"Bacterial" Genome structures
4BIM - INSA 2006

J. R. Lobry

Université Claude Bernard Lyon I – France

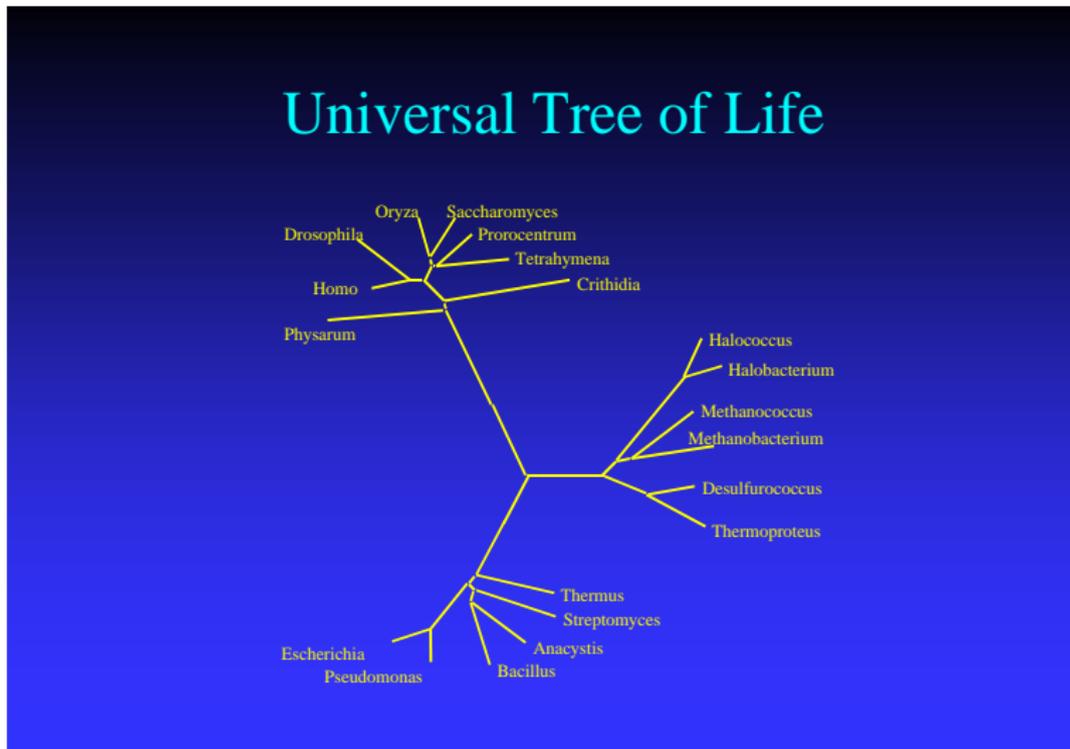18 juin 2006

Outline

Introduction

The three kingdoms

Universal Tree of Life

The three kingdoms : eucarya

Universal Tree of Life

EUCARYA

Oryza Saccharomyces
Drosophila Prorocentrum
Tetrahymena
Homo Crithidia
Physarum

Halococcus
Halobacterium
Methanococcus
Methanobacterium
Desulfurococcus
Thermoproteus

Thermus
Streptomyces
Anacystis
Escherichia Bacillus
Pseudomonas
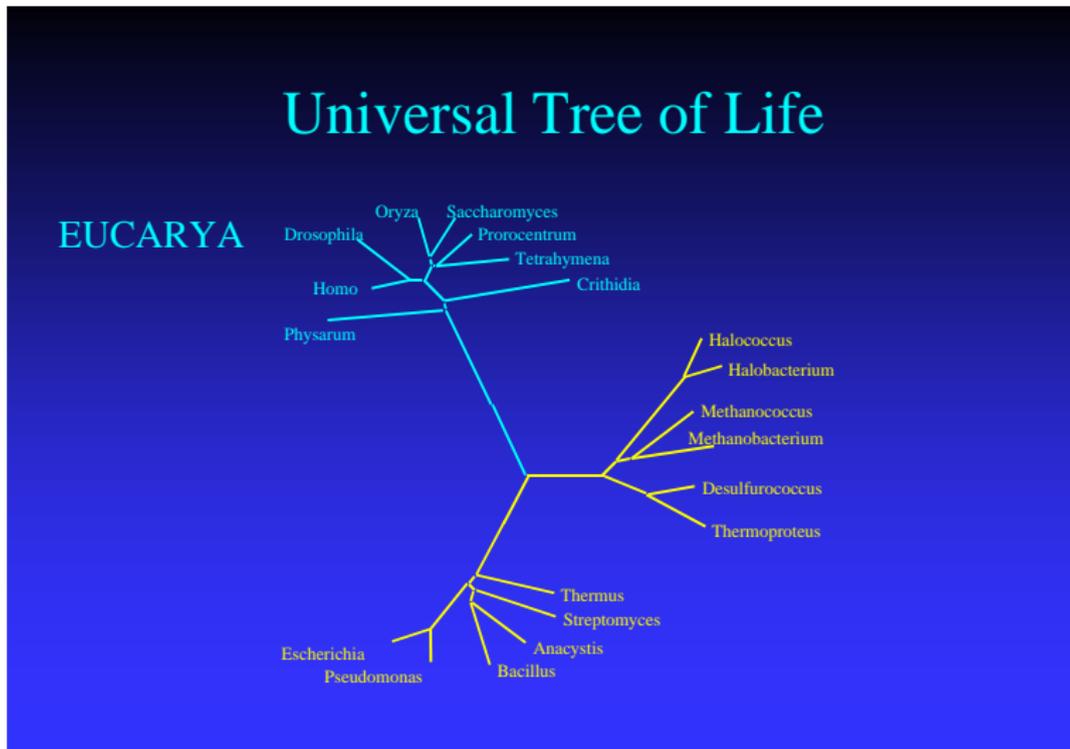
The three kingdoms : archaea

Universal Tree of Life
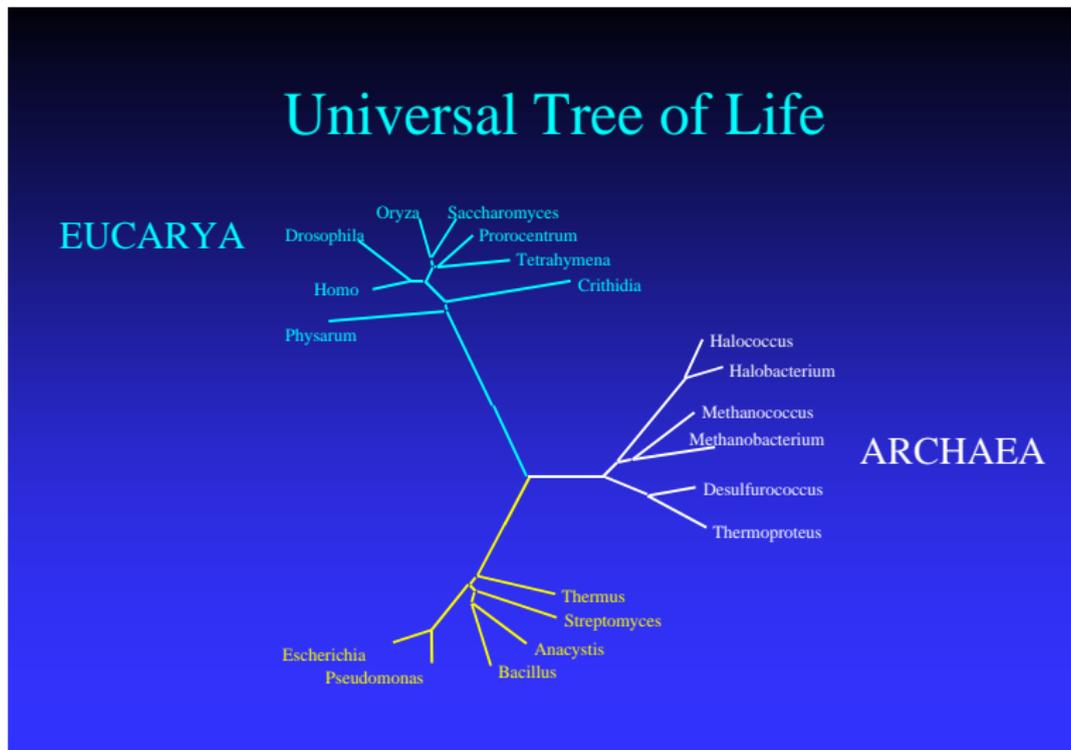
The three kingdoms : eubacteria

Universal Tree of Life
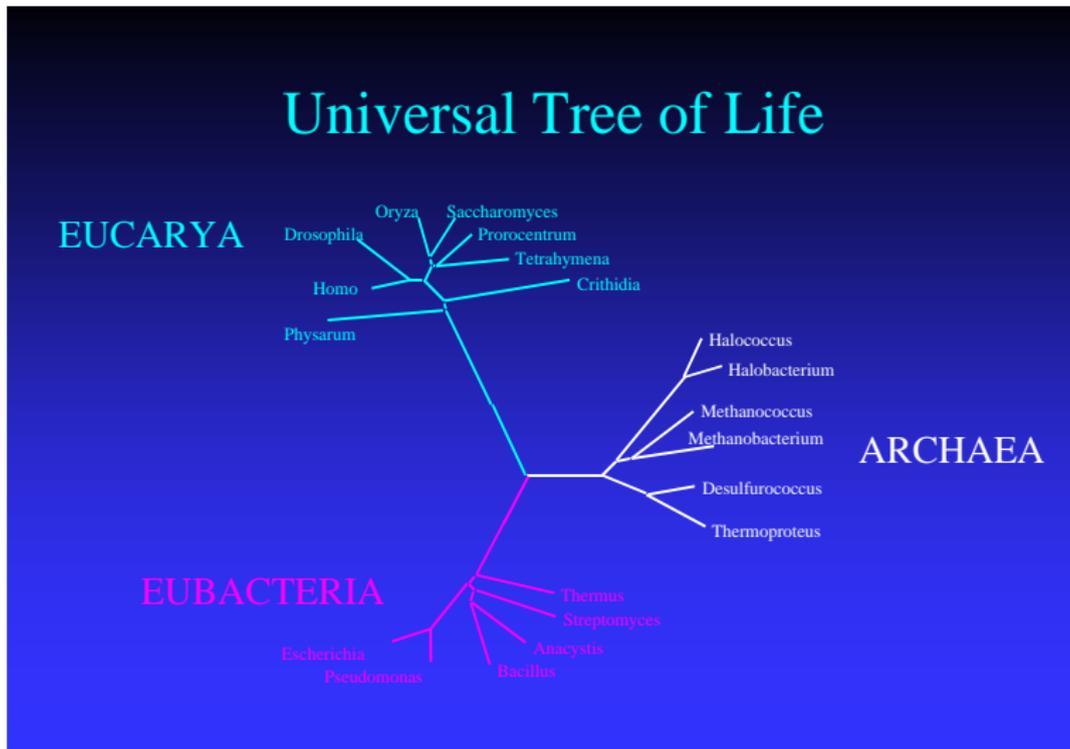
The three kingdoms : root 1
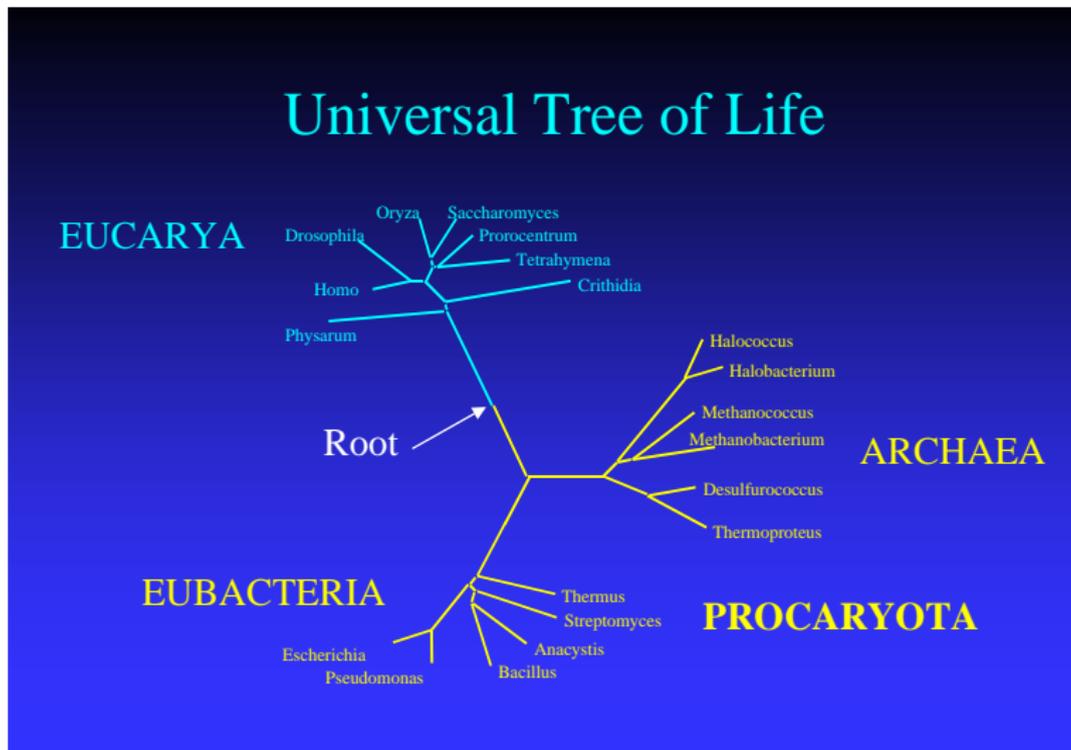
The three kingdoms : root 2

Universal Tree of Life
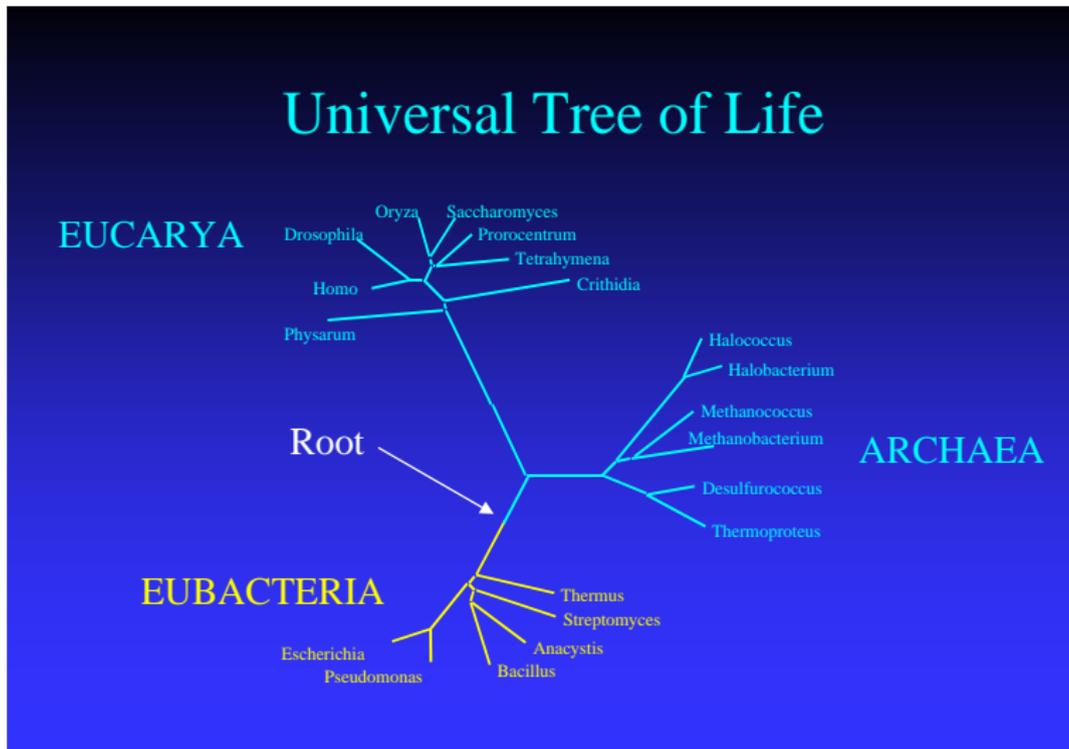
The three kingdoms : root 3
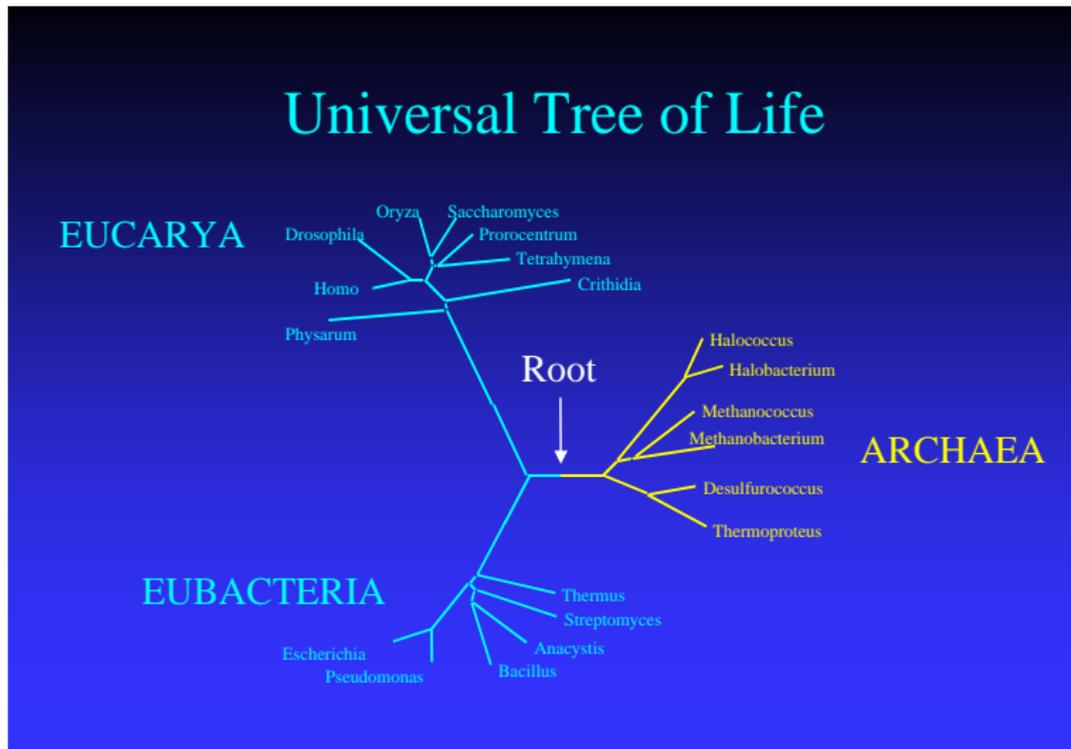
The three kingdoms : no root
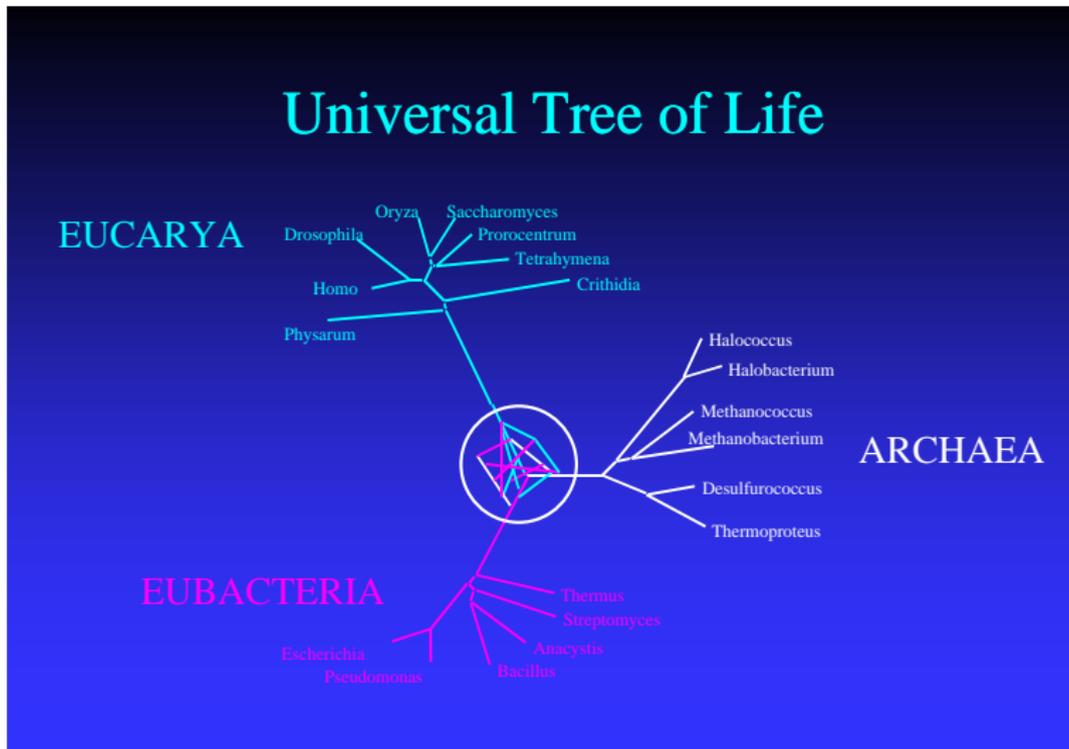
The three kingdoms : "bacteria"
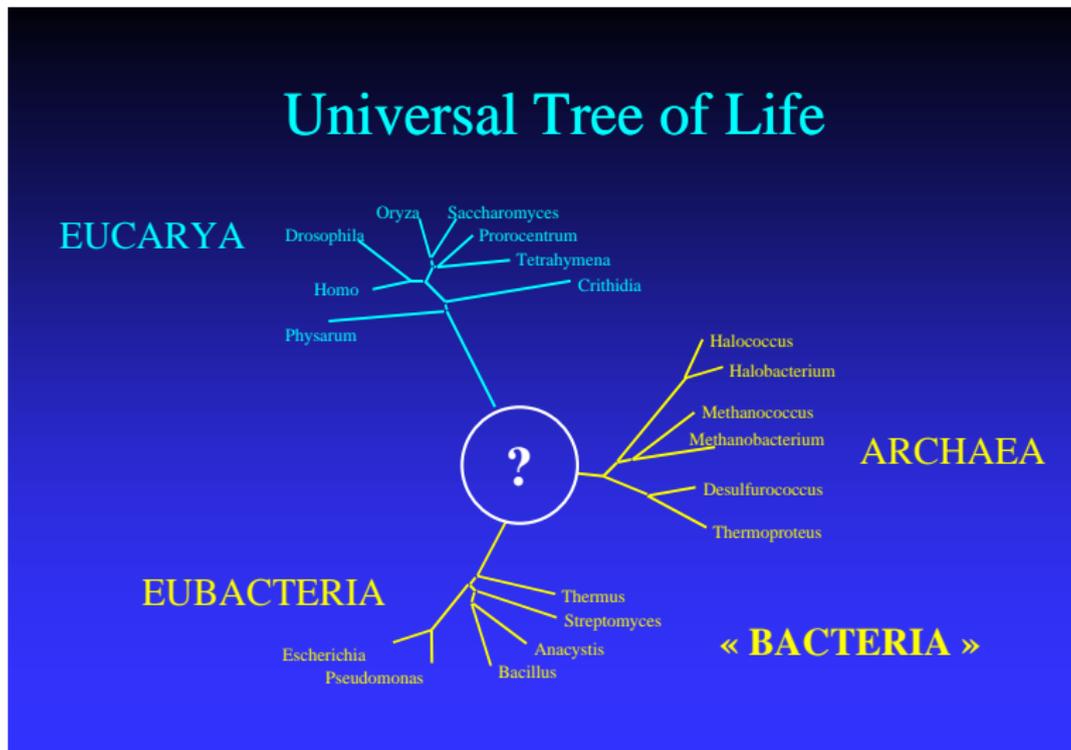
Organelles : chloroplasts & mitochondria
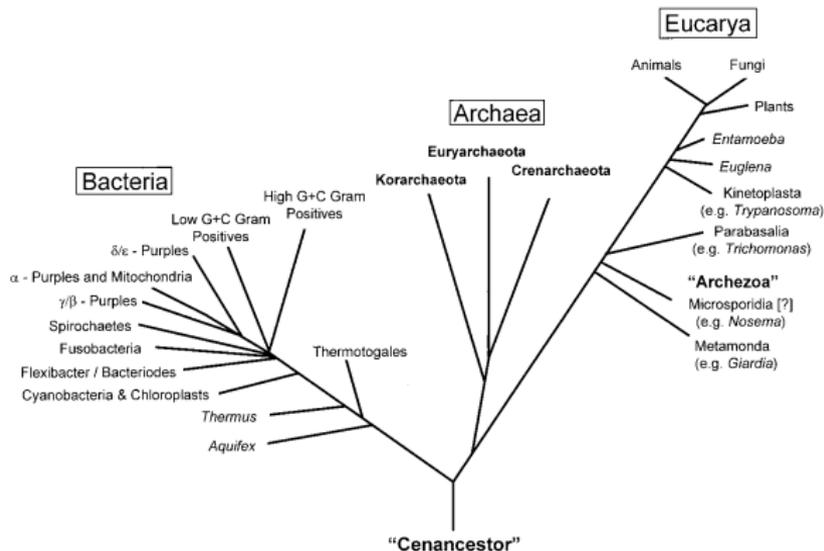
FIG. 1. Schematic drawing of a universal rRNA tree showing the relative positions of evolutionary pivotal groups in the domains *Bacteria*, *Archaea*, and *Eucarya*. The location of the root (the cenancestor) corresponds to that proposed by reciprocally rooted gene phylogenies (43, 133, 164). The question mark beside the Archezoa group Microsporidia denotes recent suggestions that it might branch higher in the eukaryotic portion of the tree. (Branch lengths have no meaning in this tree.)

