水稻含SET结构域基因家族功能分析

Functional Analysis of the Rice SET Domain Gene Family

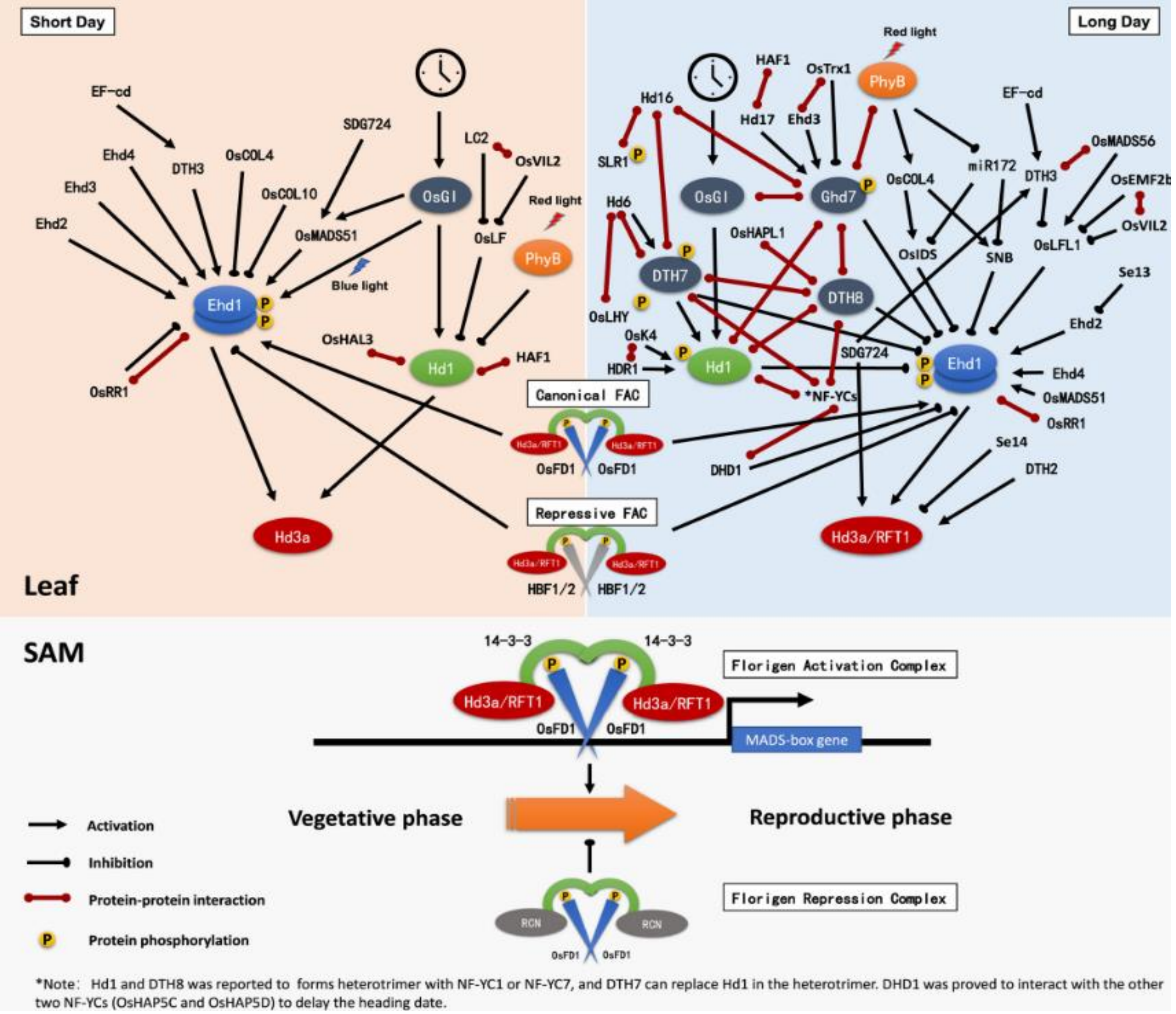
G07小组成员 杨启禹，曹新宇，李凤姣，李容博

汇报人 杨启禹

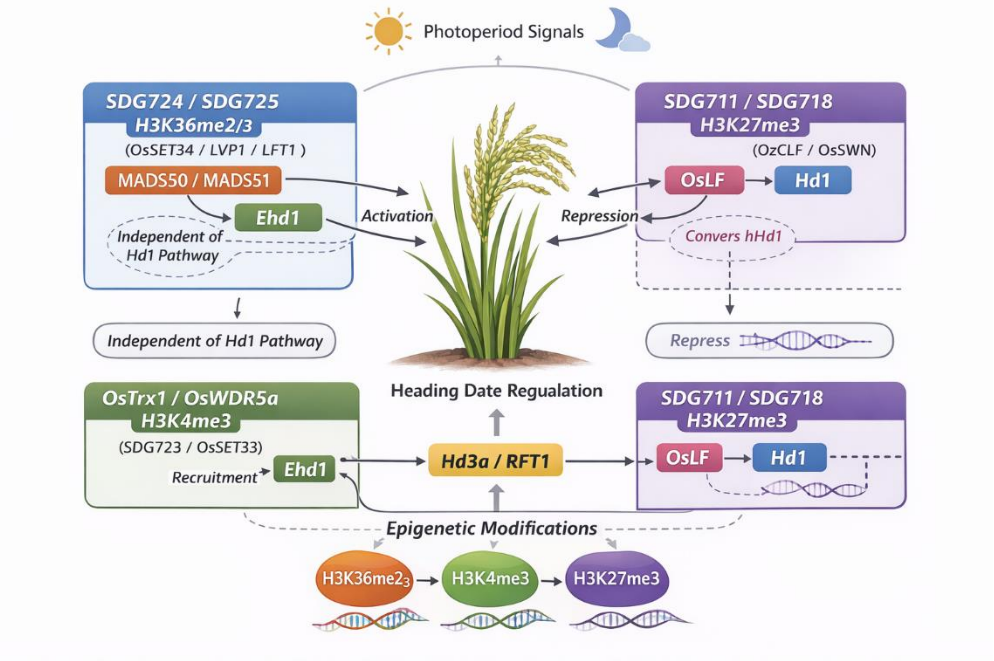
2026年1月

目录

1.1 背景介绍



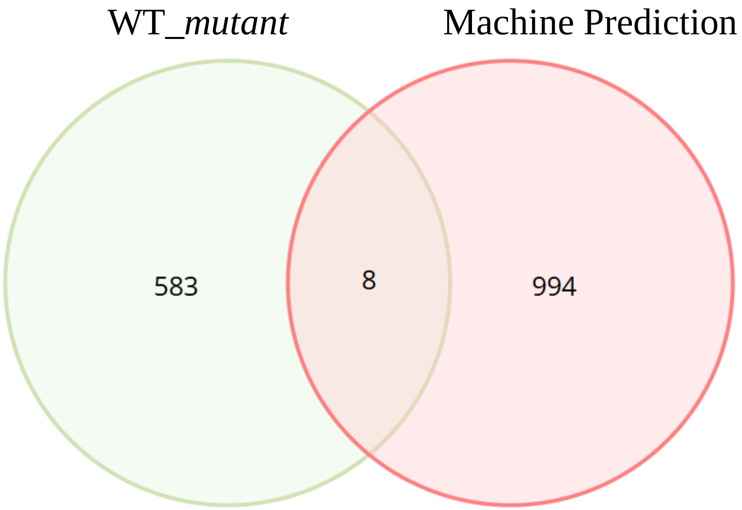
*Note: Hd1 and DTH8 was reported to forms heterotrimer with NF-YC1 or NF-YC7, and DTH7 can replace Hd1 in the heterotrimer. DHD1 was proved to interact with the other two NF-YCs (OsHAP5C and OsHAP5D) to delay the heading date.

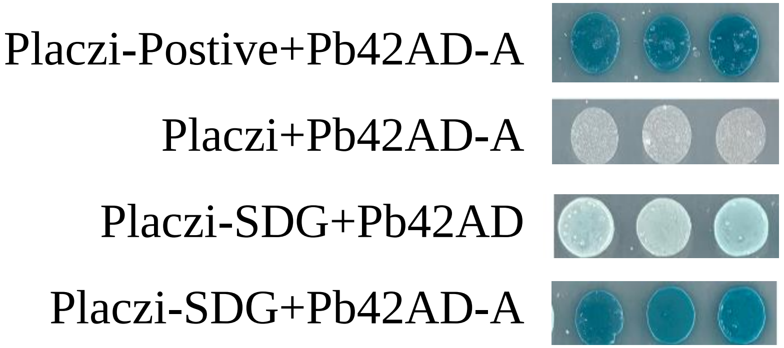


抽穗期是水稻最重要的农艺性状之一，与产量密切相关，由主效和微效 QTL / 基因共同控制。

目前，关于抽穗期基因通过组蛋白甲基化参与该过程的分子机制仍存在显著空白。

1.1 背景介绍





通过实验前期已有的RNA-seq结果，我们找到了一个组蛋白甲基转移酶；并通过后续的酵母单杂实验验证了其互作

Tools：RiceTFtarget

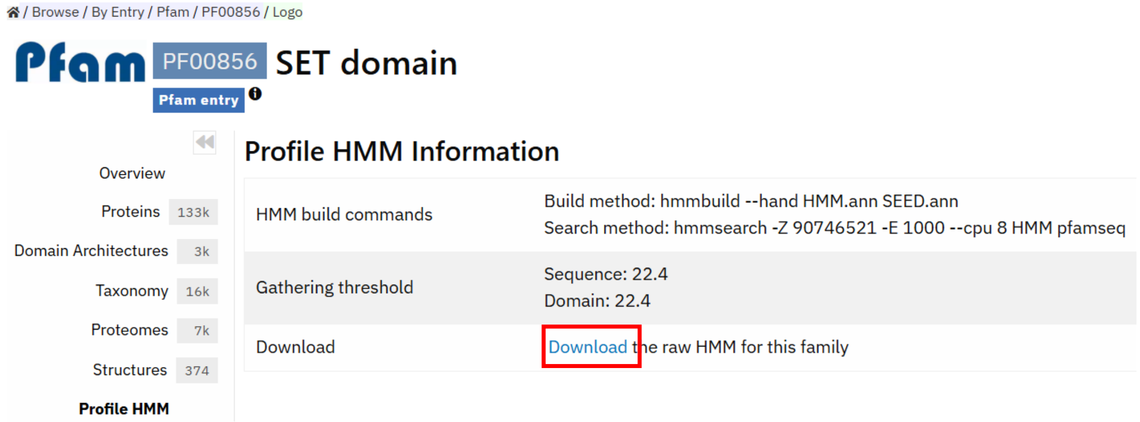
目录

一、背景介绍

二、家族蛋白查找

三、功能预测

2.1 家族蛋白查询



🏠 / Browse / By Entry / Pfam / PF00856 / Logo

Pfam PF00856 **SET domain**

Pfam entry

- Overview
- Proteins 133k
- Domain Architectures 3k
- Taxonomy 16k
- Proteomes 7k
- Structures 374
- **Profile HMM**

Profile HMM Information

HMM build commands	Build method: hmmbuild --hand HMM.ann SEED.ann Search method: hmmsearch -Z 90746521 -E 1000 --cpu 8 HMM pfamseq
Gathering threshold	Sequence: 22.4 Domain: 22.4
Download	Download the raw HMM for this family

Pfam PF00856 **SET domain**

Pfam entry

- Overview
- **Proteins** 1 3k
- Domain Architectures 3k
- Taxonomy 16k
- Proteomes 7k
- Structures 374
- Profile HMM
- AlphaFold 100k

This entry matches these proteins:

