XXX基因组项目 基因组注释分析报告

联系方式

项目名称：XXX基因组项目

项目编号：XXX

分析人员：韩露

审核人员：杨龙、刘潇潇

报告日期：202X年XX月XX日

报告单位：西安浩瑞基因技术有限公司

1 项目基本信息

指标	值
物种名称	XXX(*XXX*, XXX)
基因组大小和N50	2.83Gb (N50: 275.08Mb)
重复序列占比	86.13%
同源蛋白来源	A, B, C, E, H, L, Ls, Tk, Tm
预测基因数目	114,081（27,748个蛋白编码基因+86,333个ncRNA基因）
功能注释占比	97.42%

2 分析流程

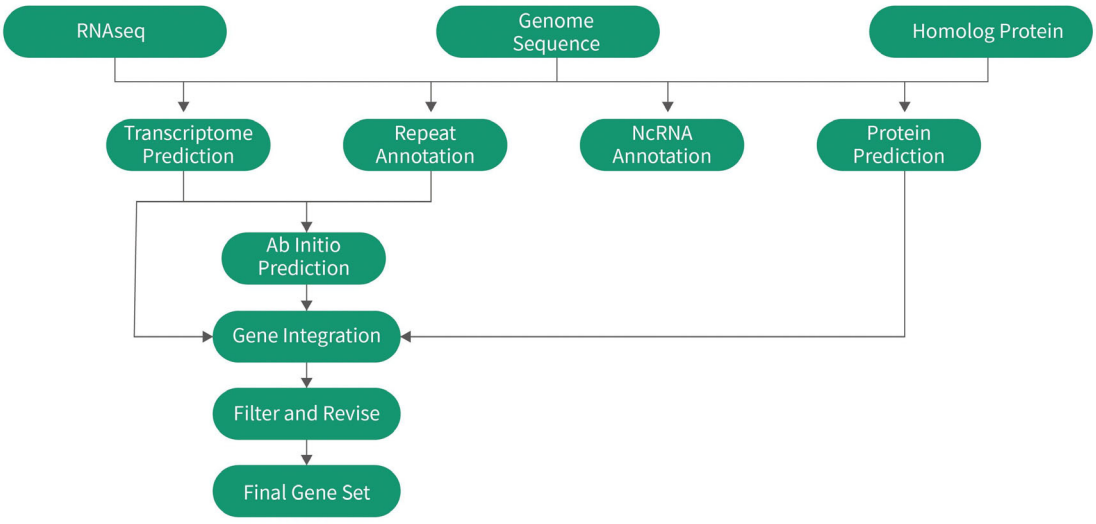

图2-1 基因组注释分析流程图

3 分析结果

3.1 基因组结构注释

3.1.1 基因组重复序列预测

3.1.1.1 串联重复分析

Krait[1]软件能识别出序列中的微卫星位点。采用该软件分析基因组中SSR序列，显示基因组中SSR序列有716,848个，总长度为19,332,462bp，占基因组总长度的比例为0.68%。按motif k-mer具体统计如下表：

表3.1-1 SSR k-mer分布统计

Motif(-mer)	Number	Length(bp)	Percentage(%)
2	463,672	13,508,948	64.68
3	203,400	4,843,209	28.37
4	38,959	698,148	5.43
5	4,889	102,565	0.68
6	5,928	179,592	0.83
Total	716,848	19,332,462	100.00

> 注：Motif(-mer)：SSR重复单元长度；Number：对应重复单元长度的SSR数量；Length(bp)：对应重复单元长度的SSR总长，以bp计；Percentage(%)：对应重复单元长度的SSR数量占比。

占比Top15的成对Motif统计如下表：

表3.1-2 Top15 Paired-Motif统计

PairedMotif	Number	Length(bp)	Percentage(%)
AT/AT	168,338	5,871,190	23.48
TA/TA	125,388	4,170,100	17.49
ATA/TAT	60,060	1,475,160	8.38
TAA/TTA	57,222	1,454,004	7.98
AC/GT	67,351	1,419,120	9.40
AAT/ATT	47,192	1,187,715	6.58
CA/TG	48,751	991,428	6.80
GA/TC	27,396	536,134	3.82
AG/CT	26,314	518,994	3.67
AAAT/ATTT	8,436	145,492	1.18
ATAA/TTAT	5,771	97,456	0.81
TAAA/TTTA	5,372	92,584	0.75
GAA/TTC	5,090	92,187	0.71

PairedMotif	Number	Length(bp)	Percentage(%)
ATAC/GTAT	3,510	73,812	0.49
AAG/CTT	3,846	69,087	0.54

> 注：PairedMotif：反向互补的SSR重复单元序列对；Number：对应重复单元序列对的SSR数量；Length(bp)：对应重复单元序列对的SSR总长，以bp计；Percentage(%)：对应重复单元序列对的SSR数量占比。

用TRF[2]软件以默认参数分析基因组中的串联重复，显示基因组中的串联重复序列有1,278,693个，总长度为200,561,204bp，占基因组总长度比例为7.10%（TRF软件利用不同k-mer做预测，不同k-mer之间存在交叠，示例文件：genome.repeat.TRF.gff3）。

表3.1-3 TRF Top15 k-mer统计

Motif(-mer)	Number	Length(bp)	Ratio(%)
2	231,311	12,757,133	18.09
21	94,990	6,243,679	7.43
22	89,773	6,663,337	7.02
3	59,899	7,990,574	4.68
20	47,579	3,256,732	3.72
18	41,413	2,193,941	3.24
19	35,357	2,039,813	2.77
23	28,192	2,104,191	2.20
24	26,647	1,996,802	2.08
15	25,553	1,265,312	2.00
17	24,089	1,435,878	1.88
16	23,161	1,272,329	1.81
28	19,477	3,171,378	1.52
14	18,888	1,338,674	1.48
12	17,898	986,647	1.40
Other	494,466	145,844,784	38.67
Total	1,278,693	200,561,204	100.00

> 注：Motif(-mer)：串联重复单元长度；Number：对应重复单元长度的串联重复数量；Length(bp)：对应重复单元长度的重复序列总长，以bp计；Ratio(%)：对应重复单元长度的串联重复数量占比。

3.1.1.2 散在重复分析

利用软件RepeatMasker[3]根据最终构建好的重复序列数据库，对该物种进行转座元件（Transposable Element，TE）预测，结果显示预测得到的TE重复序列比例为85.73%, 整合TR及其它重复序列总共2,434,041,624bp, 比例为86.13%。统计结果见下表。

表3.1-4 TE重复序列的统计结果

Class	?	Repbase/Dfam TEs	?.2	de novo TEs	?.3	Combined TEs
Type	Length(bp)	% in genome	Length(bp)	% in genome	Length(bp)	% in genome
DNA	40,743,146	1.44	132,396,328	4.68	164,696,631	5.83
LINE	19,595,982	0.69	52,838,211	1.87	54,076,726	1.91
SINE	63,183	0.00	213,197	0.01	276,380	0.01
LTR	496,429,205	17.57	1,860,266,696	65.83	1,869,734,429	66.16
Unknown	27,640	0.00	808,352,052	28.60	808,363,506	28.60
Other	12,731,022	0.45	47,812,860	1.69	56,325,843	1.99
Total	559,483,745	19.80	2,415,433,198	85.47	2,422,630,475	85.73

> 注：Repbase/Dfam TEs：RepeatMasker同源注释获得的TE长度和基因组占比；de novo TEs：RepeatModeler从头预测获得的TE长度和基因组占比；Combined TEs：整合后的TE长度和基因组占比。LTR：携带长末端重复（Long Terminal Repeat，LTR）的一类TE；SINE：短散布重复序列（Short INterspersed Elements，SINE），非LTR一类TE；LINE：长散布重复序列（Long INterspersed Elements，LINE），非LTR一类TE；DNA：DNA转座子，二类TE；表中Other是指注释分类出来但不属于以上各类的序列，Unknown是指注释无法分类的序列。

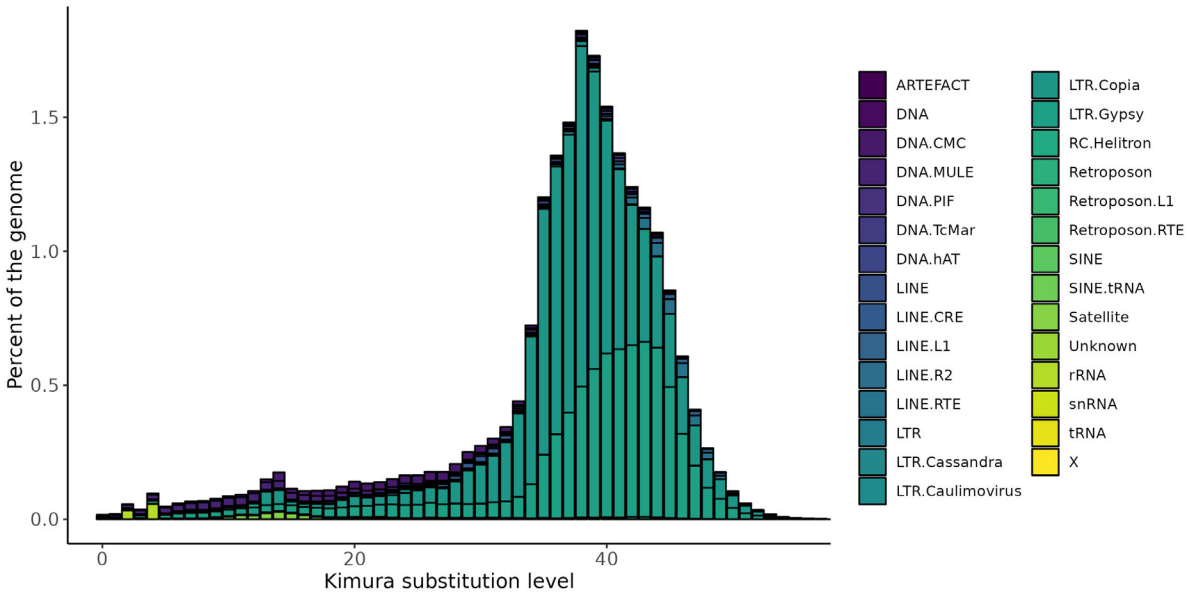

图3.1-1 转座子分化率

> 注：横坐标是基因组中注释到的TE序列与总的库文件中相应序列的分歧度，纵坐标是该分歧度下的TE序列在基因组中所占的百分比；不同的TE以不同的颜色加以标示。

3.1.2 基因结构预测

基因结构预测主要采用转录组预测，同源蛋白预测以及从头预测。

3.1.2.1 转录组预测

二代RNA-Seq转录组测序数据下机后，使用fastp[4]进行过滤和评估，质控结果如下所示：

表3.1-5 Clean data数据产量统计

Sample id	Total reads	Total bases	Clean reads	Clean bases	Q20 rate(%)	Q30 rate(%)	GC(%)
XXX-2-J	42,102,380	6,315,357,000	42,084,402	6,228,839,198	98.62	95.27	43.15
XXX-2-Y	39,425,200	5,913,780,000	39,414,822	5,845,136,592	98.61	95.19	43.52

