A New Method for EST Clustering

ZHANG Li-Da, YUAN De-Jun, ZHANG Jian-Wei, WANG Shi-Ping①, ZHANG Qi-Fa
(*National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan 430070, China*)

Abstract: We developed an EST (expressed sequence tag) clustering method, ESTClustering, to generate high-quality unique expressed sequence based on large-scale EST sequencing. The method uses consensus sequences to sequence analyze with megablast and assemble each cluster with phrap in clustering process. The clustering strategy can efficiently identify gene family and alternate splicing forms of expressed sequences. It can also reduce the adverse effects caused by sequence errors. The ESTClustering method tends to provide more expressed gene forms comparing with the UniGene clustering method of the National Center for Biotechnology Information. Analysis of the 112 256 ESTs of Arabidopsis with ESTClustering produced 23 581 EST clusters. Among these Arabidopsis EST clusters, 13 597 have corresponding genome coding sequences and this number is close to the number of genes predicted with Arabidopsis ESTs. Using this clustering method, a total of 147 191 rice ESTs were clustered into 33 896 groups.

Key words: EST clustering; consensus sequence; non-redundant cDNA library

一种新的 EST 聚类方法

张利达，袁德军，张建伟，王石平①，张启发
（华中农业大学作物遗传改良国家重点实验室，武汉 430070）

摘 要：该研究发展了一种 EST(expressed sequence tag)聚类方法(ESTClustering)，用于分析大规模 EST 测序中所产生的大量数据，以获得高质量、非重复表达序列。该方法在聚类过程中采用 MEGABLAST 工具对一致序列进行序列同源比较，并用 phrap 程序对每一 EST 簇进行拼接检验。这一聚类策略能降低测序错误带来的影响，有效识别基因家族成员，并避免选择性剪接的干扰。与 NCBI(National Center for Biotechnology Information)的 UniGene clustering 方法相比，ESTClustering 的聚类结果可以更好地反映表达序列的多样性。用 ESTClustering 对 112 256 条拟南芥 EST 聚类测试，产生 23 581 个 EST 簇，其中 13 597 个 EST 簇有对应拟南芥基因组编码序列，与该基因组中有 EST 作为依据的预测基因数目接近。应用该方法对收集的 147 191 条水稻 EST 序列进行聚类，形成 33 896 个 EST 簇。

关键词：EST 聚类；一致序列；无冗余 cDNA 文库
中图分类号：Q33 文献标识码：A 文章编号：0379-4172(2003)02-0147-07

表达序列标签(expressed sequence tag, EST)是对随机挑取的 cDNA 克隆的外源插入片段的一端或两端进行一次性测序产生的 DNA 序列[1]。每一个 EST 代表一个表达基因的部分转录片段。通过对 EST 序列的分析，从中可以获得大量的基因表达信息。为此，EST 被广泛应用于大规模基因鉴定[2]、图谱构建[3]、基因预测[4~6]、多态性分析[7,8]和表达差异[9]等研究，为基因组研究提供了一条快速、便捷的途径。由于 EST 是经一次测序产生的，序列质量相对较低，包含较多的测序错误[10]，并且大部分 EST 都未经编辑，不能提供高质量的一致序列(consensus sequence)，因而不能直接用来精确推断蛋白序列进

收稿日期：2002-05-24；修回日期：2002-12-16
基金项目：国家重点基础发展规划(973)资助项目[This research was supported by a grant from the National Key Program on Basic Resarch and Development of China (973)]
作者简介：张利达(1976-)，男，浙江省余姚市，硕士。研究方向：生物信息学。
① 通讯作者。E-mail: swang@mail.hzau.edu.cn