Bacterial genome structures
Introduction
General bacterial features

Half of biomass on earth

Bacterial genome structures
Introduction
General bacterial features

Very few species
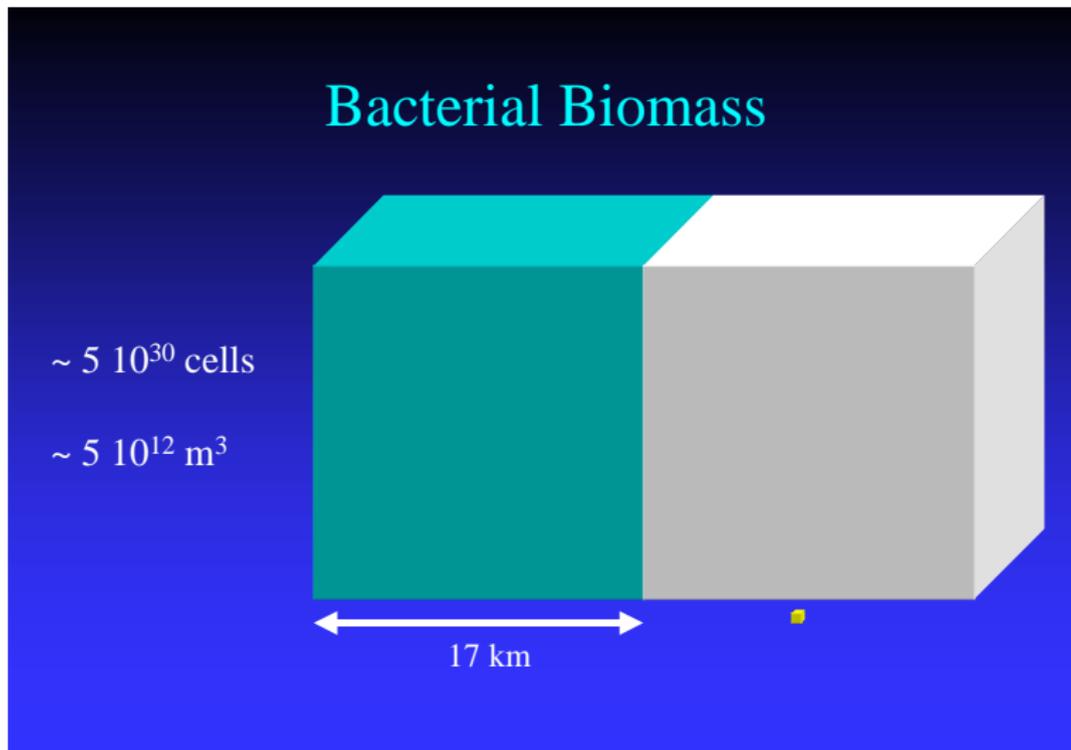
Bacterial genome structures
Introduction
General bacterial features
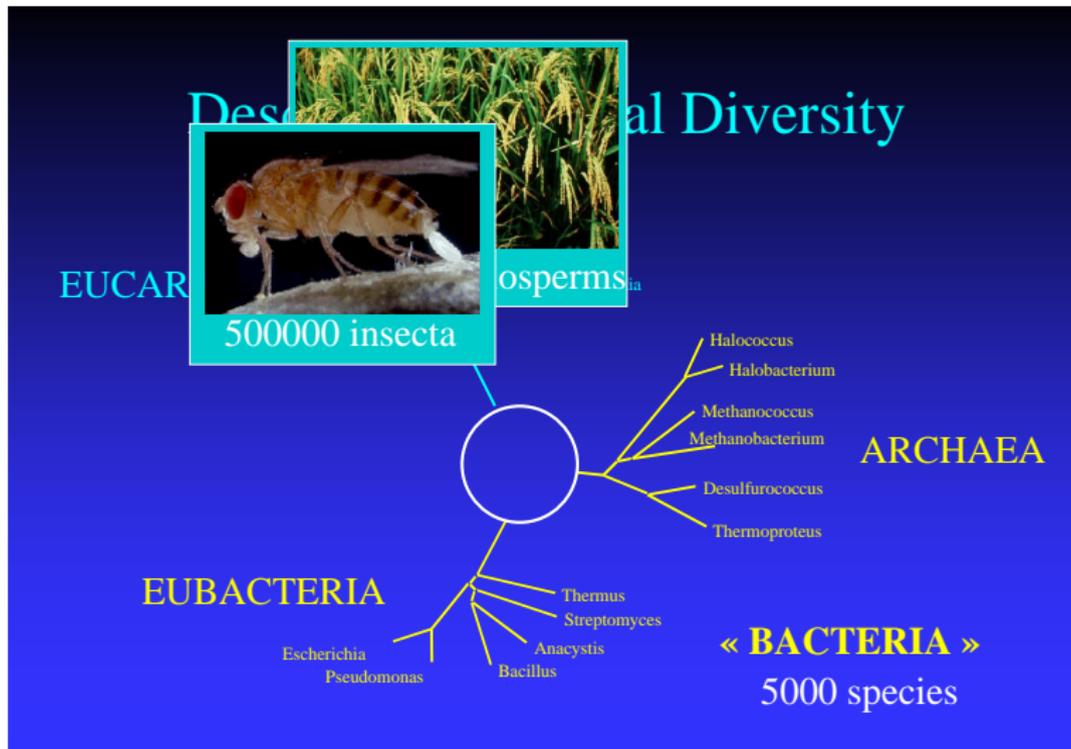
First species classification

Bacterial genome structures
Introduction
General bacterial features

Small bacteria

Bacterial genome structures
Introduction
General bacterial features

0.1 mm = 100 μm

Thickness ≈ 100 μm.

1 M€ ≈ 1 m
1 G€ ≈ 1 km

Bacterial genome structures
Introduction
General bacterial features
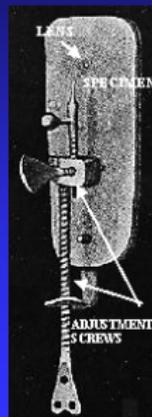
Bacterial cell size is in μm

100 μm

Naked eye
resolution

Demo

Bacterial genome structures
Introduction
General bacterial features

A giant : *Epulopiscium fishelsoni* bar is 50 μm

Bresler, V. *et al* (1998) *J. Bact.*, **180 :**5601-5611.

Bacterial genome structures
Introduction
General bacterial features

Mycoplasma genitalium bar is 0.25 μm
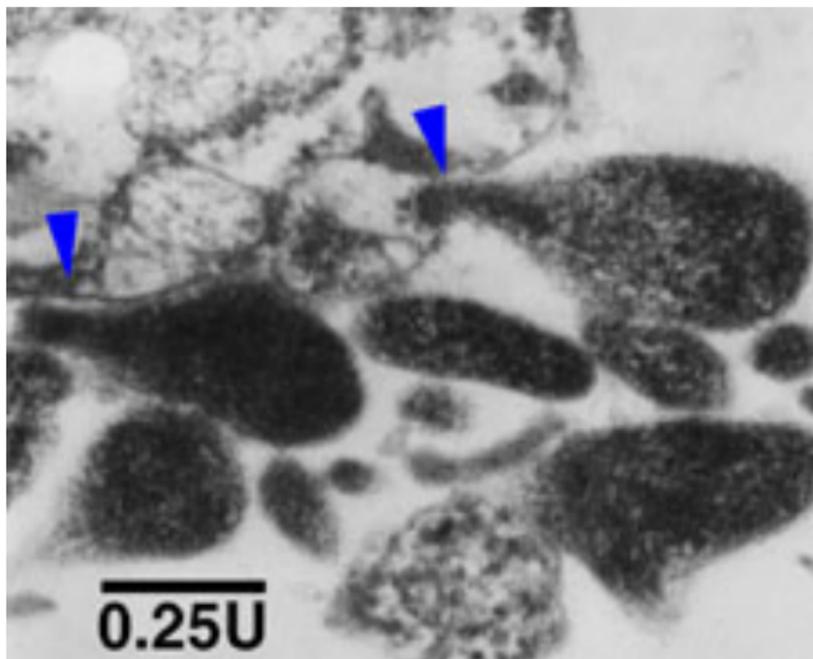
$$1 \text{ colony} \approx 10^6 \text{ cells}$$

Little Bacteria

1 colony 10^6 cells
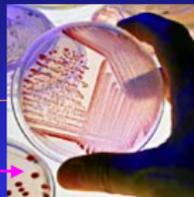
Few morphological traits

Ugly Little Bacteria

10^7 years

10^9 years

E. coli

B. subtilis

Bacterial genome structures
Introduction
General bacterial features
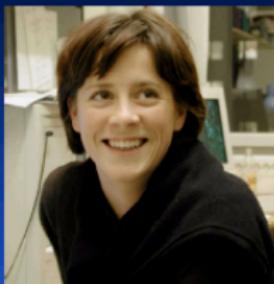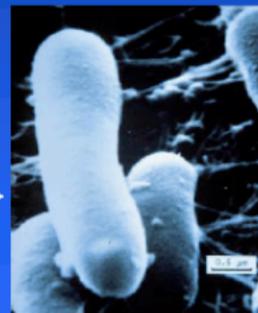
Bacterial classification

Classification of Bacteria

• Was based mainly on physiology and growth conditions

E. coli

 Most bacteria defy cultivation...

Bacterial classification : "*Candidatus* Pelagibacter ubique"

Candidatus Pelagibacter ubique HTCC1062, complete genome

Bacterial genome structures
Introduction
General bacterial features
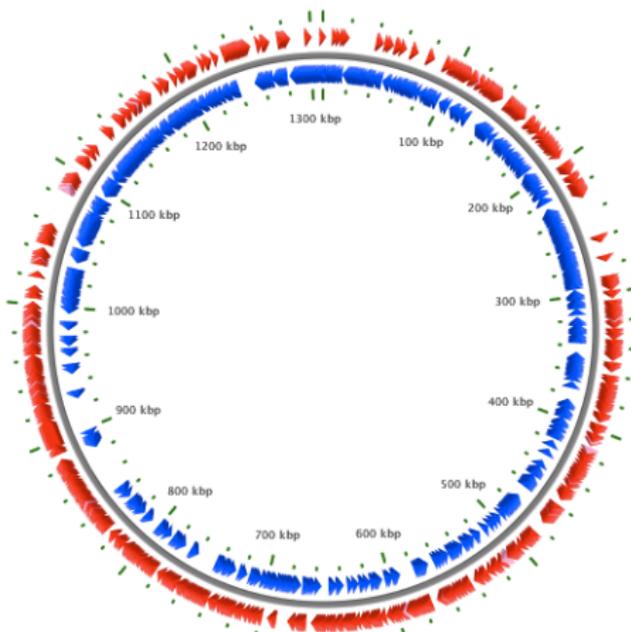
Bacterial classification : *Candidatus*

Candidatus examples :

- "*Candidatus* Arsenophonus triatominarum"
- "*Candidatus* Arthromitus"
- "*Candidatus* Blochmannia"
- "*Candidatus* Blochmannia floridanus"
- "*Candidatus* Blochmannia herculeanus"
- "*Candidatus* Burkholderia kirkii"
- "*Candidatus* Glomeribacter gigasporarum"
- "*Candidatus* Xiphinematobacter brevicolli"

Genome size

1. **Introduction**

2. **Genome size**

3. Topology & #

4. G+C content

5. Replichores

6. Gene orientation biases

7. Chrirochores

8. X-rated structure

bp : base pair

Common multiples are :

- 1 kb $= 10^3$ bp
- 1 Mb $= 10^6$ bp
- 1 Gb $= 10^9$ bp

Bacterial genomes are typically expressed in Mb

Length conversion

Dickerson *et al* (1982) *Science*, **216** :475-485.

1 bp $\approx$ 0.33 nm

- 1 kb $\approx$ 0.33 μm
- 1 Mb $\approx$ 0.33 mm
- 1 Gb $\approx$ 0.33 m

Bacterial genomes are typically in the mm range.

Mass conversion (1 pg $= 10^{-12}$ g)

Doležel *et al* (2003) *Cytometry*, **51A** :127-128.
Number of base pairs = mass in pg $\times$ 0.978 10^9

- 1 kb $\approx 10^{-6}$ pg
- 1 Mb $\approx 10^{-3}$ pg
- 1 Gb ≈ 1 pg

Bacterial genomes are typically in the 10^{-3} pg range (femtogram).

Mass conversion constant and G+C content

Base	Nucleotide	Chemical formula
A	2'-deoxyadenosine 5'-monophosphate	$C_{10}H_{14}N_5O_6P$
T	2'-deoxythymidine 5'-monophosphate	$C_{10}H_{15}N_2O_8P$
G	2'-deoxyguanosine 5'-monophosphate	$C_{10}H_{14}N_5O_7P$
C	2'-deoxycytidine 5'-monophosphate	$C_9H_{14}N_3O_7P$

Tab.: Chemical formula of the four nucleotides in DNA.

Mass conversion constant and G+C content

Evolution of the conversion constant with GC content
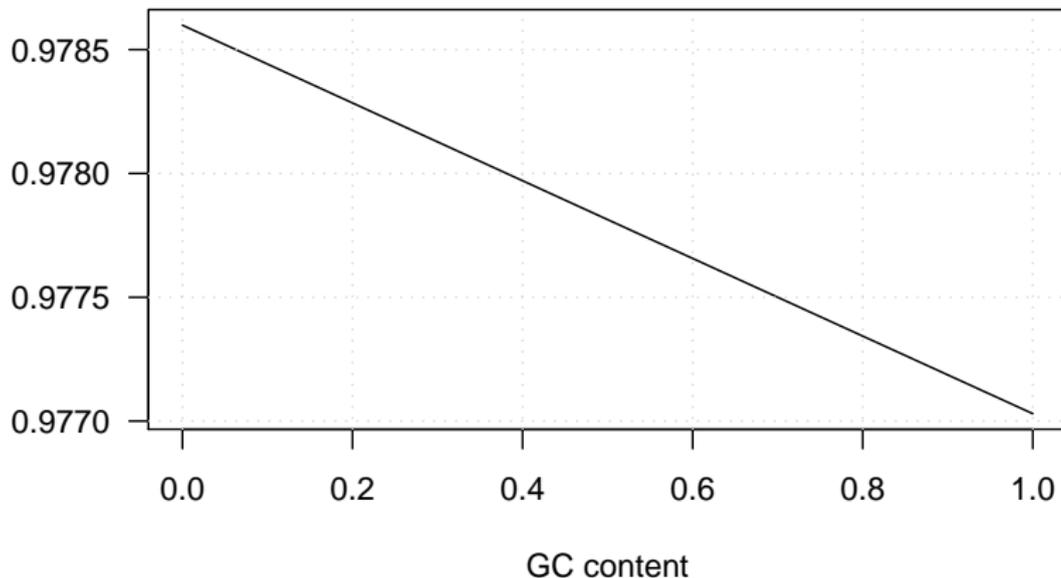
GC content

**Evolution of the conversion constant with GC content
and the fraction of methylated CpG**

Bacterial genome structures
Genome size
As compared to other
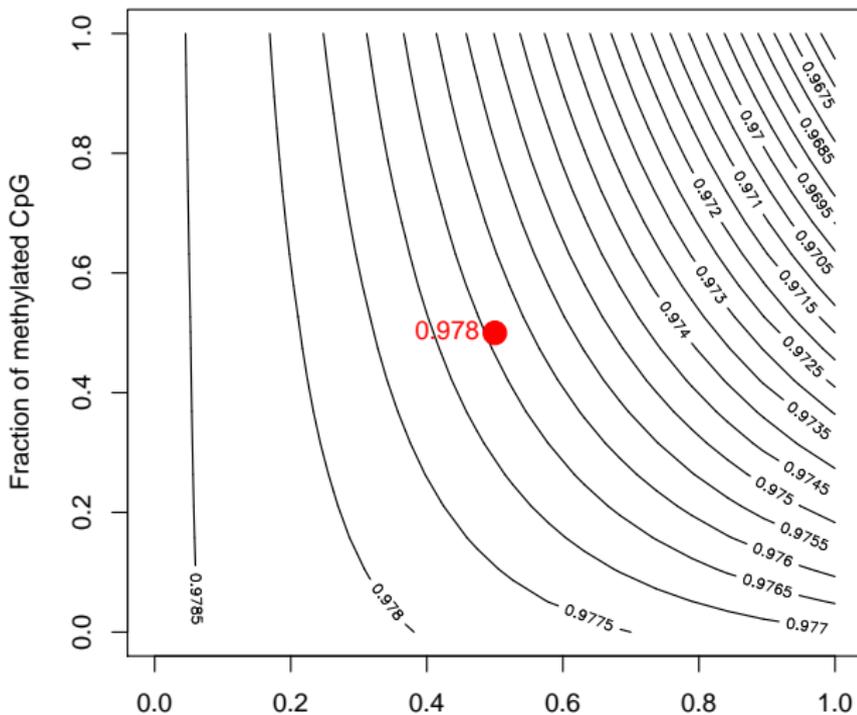
The big picture

Virus, organelles
Tiny genomes (kb)
High gene density

"Bacteria"
Small genomes (Mb)
High gene density

Eucarya
Large genomes (Gb)
Low gene density

C value paradox
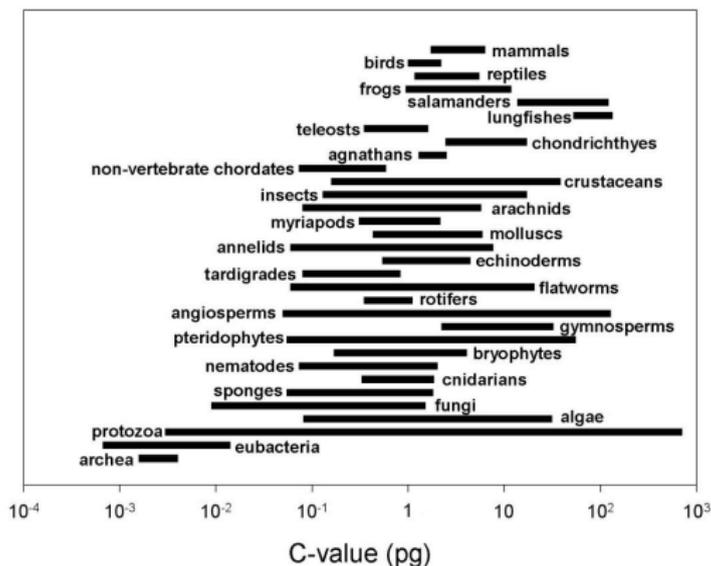
Gregory, T.R. (2004) *Paleobiology*, **30** :179-202.

