doi:10.3969/j.issn.1672-5565.2015.02.04

基于 EST 数据的水稻基因表达大规模初步分析

宋东光
(佛山科学技术学院园艺系,广东 佛山 528231)

摘　要:EST 序列代表了组织基因表达的转录信号,本研究尝试开发简单高效的大规模 EST 分析方法,从 NCBI 下载水稻(*Oryza sativa*)的所有 EST 序列并进行分析以获取水稻发育过程基因表达的重要信息。通过进行 blast 比对和 phrap 拼接分析,及利用 Unix 文本过滤方法,从 EST 序列拼接获得了 3 万多个重叠群序列。进一步将重叠群序列与 NCBI 核酸数据库进行比对获得了各个序列的注释信息。从重叠群的组织表达初步挖掘中发现花药的表达数量最多,为下一步探讨水稻发育器官特异表达基因调控打下了重要基础。

关键词:水稻;EST;Bast; Phrap; 组织特异表达
中图分类号:Q344+.13　　**文献标志码**:A　　**文章编号**:1672-5565(2015)-02-096-07

Large-scale preliminary analysis of rice gene expression mining from EST data

SONG Dongguang
(*Department of Horticulture*, *Foshan University*, *Foshan Guangdong* 528231, *China*)

Abstract: EST sequences represent transcribed signals of gene expressions in tissues. In this study, a simple and effective method for large-scale EST analysis was developed using all rice (*Oryza sativa*) ESTs downloaded from NCBI for mining important information in rice development. After the blast alignment, phrap contig joining, and Unix command-line filtering, over 30 000 contigs were obtained from EST sequences. Annotations of these contigs were returned with further alignments to NCBI nucleotide databases. Anther expressions showed the most abundant in this preliminary mining from annotations for different tissues. This lays an important foundation for further investigating tissue-specific regulation of gene expression in rice development.

Keywords: *Oryza sativa*; EST; Blast; Phrap; Tissue-specific expression

随着功能基因组学研究的广泛开展,阐明基因表达调控网络的分子机理成为了近年来分子生物学研究的主要领域之一。获得基因活动信息的方法如 EST,SAGE 分析,表达芯片分析等可以提供大量的基因活动信号,并进一步从获得的各种表达数据分析构建基因调控网络。其中,EST 分析获得的基因表达信息真实反映了细胞内基因活动的情况,包括基因的组织特异表达情况。大量的 EST 序列可以从 NCBI Genbank 数据库获取,研究者也能够从 cDNA 文库进行克隆快速测序获得,面对海量的序列数据需要有效的高通量分析工具才能提取出更多的基因表达谱信息并用于构建基因调控网络[1-3]。

EST 序列预处理如如去除载体序列、poly(A)尾巴等对于后续分析是很必要的,涉及 EST 的各种分析包括转录组、重叠群拼接,基因注释,SSR 及 SNP 多态性, ORF 确定,选择性剪接,microRNA 及非编码 RNA 分析,RNA 编辑,GO 查询,组织特异性表达谱分析以及构建基因调控网络等并取得了许多重要进展[4-10]。

本文开发了简单有效的工具以来自 NCBI 的水稻 EST 序列为材料进行大规模初步分析,包括进行 blast 比对,phrap 重叠群拼接与注释,及组织特异表达分析,为水稻生长发育过程基因表达调控网络的构建奠定重要基础。

1　材料与方法

1.1　操作系统和文本过滤工具

操作系统为 FreeBSD 10.0,由 The FreeBSD Project(http://www.freebsd.org/)开发, 利用其内嵌的 Unix 命令如 awk、sed、tr、uniq、split、comm、paste、

收稿日期:2015-01-22;**修回日期**:2015-04-15.
作者简介:宋东光,男,副教授,研究方向:生物信息学与植物分子育种;E-mail:3dsong@163.com.

join 及 sort 等进行 EST 序列预处理[11]及其他文本挖掘工作。

1.2　EST 序列及格式转换

"gz"压缩格式的 EST 序列数据从 NCBI 下载,提取其中的重要信息并转换为一行,每个字段由制表符隔开。抽取每个 EST 序列及其 id 并转换为 FASTA 格式,序列开始及末尾的长于 10nt 的 poly(A/C/G/T)通过前面的过滤命令进行去除。

