

植物基因组数据整合挖掘与应用

中国科学院微生物研究所

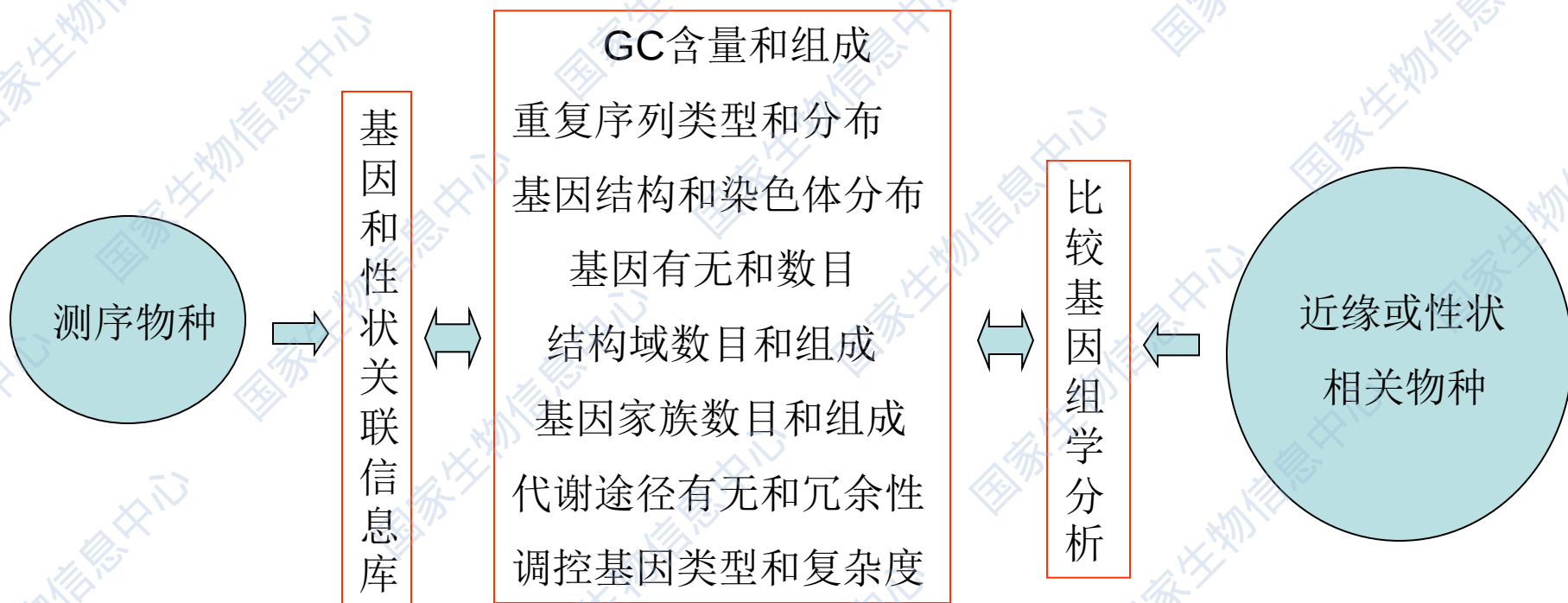
胡松年

husn@im.ac.cn

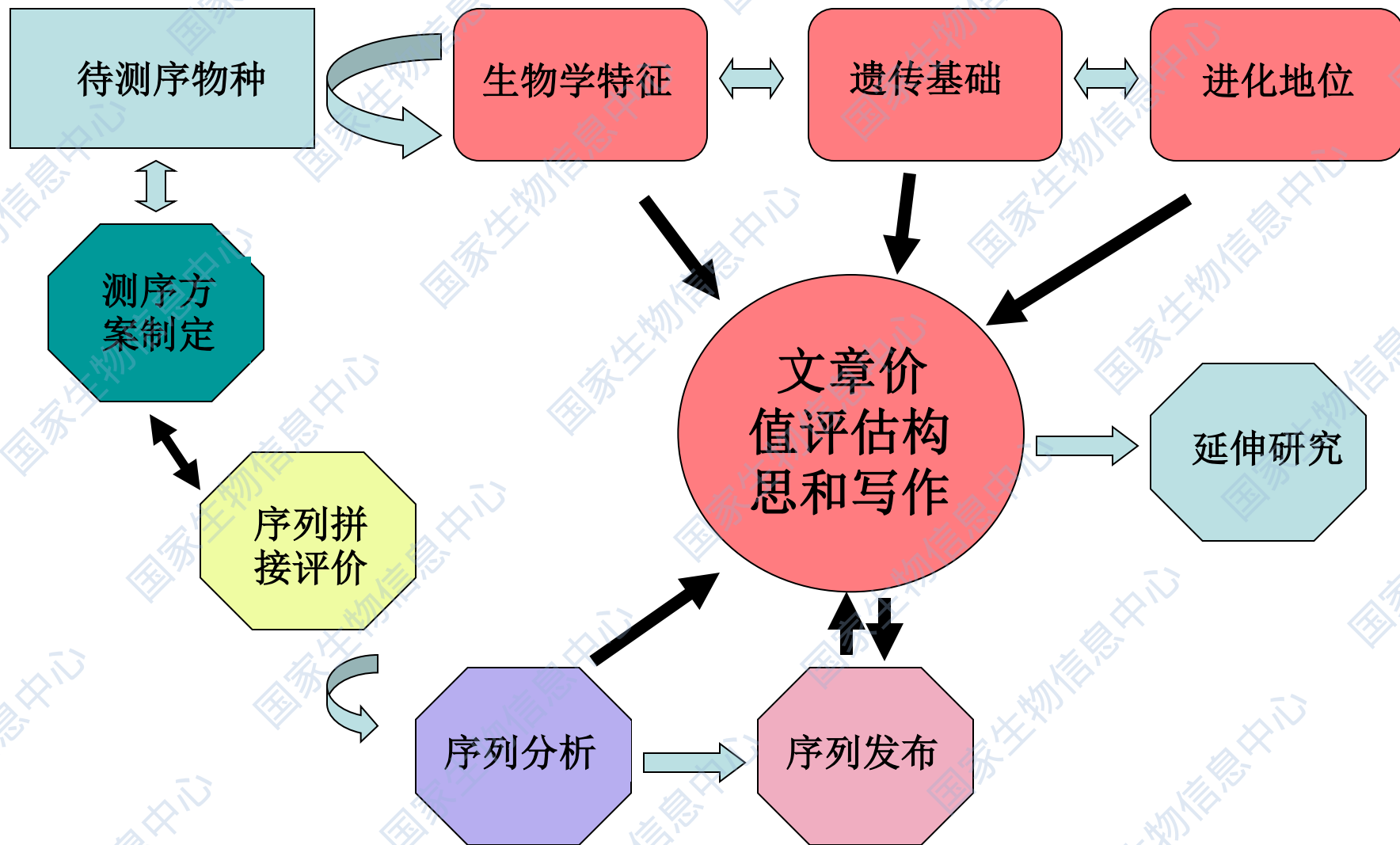
物种重要性及其测序必要性

模式生物；病原性；经济性状；进化地位

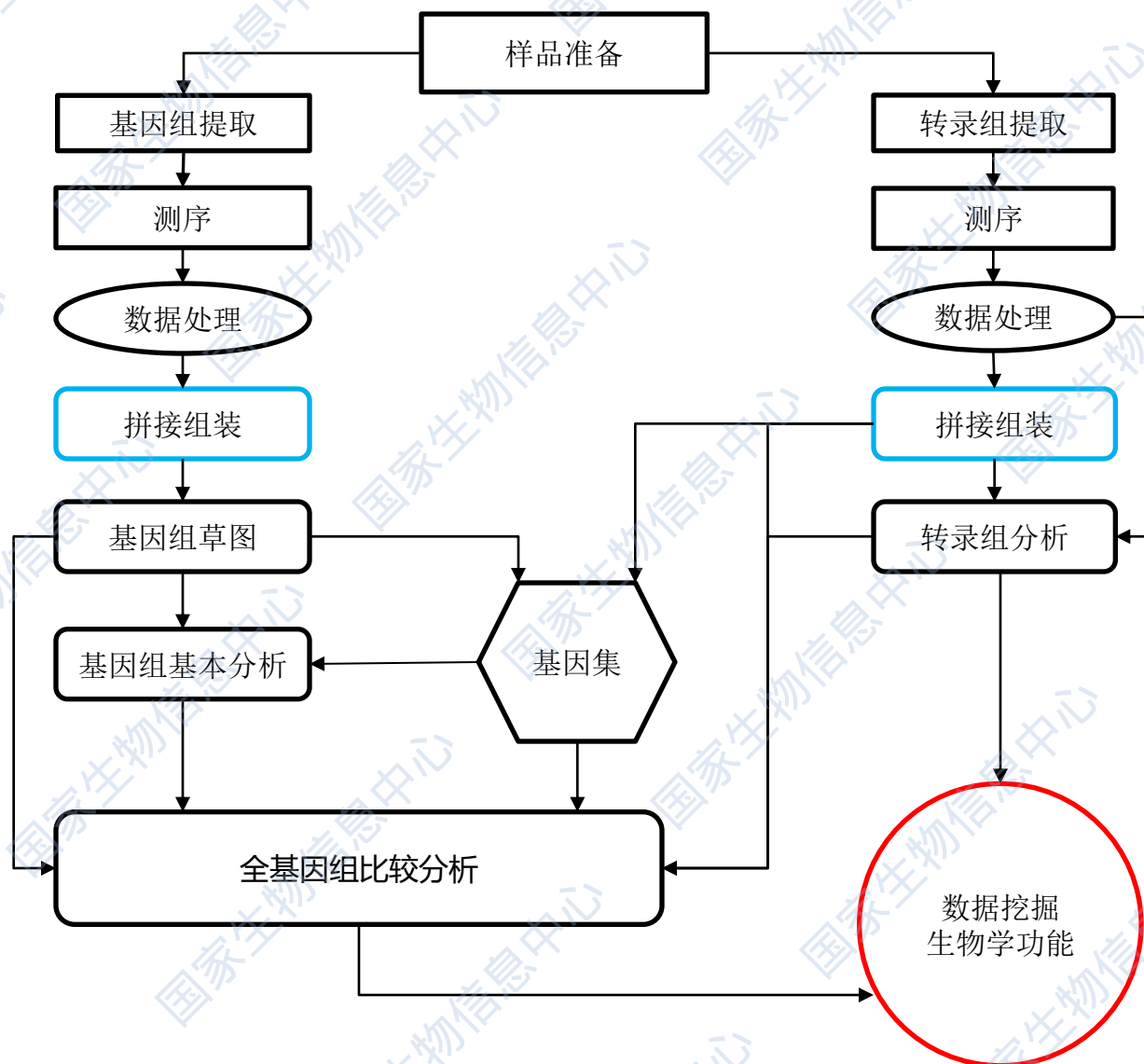
1. 基因组测序，需利用拼接，基因组注释(RNA，蛋白质和重复序列等)；
2. 关键性状的遗传因素 (功能基因组学确定性状决定的基因或基因群)；
3. 比较基因组学分析-----进化地位，性状产生和进化的基因组线索等



基因组从头测序线路图



基因组从头测序技术路线



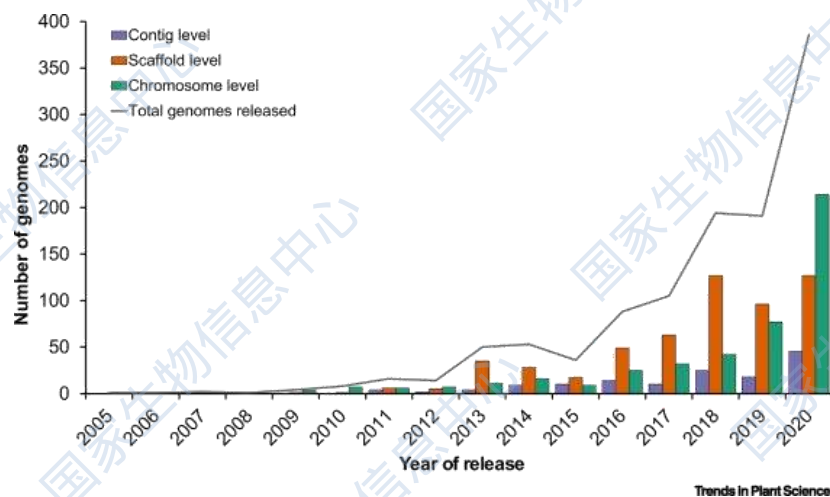
The challenges for plant genome assembly

- Gene family
- Tandem duplication
- Cluster located repeat
- High repeat content
- Large genome size
- High heterozygosity
- Recent whole genome duplication
- Polyploid



高质量参考基因组是功能基因研究和分子设计育种的基础

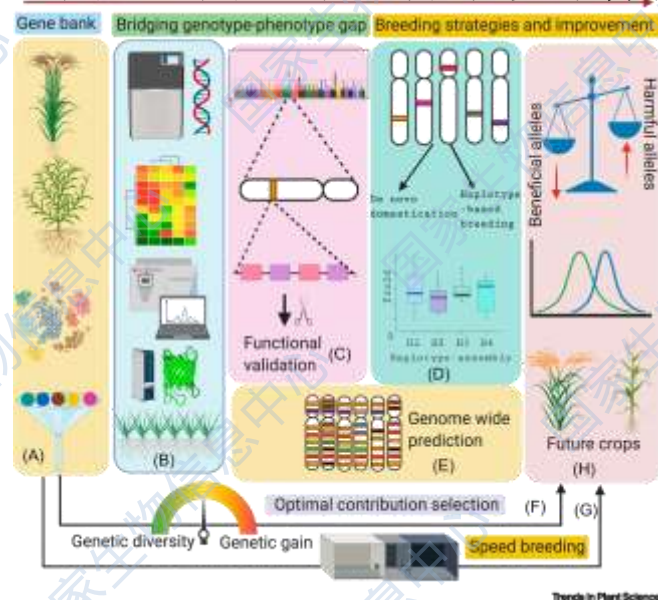
染色体水平的基因组指数级上升 ↘



同一物种的参考基因组也在不断地更新

水稻、玉米、大豆、棉花.....

基因组基因型-表型分子设计育种



Rajeev et al. 2021. Trends in Plant Science



GIGA
SCIENCE

GigaScience, 8, 2019, 1-4

doi:10.1093/gigascience/gp010
Open Access

COMMENTARY

How can a high-quality genome assembly help plant breeders?

Juliana Benevenuto, Luis Felipe V. Ferrão, Rodrigo R. Amadeu and Patricio Munoz



参考基因组质量的影响:

四倍体栽培蓝莓

(*Vaccinium corymbosum*)

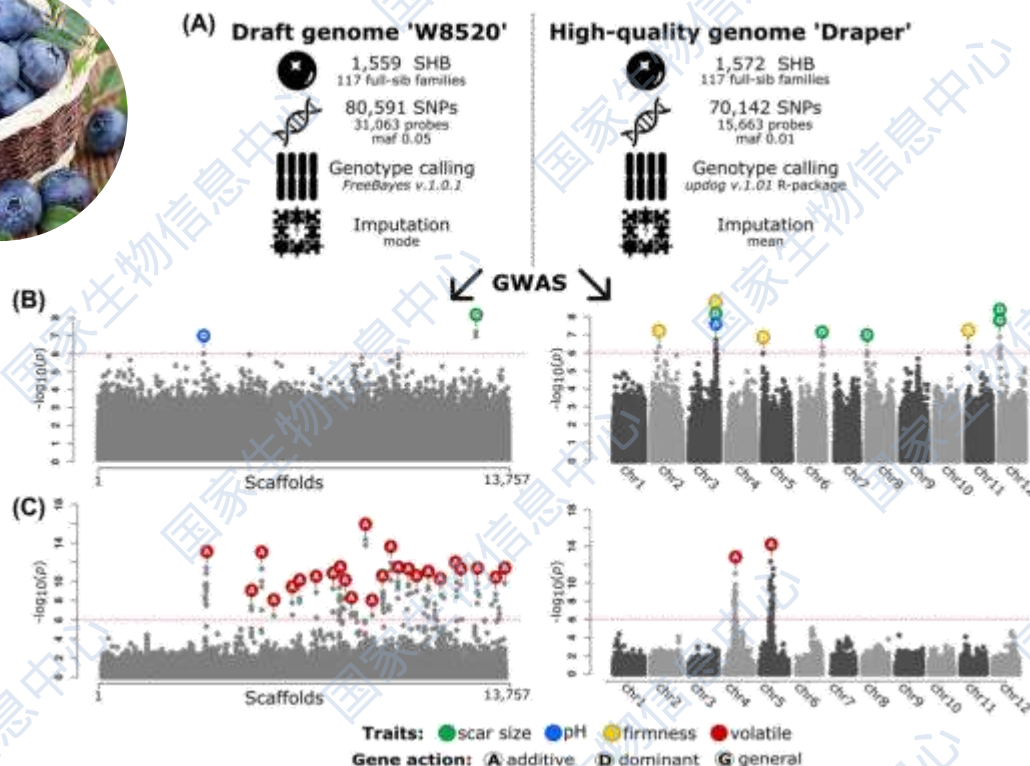
- 更少的冗余组装序列
- 更少的冗余SNP
- 更准确的GWAS峰值

Juliana et al. 2019. GigaScience

为什么需要高质量参考基因组?

454 + Illumina

Illumina + 10X Genomics + Hi-C

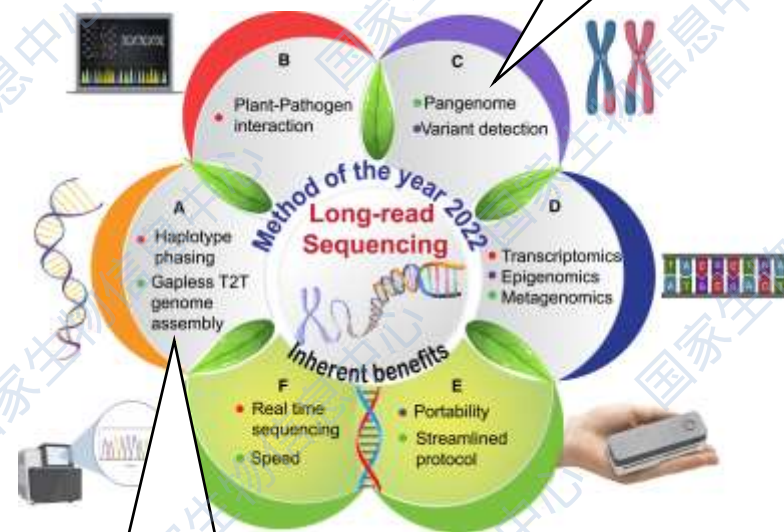
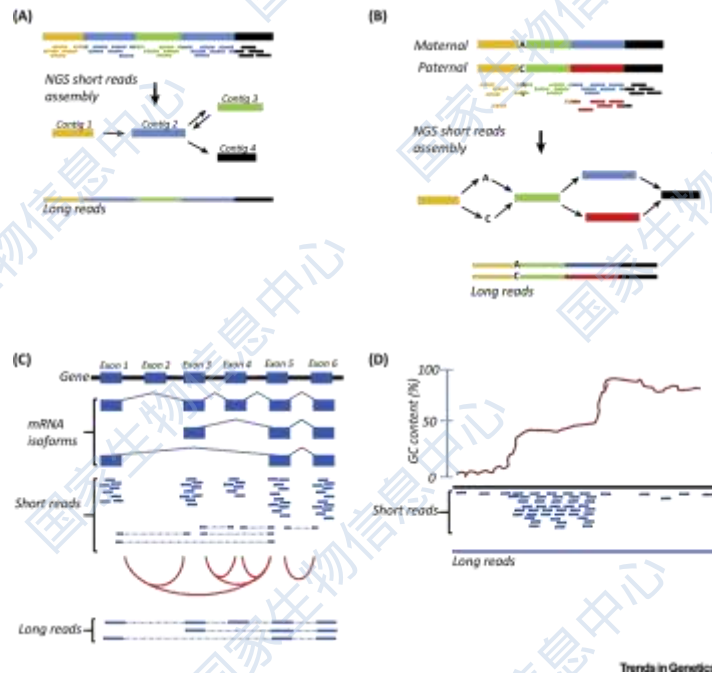


测序技术引领基因组学革命

测序技术是获得高质量基因组的根基

二代测序 (NGS) → 三代测序 (TGS)

测序通量相对于成本大幅提升，可测序更多个体



Long Reads 准确度大幅提升
PacBio HiFi(CC) Reads

Erwin *et al.* 2018. *Trends in Genetics*

Sunil *et al.* 2023. *Molecular Plant*

一些多倍体高杂合基因组组装举例 高精度、长读长测序基本已成为“标配”

Phasing strategy	Species	Karyotype	Sequencing platforms	Tools or strategies	Ref.
Reference-based variant phasing					
	<i>Litchi chinensis</i>	2 n = 2× = 30	Illumina + PacBio + 10× Genomics	HapCUT2	Nat Genet (2022)
	<i>Ipomoea batatas</i>	2 n = 6× = 90	Illumina	Ranbow	Nat Plants (2017)
De novo assembly-based haplotype phasing					
De novo phased contig tools + Hi-C scaffolding	<i>Bletilla striata</i>	2 n = 2× = 32	HiFi + Hi-C	HiFiasm + LACHESIS	Plant J (2022)
	<i>Bupleurum chinense</i>	2 n = 2× = 12	Illumina + HiFi + Hi-C	HiFiasm + 3D-DNA	Front Genet (2022)
	<i>Suaeda glauca</i>	2 n = 2× = 18	HiFi + Hi-C	HiFiasm + 3D-DNA	Front Genet (2022)
	<i>Cynodon dactylon</i>	2 n = 4× = 36	Illumina + PacBio + Bionano + Hi-C	HiFiasm + 3D-DNA	Front Plant Sci (2022)
	<i>Populus tomentosa</i>	2 n = 3× = 57	Illumina + HiFi + Hi-C	HiFiasm + 3D-DNA	Genome Biol (2022)
	<i>Malus domestica</i>	2 n = 2× = 34	Illumina + 10× Genomics + HiFi	HiFiasm + DeNovoMAGIC	Nat Genet (2020)
	<i>Manihot esculenta</i>	2 n = 2× = 36	Illumina + HiFi + Hi-C	HiFiasm + ALLHiC	Gigascience (2022)
	<i>Pogostemon cablin</i>	2 n = 4× = 64	Illumina + PacBio + Hi-C	Canu + 3D-DNA	Nat Commun (2022)
	<i>Manihot esculenta</i>	2 n = 2× = 36	Illumina + PacBio + Hi-C	FALCON + FALCON_unzip + FALCON-Phase	Plant J (2021)
	<i>Humulus lupulus</i>	2 n = 2× = 20	Illumina + PacBio + Hi-C	FALCON + FALCON-unzip	Plant Genome (2021)
	<i>Vanilla planifolia</i>	2 n = 2× = 28	Illumina + ONT	Miniasm + FALCON-Phase + LACHESIS	Nat Food (2020)
	<i>Hydrangea macrophylla</i>	2 n = 2× = 36	Illumina + PacBio + Hi-C	FALCON + FALCON_unzip + FALCON-Phase	DNA Res (2021)
	<i>Zingiber officinale</i>	2 n = 2× = 22	Illumina + PacBio + Hi-C	FALCON-Phase + 3D-DNA	Hortic Res (2021)
	<i>Saccharum spontaneum</i>	1 n = 4× = 32	Illumina + BACs + PacBio + Hi-C	Canu + ALLHiC	Nat Genet (2018)
	<i>Saccharum spontaneum</i>	2 n = 4× = 40	HiFi + Hi-C	Canu + ALLHiC + 3D-DNA	Nat Genet (2022)
	<i>Camellia sinensis</i>	2 n = 2× = 30	HiFi + Hi-C	HiFiasm + ALLHiC	Hortic Res (2021)
	<i>Solanum tuberosum</i>	2 n = 4× = 48	HiFi + Hi-C	HiFiasm + ALLHiC	PBJ (2022)
	<i>Camellia sinensis</i>	2 n = 2× = 30	Illumina + PacBio + Hi-C	Canu + ALLHiC	Nat Genet (2021)
	<i>Manihot esculenta</i>	2 n = 2× = 36	PacBio + Hi-C	Canu + Wtdbg + ALLHiC	Mol Plant (2021)
	<i>Dendrocalamus latiflorus</i>	2 n = 6× = 70	Illumina + PacBio + Hi-C	Falcon + ALLHiC	JIPB (2022)
	<i>Medicago sativa</i>	2 n = 4× = 32	Illumina + HiFi + Hi-C	Canu + ALLHiC	Nat Commun (2020)
	<i>Medicago sativa</i>	2 n = 4× = 32	PacBio + Bionano + Hi-C	Canu + MECAT + ALLHiC	Mol Plant (2020)
	<i>Medicago sativa</i>	2 n = 4× = 32	PacBio + Hi-C	Canu + ALLHiC	GPB (2022)
	<i>Artemisia annua</i>	2 n = 2× = 18	Illumina + PacBio + Bionano + Hi-C	Canu + HiFiasm + FALCON + LACHESIS	Mol Plant (2022)
Trio-binning-based de novo phasing	<i>Ananas comosus</i>	2 n = 2× = 50	PacBio + ONT + Hi-C + Illumina	Trio-binning	Plant J (2022)
	<i>Cerasus × kanzakura</i>	2 n = 2× = 16	Illumina + PacBio	Trio-binning	DNA Res (2021)
	<i>Cerasus × yedoensis</i>	2 n = 2× = 16	Illumina + PacBio	Trio-binning	DNA Res (2019)
Genetic map-based de novo phasing	<i>Pyrus bretschneideri</i>	2 n = 2× = 34	BACs + Single Cell Sequencing	Gamete binning	Genome Res (2019)
	<i>Solanum tuberosum</i>	2 n = 4× = 48	HiFi + SCS	Gamete binning	Nat Genet (2022)
	<i>Camellia sinensis</i>	2 n = 2× = 30	SCS	Gamete binning	Plant J (2021)
	<i>Prunus armeniaca</i>	2 n = 2× = 16	PacBio + SCS	Gamete binning	Genome Biol (2020)
	<i>Solanum tuberosum</i>	2 n = 2× = 24	ONT + 10× Genomics + HiFi + Hi-C + Illumina for self-population	Population resequencing	Nat Genet (2020)
	<i>Solanum tuberosum</i>	2 n = 4× = 48	HiFi + Hi-C + Illumina for self-population	Population resequencing	Mol Plant (2022)
	<i>Vitis riparia</i>	2 n = 2× = 38	Illumina + PacBio + 10×Genomics + GBS data	Population resequencing	Sci Data (2019)
	<i>Zoysia japonica</i>	2 n = 4× = 20	PacBio + GBS data	Population resequencing (PolyGembler)	Nat Genet (2020)

三大作物基因组学领域的重要进展

This timeline illustrates the significant milestones in crop genomics, focusing on three major crops: rice, maize, and wheat. The timeline is marked with years from 2002 to 2020. Key events include the release of the first draft genomes (草图) and high-resolution maps (精细图) for these crops. Major genomic projects like the 1000 Rice Genomes Project (谱系(3)), the Maize Genome Project (谱系(40)), and the Wheat Genome Project (谱系(44)) are highlighted. The timeline also shows the application of various genomic technologies, including GWAS (Genome-Wide Association Studies), QTL-seq (Quantitative Trait Locus sequencing), and CRISPR-Cas9 gene editing. The timeline is divided into three main sections: 454 high-throughput sequencing (2005-2015), Illumina high-throughput sequencing (2007-), and plant CRISPR technology (2013-). The timeline is color-coded: blue for the first section, red for the second, and green for the third.

400 Mb
2000 Mb
17000 Mb
三大作物首次应用的新技术

草图 精细图

2002 2005年 2007 2009 2010年 2012 2013 2014 2015年 2016 2017 2018 2019 2020年

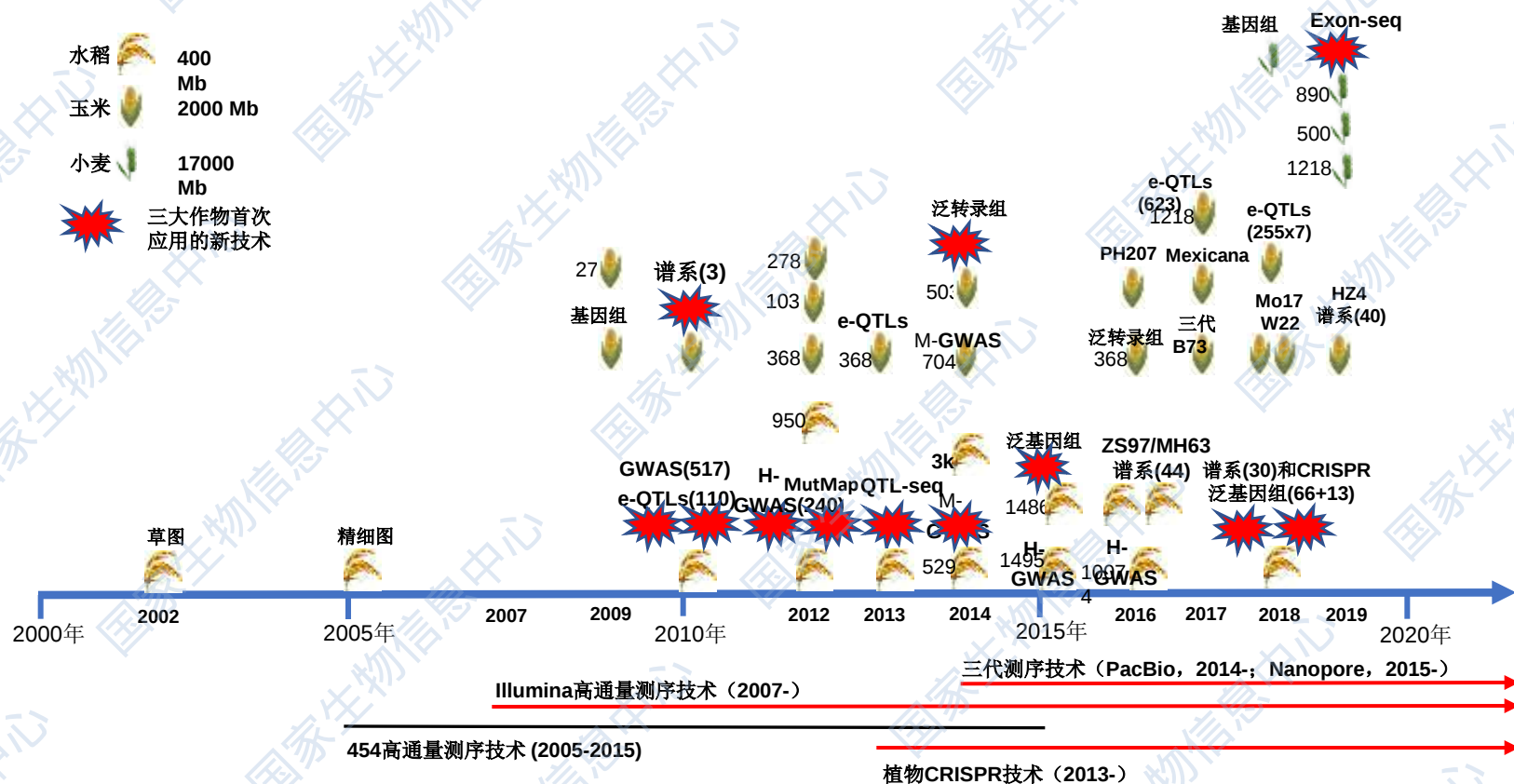
基因组 谱系(3) 278 103 368 368 50k 泛转录组 M-GWAS 704 泛基因组 泛转录组 368 PH207 Mexicana e-QTLs (623) 1218 e-QTLs (255x7) 890 500 1218 三代 B73 Mo17 W22 HZ4 谱系(40) 950 GWAS(517) e-QTLs(110) H-GWAS(240) MutMap QTL-seq 3k M-GWAS 529 泛基因组 谱系(44) 谱系(30)和CRISPR 泛基因组(66+13) 1486 1495 1007 H-GWAS 4

454高通量测序技术 (2005-2015)

Illumina高通量测序技术 (2007-)

三代测序技术 (PacBio, 2014-; Nanopore, 2015-)

植物CRISPR技术 (2013-)



基因组从头测序的三个阶段

Genome Survey阶段:

基因组大小

杂合度高低、样品污染

GC含量评估

重复序列分布

基因组框架图阶段:

基因组/转录组等测序数据累积

基因组拼接及验证

基因组基本组成（GC含量、重复序列等）

基因预测和功能注释等

进化分析（基因组内和物种间）

基因组精细图阶段阶段:

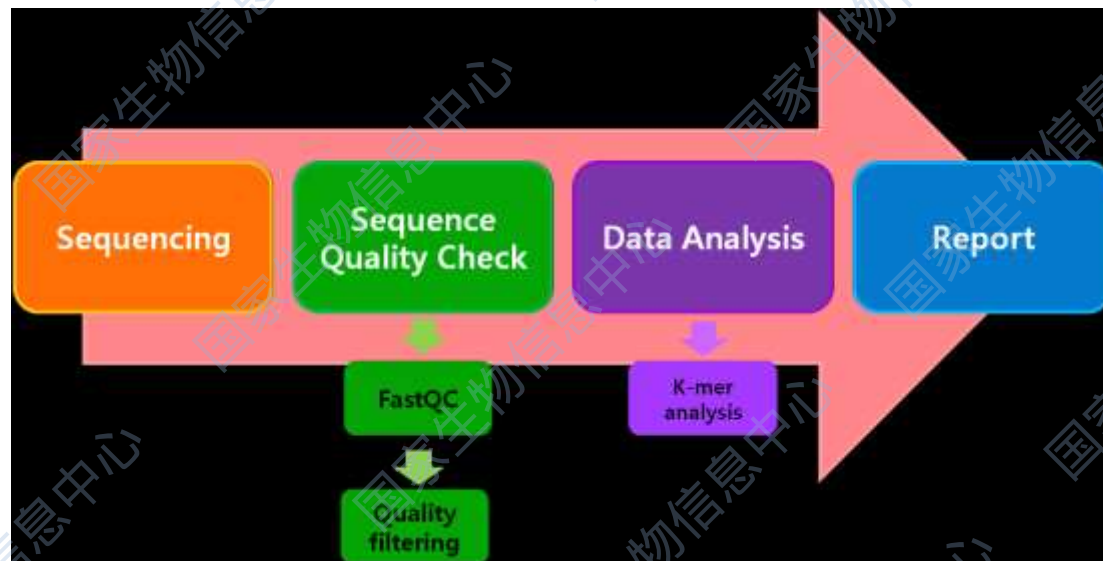
大片段文库（BAC/Fosmid）

物理图谱（BioNano/OpGen）

遗传图谱（GBS）

染色体定位（FISH）

Genome Survey流程



数据：20X以上短读长Pair-end序列

质控：FastQC (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc>)

分析：JELLYFISH (<http://www.cbcb.umd.edu/software/jellyfish/>)

Genome size estimation

- Flow cytometric analysis
- Kmer-distribution

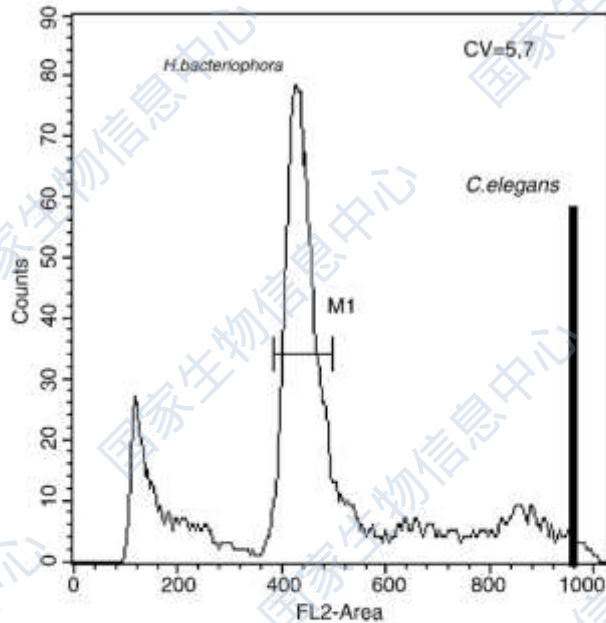


Fig. 1. Histogram of fluorescence intensity acquired by means of an analyser connected to computer for *Heterorhabditis bacteriophora*. The cytometer argon laser was adjusted at 488 nm at 20 mW in potential and the measurement was realised at 585 nm. The flux cytometer is calibrated in reference to the genome size of *C. elegans* (in bold line). An entire specimen was crushed in a solution of PBS Tween and was lysed by addition of Triton X-100 at a final concentration of 0.2%. This preparation was stained with propidium iodide at a final concentration of $10 \mu\text{g}^{-1}$. DNA content was measured immediately before vortex pellet.

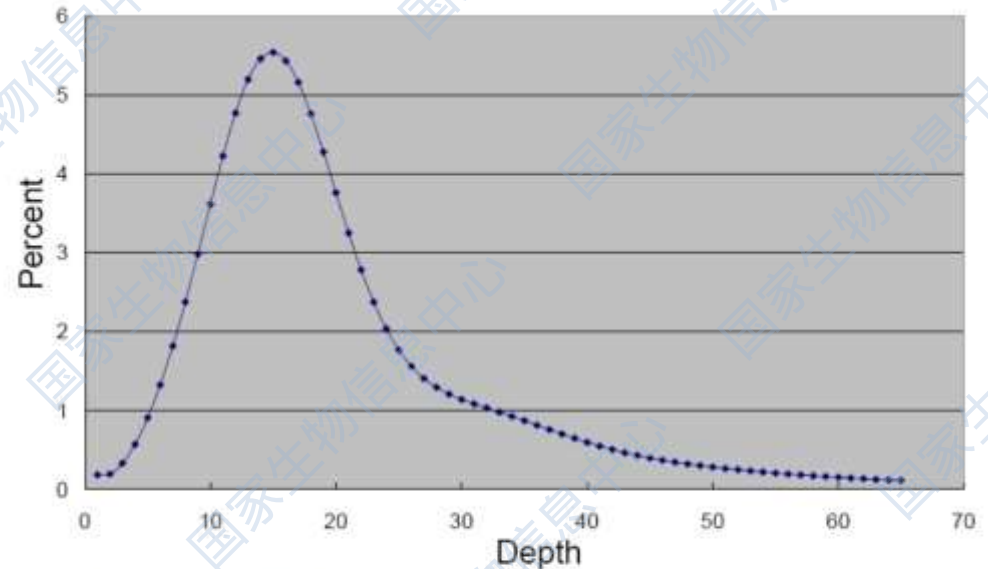
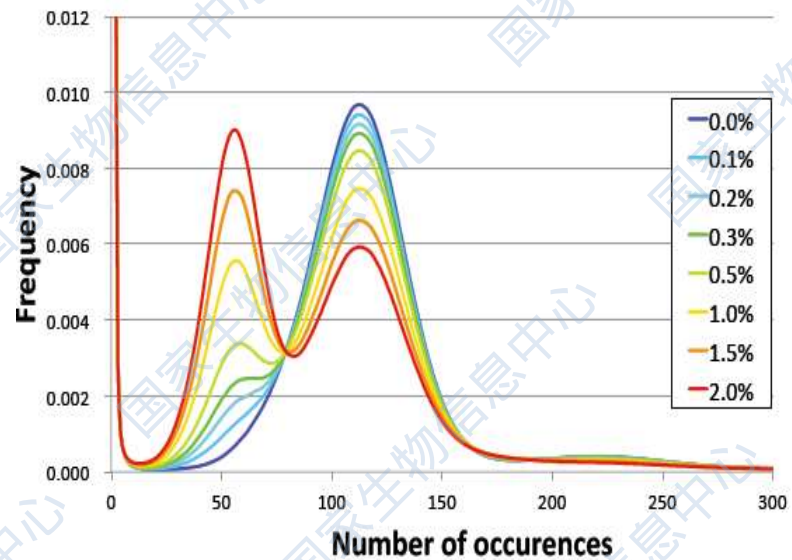
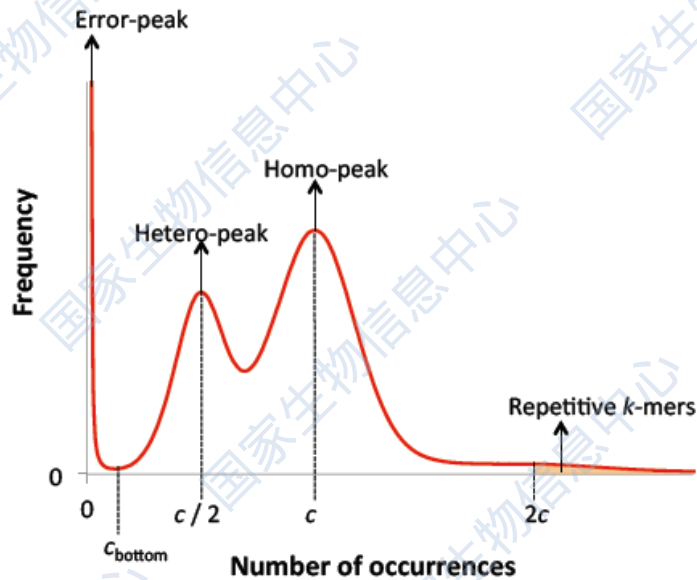


Figure S8 | Distribution of 17-mer frequency in the raw sequencing reads. We used all reads from the short insert-size libraries (<500bp). The peak depth is at 15X. The peak of 17-mer frequency (M) in reads is correlated with the real sequencing depth (N), read length (L), and kmer length (K), their relations can be expressed in a experienced formula: $M = N * (L - K + 1) / L$. Then, we divided the total sequence length by the real sequencing depth and obtained an estimated the genome size of 2.46 Gb.

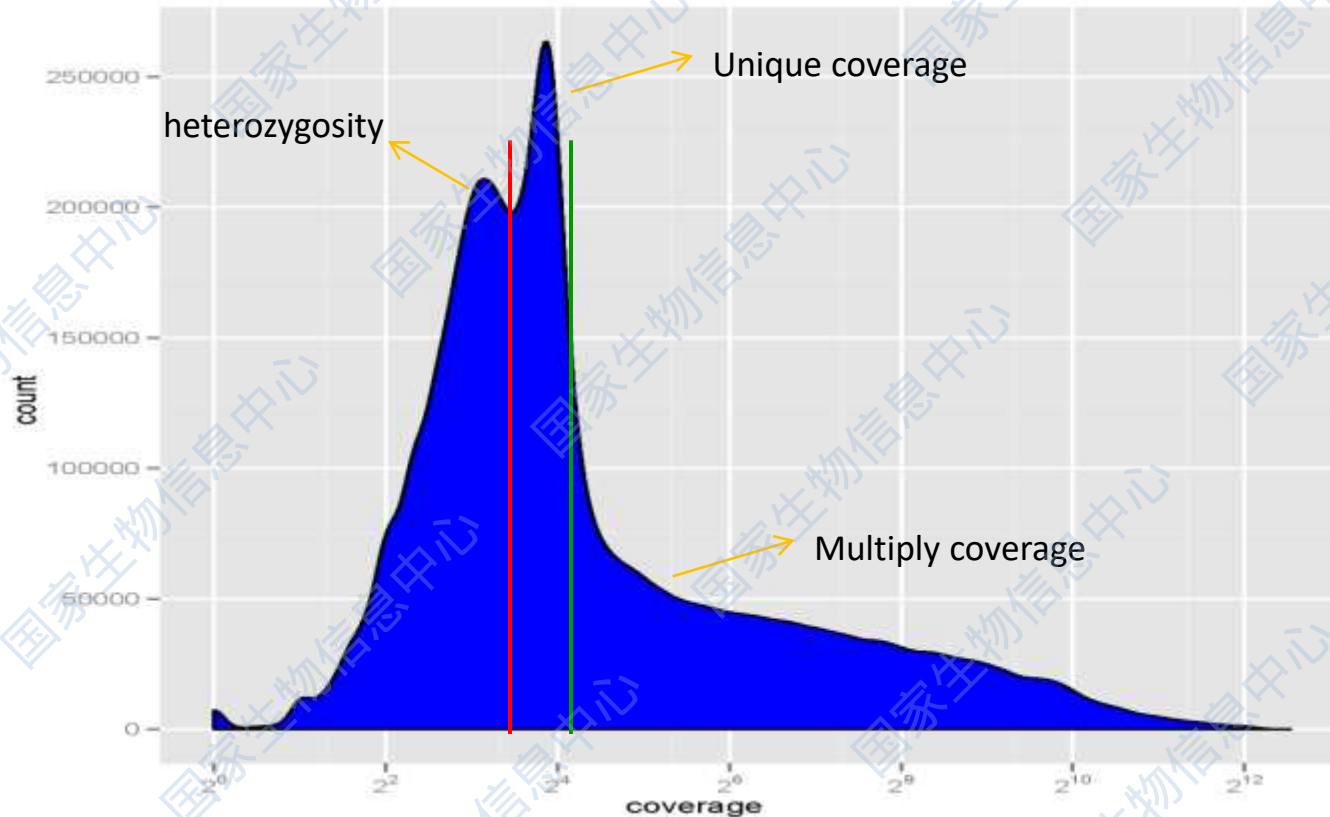
基因组拼接两大挑战：杂合与重复

De Bruijn graph



[Genome Res.](#) 2014 Aug;24(8):1384-95.

