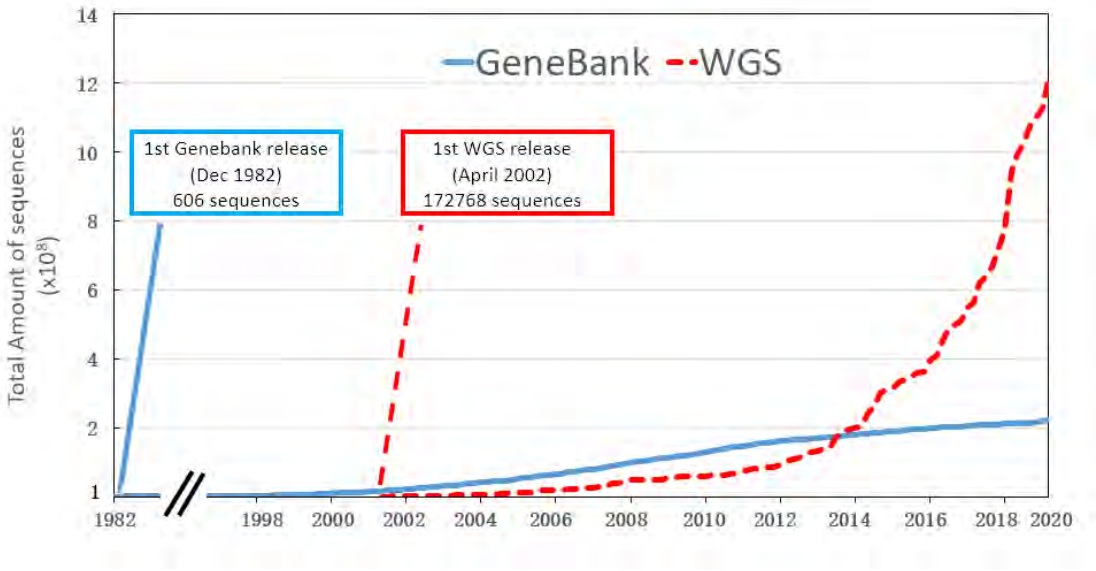


图1.1　美国国家生物技术信息中心（NCBI）核苷酸序列公共数据库GenBank 和全基因组鸟枪法测序数据库（Whole Genome Shotgun，WGS）序列数据（条数）增长情况（改自Gauthier et al.，2018；数据截至 2020 年 2 月，Release 236）

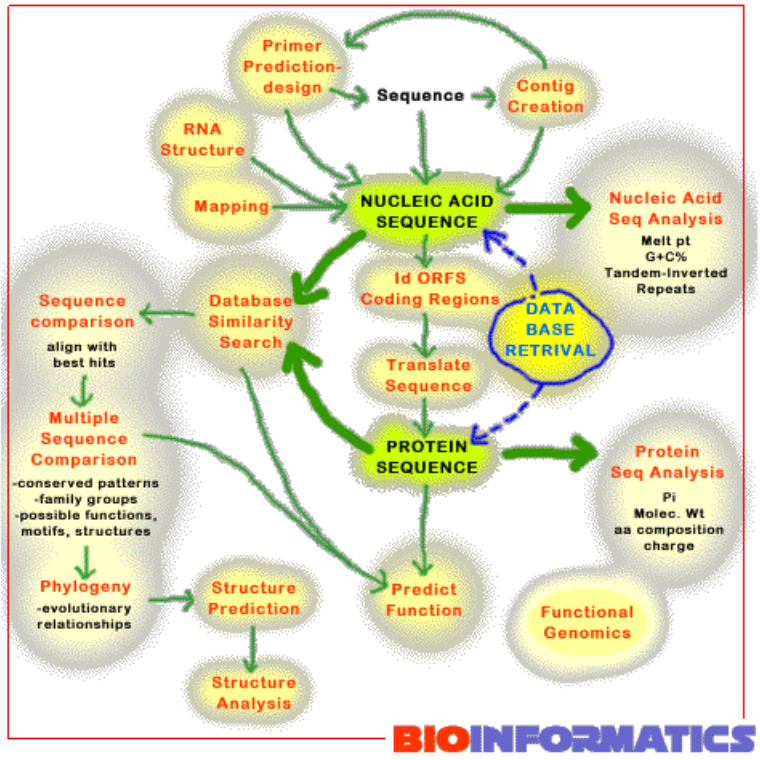


图1.2　生物信息学早期的一个“路线图”

该图给出了生物信息学主要涉及的领域

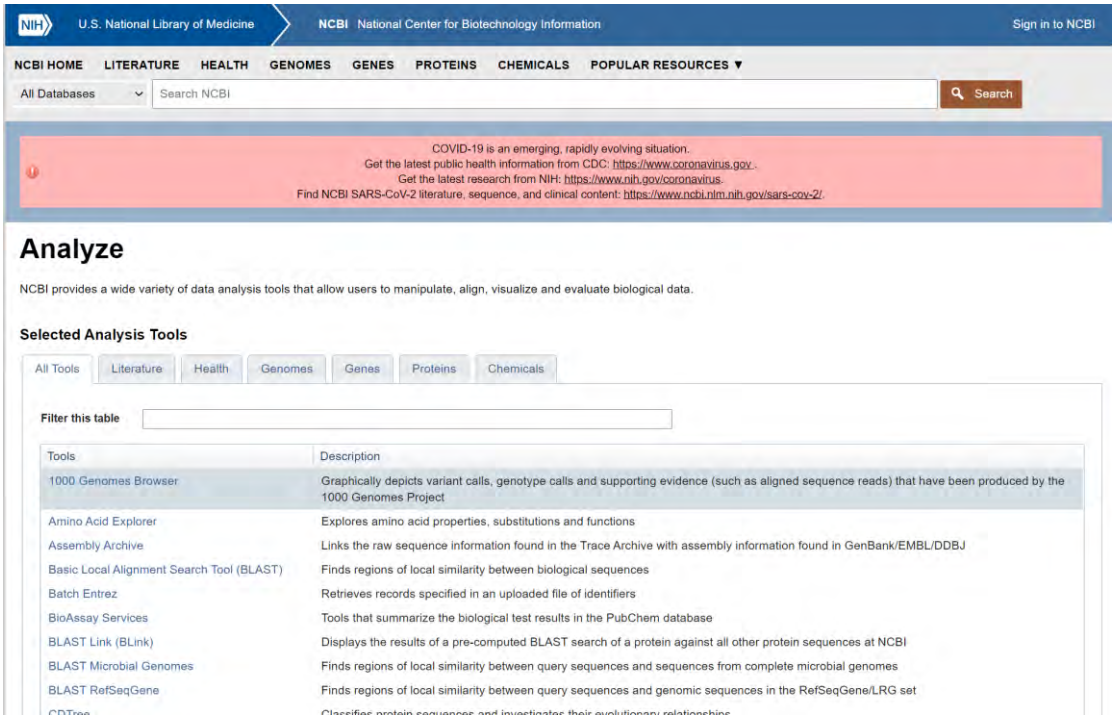


图1.3　美国国家生物技术信息中心网站数据分析工具网页

包括 BLAST、e-PCR 等工具软件

图1.4　生物信息学家们面对的是堆积如山的DNA 片段

这是在2001 年人类基因组测序完成后出现的一幅漫画。有了序列数据，接下来（图中的“Phase Two”）最重要的是如何解

读人类自身 3Gb 的基因组

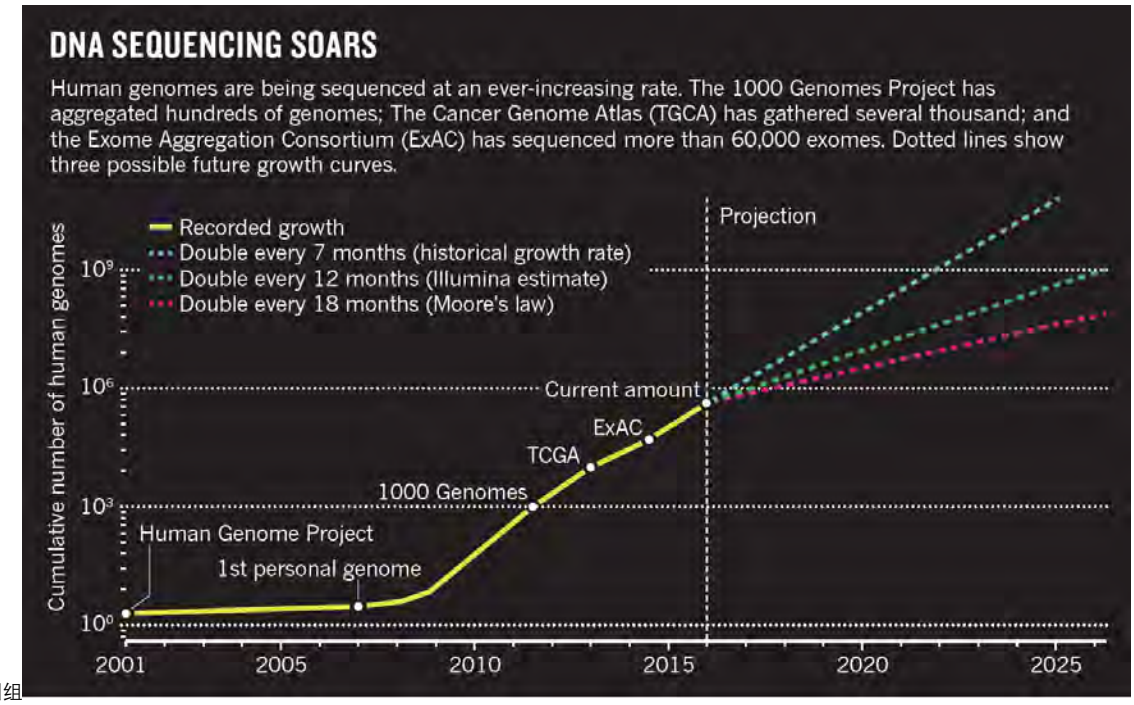


图1.5　人类基因组测序数量的增长（引自Eisenstein，　2015）

图中给出了数据增长的三个不同预测曲线。截至 2020 年实际增长率均超过预测值

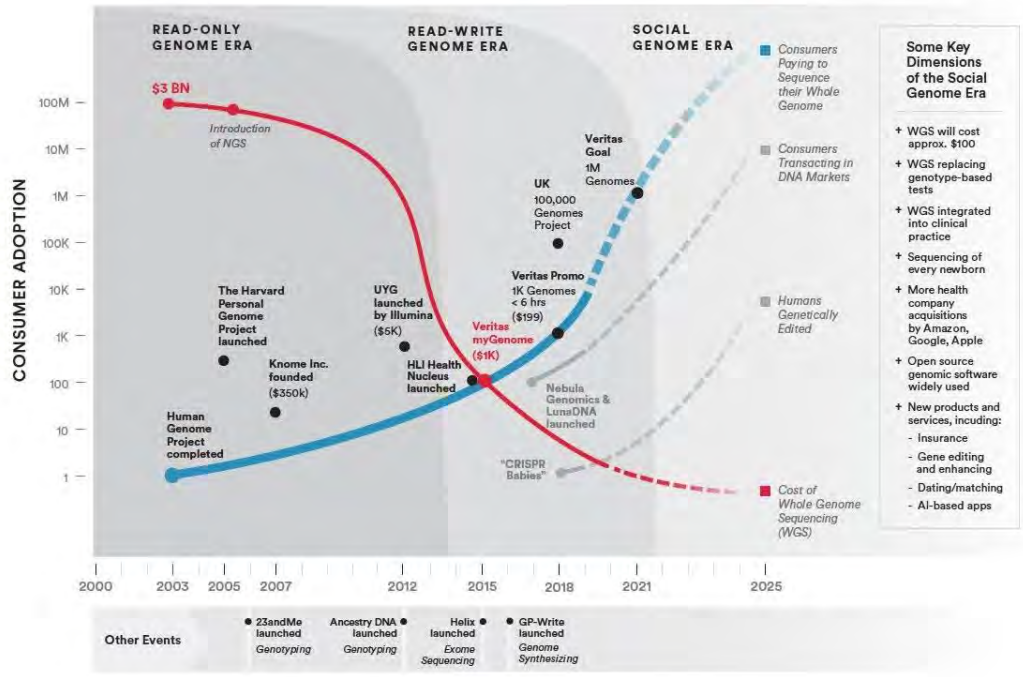


图 1.6　人类基因组研究已从读懂基因组进入书写基因组和基因组社会化时代（Veritas 公司 2019 年预测）

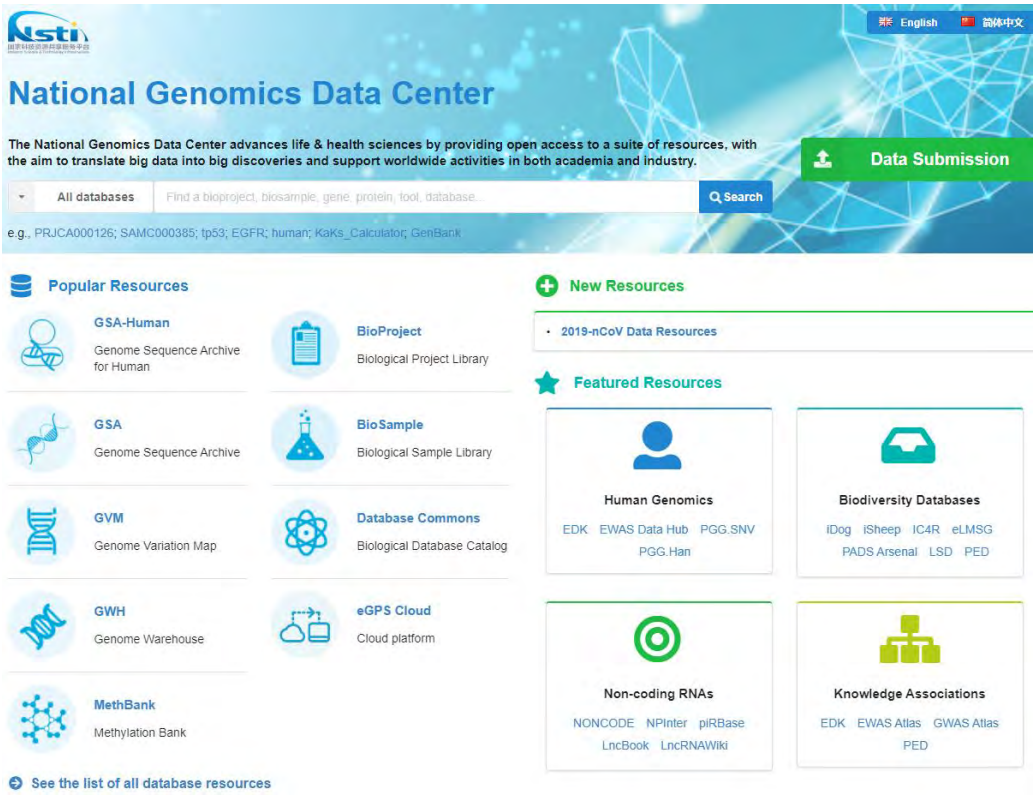


图 1.7　国家基因组科学数据中心主页（https://bigd.big.ac.cn/）

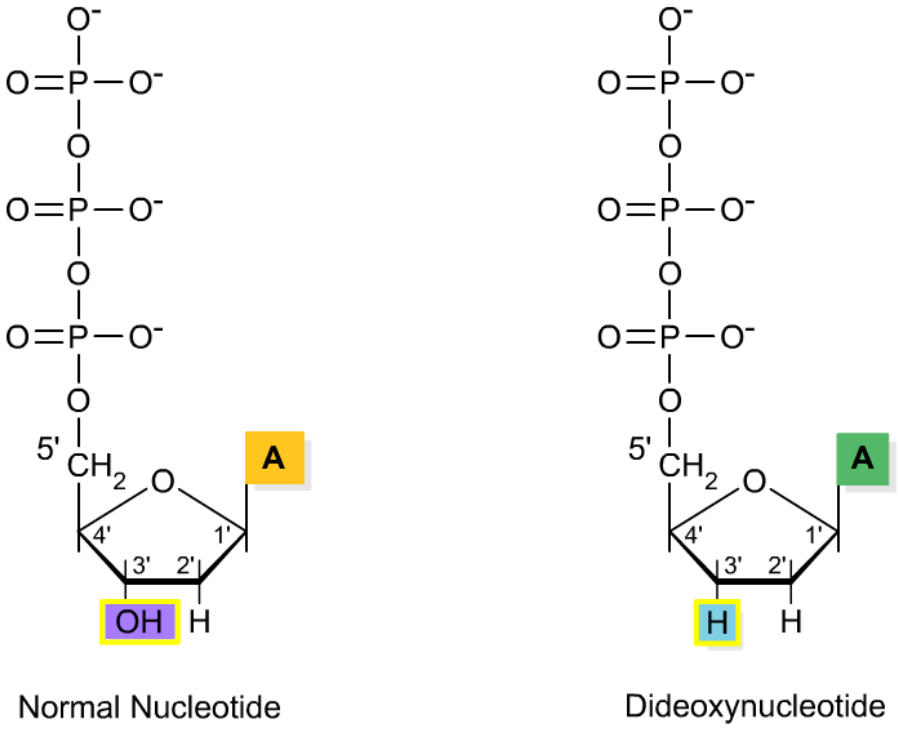


图2.1　脱氧核苷酸（左）和双脱氧核苷酸（右）

左图为正常的脱氧核苷酸（dNTP），其 3′ 位置含有羟基；右图为双脱氧核苷酸（ddNTP），其 2′ 和 3′ 位置都不含羟基

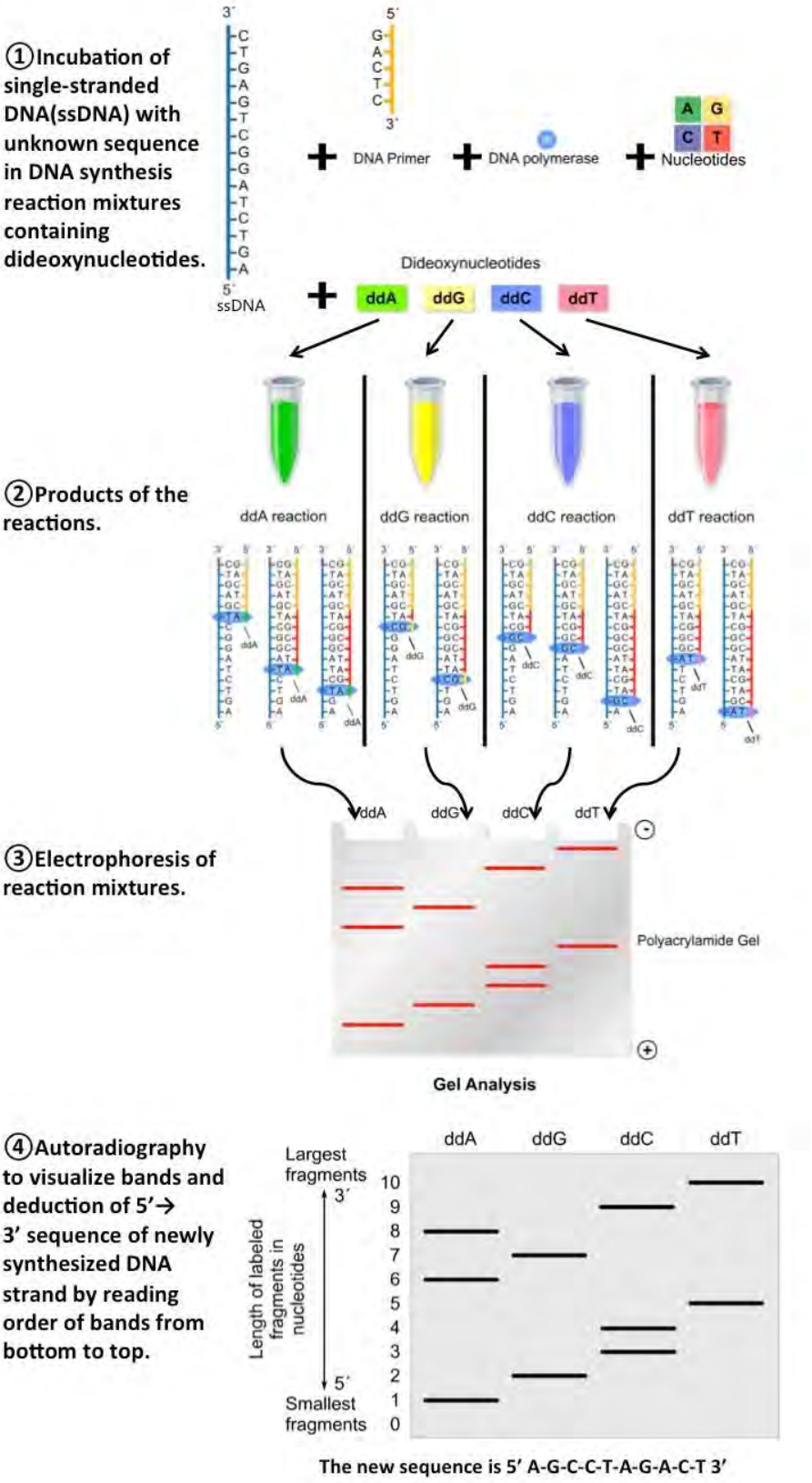


图 2.2　双脱氧链终止法（Sanger 法）测序原理

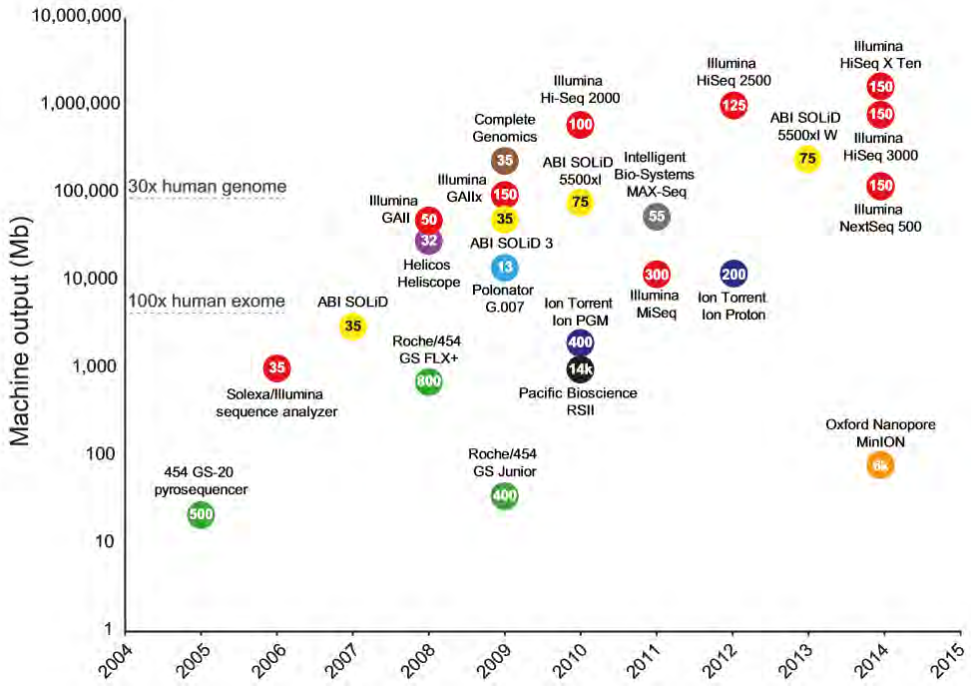


图2.3　高通量测序技术出现年代及其测序通量（引自Router et al.，　2015）

圆圈中的数字表示各测序技术产生的读序长度（nt）

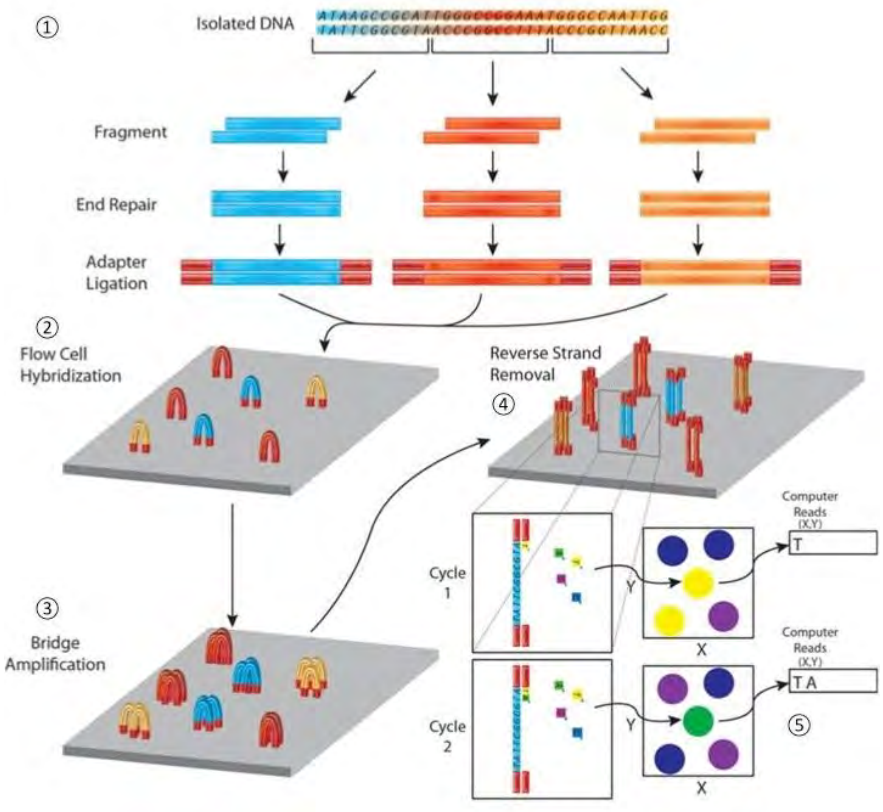


图2.4　Illumina 测序的具体流程

① DNA 文库制备；② 流动槽杂交；③ 桥式 PCR 扩增；④洗掉 DNA 反链；⑤ 测序

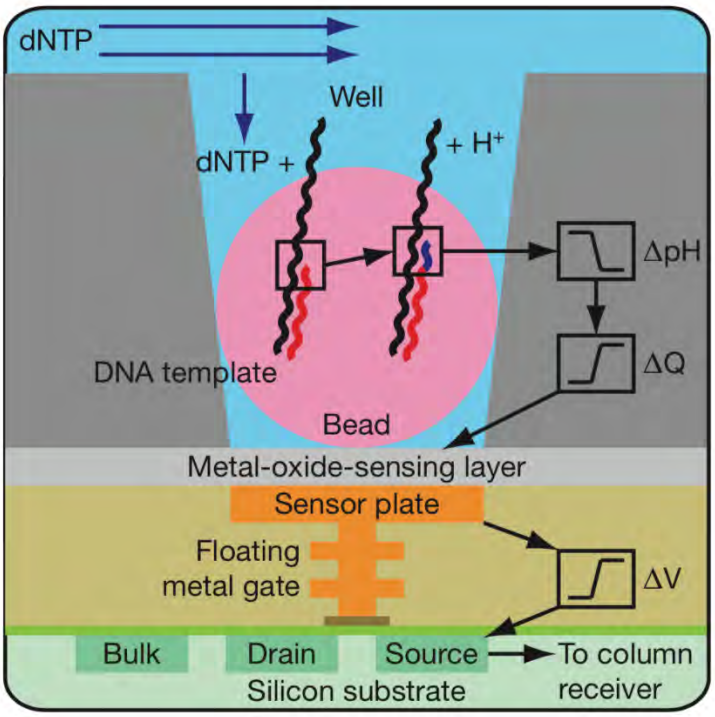


图2.5　Ion Torrent 半导体测序原理

（引自 Rothberg et al.，2011）

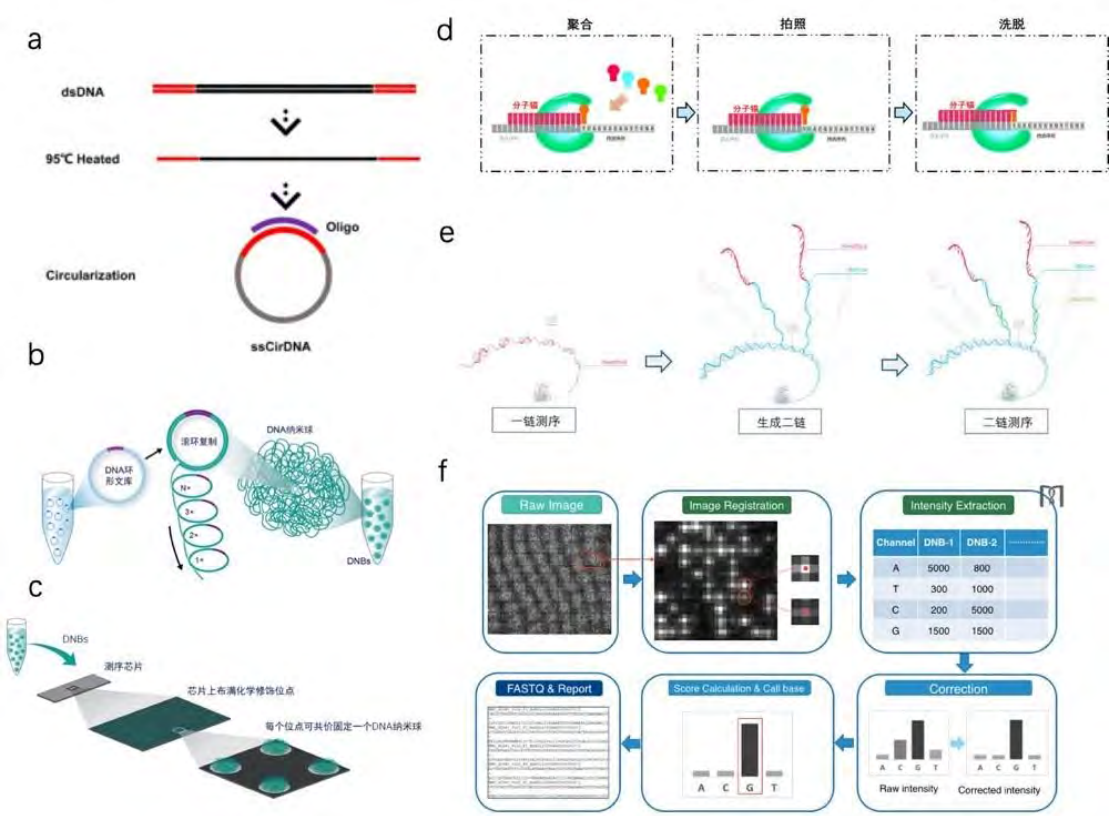


图2.6　DNA 纳米球（DNB）测序技术

A．DNA 单链环化；B．DNB 制备；C．规则阵列芯片和 DNB 加载；D．cPAS 技术；E．测序；F．碱基识别

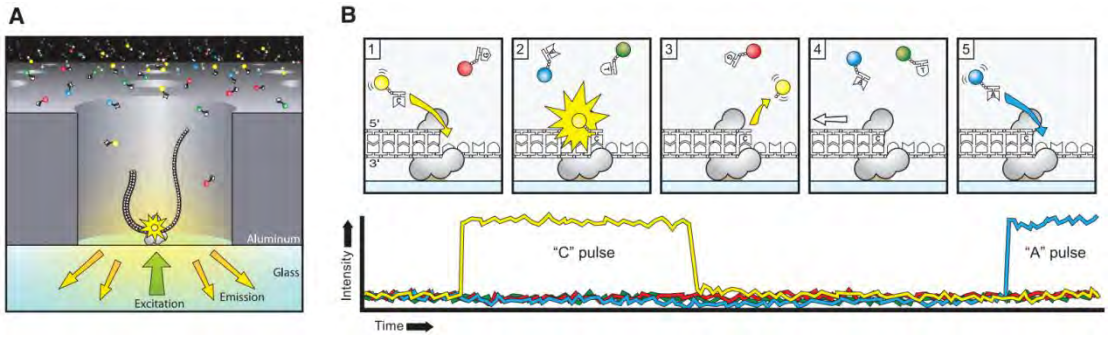


图2.7　SMRT 测序原理（引自Eid et al.，　2009）

图A 显示的是ZMW 的结构，激光进入ZMW 后呈指数级衰减，仅能照亮靠近底部的约30nm 区域，因此大部分游离的荧光标记dNTP 不会被激发，只有结合到DNA 聚合酶上的dNTP 其荧光基团被激光照亮，激发荧光；图B 显示的是DNA 合成过程中检测到的荧光信号及持续时间。结合到酶上的 dNTP 停留时间较长，信号呈脉冲式激发，因而能与噪音区分

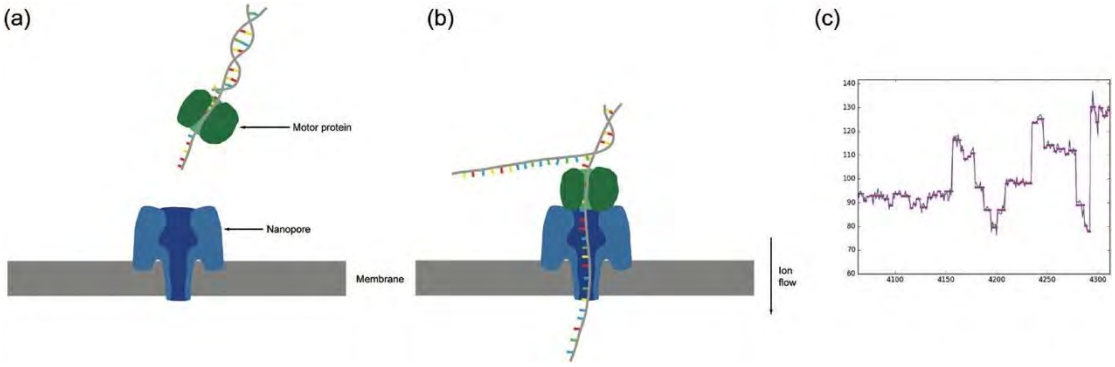


图2.8　纳米孔测序（链测序）原理（引自Leggett et al.，　2017）

A. 纳米孔嵌入在电阻合成膜中，电阻膜两侧存在施加电势差，离子流帮助核酸通过纳米孔；B. 与另一个衔接子结合的马达蛋白与纳米孔对接，并使 DNA 分子通过纳米孔；C. 纳米孔中的碱基导致电流中断，形成序列特征

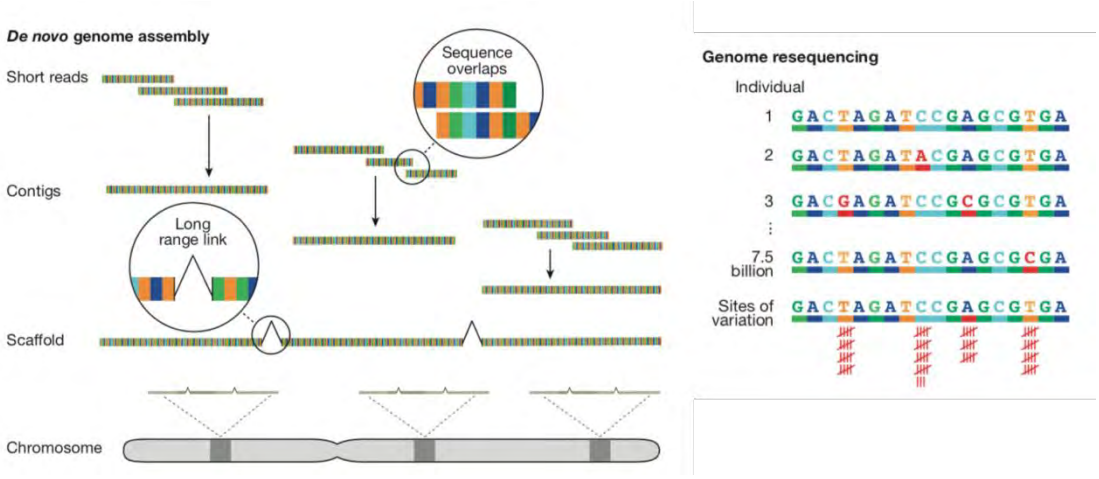


图2.9　高通量测序在基因组及基因组重测序中的相关应用（引自Shendure et al.，　2017）

contig、scaffold、染色体为基因组组装的三个层次

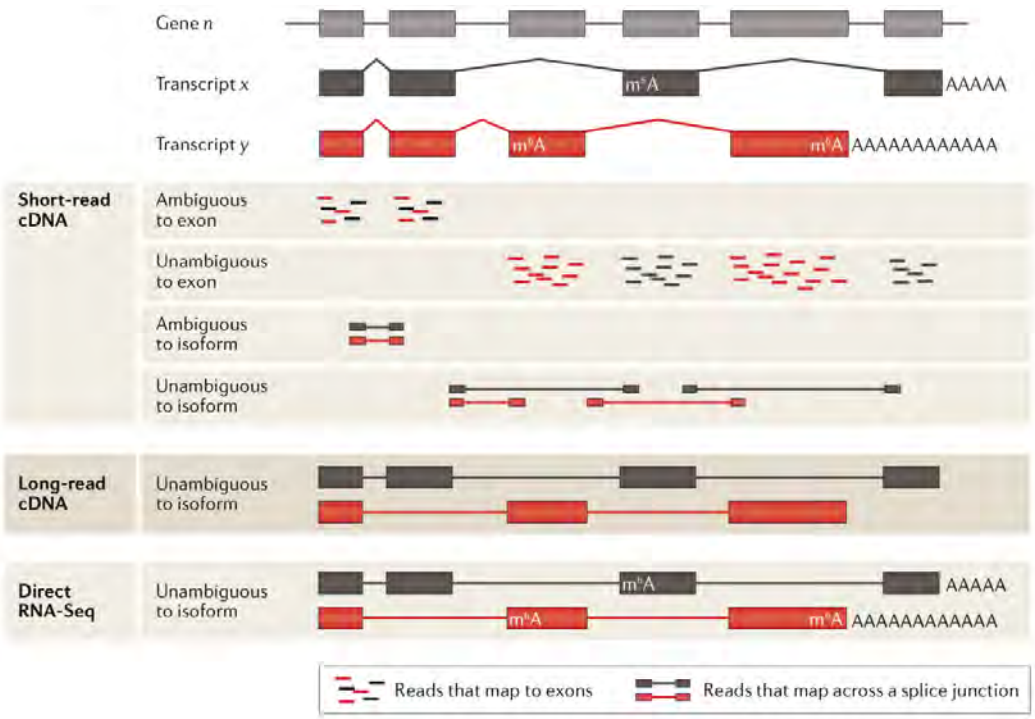


图 2.10　短读序 cDNA、长读序 cDNA 及直接 RNA 测序三类 RNA-Seq 方法比较（引自 Stark et al.， 2019）

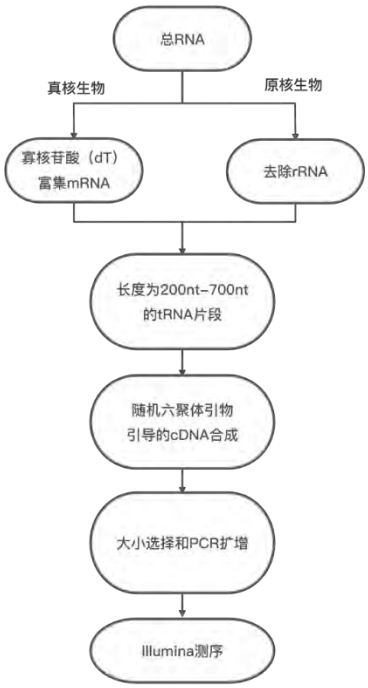


图2.11　RNA-Seq 技术流程（以Illumina 测序为例）

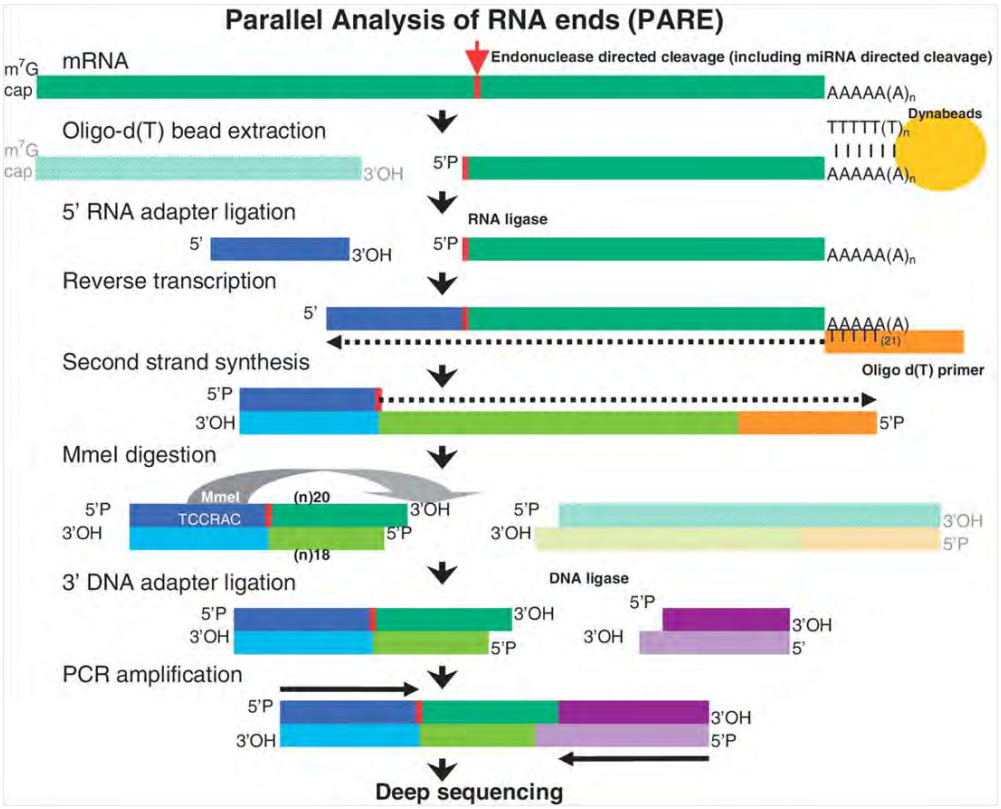


图2.12　降解组测序建库流程（引自Thomson et al.，　2011）

收集的靶基因经剪接产生2 个片段：5′ 剪接片段和3′ 剪接片段，用RNA 连接酶连接3′ 剪接片段（含有polyA 尾巴的片段），进行反转录 PCR 扩增可得到目标片段

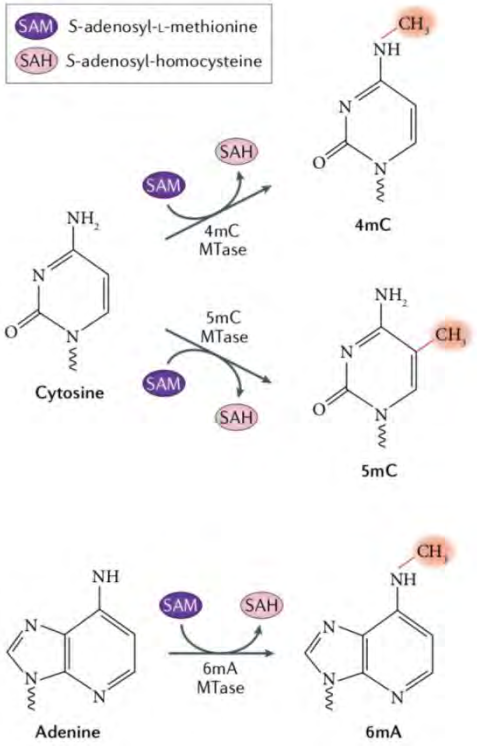


图 2.13　三种主要 DNA 甲基化修饰类型（引自 Beaulaurier et al.，　2019）

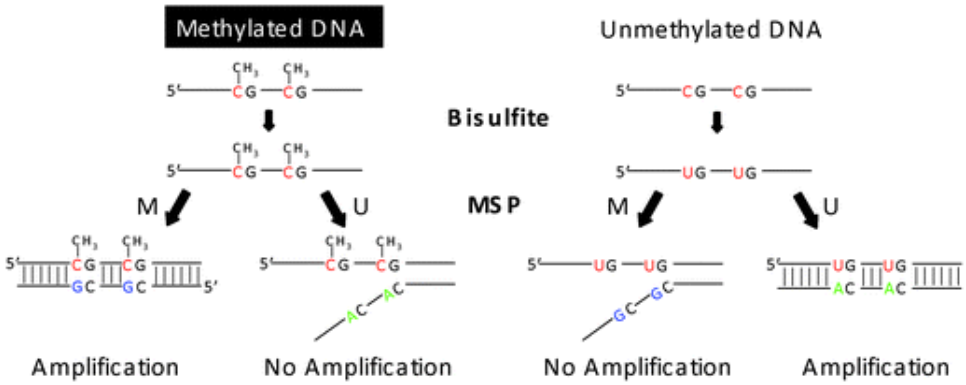


图 2.14　甲基化重亚硫酸盐测序原理（引自Zhang et al.，2009）

M（methylated specific）．甲基化特有；U（unmethylated specific）．非甲基化特有；MSP（methylation-specific PCR）．甲基化特异性 PCR

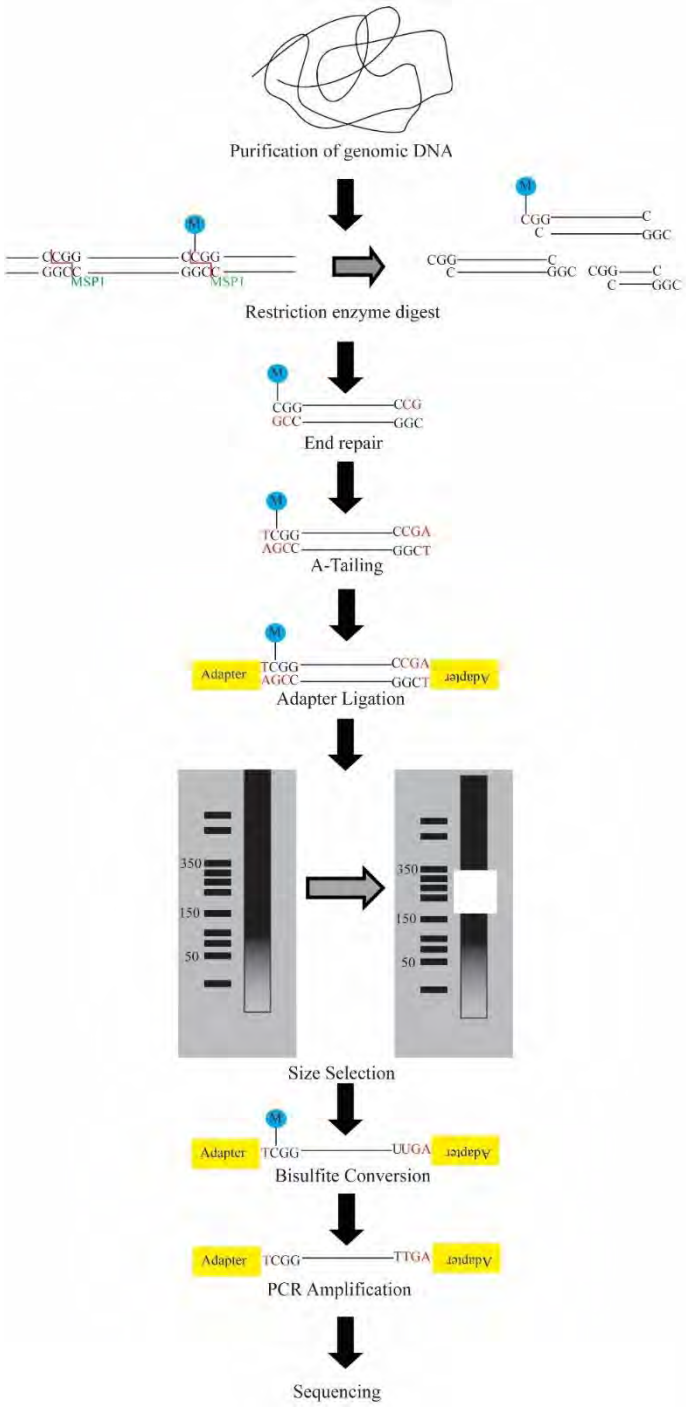


图 2.15　甲基化简化重亚硫酸盐测序（RRBS）原理（引自 Smith et al.，2009）

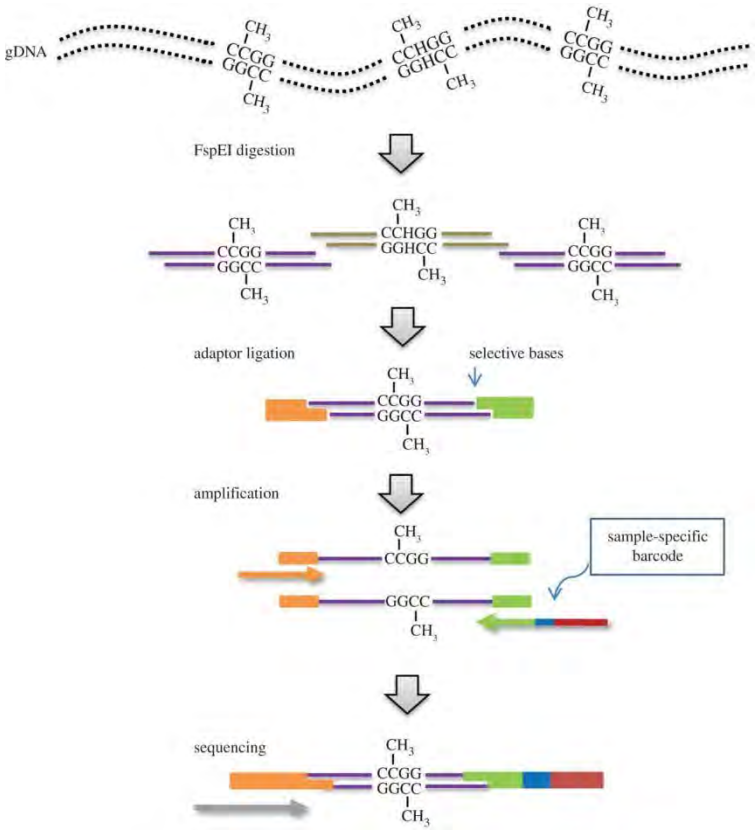


图 2.16　MethylRAD-Seq 技术测序原理（引自 Wang et al.， 2015）

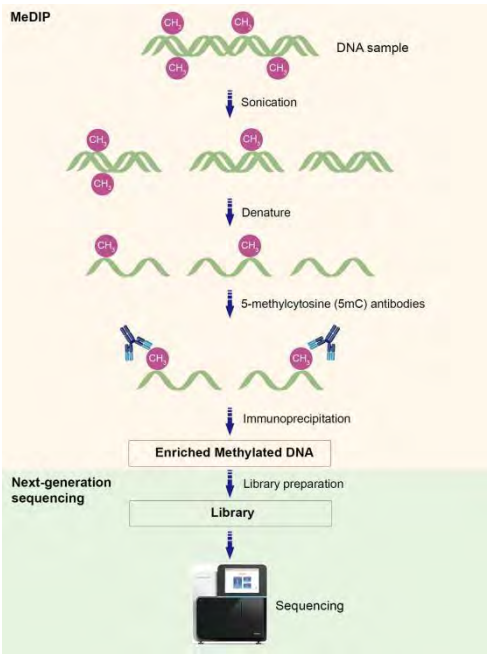


图2.17　MeDIP-Seq 方法工作流程
（www.cd-genomics.com）

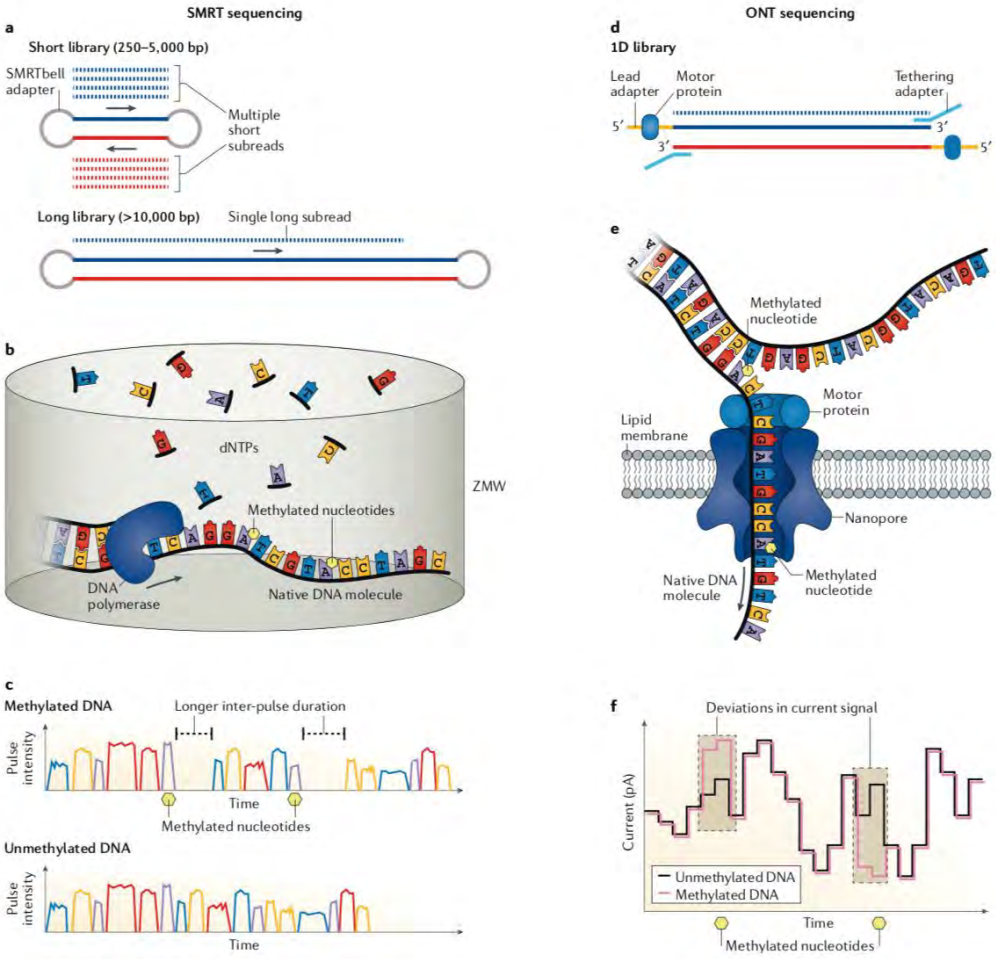


图 2.18　基于第三代测序的天然 DNA 甲基化检测技术（ 引自 Beaulaurier et al.， 2019）

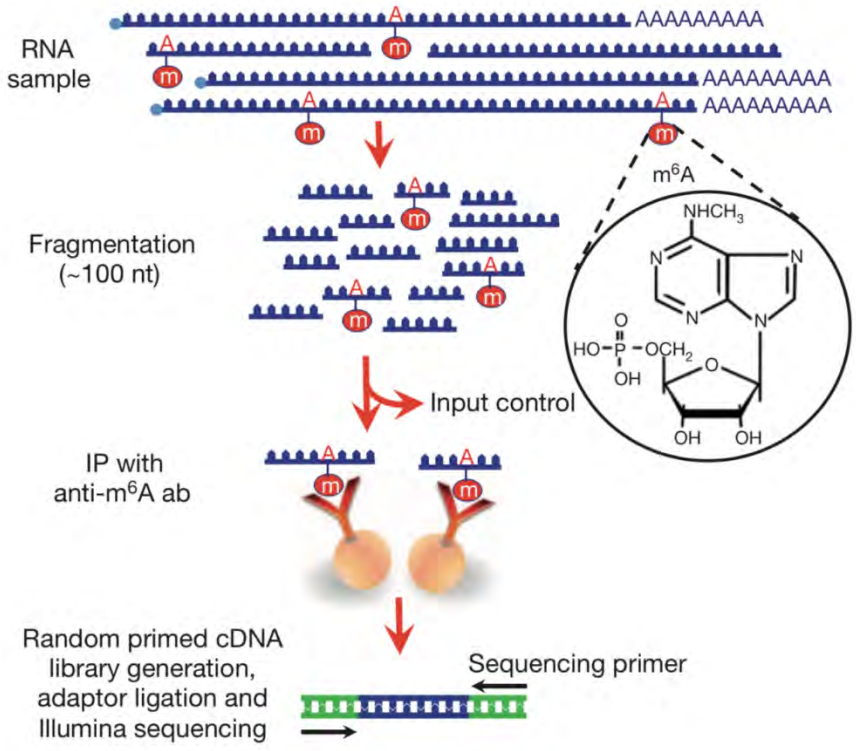


图2.19　MeRIP-Seq 测序原理（引自Dominissini et al.， 2012）

正常的免疫沉淀（IP）实验条件，添加了m6A 抗体的实验体系。通常情况下测序读段大致分布在甲基化位点附近，而对照条件下测序读段的分布与正常的每个基因的表达值正相关（没有甲基化影响）

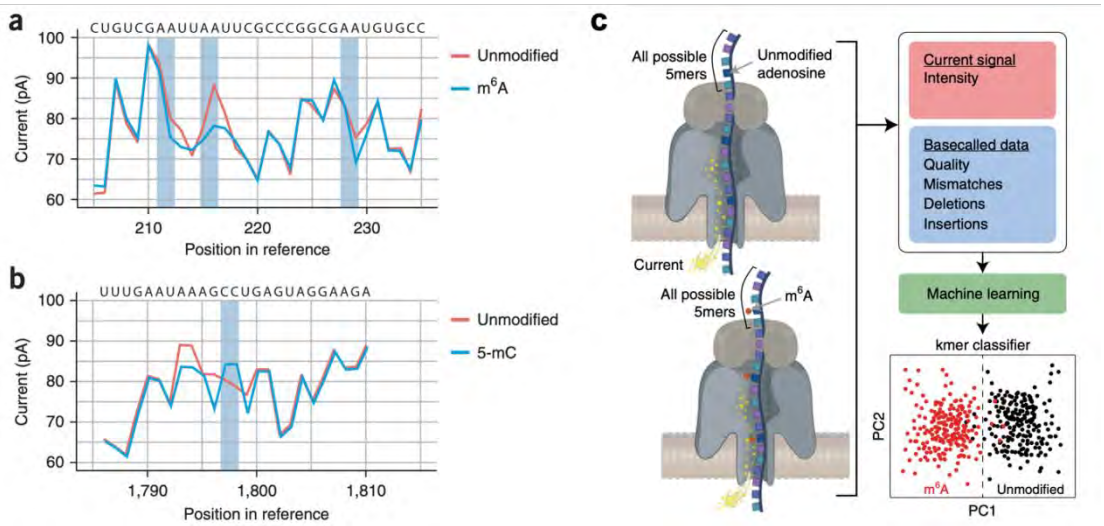


图2.20　Nanopore 直接RNA 测序鉴定碱基修饰模式图（引自Liu et al.， 2019； Garalde et al.， 2018）

A、B. 直接RNA 测序鉴定m6A 和5-mC 碱基修饰时的电信号差别；C. 直接RNA 测序鉴定碱基修饰的算法模式图，PC1 和 PC2 分别为主成分（PC）分析第 1 和第 2 主成分

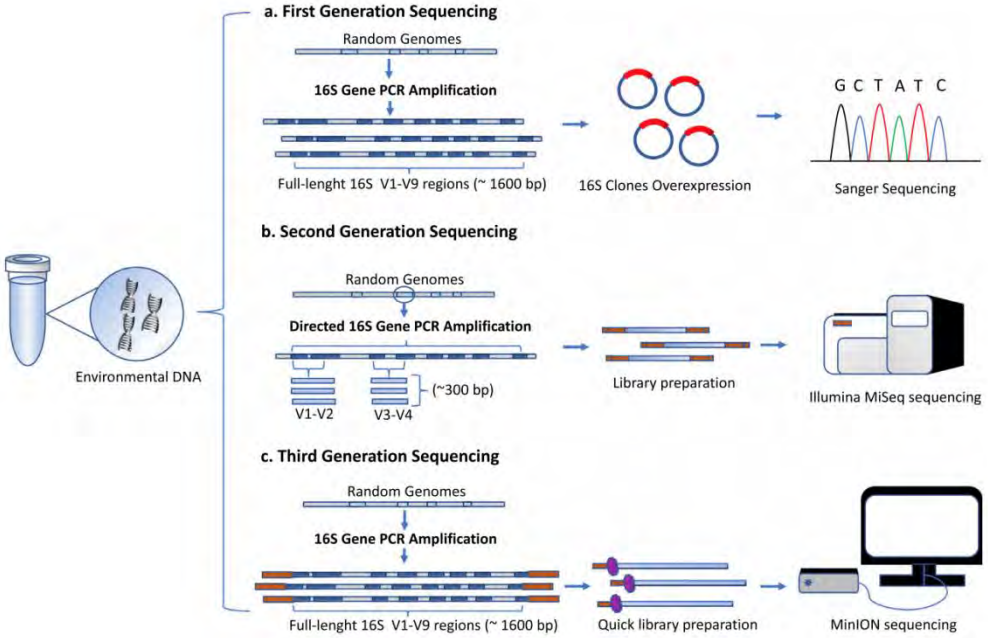


图 2.21　16S rDNA 常用测序方法及其比较（引自 Santos et al.， 2020）

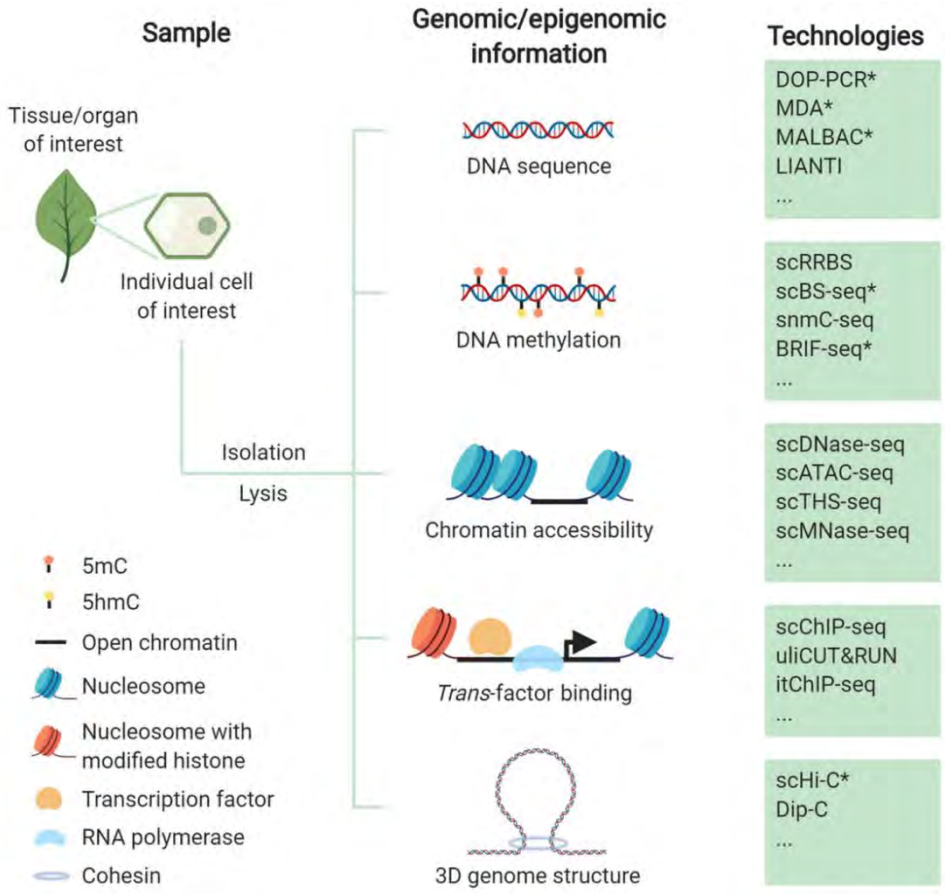


图 2.22　单细胞基因组和 DNA 甲基化测序技术平台（引自 Luo et al.， 2020）

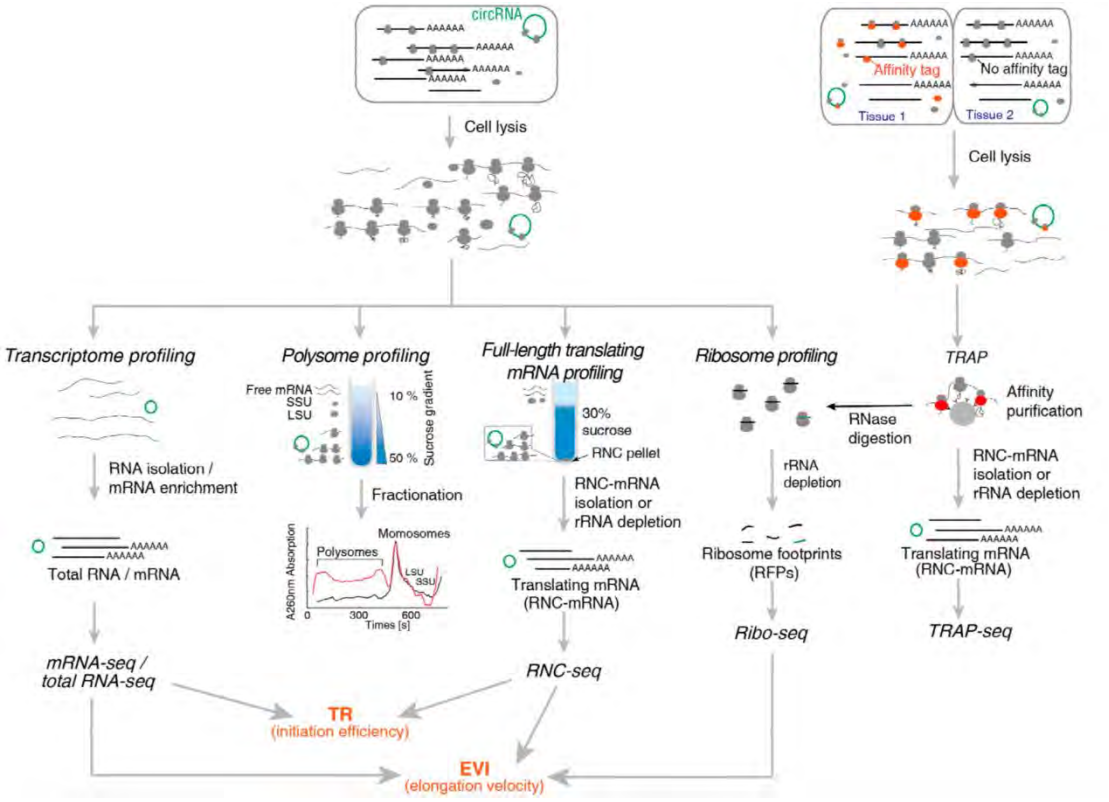


图2.23　翻译中的RNA 主要测定方法（引自Zhao et al.， 2019）

RNC．核糖体新生肽链复合物；TRAP．正在翻译的核糖体亲和纯化

```
>gi|332309159|ref|NM_181652.2| Homo sapiens peroxiredoxin 5 (PRDX5), nuclear gene encoding
mitochondrial protein, transcript variant 3, mRNA
CGCGCCTGCGCAGTGGAGGCGGCCCAGGCCCGCCTTCCGCAGGGTGTCGCCGCTGTGCCGCTAGCGGTGC
CCCGCCTGCTGCGGTGGCACCAGCCAGGAGGCGGAGTGGAAGTGGCCGTGGGGCGGGTATGGGACTAGCT
GGCGTGTGCGCCCTGAGACGCTCAGCGGGCTATATACTCGTCGGTGGGGCCGGCGGTCAGTCTGCGGCAG
CGGCAGCAAGACGGTGCAGTGAAGGAGAGTGGGCGTCTGGCGGGGTCCGCAGTTTCAGCAGAGCCGCTGC
AGCCATGGCCCCAATCAAGGTTCGGCTCCTGGCTGATCCCACTGGGGCCTTTGGGAAGGAGACAGACTTA
TTACTAGATGATTCGCTGGTGTCCATCTTTGGGAATCGACGTCTCAAGAGGTTCTCCATGGTGGTACAGG
ATGGCATAGTGAAGGCCCTGAATGTGGAACCAGATGGCACAGGCCTCACCTGCAGCCTGGCACCCAATAT
CATCTCACAGCTCTGAGGCCCTGGGCCAGATTACTTCCTCCACCCCTCCCTATCTCACCTGCCCAGCCCT
GTGCTGGGGCCCTGCAATTGGAATGTTGGCCAGATTTCTGCAATAAACACTTGTGGTTTGCGGCCATCTC
CTTGGTTAAAAAAAAA
```

图 3.1　FASTA 格式

```
@ST-E00169:117:HG2T3CCXX:4:1101:23794:1432 1:N:0:NTGTCA
NTCTAAGAATTCAATGGCTTTAAGAAGGCTTAGATAGCGAGTATTCTCAGAATTCAACGATATCTTCGCATTATCGGTTACTAGACG
+
#AA<FFFKKFKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKFKKKKFFFA<FKKKKKKFFAKKFKKFKKK
```

图 3.2　FASTQ 格式

```
##gff-version 3
##sequence-region ctg123 1 1497228
seqid  source  type start  end score strand phase attributes
ctg123 . gene 1000 9000 . + . ID=gene00001;Name=EDEN
ctg123 . TF_binding_site 1000 1012 . + . Parent=gene00001
ctg123 . mRNA 1050 9000 . + . ID=mRNA00001;Parent=gene00001
ctg123 . mRNA 1050 9000 . + . ID=mRNA00002;Parent=gene00001
ctg123 . mRNA 1300 9000 . + . ID=mRNA00003;Parent=gene00001
ctg123 . exon 1300 1500 . + . Parent=mRNA00003
ctg123 . exon 1050 1500 . + . Parent=mRNA00001,mRNA00002
```

图 3.3　GFF 格式

```
LOCUS       SCU49845               5028 bp    DNA     linear   PLN 29-OCT-2018
DEFINITION  Saccharomyces cerevisiae TCP1-beta gene, partial cds; and Axl2p
            (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION   U49845
VERSION     U49845.1
KEYWORDS    .
SOURCE      Saccharomyces cerevisiae (baker's yeast)
  ORGANISM  Saccharomyces cerevisiae
            Eukaryota; Fungi; Dikarya; Ascomycota; Saccharomycotina;
            Saccharomycetes; Saccharomycetales; Saccharomycetaceae;
            Saccharomyces.
REFERENCE   1  (bases 1 to 5028)
  AUTHORS   Roemer,T., Madden,K., Chang,J. and Snyder,M.
  TITLE     Selection of axial growth sites in yeast requires Axl2p, a novel
            plasma membrane glycoprotein
  JOURNAL   Genes Dev. 10 (7), 777-793 (1996)
   PUBMED   8846915
REFERENCE   2  (bases 1 to 5028)                                  描述部分
  AUTHORS   Roemer,T.
  TITLE     Direct Submission
  JOURNAL   Submitted (22-FEB-1996) Biology, Yale University, New Haven, CT
            06520, USA

FEATURES             Location/Qualifiers
     source          1..5028
                     /organism="Saccharomyces cerevisiae"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:4932"
                     /chromosome="IX"
     mRNA            <1..>206
                     /product="TCP1-beta"
     CDS             <1..206                                      注释部分
                     /codon_start=3
                     /product="TCP1-beta"
                     /protein_id="AAA98665.1"
                     /translation="SSIYNGISTSGLDLNNGTIADMRQLGIVESYKLKRAVVSSASEA
                     AEVLLRVDNIIRARPRTANRQHM"
     gene            <687..>3158
                     /gene="AXL2"

ORIGIN
        1 gatcctccat atacaacggt atctccacct caggtttaga tctcaacaac ggaaccattg
       61 ccgacatgag acagttaggt atcgtcgaga gttacaagct aaaacgagca gtagtcagct
      121 ctgcatctga agccgctgaa gttctactaa gggtggataa catcatccgt gcaagaccaa
      ...
     4981 tgccatgact cagattctaa ttttaagcta ttcaatttct ctttgatc
//                                                              序列部分
```

图3.4　GenBank 数据库记录格式（GBFF 格式）

A．描述部分；B．注释部分；C．序列部分

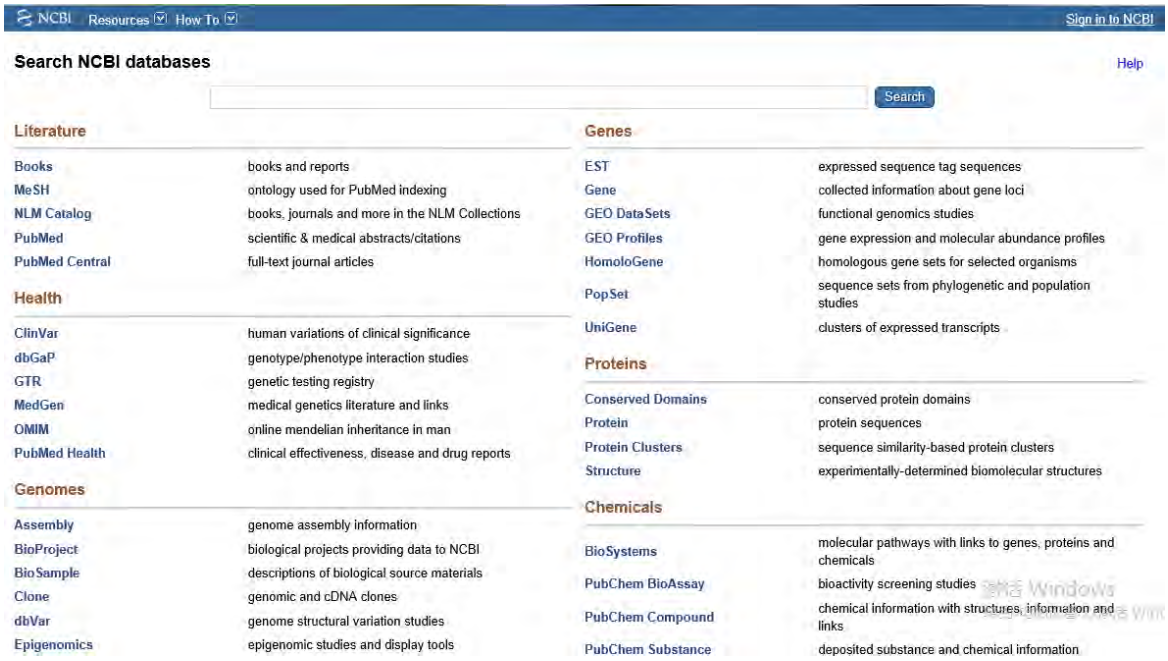
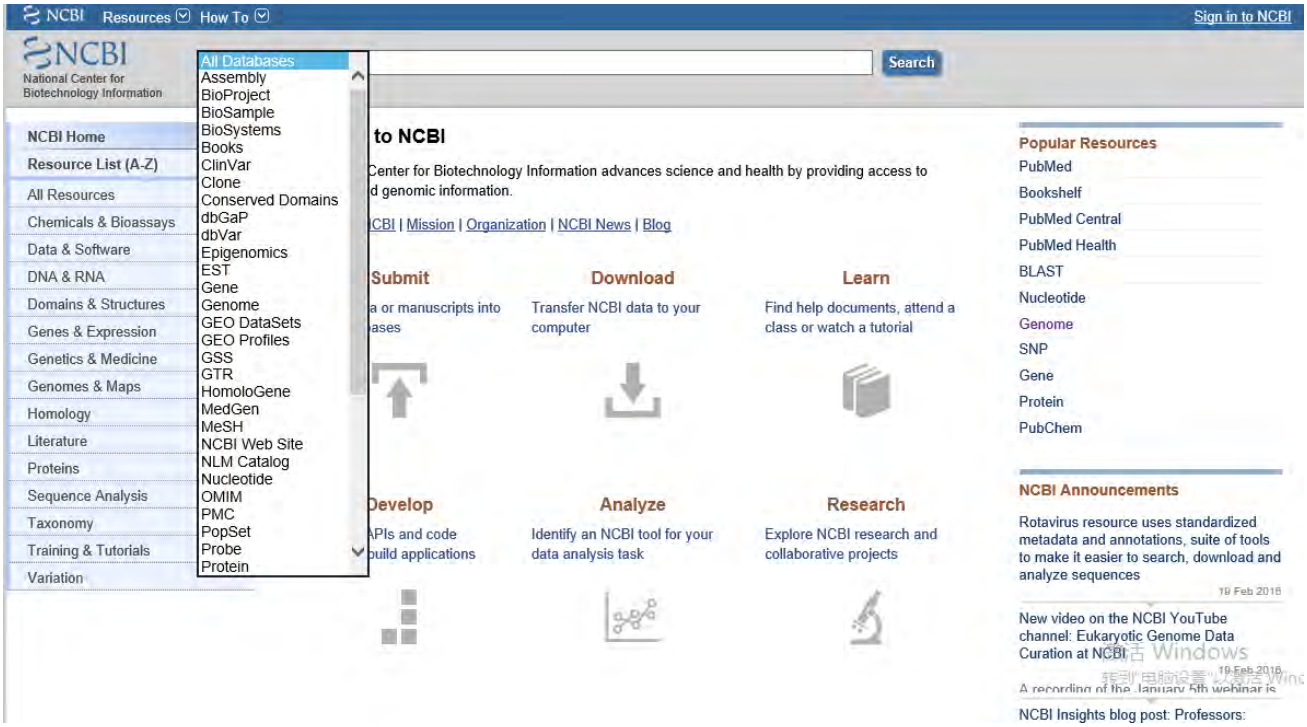