C value paradox

Gregory, T.R. (2005) *Animal Genome Size Database*

Genome size [pg] (log 10 scale)
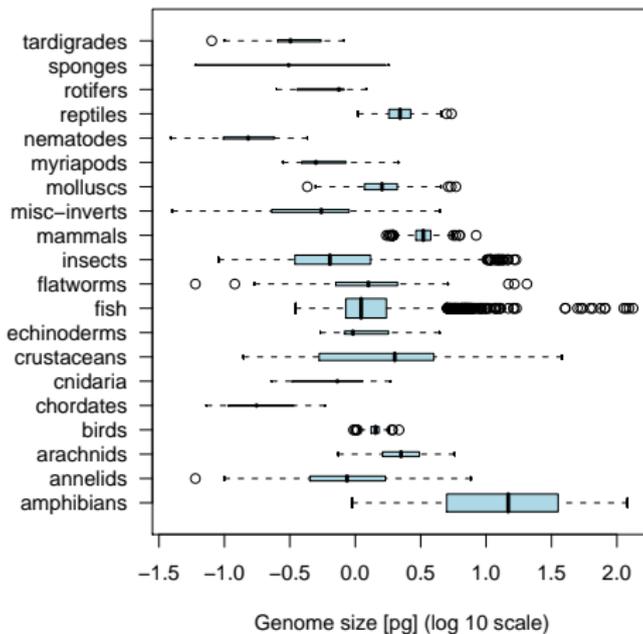
Giant virus : mimivirus 1.2 Mb
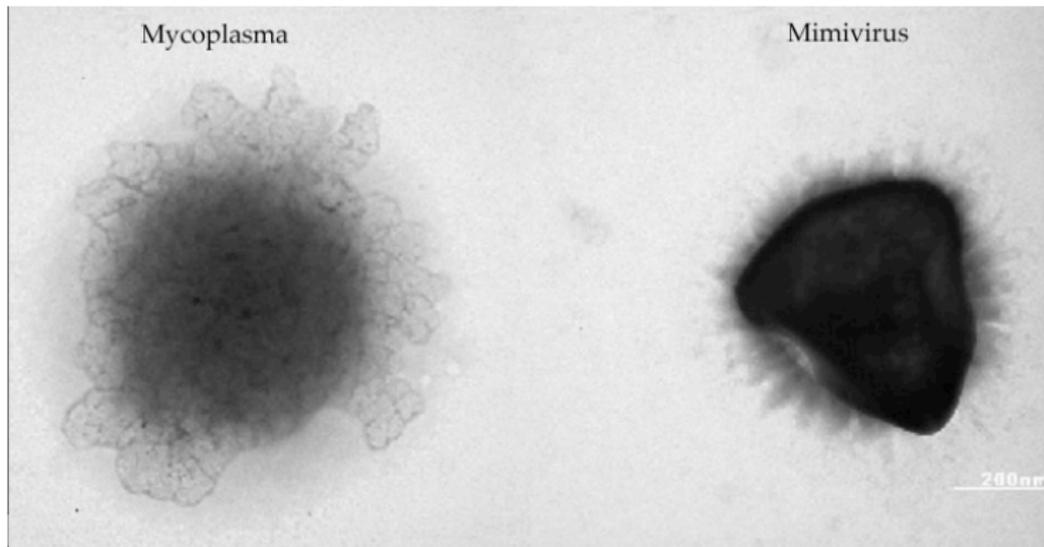
Electronic microscopy of a "bacteria" on the left (*Ureaplasma urealyticum (parvum)*) with a genome size of 0.751

Mb and mimivirus on the rigth with a genome size of 1.181 Mb. Credit : the Mimivirus picture gallery from

`http ://giantvirus.org/`. Copyright : Prof. Didier Raoult, Rickettsia Laboratory, La Timone, Marseille, France.

Pseudogenes in *Rickettsia prowazekii*

Andersson, S.G. *et al* (1998) *Nature*, **396** :133-140.
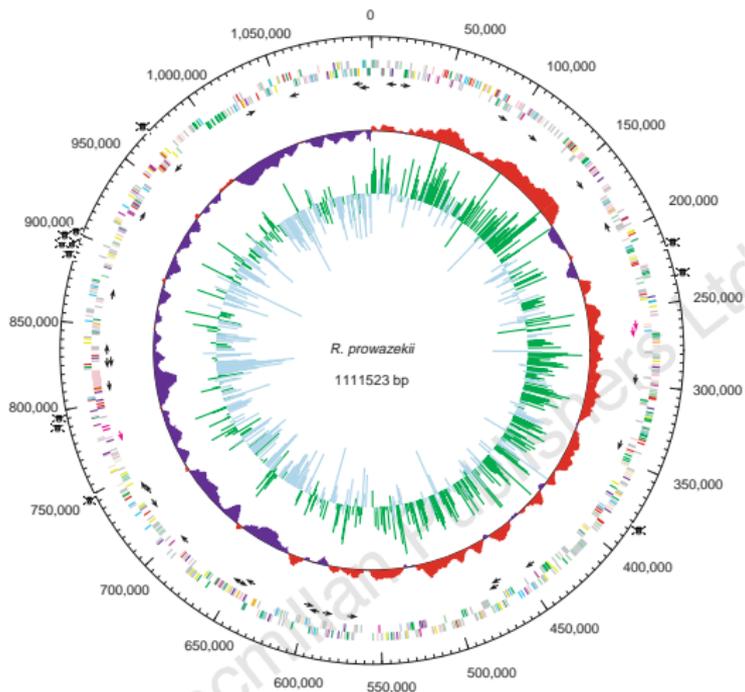
Pseudogenes in *Mycobacterium leprae*

Massive gene decay in the leprosy bacillus

S. T. Cole*, K. Eiglmeier*, J. Parkhill†, K. D. James†, N. R. Thomson†, P. R. Wheeler‡, N. Honoré*, T. Garnier*, C. Churcher†, D. Harris†, K. Mungall†, D. Basham†, D. Brown†, T. Chillingworth†, R. Connor†, R. M. Davies†, K. Devlin†, S. Duthoy*, T. Feltwell†, A. Fraser†, N. Hamlin†, S. Holroyd†, T. Hornsby†, K. Jagels†, C. Lacroix*, J. Maclean†, S. Moule†, L. Murphy†, K. Oliver†, M. A. Quail†, M.-A. Rajandream†, K. M. Rutherford†, S. Rutter†, K. Seeger†, S. Simon*, M. Simmonds†, J. Skelton†, R. Squares†, S. Squares†, K. Stevens†, K. Taylor†, S. Whitehead†, J. R. Woodward† & B. G. Barrell†

*Unité de Génétique Moléculaire Bactérienne, Institut Pasteur, 28 rue du Docteur Roux, 75724 Paris Cedex 15, France
† Sanger Centre, Wellcome Trust Genome Campus, Hinxton, CB10 1SA, UK
‡ Veterinary Laboratories Agency, Weybridge, Woodham Lane, New Haw, Addlestone, Surrey KT15 3NB, UK

Leprosy, a chronic human neurological disease, results from infection with the obligate intracellular pathogen *Mycobacterium leprae*, a close relative of the tubercle bacillus. *Mycobacterium leprae* has the longest doubling time of all known bacteria and has thwarted every effort at culture in the laboratory. Comparing the 3.27-megabase (Mb) genome sequence of an armadillo-derived Indian isolate of the leprosy bacillus with that of *Mycobacterium tuberculosis* (4.41 Mb) provides clear explanations for these properties and reveals an extreme case of reductive evolution. Less than half of the genome contains functional genes but pseudogenes, with intact counterparts in *M. tuberculosis*, abound. Genome downsizing and the current mosaic arrangement appear to have resulted from extensive recombination events between dispersed repetitive sequences. Gene deletion and decay have eliminated many important metabolic activities including siderophore production, part of the oxidative and most of the microaerophilic and anaerobic respiratory chains, and numerous catabolic systems and their regulatory circuits.

# Tiny eucaryal genome : *Guillardia theta* is only 551 kb
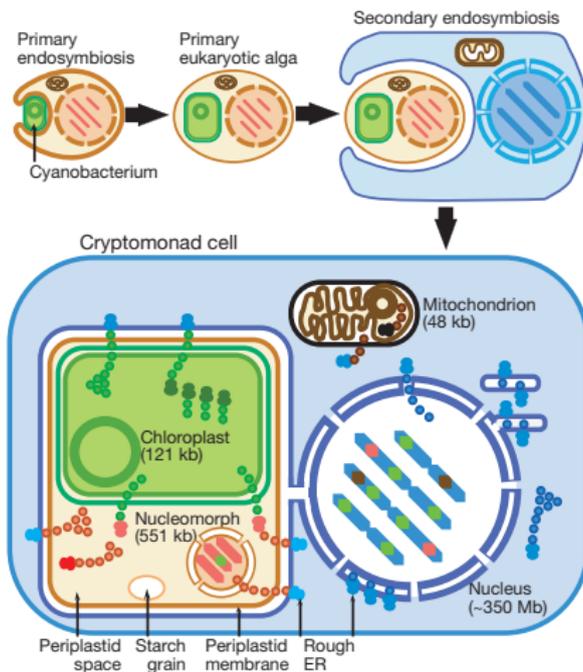
Tiny eucaryal genome : *Encephalitozoon cuniculi* is only 2.9 Mb

Towards the minimal eukaryotic parasitic genome
Christian P Vivarès* and Guy Méténier

Microsporidia are well-known to infect immunocompromised patients and are also responsible for clinical syndromes in immunocompetent individuals. In recent years, evidence has been obtained in support of a very close relationship between Microsporidia and Fungi. In some species, the compaction of the genome and genes is remarkable. Thus, a systematic sequencing project has been initiated for the 2.9 Mbp genome of *Encephalitozoon cuniculi*, which will be useful for future comparative genomic studies.

Figure 1
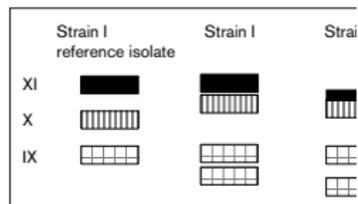
Katinka, M.D. *et al* (2001) *Nature*, **414** :450-453.

Overlap of free living forms

- Eucarya *Saccharomyces cerevisiae* is 12 Mb
- Bacteria *Sorangium cellulosum* is 13 Mb

Bacterial genome structures
Genome size
Between species variability

What is the distribution of bacterial genome size ?

Study this yourself :
http://pbil.univ-lyon1.fr/R/fichestd/tdr222.pdf

Genome size for 279 bacteria (GOLD 2002)

**Distribution de la tailles de
279 genomes bacteriens**

Densite

Taille du genome en Kb
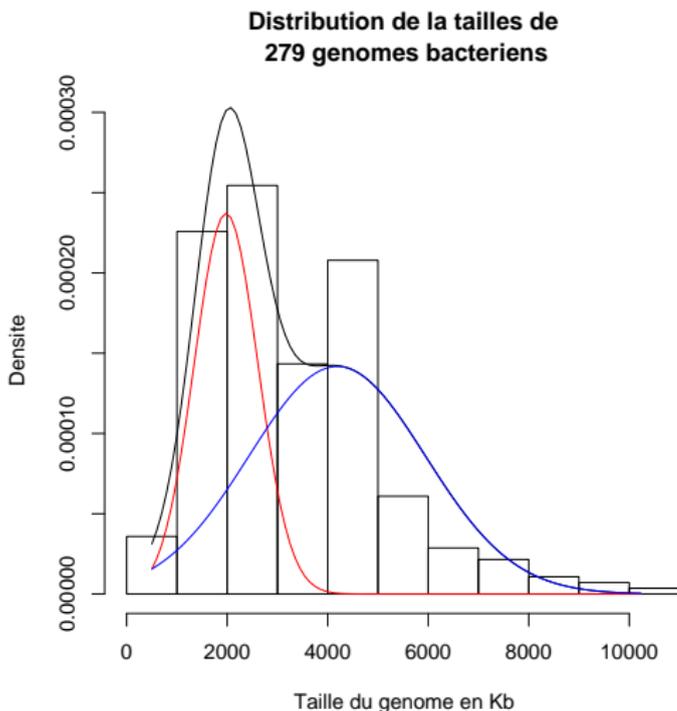
Genome size for 893 bacteria (GOLD 2006)

Genome size distribution for 893 bacterial genomes

Source of data: GOLD (Genomes OnLine Database) Tue Feb 7 11:00:24 2006
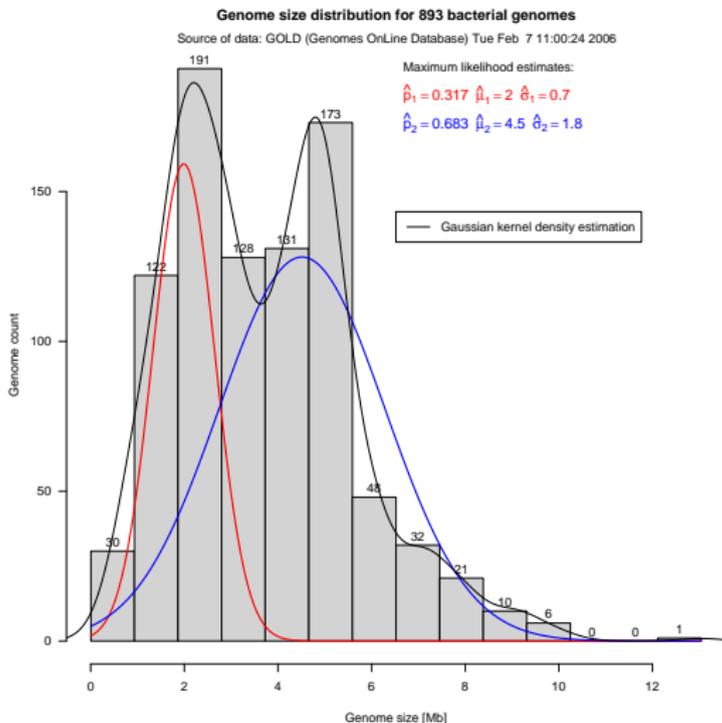
Genome size for 681 bacteria (PFGE data)

Bacterial genome structures
Genome size
Between species variability
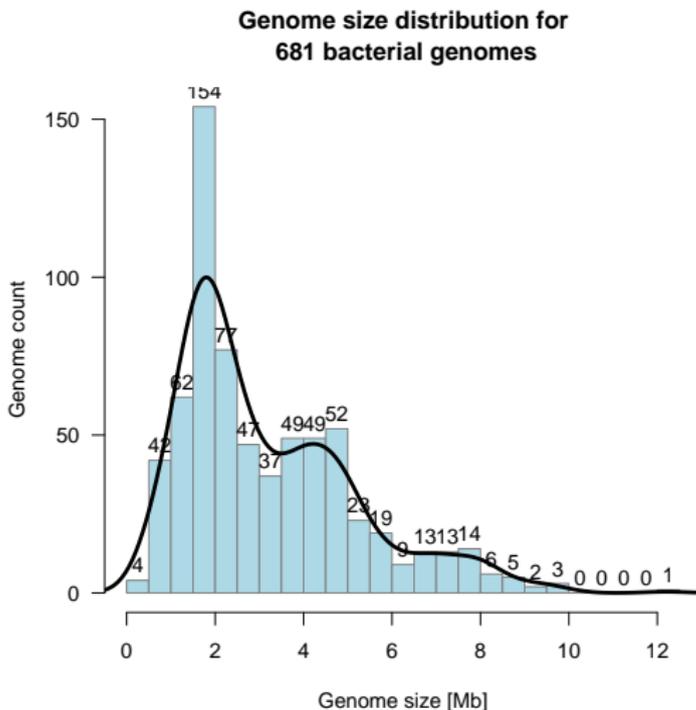
Genome size summary

- Range : from 0.45 Mb (*Buchnera*) to 13.0 Mb (*Sorangium cellulosum*).
- Three modes at 2 Mb, 4.5 Mb, and 8 Mb, respectively.

Generalists versus specialists

Giovannoni, S.J. *et al* (2005) *Science*, **309** :1242-1245.
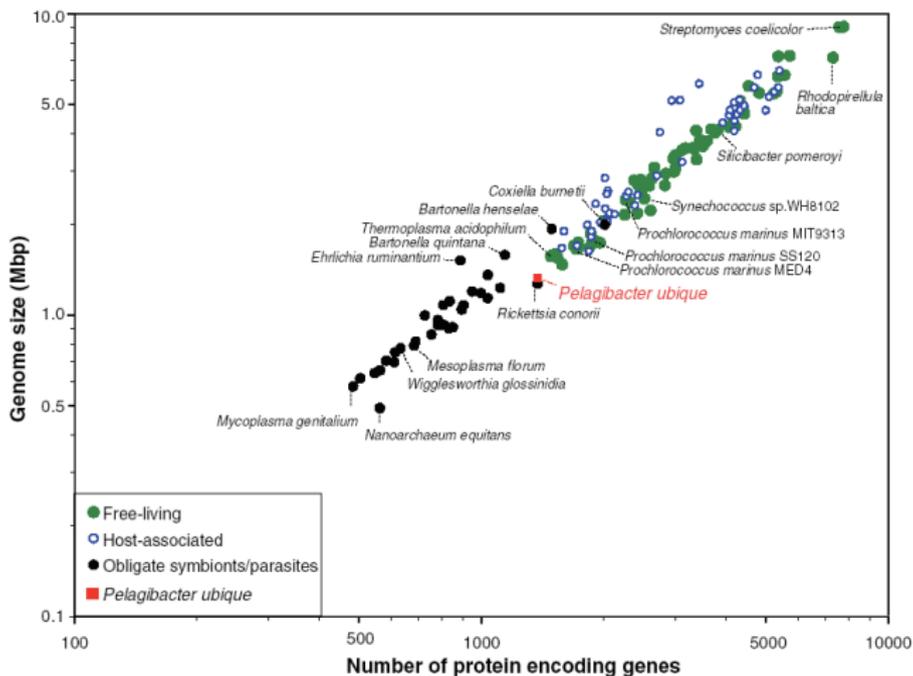
Genome size & repeat density

Genome deterioration: loss of repeated sequences and accumulation of junk DNA

A. Carolin Frank, Haleh Amiri & Siv G.E. Andersson[*]

Department of Molecular Evolution, University of Uppsala, Uppsala, S-751 36 Sweden; [*]Author for correspondence (Phone: +46-18-4714379; Fax: +46-18-471 64 04; E-mail: Siv.Andersson@ebc.uu.se)

Bacterial genome structures
Genome size
Between species variability

Genome size & repeat density
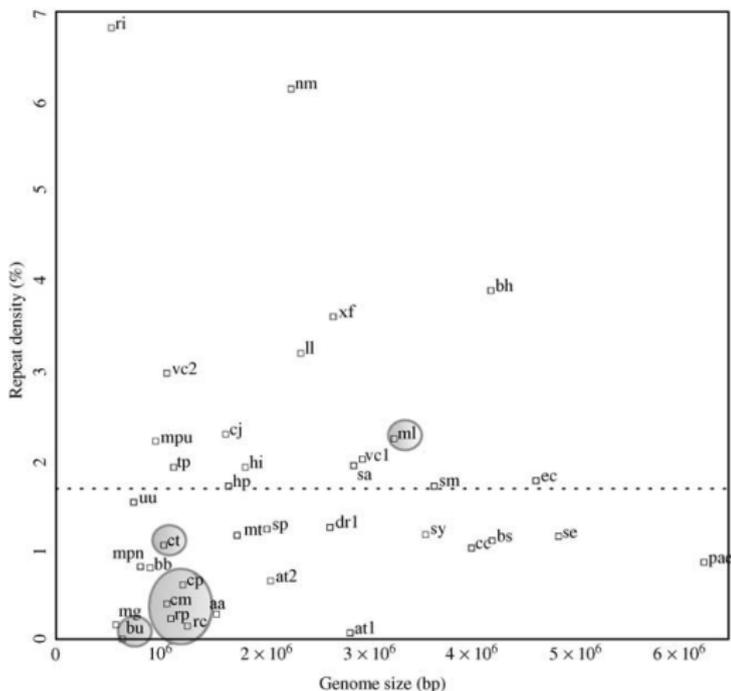
Genome size & repeat density

8

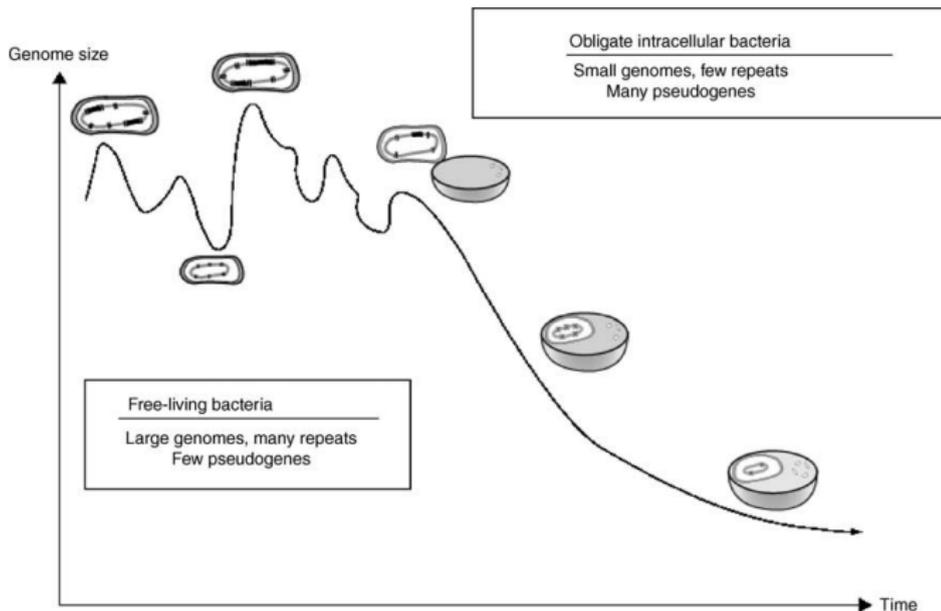*Figure 3.* Schematic illustration of genome size variations as a function of time during transitions to intracellular growth habitats. Filled boxes represent mobile genetic elements. Genomes of obligate intracellular bacteria are smaller and have a lower content of repeated sequences (//) and a higher content of pseudogenes (x) than genomes of free-living bacteria and facultative intracellular parasites.

Bacterial genome structures
  Genome size
    Within species variability

Genome size polymorphism in *E. coli*

Distribution of Chromosome Length Variation in Natural Isolates of
Escherichia coli

Ulfar Bergthorsson and Howard Ochman

Department of Biology, University of Rochester

Large-scale variation in chromosome size was analyzed in 35 natural isolates of *Escherichia coli* by physical mapping with a restriction enzyme whose sites are restricted to rDNA operons. Although the genetic maps and chromosome lengths of the laboratory strains *E. coli* K12 and *Salmonella enterica* sv. Typhimurium LT2 are highly congruent, chromosome lengths among natural strains of *E. coli* can differ by as much as 1 Mb, ranging from 4.5 to 5.5 Mb in length. This variation has been generated by multiple changes dispersed throughout the genome, and these alterations are correlated; i.e., additions to one portion of the chromosome are often accompanied by additions to other chromosomal regions. This pattern of variation is most probably the result of selection acting to maintain equal distances between the replication origin and terminus on each side of the circular chromosome. There is a large phylogenetic component to the observed size variation: natural isolates from certain subgroups of *E. coli* have consistently larger chromosomes, suggesting that much of the additional DNA in larger chromosomes is shared through common ancestry. There is no significant correlation between genome sizes and growth rates, which counters the view that the streamlining of bacterial genomes is a response to selection for faster growth rates in natural populations.

