The potato tuber transcriptome: analysis of 6077 expressed sequence tags

Meg Crookshanks[1], Jeppe Emmersen[1], Karen G. Welinder*, Kåre Lehmann Nielsen

Institut for Bioteknologi, Aalborg Universitet, Sohngaardsholmsvej 49, DK-9000 Aalborg, Denmark

Received 29 August 2001; accepted 2 September 2001

First published online 18 September 2001

Edited by Julio Celis

Abstract This is the first report of the biosynthetic potential of a tuber storage organ investigated by expressed sequence tag sequencing. A cDNA library was generated from the mature tuber of field grown potato (*Solanum tuberosum* var. Kuras). Partial sequences obtained from 6077 clones were assembled into 828 clusters and 1533 singletons. The average read length was 592 bp, and 2254 clones were full length. 5717 clones showed homology to genes from other organisms. Genes involved in protein synthesis, protein destination and cell defense predominated in tuber compared to stolon, shoot and leaf organs. 1063 clones were unique to tuber. Transcripts of starch metabolizing enzymes showed similar relative levels in tuber and stolon.

Key words: Expressed sequence tag sequencing; Potato bioinformatics; Starch related enzyme; Tuber gene; Unknown gene, characteristics of; *Solanum tuberosum*

1. Introduction

Potato (*Solanum tuberosum*) is the fourth major crop plant worldwide. It is grown both for food and as an important source of high quality starch. It is also of major interest in plant based medicine and relatively easy to modify genetically [1,2]. Here, we focus on the repertoire of genes expressed in the mature tuber of potato (var. Kuras), an organ with a key role in the storage of starch and protein [3,4]. Potato is one of approximately 200 plant species capable of tuber production. Tuberization in potato is the process by which an underground shoot, the stolon, differentiates to form a specialized storage organ, the tuber [4–8]. Over many years, there has been a great deal of effort focused on unraveling the nature of the stimuli and the molecular events underpinning the tuber growth and development [8].

This is the first paper describing the expression profiles of genes from tuber in potato (GenBank dbEST accession numbers BF153198–BF154323, BF459480–BF460469, BG349972–BG351995, BI405291–BI407114). We have sought to summarize and classify 6077 mature tuber expressed sequence tags (ESTs) using the functional classification of the Munich Information Center for Protein Sequences (MIPS) for *Arabidopsis thaliana* [9,10]. Further comparison at the species level was made by including other EST libraries from potato developing stolon, shoot and leaf [11].

2. Materials and methods

2.1. RNA extraction and library construction

Field grown potato tuber (var. Kuras) was harvested at the end of flowering, washed in 0.5% sodium dodecylsulfate, cut into pieces and frozen in liquid nitrogen in the field. RNA was extracted from 5 g as described by Scott et al. [12]. Poly(A) mRNA selection (MagneSphere technology, Promega, Madison, WI, USA), and cDNA synthesis (Stratagene) were carried out by standard procedures, except for using a different 5′ linker adapter (5′-AATTCGGCTCGAGG). cDNA was size fractionated by gel filtration and cloned unidirectionally into the λZAPII vector. The resulting DNA was packed into λ phages using Gigapack III Gold (Stratagene). From the initial plating the library was estimated to contain 10^5 clones. An aliquot of the library was amplified, followed by in vivo excision of the pBluescript SK(−) phagemid. The average insert size was 1.5 kb.

2.2. DNA sequencing

An aliquot of the excised, amplified library was used for infecting *Escherichia coli* SOLR cells of OD_{600} 1.0 and subsequently plated on LB agar containing ampicillin. The resulting colonies were picked into a 96 well culture plate and grown for 10 h at 37°C and 200 rpm. Glycerol was added to a final concentration of 15% and a backup plate was created. Plates were stored at −45°C. Polymerase chain reaction (PCR) products suitable for sequencing were generated from 0.5 μl of defrosted bacterial glycerol stock as template and T3-EST1 (AATTAACCCTCACTAAAGGG) and M13-21 (TGTAAAACGACGGCCAGT) as primers (present in the pBluescript vector arms). The PCR included 95°C 3 min, 95°C 30 s, 53°C 30 s, 72°C 105 s for 35 cycles and a final extension at 72°C 7 min. The control of size and quality of the PCR products was performed by gel electrophoresis of a representative number of samples from each plate. Excess primers and nucleotides were removed by enzymatic digestion using 5 U exonuclease I (New England Biolabs) and 0.3 U of shrimp alkaline phosphatase (Amersham Pharmacia), 37°C for 60 min, followed by inactivation of the enzymes at 80°C for 20 min. The resulting PCR product was then used as template for a sequence reaction using 5 pmol of a nested primer (GTGGCGGCCGCTCTAGAA) 38 bp upstream of the cDNA insert, and dye terminator cycle sequencing chemistry. For each reaction, 4 μl of PCR product was used (50–100 ng DNA) in a total reaction volume of 12 μl. Sequencing reactions were subjected to 95°C 20 s, 57°C 15 s, 60°C 1 min for 30 cycles and 60°C for 5 min. These were cleaned by Sephadex G50 (DNA grade, Amersham Pharmacia) in filter plates (Millipore MAHV N45) prior to capillary electrophoretic separation and detection by a MegaBace 1000 (Amersham Pharmacia).

