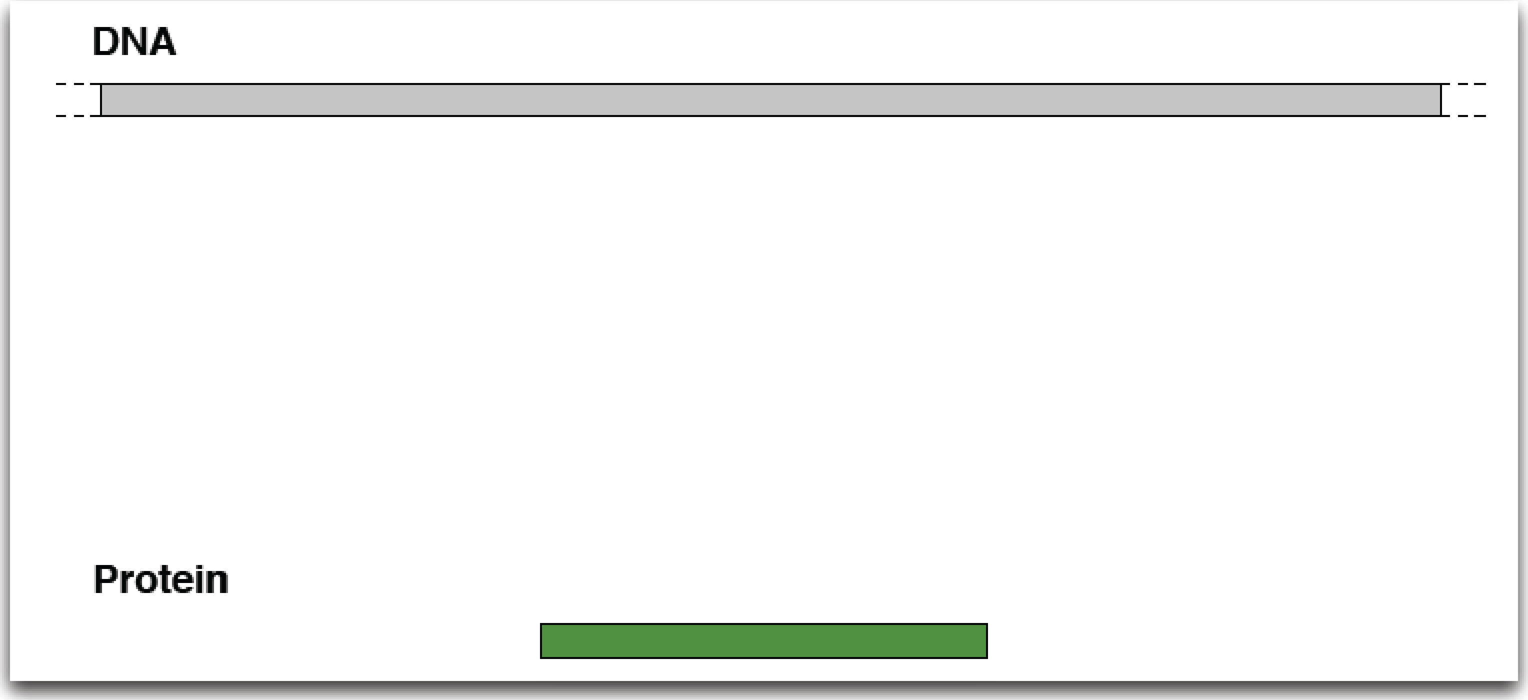
转录组学基础及研究

陈铭
mchen@zju.edu.cn

理论课内容

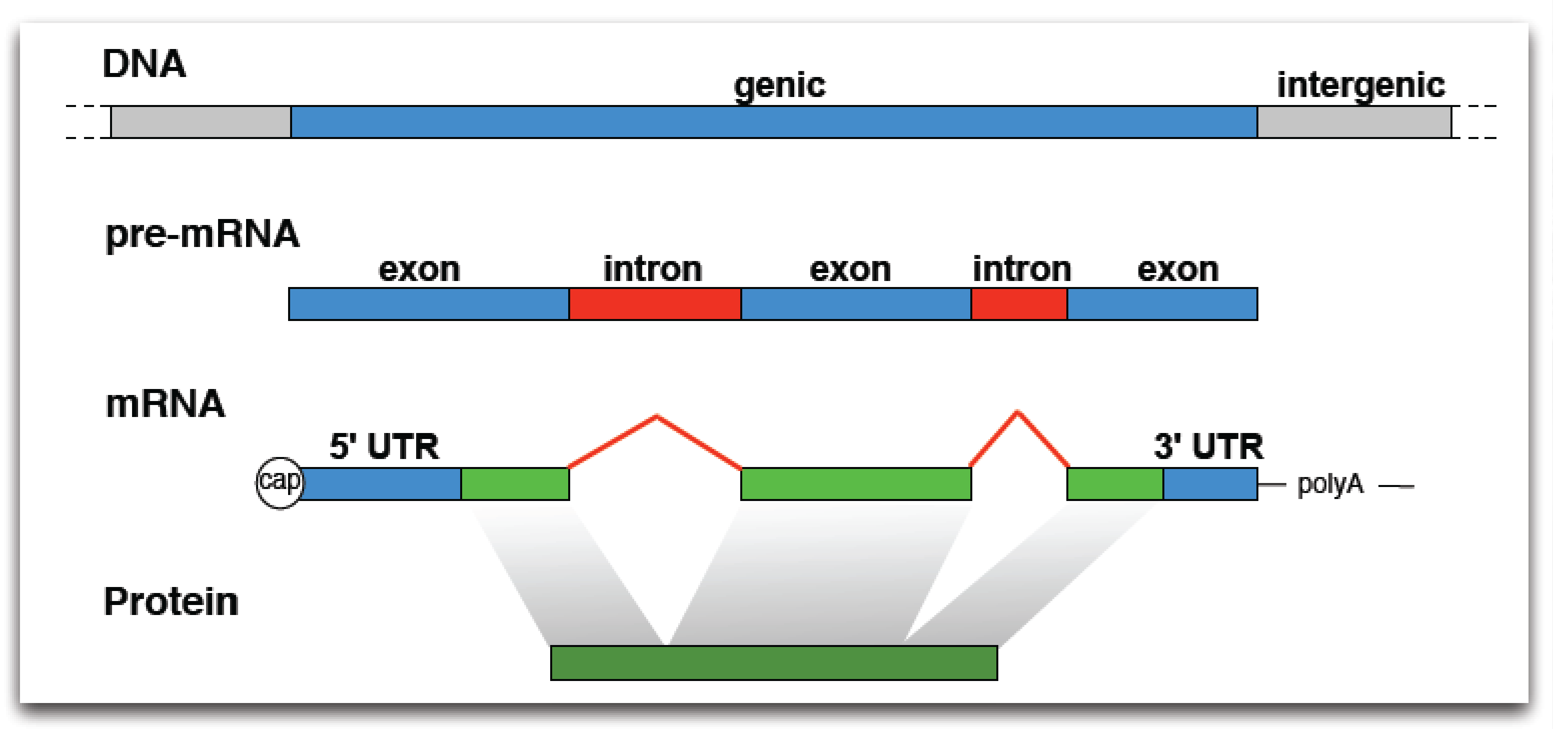
- 转录组学介绍
- 基因表达数据分析
  - 测定技术
  - 差异基因
  - 功能分析
- 几个实例
- 非编码RNA分析

Protein-coding gene

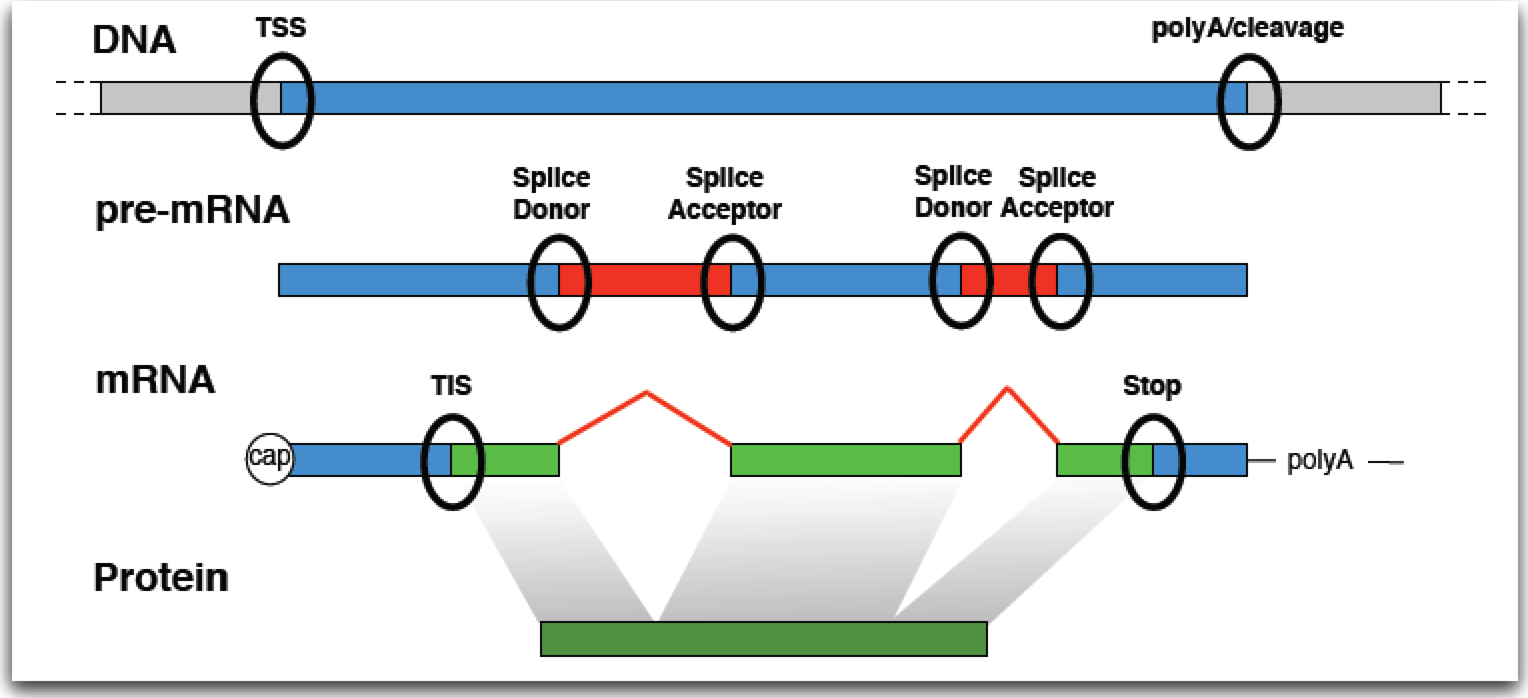


✓Gene: a functional piece of DNA sequence

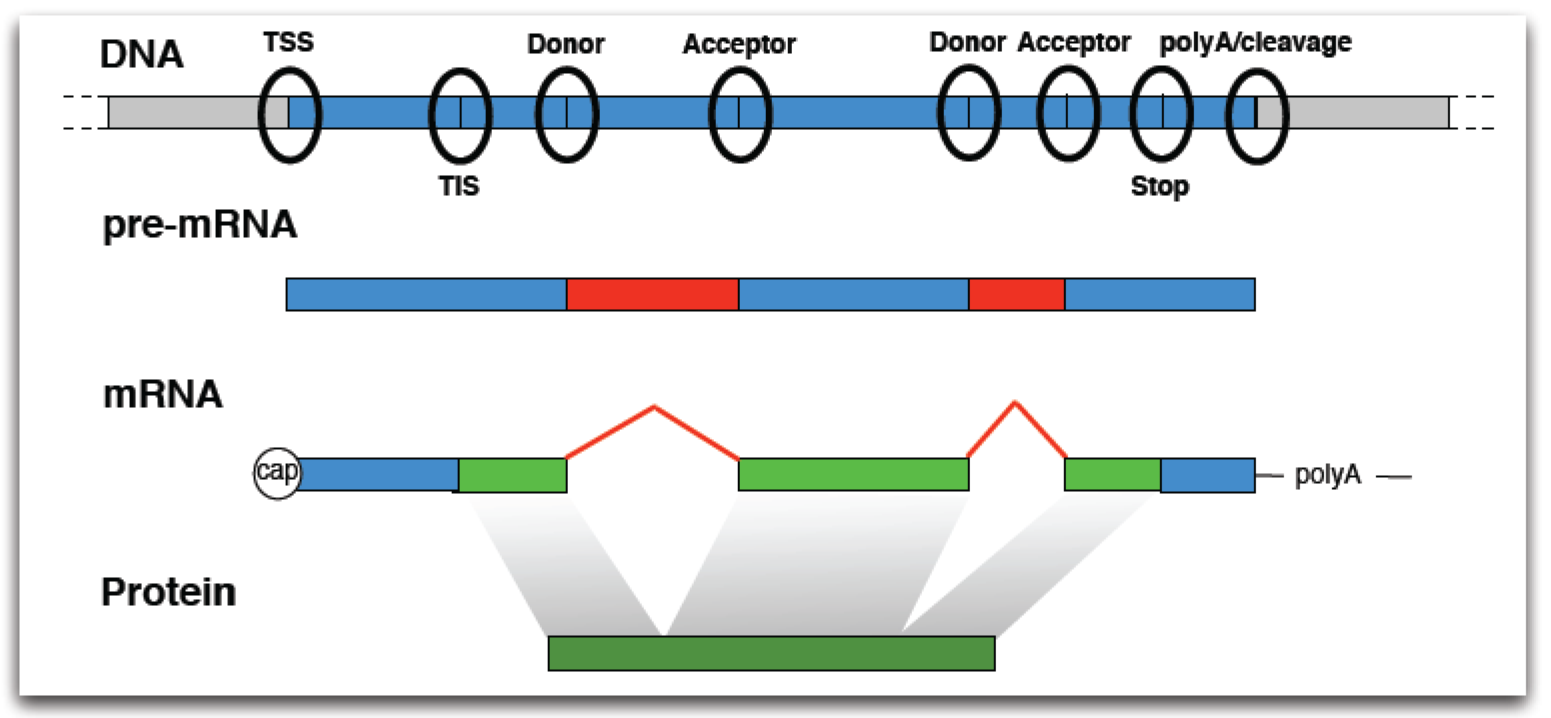
Computational Gene Finding



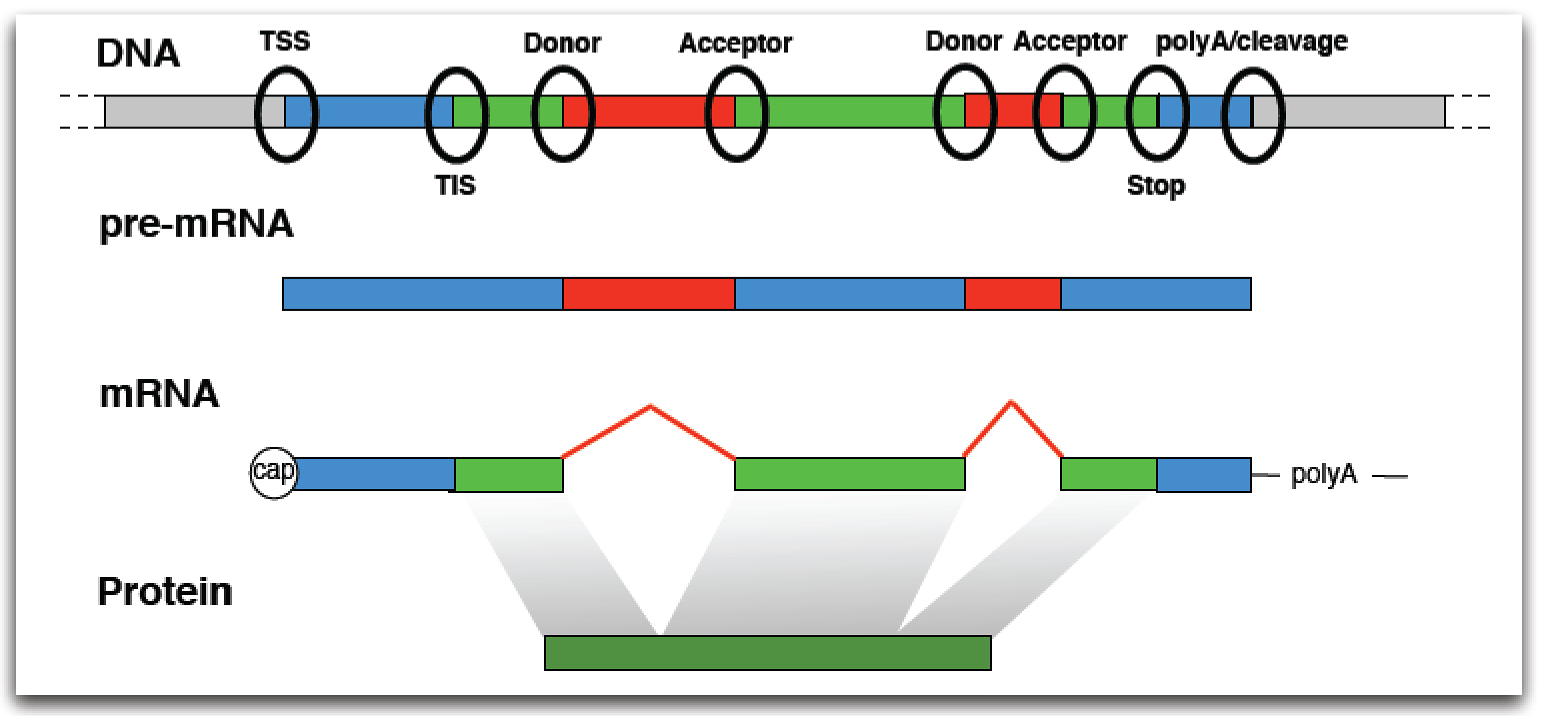
Predict signals used during processing



Predict signals used during processing



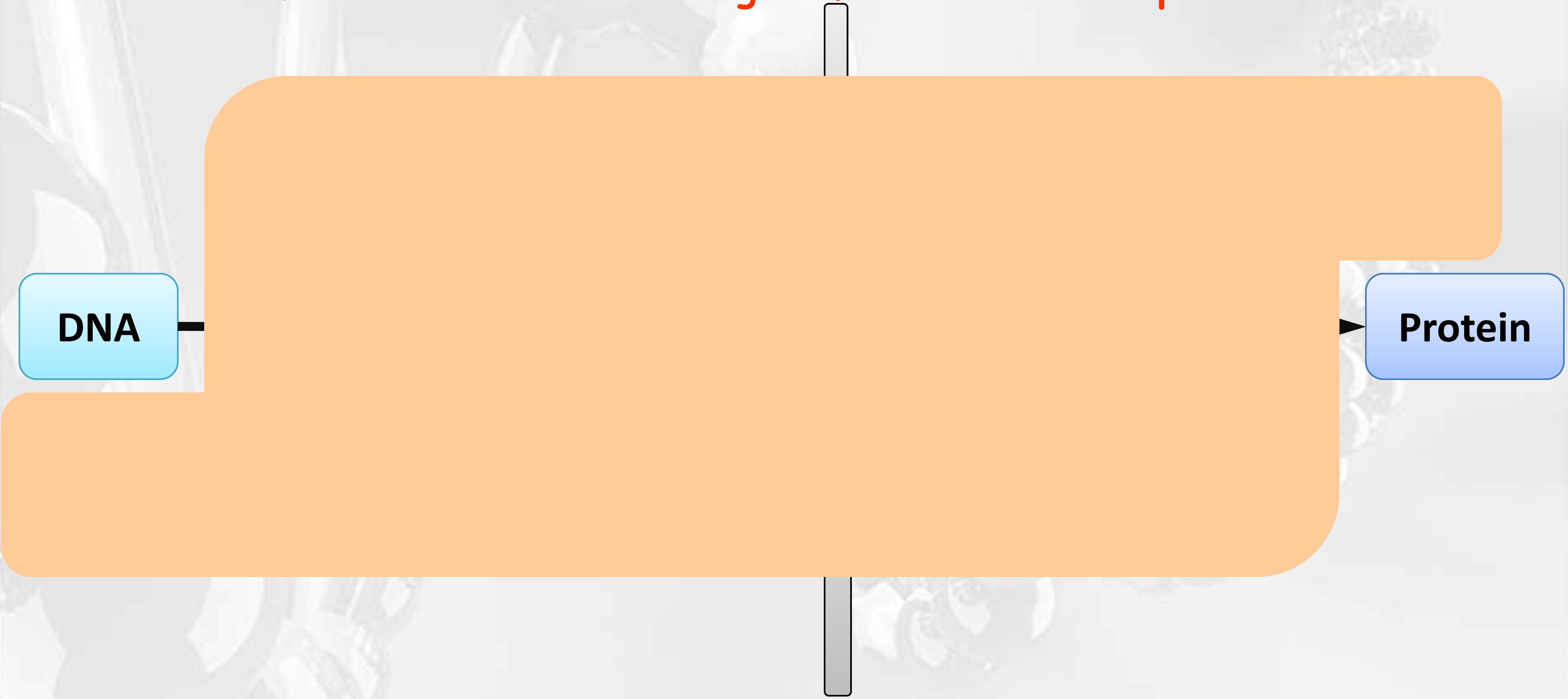
Computational Gene Finding



✓Predict the correct corresponding label sequence with labels “intergenic”, “exon”, “intron”, “5’ UTR”, etc

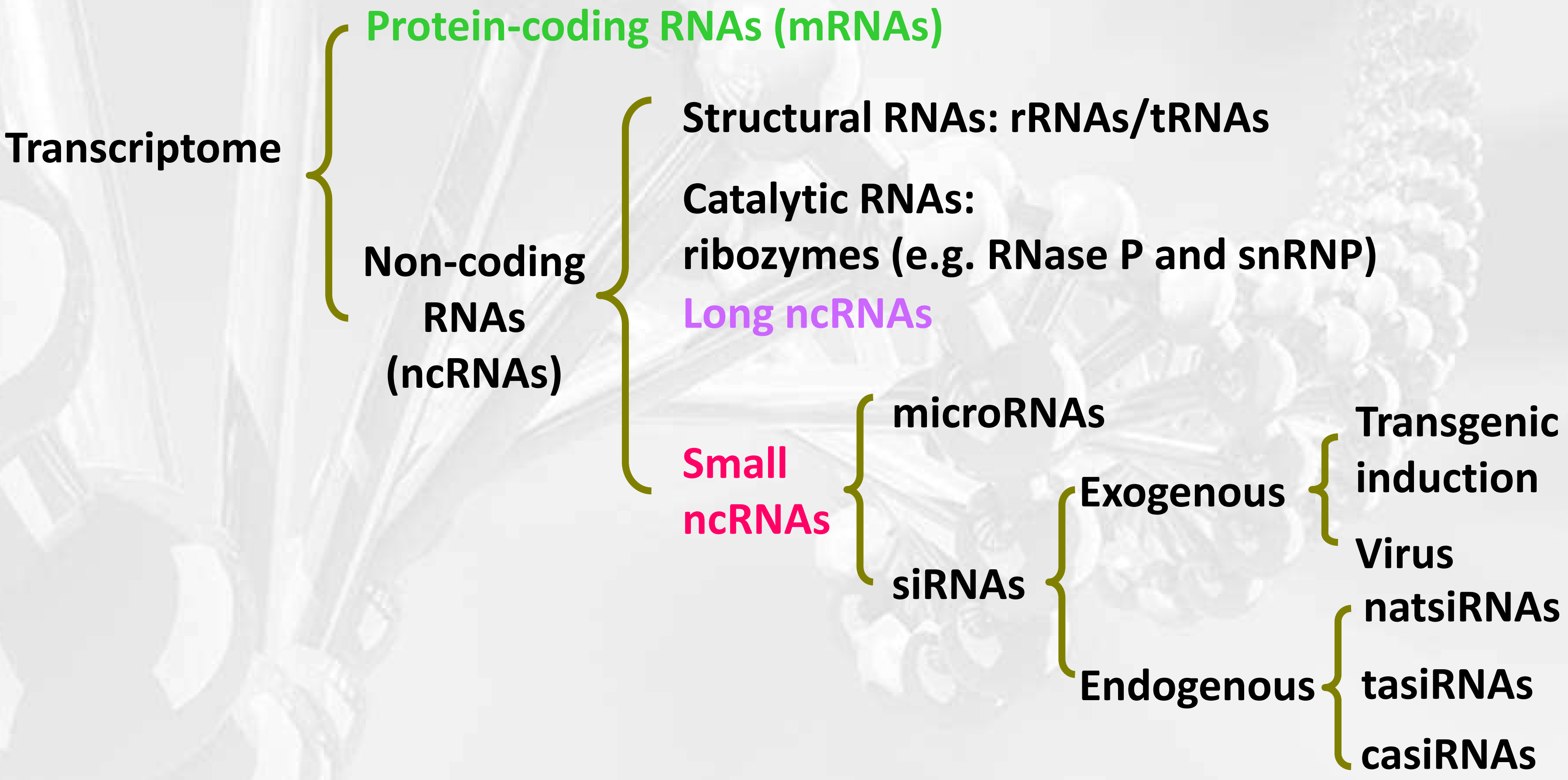
Learning about the Transcriptome

→ What is encoded on the genome and how is it processed?



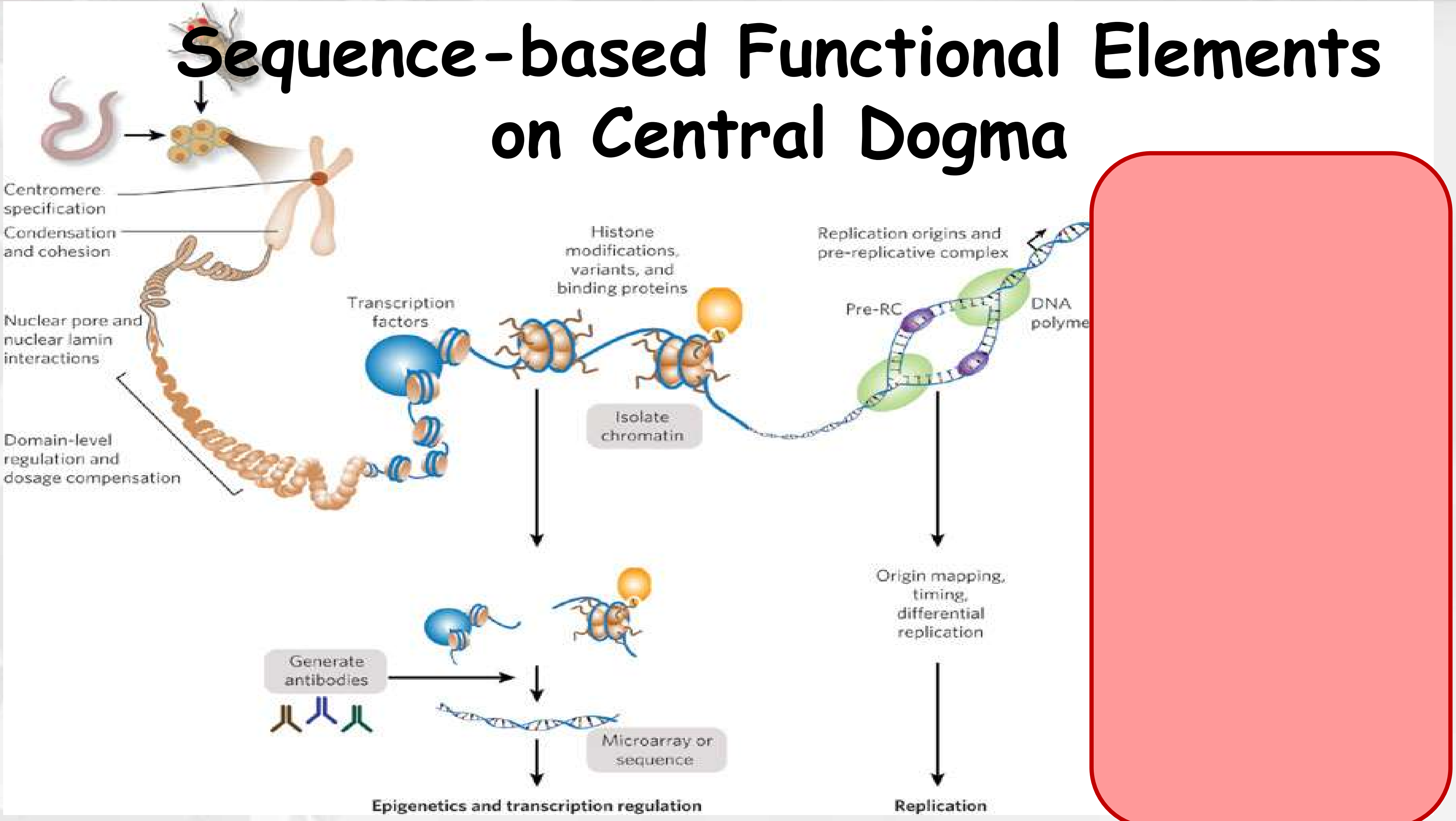
The **transcriptome** is the set of all RNA molecules, including mRNA, rRNA, tRNA, and other non-coding RNA produced in one or a population of cells.

Transcriptome



- Transcriptome
  - Protein-coding RNAs (mRNAs)
  - Non-coding RNAs (ncRNAs)
    - Structural RNAs: rRNAs/tRNAs
    - Catalytic RNAs: ribozymes (e.g. RNase P and snRNP)
    - Long ncRNAs
    - Small ncRNAs
      - microRNAs
      - siRNAs
        - Exogenous
          - Transgenic induction
          - Virus
        - Endogenous
          - natsiRNAs
          - tasiRNAs
          - casiRNAs

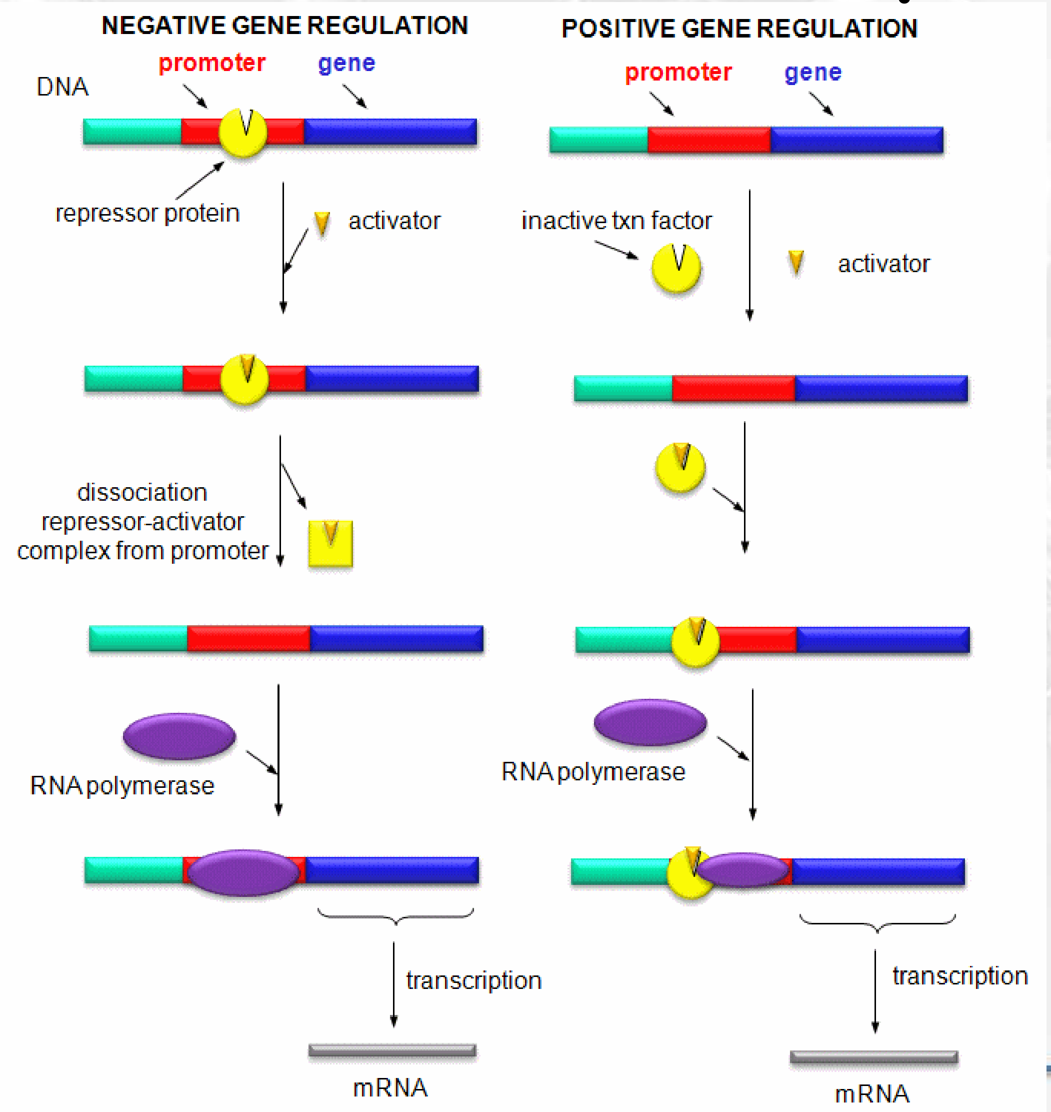
Sequence-based Functional Elements on Central Dogma


Centromere specification
Condensation and cohesion
Nuclear pore and nuclear lamin interactions
Domain-level regulation and dosage compensation
Transcription factors
Histone modifications, variants, and binding proteins
Isolate chromatin
Replication origins and pre-replicative complex
Pre-RC
DNA polyme
Generate antibodies
Microarray or sequence
Origin mapping, timing, differential replication
Epigenetics and transcription regulation
Replication

Gene expression is the process by which information from a gene is used in the synthesis of a functional gene product. These products are often proteins, but in non-protein coding genes such as rRNA, tRNA or snRNA, the product is a functional RNA.

Nature. 2009 Jun 18;459(7249):927-30.

How can gene expression be regulated at the transcriptional level?



- Chromatin domains
- Transcription
- Post-transcriptional modification
- RNA transport
- Translation
- mRNA degradation

- physiological status (nutrition, environment)
- sex and age
- various tissues and cell types
- response to stimuli (drugs, signals, toxins)
- health and disease

理论课内容

- 转录组学介绍
- 基因表达数据分析
  - 测定技术
  - 差异基因
  - 功能分析
- 几个实例
- 非编码RNA分析

Section 1: Measuring gene expression level

Quantitate gene expression level method

- Experiment-based approaches:
  a) RT-PCR
  b) Northern blot
- Hybridization-based approaches :
  a) Microarrays/chip;
  b) genomic tiling microarrays.
- Sequence-based approaches:
  a) EST: Expression Sequence Tag (~400 bp, 20-7000 bp)
  b) tag-based methods:
    - ✓ CAGE: cap analysis of gene expression (~14-20 bp, 5′ ends)
    - ✓ SAGE: serial analysis of gene expression (~14-20 bp, 3′ ends)
    - ✓ MPSS: massively parallel signature sequencing (17-20 bp)
- Next-generation Sequencing-based method:
  RNA-Seq

Nat Methods. 2008 Jul;5(7):585-7.
Annu Rev Genomics Hum Genet. 2009;10:135-51.
Nat Rev Genet. 2009 Jan;10(1):57-63.

