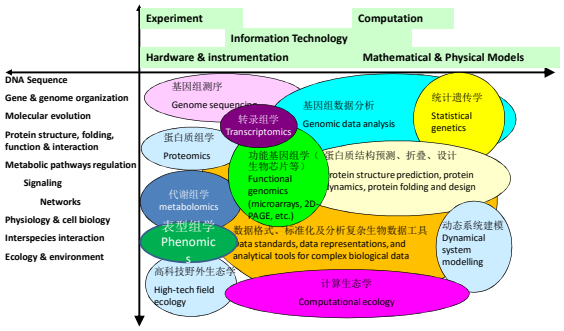
Bioinformatics
生物信息学

2. NGS & Transcriptomics

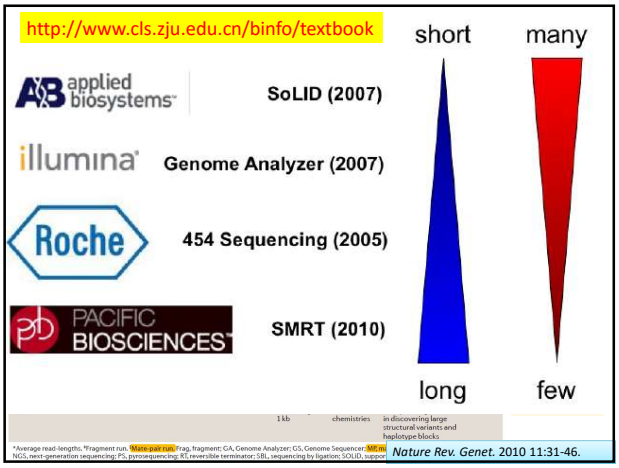
陈铭（mchen@zju.edu.cn）
2015年9月21日

Bioinformatics Review

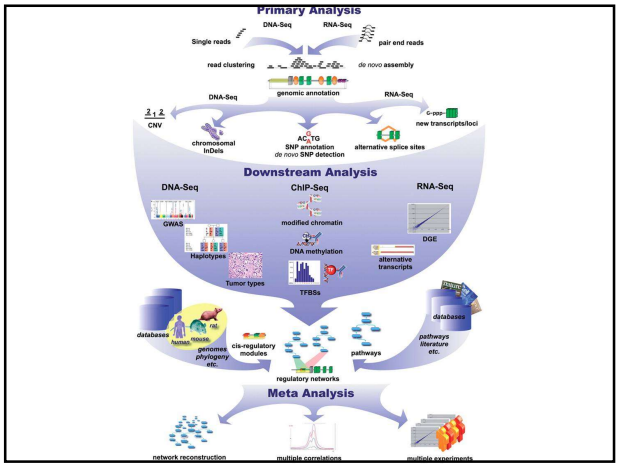


陈铭：整合生物信息学（2006）

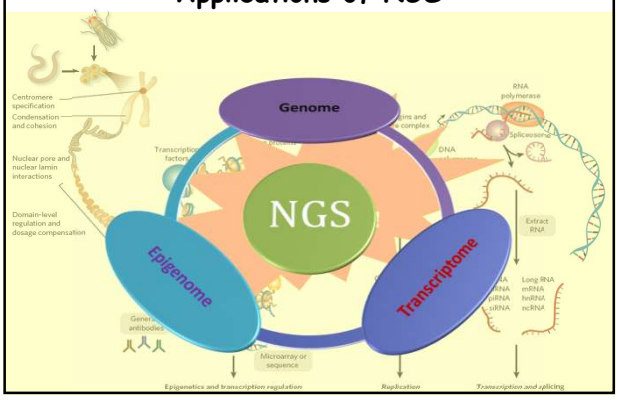
http://www.cls.zju.edu.cn/binfo/textbook



Nature Rev. Genet. 2010 11:31-46.

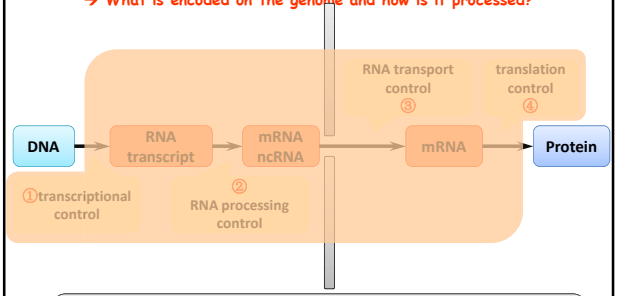


Applications of NGS



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.

6

Transcriptome

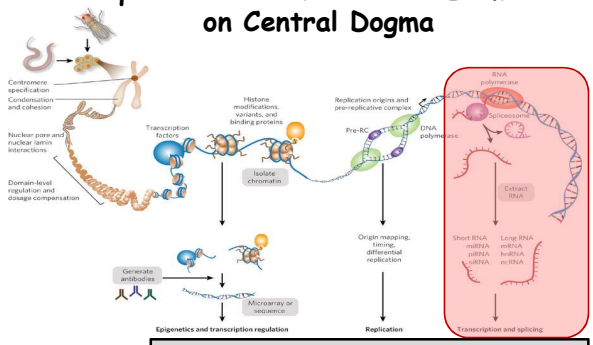
- Transcriptome
  - Protein-coding RNAs (mRNAs)
  - Non-coding RNAs (ncRNAs)
    - Structural RNAs: rRNAs/tRNAs
    - Catalytic RNAs: ribozymes (e.g. RNase P and snRNP)
    - Long ncRNAs
    - Small ncRNAs
      - microRNAs
      - siRNAs
        - Exogenous
          - Transgenic induction
        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

Gene expression data analysis

- Gene expression biology
- Measuring gene expression level
- Identifying differentially expressed genes
- Advanced analysis
- Hickory gene expression data analysis
- Training topics

Part 1: Gene expression biology

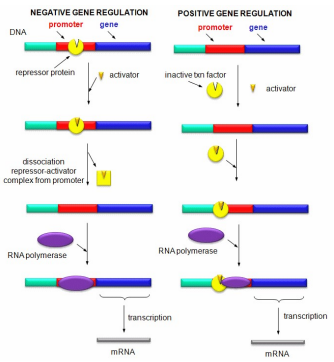
Sequence-based Functional Elements on Central Dogma



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.