The ECOR collection

TABLE 1. Standard reference strains and electromorph mobility profiles

No.	Strain — Previous designation[a]	Source — Host (sex)	Source — Location	References	Group[b]	MDH	6PG	ADK	PE2	GOT	IDH	PGI	ACO	MPI	G6P	ADH
1	RM74A	Human (F)	Iowa	8, 9, 10, 12, 13, 15, 16	I	2	6	4	5	3	2	4	7	3	2	1
2	STM1	Human (M)	New York	12, 15	I	2	6	4	5	3	2	4	7	3	2	1
3	WIR1(a)	Dog	Massachusetts	12, 15	I	2	6	4	5	3	2	4	7	3	2	1
4	RM39A	Human (F)	Iowa	8–10	I	2	15	4	7	3	2	4	6	3	2	1
5	RM60A	Human (F)	Iowa	8, 9, 12, 13, 15, 16	I	2	4	4	5	3	2	4	7	3	2	1
6	RM66C	Human (M)	Iowa	5, 6, 8, 9, 11–13, 15, 16	I	2	13	4	5	3	2	4	6	3	2	1
7	RM73C	Orangutan	Washington (zoo)	5, 8, 9, 12, 13, 15	I	2	5	4	7	3	2	4	7	3	1	1
8	RM77C (b)	Human (F)	Iowa	4, 7–9, 12, 13, 15, 16	I	2	9	4	5	3	2	4	7	3	2	1
9	FN98	Human (F)	Sweden	2, 12, 15, 16	I	2	9	4	5	3	2	4	7	3	2	1
10	AN1	Human (F)	New York	12, 15	I	2	9	4	5	3	2	4	7	3	2	1
11	C97	Human (F)	Sweden	2, 12, 15, 16	I	2	9	4	5	3	2	4	7	3	2	1
12	FN59	Human (F)	Sweden	2, 12, 15, 16	I	2	6	4	5	3	5	4	7	3	2	4
13	FN10	Human (F)	Sweden	2, 12, 15, 16	I	2	6	4	7	3	2	4	7	3	1	1

Ochman, H. and Selander, R.K. (1984) *J. Bacteriol.*, **157** :690-693.

Bacterial genome structures
Genome size
Within species variability
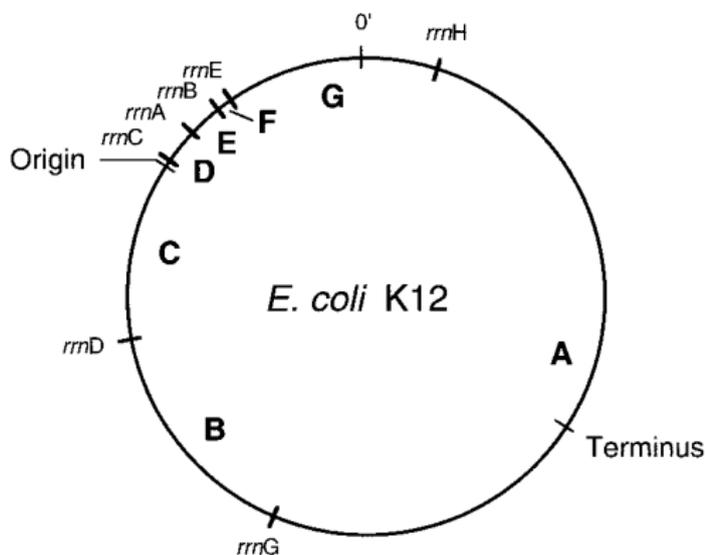
Digestion of the *E. coli* chromosome with I-*Ceu*I

FIG. 1.—Locations of I-*Ceu*I recognition sites on the *E. coli* K12 chromosome. I-*Ceu*I cleaves at the seven *rrn* genes, whose map positions are indicated. The resulting restriction fragments are designated **A** through **G**.

Bacterial genome structures
Genome size
Within species variability

Results in kb

	group	strain	Host..sex.	Location	A	B	C	D	E	F	G
1	A	ECOR4	Human (F)	Iowa	2585	707	527	90	166	38	608
2	A	ECOR5	Human (F)	Iowa	2940	743	515	90	128	38	699
3	A	ECOR11	Human (F)	Sweden	2750	824	556	90	128	38	735
4	A	ECOR13	Human (F)	Sweden	2485	680	515	90	128	38	639
5	A	ECOR14	Human (F)	Sweden	2645	735	608	90	128	38	707
6	A	ECOR15	Human (F)	Sweden	2690	735	575	90	138	38	639
7	A	ECOR18	Celebese ape	Washington	2510	699	515	90	122	38	608
8	A	ECOR19	Celebese ape	Washington	2480	699	527	90	122	38	639
9	A	ECOR20	Steer	Bali	2505	654	480	90	122	38	608
10	A	ECOR21	Steer	Bali	2505	654	480	90	122	38	608
11	A	ECOR23	Elephant	Washington	2675	807	532	90	138	38	680
12	B1	ECOR27	Giraffe	Washington	2600	707	515	90	143	38	616
13	B1	ECOR28	Human (F)	Iowa	2620	743	527	94	128	38	639
14	B1	ECOR29	Kangaroo rat	Nevada	2610	787	527	94	138	38	639
15	B1	ECOR34	Dog	Massachusetts	2500	790	515	94	138	38	680
16	B1	ECOR58	Lion	Washington	2700	743	515	94	136	38	639
17	B1	ECOR68	Giraffe	Washington	2745	843	532	94	138	38	807
18	B1	ECOR71	Human (F)	Sweden	2650	771	547	90	138	38	654
19	B1	ECOR72	Human (F)	Sweden	2635	771	532	94	138	38	680
20	B2	ECOR51	Human infant	Massachusetts	2750	810	550	112	138	38	810
...											
31	D	ECOR39	Human (F)	Sweden	2780	787	581	104	143	38	713
32	D	ECOR40	Human (F)	Sweden	2845	807	616	104	143	43	787
33	E	ECOR31	Leopard	Washington	2775	743	547	94	138	38	735
34	E	ECOR37	Marmoset	Washington	3100	787	581	94	175	38	743
35	E	ECOR42	Human (M)	Massachusetts	2735	743	616	94	143	38	699

Bacterial genome structures
    Genome size
        Within species variability

## What is the polymorphism of *E. coli* genome size?

Study this yourself:

```
pgs <- read.table("http://pbil.univ-lyon1.fr/R/donnees/polygensize.txt",
    header = TRUE, sep = "\t")
head(pgs)

  subgroup strain Host..sex. Location    A   B   C  D   E  F   G
1        A  ECOR4  Human (F)     Iowa 2585 707 527 90 166 38 608
2        A  ECOR5  Human (F)     Iowa 2940 743 515 90 128 38 699
3        A ECOR11  Human (F)   Sweden 2750 824 556 90 128 38 735
4        A ECOR13  Human (F)   Sweden 2485 680 515 90 128 38 639
5        A ECOR14  Human (F)   Sweden 2645 735 608 90 128 38 707
6        A ECOR15  Human (F)   Sweden 2690 735 575 90 138 38 639
```

- What is the distribution of genome size?
- Any relationship with the subgroup?
- What is the nice hidden structure in this dataset?

Bacterial genome structures
Genome size
Within species variability

Genome size is highly polymorphic in *E. coli*

Distribution of genome size for 35 *Escherichia coli* strains
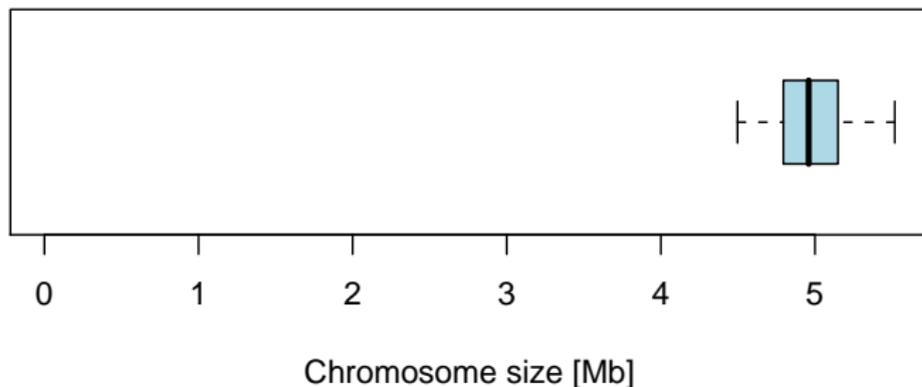
Chromosome size [Mb]

There is no meiotic constraints on chromosome length in bacteria.

Bacterial genome structures
   Genome size
     Within species variability

Genome size phylogenetic inertia

Genome size within 5 subgroups of *Escherichia coli* strains

Genome size [Mb]

Bacterial genome structures
    Genome size
       Within species variability
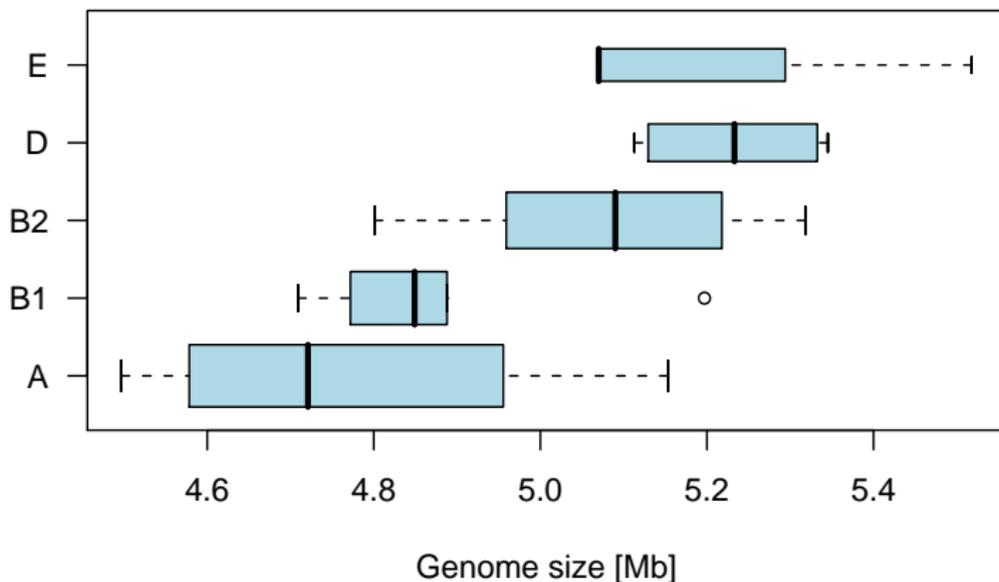
Genome size phylogenetic inertia

```
 tcs <- rowSums(pgs[, 5:11])/1000
 anova(lm(tcs ~ pgs$subgroup))

Analysis of Variance Table

Response: tcs
             Df  Sum Sq Mean Sq F value   Pr(>F)
pgs$subgroup  4 1.07556 0.26889   6.817 0.0004999 ***
Residuals    30 1.18333 0.03944
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Bacterial genome structures
Genome size
Within species variability

The nice hidden structure

Size of fragments **DEFG** in kb (right of origin) — vertical axis, 800 to 1150

Size of fragments **BC** in kb (left of origin) — horizontal axis, 1100 to 1500

The nice hidden structure

0157 :H7 *vs* EDL933

Bacterial genome structures
Genome size
Within species variability
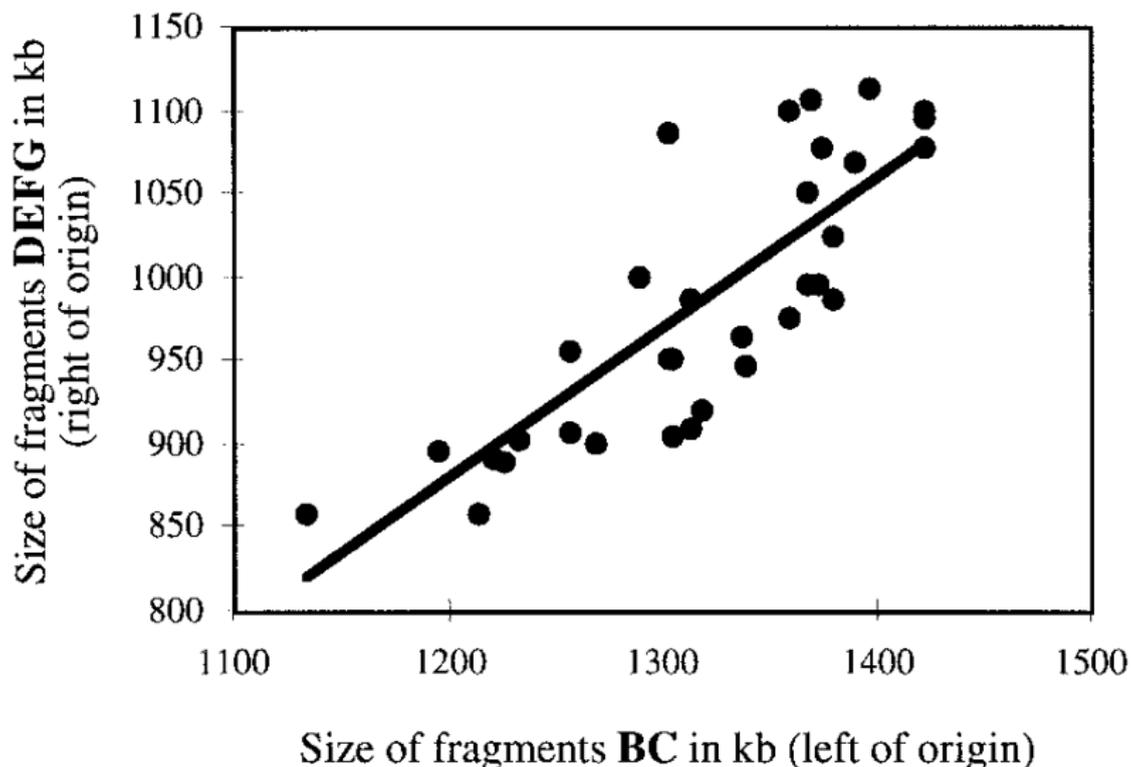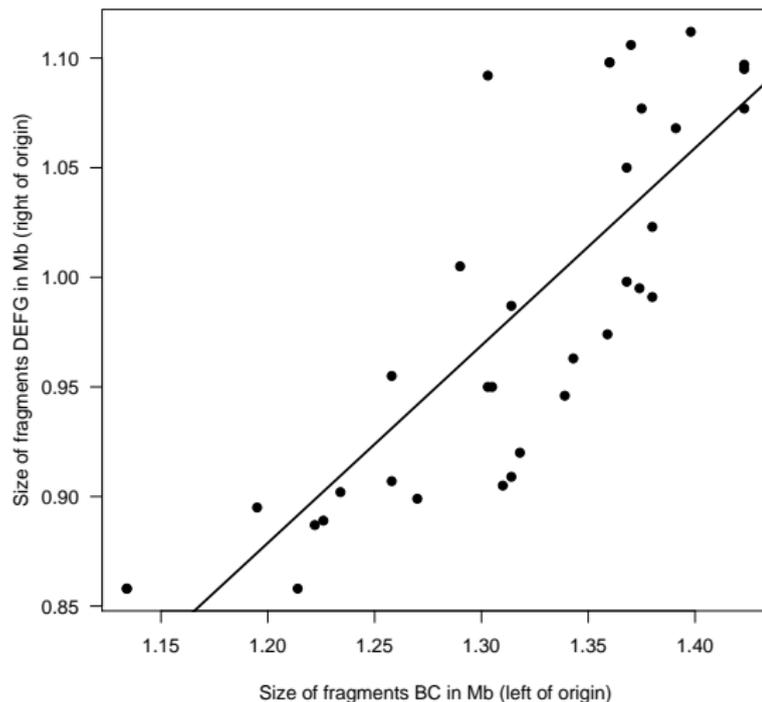
Genome size polymorphism in bacteria

**Genome size within species
with at least 10 strains**

Source: GOLD Sun Feb 26 16:28:56 2006
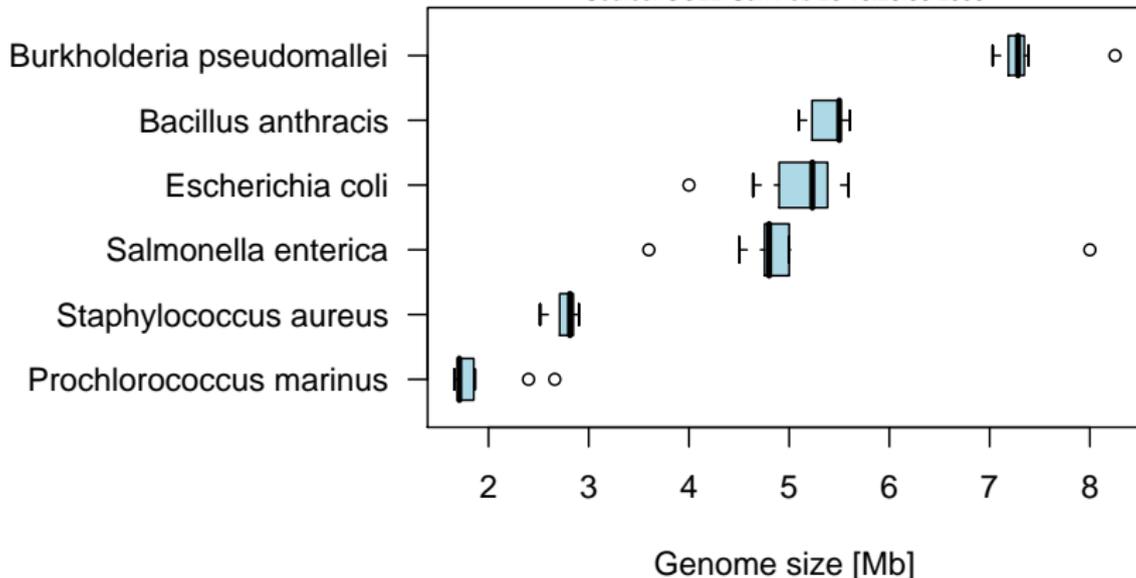
Genome size [Mb]

Topology

1. Introduction

2. Genome size

3. Topology & #

4. G+C content

5. Replichores

6. Gene orientation biases

7. Chrirochores

8. X-rated structure

The big picture

Eukaryota
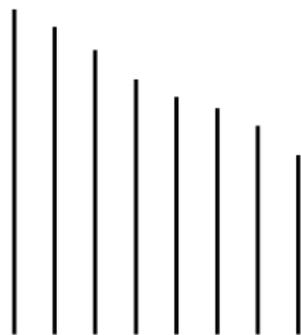
Many linear
chromosomes

Bacteria
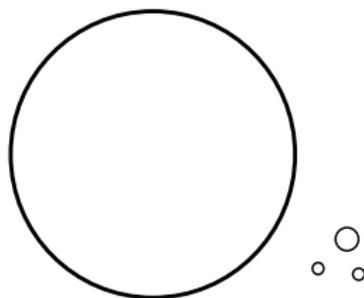
A single circular
chromosome + plasmids

V. cholerae : 2 circular chromosomes
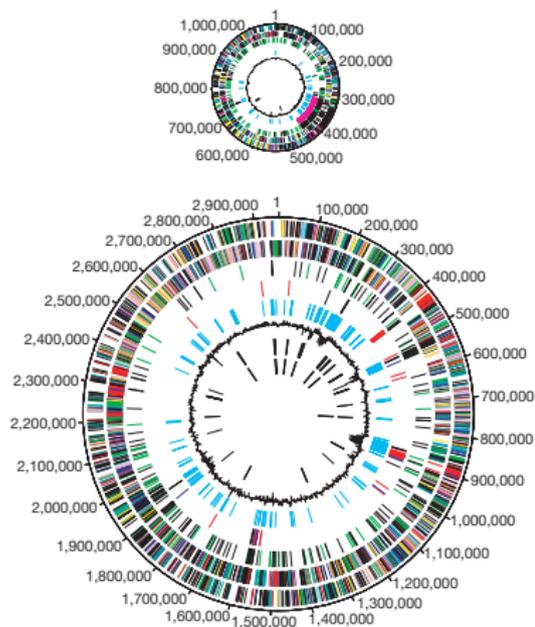
D. radiodurans : 2 circular chromosomes + 2 circular plasmids
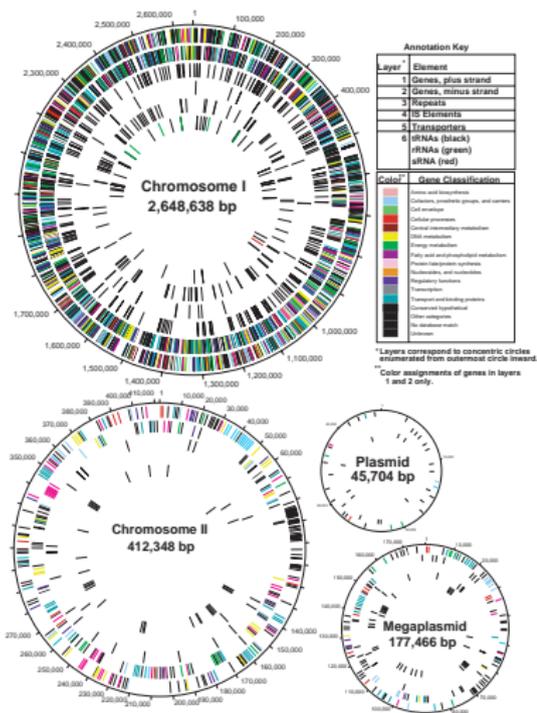
S. coelicolor : 1 linear chromosome
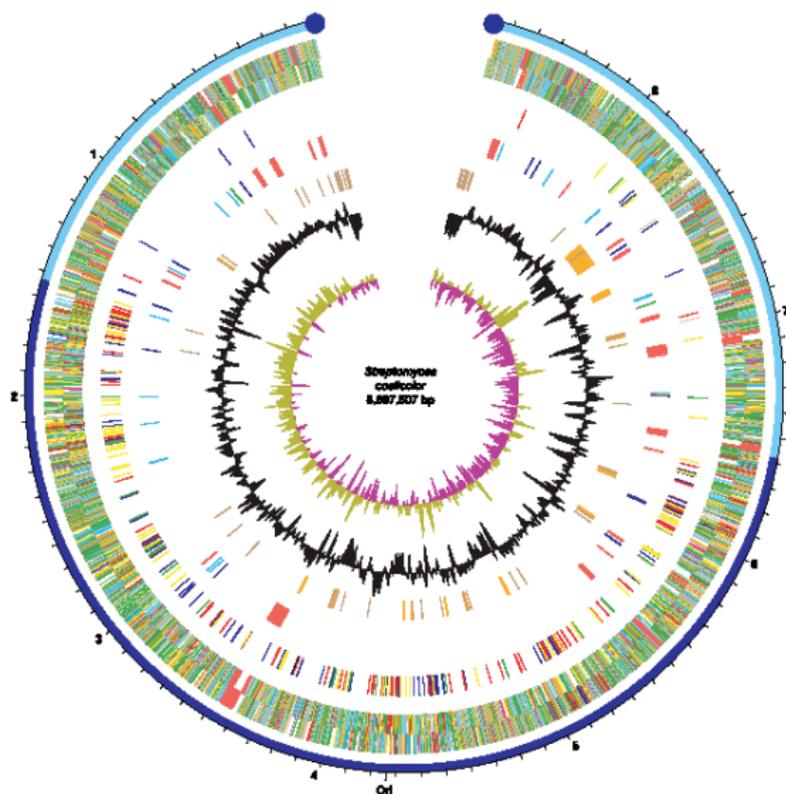
C. acetobutyliticum : 1 circular chromosome + 1 circular megaplasmid
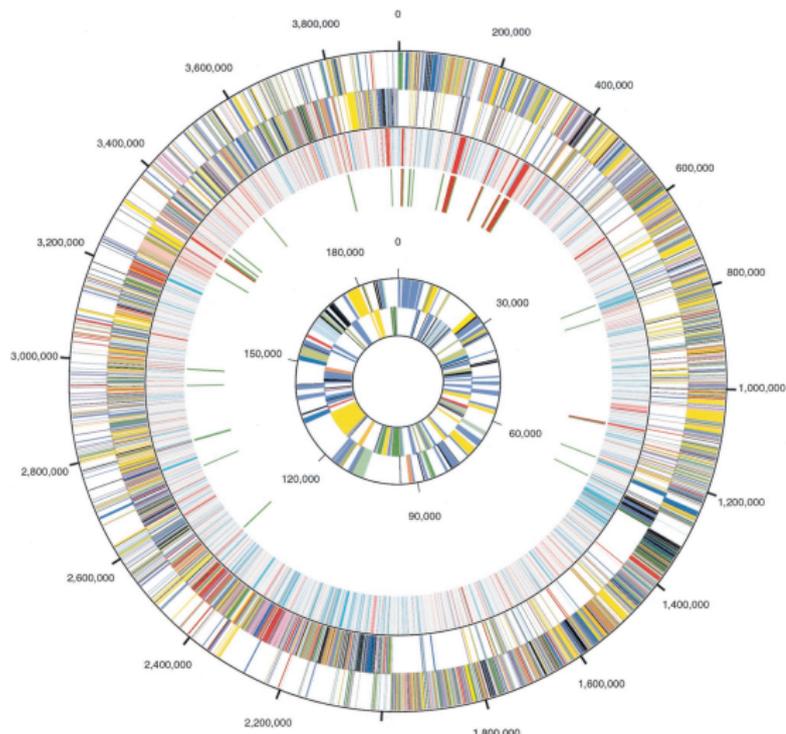
S. enterica serovar Typhimurium LT2 : 1 circular chromosome + 1 circular plasmid
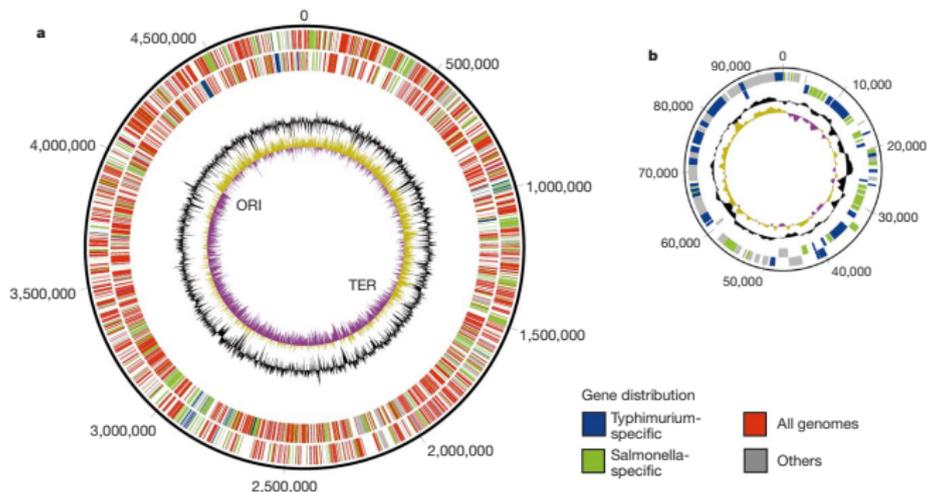
A. tumefaciens : 1 circular chromosome + 1 linear chromosome + 2 circular plasmid
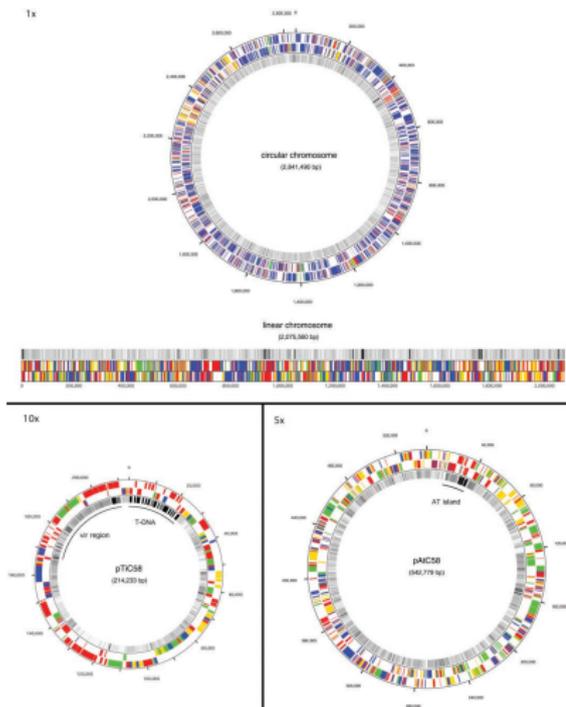
B. burgdorferi : many things !

Replicon	Geometry	Size (bp)
Chromosome[h]	Linear	910 725
cp9	Circular	9386
cp26	Circular	26 498
cp32-1	Circular	30 750
cp32-3	Circular	30 223
cp32-4	Circular	30 299
cp32-6	Circular	29 838
cp32-7	Circular	30 800
cp32-8	Circular	30 885
cp32-9	Circular	30 651
lp5[i]	Linear	5228
lp17[i]	Linear	16 928[k]
lp21[i]	Linear	18 901
lp25[i]	Linear	24 177
lp28-1[i]	Linear	28 250[k]
lp28-2	Linear	29 766
lp28-3[i]	Linear	28 601
lp28-4[i]	Linear	27 323
lp36[j]	Linear	36 849
lp38[i]	Linear	38 829
lp54	Linear	53 541
lp56 (cp32)[j]	Linear	30 349
lp56 (other)[i,j]	Linear	22 622[k]
Pseudogene plasmid [i] total		247 708
Other plasmid total		362 986
All plasmid total		610 694

Circular → linear
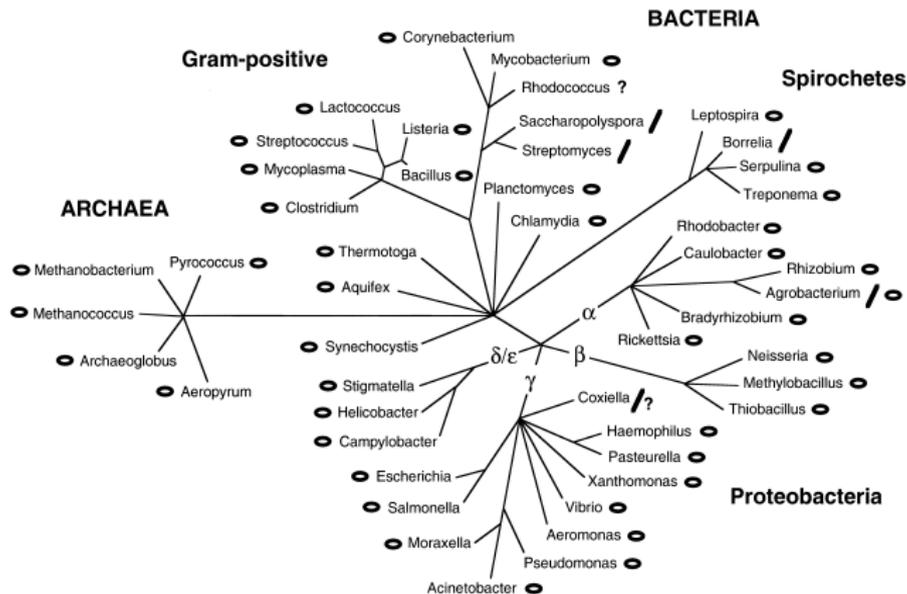
Problem with linear chromosomes

