M2 - BIM

Comparative Genomics

Introduction

Alessandra Carbone
Université Pierre et Marie Curie
(Alessandra.Carbone@lip6.fr)

What is Comparative Genomics?

Input: genomic sequences and their annotation

There are several different ways to compare genome at different resolutions

- Complete genomes
  - Alignments between genomes
  - Synteny (conservation of gene order)
  - Unusual regions
- Groups of genes
  - Gene families and species specific genes
  - Gene clustering based on functions
- Genes: variations in gene sequences
  - Codon usage, SNPs, indels, pseudogenes

Resolution

- Low resolution
  - entry: complete genomes
  - example of event: rearrangement
- High resolution
  - entry: nucleotide sequences
  - example of event: specific mutation

Evolutionary events at the scale of complete genomes

- Rearrangements
- gene/region duplication
- gene/region loss
- Chromosome ⇔ plasmid exchange
- Lateral gene transfer

Genome analysis

- Variation in
  - Genome size
  - GC content
  - Codon usage
  - Amino acid composition
  - Genome organisation
    - Single circular chromosomes
    - Linear chromosome + extra chromosomal elements

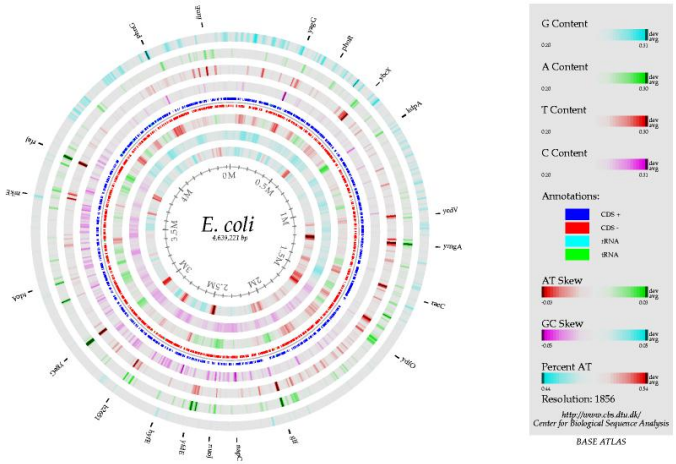
E. coli: 4.6Mbp
M. pneumoniae: 0.81Mbp
B. subtilis: 4.20Mbp

B. burgdorferi: 29%
M. tuberculosis: 68%

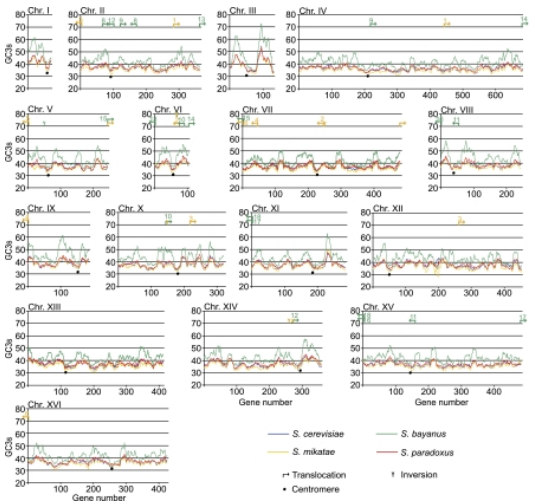
G, A, P, R: GC rich
I, F, Y, M, D: AT rich

5

Analyse statistique à large échelle



GC content distribution in yeast chromosomes



GC profiles of chromosomes in *Saccharomyces species. For Saccharomyces cerevisiae*, the values plotted are a moving average of 15 adjacent genes along each chromosome.

7

Genomes as units of comparison

- Strains of the same species
- Very close species
- Very distant species
  - Orthologs list
  - Evolution of single genes
  - Evolution of organisms

8

Identification of similarities/differences between genomes to understand:

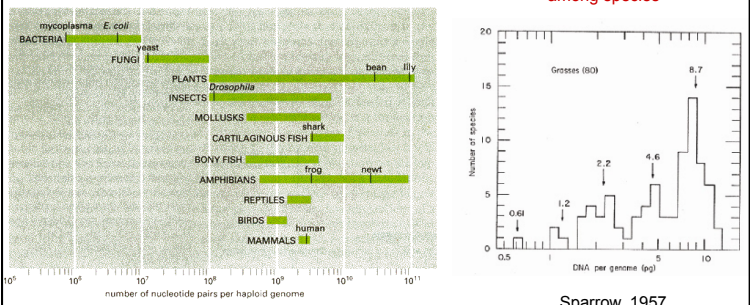
- How 2 organisms evolved/co-evolved?
- Why certain bacteria produce diseases and others do not?
- How to identify therapeutic targets?
- ……

GC analysis paved the way to answer these questions

Genomes lengths

5 decades (8 when counting viruses)

Multimodal distribution among species



Sparrow, 1957

Famous genomes and their lengths

- *Haemophilus influenzae* (1.8 Mb)
  - Human pathogen, first sequenced genome (1995)
- *Escherichia coli* (4.6 Mb)
  - Human pathogen and model organism (1997)
- *Agrobacterium tumefaciens* (6 Mb)
  - Plant pathogen and biotechnological tool (2001)

- The genome of *Plasmodium falciparum*, the malaria parasite, sizes 23 Mb.
- The human genome is around 150 times larger, the mouse genome > 100 times, and the drosophila genome > 5 times.

Questions on genome length:

How ancestral genomes could become longer and longer along evolution?

In what manner their genetic material differs?

B.Dujon lectures will address these questions

MUMmer, a tool for comparing entire genomes

Comparisons and analysis at the level of the nucleic acids or the proteins

- MUMmer (for the alignment of entire genomes)
- …..

- Comparative genomics of Parasites @ TIGR
- Microbial Genome Database (MDG) - Japan
- Comprehensive Microbial Resource page@TIGR
- …..

13

- MUMmer has been developed by Steven Salzberg's group at TIGR
  - NAR (1999) 27:2369-2376
  - Delcher AL, Phillippy A, Carlton J, Salzberg SL. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res.* 2002 Jun 1;30(11):2478-83.
  - Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. Versatile and open software for comparing large genomes. *Genome Biol.* 2004;5(2):R12
- http://mummer.sourceforge.net

14

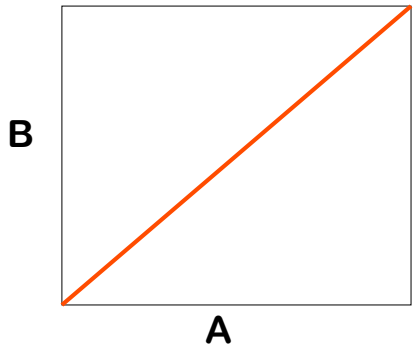
Blocks of genes as units of comparison

- Colinearity in gene order (synteny)
- Groups of conserved genes appearing with the same order in two different genomes: syntenic blocks or syntenic clusters
- Translocation: mouvement of genomic segments from a position to another in the genome.

