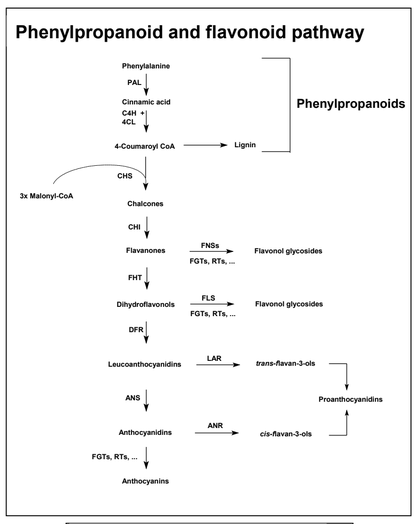
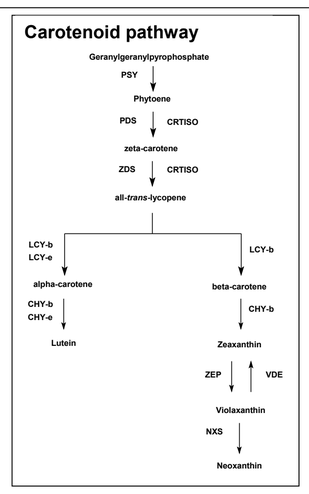
MINING THE TOBACCO GENOME INITIATIVE SEQUENCE DATABASE FOR GENES INVOLVED IN SECONDARY METABOLITE PATHWAYS

PHILIP MORRIS INTERNATIONAL

Carlo Rosati, Rutger S. van der Hoeven, Audrey Cordier, Florian Martin, Ferruccio Gadani, Paolo Donini

Philip Morris International, R&D, Applied Research and Technology Department, Quai Jeanrenaud 56, 2000 Neuchâtel, Switzerland. E-mail: carlo.rosati@pmintl.com





The North Carolina State University (NCSU) Tobacco Genome Initiative (TGI) was started in 2002 in cooperation with Philip Morris USA to gather genetic information of *Nicotiana tabacum* by means of sequencing genomic DNA and cDNA libraries of Hicks Broadleaf variety. The TGI website (http://www.tobaccogenome.org) contains related project information and a link for filing a data transfer agreement for academic researchers.

As more and more sequences are accumulated in the databases, bioinformatics analyses are needed to contribute to the understanding of genome organization and the function of genes controlling useful agronomic traits. The comparison of genetic information of tobacco and other sequenced Solanaceous and non-Solanaceous plant species will also help to unravel new genetic information on unknown and tobacco-specific genes. This effort aims at the development of effective tools to accelerate and assist conventional breeding (e.g., by marker-assisted selection) for reducing the levels of harmful constituents and improve flavour characteristics of tobacco leaf.

Specific information on genes of selected pathways leading to the formation of carotenoid or polyphenol compounds is of paramount importance for fundamental and applied research purposes. Carotenoids are metabolites playing a critical role in photosynthesis, which have been shown to generate volatile compounds upon enzymatic degradation by cleavage enzymes with different substrate specificity. Polyphenols and flavonoids are secondary metabolites playing a role in plant response to biotic and abiotic stresses, and have antioxidant properties *in vivo*.

In this study, we analysed genomic and EST sequences of structural genes of the carotenoid and flavonoid pathways, a prerequisite for understanding their metabolic pathways and developing genetic and molecular tools for improving tobacco germplasm.

The TGI sequence database

The August 2005 release of TGI dataset contains ca. 900,000 entries from both methyl-filtered genomic DNA and EST sequences from various cDNA libraries. Sequences from leaf cDNA libraries accounted for >90% total EST sequences.