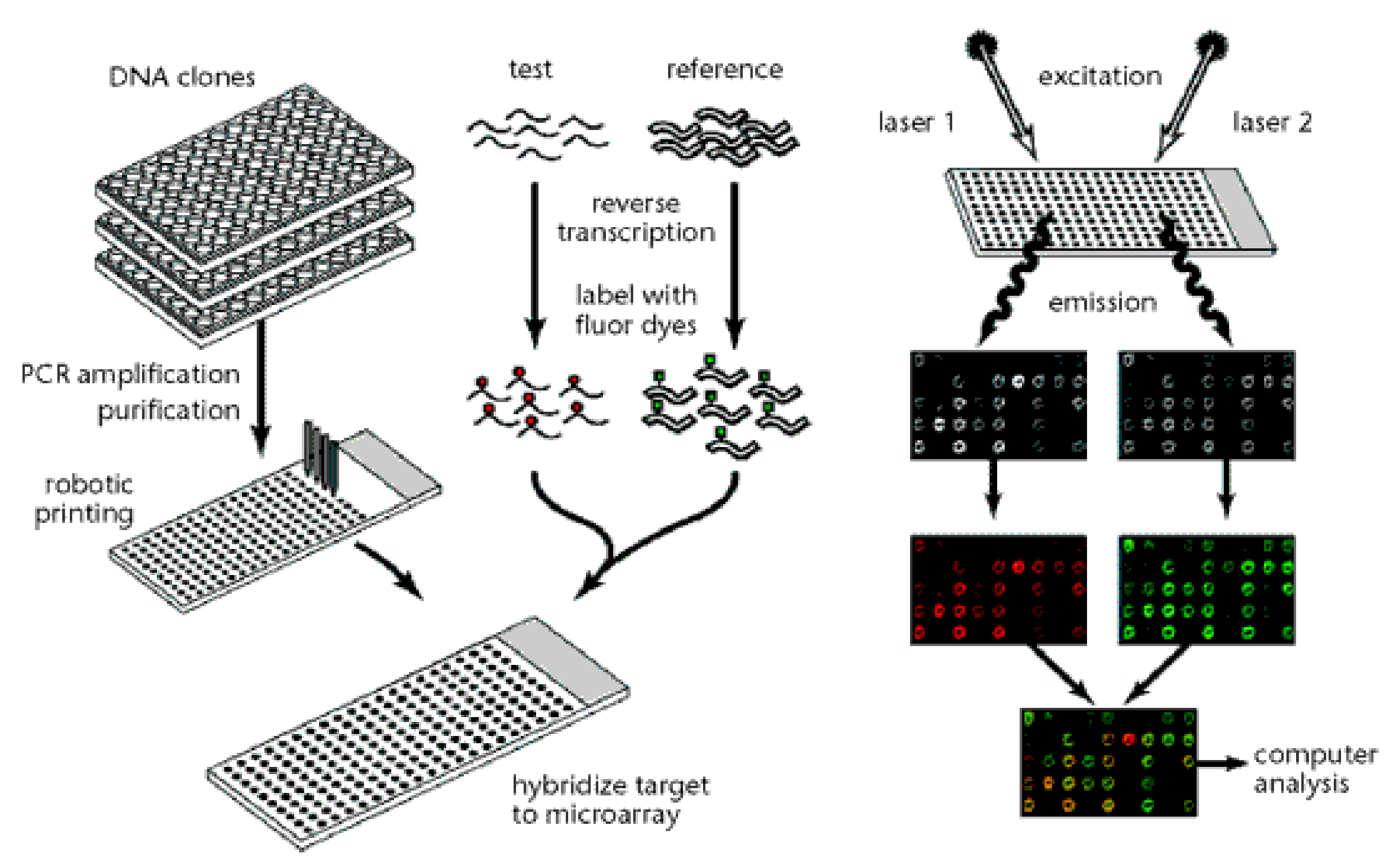
Advantages and disadvantages

- **Experiment-based approaches:**
  - Low throughput
  - expensive
- **Hybridization-based approaches :**
  - based on genome sequence;
  - cross-hybridization (high background levels);
  - limited dynamic range of detection (<1000-fold);
  - normalization problems(across different experiments).
- **Sequence-based approaches:**

  a) EST: Expression Sequence Tag (~400 bp, 20-7000 bp)
  - low throughput;
  - expensive;
  - not quantitative.

  b) tag-based methods:
  - based on expensive Sanger sequencing technology;
  - ✓ high throughput;
  - ✓ more precise;
  - a portion the short tags cannot be uniquely mapped
- **Next-generation Sequencing-based method: RNA-Seq**
  - ✓ Can be used to detect transcripts of any genome.
  - ✓ Low background, highly accurate
  - ✓ Large dynamic range of expression levels (~10000-fold)
  - ✓ High levels of reproducibility(both for technical and biological replicates)
  - ✓ Requires less RNA sample (cloning steps)
  - ✓ Lower cost

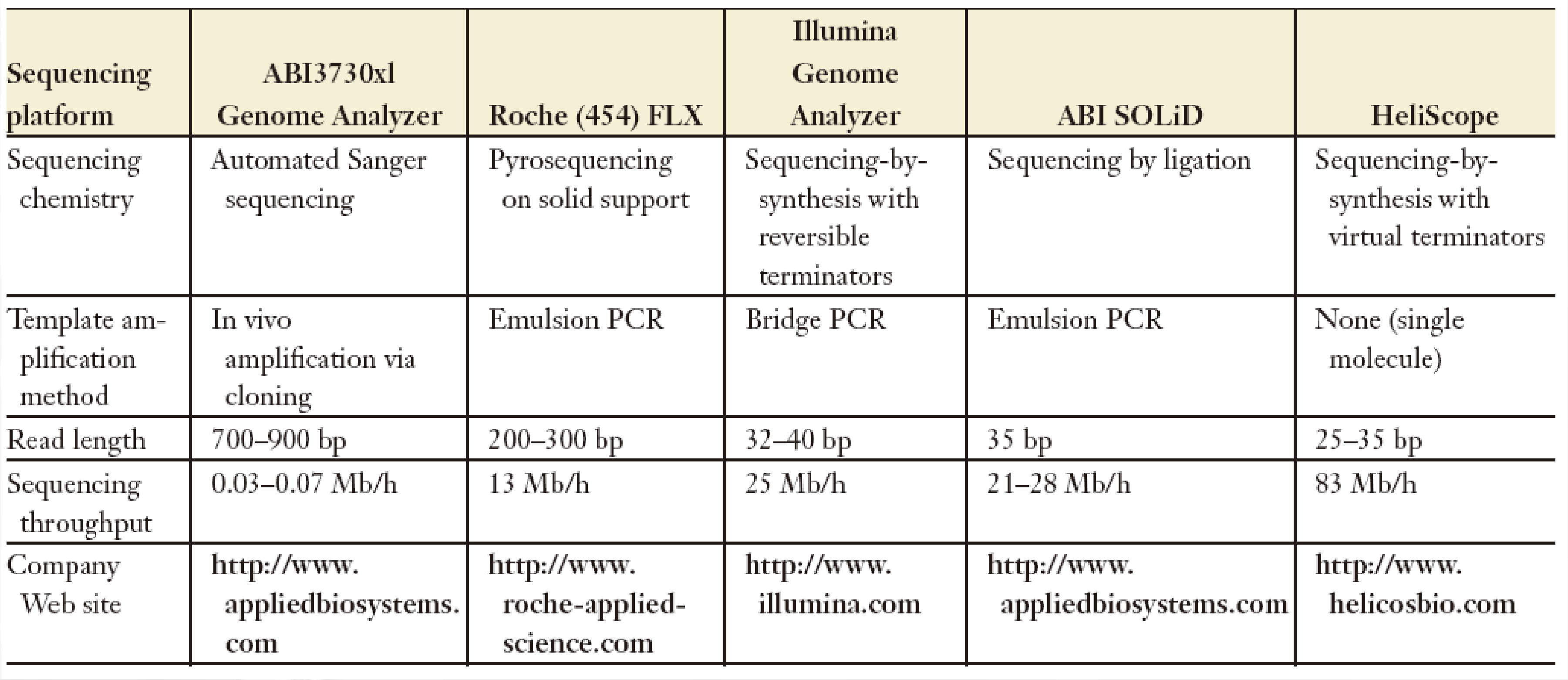
Microarray schema



From Duggan *et al. Nature Genetics* **21**, 10 – 14 (1999)

RNA-seq technologies

- Commercially available sequencing technologies used for transcriptome sequencing applications (Sep 15, 2008).



Sequencing platform	ABI3730xl Genome Analyzer	Roche (454) FLX	Illumina Genome Analyzer	ABI SOLiD	HeliScope
Sequencing chemistry	Automated Sanger sequencing	Pyrosequencing on solid support	Sequencing-by-synthesis with reversible terminators	Sequencing by ligation	Sequencing-by-synthesis with virtual terminators
Template amplification method	In vivo amplification via cloning	Emulsion PCR	Bridge PCR	Emulsion PCR	None (single molecule)
Read length	700–900 bp	200–300 bp	32–40 bp	35 bp	25–35 bp
Sequencing throughput	0.03–0.07 Mb/h	13 Mb/h	25 Mb/h	21–28 Mb/h	83 Mb/h
Company Web site	http://www.appliedbiosystems.com	http://www.roche-applied-science.com	http://www.illumina.com	http://www.appliedbiosystems.com	http://www.helicosbio.com

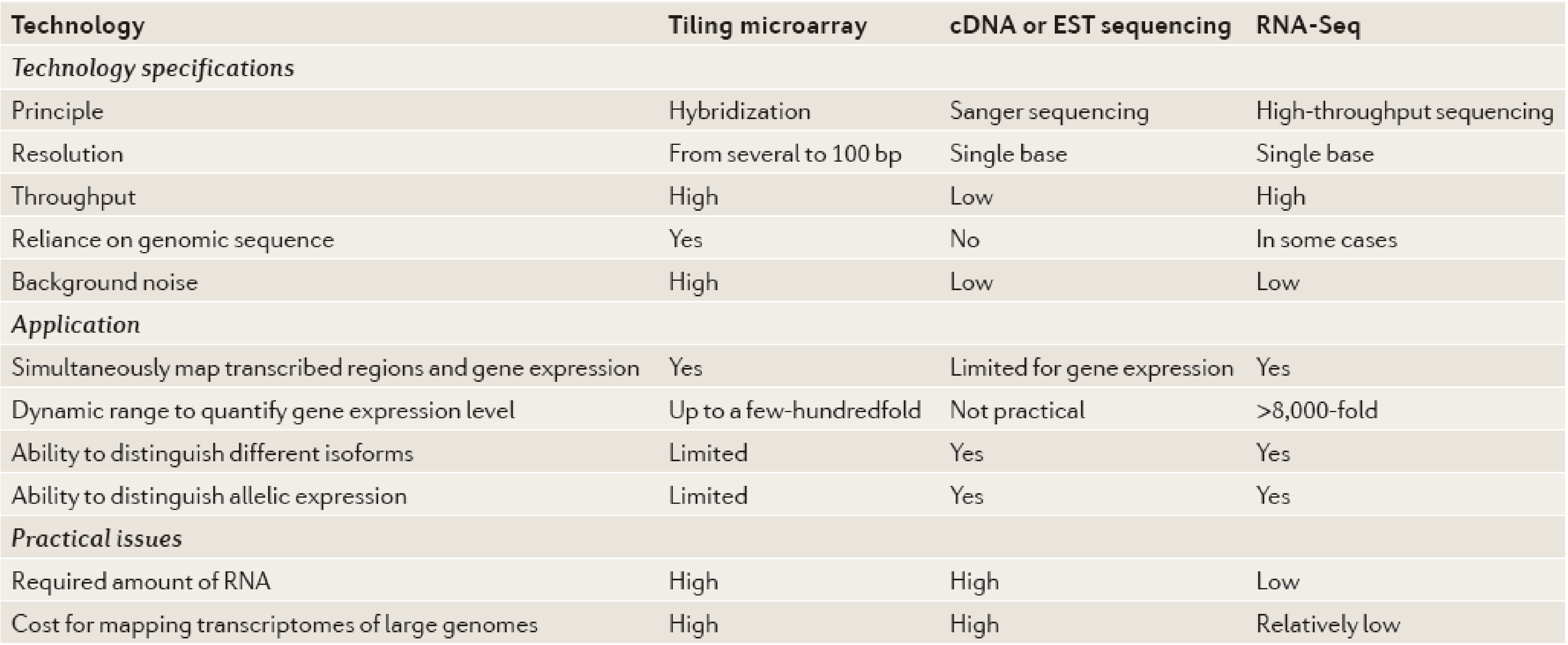
Annu Rev Genomics Hum Genet. 2009;10:135-51.

RNA-Seq: Advantages

- ◈Sequencing length: 30 - 400bp.
- ◈Advantages:
  - ➢can be used to detect transcripts of any genome.
  - ➢low background, highly accurate
  - ➢large dynamic range of expression levels (~10000-fold)
  - ➢high levels of reproducibility (both for technical and biological replicates)
  - ➢requires less RNA sample (cloning steps)
  - ➢lower cost

RNA-Seq: Advantages

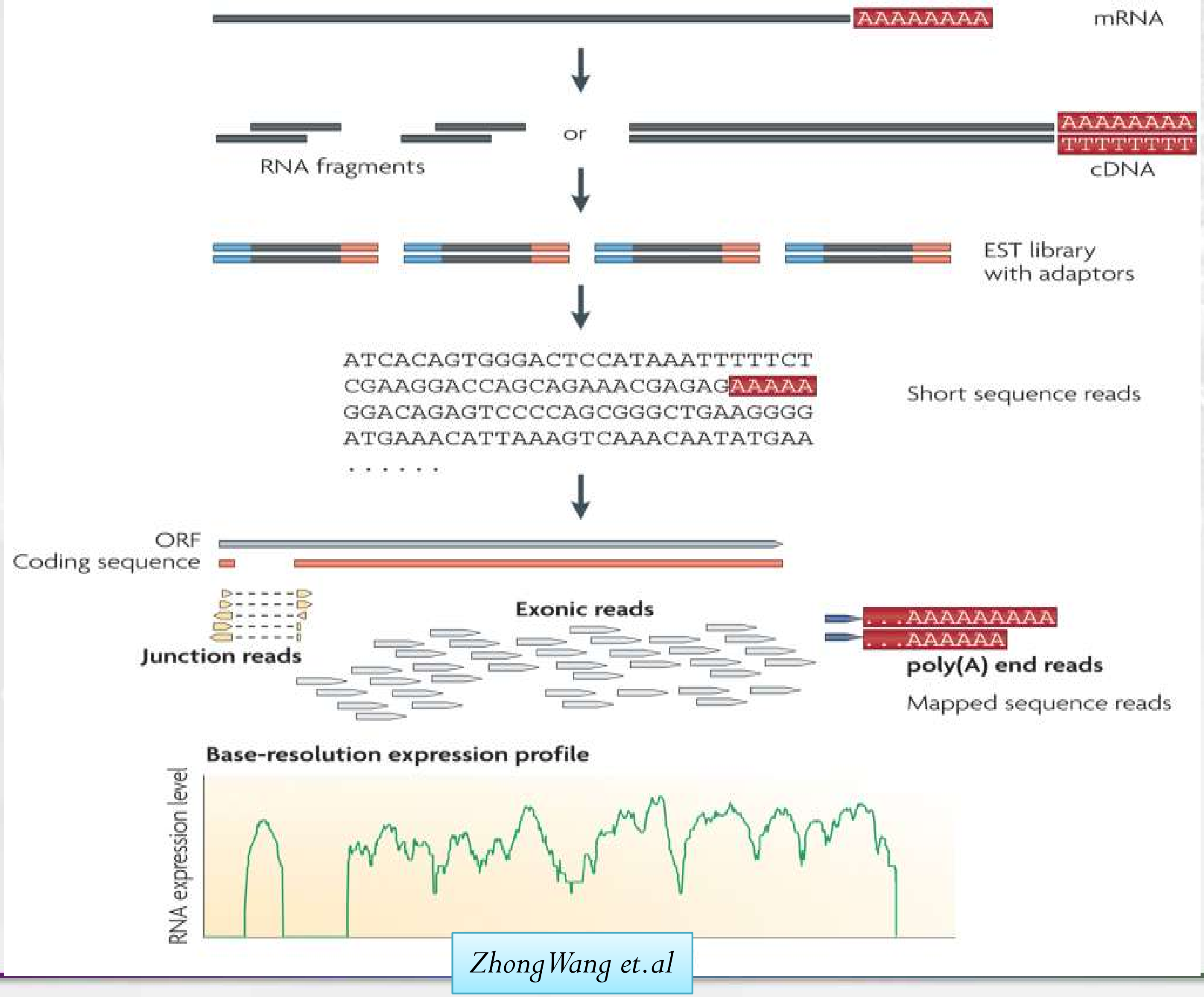
- ➢RNA-Seq v.s. other transcriptomics methods



Technology	Tiling microarray	cDNA or EST sequencing	RNA-Seq
Technology specifications			
Principle	Hybridization	Sanger sequencing	High-throughput sequencing
Resolution	From several to 100 bp	Single base	Single base
Throughput	High	Low	High
Reliance on genomic sequence	Yes	No	In some cases
Background noise	High	Low	Low
Application			
Simultaneously map transcribed regions and gene expression	Yes	Limited for gene expression	Yes
Dynamic range to quantify gene expression level	Up to a few-hundredfold	Not practical	>8,000-fold
Ability to distinguish different isoforms	Limited	Yes	Yes
Ability to distinguish allelic expression	Limited	Yes	Yes
Practical issues			
Required amount of RNA	High	High	Low
Cost for mapping transcriptomes of large genomes	High	High	Relatively low

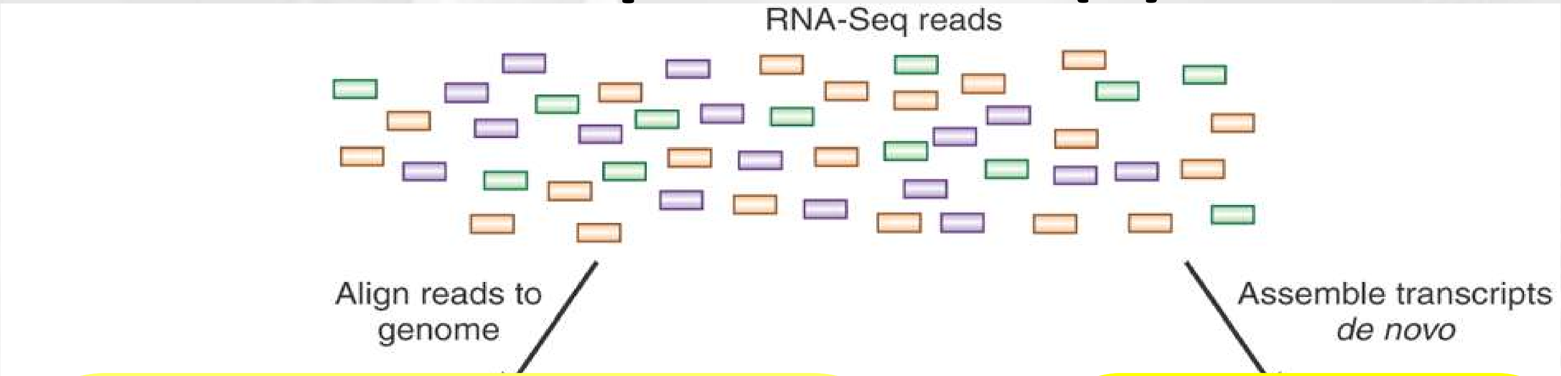
Nat Rev Genet. 2009 Jan;10(1):57-63.

RNA-seq workflow (1)



Zhong Wang et.al

RNA-seq workflow (2)



Mapping-first approaches:
Cufflinks,
Scripture

Assembly-first (*de novo*) approaches:
ABySS,
Trinity

Brian J Haas & Michael C Zody. *Nat Biotechnol.* 2010 28(5):421-3.

Gene expression level measurement for RNA-seq

✓RPKM : Reads per kilobase per million mapped reads.

$$RPKM = \frac{Total\ exon\ reads}{mapped\ reads(millions) \times exon\ length(KB)}$$

1kb transcript with 1000 alignments in a sample of 10 million reads (out of which 8 million reads can be mapped) will have RPKM = 1000/(1 * 8) = 125

✓FPKM : Fragments Per Kilobase of exon per Million fragments mapped (for paired-end sequencing).

$$\text{RPKM} = \frac{\textit{total exon reads}}{\textit{mapped reads (millions)} * \textit{exon length (KB)}}$$

假设一基因体只有两个基因，一个9 KB，一个1 KB，如今有一sample，其map 到9 KB 的read 有18 million 个，map 到1 KB 的有2 million 个，

- 对于9 KB 的基因而言，

  Total exon reads=18 million

  Mapped reads=18+2=20 million

  Exon length=9 KB

  RPKM =18million/(20*9)=0.1*10^6=10^5

- 对于1 KB 的基因而言，

  Total exon reads=2 million

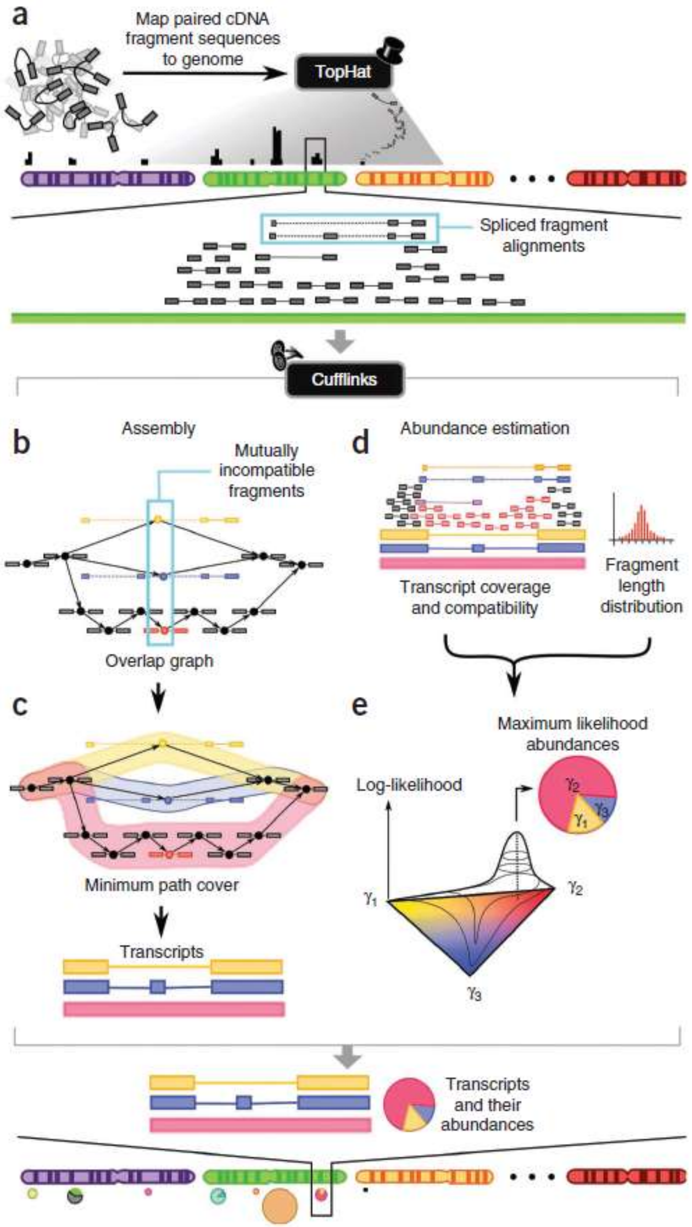
  Mapped reads=18+2=20 million

  Exon length=1 KB

  RPKM =2million/(20*1)=0.1*10^6=10^5

由此我们可以知道这两个基因表现量没有差别。

Cufflinks

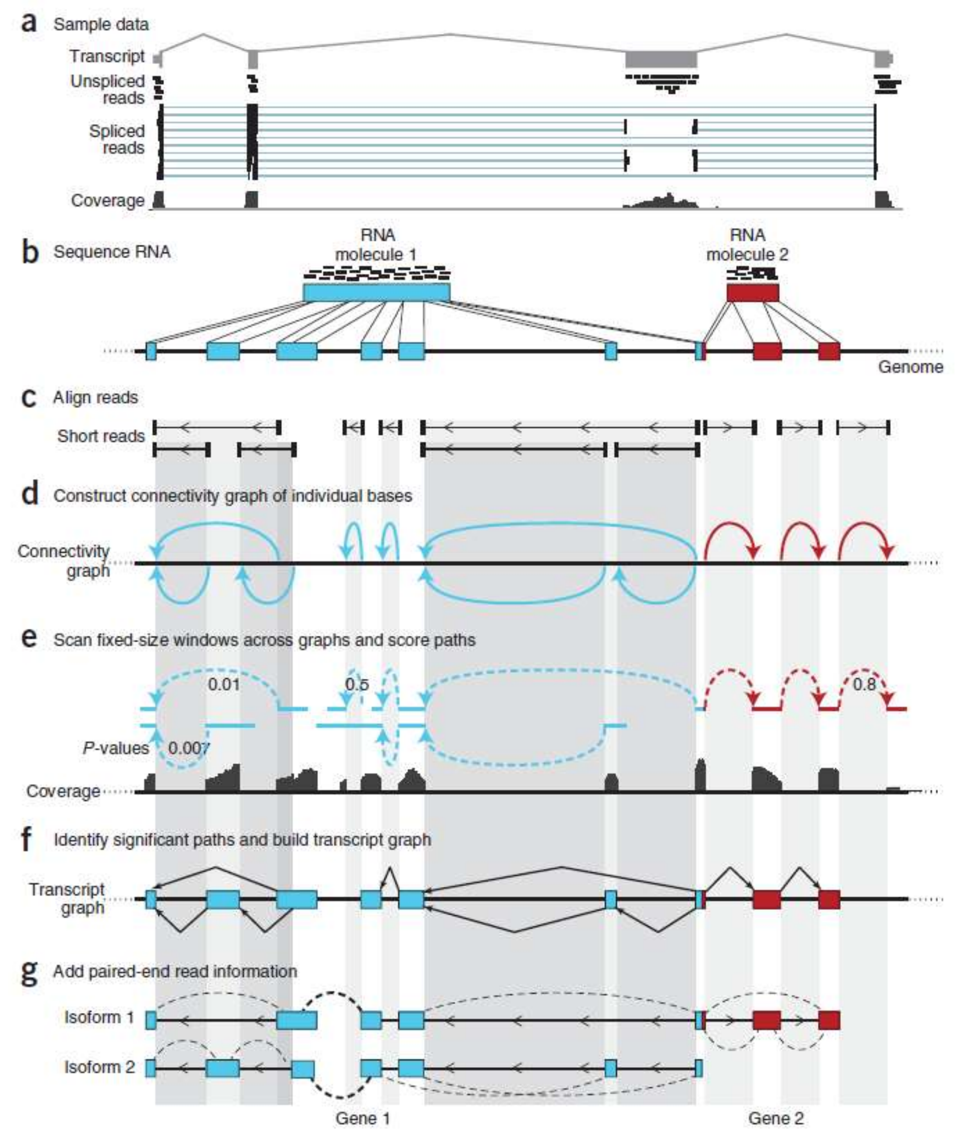

a
Map paired cDNA fragment sequences to genome
TopHat
Spliced fragment alignments
Cufflinks
b
Assembly
Mutually incompatible fragments
Overlap graph
c
Minimum path cover
Transcripts
d
Abundance estimation
Transcript coverage and compatibility
Fragment length distribution
e
Maximum likelihood abundances
Log-likelihood
γ1
γ2
γ3
Transcripts and their abundances

Cufflinks uses a rigorous mathematical model to identify the complete set of alternatively regulated transcripts at each locus and to assign coverage to each transcript.

Cufflinks 利用Tophat比对的结果（alignments）来组装转录本，估计这些转录本的丰度，并且检测样本间的差异表达及可变剪接。

Nat Biotechnol. 2010 May;28(5):511-5.

Scripture

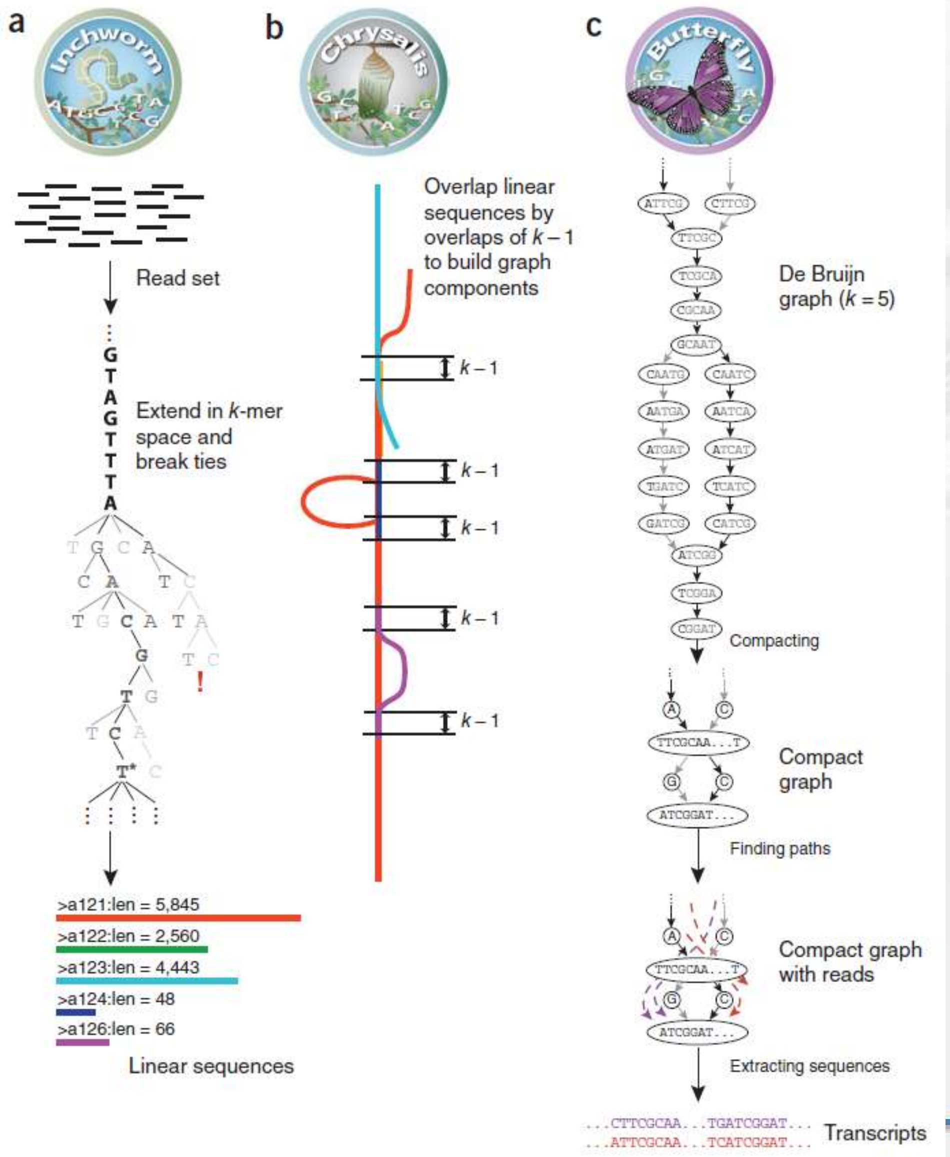


Scripture employs a statistical segmentation model to distinguish expressed loci and filter out experimental noise.

Cufflinks可根据reads映射到参考基因组的结果来预测新基因和亚型。Scripture采用统计学分段模型来区分表达位点和实验噪声。

Nat Biotechnol. 2010 May;28(5):503-10.

Trinity

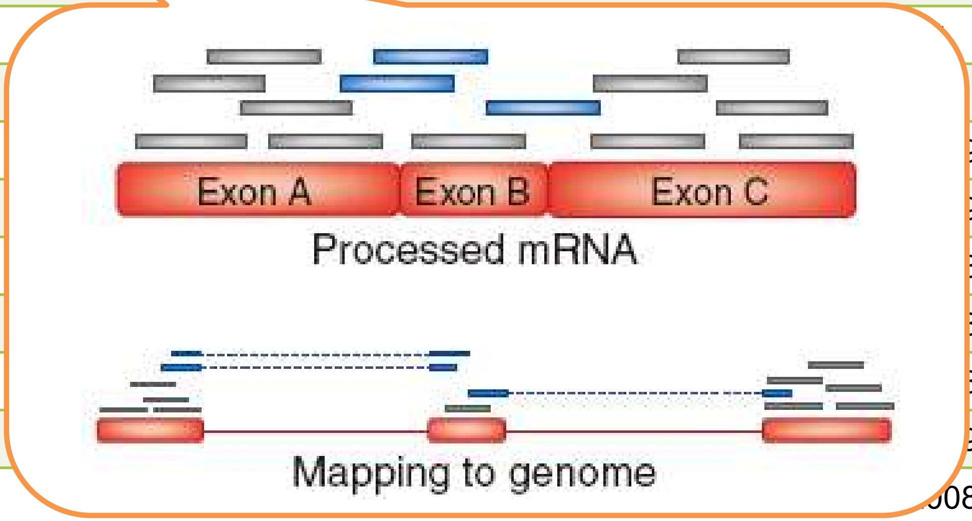


Trinity: *de novo* assembly of full-length transcripts without a reference genome, consisting of three software modules: Inchworm, Chrysalis and Butterfly

Inchworm: 将RNA-seq的原始reads数据组装成Unique序列；
Chrysalis: 将上一步生成的contigs聚类，然后对每个类构建Bruijn图；
Butterfly: 处理这些Bruijn图，依据图中reads和成对的reads来寻找路径，从而得到具有可变剪接的全长转录子，同时将旁系同源基因的转录子分开。

Nat Biotechnol. 2011 May 15.

Program	Website	Publications
BLAST	http://www.ncbi.nlm.nih.gov/blast/	1990, J. Mol. Biol.
BLAT	http://www.soe.ucsc.edu/~kent/src/	2002, Genome Research
Cross_match	http://www.phrap.org/phredphrapconsed.html	***
ELAND	http://www.illumina.com/	***
TopHat	http t.cbcb.umd.edu/	2009, Bioinformatics
Novoalign		
Mosaik		
Bowtie		9, Genome Biology
BWA		9, Bioinformatics
MAQ		3, Genome Research
SOAP/SOAP2		3/2009, Bioinformatics
ZOOM		3, Bioinformatics
PerM		9, Bioinformatics
BWT-SW		008, Bioinformatics
RMAP	http://rulai.cshl.edu/rmap/	2008, BMC Bioinformatics
SHRiMP	http://compbio.cs.toronto.edu/shrimp/	2009, PLoS Computational Biology
SeqMap	http://biogibbs.stanford.edu/~jiangh/SeqMap/	2008, Bioinformatics
MOM	http://mom.csbc.vcu.edu/	2009, Bioinformatics
ProbMatch	http://www.cs.wisc.edu/~jignesh/probematch/	2009, Bioinformatics
Exonerate	http://www.ebi.ac.uk/~guy/exonerate/	2005, BMC Bioinformatics
SSAHA2	http://www.sanger.ac.uk/Software/analysis/SSAHA2/	2001, Genome Research
Edena	http://www.genomic.ch/edena	2008, Genome Research
VCAKE	http://sourceforge.net/projects/vcake/	2007, Bioinformatics
Euler-SR	***	2007, Genome Research



Section 2: Identifying differentially expressed genes

Statistical methods for finding differentially expressed genes

- **Comparing two independent groups**
  a) T-test
  b) Linear regression model
  Normal distribution
  c) Wilcoxon rank sum test
  d) SAM
  Any distribution
- **Comparing more than two groups**
  a) F-test
  b) Linear regression model
  Normal distribution
  c) Wilcoxon rank sum test
  d) SAM
  Any distribution
- **Software: R language (Bio-conductor)**

R

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

T-test

- ✓ Suppose we want to find genes that are differentially expressed between different conditions/phenotypes, e.g. two different tumor types.

Tumor sample	1 1	1 2	1 3	1 4	2 1	2 2	2 3	2 4
gene1	X1	X2	X3	X4	Y1	Y2	Y3	Y4
gene2								
gene3								

$\bar{X}_1$ $\bar{X}_2$

[redacted]

[redacted]

- ● Need check normal assumption
- ● More arrays in each group more confidence in results

- ✓ **After a test statistic is computed, it is convenient to convert itto a p-value.** $P\ value = P(t > T(X,Y))$

➢ Linear regression model

- ✓ **Expression of gene x is made of a baseline expression level (from control group), plus the group effect.**

$$Y = Y_0 + \beta Z$$

Y_0: baseline exp. Level; β: group effect; Z: group variable (0 for control obs., 1 for group obs.)

- ✓ **P-value can be used to test group effect.**

ANOVA Table

	d.f.	Sum Sq	Mean Sq	F statistic	p-value
Group	1	29.4115	29.4115	31.323	0.000512
Residuals	8	7.5119	0.939		

- ✓ Results – one p-value per gene

Linear regression model

✓ **Expression of gene x : baseline expression level, group effect and patient age group**

$$Y = Y_0 + \beta Z + \gamma W$$

Y_0: baseline exp. Level;
β: group effect;
Z: group variable (0 for control obs., 1 for group obs.
γ: age effect
W: age variable (0 for 0-15, 1 for 16-29, 2 for 30+)

✓ **ANOVA table:**

	d.f.	Sum Sq	Mean Sq	F statistic	p-value
Treatment	1	20.6848	20.6848	25.9737	0.000263
Age	2	27.2838	13.6419	17.13	0.000305
Treatment:Age	2	0.5526	0.2763	0.3469	0.713707
Residuals	12	9.5565	0.7964		

✓ **Results: a list of p-values**

Wilcoxon rank sum test

- ✓ Non–parametric test for equality of two distributions.
- ✓ Compute the ranks of observations in the pooled sample.

  Observations: 0:3 0:5 0:8 0:9 1:3 2:4

  Ranks: 1 2 3 4 5 6

  Groups: 1 1 1 2 2 2
- ✓ The test statistic is a function of the sum of ranks in group 1;

  here, R1 = 6.
- ✓ For small sample sizes, the null distribution of the test statistic can be computed exactly. For large sample size, a normal approximation is used.
- ✓ Advantage: Non–parametric, robust against outliers

➢ SAM

- ✓ **Does not assume normal distribution.**

  --Instead, p-values computed via permutation

- ✓ **The SAM ('significance analysis of microarrays') test statistic is**

$$S = \frac{R_g}{c + SE_g}$$

R_g be the mean log ratio of the expression levels of one gene;

SE_gbe its standard error;

constant c can be taken to be the 90th percentile SEg value.

- ✓ **One p-value per gene**

Multiple testing: the problems

- ✓ **Type I: or false-positive error occurs when we declare a gene to be differentially expressed when in fact it is not.**
- ✓ **Type II: or false-negative error occurs when we fail to detect a differentially expressed gene.**
- ✓ **The available methods to address the problems:**

a) **Family-wise error-rate control:** One approach to multiple testing is to control the family-wise error rate (FWER), which is the probability of accumulating one or more false-positive errors over a number of statistical tests.

b) **False-discovery-rate control:** An alternative approach to multiple testing considers the false-discovery rate (FDR), which is the proportion of false positives among all of the genes initially identified as being differentially expressed - that is, among all the rejected null hypotheses.

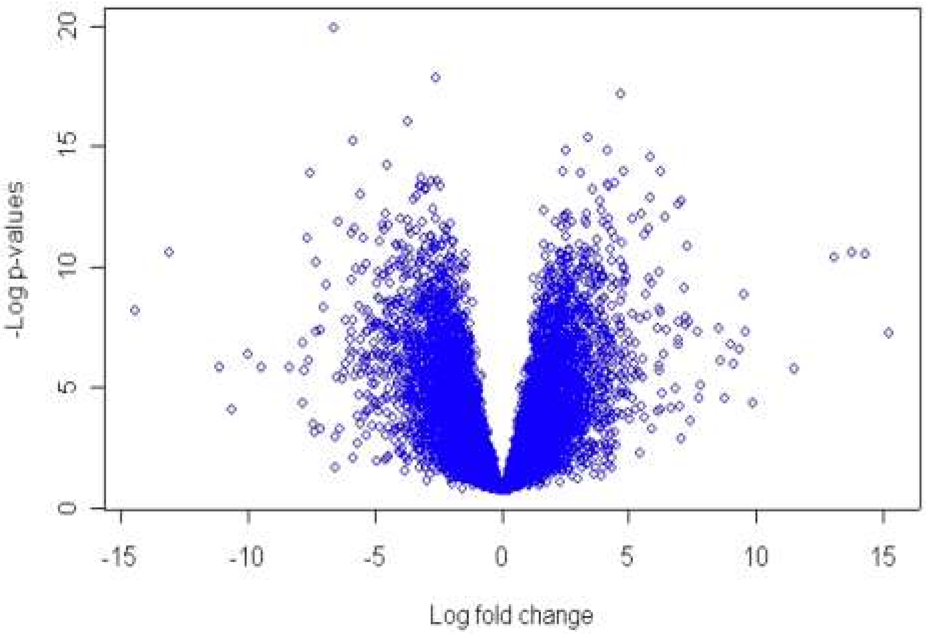
➢ P-value vs. Fold change

- ✓ P-values measure distance in terms of probability.
  – Statistical significance
- ✓ Fold changes: measure distance in arbitrary scale.

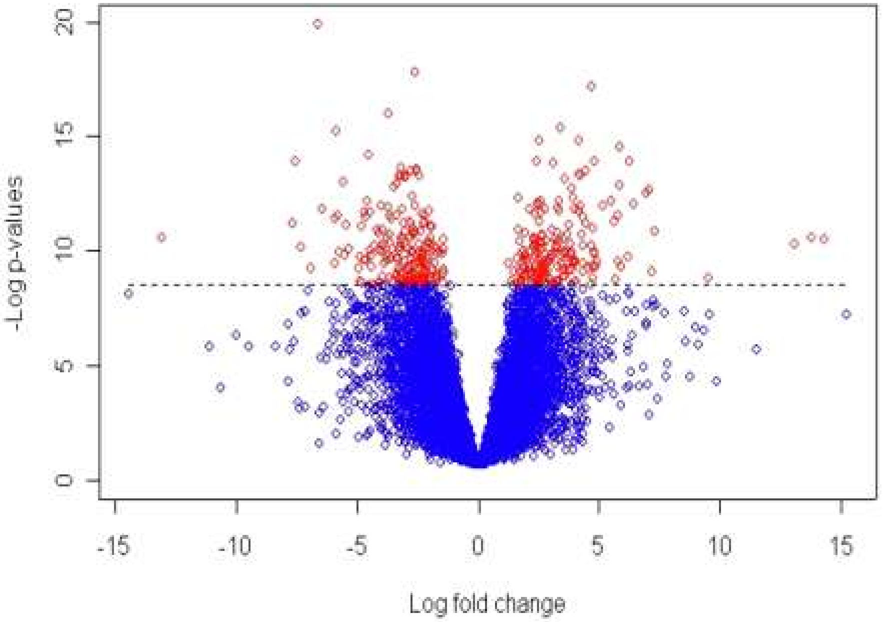
  The simplest method for identifying differentially expressed genes is to evaluate the log ratio between two conditions (or the average of ratios when there are replicates) and consider all genes that differ by more than an arbitrary cut-off value to be differentially expressed.

