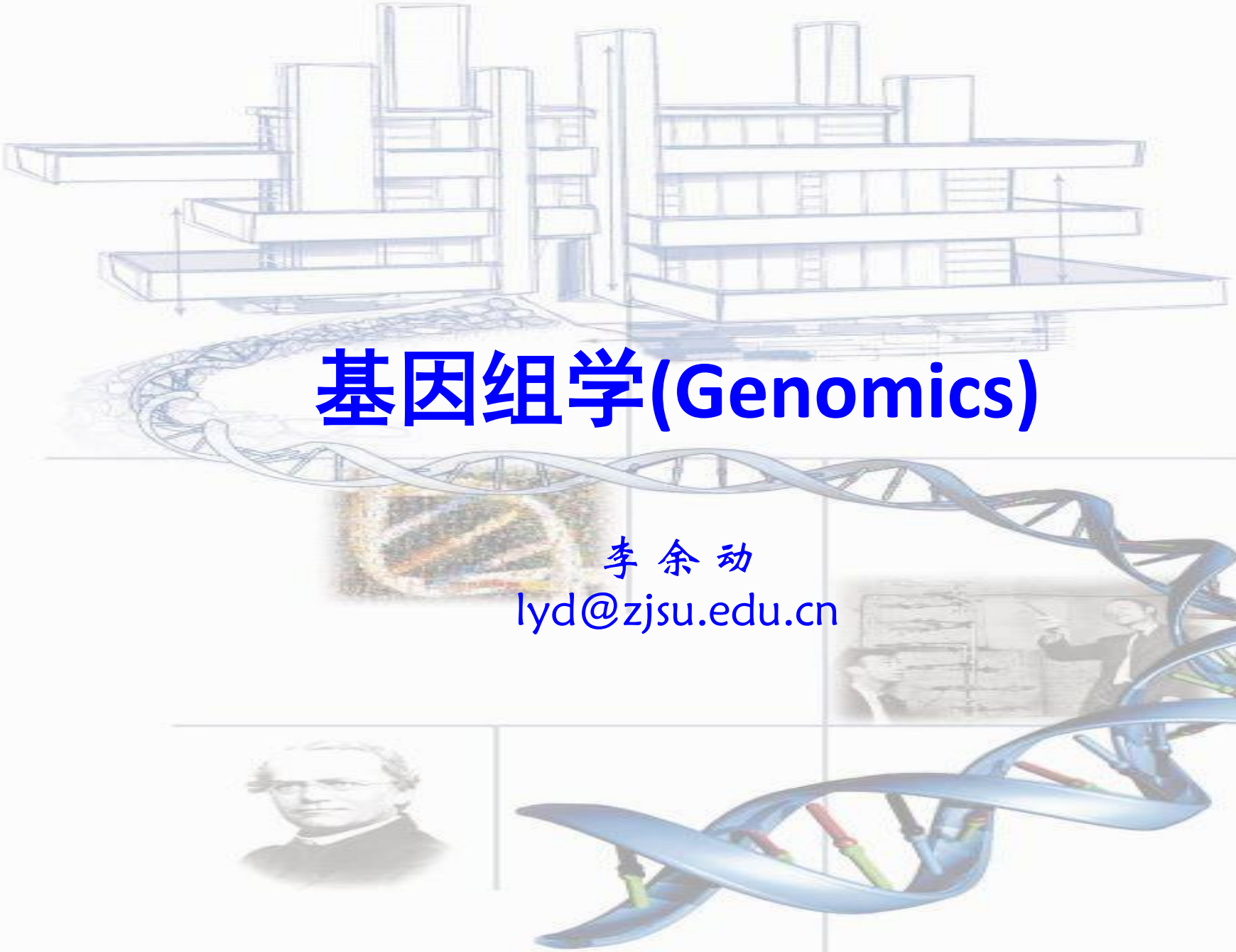


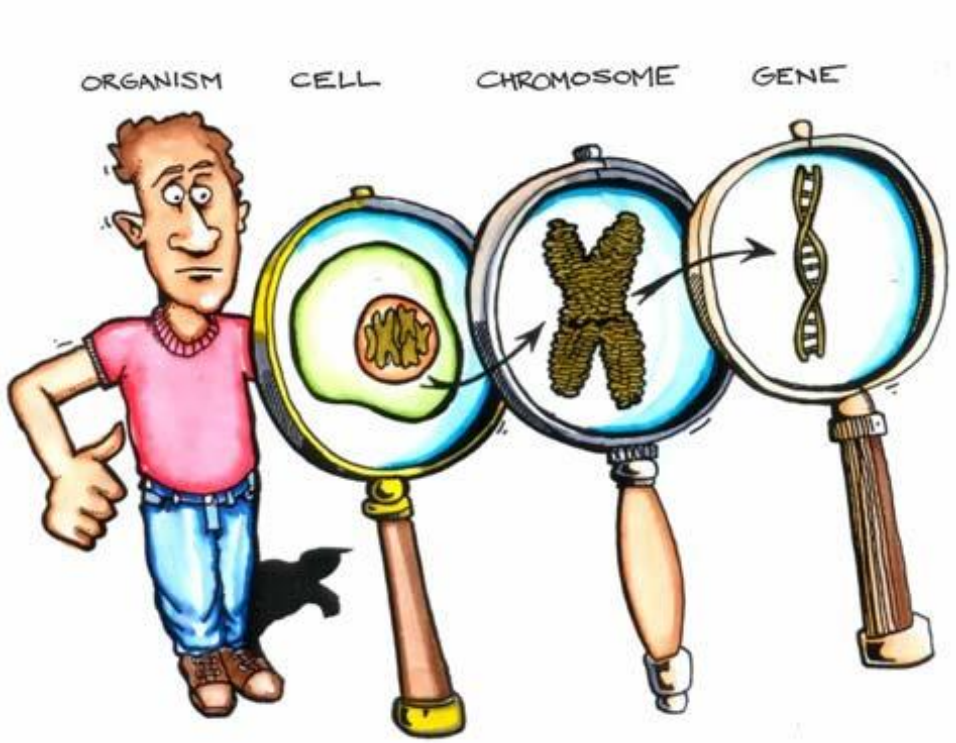
基因组学(Genomics)

李 余 动
lyd@zjsu.edu.cn

Topics

- Genomics – Introduction
- Genome Sequencing
- Genome assembly and annotation
- Comparative Genomics

Cells and DNA

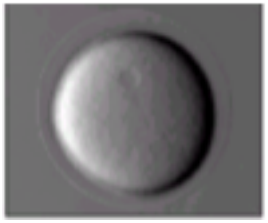


Ed Radclyffe, CSIRO

- Every cell (except a few) in an individual contains the same exact genome
- Chromosomes contain DNA
- DNA is made of 4 nucleotide bases
  - Adenine, Guanine, Cytosine & Thymine
  - AGCT sequence

Genomes to Organisms

sea urchin egg



100 µm

Sea Urchin

mouse egg

50 µm

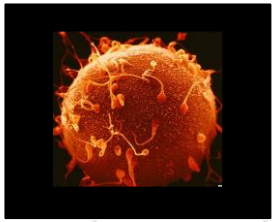
mouse

seaweed *Fucus egg*

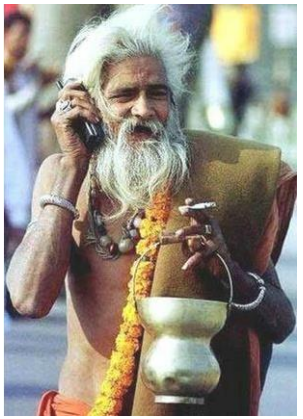
50 µm

Seaweed Fucus

Human *egg*



100 µm



Human (Wired)

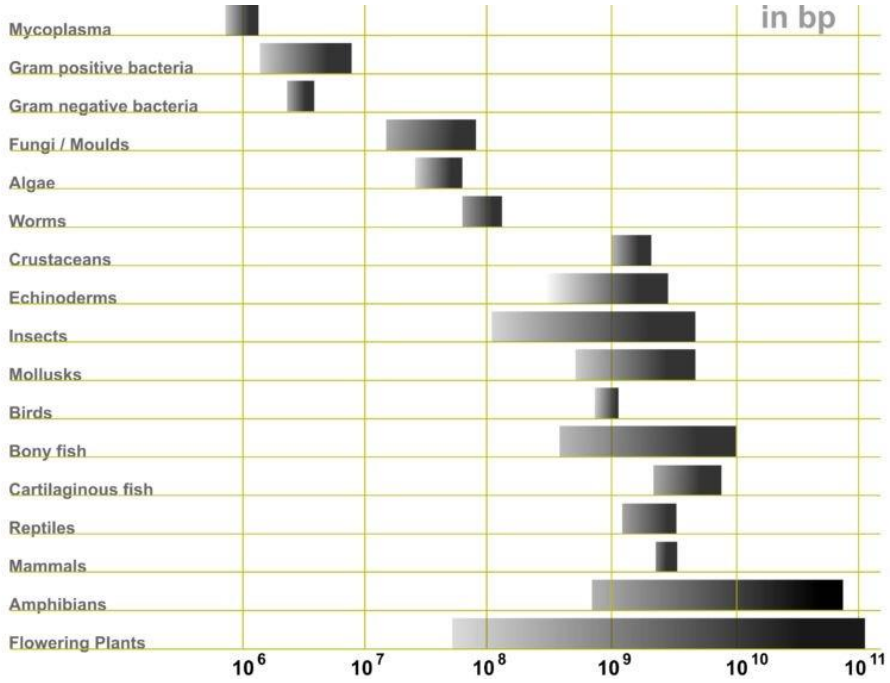
modified from MCB 4th edition

C值悖论(C-value paradox)

- C-value paradox: 基因组大小与物种的复杂性无关。
  - C-value:一个物种的核DNA含量，单位pg(皮克)
  - 基因组大小：核DNA的碱基数量，单位bp(碱基对)

1pg DNA ~ 0.978×10^9 bp

- 简单的原核/真核生物，C值大体与物种在形态学上的复杂度一致；
- 但复杂的真核生物，C值差异很大。

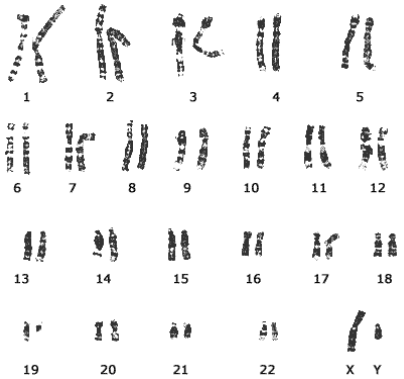


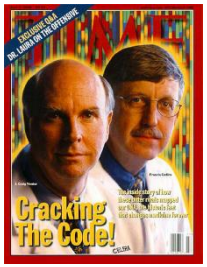
基因组学(Genomics)

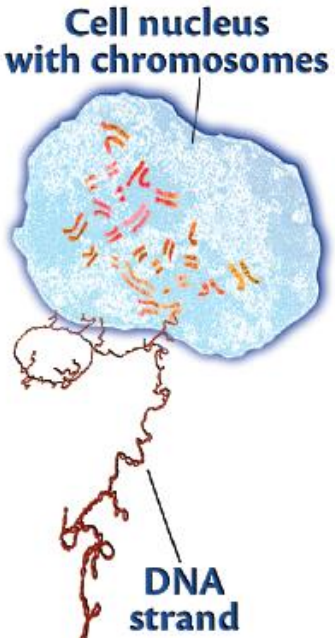
- 基因组(“genome”)是由”GENe”和”chromosOME”两个词合并而成，用于表示生物的整套染色体中的全部DNA序列。
  - 生物体的单倍体细胞中所有的遗传物质
  - 包括细胞核DNA与细胞器DNA
- Genomics:从整体水平上研究生物体全部遗传物质的结构、组成、功能及进化的学科
  - The science of genomes
  - 结构基因组学、功能基因学和比较基因组学

The Human Genome

- 23 pairs of chromosomes



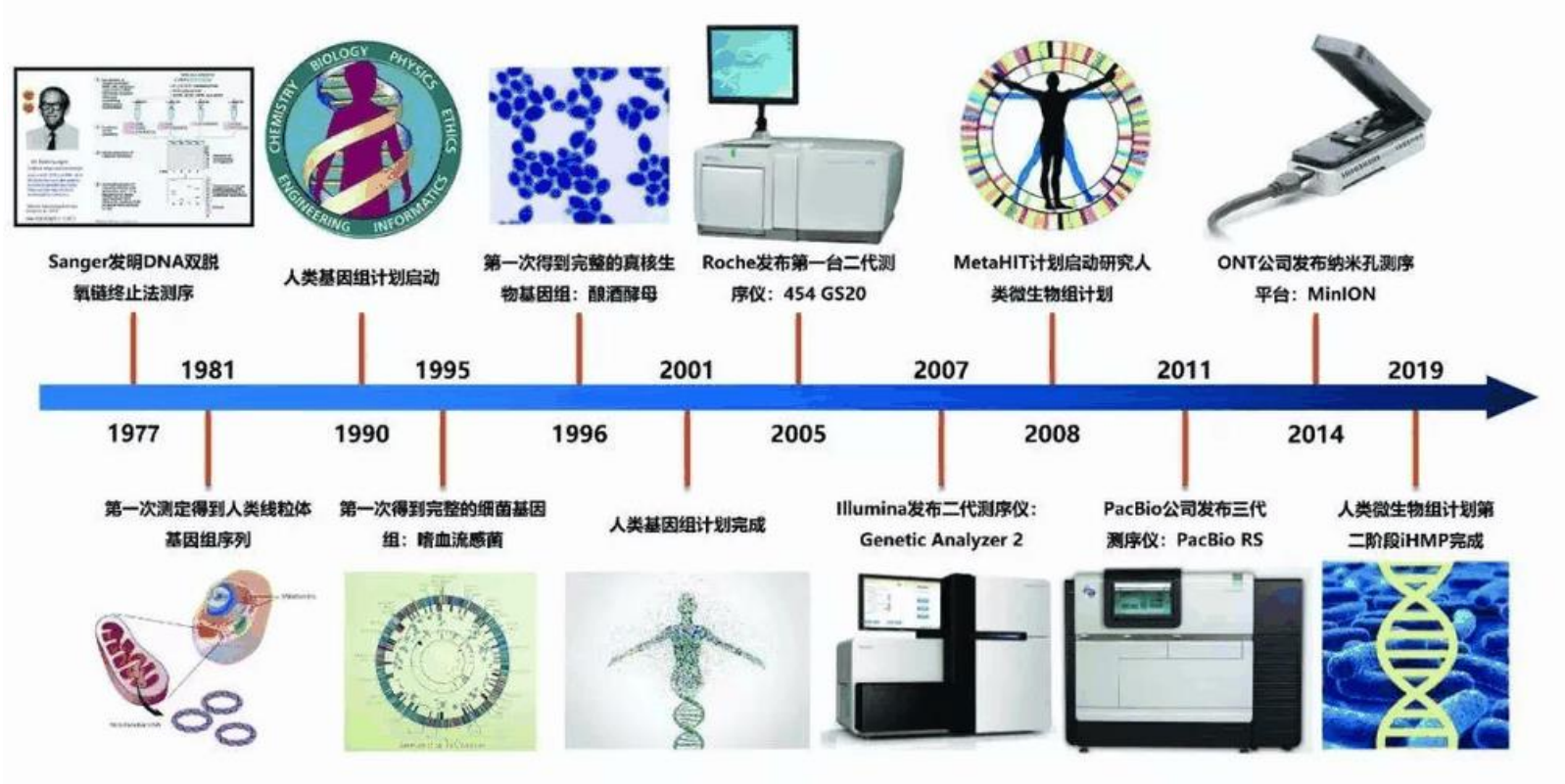




- 3 x 10^9 base pairs (or nucleotides)
  - A, G, C & T

- Human Genome Project - 10 years and cost $2.7 billion
  - ACGTGCATCTGACATTTACTG …..

