

《实用生物信息技术》

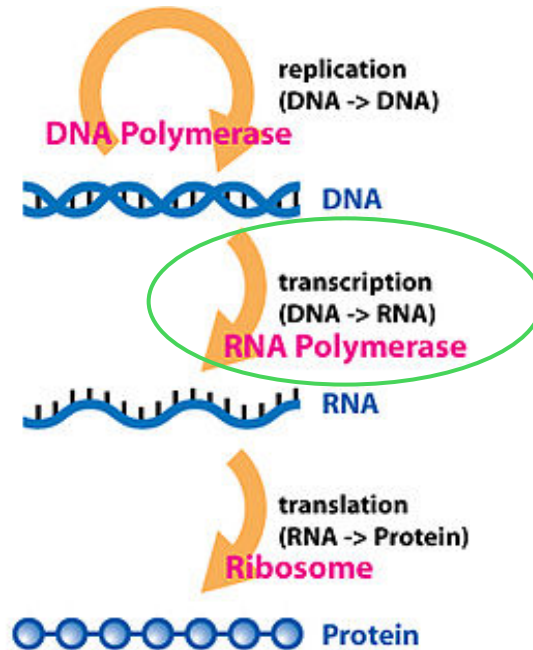


植物转录调控数据与分析平台简介

靳进朴 田丰 杨德昌

2024年10月14日

转录调控是遗传信息表达调控的关键

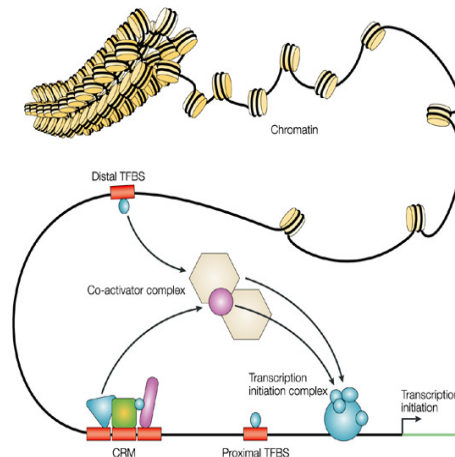


DNA是遗传信息的载体，
转录则是遗传信息表达的第一步。

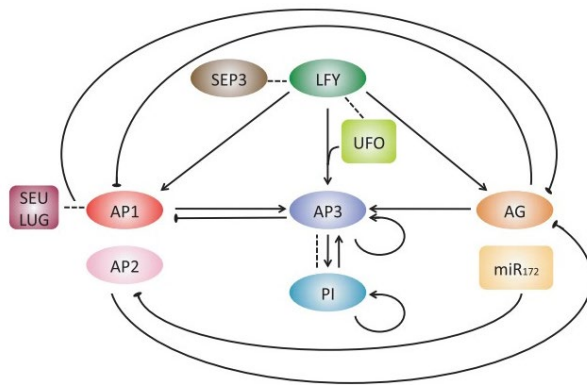
分子生物学的中心法则

转录因子及其重要性

- 转录因子是能序列特异性结合DNA并激活或抑制下游靶基因转录的蛋白
- 通过结合顺式作用元件，转录因子时空特异性地调控靶基因转录，在植物发育和应对胁迫中发挥关键作用
- 转录因子及转录调控的演化在植物形态改变和环境适应中发挥着极为重要的作用

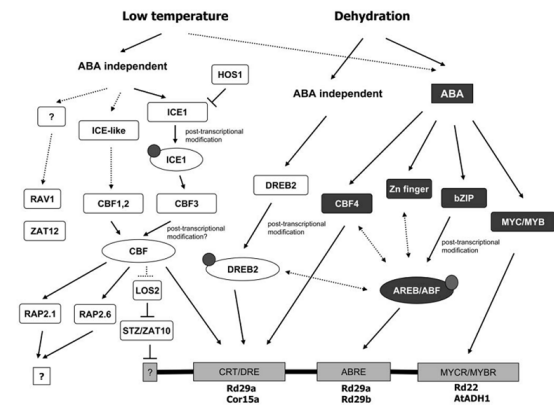


转录调控网络是理解植物发育与抗逆的基础



Irish, 2010

花分生组织分化的转录调控通路

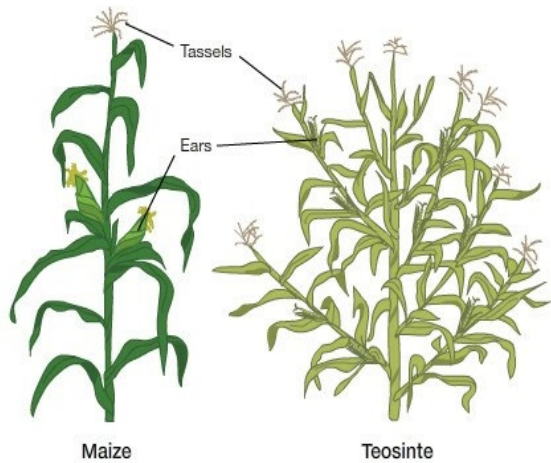


Zhang, *et al.*, 2004

低温和干旱引起的级联转录调控网络

转录调控的演变与作物驯化

分蘖



tb1 (TCP family)

Doebley, *et al.*, 1997
Wang, *et al.*, 1999

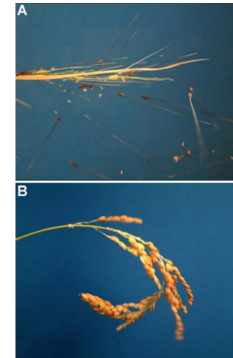
果壳包被



tga1 (SBP family)

Wang, H. *et al.*, *Nature*, 2005

落粒



sh4

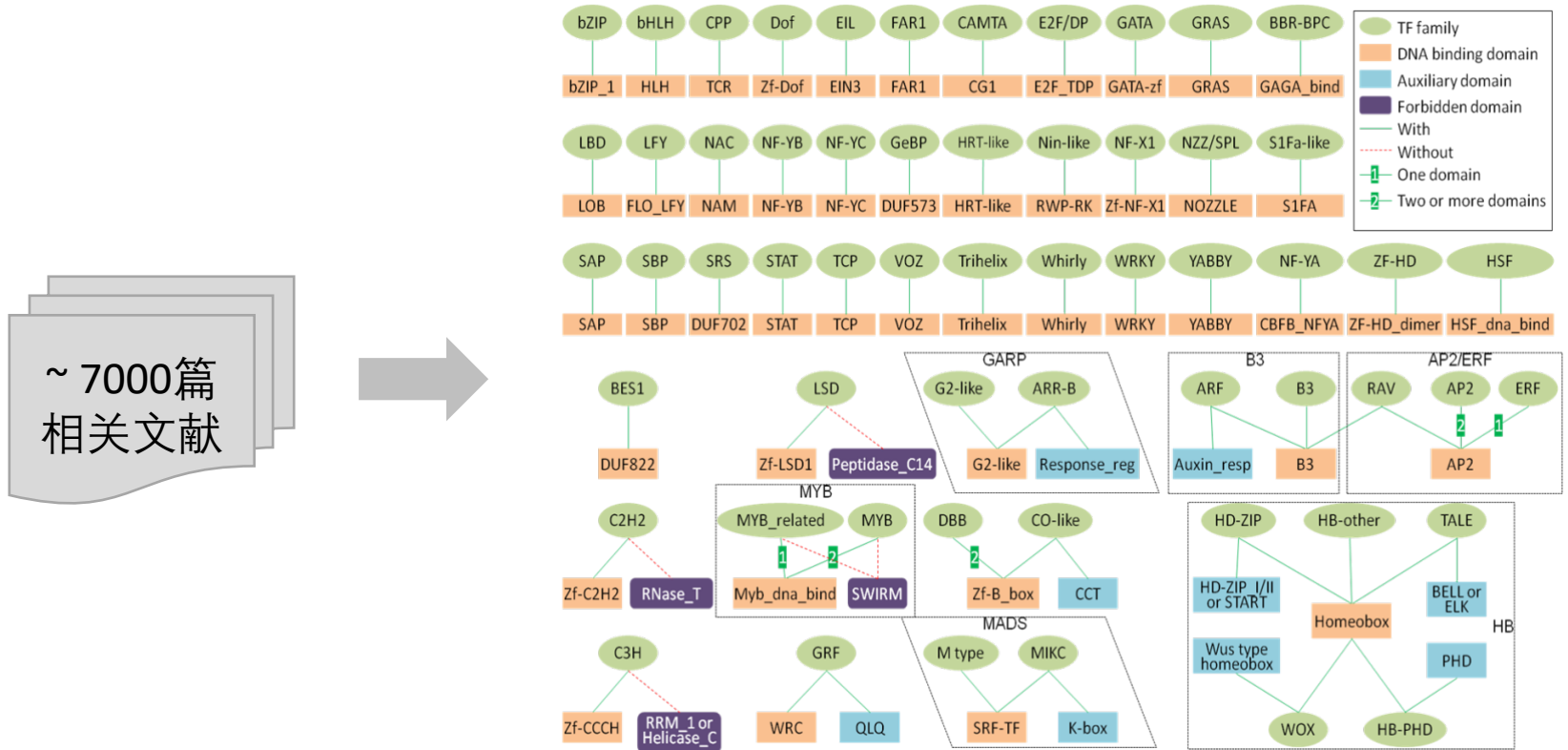
Li, C. *et al.*, *Science*, 2006

系统鉴定植物基因组中的转录因子、调控元件及二者之间的相互作用，是理解转录调控系统如何工作和演变的基础。

主要内容

- 植物转录因子分类、识别与注释 (PlantTFDB)
- 植物转录调控图谱的绘制
- 植物转录调控系统的架构原则与演变特征

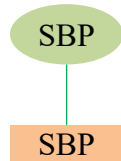
植物转录因子家族分类规则



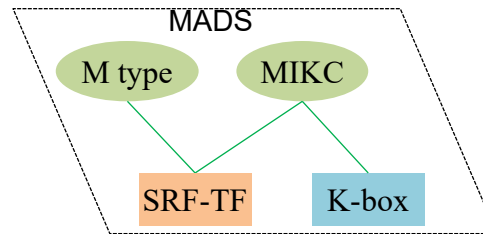
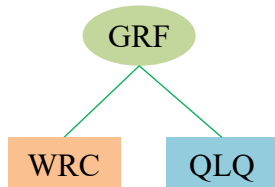
植物转录因子家族划分为58个家族