  – Biological meaning
- ✓ Differentially expressed gene selection: Need combination of these two.

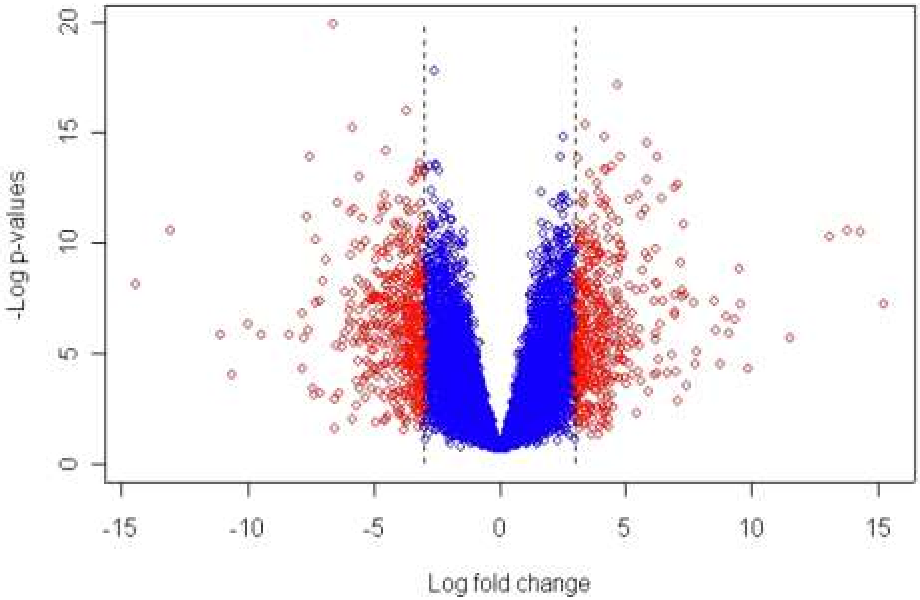
Volcano plot



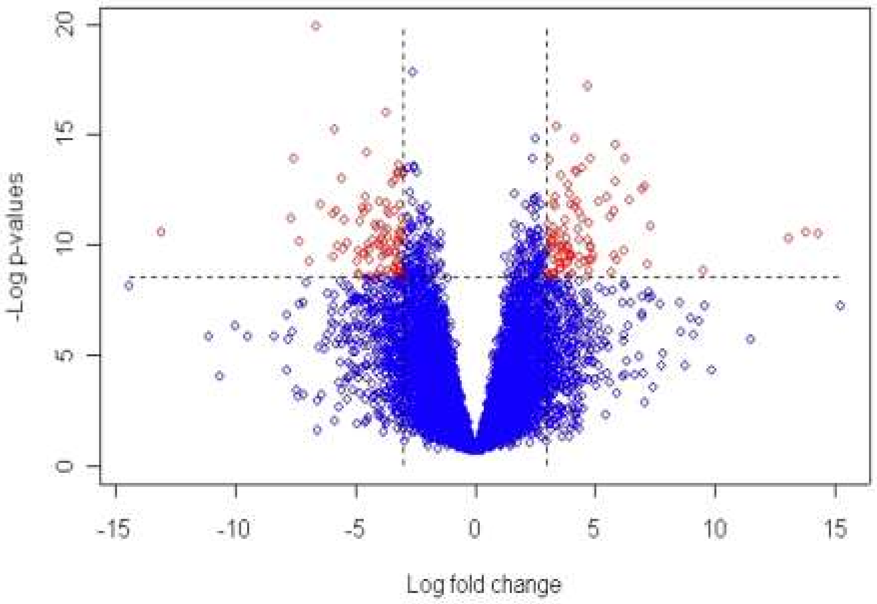
Selection statistically significant (FDR)



Selection biologically meaningful



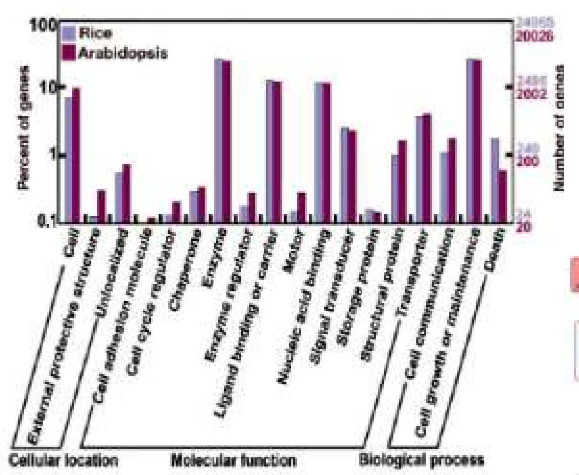
Combining the two criteria

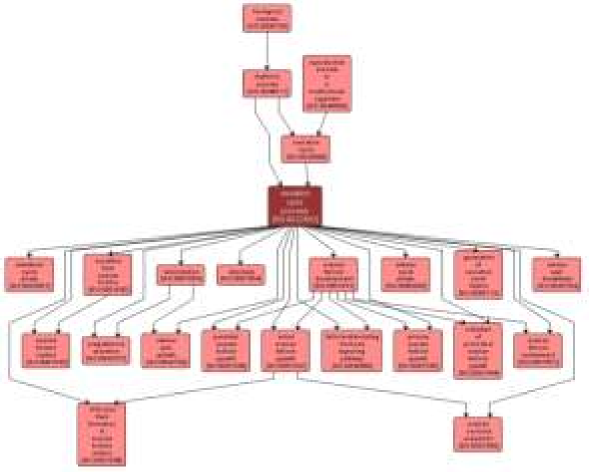


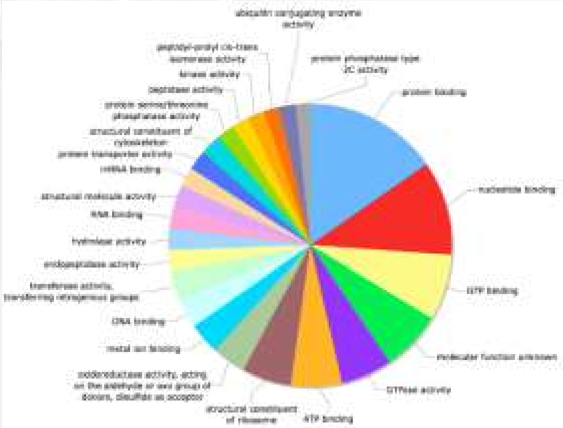
Section 3: Advanced analysis

GO analysis

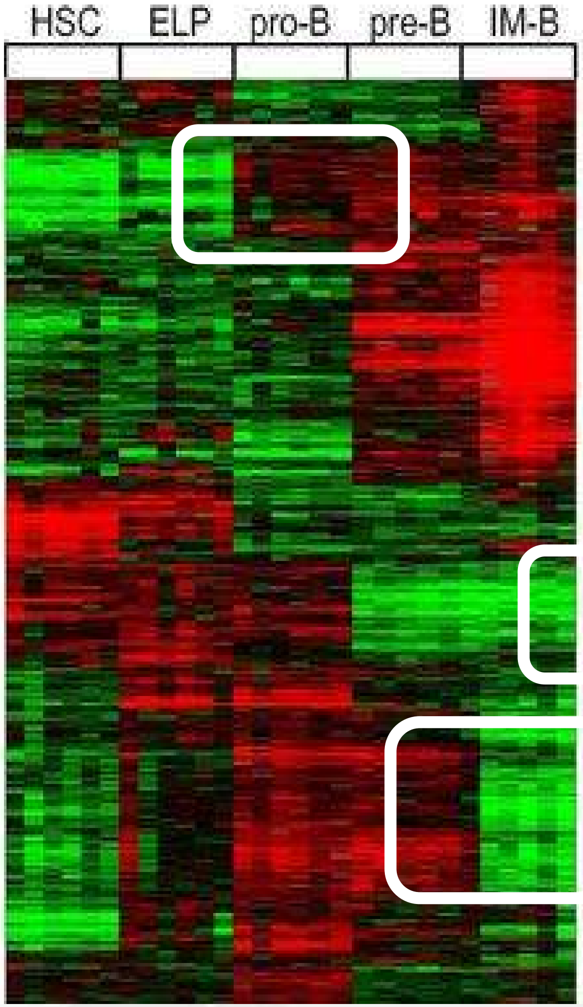
- ✓ **The Gene Ontology**, or GO, is a major bioinformatics initiative to unify the representation of gene and gene product attributes across all species.
- ✓ **Tools:** AmiGO (http://amigo.geneontology.org/cgi-bin/amigo/blast.cgi?session_id=6985amigo1343799107
  OBO-Edit (http://oboedit.org/)
  WEGO (http://wego.genomics.org.cn/cgi-bin/wego/index.pl).
- ✓ **Inputs:** FASTA file, GO number list… …
- ✓ **Outputs:** Histogram, Interactive GO graph, Pie Charts… …

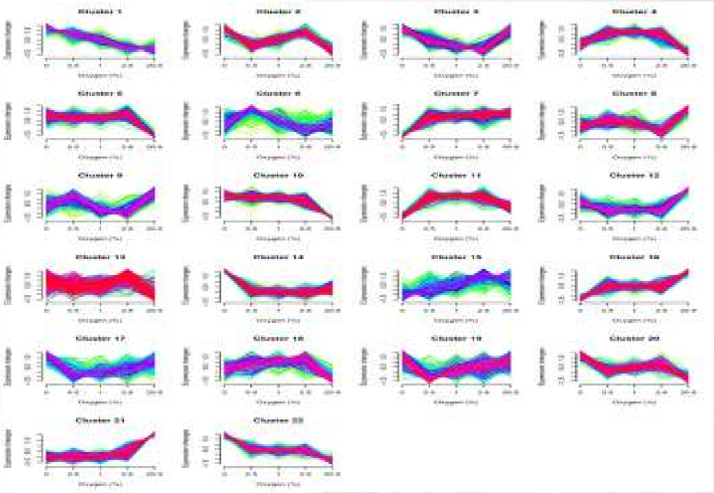


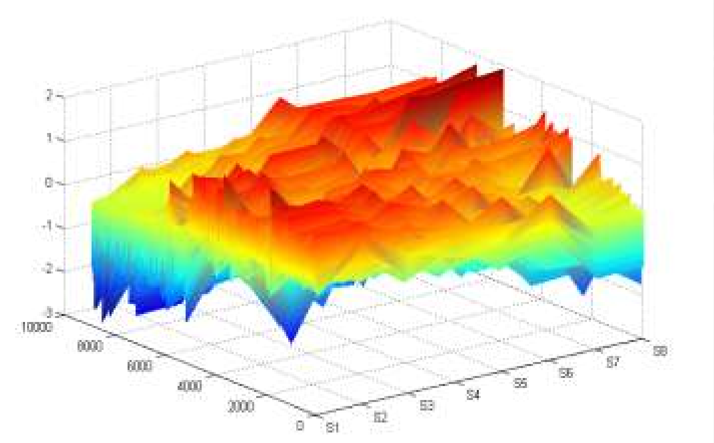



ubiquitin conjugating enzyme activity
peptidyl-prolyl cis-trans isomerase activity
kinase activity
peptidase activity
protein serine/threonine phosphatase activity
structural constituent of cytoskeleton
protein transporter activity
mRNA binding
structural molecule activity
RNA binding
hydrolase activity
endopeptidase activity
transferase activity, transferring nitrogenous groups
DNA binding
metal ion binding
oxidoreductase activity, acting on the aldehyde or oxo group of donors, disulfide as acceptor
structural constituent of ribosome
ATP binding
GTPase activity
molecular function unknown
GTP binding
nucleotide binding
protein binding
protein phosphatase type 2C activity

Clustering gene expression data







Algorithms:

a) K-means

b) Hierarchical clustering

c) K-median

d) Bi-clustering

Tools and software:

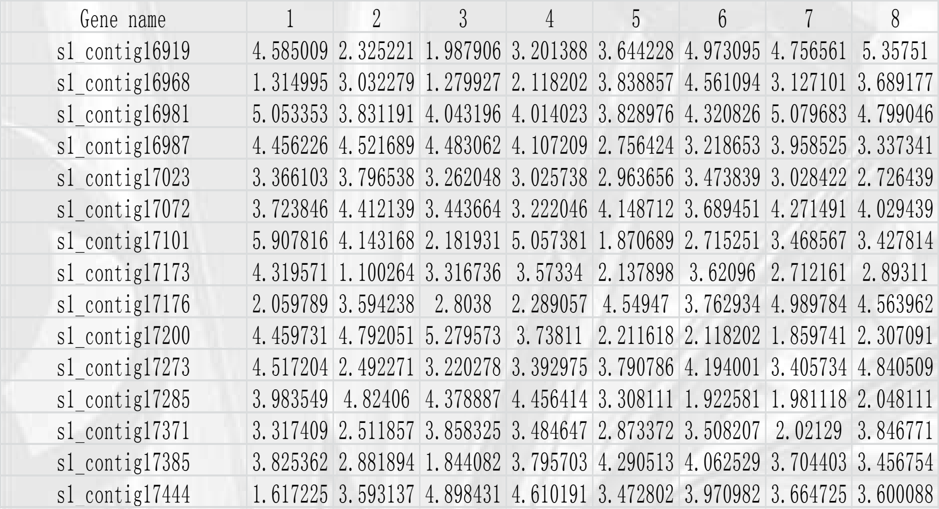
a) R language,

b) Clustal,

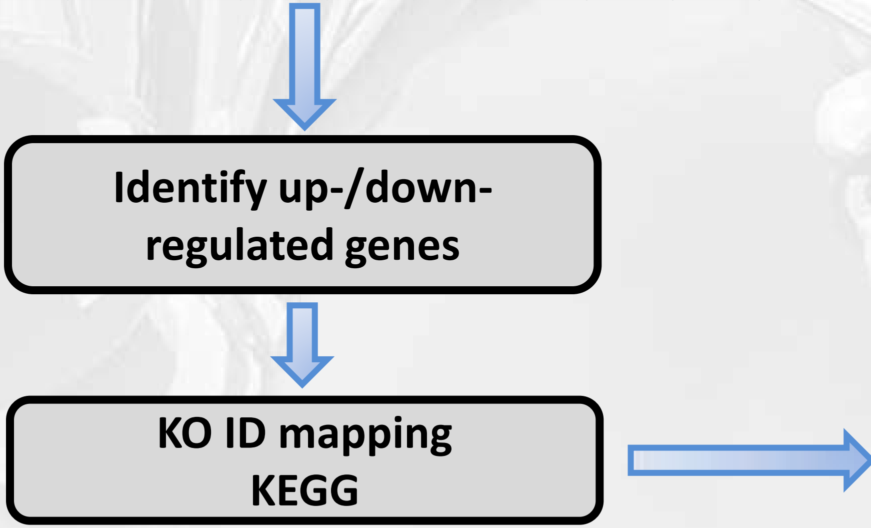
c) Mev.

If two genes are related (have similar functions or are co-regulated), their expression profiles should be similar (e.g. low Euclidean distance or high correlation).

Pathway mapping and analysis

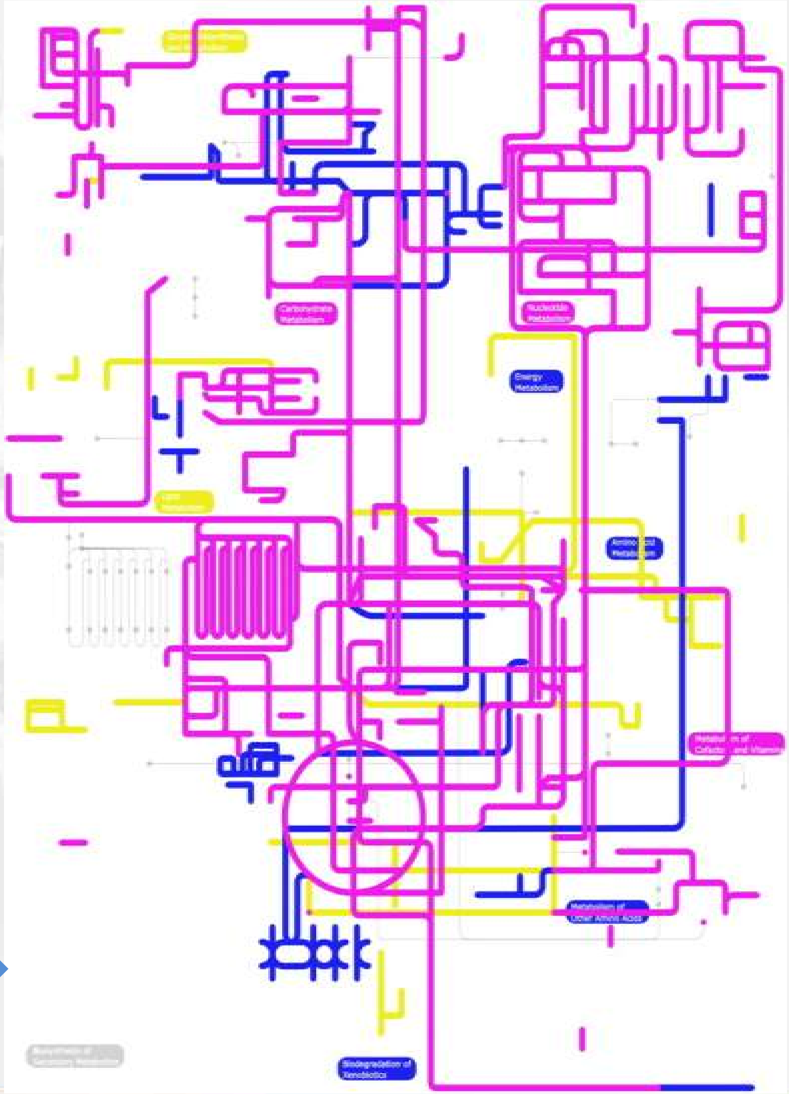


Gene name	1	2	3	4	5	6	7	8
sl_contig16919	4.585009	2.325221	1.987906	3.201388	3.644228	4.973095	4.756561	5.35751
sl_contig16968	1.314995	3.032279	1.279927	2.118202	3.838857	4.561094	3.127101	3.689177
sl_contig16981	5.053353	3.831191	4.043196	4.014023	3.828976	4.320826	5.079683	4.799046
sl_contig16987	4.456226	4.521689	4.483062	4.107209	2.756424	3.218653	3.958525	3.337341
sl_contig17023	3.366103	3.796538	3.262048	3.025738	2.963656	3.473839	3.028422	2.726439
sl_contig17072	3.723846	4.412139	3.443664	3.222046	4.148712	3.689451	4.271491	4.029439
sl_contig17101	5.907816	4.143168	2.181931	5.057381	1.870689	2.715251	3.468567	3.427814
sl_contig17173	4.319571	1.100264	3.316736	3.57334	2.137898	3.62096	2.712161	2.89311
sl_contig17176	2.059789	3.594238	2.8038	2.289057	4.54947	3.762934	4.989784	4.563962
sl_contig17200	4.459731	4.792051	5.279573	3.73811	2.211618	2.118202	1.859741	2.307091
sl_contig17273	4.517204	2.492271	3.220278	3.392975	3.790786	4.194001	3.405734	4.840509
sl_contig17285	3.983549	4.82406	4.378887	4.456414	3.308111	1.922581	1.981118	2.048111
sl_contig17371	3.317409	2.511857	3.858325	3.484647	2.873372	3.508207	2.02129	3.846771
sl_contig17385	3.825362	2.881894	1.844082	3.795703	4.290513	4.062529	3.704403	3.456754
sl_contig17444	1.617225	3.593137	4.898431	4.610191	3.472802	3.970982	3.664725	3.600088

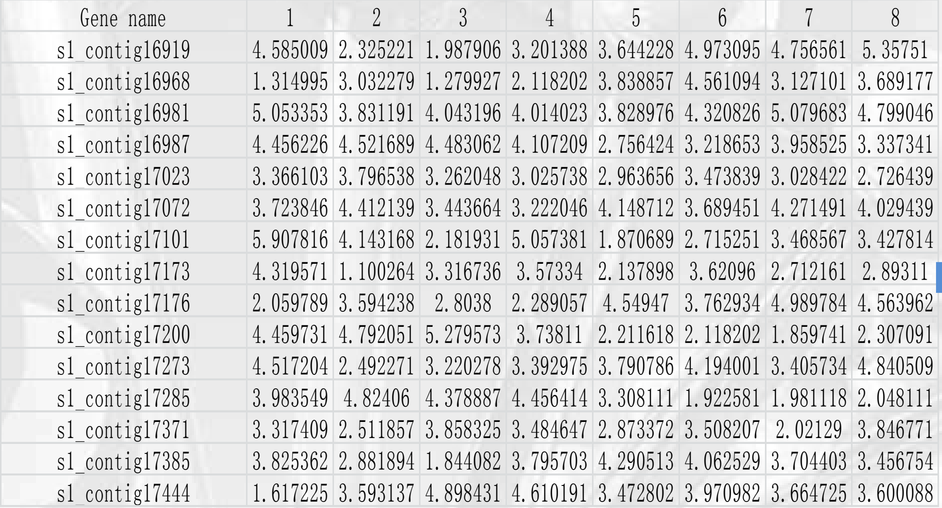


Identify up-/down-regulated genes

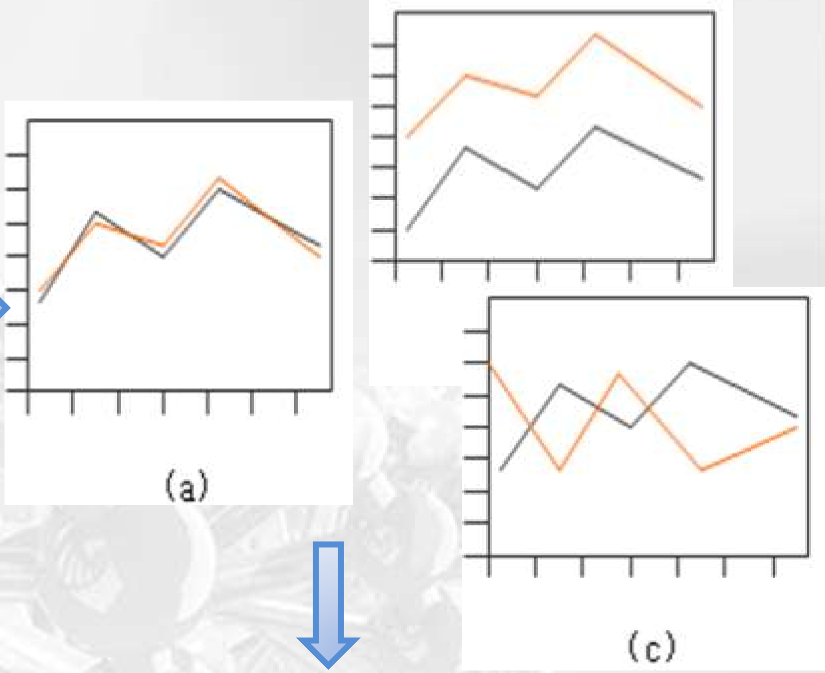
KO ID mapping
KEGG



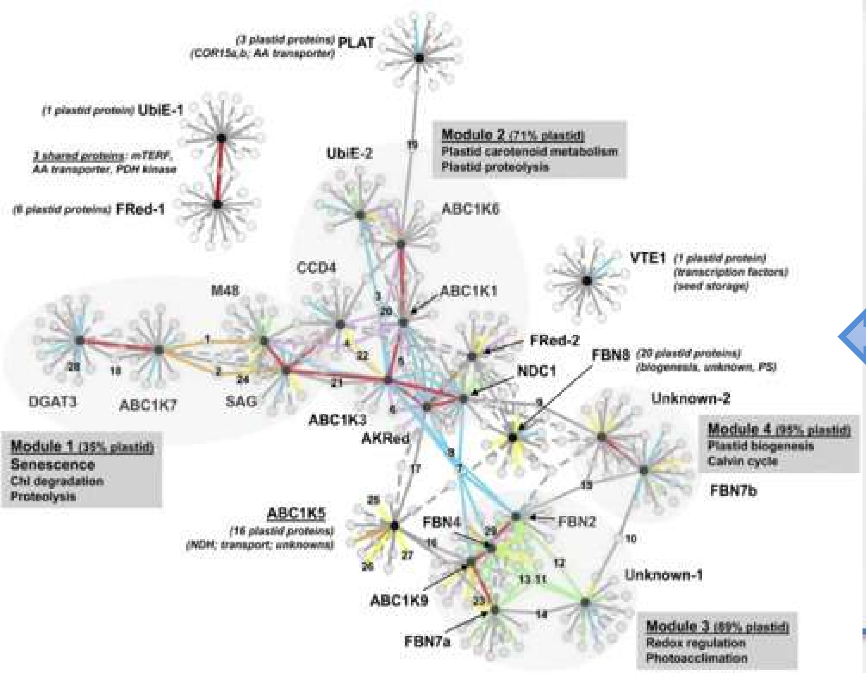
Co-expression network reconstruction



Gene name	1	2	3	4	5	6	7	8
sl_contig16919	4.585009	2.325221	1.987906	3.201388	3.644228	4.973095	4.756561	5.35751
sl_contig16968	1.314995	3.032279	1.279927	2.118202	3.838857	4.561094	3.127101	3.689177
sl_contig16981	5.053353	3.831191	4.043196	4.014023	3.828976	4.320826	5.079683	4.799046
sl_contig16987	4.456226	4.521689	4.483062	4.107209	2.756424	3.218653	3.958525	3.337341
sl_contig17023	3.366103	3.796538	3.262048	3.025738	2.963656	3.473839	3.028422	2.726439
sl_contig17072	3.723846	4.412139	3.443664	3.222046	4.148712	3.689451	4.271491	4.029439
sl_contig17101	5.907816	4.143168	2.181931	5.057381	1.870689	2.715251	3.468567	3.427814
sl_contig17173	4.319571	1.100264	3.316736	3.57334	2.137898	3.62096	2.712161	2.89311
sl_contig17176	2.059789	3.594238	2.8038	2.289057	4.54947	3.762934	4.989784	4.563962
sl_contig17200	4.459731	4.792051	5.279573	3.73811	2.211618	2.118202	1.859741	2.307091
sl_contig17273	4.517204	2.492271	3.220278	3.392975	3.790786	4.194001	3.405734	4.840509
sl_contig17285	3.983549	4.82406	4.378887	4.456414	3.308111	1.922581	1.981118	2.048111
sl_contig17371	3.317409	2.511857	3.858325	3.484647	2.873372	3.508207	2.02129	3.846771
sl_contig17385	3.825362	2.881894	1.844082	3.795703	4.290513	4.062529	3.704403	3.456754
sl_contig17444	1.617225	3.593137	4.898431	4.610191	3.472802	3.970982	3.664725	3.600088



- ✓ **Algorithms:**
- a) PCC
- b) Weighted PCC
- c) Multiple rank (MR)
- ✓ **Visualization software:**
  Cytoscape



- ✓ **GO enrichment analysis**
- ✓ **Function model analysis**

Gene regulatory network reconstruction

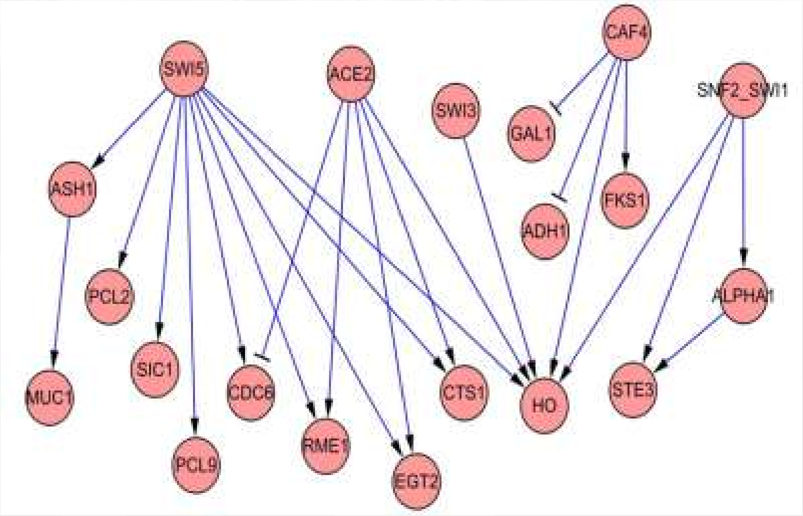
Gene name	1	2	3	4	5	6	7	8
sl_contig16919	4.585009	2.325221	1.987906	3.201388	3.644228	4.973095	4.756561	5.35751
sl_contig16968	1.314995	3.032279	1.279927	2.118202	3.838857	4.561094	3.127101	3.689177
sl_contig16981	5.053353	3.831191	4.043196	4.014023	3.828976	4.320826	5.079683	4.799046
sl_contig16987	4.456226	4.521689	4.483062	4.107209	2.756424	3.218653	3.958525	3.337341
sl_contig17023	3.366103	3.796538	3.262048	3.025738	2.963656	3.473839	3.028422	2.726439
sl_contig17072	3.723846	4.412139	3.443664	3.222046	4.148712	3.689451	4.271491	4.029439
sl_contig17101	5.907816	4.143168	2.181931	5.057381	1.870689	2.715251	3.468567	3.427814
sl_contig17173	4.319571	1.100264	3.316736	3.57334	2.137898	3.62096	2.712161	2.89311
sl_contig17176	2.059789	3.594238	2.8038	2.289057	4.54947	3.762934	4.989784	4.563962
sl_contig17200	4.459731	4.792051	5.279573	3.73811	2.211618	2.118202	1.859741	2.307091
sl_contig17273	4.517204	2.492271	3.220278	3.392975	3.790786	4.194001	3.405734	4.840509
sl_contig17285	3.983549	4.82406	4.378887	4.456414	3.308111	1.922581	1.981118	2.048111
sl_contig17371	3.317409	2.511857	3.858325	3.484647	2.873372	3.508207	2.02129	3.846771
sl_contig17385	3.825362	2.881894	1.844082	3.795703	4.290513	4.062529	3.704403	3.456754
sl_contig17444	1.617225	3.593137	4.898431	4.610191	3.472802	3.970982	3.664725	3.600088

Gene expression data Discretization

- ✓ Equal Width Discretization
- ✓ Equal Frequency Discretization
- ✓ Kmeans Discretization
- ✓ Column Kmeans Discretization
- ✓ Bikeans Discretization

Gene regulatory network reconstruction

- ✓ Greedy search
- ✓ K2
- ✓ Aracne
- ✓ Matlab
- ✓ … …



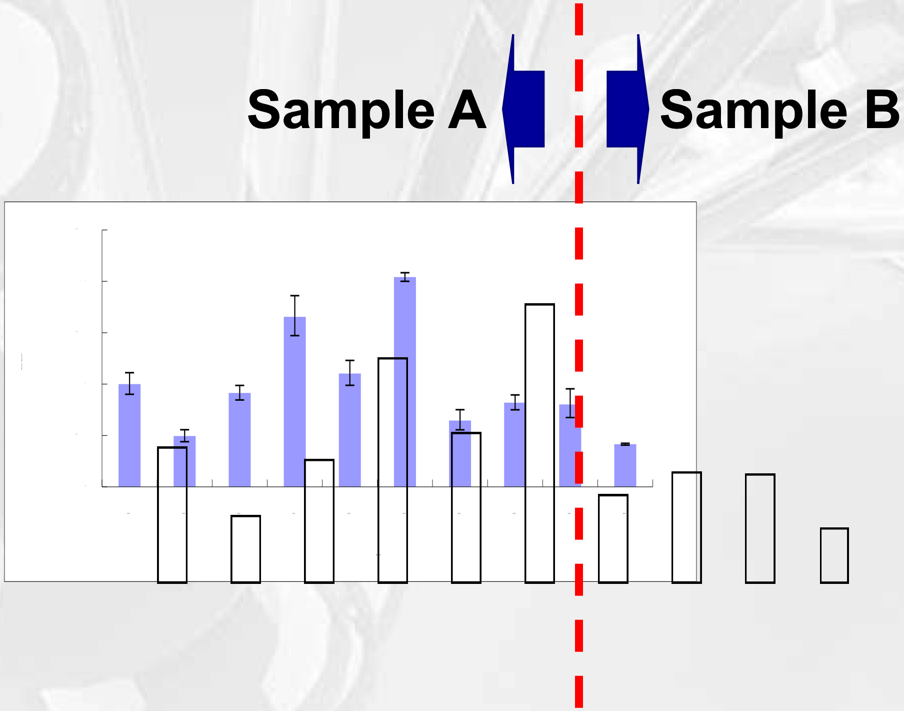
理论课内容

- 转录组学介绍
- 基因表达数据分析
  - 测定技术
  - 差异基因
  - 功能分析
- 几个实例
- 非编码RNA分析

Hickory gene expression data analysis

Materials and Methods

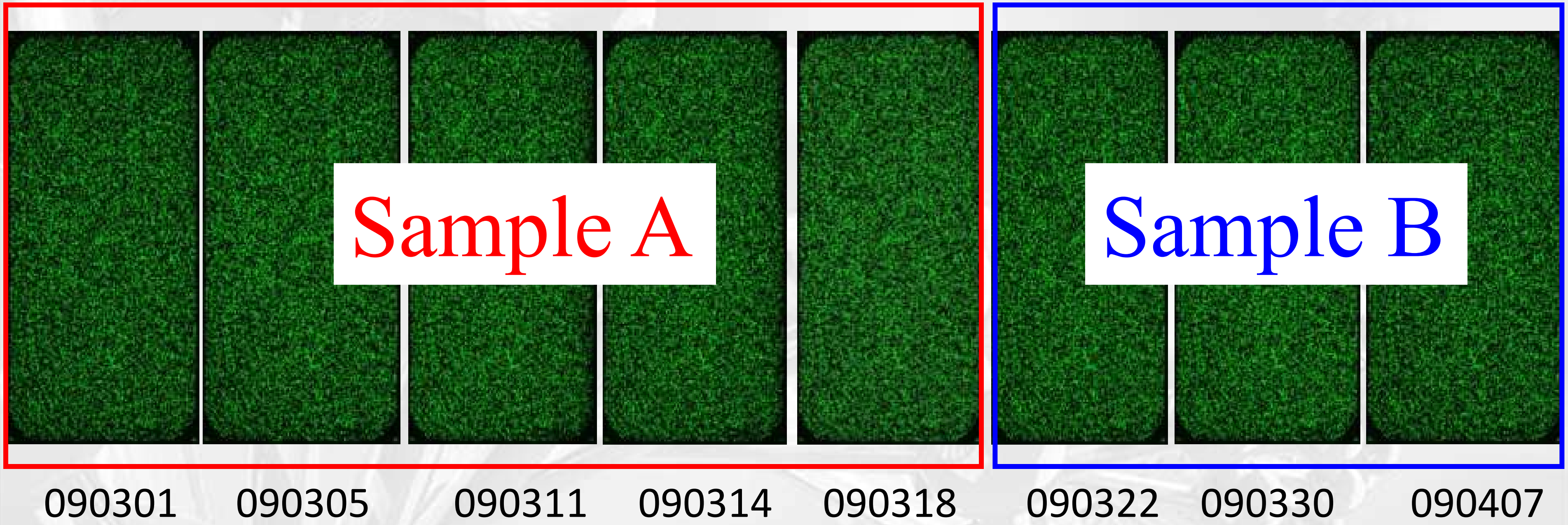
- 454 sequecing



454 Sequencing

	Sample A	Sample B
Read number	431,759	444,905
Avg. read length	332	332
contig	25339	26935
Specific gene	4951	5887
ORF number	15085	16387

Gene chips



Probe

- **454contigs: 25307 from Sample A, 7318 from Sample B**
- **Clone genes: 255**
- **Flowering Key genes: 109**
- **Positive signal hybridize with Ara: 324**

Methods

1) Flowering network construction of Arabidopsis based on literatures.

- **Key word**: flowering floral ect.
- **The total number of literatures**: About 1500.
- **Flowering genes**: 436 (Common name, Locus ID).
- **Flowering construction and visualization based on Cytoscape software.**

2) 454 sequencing analysis.

- **Contig assemble:** CAP3 software (Sample A, Sample B and All)
- **Blast analysis against Arabidopsis:** Blast software (Contigs->Ara. genes).

Result filter: Identity percent: 80%, E-value: 1e-5, Coverage: 70%.

Methods

3) Differentially expressed gene analysis.

Constraint conditions:

Fold change:4, Num(fc): 1. Signal value: except all A's

4) Gene expression pattern analysis.

Software: MeV software.

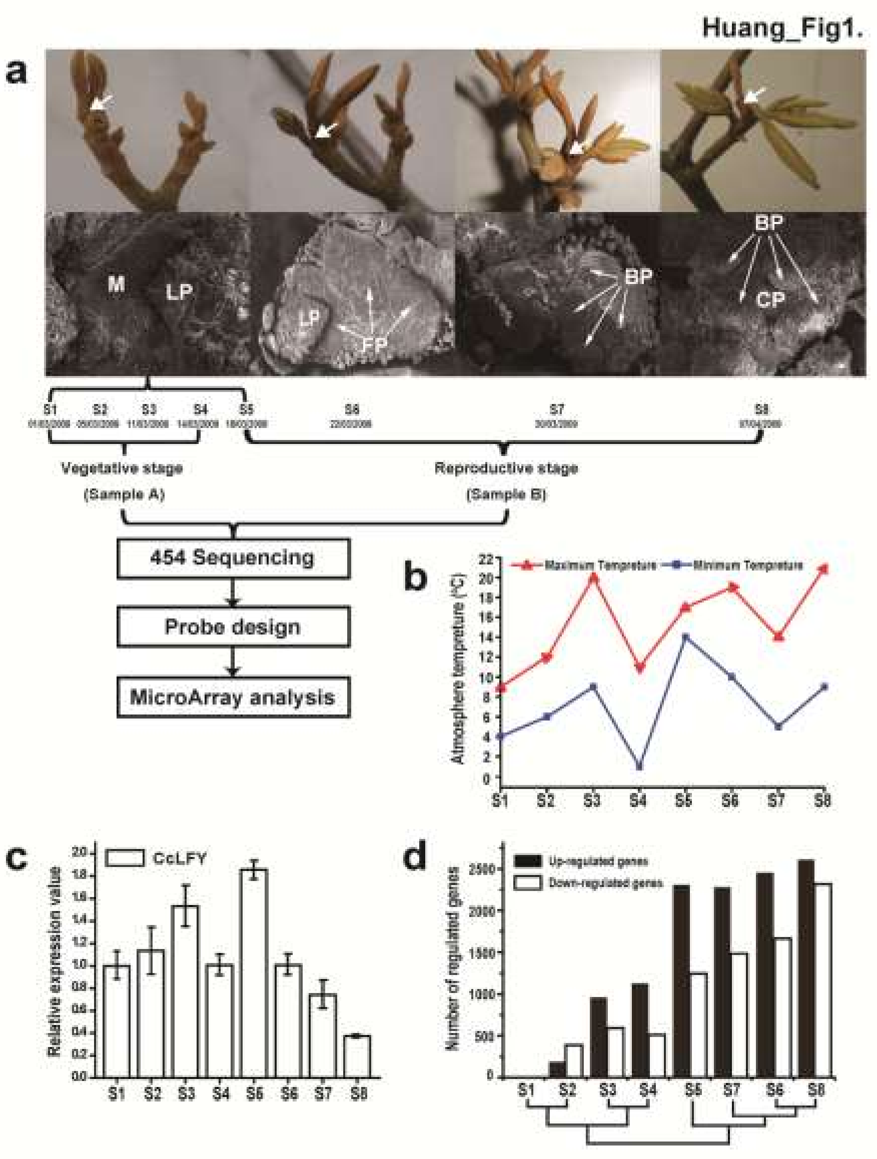
Algorithm: K-means.