图 3.5　NCBI 数据库 Entrez 搜索首页

A



B

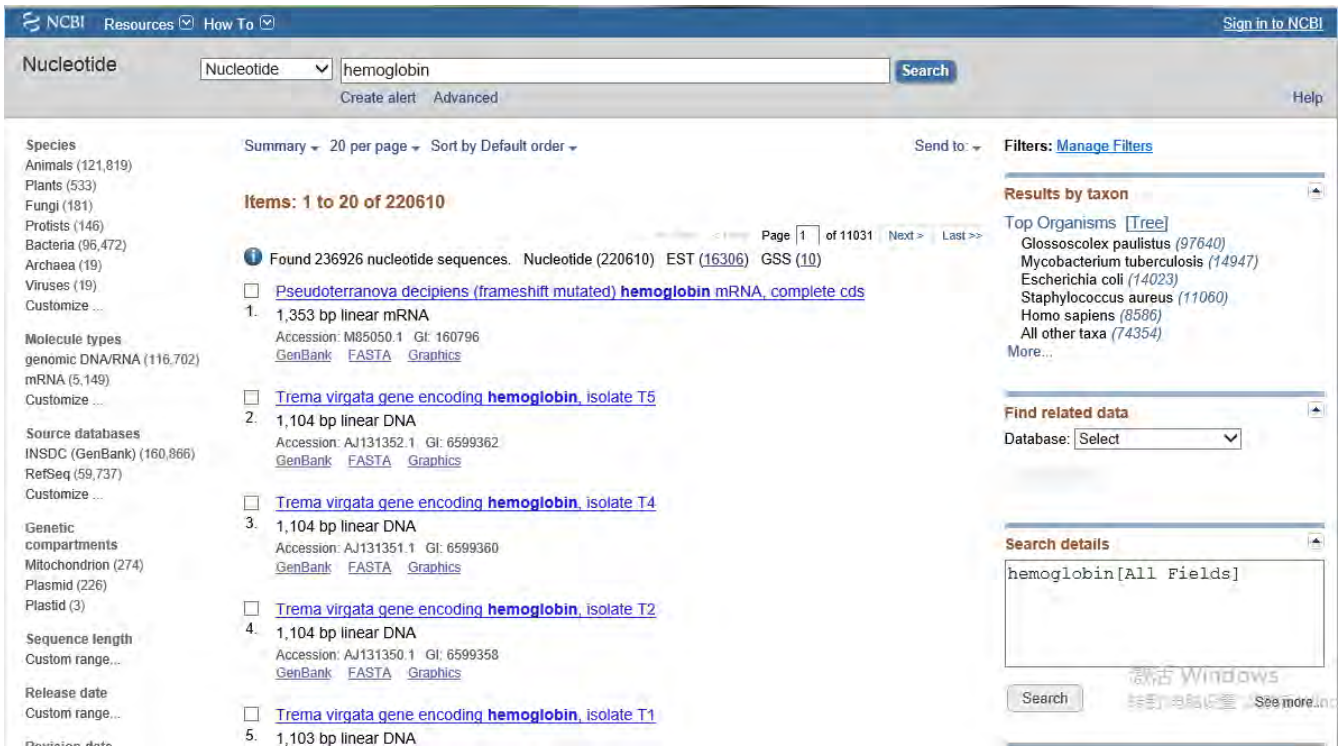


图 3.6　NCBI 数据库选项下拉界面（A）及 Gen Bank 数据库搜索界面（B）

图 3.7　GSA 核苷酸序列数据库主页

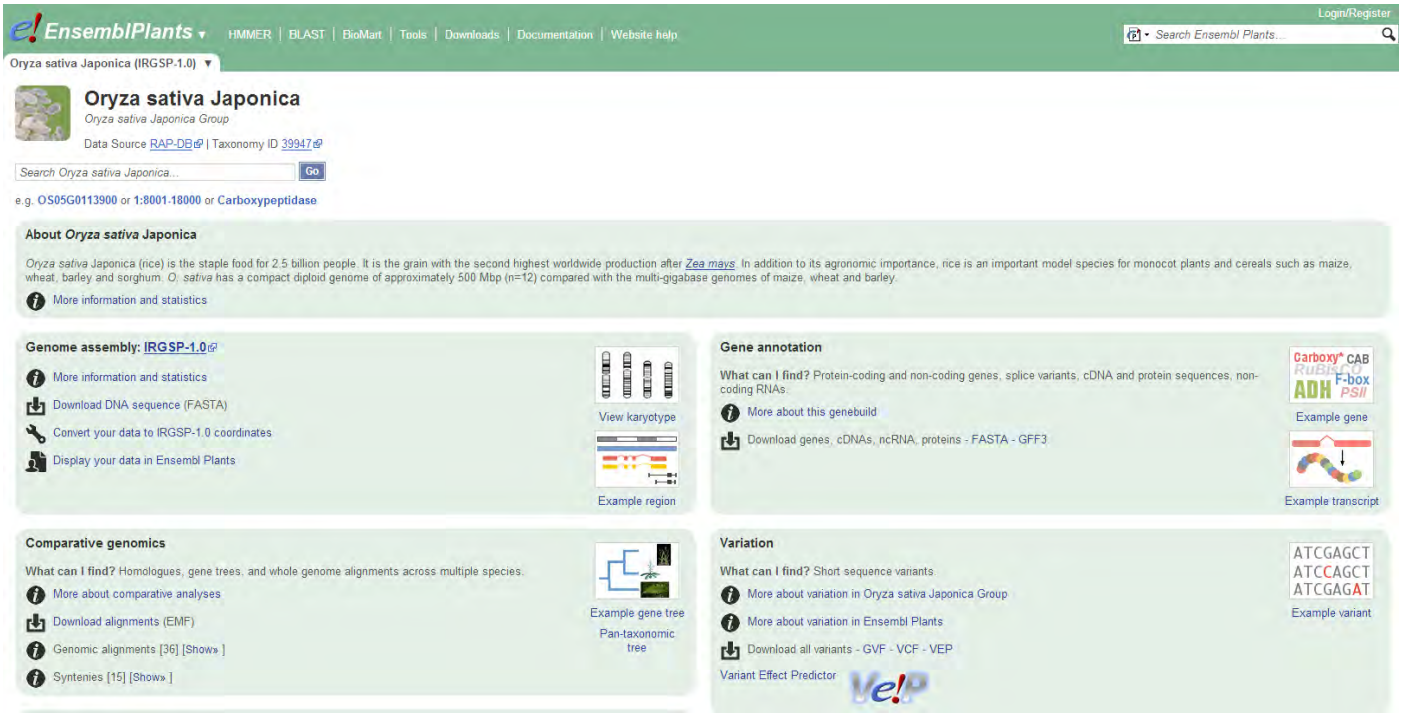
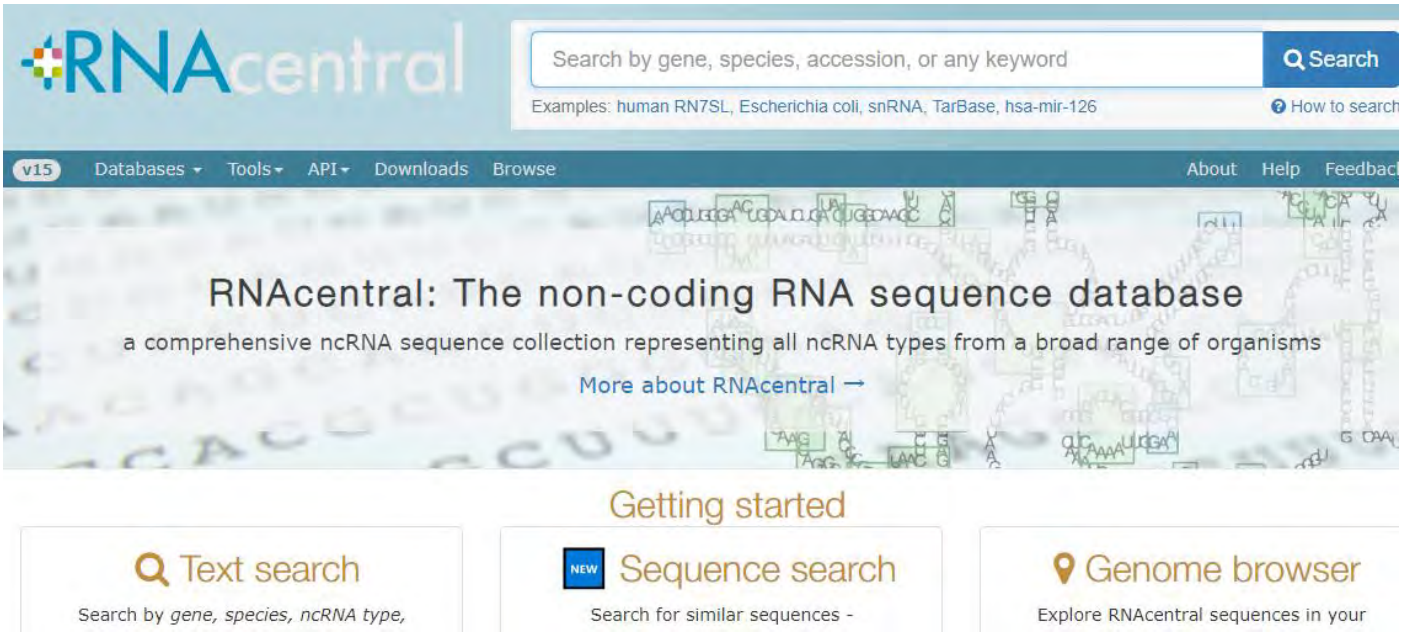


图 3.8　Ensembl Genomes 水稻基因组数据库主页

A



B

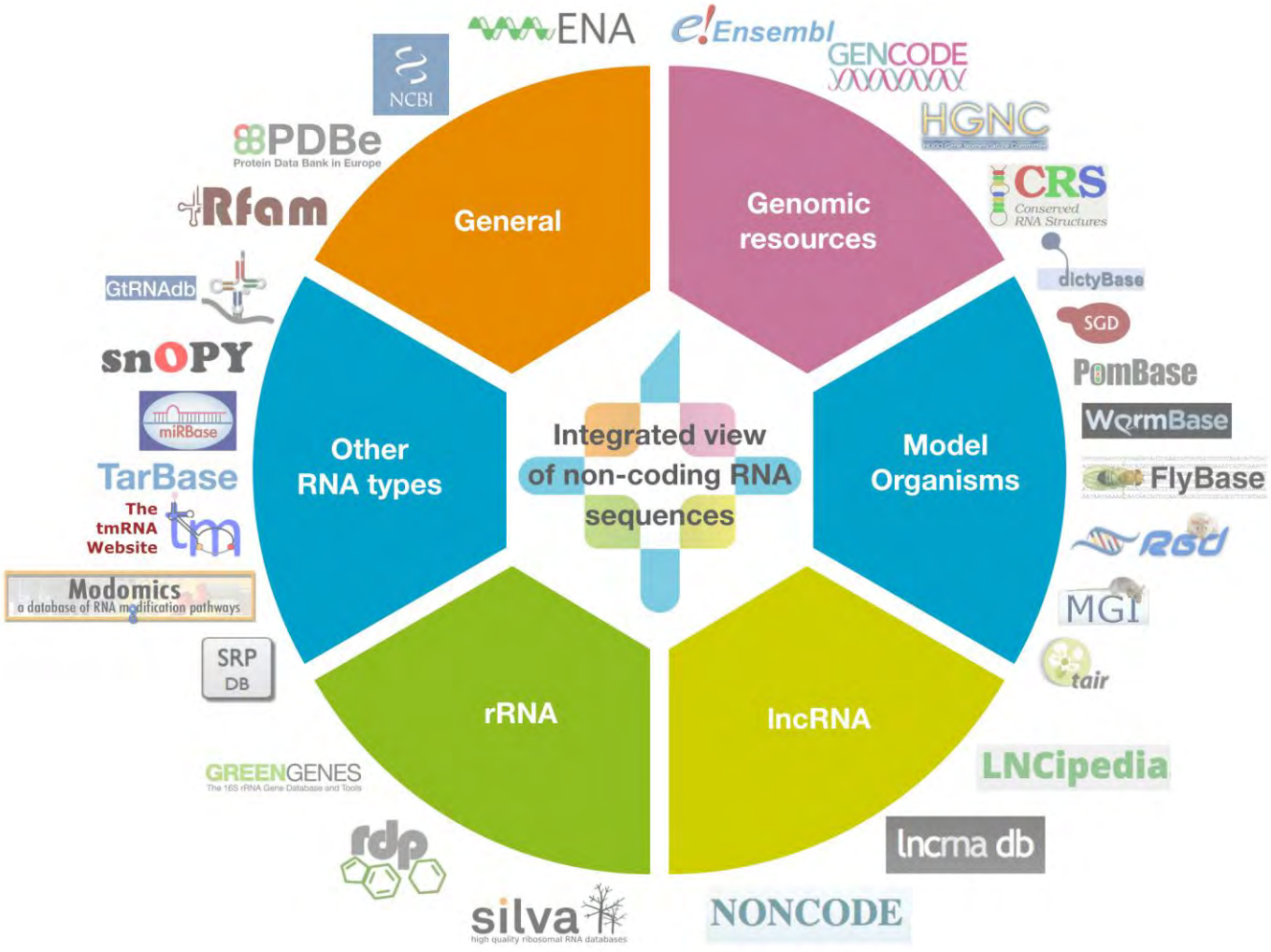


图3.9　欧洲生物信息学研究所（EBI）维护的RNAcentral
数据库主页（A）和收录的相关数据库情况（B）

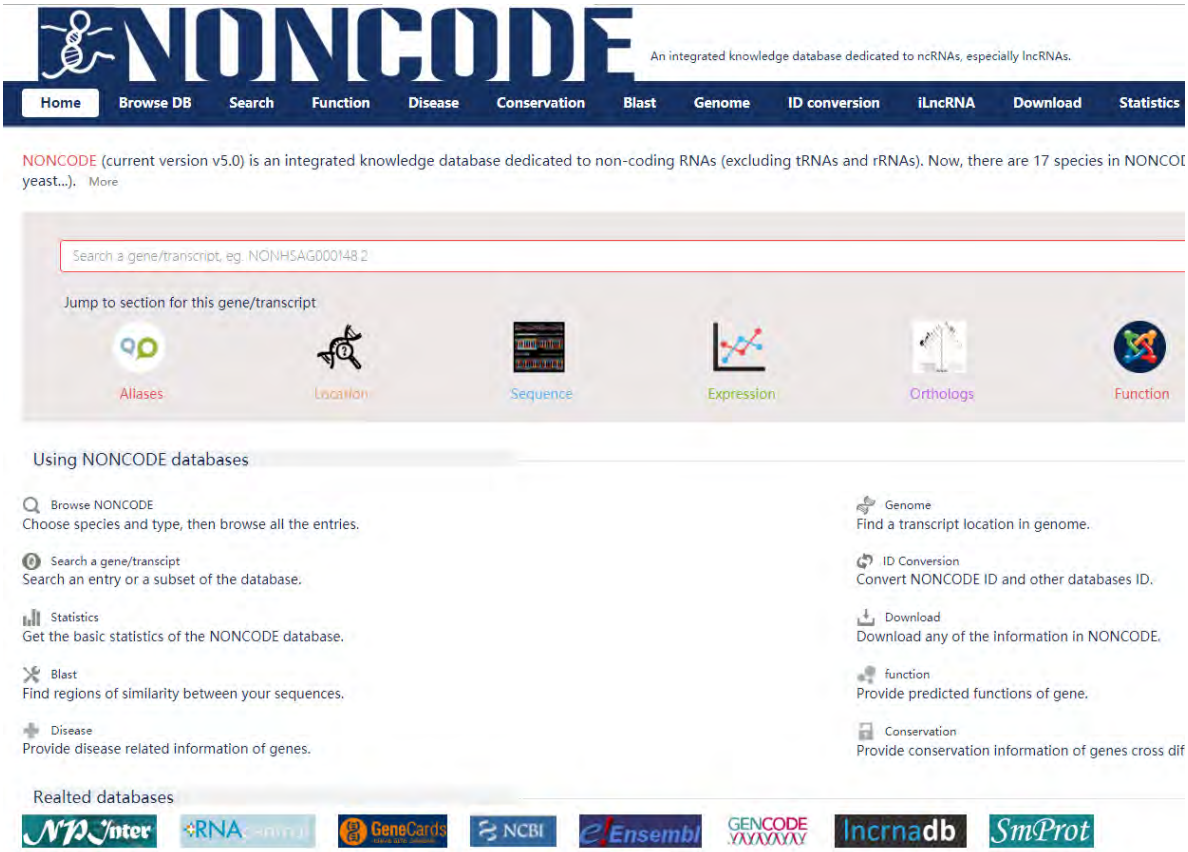


图 3.10　非编码 RNA 数据库 NONCODE 主页

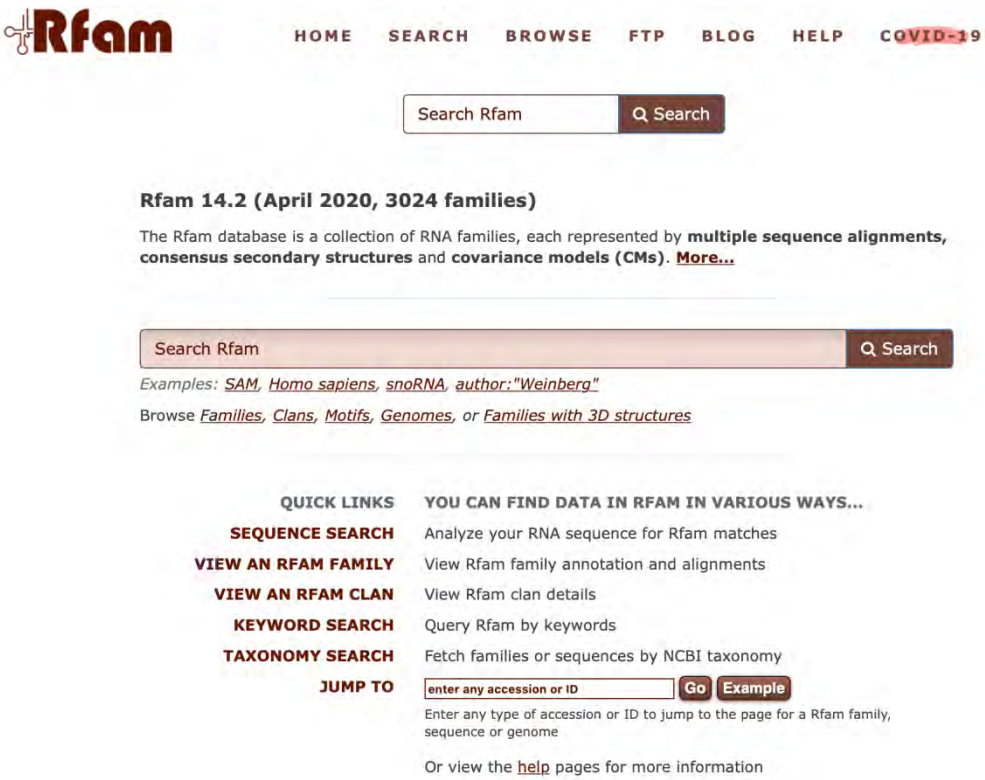


图 3.11　非编码 RNA 家族数据库 Rfam 主页

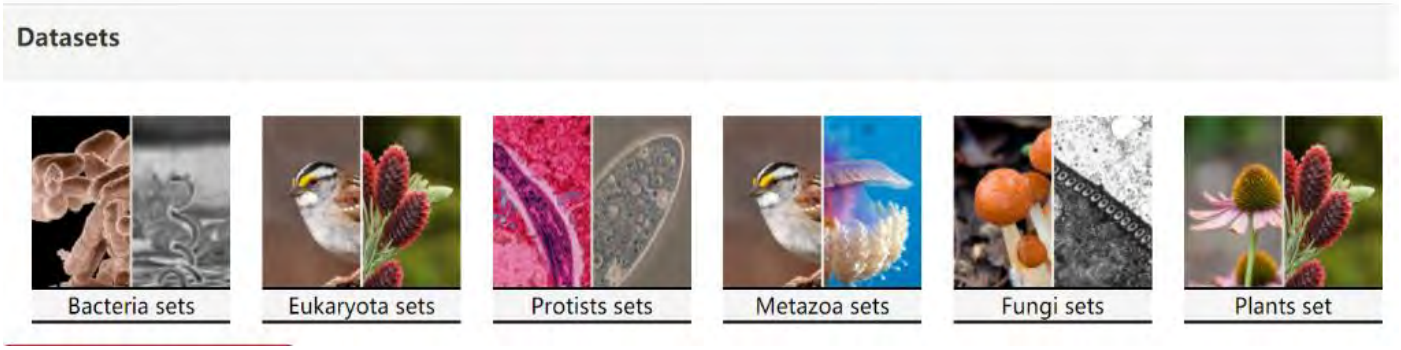


图 3.12　不同类生物同源保守基因数据库 BUSCO 主页

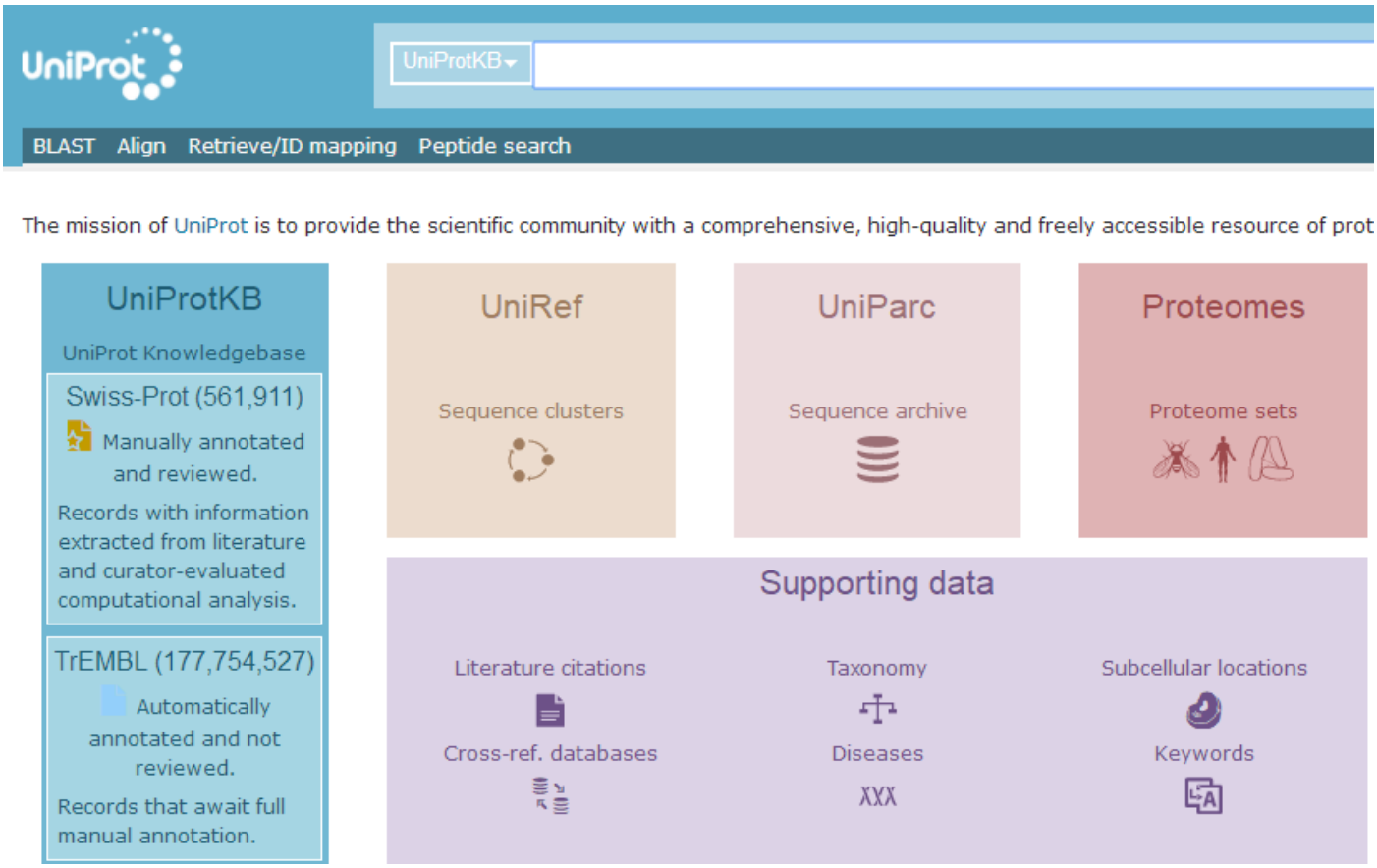


图 3.13　蛋白质数据库 UniProt（Swiss-Prot＋TrEMBL）数据库主页

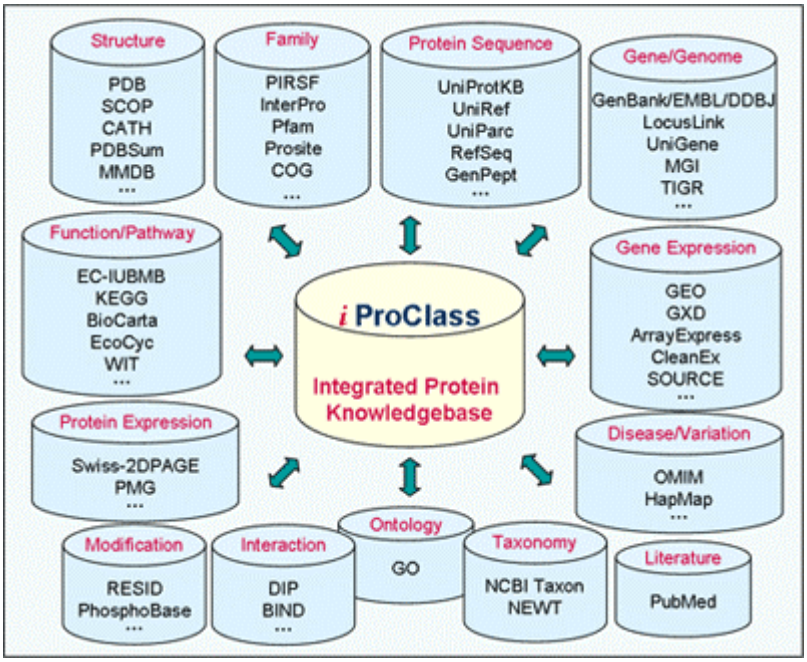


图 3.14　美国 PIR 蛋白质综合数据库 iProClass 整合了各种类型数据资源

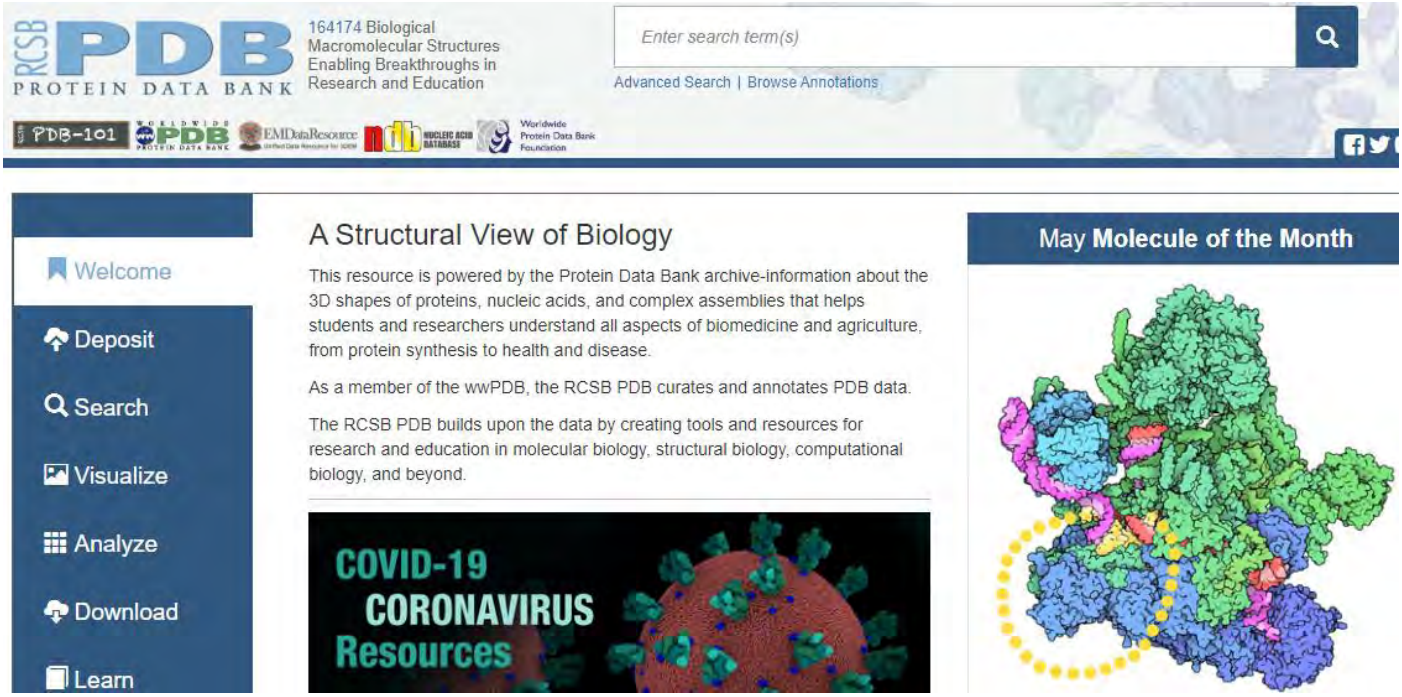


图 3.15　实验测定蛋白质结构数据库 PDB 主页

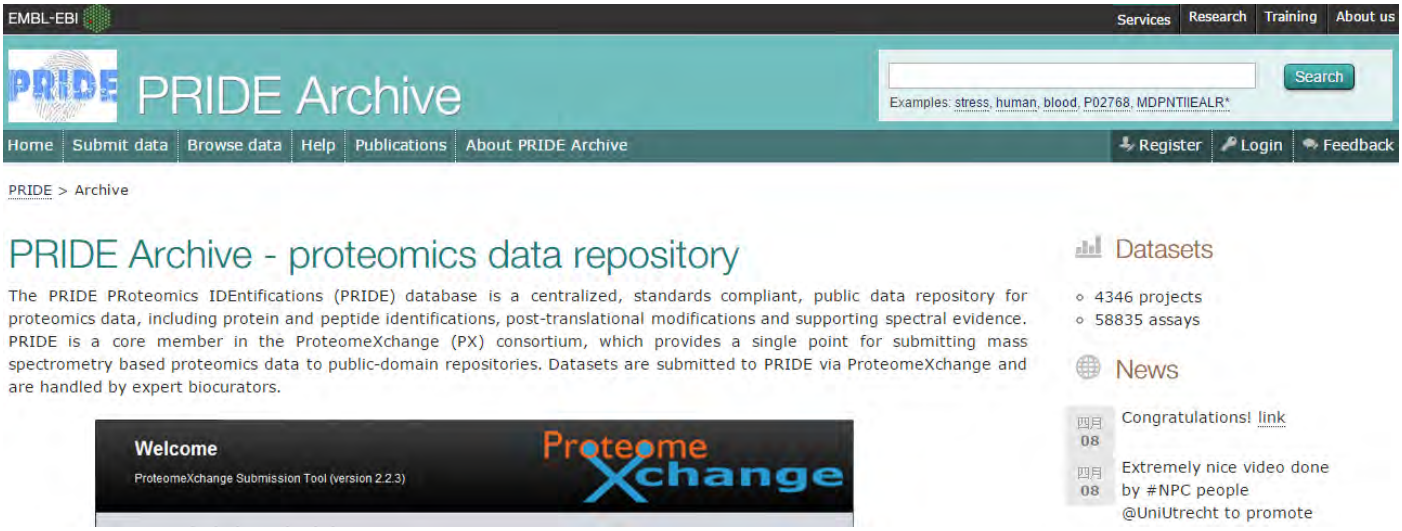


图 3.16　蛋白质组数据库 PRIDE 主页

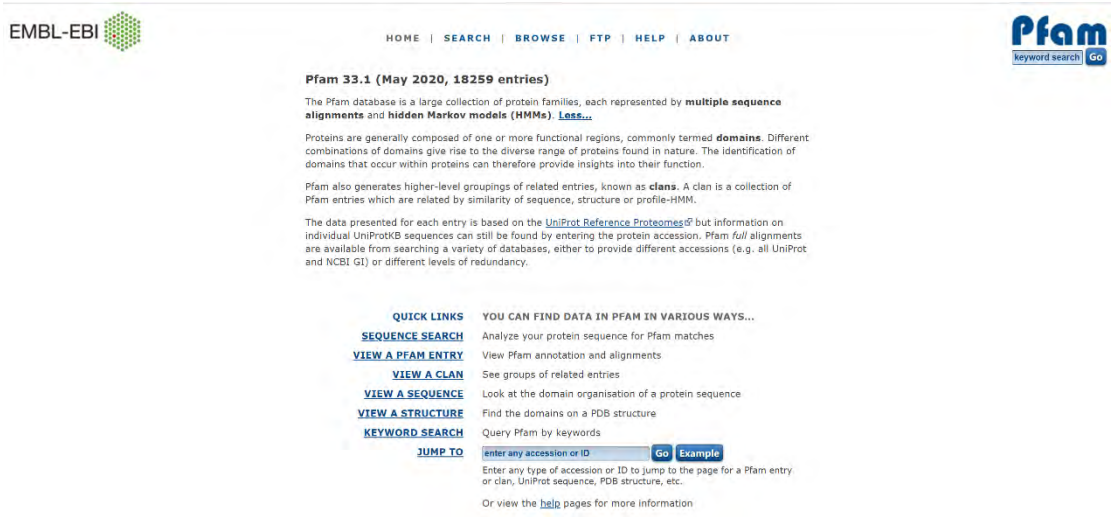


图 3.17　功能域数据库 Pfam 主页

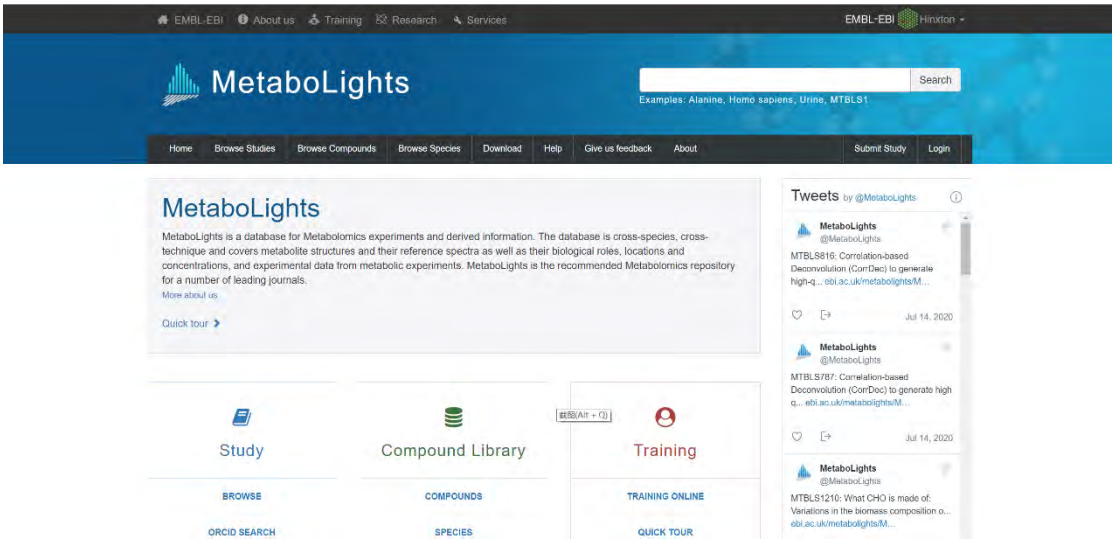


图 3.18　MetaboLights 主页

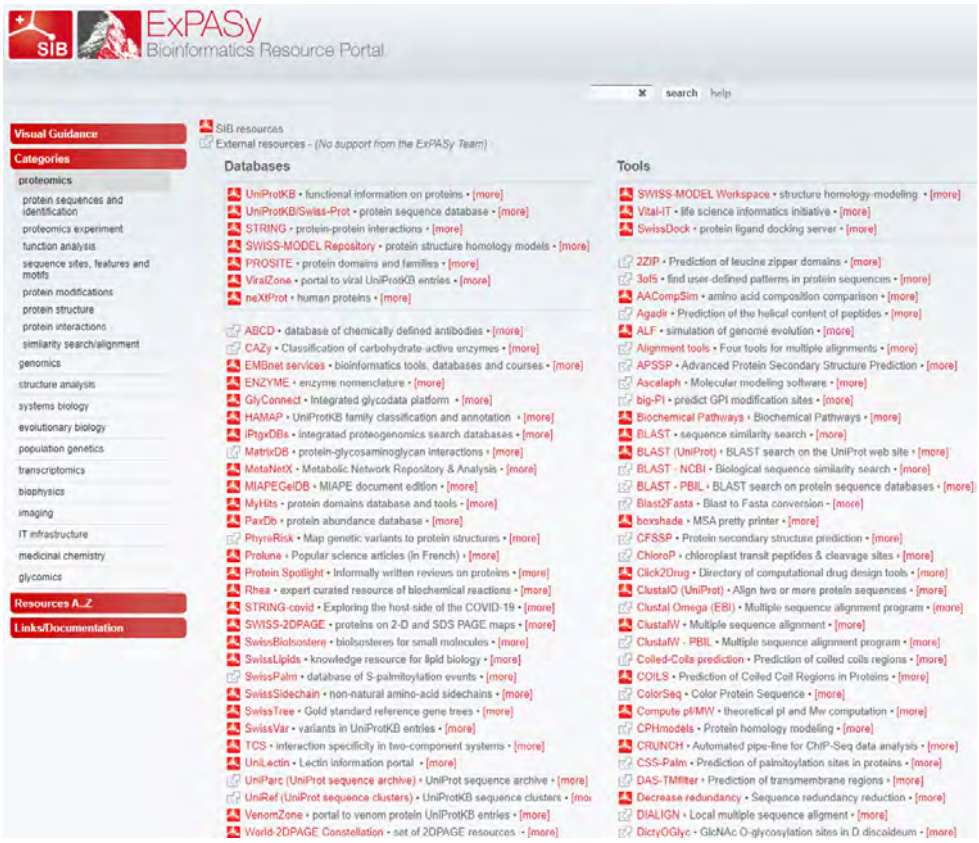


图 3.19　蛋白质序列数据资源与生物信息学分析平台 ExPASy 主页

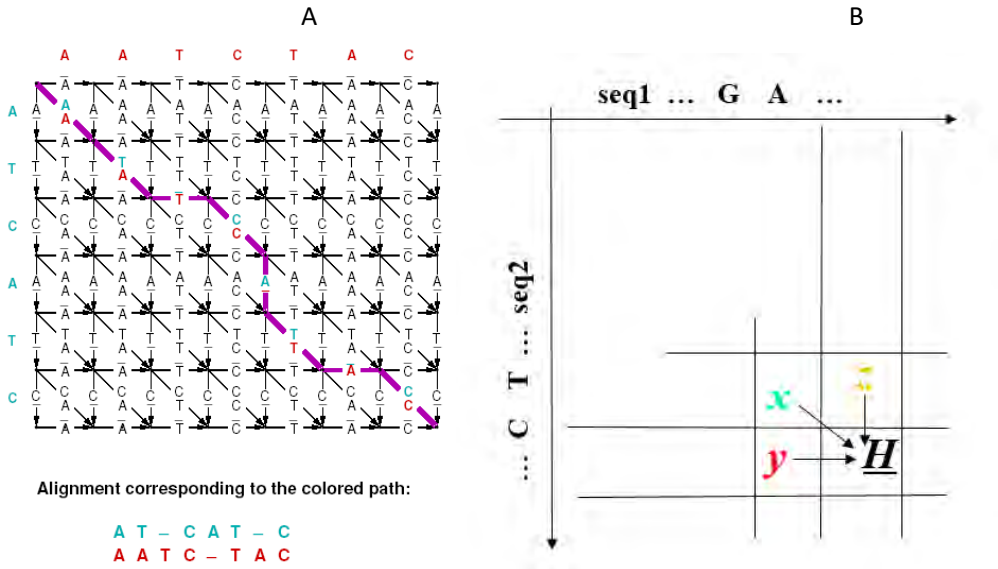


图4.1　两条序列联配方式与路径

A.　两条序列联配方形示意图。图中标出了两条联配序列（一条在最顶端，另一条在最左列）之间所有可能的联配方式（匹配、错配和引入空位）。对于全局联配，从头到尾存在许多路径（即全局联配结果），图中标出一条途径及其对应的联配结果（最下端）。B．联配方式示意图。对于任何一个联配位点，联配延伸路径仅存在三个方向，即对角线（导致匹配或错配）、横向或纵向（均引入空位）

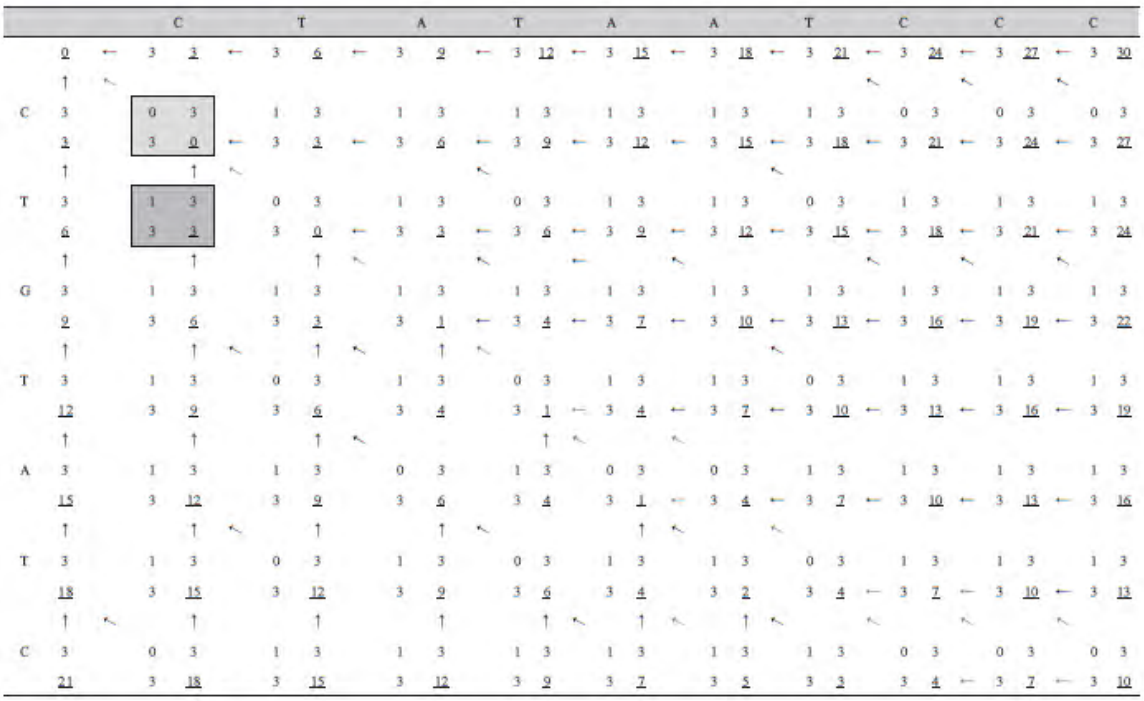


图4.2　Needleman-Wunsch算法实例

计分方式：设定碱基错配罚 1 分，单个碱基缺失或插入时罚 3 分，碱基匹配不罚分即得 0 分。图中箭头为每个位点最佳联配方向（来源）。每个单元格（特别标出三个例子）内有 4 个数字，分别为 x、y、z、H 值（H 值下画线）。两条序列的全局最优联配 H 值＝－10（即罚 10 分），其联配路径见红色箭头

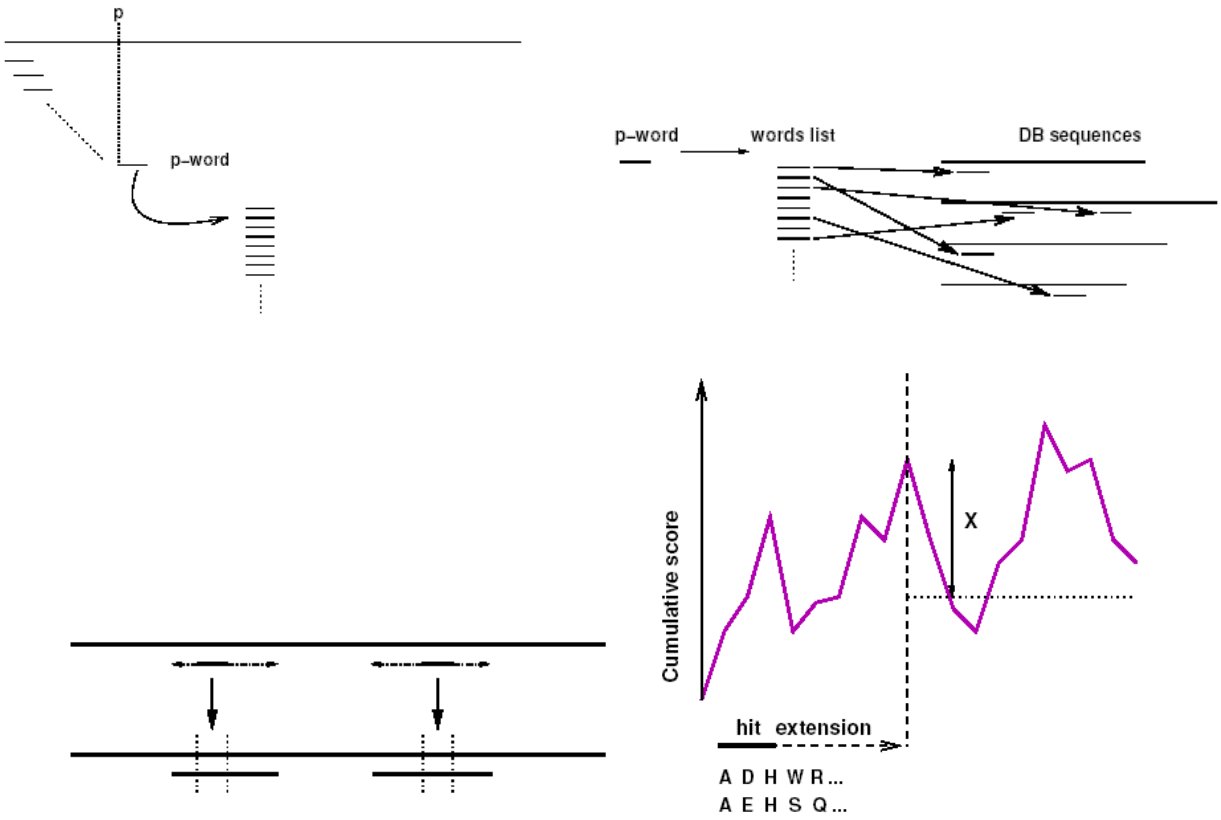


图 4.3　BLAST 算法图解

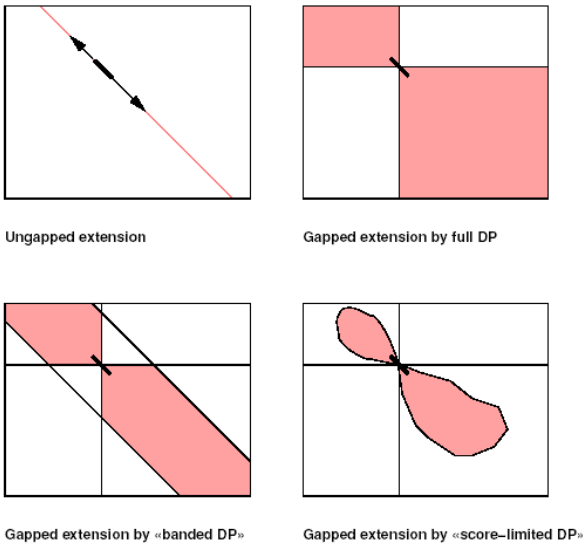


图4.4　序列动态规划联配方式

可以利用空格（是否允许）、空格数量和联配值等进行限定联配过程

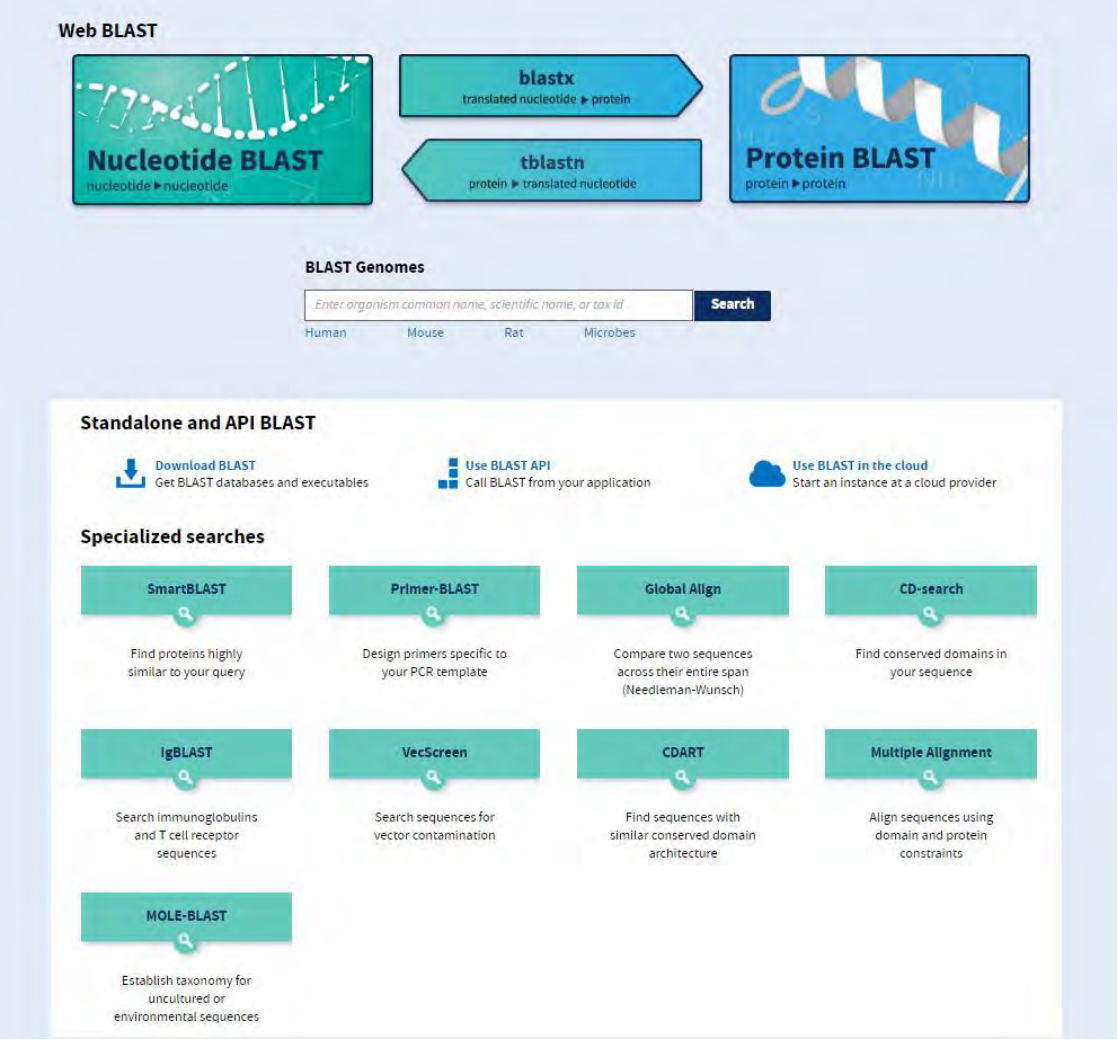


图 4.5　美国国家生物信息中心（NCBI）提供的在线 BLAST 序列搜索工具（2020 年 5 月）

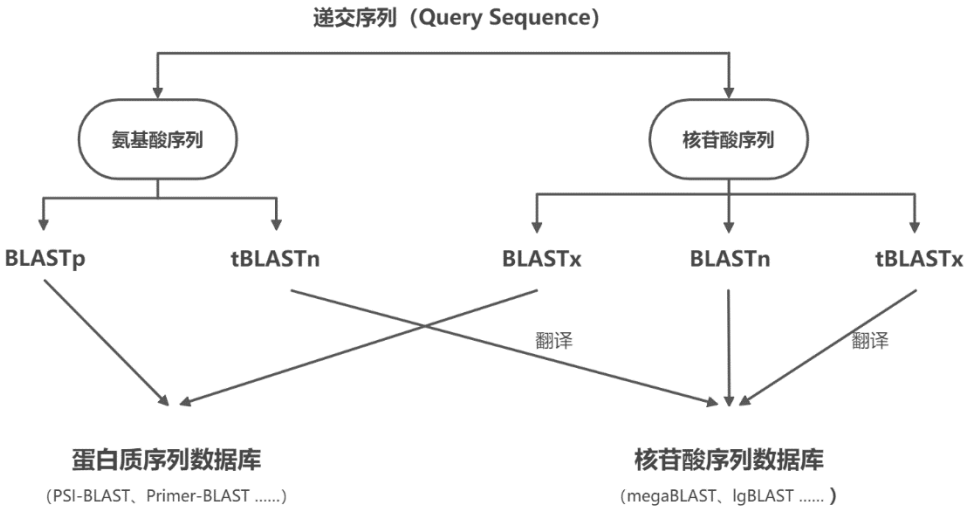


图4.6　数据库BLAST搜索工具大全

基于局部动态规划算法，除了标准 BLAST 工具（如 BLASTP 等），不断发展了新的用于蛋白质和核苷酸序列数据的 BLAST 工具（如 PSI-BLAST 等）

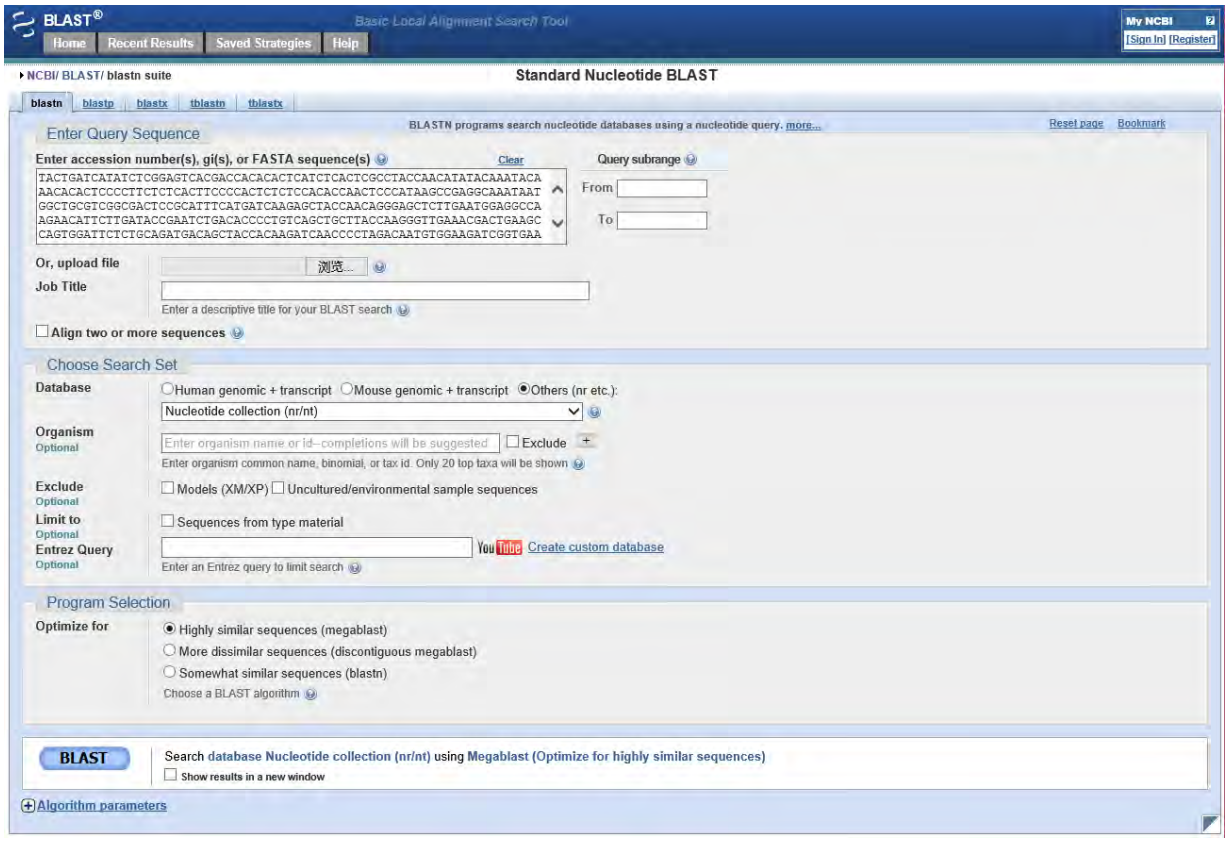


图 4.7　利用 BLASTN 工具搜索核苷酸数据库

Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenBank Graphics Distance tree of results

	Description	Max score	Total score	Query cover	E value	Ident	Accession
☐	Lilium davidii var. unicolor granule-bound starch synthase (GBSS1) mRNA, complete cds	3605	3605	95%	0.0	98%	KP179405.1
☐	Lilium davidii granule-bound starch synthase 1 gene, partial cds	965	965	25%	0.0	97%	KP751445.1
☐	PREDICTED: Musa acuminata subsp. malaccensis granule-bound starch synthase 1, chloroplastic/amyloplastic-like (LOC103996709), mRNA	798	798	70%	0.0	76%	XM_009417716.1
☐	Musa acuminata AAA Group cultivar Brazilian granule bound starch synthase (GBSSI-1) mRNA, complete cds	798	798	70%	0.0	76%	KF512020.1
☐	Musa acuminata AAA Group cultivar Tianbao granule bound starch synthase (GBSS1) mRNA, complete cds	793	793	70%	0.0	76%	HQ646360.4
☐	PREDICTED: Vitis vinifera granule-bound starch synthase 1, chloroplastic/amyloplastic (LOC100243677), mRNA	610	610	70%	4e-170	74%	XM_002273572.3
☐	Vitis vinifera clone SS0AFA25YF06	603	603	70%	7e-168	74%	FQ393574.1
☐	Vigna unguiculata granule-bound starch synthase Ib precursor, mRNA, complete cds	523	523	70%	5e-144	73%	EF472253.1
☐	Ricinus communis starch synthase, putative, mRNA	455	455	55%	2e-123	74%	XM_002524371.1
☐	Ipomoea batatas GBSSI mRNA for granule-bound starch synthase I, complete cds	427	427	72%	4e-115	72%	AB071604.1
☐	Ipomoea batatas GBSSI mRNA for granule-bound starch synthase I, complete cds, clone: 120	422	422	72%	2e-113	72%	AB524727.1
☐	Ipomoea batatas starch synthase (SPSS67) mRNA, complete cds	416	416	72%	9e-112	72%	U44126.1
☐	Ipomoea trifida clone dWx6-C9a granule bound starch synthase I (Waxy) gene, partial sequence	99.0	99.0	5%	4e-16	82%	EU192901.1
☐	Ipomoea batatas GBSSI gene for granule-bound starch synthase I, complete cds, clone: 4	95.3	95.3	5%	5e-15	81%	AB524726.1
☐	Ipomoea trifida clone dWx13-H5b granule bound starch synthase I (Waxy) gene, partial cds, alternatively spliced	95.3	95.3	5%	5e-15	81%	EU192912.1
☐	Ipomoea trifida clone dWX13A granule bound starch synthase I (Waxy) gene, partial cds, alternatively spliced	95.3	95.3	5%	5e-15	81%	EU192910.1
☐	Ipomoea batatas GBSSI gene for granule-bound starch synthase I, partial cds, clone: 1	93.5	93.5	5%	2e-14	81%	AB534171.1
☐	Ipomoea batatas GBSSI gene for granule-bound starch synthase I, complete cds, clone: 3	93.5	93.5	5%	2e-14	81%	AB524725.1

图 4.8　BLASTN 搜索结果

Lilium davidii granule-bound starch synthase 1 gene, partial cds

Sequence ID: gb|KP751445.1| **Length:** 567 **Number of Matches:** 1

Range 1: 1 to 567 GenBank Graphics Next Match Previous Match

Score	Expect	Identities	Gaps	Strand
965 bits(522)	0.0	552/567(97%)	0/567(0%)	Plus/Plus

```
Query  1024  GGGAGAAAAATAAACTGGATGAGGGCTGGAATTTTAGAATCCGACGCCGTTGTAACTGTG  1083
             |||||  |||||  ||  |||||||  |||||||||||||||||  ||||||||||||||||||||
Sbjct  1     GGGAGGAAAATCAATTGGATGAAGGCTGGAATTTTAGAAGCCGACGCCGTTGTAACTGTG  60

Query  1084  AGCCCATACTATGCTAAAGAGCTCGTCTCTGGAGAAGATAAAGGTGTTGAGTTGGACAAA  1143
              ||  ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  61    AGTCCATACTATGCTAAAGAGCTCGTCTCTGGAGAAGATAAAGGTGTTGAGTTGGACAAA  120

Query  1144  GATATAACCATGATTGGCATCAAAGGGATTGTGAATGGGATGGATATTAATTTTTGGAAT  1203
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  121   GATATAACCATGATTGGCATCAAAGGGATTGTGAATGGGATGGATATTAATTTTTGGAAT  180

Query  1204  CCATTGACAGACAAGTATATCACTGCCAATTATGATGCGACAACGGTAATGGAGGCAAAG  1263
              ||||||||||||||||||||||||||||||||||||||||||||||||||||  ||||||  |||
Sbjct  181   CCATTGACAGACAAGTATATCACTGCCAATTATGATGCGACAACGGTAACGGAGGCGAAG  240

Query  1264  CGTGTCAATAAGCAAGCACTACAAGCAGAAGTTGGCTTGCCTGTAGACCCAGACATTCCA  1323
              ||||||||||||||||||||||||||||||||||||||||||||||  ||||||||||||||||
Sbjct  241   CGTGTCAATAAGCAAGCACTACAAGCAGAAGTTGGCTTGCCTGTTGACCCAGACATTCCA  300

Query  1324  GTGATAGTCTTCGTAGGAAGGCTAGAGGAGCAGAAAGGCTCAGACATTCTCGCTGCAGCA  1383
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||  ||||||||
Sbjct  301   GTGATAGTCTTCGTAGGAAGGCTAGAGGAGCAGAAAGGCTCAGACATTCTCACTGCAGCA  360

Query  1384  ATTCCAGATTTCATTGATGAGAATGTGCAGATAATAATTCTCGGAACCGGCAAGAAAATC  1443
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  361   ATTCCAGATTTCATTGATGAGAATGTGCAGATAATAATTCTCGGAACCGGCAAGAAAATC  420

Query  1444  TTTGAAAAACAGGTCGAAGAAATAGAAGAAAAGTACCCGGACAAGGCGAGAGGAATTGCG  1503
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  421   TTTGAAAAACAGGTCGAAGAAATAGAAGAAAAGTACCCGGACAAGGCGAGAGGAATTGCG  480

Query  1504  AAATTCAATATTCCCTTAGCTCATATGATGATGGCTGGAGGTGATCTTATCATAGTTCCT  1563
              ||||||||  ||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  481   AAATTCAACATTCCCTTAGCTCATATGATGATGGCTGGAGGTGATCTTATCATAGTTCCT  540

Query  1564  AGTAGATTTGAGCCGTGTGGGCTTATT  1590
              ||||||||||  |||||  ||  ||  |||
Sbjct  541   AGTAGATTTGAACCGTGCGGTCTCATT  567
```

图 4.9　BLASTN 具体联配结果

Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenPept Graphics

Description	Max score	Total score	Query cover	E value	Ident	Accession
granule-bound starch synthase [Lilium davidii var. unicolor]	1129	1129	82%	0.0	98%	AJG44453.1
PREDICTED: granule-bound starch synthase 1, chloroplastic/amyloplastic [Phoenix dactylifera]	831	831	82%	0.0	71%	XP_008775302.1
PREDICTED: granule-bound starch synthase 1, chloroplastic/amyloplastic-like [Elaeis guineensis]	821	821	82%	0.0	69%	XP_010940833.1
PREDICTED: granule-bound starch synthase 1b, chloroplastic/amyloplastic-like [Elaeis guineensis]	812	812	79%	0.0	74%	XP_010917976.1
PREDICTED: granule-bound starch synthase 1, chloroplastic/amyloplastic-like [Musa acuminata subsp. malaccensis]	801	801	82%	0.0	69%	XP_009415991.1
granule bound starch synthase [Musa acuminata AAA Group]	796	796	82%	0.0	69%	ADZ30929.4
PREDICTED: granule-bound starch synthase 1, chloroplastic/amyloplastic [Vitis vinifera]	796	796	77%	0.0	72%	XP_002273608.1
PREDICTED: granule-bound starch synthase 1, chloroplastic/amyloplastic [Nelumbo nucifera]	794	794	79%	0.0	71%	XP_010252174.1
UDP-Glycosyltransferase superfamily protein isoform 1 [Theobroma cacao]	793	793	82%	0.0	68%	XP_007039341.1
PREDICTED: granule-bound starch synthase 1, chloroplastic/amyloplastic-like [Pyrus x bretschneideri]	792	792	77%	0.0	72%	XP_009366600.1
granule-bound starch synthase 1, chloroplastic/amyloplastic [Nelumbo nucifera]	791	791	79%	0.0	71%	NP_001289785.1
granule-bound starch synthase 1, chloroplastic/amyloplastic-like [Malus domestica]	788	788	77%	0.0	72%	NP_001280836.1
PREDICTED: granule-bound starch synthase 1, chloroplastic/amyloplastic-like [Malus domestica]	786	786	77%	0.0	72%	XP_008376222.1
PREDICTED: granule-bound starch synthase 1, chloroplastic/amyloplastic [Citrus sinensis]	783	783	77%	0.0	72%	XP_006491364.1
PREDICTED: granule-bound starch synthase 1, chloroplastic/amyloplastic-like [Musa acuminata subsp. malaccensis]	783	783	82%	0.0	68%	XP_009393091.1
granule-bound starch synthase [Codonopsis pilosula]	783	783	77%	0.0	71%	AJA91185.1
hypothetical protein CICLE_v10019346mg [Citrus clementina]	783	783	77%	0.0	72%	XP_006444732.1
PREDICTED: granule-bound starch synthase 1, chloroplastic/amyloplastic isoform X1 [Jatropha curcas]	782	782	78%	0.0	69%	XP_012086630.1
hypothetical protein PRUPE_ppa002955mg [Prunus persica]	780	780	77%	0.0	72%	XP_007218864.1
PREDICTED: granule-bound starch synthase 1, chloroplastic/amyloplastic-like isoform X1 [Gossypium raimondii]	780	780	77%	0.0	70%	XP_012439861.1
PREDICTED: granule-bound starch synthase 1, chloroplastic/amyloplastic-like isoform X2 [Gossypium raimondii]	778	778	78%	0.0	70%	XP_012439862.1
hypothetical protein CISIN_1g007224mg [Citrus sinensis]	778	778	77%	0.0	72%	KDO86605.1
unnamed protein product [Vitis vinifera]	778	778	77%	0.0	72%	CBI34608.3

granule-bound starch synthase 1, chloroplastic/amyloplastic-like [Malus domestica]

Sequence ID: ref|NP_001280836.1| Length: 615 Number of Matches: 1

See 2 more title(s)

Range 1: 43 to 615 GenPept Graphics Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
788 bits(2035)	0.0	Compositional matrix adjust.	414/573(72%)	474/573(82%)	5/573(0%)	+1

```
Query  244   YQGLKRLKPVDSLQMTATTRSTPRQC-GRSVNC----GGAISCSTGMNLVYVGTETGPHS  408
             + GL+ L  VD L++      S  RQ  G++VN       G I C +GMNLV++GTE GP S
Sbjct  43    HNGLRALNSVDELRVRIMANSVARQTRGKTVNSTRKTSGVIVCGSGMNLVFLGTEVGPWS  102

Query  409   KTgglgdvlgglPPAMAARGHRVMVVTPRYDQYKDAWDTGVVAEFKVGDKIETVRYFHLY  588
             KTGGLGDVLGGLPPAMAA GHRVM ++PRYDQYKDAWDT V  E KVGDK ETVR+FH Y
Sbjct  103   KTGGLGDVLGGLPPAMAANGHRVMTISPRYDQYKDAWDTEVTVELKVGDKTETVRFFHCY  162

Query  589   KRGVDRVFIDHPWFLEKVWGKTGGKLYGPVTGTDYDDNQLRFSLLCLAALEAPRVLNLNN  768
             KRGVDRVF+DHP FLEKVWGKT  K+YGPV G D+ DNQLRFSLLC AAL APRVLNLN+
Sbjct  163   KRGVDRVFVDHPLFLEKVWGKTASKIYGPVAGVDFKDNQLRFSLLCQAALVAPRVLNLNS  222

Query  769   SEYFSGPYGEDVVFIANDWHTGPLSCYLKSMYQAVGIYKSAKVAFCIHNIAYQGRFPFAD  948
             S+YFSGPYGE+VVFIANDWHT  L CYLK++Y+  GIYK+AKVAFCIHNIAYQGRF FAD
Sbjct  223   SKYFSGPYGEEVVFIANDWHTALLPCYLKAIYKPKGIYKTAKVAFCIHNIAYQGRFAFAD  282

Query  949   FSLLNLPdkfkssfdffdGYLKPVKGRKINWMRAGILESDAVVTVSPYYAKELVSGEDKG  1128
             F+LLNLP++FKSSFDF DGY KPVKGRKINWM+AGILESD V+TVSPYYA+ELVS  +KG
Sbjct  283   FALLNLPNEFKSSFDFIDGYNKPVKGRKINWMKAGILESDKVLTVSPYYAEELVSSVEKG  342

Query  1129  VELDKDITMIGIKGIVNGMDINFWNPLTDKYITANYDATTVMEAKRVNKQALQAEVGLpv  1308
             VELD  +    I+GIVNGMD+  WNP+TDKY T  YDA+TV +AK + K+ALQAEVGLPV
Sbjct  343   VELDNILRKSRIQGIVNGMDVQEWNPVTDKYTTVKYDASTVADAKPLLKEALQAEVGLPV  402

Query  1309  dpdipvivFVGRLEEQKGSDILAAAIPDFIDENVQIIILGTGKKIFEKQVEEIEEKYPDK  1488
             D DIPVI F+GRLEEQKGSDIL  AIP FI ENVQII+LGTGKK  EKQ+E++E +YPDK
Sbjct  403   DRDIPVIGFIGRLEEQKGSDILIEAIPHFIKENVQIIVLGTGKKPMEKQLEQLETEYPDK  462

Query  1489  ARGIAKFNIPLAHMMMAGGDLIIVPSRFEPCGLIQLEGMQYGMPVICSTTGGLVDTVKEG  1668
             ARGIAKFN+PLAHM+ AG D ++VPSRFEPCGLIQL  M+YG   I ++TGGLVDTVKEG
Sbjct  463   ARGIAKFNVPLAHMITAGADFMLVPSRFEPCGLIQLHAMRYGTVPIVASTGGLVDTVKEG  522

Query  1669  FTGFHMGAFTVNCEAIDPvdvvatvktvkkalkvYGTPAFSEMVQNCMAQDLSWKGPAKK  1848
             FTGFHMGAF V CE +DPVDV A   TV +AL  YGTPAF+E++ NCMAQDLSWKGPAKK
Sbjct  523   FTGFHMGAFNVECEVVDPVDVQAIATTVTRALGSYGTPAFTEIISNCMAQDLSWKGPAKK  582

Query  1849  WEELLLGLGVHGSQPGIDGEEIAPMSKENVATP  1947
             WEE+LL LGV  S+ GI+GEEIAP++KENVATP
Sbjct  583   WEEVLLSLGVANSELGIEGEEIAPLAKENVATP  615
```