15

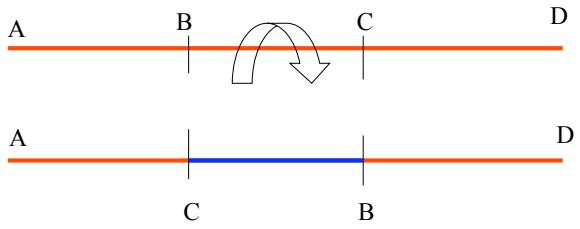
The two genomes case

If sequences were identical we would see:

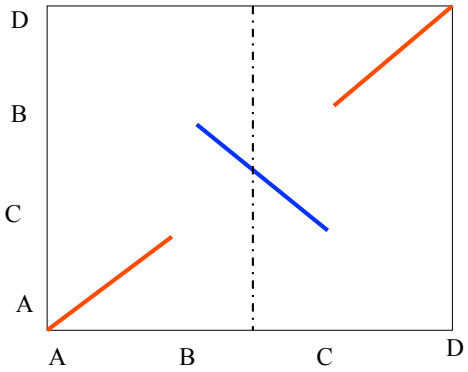


16

An *inversion*

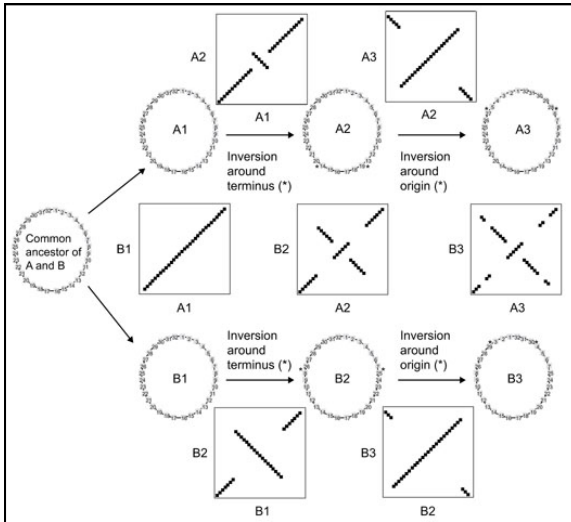


17



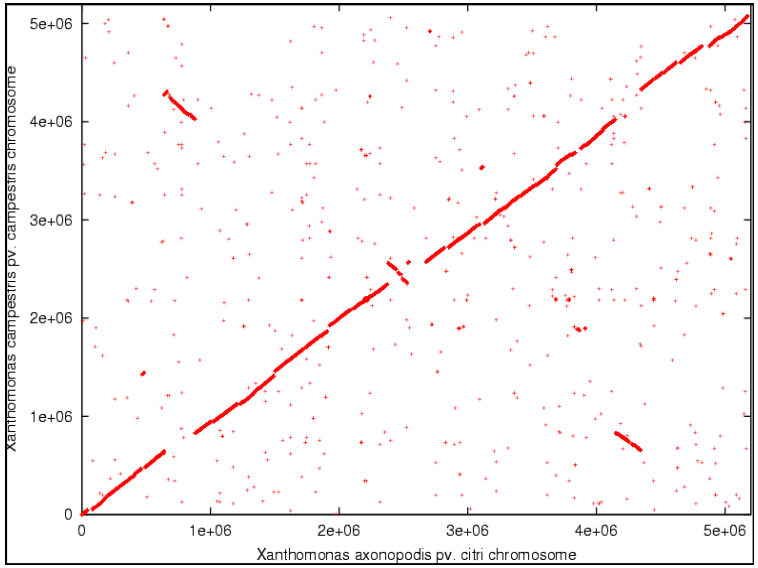
Such inversions seem to happen around the *replication origin* or the *terminus*

18



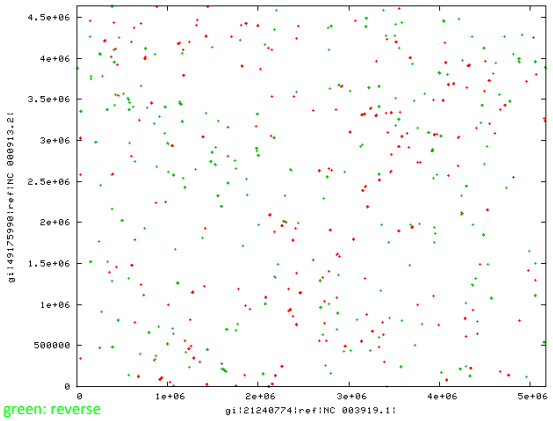
Eisen JA, Heidelberg JF, White O, Salzberg SL. Evidence for symmetric chromosomal inversions around the replication origin in bacteria. *Genome Biol.* 2000;1(6):RESEARCH0011

19



***E. coli* K12**

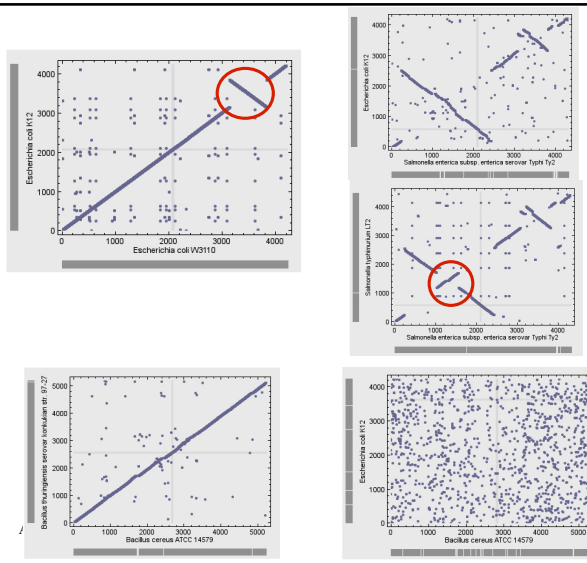
Promer alignment



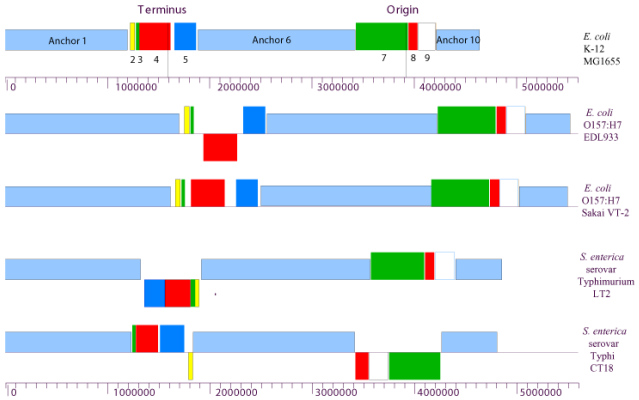
Red: direct; green: reverse

Both are γ proteobacteria!

