《Genomics》

Genome sequencing, assembly and annotation

Longjiang Fan (樊龙江)

Prof. of crop genetics and breeding
Institute of Crop Science
College of Agriculture and Biotechnology, Zhejiang University
http://cab.zju.edu.cn/nxx/

Prof. of bioinformatics
Institute of Bioinformatics and IBM Biocomputational Lab, Zhejiang University
http://ibi.zju.edu.cn

The Bioinplant Lab
(Bioinformatics and Plant Genomics Lab)
http://ibi.zju.edu.cn/bioinplant/

樊龙江教授

生物信息学专业博导

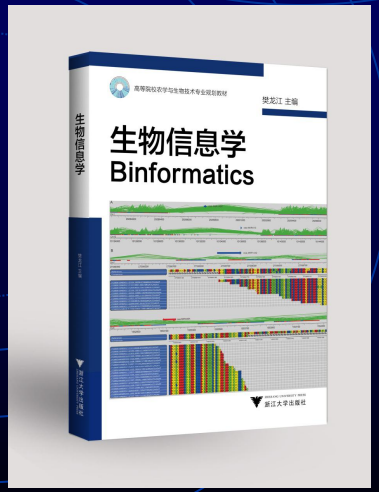
作物遗传育种专业博导

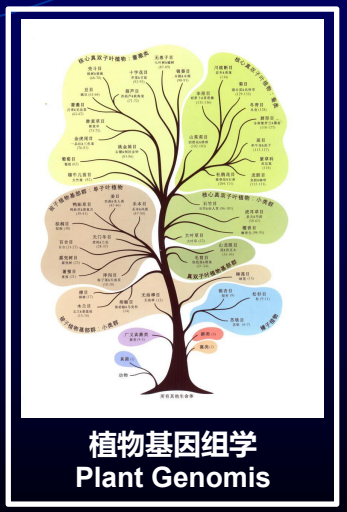
- ✓浙江大学作物科学研究所 生物信息学与大数据技术创新团队负责人
- ✓浙江大学中美作物分子育种联合实验室 主任
- ✓浙江大学生物信息学研究所 执行所长
- ✓浙江省生物信息学学会 副理事长
- ✓教育部“新世纪优秀人才支持计划”入选者

主编教材：

《生物信息学》（浙大出版社，2017）

《植物基因组学》（科学出版社，2019）





植物基因组学
Plant Genomis

《植物基因组学》简要目录

章节	内容	字数（万）
序		
前言		0.1
	第一篇 总论	
第1-1章	绪论	1.4
第1-2章	植物基因组测序与拼装	1.2
第1-3章	植物基因组构成	1.5
第1-4章	植物基因组转录	1.0
第1-5章	植物基因组表观遗传修饰	1.0
第1-6章	植物基因组进化	2.3
第1-7章	植物群体基因组	2.3
第1-8章	植物单细胞基因组	1.1
第1-9章	植物三维基因组	0.8
第1-10章	植物合成基因组	2.0
第1-11章	叶绿体基因组	1.1
第1-12章	植物线粒体基因组	1.2
第1-13章	植物功能基因组学技术	2.2
第1-14章	组学数据育种利用	2.5
第1-15章	植物基因组数据资源	0.6
	第二篇 各论	
第2-1章	拟南芥基因组	1.3
第2-2章	水稻基因组	1.7
第2-3章	玉米基因组	1.2
第2-4章	小麦基因组	1.6
第2-5章	大豆基因组	1.0
第2-6章	棉花基因组	1.1
第2-7章	油菜基因组	1.0
第2-8章	蔬菜基因组	1.1
第2-9章	林果花草基因组	2.2
第2-10章	非维管束植物基因组	2.6

Contents 目 录

第一篇 总论

2. Genome sequencing, assembly and annotation

- 2.1 Genome sequencing
- 2.2 Genome assembly
- 2.3 Genome annotation

Importance of a high-quality genome

Reference genome (2002-) → Comparative genomics

pedigree genome (maize, Lai et al. 2010)

Population genome

Full-length cDNA project ?

Rice, maize, soybean, millet
~50 rice lines (Xu et al. 2011)
~500 rice lines (Huang et al. 2009)
~1000 rice lines (Huang et al. 2012)
First haplotype map of maize (Gore et al. 2009)
~31 soybean lines (Lam et al. 2010)
~1000 millet lines (Peng et al. 2013)

Inbred line genome?

2.1 Genome sequencing

- Roadmap for studying a genome
- How to sequence a genome
- Sequencing technologies

Roadmap for studing a genome

- Genomic geography (“基因组地理”探险)
  - Four maps

A、遗传图谱(genetic map)

- 遗传学图距是以形成精子或卵子的减数分裂过程中，两个位点之间进行交换、重组百分率（cM,厘摩尔根）为单位的，反映基因遗传效应的基因图谱。
- 限制性片段长度多态性（RFLP）\微卫星标记（microsatellite marker）\SNP（single nucleotide polymorphysm）的遗传标记系统，即单核甘酸多态性

B、物理图谱(physical map)

- 物理图以Mb、kb、bp作为图距，以DNA探针的STS（sequence-tagged site，序列标签位点）序列为路标。
- 构建物理图谱的一个主要内容是把含有STS对应序列的DNA的克隆片段连接成相互重叠的“片段重叠群（conting）”。
- 在YAC载体的基础上，现有BAC（细菌人工染色体库）、PI（一种源于λ噬菌体的载体）、PAC（来源于PI的人工染色体）、MAC（一种类似于YAC，但以哺乳细胞作为宿主细胞的新型载体）、Fosmid（一种类似于BAC，来源于大肠杆菌的F基因的载体）。

C、序列图谱(sequence map-genome)

- 由全部核甘酸组成的基因组序列图。前面所谈的遗传图与物理图的构建都是为了绘制序列图而建的。
- 基因组序列
  - 1977年，人类完成对自然界第一个基因组（全长5.3kb的φx174噬菌体）全序列测序，整个测序历时将近一年时间。
  - 第一个细菌基因组全序列（1995，1.9Mb）
  - 模式生物酶母基因组全序列（1996， 12Mb）、线虫基因组全序列（1998，97Mb）、果蝇基因组全序列（1999，136Mb）
  - 人类自身全序列（2001，3286Mb）；拟南芥（2000）和水稻（2002，400M）

D、基因图谱(gene map)

- 基因图谱就是基因组中全部基因的位置、结构与功能的明细图。
- 基因图谱的意义在于它能有效地反映在正常或受控条件中表达的全基因的时空图。

Genome survey

- To get a big picture of a target genome, such genome size, GC content, repeat content, heterozygous rate, polyploid, based on genome survey sequencing
- Genomic data: 20-40 genome coverage (×) of next-generation sequencing data

Genome size estimation based on *k*-mer

- 假设存在完整连续序列G，随机选取片段长度为*K*，该片段称为*K*-mer。当达到一定覆盖度时，根据*K*-mer数量和深度估计G长度(Lander_waterman 算法)。
- Clone fringeprinting scheme for a physical map (Lander and Waterman, 1988)
- *l*-tuples (Li and Waterman, 2003)

Genomic Mapping by Fingerprinting Random Clones: A Mathematical Analysis

ERIC S. LANDER*,† AND MICHAEL S. WATERMAN‡

*Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, Massachusetts 02142; †Harvard University, Cambridge, Massachusetts 02138; and ‡Departments of Mathematics and Molecular Biology, University of Southern California, Los Angeles, California 90089

Received January 13, 1988; revised March 31, 1988

Results from physical mapping projects have recently been reported for the genomes of *Escherichia coli*, *Saccharomyces cerevisiae*, and *Caenorhabditis elegans*, and similar projects are currently being planned for other organisms. In such projects, the physical map is assembled by first "fingerprinting" a large number of clones chosen at random from a recombinant library and then inferring overlaps between clones with sufficiently similar fingerprints. Although the basic approach is the same, there are many possible choices for the fingerprint used to characterize the clones and the rules for declaring overlap. In this paper, we derive simple formulas showing how the progress of a physical mapping project is affected by the nature of the fingerprinting scheme. Using these formulas, we discuss the analytic considerations involved in selecting an appropriate fingerprinting scheme for a particular project.

available region of up to several megabases and of studying its properties. In addition, the overlapping clones comprising the physical map would constitute the logical substrate for efforts to sequence an organism's genome.

Recently, three pioneering efforts have investigated the feasibility of assembling physical maps by means of "fingerprinting" randomly chosen clones. The fin-
ger tion
frag in-
fer iffi-
cien spe-
cifi dly
spe hat
the . In
par

G = haploid genome length in bp;
L = length of clone insert in bp;
N = number of clones fingerprinted;
$\alpha = N/G$ = probability per base of starting a new clone;
T = amount of overlap in base pairs needed to detect overlap;
$\theta = T/L$;
c = redundancy of coverage $= LN/G$.

(i) Olson *et al.* (1986) fingerprinted 5000 λ clones containing approximately 15-kb inserts of genomic DNA from *Saccharomyces cerevisiae*, by measuring

基于*K-mer*的分析方法来来估计基因组大小和杂合率等，即从一段连续序列中迭代地选取长度为K个碱基的序列，若read长度为 L，*K-mer* 长度为 K，那么可以得到L-K+1个*K-mer*。

定义：

K-mer 深度分布曲线：*K-mer* 深度~*K-mer*深度频率。

假设：

1）从 read 中逐碱基取出的所有*K-mer*能够遍历整个基因组。

根据 Lander_waterman 算法，基因组大小（G）满足以下公式：

$$G = \frac{k_{num}}{k_{depth}} = \frac{b_{num}}{b_{depth}}$$

k_{num}为 *K-mer* 个数，k_{depth}为 *K-mer*期望深度，b_{num}为碱基个数，b_{depth}为碱基期望深度。

2）*K-mer* 深度频率分布服从泊松分布

$$P(X = k) = \frac{e^{-\lambda}\lambda^{k}}{k!}$$

λ 为期望值，取整后得到众数，即峰值为对应的测序*K-mer*深度，作为*K-mer*深度的估计值K_{depth}。

Genome size estimation based on *k*-mer

- Genome size = (*K*_num-*K*_unique) / peak_depth
- where *K*_num is the total number of *K*-mer, *K*_unique is the number of single or unique *K*-mer words and peak_depth is the expected value of *K*-mer depth.