植物转录因子家族分类规则

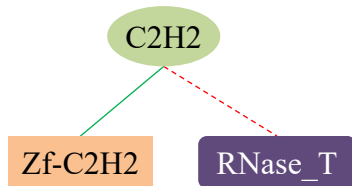
➤ 简单的分类模式



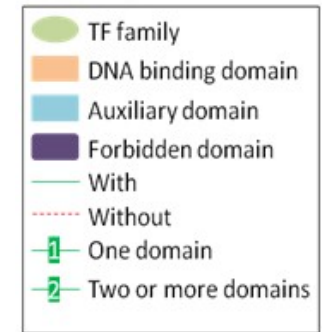
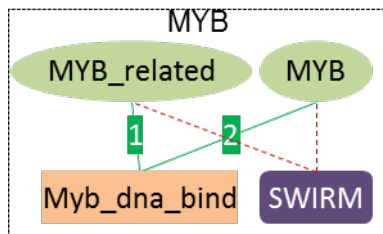
➤ 使用辅助结构域(其它类型的特征结构域)



➤ 禁止出现的结构域



➤ DNA结合结构域的数目



覆盖绿色植物各大分支的转录因子全谱

Lineage	Species	Gene	TF (%)	Family
Chlorophytae	16	10 828	129 (1.19)	36
Charophyta	1	16 215	273 (1.68)	45
Marchantiophyta	1	19 287	398 (2.06)	51
Bryophyta	2	29 932	1 091 (3.64)	55
Lycopodiophyta	1	22 285	665 (2.98)	54
Coniferophyta	2	40 634	1 511 (3.72)	55
Basal Magnoliophyta	1	26 846	900 (3.35)	58
Monocots	38	39 641	1 876 (4.73)	58
Eudicots	95	38 426	1 959 (5.10)	58

详尽的注释——转录因子个体水平

Basic Information

TF ID

Common Name

Organism

Taxonomic ID

Taxonomic Lineage

Family

Protein Properties

Description

Gene Model

AT1G53160.1

FTM6, [FTL](#)

[Arabidopsis thaliana](#)

3702

Viridiplantae; Streptophyta; Streptophyta; Embryophyta; Tracheophyta; Euphyllophyta; Spermatophyta; Magnoliophyta; eudicotyledons; core eudicotyledons; rosids; malvids; Brassicales; Brassicaceae; Camelineae; Arabidopsis

SBP

Length: 174aa MW: 20119.6 Da PI: 10.1381

squamosa promoter binding protein-like 4

Gene Model ID

Type

Source

Coding Sequence

AT1G53160.1

genome

EMBL

view CDS

Signature Domain

0

50

100

150

174

1

174

SBP

No.

Domain

Score

E-value

Start

End

HMM Start

HMM End

1

SBP

134.9

2.6e-42

54

129

2

77

-SSIT--TT-10000TT-10000-S-KETETETET-ETTSSEETTT--SS-S-STTT--S--CS

SFP 2 CysGqEadLseaykerrkdcVekhsqyvgvqgqgqEgqgqfBaLseDsekrSrrrLakhaerrddq TT

AT1G53160.1 54 CQVETATREALTGGAYCTVYVAGATGTCGLGQFQDCSTGSLFPTAASCSYSLGAGGDSKSC 129

Protein Features

0

50

100

150

174

1

174

PIRSD37575

PIR31251

PTHR31251:SF0

G3DSA:4.10.1100.10

PSS1141

SSF103612

PF03110

Database

Entry ID

E-value

Start

End

InterPro ID

Description

PIRSF

PIRSF037575

1.49E-129

1

174

[IPR017238](#)

Squamosa promoter-binding protein

PANTHER

PTHR31251

0

16

131

No hit

No description

PANTHER

PTHR31251:SF0

0

16

131

No hit

No description

Gene3D

G3DSA:4.10.1100.10

3.7E-33

49

115

[IPR004333](#)

Transcription factor, SBP-box

PROSITE profile

PSS1141

0

51

128

[IPR004333](#)

Transcription factor, SBP-box

SuperFamily

SSF103612

0

53

131

[IPR004333](#)

Transcription factor, SBP-box

Pfam

PF03110

7.6E-40

54

129

[IPR004333](#)

Transcription factor, SBP-box

Gene Ontology

GO Term

GO Category

GO Description

GO:006355

Biological Process

regulation of transcription, DNA-dependent

GO:0010321

Biological Process

regulation of negative phase change

GO:005634

Cellular Component

nucleus

GO:000377

Molecular Function

DNA binding

GO:003360

Molecular Function

sequence-specific DNA binding transcription factor activity

Plant Ontology

PO Term

PO Category

PO Description

PO:0000013

anatomy

cauline leaf

PO:0000037

anatomy

shoot apical meristem

PO:0002030

anatomy

inflorescence meristem

PO:0008019

anatomy

inflorescence

PO:0009005

anatomy

root

PO:0009006

anatomy

shoot system

PO:0009009

anatomy

plant embryo

Conserved elements

19805634

19806134

19806634

19807134

1

2

3

4

5

Gene Structure

Flowers mostCons

Crucifer CMS

Type

No.

Chromosome

Start

End

Name

Score

Flowers mostCons

1

chr1

19806339

19806353

chr1.19463708-19892909.1378

7

Flowers mostCons

2

chr1

19806557

19806573

chr1.19463708-19892909.1379

6

Flowers mostCons

3

chr1

19806636

19806647

chr1.19463708-19892909.1380

9

Flowers mostCons

4

chr1

19806669

19806678

chr1.19463708-19892909.1381

7

Flowers mostCons

5

chr1

19806681

19806707

chr1.19463708-19892909.1382

20

[illegible]

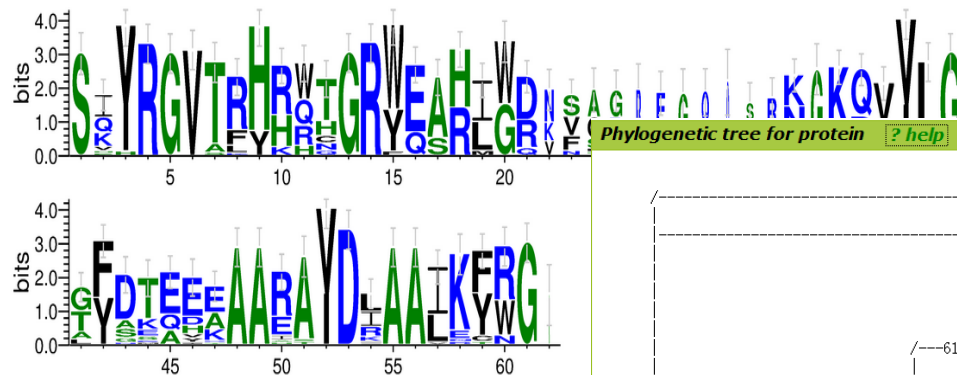
详尽的注释——家族水平

>AT1G79700.2/51-110
SPYRGVTRHRWTGRYEHLWDkNSW-NDtqtKKgRQVYLGA YDEEEAARAYDLAALKYW
G-
>AT1G79700.2/153-204
SKYRGVARHHHNGRWEARIGrVf----G----N-KYLYLGTATQEEAAIAYDIAAIEYR
G-
>AT1G72570.1/222-281
SQYRGVTRHRWTGRYEHLWDnSCKkEGq-tRRgRQVYLG GYDEEEKAARAYDLAALKYW
G-
>AT1G72570.1/324
SVYRGVTRHHQHGRWQ
G-
>AT5G17430.1/209
SIYRGVTRHRWTGRYE
G-
>AT5G17430.1/311
SIYRGVTRHHQHGRWQ
G-
>AT4G37750.1/282
SQYRGVTRHRWTGRYE
G-

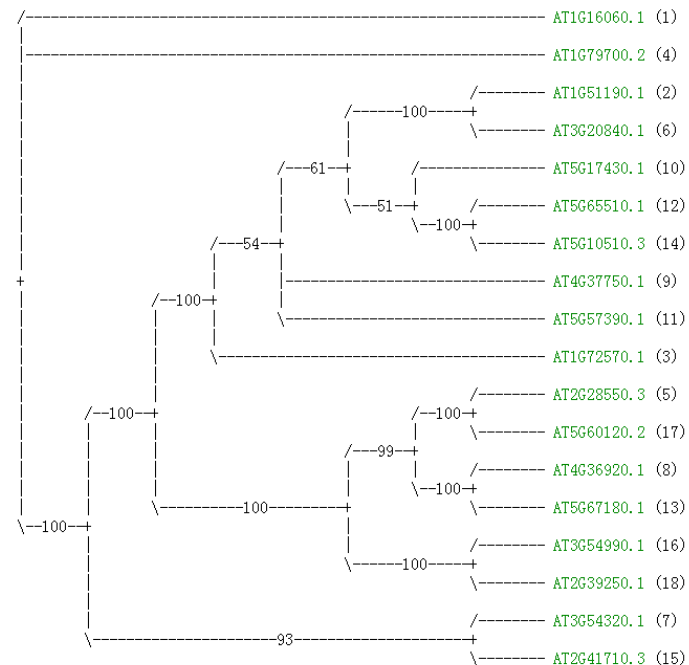
多序列比对

DNA binding domain (DBD) Alignment Logo [? help](#)

[Back to Top](#)