行蛋白质功能分析，也不能用来进行高质量的基因注释。

EST 聚类(clustering)分析通过序列同源比较或其他注释信息，把属于同一基因的 EST 聚合成一簇，以减少数据冗余程度，提高表达序列的数据质量[11]。目前，根据不同的研究目的发展了多种 EST 聚类分析方法，其中被广泛使用的有美国基因组研究所(The Institute for Genomic Research，TIGR)发展而来的 TIGR _ ASSEMBLER 方法[12] (http：//www.tigr.org/tdb/tgi.html)。该方法借助 FASTA 程序[13] 对序列进行两两比较，再根据同源比较结果用 TIGR-ASSEMBLER 工具对相关序列进行拼接，把重叠区超过 40 个碱基，且该区域的碱基同源性大于 95%的序列合并成一簇。TIGR 利用这个方法对来源 21 个物种的 5 358 611 条 EST 进行了聚类分析，分别建立了各个物种的基因索引(TIGR Gene Indices)[14]。UniGene Clustering 方法由美国国家生物技术信息中心(National Center for Biotechnology Information，NCBI)发展而来。该方法使用 MEGABLAST 程序[15] 对序列进行同源比较，采用的聚类阈值为序列间至少有 100 个碱基的重叠区，并且占 70%以上的重叠区域的碱基同源性大于 96%，依据该阈值先对已注释的基因聚类成簇，再根据 EST 与 EST 及 EST 与初始基因簇之间的序列同源性进一步进行聚类，由此产生的基因簇包括同一基因的不同剪接形式。NCBI 利用该方法建立的 UniGene 数据库，主要用于发展转录图谱，以确定基因组中的全部编码序列(http：//www.ncbi.nlm.nih.gov/UniGene/)[16,17]；南非国家生物信息研究院(South African National Bioinformatics Institute，SANBI)的 STACK-PACK 方法(http：//www.sanbi.ac.za/Dbases.html)，其主要特点是根据不同的组织来源先把 EST 分类，再根据重叠区超过 150 个碱基，且重叠区域的碱基同源性大于 96%的聚类阈值，用 d2-2cluster 程序[18] 对各类 EST 分别聚类。用 STACK-PACK 分析结果建立的 STACK 数据库可用来进行 SNPs 检测和基因特异性表达的研究[19]。

参考上述聚类分析方法，我们发展了一个新的 EST 聚类方法——ESTClustering，用于分析大规模 EST 测序中所产生的数据，以获取高质量、非重复表达序列(unique expressed sequence)。该聚类方法的主要特点是：在聚类过程中采用一致序列作为基因标准序列进行同源比较，并利用 phrap 程序[20] 对符合聚类阈值的 EST 进行序列拼接，以此分离相近的 EST 簇。同上述其他聚类分析方法相比，该聚类机制可以在一定程度上降低测序错误带来的影响，有效区分同一基因家族的不同成员及同一基因的不同剪接形式的产物。

1 聚类策略和方法评价

1.1 序列预处理

在大规模测序中，EST 数据中难免含有污染序列。如细菌基因组和核糖体的序列污染；另外，部分 EST 序列内部带有载体、接头和 poly(A/T)序列的污染。因此，在聚类分析之前需要对污染序列进行处理。我们收集了核糖体序列(ftp：//ftp.ebi.ac.uk/pub/databases/embl/)、细菌基因组序列(ftp：//ftp.ebi.ac.uk/pub/databases/embl/)、载体序列(ftp：//ftp.ncbi.nih.gov/pub/UniVec/)，以此构建了污染序列数据库。利用 Cross-Match 分析工具[21]，以污染数据库中的序列为对照，对所有待聚类分析的 EST 序列进行扫描并去除污染序列。对去除污染后的 EST 进一步筛选，舍弃序列长度小于 100 个碱基的 EST。

1.2 EST 聚类

EST 数据的测序错误率平均在 3%左右[22]。如果对同一片段分别进行 2 次测序，产生的 2 条序列之间的相似程度大约为 94%(97%×97%)。经调试，我们设定的聚类阈值为：2 条 EST 序列的重叠区超过 40 个碱基，且该区域的碱基同源性大于 94%。具体聚类过程如下：

取 1 条待分析 EST 作为检索序列，对其进行延伸。具体做法是利用 MEGABLAST 工具使其与数据库中的其余 EST 序列进行局部对齐，收集符合聚类条件的 EST；然后用 phrap 软件(默认参数)将这些 EST 拼接成一致序列。如拼接后的一致序列的有效延伸长度大于 20 个碱基，则该一致序列代替上述检索序列进行重新检索、拼接，直至获得的序列长度不能用此法继续延伸为止。将这些 EST 合并成一簇，并将它们从待分析序列的数据库中删除。另取 1 条待分析 EST，进入新的聚类循环，直至所有 EST 聚类成簇。在整个聚类分析过程中，与其他 EST 不匹配的 EST 序列单独成簇(singleton cluster)。按照匹配阈值对生成的所有 EST 簇再进行聚类，符合聚类条件的若干个簇重新合并成一簇，直至没有新簇产生(完整的聚类流程见图 1)。

