"Bacterial" Genome structures

Spring 2008 Lecture

Pr. J. R. Lobry

Université Claude Bernard Lyon I – France

Last LATEXcompilation was : February 23, 2017

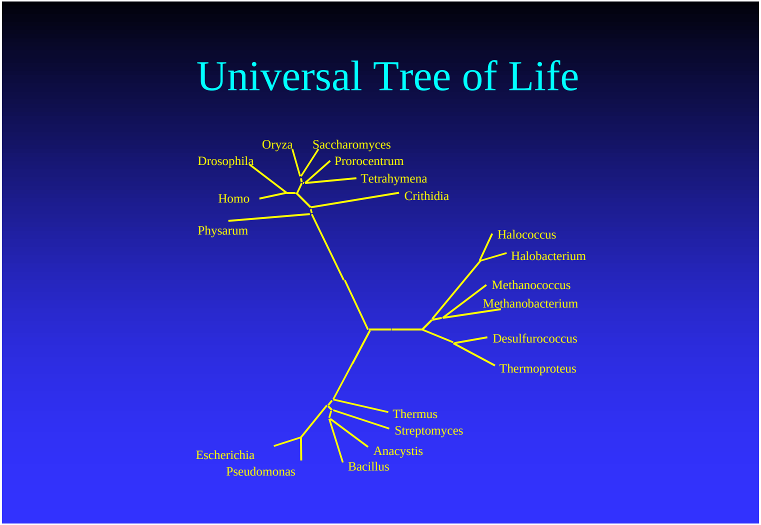
Outline

1. Introduction
2. Genome size
3. Topology & #
4. G+C content
5. Replichores
6. Gene orientation biases
7. Chirochores
8. X-rated structure

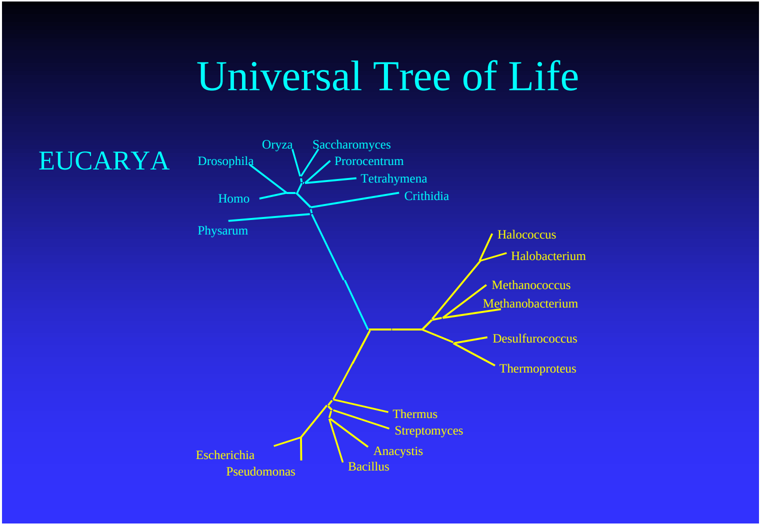
Introduction

1. Introduction
2. Genome size
3. Topology & #
4. G+C content
5. Replichores
6. Gene orientation biases
7. Chirochores
8. X-rated structure

The three kingdoms

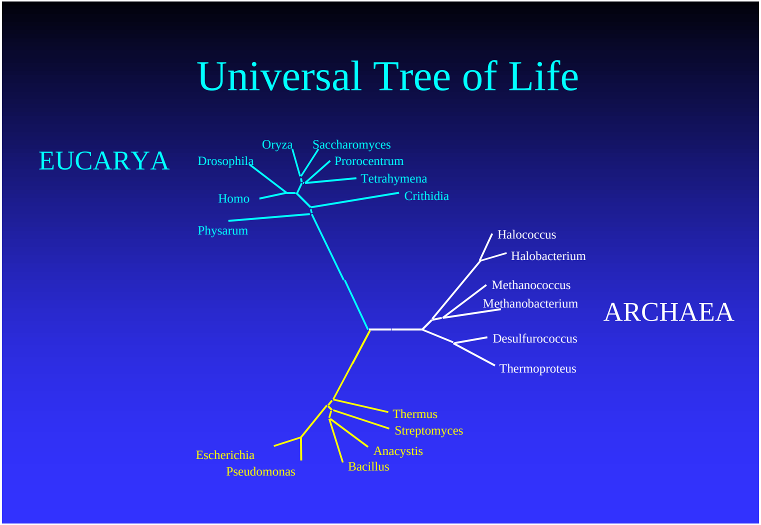


The three kingdoms: eucarya

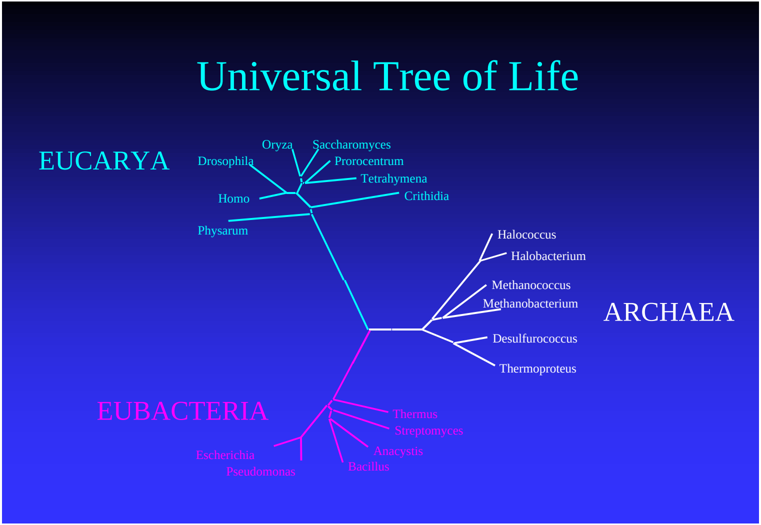


Universal Tree of Life

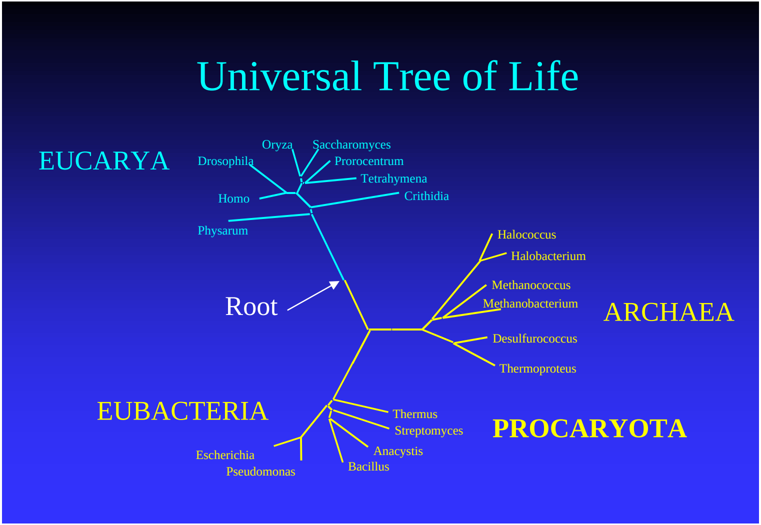
The three kingdoms: archaea



The three kingdoms: eubacteria

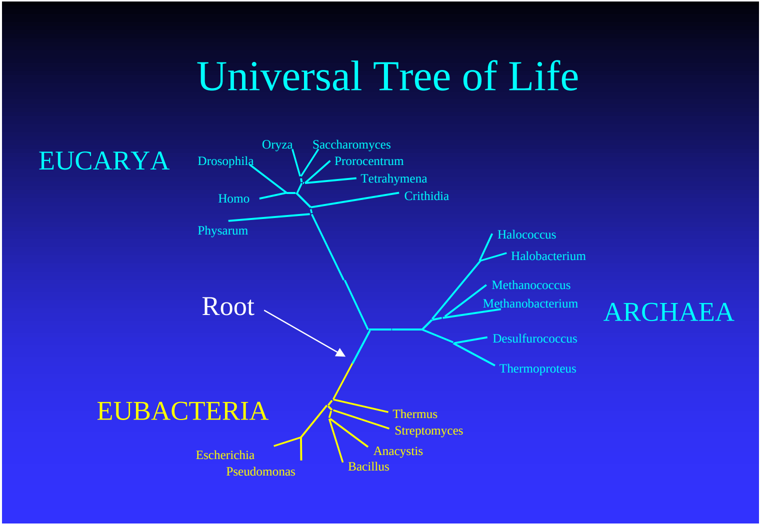


The three kingdoms: root 1



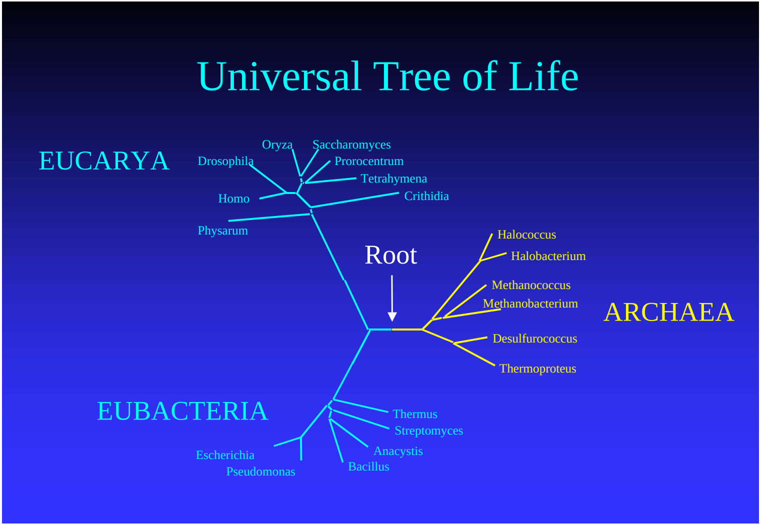
Universal Tree of Life

The three kingdoms: root 2

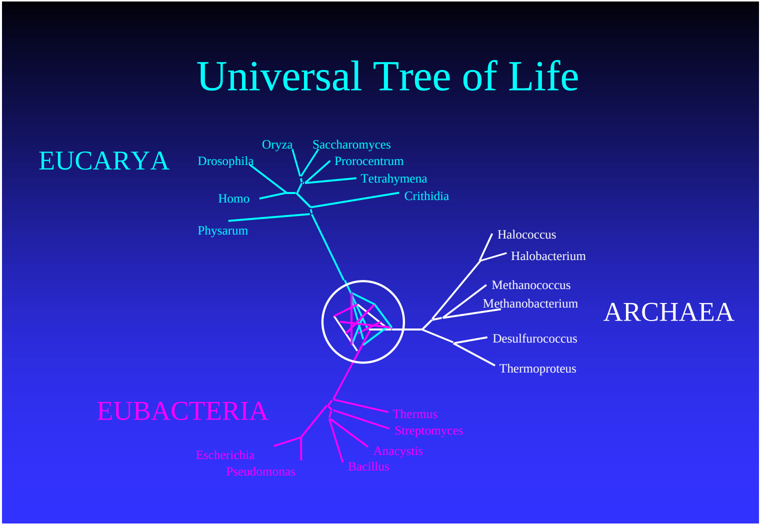


Universal Tree of Life

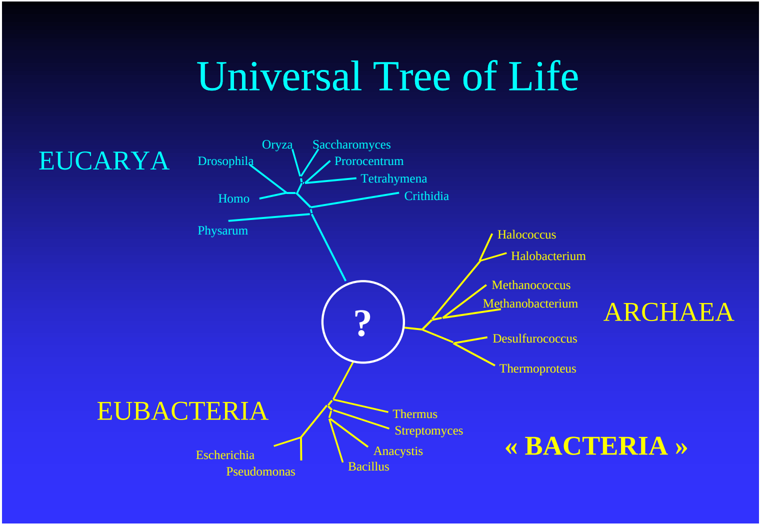
The three kingdoms: root 3



The three kingdoms: no root



The three kingdoms: ”bacteria”



Organelles: chloroplasts & mitochondria

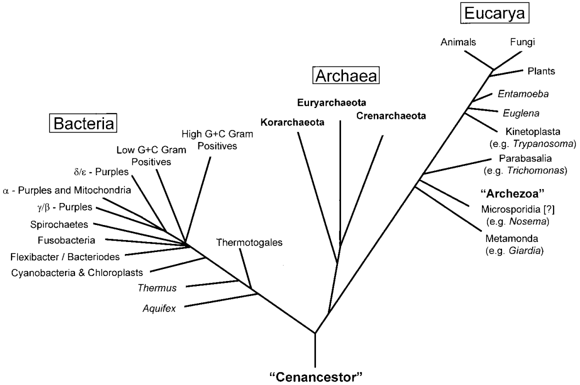
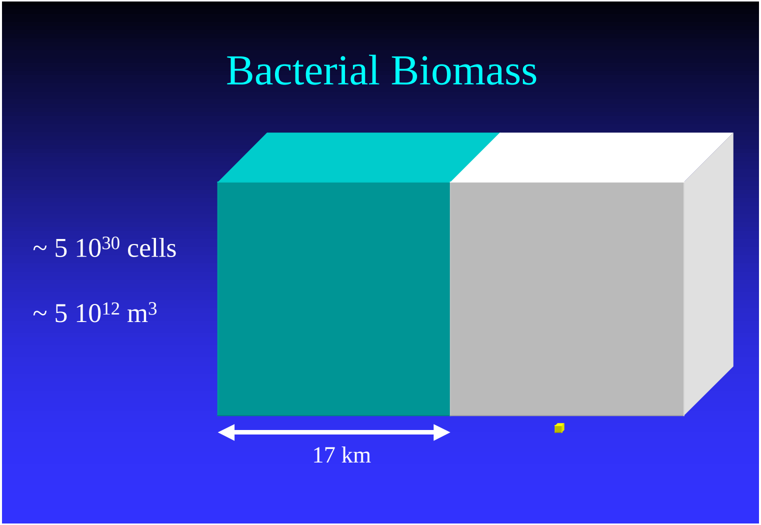


FIG. 1. Schematic drawing of a universal rRNA tree showing the relative positions of evolutionary pivotal groups in the domains *Bacteria*, *Archaea*, and *Eucarya*. The location of the root (the cenancestor) corresponds to that proposed by reciprocally rooted gene phylogenies (43, 133, 164). The question 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.)