> 注：Sample id：转录组测序样本编号；Total reads：原始的reads总数；Total bases：原始的碱基总数；Clean reads：过滤后的reads总数；Clean bases：过滤后的碱基总数；Q20 rate(%)：测序质量值大于20的碱基占比；Q30 rate(%)：测序质量值大于30的碱基占比；GC(%)：过滤后的转录组数据的GC含量。

RNA-Seq测序样本XXX-2-J的碱基质量分布图和碱基含量分布图如下所示：

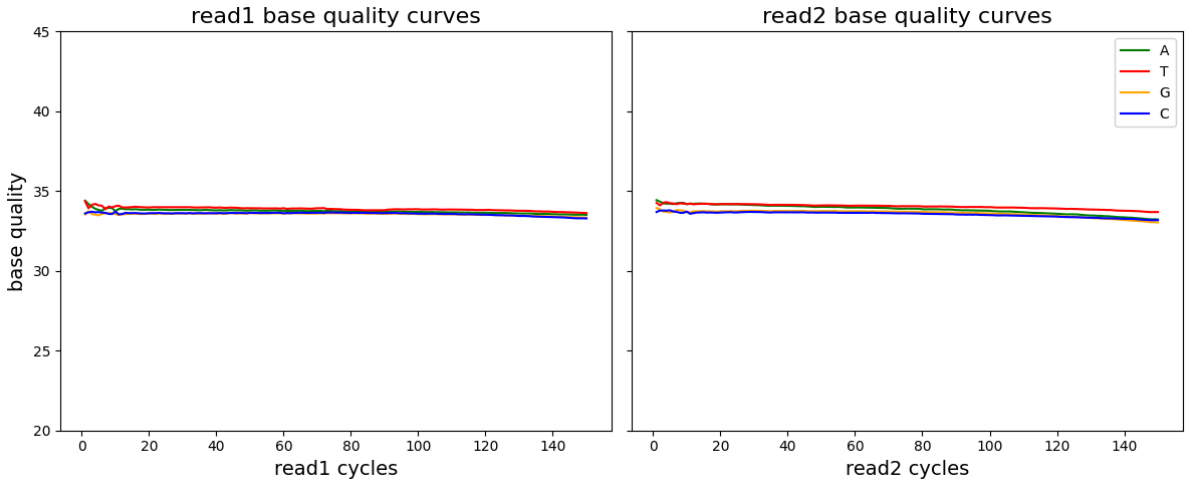

图3.1-2 碱基质量分布图

> 注：横坐标为reads的位置，纵坐标为测序质量值，不同的碱基用不用颜色标识。

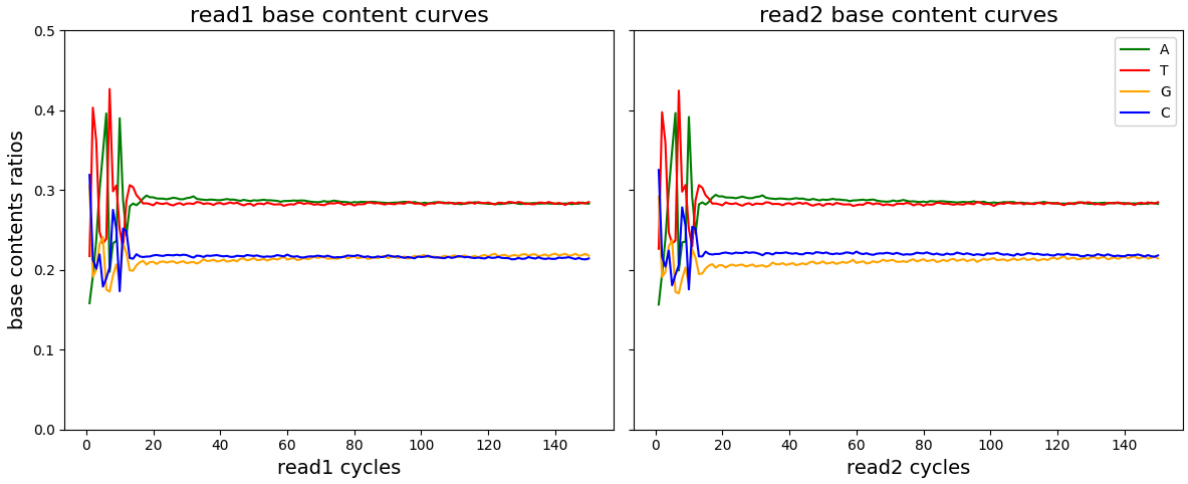

图3.1-3 碱基含量分布图

> 注：横坐标为reads的位置，纵坐标为不同碱基占比，不同的碱基用不用颜色标识，由于二代测序的本身特性，前十几个bp碱基含量会有波动。从碱基含量分布图中可以看出，在十几个bp以后，A与T、G与C含量基本一致，数据碱基含量合格。

其余样本见目录：./src/summary/3_gene_structure/RNA_seq/cleandata/。

数据质控后，我们使用HISAT2[5]软件将clean data比对到参考基因组，如果参考基因组选择合适并且相关实验不存在污染的情况下，实验所产生的测序序列的定位的百分比正常情况下会高于70%，其中具有多个定位的测序序列占总体的百分比通常不会超过20%。与参考基因组比对结果如下：

表3.1-6 与基因组比对结果统计

Sample ID	Total Valid	Total Mapped	Uniquely Mapped	Multiple Mapped
XXX-2-J	42,084,402	41,610,092(98.87%)	39,988,506(96.10%)	1,621,586(3.90%)
XXX-2-Y	39,414,822	39,002,014(98.95%)	37,227,476(95.45%)	1,774,538(4.55%)

> 注：表格各列说明如下表：

列名	说明
Sample ID	转录组样本编号
Total Valid	clean reads总数
Total Mapped	比对到基因组上的clean reads总数和占比
Uniquely Mapped	比对到基因组上唯一位置的clean reads总数和占比
Multiple Mapped	比对到基因组上多个位置的clean reads总数和占比

基于比对基因组的结果，我们使用StringTie[6]软件分别对转录本进行组装，然后合并各个样本的转录本。使用StringTie软件进行基因组的基因结构预测，共预测得到21,628个基因，并得到相应的训练模型用于从头预测。

表3.1-7 基于转录组数据预测蛋白编码基因注释统计

Gene Set	Number of Genes	Average Gene Length(bp)	Average CDS Length(bp)	Average Exons per Gene	Average Exon Length(bp)	Average Intron Length(bp)
StringTie	21,628	5,304.22	1,893.38	5.84	324.25	657.55

> 注：表格各列说明如下表：

列名	说明
Gene Set	基因集来源
Number of Genes	基因总数
Average Gene Length	基因平均长度
Average CDS Length(bp)	CDS平均长度，以bp计
Average Exons per Gene	每个基因平均外显子数
Average Exon Length(bp)	平均外显子长度，以bp计
Average Intron Length(bp)	平均内含子，以bp计

3.1.2.2 同源蛋白预测

基于近缘物种蛋白序列信息，通过GeneMark-ETP[7]（如果存在二代转录组序列，该软件会同时利用转录组数据进行注释），利用ProtHint[8]或Spaln[9]将相应蛋白信息与基因组进行比对，而后整合所有同源物种预测结果来得到对应预测基因的结构信息，结果预测得到22,569个基因。用于注释的近缘物种信息及预测结果分别见下面两个表。

表3.1-8 用于同源蛋白注释的近源物种信息及数据来源

Species	Latin Name	GFF File	pep File
Cc	*C*	-	-
Cn	*C*	-	-
Ct	*C*	-	-
Ec	*E*	-	-
Ha	*H*	-	-
Ls	*L*	-	-
Ls	*L*	-	-
Tk	*T*	-	-
Tm	*T*	-	-

> 注：Species：物种编号；Latin Name：物种拉丁学名；GFF File：基因组注释文件下载链接；pep File：蛋白序列文件下载链接。

表3.1-9 同源蛋白预测结果统计

Gene Set	Number of Genes	Average Gene Length(bp)	Average CDS Length(bp)	Average Exons per Gene	Average Exon Length(bp)	Average Intron Length(bp)
GeneMark	22,569	2,558.46	801.58	3.57	224.63	686.01

> 注：表格各列说明如下表：

列名	说明
Gene Set	基因集来源
Number of Genes	基因总数
Average Gene Length	基因平均长度
Average CDS Length(bp)	CDS平均长度，以bp计
Average Exons per Gene	每个基因平均外显子数
Average Exon Length(bp)	平均外显子长度，以bp计
Average Intron Length(bp)	平均内含子，以bp计

3.1.2.3 从头预测

基于转录组预测得到的基因，选取3,000个可靠基因通过Augustus[10]进行模型训练（具体方法见5.1.2.3）得到该物种预测模型，基于该训练模型利用Augustus进行基因结构从头预测，最终预测得到38,392个基因。选用的模型训练基因及预测结果统计见下表。

表3.1-10 基于从头预测蛋白编码基因注释统计

Gene Set	Number of Genes	Average Gene Length(bp)	Average CDS Length(bp)	Average Exons per Gene	Average Exon Length(bp)	Average Intron Length(bp)
Augustus	38,392	3,085.36	964.60	4.35	221.71	633.61

> 注：表格各列说明如下表：

列名	说明
Gene Set	基因集来源
Number of Genes	基因总数
Average Gene Length	基因平均长度
Average CDS Length(bp)	CDS平均长度，以bp计
Average Exons per Gene	每个基因平均外显子数
Average Exon Length(bp)	平均外显子长度，以bp计
Average Intron Length(bp)	平均内含子，以bp计

3.1.2.4 预测结果整合

借助整合注释分析工具BRAKER3[11]，基于StringTie的基因预测结果，GeneMark-ETP整合后的基因预测结果以及从头预测的基因预测结果，使用TSEBRA[12]将上述预测结果基于一定权重值（权重值设置的默认标准为：StringTie预测 >= GeneMark-ETP预测 >= 从头预测）进行整合，得到物种预测的基因集。

表3.1-11 蛋白编码基因注释统计

Gene Set	Number of Genes	Average Gene Length(bp)	Average CDS Length(bp)	Average Exons per Gene	Average Exon Length(bp)	Average Intron Length(bp)
XXX	27,748	5,007.13	1,175.83	5.35	219.76	589.04

> 注：表格各列说明如下表：

列名	说明
Gene Set	基因集来源
Number of Genes	基因总数
Average Gene Length	基因平均长度
Average CDS Length(bp)	CDS平均长度，以bp计
Average Exons per Gene	每个基因平均外显子数
Average Exon Length(bp)	平均外显子长度，以bp计
Average Intron Length(bp)	平均内含子，以bp计

