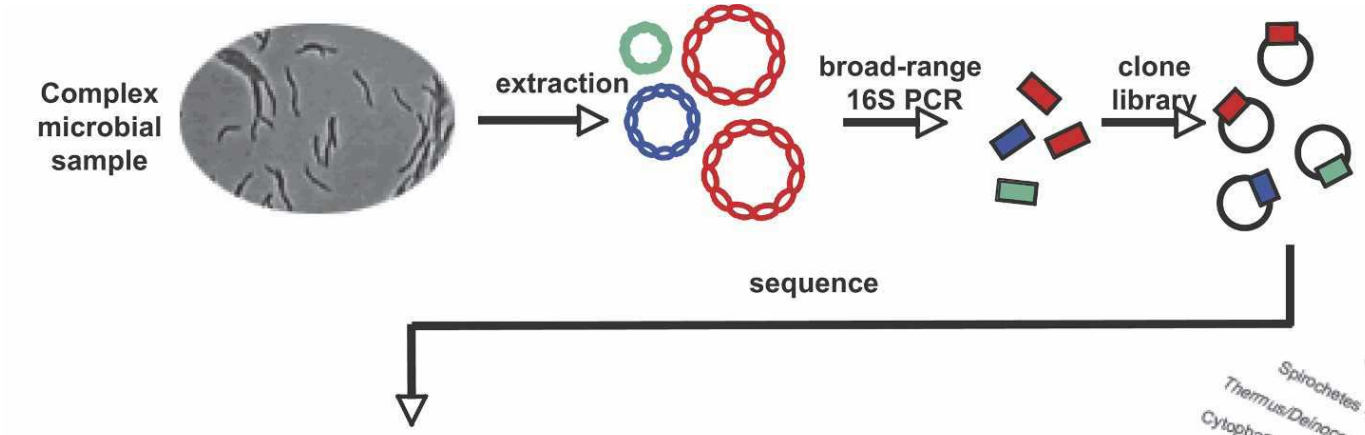
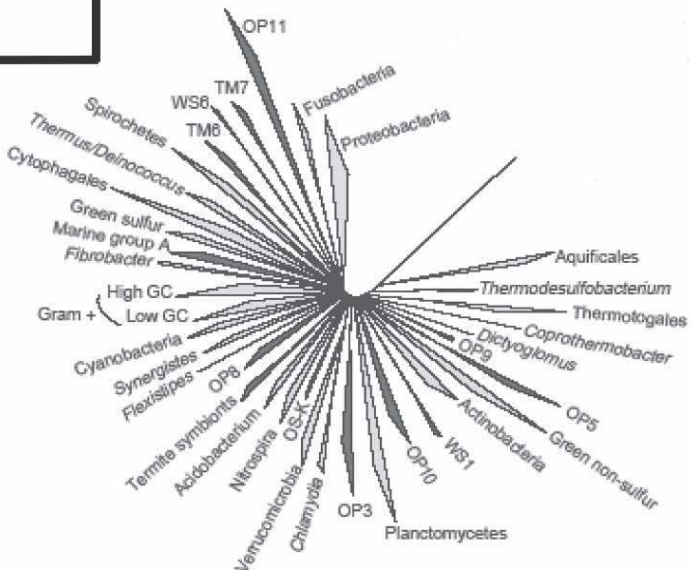
rRNA 454 datasets
and microbial biodiversity analyses

Richard Christen
Virtual Biology Laboratory
University of Nice & CNRS UMR 6543
Parc Valrose. F06108. France
christen@unice.fr

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

Microbes Environ. Vol. 23, No. 4, 000–000, 2008
http://wwwsoc.nii.ac.jp/jsme2/ doi:10.1264/jsme2.ME08525

Minireview

Global Sequencing: A Review of Current Molecular Data and New Methods Available to Assess Microbial Diversity

RICHARD CHRISTEN[1]*

PMID	Short title	Entries
18043639	Pyrosequencing enumerates and contrasts soil microbial diversity...	90110
17183309	Microbial ecology: human gut microbes associated with obesity...	18348
17699621	Molecular-phylogenetic characterization of microbial community...	15172
15831718	Diversity of the human intestinal microbial flora...	11831
18252821	Symbiotic gut microbes modulate human metabolic phenotypes...	7255
17055441	Reciprocal Gut Microbiota Transplants from Zebrafish and Mice to...	5534
16033867	Obesity alters gut microbial ecology...	3883
17409203	Loss of Bacterial Diversity During Antibiotic Treatment of...	3278
18077362	Molecular identification of bacteria in bronchoalveolar lavage...	3198
17760501	Salmonella enterica serovar typhimurium exploits inflammation to...	2897
18218029	Elevated atmospheric CO_2 affects soil microbial diversity...	2269
16741115	Metagenomic analysis of the human distal gut microbiome...	2062
17981945	Short-term temporal variability in airborne bacterial and fungal...	1966
17041161	Community structure analyses are more sensitive to differences in...	1904
16689872	Comparison of prokaryotic diversity at offshore oceanic locations...	1789
18059491	Subsurface clade of Geobacteraceae that predominates in a diversity...	1781
16033867	Obesity alters gut microbial ecology...	1692
16672518	Unexpected diversity and complexity of the guerrero negro...	1587
17124165	Effect of bowel preparation and colonoscopy on post-procedure...	1319
18033299	Metagenomic and functional analysis of hindgut microbiota of a...	1252
15505215	The gut microbiota as an environmental factor that regulates fat...	1206
15070763	Gnotobiotic zebrafish reveal evolutionarily conserved responses to...	1179
18205817	Differences in vegetation composition and plant species identity...	1075
18328082	Microbial community succession and bacterial diversity in soils...	1055

Applied Biosystems
ABI 3730XL
1 Mb / day

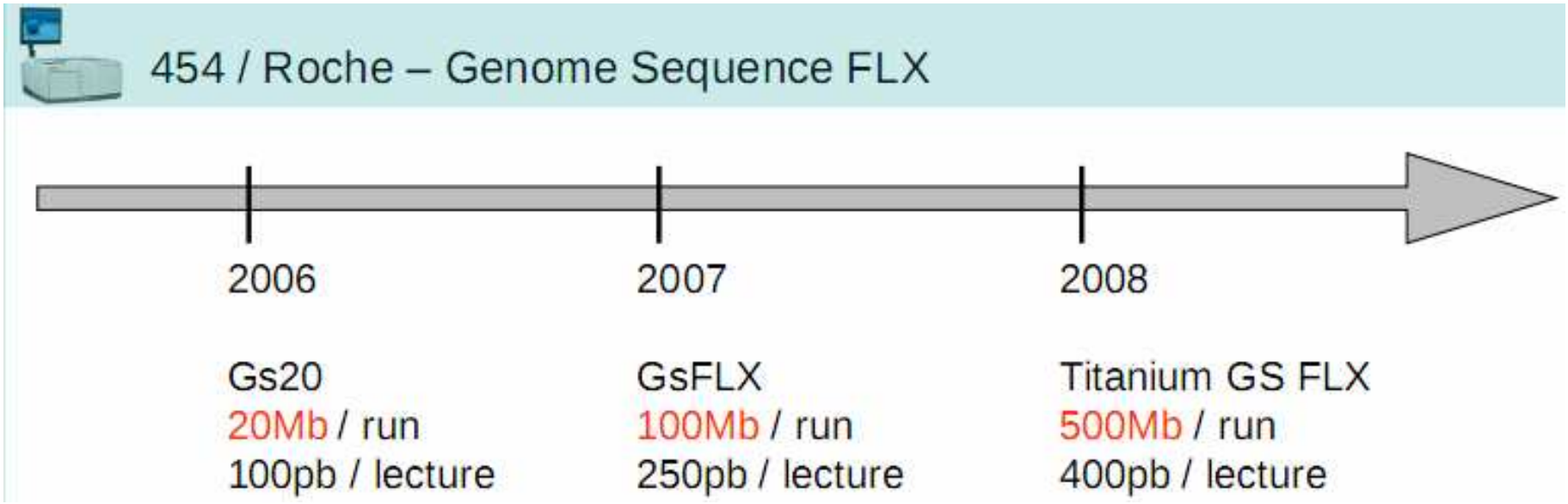
Roche / 454
Genome Sequencer FLX
100 Mb / run