Half of biomass on earth

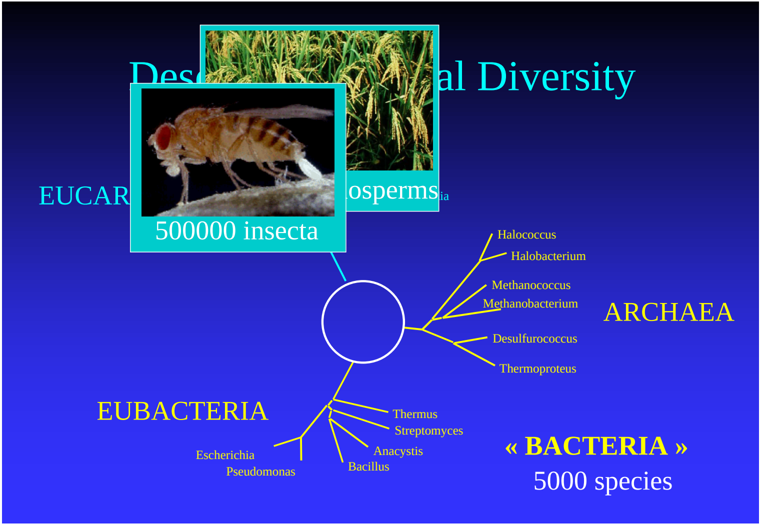


Bacterial Biomass

~ 5 10^{30} cells

~ 5 10^{12} m^3

Very few species



First species classification

An old trend

Aristotle (-384,-322)

508	Eucaryota
0	Bacteria

Small bacteria

Little Bacteria

Antony van Leeuwenhoek (1632-1723)

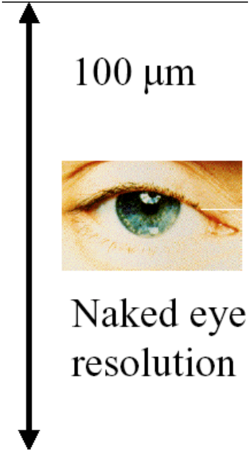
LENS
SPECIMEN
ADJUSTMENT SCREWS

0.1 mm = 100 μm

Thickness ≈ 100 μm.

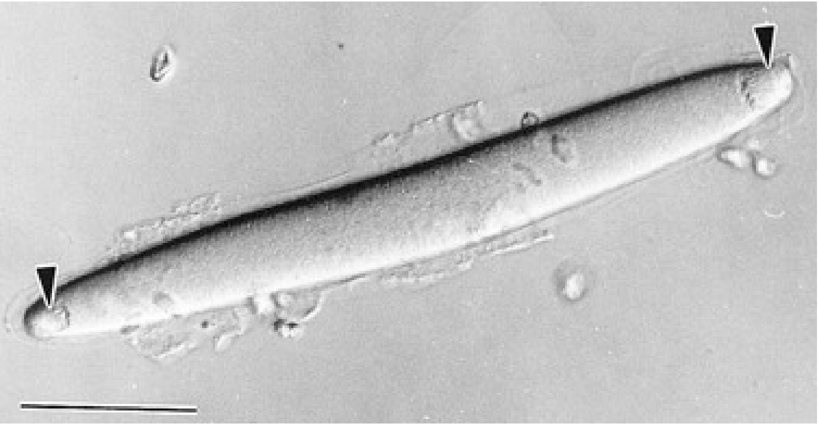
1 M€ ≈ 1 m
1 G€ ≈ 1 km

Bacterial cell size is in μm



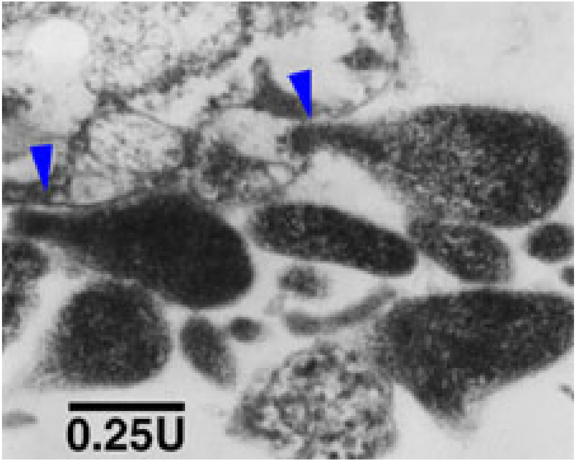
Demo

A giant: *Epulopiscium fishelsoni* bar is 50 μm

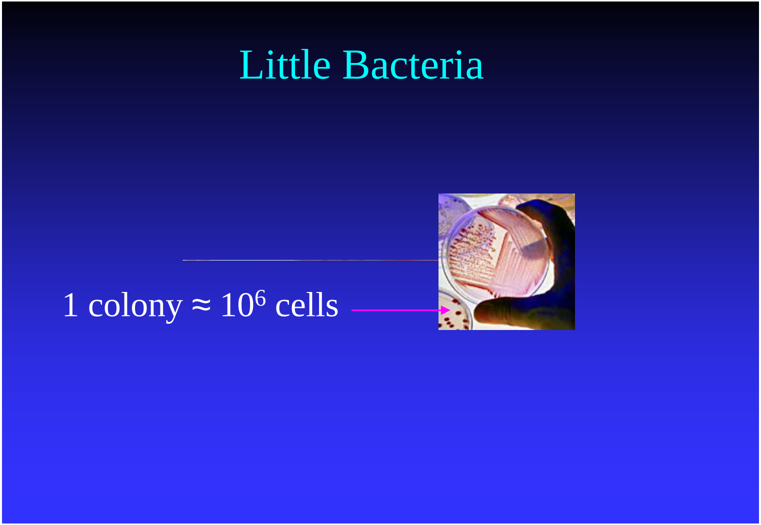


Bresler, V. *et al* (1998) *J. Bact.*, **180:**5601-5611.

Mycoplasma genitalium bar is 0.25 μm



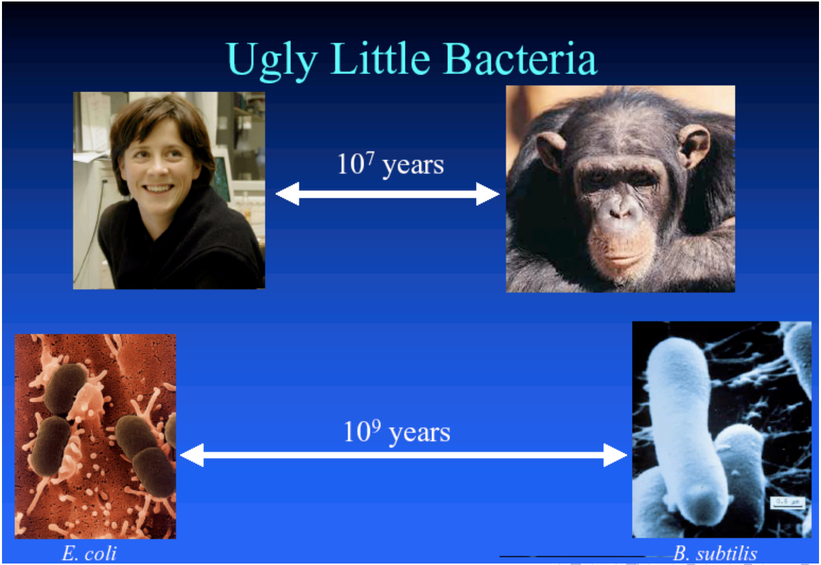
1 colony $\approx 10^6$ cells



Little Bacteria

1 colony $\approx 10^6$ cells

Few morphological traits



Ugly Little Bacteria

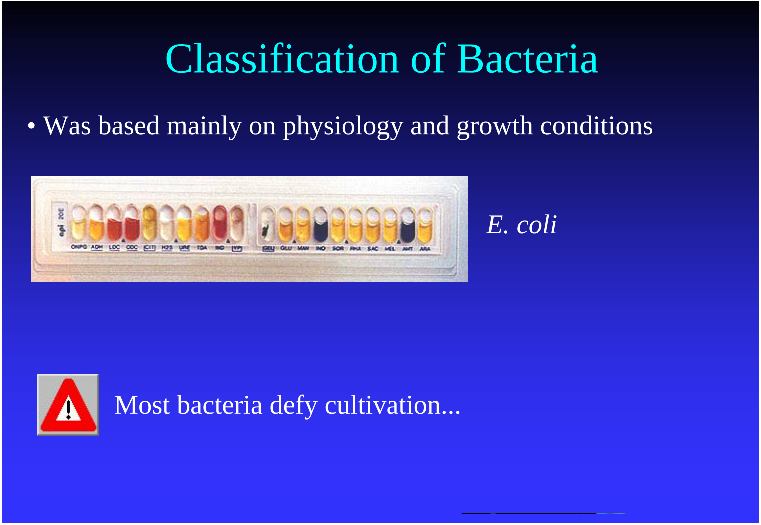
10^7 years

10^9 years

E. coli

B. subtilis

Bacterial classification



Classification of Bacteria

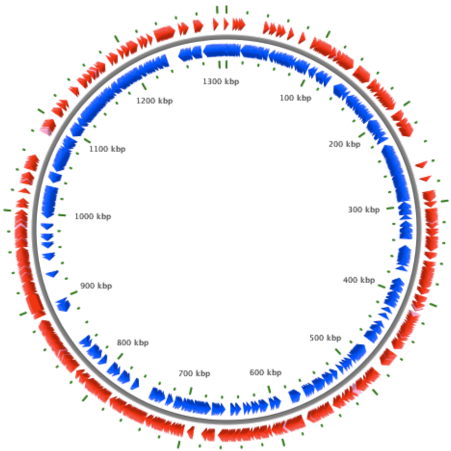
- Was based mainly on physiology and growth conditions

E. coli

Most bacteria defy cultivation...

Bacterial classification: "*Candidatus* Pelagibacter ubique"

Candidatus Pelagibacter ubique HTCC1062, complete genome



Bacterial 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. Chirochores
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 ≈ 0.33 nm

- 1 kb ≈ 0.33 μm
- 1 Mb ≈ 0.33 mm
- 1 Gb ≈ 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 $\approx$ 1 pg

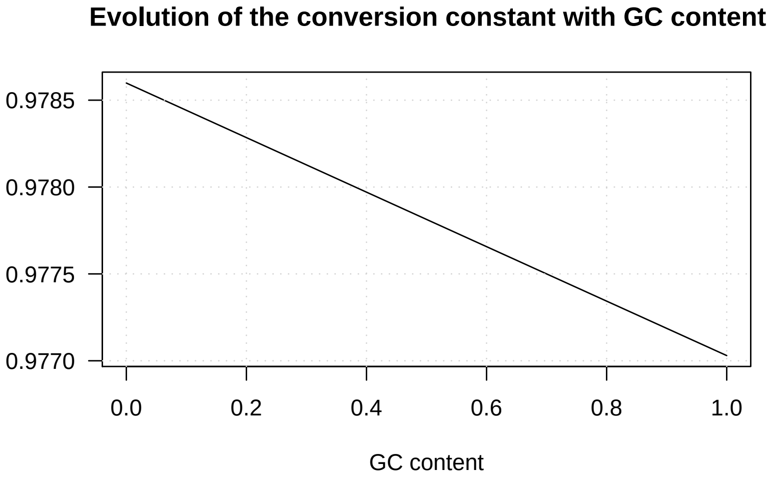
Bacterial genomes are typically in the 10^{-3} pg range (femtogram).

Mass conversion constant and G+C content

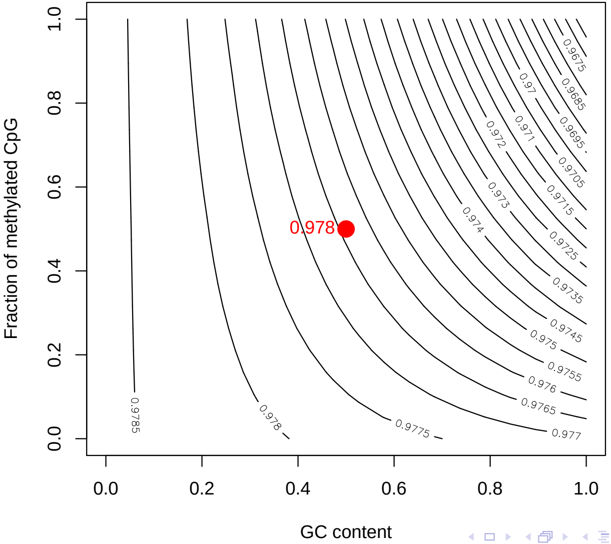
Base	Nucleotide	Chemical formula
A	2’-deoxyadenosine 5’-monophosphate	$C_{10}H_{14}N_5O_6P$
T	2’-deoxythymidine 5’-monophosphate	$C_{10}H_{15}N_2O_8P$
G	2’-deoxyguanosine 5’-monophosphate	$C_{10}H_{14}N_5O_7P$
C	2’-deoxycytidine 5’-monophosphate	$C_9H_{14}N_3O_7P$

Table: Chemical formula of the four nucleotides in DNA.

Mass conversion constant and G+C content



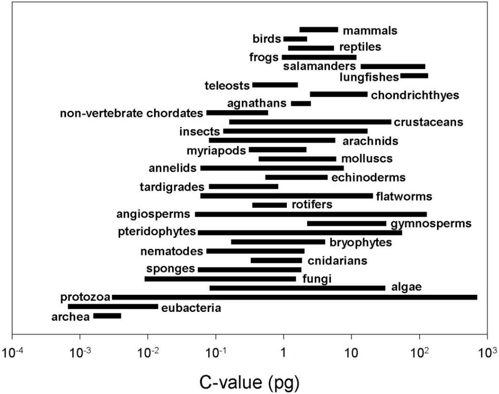
Evolution of the conversion constant with GC content and the fraction of methylated CpG



The big picture

Virus, organelles	"Bacteria"	Eucarya
Tiny genomes (kb)	Small genomes (Mb)	Large genomes (Gb)
High gene density	High gene density	Low gene density

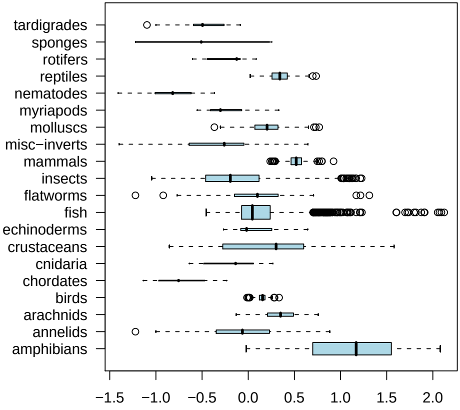
C value paradox



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

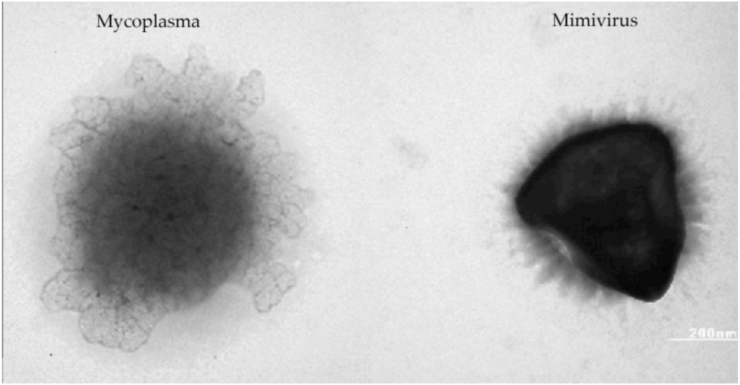
C value paradox

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



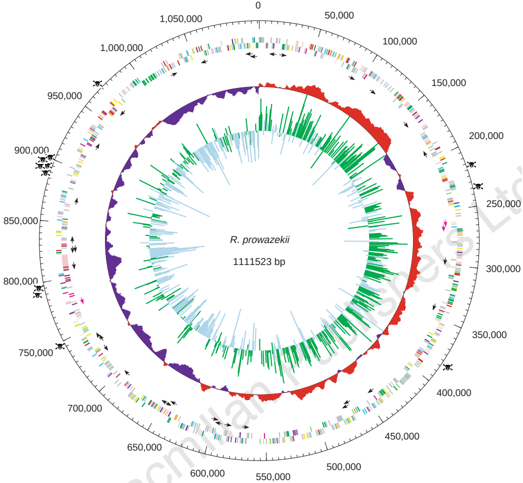
Genome size [pg] (log 10 scale)

Giant virus: mimivirus 1.2 Mb



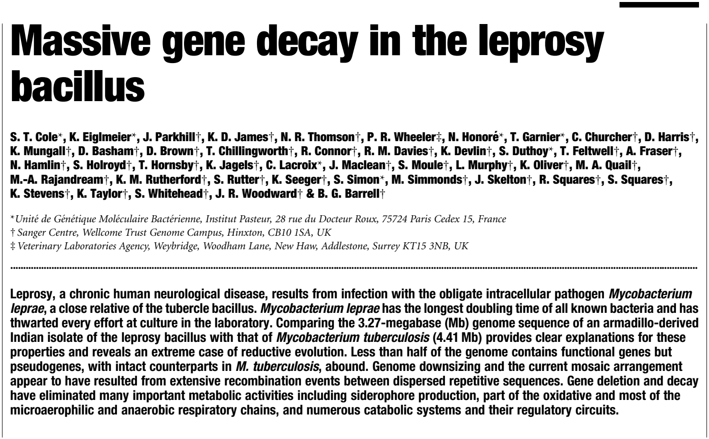
Electronic microscopy of a "bacteria" on the left (*Ureaplasma urealyticum (parvum)*) with a genome size of 0.751 Mb and mimivirus on the rigth with a genome size of 1.181 Mb. Credit: the Mimivirus picture gallery from `http://giantvirus.org/`. Copyright: Prof. Didier Raoult, Rickettsia Laboratory, La 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

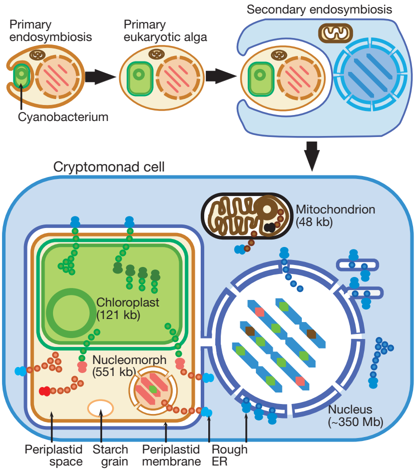
S. T. Cole[*], K. Eiglmeier[*], J. Parkhill[†], K. D. James[†], N. R. Thomson[†], P. R. Wheeler[‡], N. Honoré[*], T. Garnier[*], C. Churcher[†], D. Harris[†], K. Mungall[†], D. Basham[†], D. Brown[†], T. Chillingworth[†], R. Connor[†], R. M. Davies[†], K. Devlin[†], S. Duthoy[*], T. Feltwell[†], A. Fraser[†], 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.

