

# Bioinformatics of microbial identification and biodiversity in the Massive Parallel Sequencing era.

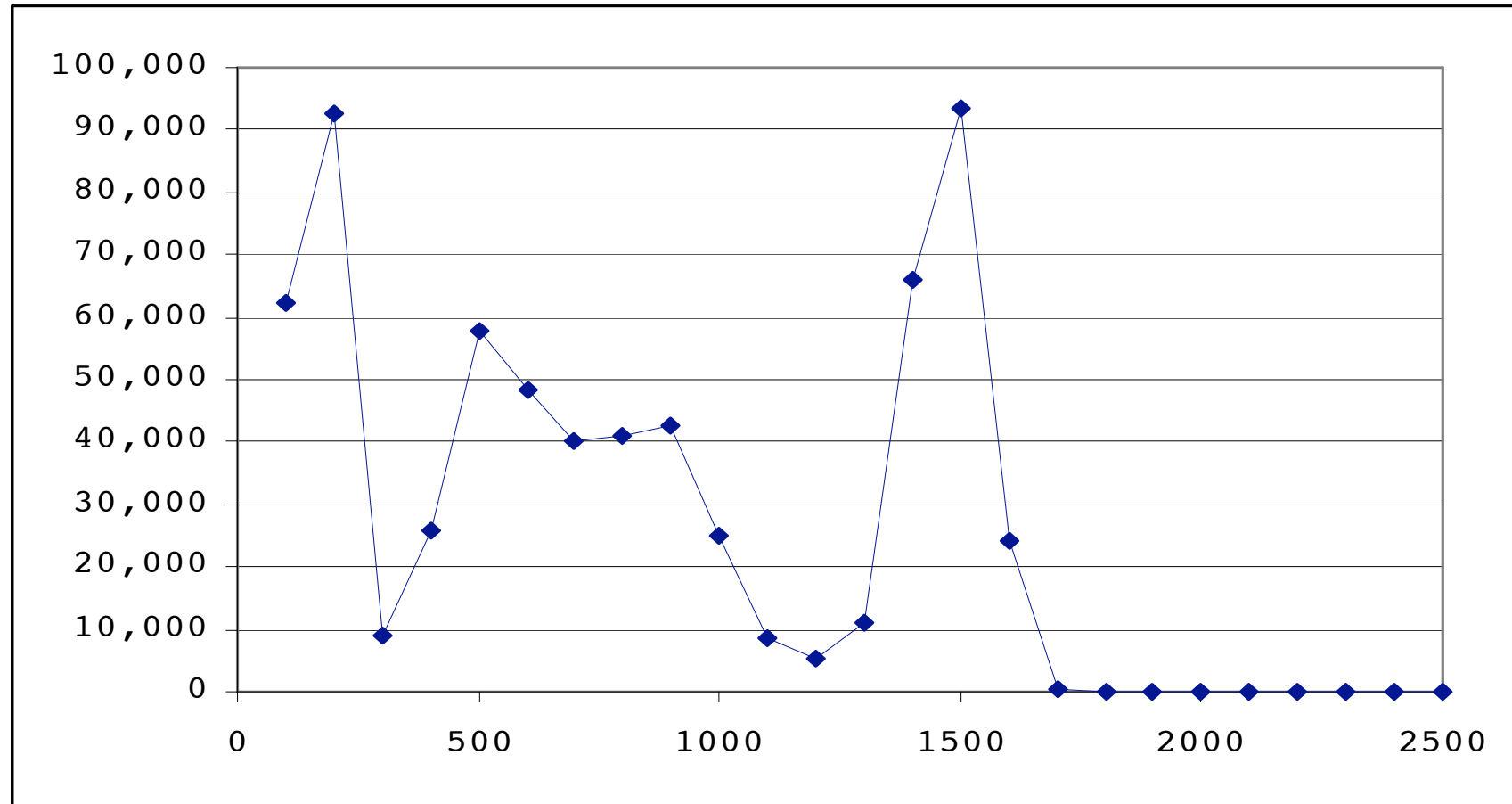
## "Global sequencing"

Richard Christen  
CNRS UMR 6543 & Université de Nice  
[christen@unice.fr](mailto:christen@unice.fr)  
<http://bioinfo.unice.fr>

# Tasks and problems

- Identification of a new isolate: the 16S “gold standard”.
- Other genes.
- Typing a strain.
- Studying biodiversity: new approaches.

## The 16S “gold standard”



Some long sequences correspond to badly annotated sequences such as Z94013, annotated with keywords "16S ribosomal RNA; 16S rRNA gene" when in fact it is a 23S rRNA sequence...

year	gb · 164	16S	% 16S	16S · named	% / 16S
1983	6	0	0	0	0
1985	3	0	0	0	0
1986	7	0	0	0	0
1987	2	0	0	0	0
1988	18	0	0	0	0
1989	28	0	0	0	0
1990	13	0	0	0	0
1991	157	52	33	43	83
1992	337	56	17	52	93
<b>1993</b>	<b>5,839</b>	<b>1,030</b>	<b>18</b>	<b>918</b>	<b>89</b>
1994	2,408	429	18	304	71
1995	3,101	807	26	442	55
1996	4,428	1,450	33	679	47
1997	5,416	1,381	25	768	56
1998	5,902	1,730	29	1011	58
1999	10,432	3,495	34	1840	53
2000	22,804	4,591	20	1832	40
2001	28,005	4,577	16	1877	41
2002	80,078	9,506	12	3263	34
2003	62,438	22,599	36	4918	22
2004	103,177	48,980	47	5107	10
2005	144,454	67,843	47	6937	10
2006	188,366	98,141	52	9403	10
2007	369,001	245,413	67	12908	5
2008	198,997	154,076	77	3574	2
total	1,235,417	666,156	54	55876	8

→ Mostly PCR derived sequences !

gb 165 (june 2008)  
bacteria  
728,358 16S rRNA seqs


“named”  
59,128 seqs >99 nt  
49,678 seqs >500nt  
39,217 seqs >1000 nt


# The 16S “gold standard”

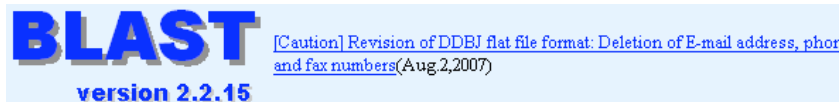
>NF001

```
CTCTCTCTCGCATTCGTCAGTGGCTGGAGGCTGTTGACCCCAACCCTTTC
TTAACGAGTGACAGTGGTTTACAACCCGAAGGCCTTCATCCCACACGCGG
CGTCGCTCCGTCAAGCTTGCGCTCATTGCGGAAGATCCTCGACTGCAGCC
TCCCGTAGGAGTTTGGGCAGTGTCTCAGTCCCAATGTGGCCGGACACCCG
CTAAGGCCGGCTACCCGTCAATGCCTTGGTGGGCCATTACCCTCACCAAC
TAGCTGATAGGACATAGATCCCTCCCCGAGCGGGAGCATCTTCAGAGGCC
TCCTTTAGTCACCGAACCAGGCGATCCAGTGACCCCATCCGGTCTTAGCT
CCGGTTTCCCGGAGTTATCCCGGTCTCGGGGGCAGGTTATCTATGCATTA
CTACCCTTCGCACTAACACCCGTATTGCTACGGTGTCCGTTTCGTCTTGCA
TGCCTAATCACGCCGCTGGCGTTCGTTCTGAGCCAGGATCCAAACTCTAT
CCGG
```

A case study: identification of a “DGGE” band using the “usual” Blast servers

EBI  EBI > Tools > Similarity & Homology > BLAST > WU-BLAST2  
**WU-BLAST2 - Nucleotide Database Query**

NCBI  BLAST Basic Local Alignment Search Tool  
Home Recent Results Saved Strategies Help

DDBJ  **BLAST** version 2.2.15  
[Caution] Revision of DDBJ flat file format: Deletion of E-mail address, phone and fax numbers(Aug.2,2007)

# NCBI

**Database:** All GenBank+EMBL+DDBJ+PDB sequences (but no EST, STS, GSS, environmental samples or phase 0, 1 or 2 HTGS sequences)  
6,904,467 sequences; 23,920,262,828 total letters

