Chapter 2

Genome Organization and Evolution

基因體系統與演化

Genomes, transcriptomes, and Proteomes

基因體，轉錄組，與蛋白質體

- Bacterial genome comes from a single DNA molecule of about 5 million characters.
- The DNA of higher organisms is organized into chromosomes: normal human cells contain 23 chromosome pairs
- The total amount of the genetic information per cell – the sequence of nucleotides of DNA – is very nearly constant for all members of a species, but varies widely between species
- The number of protein sequences cannot be easily assessed from genome size

  ➔Some genes exist in the body as multiple copies

Organism	Genome size (base pairs)
Epstein-Barr virus	0.172×10^6
Bacterium (*E. coli*)	4.6×10^6
Yeast (*S. cerevisiae*)	12.5×10^6
Nematode worm (*C. elegans*)	100.3×10^6
Thale cress (*A. thaliana*)	115.4×10^6
Fruit fly (*D. melanogaster*)	128.3×10^6
Human (*H. sapiens*)	3223×10^6

基因(Genes)

- DNA分子上一個或多個區段的核甘酸序列
- 細菌中，3N核甘酸可轉譯成N個氨基酸
- 真核細胞內，轉譯較為複雜
  - Exon轉錄形成訊息RNA(mRNA)，最後與核糖體一起作用進入轉譯蛋白質的程序
  - An exon is a stretch of DNA retained in the mature mRNA that the ribosome translates into protein.
  - An intron is an **int**ervening **r**egio**n** between two exons.
  - Many introns are very long – in some cases substantially longer than the exons
- 各種生物基因體大小見 p. 71表

Genome Sizes

Organism	Number of base pairs	Number of genes	Comment
ϕX-174	5 386	10	virus infecting *E. coli*
Human mitochondrion	16 569	37	subcellular organelle
Epstein-Barr virus (EBV)	172 282	80	cause of mononucleosis
Mycoplasma pneumoniae	816 394	680	cause of cyclic pneumonia epidemics
Rickettsia prowazekii	1 111 523	834	bacterium, cause of epidemic typhus
Treponema pallidum	1 138 011	1 039	bacterium, cause of syphilis
Borrelia burgdorferi	1 471 725	1 738	bacterium, cause of Lyme disease
Aquifex aeolicus	1 551 335	1 749	bacterium from hot spring
Thermoplasma acidophilum	1 564 905	1 509	archaeal prokaryote, lacks cell wall
Campylobacter jejuni	1 641 481	1 708	frequent cause of food poisoning
Helicobacter pylori	1 667 867	1 589	chief cause of stomach ulcers
Methanococcus jannaschii	1 664 970	1 783	archaeal prokaryote, thermophile
Haemophilus influenzae	1 830 138	1 738	bacterium, cause of middle ear infections
Thermotoga maritima	1 860 725	1 879	marine bacterium
Archaeoglobus fulgidus	2 178 400	2 437	another archaeon
Deinococcus radiodurans	3 284 156	3 187	radiation-resistant bacterium
Synechocystis	3 573 470	4 003	cyanobacterium, 'blue-green alga'

Genome Sizes (*continued*)

Vibrio cholerae	4 033 460	3 890	cause of cholera
Mycobacterium tuberculosis	4 411 532	3 959	cause of tuberculosis
Bacillus subtilis	4 214 814	4 779	popular in molecular biology
Escherichia coli	4 639 221	4 377	molecular biologists' all-time favourite
Pseudomonas aeruginosa	6 264 403	5 570	one of largest prokaryote genomes sequenced
Saccharomyces cerevisiae	12 495 682	5 770	yeast, first eukaryotic genome sequenced.
Caenorhabditis elegans	103 006 709	20 598	the worm
Arabidopsis thaliana	115 409 949	25 498	flowering plant (angiosperm)
Drosophila melanogaster	128 343 463	13 525	the fruit fly
Takiugu rubripes	329×10^6	34 080?	puffer fish (fugu fish)
Human	3223×10^6	23 000?	
Wheat	16×10^9	30 000	
Salamander	10^{11}	?	
Psilotum nudum	2.5×10^{11}	?	whisk fern—a simple plant

- The **genes** that code for proteins, and for structural RNA molecules, present only the *static picture* of the genome.

蛋白質體學和轉錄組(Proteomes and Transcriptomics)

- An organism's genome gives a ***complete specification*** of the potential life of that individual
- What reveals the activity of a cell at any instant, at the molecular level, is the ***set of RNAs being transcribed*** and ***the set of proteins synthesized***

- Inventories of RNA molecules and proteins in cells deal in a more direct and integral way with cellular status and activity.
- These data are subjects of the transcriptome and Proteome projects
- A large-scale programme dealing in an integral way with patterns of expression of proteins in biological systems, in ways that complement and extend genome projects.

蛋白質體學(Proteomes)

- In principle, a database of amino acid sequences of proteins is inherent in the database of nucleotide sequences of DNA, by virtue of the genetic code.
- New protein sequence data are now being determined by translation of DNA sequences, rather than by direct sequencing of proteins.
  - Should any distinction be made?
- First, we must assume that it is possible correctly to identify within the DNA data stream the regions that encode proteins
  - A protein inferred from a genome sequence is a hypothetical object until an experiment verifies its existence.

- The pattern recognition programs are subjected to three types of errors:
  - A genuine protein sequence may be missed entirely
  - An incomplete protein may be reported
  - A gene may be incorrectly spliced
- Second, in many cases the expression of a gene produces a molecule that must be ***modified*** with a cell, to make a mature protein that differs significantly from the one suggested by translation of the gene sequence.
  - In many cases the missing details of ***post-translational modification*** are quite important
  - In some cases, mRNA is edited before translation, creating changes in amino acid sequences that are not inferable from the genes.

Eavesdropping (偷聽) on the transmission of genetic information

- How hereditary (遺傳的) information is ***stored, passed on, and implemented*** is perhaps the fundamental problem of biology.
- 3 types of maps have been essential:
  - Linkage maps of genes (Gene map)
  - Banding patterns of chromosomes (Chromosome map)
  - DNA sequences (Sequence map)

Gene maps, chromosome maps, and sequence maps

1. A **gene map** is classically determined by observed patterns of heredity. Linkage groups and recombination frequencies can detect whether genes are on the same or different chromosomes, and, for genes on the same chromosome, how far apart they are. The principle is that the farther apart two linked genes are, the more likely they are to recombine, by crossing over during meiosis. Indeed, two genes on the same chromosome but very far apart will appear to be unlinked. The unit of length in a gene map is the Morgan, defined by the relation that 1 cM corresponds to a 1% recombination frequency. (We now know that 1 cM ~ 1×10^6 bp in humans, but it varies with the location in the genome and with the distance between genes.)
2. **Chromosome banding pattern maps** Chromosomes are physical objects. Banding patterns are visible features on them. The nomenclature is as follows: In many organisms, chromosomes are numbered in order of size, 1 being the largest. The two arms of human chromosomes, separated by the centromere, are called the p (petite = short) arm and q (= queue) arm. Regions within the chromosome are numbered p1, p2, . . . and q1, q2 . . . outward from the centromere. Subsequent digits indicate subdivisions of bands. For example, certain bands on the q arm of human chromosome 15 are labelled 15q11.1, 15q11.2, 15q12. Originally bands 15q11 and 15q12 were defined; subsequently 15q11 was divided into 15q11.1 and 15q11.2. Deletions including this region are associated with Prader-Willi and Angleman syndromes. These syndromes have the interesting feature that the alternative clinical consequences depend on whether the affected chromosome is paternal or maternal. This observation of **genomic imprinting** shows that the genetic information in a fertilized egg is not simply the bare DNA sequences contributed by the parents. Chromosomes of paternal and maternal origin have different states of methylation, signals for differential expression of their genes. The process of modifying the DNA which takes place during differentiation in development is already present in the zygote.
3. **The DNA sequence itself** Physically a sequence of nucleotides in the molecule, computationally a string of characters A, T, G and C. Genes are regions of the sequence, in many cases interrupted by non-coding regions.

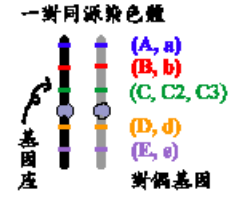
Identification of genes associated with inherited diseases

- Given a disease attributable to a defective protein:
  - If we know the protein involved, we can pursue rational approaches to therapy
  - If we know the gene involved, we can devise tests to identify sufferers or carriers
  - In many cases, knowledge of the chromosomal location of the gene is unnecessary for either therapy or detection;
    - it's required only for identifying the gene, providing a bridge between the patterns of inheritance and the DNA sequence

- 一條染色體上, 帶有很多不同大小, 種類的基因, 稱為基因聯連鎖群 (gene linkage group) , 所以, 也可以說, 人類有 46 條基因連鎖群, 而每一個基因在染色體上, 都有一個固定位置, 叫基因座 (loci).

- A chromosome contains many genes of different sizes and types, called a gene linkage group. Therefore, it can also be said that humans have 46 gene linkage groups. Each gene has a fixed location on the chromosome called a locus.

- 可是... 什麼是 "一對" 染色體呢? 誰和誰配對呢? 像人類細胞這種每一種染色體是成對的生物, 叫二倍體生物個體, 而成對的染色體叫同源染色體, 它們具有相同的基因座 (loci) ,一個基因可能有很多分身版本,稱對偶基因 (allele), 例右圖中,對偶基因 C,C2,C3, 但它們都有機會佔據相同的位置 (基因座), 但不論一個基因有多少分身版本, 人類細胞只收集二個, 擺在 "一對" 同源染色體上.