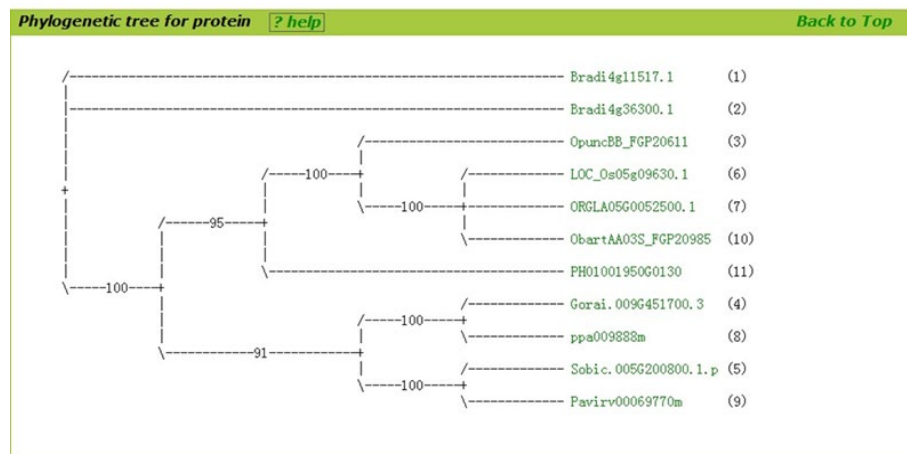
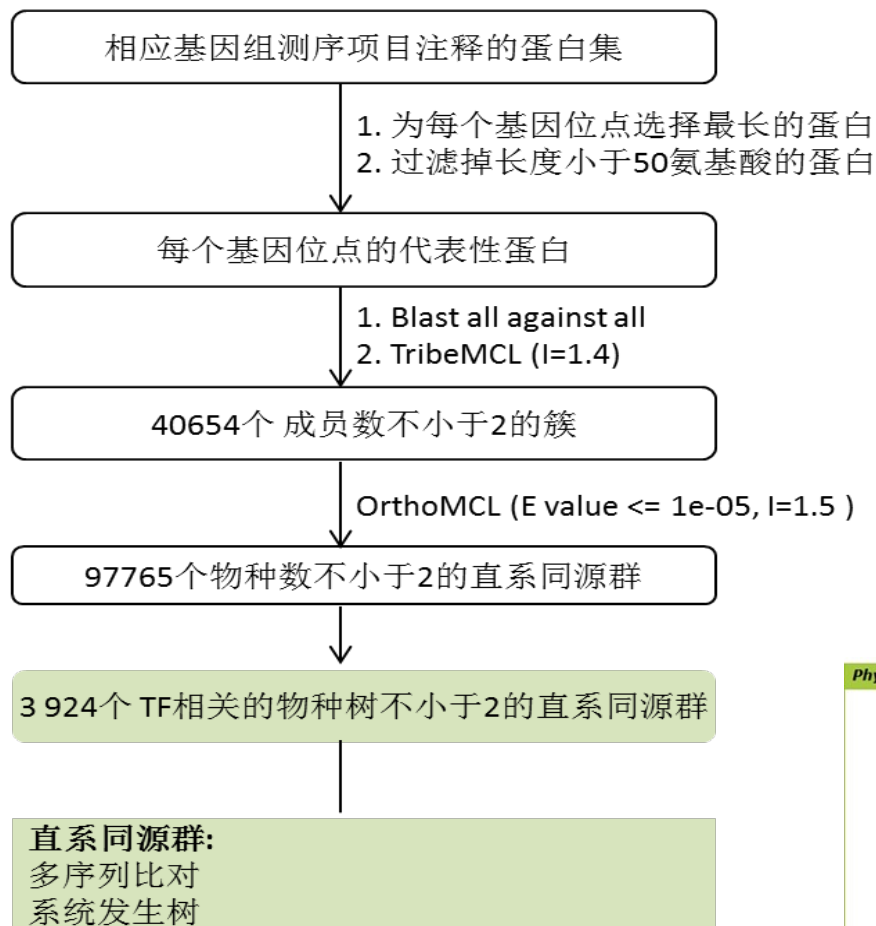
Phylogenetic tree for protein [? help](#)



Weblogo图

系统发生树

直系同源群



植物转录因子数据库PlantTFDB



PlantRegMap/PlantTFDB v5.0

Plant Transcription Factor Database

[Home](#) [TFext](#) [BLAST](#) [Prediction](#) [Download](#) [Help](#) [About](#) [Links](#) [PlantRegMap](#)

(e.g., LFY)

Note:

- The new version PlantTFDB 5.0 has been incorporated into [PlantRegMap](#).

Browse by Species

[open all](#) | [close all](#)

Taxonomic Group (165 species) (G)-species with genome sequence

☒ Chlorophytae (16 species)

☒ Charophyta (1 species)

☒ Marchantiophyta (1 species)

☒ Bryophyta (2 species)

☒ Lycopodiophyta (1 species)

☒ Coniferophyta (5 species)

☒ Basal Magnoliophyta (1 species)

☒ Monocots (38 species)

☒ Eudicots (100 species)

Quick links: [Arabidopsis thaliana](#), [Glycine max](#), [Oryza sativa](#), [Populus trichocarpa](#), [Solanum lycopersicum](#), [Zea mays](#)

Couldn't find the species interested? Please try the [extended TF repertoires](#) or identify TFs by the [prediction server](#).

包含165个物种，320370个转录因子

Jin, et al, 2017
Tian, et al, 2020

植物转录因子预测平台

Transcription Factor Prediction

The family assignment rules (see [details](#)) and thresholds determined by established methods (see [details](#)) are used to identify transcription factors from the input sequences. When you input nucleic acid sequences, [ESTScan 3.0](#) is employed to identify CDS regions of input nucleic acid sequences and translate them to protein sequences. By checking "Best hit in *Arabidopsis thaliana*", you will get the links for predicted transcription factors to the best hits in *Arabidopsis thaliana* in the result.

(See also: [TF binding site prediction](#), [Regulation prediction](#), [GO enrichment](#), and [TF enrichment](#).)

Input [protein](#) or [nucleic acid](#) sequences in [FASTA](#) format (Max file size:75M): [example](#)

Or load it from disk: No file chosen

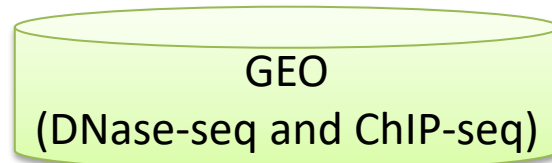
Link: ☐ Best hit in *Arabidopsis thaliana*

系统鉴定新测序植物中的转录因子

主要内容

- 植物转录因子分类、识别与注释 (PlantTFDB)
- 植物转录调控图谱的绘制
- 植物转录调控系统的架构原则与演变特征

基于实验鉴定的调控元件与相互作用

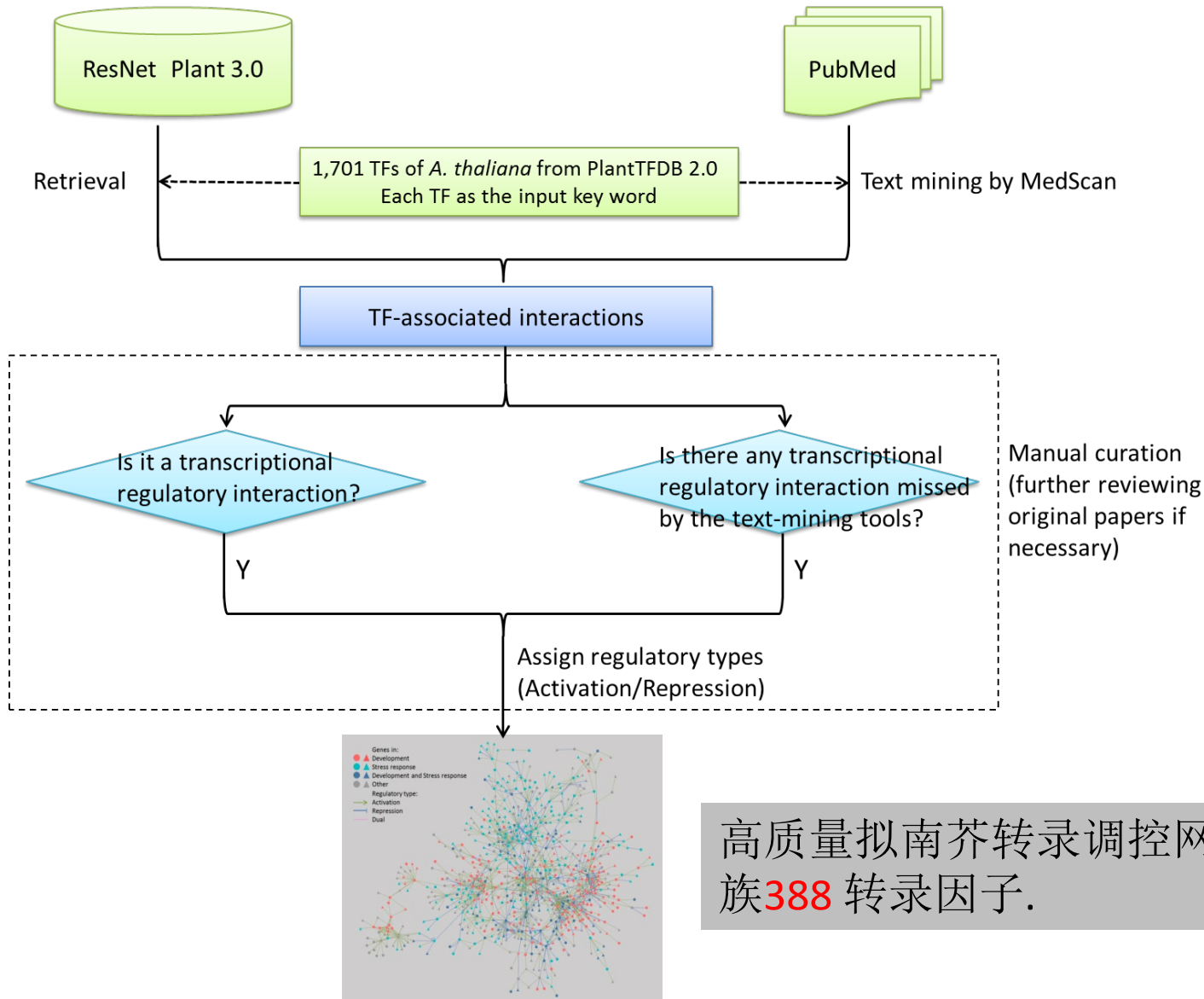


Type	Species	Organ	Peak
DNase I hypersensitive site	2	10	1 026 748
DNase I genomic TF footprints	1	4	4 794 773
TF binding peak (ChIP-seq)	2	4	345 920