k-mer counts and frequencies

GCGAGATCCAACGGTGAACAGCTGCCCAAAAGAAAAaCCGCCTGGAAGTCCGA
GGACCTTTAGTACTGTACTCTACCCCCGAACCAGCAGCCTTCGtGCCAaGCAA
GACCGCCCTTGTCCCTTTCCTTTATCCATTCCGCcTCCTTCTTTGCTTTGTTC
CAATAGAGTCTAAGGCAAAGCTAAAGTGGTTCGTaTGCCTACTTTACCTACTT
GACGAAAGGGAACGAACTTCGTTTCGTTTCCGGGTTTATGGATTGGATTCAGT
CAGCCTCACTCCTTCCTTTTTATGTTGTCGTGATGGTTACCGGCGAACGCTCC
CAAAGGCGACCCTCTCGAGTTTCCGGCTGTTTTCTAGATTGAAGTAGCCTTTC
GTCGCCCCGAAAGAAGTCACTATCAAAGAGCTCGCCCTACTGAAGTACCAAAG
GTGCGCTCAGCCCGGTGACTAAGAAATGGGTTTGCGCTTGAATTGAAGTGATG
AGGTTTTTCGAGGGAAGTAGGGCTCTTATTGACTAAAAGTGGGTTCTTCGCTT
TCCTTTAGAATGAAAGTTGCTATGAAGCCCCTACTACTTACTTTGTTTGATTC
AAAAGGCGAACGGCCCCCCAACAAGTCGTATGGGGTGGGGTGCTTGTGATAAG
CTGCCTTGGATATGAGGAATTCTCAAATTGGGAAAGCATTTCTTGATTTGAAG
AAACAAGAAAGTTAGGGTTTTTGGAATTGGATTCGGATAATGTTTGTTGTTTT
TtGTAAGTGTGAGATTAGAGGTTCACGAAATTTTGATGGG

k = 8

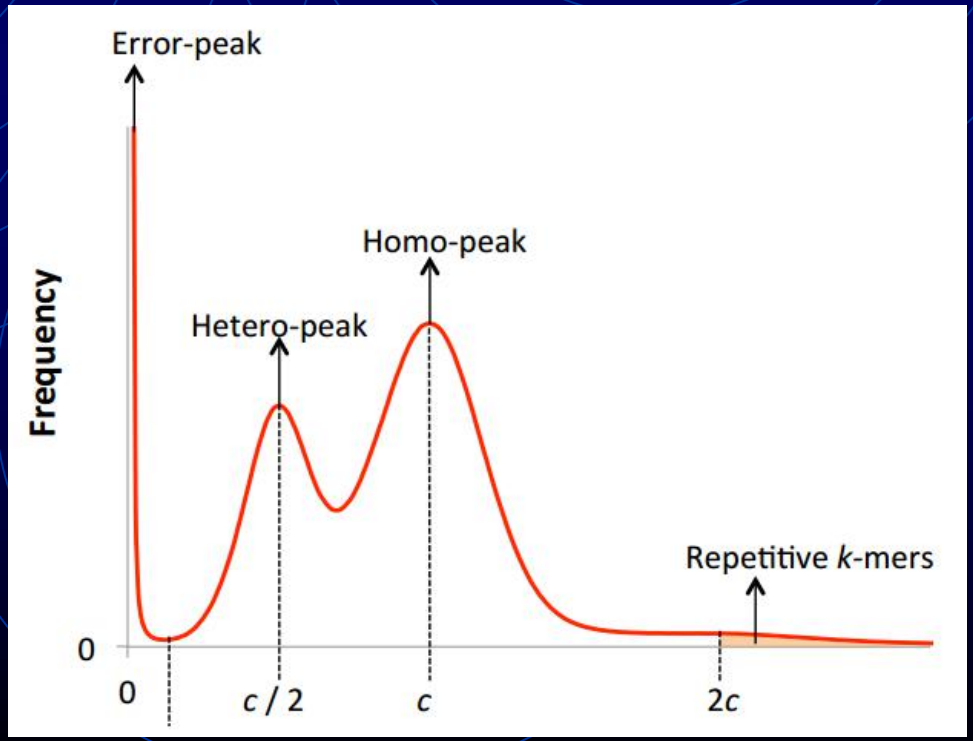
Total n = 782

8-mers = 775

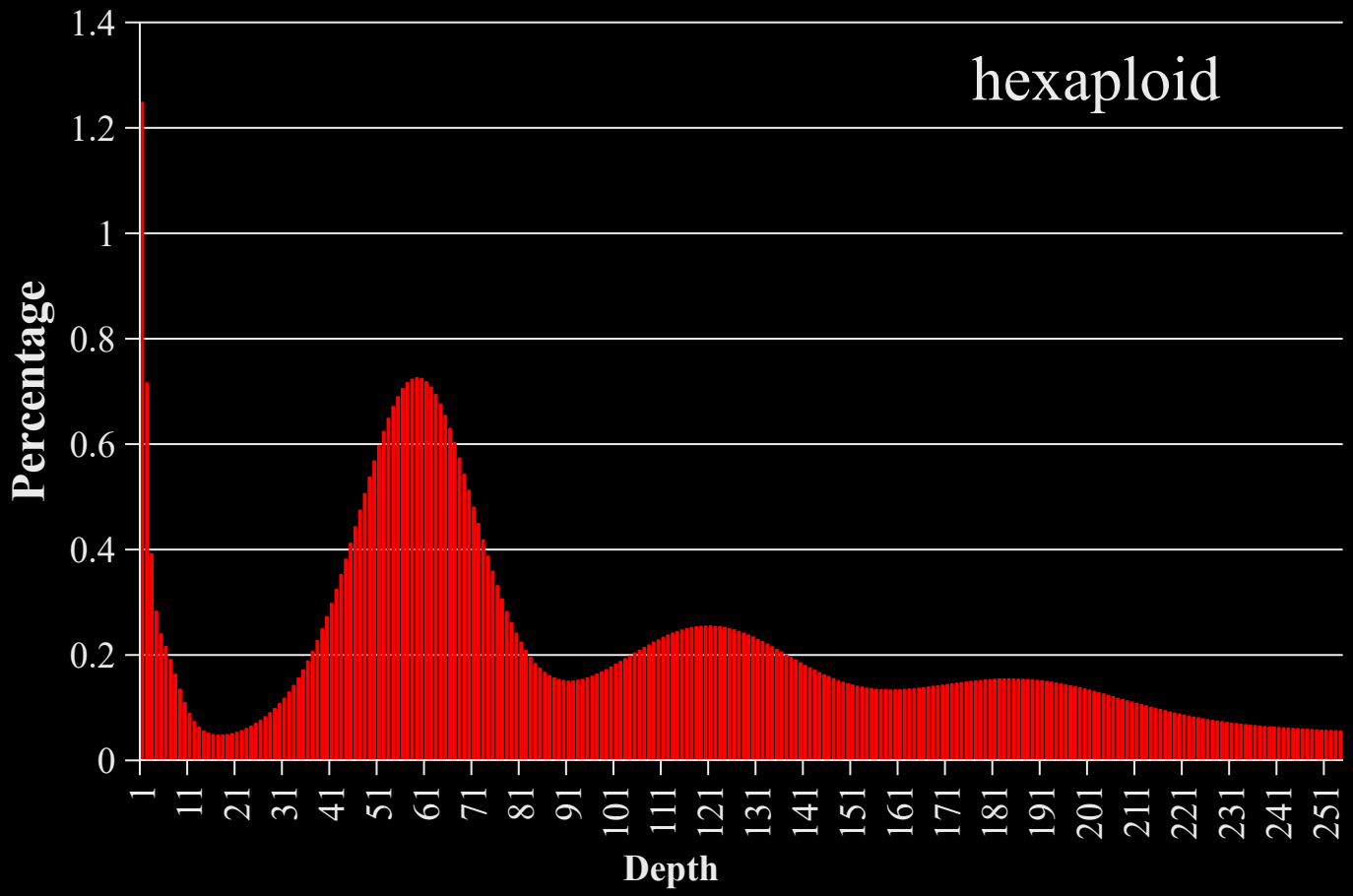
Unique = 98%

*K-mer*深度分布曲线

- 受到基因组杂合度、倍性和重复序列构成的影响，因此可以用于评价基因组杂合度和重复序列比例



Genome size estimation of barnyardgrass *E. crus-galli*.

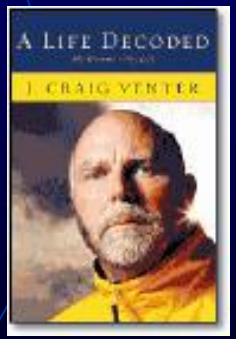


基因组测序的策略
How to sequence a genome?

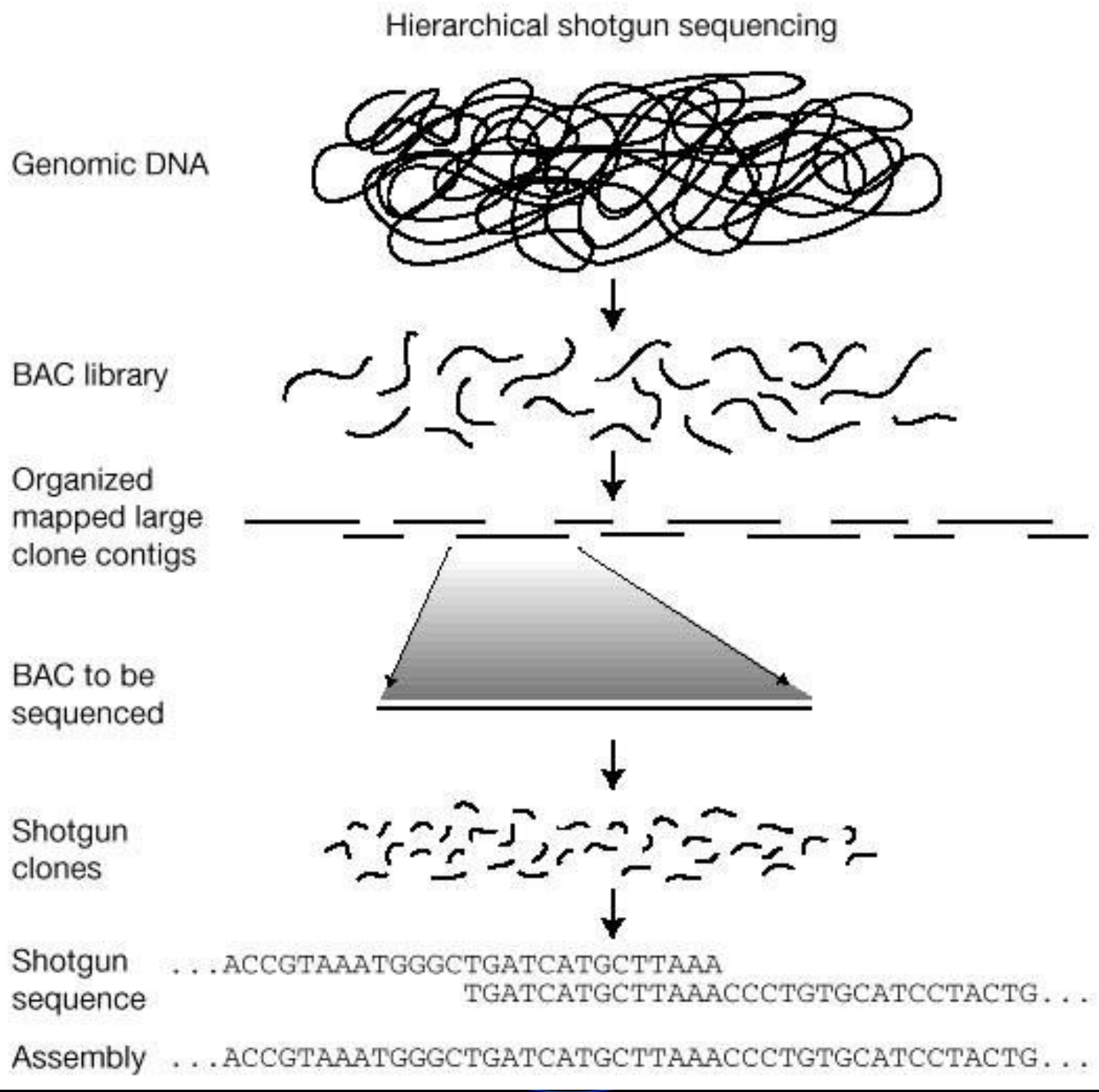
两个基因组测序策略：

逐步克隆方法
（clone by clone)

