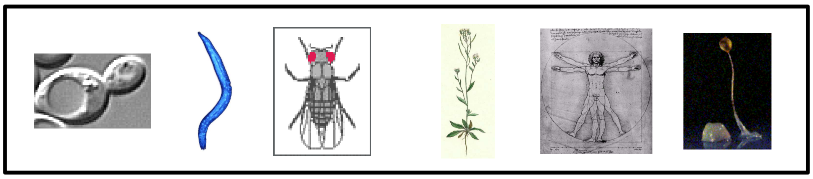
Model Organisms

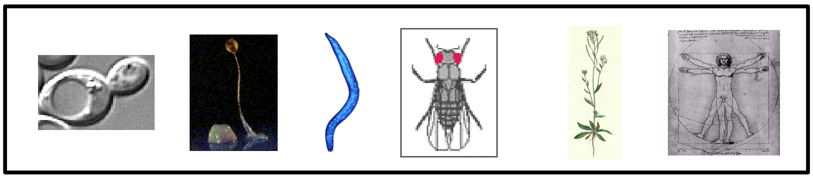
Time Scale for Sequencing

1996 1998 2000 2001 X



Increasing complexity in life systems

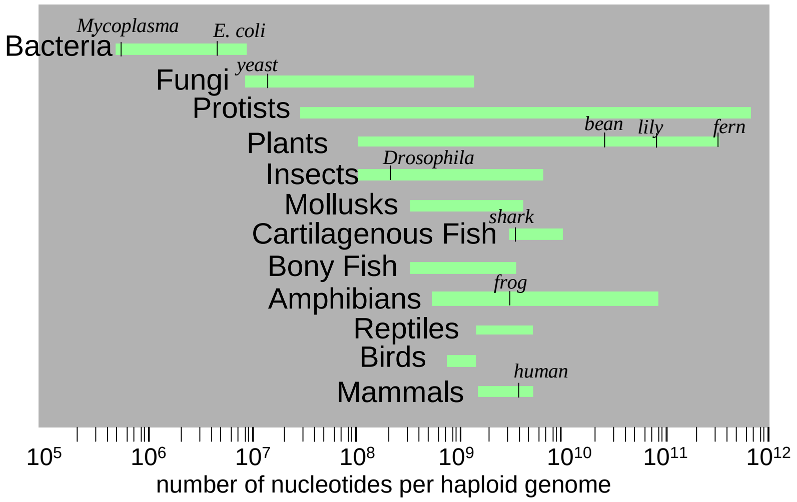
Complexity



What is a Genome?

what		**where**
nuclear genome		
	chromosomes	essential
	extrachromosomal elements	special purposes
	amplified parts of the nuclear genome	
	autonomous elements	e.g. RNA palindrome
	plasmids	mainly in bacteria, but also in eukaria
mitochondrial genome		most eukaryotes
plastid genome		algae and plants

C-Value Paradox



from: 1-38 Molecular Biology of the Cell

Steps in Structural Genome Analysis

Characterisation of the genome (size, A/T content, repeat content)

Mapping (orientation in the genome)

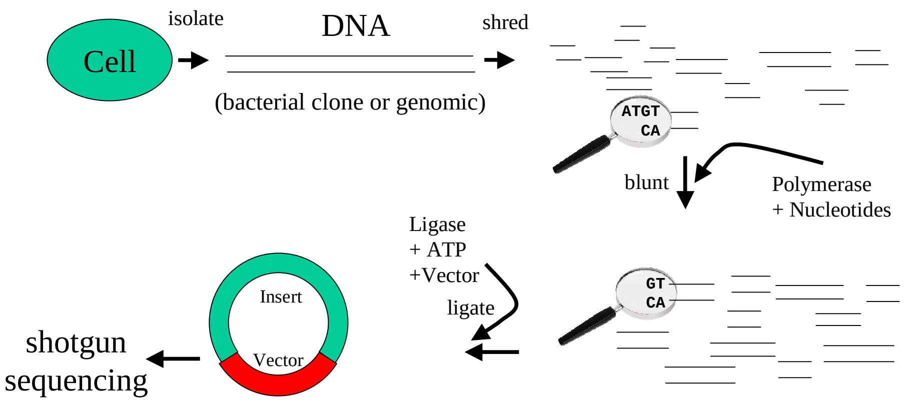
Sequencing (production phase)

Assembly (reconstruction of the genome)

Automated annotation (gene prediction, repetitive elements)

Manual annotation (confirmation of gene models)

Construction of Shotgun Libraries



Shotgun Procedure

Mapping

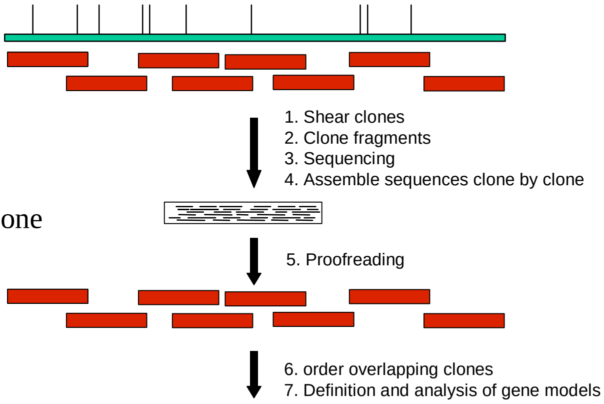
sequence ready map

Production

Assembly of each clone

Reconstruction

Annotation



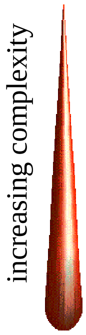
1. Shear clones
2. Clone fragments
3. Sequencing
4. Assemble sequences clone by clone