9 340 transcriptional regulatory Interactions

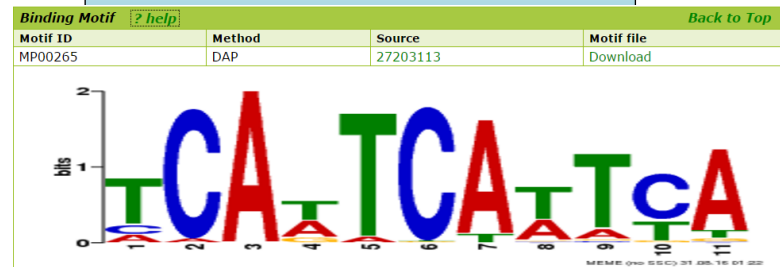
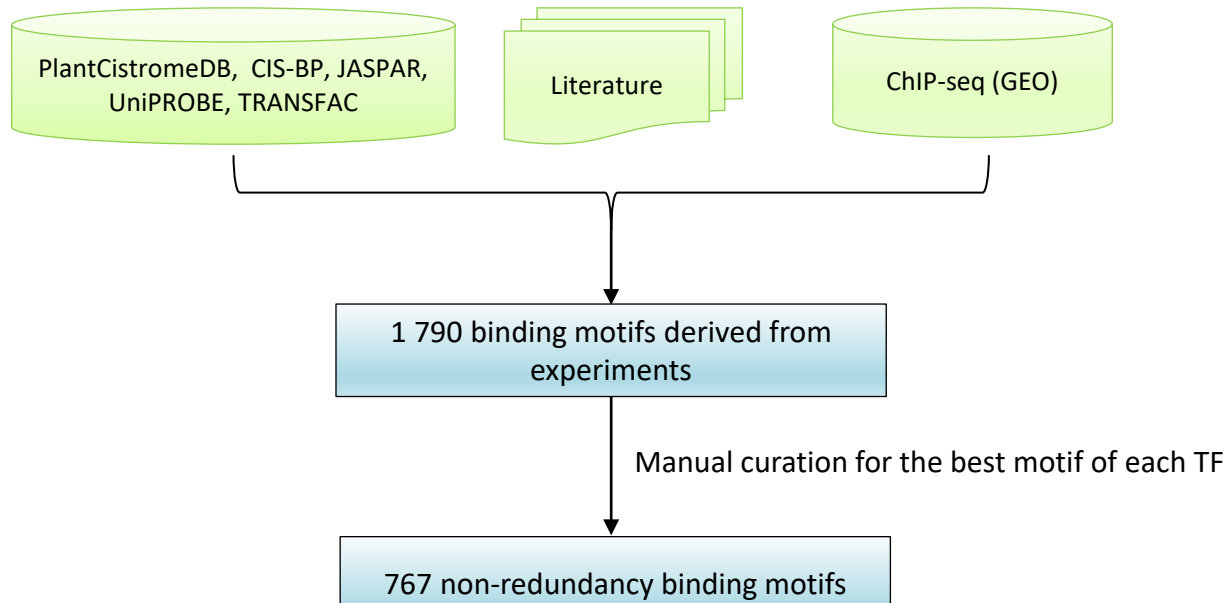
局限于模式物种（主要是拟南芥中）

基于文献挖掘构建的拟南芥转录调控网络 (ATRM)



高质量拟南芥转录调控网络，涉及47个家族388 转录因子。

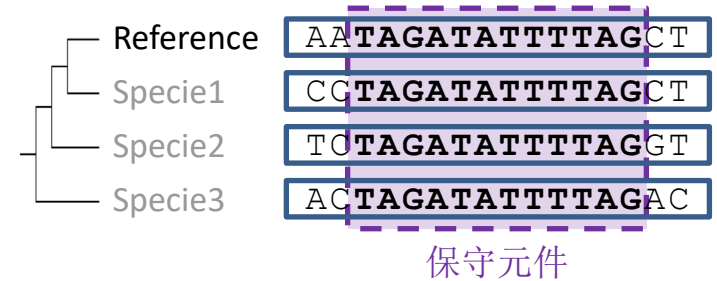
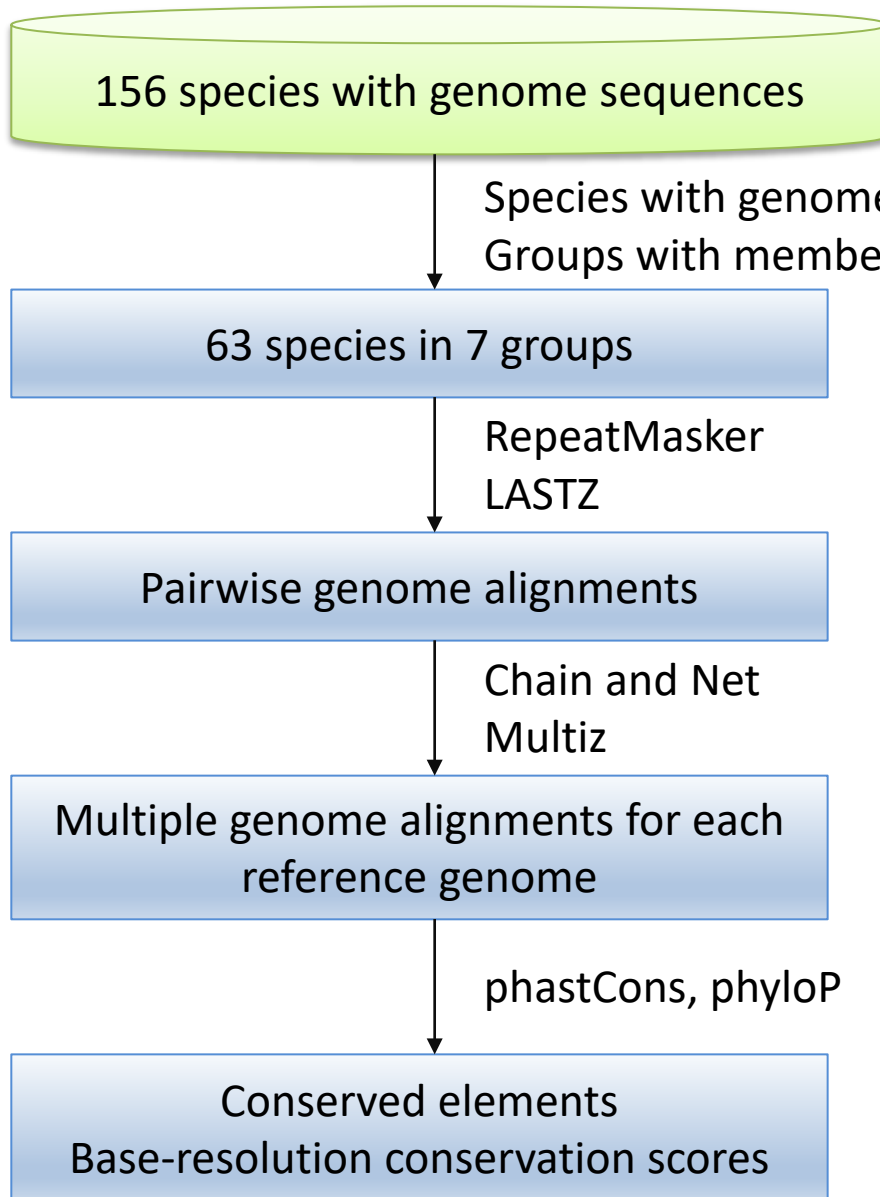
转录因子结合矩阵的整合与跨物种映射



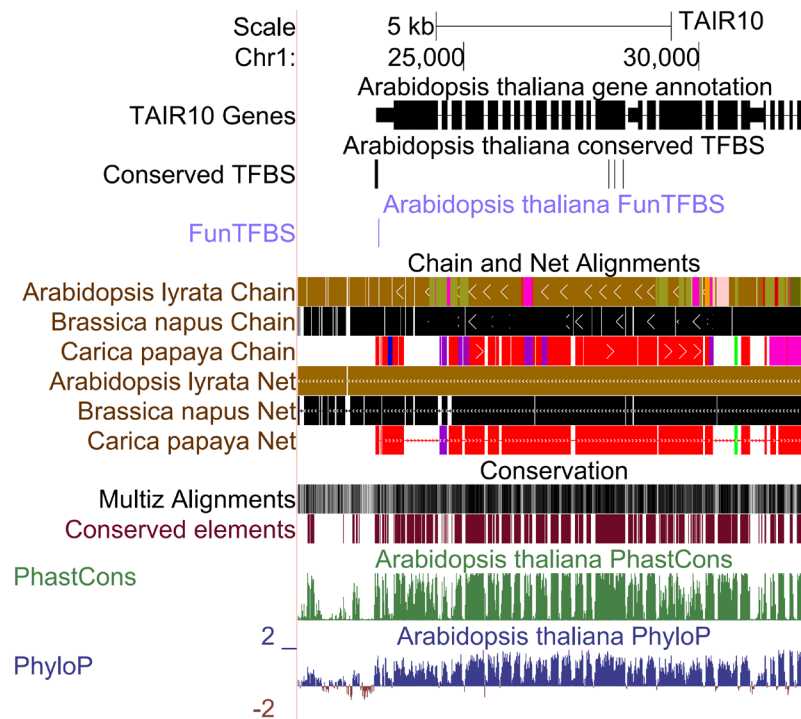
Projection (RBHs in the same family)

