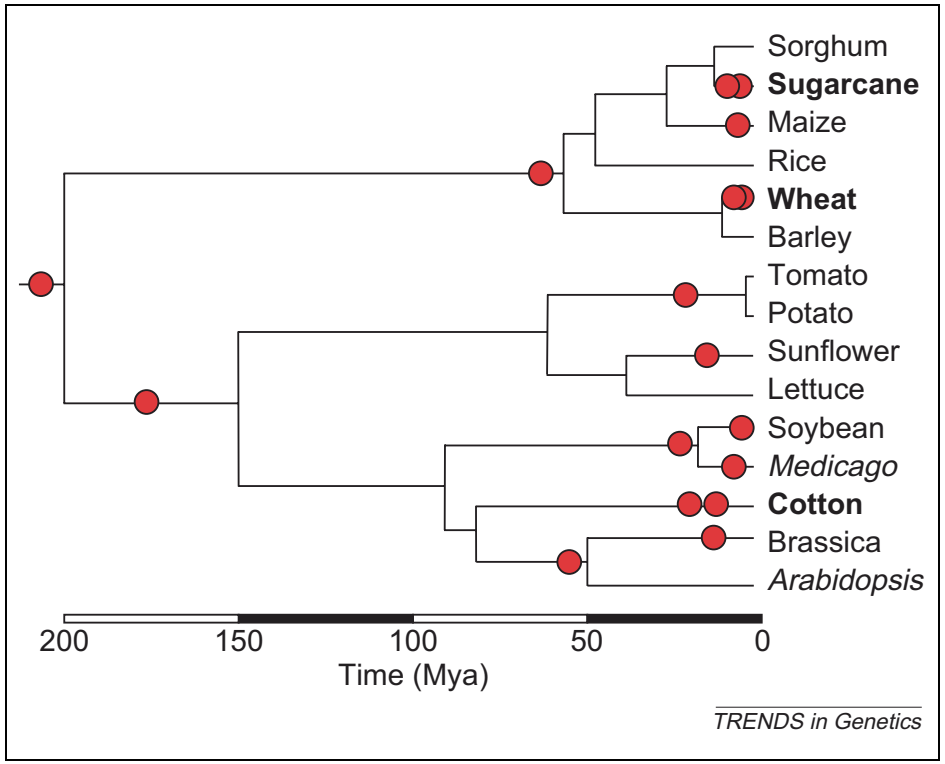
Exploring non-coding RNAs using single genome comparative analysis

Chong-Jian Chen

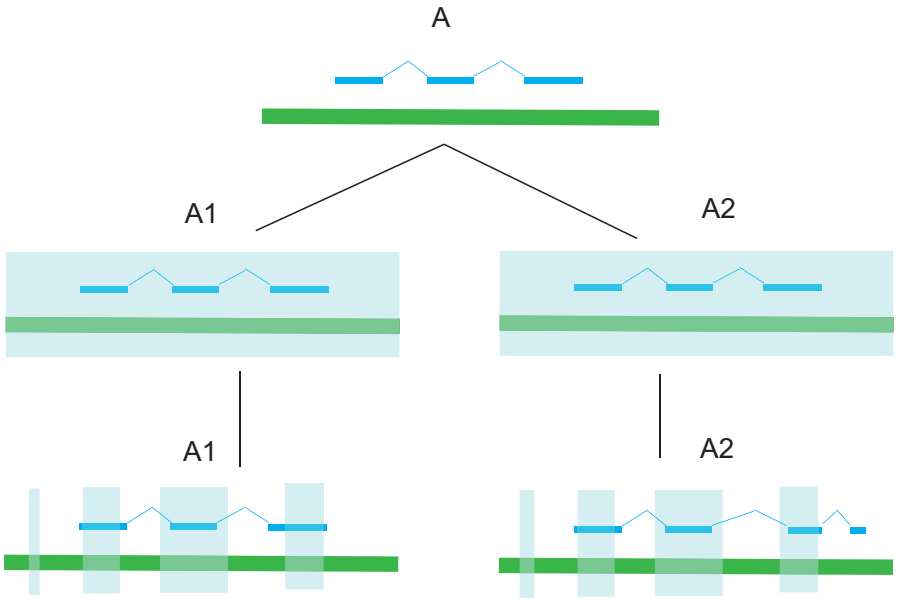
Sequence, structure and function of RNAs – Daniel Gautheret
Institut de Génétique et Microbiologie, Université Paris-Sud, France
and
Key Laboratory of Gene Engineering of the Ministry of Education – Liang-Hu Qu,
Zhongshan University, Guangzhou, China

