AP 01 - Dorlhac de Borne - A comprehensive survey of the *N. tabacum* transcriptome (The European Sequencing of Tobacco Project, ESTobacco)

Authors: DORLHAC DE BORNE F.; COATES S.; ROSS J.; VERRIER J.L.; JONES L.; JULIO E.; DELON R.

A comprehensive survey of the *N. tabacum* transcriptome (The European Sequencing of Tobacco Project, ESTobacco).

F. Dorlhac de Borne[1], S. Coates[2], J. Ross[2], J-L. Verrier[1]; L. Jones[2], E. Julio[1], R. Delon[1]
1. Altadis - Institut du Tabac, Domaine de la Tour, 769, route de Sainte-Alvère, F-24100 Bergerac, France.
2. Advanced Technologies (Cambridge) Ltd, 210 Cambridge Science Park, Milton Road, Cambridge, CB4, U.K.

Abstract

Tobacco genome research is expected in the next few years to improve our knowledge about the interactions between genes involved in the formation of undesirable compounds in cigarette smoke. A close link exists between gene expression in the tobacco plant, chemical composition of raw tobacco, and combustion products in cigarette smoke.

The aim of the ESTobacco project is to be complementary to other projects currently underway concerning the tobacco genome. Our strategy is to sequence only genes expressed in tobacco and not the whole genome. The size of the tobacco genome is too large to be totally sequenced (29 times more than *Arabidopsis thaliana*).

This project used 3 "commercial" varieties of tobacco widespread throughout the world: K326 for the Flue-cured type, Burley 21 and TN86 for the Burley type. In order to obtain the major genes, the organs of the plant (seeds, roots, stem, midrib, lamina and flowers) prepared at different stages of development (germination, young seedlings, before and after topping, maturity) were used as a basis for this work. A large tobacco expressed sequence tag (EST) dataset was obtained from 11 normalized cDNA libraries comprising 56 000 clones.

A DNA array designed with these sequences could allow the large scale study of the genes expressed in tobacco. This new tool will lead to the acceleration of programs already underway concerning the origins of risks associated with tobacco and inform strategies for harm reduction. In order to encourage a wide range of initiatives on tobacco plant genetic, as with other crops, the resulting sequences obtained during the ESTobacco project are available to the worldwide scientific community through public access databases.

Introduction

The ESTobacco project results from collaboration between Advanced Technologies (Cambridge) Ltd and the Tobacco Institute of Bergerac (Altadis Group). The goal of this project was initially to sequence a large number of expressed genes in tobacco and to identify

the tobacco genes linked to the formation of undesirable compounds in the smoke. Our aim in this project was to be complementary to other projects currently underway, and to give as soon as possible a large set of tobacco unigenes to the scientific community, to help the different teams which work on tobacco and to encourage other initiatives on tobacco plant. The strategy to obtain these sequences was to sequence only ESTs and not the whole genome. The large size of the tobacco genome was a real problem to achieve this task at a reasonable cost.

Creation of the libraries

To obtain a representative set of tobacco sequences, 11 libraries were created. Samples were prepared from seeds starting germination and 48h thereafter, seedlings 2 weeks after germination, roots before and after topping, lamina before topping, after topping and at maturity, the same mix for stem, and flowers before topping. All these samples originated from plants of the cultivar K326, one of the most commonly grown Flue-cured varieties worldwide. Two other libraries were prepared from leaf lamina at maturity using two widely used Burley cultivars, Burley 21 and TN86. At least five thousand clones were obtained in each library.

These libraries were normalized. To achieve this operation, the lamina before topping library was first used to identify the most represented sequences on a first plate, and these clones were removed from that library as well as from all the other ones.

Some problems were observed with the preparation of the 3 mature leaf samples. Quality of the cDNA was not at the same level as in other libraries, an important number of clones being short inserts.

The sequencing was achieved in 5' and the success rate was around 83%. A total of 56000 clones were sequenced and 46 546 sequences of good quality were released in Genbank in April 2006.

Analysis of the EST collections

On May 2, 2006 the total number of tobacco mRNA sequences available in Genbank was 76836 (*Fig*.1). 60% were released by the ESTobacco project, 25% by Riken in Japan, 8% by the Tobacco Science Research Institute of Yunnan in China, and 5% by Plant Systems Biology in Belgium. Other sequences represent 1.8% of the total.