覆盖156物种的高质量转录因子结合矩阵集合, 覆盖77.6% (45 of 58) 的家族。

保守元件的系统鉴定



保守元件与单碱基分辨率的序列保守值

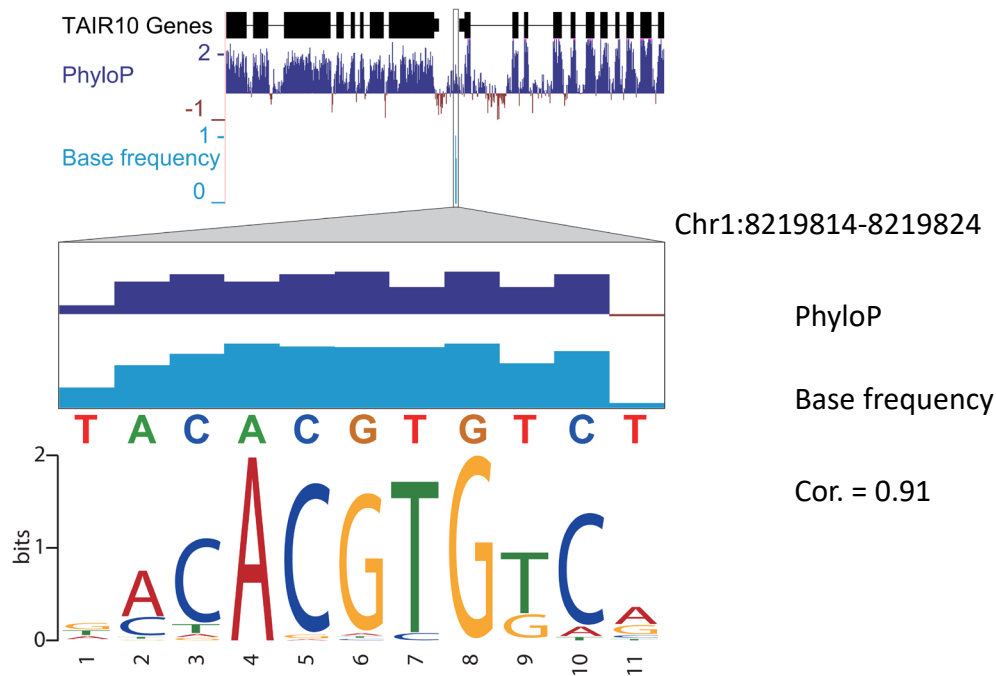


Total: **63** species in **7** groups

- 超过6700万个保守元件
- 超过220亿碱基的保守性分值

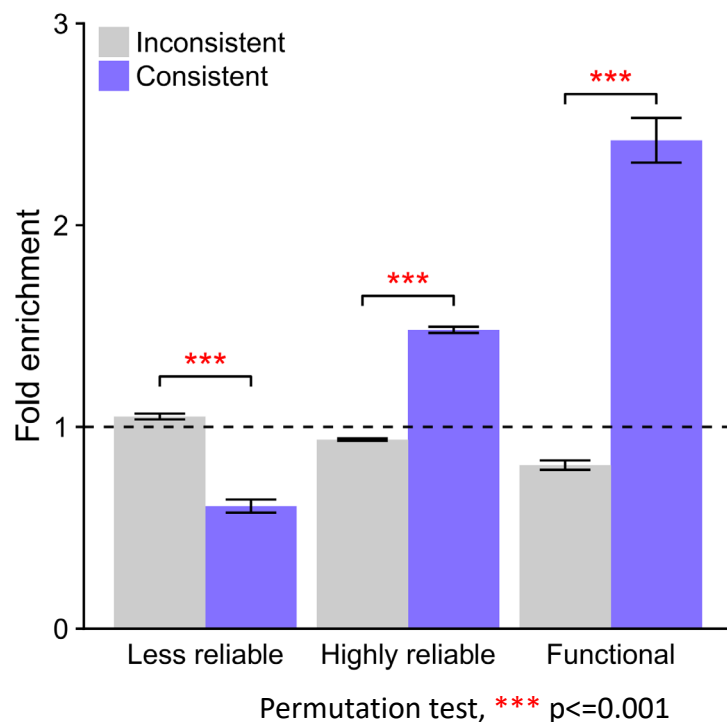
Tian, *et al. Nucleic Acids Res.*, 2020

结合矩阵碱基频率和序列保守性的相关性



Tian *et al.*, *Nucleic Acids Res*, 2020

此相关性可有效筛选功能性转录因子结合位点



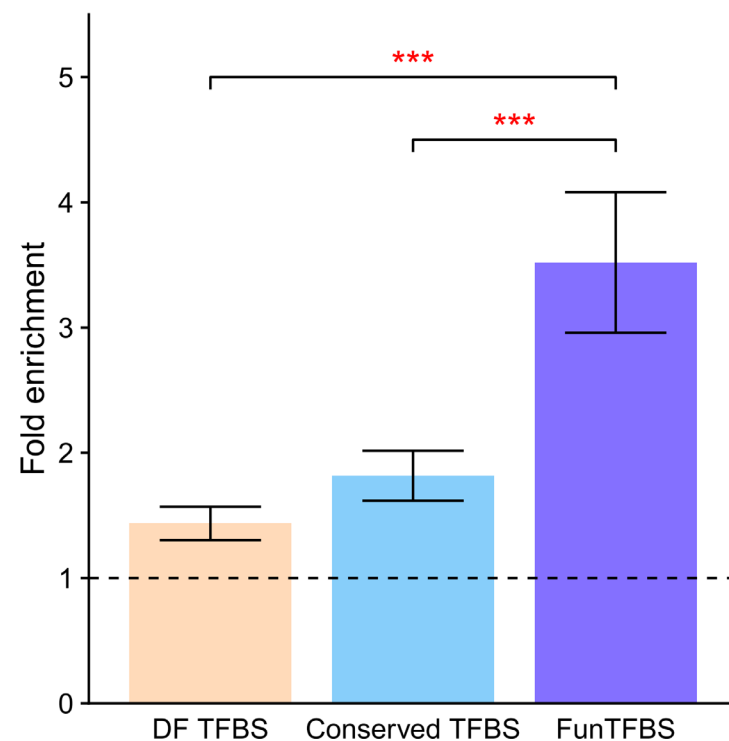
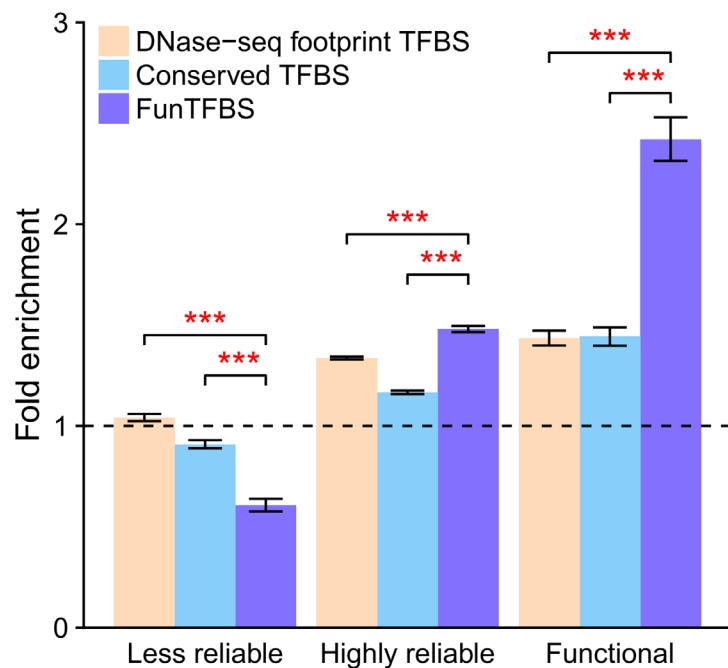
开发了FunTFBS来系统鉴定功能性转录因子结合位点

评估指标：转录因子结合位点（TFBS）中被基准数据集支持的比例。

Tian *et al.*, *Nucleic Acids Res*, 2020

FunTFBS准确地鉴定功能性转录因子结合位点与调控关系

方法	描述
DF TFBS	与基因组印迹重叠的TFBS
Conserved TFBS	与保守元件重叠的TFBS
FunTFBS	经过FunTFBS筛选后得到的TFBS



基于拟南芥转录调控网络ATRM

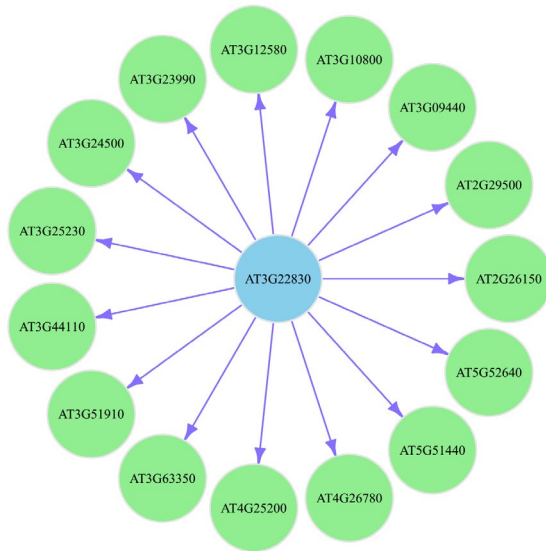
绘制基因组范围的转录调控图谱

Lineage	TF (genome-wide)	TFBS (genome-wide)	TFBS (in promoter)	Regulation
BOP clade	1,916	1,854,476	227,485	205,317
Brassicales	9,290	8,132,692	1,295,434	1,186,210
Fabales	3,587	6,196,677	440,244	356,140
Lamiales	1,085	1,281,855	76,684	61,165
Malpighiales	1,681	1,298,748	131,610	107,447
PACMAD clade	2,038	1,804,188	196,038	183,386
Rosales	1,897	1,428,865	126,082	94,897
Total	21,494	21,997,501	2,493,577	2,194,562