April 15, 2008

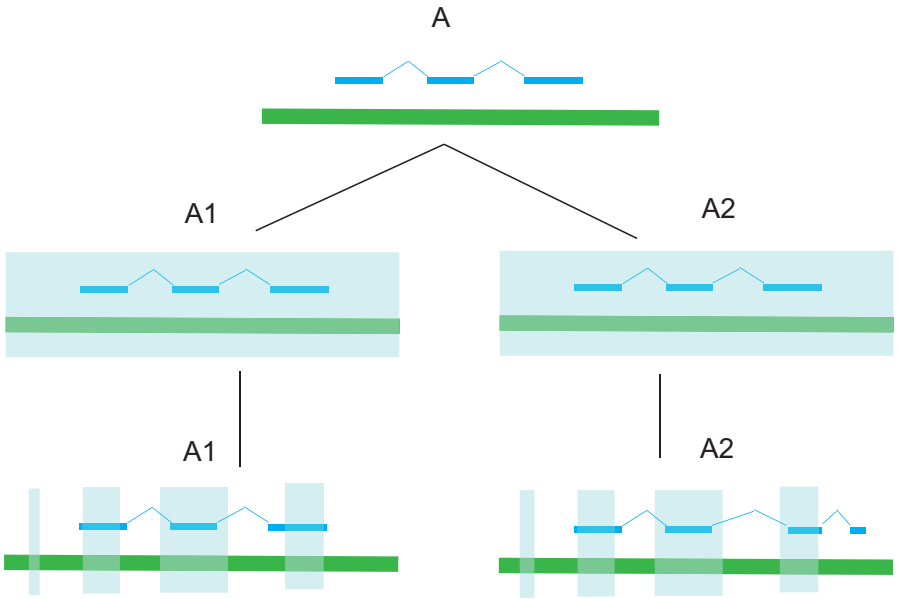
Introduction
Large-scale duplication events in phylogentic context



- At least 16 polyploid events have been documented during the evolutionary history of a respresentative sample of angiosperms



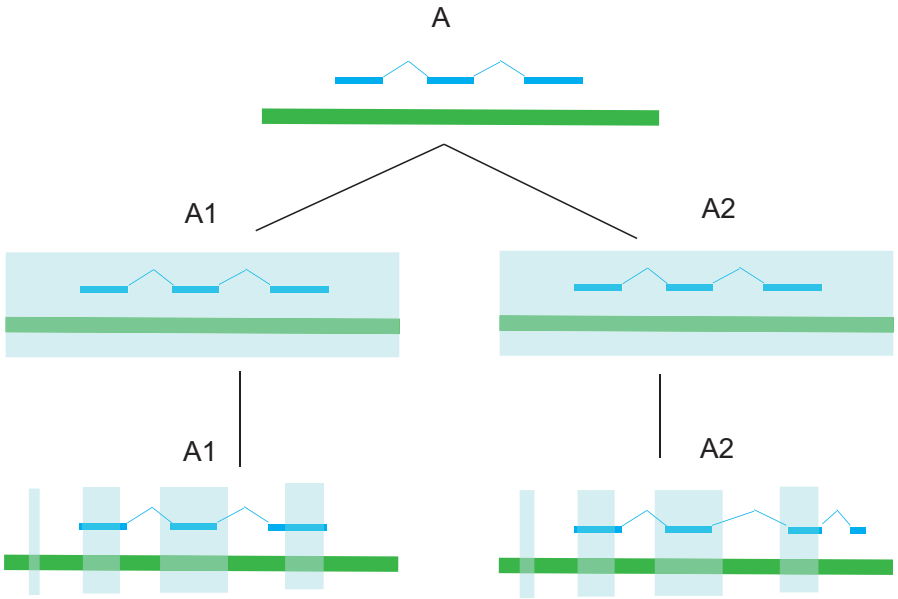
evolution of gene duplicate



evolution of gene duplicate

non-coding RNA

What is the fate of ncRNA duplicate?