```
5'-CCCCAACCCCAACCCCAACCCCAACCCCAA.........
3'-GGGGTTGGGGTTGGGGTTGGGGTTGGGGTT.........
```

After replication, strands have separated, complements are synthesized

```
5'-CCCCAACCCCAACCCCAACCCCAACCCCAA.........
3'-GGGGTTGGGGTTGGGGTTGGGGTTGGGGTT.........

5'-                      ACCCCAACCCCAA.........
3'-GGGGTTGGGGTTGGGGTTGGGGTTGGGGTT.........
```

If nothing happens to fix things, a second round of replication will yield the following:

```
5'-CCCCAACCCCAACCCCAACCCCAACCCCAA.........
3'-GGGGTTGGGGTTGGGGTTGGGGTTGGGGTT.........

5'-                      ACCCCAACCCCAA.........
3'-GGGGTTGGGGTTGGGGTTGGGGTTGGGGTT.........

5'-          ACCCCAACCCCAA......... bases
3'-          TGGGGTTGGGGTT......... lost

5'-                          CAACCCCAA.........
3'-GGGGTTGGGGTTGGGGTTGGGGTTGGGGTT.........
```

G+C content

G+C content

- Is calculated in percentage of G+C : $100\frac{[G+C]}{[A+T+C+G]}$
- First nucleic acid technology applied to bactrerial systematics
- One of the genomic characteritics recommended for the description of species and genera
- 5% and 10% are the common range found within a species and a genera, respectively
- Modulates the aminoacid content of proteins
- Source of troubles for phylogenetic inference

Bacterial genome structures
   G+C content
     Introduction

DNA double helix
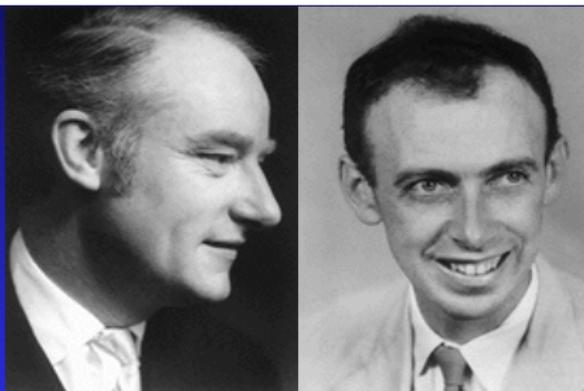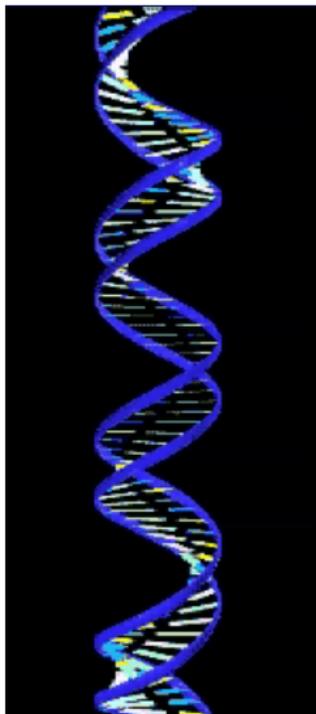
Watson and Crick, Nature, 1953

The original figure

This figure is purely diagrammatic. The two ribbons symbolize the two phosphate—sugar chains, and the horizontal rods the pairs of bases holding the chains together. The vertical line marks the fibre axis

G+C content is the same for both strands

Let

$$A_a C_c G_g T_t$$

the primary formula on one strand. Then, its complementary strand composition is given by :

$$A_t C_g G_c T_a$$

The G+C content is not affected :

$$\frac{g + c}{a + c + g + t} = \frac{c + g}{t + g + c + a}$$

DNA denaturation

ssDNA & dsDNA absorbance

Double-stranded state (dsDNA)

Single-stranded state (ssDNA)
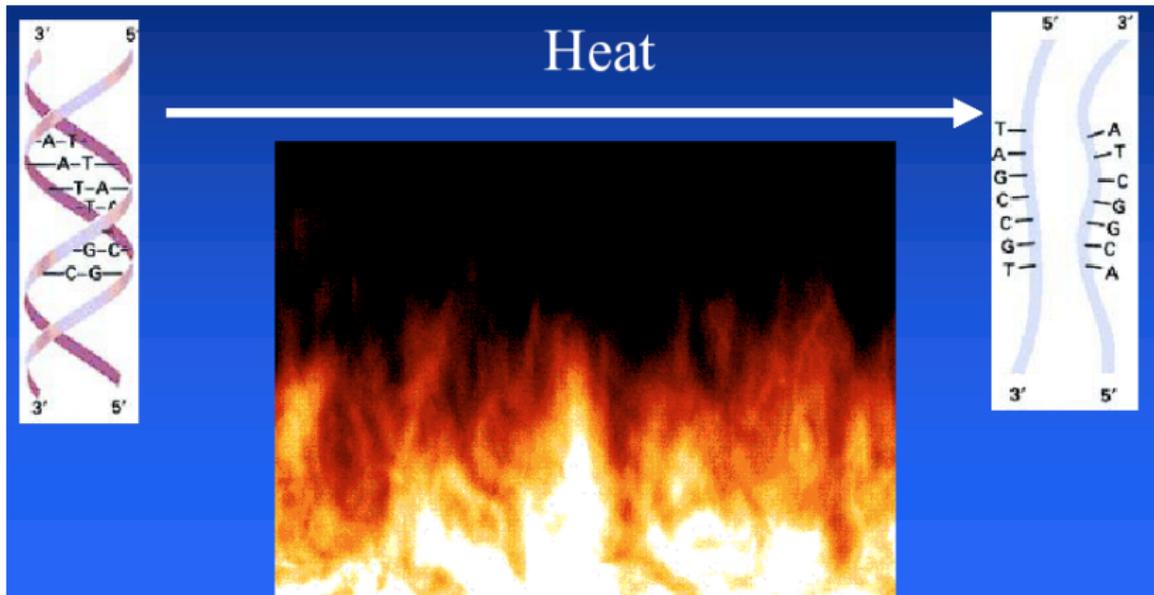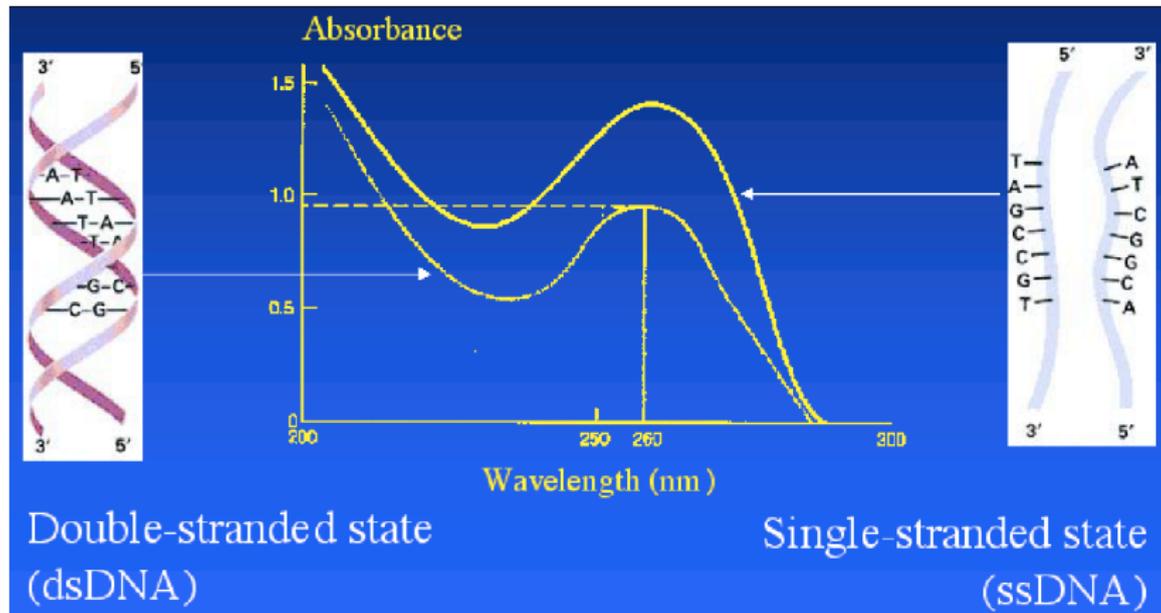
T_m is the temperature at midpoint of transition
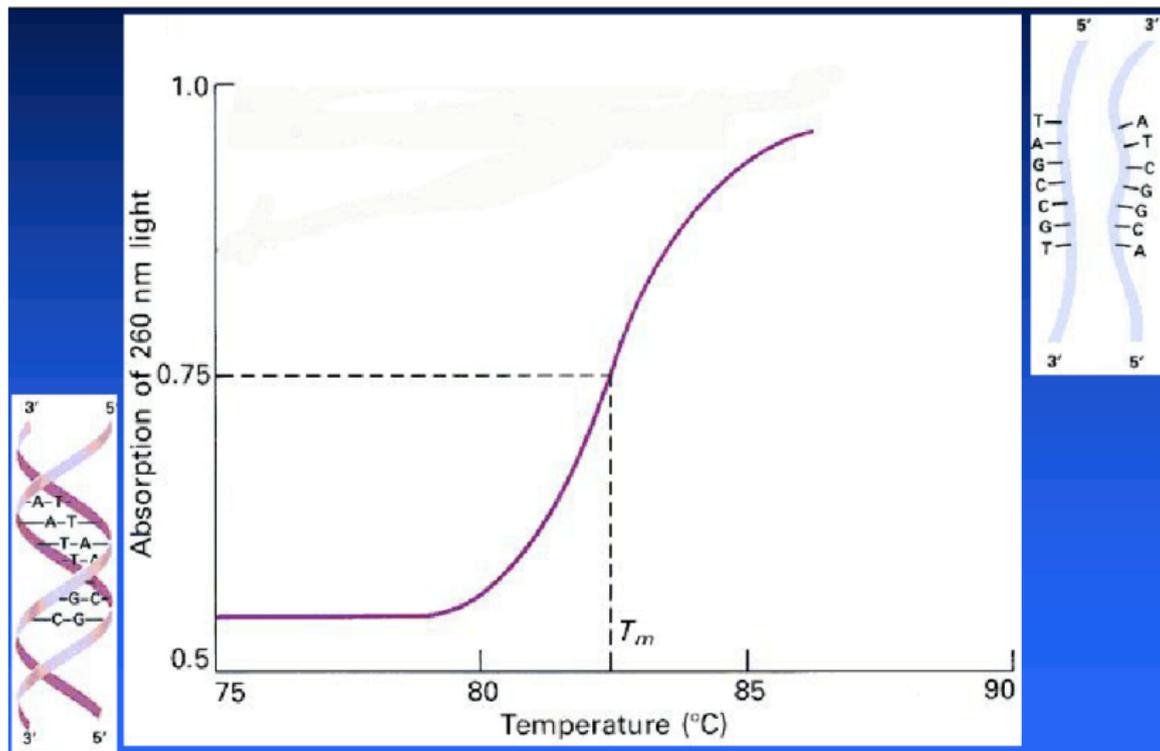
T_m increases with DNA G+C content

AT pairs:
2 hydrogen
bonds
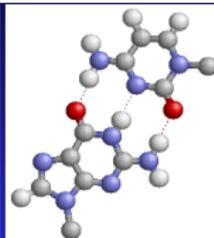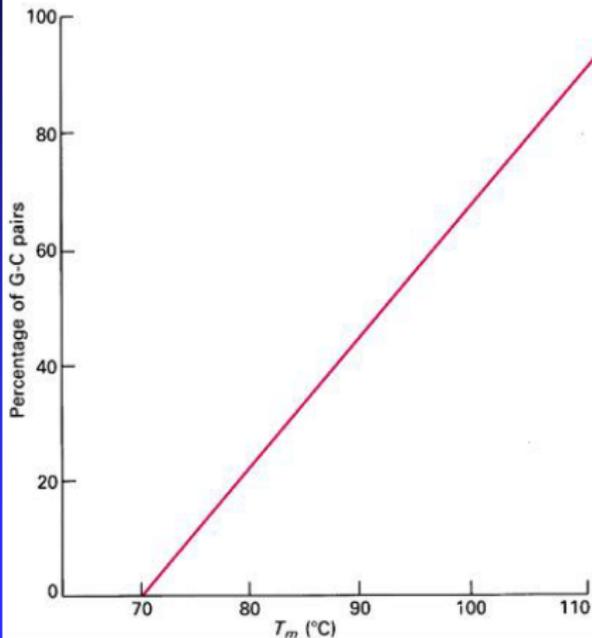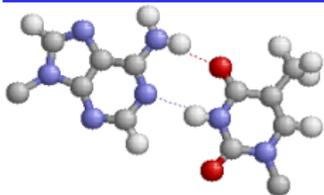
GC pairs:
3 hydrogen
bonds

Bacterial genome structures
    G+C content
        Between species variability

What is the distribution of G+C content in "bacteria"?

Study this yourself :

1. extract G+C data from GOLD
   http://www.genomesonline.org/ and study its
   distribution.

2. study data from ftp:
   //pbil.univ-lyon1.fr/pub/datasets/JME97/species.
   What is the relationship with optimum growth temperature
   T_{opt} ?

3. study data from
   http://pbil.univ-lyon1.fr/R/donnees/gc02.txt. What
   is the relationship with (an)aerobiosis ?

G+C content from GOLD

GOLD data 05–MAR–2006 (n = 682 strains)
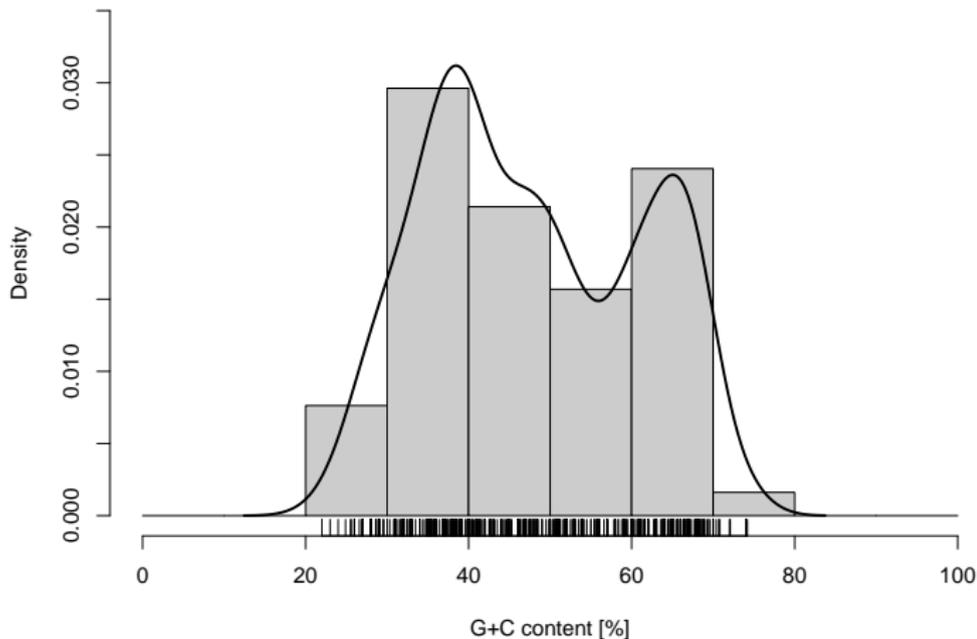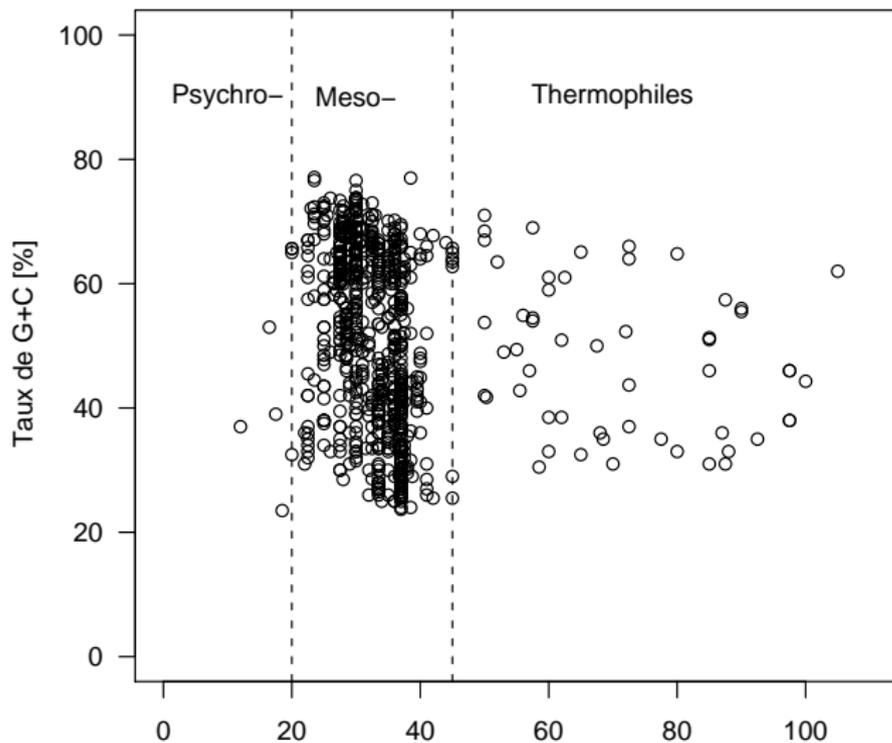
G+C content and T_{opt}

G+C content and T_{opt}

Bacterial genome structures
G+C content
Between species variability
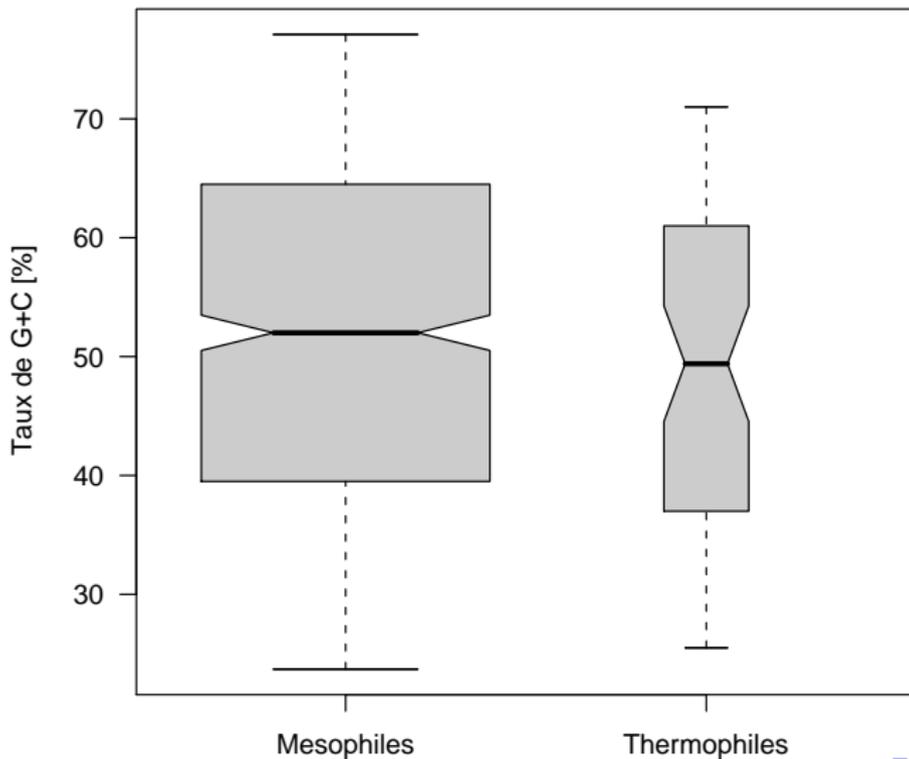
