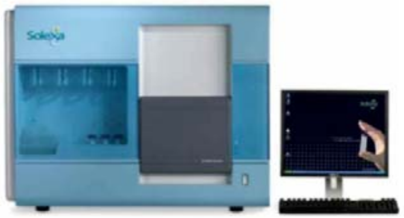
The Massive Parallel Sequencing era:

"Global sequencing"

Richard Christen
CNRS UMR 6543 & Université de Nice
christen@unice.fr
http://bioinfo.unice.fr

Applied Biosystems
ABI 3730XL
1 Mb / day

Roche / 454
Genome Sequencer FLX
100 Mb / run



Illumina / Solexa
Genetic Analyzer
2000 Mb / run

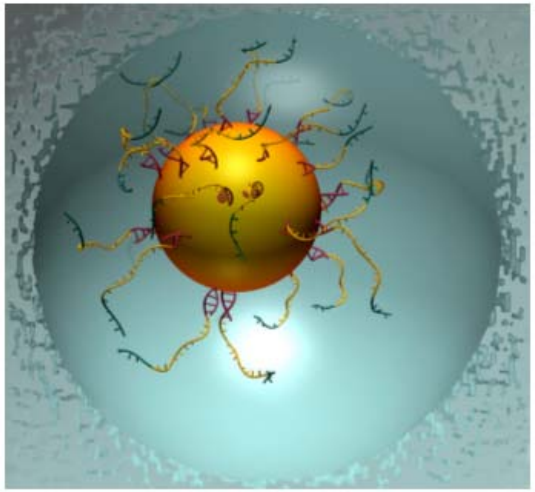
Applied Biosystems
SOLiD
3000 Mb / run

At the end of 2007, three next-generation sequencing platforms appeared: Roche/454's Genome Sequencer FLX (which succeeded a first model), Illumina's Genome Analyzer; and Applied Biosystems's SOLiD sequencer.

In many applications they will replace the "old Sanger" technology (ABI 3730XL)

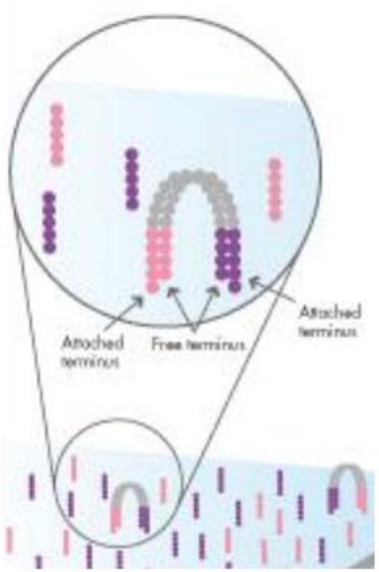
Roche / 454 : GS FLX

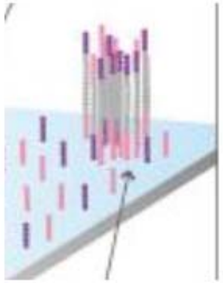
- Real Time Sequencing by Synthesis
- Chemiluminescence detection in pico titer plates
- Amplification: emulsion PCR
- Pyrosequencing
- up to 400,000 reads / run
- on average 250 bases / read
- up to 100 Mb / run



Illumina / Solexa: Genetic Analyzer

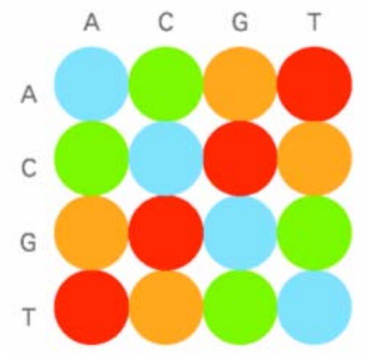
- Real Time Sequencing by Synthesis
- Clonal Single Molecule Array
- Amplification: bridging PCR
- 60 mio reads / run
- up to 50 bases / read
- 2 Gb / run
- 8 channels, app. 5 mio reads / channel
- Fluorescent labels
- Reversible 3‘OH blocking



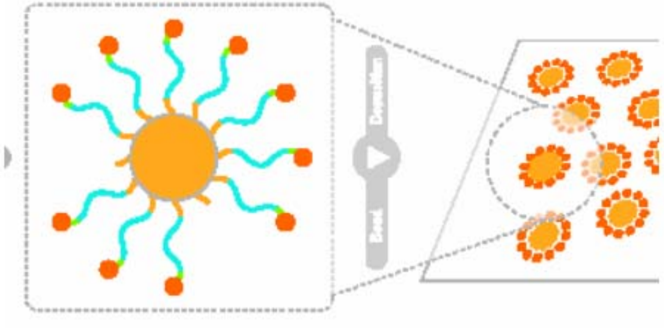


SOLiD system

- Real Time Sequencing by Ligation
- Emulsion PCR and Beads on slides
- 85 mio reads / run
- Up to 35 bases/read
- 3 Gb / run
- dual fluorescent labels
- 8 individual channels / flowcell
- 2 flowcells / run



	A	C	G	T
A	🔵	🟢	🟠	🔴
C	🟢	🔵	🔴	🟠
G	🟠	🔴	🔵	🟢
T	🔴	🟠	🟢	🔵



“The capacity and throughput of the 454 FLX system is quite similar to the Solexa system, if one can afford to run it twice a day”.

If run at maximum capacity, per year :

- consumes about **5,3 millions €** ,
- generates about **75 gigabases** of data.

➔Lower the cost of sequencing DNA.

➔Simplify the sequencing process (no cloning).

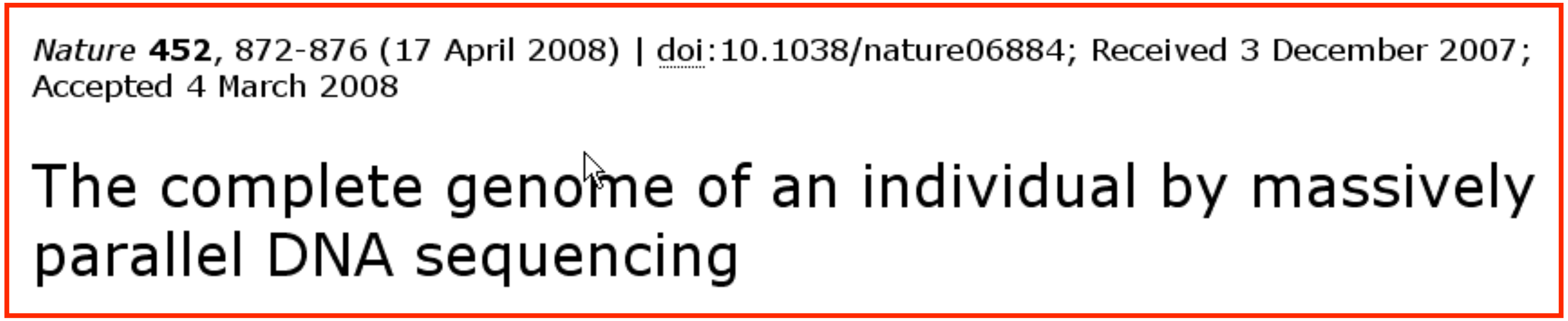
➔Produce hundreds of thousands or millions of sequences at once.

Tasks and problems

- Genomes
  - Resequencing genomes.
  - De novo sequencing a genome.
- Transcriptomes.
- Biodiversity.
  - SSU rRNA sequences
  - Metagenomes

Resequencing a genome

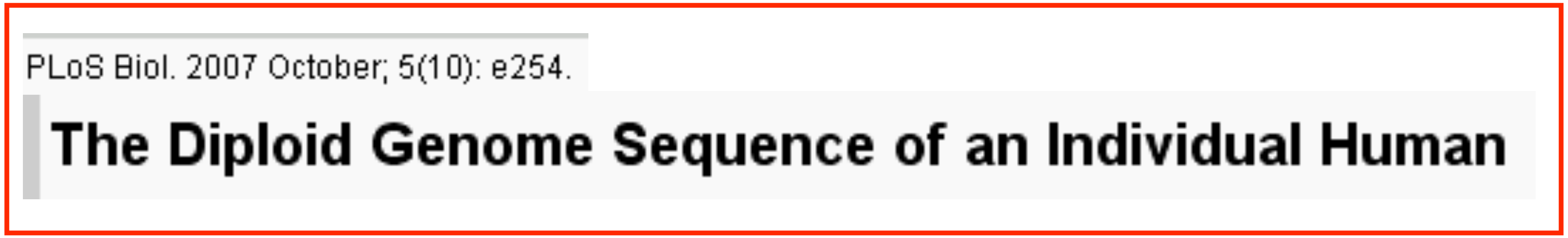
454



Nature **452**, 872-876 (17 April 2008) | doi:10.1038/nature06884; Received 3 December 2007; Accepted 4 March 2008

The complete genome of an individual by massively parallel DNA sequencing

Sanger



PLoS Biol. 2007 October; 5(10): e254.

The Diploid Genome Sequence of an Individual Human

454 : less than **1 million US $**, 7.4-fold redundancy in two months.