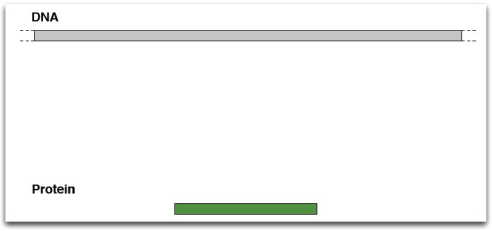
How can gene expression be regulated at the transcriptional level?



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

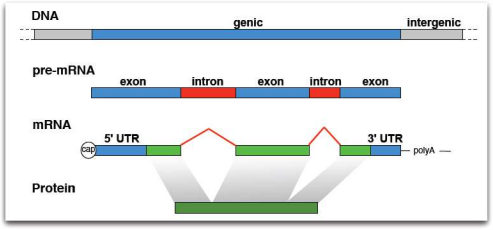
- physiological status (nutrition, environment)
- sex and age
- various tissues and cell types
- response to stimuli (drugs, signals, toxins)
- health and disease

Protein-coding gene

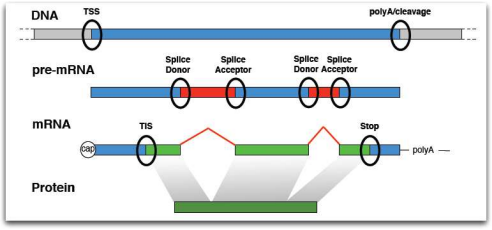


✓Gene: a functional piece of DNA sequence

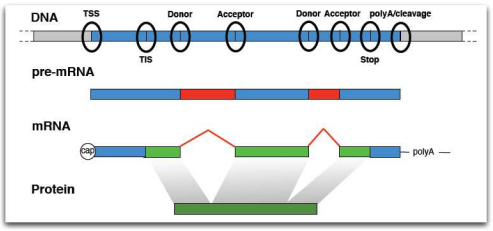
Computational Gene Finding



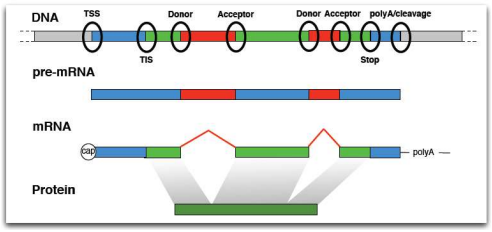
Predict signals used during processing



Predict signals used during processing



Computational Gene Finding



✓**Predict the correct corresponding label sequence with labels "intergenic", "exon", "intron", "5' UTR", etc**

Part 2: 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.

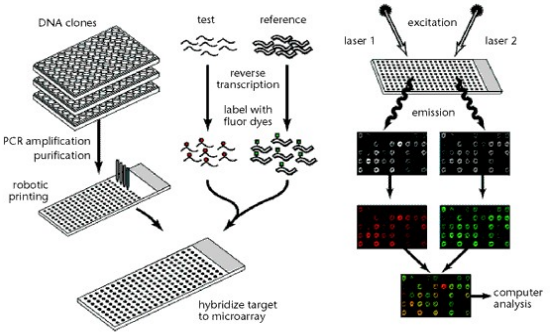
Advantages and disadvantages

- ✓ **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

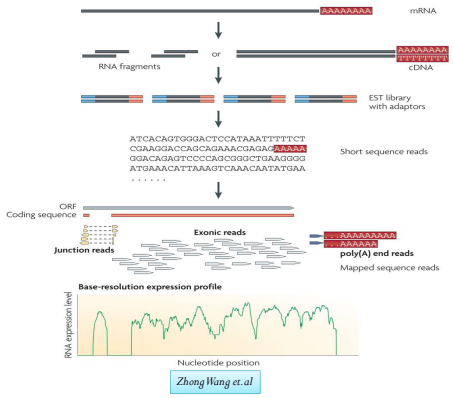
RNA-Seq: Advantages

➢RNA-Seq v.s. other transcriptomics methods

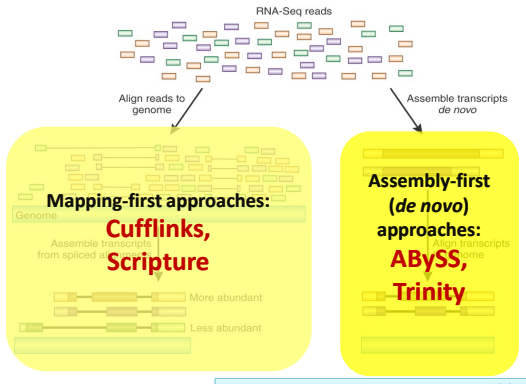
Technology	Tiling microarray	cDNA or EST sequencing	RNA-Seq
Technology specifications			
Principle	Hybridization	Sanger sequencing	High-throughput sequencing
Resolution	From several to 100 bp	Single base	Single base
Throughput	High	Low	High
Reliance on genomic sequence	Yes	No	In some cases
Background 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)



RNA-seq workflow (2)



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

Gene expression level measurement for RNA-seq

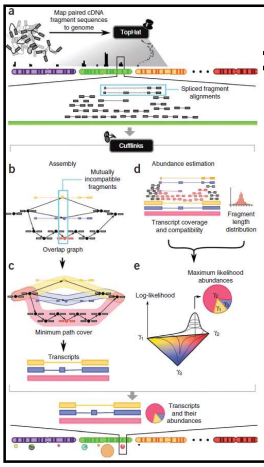
✓RPKM : Reads per kilobase per million mapped reads.