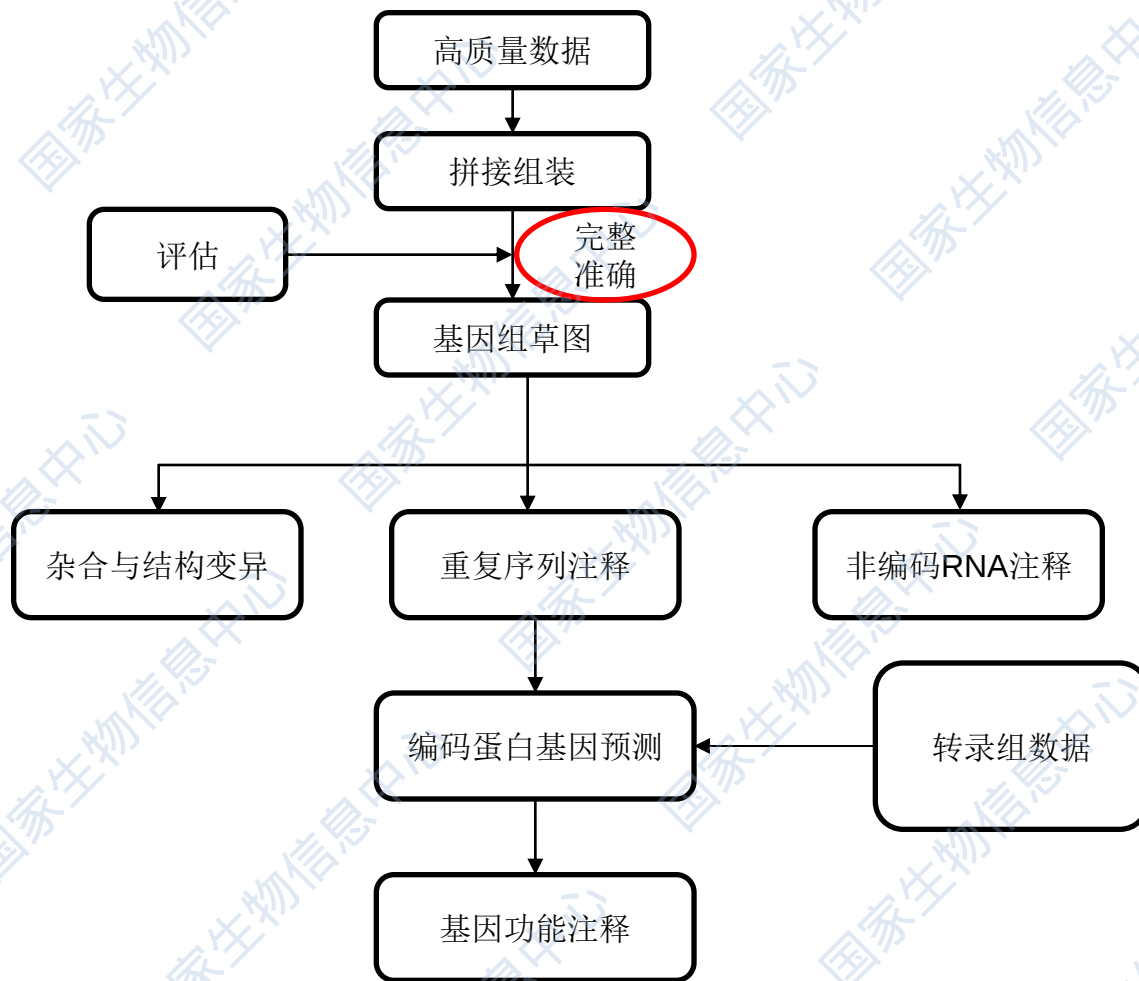
Kmer Distribution



$$104212972 \times 0.5 + 438982154 + 171206668 = 662.30 \text{ Mb.}$$

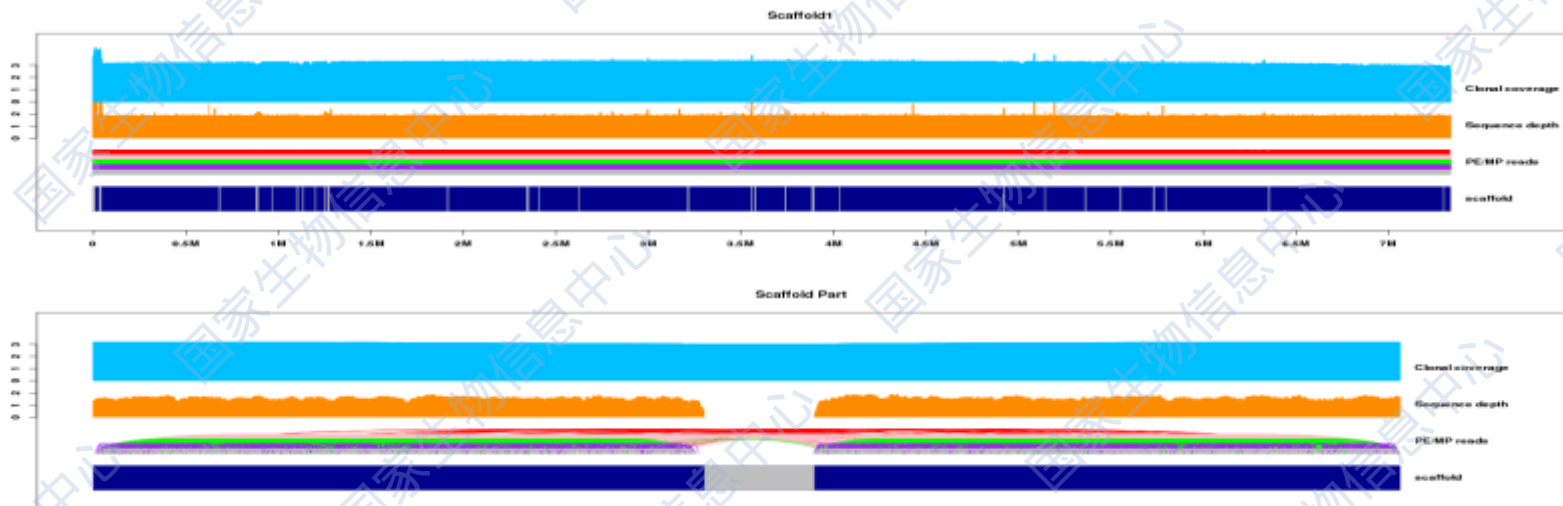


基因组基本分析

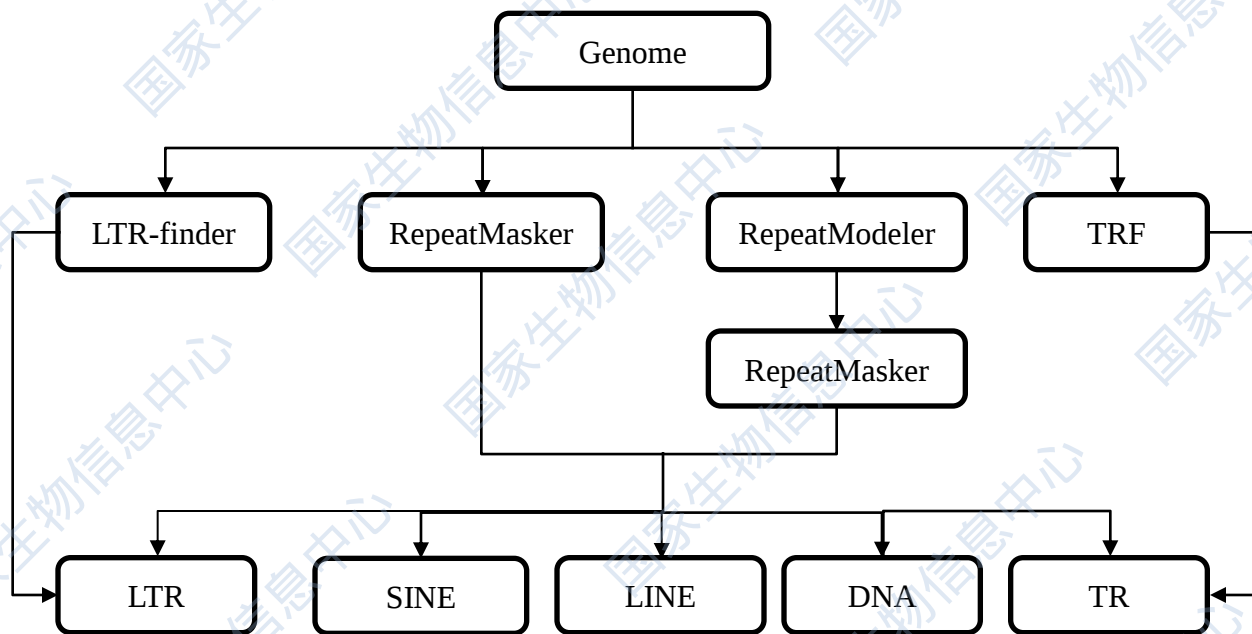


基因组拼接评估

- (1) N50/N90
- (2) Read
- (3) RNA_seq
- (4) EST/cDNA
- (5) CEGMA(Core Eukaryotic Genes Mapping Approach)
- (6) BUSCO(Benchmarking Universal Single-copy Orthologs)



基因组重复



Repetitive sequences

Supplementary Table 4. Distribution of repeats in *S. japonicum*

Repeat Family	Size (bp)	Ratio (%)
Low_complexity	2,011,687	0.51
Microsatellite	2,520,388	0.63
Minisatellite	9,815,590	2.47
Satellite	16,830,062	4.23
SINE		
<i>Sj-alpha1</i>	4,293,863	1.08
<i>Sj-alpha2</i>	3,011,537	0.76
LTR Retrotransposon		
<i>Gulliver</i>	711,318	0.18
18 Novel Sj-Retrotransposons	23,803,907	5.98
Non-LTR Retrotransposon		
<i>Sj-pido</i>	1,540,514	0.39
<i>SjR1</i>	3,358,700	0.84
<i>SjR2</i>	17,602,429	4.43
4 Novel Sj-Retrotransposons	27,743,061	6.97
3 <i>Penelope</i> -like Retrotransposons	3,938,326	0.99
Transposon	856,671	0.22
undefined repeat	41,582,296	10.45
Total repeat	159,620,349	40.13
Genome Size*	397,751,405	100

基因组ncRNA

常规ncRNA:

Rfam (<http://rfam.xfam.org/>, rfam_scan.pl)

tRNA (tRNAscan-SE)

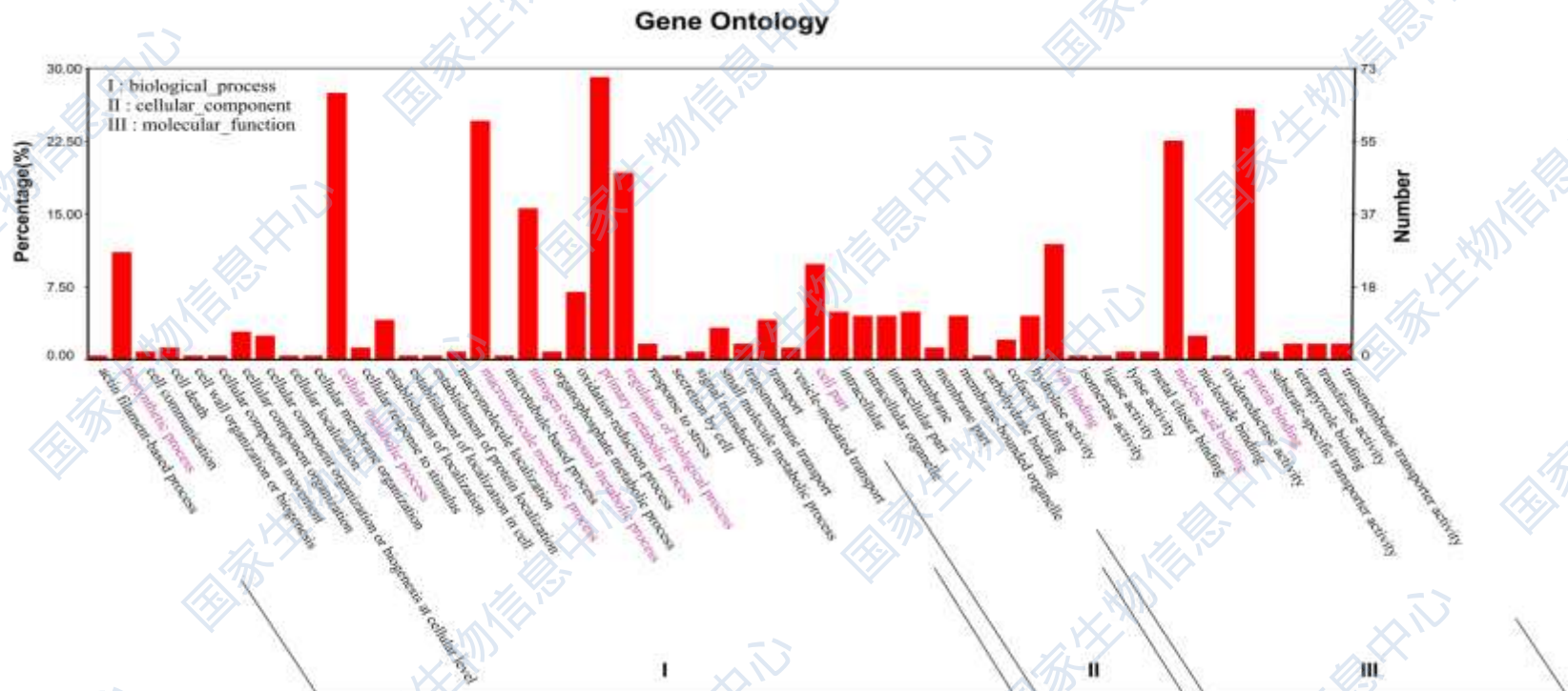
miRNA (<http://www.mirbase.org/>, ViennaRNA)

rRNA



MicroRNA and SnoRNA Prediction

- 102 date palm microRNAs (miRNAs) from 49 families, 580 genes targeted by 90 miRNAs were identified

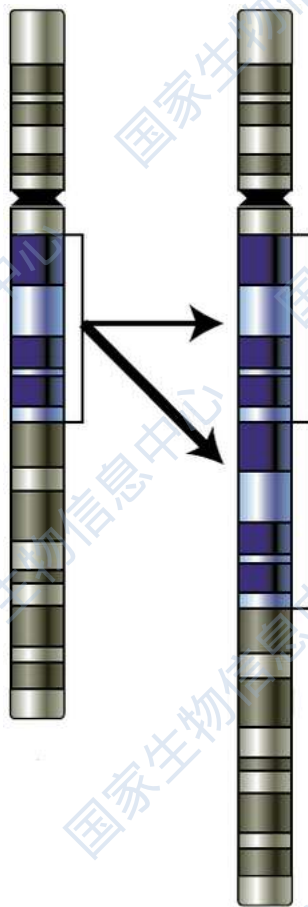


- 44 box H/ACA snoRNA genes (Table S6) and 65 box C/D snoRNA genes

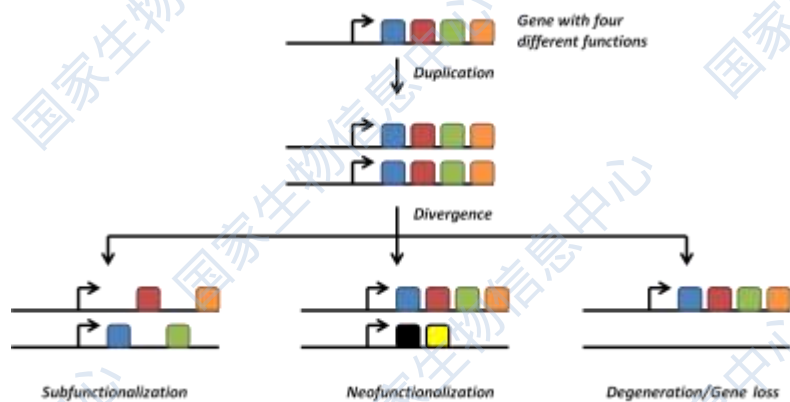
SSR identification

	n=6-11		n>11	
	% GC	% dataset	% GC	% dataset
Mononucleotides	7.88	1.76	13.23	0.09
Dinucleotids	33.25	0.14	30.23	0.11
Trinucleotides	35.67	0.04	13.18	0.01
Tetranucleotides	21.30	0.01	27.57	0.00
All periods		1.94		0.21

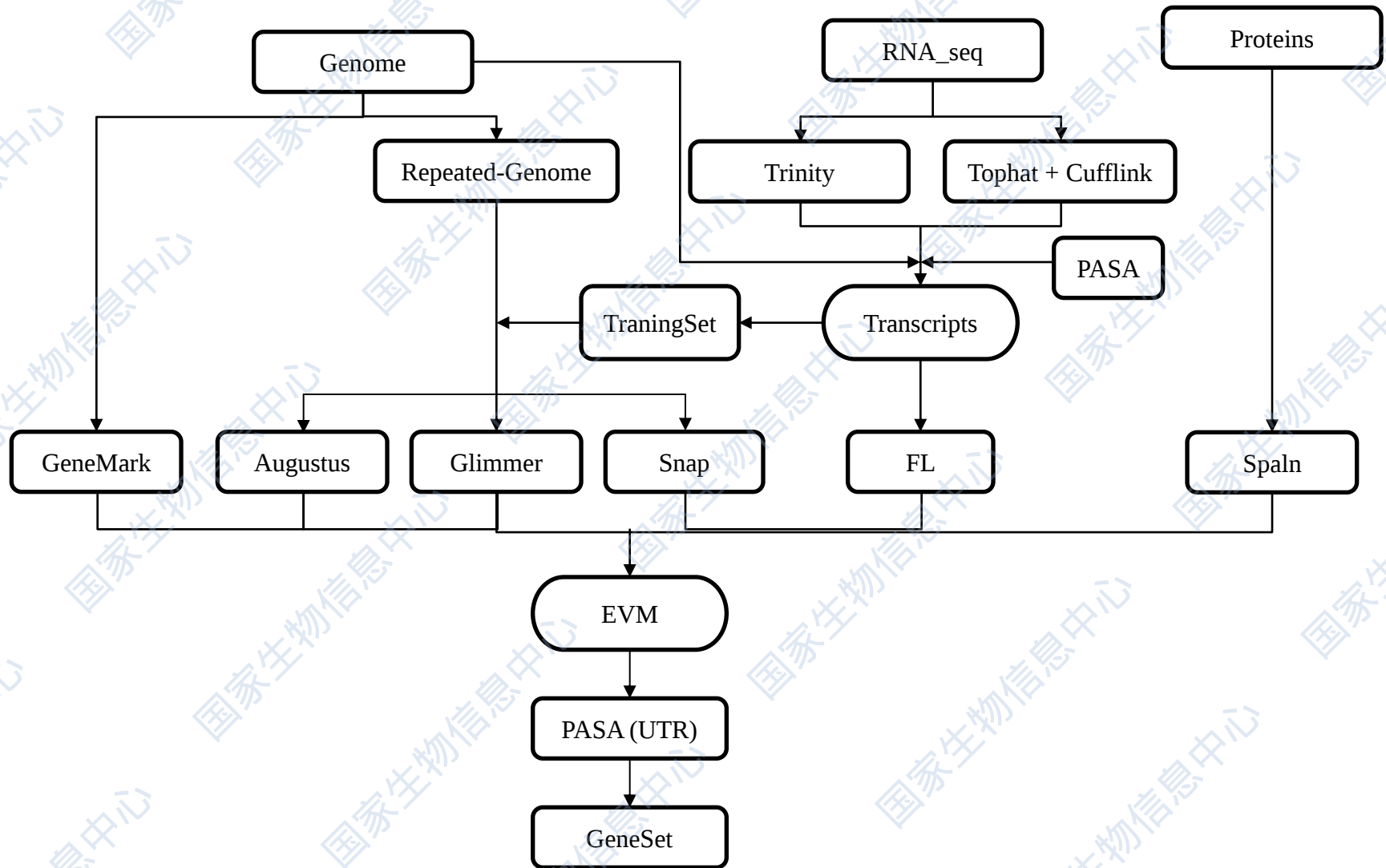
大片段扩增与基因重复



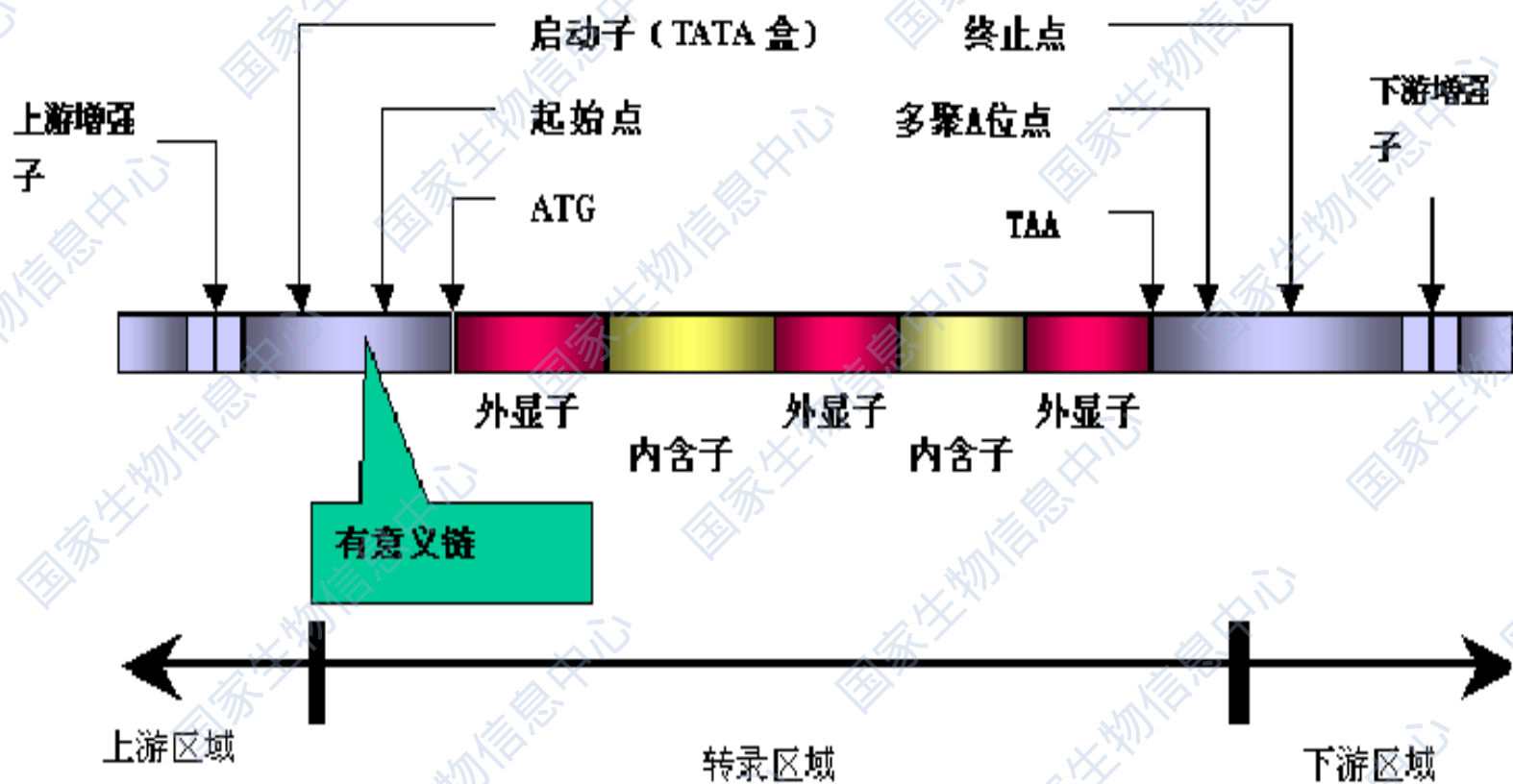
重组
复制滑移
反转座
非整倍扩增
全基因组倍增



编码蛋白预测与功能注释



真核生物基因的主要结构



基因结构分析常用软件

基因结构分析	开放读码框	GENSCAN
		GENOMESCAN
	CpG岛	CpGPlot
	转录终止信号	POLYAH
	启动子/转录起始位点	PromoterScan
	密码子偏好分析	CodonW
	mRNA剪切位点	NETGENE2
		Spidey
	选择性剪切	ASTD

基因开放阅读框/基因结构分析识别工具

ORF Finder	http://www.ncbi.nlm.nih.gov/gorf/gorf.html	NCBI	通用
BestORF	http://linux1.softberry.com/berry.phtml?topic=bestorf&group=programs&subgroup=gfind	Softberry	真核
GENSCAN	http://genes.mit.edu/GENSCAN.html	MIT	脊椎、拟南芥、玉米
Gene Finder	http://rulai.cshl.org/tools/genefinder/	Zhang lab	人、小鼠、拟南芥、酵母
FGENESH	http://linux1.softberry.com/berry.phtml?topic=fgenesh&group=programs&subgroup=gfind	Softberry	真核(基因结构)
GeneMark	http://opal.biology.gatech.edu/GeneMark/eukhmm.cgi	GIT	原核
GLIMMER	http://www.ncbi.nlm.nih.gov/genomes/MICROBES/glimmer_3.cgi http://www.cbcb.umd.edu/software/glimmer	Maryland	原核
Fgenes	http://linux1.softberry.com/berry.phtml?topic=fgenes&group=programs&subgroup=gfind	Softberry	人(基因结构)
FgeneSV	http://linux1.softberry.com/berry.phtml?topic=virus&group=programs&subgroup=gfindv	Softberry	病毒
Generation	http://compbio.ornl.gov/generation/	ORNL	原核
FGENESB	http://linux1.softberry.com/berry.phtml?topic=fgenesb&group=programs&subgroup=gfindb	Softberry	细菌(基因结构)
GenomeScan	http://genes.mit.edu/genomescan.html	MIT	脊椎、拟南芥、玉米
GeneWise2	http://www.ebi.ac.uk/Wise2/	EBI	人
GRAIL	http://grail.lsd.ornl.gov/grailexp/	ORNL	人、小鼠、拟南芥、果蝇

基因组注释

Gene Prediction

ORF Finder

GENESCAN

GENEMARK

...

S
e
q
u
e
n
c
e

Predicted Gene
Or Gene

Blastn

Fasta

Homology
Search

Transcription
Regulatory
Region

Protein Annotation

Physical&Chemical Para
(PI/MW,EXTCOEF)

Domain Identify
(HMMER,BLIMPS)

Transmembrane
(TMAP,TMHMM)

Localization Sites
(Psort)

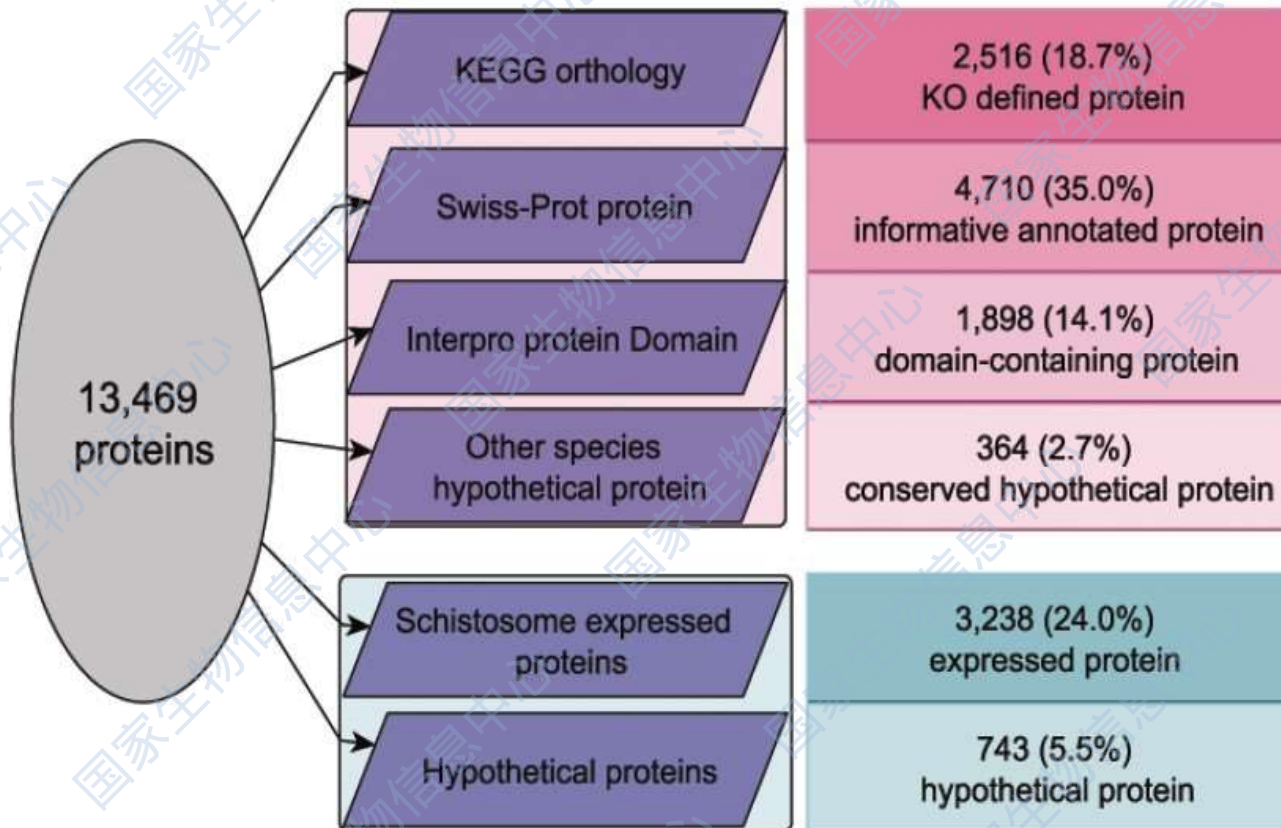
Post-translational
modifications
(NetNGlyc...)

Gene Ontology

Pathway

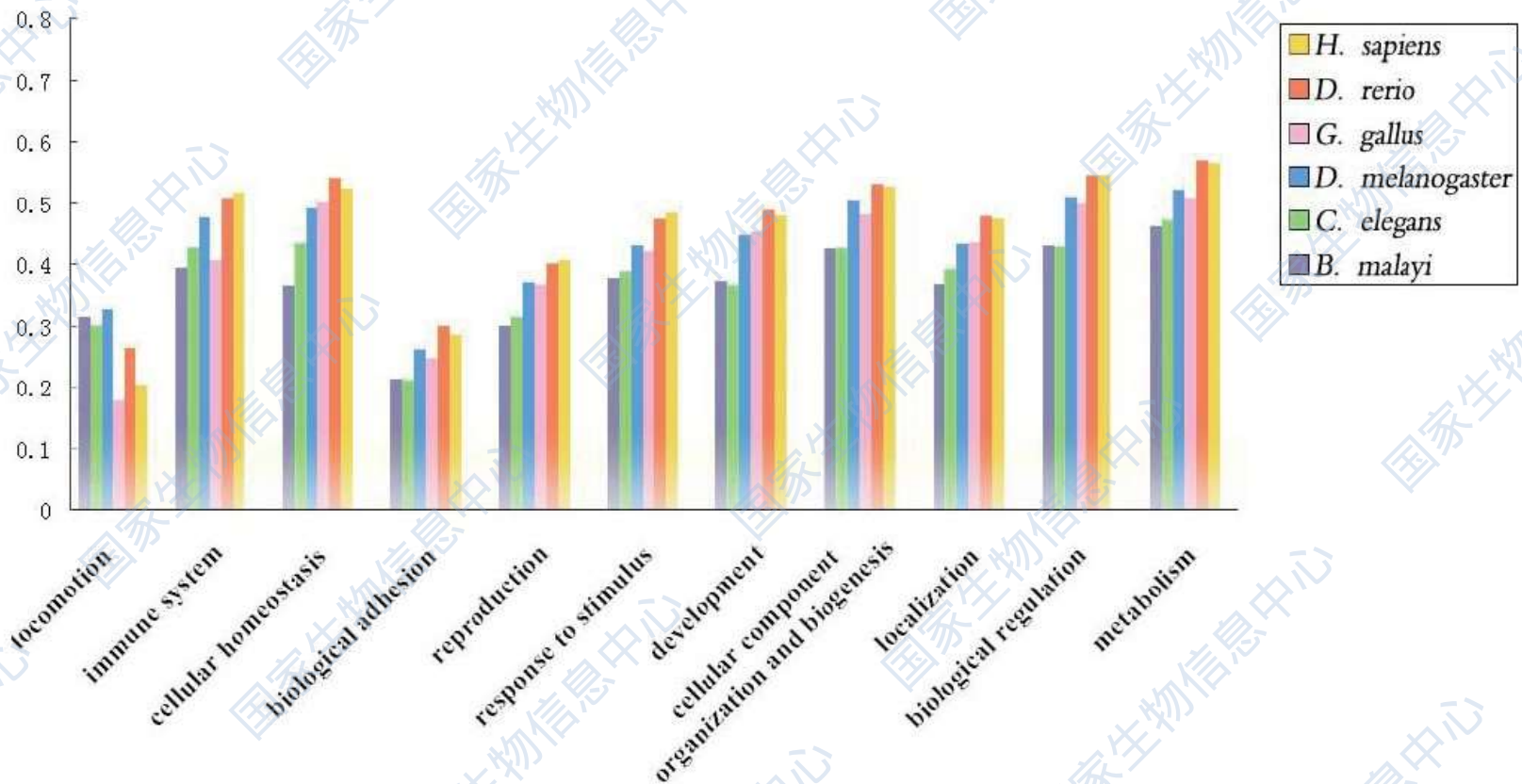
...

Gene prediction



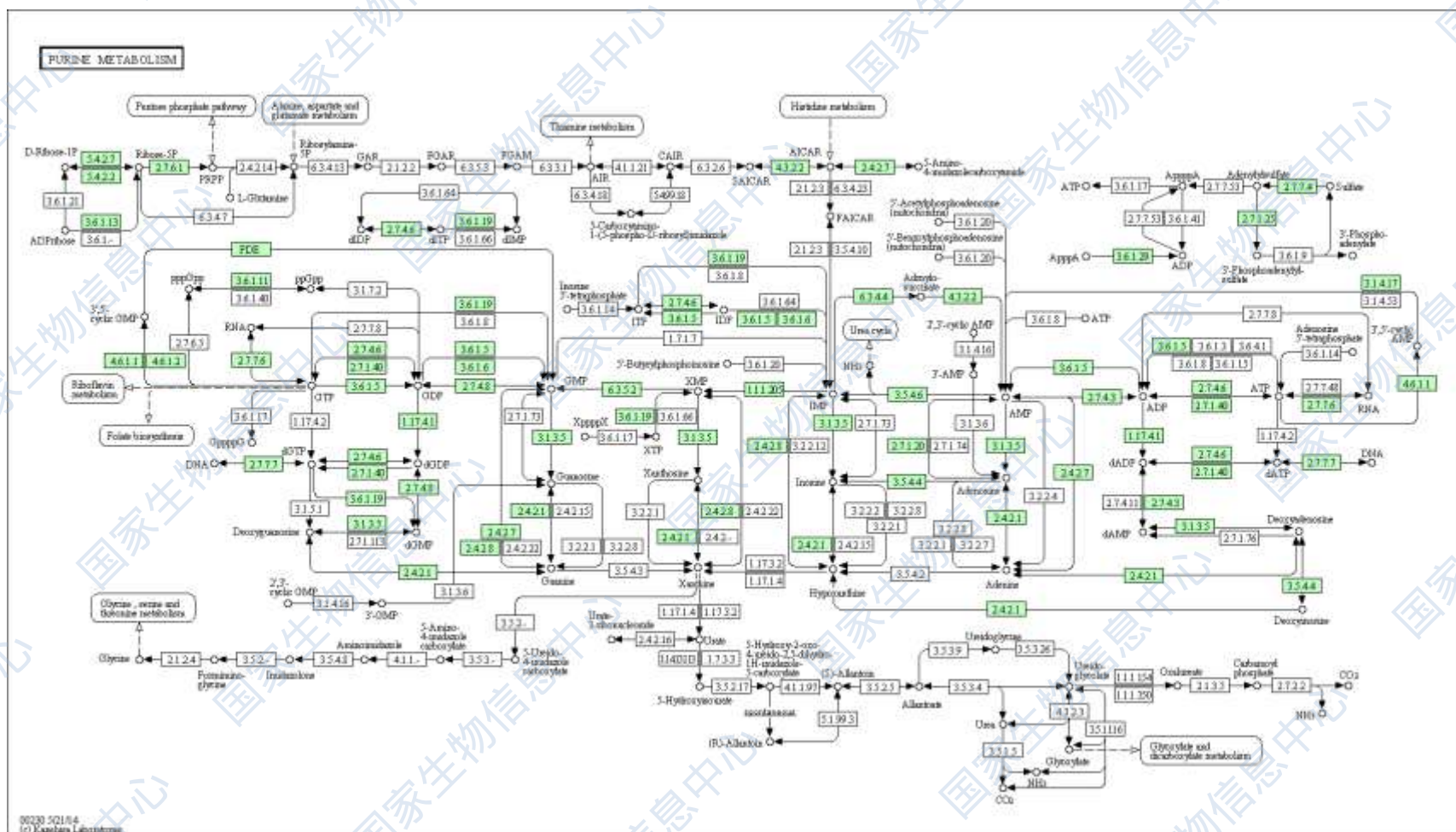
Supplementary Figure 5. Stepwise annotation of *S. japonicum* genes

Gene function comparison

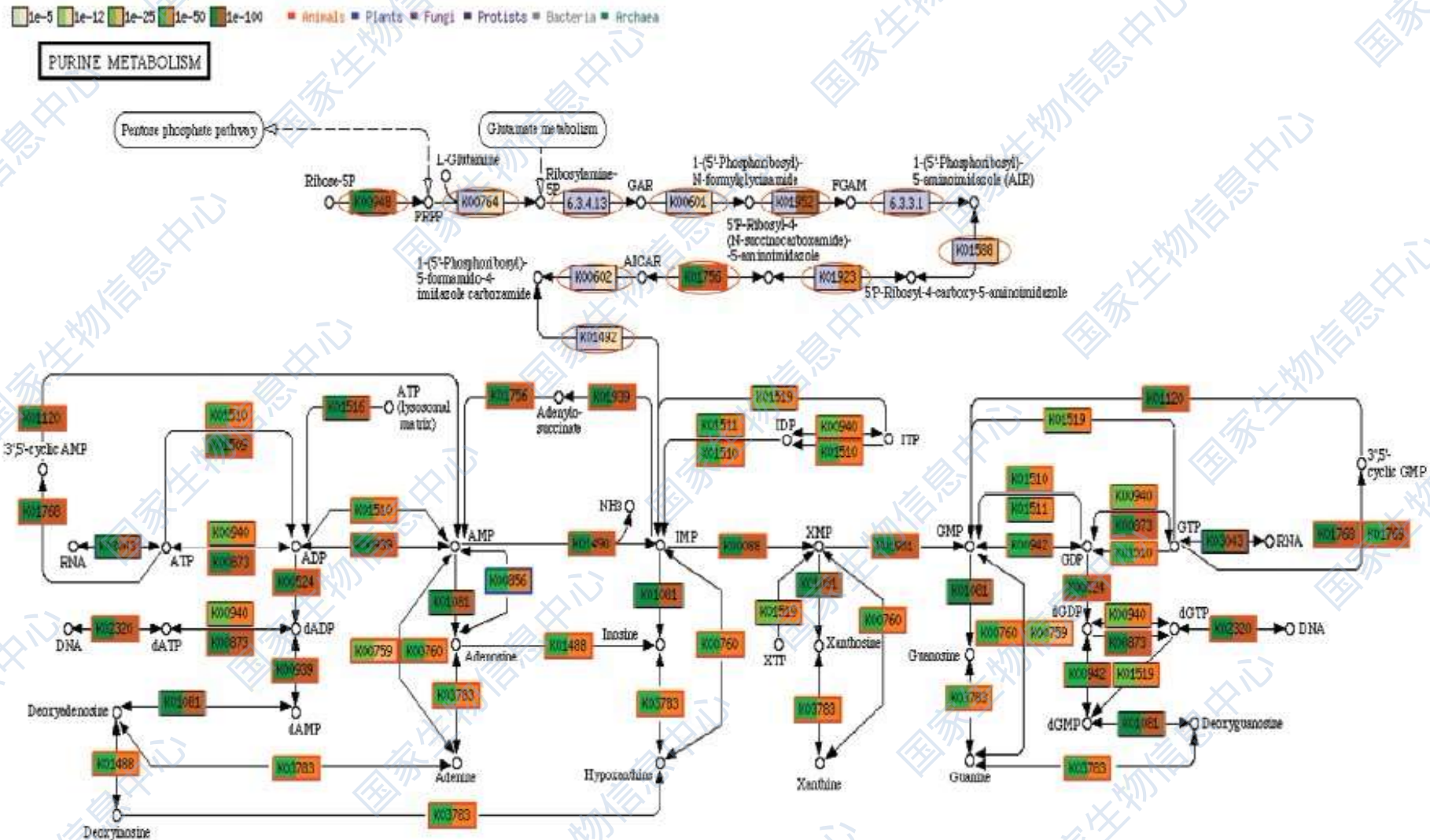


Supplementary Figure 4. Functional categorization of *S. japonicum* genes and comparative analysis with other representative organisms

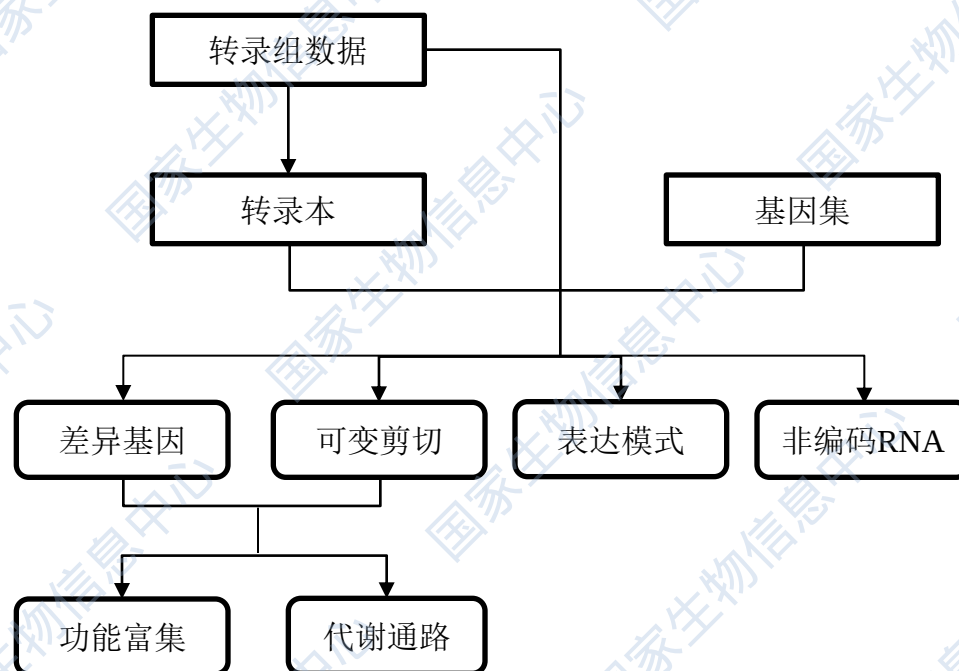
编码蛋白预测与功能注释



Species specific pathway and function



转录组分析



- RNAseq is a powerful tool to detect the whole transcriptome in cell and tissue.
- Previous RNAseq research focus on mRNA, but recent studies prove that part of functional noncoding transcript and protein-coding RNAs are lack of polyA.

Content of transcriptome

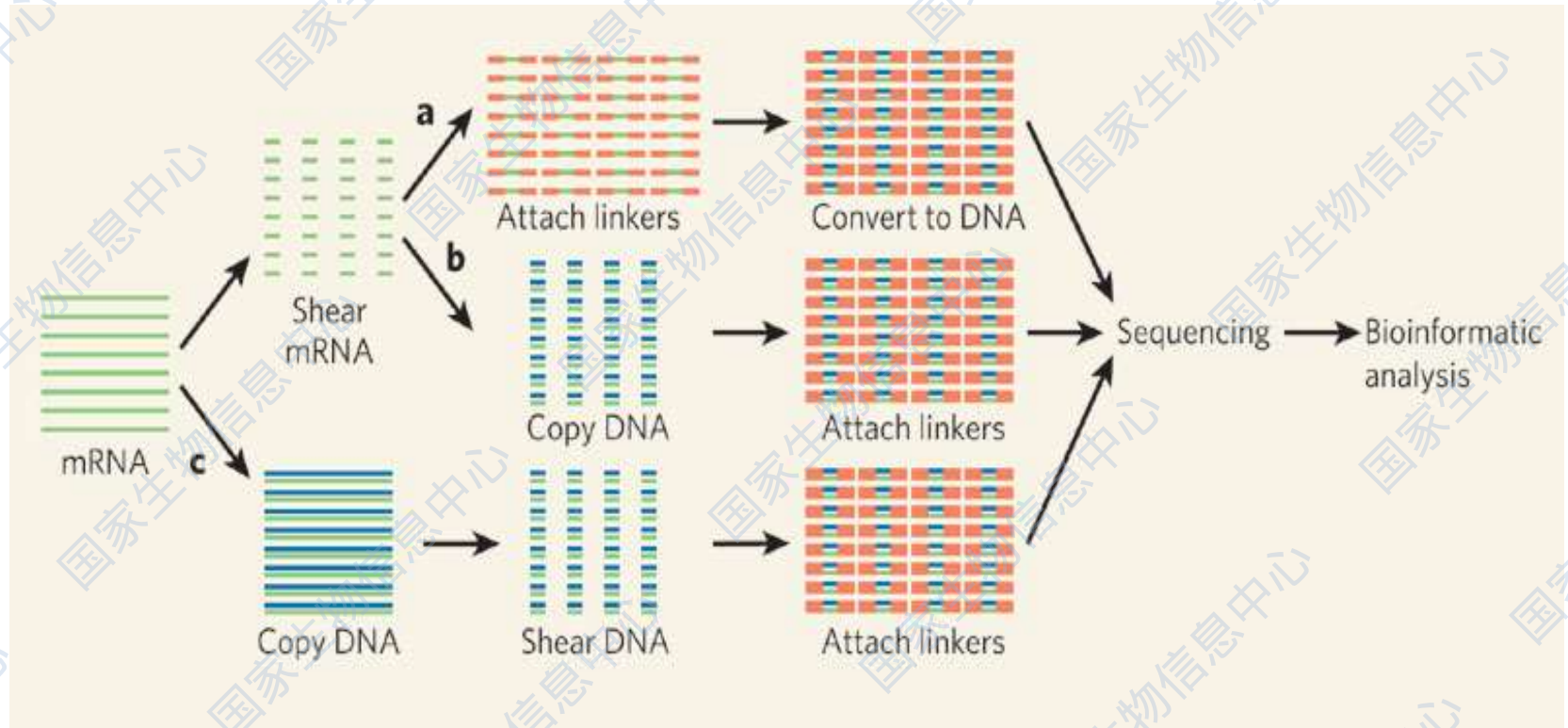
1. Genes: expression , alterante splices
2. Noncoding RNA:
snoRNA, mRNA-like ncRNA, snRNA,
some antisense transcripts,
pesudogenes, retrotransposon ,and
others functional RNAs
3. Some repeat elements

RNA-seq的生物学重复和标准

1. 至少有两个生物学重复，除非“短时间梯度取样”
(overlapping time points with high temporal resolution)
不需要技术重复
2. 对基因注释较好的物种，只定量比较研究，可用reads大于20M；用于注释基因组的转录组，大于>100M
3. 最好有浓度不同长度不同的绝对定量control (Spike-in)，以评估mapping质量、测序均匀性和RNA-seq定量效果
4. “3端/5端比值”是衡量RNA完整性的关键指标(理想值是1)，也要进行计算评估
5. 样品处理流程，文库构建流程，测序机器，测序类型，分析软件，样品评估关键指标，rpkm值关键结果完备。

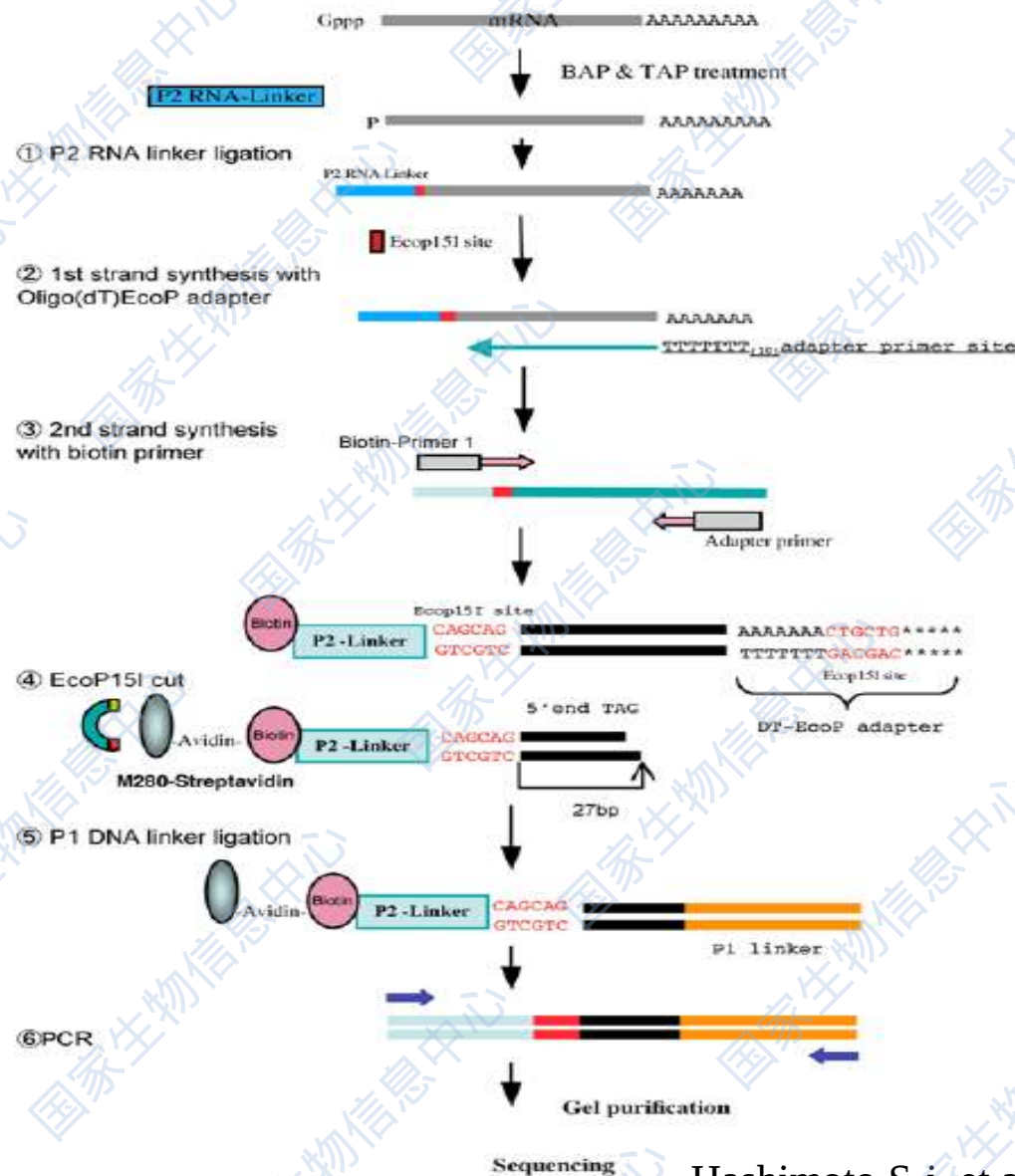
Background

mRNA-seq



Background

SAGE



rRNA depletion

Total RNA

Ribosomal RNA Depletion

18S RNA

28S RNA

rRNA-depleted RNA

Fragmentation of RNA

Ligation to Adaptor

RT and RNaseH Digestion

cDNA

cDNA Amplification

P1 Primer

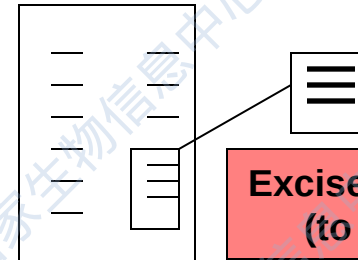
P2 Primer

Size Selection by PAGE

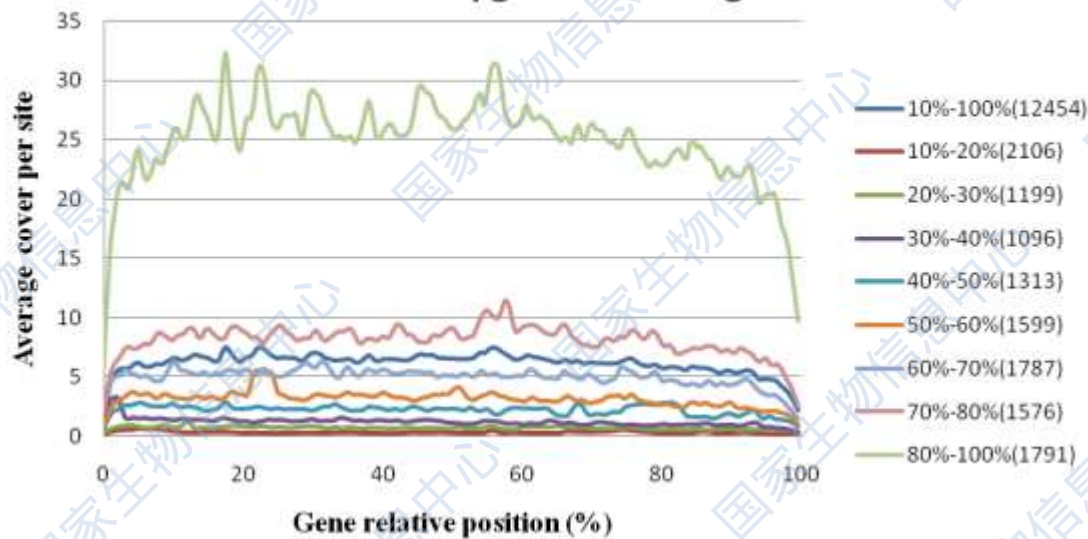
Excise cDNA ~140-200bp
(to RNA ~50-110nt)