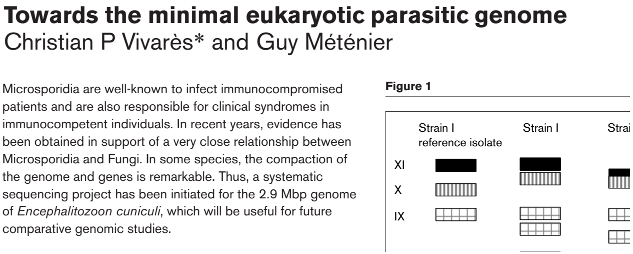
Cole, S.T. *et al* (1998) *Nature*, **409**:1007-10011.

Tiny eucaryal genome: *Guillardia theta* is only 551 kb



Douglas, S. *et al* (2001) *Nature*, **410**:1091-1096.

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

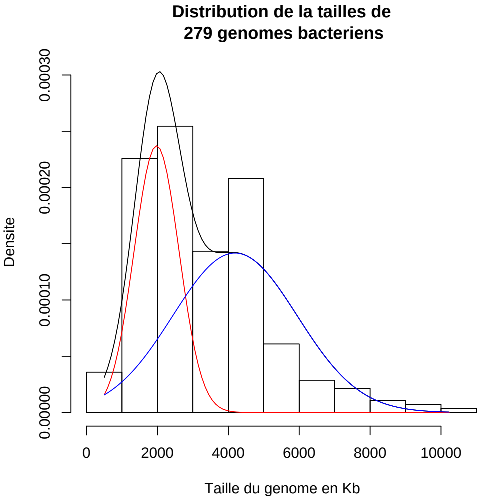
- Eucarya *Saccharomyces cerevisiae* is 12 Mb
- Bacteria *Sorangium cellulosum* is 13 Mb

What is the distribution of bacterial genome size?

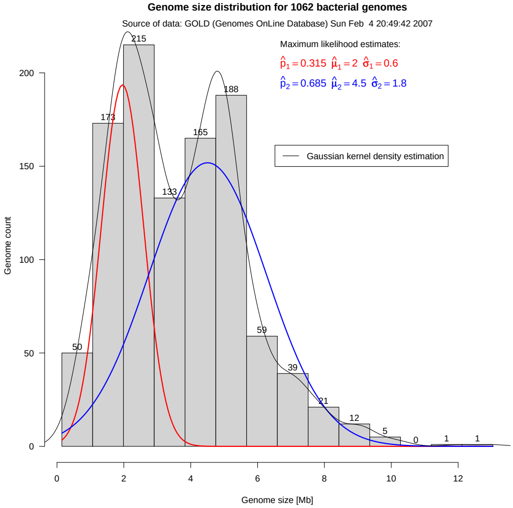
Study this yourself:

`http://pbil.univ-lyon1.fr/R/fichestd/tdr222.pdf`

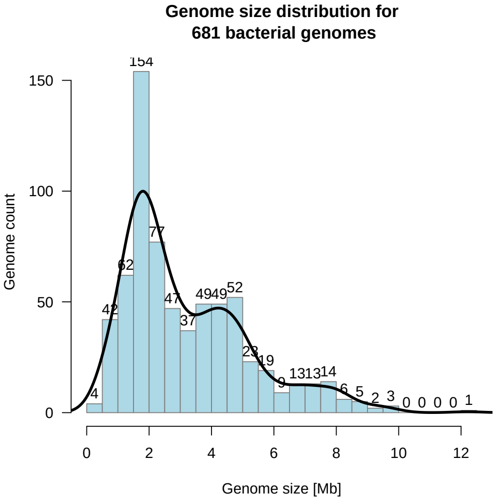
Genome size for 279 bacteria (GOLD 2002)



Genome size for 1062 bacteria (GOLD 2007)



Genome size for 681 bacteria (PFGE data)



Genome size summary

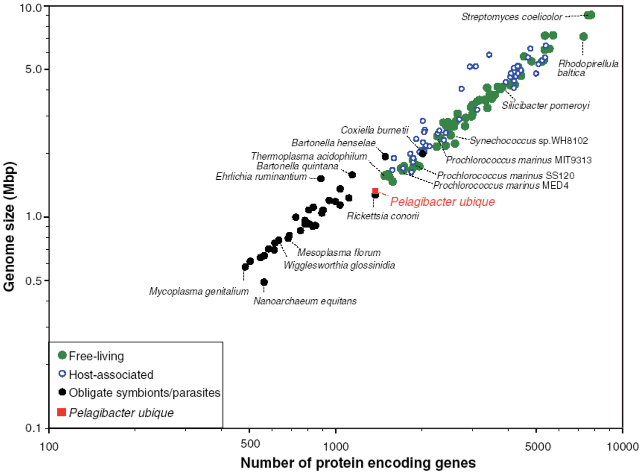
From PFGE data:

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

From complete genome data:

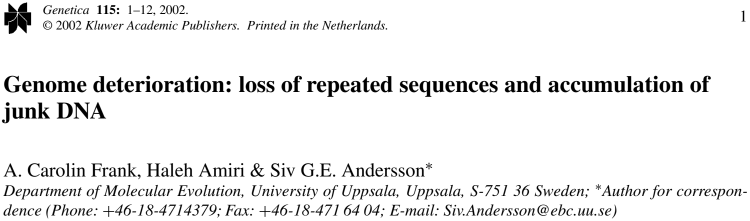
- Range: from 0.146 Mb (*Sulcia muelleri* (Wu, D. *et al.* 2006 *PLoS Biol*,**4**:e188));0.160 Mb (*Carsonella rudii* (Nakabachi, A. *et al.* 2006 *Science*,**314**:267)) to 13.0 Mb.
- Two clear modes at 2 Mb and 4.5 Mb.

Generalists versus specialists



Giovannoni, S.J. *et al* (2005) *Science*, **309**:1242-1245.

Genome size & repeat density



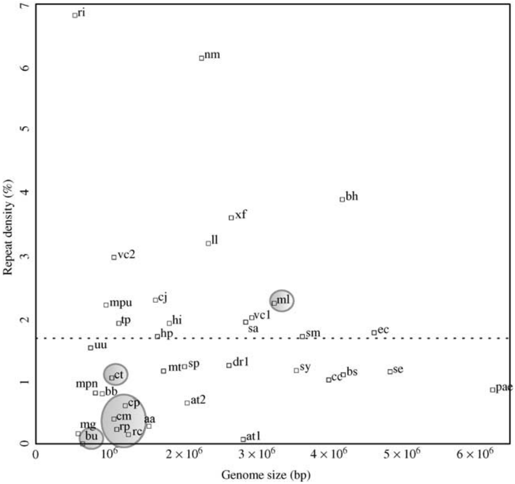
Genetica **115:** 1–12, 2002.
© 2002 *Kluwer Academic Publishers. Printed in the Netherlands.*

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)*

Genome size & repeat density



Genome size & repeat density

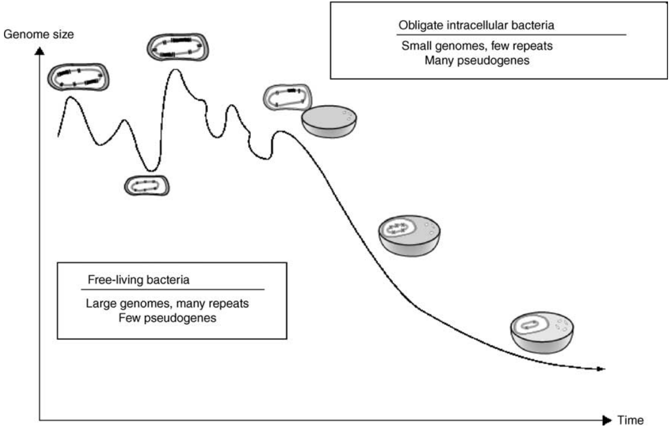
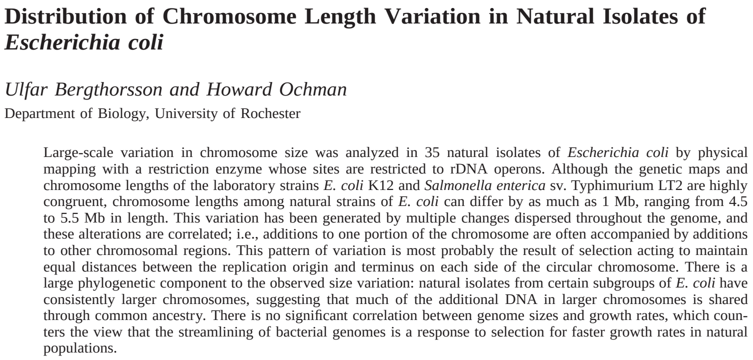

Genome size
Obligate intracellular bacteria
Small genomes, few repeats
Many pseudogenes
Free-living bacteria
Large genomes, many repeats
Few pseudogenes
Time

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.

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.

Bergthorsson, U. and Ochman H. (1998) *Mol. Biol. Evol.*, **15**:6-16.

The ECOR collection

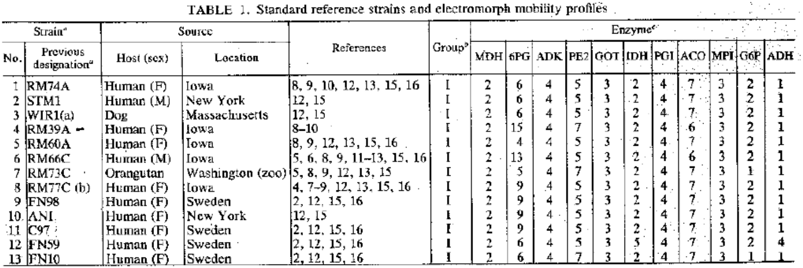


TABLE 1. Standard reference strains and electromorph mobility profiles

Strain[a]		Source		References	Group[b]	Enzyme[c]										
No.	Previous designation[a]	Host (sex)	Location			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	ANI	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.

Digestion of the *E. coli* chromosome with I-*Ceu*I

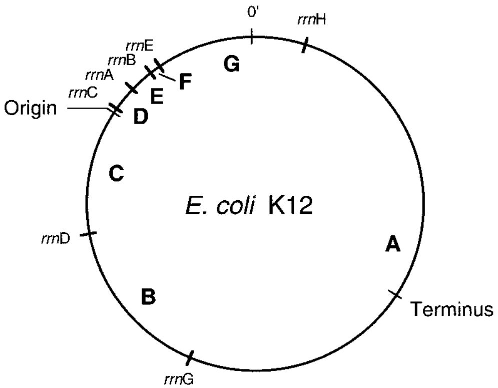


FIG. 1.—Locations of I-*Ceu*I recognition sites on the *E. coli* K12 chromosome. I-*Ceu*I cleaves at the seven *rrn* genes, whose map positions are indicated. The resulting restriction fragments are designated **A** through **G**.

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

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

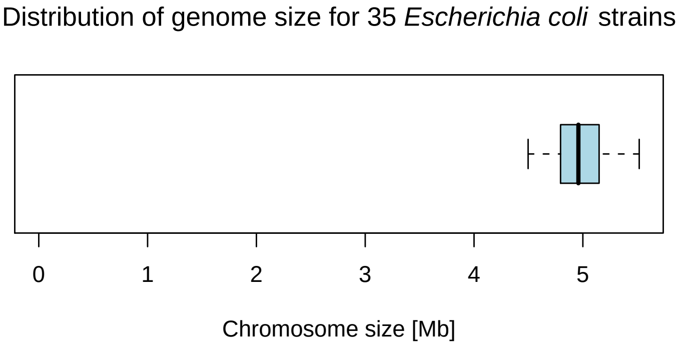
Study this yourself:

```
pgs <- read.table("https://pbil.univ-lyon1.fr/R/donnees/bgs/polygensize
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?

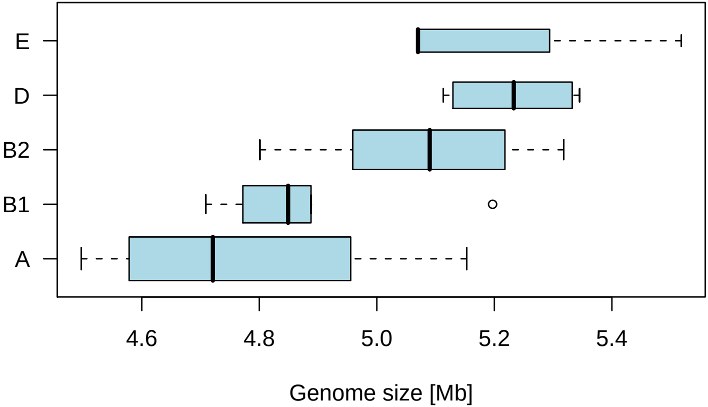
Genome size is highly polymorphic in *E. coli*



There is no meiotic constraints on chromosome length in bacteria.

Genome size phylogenetic inertia

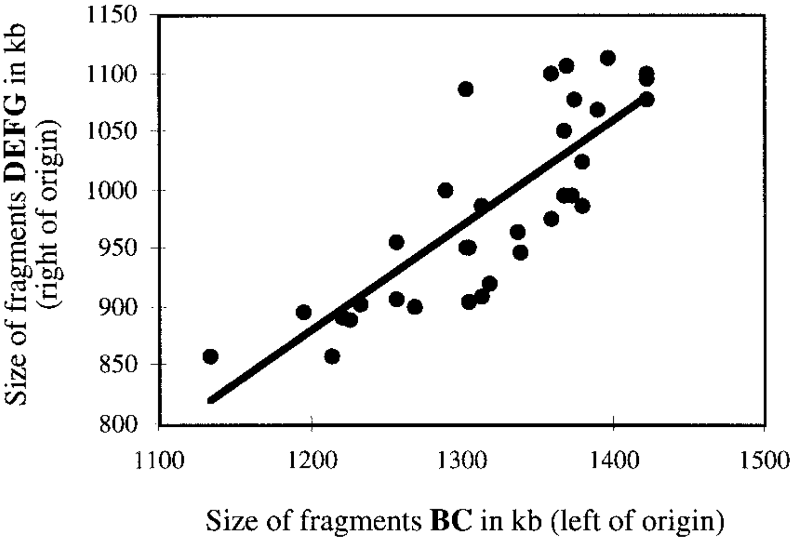
Genome size within 5 subgroups of *Escherichia coli* strains



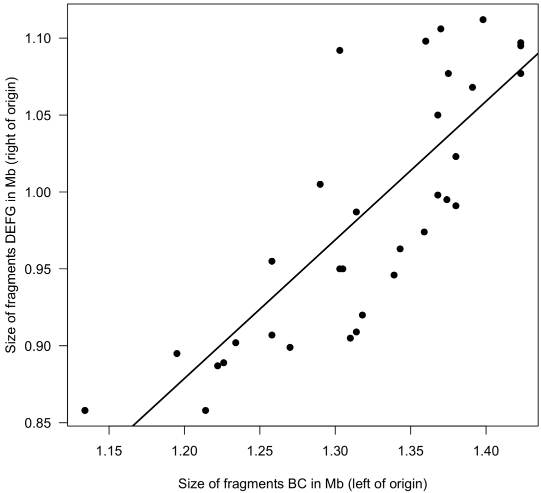
Genome size phylogenetic inertia