Illumina / Solexa
Genetic Analyzer
2000 Mb / run

Applied Biosystems
SOLiD
3000 Mb / run

In 2007, three next-generation sequencing platforms were present: 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)



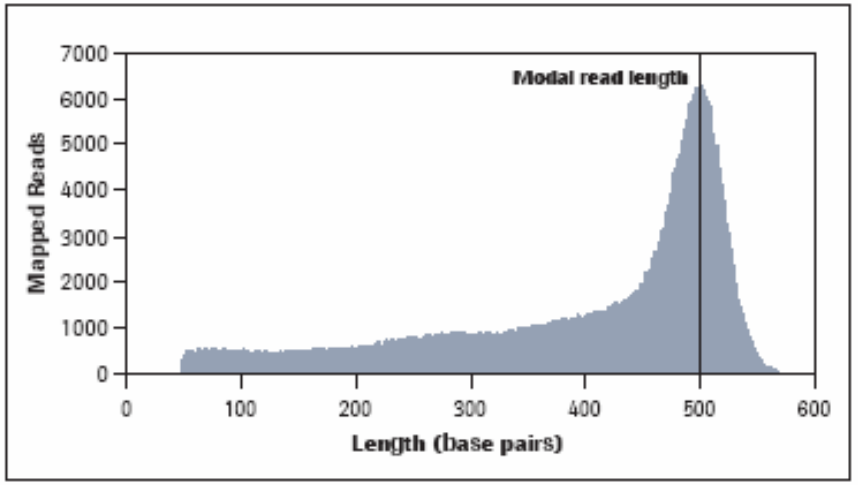


Figure 1: Example Read Length Distribution of 629,643 reads from *E. coli* K-12 (Genome size ~4.5 Mb) with a modal read length of 504 bases.

Bioinformatics of 454 datasets

Major issues (in a fast and efficient manner):

– Design “good” primers ←→ Choose domain to amplify.

– Cluster tags.

– Assign tags to a given taxonomic level.

– “Statistical analyses”

- Run biodiversity analyses on a single sample.
- Compare samples.

– Relate diversity to ecology.

The domain amplified should :

- Have “good” taxonomic properties.
  - Compare **tags extracted** to **full sequences; how many tags assign to different clades ?**

- Be present in a large number of sequences in the public databases.
  - ➔ In order to do many taxonomic assignments.
  - ➔ Compare to clone libs.

- Be “454 compatible” in length: distal primers should be reached.

Find good primers

Goal: find existing primers by searching them into published articles.

Problem: very long process !

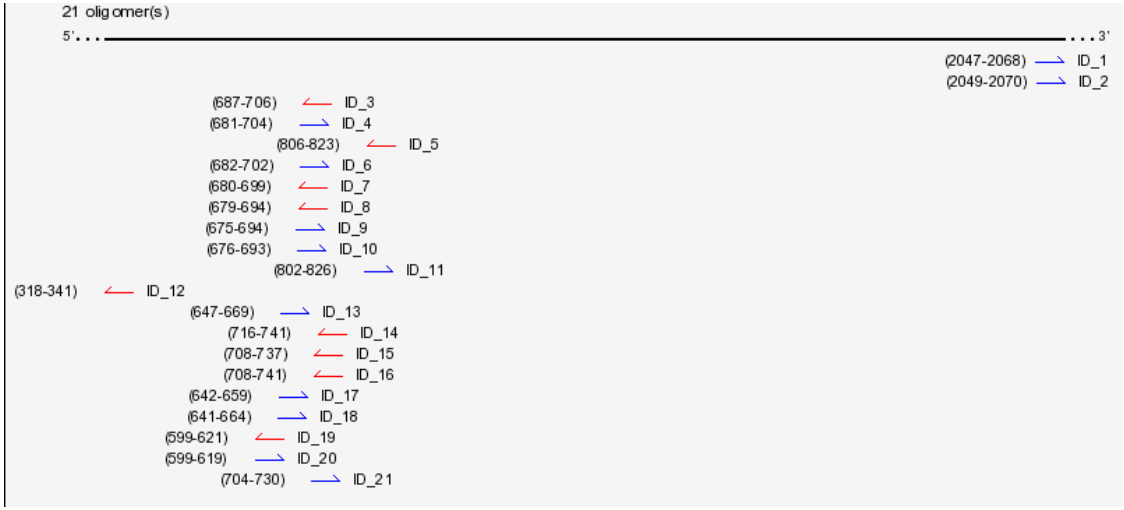
- Search for a set of relevant articles (pubmed, personal bibliography, etc.)
- Download pdf files
- Read and extract the proper primers
- Check if the primers match on the sequences we want to amplify
- Compute theoretical values (Tm values, PCR product, ...)
- Biological experiments and final validation