Sequence type	Source	# entries
Methyl-filtered clones	genomic DNA	829,969
ESTs	cDNA libraries	65,613
	of which: different leaf libraries	59,850
	roots	2,077
	flowers	1,325
	other libraries	2,361

Bioinformatic pipeline and sequence assembly

TGI database

⇩ BLASTALL based on reference sequences (cutoff: E-11)
Sequence parsing and retrieval

Selected methyl-filtered (genomic) and EST sequences

⇩ Contig assembly (>80% identity, < 50% unaligned end overlap)

Sequence analysis

Analysis of genomic sequences

Genomic sequences were found for most structural genes of both pathways. In most cases, contigs were covering partial gene regions: additional sequences are needed to allow a higher coverage and a better analysis of the gene space in the large (4.5 Gbp) tobacco genome. For several genes (e.g. PAL, FGTs, PDS), different members of gene families were identified (cf. Tables 1 and 2). The amphidiploid background of cultivated tobacco, the existence of enzyme classes (glycosyltransferases, dioxygenases, P450s), and the presence of multigene families arising from gene duplication and evolution events account for this situation. In secondary metabolism, such gene copies can have a broad range of specificities and expression patterns to produce peculiar metabolite patterns.

For the CHY gene, comparison of the 2 complete genomic sequences (Table 2) revealed a 6-aminoacid insertion/deletion in the exon 1 and several polymorphisms in coding and non-coding regions. Consensus sequences showed polymorphisms in coding and non-coding regions, identifiable as potential haplotype tags or larger insertions/deletions.

The increasing availability of promoter and intron sequences will allow a more thorough analysis of potential binding sites of regulatory *cis*-acting factors and the ability to develop molecular markers associated for genes controlling desired traits.

Table 1. Information on flavonoid genes from genomic sequences of TGI.

Gene	Contig	# clones	Length (bp)	# introns	intron length (bp)	promoter region (bp)	terminator region (bp)	cds coverage (%)	cds identity (%)	Reference sequence
PAL	a	8	1470	1	150	-	-	49	78	*L. esculentum* M83314
	b	2	948	0	-	295	-	16	70	"
	c	2	973	1	n.d.	-	-	31	91	"
	d	3	776	1	n.d.	-	-	16	83	"
	e	2	354	0	-	128	-	12	83	"
C4H	a	6	1352	1	n.d.	402	-	51	85	*C. annuum* AF212318
4CL	a	12	2010	1	n.d.	513	-	60	84	*S.tuberosum* AF150686
	b	2	1040	1	226	-	-	13	82	"
	c	2	875	2	n.d.	-	-	13	89	"
	d	3	773	0	-	123	-	31	78	"
CHS	a	9	1000	1	n.d.	517	-	15	85	*L. esculentum* X55195
	b	2	334	1	n.d.	-	-	10	83	"
	c	8	1930	1	390	211	170	100	74	"
	d	2	690	1	n.d.	-	-	44	87	"
	e	2	980	1	n.d.	422	-	15	91	"
	f	4	1089	1	n.d.	-	-	21	80	"
	g	2	929	1	n.d.	-	-	53	83	"
	like	9	1860	1	318	439	-	98	98	*N. tabacum* Y14506
CHI	a	4	668	2	99+91	-	-	58	93	*L. esculentum* AY348871
	b	2	785	0	-	-	-	20	82	"
	c	3	800	2	n.d.	-	-	30	81	"
FHT	a	3	1216	2	n.d.+453	-	-	47	83	*P.hybrida* X60512
DFR	a	2	472	1	n.d.	321	-	10	96	*P.hybrida* X15537
	b	4	1602	1	449	580	-	24	85	"
	c	3	749	3	109+94+129	-	-	36	92	"
ANS	a	4	957	1	252			53	91	*P.hybrida* X70786
ANR	a	3	1067	2	252	146	-	45	72	*M.domestica* AY830130
FGT	a	14	2134	1	84	310	-	97	78	*N.tabacum* AB176524
	b	9	1481	1	197	-	-	16	68	*S.tuberosum* AY954034
	c	4	1057	1	153	337	-	37	76	*N.tabacum* AB176524
	d	3	1083	0	-	-	-	80	84	*S.tuberosum* AY954034
	e	4	1475	1	94	73	-	86	80	*N.tabacum* AB176524
	f	4	668	0	-	125	-	35	74	*N.tabacum* AB176524
	g	3	1106	1	n.d.	349	-	35	100	*N.tabacum* AB176525
	h	2	759	0	-	-	-	24	59	*N.tabacum* AB176526
RT	a	4	1362	0	-	-	107	88	88	*P.hybrida* X71059
FLS	a	3	1102	1	459	-	-	51	91	*P.hybrida* Z22543
	b	2	798	0	-	543	-	24	78	"
	c	3	1373	1	n.d.	492	-	45	83	"
F3'5'H	a	2	789	0	-	-	474	20	89	*P.hybrida* D14588
	b	8	1621	1	521	103	-	58	87	"

Table 2. Information on carotenoid genes from genomic sequences of TGI.