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
<a href="#">EF638225.1</a>	Uncultured bacterium clone NF001 16S	<a href="#">985</a>	985	98%	0.0	100%	
<a href="#">EU280647.1</a>	Uncultured bacterium clone J005_A11 1	<a href="#">753</a>	753	95%	0.0	95%	
<a href="#">EF018297.1</a>	Uncultured bacterium clone Amb_16S_1	<a href="#">724</a>	724	96%	0.0	94%	
<a href="#">DQ450802.1</a>	Uncultured planctomycete clone G02_W	<a href="#">611</a>	611	82%	1e-171	94%	
<a href="#">EF515912.1</a>	Uncultured bacterium clone FCPO648 1	<a href="#">549</a>	549	95%	4e-153	90%	
<a href="#">AF432701.1</a>	Uncultured bacterium clone N11.129WL	<a href="#">521</a>	521	85%	9e-145	91%	
<a href="#">AB273860.1</a>	Uncultured bacterium 16S rRNA, partial	<a href="#">519</a>	519	92%	4e-144	89%	
<a href="#">AB280356.1</a>	Uncultured bacterium gene for 16S ribc	<a href="#">517</a>	517	88%	1e-143	90%	
<a href="#">EF073416.1</a>	Uncultured Firmicutes bacterium clone 1	<a href="#">498</a>	498	68%	1e-137	93%	
<a href="#">EF664035.1</a>	Uncultured Firmicutes bacterium clone 1	<a href="#">460</a>	460	71%	3e-126	91%	
<a href="#">EF692741.1</a>	Uncultured bacterium clone FW026-131	<a href="#">317</a>	317	68%	3e-83	87%	
<a href="#">EU491431.1</a>	Uncultured bacterium clone POX4b3H10	<a href="#">295</a>	347	54%	1e-76	92%	
<a href="#">AM905135.1</a>	Uncultured bacterium partial 16S rRNA	<a href="#">278</a>	278	43%	2e-71	91%	
...							
<a href="#">AJ871741.1</a>	Uncultured planctomycete partial 16S r	<a href="#">244</a>	280	47%	3e-61	95%	
<a href="#">AJ871740.1</a>	Uncultured planctomycete partial 16S r	<a href="#">244</a>	280	47%	3e-61	95%	
<a href="#">AJ871739.1</a>	Uncultured planctomycete partial 16S r	<a href="#">244</a>	280	47%	3e-61	95%	
<a href="#">AJ871738.1</a>	Uncultured planctomycete partial 16S r	<a href="#">244</a>	280	47%	3e-61	95%	
<a href="#">AJ871737.1</a>	Uncultured planctomycete partial 16S r	<a href="#">244</a>	280	47%	3e-61	95%	
<a href="#">AJ871735.1</a>	Uncultured planctomycete partial 16S r	<a href="#">244</a>	280	47%	3e-61	95%	

# EBI improved

PROGRAM: BLASTN

DATABASE: Nucleic Acid (selected), EMBL Release (highlighted), EMBL HTC Plant, EMBL HTG Plant, EMBL Patent Plant, EMBL Standard Plant, EMBL STS Plant, EMBL TPA Plant, EMBL Prokaryote, EMBL EST Prokaryote, EMBL GSS Prokaryote, EMBL HTC Prokaryote, EMBL HTG Prokaryote, EMBL Patent Prokaryote, EMBL Standard Prokaryote (highlighted), EMBL STS Prokaryote, EMBL TPA Prokaryote, EMBL Rodent, EMBL EST Rodent, EMBL GSS Rodent, EMBL HTC Rodent, EMBL HTG Rodent

RESULTS: interactive

SEARCH TITLE: Sequence

YOUR EMAIL: [empty]

MATRIX: default

SCORES: default

VIEW FILTER: [empty]

SENSITIVITY: normal

DNA STRAND: both

STATS: [empty]

topcomboN

FORMAT: Default

Enter or Paste

>NF001  
CTCTCTCTCGC  
TTAACGAGTGAC  
CGTCGCTCCGT  
TCCCGTAGGAG  
CTAAGGCCGGC  
TAGCTGATAGG  
TCCTTTAGTCA  
CCGGTTCCCGC

CCCCAACCCTTTC  
ATCCACACGCGG  
CTCGACTGCAGCC  
GGCCGGACACCCG  
TACCCTCACCAAC  
ATCTTCAGAGGCC  
TCCGGTCTTAGCT  
TATCTATGCATTA

Select the database excluding sequences from the “ENV” division

# EBI improved

EM_PRO:	<a href="#">DQ095862</a> ;	DQ095862	Thermolithobacter carboxydivora...	1118	9.8e-52	2
EM_PRO:	<a href="#">AJ508927</a> ;	AJ508927	Propionispora hippei partial 16...	1260	9.0e-51	1
EM_PRO:	<a href="#">AB308475</a> ;	AB308475	Caldaterra yamamurae gene for 1...	1089	2.0e-50	2
EM_PRO:	<a href="#">AF503917</a> ;	AF503917	Agromyces albus 16S ribosomal R...	811	2.2e-50	2
EM_PRO:	<a href="#">AJ276701</a> ;	AJ276701	Desulfitobacterium frappieri 16...	1052	4.7e-50	2
EM_PRO:	<a href="#">AB250968</a> ;	AB250968	Caldaterra satsumae gene for 16...	1087	6.1e-50	2
EM_PRO:	<a href="#">EF468656</a> ;	EF468656	Arthrobacter <b>sp.</b> TD4 16S riboso...	819	6.7e-50	2
EM_PRO:	<a href="#">EU402968</a> ;	EU402968	Arthrobacter nicotianae strain ...	819	1.2e-49	2
EM_PRO:	<a href="#">EF379937</a> ;	EF379937	Arthrobacter <b>sp.</b> TD2 16S riboso...	819	1.7e-49	2
EM_PRO:	<a href="#">AB062280</a> ;	AB062280	Thermoanaerobacter <b>sp.</b> ToBE gen...	1070	1.8e-49	2
EM_PRO:	<a href="#">L40620</a> ;	L40620	Geodermatophilus obscurus obscurus ...	812	2.1e-49	2
EM_PRO:	<a href="#">DQ119659</a> ;	DQ119659	Planifilum yunnanesis 16S ribos...	1229	2.3e-49	1
EM_PRO:	<a href="#">EU327526</a> ;	EU327526	Arthrobacter <b>sp.</b> ArthroaeroA3 1...	819	2.7e-49	2
EM_PRO:	<a href="#">EU264108</a> ;	EU264108	Arthrobacter <b>sp.</b> A-1 16S riboso...	810	3.0e-49	2
EM_PRO:	<a href="#">U40078</a> ;	U40078	Desulfitobacterium frappieri 16S ri...	1024	3.3e-49	2
EM_PRO:	<a href="#">EU099409</a> ;	EU099409	Arthrobacter <b>sp.</b> WBF35 16S ribo...	819	3.6e-49	2
EM_PRO:	<a href="#">AB088361</a> ;	AB088361	Planifilum fulgidum gene for 16...	1224	3.9e-49	1
EM_PRO:	<a href="#">DQ777749</a> ;	DQ777749	Dehalobacter <b>sp.</b> 1,1-DCA1 16S r...	1042	4.3e-49	2
EM_PRO:	<a href="#">AB262671</a> ;	AB262671	<b>Firmicutes bacterium EG24</b> gene ...	1051	4.4e-49	2
EM_PRO:	<a href="#">AF357919</a> ;	AF357919	Desulfitobacterium <b>sp.</b> Viet-1 1...	1032	4.8e-49	2
EM_PRO:	<a href="#">AM056027</a> ;	AM056027	<b>planctomycete A-2 partial</b> 16S r...	1221	5.4e-49	1
EM_PRO:	<a href="#">AJ551163</a> ;	AJ551163	Arthrobacter ardleyensis partia...	810	5.7e-49	2
EM_PRO:	<a href="#">AJ715981</a> ;	AJ715981	Arthrobacter ardleyensis partia...	810	6.0e-49	2
EM_PRO:	<a href="#">AJ508928</a> ;	AJ508928	Propionispora hippei partial 16...	1051	6.1e-49	2
EM_PRO:	<a href="#">AB098572</a> ;	AB098572	Arthrobacter sp. TUT1004 gene f...	820	6.1e-49	2
EM_PRO:	<a href="#">EF376010</a> ;	EF376010	Arthrobacter sp. TD-1 16S ribos...	810	6.4e-49	2
EM_PRO:	<a href="#">EF376011</a> ;	EF376011	Arthrobacter sp. HWTW-22 16S rib...	810	6.4e-49	2

# Blast on cultured strains

## Blast server for bacterial identification

Specific databases of 16S rRNA sequences for cultured strains only

**Blast** (classic interface from NCBI) **Viroblast** (a blast with more features) **Documentation** (explanation, tutorials and more) **Blast to TreeDyn** (export Blast results to TreeDyn software)

Choose program to use and database to search:

[Program](#)  [Database](#)

Enter sequence below in [FASTA](#)

```
CTAAGGCCGGCTACCCGTCAATGC
TAGCTGATAGGACATAGATCCCTC
TCCTTTAGTCACCGAACCAGGCGA
CCGGTTTCCCGGAGTTATCCCGGT
CTACCCTTCGCACTAACACCCGTA
TGCCTAATCACGCCGCTGGCGTTC
CCGG
```