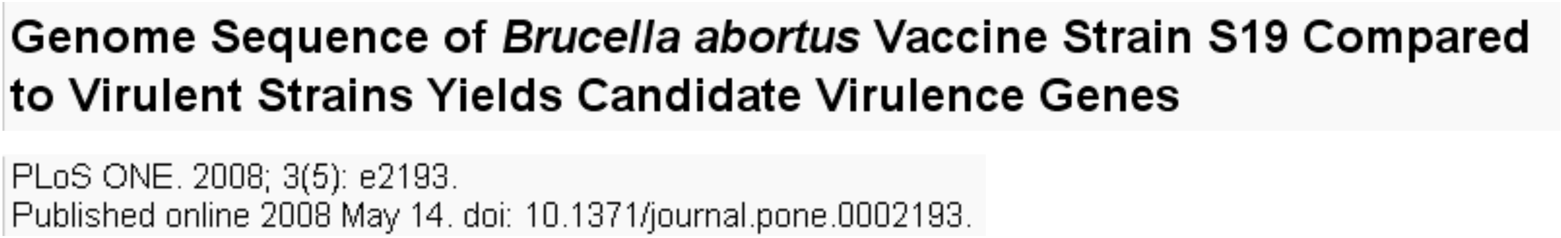
Sanger : approximately **100 million $**...

234 runs of 454 produced over 105 million bases per run.

➔ 3.3 million mutations, of which 10,654 cause changes in proteins.

Resequencing genomes

454



Genome Sequence of *Brucella abortus* Vaccine Strain S19 Compared to Virulent Strains Yields Candidate Virulence Genes

PLoS ONE. 2008; 3(5): e2193.
Published online 2008 May 14. doi: 10.1371/journal.pone.0002193.

A total of two, four-hour runs were performed to generate a total of ~800 thousand sequences with an average length of about 100 bases, resulting in more than 20X coverage of the whole genome of the strain.

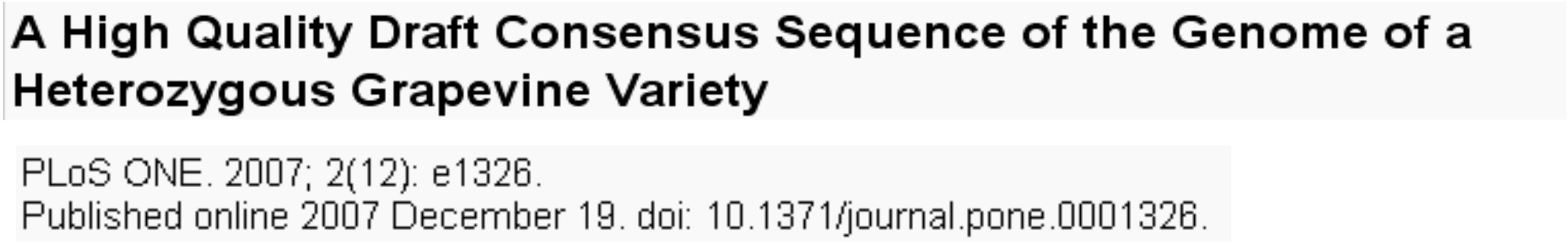
The functional analyses of the differences have revealed a total of 24 genes that may be associated with the loss of virulence

Tasks and problems

- Genomes
  - Resequencing genomes.
  - De novo sequencing a genome.
- Transcriptomes.
- Biodiversity.
  - SSU rRNA sequences
  - Metagenomes

Sequencing new genomes

454 & Sanger



A High Quality Draft Consensus Sequence of the Genome of a Heterozygous Grapevine Variety

PLoS ONE. 2007; 2(12): e1326.
Published online 2007 December 19. doi: 10.1371/journal.pone.0001326.

454 : In total, 12.5 million reads corresponding to 2.1 billions bases were produced.

Sanger: 6.2 million reads for a total of 3.5 billions bases were produced by Sanger sequencing from 43 libraries

The genome size of *V. vinifera* is 504.6 Mb

Problems

- Genomes
  - Resequencing genomes.
    - Assemble fragments with the help of the known reference genome. ➔ Easy & Known
  - De novo sequencing a genome.
    - Assemble fragments without the help of the known reference genome. ➔ More difficult & Known
  - Identification of genes, regulatory regions, mutations,...
    - Difficult but Known

A flood of data to come

Genomes : assembling the tags

- 2008
- Zerbino, D. R., and E. Birney. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18:821-829.
- Butler, J., I. MacCallum, M. Kleber, I. A. Shlyakhter, M. K. Belmonte, E. S. Lander, C. Nusbaum, and D. B. Jaffe. 2008. ALLPATHS: de novo assembly of whole-genome shotgun microreads. Genome Res. 18:810-820.
- Hernandez, D., P. Francois, L. Farinelli, M. Osteras, and J. Schrenzel. 2008. De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. Genome Res. 18:802-809.
- Chaisson, M. J., and P. A. Pevzner. 2008. Short read fragment assembly of bacterial genomes. Genome Res. 18:324-330.

- 2007
- Dohm, J. C., C. Lottaz, T. Borodina, and H. Himmelbauer. 2007. SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing. Genome Res. 17:1697-1706.

Conclusions :

- The work is “as before” excepted that sequences to assemble are shorter and in great abundance.
- According to publications, this seems to be a very active field.

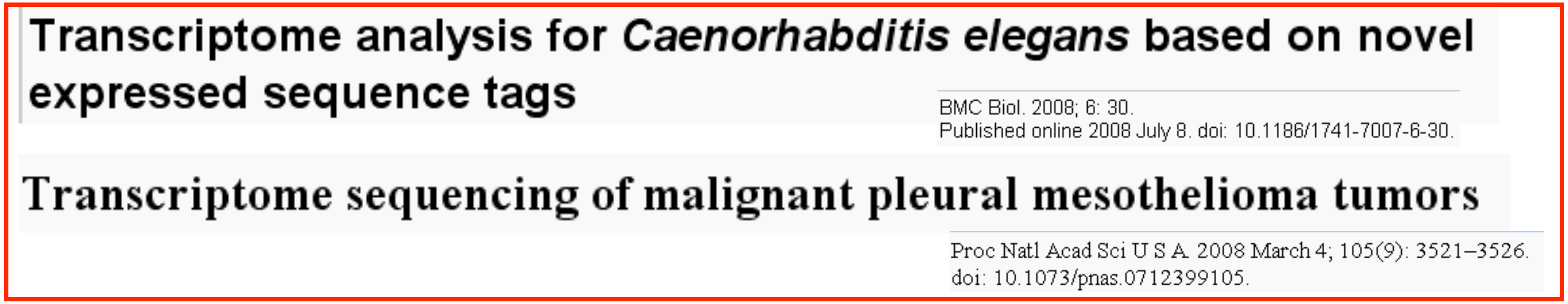
A flood of data to come

Tasks and problems

- Genomes
  - Resequencing genomes.
  - De novo sequencing a genome.
- Transcriptomes.
- Biodiversity.
  - SSU rRNA sequences
  - Metagenomes

Gene expression analyses

454



Transcriptome analysis for *Caenorhabditis elegans* based on novel expressed sequence tags

BMC Biol. 2008; 6: 30.
Published online 2008 July 8. doi: 10.1186/1741-7007-6-30.

Transcriptome sequencing of malignant pleural mesothelioma tumors

Proc Natl Acad Sci U S A. 2008 March 4; 105(9): 3521–3526.
doi: 10.1073/pnas.0712399105.

Over 30 million bases of cDNA from first larval stage worms. Approximately **14% of the newly sequenced expressed sequence tags do not map to annotated genes** ➔ these are novel genetic ***structures***.

Approximately 15 millions cDNA sequence reads with lengths of ≈105 bp each ➔ rapid and efficient analysis of gene expression in tumors.

Gene expression analyses

These new data sets are **very much similar to the previous technology** such as EST (Expressed Sequence Tags), excepted that :

- Sequences are a shorter (but not that much with 454 technology).
- There are much **much more** sequences (in the range 100-1000 fold)

Remarks :

Most labs use bioinformatic tools that are not well adapted, in particular Blast (or Blat) which was written in 1990 with much fewer sequences in mind.

Biologists are in need of tools to :

- Assemble tags into a cDNA (not always).
- Map the tags onto a reference genome.
- Make sense of the data (compare samples, cluster tags & samples, link to knowledge database).

Some tools simply need to be improved from previous ones developed for EST, SAGE and DNA chip technologies.