基因组学的产生

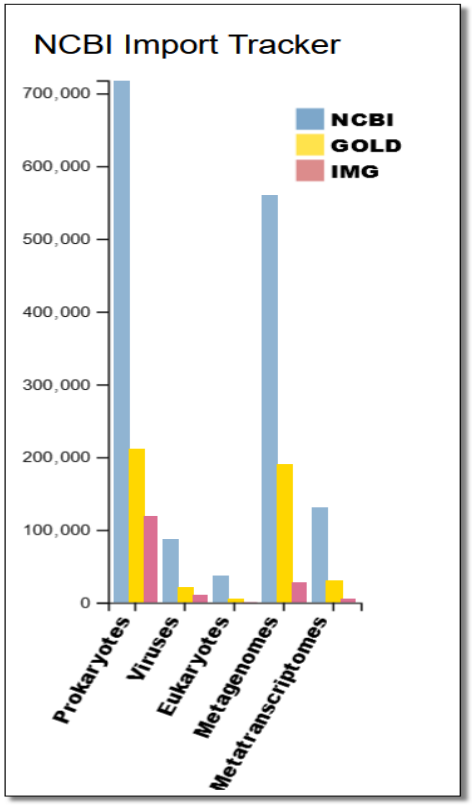


- 1977年，Sanger等提出链终止测序法，并完成噬菌体ΦX174全基因组测序
- 1996年，第一个真核生物酿酒酵母基因组完成测序
- 2001年，人类基因组计划(HGP)人类基因组草图公布
- 2005年，高通量测序技术的诞生，DNA测序片段长度和测序通量不断提高，测序成本显著下降，基因组学因此得到迅猛发展。

基因组在线数据库

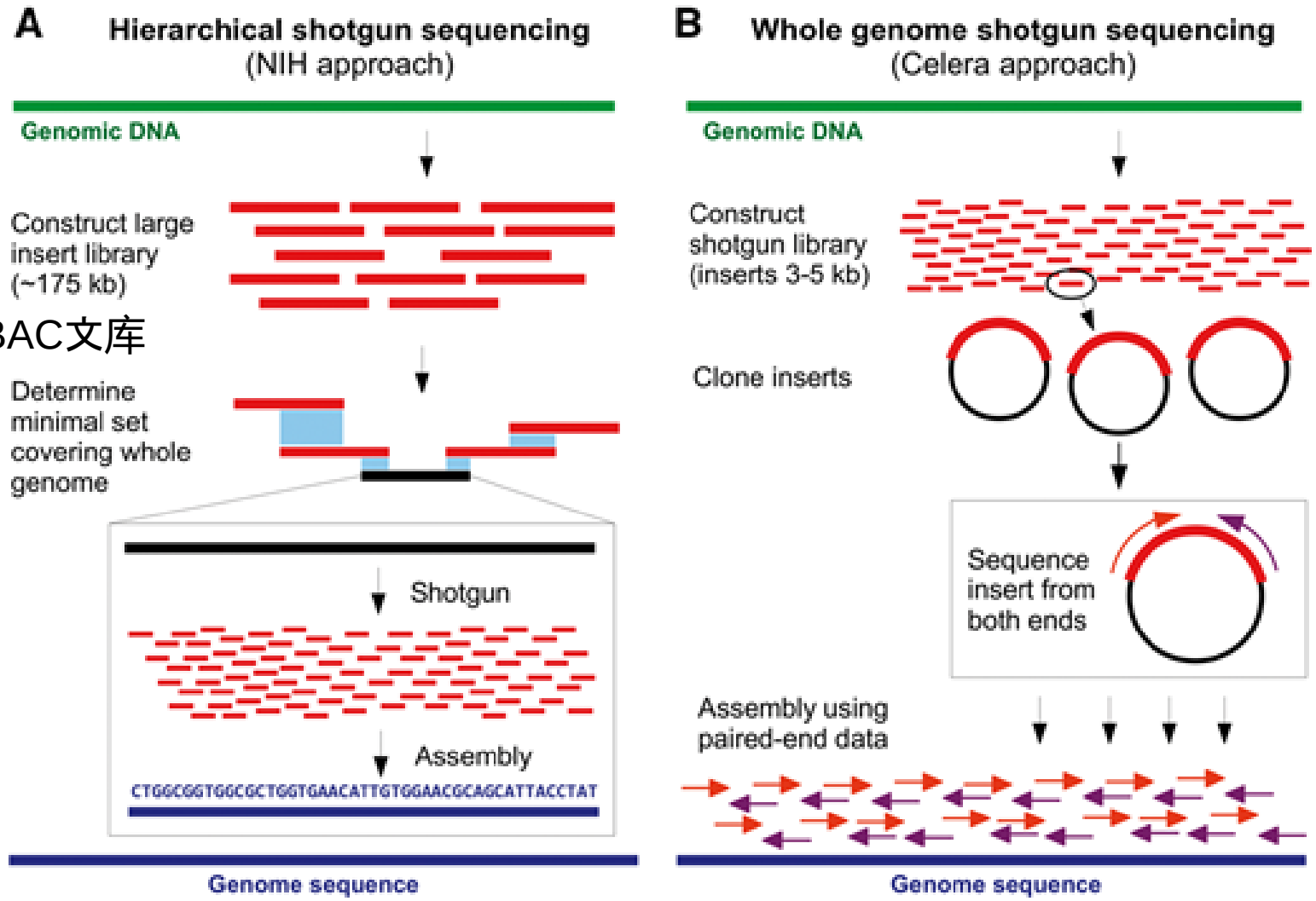
- GOLD(Genomes OnLine Database)
  - 查询基因组项目的信息，如测序类型、进展等
  - GOLD 网址：https://gold.jgi.doe.gov/

- 原核生物
- 病毒
- 真核生物
- 宏基因组
- 宏转录组



基因组测序的策略

逐步克隆法(Clone by Clone)　　全基因组鸟枪法(Whole Genome Shotgun, WGS)



基因组测序的两种策略

逐步克隆法

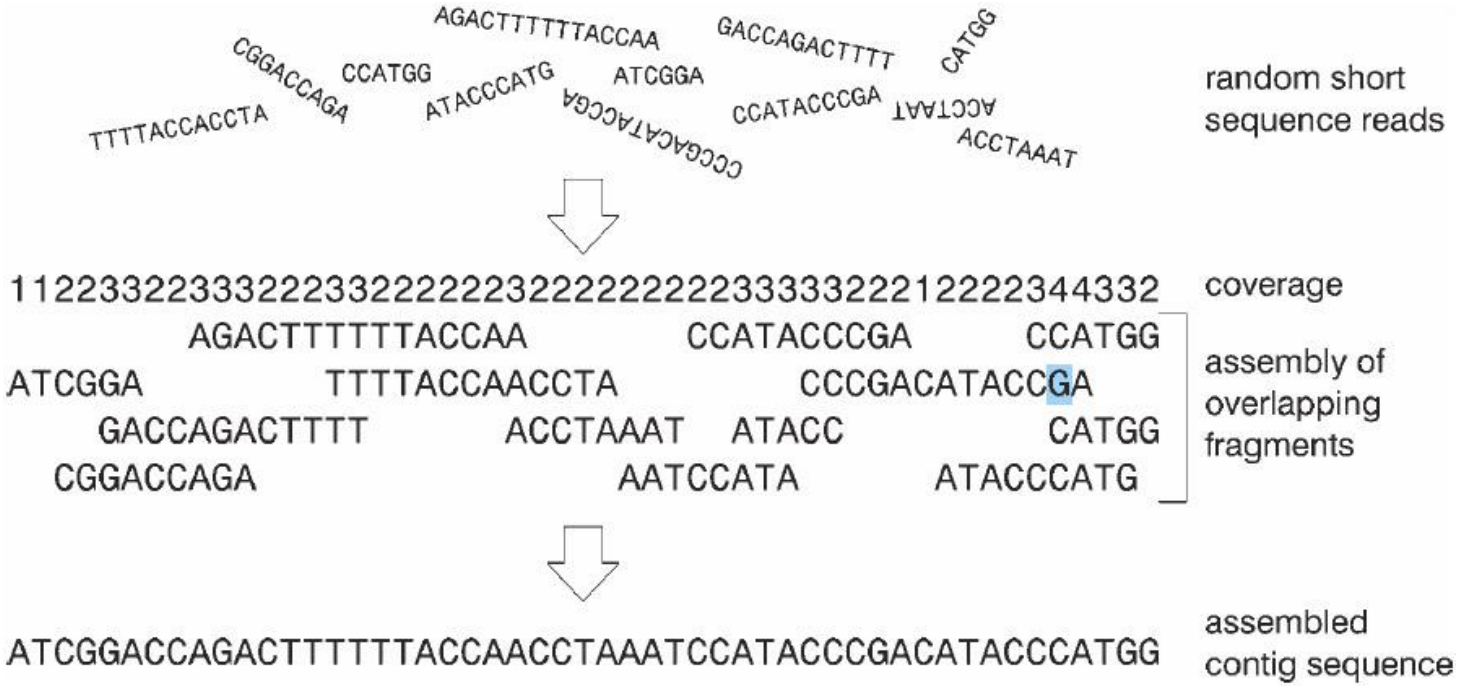
按照大分子DNA克隆绘制的物理图谱分别在单个DNA克隆内部进行测序与组装，然后将彼此相连的大分子克隆按次序搭建支架，最后以分子标记为向导将搭建好的支架锚定在基因组整合图上。

全基因组鸟枪法

将整个基因组DNA打断成小片段后将其克隆到载体中，然后随机挑取克隆进行测序，以获得的序列构建重叠群，进一步搭建序列支架，最后以分子标记为向导将序列支架锚定到基因组整合图上。

基因组组装(Genome assembly)

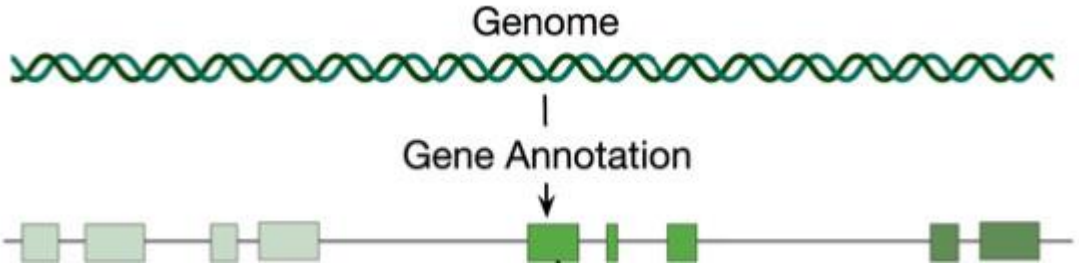
- 组装(Assembly)指通过利用一代或二代测序所得的多个读段(reads)之间的相互重叠(overlap)关系进行短序列的延长，直至无法进一步延长，得到连续序列(continue sequences)，或又称重叠群(Contigs)。



具体组装算法见第13章

基因组注释 (Genome annotation)

- 基因组注释的目标是尽可能也确定基因组中每一个核苷酸的生化和生物学功能(Brent, 2008)
  - Structural Annotation, 寻找基因（基因预测）
    - 编码基因、RNA基因、假基因等
  - Functional Annotation, 确定基因功能



计算机分析 + 实验

基因组注释信息

- GENBANK数据库的GBK格式

```
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
                     /codon_start=3
                     /product="TCP1-beta"
                     /protein_id="AAA98665.1"
                     /translation="SSIYNGISTSGLDLNNGTIADMRQLGIVESYKLKRAVVSSASEA
                     AEVLLRVDNIIRARPRTANRQHM"
     gene            <687..>3158
                     /gene="AXL2"
     mRNA            <687..>3158
                     /gene="AXL2"
                     /product="Axl2p"
     CDS             687..3158
                     /gene="AXL2"
                     /note="plasma membrane glycoprotein"
                     /codon_start=1
                     /product="Axl2p"
                     /protein_id="AAA98666.1"
                     /translation="MTQLQISLLLTATISLLHLVVATPYEAYPIGKQYPPVARVNESF
                     TFQISNDTYKSSVDKTAQITYNCFDLPSWLSFDSSSRTFSGEPSSDLLSDANTTLYFN
```

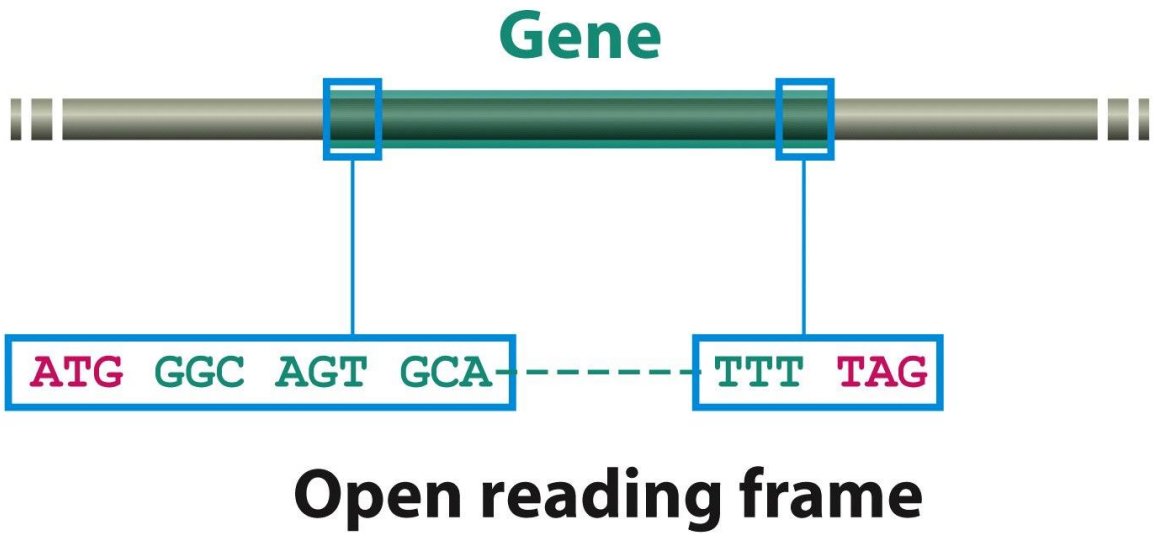
https://www.ncbi.nlm.nih.gov/genbank/samplerecord/

基因预测(Gene Prediction)方法

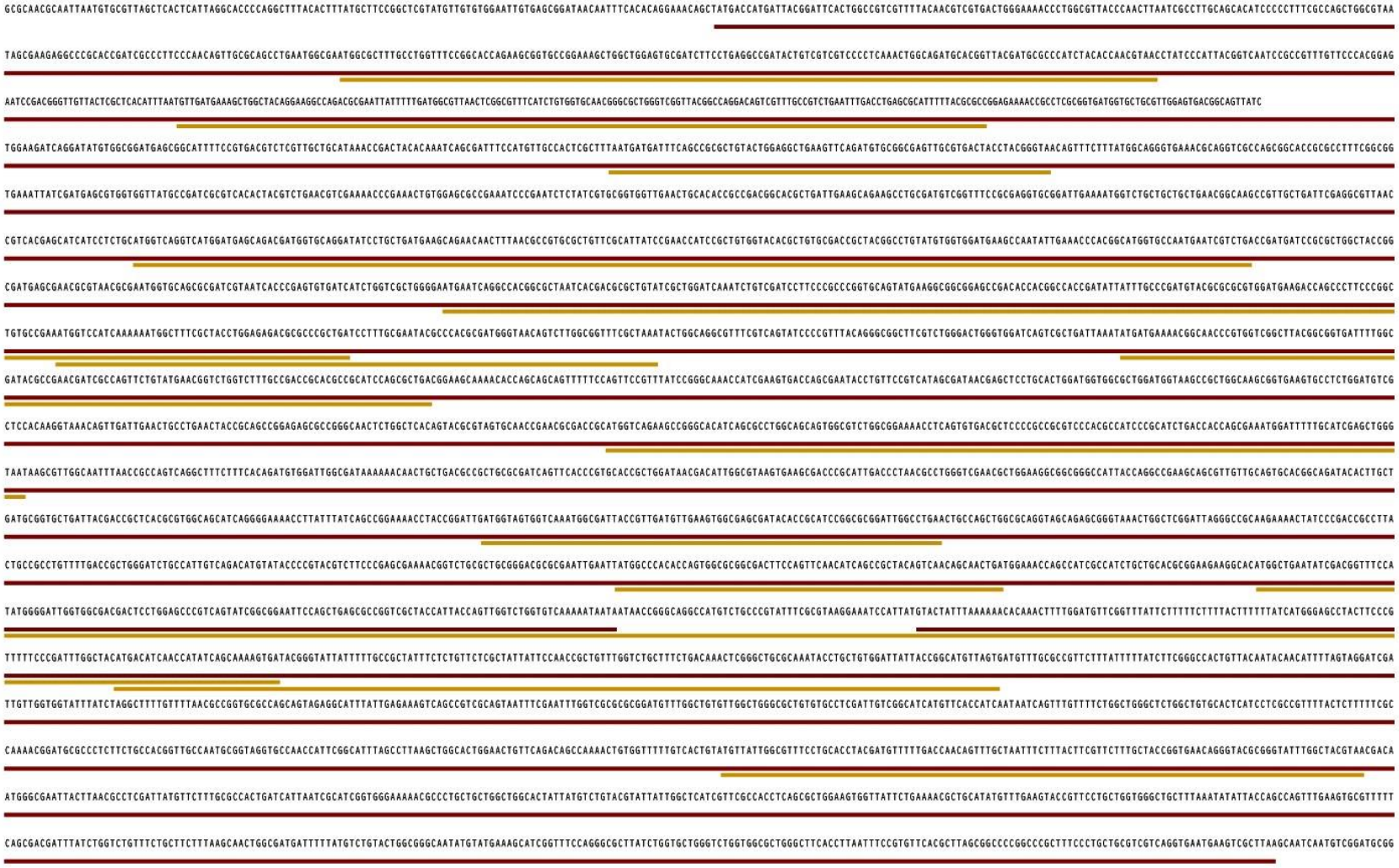
- 基因预测指通过DNA序列的信息来确定基因结构。
  - 基于基因结构特征从头预测(*ab initio* method)
    - 根据生物体内转录和翻译的信号特征来识别基因，即基因不是核苷酸的随机排列而是具有明显特征，如ORF、TATA框等
    - 软件GeneScan, Augustus, GlimmerHMM
  - 基于同源基因搜索的同源预测（Homology method）
    - 根据与已知基因的序列相似性来识别基因，主要借助同源物种的基因组序列及注释文件
    - 软件GeMoMa