图 4.10　利用 BLASTX 工具搜索的结果

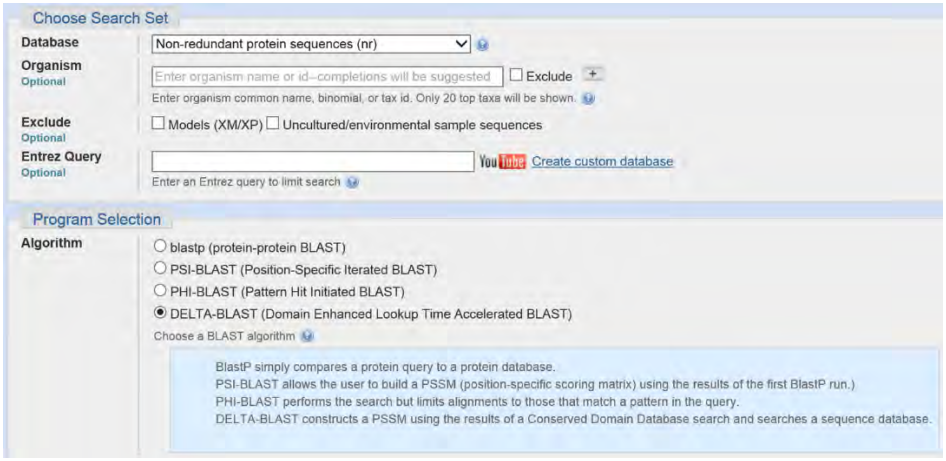


图 4.11　NCBI 的 BLAST 主页

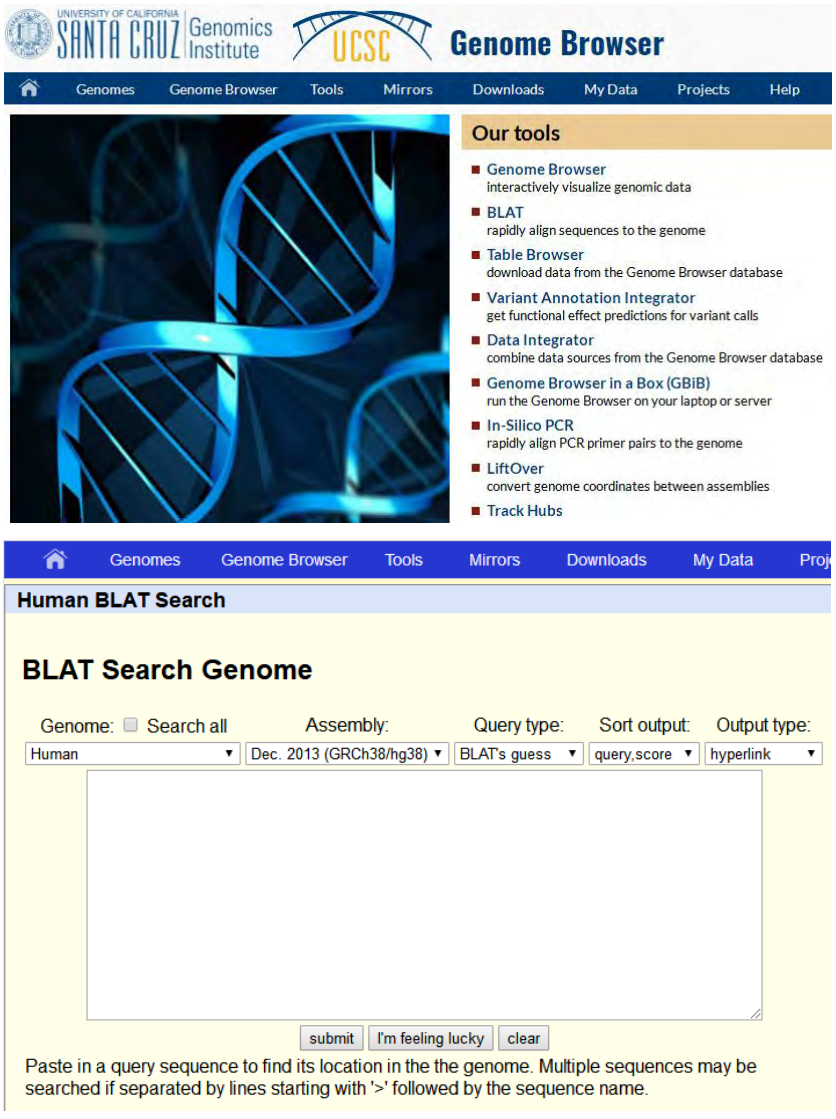


图 4.12　美国加州大学基因组学研究所 BLAT 主页

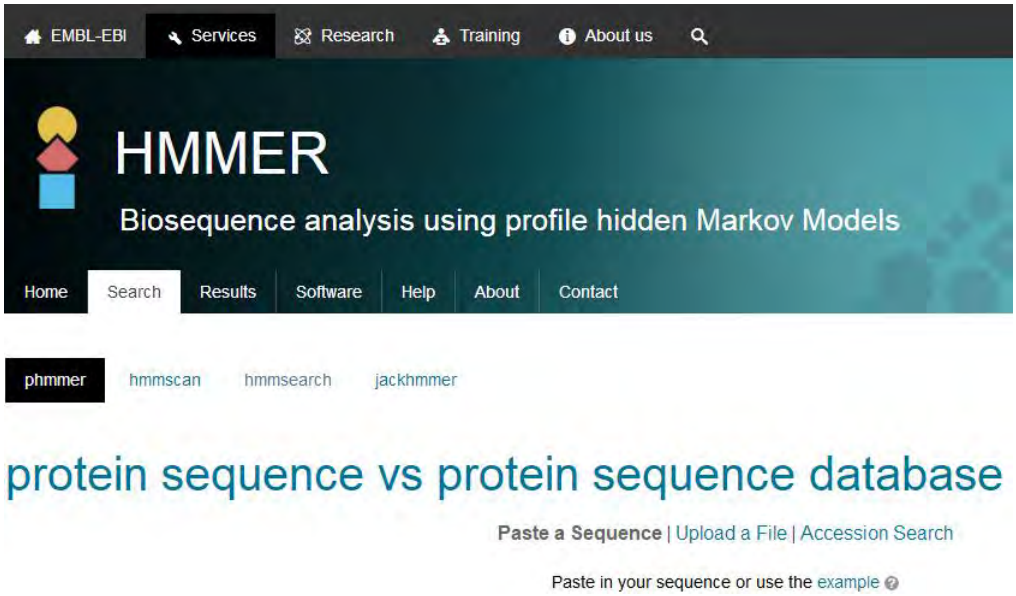


图 4.13　欧洲生物信息学研究所提供的数据库 HMMER 在线搜索界面

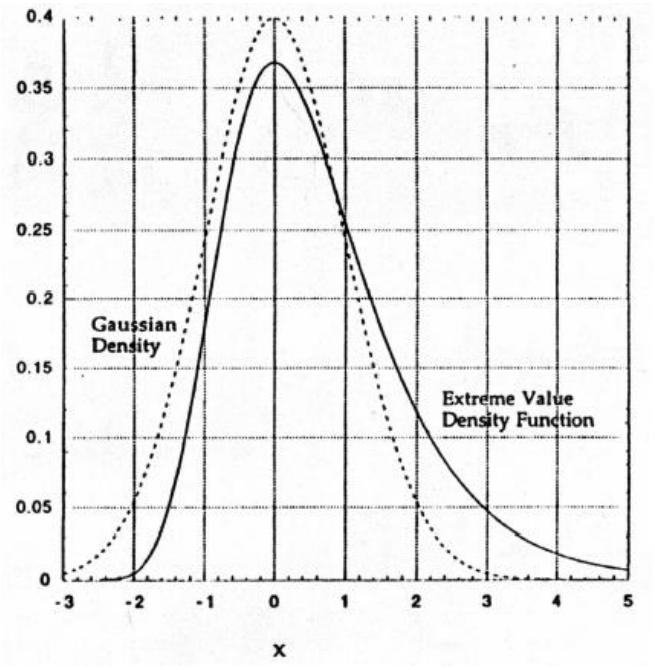


图4.14　概率密度函数正态分布（虚线）和极值分布（实线）比较

x 表示变量

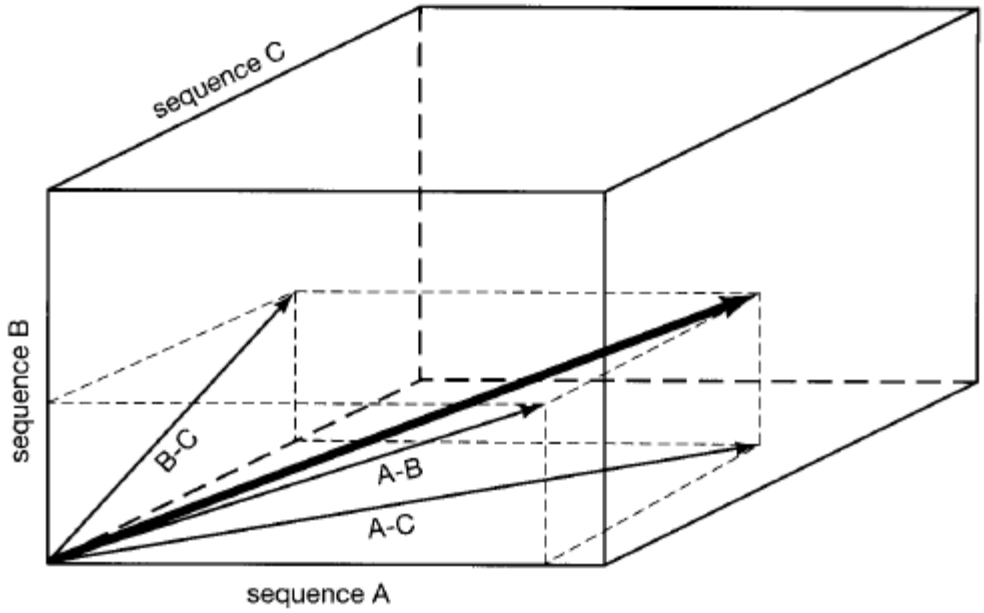


图 5.1　三条序列全局联配路径空间示意图

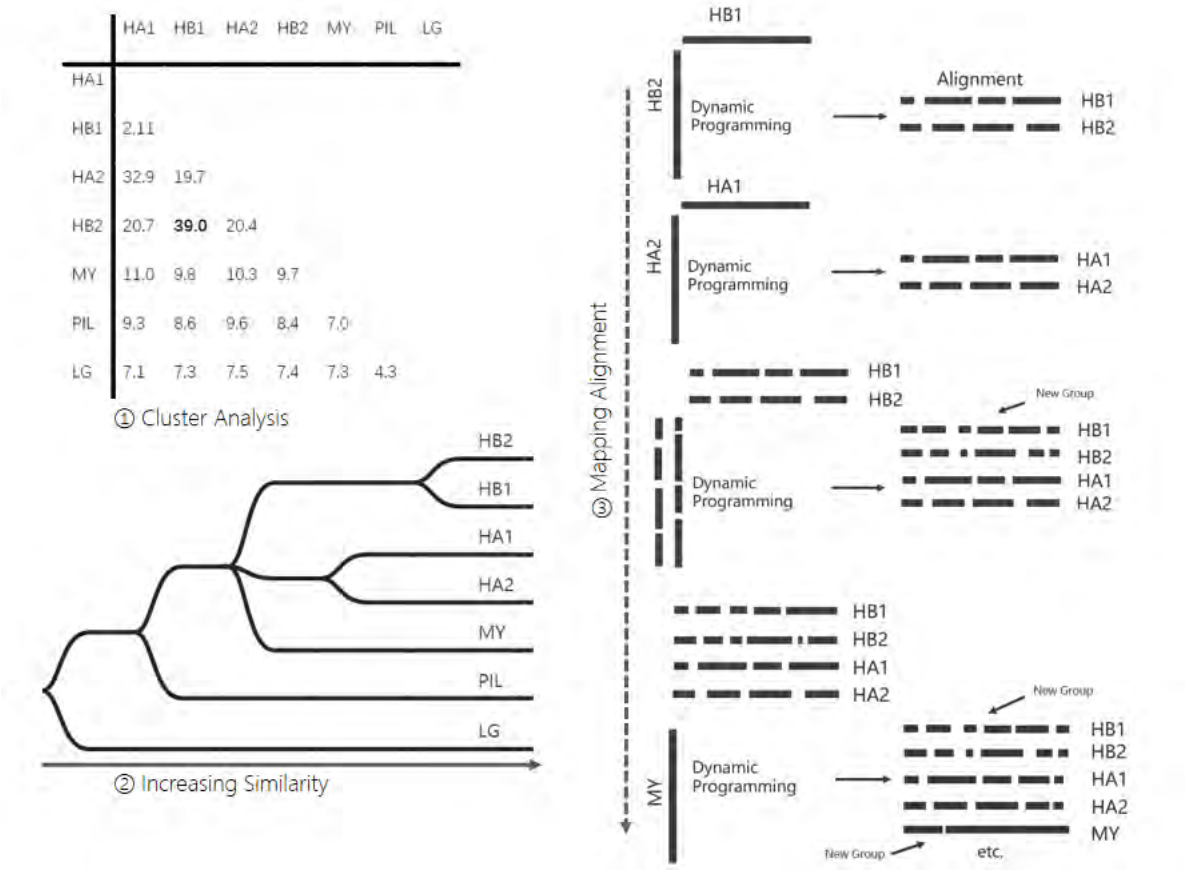


图 5.2　多序列渐进式全局联配算法案例（改自 Baxevanis and Ouellette，2001）

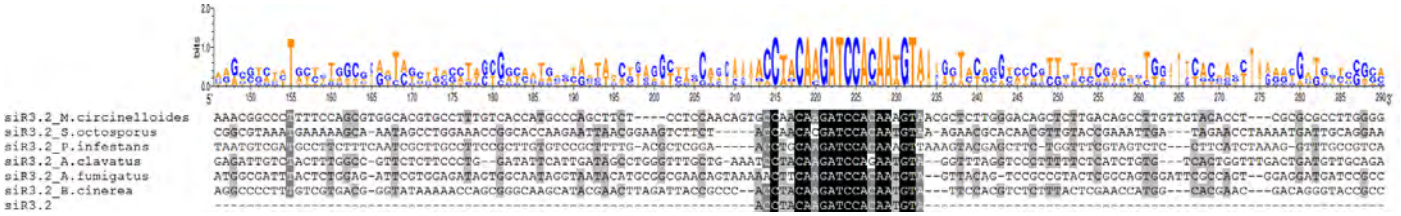


图5.3　多序列局部保守性

7 条序列中有一段序列高度保守（黑色区域）；图最上方为相对信息量图，表示各列信息量大小

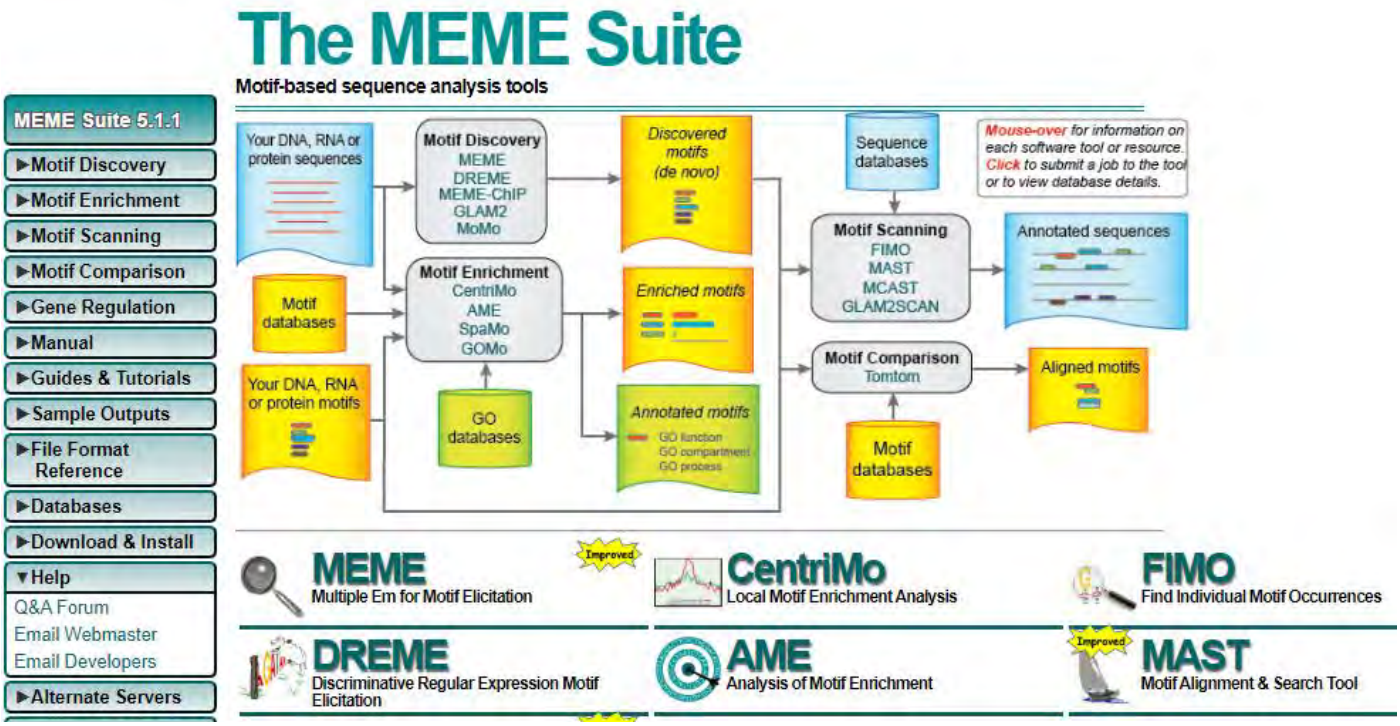


图 5.4　MEME 软件包在线分析平台（http://meme-suite.org/）

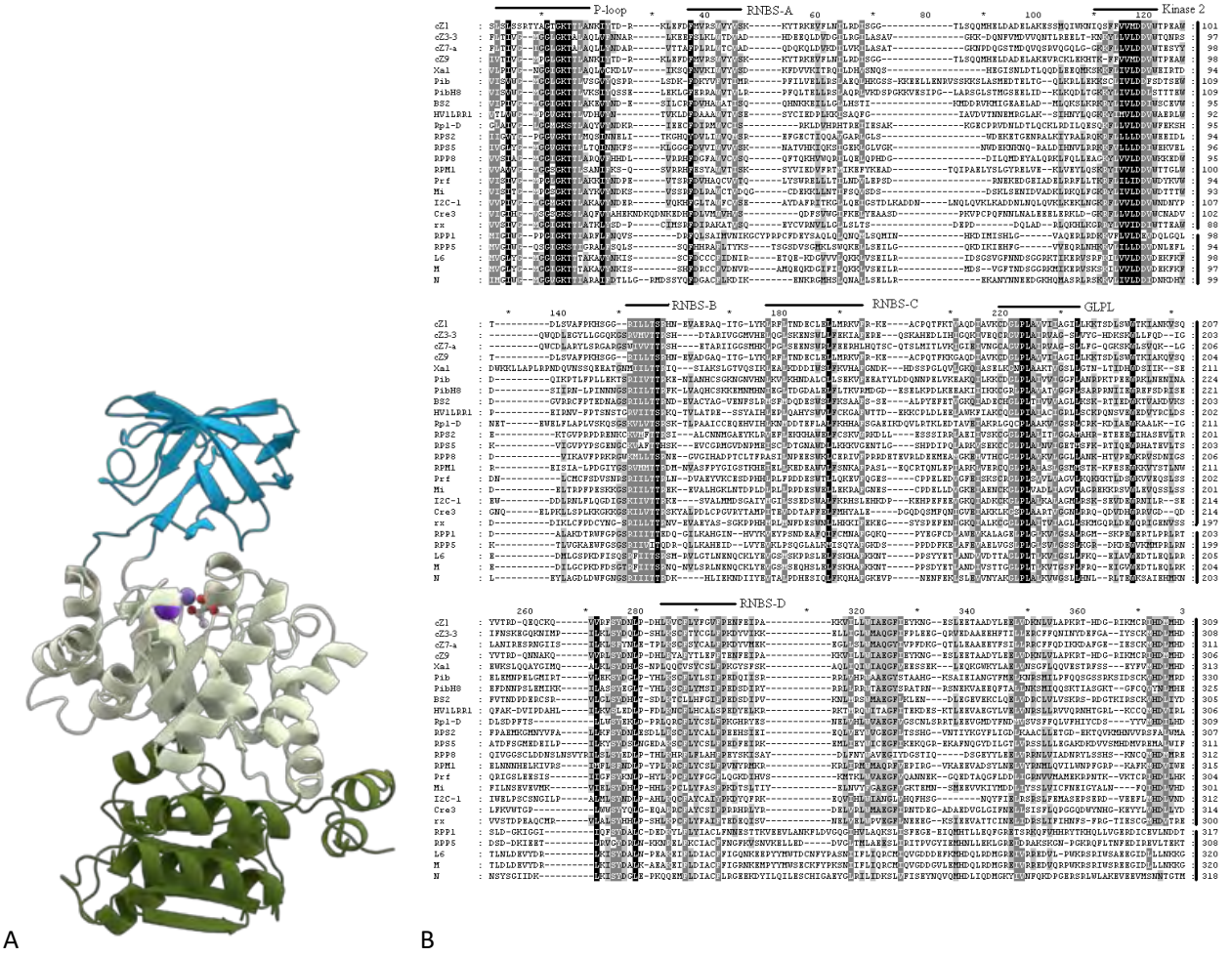
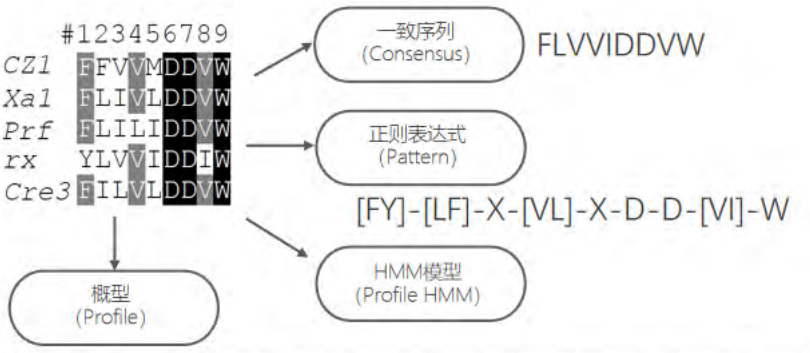


图5.5　蛋白质功能域举例

A．蛋白质结构水平上的功能域：丙酮酸激酶（pruvatekinase，PDB 数据库：1PKN）的三个功能域；B．蛋白质序列水平上的功能域：植物 NBS 类抗性基因的 NBS 功能域（引自 Tian et al.，2004），图中标出了该功能域的若干基序（如 Kinase2、GLPL 等）



氨基酸	#1	#2	#3	#4	#5	#6	#7	#9	#9
F	0.80	0.20	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Y	0.20	0.80	0.00	0.00	0.00	0.00	0.00	0.00	0.00
L	0.00	0.00	0.20	0.20	0.40	0.00	0.00	0.00	0.00
I	0.00	0.00	0.40	0.00	0.40	0.00	0.00	0.20	0.00
V	0.00	0.00	0.40	0.80	0.00	0.00	0.00	0.80	0.00
M	0.00	0.00	0.00	0.00	0.20	0.00	0.00	0.00	0.00
D	0.00	0.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00
W	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
其他	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

图 5.6　功能域的 4 种表述方法/ 模型（以 NBS 功能域的 Kinase2 基序为例）

Entry: PS50822

General information about the entry

Entry name [info]	PIWI
Accession [info]	PS50822
Entry type [info]	MATRIX
Date [info]	MAY-2002 (CREATED); OCT-2013 (DATA UPDATE); MAR-2016 (INFO UPDATE).
PROSITE Doc. [info]	PDOC50822
Associated ProRule [info]	PRU00150

Name and characterization of the entry

Description [info]	Piwi domain profile.
Matrix / Profile [info]	

```
/GENERAL_SPEC: ALPHABET='ABCDEFGHIKLMNPQRSTVWYZ'; LENGTH=297;
/DISJOINT: DEFINITION=PROTECT; N1=6; N2=292;
/NORMALIZATION: MODE=1; FUNCTION=LINEAR; R1=2.1406; R2=0.01583904; TEXT='NScore';
/CUT_OFF: LEVEL=0; SCORE=402; N_SCORE=8.5; MODE=1; TEXT='!';
/CUT_OFF: LEVEL=-1; SCORE=276; N_SCORE=6.5; MODE=1; TEXT='?';
/DEFAULT: MO=-8; D=-20; I=-20; B1=-60; E1=-60; MI=-105; MD=-105; IM=-105; DM=-105;

                A   B   C   D   E   F   G   H   I   K   L   M   N   P   Q   R   S   T   V   W   Y   Z
/I:          B1=0; BI=-105; BD=-105;
/M: SY='L': M=-11,-16,-11,-18,-19, -4,-19,-18,  7,-23, 11,  3,-16,-25,-18,-22,-17,-11,  2,-21, -1,-20;
/M: SY='L': M= -8,-30,-19,-33,-26, 13,-32,-25, 27,-27, 28, 15,-27,-28,-25,-22,-21, -7, 24,-20,  0,-26;
/M: SY='M': M=  3,-20,-20,-26,-14,  1,-22,-17, 14,-17, 11, 15,-18,-19,-13,-18,-12, -7, 10,-19, -4,-14;
/M: SY='V': M= -8,-25,  6,-29,-24,  3,-30,-25, 11,-20, 11,  5,-23,-29,-24,-18,-17, -7, 17,-25, -6,-24;
/M: SY='I': M= -6,-27,-19,-31,-25,  8,-30,-25, 26,-25, 16, 11,-22,-25,-23,-22,-13, -4, 26,-22, -2,-25;
/M: SY='L': M= -9,-29,-21,-32,-23,  6,-31,-21, 28,-27, 37, 23,-27,-27,-19,-21,-25, -9, 18,-21, -1,-22;
/M: SY='P': M= -7, -3,-29,  2,  3,-27,  1,-13,-27, -5,-26,-19, -3, 19, -5, -7,  2, -3,-24,-29,-23, -3;
/M: SY='E': M= -4,  8,-21, 10, 14,-24,-12, -7,-21,  8,-24,-16,  8,-11,  4,  0,  9,  0,-15,-33,-17,  9;
/M: SY='B': M=-12,  9,-26,  9,  8,-12,-17,  3,-15, -3,-17,-14,  9, -8, -2, -7, -4, -6,-18,-20,  2,  2;
/I:          I=-4; MI=0; MD=-19; IM=0; DM=-19;
/M: SY='N': M= -7,  7,-22,  0, -3,-19, -2, -5,-16,  7,-21,-11, 16,-15, -1,  6,  2, -4,-16,-25,-14, -3; D=-4;
/I:          I=-4; DM=-19;
/M: SY='D': M= -3,  7,-23, 12,  8,-26, -3, -8,-26,  3,-23,-17,  3,  0,  0,  0,  3, -6,-20,-26,-18,  3; D=-4;
/I:          I=-4; DM=-19;
/M: SY='E': M= -8, 10,-22,  8, 13,-12,-13, -5,-16, -5,-18,-15, 13,-12,  0, -8,  7, -1,-17,-30,-13,  6;
/M: SY='K': M=-10, -6,-23, -6, -4, -8,-20,-13,-14, -1, -7, -7, -7, -7, -7, -7,  1, -5,  0,-11,-24, -8, -6;
/M: SY='Y': M=-16,-22,-30,-20,-15, 12,-28,  3, -1,-14,  3, -1,-22, -2,-12,-14,-20,-10,-11,  6, 39,-18;
```

图 5.7　概型举例

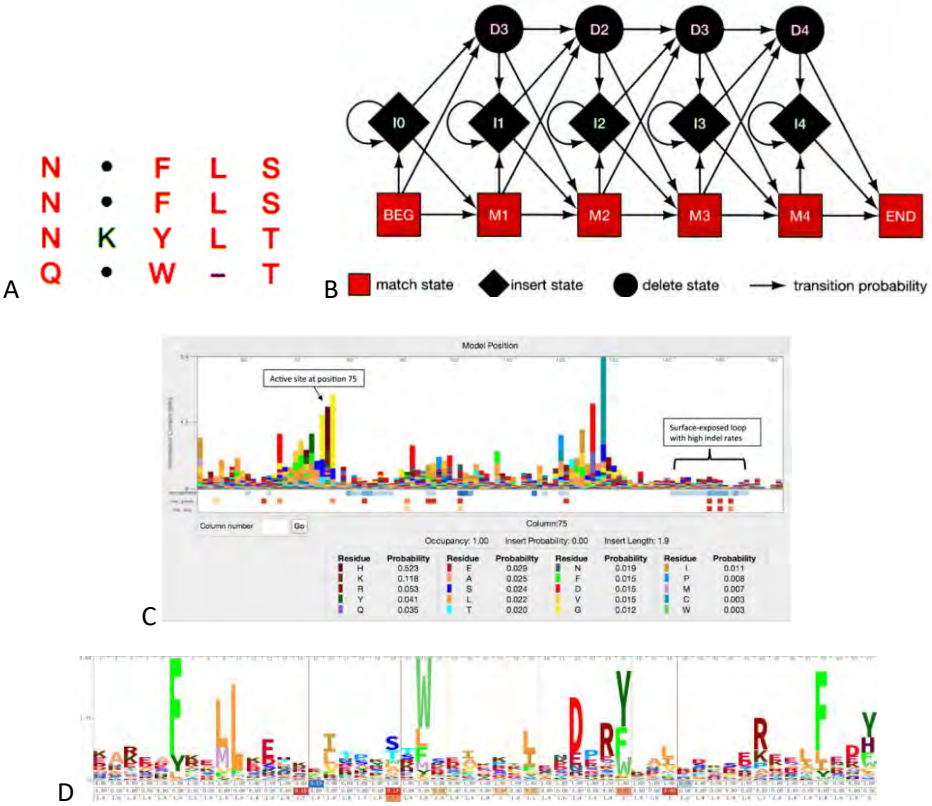


图 5.8　功能域 HMM 模型举例（部分引自 Mount，2004）

A．一个多序列联配结果：对于一个多序列联配结果，可能存在氨基酸匹配或错配（如第 1、3 和 5 列）、插入（第 2 列）和删 除（第 4 列）。B．多序列联配的马尔可夫模型：方形表示氨基酸联配状态，菱形表示插入状态，圆形表示缺失状态，箭头表 示转移概率。C．功能域数据库 Pfam 的一个记录（http://pfam.xfam.org/）：该记录以类似序列徽标的形式呈现，徽标图内包括 每个位点上 20 种氨基酸的出现概率大小（用比特表示），徽标图下面三行给出了 HMM 模型三个态（匹配或错配、插入和删 除）之间的转移概率（有关信息量和序列徽标详见下节说明）。作为一个说明，图中下方特别给出了第 75 位氨基酸的相应转 移概率和 20 种氨基酸的出现概率。根据该序列徽标，可以清楚看到该功能域蛋白质序列的保守区域，如第 75 位点附近区域

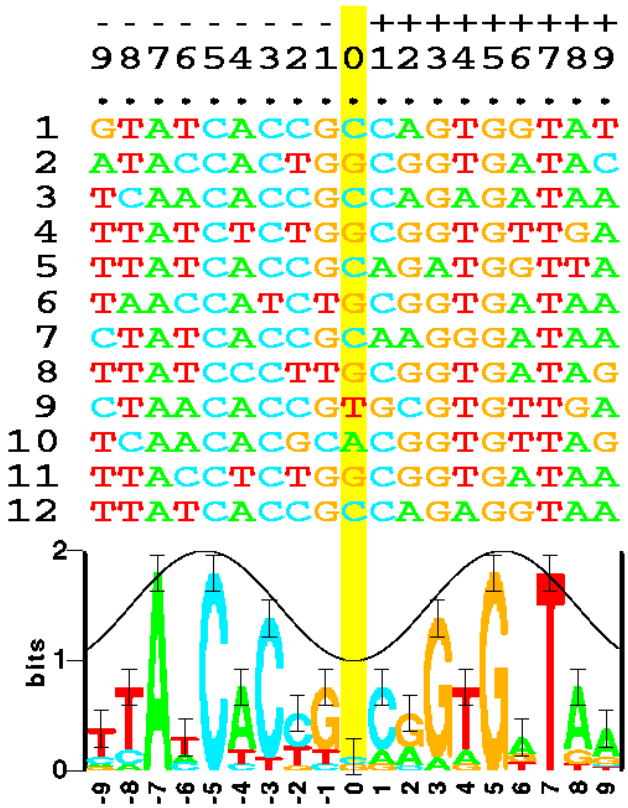


图5.9　多序列联配结果的序列徽标举例

图中为 12 条 λ 噬菌体 *cI* 和 *cro* 蛋白结合位点序列联配结果（上）和各位点信息量情况（下）

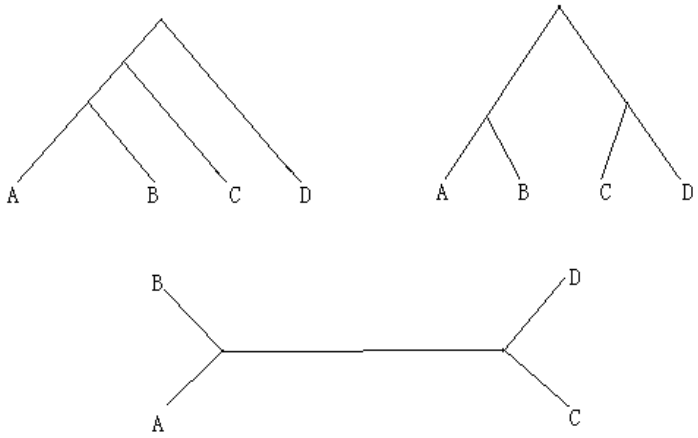


图 6.1　4 个物种（A～D）的 2 种有根树（上） 和 1 种无根树（下）形式

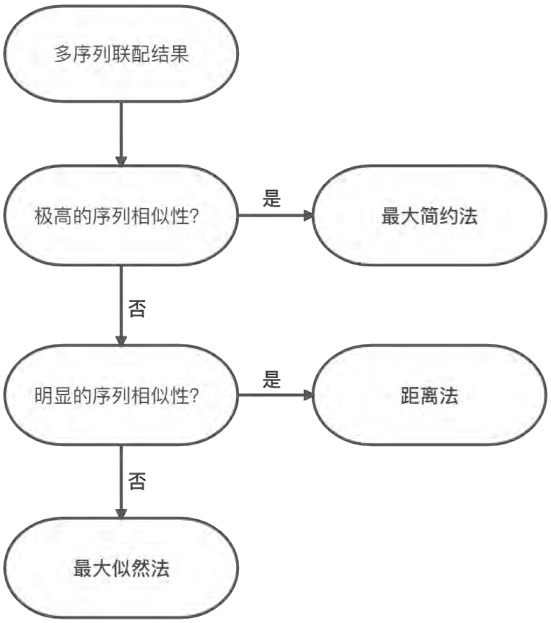


图 6.2　距离矩阵法、最大简约法和最大似然法的建树方法适用性

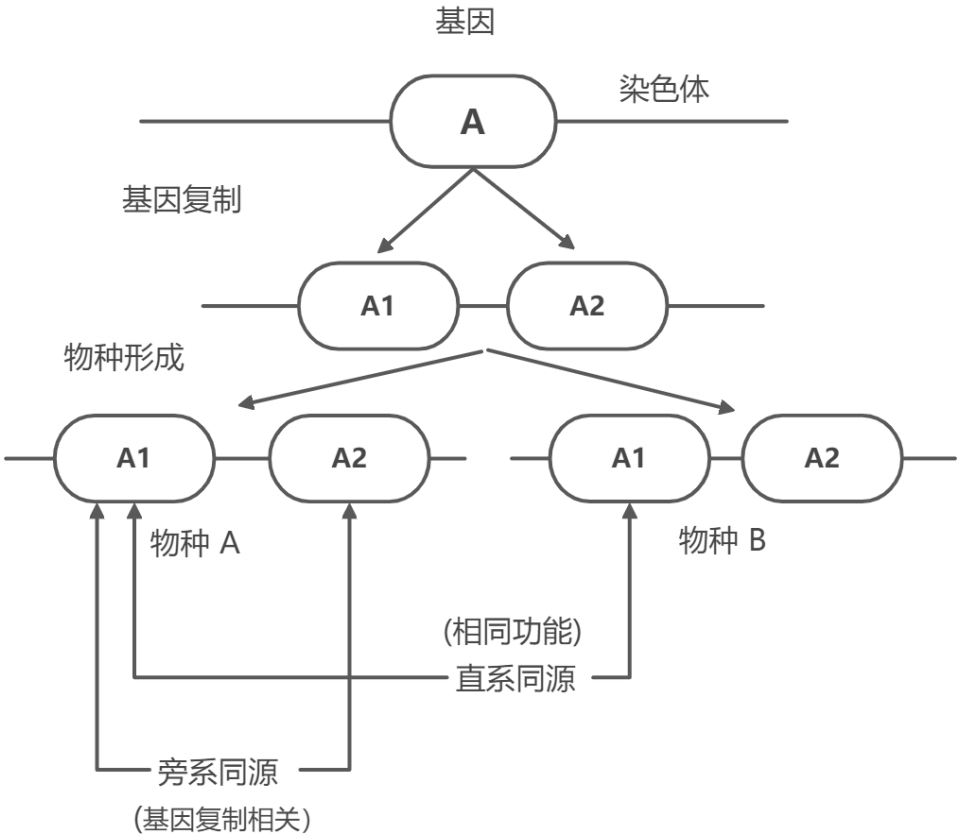


图 6.3　旁系同源基因和直系同源基因的产生机制

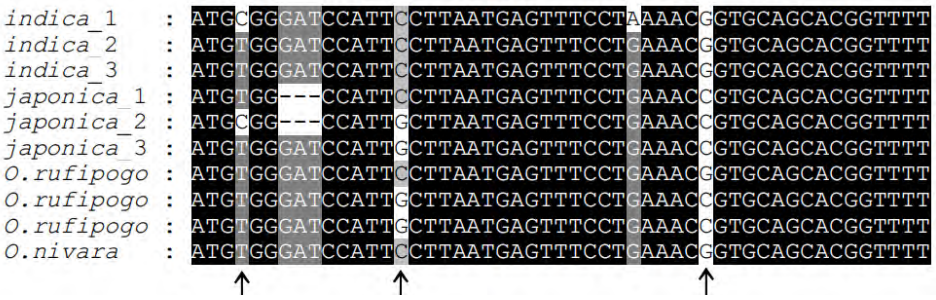


图6.4　水稻及其祖先野生种碱基变异情况

其中 3 个位点被定义为 SNP 位点（箭头处），一个位点发生插删

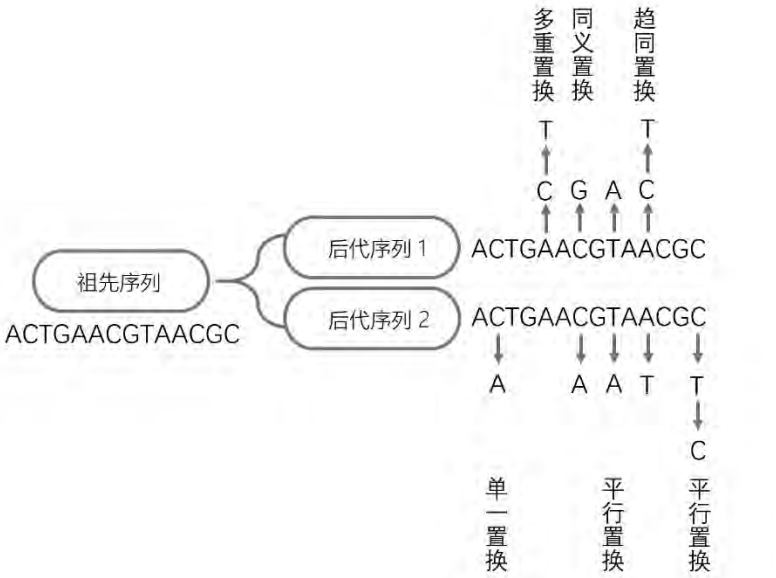


图 6.5　同源序列间的核苷酸变异机制

图 6.6　5 条生物线粒体 DNA 序列

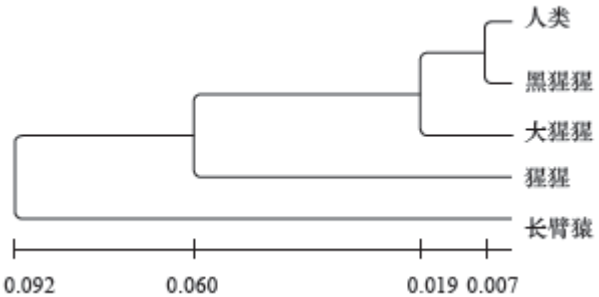


图 6.7　非加权平均连接聚类法（UPGMA）系统树

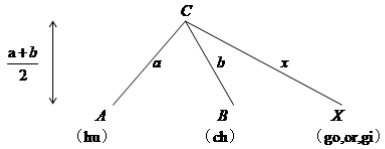


图 6.8　将 Fitch-Margoliash 法应用于图 6.6 线粒体序列分析的初始步骤

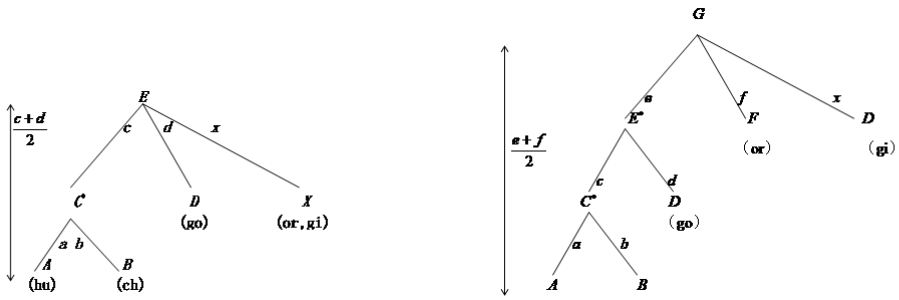


图 6.9　将 Fitch-Margoliash 法应用于图 6.6 线粒体序列分析的中间步骤

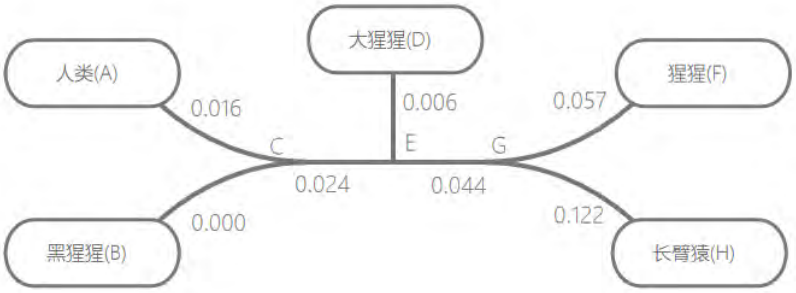


图 6.10　图 6.6 所列线粒体序列的 Fitch-Margoliash 无根系统树

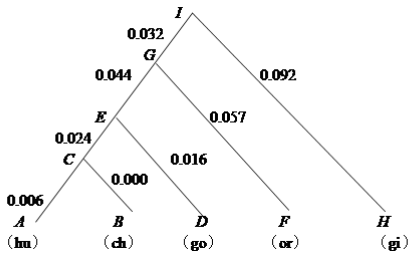


图 6.11　基于图 6.6 所列线粒体序列的 Fitch-Margoliash 有根系统树

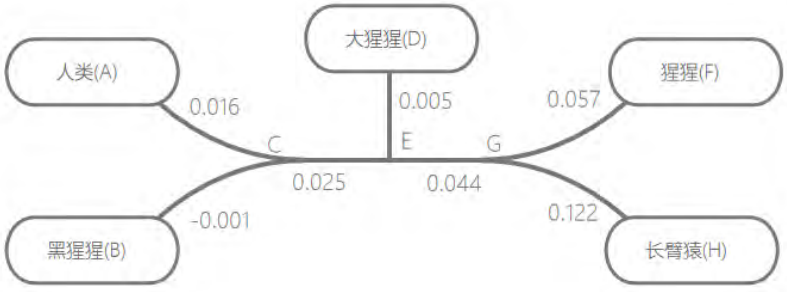


图 6.12　利用邻接法构建的 5 条线粒体序列无根系统发生树

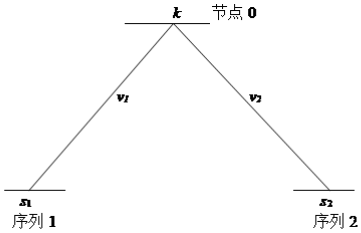


图6.13　两条序列的有根树状图

在 j 位点，两条序列具有碱基 s_1 和 s_2，相应节点具有碱基 k

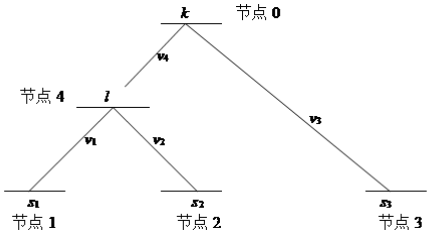


图6.14　三条序列的一种有根系统发生树

在位点 j，三个序列具有碱基 s_1、s_2、s_3，节点 0 和节点 4 具有碱基 k 和 l

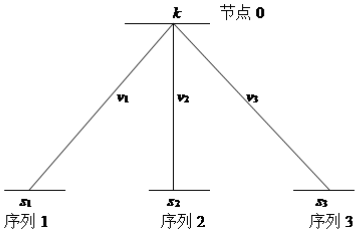


图6.15　三条序列的星状系统发生树

三条序列来自同一祖先序列 0

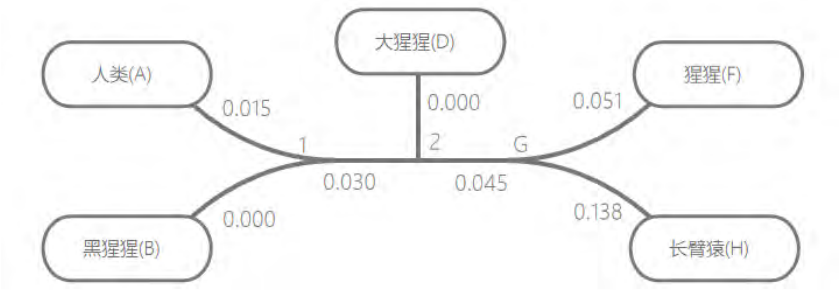


图 6.16　利用 PHYLIP 软件构建图 6.6 5 条线粒体序列的最大似然树

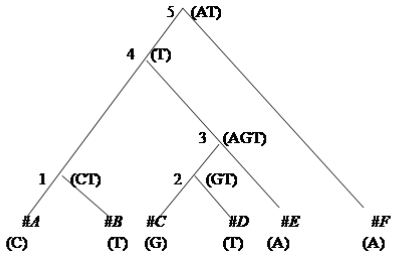


图 6.17　基于 6 条序列（物种#A～#F）一个位点碱基变异确定最简约树的过程

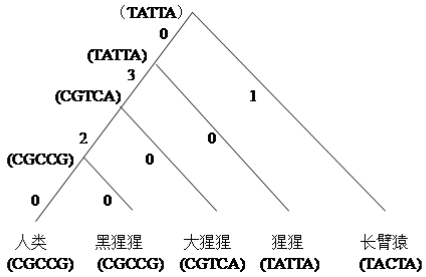


图6.18　基于图6.6 中5 个物种线粒体序列的最简约系统树

图中数字为节点间的碱基变更数

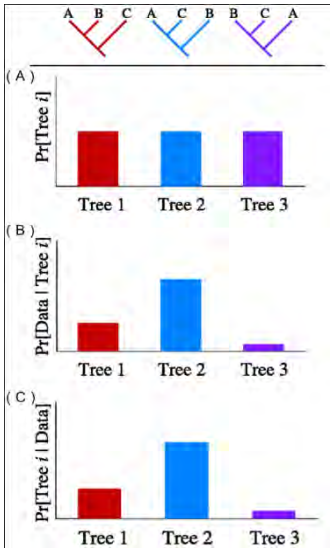
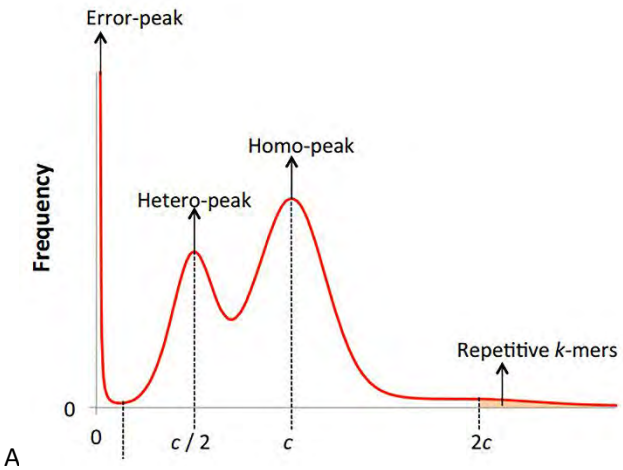
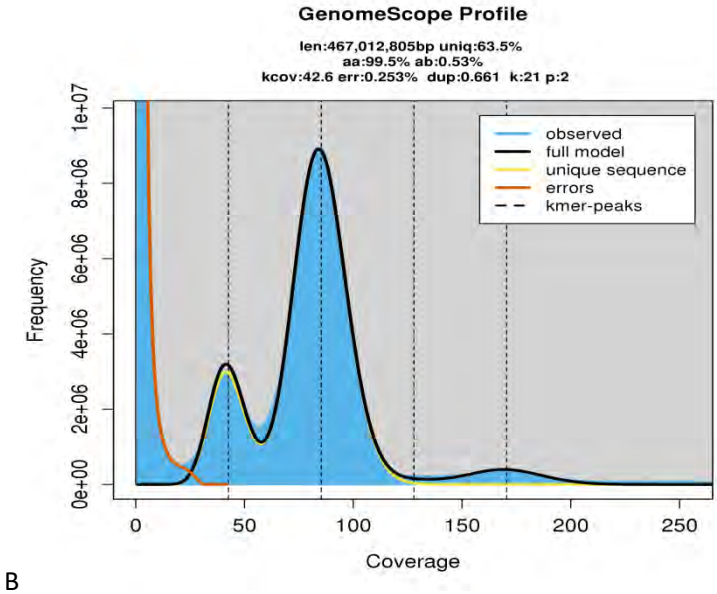


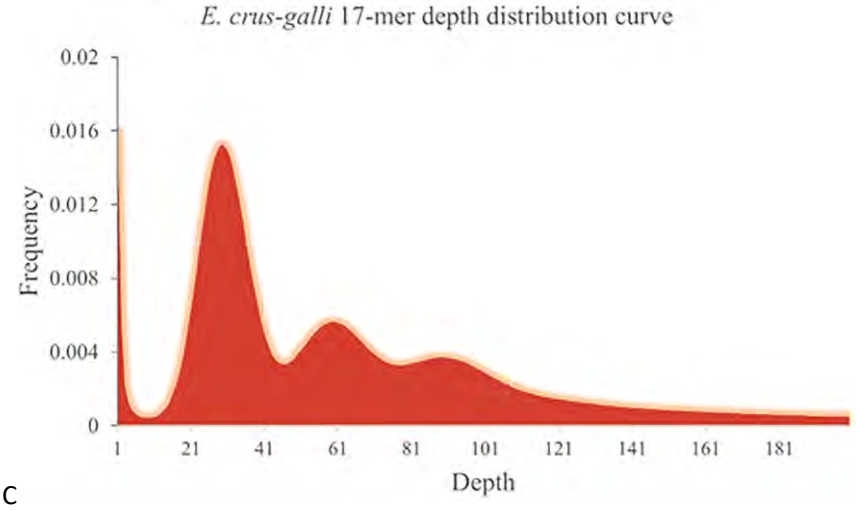
图 6.19　贝叶斯法建树的主要要素（引自 Huelsenbeck et al.，2001）



A



B



C

图7.1　复杂基因组*K*-mer 深度分布图及其特征峰

A．一个复杂基因组 *K*-mer 深度分布模式图，基因组杂合性和重复序列会在 *K*-mer 深度分布曲线上产生特征峰。B．一个杂合物种实例（*Z. latifolia*）。len. 长度；uniq. 特异序列；aa. 纯合率；ab. 杂合性；kcov. *K*-mer 覆盖深度；err. 错误率；dup. 重复率； *K*．字符串长度；*P*. 倍性。C．多倍化基因组 *K*-mer 深度分布图，以一个六倍体物种（*E. crus-galli*）为例

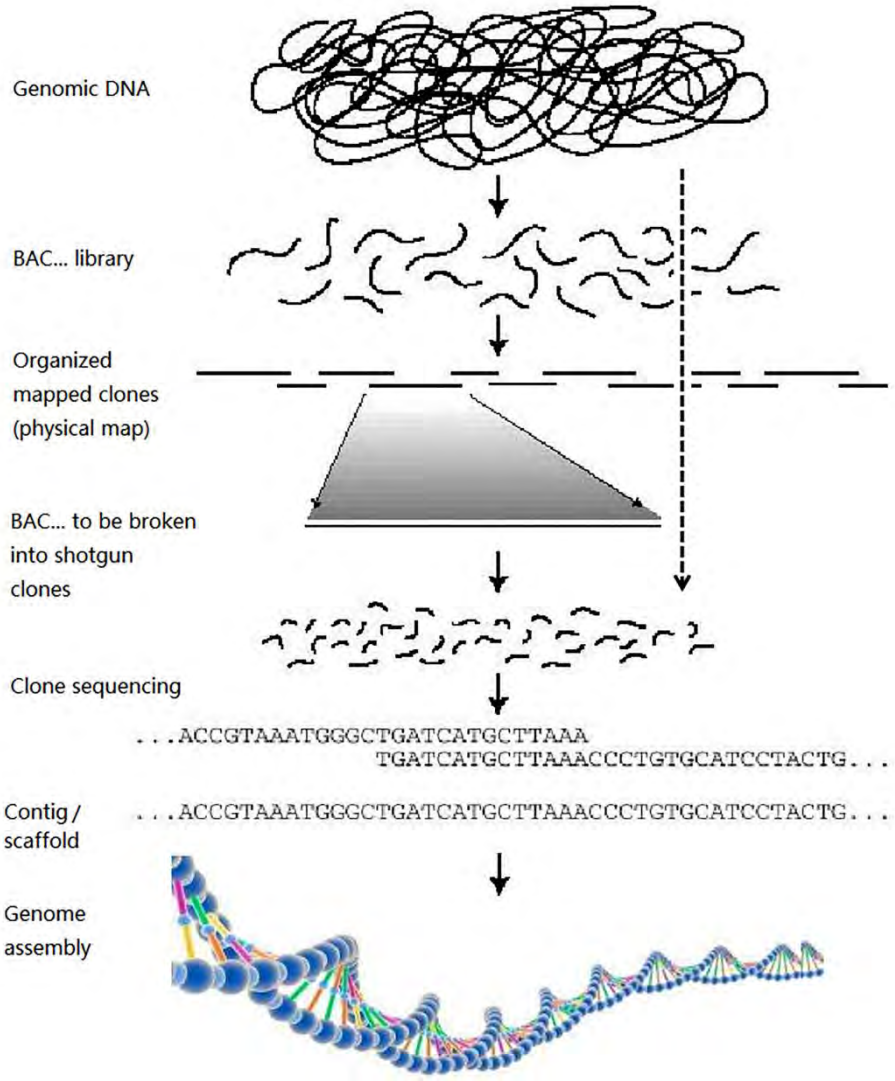


图7.2　全基因组的两种测序和拼装策略

逐步克隆法（实线箭头）需遗传图谱和物理图谱构建构成，然后基于每个 BAC 测序数据获得全基因组序列；全基因组鸟枪法（虚线箭头）直接将基因组 DNA 打碎进行测序，基于大量短读序及其相互关联数据进行全基因组拼装

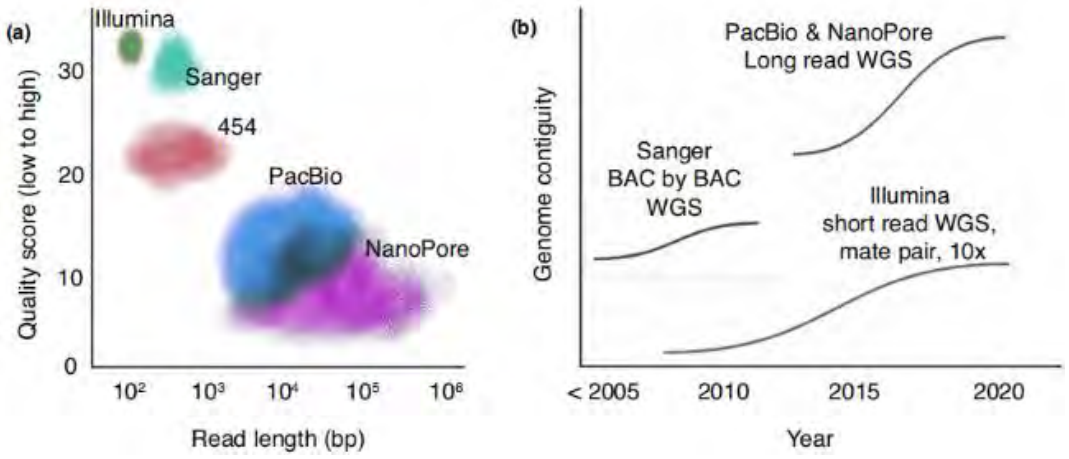


图7.3　不同测序技术对基因组拼接长度的影响（引自Michael and van Buren，2020）

A．传统测序技术和高通量测序技术的读序长度和测序质量值；B．不同测序技术用于基因组拼接的片段连续性比较

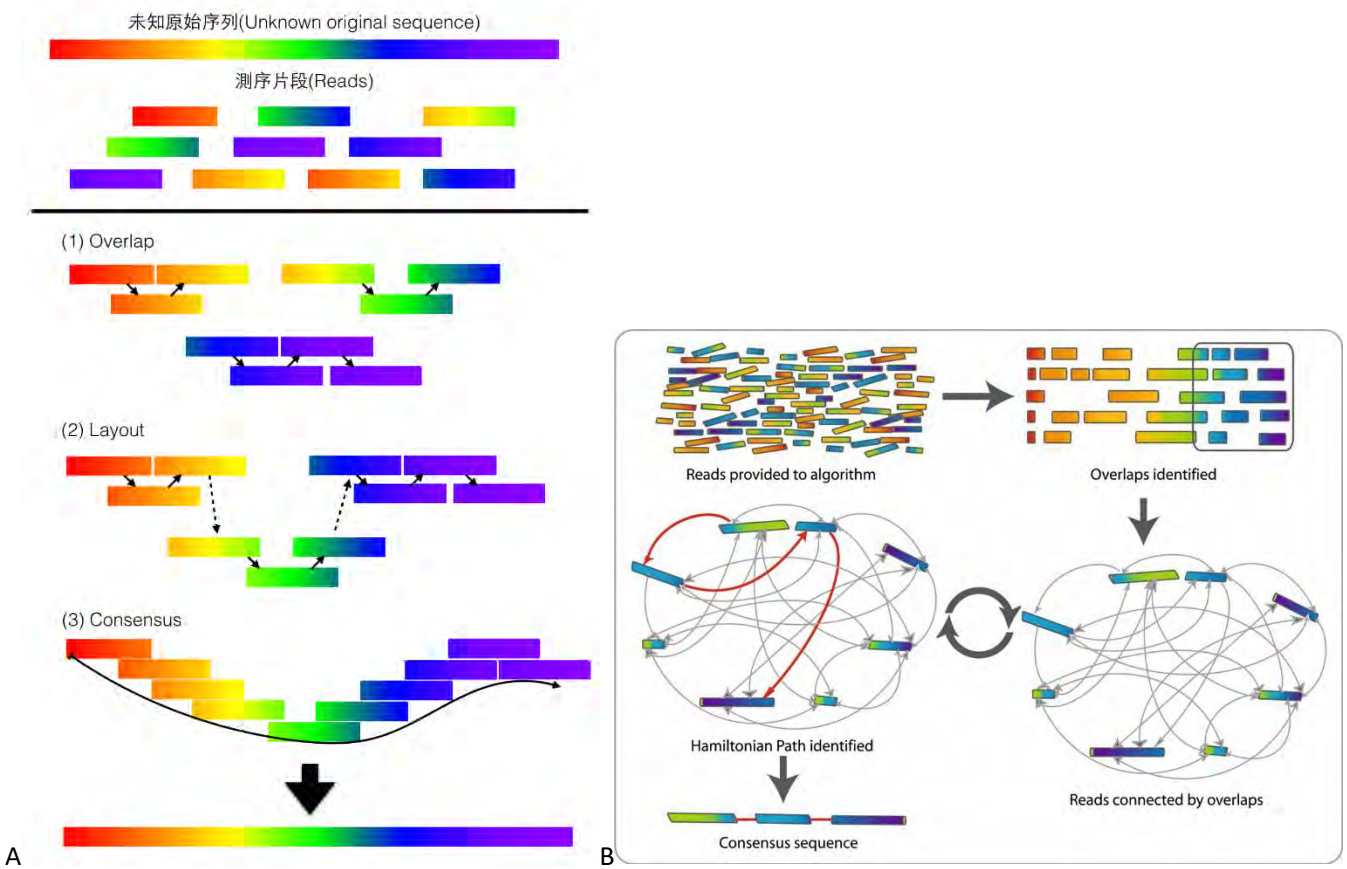


图7.4　OLC 拼接算法模式图

A．基于重叠序列联配；B．基于构图路径

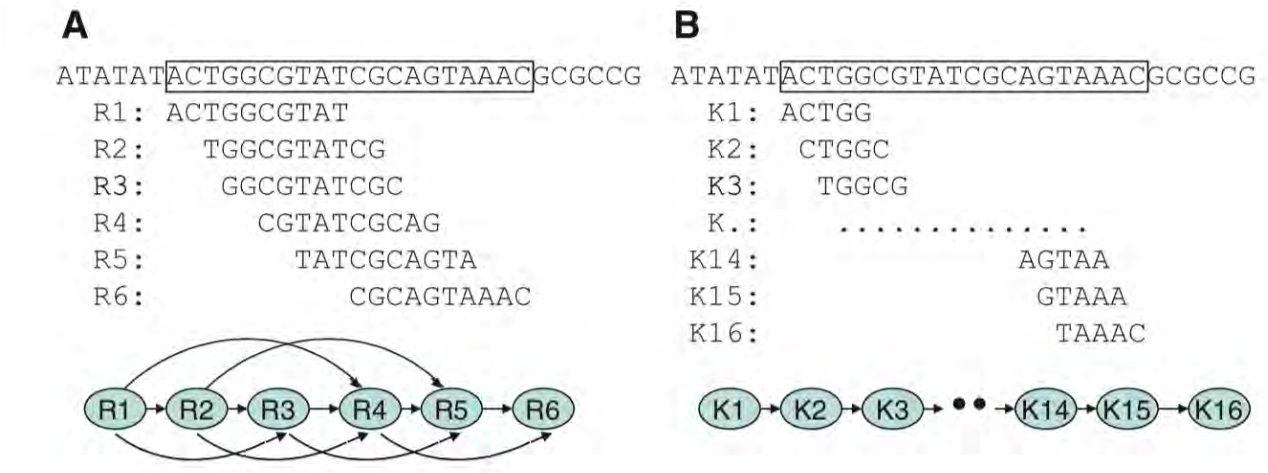


图7.5　OLC 和DBG 算法序列拼接模式图（引自Li et al.，2012a）

A．OLC 算法序列拼接模式图，在该 20bp 的区域产生了 6 个读段（R1～R6），读取长度（*L*）为 10 bp，重叠长度的截止值为 5 bp。读序按照其起始位置和相应的 OLC 图（下方所示）沿基因组有序排列，大部分节点都有超过一个进入的边和出去的边。B．DBG 算法序列拼接模式图，将这些读段切成 *K*-mer（*K*＝5），共有 16 种不同的 *K*-mer，其中大多数发生在一个以上的读段中。*K*-mer 根据其起始位置和基因组 DBG 图的结构，沿基因组排列，大部分的节点都仅有一个进入的边和出去的边

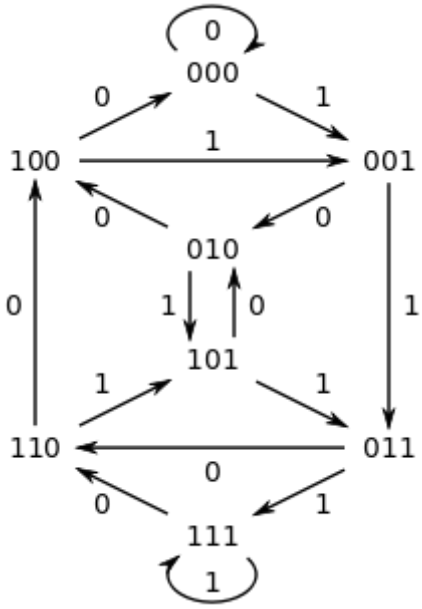


图7.6　德布鲁因路径举例

该图为三维德布鲁因图（顶点由 3 个数字组成，边由 4 个数字组成）

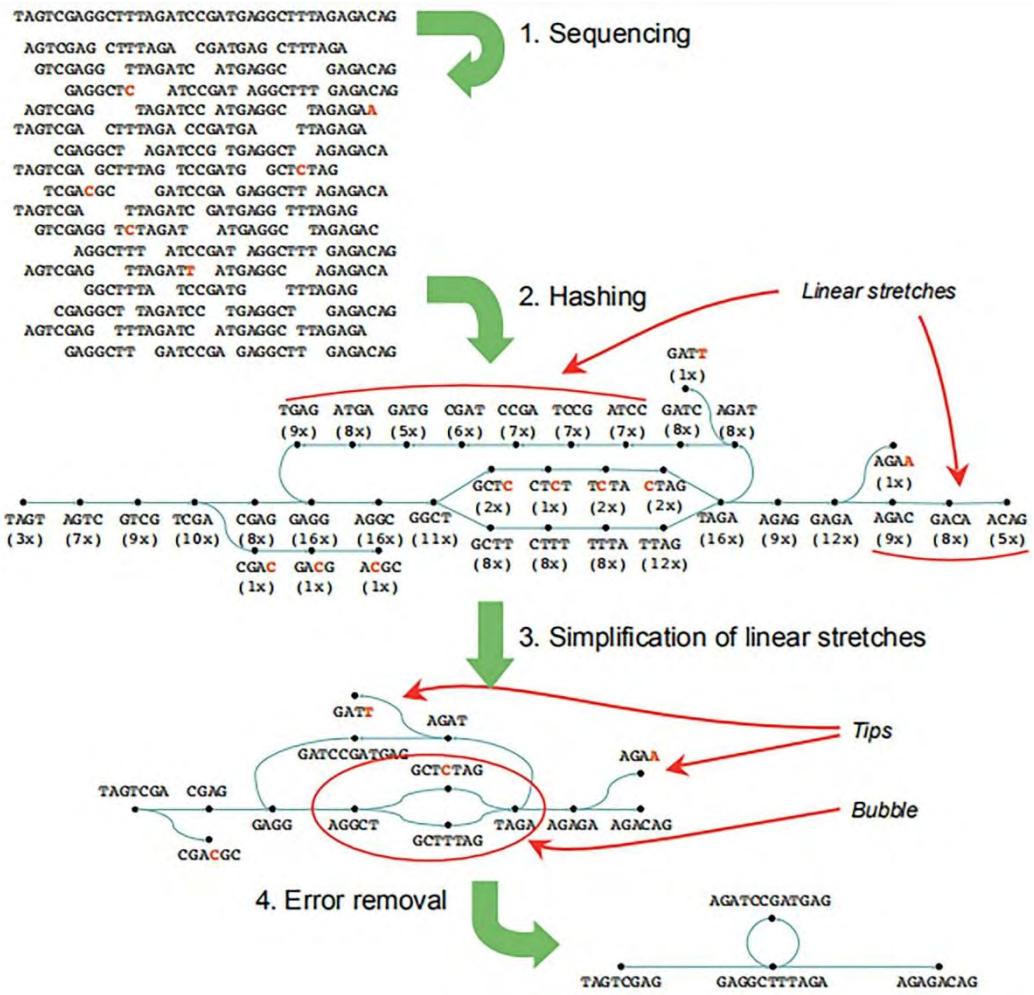


图7.7　基于基因测序读序构建德布鲁因图举例

构图包括四步：①一段基因组序列经测序，获得大量 7nt 长度读序（红色碱基为测序误差）；②基于读序的 4-mer 构建德布鲁因图；③通过 4-mer 延伸简化德布鲁因图；④进一步去除由于测序误差导致的分叉和小包，获得最终图。基于该图的欧拉路径（各边遍历一次）获得原始基因组序列

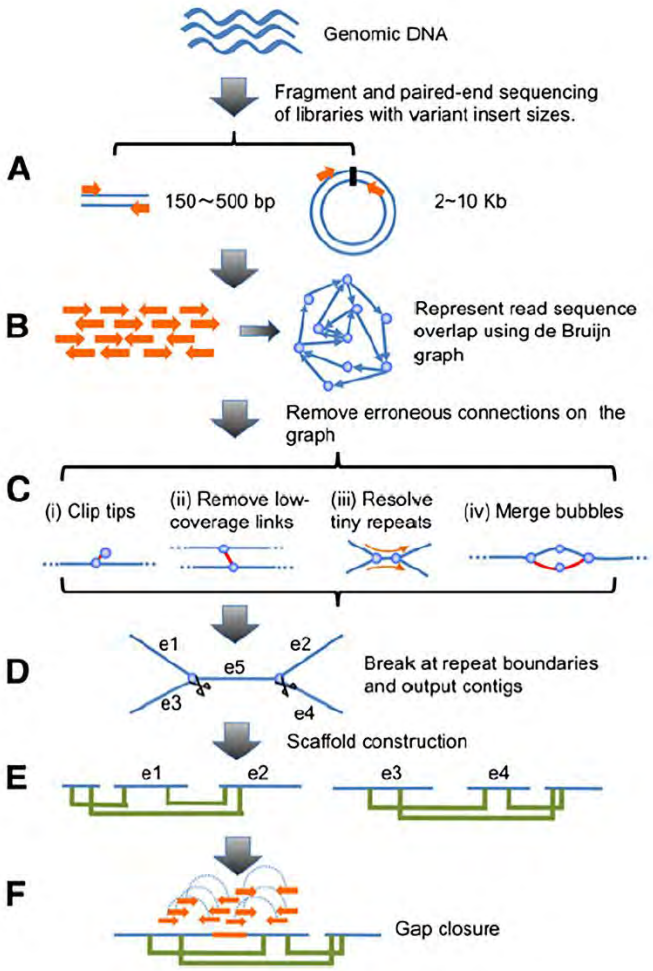


图 7.8　基于德布鲁因图的基因组拼接算法举例（以 SOAPdenovo 为例，引自 Li et al.，2010）

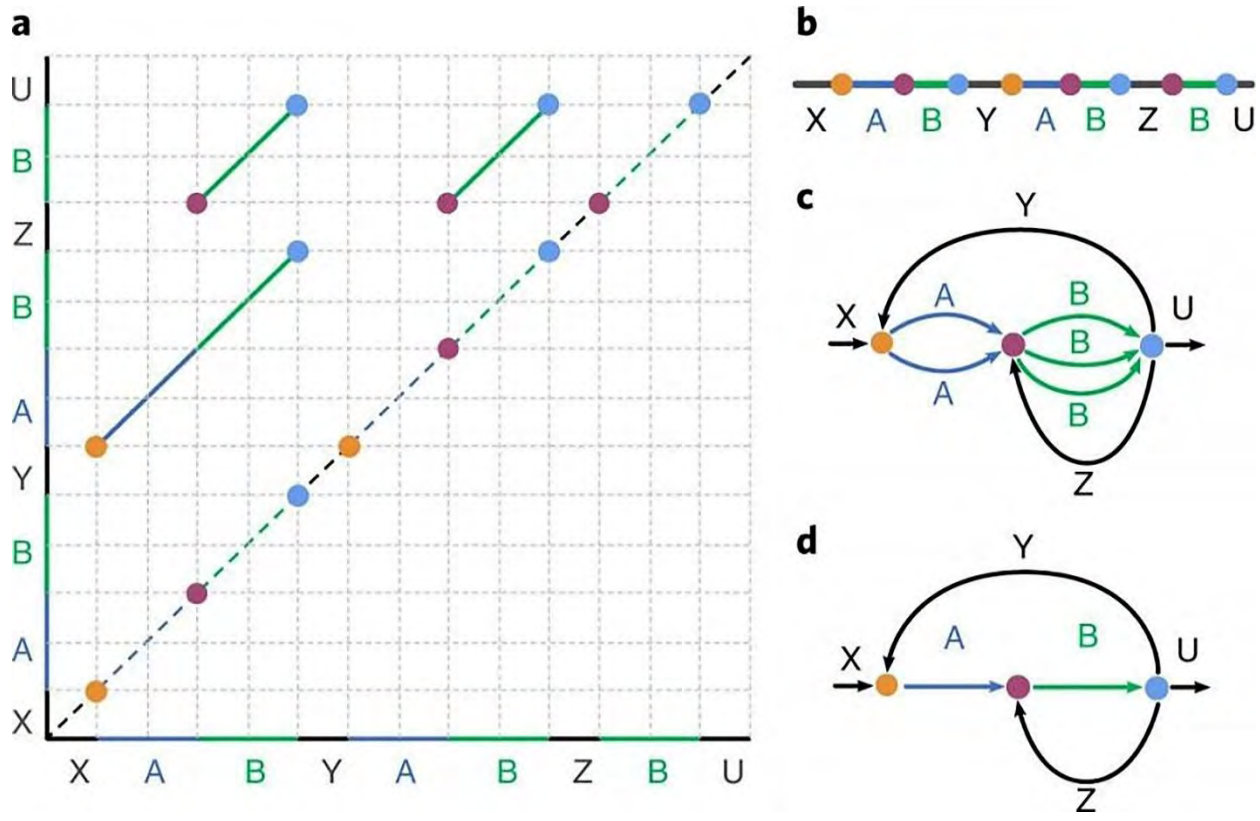


图7.9　基因组片段近似重复图的构建过程（引自Kolmogorov et al.，2019）

A 图．基因组片段序列自身比较的点阵图，该片段中，A 和 B 为重复序列，其他（X/Y/Z/U）为非重复片段；　B 图．基因组片段构成；C 图．构建重复图；D 图．近似重复图

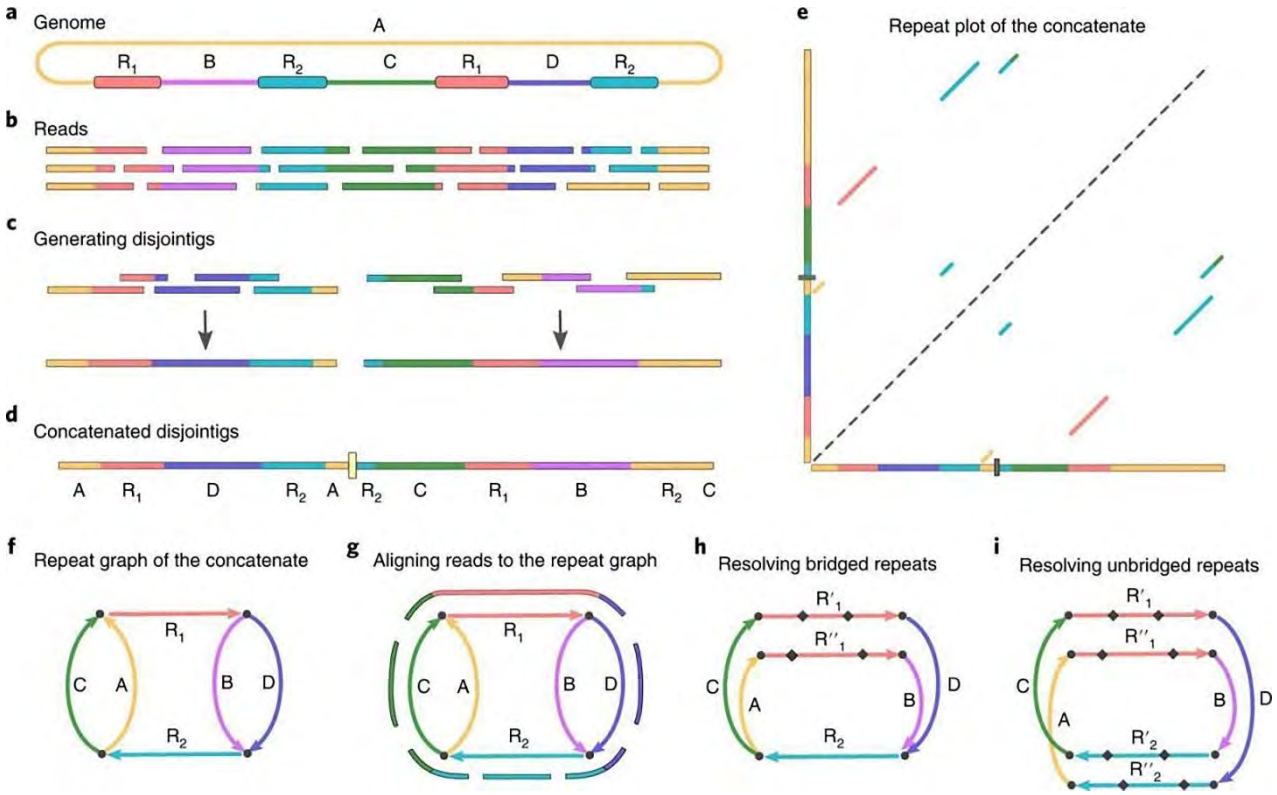


图7.10　基因组拼接Flye 算法（引自Kolmogorov et al.，2019）

A 图．基因组中存在两个重复序列（R_1 和 R_2），分别有两个拷贝，这两个拷贝的序列构成均 99% 一致，A～D 为非重复片段；　B图．长读序及其定位结果；C图．构建随机基因组拼接序列（disjointig），图中构建出两条“disjointig”序列；D、E 图．串连两条“disjointig”序列及该序列的比较点阵图；F 图．基于点阵图构建其近似重复图；G 图．长读序在重复图中的定位情况；H、I 图．基于重复序列不同拷贝的细微变异，重构完整重复图路径