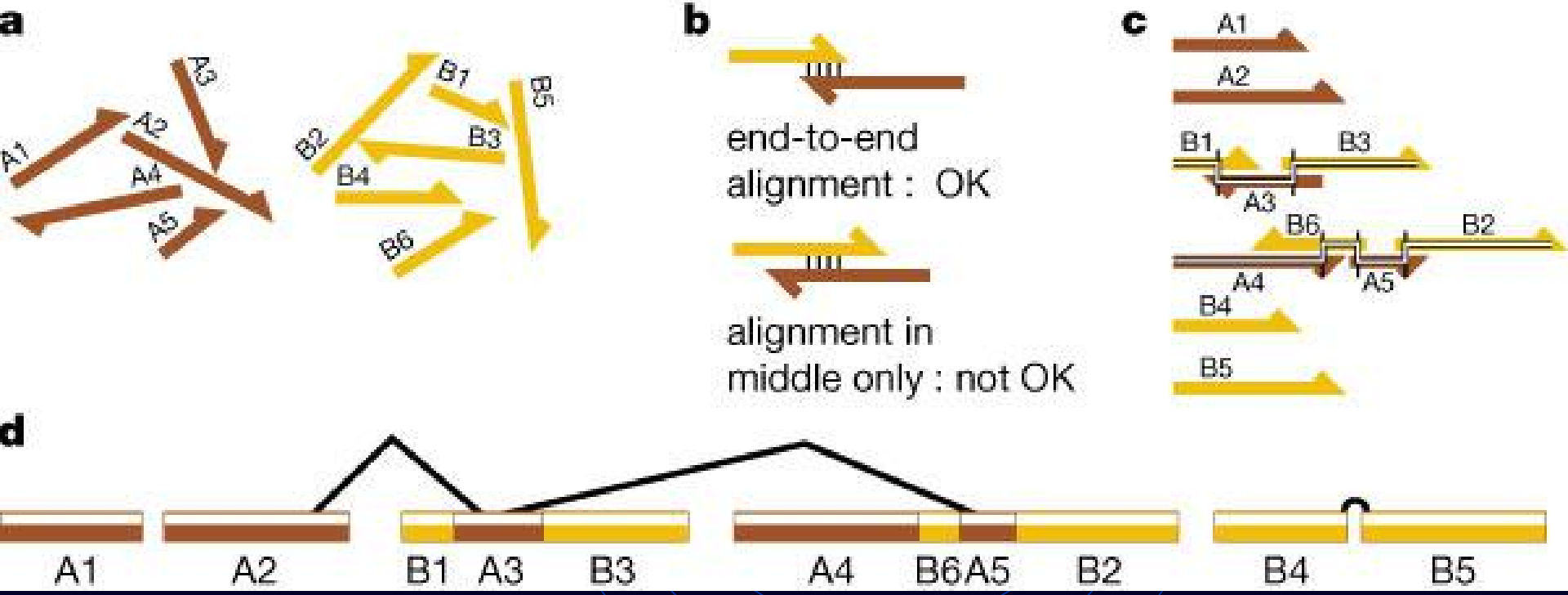
全基因组鸟枪方法
（whole genome shotgun, WGS)



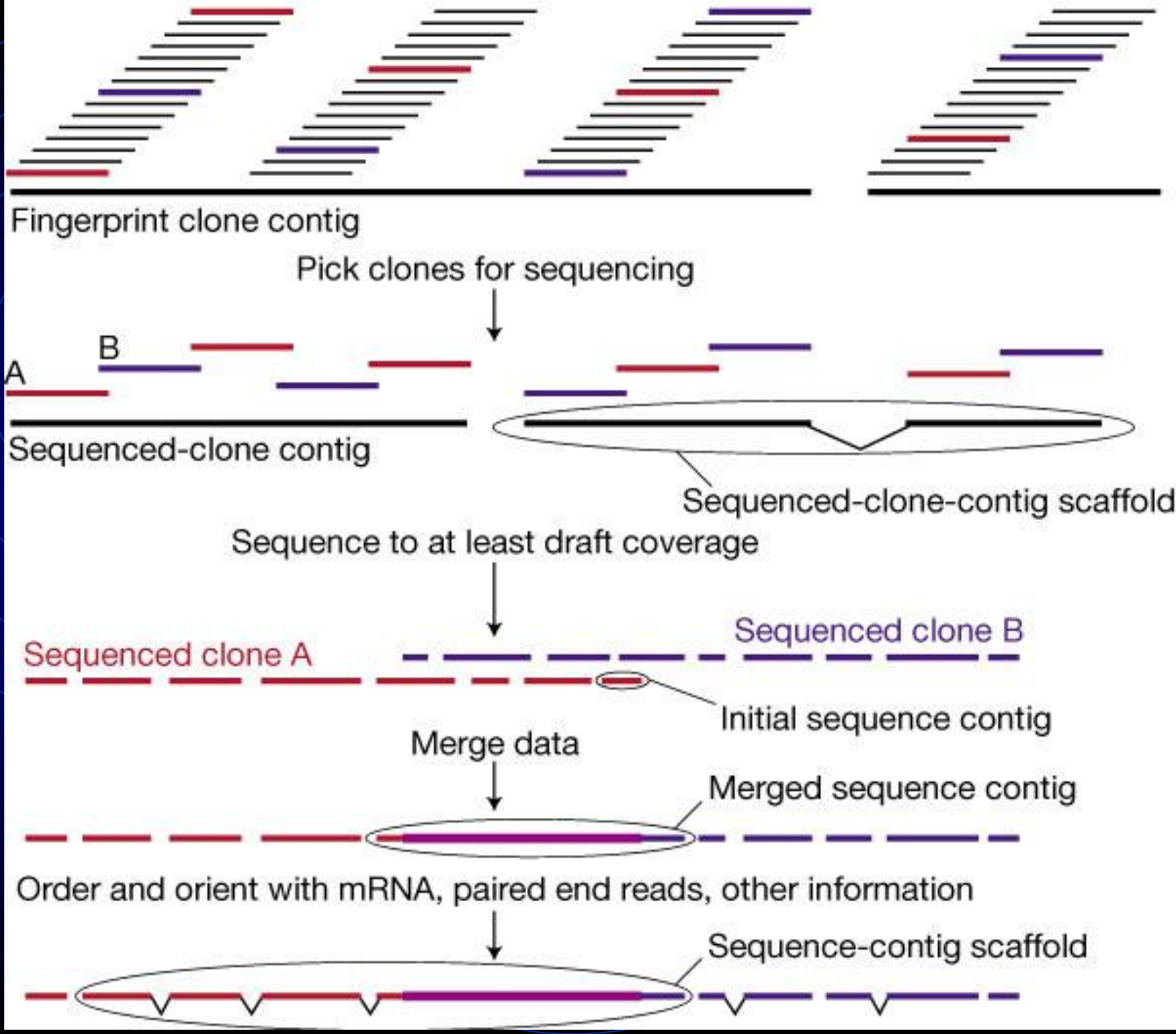
Clone by Clone Approach



The key steps in assembling individual sequenced clones into the draft genome sequence



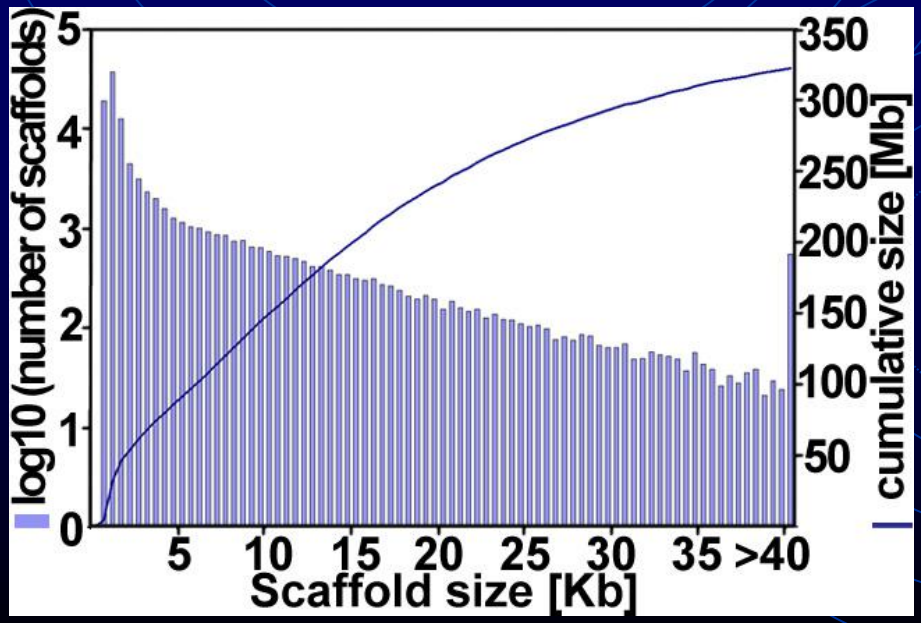
Contig and Scaffold



Quality

- ✓ Genome coverage
- ✓ Functional coverage
- ✓ Assembly

Contigs:127,550 (N50=6,688 bp)



Scaffolds: 102,444 (N50=11,764 bp)

Quality estimation

- Genomic coverage: flow cytometer; genome size estimation by *k*-mers.
- Functional coverage: traditional ESTs; reference genomes
- Assembly quality: BAC/PAC/FOSMID clone sequencing; BUSCO-a set of single copy orthologs

Genome sequencing

- Wikipedia: Whole genome sequencing (WGS), complete genome sequencing, or entire genome sequencing
- Sequencing technology
  - Sanger method
    - ABI3730: 700-900bp per read
  - high-throughput approaches
    - Illumina Geome Analyzer II System/ HiSeq 2000
    - Applied Biosystems SOLID System
    - 454/Roche GS FLX
    - PacBio
    - Nanopore

Read Length	Run Time	Output
1 x 35 bp	~1.5 days	26-35 Gb
2 x 50 bp	~4 days	75-100 Gb
2 x 100 bp	~8 days	150-200 Gb

Throughput
Up to 25 Gb per day for a 2 x 100 bp run.