一、ORF预测

- ORF(开放阅读框)：指从起始密码子(ATG)到终止密码子(TAA,TAG,TGA)的一段序列，通常代表一个编码蛋白质的基因。
- 由于遗传密码一共有64个密码子，其中3个是终止密码子(TAG/TAA/TGA)。因此，如果一条核酸序列是均匀随机分布的，那么终止密码子出现的期望次数为每21（64/3）个密码子出现一次终止密码子。
- 一般ORF >50 codons (*E. coli* ~317 codons, yeast ~483, human ~450)

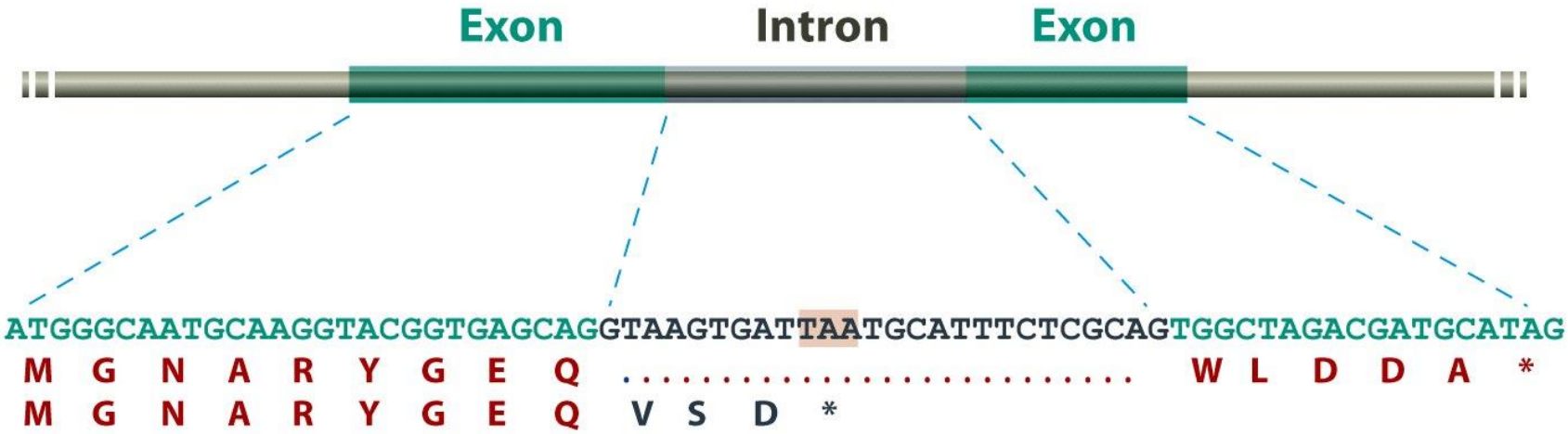


- ORF scanning is an effective way of locating genes in a bacterial genome
  - 原核生物基因无内含子，基因间DNA少，很少有重叠基因等



ORF scan are less effective with DNA of eukaryotes

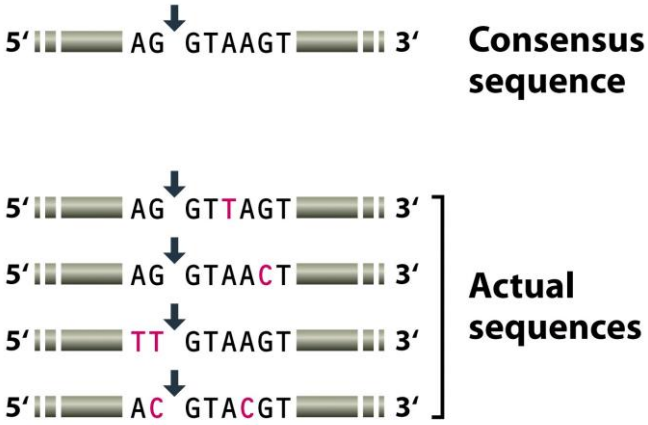
- The genes of a higher eukaryote are split by introns, and continuing the reading frame into an intron usually leads to a termination sequence.



- 基因间有大量的非编码序列
- 基因通常含有非编码的内含子，外显子长度往往小于100个密码子

真核生物基因识别的其它信号

- 外显子-内含子边界(donor/acceptor splice sites)
  - 内含子的剪切规则: GU-AG rule



很多外显子-内含子边界序列并不是上述序列，所以此规则只适用于一定的范围。

- 上游调控序列(Upstream regulatory sequence)
  - CpG islands，核糖体结合位点(RBS)

- 密码子偏倚(Condon bias)
  - 指特定生物体的基因中并不是所有密码子的使用频率都是相同的，如酵母中精氨酸的6种密码子，但其48%为AGA

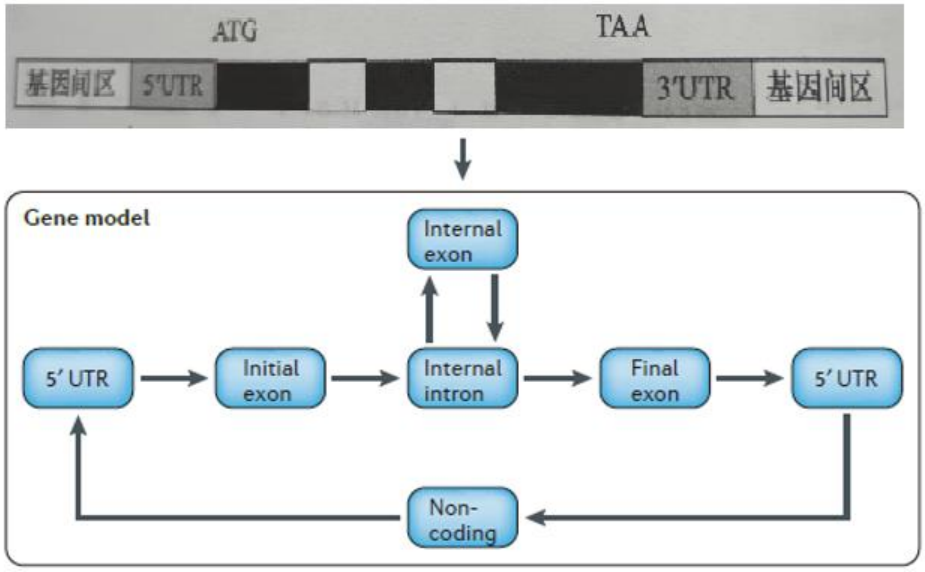
人类、果蝇和大肠杆菌中精氨酸密码使用频率的比较

Codon	Human	Drosophila	E.coli
Arginine:			
AGA	22%	10%	1%
AGG	23%	6%	1%
CGA	10 %	8%	4%
CGC	22 %	49%	39%
CGG	14 %	9%	4 %
CGU	9 %	18%	49%
Total number of arginine codons	2403	506	149
Total number of genes	195	46	149

- 预期真正的外显子有密码子偏倚，而非编码区的三联核苷酸随机排列，碱基平均分布，不会有密码子偏倚现象。
- 基因预测软件可根据已有的生物密码子偏倚的信息预测基因，所以许多基因注释程序会写明适用于哪些物种。

基因预测(Gene Prediction)算法

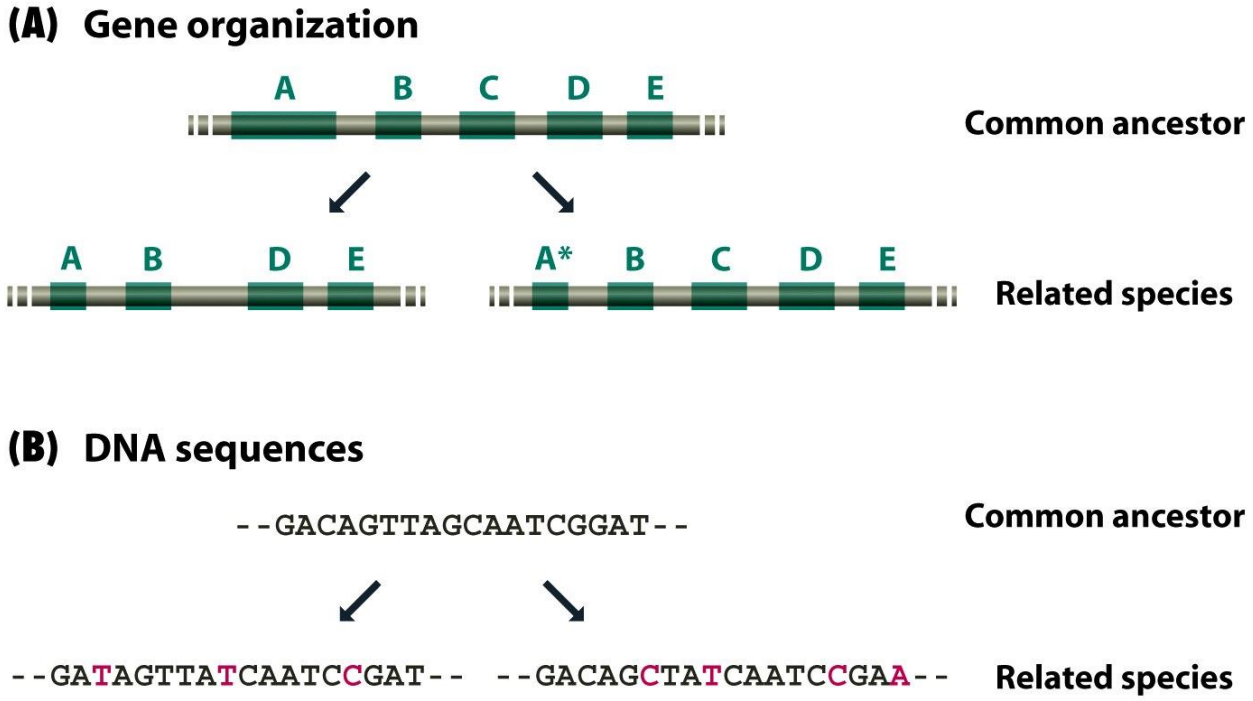
- 基因预测软件通过整合一系列不同基因特征信息，并使用动态规划算法(dynamic programming, DP)或隐马尔可夫模型(hidden markov model, HMM)来生成基因结构。
  - 如Genscan、Glimmer、GeneMark等



基因结构与基因预测HMM模型

二、同源搜索(homology search)用于识别基因

- Homology：同源性，指他们有进化相关性
  - 同源只有是或否情况，不能有百分比情况
- Similarity: 相似性，指序列相似，如80%相同
- Synteny：共线性(colinearity)，指基因的位置保守



用氨基酸序列比对有助于搜寻同源基因

- DNA只有4个碱基（A、C、T、G）
- 蛋白质有20个常用氨基酸

```
Sequence 1  GGTGAGGGTATCATCCCATCTGACTACACCTCATCGGGAGACGGAGCAGT
Sequence 2  GGTCAGGATATGATTCCATCACACTACACCTTATCCCGAGTCGGAGCAGT
Identities  *** *** *** ** *****  ********* ***  *** *********
```

两条DNA序列有80%的序列一致性

```
            G  A  P  G  M  W  L  R  L  A  A  G  S  F  E  H  A  G
Sequence 1  GGTGCACCCGGTATGTGACTGCGATTAGCAGCGGGATCATTTCAGCATGCAGGG
            * * ***** **** **** ** *** **** ***** *** ** **** ** *
Sequence 2  GATACACCCCGTATTTGACAGCAATTTGCAGGGGGATGATTGCACCATGGAGCG
            D  T  P  R  I  W  E  E  F  A  G  G  W  L  H  H  G  A
```

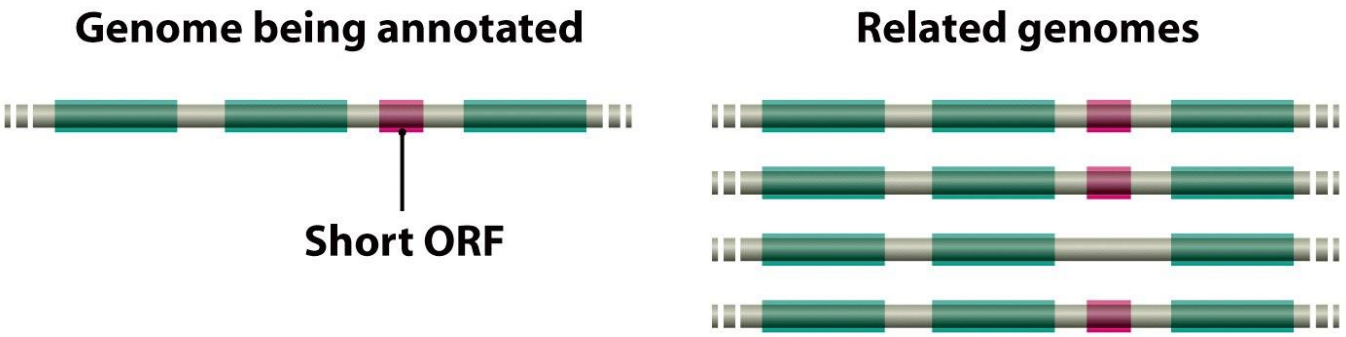
当用氨基酸序列比对时，两条序列缺少同源性就更明显(NT 78%; AA 28%)

一般认为氨基酸序列的一致性或相似性大于25%可视为同源基因。

同源搜索(homology searching)识别基因

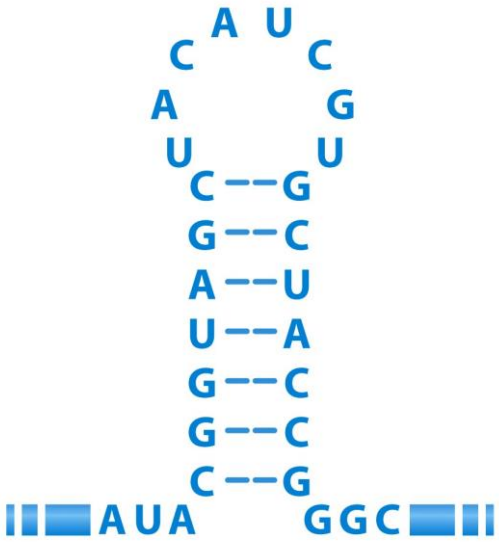
- 通过待查基因组序列与DNA数据库中的已知基因序列进行比对，从中查找与之匹配的碱基序列或蛋白质序列及其比例用于界定基因的方法。
  - 常用程序为BLAST，此程序能鉴别相似性大于30%-40%的同源基因

- 比较基因组学是一种更准确的同源基因搜寻方法
  - 运用基因组之间的同线性可以检测短ORF的真实性



三、RNA基因预测

- RNA基因指不编码蛋白质的基因，又称非编码RNA(non-coding RNAs)
  - 功能RNA：tRNA, rRNA, miRNA, lncRNA…
- RNA基因缺少显著的编码序列特征，主要是单链RNA分子内部碱基配对形成茎环结构(stem-loops)



- Cis-reg;
  - Cis-reg; IRES;
  - Cis-reg; frameshift_element;
  - Cis-reg; leader;
  - Cis-reg; riboswitch;
  - Cis-reg; thermoregulator;
- Gene;
  - Gene; CRISPR;
  - Gene; antisense;
  - Gene; miRNA;
  - Gene; rRNA;
  - Gene; ribozyme;
  - Gene; sRNA;
  - Gene; snRNA;
  - Gene; snRNA; snoRNA; CD-box;
  - Gene; snRNA; snoRNA; HACA-box;
  - Gene; snRNA; snoRNA; scaRNA;
  - Gene; snRNA; splicing;
  - Gene; tRNA;
- Intron;

RNA基因预测工具

- Rfam(https://rfam.org/)是目前最全面的RNA序列数据库
  - 通过对Rfam中上千个基因家族的多序列比对和结构特征建模，获得特定RNA家族的协方差模型(co-variance model, CM)。
- Infernal软件利用Rfam家族的协方差模型，预测miRNA, snRNA序列等ncRNAs。
- RNAmmer软件用于基因组中rRNA基因的预测注释
  - predicts 5s/8s, 16s/18s, and 23s/28s ribosomal RNA.
- tRNAscan-SE软件通过tRNA的CM模型进行tRNA预测，是基因组分析的标准软件。

Sequence of one of the *Escherichia coli* $tRNA^{leu}$ genes

5' GCCGAAGTGGCGAAATCGGTAGTCGCAGTTGATTCAAAATCAACCGTAGAAATACGTGCCGGTTCGAGTCCGGCCTTCGGCACCA 3'

四、基因功能注释

- 现存生物的不同种属之间具有功能或结构相似的同源基因，它们在起源上一脉相承，其间存在保守的序列组成。

序列保守性 ⟷ 功能保守性

- e.g., the yeast gene *SGS1* coded for DNA helicase that are required for transcription of rRNA genes and for DNA replication. Yeasts with a mutant SGS1 gene live for shorter periods than normal yeasts and display accelerated onset-of-aging indicators such as sterility.