- What is a "pair" of chromosomes? Who pairs with whom? Each chromosome in human cells is a paired organism, called a diploid organism. Pairs of chromosomes are called homologous chromosomes, and they have the same genetic locus (loci). A gene may have many copies, called alleles. In the figure, the dual genes C, C2, and C3 all have the chance to occupy the same position (gene locus). But no matter how many copies a gene has, human cells only collect two and place them on a "pair" of homologous chromosomes.

- 所以一對同源染色體，它們所帶有的所有基因大致是相同的，只是可能雙方版本不同而已。

- Therefore, in a pair of homologous chromosomes, all the genes they carry are roughly the same, but they may have different versions.

Picking out genes in genomes
從基因體中挑選基因

- ORF (***open reading frame***): a region of DNA sequence that begins with an initiation codon (ATG) and ends with a stop codon.
- An ORF is a potential protein-coding region
- Approaches to identifying protein-coding regions choose from or combine 2 possible ways:
  1. Detection of regions similar to known coding regions from other organisms
  2. Ab initio (從頭算) methods, that seek to identify genes from the properties of the DNA sequences themselves

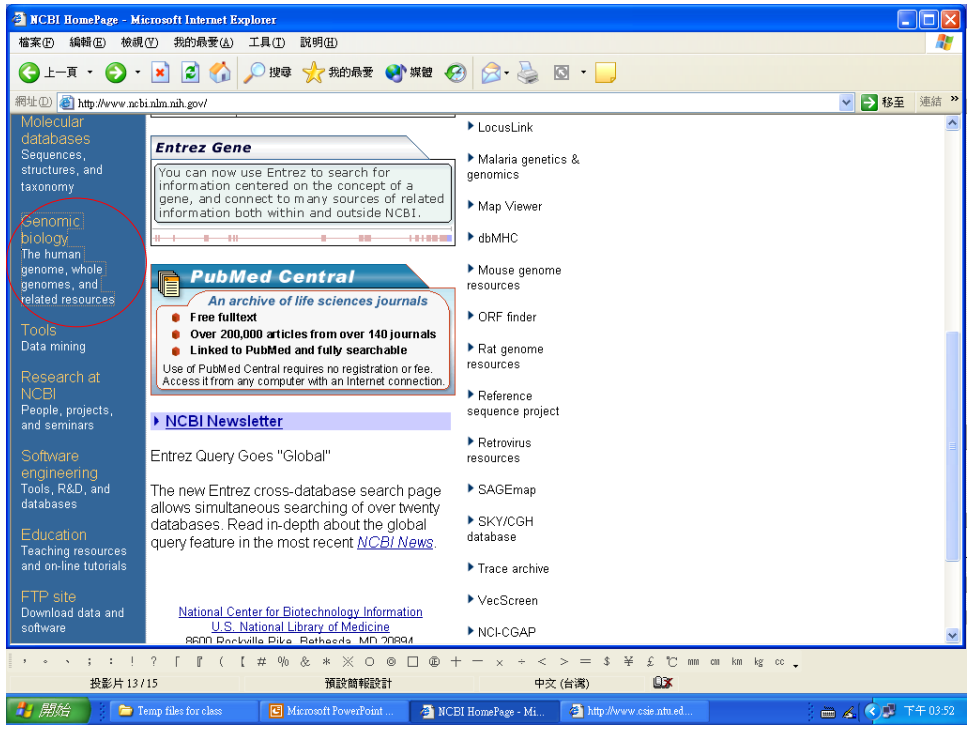
- 20230928

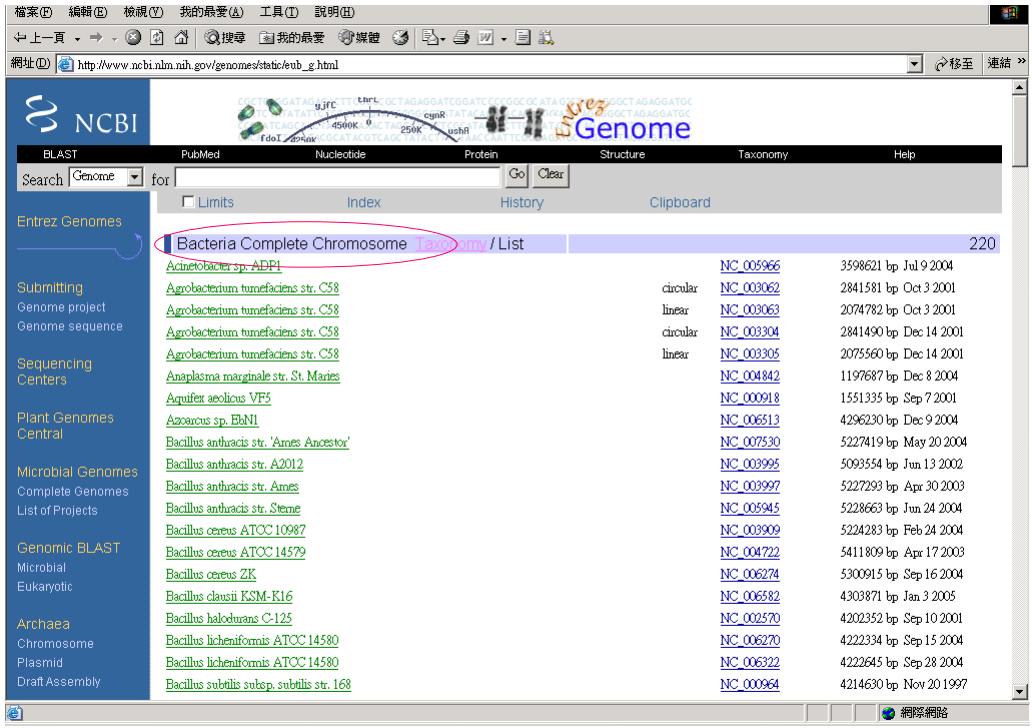
原核生物基因體 Genomes of prokaryotes

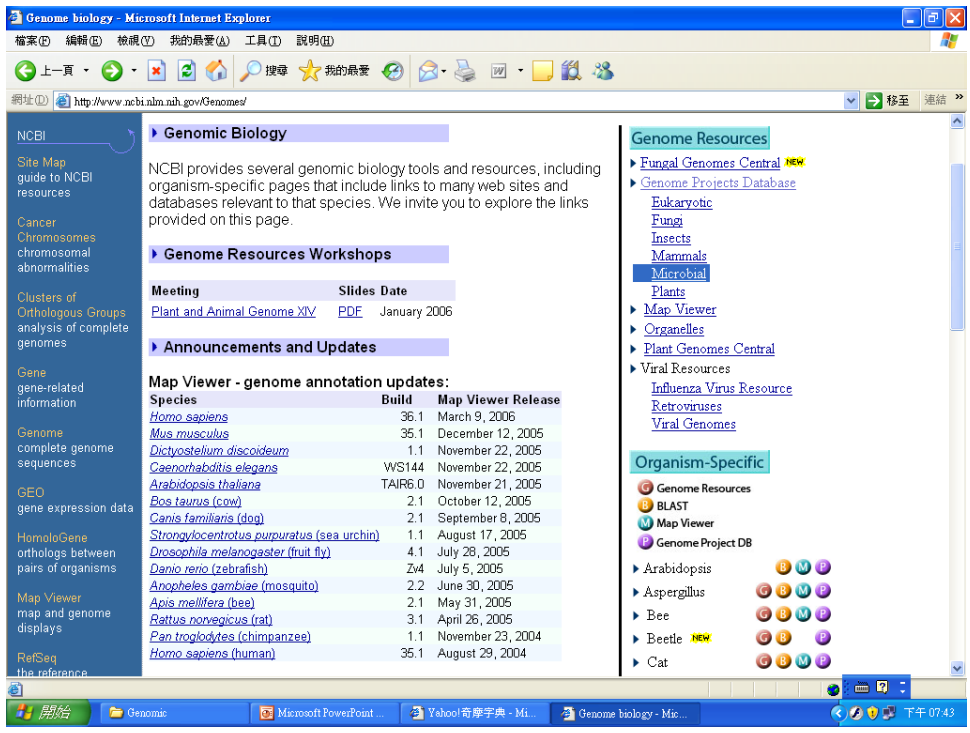
- 細菌基因體之蛋白質編碼區段內並不含intron
- (The protein-coding sequences of the bacterial genome do not contain intron)
- 在典型原核生物的基因組內，相對於真核生物，僅具有非常少量的未編碼DNA片段。
- (In the genome of typical prokaryotes, there are only a very small number of uncoded DNA segments compared to eukaryotes.)