1.3　Blast 比对分析

NCBI 开发的 blast 程序 blast-2.2.22-ia32-freebsd 用于 EST 序列的本地 blast 比对分析,每个 EST 序列彼此间进行相似性比对找出得分大于 100 的去除重复后合并其 id 于一行。

1.4　重叠群拼接分析

phrap 程序(由 Washington 大学的 Phil Green 开发,http://www.phrap.org)用于将相似性较高的 EST 序列重叠拼接获得重叠群(contig)。

1.5　重叠群注释

将以上的拼接重叠群进行远程 NCBI 网络 blast 比对以获得重叠群的注释,每次可以进行 200 个重叠群(FASTA 格式),返回结果保存为".txt"格式并只提取注释信息。

1.6　组织特异表达谱分析

不同组织表达的 EST 可以通过比较组织表达的 EST id 及拼接注释后的重叠群 id 得到。

2　结　果

2.1　大规模 EST 分析通路

本文的 EST 大规模分析流程图参照图 1。

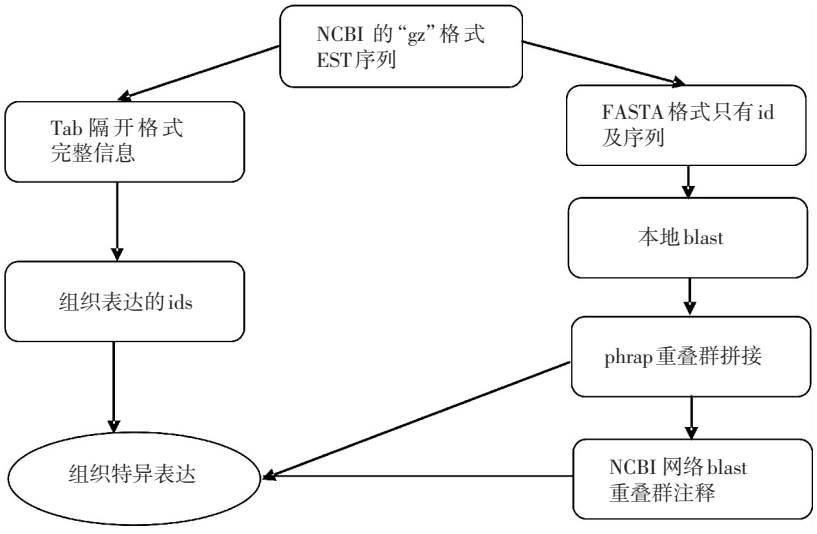

图 1　本文 EST 大规模分析流程图(具体过程见方法)

Fig.1　Flow-chart for EST large-scale analysis in this study(details in Methods)

EST 序列下载后将其从"gz"格式解压缩,提取必要信息并将转换为由制表符隔开的一行数据库录入格式,含 6 个字段即 GI-GenBank 数据库中的唯一标识号, DEFINITION-EST 数据定义信息,TITLE-测序记录号, /organism/-物种名, FEATURES-EST 序列简单介绍, ORIGIN-EST 核苷酸序列。典型的一个 EST 序列见图 2。

88908526　CI713085 Oryza sativa (japonica cultivar_group) 1 nuclei stage anther Oryza sativa Japonica Group cDNA clone J08B0009E04T3 5′, mRNA sequence.　Collection and mapping of over 30 000 transcription units by the rice full-length cDNA project from japonica rice　Oryza sativa Japonica Group　"mRNA" /cultivar="Nipponbare" /db_xref="taxon:39947" /clone="J08B0009E04T3" /tissue_type="1 nuclei stage anther" /clone_lib="LIBEST_019097 Oryza sativa (japonica cultivar_group) 1 nuclei stage anther
accggcgtgttcgggccaccggcggtggacgccaaggcggcggcggccgttggcgcggcgaaggccggcgcgaatggtggcgcctccgtgctcgaccagaaggtttggttccgccccctcgaggacgttga
gaagcctcccgtcgcctgatcgacaggcctaatacgagccttggctctgtgccacctaccattctcttcagctggggctccttgattattattactgtactagtagtgtggttctgctaatatagtatcagcagatcga
catggcgtgcatcggtagttacagtgaggcgatgctgatgcctaaaataagaacaagcaaaagacgtgtgctcatgttgcttgtggctgtaatgtacttct