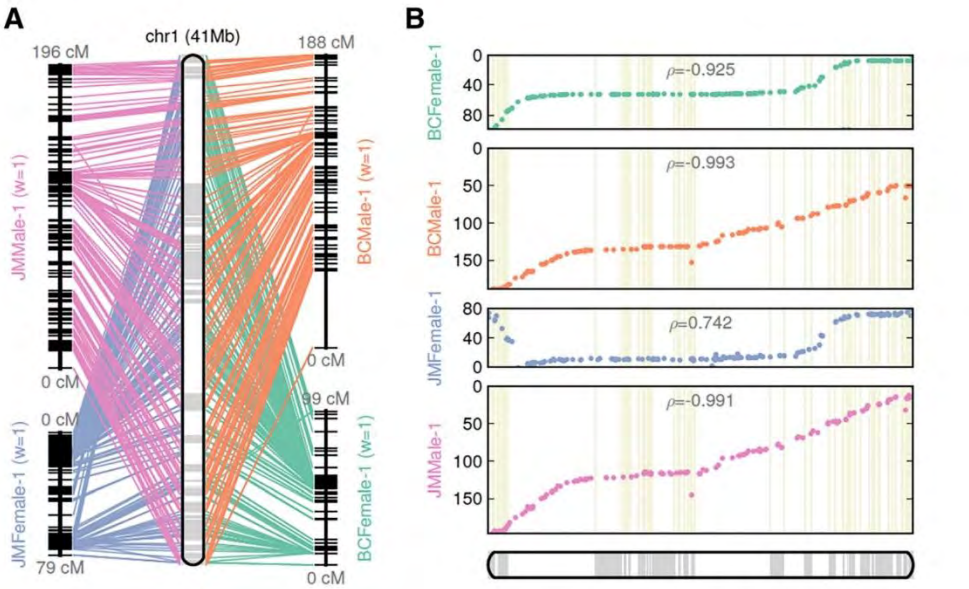


图7.11　利用ALLMAPS 重建的黄鱼基因组一号准染色体（引自Tang et al.，2015）

A．图两侧代表的是 4 个遗传图谱，采用 BCFemale、BCMale、JMFemale 和 JMMale 4 个遗传图谱，权重（w）相同，黑色横线代表遗传标记，中间代表 1 号准染色体序列，灰白相间代表连接的每一条 scaffold，遗传图谱和染色体序列之间的线连接重建染色体上的物理位置和图谱上的遗传距离。B．4 个散点图，点分别代表染色体上的物理位置（x 轴）与遗传图谱的遗传距离（y 轴）。散点图上的 ρ 值代表 Pearson 相关系数，其值为－1～1，接近－1 或 1 的值表示接近完美的共线性

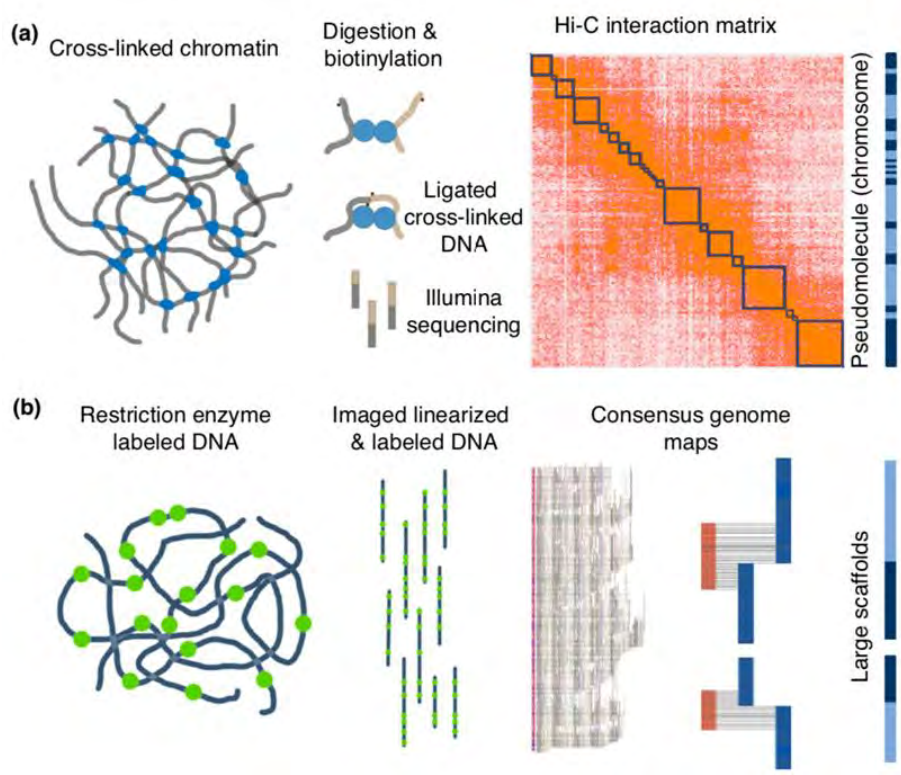


图7.12　长跨度技术辅助准染色体重构（引自Michael and van Buren，2020）

A． Hi-C 辅助准染色体重构模式图，Hi-C 数据获取自交联染色质的相互作用，然后通过 contig/scaffold 上位点之间的相互作用来构建矩阵，进而确认它们之间的位置关系，以 contig/scaffold 重建准染色体。B．光学图谱辅助构建准染色体模式，光学图谱利用限制酶位点和单分子成像来创建基因组的物理图。通过使用限制酶对 DNA 的长片段进行刻痕并标记，DNA 分子被线性化和成像，并且每个分子的指纹被合并以创建一个共有的基因组图。基于比对，contig/scaffold 被覆盖在基因组图谱上，并锚定在支架或假分子中，进而构成准染色体

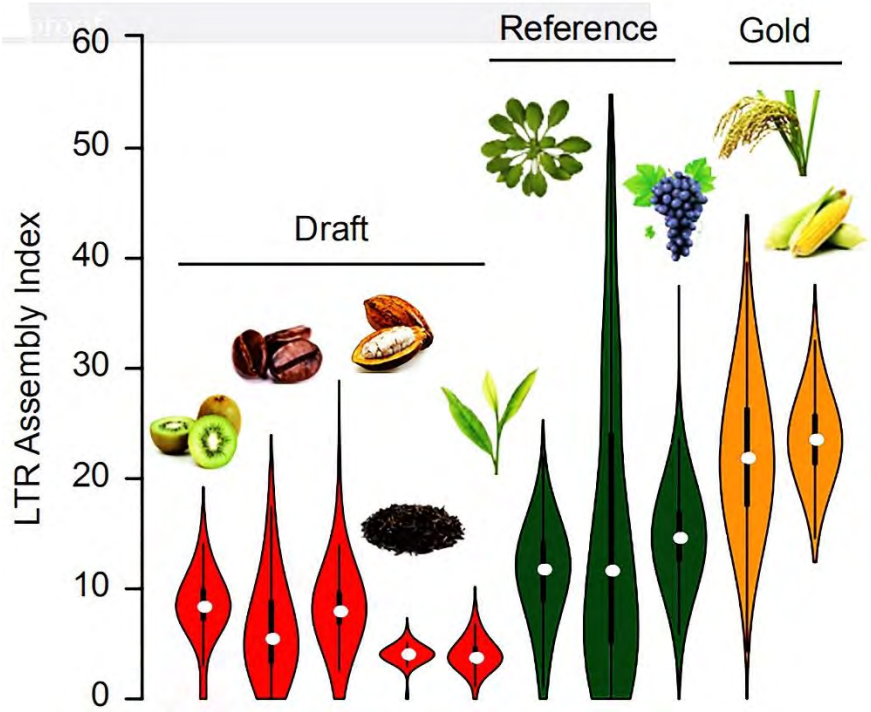


图7.13　利用LTR 重复序列对不同植物物种基因组序列完整度进行评价（引自Xia et al.，2020）

基于 LAI 指标值大小，可以根据基因组完整程度分为草图（≤10）和参考基因组（＞10）

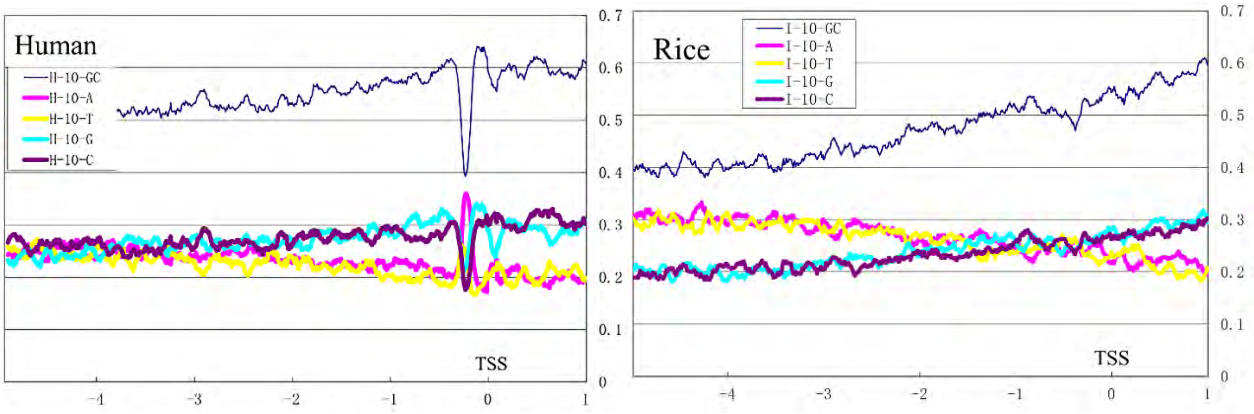


图7.14　人类（左）和水稻（右）基因转录起始位点（TSS）附近碱基分布变化

随机挑选 100 条基因序列以转录起始位点对齐，按照 10nt 窗口长度逐碱基计算 4 种碱基频率

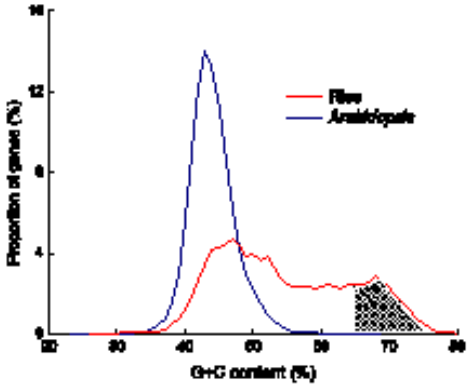


图 7.15　水稻和拟南芥具有不同 GC 含量蛋白质编码基因分布图 （引自 Guo et al.，2007）

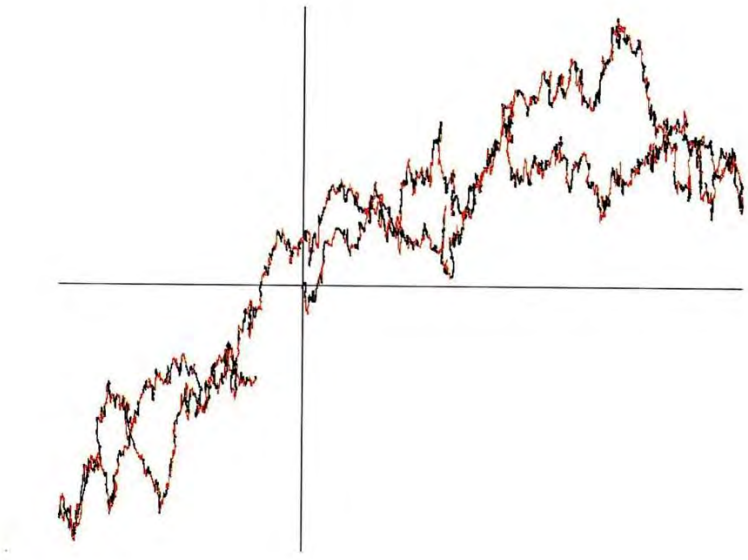


图 7.16　大肠杆菌 K12 菌株基因组序列的二维 DNA 行走曲线（引自郝柏林，2015）

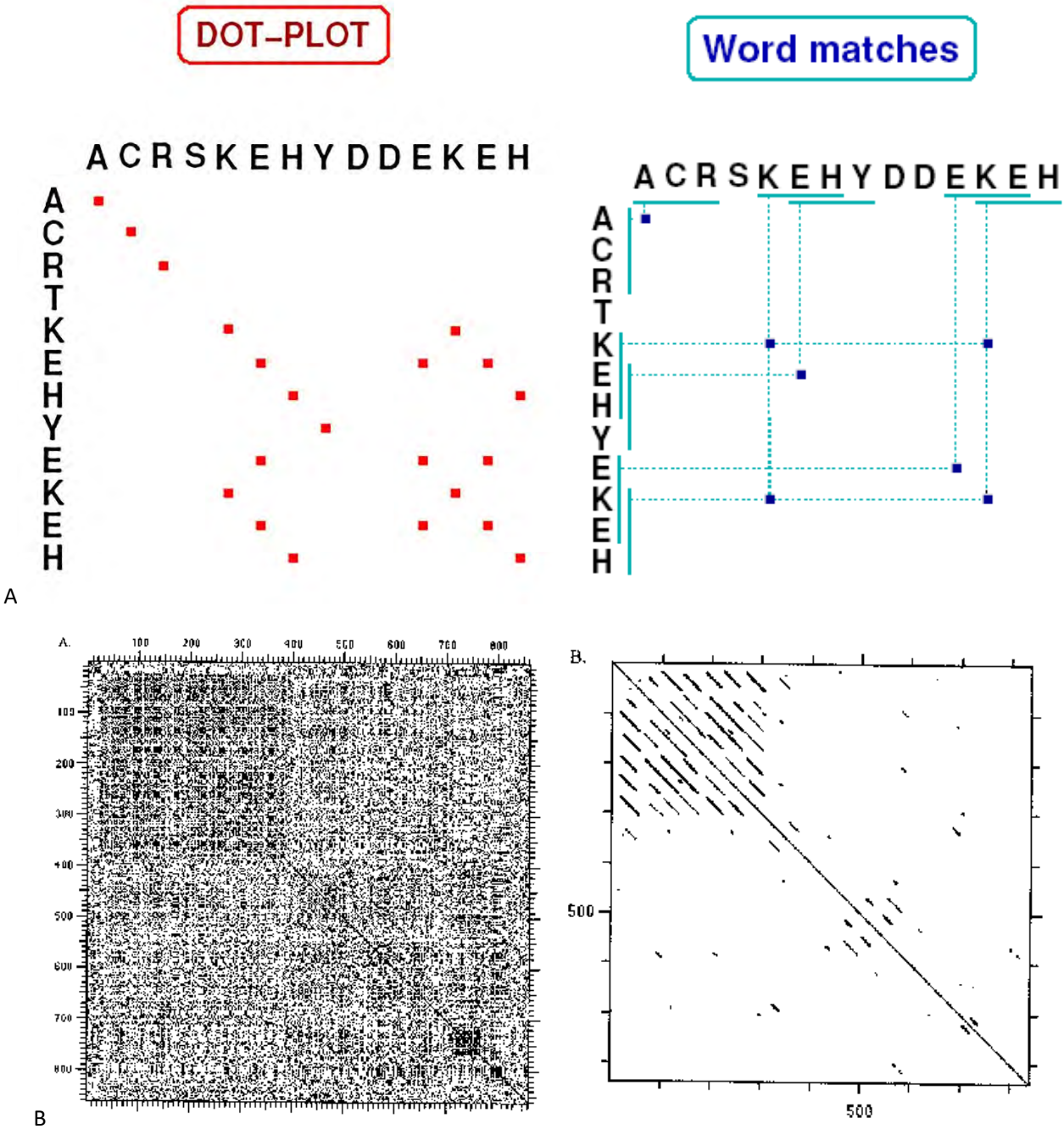
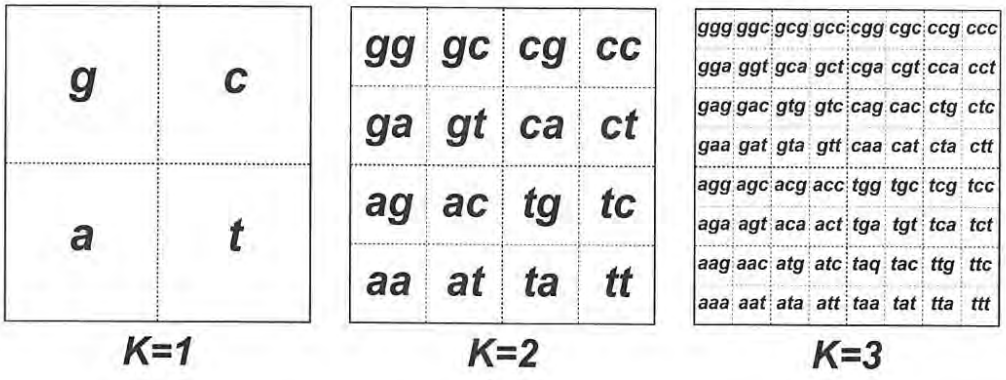


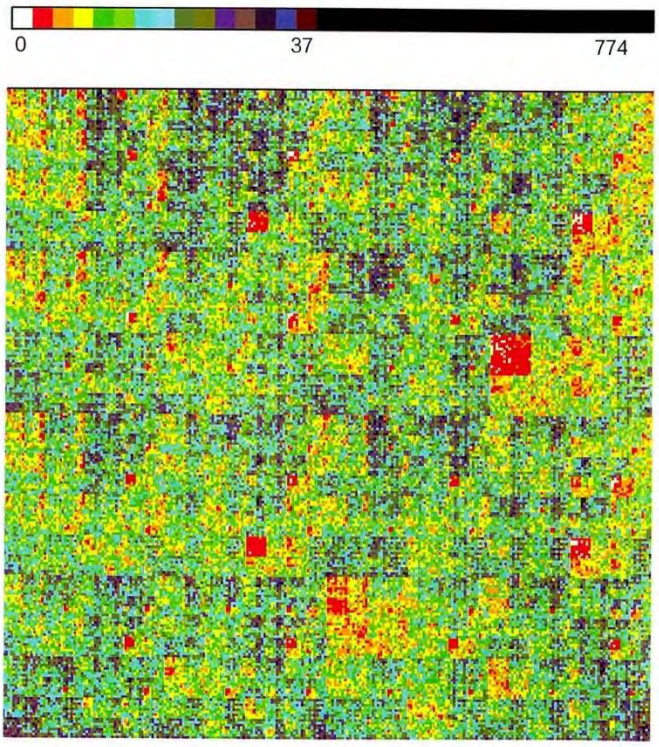
图7.17　利用点阵图（dot matrix）方法进行序列保守性分析举例（引自Zhang et al.，2005）

A．两条蛋白质序列不同字长（单个和三个氨基酸）的分析效果；

B．人类 LDL 受体基因不同字长（单碱基和 23nt）的自身比较图



A



Escherichia coli strain K12 (*K*=8)

B

图 7.18　基因组 *K* 框架（A）及大肠杆菌 K12 基因组“肖像”（B）（引自郝柏林，2015）

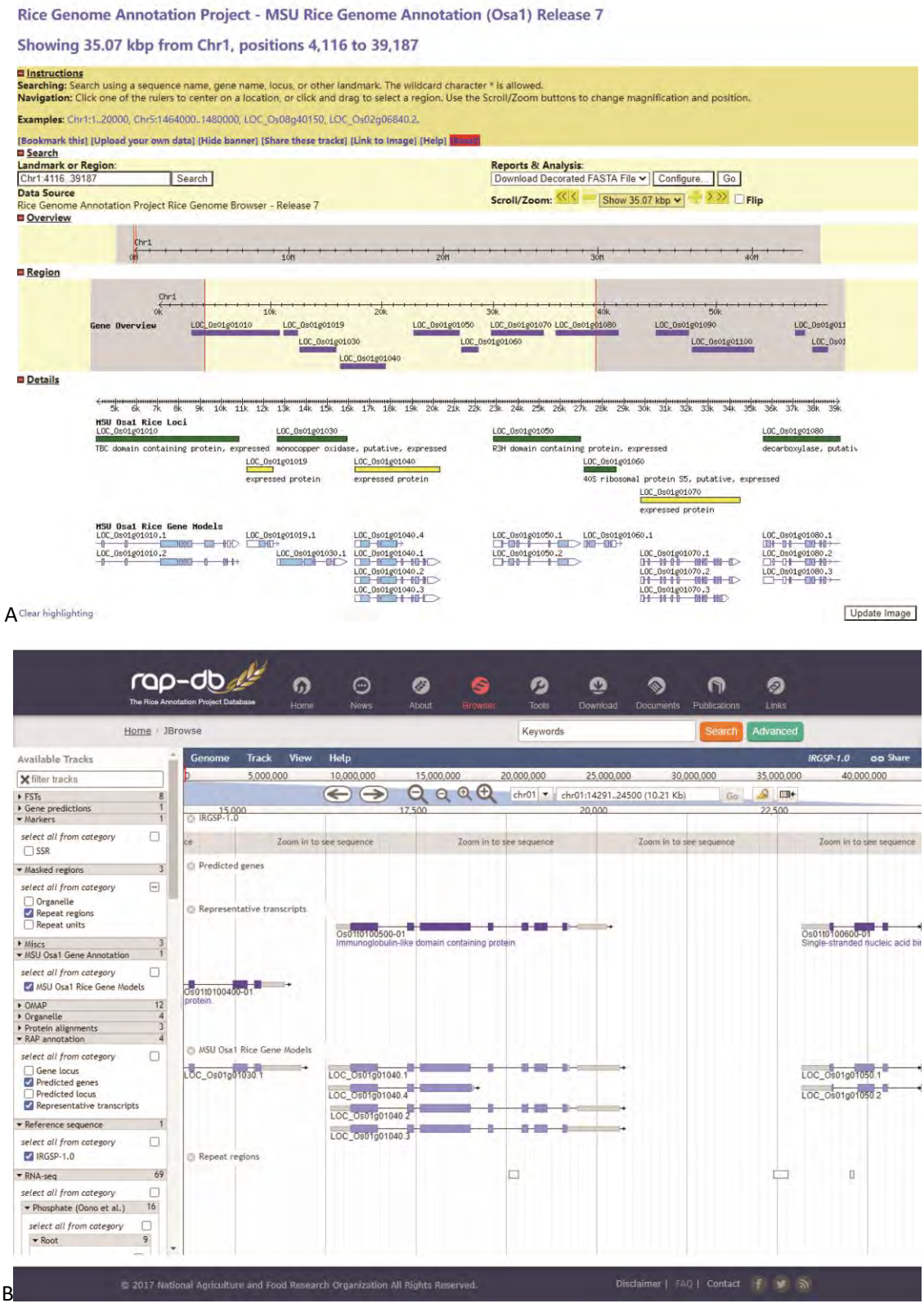


图7.19　基因组浏览器应用举例

A．以 GBrowse 为基础构建的水稻数据库（Rice Genome Annotation Project）基因组浏览器；B．以 JBrowse 为基础构建的水稻数据库（RAP-DB）基因组浏览器

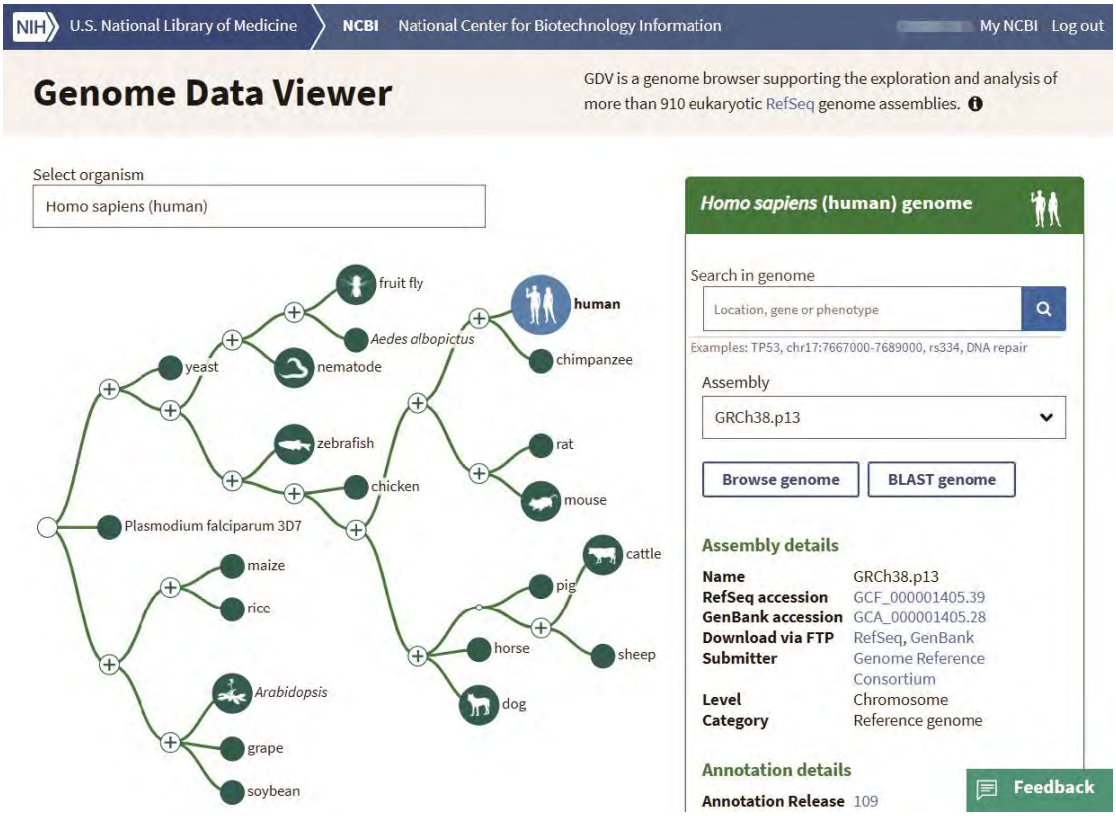


图 7.20　Genome Data Viewer 基因组浏览器

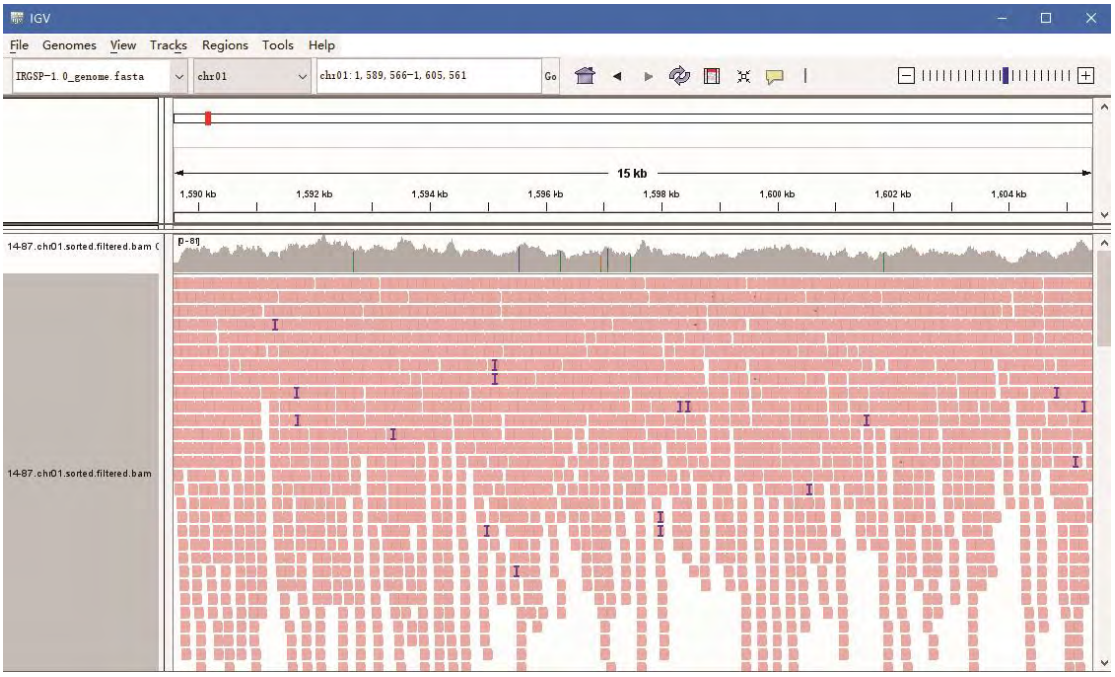


图 7.21　交互式基因组可视化工具 IGV

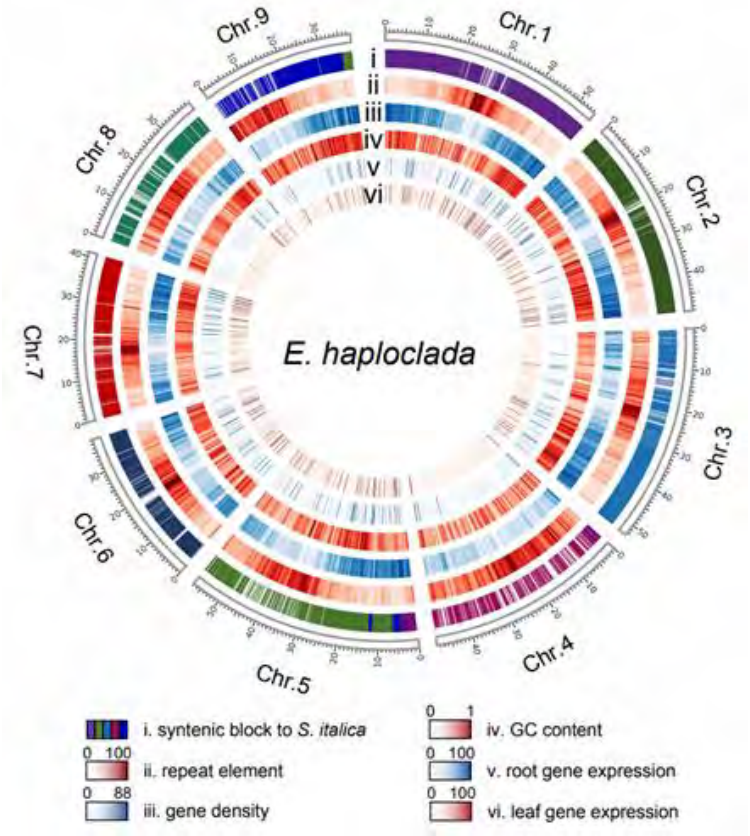


图 7.22　Circos 图举例——稗属物种 *Echinochloa haploclada* 基因组及其相关信息（引自 Ye et al.，2020）

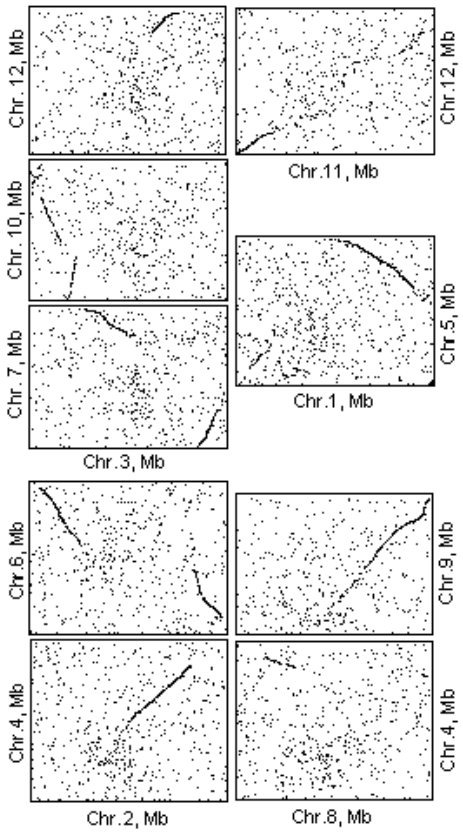


图7.23　利用点阵图方法进行染色体之间的比较（引自Zhang et al.，2005）

图示水稻 12 条染色体之间的比较，图中可见基因组共线性区域

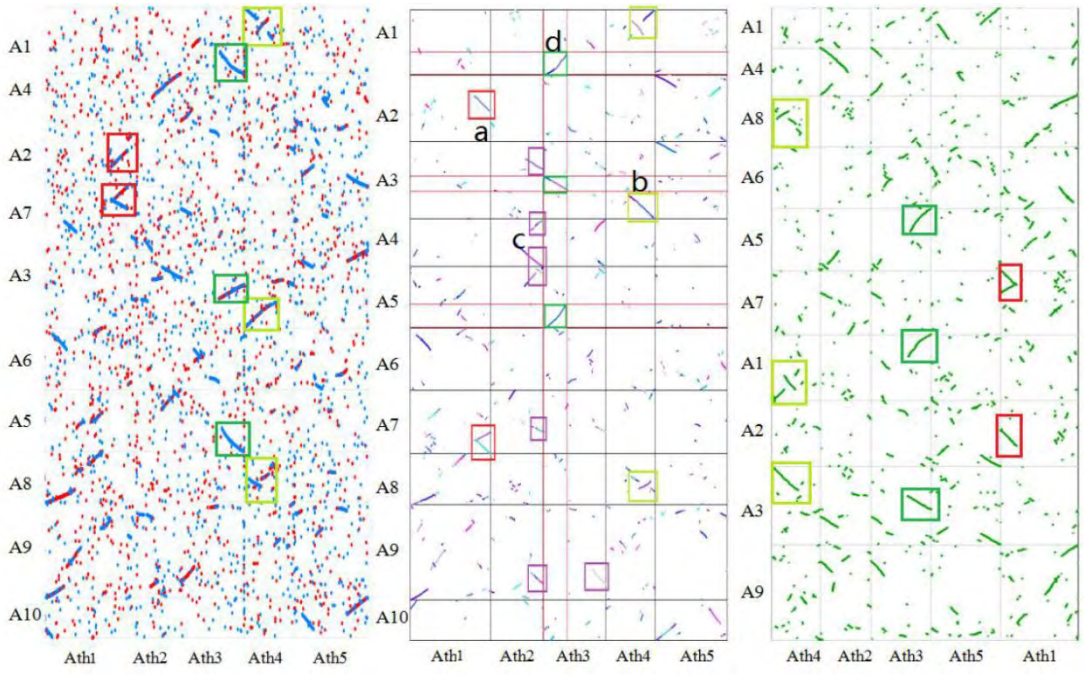


图7.24　基于三种基因组共线性分析方法的拟南芥（*x* 轴，染色体Ath1～Ath5）与 白菜（*y* 轴，染色体A1～A10）基因组共线关系

从左到右：MUMmer（基于基因组 DNA 序列比对）、MCScanX（基于同源基因顺序关系比对）和 COGE（基于基因组编码区序列比对）

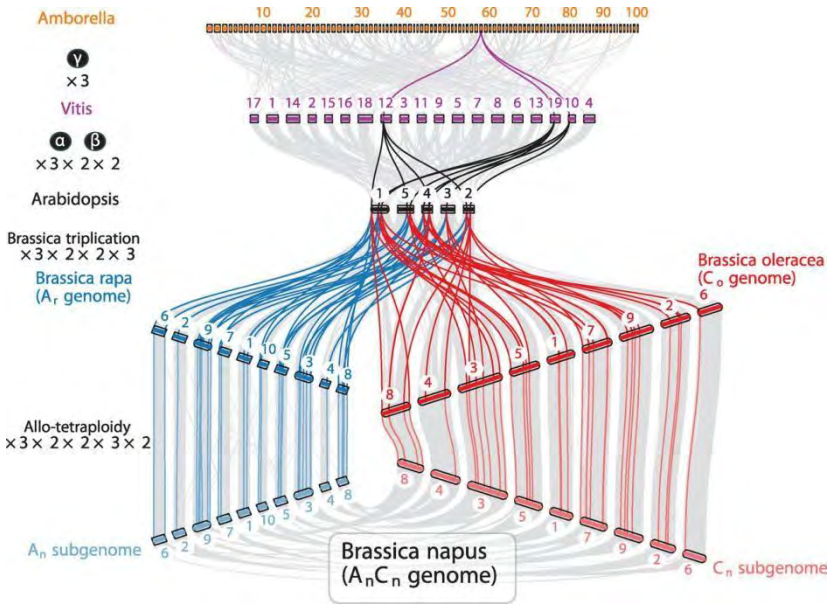


图7.25　使用JCVI 绘制的甘蓝型油菜进化过程中发生的多次基因组加倍事件（引自Chalhoub et al.，2014）

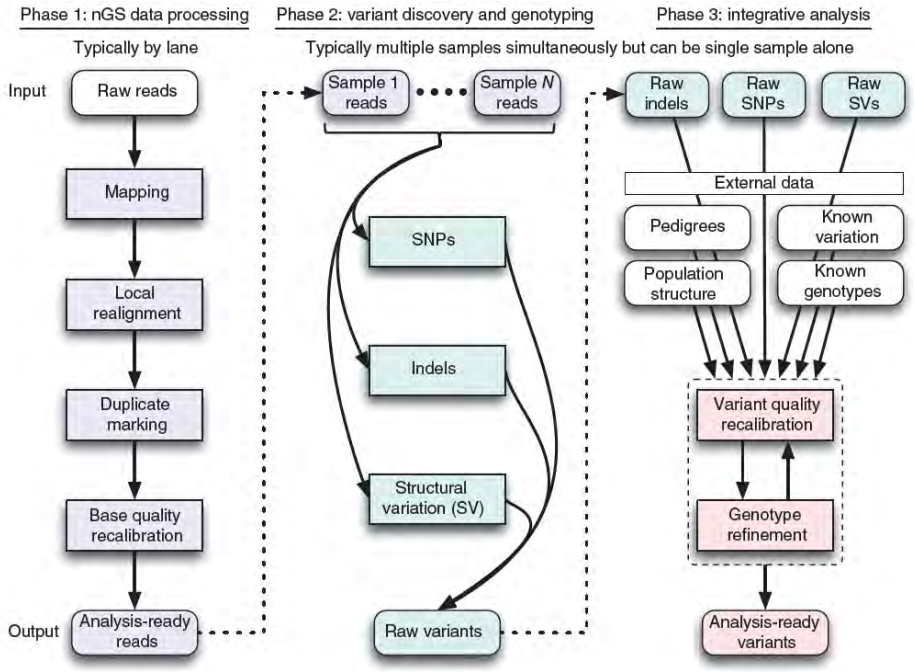


图 7.26　基因组重测序数据变异检测流程及其应用（引自 DePristo et al.，2011）

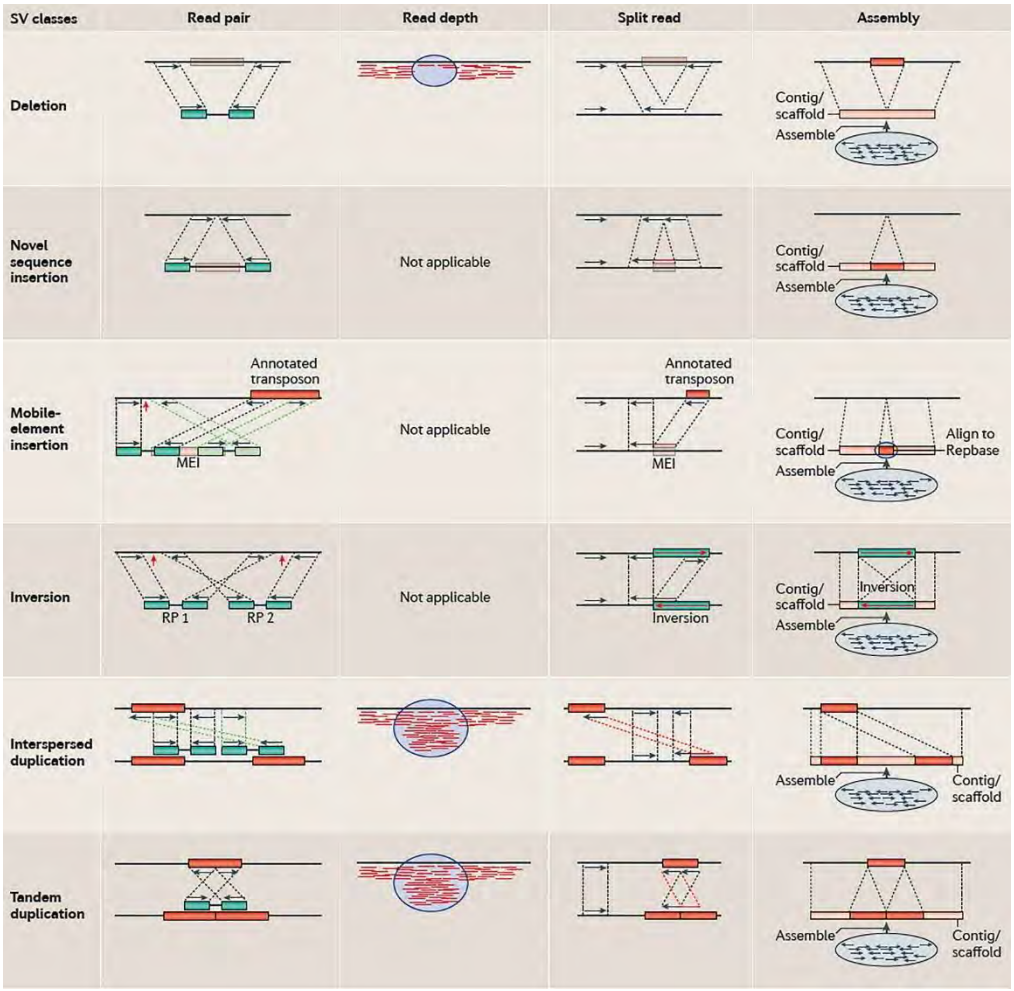


图7.27　4 种基因组结构变异主要检测策略（引自Alkan et al.，2011）

理论上双端读序联配法、拆分读序法、从头拼接法这三种方法可以检测各种类型的结构变异，不过每种方法都会存在一定的错误偏向

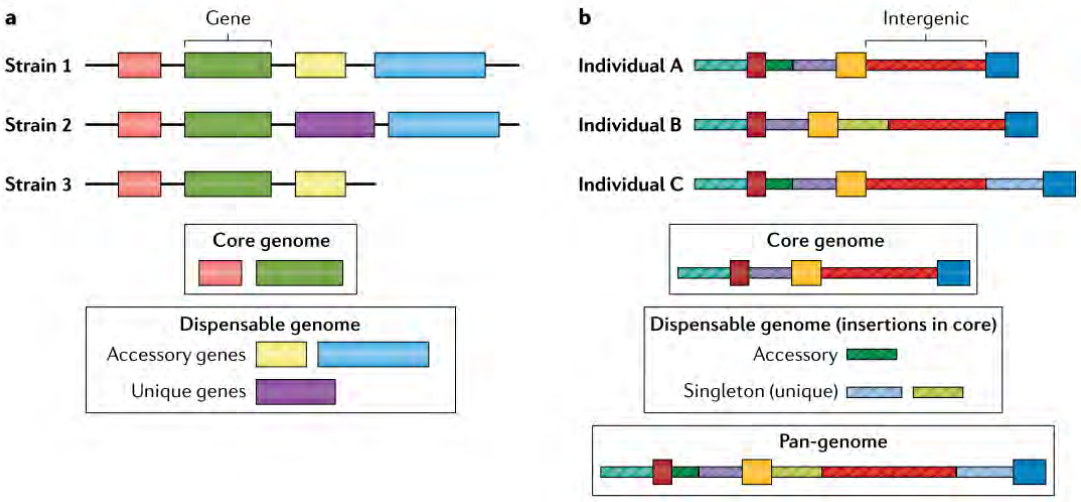


图7.28　泛基因组的构成（引自Sherman et al.，2020）

A.　细菌及其他原核生物基因组主要由基因构成，因此其泛基因组定义为一个物种所有基因的总和；B.　真核生物基因组中存在很高比例的基因间区，其泛基因组定义为所有 DNA 序列的集合，包括基因区和基因间区

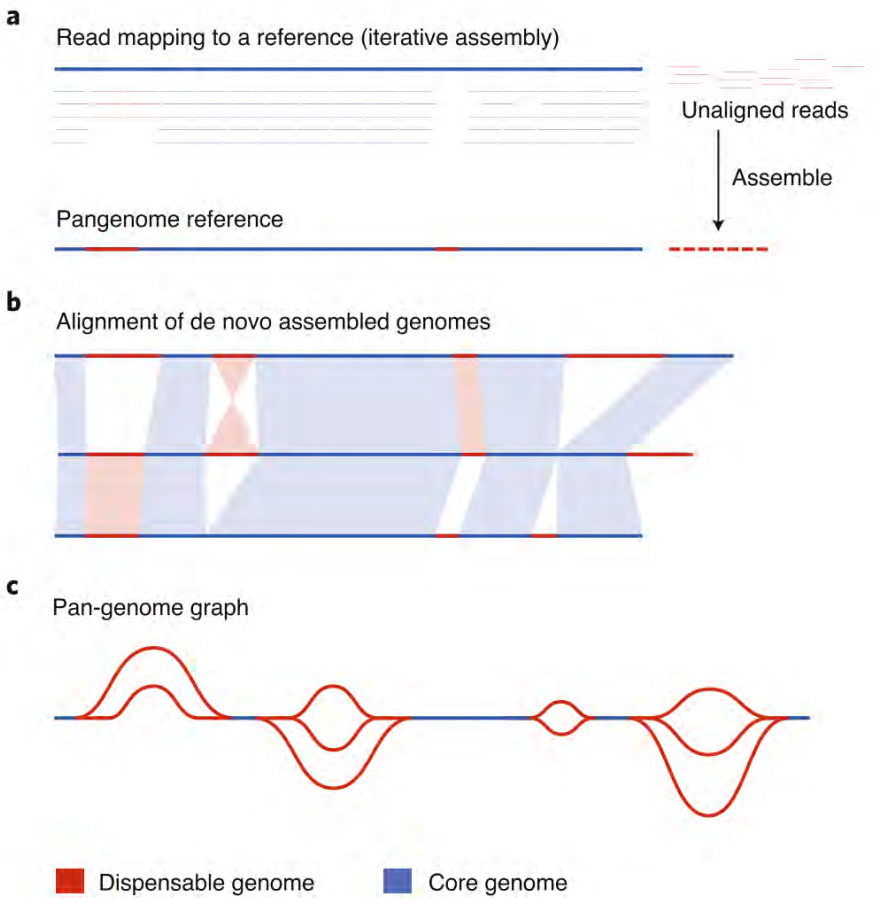


图 7.29　泛基因组组装三种策略（引自 Bayer et al.，2020）

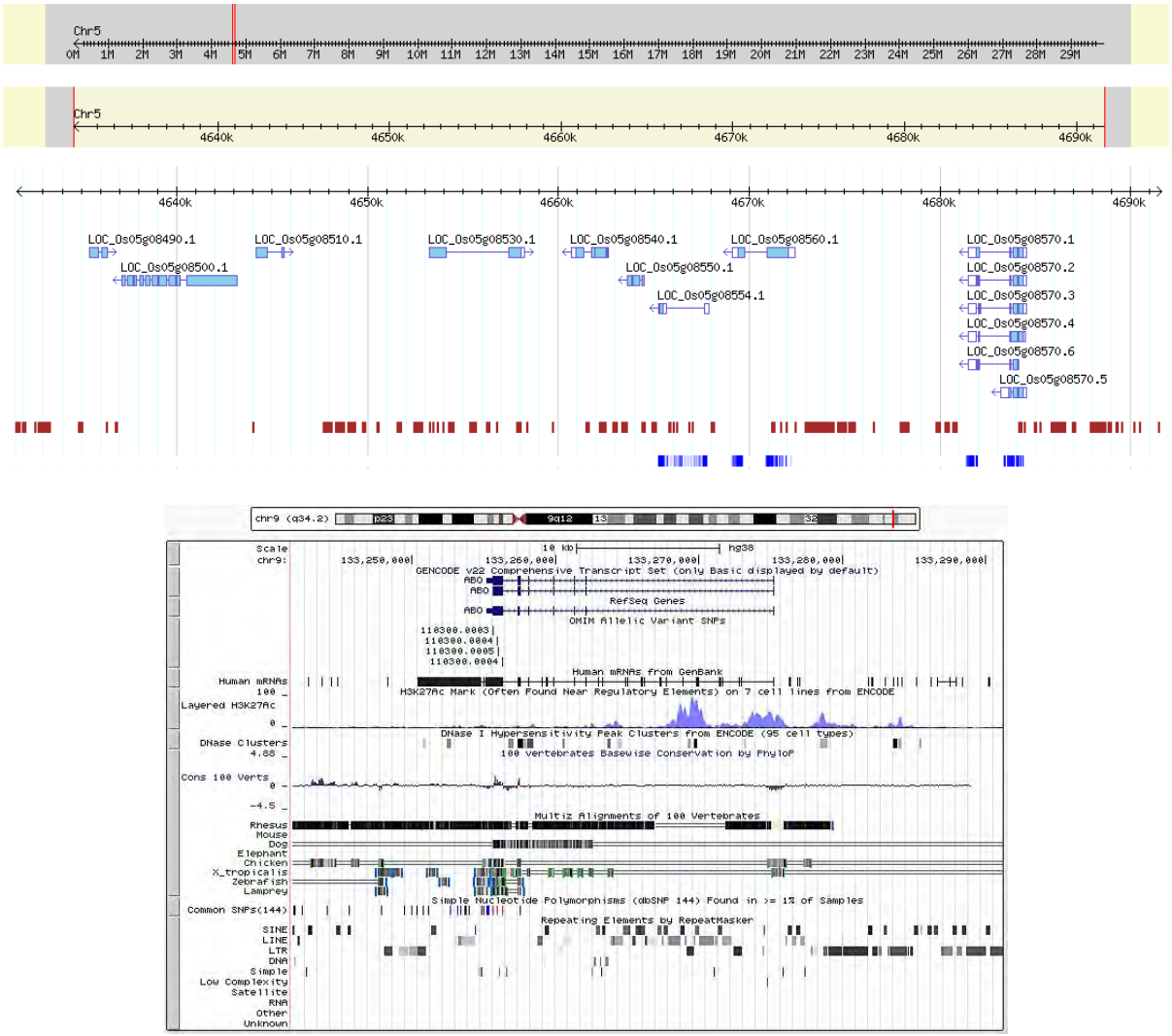


图 8.1　水稻（上）和人类（下）基因组序列构成举例

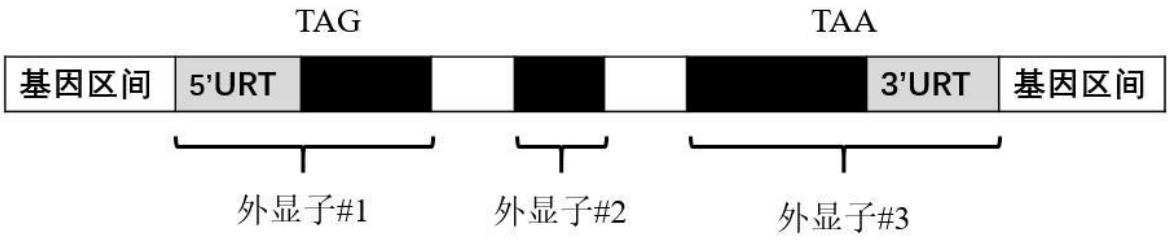


图8.2　一种典型蛋白质编码基因的结构示意图

蛋白质编码区（CDS，黑色区域）包括大部分外显子序列（除了两端 UTR 序列），自起始密码子（TAG） 开始，到终止密码子（TAA 等）结束

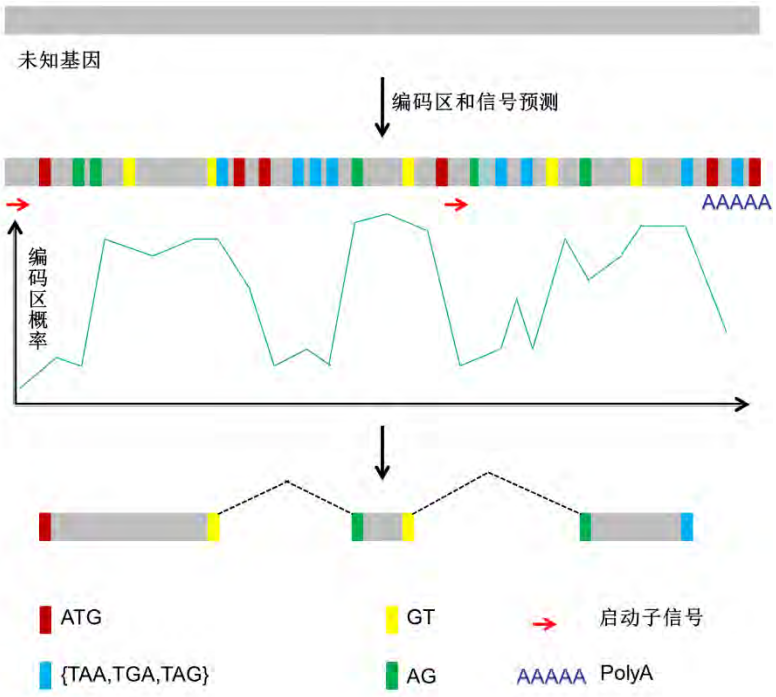


图8.3　基因从头预测方法模式图

其中编码区概率估计往往根据一定的概率模型（如 HMM）

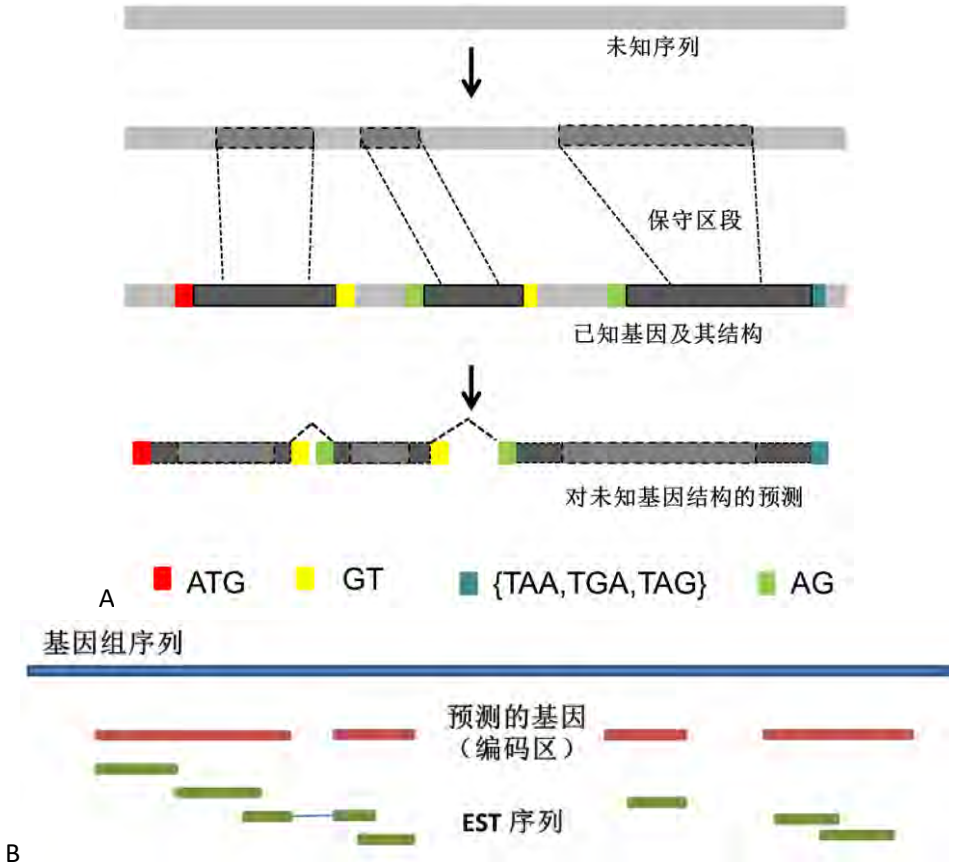


图8.4　同源比对预测基因模式图

A．基于近缘物种已知基因结构；B．基于基因表达序列（如 EST）

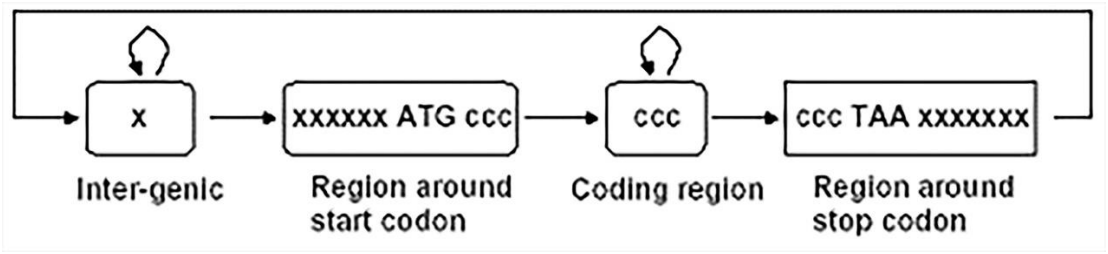


图 8.5　蛋白质编码基因预测 HMM 模型举例

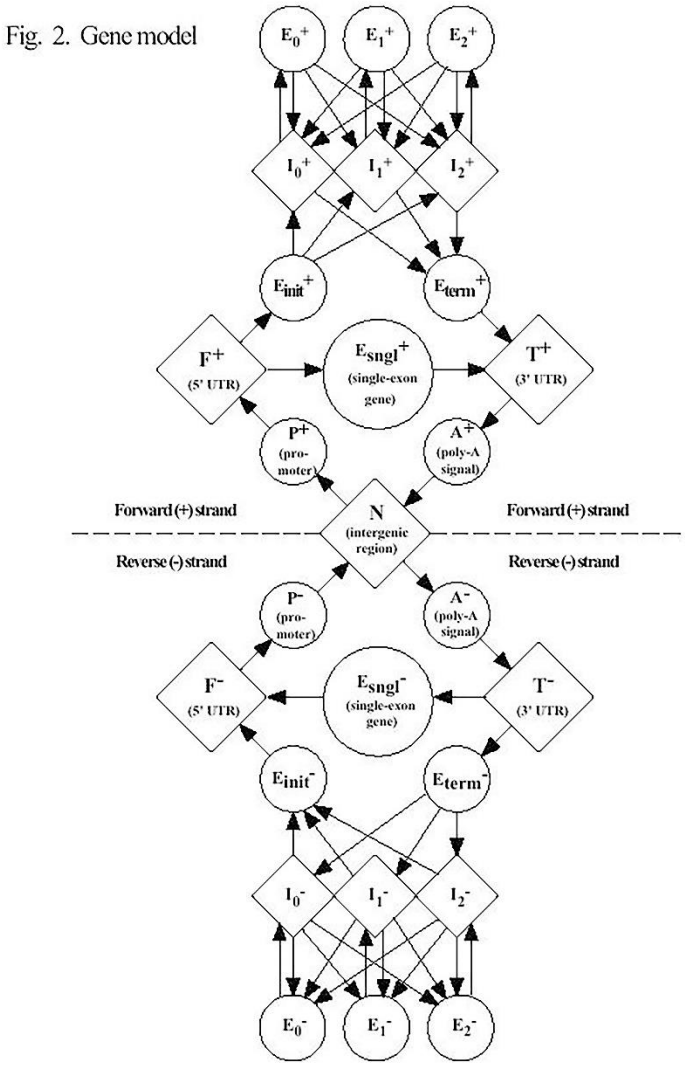
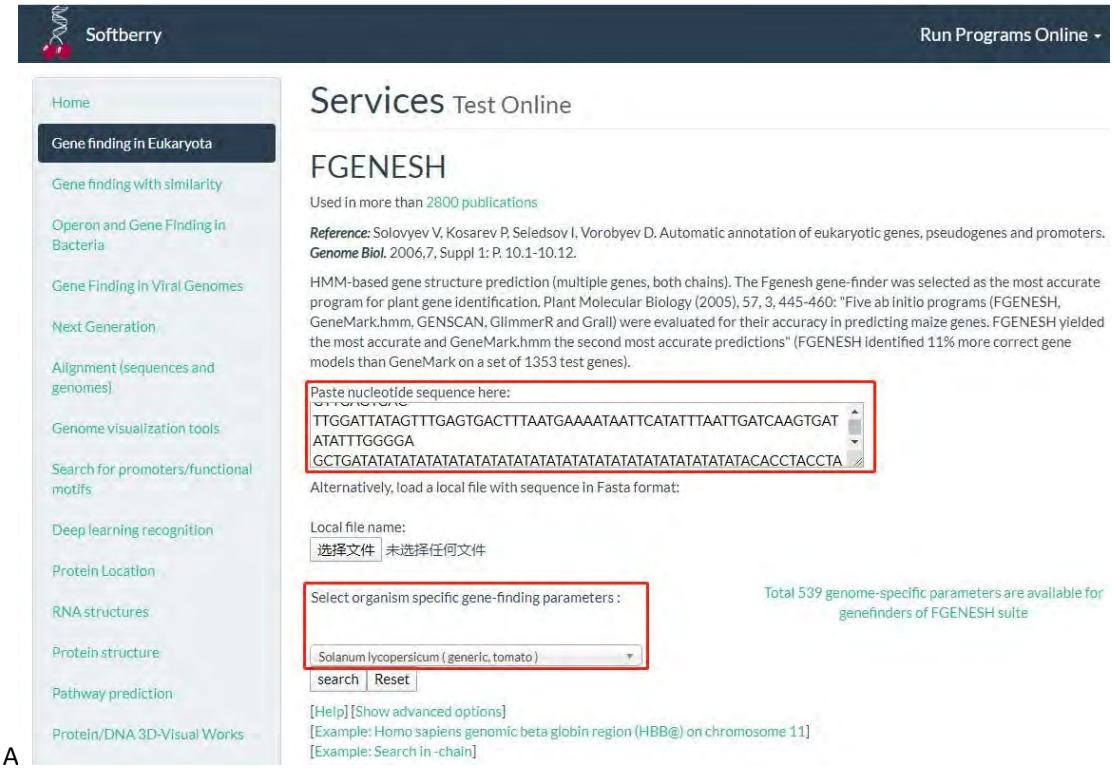


图8.6　基因预测工具GENSCAN的HMM模型

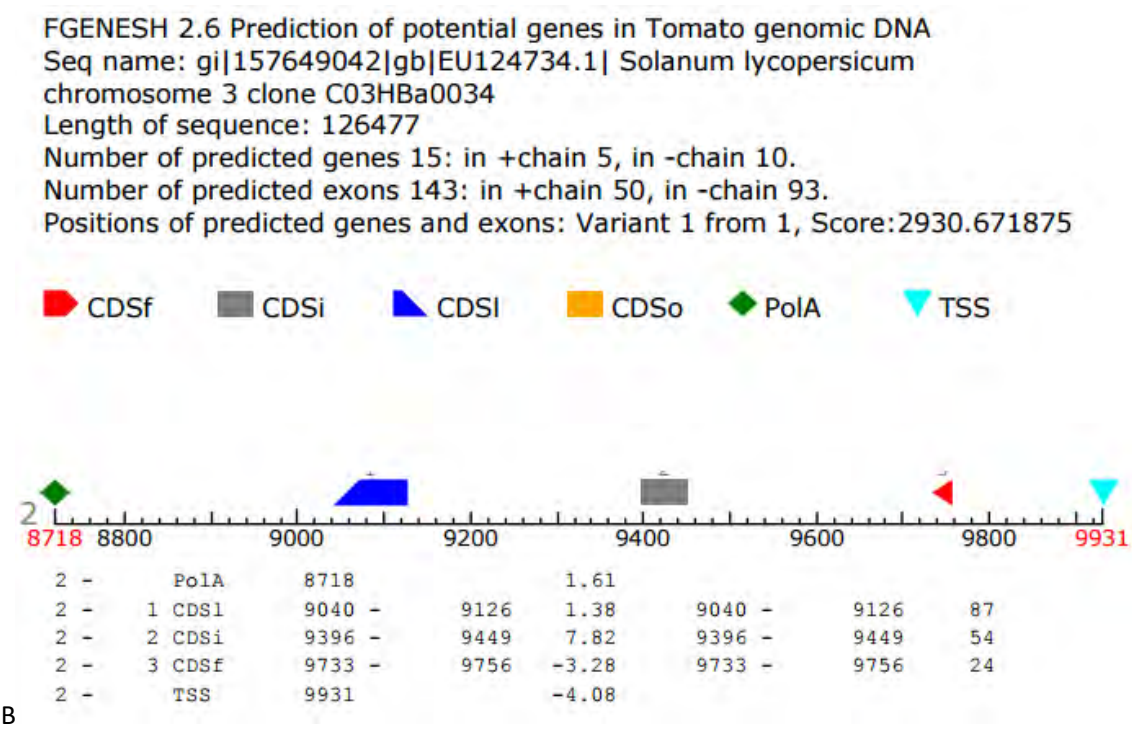
（引自Burge and Karlin，1997）

图中E、I、F、T等字母均表示特定区域，例如，E表示外显子、

I表示内含子、F和T分别表示5′和3′非转录区域（UTR）等，其上标正/负号表示序列正/负链，下标表示单个（sngl）、起始（init）、终止（term）和中间（0～2）外显子/内含子



A



B

图8.7　基于HMM 模型的主流基因预测工具FGENESH

A. FGENESH 提供的在线基因预测服务平台主页。B. 利用 FGENESH 进行基因预测的结果举例。一段来自番茄约 120kb 基因组序列（EU124734）的预测结果。共 15 个基因被预测出，图中仅列出其中一个基因的具体预测结果。图中红色的 CDSf 代表基因模型中的第一个外显子；灰色的 CDSi 代表中间的外显子；蓝色的 CDSl 表示最后一个外显子；如果仅有一个外显子则用橙色的 CDSo 表示；淡蓝色的 TSS 和深绿色的 PolA 分别代表转录起始位点和 poly A 尾巴结构

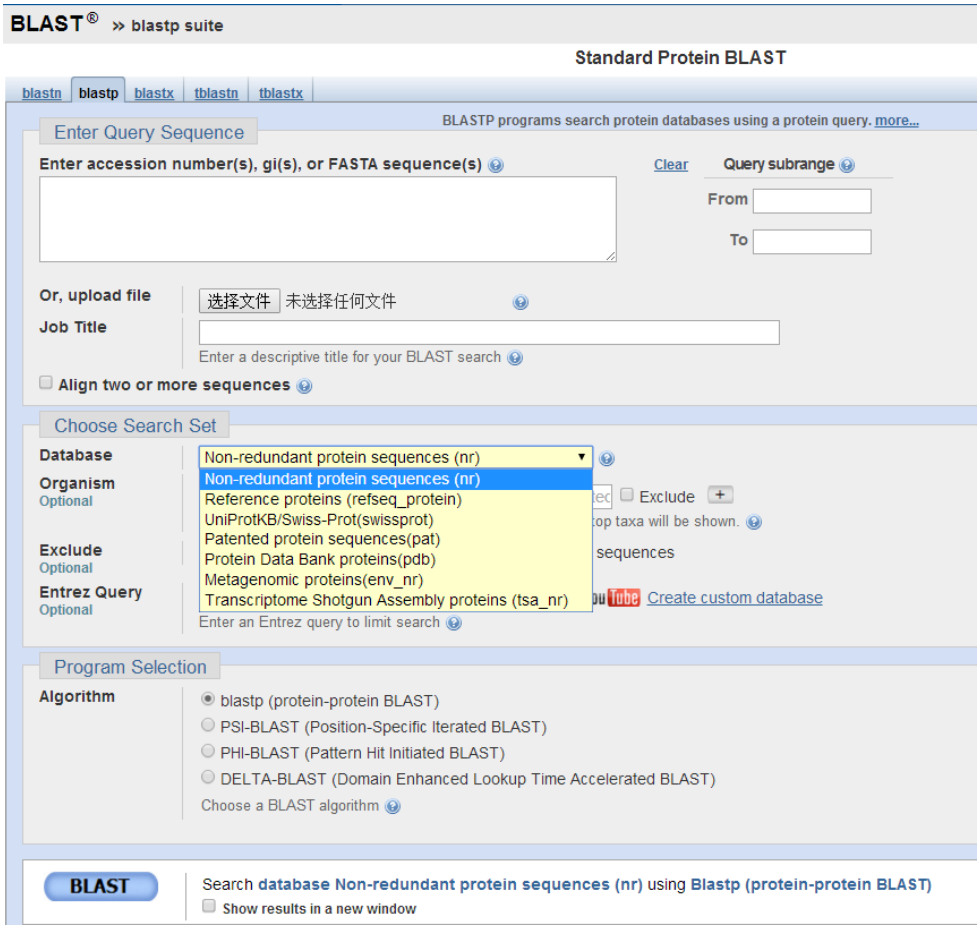


图8.8　通过BLAST和选用不同蛋白质数据库进行基因功能注释

图 8.9　KAAS（KEGG Automatic Annotation Server）注释服务器主页

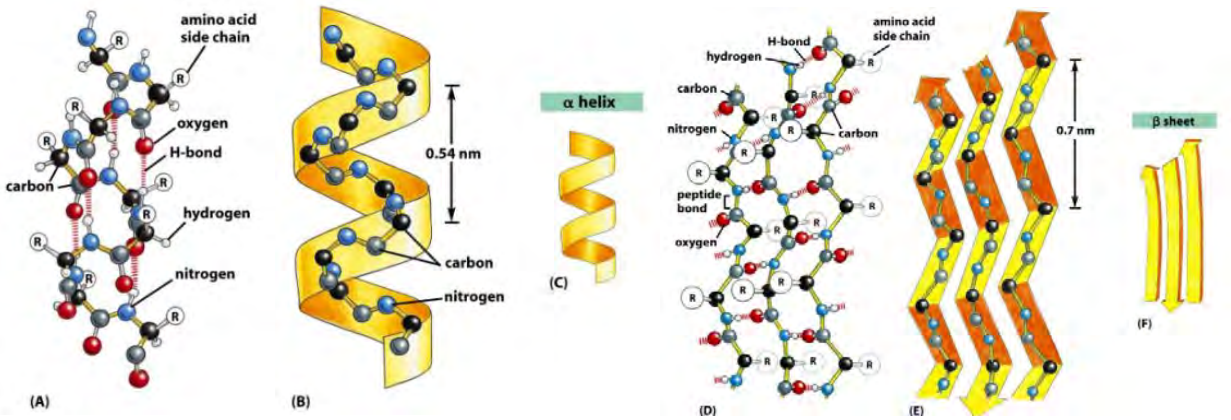


图 8.10　α 螺旋（A～C）和 β 折叠（D～F）结构示意图（引自 Alberts et al.，2007）

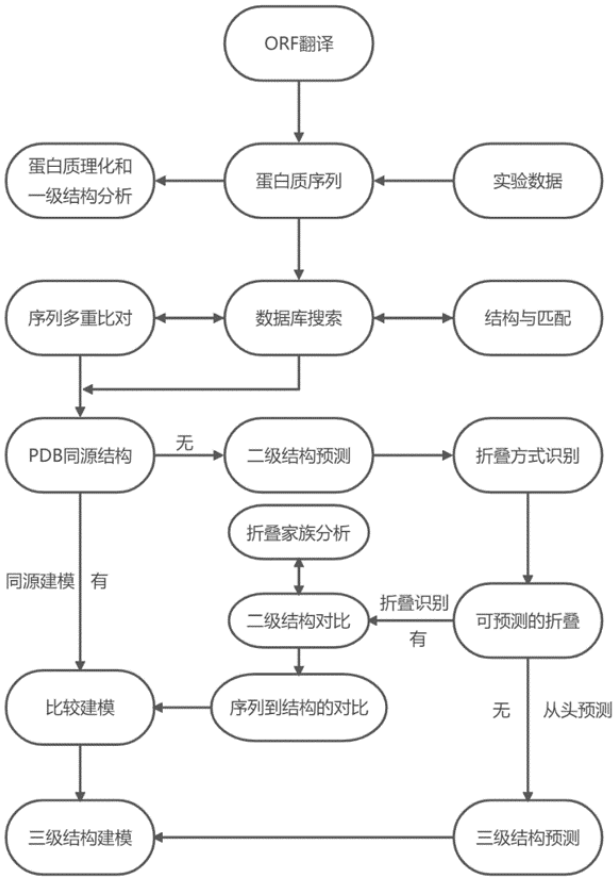


图8.11　蛋白质结构预测方法概述

目前蛋白质三级结果预测主要包括从头预测、同源建模和折叠识别三种方法

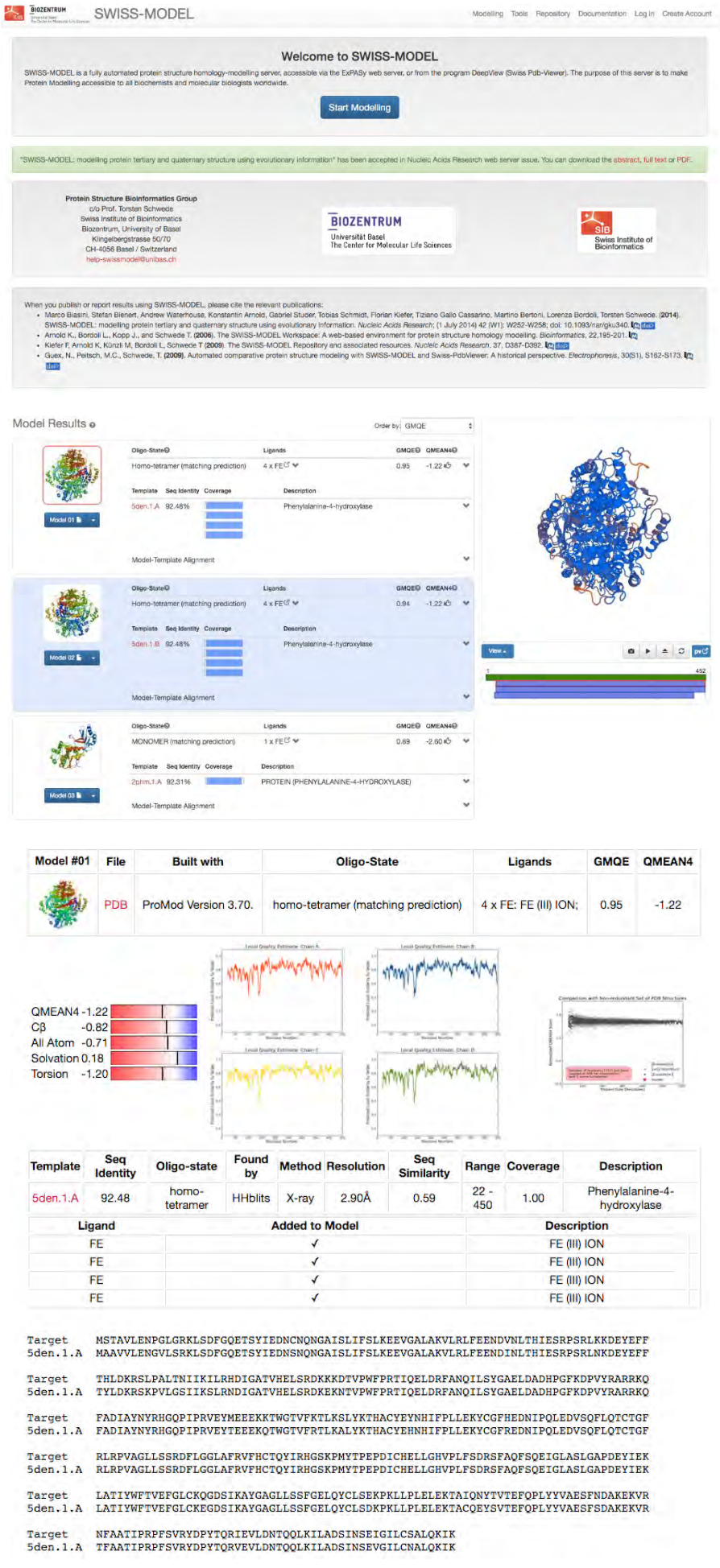


图 8.12　利用 SWISS-MODEL 算法进行蛋白质结构预测（以蛋白质序列 NP_000268 为例）

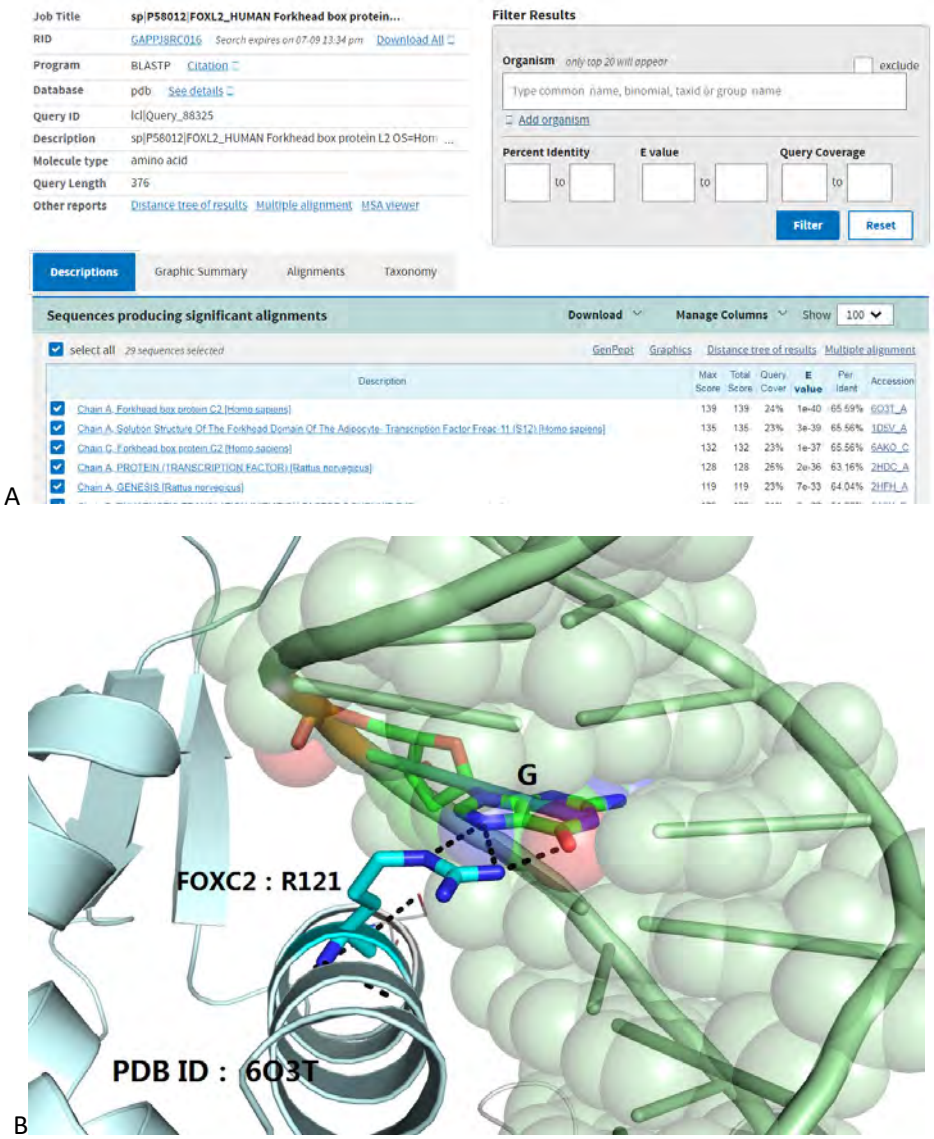
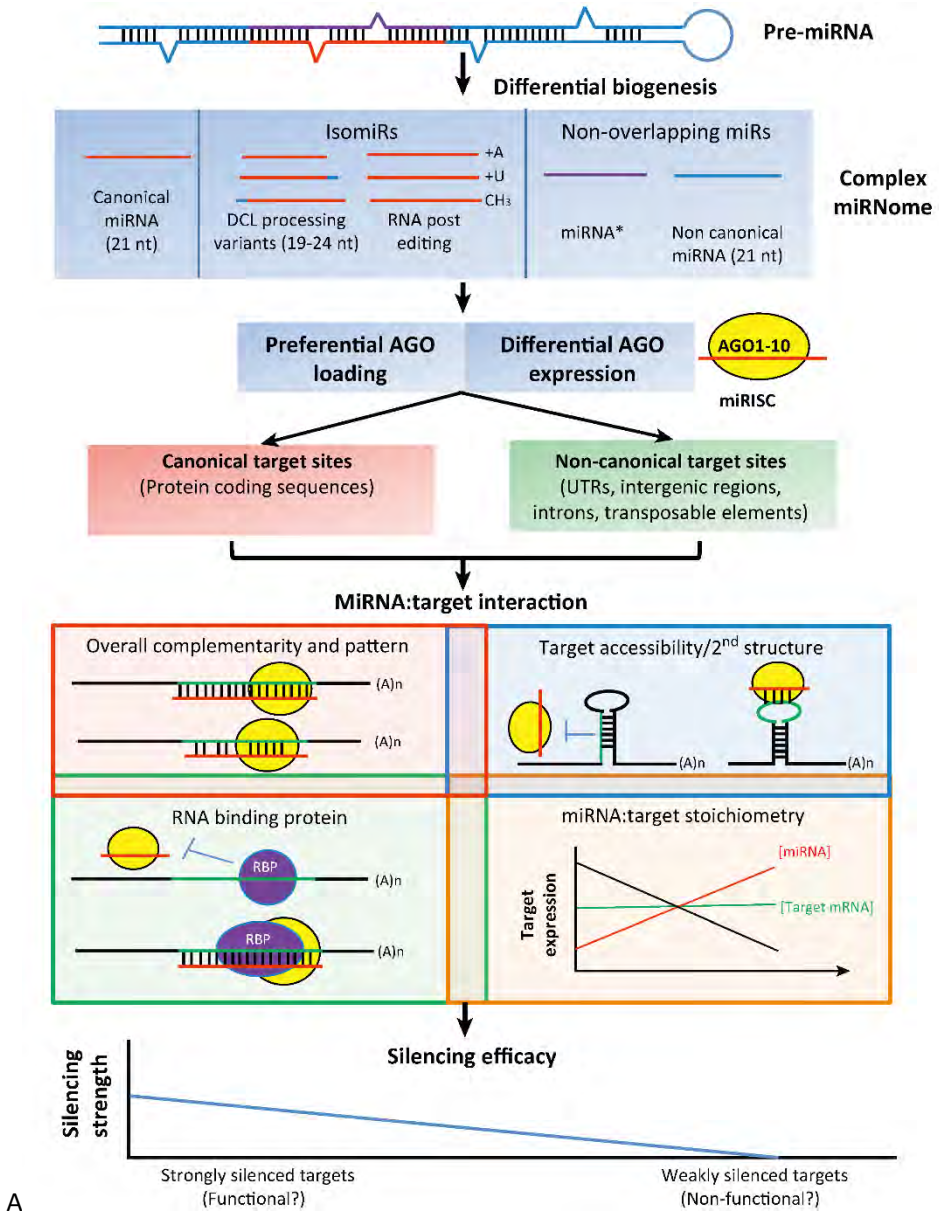