```
 tcs <- rowSums(pgs[,5:11])/1000
 options(show.signif.stars = FALSE)
 anova(lm(tcs~pgs$subgroup))
Analysis of Variance Table
Response: tcs
             Df Sum Sq  Mean Sq F value    Pr(>F)
pgs$subgroup  4 1.0756 0.268891   6.817 0.0004999
Residuals    30 1.1833 0.039444
```

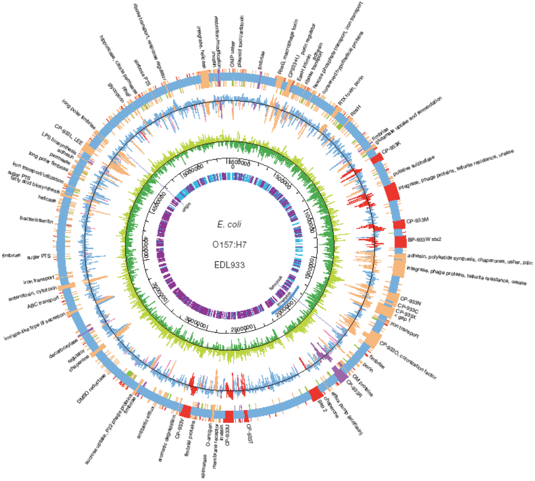
The nice hidden structure



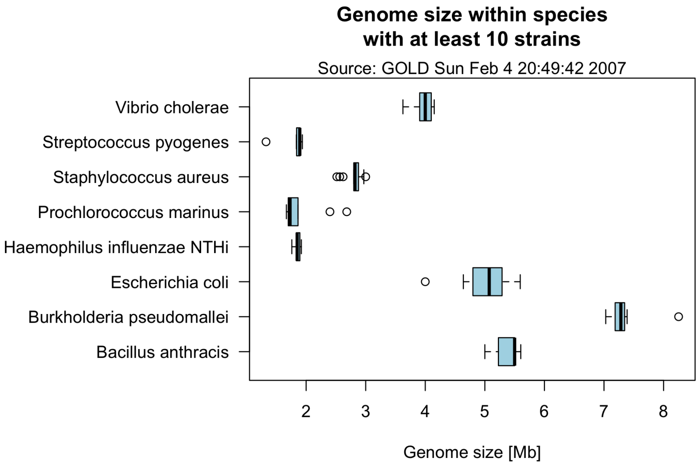
The nice hidden structure



0157:H7 *vs* EDL933



Genome size polymorphism in bacteria

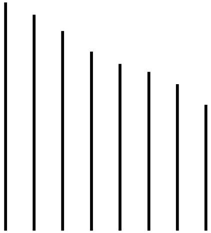


Topology

1. Introduction
2. Genome size
3. Topology & #
4. G+C content
5. Replichores
6. Gene orientation biases
7. Chirochores
8. X-rated structure

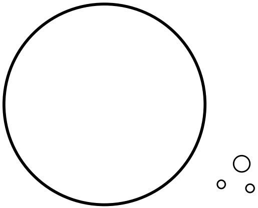
The big picture

Eukaryota



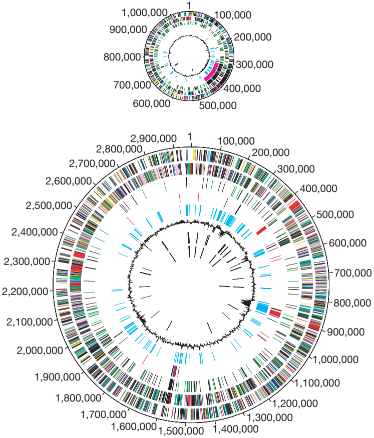
Many linear
chromosomes

Bacteria

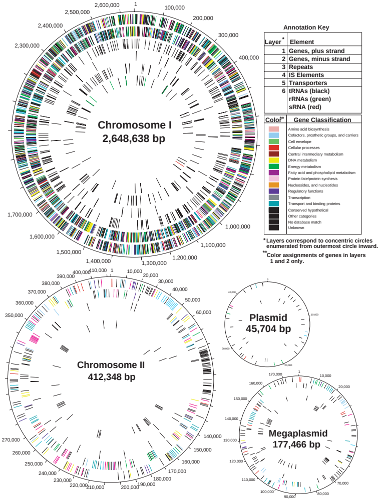


A single circular
chromosome + plasmids

V. cholerae : 2 circular chromosomes



D. radiodurans : 2 circular chromosomes + 2 circular plasmids



Annotation Key

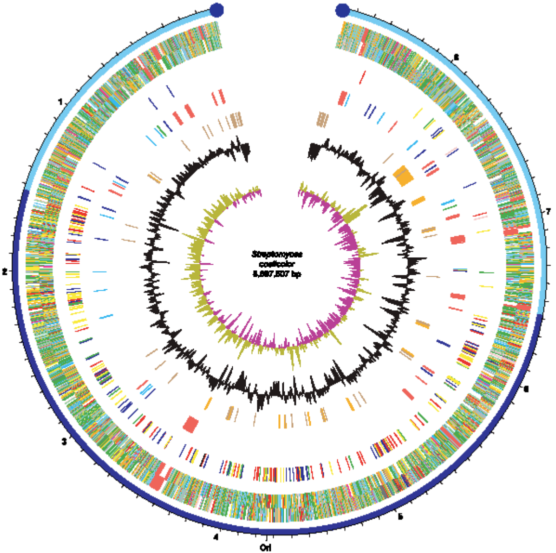
Layer*	Element
1	Genes, plus strand
2	Genes, minus strand
3	Repeats
4	IS Elements
5	Transporters
6	tRNAs (black) rRNAs (green) sRNA (red)

Color**	Gene Classification
🟥	Amino acid biosynthesis
🟦	Cofactors, prosthetic groups, and carriers
🟩	Cell envelope
🟥	Cellular processes
🟫	Central intermediary metabolism
🟨	DNA metabolism
🟩	Energy metabolism
🟪	Fatty acid and phospholipid metabolism
🟥	Protein fate/protein synthesis
🟧	Nucleosides, and nucleotides
🟪	Regulatory functions
⬜	Transcription
🟦	Transport and binding proteins
⬛	Conserved hypothetical
⬛	Other categories
⬛	No database match
⬛	Unknown

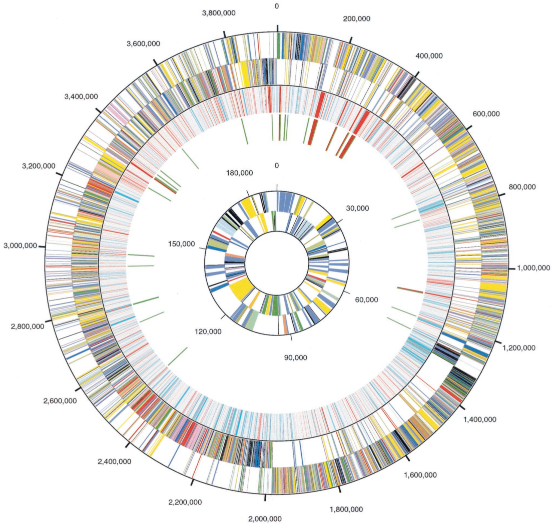
*Layers correspond to concentric circles enumerated from outermost circle inward.

**Color assignments of genes in layers 1 and 2 only.

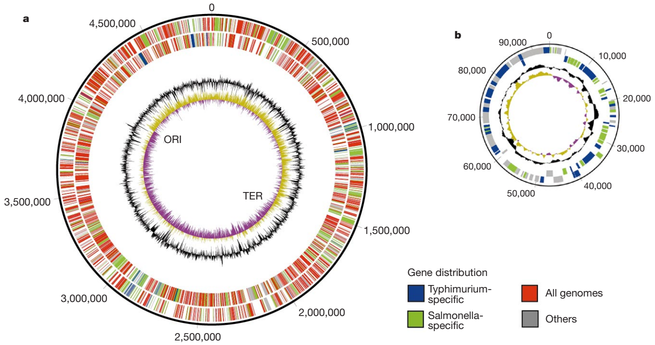
S. coelicolor : 1 linear chromosome



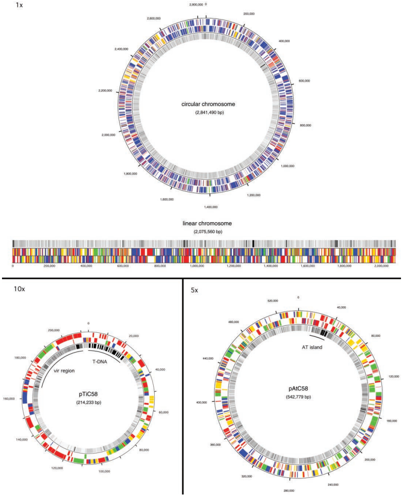
C. acetobutyliticum : 1 circular chromosome + 1 circular megaplasmid



S. enterica serovar Typhimurium LT2 : 1 circular chromosome + 1 circular plasmid



A. tumefaciens : 1 circular chromosome + 1 linear chromosome + 2 circular plasmid



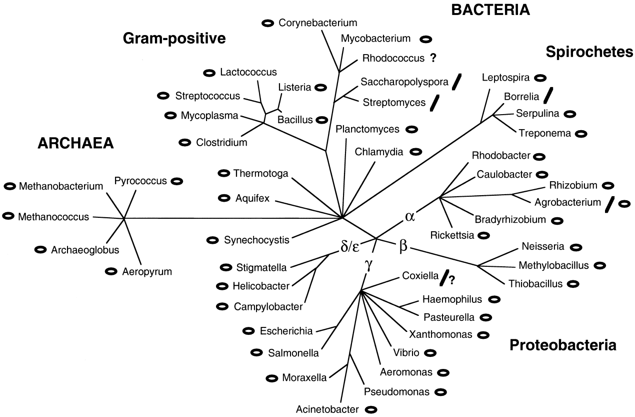
B. burgdorferi : many things !

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

144 J.-N. Volff, J. Altenbuchner / FEMS Microbiology Letters 186 (2000) 143–150



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.........
```

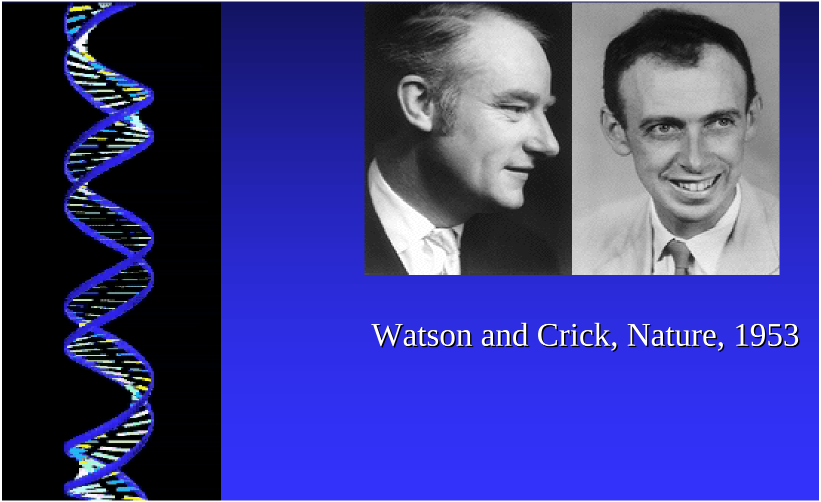
G+C content

1. Introduction
2. Genome size
3. Topology & #
4. G+C content
5. Replichores
6. Gene orientation biases
7. Chirochores
8. X-rated structure

G+C content

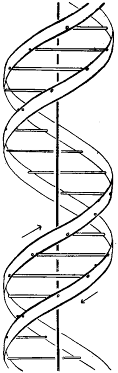
- Is calculated in percentage of G+C : $100\frac{[G+C]}{[A+T+C+G]}$
- First nucleic acid technology applied to bactrerial systematics
- One of the genomic characteritics recommended for the description of species and genera
- 5% and 10% are the common range found within a species and a genera, respectively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

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

$$\mathrm{A_aC_cG_gT_t}$$

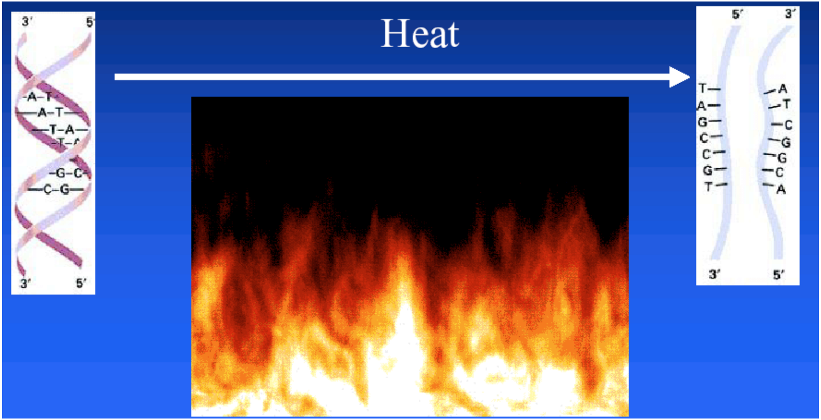
the primary formula on one strand. Then, its complementary strand composition is given by :

$$\mathrm{A_tC_gG_cT_a}$$

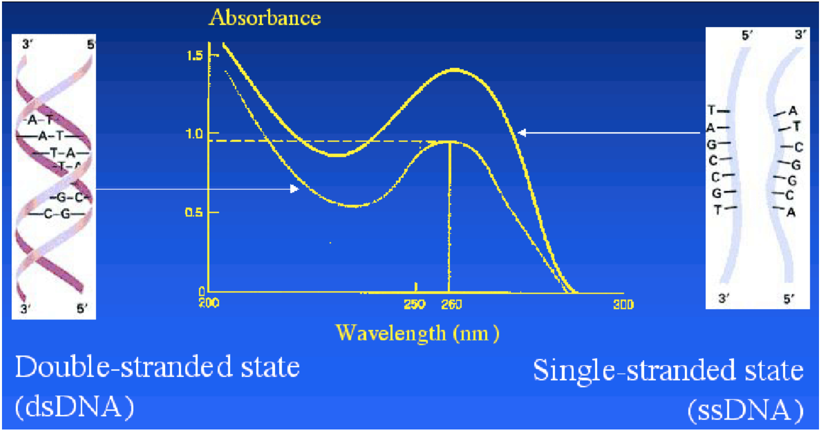
The G+C content is not affected :