Or load it from disk

Set subsequence: From

- Bacteria 16S (length > 1200)
- Bacteria 16S (length > 1000)
- Bacteria 16S (length > 800)
- Bacteria 16S (length > 500)
- Bacteria 16S (length > 50)
- Bacteria 16S (2 seq/taxa length > 1200)
- Bacteria 16S (2 seq/taxa length > 1000)
- Bacteria 16S (2 seq/taxa length > 800)
- Bacteria 16S (2 seq/taxa length > 500)
- Archea 16S (length > 1200)
- Archea 16S (length > 1000)
- Archea 16S (length > 800)
- Archea 16S (length > 500)
- Archea 16S (length > 50)
- Sequences Laure

<http://bioinfo.unice.fr/blast/>

Select by minimal length  
Select two sequences only by species

# Blast on cultured strains


Sequences producing significant alignments:	Score (bits)	E Value
U76364  Thermoterrabacterium ferrireducens 17_22_37_35_132_1614 ...	<u>214</u>	7e-55
EF542810  Carboxydotherrmus siderophilus 17_22_37_35_132_304 len=...	<u>214</u>	7e-55
CP000141.3  Carboxydotherrmus hydrogenoformans 17_22_37_35_132_30...	<u>214</u>	7e-55
CP000141.2  Carboxydotherrmus hydrogenoformans 17_22_37_35_132_30...	<u>214</u>	7e-55
CP000141.1  Carboxydotherrmus hydrogenoformans 17_22_37_35_132_30...	<u>214</u>	7e-55
CP000141.0  Carboxydotherrmus hydrogenoformans 17_22_37_35_132_30...	<u>214</u>	7e-55
EF554597  Fervidomicrobium thiophilum 15_37_58_55_67_575 len=1536	<u>206</u>	2e-52
X84257  Corynebacterium glutamicum 1_4_8_150_47_389 len=1412	<u>204</u>	6e-52
X82061  Corynebacterium glutamicum 1_4_8_150_47_389 len=1402	<u>204</u>	6e-52
EU231610  Corynebacterium glutamicum 1_4_8_150_47_389 len=1517	<u>204</u>	6e-52
DQ173748  Corynebacterium glutamicum 1_4_8_150_47_389 len=1473	<u>204</u>	6e-52
DQ173747  Corynebacterium glutamicum 1_4_8_150_47_389 len=1454	<u>204</u>	6e-52
DQ171711  Corynebacterium glutamicum 1_4_8_150_47_389 len=1472	<u>204</u>	6e-52
BX927156.0  Corynebacterium glutamicum 1_4_8_150_47_389 len=1524	<u>204</u>	6e-52
BX927155.1  Corynebacterium glutamicum 1_4_8_150_47_389 len=1524	<u>204</u>	6e-52
BX927155.0  Corynebacterium glutamicum 1_4_8_150_47_389 len=1524	<u>204</u>	6e-52
BX927152.0  Corynebacterium glutamicum 1_4_8_150_47_389 len=1524	<u>204</u>	6e-52
BX927150.0  Corynebacterium glutamicum 1_4_8_150_47_389 len=1525	<u>204</u>	6e-52
BX927148.0  Corynebacterium glutamicum 1_4_8_150_47_389 len=1524	<u>204</u>	6e-52

The taxonomy bar-code :

- 17 Bacteria; Firmicutes;
- 15 Bacteria; Fervidomicrobium.
- 1 Bacteria; Actinobacteria

# Blast on type strains

**EzTaxon.org Server**  
 Easiest way to the accurate identification of prokaryotes


















account | [logout](#) 

HOME SEARCH SIMTABLE RESULTS BATCH SEQLIST UTIL HOW TO CITE FAQ LINKS

BLAST Results of gb|U89823|Rhizobium sp. USDA 1920 USDA 1920

<http://210.218.222.43:8080>

Exclude "unclutured" sequences

Count	Name/Title 	Type strain in DB	Accession	BLAST score 	BLAST e val 	Pairwise Similarity 
1	<i>Arthrobacter viscosus</i> 	LMG 16473T	AJ639832	2498	0	99.57
2	<i>Rhizobium sllae</i> 	DSM 14623T	Y10170	2399	0	99.27
3	<i>Rhizobium gallicum</i> 	R602spT	U86343	2453	0	98.66
4	<i>Rhizobium mongolense</i> 	USDA 1844T	U89817	2446	0	98.59
5	<i>Rhizobium indigoferae</i> 	CCBAU 71042T	AF364068	2432	0	98.51
6	<i>Rhizobium leguminosarum</i> 	IAM 12609T	D12782	2444	0	98.44
7	<i>Rhizobium etli</i> 	CFN 42T	U28916	2426	0	98.24
8	<i>Rhizobium trifolii</i> 	ATCC 14480T	AY509900	2425	0	98.17
9	<i>Rhizobium yanglingense</i> 	SH 22623T	AF003375	2403	0	98.16
10	<i>Sinorhizobium americanum</i> 	CFNEI 156T	AF506513	2333	0	97.24
11	<i>Ensifer kostiensis</i> 	LMG 19227T	AM181748	2296	0	97.15
12	<i>Ensifer fredii</i> 	ATCC 35423T	D14516	2315	0	97.10
13	<i>Ensifer adhaerens</i> 	LMG 20216T	AM181733	2318	0	97.01

This Blast does not take parameters



# Blast 2 TreeDyn

**Blast2TreeDyn** <http://bioinfo.unice.fr/blast>

- Paste sequences from the blast results (select text and Ctrl+C) into this box (Ctrl + V):

```
AY794056 | Corynebacterium  
glutamicum|1_4_8_150_47_389|len=1472  
204 6e-52
```

Add to the list

- Or, upload a file with blast result:  Parcourir...

List of sequences:

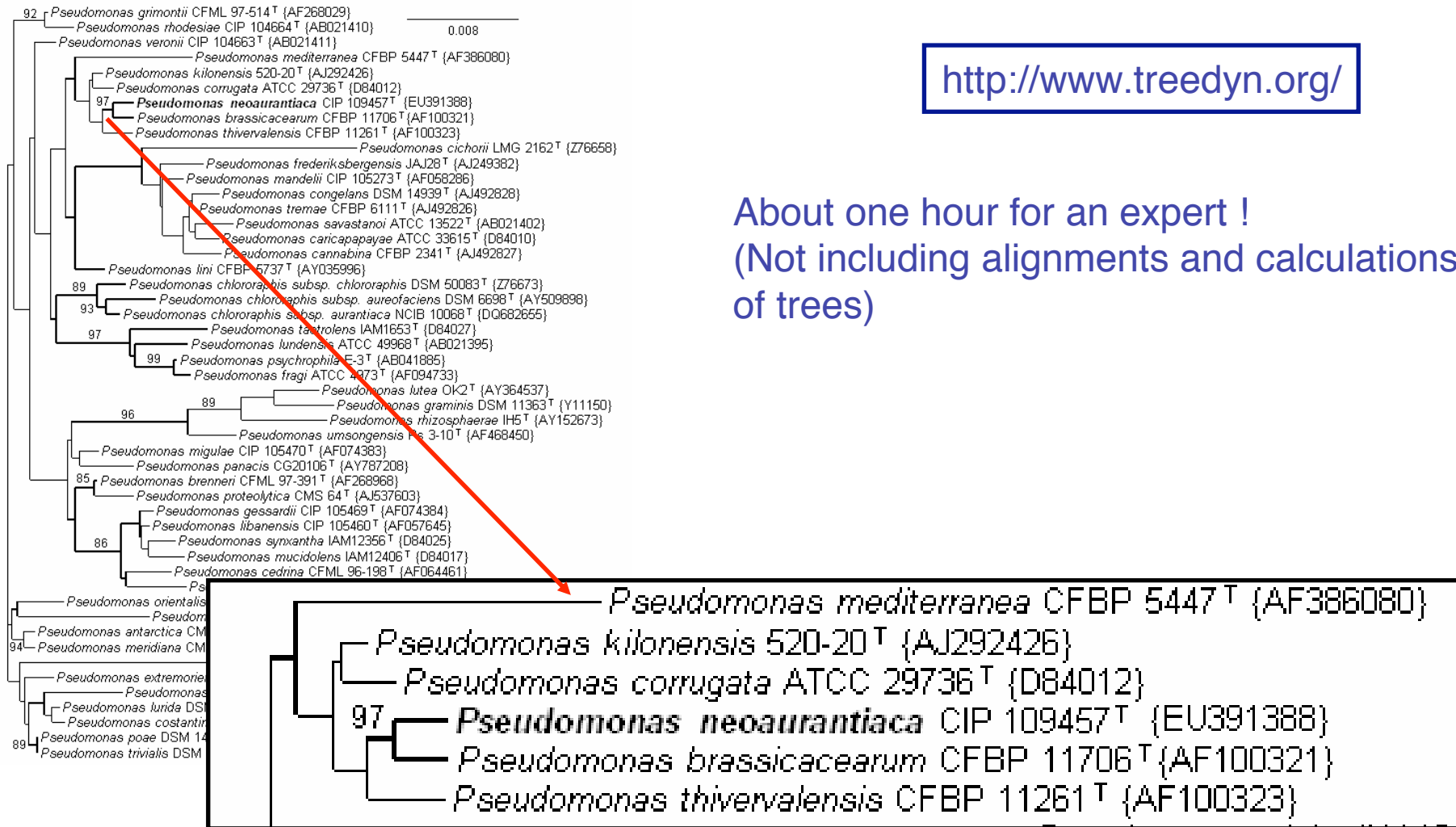
Change accession identifiers to incremental identifiers

Download sequences and **annotations**

# Clustal - Phylip - TreeDyn

<http://www.treedyn.org/>

About one hour for an expert !  
(Not including alignments and calculations of trees)



Ready for publication !