大腸桿菌基因體(Genome of the bacterium E. Coli)：

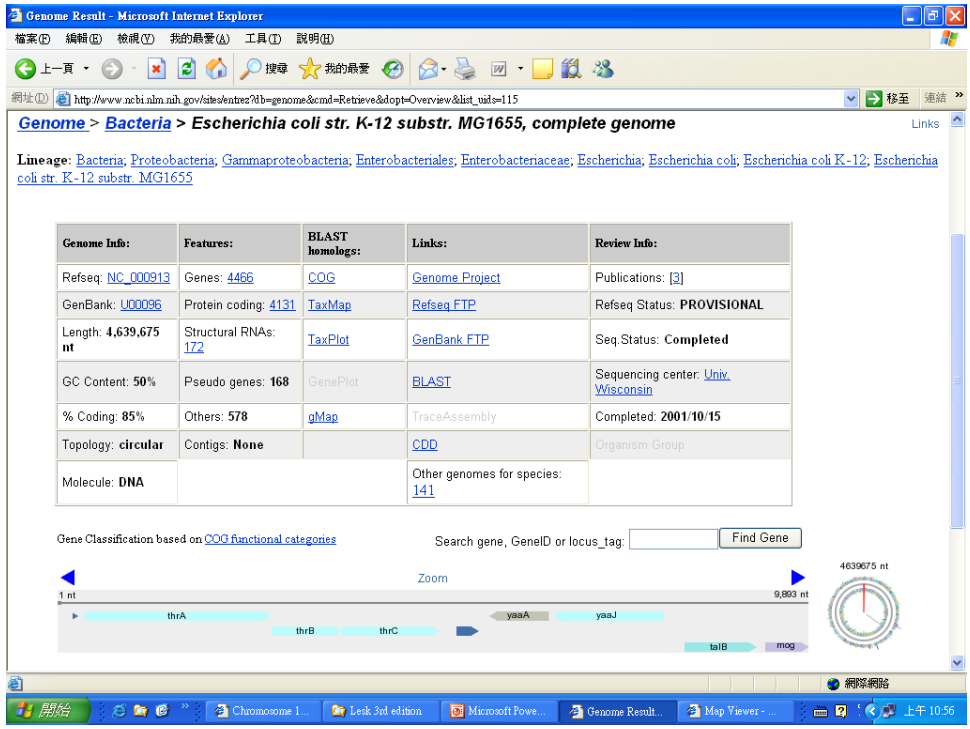
- 4,639,221 bp單一環狀DNA分子 (a single circular DNA molecule, with no plastids)
- 89%序列密碼用在蛋白質或結構DNA。(89% of the sequence codes for proteins or structural RNAs)
- 4824 protein-coding genes
- 122 structural RNA genes
- 相當密集的基因，ORF平均大小為317 Amino acids (Average size of an ORF is 317 amino acids)
- 最大的蛋白質分類為酵素，接近30%。(the largest class of proteins is the enzymes, 30% of total genes)

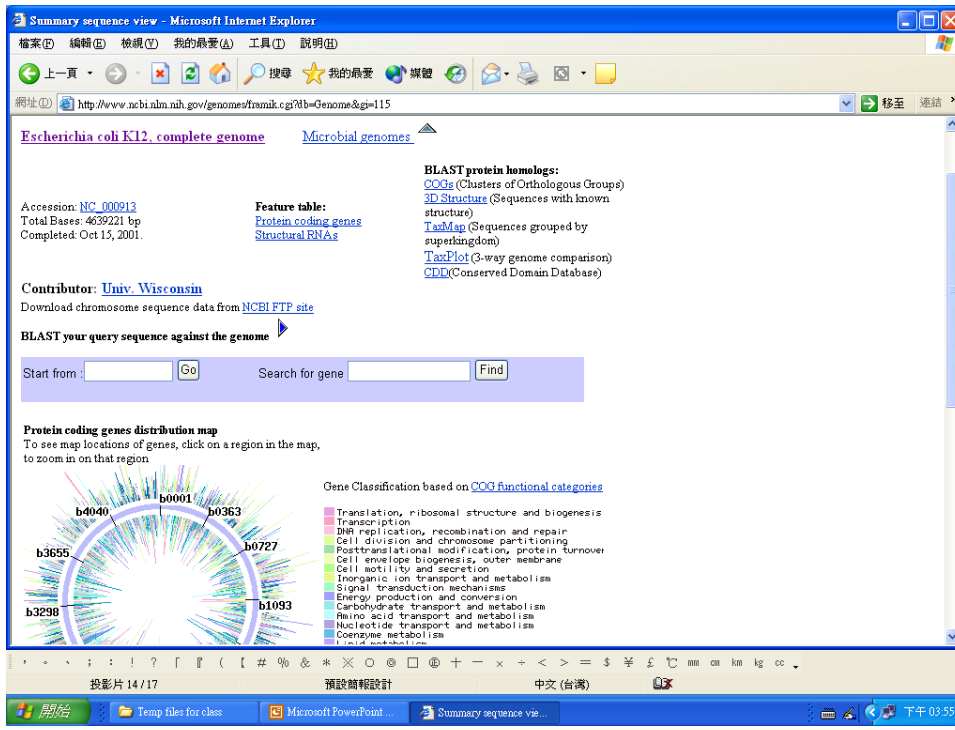



NCBI
Entrez Genome
BLAST PubMed Nucleotide Protein Structure Taxonomy Help
Search Genome for Go Clear
Limits Index History Clipboard
Entrez Genomes
Submitting
Genome project
Genome sequence
Sequencing Centers
Plant Genomes Central
Microbial Genomes
Complete Genomes
List of Projects
Genomic BLAST
Microbial
Eukaryotic
Archaea
Chromosome
Plasmid
Draft Assembly
Bacteria Complete Chromosome Taxonomy / List 220
Acinetobacter sp. ADP1 NC_005966 3598621 bp Jul 9 2004
Agrobacterium tumefaciens str. C58 circular NC_003062 2841581 bp Oct 3 2001
Agrobacterium tumefaciens str. C58 linear NC_003063 2074782 bp Oct 3 2001
Agrobacterium tumefaciens str. C58 circular NC_003304 2841490 bp Dec 14 2001
Agrobacterium tumefaciens str. C58 linear NC_003305 2075560 bp Dec 14 2001
Anaplasma marginale str. St. Maries NC_004842 1197687 bp Dec 8 2004
Aquifex aeolicus VF5 NC_000918 1551335 bp Sep 7 2001
Azoarcus sp. EbN1 NC_006513 4296230 bp Dec 9 2004
Bacillus anthracis str. 'Ames Ancestor' NC_007530 5227419 bp May 20 2004
Bacillus anthracis str. A2012 NC_003995 5093554 bp Jun 13 2002
Bacillus anthracis str. Ames NC_003997 5227293 bp Apr 30 2003
Bacillus anthracis str. Sterne NC_005945 5228663 bp Jun 24 2004
Bacillus cereus ATCC 10987 NC_003909 5224283 bp Feb 24 2004
Bacillus cereus ATCC 14579 NC_004722 5411809 bp Apr 17 2003
Bacillus cereus ZK NC_006274 5300915 bp Sep 16 2004
Bacillus clausii KSM-K16 NC_006582 4303871 bp Jan 3 2005
Bacillus halodurans C-125 NC_002570 4202352 bp Sep 10 2001
Bacillus licheniformis ATCC 14580 NC_006270 4222334 bp Sep 15 2004
Bacillus licheniformis ATCC 14580 NC_006322 4222645 bp Sep 28 2004
Bacillus subtilis subsp. subtilis str. 168 NC_000964 4214630 bp Nov 20 1997


Genome biology - Microsoft Internet Explorer
http://www.ncbi.nlm.nih.gov/Genomes/
NCBI
Site Map guide to NCBI resources
Cancer Chromosomes chromosomal abnormalities
Clusters of Orthologous Groups analysis of complete genomes
Gene gene-related information
Genome complete genome sequences
GEO gene expression data
HomoloGene orthologs between pairs of organisms
Map Viewer map and genome displays
RefSeq
Genomic Biology
NCBI provides several genomic biology tools and resources, including organism-specific pages that include links to many web sites and databases relevant to that species. We invite you to explore the links provided on this page.
Genome Resources Workshops
Meeting Slides Date
Plant and Animal Genome XIV PDF January 2006
Announcements and Updates
Map Viewer - genome annotation updates:
Species Build Map Viewer Release
Homo sapiens 36.1 March 9, 2006
Mus musculus 35.1 December 12, 2005
Dictyostelium discoideum 1.1 November 22, 2005
Caenorhabditis elegans WS144 November 22, 2005
Arabidopsis thaliana TAIR6.0 November 21, 2005
Bos taurus (cow) 2.1 October 12, 2005
Canis familiaris (dog) 2.1 September 8, 2005
Strongylocentrotus purpuratus (sea urchin) 1.1 August 17, 2005
Drosophila melanogaster (fruit fly) 4.1 July 28, 2005
Danio rerio (zebrafish) Zv4 July 5, 2005
Anopheles gambiae (mosquito) 2.2 June 30, 2005
Apis mellifera (bee) 2.1 May 31, 2005
Rattus norvegicus (rat) 3.1 April 26, 2005
Pan troglodytes (chimpanzee) 1.1 November 23, 2004
Homo sapiens (human) 35.1 August 29, 2004
Genome Resources
Fungal Genomes Central NEW
Genome Projects Database
Eukaryotic
Fungi
Insects
Mammals
Microbial
Plants
Map Viewer
Organelles
Plant Genomes Central
Viral Resources
Influenza Virus Resource
Retroviruses
Viral Genomes
Organism-Specific
Genome Resources
BLAST
Map Viewer
Genome Project DB
Arabidopsis
Aspergillus
Bee
Beetle NEW
Cat