5. Proofreading

6. order overlapping clones
7. Definition and analysis of gene models

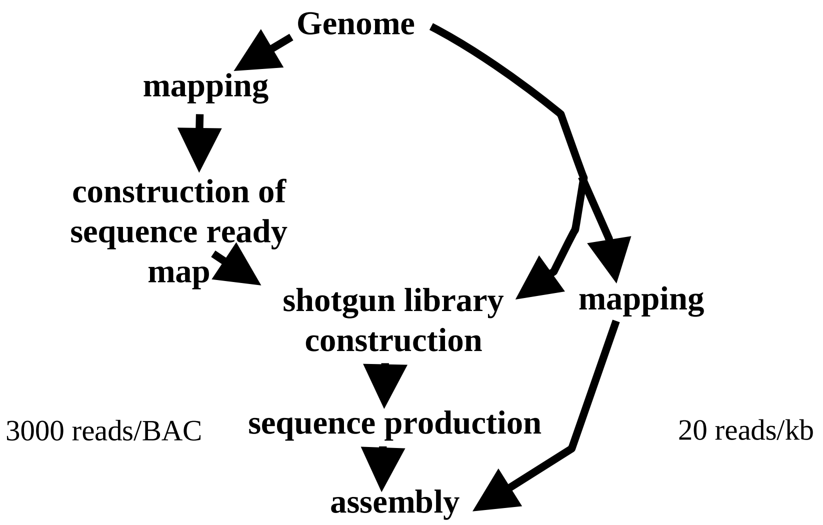
Completely annotated sequence of chromosome

Sequencing Target Complexity



clone	insert size	comment
lambda	20 kb	size limited by phage head
cosmid	40 kb	"
P1	90 kb	"
PAC	150 kb	
BAC	250 kb	
YAC	> 500 kb	
chromosome/ genome	> 1MB	

'clone by clone' versus WCS Strategy



Genome

mapping

construction of sequence ready map

shotgun library construction

mapping

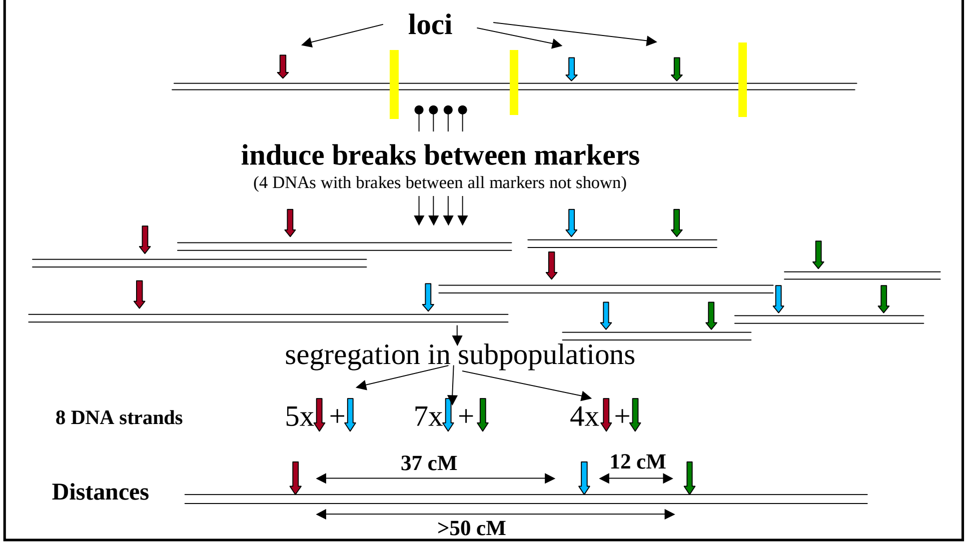
sequence production

assembly

3000 reads/BAC

20 reads/kb

The Basics of Mapping



loci

induce breaks between markers

(4 DNAs with brakes between all markers not shown)

segregation in subpopulations

8 DNA strands

5x + 7x + 4x +

Distances

37 cM

12 cM

>50 cM

Mapping Methods

Genetic

use of meiotic or mitotic crossover events

conjugation (bacteria)

Physical

radiation hybrids (including happy mapping)

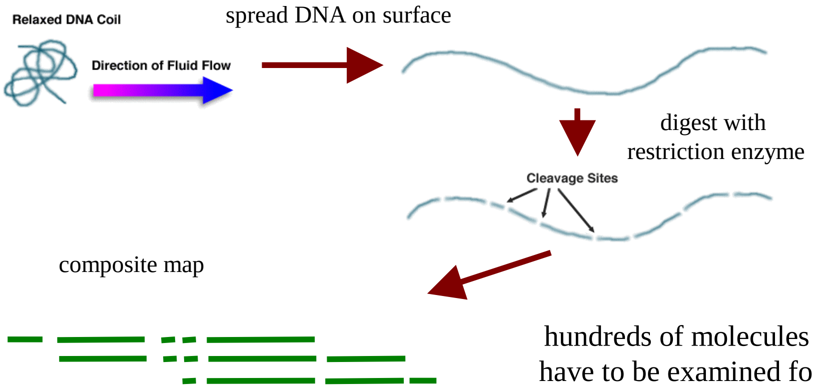
clone map (fingerprinting)

hybridisation of probes to chromosomes or restriction fragments

optical map

sequence

Optical mapping



spread DNA on surface

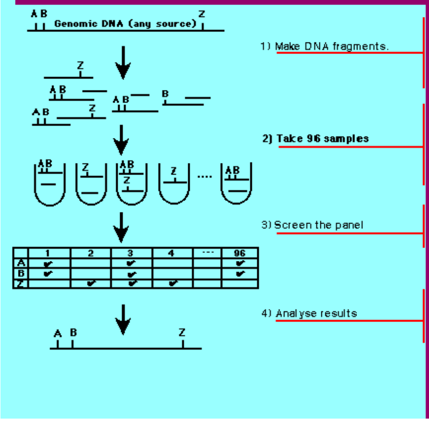
digest with restriction enzyme

composite map

hundreds of molecules have to be examined for a correct map!