# Identify 16S rRNA sequences: LOL ?

16 RRNA  
16 RRNA GENE

16S GENE  
16S R-RNA  
16S RDNA

16S RBOSOMAL RNA  
16S RIBBOSOMAL DNA  
16S RIBISOMAL RNA  
16S RIBOBOMAL DNA  
16S RIBOBSOMAL RNA  
16S RIBOOSMAL RNA  
16S RIBOOSOMAL DNA  
16S RIBOOSOMAL RNA  
16S RIBOSAMAL RNA

16S RIBOSMAL RNA  
16S RIBOSMOMAL RNA  
16S RIBOSOAMAL RNA  
16S RIBOSOAML RNA  
16S RIBOSOMAL

16S RIBOSOMAL DNA  
16S RIBOSOMAL GENE  
16S RIBOSOMAL RNA

16S RIBOSOMIAL RNA  
16S RIBOSOML RNA  
16S RIBOSOMMAL RNA  
16S RIBOSONAL RNA  
16S RIBOSORMAL RNA  
16S RIBOSOSMAL RNA

16S RIBSOMAL RNA

16S RIBSOSOMAL RNA

16S RIBOSOMAL RNA [  
16S ROBOSOMAL RNA  
16S ROSOMAL DNA

16S RNA  
16S RNA GENE

16S RNRA

16S RRNA

16S (LSU) RIBOSOMAL RNA

16S LARGE RIBOSOMAL

16S LARGE SUBUNIT RIBOOSMAL

16S LARGE SUBUNIT RIBOSOMAL RNA (2,986 entries)

?

# Tasks and problems

- Identification of a new isolate: the 16S “gold standard”.
- Other genes.
- Typing a strain.
- Studying biodiversity: new approaches.

# MLSA

gene	sequences	gene product
nifh	9530	alpha subunit of dinitrogenase reductase
gyrb	7080	subunit b protein of dna gyrase
rpob	6629	rna polymerase beta subunit
reca	5052	homologous recombination factor reca
amoA	4226	ammonia monooxygenase subunit a
nirS	3406	cytochrome cd1 nitrite reductase
nirK	3272	dissimilatory nitrite reductase
pmoA	2668	particulate methane monooxygenase subunit a
dsrA	2186	dissimilatory sulfite reductase subunit alpha
atpA	2182	proton-translocating atpase alpha subunit
cpn60	2173	60 kda chaperonin
gyrA	2168	dna gyrase a subunit
nosZ	2018	nitrous oxide reductase
mdh	2004	malate dehydrogenase
<b>gene product</b>		
hypothetical protein	331537	
conserved hypothetical protein	229436	
unknown	14976	
putative membrane protein	12359	
transcriptional regulator, lysr family	7599	
abc transporter related	7221	
unknown protein	6674	
transposase	6550	
dinitrogenase reductase	5191	

Multi Locus Sequence Analysis : most sequenced genes and gene products.

# MLSA *Vibrios*

Domains sequenced for *rpoB*, *gyrB*, and *recA* gene sequences (*Vibrios*)

Length	lb	rb	Number
<b>rpoB</b>			
1400	0	1400	2
1370	30	1400	3
400	700	1100	15
400	400	800	1
300	700	1000	2
<b>300</b>	<b>400</b>	<b>700</b>	<b>65</b>
200	1100	1300	1
100	1100	1200	9
100	600	700	1
<b>gyrB</b>			
800	0	800	5
410	90	500	47
400	100	500	4
380	20	400	2
370	30	400	1
340	60	400	16
310	90	400	1
300	100	400	69
<b>200</b>	<b>200</b>	<b>400</b>	<b>212</b>
200	100	300	162
100	200	300	1

<b>recA</b>			
400	0	400	1
360	40	400	1
350	50	400	7
340	60	400	6
330	70	400	8
240	60	300	1
230	70	300	94
220	80	300	141
210	90	300	47
<b>200</b>	<b>100</b>	<b>300</b>	<b>300</b>
130	70	200	63
120	80	200	4
100	100	200	22

This table is sorted by decreasing sequence lengths (column 1), lb and rb correspond to left and right boundaries of sequences in alignments (sequence starts aligning at or after lb position and ends before rb), number is the number of sequences in each category (*rpoB* 99 sequences, 26 species; *gyrB* 520 sequences, 59 species; *recA* 695 sequences, 58 species). In bold are the most sequenced domains.

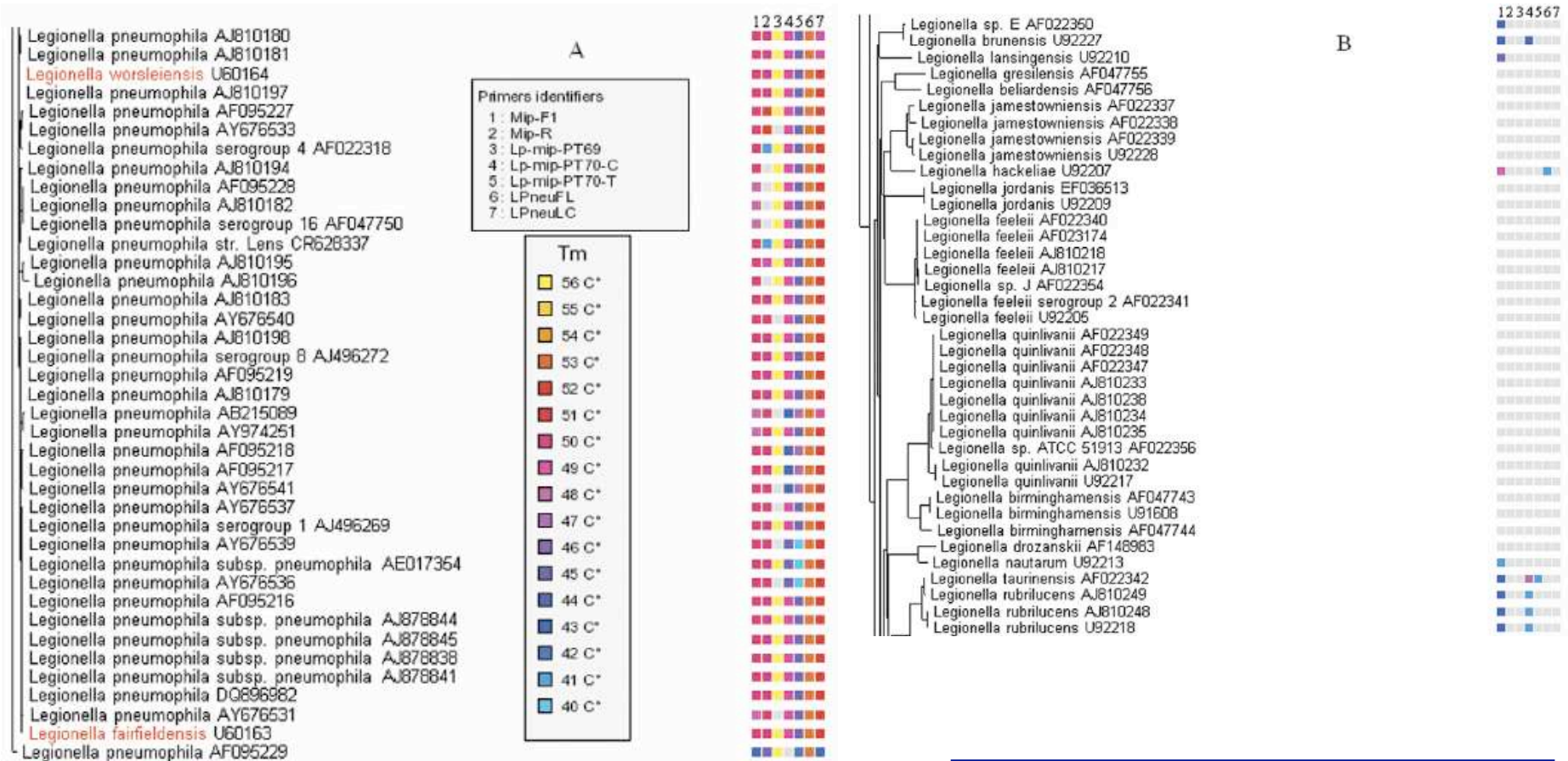
Identifications of pathogens—a bioinformatic point of view  
Richard Christen

Current Opinion in Biotechnology 2008, 19:1–8

Mostly short PCR sequences !