2.3. Sequence processing and analysis

A custom PERL script processed sequence files automatically. This script linked sequence backup, basecalling by Phred (trimming option on, cut-off set to 0.05; CodonCode), discarding sequences shorter than 150 bp, and vector trimming by Cross Match (CodonCode) into one routine. DNATools [13] was used to automatically BLAST and analyze results, build EST submission files for the dbEST (GenBank dbEST accession numbers BF153198–BF154323, BF459480–

*Corresponding author. Fax: (45)-9814 1808.
E-mail address: welinder@bio.auc.dk (K.G. Welinder).

[1] These authors contributed equally to this work.

Abbreviations: EST, expressed sequence tag; BLAST, basic local alignment search tool; MIPS, Munich Information Center for Protein Sequences; TIGR, The Institute for Genomic Research

0014-5793/01/$20.00
PII: S0014-5793(01)02888-5

BF460469, BG349972–BG351995, BI405291–BI407114), and edit sequences. It was also used to build a searchable flat database containing sequences and BLAST results. BLASTX searches and putative identification were carried out locally because of speed. A 600 bp sequence was blasted against 660 000 non-redundant GenBank protein entries in 25 s using a 1100 MHz AMD CPU with 768 MB RAM. Inverted sequences and sequences originating from *E. coli* and *Lambda* inserts were removed. Contigs were built with the edited sequences using Phrap (CodonCode): phrap > readslog.txt-revise_greedy-confirm_score 40-vector_bound 10-maxgap 10.

2.4. Functional analysis of EST sequences

The MIPS functional classification applied to *Arabidopsis* genes [9,10] was adapted for potato. Translated potato ESTs were sorted into 12 functional groups and an unclassified group by sequence comparison to classified *Arabidopsis* proteins using an *E*-value cut-off at 10^{-5} [14]. Some were assigned more than one function in agreement with the homologous *Arabidopsis* proteins. All *Arabidopsis* protein sequences were downloaded from TIGR in batches of separate functional class [11]. These flat file databases were concatenated into one file, and a BLAST searchable database called At-Class was built with formatdb.

3. Results and discussion

3.1. Data quality, statistics, clustering and characteristics of unidentified genes

Sequencing of the potato tuber cDNA library and sequence processing gave rise to 6077 high quality ESTs (Table 1). The average Phred score for the library was 25, therefore the probability of one base being called correctly was between 99% and 99.9%. The tuber dbEST represents up to 2361 different genes, 828 clusters assembled from two or more ESTs and 1533 singletons. The tuber sequences have been compared to recent potato ESTs from stolon, shoot and leaf libraries (Table 1). In combination, the four libraries comprise nearly 11683 different potato genes. If potato has the same number of protein encoding genes as *Arabidopsis* (25 498) [10], then the present collection represents 46% of all potato genes. The number of active genes in the mature tuber is rather low as expected for a storage organ, i.e. approximately half of those in the vegetative stolon, shoot and leaf organs.

We found an unexpected bias in average read length and G+C content of 'unknown' genes. When setting the *E*-value to a cut-off at 10^{-2}, 360 or 6% of the tuber sequences had no homology to any sequences in the non-redundant database.