Complete Microbial Genomes - Microsoft Internet Explorer
http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi
Ehrlichia ruminantium str. Welgevonden B Alphaproteobacteria 1.51 27.5 1 CR925678 NC_006832 02/04/2005 CIRAD, France
Ehrlichia ruminantium str. Welgevonden B Alphaproteobacteria 1.52 27.5 1 CR767821 NC_005295 01/07/2005 University of Pretoria, South Africa
Enterococcus faecalis V583 B Firmicutes 3.36 37.4 1 3 AE016830 NC_004668 03/29/2003 TIGR
Erwinia carotovora subsp. atroseptica SCRI1043 B Gammaproteobacteria 5.06 51 1 BX950851 NC_004547 07/21/2004 Welcome Trust Sanger Institute
Erythrobacter litoralis HTCC2594 B Alphaproteobacteria 3.05 63.1 1 CP000157 NC_007722 03/11/2005 J. Craig Venter Institute
Escherichia coli CFT073 B Gammaproteobacteria 5.23 50 1 AE014075 NC_004431 12/10/2002 University of Wisconsin-Madison, USA
Escherichia coli K12 B Gammaproteobacteria 4.64 50 1 U00096 NC_000913 09/05/1997 University of Wisconsin-Madison, USA
Escherichia coli O157:H7 B Gammaproteobacteria 5.59 50 1 2 BA000007 NC_002695 03/29/2000 Osaka Univ.
Escherichia coli O157:H7 EDL933 B Gammaproteobacteria 5.62 50 1 1 AE005174 NC_002655 02/24/2001 University of Wisconsin-Madison, USA
Escherichia coli W3110 B Gammaproteobacteria 4.65 1 AP009048 AC_000091 01/23/2006 Nara Institute of Science and Technology
Francisella tularensis subsp. holarctica B Gammaproteobacteria 1.9 32.2 1 AM233362 NC_007880 03/02/2006 Lawrence Livermore National Laboratory
Francisella tularensis subsp. tularensis SCHU S4 B Gammaproteobacteria 1.89 32.3 1 AJ749949 NC_006570 12/17/2004 Swedish Defence Research Agency
Frankia sp. CcI3 B Actinobacteria 5.43 70.1 1 CP000249 NC_007777 06/20/2005 DOE Joint Genome Institute
Fusobacterium nucleatum subsp. nucleatum ATCC 25586 B Fusobacteria 2.17 27 1 AE009951 NC_003454 03/13/2002 Integrated Genomics

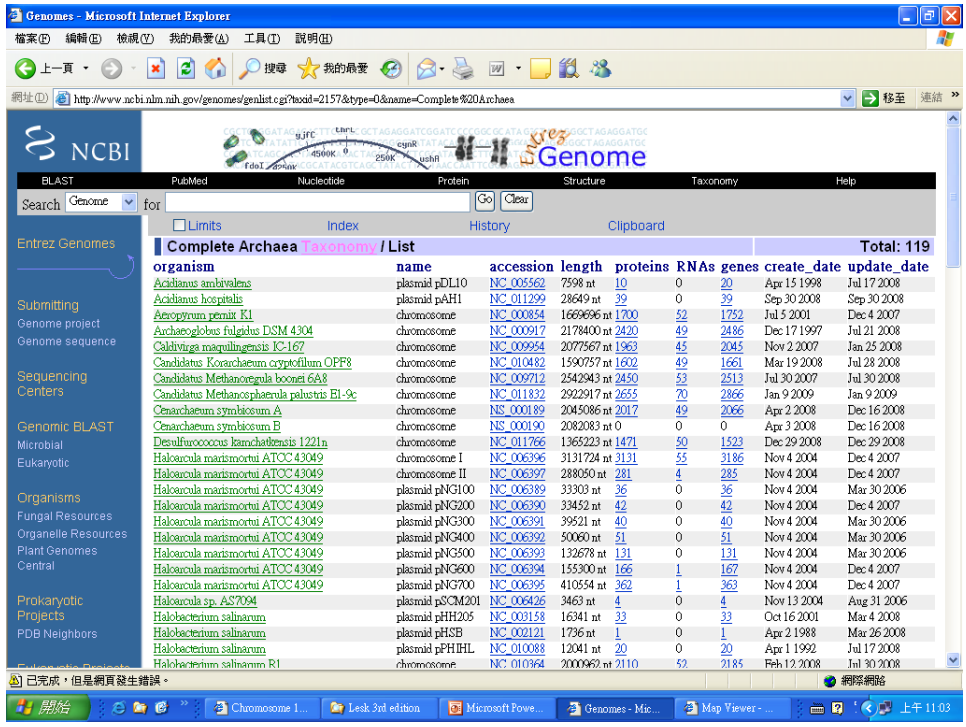

Genome Result - Microsoft Internet Explorer
http://www.ncbi.nlm.nih.gov/sites/entrez?db=genome&cmd=Retrieve&dopt=Overview&list_uids=115
Genome > Bacteria > Escherichia coli str. K-12 substr. MG1655, complete genome
Links
Lineage: Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae; Escherichia; Escherichia coli; Escherichia coli K-12; Escherichia coli str. K-12 substr. MG1655
Genome Info: | Features: | BLAST homologs: | Links: | Review Info:
Refseq: NC_000913 | Genes: 4466 | COG | Genome Project | Publications: [3]
GenBank: U00096 | Protein coding: 4131 | TaxMap | Refseq FTP | Refseq Status: PROVISIONAL
Length: 4,639,675 nt | Structural RNAs: 172 | TaxPlot | GenBank FTP | Seq.Status: Completed
GC Content: 50% | Pseudo genes: 168 | GenePlot | BLAST | Sequencing center: Univ. Wisconsin
% Coding: 85% | Others: 578 | gMap | TraceAssembly | Completed: 2001/10/15
Topology: circular | Contigs: None | CDD | Organism Group
Molecule: DNA | Other genomes for species: 141
Gene Classification based on COG functional categories
Search gene, GeneID or locus_tag:
Find Gene
Zoom
1 nt
9,893 nt
4639675 nt
thrA
thrB
thrC
yaaA
yaaJ
talB
mog

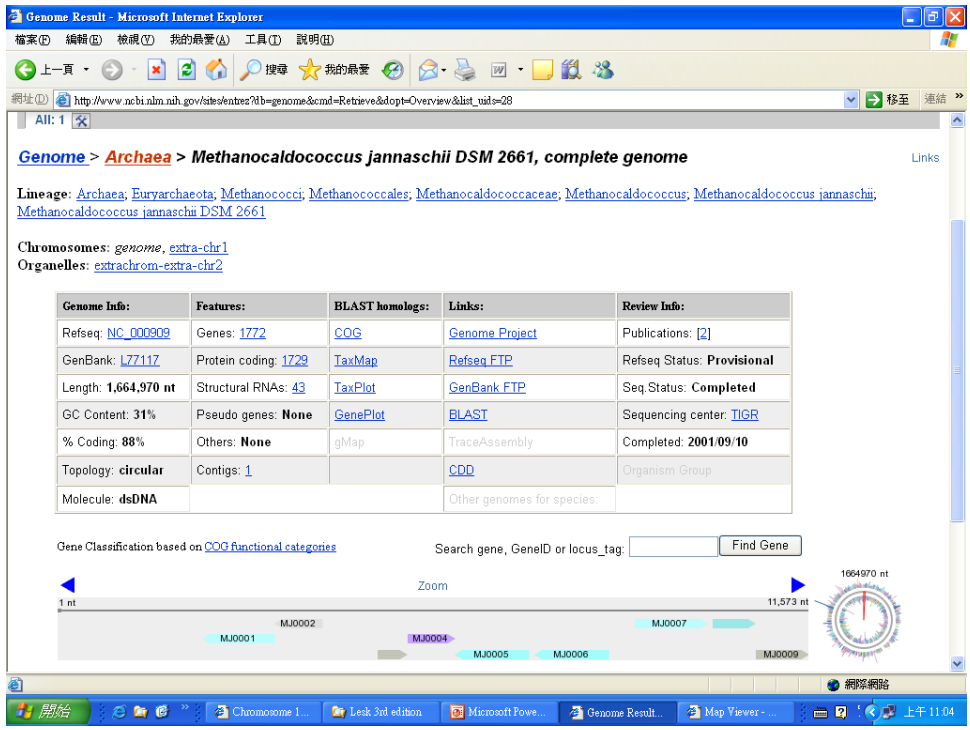


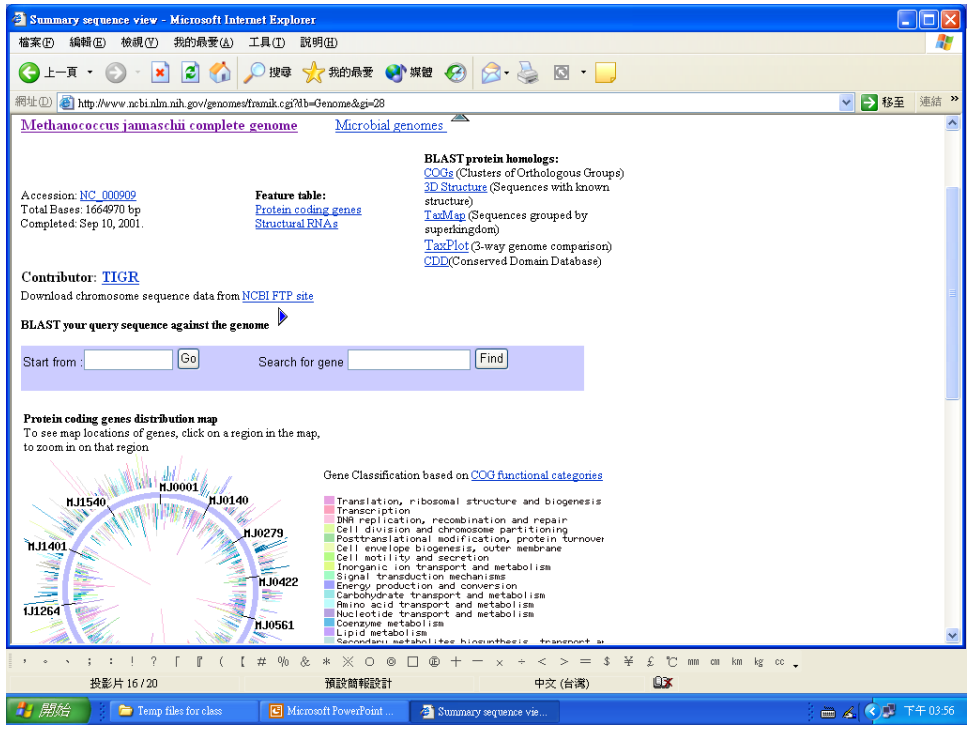
Archaea 古生菌

- 古甲烷球菌基因體：
  - 屬於古細菌，原核生物
  - 可適應於嚴苛環境條件如高溫、高壓、或高鹽濃度
  - 比起細菌，在某方面上古細菌更與真核生物具有親密關係
  - 蛋白質所涉入的transcription, translation和調節作用與真核生物較為接近
  - 但是代謝作用則與細菌較為接近