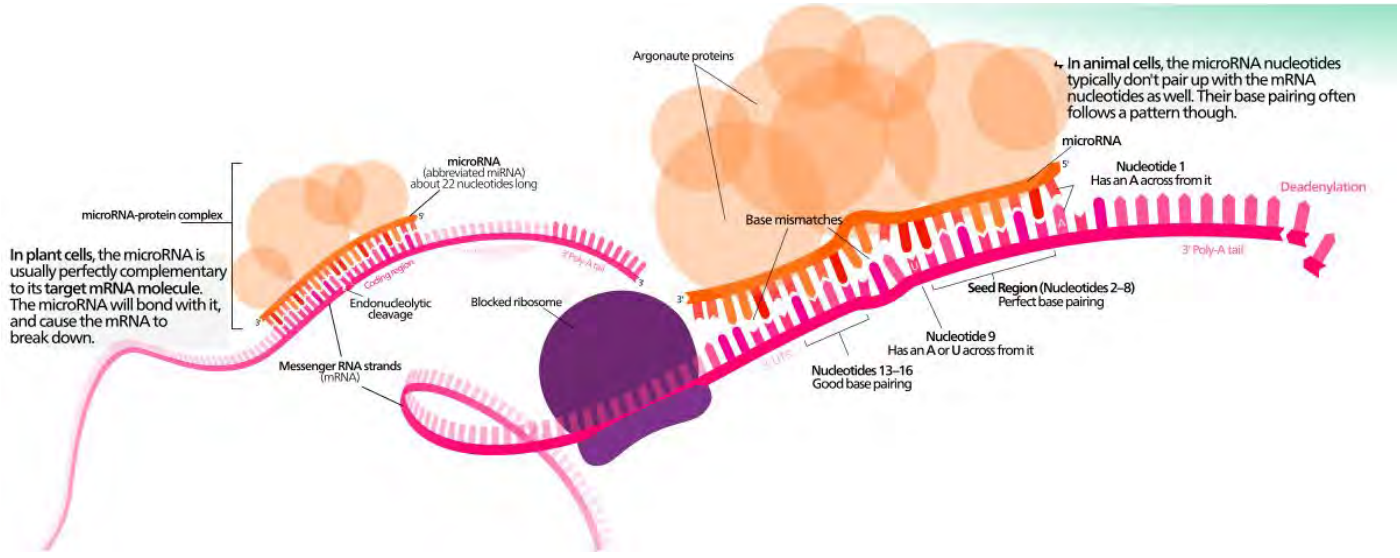
图8.13　蛋白质FOXL2（记录号P58012）R103C 突变的三维结构分析

A. FOXL2 蛋白序列在PDB 数据库中搜索获得的同源序列情况，结果显示其最佳同源蛋白为FOXC2（记录号6O3T）；

B. FOXC2 的三维结构及其与 FOXL2 的 R103C 突变对应位点（R121）与底物 DNA 结合情况（利用 PyMOL 可视化工具）



A



B

图9.1　植物miRNA调控机制及其与动物的差异

A．植物 miRNA 的产生及其调控机制（引自 Li et al.，2014）；B．动植物 miRNA 靶向调控机制比较

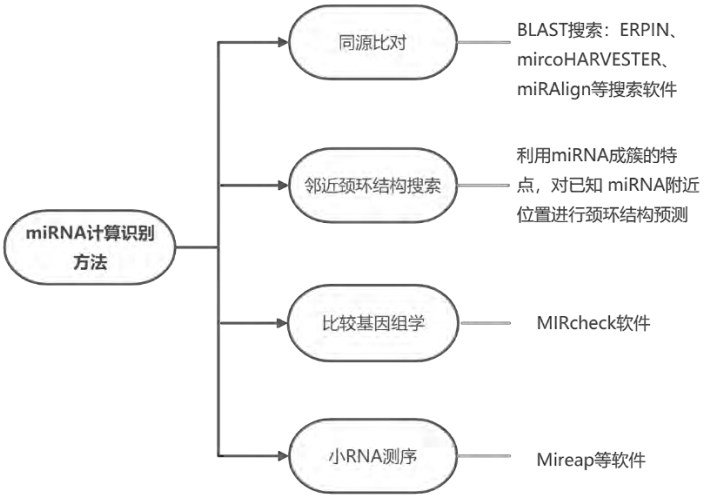


图 9.2　miRNA 计算识别方法

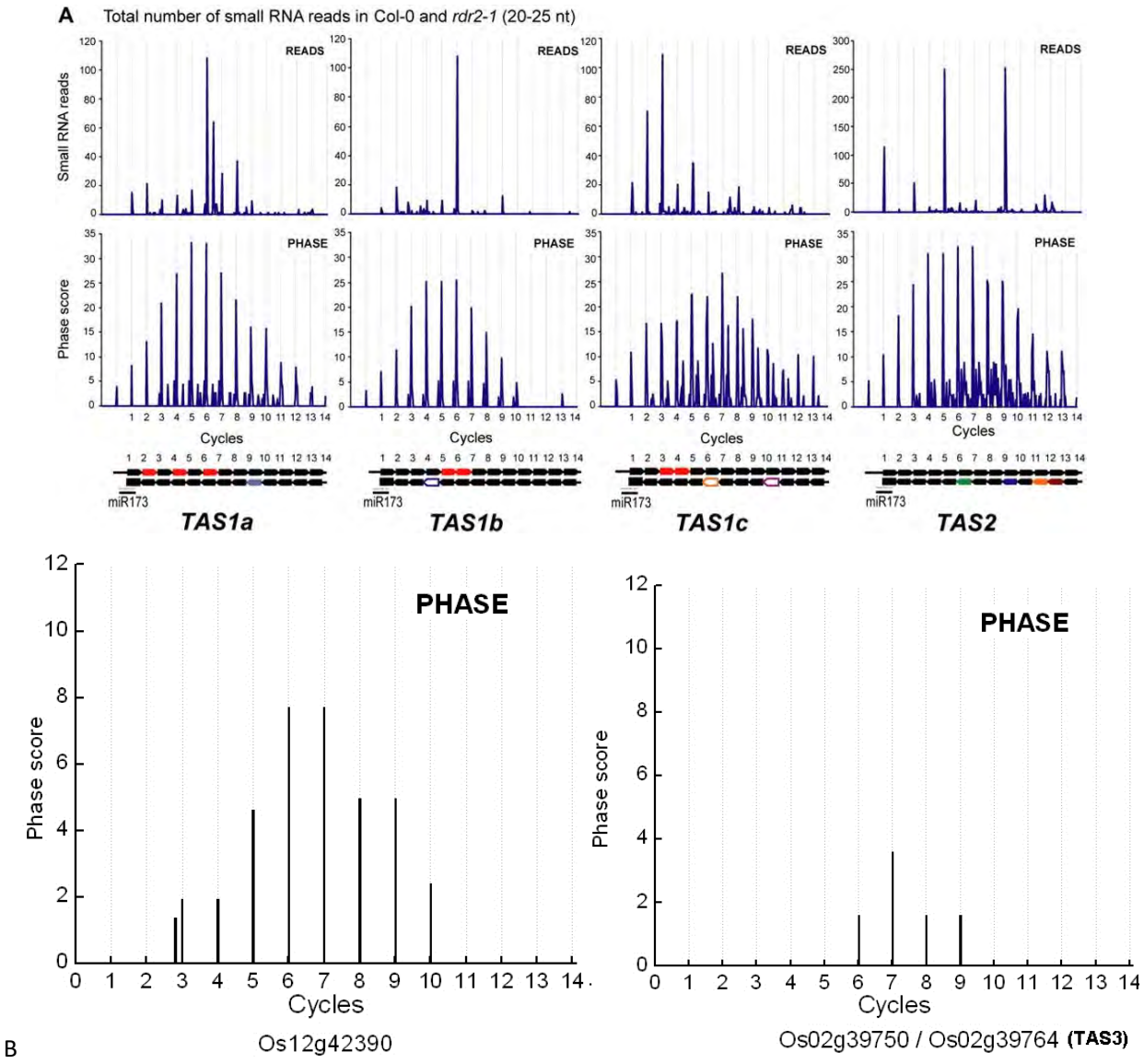


图9.3　植物phasiRNA基因位点鉴定举例

A. 4 个拟南芥 phasiRNA 基因（*TAS*）位点 21nt miRNA 读序分布及相位信号（引自 Howell et al.，2007）；B.水稻 *TAS* 基因 21nt miRNA 读序的相位值分布（引自 Shen et al.，2009）

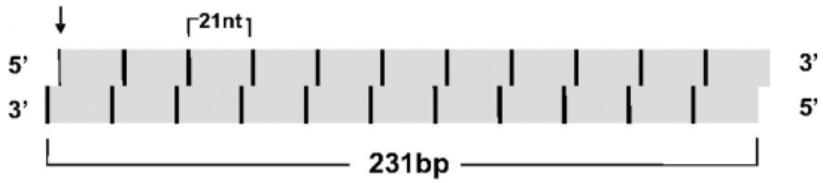
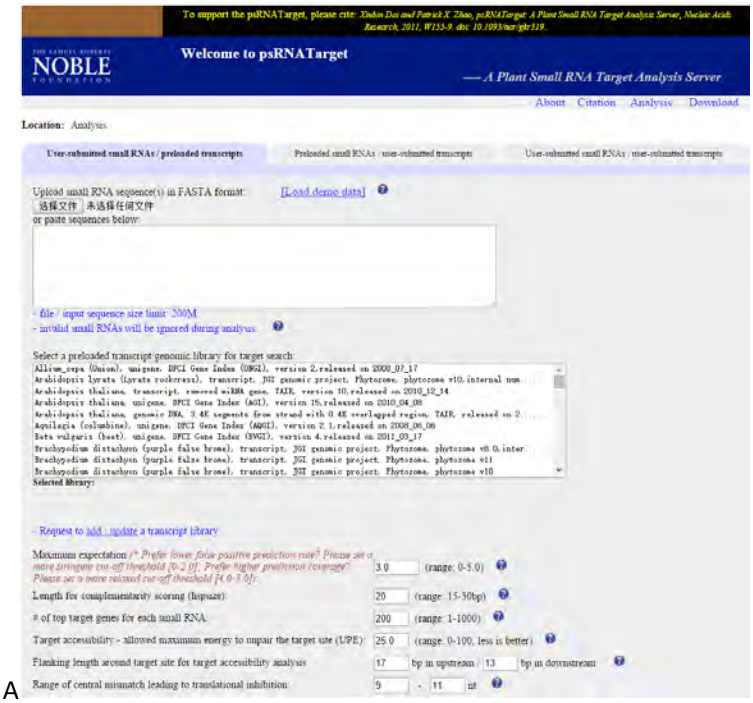
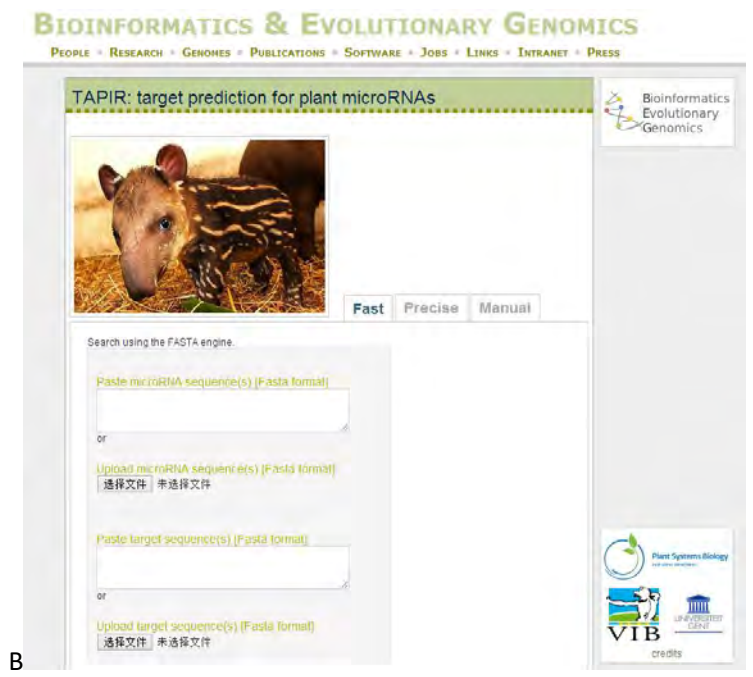


图 9.4　相位基因 *TAS* 基因的相位siRNA分布示意图（引自Chen et al.，2007）

箭头表示 phasiRNA 的起始位点，竖线表示与起始位置相距 21nt 的 siRNA 的相对位置



A



B

图 9.5　植物 miRNA 靶基因预测工具 psRNATarget（A）和 TAPIR（B）在线平台界面

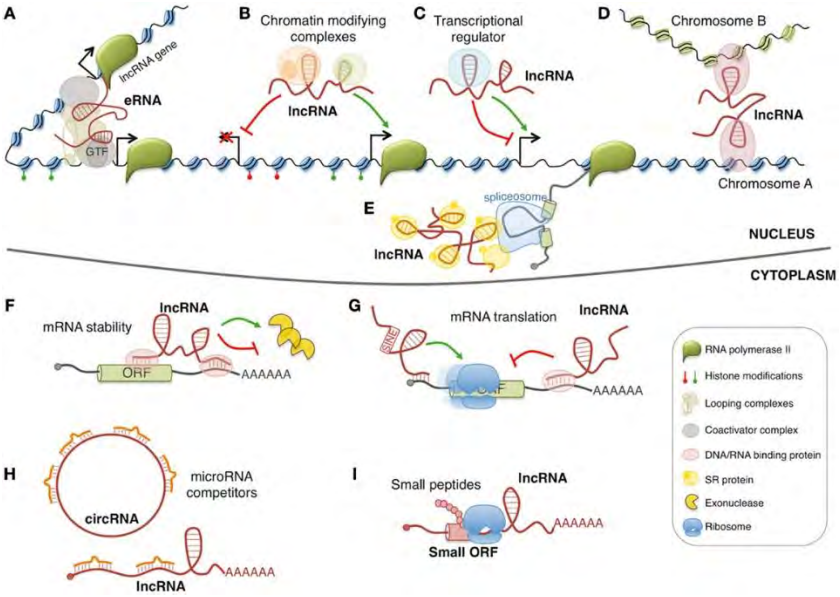


图9.6　lncRNA的已知功能（引自Morlando et al.，2015）

A．lncRNA 作为增强子来调节 mRNA 的转录；B．lncRNA 招募染色质修饰复合体来调节转录；C．lncRNA 通过调节转录因子的活性来调节转录；D．lncRNA 通过改变染色体的空间结构来调节基因的表达；E．lncRNA 通过影响 mRNA 前体的剪接来影响基因的表达；F．lncRNA 通过调节 mRNA 的稳定性来调节 mRNA 的表达；G．lncRNA 通过调节 mRNA 的翻译来调节 mRNA 的表达；H．lncRNA 通过竞争性结合 miRNA 来调节 mRNA 的表达；I．一部分含有开放阅读框的 lncRNA 可以被翻译形成小肽

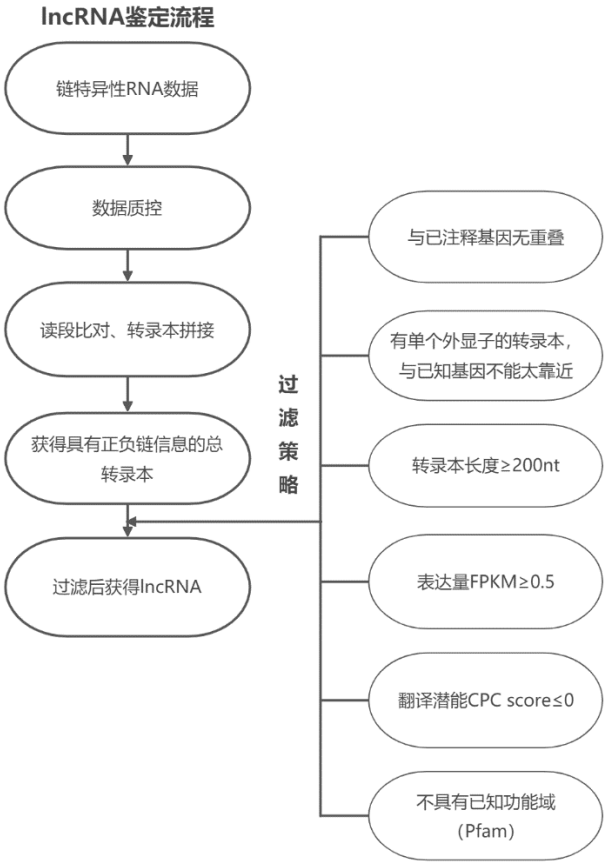


图 9.7　lncRNA 鉴定流程

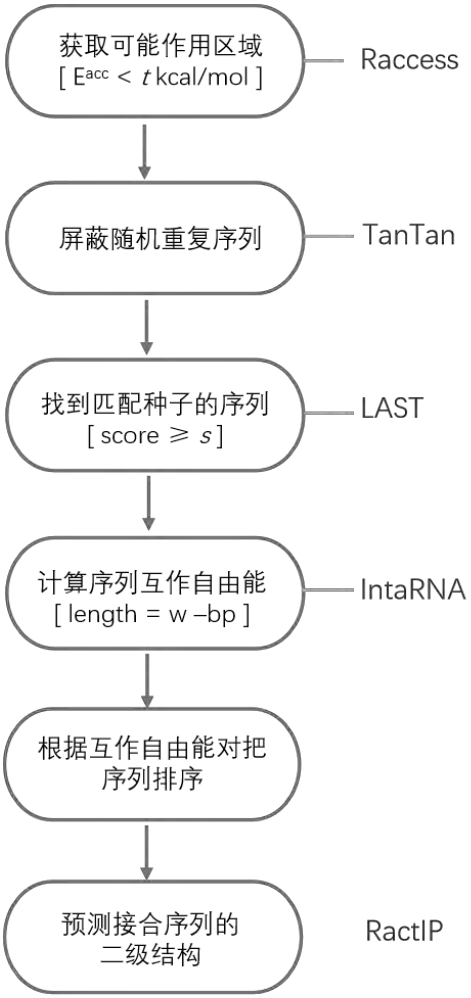


图 9.8　预测 lncRNA 与 RNA 互作的综合算法流程 （引自 Terai et al.，2015）

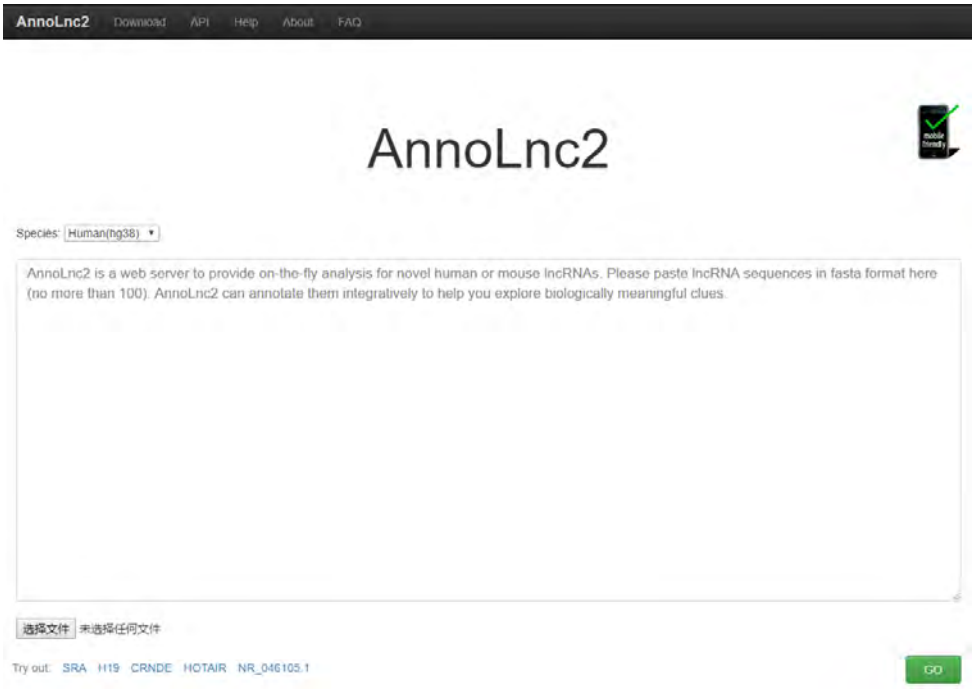


图 9.9　长非编码 RNA 在线注释工具 AnnoLnc2 主页（http://annolnc.gao-lab.org/）

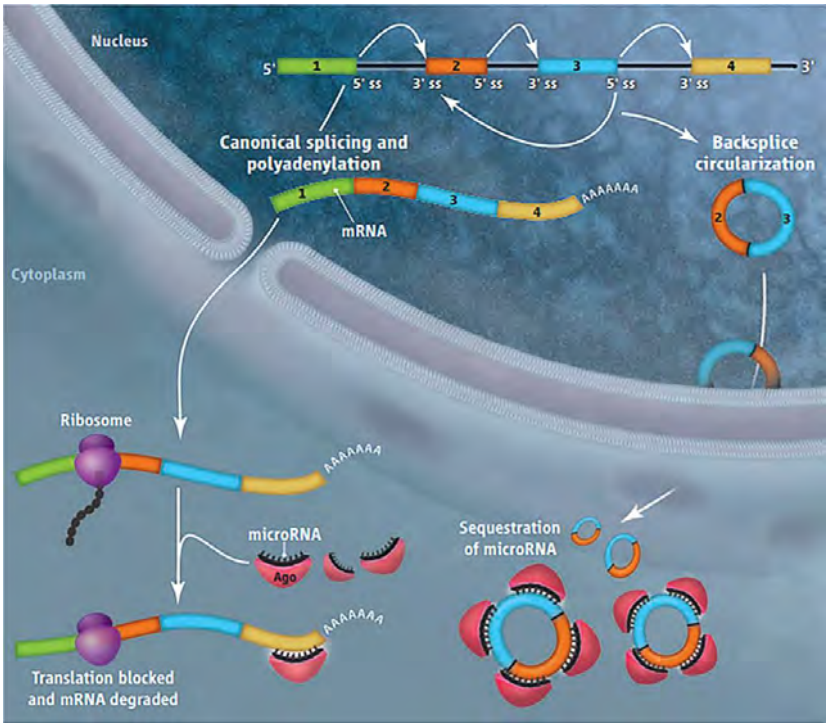
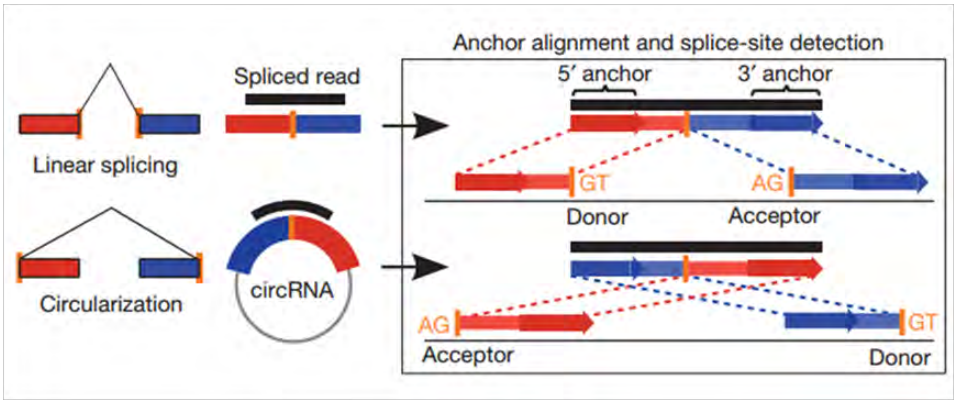


图 9.10　环状 RNA 分子的形成（引自 Bolisetty and Graveley，2013）



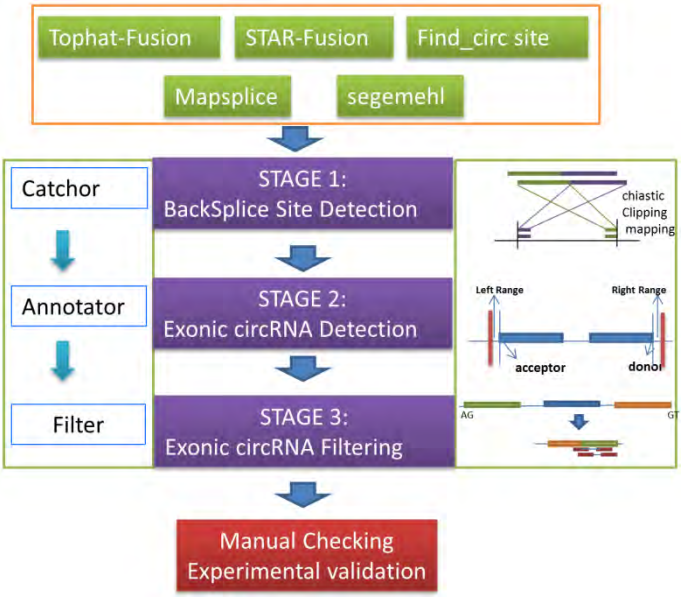


图9.11　环状RNA预测工具

A．find_circ 预测环状 RNA 算法（引自 Memczak et al.，2013）；B．Pcirc_finder 算法（引自 Chen et al.，2016a）

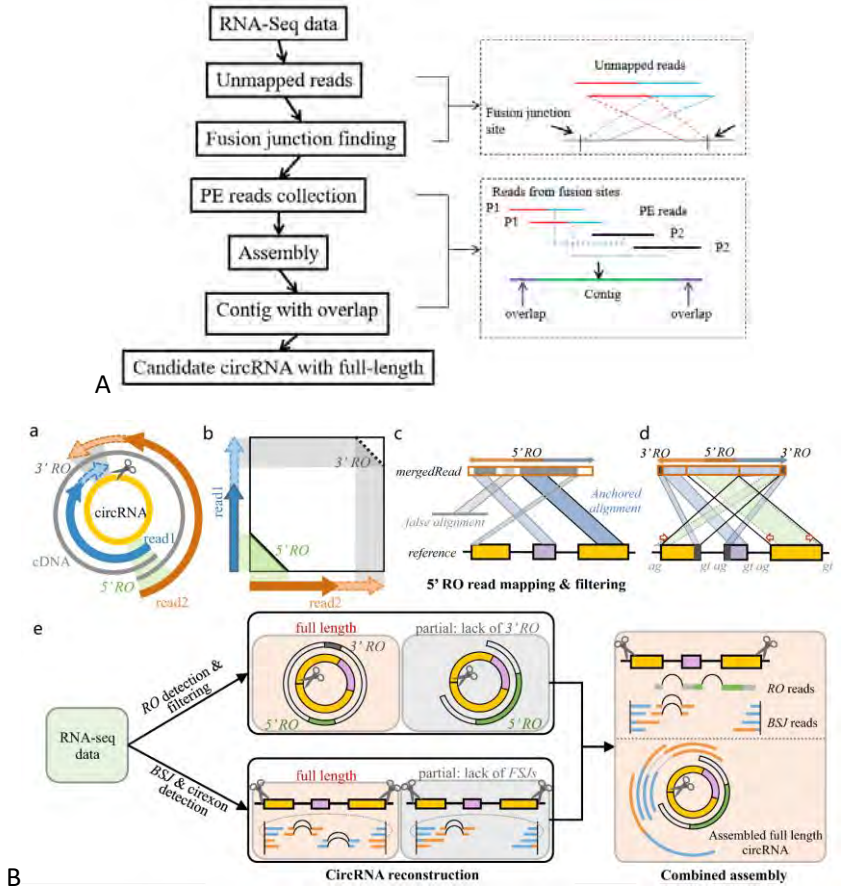


图9.12　环状RNA全长拼接算法

A．circseq_cup 流程（引自 Ye et al.，2017）。B．CIRI_full 流程（引自 Zheng et al.，2019）。a、b．读序的反向重叠（RO）；c、d．基于 5′-RO，将两条读序合并，并比对至基因组以确定其准确位置；e．环状 RNA 全长序列的拼接。当 5′-RO 和 3′-RO 均能发现时，或能被支持反向剪接位点的读序（BSJ）全部覆盖时，其全长序列可直接拼接获得；当 3′-RO 没有，或者支持反向剪接位点的读序不能全部覆盖时，将结合 5′-RO 和 BSJ 两者拼接环状 RNA 的全长序列

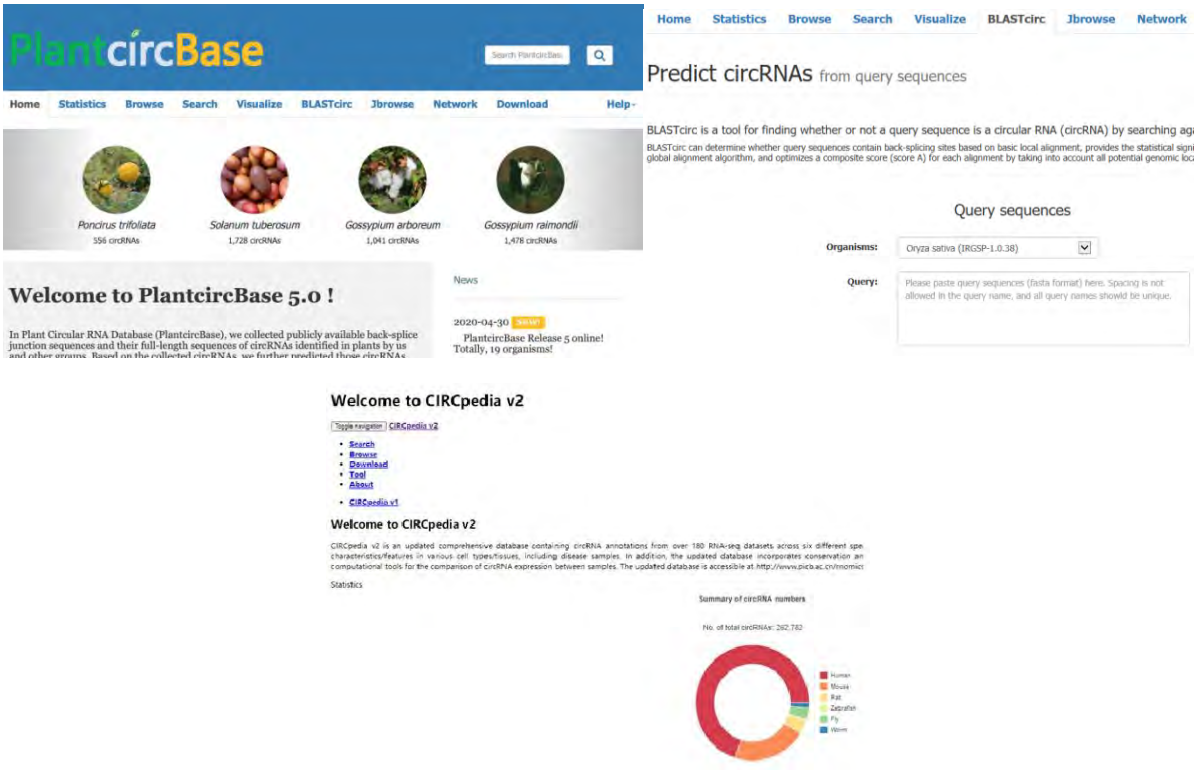


图9.13　环状RNA数据库

A．植物环状RNA数据库PlantcircBase主页及其提供的环状RNA搜索工具BLASTcirc（引自Chu et al.，2017；2018）；

B.人类及模式动物环状 RNA 数据库 CIRCpedia（引自 Zhang et al.，2016）

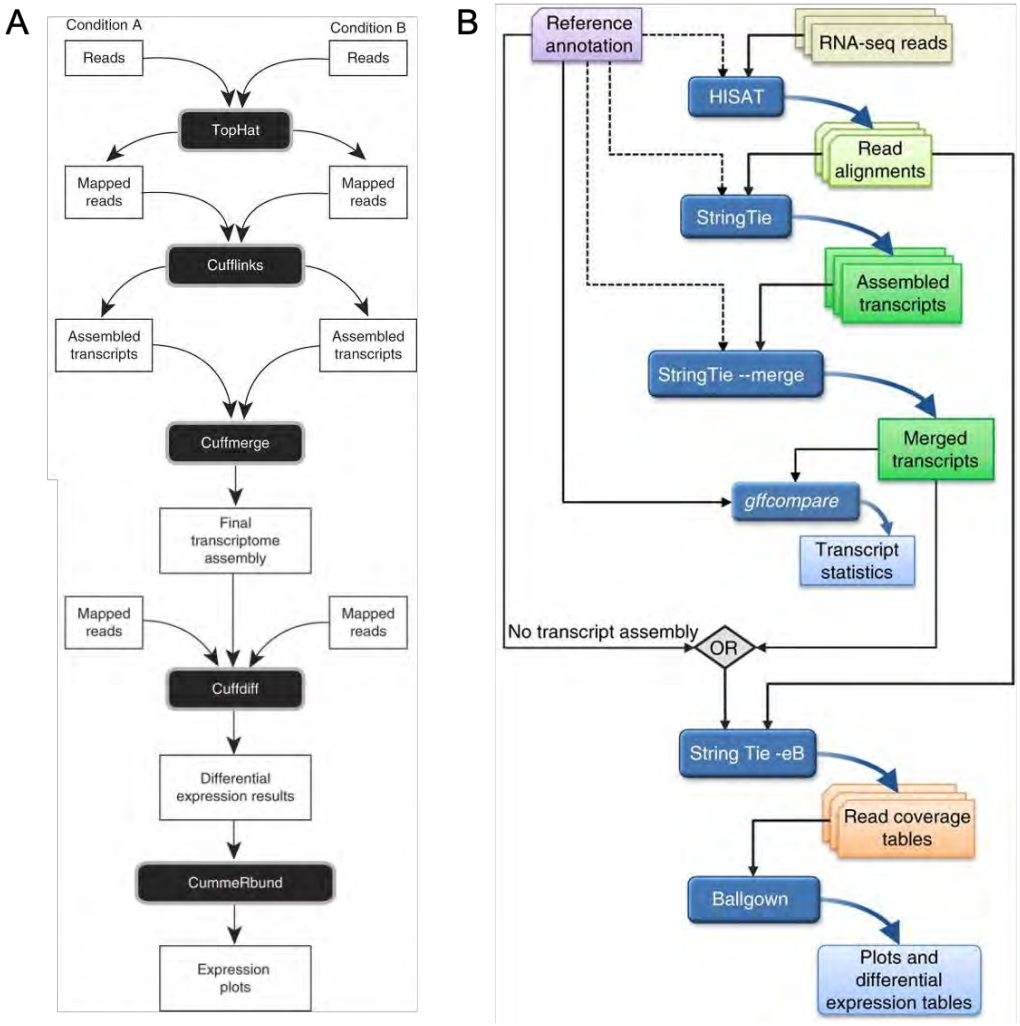


图10.1　转录组数据分析的两个经典分析流程

A. TCC 流程，即 TopHat 比对、Cufflinks 拼接和定量、Cuffdiff 差异分析、CummeRbund 可视化的转录组分析流程（引自 Trapnell et al.，2012）；B. HSB 流程，即 HISAT 比对、StringTie 拼接定量和差异分析、Ballgown 可视化的转录组分析流程（引自 Pertea et al.，2016）

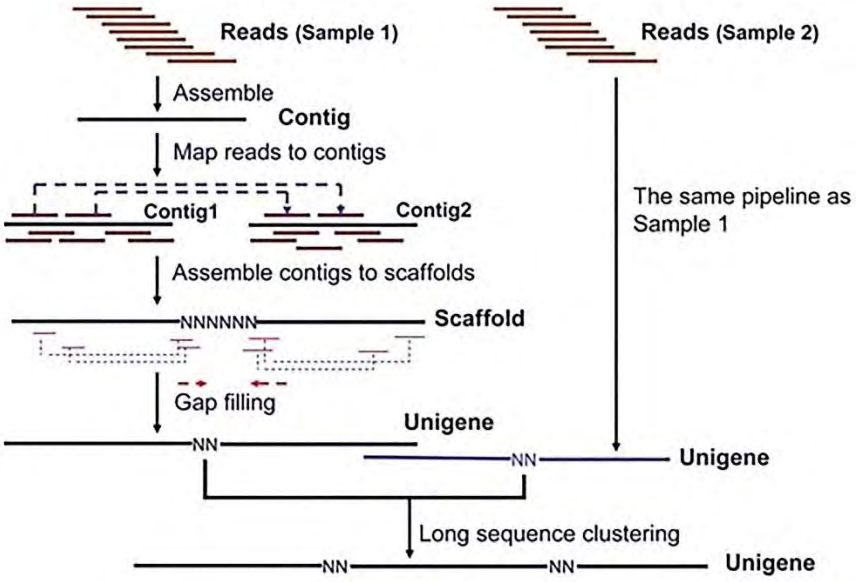


图 10.2　RNA-Seq 序列拼接流程

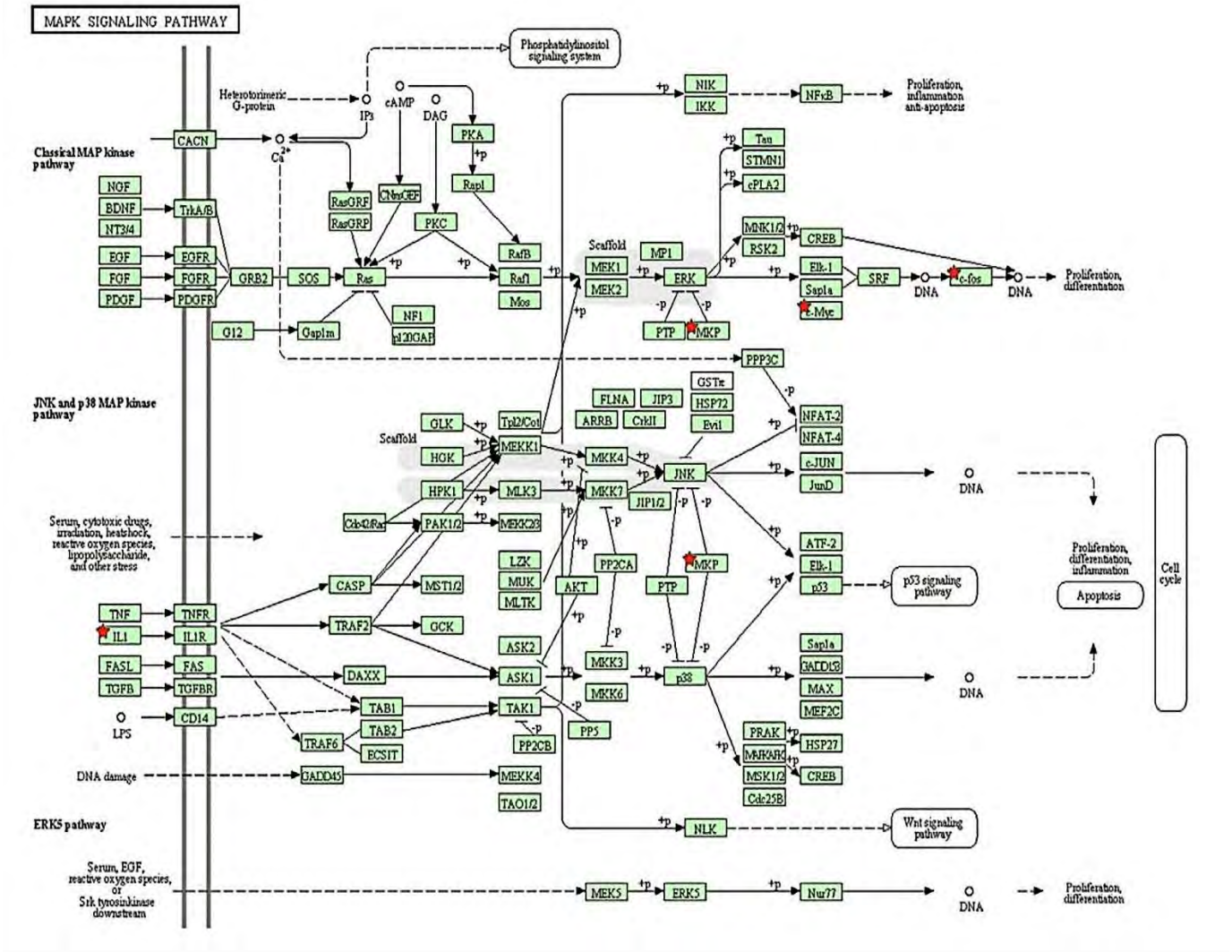


图 10.3　烟气处理与对照（空气处理）之间肺支气管细胞差异表达基因的 KEGG 富集通路（引自 Shen et al.，2016）

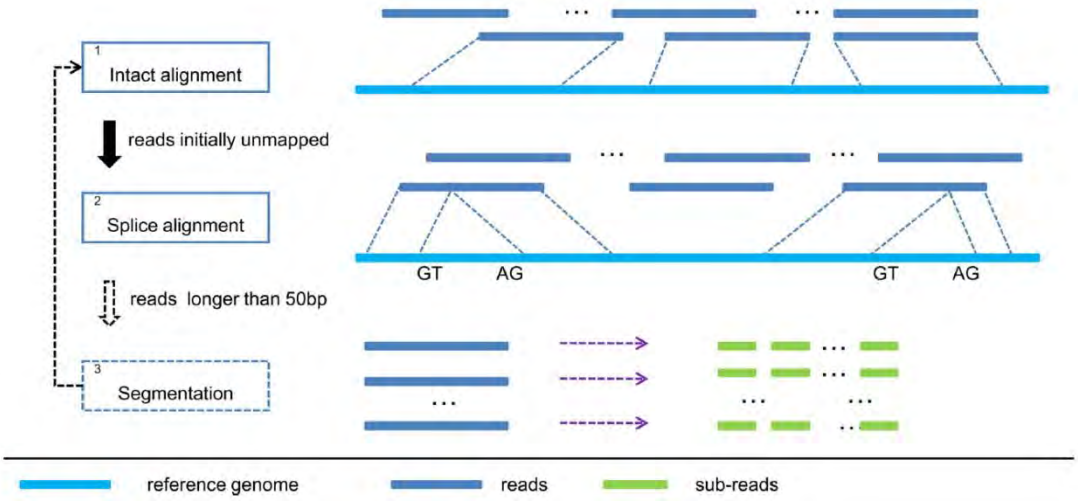


图 10.4　SOAPsplice 软件鉴定可变剪接的原理（引自 Huang et al.，2011）

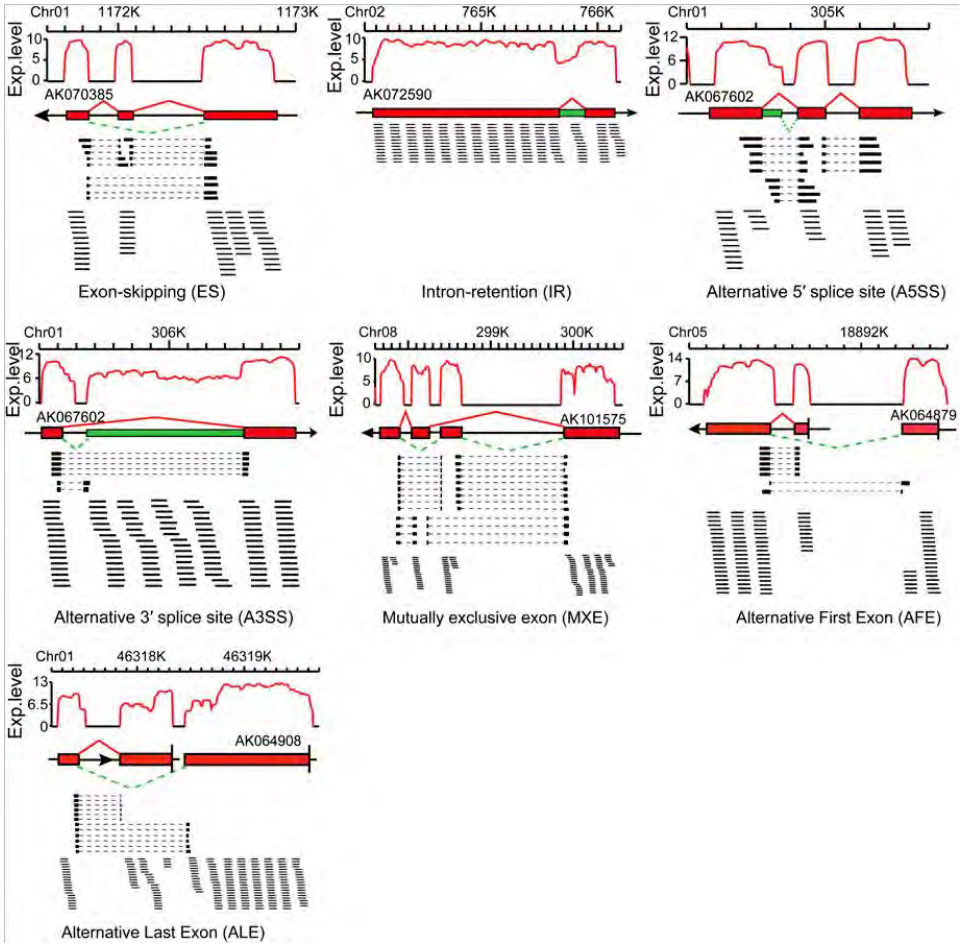


图10.5　基于转录组数据发现水稻中存在至少7 种不同的基因可变剪接形式（引自Zhang et al.，2010）

曲线表示表达量值；方形图表示外显子；虚线表示外显子连接的方式；比对上的读序用黑色的短线表示。7 种可变剪接方式为外显子跳跃（ES）、内含子保留（IR）、5′ 端可变剪接（A5SS）、3′ 端可变剪接（A3SS）、外显子排除（MEE）、第一外显子可变剪接（AFE）和最末外显子可变剪接（ALE）

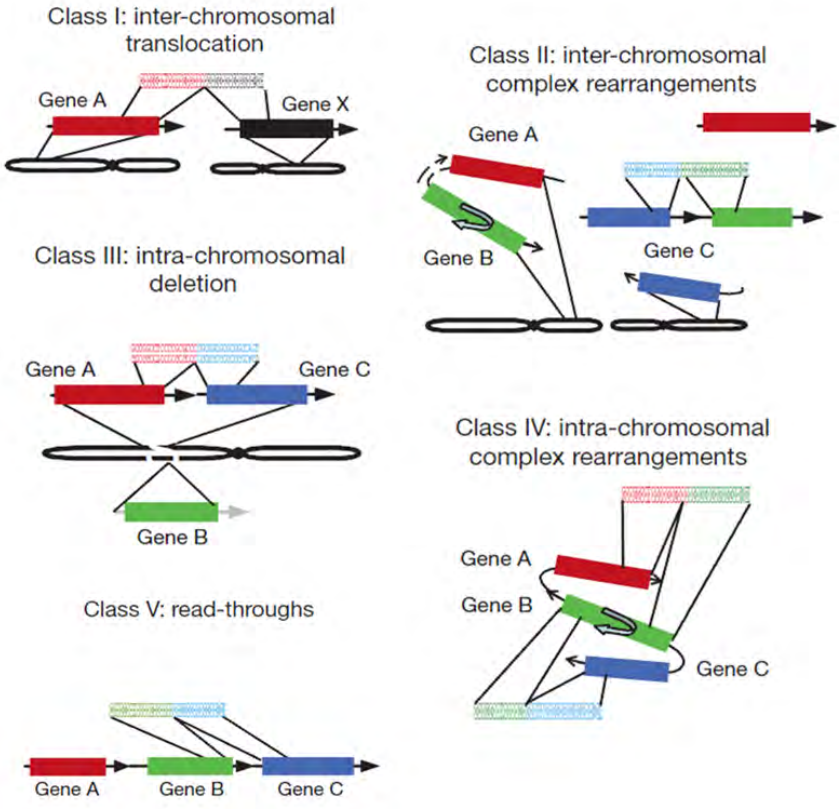


图 10.6　5 种不同类型（Ⅰ～Ⅴ）融合基因（引自 Maher et al.，2009）

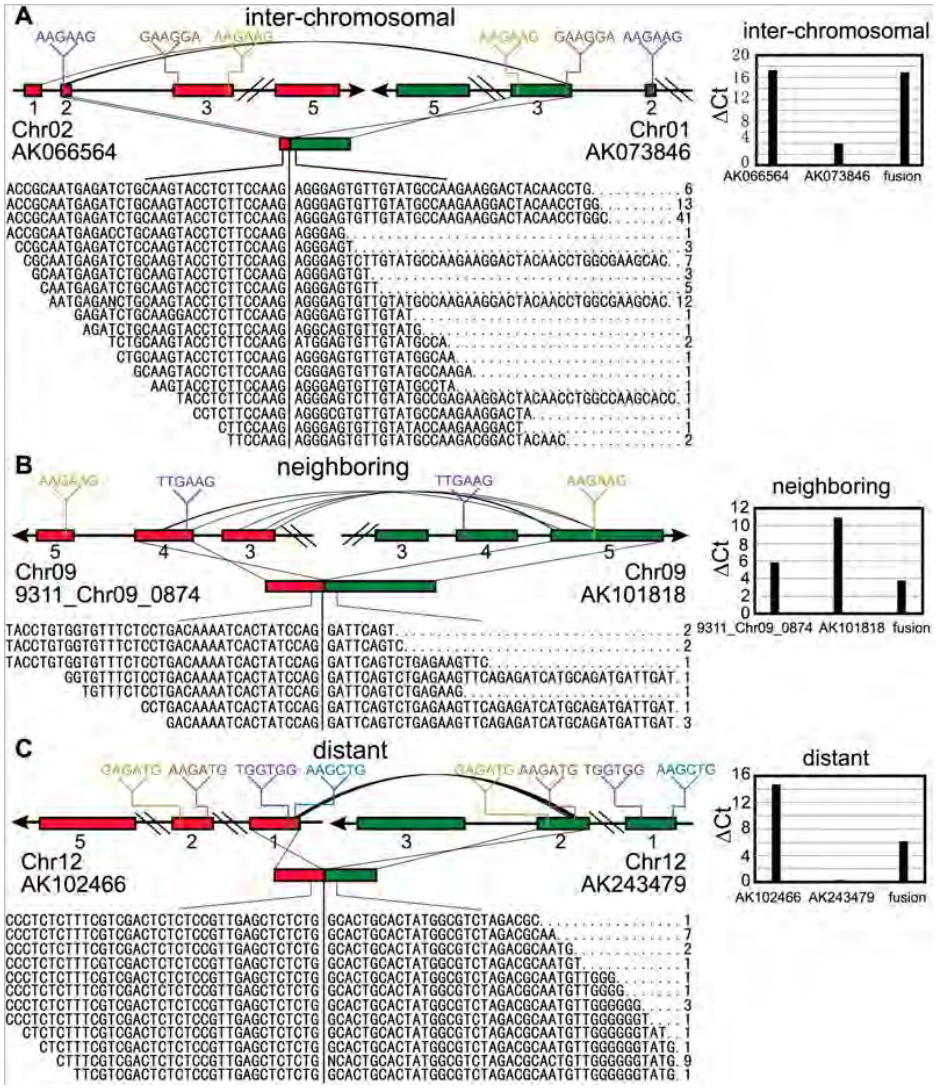


图10.7　水稻中鉴定发现的基因融合现象（引自Zhang et al.，2010）

通过黑线连接的红色和绿色柱形图表示不同外显子区域，右侧的柱状图表示两个基因及其融合基因表达量（定量PCR 结果）， 图中显示的是鉴定到的三种不同水稻基因融合方式：A．两条染色体间的基因融合现象；B．相邻基因间的基因融合现象；C．在同一条染色体上远距离基因间的融合现象

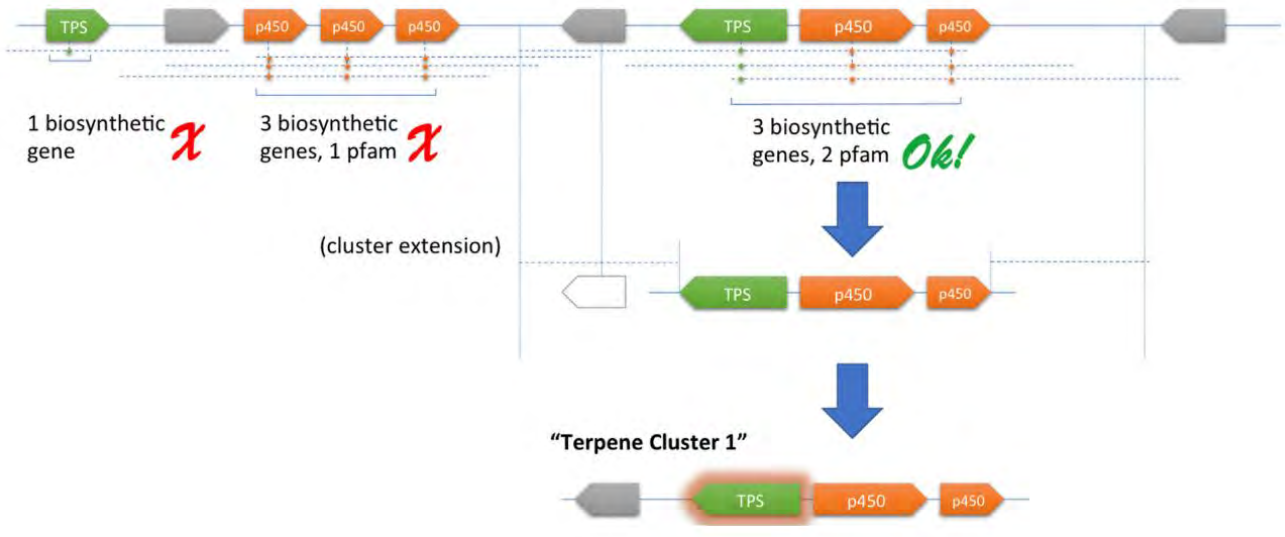


图10.8　植物次生代谢合成基因簇鉴定算法plantiSMASH的原理（引自Kautsar et al.，2017）

TPS．萜烯合成酶基因；p450．细胞色素 P450 基因

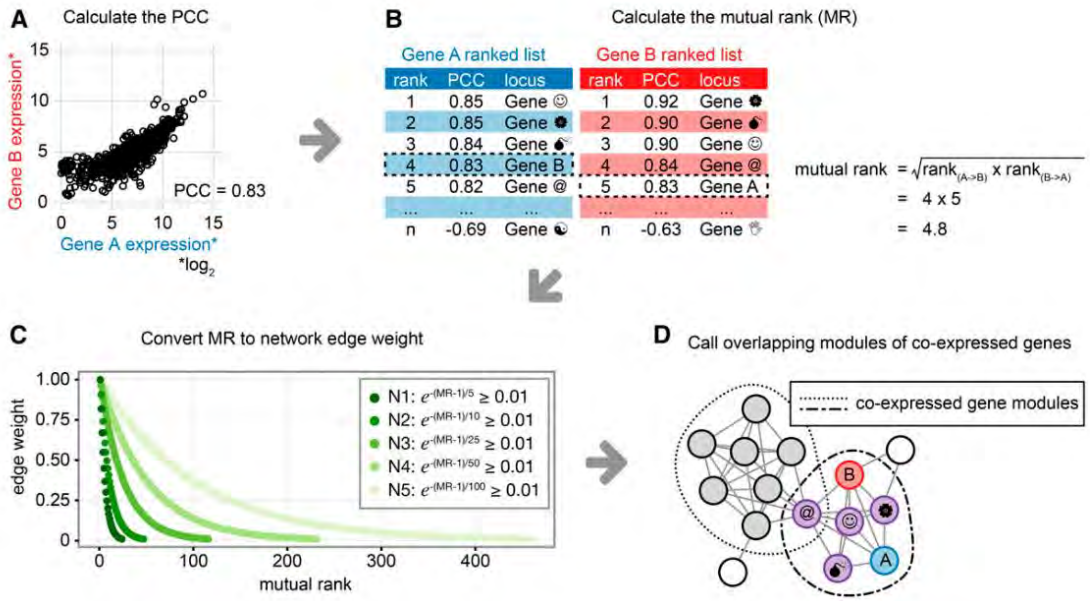


图10.9　植物次生代谢物合成途径共表达网络分析流程（引自Wisecaver et al.，2017）

A.　计算基因组中每个基因对表达的皮尔森相关系数（如基因 A 和基因 B 之间的相关性为 0.83）；B.对相关性进行排序，并计算每个基因对的互排位（mutual rank，MR）值（如基因 A 和 B 的 MR 值为 4.8）；C.使用一个或多个指数衰减函数将 MR 值转换为网络边权值（network edge weight），这里评估了 5 种不同的衰变率，得到了 5 种不同的网络（N1～N5）；D.使用 ClusterONE 可视化基因共表达模块交集，在本例中， 基因 A 和基因 B 及 4 个其他基因（紫色圆圈）组成了一个模块，基因可以属于单个或多个（如 Gene@）共表达模块

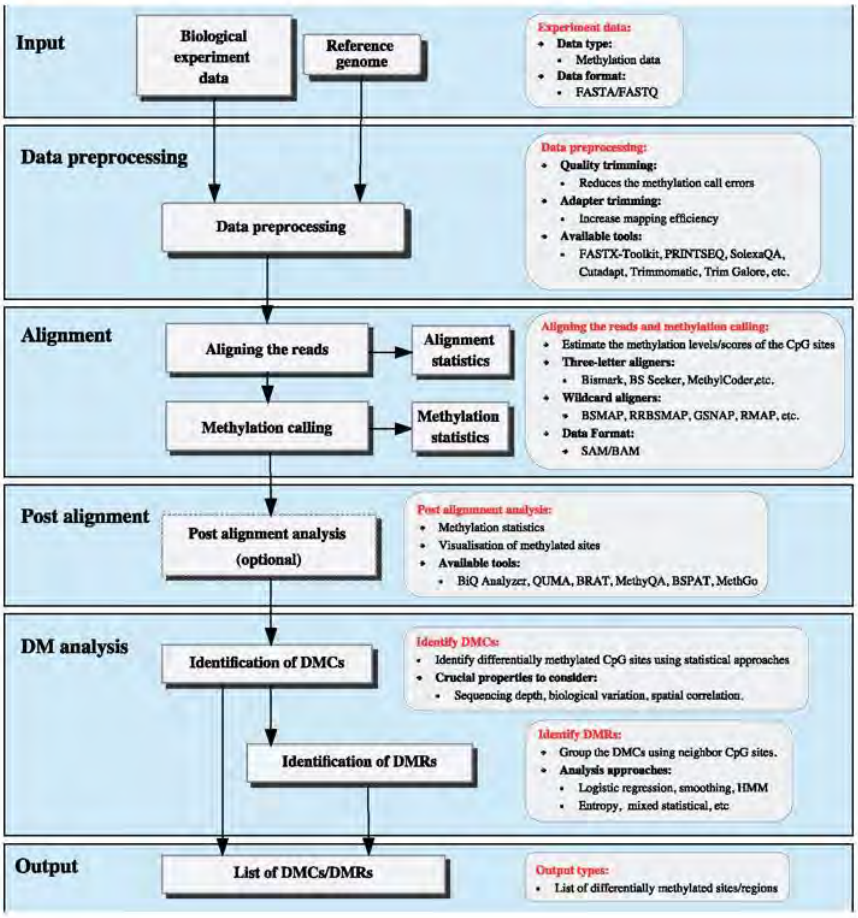


图10.10　DNA甲基化数据处理和分析流程图（引自Shafi et al.，2018）

DMC．甲基化差异位点；DMR．甲基化差异区域

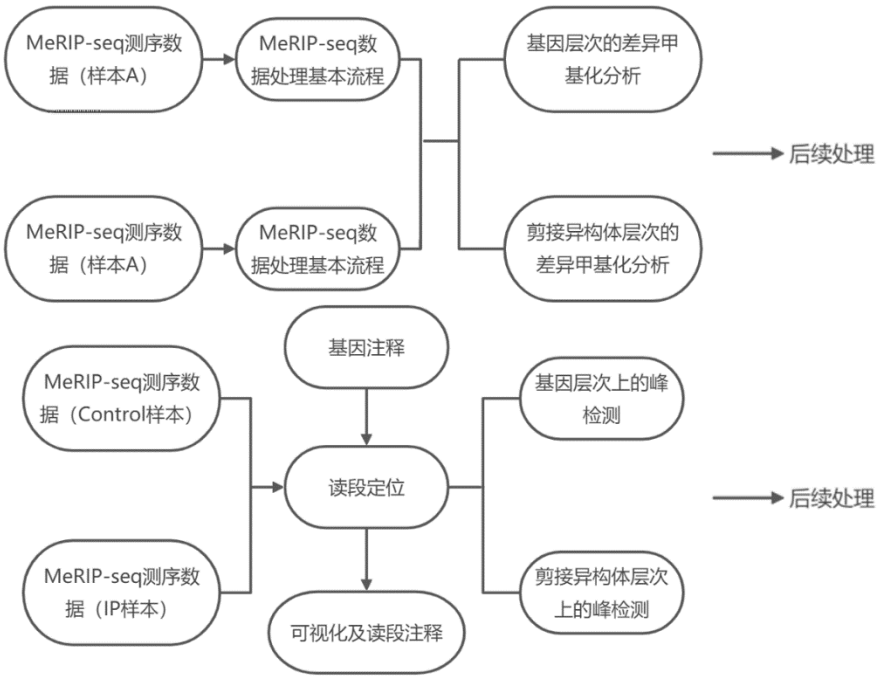


图10.11　RNA甲基化MeRIP-Seq数据分析流程（引自Liu et al.，2015）

A.单样本 MeRIP-Seq 数据分析；B.双样本 MeRIP-Seq 数据分析

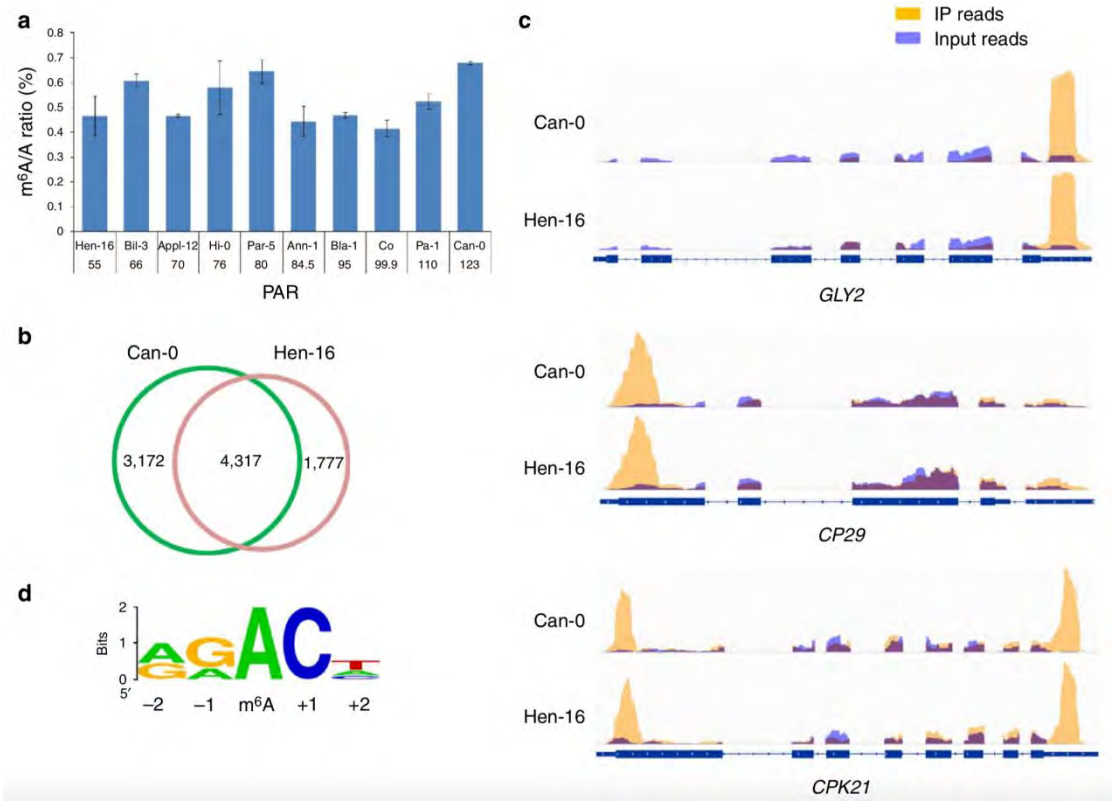


图10.12　拟南芥中的m6A 修饰（引自Luo et al.，2014）

A. 不同拟南芥材料中 m6A 修饰的比例；B. 拟南芥‘Can-0’和‘Hen-16’材料中 m6A 修饰位点的数量，其中有 4317 个为两者共有；C. 材料‘Can-0’和‘Hen-16’在若干基因中保守的甲基化位点（m6A 峰）举例，其中黄色的读序来自 MeRIP-Seq，蓝色等其他颜色来自常规 RNA-Seq；D. 修饰位点（即 m6A 峰区域）序列呈现“RRACH”保守性

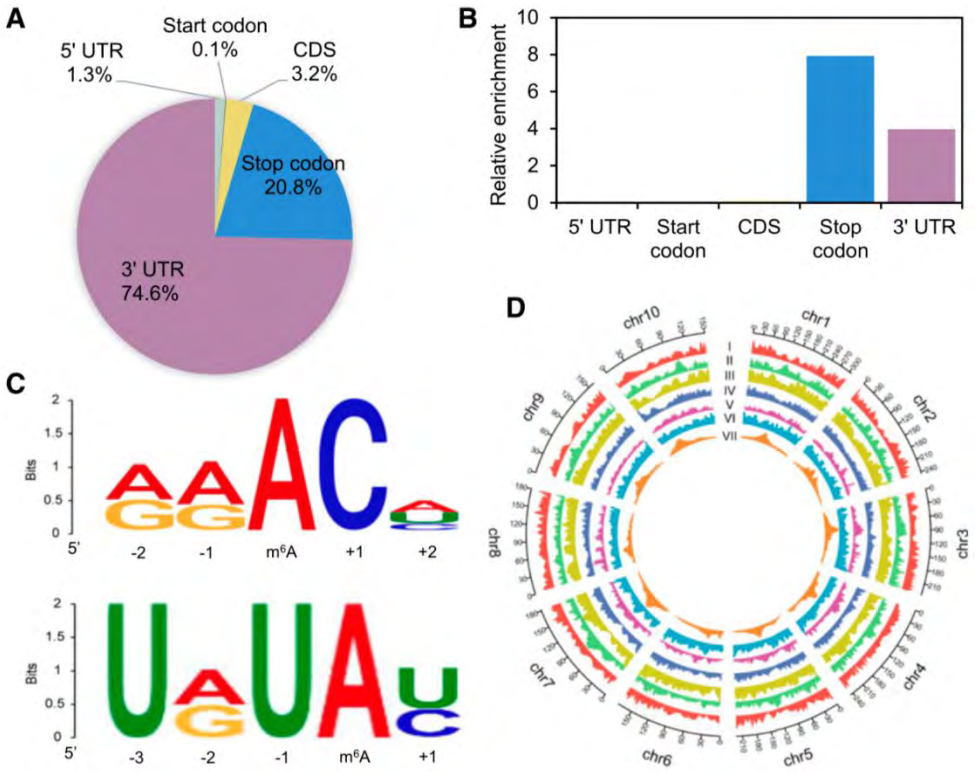


图10.13　玉米中m6A 修饰情况（引自Miao et al.，2020）

A. m6A 修饰峰在全转录组上的分布，CDS．蛋白质编码区，UTR．非翻译区；B.m6A 修饰峰在转录本不同片段上的密度 分布；C. 大部分 m6A 修饰峰所具有的经典保守序列“RRACH”（上）和“URUAY”保守区域（下，该保守片段是植物上特有的一段保守序列并且可以被 m6A 阅读蛋白 ECT2 识别）；D.m6A 修饰基因在玉米基因组 10 条染色体上的分布，从外向内每圈分别代表 m6A 修饰基因的出现频率（Ⅰ）、基因平均长度（Ⅱ）、平均 GC 含量（Ⅲ）、平均外显子长度（Ⅳ）、平均内含子长度（Ⅴ）、平均外显子数量（Ⅵ）、到邻近基因的平均距离（Ⅶ）