图 2　提取转换格式后的一条典型 EST 序列,含 6 个字段由制表符分隔,即 GI, DEFINITION, TITLE, organism, FEATURES, ORIGIN, 分别与 NCBI 的 GenBank 数据库的内容对应

Fig.2　A typical extracted EST sequence data contains 6 Tab-separated fields of GI, DEFINITION, TITLE, organism, FEATURES, ORIGIN, which corresponds to GenBank data from NCBI

2.2 NCBI 记录的不同物种 EST 序列统计

截止 2014 年 2 月 14 日从 NCBI 下载的所有“gz”格式的 EST 序列提取其 GI 及 organism 后统计了各个物种的 EST 总数。119 个物种 EST 记录数超过 10 万条,但其中只有 63 个物种数量超过了 20 万条(见表 1,只列出了部分物种)。这其中,人(*Homo sapiens*)和家鼠(*Mus musculus*)记录数最多,分别达到了 8 千 7 百万和 4 千 8 百多万条,排在第三位的是玉米(*Zea mays*)有 2 百多万条,水稻(*Oryza sativa*)为 1 百多万条,包括了籼稻和粳稻(见表 1)。

表 1 截止 2014 年 2 月 14 日从 NCBI 下载的所有物种记录数

Table 1 Total EST records downloaded from NCBI dated Feb.14 2014

物种	数量	物种	数量	物种	数量
Acyrthosiphon pisum	214 833	*Gallus gallus*	600 434	*Ovis aries*	338 551
Aedes aegypti	301 595	*Gasterosteus aculeatus*	277 051	*Panicum virgatum*	720 590
Alvinella pompejana	218 454	*Glycine max*	1 461 723	*Phaseolus coccineus*	391 150
Aplysia californica	255 604	*Gossypium hirsutum*	297 540	*Physcomitrella patens*	382 587
Arabidopsis thaliana	1 529 700	*Homo sapiens*	8 704 880	*Picea glauca*	313 109
Bombyx mori	568 825	*Hordeum vulgare*	415 599	*Pimephales promelas*	258 504
Bos taurus	1 559 496	*Ictalurus punctatus*	354 516	*Pinus taeda*	328 661
Brachypodium distachyon	206 255	*Linum usitatissimum*	286 856	*Porphyridium purpureum*	386 903
Branchiostoma floridae	334 502	*Lottia gigantea*	252 091	*Rattus norvegicus*	1 103 576
Brassica napus	643 881	*Lotus japonicus*	242 432	*Saccoglossus kowalevskii*	202 190
Caenorhabditis elegans	396 687	*Macropus eugenii*	280 713	*Salmo salar*	498 245
Callithrix jacchus	292 992	*Malus domestica*	325 020	*Schistosoma mansoni*	208 557
Canis lupus familiaris	382 636	*Medicago truncatula*	269 498	*Solanum lycopersicum*	299 199
Chlamydomonas reinhardtii	204 075	*Mimulus guttatus*	231 095	*Solanum tuberosum*	249 969
Ciona intestinalis	1 205 673	*Mus musculus*	4 853 497	*Sorghum bicolor*	209 835
Citrus sinensis	214 598	*Neurospora crassa*	277 147	*Sus scrofa*	1 669 347
Crassostrea gigas	206 388	*Nicotiana tabacum*	335 082	*Triticum aestivum*	1 286 996
Culex quinquefasciatus	205 275	*Oncorhynchus mykiss*	287 565	*Vitis vinifera*	446 664
Danio rerio	1 488 339	*Oryza sativaIndica Group*	204 022	*Xenopus tropicalis*	1 271 480
Drosophila melanogaster	821 005	*Oryza sativaJaponica Group*	987 327	*Xenopus laevis*	677 911
Gadus morhua	257 218	*Oryzias latipes*	666 891	*Zea mays*	2 019 522