G+C content and aerobiosis

Bacterial genome structures
  G+C content
    Between species variability
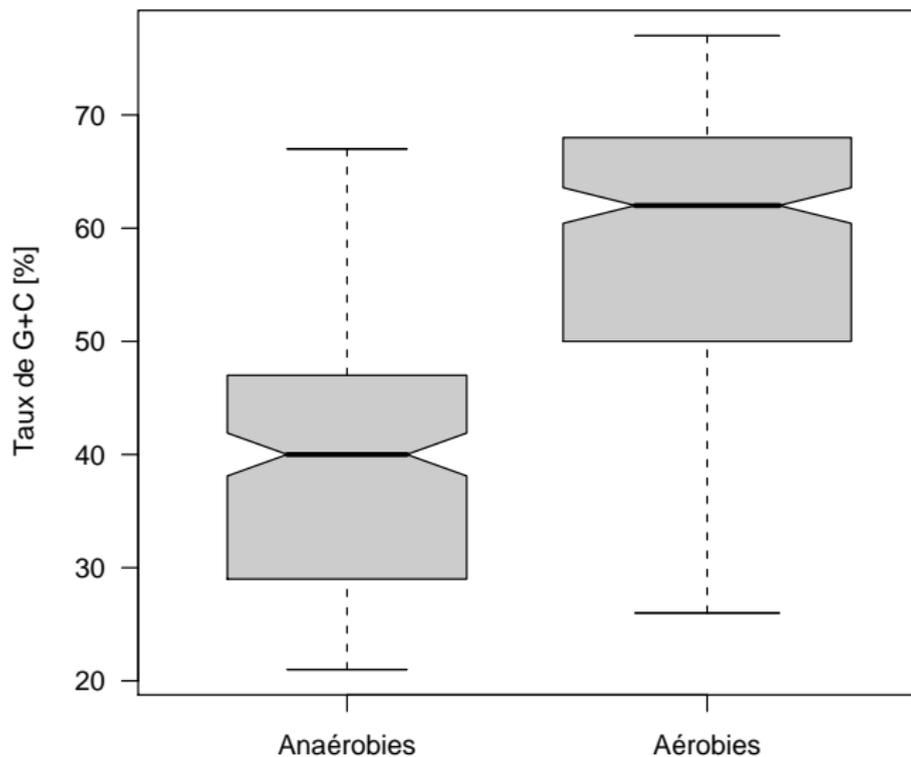
The distribution of G+C content in "bacteria"

Results from your study :

1. G+C content ranges from $\approx 25\%$ to 75% in "bacteria".

2. G+C content is not correlated with T_{opt}.

3. G+C content is correlated with aerobiosis : aerobic "bacetria" have a significant higher G+C than anerobic "bacteria".

Bacterial genome structures
G+C content
Between species variability

Underlying mechanism

Symmetric Directional Mutation Pressure :

Noboru Sueoka

AT pair

GC pair

Mutation rate

v

u
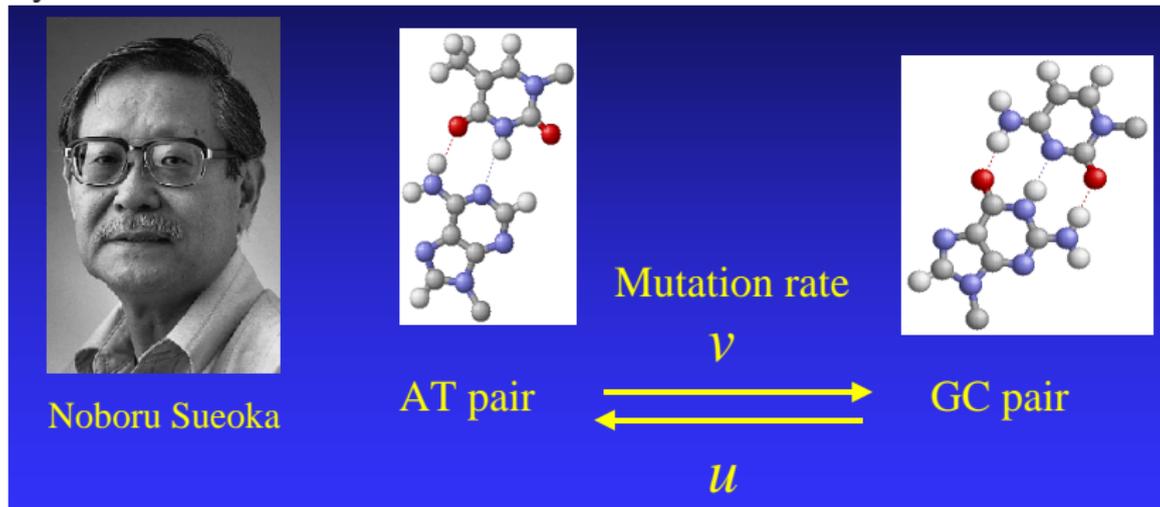
Sueoka, N. (1962) On the genetic basis of variation and heterogeneity of DNA base composition. *Proc. Natl. Acad. Sci. USA*, **48** :582-592

Bacterial genome structures
  G+C content
    Between species variability

Underlying mechanism

Direct experimental evidence :

Cox, E.C., Yanofsky, C. (1967) Altered base ratios in the DNA of an *Escherichia coli* mutator strain. *Proc. Natl. Acad. Sci. USA*, **58** :1895-1902

Accelerated evolution experiment with a mutator strain : G+C content variation visible at a lab time scale.

Bacterial genome structures
G+C content
Between species variability

Underlying mechanism

$$\frac{d\theta}{dt} = v(1 - \theta) - u\theta$$

$$\theta(t) = \left(\theta_0 - \frac{v}{u+v}\right) e^{-(u+v)t} + \frac{v}{u+v}$$

G+C content at equilibrium :

$$\theta^\star = \theta(+\infty) = \frac{v}{u+v}$$

Underlying mechanism
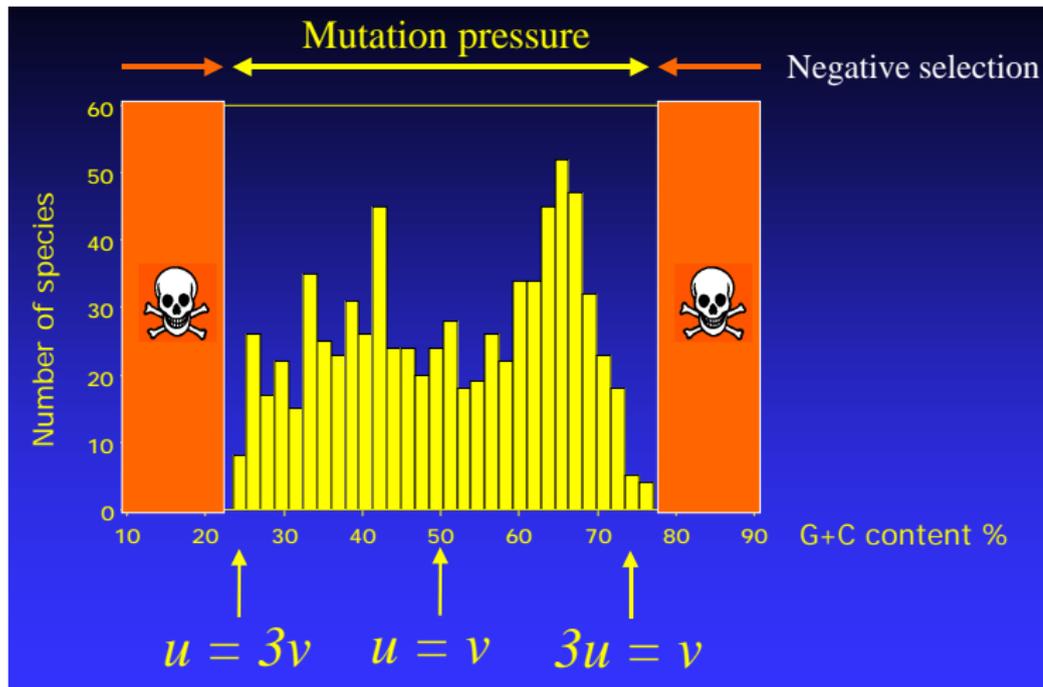
G+C content and aa content

The impact of G+C content on the amino-acid composition of proteins was known even before the deciphering of the genetic code :
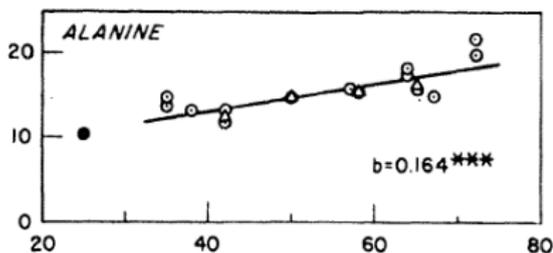
Sueoka, N. (1961) Correlation between base composition of deoxyribonucleic acid and amino acid composition of protein. *Proc. Natl. Acad. Sci. USA*, **48** :582-592

- Was used as a clue to crack the genetic code (*e.g.* here Ala codons are expected to be G+C rich, and yes indeed GCN codons are G+C rich !).
- First evidence that the genetic code is (almost) universal.

Bacterial genome structures
G+C content
G+C content and amino-acid content in proteins

A null hypothesis for the aa content of proteins

- CDS are build by random sampling from an urn with a given G+C content θ.
- We assume for the shake of simplicity that $C = G = \frac{\theta}{2}$ and $A = T = \frac{1-\theta}{2}$.
- What would be the amino-acid composition of proteins under this simplistic model ?

Bacterial genome structures
G+C content
G+C content and amino-acid content in proteins

A null hypothesis for the aa content of proteins

Let $X_i \in \{A, C, G, T\}$ a random variable for the result of outcome number i. Note $P_A = P(X_i = A)$, $P_C = P(X_i = C)$, $P_G = P(X_i = G)$, $P_T = P(X_i = T)$ the probabilities for the four bases. We have assumed that $P_C = P_G = \frac{\theta}{2}$ and $P_A = P_T = \frac{1-\theta}{2}$.

The probability for codon GAA is for instance :
$P(GAA) = P(X_1 = G \cap X_2 = A \cap X_3 = A) = P_G P_A P_A$
In coding sequences there are no stop codons (TAA, TAG or TGA) so that :