evolution of gene duplicate

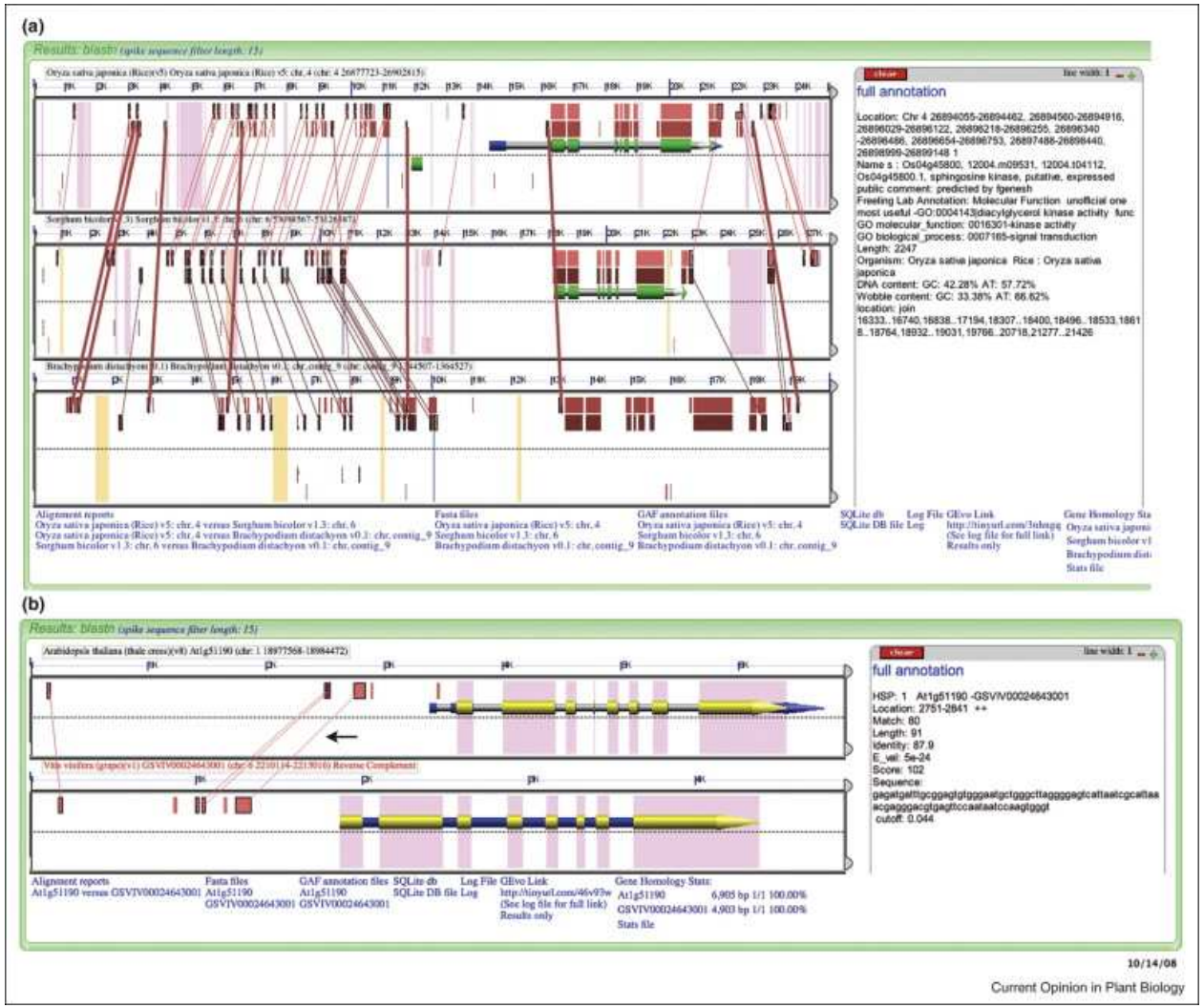
non-coding RNA

What is the fate of ncRNA duplicate?

Hypothese

- like protein coding gene, paralogues of non-coding elements experience purifying selection in their history, while that of surrounding sequences evolve in a neutral manner

Introduction
Consequence of genome sequence after duplication



Graphic representations of plant syntenic regions

- Michael Freeling (PNAS 2006) use bl2seq to identify gene-associated conserved noncoding RNA in Arabidopsis
- After strictly sorting, they obtained
  - 14,944 intragenomic Arabidopsis CNSs
  - The mean CNS length is 31 bp, ranging from 15 to 285 bp
  - $\sim$1.7 CNSs associated with a typical gene

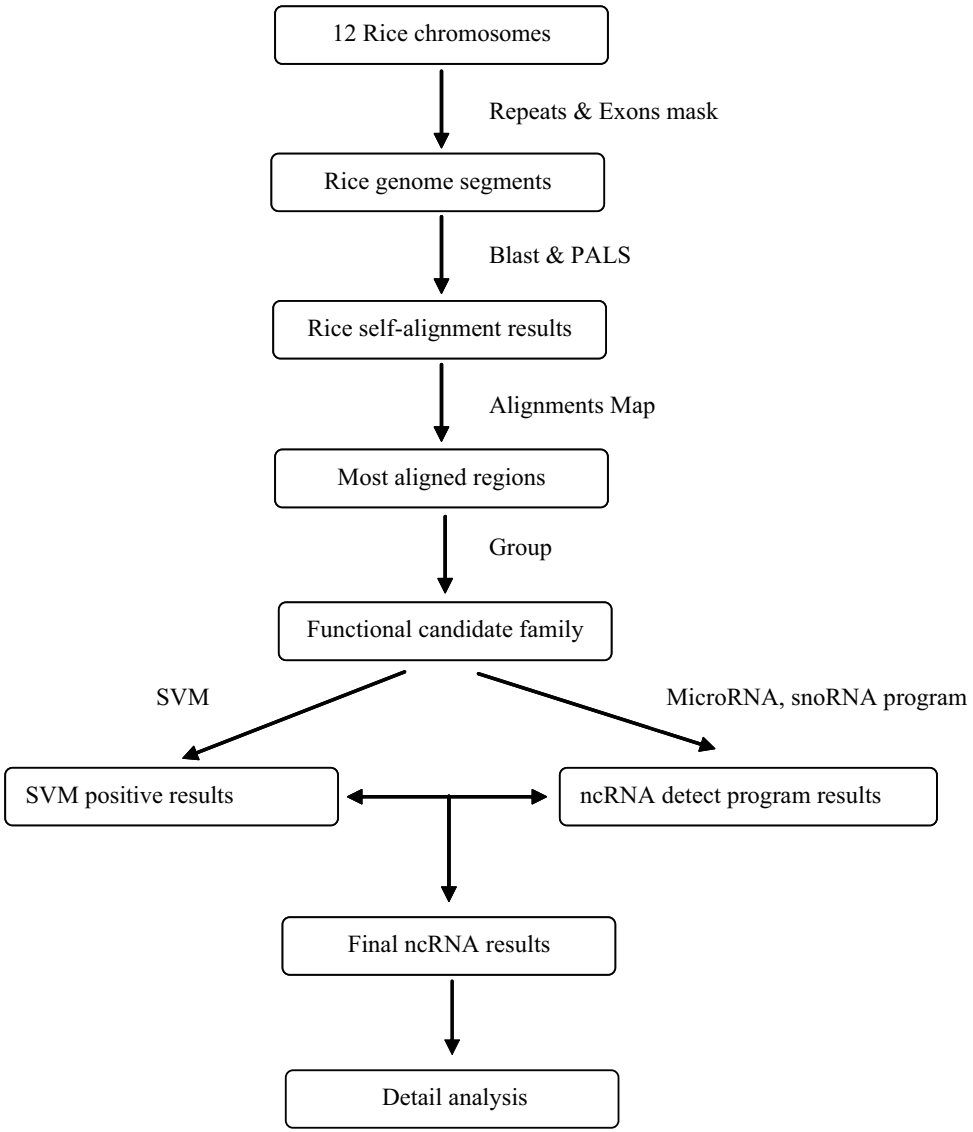
- Michael Freeling (PNAS 2006) use bl2seq to identify gene-associated conserved noncoding RNA in Arabidopsis
- After strictly sorting, they obtained
  - 14,944 intragenomic Arabidopsis CNSs
  - The mean CNS length is 31 bp, ranging from 15 to 285 bp
  - $\sim$1.7 CNSs associated with a typical gene