Reads
Up to one billion clusters passing filter and up to two billion paired-end reads

Performance
HiSeq 2000 provides the greatest yield of perfect reads and bases greater than Q30

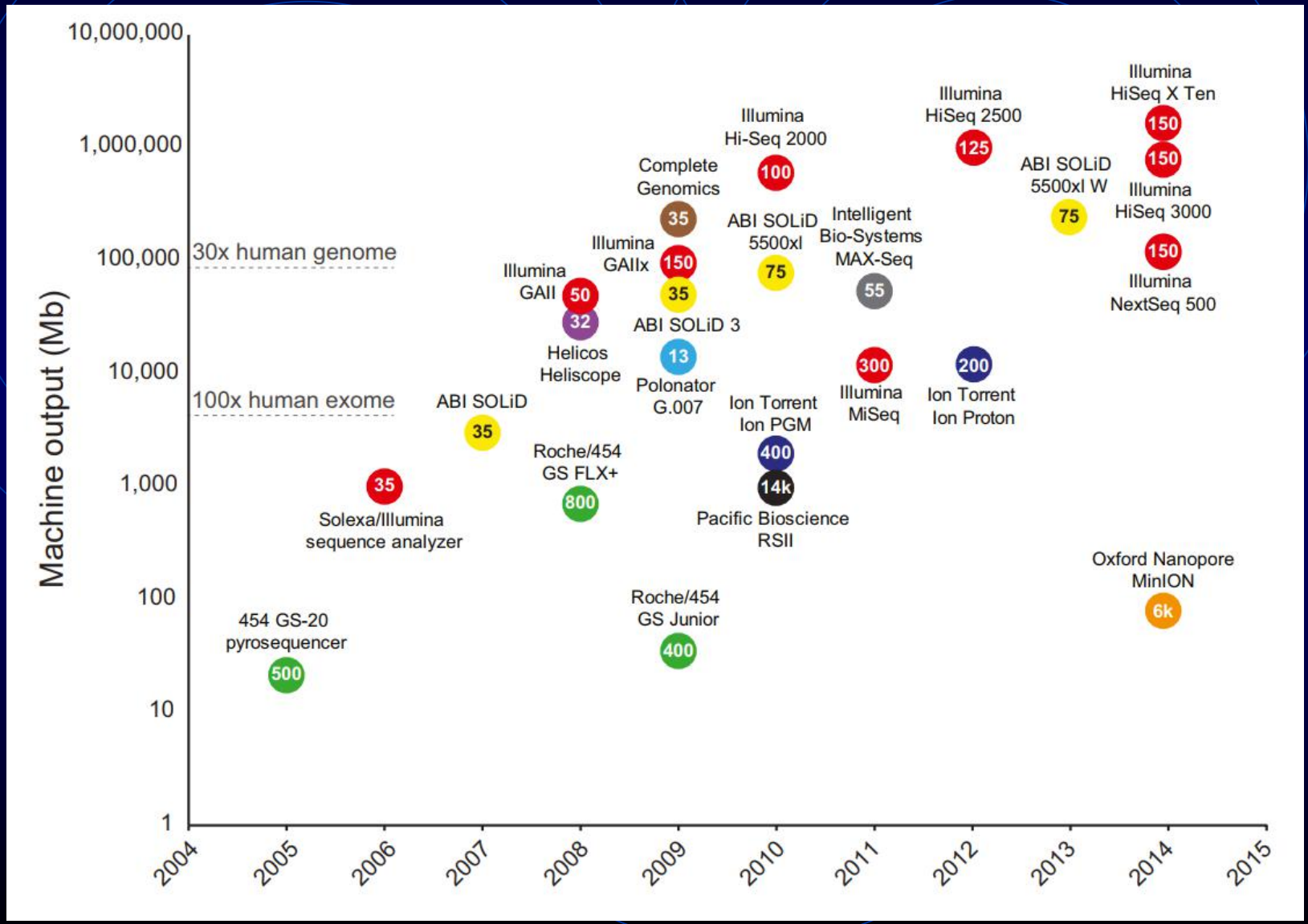
- Greater than 90% bases higher than Q30 at 2 x 50 bp*
- Greater than 85% bases higher than Q30 at 2 x 100 bp*

*Typical performance for sequencing output generated using TruSeq SBS-HS Kit with an Illumina PhiX library and cluster densities between 260 – 347K/mm2 that pass filtering on a HiSeq system. Performance may vary based on sample quality, cluster density, and other experimental factors. Paired 100 bp runs may vary in the range of 80 to 90% of bases higher than Q30 and paired 50 bp runs may vary in the range of 85 to 95% bases above Q30 based on the above factors.

Services and Support
Illumina will ensure that your HiSeq 1000 is properly installed and qualified, and will provide ongoing maintenance and service. This industry-leading support is available in North America, Europe, and Asia

Illumina: the sequencing-by-synthesis (SBS)

高通量测序技术出现年代及其测序通量变化趋势（Reuter 等，2015）



Genome projects

- **Genome projects** are scientific endeavours that ultimately aim to determine the complete genome sequence of an organism and to annotate protein-coding genes and other important genome-encoded features.

- **Genome assembly**
- **Genome annotation**
  - It consists of two main steps: identifying elements on the genome, a process called gene prediction, and attaching biological information to these elements.
  - These steps may involve both biological experiments and *in silico* analysis

Genome re-sequencing

- Genome re-sequencing:
  - Deep sequencing: 30-50X
    - SNP calling/ de novo assembly
  - Germplasm survey: 10-15X
    - SNP calling

Metagenomics: environmental samples

- **Metagenomics** is the study of **metagenomes**, genetic material recovered directly from environmental samples
- Traditional microbiology and microbial genome sequencing rely upon cultivated clonal cultures environmental samples.
- Early environmental gene sequencing cloned specific genes (often the 16S rRNA gene) to produce a profile of diversity in a natural sample. Such work revealed that the vast majority of microbial biodiversity had been missed by cultivation-based methods.
- Recent studies use "shotgun" Sanger sequencing or massively parallel pyrosequencing to get largely unbiased samples of all genes from all the members of the sampled communities.

2.2 Genome assembly

- About assembly
- Assembly Algorithms

Influence of technological changes

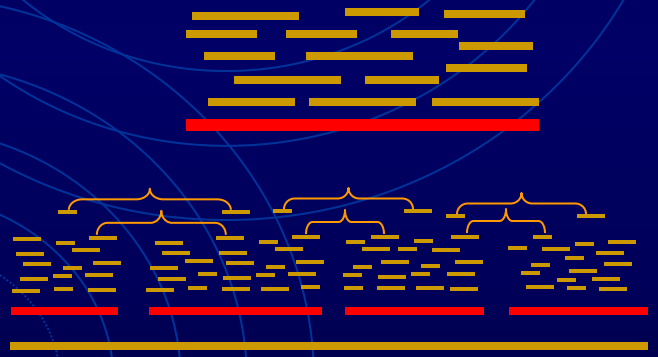
- The complexity of sequence assembly is driven by two major factors: the number of fragments and their lengths.
  - PHRAD: Sanger sequencing
  - NGS: 454/Illumina/PacBio
- The complexity of sequence assembly is also driven by other several factors: repeat; sequencing errors, high heterozygous rate, polyploid, etc.

Overlap-Layout-Consensus

Assemblers: ARACHNE, PHRAP, CAP, TIGR, CELERA

Overlap: find potentially overlapping reads

Layout: merge reads into contigs and contigs into supercontigs



Consensus: derive the DNA sequence and correct read errors

..ACGATTACAATAGGTT..

短序列（读序）拼装难题

- **短序列拼装几乎是近年来NGS（next-generation sequencing）最热门的话题。简单来说，就是把基因组长长的序列打断(shotgun sequencing)，因为我们不知道基因组整条序列是如何排列成为一条染色体的（如何区分不同染色体），而我们又无法实现一次把整条长序列完整测序。**
- **需要通过算法，把这些短的序列组装起来成为一条完整有序的序列。**

- 就好比我们有这样一句话：　it is just a hypothesis, so don't be seriously！
假设，我们现在不知道这句话到底是什么，就像我们有一个box，我们抽到一张纸，但没打开，我们把这张纸撕成pieces，当然可能还发生了变化，所有的空格和标点都消失了（魔术！）我们得到：
itis ypo stah the sodo eriou siss ju ntbes sly……
因为我们测了几次，为了增加覆盖度，这样我们能通过高覆盖度而提高置信度：
itis ypo stah the sodo eriou siss ju ntbes sly tis yopth sodon beser beser ssod iti sju……
另外，我们又发明了一种称作为paired-ends的序列测序方法，即两头定长，中间插入片段一定的序列，像这样：
iti*****ahyp sju*****pot the*****don sod*****ser bes*****sly ……
这样我们根据如下图的方法，我们可以把这句话拼回来：
itisjustahypothesissodontbeseriously
但它不是最终结果，我们根据我们的现有的语法习惯，我们给它们加上空格（gap)和标点（遗漏的关键东西），我们能够还原原话！

为什么需要新方法处理高通量数据？

- OLC: for very short reads, it is hard to distinguish correct assembly from repetitive sequence overlap due to there being only a very short sequence overlap between these short reads. Also, in practice, it is unrealistic to record into a computer memory all the sequence overlap information from deep sequencing.
- **The de bruijn graph** data structure, introduced in the EULER (Pevzner et al. 2001) assembler, is particularly suitable for representing the short read overlap relationship. The advantage of the data structure is that it uses *K*-mer as vertex, and read path along the *K*-mers as edges on the graph. Hence the graph size is determined by the genome size and repeat contents of the sequenced sample, and in principle, will not be affected by the high redundancy of deep read coverage.

Li et al. 2010, Genome Res.

Assembly Algorithms

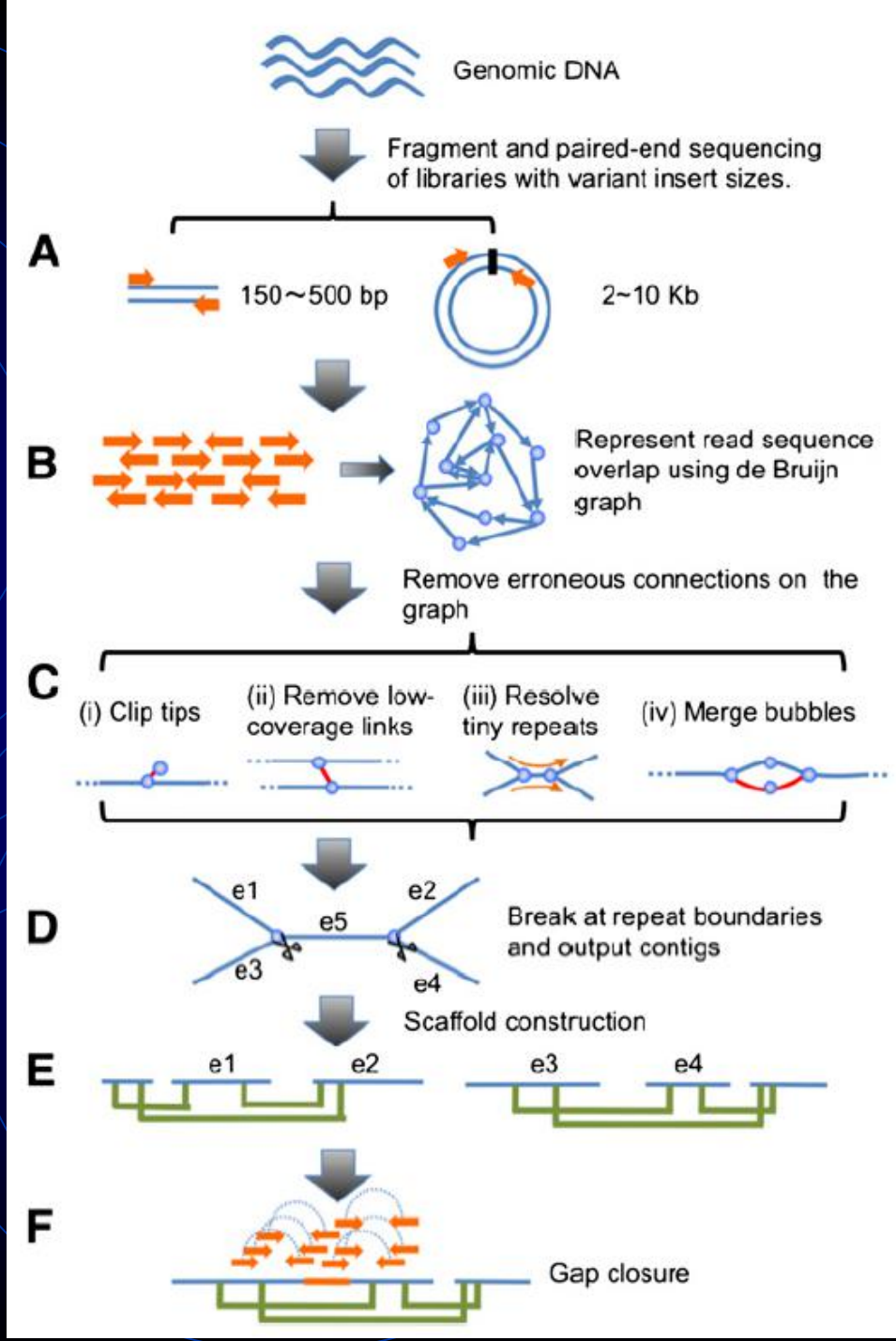
- Overlap–layout–consensus (OLC)
- de Bruijn Graph