$P(GAA|\mathrm{not-stop}) = \frac{P(GAA)}{P(\mathrm{not-stop})} = \frac{P_G P_A P_A}{1-(P_T P_A P_A + P_T P_A P_G + P_T P_G P_A)}$

Bacterial genome structures
G+C content
G+C content and amino-acid content in proteins

A null hypothesis for the aa content of proteins

At the amino-acid level, Glu is encoded by GAA or GAG, so that :

$$P(Glu) = P(GAA \cup GAG | \mathrm{not-stop}) =$$

$$P(GAA | \mathrm{not-stop}) + P(GAG | \mathrm{not-stop})$$

In a similar way, we can deduce of the expected frequencies for all amino-acids under the model.

A null hypothesis for the aa content of proteins

$$P(\theta, aa) = \frac{f(\theta)}{8 - (1 - \theta)^2 (1 + \theta)}$$

$$f(\theta) = \begin{cases} (1 - \theta)^2 (2 - \theta) & \text{si } aa \in \{\text{Ile}\} \\ (1 - \theta)^2 & \text{si } aa \in \{\text{Phe, Lys, Tyr, Asn}\} \\ 1 - \theta^2 & \text{si } aa \in \{\text{Leu}\} \\ (1 - \theta)^2 \theta & \text{si } aa \in \{\text{Met}\} \\ (1 - \theta)\theta & \text{si } aa \in \{\text{Asp, Glu, His, Gln, Cys}\} \\ 2(1 - \theta)\theta & \text{si } aa \in \{\text{Val, Thr}\} \\ 3(1 - \theta)\theta & \text{si } aa \in \{\text{Ser}\} \\ (1 - \theta)\theta^2 & \text{si } aa \in \{\text{Trp}\} \\ \theta(\theta + 1) & \text{si } aa \in \{\text{Arg}\} \\ 2\theta^2 & \text{si } aa \in \{\text{Gly, Pro, Ala}\} \end{cases}$$

Bacterial genome structures
G+C content
G+C content and amino-acid content in proteins

What is the impact of G+C on the aa content ?

Study this yourself :
ftp://pbil.univ-lyon1.fr/pub/datasets/GENE97/README

- Differentiate cytoplasmic proteins and integral membrane proteins (by color or point style).
- Show the influence of G+C content for Ala, Lys, and Glu (at least).
- Add the null model as a line.

Replichores

1. Introduction

2. Genome size

3. Topology & #

4. G+C content

5. Replichores

6. Gene orientation biases

7. Chrirochores

8. X-rated structure

Replichores : origin and terminus of replication

Eukaryota

Bacteria

Many origins
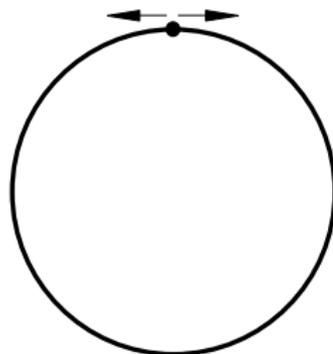
A single origin

There are two replichores per chromosome

$\approx \pi$
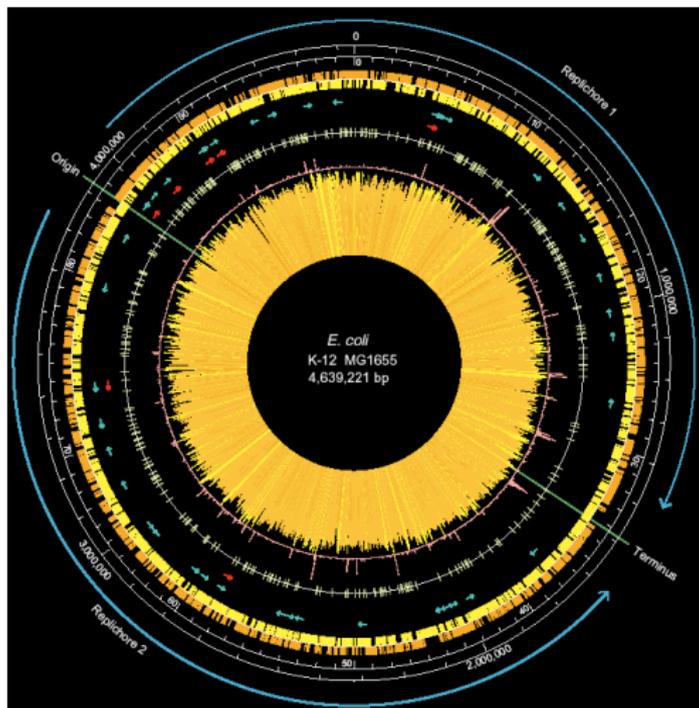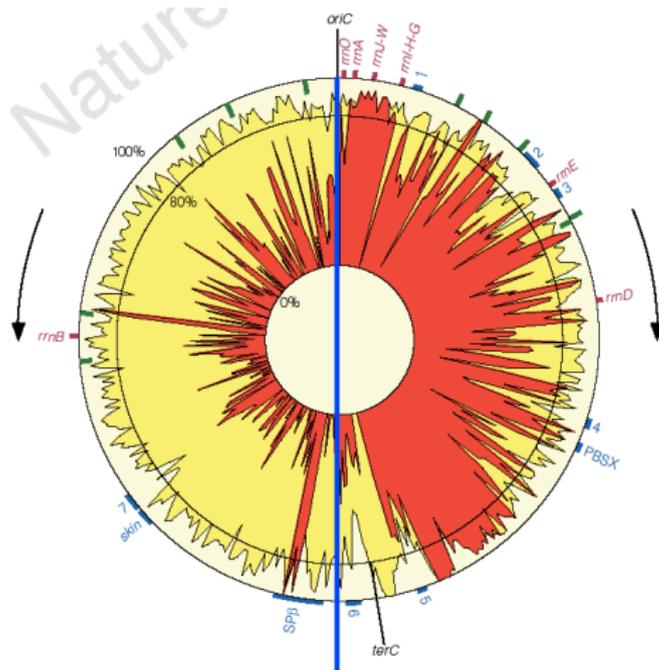
Archae & Bacteria

Small circular genome with a single origin of replication

Similar replication factors

Archaea

Bacteria

Eukarya
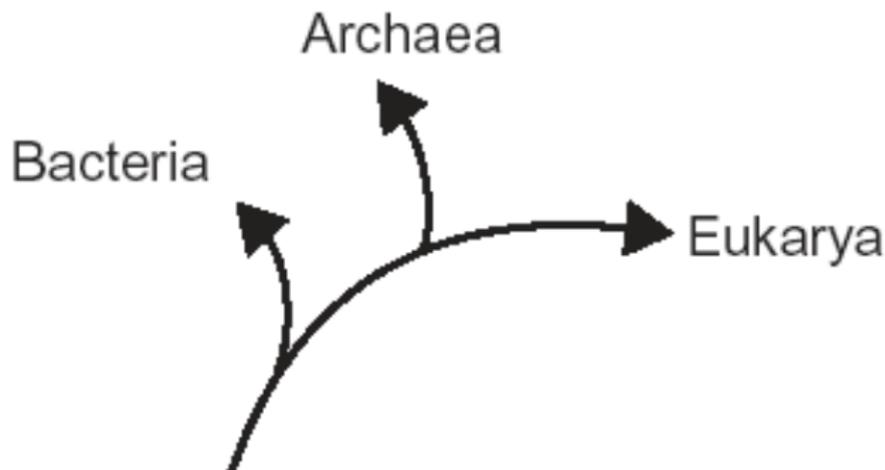
Looking for the origin
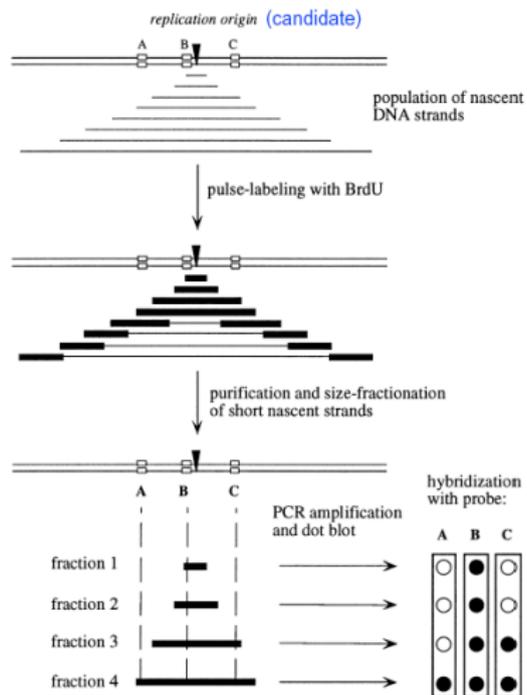
Looking for the origin in *E. coli*
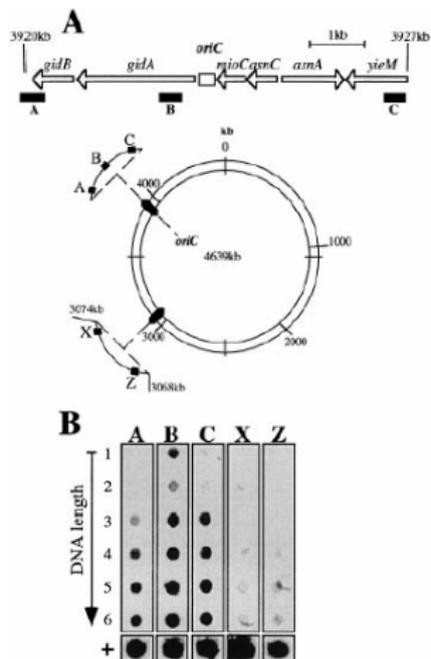
Looking for the origin in *B. burgdorferi*
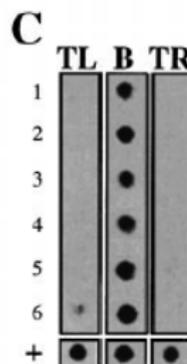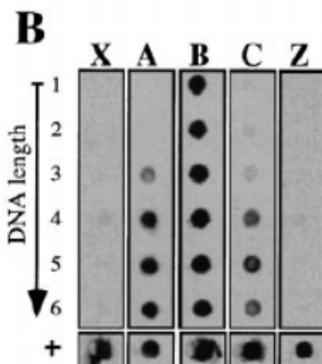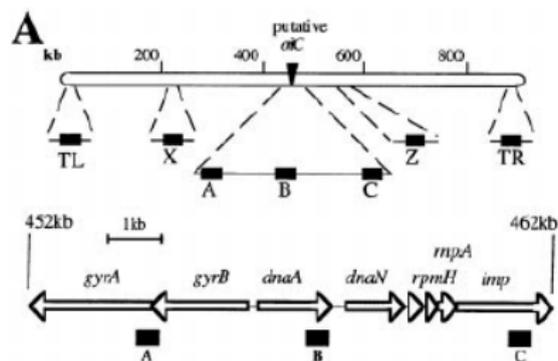
Zoom at the origin
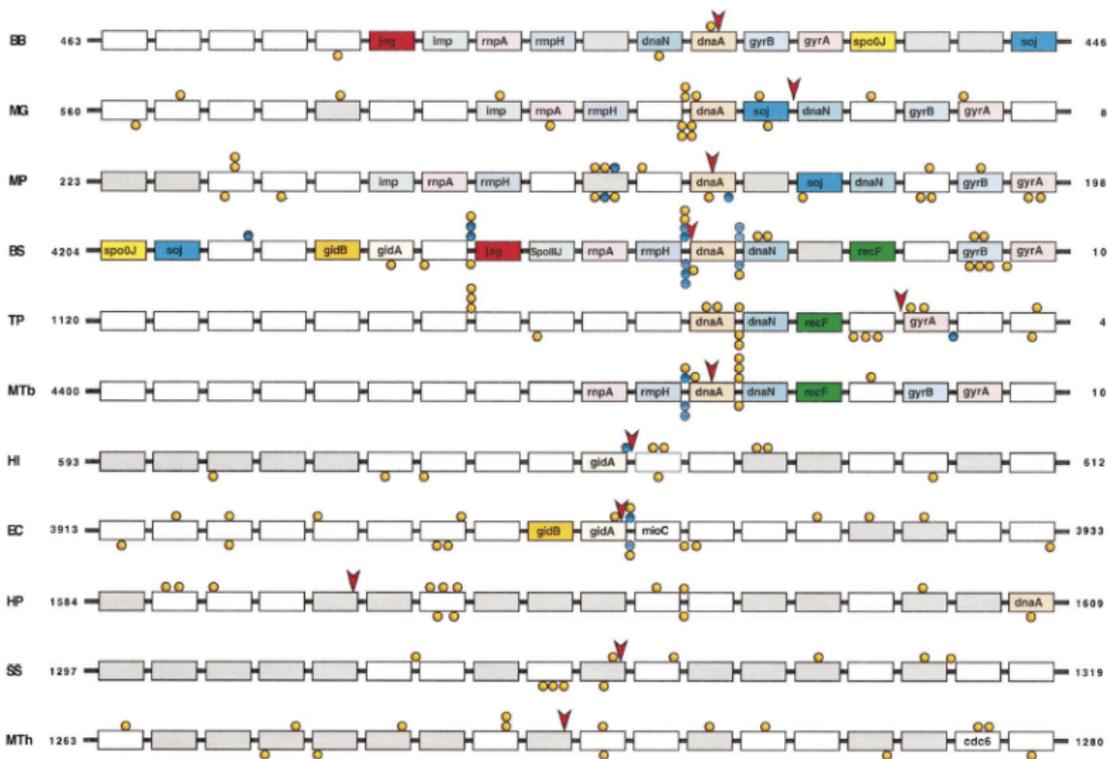
Gene orientation biases

Leading and lagging CDS
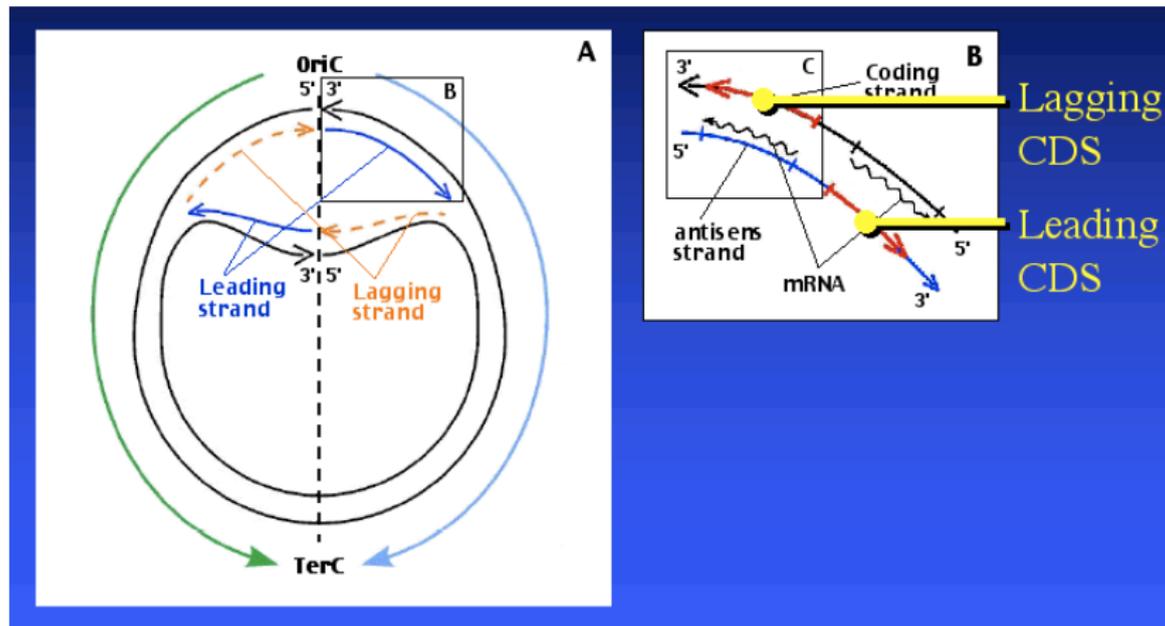
More leading CDS than lagging CDS
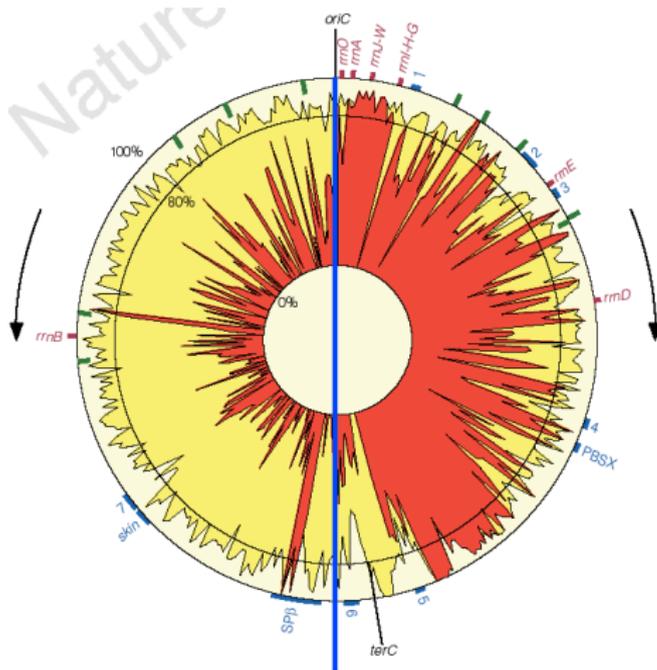
Lactobacillus plantarum
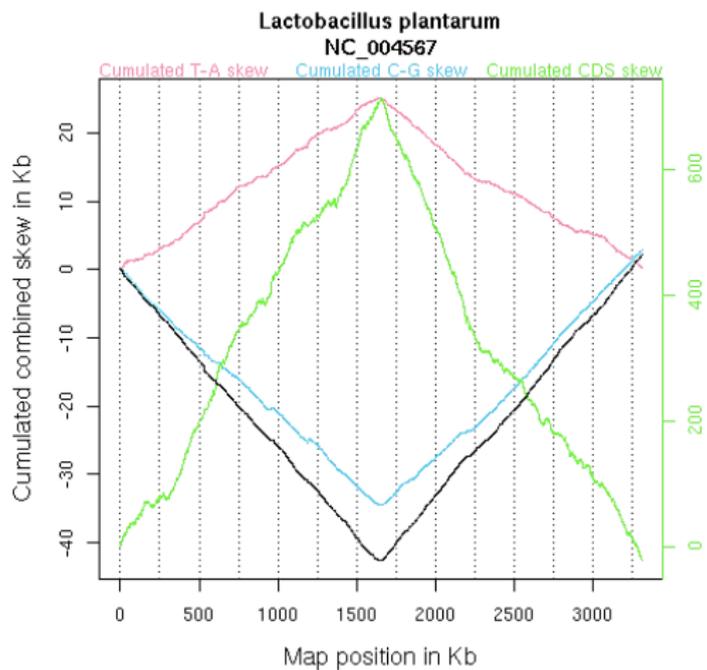
Legionella pneumophila
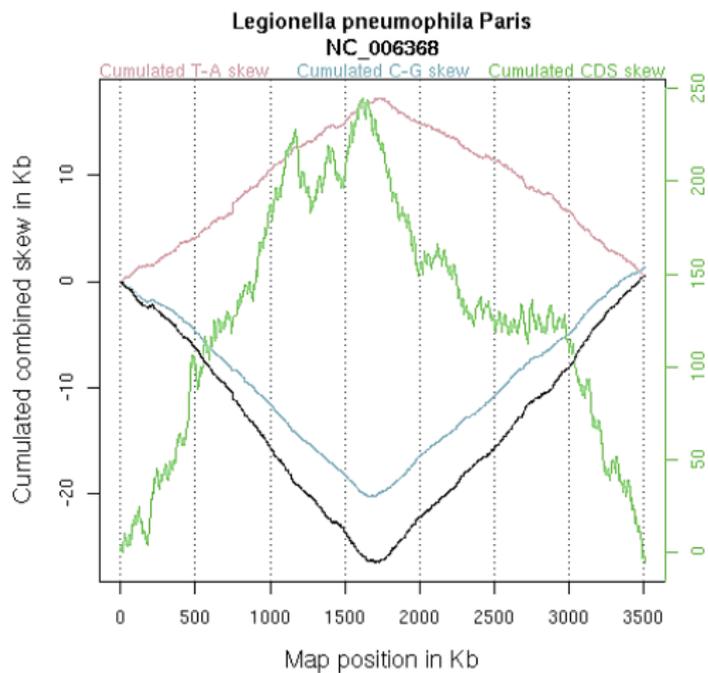
Nostoc sp
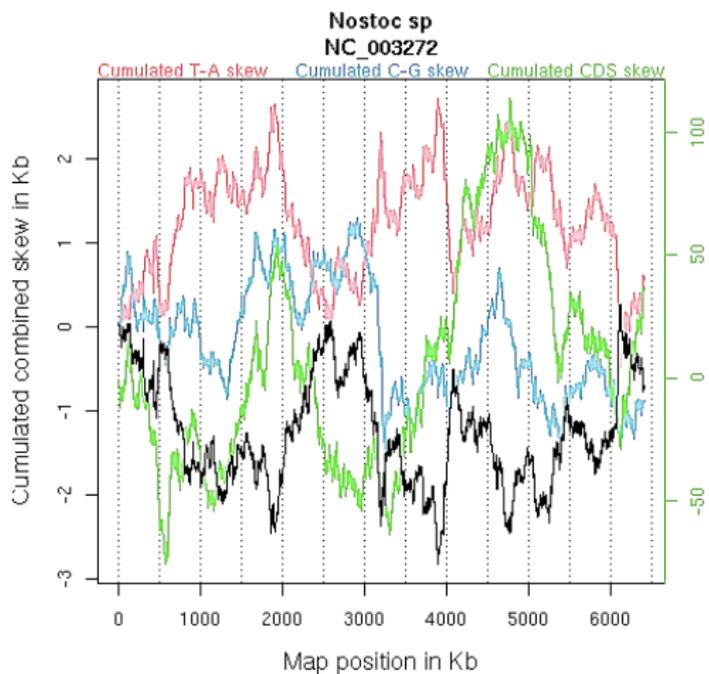
Collisions between polymerases

Related Articles, Books

Consequences of replication fork movement through transcription units in vivo.

French S.

Department of Biology, University of Virginia, Charlottesville 22903.