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

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

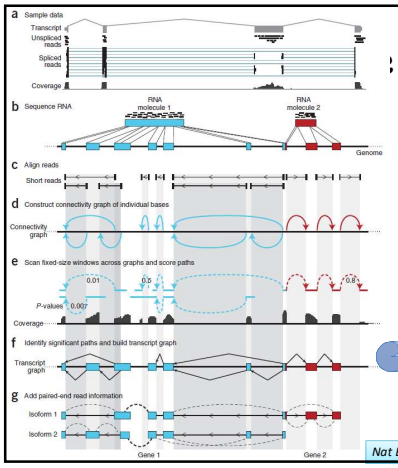
✓FPKM : Fragments Per Kilobase of exon per Million fragments mapped (for paired-end sequencing).

Cufflinks



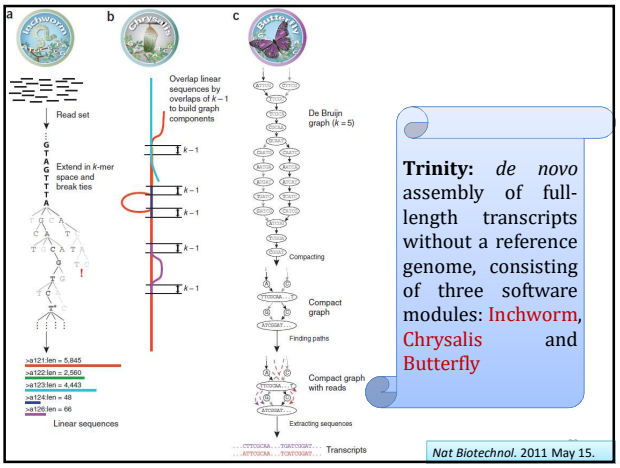
Cufflinks uses a rigorous mathematical model to identify the complete set of alternatively regulated transcripts at each locus and to assign coverage to each transcript.

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



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

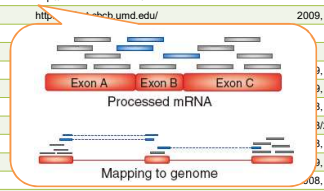
Nat Biotechnol. 2010 May;28(5):503-10.



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

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...hcb.umd.edu/	2009, Bioinformatics
Novoalign		
Mosaik		
Bowtie		), Genome Biology
BWA		), Bioinformatics
MAQ		), Genome Research
SOAP/SOAP2		9/2009, Bioinformatics
ZOOM		), Bioinformatics
PerM		), Bioinformatics
BWT-SW		08, 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



Part 3: 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)**

T-test

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

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

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

$$S^2 = \frac{1}{n_1+n_2-2}\left(\sum_{i=1}^{n_1}(X_i-\overline{X_1})^2+\sum_{i=1}^{n_1}(Y_i-\overline{Y_2})^2\right)$$

$$T(X,Y)=\frac{\bar{X}-\bar{Y}}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

- 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_g be 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.

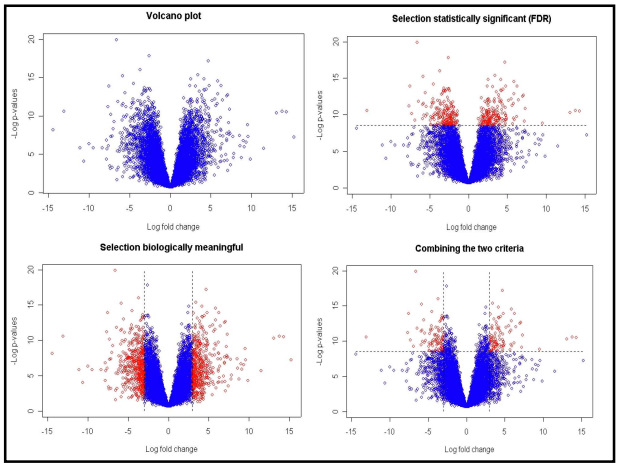
b) False-discovery-rate control: An alternative approach to multiple testing considers the false-discovery rate (FDR), which is the proportion of false positives among all of the genes initially identified as being differentially expressed - that is, among all the rejected null hypotheses.

➢ 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.

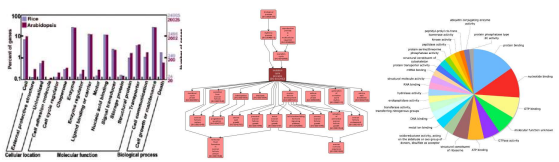
  – Biological meaning
- ✓ Differentially expressed gene selection: Need combination of these two.



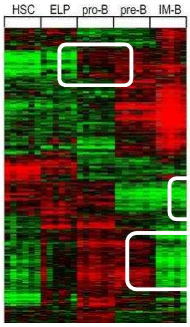
Part 4: Advanced analysis

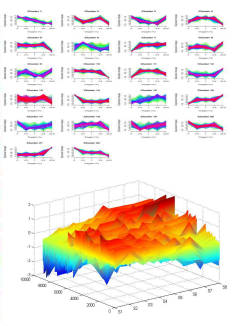
GO analysis

- ✓ **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… …



Clustering gene expression data

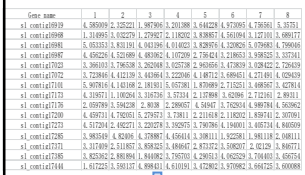




Algorithms:
a) K-means
b) Hierarchical clustering
c) K-median
d) Bi-clustering
Tools and software:
a) R language,
b) Clustal,
c) Mev.

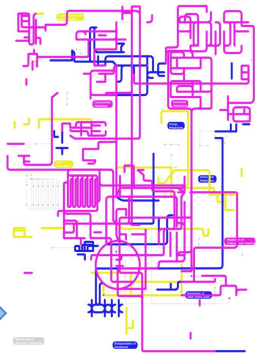
If two genes are related (have similar functions or are co-regulated), their expression profiles should be similar (e.g. low Euclidean distance or high correlation).