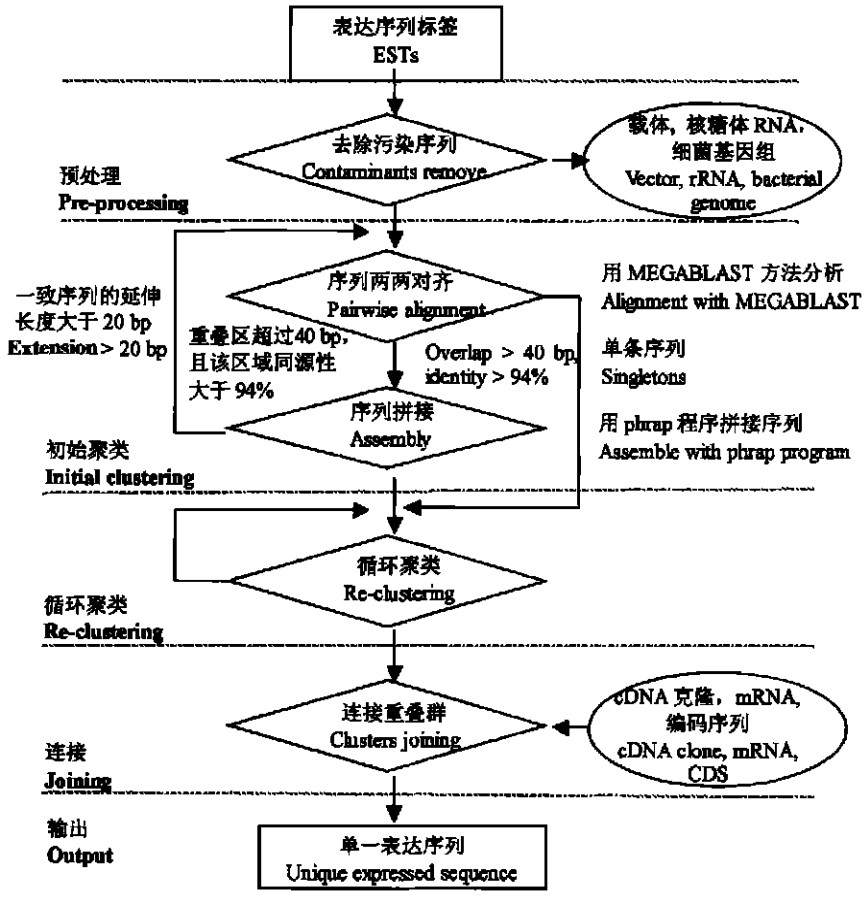

图 1　EST 聚类流程

Fig. 1　The process of EST clustering

1.3　EST 簇的连接

对于属于同一个基因但相互间无重叠区的 EST 簇, 则借助 mRNA、预测的基因组完整编码序列(coding sequence, CDS)及 cDNA 克隆信息进行连接。当多个 EST 簇与同一基因的完整编码序列具有同源性(图 2), 且同源区长度覆盖由 EST 簇拼接成的一致序列的 80%以上、碱基同源性大于 97%, 则将这些 EST 簇合并成一簇, 并将相对应的基因完整编码序列作为这个 EST 簇的表达序列。对于含有 cDNA 克隆信息(克隆的 5′和/或 3′末端序列已知)的 EST 簇, 聚类分析时以此为媒介, 将与同一 cDNA 克隆同源的独立 EST 簇合并成一簇, 相应各簇的一致序列通过 20 个"N"的接头按 5′→3′方向重新连接成表达序列[23](图 3)。

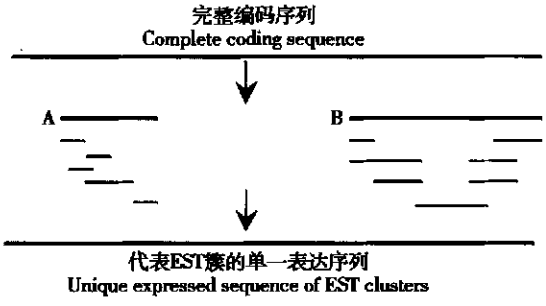

图 2　利用因组编码序列连接属于同一个基因而无重叠区的 EST 簇

A 和 B 分别代表两个 EST 簇和 EST 簇的一致序列。

Fig. 2　Link of non-overlapping EST clusters by coding sequence

A and B. EST clusters A and B and the consensus sequences of the clusters.

1.4　测试结果分析

1.4.1　测序错误对一致序列质量的影响

我们首先用计算机程序模拟测序错误和序列覆盖倍数对一致序列的准确性所带来的影响。以水稻抗病基因 *Xa21*(GenBank 注册号 U37133, 序列长度 3 078 bp)为测试序列, 用计算机程序对序列随机打

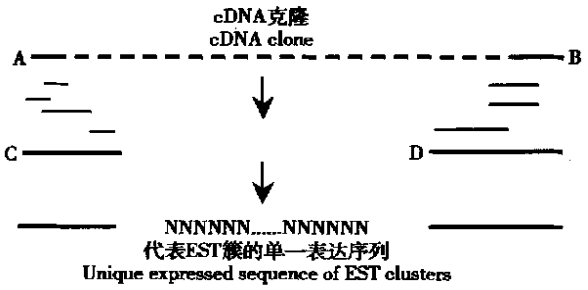

图 3 利用 cDNA 克隆信息连接属于同一个基因而无重叠区的 EST 簇

A 和 B 分别代表同一 cDNA 克隆的两端序列; C 和 D 分别代表序列 A 和 B 所在的两个 EST 簇和 EST 簇的一致序列。

Fig. 3 Link of non-overlapping EST clusters by cDNA clone

A and B, 5′ and 3′ sequences of the same cDNA clone; C and D, EST clusters C and D and the consensus sequences of the clusters.