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

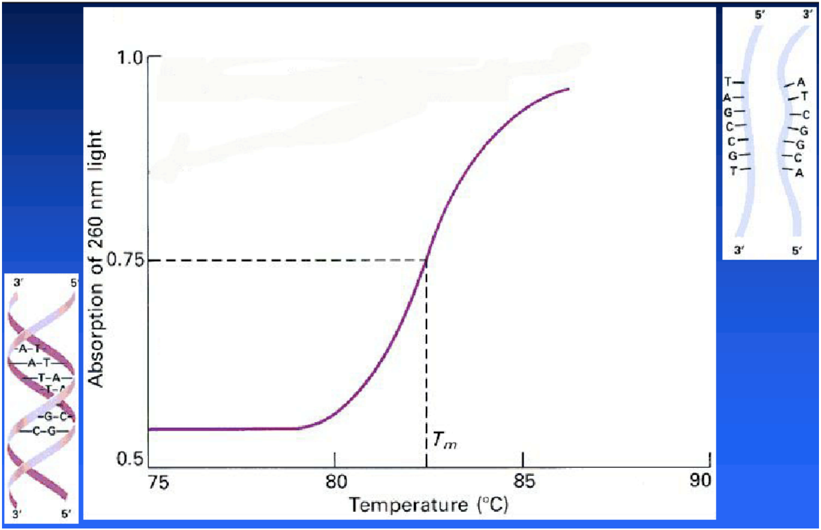
DNA denaturation



ssDNA & dsDNA absorbance

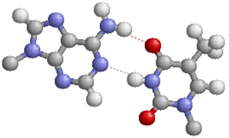


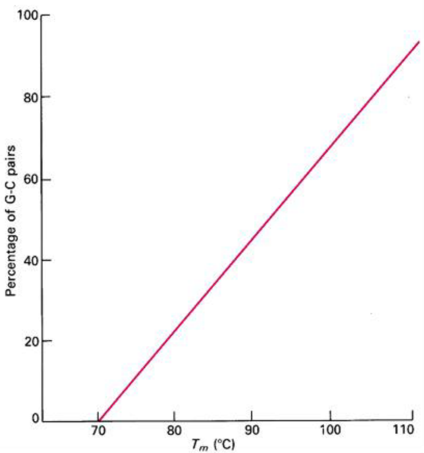
T_m is the temperature at midpoint of transition

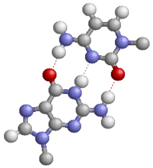


T_m increases with DNA G+C content

AT pairs:
2 hydrogen bonds







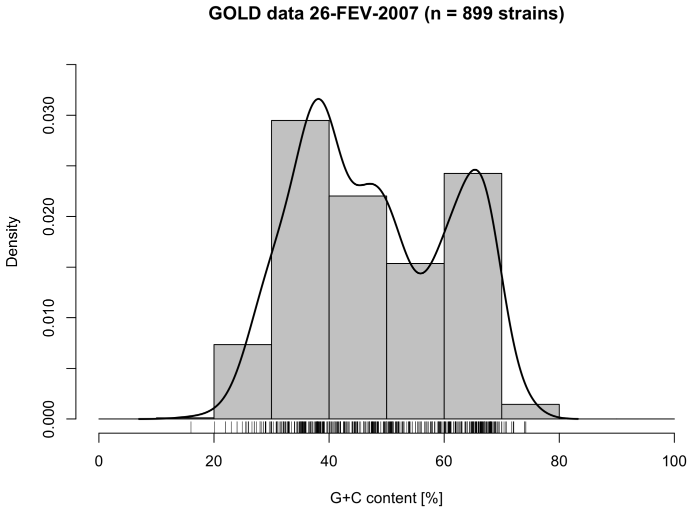
GC pairs:
3 hydrogen bonds

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

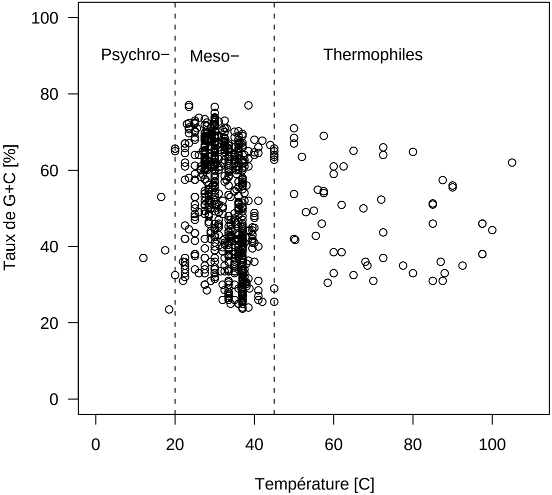
Study this yourself:

1. extract G+C data from GOLD `http://www.genomesonline.org/` and study its distribution.
2. study data from `http://pbil.univ-lyon1.fr/R/donnees/gctopt.RData`. What is the relationship with optimum growth temperature $\mathrm{T}_{\mathrm{opt}}$?

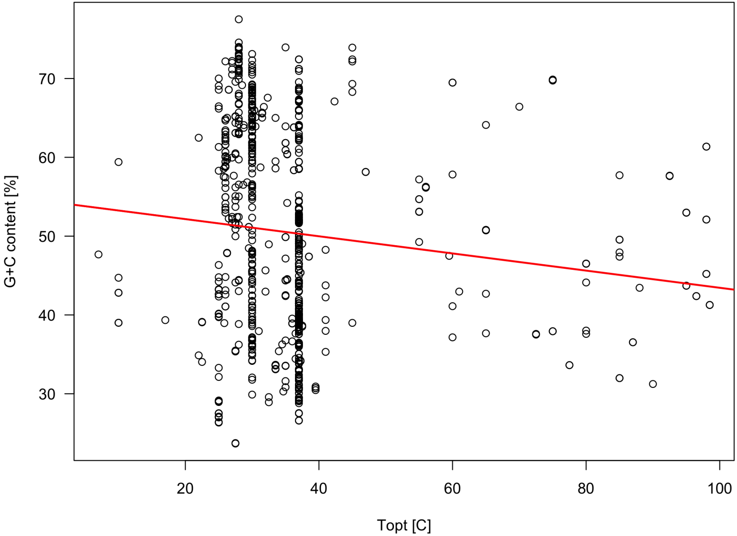
G+C content from GOLD



G+C content and T_{opt}



G+C content and T_{opt} (n = 739)



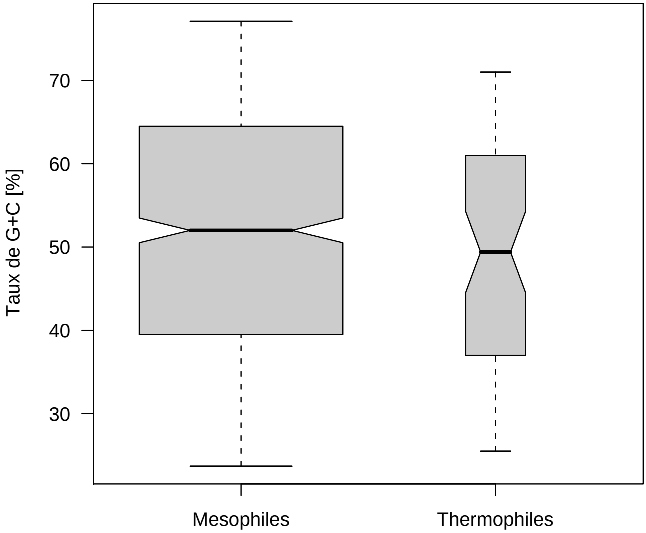
G+C content and T_{opt} ($n = 739$)

```
 shapiro.test(gctopt$topt)
        Shapiro-Wilk normality test
data:  gctopt$topt
W = 0.6389, p-value < 2.2e-16
 shapiro.test(gctopt$gc)
        Shapiro-Wilk normality test
data:  gctopt$gc
W = 0.957, p-value = 6.938e-14
```

G+C content and $\mathrm{T}_{\mathrm{opt}}$ $(n = 739)$

```
 cor.test(gctopt$topt, gctopt$gc, method = "spearman", alternative = "le
        Spearman's rank correlation rho
data:  gctopt$topt and gctopt$gc
S = 82714000, p-value = 1.321e-10
alternative hypothesis: true rho is less than 0
sample estimates:
       rho
-0.2297003
 cor.test(jitter(gctopt$topt), jitter(gctopt$gc), method = "spearman", a
        Spearman's rank correlation rho
data:  jitter(gctopt$topt) and jitter(gctopt$gc)
S = 82694000, p-value = 1.573e-10
alternative hypothesis: true rho is less than 0
sample estimates:
       rho
-0.2294015
```

G+C content and $\mathrm{T_{opt}}$

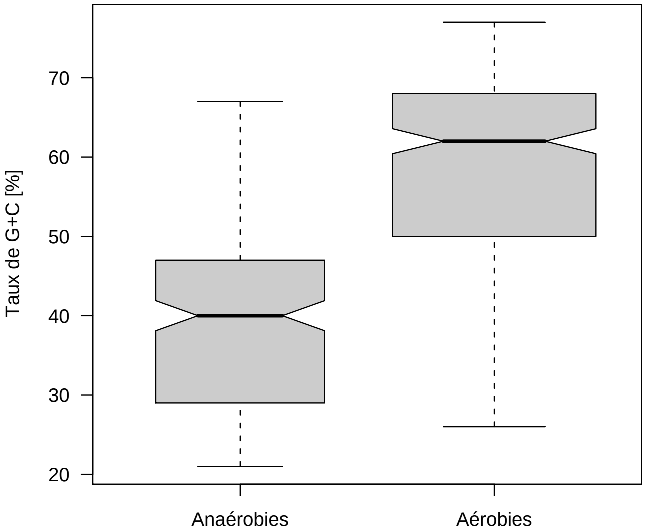


G+C content and aerobiosis

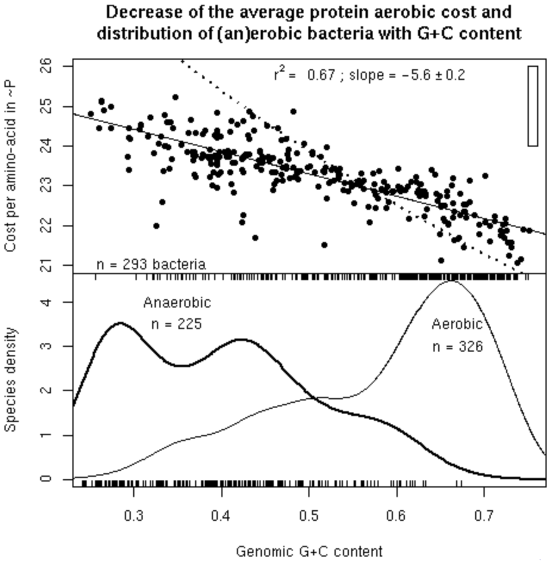
Study this yourself:

1. study data from `http://pbil.univ-lyon1.fr/R/donnees/gc02.txt`. What is the relationship with (an)aerobiosis ?
2. study data from figure 3 at `http://pbil.univ-lyon1.fr/members/lobry/repro/lncs04/`.

G+C content and aerobiosis



G+C content and aerobiosis



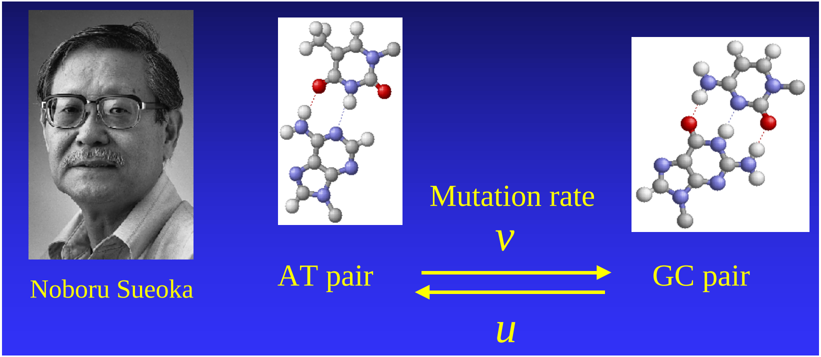
The distribution of G+C content in ”bacteria”

Results from your study:

1. G+C content ranges from $\approx$ 25% to 75% in ”bacteria”.
2. G+C content is not correlated with $\mathrm{T}_{\mathrm{opt}}$.
3. G+C content is correlated with aerobiosis : aerobic ”bacetria” have a significant higher G+C than anerobic ”bacteria”.

Underlying mechanism

Symmetric Directional Mutation Pressure :



Noboru Sueoka

AT pair

Mutation rate

v

$$\text{AT pair} \underset{u}{\overset{v}{\rightleftarrows}} \text{GC pair}$$

u

GC pair

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

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.

Underlying mechanism

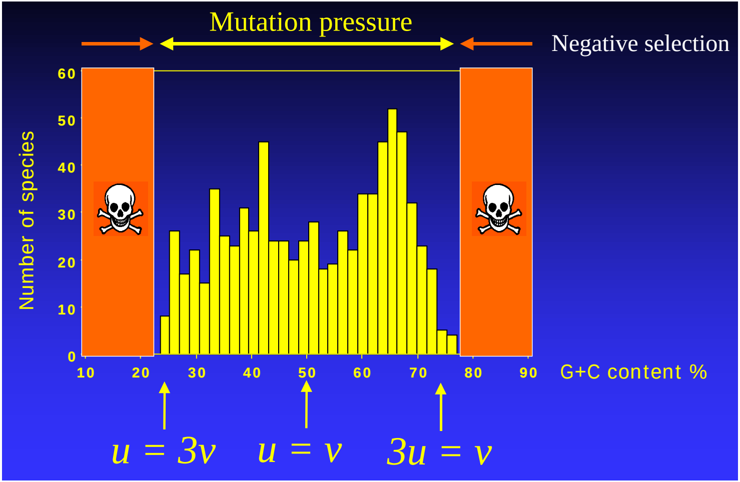
$$\frac{d\theta}{dt} = v(1-\theta) - u\theta$$

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

G+C content at equilibrium :

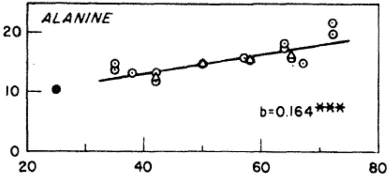
$$\theta^{\star} = \theta(+\infty) = \frac{v}{u+v}$$

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.

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 ?

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:
$P(GAA|\mathrm{not} - \mathrm{stop}) = \frac{P(GAA)}{P(\mathrm{not}-\mathrm{stop})} = \frac{P_G P_A P_A}{1-(P_T P_A P_A + P_T P_A P_G + P_T P_G P_A)}$

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

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

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

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

Study this yourself:

```
 load(url("https://pbil.univ-lyon1.fr/R/donnees/bgs/uco739.RData"))
 dim(uco739)
[1] 739  61
 uco739[1:5,1:5]
                                 aaa  aac  aag  aat  aca
ACHROMOBACTER DENITRIFICANS      216  417  691  149  134
ACHROMOBACTER XYLOSOXIDANS       349  807 1225  283  169
ACIDIANUS AMBIVALENS             756  330  625  519  301
ACIDITHIOBACILLUS FERROOXIDANS   732  897 1252  662  270
ACINETOBACTER BAUMANNII         3442 1233 1429 2590 1271
```

This is a dataset of codon counts in 739 bacterial species.