Pathway mapping and analysis

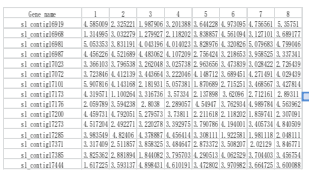


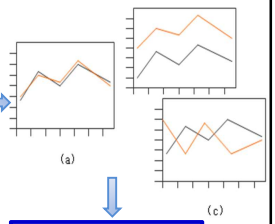
Identify up-/down-regulated genes

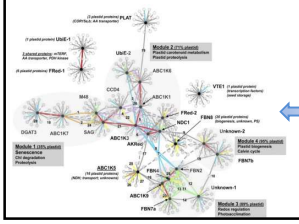
KO ID mapping KEGG



Co-expression network reconstruction



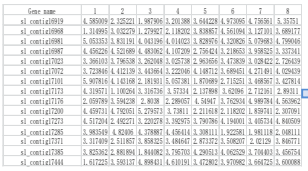




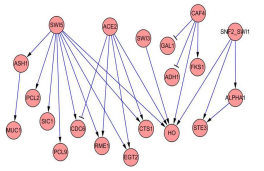
- ✓ **Algorithms:**
  a) PCC
  b) Weighted PCC
  c) Multiple rank (MR)
- ✓ **Visualization software:** Cytoscape

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

Gene regulatory network reconstruction



Gene expression data Discretization
- ✓ Equal Width Discretization
- ✓ Equal Frequency Discretization
- ✓ Kmeans Discretization
- ✓ Column Kmeans Discretization
- ✓ Bikeans Discretization



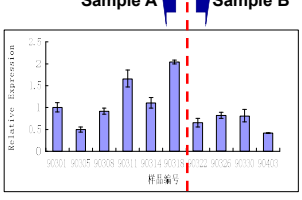
Gene regulatory network reconstruction
- ✓ Greedy search
- ✓ K2
- ✓ Aracne
- ✓ Matlab
- ✓ … …

Part 4: gene expression data analysis: Examples

Hickory gene expression data analysis

Materials and Methods

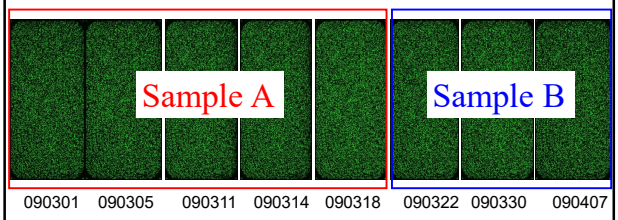
➢ **454 sequecing**



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

Results

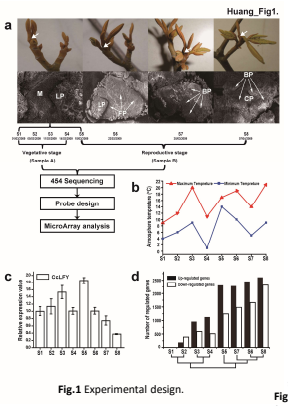


Fig.1 Experimental design.

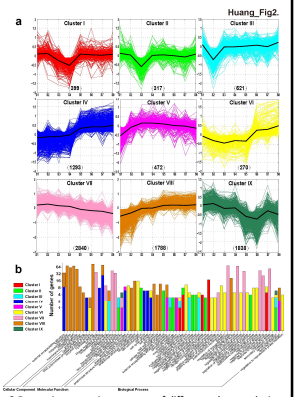
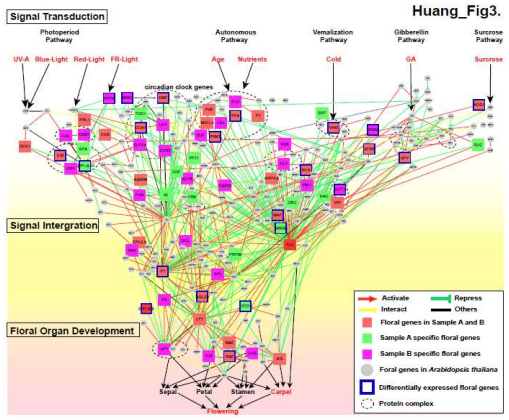


Fig. 2 Dynamic expression pattern of different clusters during flower development and GO function enrichment analysis.



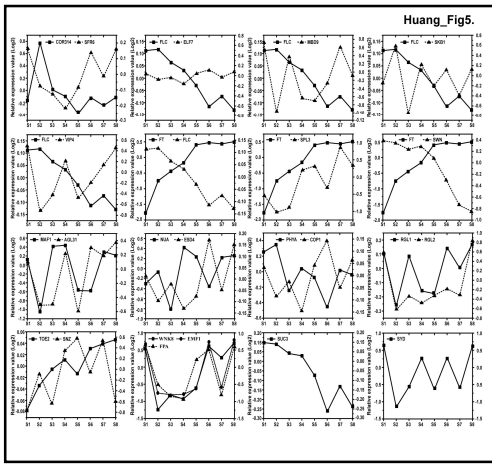
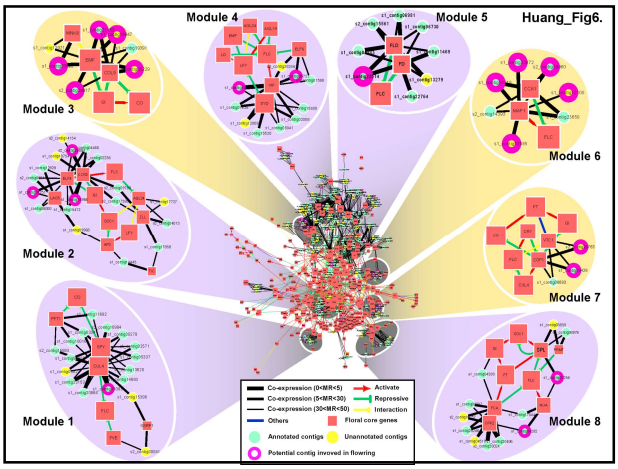


Fig. 5 Transcriptional regulation of differentially expressed genes in floral development in hickory.