Filter By

▾ UniProt Curation

- ● All 132k
- ○ Reviewed 314
- ○ Unreviewed 131k

1 - 20 of **156** proteins

Search: Oryza sativa subsp

Accession	Name	Species	Gene	Predicted Structure	Matches
A0A0P0UZP2	Os01g0218800 protein	Oryza sativa subsp. japonica (Rice)	Os01g0218800	AlphaFold	

Legends

- 🟪 bacteria
- 🟥 viruses
- 🟨 archaea
- 🟩 eukaryota
- 🟦 Other

Weight Segments by

Number of sequences ▾

Font Size

14 ▾

Sunburst Depth

7 rings

2 ——●— 8

Selected Taxon

Name

Oryza

Number of sequences

654

Number of species

14

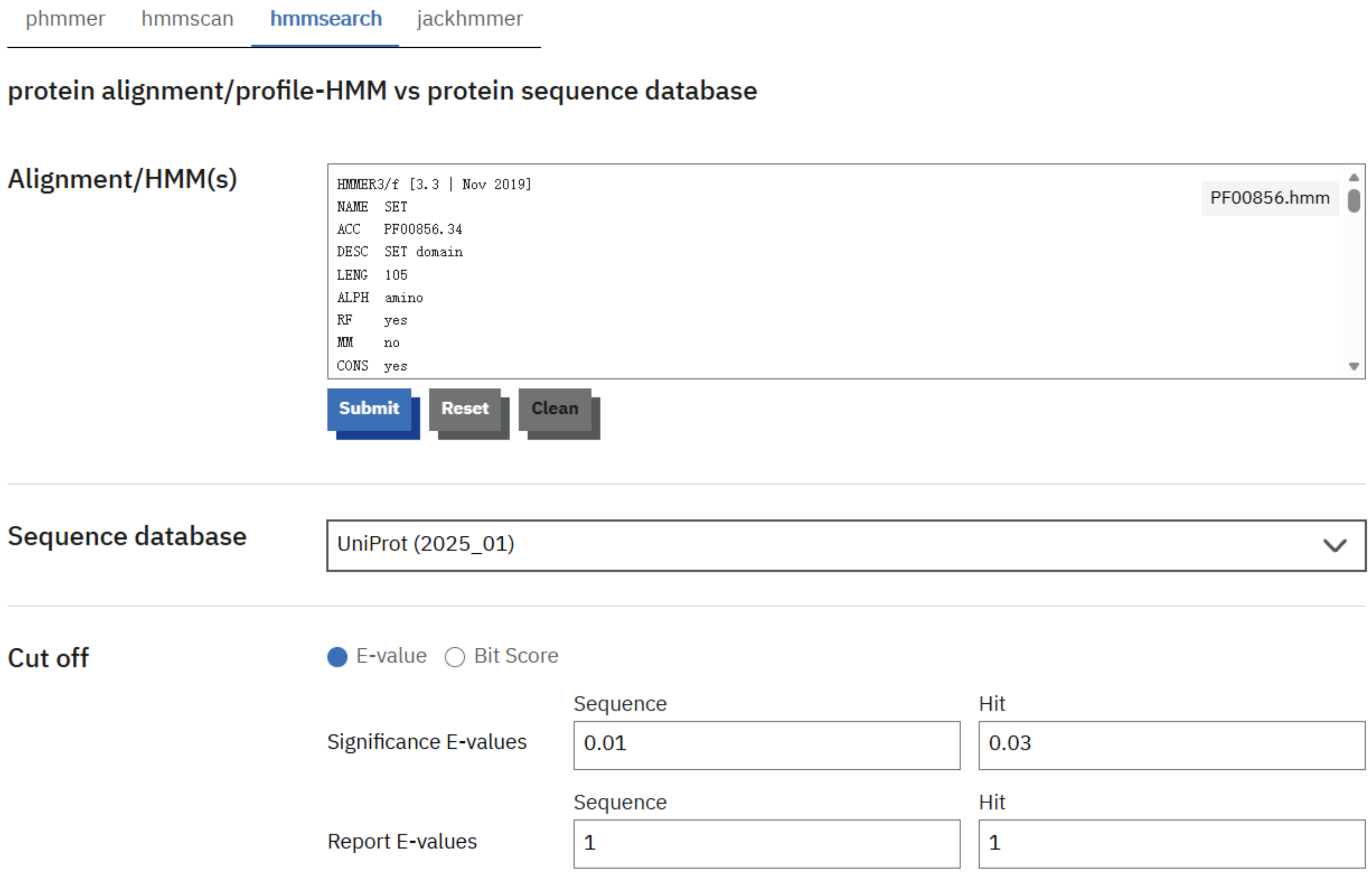
Lineage

root; *Eukaryota*; *Viridiplantae*; *Streptophyta*; *Magnoliopsida*; *Poales*; *Poaceae*; *Oryza*;

通过domain的PF号，下载该结构域的HMM模型，并观察该结构域的物种分布。

Tools： Pfam – Profile HMM

2.2 家族蛋白查询



protein alignment/profile-HMM vs protein sequence database

Alignment/HMM(s)

```
HMMER3/f [3.3 | Nov 2019]
NAME  SET
ACC   PF00856.34
DESC  SET domain
LENG  105
ALPH  amino
RF    yes
MM    no
CONS  yes
```

PF00856.hmm

Sequence database: UniProt (2025_01)

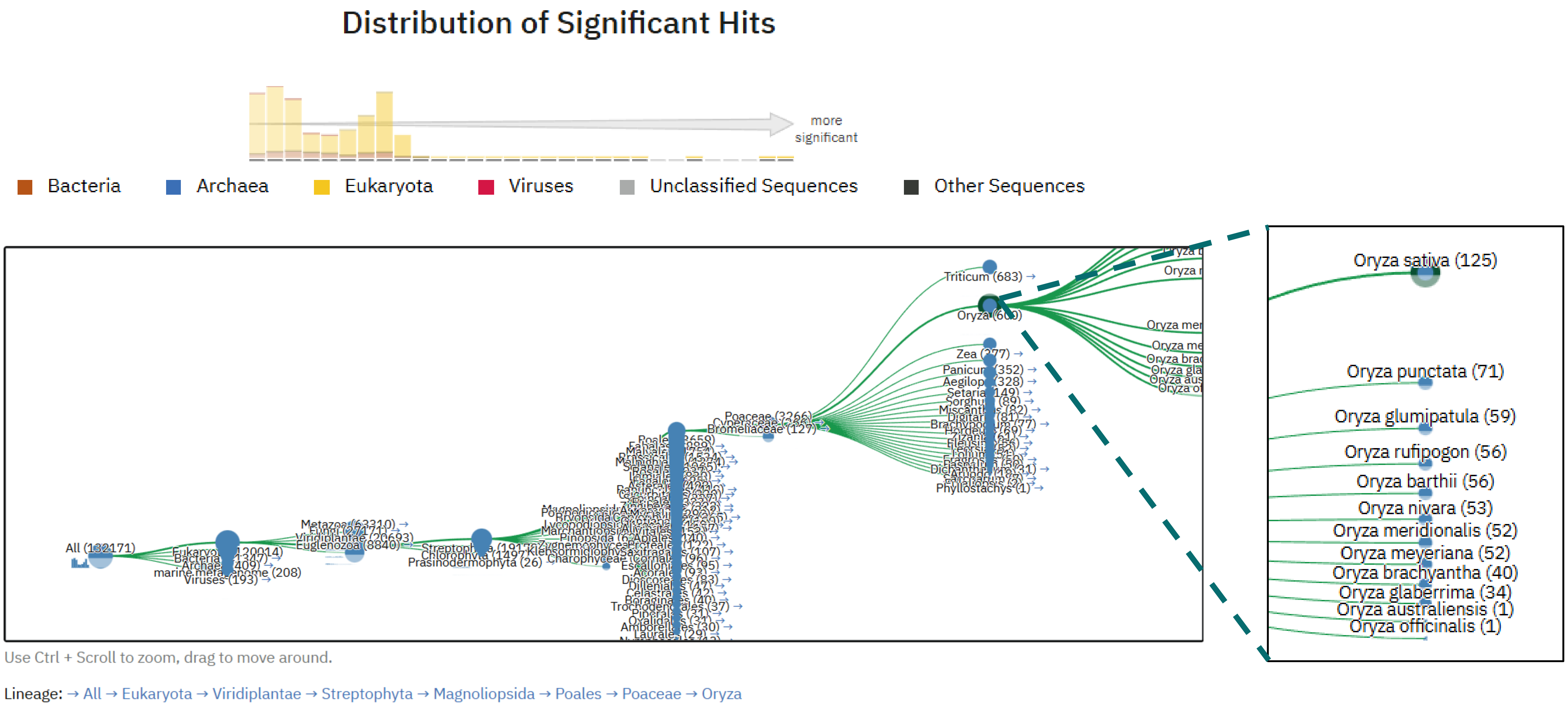
Cut off: ● E-value ○ Bit Score

	Sequence	Hit
Significance E-values	0.01	0.03
Report E-values	1	1

通过HMMERsearch，进行该结构域的检索，注意所选数据库。

Tools：EMBL – HMMER

2.1 家族蛋白查询 —— HMMER查找包含domain的蛋白



Distribution of Significant Hits

Use Ctrl + Scroll to zoom, drag to move around.

Lineage: → All → Eukaryota → Viridiplantae → Streptophyta → Magnoliopsida → Poales → Poaceae → Oryza

Tools：EMBL – HMMER

2.3 冗余蛋白序列删除

* fasta

Oryza125FulllengthFASTA.fa

注意：

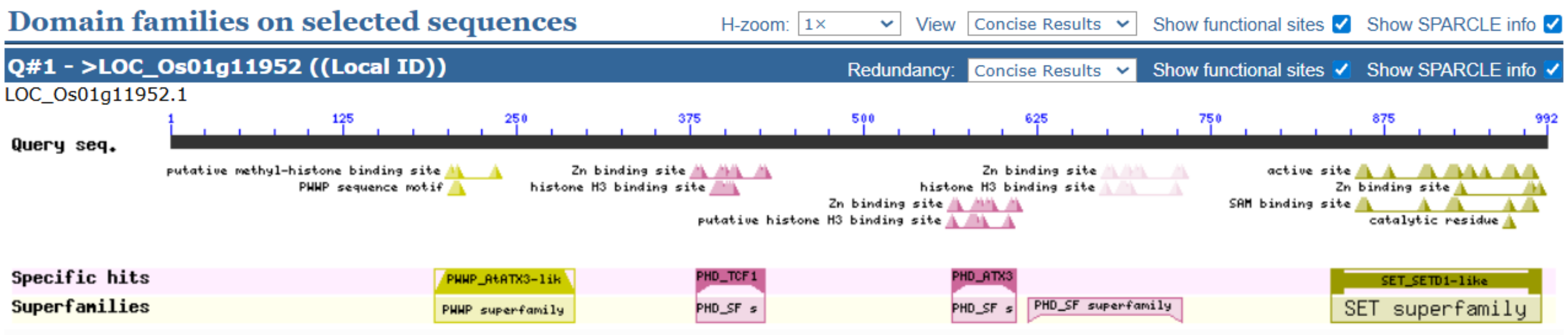
*上传的文件必须为FASTA格式的文件。文件后缀是：.fa,.fasta,.fna

* identity: 0.9

* threads: 10

* word_length: 5

* Email: 2239652834@qq.com



Domain families on selected sequences

H-zoom: 1×　View: Concise Results　Show functional sites ☑　Show SPARCLE info ☑

Q#1 - >LOC_Os01g11952 ((Local ID))