如上表所示，基因组共预测得到基因27,748个，平均基因长度为5,007.13bp，平均CDS长度为1,175.83bp，每个基因中平均含有5.35个外显子，平均外显子长度为219.76bp，平均内含子长度为589.04bp。

3.1.3 ncRNA注释

基于已组装好的基因组序列，使用软件Infernal[13]与Rfam[14]数据库比对来预测基因组ncRNA，同时利用tRNAscan-SE[15]预测tRNA并利用RNAmmer[16]或Barrnap[17]构建模型预测rRNA及其各类亚基，上述结果经进一步整合获得该基因组中ncRNA预测结果，具体信息见下表。

表3.1-12 非编码RNA注释统计

Type	Number	Average Length(bp)	Total Length(bp)	Base Ratio(%)
28S rRNA	1,674	4,402	7,370,293	0.2608
18S rRNA	1,675	1,620	2,715,129	0.0961
5.8S rRNA	531	155	82,583	0.0029
5S rRNA	66,527	116	7,766,686	0.2748
tRNA	7,115	74	527,366	0.0187
miRNA	184	132	24,332	0.0009
snoRNA	3,688	105	390,832	0.0138
catalytic RNA	1,188	58	68,961	0.0024
spliceosomal RNA	219	143	31,382	0.0011
other sRNA	1	164	164	0.0000
cis regulator	44	48	2,149	0.0001
other ncRNA	3,487	101	352,280	0.0125

> 注：表格各列说明如下表：

列名	说明
Type	ncRNA类型
Number	对应类型总数
Average Length(bp)	对应类型平均长度，以bp计
Total Length(bp)	对应类型总长，以bp计
Base Ratio(%)	对应类型占该基因组的比例

如上表所示，预测共得到86,333个ncRNA基因，其中包括70,407个rRNA基因，7,115个tRNA基因，184个miRNA基因。

3.2 基因功能注释

3.2.1 NR数据库注释

基于预测得到的基因的蛋白序列，使用diamond[18]比对非冗余蛋白序列数据库NR[19]，检索得到相应基因的功能信息，并对功能注释结果中包含的物种信息进行分类。NR数据库注释结果中物种比对结果如下表（只展示部分）和下图所示。

表3.2-1 NR数据库注释统计

query_id	subject_id	annotation
g1.t1	KAI3786253.1	hypothetical protein L1987_45388 [Smallanthus sonchifolius]
g1.t2	KAI3786253.1	hypothetical protein L1987_45388 [Smallanthus sonchifolius]
g2.t1	KAI3759844.1	hypothetical protein L6452_07945 [Arctium lappa]
g2.t2	KAI3759844.1	hypothetical protein L6452_07945 [Arctium lappa]
g2.t3	KAI3759844.1	hypothetical protein L6452_07945 [Arctium lappa]
g2.t4	KAI3759844.1	hypothetical protein L6452_07945 [Arctium lappa]
g3.t1	XP_043617657.1	uncharacterized protein LOC122589427 [Erigeron canadensis]
g4.t1	KAI3759840.1	hypothetical protein L6452_07939 [Arctium lappa]
g4.t2	KAI3759840.1	hypothetical protein L6452_07939 [Arctium lappa]
g5.t1	KAD0708999.1	hypothetical protein E3N88_43824 [Mikania micrantha]
g5.t2	KAD0708999.1	hypothetical protein E3N88_43824 [Mikania micrantha]
g6.t1	XP_024969255.1	uncharacterized protein LOC112508750 isoform X6 [Cynara cardunculus var. scolymus]
g7.t1	XP_024969249.1	uncharacterized protein LOC112508750 isoform X1 [Cynara cardunculus var. scolymus]
g8.t1	XP_022001951.1	uncharacterized protein LOC110899373 isoform X2 [Helianthus annuus]
g9.t1	XP_024971026.1	uncharacterized protein LOC112510010 [Cynara cardunculus var. scolymus]
g9.t2	XP_024971026.1	uncharacterized protein LOC112510010 [Cynara cardunculus var. scolymus]
g10.t1	XP_023758964.1	uncharacterized membrane protein At1g16860 [Lactuca sativa]
g11.t1	KAI3759830.1	hypothetical protein L6452_07925 [Arctium lappa]
g11.t2	KAI3759830.1	hypothetical protein L6452_07925 [Arctium lappa]

> 注：query_id：注释的基因/蛋白的编号；subject_id：数据库中蛋白的编号；annotation：注释信息描述。

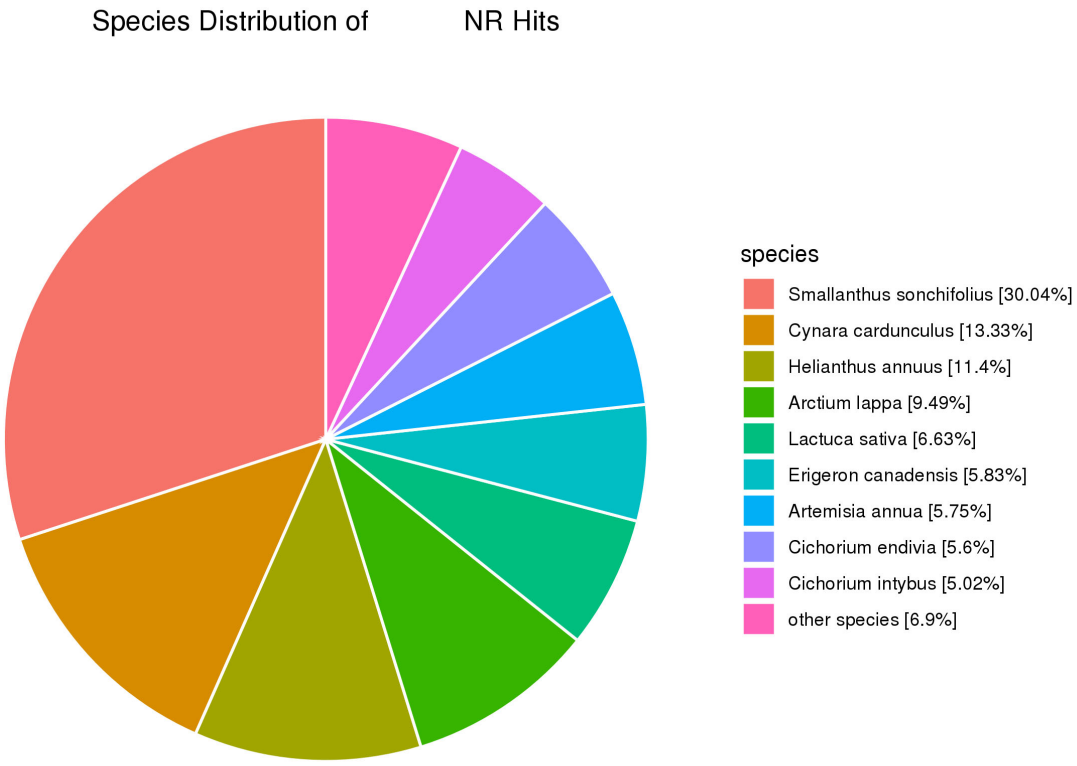

图3.2-1 NR库注释结果对应物种分类图

3.2.2 KEGG数据库注释

基于预测得到的基因的蛋白序列，通过比对京都基因与基因组百科全书数据库（Kyoto Encyclopedia of Gene and Genomes，KEGG[20]）数据库，进一步补充注释基因的功能信息并对蛋白序列注释到的生物学通路进行统计。下图为KEGG注释结果的统计信息图。

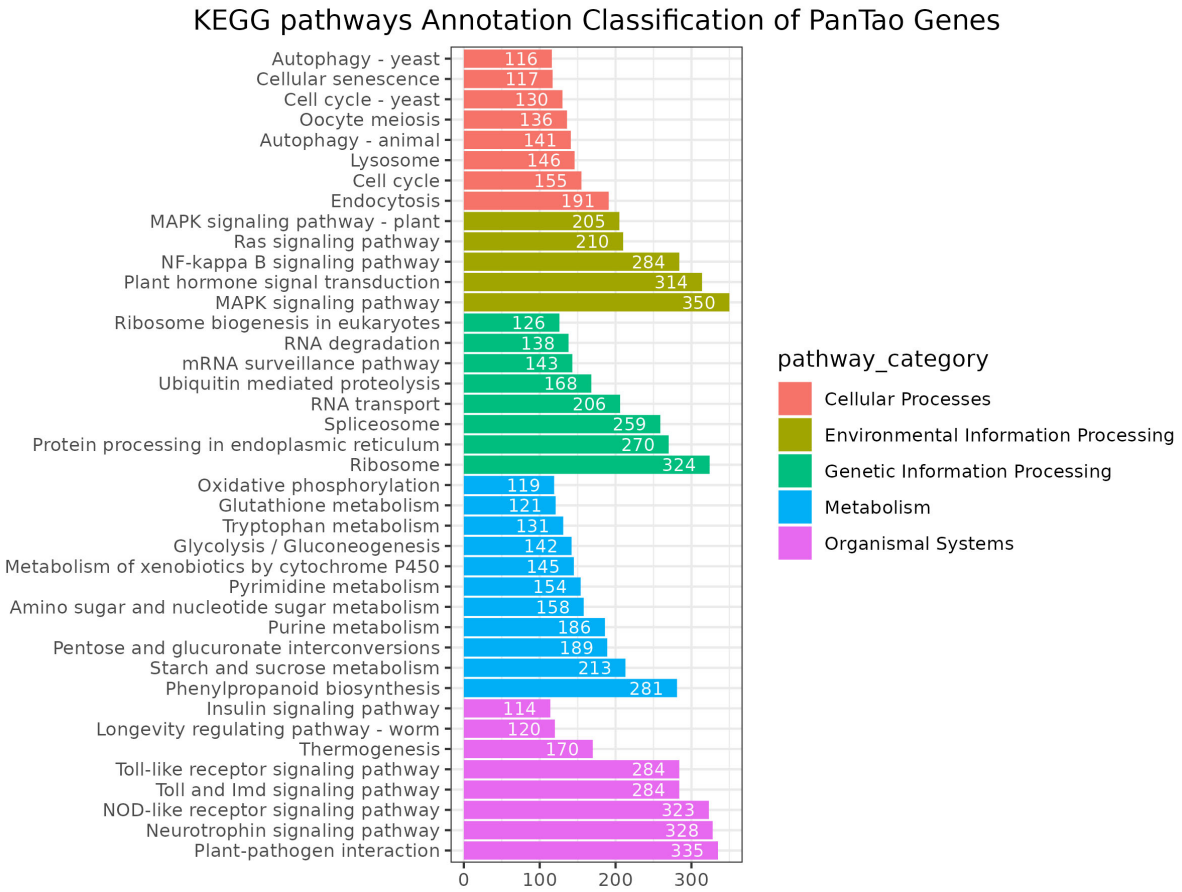

图3.2-2 KEGG数据库注释结果分类图

3.2.3 KOG数据库注释

基于预测得到的基因的蛋白序列，使用EggNOG-mapper[21]比对真核同源蛋白簇数据库（Eukaryotic Orthologous Groups of protein，KOG[22]）数据库，根据蛋白序列注释到的KOG信息，推测该序列的功能。如下为KOG注释结果统计图。