Question

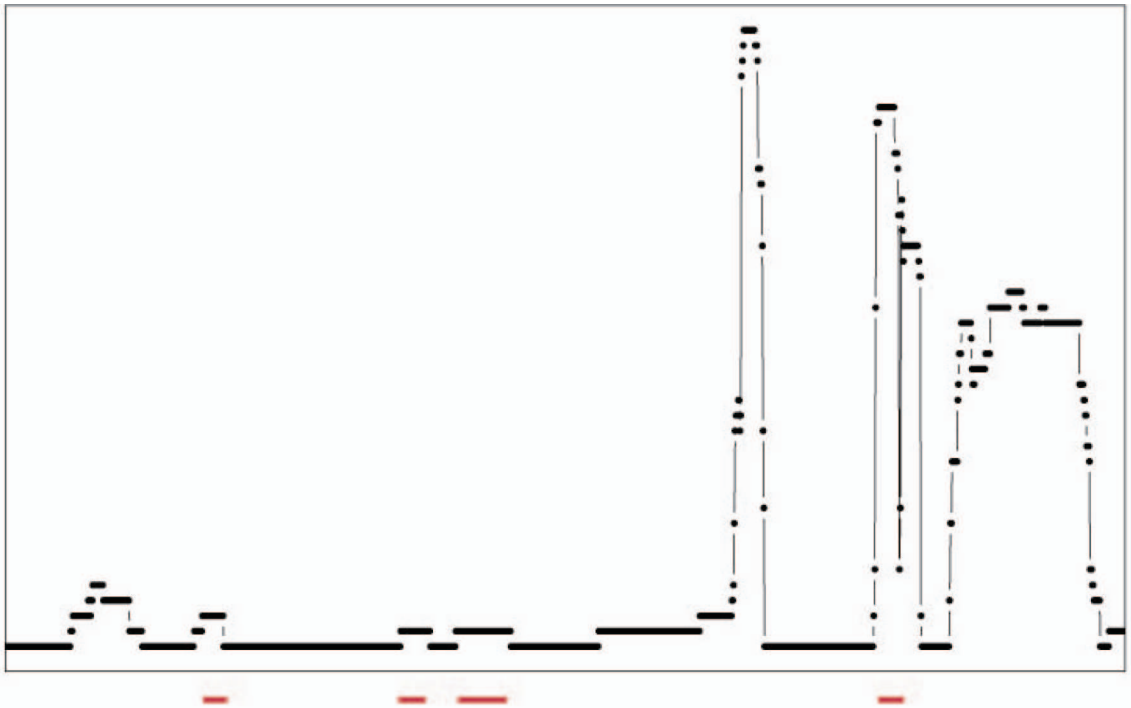
What about finding non-syntenic conserved non-coding RNAs?