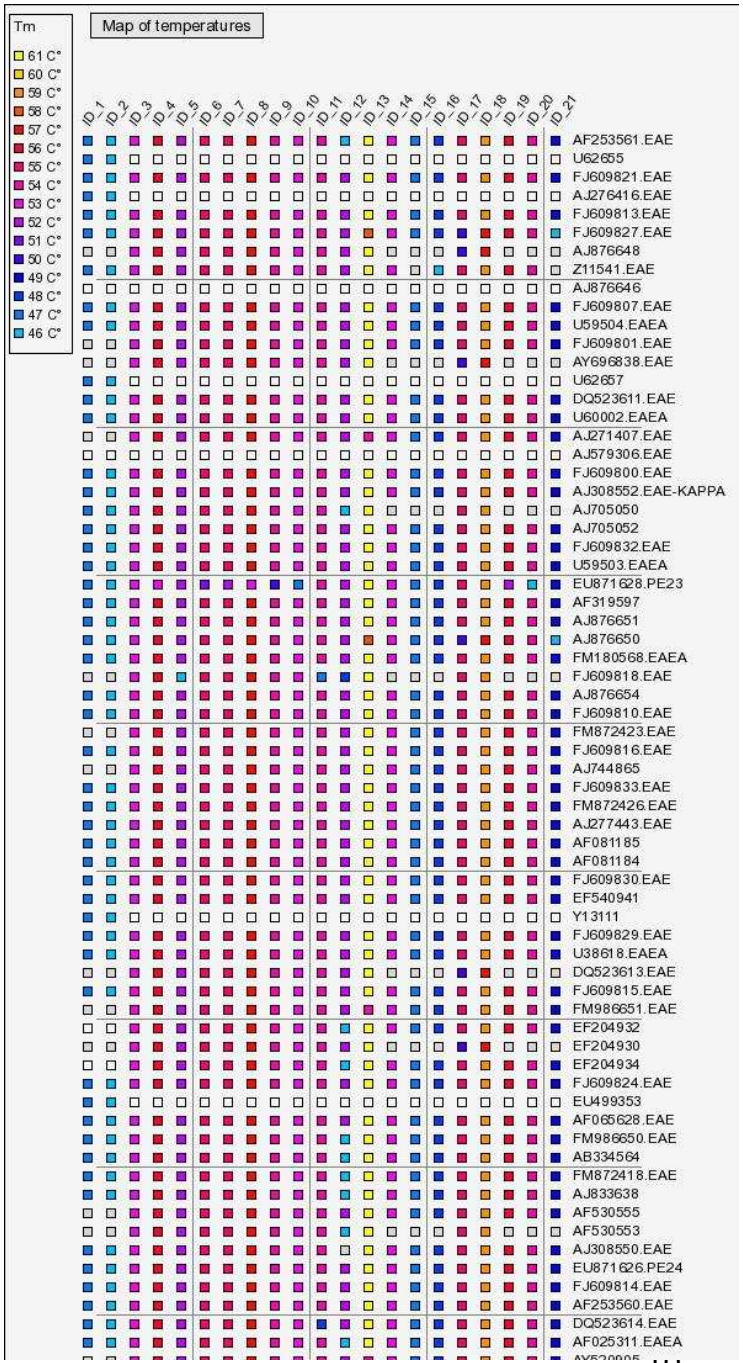
Automatic process -> new sofware

- Search for relevant articles → - Get a list of articles from Pubmed, Jane and EtBlast (keywords)
  - Add personal librairies
- Download pdf files → Download every available article as pdf files (and/or html files) from the Internet (websites of the publishers, miscellaneous repositories, etc.)
- Extract the primers → Automated extraction of each oligomer from each article (search engine, filters).
- Check the presence of primers → Comparison with a set of sequences to amplify:
  → keep only oligomers present in sequences
- Check for theoretical values → **OHM** (http://bioinfo.unice.fr/ohm)

- Biological experiments

Examples





Use of OHM to provide an overview of Tms

http://bioinfo.unice.fr/ohm

Check for “good” primers

&

Choose domain to amplify.

PrimerExplorer

Design of Primers

- PrimerExplorer
  - Universal
    - GESOL & FONCTIOMIC-RMQS (INRA)
    - BioMarks
  - Group specific
    - CEA Cadarache
    - BioMarks

- Able to analyse 100 couples / 800 000 sequences per 24 hours
- Takes the IUPAC code
- Allows k more differences between a primer and a sequence

PrimerExplorer

Inputs :

- a file of primers,
- a file of fasta sequences
- a value of k for F and R primers

Outputs:

- Every couple of primers found at k differences.
- Every tag that is amplified in these conditions.
- The taxonomic descriptions of amplicons.

Variable domains in the 16S rRNA gene sequences

Table 1: 16S variable region range definitions.

Variable region	*E. coli* 16S rDNA range			5' primer	3' primer
	start	end	length		
V1	8	120	113	5'-AGAGTTTGATCMTGGCTCAG	5'-TTACTCACCCGTICGCCRCT
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'-CCGTCAATTCMTTTGAGTTT
V6	907	1073	167	5'-AAACTCAAAKGAATTGACGG	5'-ACGAGCTGACGACARCCATG
V7 & VV8	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.

BMC Microbiol. 2007; 7: 108.

Bacterial flora-typing with targeted, chip-based Pyrosequencing

Sundquist, Bigdeli,Jalili, Druzin, Waller, Pullen, El-Sayed, Taslimi, Batzoglou and Ronaghi.

Variable domains in the 16S rRNA gene sequences

Table 1: 16S variable region range definitions.

Variable region	*E. coli* 16S rDNA range			5' primer	3' primer
	start	end	length		
V1	8	120	113	5'-AGAGTTTGATCMTGGCTCAG	5'-TTACTCACCCGTTCGCCRCT
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'-CCGTCAATTCMTTTGAGTTT
V6	907	1073	167	5'-AAACTCAAAKGAATTGACGG	5'-ACGAGCTGACGACARCCATG
V7 & VV8	1054	1406	353	5'-CATGGYTGTCGTCAGCTCGT	5'-ACGGGCGGTGTGTAC
V9	1392	1507	116	5'-GTACACACCGCCCGT	5'-TACCTTGTTACGACTT

domain¤	left¤	rigth¤	length¤	extracted·k¤	%¤	extracted¤	%¤
V1¤	3¤	95¤	71¤	129,671¤	29.4¤	83360¤	18.9¤
V2¤	74¤	317¤	223¤	356,400¤	80.9¤	229978¤	52.2¤
V3¤	305¤	473¤	148¤	388,054¤	88.1¤	332483¤	75.5¤
V4¤	458¤	728¤	252¤	1,248¤	0.3¤	323¤	0.1¤
V5¤	645¤	763¤	99¤	1,024¤	0.2¤	188¤	0.0¤
V6¤	811¤	958¤	127¤	358,323¤	81.4¤	315315¤	71.6¤
V7·&·VV8¤	978¤	1316¤	317¤	251,597¤	57.1¤	184965¤	42.0¤
V9¤	1312¤	1411¤	84¤	95,982¤	21.8¤	85108¤	19.3¤

Calculation times for analysis of **440,390** bacterial 16S rRNA sequences longer than 800 nt (at 0 difference 749 seconds, at 1 difference 757 seconds , at 2 differences 695 seconds, at 3 differences 739 seconds = **10 minutes, almost 1 minute per couple of primers**).

Conserved domains in the 16S rRNA gene sequences

Primers for domain V2
nbr extracted tags at 2 differences : 356,400 (229,978 exact)
min length=42, max length = 1060, mean length=223

F primers

AGYGGCGIACGGGTGAGTAA	244493	31.8
A**X**YGGCGIACGGGTGAGTAA	26738	3.5
AGYGGC**X**IACGGGTGAGTAA	19778	2.6
AGYGGCGIACGGGTG**X**GTAA	11116	1.4
AGY**X**GCGIACGGGTGAGTAA	9337	1.2
A**X**YGGCGIACGGGTG**X**GTAA	7890	1.0
AGYGGCGIACGGGTGAG**X**AA	6376	0.8
AGYGGCGIAC**X**GGTGAGTAA	4184	0.5
AGYGGCGIA**X**GGGTGAGTAA	3160	0.4
AGYGGCGIACGGGT**X**AGTAA	3020	0.4
AGYGGCGIACGG**X**TGAGTAA	2251	0.3
AGYGGCGI**X**CGGGTGAGTAA	1938	0.3
AGYGGCGIACGGGTGAGTA**X**	1816	0.2

R primers

CYIACTGCTGCCTCCCGTAG	328877	42.7
CYIACTGCTGCCTCCCG**X**AG	4935	0.6
XYIACTGCTGCC**X**CCCGTAG	3480	0.5
CYIACTGCTGCCXCCCGTAG	3034	0.4
CYI**X**CTGCTGCCTCCCGTAG	2485	0.3
XYI**X**CTGCTGCCTCCCGTAG	2409	0.3
XYIACTGC**X**GCCTCCCGTAG	1379	0.2
XYIACTGCTGCCTCCCGTAG	1174	0.2
CYIACTGC**X**GCCTCCCGTAG	1011	0.1
CYIACTGCTGCCT**X**CCGTAG	999	0.1
CYIACTGCTGXCTCCCGTAG	750	0.1
CYI**X**CTGC**X**GCCTCCCGTAG	649	0.1
AGYGGCGIACGGGTGAGTAA	589	0.1

➔ Quickly improve primers.

