Computational analysis of the cassava transcriptome, gene discovery and regulatory element prediction

Germán Plata[1], Fausto Rodríguez-Zapata[1], Tetsuya Sakurai[2], Motoaki Seki[2], Andrés Salcedo[1], Joe Tohme[1], Yoshiyuki Sakaki[2], Atsushi Toyoda[2], Atsushi Ishiwata[2], Kazuo Shinozaki[2] and Manabu Ishitani[1].

[1]International Center for Tropical Agriculture (CIAT), A.A. 6713, Cali, Colombia
[2]RIKEN Research Institute, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, 230-0045, Japan
Contact: gaplata@cgiar.org

Introduction

The first full-length enriched cDNA library of abiotic stress induced genes in cassava (Sakurai et al, 2007) was sequence-characterized using ESTs. The analysis of these sequences revealed high gene diversity in the library, including many elements involved in stress response and several biochemical pathways relevant for agronomic traits. Further analysis of these sequences, including the prediction of transcription factor families and small-RNA binding sites, may provide a comprehensive list of candidate genes for the dissection of gene regulation in cassava and the improvement of this crop.

Methods

Table 1. Conditions used for mRNA extraction.

Treatment	Age	Tissue	Duration of treatment (h)
No treatment	9, 11, 12 weeks	leaf	
No treatment	9 month	root	
Drought shock	7 weeks	leaf	3,6,24,72
Heat	9 weeks	leaf	3,6,24,72
PPD	9 month	root	24,48,120
High Al, low pH	9 weeks	leaf	3,6,24,72
High Al, low pH	9 month	root	6,24,48

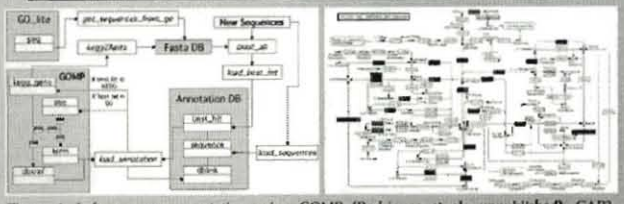


Figure 1. Left, sequence annotation using GOMP (Rodriguez et al., unpublished). CAP3 assembled sequences were compared to known protein sequences and mapped to Gene Ontology terms and KEGG pathways using BLASTX, the annotation was transferred and stored onto a mySQL relational database for querying. Right, starch metabolism pathway with cassava genes (red) mapped to it, Arabidopsis genes not captured in the library are shown in green.

Comparative genomics - Reciprocal runs of TBLASTX were performed between the cassava sequences and the transcripts of *Ricinus communis*, *Populus trichocarpa* and *Arabidopsis thaliana* (figure 2, left). Recent gene duplications were defined as motifs in the network of best blast hits with two or more cassava sequences sharing the same best hit in at least two other species (Figure 2, right).

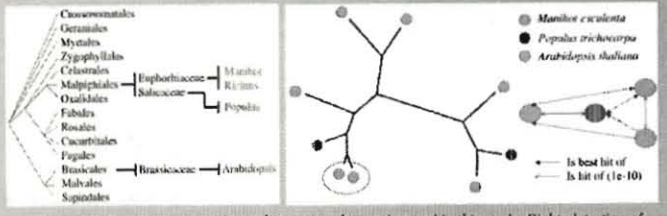


Figure 2. Left, evolutionary relationship among the species used in this study. Right, detection of recent duplications in cassava. The unrooted parsimony tree for Monodehydroascorbate reductase is shown as an example.

Transcription factor and miRNA target predictions – Sequences with significant BLAST hits to plant transcription factors (TFs) were scanned for Pfam TF protein motifs using Hmmer. TF families were then built according to the definitions used by the Plant Transcription Factor Database (http://plntfdb.bio.uni-potsdam.de/v2.0/). For miRNA target prediction 339 plant miRNAs from mirBase and 34 miRNA precursors from a computational prediction pipeline were compared to the cassava sequences using Vmatch (www.vmatch.de).

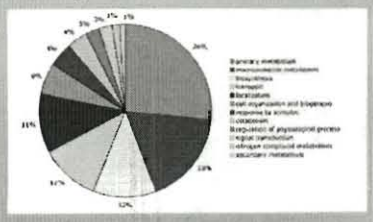
Results

Functional annotation - The ESTs from the full-length cDNA library were assembled into 10577 scaffolds, out of wich a function was assigned to 8227 (78%). There are 114 KEGG pathways for Arabidopsis thaliana, using these as reference, we were able to map cassava transcripts to 101 of them (Figure 3).

We found that 5'UTRs of the cassava transcripts had on average lower folding free energies, were longer and had higher GC% than those of Populus and Arabidopsis; altogether this suggest a higher post-transcriptional regulation of cassava stress genes.

Figure 3. Percentage of cassava genes mapped to Arabidopsis pathways. On average 61% of the enzymes on each pathway were captured in the full-length cDNA library. The most complete pathways include: Glycolysis/Gluconeogenesis (100%), Pentose phosphate pathway (93%), Carbon fixation (92%), Starch and sucrose metabolism (76%), Stilbene, coumarine and lignin biosynthesis (73%) and Biosynthesis of steroids (70%).