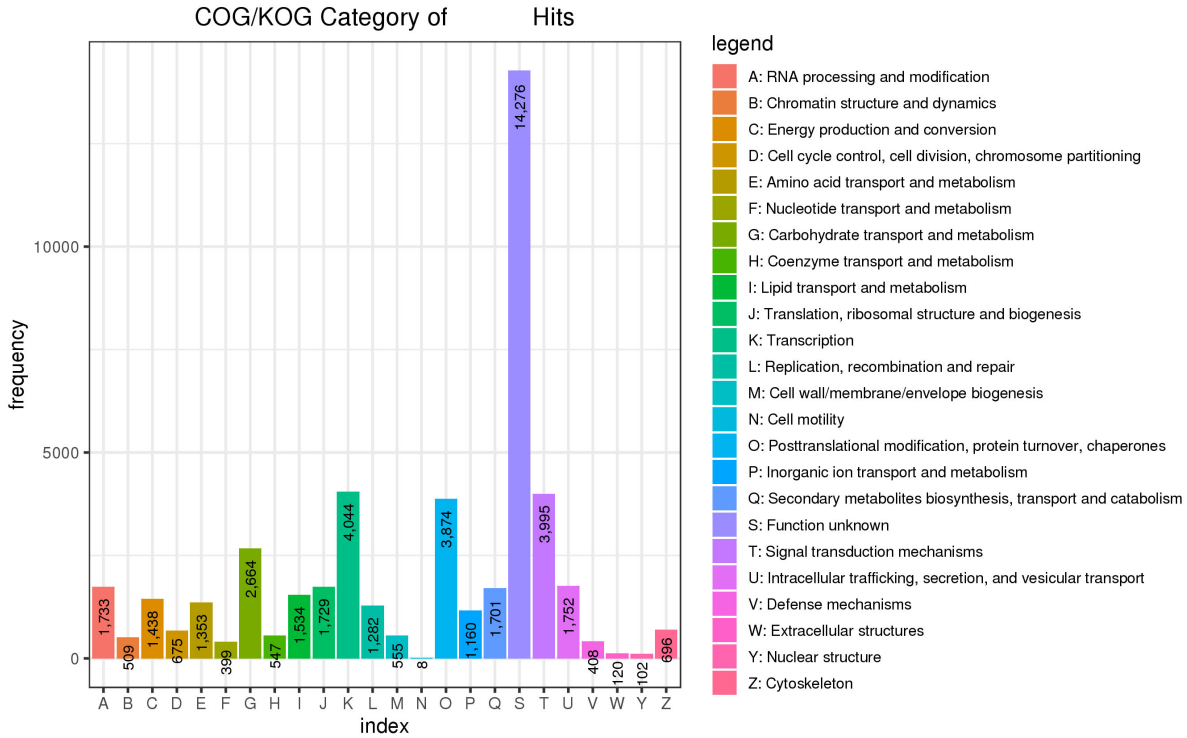

图3.2-3 KOG数据库注释结果分类图

3.2.4 GO数据库注释

基于预测得到的基因的蛋白序列，使用EggNOG-mapper软件进行GO[23]功能注释。如下图为GO注释结果的统计信息。

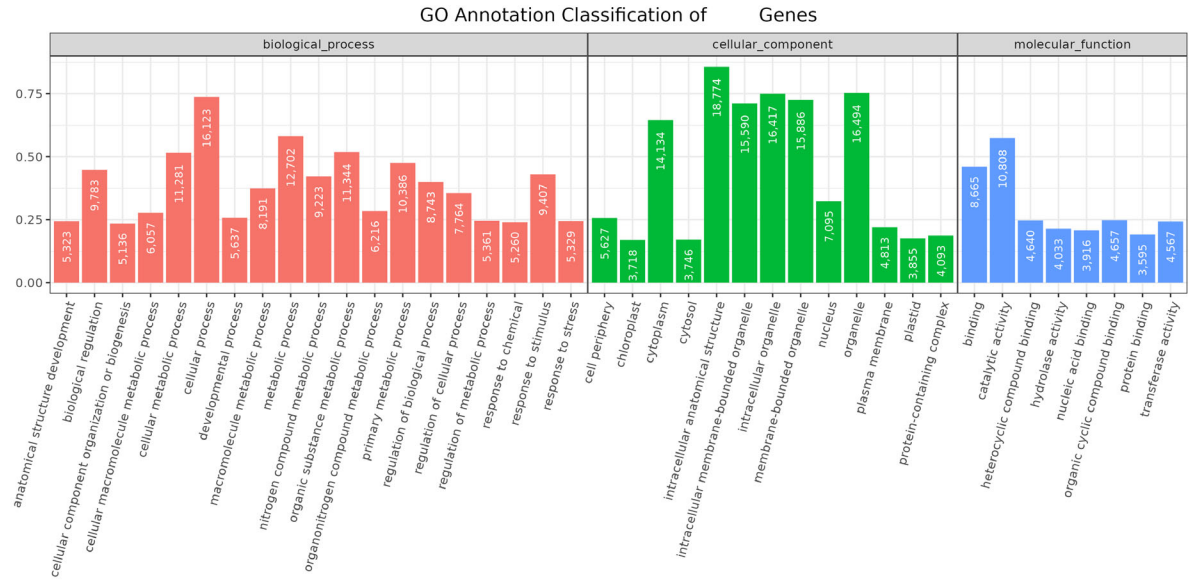

图3.2-4 GO数据库注释结果分类图

3.2.5 Swiss-Prot和TrEMBL数据库注释

基于预测得到的基因的蛋白序列，比对Swiss-Prot[24]和TrEMBL[25]数据库，并统计基因组蛋白序列注释到的数据库蛋白信息，结果如下（只展示部分）。

表3.2-2 Swiss-Prot数据库注释统计

query_id	subject_id	annotation
g1.t1	sp_Q8W4D6_HC173_ARATH	Protein HIGH CHLOROPHYLL FLUORESCENCE PHENOTYPE 173, chloroplastic OS=Arabidopsis thaliana (Mouse-ear cress) OX=3702 GN=F19K19.14 PE=1 SV=1
g1.t2	sp_Q8W4D6_HC173_ARATH	Protein HIGH
g5.t1	sp_Q9SGU2_SAU71_ARATH	Auxin-responsive protein SAUR71 OS=Arabidopsis thaliana (Mouse-ear cress) OX=3702 GN=F14G9.23 PE=2 SV=1
g5.t2	sp_Q9SGU2_SAU71_ARATH	Auxin-responsive protein
g7.t1	sp_B6SFA4_MAA3_ARATH	Probable helicase MAGATAMA 3 OS=Arabidopsis thaliana (Mouse-ear cress) OX=3702 GN=MAA3 PE=2 SV=1
g10.t1	sp_Q9FZ45_Y1686_ARATH	Uncharacterized membrane protein At1g16860 OS=Arabidopsis thaliana (Mouse-ear cress) OX=3702 GN=At1g16860 PE=1 SV=1
g11.t1	sp_Q9LK31_Y3272_ARATH	Kelch repeat-containing protein At3g27220 OS=Arabidopsis thaliana (Mouse-ear cress) OX=3702 GN=At3g27220 PE=2 SV=1

query_id	subject_id	annotation
g11.t2	sp_Q9LK31_Y3272_ARATH	Kelch repeat-containing protein At3g27220 OS=Arabidopsis thaliana (Mouse-ear cress) OX=3702 GN=At3g27220 PE=2 SV=1
g12.t1	sp_Q9C8K7_FBK21_ARATH	F-box/kelch-repeat protein At1g51550 OS=Arabidopsis thaliana (Mouse-ear cress) OX=3702 GN=At1g51550 PE=2 SV=1
g12.t2	sp_Q9C8K7_FBK21_ARATH	F-box/kelch-repeat protein At1g51550 OS=Arabidopsis thaliana (Mouse-ear cress) OX=3702 GN=At1g51550 PE=2 SV=1
g13.t1	sp_Q9C512_MNS1_ARATH	Mannosyl-oligosaccharide 1,2-alpha-mannosidase MNS1 OS=Arabidopsis thaliana (Mouse-ear cress) OX=3702 GN=F19C24.18, PE=1 SV=1
g14.t1	sp_Q9LU41_ACA9_ARATH	Calcium-transporting ATPase 9, plasma membrane-type OS=Arabidopsis thaliana (Mouse-ear cress) OX=3702 GN=ACA9 PE=2 SV=2
g14.t2	sp_Q9LU41_ACA9_ARATH	Calcium-transporting ATPase 9, plasma membrane-type OS=Arabidopsis thaliana (Mouse-ear cress) OX=3702 GN=ACA9 PE=2 SV=2
g15.t1	sp_Q0WPA5_MSR2_ARATH	Protein MANNAN SYNTHESIS-RELATED 2 OS=Arabidopsis thaliana (Mouse-ear cress) OX=3702 GN=F19C24.14 PE=2 SV=1
g15.t2	sp_Q0WPA5_MSR2_ARATH	Protein MANNAN SYNTHESIS-RELATED 2 OS=Arabidopsis thaliana (Mouse-ear cress) OX=3702 GN=F19C24.14 PE=2 SV=1
g16.t1	sp_Q9LMJ4_E70B2_ARATH	Exocyst complex component EXO70B2 OS=Arabidopsis thaliana (Mouse-ear cress) OX=3702 GN=F10K1.28 PE=1 SV=1
g16.t2	sp_Q9LMJ4_E70B2_ARATH	Exocyst complex component EXO70B2 OS=Arabidopsis thaliana (Mouse-ear cress) OX=3702 GN=F10K1.28 PE=1 SV=1
g17.t1	sp_Q9LU39_GLUBP_ARATH	Glutamyl-tRNA reductase-binding protein, chloroplastic OS=Arabidopsis thaliana (Mouse-ear cress) OX=3702 GN=MXL8.5 PE=1 SV=1
g17.t2	sp_Q9LU39_GLUBP_ARATH	Glutamyl-tRNA reductase-binding protein, chloroplastic OS=Arabidopsis thaliana (Mouse-ear cress) OX=3702 GN=MXL8.5 PE=1 SV=1