断,分别产生相对于 *Xa21* 基因全序列覆盖倍数为 5 倍、10 倍、15 倍和 30 倍的不同长度的片段(300~500 bp),并对产生的片段按替换、插入、缺失比为 3∶1∶1 的频率进行随机突变[23],模拟错误率在 1%~8%[24] 的测序错误。分别用 phrap(默认参数)软件对不同覆盖倍数的序列进行拼接。测试结果表明,拼接成的一致序列随着覆盖倍数的增加而更接近基因的真实序列,但覆盖倍数超过一定程度后(覆盖倍数>15),一致序列会因为插入突变的积累反而影响一致序列的质量(表 1)。水稻 EST 序列的覆盖倍数远没有达到 15,为此在 EST 聚类分析过程中借助一致序列作为标准序列来进行相似性检索,可以减小由测序错误而引起的错误聚类。

表 1 用 phrap 软件对 *Xa21* 基因序列不同覆盖倍数的序列片段拼接结果

Table 1 Assembly of *Xa21* sequence fragments with different coverage by phrap program

覆盖倍数 Coverage	5	10	15	30
一致序列(bp) Consensus sequence (bp)	2 974	3 084	3 102	3 149
相似性(%) Identity (%)	95.3	98.4	98.5	97.2

1.4.2 ESTClustering 对基因家族的识别测试

EST 聚类分析的一个重要方面就是对基因家族不同成员的识别分离。仍以水稻抗病基因 *Xa21* 为测试序列,用计算机程序对 *Xa21* 序列进行随机替换突变,产生出与 *Xa21* 序列相似程度为 98%、97%、96%、95%和 90%的 5 个计算机模拟的基因家族成员。再用计算机程序分别将 *Xa21* 序列和这 5 个家族成员序列随机打断成 300~500 bp 不同长度的片段,覆盖率都为 10 倍。用 ESTClustering 对所有随机打断生成的片段一起进行聚类分析。测试结果表明,由相似程度 97%、96%、95%和 90%的 4 个计算机模拟的家族成员而产生的片段被各自聚合成簇;由相似程度 98%的家族成员而产生的片段不能被聚类程序所分离,与未突变的 *Xa21* 序列生成的片段聚到同一簇。可见,在没有测序错误的理想情况下,ESTClustering 能很好地区分两两之间相似程度小于 98%的家族成员,而且拼接成的一致序列能很好地反映各家族成员的本来面目。

1.4.3 ESTClustering 与 NCBI 的 UniGene clustering 方法比较

从 NCBI(ftp://ftp.ncbi.nlm.nih.gov/repository/UniGene/)下载水稻 UniGene 数据库(UniGene Release *Oryza sativa* #12, 2002-03-28),该数据库有 18 344 个水稻单一基因簇,共包括 84 114 个水稻基因片段,其中最大的一个基因簇的编号为 Os.15824,具有 1 273 个序列片段。用 ESTClustering 对所有 84 114 个基因片段进行重新聚类,生成 20 350 个基因簇,其中最大的基因簇共包含 429 个序列片段。与 UniGene 相比,单一序列数增加了 7%(从 10 287 增至 11 010),总簇数则增加了 10.9%(从 18 344 增至 20 350),并通过序列对齐的方法对这些新增基因簇进行了确认。两种方法的聚类结果的比较如图 4 所示,EST-Clustering 所采用的聚类机制与 NCBI 的 UniGene clustering 相比可以更好地反映表达序列的多样性,相应地使聚类结果中簇的总数增加,大簇数目减少。

1.4.4 ESTClustering 对拟南芥 EST 的聚类分析结果

从 EBML(European Molecular Biology Laboratory)核苷酸数据库分离获得 112 256 条符合分析条件的拟南芥 EST。拟南芥所有 26 182 个基因(hypothetical, predicted and experimentally verified)的完整编码序列则来自国际拟南芥组织(ftp://tairpub:tairpub@ftp.arabidopsis.org/home/tair/Sequences/blast-datasets/ATH1.cds.01072002, 2002-01-07)。用 ESTClustering 对 112 256 条拟南芥 EST 进行分析,聚合成 23 581 个 EST 簇,其中 13 597 个 EST 簇有对应拟南芥基因完整编码序列,这与拟南芥基因组中有 15 349 个预测基因有 EST 作为依据[25] 的数据较为接近。