- Genome of archaeon Methanococcus :
  - Belongs to archaea, prokaryotes
  - Can adapt to harsh environmental conditions such as high temperature, high pressure, or high salt concentration
  - Archaea are in some ways more closely related to eukaryotes than bacteria
  - The transcription, translation and regulatory functions involved in proteins are closer to those of eukaryotes
  - But the metabolism is closer to that of bacteria

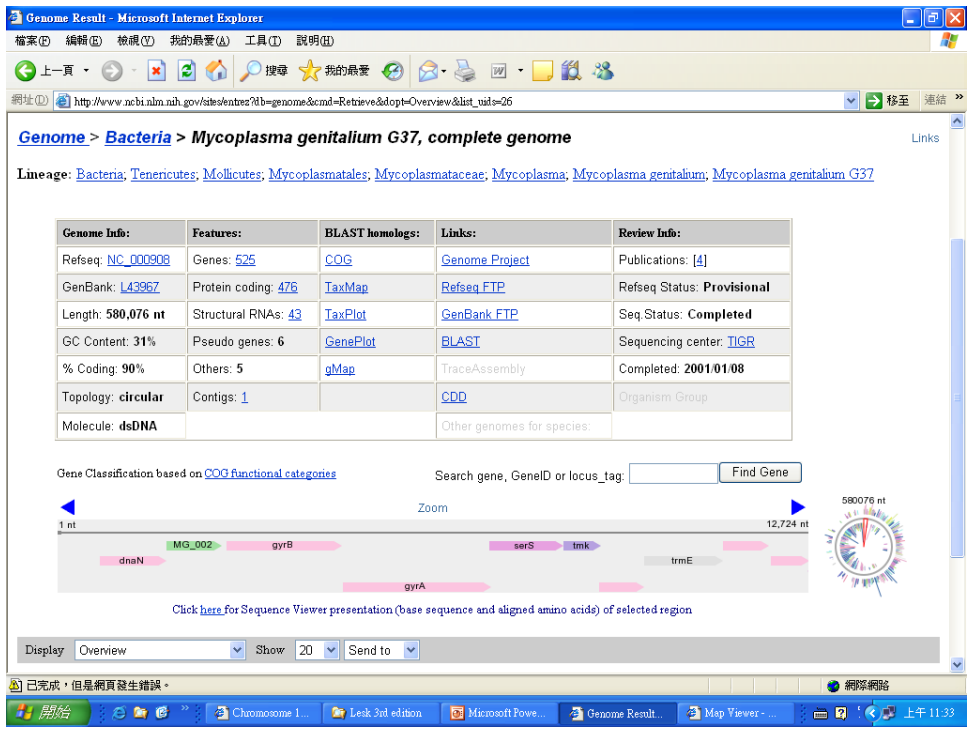


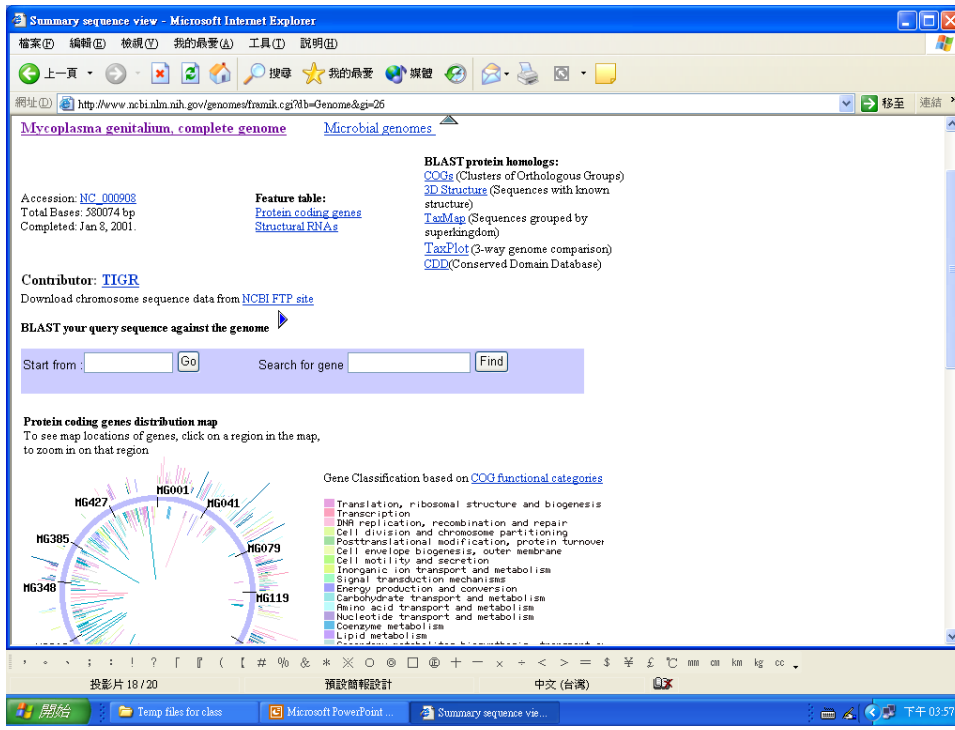

Genome Result - Microsoft Internet Explorer
http://www.ncbi.nlm.nih.gov/sites/entrez?db=genome&cmd=Retrieve&dopt=Overview&list_uids=28
Genome > Archaea > Methanocaldococcus jannaschii DSM 2661, complete genome
Links
Lineage: Archaea; Euryarchaeota; Methanococci; Methanococcales; Methanocaldococcaceae; Methanocaldococcus; Methanocaldococcus jannaschii; Methanocaldococcus jannaschii DSM 2661
Chromosomes: genome, extra-chr1
Organelles: extrachrom-extra-chr2
Genome Info:
Features:
BLAST homologs:
Links:
Review Info:
Refseq: NC_000909
Genes: 1772
COG
Genome Project
Publications: [2]
GenBank: L77117
Protein coding: 1729
TaxMap
Refseq FTP
Refseq Status: Provisional
Length: 1,664,970 nt
Structural RNAs: 43
TaxPlot
GenBank FTP
Seq.Status: Completed
GC Content: 31%
Pseudo genes: None
GenePlot
BLAST
Sequencing center: TIGR
% Coding: 88%
Others: None
gMap
TraceAssembly
Completed: 2001/09/10
Topology: circular
Contigs: 1
CDD
Organism Group
Molecule: dsDNA
Other genomes for species:
Gene Classification based on COG functional categories
Search gene, GeneID or locus_tag:
Find Gene
Zoom
1 nt
11,573 nt
1664970 nt
MJ0001
MJ0002
MJ0004
MJ0005
MJ0006
MJ0007
MJ0009


Summary sequence view - Microsoft Internet Explorer
http://www.ncbi.nlm.nih.gov/genomes/framik.cgi?db=Genome&gi=28
Methanococcus jannaschii complete genome
Microbial genomes
BLAST protein homologs:
COGs (Clusters of Orthologous Groups)
3D Structure (Sequences with known structure)
TaxMap (Sequences grouped by superkingdom)
TaxPlot (3-way genome comparison)
CDD(Conserved Domain Database)
Accession: NC_000909
Total Bases: 1664970 bp
Completed: Sep 10, 2001.
Feature table:
Protein coding genes
Structural RNAs
Contributor: TIGR
Download chromosome sequence data from NCBI FTP site
BLAST your query sequence against the genome
Start from :
Go
Search for gene
Find
Protein coding genes distribution map
To see map locations of genes, click on a region in the map, to zoom in on that region
Gene Classification based on COG functional categories
MJ0001
MJ0140
MJ0279
MJ0422
MJ0561
MJ1264
MJ1401
MJ1540
Translation, ribosomal structure and biogenesis
Transcription
DNA replication, recombination and repair
Cell division and chromosome partitioning
Posttranslational modification, protein turnover
Cell envelope biogenesis, outer membrane
Cell motility and secretion
Inorganic ion transport and metabolism
Signal transduction mechanisms
Energy production and conversion
Carbohydrate transport and metabolism
Amino acid transport and metabolism
Nucleotide transport and metabolism
Coenzyme metabolism
Lipid metabolism

最簡單微生物之一：黴漿菌基因體
One of the simplest microorganisms: Mycoplasma genome

- 是一種感染細菌(an infectious bacterium)
- 基因體為單一DNA分子，含有580070 bp (The genome is a single DNA molecule containing 580070 bp)
- 至今日為止，黴漿菌是被解碼的基因體中最小的(To date, Mycoplasma has the smallest genome to be decoded)
- 最接近微小生物、最小且具獨立的生命體(The smallest and independent living organism closest to microorganisms)
- 85%序列具有編碼功能，約有468個基因(85% of the sequences have coding functions, about 468 genes)