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

Compute the G+C content from codon counts. From `colnames(uco739)`, make a vector of G+C content in each codon:

```
aaa aac aag aat aca acc acg act aga agc agg agt ata atc atg att caa cac
  0   1   1   0   1   2   2   1   1   2   2   1   0   1   1   0   1   2
ccc ccg cct cga cgc cgg cgt cta ctc ctg ctt gaa gac gag gat gca gcc gcg
  3   3   2   2   3   3   2   1   2   2   1   1   2   2   1   2   3   3
ggg ggt gta gtc gtg gtt tac tat tca tcc tcg tct tgc tgg tgt tta ttc ttg
  3   2   1   2   2   1   1   0   1   2   2   1   2   2   1   0   1   1
```

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

Now, thanks to matrix multiplication (%*%), in one line compute the G+C content in percent:

```
                                   [,1]
ACHROMOBACTER DENITRIFICANS    61.88166
ACHROMOBACTER XYLOSOXIDANS     63.37462
ACIDIANUS AMBIVALENS           37.59371
ACIDITHIOBACILLUS FERROOXIDANS 58.72497
ACINETOBACTER BAUMANNII        43.92444
ACINETOBACTER CALCOACETICUS    42.96045
```

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

Compute the amino-acid content from codon counts. From `colnames(uco739)`, make a vector of the corresponding amino-acid:

```
  aaa   aac   aag   aat   aca   acc   acg   act   aga   agc   agg   agt
"Lys" "Asn" "Lys" "Asn" "Thr" "Thr" "Thr" "Thr" "Arg" "Ser" "Arg" "Ser"
  atg   att   caa   cac   cag   cat   cca   ccc   ccg   cct   cga   cgc
"Met" "Ile" "Gln" "His" "Gln" "His" "Pro" "Pro" "Pro" "Pro" "Arg" "Arg"
  cta   ctc   ctg   ctt   gaa   gac   gag   gat   gca   gcc   gcg   gct
"Leu" "Leu" "Leu" "Leu" "Glu" "Asp" "Glu" "Asp" "Ala" "Ala" "Ala" "Ala"
  ggg   ggt   gta   gtc   gtg   gtt   tac   tat   tca   tcc   tcg   tct
"Gly" "Gly" "Val" "Val" "Val" "Val" "Tyr" "Tyr" "Ser" "Ser" "Ser" "Ser"
  tgt   tta   ttc   ttg   ttt
"Cys" "Leu" "Phe" "Leu" "Phe"
```

Useful functions are `s2c()`, `translate()` and `aaa()` in the seqinr package.

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

Now, in one line, thanks to `apply()` and `tapply()`, compute the amino-acid content in each proteome:

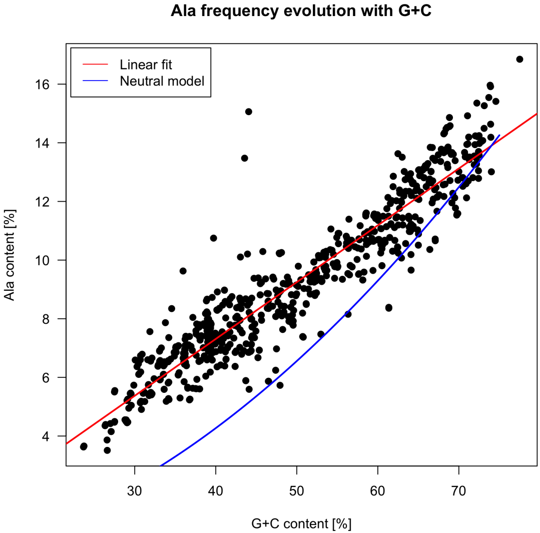
```
                               Ala  Arg  Asn  Asp Cys
ACHROMOBACTER DENITRIFICANS   2783 1601  566 1204 206
ACHROMOBACTER XYLOSOXIDANS    5031 2828 1090 2062 330
ACIDIANUS AMBIVALENS          1297  700  849  853 218
ACIDITHIOBACILLUS FERROOXIDANS 5876 3794 1559 2705 623
ACINETOBACTER BAUMANNII       7428 4257 3823 4476 745
```

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

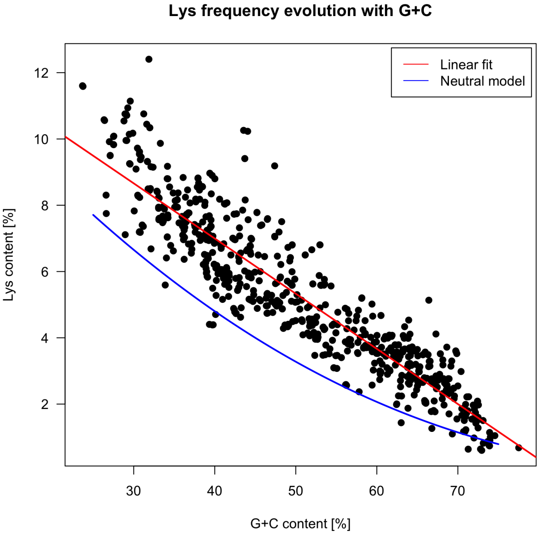
Plot the results :

- Show the influence of G+C content for Ala, Lys, and Glu (at least).
- Add the linear fit
- Add the neutral model as a line.

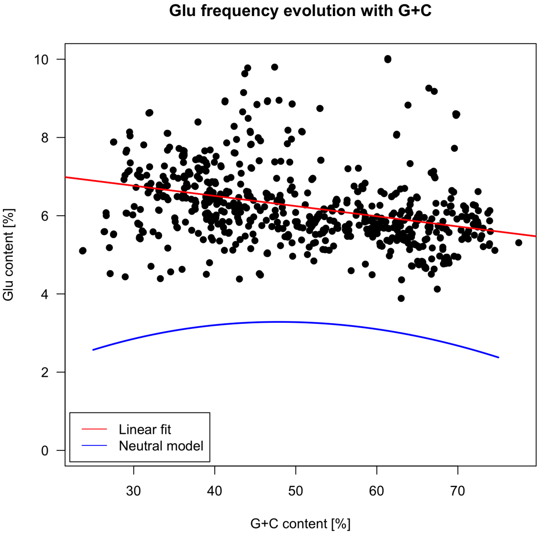
What should be obtained for Ala:



What should be obtained for Lys:



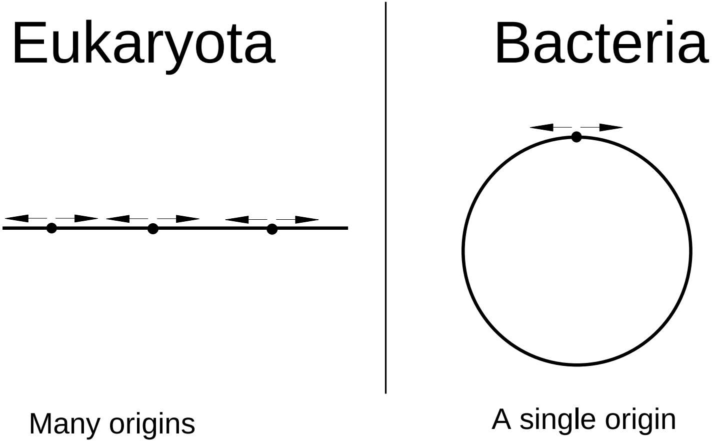
What should be obtained for Glu:



Replichores

1. Introduction
2. Genome size
3. Topology & #
4. G+C content
5. Replichores
6. Gene orientation biases
7. Chirochores
8. X-rated structure

Replichores: origin and terminus of replication



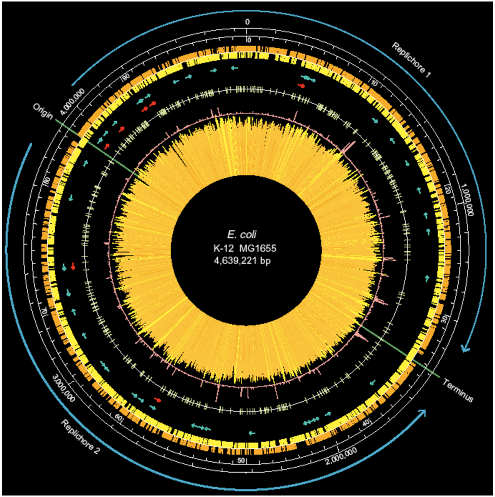
Eukaryota

Bacteria

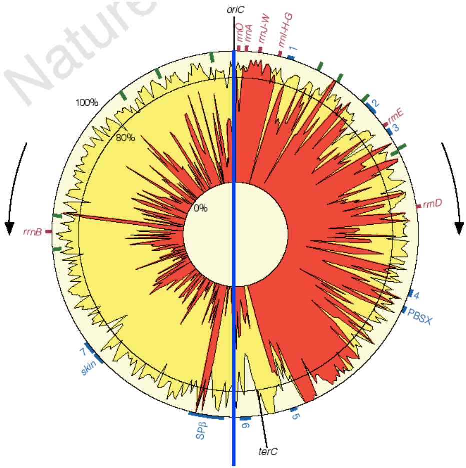
Many origins

A single origin

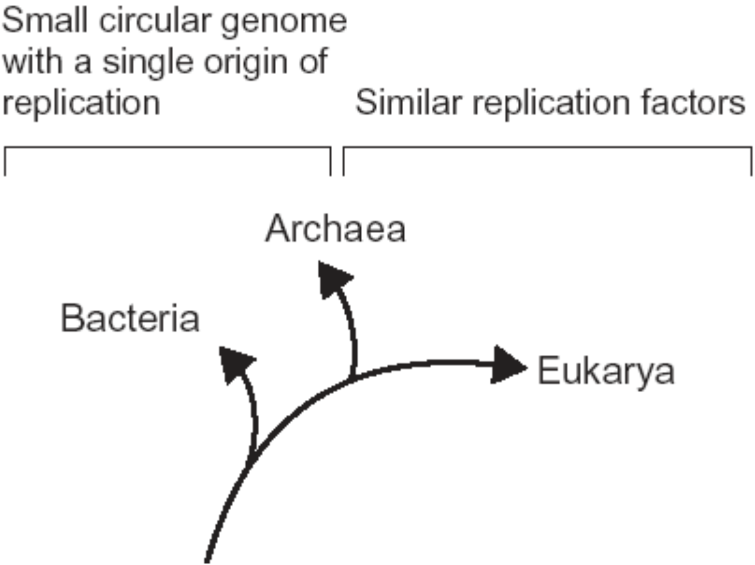
There are two replichores per chromosome



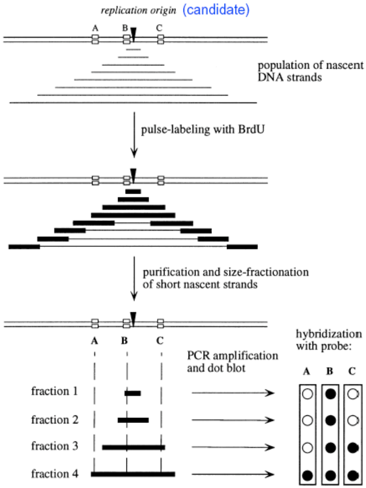
$\approx \pi$


Nature
oriC
rrnO
rrnA
rrnJ-W
rrnI-H-G
1
2
rrnE
3
rrnD
4
PBSX
5
6
terC
SPβ
skin
7
rrnB
100%
80%
0%

Archae & Bacteria

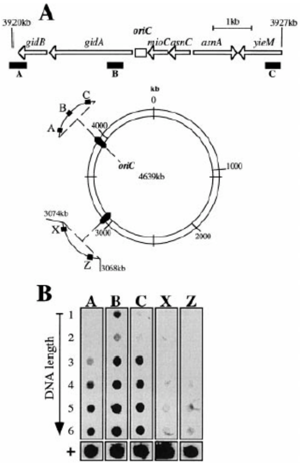

Small circular genome with a single origin of replication
Similar replication factors
Archaea
Bacteria
Eukarya

Looking for the origin

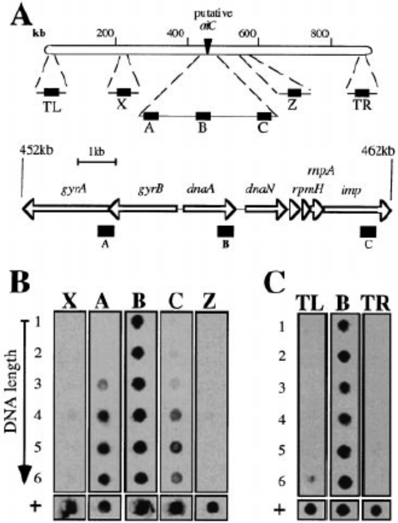

replication origin (candidate)
A
B
C
population of nascent DNA strands
pulse-labeling with BrdU
purification and size-fractionation of short nascent strands
A
B
C
PCR amplification and dot blot
hybridization with probe:
fraction 1
fraction 2
fraction 3
fraction 4

	A	B	C
fraction 1	○	●	○
fraction 2	○	●	○
fraction 3	○	●	●
fraction 4	●	●	●

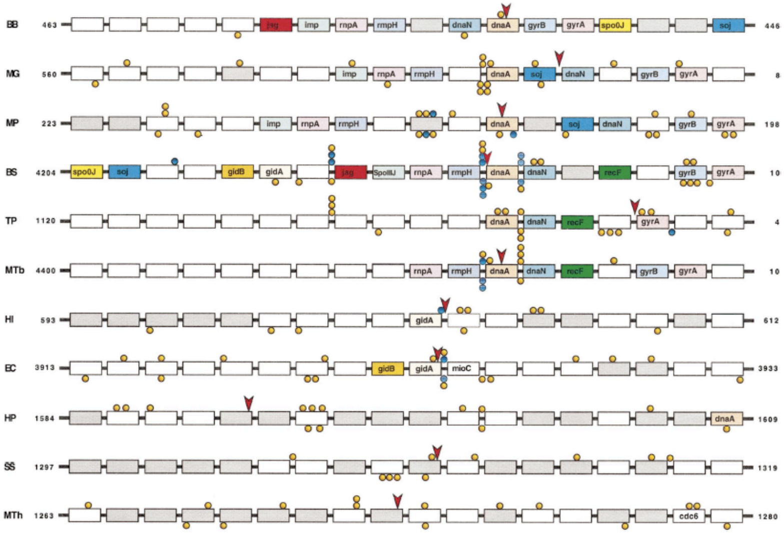
Looking for the origin in *E. coli*



Looking for the origin in *B. burgdorferi*



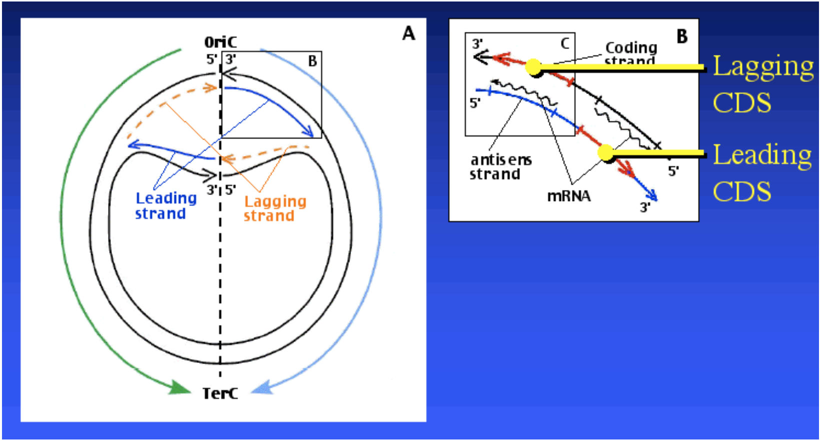
Zoom at the origin


BB 463 jag imp rnpA rmpH dnaN dnaA gyrB gyrA spo0J soj 446
MG 560 imp rnpA rmpH dnaA soj dnaN gyrB gyrA 8
MP 223 imp rnpA rmpH dnaA soj dnaN gyrB gyrA 198
BS 4204 spo0J soj gidB gidA jag SpoIIJ rnpA rmpH dnaA dnaN recF gyrB gyrA 10
TP 1120 dnaA dnaN recF gyrA 4
MTb 4400 rnpA rmpH dnaA dnaN recF gyrB gyrA 10
HI 593 gidA 612
EC 3913 gidB gidA mioC 3933
HP 1584 dnaA 1609
SS 1297 1319
MTh 1263 cdc6 1280