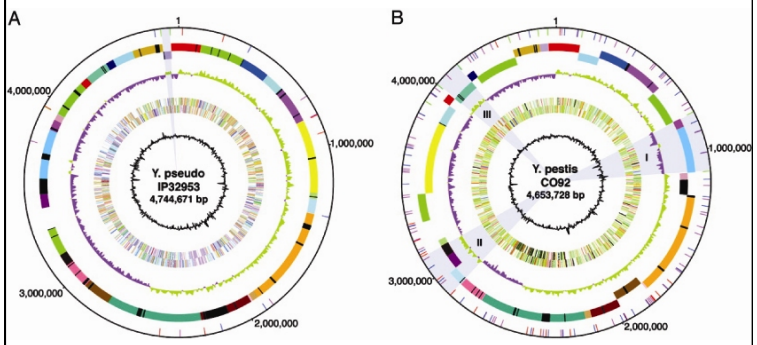
Xanthomonas axonopodis* pv *citri



Rearrangements of Locally Collinear Blocks between three Escherichia and two Salmonella strains



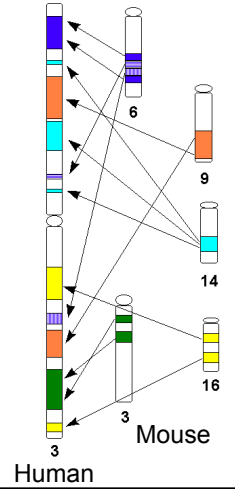
Rearrangement in bacteria :
Yersinia pseudotuberculosis versus *Yersinia pestis*



High virulence

Human and mouse

- There is a meaningful quantity of genome rearrangement between human and mouse.
- Here we see a chart of the human chromosome 3.
- It contains sequences that are homologous to at least 5 chromosomes in the mouse.



An example : human and mouse

- The mouse has 2.1×10^9 bp vs 2.9×10^9bp in human.
- About 95% of the genetic material is shared.
- 99% of shared genes on a total of 30,000.
- 300 homologous genes in the two species are involved into immunity, detoxification, smell and sex.

Genome rearrangements are rare events compared to **pointwise mutations**:

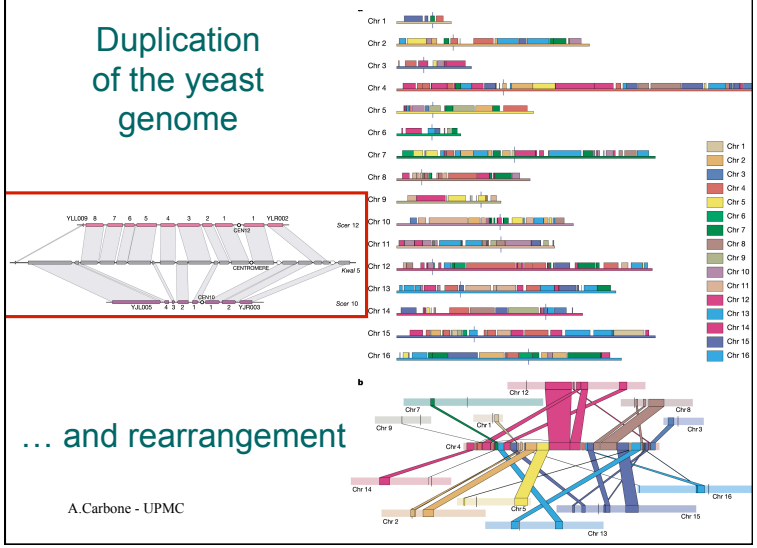
10 substitutions per generation of an organism
1 non fatal rearrangement every 5-10 millions years

The low frequency of rearrangement events allows to establish the existence of the evolutionary process. As a consequence, passing through the reconstruction of all rearrangements, we can reconstruct evolutionary hypothesis.

Search for synteny

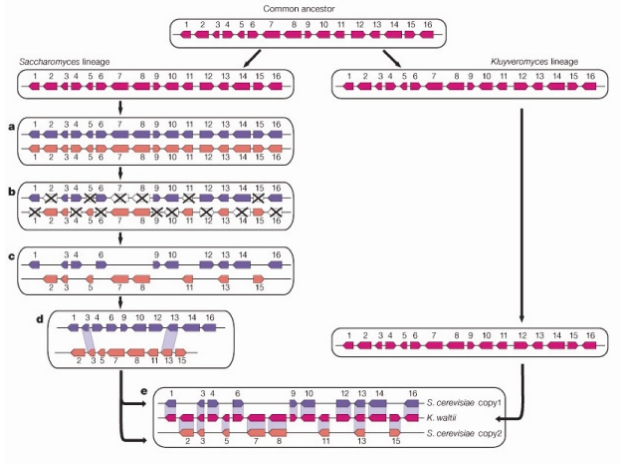
- Considers regions in the two genomes showing important similarities in terms of
  - Sequence
  - Conservation of gene order
- Most probably it relies on the existence of a common ancestor.