5) GO Enrichment analysis

6) Co-expression network reconstruction for flowering genes.

Algorithm: Mutual Rank (MR) (2008, NAR)

7) Real time quantitative PCR



454 Sequencing

Probe design

MicroArray analysis

Fig.1 Experimental design.

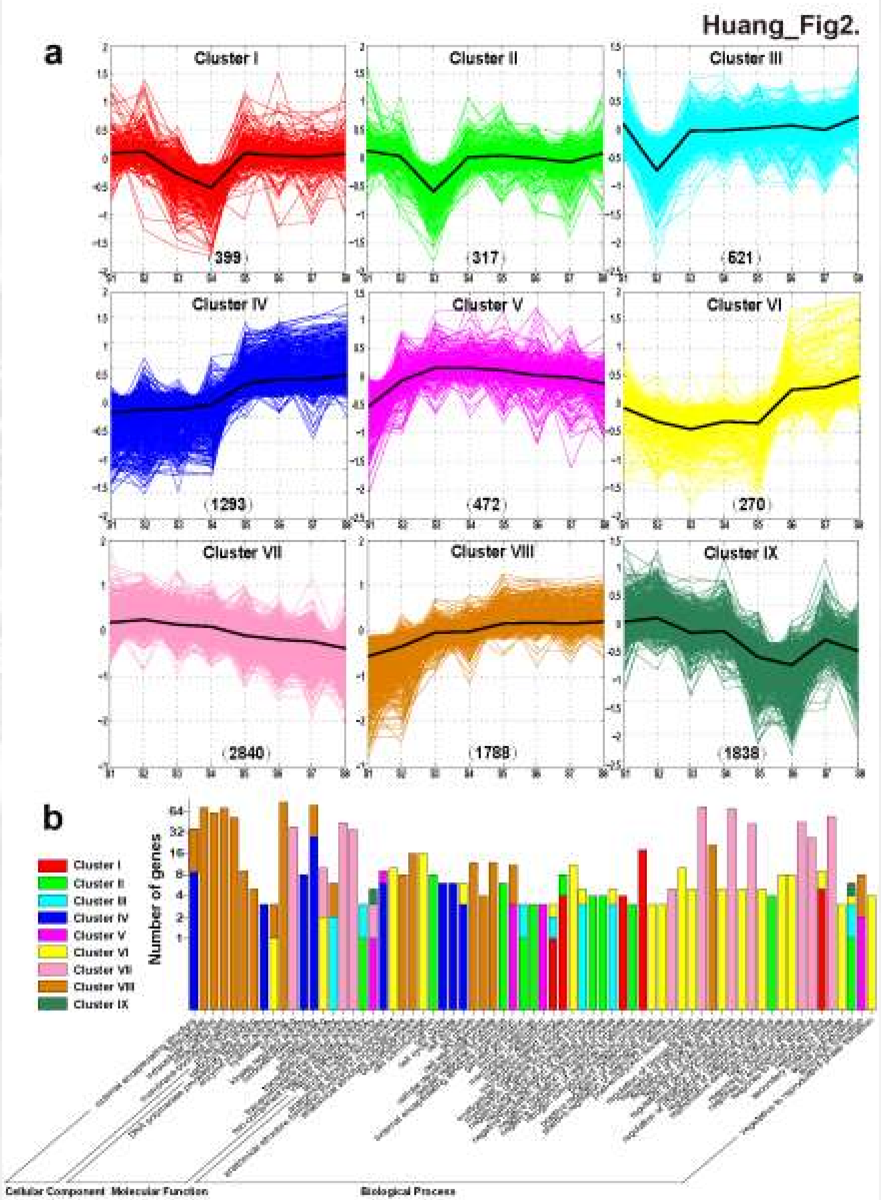
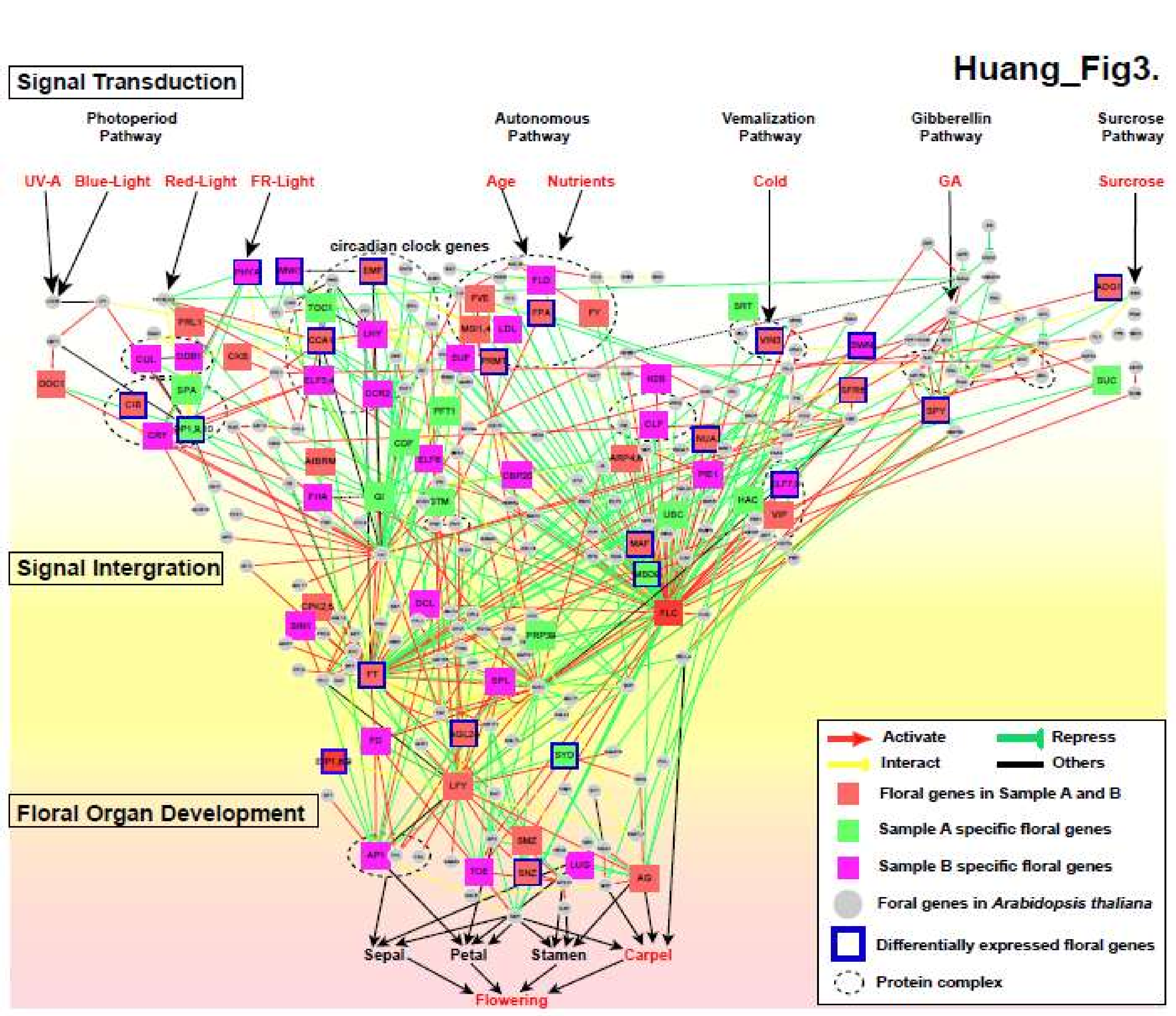


Fig. 2 Dynamic expression pattern of different clusters during flower development and GO function enrichment analysis.


Huang_Fig3.
Signal Transduction
Photoperiod Pathway
Autonomous Pathway
Vemalization Pathway
Gibberellin Pathway
Surcrose Pathway
UV-A
Blue-Light
Red-Light
FR-Light
Age
Nutrients
Cold
GA
Surcrose
circadian clock genes
PHYA
EMF
FLD
FVE
FPA
FY
TOC1
CCA1
LHY
ELF3/4
ELF3
CDF
GI
FT
FLC
LFY
AP1
SMZ
SNZ
TOE
AG
SPL
SPA
CRY
CUL
DDB1
CIR
PRR1
DOC1
CKB
SPY
SUC
VIN3
SRT
ARP4/6
MAF
HAC
VIP
UBC
PIE1
CLF
HUA
NHB
DCL
PRP39
FD
SYD
LUG
STM
DBR2
Signal Intergration
Floral Organ Development
Sepal
Petal
Stamen
Carpel
Flowering
Activate
Repress
Interact
Others
Floral genes in Sample A and B
Sample A specific floral genes
Sample B specific floral genes
Foral genes in *Arabidopsis thaliana*
Differentially expressed floral genes
Protein complex

Huang_Fig5.

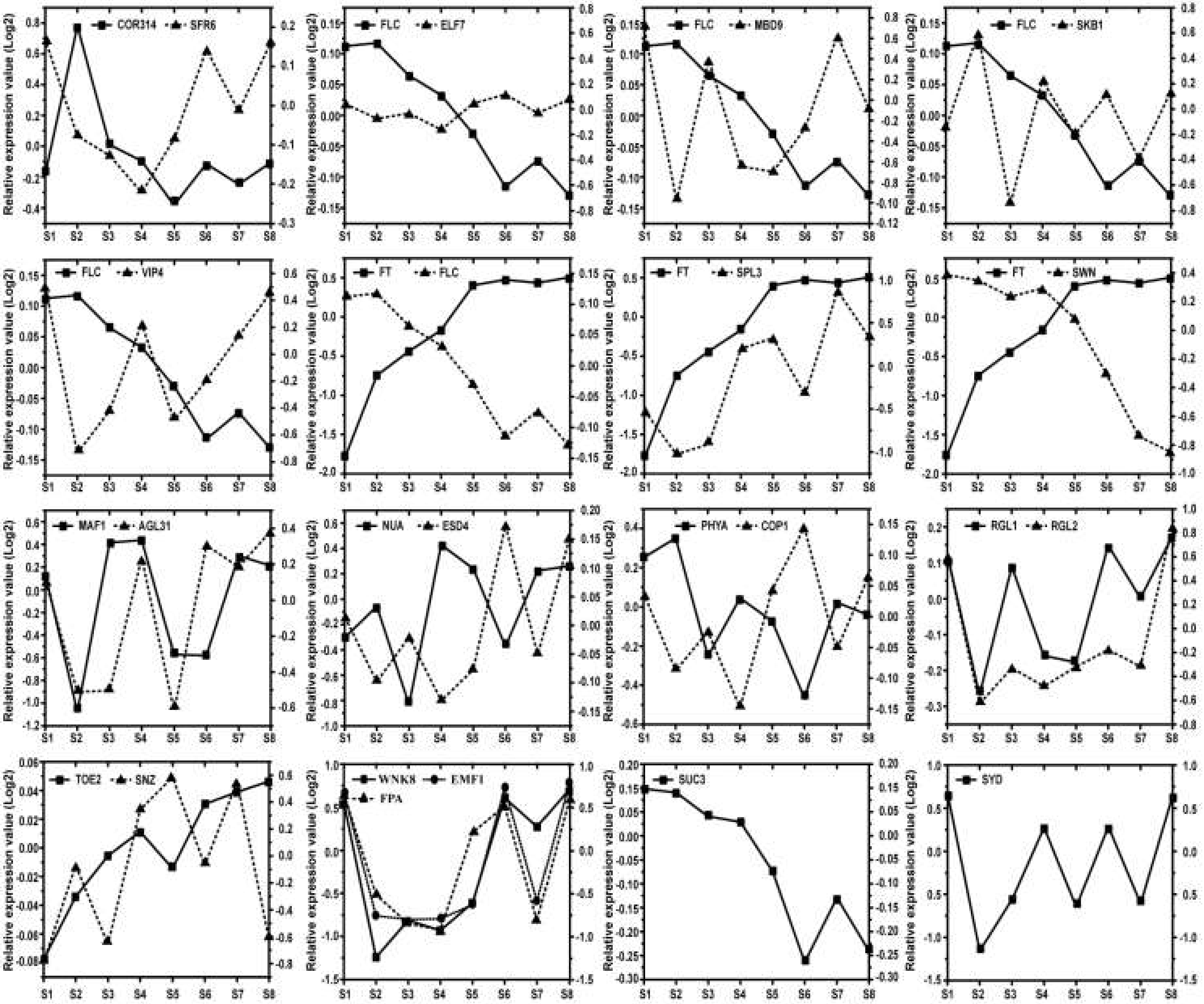
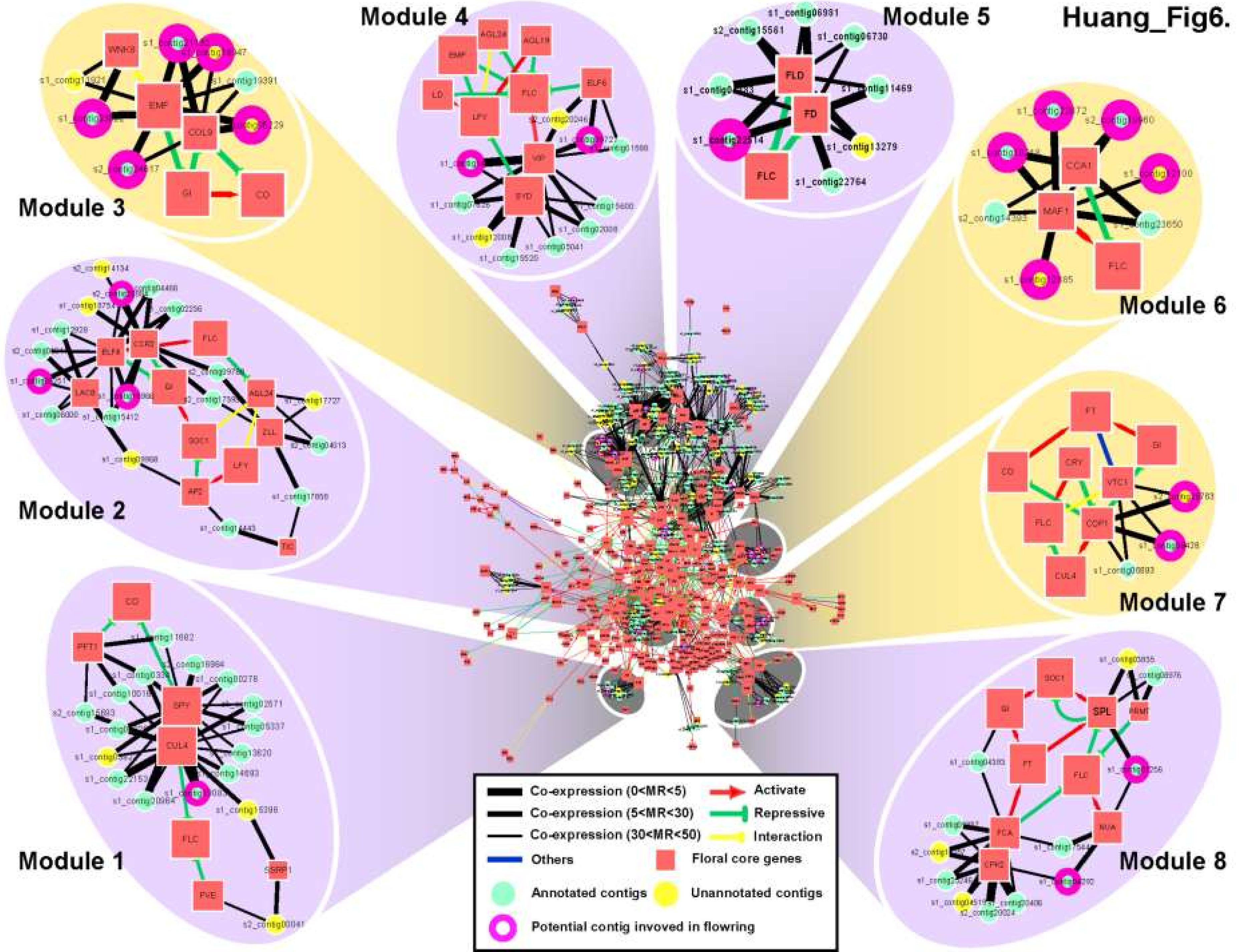


Fig. 5 Transcriptional regulation of differentially expressed genes in floral development in hickory.

Huang_Fig6.



Module 1
CO
PFT1
SPY
CUL4
FLC
PIE
SSRP1
s1_contig11682
s1_contig16964
s1_contig00278
s1_contig02671
s1_contig05337
s1_contig13620
s1_contig14693
s1_contig15396
s1_contig22152
s1_contig20964
s2_contig15693
s2_contig00041
Module 2
ELF4
CCR2
FLC
LACB
GI
AGL24
SOC1
ZLL
LFY
AP2
TFL
s2_contig14134
s1_contig04466
s1_contig02256
s1_contig12928
s1_contig17727
s1_contig15412
s1_contig06006
s1_contig09908
s2_contig04013
s1_contig17656
Module 3
WNK8
EMF
COL9
GI
CO
s1_contig19391
s1_contig11921
s2_contig04617
Module 4
AGL24
AGL19
EMF
LD
FLC
ELF6
LFY
VIP
SYD
s2_contig20246
s1_contig15600
s1_contig02008
s1_contig05041
s1_contig16525
Module 5
FLD
FD
FLC
s1_contig06991
s2_contig15561
s1_contig06730
s1_contig11469
s1_contig13279
s1_contig22764
Module 6
CCA1
MAF1
FLC
s2_contig14383
s1_contig23650
Module 7
FT
GI
CO
CRY
VTC1
FLC
COP1
CUL4
s1_contig00893
Module 8
SOC1
SPL
PRMT
GI
FT
FLC
FCA
VUA
CPK5
s1_contig00835
s1_contig08976
s1_contig04383
s1_contig17564
s2_contig20024
Co-expression (0<MR<5)
Co-expression (5<MR<30)
Co-expression (30<MR<50)
Others
Activate
Repressive
Interaction
Floral core genes
Annotated contigs
Unannotated contigs
Potential contig invoved in flowring

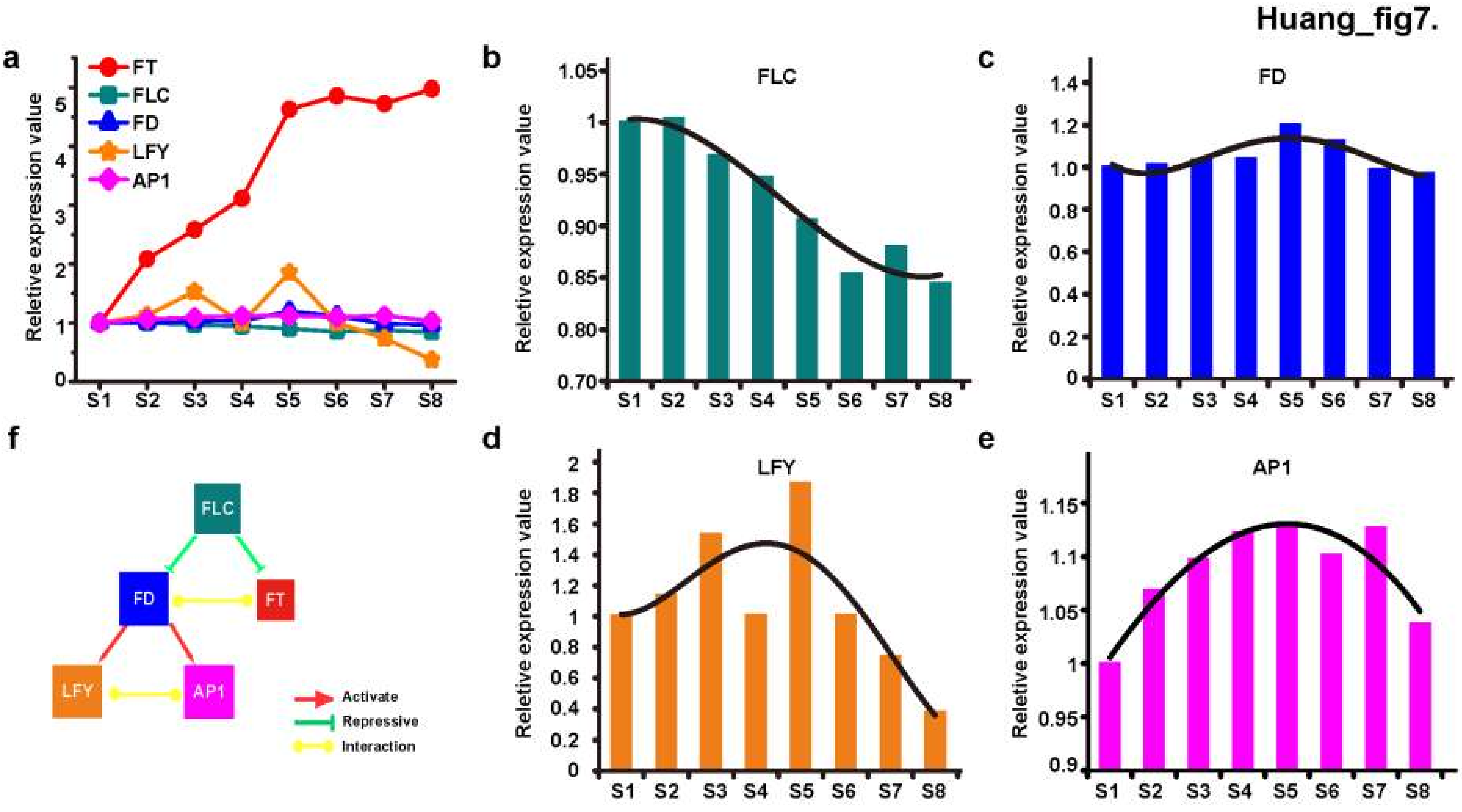
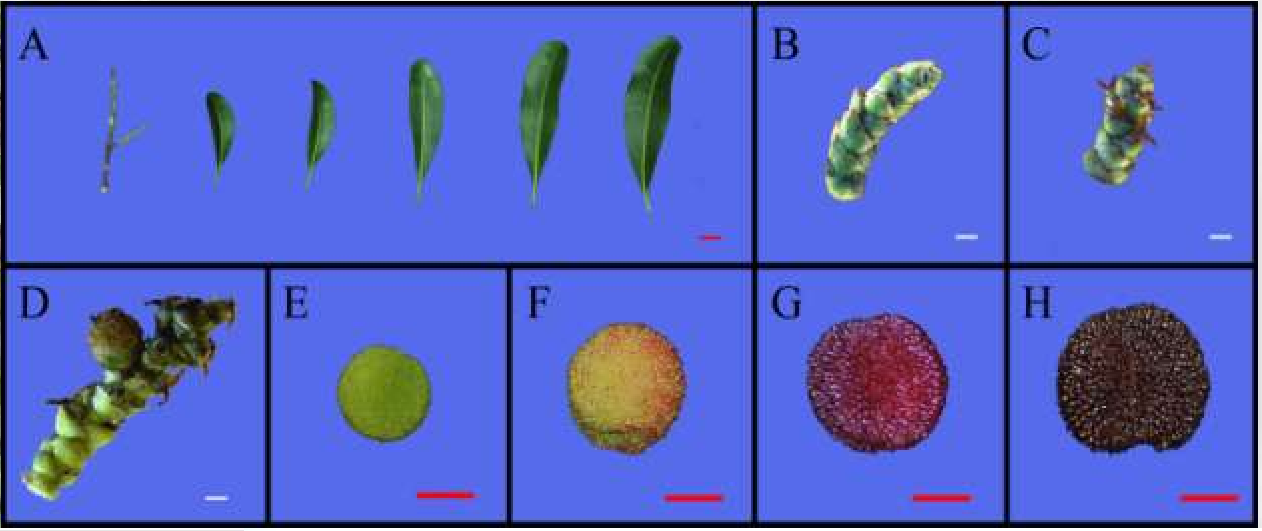
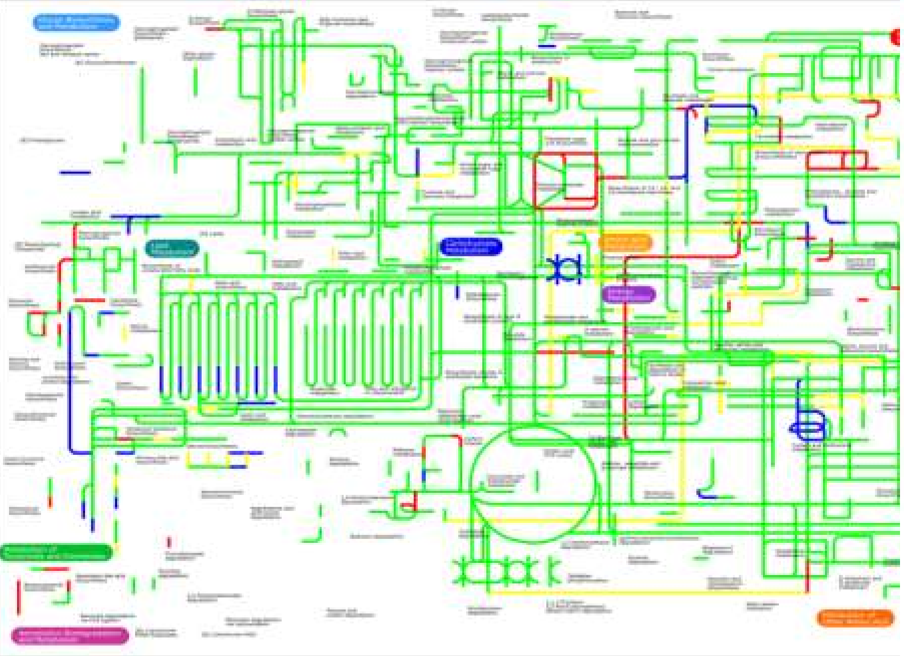


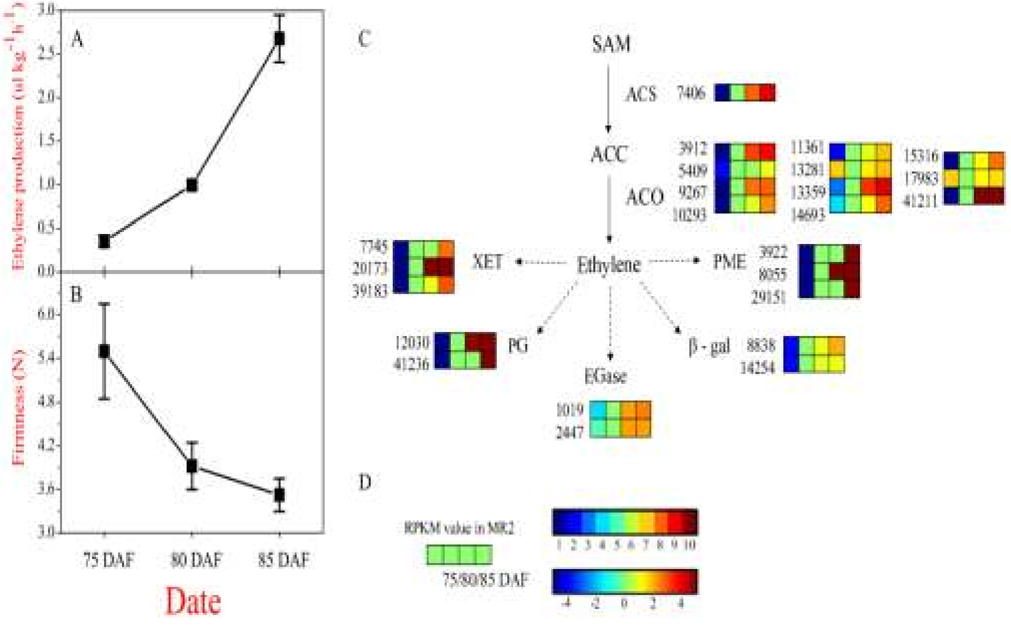
Fig. 7 Expression and regulation relationship of floral integrators in hickory.

Chinese bayberry

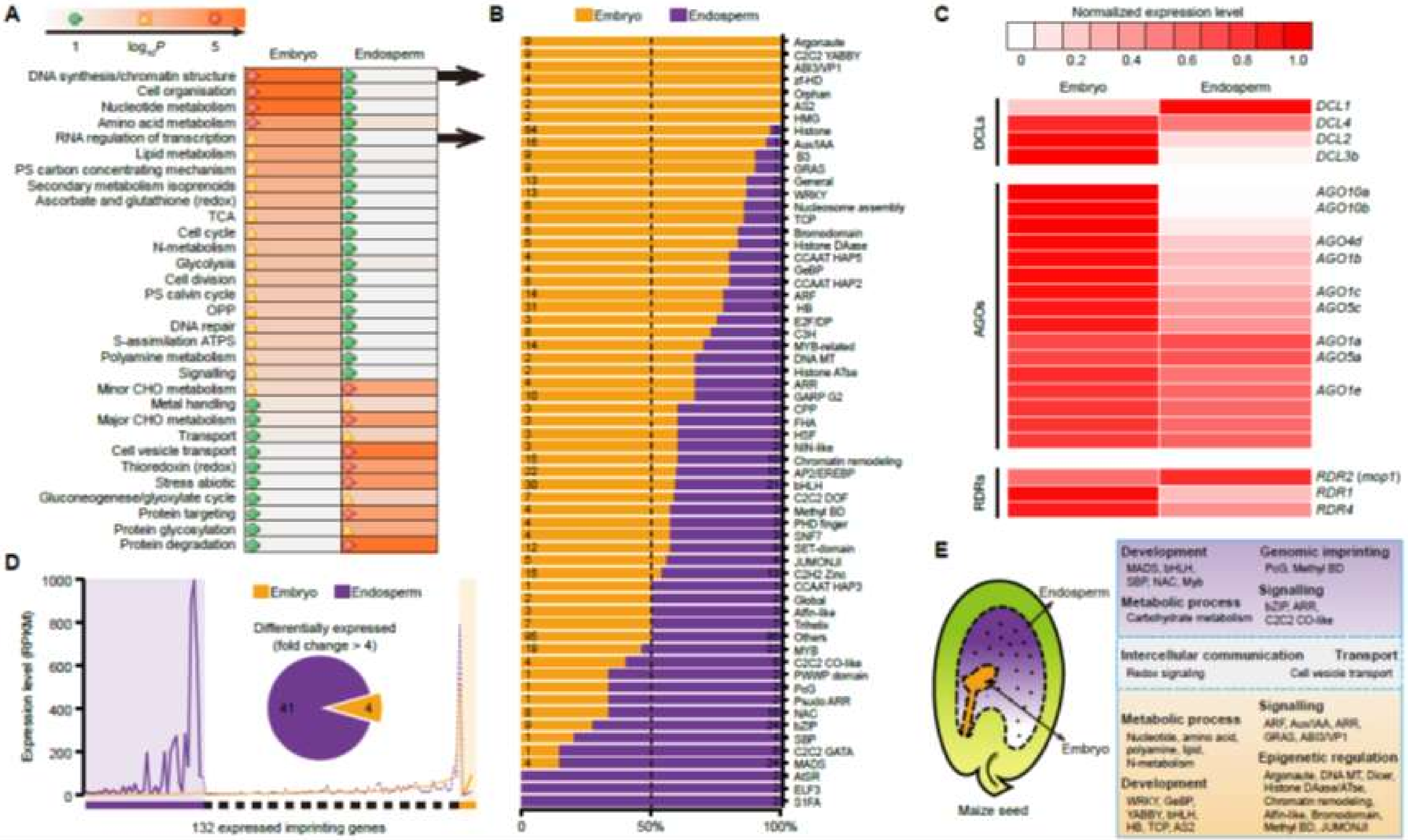
Highly accessed







Dynamic progression map of seed transcriptome



理论课内容

- 转录组学介绍
- 基因表达数据分析
  - 测定技术
  - 差异基因
  - 功能分析
- 几个实例
- 非编码RNA分析

Small RNA transcriptome analysis

Fine-scale methods:

QRT-PCR, *in situ* hybridization/RT-PCR, Northern blot…

Low-throughput, tedious, not sensitive enough (Northern)…

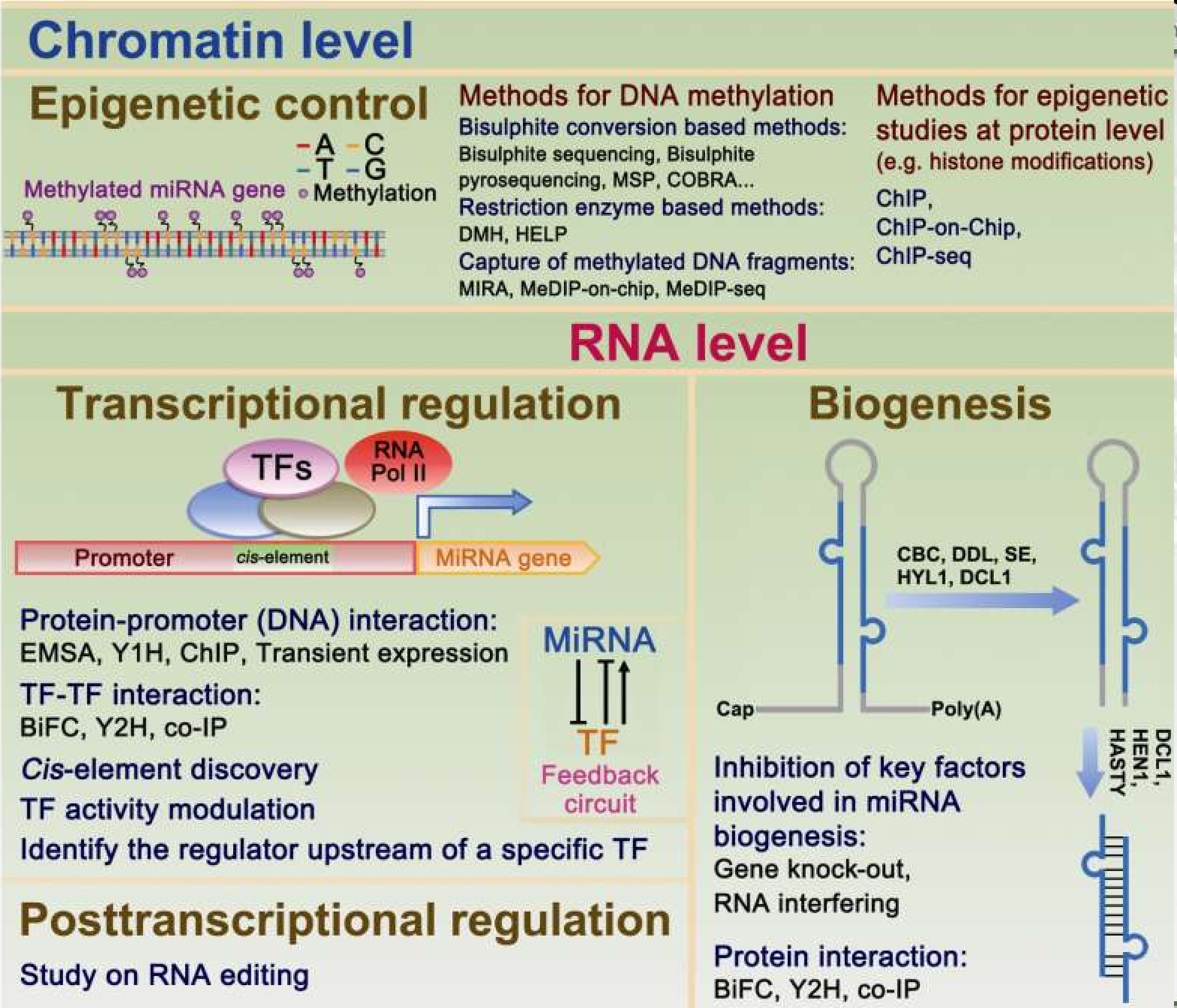
High-throughput methods:

microarray, next-generation sequencing (NGS)

High-throughput, expensive
cross-hybridization & limited sensitivity (microarray)

Given the in-depth (sensitive) and quantitative feature, many plant transcriptome analyses were promoted by NGS.

Upstream analysis



Chromatin level

Epigenetic control

Methods for DNA methylation

Bisulphite conversion based methods:

Bisulphite sequencing, Bisulphite pyrosequencing, MSP, COBRA...