Happy Mapping: Procedure

Happy mapping is a *in vitro* method



DNA fragments broken at random

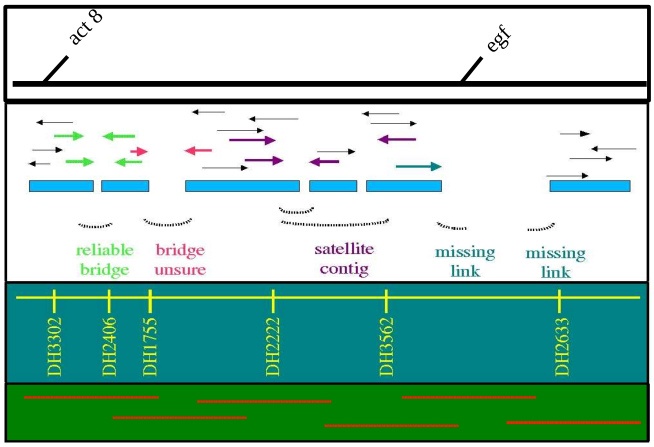
Each well contains < 1 genome equivalent

fragment length determines resolution

PCR screen

Co-segregation frequency determines distance

The Composite Map



genetic markers

scaffolds

STS marker

Clones

Assembly Calculations

probability for a base to be not sequenced: $\mathbf{P_0=e^{-c}}$

total gap length: $\mathbf{G_L=T_Le^{-c}}$

number of gaps $\mathbf{G_N=Ne^{-c}}$

where c=fold coverage
e=2.718
P_0=probability to be not sequenced
G_L=gap length
T_L=target length
G_N=number of gaps
$N=T_L/R_L$=read number for given coverage

Example I

150 kb 500 bases mean read length

fold coverage	Total bases sequenced	e^{-c}	total gap length in bases $=G_Le^{-c}$	Number of Gaps = Ne^{-c}	Gap Length/# gaps= # bases per gap	% complete
1	150000	0.37	55,500	111	500	63
2	300000	0.135	20,250	81	250	87.5
3	450000	0.05	7,500	45	167	95
4	600000	0.018	2,700	22	123	98.2
5	750000	0.0067	1,005	10	101	99.4
6	900000	0.0025	375	5	75	99.75
7	1050000	0.0009	135	2	68	99.91
8	1200000	0.0003	45	1	45	99.97
9	1350000	0.0001	15	1	15	99.99
10	1500000	0.000045	6	1	6	99.995

Example II

4MB 500 bases mean read length

fold	Total bases		total gap length	Number of	Gap Length/# gaps=	%
coverage	sequenced	e^{-c}	in bases $=G_L e^{-c}$	Gaps = Ne^{-c}	# bases per gap	complete
1	4000000	0.37	1,480,000	2960	500	63
2	8000000	0.135	540,000	2160	250	87.5
3	12000000	0.05	200,000	1200	167	95
4	16000000	0.018	72,000	576	125	98.2
5	20000000	0.0067	26,800	268	100	99.4
6	24000000	0.0025	10,000	120	83	99.75
7	28000000	0.0009	3,600	50	72	99.91
8	32000000	0.0003	1,200	19	63	99.97
9	36000000	0.0001	400	7	57	99.99
10	40000000	0.000045	180	4	45	99.995

Genome Sequencing, Assembly, and Mapping

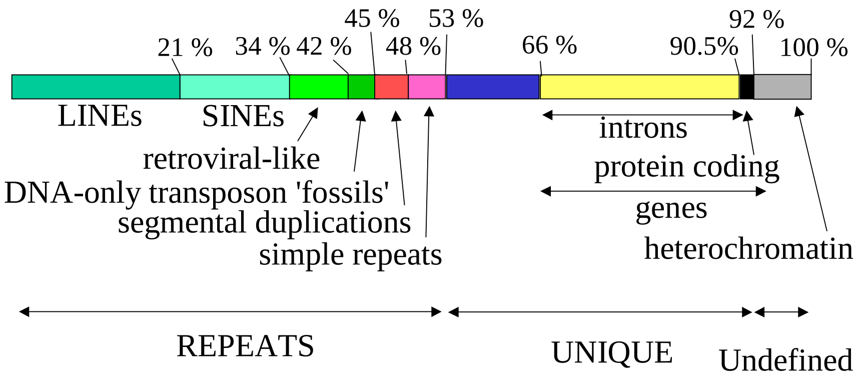
genome features

size, chromosomes, repetitive elements

genome processing

sequencing	**library construction, shogun sequencing**
assembly	**methods, problems**
mapping	**reasons for m., methods**

Percentage of 'Junk' in the Human Genome



21 %
34 %
42 %
45 %
48 %
53 %
66 %
90.5%
92 %
100 %

LINEs
SINEs
retroviral-like
DNA-only transposon 'fossils'
segmental duplications
simple repeats
introns
protein coding
genes
heterochromatin

REPEATS
UNIQUE
Undefined

Automated Annotation

Identification of physical properties

  GC content, triplet usages, etc.

Identification of repetitive elements

  complex repetitive elements, tandem and inverted repeats,

  hairpin structures, etc.

Definition of gene models

  use of different gene prediction programs (sensitive and specific)

EST analysis, mapping onto the genome

Discrimination between Coding and Non-Coding Regions

intergenic regions reflect the overall GC bias of a genome
(nearly random distribution of nucleotides)
genic regions underlie natural selection pressures
(maintenance of functional codons and evolution of function)

Genomic attributes for prediction of genes:

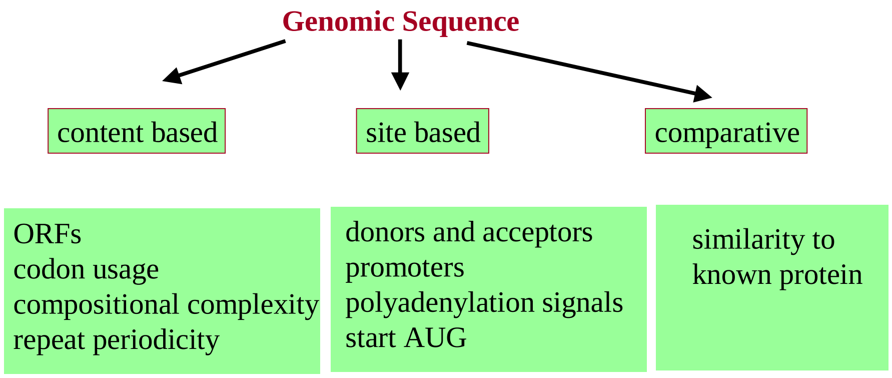
species independent

Base composition differences

species specific

Codon preference, splice site composition

Gene Finding Strategies



Genomic Sequence

content based

site based

comparative

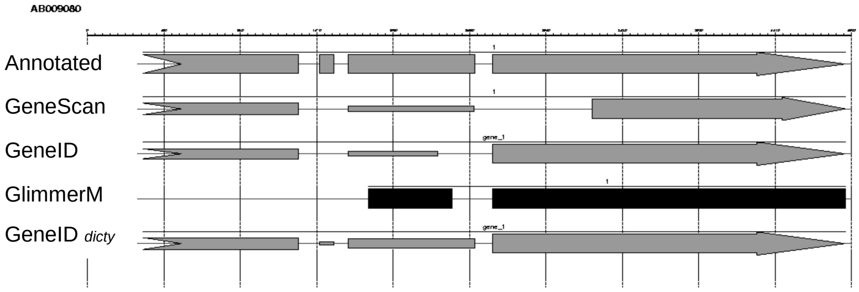
ORFs
codon usage
compositional complexity
repeat periodicity

donors and acceptors
promoters
polyadenylation signals
start AUG

similarity to
known protein

Gene Prediction in Eukaryotes

GlimmerM:	partially trained
GeneID:	fine tuned *Dictyostelium* version

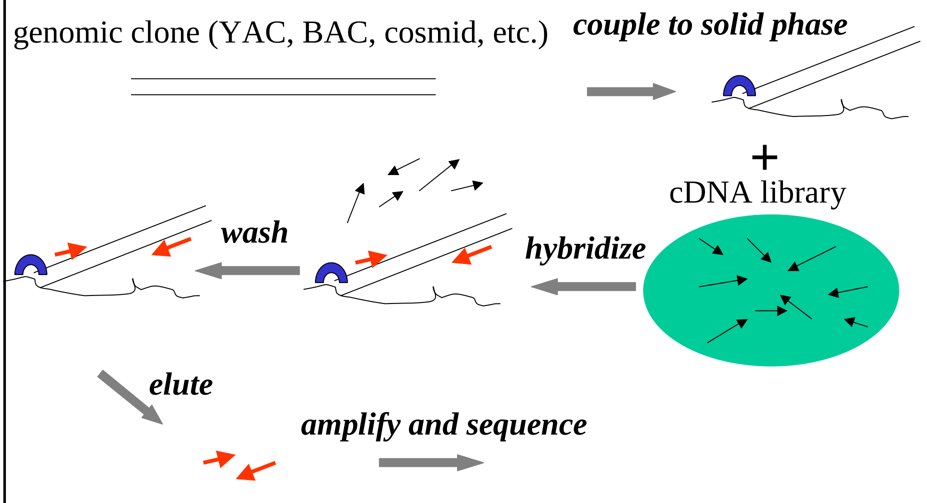


Not all gene structures can be predicted accurately

Experimental Methods for Gene Detection and Verification

Southern +Northern blot	closely related species required, small test sets
cDNA selection	enrichment of specific transcribed sequences
Exon Trapping	many artificial results
isolation of CpG islands	restricted to mammals and birds
temperature sensitive degradation	enrich for high GC DNA

cDNA Selection



genomic clone (YAC, BAC, cosmid, etc.)

couple to solid phase

+

cDNA library

hybridize

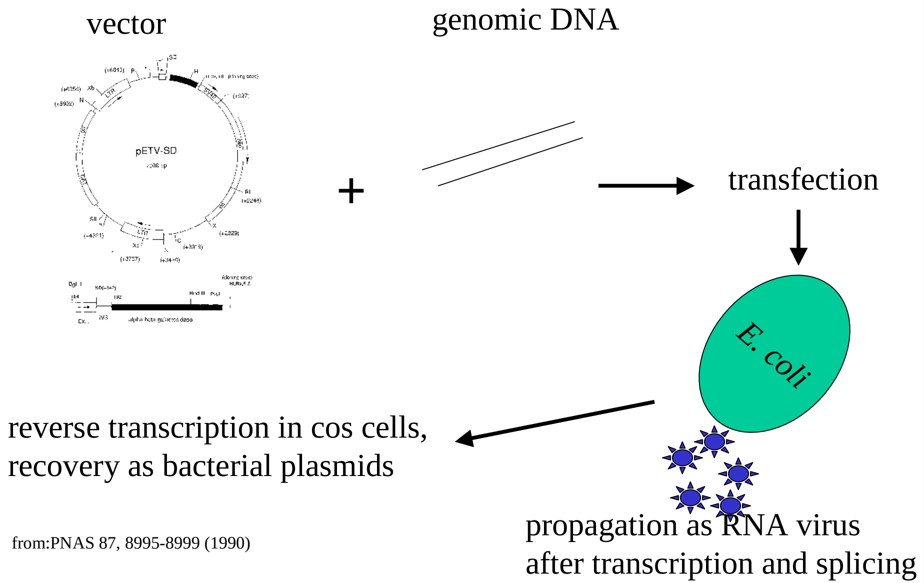
wash

elute

amplify and sequence

from : PNAS 88, 9623-9627 (1991)