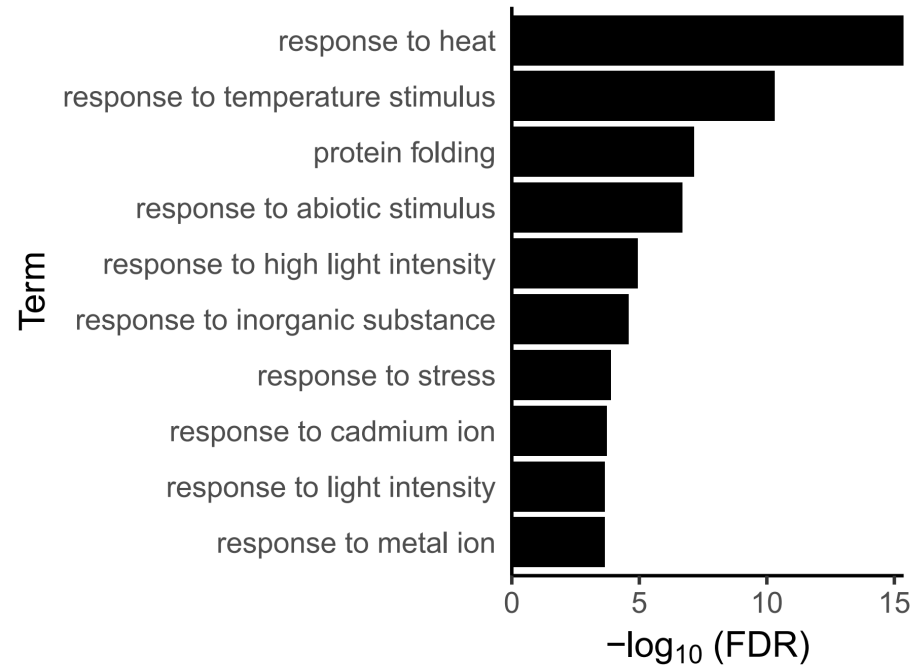
覆盖7个植物类群63个物种

应用示例：预测转录因子功能

AT3G22830: 与热刺激响应相关的转录因子（HSF家族）



靶基因的GO富集分析结果



转录调控预测与分析工具

➤ 转录因子结合位点预测

基于结合矩阵扫描并鉴定用户输入序列中的转录因子结合位点。

➤ 转录调控网络预测

根据用户输入的基因组区域预测转录因子与靶基因之间的调控关系。

➤ 转录因子富集

根据用户输入的基因列表鉴定上游共同的调控基因。

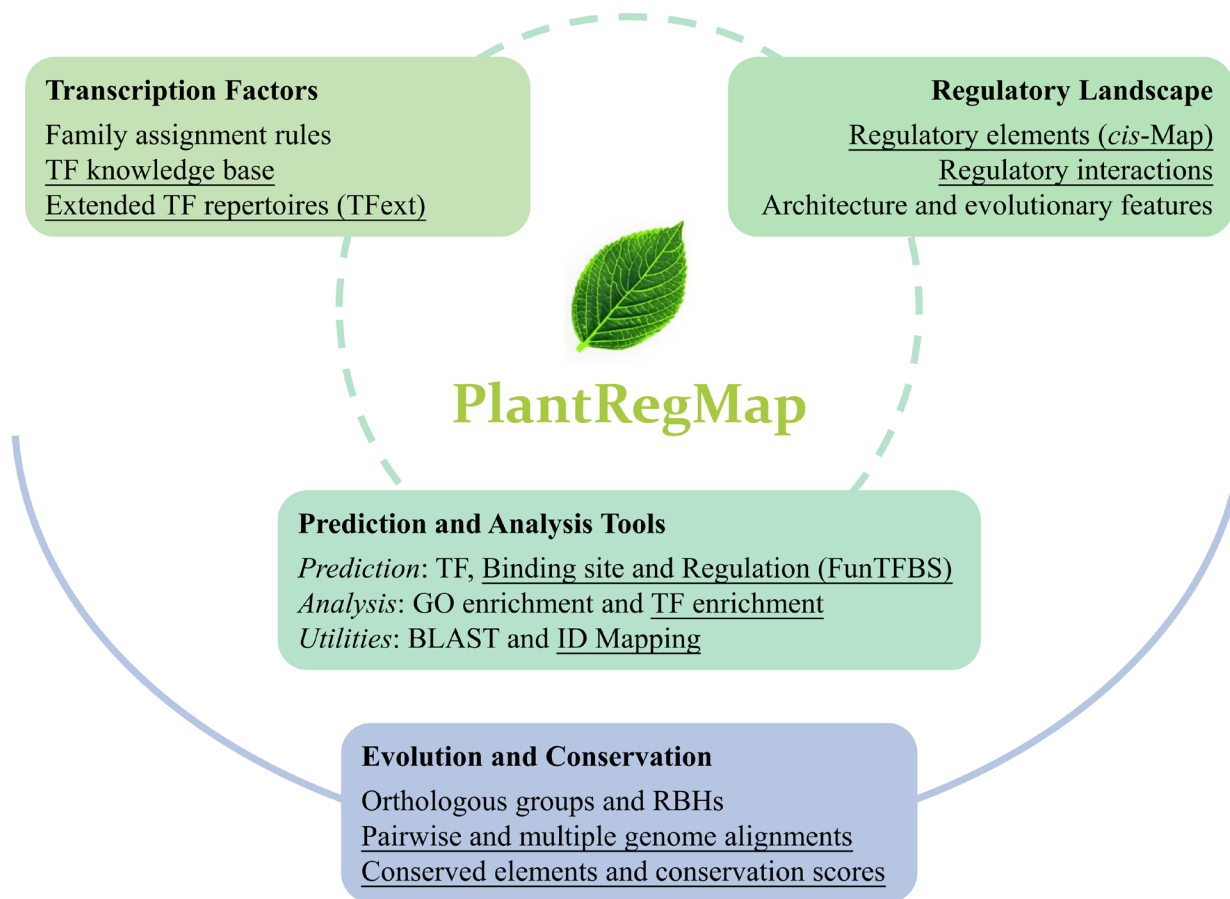
➤ GO富集

根据用户输入的基因列表鉴定富集的GO条目。

➤ ID映射

基于双向最优BLAST将用户输入的序列ID对应到数据库中的相应基因ID。

植物转录调控数据与分析平台



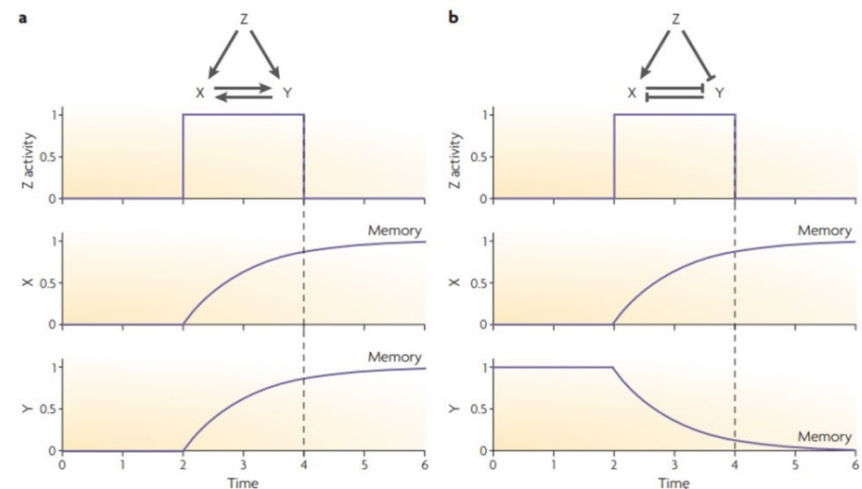
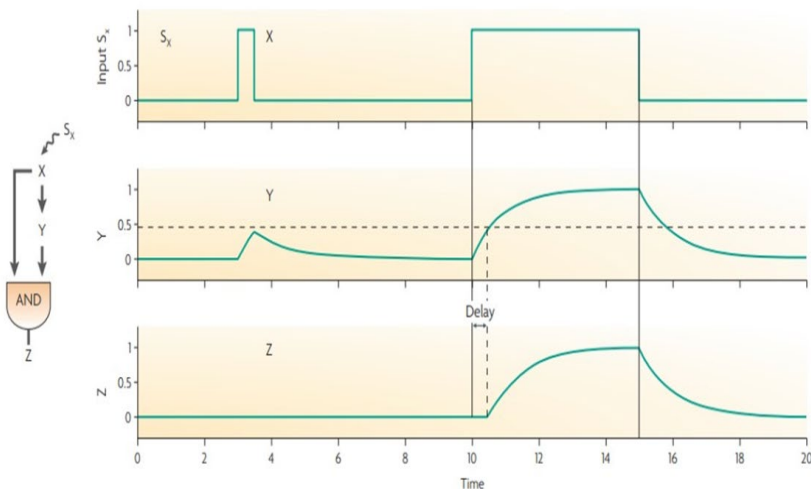
年访问量逾四千万，总引用次数超四千。