Table 5.1 Examples of human disease genes that have homologs in *Saccharomyces cerevisiae*

Human disease gene	Yeast homolog	Function of the yeast gene
Amyotrophic lateral sclerosis	*SOD1*	Protection against superoxide (O_2^-)
Ataxia telangiectasia	*TEL1*	Codes for a protein kinase
Colon cancer	*MSH2, MLH1*	DNA repair
Cystic fibrosis	*YCF1*	Metal resistance
Myotonic dystrophy	*YPK1*	Codes for a protein kinase
Type 1 neurofibromatosis	*IRA2*	Codes for a regulatory protein
Bloom's syndrome, Werner's syndrome	*SGS1*	DNA helicase
Wilson's disease	*CCC2*	Copper transport?

同源搜索(homology searching)进行基因功能注释

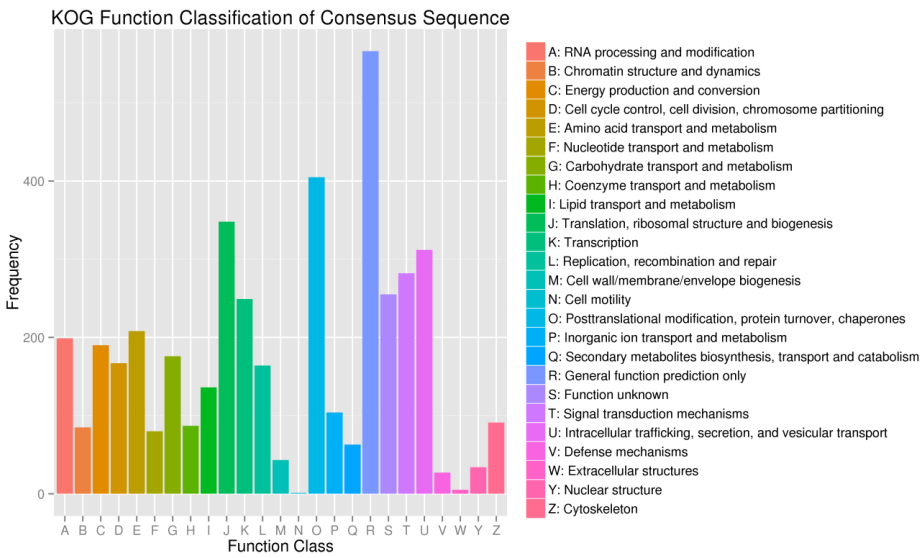
- 通过将待注释序列与数据库中已知功能的序列进行BLAST比对分析，依据序列的相似性初步判断其功能。

- 基于假设: 序列相似 = 同源 = 功能相似。
  - 此假设造成大量错误，可通过选择更严格的同源性指标(如Identity、 E-value、Coverage等)的阈值。

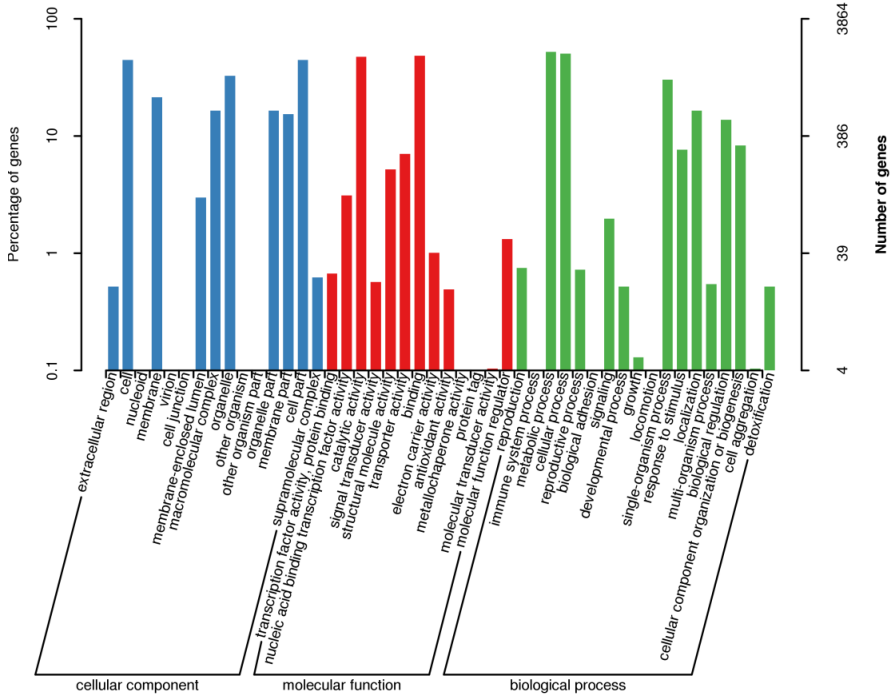
- 常用的功能注释数据库有GenBank的NR (non-redundant)，Swiss-Prot，InterPro，COG (Clusters of Orthologs)，GO (Gene Ontology)，KEGG(Kyoto Encyclopedia of Genes and Genomes)等。

基因功能注释

COG/KOG注释



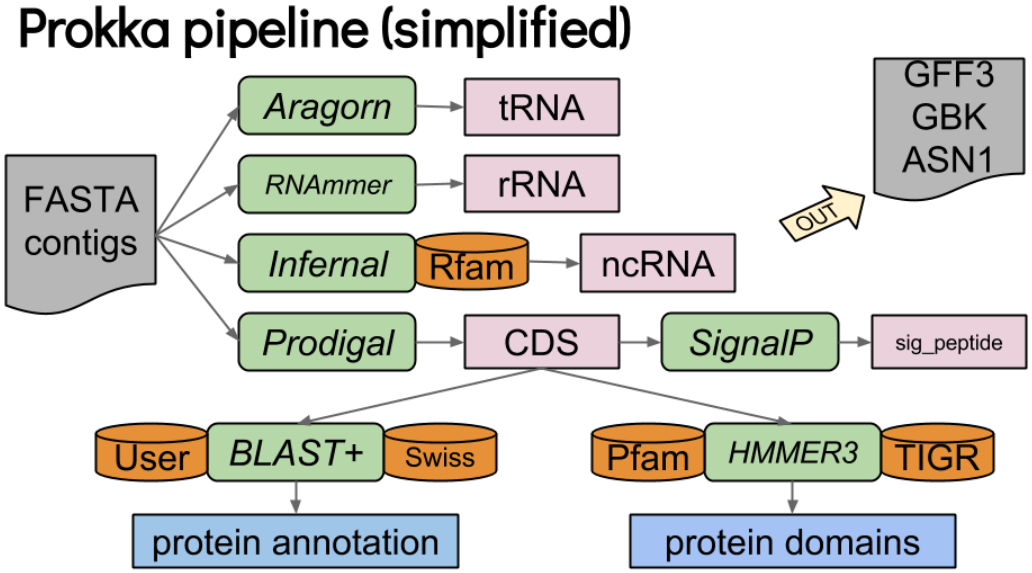
GO注释



GO(Gene Ontology)是一种用于描述基因和蛋白质功能的分类系统，提供了一套标准化的词汇和概念，涵盖了细胞组分(Cellular component)、分子功能(Molecular function)和生物学过程(Biological process)三个方面。

原核基因组注释流程： Prokka

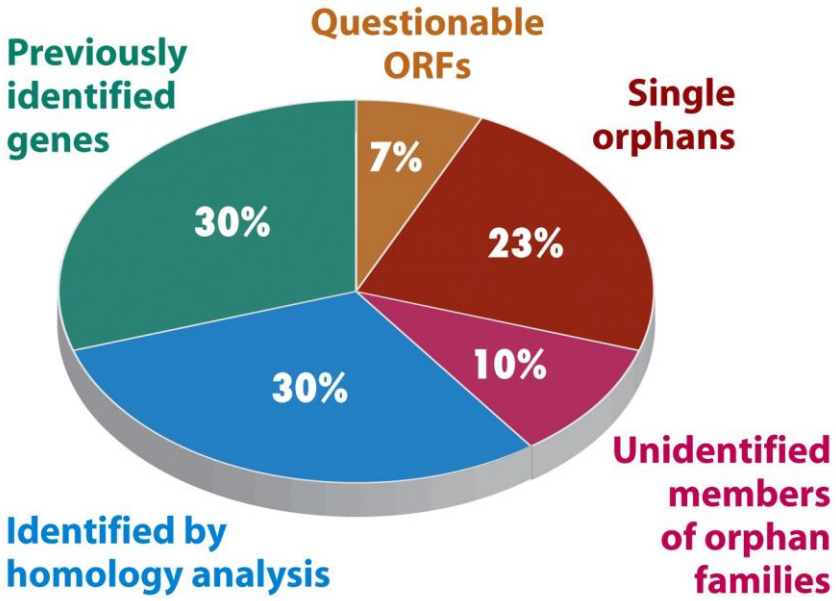
- Prokka原核基因组注释的分析流程，包括基因鉴定、功能注释和基因组配套注释文件生成等。
- Prokka协调一套现有的软件工具，可以对原核基因组和宏基因组进行快速高效的功能注释。



Torsten S., prokka: rapid prokaryotic genome annotation, *Bioinformatics* 2014

Case study: Annotation of the *Saccharomyces cerevisiae* genome sequence

- The *S. cerevisiae* genome was completed in 1996, initial analysis identified 6274 ORFs of 100 codons and longer, but it's reduced to 6120 now.
- **Orphan families**: the yeast genes had homologs in the databases, but the functions of these homologs were unknown.
- **Single orphans**: the yeast genes had no homologs in the databases, but they looked like genes and were unique.



Case study: Annotation of the *Saccharomyces cerevisiae* genome sequence

- Although there are just 6274 ORFs of 100 codons or longer in yeast genome, there are over 100,000 ORFs of 15 codons or more.
  - A few short genes have been identified using comparison with other yeast genomes.

- Approximately 55% of all yeast genes now have a well-characterized function.
- Another 2000 genes (33%) have functions that been assigned on the basis of homology analysis.
- 500 ORFs are thought to be genuine genes but have no assigned function.
- 300 ORFs that may not be real genes.

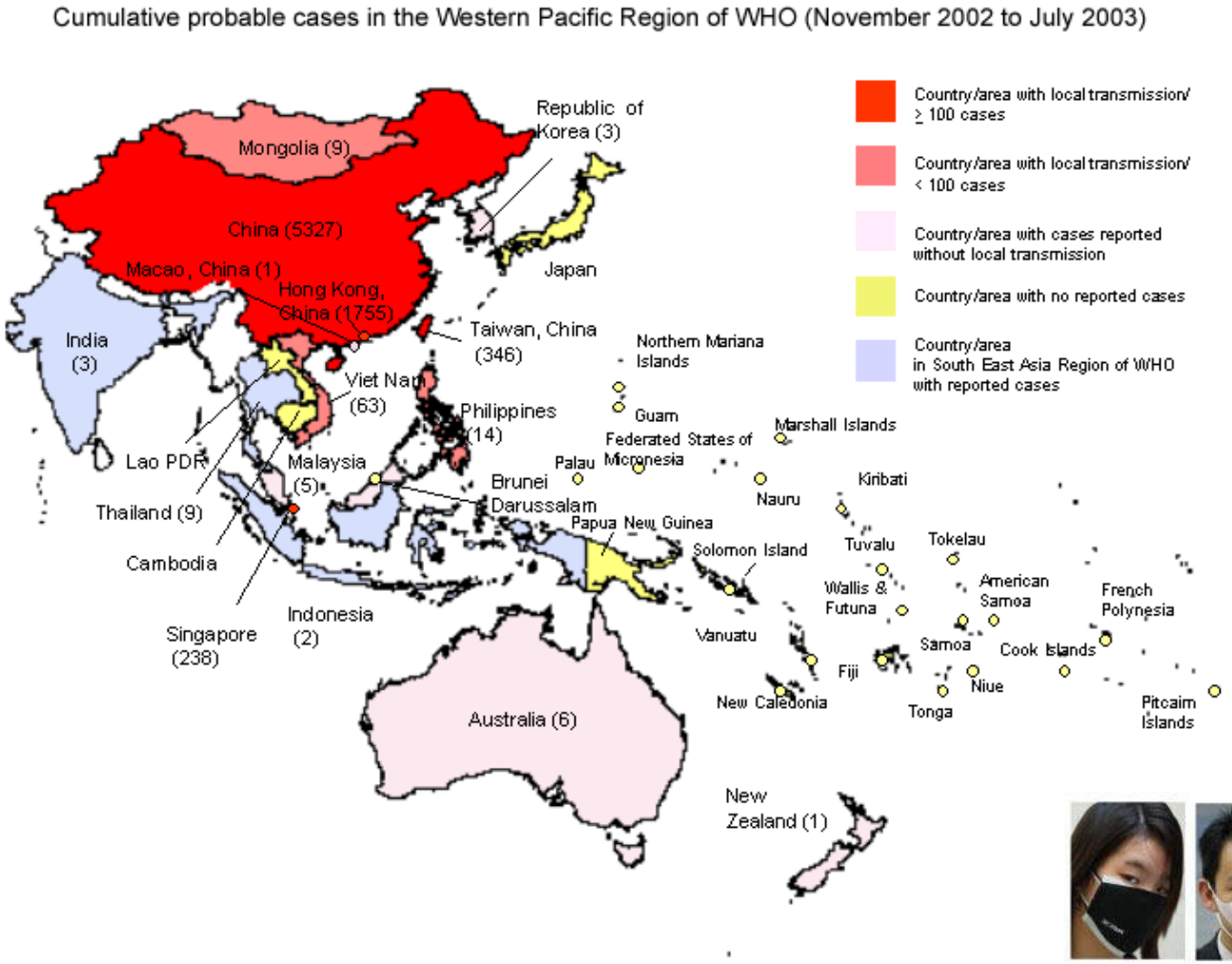
Topics

- Genomics – Introduction
- Genome sequencing
- Genome assembly and annotation
- **Comparative Genomics**

Comparative Genomics

- **比较基因组学(Comparative Genomics)是基于基因组图谱和测序基础上，对已知的基因和基因组结构进行比较，来了解基因的功能、表达机理和物种进化的学科。**

SARS病毒疫情



SARS流行病分布图

The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 MAY 15, 2003 VOL. 348 NO. 20

A Novel Coronavirus Associated with Severe Acute Respiratory Syndrome

Thomas G. Ksiazek, D.V.M., Ph.D., Dean Erdman, Dr.P.H., Cynthia S. Goldsmith, M.S., Sherif R. Zaki, M.D., Ph.D., Teresa Peret, Ph.D., Shannon Emery, B.S., Suxiang Tong, Ph.D., Carlo Urbani, M.D.,* James A. Comer, Ph.D., M.P.H., Wilina Lim, M.D., Pierre E. Rollin, M.D., Scott F. Dowell, M.D., M.P.H., Ai-Ee Ling, M.D., Charles D. Humphrey, Ph.D., Wun-Ju Shieh, M.D., Ph.D., Jeannette Guarner, M.D., Christopher D. Paddock, M.D., M.P.H.T.M., Paul Rota, Ph.D., Barry Fields, Ph.D., Joseph DeRisi, Ph.D., Jyh-Yuan Yang, Ph.D., Nancy Cox, Ph.D., James M. Hughes, M.D., James W. LeDuc, Ph.D., William J. Bellini, Ph.D., Larry J. Anderson, M.D., and the SARS Working Group†

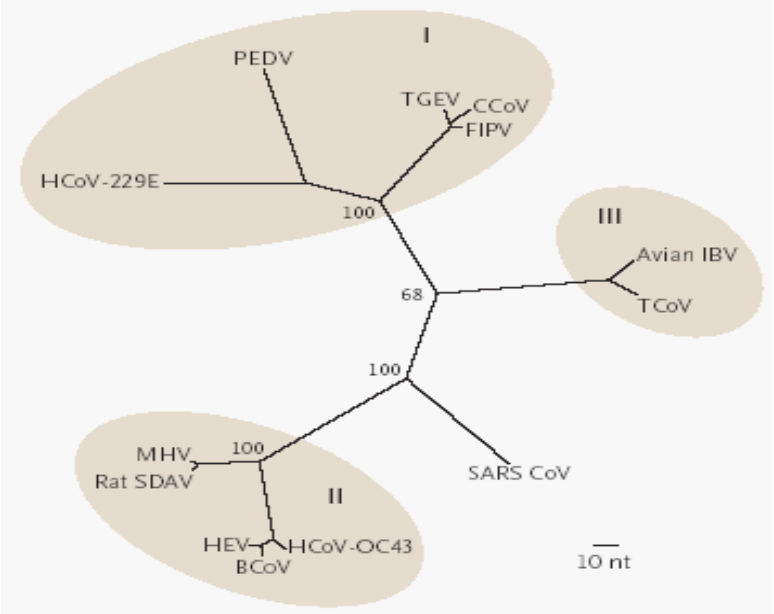


Figure 3. Estimated Maximum-Parsimony Tree Based on the Sequence Alignment of 405 Nucleotides of the Coronavirus Polymerase Gene Open Reading Frame 1b (Nucleotide Numbers 15173 to 15578 Based on Bovine Coronavirus Complete Genome Accession Number NC_003045) Comparing SARS Coronavirus with Other Human and Animal Coronaviruses.