Duplication of the yeast genome



… and rearrangement

A.Carbone - UPMC



A.Carbone - UPMC 30

Whole Genome Duplications in Evolution

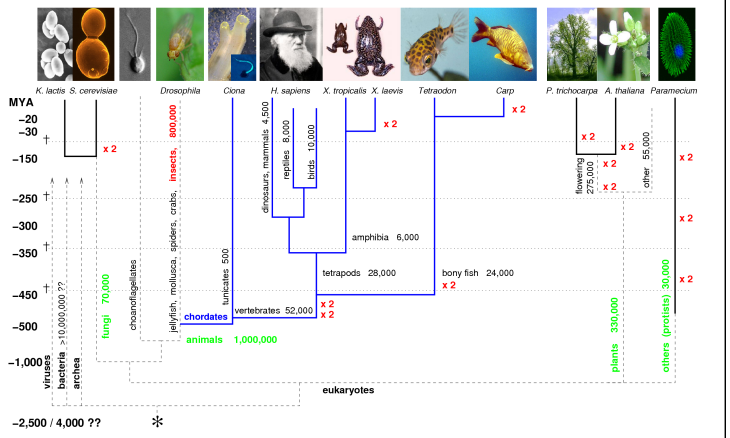


Image de Herve Isambert

Genes as units of comparison

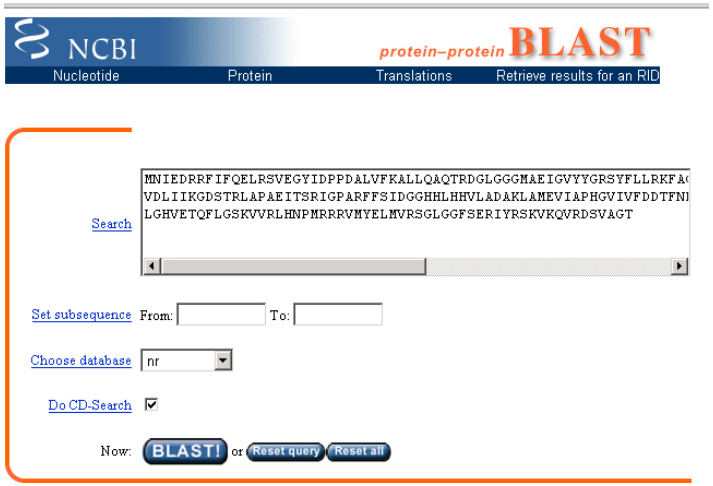
- Number
- Content (sequence)
- Localisation (position)
- Gene order
- Gene clustering (genes belonging to the same metabolic pathway belong to the same cluster)

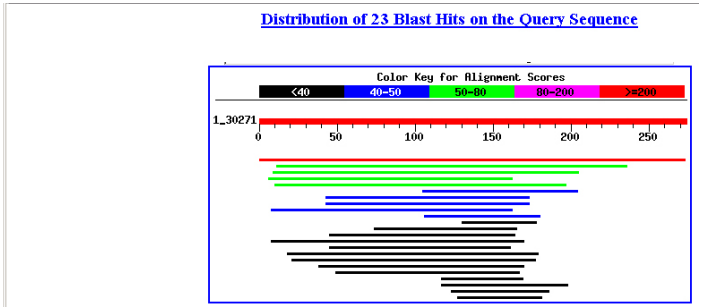
32

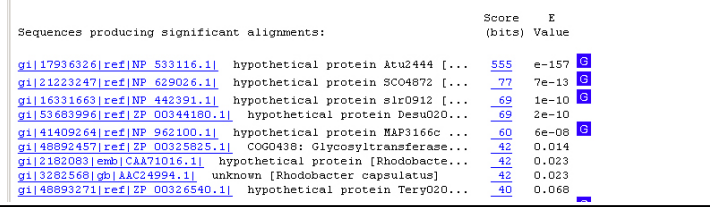
How to search for genes in different species

- DNA databases are much larger than protein databases.
- Translation of a DNA sequence into a protein sequence induces an information loss.
- Protein sequences are more conserved than DNA sequences.

The translation of a DNA sequence into a protein sequence brings always better results!







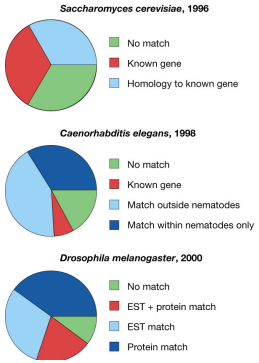
The notion of Reciprocal Best Hit

Take a sequence S_1
BLAST S_1 and get the 1st BLAST choice S_2 (sequence with best E-value)
BLAST S_2 and get the 1st BLAST choice S_3 (sequence with best E-value)
If S_3 is S_2 then we say that S_2 and S_2 are reciprocal best hits (RBH)

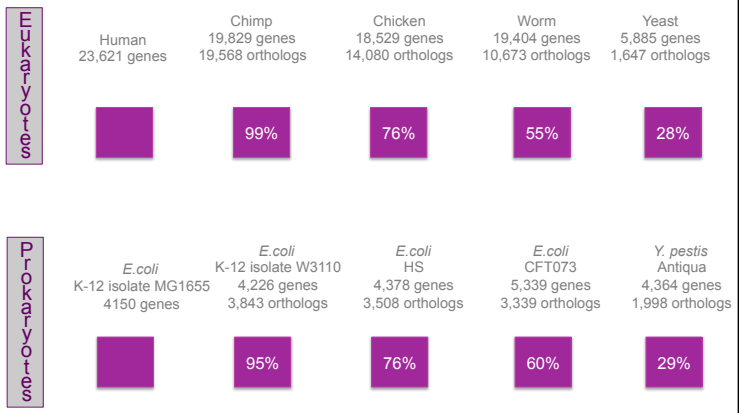
Weaker notions of homology exist:
Take a sequence S_1
BLAST S_1 and get the 1st BLAST choice S_2 (sequence with best E-value) if it is at >30% sequence identity from S_1

Old paradigm: few genes, the same genes

Organism	Year	Millions of bases sequenced	Total coverage (%)	Coverage of euchromatin (%)	Predicted number of genes	Number of genes per million bases sequenced
Saccharomyces cerevisiae	1996	12	93	100	5,800	483
Caenorhabditis elegans	1998	97	99	100	19,099	197
Drosophila melanogaster	2000	116	64	97	13,601	117
Arabidopsis thaliana	2000	115	92	100	25,498	221
Human chromosome 21	2000	34	75	100	225	7
Human chromosome 22	1999	34	70	97	545	16
Human genome rough draft (public sequence)	2001	2,693	84	90	31,780	12
Human genome rough draft (Celera sequence)	2001	2,654	83	88–93	39,114	15



Intra-species variations



1 genome is not enough to represent the species ⟶ PAN- GENOME

Comparison of several *Ecoli* strains

Organised Genome Dynamics in the *Escherichia coli* Species Results in Highly Diverse Adaptive Paths
PLoS Genetics, 2009