# Using a pathogenicity gene as target

Analyses of 2006 publications !



*Legionella pneumophila*: the mip gene.

Identifications of pathogens—a bioinformatic point of view

Richard Christen

URL: <http://bioinfo.unice.fr/ohm>

Current Opinion in Biotechnology 2008, 19:1-8

# Using a pathogenicity gene as target

Evaluation of primers and probes recently used for the identification of the <i>mip</i> genes in <i>Legionella</i>		
Primer/Tm	Sequence variant	Number of sequences
Mip-F1	tttgtattgcaaaccacttggc	F
50.32	.....	130
48.05	.....C.....	5
47.74	.....g.....	6
45.4	.....t...	4
42.52	.....C.....t...	14
42.5	a.....t...	4
37.14	.....C..CC.....t...	15
34.05	a.....g..t..g	16
22.4	a..C..C..ac.t.....	10
		277
Mip-R1	ctcgacagtgactgtatccgattt	R
52.26	.....	3
49.79	t.....	122
31.44	t..a.....t.....g...c...	9
26.69	t..c.....c.....g.....	6
21.29	...a.....c..g.....atc...	9
21.09	t..a.....t..a.....t.....	14
16.44	...a..c.....c..g..atcc..	11
15.09	.....g..t..c..g..ttc...	12

Lp-mip-	Sequence variant	Number of sequences
PT69	ccaaatcggcaccaatgc	F
55.76	.....	129
32.96	.t.....t.....	16
32.68	.....c..t.....a.	7
22.34	.t.....t.....a.	59
3.27	.t.....g..t.....a.	15
LPneuLC	ccattgcttcggattaacatctatgcc	R
50.85	.....	138
49.37	.....g.....	3
24.13	.....g.g.....t..a..t..	10
20.52	.....cgaa...c.g.tt..g....	13
17.86	.t.a..g.....t..a..a..	15
11.77	.....agaa...c.g.tt..a..c..	12
0	.t.a.a.g..t..g....tg..a.....	10

For each oligomer: column (1) Tm in °C estimated for each *mip* sequence variant; column (2) the variant sequence; column (3) the number of such sequences (about 270 *mip* sequences available, only excerpts shown). F: forward primer, R: reverse primer.

Wrong primer used in publications of year 2006 !

Identifications of pathogens—a bioinformatic point of view  
Richard Christen

Current Opinion in Biotechnology 2008, 19:1–8

# Tasks and problems

- Identification of a new isolate: the 16S “gold standard”.
- Other genes.
- Typing a strain.
- Studying biodiversity: new approaches.

# Use tandem repeat sequences

**G P M S**

**Genomes, Polymorphism and Minisatellites** <http://minisatellites.u-psud.fr>

with a focus on Molecular Epidemiology using Tandem Repeats

## VNTRDB: a bacterial variable number tandem repeat locus database

Chia-Hung Chang,<sup>1,2</sup> Yu-Chung Chang,<sup>3</sup> Anthony Underwood,<sup>4</sup> Chien-Shun Chiou,<sup>5</sup> and Cheng-Yan Kao<sup>1,6\*</sup>

Nucleic Acids Res. 2007 January; 35(Database issue): D416–D421.  
Published online 2006 December 14. doi: 10.1093/nar/gkl872.

## Characterization of a Trinucleotide Repeat Sequence (CGG)<sub>n</sub> and Potential Use in Restriction Fragment Length Polymorphism Typing of *Mycobacterium tuberculosis*

J Clin Microbiol. 2004 August; 42(8): 3538–3548.  
doi: 10.1128/JCM.42.8.3538-3548.2004.

### Comparative Genomics

**Editor(s):** Nicholas H. Bergman<sup>1</sup>

**Affiliation(s):** (1)Bioinformatics Program and Department of Microbiology and Immunology, University of Michigan Medical School, Ann Arbor MI

**Series:** Methods in Molecular Biology | **Volume No.:** 396

**Print ISBN:** 978-1-934115-37-4

### Variable Number Tandem Repeat Typing of Bacteria

**By:** Siamak P. Yazdankhah<sup>2</sup> [Contact Information](#), Bjørn-Arne Lindstedt<sup>2</sup>

#### Variable Number Tandem Repeat Typing of Bacteria

Book Series	Methods in Molecular Biology™
ISSN	1064-3745
Volume	Volume 396
Book	Comparative Genomics
Éditeur	Humana Press
DOI	10.1007/978-1-59745-515-2

Tracing isolates of bacterial species by multilocus variable number of tandem repeat analysis (MLVA)

VAN BELKUM Alex (1) ;

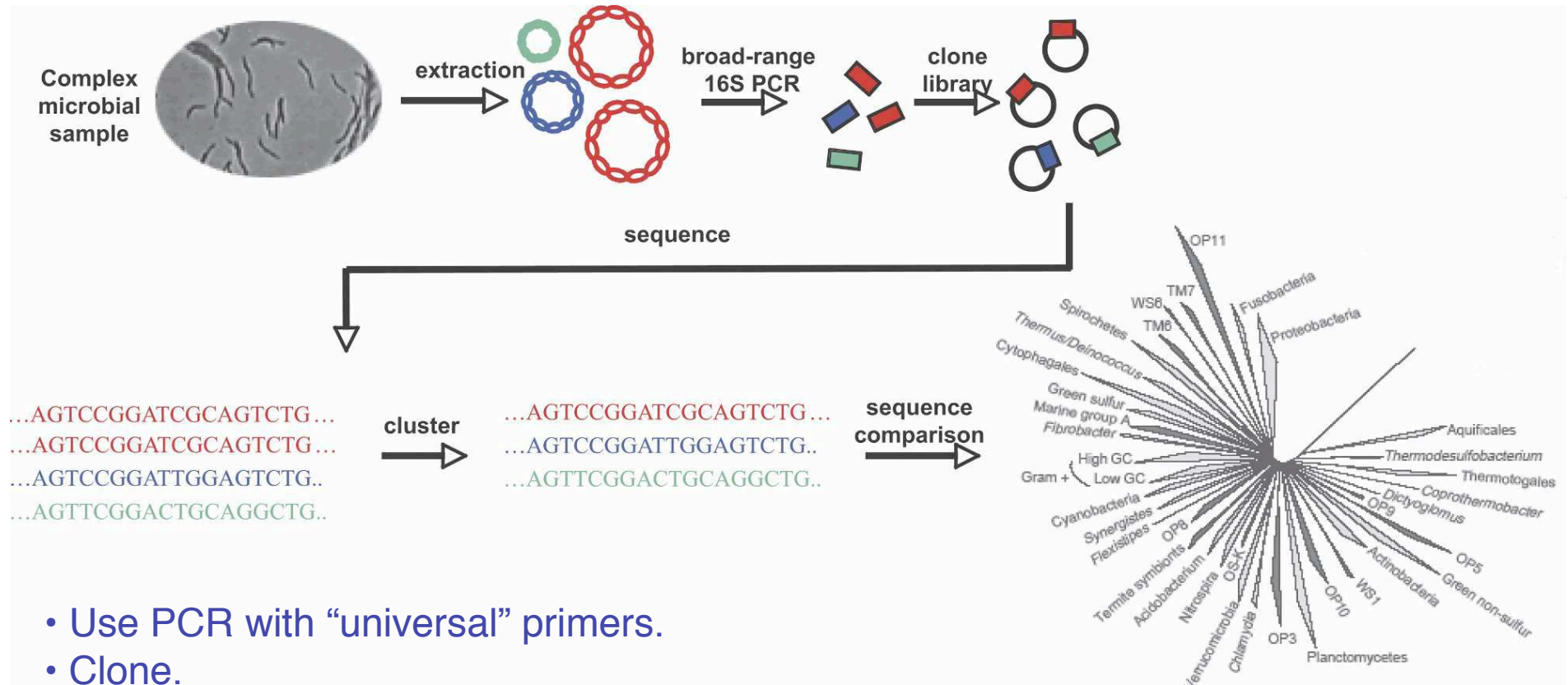
FEMS immunology and medical microbiology **ISSN** 0928-8244

2007, vol. 49, no1, pp. 22-27

# Tasks and problems

- Identification of a new isolate: the 16S “gold standard”.
- Other genes.
- Typing a strain.
- Studying biodiversity: new approaches.

# The “classic” approach



- Use PCR with “universal” primers.
- Clone.
- Random sequence ... 200 clones.