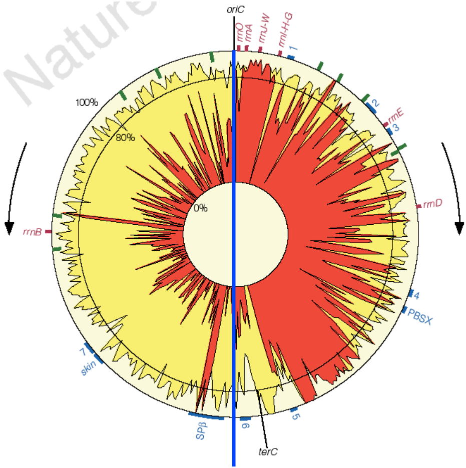
Gene orientation biases

1. Introduction
2. Genome size
3. Topology & #
4. G+C content
5. Replichores
6. Gene orientation biases
7. Chirochores
8. X-rated structure

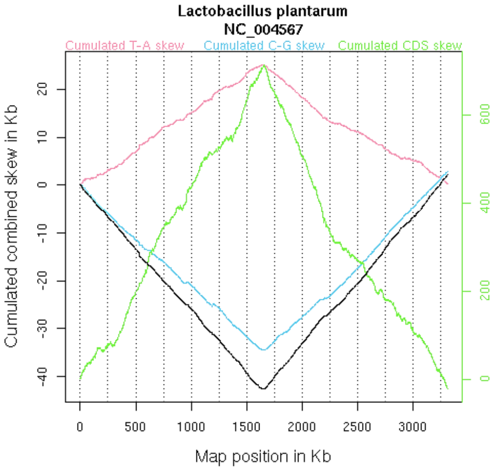
Leading and lagging CDS



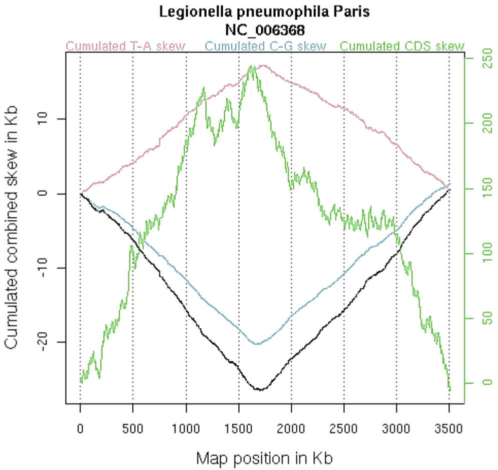
More leading CDS than lagging CDS



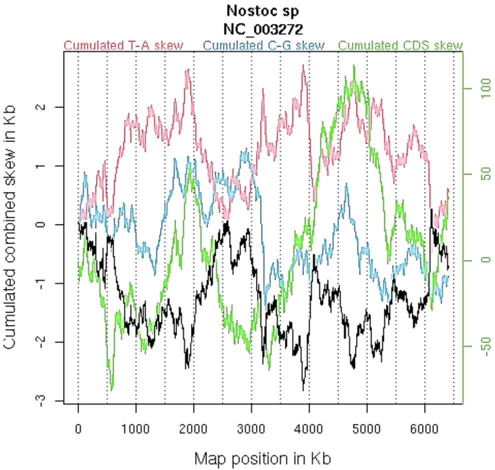
Lactobacillus plantarum



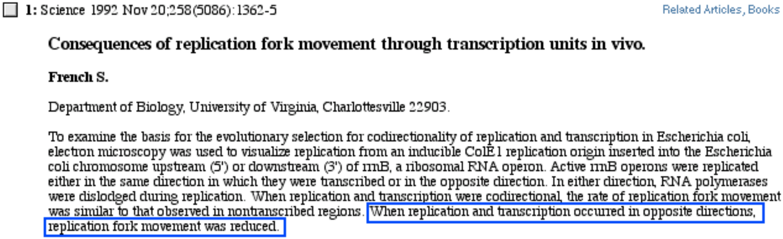
Legionella pneumophila



Nostoc sp


Nostoc sp
NC_003272
Cumulated T-A skew
Cumulated C-G skew
Cumulated CDS skew
Cumulated combined skew in Kb
2
1
0
-1
-2
-3
100
50
0
-50
0
1000
2000
3000
4000
5000
6000
Map position in Kb

Collisions between polymerases



□ **1:** Science 1992 Nov 20;258(5086):1362-5

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

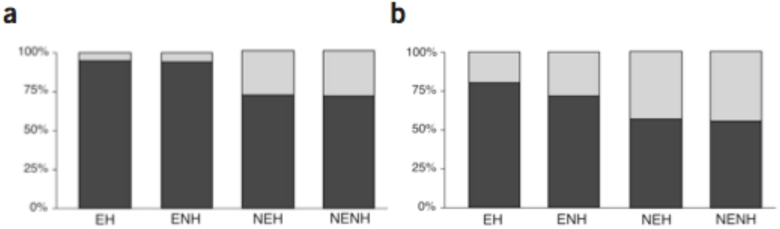


Figure 1 Distribution of genes between the leading (dark gray) and the lagging (light gray) strands of the genome of *B. subtilis* (**a**) and *E. coli* (**b**). H, highly expressed; NH, non-highly expressed; E, essential; NE, non-essential.

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

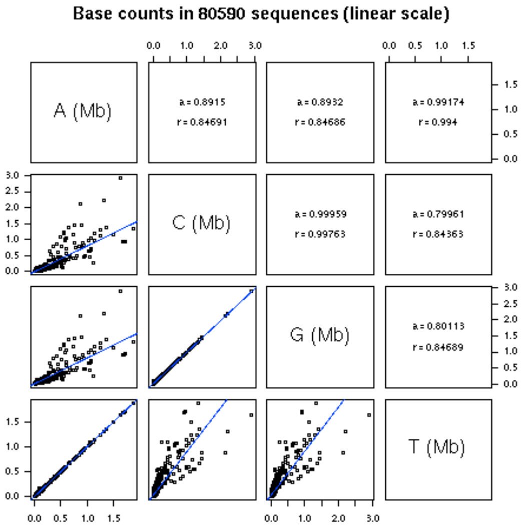
Chrirochores: base composition biases

1. Introduction
2. Genome size
3. Topology & #
4. G+C content
5. Replichores
6. Gene orientation biases
7. Chirochores
8. X-rated structure

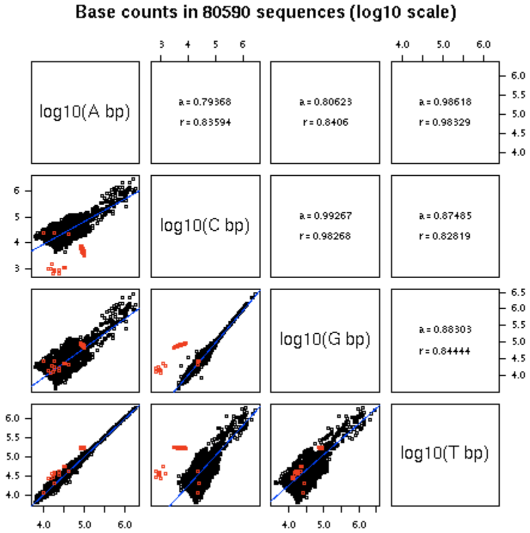
PR2 parity rule number 2

- In double-stranded DNA we have exactly $\mathrm{A} = \mathrm{T}$ and $\mathrm{C} = \mathrm{G}$ as a direct consequence of Watson-Crick base pairing rules.
- More surprisingly, in non artificial single-stranded DNA we have approximately $\mathrm{A} \approx \mathrm{T}$ and $\mathrm{C} \approx \mathrm{G}$.
- Parity rule number 2, or PR2 state, refers to the second assertion.
- The following examples are from the base counts in all ssDNA sequences ($>$ 50 Kb and $<$ 1 % of ambiguous bases) from GenBank (24-NOV-2004).

PR2 illustration (linear scale)



PR2 illustration (log scale, synthetic sequences in red)



PR1 parity rule number 1

- PR1 parity rule number 1 is an hypothesis about the process of evolution of the DNA sequences.
- PR1 hypothesis is that substitution rates are symmetric with respect to the two DNA strands.
- PR1 hypothesis doesn't mean that the substitution matrix itself is symmetric.

PR1 derivation

In the general case, let

$$\mathrm{r}(\mathrm{X} \rightarrow \mathrm{Y})$$

be the substitution rate from basis X to Y on one strand, and

$$\overline{\mathrm{r}}(\overline{\mathrm{X}} \rightarrow \overline{\mathrm{Y}})$$

the substitution rate for the complementary event on the other strand. The apparent substitution rate on one strand is equal to the sum of these two substitution rates:

$$\mathrm{R}(\mathrm{X} \rightarrow \mathrm{Y}) = \mathrm{r}(\mathrm{X} \rightarrow \mathrm{Y}) + \overline{\mathrm{r}}(\overline{\mathrm{X}} \rightarrow \overline{\mathrm{Y}})$$

PR1 derivation

Still in the general case, consider the complementary event:

$$\mathrm{R}(\overline{\mathrm{X}} \to \overline{\mathrm{Y}}) = \mathrm{r}(\overline{\mathrm{X}} \to \overline{\mathrm{Y}}) + \bar{\mathrm{r}}(\overline{\overline{\mathrm{X}}} \to \overline{\overline{\mathrm{Y}}})$$

Since

$$\overline{\overline{\mathrm{N}}} = \mathrm{N}$$

this can be rewritten as

$$\mathrm{R}(\overline{\mathrm{X}} \to \overline{\mathrm{Y}}) = \mathrm{r}(\overline{\mathrm{X}} \to \overline{\mathrm{Y}}) + \bar{\mathrm{r}}(\mathrm{X} \to \mathrm{Y})$$

PR1 derivation

We introduce now PR1 hypothesis:

PR1 hypothesis:

$$\forall X, Y \in N : r(X \to Y) = \bar{r}(X \to Y)$$

In general we had:

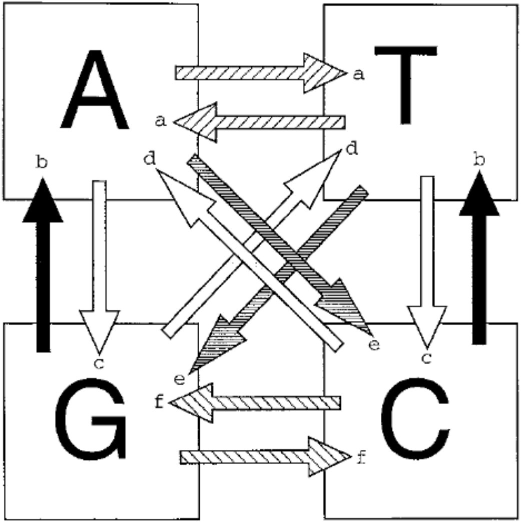
$$R(X \to Y) = r(X \to Y) + \bar{r}(\overline{X} \to \overline{Y})$$

$$R(\overline{X} \to \overline{Y}) = r(\overline{X} \to \overline{Y}) + \bar{r}(X \to Y)$$

So that under PR1 hypothesis we have:

$$R(X \to Y) = R(\overline{X} \to \overline{Y})$$

PR1 graphically



PR1 in matrix notations

$$\mathbf{X} = \begin{pmatrix} A(t) \\ T(t) \\ G(t) \\ C(t) \end{pmatrix}$$

$$\frac{d\mathbf{X}}{dt} = \mathbf{RX}$$

$$\mathbf{R} = \begin{pmatrix} -a-e-c & a & b & d \\ a & -a-e-c & d & b \\ c & e & -b-d-f & f \\ e & c & f & -b-d-f \end{pmatrix}$$

Relationship between PR1 and PR2

- PR2 state is an asymptotic property of systems evolving under PR1 hypothesis.
- This true even for non-autonomous systems $\frac{d\mathbf{X}}{dt} = \mathbf{R}(t)\mathbf{X}$ (*Mol. Biol. Evol.* **16**:719-723).
- If PR2 is not observed for natural ssDNA sequences, PR1 can be rejected safely.

AT and GC skews

The AT skew is the deviation from A = T:

$$\mathrm{AT}_{\mathrm{skew}} = \frac{\mathrm{A} - \mathrm{T}}{\mathrm{A} + \mathrm{T}}$$

The GC skew is the deviation from C = G:

$$\mathrm{GC}_{\mathrm{skew}} = \frac{\mathrm{C} - \mathrm{G}}{\mathrm{C} + \mathrm{G}}$$

Skews are not the same for both strands

Let

$$\mathrm{A_aC_cG_gT_t}$$

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

$$\mathrm{A_tC_gG_cT_a}$$

The AT and GC skews are affected :

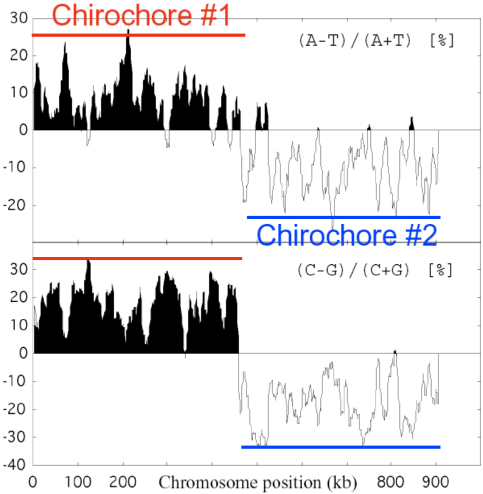
$$\frac{\mathrm{a-t}}{\mathrm{a+t}} = -\frac{\mathrm{t-a}}{\mathrm{t+a}}$$

$$\frac{\mathrm{c-g}}{\mathrm{c+g}} = -\frac{\mathrm{g-c}}{\mathrm{g+c}}$$

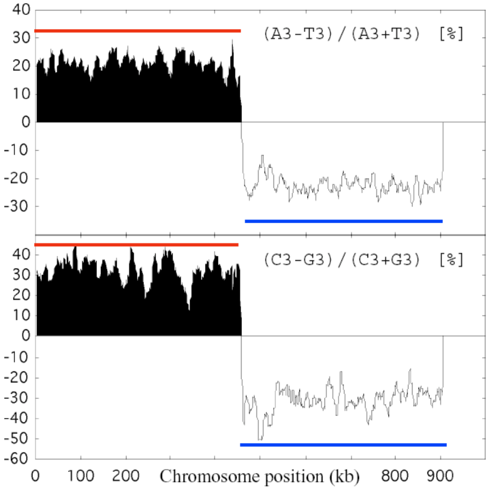
Chirochore: definition

- A chirochore is a segment of ssDNA homogeneous for its deviation from PR2 state.
- A chirochore is therefore characterized by constant AT and GC skews.
- Note the difference with isochores that are characterized by a constant G+C content.