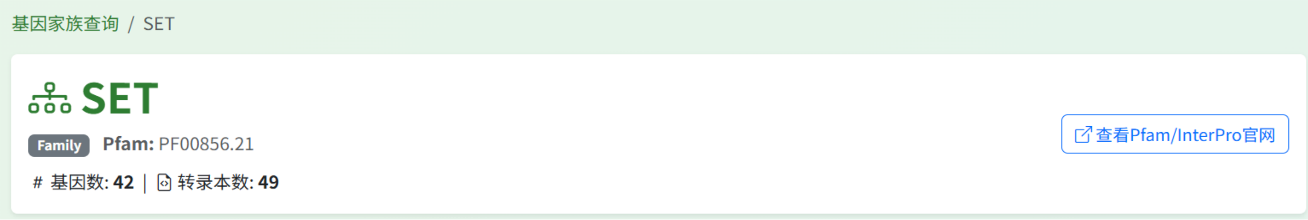
Redundancy: Concise Results　Show functional sites ☑　Show SPARCLE info ☑

LOC_Os01g11952.1

经过CD-HIT处理后，对结果取并集，并进行NCBI-CDD的重新注释和筛选，确定最终家族成员。

Tools：NMDC – CD_hit；NCBI - CDD

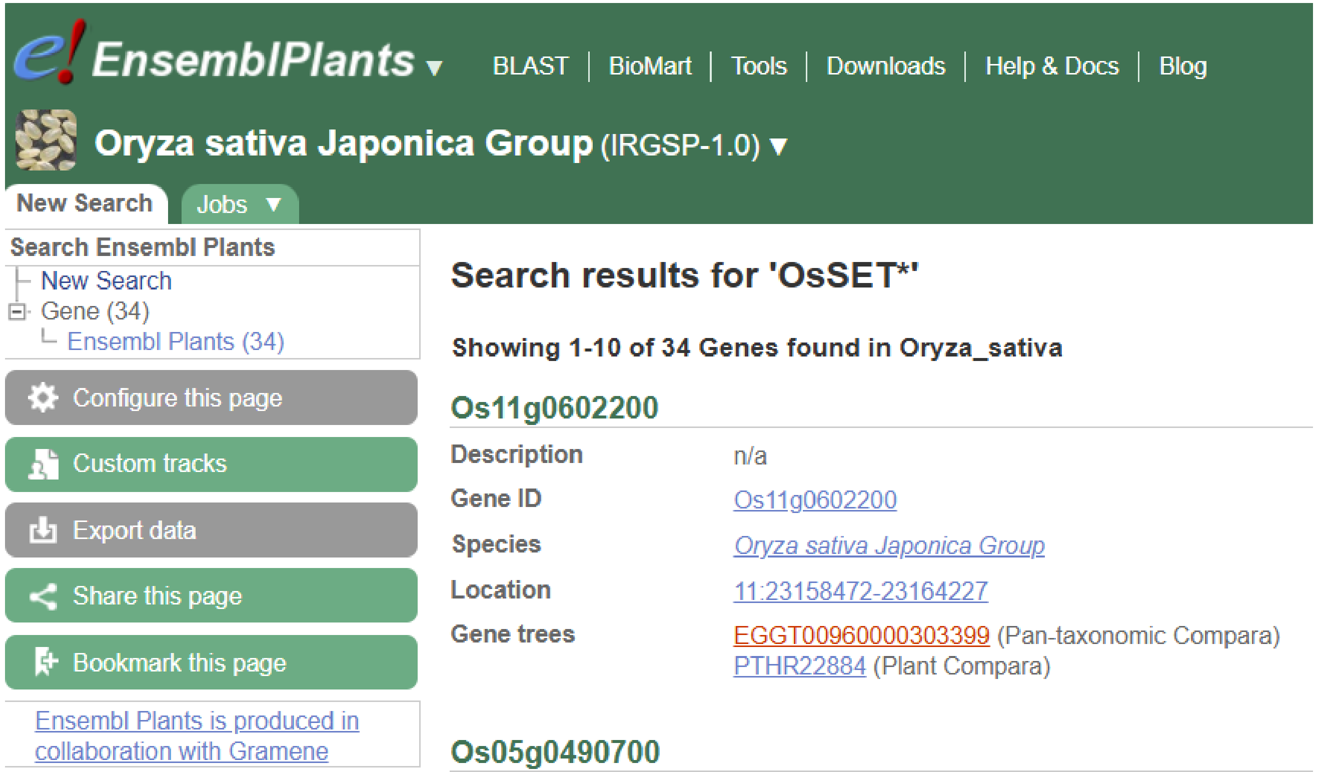
2.4 家族蛋白成员确认



SET

Family **Pfam:** PF00856.21

基因数: **42** | 转录本数: **49**



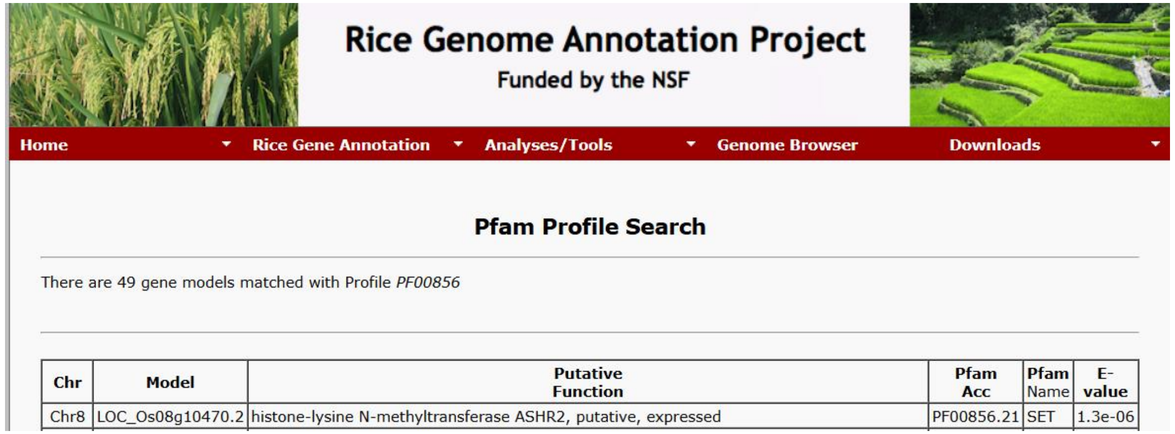
Search results for 'OsSET*'

Showing 1-10 of 34 Genes found in Oryza_sativa

Os11g0602200

Description	n/a
Gene ID	Os11g0602200
Species	*Oryza sativa Japonica Group*
Location	11:23158472-23164227
Gene trees	EGGT00960000303399 (Pan-taxonomic Compara) PTHR22884 (Plant Compara)

Os05g0490700



Pfam Profile Search

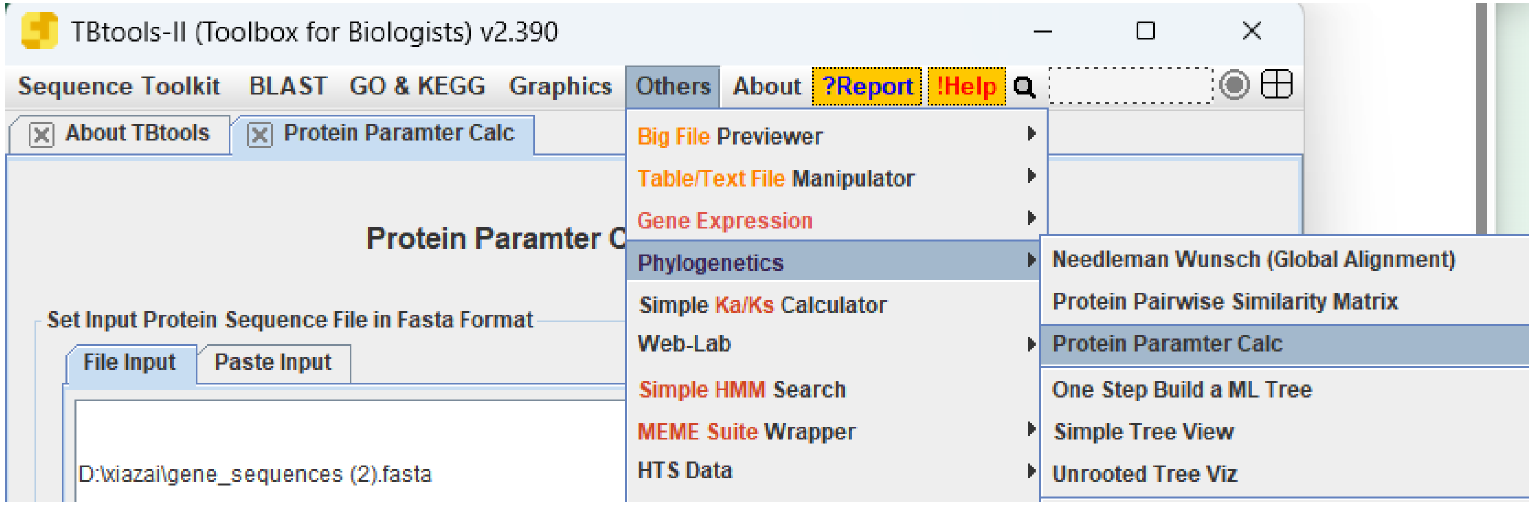
There are 49 gene models matched with Profile *PF00856*

Chr	Model	Putative Function	Pfam Acc	Pfam Name	E-value
Chr8	LOC_Os08g10470.2	histone-lysine N-methyltransferase ASHR2, putative, expressed	PF00856.21	SET	1.3e-06

结合已有植物专用数据库（RGAP，EnsemblPlants等），进行比对，最终确认该基因家族共有42个成员

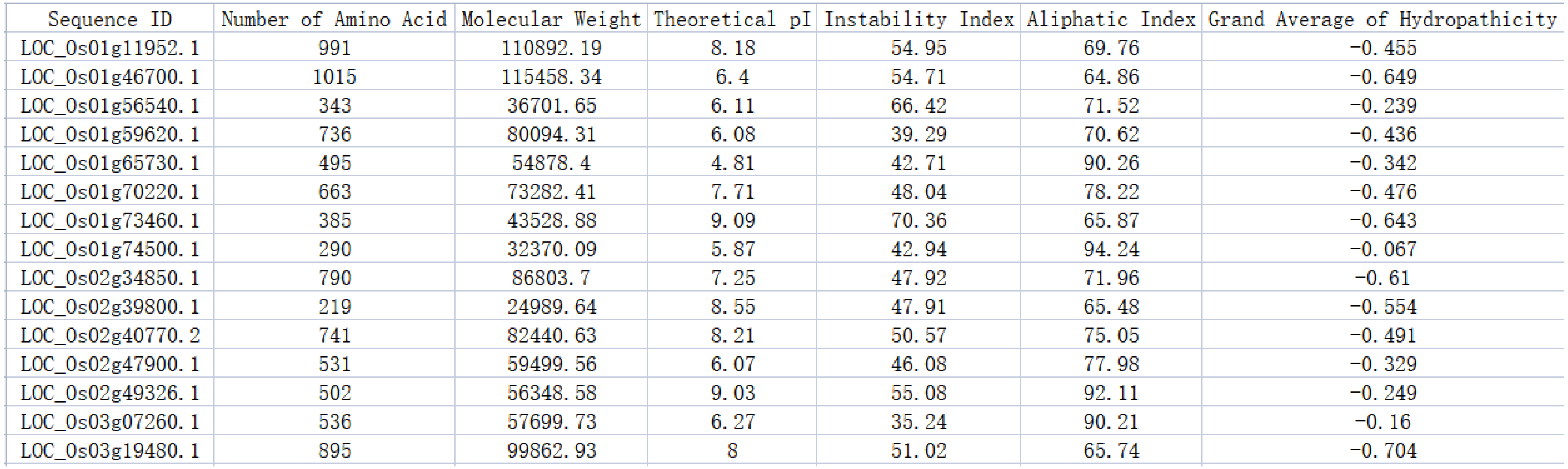
目录

3.1 基础理化性质分析



D:\xiazai\gene_sequences (2).fasta

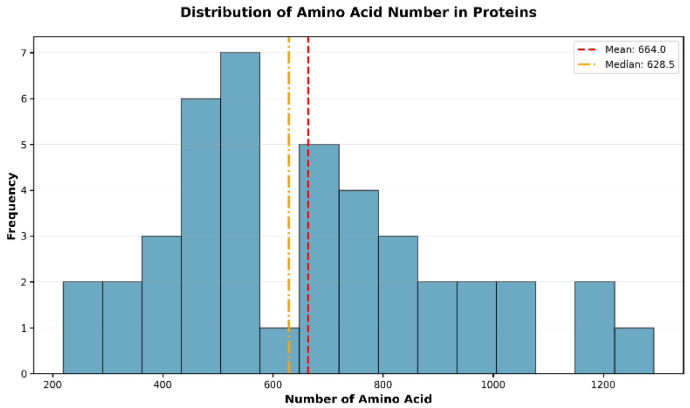
通过Tbtools的Protein Paramter Calc功能完成家族蛋白的基本理化性质结果汇总