> 注：query_id：注释的基因/蛋白的编号；subject_id：数据库中蛋白的编号；annotation：注释信息描述。

3.3 基因组注释结果评估

3.3.1 BUSCO评估

利用BUSCO对预测得到的基因集进行评估，即根据OrthoDB数据库[26]中进化分支eudicots的**通用单拷贝直系同源基因集**（Benchmarking Universal Single-Copy Orthologs，**BUSCOs**[27]）预测转录组现有序列的基因情况，进而评估预测基因集的完整性，详细评估结果如下表所示：

表3.3-1 BUSCO预测统计

Type	Number	Percent(%)
Complete BUSCOs (C)	2,268	97.51
Complete and single-copy BUSCOs (S)	2,153	92.56
Complete and duplicated BUSCOs (D)	115	4.94
Fragmented BUSCOs (F)	11	0.47
Missing BUSCOs (M)	47	2.02
Total BUSCO groups searched	2,326	100.00

> 注：表格第一列为BUSCO统计指标，第二列为BUSCO数目，第三列为BUSCO占比。表格各行统计指标说明如下表：

行名	说明
Complete BUSCOs (C)	序列完全比对上BUSCO；
Complete and single-copy BUSCOs (S)	一个BUSCO比对上一个基因；
Complete and duplicated BUSCOs (D)	一个BUSCO比对上多个基因；
Fragmented BUSCOs (F)	部分序列比对上BUSCO；
Missing BUSCOs (M)	未比对上BUSCO；
Total BUSCO groups searched	总BUSCO集。

基于BUSCO预估结果进行绘图得到图3.3-1。

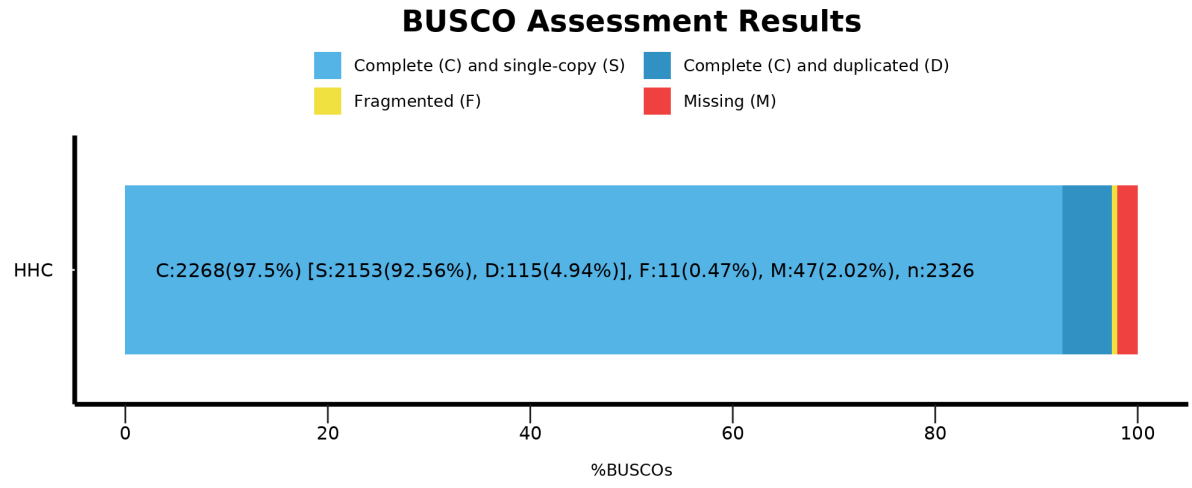

图3.3-1 BUSCO预测统计图

从图表中可以看到，注释基因集中可以找到约97.51%的完整基因元件，说明绝大部分保守基因预测比较完整，从侧面反映预测结果可信度较高。

3.3.2 基因功能注释评估

基于不同数据库得到的基因功能注释结果进行统计，结果显示能够注释到功能数据库的基因蛋白数目为49,109个。占总基因蛋白数的97.42%。具体信息见下表。

表3.3-2 基因功能注释统计

Database	Number	Ratio(%)
COG/KOG	47,181	93.60
KEGG	23,860	47.33
GO	25,424	50.44
Pfam	43,625	86.54
NR	49,080	97.36
Swiss-Prot	35,738	70.90
TrEMBL	48,995	97.19
Overall	49,109	97.42
Query	50,409	100.00

> 注：Database：数据库类；Number：注释到对应数据库的基因蛋白数；Ratio(%)：注释到的基因蛋白占所有预测基因蛋白的百分比。

基于不同数据库得到的基因功能注释结果进行Upset图交叠统计绘图，具体结果见下图。

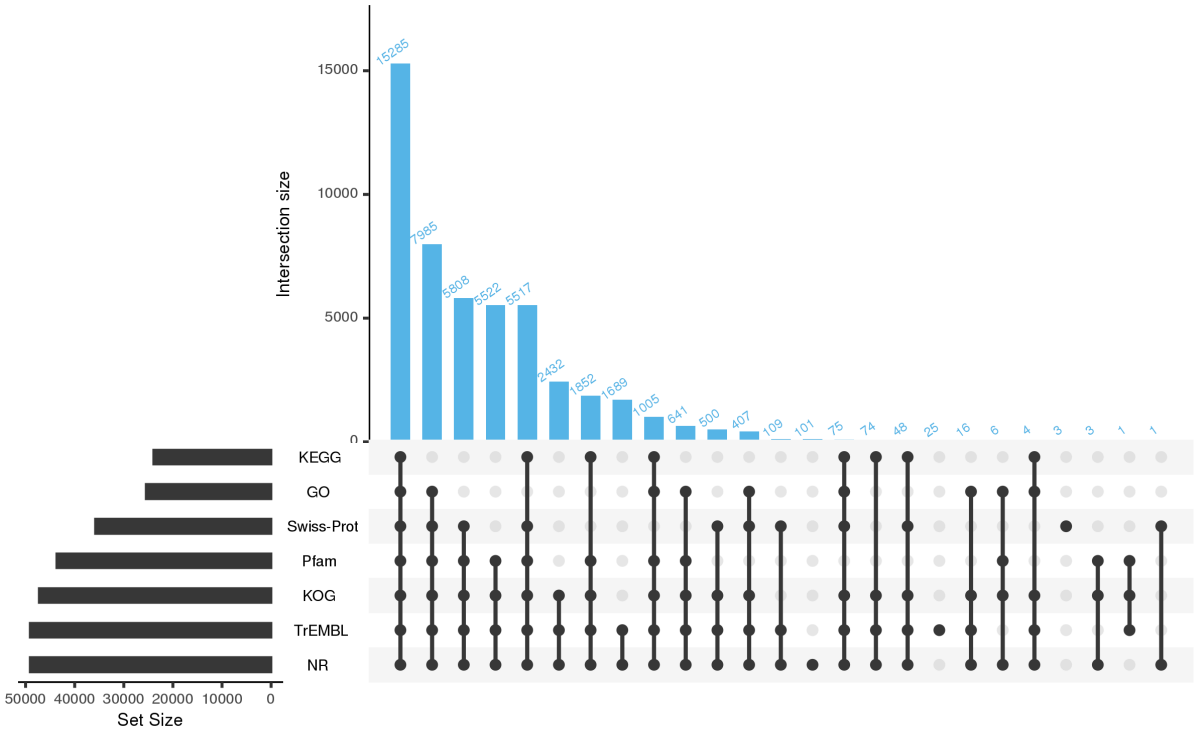

图3.3-2 各数据库注释结果Upset图

3.3.3 基因表达水平分析

基于基因结构注释得到的基因信息，我们通过定位到基因组区域或基因外显子区的测序序列(reads)的计数来估计基因的表达水平。Reads计数除了与基因的真实表达水平成正比外，还与基因的长度和测序深度成正相关。为了使不同基因、不同实验间估计的基因表达水平具有可比性，我们基于以下公式将Reads计数转换为FPKM值（每1百万个map上的reads中map到外显子的每1K个碱基上的片段个数），详细结果如下表所示，

表3.3-3 转录组支持率统计

sample_id	FPKM~[0,0.1)	FPKM~[0.1,)	FPKM~[0.1,3.75)	FPKM~[3.75,15)	FPKM~[15,)
XXX-2-J	5,468(19.71%)	22,280(80.29%)	7,252(26.14%)	7,484(26.97%)	7,544(27.19%)
XXX-2-Y	5,848(21.08%)	21,900(78.92%)	8,331(30.02%)	7,353(26.50%)	6,216(22.40%)

> 注：表格各列说明如下表：

列名	说明
sample_id	RNA-Seq数据样本名称
FPKM~[0,0.1)	FPKM值小于0.1的基因数据及占比
FPKM~[0.1,)	FPKM值大于等于0.1的基因数据及占比
FPKM~[0.1,3.75)	FPKM值大于等于0.1且小于3.75的基因数据及占比
FPKM~[3.75,15)	FPKM值大于等于3.75且小于15的基因数据及占比
FPKM~[15,)	FPKM值大于等于15的基因数据及占比

更多基因表达水平分析结果见目录：./src/summary/3_gene_structure/RNA_seq/expression/。

3.3.4 近缘物种基因信息比较统计

统计分析物种与相应近缘物种信息，该物种注释得到27,748个基因，基因平均长度为5,007.13bp，平均CDS长度为1,175.83bp，平均外显子个数为5.35。进行相关分布统计发现注释物种与近缘物种的分布趋势一致，表明注释结果可靠。

表3.3-4 基因组与其它物种基因集比较

Gene Set	Number of Genes	Average Gene Length(bp)	Average CDS Length(bp)	Average Exons per Gene	Average Exon Length(bp)	Average Intron Length(bp)
XXX	27,748	5,007.13	1,175.83	5.35	219.76	589.04
Cc	26,505	5,397.54	1,216.35	5.47	292.75	850.96
Ec	28,703	3,764.55	1,341.86	5.42	299.50	703.71
Ct	23,219	6,449.00	1,172.75	6.05	264.86	935.10
Tk	45,224	2,629.67	996.22	4.89	265.17	545.41
Tm	45,551	2,532.05	1,018.47	4.44	298.06	565.38
Ls	38,293	2,473.71	1,062.32	4.54	281.46	339.66

Gene Set	Number of Genes	Average Gene Length(bp)	Average CDS Length(bp)	Average Exons per Gene	Average Exon Length(bp)	Average Intron Length(bp)
Ha	70,864	2,995.34	920.34	4.09	321.72	545.64
Ls	42,719	2,521.94	1,110.62	5.00	261.39	512.23
Cn	64,257	3,530.17	1,069.34	5.03	234.02	681.56

> 注：表格各列说明如下表：

列名	说明
Gene Set	物种编号
Number of Genes	基因总数
Average Gene Length	基因平均长度
Average CDS Length(bp)	CDS平均长度，以bp计
Average Exons per Gene	每个基因平均外显子数
Average Exon Length(bp)	平均外显子长度，以bp计
Average Intron Length(bp)	平均内含子，以bp计