A flood of data to come

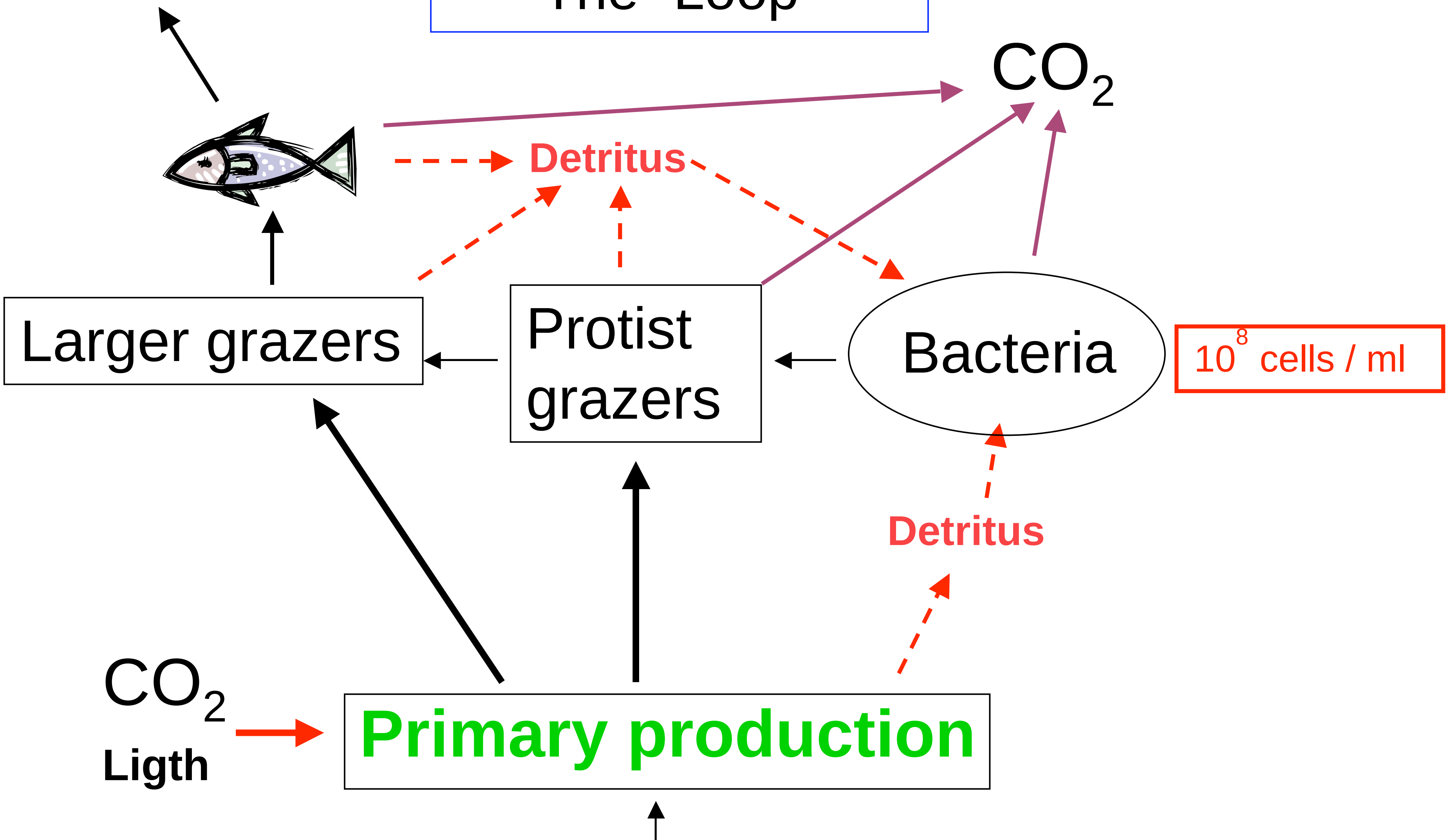
Tasks and problems

- Genomes
  - Resequencing genomes.
  - De novo sequencing a genome.
- Transcriptomes.
- Biodiversity.
  - SSU rRNA sequences
  - Metagenomes

Studying biodiversity, why ?

- Most of the earth’s biomass is not visible to the naked eye.
- These prokaryotes or protists are very difficult (impossible) to identify under a microscope.
- They produce more than 50% of the oxygen, and almost entirely recycle the inorganic matter on earth (Nitrogen, Phosphates, ...).
- They could play a significant role in the process of “Global Warming”.
- But : we have almost no idea of how many species there are and of which is doing what and when...

The “Loop”



