2013 CORESTA Congress, Italy

The analysis of molecular marker and evolution based on tobacco EST

Gong daping, xie minmin, Sun yuhe

October 17, 2013

Contents

- Why we chose the Expressed sequence tags (ESTs) sequencing?

- Construction of two normalized full-length enriched cDNA libraries from *N.sylvestris* and *N.tomentosiformis*

- Annotation of data and identification of unigenes

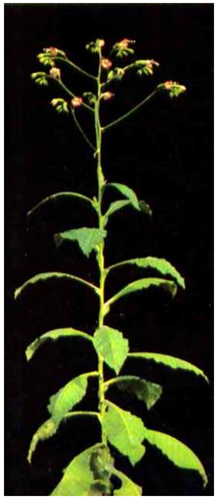
- Identification of the molecular markers (single nucleotide polymorphisms (SNPs) and Microsatellites (SSRs))

Origin of Tobacco

N. sylvestris

2n = 24

x



N. tomentosiformis

2n = 24

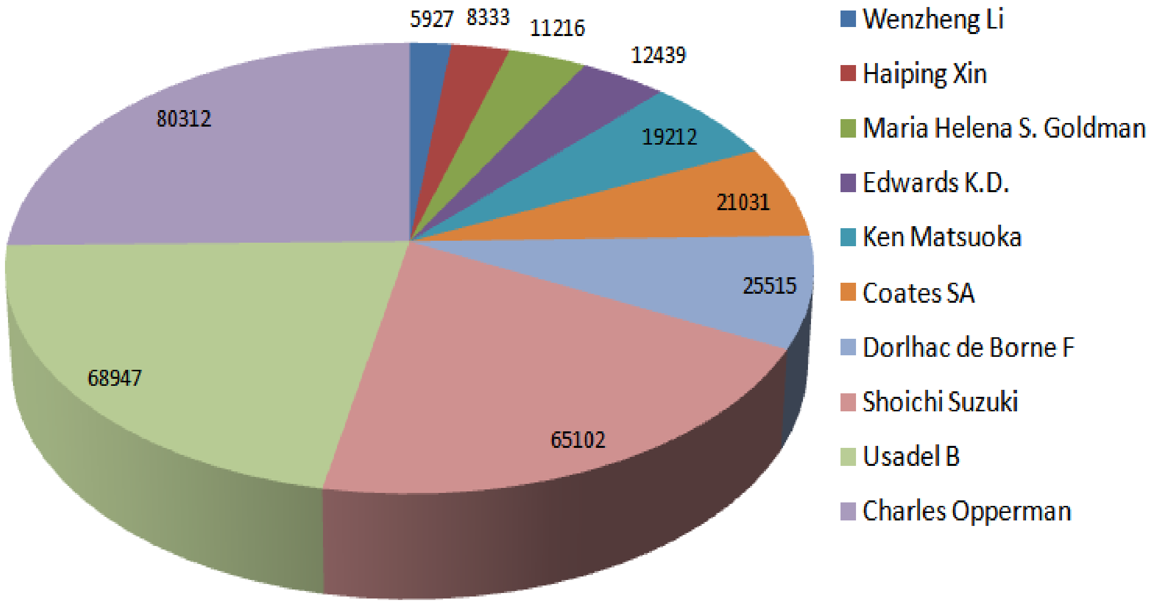
→

N. tabacum

2n = 48

- Tobacco is an allotetraploid (2n = 48) and most likely derived from the interspecific hybridization of the diploid species N. *sylvestris* (2n = 24) and N. *tomentosiformis* (2n = 24).
- Tobacco have a bigger genome size of approximately 4.5 Gb.

Tobacco EST in Genbank



- The de novo sequencing of the whole genome of tobacco is a challenging task.
- Expressed sequence tags (ESTs) are a less expensive alternative for gaining transcriptionally active genes.
- Currently, Over 300,000 EST sequences were available at Genbank.