The three major coronavirus antigenic groups (I, II, and III), represented by human coronavirus 229E (HcoV-229E), canine coronavirus (CCoV), feline infectious peritonitis virus (FIPV), porcine transmissible gastroenteritis virus (TGEV), porcine epidemic diarrhea virus (PEDV), human coronavirus OC43 (HCoV-OC43), bovine coronavirus (BCoV), porcine hemagglutinating encephalomyelitis virus (HEV), rat sialodacryoadenitis virus (SDAV), mouse hepatitis virus (MHV), turkey coronavirus (TCoV), and avian infectious bronchitis virus (IBV), are shown shaded. Bootstrap values (from 100 replicates) obtained from a 50 percent majority rule consensus tree are plotted at the main internal branches of the phylogram. Branch lengths are proportionate to nucleotide differences.

Isolation and Characterization of Viruses Related to the SARS Coronavirus from Animals in Southern China

Y. Guan,[1*†] B. J. Zheng,[1*] Y. Q. He,[2] X. L. Liu,[2] Z. X. Zhuang,[2] C. L. Cheung,[1] S. W. Luo,[1] P. H. Li,[1] L. J. Zhang,[1] Y. J. Guan,[1] K. M. Butt,[1] K. L. Wong,[1] K. W. Chan,[3] W. Lim,[4] K. F. Shortridge,[1] K. Y. Yuen,[1] J. S. M. Peiris,[1] L. L. M. Poon[1]

A novel coronavirus (SCoV) is the etiological agent of severe acute respiratory syndrome (SARS). SCoV-like viruses were isolated from Himalayan palm civets found in a live-animal market in Guangdong, China. Evidence of virus infection was also detected in other animals (including a raccoon dog, *Nyctereutes procyonoides*) and in humans working at the same market. All the animal isolates retain a 29-nucleotide sequence that is not found in most human isolates. The detection of SCoV-like viruses in small, live wild mammals in a retail market indicates a route of interspecies transmission, although the natural reservoir is not known.

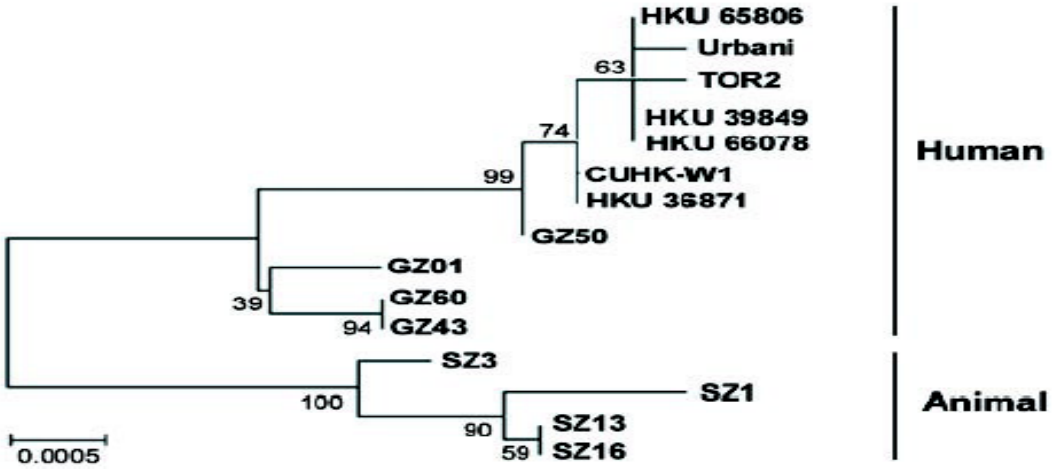


Fig. 2. Phylogenetic analysis of the nucleotide acid sequence of the spike gene of SCoV-like viruses. Nucleotide sequences of representative SCoV S genes (S gene coding region 21477 to 25244, 3768 bp) were analyzed. The phylogenetic tree was constructed by the neighbor-joining method with bootstrap analysis (1000 replicates) using MEGA 2 (*10*). Number at the nodes indicates bootstrap values in percentage. The scale bar shows genetic distance estimated using Kimura's two-parameter substitution model (*11*). In addition to viruses sequenced in the present study, the other sequences used in the analysis could be found in GenBank with accession number: from AY304490 to AY304495, AY278741, AY278554, AY278491, AY274119, and AY278489.

德国耐药肠道菌疫情

- 2011年5月，德国出现了由“肠出血性大肠杆菌”引发的“溶血性尿毒综合征”，迅速发生的疫情让4000多人染病，53人死亡。由于首先怀疑是一些人吃了黄瓜而致病，因而这一疫情被称为“**德国毒黄瓜事件**”。
- 在出现大量溶血性尿毒综合征病人后，德国和其他一些国家的研究人员马上投入到追查病原体的战斗中。 仅用了一个月时间，华大基因和德国研究人员合作的基因组测序结果“产志贺毒素大肠杆菌O104：H4的开源基因组分析”就在网站上公布出来，并随后发表在《新英格兰医学杂志》网络版上。
- 最终证据证明农场出产的豆芽是病菌源头。

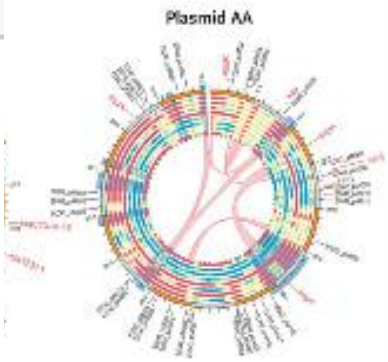
The NEW ENGLAND JOURNAL of MEDICINE

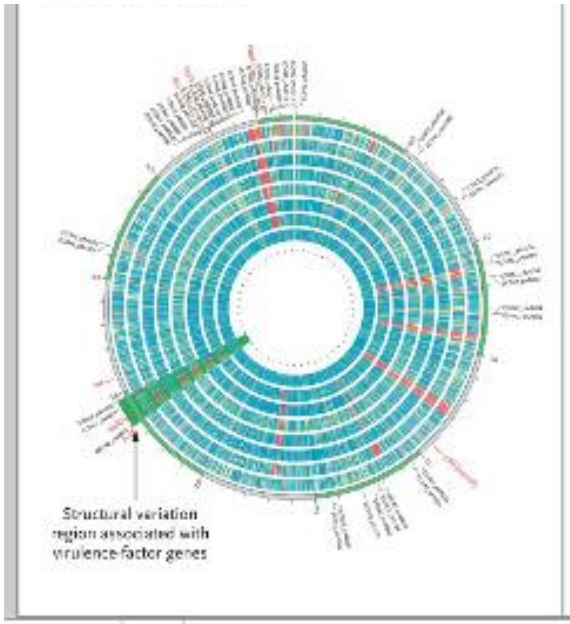
ORIGINAL ARTICLE

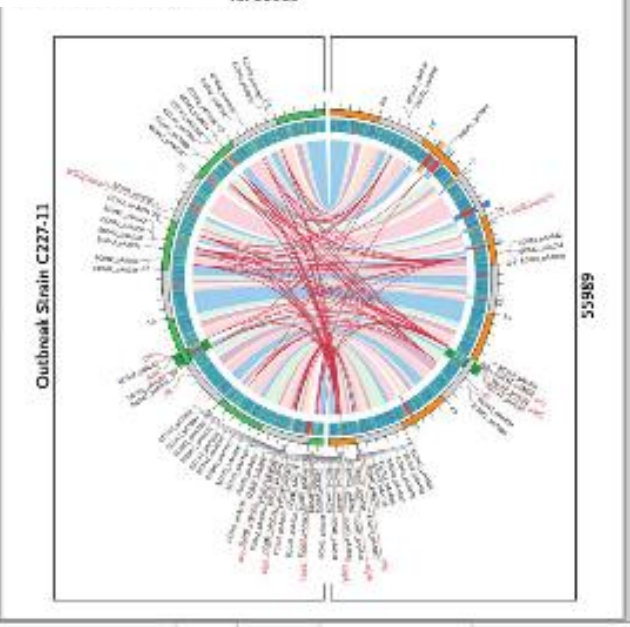
Origins of the *E. coli* Strain Causing an Outbreak of Hemolytic–Uremic Syndrome in Germany

David A. Rasko, Ph.D., Dale R. Webster, Ph.D., Jason W. Sahl, Ph.D., Ali Bashir, Ph.D., Nadia Boisen, Ph.D., Flemming Scheutz, Ph.D., Ellen E. Paxinos, Ph.D., Robert Sebra, Ph.D., Chen-Shan Chin, Ph.D., Dimitris Iliopoulos, Ph.D., Aaron Klammer, Ph.D., Paul Peluso, Ph.D., Lawrence Lee, Ph.D., Andrey O. Kislyuk, Ph.D., James Bullard, Ph.D., Andrew Kasarskis, Ph.D., Susanna Wang, B.S., John Eid, Ph.D., David Rank, Ph.D., Julia C. Redman, B.S., Susan R. Steyert, Ph.D., Jakob Frimodt-Møller, M.Sc.Eng., Carsten Struve, Ph.D., Andreas M. Petersen, Ph.D., Karen A. Krogfelt, Ph.D., James P. Nataro, M.D., Ph.D., M.B.A., Eric E. Schadt, Ph.D., and Matthew K. Waldor, M.D., Ph.D.

N Engl J Med 2011; 365:709-717 | August 25, 2011 | DOI: 10.1056/NEJMoa1106920







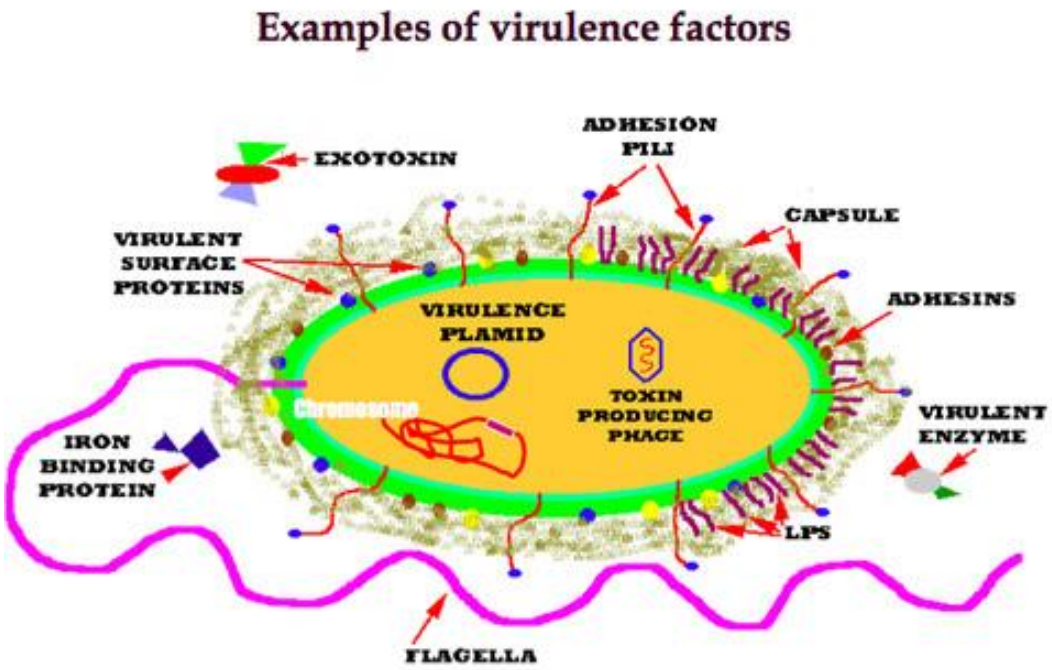
大肠埃希菌 (*Escherichia coli*)

- 大多E. coli是人肠道中的正常菌，对人体无害，如菌株E. coli K12

- 有些E. coli为人类致病菌，主要有：
  - ①肠产毒性大肠杆菌(ETEC)；
  - ②肠出血性大肠杆菌(EHEC)；
    - 致病物质主要为志贺样毒素(Shiga toxins)；
    - EHEC的主要血清型是O157：H7
  - ③肠侵袭性大肠杆菌(EIEC)；
  - ④肠致病性大肠杆菌(EPEC)。

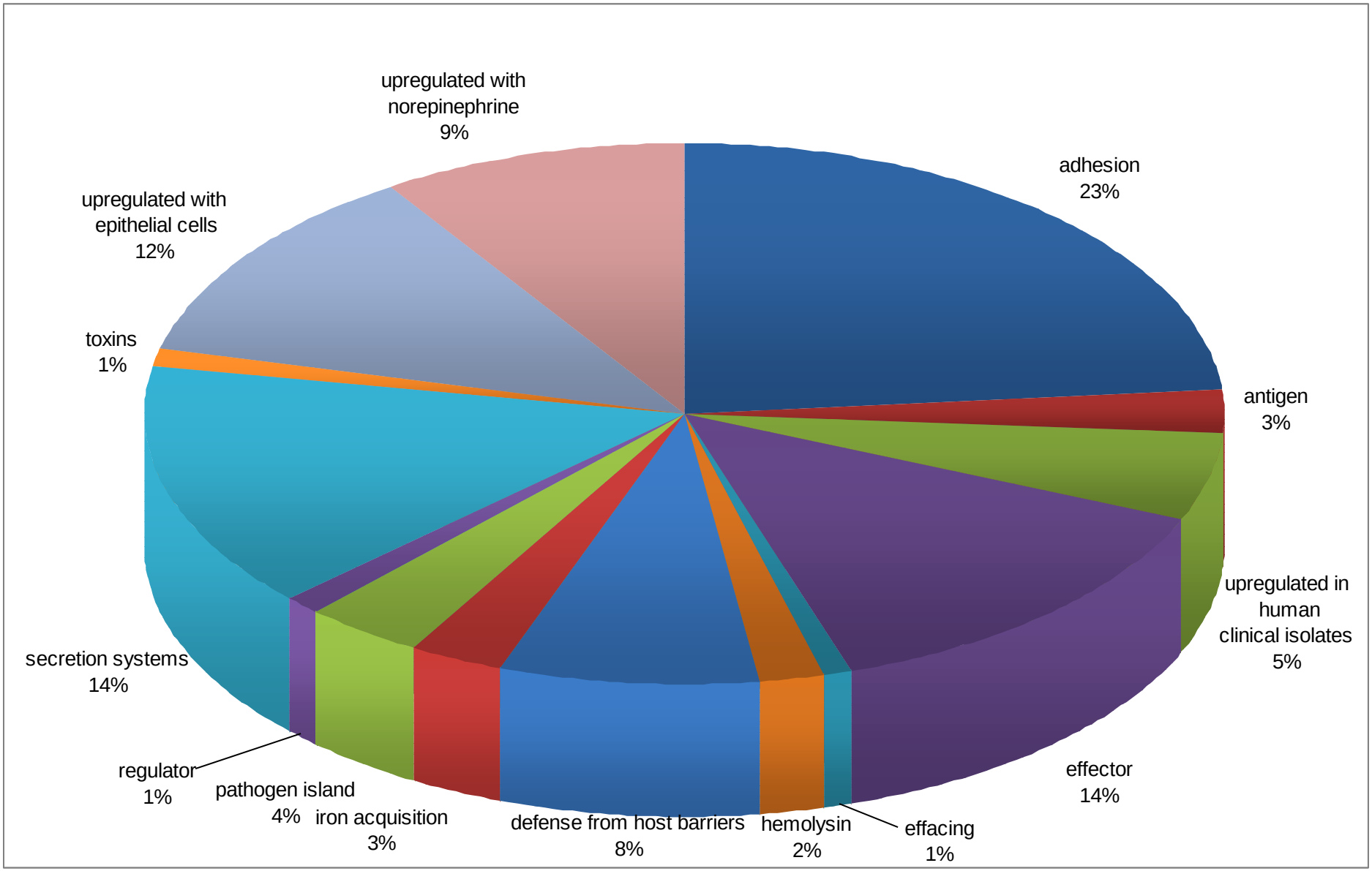
- 病原菌的传播主要通过污染的食品与水等，**主要污染肉、乳、水产品、蔬菜（低温保存、煮熟才吃）**。