CO_2

Detritus

Larger grazers

Protist grazers

Bacteria

10^8 cells / ml

Detritus

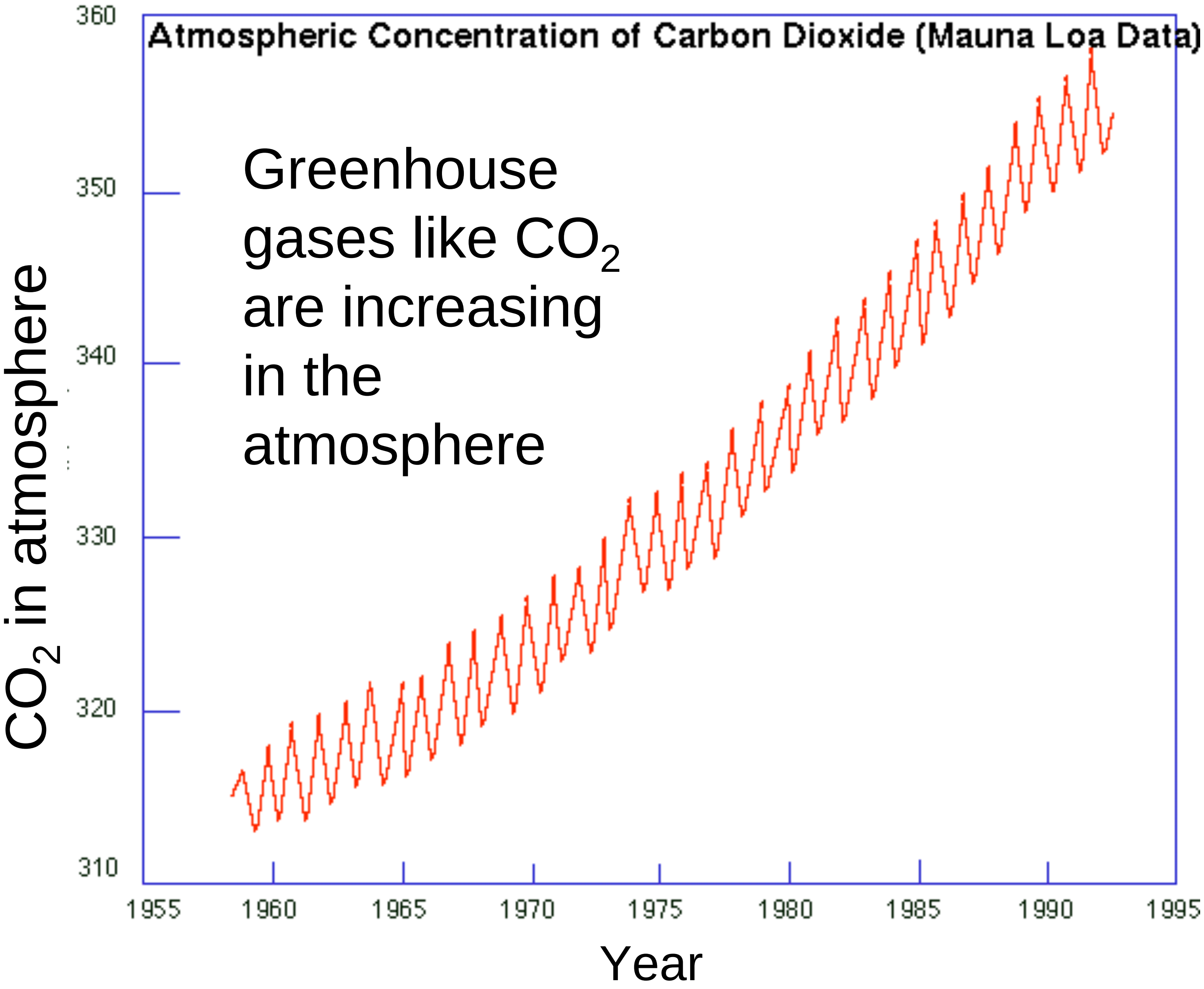
CO_2

Ligth

Primary production

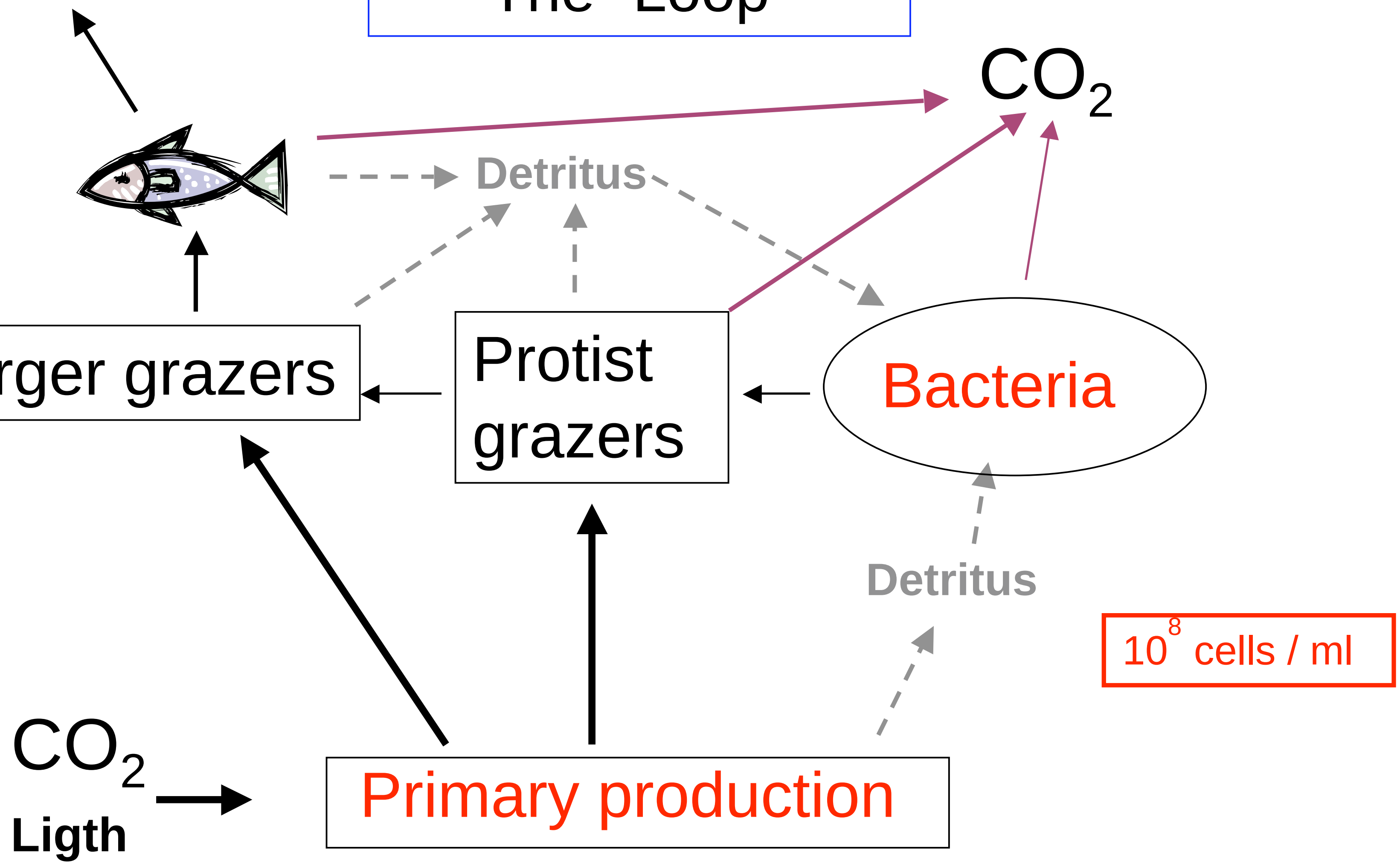
mostly in oceans, mostly microbes

The loop has been near equilibrium for a long time



Greenhouse gases like CO_2 are increasing in the atmosphere

The “Loop”



How will the loop react to increased CO2 ?

The identification of microbes

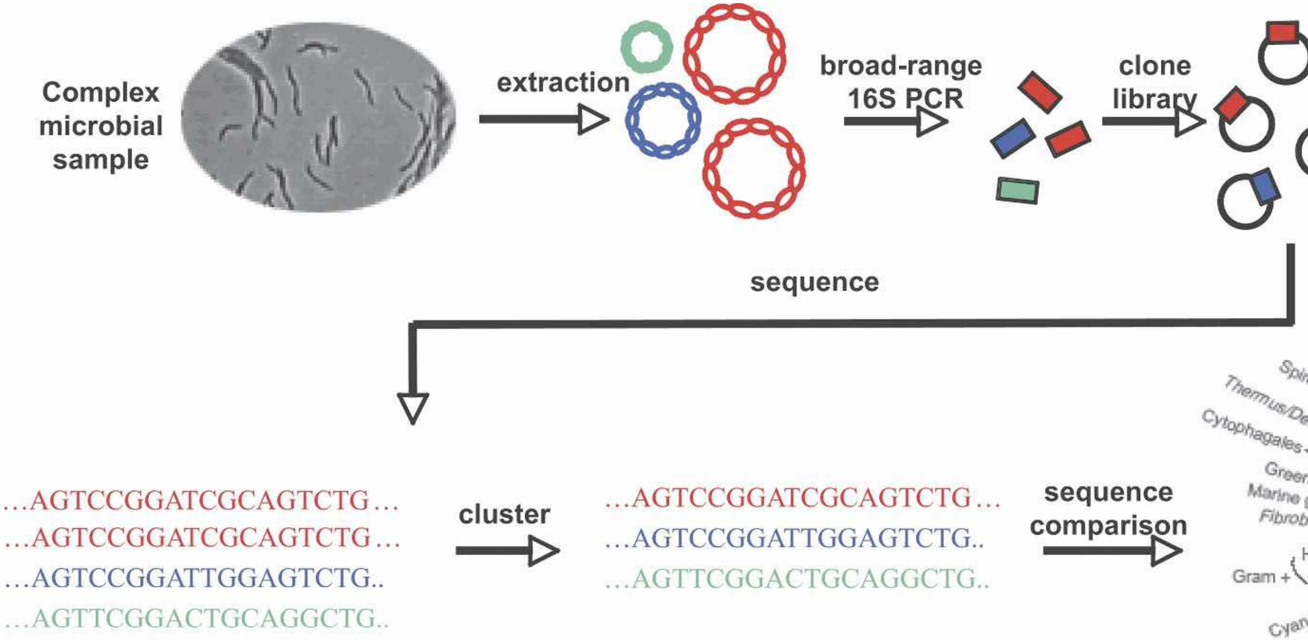
- Culture them ➔ not possible.

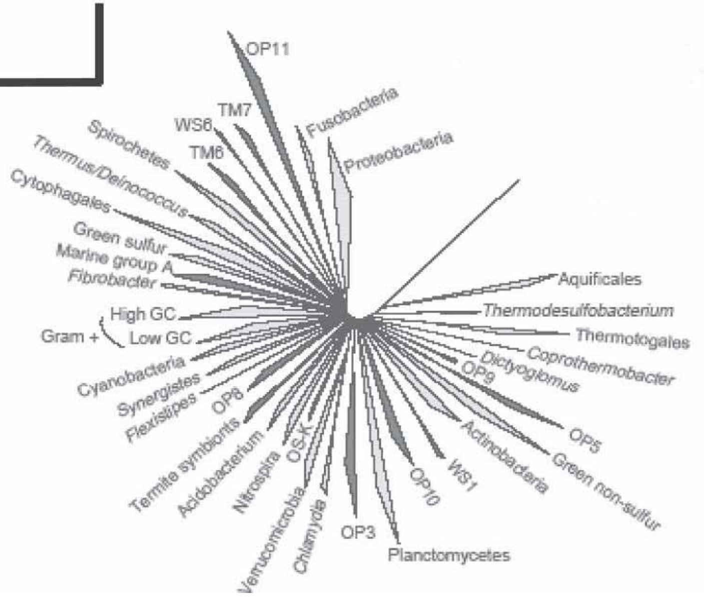
- Sequence their genomes ➔ not feasible.

- Use a gene present in the genome of every cell.
  - First done in 1977
  - Now the procedure of choice in every lab in the world.
    - Human gut, mouth, wounds,...
    - Sea water, earth fields, deep earth, ice, very hot waters (>100 °C), ...
      - ➔ they are many, everywhere
    - Industry & agriculture.

  - The gene used is coding for the ribosomal RNAs (that structures the machinery to make proteins).