SOLiD emulsion PCR

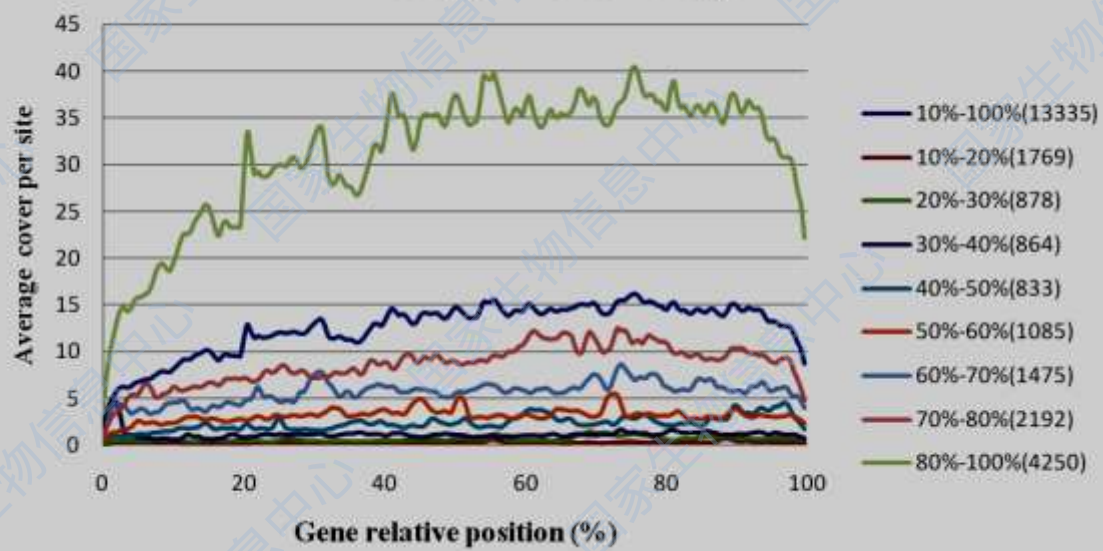
SOLiD Sequencing



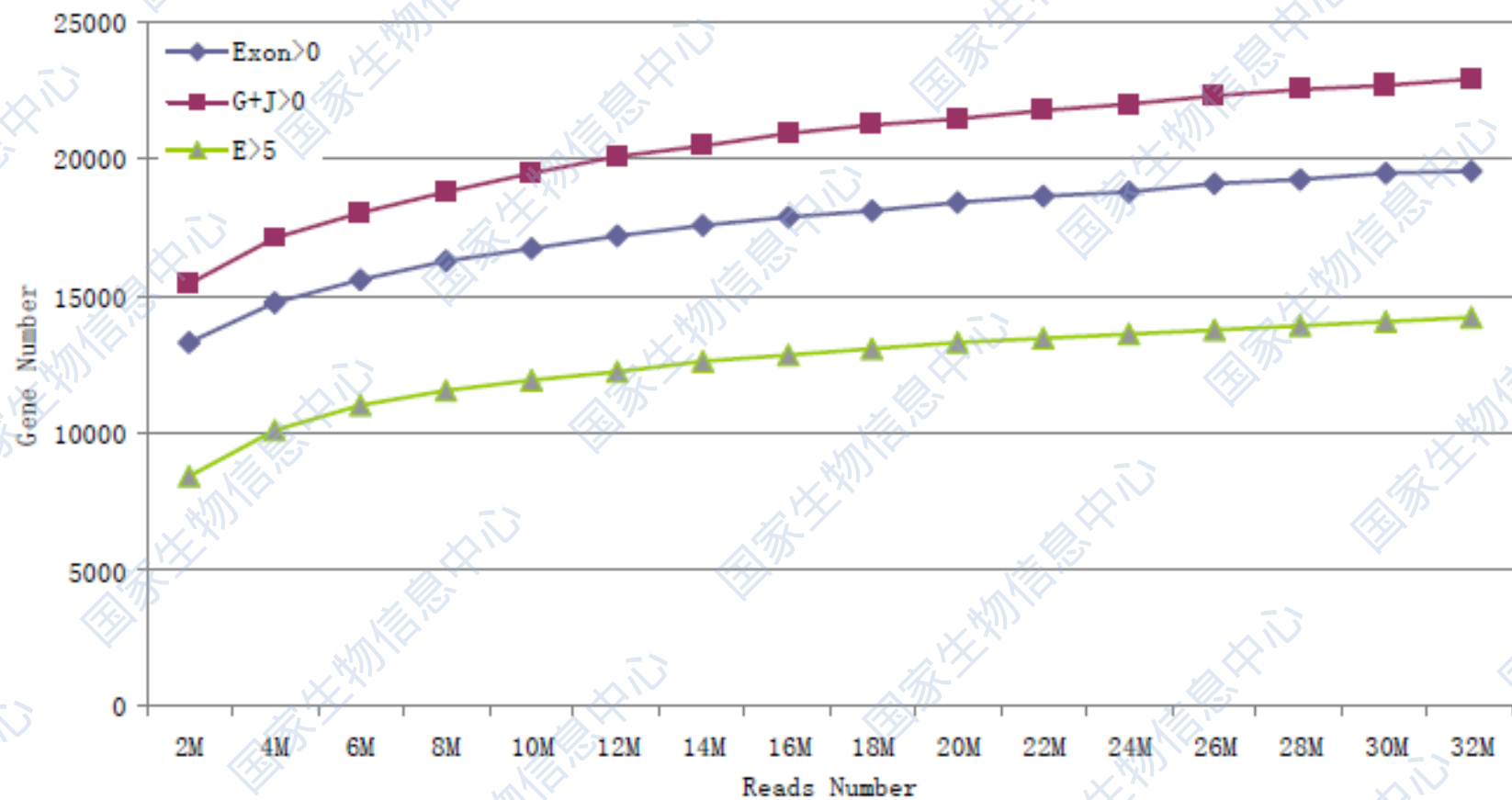
Ribominus RNA-seq genes coverage



mRNA-seq genes coverage



测序饱和度



Gene expression normalization

Fragment Reads:

RPKM: quantified transcript levels in reads per kilobase of exon model per million mapped reads

RPKM统计的是每个基因中的read个数

Pair-end Reads:

FPKM: quantified transcript levels in fragment reads per kilobase of exon model per million mapped reads

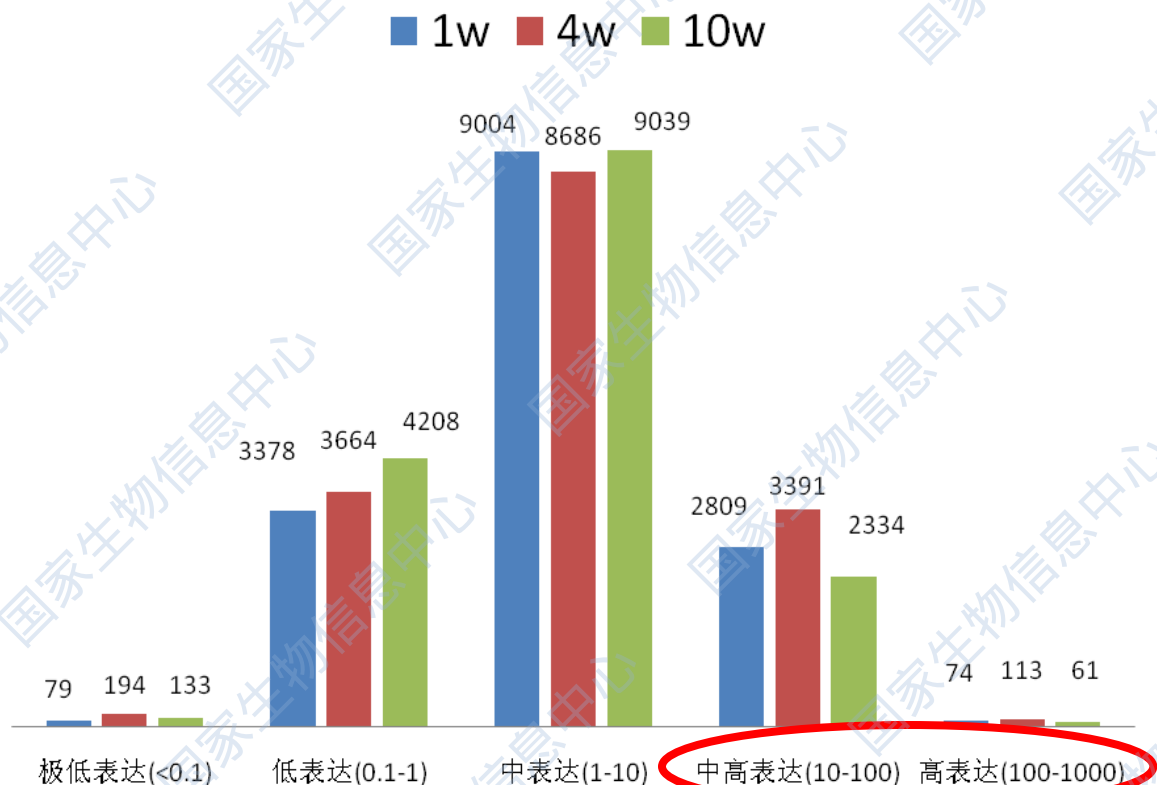
(fragment is a pair of reads for paired end data)

FPKM统计的是每个转录本isoform中的fragment个数

基因表达谱的分析

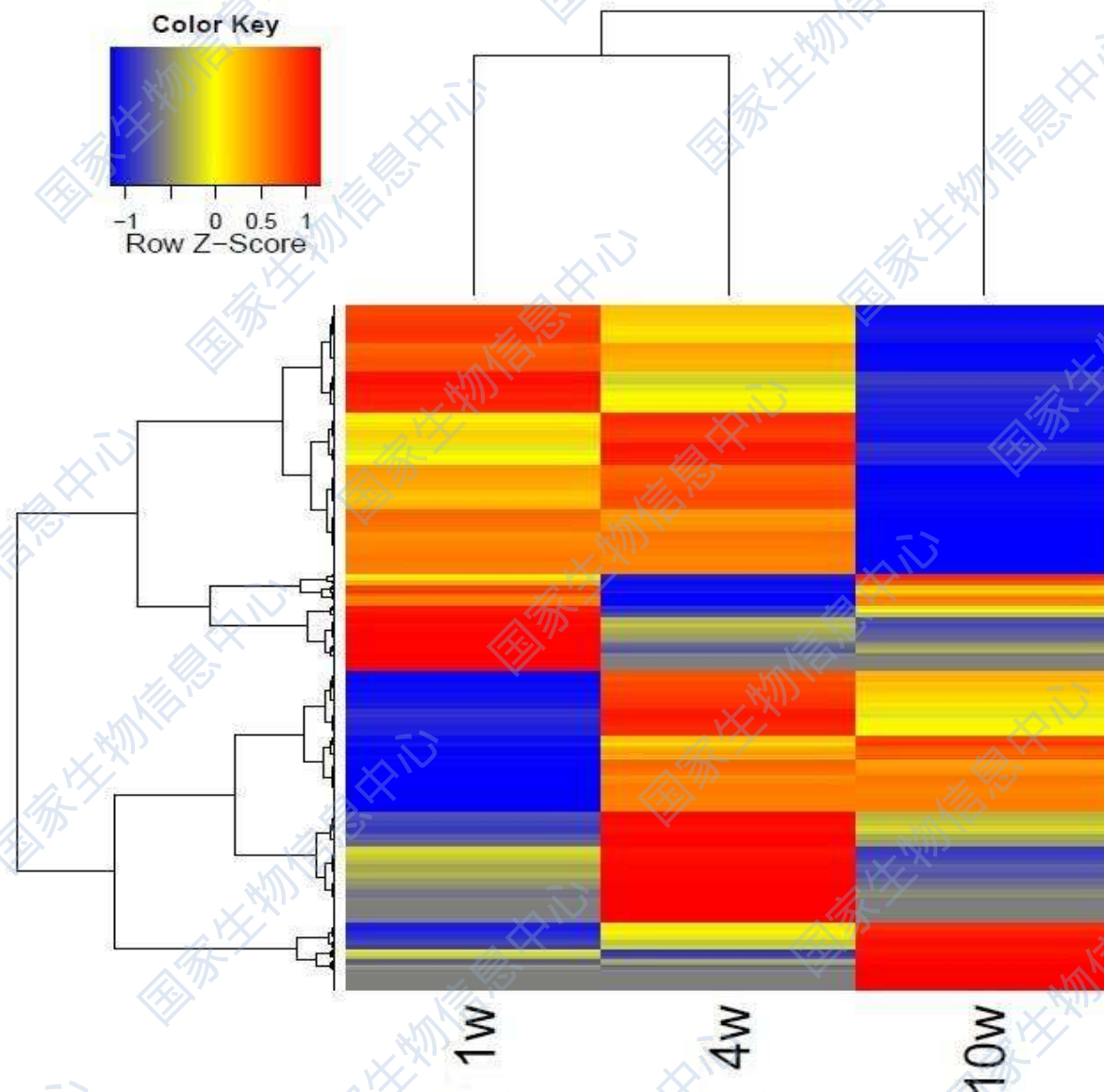
基因表达丰度：

$$\text{RPKM} = \frac{\text{专一比对到外显子的序列数} \times 10^9}{\text{外显子长度} \times \text{专一比对的所有序列数}}$$



基因表达谱的分析

三个时期表达基因的聚类分析



RNA-seq技术与芯片技术的比较

方法	表达芯片	Tilling芯片 与基因组大小相关 (人/小鼠 $\geq 35\text{bp}$)	RNA-seq
分辨率	N/A	低-高, 与芯片密度相关	1bp
成本	低	<2 数量级	高
检测范围	<4个数量级	与测序量及实际表达量相关	高
灵敏度	中	有限	可以
新转录本	不可以	有限	可以
可变剪接	不可以	可以	可以
可变调控区	N/A	N/A	可以
反义转录本	有限	有限	可以
SNPs, 突变	有限	有限	可以
等位表达差异	0.01-0.05Gb	0.1-1Gb	1-15Tb
数据量	低	高	极高
生物信息分析			

-----引自Wilhelm, B.等, 2009

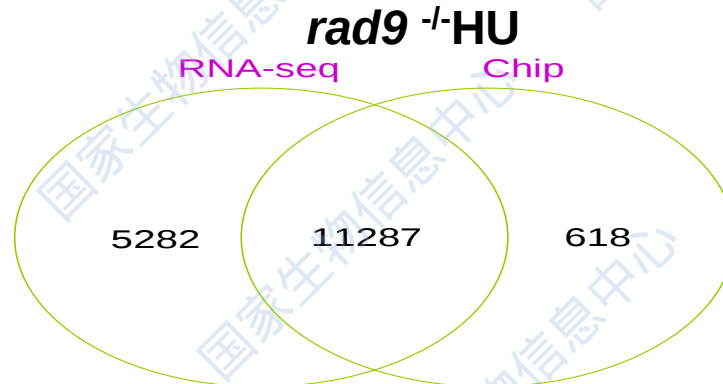
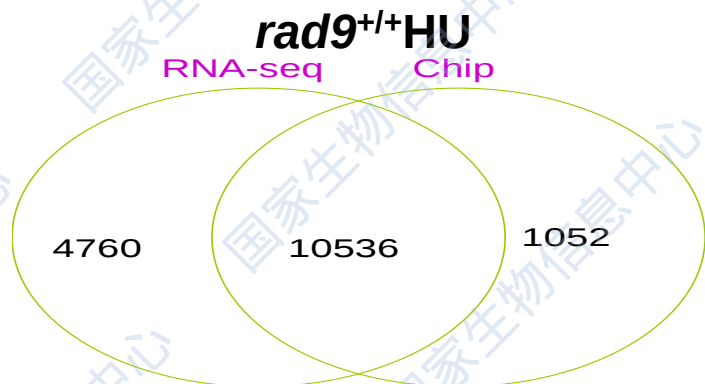
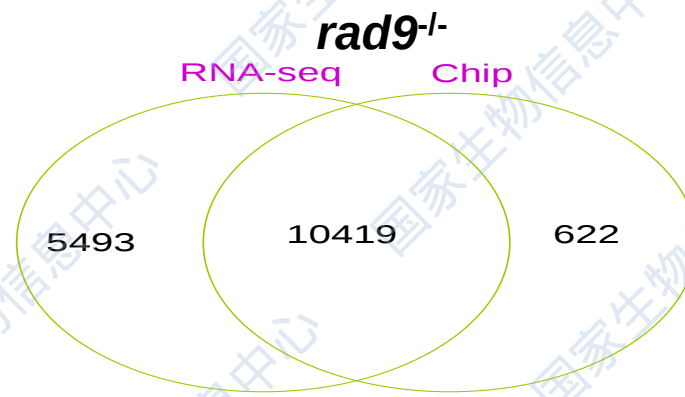
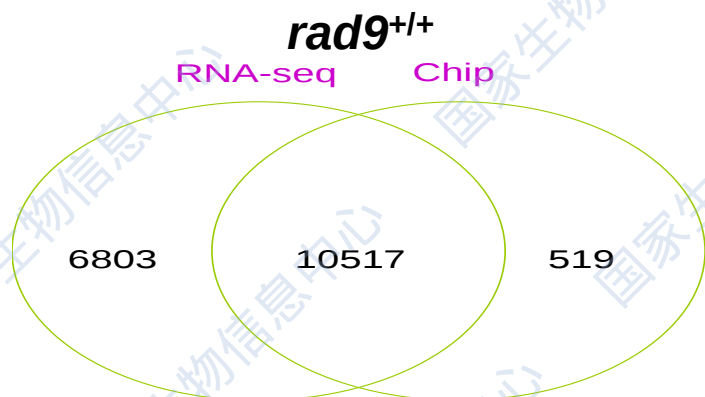
基因表达谱的分析

表达基因的种类

文库	<i>rad9</i> ^{+/+}	<i>rad9</i> ^{+/+} HU	<i>rad9</i> ^{-/-}	<i>rad9</i> ^{-/-} HU
RNA-seq (>5tags)	17320	15296	15912	16569
DNA-chip	11036	11588	11041	11905

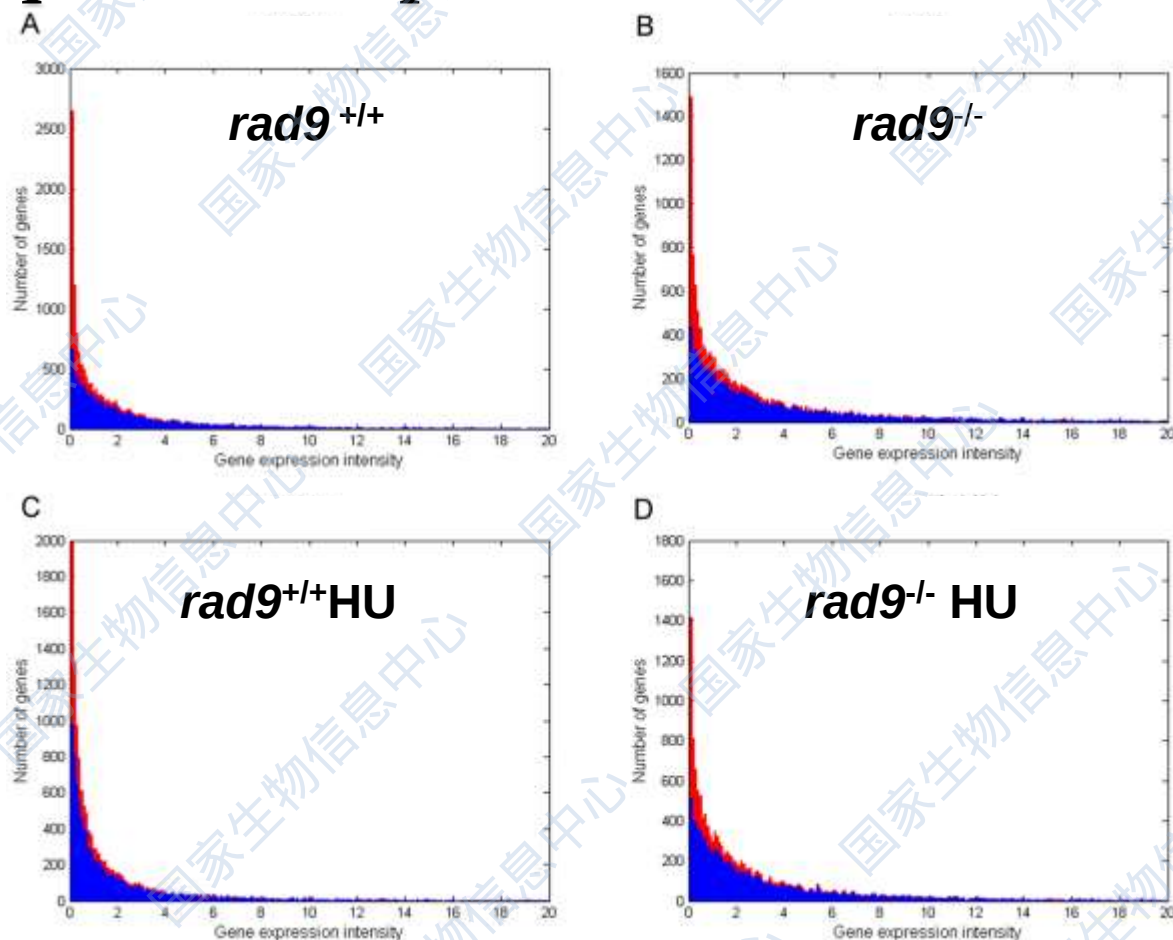
基因表达谱的分析(续)

RNA-seq与DNA-chip对基因检测能力分析



基因表达谱的分析(续)

RNA-seq与DNA-chip对检测到的基因表达量的分布



基因功能分类

◆ 手工分类

大部分以Adams 95年的文章中的采用分类体系为标准。

【Adams. MD, et al. Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. Nature. 1995 377(6547 Suppl):3-174 】

◆ 计算机批量处理

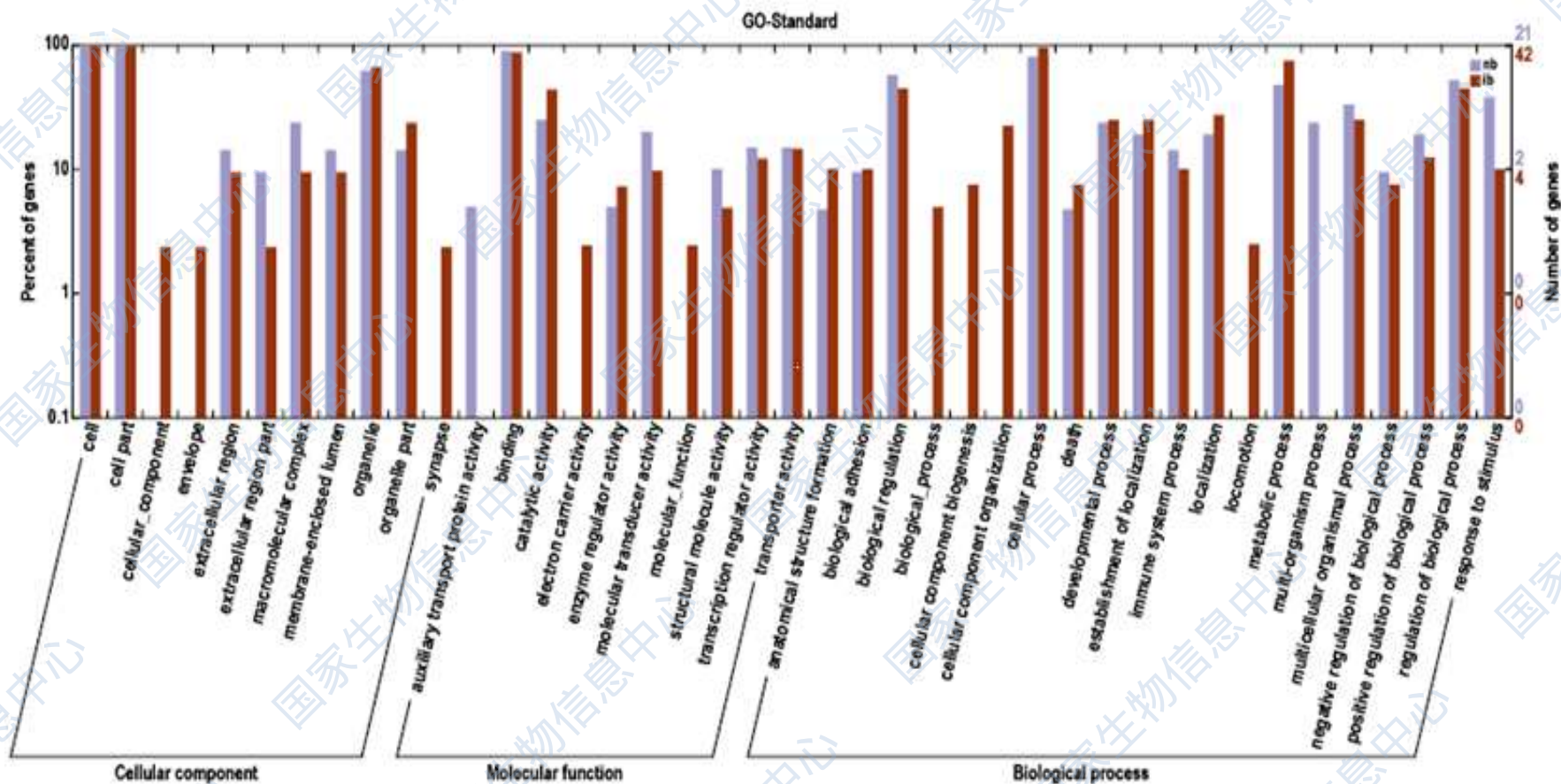
利用标准基因词汇体系Gene Ontology，进行近似的分类(分子功能、生物学过程、分子组分)。(<http://www.geneontology.org/>)

◆ 基因产物直系同源簇的分析（COG）

(<http://www.ncbi.nlm.nih.gov/COG/>)

Data Analysis Result for fragment data

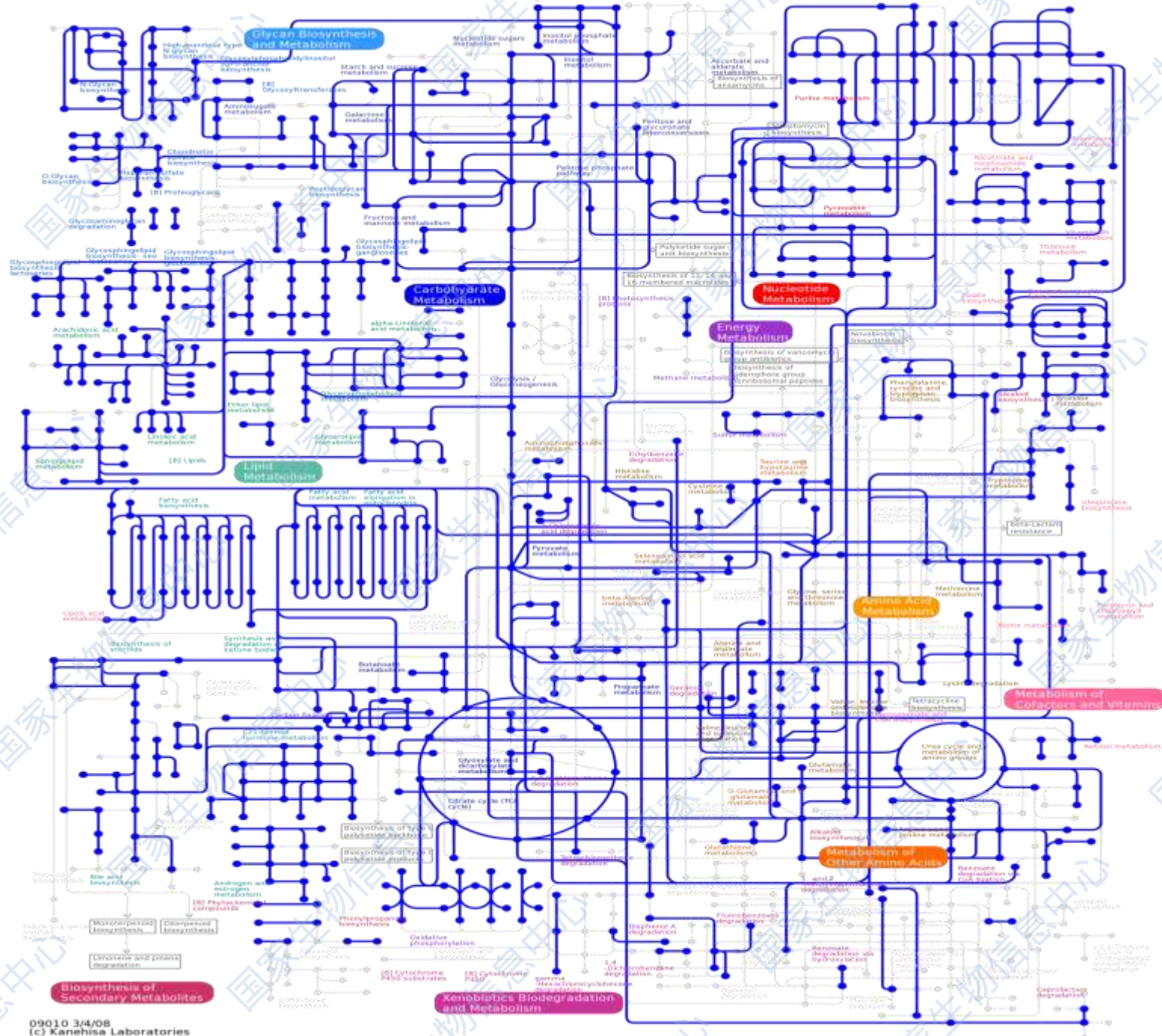
GO functional classification



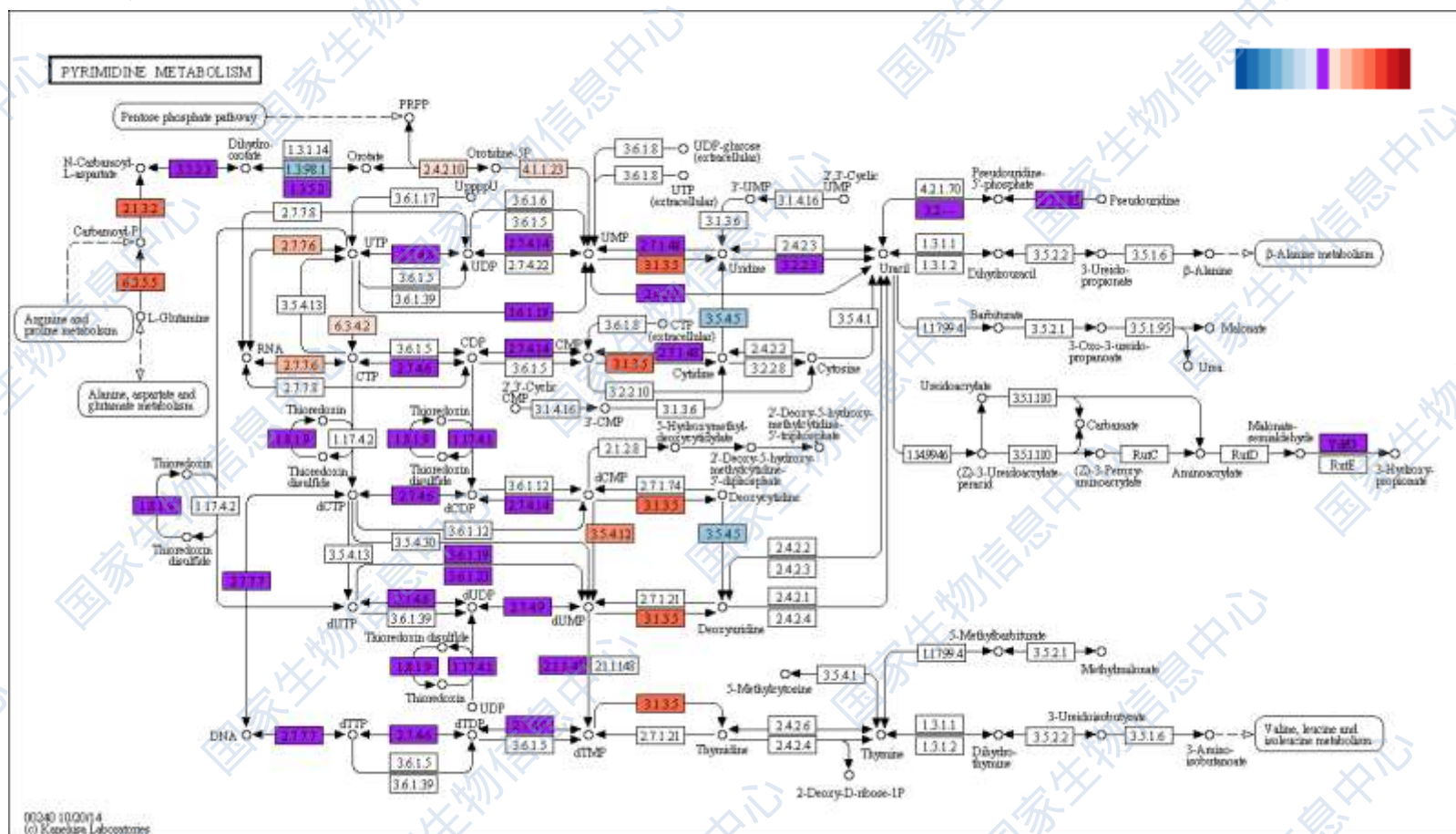
EST的代谢途径分析(KEGG)

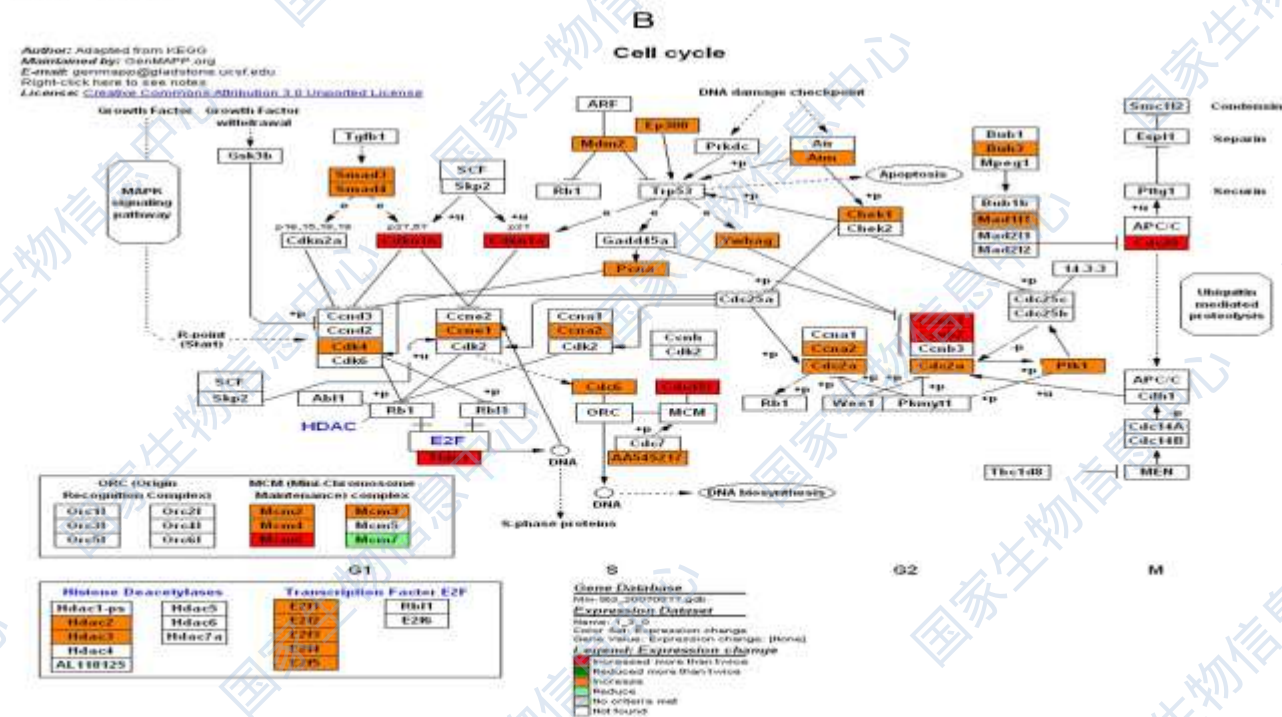


<http://www.genome.ad.jp/kegg/>

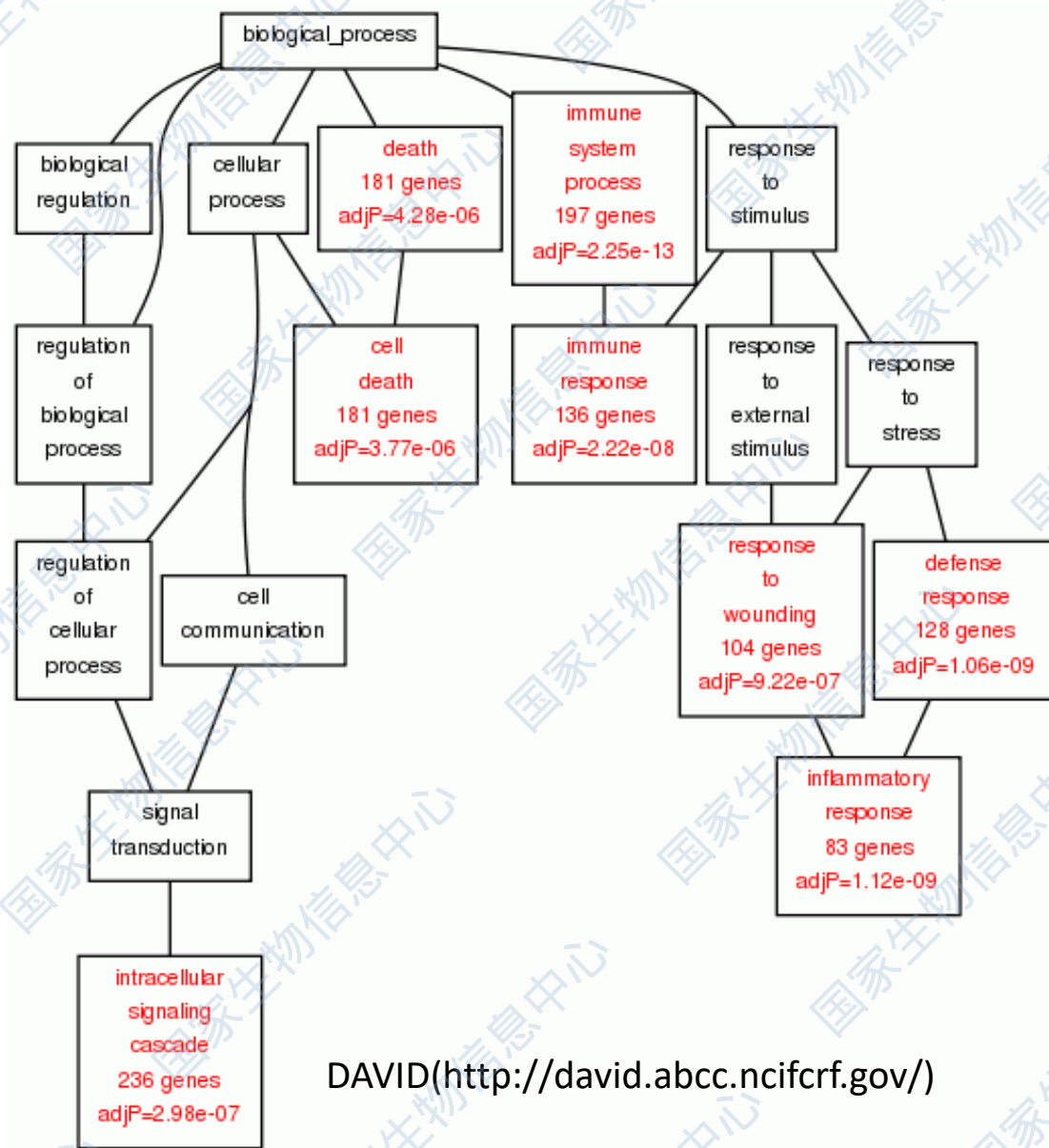


差异表达基因代谢途径





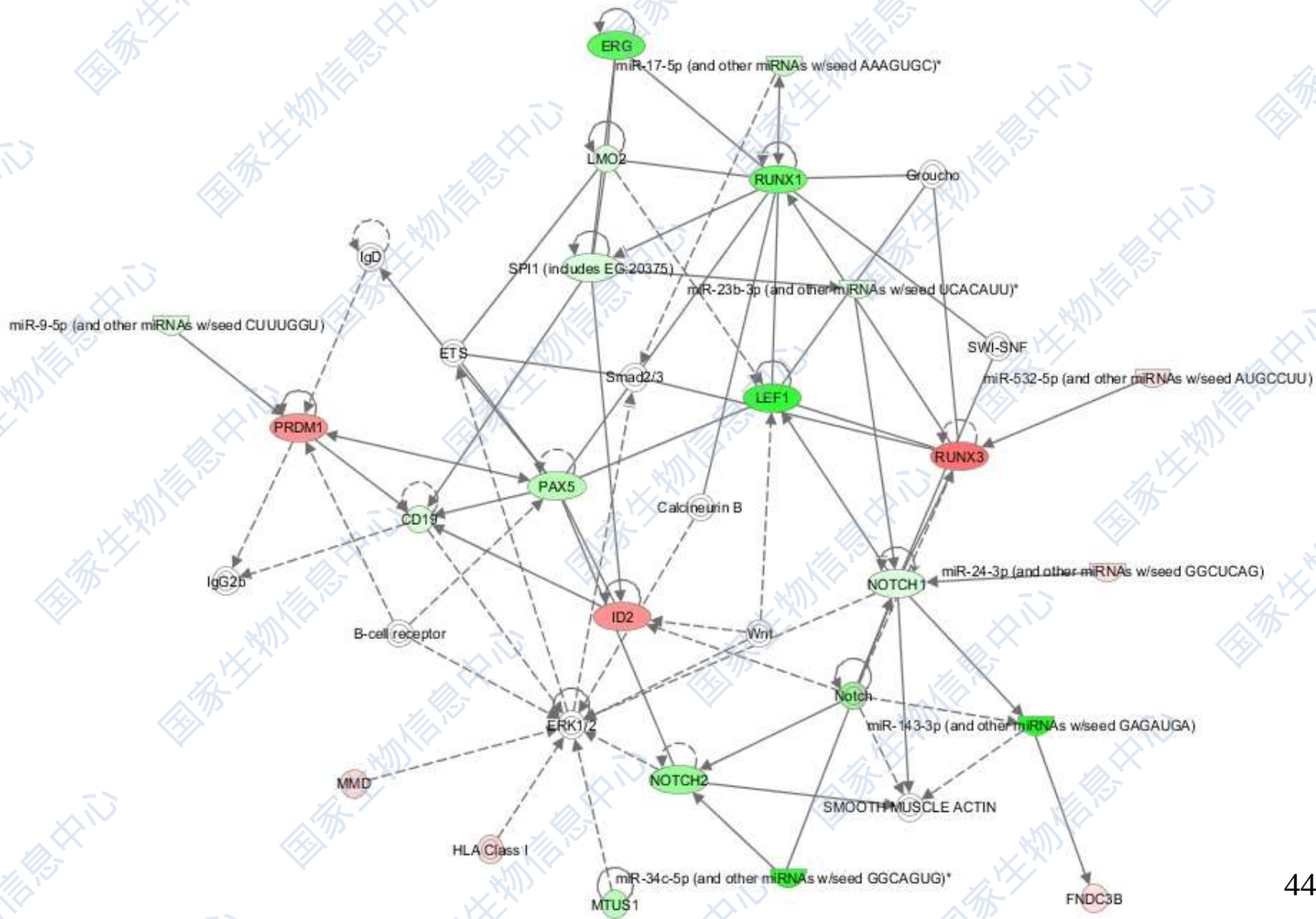
差异表达基因的功能富集



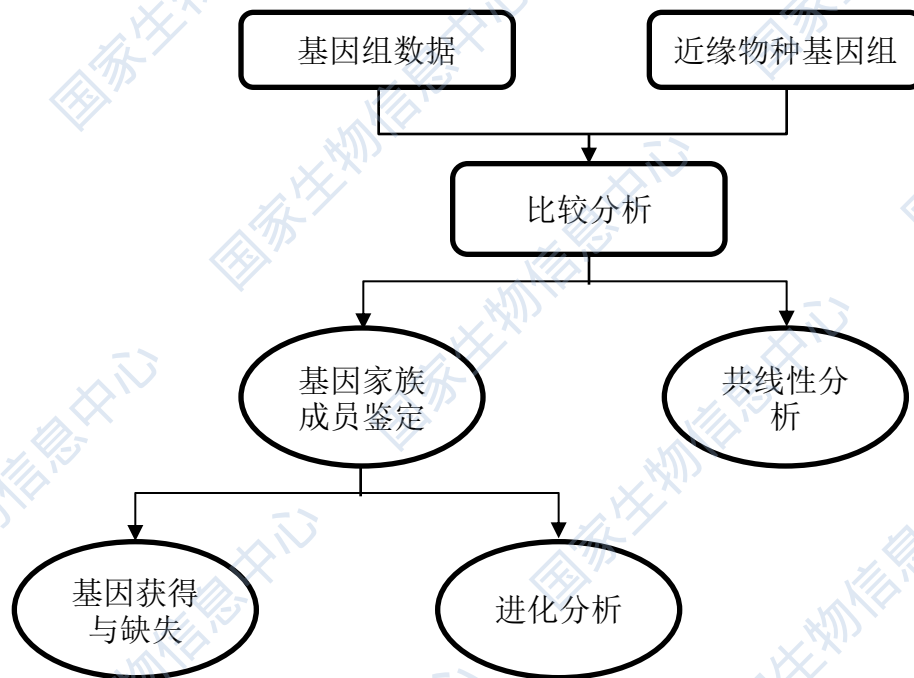
DAVID(<http://david.abcc.ncifcrf.gov/>)