Method

Our pipeline to detect conserved non-coding RNAs



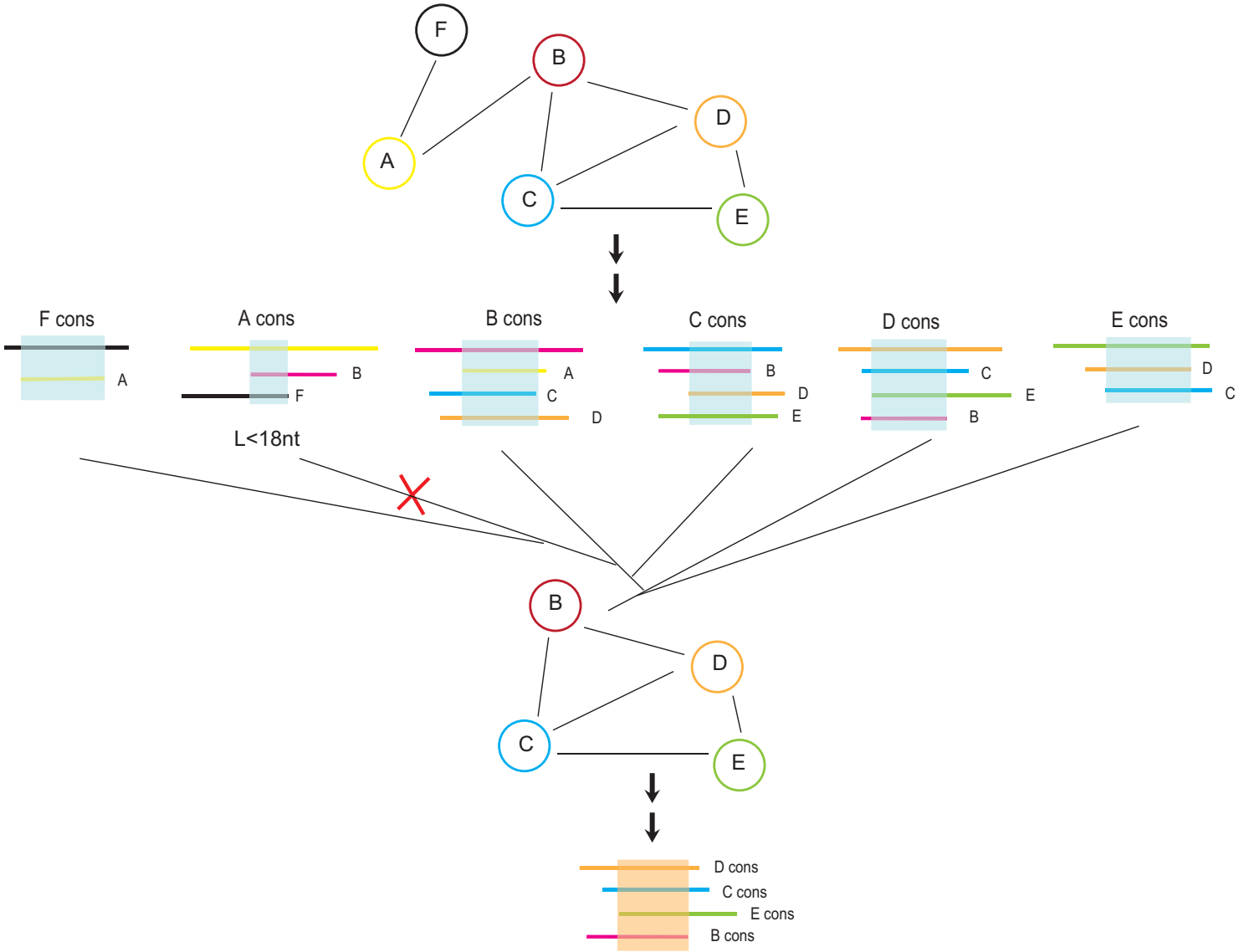
Method
Definition of high-frequency areas



- We mapped all Blast/Pals hits back to the genome and plotted hit numbers at each position to generate frequency maps at all overlapping aligned regions. We extracted high-frequency areas larger than 18nt

Method

Clustering aligned hits into families



- We calculated the conservation score of each family in 15bp window, and retained region with $\pi \geq 0.7$

$$\pi = \frac{2 \sum_{\substack{0<i<n \\ i<j\leqslant n}} C_{ij}}{L(n-1)n}$$

n: size of family
C_{ij}: number of identical nucleotides between sequence i and j
L: size of window 15

Results
Putative ncRNA family after clustering

- 122,853 families were identified, in which 120 known non-coding RNA families were present
  - Box C/D snoRNA: 50
  - Box H/ACA snoRNA: 19
  - microRNA: 18
  - tRNA: 12
  - rRNA: 21
- The mean length of hits is 35 bp, ranging from 14 to 200 bp
- The mean family size is 2.6, ranging from 2 to 219

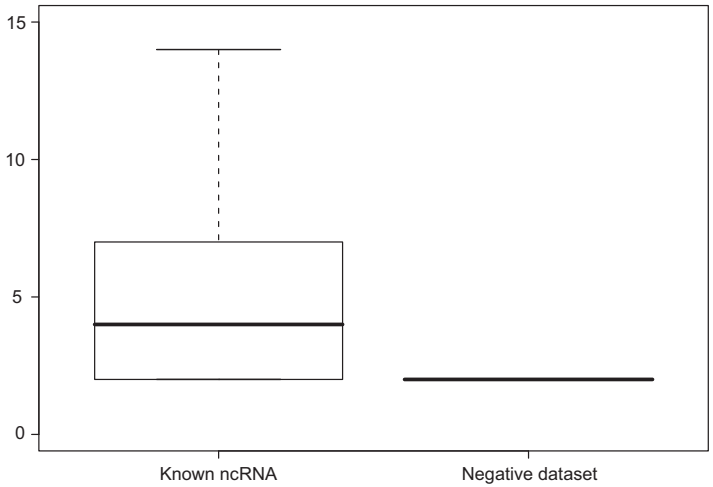
- A negative set of 35,725 families was built from low complexity and repeated elements
  - either single nucleotide percentage $\leqslant 0.1$ or ≥ 0.5, and either di-nucleotide percentage $\leqslant 0.2$ or ≥ 0.8
  - percentage of repeats in family ≥ 0.5
- Measure F score of 15 features: 120 known RNA families and 6000 random selected negative families

