.it.

normalization both in RT-Real time PCR (Table 2) and arrays (Table 3 for GeneChip details). To the best of our knowledge, there are few examples of studies specifically concerned with housekeeping gene expression analysis in plants and very often they are devoted to the evaluation or validation, in the specific species and experimental condition of interest, a list of literature-based, well known reference genes. Recently novel internal controls for normalization have been identified in *Arabidopsis* (Czechowski *et al.*, 2005) via a genome-wide screening, revealing that there are many genes other than the ones traditionally used that are more stably expressed. In the procedure described by Faccioli *et al.* (2007), the analytical approach for the identification of candidate reference genes is effective and very simple conceptually and has several advantages. Firstly, it does not start from a list of literature-based potential housekeeping genes. Furthermore genes without a known function can be selected from the TC collection. Secondly, the necessary calculations are very simple and are based on a plain frequency counting but, despite of this simplicity, the results are very encouraging as demonstrated by lab-based validation. The procedure can be performed in several species for which a Gene Index, organized on a significant number of cDNA libraries and ESTs sequences, is available.

Table 2. Common housekeeping genes used as references in RT-real Time PCR

Housekeeping gene	Species	References
Tubulin	Barley	Close *et al.* Plant Physiology (2004), 134: 960-968
		Ozturk *et al.* Plant Molecular Biology (2002), 48: 551-573
		Burton *et al.* Plant Physiology (2004), 134: 224-237
		Suprunova *et al.* Plant, Cell and Environment (2004), 27: 1297-1308
		Svensson *et al.* Plant Physiology (2006), 141: 257-271
	Wheat	Remoto and Sasakuma. Phytochemistry (2002), 61: 129-133
	Potato	Nicot *et al.* Journal of Experimental Botany (2005), 56 (No. 421): 2907-2914
	Sugarcane	Iskandar *et al.* Plant Molecular Biology Reporter (2004), 22: 325-338
	Arabidopsis	Czechowski *et al.* Plant Physiology (2005), 139: 5-22
GAPDH	Barley	Close *et al.* Plant Physiology (2004), 134: 960-969
		Svensson *et al.* Plant Physiology (2006), 141: 257-270
		Burton *et al.* Plant Physiology (2004), 134: 224-236
	Wheat	Travella *et al.* Plant Physiology (2006), 142: 6-20
		Crismali *et al.* BMC Genomics (2006), 7: 267
	Rice	Bo-Ra *et al.* Biotechnology Letters (2003), 25: 1869-1873
	Sugarcane	Iskandar *et al.* Plant Molecular Biology Reporter (2004), 22: 325-339
	Arabidopsis	Czechowski *et al.* Plant Physiology (2005), 139: 5-17
Actin	Barley	Close *et al.* Plant Physiology (2004), 134: 960-969
		Svensson *et al.* Plant Physiology (2006), 141: 257-271
	Wheat	Crismani *et al.* BMC Genomics (2006), 7: 267
	Soybean	Byfield *et al.* Crop Science (2006), 46: 840-846
	Sunflower	Clèment *et al.* Plant Molecular Biology (2003), 52: 1025-1036
	Potato	Nicot *et al.* Journal of Experimental Botany (2005), 56 (No. 421): 2907-2919
	Sugarcane	Iskandar *et al.* Plant Molecular Biology Reporter (2004), 22: 325-337
	Tomato	Coker *et al.* Physiologia Plantarum (2005), 124: 311-322
	Arabidopsis	Czechowski *et al.* Plant Physiology (2005), 139: 5-20
Translation initiation factor 5A	Rice	Close *et al.* Plant Physiology (2004), 134: 960-970
Elongation factor 1alfa	Wheat	Crismani *et al.* BMC Genomics (2006), 7: 267
	Rice	Jain *et al.* Biochemical and Biophysical Research Communications (2006), 345: 646-652
	Potato	Nicot *et al.* Journal of Experimental Botany (2005), 56 (No. 421): 2907-2915
	Arabidopsis	Czechowski *et al.* Plant Physiology (2005), 139: 5-18
Ribosomal protein L2	Potato	Nicot *et al.* Journal of Experimental Botany (2005), 56 (No. 421): 2907-2915
18 S rRNA	Barley	Walia *et al.* Functional Integrative Genomics (2006), 6: 143-156
	Rice	Bo-Ra *et al.* Biotechnology Letters (2003), 25: 1869-1872
	Potato	Nicot *et al.* Journal of Experimental Botany (2005), 56 (No. 421): 2907-2919
Adenine phosphoribosyl transferase	Potato	Nicot *et al.* Journal of Experimental Botany (2005), 56 (No. 421): 2907-2919
Cyclophilin	Barley	Burton *et al.* Plant Physiology (2004), 134: 224-239
	Wheat	Crismani *et al.* BMC Genomics (2006), 7: 267
	Potato	Nicot *et al.* Journal of Experimental Botany (2005), 56 (No. 421): 2907-2919
Polyubiquitinin	*Arabidopsis*	Czechowski *et al.* Plant Physiology (2005), 139: 5-19
Ubiquitin 5	Rice	Jain *et al.* Biochemical and Biophysical Research Communications (2006), 345: 646-651
Heat shock protein 70	Barley	Burton *et al.* Plant Physiology (2004), 134: 224-239