Strains	Host	Sample	Clinical condition (Pathotype[a])	Phylogenetic group[b]	Extraintestinal mouse model phenotype[c] (Number of mice killed out of 10)	Genome sequence reference
K-12 MG1655	Human	Faeces	Commensal	A	NK (0)	[115]
K- 12 W3110	Human	Faeces	Commensal	A	NK (0)	Nara Institute of Science and Technology
IAI1	**Human**	**Faeces**	**Commensal**	**B1**	**NK (0)**	**This work**
55989	**Human**	**Faeces**	**Diarrhoea (EAEC)**	**B1**	**K (10)**	**This work**
S. boydii 4 227 (Sb 227)	Human	Faeces	Shigellosis	S1	ND[d]	[116]
S. sonnei 046 (Ss 046)	Human	Faeces	Shigellosis	SS	ND	[116]
S. flexneri 2a 301 (Sf 301)	Human	Faeces	Shigellosis	S3	ND	[117]
S. flexneri 2a 2457T (Sf 2457T)	Human	Faeces	Shigellosis	S3	NK (0)	[118]
S. flexneri Sb 8401 (Sf 8401)	Human	Faeces	Shigellosis	S3	ND	[119]
S. dysenteriae 1 197 (Sd 197)	Human	Faeces	Shigellosis	SD1	ND	[116]
O157:H7 EDL933	Human	Faeces	Diarrhoea (EHEC)	E	NK (1)	[120]
O157:H7 Sakai	Human	Faeces	Diarrhoea (EHEC)	E	NK (1)	[121]
UMN026	**Human**	**Urine**	**Cystitis (ExPEC)**	**D**	**K (10)**	**This work**
IAI39	**Human**	**Urine**	**Pyelonephritis (ExPEC)**	**D**	**K (8)**	**This work**
UTI89	Human	Urine	Cystitis (ExPEC)	B2	K (10)	[122]
APEC O1	Chicken	Lung	Colisepticemia (ExPEC)	B2	K (10)	[123]
S88	**Human**	**Cerebro-spinal fluid**	**New born meningitis (ExPEC)**	**B2**	**K (10)**	**This work**
CFT073	Human	Blood	Pyelonephritis (ExPEC)	B2	K (10)	[30]
ED1A	**Human**	**Faeces**	**Healthy subject**	**B2**	**NK (0)**	**This work**
536	Human	Urine	Pyelonephritis (ExPEC)	B2	K (10)	[124]
E. fergusonii	**Human**	**Faeces**	**Unknown**	**Outgroup**	**NK (1)**	**This work**

The strains in bold correspond to the strains sequenced in this work.
[a]EAEC (enteroaggregative *E. coli*), EHEC (enterohaemorrhagic *E. coli*), ExPEC (extraintestinal pathogenic *E. coli*).
[b]The *E. coli* and *Shigella* phylogenetic groups are as defined in [22] and [6], respectively.
[c]K, killer; NK, Non Killer [32].
[d]ND, not determined.
doi:10.1371/journal.pgen.1000344.t001

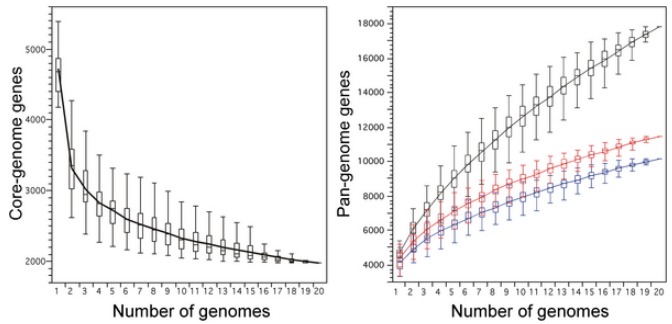
Chromosome features	*E. coli* K-12 MG1655	*E. coli* strains						*E. fergusonii* ATCC
		55989	IAI1	ED1a	S88	IAI39	UMN026	
Genome Size (bp)	4 639 675	5 154 862	4 700 560	5 209 548	5 032 268	5 132 068	5 202 090	4 588 711
G+C content (%)	50.8	50.7	50.8	50.7	50.7	50.6	50.7	49.9
rRNA operons	7 (+5S)	7 (+5S)	7 (+5S)	7 (+5S)	7 (+5S)	7 (+5S)	7 (+5S)	7 (+5S)
tRNA genes	86	94	86	91	91	88	88	87
Total Protein-coding genes[a]	4306	4969	4491	5129	4859	4906	4918	4336
Pseudogenes[b] (nb)	81	79	51	95	90	80	45	22
Protein coding density[c]	85.7	87.4	87.6	86.2	87	86.1	87.8	84.7
Assigned function[d] (%)	80	74	77	74	77	78	76.5	77
Conserved hypothetical (%)	12.5	23	21.5	23	22	20	22	20
Orphans (%)	7.5	3	1.5	3	1	2	1.5	3
IS-like genes (nb)	66	150	42	118	47	224	92	29
Phage-associated genes (nb)	231	406	201	657	507	393	429	235

[a]The number of protein-coding genes is given without the number of coding sequences annotated as artifactual genes (Supplementary Table 2A).
[b]The number of pseudogenes computed for each genome corresponds to the real number of genes that are pseudogenes: one pseudogene can be made of only one CDS (in this case the gene is partial compared to the wild type form in other *E. coli* strains) or of several CDSs (generally two or three CDSs corresponding to the different fragments of the wild type form in other *E. coli* strains). These lists of pseudogenes are available in Supplementary Table 1.
[c]The computed protein coding density takes into account the total length of protein genes excluding overlaps between genes, artifacts, and RNA genes.
[d]Protein genes with assigned function include the total number of definitive and putative functional assignments.
doi:10.1371/journal.pgen.1000344.t002

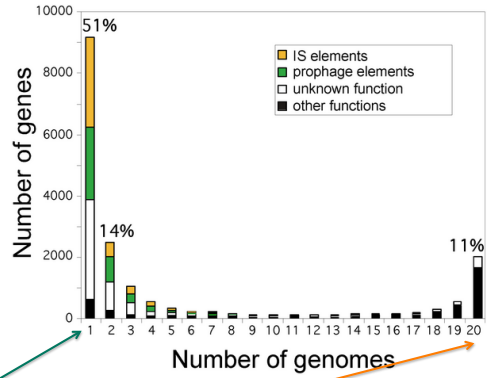
D'autres informations manquent : ARN non codants par exemple

Plasmid features	*E. coli* strains					*E. fergusonii* ATCC
	55989	ED1a	S88	UMN026		
Genome Size (bp)	72 482	119 594	133 853	122 301	33 809	55 150
G+C content (%)	46.1	49.2	49.3	50.5	42	48.5
Total Protein-coding genes[a]	100	150	144	149	49	54
Pseudogenes[b] (nb)	7	11	9	8	0	5
Protein coding density[c]	75.6	86.2	87	79.4	87.5	88.7
Assigned function[d] (%)	74	53	65	65.7	35.4	46.6
Orphans (%)	17	31.5	25.8	27.8	12.5	20.7
Hypothetical (%)	9	15.5	9.2	6.5	52.2	32.7
IS-like genes (nb)	18	14	14	15	0	4