Exon Trapping



vector

+

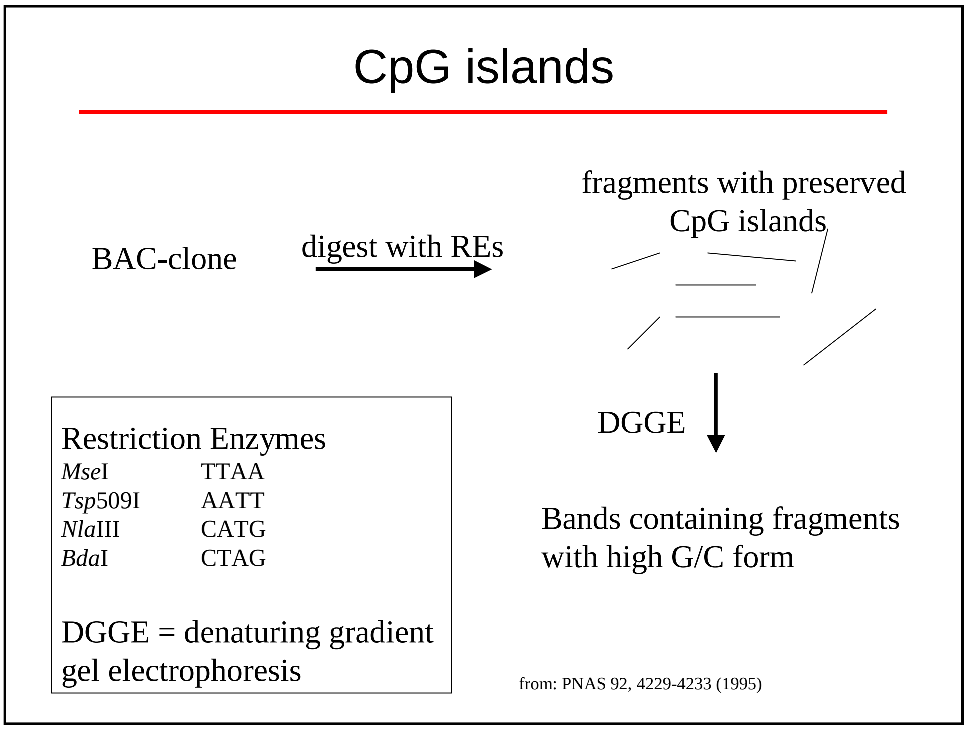
genomic DNA

transfection

reverse transcription in cos cells,
recovery as bacterial plasmids

propagation as RNA virus
after transcription and splicing

from:PNAS 87, 8995-8999 (1990)



CpG islands

BAC-clone

digest with REs

fragments with preserved CpG islands

DGGE

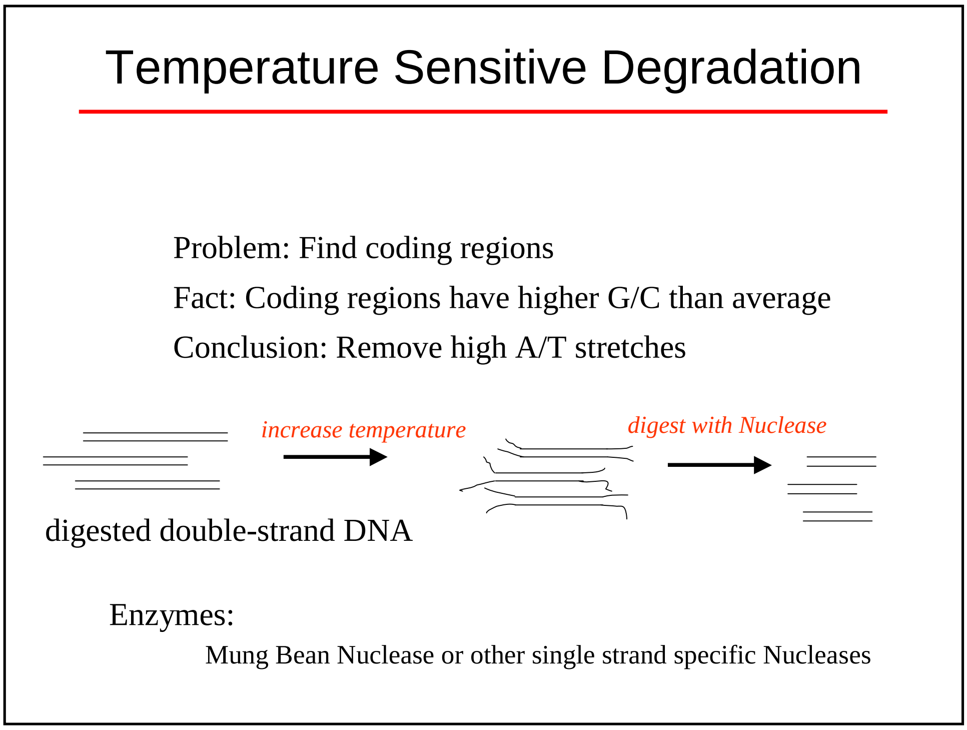
Bands containing fragments with high G/C form

Restriction Enzymes

*Mse*I	TTAA
*Tsp*509I	AATT
*Nla*III	CATG
*Bda*I	CTAG

DGGE = denaturing gradient gel electrophoresis

from: PNAS 92, 4229-4233 (1995)



Temperature Sensitive Degradation

Problem: Find coding regions

Fact: Coding regions have higher G/C than average

Conclusion: Remove high A/T stretches

digested double-strand DNA

Enzymes:

Mung Bean Nuclease or other single strand specific Nucleases

Annotation

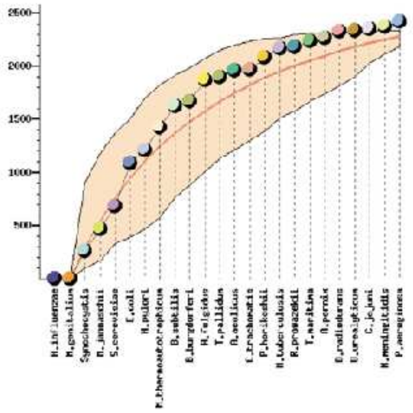
Annotation tools

Data banks		
	GenBank+Embl	DNA and Protein databases
	SwissProt + PIR	annotated proteins database
Clustering		
	COG	clusters of orthologous groups
	Prosite	motif search
	Pfam	protein family domains
	IPR	combination of motif databases
Classification		
	GO	classification system
	MIPSyeast	classification system based on yeast
Structures		
	Brookhaven structure database	