调控网络的构建 - IPA

细胞分化过程相关基因调控网络

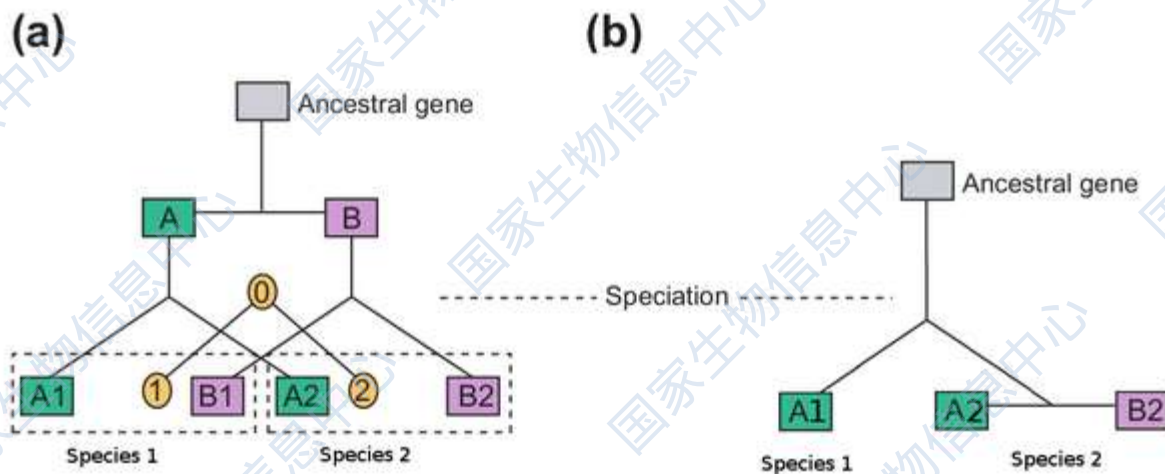


4.基因组比较分析

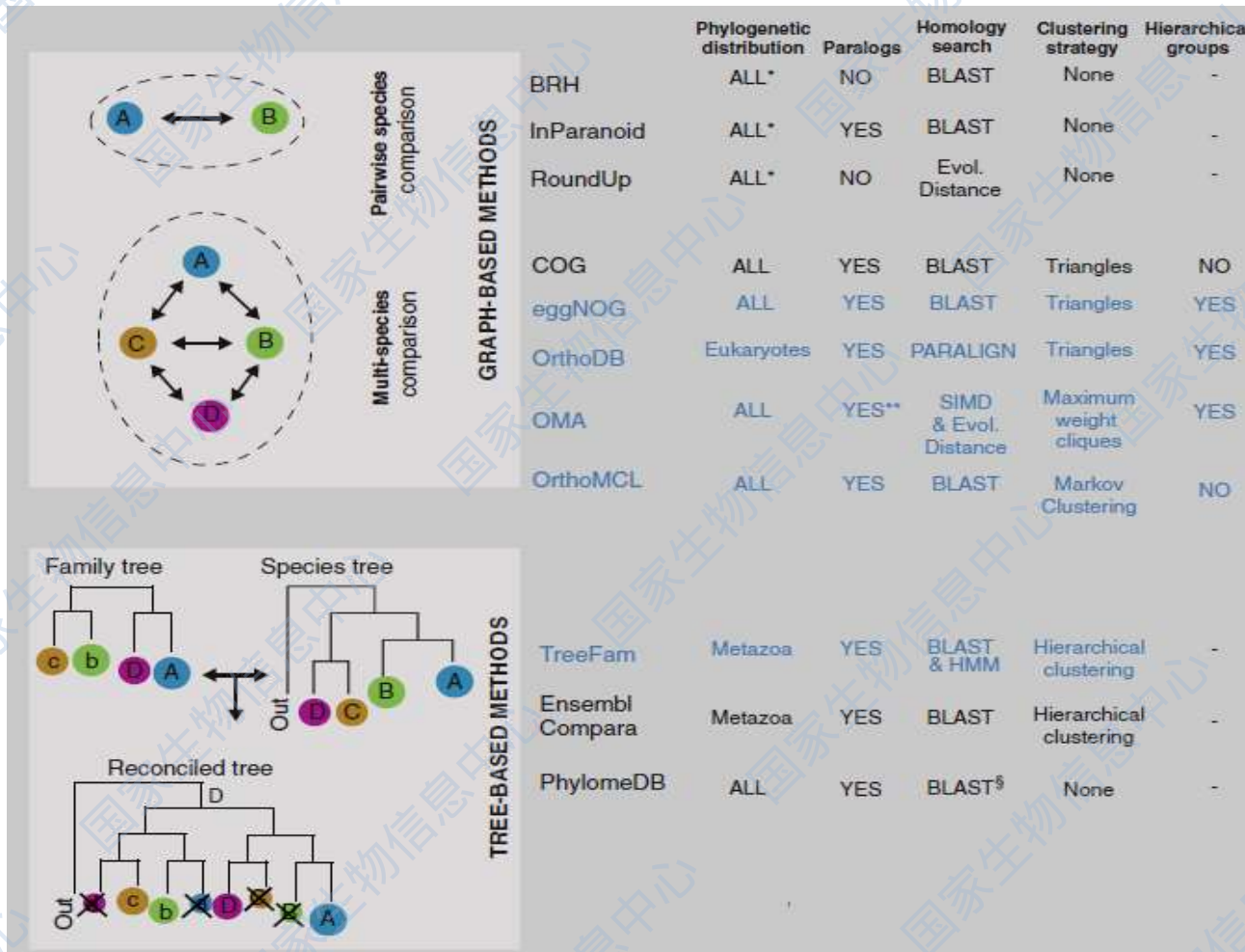


基因家族

直系同源（Orthologs）和旁系同源（Paralogs）

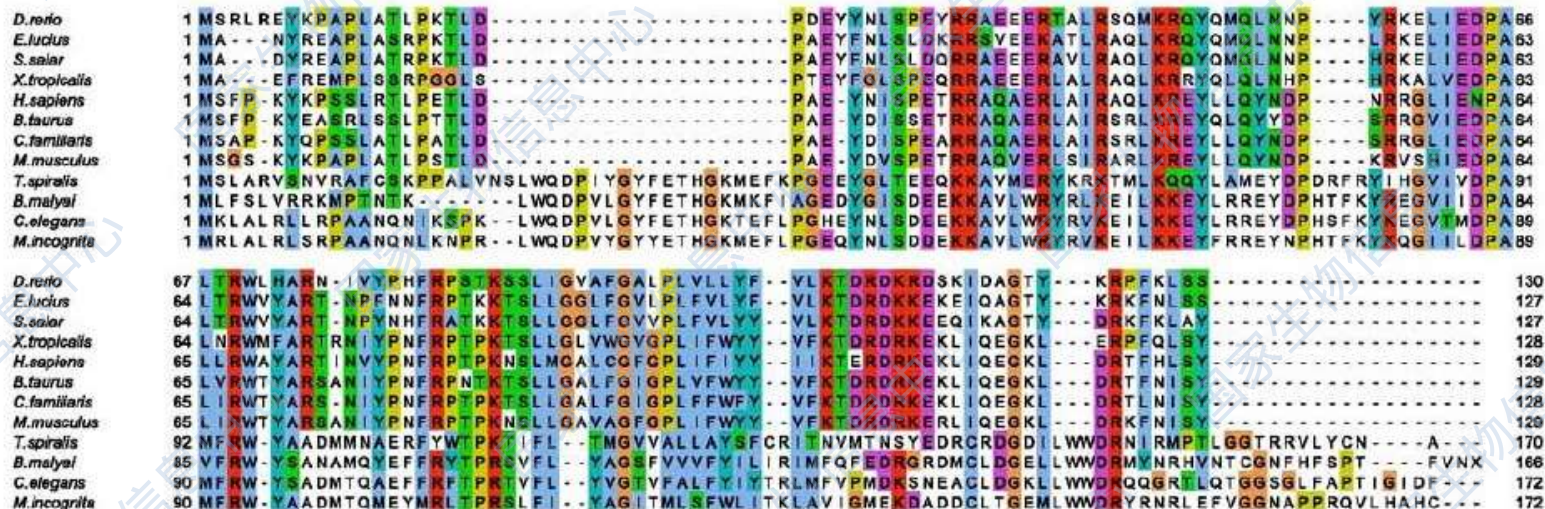


基因家族



Orthologues evolution and conservation

a



b

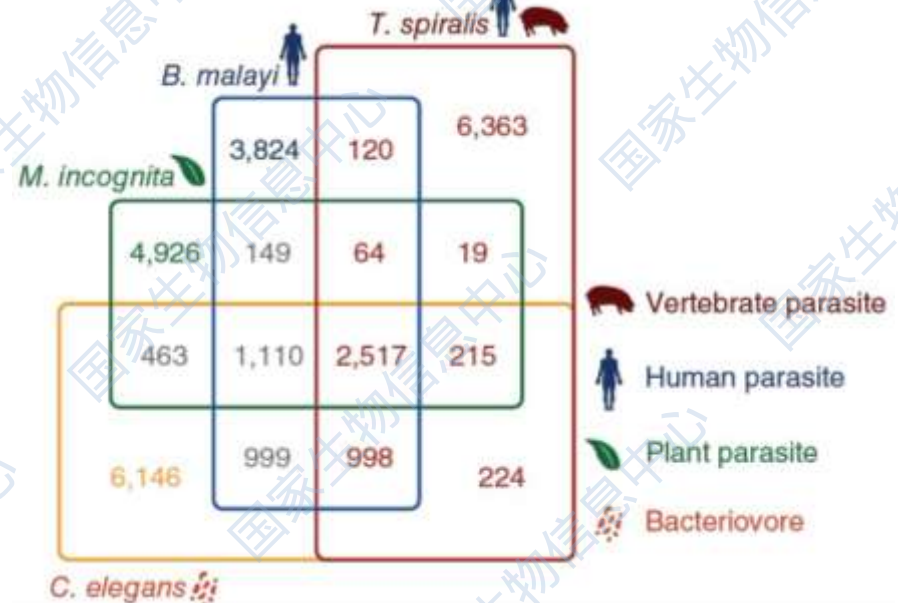
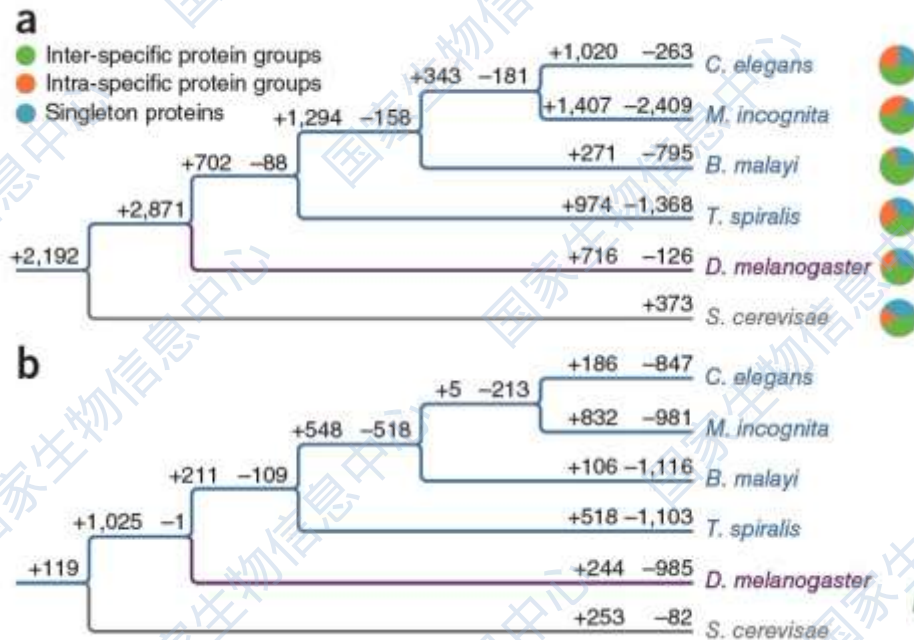
Nematode proteins



Non-nematode proteins

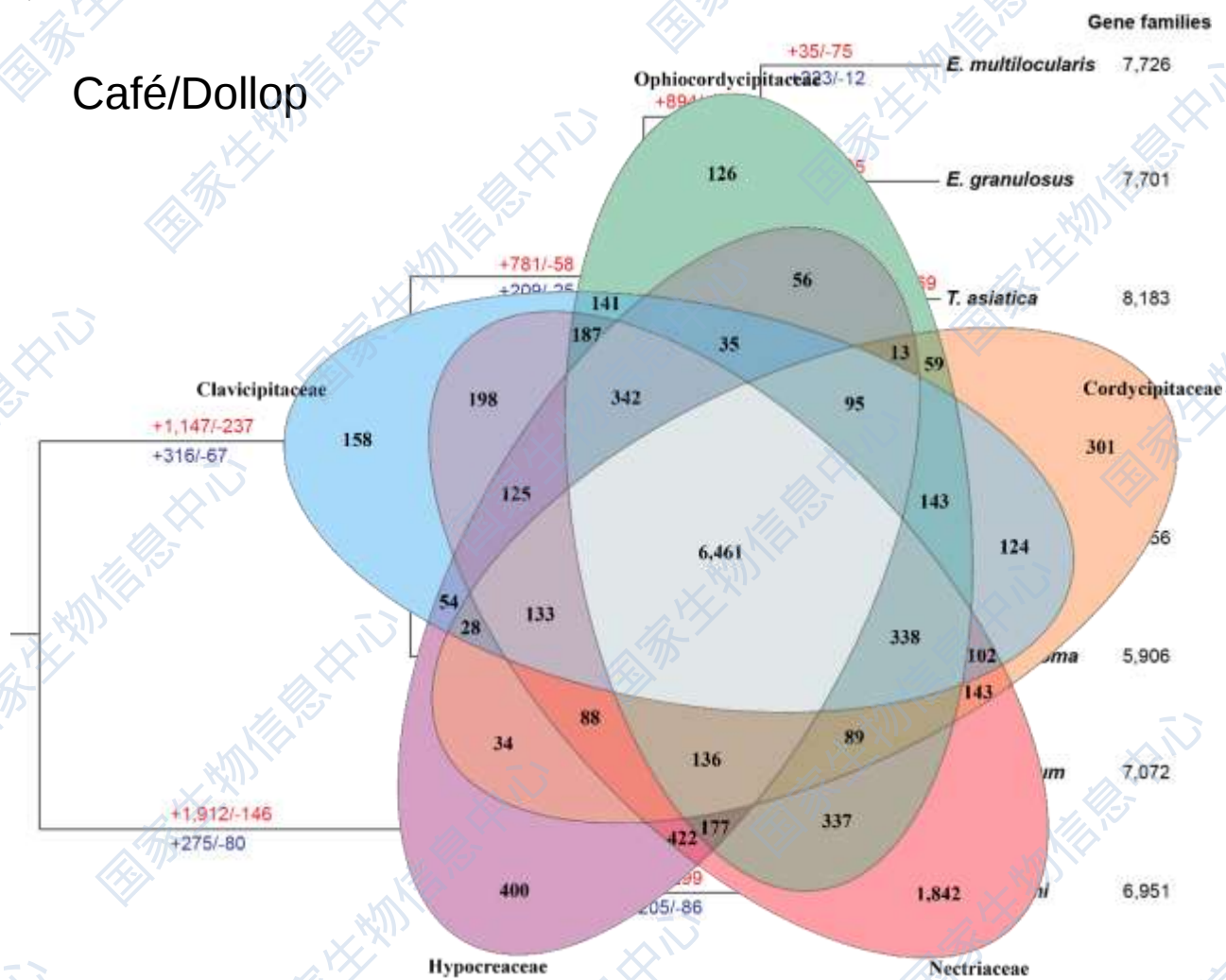


Gene tree analysis of gene expansions

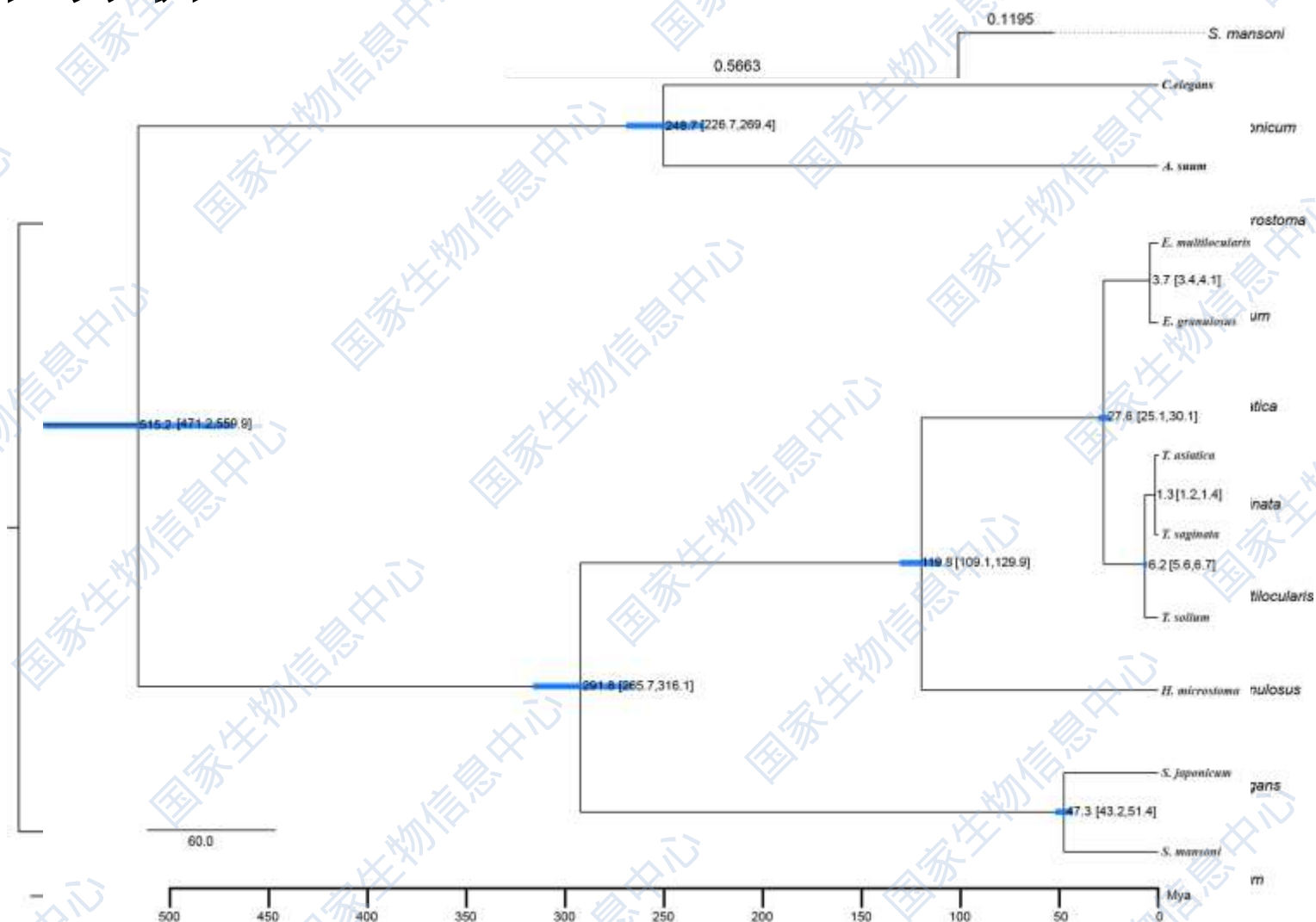


基因家族

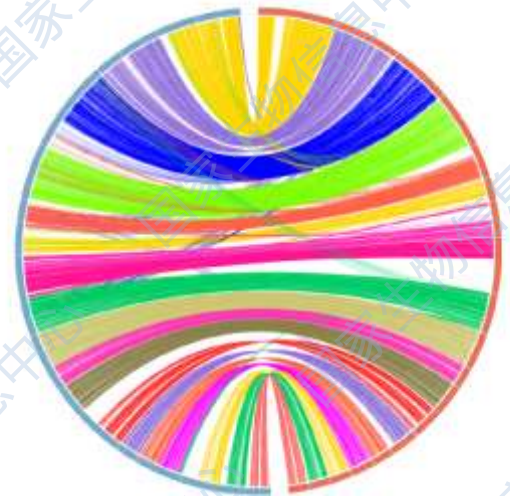
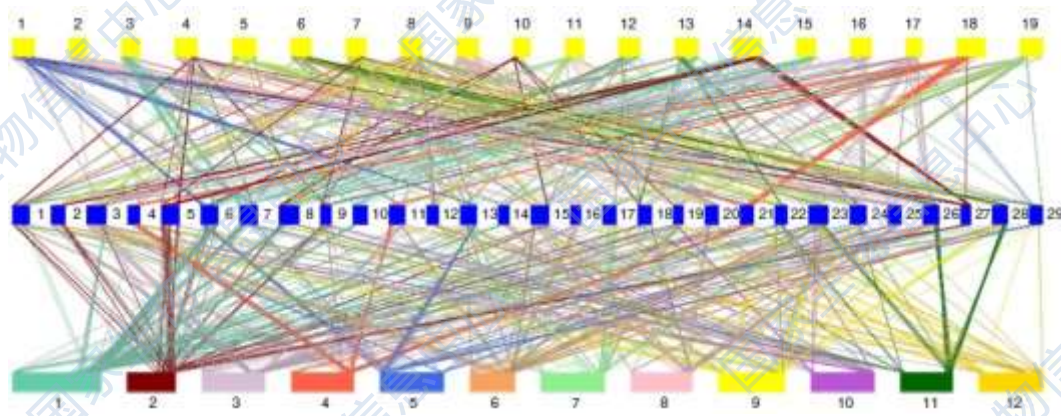
Café/Dollop



进化分析



共线性分析



[Nat Commun.](#) 2013;4:2640



橡胶树属(Hevea)

1. 大戟科 (Euphorbiaceae) 300属8000种, 我国61属360种。
2. 特征: 高达乔木, 有乳状汁。三出复叶; 叶柄顶端有腺体。花小, 单性同株, 成圆锥状聚伞花序; 萼5齿裂, 无花瓣。蒴果

3. 其他属 :

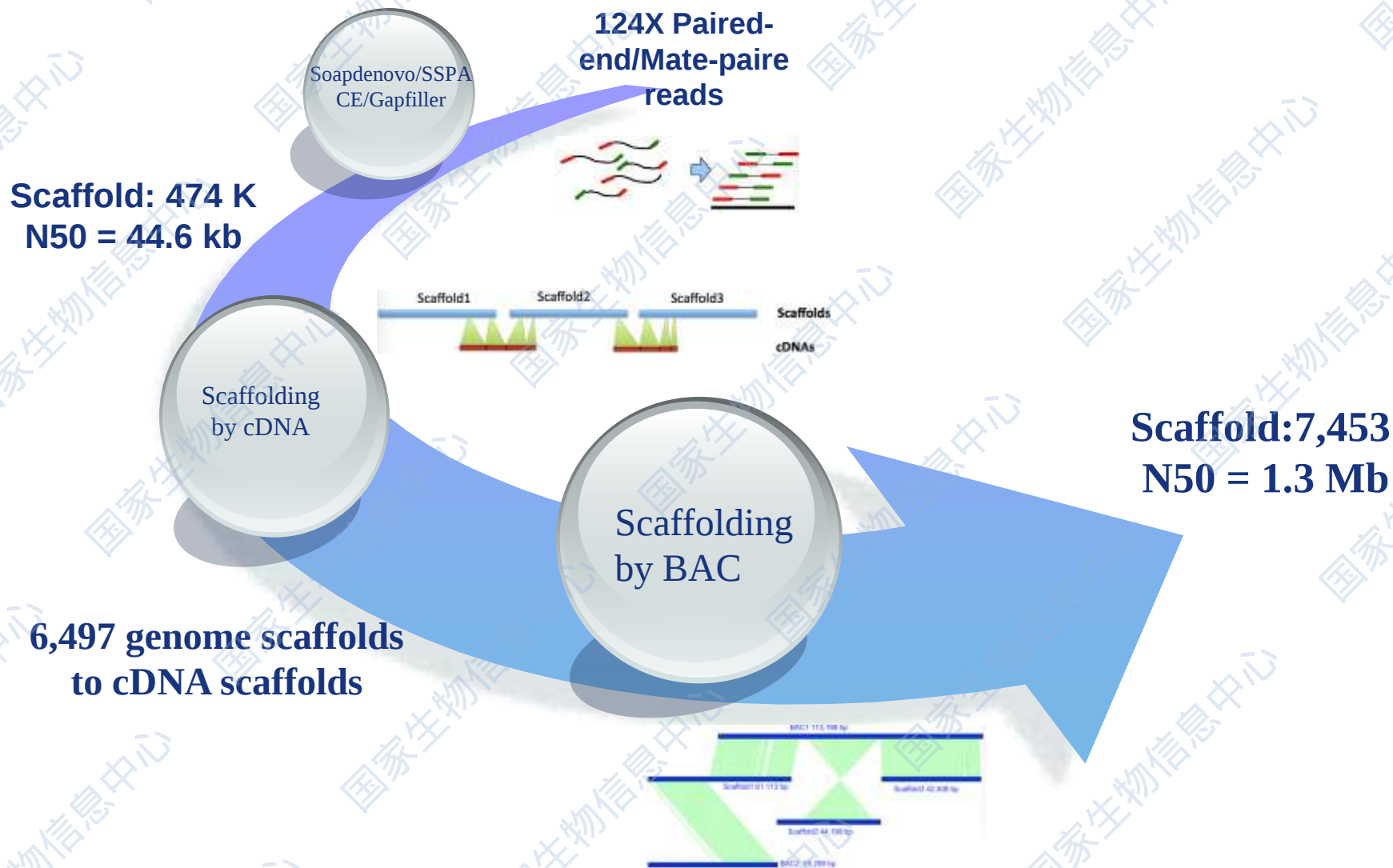
- 1). 蓖麻属(Ricinus)
- 2). 油桐属(Aleurites)
- 4). 大戟属(Euphorbia)
- 5). 铁苋菜属(Acalypha)



4. 热研7-33-97系中国热带农业科学院橡胶所培育的高产抗风品种, 海南植胶区大规模推广, 为RRIM600×PR107的杂交后代。
5. RRIM600 (马来西亚品种, 海南植胶区大规模推广, 产量较高, 但抗逆性较差); PR107(印尼品种, 我国植胶区大规模推广, 晚熟品种, 耐乙烯利刺激, 后期产量高, 抗风性较强)。

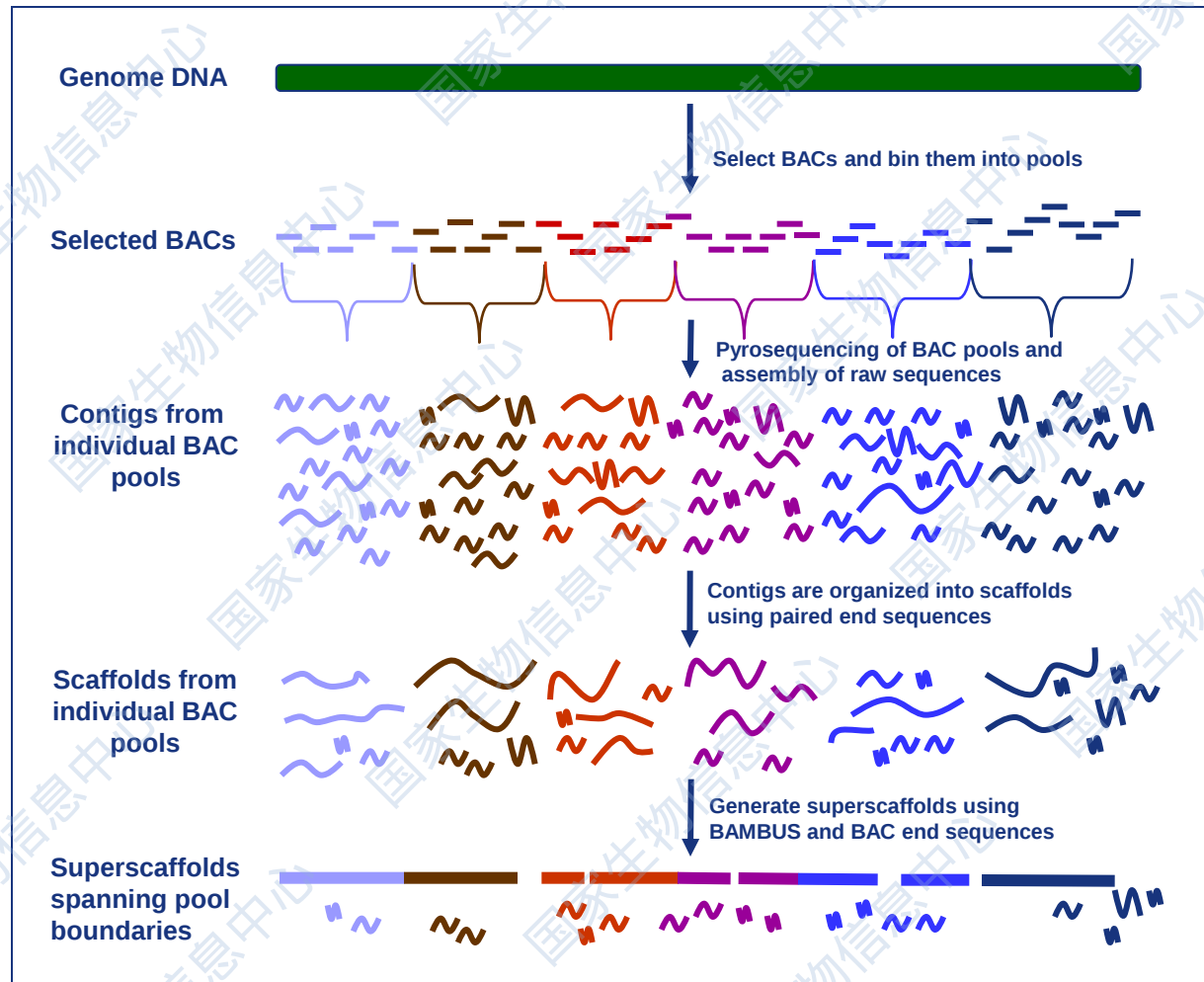
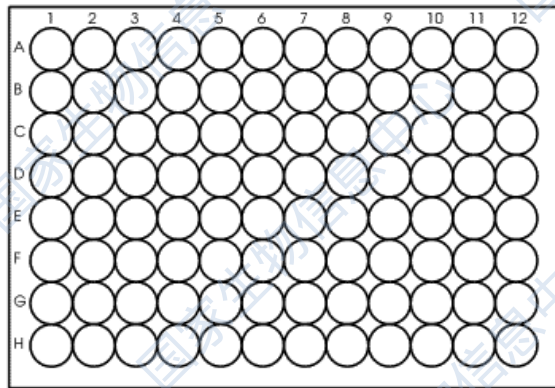
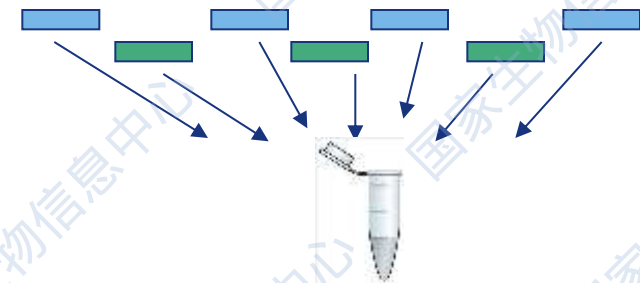


基因组组装过程





基因组组装BAC pooling 示意图



From Rounsley et al. (2009)



基因组组装验证

1

BACs

(18/18)

2

Proteins

(944/944)

3

ESTs

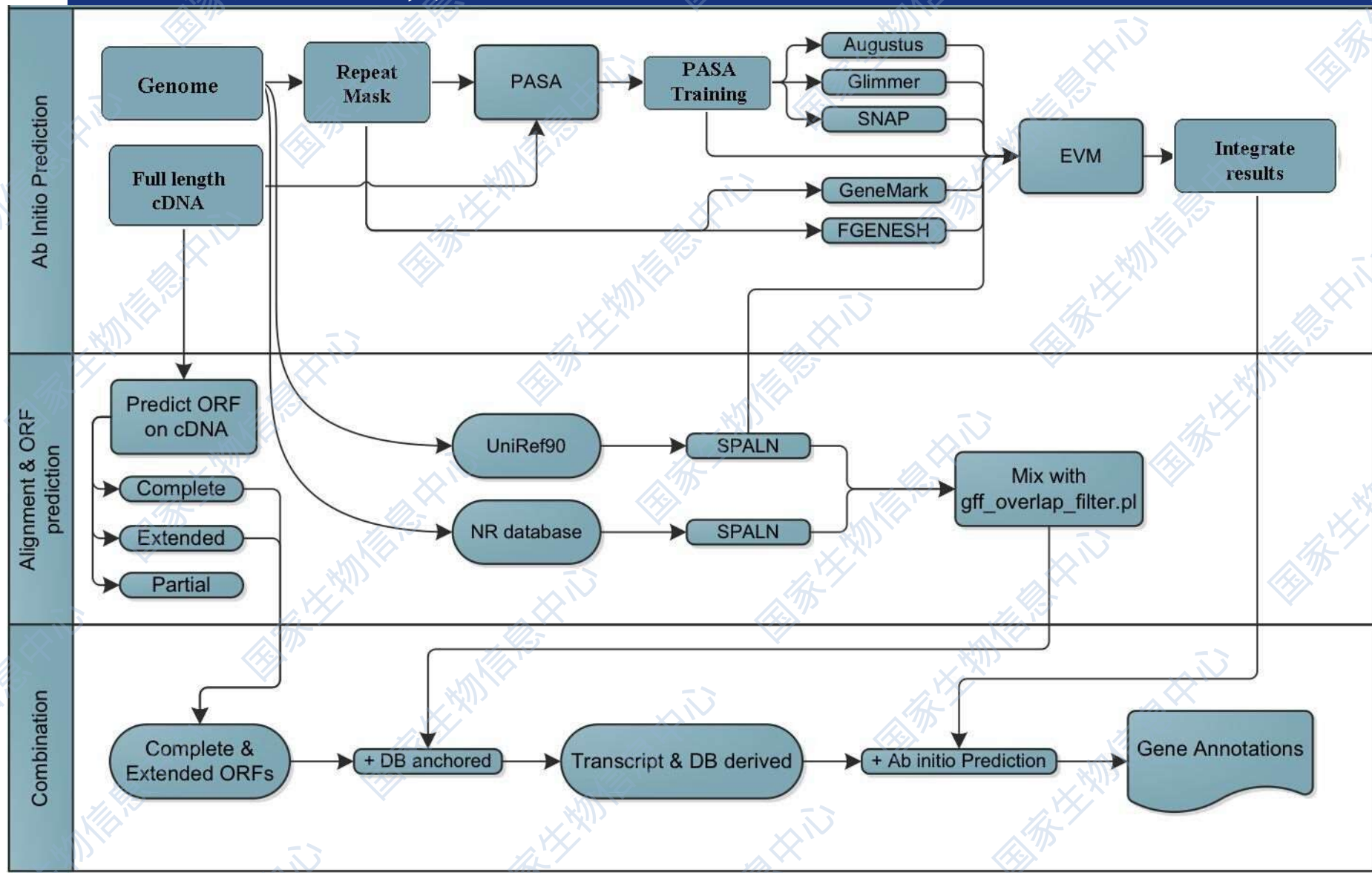
(50,534/51,631)

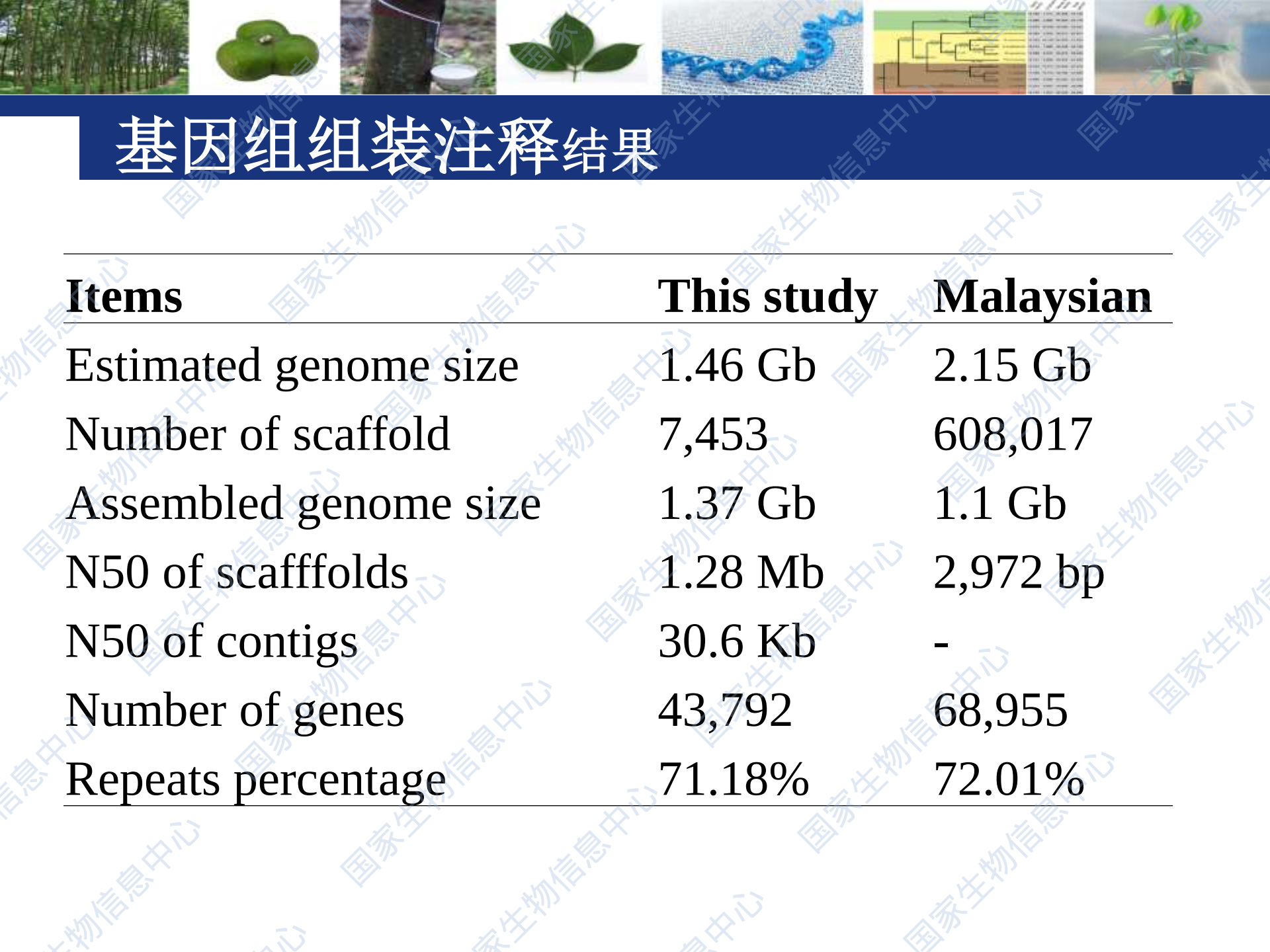
4

**Rubber
sequences**

(22,331/24,401)

基因组注释流程



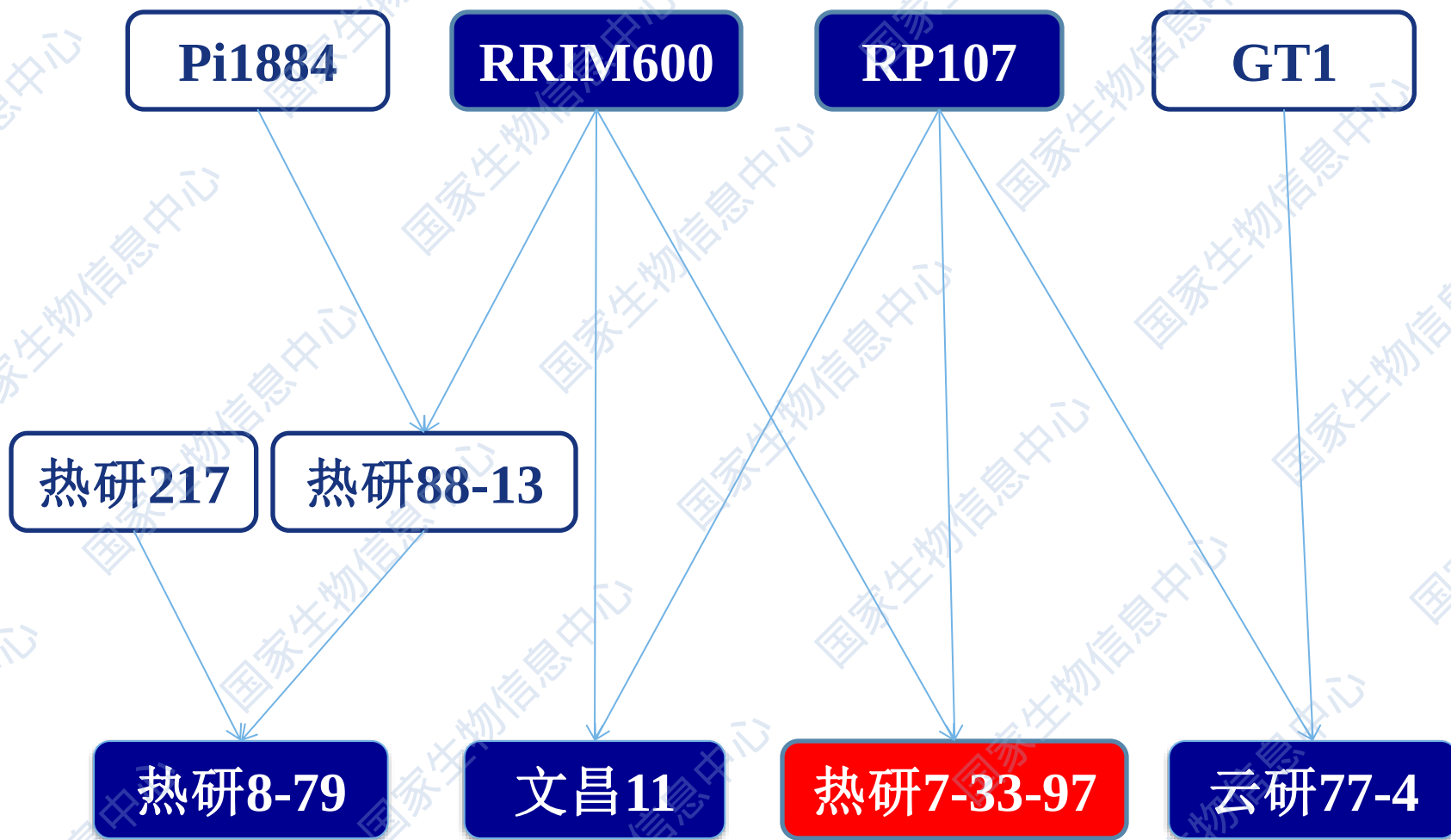


基因组组装注释结果

Items	This study	Malaysian
Estimated genome size	1.46 Gb	2.15 Gb
Number of scaffold	7,453	608,017
Assembled genome size	1.37 Gb	1.1 Gb
N50 of scaffolds	1.28 Mb	2,972 bp
N50 of contigs	30.6 Kb	-
Number of genes	43,792	68,955
Repeats percentage	71.18%	72.01%