The full-length libraries

- Two normal full-length cDNA libraries were constructed
- Two diploid ancestral species *(N.tomentosiformis* and *N.sylvestris*) of N.tabaccum
- Mixed entire seedlings, root, stem, leaves, flowers, buds
- High quality full length libraries

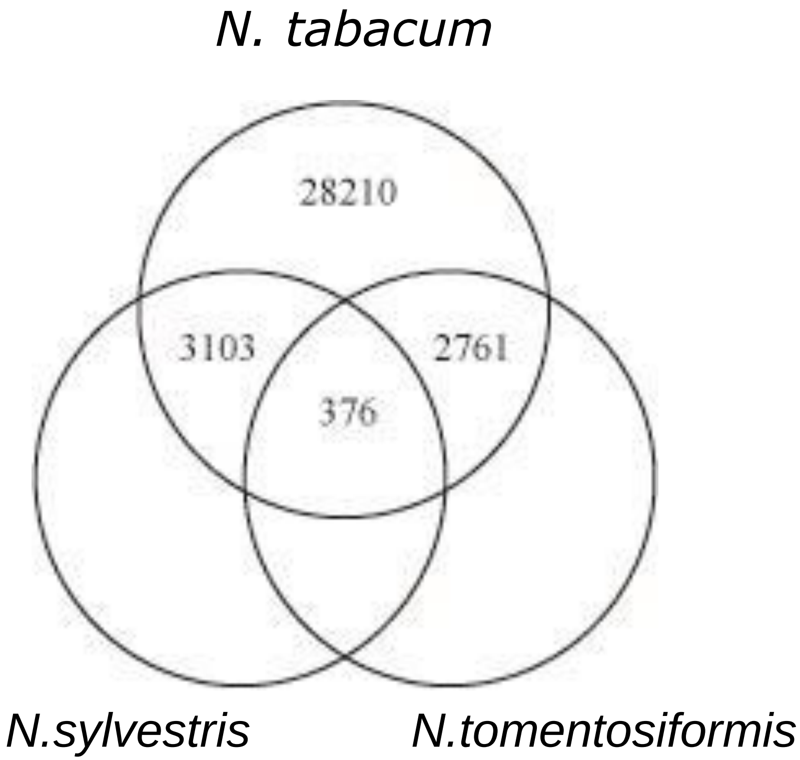
Sequencing results

- 22,525 clones were sequenced (3’ end)
- The success rate of sequencing was approximately 93%
  - 10,503 sequences were from *N.sylvestris*
  - 10,450 sequences were from *N.tomentosiformis*
- Average length of each ESTs was 656 bp.

EST assembly characteristics

- A total of 347,022 Nicotiana ESTs were assembled into 34450 contigs and 123,511 singletons (15,7961 unigenes)
- The length of each contigs ranged from 107 to 3502 bp with an average length of 879 bp
- Median number of ESTs per contig was 14.8 (min= three, max= 1158)
- The average depth of coverage for each assembled nucleotide was 4.4

The co-assembly of EST



Our result shows that the transcripts from resident T- and S-genomes in the allotetraploid nucleus are more closely related to their diploid homologs rather than to each other.