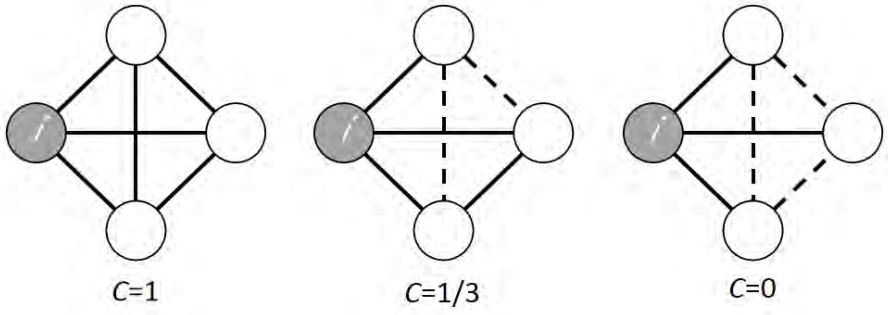


图 10.14　具有不同聚合系数的网络图举例

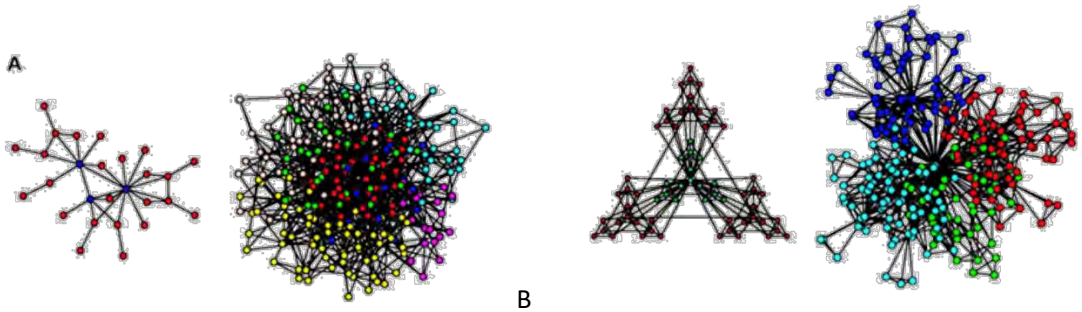


图10.15　复杂网络模型（引自E. Ravasz et al.，2002）

无标度网络（A）和阶层网络（B）

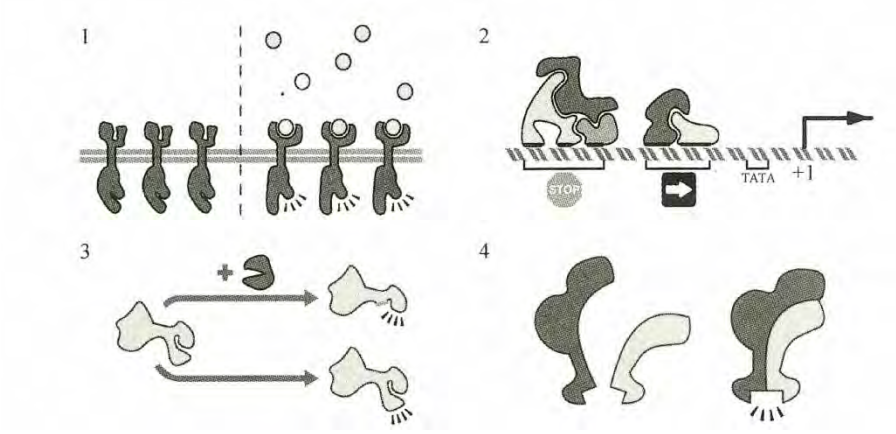
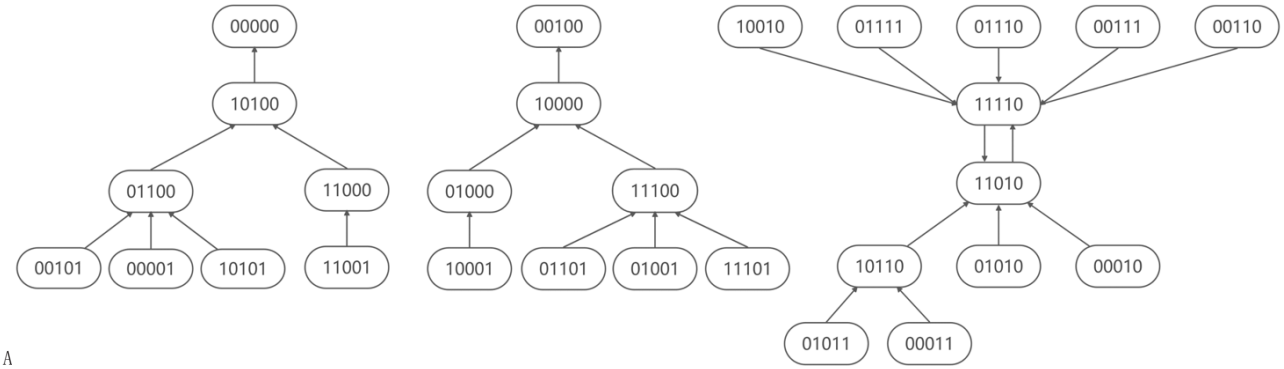


图10.16　细胞内几种相互作用方式（引自Fernandez and Sole，2006）

A．信号传导；B．基因的转录；C．剪接修饰；D．蛋白质相互作用形成复合体

A



B

$x_1x_2x_3$	$f_1^{(1)}$	$f_2^{(1)}$	$f_1^{(2)}$	$f_1^{(3)}$	$f_2^{(3)}$
000	0	0	0	0	0
001	1	1	1	0	0
010	1	1	1	0	0
011	1	0	0	1	0
100	0	0	1	0	0
101	1	1	1	1	0
110	1	1	0	1	0
111	1	1	1	1	1
$c_j^{(i)}$	0.6	0.4	1	0.5	0.5

图10.17　基因调控网络布尔模型举例（引自Shmulevich and Dougherty，2007）

A.　一个包含 5 个基因的布尔网络状态转化图；B．一个包含 3 个基因的概率布尔网络（PBN）的函数真值表

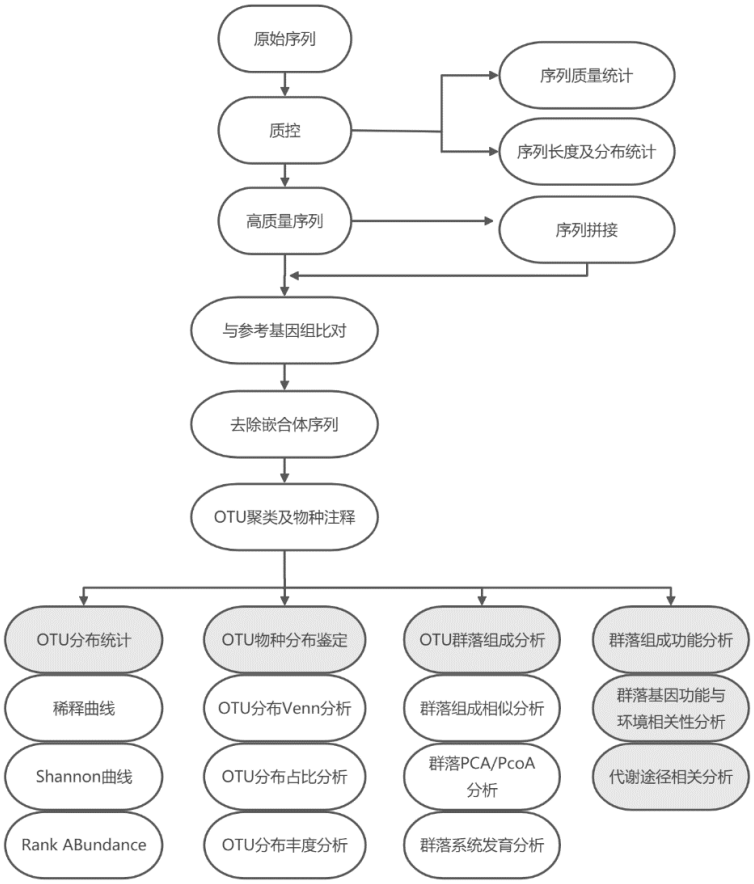


图 11.1　16S rDNA 生物信息学分析流程

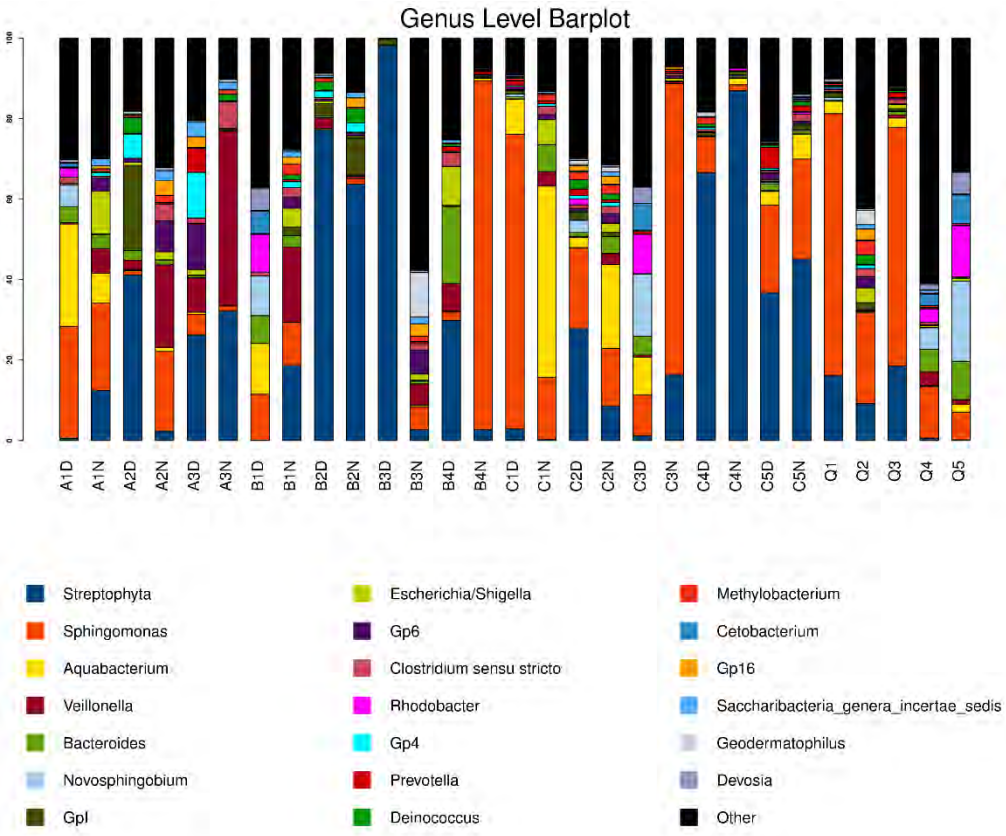


图 11.2　多样本物种多样性及其在属分类水平上的占比分析

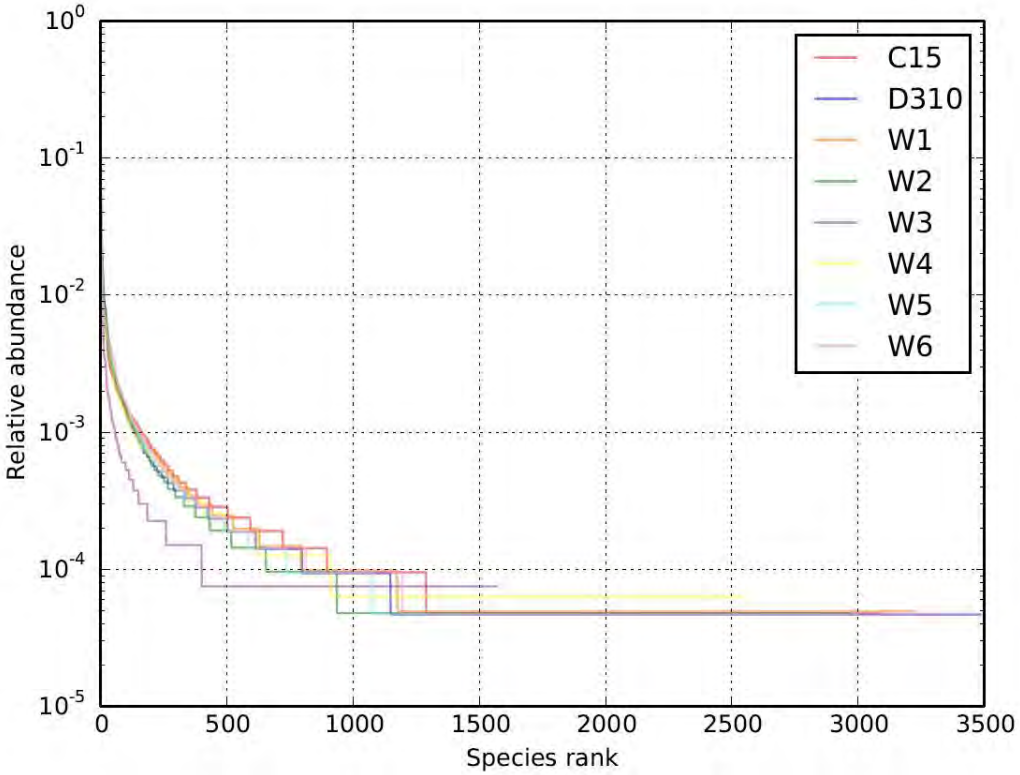


图 11.3　8 个环境样本 16S rDNA 排名丰度曲线图

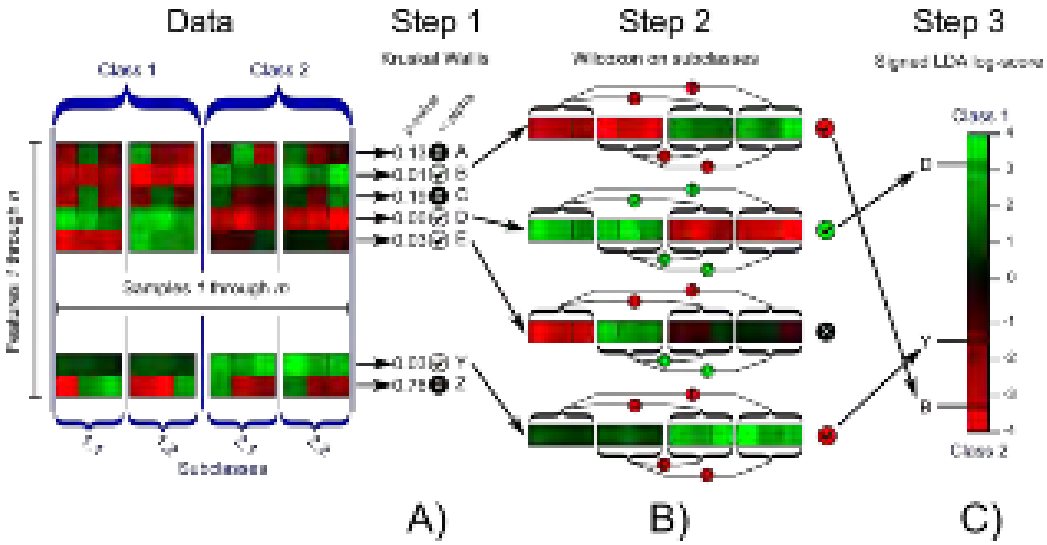


图 11.4　LEfse 差异分析流程（引自 Segata et al.， 2011）

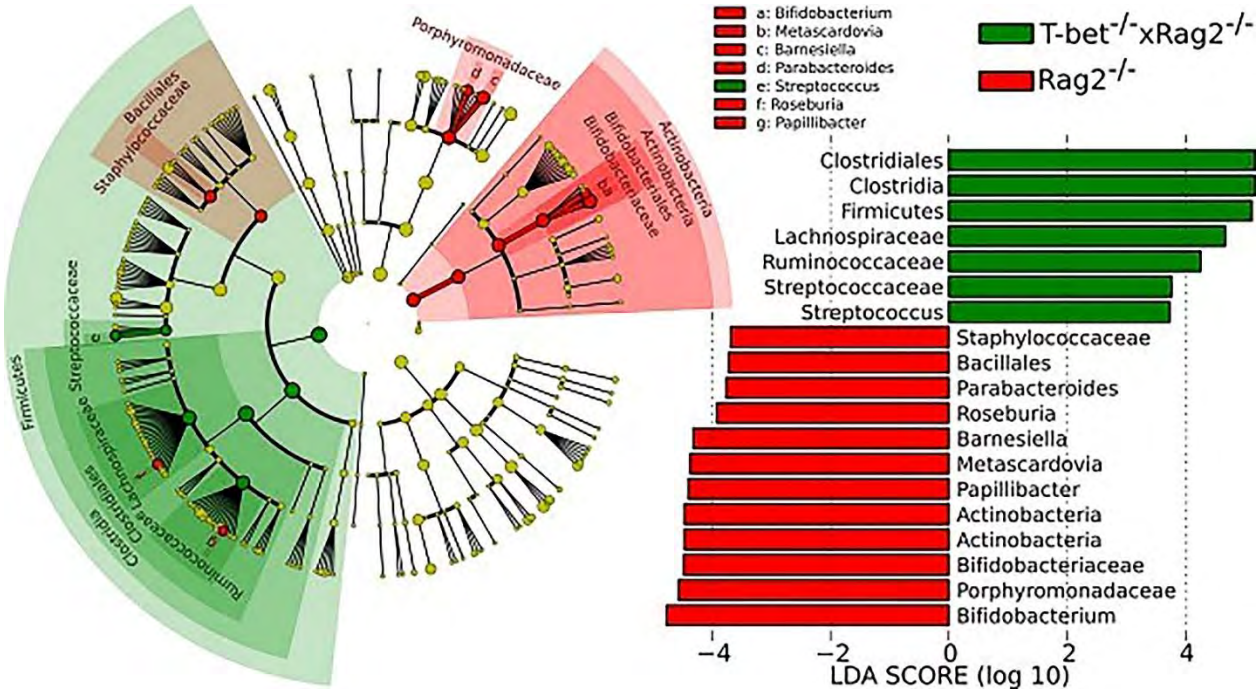


图 11.5　LEfse 结果可视化呈现（引自 Segata et al., 2011）

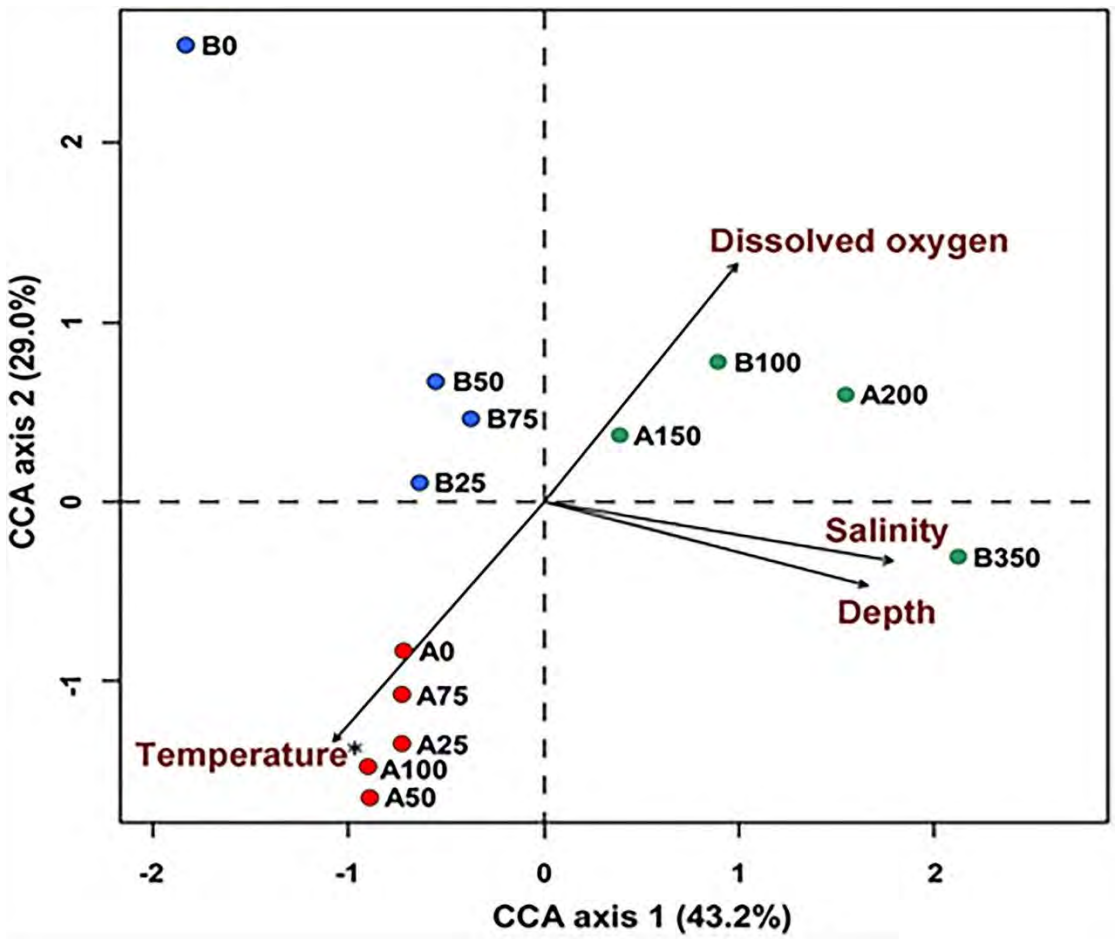


图11.6　菌群分布与环境因子之间关系的CCA分析图（引自Yu et al.， 2015）

图中箭头代表不同的环境因子，射线越长表示该环境因子影响越大；环境因子之间的夹角为锐角时表示两个环境因子之间呈正相关关系，钝角时表示呈负相关关系

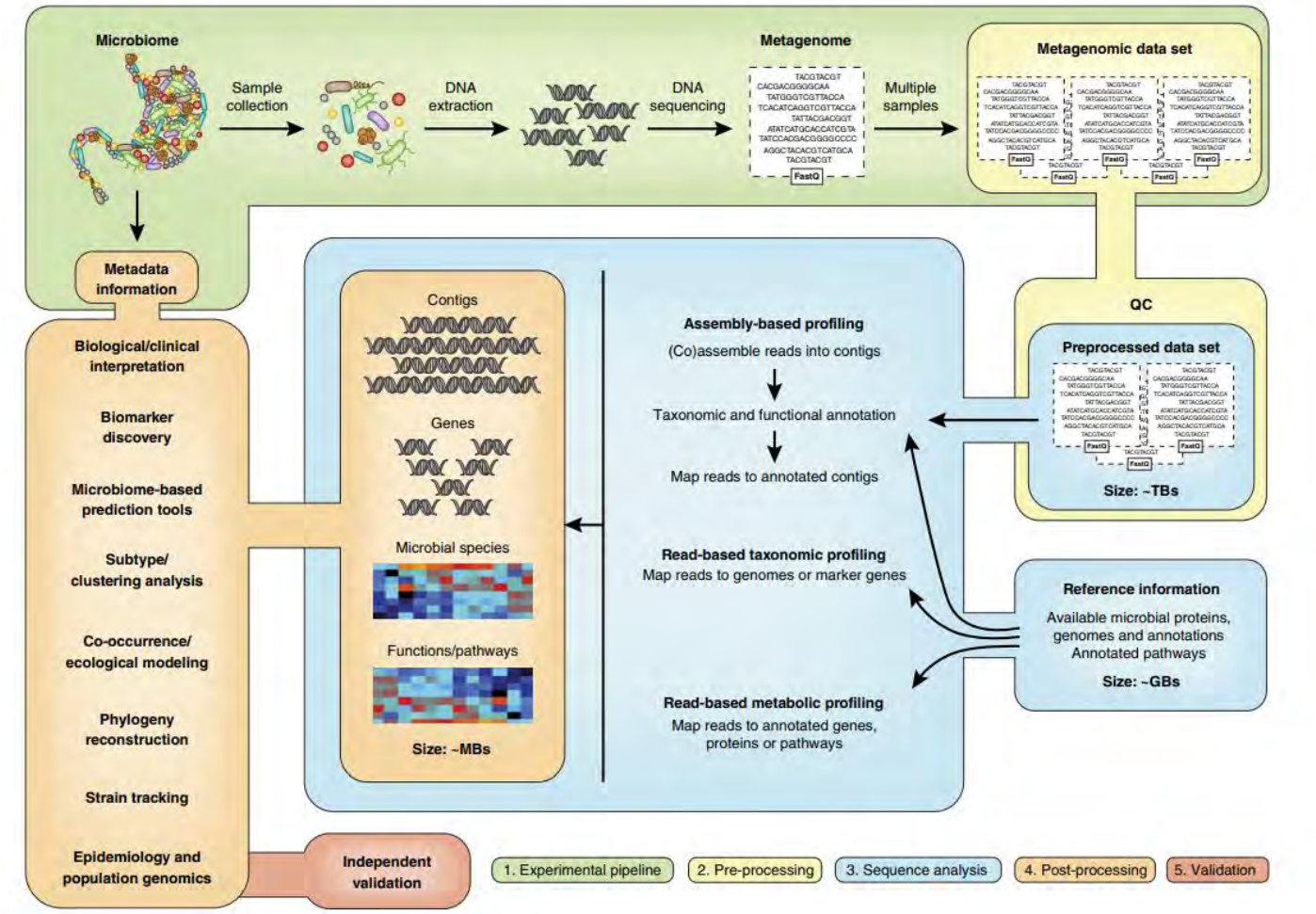
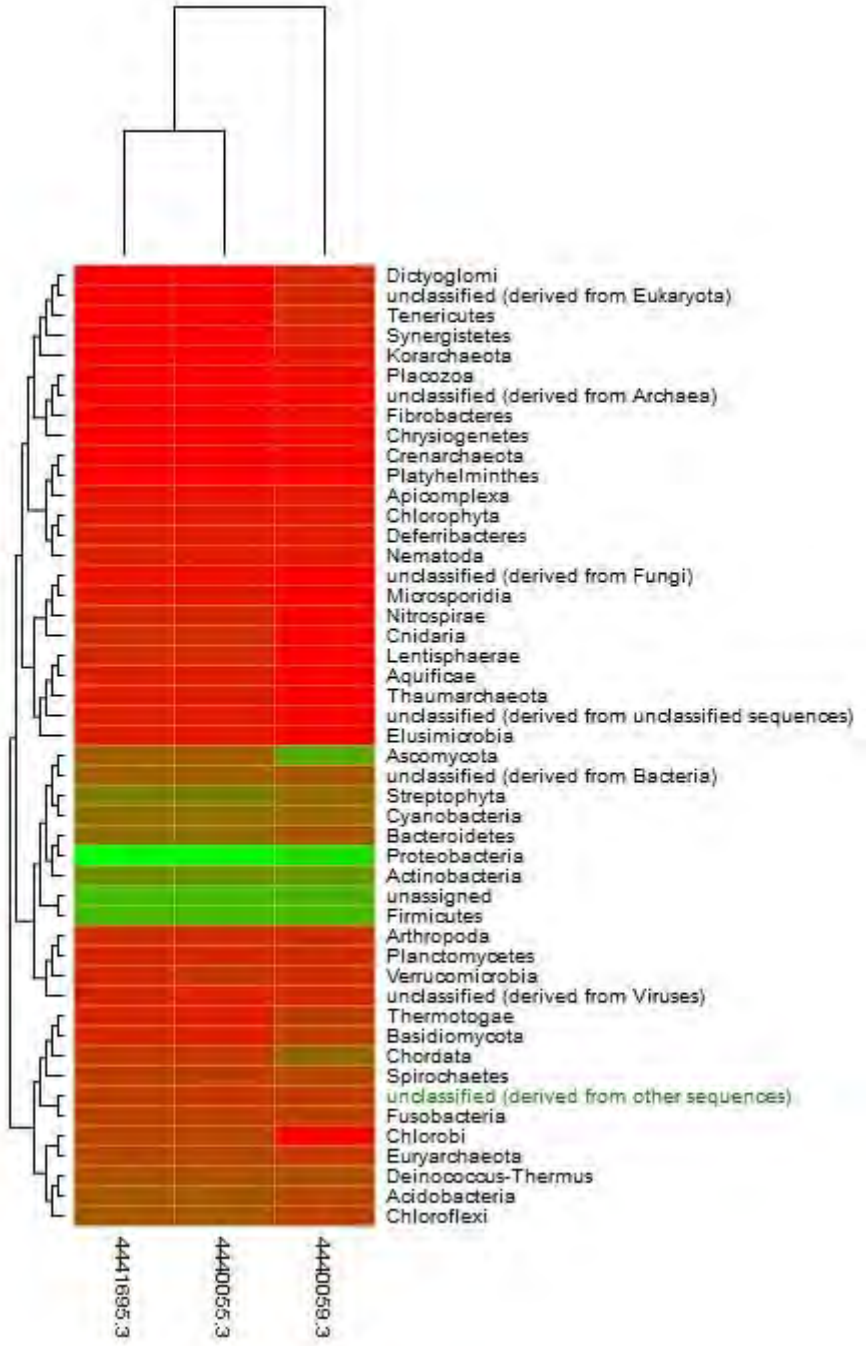
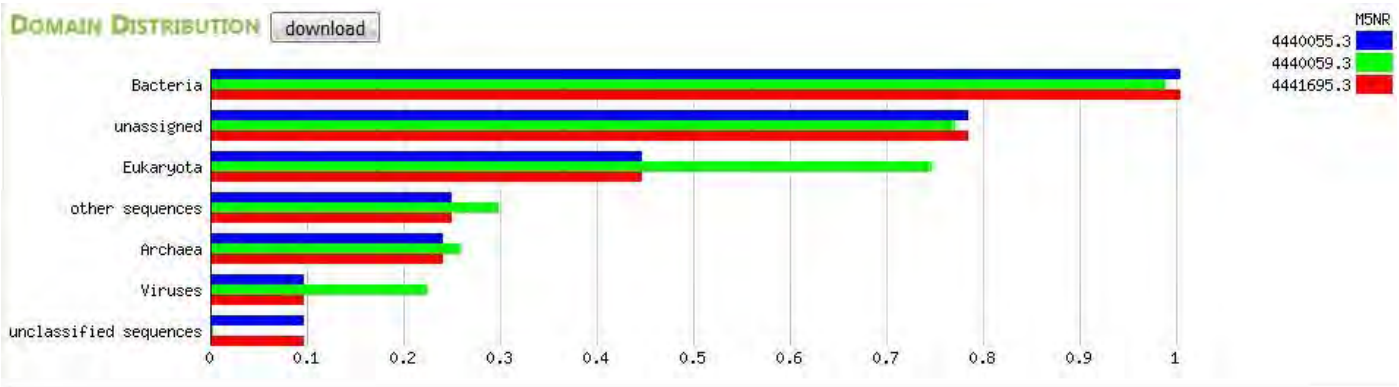


图11.7　基于全基因组数据的宏基因组分析流程（引自Quince， 2017）

主要步骤包括：①研究内容和实验设计；②分析预处理，通过质控步骤以降低原始测序偏好性和不需要的信息；③序列分析， 根据实验目标采用基于读序和基于拼接的方法；④后续处理，根据研究内容选择应用各种多元变量统计方法；⑤进行相应验证


Dictyoglomi
unclassified (derived from Eukaryota)
Tenericutes
Synergistetes
Korarchaeota
Placozoa
unclassified (derived from Archaea)
Fibrobacteres
Chrysiogenetes
Crenarchaeota
Platyhelminthes
Apicomplexa
Chlorophyta
Deferribacteres
Nematoda
unclassified (derived from Fungi)
Microsporidia
Nitrospirae
Cnidaria
Lentisphaerae
Aquificae
Thaumarchaeota
unclassified (derived from unclassified sequences)
Elusimicrobia
Ascomycota
unclassified (derived from Bacteria)
Streptophyta
Cyanobacteria
Bacteroidetes
Proteobacteria
Actinobacteria
unassigned
Firmicutes
Arthropoda
Planctomycetes
Verrucomicrobia
unclassified (derived from Viruses)
Thermotogae
Basidiomycota
Chordata
Spirochaetes
unclassified (derived from other sequences)
Fusobacteria
Chlorobi
Euryarchaeota
Deinococcus-Thermus
Acidobacteria
Chloroflexi
4441695.3
4440055.3
4440059.3


DOMAIN DISTRIBUTION
download
M5NR
4440055.3
4440059.3
4441695.3
Bacteria
unassigned
Eukaryota
other sequences
Archaea
Viruses
unclassified sequences
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

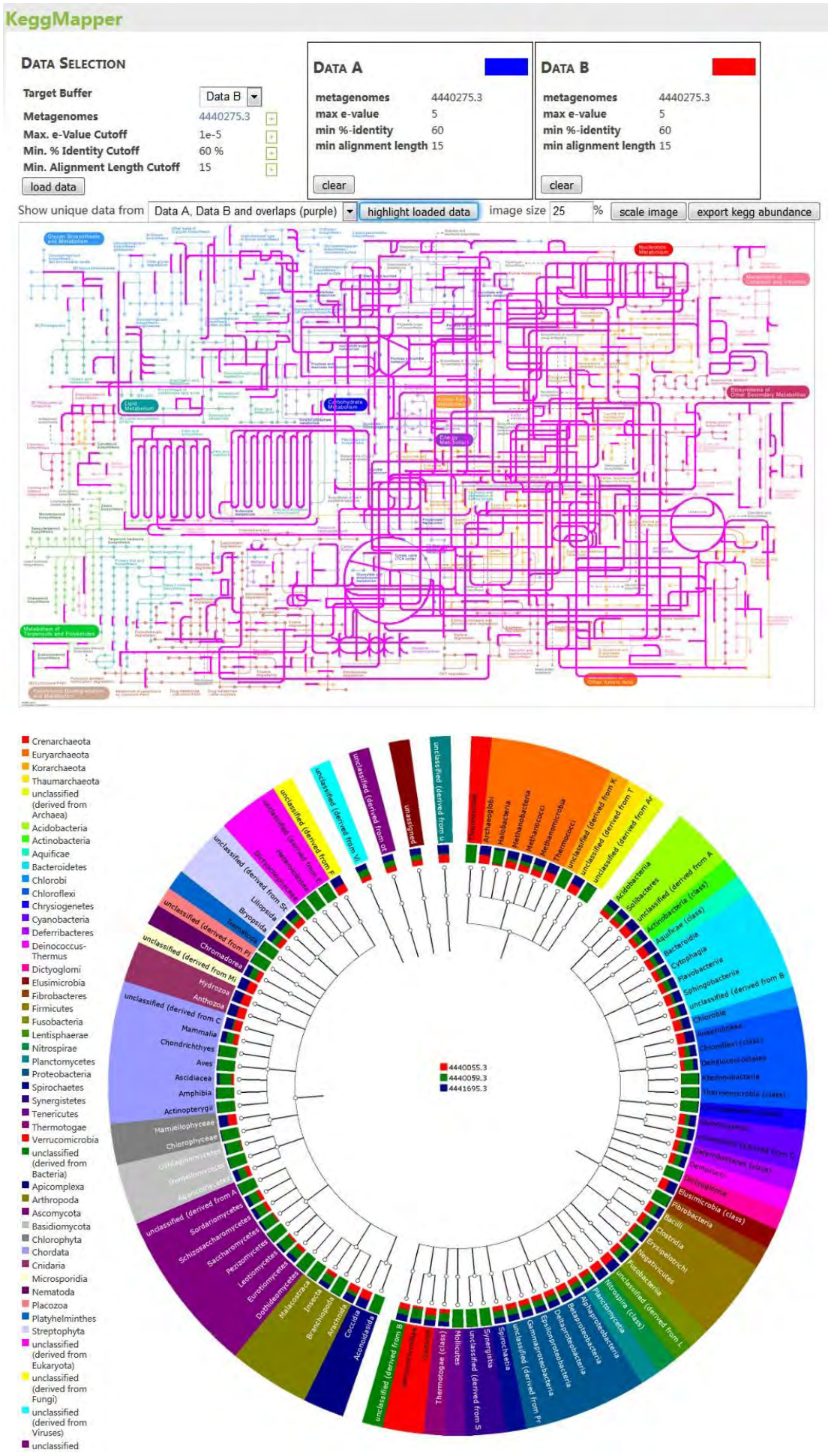


图 11.8　MG-RAST 宏基因组注释系统的分析结果界面

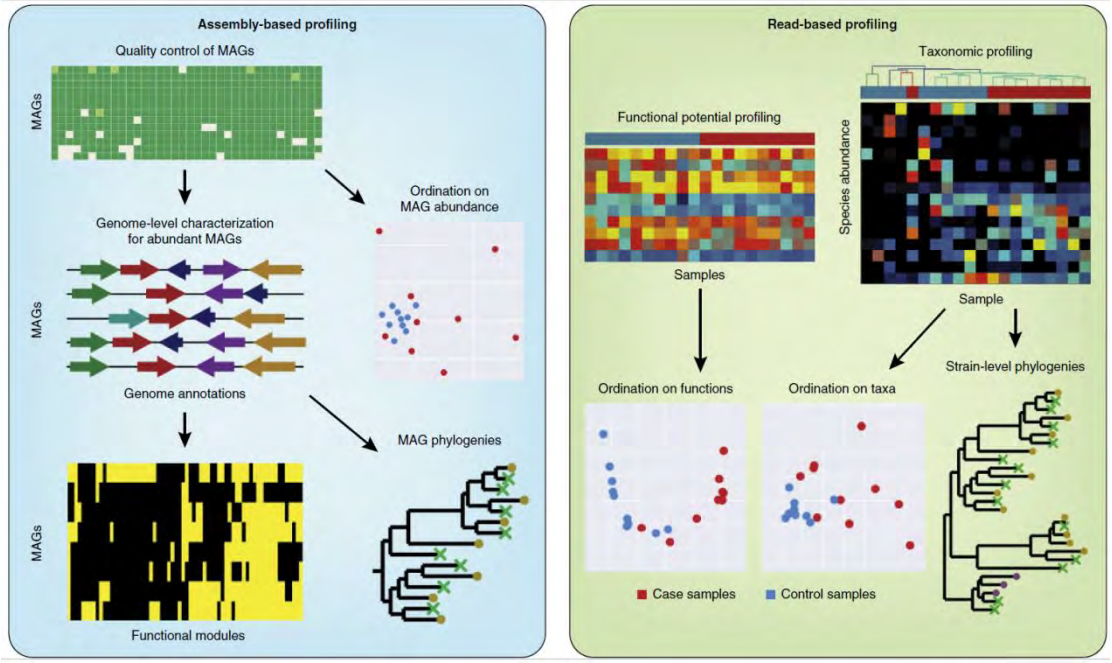


图 11.9　基于读序和基因组拼接的基因组重构方法比较（引自 Quince， 2017）

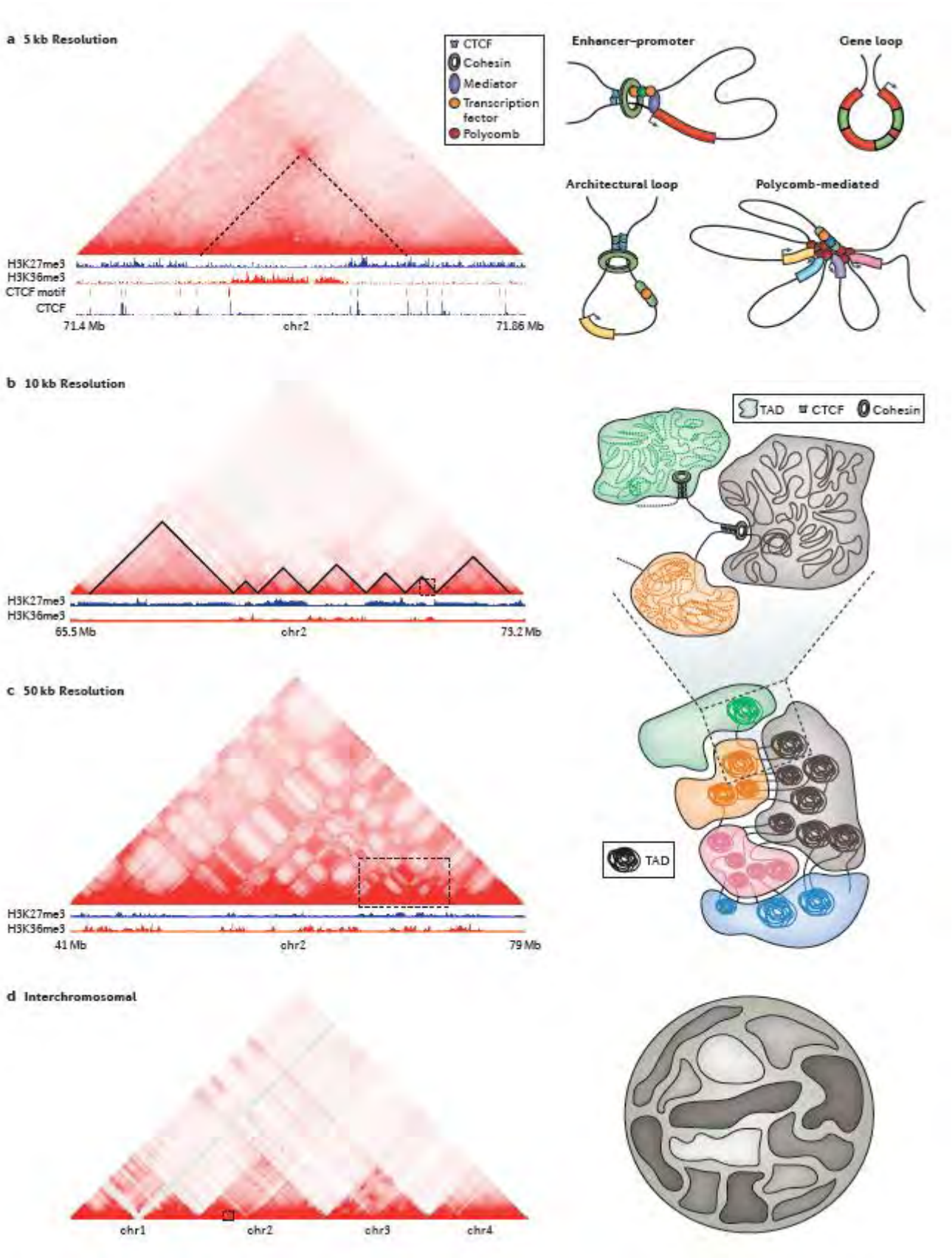


图12.1　染色体分级结构（引自Bonev and Cavalli，2016）

A～D 分别对应 5kb（染色质环）、10kb（拓扑关联域）、50kb（活性/ 惰性区室）、Mb（染色体疆域）这几个层次的不同分辨率下的染色质结构

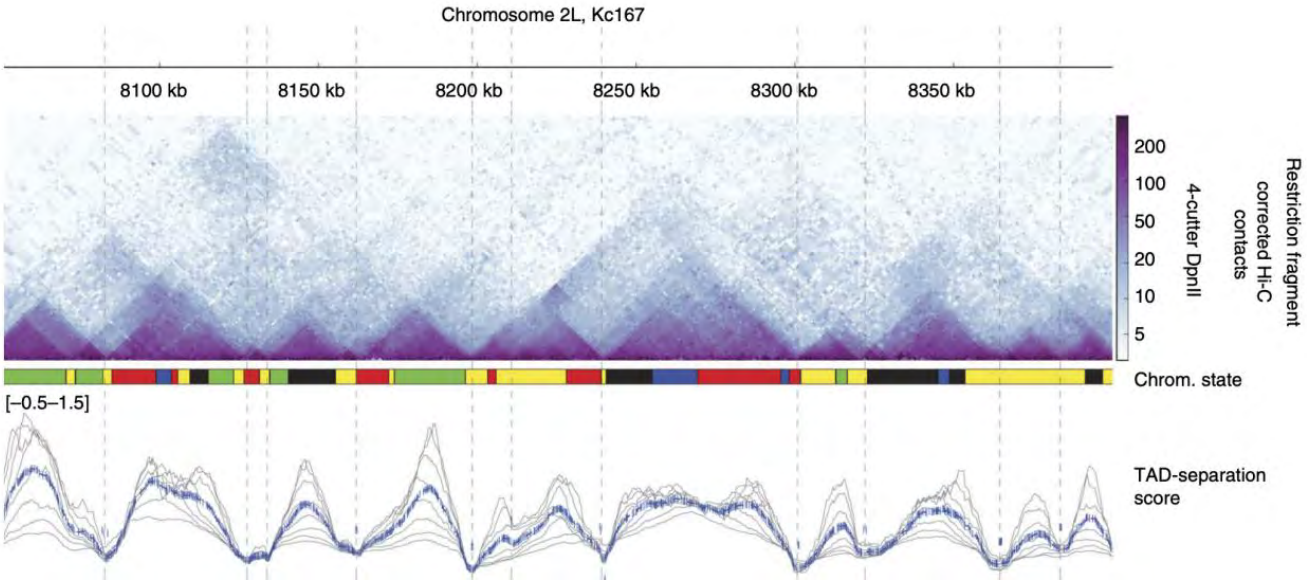


图12.2　高分辨率TAD 鉴定方法举例（引自Fidel et al.，　2018）

图上部是来自果蝇2L 染色体一段350kb 区域的Hi-C 接触矩阵；中间区域代表染色质状态，红色和黄色代表活性染色质，黑色、蓝色和绿色代表非活性染色质；底部代表 TAD 分类得分情况，垂直虚线为预测的 TAD 边界

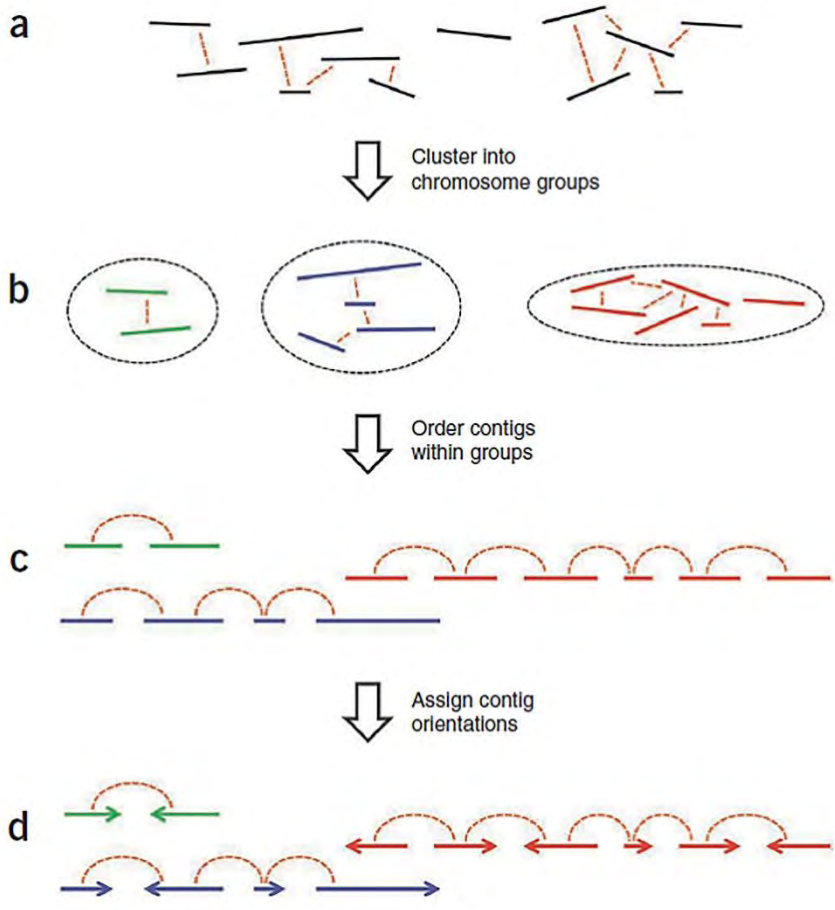


图12.3　经典Hi-C 组装软件Lachesis 的组装流程

（引自 Burton et al.，　2013）

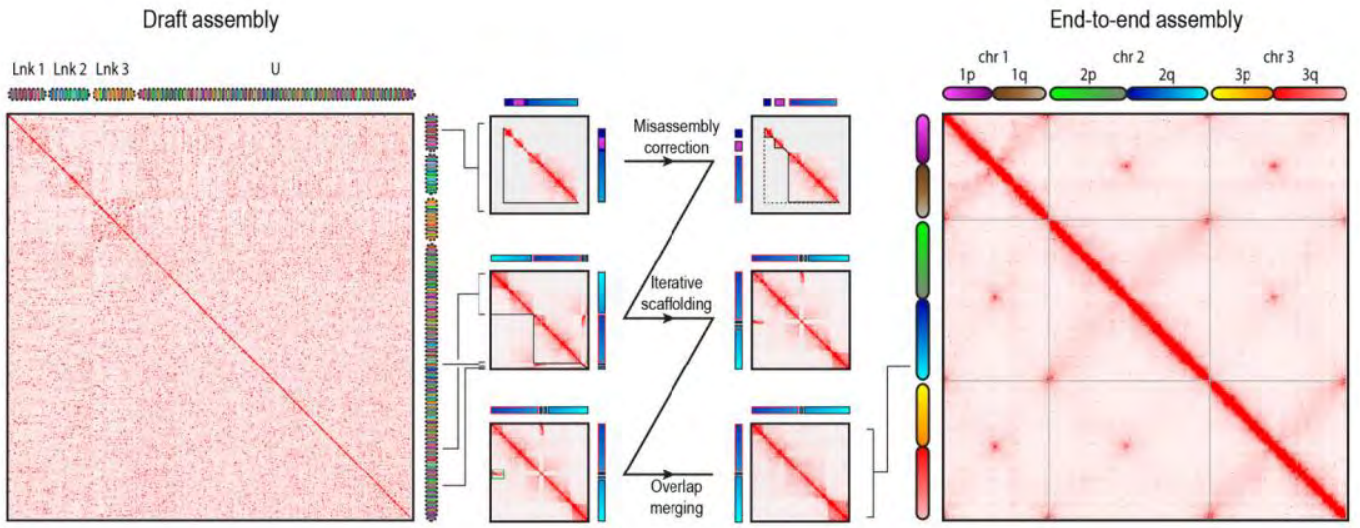


图12.4　Hi-C 组装软件3d-dna 流程图（引自Dudchenko et al.， 2017）

左图代表未通过 3d-dna 进行矫正组装前的 Hi-C 互作矩阵图谱，包含连锁群 link1、link2、link3 和未分类区域（U）；中图代表 3d-dna 利用 Hi-C 数据进行矫正和组装的流程，包含 Hi-C 数据矫正、迭代搭建框架（scaffolding）和合并重叠区域三个步骤； 右图为通过 3d-dna 完成矫正组装后的 Hi-C 互作矩阵图谱。可以明显发现，获得染色体 1、染色体 2 和染色体 3 三个连锁群， 证明了 3d-dna 的实际效果

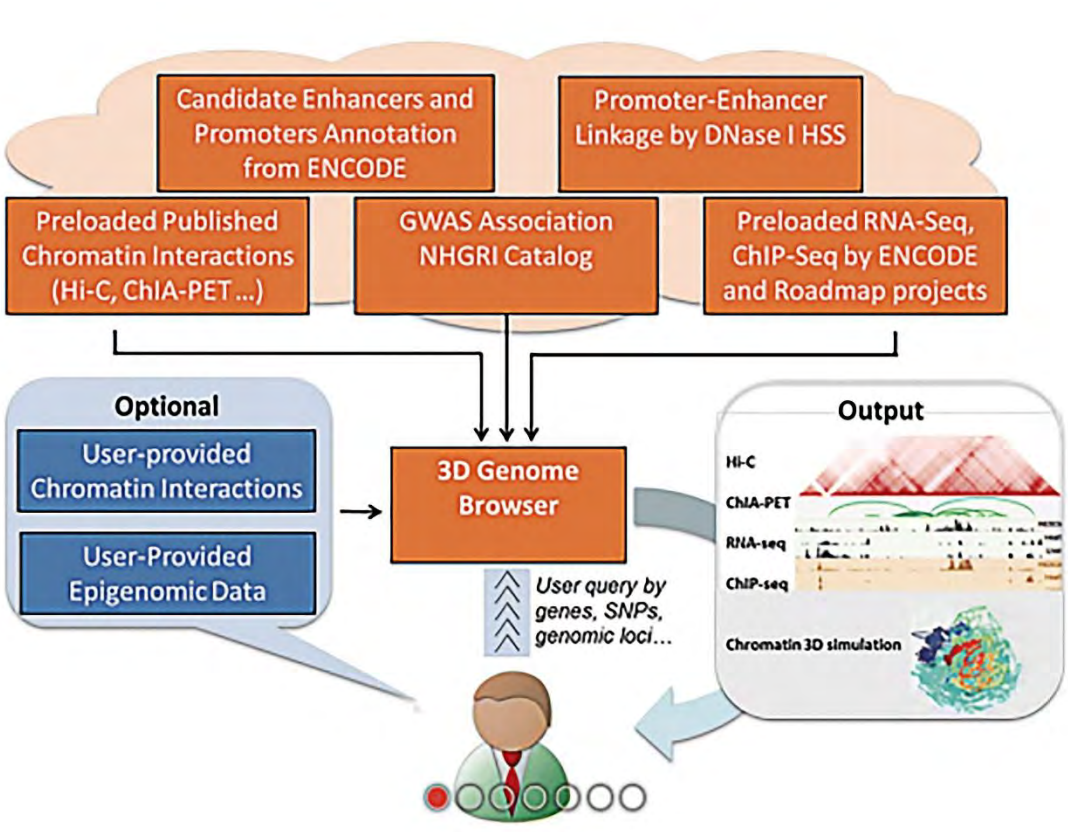


图12.5　三维基因组浏览器3D Genome Browser的整体设计（引自Wang et al.， 2018）

ENCODE（encyclopedia of DNA elements）．美国 NHGRI 资助的一个国际合作项目；NHGRI（National Human Genome Research Institute）．美国 NIH 下属研究机构之一；DNase Ⅰ HSS．脱氧核糖核酸酶Ⅰ超敏感位点

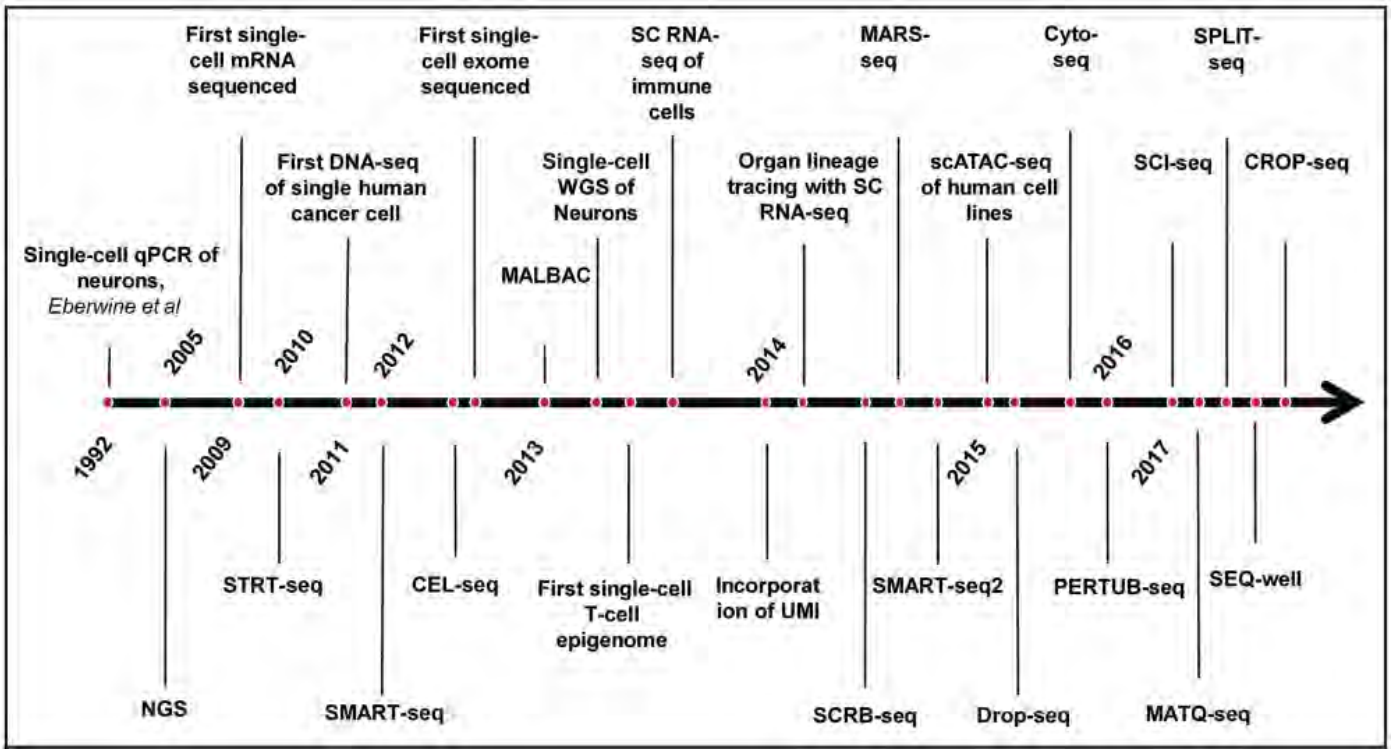


图 12.6　单细胞组学研究历程（引自 Paolillo et al.， 2019）

图 12.7　单细胞 RNA（scRNA）研究进展（引自 Chen et al.，2021a）

图中列出了最近 5 年利用不同类物种和不同单细胞技术平台发表的 1244 篇论文情况。植物单细胞研究细胞数单独框出

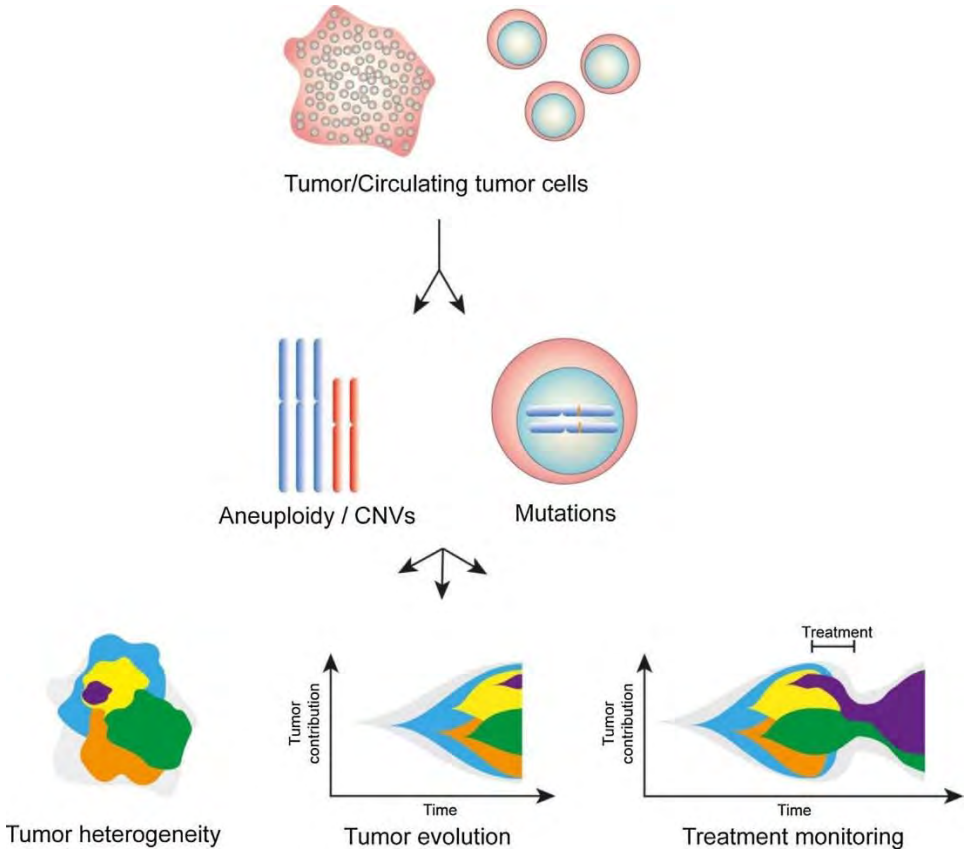


图12.8　单细胞DNA测序的应用——以肿瘤生物学研究为例（引自Bos et al.，2018）

通过测序肿瘤细胞或循环肿瘤细胞，可以鉴定非整倍性、拷贝数变异和/或突变。这些可用来表征肿瘤的异质性，确定其进化路径，并有利于监测治疗

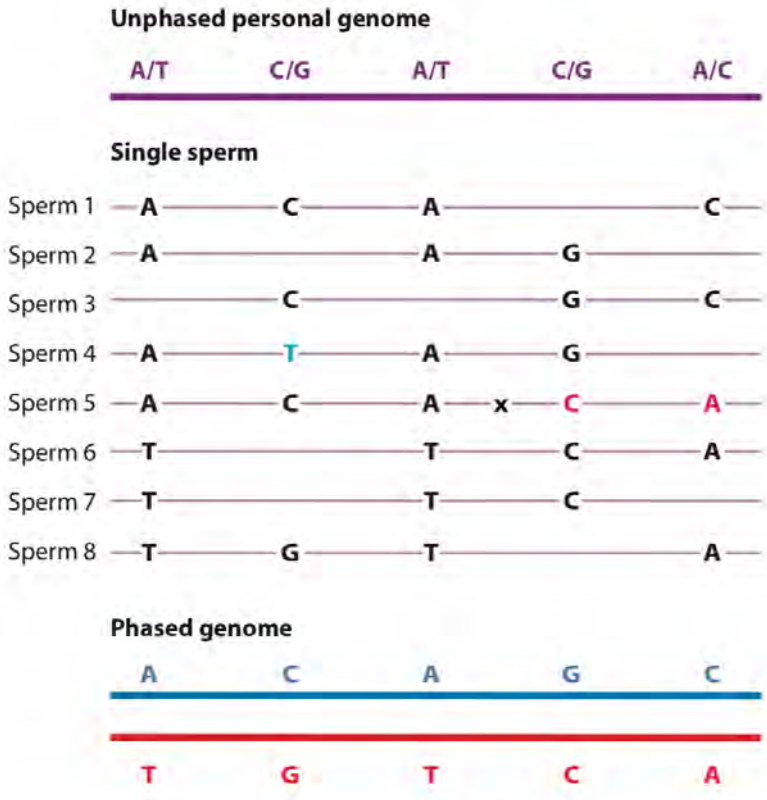


图12.9　通过单倍体单细胞基因组SNP连锁关系获得个体基因组相位分型信息（引自Huang et al.， 2015）

A．对二倍体进行普通基因组测序，发现其包含 5 个杂合单核苷酸多态性（SNP）位点。B．对单个精子细胞进行基因组测序，其中蓝色“T”表示全基因组扩增或测序产生的错误；黑色“x”代表重组中的交叉点，即父本与母本 DNA 之间的转换点。C．最终利用每个单个精子细胞的 SNP 连锁信息，确定基因组相位

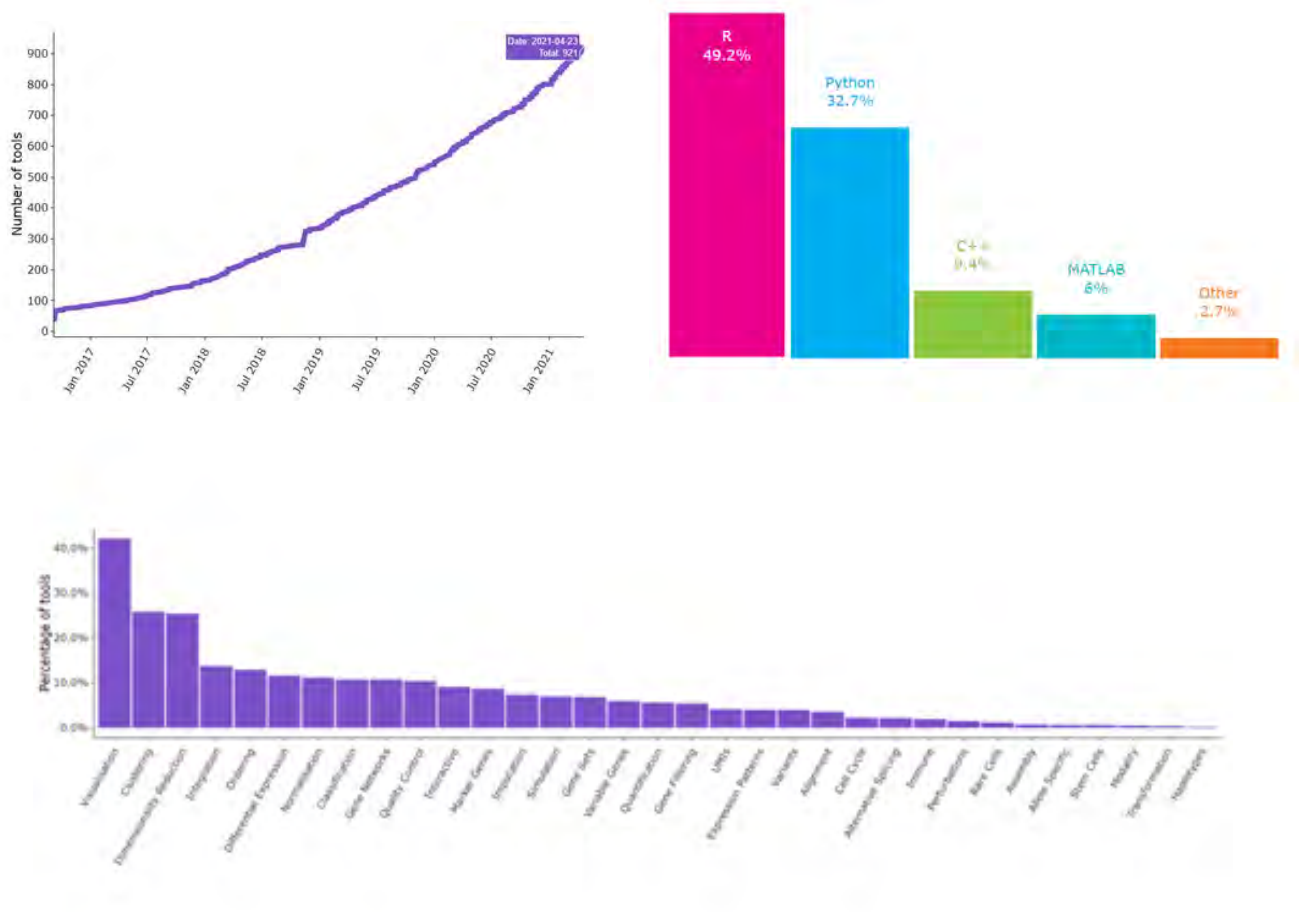


图12.10　单细胞转录组数据分析软件统计（资料来源：https://www.scrna-tools.org/table）

A．单细胞转录组数据分析软件数量增长情况；B．分析软件所用计算机语言；C．单细胞转录组数据分析涉及主题和比例

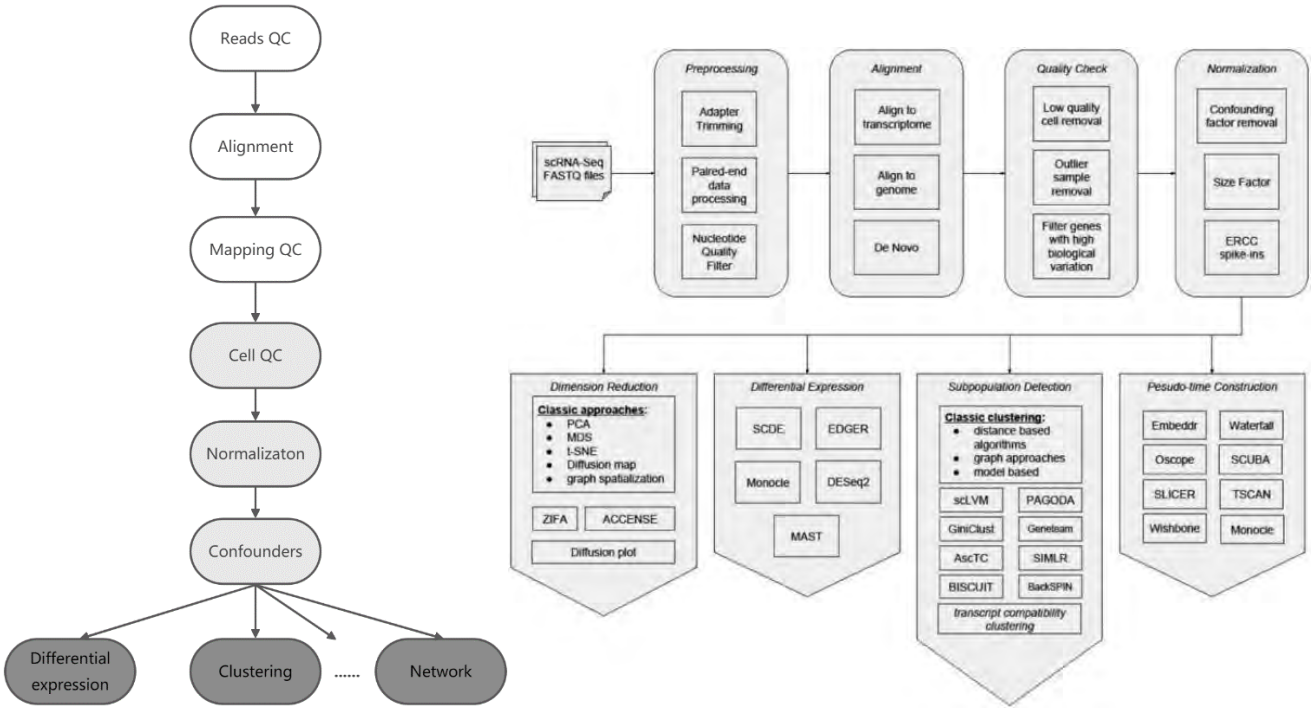
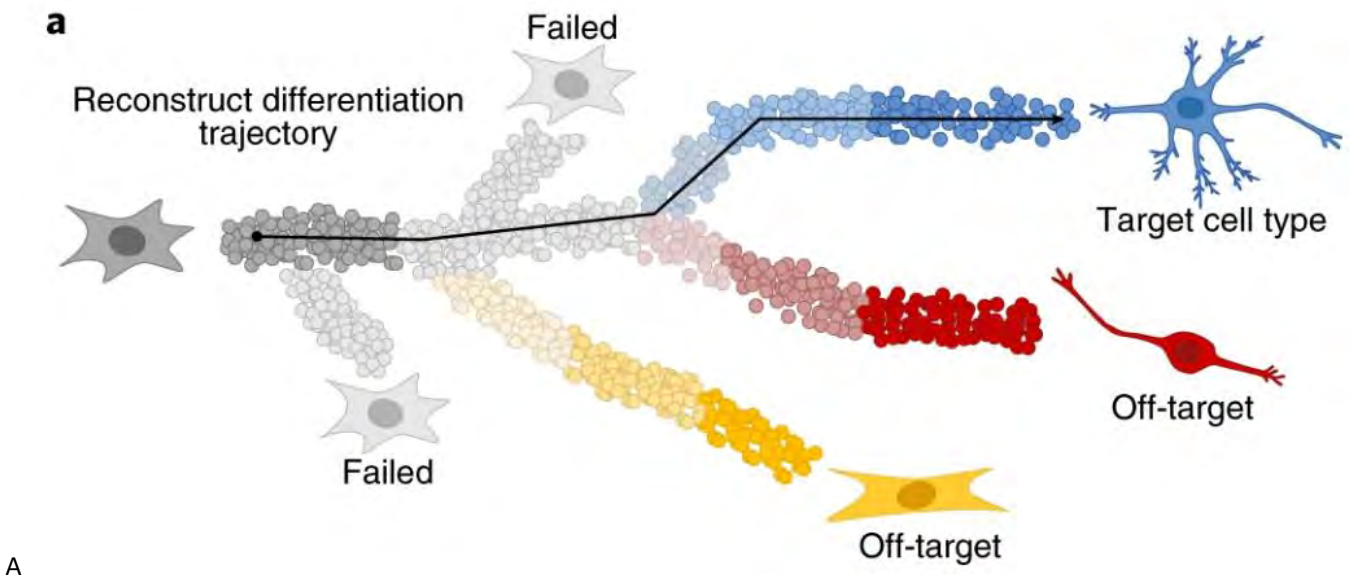


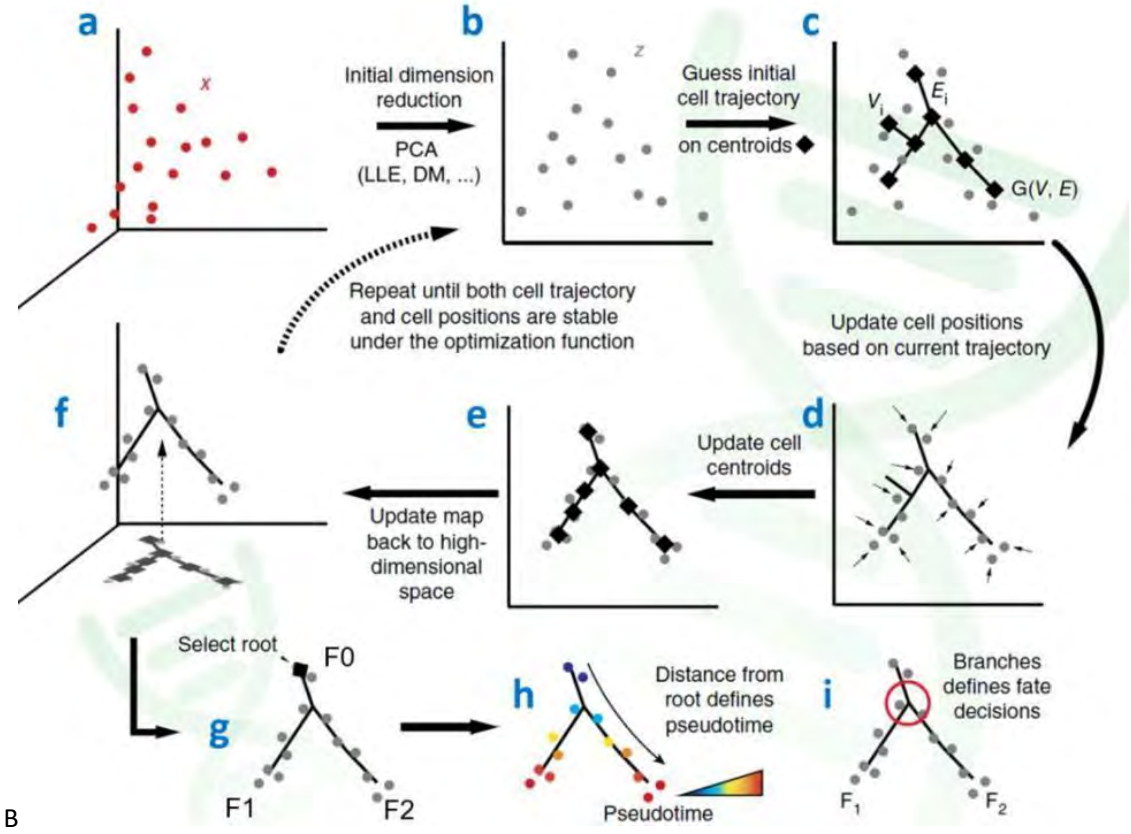
图12.11　单细胞转录组分析总体流程（A）和具体分析内容及其生物信息学工具（B）

（引自Poirion et al.，2016；Luecken et al.，2019）

PCA．主成分分析；MDS．多维尺度变换；*t*-SNE．*t*-随机邻近嵌入；ERCC（external RNA control consortium）．外源 RNA 定量协作组；ZIFA（zero-inflated dimensionality reduction algorithm）．零膨胀降维算法；ACCENSE（automatic classification of cellular expression by nonlinear stochastic embedding）．基于非线性随机嵌入的细胞表达式自动分类



A



B

图12.12　拟时序分析的意义和算法

A.　拟时序分析的意义（引自 Camp et al.， 2018）。通过分化轨迹重构，可以发现现有方法未发现（off-target）或无法发现（failed）的发育细胞类型。B．拟时序分析算法流程图（引自 Carter et al.， 2018）。a．每个细胞均表示为高维空间 X 中的一个点， 其中每个维度对应于有序基因的表达水平；b．通过降维方法（如 PCA）将数据投影到较低维的空间 Z 中；c．使用 K-means 等聚类方法自动选择的一组质心（菱形点）来构建最小生成树；d．然后将其余细胞移向最近的树顶点；e．同时将顶点位置进行更新，学习新的生成树；f．然后迭代该过程，直到树和单元格均收敛为止；g．根据背景知识选择一个树的尖端作为“根”（即发育起点）；h．计算每个像元的拟时序作为其沿树到根的测地距离；i．根据主图自动分配其分支