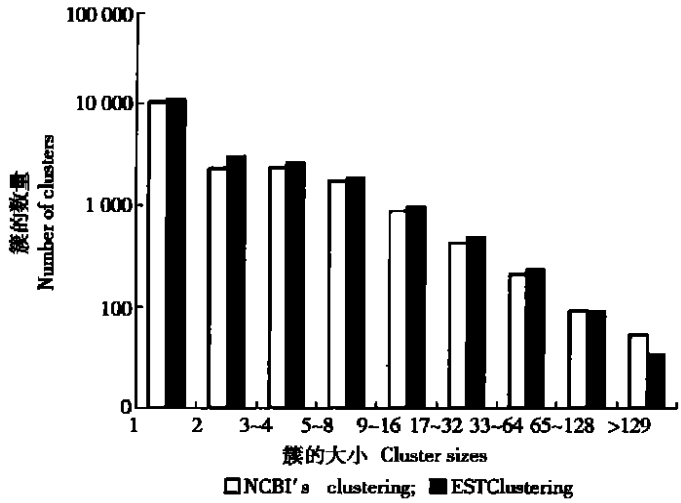

图 4　用 NCBI 的 UniGene clustering 与 ESTClustering 对 84 114 个水稻基因片段聚类结果比较

Fig. 4　EST Clusters produced by NCBI's UniGene clustering and ESTClustering using 84 114 rice gene fragments

2　聚类方法在水稻 EST 中的应用

2.1　数据来源

用于分析的 150 620 条水稻 EST 序列的数据来源: 本实验室和中科院国家基因研究中心对本室 7 个不同 cDNA 文库部分克隆进行 5′端测序获得的 EST 约 29 670 条, 中国科学院遗传与发育所及国家基因研究中心提供的 20 240 条 EST 以及从 EMBL 数据库中分离获得的水稻 EST 序列近 100 710 条。用于 EST 簇连接的 7 164 条水稻基因完整编码序列也从 EMBL 数据库中获取。

2.2　聚类分析结果

通过预处理后, 获得 147 191 条待分析 EST 序列, 其中包括来自本实验室 cDNA 文库的 28 384 条 EST。被剔除的水稻核糖体序列和细菌基因组污染序列等不合格序列共 3 429 条, 占序列总数的近 2.3%。

通过聚类分析初步合并成 40 006 个 EST 簇, 再以 7 164 条水稻基因完整编码序列和 cDNA 克隆为媒介最终合并成 33 896 个 EST 簇, 其中单一序列簇占总簇数的 55.9%。簇的大小分布情况具体见图 5。

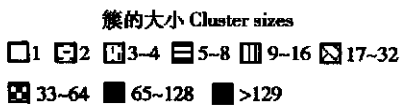

图 5　ESTClustering 对水稻 EST 聚类结果

Fig. 5　Summary of rice EST clusters analyzed by ESTClustering

聚类分析使表达序列总数由 147 191 条降至 33 896 条, 重复信息减少了近 4/5, 大大降低了 EST 数据的冗余程度, 拼接成的非重复表达序列的质量得到了提高。在聚类分析中, 通过 EST 簇连接, 合并了 6 110 个 EST 簇, 减少了 15.2%的重复簇, 说明水稻 EST 测序的覆盖率还没达到拼接成完整表达序列的水平, 部分相互不重叠的簇是同一表达序列的不同片段, 预计这一百分比将随着 EST 数据的增加而逐渐降低。

根据聚类的最终结果, 确定了包含本实验室 EST 的 EST 簇共 11 566 个; 再从原始文库中挑取每个簇的特征克隆, 重新整理成无冗余的水稻 cDNA 文库。

3　讨　论

3.1　EST 聚类分析的注意点

ESTClustering 方法的主要特点是在聚类过程中采用一致序列作为基因标准序列进行同源比较, 可以降低测序错误带来的影响; 对符合聚类阈值的 EST 再进行序列拼接可以更好地区分基因家族的不同成员和基因不同剪接形式的产物。采取这种聚类机制保证了聚类分析后的 EST 数据在有效地减小冗余性的同时, 最大限度地保留表达序列的多样性。聚类后 EST 簇连接可以减少不重叠的非独立 EST 簇的数量, 但在连接时要注意以下两点, 一是用编码序列作桥梁, 匹配条件不能太宽松, EST 簇间不要出

现较大的缺口(一般小于 10 个碱基),否则聚类结果会丧失一些相似程度较高的基因家族成员或因选择性剪接而产生的不同剪接形式。第二个注意点是 cDNA 文库中大约含有 1%左右的嵌合体克隆[26],因此,以克隆为媒介合并的 EST 簇,会产生一定的错误连接,最好的解决办法就是用基因组序列进行验证区分。

聚类分析采用的条件可视具体情况进行适当调整,严紧的聚类条件可以有效区分基因家族的不同成员和基因选择性剪接的不同产物,分析结果会产生更多、更小的 EST 簇,拼接成的一致序列质量较高,但长度较短。在分析序列覆盖倍数较小、质量不高的 EST 数据时,可以适当放宽聚类条件,以免测序错误给聚类结果带来较大干扰。