Studying biodiversity, the “classic” approach

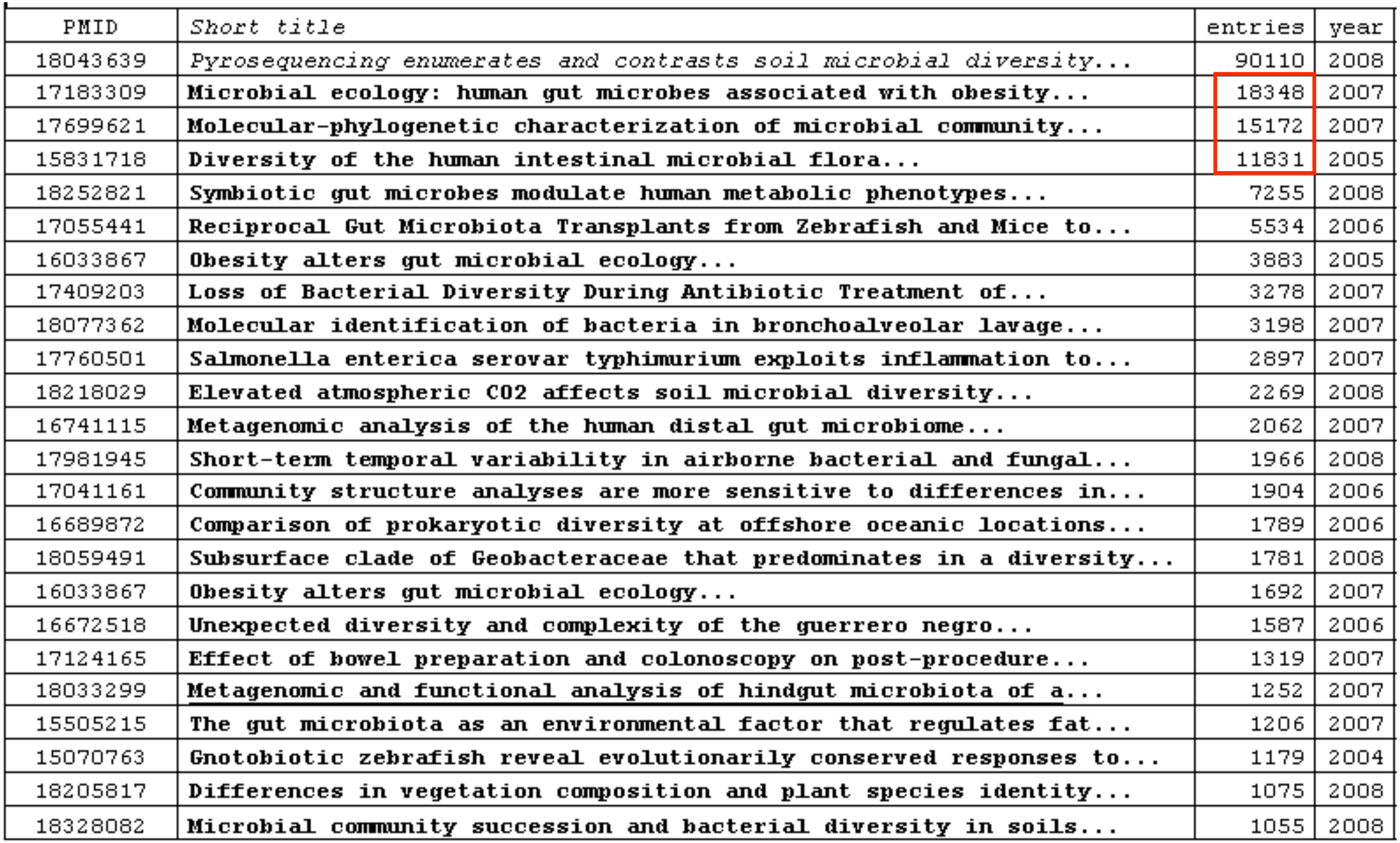




1. Purify the DNA
2. Extract all the ribosomal gene sequences.
3. **Clone the ribosomal RNAs of every cell.**
4. Random sequence ... as many clones as possible.
5. Analyse results, compare samples.
6. Publish you results ☺

Genome Res. 2006 16: 316-322

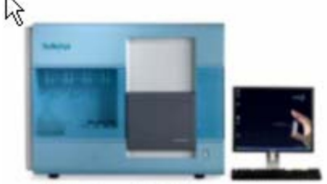
Biodiversity analyses - classic



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

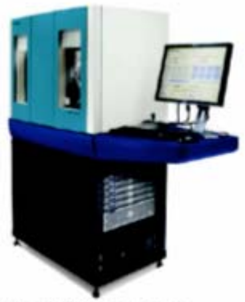
PCR – clone - sequence : too tedious for most labs !

Applied Biosystems
ABI 3730XL
1 Mb / day



Illumina / Solexa
Genetic Analyzer
2000 Mb / run

Roche / 454
Genome Sequencer FLX
100 Mb / run



Applied Biosystems
SOLiD
3000 Mb / run

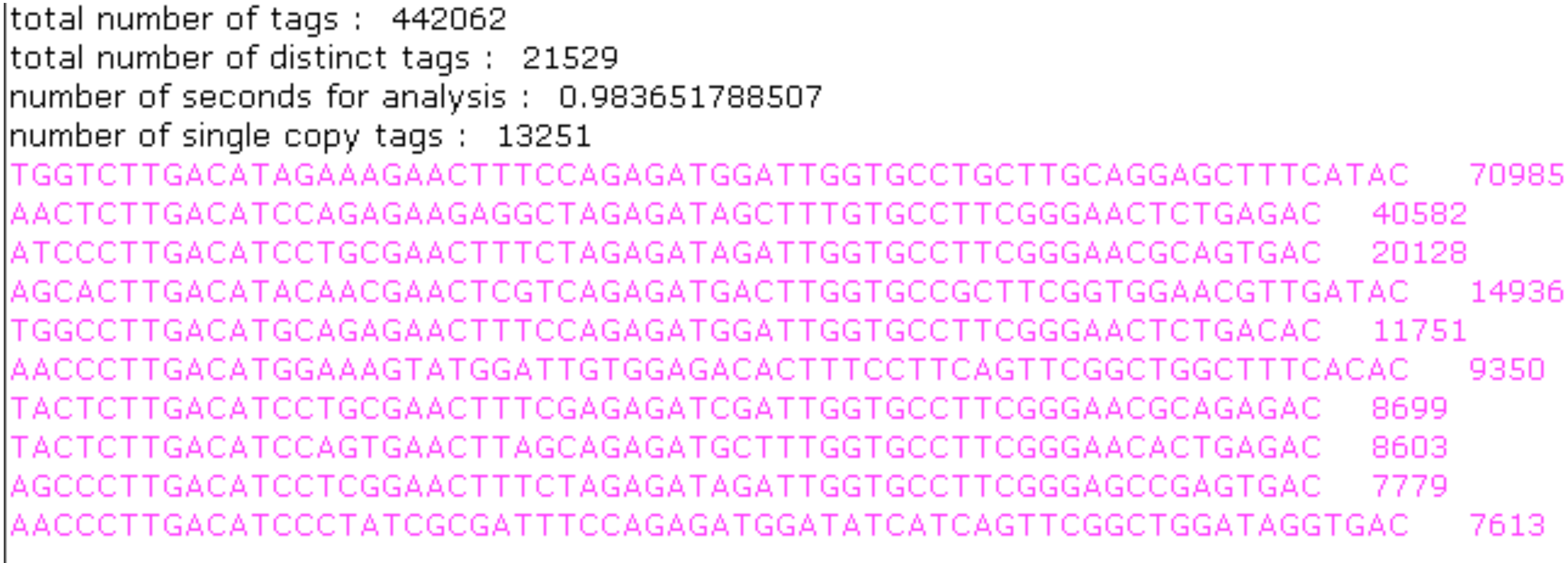
~~Clone~~ & sequence

Sequence every gene isolated :
> 400,000 sequences per day

Biodiversity, case studies

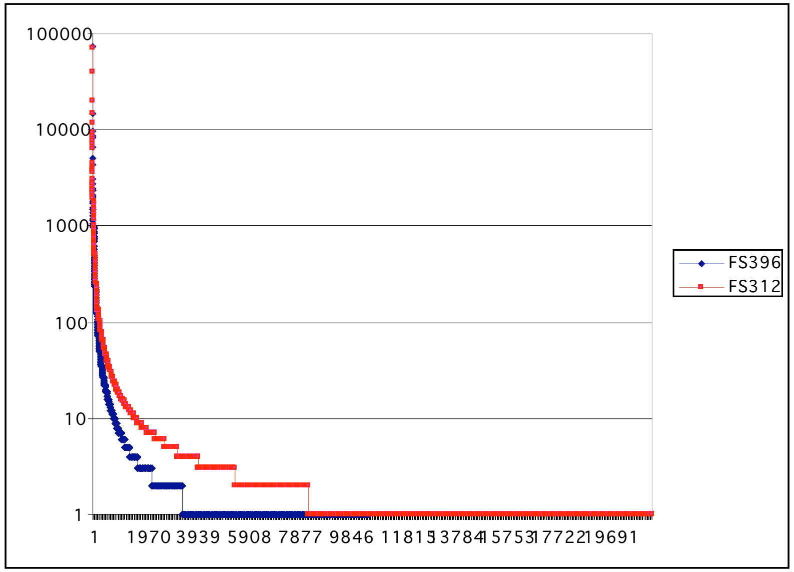
- Huber, J. A., D. B. Mark Welch, et al. (2007). "Microbial population structures in the deep marine biosphere." Science 318(5847): 97-100.
- Sogin, M. L., H. G. Morrison, et al. (2006). "Microbial diversity in the deep sea and the underexplored "rare biosphere"." Proc. Natl. Acad. Sci. U S A 103(32):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

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
AGCACTTGACATACAACGAACTCGTCAGAGATGACTTGGTGCCGCTTCGGTGGAACGTTGATAC   14936
TGGCCTTGACATGCAGAGAACTTTCCAGAGATGGATTGGTGCCTTCGGGAACTCTGACAC   11751
AACCCTTGACATGGAAAGTATGGATTGTGGAGACACTTTCCTTCAGTTCGGCTGGCTTTCACAC   9350
TACTCTTGACATCCTGCGAACTTTCGAGAGATCGATTGGTGCCTTCGGGAACGCAGAGAC   8699
TACTCTTGACATCCAGTGAACTTAGCAGAGATGCTTTGGTGCCTTCGGGAACACTGAGAC   8603
AGCCCTTGACATCCTCGGAACTTTCTAGAGATAGATTGGTGCCTTCGGGAGCCGAGTGAC   7779
AACCCTTGACATCCCTATCGCGATTTCCAGAGATGGATATCATCAGTTCGGCTGGATAGGTGAC   7613