*Genome Res.* 2006 16: 316-322

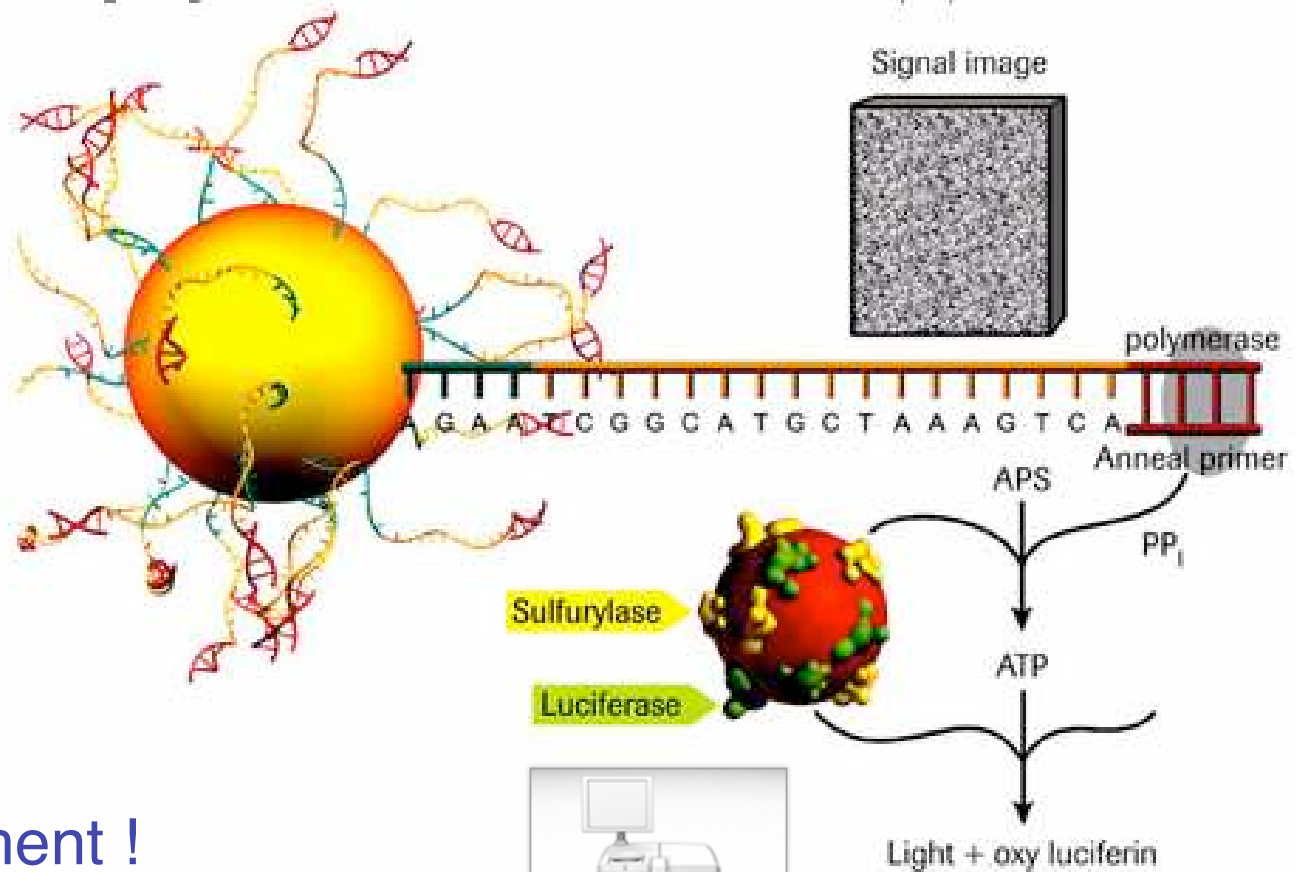
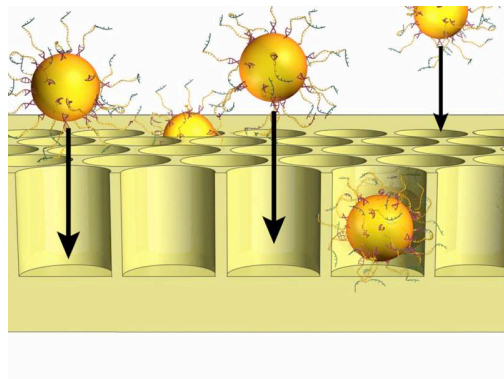
# Biodiversity analyses - classic

PMID	Short title	entries	year
18043639	<i>Pyrosequencing enumerates and contrasts soil microbial diversity...</i>	90110	2008
17183309	<b>Microbial ecology: human gut microbes associated with obesity...</b>	18348	2007
17699621	<b>Molecular-phylogenetic characterization of microbial community...</b>	15172	2007
15831718	<b>Diversity of the human intestinal microbial flora...</b>	11831	2005
18252821	<b>Symbiotic gut microbes modulate human metabolic phenotypes...</b>	7255	2008
17055441	<b>Reciprocal Gut Microbiota Transplants from Zebrafish and Mice to...</b>	5534	2006
16033867	<b>Obesity alters gut microbial ecology...</b>	3883	2005
17409203	<b>Loss of Bacterial Diversity During Antibiotic Treatment of...</b>	3278	2007
18077362	<b>Molecular identification of bacteria in bronchoalveolar lavage...</b>	3198	2007
17760501	<b>Salmonella enterica serovar typhimurium exploits inflammation to...</b>	2897	2007
18218029	<b>Elevated atmospheric CO2 affects soil microbial diversity...</b>	2269	2008
16741115	<b>Metagenomic analysis of the human distal gut microbiome...</b>	2062	2007
17981945	<b>Short-term temporal variability in airborne bacterial and fungal...</b>	1966	2008
17041161	<b>Community structure analyses are more sensitive to differences in...</b>	1904	2006
16689872	<b>Comparison of prokaryotic diversity at offshore oceanic locations...</b>	1789	2006
18059491	<b>Subsurface clade of Geobacteraceae that predominates in a diversity...</b>	1781	2008
16033867	<b>Obesity alters gut microbial ecology...</b>	1692	2007
16672518	<b>Unexpected diversity and complexity of the guerrero negro...</b>	1587	2006
17124165	<b>Effect of bowel preparation and colonoscopy on post-procedure...</b>	1319	2007
18033299	<b>Metagenomic and functional analysis of hindgut microbiota of a...</b>	1252	2007
15505215	<b>The gut microbiota as an environmental factor that regulates fat...</b>	1206	2007
15070763	<b>Gnotobiotic zebrafish reveal evolutionarily conserved responses to...</b>	1179	2004
18205817	<b>Differences in vegetation composition and plant species identity...</b>	1075	2008
18328082	<b>Microbial community succession and bacterial diversity in soils...</b>	1055	2008

PCR – clone - sequence : too tedious for most labs !

# High-throughput sequencing

- High-throughput sequencing technologies are intended to lower the cost of sequencing DNA libraries
- Many of the new high-throughput methods use methods that parallelize the sequencing process, producing thousands or millions of sequences at once.



No cloning !  
One day experiment !



FIGURE 1

# Advantages and Disadvantages

- 454 Sequencing runs at 20 megabases per 4.5-hour run (1 day: from sampling to sequences).
- G-C rich content is not as much of a problem.
- Unclonable segments are not skipped.
- Detection of mutations in an amplicon pool at a low sensitivity level.
  
- Each read of the GS20 is only 100 base pairs long (2005-2006);
- The new FLX system does 200-300 base pairs (2007)
- 454 has said they expect 500 in '08.

# Biodiversity, case studies

- Huber, J. A., D. B. Mark Welch, et al. (2007). "Microbial population structures in the deep marine biosphere." *Science* 318(5847): 97-100.
- Sogin, M. L., H. G. Morrison, et al. (2006). "Microbial diversity in the deep sea and the underexplored "rare biosphere"." *Proc. Natl. Acad. Sci. U S A* 103(32): 12115-20.
- Roesch, L. F., R. R. Fulthorpe, et al. (2007). "Pyrosequencing enumerates and contrasts soil microbial diversity." *ISME J.* 1(4): 283-90.

## Possible variable domains in the 16S rRNA gene sequences

Table 1: 16S variable region range definitions.