3.2 根据 EST 聚类结果重建无冗余 cDNA 文库

常规方法构建的 cDNA 文库含有大量的冗余克隆,采用平衡化和差减技术也只能在一定程度上减小文库的冗余程度[27]。在水稻 cDNA 文库构建过程中,我们首先用平衡化或差减技术尽量减小原始 cDNA 文库的冗余信息,再对文库进行大规模测序,并把测序产生的 EST 序列进行聚类分析,并根据聚类结果重新整理 cDNA 文库。采用这种反向消除方法不仅可以彻底解决 cDNA 文库所出现的冗余问题,而且还清除了核糖体和外来基因组污染的克隆。整理后 cDNA 所包含的克隆只占原克隆数的2/5,这为cDNA 文库的保存以及后期的利用、分析提供了极大的便利,尤其有利于用以制作 cDNA 芯片的研究。

参考文献(References):

[1] Adams M D, Kelley J M, Gocayne J D, Dubnick M, Polymeropoulos M H, Xiao H, Merril C R, Wu A, Olde B, Moreno R F. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, 1991, 252: 1651~1656.

[2] Boguski M S, Tolstoshev C M, Bassett D E Jr. Gene discovery in dbEST. *Science*, 1994, 265: 1993~1994.

[3] Boguski M S, Schuler G D. ESTablishing a human transcript map. *Nat Genet*, 1995, 10: 369~371.

[4] Bailey L C Jr, Searls D B, Overton G C. Analysis of EST-driven gene annotation in human genomic sequence. *Genome Res*, 1998, 8: 362~376.

[5] Lin X, Kaul S, Rounsley S, Shea T P, Benito M I, Town C D, Fujii C Y, Mason T, Bowman C L, Barnstead M, Feldblyum T V, Buell C R, Ketchum K A, Lee J, Ronning C M, Koo H L, Moffat K S, Cronin L A, Shen M, Pai G, Van Aken S, Umayam L, Tallon L J, Gill J E, Adams M D, Carrera A J, Creasy T H, Goodman H M, Somerville C R, Copenhaver G P, Preuss D, Nierman W C, White O, Eisen J A, Salzberg S L, Fraser C M, Venter J C. Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. *Nature*, 1999, 402: 761~768.

[6] Mayer K, Schuller C, Wambutt R, Murphy G, Volckaert G, Pohl T, Dusterhoft A, Stiekema W, Entian K D, Terryn N, Harris B, Ansorge W, Brandt P, Grivell L, Rieger M, Weichselgartner M, Simone V, Obermaier B, Mache R, Muller M, Kreis M, Delseny M, Puigdomenech P, Watson M, Schmidtheini T, Reichert B, Portatelle D, Perez-alonso M, Boutry M, Bancroft I, Vos P, Hoheisel J, Zimmermann W, Wedler H, Ridley P, Langham S A, Mccullagh B, Bilham L, Robben J, Schueren J V D, Grymonprez B, Chuang Y J, Vandenbussche F, Braeken M, Weltjens I, Voet M, Bastiaens I, Aert R, Defoor E, Weitzenegger T, Bothe G, Ramsperger U, Hilbert H, Braun M, Holzer E, Brandt A, Peters S, Staveren M V, Dirkse W, Mooijman P, Lankhorst R K, Rose M, Hauf J, Kötter P, Berneiser S, Hempel S, Feldpausch M, Lamberth S, Daele H V D, Keyser A D, Buysshaert C, Gielen J, Villarroel R, Clercq R D, Montagu M V, Rogers J, Cronin A, Quail M, Bray-Allen S, Clark L, Doggett J, Hall S, Kay M, Lennard N, Mclay K, Mayes R, Pettett A, Rajandream M A, Lyne M, Benes V, Rechmann S, Borkova D, Blöcker H, Scharfe M, Grimm M, Löhnert T H, Dose S, Haan M D, Maarse A, Schäfer M, Müller-Auer S, Gabel C, Fuchs M, Fartmann B, Granderath K, Dauner D, Herzl A, Neumann S, Argiriou A, Vitale D, Liguori R, Piravandi E, Massenet O, Quigley F, Clabauld G, Mündlein A, Felber R, Schnabl S, Hiller R, Schmidt W, Lecharny A, Aubourg S, Chefdor F, Cooke R, Berger C, Montfort A, Casacuberta E, Gibbons T, Weber N, Vandenbol M, Bargues M, Terol J, Torres A, Perez-Perez 27 A, Purnelle B, Bent E, Johnson S, Tacon D, Jesse T, Heijnen L, Schwarz S, Scholler P, Heber S, Francs P, Bielke C, Frishman D, Haase D, Lemcke K, Mewes H W, Stocker S, Zaccaria P, Bevan M, Wilson R K, Bastide M D L, Habermann K, Parnell L, Dedhia N, Gnoj L, Schutz K, Huang E, Spiegel L, Sehkon M, Murray J, Sheet P, Cordes M, Abu-Threideh J, Stoneking T, Kalicki J, Graves T, G. Harmon G, Edwards J, Latreille P, Courtney L, Cloud J, Abbott A, Scott K, Johnson D, Minx P, Bentley D, Fulton B, Miller N, Greco T, Kemp K, Kramer J, Fulton L, Mardis E, Dante M, Pepin K, Hillier L, Nelson J, Spieth J, Ryan E, Andrews S, Geisel C, Layman D, Du H, Ali J, Berghoff A, Jones K, Drone K, Cotton M, Joshu C, Antonoiu B, Zidanic M, Strong C, Sun H, Lamar B, Yordan C, Ma P, Zhong J, Preston R, Vil D, Shekher M, Matero A, Shah R, Swaby I, O' shaughnessy A, Rodriguez M, Hoffman J, Till S, Granat S, Shohdy N, Hasegawa A, Hameed A, Lodhi M, Johnson A, Chen E, Marra M, Martienssen R, McCombie W R. Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana*. *Nature*, 1999, 402: 769~777.