Gene identification and annotation

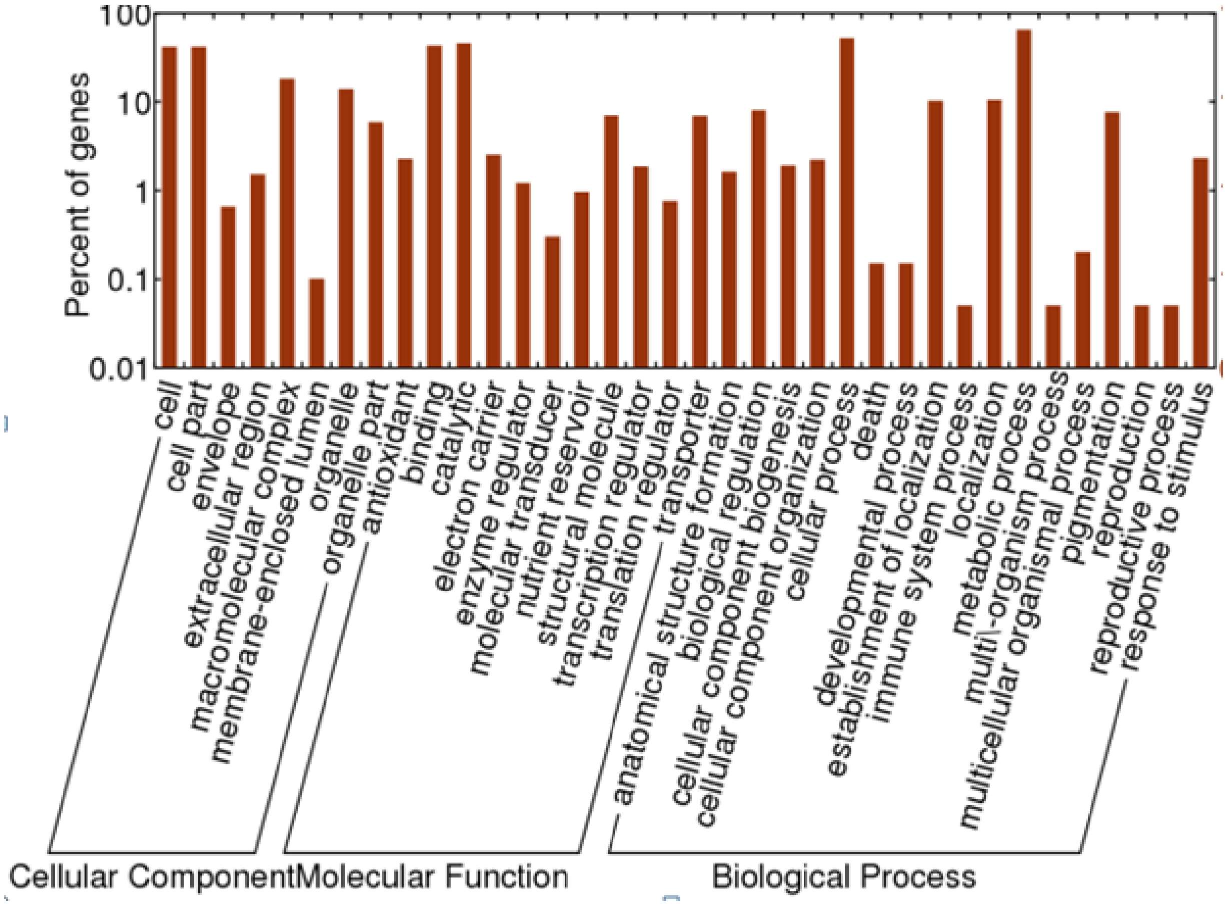
- Open reading frame (ORF) were found for 104,915 (66.4%) unigenes with an average length of 524 bp (min = 150, max = 3486) with ESTscan

- 73,670 protein products had at least one annotated Protein family domain

- The most abundant domain found was protein kinase

- Other dominant Pfam annotations include RNA recognition motif, leucine rich repeat, WD40 repeat, Zinc finger domains and P450