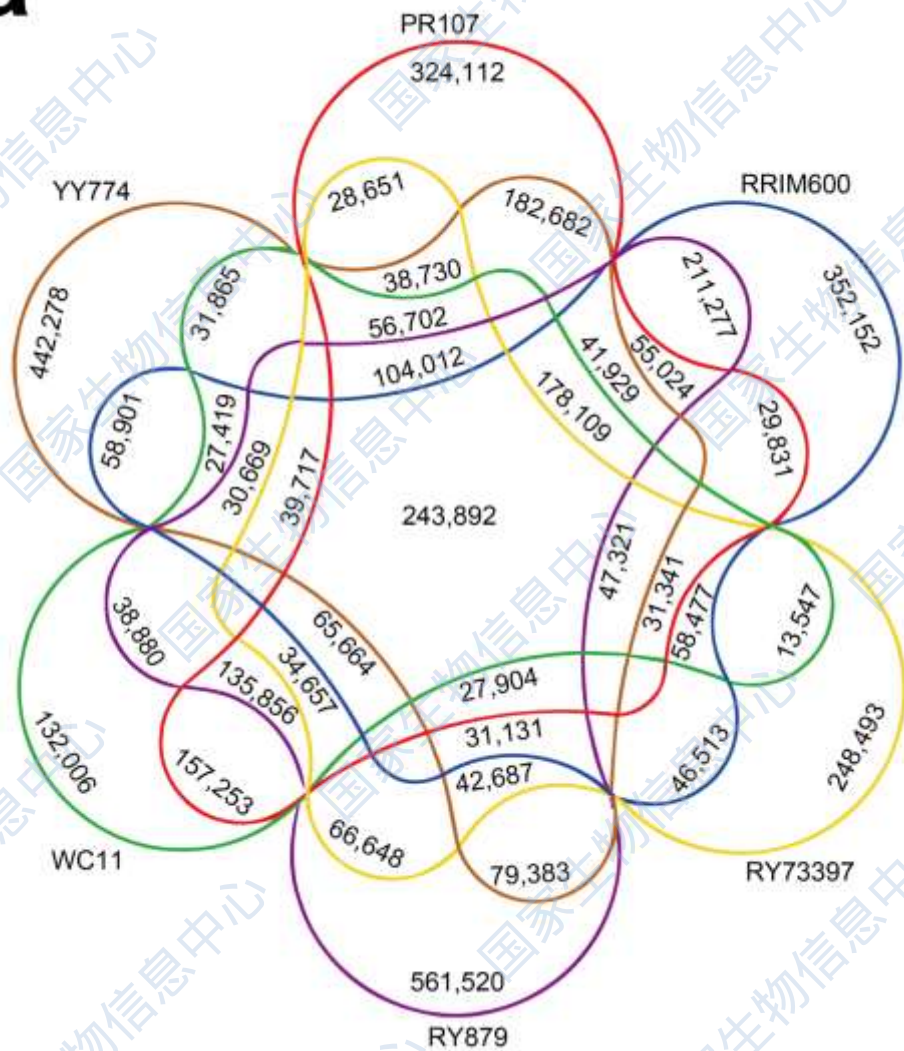
基因组重测序品系



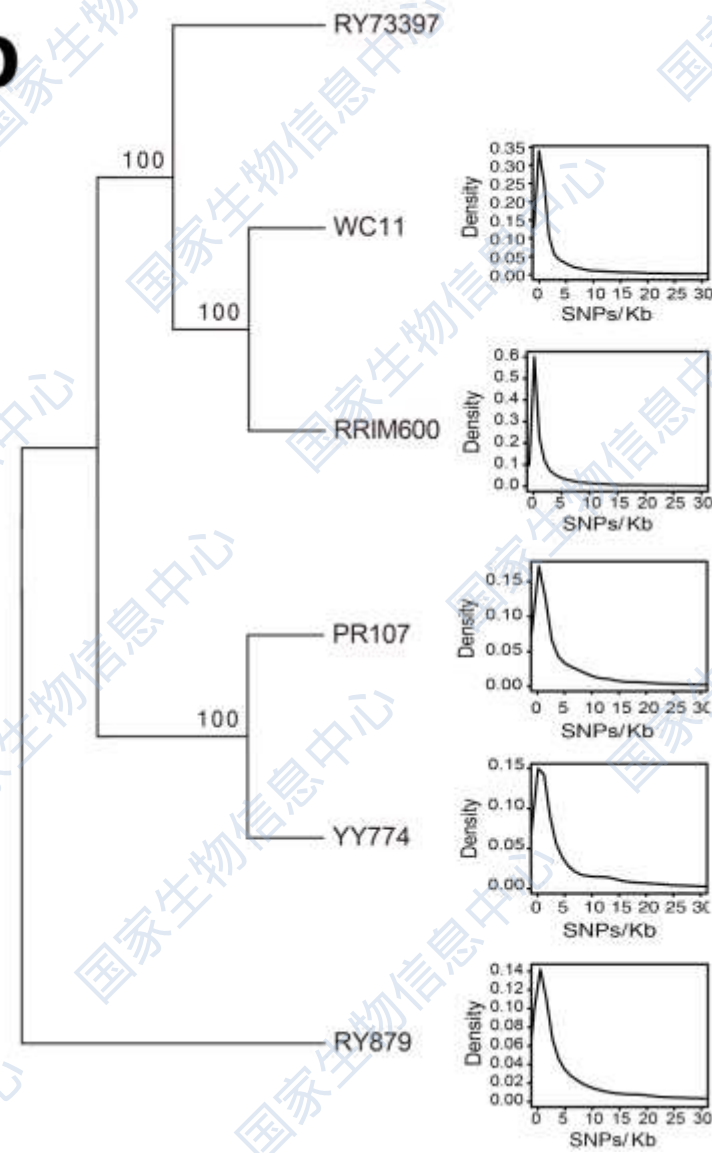


基因组重测序遗传多样性分析

a



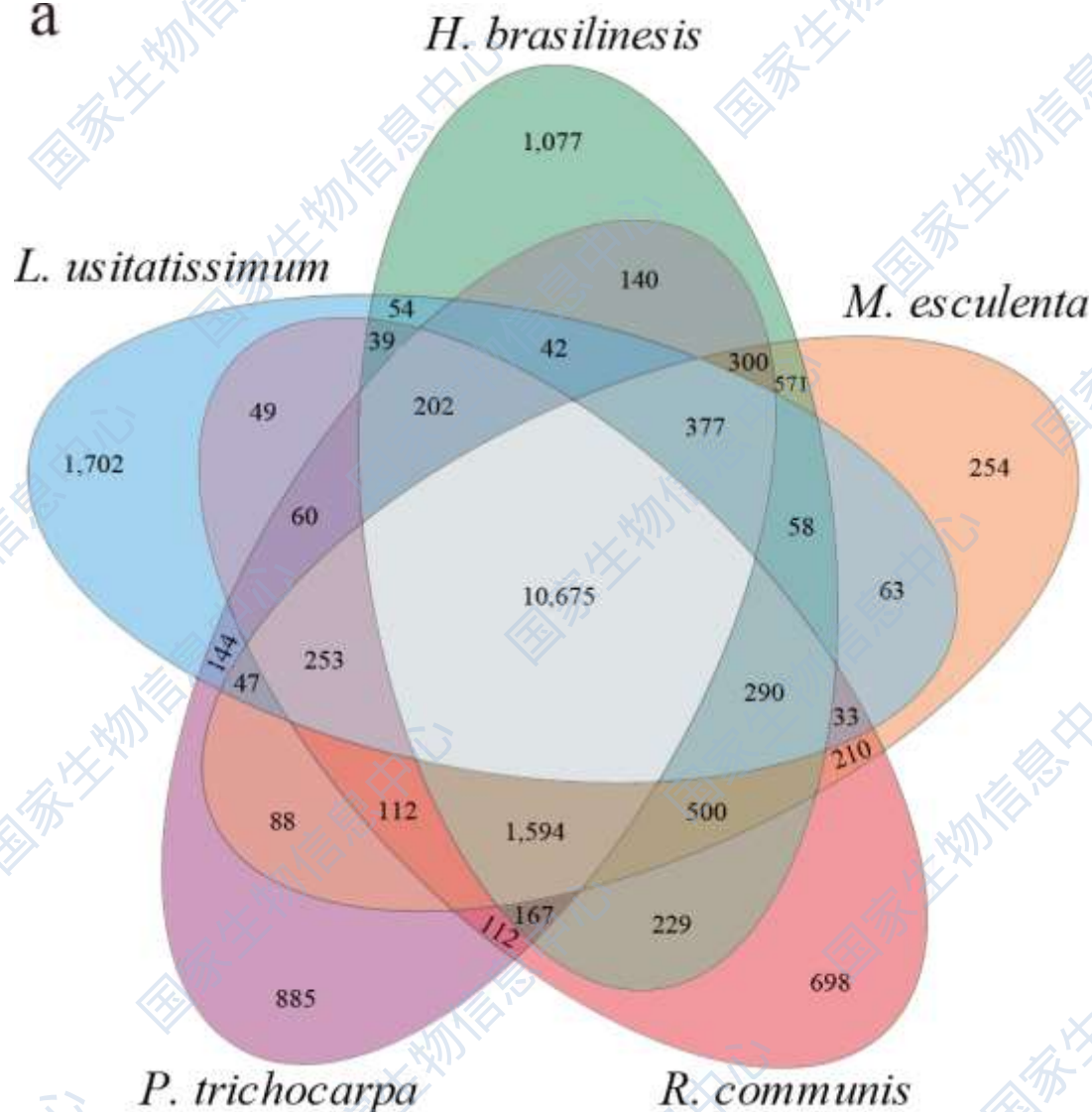
b





比较基因组分析基因分布

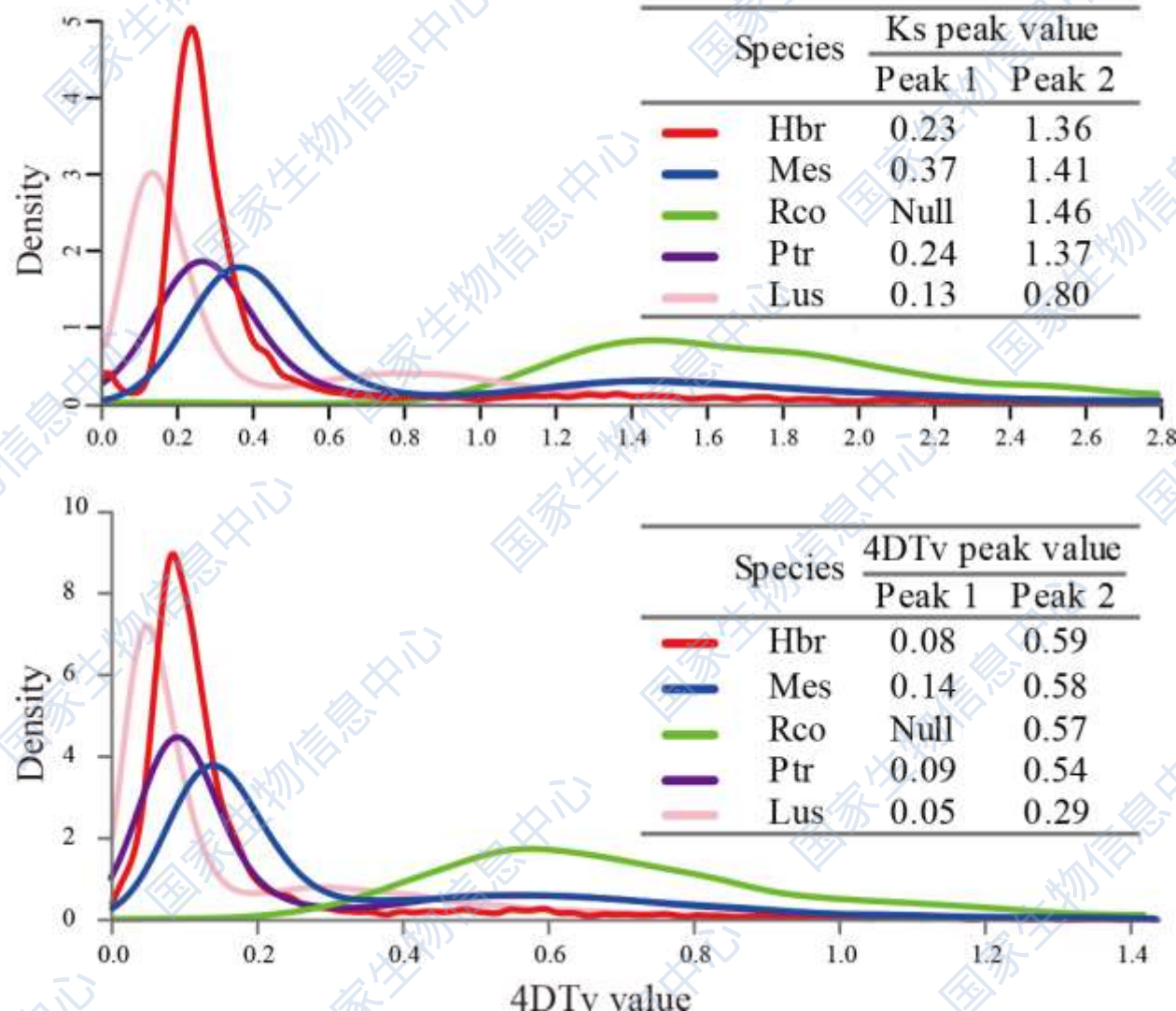
a

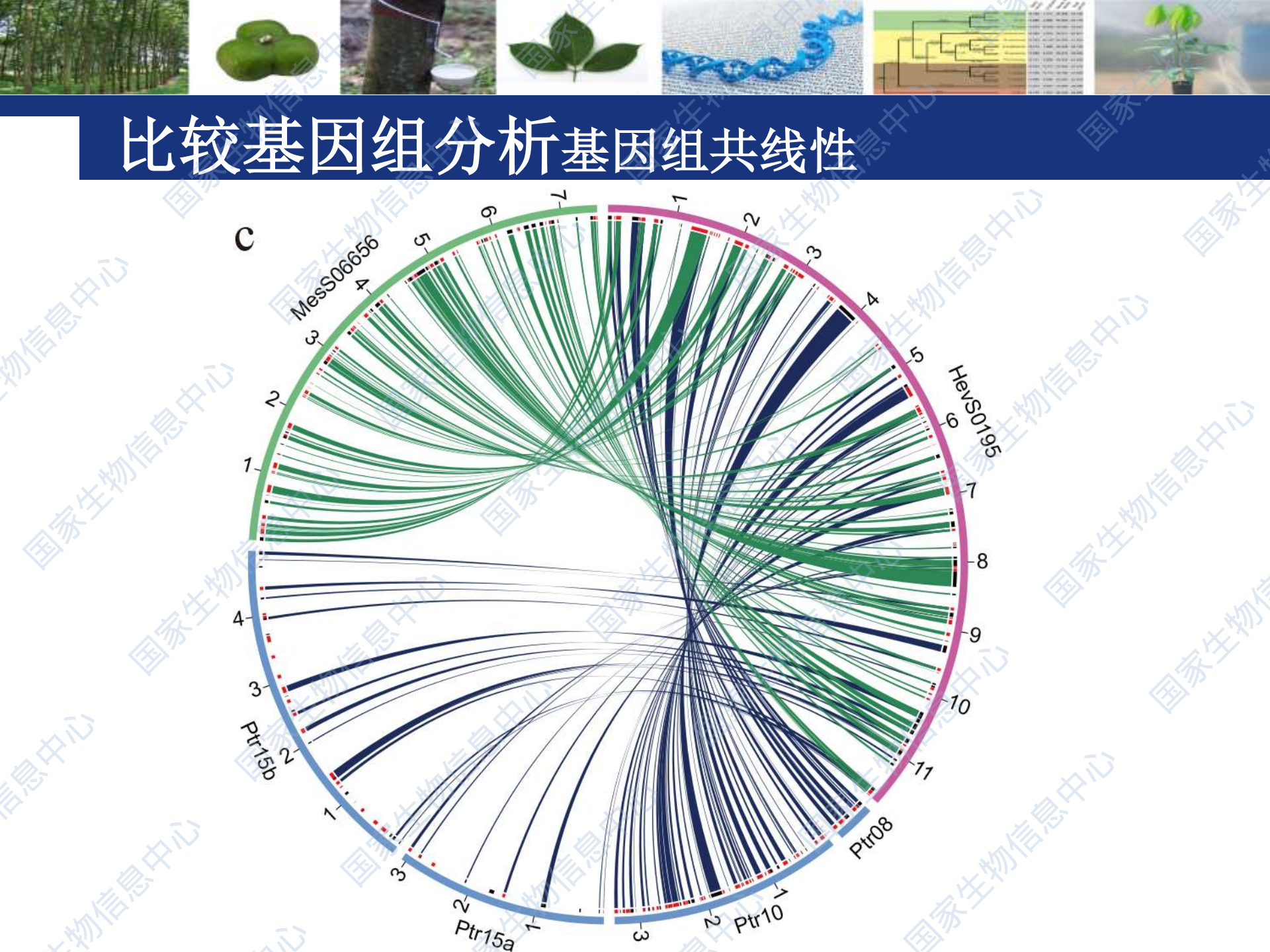




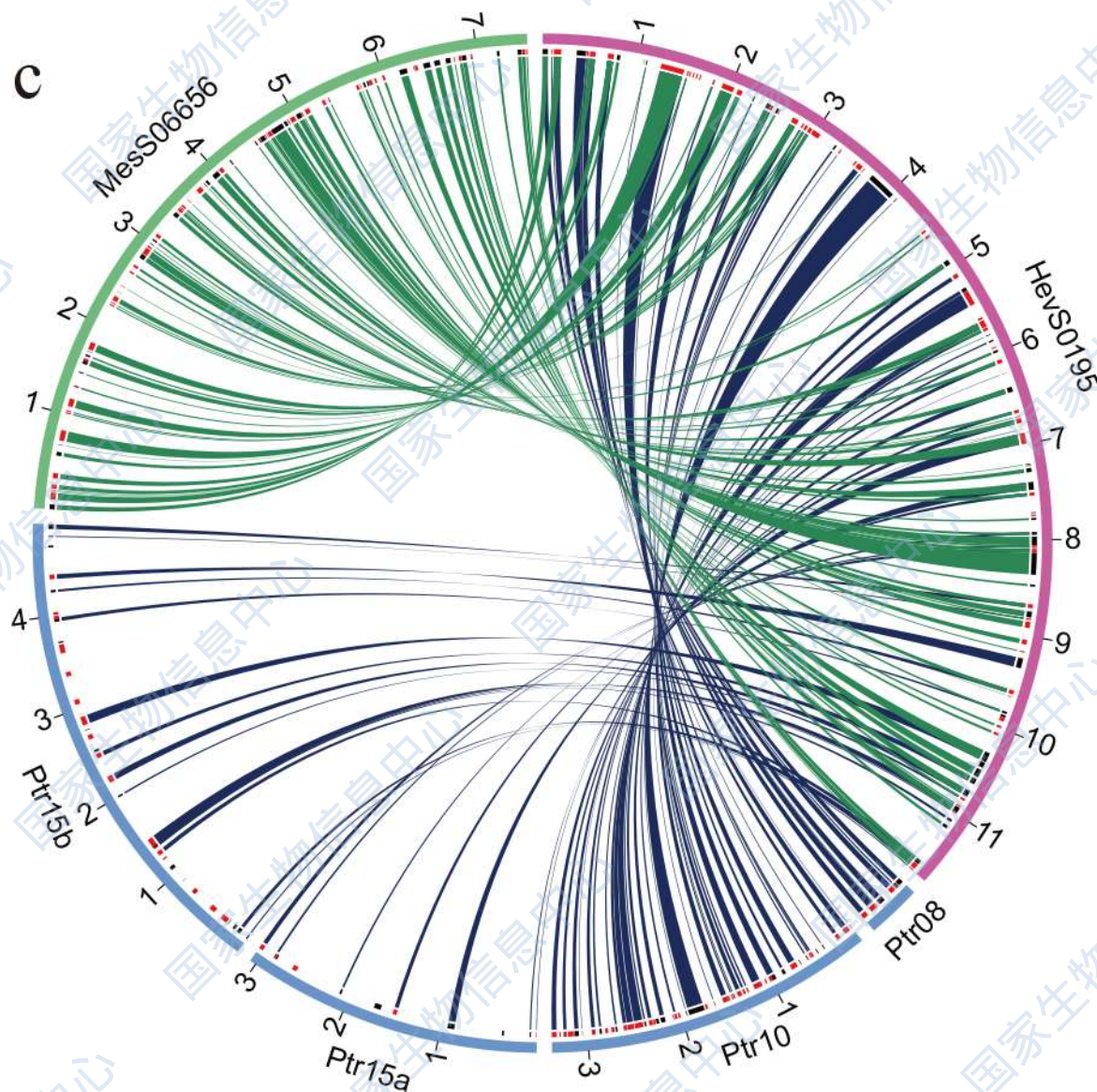
比较基因组分析基因组进化

b





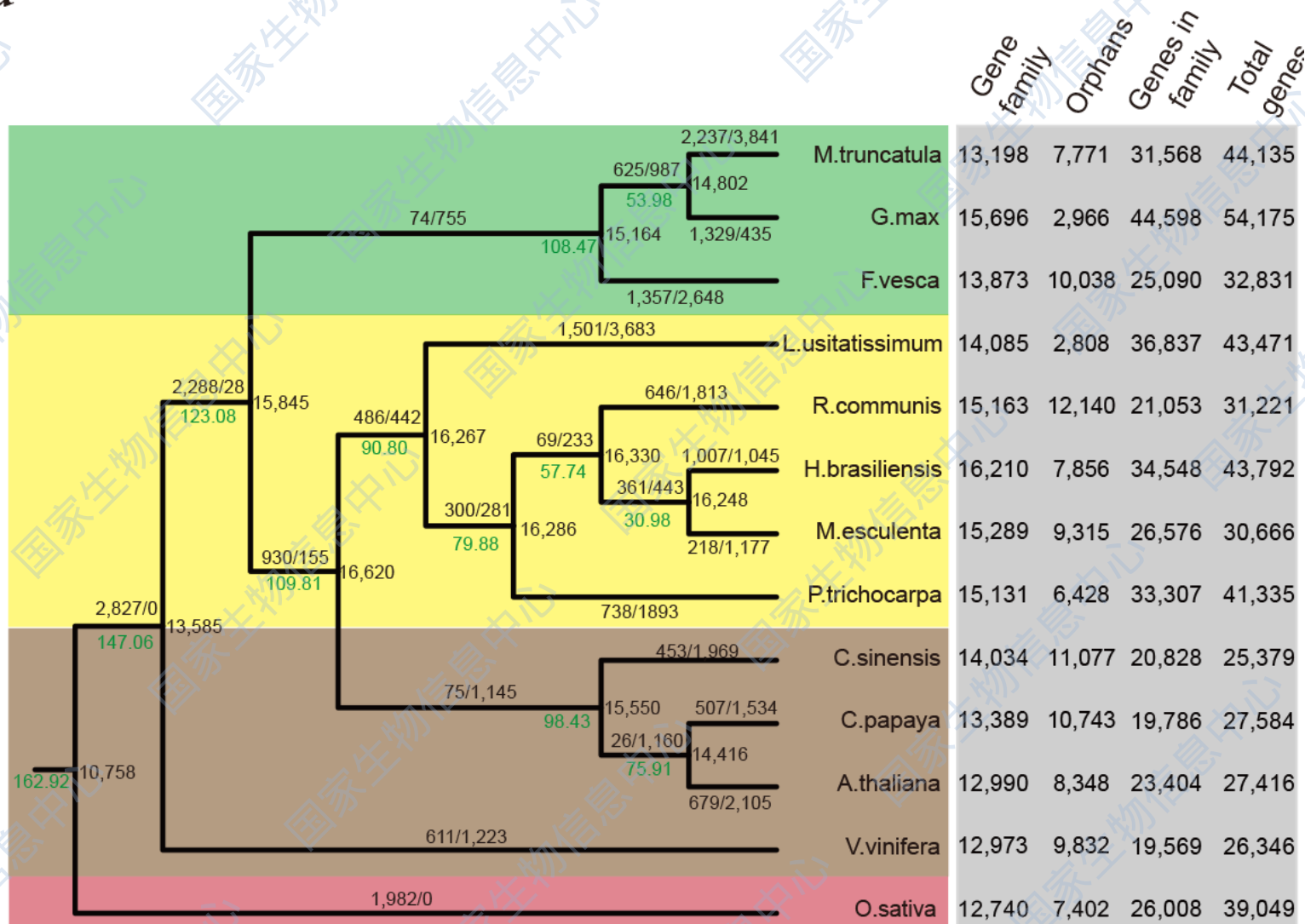
比较基因组分析基因组共线性





比较基因组分析物种进化

d





产胶生物学分析

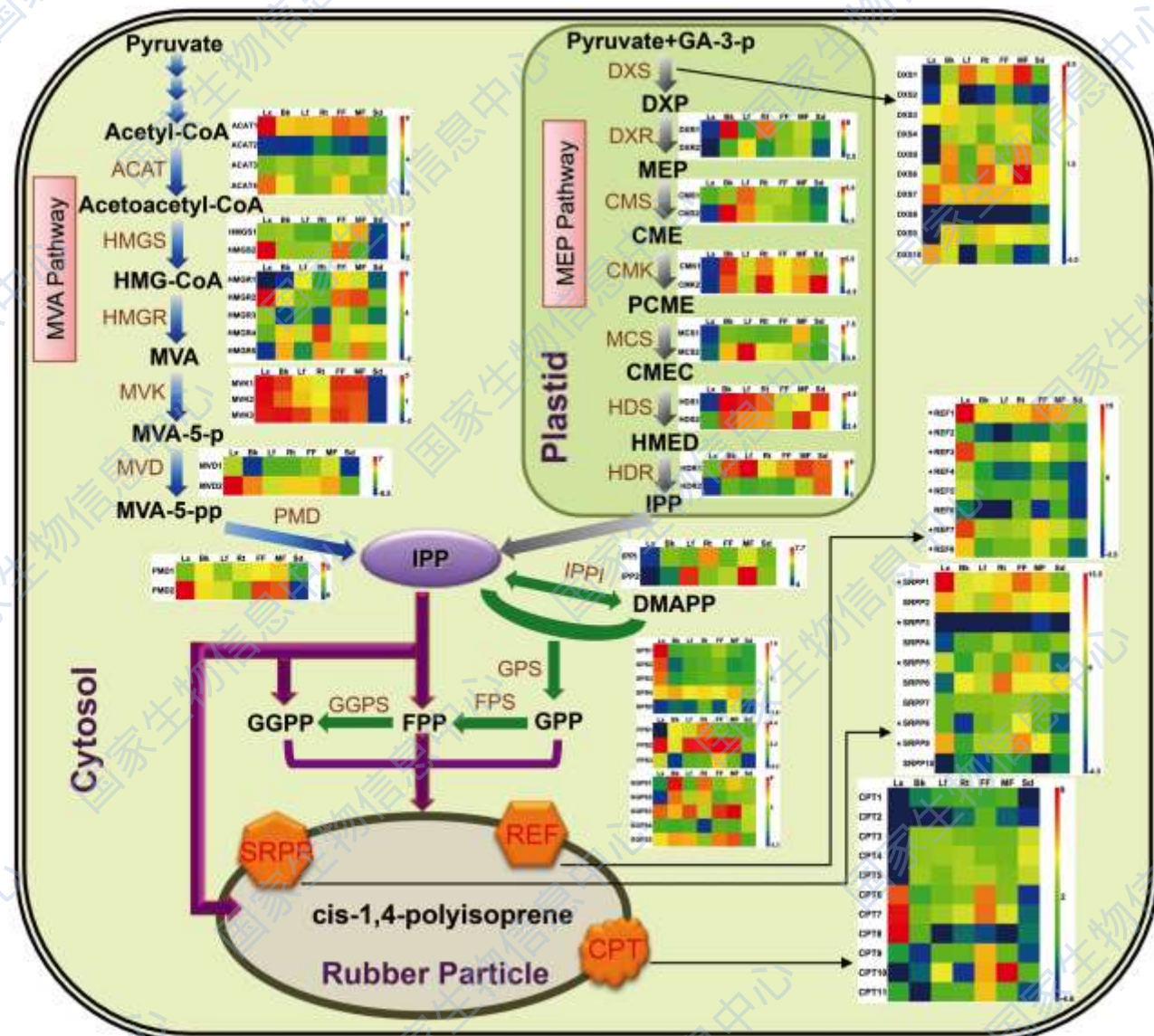
Gene duplication

REF/SRPP expansion
& divergence

High Rubber
Yielding

Active ethylene
signaling

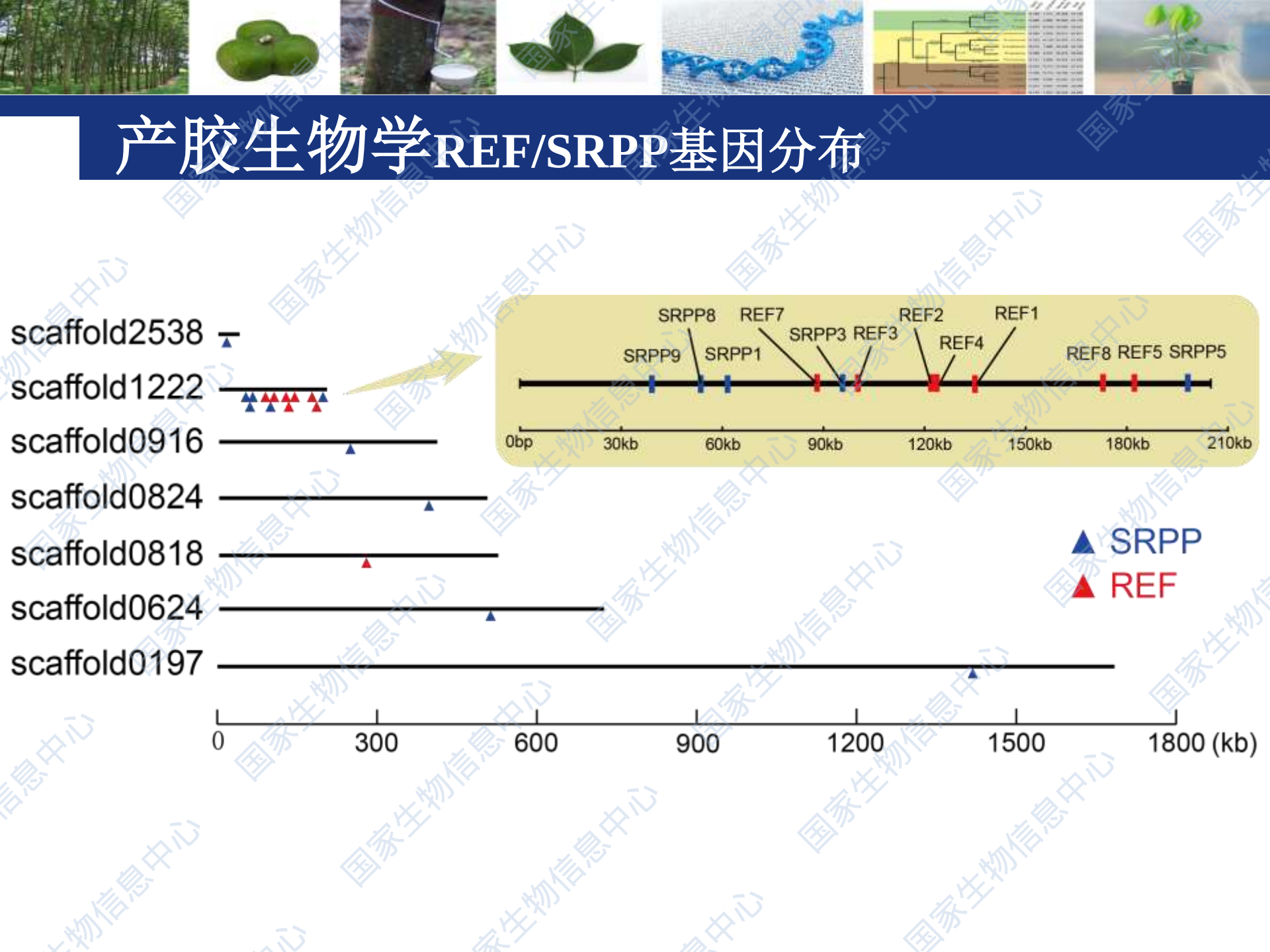
产胶生物学橡胶生物合成



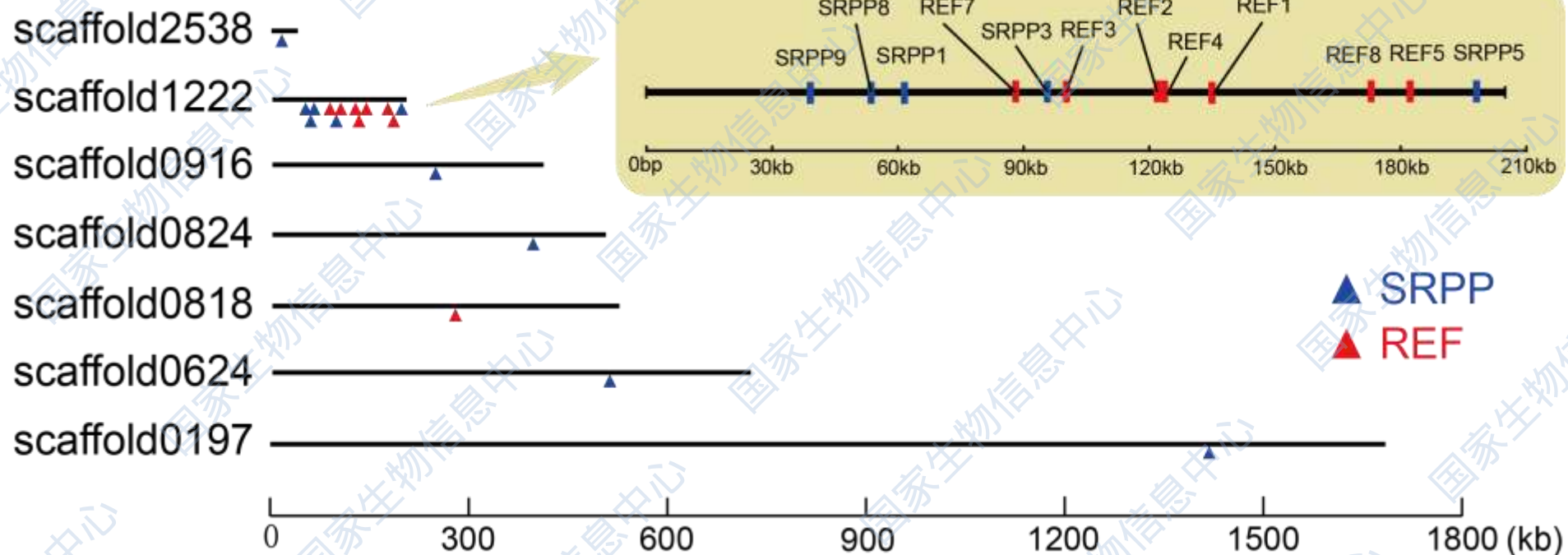


产胶生物学REF/SRPP基因数量

	Plant species	Genome size	Data origin	REF/SRPP number
Rubber-producing	<i>Hevea brasiliensis</i>	1.5 Gb	This study	18
	<i>Lactuca sativa</i>	2.59 Gb ¹	NCBI	8
	<i>Parthenium argentatum</i>	NA	NCBI (mRNA, EST)	6
	<i>Taraxacum</i>	NA	NCBI (Proteins) and reference ²⁰	6
	<i>brevicorniculatum</i>			
	<i>Taraxacum officinale</i>	1.25 Gb ²	NCBI (mRNA, EST)	3
Latex-producing	<i>Manihot esculenta</i>	742 Mb ³	Phytozome	6
	<i>Ricinus communis</i>	350.6 Mb ⁴	Phytozome	3
	<i>Jatropha curcas</i>	410 Mb ⁵	NCBI	3
	<i>Carica papaya</i>	370 Mb ⁶	Phytozome	3
	<i>Cannabis sativa</i>	534 Mb ⁷	Phytozome	2
Non rubber- or latex-producing	<i>Arabidopsis thaliana</i>	125 Mb ⁸	Phytozome	4
	<i>Populus trichocarpa</i>	480 Mb ⁹	NCBI	6
	<i>Gossypium hirsutum</i>	2.25-2.43 Gb ¹⁰	CGP	10
	<i>Gossypium arboreum</i>	1.7 Gb ¹¹	CGP	8
	<i>Gossypium raimondii</i>	775 Mb ¹²	CGP	7
	<i>Capsicum annuum</i>	3.48 Gb ¹³	Phytozome	3
	<i>Vitis vinifera</i>	490 Mb ¹⁴	Phytozome	2
	<i>Linum usitatissimum</i>	373 Mb ¹⁵	Phytozome	4
	<i>Oryza sativa</i>	466 Mb ^{16,17}	Phytozome	2
	<i>Zea mays</i>	2.3 Gb ¹⁸	Phytozome	1
	<i>Triticum aestivum</i>	17 Gb ¹⁹	IWGSC	4

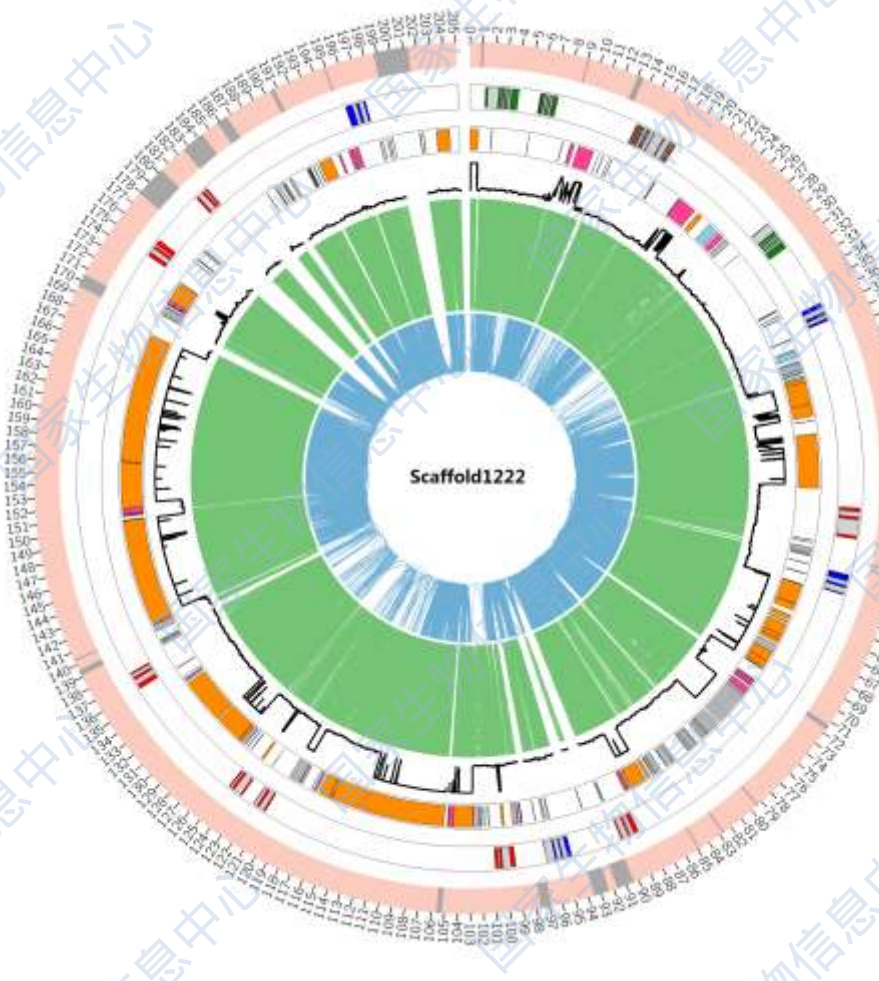


产胶生物学REF/SRPP基因分布

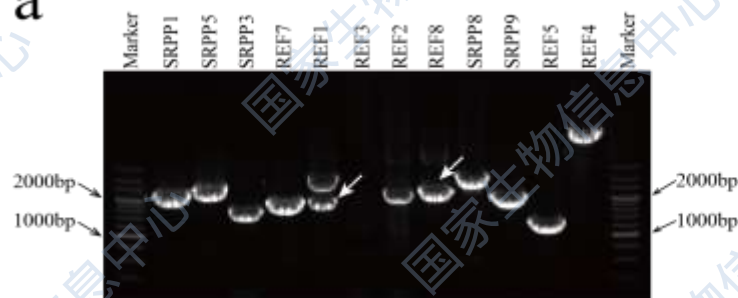




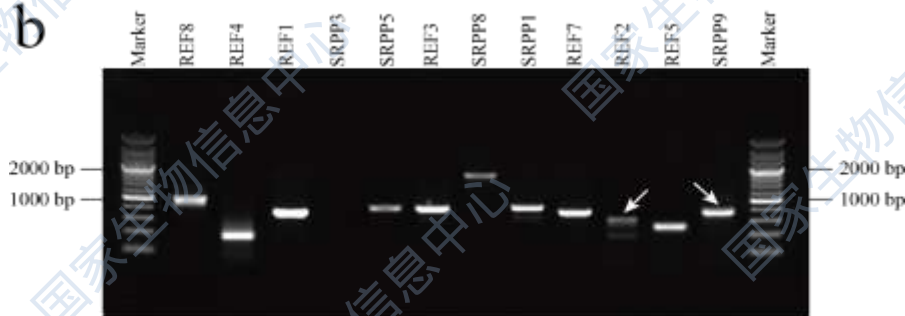
产胶生物学REF/SRPP基因簇验证



a

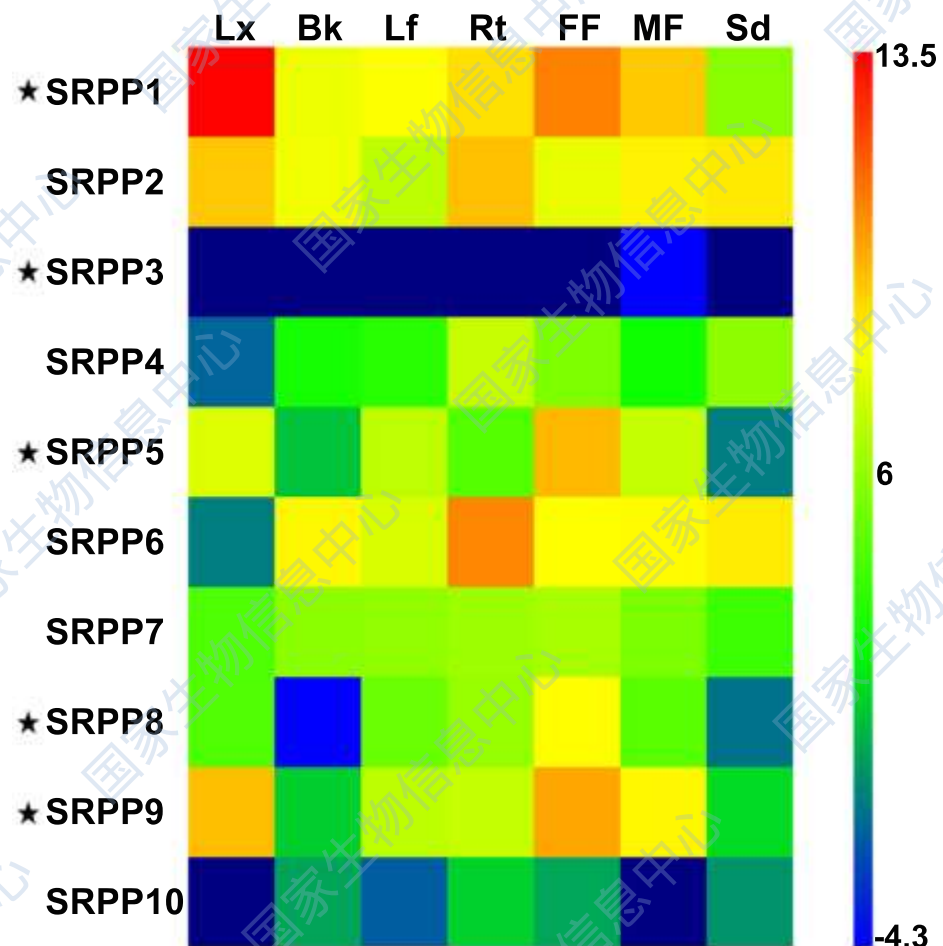
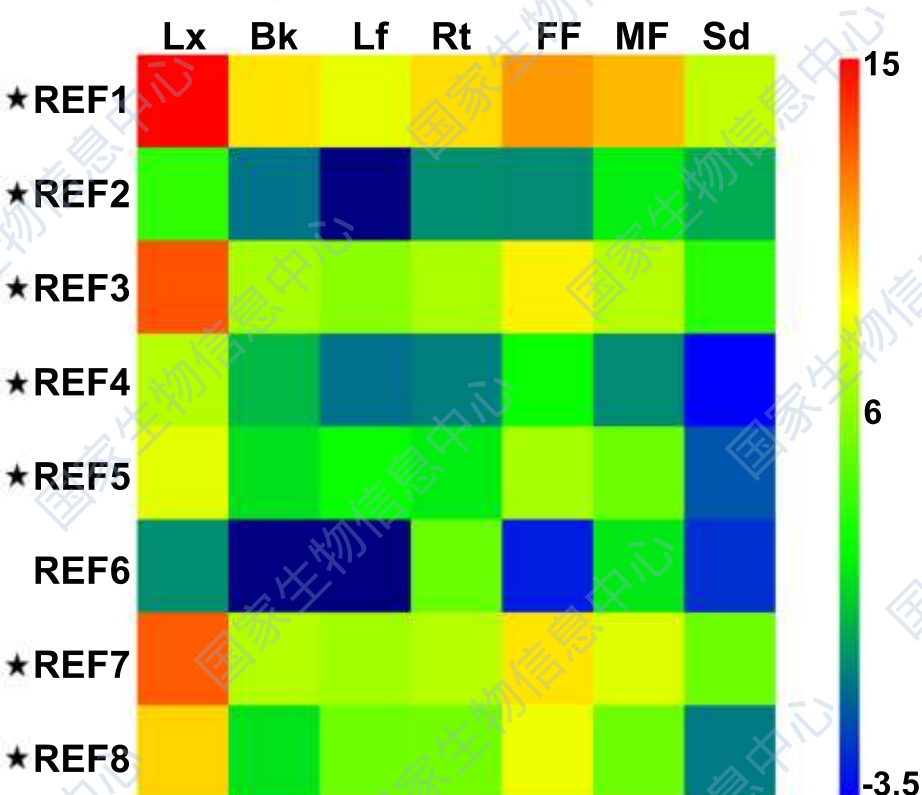