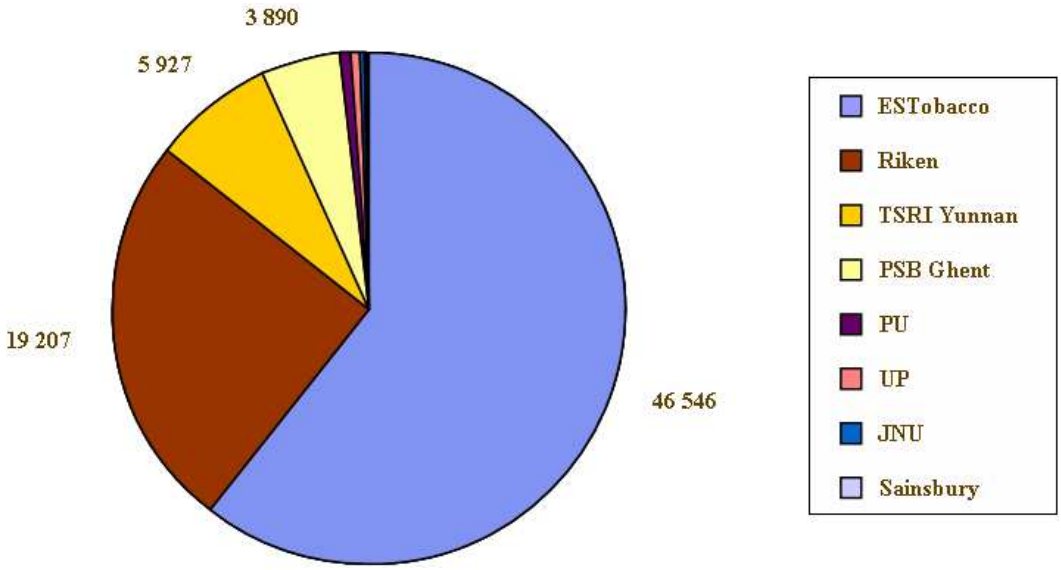

Figure 1. Tobacco sequences available in Genbank (May 2, 2006).

The average length of this total collection is around 586 bp, with 45 millions of bases sequenced. The ESTobacco project contributed with sequences having an average length of 653 bp, for a total length of more than 30 millions bp, and 41 undetermined bases. The average length was lower on the collections released by Riken and the TSRI (below 500 bp), with more than 26 000 undetermined bases.

An online pipeline analysis system provided by Kegg was used to analyze these data. The clustering is operated in 5 steps. A first step to clean the sequences, to filter polyA, low complexity and low quality sequences. A second step to mask microsatellites, transposable elements, and pseudogenes. A third step to mask vectors and other contaminations. A fourth step to mask organelles and the last step of assembling with Cap3. The parameters adopted with this software were an overlap percent identity cutoff of 95% and an overlap length cutoff of 40 bp.

Transposable elements were identified during the second step of the analysis. A high rate of retrotransposons in the Riken collection (0.82%) was observed compared to the ESTobacco collection (0.05%). This observation may be related to retrotransposition phenomenons during in vitro cultivation of the BY02 cell line from which the Riken collection is derived. A total of 425 retrotransposons were identified.

Microsatellites were identified with an additional system: the GDR SSR Server. The minimum of identified repeats was 5 for dinucleotides, 4 for trinucleotides and 3 for tetra to hexanucleotides. The total of unique sequences with at least one repeat was 3 339. A majority of trinucleotides 48.2% and dinucleotides 27.6% was observed. The percentage of polymorphism was estimated on 93 microsatellites using the tobacco genotypes studied by Julio *et al*. (2006) to create a genetic map of tobacco based on AFLP markers. 22.6% of polymorphism was observed with these varieties.

After assembling, 16 987 unigenes were obtained in the ESTobacco collection, 6 568 contigs and 10419 singletons for an average length of 769 bp. For the total set of sequences available

in Genbank, 35 759 unigenes were identified, 9598 contigs and 26161 singletons for an average length of 641 and 23 millions of bp.