毒力因子(virulence factors)



- 一般毒力因子
  - LPS(脂多糖)、荚膜：保护细菌表面
    - III型分泌系统：细菌向真核细胞内输送毒性基因产物的效应系统。
- 特殊毒力因子
  - 黏附素（CFA、AAF、BfP、紧密素、Ipa等）：黏附细胞表面
  - 外毒素（Stx、ST、LT等）：最主要毒力因子

E. coli O157:H7的virulence factors分类(n=394 genes)



大肠杆菌不同菌株全基因组比较

- 问题：对以下三个肠道杆菌的全基因组序列进行比对，分析其可能致病因子(Virulence Factor)
  1、 Escherichia coli str. K-12 substr. MG1655
  2、 Escherichia coli O157:H7 str. EC4115
  3、 Escherichia coli EDL933

ftp://ftp.ncbi.nlm.nih.gov/genbank/genomes/Bacteria

Mauve软件与基因组数据

1) 下载软件Mauve (version 2.3.1)：
http://gel.ahabs.wisc.edu/mauve/download.php

2) 下载压缩文件 (3 O157 alignments.zip)：
http://gel.ahabs.wisc.edu/~baumler/

3) 解压缩文件到当前目录。

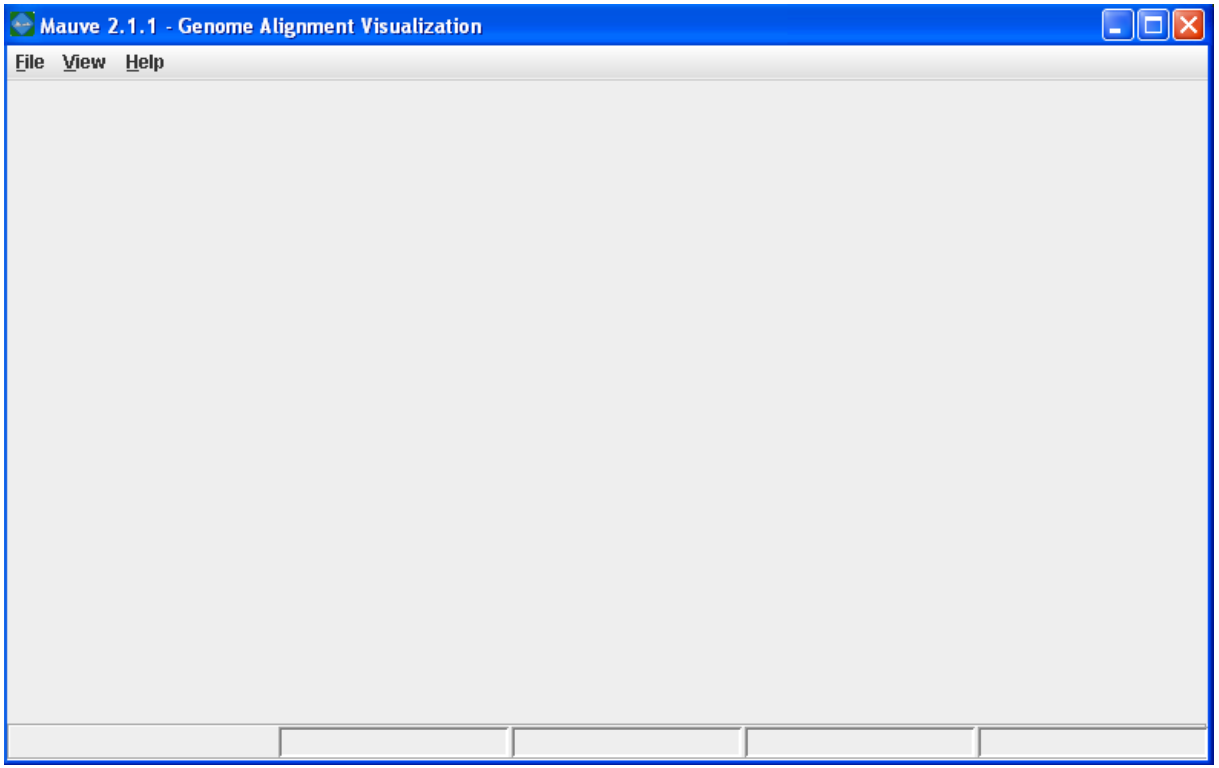
基因组GenBank注释文件

Mauve使用说明

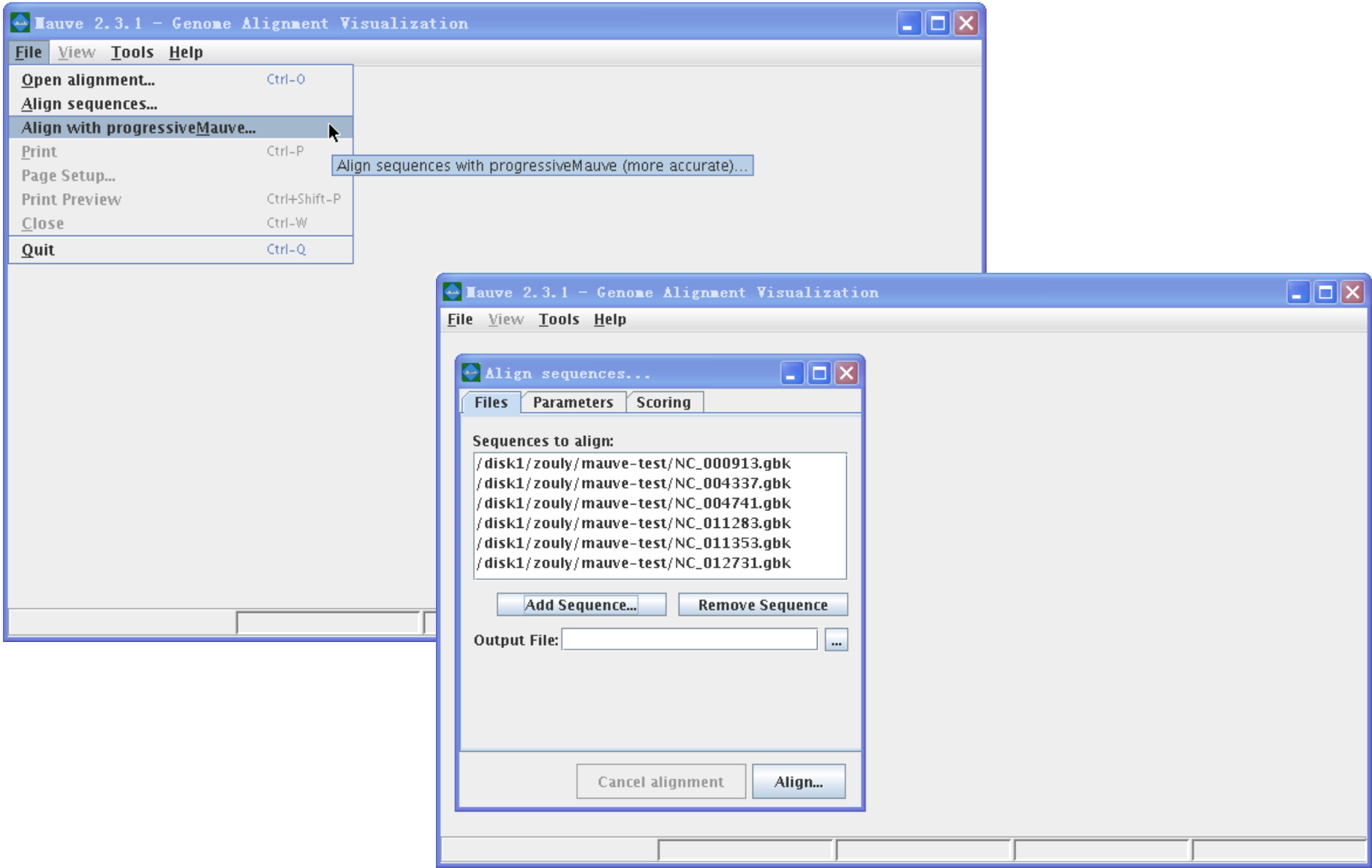