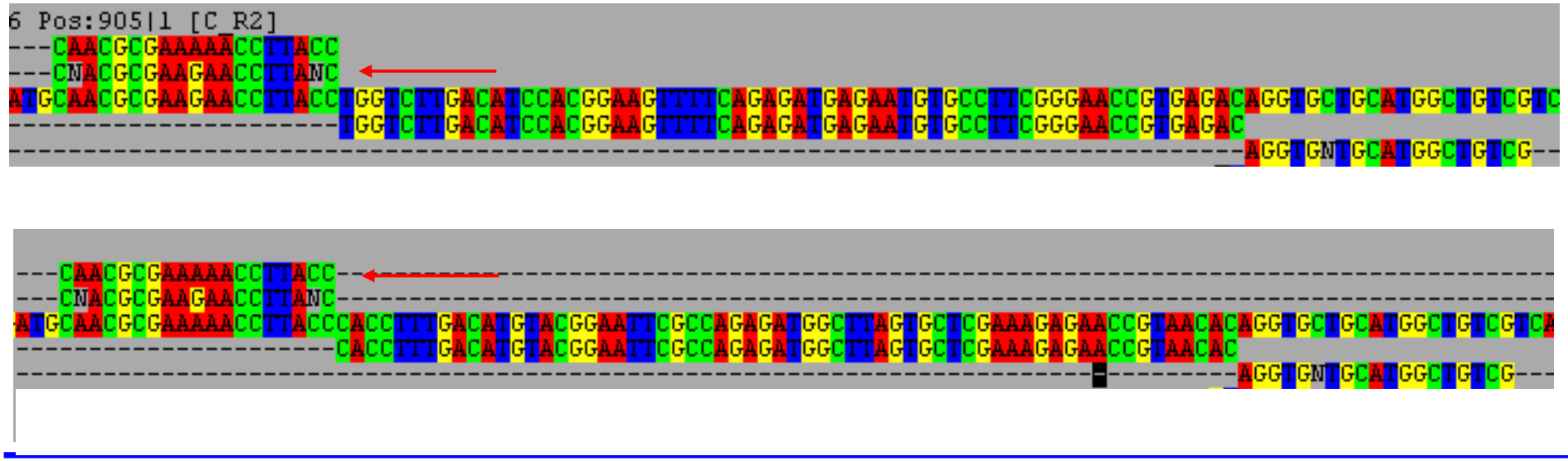
Automated taxonomic validations

	V2	V3	V6	V7_V8
Acidobacteria	26.7	31.3	19.1	13.4
Actinobacteria	46	51.9	47.8	36.9
Aquificae	63.3	68.8	70.6	27.4
Bacteroidetes	29.1	38.4	31.5	22.6
Caldus	-	100	100	100
Chlamydiae	30.5	38.3	62.5	48.8
Chlorobi	6.1	34.8	37.4	22.9
Chloroflexi	37.7	45	47.8	21.4
Chrysiogenetes	20	20	20	20
Cyanobacteria	20.6	24.4	27.9	20.8
Deferribacteres	55.7	58.2	55.7	47.5
Deinococcus-Thermus	52.8	55	57.2	48
Dictyoglomi	38.2	44.1	29.4	29.4
Fibrobacteres	84.9	96.3	87.6	74.9
Firmicutes	45.5	48.6	40	32.7
Fusobacteria	1.2	33.4	25.2	21.2
Planctomycetes	8.5	16.3	26.1	16.8
Proteobacteria	44.4	48.5	43.3	33.1
Spirochaetes	59.1	68.6	70.8	45.1
Synergistetes	75.5	76.1	44.7	29.8
Thermotogae	84.2	85.6	83.7	30
Verrucomicrobia	37.6	40.9	29.8	20.9

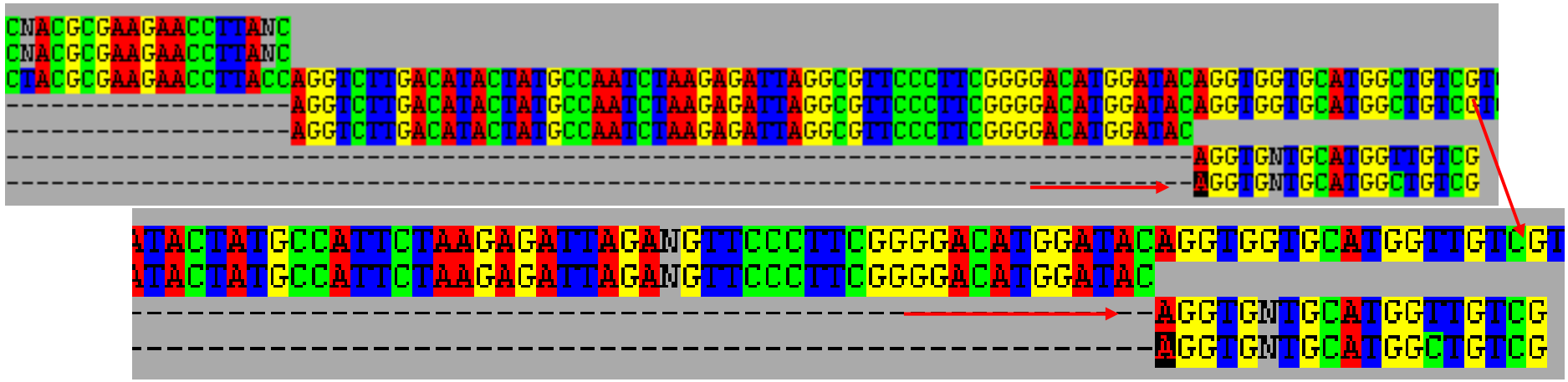
% of sequences amplified
at 2 differences,
at least 200,000 extracted tags

Multiple occurences of couples

Positions 847 & 1418 in AY706434



AY594276



Universal primers ?

Problems :

- no such thing ?
- the yield of “universal primers” is context-dependent (sequence of the domain amplified).

➔ Clade specific primers:

- Specificity ?
- Generality ?

Deinoccus specific primers

49793	31	TTTGATCCTGGCTCAGGX	1	Firmicutes
28172	17	TTTGATCCTGGCTCAGXX	2	Proteobacteria
13430	8	TTTGATCCTGGCTCAGGX	1	Bacteroidetes
13294	8	TTTGATCXTGGCTCAGXX	3	Proteobacteria
6615	4	TTTGATCCTGGCTCAGGX	1	Actinobacteria
3998	2	TTTGATCCTGGCTCAGXG	1	Proteobacteria
3584	2	TTTGATCXTGGCTCAGGX	2	Firmicutes
2019	1	TTTGATCXTGGCTCAGGX	2	Bacteroidetes
1849	1	TTTGATCCTGGCTCAGGX	1	Cyanobacteria
1831	1	TTTGATCXTGGCTCAGGX	2	Actinobacteria
1798	1	TTTGATXCTGGCTCAGGX	2	Firmicutes
1554	0	TTTGATCCTGGCTCAGXX	2	Acidobacteria
1313	0	TTTGATCCTGGCTCAGXX	2	Verrucomicrobia
1065	0	TTTGATCXTGGCTCAGXG	2	Proteobacteria
1003	0	TTTGATXXTGGCTCAGGX	3	Firmicutes
982	0	TTTGATCCTGGCTCAGGX	1	Chloroflexi
962	0	TTTGATCCTGGCTCAXGX	2	Firmicutes
893	0	TTTGATCXTGGCTCAGGX	2	Cyanobacteria
801	0	TTTGATCCTGGCTCAGXX	2	Planctomycetes

Clustering the tags

- Objectives :
  - Reduce the number of sequences for further analyses.
  - Group together sequences that may represent a unique clade.
  - Compare samples.
  - Calculate diversity indexes.
  - ...