Results
Attributes used in SVM

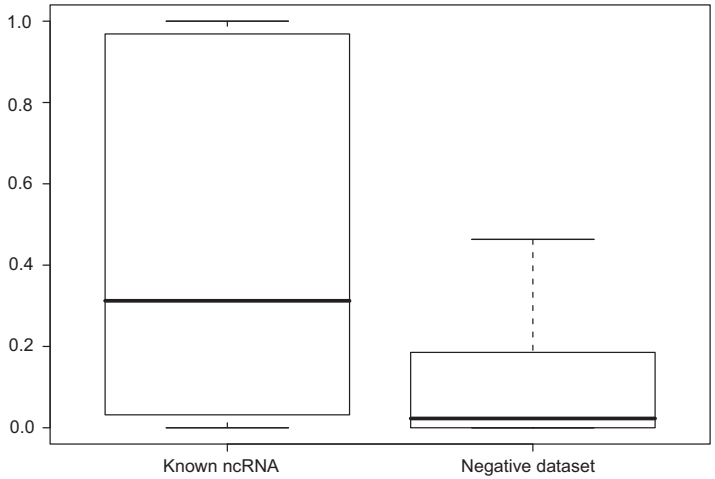
Number	Feature	F Score
1	size of family	0.0488
2	RNAz P-value	0.0330
3	length of candidate	0.0227
4	deep sequencing-based expression	0.0281
5	size of largest positional cluster	0.0182
6	candidates in clusters/size of family	0.0069
7	sequence identity of family	0.0037
8	size of family/max number of mapped hits	0.0028
9	tilling array-based expression	0.0017
10	size of largest duplicate	0.0007
11	average interval in cluster	0.0005
12	percentage of candidates in intron	0.0004
13	percentage of candidates in UTR	0.0004
14	number of positional cluster/size of family	0.0002
15	average size of large duplicate	0.0001