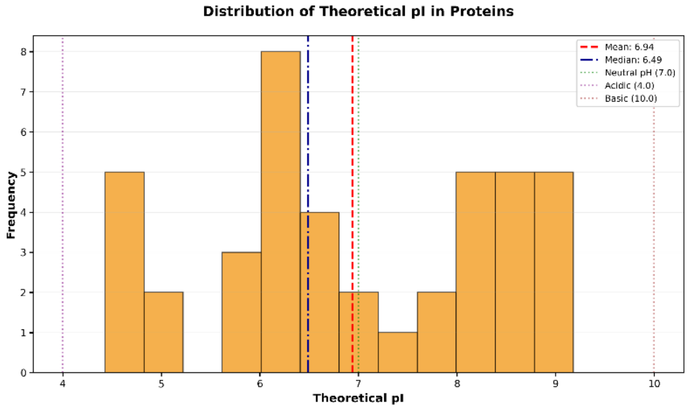


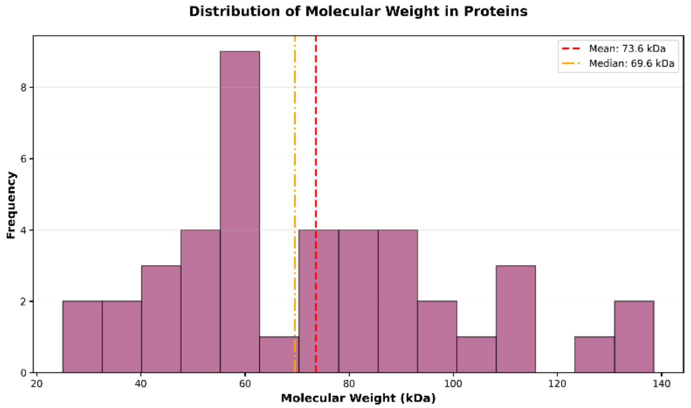
Sequence ID	Number of Amino Acid	Molecular Weight	Theoretical pI	Instability Index	Aliphatic Index	Grand Average of Hydropathicity
LOC_Os01g11952.1	991	110892.19	8.18	54.95	69.76	-0.455
LOC_Os01g46700.1	1015	115458.34	6.4	54.71	64.86	-0.649
LOC_Os01g56540.1	343	36701.65	6.11	66.42	71.52	-0.239
LOC_Os01g59620.1	736	80094.31	6.08	39.29	70.62	-0.436
LOC_Os01g65730.1	495	54878.4	4.81	42.71	90.26	-0.342
LOC_Os01g70220.1	663	73282.41	7.71	48.04	78.22	-0.476
LOC_Os01g73460.1	385	43528.88	9.09	70.36	65.87	-0.643
LOC_Os01g74500.1	290	32370.09	5.87	42.94	94.24	-0.067
LOC_Os02g34850.1	790	86803.7	7.25	47.92	71.96	-0.61
LOC_Os02g39800.1	219	24989.64	8.55	47.91	65.48	-0.554
LOC_Os02g40770.2	741	82440.63	8.21	50.57	75.05	-0.491
LOC_Os02g47900.1	531	59499.56	6.07	46.08	77.98	-0.329
LOC_Os02g49326.1	502	56348.58	9.03	55.08	92.11	-0.249
LOC_Os03g07260.1	536	57699.73	6.27	35.24	90.21	-0.16
LOC_Os03g19480.1	895	99862.93	8	51.02	65.74	-0.704

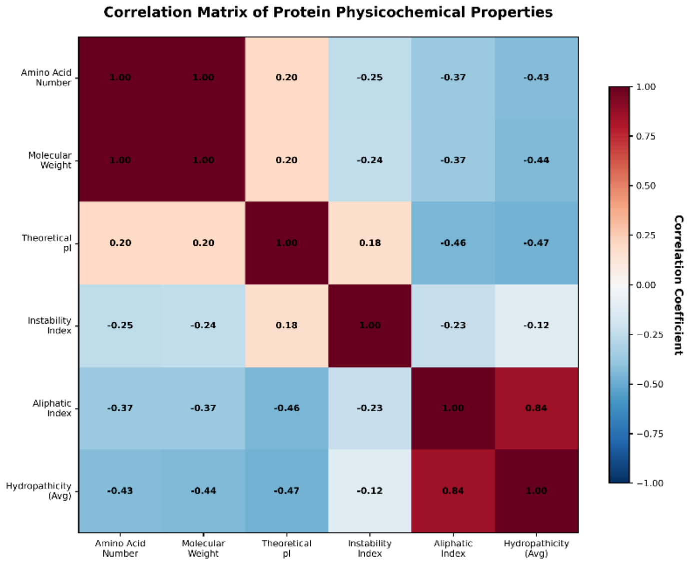
Tools：ProtParam；TBtools

3.1 基础理化性质分析

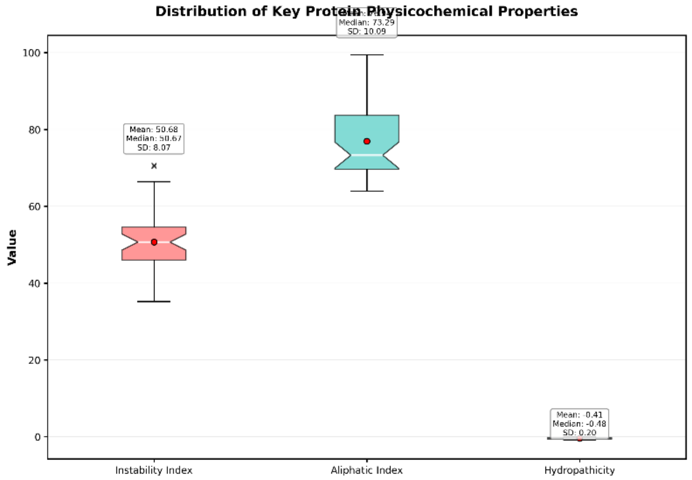

Distribution of Amino Acid Number in Proteins

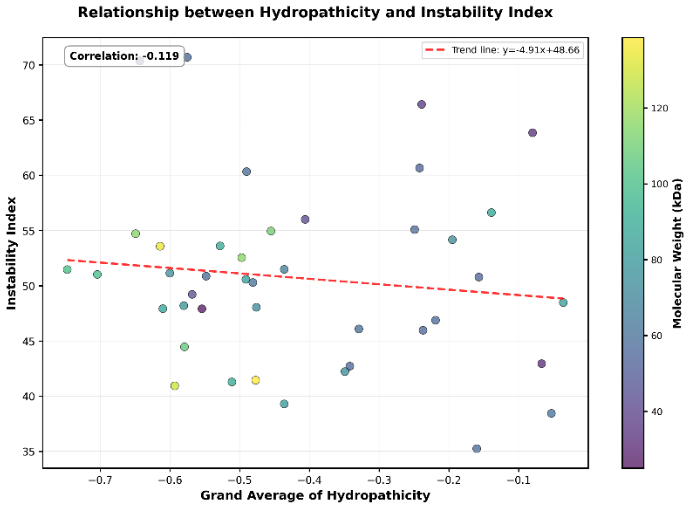

Distribution of Theoretical pI in Proteins


Distribution of Molecular Weight in Proteins


Correlation Matrix of Protein Physicochemical Properties

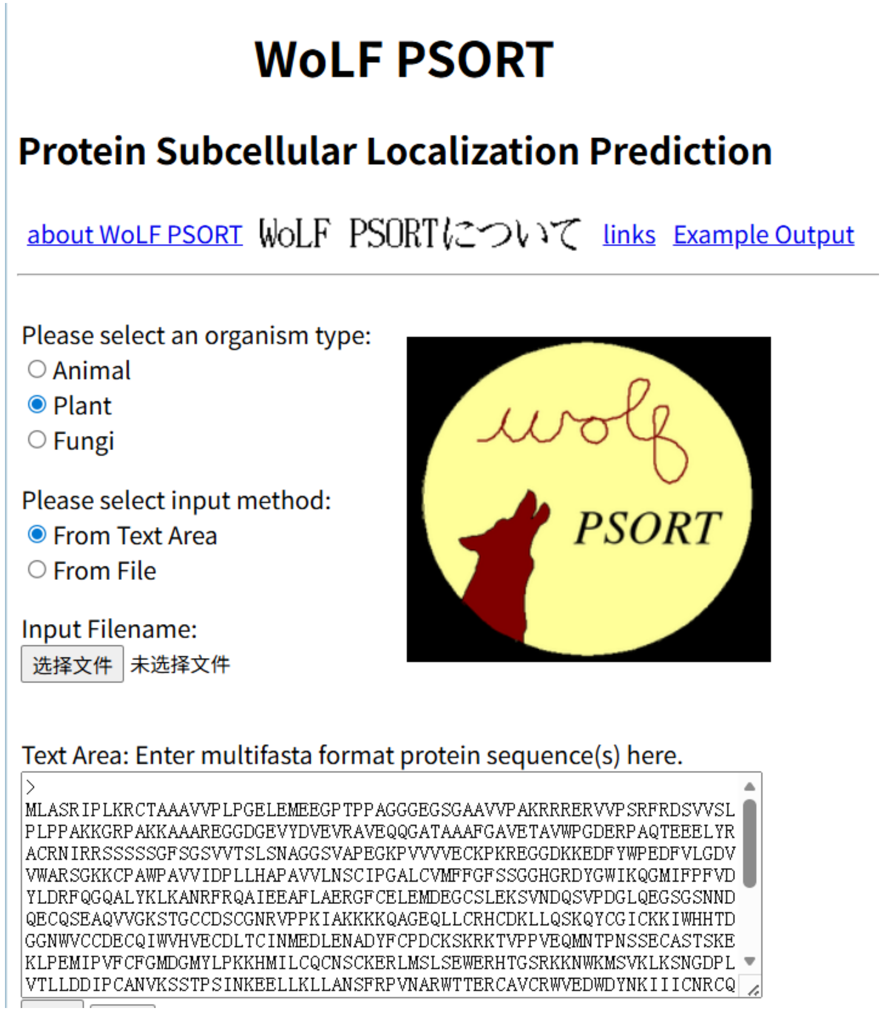
	Amino Acid Number	Molecular Weight	Theoretical pI	Instability Index	Aliphatic Index	Hydropathicity (Avg)
Amino Acid Number	1.00	1.00	0.20	-0.25	-0.37	-0.43
Molecular Weight	1.00	1.00	0.20	-0.24	-0.37	-0.44
Theoretical pI	0.20	0.20	1.00	0.18	-0.46	-0.47
Instability Index	-0.25	-0.24	0.18	1.00	-0.23	-0.12
Aliphatic Index	-0.37	-0.37	-0.46	-0.23	1.00	0.84
Hydropathicity (Avg)	-0.43	-0.44	-0.47	-0.12	0.84	1.00