Clustering of tags
Underlying hypotheses

- % of differences are rather meaningless.
  - we don’t have good substitution matrices.
  - We don’t know the penalties for gap & extension.
- Number of differences between two sequences is meaningful.
- PCR & 454 introduce errors, there will be a true sequence and error sequences.
  - The true sequence will have many occurences.
  - The error sequences will be rare (even more as tags are longer, not twice the same error at the same place by chance).
  - ➔ Seed the alignment starting with most abundant tags, not on longest tags as done by cd-hit or uclust !

Which algorithm

- Clustering by word counting:
  - CD-HIT
  - UCLUST

CH-HIT is very fast, UCLUST is very very fast.

They were designed to cluster protein coding sequences (banded alignments) ➔ not good for rRNA sequences (indels).

- Clustering by alignment :
  - Crunclust

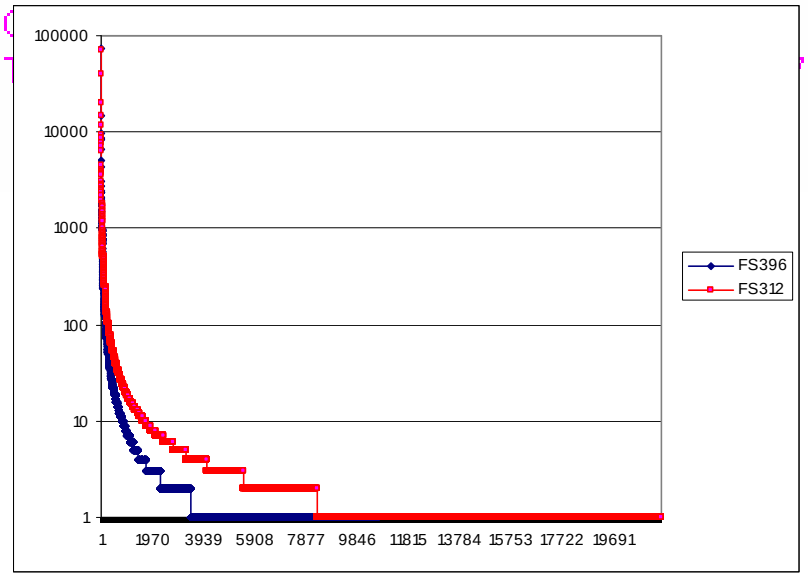
Crunchclust is fast (now faster than CD-HIT)

It was designed **specifically** to cluster 454 PCR tags.

Tag strict 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
AGCCCTTGACATCCTCGGAACTTTCTAGAGATAGATTC
AACCCTTGACATCCCTATCGCGATTTCCAGAGATGGAT                          7613