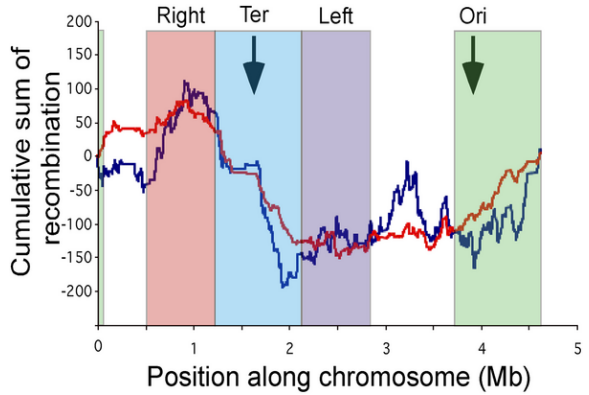
[a]The number of protein-coding genes is given without the number of coding sequences annotated as artifactual genes (Supplementary Table 2A).
[b]The number of pseudogenes computed for each genome corresponds to the real number of genes that are pseudogenes: one pseudogene can be made of only one CDS (in this case the gene is partial compared to the wild type form in other *E. coli* strains) or of several CDSs (generally two or three CDSs corresponding to the different fragments of the wild type form in other *E. coli* strains). These lists of pseudogenes are available in Supplementary Table 1.
[c]The computed protein coding density takes into account the total length of protein genes excluding overlaps between genes, artifacts, and RNA genes.
[d]Protein genes with assigned function include the total number of definitive and putative functional assignments.
doi:10.1371/journal.pgen.1000344.t003



***Escherichia coli* core and pan-genome evolution according to the number of sequenced genomes.** Number of genes in common (left) and total number of non-orthologous genes (right) for a given number of genomes analysed for the different strains of *E. coli*. At 20 sequenced genomes, the core-genome had 1976 genes (11% of the pan-genome), whereas the pan-genome had (i) 17 838 total genes (**black**), (ii) 11 432 genes (**red**) with no strong relation of homology (<80% similarity in sequence), and (iii) 10 131 genes (**blue**) after removing insertion sequence-like elements (3834, 21% of all genes) and prophage-like elements (3873, 22% of all genes).



Frequency of genes within the 20 analysed *Escherichia coli* genomes. At one extreme of the x-axis are the genes present in a single genome which are regarded as strain specific genes (9054 genes: 51% of the pan-genome), while at the opposite end of the scale are situated the genes found in all 20 genomes, which represent the *E. coli* core-genome (1976 genes: 11% of the pan-genome). Coloured rectangles represent the proportion of insertion sequence (IS)-like elements (**yellow**), prophage-like elements (**green**), and genes of unknown/unclassified function (**white**). **Black** rectangles represent genes for which a function can be assigned. Strain-specific genes correspond to 2885 IS-like elements (32%), 2352 prophage-like elements (26%), and 3220 genes of unknown/unclassified function (35%).



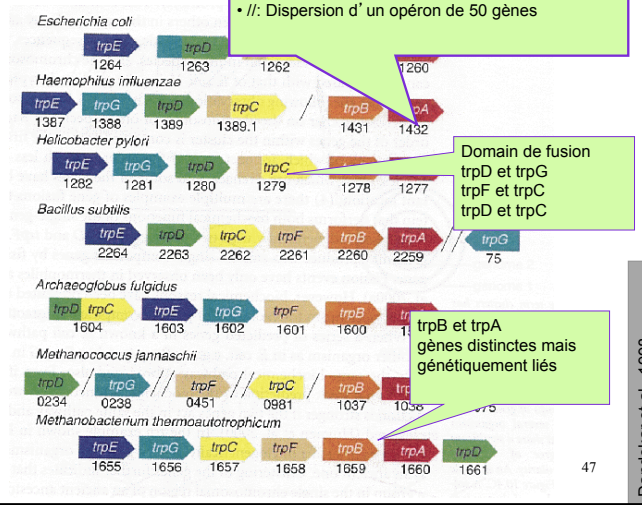
Gene conversion rate (i.e., probability of being involved in a gene conversion event $C_{gc}.L_{gc}$) is shown in **blue**, and G+C content in **red**. A decrease in the cumulative sum reflects regions of lower-than-expected values of the statistics. Around the terminus domain, we found a decrease in both recombination and G+C content. Coloured boxes represent the 4 different organisation macrodomains (Right, Ter, Left, Ori). The arrows point towards the origin and terminus of replication.

From **gene order** to **gene function**

- Gene order is very conserved for close species but it changes by rearrangement.
- By considering larger evolutionary distances, we loose the correspondance between gene order for orthologous genes.
- Groups of genes with similar biochemical functions have the tendency to remain localized.
  - Genes involved in tryptophane synthesis (trp genes) in *E. coli* and other prokaryotes

Structure de

- Nombres: nombre du gène
- fleches: direction de la transcription
- //: Dispersion d' un opéron de 50 gènes

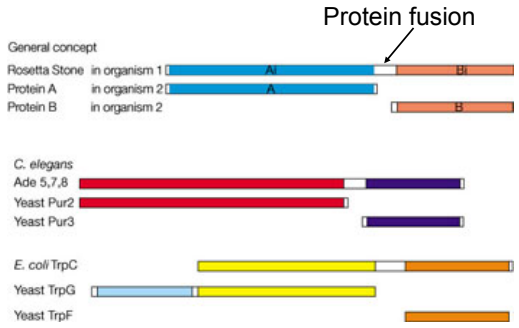


Domain de fusion
trpD et trpG
trpF et trpC
trpD et trpC

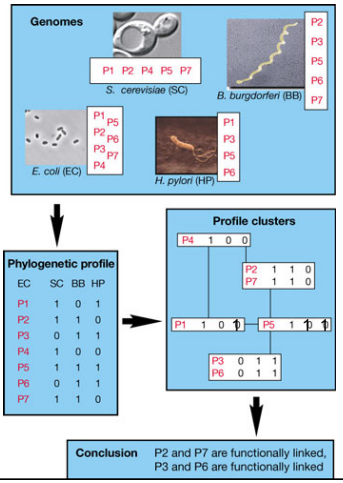
trpB et trpA
gènes distinctes mais
génétiquement liés