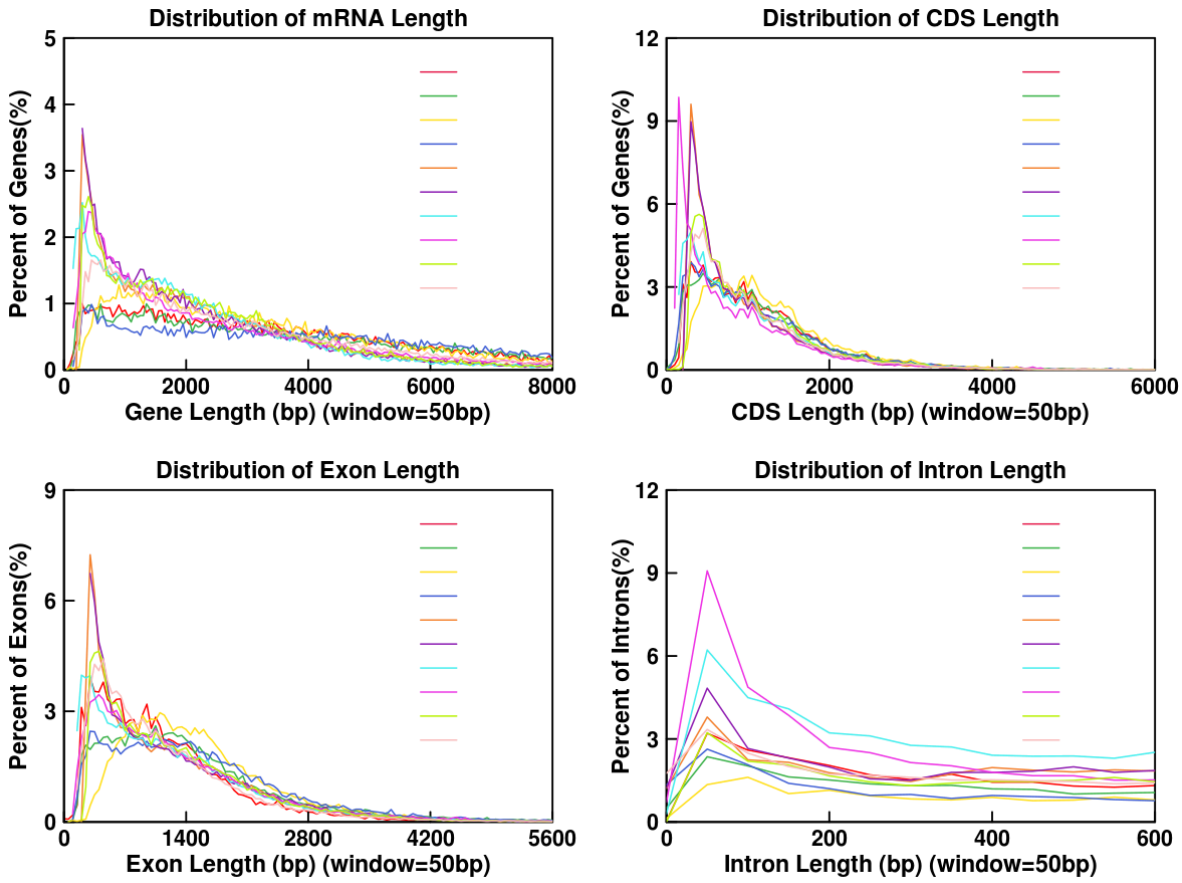

图3.3-3 动物与其它近缘物种基因各元件分布图

4 软件及数据库

表4.1 分析所用软件信息表

分析模块	分析内容	工具名称	版本
重复序列注释	SSRs(STRs)序列预测	Krait	v1.4.0
重复序列注释	串联重复序列预测	TRF	v4.09
重复序列注释	MITE序列从头预测	MITE-Tracker	v1.0
重复序列注释	转座元件从头预测	RepeatModeler	v2.0.2a
重复序列注释	转座元件同源注释	RepeatMasker	v4.1.2
重复序列注释	转座元件归类注释	TEsorter	v1.4.6
非编码基因注释	tRNA基因从头预测	tRNAscan-SE	v1.23
非编码基因注释	rRNA基因从头预测	RNAmmer	v1.2
非编码基因注释	rRNA基因从头预测	Barrnap	v0.8
非编码基因注释	ncRNA基因同源注释	Infernal	v1.1.4
编码基因结构注释	基因结构从头预测	Augustus	v3.4.0
编码基因结构注释	转录组预测-数据质控	fastp	v0.22.0
编码基因结构注释	转录组预测-比对	HISAT2	v2.2.1
编码基因结构注释	转录组预测-组装	StringTie	v2.2.1
编码基因结构注释	转录本UTRs预测	GUSHR	v1.0.0
编码基因结构注释	转录本UTRs预测	GeMoMa	v1.6.2
编码基因结构注释	同源蛋白预测-比对	ProtHint	v2.6.0
编码基因结构注释	同源蛋白预测-比对	Spaln	v2.4.13f
编码基因结构注释	同源蛋白+转录组预测	GeneMark-ETP	v1.00
编码基因结构注释	预测结果整合	TSEBRA	v1.1.1
编码基因结构注释	综合预测流程	BRAKER3	v3.0.3
基因功能注释	蛋白库同源比对	diamond	v2.0.15
基因功能注释	基因功能整合注释	eggNOG-mapper	v2.1.0
基因组注释评估	基因集完整度评估	BUSCO	v5.2.2

表4.2 分析所用数据库信息表

分析内容	数据库名称	版本
非编码基因注释	Rfam	v14.7
重复序列注释	Dfam	v3.5_curated

分析内容	数据库名称	版本
重复序列注释	RepBase	v20181026
编码基因注释	OrthoDB	v11
基因功能注释	eggNOG	v5.0
基因功能注释	GO	r20211215
基因功能注释	KEGG	r100.0
基因功能注释	Swiss-Prot	r2022.02
基因功能注释	TrEMBL	r2022.02
基因功能注释	NR	v20220801
基因功能注释	PFAM	v35.0

5 材料方法

5.1 基因组结构注释

5.1.1 重复序列注释

基因组中存在大量的重复序列，按照在基因组中分布的方式主要可分为两大类，分别是串联重复序列与散在重复序列。

串联重复（Tandem Repeat, TR）指DNA中的一个或多个核苷酸前后相连接的重复。串联重复又分为卫星DNA（Satellite DNA）、小卫星DNA（Minisatellite DNA）、微卫星DNA（Microsatellite DNA）。微卫星在动物里面一般称为短串联重复序列(Short Tandem Repeats, STRs)，在植物里面一般称为（Simple Sequence Repeats，SSRs）。由于STR/SSR在基因组中具有多态性的特点，STR分析法已经成为法医学领域个体识别和亲权鉴定的重要分析方法，可应用于司法案件调查，也就是遗传指纹分析。SSR在植物中也经常被用作遗传标记使用。

散在重复序列不同于串联重复序列，它在基因组中是分散存在的。具有在基因组中移动的能力，因此也被称为转座元件（Transposable Element，TE）。TE按照它在基因组中移动的方式可以分为两类，一类是通过RNA介导的“复制-粘贴”方式实现移动，二类不依赖RNA通过“剪切-粘贴”的方式实现转座。由于其可复制可转座的特点，TE在基因组中有时非常丰富，在有些物种中占到了80%以上。这经常给编码基因的预测和注释带来困难，因此通常在进行编码基因预测和注释之前需要将TE进行屏蔽。

5.1.1.1 串联重复分析

鉴于串联重复的重要性，我们采用两款软件Krait[1]和Tandem Repeats Finder（TRF[2]）以默认参数在全基因组范围内搜索串联重复序列。Krait主要搜索重复单元较短的简单重复序列（SSRs），TRF搜索全部重复单元的串联重复。将基因组进行SSR软屏蔽（碱基用小写标记）后，用TRF软件再搜索串联重复，TRF软件能覆盖所有重复单元的TR, 在进行TE注释之前我们将串联重复进行软屏蔽，以减小TE注释的搜索范围，同时避免TR和TE的冲突。

5.1.1.2 散在重复分析

基于TR软屏蔽的基因组进行TE搜索。我们首先采用MITE-Tracker[28]在基因组中搜索一种称为微型反向重复转座子（Miniature Inverted Transposable Elements，**MITE**）的小转座子，这类转座子在动植物中都存在。先形成一个MITE库文件。

针对大多数的植物我们还会用LTR_finder[29]和LTRharvest[30]两款软件分别搜索重复序列，再利用他们各自的结果用LTR_retriever[31]构建一个LTR重复序列库文件。

将以上两种库整合起来形成一个TE库文件，并对基因组进行一次硬屏蔽（碱基标记为N），使用RepeatModeler[3]进一步进行*de novo*搜索重复序列，形成*de novo*库文件(RepMod.lib)。考虑到RepMod.lib中包含较多未知重复元件，进而使用TEsorter[32]进行分类。最后将从头预测库和Dfam[33]及Repbase[34]库进行整合，形成一个总的库文件，利用该库文件和RepeatMasker对全基因组进行重复序列的搜索。

5.1.2 基因结构注释

基因结构预测采用从头预测，基于同源预测和转录组预测三种方法相结合，最后通过TSEBRA[12]软件整合，获得基因编码区结构注释结果。然后，基于转录组数据，利用GUSHR[35]和GeMoMa[36]将转录本比对至基因组，预测基因集的UTR区域。

5.1.2.1 转录组预测

测序得到的原始图像数据经碱基识别（base calling）转化为序列数据，称之为raw data或raw reads，以FASTQ[36]文件格式存储。FASTQ文件为用户得到的最原始文件，里面存储reads的序列以及reads的测序质量。FASTQ格式文件中每个read由四行描述：

```
@HWI-ST966:160:C28PYACXX:8:1101:1474:2178 1:N:0:GCCAAT
CAAAAACTGAAAGCATTTGGTTTCTACGGAACATACATATCCAGCAACCAGGCCAACTTAATTAACTCGGGGTCTAACGAAAGCTGC
GTTCTTCTCTTAG
+
BBBFFFFFFFFFFFIIIIIIIFIIBFIIIIFIIIIIIIFFIIIIIBFFIIIIIIFIIIFBFIBFFFIFFFFFFFBBBBBFBBBBBFBF
F7BBBBBBBBBBF
```

其中第一行以“@”开头，随后为测序标识别符（Sequence Identifiers）和描述文字（可选）；第二行是碱基序列；第三行以“+”开头，随后为可选的测序标识别符和描述性文字；第四行是对应碱基的测序质量编码序列。

第四行每个字符对应的ASCII值减去33，即为该碱基的测序质量值，比如B对应的ASCII值为66，那么其对应的碱基质量值是33。如果测序错误率用E表示，碱基质量值用sQ表示，则有下列关系：`sQ = -10log10(E)`。测序错误率与测序质量值简明对应关系如下面：

表5.1-1 测序错误率与测序质量值简明对应关系

测序错误率	测序质量值	对应字符
5%	Q13	.
1%	Q20	5
0.1%	Q30	?