Table 3. The Affymetrix GeneChips are designed specifically to monitor gene expression in several model plants and crops. The majority of these arrays were created in collaboration with leading researchers through the Affymetrix GeneChip ®Consortia Program. The sequence information for the majority of these arrays were selected from EST and cDNA clustering databases. In addition to GeneChip arrays that quantitate quantify known and annotated transcripts, a GeneChip® Arabidopsis Tiling 1.0R Array is designed for whole-genome experiments

Plant species	Product name	Probe pairs/probe set	Number of genes or TCs or transcripts	Reference database	Housekeeping/control genes
Arabidopsis thaliana	Arabidopsis Genome Array	16	8,300 genes	GenBank	Actin, GAPDH, 25SrRNA, 5SrRNA.
Arabidopsis thaliana	Arabidopsis ATH1 Genome Array	11	24,000 genes	TIGR (ATH1-121501)	Actin, GAPDH, ubiquitin
Hordeum vulgare	GeneChip® Barley Genome Array	11	25,500 contigs and singletons	HarvEST Triticeae v0.95 and higher	Ubiquitin, GAPDH, tubulin, translation initiation factor 5A
Citrus	GeneChip® Citrus Genome Array	11	33,879 Citrus transcripts	Citrus HarvEST EST and cDNA clustering db	GAPC, ß-actin, UBQ11
Gossypium hirsutum, *G. raimondii,* *G. arboretum,* *G. barbadense*	GeneChip® Cotton Genome Array	11	21,854 transcripts	*Gossypium hirsutum* Unigene (2 August 2006), *Gossypium raimondii* UniGene (2 September 2005), GenBank, dbEST, RefSeq.	Sucrose synthase, actin, polyubiquitin
Zea mays	GeneChip® Maize Genome Array	15	14,850 genes	NCBI's GenBank (up to September 29, 2004), *Zea mays* UniGene Build (July 23, 2004)	GAPDH, actin, cyclophilin, ubiquitin, 18SrRNA, ef1a
Medicago truncatula, *M. sativa,* *Sinorhizobium meliloti*	GeneChip® Medicago Genome Array	11	Not specified	TIGR *M. truncatula* Gene Index (January 2005), International *Medicago* Genome Annotation Group (IMGAG), *S. meliloti* genome, *M. sativa* EST (TIGR)	ß-actin, GAPDH, glutathione-S-transferase, ubiquitin
Populus sp.	GeneChip® Poplar Genome Array	11	56,000 transcripts and gene predictions	UniGene Build #6 (March 16, 2005), Gene Bank mRNAs and ESTs for all *Populus* species (April 26, 2005), Gene set v.1.1 from *Populus* genome project (*P. trichocarpa*) from JGI US Dpt Energy	GAPDH, ACTB, ef1A1

Table 3 (cont.). The Affymetrix GeneChips are designed specifically to monitor gene expression in several model plants and crops. The majority of these arrays were created in collaboration with leading researchers through the Affymetrix GeneChip ®Consortia Program. The sequence information for the majority of these arrays were selected from EST and cDNA clustering databases. In addition to GeneChip arrays that quantitate quantify known and annotated transcripts, a GeneChip® Arabidopsis Tiling 1.0R Array is designed for whole-genome experiments

Plant species	Product name	Probe pairs/probe set	Number of genes or TCs or transcripts	Reference database	Housekeeping/control genes
Oryza sativa (*japonica* and *indica* varieties)	GeneChip® Rice Genome Array	11	51,279 transcripts	GenBank mRNAs, TIGR Os1 v.2, NCBI UniGene Build #52 (May 7, 2004), International Rice Genome Sequencing.	GAPDH, actin, cyclophilin, ubiquitin, 18SrRNA, 27SrRNA, ef1a, 25SrRNA, 5.8SrRNA
Glycine max, Phytophtora sojae, Heterodera glycines.	GeneChip® Soybean Genome Array	11	37,500 soybean transcripts, 15,800 *P. sojae* transcripts, 7,500 *H. glycines* transcripts	GenBank, dbEST, UniGene Build 13 (November 5, 2003)	18SrRNA, actin, GSTA A, cyt.P450, SBP, ubiquitin
Saccharum officinarum	GeneChip® Sugar Cane Genome Array	11	6,024 genes	*S. officinarum* UniGene Build 5 (August 27, 2004), GenBank mRNAs (up to November 2, 2004)	Actin, ef1a, GAPDH
Lycopersicon esculentum	GeneChip® Tomato Genome Array	11	9,200 transcripts	*L. esculentum* UniGene Build#20 (October 3, 2004), GenBank mRNAs (up to November 5, 2004)	ß-actin, GAPDH, elongation factor 1, 17SrRNA, 25SrRNA, glutathione-S-transferase, phytocrome B2, ubiquitin
Vitis vinifera, other *Vitis* species	GeneChip® *Vitis vinifera* Genome Array	16	14,000 *V. vinifera* transcripts, 1,700 other *Vitis* species transcripts.	GenBank, dbEST, RefSeq, UniGene (October 7, 2003)	ß-actin, GAPDH, elongation factor 1-α
Triticum aestivum, T. monococcum, T. turgidum, Aegilops tauschii	GeneChip® Wheat Genome Array	11	55,052 transcripts	*T. aestivum* UniGene Build#38 (April 24, 2004), GenBank mRNAs from other species (May 18, 2004)	Ubiquitin, 18SrRNA, G6PDH, cyt. P450, sucrose synthase, actin, ef1a, GAPDH