2.3 水稻 EST 序列彼此间的 blast 比对

水稻的 125 万条 EST 序列(截止 2010 年 3 月 24 日,包括籼稻和粳稻)经过预处理去除了 poly(A/T/G/C)后利用本地的 blast 程序进行了比对,比对工作连续进行约用时 1 个多月,之后将彼此比对打分达到 100 以上的序列 ids(即 GI 号)合为一行,得到 1 237 411 行 id 组,部分示例列于图 3。

```
10119739 163920841 4715115 89177972 88266751 86850863 29629682 89189895 88263899 86506006 44674929 58609385 88445856 86870708 88846158 88384942 88960745 88252014
11065615 218488962 86601392 88536975 88408619 88560924 88328458 88423949 88467385 88397705 88540515 88526571 88546397 88903804
11065616 29617906 29617870 4715718 58662804 29618278 10423644 87007400 86538713 88215996 58674681 86260828 86260807 88962214 66922123 33381288 86495568 33381360 88541471 88524563 88542928 86540422 88536287 88546041 88525981 88560343 88535202 88552618 88533187 88522745 88533840 88560552 88539761 88558894 88557633 23513945 30226736 29669005 117251641 88728629 88693206 33794374 29682747 29682379 29667657 29618608
11065617 218479325 10109130 117230561 163920246 10423610 86594434 66912947 27545704 86545576 86524466 88200457 66926434 86522341 88297987 88762546 88526867
13895815 86533063 87023985 86594565 38568558 88689643 88459057
```

图 3 相似性比对(打分 100 及以上的)EST 序列其 id 合为一行

Fig.3 Similar aligned(scored above 100) EST sequence joined in one-line by ids

blast 比对是用水稻的每个 EST 序列与所有的 EST 进行两两比对得到的结果,上述结果需要去除重复的相同行,并合并不同行中的相同 ids。去除重复行得到 543 460 行,然后每行内的 id 排序后将每行第一个 id 相同的行进行合并,得到 76 337 行,再次进行每行第一个 id 排序合并后得到 39 572 行。然后可以将每行内 id 代表的各个序列下一步用 phrap 获得重叠群,结果见表 2.

表 2　Blast 比对水稻所有的 EST 两两序列并合并序列相似性打分达到 100 以上序列 ids

Table 2　Blast total rice EST sequences against each other and joined together scored above 100

合并 blast 行	项目			
	两两比对,打分 100	排序去除重复行	首位 id 排序,合并	首位 id 再次排序
返回的行数	1 257 003, 1 237 411	543 460	76 337	39 572