使用fastp[4]软件对原始数据进行过滤，过滤原则为：1. 去除包含接头（adaptor）的reads；2. 去除N比例10%以上的reads；3. 去除低质量（质量值小于20）的碱基占比大于50%的reads。然后clean data进行质控，统计clean data产量。若下机数据质控合格，则进行后续分析。

数据质控后，使用HISAT2[5]软件将clean data比对到参考基因组，如果参考基因组选择合适并且相关实验不存在污染的情况下，clean data的基因组比对率将大于70%，其中具有多个定位的测序序列占总体的百分比通常不会超过10%。最后使用StringTie[6]基于默认参数将比对信息转为转录本坐标，利用转录本坐标提取对应区域的序列信息，即可得到二代转录本序列信息。

5.1.2.2 同源蛋白预测

GeneMark-ETP[7]是基于同源性进行基因预测的软件。通过近缘物种蛋白编码基因对本物种的序列进行基因结构的预测，主要依据氨基酸和内含子的保守性。此外，还整合转录组比对数据进行可变剪接位点的分析。

基于提供的近缘物种信息，获得对应物种的蛋白序列，通过ProtHint[8]或Spaln[9]将蛋白序列回比到基因组，确定相应蛋白比对的位置信息，进而得到基于同源蛋白注释的基因信息。

5.1.2.3 从头预测

Augustus[10]是真核基因组结构从头预测软件，主要运用广义隐马尔可夫的概率模型（Generalized Hidden Markov Model，GHMM[38]）进行基因结构的预测。通过分析DNA序列在概率模型中最有可能的基因结构，从而发现目标DNA序列中的基因。此外软件自身包含常见模式生物训练集，可利用这些物种直接进行基因预测；也可以利用同源预测和转录组最优结果生成训练集，预测基因。

基于StringTie利用转录组序列注释得到的基因信息，采用GeneMark-ES[39]软件对基因模型进行半监督自我训练，再进行预测。基于过滤后的结果选取GeneMark-ES比对得分最高的3,000个基因作为训练集用于Augustus模型训练。最后基于预测模型利用Augustus对基因组的基因进行预测。

5.1.2.4 注释结果整合

从头预测、同源注释和转录组预测得到不同的预测基因集。通过TSEBRA软件对这些数据进行整合，获得非冗余外显子集合，从而定义出更加可靠的基因结构。将上述三种方法预测得到的基因集信息文件转换为TSEBRA所接受的文件格式，以默认参数对基因信息文件进行整合，得到初始非冗余基因集。

5.1.3 ncRNA注释

非编码RNA(non-coding RNA，ncRNA)是指不编码蛋白质的RNA。包括rRNA、tRNA、snRNA、snoRNA和microRNA等多种已知功能的RNA，及其他未知功能的RNA。按照长度来划分，可以分为三大类：1. 小于50nt，主要包括microRNA、siRNA和piRNA，2. 在50nt-500nt之间，包括含量最高且最常见的是rRNA和tRNA，以及snRNA和snoRNA等。3. 大于500nt，主要指lncRNA (long noncoding RNA)。目前对于ncRNA的预测主要采用以下三种方法，然后再对这三种方法进行总的整合得到最终结果：

1. rRNA、snRNA和miRNA等预测是基于与Rfam[14]数据库进行比对，其比对方法是调用软件Infernal[13]中的程序中cmscan，将提交的序列在Rfam.cm数据库中进行检索，从而得到其比对的结果。
2. 对基因组中tRNA序列的预测，一般采用软件tRNAscan-SE[15]进行预测；
3. 通过软件RNAmmer[16]或Barrnap[17]构建模型来预测rRNA及其各类亚基软件。

5.2 基因功能注释

5.2.1 NR数据库注释

非冗余蛋白序列数据库（Non-Reduntant Protein Database，NR[19]）由NCBI建立及维护，包含了所有GenBank、EMBL、DDBJ和PDB的非冗余蛋白序列，对于所有已知的编码序列，NR数据库中都给出了相应的氨基酸序列，并且在注释结果中包含物种信息，可用于物种分类。

注释方法：使用diamond[18]软件将基因组蛋白序列比对到NR数据库，统计蛋白序列注释到的NR信息并对注释到的物种信息进行分类。

5.2.2 KEGG数据库注释

京都基因与基因组百科全书数据库（Kyoto Encyclopedia of Gene and Genomes，KEGG[20]）是日本京都大学生物信息学中心于1995年建立的数据库，该数据库描述了生物体中复杂的生物学通路，其丰富的通路信息能够帮助我们系统地了解蛋白的生物学功能，如代谢通路、遗传信息传递以及细胞过程等一些复杂的生物功能。

注释方法：使用EggNOG-mapper软件将基因组蛋白序列比对到KEGG数据库，并统计蛋白序列注释到的KEGG通路信息。

5.2.3 KOG数据库注释

真核同源蛋白簇数据库（Eukaryotic Orthologous Groups of protein，KOG[22])由NCBI创建并维护，根据真核生物完整基因组的编码蛋白系统进化关系分类构建而成。通过比对可将某个蛋白序列注释到某一个KOG中，每一簇KOG由直系同源序列构成，从而推测该序列的功能。

注释方法：使用EggNOG-mapper软件将基因组蛋白序列比对到KOG数据库，并统计蛋白序列注释到的KOG信息。

5.2.4 GO数据库注释

基因本体数据库（Gene Ontology，GO[23]）由基因本体联合会建立的将所有与基因有关的研究结果进行分类汇总的综合数据库。利用GO数据库可以将基因按照其参与的生物过程（Biological Process，BP）、细胞组分（Cellular Component，CC）和分子功能（Molecular Function，MF）三个方面进行分类注释。

注释方法：使用EggNOG-mapper软件，以默认参数将基因组蛋白序列比对到eggNOG数据库，将比对上的NOG编号转换为对应的GO注释信息，并按生物过程、细胞组分和分子功能三个方面进行分类注释。

5.2.5 Swiss-Prot和TrEMBL数据库注释

Swiss-Prot[24]和TrEMBL[25]隶属于UniProt数据库，Swiss-Prot包含经过注释和验证的严格去冗余的蛋白序列数据库，提供了蛋白序列详尽的注释信息。TrEMBL是通过计算分析方法获得的蛋白序列注释，为人工审编提供候选参考。

注释方法：使用diamond软件将基因组蛋白序列比对到Swiss-Prot和TrEMBL数据库，并统计蛋白序列注释到的Swiss-Prot和TrEMBL的蛋白信息。

5.3 基因组注释结果评估

5.3.1 BUSCO评估

通用单拷贝直系同源基因集（Benchmarking Universal Single-Copy Orthologs，BUSCO[27]）评估是指在基因含量层面来评估其基因集完整性。首先，构建好进化中各分支BUSCO的数据库（例如真核生物数据库，真菌数据库，鸟类数据库等等）。其中，运行BUSCO需要python、hmmer和Augustus等软件的支持。其次，对注释得到的蛋白序列进行比对评估，利用HMMER3[40]进行比对，从而评估基因集完整性。如果匹配长度在BUSCO配置文件匹配长度的预期范围内，则将其归类为“完成”。如果不止一次发现它们，则它们被归类为“重复的”。仅部分恢复的匹配被分类为“碎片”，没有匹配上的被分类为“缺失”。

将注释得到基因的氨基酸序列或CDS作为输入文件，比对选定的物种分支库，获得BUSCO完整度分析结果。

5.3.2 功能注释评估

功能注释评估是指用结构注释预测的基因与功能数据库进行比对来评估注释结果的准确性。主要的数据库有Swiss-Prot、TrEMBL、KOG、NR、KEGG和GO。基于所用数据注释结果信息进行交叠统计，得到每个数据库注释得到的基因数量及其占比，同时对所有数据库注释结果取并集进行统计，查看最终所有数据库（并集）的注释率。

5.3.3 转录组表达评估

转录组表达评估是指结构注释出来的基因与转录组数据进行比对，判断基因是否有转录组数据支持来评估注释结果的可靠性。首先，利用fastp[4]将转录组数据进行过滤（去除低质量reads和接头序列），然后利用HISAT2将过滤后的转录组数据与基因组进行比对得到比对文件，随后利用SAMtools[41]对得到的比对结果进行排序，然后再基于排序后的BAM文件利用StringTie软件计算基因的表达量FPKM值（Fragments per Kilobase Million，即每1百万个可比对的PE-reads中比对到该基因单位长度上的数量）。最后对注释得到的基因的FPKM值分布进行统计，FPKM值大于0.1或1则认为这个基因有转录组数据支持。

5.3.4 近缘物种基因信息比较统计

注释物种与近源物种在分类关系上一般属于同科或同属，亲缘关系较近。在编码基因层面上有着相似的结构与功能。因此对于注释后的基因，我们统计了基因信息统计包括基因数目、基因长度、平均CDS长度、平均外显子个数等。如果注释物种与近源物种的分布趋势一致，则说明注释结果的准确（如提供的参考物种较远或其注释结果较差，相互比对的参考意义不大）。

6 参考英文流程

6.1 Annotation of non-coding RNAs (ncRNAs)

To obtain the ncRNA (non-coding RNA), two strategies were used: searching against database and prediction with model. Transfer RNAs (tRNAs) were predicted using tRNAscan-SE with eukaryote parameters. MicroRNA, rRNA, small nuclear RNA, and small nucleolar RNA were detected using cmscan in Infernal toolkits to search the Rfam database. The rRNAs and their subunits were predicted using RNAmmer or Barrnap.

6.2 Repet Annotation

We first annotation the tandem repeats using the software Krait and Tandem Repeats Finder(TRF） where Krait identifies the simple repeat sequences (SSRs) and TRF recognizes all tandem repeat elements in the whole genome. Transposable elements (TE) in the XXX genome were then identified using a combination of *ab initio* and homology-based methods. Briefly, an *ab initio* repeat library for XXX was first predicted using MITE-Tracker and RepeatModeler with default parameters, in which LTR_Finder, LTR_harverst and LTR_retriever were also be included for plant genome. The obtainted library was then aligned to TEsorter Repbase (http://www.girinst.org/repbase) to classify the type of each repeat family. For further identification of the repeats throughout the genome, RepeatMasker was applied to search for known and novel TEs by mapping sequences against the de novo repeat library and Dfam/Repbase TE library. Overlapping transposable elements belonging to the same repeat class were collated and combined.