Gene duplications - 230 gene duplications were detected, the annotation of these sequences revealed that many of them were in the biosynthesis and response to stimulus categories (Figure 4, left), in particular many enzymes related to oxidative stress, a common theme in plant response to heat, drought, acidic soil-aluminum and post-harvest physiological deterioration (PPD) stresses were included in this group. Important enzymes in H_2O_2 scavenging were found to be duplicated as well as enzymes involved in signal transduction and stress reaction (Figure 4, right).



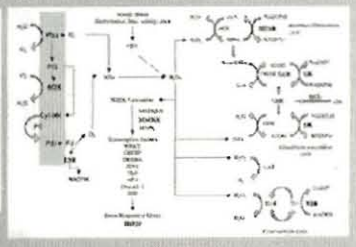


Figure 4. Left, Gene Ontology anotatin summary for 230 candidate gene duplications in cassava. Right, Reactive oxygen species metabolism, cassava gene duplications are shown in bold and underlined, these include: AOX, Alternative Oxidase; FNR, Ferredoxin NADPH Reductase; MAPKK, Mitogen Activated Protein Kinase Kinase; MDAR, Monodehydroascorbate reductase; GLR, Glutaredoxin; GR, Glutathione reductase; GCL, Glutamate Cysteine Ligase; NTR, NADPH Thioredoxin Reductase (Based on Mittler *et al.* 2004).

Transcription factors and miRNA targets - We found 355 sequences showing significant similarity to known transcription factors and with at least one TF protein motif identified through the HMM search. These included members of 48 out of 68 TF families; some of the largest gene families identified inluded NAC, AP2-EREBP and MYB, whose members are known to play role in stress tolerance (Figure 5, left).

CASSAVA FL-cDNA TRANCRIPTION FACTOR FAMILIES

ABI3VP1	5	C3H	26	HSF	8	SBP	0
AlfIn-like	0	CAMTA	2	LFY	0	Sigma70-like	5
AP2-EREBP	24	CCAAT	13	LIM	4	SRS	1
ARF	1	CPP	0	MADS	6	TAZ	0
ARR-B	6	CSD	6	MYB	32	TCP	6
BBR/BPC	0	DBP	11	MYB-related	2	Trihelix	0
BES1	3	E2F-DP	1	NAC	21	TUB	9
bHLH	11	EIL	1	NOZZLE	0	ULT	0
bHSH	0	FHA	2	Orphans	0	VOZ	0
bZIP	9	G2-like	0	PBF-2-like	1	WRKY	7
C2C2-CO-like	19	GeBP	3	PLATZ	7	zf-HD	3
C2C2-Dof	9	GRAS	11	Pseudo	0	ZIM	0
C2C2-GATA	7	GRF	0	RWP-RK	0		
C2C2-YABBY	5	HB	10	S1Fa-like	2		
C2H2	13	HRT	0	SAP	0		

OTHER TRANCRIPTIONAL REGULATORS

ARID	0	HMG	6	MBF1	4	SET	3
AUX/IAA	0	Jumonji	4	PHD	10	SNF2	1
DOT	0	LUG	0	RB	0		

TOTAL transcripts in families: 355

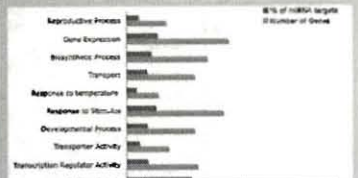


Figure 5. Left, members of transcription factor families captured in the cassava full-length cDNA library. Right, main functional categories of putative microRNA targets in cassava.

Using Vmatch to pair both, mirBase and computationally predicted microRNAs to cassava ESTs, we detected 160 microRNAs or microRNA precursors that could regulate the expression of 328 putative targets. The functional annotation of these genes revealed an important proportion of transcription factors homologues and proteins involved in stress response (Figure 5, right).

Conclusions and perspectives

Large-scale sequence analysis of cassava genes provides not only a catalog of annotated sequences that are suited for functional analyses and association studies, but also information that can be mined for pathway reconstruction, regulatory element prediction, training of gene prediction software and experiment design for expression analysis. This library as well as a second full-length cDNA library that is on the make, should prove valuable for the post-genomic era of cassava research.

Acknowledgements

We thank all of the technical staff of the Sequencing Technology Team at RIKEN GSC for their assistance. We also thank Martin Fregene and Hernán Ceballos for providing information related to the used genotypes.

References

- Sakurai, T., Plata, G., Rodriguez, F., Seki, M., Salcedo, A., Toyoda, A., Ishiwata, A., Tohme, J., Sakaki, Y., Shinozaki, K., Ishitani, M. 2007. Sequencing analysis of 20,000 full-length cDNA clones from cassava reveals lineage specific expansions in gene families related to stress response. - *BMC Plant Biology* 7:66
- Rodríguez-Zapata, *et al.* 2008. GOMP: Gene ontology and metabolic pathway integration. *In preparation*
- Mittler, *et al.* 2004. Reactive oxygen gene network of plants. *Trends in Plant Science*, 19:490-498