Restriction enzyme based methods:

DMH, HELP

Capture of methylated DNA fragments:

MIRA, MeDIP-on-chip, MeDIP-seq

Methods for epigenetic studies at protein level (e.g. histone modifications)

ChIP,
ChIP-on-Chip,
ChIP-seq

RNA level

Transcriptional regulation

Protein-promoter (DNA) interaction:

EMSA, Y1H, ChIP, Transient expression

TF-TF interaction:

BiFC, Y2H, co-IP

Cis-element discovery

TF activity modulation

Identify the regulator upstream of a specific TF

Posttranscriptional regulation

Study on RNA editing

Biogenesis

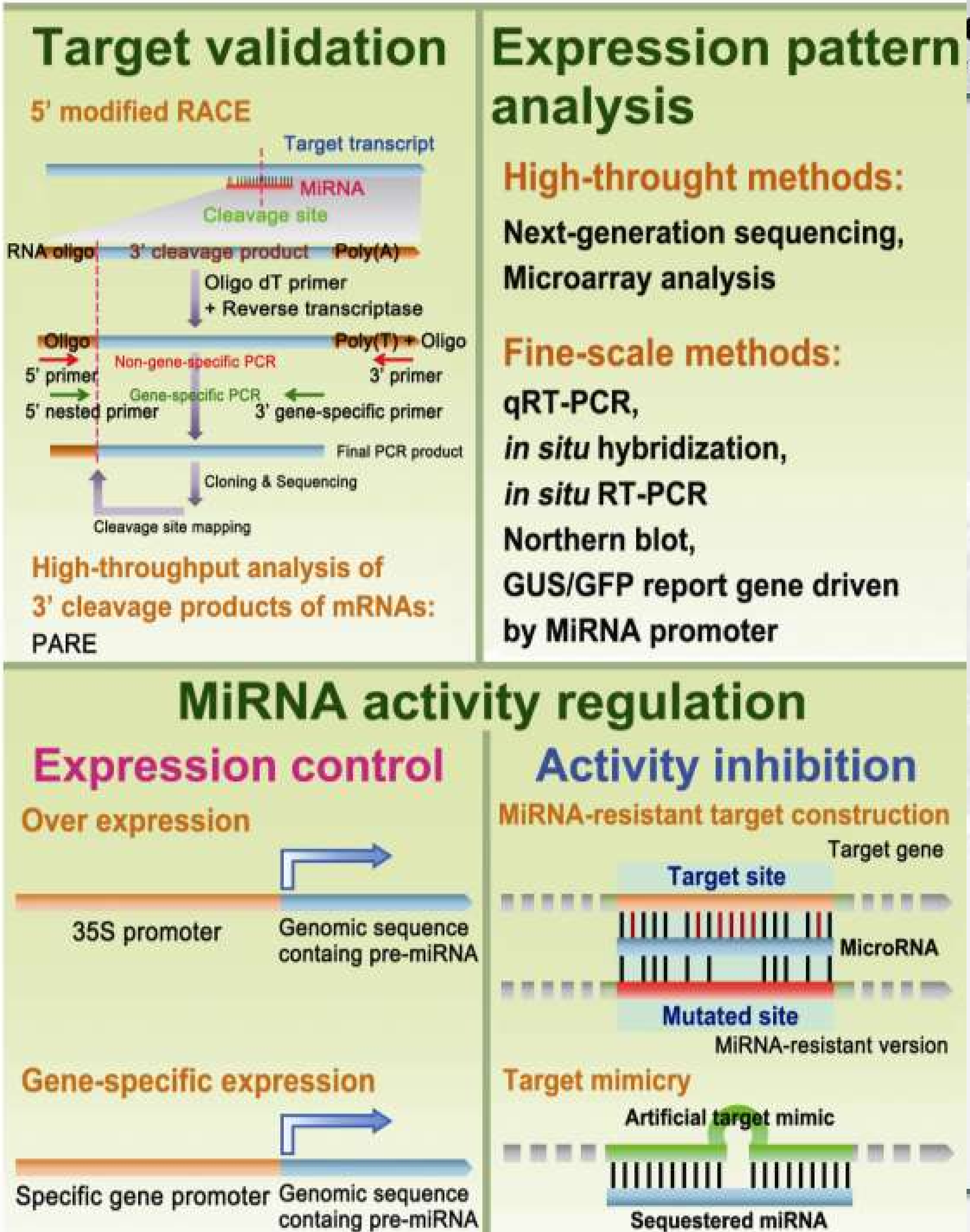
Inhibition of key factors involved in miRNA biogenesis:

Gene knock-out,
RNA interfering

Protein interaction:

BiFC, Y2H, co-IP

J Exp Bot, 2010



Target validation

5' modified RACE

High-throughput analysis of 3' cleavage products of mRNAs:

PARE

Expression pattern analysis

High-throught methods:

Next-generation sequencing,
Microarray analysis

Fine-scale methods:

qRT-PCR,
in situ hybridization,
in situ RT-PCR
Northern blot,
GUS/GFP report gene driven by MiRNA promoter

MiRNA activity regulation

Expression control

Over expression

Gene-specific expression

Activity inhibition

MiRNA-resistant target construction

Target mimicry

Topics

- Plant ncRNAs
- biogenesis,
- characteristics,
- expressions,
- interactions,
- regulations,
- even dynamic functions, 3D…

Small RNAs in angiosperms: sequence characteristics, distribution and generation

Eudicots (16)

Arabidopsis拟南芥
Tomato西红柿
Medicago苜蓿
Pepper胡椒
Pumpkin南瓜
Sweet orange甜橙
Tree cotton木棉
Cultivated lettuce莴苣
Common monkey-flower猴面花
Tobacco烟草
Petunia矮牵牛花
Poplar白杨
White campion白花蝇子草
Potato土豆
Grapevine葡萄
Papaya木瓜

Monocots (10)

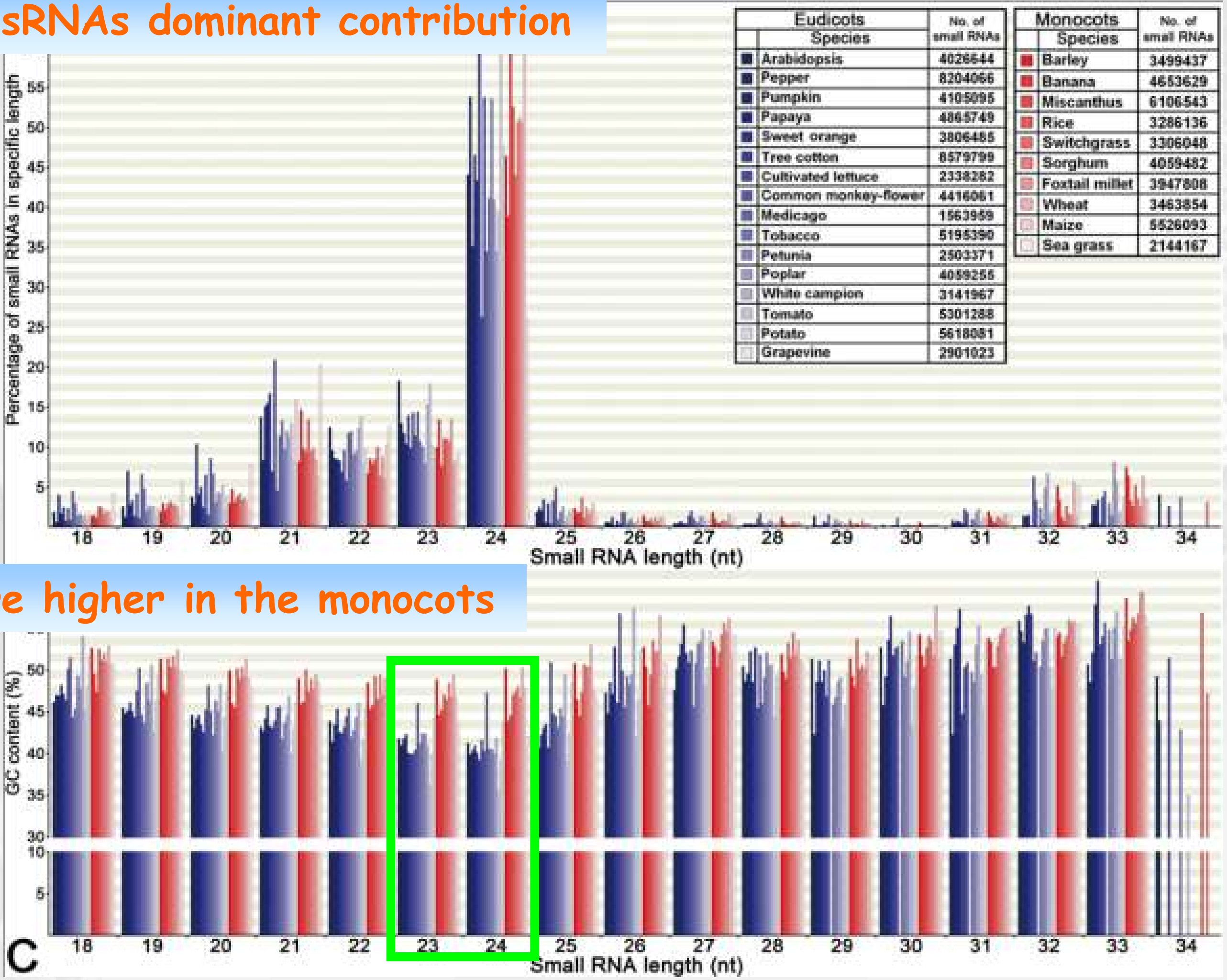
水稻Rice
玉米Maize
大麦Barley
香蕉Banana
柳枝稷Switchgrass
高粱Sorghum
小麦Wheat
海草Sea grass
芒草Miscanthus
谷子Foxtail millet

Bioinformatics, 2010

Small RNA序列特征分析（1）

The 21-24-nt sRNAs dominant contribution

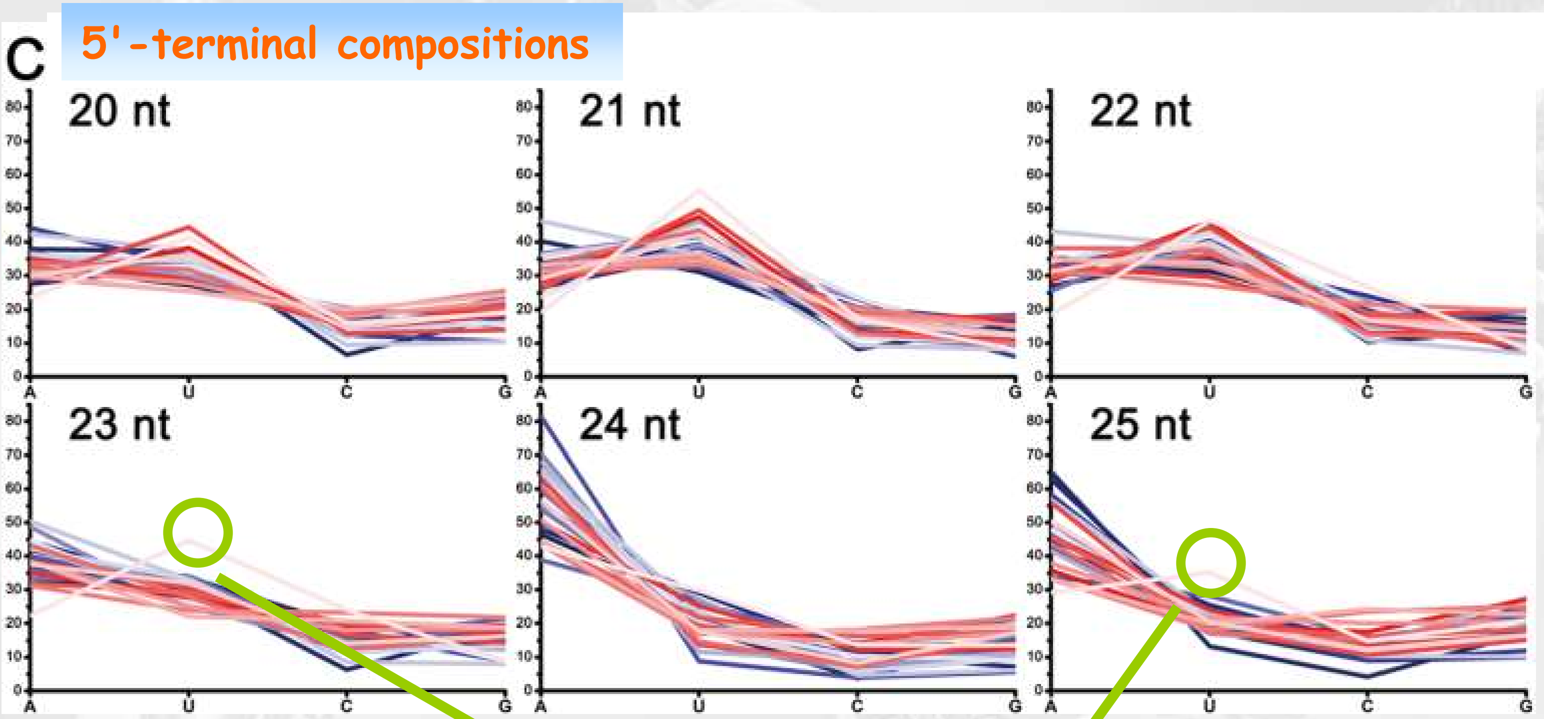
GC contents are higher in the monocots



	Eudicots Species	No. of small RNAs
■	Arabidopsis	4026644
■	Pepper	8204066
■	Pumpkin	4105095
■	Papaya	4865749
■	Sweet orange	3806485
■	Tree cotton	8579799
■	Cultivated lettuce	2338282
■	Common monkey-flower	4416061
■	Medicago	1563959
■	Tobacco	5195390
■	Petunia	2503371
■	Poplar	4059255
■	White campion	3141967
■	Tomato	5301288
□	Potato	5618081
□	Grapevine	2901023

	Monocots Species	No. of small RNAs
■	Barley	3499437
■	Banana	4653629
■	Miscanthus	6106543
■	Rice	3286136
■	Switchgrass	3306048
■	Sorghum	4059482
■	Foxtail millet	3947808
■	Wheat	3463854
□	Maize	5526093
□	Sea grass	2144167

Small RNA序列特征分析（2）

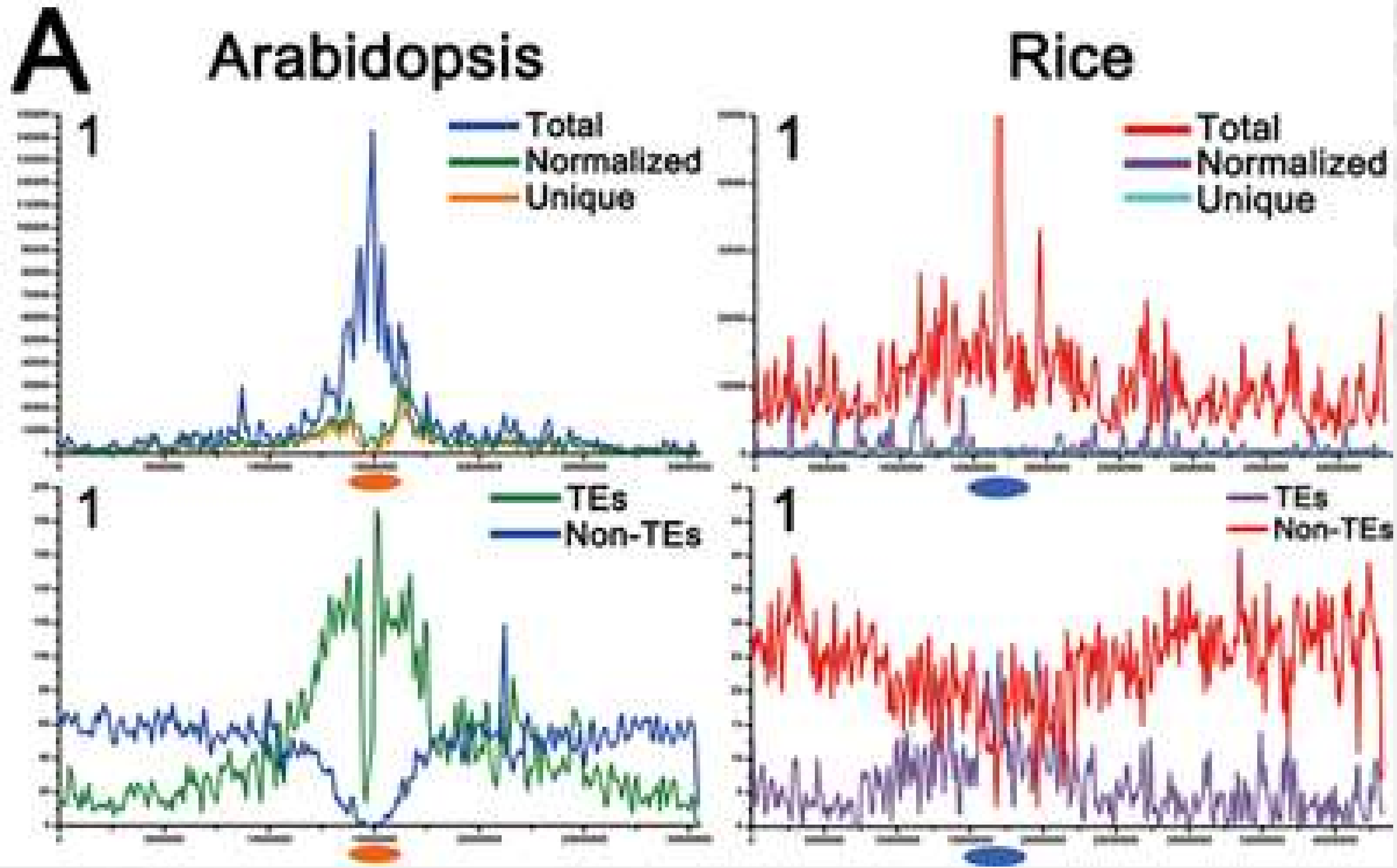


The 5’-terminal composition patterns are similar between the eudicots and the monocots.

单子叶植物中的海草（sea grass），其sRNAs 5’端碱基组成比较特别，是否与其水生环境有关？

Small RNAs及基因在染色体上分布模式比较：sRNAs大量重复分布于着丝粒及附近区域，pattern和转座子十分相似；而转座子序列本身又包含了大量的重复序列；为控制其转座活性、维持染色体结构序列上的稳定性，大量内源siRNAs用于控制转座子转座。

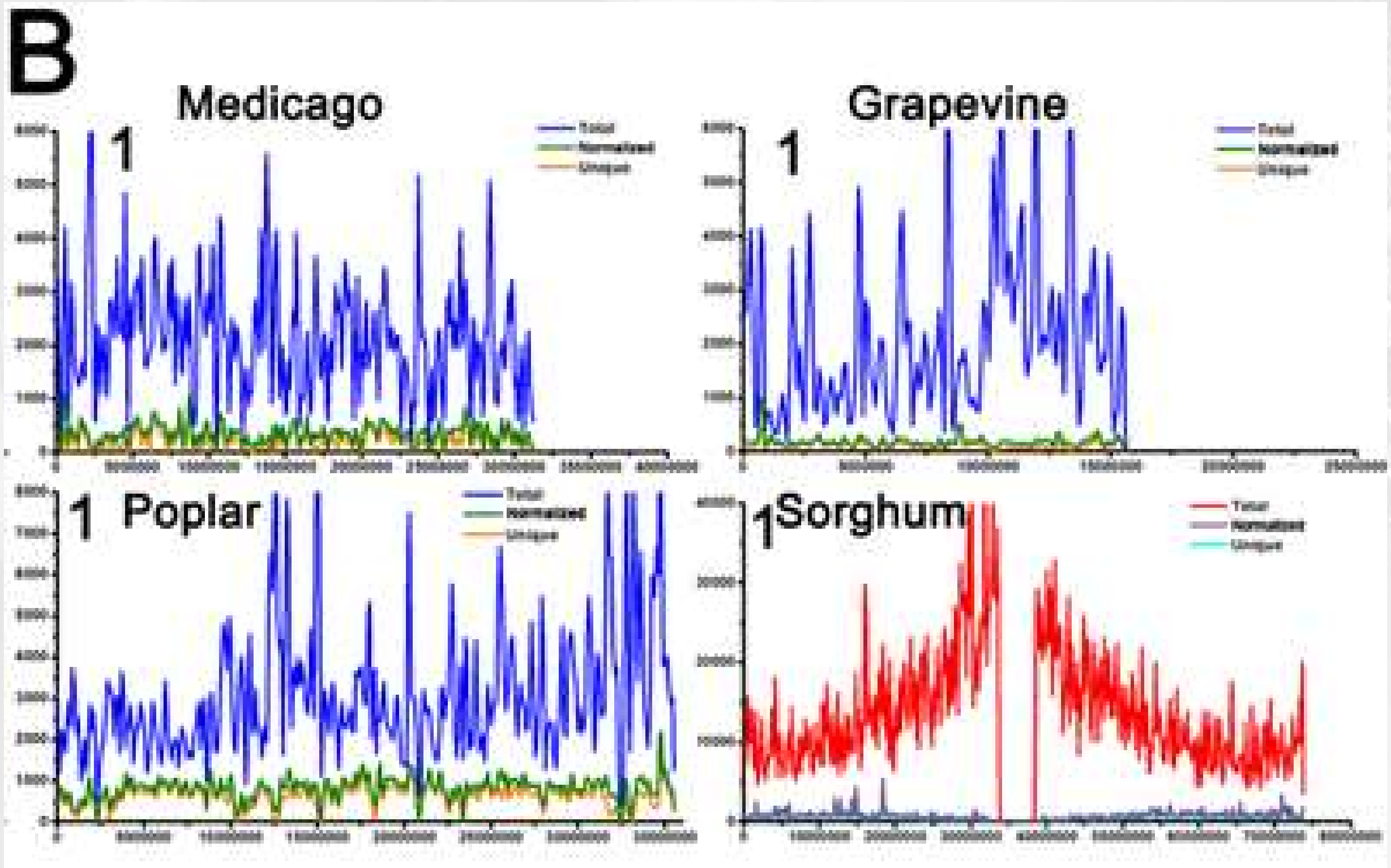
Chromosome-wide distribution patterns



The "total" locus distribution was similar to that of the Transposed Elements (TEs), but complementary to the non-TEs'.

Potential role in TE transposition control?

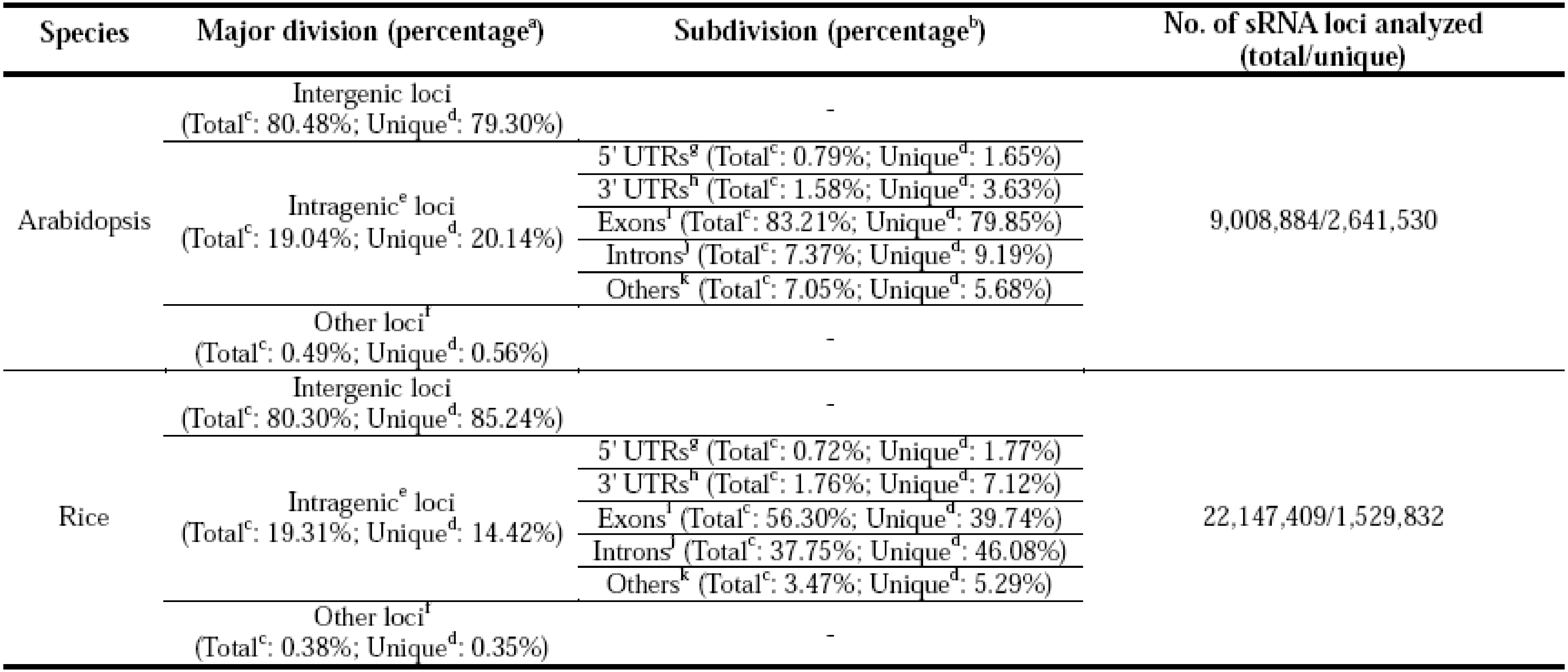
Extensive sRNA enrichment was detected on all the sorghum chromosomes.



高粱（sorghum）的sRNAs染色体分布在全部10条染色体上十分一致，起峰位置十分明显。根据拟南芥、水稻的分析经验，可以作为未拼接完成的高粱染色体组着丝粒位置的大致界定参考依据。

Small RNAs derived from gene models

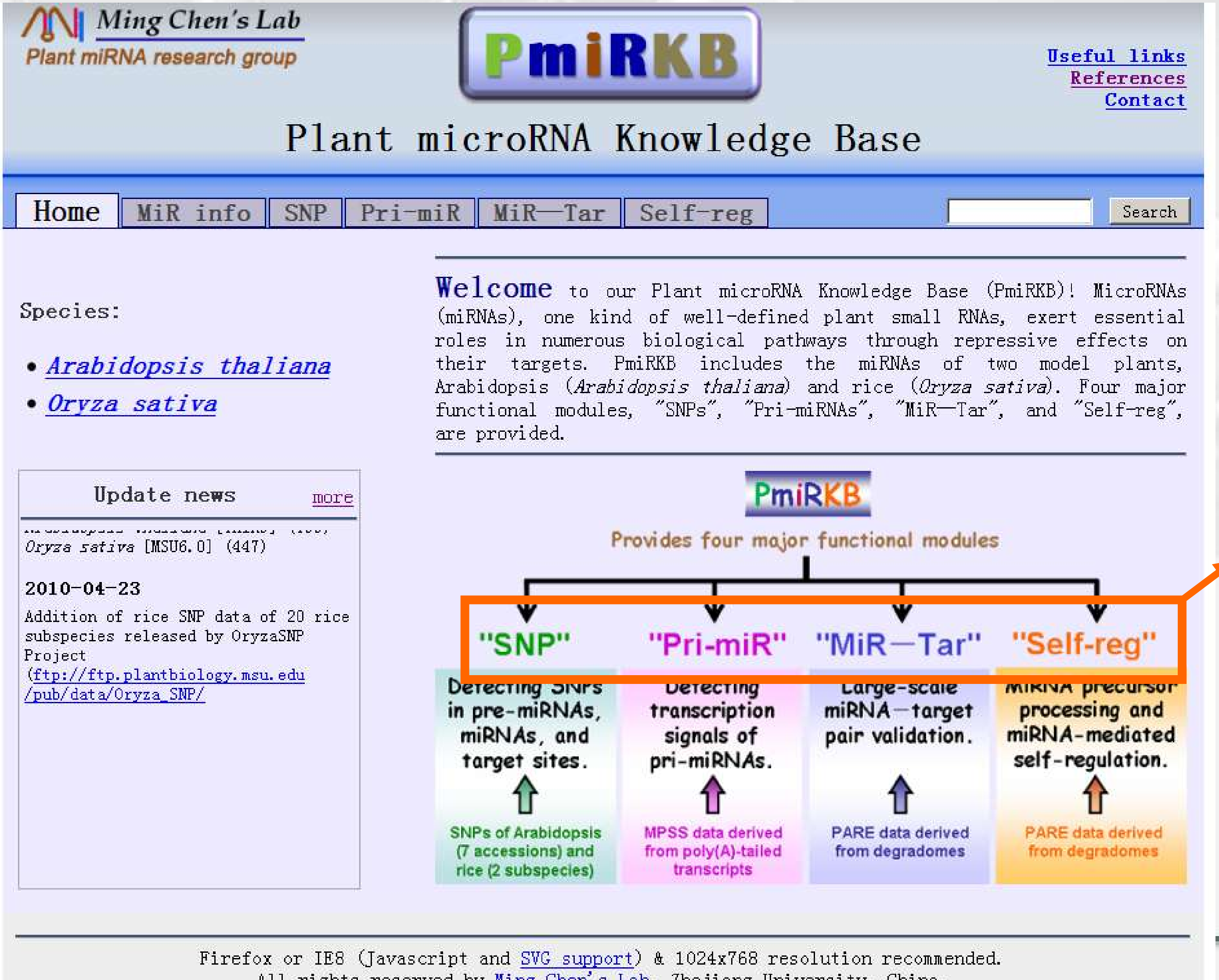
Bioinformatics, 2010



Species	Major division (percentage[a])	Subdivision (percentage[b])	No. of sRNA loci analyzed (total/unique)
Arabidopsis	Intergenic loci (Total[c]: 80.48%; Unique[d]: 79.30%)	-	9,008,884/2,641,530
	Intragenic[e] loci (Total[c]: 19.04%; Unique[d]: 20.14%)	5' UTRs[g] (Total[c]: 0.79%; Unique[d]: 1.65%)	
		3' UTRs[h] (Total[c]: 1.58%; Unique[d]: 3.63%)	
		Exons[i] (Total[c]: 83.21%; Unique[d]: 79.85%)	
		Introns[j] (Total[c]: 7.37%; Unique[d]: 9.19%)	
		Others[k] (Total[c]: 7.05%; Unique[d]: 5.68%)	
	Other loci[f] (Total[c]: 0.49%; Unique[d]: 0.56%)	-	
Rice	Intergenic loci (Total[c]: 80.30%; Unique[d]: 85.24%)	-	22,147,409/1,529,832
	Intragenic[e] loci (Total[c]: 19.31%; Unique[d]: 14.42%)	5' UTRs[g] (Total[c]: 0.72%; Unique[d]: 1.77%)	
		3' UTRs[h] (Total[c]: 1.76%; Unique[d]: 7.12%)	
		Exons[i] (Total[c]: 56.30%; Unique[d]: 39.74%)	
		Introns[j] (Total[c]: 37.75%; Unique[d]: 46.08%)	
		Others[k] (Total[c]: 3.47%; Unique[d]: 5.29%)	
	Other loci[f] (Total[c]: 0.38%; Unique[d]: 0.35%)	-	

Plant microRNA knowledge base

Nucleic Acids Res, 2011



Ming Chen's Lab
Plant miRNA research group

PmiRKB

Useful links
References
Contact

Plant microRNA Knowledge Base

Species:

- *Arabidopsis thaliana*
- *Oryza sativa*

Update news more

Oryza sativa [MSU6.0] (447)

2010-04-23

Addition of rice SNP data of 20 rice subspecies released by OryzaSNP Project (ftp://ftp.plantbiology.msu.edu/pub/data/Oryza_SNP/

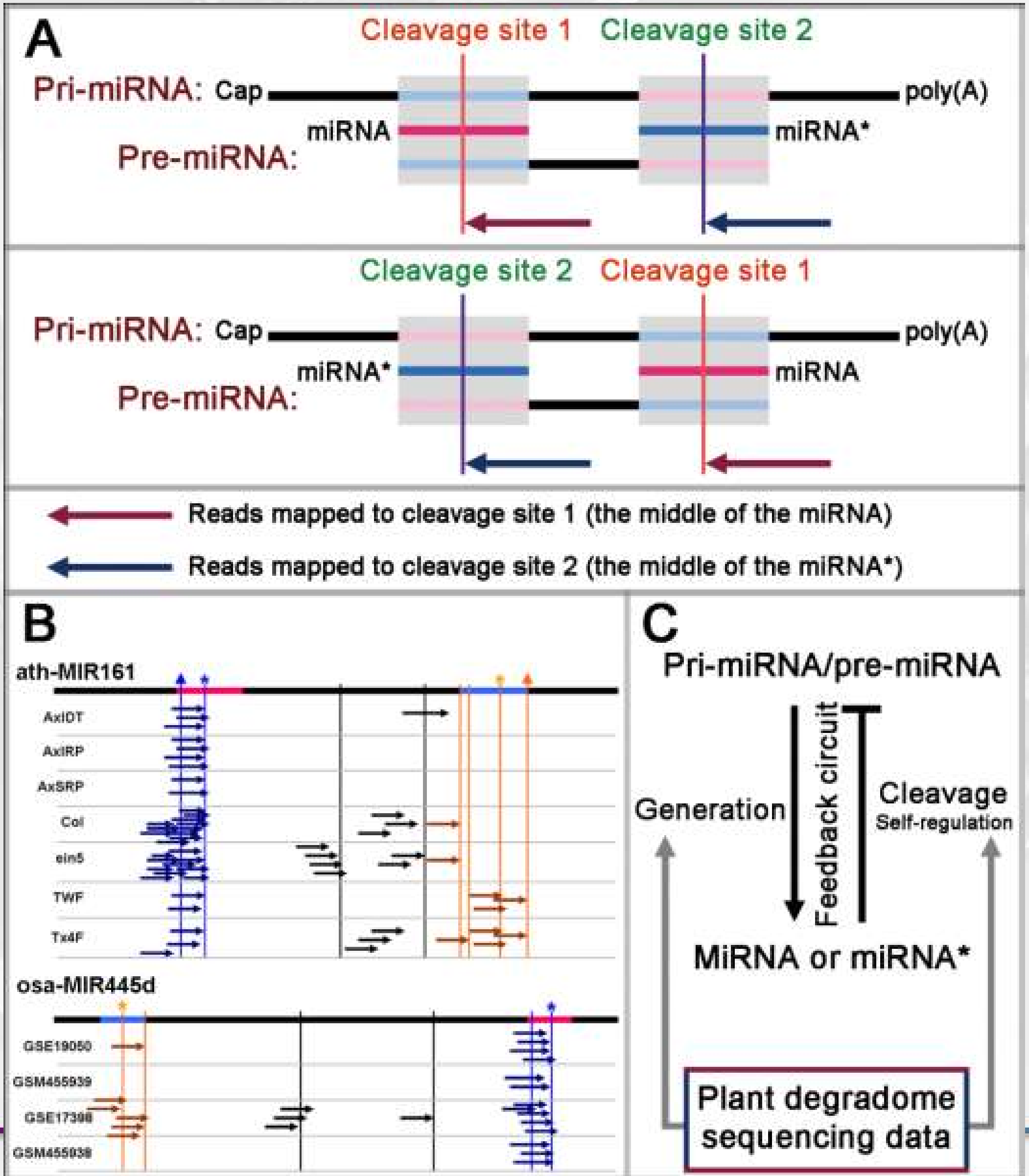
Welcome to our Plant microRNA Knowledge Base (PmiRKB)! MicroRNAs (miRNAs), one kind of well-defined plant small RNAs, exert essential roles in numerous biological pathways through repressive effects on their targets. PmiRKB includes the miRNAs of two model plants, Arabidopsis (*Arabidopsis thaliana*) and rice (*Oryza sativa*). Four major functional modules, "SNPs", "Pri-miRNAs", "MiR—Tar", and "Self-reg", are provided.

PmiRKB
Provides four major functional modules
"SNP"
"Pri-miR"
"MiR—Tar"
"Self-reg"
Detecting SNPs in pre-miRNAs, miRNAs, and target sites.
Detecting transcription signals of pri-miRNAs.
Large-scale miRNA—target pair validation.
miRNA precursor processing and miRNA-mediated self-regulation.
SNPs of Arabidopsis (7 accessions) and rice (2 subspecies)
MPSS data derived from poly(A)-tailed transcripts
PARE data derived from degradomes
PARE data derived from degradomes

Firefox or IE8 (Javascript and SVG support) & 1024x768 resolution recommended.
All rights reserved by Ming Chen's Lab, Zhejiang University, China.