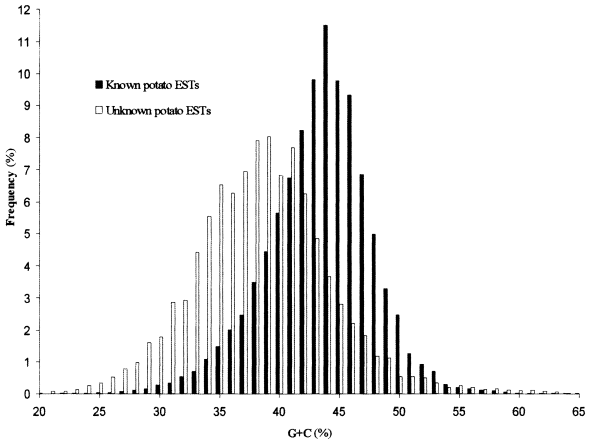

Fig. 1. Differential distribution in the G+C content of all known and unknown potato EST sequences. 'Known' indicates that the translated sequences match a protein from another organism with an *E*-value of 10^{-2} or lower.

The average length and G+C composition of the unknowns in all four libraries were significantly lower than for the total number of ESTs (Table 1). Fig. 1 illustrates the different G+C distributions in known and unknown potato genes that amounted to 4% on the average. In *Arabidopsis* coding and non-coding DNA, the G+C content was approximately 44 and 33%, respectively [10]. The difference in frequency distribution indicates that a substantial portion of the unknown sequences in all the potato EST databases may contain non-coding regions such as introns or untranslated regions. Presumably this is true not only for potato.

3.2. Function of expressed potato genes

Approximately 40% of the potato ESTs were assigned a function by aligning to *Arabidopsis* proteins with an *E*-value of 10^{-5} or lower (Fig. 2). The frequency of sequences which either did not align to any translated *Arabidopsis* gene or aligned to the unclassified group for *Arabidopsis* (31% of *Arabidopsis* genes [10]) varied from 54% in mature tuber to 63% in leaves. Even though this large fraction was not assigned a functional role, the general result was that this classification produced physiologically meaningful differences among the

Table 1
Analysis of potato EST libraries

Library	Tuber	Stolon	Shoot	Leaf	All
Total ESTs	6077	10388	8733	10451	35648
Clusters	828	1911	1582	1851	5463
Singletons	1533	2986	2995	3152	6220
Redundancy (%)[a]	74.8	71.0	66.0	70.0	82.5
Full length[b]	2254	4159	3471	3529	13413
ESTs unique to organ	1063	2464	1999	3833	
Homologues (EST #)					
Arabidopsis	4358	7192	6612	6984	25146
All organisms	5717	8784	7789	8605	30895
Unknowns	360	1604	944	1846	4753
Average length (bp)					
Total ESTs	592	455	634	450	520
Unknowns	416	408	559	402	436
G+C content (%)					
Total ESTs	43.5	43.2	43.0	42.0	43.1
Unknowns	41.1	38.9	39.2	38.2	39.3

Tuber: this work. Stolon, shoot, leaf: www.tigr.org, May 2001.
[a]Redundancy = ESTs assembled in clusters/total ESTs.
[b]Full length EST starts within 10 amino acids from the initiating Met of the homologous protein. Unknowns not included.