基因组拼接主要利用两类算法

- 一是OLC算法(Overlap-Layout-Consensus)，适用于第一代测序技术等获得的长测序读序（read），但不适用于第二代测序数据的短读序（长度100BP左右）。OLC算法对于短读序，由于重复序列问题，很难基于序列重叠（overlap）获得一个正确的拼接结果，而且在实际运算过程中，大量重叠关系的读序信息需要大量内存，目前计算机能力难以承受。
- 另一类为基于德布鲁因图的算法，是目前用于高通量测序数据拼接的主要算法。基于德布鲁因图的数据结构，特别适合处理大量具有重叠关系的短度序。该数据结构利用*K*-mer作为顶点，读序作为边，这样总体上说，图的大小就是由目标基因组大小和重复序列含量决定，而与读序覆盖深度无法。

（Li et al., 2010）

SOAPdenov



Li et al. 2010, Genome Res.

Read, *k*-mer and de Bruijn Graph

表 1　一个 read 划分为 k-mer 的例子

Read:	A	G	A	T	A	C	T
k-mer:	A	G	A				
		G	A	T			
			A	T	A		
				T	A	C	
					A	C	T

K=3

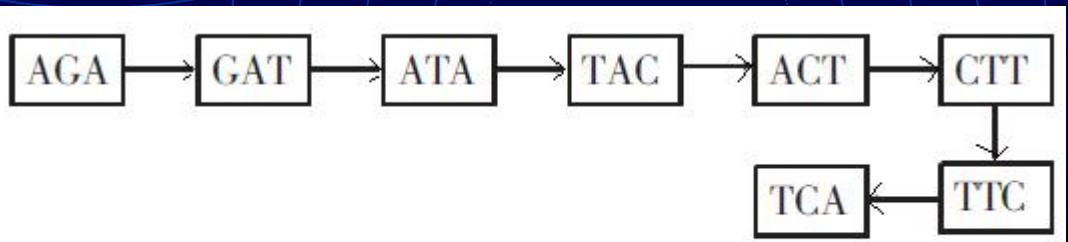


图 1　一个 de Bruijn 图实例

More complicated de Bruijn Graphs

ATCTTATTCG

ATCTAATTCG

ATCTTCCG

ATCTTATTCC

?

GACCTACA
ACCTACAA
CCTACAAG
CTACAAGT
TACAAGTT
ACAAGTTA
CAAGTTAG
TACAAGTC
ACAAGTCC
CAAGTCCG

repeat

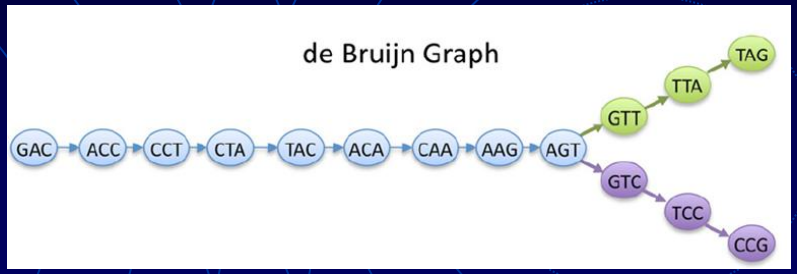
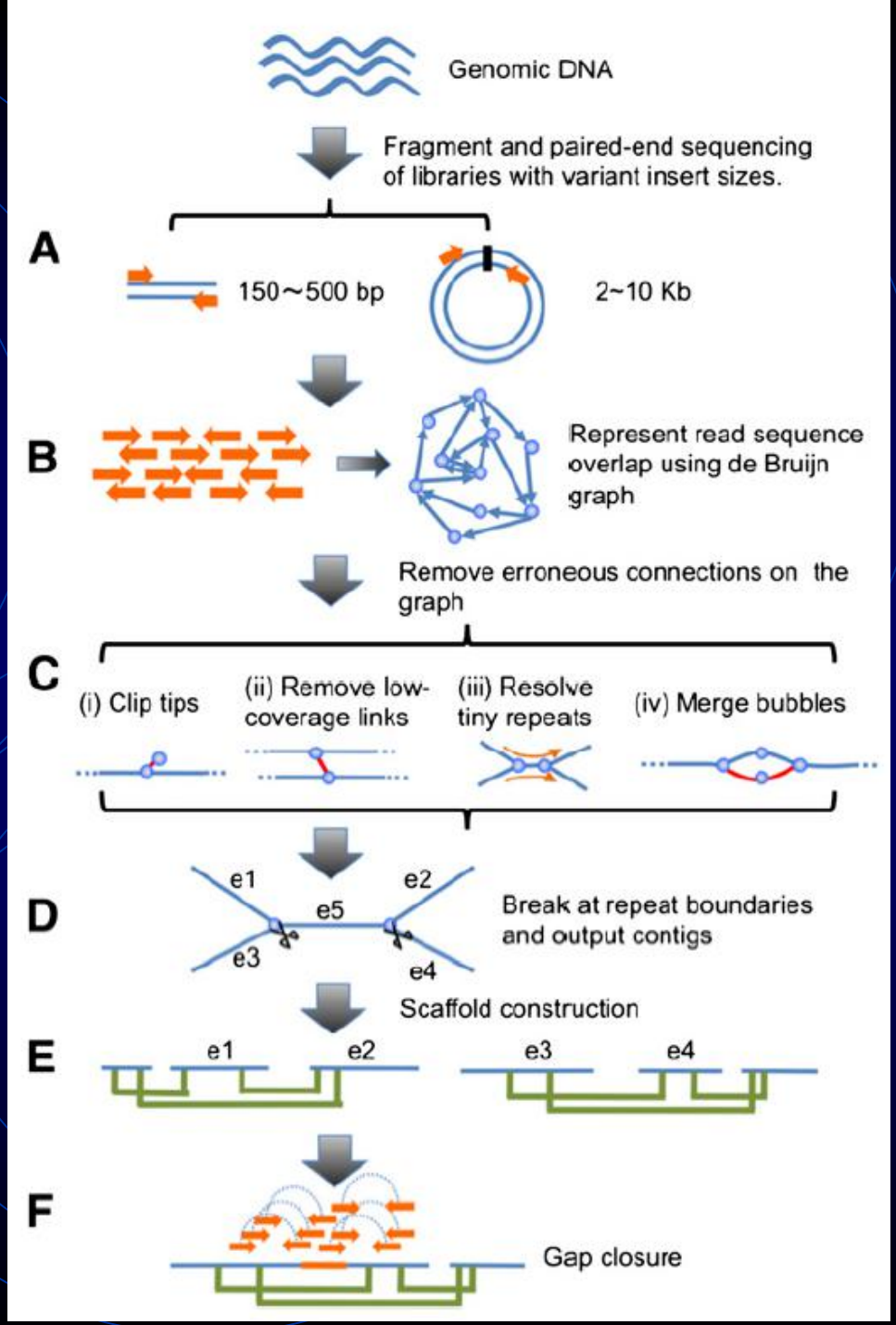


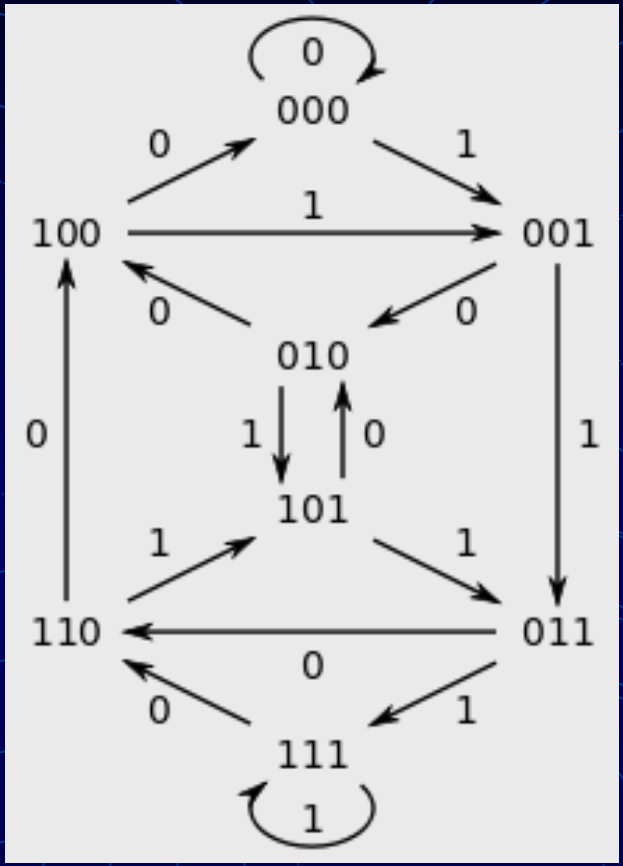
图 10　de Bruijn 图中分支结构简图

图 10　de Bruijn 图中分支结构简图

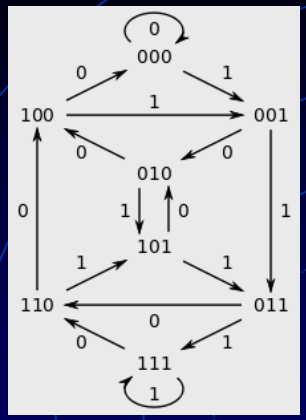
SOAPdenov

(Li et al. 2010, Genome Research)





德布鲁因（De Bruijn）图列举
(该图为3维*B*(2,4)序列图).