To examine the basis for the evolutionary selection for codirectionality of replication and transcription in Escherichia coli, electron microscopy was used to visualize replication from an inducible ColE1 replication origin inserted into the Escherichia coli chromosome upstream (5') or downstream (3') of rrnB, a ribosomal RNA operon. Active rrnB operons were replicated either in the same direction in which they were transcribed or in the opposite direction. In either direction, RNA polymerases were dislodged during replication. When replication and transcription were codirectional, the rate of replication fork movement was similar to that observed in nontranscribed regions. When replication and transcription occurred in opposite directions, replication fork movement was reduced.

Connection with essentiality
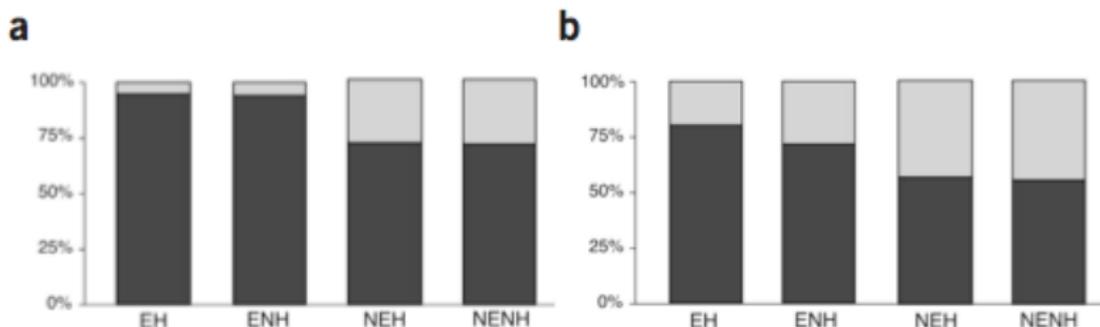
Figure 1 Distribution of genes between the leading (dark gray) and the lagging (light gray) strands of the genome of *B. subtilis* (**a**) and *E. coli* (**b**). H, highly expressed; NH, non-highly expressed; E, essential; NE, non-essential.

Rocha, E.P.C. & Danchin, A. (2003) *Nature Genetics*, **34** :377-378.

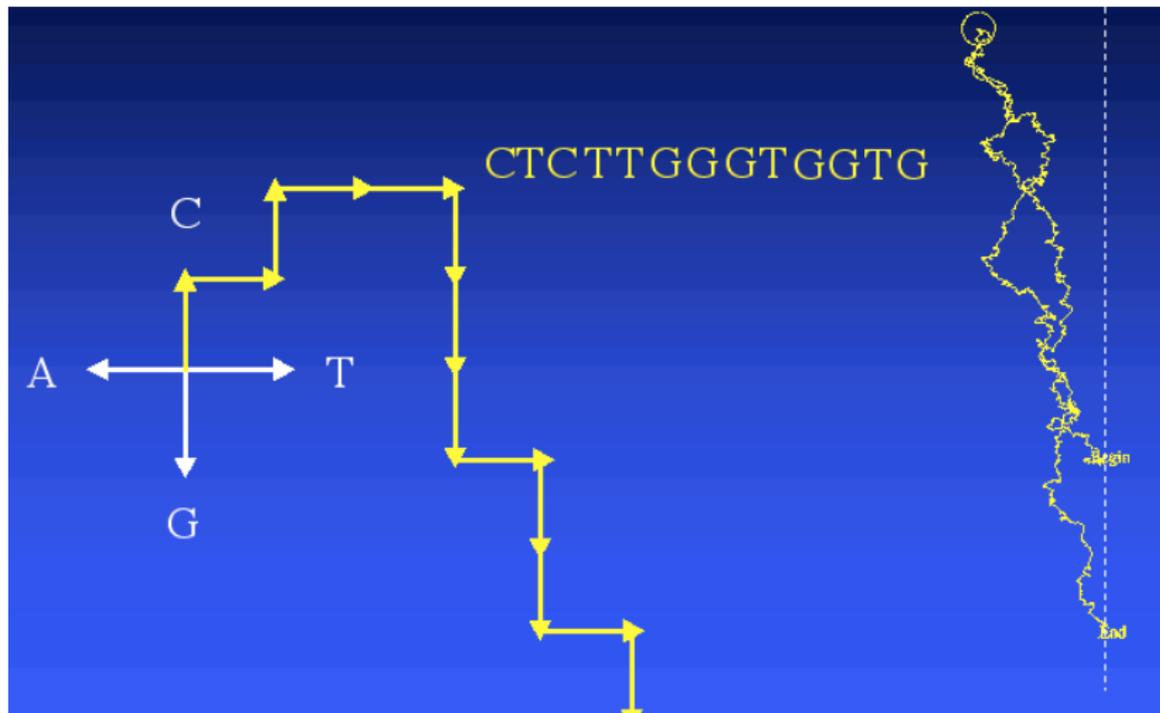# Chrirochores : base composition biases

Chirochores in *B. burgdorferi*
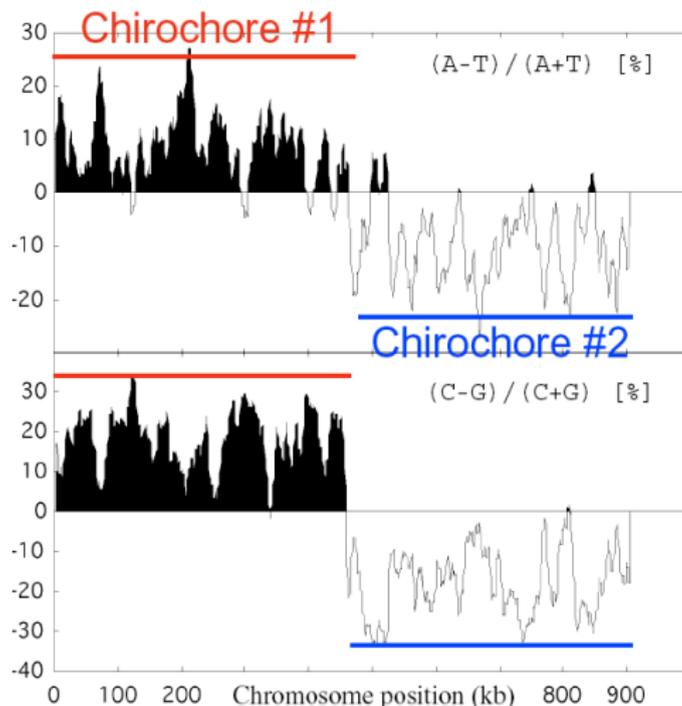
Chirochores in *B. burgdorferi* (third codon positions)
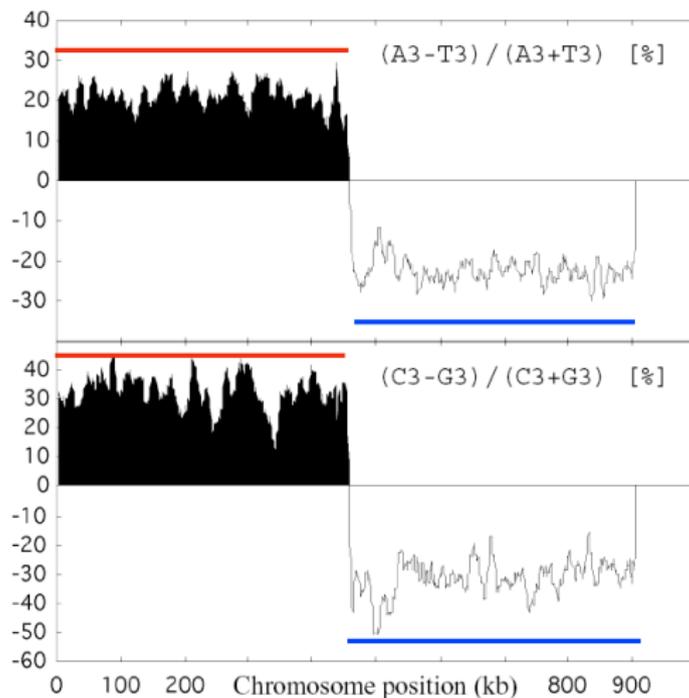
Usually GC skew $>$ AT skew *e.g. E. coli*
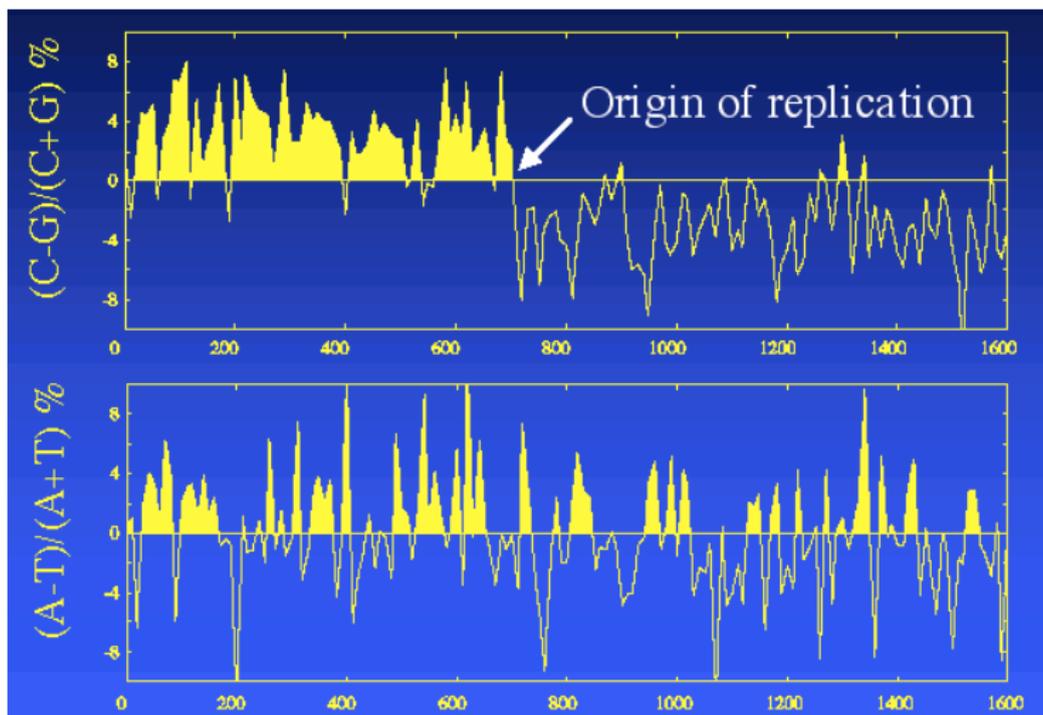
Simple DNA walk *E. coli*

Oriloc

Chlamydia trachomatis
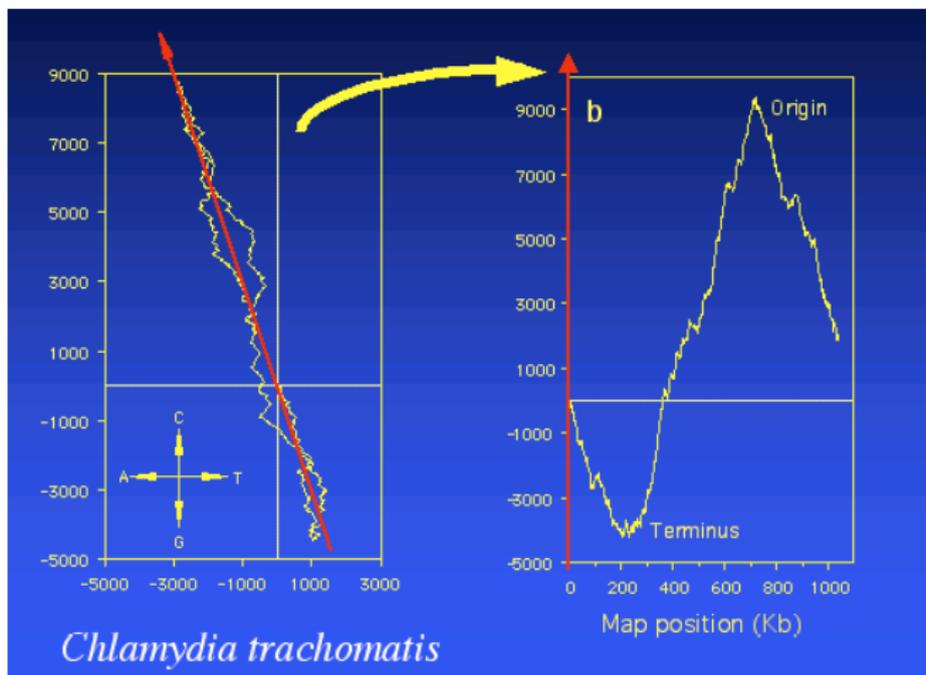
Cytosine deamination theory

Fig. 6.

Buchnera aphidicola

Klasson, L. & Anderson, S.G.E. (2006) *Mol. Biol. Evol*, **23** :1031-1039.
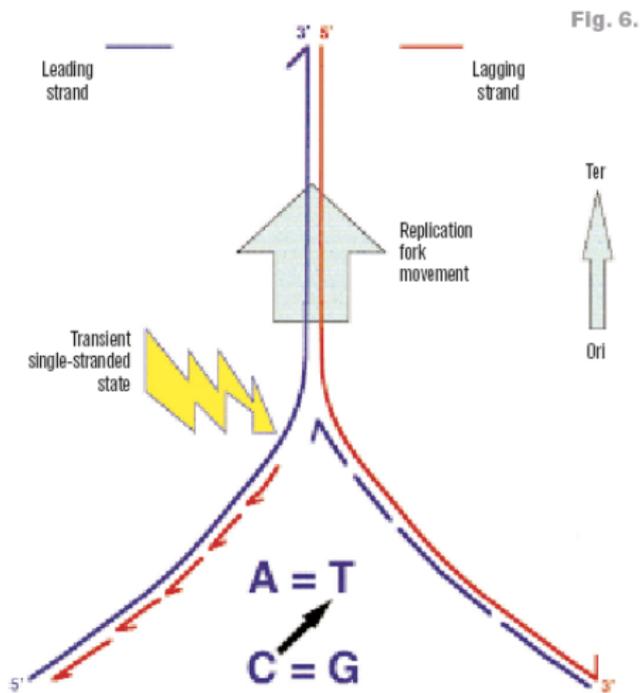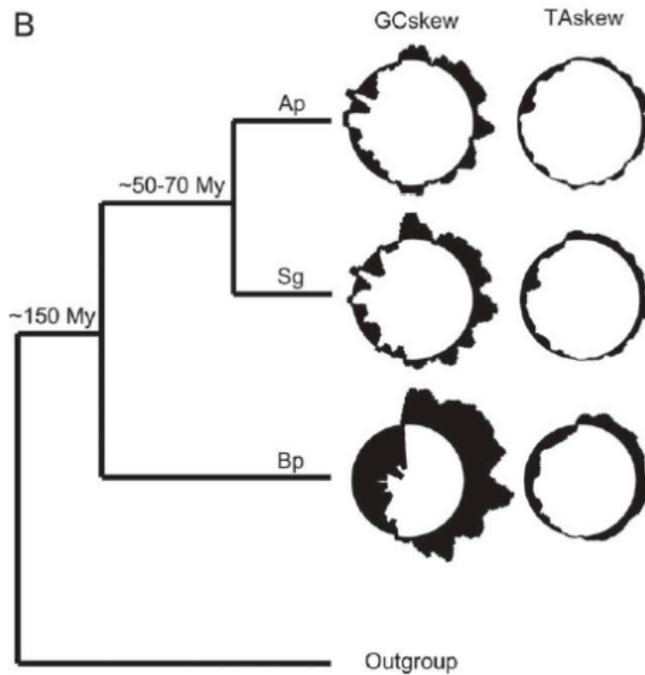
Chirochores practical

Study this yourself :

- Plot the results of oriloc() for *Chlamydia trachomatis*
  (library(seqinr))
- Have a look at oriloc() outputs there :
  http://pbil.univ-lyon1.fr/software/Oriloc/.

X-rated structure : gene order evolution

Tsuzumi drum

Gene order comparison in bacteria
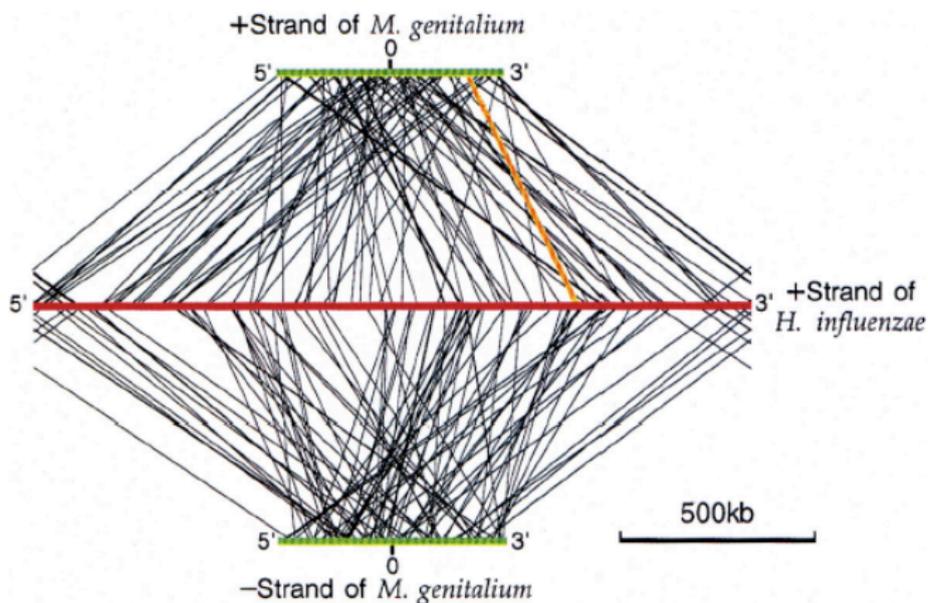
Watanabe *et al.* (1997) *J. Mol. Evol.*, **44** :s57-s64.
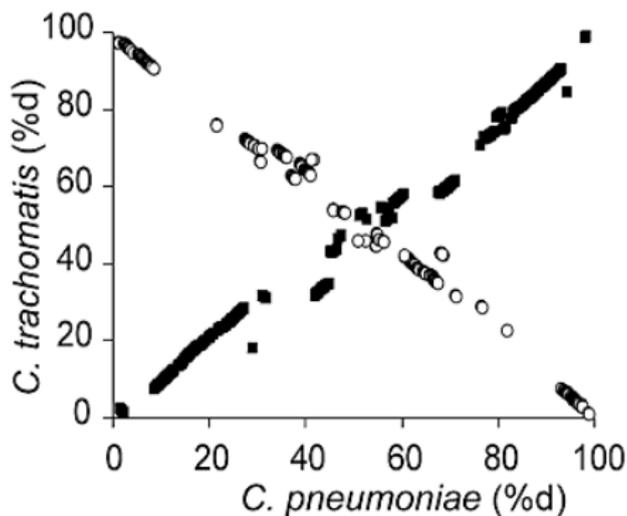
Genomic dot plots (*E. coli vs. V. cholerae*)
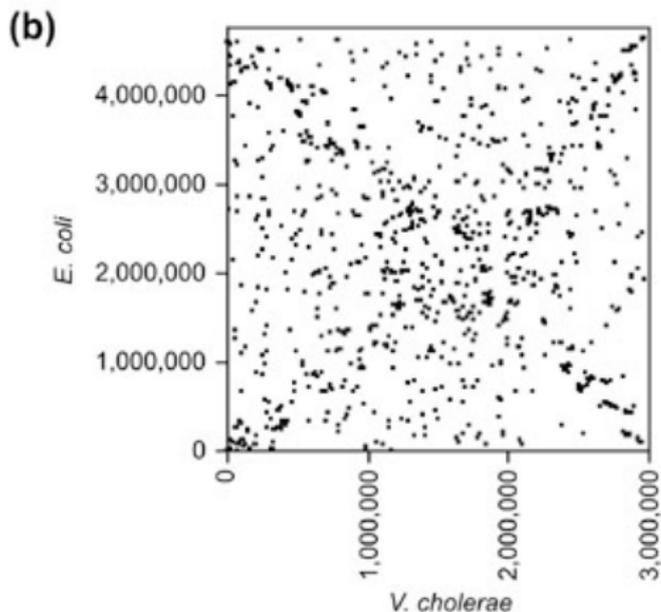
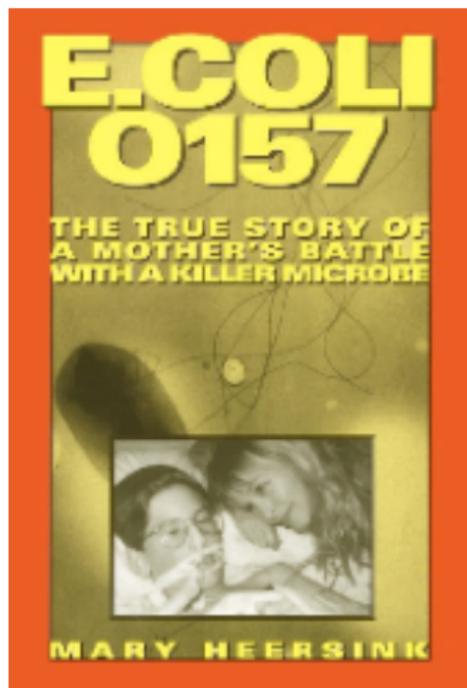Eisen *et al.* (2000) *Genome Biology*, **1** :research0011.1-9.

Genomic dot plots (*C. pneumoniae vs. C. trachomatis*)
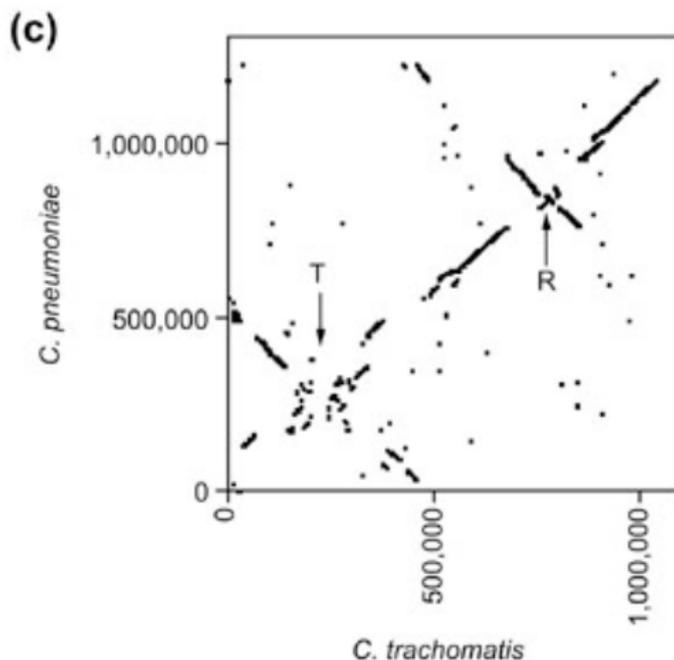
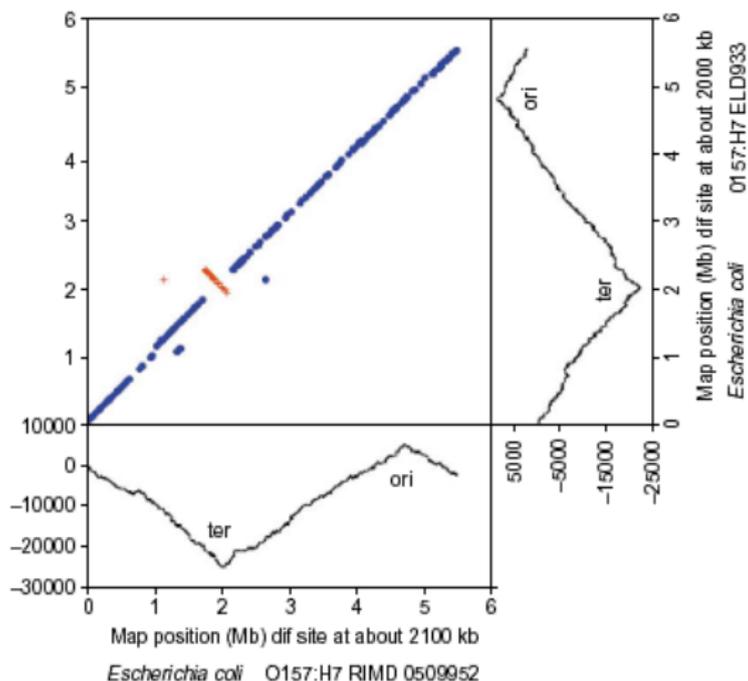Eisen *et al.* (2000) *Genome Biology*, **1** :research0011.1-9

Genomic dot plots (*C. pneumoniae vs. C. trachomatis*)

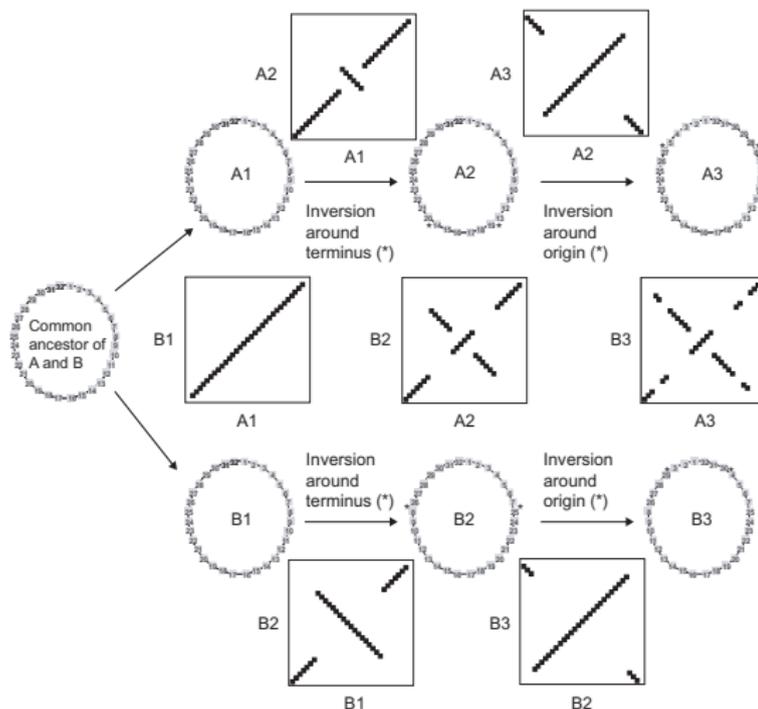Tillier & Collins (2000) *Nature Genetics*, **26** :195-197.

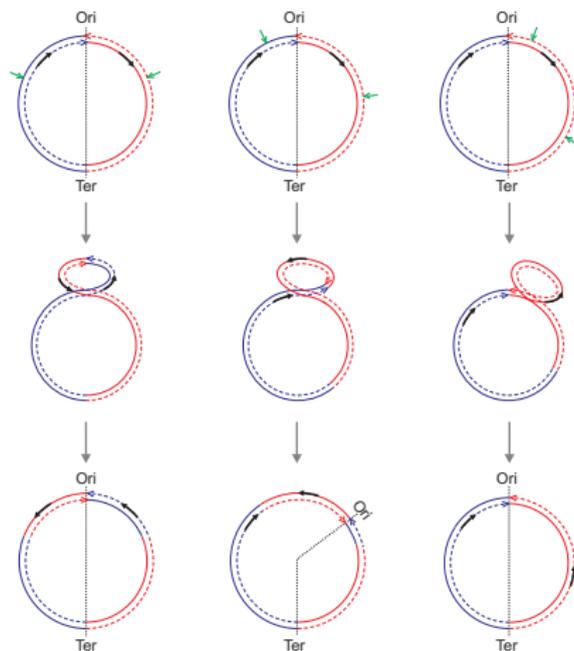## Genomic dot plots (0157 :H7 *vs.* O157 :H7)

Genomic dot plots (0157 :H7 *vs.* O157 :H7)

Lobry & Louarn (2003) *Curr. Op. Microbiol.*, **6** :101-108.

Simulation of symmetric inversions

Three models for inversions

Mackiewicz *et al.* (2001) *Genome Biology*, **2** :interactions1004.1-4.

"Bacterial" Genome structures
4BIM - INSA 2006

J. R. Lobry

Université Claude Bernard Lyon I – France

18 juin 2006