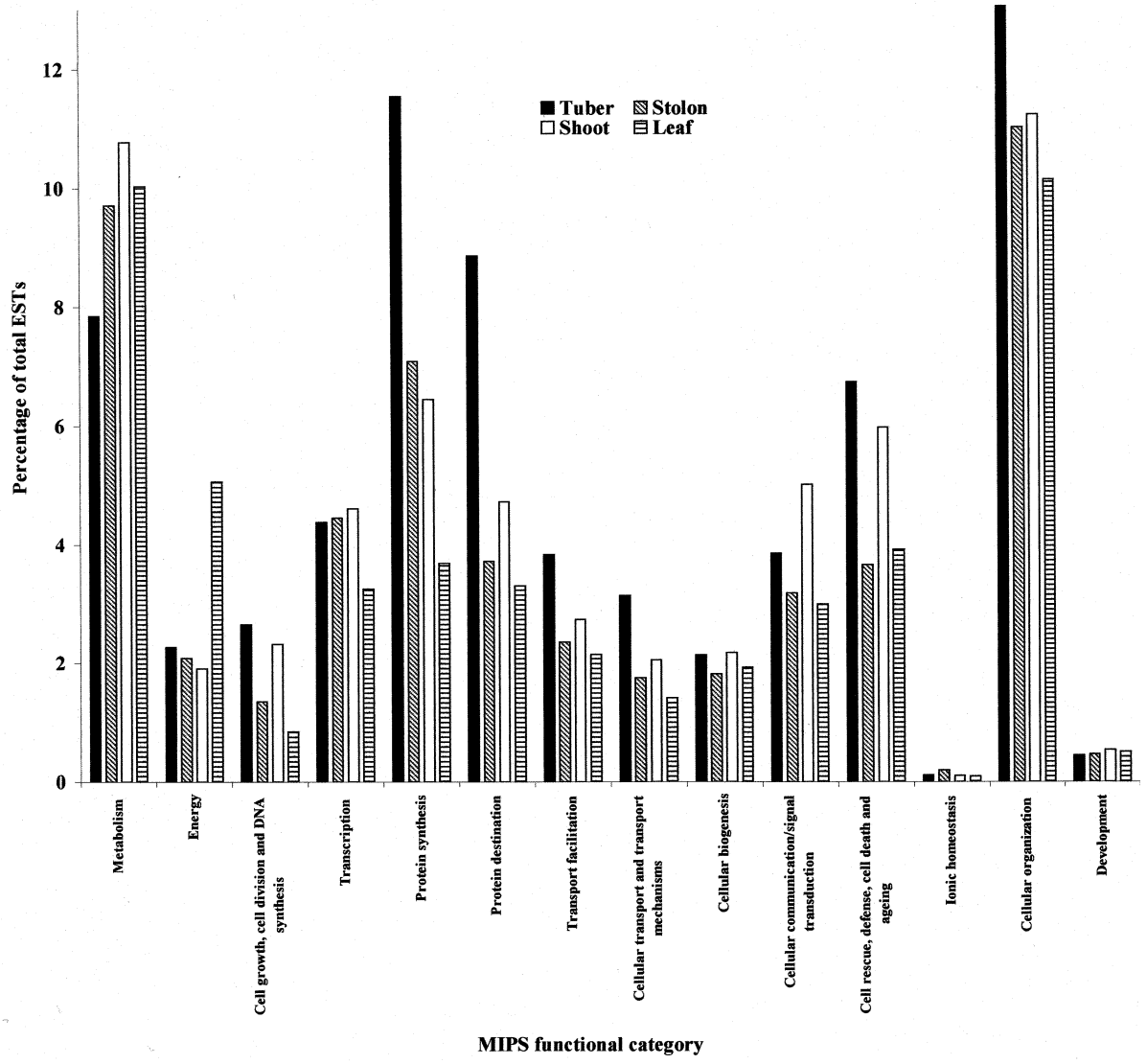

Fig. 2. Distribution of EST sequences from *S. tuberosum* libraries among functional categories. Approximately 40% of potato ESTs were assigned a function by alignment to *Arabidopsis* genomic protein sequences using an *E*-value cutoff at 10^{-5}. However, 31% of *Arabidopsis* proteins remain unassigned [10]. 28% of tuber, 31% of stolon, 24% of shoot, and 33% of leaf ESTs had no *Arabidopsis* homologue.

four different organs. Mature tuber showed a relatively higher percentage of ESTs in the functional categories Protein destination, Protein synthesis, and Cell defense (shared with shoot), with a relatively lower percentage of sequences in Metabolism. This finding was consistent with the function of the tuber as a storage organ of both starch and protein [3]. The high percentage of leaf ESTs in the Energy class was consistent with the physiological role of the leaves in photosynthesis, the main organ of energy capture and metabolism.

It should be noticed that some sequences were assigned multiple functions and, therefore, the sum of functional class percentages was 120% for tuber, 113% for stolon, 118% for shoot, and 112% for leaf. A striking example of multiple class assignment was Cellular organization, this represented a total of 4.8% of the mature tuber EST clones if the ESTs were only annotated with the highest scoring *Arabidopsis* sequence (data not shown). When multiple functions were assigned to one gene, this class represented a total of 12.5% mature tuber

Table 2
Transcripts predominant in the tuber EST library

Protein	Accession number	ESTs in cluster (%)
Proteinase inhibitor, PIE	T07414	209 (3.4)
Patatin class I	CAA25592	185 (3.0)
Elongation factor 1α	P17786	155 (2.6)
Aspartic proteinase inhibitor	S24186	143 (2.4)
DnaJ-like	T07371	126 (2.1)
Probable cysteine proteinase inhibitor 8	T07750	121 (2.0)