2.4　用 phrap 拼接获得 EST 重叠群

根据前述方法用 phrap 程序从前面的 blast 比对结果进行重叠群拼接,获得只有一个重叠序列的重叠群为 27 556 个,两个以上超过一个重叠序列的为 7 413 个,所有重叠群序列总数达到 171 698 个(见图 4)。为了找出更合适的比对重叠群,将获得的重叠群两两进行了 blast 但打分大于 250,这样获得了 34 969 个比对结果,其中 16 900 个为单一序列(见图 5),这样为下一步进行 clustalw 比对分析很有帮助(本文未附)。

```
>10109107.Cotig1
CACCCGCCGCGCTTAATCCCCCGCTTACAGGGCGCNTCCCATTCGCCATTTCAGCTCCGCAACTGTTGGGAAGGCCNATCGGTGCGGC
CCTCTTCGCTATTACCCAGCTGGCGAAAGGGGGATGTGCTGCAAGGCGATTAAGTTGGGTAACGCCAGGGTTTTCCCAGTCACGACG
TTGTAAAACGACGGCCNGTGAGCGCGCGTAATACGACTCACTATAGGGCGAATTGGGTACCGGGCCCCCCCTCGAGGTCGACCCACG
CGTCCGCTAGGGTTTCGCCGCCGCCTCTTCGTCCCCGCCGCGACGATGCCGTCCCACAAGACCTTCCGCATCAAGAAGAAGCTGGCG
AAGAAGATGCGCCAGAACCGCCCCATCCCCTACTGGATCCGCATGCGGACCGACAACACCATCAGGTATAACGCGAAGCGCAGGCA
CTGGCGCCCGCACCAAGCTTGGGTTCTGAGCAGGAGGTGGCACCGGCTGCTGCCTGCGGATTTTGTTTTATGGATGTTTCAGTGTTTAT
GCAAGCTTAGGATGTGATTCCATGATGCTTTTTATTAAGGTTTTCTGAAATGTTAAGGGTAGTAGTACTCTGGTTTACTGAGGATACC
ATGCTATGCTGTGCTGAGTAAAAGAACCATTCTAATATATCTGCCTTGTTGAAGTTAAAAAAAAAAAAAAAAA
>10109107.Contig2
GGGCGCTGGCAAGTGTAGCGGTTCACTGCGCGTAACCACCACACCCGCCGCGCTTAATGCGCCGCTACAGGGCGCGTCCCATTCGCC
ATTCAGNTGCGCAACTGTTGGGAAGGGCGATCGGTGCGGGCCTCTTCGCTATTACGCCAGCTGGCGAAAGGGGGATGTGCTGCAAGG
CGATTAAGTTGGGTAACGCCAGGGTTTTCCCAGTCACGACGTTGTAAAACGACGGCCAGTGAGCGCGCGTAATACGACTCACTATAG
GGCGAATTGGGTACCGGGCCCCCCCTCGAGGTCGACCCACGCGTNCCGCGCGAAGCGCAGGCACTGGCGCCGCACCAAGCTTGGGTT
CTGAGCAGGAGGTAGCACCGGTTGCTGCCTGCGGAGGCTAGGGTTACTGAGTTATCTGCATCATTTTGTTTTATGGATGTTTCAGTGT
TTATGCAAGCTTAGGATGTGATTCCAGGATGCTTTTTATTAAGGTTTTCTGAAATGTTAAGTGTAGTTGTACTCTGGTTTACTGAGGAT
ACCATGCTATGCTGTGCTGAGTAAAGGAACCATTCCAATATATGTCTTTTGANGTTATTTCAAAAAAAAAAAAAAAAAAAAAAA
```

图 4　Blast 结果用 phrap 进行重叠群拼接

Fig.4　Contigs built from phrap joining of blast results

```
1036898.Contig1 11025574.Contig1 117239431.Contig2 117227446.Contig1 12405249.Contig2 218499895.Contig1 15857617.Contig4
117234368.Contig3 109713088.Contig1 226698423.Contig1 117252578.Contig1 163918222.Contig4 15847190.Contig1 86522348.Con-
tig2 163927916.Contig2 117247626.Contig2 117222284.Contig3 11171554.Contig1 29677590.Contig2 11610391.Contig1 117228922.
Contig1 12623322.Contig1 117251684.Contig1 117231274.Contig1 16577643.Contig1 11025496.Contig3 117249884.Contig1
1036902.Contig1 10611525.Contig1
1036912.Contig1
```

图 5　Phrap 得到的重叠群进行 blast 比对,显示了 3 行,每行超过一个重叠群的彼此相似性打分超过 250

Fig.5　Blast results from phrap joined contigs showing 3 lines, lines with more than one contigs had scored value above 250

2.5　重叠群与 NCBI nt 数据库比对进行注释

获取重叠群的注释尤为重要,将重叠群与 NCBI nt 核酸数据库进行比对后从返回的信息中挖掘各个重叠群的注释。全部的 34 969 个重叠群与 NCBI nt 数据库进行 blast 比对后,1 971 个没有返回比对结果,注释内容提取合为一行如图 6 所示。去除重复行后注释行总数为 211 351,但其中还有相当部分为未注释的行,如在含有 chromosome, cultivar:, genomic sequence, clone, mRNA sequence, unknown, hypothetical protein, DNA, Cosmid, vector, cDNA,