SNPs or insertion/deletion were then detected in the contigs. The SNP discovery pipeline was used to identify them and verified with Contigviewer. Only 74 SNPs and 38 InDels were identified with this system. This low number is the consequence of the high number of sequences coming from K326 and BY02. On these SNPs, only 6.2% of polymorphism is detected using the same genotypes as previously.

The percentage of annotations in unigenes and the absence of open reading frame were determined by Blast2GO and ORF Predictor (*Fig*. 2). In Riken and TSRI sets around 6% of the unigenes have no open reading frame compared to 2% in ESTobacco.

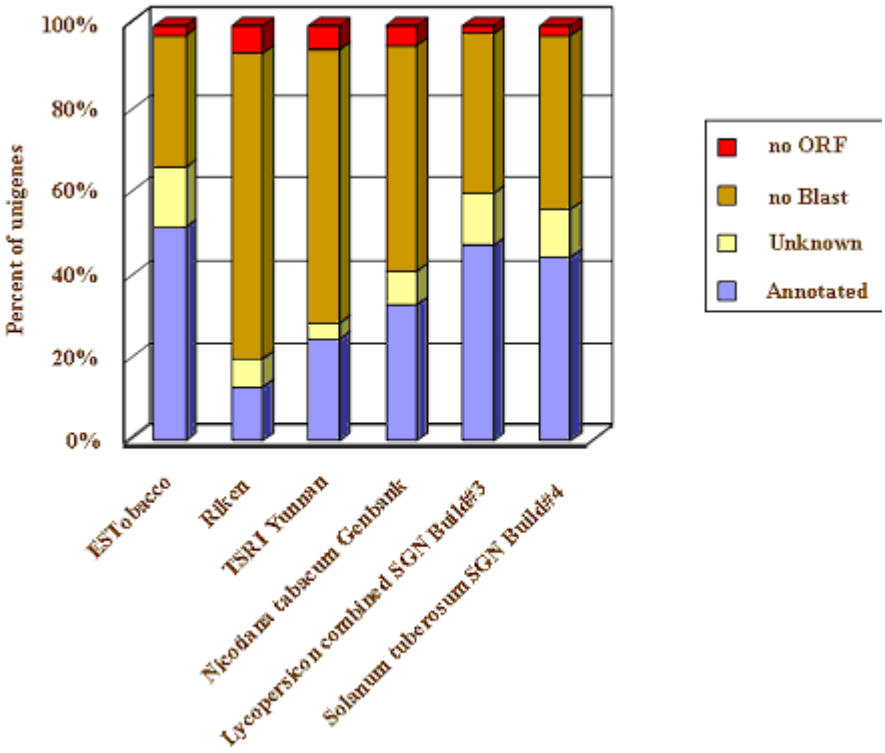

Figure 2. Unigene annotations (BlastX with an eValue < $1e^{-25}$) and ORF predictions (ORF Predictor).

The percentage of unigenes annotated was estimated after BlastX with an eValue < $1e^{-25}$. Results from ESTobacco, the SGN *lycopersicon* and the SGN *Solanum tuberosum* unigene sets are very similar, 51.4, 47.3 and 44.3% respectively. In the Riken and TSRI sets the percentages of annotated unigenes are lower, 13% and 24.5% respectively. In the total set of *Nicotiana tabacum*, the percentage of annotations is 32.7%.

Three different groups of GOslim annotations were obtained, cellular location, molecular function and biological process. A high similarity between *Lycopersicon*, *S. tuberosum* and the *N. tabacum* set is observed in the distribution inside these different groups, except for structural molecules.

The KAAS server was used to identify and place genes in metabolic pathways. This analysis was operated on the Arabidopsis data set. 161 metabolic pathways were generated. For the major metabolic pathways half of the genes were identified.

The ORF obtained with ORFpredictor were submitted to the Sanger Institute to identify PFAM domains. Except for MYB and GRAS, a high similarity is observed between *N. tabacum*, *Lycopersicon* combined and *S. tuberosum* for the 10 most common transcription factor domains (*Fig*. 3).