注意:由于mauve使用Java平台进行序列比对，电脑必须安装有Java。

第1步：启动mauve程序

Mauve主窗口

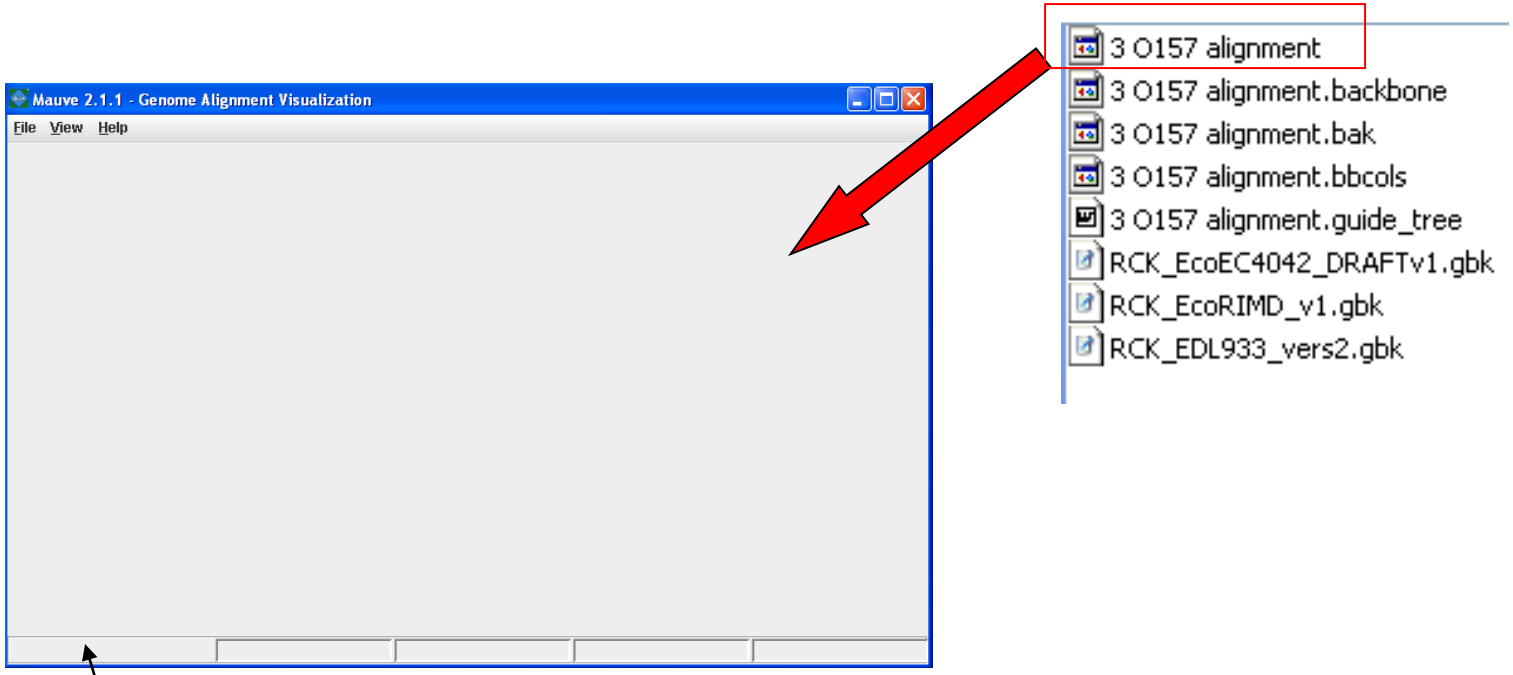


第2步：导入序列，执行比对



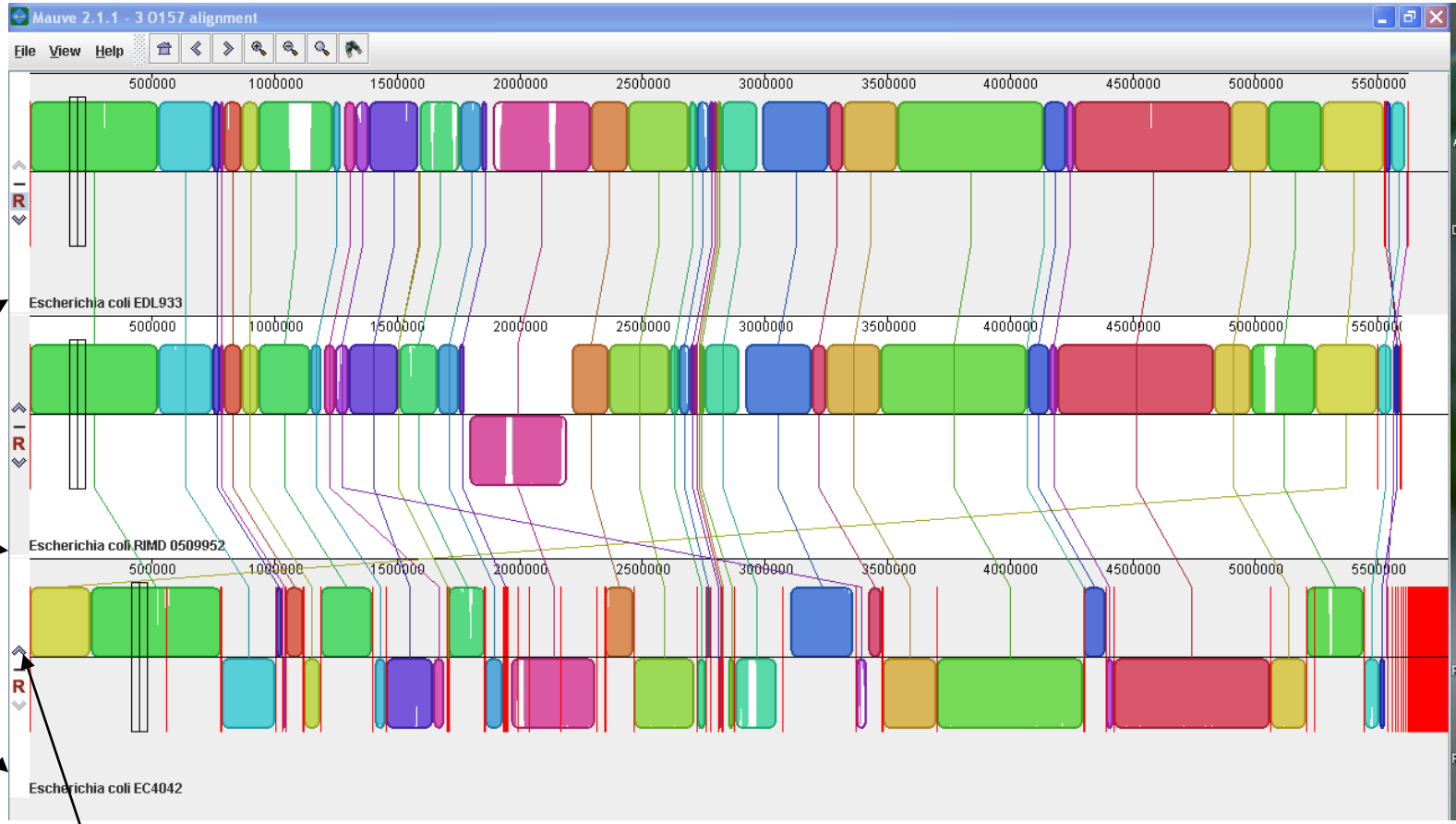
第2步：导入已比对好的数据

可以拖动第一个数据(3 O157 alignment)到窗口内打开



状态栏显示读取序列进度

第3步：显示和分析基因组比对结果



物种名称

上下箭头可以调整基因组的排列位置

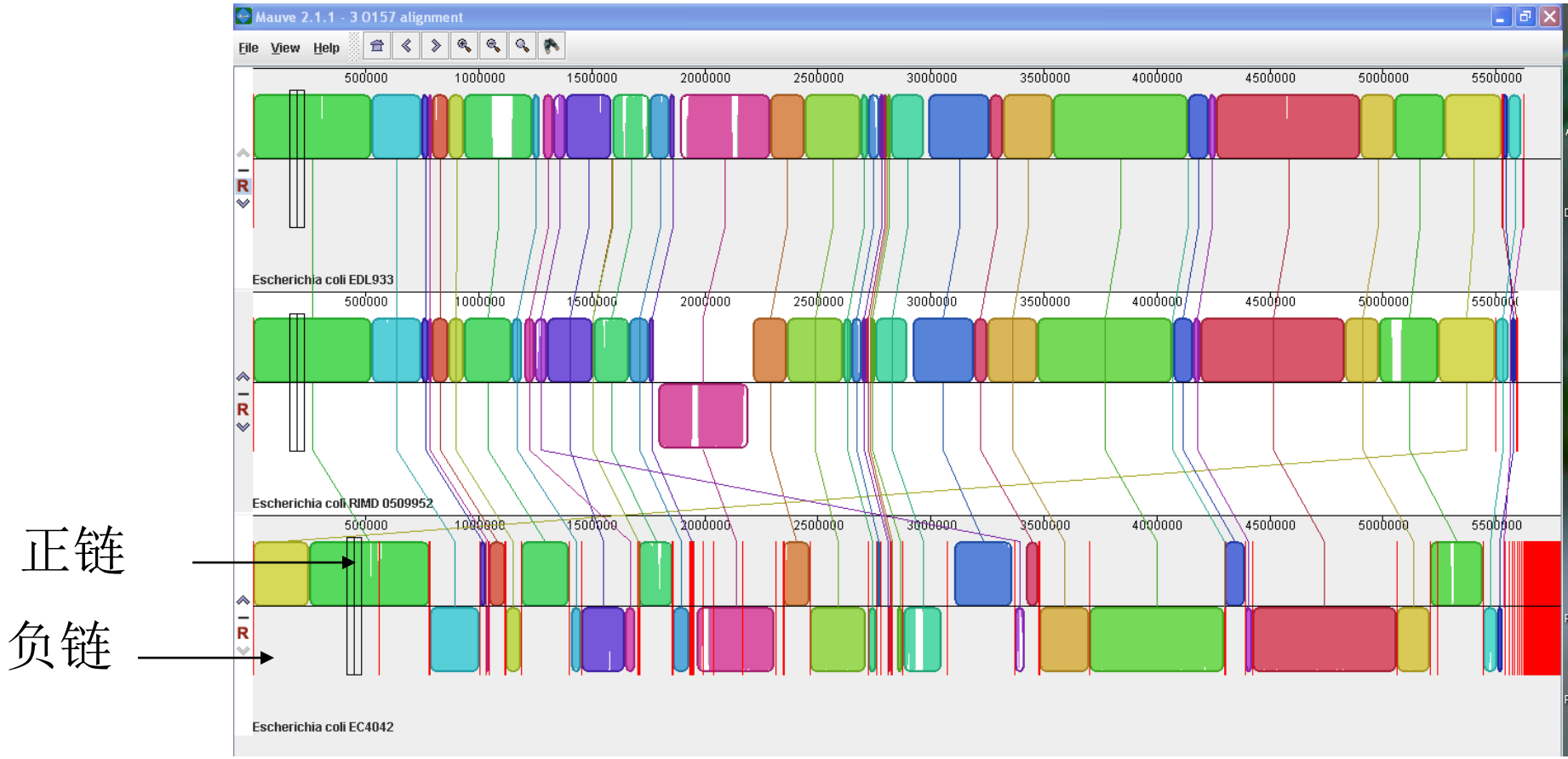
工具栏

回到起始视图

向左/右移动

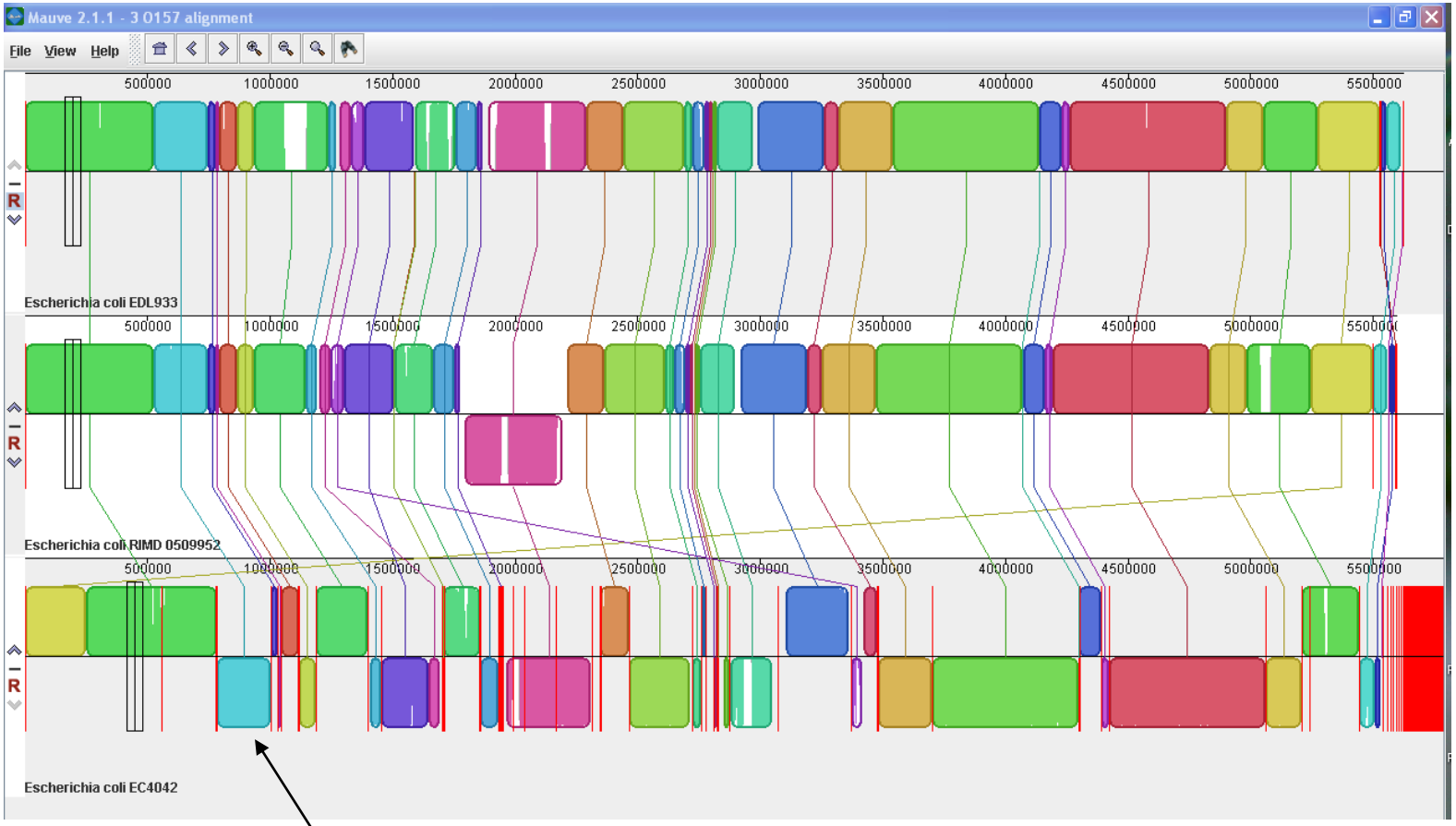
放大/缩小

查找特征
(features)

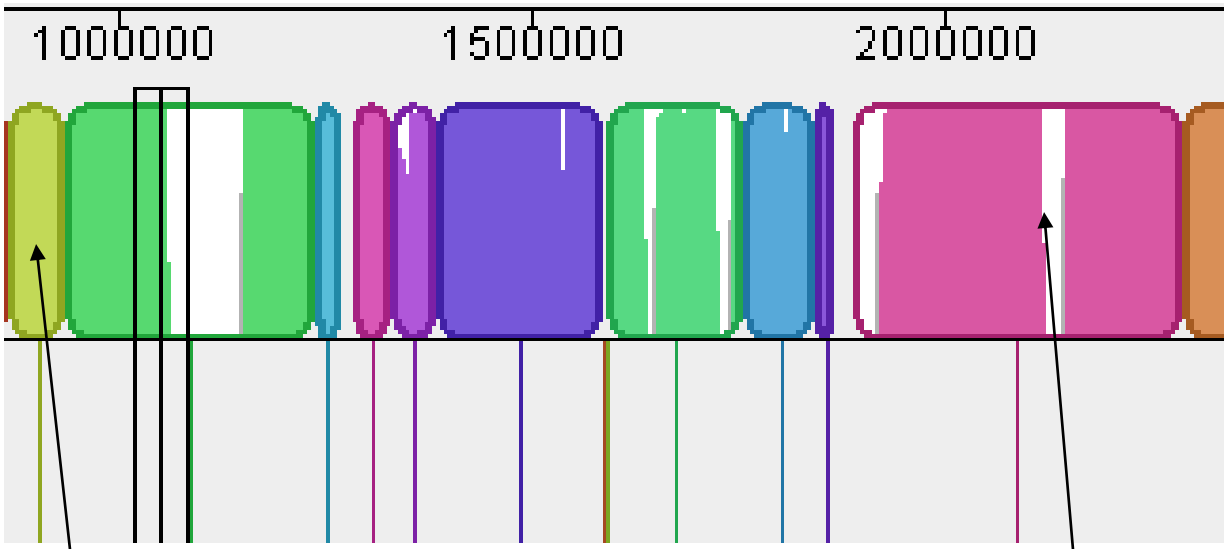


颜色块是代表Mauve鉴定的保守区域，称作 local colinear blocks (LCB's)。

不同物种的LCBs间以线连接, 注意有些在其它物种基因组中的位置不一样，有些是倒置的（位于负链）。

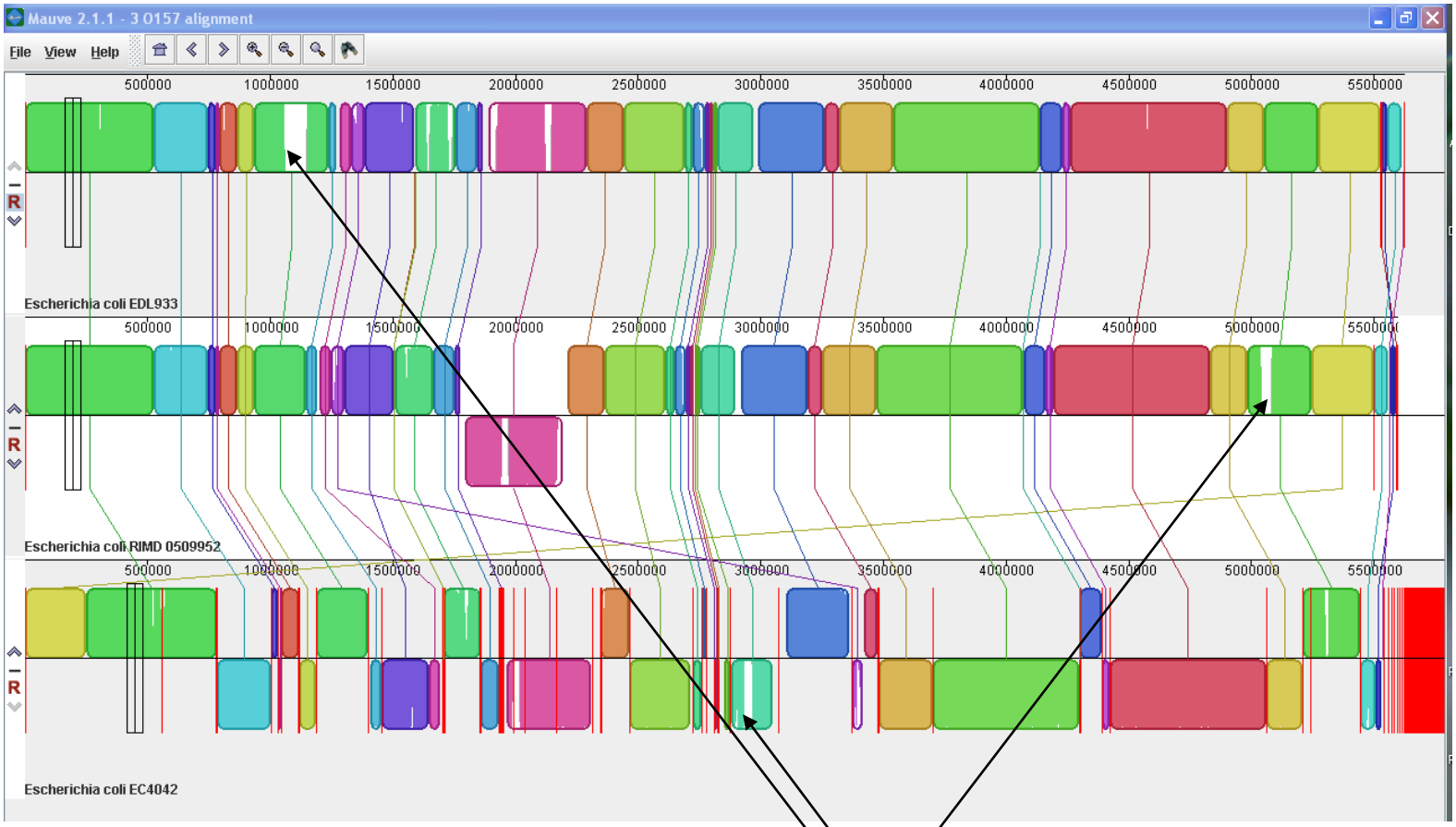


当你移动鼠标到一个基因组，所有三个基因组的相应区域就会显示一个黑框。

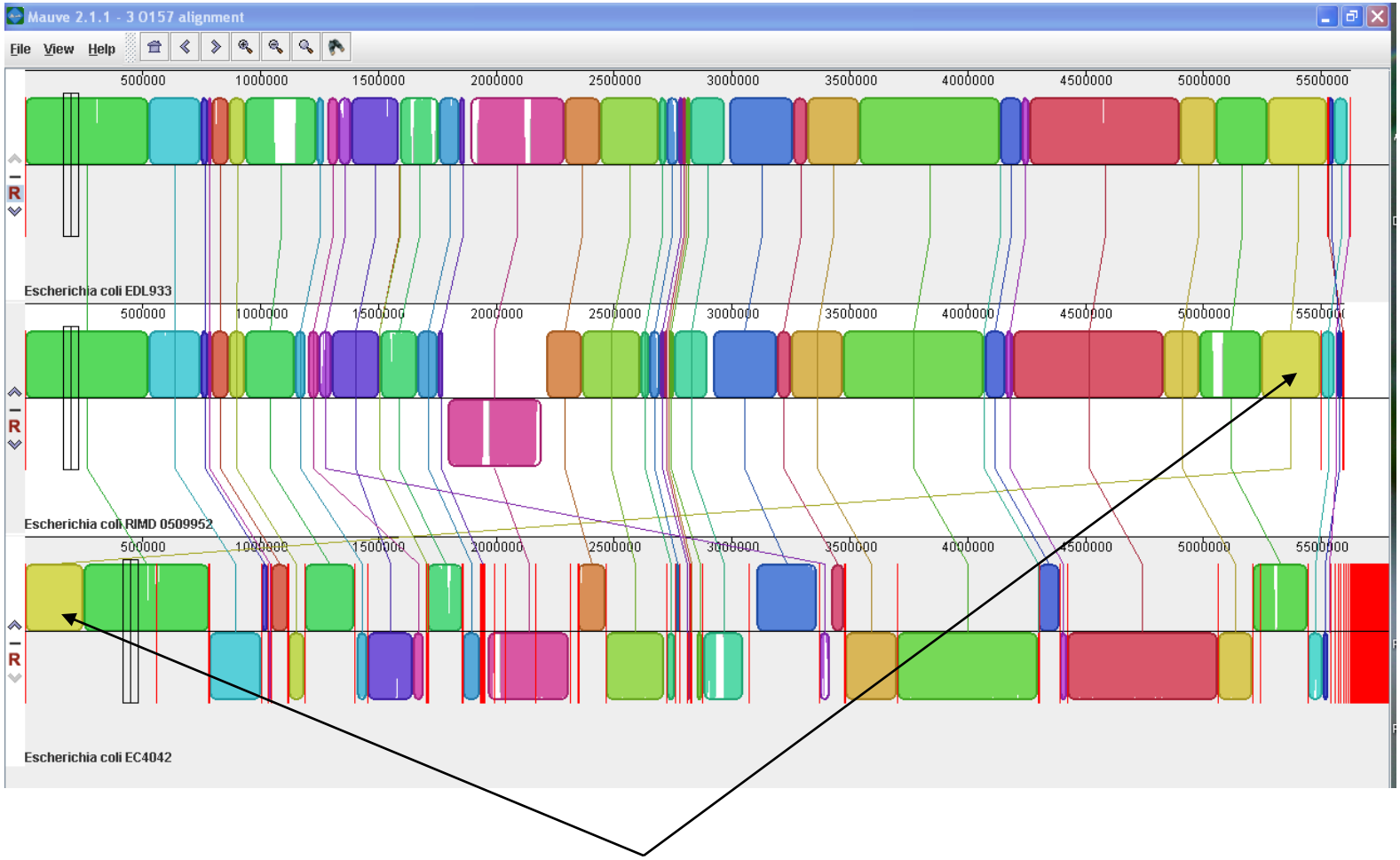


全色LCB代表高度保守的区域
(conserved/identical)

LCB中白色部分代表是某基因组中独特的区域
(unique/variable)