BAC clone, marker 等的比对结果中大部分没有有用的注释信息,还需要进一步去除约只有一半为有用的注释行,见图 7 示例.这些注释内容需要与前面的重叠群进行匹配后进一步进行挖掘。

```
Query= 10109112.Contig1
>ref|NM_001054196.1| Oryza sativa Japonica Group Os02g0664300 (Os02g0664300) mRNA, Score =2 627 bits (1 422), Expect = 0.0 Identities = 1 466/1 484 (99%), Gaps = 15/1 484 (1%)
>dbj|AK067099.1| Oryza sativa Japonica Group cDNA clone:J013096M07, full insert Score =2 621 bits (1 419),Expect = 0.0 Identities = 1 465/1 484 (99%), Gaps = 15/1 484 (1%)
>dbj|AK103515.1| Oryza sativa Japonica Group cDNA clone:J033131G16, full insert Score =2 132 bits (1154),Expect = 0.0 Identities = 1 198/1 216 (99%), Gaps = 15/1 216 (1%)
>ref|XM_006647584.1| PREDICTED: Oryza brachyantha tripeptidyl-peptidase 2-like (LOC102701235),Score =1 531 bits (829),Expect = 0.0 Identities = 1 120/1 250 (90%), Gaps = 62/1 250 (5%)
```

图 6 重叠群与 NCBI nt 数据库进行 blast 比对后提取的注释行示例

Fig.6 Sample annotated lines taken out from blast alignments against NCBI nt databases

```
AF014663 Human T-cell lymphotropic virus type 1 isolate BRASP31, 5′ long
AF015302 Oryza sativa RbohAOsp mRNA, partial cds
AF015522 Zea mays ribsomal protein S4 (rps4) mRNA, complete cds
AF015771 Magnaporthe grisea putative transcriptional regulator (CON7)
AF015784 Phaseolus vulgaris TATA-box binding protein (PVTBP1) mRNA, complete
AF015785 Phaseolus vulgaris TATA-box binding protein (PVTBP2) mRNA, complete
AF016404 Chloroplast transformation vector pEZC2040.2 for Euglena gracilis,
AF016889 Cloning vector pWSK29, complete sequence
AF016896 Oryza sativa GDP dissociation inhibitor protein OsGDI1 (OsGDI1)
AF016897 Oryza sativa GDP dissociation inhibitor protein OsGDI2 (OsGDI2)
AF016899 Schistosoma mansoni alpha-(1,3)-fucosyltransferase VII (SmFuct)
AF017063 Cloning vector pMECA, complete sequence
AF017356 Oryza sativa low molecular early light-inducible protein mRNA,
AF017357 Oryza sativa low molecular early light-inducible protein mRNA,
AF017360 Oryza sativa lipid transfer protein LPT III mRNA, complete cds
AF017362 Oryza sativa aldolase mRNA, complete cds
```

图 7 从注释行中去除非注释行获得的注释行示例,参见上下文分析

Fig.7 Sample annotations after removing unannotated ones as described in the context

物种关联的注释可提供一些有意义的信息,尤其是对于比较基因组学分析。从比对结果中找出了 939 个物种与水稻重叠群有关联,只有 82 个物种出现的注释超过 100 条,而其中仅仅 10 个超过了1 000 条。玉米与水稻的比对注释最多达到了36 804 条,大多数为 mRNA/cDNA/protein 的注释也许可以提供与基因功能相关的有用信息。剩下的 9 个中只有 Brachypodium distachyon 超过 1 万条,为 11 610 条,见图 8。