主要内容

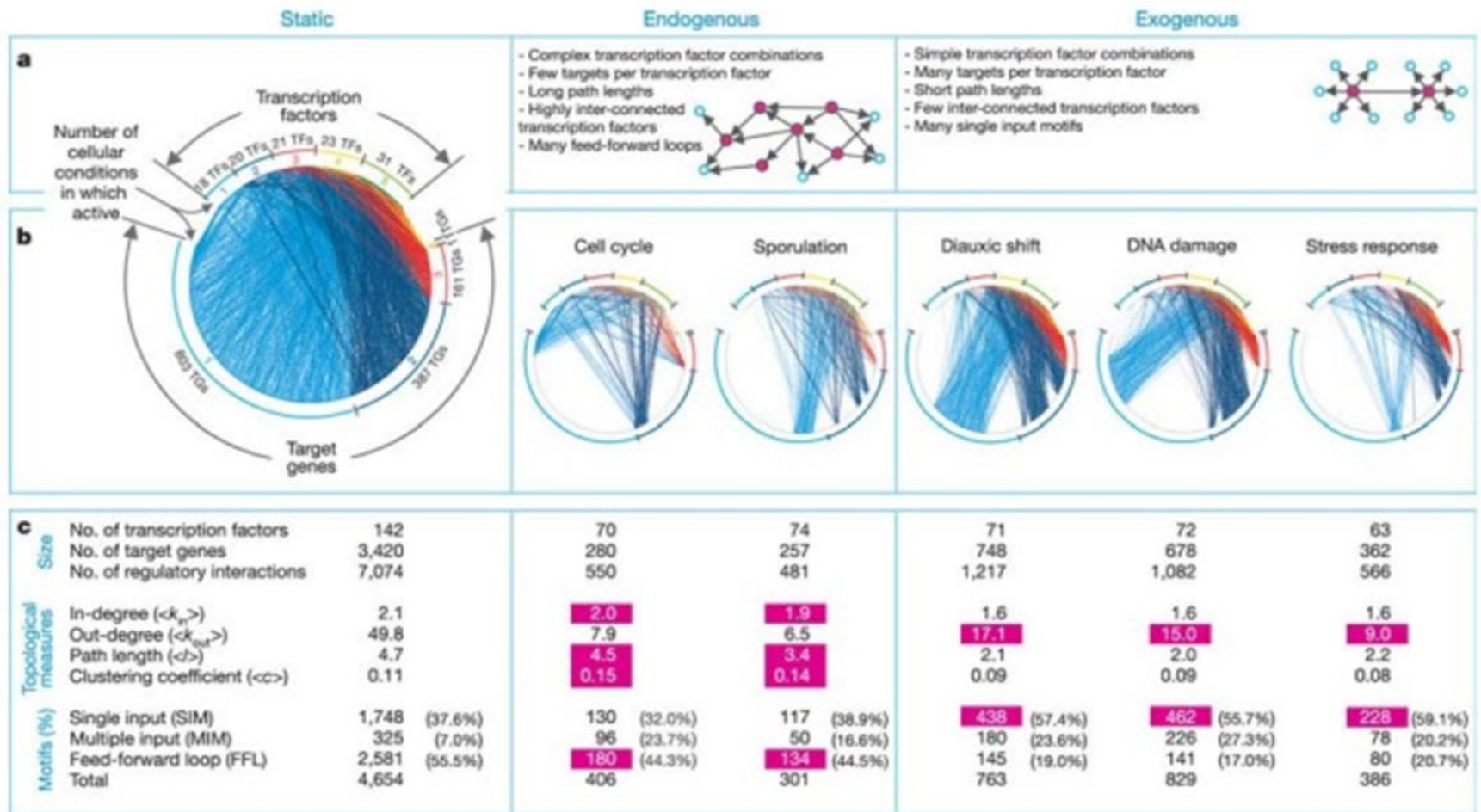
- 植物转录因子分类、识别与注释 (PlantTFDB)
- 植物转录调控图谱的绘制
- 植物转录调控系统的架构原则与演变特征

转录调控网络由一些结构元件组成

- 在大肠杆菌、酿酒酵母等单细胞模式生物中的研究发现：它们的转录调控网络由一些显著多于相应随机网络的调控模式组成，这些显著出现的模式即**结构元件**（Network Motif）
- 在高等生物的发育通路中，还存在一类结构元件——反馈环（Feedback Loop）
- 实验和动力学模拟表明这些结构元件是能完成特定生物功能的功能元件



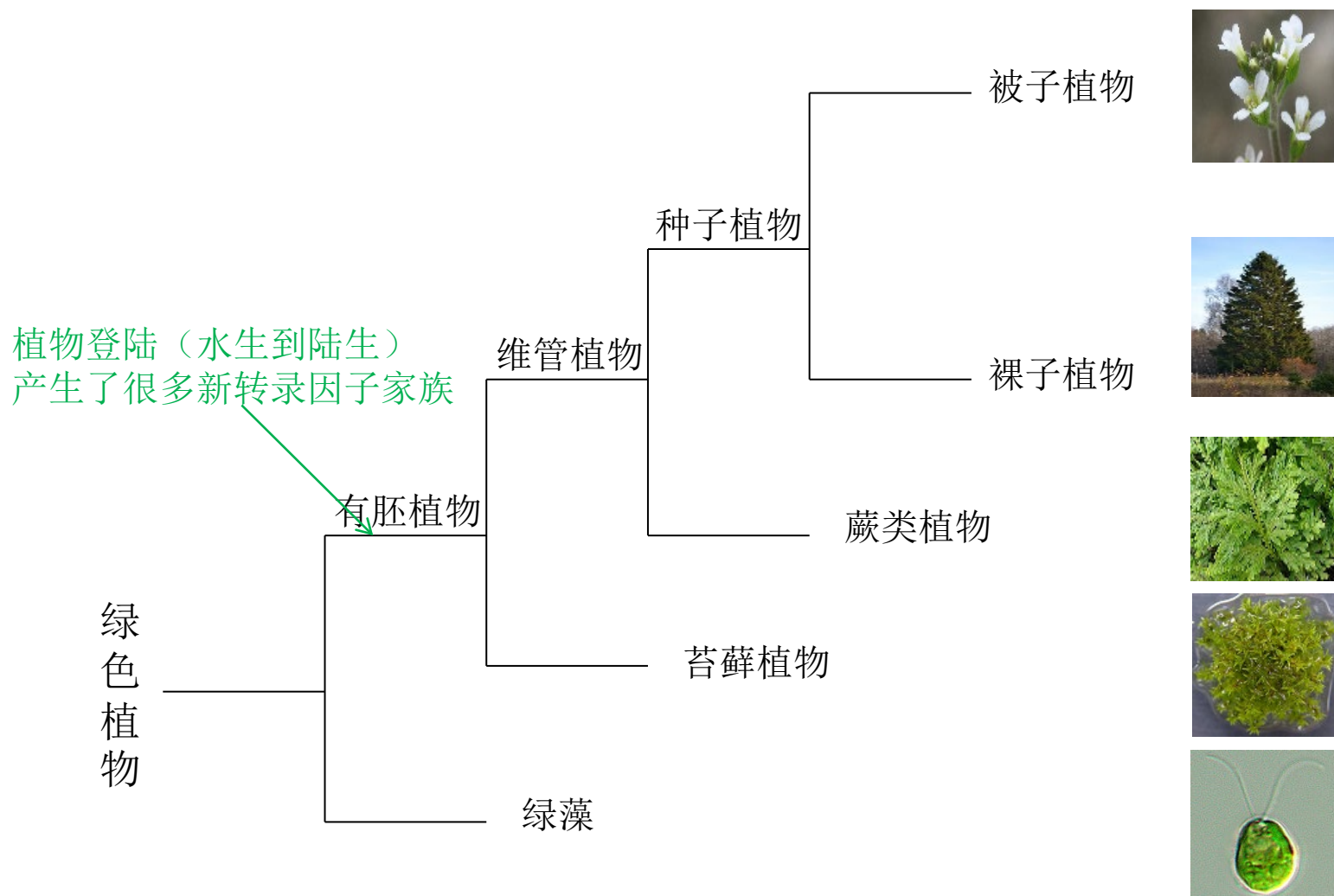
酵母内、外源系统在网络结构上存在显著差异



内源性系统使用更加复杂的调控精确地推进各生命过程

外源性系统使用更具影响力的转录因子通过简单的调控快速响应外部的各种刺激

水生到陆生——植物演化历程中的重大事件

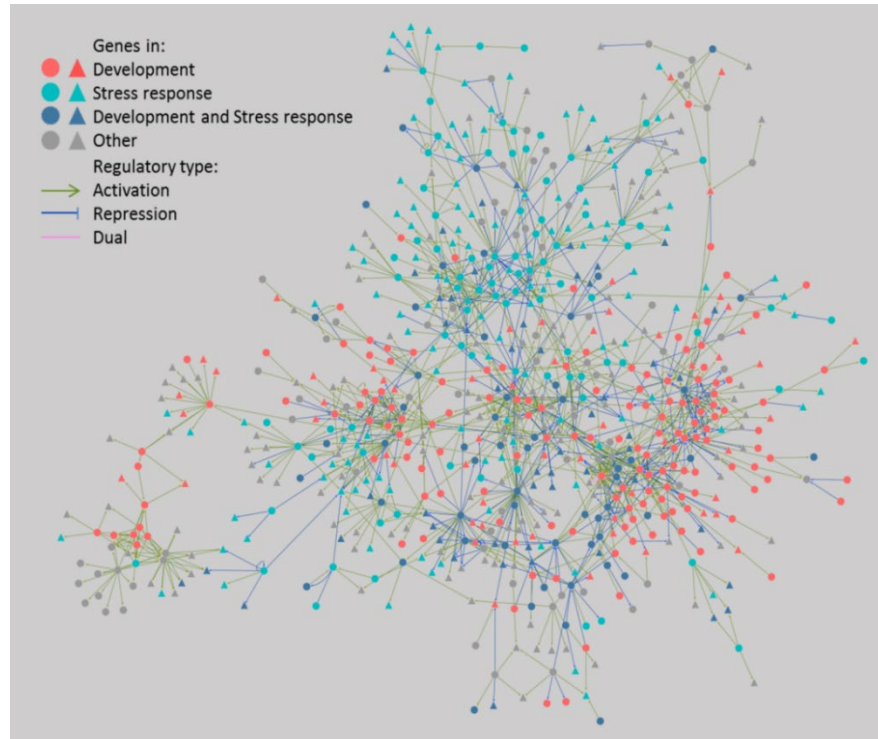


形态和生境的剧烈改变

科学问题

- 植物转录调控网络是如何架构的？发育子网络与应激子网络在全局拓扑结构和局部结构元件组成方面是否有所不同？
- 与单细胞生物相比，多细胞植物的调控网络是否有新的结构元件？如有，参与何功能？
- 新转录因子家族是何时产生的？它们在性质及网络位置上是否有所偏好？以及如何促进转录调控系统演变的？

发育子网络与应激子网络的划分



发育子网络：ATRM中参与发育过程的基因之间的相互调控组成的网络

应激子网络：ATRM中参与应激过程的基因之间的相互调控组成的网络

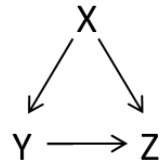
发育子网络和应激子网络在全局拓扑结构上的差异

	发育	应激	
<Targets per TF>	2.96	< 3.77	→ 转录调控的瞬时影响力
<TFs per target>	2.75	> 1.97	→ 调控的复杂性
< Path length >	3.77	> 1.74	→ 调控的路径长度
< Clustering coefficient >	0.21	> 0.08	→ 调控的复杂性