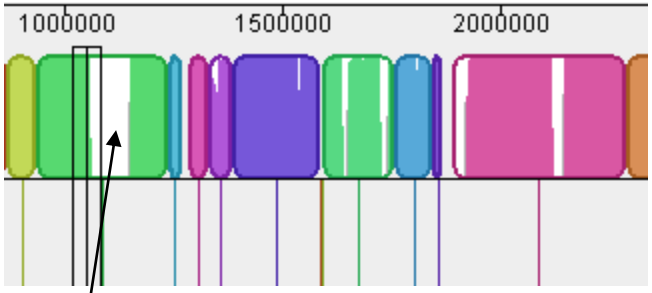


注意，当你把鼠标慢慢移动到白色区域，其它基因组中的黑框就会消失；当鼠标移过时，黑框又会出现。



如果一个LCB在不同的位置，你可以通过鼠标单击其中任一个LCB，就可以把所有三个LCB显示在一块。

1)首先点击home图标回到基因组比对的原始视图

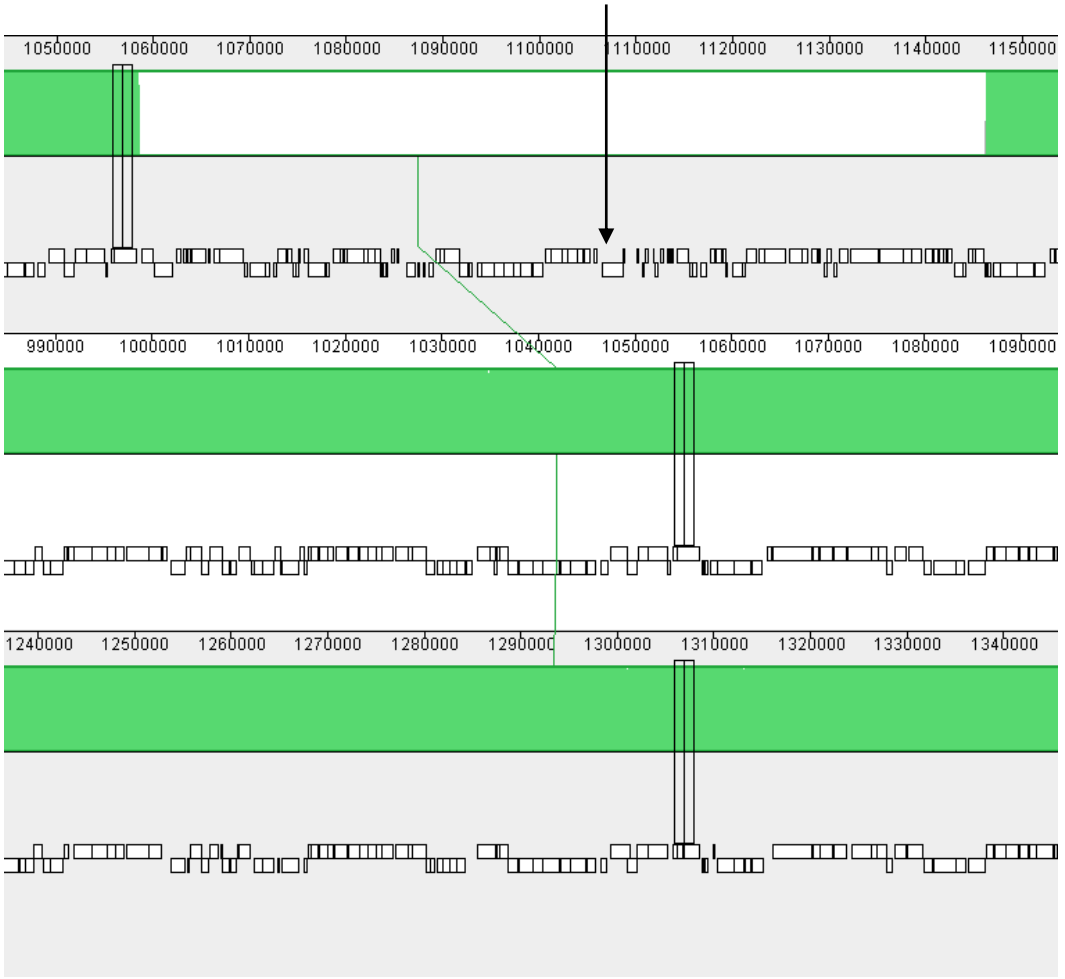


2)鼠标点击基因组中白色区域，并移动此区域到屏幕中央。后通过多次放大直到显示LCB下面的黑框。

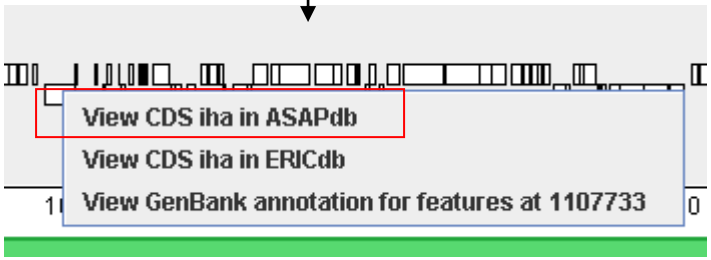
3)这些黑框代表预测的基因，即阅读框Open Reading Frames (ORFs).

鼠标移动到某个ORF可以显示一个跳出窗口，并显示基因的信息。 利利用此方法，你可以查看所有在 EDL933基因组中存在，但不存在于其它菌的基因或区域。

4)现在把鼠标移动到一个ORFs, 如*iha:*irgA homolog adhesion.



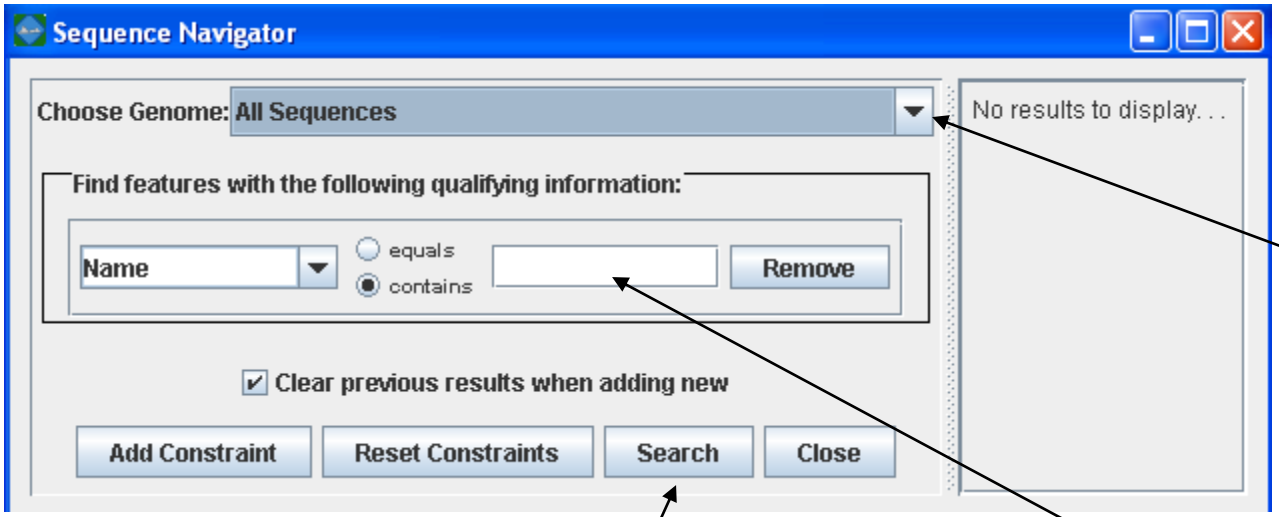
5) 点击 ORF, 从跳出的窗口中选择” View CDS iha in ASAPdb”.



6)这样就打开了 ASAP 数据库网页, 包括了这个基因的所有注释。 你可以查看这些注释是否会说明这个基因跟毒性virulence有关。

寻找基因组特征:

1) 点击”find features”工具栏图标



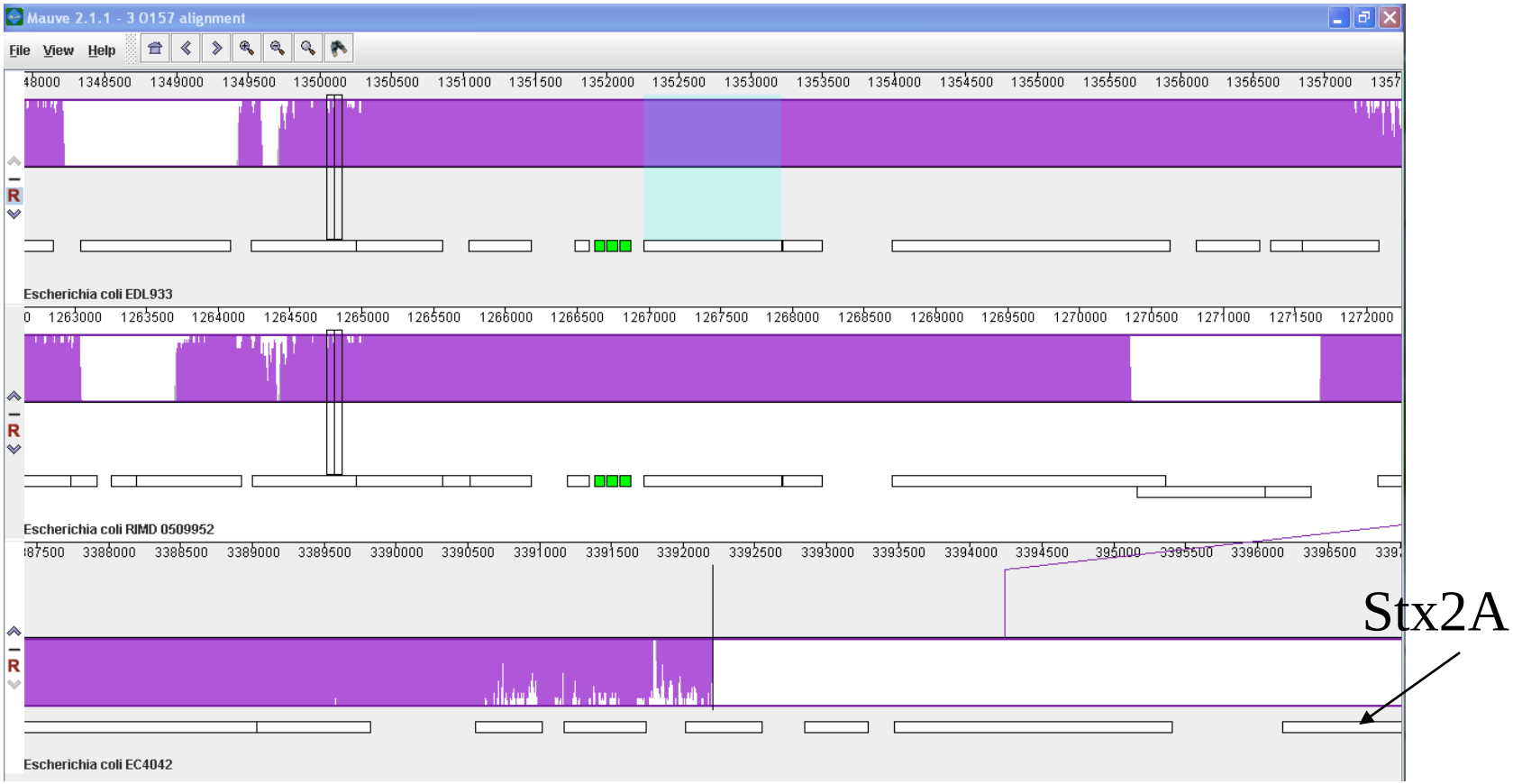
2) 选择一个基因组 (如EDL933)

3) 输入基因名称 (如*stx2A*)

4) 点击search

注意找到的基因*stx2A*以蓝色显示,可以看到它也存在于RIMD菌.

然而，EC4042菌的相同位置不存在此基因, 但它可以在 EC4042基因组的右边找到，有的基因在基因组中有许多份拷贝。



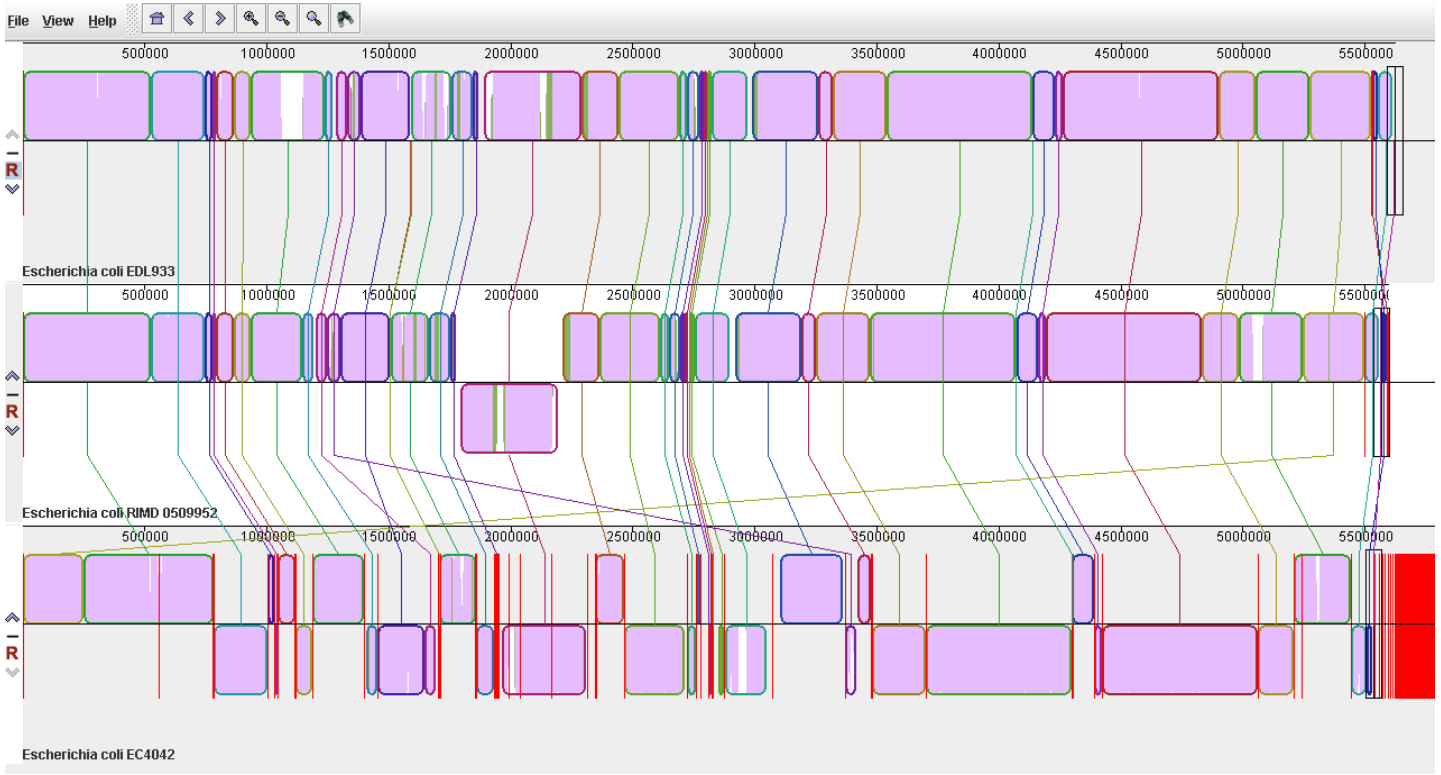
查找只在部分菌株中存在的基因组区域

视图模式从LCB切换为 backbone view.

1）按Home键

2）从视图菜单(View)选择color scheme 为 backbone color

现在基因比对显示为backbone color, 在三个基因组里都存在的区域显示为淡紫色( ) . 其它颜色代表在其它 只在部分基因组中保守的区域.



区域只在菌株EDL933 与 RIMD： olive green ( )

区域只在菌株EDL933 and EC4042： maroon ( )

区域只在菌株RIMD and EC4042： tan/brown ( )

不同的颜色代表只在其中两个基因组中保守的区域。

导入当前视图为图片:

快捷键： Ctrl+E (或通过菜单Tools-Export Image)

其它功能可以查阅软件说明书

作业

根据本章分析实例中3个*E. coli*菌株基因组的Mauve比对结果，试回答以下问题：

1、利用三个基因组比较结果，鉴定某个*E. coli*致病菌株独特的基因组区域(island)，并简要描述处于这个基因组区域的基因产物（要求显示分析区域的图片）；

2、同上，鉴定两个病原*E.coli*菌株中共有，在另一菌株中不存在的区域，并简要描述处于这个基因组区域的基因产物（要求显示分析区域的图片）；

3、综上分析，你觉得哪些区域可能与病原菌株的致病性有关，或可能是致病因子(Virulence Factor)?

4、选择一个可能是致病因子的序列，通过BLAST搜索此致病因子在其它细菌或古菌中的同源基因，并以表列出前5个最佳比对序列，含基因、物种名称及相似度(%)。