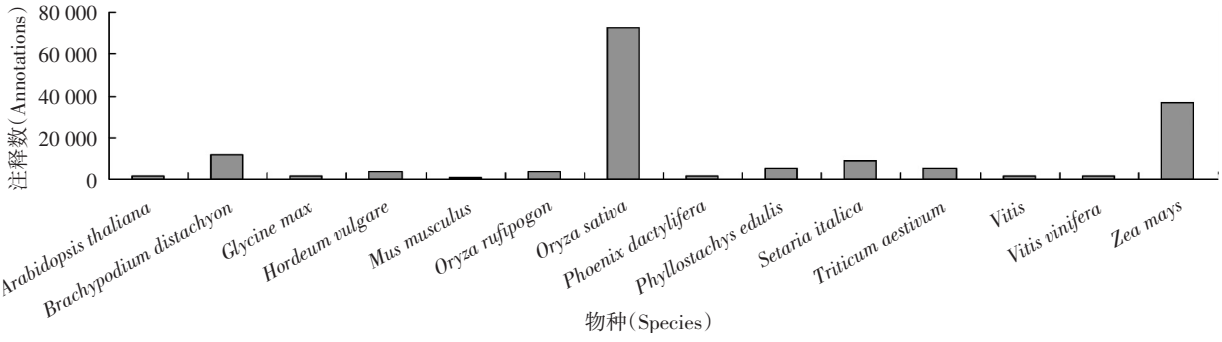

图 8 不同物种与水稻重叠群比对返回超过 1 000 条的注释数

Fig.8 Number of annotations above 1 000 counted for different species as to Oryza sativa contigs

每个重叠群比对结果出现推测基因功能的注释对于进一步的功能基因组学分析特别是构建基因调控网络是很有帮助的,这将是我们下一步的研究目标。

2.6 水稻发育过程组织特异表达

确定组织特异转录谱对于分析基因表达模式及

构建基因调控网络是很重要的。所有的 EST 记录中见图 9,其中花药的记录数最多。从比对的39 572 个 EST id(见图 2)组找出了各个组织的表达重叠群,结果见图 10,虽然表达重叠群中可能含有相似的重叠群,如图 5 所示。从图 10 可以清楚看出,花药的表达重叠群最多达到了最高重叠群数。这并奇怪,因为从花药的 EST 总数 977 141(见图 9)可以预见(分析的 EST 序列总数只有 125 万条),其他的组织都少于 20 万条。从以上结果尚不能完全的获得组织特异表达谱(见图 10),但是很显然组织特异表达谱对于构建水稻发育过程基因调控网络是很重要的,我们将在今后继续进行探讨。

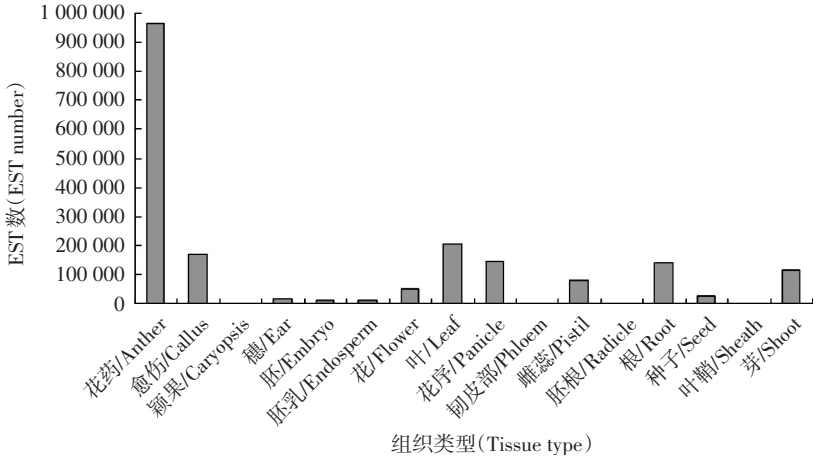

图 9　本文引用的 NCBI 来源水稻不同组织 EST 序列数

Fig.9　Total EST sequence records in different rice tissues from NCBI used in this study

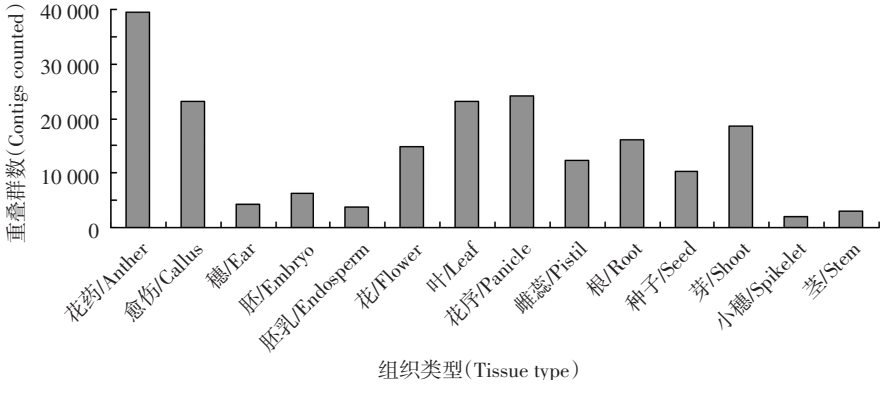

图 10　从比对后 EST id 组(见图 2)获取的不同组织表达重叠群计数

Fig.10　Counted tissue-expressed contigs from aligned EST id sets described in Fig.2