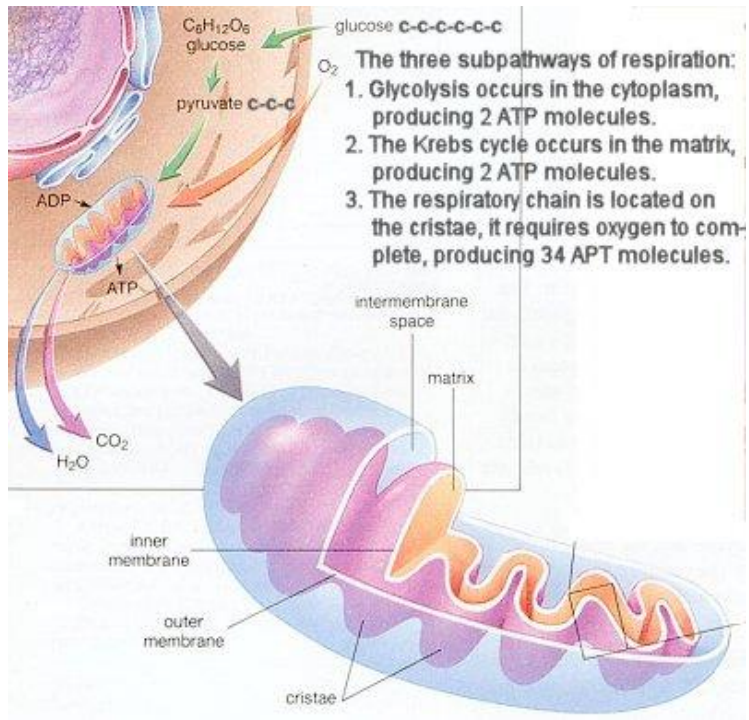


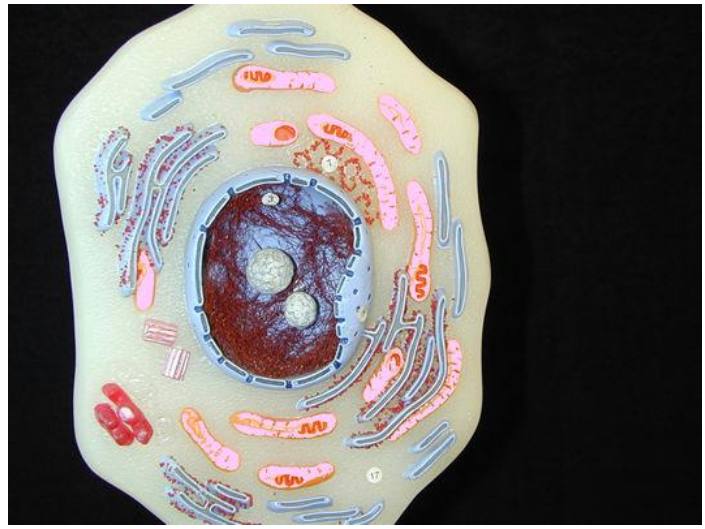


真核生物基因體 (Genome of eukaryotes)

- 酵母菌 (baker's yeast)基因體
- 線蟲 (C. elegans)基因體
- 果蠅 (Fruit fly, Drosophila melanogaster)基因體
- 阿拉伯芥(Arabidopsis thaliana)基因體

- Majority of DNA is in the nucleus, separated into bundles of nucleoprotein, the chromosomes.
  - Each chromosome contains a single double-stranded DNA molecule.
  - Smaller amounts of DNA appear in organelles (胞器) – mitochondria (粒線體) and chloroplasts (葉綠體)
- Nuclear genomes of different species vary widely in size
  - Correlation between genome size and complexity of the organism is very rough
  - In many cases, reflect different amounts of simple repetitive sequences, often referred to as “junk DNA.”
    - Garbage you throw away, junk you keep around!





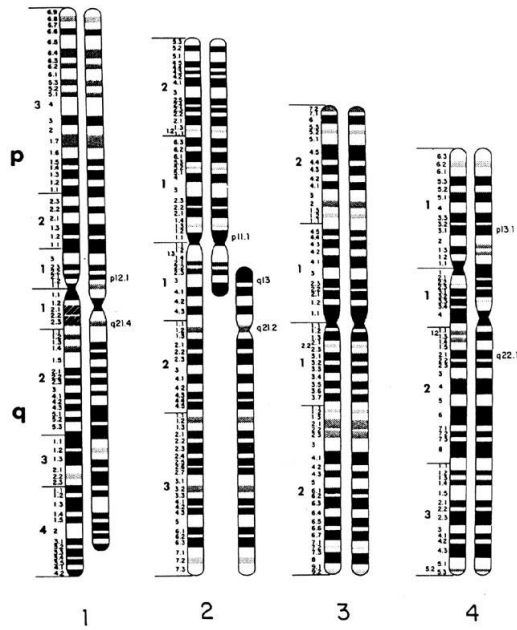
Inventory (細目) of an eukaryotic genome

- 調節性重複DNA (moderately repetitive DNA)
  - 功能性 functional
    - 分散基因家族 dispersed gene families
    - 串聯基因家族陣列 tandem gene family arrays
  - 未知功能 without known function
    - SINES: short interspersed elements
      - 200—300 bp long
      - 散落配置 scattered locations
    - LINES: long interspersed elements
      - 1—5 kbp
      - 每個基因有10—10000複製體 (10-10000 copies per gene)
      - 偽基因 pseudogenes

- 高度重複DNA (highly repetitive DNA)
  - 小衛星基座 minisatellites
    - 14—500 bp片段長的重複基因所組成 (composed of repeats of 14-500 bp segments)
    - 1—5kbp long
    - Many different ones
    - Scattered throughout the genome
  - 微衛星基座 microsatellites
    - Composed of repeats of up to 13 bp
    - ~100s of kb long
    - ~10^6 copies/genome
  - 端粒 telomeres
    - Contain a short repeat unit
    - 250—1000 repeats at the end of each chromosome

Difference with genome of eukaryotes

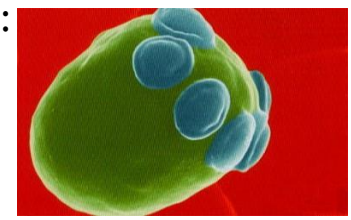
- Variation in DNA content
- Vary in the number of chromosomes and distribution of genes among them
  - Some differences in the distribution of genes among chromosomes involve translocations （跳躍作用）, or chromosome fragmentations (分裂) or joins (結合).
  - For example, human (23 pairs) and chimpanzees (24 pairs), see Fig. 2.3



Human and Chimpanzee Chromosomes 1-4
(Chromosome on left of each pair is human)

酵母菌基因體
Saccharomyces cerevisiae (baker's yeast)

- One of the simplest known eukaryotic organisms
- Its cells contain a nucleus and other specialized intracellular compartments
- Sequencing of its genome: 1992
- 12057500 bp distributed over 16 chromosomes
- 6172 predicted protein-coding genes:
  - 140 for ribosomal RNAs
  - 40 for small nuclear RNAs
  - 275 transfer RNA genes



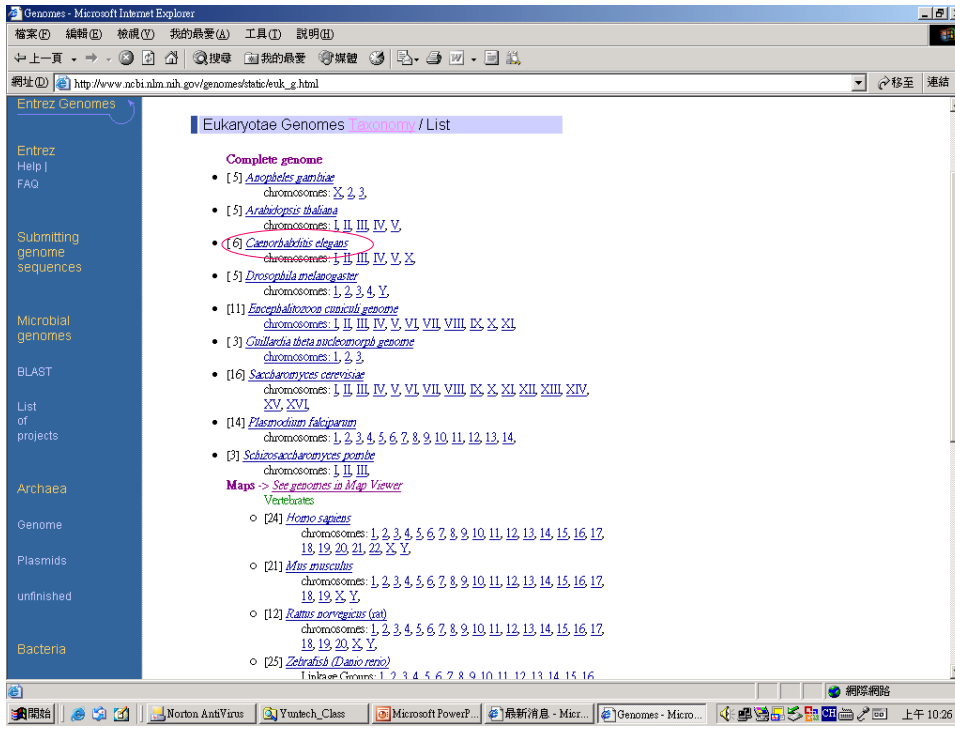
- Yeast genome is denser in coding regions than the known genomes of the more complex eukaryotes (線蟲，人類，果蠅) in two respects:
  - Introns are relatively rare, and relatively small. Only 231 genes in yeast contain introns
  - There are fewer repeat sequences compared with more complex eukaryotes.
- Of 6172 potential genes,
  - 4000~5000, a function can be assigned, with varying degree of confidence
  - ~1000 contain some similarity to known proteins in other species
  - Another ~800 are similar to ORFs in other genomes that correspond to known proteins
    - Many of these homologues appear in prokaryotes
    - Only ~1/3 of yeast proteins have identifiable homologues in the human genome