Interpro Domains

Domain	Description	DD	SC	AT	CE	DM	HS
IPR001687	ATP/GTP-binding site motif A (P-loop)	6.07%	0.57%	0.61%	0.32%	0.46%	0.33%
IPR000694	Proline-rich region	3.72%	NA	NA	NA	NA	NA
IPR000561	EGF-like domain	2.18%	0.02%	0.16%	0.68%	0.62%	1.28%
IPR000719	Eukaryotic protein kinase	1.93%	1.91%	4.07%	2.34%	1.79%	2.64%
IPR002290	Serine/Threonine protein kinase	1.89%	1.83%	3.34%	1.33%	1.22%	1.83%
IPR001245	Tyrosine protein kinase	1.71%	0.05%	1.84%	0.84%	0.65%	1.22%
IPR001680	G-protein beta WD-40 repeats	1.11%	1.63%	1.02%	0.80%	1.31%	1.34%
IPR003593	AAA ATPase superfamily	1.11%	0.95%	0.90%	0.40%	0.56%	0.46%
IPR000051	SAM nucleotidebinding motif	0.89%	0.33%	0.40%	0.25%	0.28%	0.20%
IPR001849	Pleckstrin homology (PH) domain	0.89%	0.47%	0.12%	0.41%	0.54%	1.24%
IPR002048	EF-hand	0.86%	0.26%	0.85%	0.65%	0.93%	1.15%
IPR001841	RING finger	0.82%	0.65%	1.82%	0.81%	0.85%	1.20%
IPR002085	Zinc-containing alc. dehyd. superfamily	0.82%	0.34%	0.15%	0.06%	0.07%	0.08%
IPR000794	Beta-ketoacyl synthase	0.79%	0.03%	0.02%	0.02%	0.03%	0.01%

COG Database



Each organism adds new COGs

from: NAR 29, 22-28 (2001)

What is a Model Organism?

A species which qualifies as representative for certain functions/behaviours

trait/function	**exampel**
molecular function	primary metabolism
cell structure	cytoskeleton
motility	flagella
QTL	body weight

Relationships between organisms - the phylogeny - must be known!

Problems Associated with Phylogeny

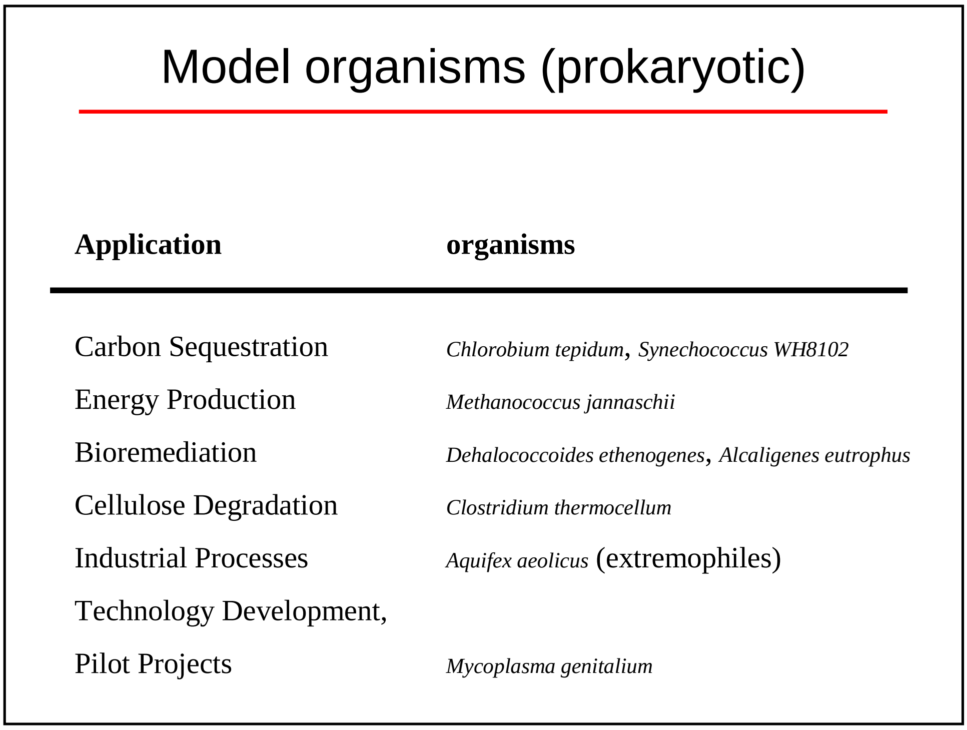
Prokaryotes

Gene duplications, gene losses
horizontal gene transfer
conserved synteny as evolutionary measure