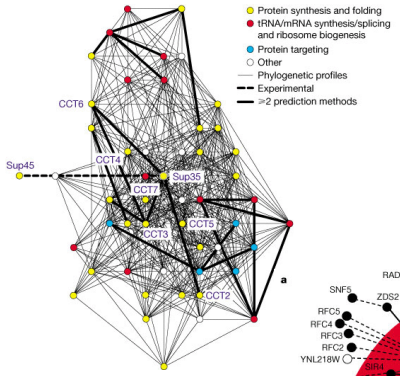
Dandekar et al., 1998

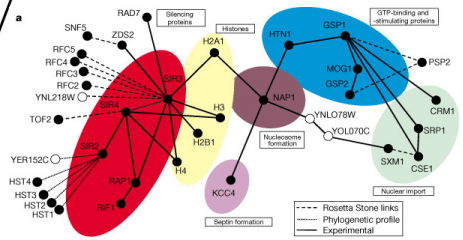
Search of genes in different species: criteria for the detection of their functional relation



Phylogenetic profiles







Complex networks of interaction: **protein networks**

Percentage of proteins with unknown function in different genomes

Kingdom	Clade	Species	number of proteins	% of unknown proteins
Eukaryota	Metazoa	*Brugia malayi*	11472	20
		Caenorhabditis elegans	26047	24
		Drosophila melanogaster	27752	19
		Anopheles Gambiae	14576	22
		Ciona intestinalis	4122	25
	Fungi	*Saccharomyces cerevisiae*	6607	23
	Amoebozoa	*Entamoeba histolytica*	8201	44
		Plasmodium falciparum	5491	30
		Dictyostelium discoideum	12646	25
	Diplomonadida	*Giardia intestinalis*	5012	49
	Cryptophyta	*Giardia theta*	24822	51
		Bigelowiella natans	21000	47
	Stramenopiles	*Phaeodactylum tricornutum*	10408	20
	Parabasalia	*Trichomonas vaginalis*	59681	43
Bacteria	Proteobacteria	*Salmonella enterica*	4720	33
		Helicobacter pylori	1593	32
	Elusimicrobia	*Elusimicrobium minutum*	1548	35
	Firmicutes	*Staphylococcus aureus*	2620	62
		Enterococcus faecalis	3278	27
		Streptococcus pneumoniae	1990	23
	Actinobacteria	*Mycobacterium tuberculosis T46*	4134	42
	Cyanobacteria	*Microcystis aeruginosa*	5356	50
Archaea	Euryarchaeota	*Methanobrevibacter smithii*	1710	25
		Methanococcus maripaludis	1807	31
		Halobacterium salinarum	2749	47
	Korarchaeota	*Candidatus Korarchaeum cryptofilum*	1612	25
	Thaumarchaeota	*Cenarchaeum symbiosum A*	2017	48
	Crenarchaeota	*Pyrobaculum oguniense TE7*	2869	38

Génomique comparative des phages

A. Carbone. Codon bias is a major factor explaining phage evolution in translationally biased hosts, *Journal of Molecular Evolution*, 66(3):210--23, 2008.

Synthetic biology | Bacteria and environment

Phages ⟷ Bacteria **are** the phage environment

A significant fraction of the procaryotic community is infected by phages

The total number of viruses, which is much larger than the total prokaryotic abundance, varies strongly in different environments and it is correlated with bacterial abundance and activity

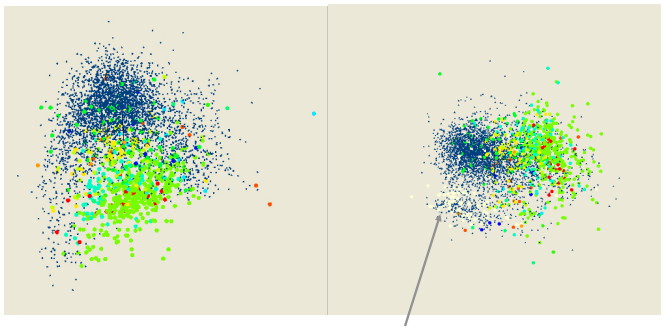
Codon bias is a major factor explaining phage evolution in transl biased hosts

116 phages of 11 translationally biased host bacteria :

actinobacteria	*Mycobacterium smegmatis*
	Mycobacterium tuberculosis
proteobacteria gamma	*Escherichia coli*
	Vibrio cholerae
	Salmonella thyphimurium
firmicutes bacillales	*Bacillus subtilis*
	Listeria monocytogenes
	Staphylococcus aureus
firmicutes lactobacillales	*Lactococcus lactis*
	Streptococcus pyogenes
chlamydiales	*Chlamydophila caviae*

102 of these phages display at least one gene with high bias

Vibrio cholerae + 10 phages



Membrane proteins and permease proteins («hydrophobic» aa)

Functional classification of highly biased phage genes

Phage	cap	mor	tai	lys	hyp	oth	hig	Phage	cap	mor	tai	lys	hyp	oth	hig
Enterobacteria phage RB43	5			4		35	7	51	Lactococcus phage P335	2				2	1		5
Enterobacteria phage RB49	3	2	2			26	6	39	Lactococcus phage bIL170	2					3		5
Enterobacteria phage RB69	4	2	1	1		8	4	20	Lactococcus phage r1t	1					4		5
Enterobacteria phage T4	2	2				3	3	10	Lactococcus phage BK5-T	2				1		1	4
Enterobacteria phage T7	2	1				2	3	8	Lactococcus phage ljj50	2				1	2		4
Enterobacteria phage K1E	1			1		3		5	Lactococcus phage $\phi LC3$	1	1				2		4
Enterobacteria phage K1F	1	1				1	1	4	L.lactis phage ul36	1	1			1			3
Enterobacteria phage Mu	1					2		3	Lactococcus phage P008	2					1		3
Enterobacteria phage P1	1		1				1	3	Lactococcus phage 712	2							2
Enterobacteria phage N15	2					1		3	Lactococcus phage c2	1					1		2
Enterobacteria phage 03	1		1					2	Lactococcus phage Tuc2009	1					1		2
Enterobacteria phage λ	1					1		2	Listeria phage A118	1				1	4		6
Enterobacteria phage P2	1	1						2	Listeria phage P100						5	1	6
Enterobacteria phage 186	1							1	Streptococcus phage 315-5					1	6		7
Enterobacteria phage G4	1							1	Streptococcus phage 315-2						3	1	4
Enterobacteria phage HK022	1							1	Streptococcus phage 315-1	1		1			1		3
Enterobacteria phage I2-2	1							1	Staphylococcus phage K	1		1	2		18	3	25
Enterobacteria phage Ike	1							1	S.aureus phage $\phi P68$	3					3	1	7
Enterobacteria phage ϕK	1							1	Staphylococcus phage 44AHJD	3					3	1	7
Enterobacteria phage ϕX174							1	1	S.aureus phage $\phi NM3$ provirus						4	1	5
Enterobacteria phage 933W							1	1	S.aureus phage PVL provirus						33		3
Enterobacteria phage HK97	1							1	S.aureus phage phi11 provirus	2						1	3
Enterobacteria phage ϵ15					6	1	7	Staphylococcus phage ϕETA	1				1	1		3
Salmonella phage SETP3	1				2		3	Staphylococcus phage ϕNM	1				1	1		3
Enterobacteria phage P22	1				1		2	Vibrio phage KVP40	2	1	4			178	57	242
Enterobacteria phage S13							1	1	Vibriophage VP4	4	1	2			4	15	25
Bacillus phage SPBc2					50	9	59	Vibrio phage VP2						12	1	13
Bacillus phage GA-1 virus	1	1		2	8	3	15	Vibrio phage K139	1	1	1					3
Bacillus phage PZA	1	1		1	4		7	Vibrio phage $VGJ\phi$	1					1	1	3
Bacillus phage B103	1		2	1	1	1	6	Chlamydia phage chp1	4					3	2	9
Mycobacterium phage D29	1		5	1	38	3	48	Chlamydia phage chp2	1							1
Mycobacterium phage Che12	2	1	4	3	33	9	52	Chlamydia phage $\phi CPG1$	1							1