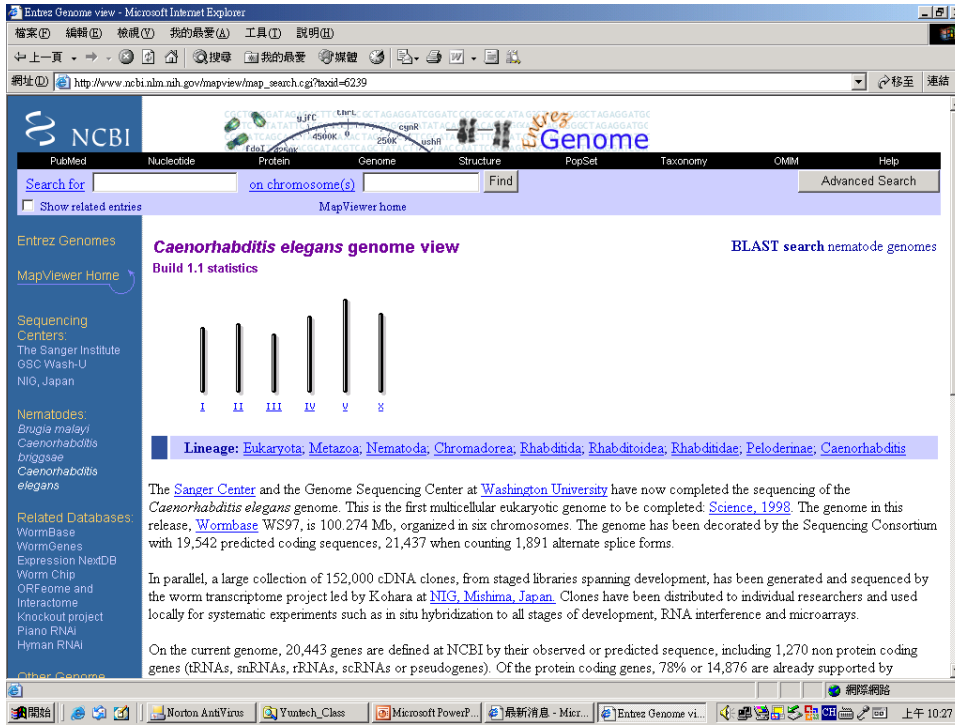
Yeast Genes classification

- http://mips.gsf.de/genre/proj/yeast/

線蟲基因體 Caenorhabditis elegans

- Potential: a sufficiently complex organism to be interesting but simple enough to permit complete analysis, at the cellular level, of its development and neural circuitry.
- Genome was completed in 1998, provided the first full DNA sequence of a multicellular organism.
- ~97 Mbp of DNA (see box, no Y chromosome)

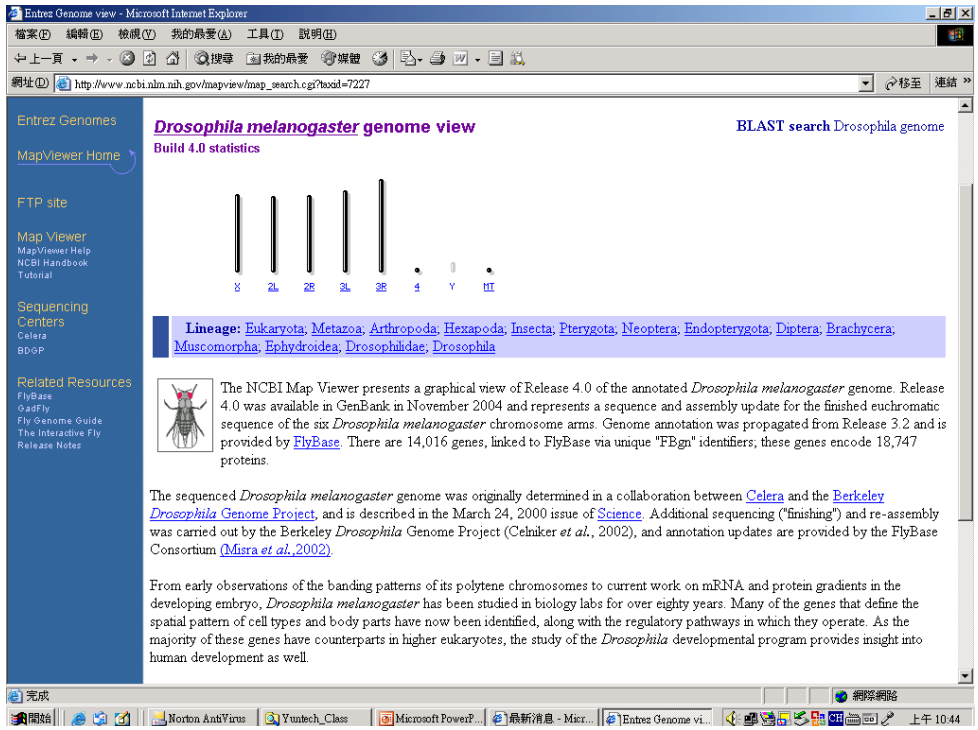

Genomes - Microsoft Internet Explorer
http://www.ncbi.nlm.nih.gov/genomes/static/euk_g.html
Entrez Genomes
Entrez
Help |
FAQ
Submitting genome sequences
Microbial genomes
BLAST
List of projects
Archaea
Genome
Plasmids
unfinished
Bacteria
Eukaryotae Genomes Taxonomy / List
Complete genome
[5] Anopheles gambiae
chromosomes: X, 2, 3,
[5] Arabidopsis thaliana
chromosomes: I, II, III, IV, V,
[6] Caenorhabditis elegans
chromosomes: I, II, III, IV, V, X,
[5] Drosophila melanogaster
chromosomes: 1, 2, 3, 4, Y,
[11] Encephalitozoon cuniculi genome
chromosomes: I, II, III, IV, V, VI, VII, VIII, IX, X, XI,
[3] Guillardia theta nucleomorph genome
chromosomes: 1, 2, 3,
[16] Saccharomyces cerevisiae
chromosomes: I, II, III, IV, V, VI, VII, VIII, IX, X, XI, XII, XIII, XIV, XV, XVI,
[14] Plasmodium falciparum
chromosomes: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14,
[3] Schizosaccharomyces pombe
chromosomes: I, II, III,
Maps -> See genomes in Map Viewer
Vertebrates
[24] Homo sapiens
chromosomes: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, X, Y,
[21] Mus musculus
chromosomes: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, X, Y,
[12] Rattus norvegicus (rat)
chromosomes: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, X, Y,
[25] Zebrafish (Danio rerio)
Linkage Groups: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16
網際網路
開始
Norton AntiVirus
Yuntech_Class
Microsoft PowerP...
最新消息 - Micr...
Genomes - Micro...
上午 10:26

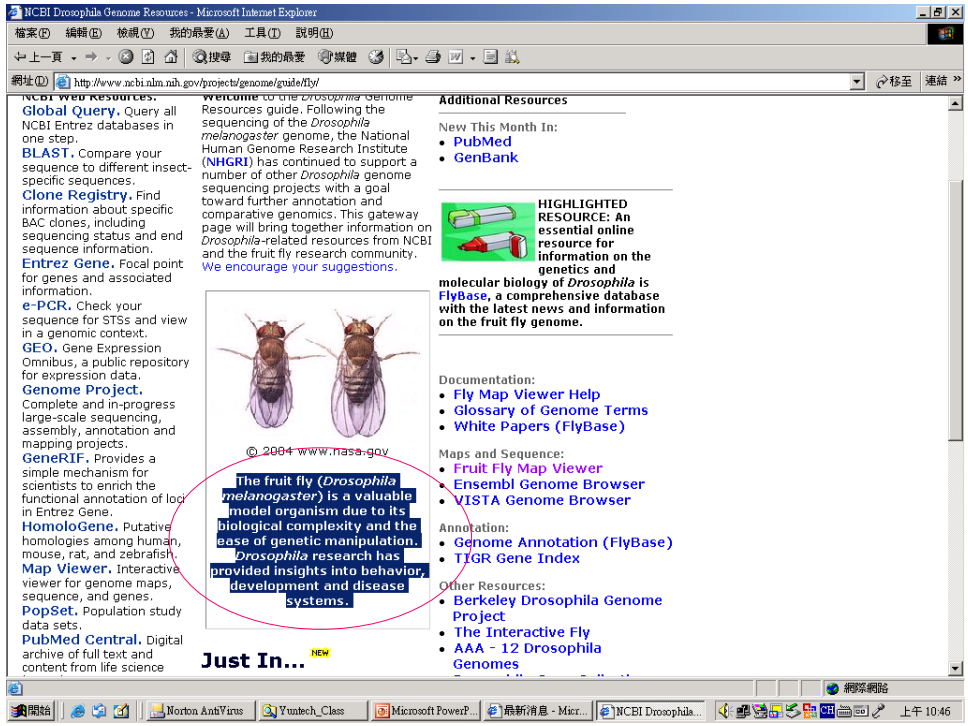

Entrez Genome view - Microsoft Internet Explorer
http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=6239
NCBI
Entrez Genome
PubMed
Nucleotide
Protein
Genome
Structure
PopSet
Taxonomy
OMIM
Help
Search for
on chromosome(s)
Find
Advanced Search
Show related entries
MapViewer home
Entrez Genomes
MapViewer Home
Sequencing Centers:
The Sanger Institute
GSC Wash-U
NIG, Japan
Nematodes:
Brugia malayi
Caenorhabditis briggsae
Caenorhabditis elegans
Related Databases:
WormBase
WormGenes
Expression NextDB
Worm Chip
ORFeome and Interactome
Knockout project
Piano RNAi
Hyman RNAi
Caenorhabditis elegans genome view
Build 1.1 statistics
BLAST search nematode genomes
I
II
III
IV
V
X
Lineage: Eukaryota; Metazoa; Nematoda; Chromadorea; Rhabditida; Rhabditoidea; Rhabditidae; Peloderinae; Caenorhabditis
The Sanger Center and the Genome Sequencing Center at Washington University have now completed the sequencing of the Caenorhabditis elegans genome. This is the first multicellular eukaryotic genome to be completed: Science, 1998. The genome in this release, Wormbase WS97, is 100.274 Mb, organized in six chromosomes. The genome has been decorated by the Sequencing Consortium with 19,542 predicted coding sequences, 21,437 when counting 1,891 alternate splice forms.
In parallel, a large collection of 152,000 cDNA clones, from staged libraries spanning development, has been generated and sequenced by the worm transcriptome project led by Kohara at NIG, Mishima, Japan. Clones have been distributed to individual researchers and used locally for systematic experiments such as in situ hybridization to all stages of development, RNA interference and microarrays.
On the current genome, 20,443 genes are defined at NCBI by their observed or predicted sequence, including 1,270 non protein coding genes (tRNAs, snRNAs, rRNAs, scRNAs or pseudogenes). Of the protein coding genes, 78% or 14,876 are already supported by
網際網路
開始
Norton AntiVirus
Yuntech_Class
Microsoft PowerP...
最新消息 - Micr...
Entrez Genome vi...
上午 10:27