[7] Buetow K H, Edmonson M N, Cassidy A B. Reliable identification of large numbers of candidate SNPs from public EST data. *Nat Genet*, 1999, 21: 323~325.

[8] Garg K, Green P, Nickerson D A. Identification of candidate coding region single nucleotide polymorphisms in 165 human genes using assembled expressed sequence tags. *Genome Res*, 1999, 9: 1087 ~ 1092.

[9] Okubo K, Hori N, Matoba R, Niiyama T, Fukushima A, Kojima Y, Matsubara K. Large scale cDNA sequencing for analysis of quantitative and qualitative aspect of gene expression. *Nat Genet*, 1992, 2: 173 ~ 179.

[10] Aronson J S, Eckman B, Blevins R A, Borkowski J A, Myerson J, Imran S, Elliston K O. Toward the development of a gene index to the human genome: An assessment of the nature of high-throughput EST sequence data. *Genome Res*, 1996, 6: 829 ~ 845.

[11] Hide W, Miller R, Ptitsyn A, Kelso J, Gopallakrihnan C, Christoffels A. EST Clustering Tutorial, ISMB in Heidelberg, Germany, 1999, 6.

[12] Sutton G, White O, Adams M D, Kerlavage A R. TIGR Assembler: A new tool for assembling large shotgun sequencing projects. *Genome Sci Technol*, 1995, 1: 9 ~ 18.

[13] Pearson W R, Lipman D J. Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA*, 1988, 85: 2444 ~ 2448.

[14] Quackenbush J, Cho J, Lee D, Liang F, Holt I, Karamycheva S, Parvizi B, Pertea G, Sultana R, White J. The TIGR Gene Indices: Analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Res*, 2001, 29: 159 ~ 164.

[15] Zhang Z, Schwartz S, Wagner L, Miller W. A greedy algorithm for aligning DNA sequences. *J Comput Biol*, 2000, 7: 203 ~ 214.

[16] Boguski M S, Schuler G D. ESTablishing a human transcript map. *Nat Genet*, 1995, 10: 369 ~ 371.

[17] Schuler G D, Boguski M S, Stewart E A, Stein L D, Gyapay G, Rice K, White R E, Rodriguez-Tome P, Aggarwal A, Bajorek E, Bentolila S, Birren B B, Butler A, Castle A B, Chiannilkulchai N, Chu A, Clee C, Cowles S, Day P J, Dibling T, Drouot N, Dunham I, Duprat S, East C, Edwards C, Fan J B, Fang N, Fizames C, Garrett C, Green L, Hadley D, Harris M, Harrison P, Brady S, Hicks A, Holloway E, Hui L, Hussain S, Louis-Dit-Sully C, Ma J, MacGilvery A, Mader C, Maratukulam A, Matise T C, McKusick K B, Morissette J, Mungall A, Muselet D, Nusbaum H C, Page D C, Peck A, Perkins S, Piercy M, Qin F, Quackenbush J, Ranby S, Reif T, Rozen S, Sanders C, She X, Silva J, Slonim D K, Soderlund C, Sun W L, Tabar P, Thangarajah T, Vega-Czarny N, Vollrath D, Voyticky S, Wilmer T, Wu X, Adams M D, Auffray C, Walter N A R, Brandon R, Dehejia A, Goodfellow P N, Houlgatte R, Hudson J r, Ide S E, Iorio K R, Lee W Y, Seki N, Nagase T, Ishikawa K, Nomura N, Phillips C, Polymeropoulos M H, Sandusky M, Schmitt K, Berry R, Swanson K, Torres R, Venter J C, Sikela J M, Beckmann J S, Weissenbach J, Myers R M, Cox D R, James M R, Bentley D, Deloukas P, Lander E S, Hudson T J. A gene map of the human genome. *Science*, 1996, 274: 540 ~ 546.