complete analysis in seconds :   1.04010820515
```



Problems :

- Strict dereplication ?
- Loose dereplication ?

Clustering tags into OTU

Operational Taxonomic Unit : cluster together tags that are similar.

- How to define similarity ? i.e. how to calculate distances ?
- How to cluster ?

- Usual manner for few long sequences :
  - Do a multiple alignement.
  - Compute phylogenetic distances.
  - Phylogeny or various clustering methods.

- But :
  - Too many sequences to align.
  - Domains are too divergent for present multiple alignements methods.

  - ➔Cluster according to words frequencies (ex. words of 5 nt) ?
    - No alignement, much faster, much better ?
  - ➔ ???

We need cleaned experimental data sets to evaluates methods & algorithms

Assign each tag to a taxon

Clustering may be fine for comparing samples, but it provides no hint about :

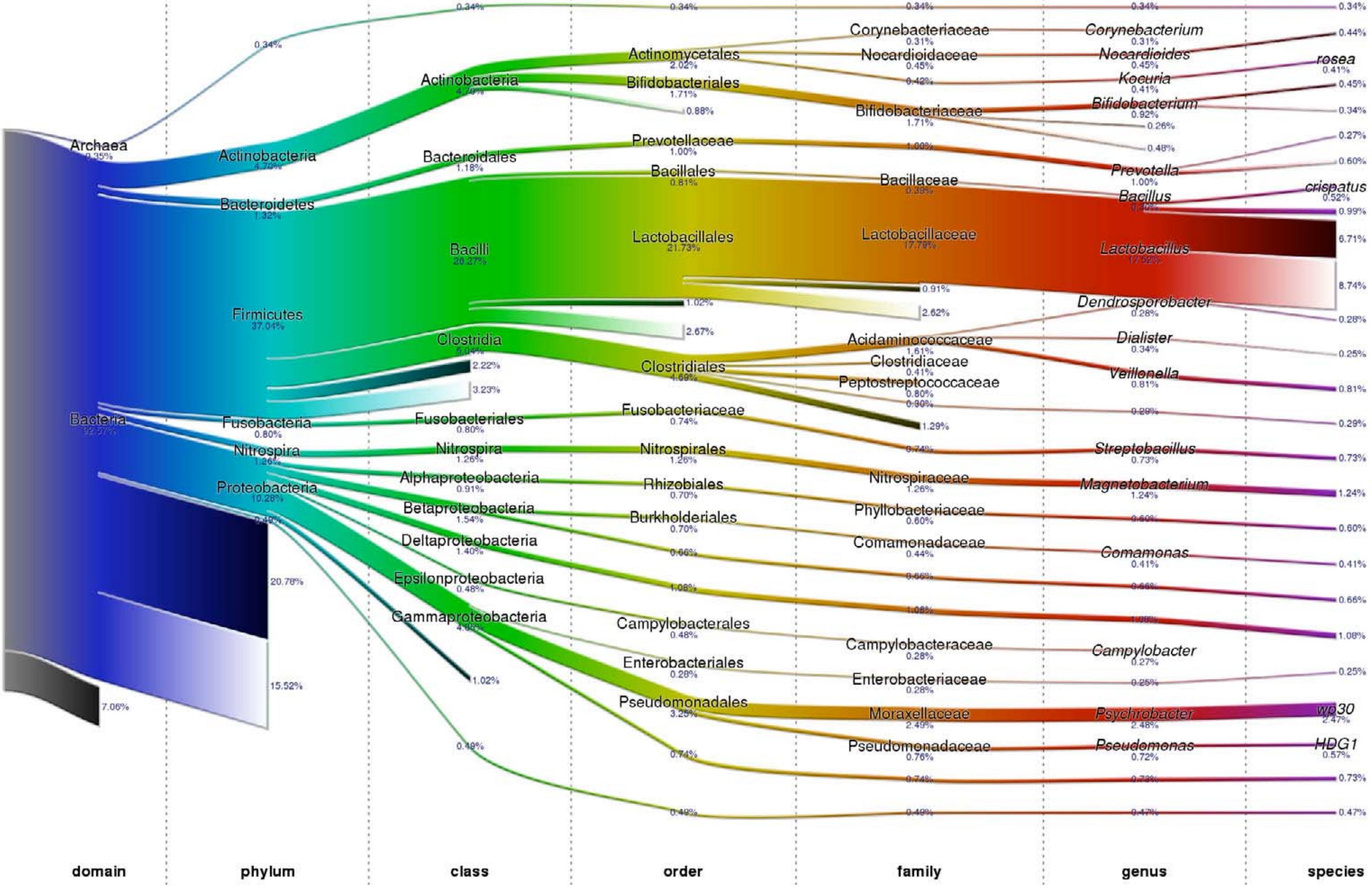
- Which are the species present ?
- What do they do ?
- What is the significance of a change in composition over time or space ?

We need to assign each tag or each OTU to a name, the best would be to assign as much as possible :

1. To a known species (which is in culture somewhere).
2. To an unknown but sequenced species (genome sequenced, but no culture).
3. To a sequence found elsewhere.

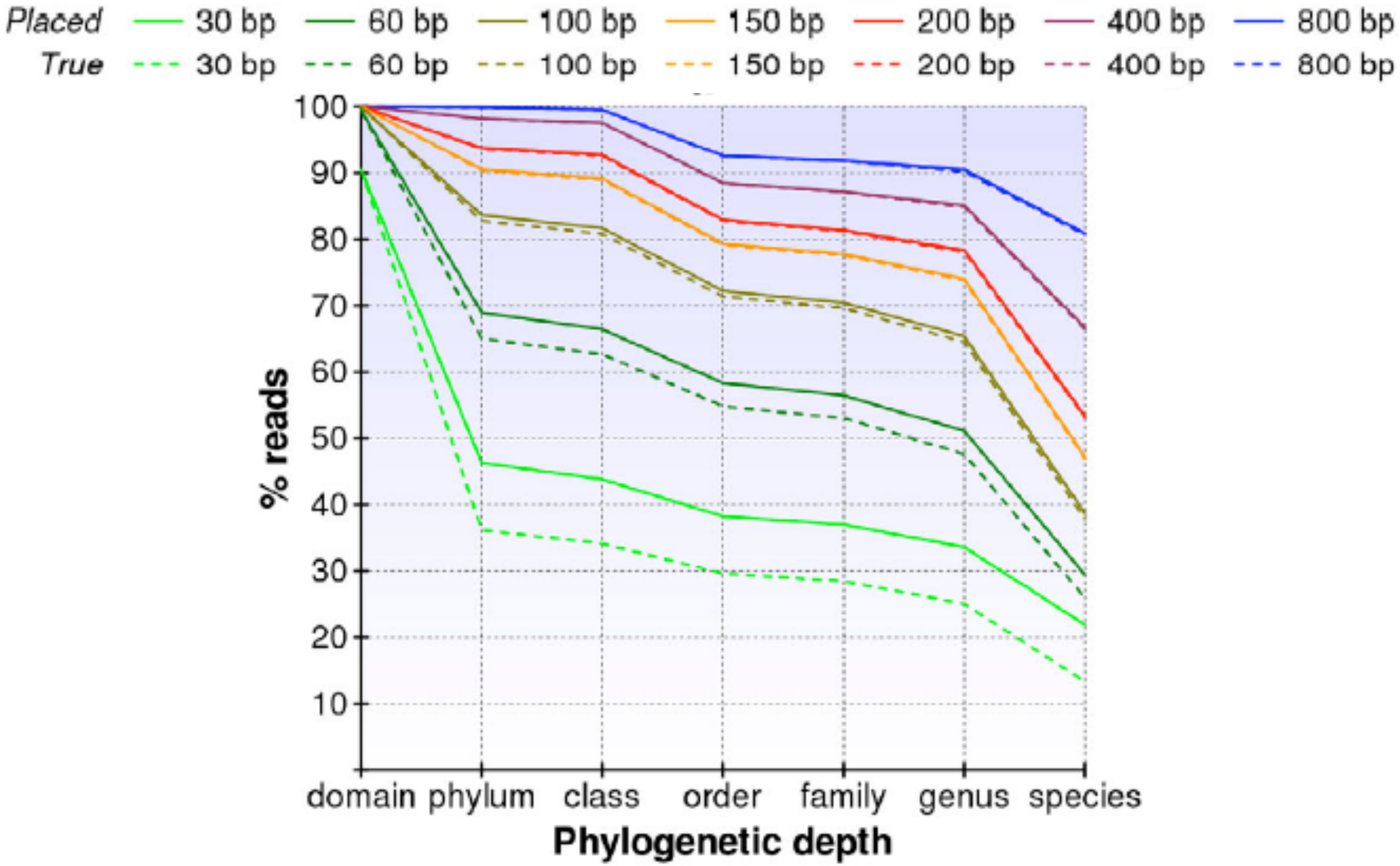
Assignments are done by similarity to the public sequences database (Blast).

Assign each tag to a taxon



BMC Microbiology 2007, **7**:108

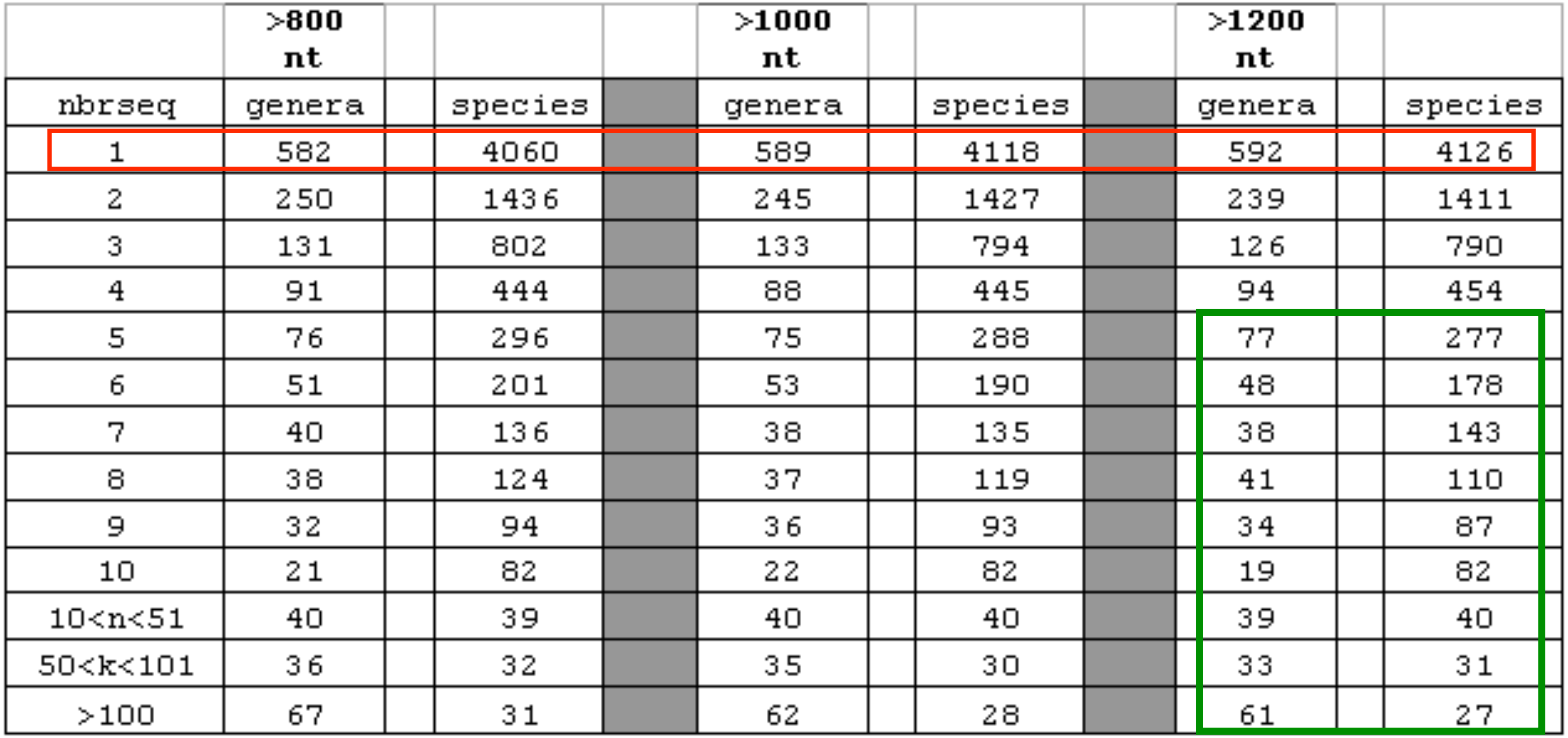
Assign each tag to a taxon



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

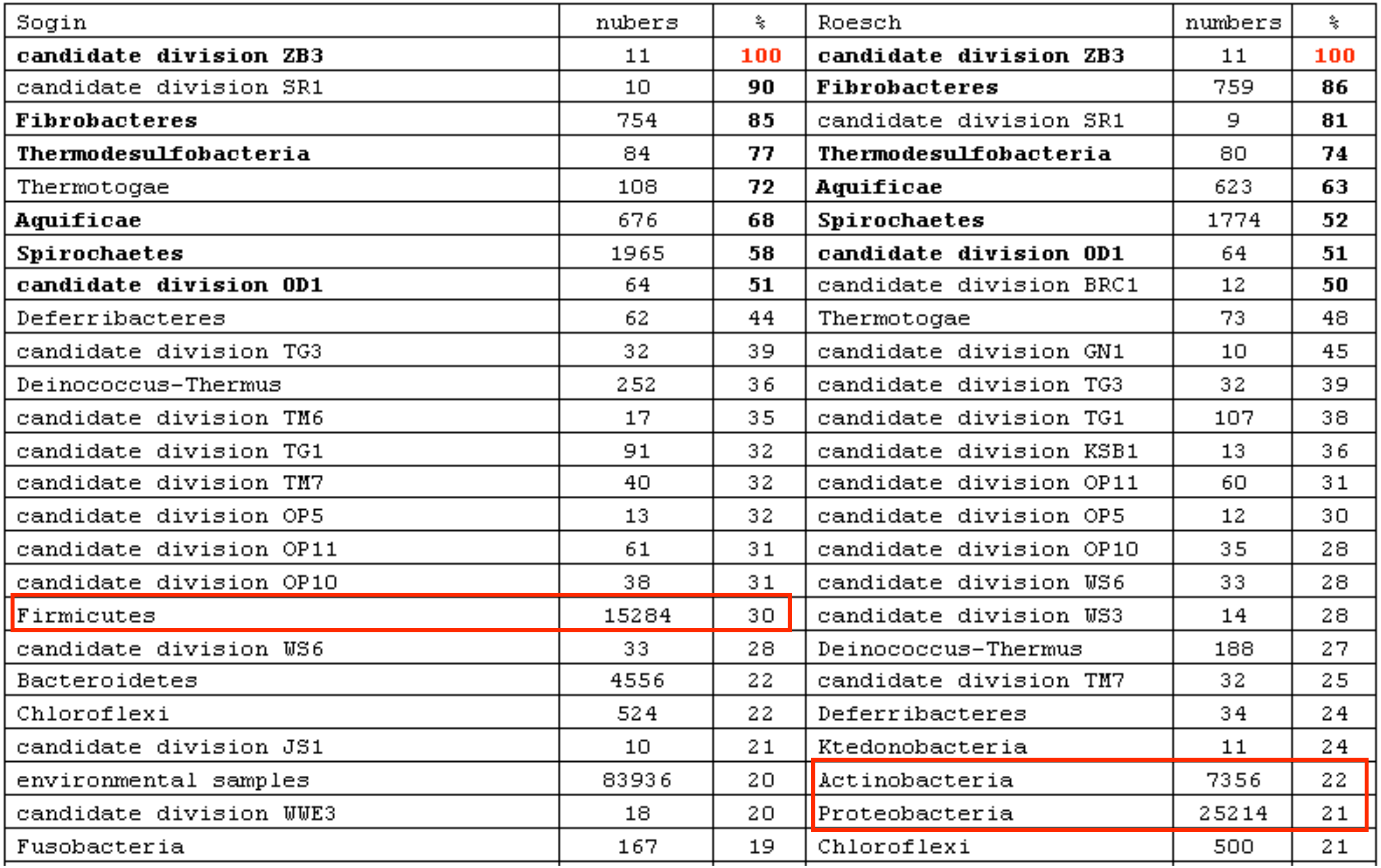
Only 8,000 species in cultures !

Most species are known from a single sequence !

➔ Tags taxonomic specificities are over-evaluated.

➔ Most species have not been sequenced at all.

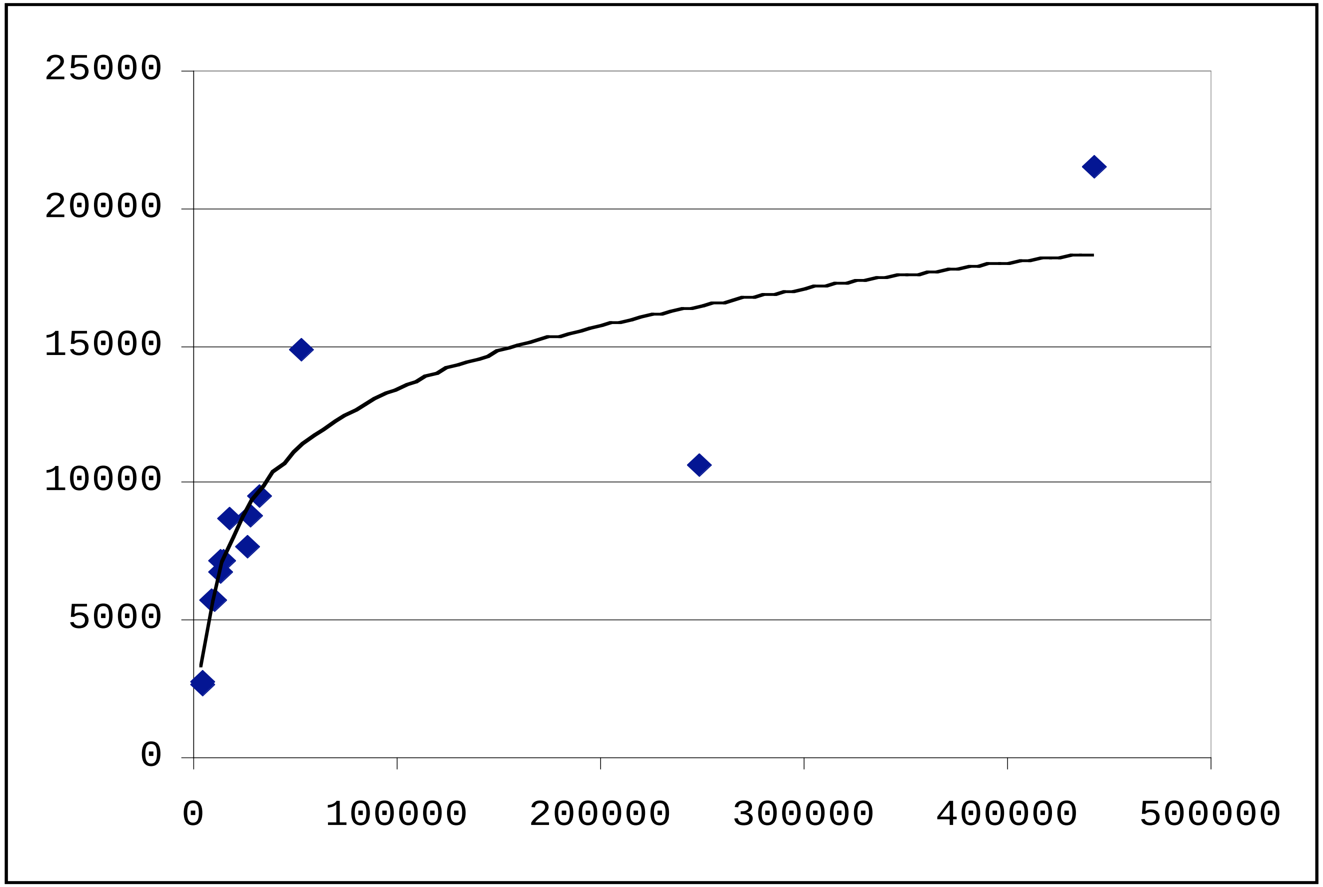
Main taxa that were not amplified



Sogin	nubers	%	Roesch	numbers	%
candidate division ZB3	11	**100**	**candidate division ZB3**	11	**100**
candidate division SR1	10	**90**	**Fibrobacteres**	759	**86**
Fibrobacteres	754	**85**	candidate division SR1	9	**81**