Distribution of Key Protein Physicochemical Properties


Relationship between Hydropathicity and Instability Index

Tools：ProtParam；TBtools

3.2 亚细胞定位预测



WoLF PSORT

Protein Subcellular Localization Prediction

Please select an organism type:

○ Animal
● Plant
○ Fungi

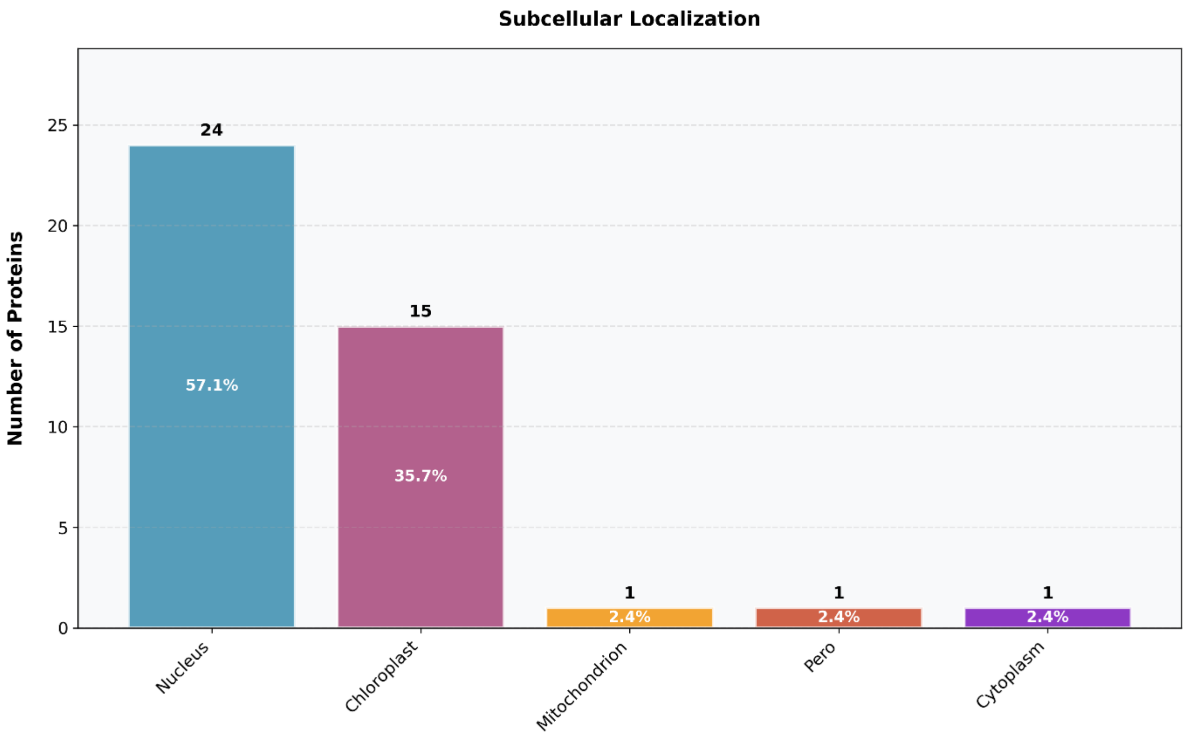
Please select input method:

● From Text Area
○ From File

Input Filename:

Text Area: Enter multifasta format protein sequence(s) here.

```
>
MLASRIPLKRCTAAAVVPLPGELEMEEGPTPPAGGGEGSGAAVVPAKRRRERVVPSRFRDSVVSL
PLPPAKKGRPAKKAAAREGGDGEVYDVEVRAVEQQGATAAAFGAVETAVWPGDERPAQTEEELYR
ACRNIRRSSSSSGFSGSVVTSLSNAGGSVAPEGKPVVVVECKPKREGGDKKEDFYWPEDFVLGDV
VWARSGKKCPAWPAVVIDPLLHAPAVVLNSCIPGALCVMFFGFSSGGHGRDYGWIKQGMIFPFVD
YLDRFQGQALYKLKANRFRQAIEEAFLAERGFCELEMDEGCSLEKSVNDQSVPDGLQEGSGSNND
QECQSEAQVVGKSTGCCDSCGNRVPPKIAKKKKQAGEQLLCRHCDKLLQSKQYCGICKKIWHHTD
GGNWVCCDECQIWVHVECDLTCINMEDLENADYFCPDCKSKRKTVPPVEQMNTPNSSECASTSKE
KLPEMIPVFCFGMDGMYLPKKHMILCQCNSCKERLMSLSEWERHTGSRKKNWKMSVKLKSNGDPL
VTLLDDIPCANVKSSTPSINKEELLKLLANSFRPVNARWTTERCAVCRWVEDWDYNKIIICNRCQ
```

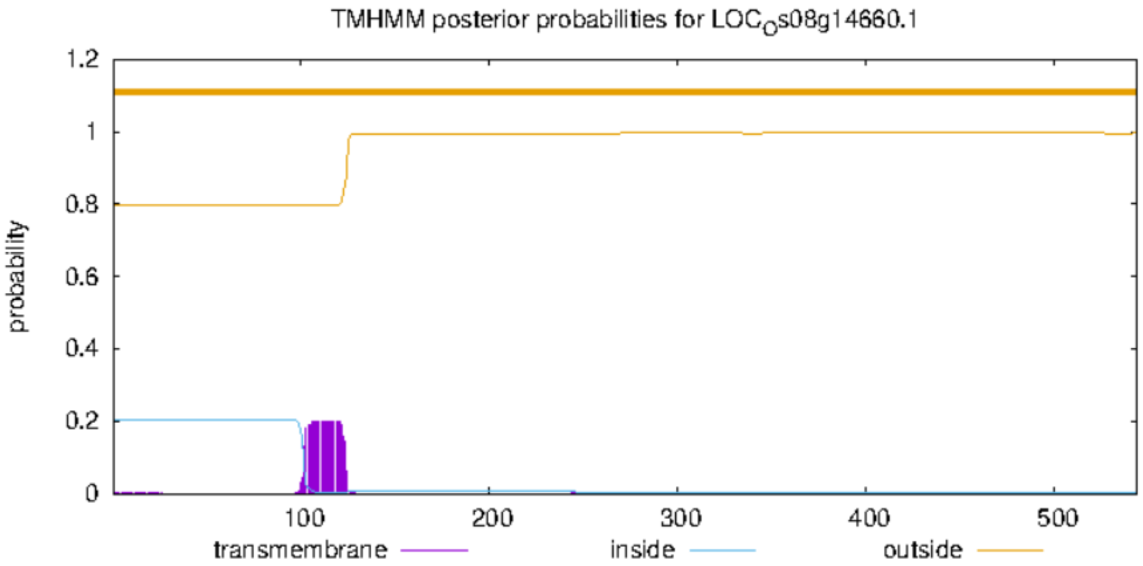


大部分蛋白预测位于细胞核和叶绿体中，预示着参与了基因表达调控、转录过程、染色质修饰等细胞核核心功能；结合叶绿体功能，也预示着部分蛋白参与了光合作用、叶绿体代谢、叶绿体基因表达等关键功能

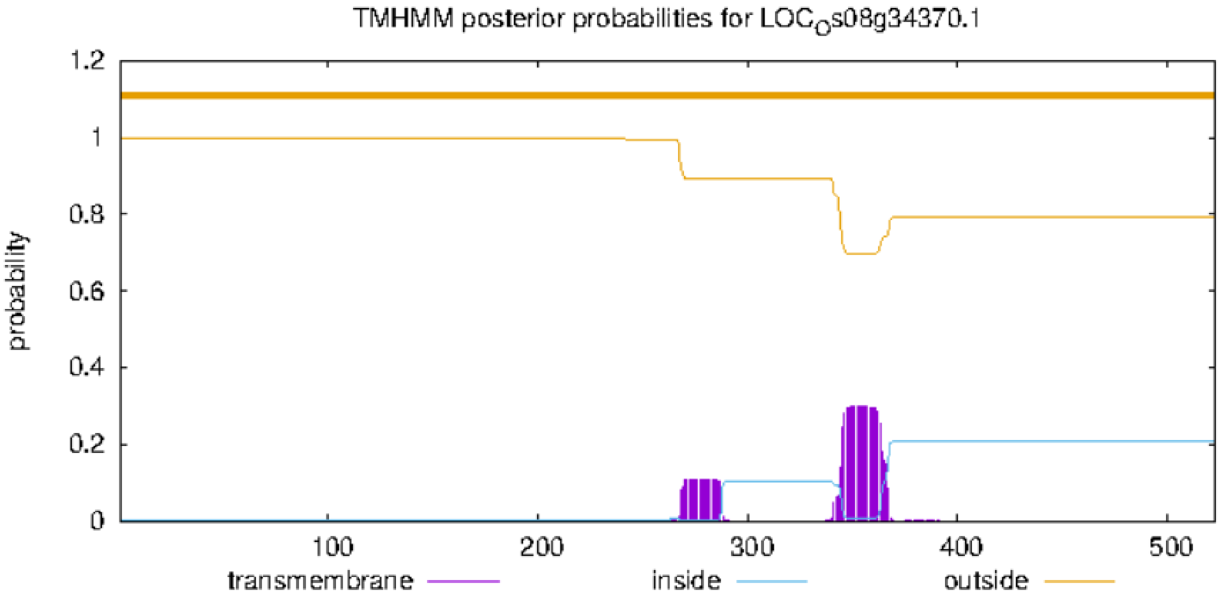
Tools：WoLF PSORT

3.3 跨膜结构预测

```
# LOC_Os08g14660.1 Length: 544
# LOC_Os08g14660.1 Number of predicted TMHs:  0
# LOC_Os08g14660.1 Exp number of AAs in TMHs: 4.67808999999998
# LOC_Os08g14660.1 Exp number, first 60 AAs:  0.01955
# LOC_Os08g14660.1 Total prob of N-in:        0.20268
LOC_Os08g14660.1        TMHMM2.0        outside      1   544
```



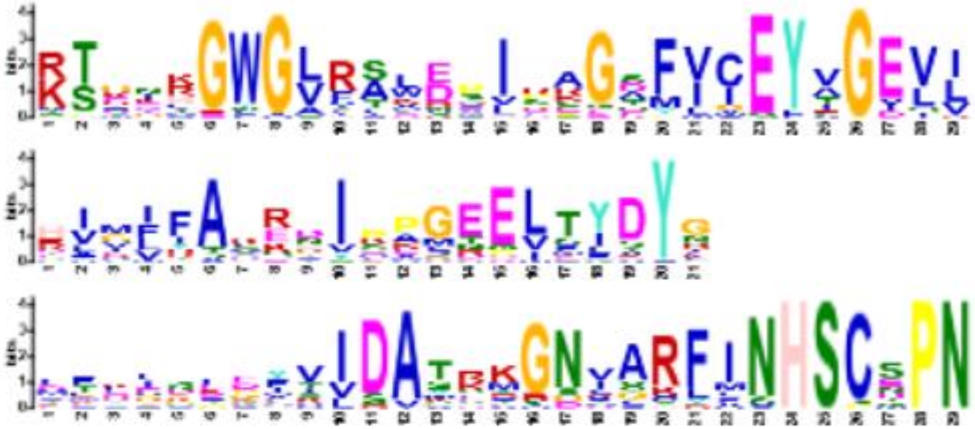
```
# LOC_Os08g34370.1 Length: 523
# LOC_Os08g34370.1 Number of predicted TMHs:  0
# LOC_Os08g34370.1 Exp number of AAs in TMHs: 8.56308999999999
# LOC_Os08g34370.1 Exp number, first 60 AAs:  0
# LOC_Os08g34370.1 Total prob of N-in:        0.00298
LOC_Os08g34370.1        TMHMM2.0        outside      1   523
```