以图例构建$B(2,4)$序列。这是一个3维德布鲁因图，顶点由3个数字组成的子序列，边为4个数字组成。假如我们沿着如下欧拉路径行走：
000,000,001,011,111,111,110,101,011,110,100,001,010,101,010,100,000

这样就形成如下k长度为4的子序列串：
0000
_0001
__0011
……

对应的德布鲁因序列为“0000111101100101”

德布鲁因序列，$B(k, n)$，是由k个元素长度为n的亚序列构成的循环序列。以DNA序列拼接为例，$B(4, 20)$是指4个碱基k-mer长度为20的德布鲁因序列（contig）

确定德布鲁因序列 (de Bruijn sequence)

德布鲁因序列可以通过确定*n*维德布鲁因图的哈密顿路径或*n*-1维德布鲁因图的欧拉路径进行构建:

如果一条路径恰好通过每条边一次并且回到起始点，那么每四个数的亚序列（图论的边）会出现正好一次（该路径为欧拉回路）；如果路径刚好访问每个节点一次，那么每三个字的亚序列（图论的顶点）出现正好一次（该路径为汉密顿路径圈）。

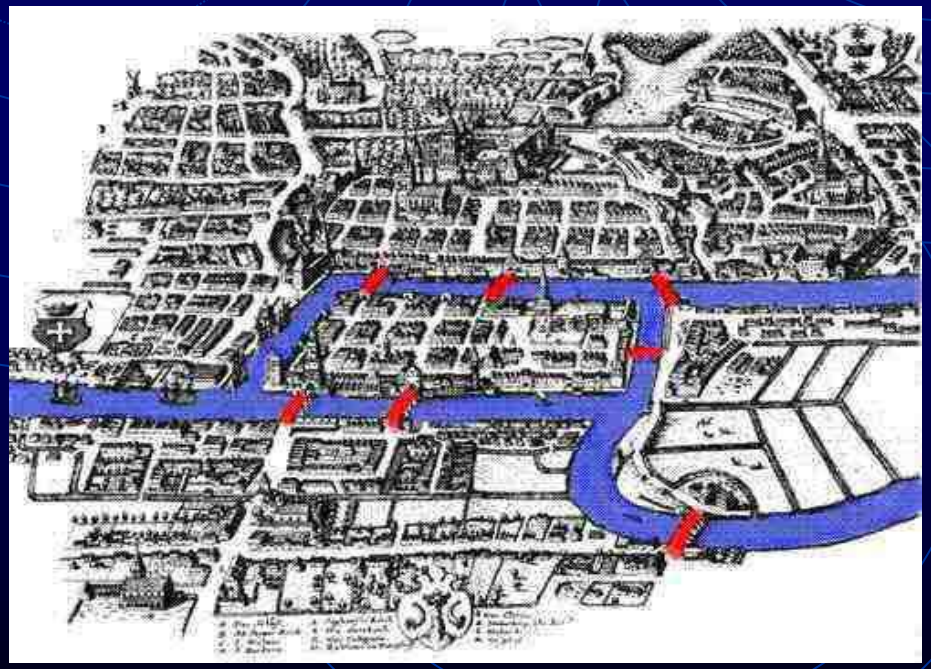
de Bruijn Graph

- In graph theory, an *n*-dimensional **De Bruijn graph** of *m* symbols is a directed graph representing overlaps between sequences of symbols. It has m^n vertices, consisting of all possible length-*n* sequences of the given symbols; the same symbol may appear multiple times in a sequence.
- Although De Bruijn graphs are named after Nicolaas Govert de Bruijn, they were discovered independently by both De Bruijn and I. J. Good. （1946）

Wikipedia

The Bridge Obsession Problem

Find a tour crossing every bridge just once
Leonhard Euler, 1735



Bridges of Königsberg

Pavel A. Pevzner

- Pavel A. Pevzner, Haixu Tang, and Michael S. Waterman. 2001. An Eulerian path approach to DNA fragment assembly. PNAS. 98: 9748–9753

2.3 Genome annotation

- Genome survey
  - Genome size/GC content/repeat content
- Gene finding
  - Coding genes
  - non-coding small RNAs
- Repeat annotation

Gene finding

- Coding genes
- non-coding small RNAs

Appearance of Genome

- One to many chromosomes
- Repeat sequences common in some genomes e.g. 35% of human are transposable elements -10% *Alu*, 14.6% LINE1 sequences
- Gene structure varies – no. and length of introns

What does 50 kb of sequence look like?

Intron-exon components of a gene

Human – very few genes - repeats

Maize – mostly repeats

Yeast – many genes (~25) – few repeats

Rice – not many gene - not few repeats

Protein-coding and non-coding sequences in genome

基因组包括基因和非编码DNA

Non-coding sequences: small RNAs (microRNA and siRNA) and long non-coding RNAs (lncRNA)

Gene finding

- Given the sequence of a genome, we would like to be able to identify:
  - Genes
  - Exon boundaries & splice sites
  - Beginning and end of translation
  - Alternative splicings
  - Regulatory elements (e.g. promoters)
- Only certain way to do this is experimentally, but computational methods can achieve reasonable accuracy quickly, and help direct experimental approaches.

Gene finding strategies

There is no (yet known) perfect method for finding genes. All approaches rely on combining various "weak signals" together and assemble into a consistent gene model

Homology method

- Gene structure can be deduced by homology
- Requires a not too distant homologous sequence

Ab initio method

- Requires two types of information
  - compositional information
  - signal information

LOCUS OSJN00244 151936 bp DNA linear PLN 14-NOV-2003
DEFINITION Oryza sativa genomic DNA, chromosome 4, BAC clone:
OSJNBa0053B21, complete sequence.

COMMENT

---------- Summary Statistics ----------

Assembly program: phrap

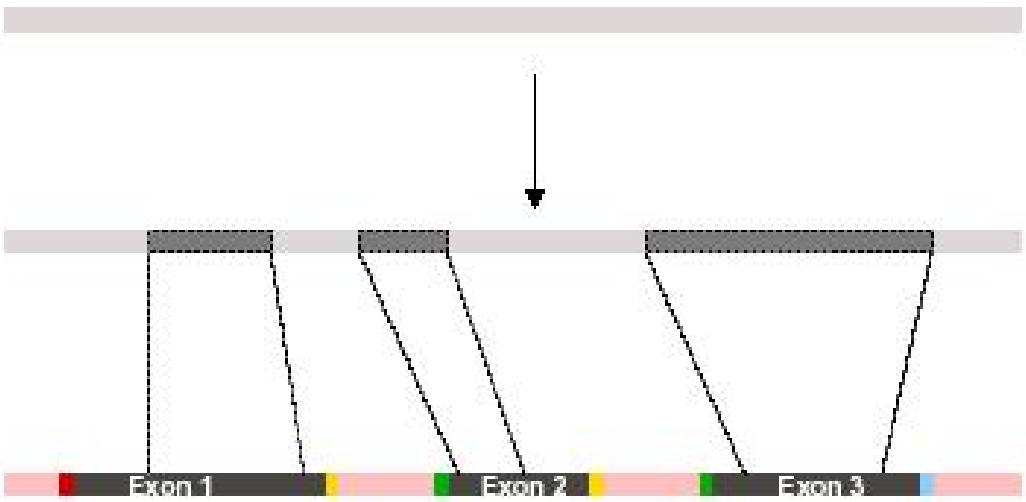
Genes were identified by a combination of several methods: Gene prediction programs including Fgenesh (http://www.softberry.com/), genscan (http://CCR-081.mit.edu/GENSCAN.html), GeneMarkHMM (http://genemark.biology.gatech.edu/GeneMark/), tRNAscan-SE (Sean Eddy, http://genome.wustl.edu/eddy/tRNAscan-SE/), searches of the complete sequence against NCBI none redundant protein database (nr) (ftp://ncbi.nlm.nih.gov/blast/db) and the EST database at NCGR.

Homology method

Principles of the homology method:

- **Coding regions evolve slower than non-coding regions**, i.e. local sequence similarity can be used as a gene finder.
- Homologous sequences reflect a common evolutionary origin and possibly a common gene structure, i.e. gene structure can be solved by homology (mRNAs, ESTs, proteins, domains).
- Standard homology search methods can be used (BLAST, Smith-Waterman, ...).
- Include ”gene syntax” information (start/stop codons, ...).

A simple view



Gene of unknown structure

Homology with a gene of known structure

Find DNA signals

ATG

{TAA,TGA,TAG}

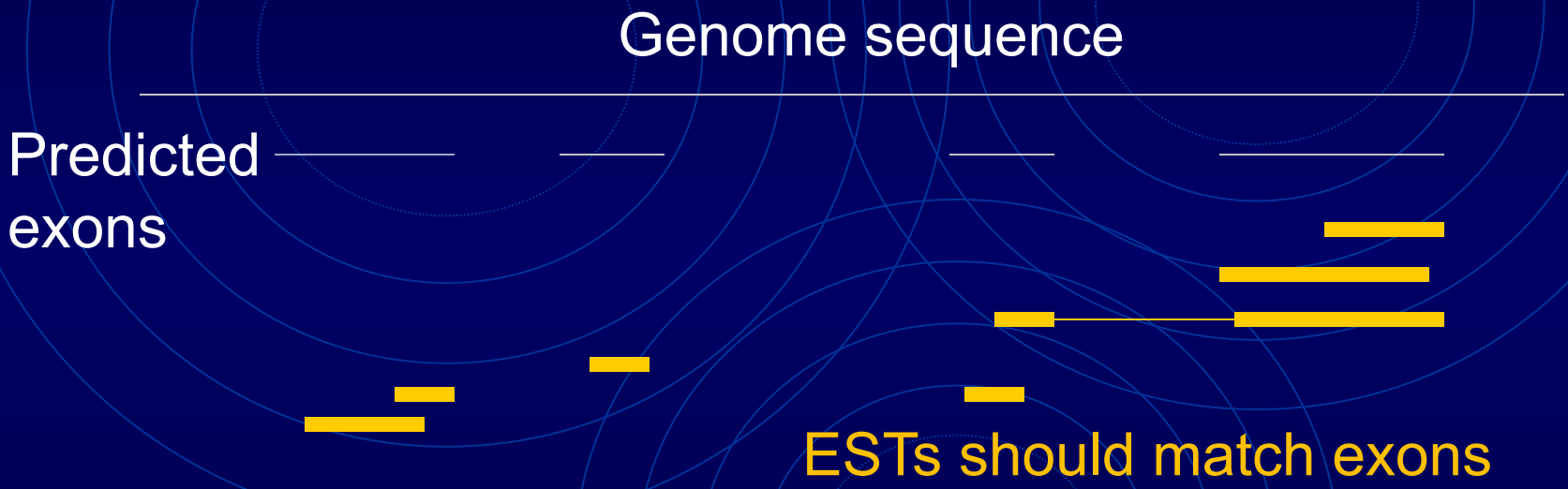
GT

AG

Inference by homology

- For exon finding, we need to find matches to
  - mRNA/cDNA sequences
  - ESTs
  - Known exons

EST/RNA-Seq reads can be helpful in confirming a gene model



may need to fill in gaps by RT PCR and often need to obtain the whole cDNA sequence

same mRNA may be spliced differently in different tissues giving a different protein or mRNA may be edited to change sequence