Thermodesulfobacteria	84	**77**	**Thermodesulfobacteria**	80	**74**
Thermotogae	108	**72**	**Aquificae**	623	**63**
Aquificae	676	**68**	**Spirochaetes**	1774	**52**
Spirochaetes	1965	**58**	**candidate division OD1**	64	**51**
candidate division OD1	64	**51**	candidate division BRC1	12	**50**
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
Firmicutes	15284	30	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	Ktedonobacteria	11	24
environmental samples	83936	20	Actinobacteria	7356	22
candidate division WWE3	18	20	Proteobacteria	25214	21
Fusobacteria	167	19	Chloroflexi	500	21

Primers need to be better designed !

New tags as a function of sequencing effort
Saturation curve



Even when sequencing 400,000 tags, we were not able to sequence every present species ... We are still missing the rare ones.

The singletons !

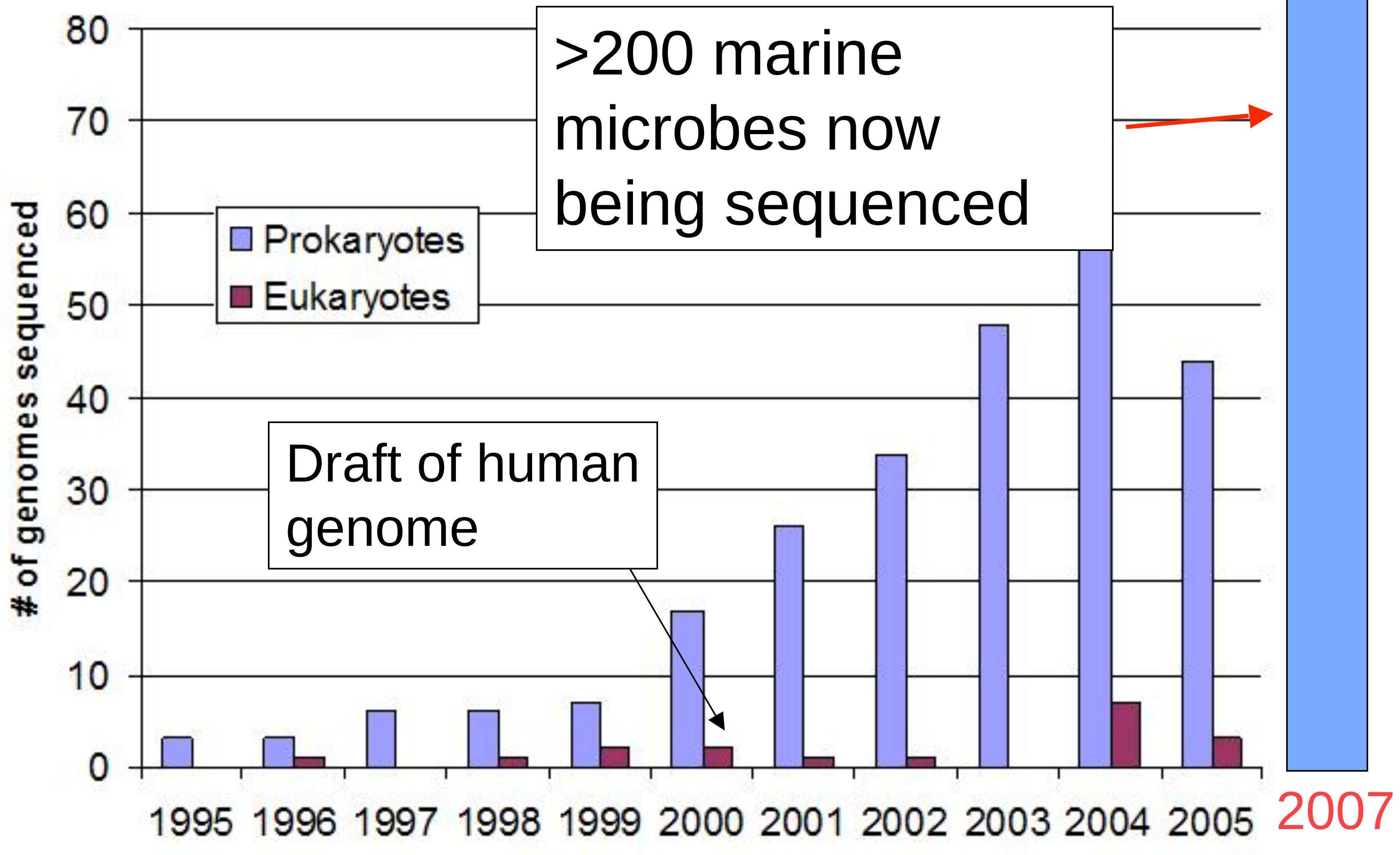
- A singleton is a sequence which was found only once !
- ➔ How many singletons in these experiments ?

Experiment	Il	Br	Ca	Fl						
Total tags	**31745**	**26115**	**53245**	**28247**						
unique tags	9486	7683	14885	8779						
singletons tags	7337	5598	11638	6792						
% Singletons	**23**	**21**	**22**	**24**						
Experiment	53R	55R	112R	115R		138	FS396	FS312	FS396	FS312
Total tags	**4999**	**13901**	**9281**	**11004**		**14373**	**17665**	**4834**	**247825**	**442061**
unique tags	2655	7186	5751	5776		7167	8699	2769	10613	21529
singletons tags	2297	6217	5040	5009		6237	7587	2396	7185	13251
% Singletons	**46**	**45**	**54**	**46**		**43**	**43**	**50**	**3**	**3**

Tasks and problems

- Genomes
  - Resequencing genomes.
  - De novo sequencing a genome.
- Transcriptomes.
- Biodiversity.
  - SSU rRNA sequences
  - Metagenomes

Many genomes are now sequenced.



>200 marine microbes now being sequenced

Draft of human genome

What is a metagenome ?

- Metagenome experiments consist in :

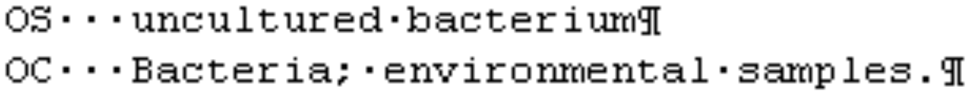
1. **Extract the DNA from a given sample.**
2. **Sequence it all.**
3. Try to assemble these pieces to reconstitute the different genomes that were present in the sample.