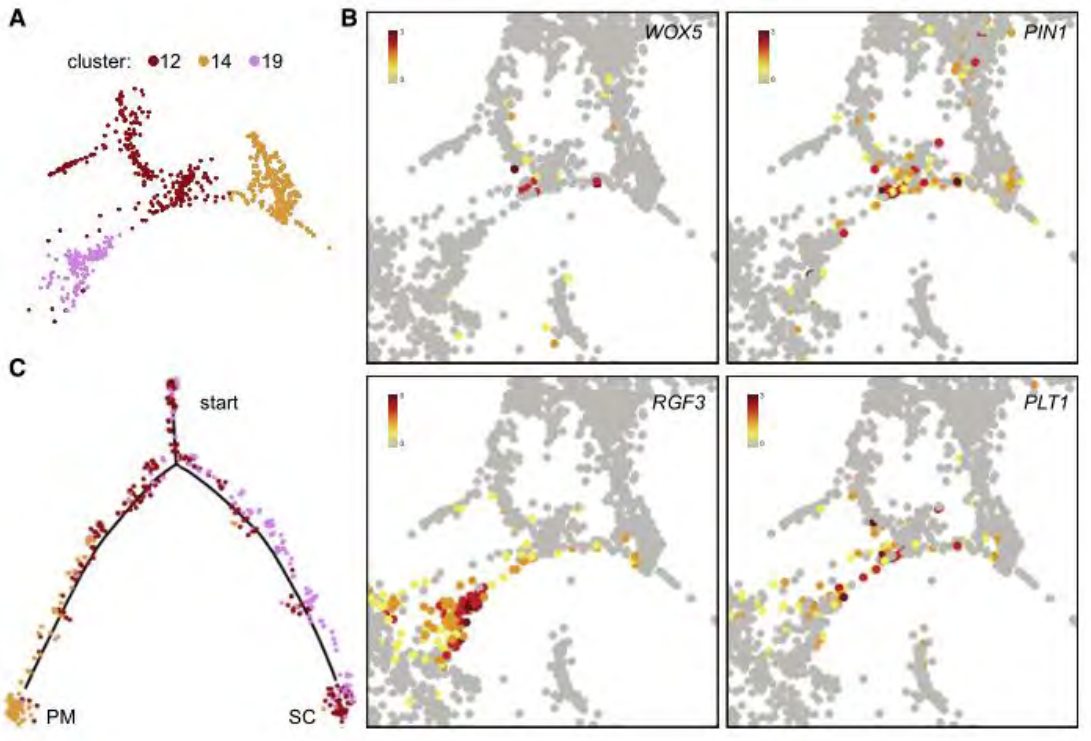
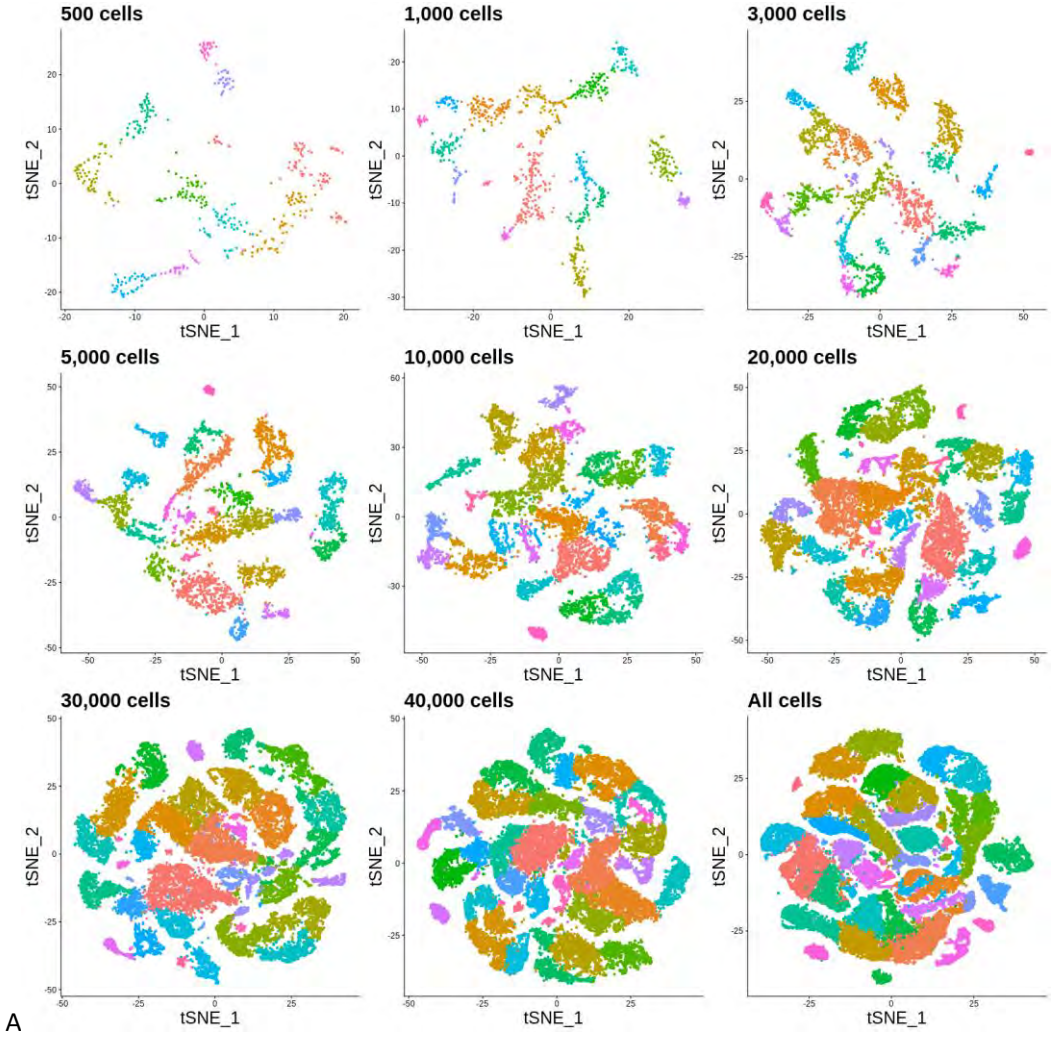


图12.13　拟南芥根部单细胞拟时序分析结果（引自Zhang et al.， 2019c）

A.　干细胞生态位（SCN）细胞 UMAP 降维聚类获得近端分生组织（PM）、干细胞（SC）类群（cluster12、14 和 19）；B．根分生组织标记基因（*WOX5*、*PIN1*、*RGF3* 和 *PLT1*）的表达模式，颜色代表 UMAP 图上单个细胞中这些基因的表达水平，彩色条指示相对表达水平；C．SCN 细胞拟时序分析



A

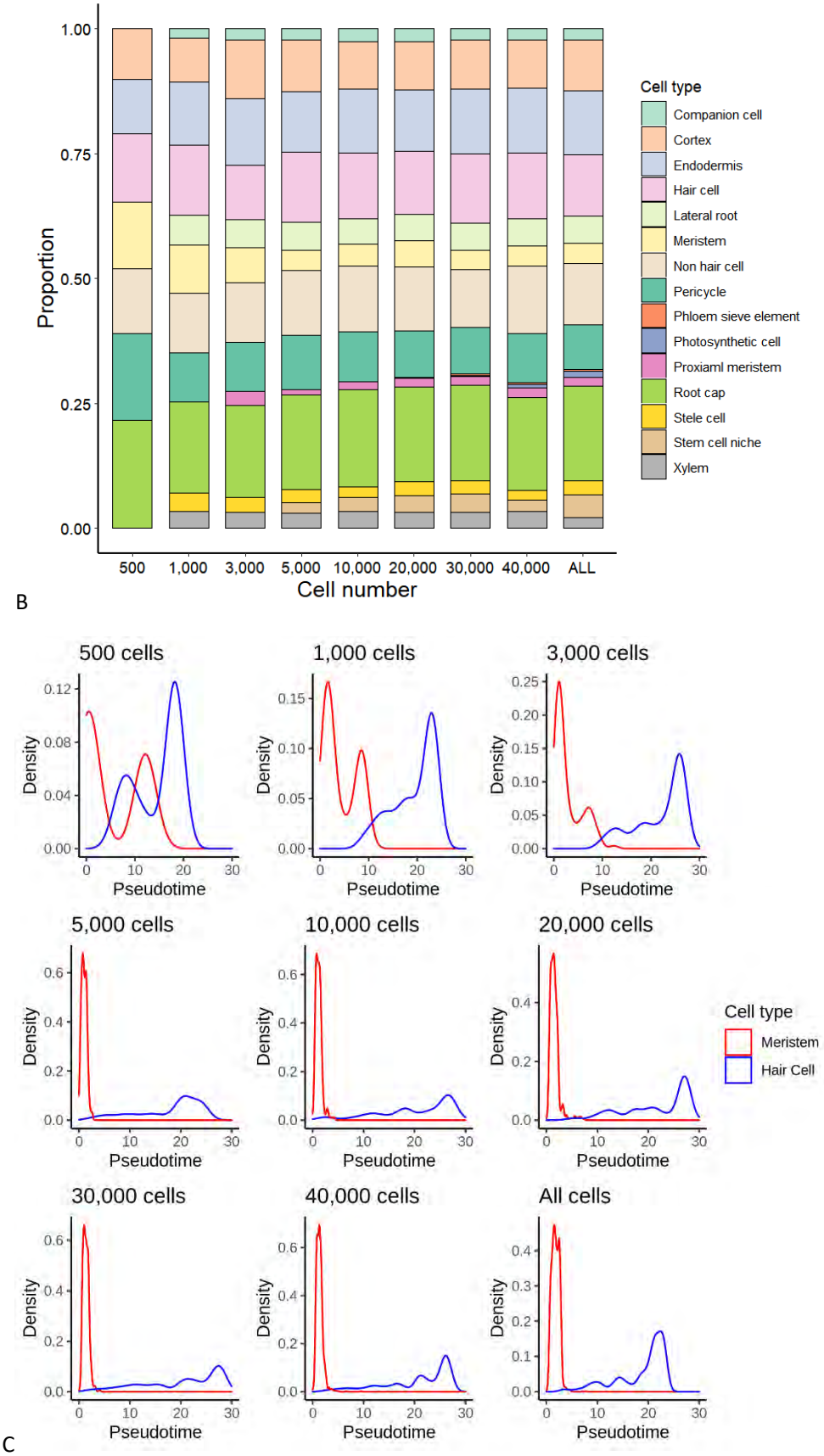


图 12.14　不同单细胞数量对细胞类型鉴定及拟时序分析结果的影响（引自 Chen et al.， 2021a）

A．基于不同拟南芥根单细胞测序数量进行 *t*-SNE 聚类的结果比较，单细胞测序数量从 500 个依次到 40 000 个和全部 （约 57 000 个）细胞（all cells）；B．不同细胞数下鉴定到的拟南芥根细胞类型比较；C．不同细胞数下，拟南芥根分生组织和 根毛细胞发育拟时序分析结果比较

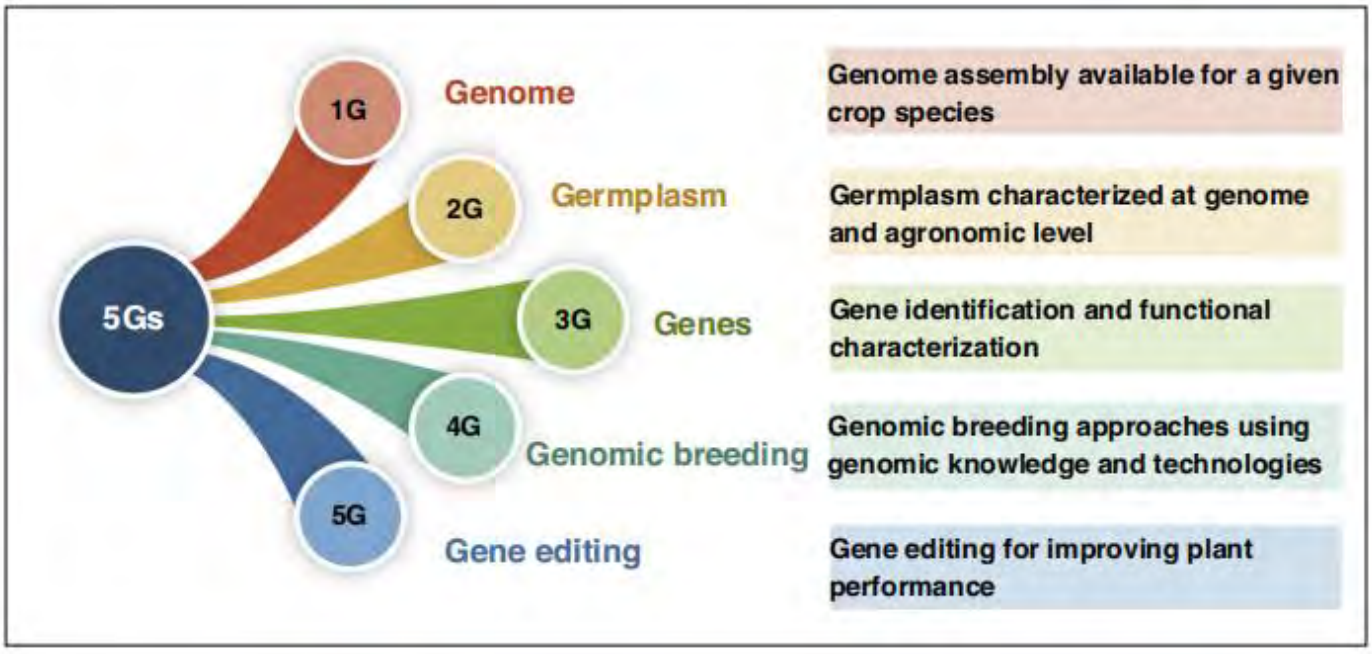


图12.15　以生物信息学技术为基础的作物“5G”遗传改良技术或策略（引自Varshney et al.， 2020）

所谓“5G”育种策略，第一个“G”是作物物种的基因组测序组装（genome assembly），第二个“G”是种质资源基因组特征和农艺性状调查（germplasm characterization），第三个“G”是基因及其功能鉴定（gene identification and functions），第四个“G”是基因组育种方法（genomic breeding methodologies），第五个“G”是基因编辑技术（gene editing）。这 5G 的实现都离不开生物信息学方法和工具

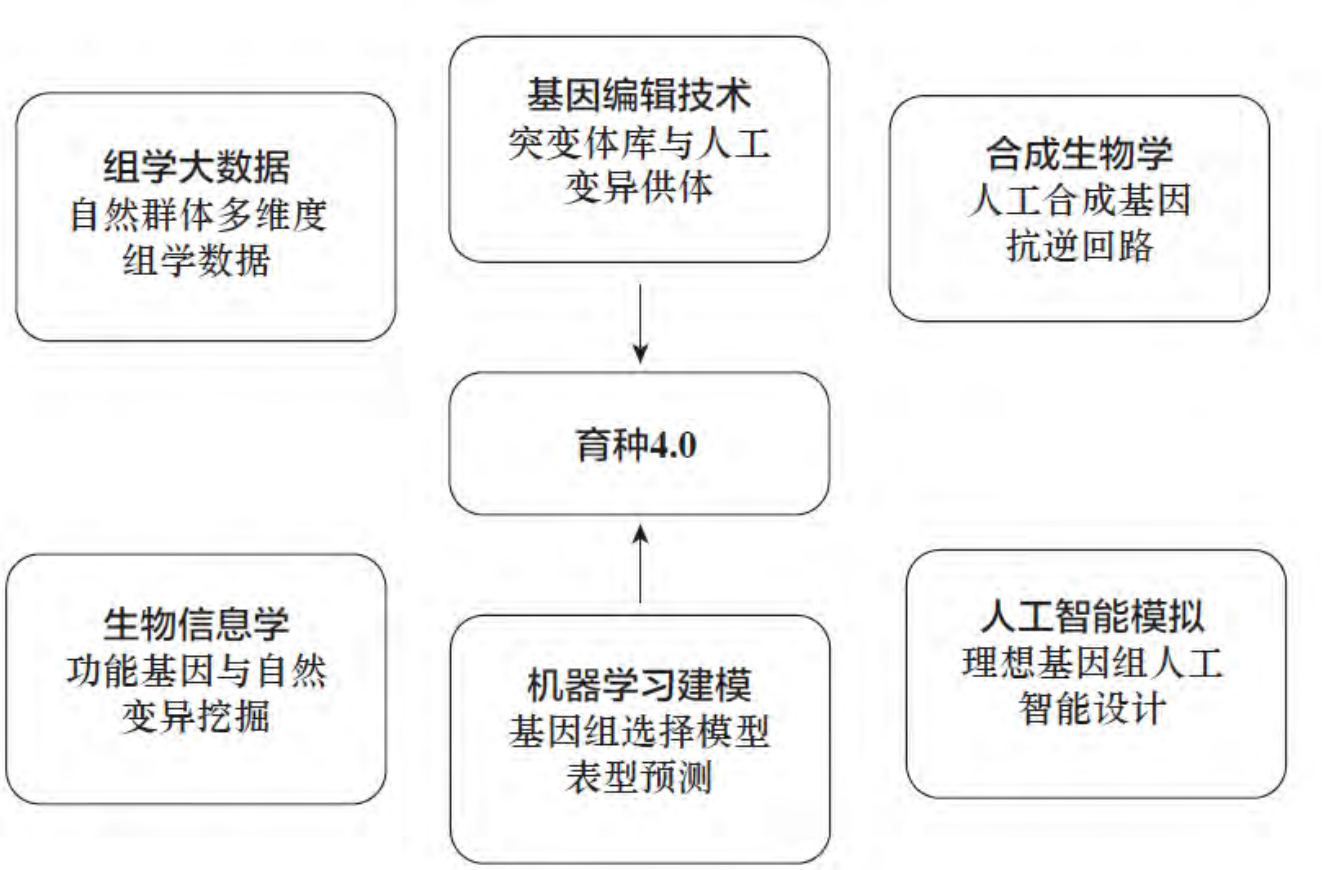


图12.16　依托组学大数据及其多层次生命科学与信息科学技术的现代育种技术——育种4.0（引自王向峰和才卓， 2019）

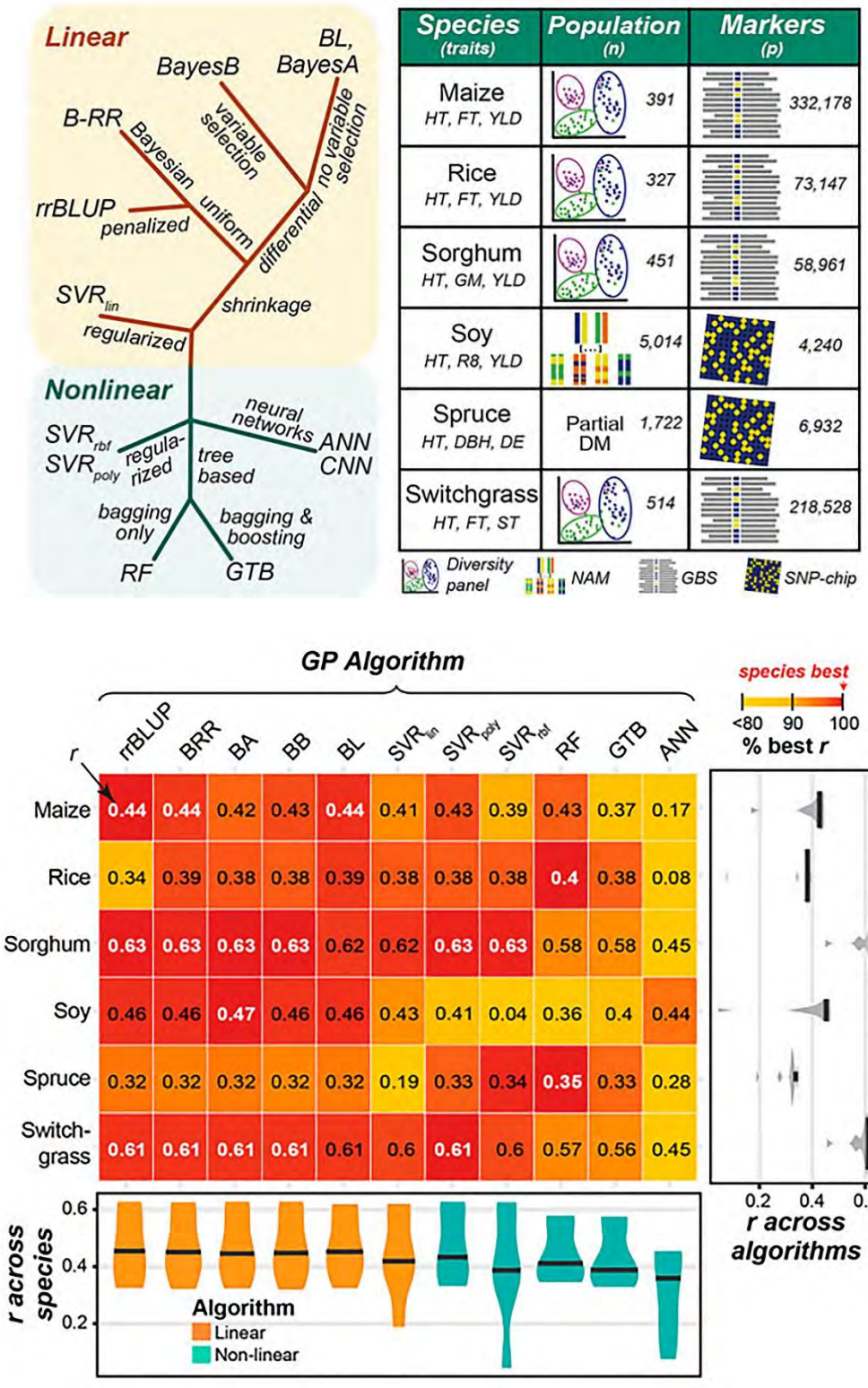


图12.17　复杂农艺性状的基因组预测方法及其应用（引自Azodi et al.， 2019）

A．基因组预测方法及其应用物种训练数据情况。左图：主要基因组预测方法及其相互关系。树中棕色线条表明该方法只用于线性关系的预测，绿色表明该方法同时适用于线性关系、非线性关系的预测。树中各个方法的位置表明其各自之间的关系。分支上的标签介绍了该方法与其他方法的区别。各个预测方法名称为：岭回归最佳线性无偏预测（ridge regression BLUP，RR-BLUP）；贝叶斯岭回归（BRR，Bayesian ridge regression）；贝叶斯 A（BA，Bayes A）；贝叶斯 B（BB，Bayes B）；贝叶斯 LASSO 回归（Bayesian LASSO，BL）；支持向量回归（SVR，support vector regression）；随机森林（RF，random forest）；梯度树提升（GTB，gradient tree boosting）；人工神经网络（ANN，artificial neural network）；卷积神经网络（CNN，convolutional neural network）。右图：应用涉及的植物物种及其性状和训练数据集群体类型、大小和标记数字。性状：株高（HT）、花期（FT）、产量（YLD）、谷物含水量（GM）、R8 发育时间（R8）、树径（DBH）、木材密度（DE）、直立度（ST）。群体类型：资源群体、巢式关联作图（NAM）和部分双列杂交（partial DM）。标记测定方法：简化测序（GBS）和单核苷酸多态性芯片（SNP-chip）。B．不同基因组预测算法预测玉米等物种株高总体效果。网格中数字为效果均值（皮尔逊相关系数，r），网格颜色表示最佳效果均值的相对比例（最佳 r 为红色），白色数字为最佳效果均值。小提琴图显示每个特征（右）和算法（下）的 r 值的中位数和分布。GP．基因组预测

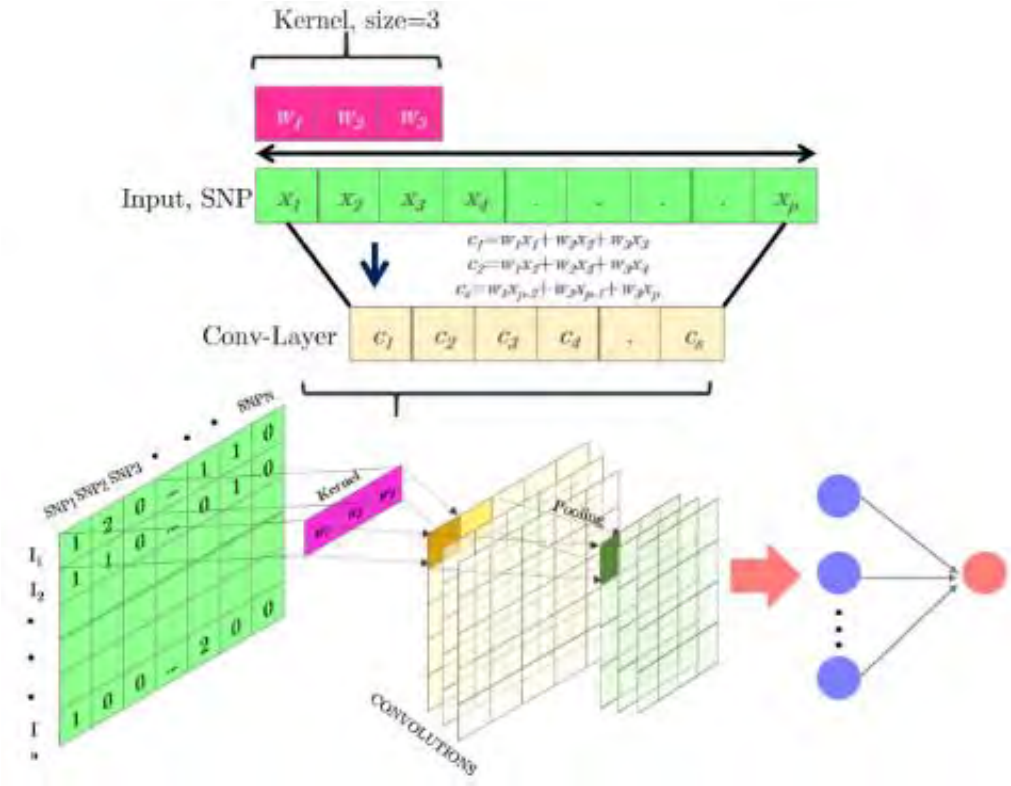


图12.18　基因组育种利用的一维（1D）卷积神经网络结构示意图
（引自 Pérez-Enciso and Zingaretti， 2019）

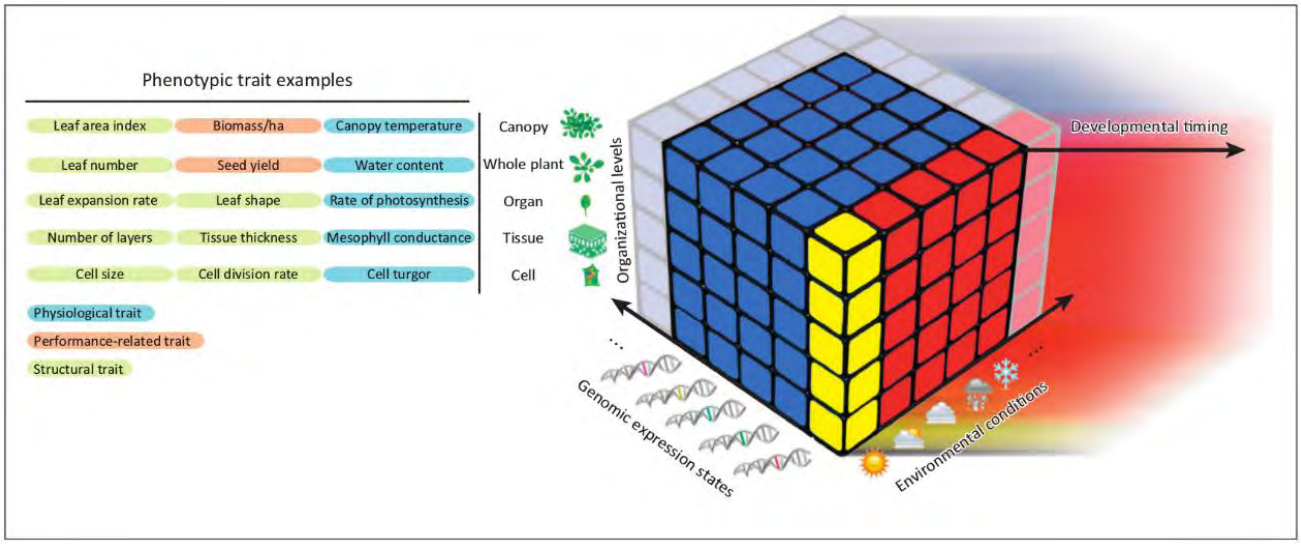


图 12.19　从植物表型到表型组学（引自 Dhondt et al.， 2013）

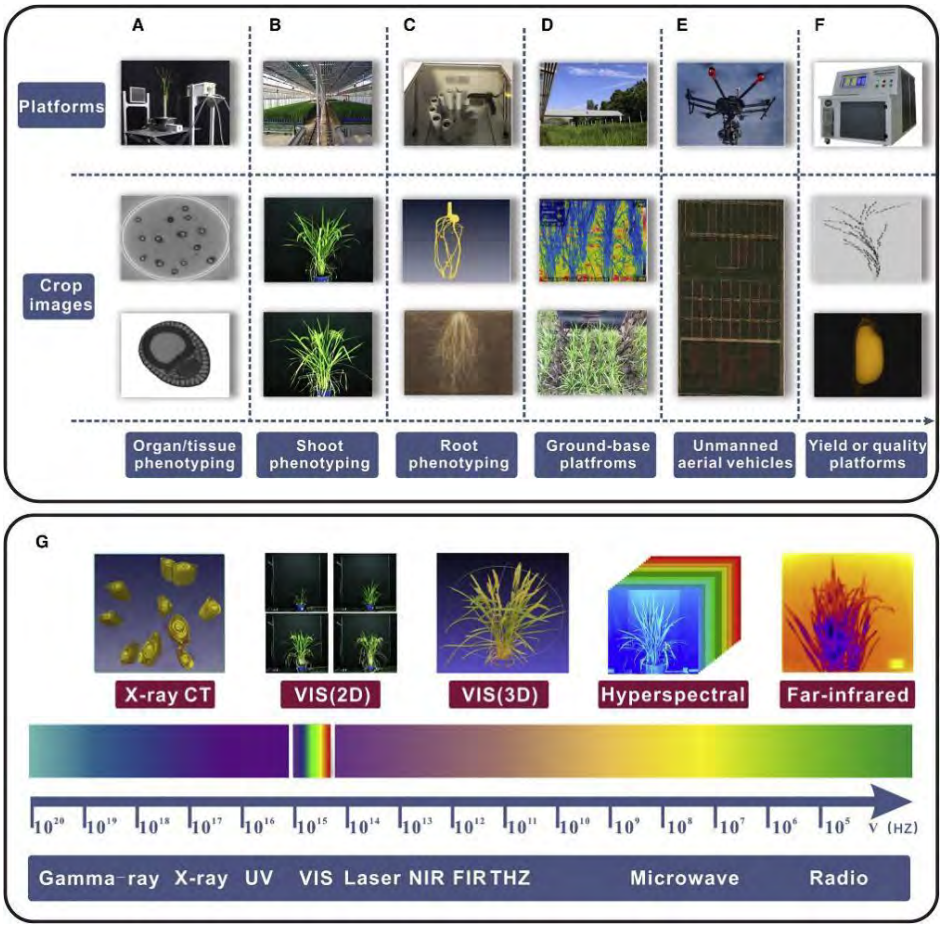


图12.20　作物中不同尺度下的表型获取平台概览（引自Yang et al.， 2020）

CT．断层扫描；VIS．可见光；VIS（2D）．可见光2D 成像；VIS（3D）．可见光3D 成像；Hz．赫兹；UV．紫外光；NIR．近红外光谱；FIR．远红外光谱；THZ．太赫兹射线；V．频率

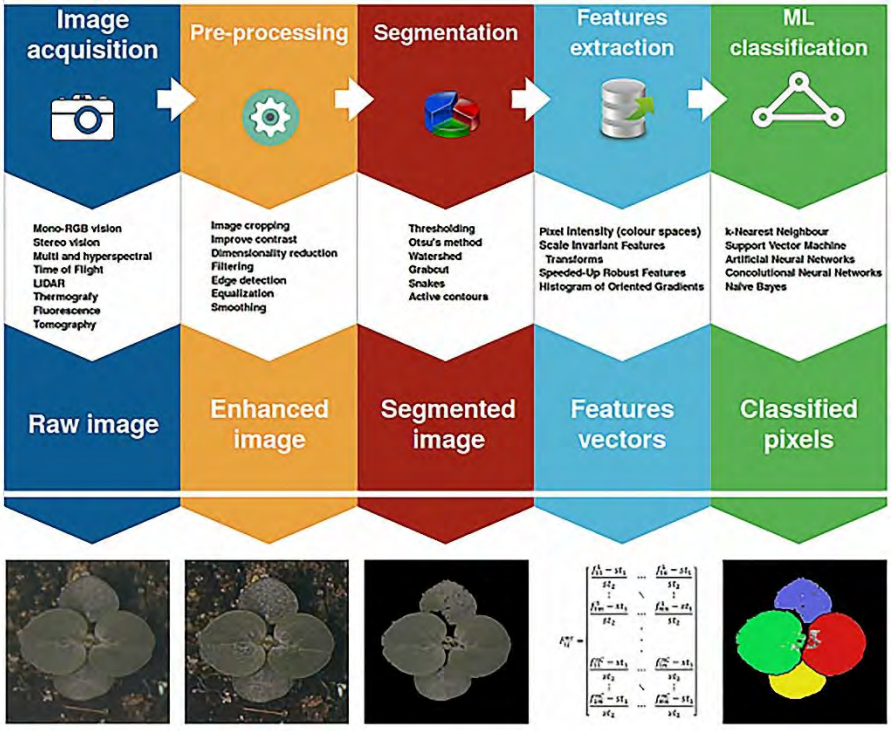


图12.21　图像识别分析流程（改自Perez-Sanz et al.， 2017）

上图：图像识别主要流程及其具体方法，即图像采集、预处理、分割、特征提取和机器学习分类；下图：以植物为例给出其表型鉴定过程，即从原始图像开始，依次获得强化图像、分割图像、特征向量和最终像素分类后的图像

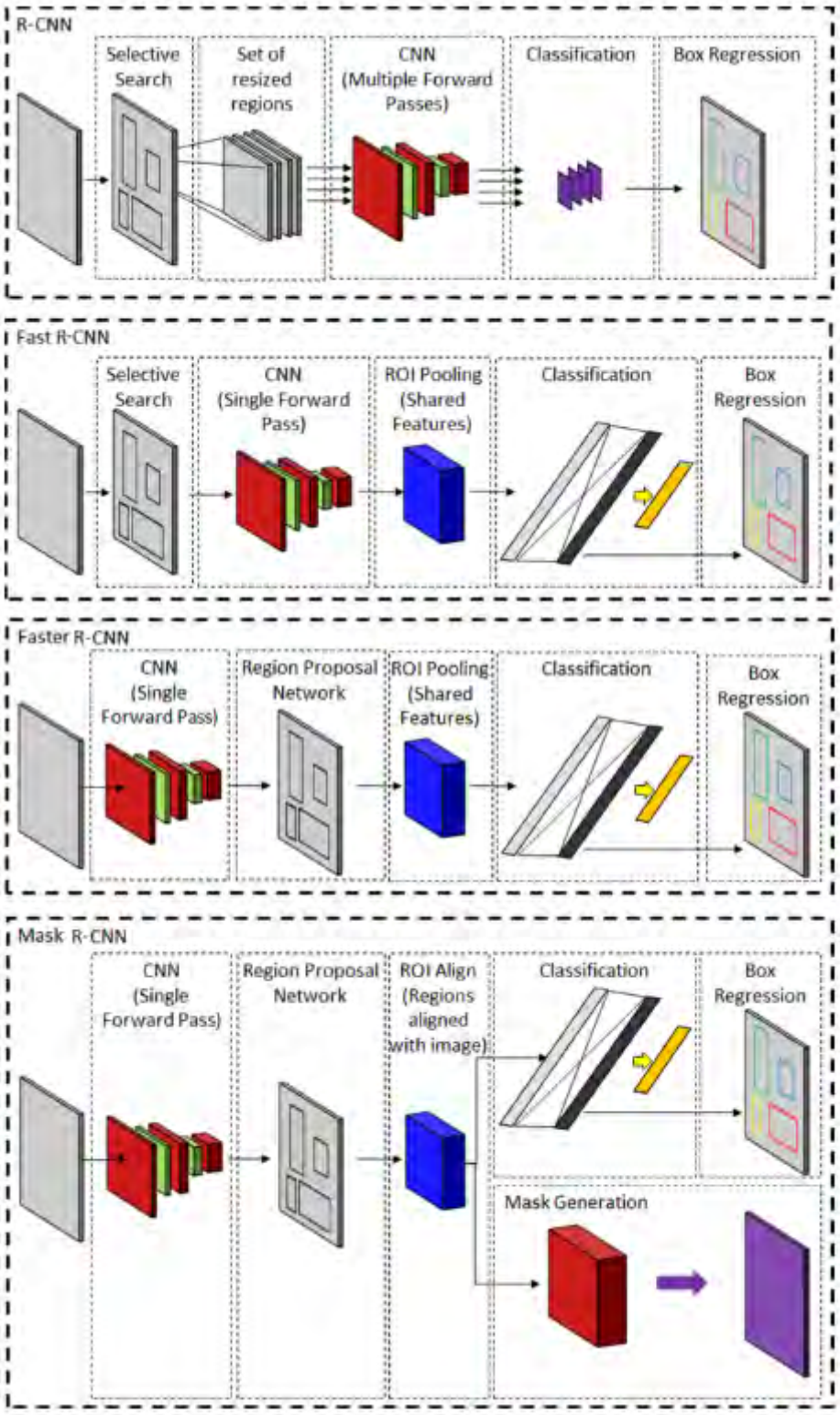


图12.22　基于R-CNN 发展的图像分割算法

（引自 Ghosh et al.， 2019）

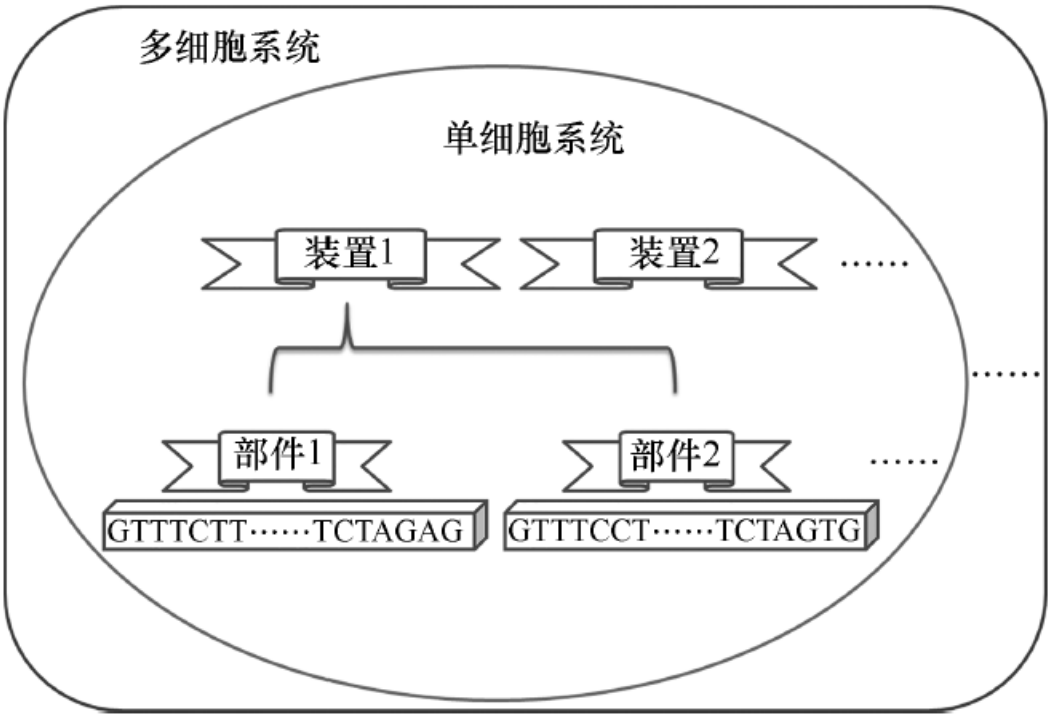


图 12.24　人工合成生物系统的层次化结构示意图

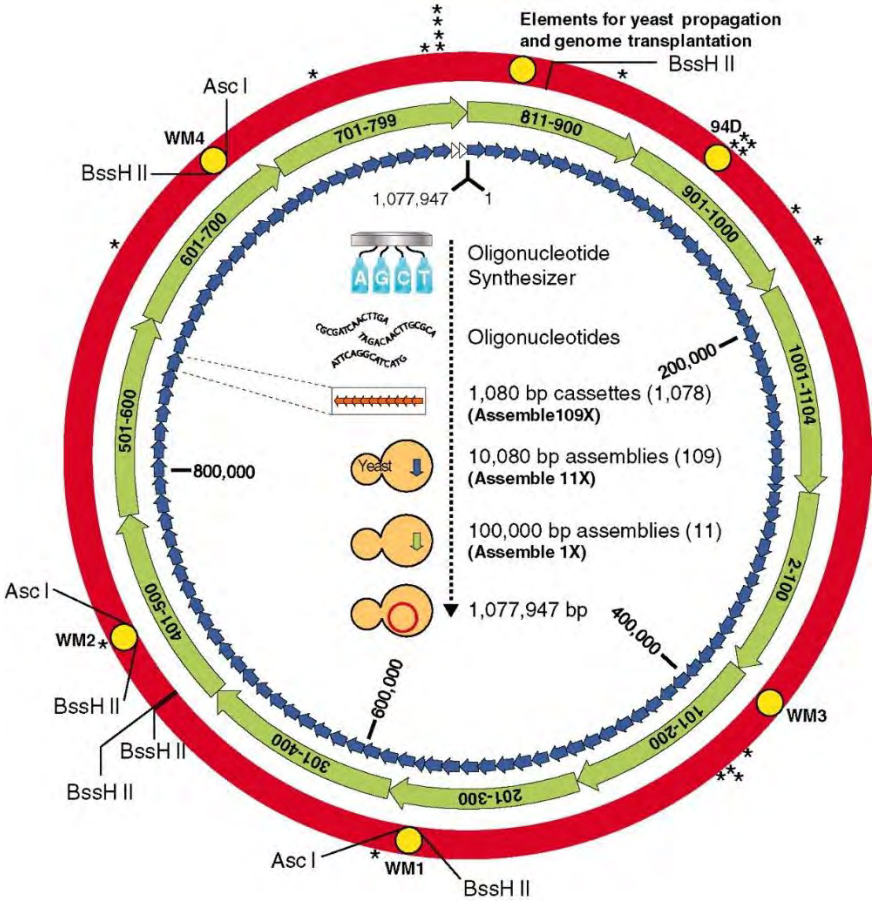


图12.25　*M. mycoides*人工基因组的设计与合成路线（引自Gibson et al.， 2010）

组装完成的人工基因组（红色外圈）与 *M. mycoides* 原始基因组唯一的不同是一些基因组经过刻意设计（外圈中黄点位置）：4 个加入的水印区域（WM1～4）、一个 4kb 删除区域（94D）和一个加入的酵母繁殖和基因组移植元件区域（EYPGT）。同时，人工基因组上也标出了内切酶（*Asc* Ⅰ和 *Bss*H Ⅱ）的酶切位点和存在遗传多态性的区域（星号）

图 12.26　在线植物合成生物学设计平台 GoldenBraid 主页

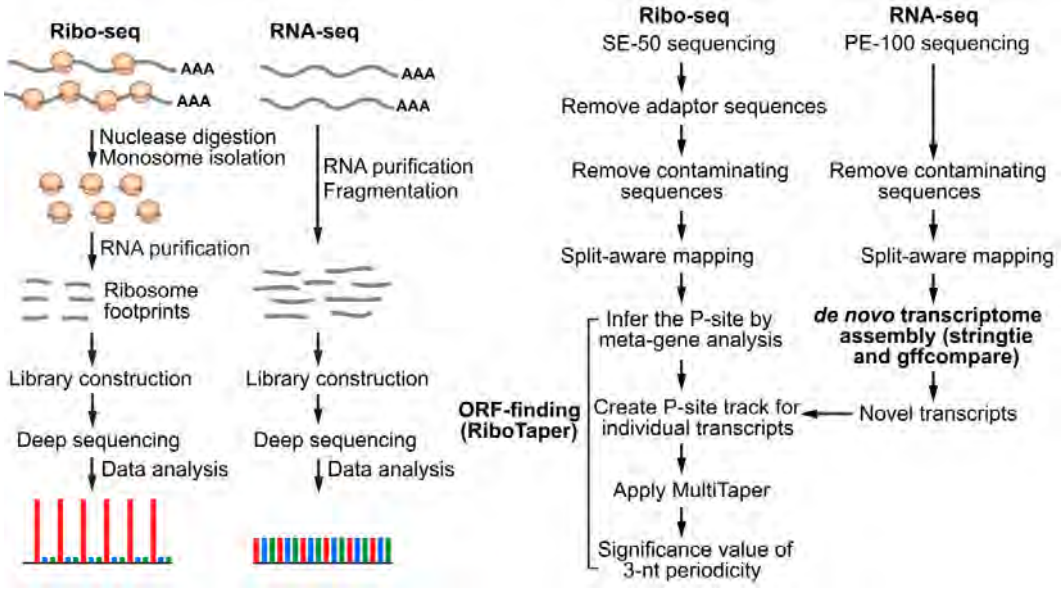


图 12.27　Ribo-Seq 与 RNA-Seq 的实验流程（A）和数据分析流程（B）比较（引自 Wu et al.，2019）

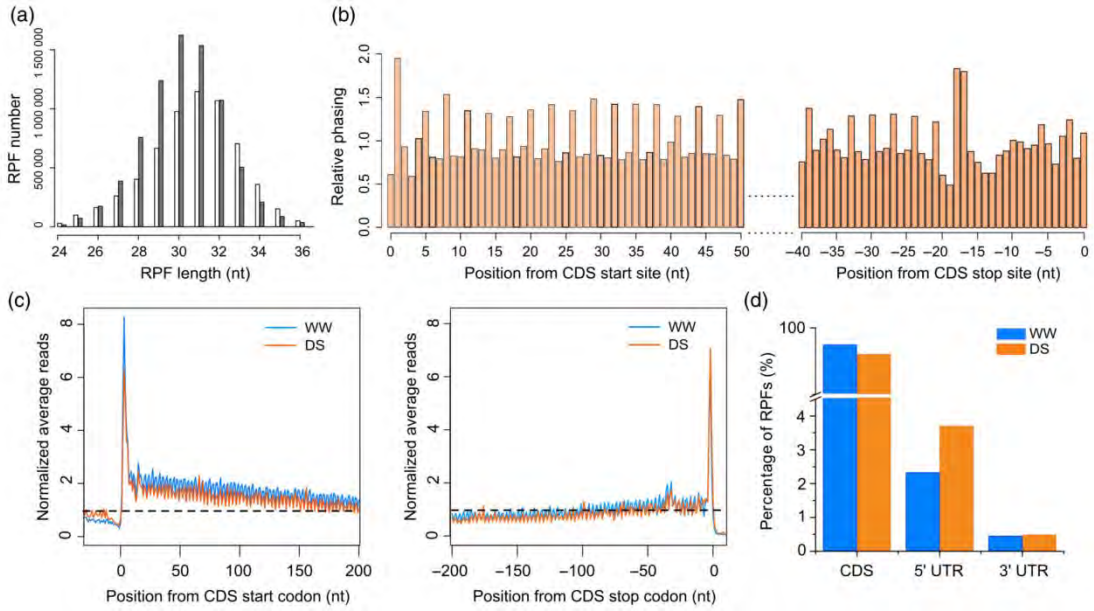


图12.28　玉米幼苗在正常（WW）和干旱（DS）条件下Ribo-Seq 数据的特征（引自Lei et al.，2015）

A．核糖体保护片段（RPF）的长度分布；B．编码序列（CDS）干旱条件下前50 个和后40 个碱基的三核苷酸周期性；C．CDS 起始和终止位置的 RPF 的密度分布；D．RPF 在 CDS、5′ UTR 和 3′ UTR 区域的读段分布

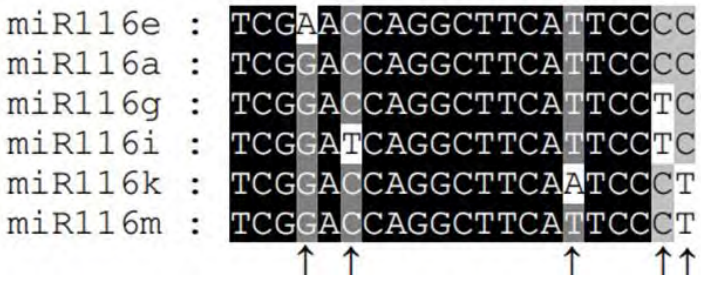
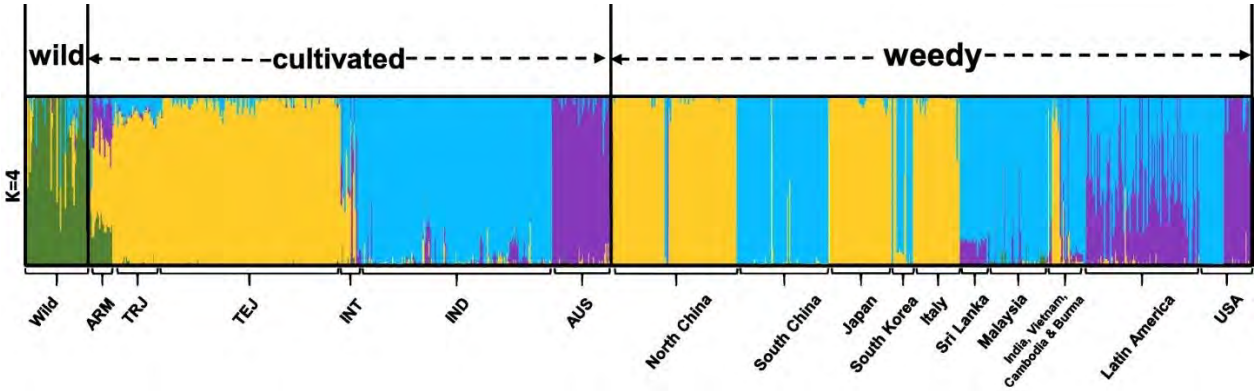


图13.1　群体个体序列联配结果及其遗传多态性

箭头所指位点存在遗传变异，为分离位点



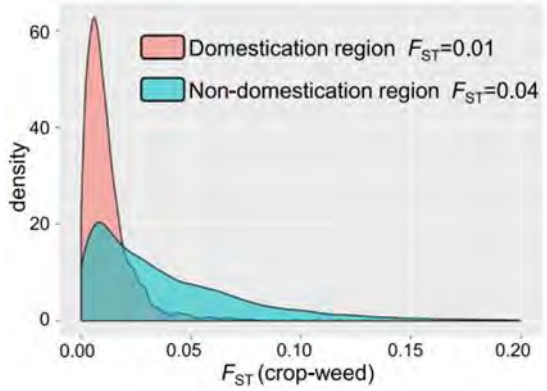


图13.2　水稻及其野化群体分析（引自Qiu et al.，2020）

A.　全球稻区杂草稻与当地栽培稻和野生稻群体结构分析结果，图中列出了最佳分组数量（K ＝4）下各个地区水稻及其野化群体情况。ARM. 香米；TRJ. 热带粳稻；TEJ. 温带粳稻；INT. 中间类型；IND. 籼稻；AUS. 秋稻。稻区依次为：华北、华南，以及日本、韩国、意大利、斯里兰卡、马来西亚、亚洲其他地区、拉丁美洲、美国。B. 基于驯化和非驯化基因区域的水稻和其野化群体分化情况

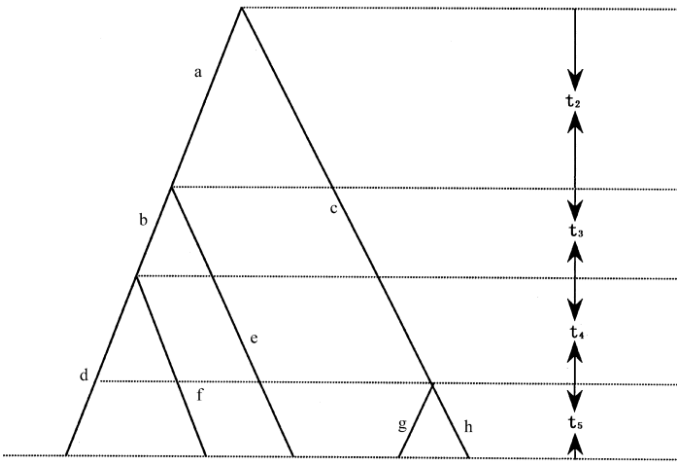


图13.3　由5条序列构建的系统发生树（引自周琦和王文，2004）

图中每一个结点代表两条 DNA 序列的共同祖先，由上至下意味着时间上的由古至今。t_m（m=2，…，5）代表由 m 条序列回溯至（m−1）条序列所需代数（generation time）

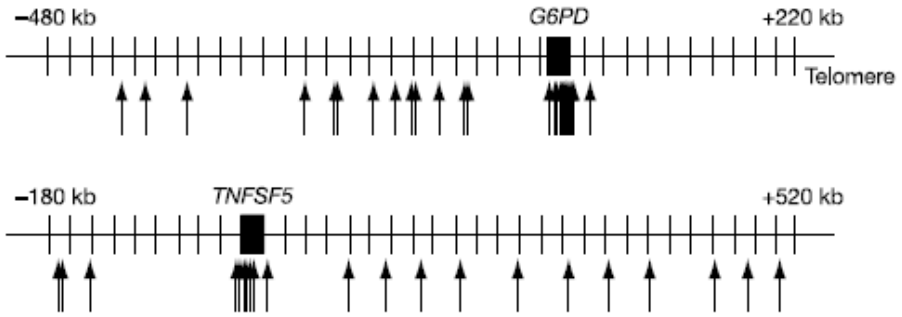


Figure 1 Experimental design of core and long-range SNPs for *G6PD* and *TNFSF5*. The core region is highlighted by a cluster of densely spaced SNPs (arrows) at the gene. Additional, widely separated flanking SNPs, used to examine the decay of LD from each core haplotype, are also shown. Markers distal to *G6PD* were within repetitive subtelomeric sequence and could not be genotyped.

图13.4　核心单倍体型举例（引自Sabeti et al.，2002）

图中列出了 *G6PD* 和 *TNFSF5* 两个核心单倍体型与周边 SNP（箭头所示）的情况

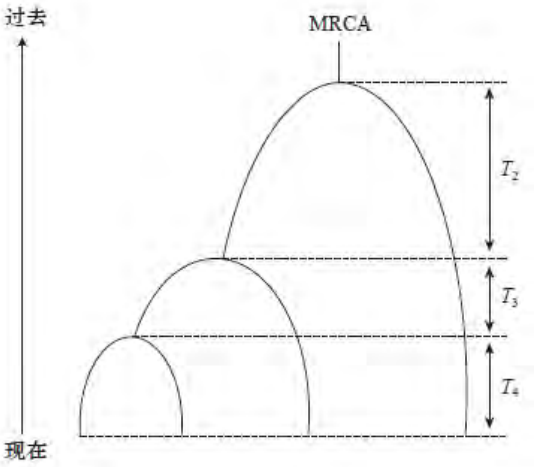


图13.5　4个个体溯祖过程（引自高峰和李海鹏，2016）

树的枝长 T_k 表示在当前有 k 个枝的情况下发生下一次溯祖事件的时间；MRCA.最近共同祖先

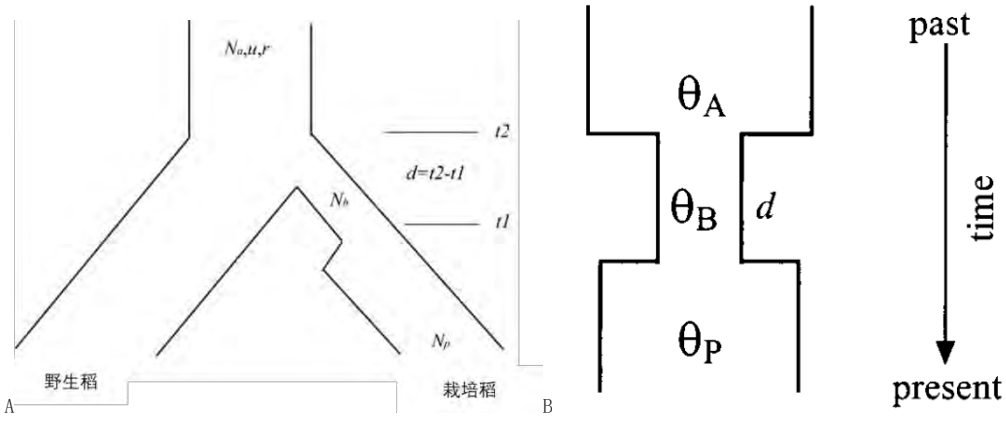


图13.6　用于作物人工选择溯祖模拟测验的进化模型举例

A.水稻：N_a 为野生稻有效群体大小；u 为碱基突变速率；r 为等位基因频率相关系数；经历一段时间（t_2-t_1）瓶颈效应后的群体大小为 N_b，而现在的栽培稻群体大小为 N_p（引自 Zhu et al.，2007）。B.玉米：θ_A、θ_B、θ_P 分别表示祖先群体、经历选择瓶颈效应的群体和当代群体；d 为瓶颈效应时间（引自 Eyre-Walker et al.， 1998）

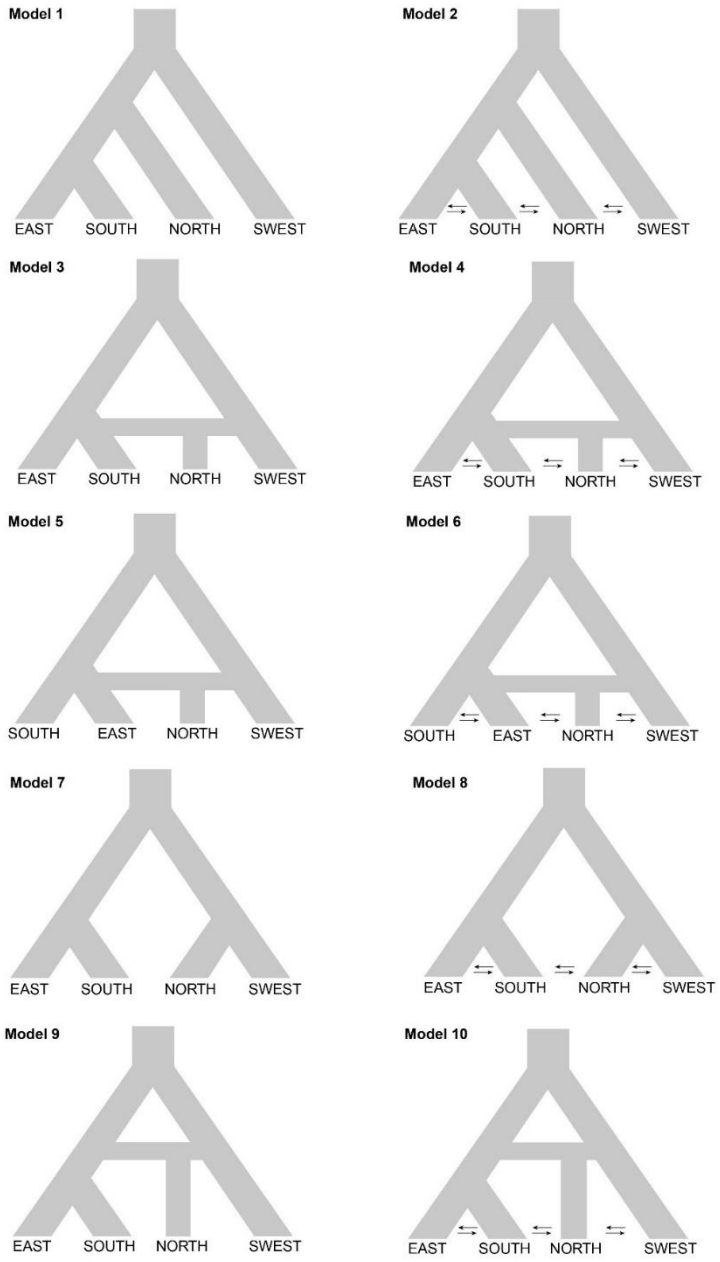


图13.7　银杏群体进化溯祖模拟测验预设的10 个起源模型（引自Zhao et al.， 2019）

共设置了 4 个谱系（东部、南部、北部和西南部）； 表示谱系间基因流

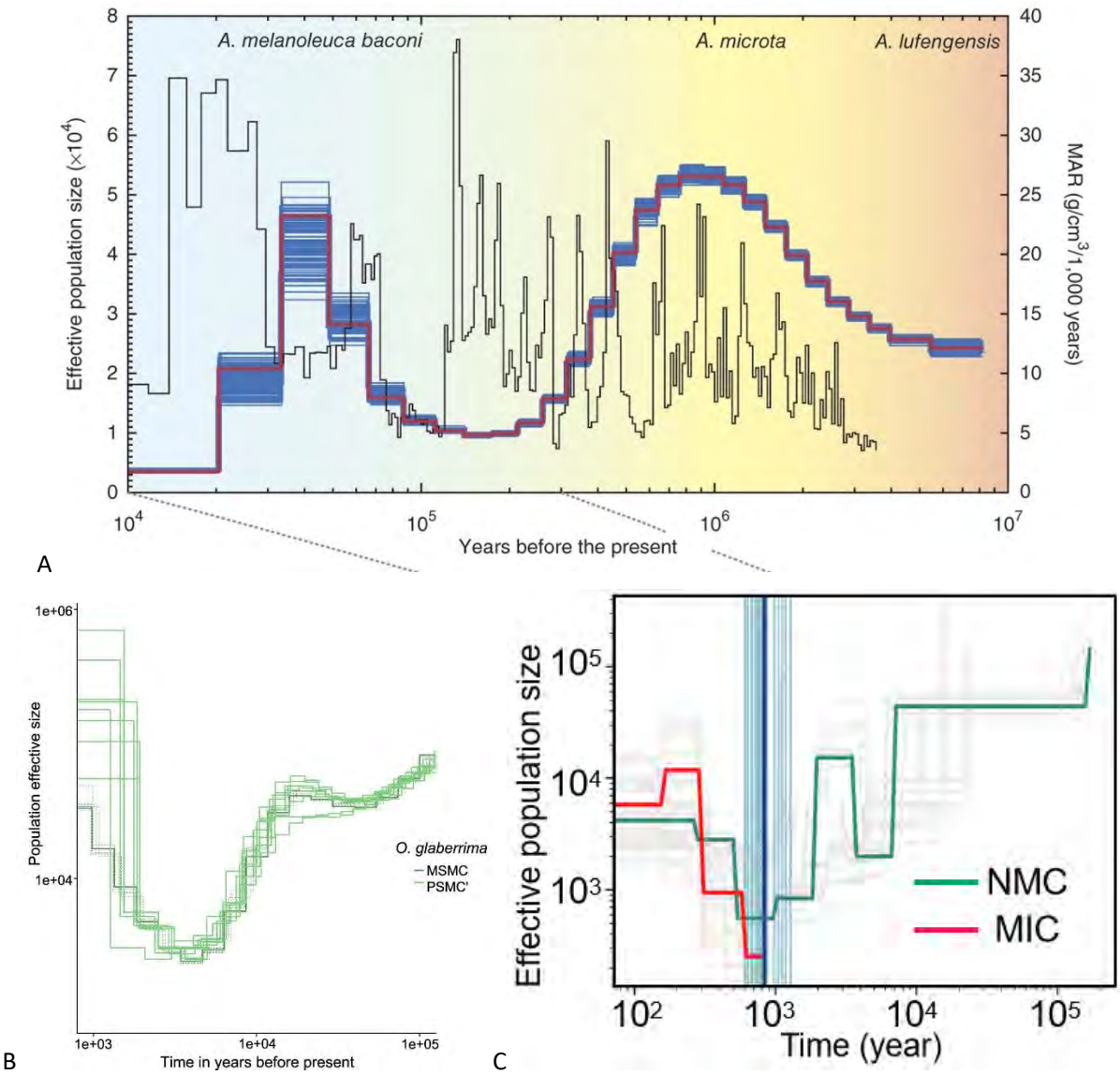


图13.8　基于溯祖理论进行有效种群大小估计举例

A.利用 PSMC 推测大熊猫有效种群大小历史动态（引自 Zhao et al.， 2013）。图中拉丁名从左到右依次表示巴氏大熊猫、大熊猫小种和始熊猫；B.利用 MSMC 和 PSMC 两种方法推测非洲栽培稻（*O. glaberrima*）的有效种群大小变化（引自 Philippeet al.， 2018）；C.利用 SMC＋＋推测长江流域稻田拟态稗草（MIC）和非拟态稗草（NMC）的有效群体大小变化（引自 Yeet al.， 2019）。时间估计（横坐标）基于繁殖代数（每代 ＝1 年）

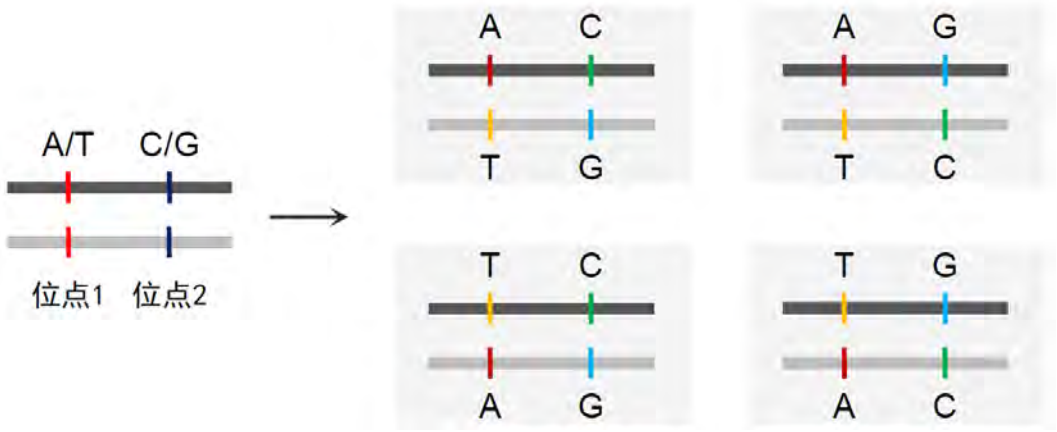


图13.9　基因组杂合位点的基因型定相

两个杂合位点存在 4 种可能的基因型组合

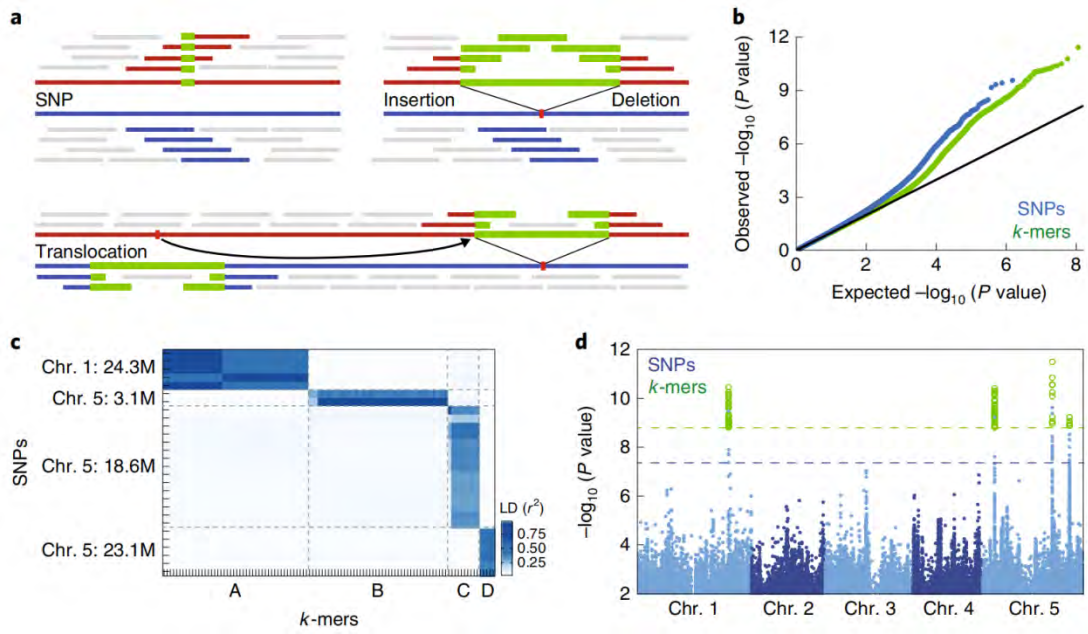


图13.10　基于SNP和*K*-mer序列的GWAS分析结果比较——以拟南芥花期为例（引自Voichek and Weigel，2020）

A. *K*-mer 与其他遗传变异。红和蓝长线表示两个不同个体基因组，彩色短线表示各个单个基因组特有的 *K*-mer 序列，灰色短线表示共有 *K*-mer 序列；B. *K*-mer 与 SNP 的 Q-Q 曲线分布图比较；C、D. 基于 *K*-mer 与 SNP 获得的花期 GWAS 分析候选关联位点比较

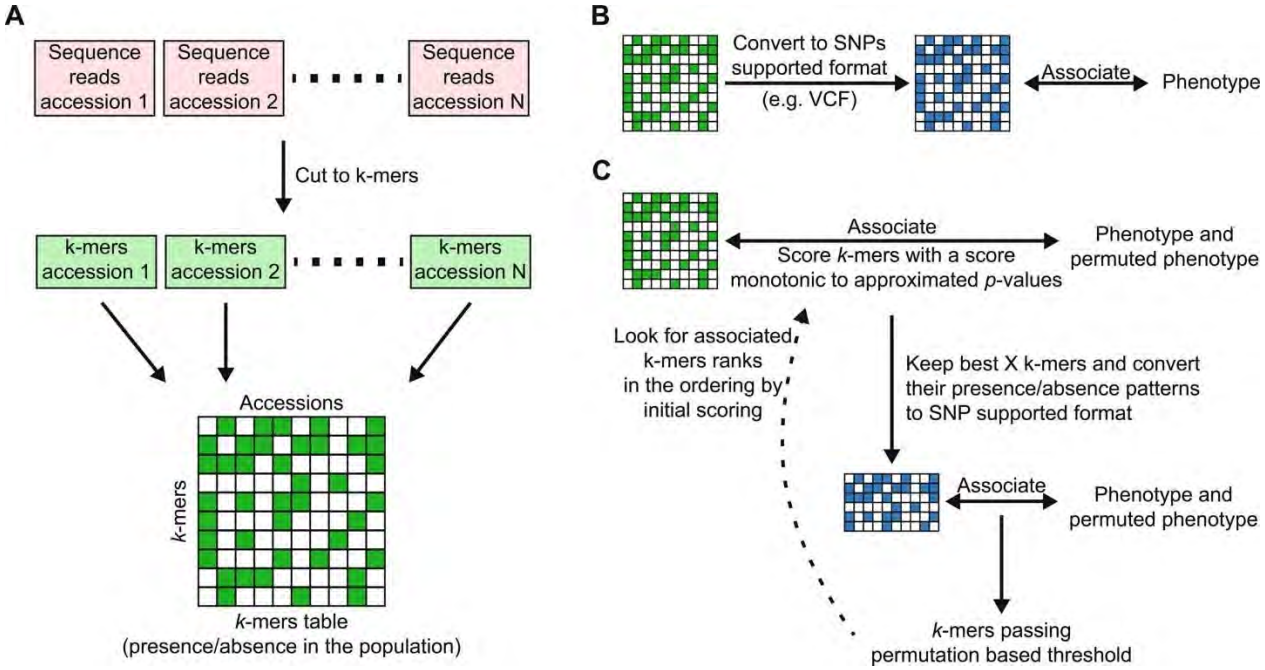


图13.11　基于*K*-mer 序列的GWAS 分析流程（引自Voichek and Weigel， 2020）

A.　*K*-mer 序列变异检测；B、C. 基于 *K*-mer 序列的 GWAS 分析和优化过程

A

normal.ped

1	1	0	0	1	1	A	A	G	T
2	1	0	0	1	1	A	C	T	G
3	1	0	0	1	1	C	C	G	G
4	1	0	0	1	2	A	C	T	T
5	1	0	0	1	2	C	C	G	T
6	1	0	0	1	2	C	C	T	T

normal.map

1	snp1	0	5000650
1	snp2	0	5000830

B

trans.tped

1	snp1	0	5000650	A	A	A	C	C	C	A	C	C	C	C	C
1	snp2	0	5000830	G	T	G	T	G	G	T	T	G	T	T	T

trans.tfam

1	1	0	0	1	1
2	1	0	0	1	1
3	1	0	0	1	1
4	1	0	0	1	2
5	1	0	0	1	2
6	1	0	0	1	2

图 13.12　常见的两种 Plink 数据格式

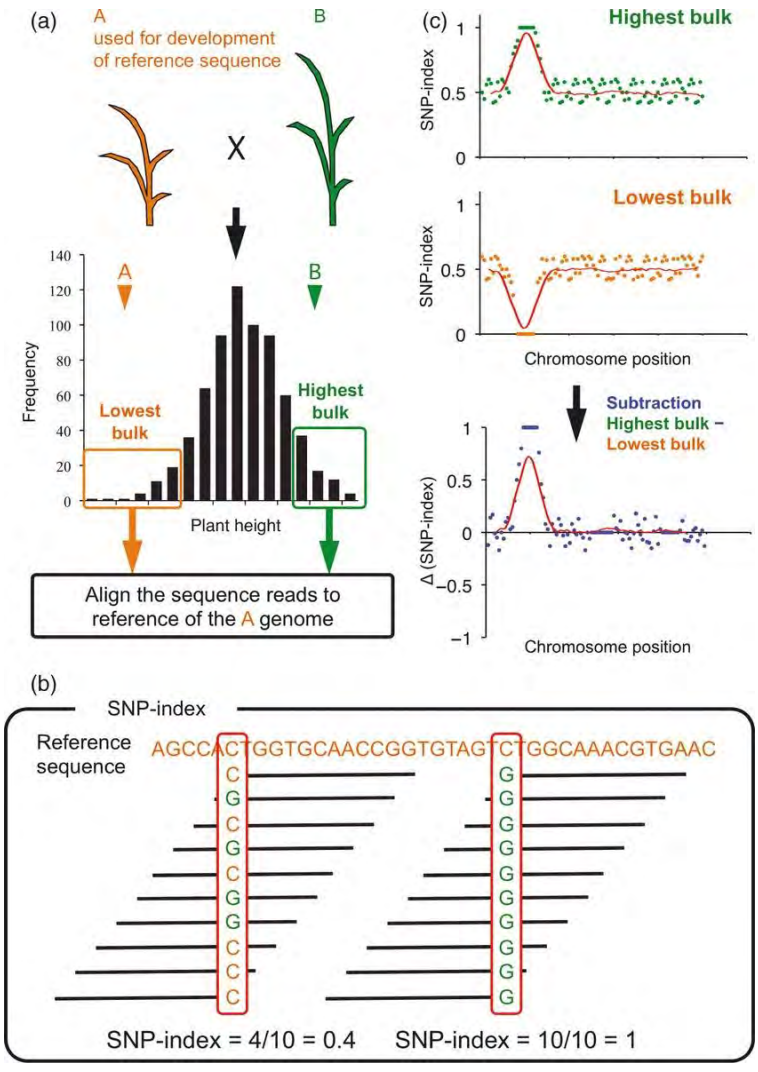


图13.13　BSA定位工具QTL-Seq的主要技术流程（引自Takagi et al.，2013）

A.定位材料的选择及其性状（株高为例）分离情况，其中材料 a 基因组作为参照基因组用于后续 SNP 分析；B.性状分离群体（矮或高株池）在基因组不同位点 SNP 指数（SNP-index）变化情况；C. SNP 指数在不同性状分离群体染色体上的变化情况和候选位点（极端差值区域）