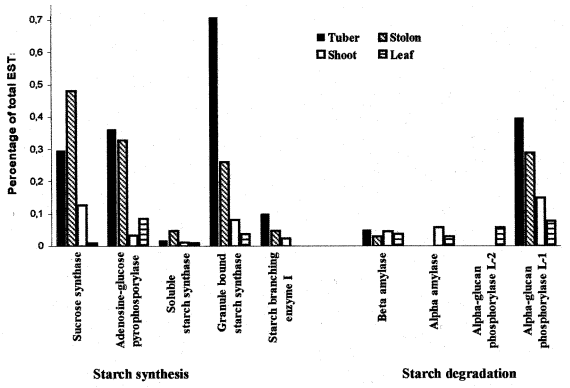

Fig. 3. Relative abundance of enzymes involved in starch metabolism. Only starch branching enzyme I was found. Starch branching enzyme II was not seen in any library.

EST clones. The family of elongation factors was assigned to both Cellular organization and Protein synthesis.

3.3. Genes highly expressed in tuber, and genes involved in starch metabolism

A major purpose of the EST analysis of potato tuber was to generate a global view of the biosynthetic capacity of this storage organ. Table 2 lists the six most highly expressed genes observed among the randomly picked EST clones. Other genes constituted ≤1%. Three types of protease inhibitors accounted for a total of 7.8% of EST clones. Patatin class I genes accounted for 3.0%, which is markedly less than the 25–40% patatin protein that can be found in tuber [15,16]. Overall, it appears that the tuber was still capable of storage protein production.

Fig. 3 compares the abundance of ESTs for enzymes involved in starch biosynthesis and degradation in the four organs. The mature tuber has significant mRNA levels of granule bound starch synthase, ADP-glucose pyrophosphorylase, sucrose synthase and α-glucan phosphorylase L-1, very similar to stolon. Furthermore, tuber and stolon share similar low levels of soluble starch synthase, starch branching enzyme I and β-amylase.

4. Conclusions

The EST data provides a global perspective to the biology and biosynthetic capacity of the potato plant, and a new foundation for analysis of gene function with clones for more than 11 000 different potato genes available. Rational plant improvement, the design of improved starch composition, novel biomolecules including carbohydrates and recombinant proteins for medical uses have come closer.

Acknowledgements: M.C. was funded by the Danish Natural Science Research Council (9901834), and the project was supported by EFRU (1999-503/0009-146) and the Danish Technical Research Council (26-00-0141). Technical assistance was provided by J. Poulsen and A. Nygaard.

References

[1] Yu, J. and Langridge, W.H. (2001) Nature Biotechnol. 19, 548–552.
[2] Horsch, R.B., Fry, J.E., Hoffmann, N.L., Eichholtz, D., Rogers, S.G. and Fraley, R.T. (1985) Science 227, 1229–1231.
[3] Hannapel, D.J. (1990) Plant Physiol. 94, 919–925.
[4] Hannapel, D.J. (1991) Physiol. Plant 83, 568–573.
[5] Gregory, L.E. (1956) Am. J. Bot. 43, 281–288.
[6] Chapman, H.W. (1958) Physiol. Plant 11, 215–224.
[7] Prat, S., Frommer, W.B., Hofgen, R., Keil, M., Kossmann, J., Koster-Topfer, M., Liu, X.J., Muller, B., Pena-Cortes, H. and Rocha-Sosa, M. et al. (1990) FEBS Lett. 268, 334–338.
[8] Jackson, S.D. (1999) Plant Physiol. 19, 1–8.
[9] Frishman, D., Albermann, K., Hani, J., Heumann, K., Metanomski, A., Zollner, A. and Mewes, H.W. (2001) Bioinformatics 17, 44–57.
[10] The Arabidopsis Genome Initiative (AGI) (2000) Nature 408, 796–815.
[11] http//www.tigr.org.
[12] Scott Jr., D.L., Clark, C.W., Deahl, K.L. and Prakash, C.S. (1998) Plant Mol. Biol. Rep. 16, 3–8.
[13] Rasmussen, S.W. (2001) www.dnatools.dk.
[14] Wilson, C.A., Kreychman, J. and Gerstein, M. (2000) J. Mol. Biol. 297, 233–249.
[15] Paiva, E., Lister, R.M. and Park, W.D. (1983) Plant Physiol. 71, 161–168.
[16] Tweel, D. and Ooms, G. (1988) Mol. Gen. Genet. 212, 325–336.