➢ phylogenetic species concept

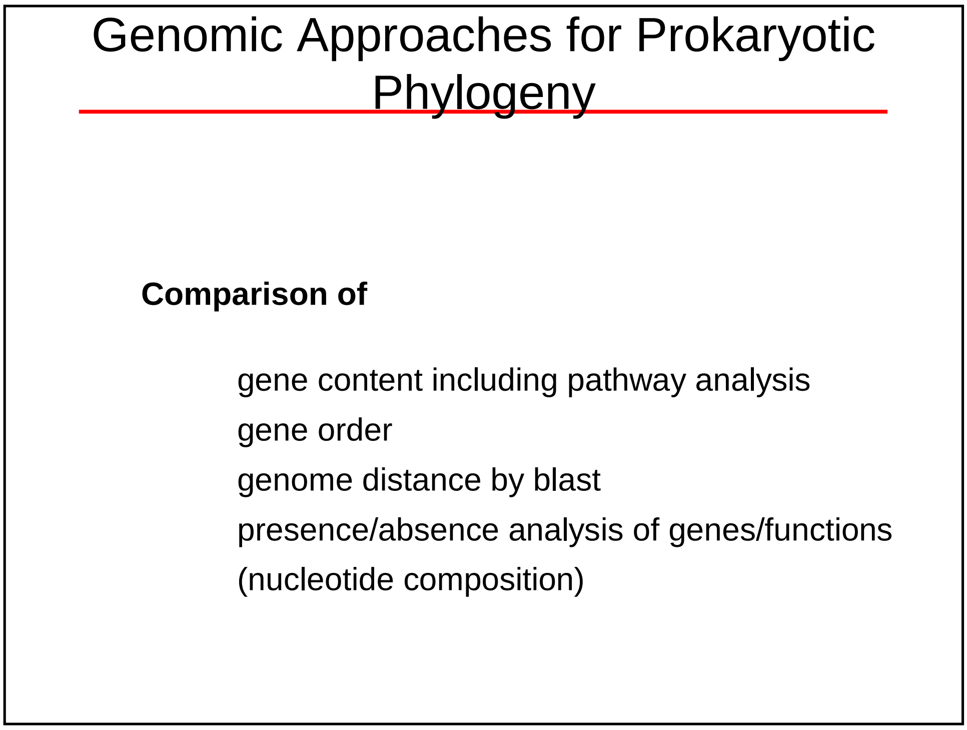
Eukaryotes

genome wide phylogeny hindered by
unclear orthologous relationships caused by
individual domain combinations, adaptations, gene family expansions etc.



Model organisms (prokaryotic)

Application	**organisms**
Carbon Sequestration	*Chlorobium tepidum*, *Synechococcus WH8102*
Energy Production	*Methanococcus jannaschii*
Bioremediation	*Dehalococcoides ethenogenes*, *Alcaligenes eutrophus*
Cellulose Degradation	*Clostridium thermocellum*
Industrial Processes	*Aquifex aeolicus* (extremophiles)
Technology Development,	
Pilot Projects	*Mycoplasma genitalium*



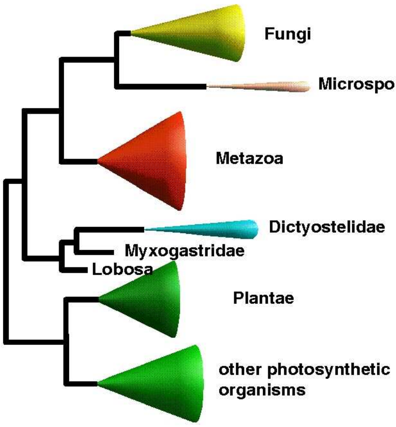
Genomic Approaches for Prokaryotic Phylogeny

Comparison of

gene content including pathway analysis
gene order
genome distance by blast
presence/absence analysis of genes/functions
(nucleotide composition)

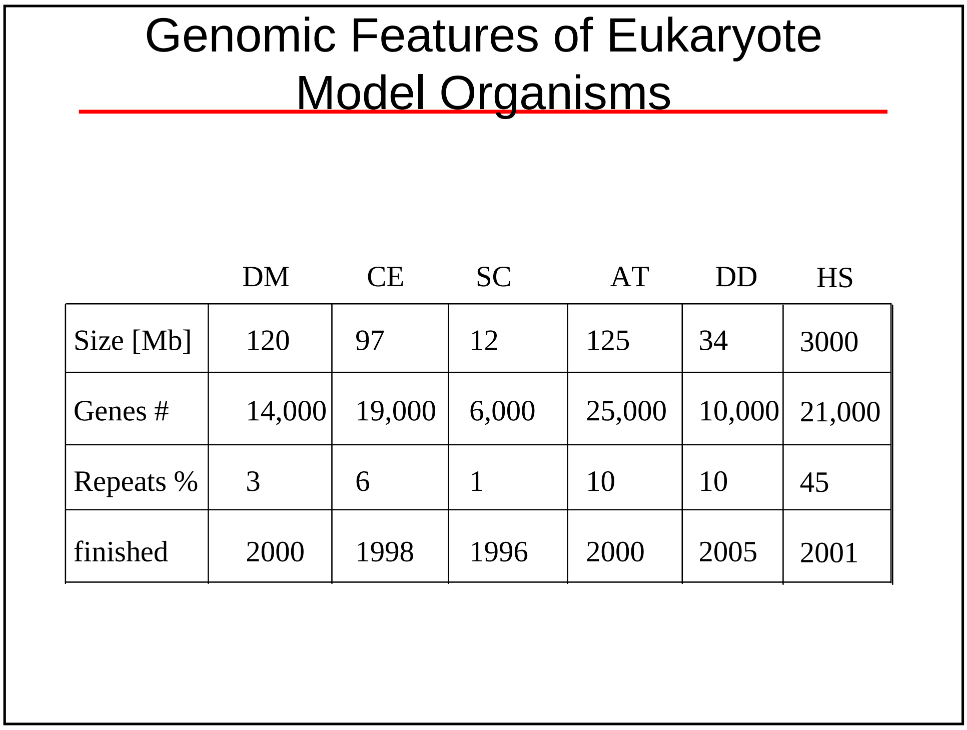
Eukaryotic Phylogeny

[adapted from Baldauf et al. (2000) Science 290, 972-977]