Variable region	<i>E. coli</i> 16S rDNA range			5' primer	3' primer
	start	end	length		
V1	8	120	113	5'-AGAGTTTGATCMTGGCTCAG	5'-TACTCACCCGTICGCCRCT
V2	101	361	261	5'-AGYGGCGIACGGGTGAGTAA	5'-CYIACTGCTGCCTCCCGTAG
V3	338	534	197	5'-ACTCCTACGGGAGGCAGCAG	5'-ATTACCGCGGCTGCTGG
V4	519	806	288	5'-TGCCAGCAGCCGCGGTAA	5'-GGACTACARGGTATCTAAT
V5	787	926	140	5'-ATTAGATACCYTGTAGTCC	5'-CCGTCAATTCMTTGTAGTTT
V6	907	1073	167	5'-AAACTCAAAGAATTGACGG	5'-ACGAGCTGACGACARCCATG
V7 & V8	1054	1406	353	5'-CATGGYTGTCGTCAGCTCGT	5'-ACGGGCGGTGTGTAC
V9	1392	1507	116	5'-GTACACACCGCCCGT	5'-TACCTTGTTACGACTT

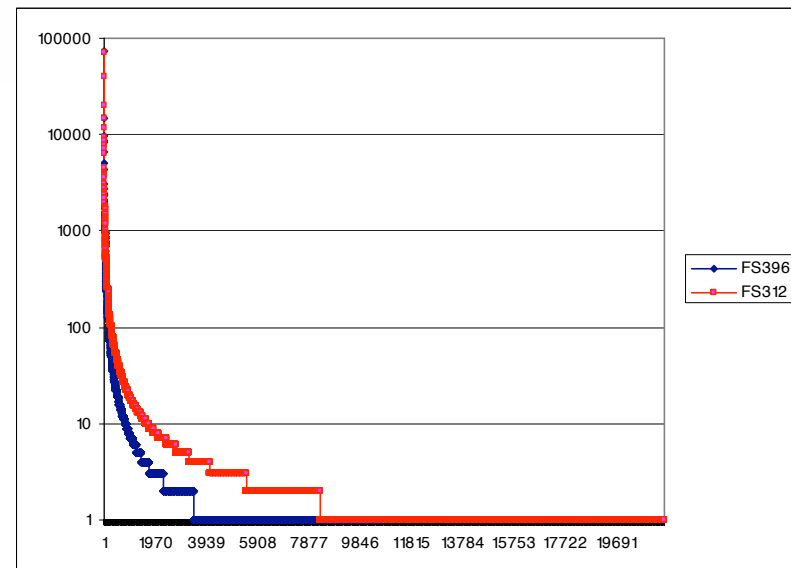
Regions were chosen to be mostly non-overlapping, each containing one or two variable regions. Coordinates are given relative to the 1542 bp *E. coli* K12 16S rDNA sequence.

# Tag dereplication

total number of tags : 442062  
total number of distinct tags : 21529  
number of seconds for analysis : 0.983651788507  
number of single copy tags : 13251

TGGTCTTGACATAGAAAGAACTTTCCAGAGATGGATTGGTGCCTGCTTGCAGGAGCTTTCATAC 70985  
AACTCTTGACATCCAGAGAAGAGGCTAGAGATAGCTTTGTGCCTTCGGGAACTCTGAGAC 40582  
ATCCCTTGACATCCTGCGAACTTTCTAGAGATAGATTGGTGCCTTCGGGAACGCAGTGAC 20128  
AGCACTTGACATAACAACGAACTCGTCAGAGATGACTTGGTGCCTTCGGTGGAAACGTTGATAC 14936  
TGGCCTTGACATGCAGAGAACTTTCCAGAGATGGATTGGTGCCTTCGGGAACTCTGACAC 11751  
AACCTTGACATGGAAAGTATGGATTGTGGAGACACTTTCCTTCAGTTCGGCTGGCTTTCACAC 9350  
TACTCTTGACATCCTGCGAACTTTCGAGAGATCGATTGGTGCCTTCGGGAACGCAGAGAC 8699  
TACTCTTGACATCCAGTGAAGTATAGCAGAGATGCTTTGGTGCCTTCGGGAACACTGAGAC 8603  
AGCCCTTGACATCCTCGGAACTTTCTAGAGATAGATTGGTGCCTTCGGGAGCCGAGTGAC 7779  
AACCTTGACATCCCTATCGCGATTTCAGAGATGGATATCATCAGTTCGGCTGGATAGGTGAC 7613

complete analysis in seconds : 1.04010820515



# Clustering tags into OTU

- Usual manner : align (Muscle), compute distances, phylogeny or cluster.
- Better : cluster according to words frequencies (ex. words of 5 nt)
  - No alignment
  - Much faster
  - Much better

Total calculation time : 7 minutes

clustering 21 529 sequences		
clustering 1 349 533 nt		
longest tag 100; shortest tag 50		
simil %	nbr clusters	
80	3 175	
85	3 931	
88	4 956	
90	5 683	
92	6 670	
95	9 690	
96	8 223	
97	21 529	
98	11 246	
99	13 008	

# Assign each tag to a taxon

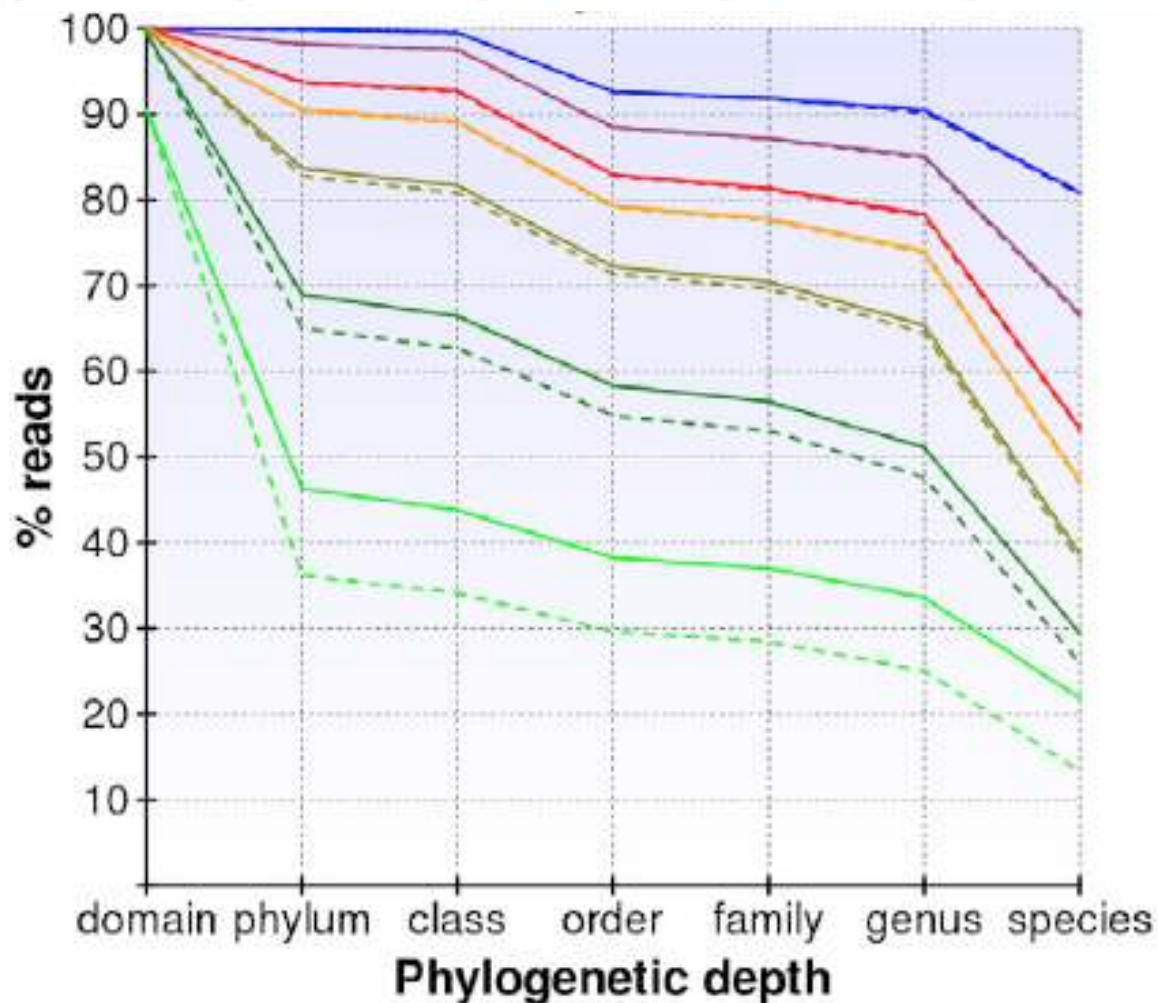
- **GreenGenes**. The greengenes web application provides access to a 16S rRNA gene sequence alignments for browsing, blasting, probing, and downloading. URL: <http://greengenes.lbl.gov>
- **RDP**. The Ribosomal Database Project provides ribosome related data services, including online data analysis, rRNA derived phylogenetic trees, and aligned and annotated rRNA sequences. URL: <http://rdp.cme.msu.edu/>
- **SILVA**. SILVA provides comprehensive, quality checked and regularly updated databases of aligned small and large rRNA sequences for all three domains of life (*Bacteria*, *Archaea* and *Eukarya*). URL: <http://www.arb-silva.de/>

Assignments done using first hit of blast.