Results

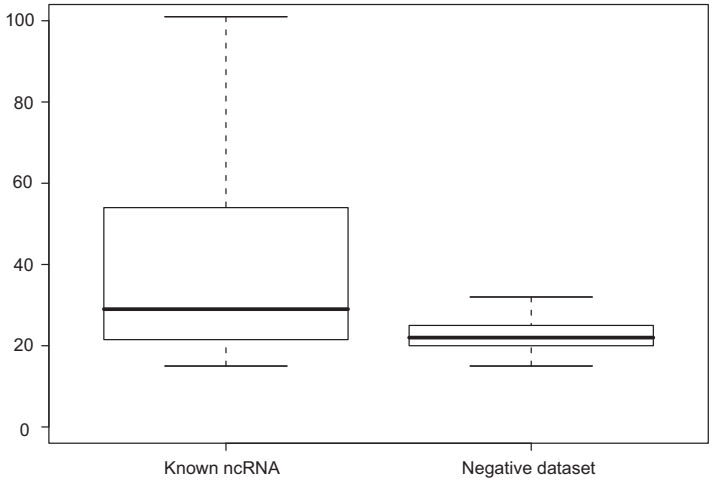
Boxplots of four most useful attributes



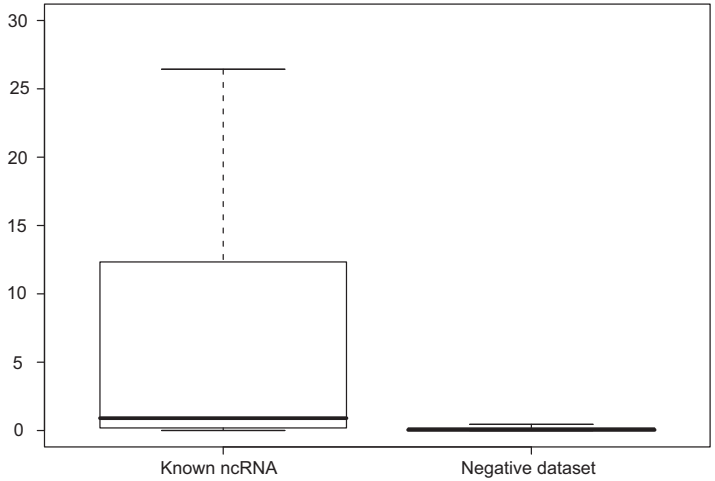
1. size of family



2. RNAz P-value

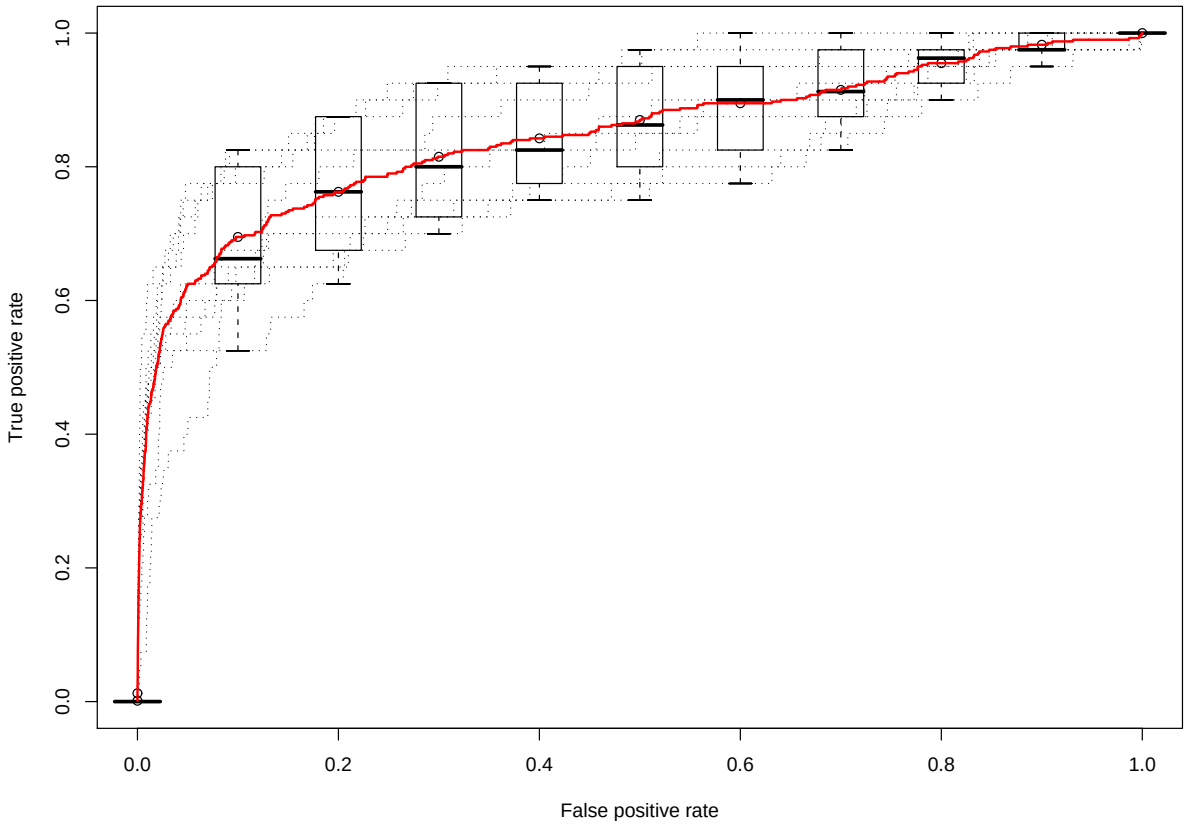


3. length of candidate



4. deep sequencing-based expression

Results
ROC curve

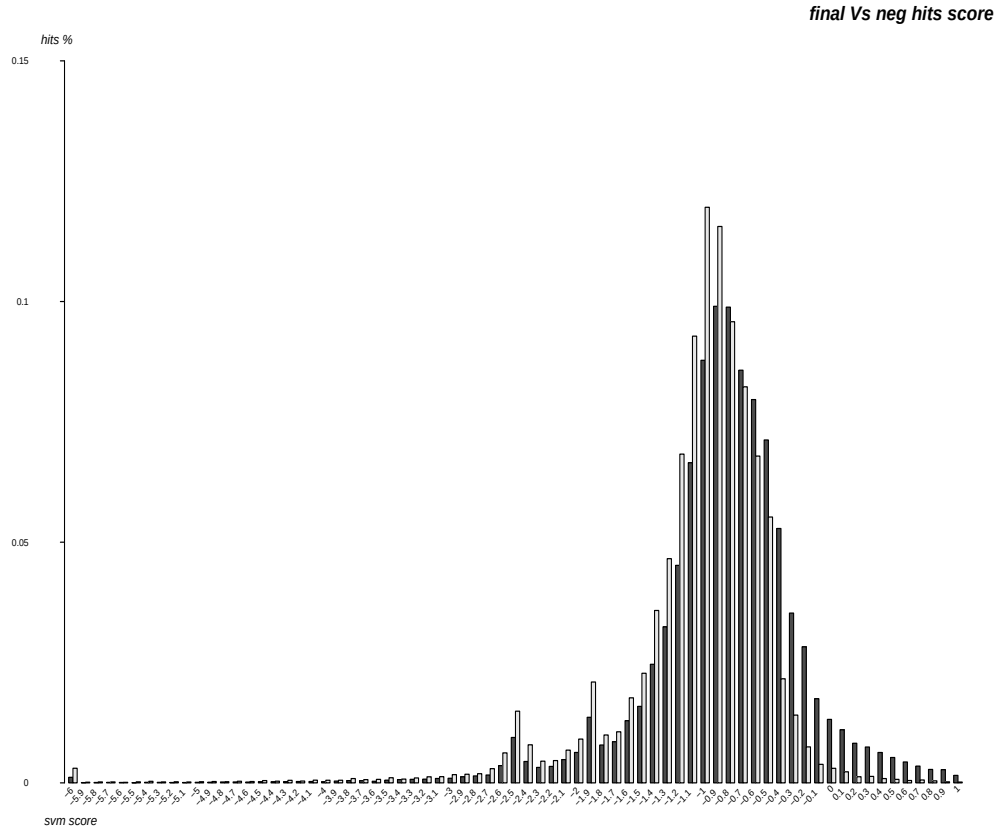


- After ten bootstraps the SVM achieved a mean Roc curve area (AUC) of 0.84

Method
Utilize SVM to predict ncRNA

- A negative set of 35,725 families was built from low complexity and repeated elements
  - either single nucleotide percentage $\leqslant 0.1$ or ≥ 0.5, and either di-nucleotide percentage $\leqslant 0.2$ or ≥ 0.8
  - percentage of repeats in family ≥ 0.5
- Measure F score of 15 features: 120 known RNA families and 6000 random selected negative families