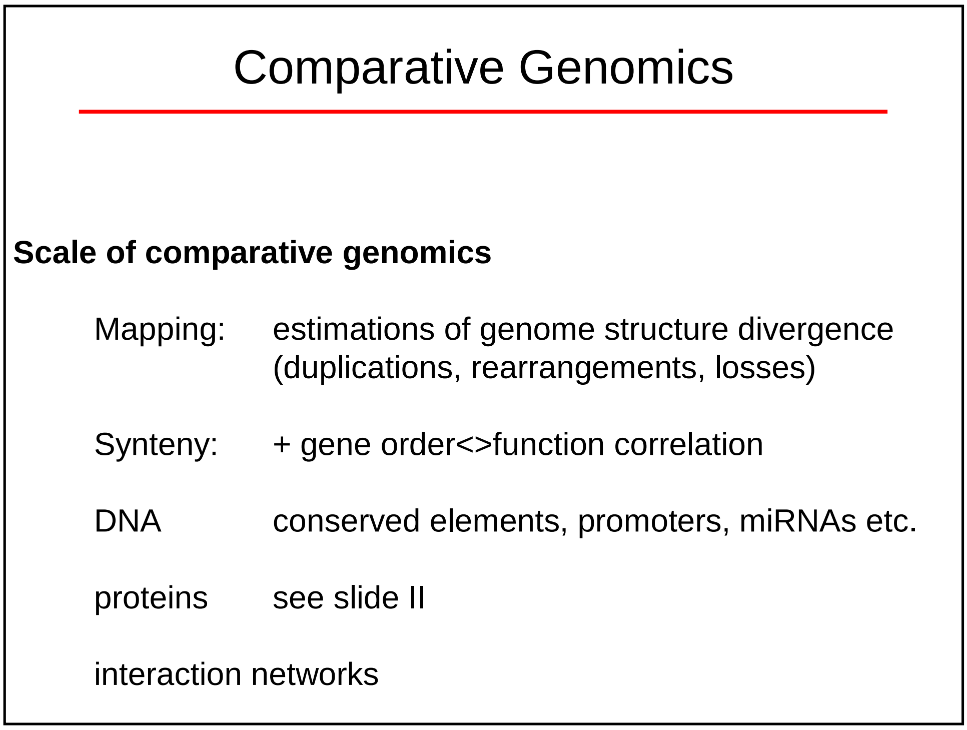
Eukaryotic Model Organisms

Saccharomyces cerevisiae	single eukarytic cell
Dictyostelium discoideum	cell movement, signalling, multicellularity
Caenorhabditis elegans	multicellular organism
Chlamydomonas reinhardtii	'green yeast'
Arabidopsis thaliana	vascular plant
Physcomitrella patens	moss
Danio rerio	vertebrate, development
Fugu rubripes	“, comparative genomics in vertebrates
Rattus rattus	mammal, physiology more similar to Hs than mouse
Mus musculus	"
Homo sapiens	primate



Genomic Features of Eukaryote Model Organisms

	DM	CE	SC	AT	DD	HS
Size [Mb]	120	97	12	125	34	3000
Genes #	14,000	19,000	6,000	25,000	10,000	21,000
Repeats %	3	6	1	10	10	45
finished	2000	1998	1996	2000	2005	2001



Comparative Genomics

Scale of comparative genomics

Mapping:	estimations of genome structure divergence (duplications, rearrangements, losses)
Synteny:	+ gene order<>function correlation
DNA	conserved elements, promoters, miRNAs etc.
proteins	see slide II
interaction networks	

Comparative Genomics II

➢lineage specific genes

➢species specific genes

➢phenotype related traits (multi-species comparisons)

➢gene losses on evolutionary lines

➢new inventions

Yeast

Many genes from yeast have **orthologues genes** in higher eukaryotes. In many cases, functions are strictly conserved, meaning that a human ortholog will function in yeast.

> Cell cycle genes and components of the basal gene expression machinery

Others: similar function, but **specific biological context** and role **differs** between organisms and cell types.

> MAP kinases and other signaling pathway components

Yeast Genome

small genome (12 MB)

~5600 genes

free living single cell

reduced abilities (no motility, phagocytosis, etc.

wide spectrum of manipulation methods

Other model fungi:
Schizosaccharomyces pombe (fission yeast), *Candida albicans*

Caenorhabditis elegans

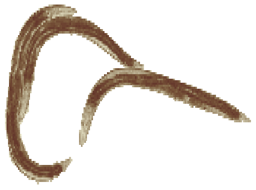
hermaphroditic nematode
developed as model in the 1960s

less than 1,000 constituent cells form an individual animal.
Genetics of development and neurobiology.

ACEDB was developed for the sequencing project

special techniques: RNAi

19.000 genes on 100 MB



Drosophila melanogaster

Originally, it was **the species** to study genetics:

e.g. genes are related to proteins,

the rules of genetic inheritance

Standard map of polytene chromosomes: 102 bands

band 57

Mutant flies of several thousand genes are available

Today: embryogenesis (spatial and temporal patterns),
eye development, behaviour, neuronal development

14.000 genes on 120 MB

Arabidopsis thaliana

Model plant:

small plant

small genome size (125 MB)

related to important crop plants (Brassicaceae)

Many duplications (70 % of the genes)
As with all plants: not easy to manipulate

Alternative: *Physcomitrella patens* can be transformed

A. ROCKENTRAV, TURRITIS GLABRA L.
B. BACKTRAV, ARABIDOPSIS THALIANA (L.) SCHUR.

Repetitive Elements

Classes

- complex
  - LTR- and Non-LTR RNA elements; DNA elements
- simple
  - tandem, inverse repeats; monotone triplet repeats

Impact on genome

- complex
  - contribute to plasticity, potentially involved in speciation
  - genome size
- simple
  - used for genome characterisation (forensics)
  - expansions can cause diseases

Bioinformatics

Tools for genome characterisation

- **prediction**
  - gene finder, promoter analysis, repeat finder
- **protein clustering and domain definition**
  - COG, PFam, Prosite etc.
- **categorisation**
  - MIPS yeast, GO, etc.
- **pathways**
  - KEGG, Biocyc, etc.

Metagenomics

Shotgun analysis of environmental samples

benefits

overview of (unculturable) species in undefined samples
revailing species genome analysis
hypothesis on common prerequisites in a certain niche

drawbacks

overestimation of species numbers due to fragmentation
species with low abundance not well defined
species contigs without relatives are not easily categorised

→ should be flanked by sequencing of cultured species

Science

22 December 2000

Vol. 290 No. 5500
Pages 2201–2372 $8

SEQUENCED GENOMES

Breakthrough of the Year

American Association for the Advancement of Science