b



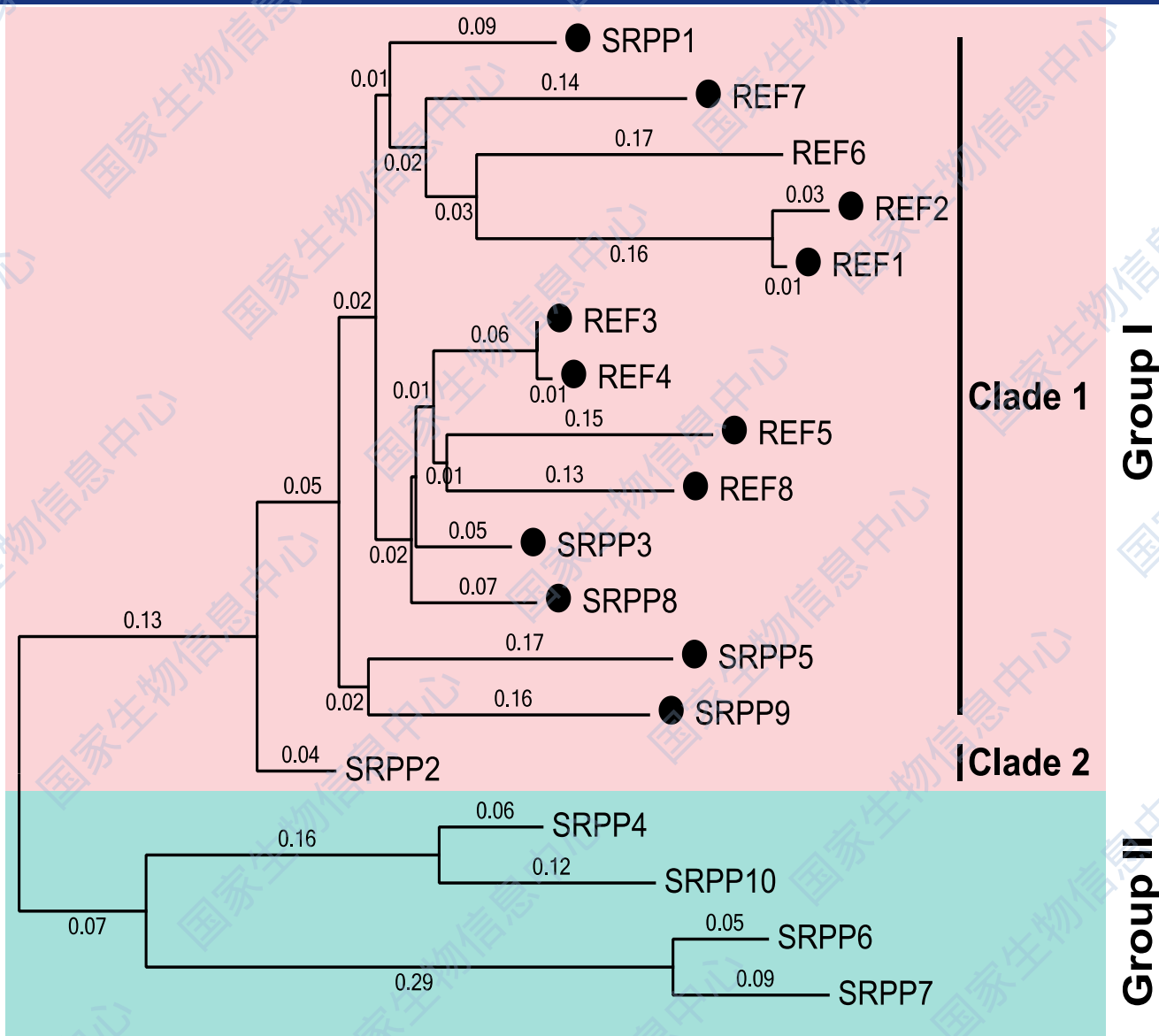


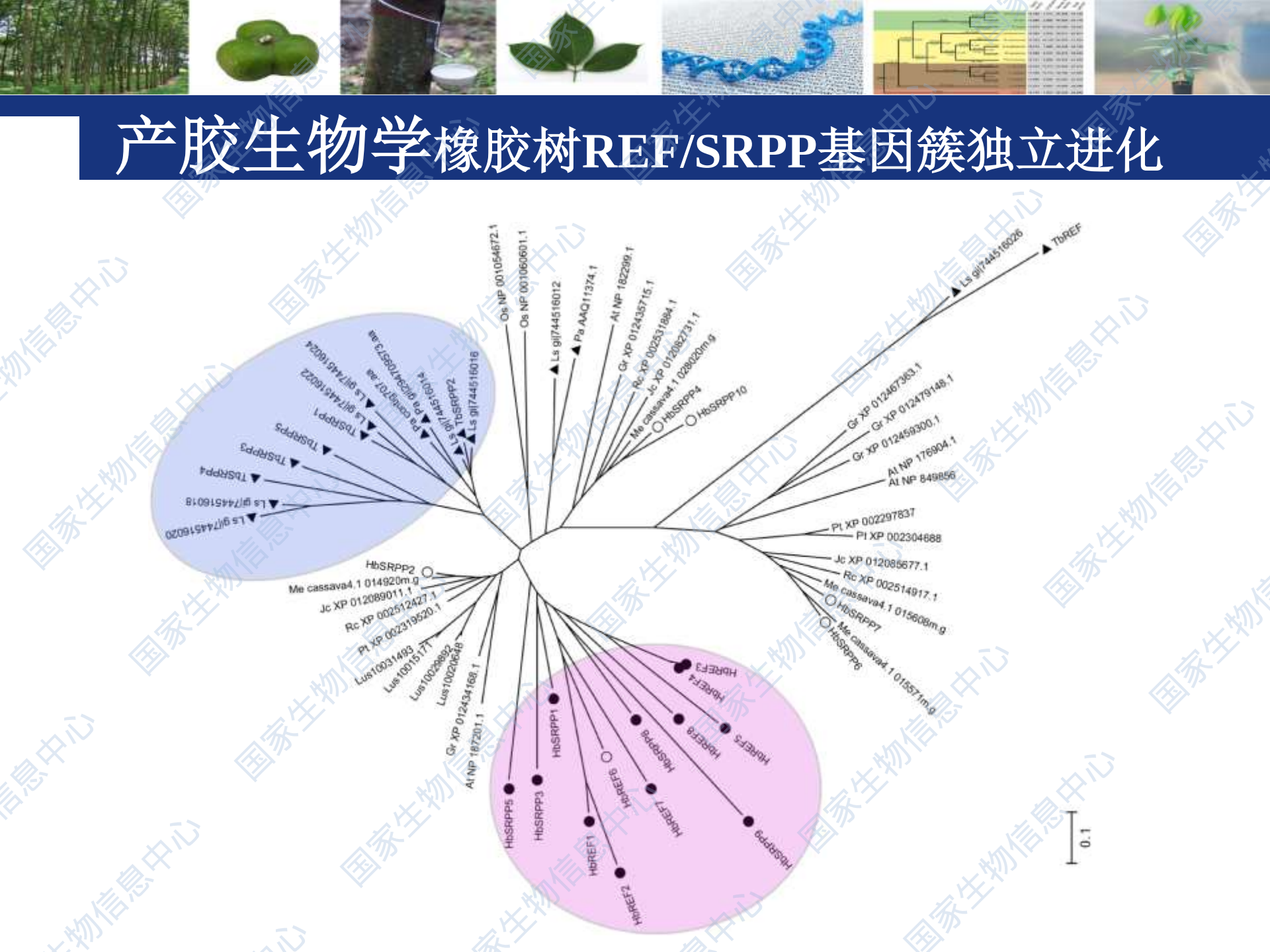
产胶生物学REF/SRPP基因乳管特异分化



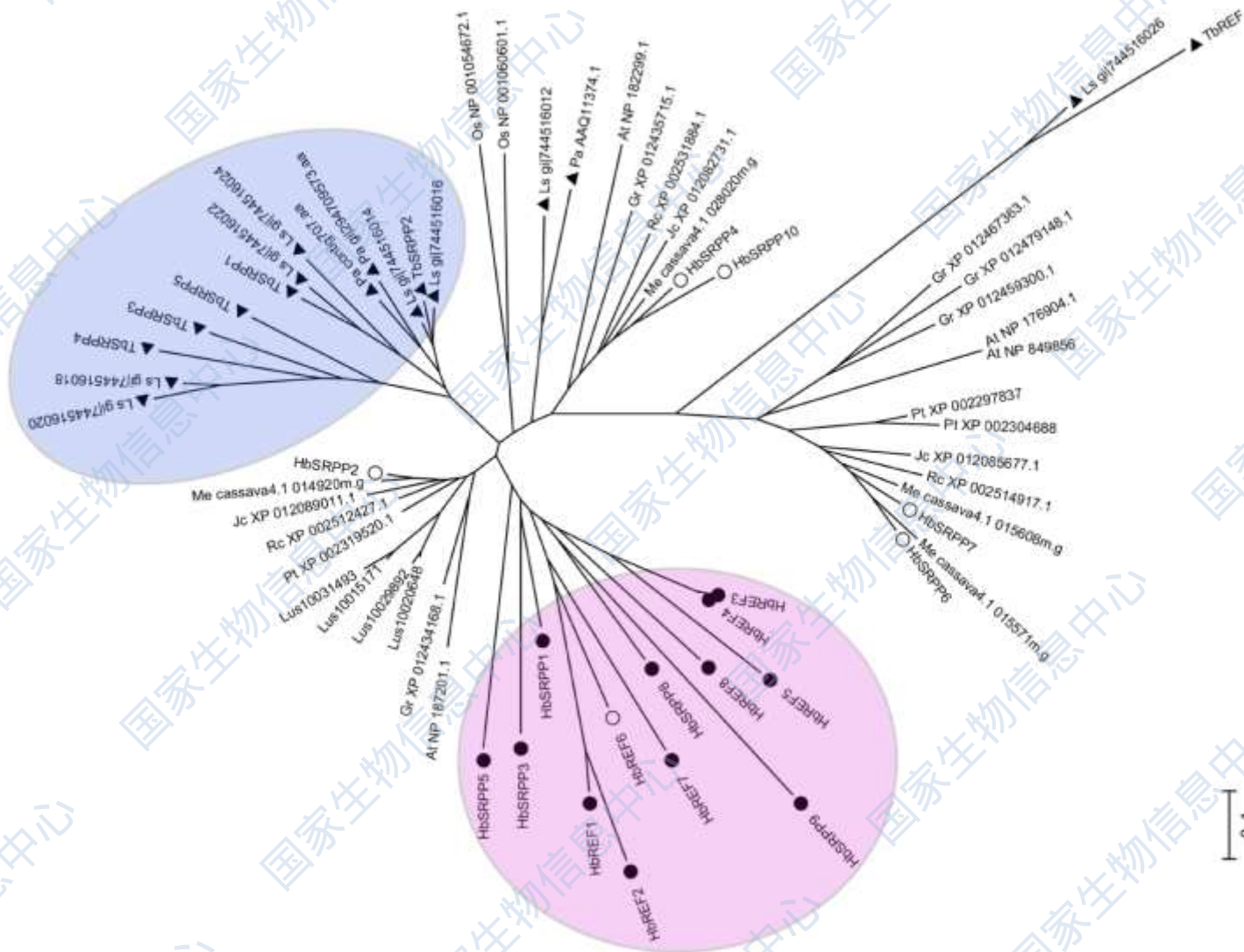


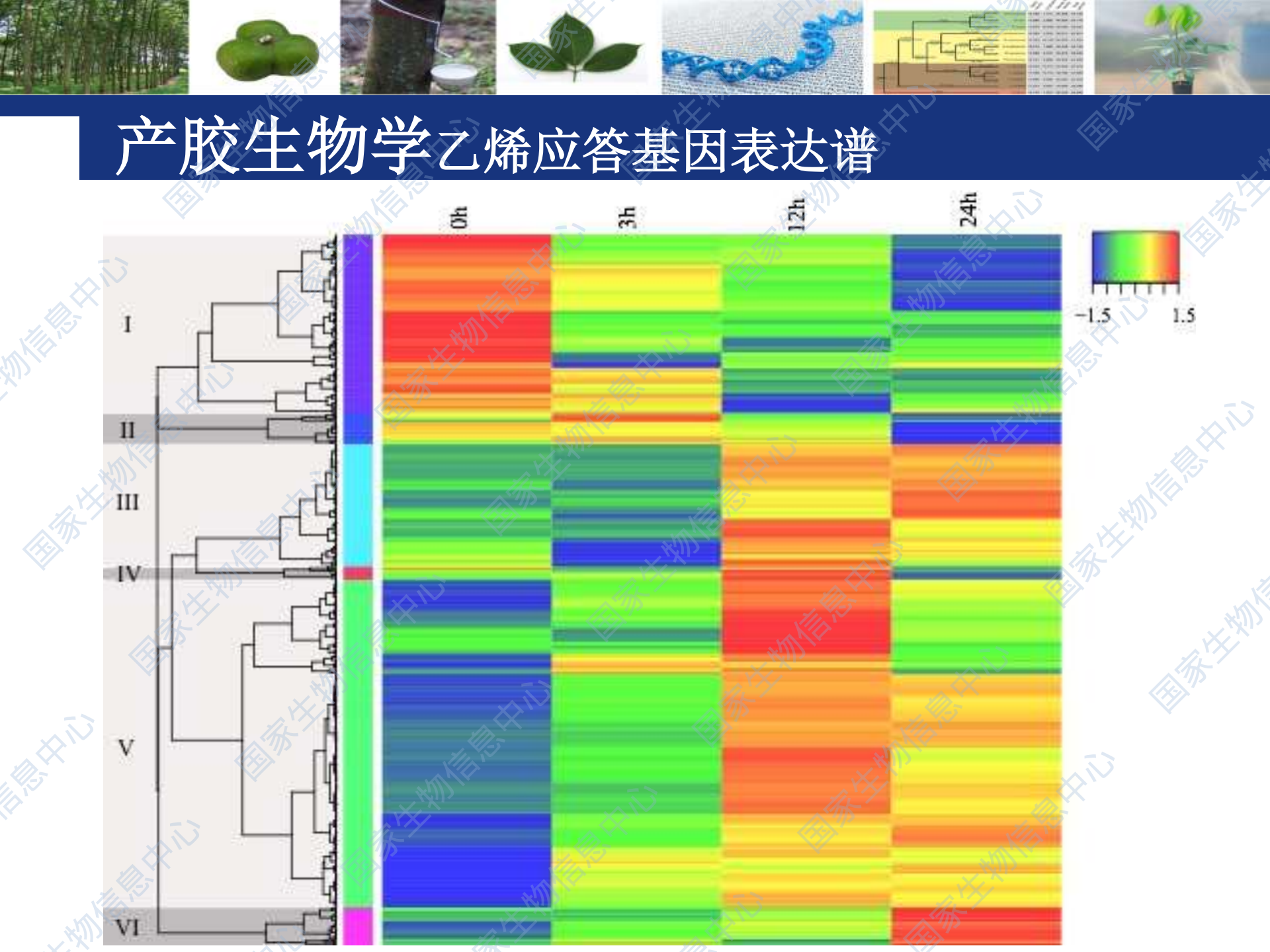
产胶生物学REF/SRPP家族进化



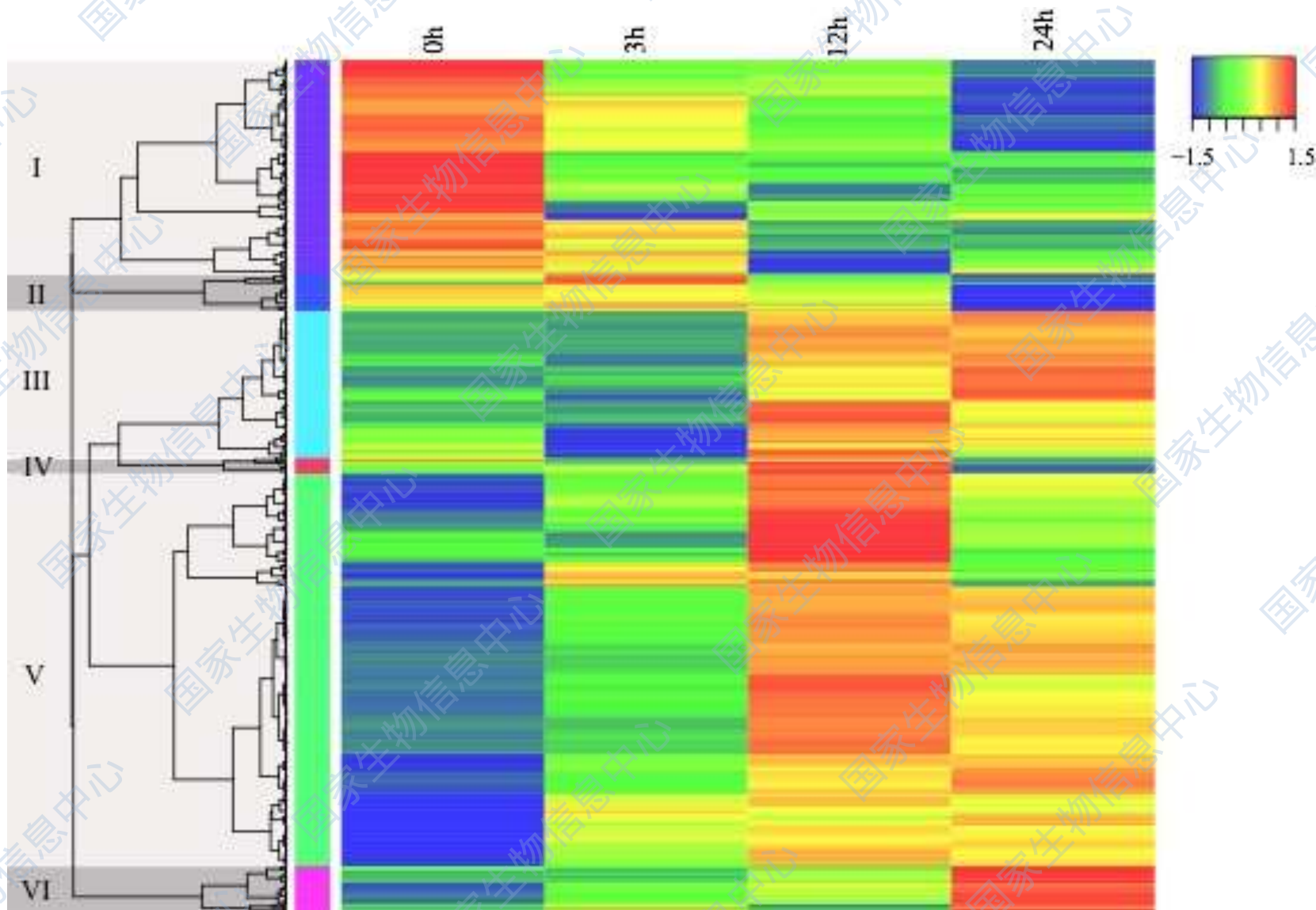


产胶生物学橡胶树REF/SRPP基因簇独立进化

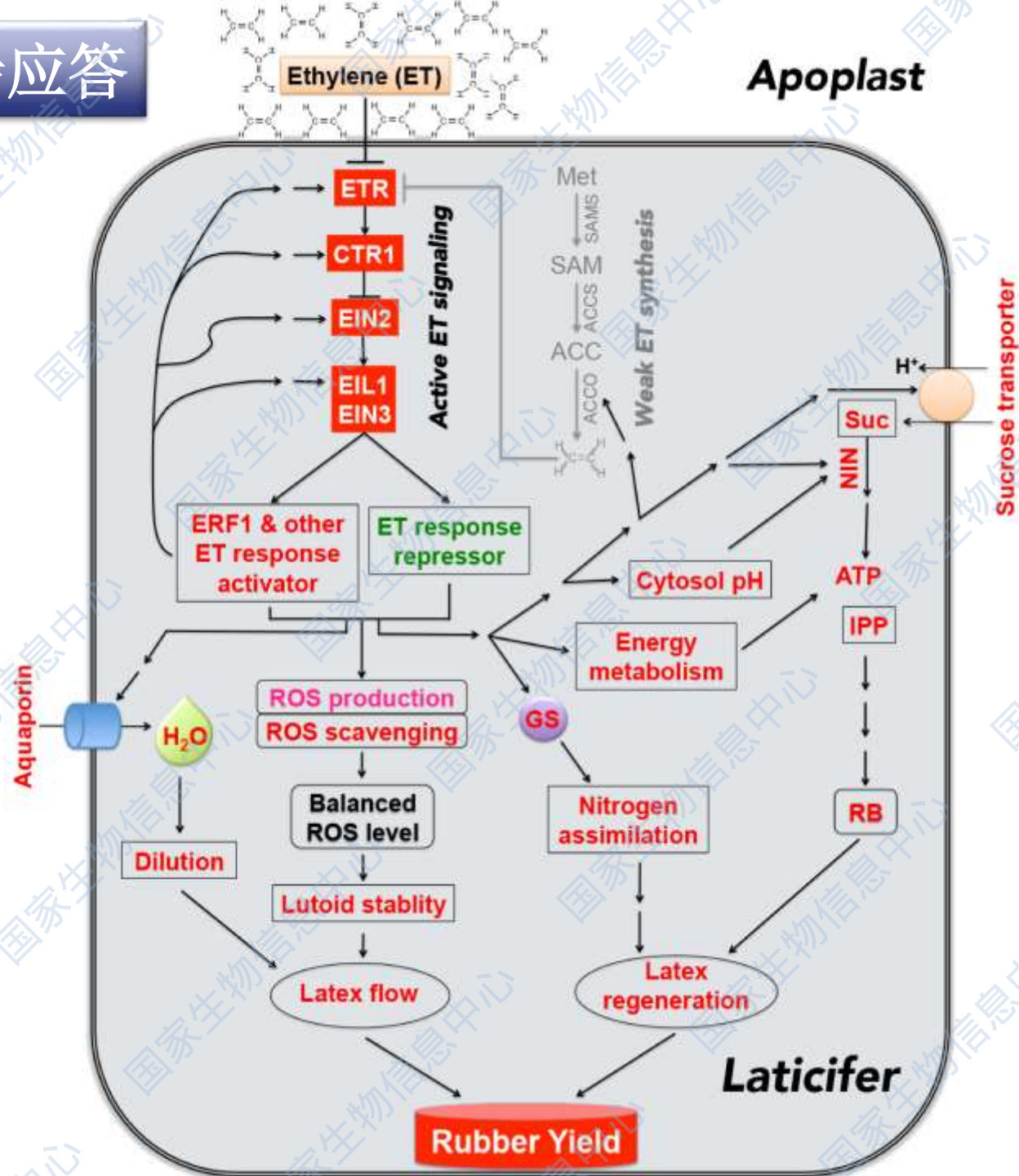




产胶生物学乙烯应答基因表达谱



乳管乙烯应答



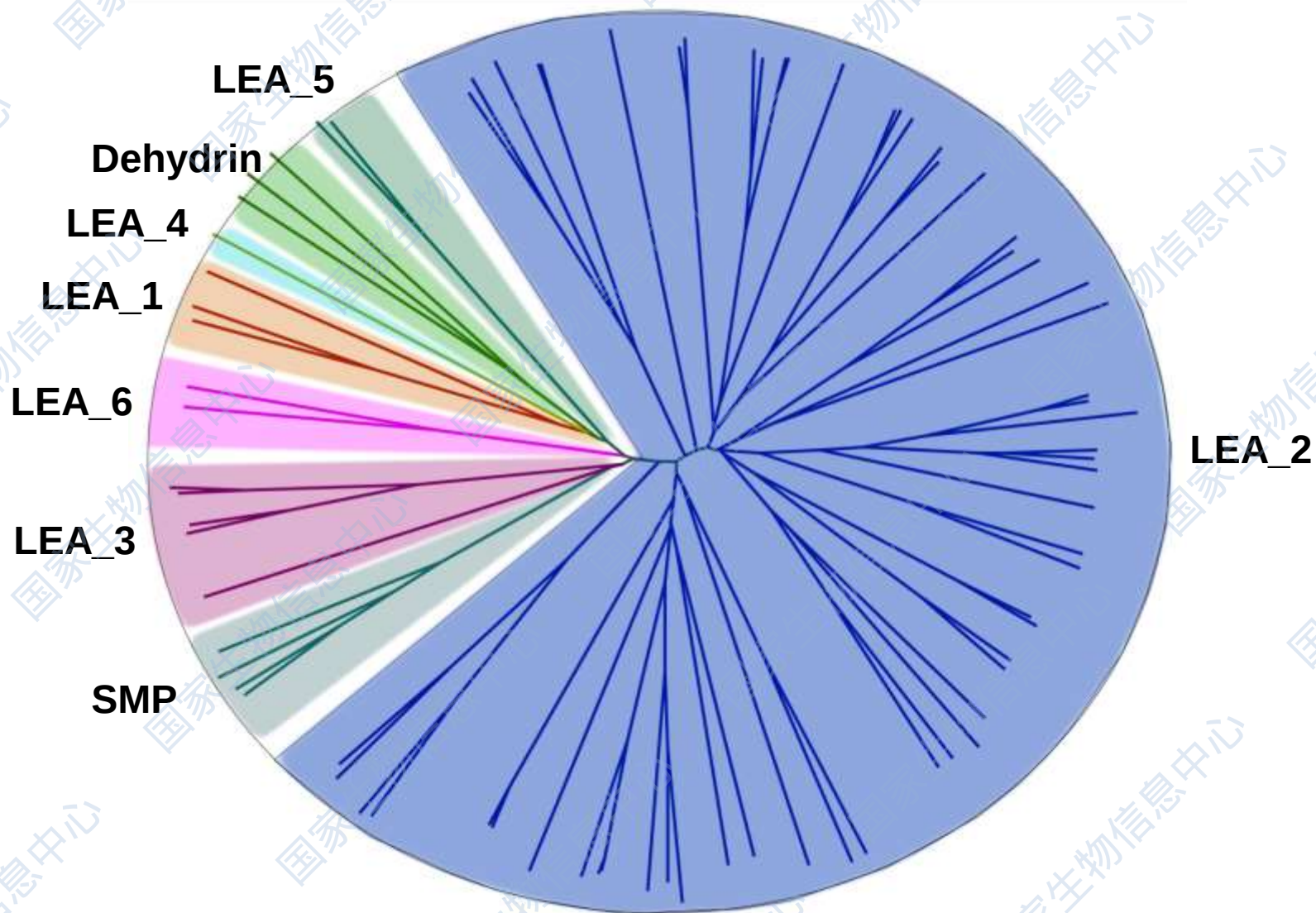
Date Palm (*Phoenix dactylifera* L.)



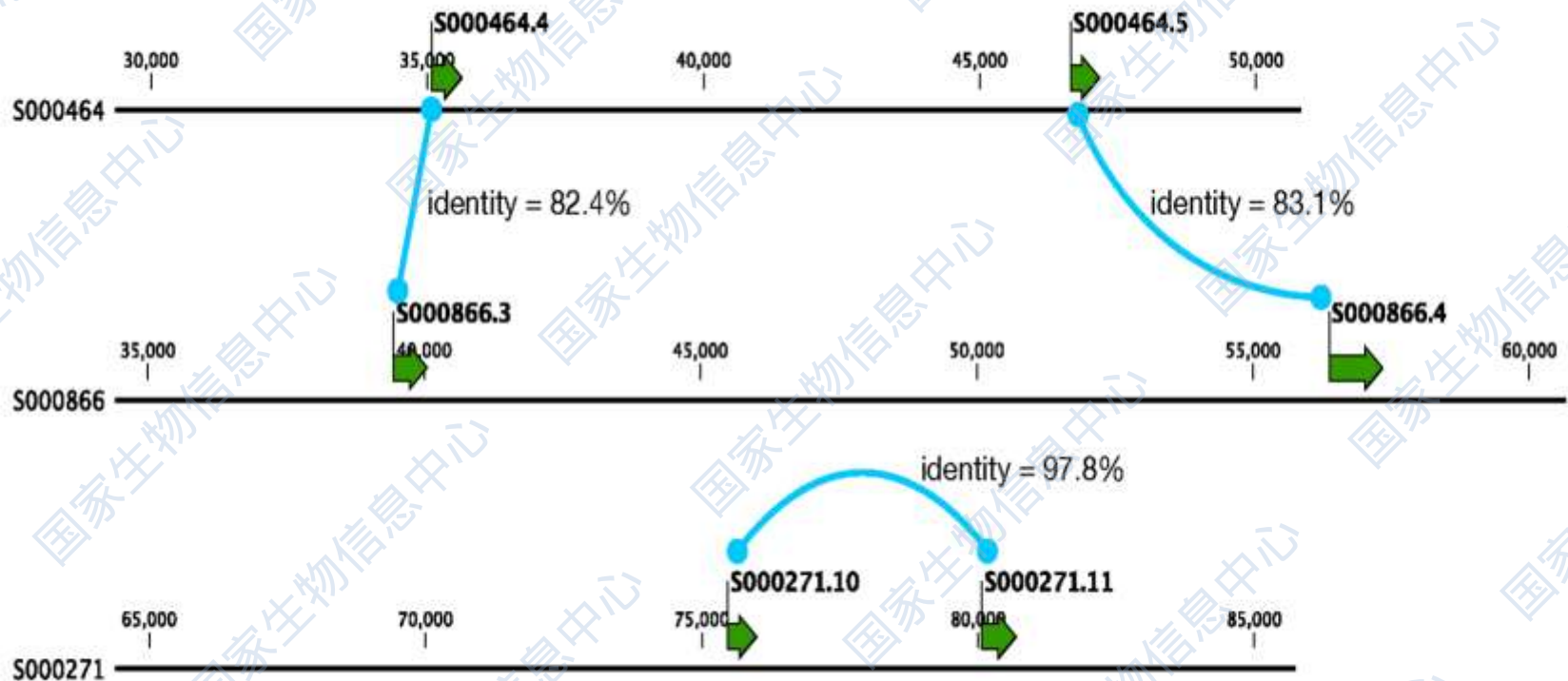
LEA family Found in Date Palm Genome

LEA family	Exon	<i>P. dactylifera</i>	<i>O.sativa</i>	<i>A.thaliana</i>	Size	MW(KDa)	Top4 AA(%)
Dehydrin	2.00	3	4	3	171	18.86	E(15.47)/G(14.78)/K(14.14.76)/H(9.37)
LEA_1	2.00	3	4	3	155.33	16.66	A(17.55)/E(11.83)/K(11.38)/T(8.06)
LEA_2	1.62	59	5	4	237.38	26.06	L(9.57)/V(9.03)/A(8.36)/S(8.35)
LEA_3	2.00	5	6	18	96.6	10.12	A(19.26)/S(12.12)/R(10.32)/L(7.24)
LEA_4	3.00	1	2	2	473.00	51.63	E(19.03)/A(15.86)/K(15.43)/G(8.46)
LEA_5	1.00	2	0	0	87.50	9.46	G(16.54)/E(15.40)/R(10.82)/A(8.60)
LEA_6	1.00	2	8	10	90.50	9.64	G(13.81)/E(8.86)/T(8.85)/P(8.30)
SMP	2.75	4	5	6	256.75	26.69	A(17.36)/G(9.72)/V(9.54)/R(8.31)
Total		79	34	51			

LEA family



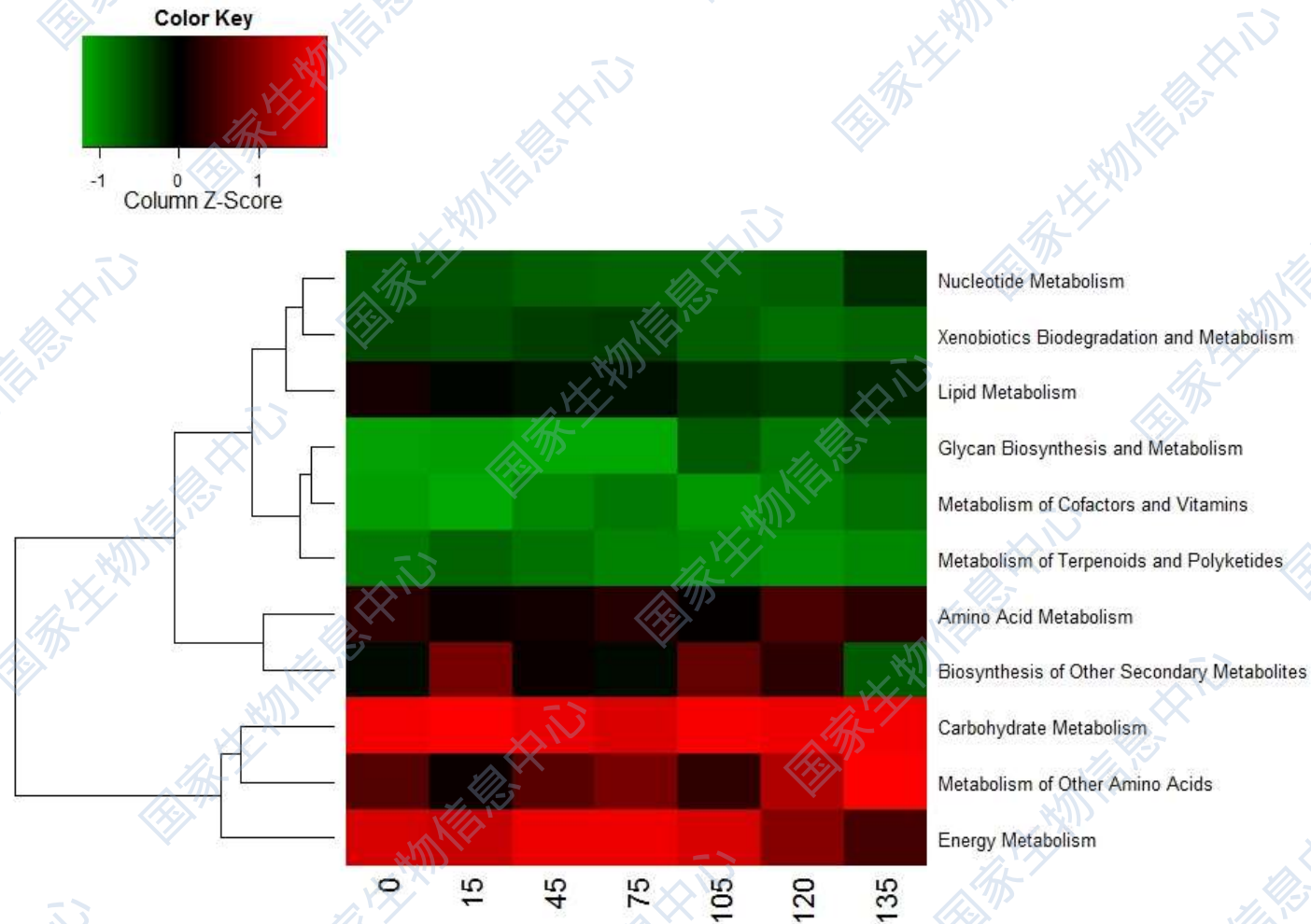
Expansion of LEA family



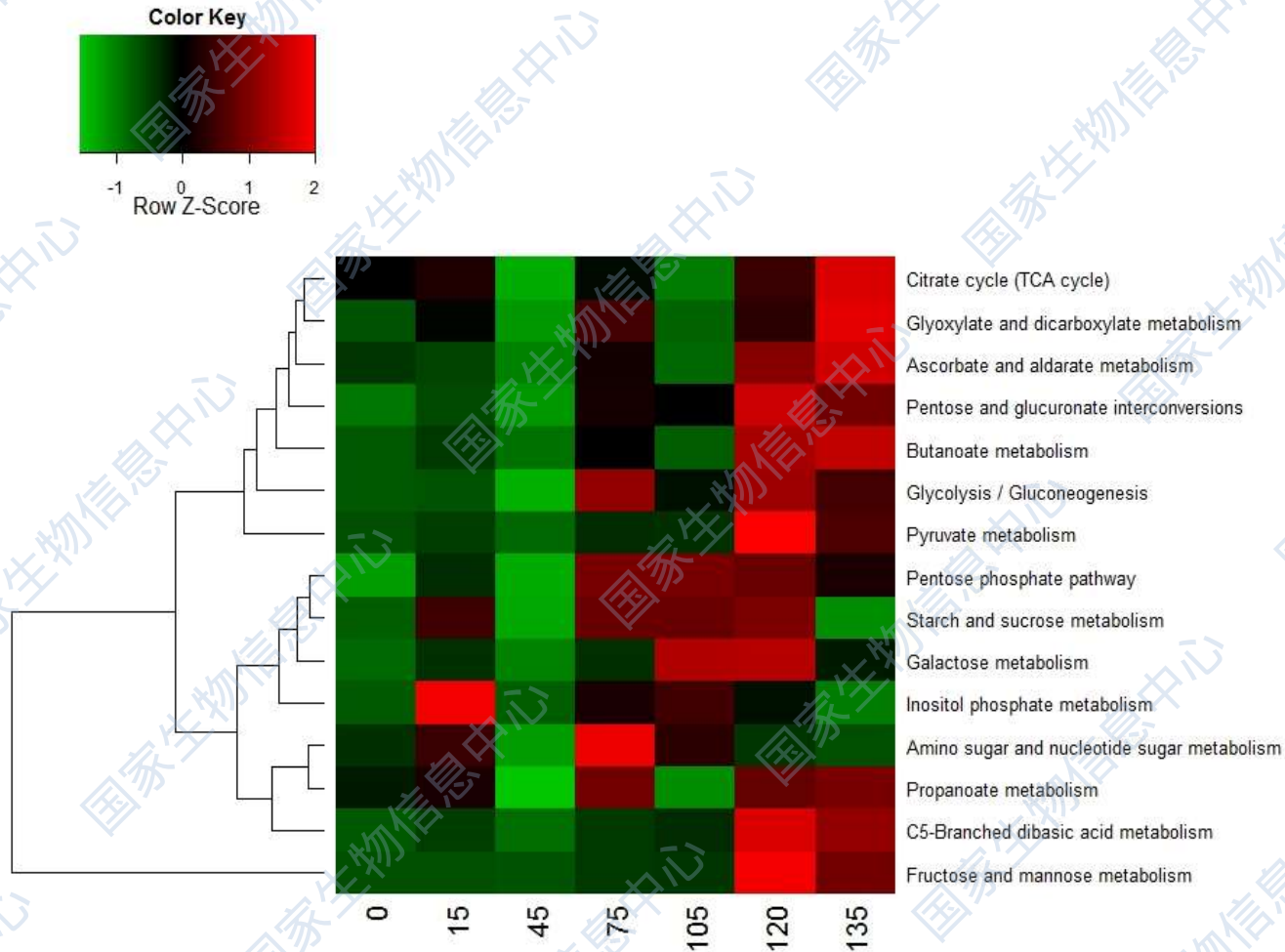
Overview of fruit transcriptomic data

Tissue	Total reads	Mapped reads	Unique mapped reads	Expressed mRNAs
0DPP	76,515,395	53,825,507	11,396,571	17,464
15DPP	83,950,995	59,009,944	9,903,023	17,570
45DPP	77,132,786	53,999,668	12,355,849	15,920
75DPP	74,369,980	55,665,445	14,302,703	18,329
105DPP	71,372,371	52,916,490	12,403,501	16,349
120DPP	83,506,278	62,428,925	14,407,421	15,709
135DPP	98,713,383	70,080,905	10,247,674	15,978
Green leaf	94,042,176	45,735,383	14,697,605	16,022

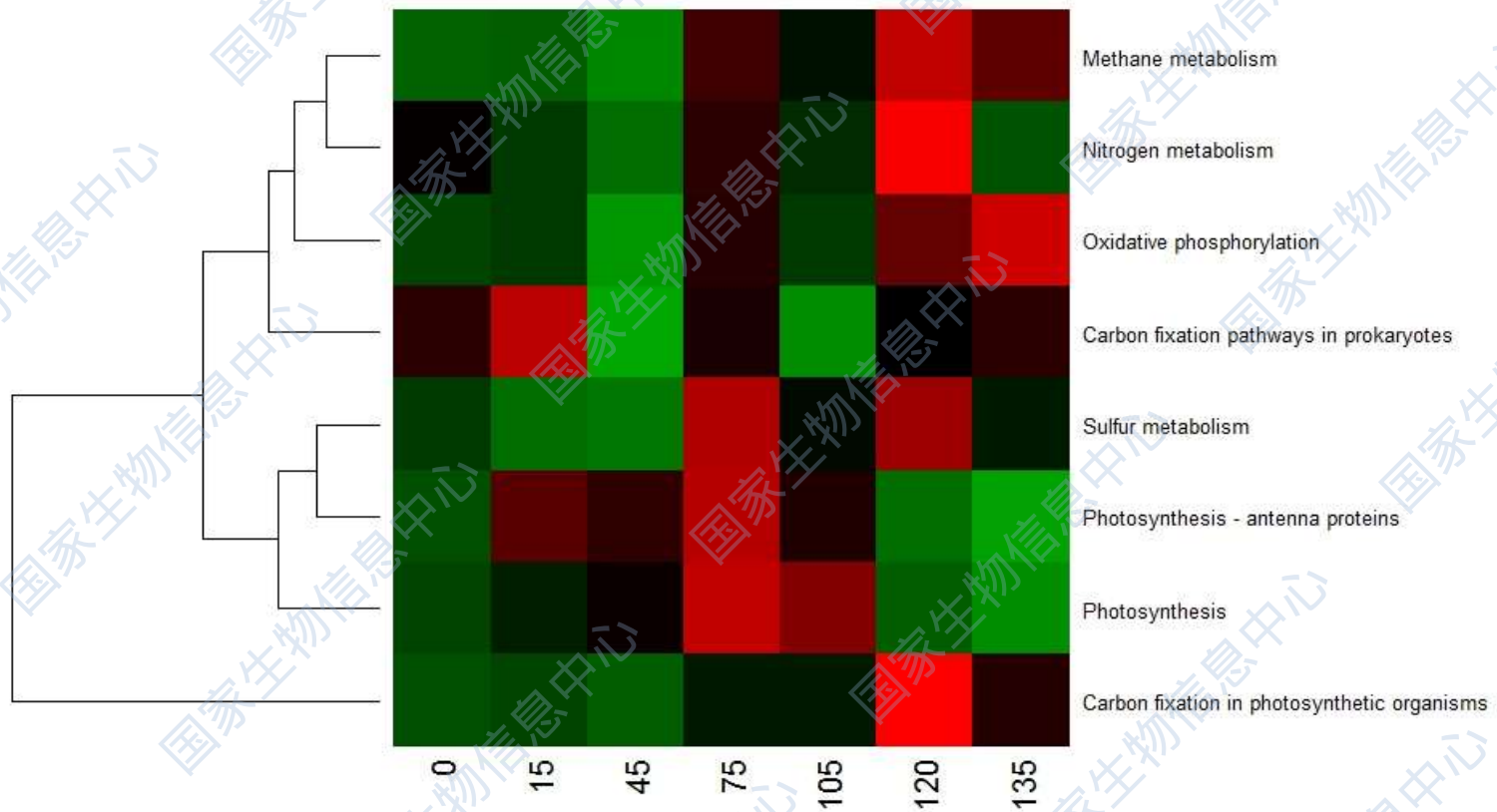
Gene Expression Profile of Metablism Pathway During Date Palm Fruit Ripening



Gene Expression Profile of Carbon Metabolic Pathway During Date Palm Fruit Ripening



Gene Expression Profile of Energy Metabolic Pathway During Date Palm Fruit Ripening

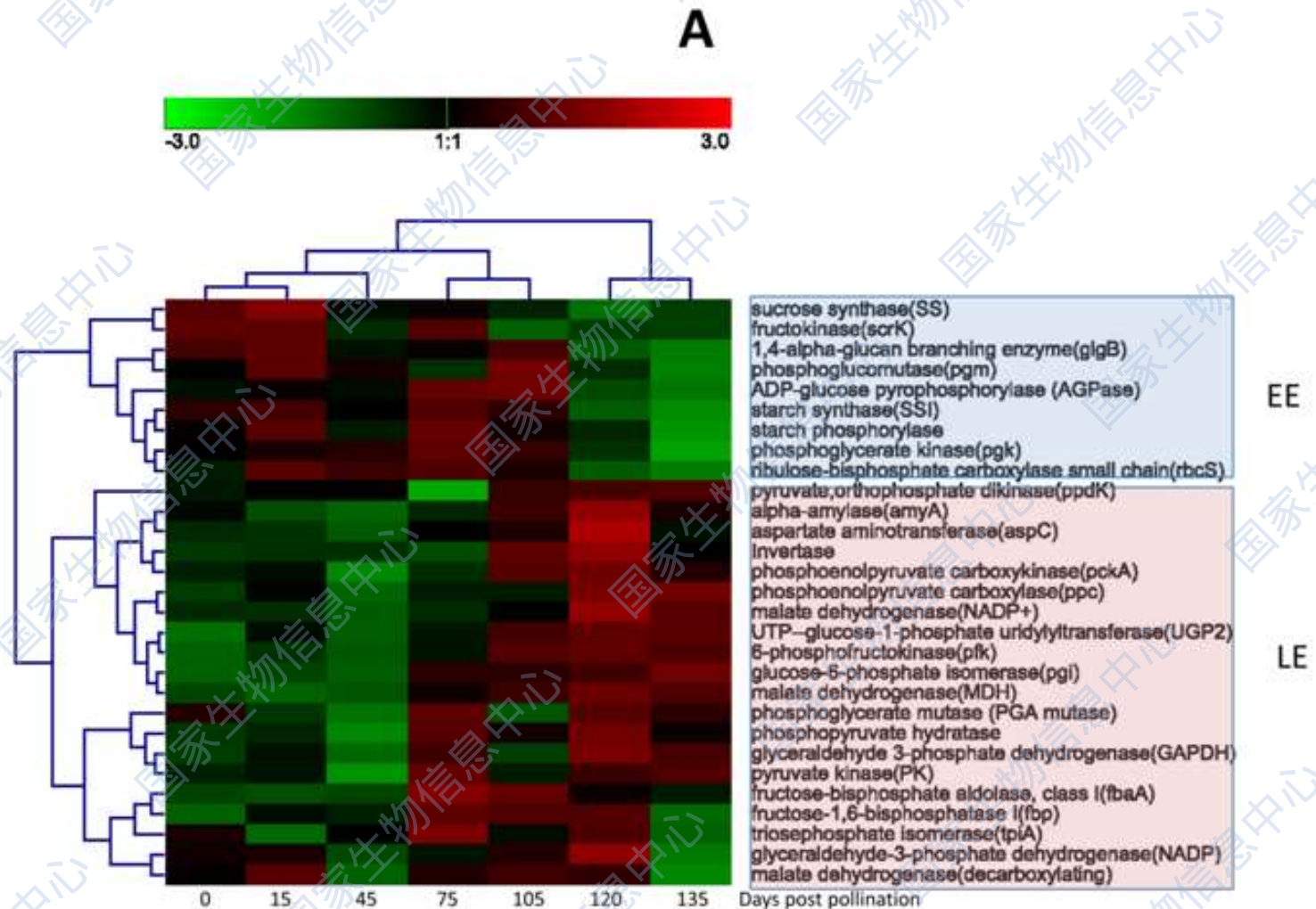


Enzymes Involved in Sugar Metabolism During Date Palm Fruit Development

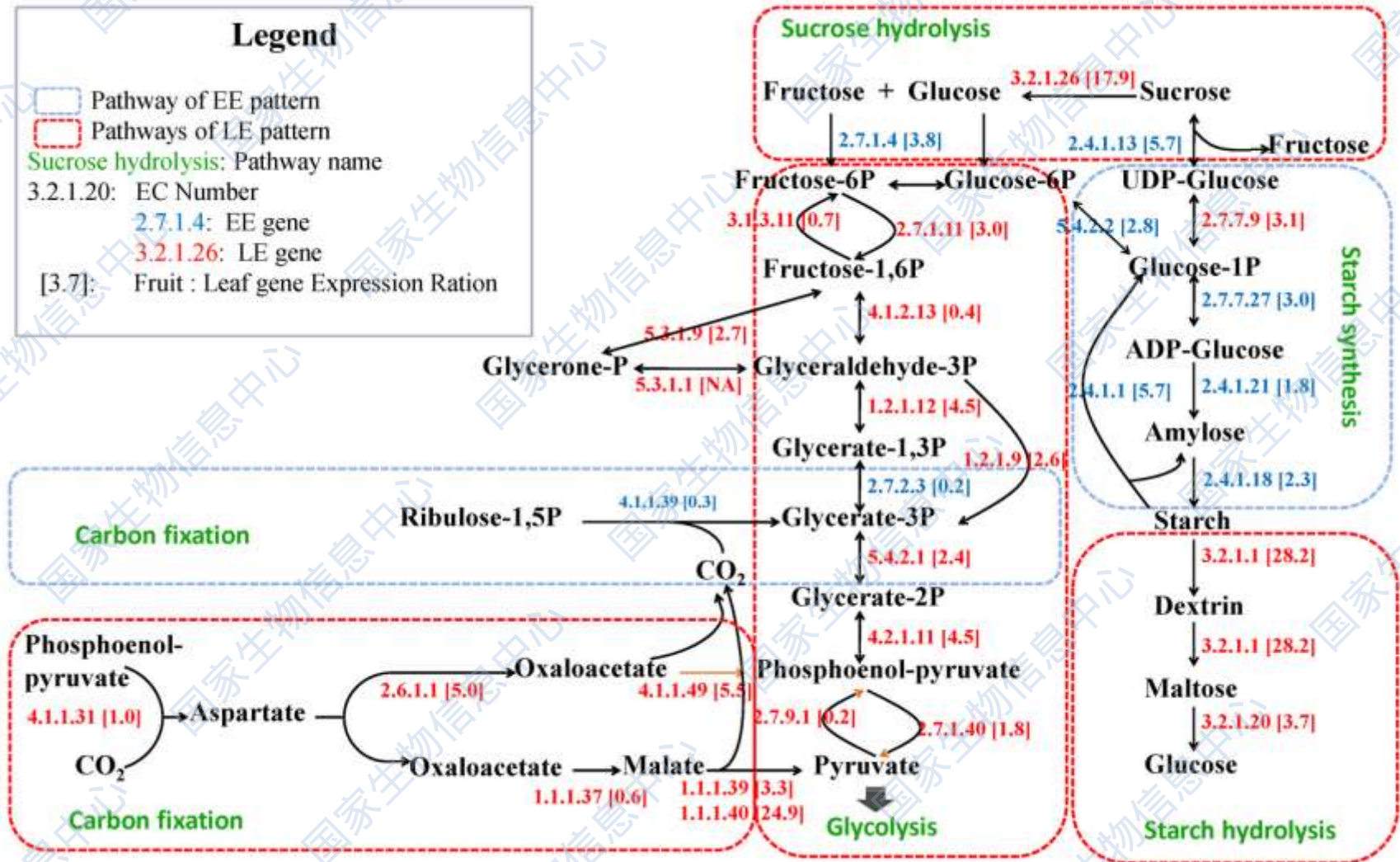
EC	Pathway	KO	Enzyme name	Days Post Pollination (RPKM normalized)							green leaf	Fruit/Leaf ¹
				0	15	45	75	105	120	135		
2.4.1.13	Sucrose metaoblism	K00695	sucrose synthase(SS)	191.23	322.32	58.85	85.68	46.68	19	33.66	56.49	5.7
3.2.1.26	Sucrose metaoblism	K01193	Invertase	20.96	23.5	23.42	15.47	239	542.56	51.76	30.25	17.9
2.7.7.9	Starch synthesis	K00963	UTP--glucose-1-phosphate uridylyltransferase (UGP2)	109.25	176.88	118.72	167.85	262.16	301.77	291.13	97.1	3.1
2.7.7.27	Starch synthesis	K00975	ADP-glucose pyrophosphorylase (AGPase)	14.81	17.9	15.21	30.21	32.46	10.64	8.15	10.59	3.1
2.4.1.21	Starch synthesis	K00703	starch synthase (SSI)	38.66	45.85	26.11	51.78	38.11	12.16	8.17	29.48	1.8
2.4.1.18	Starch synthesis	K00700	1,4-alpha-glucan branching enzyme (glgB)	31.7	38.05	16.98	20.89	32.57	9.99	7.76	16.3	2.3
3.2.1.1	Starch hydrolysis	K01176	alpha-amylase (amyA)	6.66	1.71	1.15	3.65	24.32	120.89	13.63	4.29	28.2
2.4.1.1	Starch hydrolysis	K00688	starch phosphorylase	83.19	150.47	63.02	168.5	89.41	57.41	24.67	29.6	5.7
2.7.1.4	Glycosis	K00847	fructokinase (scrK)	30.73	30.06	12.04	28.36	8.46	10.77	10.45	8.15	3.8
4.1.2.13	Glycosis	K01623	fructose-bisphosphate aldolase, class I (fbaA)	220.99	215	220.84	374.67	344.3	276.53	246.91	997.93	0.4
5.4.2.2	Glycosis	K01835	phosphoglucomutase (pgm)	58.61	106.72	69.4	44.21	124.85	45.27	26.48	44.05	2.8
5.3.1.9	Glycosis	K01810	glucose-6-phosphate isomerase (pgi)	76.24	89.15	77.65	165.75	162.54	182.81	235.56	86.18	2.7
2.7.1.11	Glycosis	K00850	6-phosphofructokinase (pfk)	91.64	110.56	94.81	124.96	190.71	200.22	186.49	67.46	3
3.1.3.11	Glycosis	K03841	fructose-1,6-bisphosphatase I (fbp)	9.17	14.58	13.28	26.95	24.9	24.9	8.73	41.34	0.7
5.4.2.1	Glycosis	K01834	phosphoglycerate mutase (PGA mutase)	131.74	84.14	62.41	172.6	72.72	169.65	136.45	72.93	2.4
4.2.1.11	Glycosis	K01689	phosphopyruvate hydratase	206.47	218.57	156.91	337.4	268.9	368.99	262.89	81.9	4.5
5.3.1.1	Glycosis	K01803	triosephosphate isomerase (tpiA)	153.58	90.26	137.05	231.68	129.81	190.8	92.43	0	/
2.7.2.3	Glycosis	K00927	phosphoglycerate kinase (pgk)	10.54	12.95	14.4	22.93	16.46	6.96	2.64	92.5	0.2
1.2.1.12	Glycosis	K00134	glyceraldehyde 3-phosphate dehydrogenase (GAPDH)	556.12	639.58	485.47	836.38	557.07	1025.21	855.87	225.72	4.5
1.2.1.9	Glycosis	K00131	glyceraldehyde-3-phosphate dehydrogenase (NADP)	13.73	14.25	8.4	11.18	15.08	25.97	6.64	9.84	2.6
4.1.1.39	Carbon fixation	K01602	ribulose-bisphosphate carboxylase small chain (rbcS)	14.69	82.07	53.07	72.29	33.71	5.84	3.46	263.17	0.3
4.1.1.31	Carbon fixation	K01595	phosphoenolpyruvate carboxylase (ppc)	115.76	148.45	96.71	133.59	134.51	255.14	251.61	258.37	1
1.1.1.40	Carbon fixation	K00029	malate dehydrogenase (NADP+)	233.2	218.7	117.36	231.55	398.52	3486.55	1446.91	140.14	24.9
2.7.9.1	Carbon fixation	K01006	pyruvate, orthophosphate dikinase (ppdK)	2.55	3.24	3.42	0.53	6.02	7.37	7.98	53.13	0.2
2.6.1.1	Carbon fixation	K00813	aspartate aminotransferase (aspC)	124.16	116.24	88.96	168.96	202.28	547.91	146.27	108.83	5
4.1.1.49	Carbon fixation	K01610	phosphoenolpyruvate carboxykinase (pckA)	15.9	21.39	3.28	13.42	102.73	153.29	47.2	27.66	5.5
2.7.1.40	Carbon fixation	K00873	pyruvate kinase (PK)	162.16	168.8	115.19	243.93	164.98	215.26	228.51	133.91	1.8
1.1.1.37	Carbon fixation	K00026	malate dehydrogenase (MDH)	54.62	48.07	51.03	81.53	99.59	123.77	109.35	201.3	0.6
1.1.1.39	Carbon fixation	K00028	malate dehydrogenase (decarboxylating)	77.97	100.75	55.81	85.24	93.49	89.59	49.19	30.59	3.3

Sugar Metabolism Related Genes

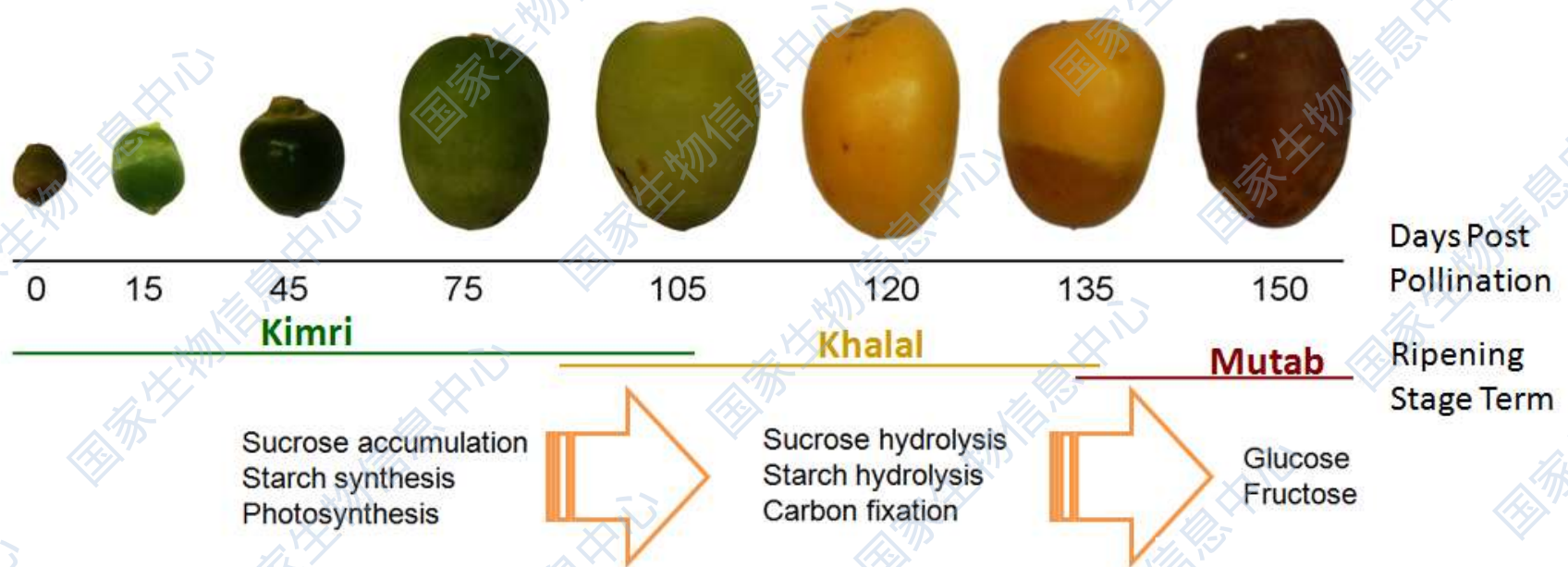
Clustering



Sugar Metablism in Dates Ripening



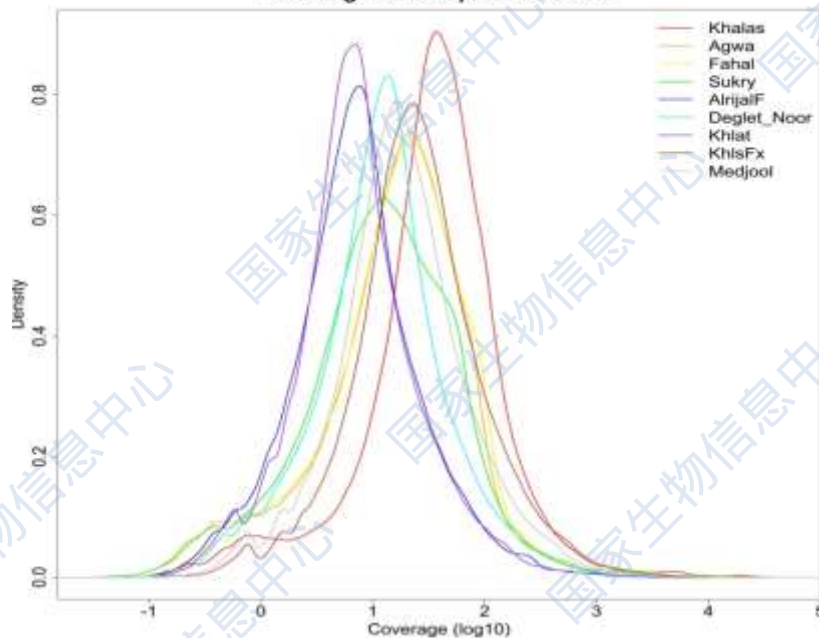
Sugar Accumulation in Date Palm Fruits



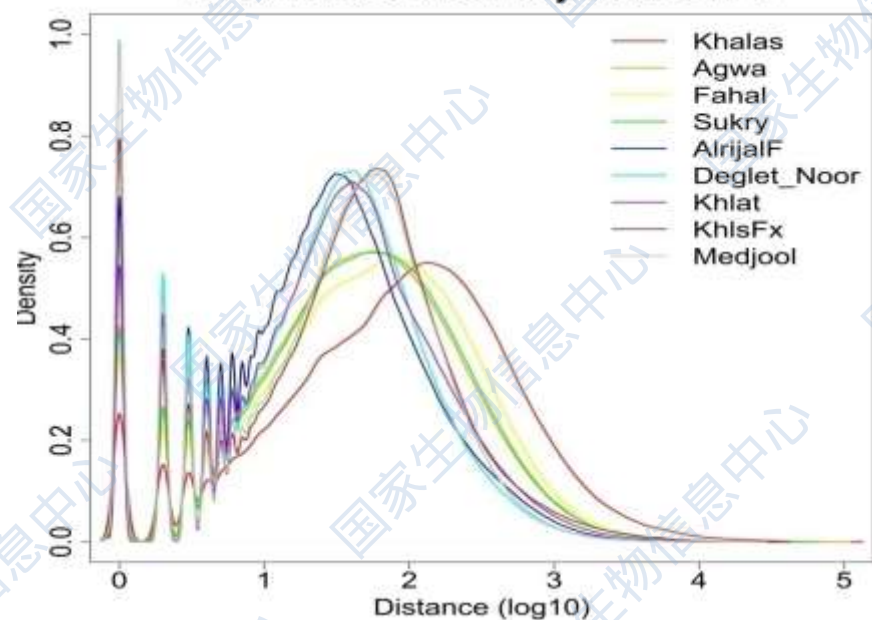
Summary of diversity among 9 date palm cultivars

Cultivars	Origin	Sample collection site	Coverage	SNP	SNP_Filter	SNP frequency (SNPs/Kb)	Indel	Indel_filter	Indelfrequency (indels/Kb)
Khalas	KSA	Al-hssa oasis,KSA	114	1702677	1370243	2.43303354	696625	56463	0.100257
Agwa	KSA	Al-hssa oasis,KSA	85	3982346	3429964	6.09031934	909160	138270	0.245515
Fahal(M)	KSA	Al-hssa oasis,KSA	92	3729689	3046293	5.40906469	770735	86495	0.153582
Sukry	KSA	Al-hssa oasis,KSA	64	4090785	3433597	6.09677017	704840	113057	0.200746
AlrijalF	Qatar	Qatar	15	9133871	6436196	11.4282508	-	-	-
Deglet_Noor	North Africa	California, USA	24	11395695	7054985	12.5269862	-	-	-
Khalt(M)	Qatar	Qatar	12	7960218	5085547	9.03000884	-	-	-
KhalsFx	KSA/USA	California, USA	39	6789130	4257243	7.55925408	-	-	-
Medjool	North Africa	California, USA	28	10211201	6175899	10.9660618	-	-	-

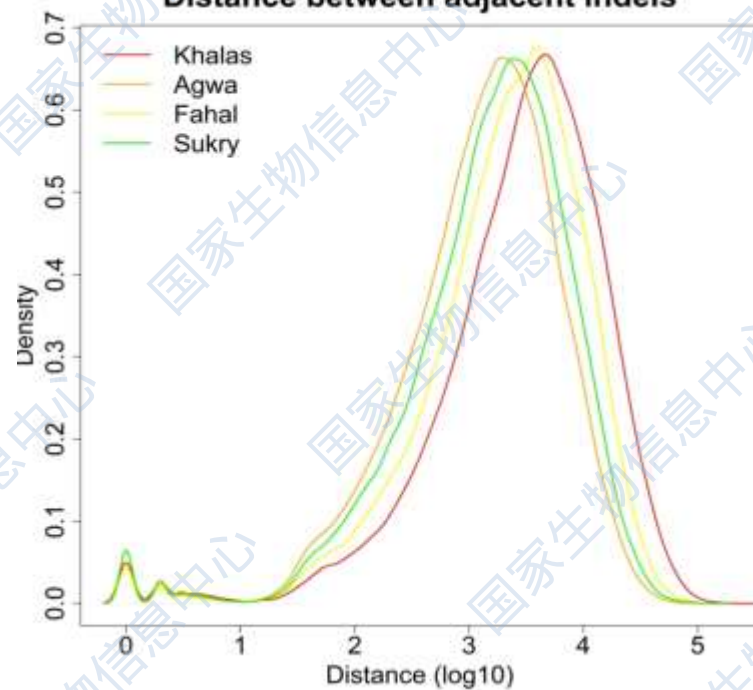
Coverage of date palm cultivars



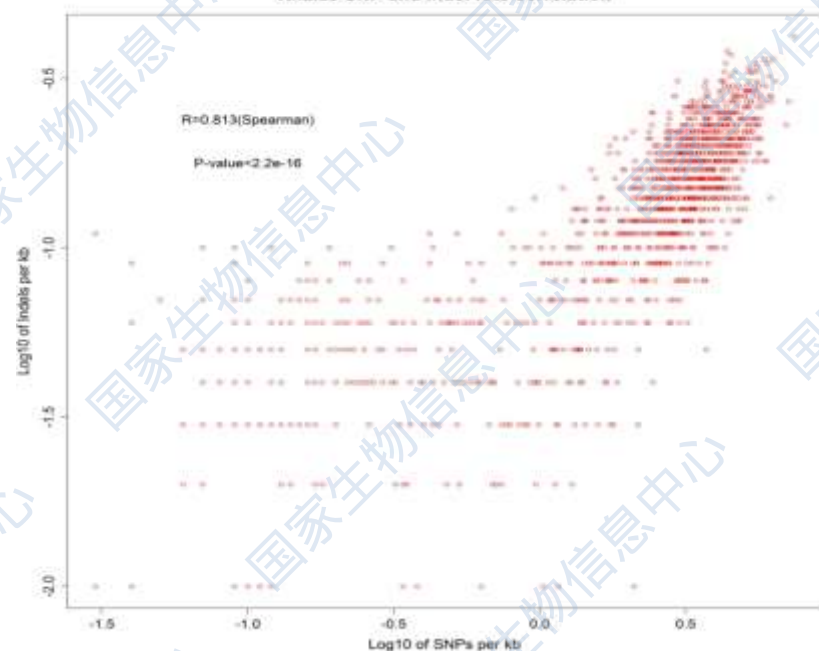
Distance between adjacent SNPs



Distance between adjacent indels



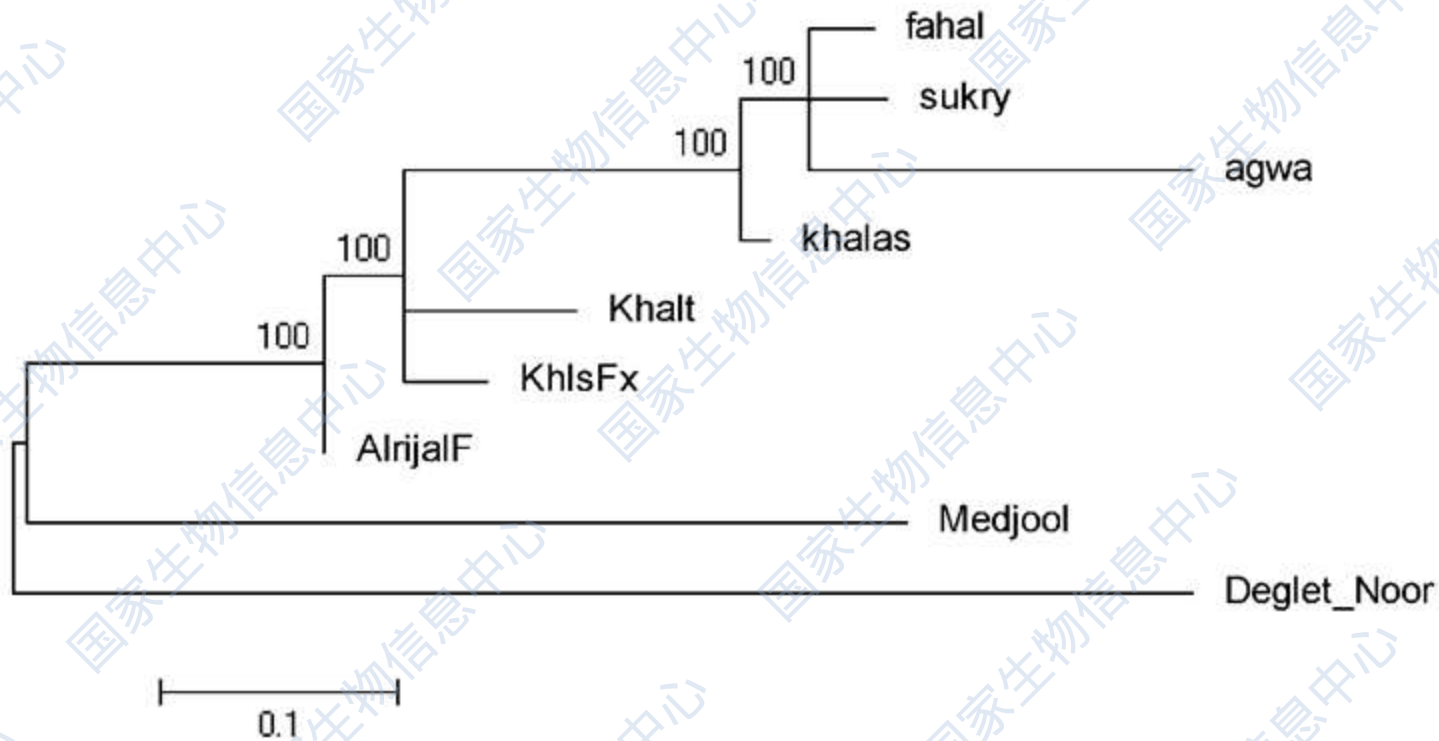
Khalas SNP and indel rate correlation



Overview of SNP transition, SNP transversion, insertion and deletion among 9 cultivars