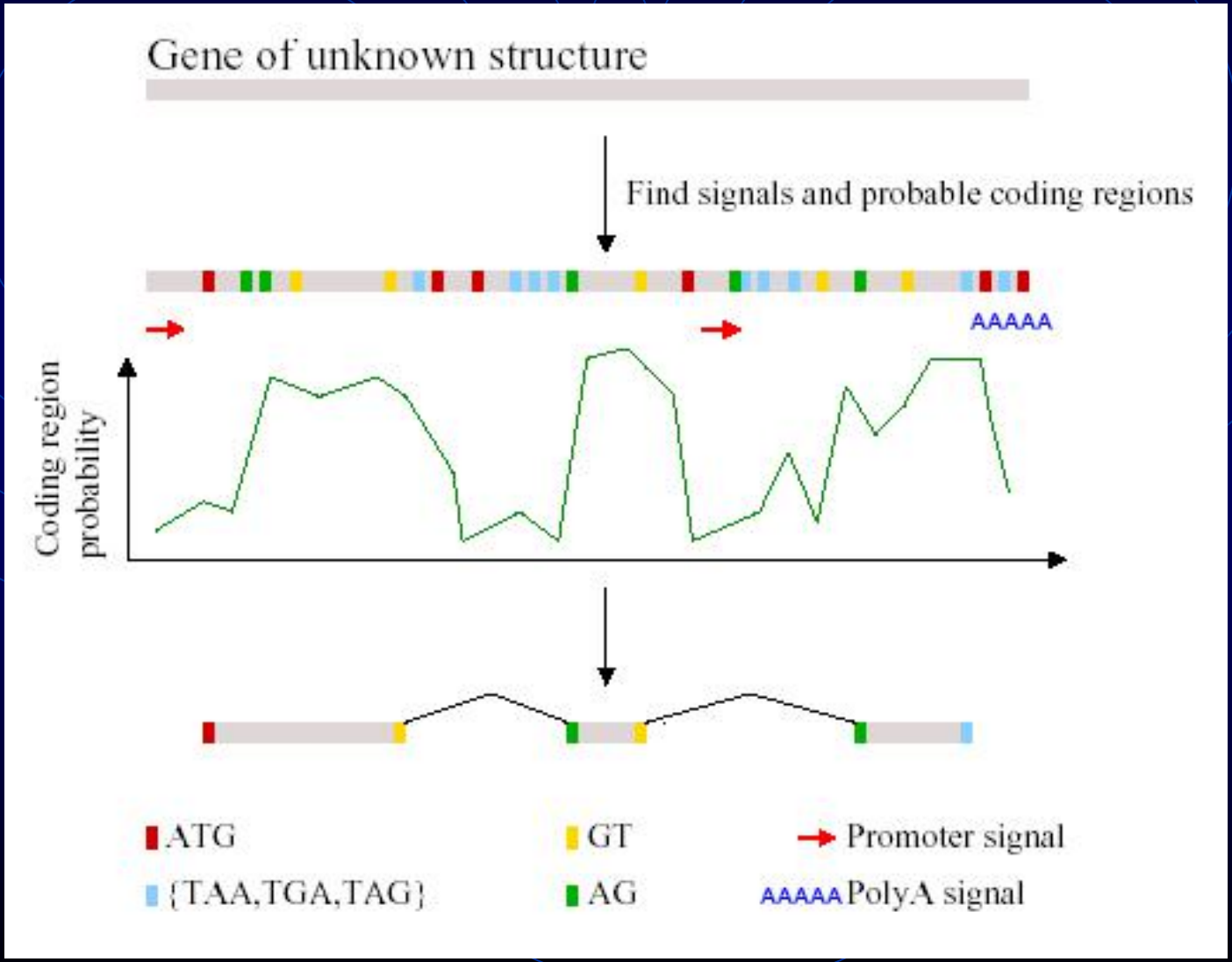
Ab initio method

Principles of the *ab initio* methods

- Integration of signal detection and coding statistics
- Signal detection and coding statistics are deduced from a training set
- Probabilistic frameworks are used to infer a probable gene structure
- A solid scoring system can be used to evaluate the predictions

•AUGUSTUS / GeneMark.hmm / FGENESH a

A simple review

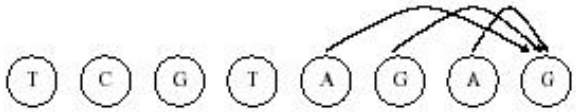


HMM

Markov Model (MM)

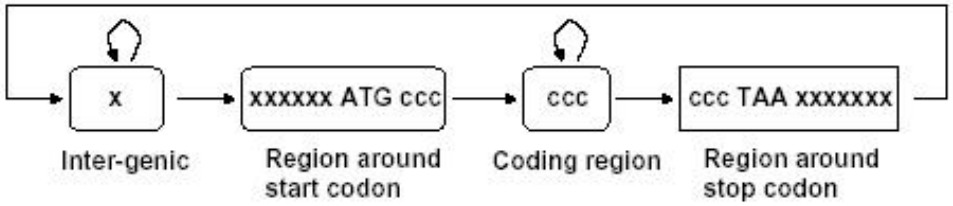
- Biological sequences can be modeled as the output of a stochastic process in which the probability for a given nucleotide to occur at position p depends on the k previous positions. This representation is called k-order Markov Model.

$$P(x_i|x_1, x_2, \ldots, x_{i-1}) = P(x_i|x_{i-k}, x_{i-(k-1)}, \ldots, x_{i-1})$$



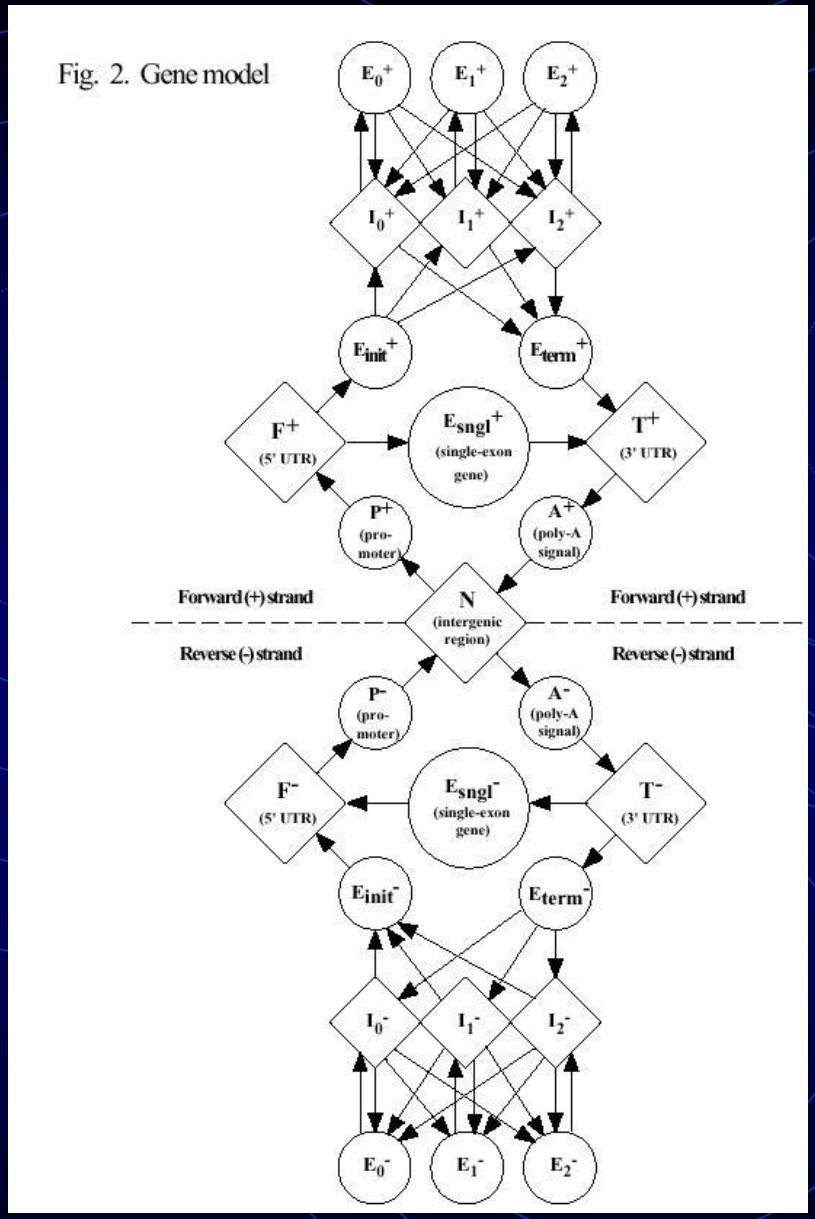
Hidden Markov Model (HMM)

- In a HMM the biological sequences are modeled as the output of a stochastic process that progresses through a series of discrete states. Each state model correspond to a Markov Model.



(Krog, 1998)

（参见BS



GENSCAN基因预测HMM模型（Burge和Karlin, 1997）

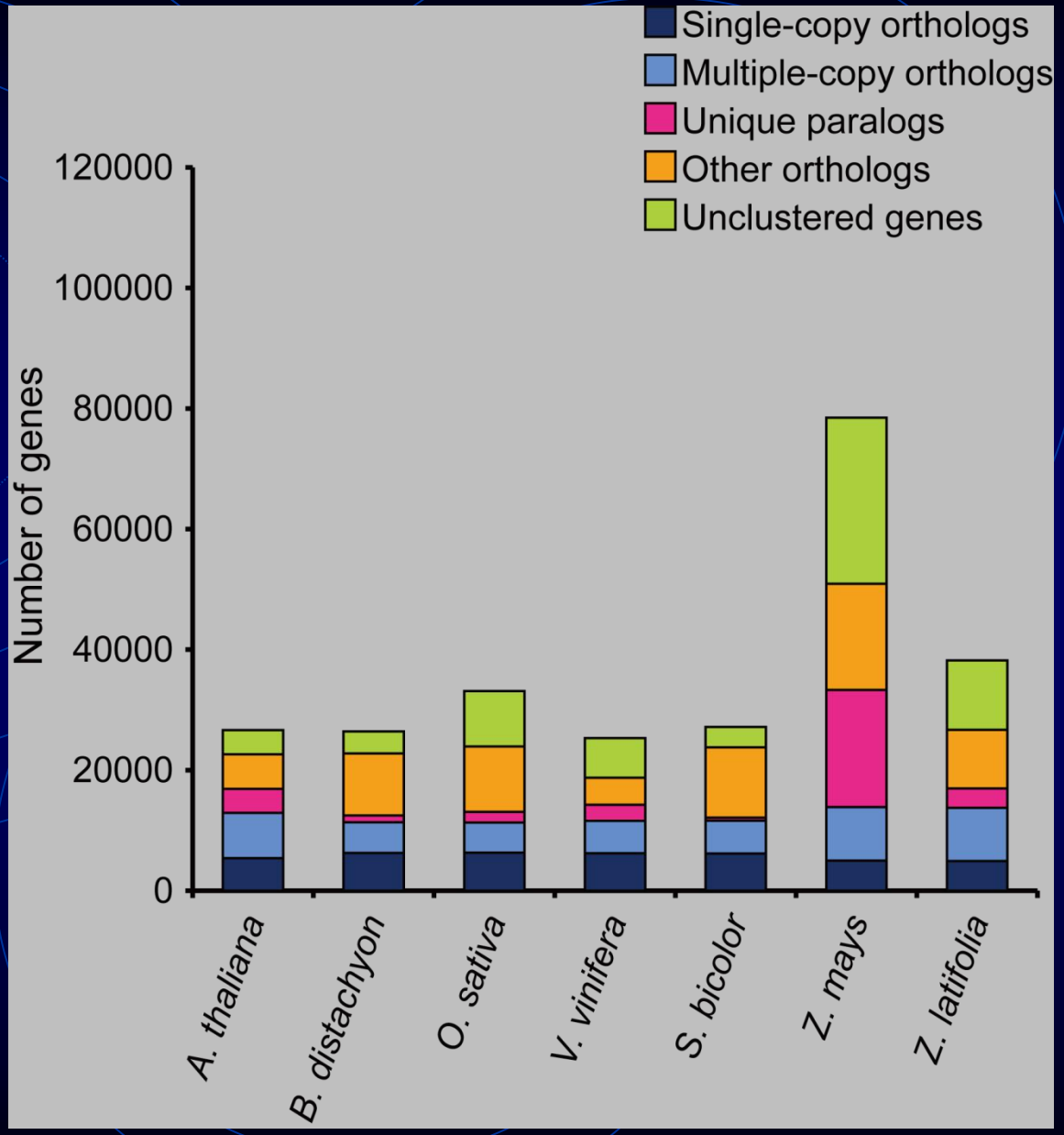
Fgenesh (www.softberry.com)