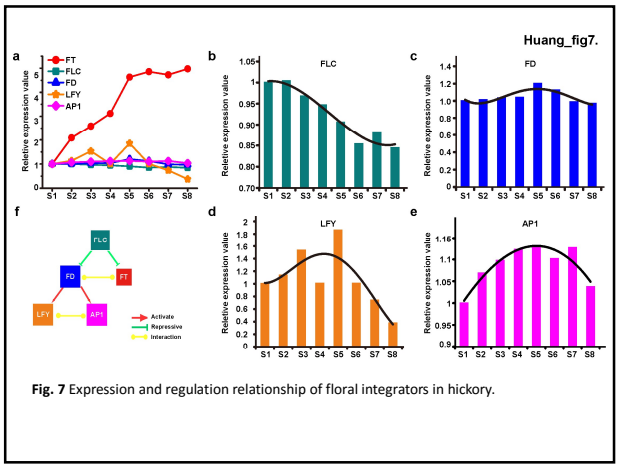
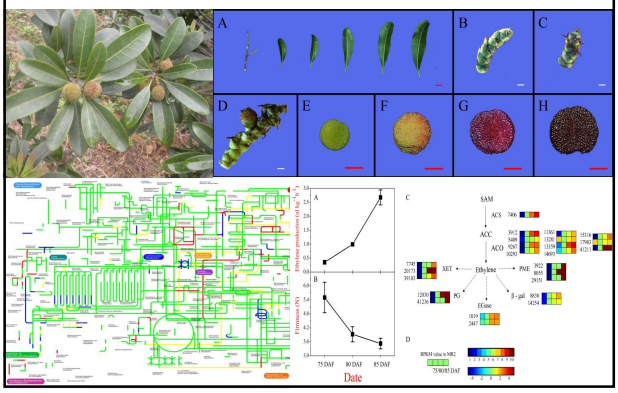


Fig. 7 Expression and regulation relationship of floral integrators in hickory.

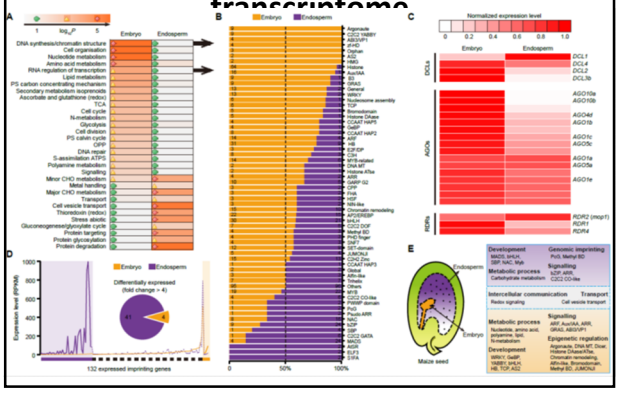
Chinese bayberry

BMC Genomics, 2012, 13: 19. (IF=4.073)

Highly accessed



Dynamic progression map of seed transcriptome



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:

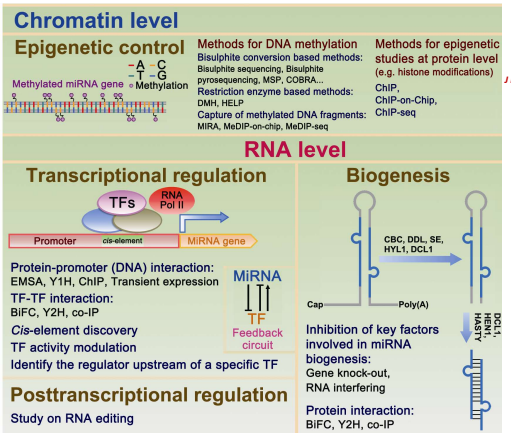
microarray, next-generation sequencing (NGS)

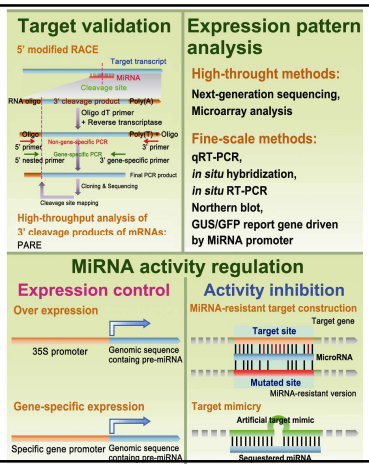
High-throughput, expensive
cross-hybridization & limited sensitivity (microarray)

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

small RNAs: Move from microarray to Next-Generation Sequencing (NGS)

Upstream analysis





Research work of my lab

- 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

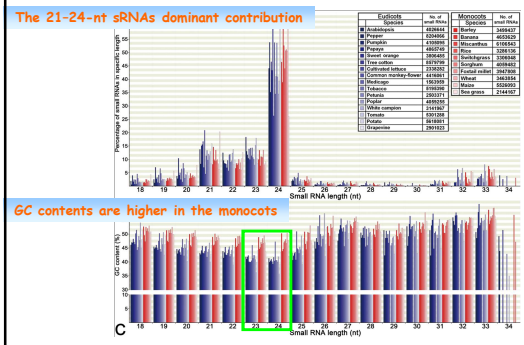
Monocots (10)

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）

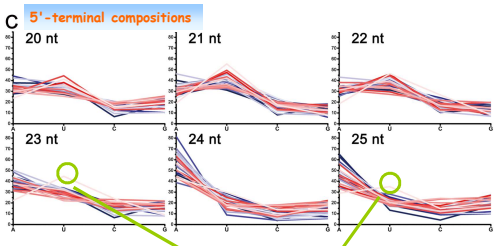
The 21-24-nt sRNAs dominant contribution



GC contents are higher in the monocots

Small RNA序列特征分析（2）

C 5'-terminal compositions

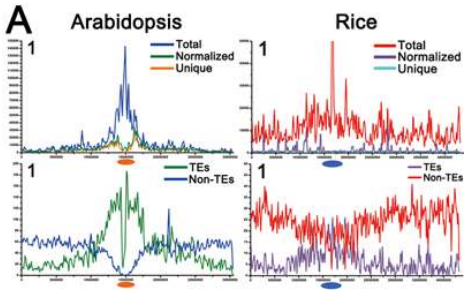


The 5'-terminal composition patterns are similar between the eudicots and the monocots.