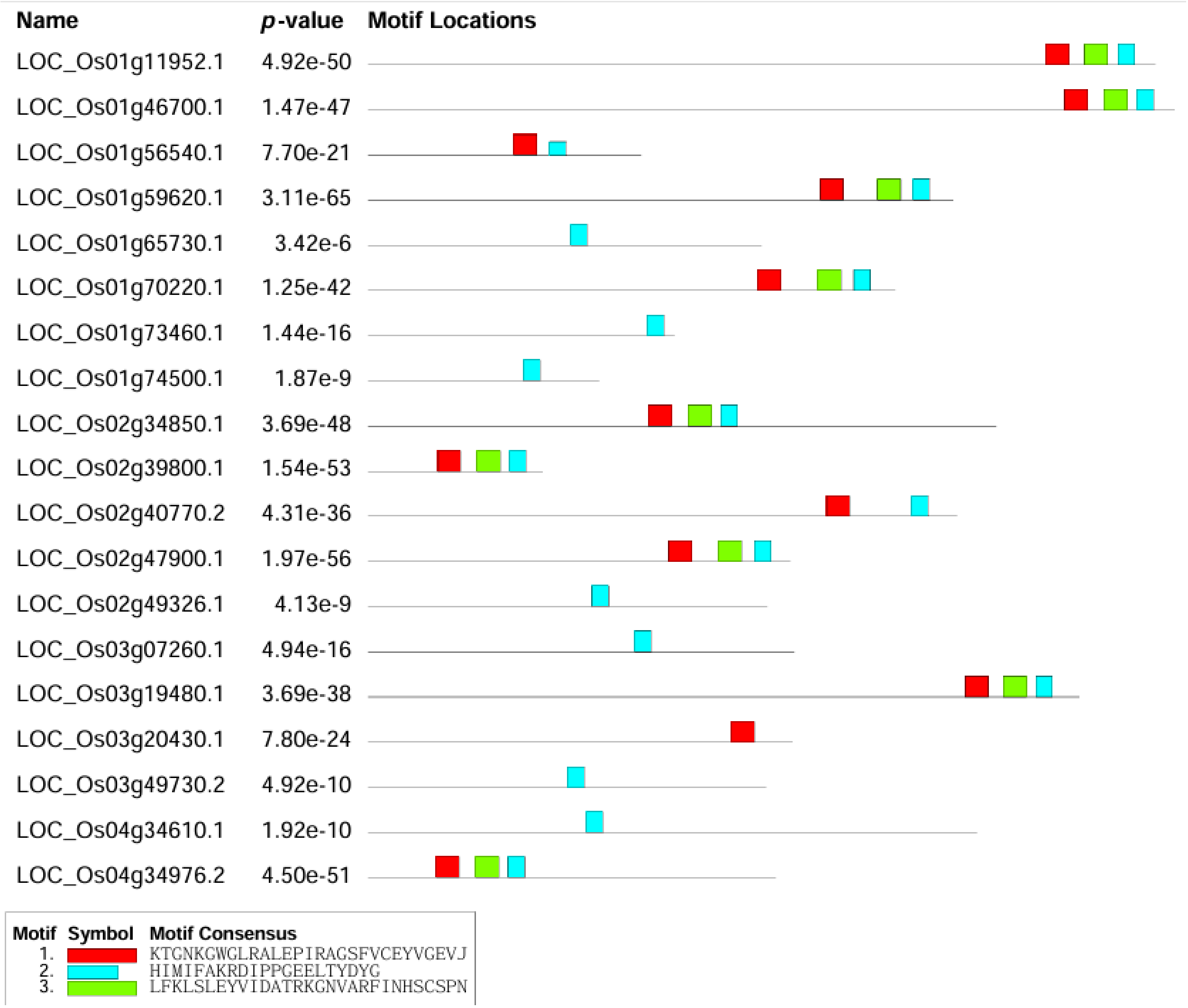


通过TMHMM预测，仅有两个蛋白含有跨膜结构。

Tools：TMHMM

3.4 保守 Motif 分析





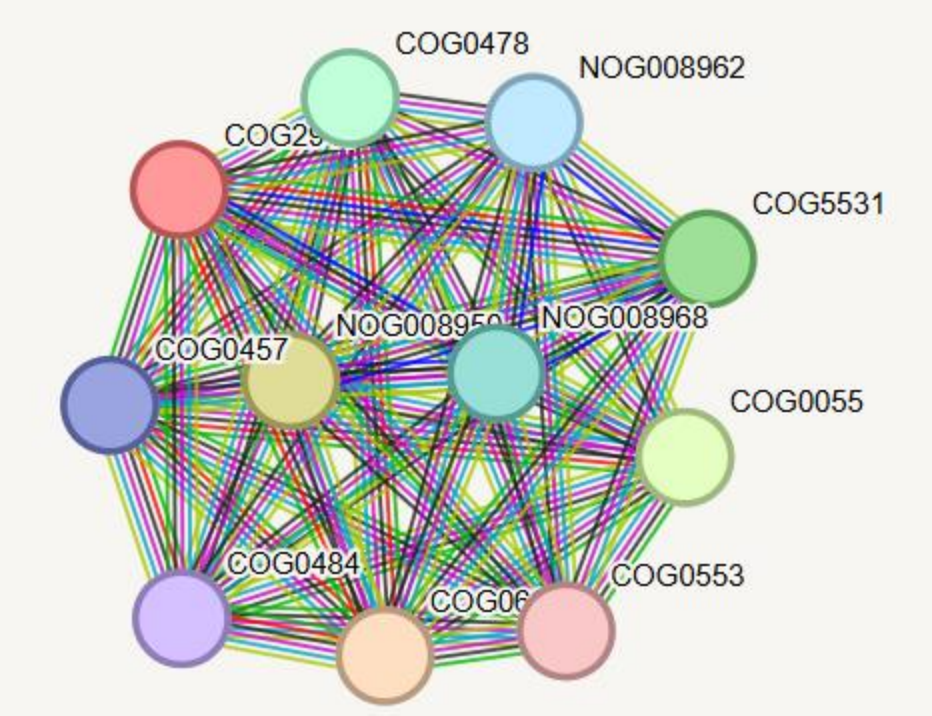
Name	*p*-value	Motif Locations
LOC_Os01g11952.1	4.92e-50	🟥 🟩 🟦
LOC_Os01g46700.1	1.47e-47	🟥 🟩 🟦
LOC_Os01g56540.1	7.70e-21	🟥 🟦
LOC_Os01g59620.1	3.11e-65	🟥 🟩 🟦
LOC_Os01g65730.1	3.42e-6	🟦
LOC_Os01g70220.1	1.25e-42	🟥 🟩 🟦
LOC_Os01g73460.1	1.44e-16	🟦
LOC_Os01g74500.1	1.87e-9	🟦
LOC_Os02g34850.1	3.69e-48	🟥 🟩 🟦
LOC_Os02g39800.1	1.54e-53	🟥 🟩 🟦
LOC_Os02g40770.2	4.31e-36	🟥 🟦
LOC_Os02g47900.1	1.97e-56	🟥 🟩 🟦
LOC_Os02g49326.1	4.13e-9	🟦
LOC_Os03g07260.1	4.94e-16	🟦
LOC_Os03g19480.1	3.69e-38	🟥 🟩 🟦
LOC_Os03g20430.1	7.80e-24	🟥
LOC_Os03g49730.2	4.92e-10	🟦
LOC_Os04g34610.1	1.92e-10	🟦
LOC_Os04g34976.2	4.50e-51	🟥 🟩 🟦

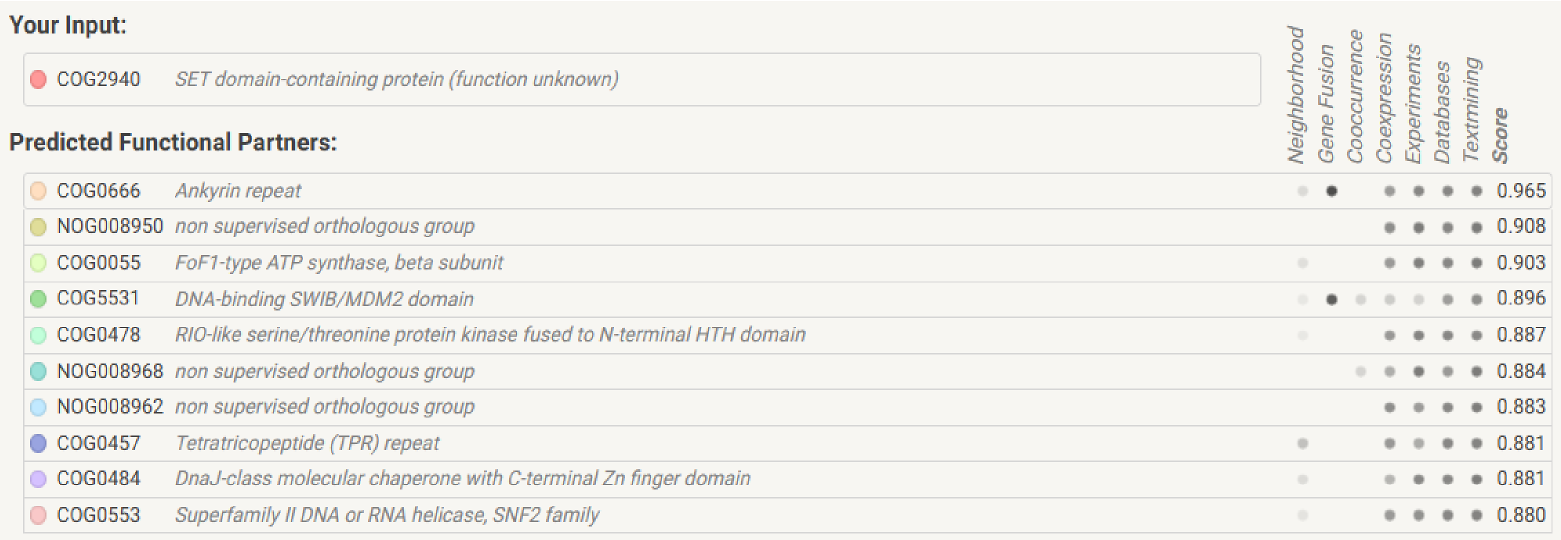
Motif	Symbol	Motif Consensus
1.	🟥	KTGNKGWGLRALEPIRAGSFVCEYVGEVJ
2.	🟦	HIMIFAKRDIPPGEELTYDYG
3.	🟩	LFKLSLEYVIDATRKGNVARFINHSCSPN

对家族蛋白的蛋白序列进行保守Motif分析，便于开展后于酶活性位点分析工作。

Tools：MEME

3.5 基于结构域互作预测





Your Input:

🟥 COG2940 *SET domain-containing protein (function unknown)*

Predicted Functional Partners:

			Neighborhood	*Gene Fusion*	*Cooccurrence*	*Coexpression*	*Experiments*	*Databases*	*Textmining*	***Score***
●	COG0666	*Ankyrin repeat*	●	●		●	●	●	●	0.965
●	NOG008950	*non supervised orthologous group*				●	●	●	●	0.908
●	COG0055	*FoF1-type ATP synthase, beta subunit*	●			●	●	●	●	0.903
●	COG5531	*DNA-binding SWIB/MDM2 domain*	●	●	●	●	●	●	●	0.896
●	COG0478	*RIO-like serine/threonine protein kinase fused to N-terminal HTH domain*	●			●	●	●	●	0.887
●	NOG008968	*non supervised orthologous group*			●	●	●	●	●	0.884
●	NOG008962	*non supervised orthologous group*				●	●	●	●	0.883
●	COG0457	*Tetratricopeptide (TPR) repeat*	●			●	●	●	●	0.881
●	COG0484	*DnaJ-class molecular chaperone with C-terminal Zn finger domain*	●			●	●	●	●	0.881
●	COG0553	*Superfamily II DNA or RNA helicase, SNF2 family*	●			●	●	●	●	0.880