- Fgenes (Find genes) is the multiple gene prediction program based on dynamic programming;
- Fgenesh: Hidden Markov Model (HMM)-based gene prediction program (Salamov and Solovyev 2000, Genome Res)
- Fgenesh+: is a version of Fgenesh, which uses additional information from the available protein homolog. When exons predicted by Fgenesh show high similarity to a protein from the database, it is often advantageous to use this information to improve the prediction accuracy.

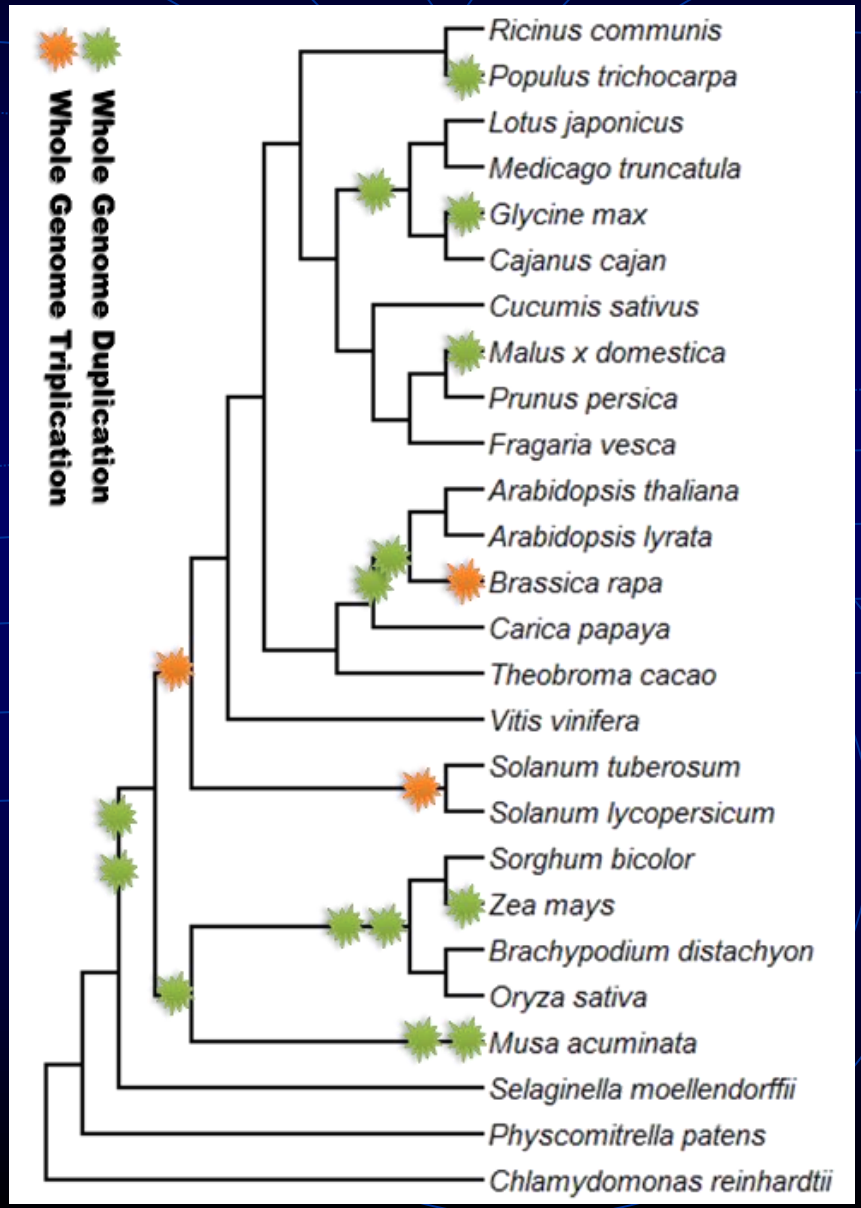
Non-coding gene finding

- microRNA(miRNA)
  - miRNA-like long hairpin
- siRNA
  - *trans*-acting siRNA (ta-siRNA)
  - Phase siRNA (phasiRNA)
- long non-coding RNA (lncRNA)
- circular RNA (circRNA)

Gene families in genome



Genome duplication/triplications in plants



Repeat annotation

- 基因组重复序列比例
- 重复序列类别

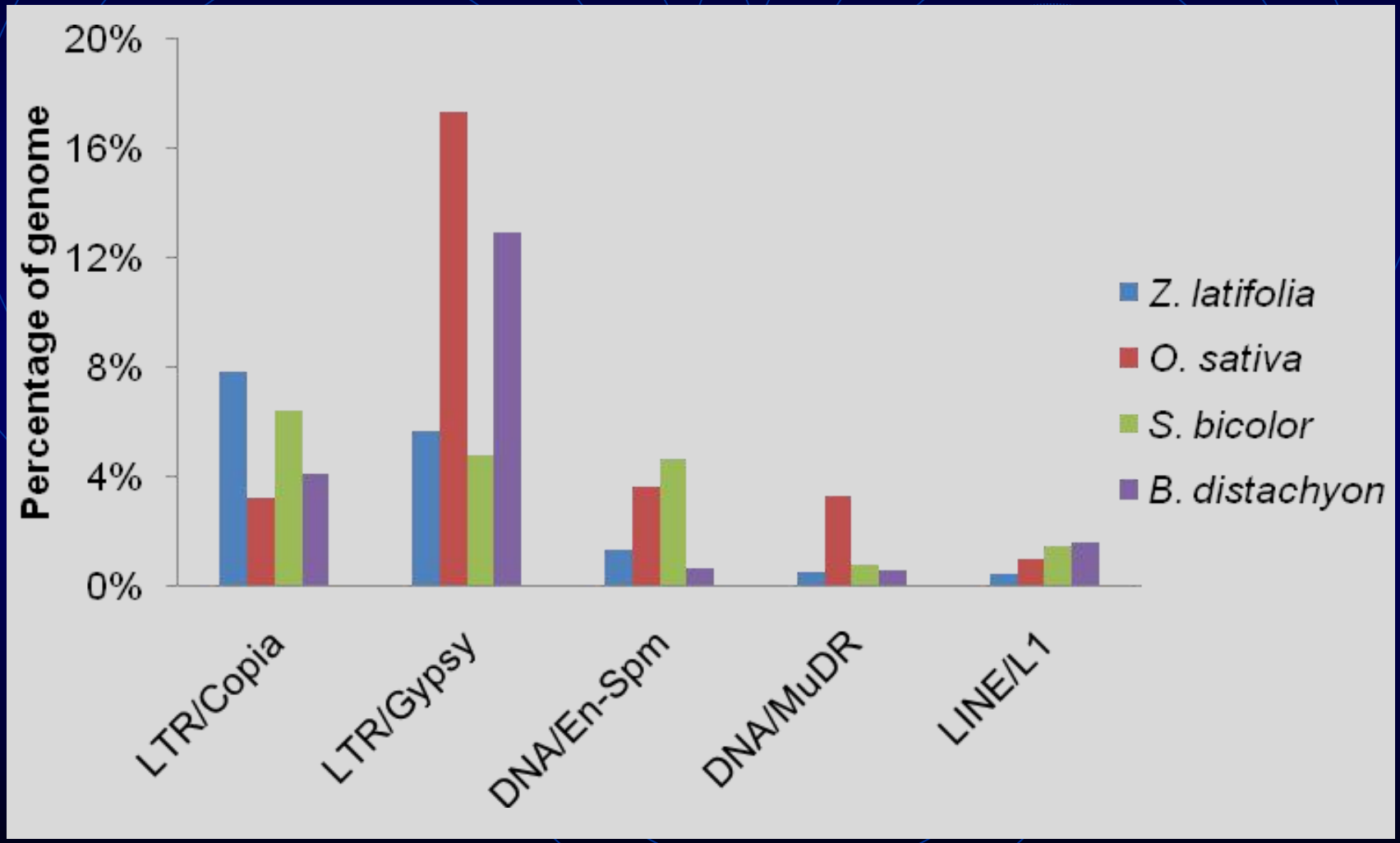
Percentage of repeat sequences in grass genomes

Super-family		*O. staiva*	*S. bicolor*	*Z. mays*	*B. distachyon*	*S. italica*	*E. crus-galli*
Retroelements	LTR	18.18	54.43	74.6	21.39	29.58	17.71
	LINE	1.12	0.04	1	1.94	1.81	1.75
	SINE	0.06	0	0	0	0.17	0.18
DNA transposons		12.96	7.46	8.6	4.77	9.38	5.58
Unknown (unclassified)		1.8	0.12	-	-	5.39	9.27
Total TEs		**28.81**	**62.0**	**84.2**	**28.1**	**46.44**	**34.5**

Statistics of TEs in the *Zizania latifolia* 'HSD2' genome

	Repbase TEs		*de novo* TEs		Combined TEs	
Type	**Length (Mb)**	**% in genome**	**Length (Mb)**	**% in genome**	**Length (Mb)**	**% in genome**
DNA	18.26	3.02	34.39	5.69	42.94	7.11
LINE	3.03	0.50	6.10	1.01	7.23	1.20
SINE	0.03	0.01	0.28	0.05	0.30	0.05
LTR	81.58	13.51	177.83	29.44	180.03	29.80
Other	0.02	0.00	0.00	0.00	0.02	0.00
Unknown	0.00	0.00	10.46	1.73	10.46	1.73
Total	102.73	17.01	221.27	36.63	227.45	37.65

Repeat elements and content



Summary

- 基因组调查测序及其目的
- 基因组测序策略
- 基因组构成及其基因注释的一般方法

- *K*-mer genome survey
- Two ways to sequence a genome
- Two ways to annotate a genome

Question/homework

- Any difference of plant genomes to human/animal genomes?