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双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；为检测其转座活性、维持染色体结构序列上的稳定性，大量内源siRNAs用于抑制转座子转座。

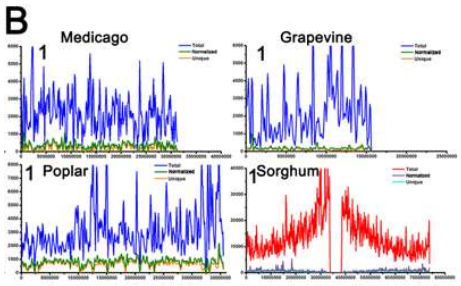
Chromosome-wide distribution patterns



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，短峰位置十分明显。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蛋白编码位置的大致界定参考依据。

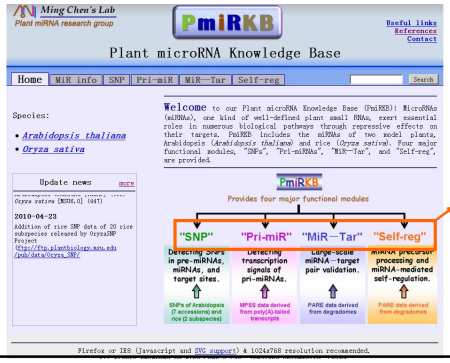
Small RNAs derived from gene models

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

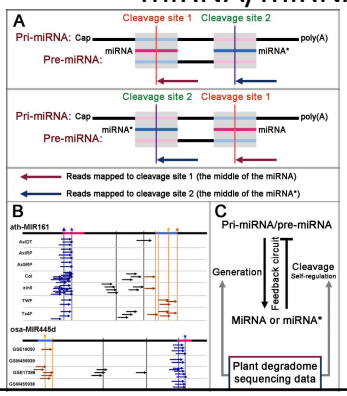
Nucleic Acids Res, 2011



包含4个主要功能模块

miRNA/miRNA*

J Exp Bot, 2010

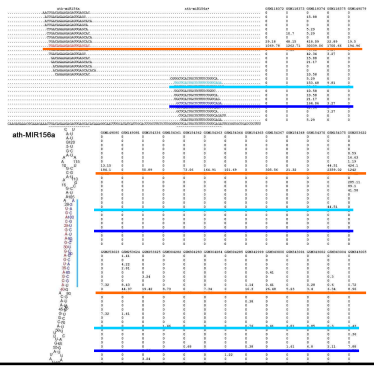


基于degradome测序数据分析可阐释：

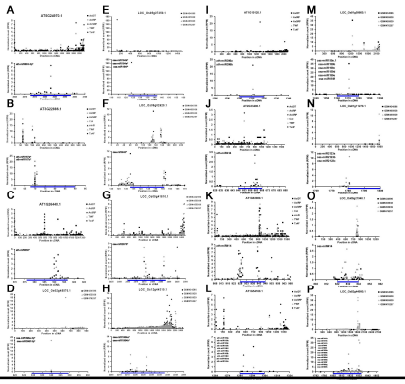
- 从miRNA前体到miRNA成熟体的加工过程
- miRNAs/miRNA*s介导的自调控过程。

miRNA/miRNA* detection

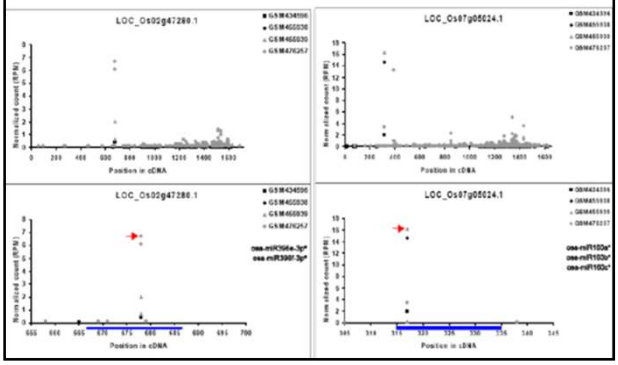
RNA Biology, 2011



miRNAs/miRNA*s targets

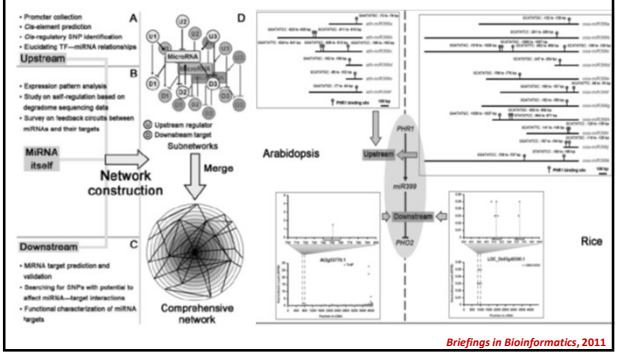


Target determination



miRNA regulatory network

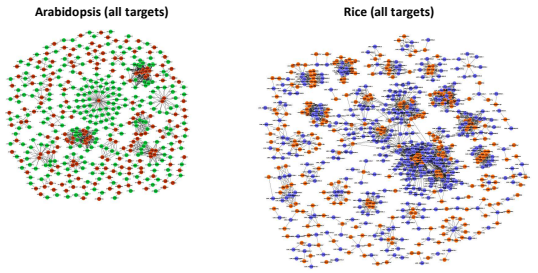
miRNA介导的基因调控网络构建思路　　对已有*PHR1—miR399—PHO2*调控通路进行验证性重建