- Training SVM model
  - Training dataset: 80 known RNA families and 4000 random selected negative families
  - Test dataset: remaining positive and negative datasets

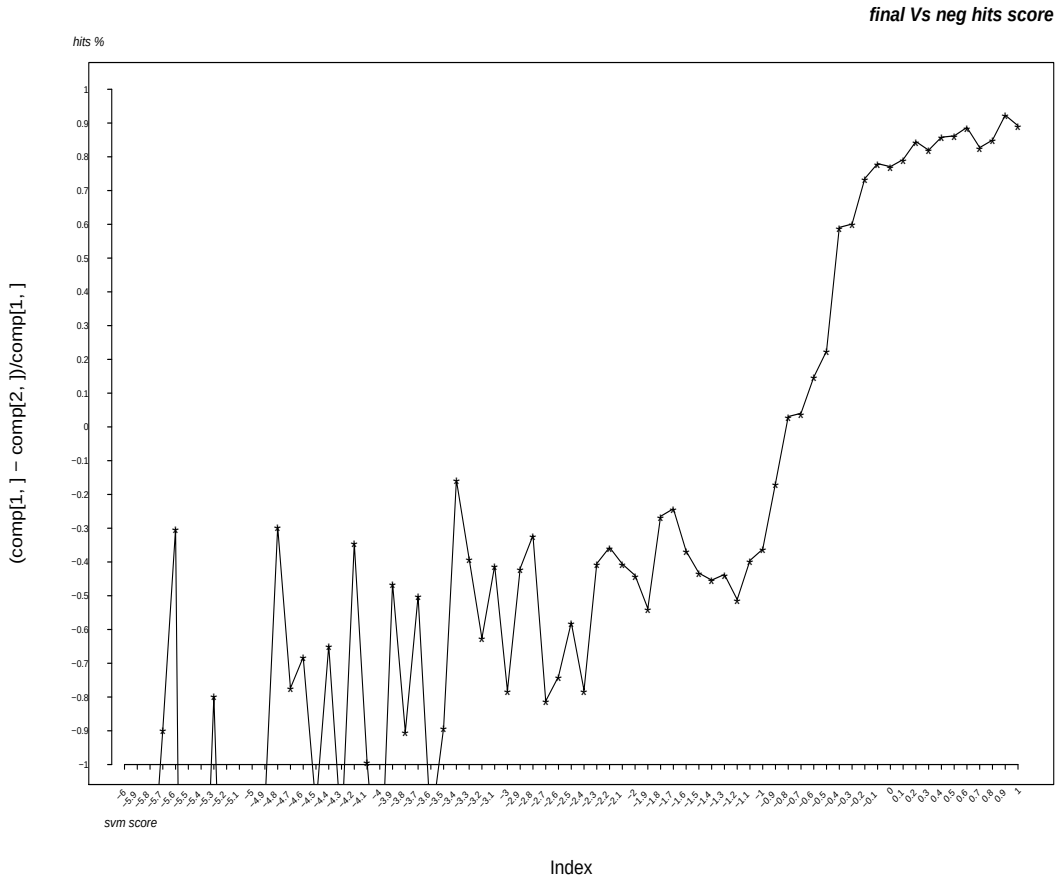
Results
Score distibution of Non-negative and Negative dataset



X axis: $1+log_{10}^{(Pvalue)}$

Y axis: percentage of families

Results
Score distibution of Non-negative and Negative dataset



X axis: $1+log_{10}^{(Pvalue)}$

Y axis: $(P_T\text{-}P_N)/P_T$

- At a Pval$\geq$ 0.5, we finally obtained 1049 putative non-coding RNA familes with an estimated specificity of 88%

To be continuing

1 2 3 4 ...

- scan snoRNAs in putative RNA families
- scan microRNAs in putative RNA families
- Use Rfam to catalog ncRNA families
- Use RNAclust to find new structural ncRNA families
- ...


Thank You !