- C. elegans genome is about 8 times larger than that of yeast
- 19,099 predicted coding sequences, 21,437 when counting 1,891 alternate splice forms, ~ 3 times the number in yeast.
- Gene density is relatively low: ~ 1 gene/5 kb
- Exons cover ~27% of the genome; genes contain an average of 5 introns each.
- 42% proteins have homologues outside the phylum (門種)
- 34% are homologous to proteins of other nematodes (線蟲物種)
- 24% have no known homologues outside C. elegans itself.
- Genome contains many repeat sequences

果蠅 Drosophila melanogaster

- Genome sequence was announced in 1999
- Chromosomes are nucleoprotein complexes
  - ~1/3 is contained in heterochromatin (異染色質), highly coiled and compact regions
  - The other 2/3 is euchromatin (真染色質), a relatively uncoiled, less-compact form, most of the active genes are in.
- Total chromosomal DNA contains about 180 Mbp
- 13601 genes, ~double the number in yeast, but are fewer than in C. elegans


Entrez Genome view - Microsoft Internet Explorer
http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=7227
Entrez Genomes
MapViewer Home
FTP site
Map Viewer
MapViewer Help
NCBI Handbook
Tutorial
Sequencing Centers
Celera
BDGP
Related Resources
FlyBase
GadFly
Fly Genome Guide
The Interactive Fly
Release Notes
Drosophila melanogaster genome view
Build 4.0 statistics
BLAST search Drosophila genome
X 2L 2R 3L 3R 4 Y MT
Lineage: Eukaryota; Metazoa; Arthropoda; Hexapoda; Insecta; Pterygota; Neoptera; Endopterygota; Diptera; Brachycera; Muscomorpha; Ephydroidea; Drosophilidae; Drosophila
The NCBI Map Viewer presents a graphical view of Release 4.0 of the annotated Drosophila melanogaster genome. Release 4.0 was available in GenBank in November 2004 and represents a sequence and assembly update for the finished euchromatic sequence of the six Drosophila melanogaster chromosome arms. Genome annotation was propagated from Release 3.2 and is provided by FlyBase. There are 14,016 genes, linked to FlyBase via unique "FBgn" identifiers; these genes encode 18,747 proteins.
The sequenced Drosophila melanogaster genome was originally determined in a collaboration between Celera and the Berkeley Drosophila Genome Project, and is described in the March 24, 2000 issue of Science. Additional sequencing ("finishing") and re-assembly was carried out by the Berkeley Drosophila Genome Project (Celniker et al., 2002), and annotation updates are provided by the FlyBase Consortium (Misra et al.,2002).
From early observations of the banding patterns of its polytene chromosomes to current work on mRNA and protein gradients in the developing embryo, Drosophila melanogaster has been studied in biology labs for over eighty years. Many of the genes that define the spatial pattern of cell types and body parts have now been identified, along with the regulatory pathways in which they operate. As the majority of these genes have counterparts in higher eukaryotes, the study of the Drosophila developmental program provides insight into human development as well.
完成
開始
上午 10:44


NCBI Drosophila Genome Resources - Microsoft Internet Explorer
http://www.ncbi.nlm.nih.gov/projects/genome/guide/fly/
Global Query. Query all NCBI Entrez databases in one step.
BLAST. Compare your sequence to different insect-specific sequences.
Clone Registry. Find information about specific BAC clones, including sequencing status and end sequence information.
Entrez Gene. Focal point for genes and associated information.
e-PCR. Check your sequence for STSs and view in a genomic context.
GEO. Gene Expression Omnibus, a public repository for expression data.
Genome Project. Complete and in-progress large-scale sequencing, assembly, annotation and mapping projects.
GeneRIF. Provides a simple mechanism for scientists to enrich the functional annotation of loci in Entrez Gene.
HomoloGene. Putative homologies among human, mouse, rat, and zebrafish.
Map Viewer. Interactive viewer for genome maps, sequence, and genes.
PopSet. Population study data sets.
PubMed Central. Digital archive of full text and content from life science
Welcome to the Drosophila Genome Resources guide. Following the sequencing of the Drosophila melanogaster genome, the National Human Genome Research Institute (NHGRI) has continued to support a number of other Drosophila genome sequencing projects with a goal toward further annotation and comparative genomics. This gateway page will bring together information on Drosophila-related resources from NCBI and the fruit fly research community.
We encourage your suggestions.
© 2004 www.nasa.gov
The fruit fly (Drosophila melanogaster) is a valuable model organism due to its biological complexity and the ease of genetic manipulation. Drosophila research has provided insights into behavior, development and disease systems.
Just In...
Additional Resources
New This Month In:
PubMed
GenBank
HIGHLIGHTED RESOURCE: An essential online resource for information on the genetics and molecular biology of Drosophila is FlyBase, a comprehensive database with the latest news and information on the fruit fly genome.
Documentation:
Fly Map Viewer Help
Glossary of Genome Terms
White Papers (FlyBase)
Maps and Sequence:
Fruit Fly Map Viewer
Ensembl Genome Browser
VISTA Genome Browser
Annotation:
Genome Annotation (FlyBase)
TIGR Gene Index
Other Resources:
Berkeley Drosophila Genome Project
The Interactive Fly
AAA - 12 Drosophila Genomes
開始
上午 10:46

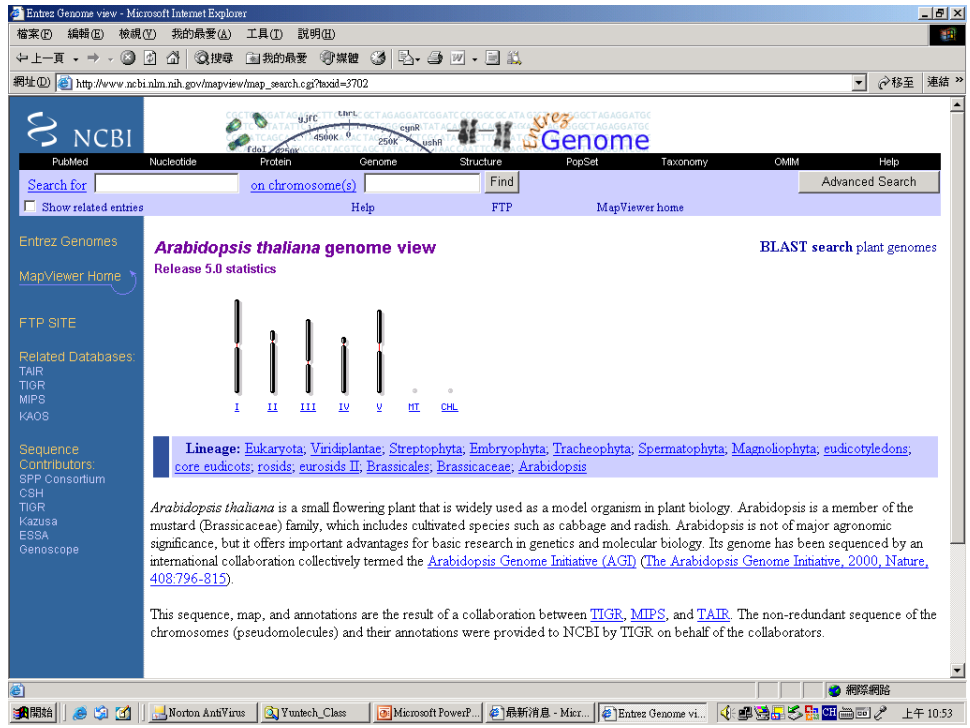
- Contains homologues of 289 human genes implicated in various diseases.
- Non-coding regions must contain regions controlling spatiotemporal patterns of development (時空發展模式)
- It is an organism in which the study of the genomics of development will prove extremely informative.

阿拉伯芥Arabidopsis thaliana

- ~146 Mbp DNA
- 5 pairs of chromosomes, containing 25498 predicted genes
- Genome is relatively compact, with 1 gene/4.6kbp on the average
- Genes are relatively small, exons are typically 250 bp long, and introns relatively small, with mean length 170 bp

Arabidopsis Thaliana

- Typical of plant gene is an enrichment of coding regions in GC content.
- Most proteins have homologues in animals, but some systems are unique to plants, including cell wall production and photosynthesis.
- Traces of large- and small-scale duplication events appear in the genome
- 58% of genome contains 24 duplicated segments, >=100 kbp in length.



Homework #1

- Problem 2.2: 肺炎漿球菌 ***M. genitalium***和嗜血流桿菌***H. influenzae***, what are the values of (a) gene density in genes/kbp, (b) average gene size in bp, (c) number of genes? (d) Which factor contributes most highly to the reduction of genome size in *M. genitalium* relative to *H. influenzae*?

- Weblem 2-19：What chromosome of the ***cow*** contains a region homologous to human chromosome region 8q21.12?

- Please describe the sources of your answers for these two problems.