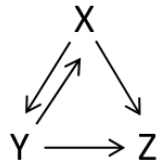
发育子网络具有更多转录因子与转录因子之间的复杂调控，应激子网络具有较短的调控路径和较简单的调控模式。

Jin, *et al.*, *Mol. Biol. Evol.*, 2015

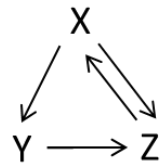
网络结构元件组成



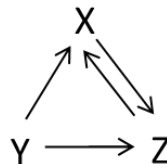
Motif 5 (303)



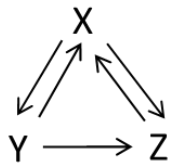
Motif 6 (53)



Motif 10 (22)



Motif 11 (34)



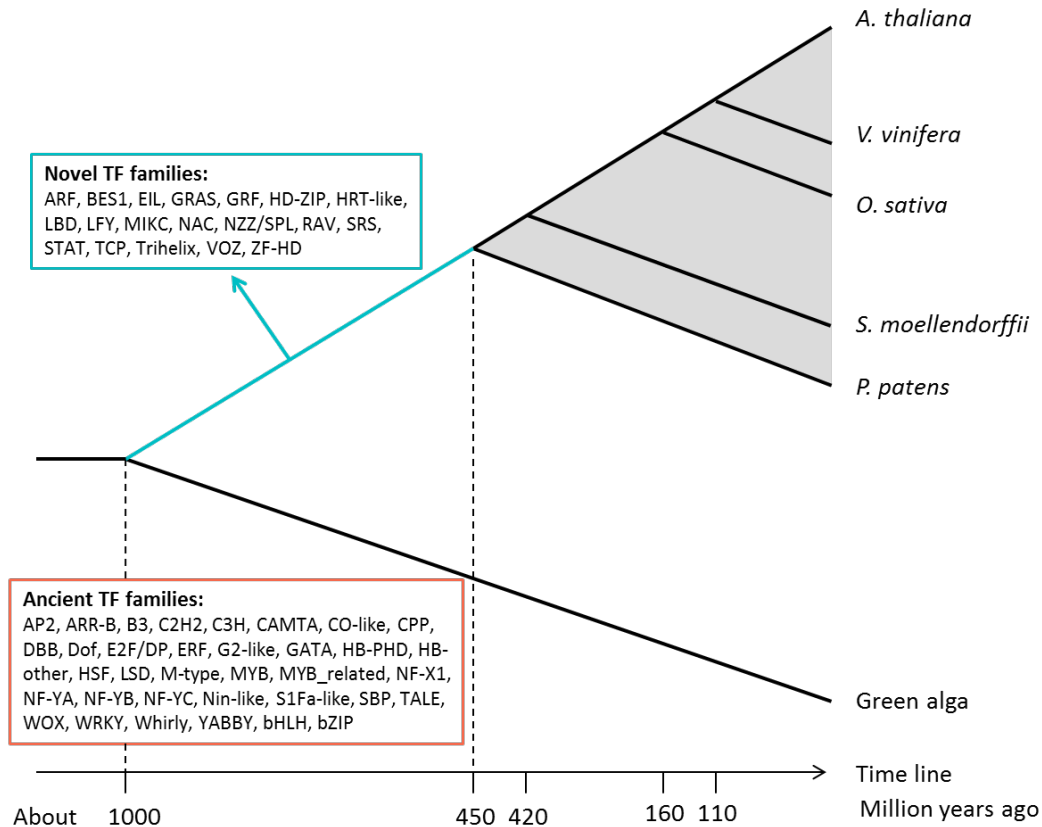
Motif 12 (25)

Motif	Development	Stress response
Motif 5	107	54
Motif 6	25	4
Motif 10	17	0
Motif 11	32	0
Motif 12	23	0

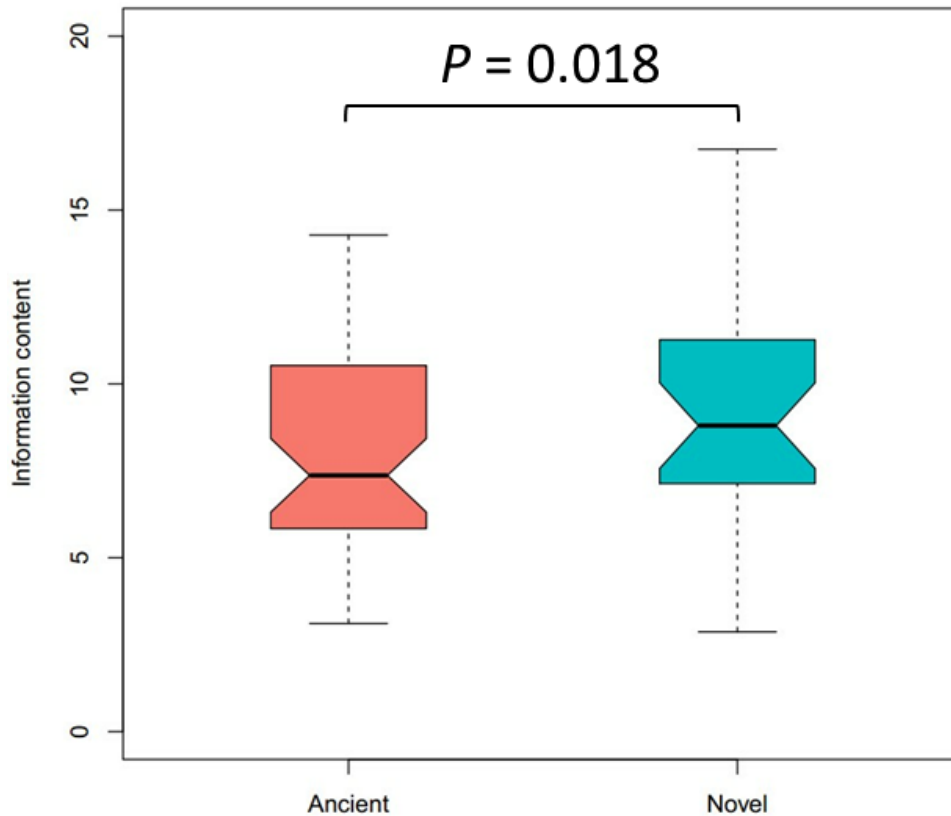
GO ID	GO term	P
GO:0048731	systemdevelopment	3.2e-16
GO:0048856	anatomical structure development	7.9e-16
GO:0007275	multicellular organismal development	3.6e-15
GO:0032501	multicellular organismal process	7.9e-15

与单细胞生物相比，拟南芥转录调控网络具有三个新的结构元件，它们
在多细胞发育中发挥至关重要的作用。

植物登陆后产生19个新转录因子家族



新类型转录因子具有更高的DNA结合特异性



结合特异性：结合矩阵的信息量
(information content)

新类型转录因子倾向于参与发育过程的构建

Type	Development	Stress response
Novel	85	19
Ancient	157	150

One-tailed Fisher's exact test: $P = 1.36\text{e-}08$, odds ratio= 4.26

GO ID	GO term	<i>P</i>
GO:0048731	system development	3.9e-14
GO:0048856	anatomical structure development	5.8e-14
GO:0048513	organ development	1.3e-12
GO:0032502	developmental process	1.8e-11
GO:0007275	multicellular organismal development	1.8e-11
GO:0032501	multicellular organismal process	2.4e-11
GO:0022621	shoot system development	1.4e-10
GO:0048367	shoot development	1.4e-10
GO:0048366	leaf development	5.5e-09
GO:0048827	phyllome development	1.3e-08
GO:0010016	shoot morphogenesis	1.9e-08

新类型转录因子参与网络构建的偏好性

	Targets per TF		TFs per target		Motifs (5, 6)		Motifs (10, 11, 12)		Proportion of TFs in targets	
Mean	7.13		2.76		5.13		1.46		0.42	
</> mean	<	>	<	>	<	>	<	>	<	>
Novel	33	17	29	21	41	9	33	17	12	30
Ancient	79	33	68	44	75	37	94	18	57	33
Odds ratio	0.81		0.89		2.24		0.37		0.23	
<i>p</i> *	0.34		0.44		0.04		0.01		1.80e-04	

* One-tailed Fisher's exact test

与古老类型的转录因子相比，新类型转录因子倾向于调控其它转录因子以及参与单细胞生物所不具备的新网络结构元件的构建。

进一步分析揭示了转录因子结合特异性与其参与网络连接偏好性之间的内在关联，一定程度上解释了为什么在拟南芥中新类型转录因子倾向于参与发育过程。

总结

- 通过纵览7000余篇植物转录因子相关的文献，整理出一套完整的植物转录因子家族分类规则。
- 系统识别了156种植物基因组中的转录因子，首次发布了覆盖绿色植物各大分支的转录因子全谱。
- 对识别出的转录因子做了详尽的功能和演化注释，建成植物转录因子的知识库。
- 基于DNase-seq、ChIP-seq等高通量测序数据，系统鉴定了植物基因组中的调控元件。
- 通过文献挖掘和人工校验、ChIP-seq以及使用转录因子结合矩阵预测等方式构建了覆盖绿色植物主要分支的转录调控网络。
- 开发出一系列植物转录调控预测与分析的工具，包括转录因子识别、结合位点预测、调控预测和功能富集分析等。
- 通过系统分析上述转录调控图谱，揭示了发育子网络和应激子网络、单细胞生物和多细胞生物、植物和动物在网络结构上的差异，展示了网络结构与功能之间的内在关联。
- 揭示了新类型转录因子在性质和网络位置上的偏好性及其内在原因，提出模型解释了新类型转录因子在多细胞生物起源中的重要贡献。

感谢参与项目和平台
建设的所有人员！

谢谢大家！