根据Protein Family by Name（COG2940）进行结构域互作预测。

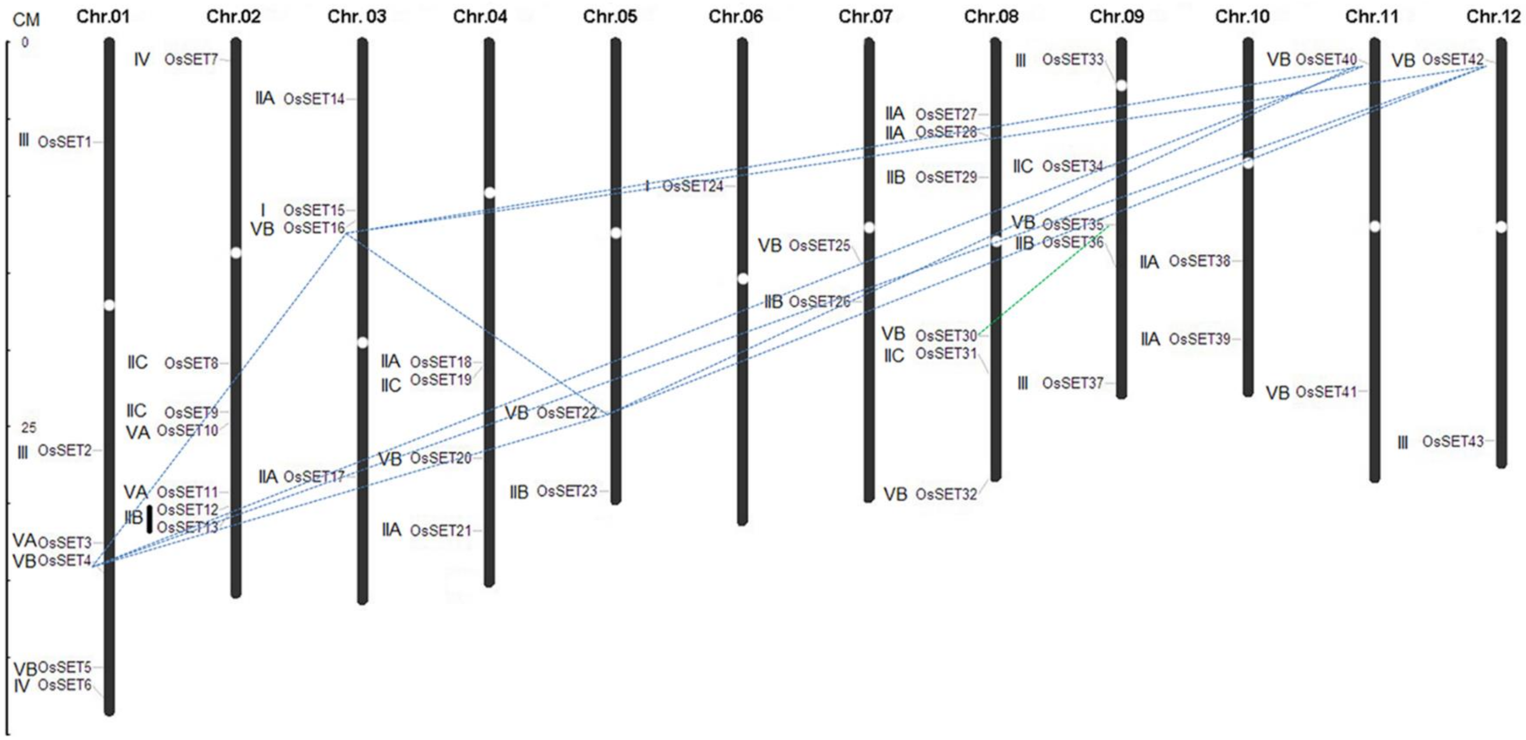
Tools：STRING

3.6 基因组染色体分布

Genome-Wide Identification, Phylogenetic and Co-Expression Analysis of *OsSET* Gene Family in Rice

Zhanhua Lu, Xiaolong Huang, Yidan Ouyang, Jialing Yao

Published: June 7, 2013 • https://doi.org/10.1371/journal.pone.0065426

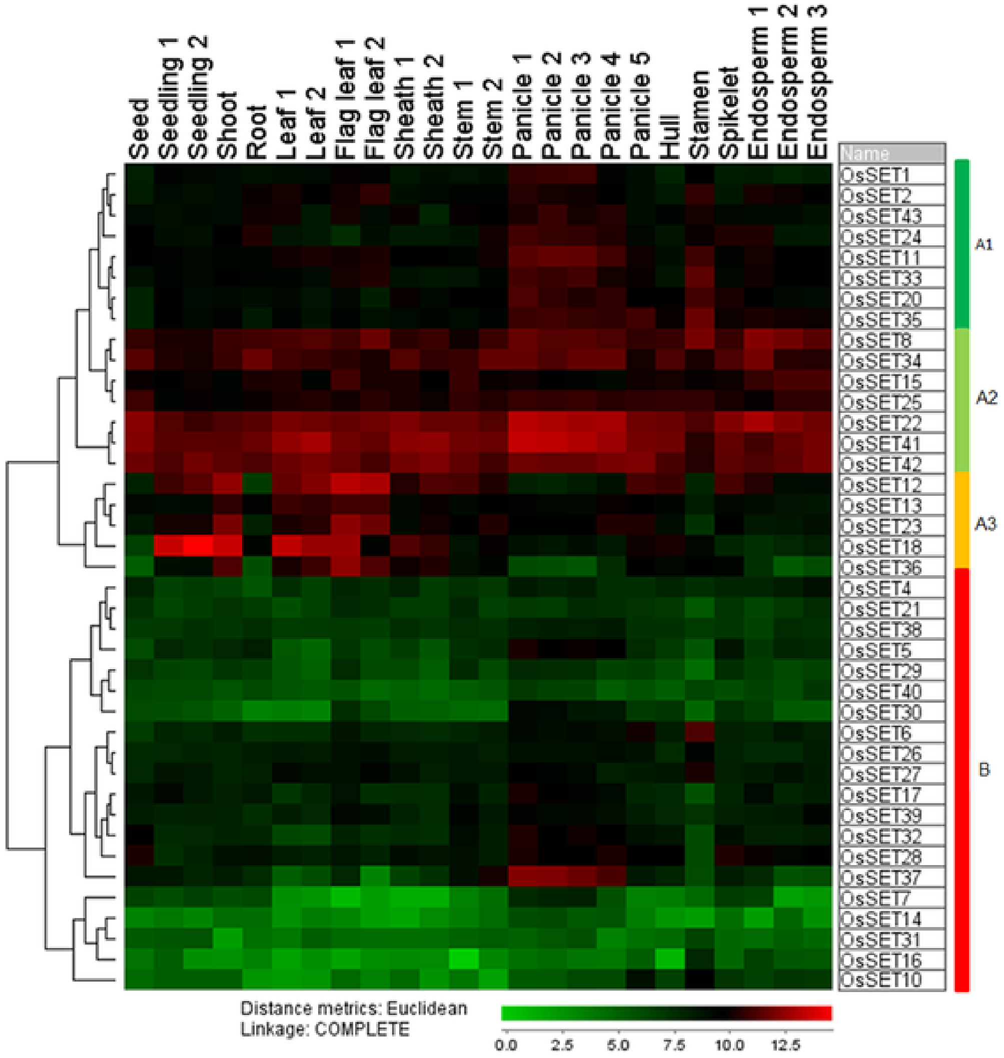


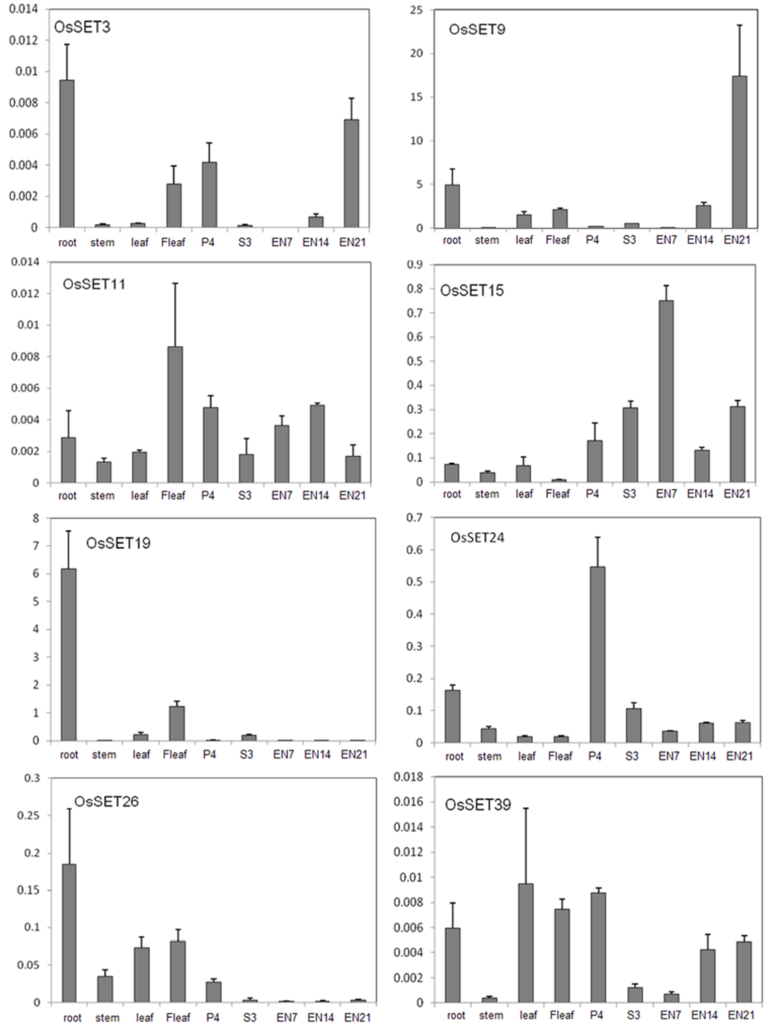
OsSET基因均未由串联重复产生。

水稻中OsSET基因家族的多样性主要源于片段复制事件。

Data Source：Lu Z *et al.*. Genome-wide identification, phylogenetic and co-expression analysis of OsSET gene family in rice. *PLoS One*. 2013;8(6):e65426. Published 2013 Jun 7. doi:10.1371/journal.pone.0065426