complete analysis in seconds :  1.04010820515
```



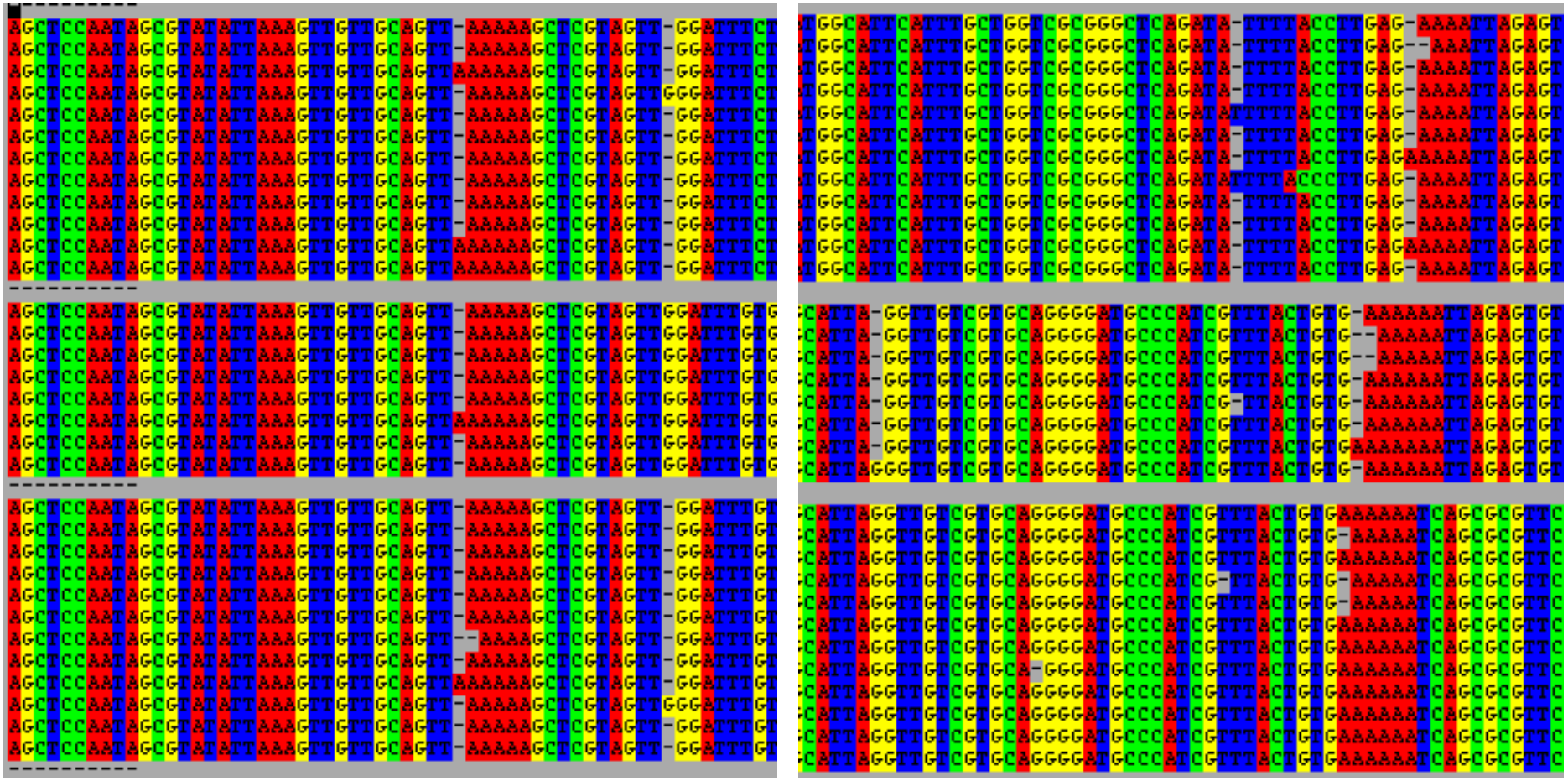
Cluster tags at k differences.

- Very fast (seconds).
- Does not require complex post analyses (Blast).
- Contrarily to Multiple Sequences Alignements, does no error.
- Allows to correct for almost 50% of 454 errors.

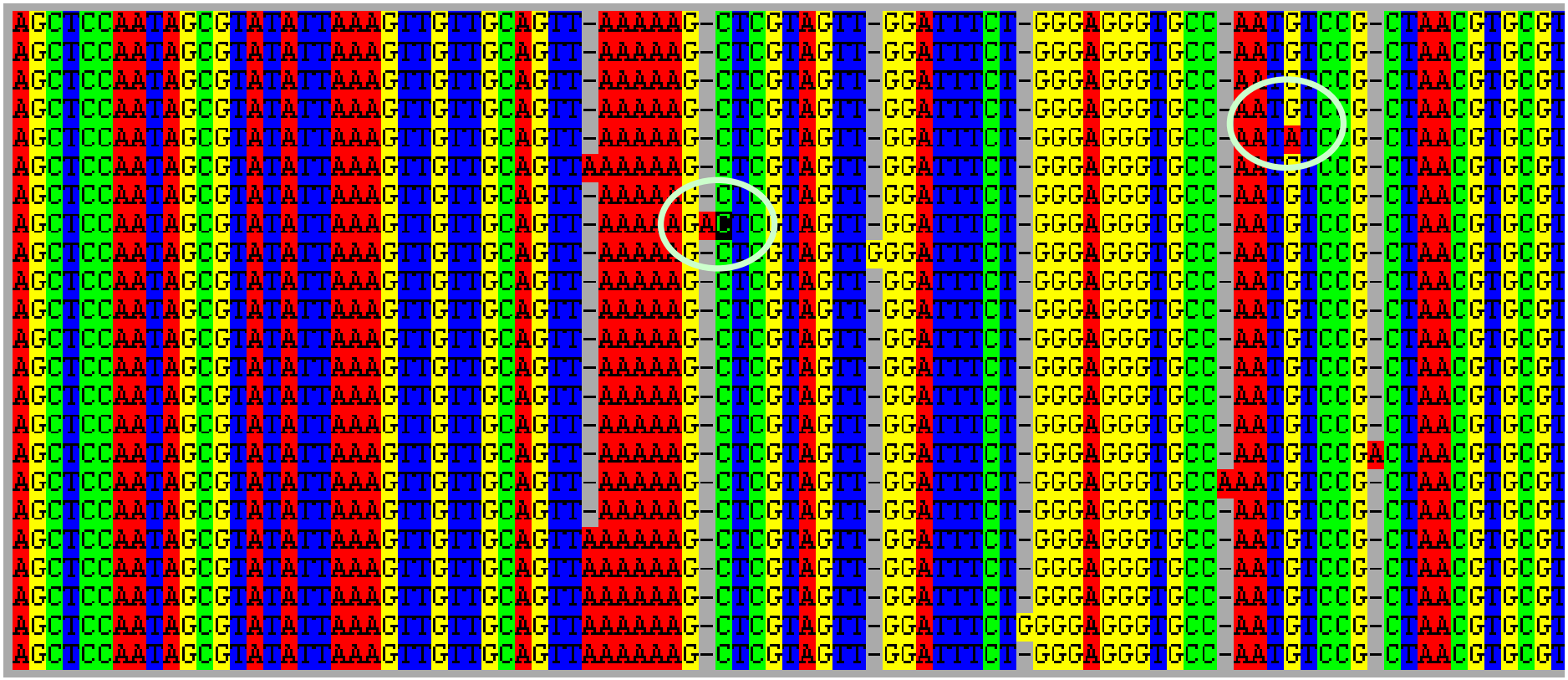
- Run on a single sample or include several samples.
  - ➔ Rank abundances.
  - ➔ Saturation curves.
  - ➔ ... In minutes.

➔ Demonstrates systematic 454 errors.

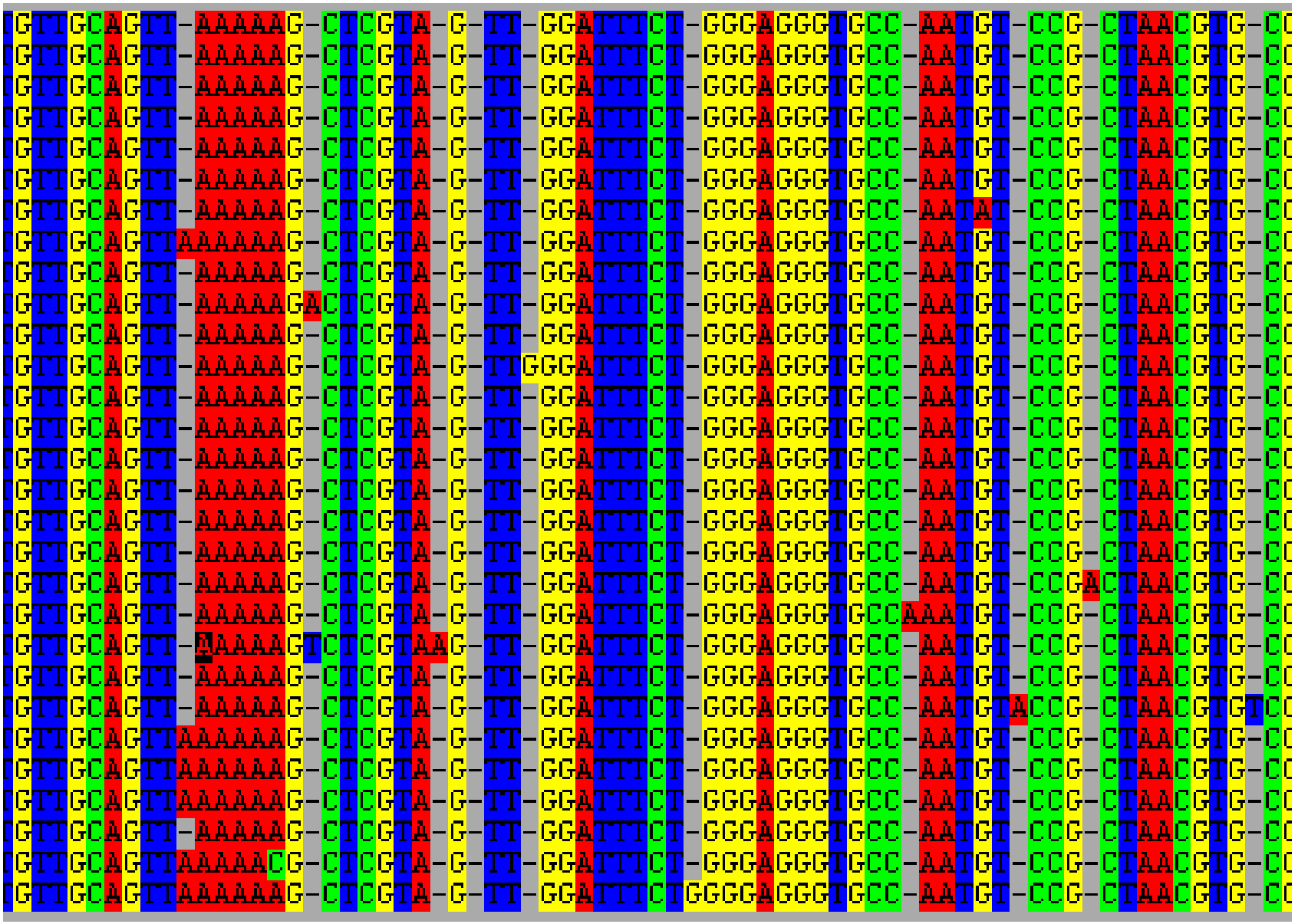
K=0



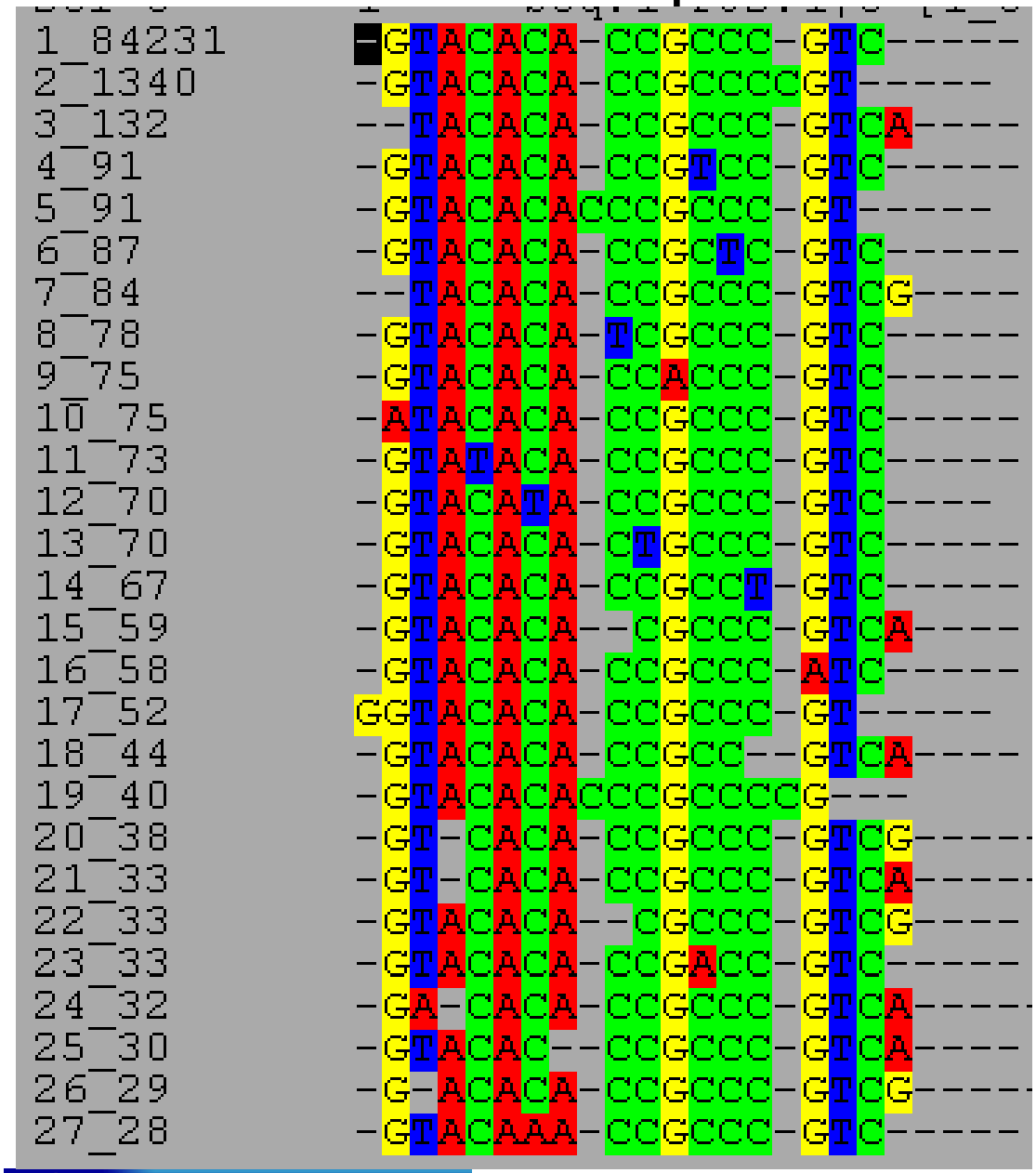
K=1



K=3

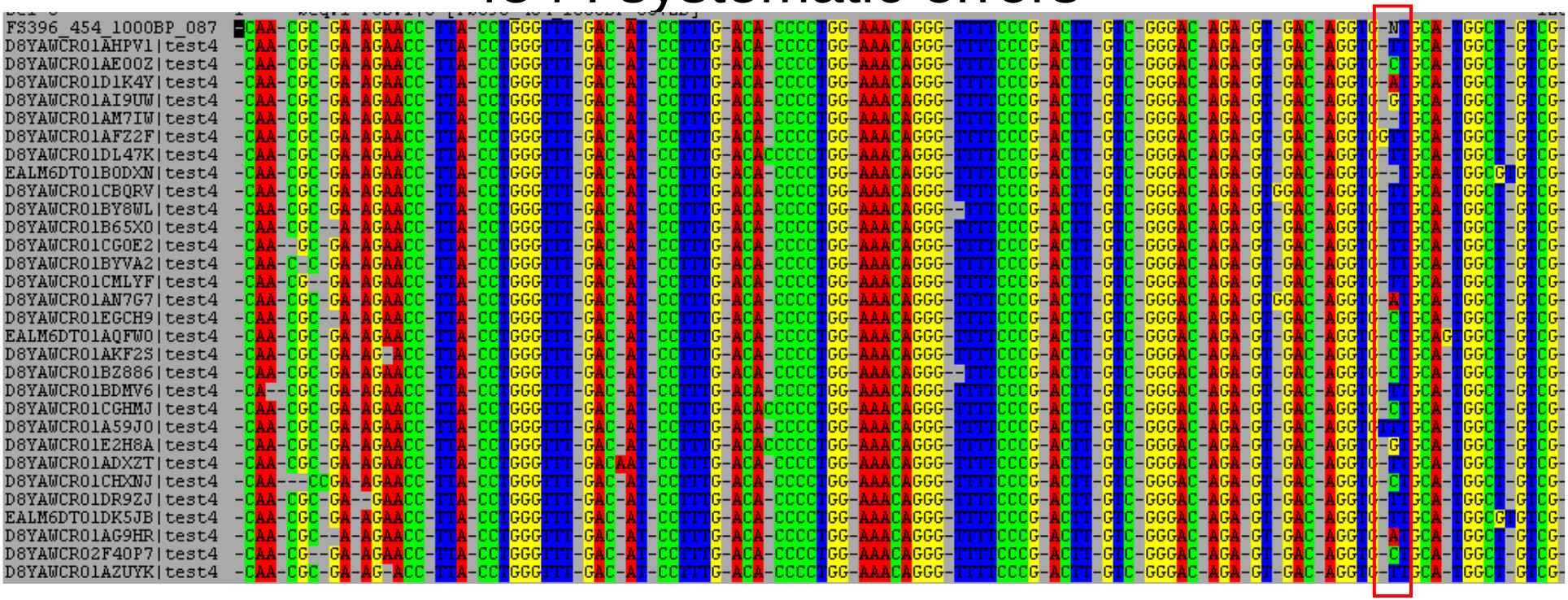