Gene	Contig	# clones	Length (bp)	# introns	intron length (bp)	promoter region (bp)	terminator region (bp)	cds coverage (%)	cds identity (%)	Reference sequence
PSY	a	4	1398	3	n.d.	146	-	52	88	*L.esculentum* L23424
	b	2	704	1	404	-	-	24	94	"
PDS	a	7	1679	3	165-106-307	-	565	53	92	*L.esculentum* X78271
	b	8	3302	1	86	2604	-	20	89	"
ZDS	a	5	2413	2	541-151	644	-	29	94	*L.esculentum* AF195507
CRTISO	a	7	1724	4	439-92-137-79	592	-	53	90	*L.esculentum* AF416727
	b	2	502	1	n.d.	-	304	4	89	"
	c	3	1207	3	n.d.	-	-	10	93	"
LCYβ	a	7	1072	0	-	156	-	57	84	*L.esculentum* AF416727
CHY	a	22	2177	6	98-295-120-203-147-112	90	180	100	82	*L.esculentum* Y14809
	b	16	2779	6	93-359-151-109-119-143	504	404	100	78	"
ZEP	a	5	1233	1	n.d.	687	-	26	85	*L.esculentum* Z83835
	b	5	1821	3	107-85-90	-	1156	15	87	"
VDE	a	7	2303	2	n.d.	-	801	46	98	*N.tabacum* U34817
	b	2	1017	2	n.d.	-	-	6	98	"

EST sequence analysis

Table 3. Frequency of EST sequences of phenylpropanoid and flavonoid genes obtained from cDNA libraries.

Gene	Pathway	Total ESTs	Contigs	ESTs in contigs	Singlets
FGT	both	30	3	14	16
PAL	Phenylpropanoid	29	2	28	1
4CL	Phenylpropanoid	17	4	14	3
C4H	Phenylpropanoid	14	2	13	1
DFR	Flavonoid	5	0		5
ANS	Flavonoid	2	0		2
LAR-ANR	Flavonoid	2	0		2
CHI	Flavonoid	1	0		1
CHS	Flavonoid	1	0		1
FLS	Flavonoid	1	0		1
RT	Flavonoid	1	0		1
FHT	Flavonoid	1	0		1

Sequencing of cDNA libraries produced a relatively low number of EST sequences from the carotenoid pathway. This result can be explained by the large amount of sequences originating from leaf libraries. In green leaves, carotenogenesis is already accomplished, and the high proportion of early carotenoid pathway genes (PDS, PSY, ZDS; Table 4) could be attributed to “housekeeping” functions in maintaining the carotenoid levels as a support to photosynthesis.

As for polyphenol metabolism, a clear difference between the number of EST related to phenylpropanoid and flavonoid pathway genes was found, the former being expressed to a higher rate (Table 3). A high expression of PAL, C4H and 4CL genes could be responsible also for the formation of other phenolic compounds (e.g., chlorogenic acid) and lignin precursors. Homeostasis can account for the low expression of flavonoid genes, except for different flavonoid glycosyltransferases, responsible for the final modification of flavonoid and phenylpropanoid end-products, which were highly represented (Table 3).

Table 4. Frequency of EST sequences of carotenoid genes obtained from cDNA libraries.

Gene	Total ESTs	Contigs	ESTs in contigs	Singlets
PDS	10	1	10	
LYβCY	9	2	9	
PSY	6	1	5	1
ZDS	3	1	3	
LYεCY	3	1	3	
ZEP	1			1
VDE	1			1
CHY	1			1

TOBACCO GENOME INITIATIVE

http://www.tobaccogenome.org

Acknowledgements

NCSU is acknowledged for contribution to the TGI. Irfan Gunduz (PM USA) is acknowledged for useful discussions and for implementing some bioinformatics tools used in this study.