4. Try to make sense of this assembly

1. No problem.
2. Now almost feasible.
3. Works only for samples with few different genomes (presently less than 10).
4. Presently impossible.

NOTE : the first metagenome (Sargasso sea sample) provided more protein sequences than was already known.
➔ This required to build a new division for storage in the public database ...

Technical problems

- Lack of complete sequences to evaluate primers.
- 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) !

Conclusions in Biology

- 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 sequencing can now be done in a **single operation** using bench-top instruments.

- **Global sequencing** will soon **replace any other method for estimating biodiversity and in transcriptome studies**.

- A wide and generalized sequencing effort 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**.

Conclusions in Bioinformatics

- A wide and generalized sequencing effort of **ontology** building of **well-identified strains** deposited in collections worldwide is required to form the basis of derived annotations of environmental sequences.

- New formats need to be developed to store the flood of data soon to come, how to store efficiently :
  - The raw data.
  - Data with final annotations.
  - **Intermediate calculations and results.**

- New tools are required to efficiently query these hudge datasets.
  - Entrez is nearly not usable.
  - SRS is problematic.
  - ACNUC works quite well but is not widely supported.

Conclusions in Informatics

- Efficient algorithms (computer clusters ?) to assemble genomes.
    - Already a blooming field !
- Efficient algorithms to analyse transcriptomic data.
    - Already a blooming field !
    - Most developments are derivatives from earlier methods.
- A query system linking knowledge datases (ontologies) and sequence annotations needs to be developed.

- New methods to classify short & divergent sequences are needed.
- New methods to search sequences by similarity ?
- Is there a better solution than simply flat files or SQL databases to store these hudge data sets?