Cap = capsid proteins
Mor = morphogenesis
tai = tail proteins
lys = lysis proteins
hyp = hypothetical proteins
oth = other functions

Functional classification of biased coliphage genes

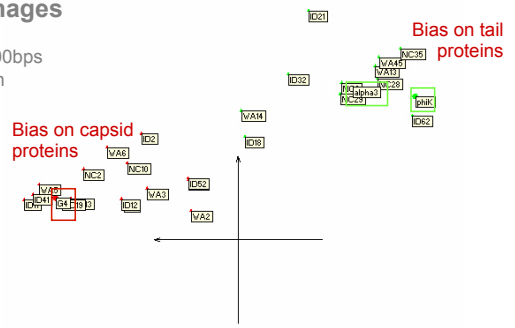
	RB43	T4	RB69	RB49	P1	VT2-sa	933W	T1	λ	K1E	HK97	T7	N15	HK620	HK022	Mu	K1F	P2	186	P4	ϕX174	G4	M13	03	1f1	Ike	ϕK	I2-2
number of genes	292	288	275	274	110	83	80	78	71	62	61	60	60	58	57	55	53	43	42	14	11	10	10	10	10	10	10	9
number of biased genes	168	57	87	113	16	11	11	22	20	19	10	20	19	10	5	17	18	11	16	2	4	4	4	4	1	3	4	4
percentage of biased genes	57	19	31	41	14	13	13	28	28	30	16	33	31	17	9	31	42	38	38	14	36	40	40	40	10	30	40	44
capsid/head/coat proteins	5	5	5	4	2			1	2	1		1	2	1	1	1	1	1	1	2		1	1	2	1	1	2	2
portal proteins	1			1	1											1		1	1									
terminase subunits	1																	2	2									
capsid protease	1															1												
head-tail joining proteins	1			1						1																		
tape measure proteins													1															
virion morphogenesis	3	3	2	5	4			1	1			1	4	2	1	1	3	2	2		2	1		1			1	1
lysis	1	4	4	1	1	2	2	1	1	2	1					1	4		2									
tails/tail fibers proteins	12	2	2	3	3			1	5		2		3		1	1	1	2	3			1		1			1	
Other functions																												
host interaction and inhibitors of host functions	4	3	2	2	2	1	1	1	1		2		1			1		2										
inhibitors of phage functions	1		2	1																				1				
transcription regulation	2	3	3	2	1			1						1		1		1										
translational regulator		1	2																									
DNA condensation and stabilization								2														1		1				
DNA destabilization (ssDNA binding)	1	1	1								1					1					1		1		1		1
dsDNA binding protein	1	1	1	1																								
exonuclease ∞	1					1	1	1	1		1					1												
endonuclease	1	1	1	1					1							1												
RNA replicase																												
dUTPase, dCTPase	1	1	1	1																								
RNA polymerase and RNA pol. binding proteins	1	1	1	2					1		1					1												
DNA polymerase accessory proteins	3	2	3	2							1																	
reductase	3		1	4																								
phosphatase	1	1	1	1																								
recombination protein	1		2	2		1	1	1	1																			
DNA topoisomerase	1	2	1																									
DNA replication protein							1		1				1															
ligase (DNA or RNA)	2											1																
RNAse ribonuclease			1	1																								
dCMP protein		1		1																								
replicase																												
endoprotease																												
DNA primase																				1								
transposase																1												
synthetase	1																											
helicase (DNA or RNA)	3																											
thioredoxin				1																								
unknown function	115	25	51	76	2	6	5	18	4	12	4	9	8	5	2	9	4	2	2	1	0	0	0	0	0	0	0	0

If phage genomes were to contain a pool of essential genes, these functional classes could suggest appropriate candidate genes

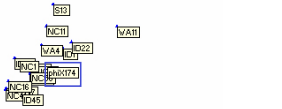
Adaptation explains phage classification

Microviridae phages

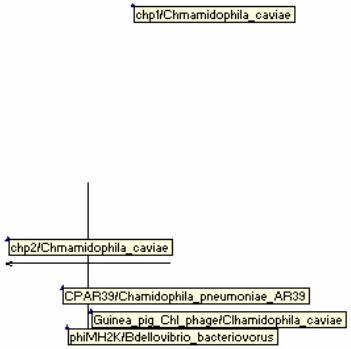
Host: *E.coli*
Genome length: ~6000bps
Genes: 10 in common



Coordinates
=
10-dim vector
of SCCI

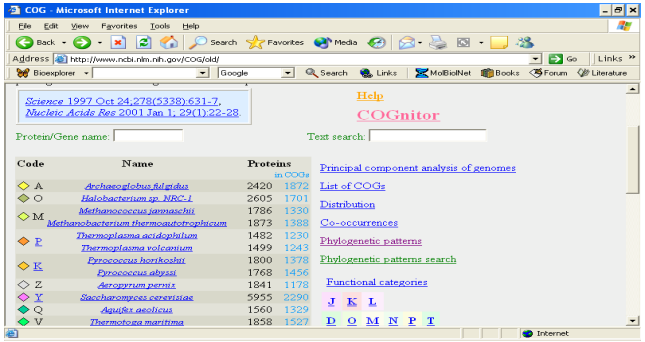


Patterns of adaptation go beyond host species



Phage classification does not reflect host phylogeny

COGs:
Classification phylogénétique des protéines codées dans les génomes complets



62