# Genome size

# 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 $\mu$m
- 1 Mb $\approx$ 0.33 mm
- 1 Gb $\approx$ 0.33 m

Bacterial genomes are typically in the mm range (and therefore 1000x bigger than the typical bacterial size).

# 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 $10^9$

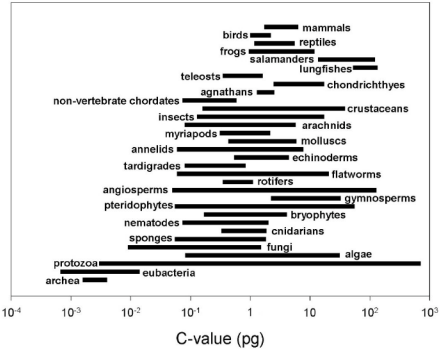
- 1 kb $\approx 10^{-6}$ pg
- 1 Mb $\approx 10^{-3}$ pg
- 1 Gb $\approx$ 1 pg

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

# 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 |
| Bacteriophages: | *E. coli*: | *Homo Sapiens*: |
| 10-100 genes | $\sim$ 5000 genes | $\sim$ 25000 genes |

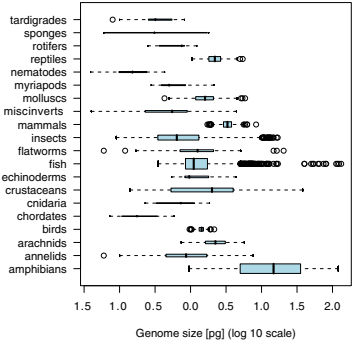
# C value paradox




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

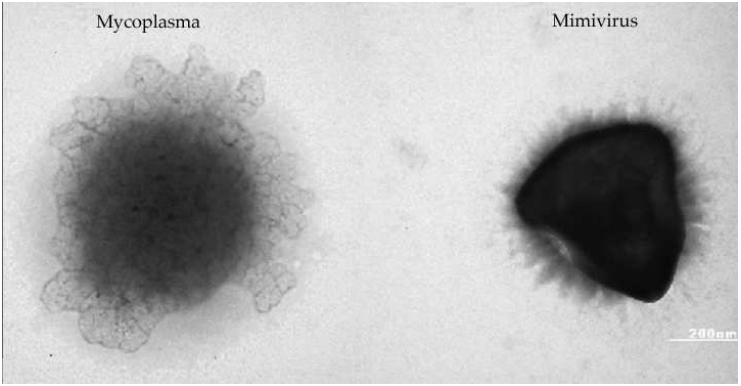
# C value paradox: who has the biggest genome?

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



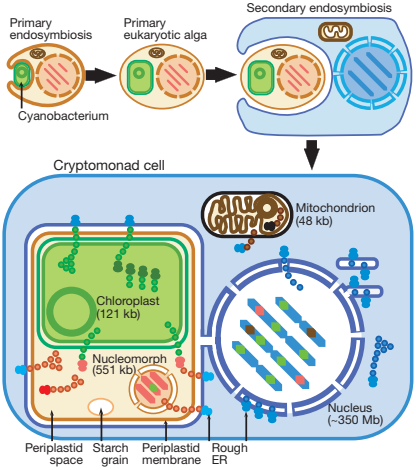

Hard Quizz : what makes the humans being "biologically" different from other animals, if not a bigger genome?

# 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.

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




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

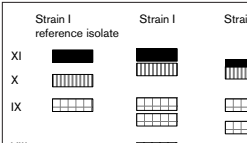
# Tiny eucaryal genome: *Encephalitozoon cuniculi* is only 2.9 Mb

**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

# Nothing in Biology Makes Sense Except in the Light of Evolution (*T. Dobzhansky*)

**Principle**

Species evolve through random changes which are submitted to natural selection

- Variability
- Natural selection

# What is the distribution of bacterial genome size...

...and what do you expect if it is a character under selection? Not under selection?

Study this yourself:
If you do not remember in details what is a mixture of gaussian laws, read first – and quickly:
`http://pbil.univ-lyon1.fr/R/fichestd/tdr221.pdf`

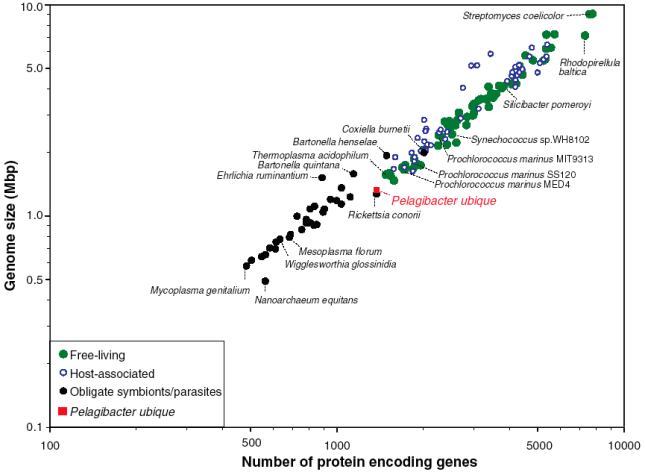
Then:
`http://pbil.univ-lyon1.fr/R/fichestd/tdr222.pdf`
(use the file `goldtable.txt` already downloaded for the last part)

# Genome size for 279 bacteria (GOLD 2002)

# Genome size for 1062 bacteria (GOLD 2007)

# Genome size summary

# Generalists versus specialists




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

# Genome size & repeat density

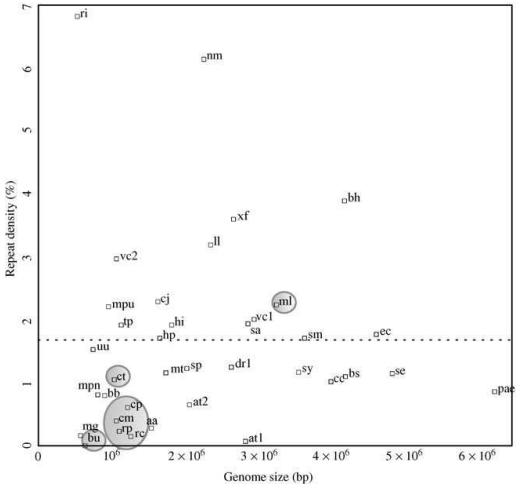
*Genetica* **115:** 1–12, 2002.