3　讨　论

EST 大数据包含了大量基因表达信息,EST 数据大规模分析有助于发现基因调控的活动情况,并可以用于构建基因调控网络。本文从 NCBI 下载了水稻的 125 万条 EST 序列并进行了基因表达分析。所有的分析工作都是通过 FreeBSD 操作系统完成的,主要工具包括 Unix 命令,及本地 blast,phrap 及远程 blast 程序(见方法)。经过 blast 比对,phrap 重叠群拼接及再比对,获得了 34 969 重叠群,其中约一半只有一个重叠群序列(见图 5)。进一步我们将重叠群序列与 NCBI 全长 cDNA 获取的单一基因进行比对以获得水稻的完整转录组。以上结果表明,我们进行的大规模 EST 分析是有效且快捷,与其他方法相比并不需要复杂的算法[3,10]。

本文初步分析了水稻的组织特异表达谱,发现花药表达的 EST 重叠群数量最多,其他组织较少些,原因尚未进一步分析(见图 10)。通过与 NCBI 核酸数据库进行远程比对,从返回结果中提取了每个重叠群的注释信息(见图 6~图 8),今后我们将着重挖掘特异表达基因并进一步构建水稻发育过程的基因调控网络。

参考文献(References)

[1] GIALLOURAKIS C C, BENITA Y, MOLINIE B, et al. Genome-wide analysis of immune system genes by expressed sequence Tag profiling[J]. J Immunol, 2013, 190(11):5578-87.

[2] SHA A H, LI C, YAN X H, et al. Large-scale sequencing of normalized full-length cDNA library of soybean seed at different developmental stages and analysis of the gene expression profiles based on ESTs [J]. Mol Biol Rep, 2012,39(3):2867-74.

[3] MENON R, GARG G, GASSER R B, et al. TranSeqAnnotator: large-scale analysis of transcriptomic data[J]. BMC Bioinformatics, 2012, 13(Suppl 17): S24.

[4] ZHU W, BUELL C R. Improvement of whole-genome annotation of cereals through comparative analyses[J]. Genome Res, 2007, 17(3):299-310.

[5] WARD J A, PONNALA L, WEBER C A. Strategies for transcriptome analysis in nonmodel plants[J]. Am J Bot, 2012, 99(2):267-76.

[6] LUO H, SUN C, LI Y, et al. Analysis of expressed sequence tags from the Huperzia serrata leaf for gene discovery in the areas of secondary metabolite biosynthesis and development regulation[J]. Physiol Plant, 2010, 139(1):1-12.

[7] FRAZIER T P, ZHANG B. Identification of plant microRNAs using expressed sequence tag analysis[J]. Methods Mol Biol, 2011, 678:13-25.

[8] VICTORIA F C, DA MAIA L C, DE OLIVEIRA A C. In silico comparative analysis of SSR markers in plants[J]. BMC Plant Biol, 2011, 11:15.

[9] XIE F, SUN G, STILLER J W, et al. Genome-wide functional analysis of the cotton transcriptome by creating an integrated EST database[J]. PLoS One, 2011, 6 (11):e26980.

[10] LI Y, GONG P, PERKINS E J, et al. RefNetBuilder: a platform for construction of integrated reference gene regulatory networks from expressed sequence tags[J]. BMC Bioinformatics, 2011, 12(Suppl 10):S20.

[11] SONG D G, ZHANG H S, HUANG L X, et al. Localization, Updating and Sequence Preprocessing of EST Database under Unix Environment[J]. Chinese Journal of Bioinformatics, 2010,8(1):52-56.