3.7 基因表达分析

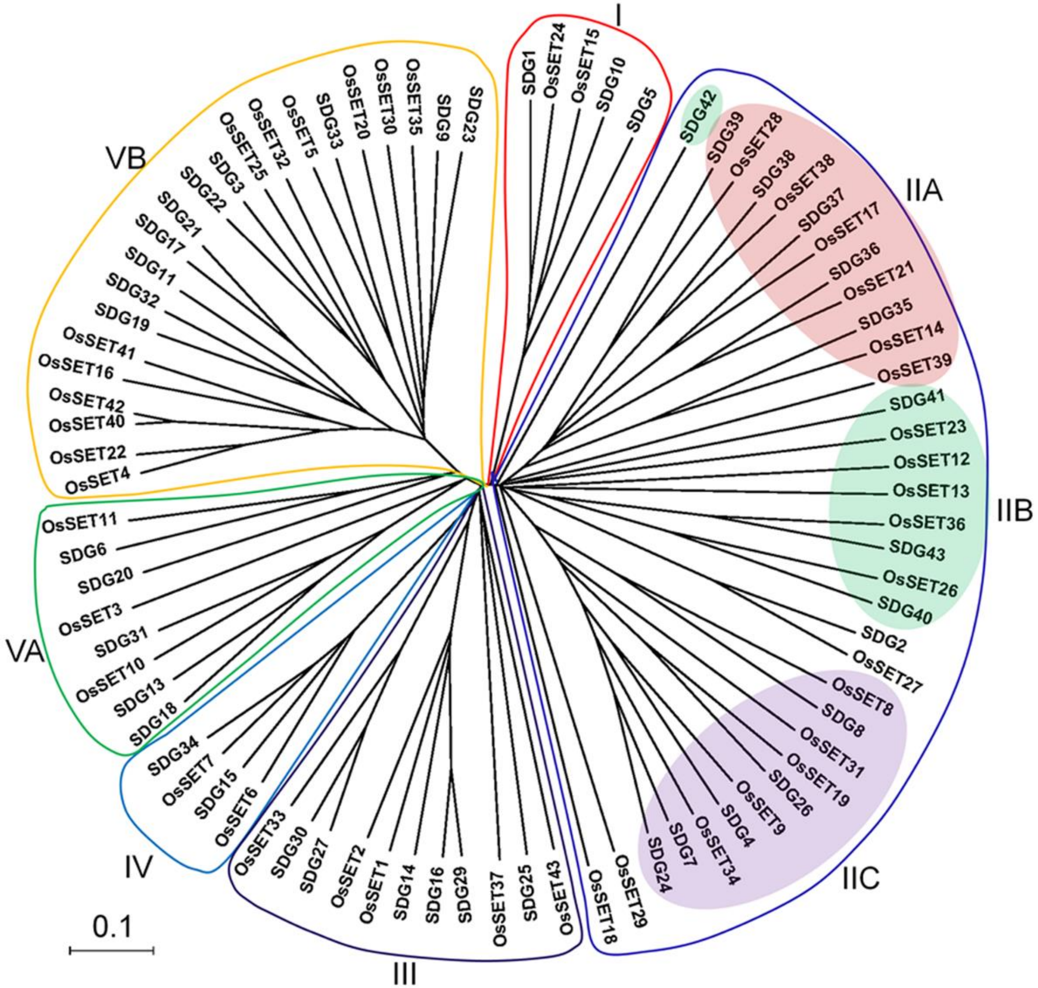


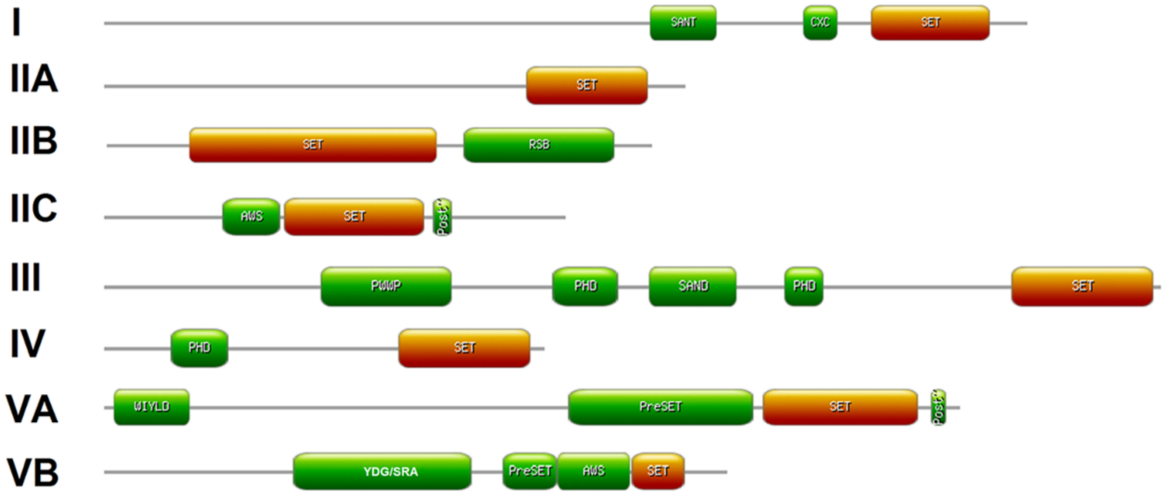


不同OsSET基因的表达具有组织特异性，也预示着该基因家族参与多种生命进程。

Data Source：Lu Z *et al.*. Genome-wide identification, phylogenetic and co-expression analysis of OsSET gene family in rice. *PLoS One*. 2013;8(6):e65426. Published 2013 Jun 7. doi:10.1371/journal.pone.0065426

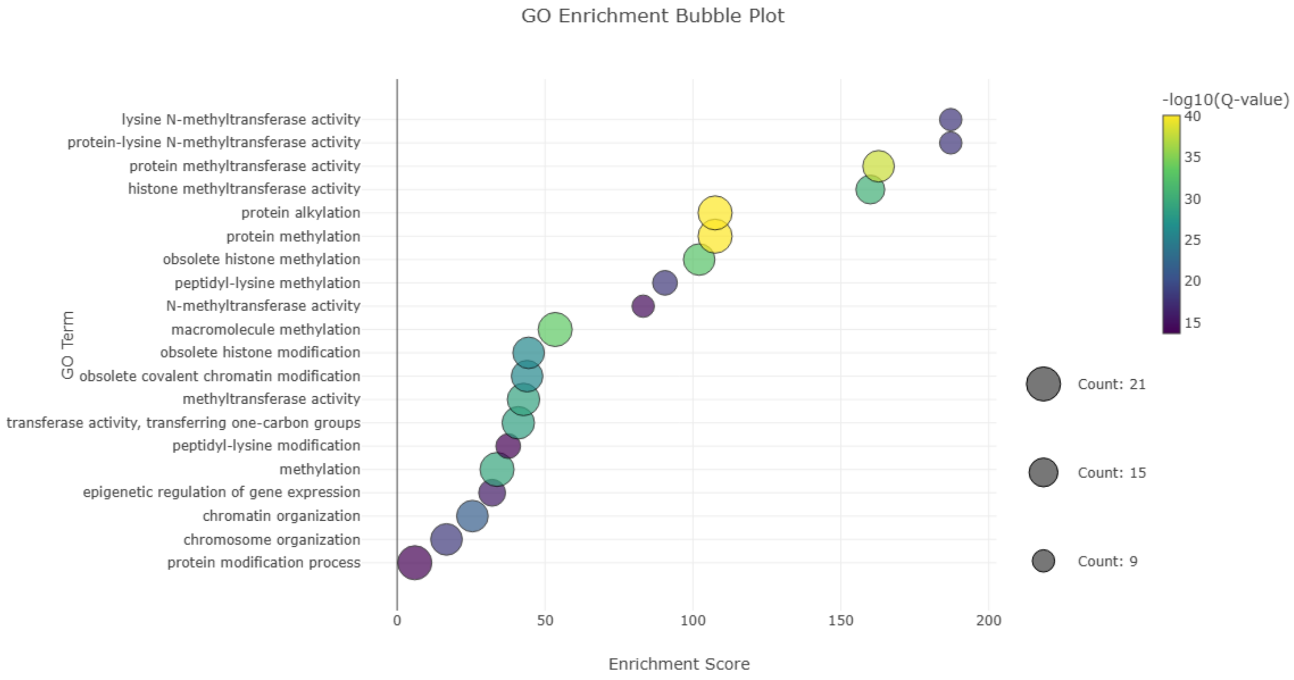
3.7 基因家族进化分析

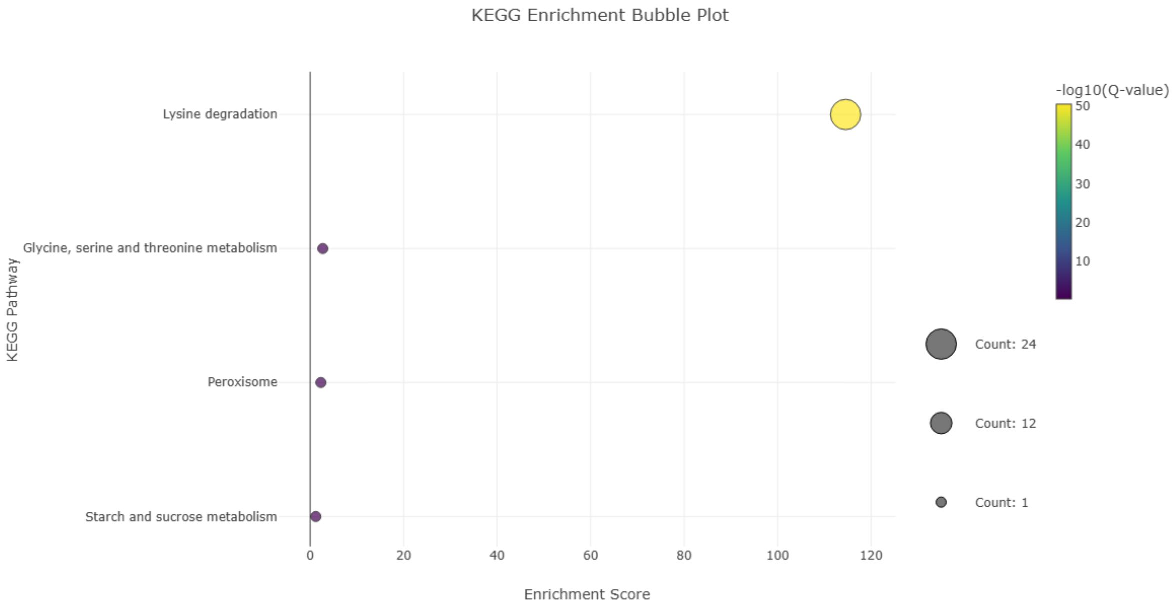




Data Source：Lu Z *et al*.. Genome-wide identification, phylogenetic and co-expression analysis of OsSET gene family in rice. *PLoS One*. 2013;8(6):e65426. Published 2013 Jun 7. doi:10.1371/journal.pone.0065426

3.8 家族蛋白GO和KEGG富集





家族蛋白具有组蛋白甲基转移酶活性并参与参与赖氨酸的蛋白修饰过程（GO），也参与赖氨酸的代谢降解过程（KEGG）。

小结与展望

小结:

1.已完成家族成员系统鉴定：通过结构域HMM 模型，并结合 RGAP、EnsemblPlants 等植物专用数据库比对，最终确认水稻含 SET 结构域基因家族共有 42 个成员，明确了其物种分布与序列特征。

2.初步多维功能预测分析：利用 TBtools、等网站工具，系统分析了家族成员的基础理化性质等基础信息；结合已有文献数据， GO 和 KEGG 富集分析，预测该家族基因可能表观遗传调控水稻抽穗期。

下一步工作计划:

1.功能预测工作补充：补充启动子顺式调控元件分析，PPI互作网络，亚家族功能机制细致划分等工作，构建抽穗期调控通路网络。

2. 家族蛋白结构分析：预测保守结构域的蛋白结构预测，观察家族蛋白结构的异同。

3.环境适应性与应用研究：分析不同环境条件下家族基因的表达响应模式，探究在环境适应性中的作用；

4.跨物种比较与进化研究：扩展研究范围至其他禾本科作物（如小麦、玉米）的 SET 结构域基因家族，通过跨物种比较分析，揭示该基因家族的进化规律与功能保守性，为作物遗传改良提供更广泛的理论参考。

小组成员

G07A： 杨启禹　课题提供，材料补充，PPT制作

G07B： 曹新宇　基因家族成员鉴定

G07C： 李凤姣　基因家族功能预测分析

G07D： 李容博　基因家族功能预测分析

中国农业科学院
CHINESE ACADEMY OF AGRICULTURAL SCIENCES

THANKS FOR LISTENING!