包含4个主要功能模块

miRNA/miRNA*



J Exp Bot, 2010

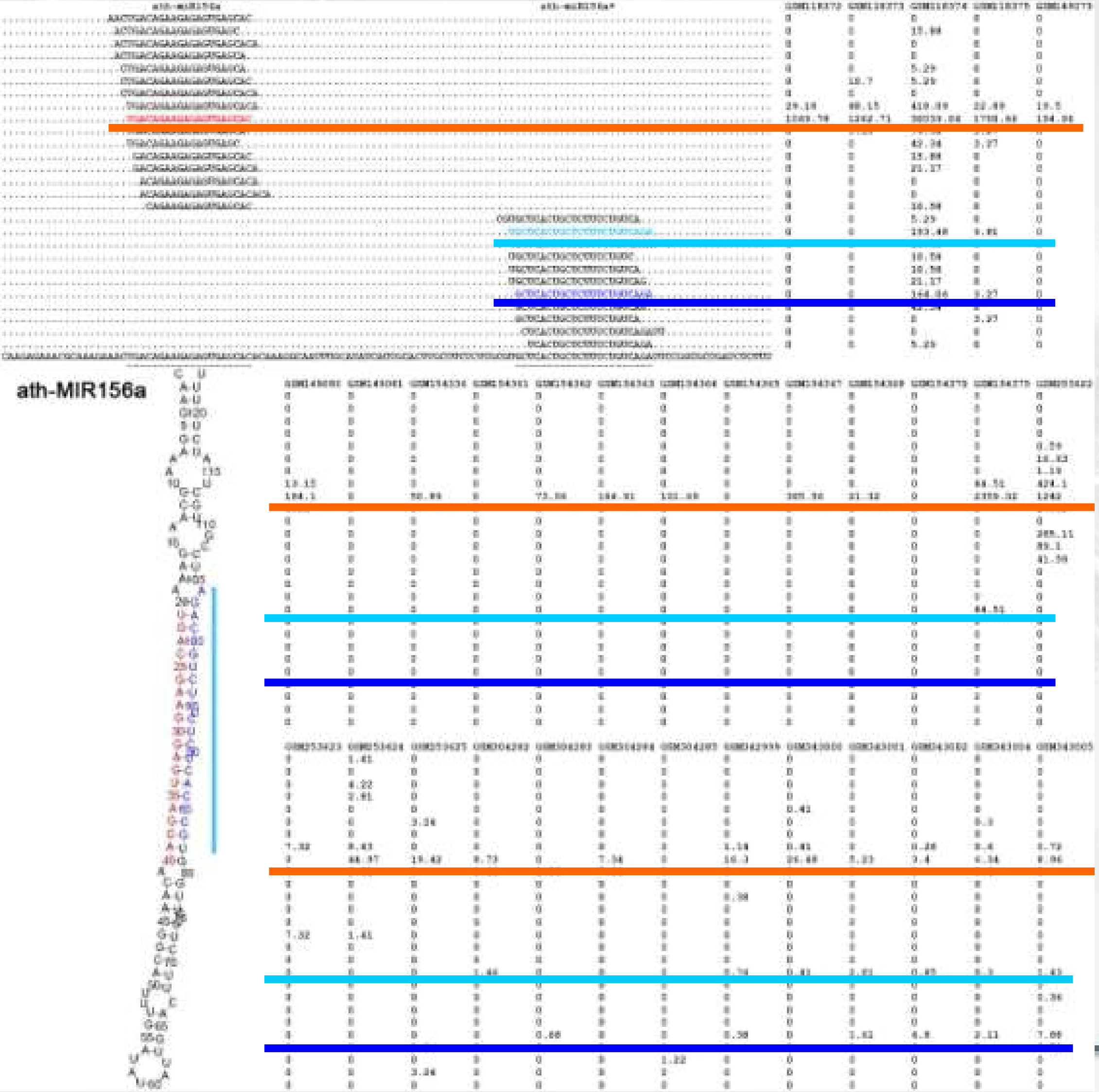
基于degradome测序数据分析可阐释：

•从miRNA前体到miRNA成熟体的加工过程

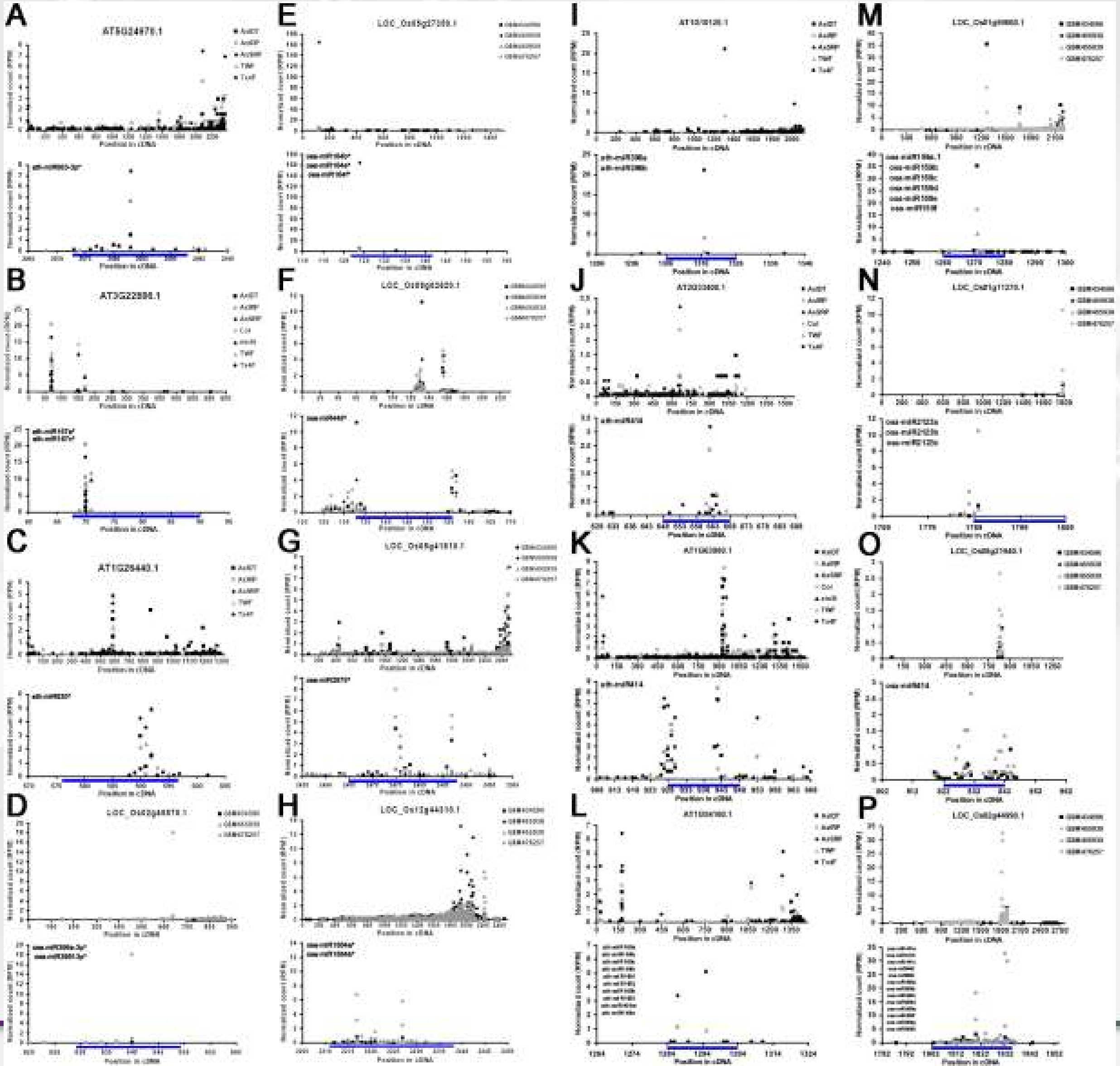
•miRNAs/miRNA*s介导的自调控过程。

miRNA/miRNA* detection

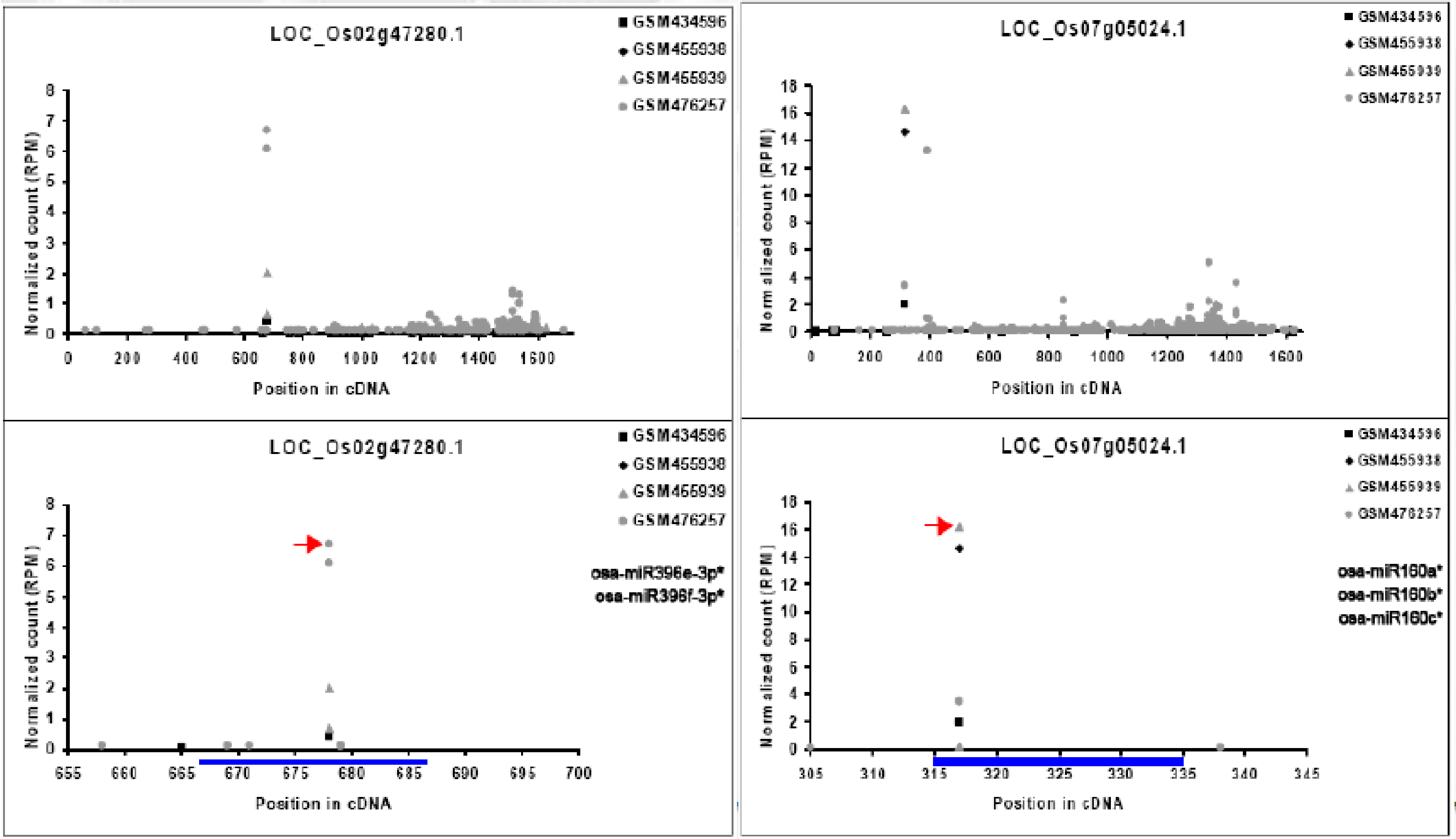
RNA Biology, 2011


ath-miR156a
ath-miR156a*
ath-MIR156a

miRNAs/miRNA*s targets



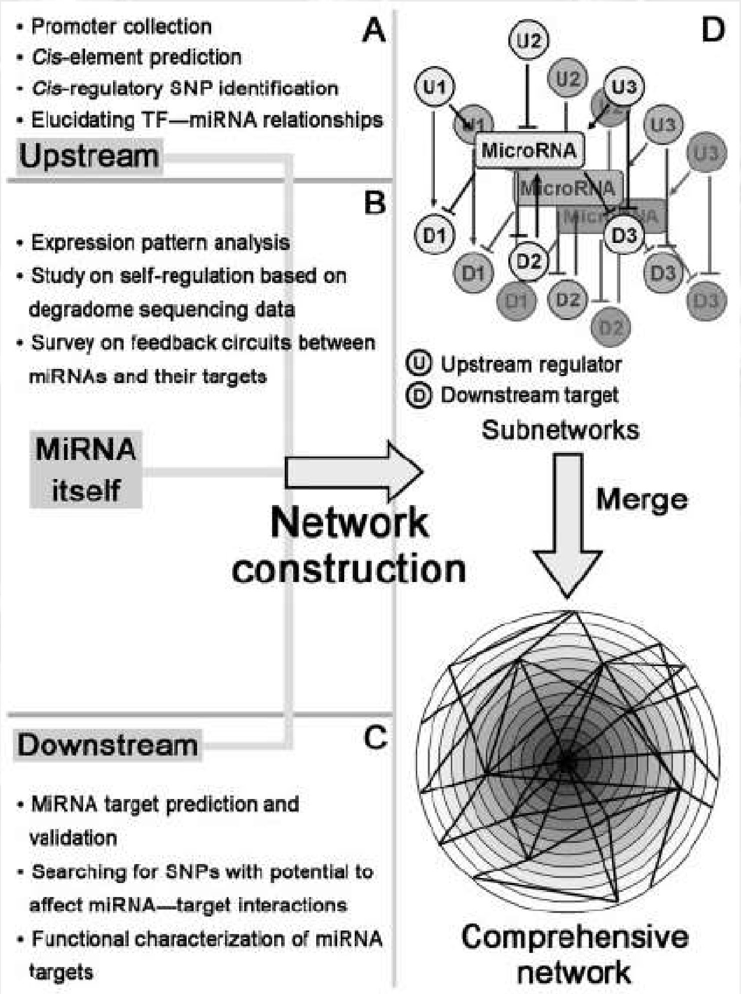
Target determination



miRNA regulatory network

miRNA介导的基因调控网络构建思路

对已有*PHR1—miR399—PHO2*调控通路进行验证性重建



Upstream

- Promoter collection
- *Cis*-element prediction
- *Cis*-regulatory SNP identification
- Elucidating TF—miRNA relationships

MiRNA itself

- Expression pattern analysis
- Study on self-regulation based on degradome sequencing data
- Survey on feedback circuits between miRNAs and their targets

Downstream

- MiRNA target prediction and validation
- Searching for SNPs with potential to affect miRNA—target interactions
- Functional characterization of miRNA targets

Network construction

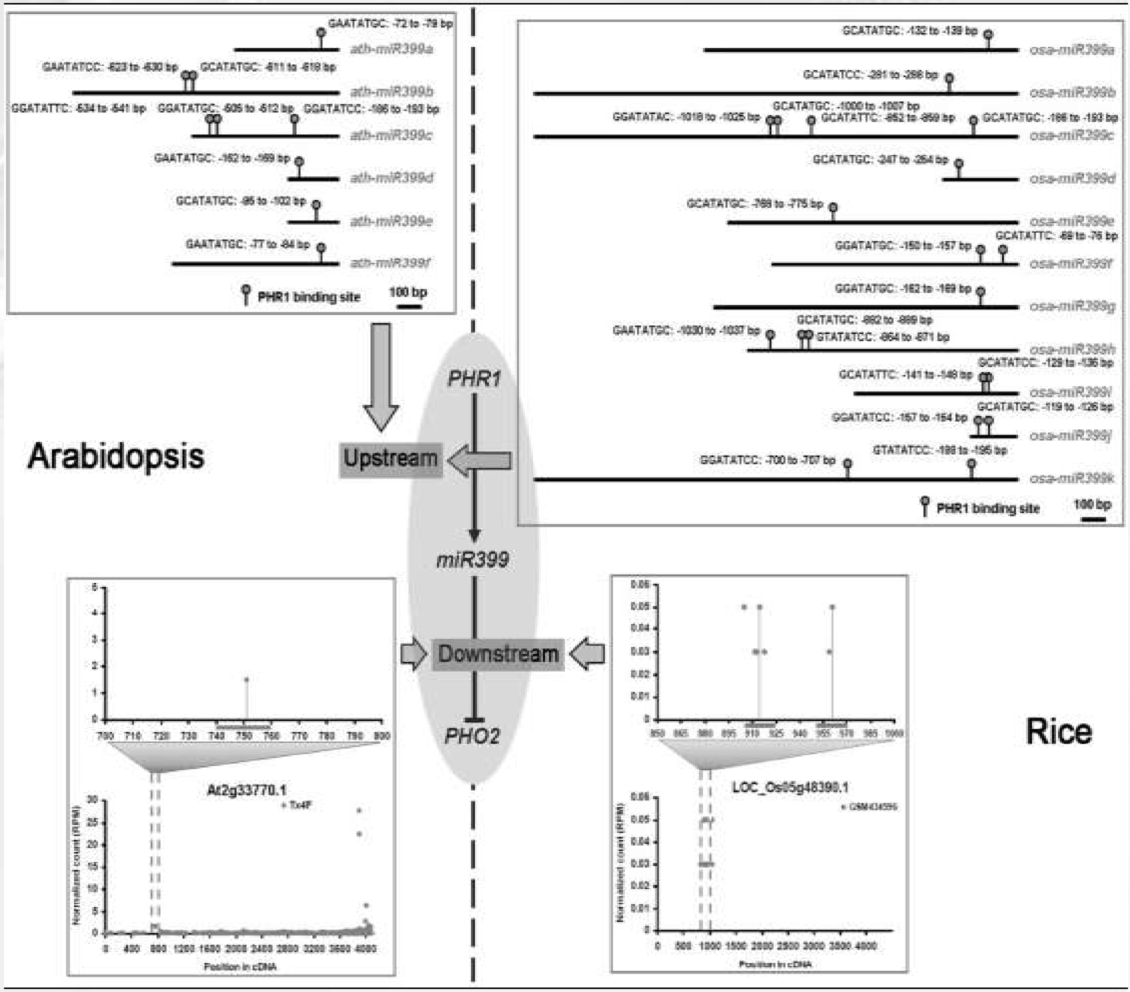
Ⓤ Upstream regulator
Ⓓ Downstream target

Subnetworks

Merge

Comprehensive network

A
B
C
D
U1
U2
U3
D1
D2
D3
MicroRNA

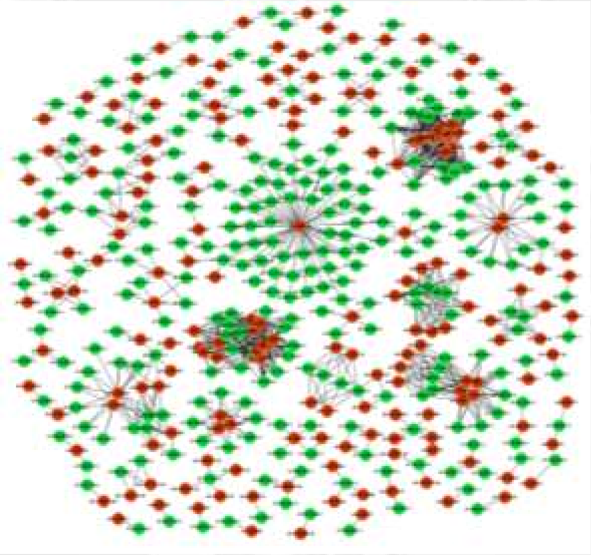


GAATATGC: -72 to -79 bp
ath-miR399a
GAATATCC: -623 to -630 bp
GCATATGC: -611 to -618 bp
ath-miR399b
GGATATTC: -534 to -541 bp
GGATATGC: -505 to -512 bp
GGATATCC: -186 to -193 bp
ath-miR399c
GAATATGC: -162 to -169 bp
ath-miR399d
GCATATGC: -95 to -102 bp
ath-miR399e
GAATATGC: -77 to -84 bp
ath-miR399f
PHR1 binding site
100 bp
Arabidopsis
Upstream
PHR1
miR399
Downstream
PHO2
GCATATGC: -132 to -139 bp
osa-miR399a
GCATATCC: -281 to -288 bp
osa-miR399b
GGATATAC: -1018 to -1025 bp
GCATATGC: -1000 to -1007 bp
GCATATTC: -852 to -859 bp
GCATATGC: -186 to -193 bp
osa-miR399c
GCATATGC: -247 to -254 bp
osa-miR399d
GCATATGC: -768 to -775 bp
osa-miR399e
GGATATGC: -150 to -157 bp
GCATATTC: -69 to -76 bp
osa-miR399f
GGATATGC: -162 to -169 bp
osa-miR399g
GAATATGC: -1030 to -1037 bp
GCATATGC: -882 to -889 bp
GTATATCC: -864 to -871 bp
osa-miR399h
GCATATCC: -129 to -136 bp
GCATATTC: -141 to -148 bp
osa-miR399i
GCATATGC: -119 to -126 bp
GGATATCC: -157 to -164 bp
osa-miR399j
GTATATCC: -188 to -195 bp
GGATATCC: -700 to -707 bp
osa-miR399k
PHR1 binding site
100 bp
At2g33770.1
Normalized count (RPM)
Position in cDNA
LOC_Os05g48390.1
Normalized count (RPM)
Position in cDNA
Rice

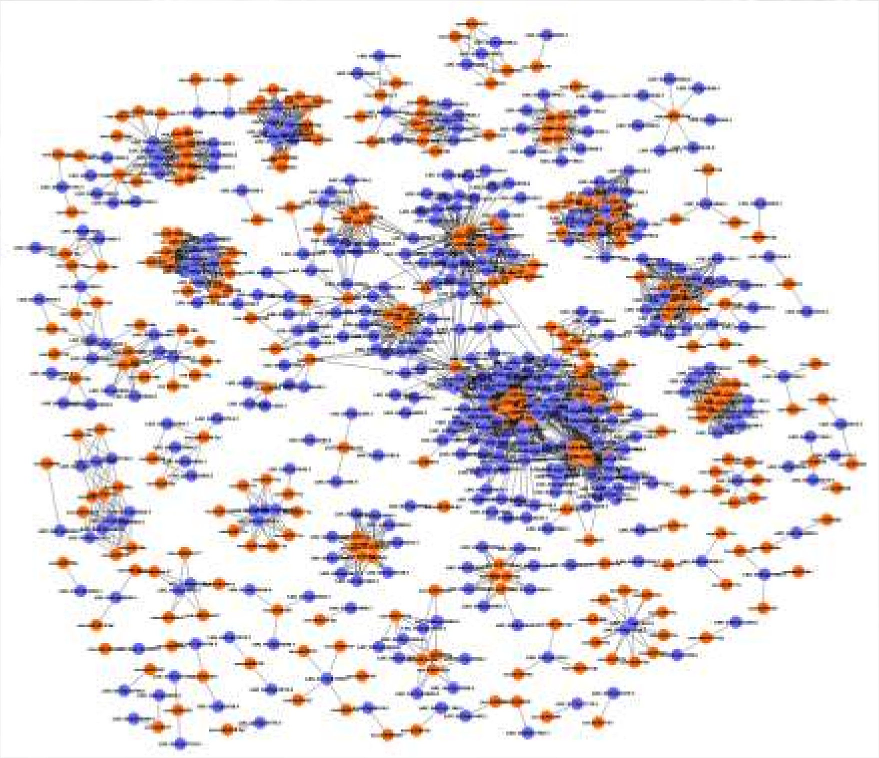
miRNA/miRNA* regulatory network

基于miRNA target lists、miRNA* target lists和co-regulated target lists构建network

Arabidopsis (all targets)

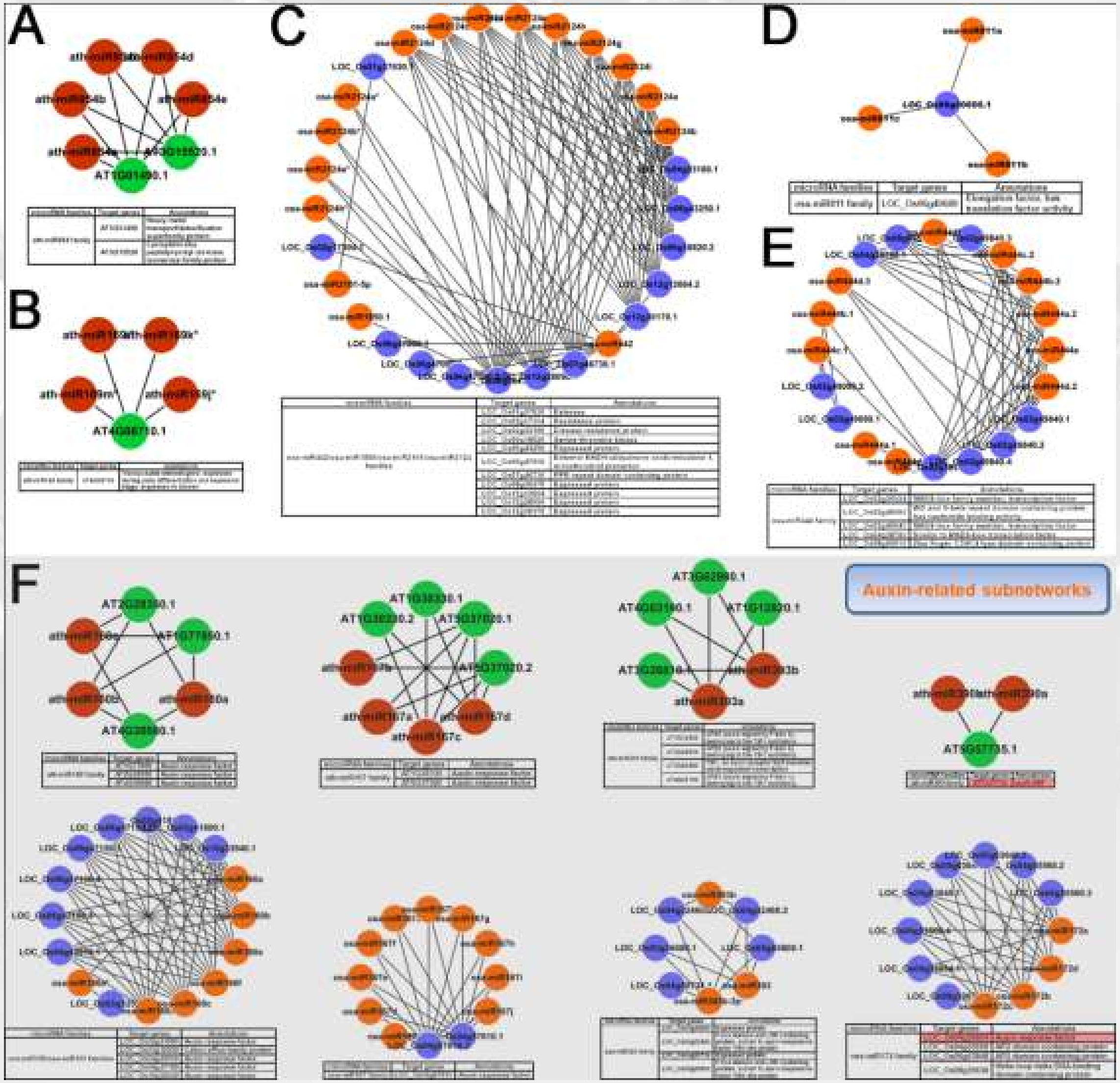


Rice (all targets)



基于降解组测序数据，对拟南芥、水稻中已注释miRNAs的靶基因预测和大规模鉴定；利用sRNA高通量测序数据，基于表达量鉴定了所有miRNAs对应的miRNA*s，并对miRNA*s的潜在的靶基因进行了预测鉴定；最终构建了由miRNAs/miRNA*s介导的基因调控网络

Subnet analysis

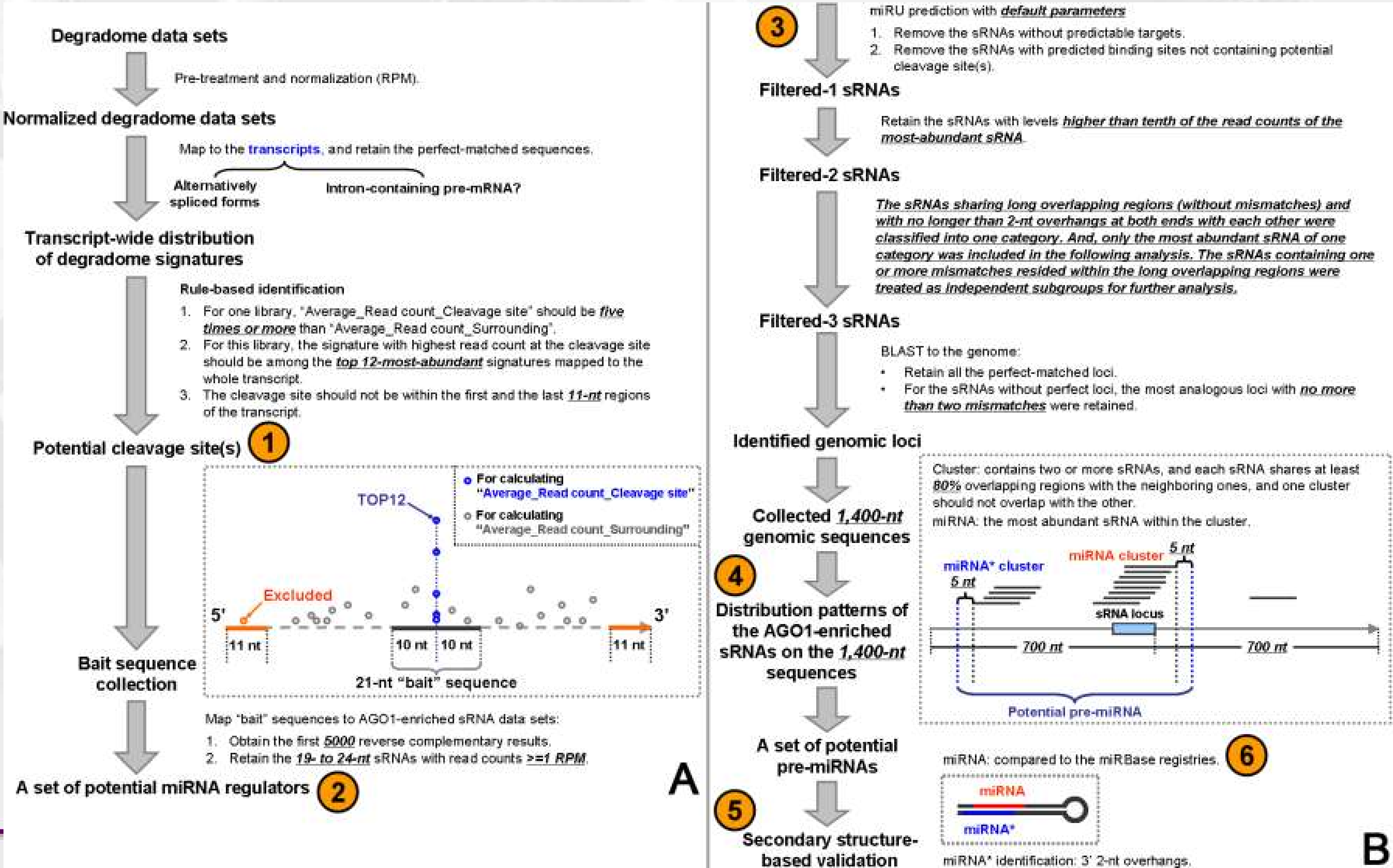


发现一些可能具有重要生物学功能（参与重金属胁迫应答、植物抗病相关）的子网络

生长素信号相关子网络在拟南芥、水稻中具有高度保守性，但也发现了一些物种特异的调控关系（红色背景）

A reverse framework

BiB 2012



Degradome data sets

Pre-treatment and normalization (RPM).

Normalized degradome data sets

Map to the transcripts, and retain the perfect-matched sequences.

Alternatively spliced forms

Intron-containing pre-mRNA?

Transcript-wide distribution of degradome signatures

Rule-based identification

1. For one library, "Average_Read count_Cleavage site" should be *five times or more* than "Average_Read count_Surrounding".
2. For this library, the signature with highest read count at the cleavage site should be among the *top 12-most-abundant* signatures mapped to the whole transcript.
3. The cleavage site should not be within the first and the last *11-nt* regions of the transcript.

Potential cleavage site(s) 1

For calculating "Average_Read count_Cleavage site"
For calculating "Average_Read count_Surrounding"
TOP12
Excluded
5'
3'
11 nt
10 nt
10 nt
11 nt
21-nt "bait" sequence

Bait sequence collection

Map "bait" sequences to AGO1-enriched sRNA data sets:

1. Obtain the first *5000* reverse complementary results.
2. Retain the *19- to 24-nt* sRNAs with read counts *>=1 RPM*.

A set of potential miRNA regulators 2

3 miRU prediction with *default parameters*

1. Remove the sRNAs without predictable targets.
2. Remove the sRNAs with predicted binding sites not containing potential cleavage site(s).

Filtered-1 sRNAs

Retain the sRNAs with levels *higher than tenth of the read counts of the most-abundant sRNA*.

Filtered-2 sRNAs

The sRNAs sharing long overlapping regions (without mismatches) and with no longer than 2-nt overhangs at both ends with each other were classified into one category. And, only the most abundant sRNA of one category was included in the following analysis. The sRNAs containing one or more mismatches resided within the long overlapping regions were treated as independent subgroups for further analysis.

Filtered-3 sRNAs

BLAST to the genome:

- Retain all the perfect-matched loci.
- For the sRNAs without perfect loci, the most analogous loci with *no more than two mismatches* were retained.

Identified genomic loci

Collected *1,400-nt* genomic sequences

4 Distribution patterns of the AGO1-enriched sRNAs on the *1,400-nt* sequences

Cluster: contains two or more sRNAs, and each sRNA shares at least *80%* overlapping regions with the neighboring ones, and one cluster should not overlap with the other.

miRNA: the most abundant sRNA within the cluster.

A set of potential pre-miRNAs

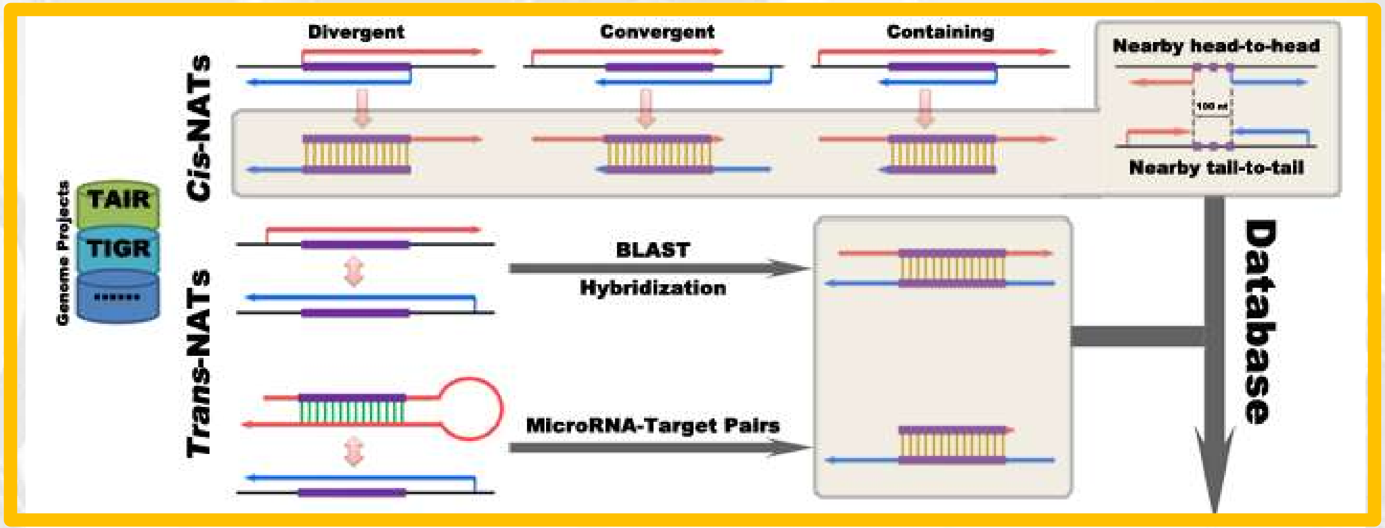
5 Secondary structure-based validation

miRNA: compared to the miRBase registries. 6

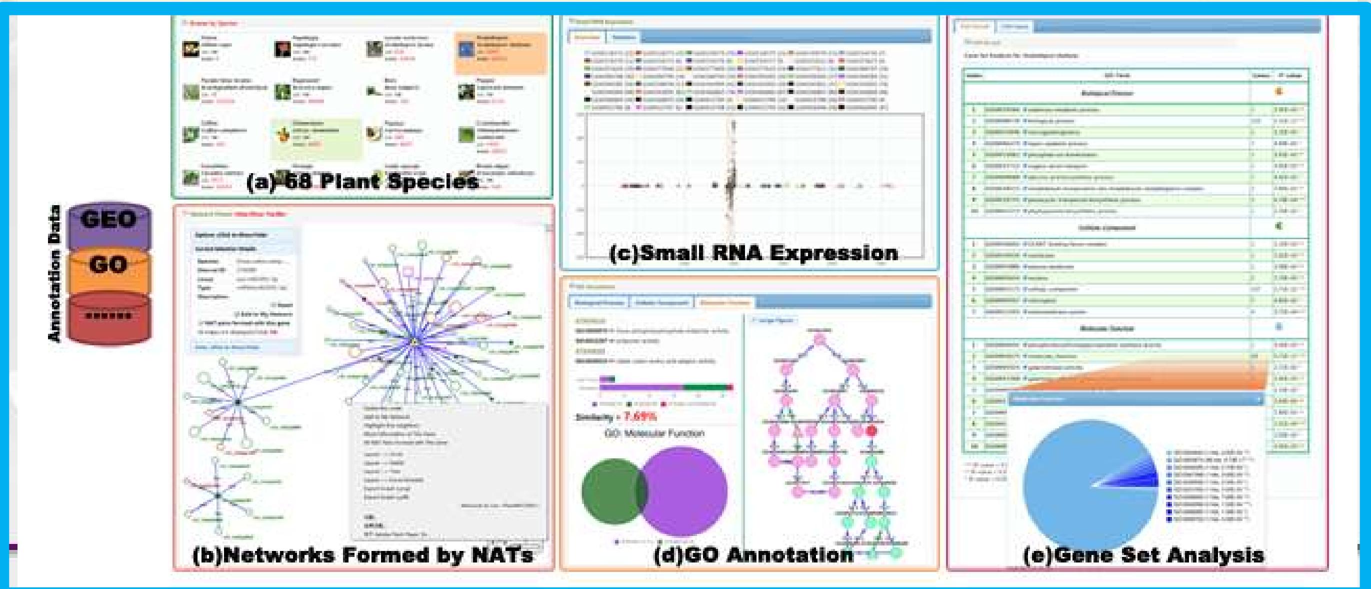
miRNA* identification: 3' 2-nt overhangs.

NATs

Natural antisense transcripts (NATs): *cis*-NATs and *trans*-NATs (Wang *et al.*, 2006; Zhou *et al.*, 2009)



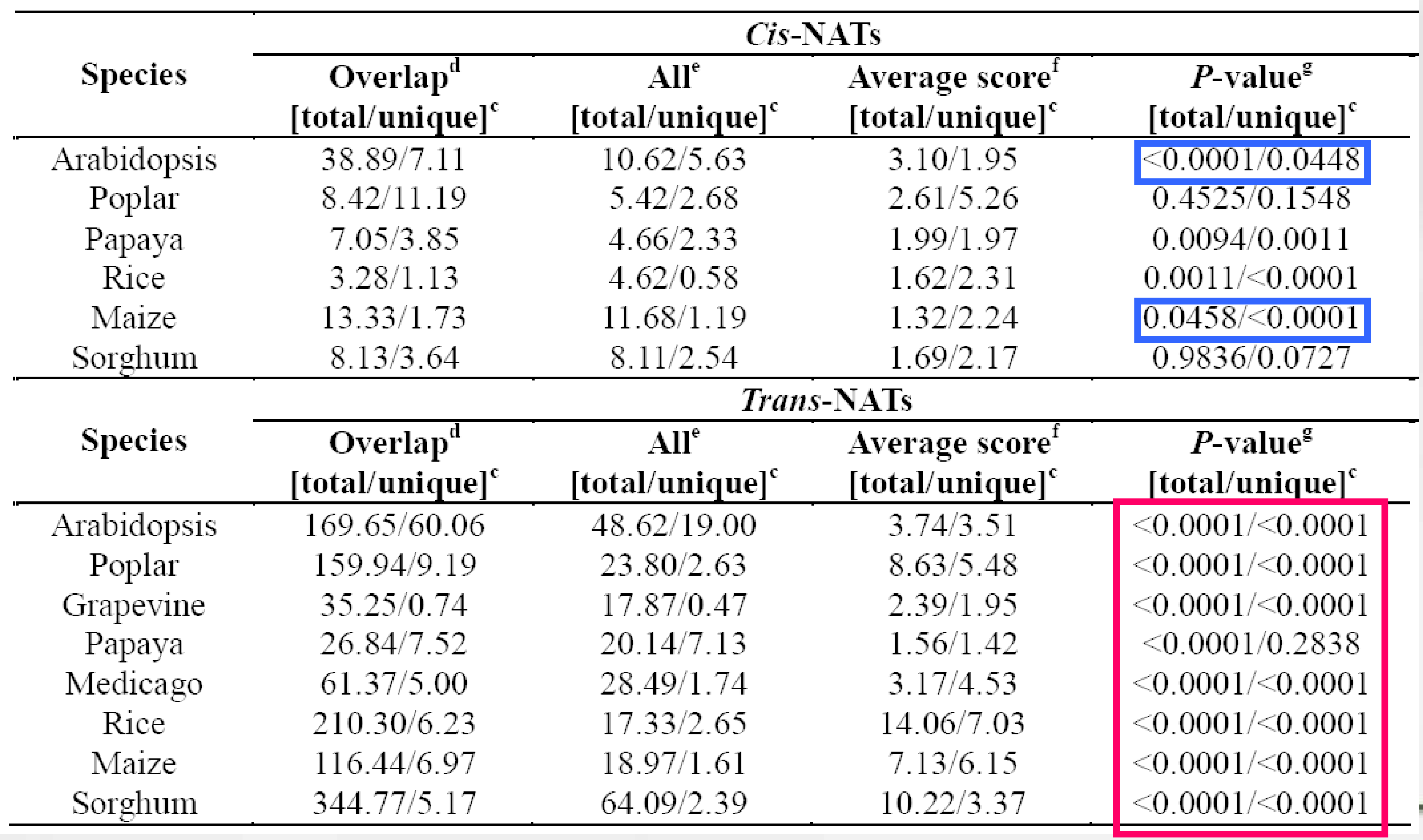
Prediction of NATs



Database Implementation

NATs Generated Small RNAs

sRNA loci are enriched in the overlapping regions of *trans*-NATs, but not for *cis*-NATs.