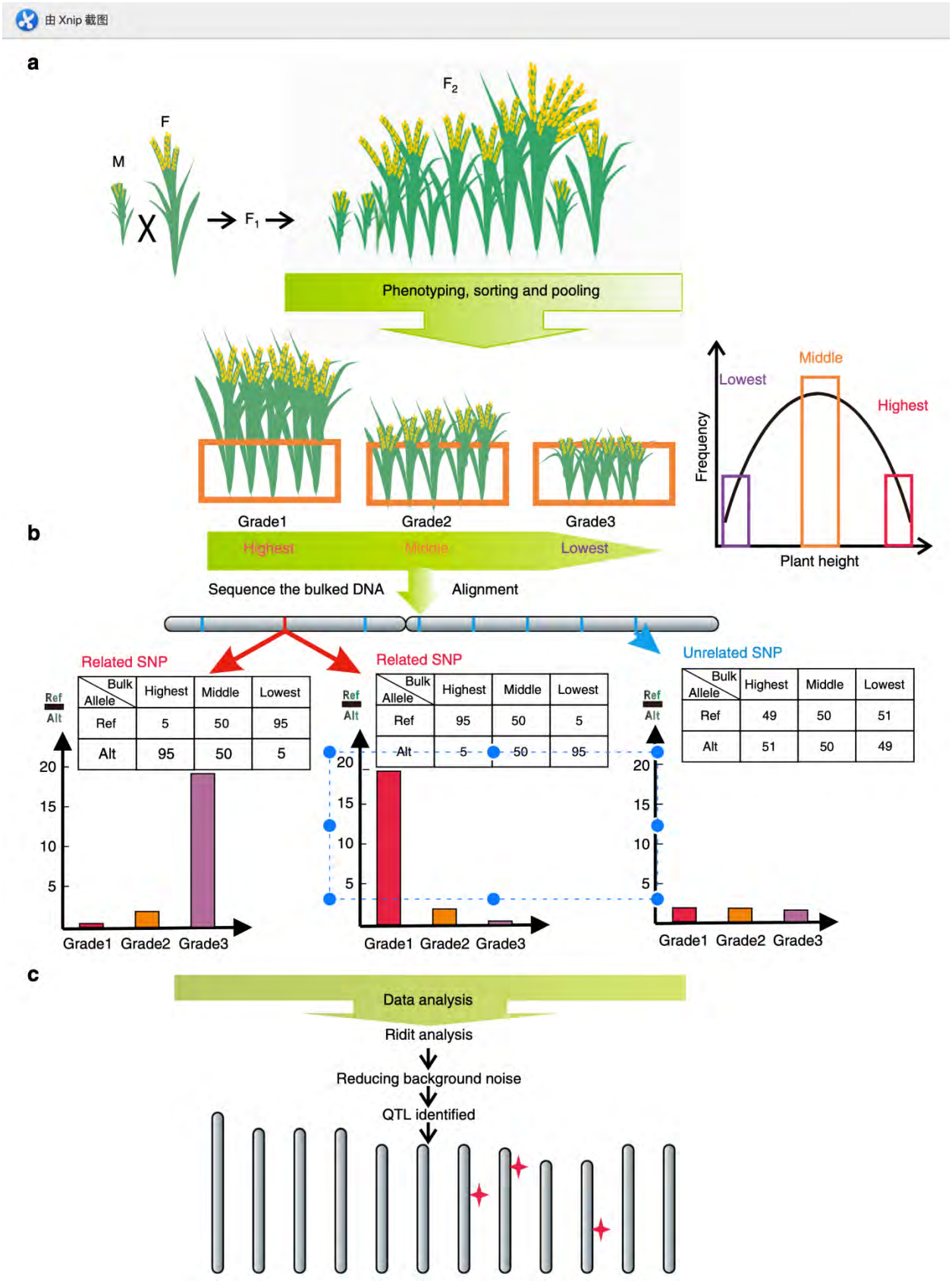


图13.14　GradedPool-Seq 技术路线（引自Wang et al.，2019）

A. 定位群体材料的构建及其性状（株高）分离情况，由此可以获得不同等级目标性状的亚群体。B. 不同性状群体材料测序及其不同等位基因频率情况。基于100× 测序深度，如果某一遗传变异（SNP）与性状无关，则其定位到参照基因组上的两个基因型读序平均各 50%（不相关 SNP），否则表现出明显差异（相关 SNP）。图中给了三个 SNP 案例（2 个性状相关，1 个性状不相关），表中列出了具有参考基因组相同碱基序列（Ref）和变异（Alt）碱基序列的个体数量。C. 进一步数据分析，包括序数 Ridit 分析和背景降噪算法，最终确定候选 QTL 位点

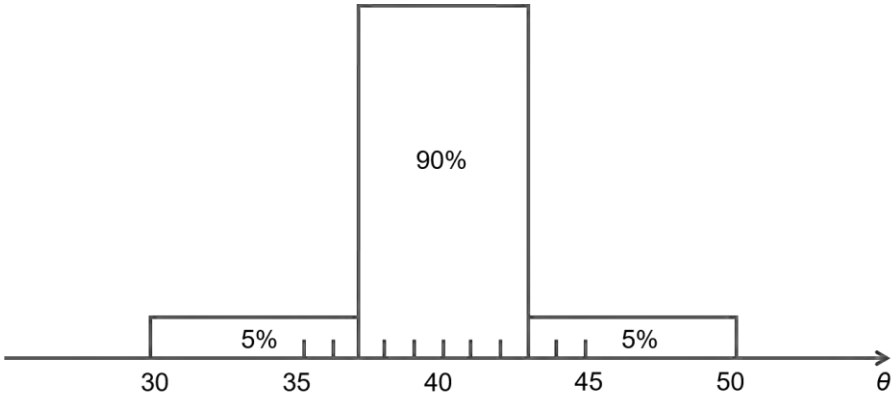


图 14.1　学生判断新教师年龄的先验分布

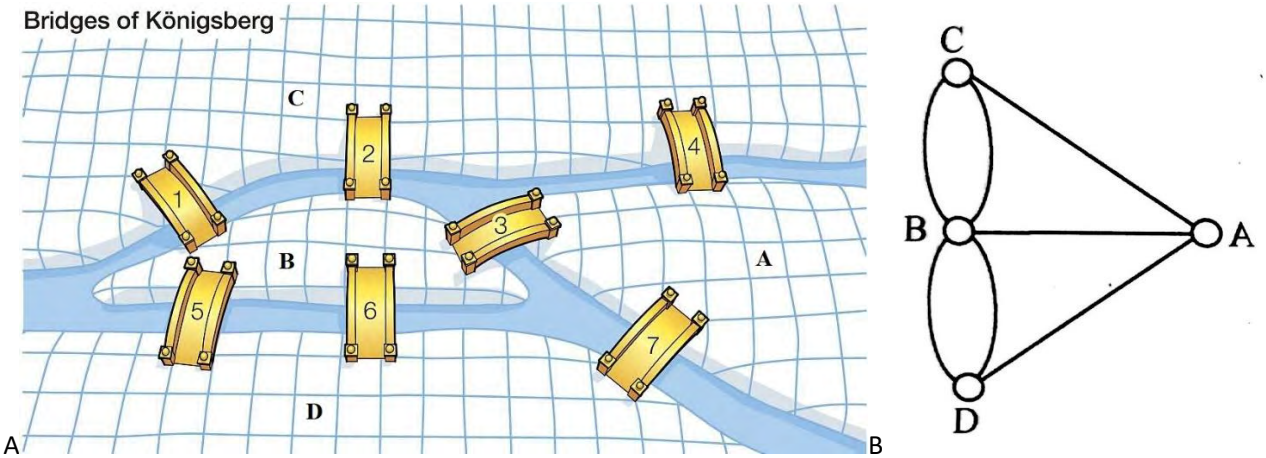


图 14.2　哥尼斯堡七桥问题（A）及其抽象图解（B）

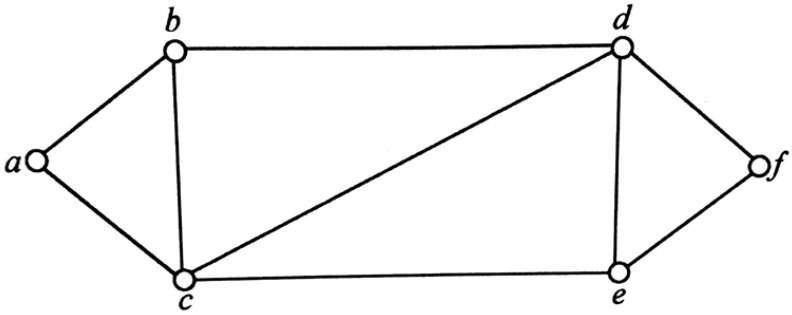


图 14.3　一个具有 6 个顶点的连通图

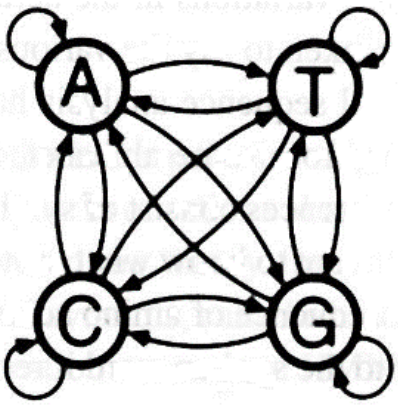


图 14.4　DNA 序列马尔可夫模型

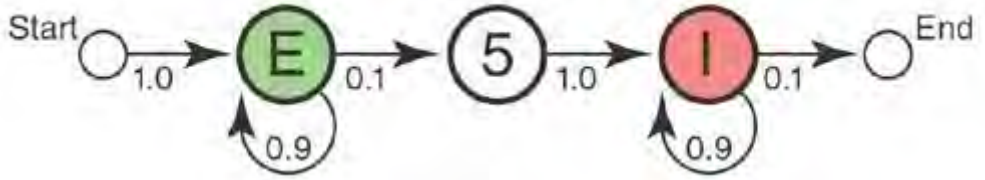


图14.5　外显子（E）与内含子（I）剪接位点（5′端）马尔可夫模型（引自Eddy， 2004b）

图中数字表示转移概率

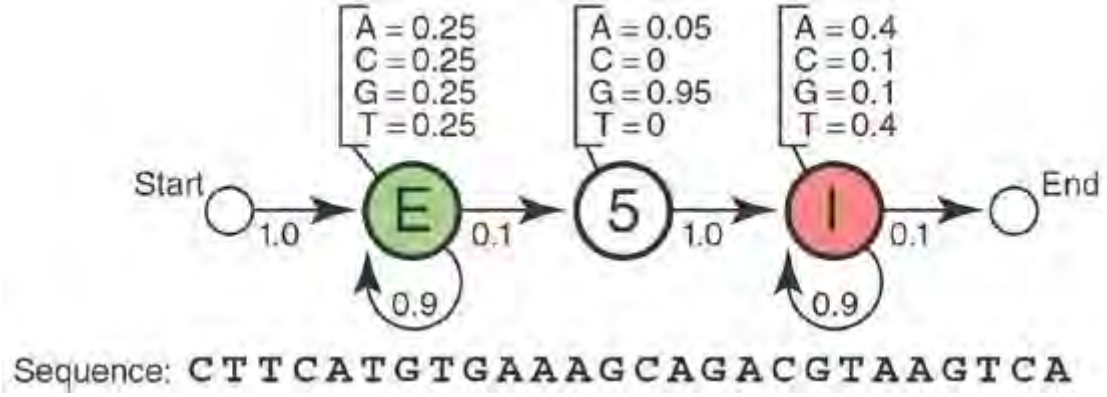


图 14.6　外显子与内含子剪接位点（5′ 端） HMM 模型（引自 Eddy， 2004b）

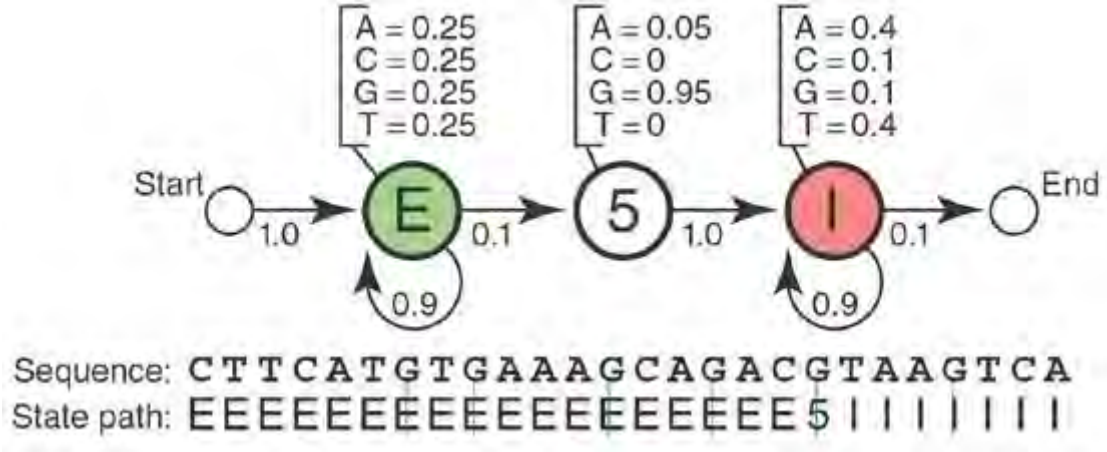


图 14.7　图 14.6 中 HMM 模型的一条马尔可夫“态”链（即一种剪接方式）

图14.8　生物序列的k 阶马尔可夫模型

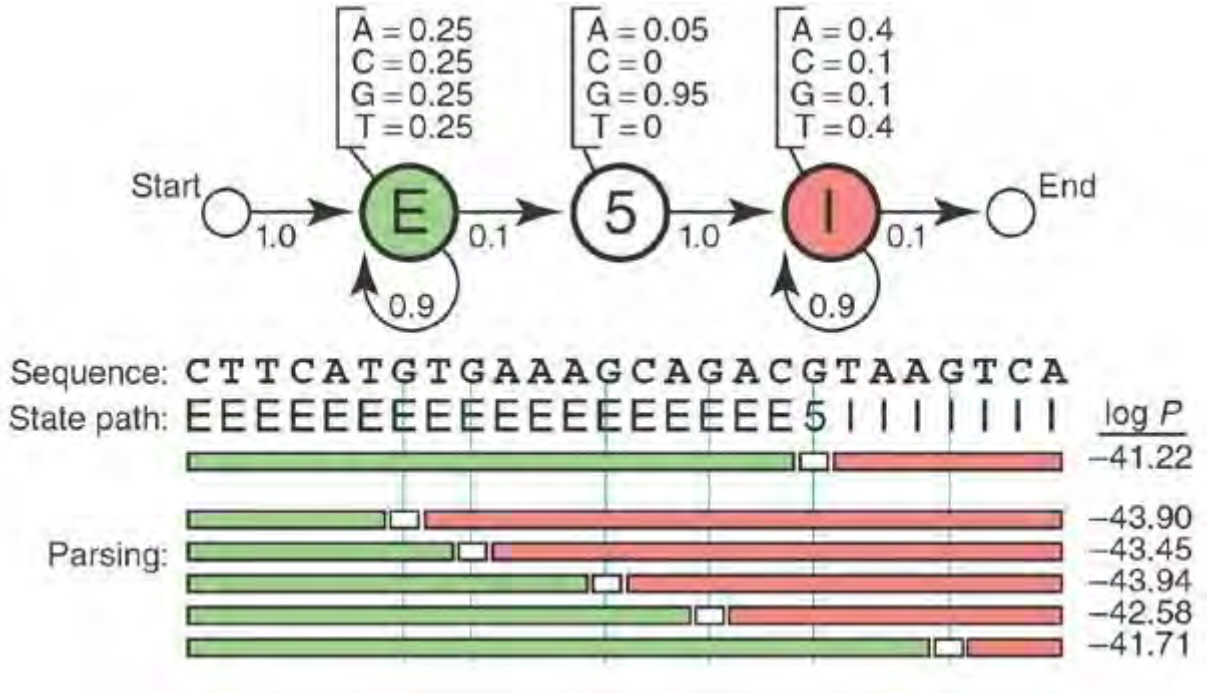


图 14.9　根据给定序列和马尔可夫链状态路径预测基因结构

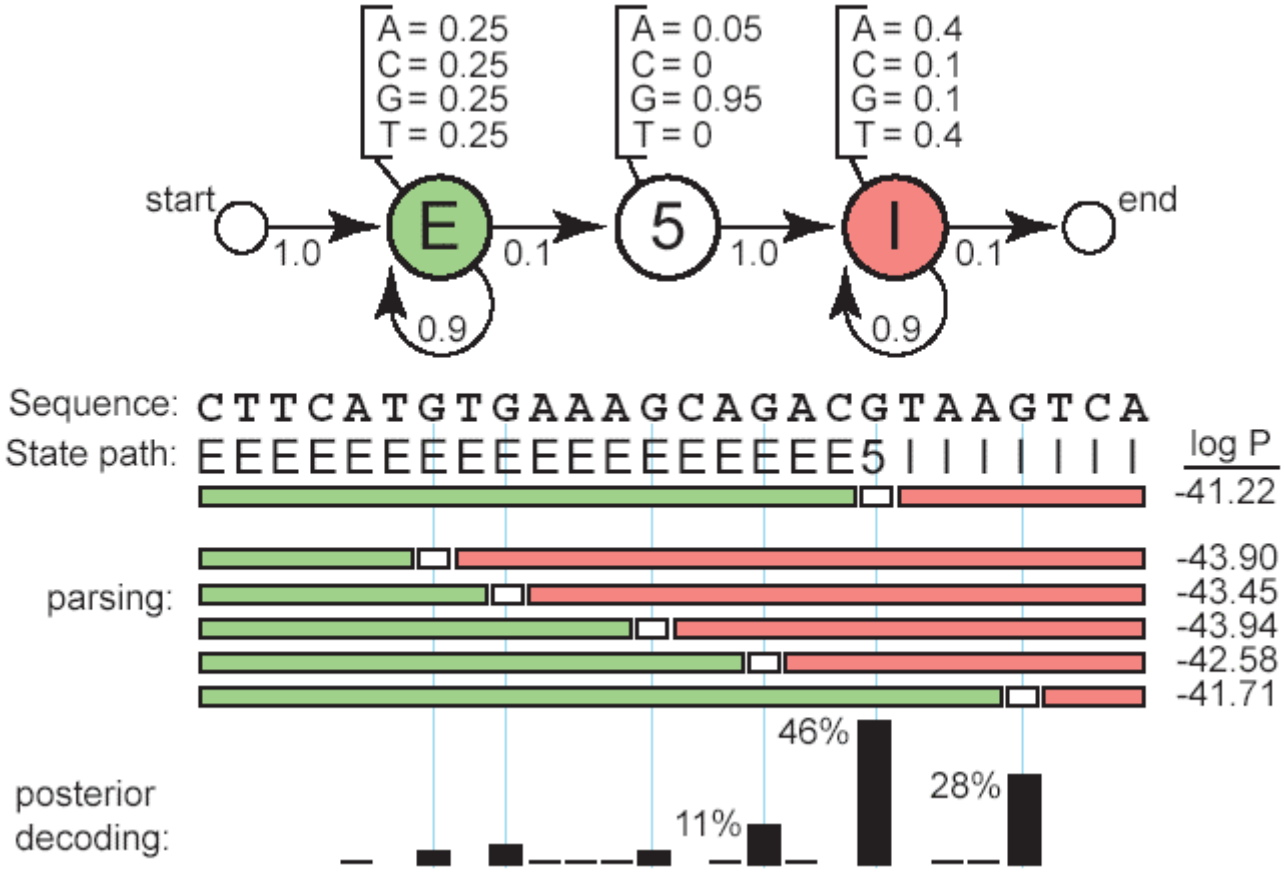


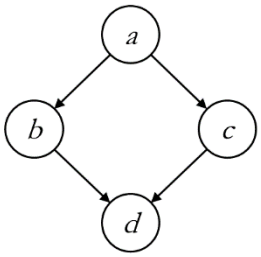
图 14.10　HMM 模型应用贝叶斯统计案例（引自 Eddy，2004b）

$P(a)$:

$P(a=U)$	$P(a=D)$
0.7	0.3

$P(b|a)$:

a	$P(b=U)$	$P(b=D)$
U	0.80	0.20
D	0.50	0.50



$P(c|a)$:

a	$P(c=U)$	$P(c=D)$
U	0.60	0.40
D	0.90	0.10

$P(d|b,c)$:

b	c	$P(b=U)$	$P(b=D)$
U	U	1.00	0.00
U	D	0.70	0.30
D	U	0.60	0.40
D	D	0.50	0.50

图14.11　一个包括4个变量或节点（a～d）的贝叶斯网络（改自林标扬，2012）

U 和 D 表示两种不同的状态

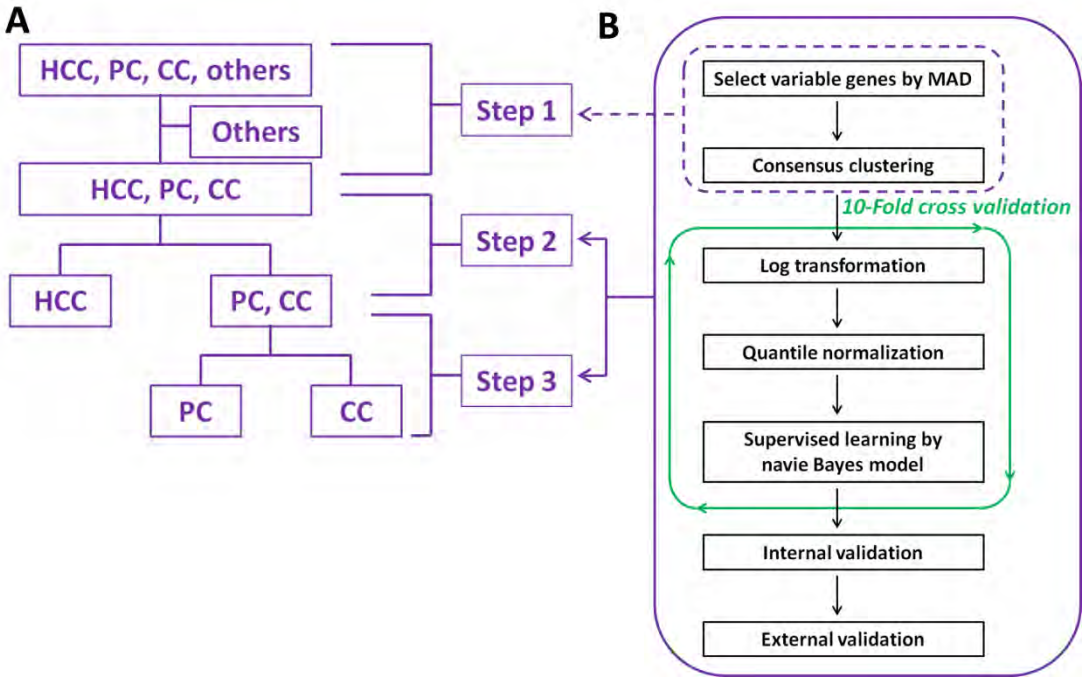


图14.12　利用朴素贝叶斯分类算法诊断多灶性肝胆胰肿瘤组织起源算法流程图（引自Jiang et al.， 2017）

HCC.肝细胞癌；PC.胰腺癌；CC.胆管癌；MAD.绝对中位差

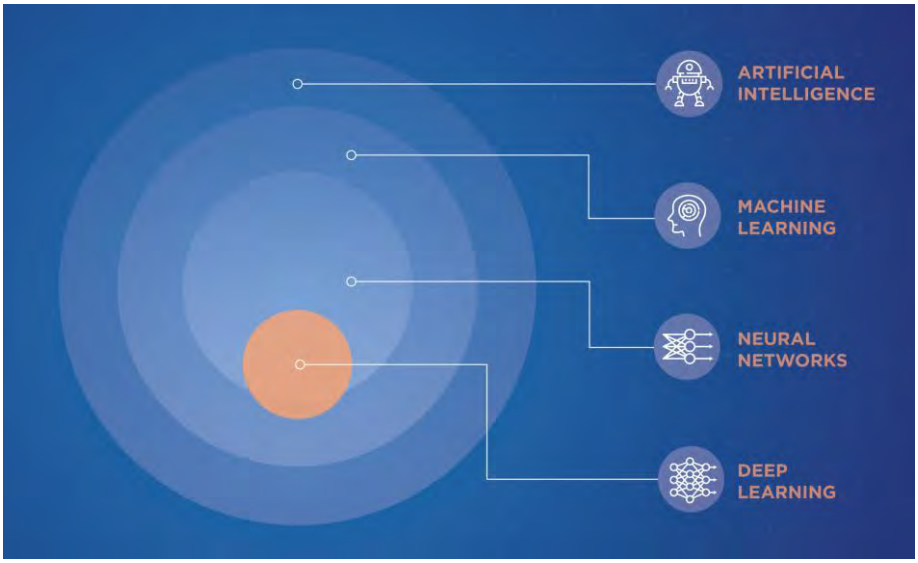
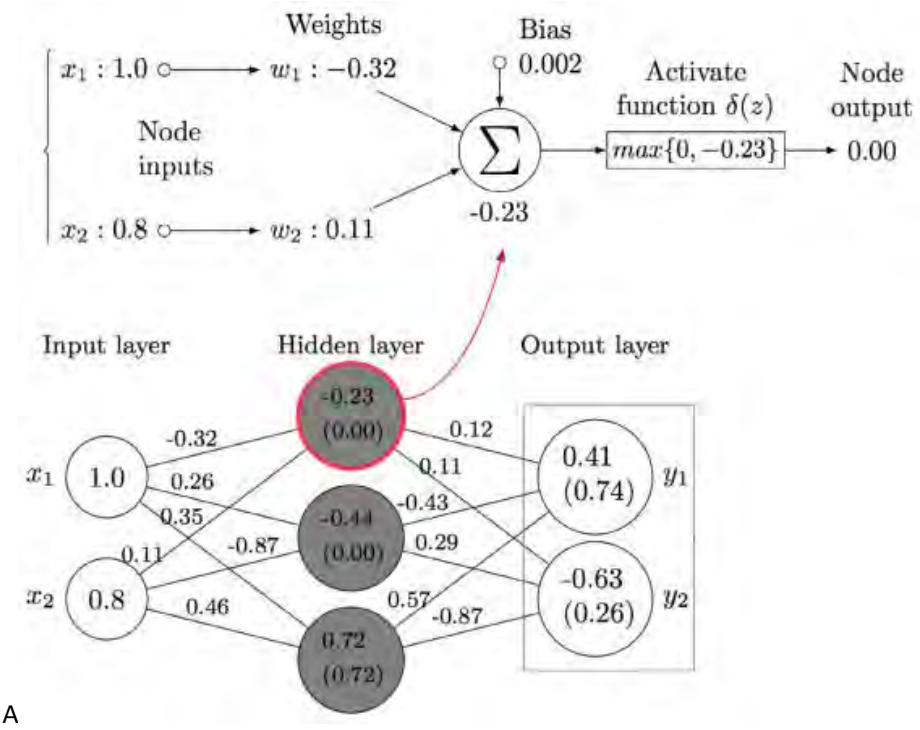
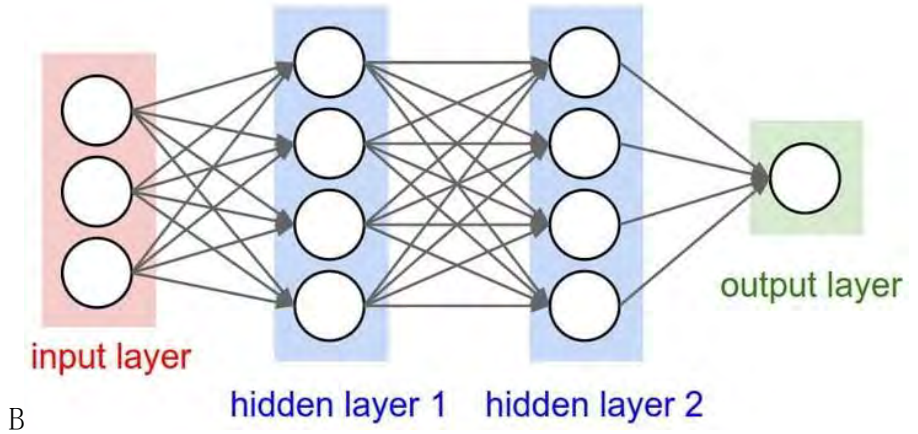


图 14.13　人工神经网络与人工智能、机器学习和深度学习的关系（引自 Akst，2019）



A



B

图 14.14　双层神经网络模型（A）和多层神经网络（B）举例（引自 Li et al.， 2019b）

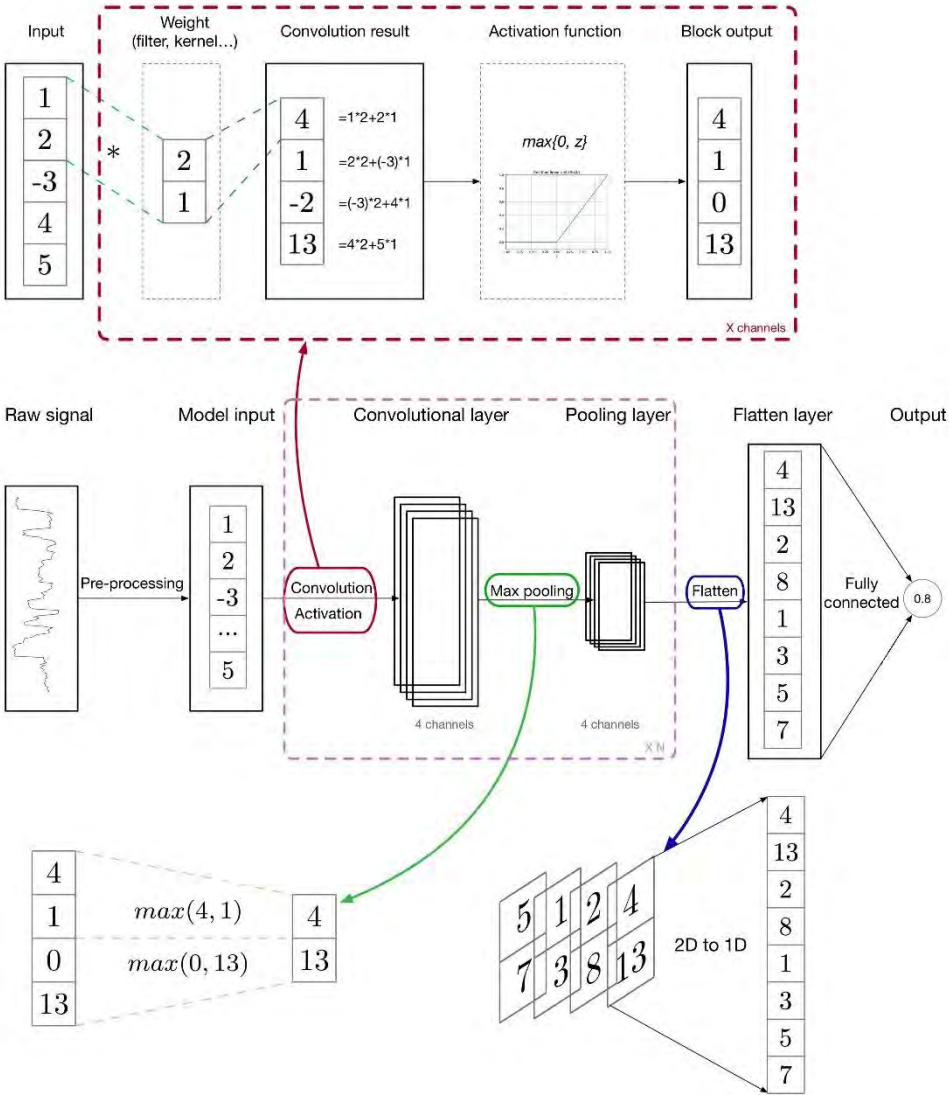


图 14.15　卷积神经网络示意图及其数据分析流程举例（引自 Li et al.， 2019b）

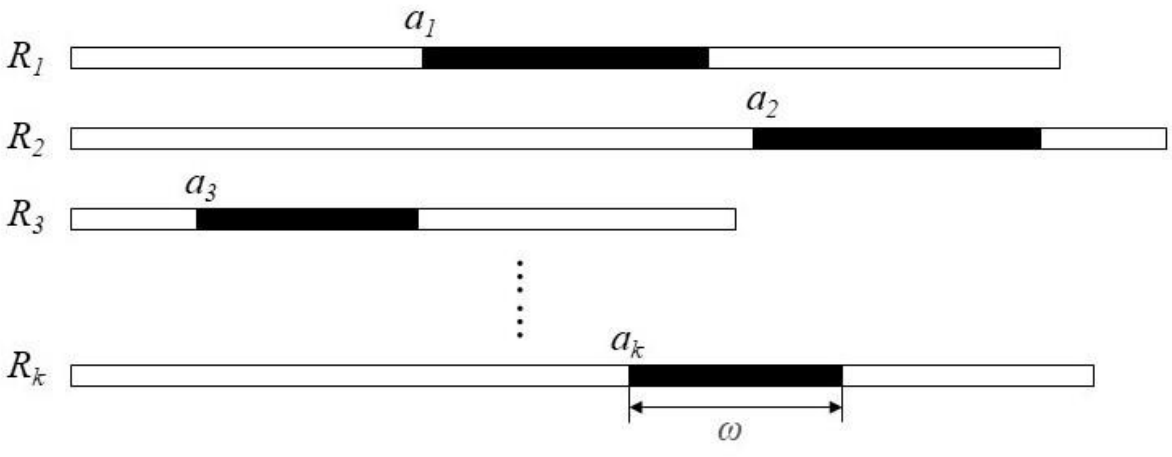


图14.16　在多重序列（R）中找基序位置的示意图

每条序列中加黑的片段（宽度 ω）表示这些序列期待的“共同模式”（基序）；它的位置（a）和内容都是未知的

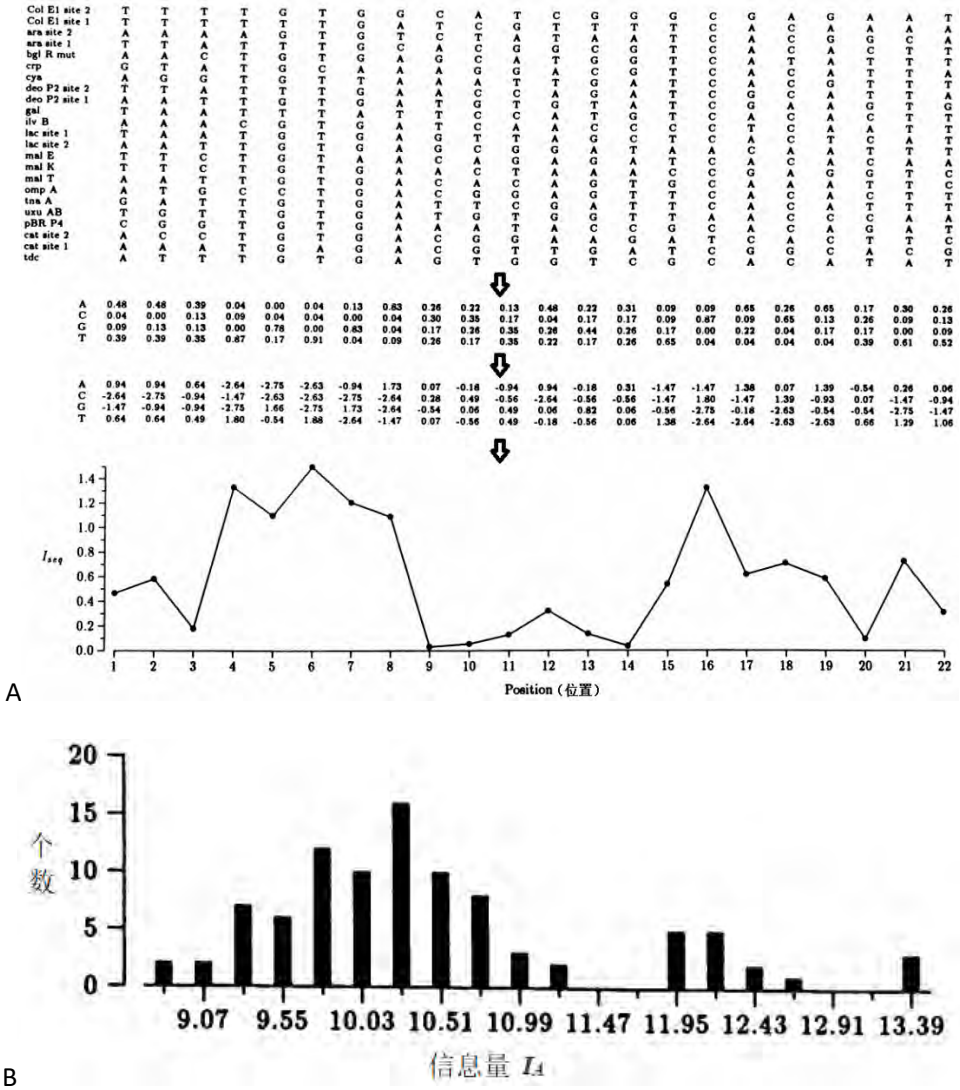


图14.17　多序列中寻找基序算法举例（引自Stormo and Hartzell， 1989）

A.来自 *E. coli* 的 18 个基因上游非编码区 DNA 序列的一个联配结果（总 22 列），图中列出了基于该联配结果的每列碱基构成比例、对数转换（$\log_2$）数值和信息量；B. 基于 Stormo-Hartzell 算法获得的每个 PSSM 信息量数值分布图

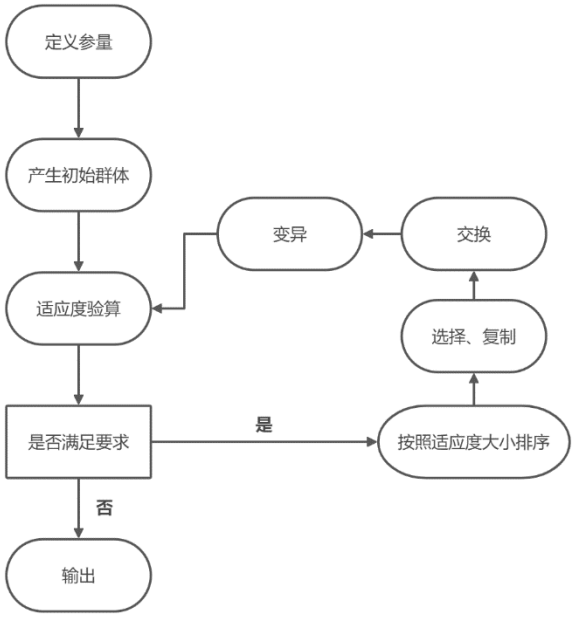
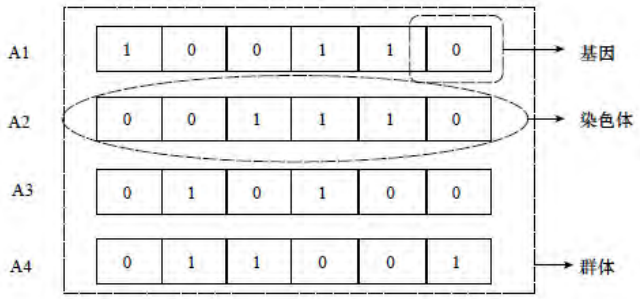


图 14.18　遗传算法基本运算流程图



染色体	生存值	占比/%
A1	28	28.9
A2	23	23.7
A3	12	12.4
A4	34	35.1

图 14.19　背包问题二进制编码的初始群体

图中给出了基因、染色体和群体的具体定义，其中涉及 4 个个体（即染色体）背包中的物品（即基因）情况、每个个体生存值及其在群体中所占比例

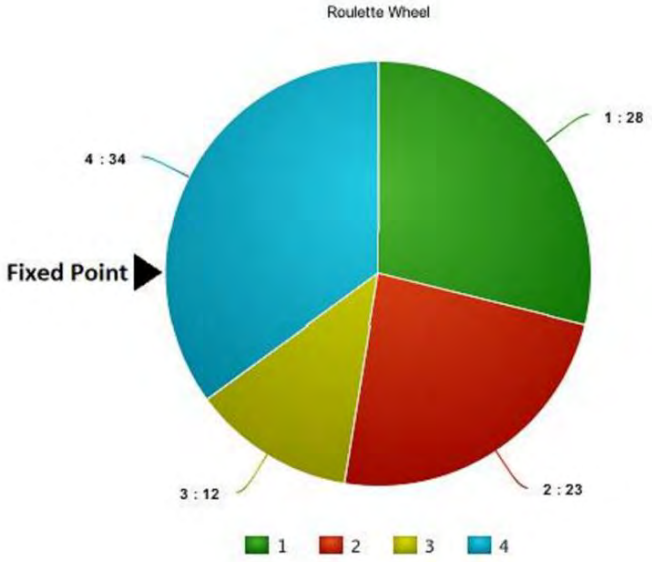


图 14.20　轮盘形式随机选择父母本　以图 14.19 背包问题为例，涉及 4 条染色体（A1～A4）

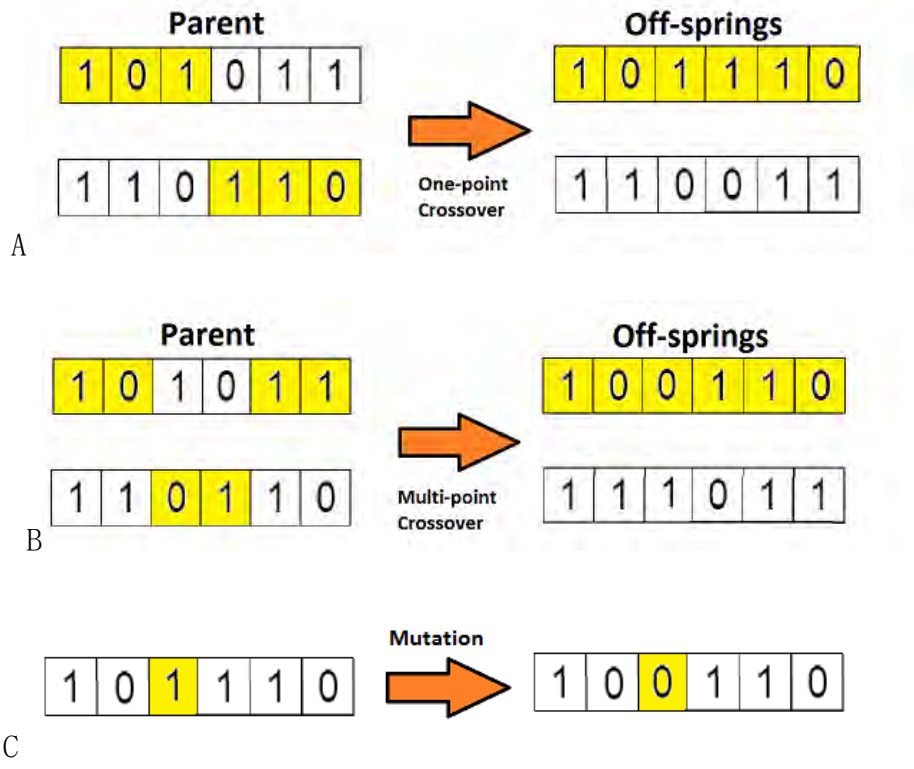


图 14.21　个体间基因交换（单点和多点交换）（A、B）和变异（突变）（C）导致的后代染色体变化

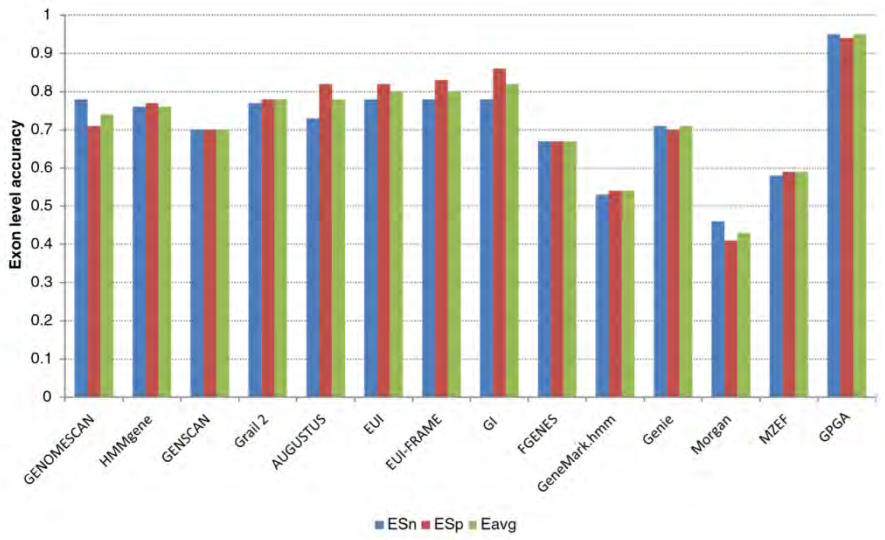


图14.22　GPGA 与其他基因预测工具在外显子水平预测准确性比较（引自Chowdhury et al.， 2017）

ESn. 敏感性；ESp. 准确性；Eavg. 敏感性与准确性的平均值

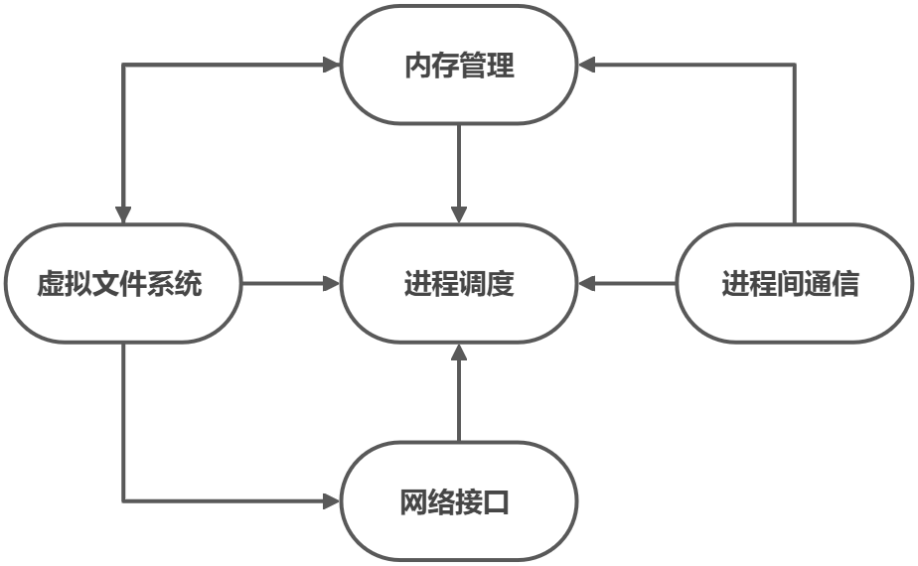


图15.1　Linux 内核组成部分与关系

箭头表示依赖关系

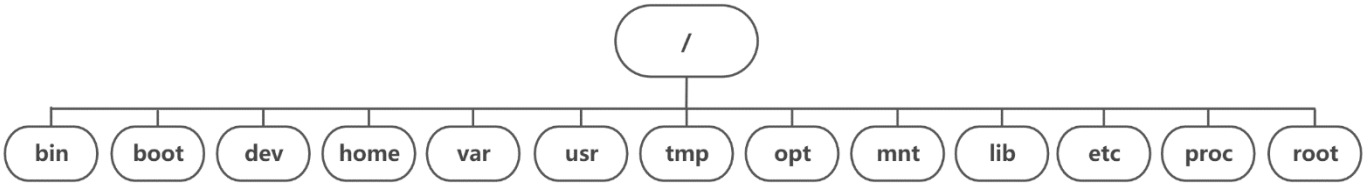


图15.2　Linux 文件系统树型结构

bin．存放二进制可执行文件（ls，cat，mkdir 等）；boot．存放用于系统引导时使用的各种文件；dev．用于存放设备文件；home．存放所有用户文件的根目录；var．用于存放运行时需要改变数据的文件；usr．用于存放系统应用程序，比较重要的目录是/usr/local 本地管理员软件安装目录；tmp．用于存放各种临时文件；opt．额外安装的可选应用程序包所放置的位置；mnt．系统管理员安装临时文件系统的安装点；lib．存放文件系统中的程序运行所需要的共享库及内核模块；

etc．存放系统配置文件；proc．虚拟文件系统，存放当前内存的映射；root．超级用户目录

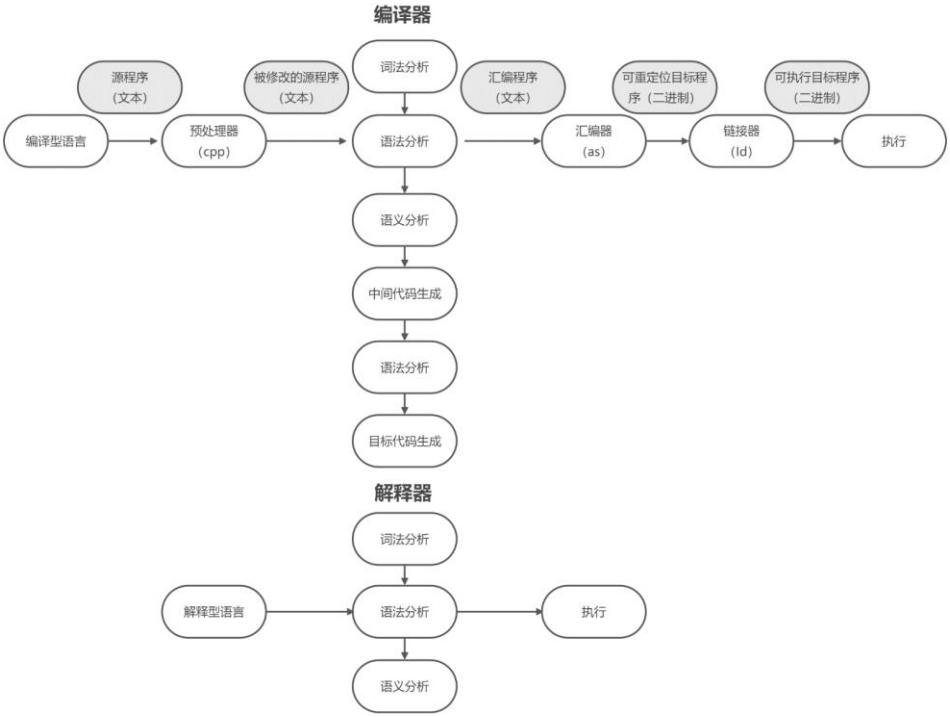


图 15.3　编译型语言与解释型语言的编译原理比较

```
#!/usr/bin/python
import sys
from Bio import Entrez
from Bio import SeqIO
file_in_name="test.id"
file_out_name="result.fasta"
    Entrez.email = '0014289@zju.edu.cn'    #email
input_file=open(file_in_name,"r")
output_file=open(file_out_name,"a")
for record_id in input_file:
result_handle = Entrez.efetch(db="nucleotide", rettype="gb", id=record_id)
    seqRecord = SeqIO.read(result_handle, format='gb')
    result_handle.close()
    output_file.write(seqRecord.format('fasta'))
    output_file.close()
    input_file.close()
```

图 15.4　利用 Biopython 批量下载目标基因序列并保存为 FASTA 格式文件

```
>>> from Bio.Blast.Applications import NcbiblastxCommandline
>>> help(NcbiblastxCommandline)
...
>>> blastx_cline = NcbiblastxCommandline(query="result.fasta", db="nr", evalue=0.001,
outfmt=5, out="opuntia.xml")
>>> blastx_cline
NcbiblastxCommandline(cmd='blastx', out='result.xml', outfmt=5, query='result.fasta',
db='nr', evalue=0.001)
>>> print blastx_cline
blastx -out opuntia.xml -outfmt 5 -query opuntia.fasta -db nr -evalue 0.001
>>> stdout, stderr = blastx_cline()
```

图 15.5　利用 Biopython 中 Bio.Blast.Applications 模块实现 BLASTX 序列注释

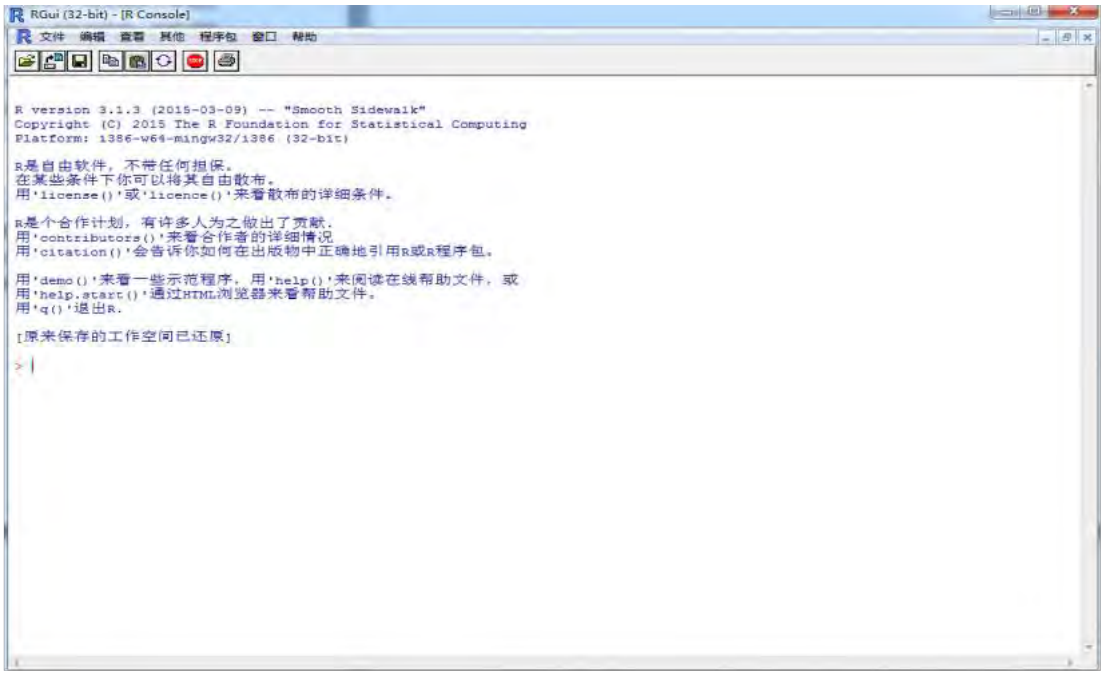


图15.6　R软件简介

A．CRAN 镜像点上 R 程序包及其功能介绍；B．R 语言工作界面

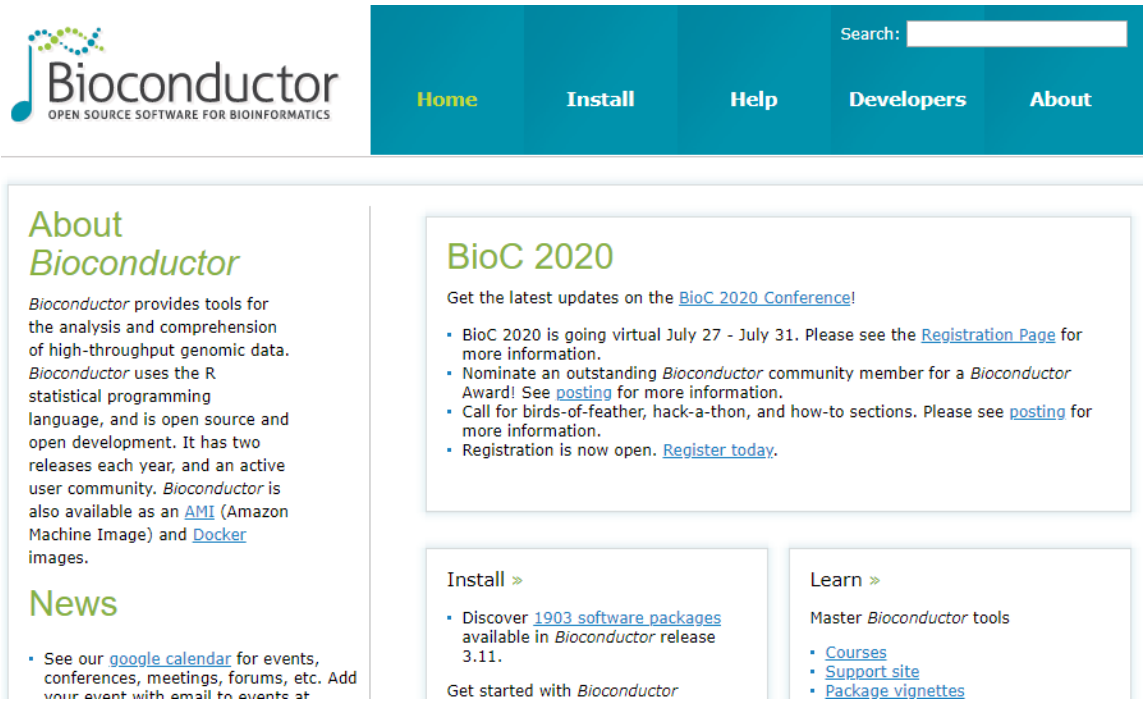


图 15.7　Bioconductor 主页

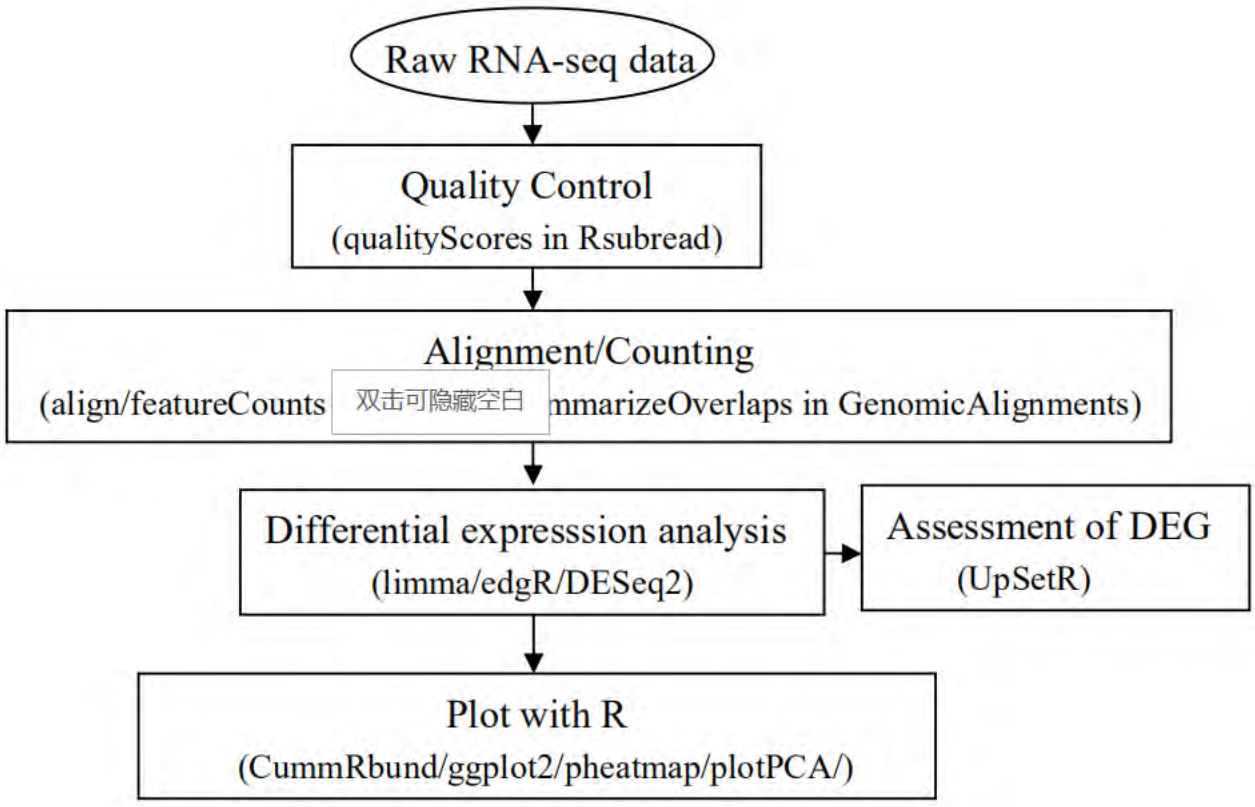


图15.8　基于R 语言的RNA-Seq 有参考基因组完整解决方案

针对 RNA 测序分析相关内容（每个框内），括号里分别提供了这些分析用到的 R 包及函数

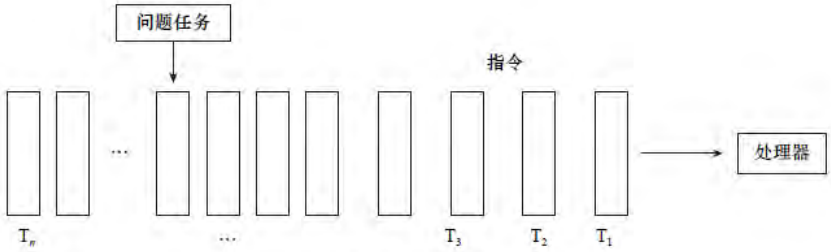


图 15.9　串行式计算原理

T 表示第 n 个指令 / 操作 / 任务

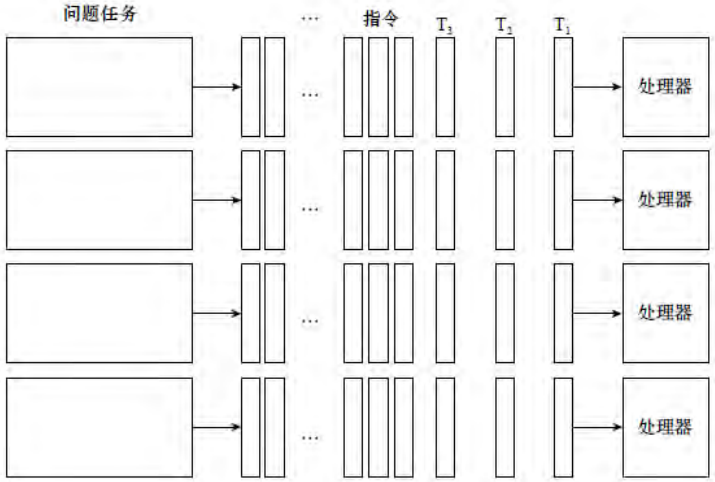


图 15.10　并行式计算原理

T 表示第 n 个指令 / 操作 / 任务

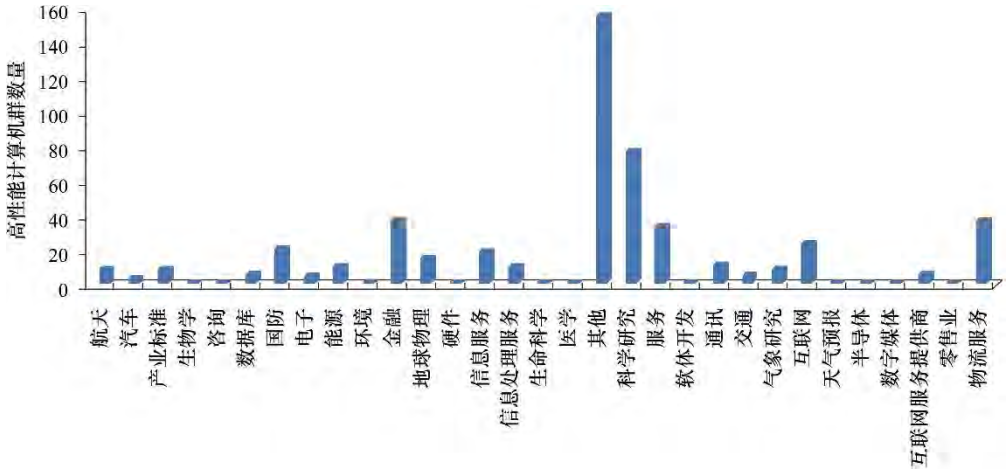


图 15.11　世界 Top500 高性能计算应用领域（图片来自美国劳伦斯利弗莫尔国家实验室）

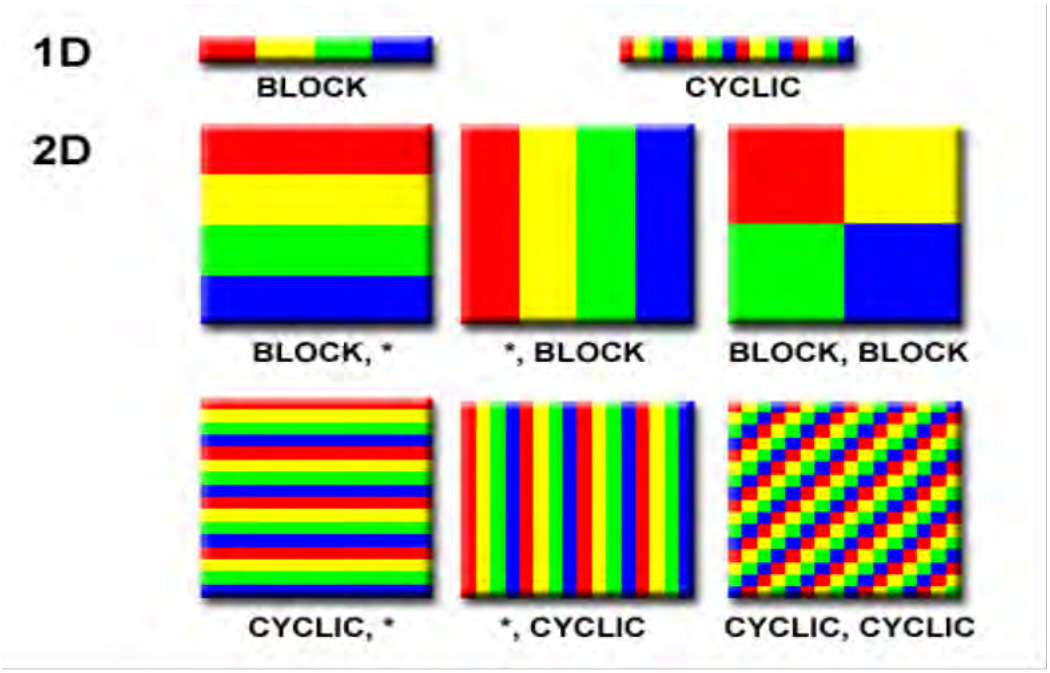


图15.12　不同方法分解数据

（图片来自美国劳伦斯利弗莫尔国家实验室）

```
#include <stdio.h>

Int main( int argc, char *argv[] )
{
    Printf("Hello, World!\n");
}
```

编译并运行：

```
gcc–o hello hello.c
./hello
```

```
#include <stdio.h>
#include "mpi.h"

main( int argc, char *argv[] )
{
    MPI_Init(&argc, &argv );
    printf("Hello, World!\n");
        MPI_Finalize();
```

编译并运行：

```
mpicc–O2–o hello hello.c
mpirun–np 4 hello
```

(-np:指定运行程序的进程数)

```
program main
include'mpif.h'
integer ierr

call MPI_INIT( ierr )
print *, 'Hello, World!'
call MPI_FINALIZE( ierr )
end
```

编译并运行：

```
mpicc–O2–o hello hello.c
mpirun–np 4 hello
```

图 15.13　串行 C（左）、并行 C（中）和 Fortran（右）程序比较

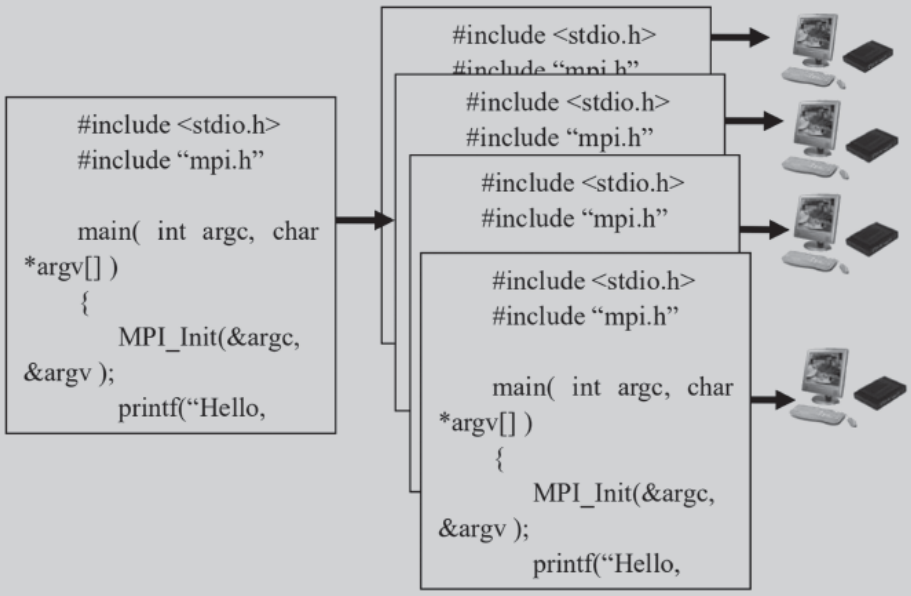


图 15.14　并行计算运行流程

图 15.15　数据的可视化

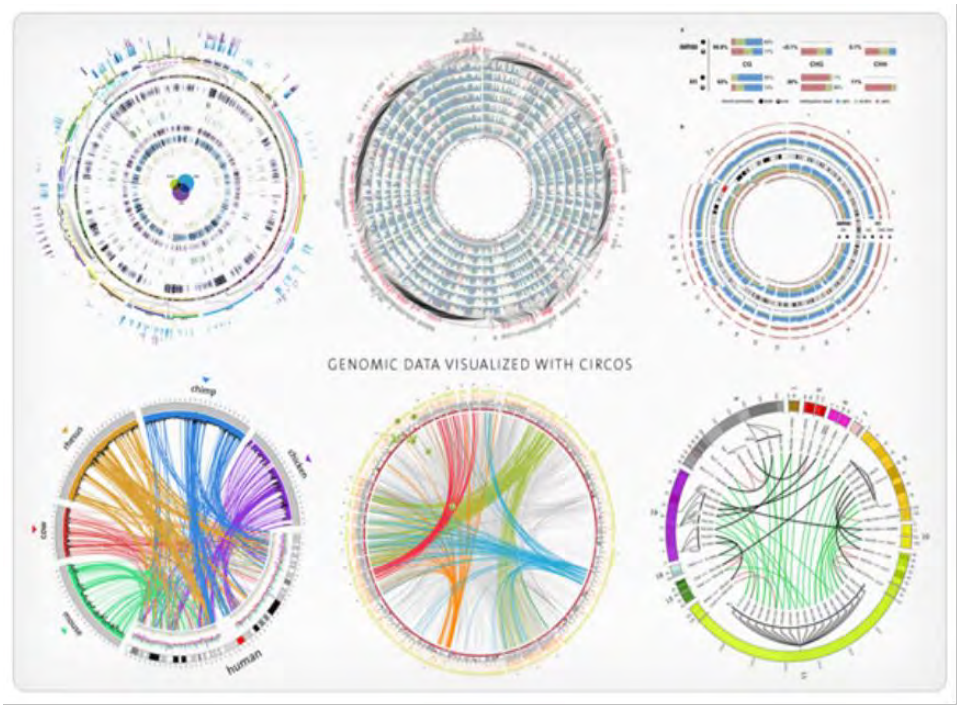


图 15.16　Circos 可视化工具

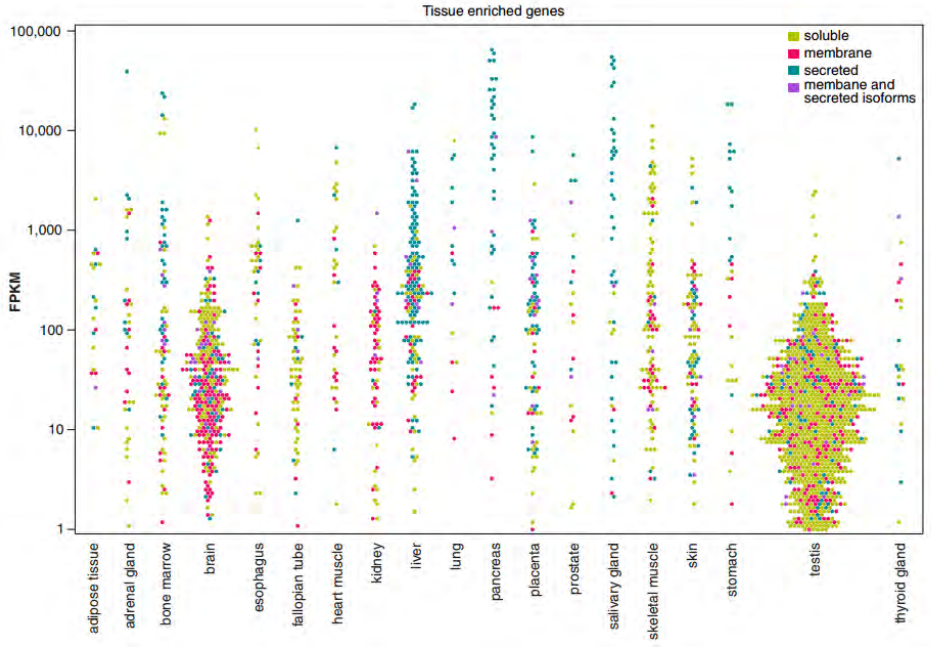


图 15.17　R 语言的可视化（图片来自 http://science.sciencemag.org/content/347/6220/1260419）

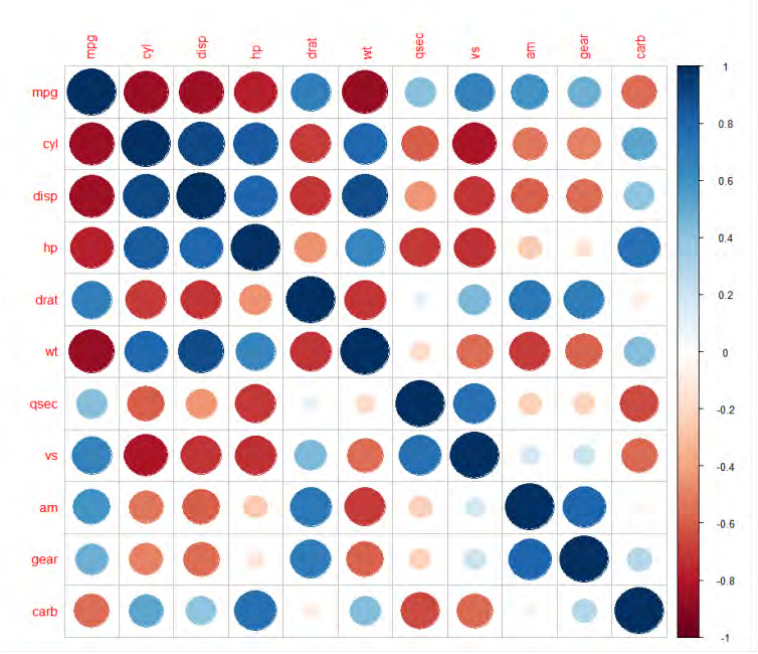


图 15.18　ggplot2 中的 corrplot 包示例

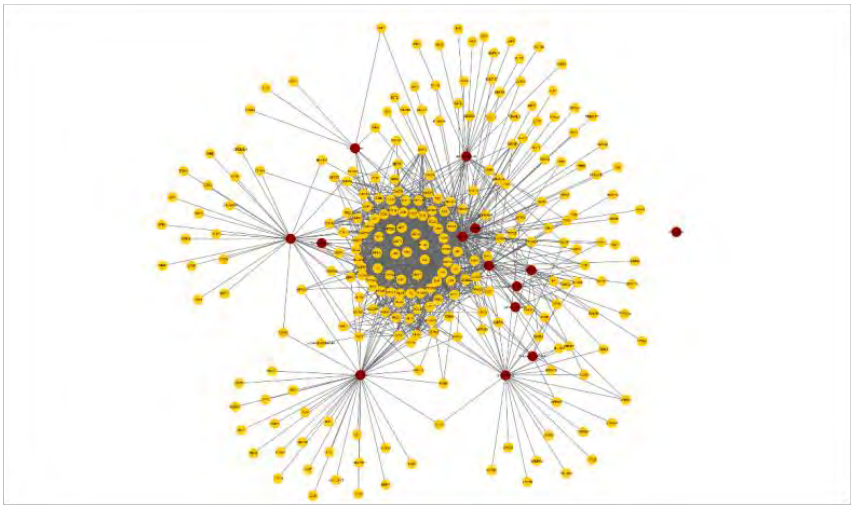


图 15.19　R 包的网络图