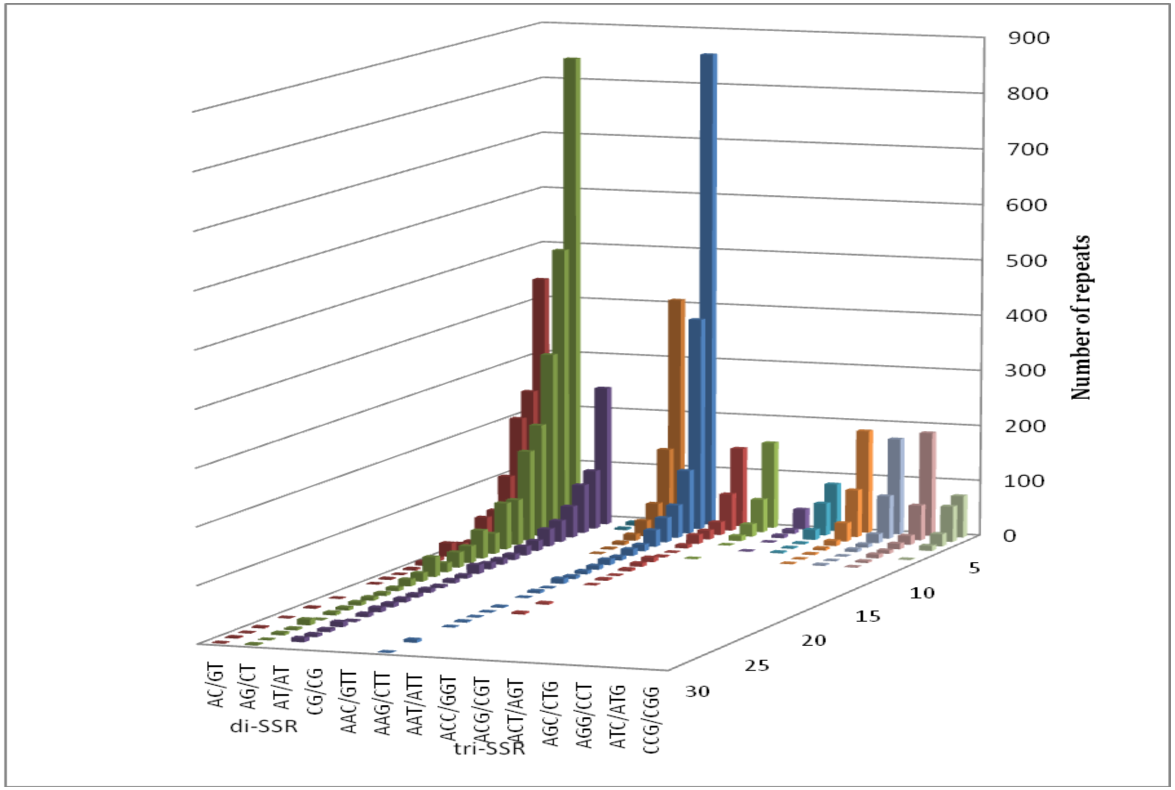
Functional classification of genes



Identification of SSR markers

Total number of unique sequences	152,891
Total number of identified SSRs	11,869
Number of sequences containing SSR	10424
Number of sequences containing more than 1 SSR	1209
Number of SSRs present in compound formation	945
Di-nucleotide repeats	4434
Tri-nucleotide repeats	4054
Tetra-nucleotide repeats	516
Penta-nucleotide repeats	1273
Hexa-nucleotide repeats	1592

Frequency distribution of SSRs



- The most frequent SSR motifs were AG and AAG.
- The longest number of repeats was observed in AG motif having 53 repeats.
- The SSR density is 12.5 SSRs per 10 kb

Identification of SNP/InDel markers

- A total of 103,027 putative SNPs and 24,760 putative insertions/deletions (indels) were identified
- These SNP included 64349 transitions and 38678 transversions, respectively
- SNP frequency observed was 4.3 SNPs per kb of transcribed sequences
- Majority of filtered SNPs were identified from the contigs containing more than 3 sequences

Future prospects

- The newly identified SNPs and SSRs can be used to generate genetic map, locate genes of economically important traits and Marker-assisted-selection (MAS) in breeding programme

- The tobacco genome will be completed in the near future and will allow comprehensive and large scale functional genomic study

- The advanced high-throughput sequencing technology and the availability of the reference tobacco genome will make it feasible to re-sequence tobacco genome and thereby allow the genome-wide survey of genetic variation

Thank You !