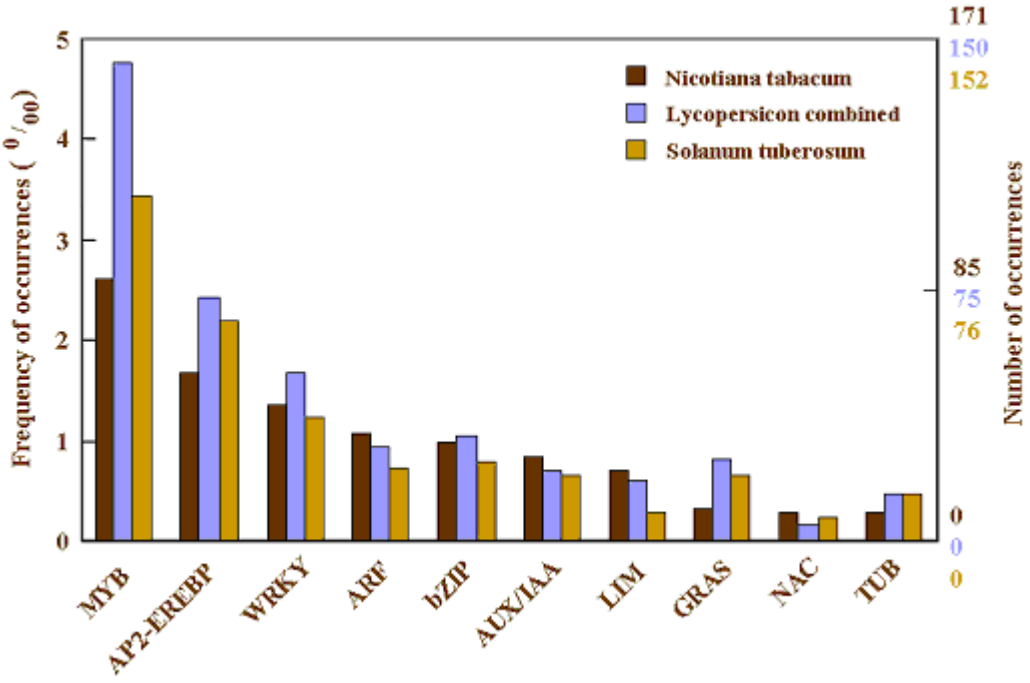

Figure 3. Most common transcription factors in *Solanaceae*.

An analysis was performed to identify for candidate COS markers. Sequences of *N. tabacum*, *Lycopersicon* combined and *S. tuberosum* were compared by BlastP on ORF. A total of 11 698 sequences are shared between *N. tabacum* and the two taxa with an eValue < $1e^{-25}$.

N. tabacum appears closer to *Lycopersicon* than *S. tuberosum*. 1409 other genes are common with *Lycopersicon* and only 1122 with *S. tuberosum*. These sequences were compared with 3 714 unique sequences of *A. thaliana*. 550 candidate COS markers were selected to have a single BlastP hit to *A. thaliana* genes. It will be interesting to use these COS markers when the genetic map will be enriched with microsatellites for a comparative study among these *Solanaceae*.

Conclusion

Global results confirm the quality of the unigene set obtained during the ESTobacco project and the high potential available to identify genes linked with harm reduction.

Candidate markers (microsatellites, SNPs and COS) will give the possibility to complete the

genetic map already available to identify new QTLs.

New tools to study transcriptome will be developed. Particularly an Affymetrix DNA microarray to study expressed genes in tobacco. Our idea is to identify as soon as possible interesting genes and to share this information with the scientific community.

To encourage a wide range of initiatives on tobacco plant genetics, as for other crops, the results obtained during this project are freely available to the worldwide scientific community on the ESTobacco website. Bacterial clones will be available soon at the CNRGV Toulouse, France.

Reference

JULIO E., VERRIER J.L. and DORLHAC de BORNE F.. Development of SCAR markers linked to three disease resistances based on AFLP within Nicotiana tabacum L. Theor. Appl. Genet., 335-346, 2006.

Acknowledgments

Martin Ward, Bernard Duméry and Jean-Paul Lebondidier for their help to initiate this project. Georges Freyssinet, Xavier Sarda, Denis Scala and all the members of Biogemma Evry to prepare the clones. All the members of ATC Cambridge and Tobacco Institute of Bergerac in the preparation of the samples.