1

## 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

8

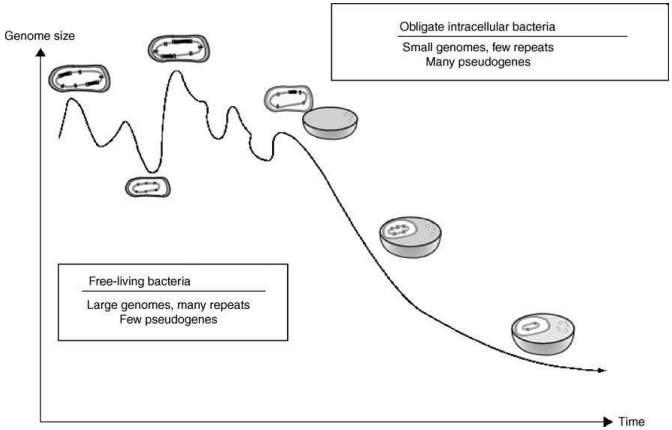
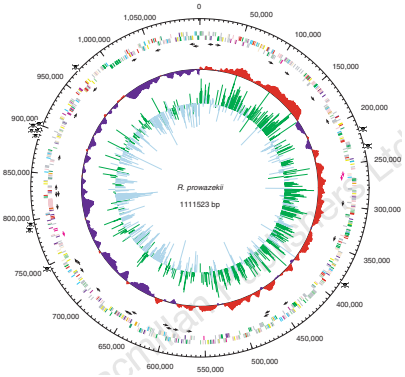



*Figure 3.* Schematic illustration of genome size variations as a function of time during transitions to intracellular growth habitats. Filled boxes represent mobile genetic elements. Genomes of obligate intracellular bacteria are smaller and have a lower content of repeated sequences (//) and a higher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

# Pseudogenes in *Rickettsia prowazekii*




Andersson, S.G. *et al* (1998) *Nature*, The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria **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.

# 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 | ANL | 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

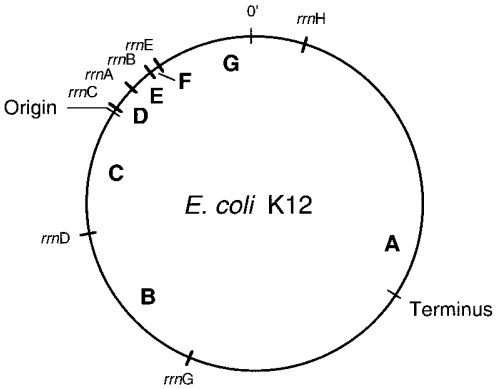



FIG. 1.—Locations of I-*Ceu*I recognition sites on the *E. coli* K12 chromosome. I-*Ceu*I cleaves at the seven *rrn* genes, whose map positions are indicated. The resulting restriction fragments are designated **A** through **G**.

# 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("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?

# Genome size is highly polymorphic in *E. coli*

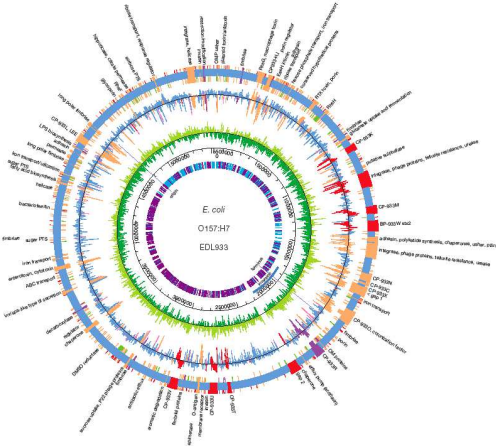
# Genome size phylogenetic inertia

# Genome size phylogenetic inertia

# The nice hidden structure

# The nice hidden structure (II)

# 0157:H7 EDL933 *vs* MG1655




Red parts on the outer circle represent insertion sequences in the pathogenic bacteria.

# Insertion Sequences (IS)

- One of the main reason of within-species genome polymorphism
- IS are DNA sequences inserted in the genome, present among certain individuals in a population
- Typically IS can be lysogenic phages or sequences acquired by horizontal transfer

# Horizontal transfert

There is 3 main ways of acquiring sequences by horizontal transfert for bacteria:

- Transformation : acquisition of external DNA sequences by "competent" bacteria
- Conjugation : exchange of DNA sequences between individuals in a population
- Transduction : phage-mediated transfer of DNA

# Insertion Sequences and pathogenicity

Many bacteria are pathogens because of acquired IS, e.g. from phages (prophages). Examples:

- *Y. pestis* acquired the toxicity protein from a phage
- *E. coli O157:H7* is a pathogenic strain of *E. coli*, only because of added IS.
- *P. aeruginosa PA01* contains inserted sequences of bacteriocins, designed by phages to kill bacteria

# Genome size polymorphism in bacteria