Briefings in Bioinformatics, 2011

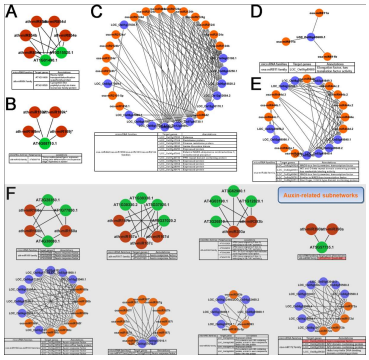
miRNA/miRNA* regulatory network

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



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

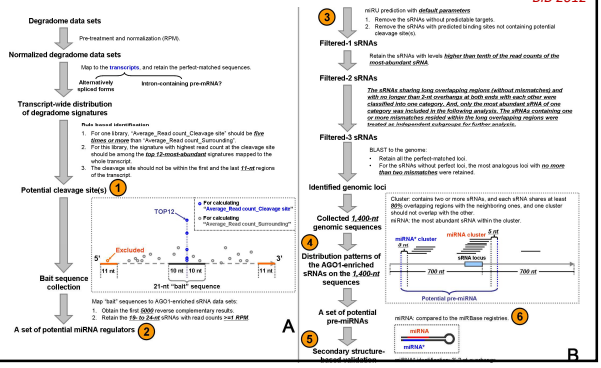


发现一些可能具有重要生物学功能（参与重金属胁迫应答、植物抗病相关）的子网络

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）

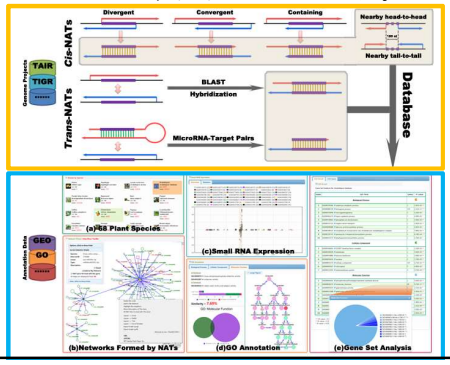
A reverse framework

BiB 2012



NATs

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



NATs Generated Small RNAs

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

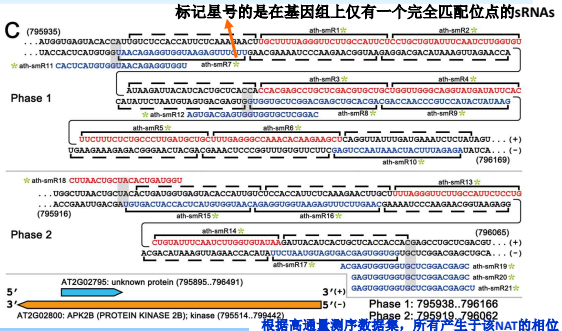
Species	*Cis*-NATs Overlap[d] [total/unique][c]	All[e] [total/unique][c]	Average score[f] [total/unique][c]	*P*-value[g] [total/unique][c]
Arabidopsis	38.89/7.11	10.62/5.63	3.10/1.95	<0.0001/0.0448
Poplar	8.42/11.19	5.42/2.68	2.61/5.26	0.4525/0.1548
Papaya	7.05/3.85	4.66/2.33	1.99/1.97	0.0094/0.0011
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

Species	*Trans*-NATs 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

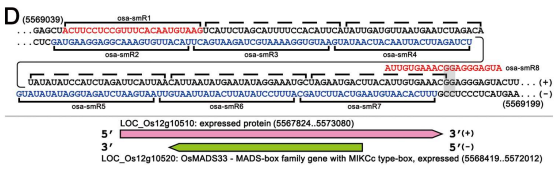


Exclusively cloned from floral organs

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。

Organ specific - rice

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

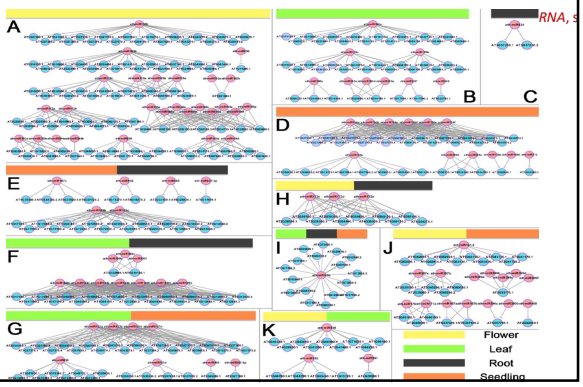


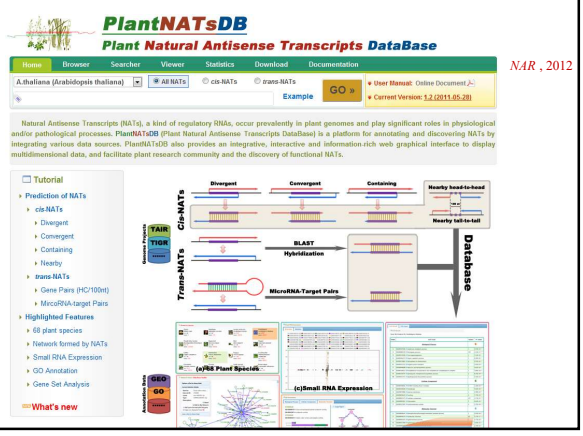
Exclusively cloned from grains

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*





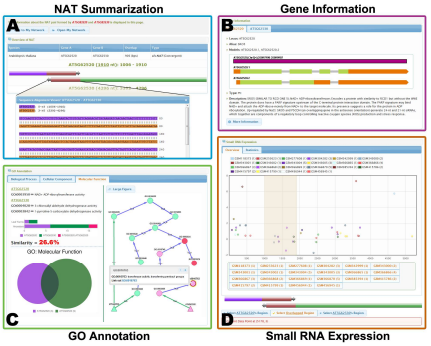
NAR , 2012

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	aco	*Allium cepa*	*NA*	4063 (10)	*NA*	5 (*NA*)	5
2	aqc	*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	eea	*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

An example



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%)	~1.8%
		Others[k] (Total[c]: 7.05%; Unique[d]: 5.68%)	
Rice	Intergenic loci (Total[c]: 80.48%; Unique[d]: 79.30%)	-	
	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%)	~6.6%
		Others[k] (Total[c]: 3.47%; Unique[d]: 5.29%)	
	Other loci[f] (Total[c]: 0.38%; Unique[d]: 0.35%)	-	