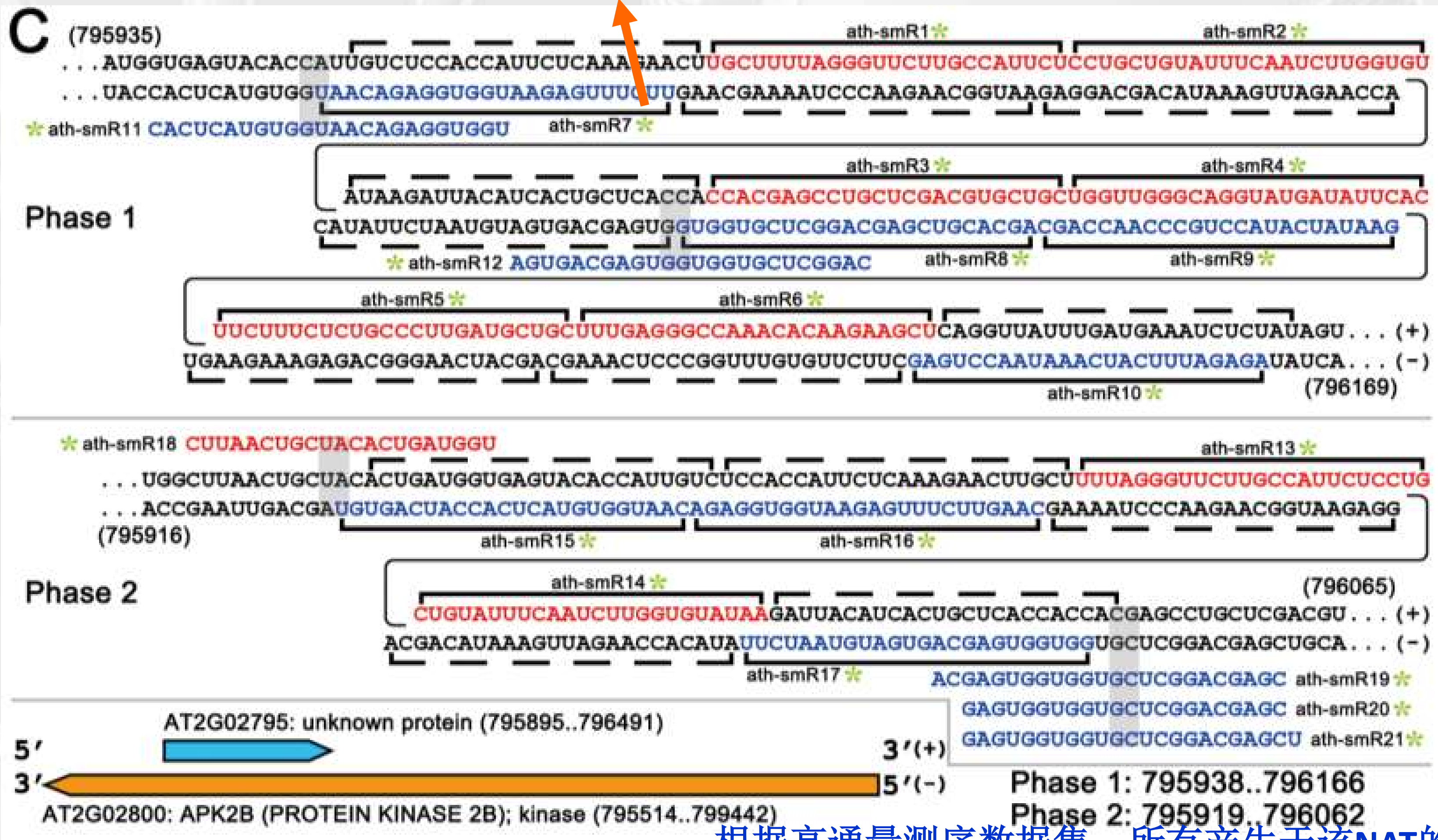
	Cis-NATs			
Species	Overlap[d] [total/unique][c]	All[e] [total/unique][c]	Average score[f] [total/unique][c]	*P*-value[g] [total/unique][c]
Arabidopsis	38.89/7.11	10.62/5.63	3.10/1.95	<0.0001/0.0448
Poplar	8.42/11.19	5.42/2.68	2.61/5.26	0.4525/0.1548
Papaya	7.05/3.85	4.66/2.33	1.99/1.97	0.0094/0.0011
Rice	3.28/1.13	4.62/0.58	1.62/2.31	0.0011/<0.0001
Maize	13.33/1.73	11.68/1.19	1.32/2.24	0.0458/<0.0001
Sorghum	8.13/3.64	8.11/2.54	1.69/2.17	0.9836/0.0727

	Trans-NATs			
Species	Overlap[d] [total/unique][c]	All[e] [total/unique][c]	Average score[f] [total/unique][c]	*P*-value[g] [total/unique][c]
Arabidopsis	169.65/60.06	48.62/19.00	3.74/3.51	<0.0001/<0.0001
Poplar	159.94/9.19	23.80/2.63	8.63/5.48	<0.0001/<0.0001
Grapevine	35.25/0.74	17.87/0.47	2.39/1.95	<0.0001/<0.0001
Papaya	26.84/7.52	20.14/7.13	1.56/1.42	<0.0001/0.2838
Medicago	61.37/5.00	28.49/1.74	3.17/4.53	<0.0001/<0.0001
Rice	210.30/6.23	17.33/2.65	14.06/7.03	<0.0001/<0.0001
Maize	116.44/6.97	18.97/1.61	7.13/6.15	<0.0001/<0.0001
Sorghum	344.77/5.17	64.09/2.39	10.22/3.37	<0.0001/<0.0001

Organ specific - Arabidopsis

Phase-distributed sRNA in the overlapping region of a *cis*-NAT in Arabidopsis

标记星号的是在基因组上仅有一个完全匹配位点的**sRNAs**

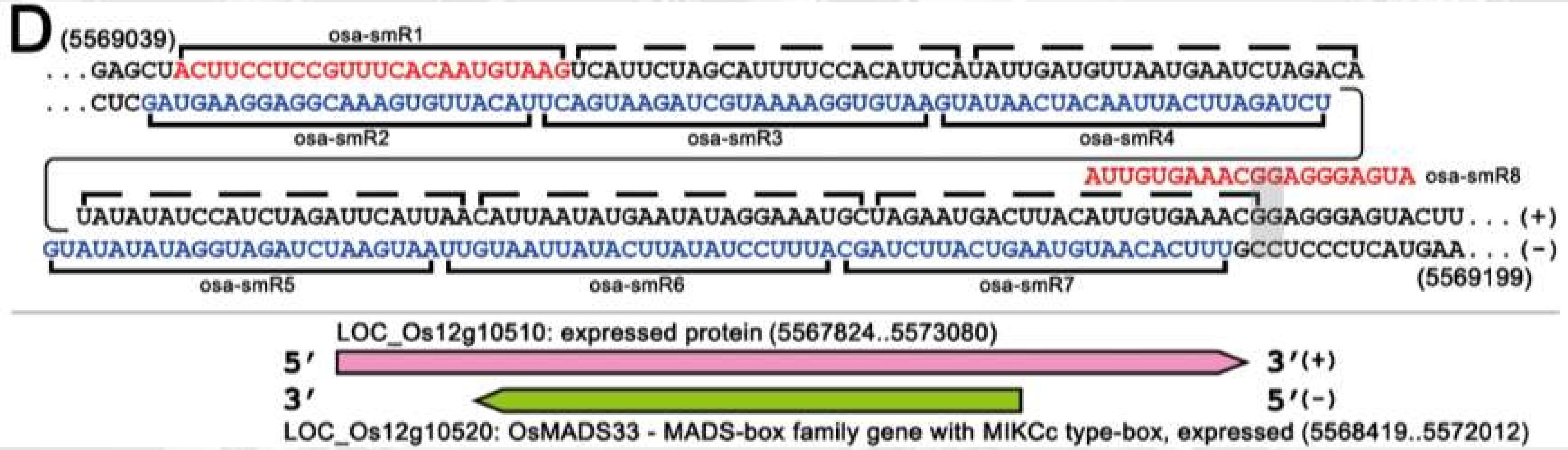

C
(795935)
ath-smR1
ath-smR2
...AUGGUGAGUACACCAUUGUCUCCACCAUUCUCAAAGAACUUGCUUUUAGGGUUCUUGCCAUUCUCCUGCUGUAUUUCAAUCUUGGUGU
...UACCACUCAUGUGGUAACAGAGGUGGUAAGAGUUUCUUGAACGAAAAUCCCAAGAACGGUAAGAGGACGACAUAAAGUUAGAACCA
ath-smR11 CACUCAUGUGGUAACAGAGGUGGU
ath-smR7
ath-smR3
ath-smR4
Phase 1
AUAAGAUUACAUCACUGCUCACCACCACGAGCCUGCUCGACGUGCUGCUGGUUGGGCAGGUAUGAUAUUCAC
CAUAUUCUAAUGUAGUGACGAGUGGUGGUGCUCGGACGAGCUGCACGACGACCAACCCGUCCAUACUAUAAG
ath-smR12 AGUGACGAGUGGUGGUGCUCGGAC
ath-smR8
ath-smR9
ath-smR5
ath-smR6
UUCUUUCUCUGCCCUUGAUGCUGCUUUGAGGGCCAAACACAAGAAGCUCAGGUUAUUUGAUGAAAUCUCUAUAGU... (+)
UGAAGAAAGAGACGGGAACUACGACGAAACUCCCGGUUUGUGUUCUUCGAGUCCAAUAAACUACUUUAGAGAUAUCA... (-)
ath-smR10
(796169)
ath-smR18 CUUAACUGCUACACUGAUGGU
ath-smR13
...UGGCUUAACUGCUACACUGAUGGUGAGUACACCAUUGUCUCCACCAUUCUCAAAGAACUUGCUUUUAGGGUUCUUGCCAUUCUCCUG
...ACCGAAUUGACGAUGUGACUACCACUCAUGUGGUAACAGAGGUGGUAAGAGUUUCUUGAACGAAAAUCCCAAGAACGGUAAGAGG
(795916)
ath-smR15
ath-smR16
Phase 2
ath-smR14
(796065)
CUGUAUUUCAAUCUUGGUGUAUAAGAUUACAUCACUGCUCACCACCACGAGCCUGCUCGACGU... (+)
ACGACAUAAAGUUAGAACCACAUAUUCUAAUGUAGUGACGAGUGGUGGUGCUCGGACGAGCUGCA... (-)
ath-smR17
ACGAGUGGUGGUGCUCGGACGAGC ath-smR19
GAGUGGUGGUGCUCGGACGAGC ath-smR20
GAGUGGUGGUGCUCGGACGAGCU ath-smR21
AT2G02795: unknown protein (795895..796491)
5′
3′(+)
3′
5′(-)
AT2G02800: APK2B (PROTEIN KINASE 2B); kinase (795514..799442)
Phase 1: 795938..796166
Phase 2: 795919..796062

根据高通量测序数据集，所有产生于该**NAT**的相位分布**sRNAs**均只在拟南芥花器官中被克隆到。

Exclusively cloned from floral organs

Organ specific - rice

Phased sRNA in the overlapping region of a *cis*-NAT in rice


D
(5569039)
osa-smR1
...GAGCUACUUCCUCCGUUUCACAAUGUAAGUCAUUCUAGCAUUUUCCACAUUCAUAUUGAUGUUAAUGAAUCUAGACA
...CUCGAUGAAGGAGGCAAAGUGUUACAUUCAGUAAGAUCGUAAAAGGUGUAAGUAUAACUACAAUUACUUAGAUCU
osa-smR2
osa-smR3
osa-smR4
AUUGUGAAACGGAGGGAGUA osa-smR8
UAUAUAUCCAUCUAGAUUCAUUAACAUUAAUAUGAAUAUAGGAAAUGCUAGAAUGACUUACAUUGUGAAACGGAGGGAGUACUU... (+)
GUAUAUAUAGGUAGAUCUAAGUAAUUGUAAUUAUACUUAUAUCCUUUACGAUCUUACUGAAUGUAACACUUUGCCUCCCUCAUGAA... (−)
(5569199)
osa-smR5
osa-smR6
osa-smR7
LOC_Os12g10510: expressed protein (5567824..5573080)
5′
3′(+)
3′
5′(−)
LOC_Os12g10520: OsMADS33 - MADS-box family gene with MIKCc type-box, expressed (5568419..5572012)

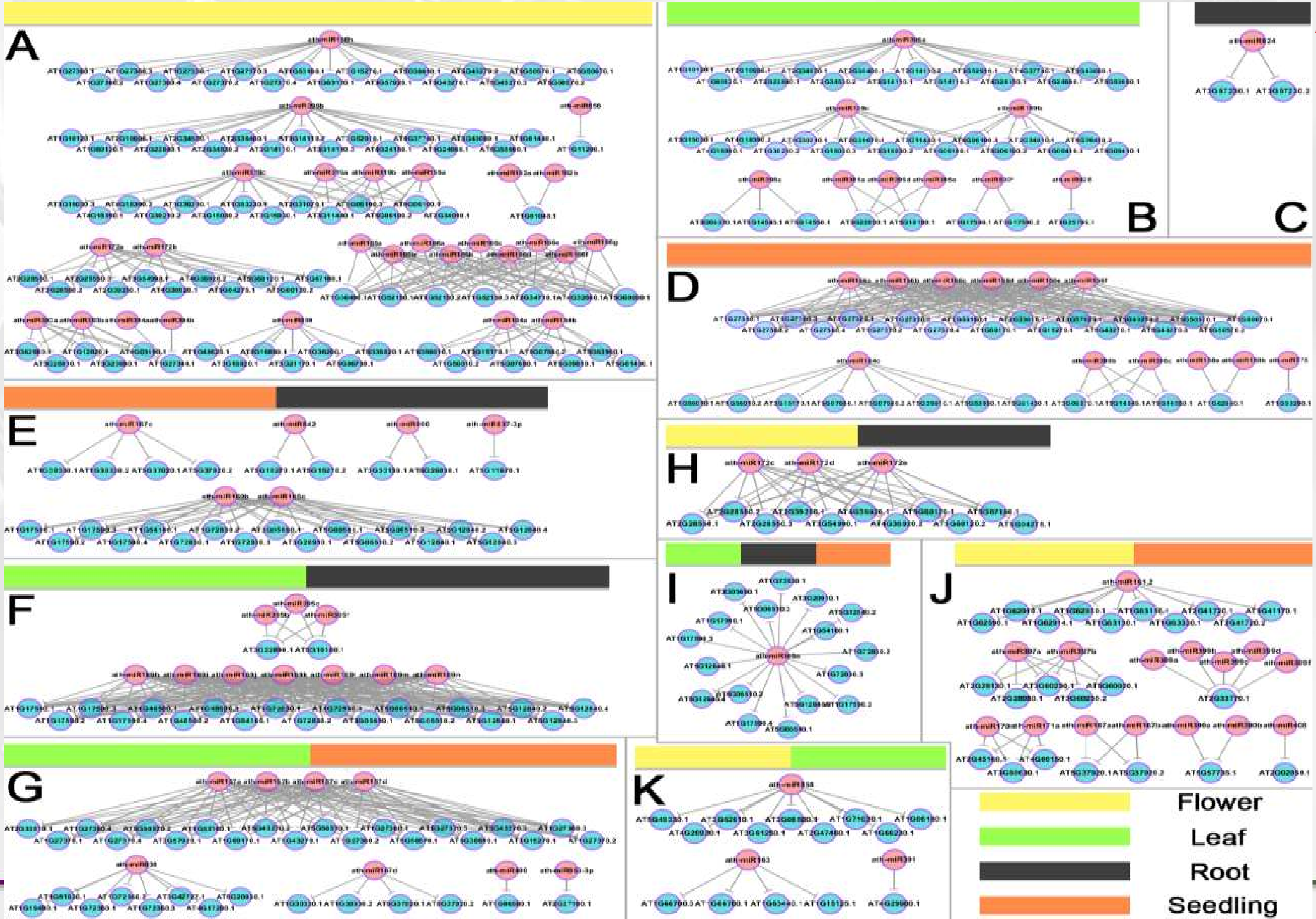
Exclusively cloned from grains

根据高通量测序数据集，所有产生于该NAT的相位分布sRNAs均只在水稻谷粒中被克隆到。

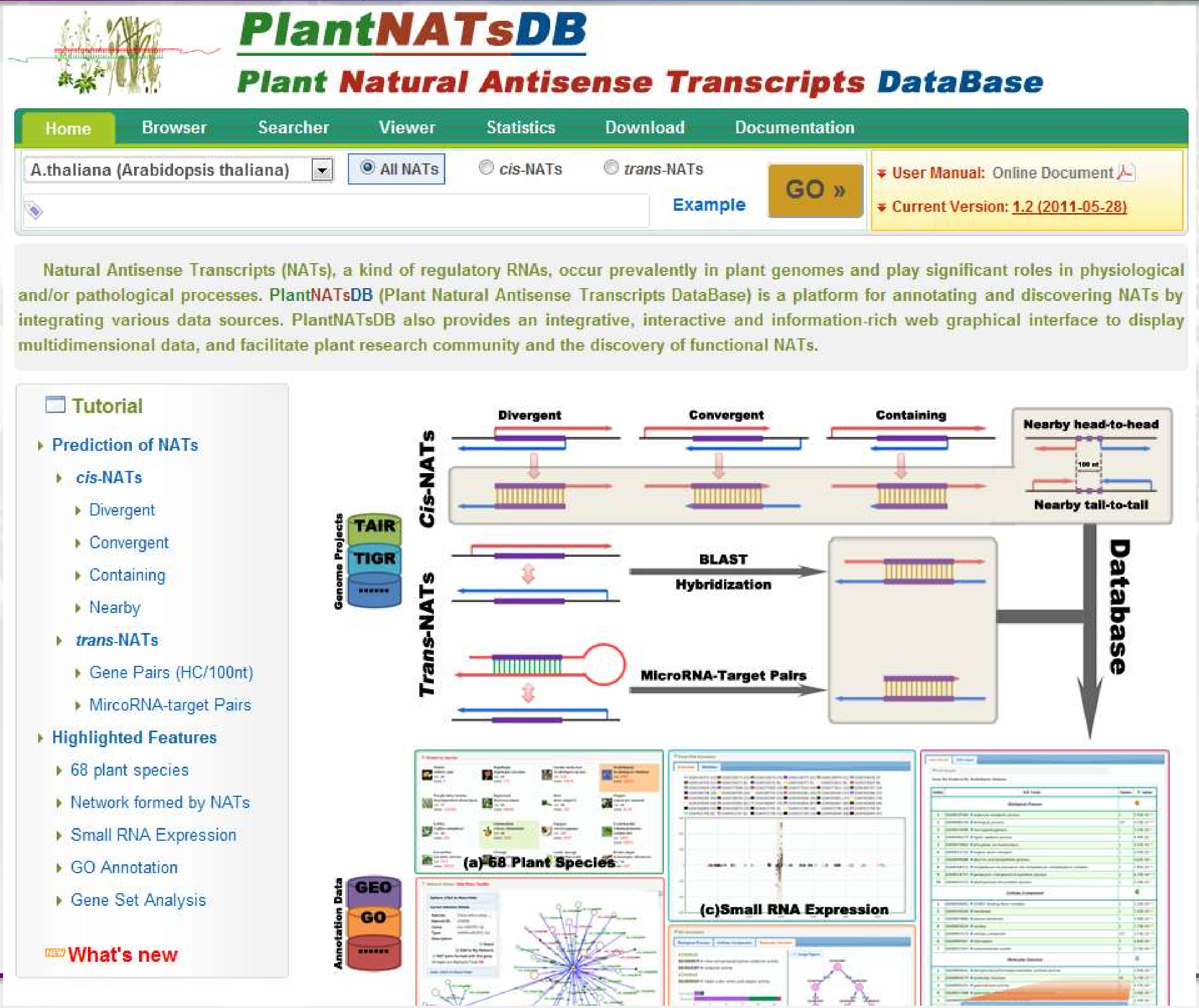
Organ-specific regulatory role?

Organ-specific miRNAs in *Arabidopsis*

RNA, s



NAR , 2012



PlantNATsDB

Plant Natural Antisense Transcripts DataBase

Home | Browser | Searcher | Viewer | Statistics | Download | Documentation

A.thaliana (Arabidopsis thaliana) ▾ ● All NATs ○ *cis*-NATs ○ *trans*-NATs

Example GO »

User Manual: Online Document

Current Version: 1.2 (2011-05-28)

Natural Antisense Transcripts (NATs), a kind of regulatory RNAs, occur prevalently in plant genomes and play significant roles in physiological and/or pathological processes. PlantNATsDB (Plant Natural Antisense Transcripts DataBase) is a platform for annotating and discovering NATs by integrating various data sources. PlantNATsDB also provides an integrative, interactive and information-rich web graphical interface to display multidimensional data, and facilitate plant research community and the discovery of functional NATs.

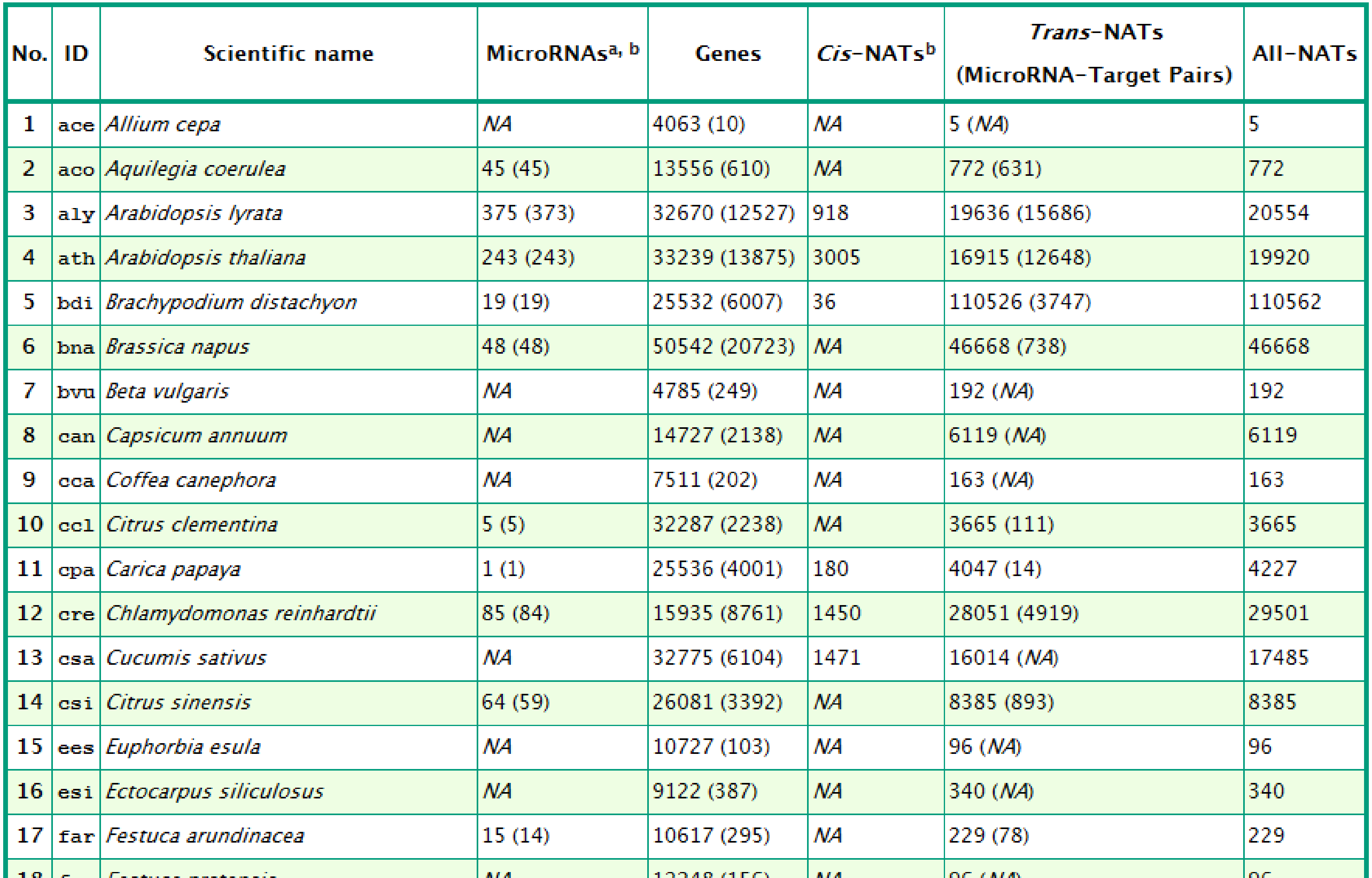
Tutorial

- Prediction of NATs
  - *cis*-NATs
    - Divergent
    - Convergent
    - Containing
    - Nearby
  - *trans*-NATs
    - Gene Pairs (HC/100nt)
    - MircoRNA-target Pairs
- Highlighted Features
  - 68 plant species
  - Network formed by NATs
  - Small RNA Expression
  - GO Annotation
  - Gene Set Analysis

What's new

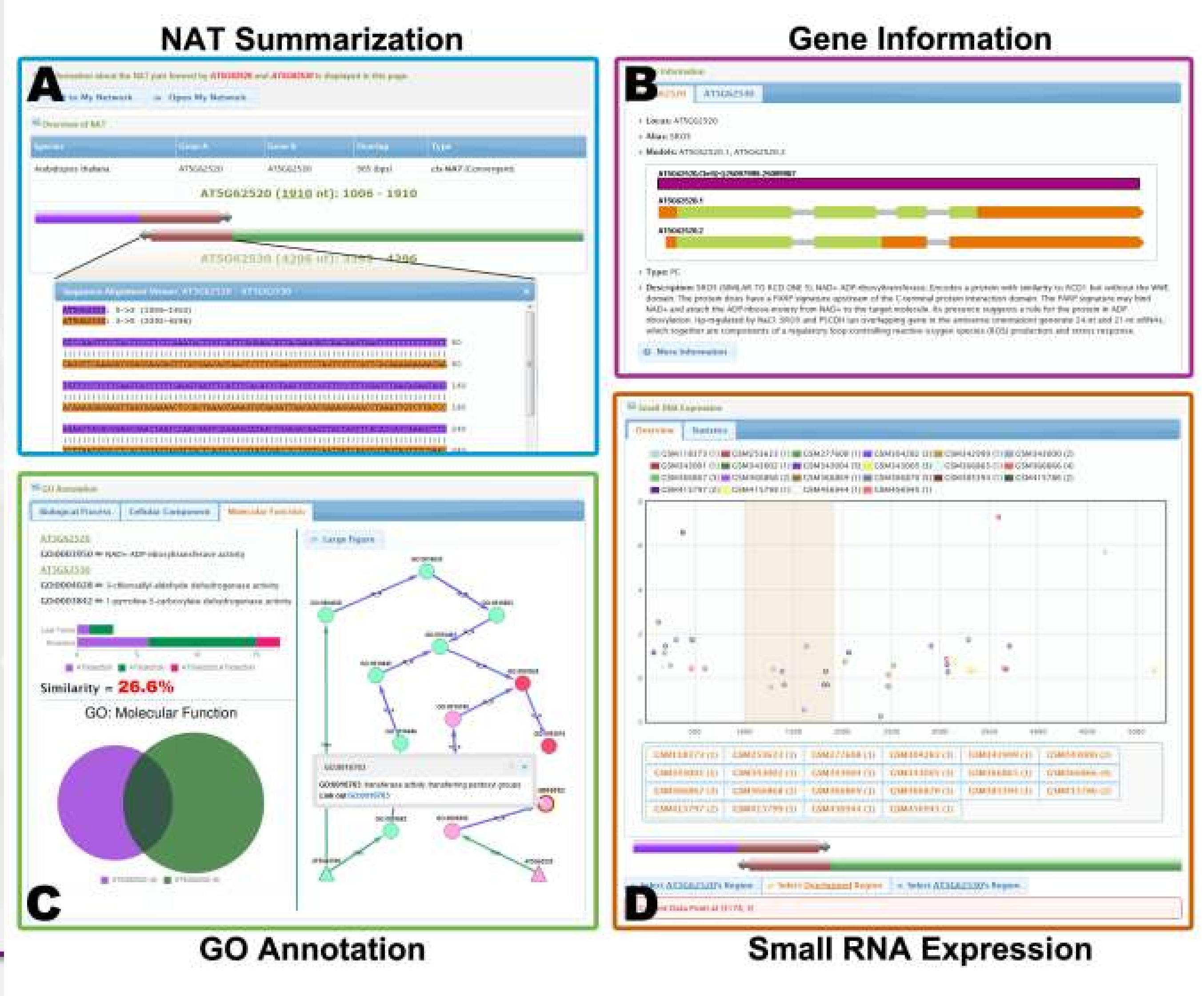
Statistics

PlantNATsDB predicted 2,066,720 NATs from 69 plant species

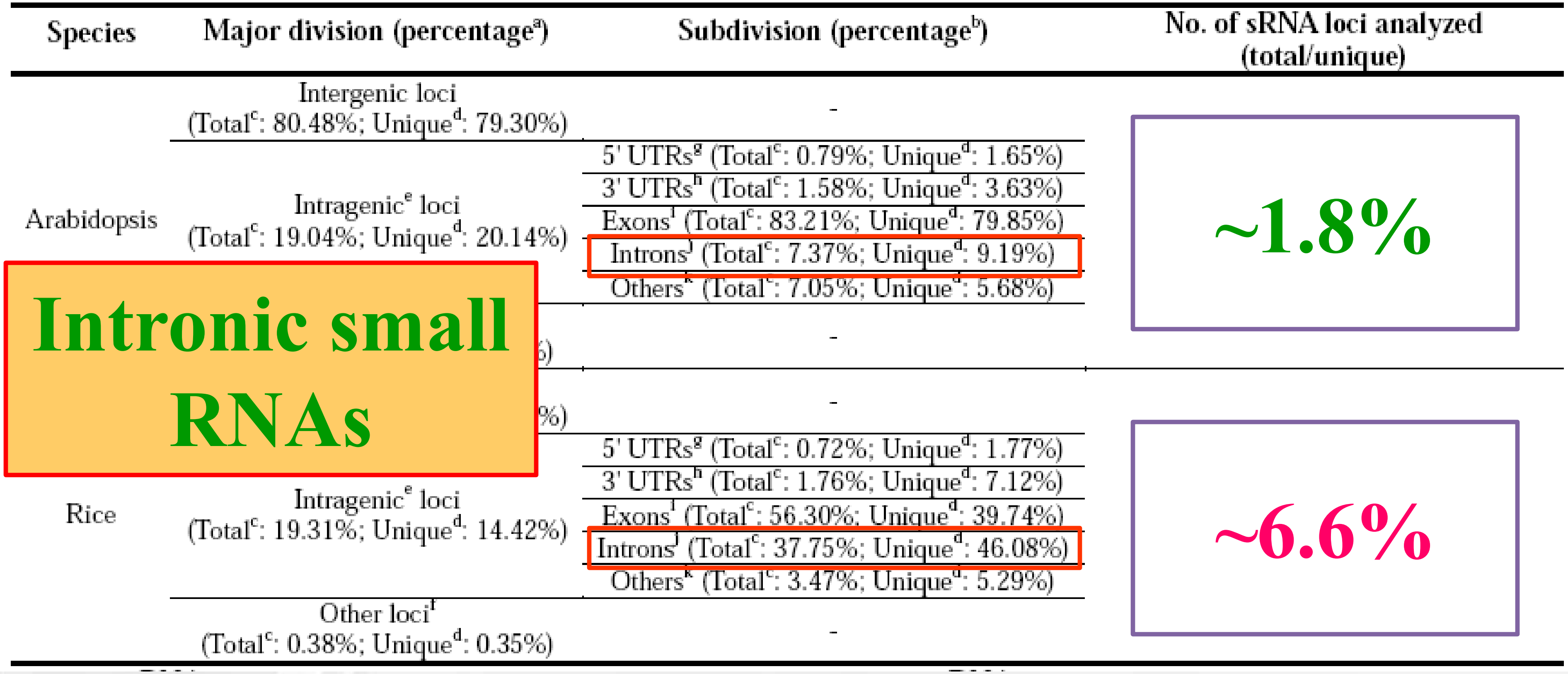


No.	ID	Scientific name	MicroRNAs[a, b]	Genes	*Cis*-NATs[b]	*Trans*-NATs (MicroRNA-Target Pairs)	All-NATs
1	ace	*Allium cepa*	*NA*	4063 (10)	*NA*	5 (*NA*)	5
2	aco	*Aquilegia coerulea*	45 (45)	13556 (610)	*NA*	772 (631)	772
3	aly	*Arabidopsis lyrata*	375 (373)	32670 (12527)	918	19636 (15686)	20554
4	ath	*Arabidopsis thaliana*	243 (243)	33239 (13875)	3005	16915 (12648)	19920
5	bdi	*Brachypodium distachyon*	19 (19)	25532 (6007)	36	110526 (3747)	110562
6	bna	*Brassica napus*	48 (48)	50542 (20723)	*NA*	46668 (738)	46668
7	bvu	*Beta vulgaris*	*NA*	4785 (249)	*NA*	192 (*NA*)	192
8	can	*Capsicum annuum*	*NA*	14727 (2138)	*NA*	6119 (*NA*)	6119
9	cca	*Coffea canephora*	*NA*	7511 (202)	*NA*	163 (*NA*)	163
10	ccl	*Citrus clementina*	5 (5)	32287 (2238)	*NA*	3665 (111)	3665
11	cpa	*Carica papaya*	1 (1)	25536 (4001)	180	4047 (14)	4227
12	cre	*Chlamydomonas reinhardtii*	85 (84)	15935 (8761)	1450	28051 (4919)	29501
13	csa	*Cucumis sativus*	*NA*	32775 (6104)	1471	16014 (*NA*)	17485
14	csi	*Citrus sinensis*	64 (59)	26081 (3392)	*NA*	8385 (893)	8385
15	ees	*Euphorbia esula*	*NA*	10727 (103)	*NA*	96 (*NA*)	96
16	esi	*Ectocarpus siliculosus*	*NA*	9122 (387)	*NA*	340 (*NA*)	340
17	far	*Festuca arundinacea*	15 (14)	10617 (295)	*NA*	229 (78)	229
18	f	*Festuca pratensis*	*NA*	12248 (156)	*NA*	96 (*NA*)	96

An example


NAT Summarization
Gene Information
A
B
AT5G62520 (1910 nt): 1006 - 1910
Similarity = 26.6%
GO: Molecular Function
C
D
GO Annotation
Small RNA Expression

Small RNAs derived from gene models



Species	Major division (percentage[a])	Subdivision (percentage[b])	No. of sRNA loci analyzed (total/unique)
Arabidopsis	Intergenic loci (Total[c]: 80.48%; Unique[d]: 79.30%)	-	
	Intragenic[e] loci (Total[c]: 19.04%; Unique[d]: 20.14%)	5' UTRs[g] (Total[c]: 0.79%; Unique[d]: 1.65%)	
		3' UTRs[h] (Total[c]: 1.58%; Unique[d]: 3.63%)	
		Exons[i] (Total[c]: 83.21%; Unique[d]: 79.85%)	
		Introns[j] (Total[c]: 7.37%; Unique[d]: 9.19%)	
		Others[k] (Total[c]: 7.05%; Unique[d]: 5.68%)	
	[illegible]	-	
Rice	[illegible]	-	
	Intragenic[e] loci (Total[c]: 19.31%; Unique[d]: 14.42%)	5' UTRs[g] (Total[c]: 0.72%; Unique[d]: 1.77%)	
		3' UTRs[h] (Total[c]: 1.76%; Unique[d]: 7.12%)	
		Exons[i] (Total[c]: 56.30%; Unique[d]: 39.74%)	
		Introns[j] (Total[c]: 37.75%; Unique[d]: 46.08%)	
		Others[k] (Total[c]: 3.47%; Unique[d]: 5.29%)	
	Other loci[f] (Total[c]: 0.38%; Unique[d]: 0.35%)	-	

Intronic small RNAs

~1.8%

~6.6%

Identification of intronic long hairpins

RNA, 2011

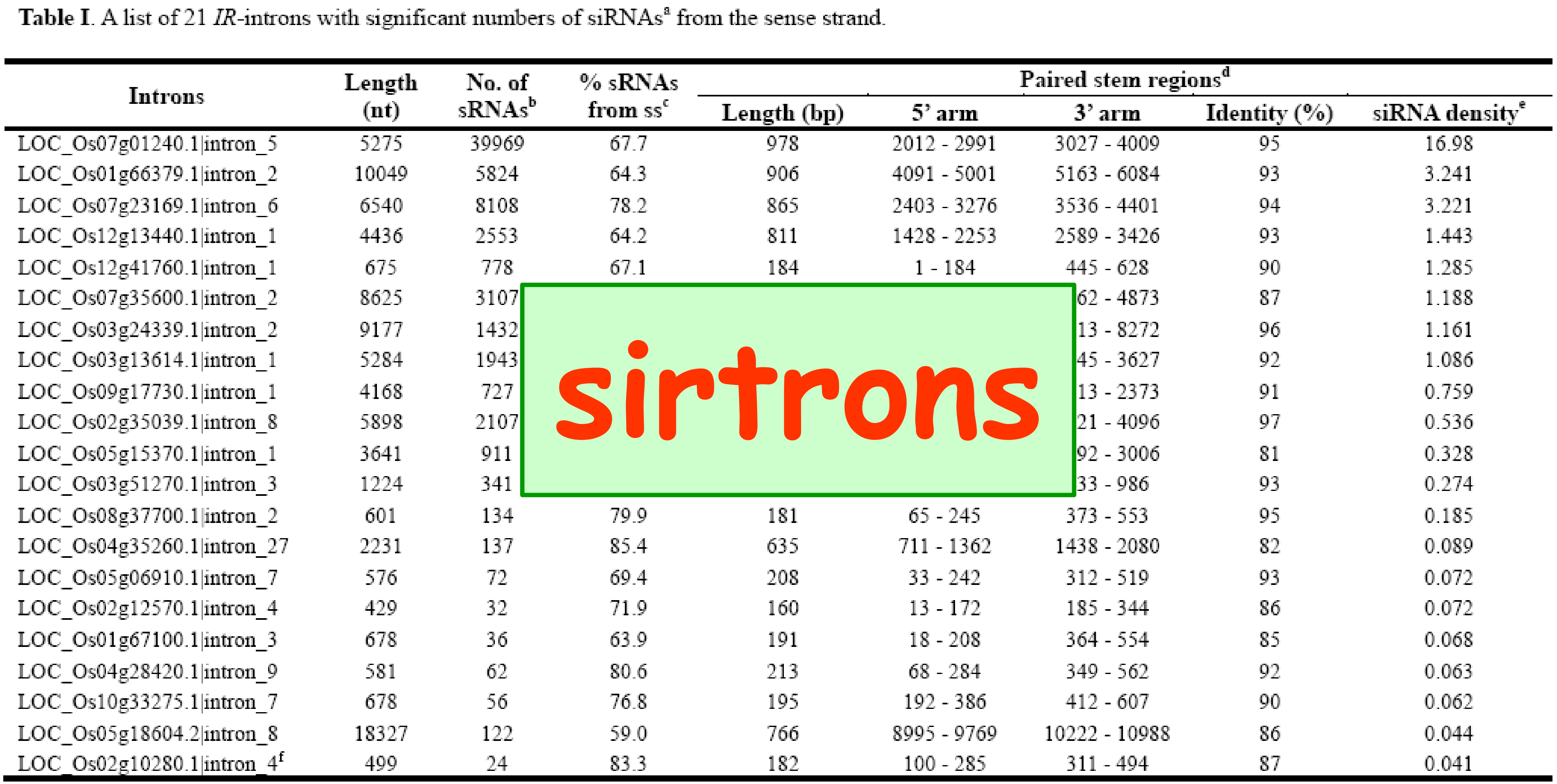
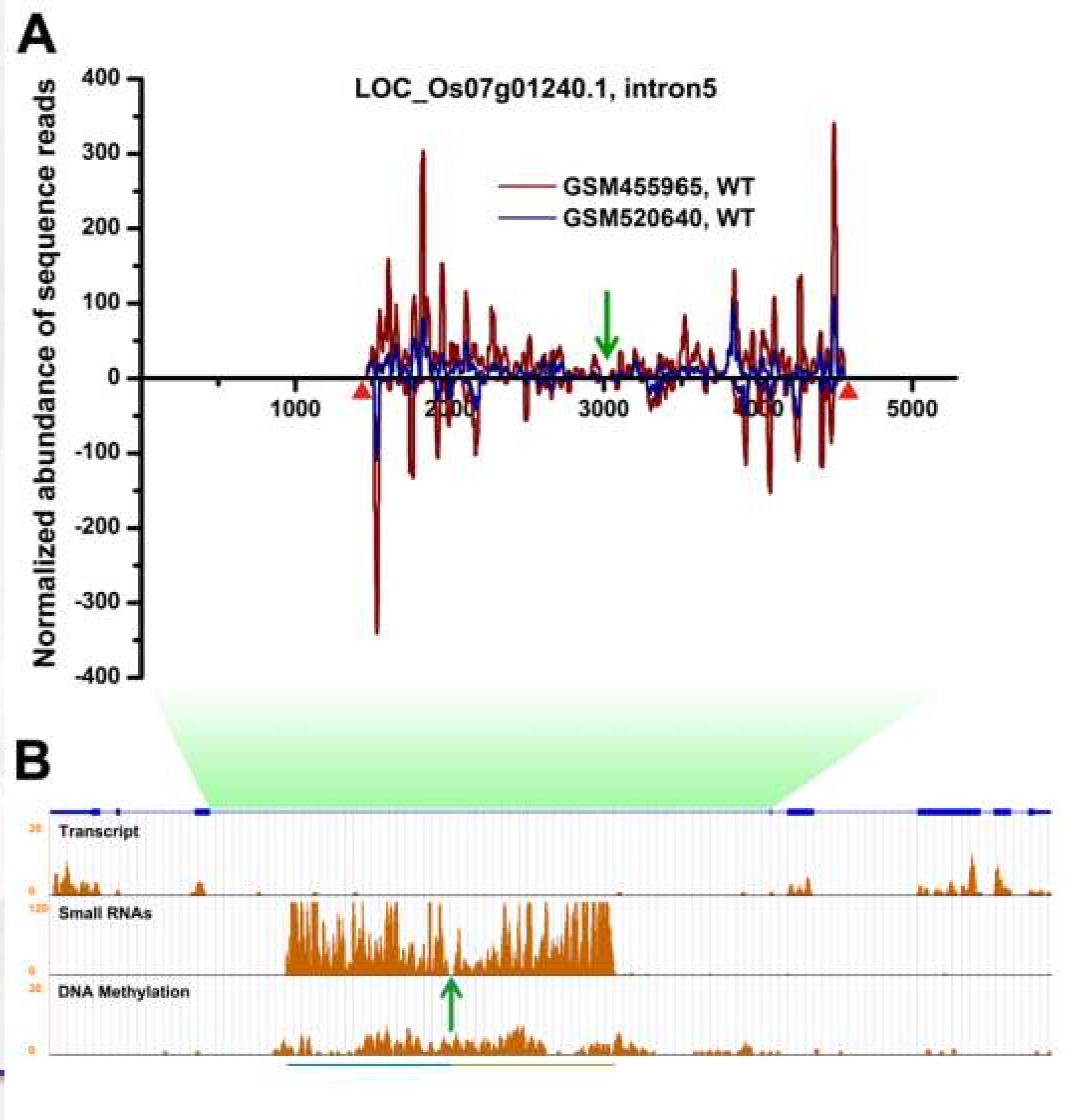


Table I. A list of 21 *IR*-introns with significant numbers of siRNAs[a] from the sense strand.