Moreover, previously suggested statistical methods can be applied on the selected set of candidate housekeeping genes, for the identification of genes showing minimal variation across a variety of experimental conditions (Table 4).

Because a "good reference gene for all experiments" does not exist, lab validation of each reference gene on the specific physiological condition/tissue of interest is necessary to avoid unexpected changes in gene expression that could result in erroneous conclusions, particularly when subtle differences are considered.

Table 4. Some examples of freely available software based on excel platform that allow the assessment of multiple reference genes for real-time RT-PCR normalisation

Program	How does it work?	Reference	Website
geNorm	geNorm determines the most stable housekeeping genes from a set of tested genes in a given cDNA sample panel, and calculates a gene expression normalization factor for each tissue sample based on the geometric mean of a user-defined number of housekeeping genes	Vandensompele, J., De Preter, K., Pattyn, F., Poppe, B., Van Roy, N., De Paepe, A. and Speleman, F. (2002). Accurate normalization of real-time quantitative RT-PCR data by geometric avering of multiple internal control genes. *Genome Biol.*, 3: 0034.I-0034.II.	http://medgen.ugent.be/~jvdesomp/genorm/
BestKeeper	BestKeeper determines the best suited standards, out of ten candidates, and combines them into an index. The index can be compared with further ten target genes to decide, whether they are differentially expressed under an applied treatment. The software uses geometric mean of raw data	Pfaffl, M.W., Tichopad, A., Prgomet, C. and Neuvians, T.P. (2004). Determination of stable housekeeping genes, differentially regulated target genes and sample integrity: BestKeeper-Excel-based tool using pair-wise correlations. *Biotechnol Lett.*, 26: 509-515.	http://www.gene-quantification.info/
Norm-Finder	Norm-Finder measures the variation and ranks the potential reference genes in different experimental conditions	Andersen, C.L., Jensen, J.L. and Orntoft, T.F. (2004). Normalization of real-time quantitative reverse transcription-PCR data: A model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets. *Cancer Res.*, 64: 5245-5250.	http://www.mdl.dk/publicationsnormfinder.htm

References

Czechowski, T., Stitt, M., Altmann, T., Udvardi, M.K. and Scheible, W. (2005). Genome-wide identification and testing of superior reference genes for transcript normalization in *Arabidopsis*. *Plant Physiology*, 139: 5-17.

Faccioli, P., Ciceri, G.P., Provero, P., Stanca, A.M., Morcia, C. and Terzi, V.A. (2007). Combined strategy of "in silico" transcriptome analysis and web search engine optimization allows an agile identification of reference genes suitable for normalization in gene expression studies. *Plant Molecular Biology, Plant Molecular Biology*, 63: 679-688.

Faccioli, P., Lagonigro, M.S., De Cecco, L., Stanca, A.M., Alberici, R. and Terzi, V. (2002). Analysis of differential expression of barley ESTs during cold acclimatization using microarray technology. *Plant Biology*, 4: 630-639.

Faccioli, P., Pecchioni, N., Cattivelli, L., Stanca, A.M. and Terzi, V. (2001). Expressed sequence tags from cold-acclimatized barley identify novel plant genes. *Plant Breeding*, 120: 497-502.

Faccioli, P., Provero, P., Herrmann, C., Stanca, A.M., Morcia, C. and Terzi, V. (2005). From single genes to co-expression networks: Extracting knowledge from barley functional genomics. *Plant Molecular Biology*, 58(5): 1-12.

Marra, M.A., Hillier, L. and Waterston, R.H. (1998). Expressed sequence tags – ESTablishing bridges between genomes. *Trends in Genetics*, 14: 4-7.

Quackenbush, J., Cho, J., Lee, D., Liang, F., Holt, I., Karamycheva, S., Parvizi, B., Pertea, G., Sultana, R. and White, J. (2001). The TIGR Gene Indices: Analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Research*, 29(1): 159-164.

Stekel, D.J., Git, Y. and Falciani, F. (2000). The comparison of gene expression from multiple cDNA libraries. *Genome Research*, 10: 2055-2061.