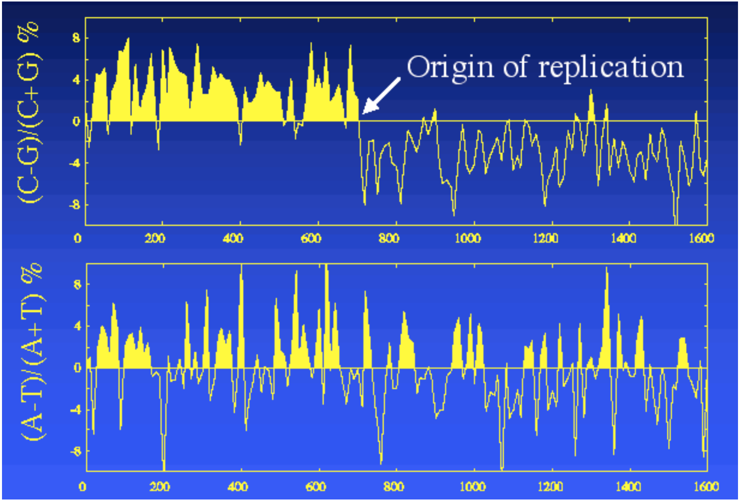
Chirochores in *B. burgdorferi*



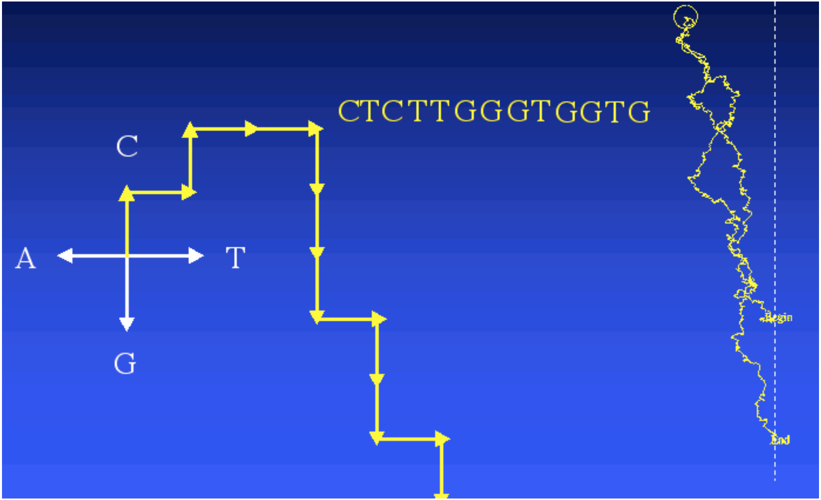
Chirochores in *B. burgdorferi* (third codon positions)



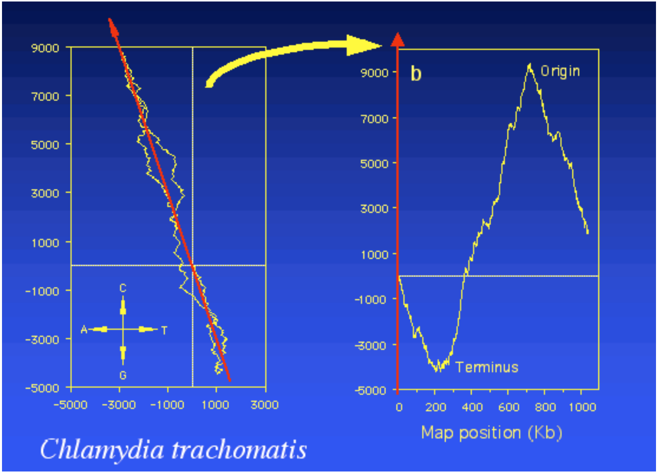
Usually GC skew > AT skew *e.g. E. coli*



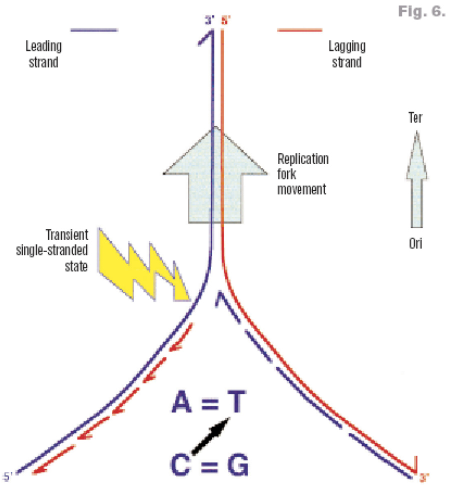
Simple DNA walk *E. coli*



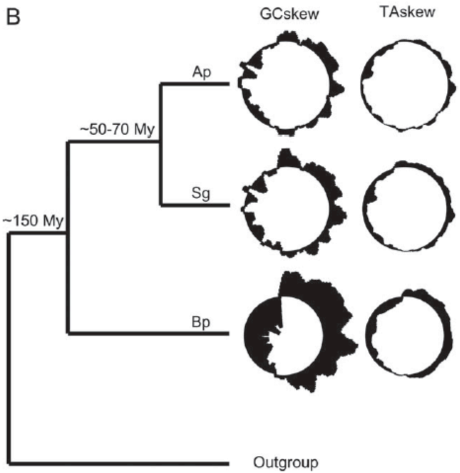
Oriloc


9000
7000
5000
3000
1000
-1000
-3000
-5000
-5000
-3000
-1000
1000
3000
C
A
T
G
Chlamydia trachomatis
b
Origin
Terminus
9000
7000
5000
3000
1000
-1000
-3000
-5000
0
200
400
600
800
1000
Map position (Kb)

Cytosine deamination theory



Buchnera aphidicola



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

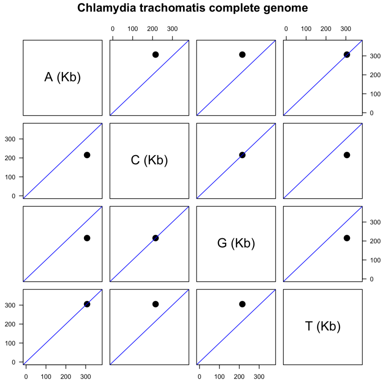
Chirochore practical

Study this yourself with the complete genome from *Chlamydia trachomatis*:

```
 library(seqinr)
 ctf <- system.file("sequences/ct.fasta.gz", package = "seqinr")
 myseq <- read.fasta(ctf)[[1]]
 length(myseq)
[1] 1042519
 head(myseq)
[1] "g" "c" "g" "g" "c" "c"
 sum(myseq == "a")
[1] 306721
```

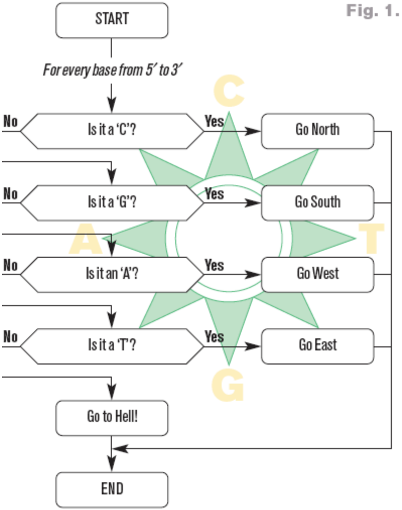
Chirochore practical

With the function `pairs()` show how close is this genome to PR2



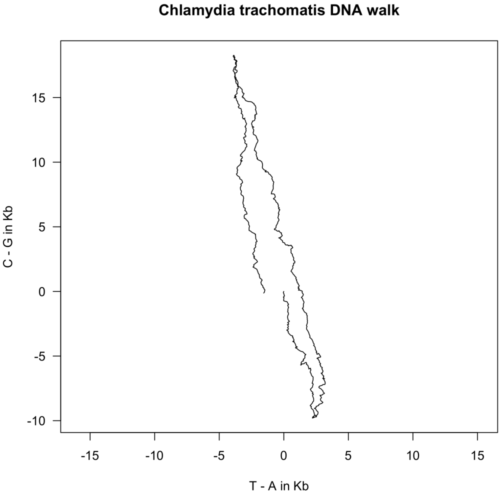
state:

Chirochore practical



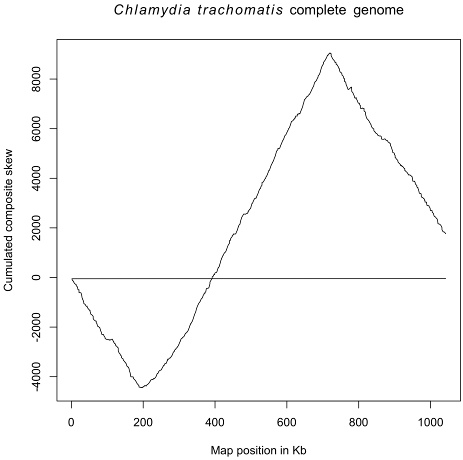
Make a simple DNA walk on this genome. Use the functions `ifelse()`, `cumsum()` and plot a point every Kb with the same scale (also in Kb) for both axes.

Chirochore practical



Chirochores practical

Plot the results of `oriloc()`.

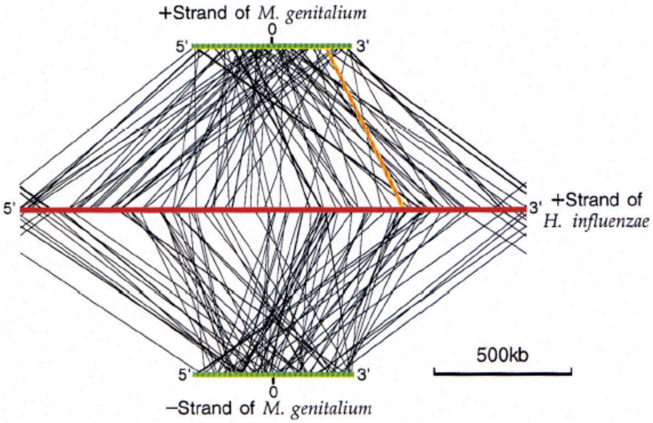


X-rated structure: gene order evolution

1. Introduction
2. Genome size
3. Topology & #
4. G+C content
5. Replichores
6. Gene orientation biases
7. Chirochores
8. X-rated structure

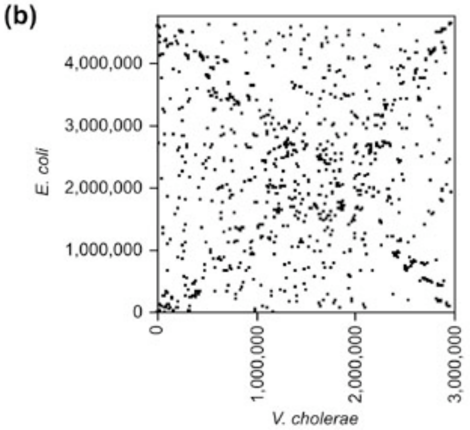
Tsuzumi drum

Gene order comparison in bacteria


+Strand of *M. genitalium*
0
5'
3'
5'
3'
+Strand of *H. influenzae*
500kb
5'
3'
0
−Strand of *M. genitalium*

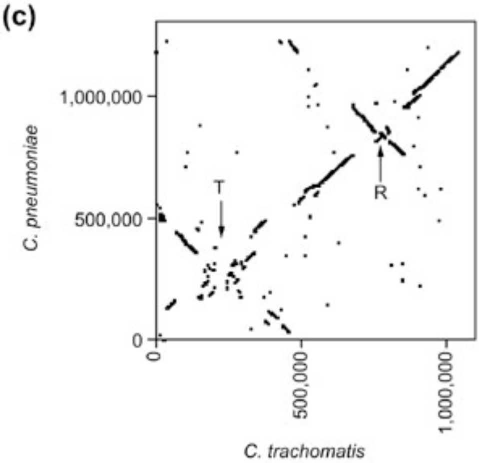
Watanabe *et al.* (1997) *J. Mol. Evol.*, **44**:s57-s64.

Genomic dot plots (*E. coli vs. V. cholerae*)



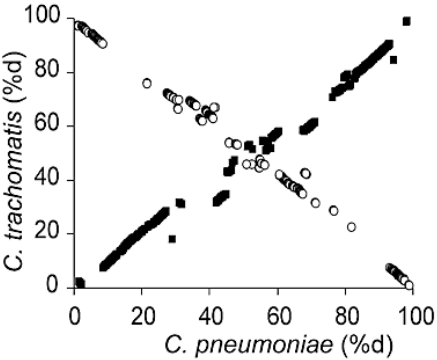
Eisen *et al.* (2000) *Genome Biology*, **1**:research0011.1-9.

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



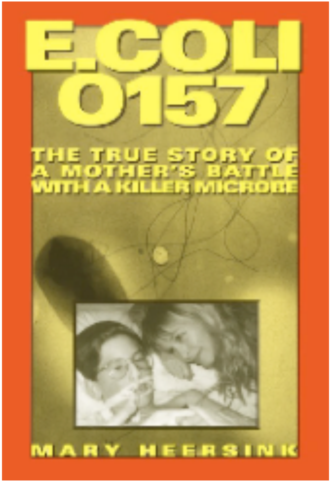
Eisen *et al.* (2000) *Genome Biology*, **1**:research0011.1-9.

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

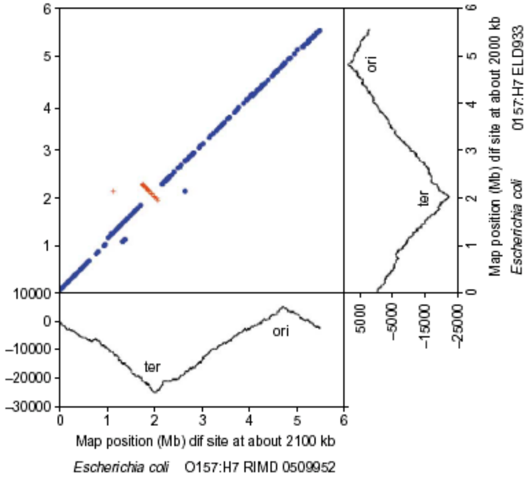


Tillier & Collins (2000) *Nature Genetics*, **26**:195-197.

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

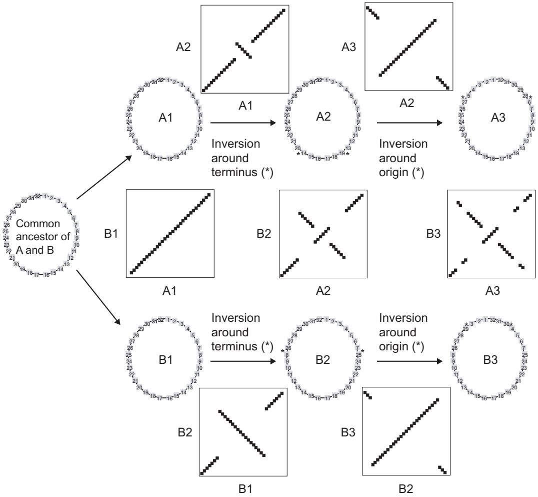


Genomic dot plots (0157:H7 *vs.* O157:H7)



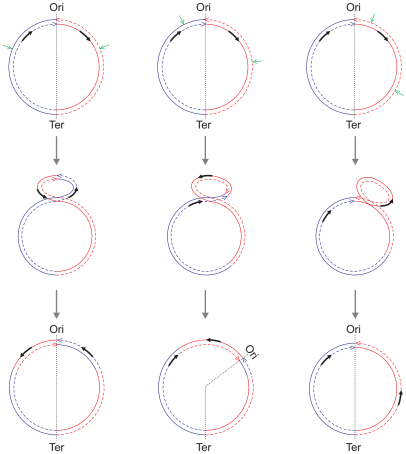
Lobry & Louarn (2003) *Curr. Op. Microbiol.*, **6**:101-108.

Simulation of symmetric inversions


A2
A1
A3
A2
A1
A2
A3
Inversion around terminus (*)
Inversion around origin (*)
Common ancestor of A and B
B1
A1
B2
A2
B3
A3
Inversion around terminus (*)
Inversion around origin (*)
B1
B2
B3
B2
B1
B3
B2

Eisen *et al.* (2000) *Genome Biology*, **1**:research0011.1-9.

Three models for inversions



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

"Bacterial" Genome structures

Spring 2008 Lecture

Pr. J. R. Lobry

Université Claude Bernard Lyon I – France

Last LATEXcompilation was : February 23, 2017