Introns	Length (nt)	No. of sRNAs[b]	% sRNAs from ss[c]	Paired stem regions[d]				
				Length (bp)	5' arm	3' arm	Identity (%)	siRNA density[e]
LOC_Os07g01240.1|intron_5	5275	39969	67.7	978	2012 - 2991	3027 - 4009	95	16.98
LOC_Os01g66379.1|intron_2	10049	5824	64.3	906	4091 - 5001	5163 - 6084	93	3.241
LOC_Os07g23169.1|intron_6	6540	8108	78.2	865	2403 - 3276	3536 - 4401	94	3.221
LOC_Os12g13440.1|intron_1	4436	2553	64.2	811	1428 - 2253	2589 - 3426	93	1.443
LOC_Os12g41760.1|intron_1	675	778	67.1	184	1 - 184	445 - 628	90	1.285
LOC_Os07g35600.1|intron_2	8625	3107	[illegible]	[illegible]	[illegible]	62 - 4873	87	1.188
LOC_Os03g24339.1|intron_2	9177	1432	[illegible]	[illegible]	[illegible]	13 - 8272	96	1.161
LOC_Os03g13614.1|intron_1	5284	1943	[illegible]	[illegible]	[illegible]	45 - 3627	92	1.086
LOC_Os09g17730.1|intron_1	4168	727	[illegible]	[illegible]	[illegible]	13 - 2373	91	0.759
LOC_Os02g35039.1|intron_8	5898	2107	[illegible]	[illegible]	[illegible]	21 - 4096	97	0.536
LOC_Os05g15370.1|intron_1	3641	911	[illegible]	[illegible]	[illegible]	92 - 3006	81	0.328
LOC_Os03g51270.1|intron_3	1224	341	[illegible]	[illegible]	[illegible]	33 - 986	93	0.274
LOC_Os08g37700.1|intron_2	601	134	79.9	181	65 - 245	373 - 553	95	0.185
LOC_Os04g35260.1|intron_27	2231	137	85.4	635	711 - 1362	1438 - 2080	82	0.089
LOC_Os05g06910.1|intron_7	576	72	69.4	208	33 - 242	312 - 519	93	0.072
LOC_Os02g12570.1|intron_4	429	32	71.9	160	13 - 172	185 - 344	86	0.072
LOC_Os01g67100.1|intron_3	678	36	63.9	191	18 - 208	364 - 554	85	0.068
LOC_Os04g28420.1|intron_9	581	62	80.6	213	68 - 284	349 - 562	92	0.063
LOC_Os10g33275.1|intron_7	678	56	76.8	195	192 - 386	412 - 607	90	0.062
LOC_Os05g18604.2|intron_8	18327	122	59.0	766	8995 - 9769	10222 - 10988	86	0.044
LOC_Os02g10280.1|intron_4[f]	499	24	83.3	182	100 - 285	311 - 494	87	0.041

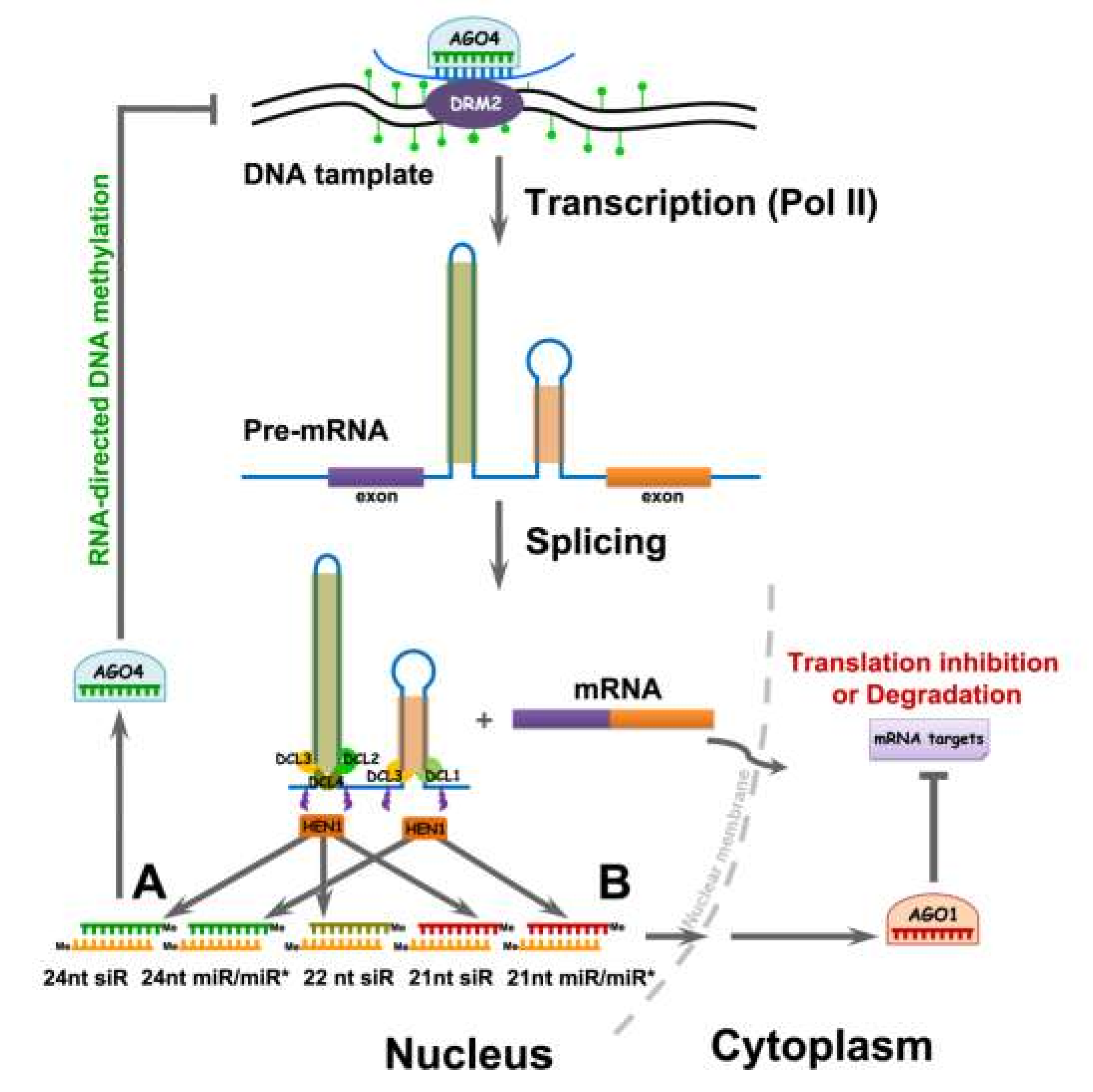
sirtrons

An example of sirtrons

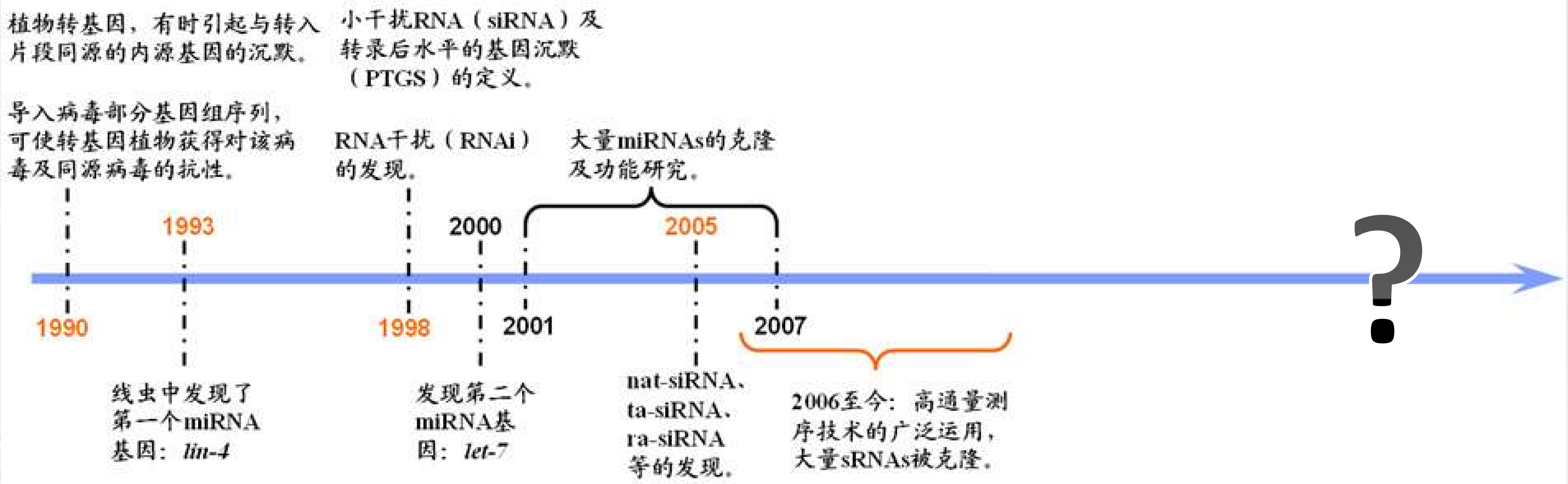


A proposed self-regulation model

RNA , 2012



Prospective



植物转基因，有时引起与转入片段同源的内源基因的沉默。

导入病毒部分基因组序列，可使转基因植物获得对该病毒及同源病毒的抗性。

小干扰RNA（siRNA）及转录后水平的基因沉默（PTGS）的定义。

RNA干扰（RNAi）的发现。

大量miRNAs的克隆及功能研究。

线虫中发现了第一个miRNA基因：*lin-4*

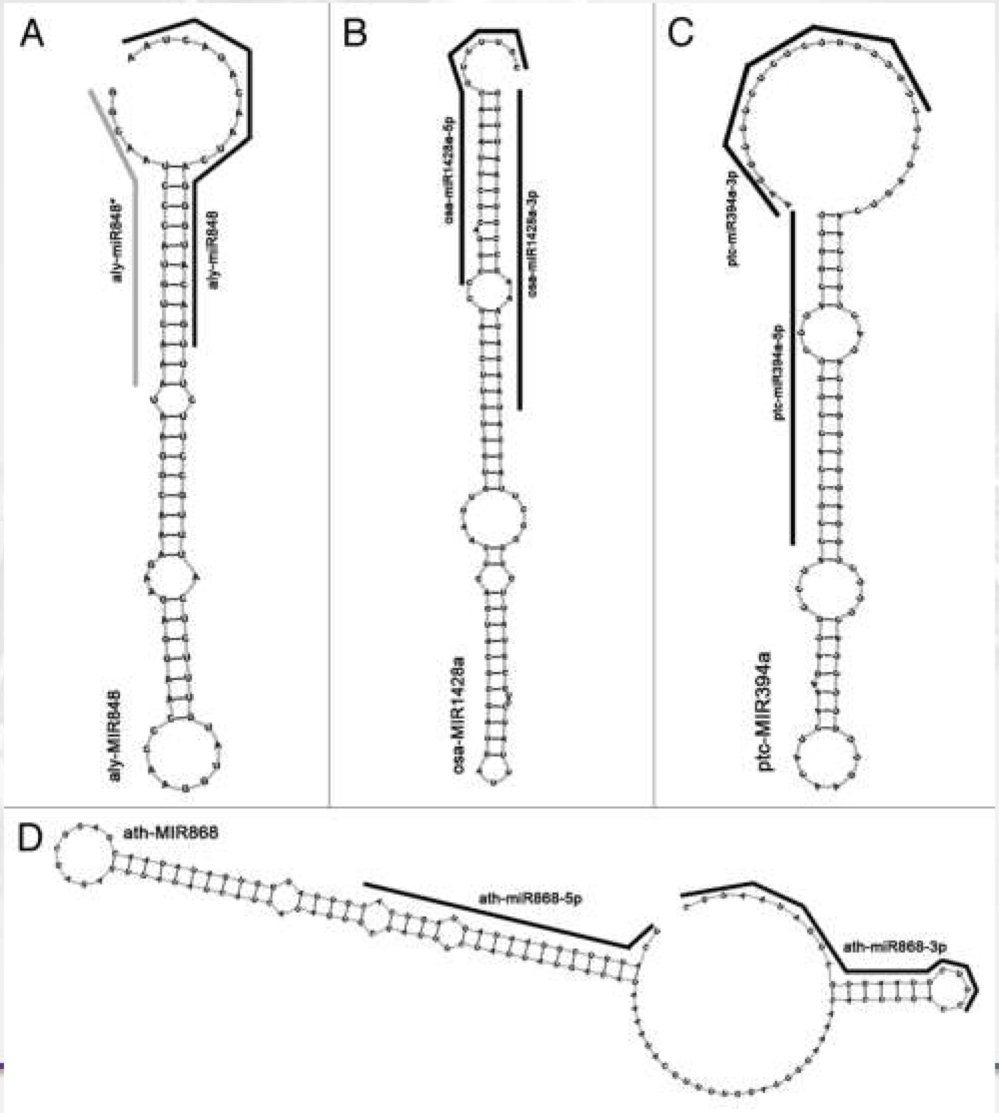
发现第二个miRNA基因：*let-7*

nat-siRNA、ta-siRNA、ra-siRNA等的发现。

2006至今：高通量测序技术的广泛运用，大量sRNAs被克隆。

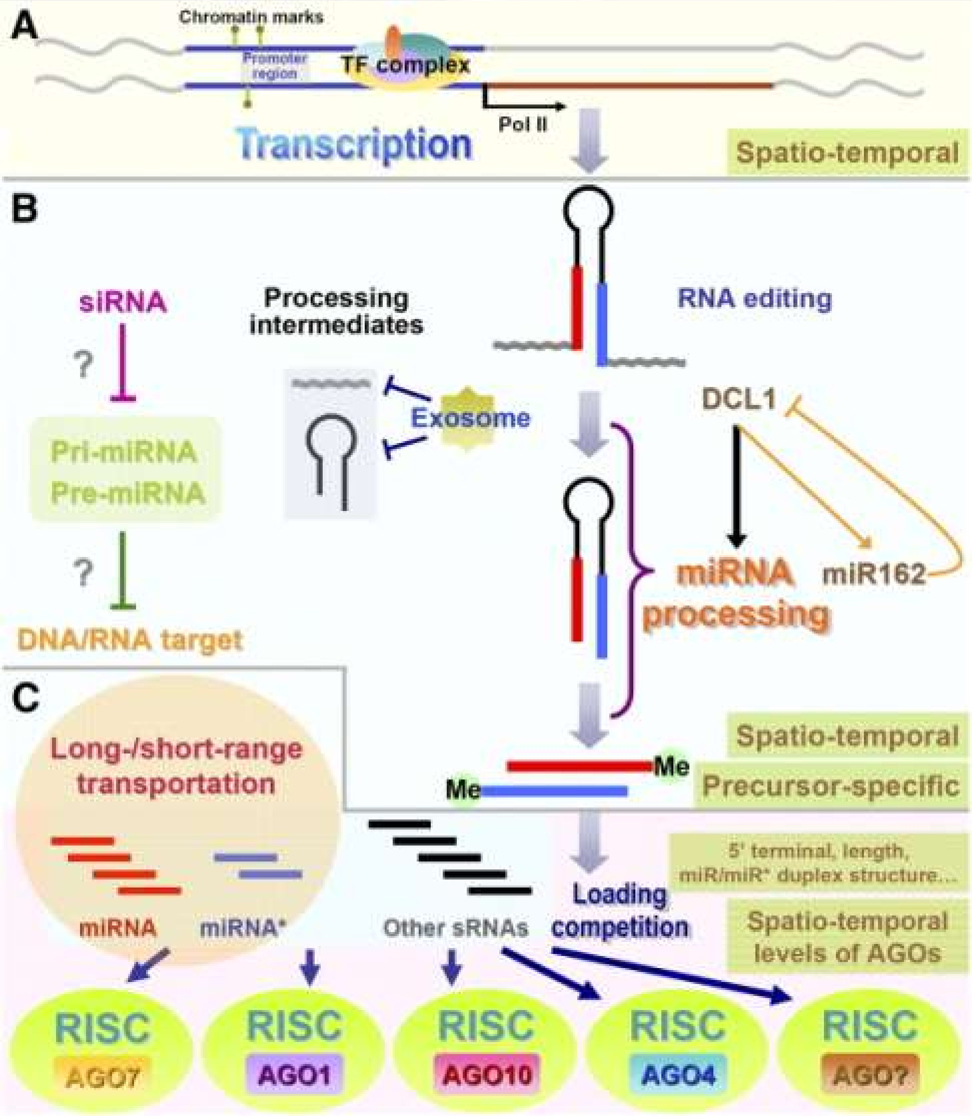
Are all the miRBase-registered microRNAs true?

RNA Biology, 2012

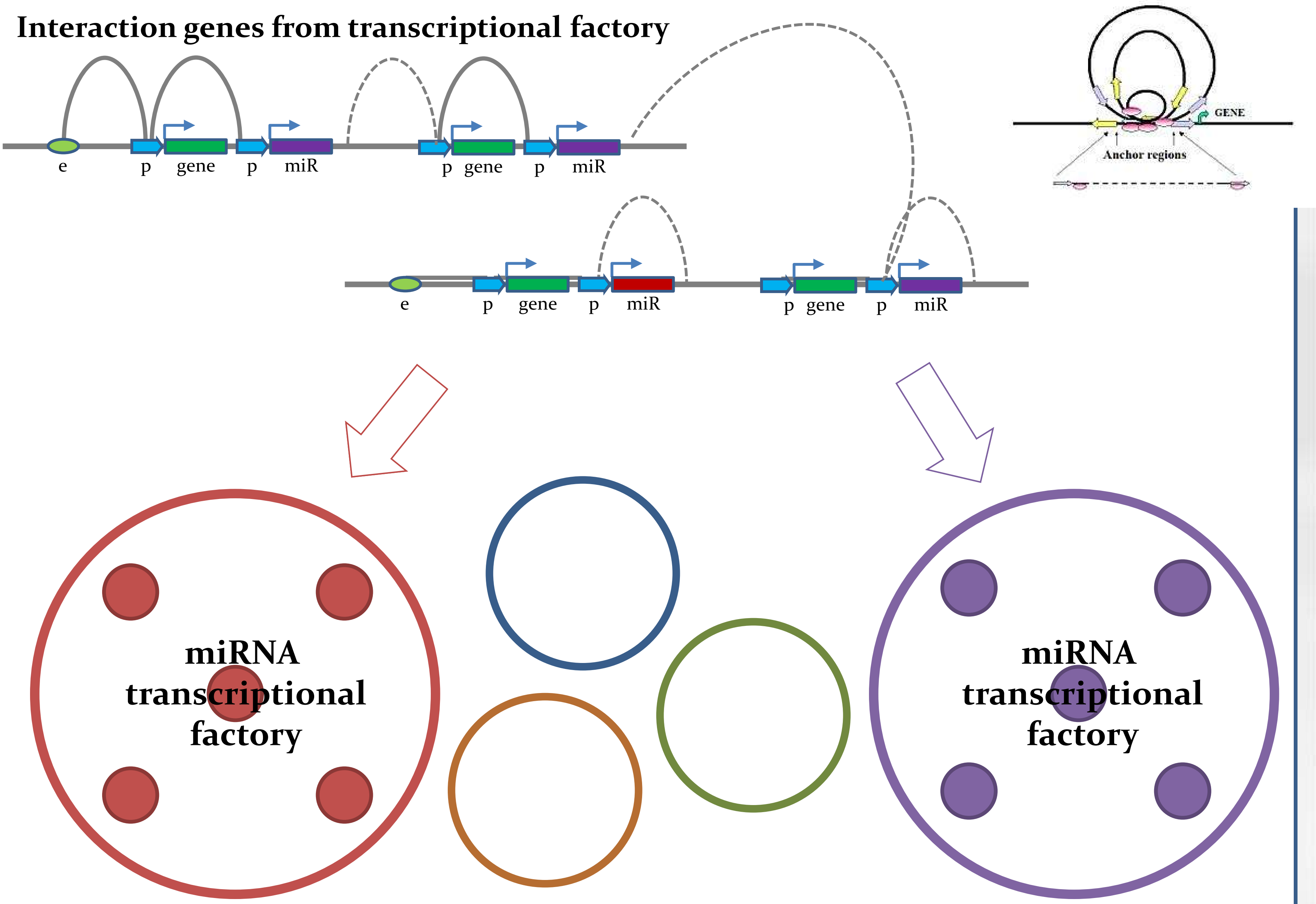


Dynamic nature of miRNA biogenesis

Plant Physiology, 2011



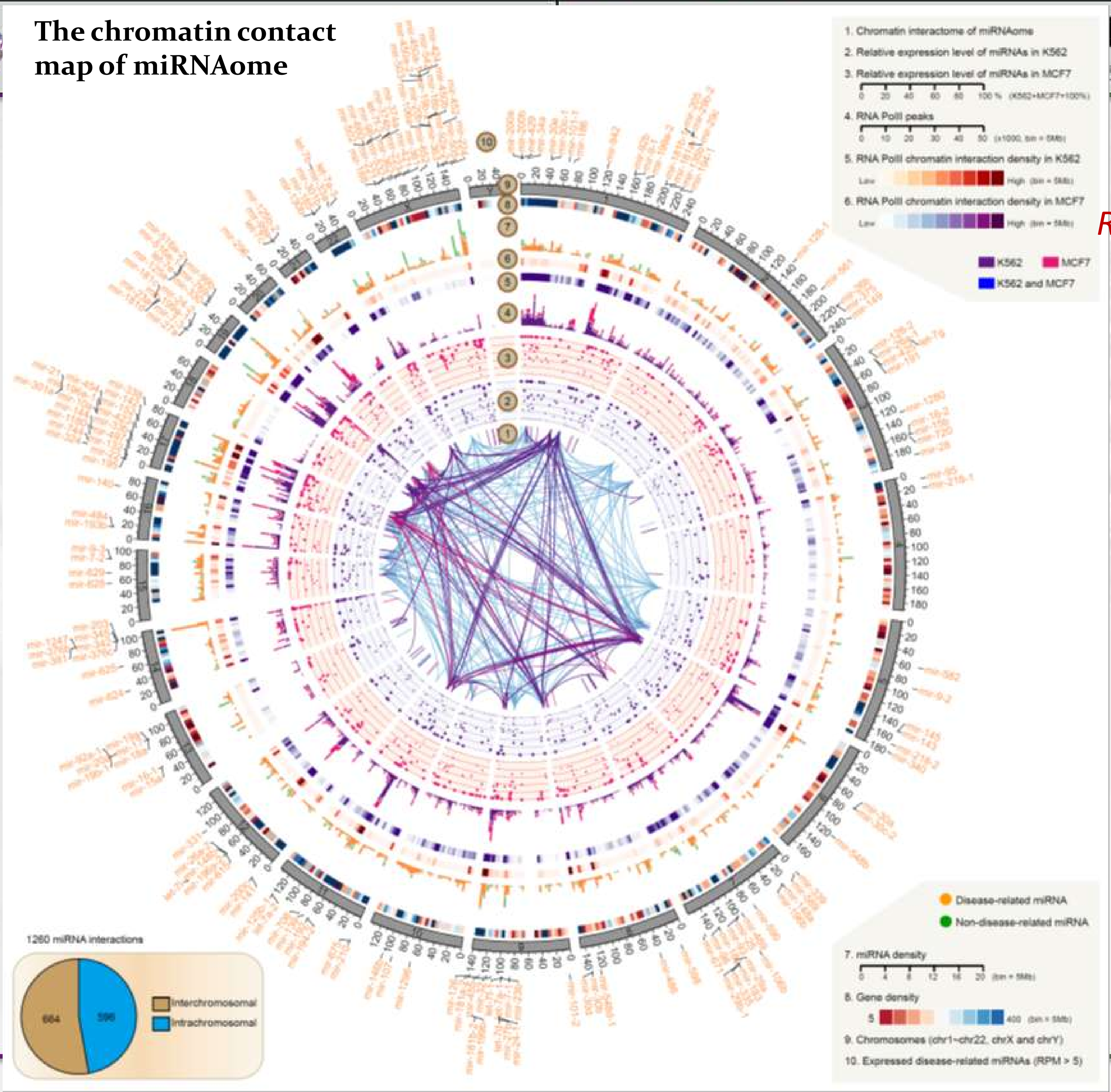
Interaction genes from transcriptional factory



miRNA transcriptional factory

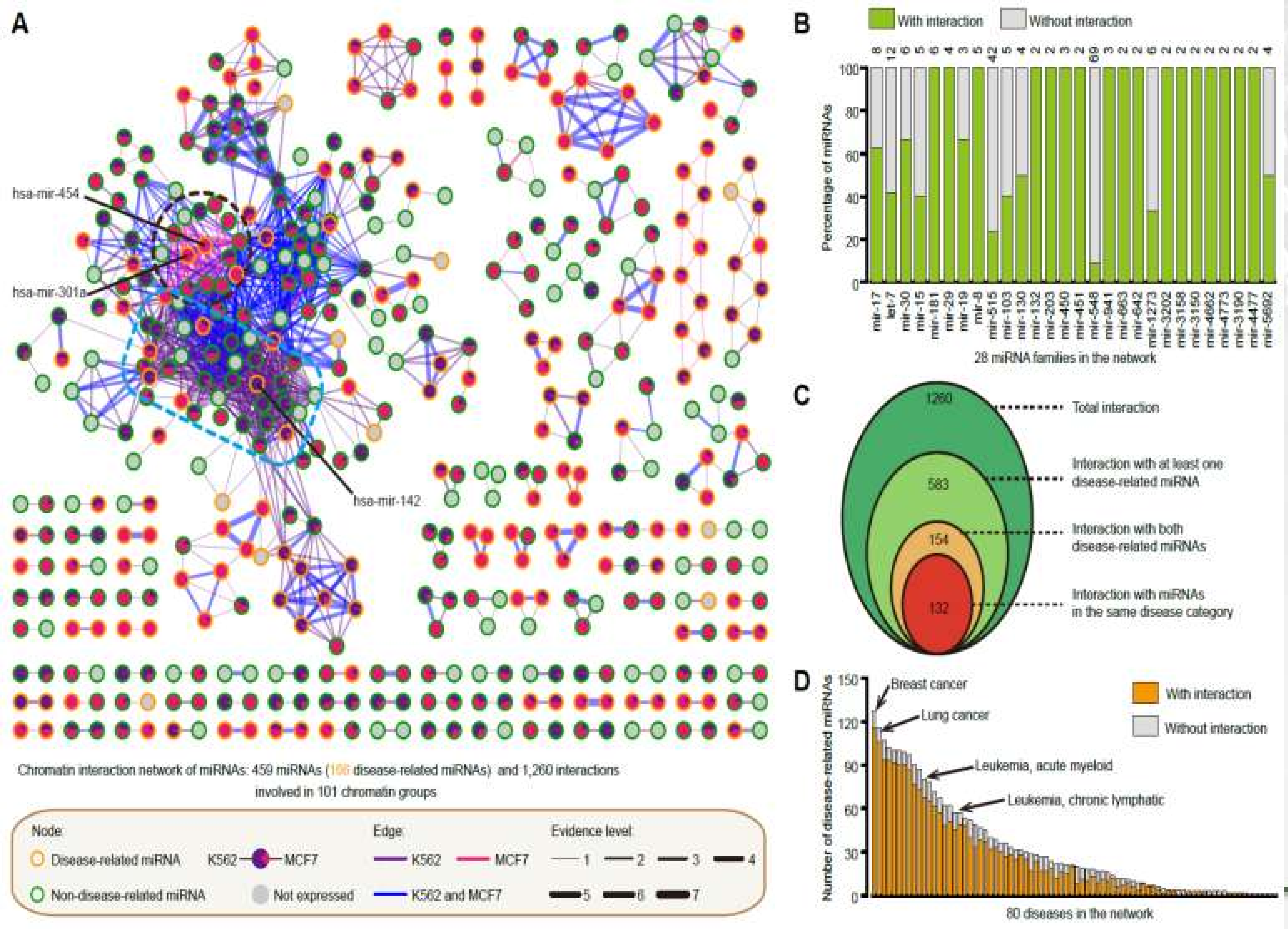
miRNA transcriptional factory

The chromatin contact map of miRNAome



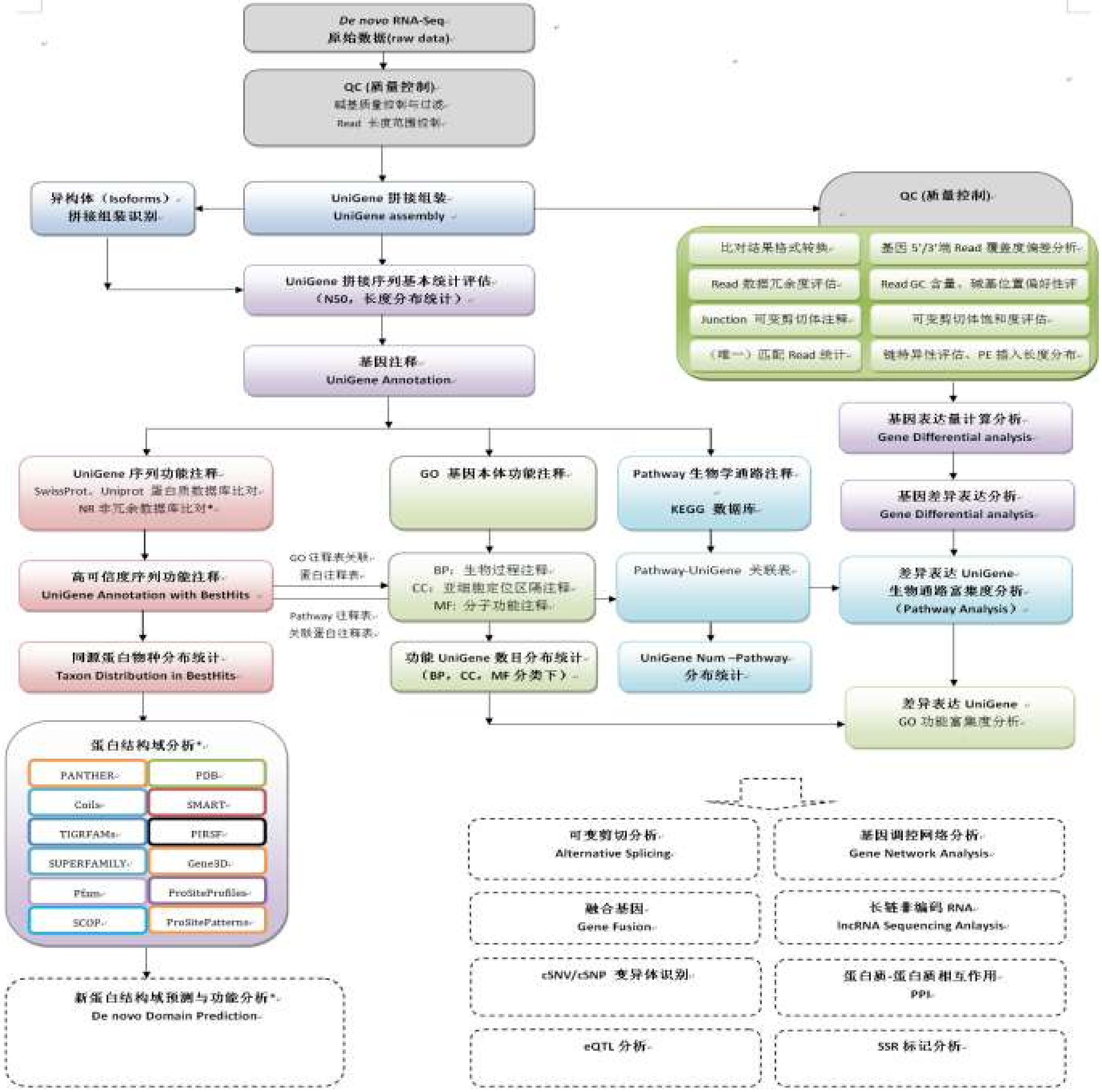
RNA , 2014

chromatin interactome networks



理论课内容

- ✓转录组学介绍
- ✓基因表达数据分析
  - 测定技术
  - 差异基因
  - 功能分析
- ✓几个实例
- ✓非编码RNA分析



De novo RNA-Seq
原始数据(raw data)

QC (质量控制)
碱基质量控制与过滤
Read 长度范围控制

异构体（Isoforms）
拼接组装识别

UniGene 拼接组装
UniGene assembly

UniGene 拼接序列基本统计评估
（N50，长度分布统计）

基因注释
UniGene Annotation

QC (质量控制)

比对结果格式转换

基因 5'/3'端 Read 覆盖度偏差分析

Read 数据冗余度评估

Read GC 含量，碱基位置偏好性评

Junction 可变剪切体注释

可变剪切体饱和度评估

（唯一）匹配 Read 统计

链特异性评估、PE 插入长度分布

基因表达量计算分析
Gene Differential analysis

基因差异表达分析
Gene Differential analysis

UniGene 序列功能注释
SwissProt、Uniprot 蛋白质数据库比对
NR 非冗余数据库比对*

GO 基因本体功能注释

Pathway 生物学通路注释
KEGG 数据库

高可信度序列功能注释
UniGene Annotation with BestHits

GO 注释表关联
蛋白注释表

Pathway 注释表
关联蛋白注释表

BP：生物过程注释
CC：亚细胞定位区隔注释
MF：分子功能注释

Pathway-UniGene 关联表

差异表达 UniGene
生物通路富集度分析
（Pathway Analysis）

同源蛋白物种分布统计
Taxon Distribution in BestHits

功能 UniGene 数目分布统计
（BP，CC，MF 分类下）

UniGene Num –Pathway
分布统计

差异表达 UniGene
GO 功能富集度分析

蛋白结构域分析*

PANTHER

PDB

Coils

SMART

TIGRFAMs

PIRSF

SUPERFAMILY

Gene3D

Pfam

ProSiteProfiles

SCOP

ProSitePatterns

新蛋白结构域预测与功能分析*
De novo Domain Prediction

可变剪切分析
Alternative Splicing

基因调控网络分析
Gene Network Analysis

融合基因
Gene Fusion

长链非编码 RNA
lncRNA Sequencing Anlaysis

cSNV/cSNP 变异体识别

蛋白质-蛋白质相互作用
PPI

eQTL 分析

SSR 标记分析