6.3 Gene Prediction

Three independent approaches, including *ab initio* prediction, homology-based search, and transcriptome-guided identify, were used for gene prediction in a TE-masked genome. In detail, GeneMark-ETP was used to align the homologous peptides from closely-related species to the assembly and then got the gene struture information, which was homolog prediction. For RNAseq-based gene prediction, filtered mRNA-seq reads were aligned to the reference genome using HISAT2. The transcripts were then assembled using stringtie. For the transcriptome-guided prediction, RNA-seq reads were assembled using stringtie and analyzed with StringTie to produce a training set. Augustus with default parameters were then utilized for *ab initio* gene prediction with the training set. Finally, TSEBRA was used to produce an integrated gene set. Untranslated regions (UTRs) and alternative splicing regions were determined using GUSHR and GeMoMa based on RNA-seq alignments. We retained the longest transcripts for each locus, and regions outside of the ORFs were designated UTRs.

6.4 Functional annotation of gene models

Gene function information, motifs and domains of their proteins were assigned by comparing with public databases including Swiss-Prot, TrEMBL, NR, KEGG, KOG and Gene Ontology. The putative domains and GO terms of genes were identified using the EggNOG-mapper program with default parameters. For the other four databases, diamond was used to compare the TSEBRA-integrated CDS or protein sequences against the four well-known public protein database with an E-value cutoff of 1e−05 and the results with the hit with lowest E-value was retained. Results from the six database searches were concatenated.

> 此内容为通用流程，如需用于论文发表，请根据实际内容进行修正并注意语言修改。

7 参考文献

1. Du L, Zhang C, Liu Q, Zhang X, Yue B, Hancock J. **Krait: an ultrafast tool for genome-wide survey of microsatellites and primer design**. *Bioinformatics*. 2018;34(4):681-683. doi:10.1093/bioinformatics/btx665
2. Benson G. **Tandem repeats finder: a program to analyze DNA sequences**. *Nucleic Acids Res*. 1999;27(2):573-580. doi:10.1093/nar/27.2.573
3. Tarailo-Graovac M, Chen N. **Using RepeatMasker to identify repetitive elements in genomic sequences**. *Curr Protoc Bioinformatics*. 2009;Chapter 4:. doi:10.1002/0471250953.bi0410s25
4. Chen S, Zhou Y, Chen Y, Gu J. **fastp: an ultra-fast all-in-one FASTQ preprocessor**. *Bioinformatics*. 2018 Sep 1;34(17):i884-i890. doi: 10.1093/bioinformatics/bty560. PMID: 30423086; PMCID: PMC6129281.
5. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. **Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype**. *Nat Biotechnol*. 2019;37(8):907-915. doi:10.1038/s41587-019-0201-4
6. Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, Pertea M. **Transcriptome assembly from long-read RNA-seq alignments with StringTie2**. *Genome Biol*. 2019;20(1):278. doi:10.1186/s13059-019-1910-1
7. Bruna T, Lomsadze A, Borodovsky M. **GeneMark-ETP: Automatic Gene Finding in Eukaryotic Genomes in Consistence with Extrinsic Data**. *bioRxiv*. 2023;2023.01.13.524024. doi:10.1101/2023.01.13.524024
8. Brůna T, Lomsadze A, Borodovsky M. **GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins**. *NAR Genom Bioinform*. 2020;2(2):lqaa026. doi:10.1093/nargab/lqaa026
9. Iwata H, Gotoh O. **Benchmarking spliced alignment programs including Spaln2, an extended version of Spaln that incorporates additional species-specific features**. *Nucleic Acids Res*. 2012;40(20):e161. doi:10.1093/nar/gks708
10. Stanke M, Diekhans M, Baertsch R, Haussler D. **Using native and syntenically mapped cDNA alignments to improve de novo gene finding**. *Bioinformatics*. 2008;24(5):637-644. doi:10.1093/bioinformatics/btn013
11. Gabriel L, Brůna T, Hoff KJ, et al. **BRAKER3: Fully Automated Genome Annotation Using RNA-Seq and Protein Evidence with GeneMark-ETP, AUGUSTUS and TSEBRA**. *bioRxiv*. 2023;2023.06.10.544449. doi:10.1101/2023.06.10.544449
12. Gabriel L, Hoff KJ, Brůna T, Borodovsky M, Stanke M. **TSEBRA: transcript selector for BRAKER**. *BMC Bioinformatics*. 2021;22(1):566. doi:10.1186/s12859-021-04482-0
13. Nawrocki EP, Eddy SR. **Infernal 1.1: 100-fold faster RNA homology searches**. *Bioinformatics*. 2013;29(22):2933-2935. doi:10.1093/bioinformatics/btt509
14. Kalvari I, Nawrocki EP, Ontiveros-Palacios N, et al. **Rfam 14: expanded coverage of metagenomic, viral and microRNA families**. *Nucleic Acids Res*. 2021;49(D1):D192-D200. doi:10.1093/nar/gkaa1047
15. Lowe TM, Eddy SR. **tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence**. *Nucleic Acids Res*. 1997 Mar 1;25(5):955-64. doi: 10.1093/nar/25.5.955. PMID: 9023104; PMCID: PMC146525.
16. Lagesen K, Hallin P, Rødland EA, Staerfeldt HH, Rognes T, Ussery DW. **RNAmmer: consistent and rapid annotation of ribosomal RNA genes**. *Nucleic Acids Res*. 2007;35(9):3100-3108. doi:10.1093/nar/gkm160
17. Seemann T (2013). **barrnap 0.8 : rapid ribosomal RNA prediction**. https://github.com/tseemann/barrnap
18. Buchfink B, Xie C, Huson DH. **Fast and sensitive protein alignment using DIAMOND**. *Nat Methods*. 2015;12(1):59-60. doi:10.1038/nmeth.3176
19. Deng YY, Li JQ, Wu S F, Zhu YP, et al. **Integrated NR Database in Protein Annotation System and Its Localization**. *Computer Engineering* 2006.,32(5):71-74.
20. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. **KEGG: new perspectives on genomes, pathways, diseases and drugs**. *Nucleic Acids Res*. 2017;45(D1):D353-D361. doi:10.1093/nar/gkw1092
21. Huerta-Cepas J, Forslund K, Coelho LP, et al. **Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper**. *Mol Biol Evol*. 2017;34(8):2115-2122. doi:10.1093/molbev/msx148
22. Tatusov RL, Fedorova ND, Jackson JD, et al. **The COG database: an updated version includes eukaryotes**. *BMC Bioinformatics*. 2003;4:41. doi:10.1186/1471-2105-4-41

23. Gene Ontology Consortium. **The Gene Ontology resource: enriching a GOld mine**. *Nucleic Acids Res*. 2021;49(D1):D325-D334. doi:10.1093/nar/gkaa1113

24. McMillan LE, Martin AC. **Automatically extracting functionally equivalent proteins from Swiss-Prot**. *BMC Bioinformatics*. 2008;9:418. doi:10.1186/1471-2105-9-418

25. O'Donovan C, Martin MJ, Gattiker A, Gasteiger E, Bairoch A, Apweiler R. **High-quality protein knowledge resource: SWISS-PROT and TrEMBL**. *Brief Bioinform*. 2002;3(3):275-284. doi:10.1093/bib/3.3.275

26. Kuznetsov D, Tegenfeldt F, Manni M, et al. **OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity**. *Nucleic Acids Res*. 2023;51(D1):D445-D451. doi:10.1093/nar/gkac998

27. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. **BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs**. *Bioinformatics*. 2015 Oct 1;31(19):3210-2. doi: 10.1093/bioinformatics/btv351.

28. Crescente JM, Zavallo D, Helguera M, Vanzetti LS. **MITE Tracker: an accurate approach to identify miniature inverted-repeat transposable elements in large genomes**. *BMC Bioinformatics*. 2018;19(1):348. doi:10.1186/s12859-018-2376-y

29. Ou S, Jiang N. LTR_FINDER_parallel: parallelization of LTR_FINDER enabling rapid identification of long terminal repeat retrotransposons. Mob DNA. 2019;10:48. doi:10.1186/s13100-019-0193-0

30. Ellinghaus D, Kurtz S, Willhoeft U. **LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons**. *BMC Bioinformatics*. 2008;9:18. doi:10.1186/1471-2105-9-18

31. Ou S, Jiang N. **LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons**. *Plant Physiol*. 2018;176(2):1410-1422. doi:10.1104/pp.17.01310

32. Zhang RG, Li GY, Wang XL, et al. **TEsorter: an accurate and fast method to classify LTR-retrotransposons in plant genomes**. *Hortic Res*. 2022;9:uhac017. doi:10.1093/hr/uhac017

33. Hubley R, Finn RD, Clements J, et al. **The Dfam database of repetitive DNA families**. *Nucleic Acids Res*. 2016;44(D1):D81-D89. doi:10.1093/nar/gkv1272

34. Bao W, Kojima KK, Kohany O. **Repbase Update, a database of repetitive elements in eukaryotic genomes**. *Mob DNA*. 2015;6:11. doi:10.1186/s13100-015-0041-9

35. Keilwagen J, Hartung F, Paulini M, Twardziok SO, Grau J. **Combining RNA-seq data and homology-based gene prediction for plants, animals and fungi**. *BMC Bioinformatics*. 2018;19(1):189. doi:10.1186/s12859-018-2203-5

36. Keilwagen J, Hartung F, Grau J. **GeMoMa: Homology-Based Gene Prediction Utilizing Intron Position Conservation and RNA-seq Data**. *Methods Mol Biol*. 2019;1962:161-177. doi:10.1007/978-1-4939-9173-0_9

37. Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM. **The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants**. *Nucleic Acids Res*. 2010;38(6):1767-1771. doi:10.1093/nar/gkp1137

38. Stanke M, Schöffmann O, Morgenstern B, Waack S. **Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources**. *BMC Bioinformatics*. 2006;7:62. doi:10.1186/1471-2105-7-62

39. Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M. **Gene identification in novel eukaryotic genomes by self-training algorithm**. *Nucleic Acids Res*. 2005 Nov 28;33(20):6494-506. doi: 10.1093/nar/gki937.

40. Eddy SR. **Accelerated Profile HMM Searches**. *PLoS Comput Biol*. 2011;7(10):e1002195. doi:10.1371/journal.pcbi.1002195

41. Li H, Handsaker B, Wysoker A, et al. **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics*. 2009;25(16):2078-2079. doi:10.1093/bioinformatics/btp352

8 联系我们

西安浩瑞基因成立于2019年，公司引入了三代测序平台--3台PacBio Revio和7台Sequell设备，致力于深耕动植物基因组学、转录组和微生物组学研究的科研技术服务。2024年，与华大智造携手共建西北首家DCSLab组学前沿实验室，引入DNBSEO-T7测序平台，开展基于二代测序的单细胞转录组、时空转录组等前沿技术服务。凭借专业的一站式多组学技术，为广大科研客户提供专业、高效、可靠的组学科研技术服务。

联系方式

热线电话：+86 029-89303503

官方网站：www.xahorizon.cn

邮　箱：project@xahorizon.cn

地　址：陕西省西安市沣东新城中兴深蓝科技产业园A区2号楼3层