Intronic small RNAs

Identification of intronic long hairpins

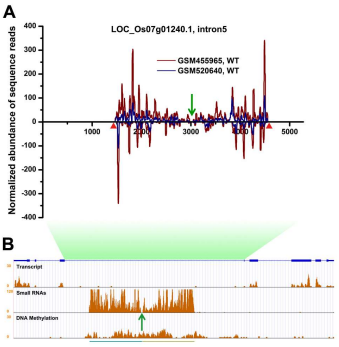
RNA, 2011

Table 1. A list of 21 *IR*-intron 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				62 - 4873	87	1.188
LOC_Os03g24339.1|intron_2	9177	1432				13 - 8272	96	1.161
LOC_Os03g13614.1|intron_1	5284	1943				45 - 3627	92	1.086
LOC_Os09g17730.1|intron_1	4168	727				13 - 2373	91	0.759
LOC_Os02g35039.1|intron_8	5898	2107				21 - 4096	97	0.536
LOC_Os05g15370.1|intron_1	3641	911				92 - 3006	81	0.328
LOC_Os03g51270.1|intron_3	1224	341				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

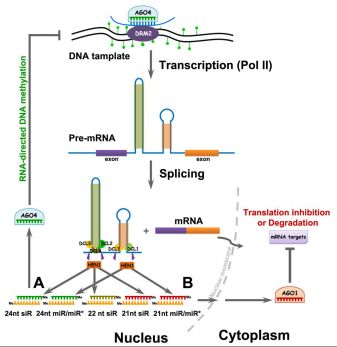
sirtrons

An example of sirtrons

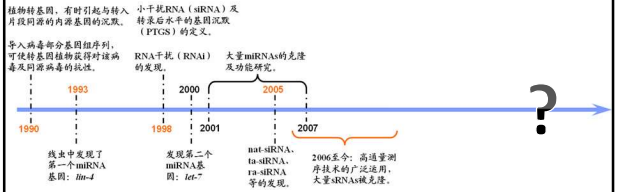


A proposed self-regulation model

RNA , 2012

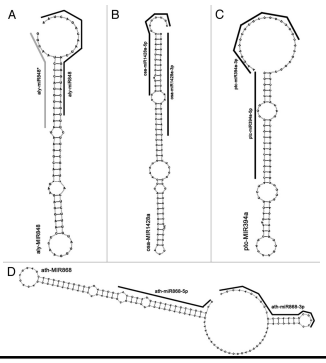


Prospective



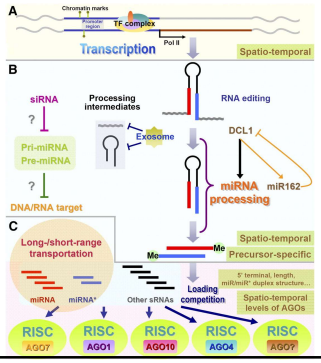
Are all the miRBase-registered microRNAs true?

RNA Biology, 2012

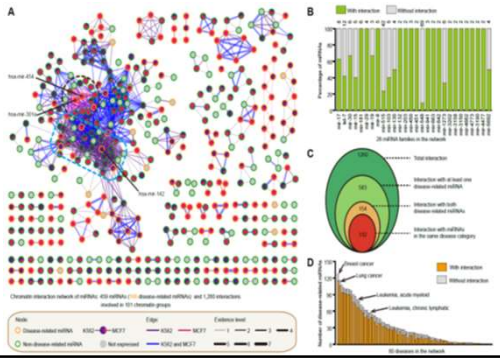


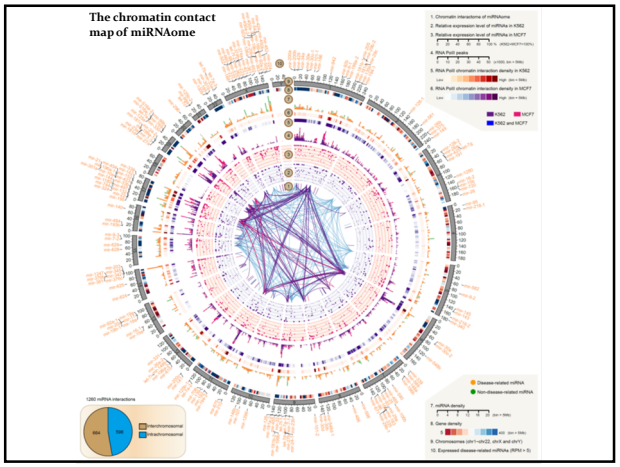
Dynamic nature of miRNA biogenesis

Plant Physiology, 2011

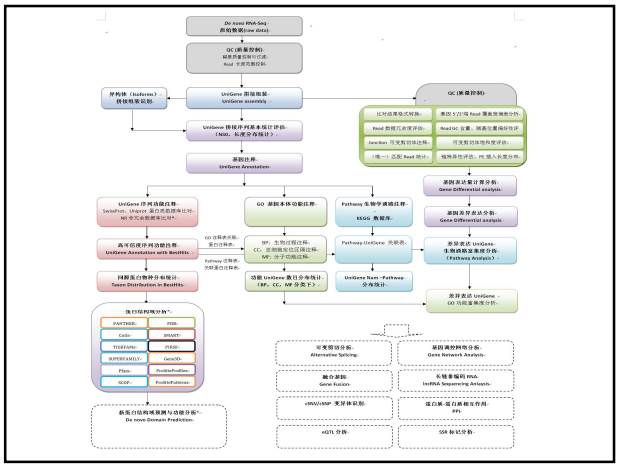


chromatin interactome networks





Part 4: Practice



Training Topics

One: Using R language find differentially expressed genes.

Two: Do the GO analysis for identified differentially expressed genes.

Three: Do the pathway analysis for the differentially expressed genes.

http://www.cls.zju.edu.cn/binfo/links.htm