[18] Burke J, Davison D, Hide W. d2-cluster: a validated method for clustering EST and full-length cDNA sequences. *Genome Res*, 1999, 9: 1135 ~ 1142.

[19] Miller R T, Christoffels A G, Gopalakrishnan C, Burke J, Ptitsyn A A, Broveak T R, Hide W A. A comprehensive approach to clustering of expressed human gene sequence: The sequence tag alignment and consensus knowledge base. *Genome Res*, 1999, 9: 1143 ~ 1155.

[20] Green P. http: //bozeman. genome. washington. edu/phrap. docs/phrap. html. phrap 2002.

[21] Green P. http: //bozeman. genome. washington. edu/phrap. docs/general. html. crossmatch 2002.

[22] Boguski M S, Lowe T M, Tolstoshev C M. dbEST——database for ‘expressed sequence tags.’ *Nat Genet*, 1993, 4: 332 ~ 333.

[23] Hide W, Miller R, Ptitsyn A, Kelso J, Gopallakrishnan C, Christoffels A. EST Clustering Tutorial, ISMB in Heidelberg, Germany, 1999, 15.

[24] Liang F, Holt I, Pertea G, Karamycheva S, Salzberg S L, Quackenbush J. An optimized protocol for analysis of EST sequences. *Nucleic Acids Res*, 2000, 28: 3657 ~ 3665.

[25] Kaul S, Koo H L, Jenkins J, Rizzo M, Rooney T, Tallon L J, Feldblyum T, Nierman W, Benito M, Lin X, Town C D, Venter J C, Fraser C M, Tabata S, Nakamura Y, Kaneko T, Sato S, Asamizu E, Kato T, Kotani H, Sasamoto S, Ecker J R, Theologis A, Federspiel N A, Palm C J, Osborne B I, Shinn P, Conway A B, Vysotskaia V S, Dewar K, Conn L, Lenz C A, Kim C K, Hansen N F, Liu S X, Buehler E, Altafi H, Sakano H, Dunn P, Lam B, Pham P K, Chao Q, Uyen M, Yu G, Chen H, Youthwick A, Lee J M, Miranda M, Toriumi M J, Davis R W, Wambutt R, Murphy G, Düsterlöft A, Stiekema W, Pohl T, Entian K D, Terryn N, Volckaert G, Salanoubat M, Choisne N, Rieger M, Ansorge M, Unseld M, Fartmann B, Valle G, Artiguenave F, Weissenbach J, Quetier F, Wilson R K, Bastide M, Sekhon M, Huang E, Spiegel L, Gnoj L, Pepin K, Murray J, Johnson D, Habermann K, Dedhia N, Parnell L, Preston R, Hillier L, Chen E, Marra M, Martienssen R, McCombie W R, Mayer K, White O, Bevan M, Lemcke K, Creasy T H, Bielke C, Haas B, Haase D, Maiti R, Rudd S, Peterson J, Schoof H, Frishman D, Morgenstern B, Zaccaria P, Ermolaeva M, Pertea M, Quackenbush J, Volfovsky N, Wu D, Lowe T M, Salzberg S L, Mewes H, Rounsley S, Bush D, Subramaniam S, Levin I, Norris I, Schmidt R, Acarkan A, Bancroft I, Quetier F, Brennicke F, Eisen J A, Bureau T, Legault B A, Le Q H, Agrawal N, Yu Z, Martienssen R, Copenhaver D P, Luo S, Pikaard C s, Preuss D, Paulsen I T, Sussman M, Britt A B, Eisen J A, Selinger D A, Pandey R, Mount D W, Chandler V L, Jorgensen R A, Pikaard G, Juergens G, Meyerowitz E M, Ecker J R, Theologis A, Dangl J, Jones J D G, Chen M, Chory J, Somerville C. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, 2000, 408: 796 ~ 815.

[26] Hillier L D, Lennon G, Becker M, Bonaldo M F, Chiapelli B, Chissoe S, Dietrich N, DuBuque T, Favello A, Gish W, Hawkins M, Hultman M, Kucaba T, Lacy M, Le M, Le N, Mardis E, Moore B, Morris M, Parsons J, Prange C, Rifkin L, Rohlfing T, Schellenberg K, Marra M. Generation and analysis of 280000 human expressed sequence tags. *Genome Res*, 1996, 6: 807 ~ 828.

[27] Bonaldo M F, Lennon G, Soares M B. Normalization and subtraction: Two approaches to facilitate gene discovery. *Genome Res*, 1996, 6: 791 ~ 806.

（责任编辑：周　素）