Percent in SNPs/indels									
Type	Khalas	Agwa	Fahal	Sukry	AlrijalF	Deglet_Noor	Khalt	KhalsFx	Medjool
Transitions	83.5%	84.1%	84.4%	83.3%	86.1%	84.3%	86.1%	84.3%	83.7%
AC	3.7%	3.6%	3.7%	3.7%	3.5%	4.1%	3.3%	3.9%	4.3%
CA	4.0%	3.8%	3.7%	2.8%	3.7%	3.0%	3.4%	2.7%	3.0%
AG	16.6%	17.3%	17.6%	17.5%	16.6%	17.3%	16.1%	17.9%	17.3%
GA	17.5%	17.3%	17.2%	17.2%	19.3%	17.5%	20.2%	17.3%	16.9%
CT	17.4%	17.3%	17.2%	17.2%	19.3%	17.5%	20.2%	17.2%	17.0%
TC	16.6%	17.4%	17.7%	17.6%	16.6%	17.5%	16.2%	18.1%	17.5%
GT	4.1%	3.7%	3.7%	3.7%	3.6%	3.5%	3.4%	3.5%	3.6%
TG	3.7%	3.6%	3.7%	3.7%	3.5%	3.9%	3.3%	3.8%	4.1%
Transversions	16.5%	15.9%	15.6%	16.7%	13.9%	15.7%	13.9%	15.7%	16.3%
AT	5.3%	5.1%	4.9%	5.1%	4.4%	4.6%	4.5%	4.7%	4.8%
TA	5.4%	5.1%	4.9%	5.1%	4.4%	4.6%	4.5%	4.7%	4.8%
CG	2.9%	2.8%	2.9%	2.8%	2.5%	3.0%	2.4%	2.7%	3.0%
GC	2.9%	2.8%	2.9%	3.7%	2.5%	3.5%	2.4%	3.5%	3.6%
Insertions	41.9%	39.3%	39.8%	38.7%	-	-	-	-	-
A	13.5%	12.9%	12.4%	12.2%	-	-	-	-	-
C	5.9%	5.8%	6.1%	6.3%	-	-	-	-	-
G	5.9%	6.0%	6.1%	6.4%	-	-	-	-	-
T	13.1%	12.6%	12.3%	11.9%	-	-	-	-	-
N	3.5%	2.0%	2.9%	1.8%	-	-	-	-	-
Deletions	58.1%	60.7%	60.2%	61.3%	-	-	-	-	-
A	16.7%	18.5%	18.0%	18.2%	-	-	-	-	-
C	9.3%	10.1%	10.0%	10.4%	-	-	-	-	-
G	9.5%	10.3%	10.3%	10.6%	-	-	-	-	-
T	16.4%	18.4%	18.0%	18.1%	-	-	-	-	-
N	6.3%	3.3%	3.9%	4.0%	-	-	-	-	-

Phylogenetic tree among 9 date palm cultivars based on SNP site



Allele frequency $\geq 80\%$, coverage ≥ 4 , totally ~ 200,000 SNPs

泛基因组时代已经来临

- 测序高质量的基因组——获得更准确的结果
- 大量同一物种的群体基因组测序的意义是什么？
 - 泛基因组：某个生物分支（物种）的全部基因组总和
 - 目的：整体研究物种中共有和变化的基因组序列

Published: 15 February 2001

Initial sequencing and analysis of the human genome

International Human Genome Sequencing Consortium

Nature 409, 860–921 (2001) | [Cite this article](#)

313k Accesses | 15911 Citations | 1538 Altmetric | [Metrics](#)



x 1



Article | [Open Access](#) | Published: 10 May 2023

A draft human pangenome reference

Wen-Wei Li, Mohin Azei, Jana Ebler, Daniel Doern, Marina Haukness, Glenn Hickey, Shuangjie Lu, Julian S. Lucas, Jean Montlong, Haley J. Abel, Silvia Buonomo, Xian H. Chiang, Haoyu Cheng, Justin Chu, Vincenzo Colonna, Jordan M. Eizenga, Xiaowen Feng, Christian Fischer, Robert S. Fulton, Shilpa Garg, Cristian Groza, Andrea Guarnacino, William T. Hickey, Simon Houmans, — Benedict Paten

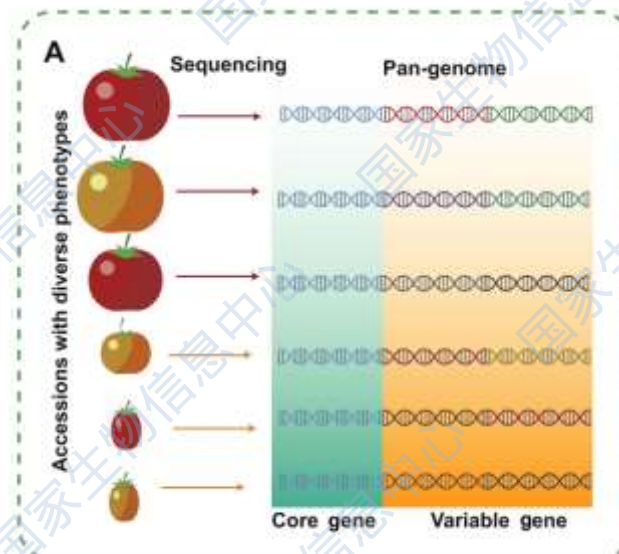
Nature 617, 312–324 (2023) | [Cite this article](#)

118k Accesses | 7 Citations | 3151 Altmetric | [Metrics](#)



x 47

单个基因组不足以
代表整个物种

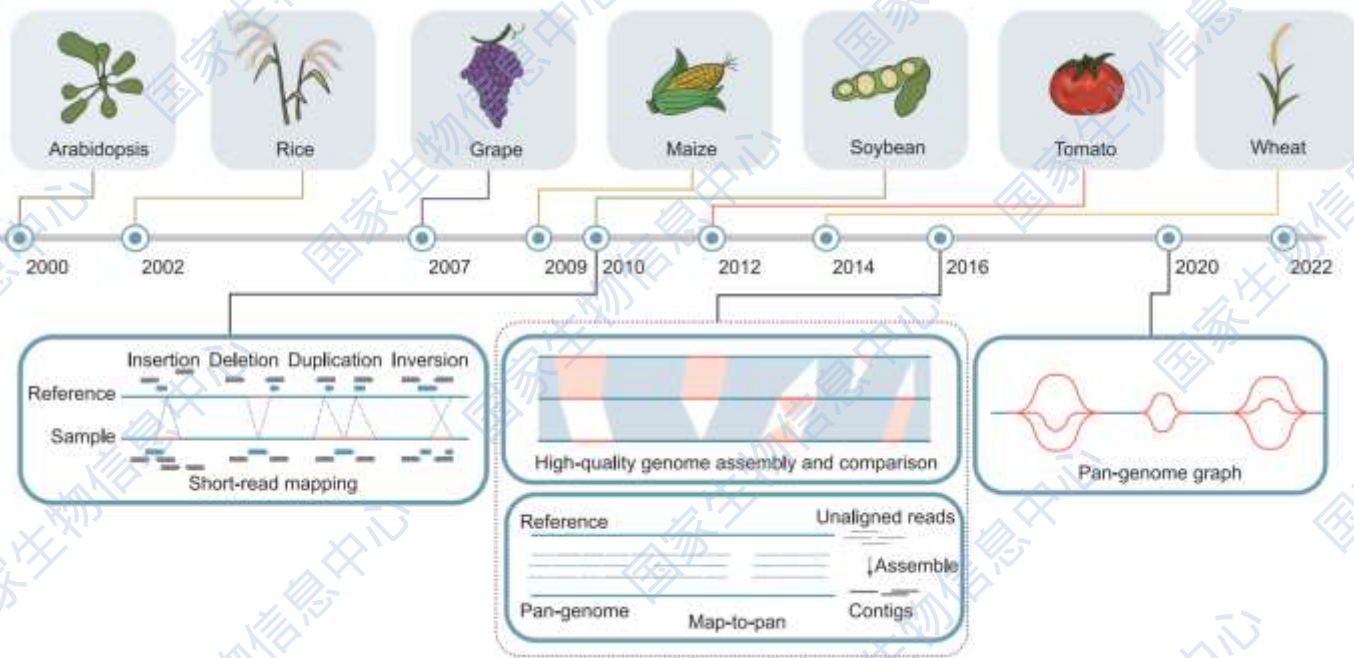


Li et al. 2022. *Journal of Genetics and Genomics*

Plant pan-genomics and its applications

Junpeng Shi^{1,2}, Zhidi Tian², Jinsheng Lai² and Xuehui Huang^{1,2}

测序技术进步催生泛基因组基因型鉴定



群体重测序
(单基因组)

比较泛基因组
(单基因组+附加序列)

图 (Graph) 泛基因组
(多基因组)

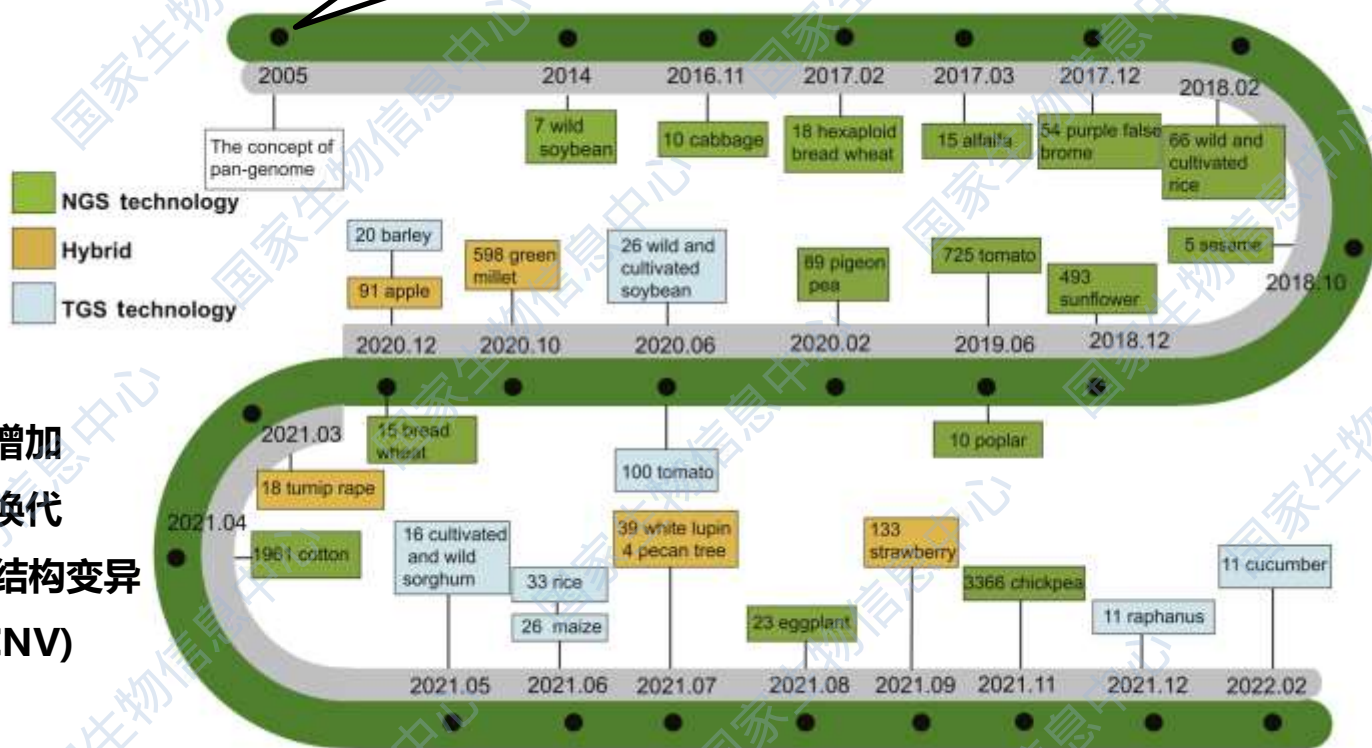
Review

Plant pan-genomics: recent advances, new challenges, and roads ahead

Wei Li¹, Jianan Liu¹, Hongyu Zhang, Ze Liu, Yu Wang, Longsheng Xing, Qiang He, Hailong Du^{*}

泛基因组概念在微生物中提出

作物泛基因组研究的发展



- 群体数量不断增加
- 测序技术更新换代
- 从SNP到复杂结构变异 (PAV、SV、CNV)

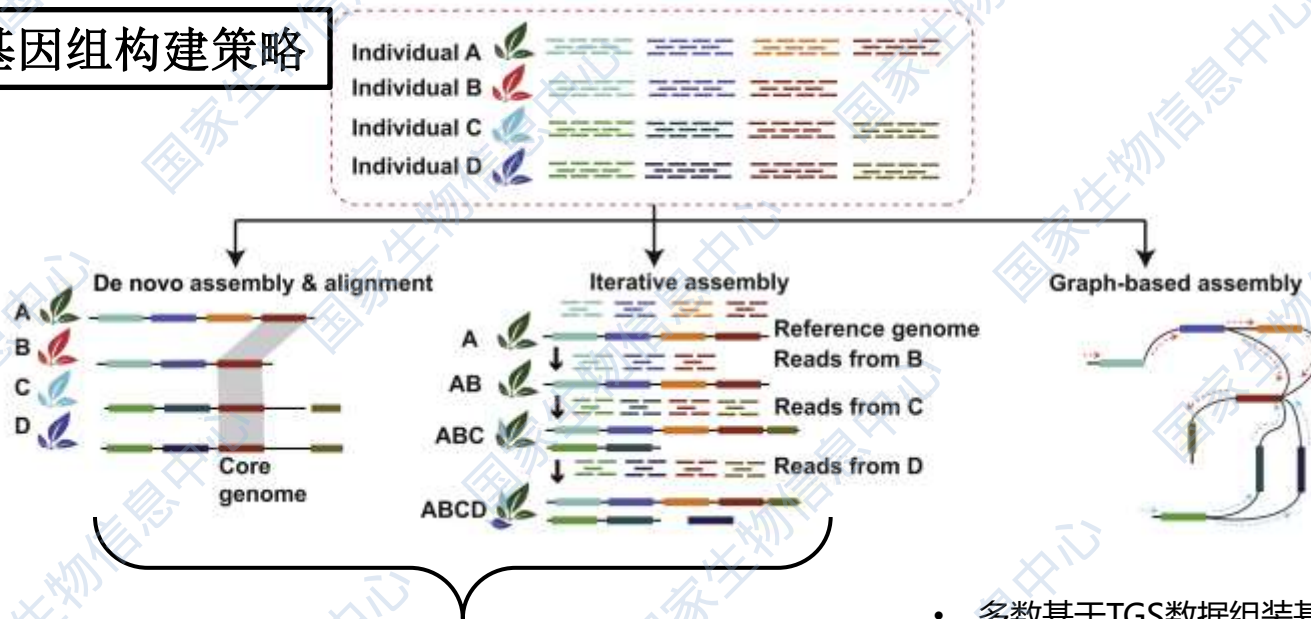
Li et al. 2022. *Journal of Genetics and Genomics*

Review

Plant pan-genomics: recent advances, new challenges, and roads ahead

Wei Li¹, Jianan Liu¹, Hongyu Zhang, Ze Liu, Yu Wang, Longsheng Xing, Gang He, Huilong Du^{*}

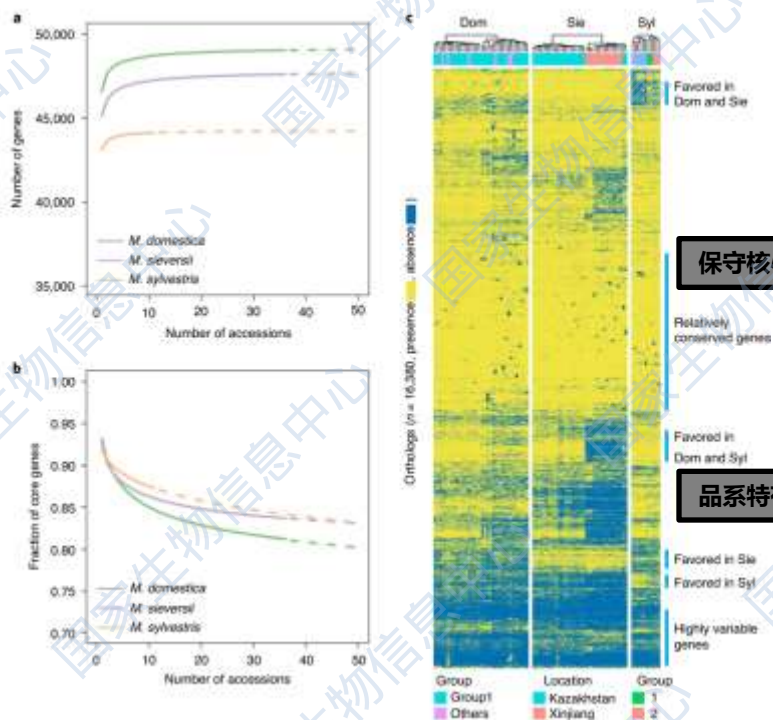
泛基因组构建策略



- 多数基于NGS数据组装基因组构建
- 基于个体基因组的注释基因构建
- 核心 (core) 与非核心 (variable) 基因
- 泛基因的存在缺失变异 (PAV) 分析
- 多数基于TGS数据组装基因组构建
- 整合多个基因组的变异信息
- SNP和结构变异 (SV) 基因型鉴定
- 允许使用短reads鉴定复杂SV
- SV与表型关联 (Pan-GWAS) 分析

泛基因组基因存在缺失变异 (PAV) 分析

1. 寻找区分不同品系的marker genes
2. 寻找核心与可变基因的功能富集



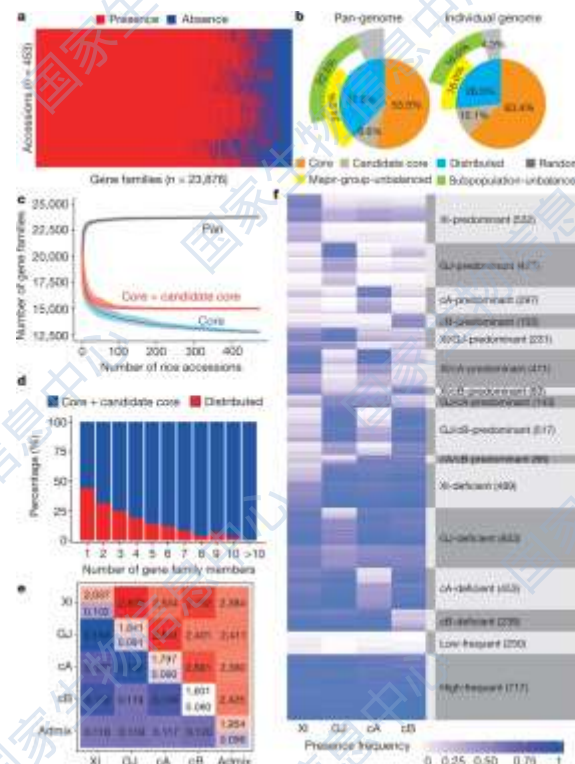
保守核心基因

品系特有基因

苹果 (*Malus*) 泛基因组
(91个体, 16380泛基因, NGS)

Sun *et al.* 2020. *Nature Genetics*

Wang *et al.* 2018. *Nature*



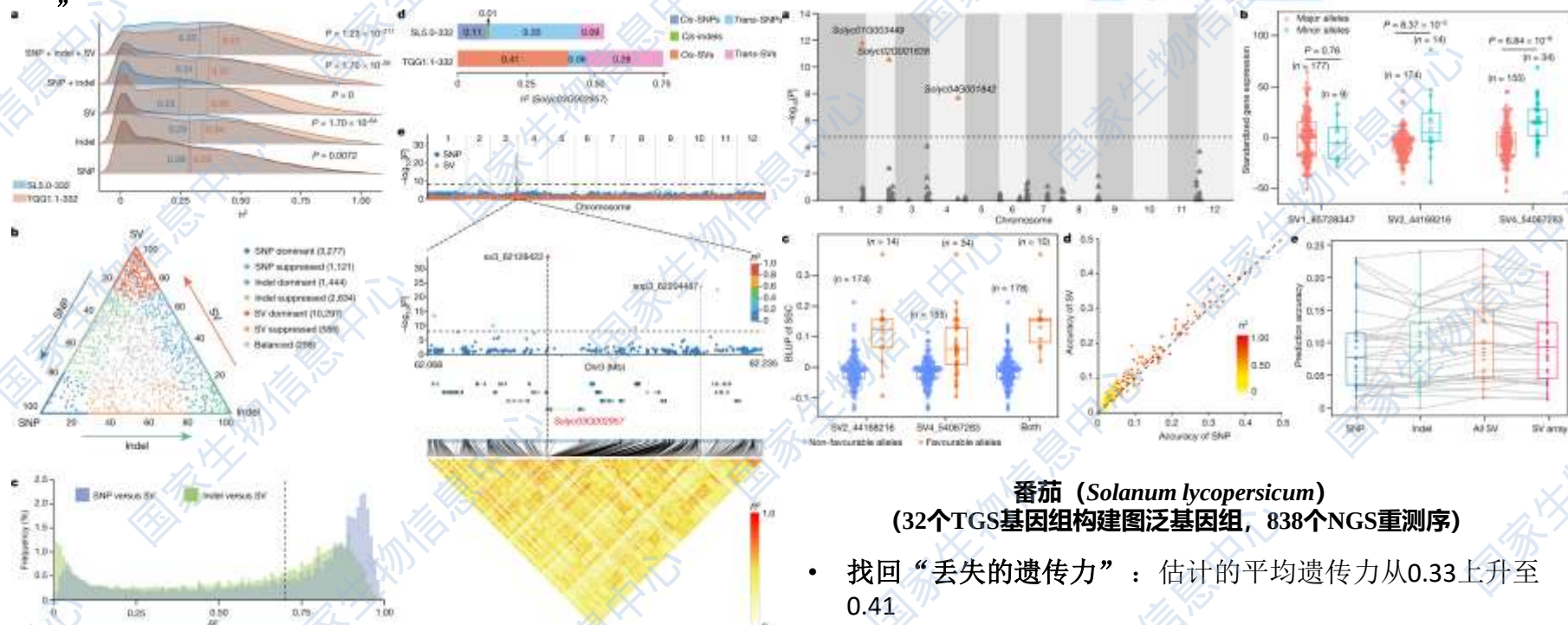
品系特有基因

保守核心基因

亚洲栽培稻 (*Oryza sativa*) 泛基因组
(453个体, 23876泛基因, NGS)

基于泛基因组的基因型鉴定和表型关联 (Pan-GWAS) 分析

Nature评述: “这项工作是对图泛基因组概念最全面的分析。图泛基因组将可能成为基因组分析和作物基因组育种的标准, 在这个意义上, 这篇论文是奠基性的 (foundational)”



番茄 (*Solanum lycopersicum*) (32个TGS基因组构建图泛基因组, 838个NGS重测序)

- 找回“丢失的遗传力”: 估计的平均遗传力从0.33上升至0.41
- 鉴定出2个与可溶性固形物 (SSC) 含量高度相关的SVs

“遗传力丢失” (Missing heritability) 是重要的数量遗传学问题, 即通过遗传标记估计的遗传力以及通过全基因组关联分析 (GWAS) 发现的所有相关基因所贡献的遗传力总和均低于实际的遗传力。

Zhou et al. 2022. Nature

Article | Open Access | Published: 08 June 2022

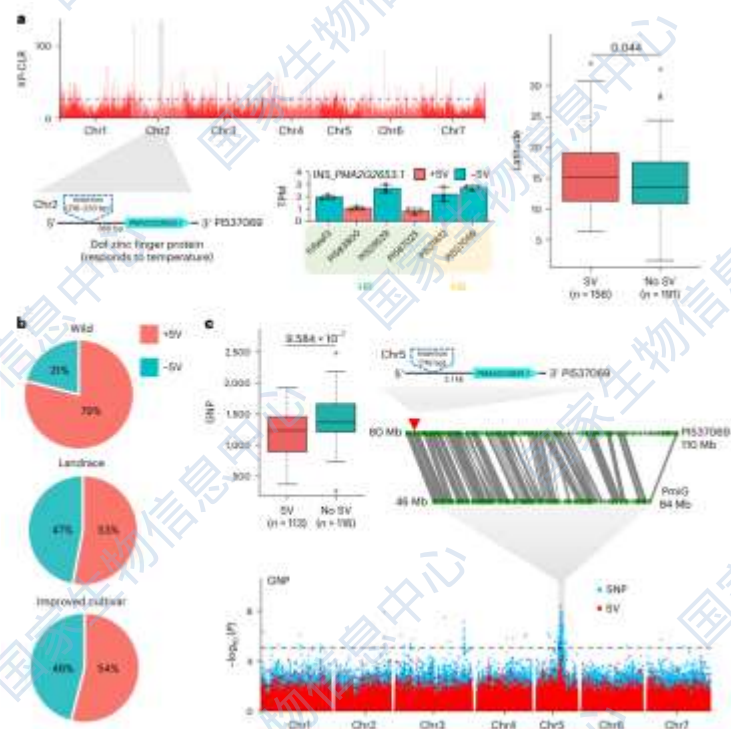
Graph pangenome captures missing heritability and empowers tomato breeding

Yao Zhou, Zhiyang Zhang, Zhiqiang Bao, Hongbo Li, Yang Liu, Yanjun Zou, Yeyan Wu, Lin Cheng, Yuhui Fang, Kun Wu, Jiehe Zhang, Hongyan Yao, Tao Lin, Qiang Gao, Surya Saha, Lukas Mueller, Zhongjun Fan, Thomas Städler, Shitong Xu, Zhiyu Zhang, Doug Speed & Sarwan Huang

Nature 606, 527–534 (2022) | Cite this article

基于SV的泛基因组表型关联 (Pan-GWAS) 分析

受益于SV鉴定，泛基因组可大幅提升基因型的检出效率



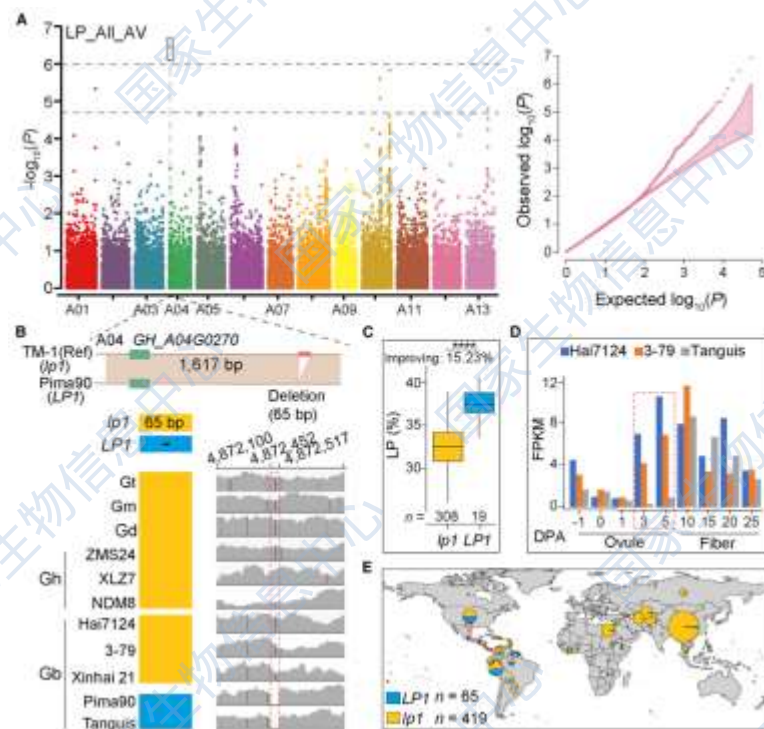
珍珠粟 (*Pennisetum glaucum*)

(10个TGS基因组构建图泛基因组, 1036个NGS重测序)

从424 085个结构变异 (SV) 中鉴定出与耐热相关的SV

Yan *et al.* 2023. *Nature Genetics*

Jin *et al.* 2023. *Molecular Plant*

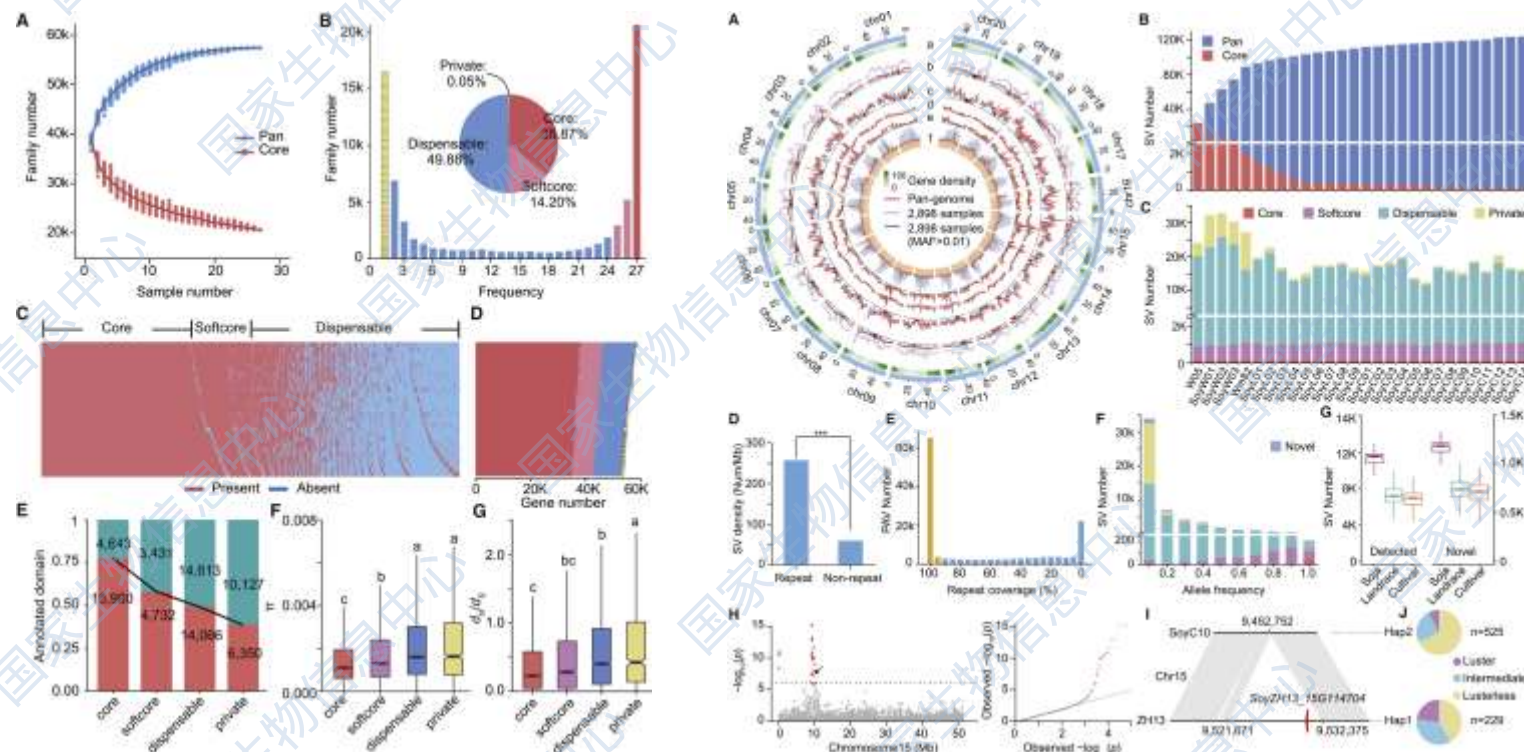


异源四倍体棉花 (*Gossypium barbadense*)

(12个TGS基因组构建图泛基因组, 333个NGS重测序)

从182 593个SV中鉴定出与纤维质量 (衣分率) 相关的SV
lint percentage

综合基于SV和PAV的泛基因组表型关联 (Pan-GWAS) 分析



大豆 (*Glycine max*) 泛基因组

(29个TGS基因组构建泛基因组, 2898个体重测序)

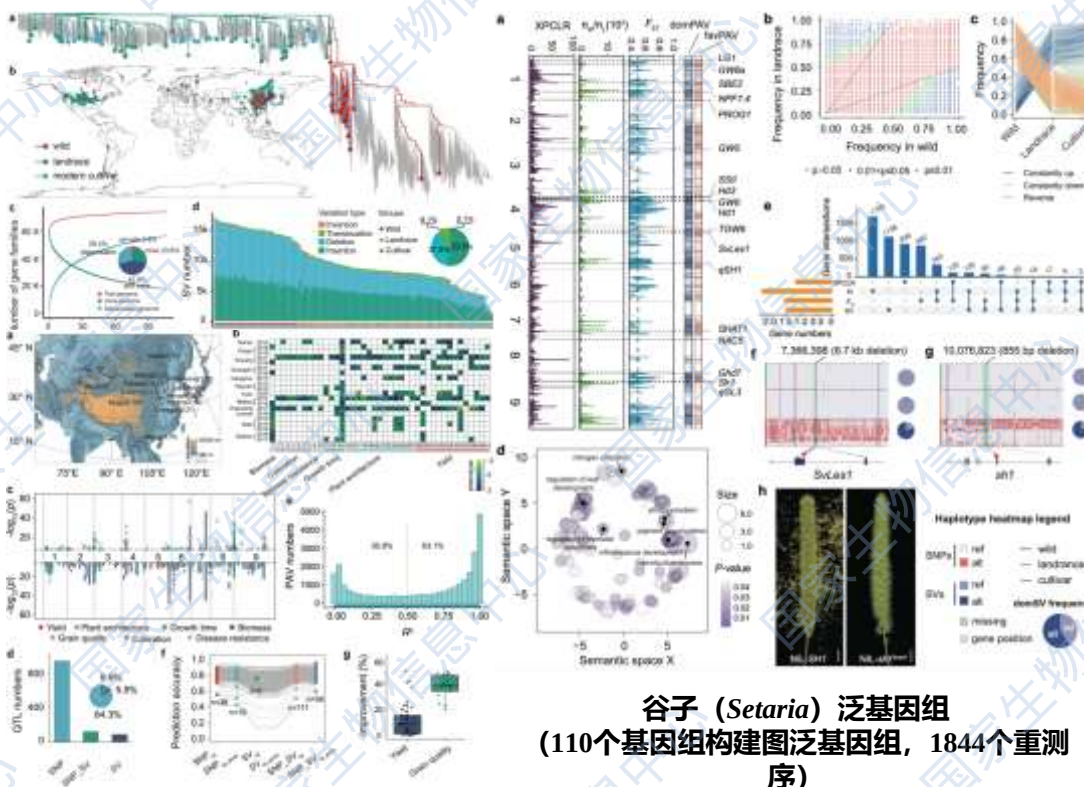
从723 862个PAV中鉴定出与种子光泽(seed luster)相关的基因PAV

(SoyZH13_15G114704, an HPS encoding gene)

Liu *et al.* 2020. *Cell*

综合基于SV和PAV的泛基因组表型关联（Pan-GWAS）分析

“开启大规模种质资源发掘利用新篇章”



- 应用泛基因组发现在栽培种中已被完全驯化的基因（丢失多态性问题）
- 引入结构变异（SV）可以使167组表型（73.9%）的预测精度提升0.04%~12.6%

He et al. 2023-06-23. *Nature Genetics*

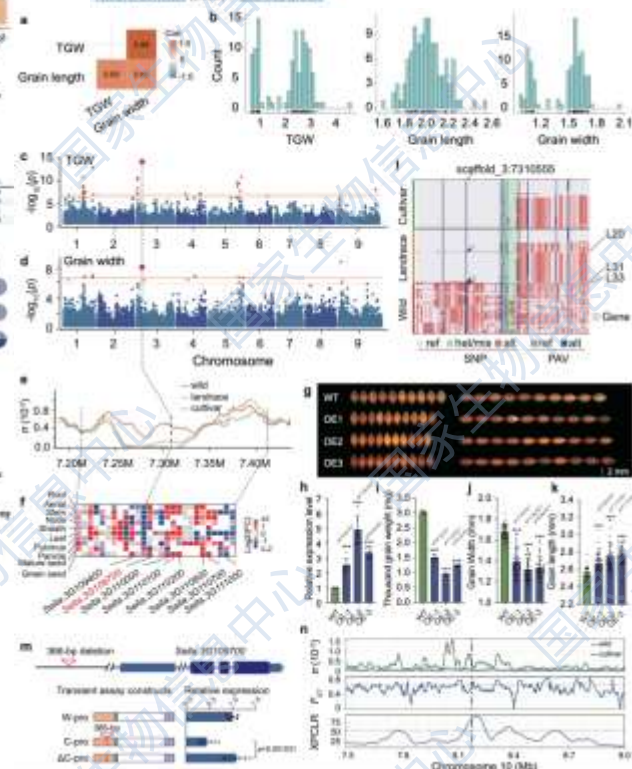
Article | Open Access | Published: 08 June 2023

A graph-based genome and pan-genome variation of the model plant *Setaria*

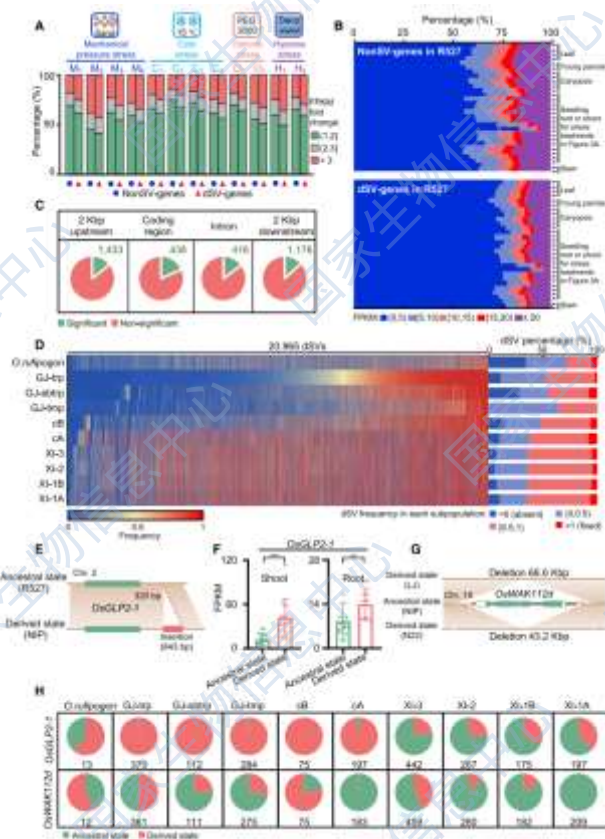
Qiang He, Sha Tang, Hui Zhi, Anfeng Chen, Jia Zhang, Honghui Wang, Qianqian Alam, Hongbo Li, Hui Zhang, Lili Xing, Xuhai Li, Wei Zhang, Hailong Wang, Junceng Shi, Huikang Du, Hongbo Wu, Liwei Wang, Bing Yang, Lu Xing, Hongshan Yan, Zhongqiang Song, Jinxiong Liu, Hailong Wang, Xiang Tian, Xianmin Zhao

Show authors

Nature Genetics (2023) | Cite this article



基于CNV的泛基因组表型关联分析

Qin *et al.* 2021. *Cell*

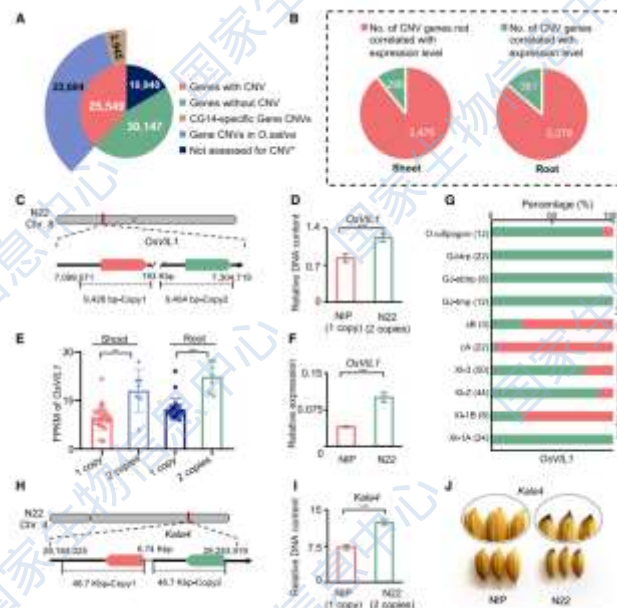
Cell

CellPress

Abstract

Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations

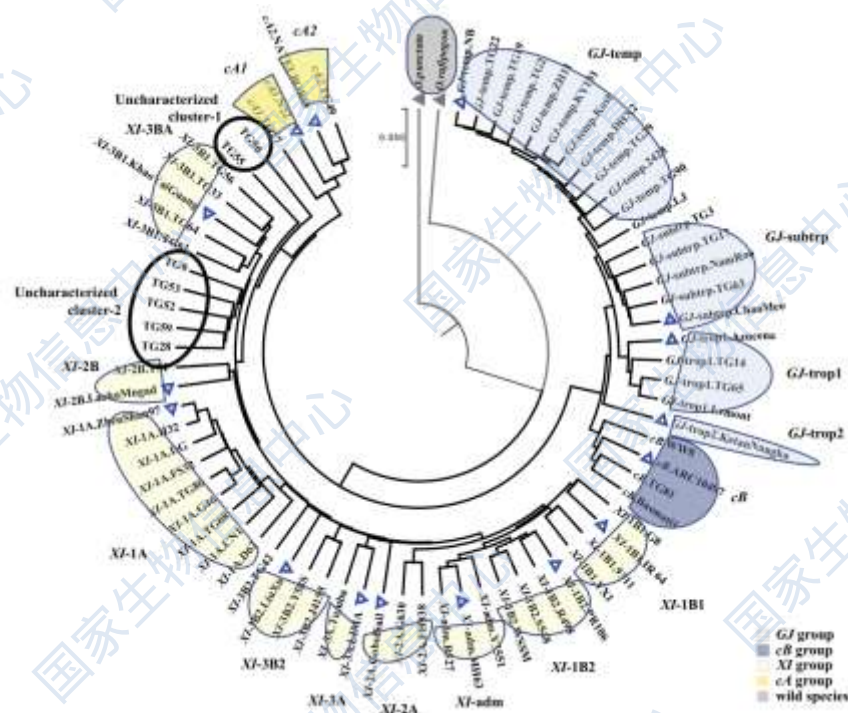
Yong Chen,^{1,2} Hongwei Lu,^{1,2,3} Hailong Du,^{1,2,3,4} Xue Wang,^{1,2} Wotian Chen,^{1,2} Zhao Chen,^{1,2,3} Deng He,¹ Shupin Gu,¹ Hongye Zhang,¹ Ruijie Li,¹ Shuo Li,^{1,2} Yanli Yu,^{1,2} Gang Gao,¹ Min Tu,¹ Hua Yuan,¹ Jinghui Ma,¹ Tingting Wang,¹ Tingting Guo,¹ Enjie Fan,¹ Wenbo Li,¹ Jing Wang,¹ Min He,¹ Junyi Fu,¹ Ting Li,¹ Ning Jiang,¹ Ruiwen Chen,¹ Zhongqi Jiang,¹ and Shugu Li^{1,2}



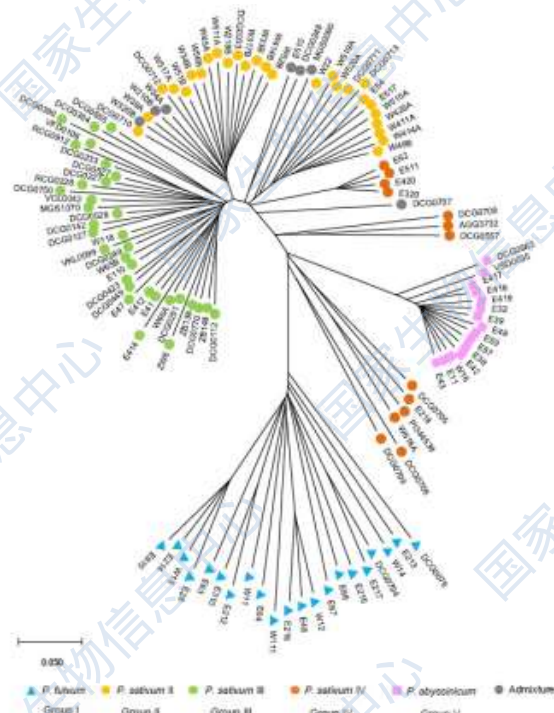
- **水稻SV+CNV与表型的关联分析**
- **基于33个高质量水稻基因组**
- **尚无大规模CNV-GWAS相关成果（暂时受限
于方法）**

泛基因组分析遗传和进化关系

泛基因组可揭示基因和SNP “无法看到” 的遗传进化关系



基于79个体水稻泛基因组
倒置变异 (Inversion) 构建的进化关系



基于116个体豌豆泛基因组
存在缺失变异 (PAV) 构建的进化关系

Zhou *et al.* 2023. *Nature Communications*
Yang *et al.* 2022. *Nature Genetics*

其他一些植物泛基因组研究举例

受限于技术和成本，泛基因组研究仍在“循序渐进”，但已“渐入佳境”

Species	Sequencing platform	Construction methods	Population size	Pan-genome representations ^a	References
<i>Arabidopsis thaliana</i>	PacBio	Genome assembly comparison	8	Pan-genome of ~135 Mb and ~30 000 pan-genes	Jiao and Schneeberger (2020)
Rice	Illumina	Map-to-pan	3010	Pan-genome of ~642 Mb	Wang et al. (2018)
	Illumina	Genome assembly comparison	66	42 580 pan-genes	Zhao et al. (2018)
	PacBio	Genome assembly comparison	16	~33.7 Mb novel sequences in each line	Zhou et al. (2020b)
	PacBio	Assembly-based graph genome	31	66 636 pan-genes	Qin et al. (2021)
	Nanopore	Genome assembly comparison	111	Pan-genome of ~1250 Mb and 75 305 pan-genes	Zhang et al. (2022a)
	Nanopore	Assembly-based graph genome	251	Pan-genome of ~1.52 Gb and 51 359 pan-genes	Shang et al. (2022)
Maize	Illumina	Map-to-pan (transcriptome)	503	8681 novel transcripts	Hirsch et al. (2014)
	Illumina	Genome assembly comparison	6	34 352 core genes	Haberer et al. (2020)
	PacBio	Genome assembly comparison	26	103 000 pan-genes	Hufford et al. (2021)
	Illumina	Assembly-based graph genome	721	Pan-genome of ~6.71 Gb and 58 944 pan-genes	Gui et al. (2022)
Wheat	Illumina	Map-to-pan	18	140 500 pan-genes	Montenegro et al. (2017)
	Illumina	Genome assembly comparison	15	~13 000 PAV genes	Walkowiak et al. (2020)
Soybean	Illumina	Genome assembly comparison	7	2.3 to 3.9 Mb PAV sequences and 338 PAV genes	Li et al. (2014)
	PacBio	Assembly-based graph genome	29	~167 Mb specific sequences in each accession	Liu et al. (2020c)
	Illumina	Genome assembly comparison	204	Pan-genome of ~1086 Mb and 54 531 pan-genes	Torkamaneh et al. (2021)
	PacBio	Genome assembly comparison	26	129 006 pan-genes	Zhuang et al. (2022)
Barley	Illumina	Genome assembly comparison	20	638.6 Mb PAV sequences	Jayakodi et al. (2020)
Sorghum	PacBio	Assembly-based graph genome	16	Pan-genome of 984.8 Mb and 44 079 pan-genes	Tao et al. (2021)
	Illumina	Map-to-pan	176	Pan-genome of 883.3 Mb and 35 719 pan-genes	Ruperao et al. (2021)
Cotton	Illumina	Map-to-pan	1807	Pan-genome of 3388 Mb and 102 768 genes (<i>G. hirsutum</i>)	Li et al. (2021a)
				Pan-genome of 2575 Mb and 80 148 genes (<i>G. barbadense</i>)	
Tomato	Illumina	Map-to-pan	725	Pan-genome of 1179 Mb and 40 396 pan-genes	Gao et al. (2019)
	Nanopore	Long-read mapping	100	238 490 SVs	Alonge et al. (2020)
Cucumber	PacBio	Assembly-based graph genome	12	26 822 pan-genes	Li et al. (2022a)
Rapeseed	PacBio	Genome assembly comparison	8	Pan-genome of ~1.8 Gb and 152 185 pan-genes	Song et al. (2020)
Mung bean	Illumina	Genome assembly comparison	217	Pan-genome of ~762.9 Mb and 43 462 pan-genes	Liu et al. (2022a)

Shi *et al.* 2022. *Molecular Plant*

高质量中国豌豆基因组和泛基因组研究



- **豌豆**(*Pisum sativum* L.)是北方一年生和南方越年生冷季食用豆类作物，具有丰富的营养价值和显著的生态优势，在世界范围内广泛种植。**粮菜饲绿肥兼用**。
- 自孟德尔发现遗传定律以来，豌豆作为遗传研究的模式植物，受到了学界和公众的广泛关注。
- 然而，豌豆基因组的研究远远滞后于其他作物.....