# Assign each tag to a taxon

*Placed*    — 30 bp    — 60 bp    — 100 bp    — 150 bp    — 200 bp    — 400 bp    — 800 bp  
*True*        - - 30 bp    - - 60 bp    - - 100 bp    - - 150 bp    - - 200 bp    - - 400 bp    - - 800 bp



Simulated resolution at increasing read-lengths

*BMC Microbiology* 2007, 7:108

# Numbers of 16S rRNA sequences per species

	>800 nt			>1000 nt			>1200 nt		
nbrseq	genera	species		genera	species		genera	species	
1	582	4060		589	4118		592	4126	
2	250	1436		245	1427		239	1411	
3	131	802		133	794		126	790	
4	91	444		88	445		94	454	
5	76	296		75	288		77	277	
6	51	201		53	190		48	178	
7	40	136		38	135		38	143	
8	38	124		37	119		41	110	
9	32	94		36	93		34	87	
10	21	82		22	82		19	82	
10<n<51	40	39		40	40		39	40	
50<k<101	36	32		35	30		33	31	
>100	67	31		62	28		61	27	

- Most species are known from a single sequence !
- ➔ Tags taxonomic specificities are over-evaluated.
  - ➔ Most species have not been sequenced at all.

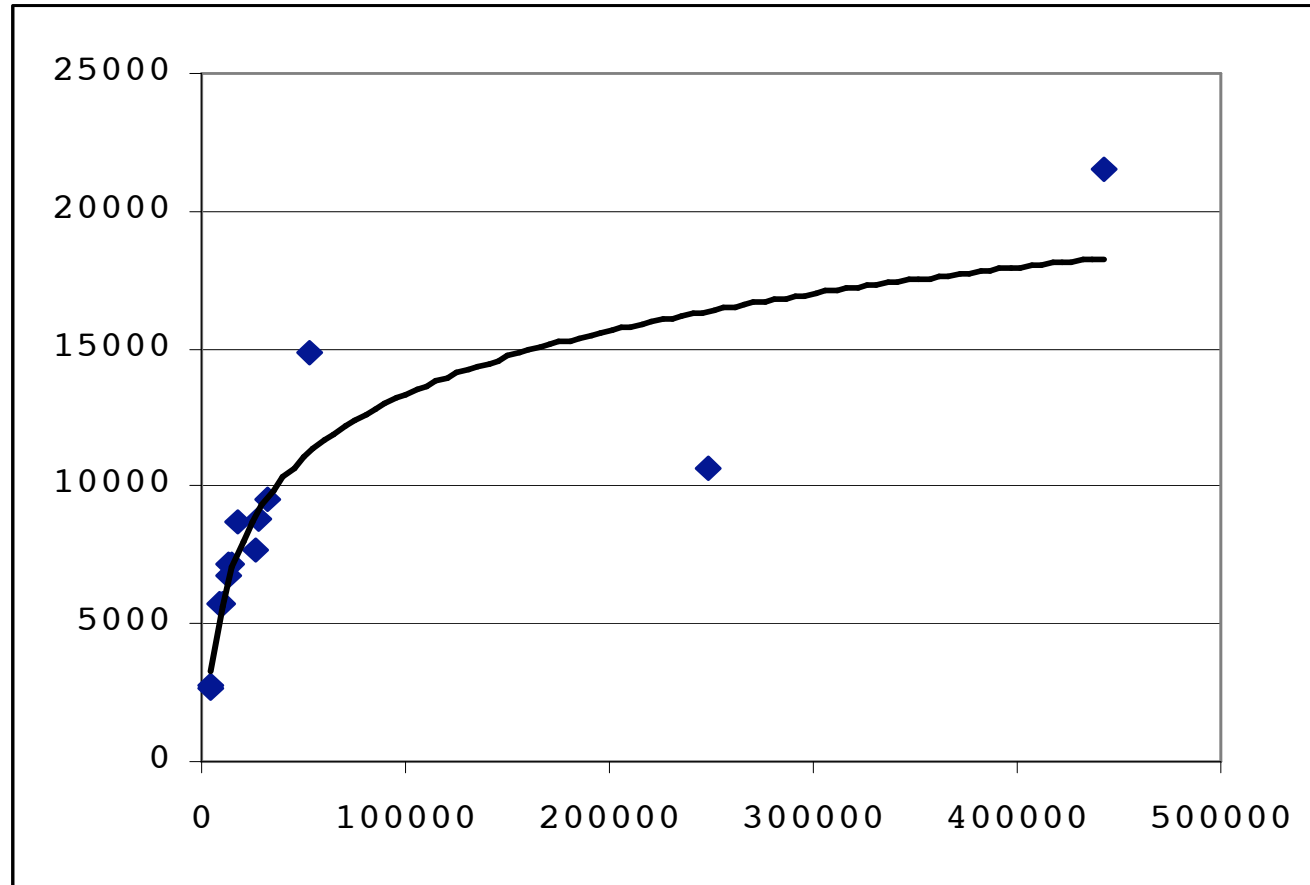
# Main taxa that were **not** amplified

Sogin	numbers	%	Roesch	numbers	%
<b>candidate division ZB3</b>	11	<b>100</b>	<b>candidate division ZB3</b>	11	<b>100</b>
candidate division SR1	10	<b>90</b>	<b>Fibrobacteres</b>	759	<b>86</b>
<b>Fibrobacteres</b>	754	<b>85</b>	candidate division SR1	9	<b>81</b>
<b>Thermodesulfobacteria</b>	84	<b>77</b>	<b>Thermodesulfobacteria</b>	80	<b>74</b>
Thermotogae	108	<b>72</b>	<b>Aquificae</b>	623	<b>63</b>
<b>Aquificae</b>	676	<b>68</b>	<b>Spirochaetes</b>	1774	<b>52</b>
<b>Spirochaetes</b>	1965	<b>58</b>	<b>candidate division OD1</b>	64	<b>51</b>
<b>candidate division OD1</b>	64	<b>51</b>	candidate division BRC1	12	<b>50</b>
Deferribacteres	62	44	Thermotogae	73	48
candidate division TG3	32	39	candidate division GN1	10	45
Deinococcus-Thermus	252	36	candidate division TG3	32	39
candidate division TM6	17	35	candidate division TG1	107	38
candidate division TG1	91	32	candidate division KSB1	13	36
candidate division TM7	40	32	candidate division OP11	60	31
candidate division OP5	13	32	candidate division OP5	12	30
candidate division OP11	61	31	candidate division OP10	35	28
candidate division OP10	38	31	candidate division WS6	33	28
<b>Firmicutes</b>	<b>15284</b>	<b>30</b>	candidate division WS3	14	28
candidate division WS6	33	28	Deinococcus-Thermus	188	27
Bacteroidetes	4556	22	candidate division TM7	32	25
Chloroflexi	524	22	Deferribacteres	34	24
candidate division JS1	10	21	<b>Ktedonobacteria</b>	<b>11</b>	<b>24</b>
environmental samples	83936	20	<b>Actinobacteria</b>	<b>7356</b>	<b>22</b>
candidate division WWE3	18	20	<b>Proteobacteria</b>	<b>25214</b>	<b>21</b>
Fusobacteria	167	19	Chloroflexi	500	21

Primers need to be better designed !

# New tags as a function of sequencing effort

## Saturation curve



MPS will sequence every PCR product present.  
But has PCR amplified every gene present in sample ?

# Conclusions

- Identification using **16S rRNA gene sequences** is now easy.
- **MLSA**: there is a lack of complete sequences to evaluate published primers.
- **MPS on 16S**:
  - Lack of complete sequences to evaluate primers (almost complete seqs).
  - A single sequence available for a majority of species.
  - Most sequences have a poorly annotated taxonomy.
    - 112,509 (**16.8 %**) only of the 670,401 bacterial 16S rRNA gene sequences of length >100 nt presently deposited have a taxonomic description **down to the genus level**, while 383,570 sequences (57 %) have "environmental samples" as sole description.
    - OS...uncultured.bacterium¶  
OC...Bacteria;.environmental.samples.¶
  - MPS technologies have **not been validated** against samples of known compositions.
  - MPS machines **are not calibrated** before, during or after a run.
  - MPS experiments to estimate diversity **are not reproduced** (duplicated) !
  - Primers have to be improved.
  - Degenerated primers should NOT be mixed (competition).

# Final conclusions

- MPS technologies enable researchers to undertake the process of global sequencing in a **single operation** using bench-top instruments.
- The term ‘post-genomics’ has been prematurely coined and we are in fact on the beginning of a **global sequencing era**, which opens a long journey that will occupy a broad spectrum of the scientific community for decades.
- **Global (pyro)sequencing will replace any other method for estimating biodiversity.**
- **Global (pyro)sequencing will replace any other method in transcriptome studies.**

**16S rRNA sequences do not offer a direct link to phenotypes or functions.**

- A wide and generalized sequencing effort and **ontology** building of **well-identified strains** deposited in collections worldwide is required to form the basis of derived annotations of environmental sequences.
- Developing ecosystem predictive models is fundamental, but this is still a long-term objective, as **connection of taxonomy to functions is still missing in most cases.**