孟德尔诞辰200周年



Diamond 2002; Doležel *et al.* 2010; Smýkal 2012; Macwilliam *et al* 2014; FAO 2019; 宗绪晓等 2009; 刘旭等 2018

PeaCaméor

Illumina + 遗传图谱

PeaZW6

PacBio + Illumina
+ BioNano
+ 10X Genomics
+ Hi-C
+ 遗传图谱

PeaZW6和PeaCaméor基因组质量比较

单粒传3代的中国豌豆主栽品种中豌6号(ZW6)



基因组大小4.2Gb, >80% 重复序列 (LTR-RTs)

Table 1 | Summary of pea genome assembly

Genome feature	PeaZW6		PeaCaméor	
	Number	Size (Mb)	Number	Size (Mb)
Assembled scaffolds	1,579	3,796.7	24,623	3,920.1
Superscaffolds	7	3,719.6	7	3,234.4
Remaining scaffolds	1,572	77.1	14,266	685.4
N50 remaining scaffolds	289	0.074	1,411	0.13
Contigs	2,402	3,786.4	218,010	3,159.3
N50 contigs	118	8.98	32,663	0.037
Remaining contigs	1,586	76.6	69,733	572.1
N50 remaining contigs	298	0.073	6,466	0.023
Protein-coding gene models	47,526	121.84	44,756	124.6
Genes in pseudomolecules	46,607	119.68	38,312	102.3

Assembly	k-mers uniquely found only in the assembly	k-mers found in both assembly and the read set	QV	Error Rate (%)	solid k-mers in the assembly	Total solid k-mers in the read set	Completeness (%)
PeaZW6	2,832,546	3,786,410,593	44.48	0.0036	999,035,347	1,025,354,019	97.43
PeaCaméor	234,931,999	3,153,805,448	24.34	0.3680	971,124,901	1,023,752,156	94.86

Chr	Length (bp)	PeaZW6		LAI	Length (bp)	PeaCaméor		LAI
		Intact LTR-RTs content (%)	Total LTR-RTs content (%)			Intact LTR-RTs content (%)	Total LTR-RTs content (%)	
Chr1	463,644,500	10.79	77.56	13.91	372,165,146	1.30	57.49	2.26
Chr2	492,729,000	10.31	81.69	12.62	427,597,918	1.25	59.62	2.09
Chr3	535,674,373	10.43	80.86	12.92	437,560,801	1.18	58.13	2.03
Chr4	501,613,000	10.71	79.11	13.54	446,350,677	1.16	58.20	1.99
Chr5	652,929,960	10.76	79.70	13.50	379,269,071	1.30	59.13	2.19
Chr6	523,395,090	10.46	78.81	13.28	480,417,049	1.33	57.56	2.31
Chr7	549,106,000	10.59	79.03	13.40	491,380,962	0.98	55.87	1.76
Total	3,719,091,923	10.58	79.54	13.31	3,234,741,624	1.21	58.09	2.09

Searching model	Protein categories	BUSCO			
		PeaZW6		PeaCaméor	
		Number	Percentage	Number	Percentage
Genome	Complete BUSCOs (C)	1604	99.38	1562	96.78
	Complete and single-copy BUSCOs (S)	1537	95.23	1518	94.05
	Complete and duplicated BUSCOs (D)	67	4.15	44	2.73
	Fragmented BUSCOs (F)	6	0.37	40	2.48
	Missing BUSCOs (M)	4	0.25	12	0.74
Total BUSCO groups searched		1614	100.00	1614	100.00
Protein	Complete BUSCOs (C)	1578	97.77	1517	93.90
	Complete and single-copy BUSCOs (S)	1516	93.93	1481	91.76
	Complete and duplicated BUSCOs (D)	62	3.84	36	2.23
	Fragmented BUSCOs (F)	15	0.93	67	4.15
	Missing BUSCOs (M)	21	1.30	30	1.86
Total BUSCO groups searched		1614	100.00	1614	100.00

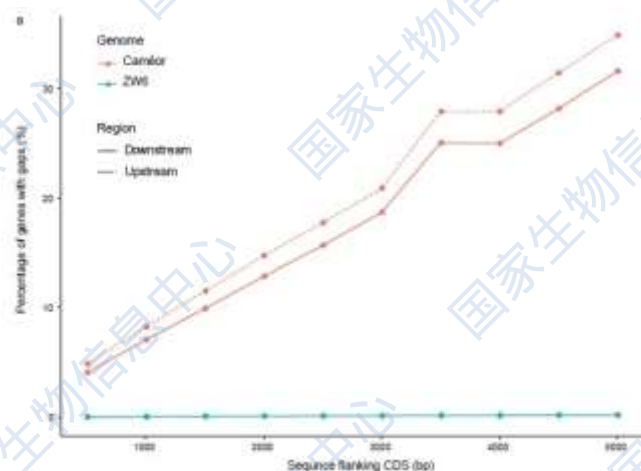
Kreplak *et al.* 2019. *Nature Genetics*

Yang *et al.* 2022. *Nature Genetics*

- 增加了~600Mb有效序列
- N50从37kb增至8.98Mb

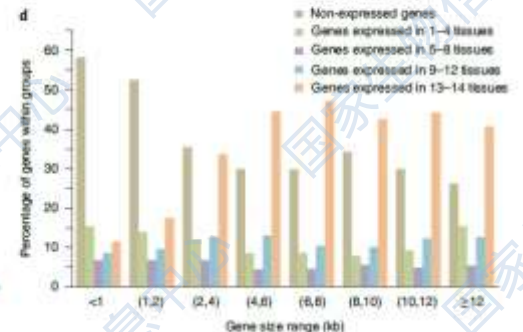
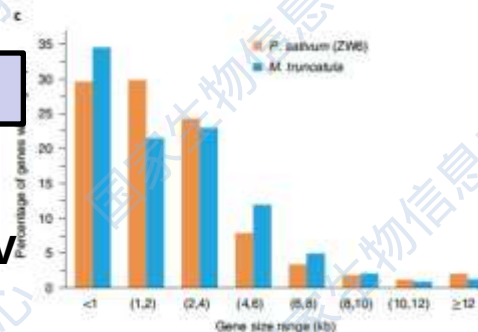
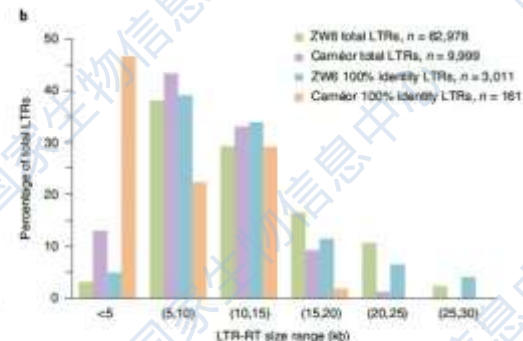
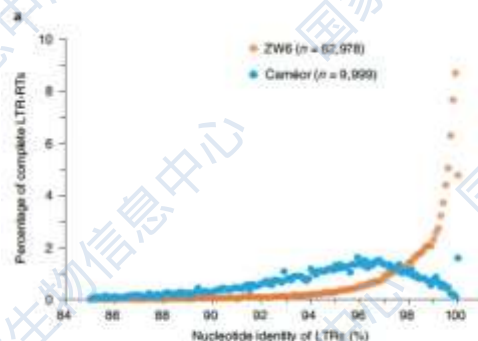
PeaZW6和PeaCaméor基因组质量比较

相对更多、更完整的LTR，特别是近期产生的



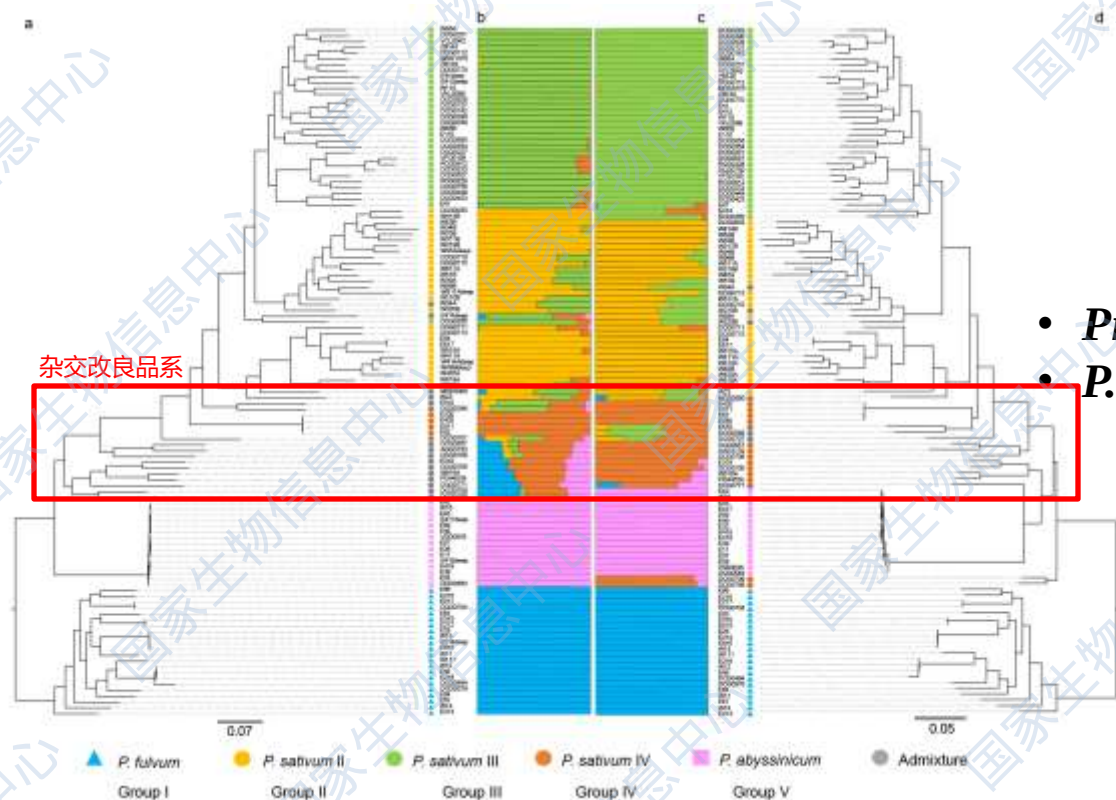
在ZW6中几乎不存在基因内部的Gap

- NGS组装的Gap多与LTR相关
- 长基因中可能包含更多LTR，累积更多SV
- 高质量的基因组对未来发现SV相关的基因型-表型关联是必不可少的

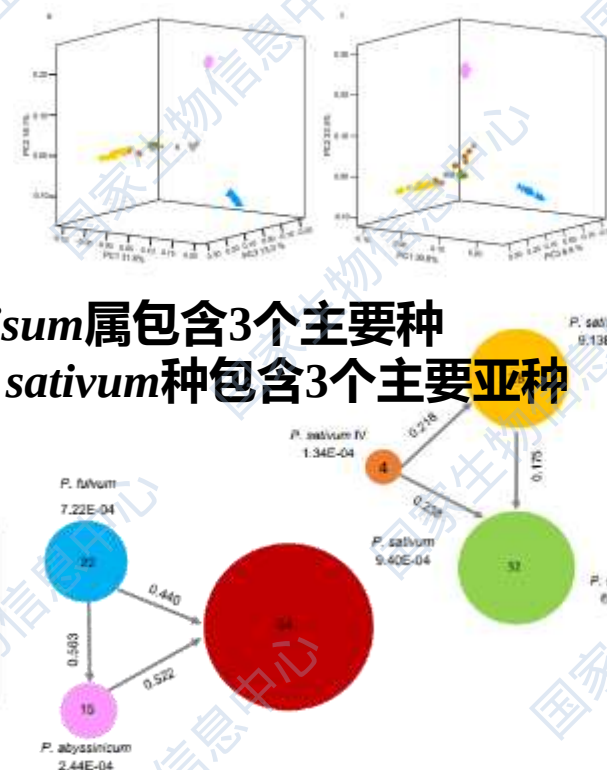


短基因（较少的intron）占比更高，但长基因更house-keeping

群体重测序 SNP (左) / SV (右) 重现豌豆驯化历史和遗传多样性

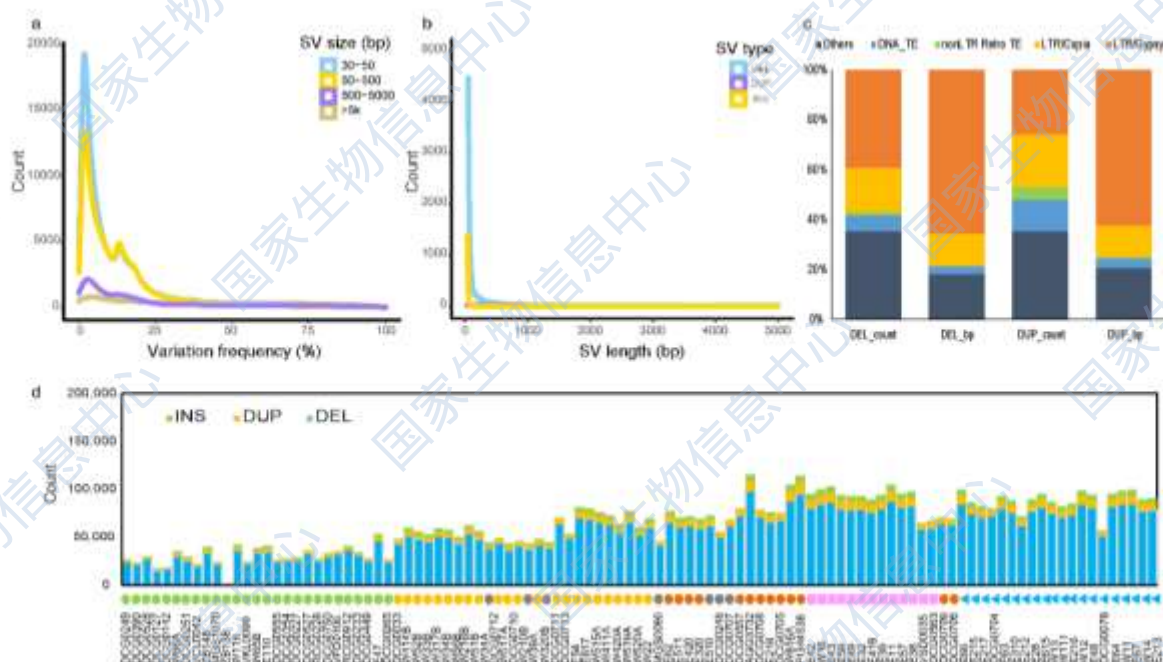


- *Pisum* 属包含3个主要种
- *P. sativum* 种包含3个主要亚种



Yang *et al.* 2022. *Nature Genetics*

群体重测序鉴定结构变异 (SV)

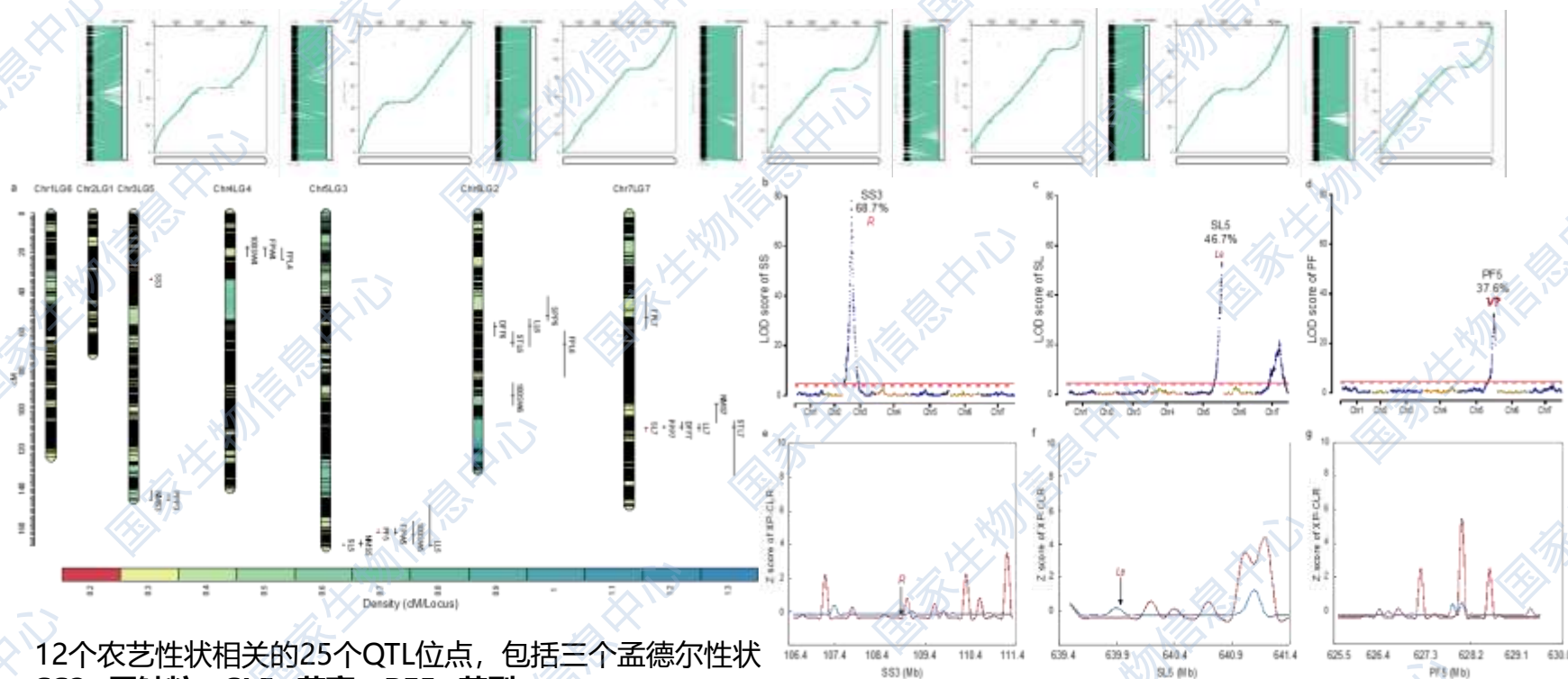


- 以PeaZW6为参考基因组，相比*P. sativum*，*P. fulvum*和*P. abyssinicum*具有更多的SV
- 提示应当进一步组装测序高质量的*P. fulvum*和*P. abyssinicum*基因组（已在进行中）

Yang *et al.* 2022. *Nature Genetics*

遗传图谱+QTL定位“重新发现了”孟德尔基因 ——圆皱粒、茎高和荚型

SS3、SL5、PF5位点内基因的同源比对和功能注释发现了两个先前已知的控制粒型和株高的孟德尔基因 R 和 Le ，以及一个与荚型相关的候选基因。



12个农艺性状相关的25个QTL位点，包括三个孟德尔性状
SS3=圆皱粒；**SL5=茎高**；**PF5=荚型**

Yang *et al.* 2022. *Nature Genetics*

基于116个豌豆基因组基因的泛基因组

Supplementary data 16 Summary of pea pan-genomes based on species and genetic groups

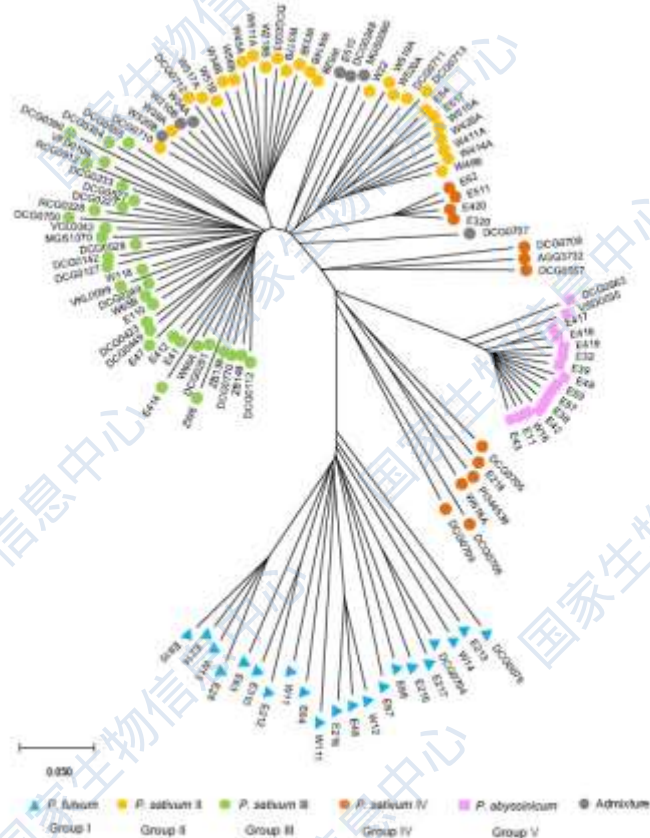
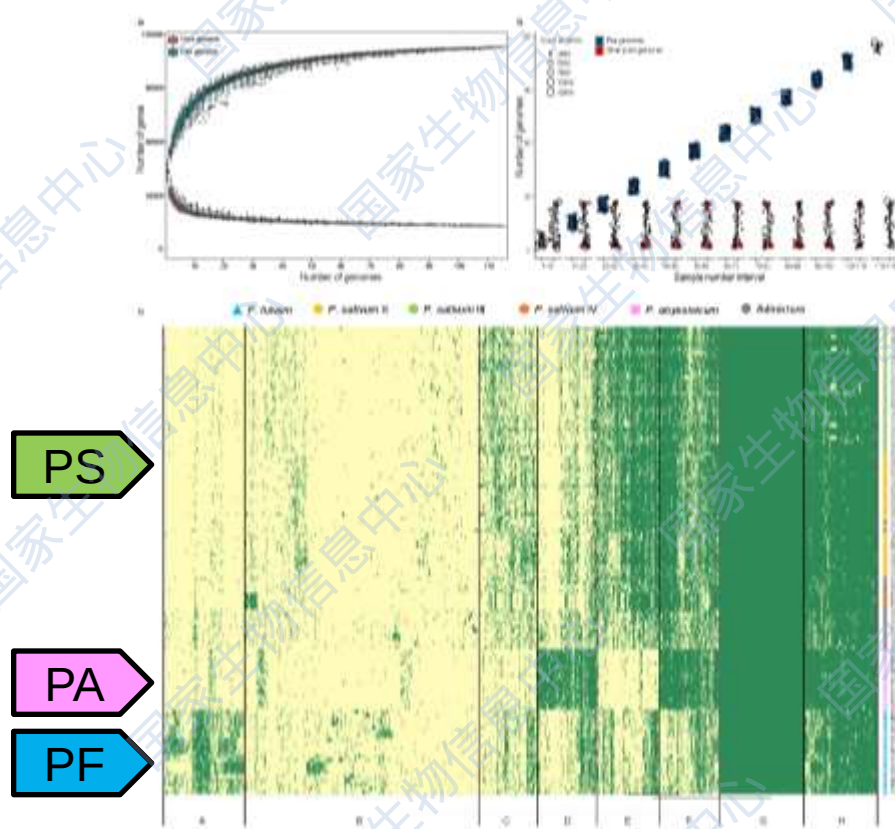
Genomes/genetic groups	Pan-genome		Core-genes (>=90% genomes)			Soft-core-genes (>=90% and <99% genomes)			Shell-genes (>=15% and <90% genomes)			Cloud-genes (<15% genomes)		
	HOGs	Genes	HOGs	Genes	Percent of pan-genes	HOGs	Genes	Percent of pan-genes	HOGs	Genes	Percent of pan-genes	HOGs	Genes	Percent of pan-genes
All genomes (n=116) ^a	112,776	6,693,819	15,470	2,355,605	35.19%	6,170	1,040,106	15.54%	41,028	2,963,843	44.28%	30,108	334,265	4.99%
<i>P. sativum</i> (n=21)	87,218	1,009,847	18,467	444,627	44.03%	4,913	111,262	11.02%	39,162	407,148	40.12%	24,676	46,810	4.64%
<i>P. abyssinicum</i> (n=15)	56,615	819,534	30,944	618,915	75.52%	4,461	74,053	9.04%	12,786	114,802	14.01%	8,424	11,764	1.44%
<i>P. sativum</i> II (n=28)	74,829	1,740,968	24,611	1,038,940	59.68%	5,038	194,157	11.15%	27,079	469,844	26.99%	18,101	38,027	2.18%
<i>P. sativum</i> III (n=32)	76,727	1,976,752	19,942	982,407	49.70%	9,492	420,592	21.28%	27,630	533,270	26.98%	19,663	40,483	2.05%
<i>P. sativum</i> IV (n=4)	48,623	230,607	38,403	209,371	90.79%	-	-	0.00%	10,220	21,236	9.21%	-	-	0.00%
<i>P. sativum</i> (n=64)	85,840	3,948,327	17,797	1,718,146	43.52%	10,422	965,289	24.45%	30,249	1,166,847	29.55%	27,372	98,045	2.48%
Admixture (n=16)	80,928	854,920	18,846	399,406	46.72%	3,581	74,955	8.77%	37,226	349,504	40.88%	21,275	31,055	3.63%

HOG = phylogenetic Hierarchical OrthoGroup
^a Three accessions DCG0494, DCG 0771 and DCG0265 with low percent (<90%) of complete BUSCOs (Supplementary data 9) were excluded for pan-genome analysis.

- 经过质量控制后，将来自PeaZW6和115个豌豆属基因组的基因进行基因泛基因组分析
- 共聚类成 112,776个泛基因，以Orthofinder中的直系同源簇（HOG）来表示。

Yang et al. 2022. Nature Genetics

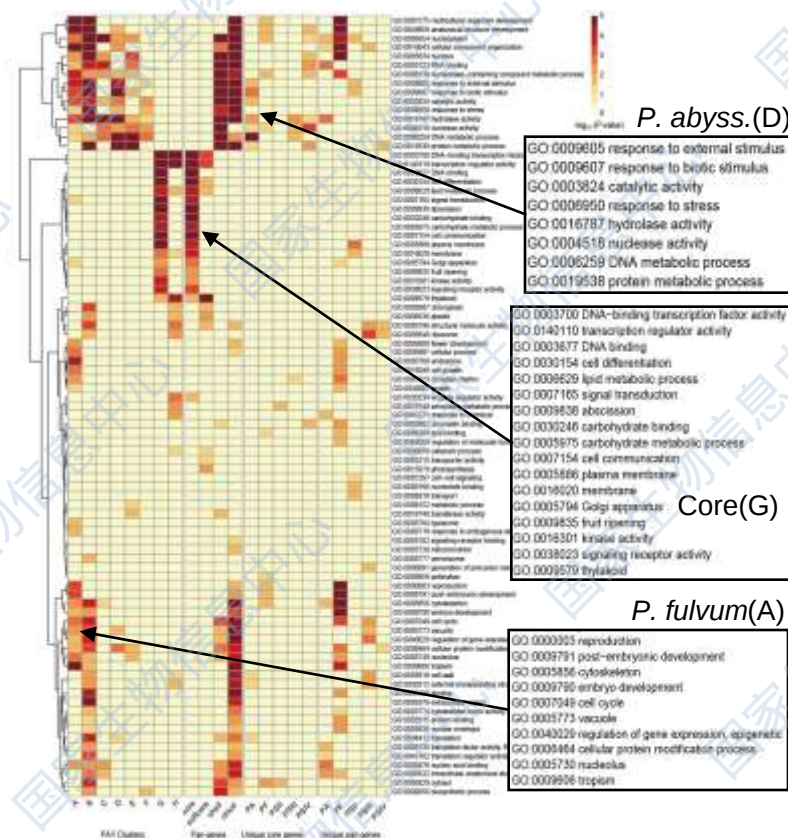
基于116个豌豆基因组基因的泛基因组



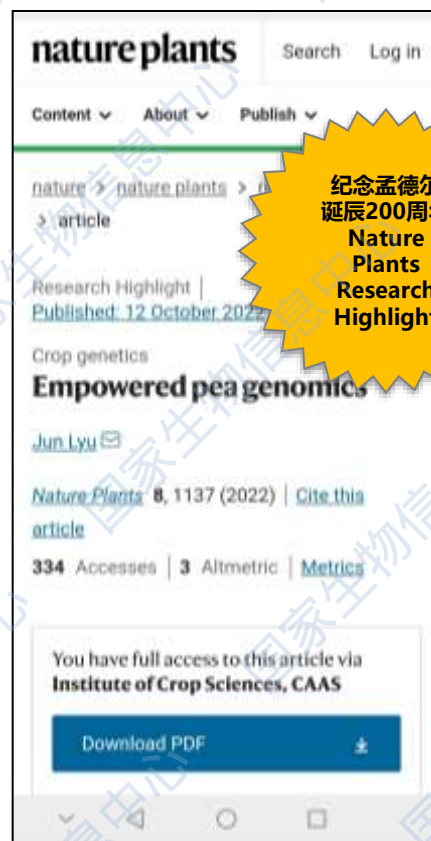
基于116个豌豆基因组构建的泛基因组
基因存在缺失变异 (PAV) (左) 及基于PAV构建的进化关系(右)

基于116个豌豆基因组基因的泛基因组

- GO 分析显示出保守基因 (core和soft-core基因) 和可变基因 (shell和cloud基因) 之间的不同功能富集。
- 保守基因富含碳水化合物和脂质代谢过程等功能
- 可变基因富含诸如逆境胁迫和应激反应等功能
- P. abyssinicum*特有泛基因在应激和化学反应上富集
- P. fulvum*特有泛基因在发育、生长、繁殖、细胞骨架上富集
- 提示*P. abyssinicum* 和 *P. fulvum*作为育种材料在未来提高豌豆品种抗性和产量的潜在价值
- 在其他作物中也存在类似的规律

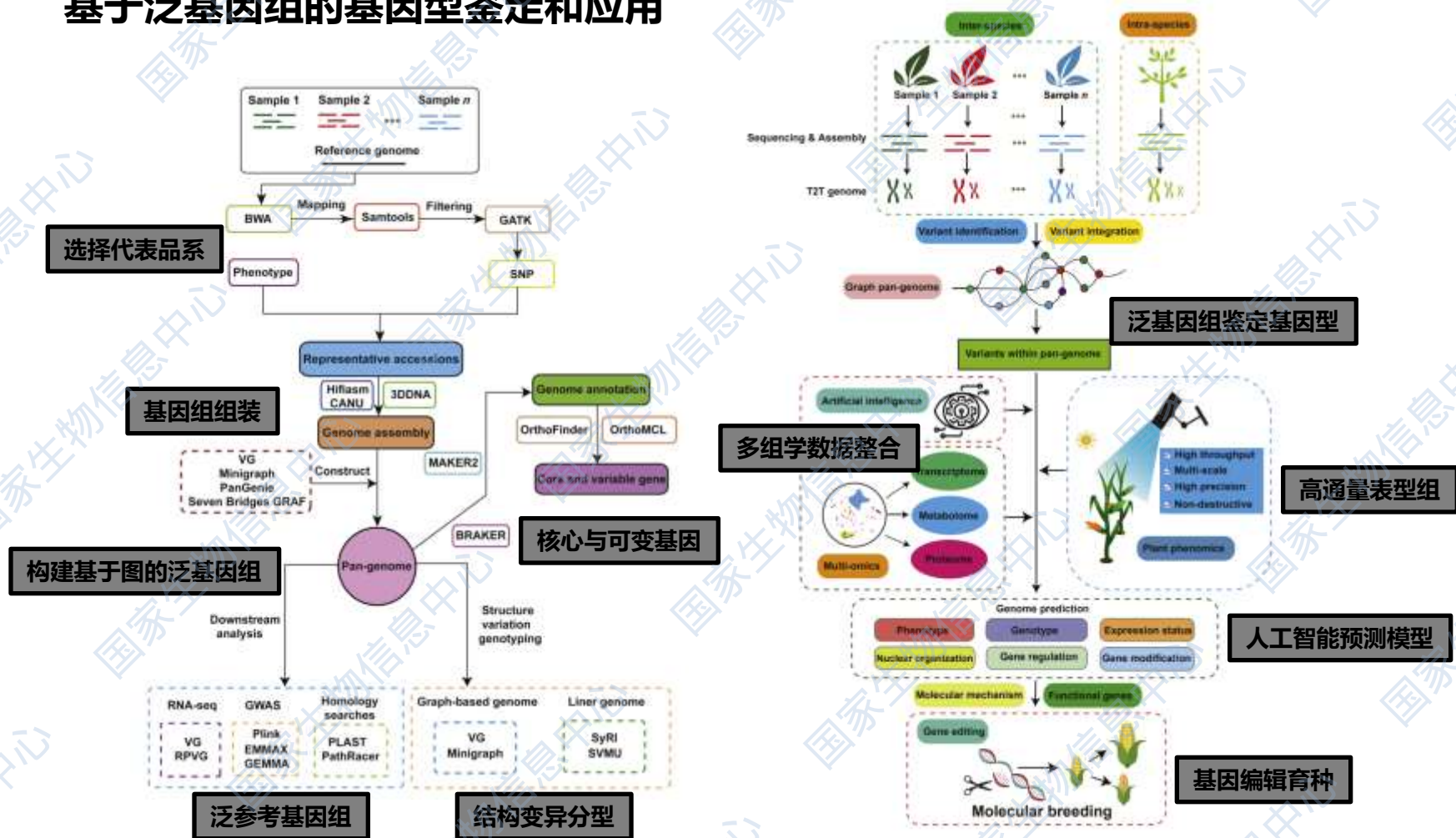


高质量中国豌豆基因组和泛基因组研究

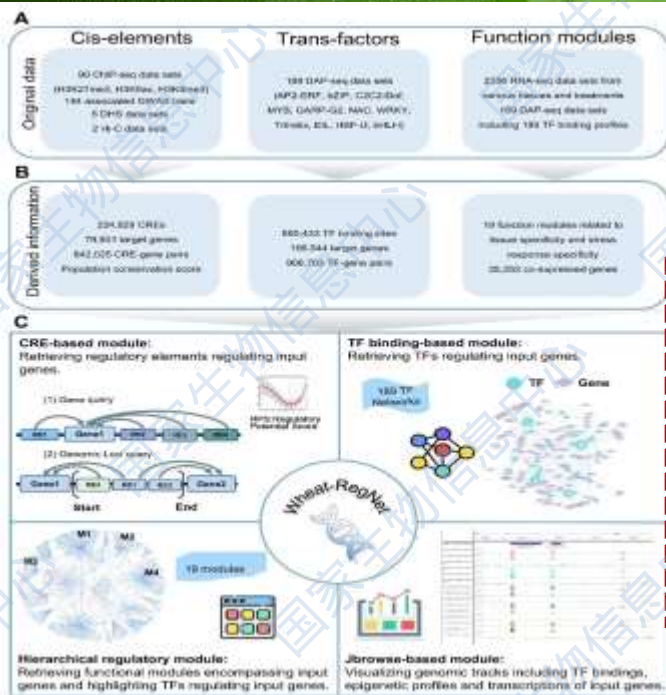
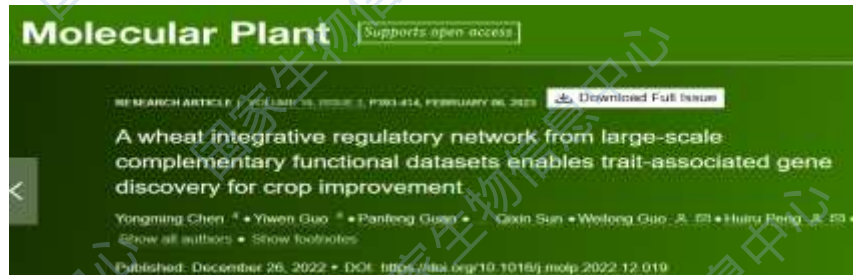


- 获得了显著改进的豌豆高质量基因组和注释, 在复杂重复区域和转座子连续性、准确性和完整性方面都表现优异。
- 构建了包含116个栽培和野生豌豆种质的泛基因组, 差异显著的存在/缺失变异 (PAV) 模式和功能富集, 表明豌豆近缘野生种和地方种质作为未来豌豆育种资源的重要潜力。
- 受限于短序列重测序数据, 且只有1个高质量基因组, 因此本研究只构建了基于基因的泛基因组。
- 未来将基于高质量的 *P. abyssinicum* 和 *P. fulvum* 基因组构建基于图的泛基因组, 在豌豆属内实现基于SV的表型关联研究。

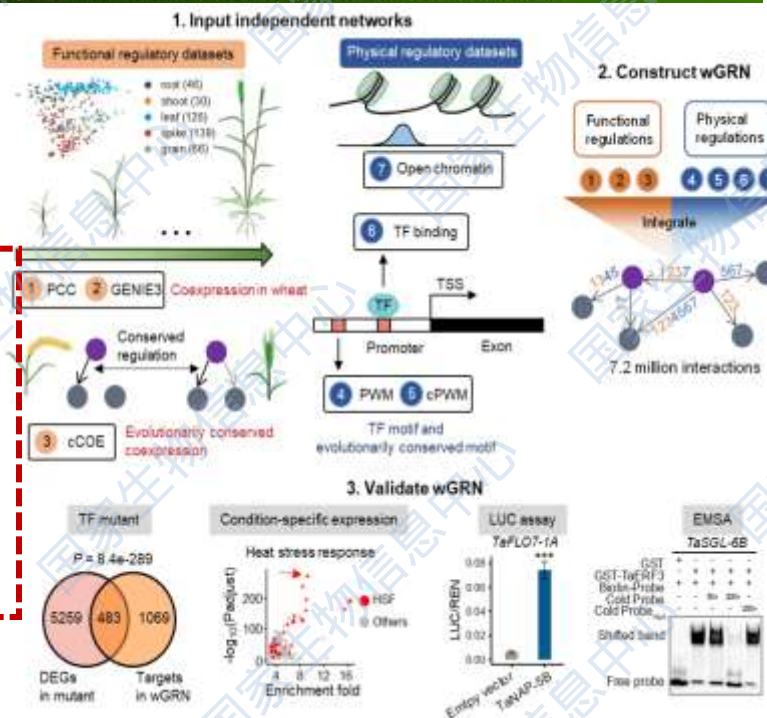
基于泛基因组的基因型鉴定和应用



小麦多组学数据整合与应用（数据整合）



保守motif
RNA-seq
TF结合位
点
染色体状
态



玉米多组学数据整合与应用（数据整合+实验技术，整体分析）

nature genetics

Article

<https://doi.org/10.1038/s41588-022-01262-1>

A multi-omics integrative network map of maize

华中农大

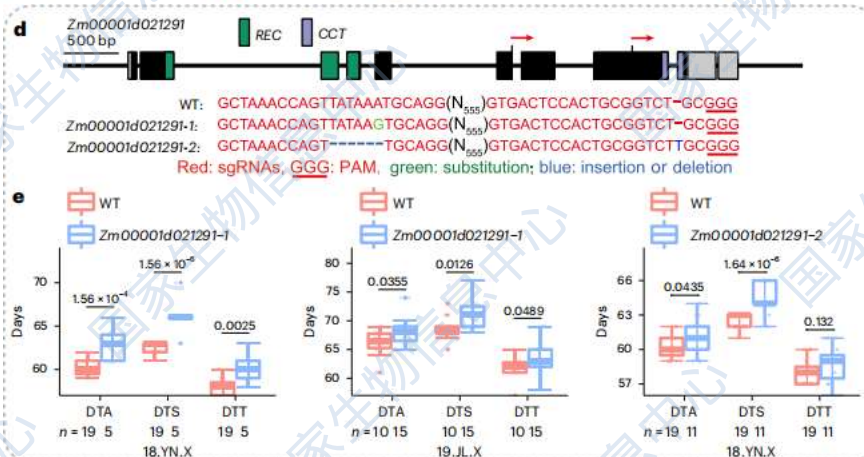
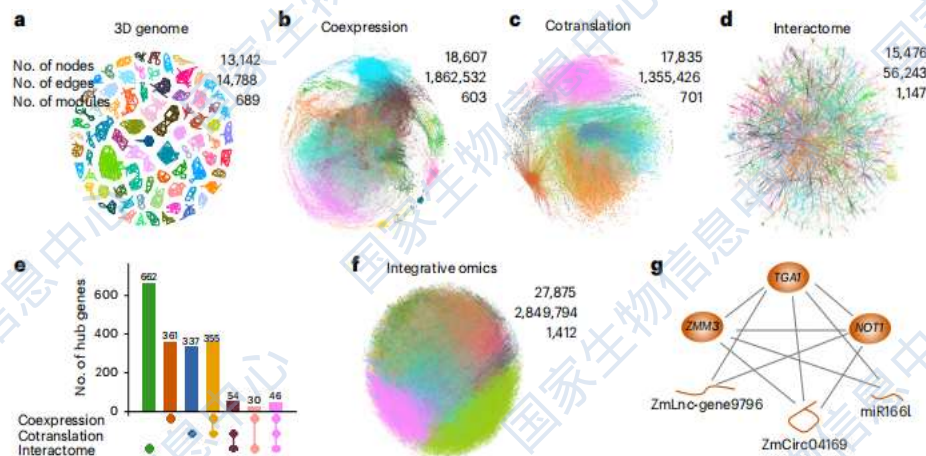
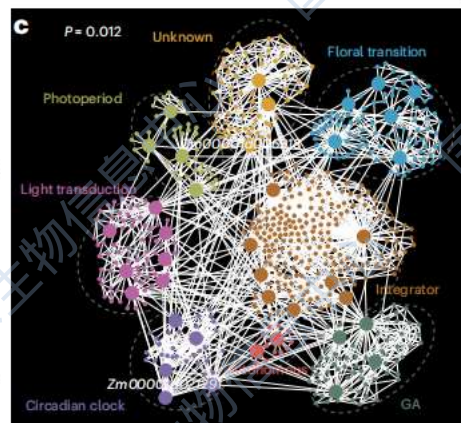
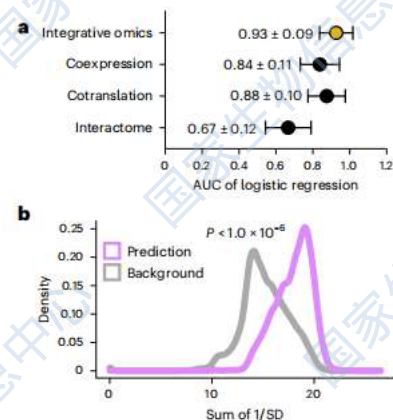
Received: 12 July 2021

Accepted: 3 November 2022

Published online: 29 December 2022

Check for updates

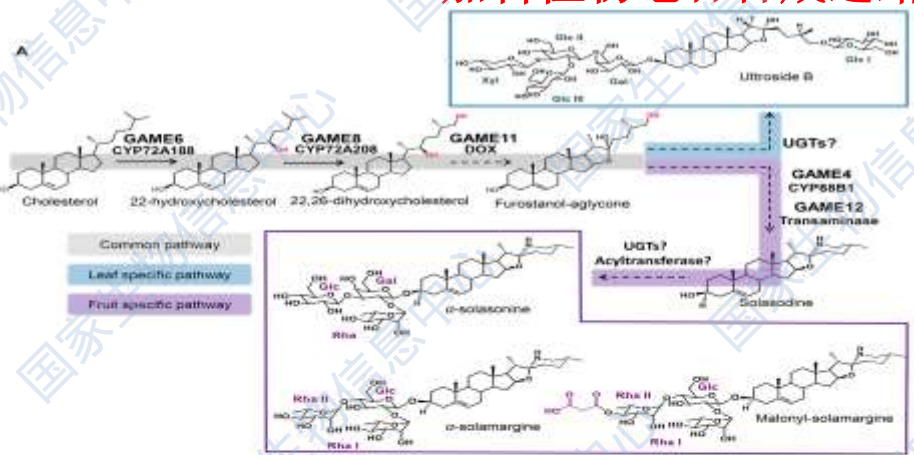
Linqian Han^{1,2,8}, Wanshun Zhong^{1,2,8}, Jia Qian^{1,2,8}, Minliang Jin^{1,2,8}, Peng Tian^{1,2,8}, Wanchao Zhu^{1,2,8}, Hongwei Zhang^{1,4}, Yonghao Sun^{1,2,8}, Jia-Wu Feng¹, Xiangguo Liu⁴, Guo Chen^{1,5}, Babar Farid^{1,6}, Ruonan Li¹, Zimo Xiong¹, Zhihui Tian¹, Juan Li^{1,2}, Zi Luo^{1,2}, Dengxiang Du¹, Sijia Chen¹, Qixiao Jin^{1,2}, Jiaxin Li¹, Zhao Li^{1,2}, Yan Liang¹, Xiaomeng Jin¹, Yong Peng¹, Chang Zheng¹, Xinnan Ye¹, Yuejia Yin⁴, Hong Chen¹, Weifu Li¹, Ling-Ling Chen¹, Qing Li^{1,2}, Jianbing Yan^{1,2,8}, Fang Yang^{1,2,8} & Lin Li^{1,2,8}



A scaffold protein manages the biosynthesis of steroidal defense metabolites in plants

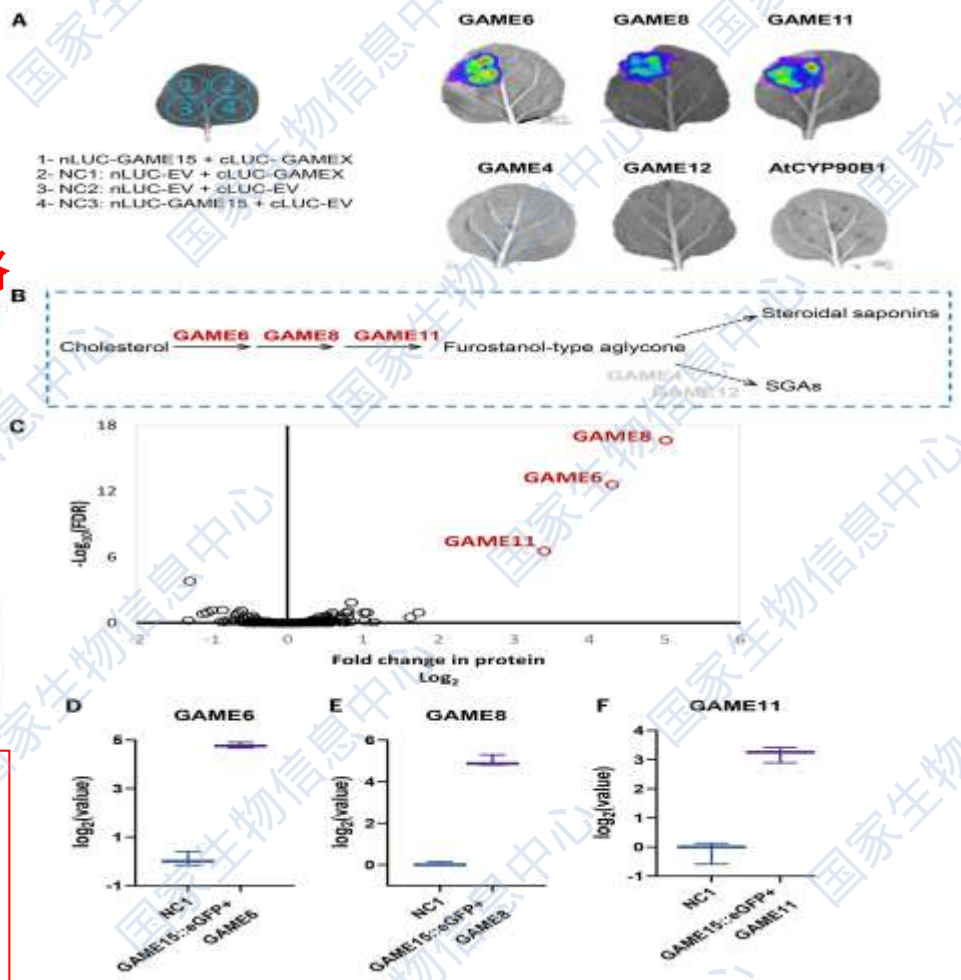
MARINNA BOCCA, DANNY KESSLEY, NIKKI DEBT, YET GRABE, CARLOS E. RODRIGUEZ LOPEZ, DANNY DRUECH, SARAH HEINICKE, SARAH O'CONNOR, ANIRUJHANTH S. SUNDARAM, Authors Info & Affiliations

SCIENCE - 77 (2024) - First Release - DOI:10.1126/science.adc3492



茄科植物皂苷合成通路

多物种共表达网络



- 体外重建代谢通路失败
- 跨物种基因共表达+基因簇+功能实验
- 生态意义
- 合成生物学

植物基因组学领域研究展望

- ◆ 高通量测序技术成为基础性技术
- ◆ 多组学数据整合和信息挖掘
- ◆ 作物有用基因的高通量挖掘
- ◆ 基因组的合成生物学改造
- ◆ 资源收集、育种、基因和基因组研究的统一

谢谢!