F Primers sequences



AA	A	1	0.101 %
AA	*AA*	*986*	
AA	AAA	4	0.406 %
AAA	*AAA*	*240*	
AAA	**AAAA**	**4**	**1.667 %**
CC	*CC*	*574*	
CC	CCC	5	0.871 %
CCC	**CC**	**2**	**2.469 %**
CCC	*CCC*	*81*	
GG	G	4	0.446 %
GG	*GG*	*897*	
GG	GGG	7	0.780 %
GGG	**GG**	**5**	**1.553 %**
GGG	*GGG*	*322*	
GGG	GGGG	1	0.311 %
GGGGG	**GGGG**	**5**	**3.876 %**
GGGGG	*GGGGG*	*129*	
GGGGG	**GGGGGG**	**31**	**24.031 %**
TT	T	4	0.985 %
TT	*TT*	*406*	
TT	TTT	2	0.493 %
TTT	*TTT*	*80*	
TTT	**TTTT**	**1**	**1.250 %**

454 : systematic errors



Accuracy and quality of massively parallel DNA pyrosequencing
Huse, Huber, Morrison,Sogin, and Welch. Genome Biol. 2007; 8(7): R143

➔Most errors are corrected at 1 difference.

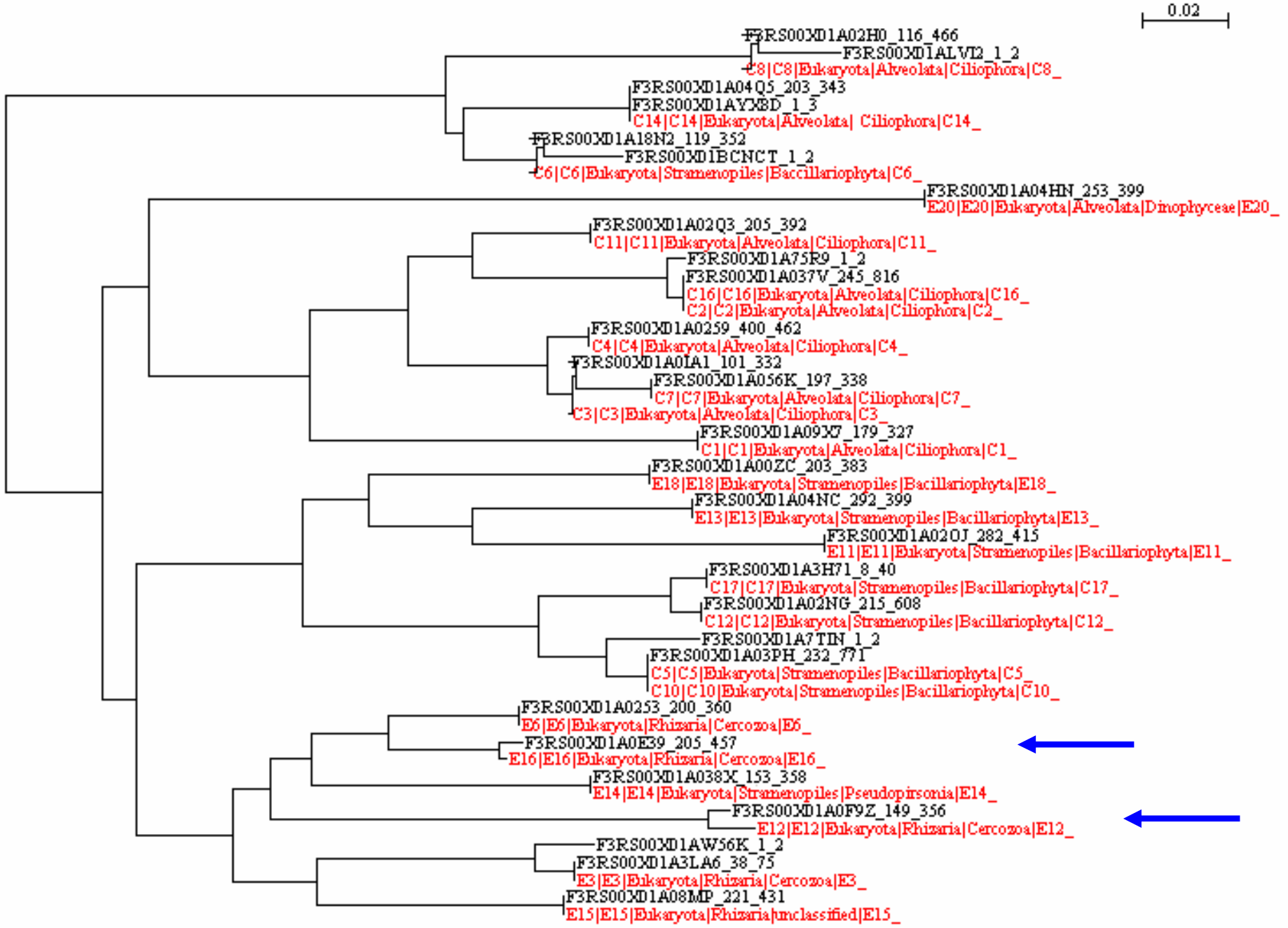
➔Discard single singletons at 1 difference.

- **Singleton : a tag which is found only once in experiment(s).**
- **Single singleton : a cluster at k (1) difference(s) that:**
  - **Contains a single member.**
  - **This member is a singleton.**

CC/UC V4 tita

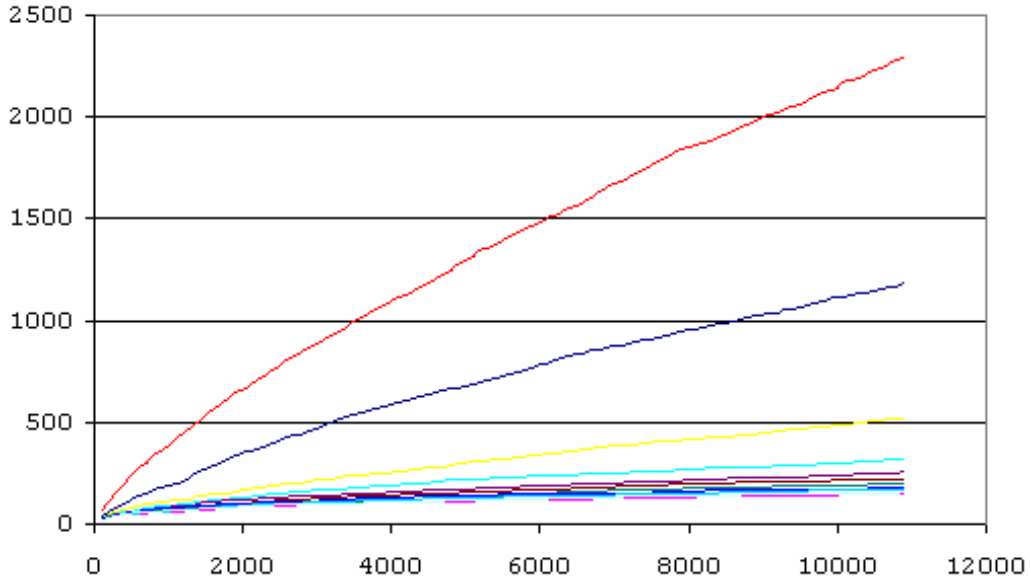
unique/occur		1085/8916		cc	target	cc/ta	theorey %						
k0	100	645	514	131	24	5.5	445.8						
k1	99.5	370	254	116	24	4.8	383.3						
k2	99	174	108	66	23	2.9	175.0						
k3	98.5	98	58	40	23	1.7	66.7						
k4	98	61	30	31	22	1.4	29.2	62	18	44	21	2.1	83.3
k5	97.5	51	23	28	21	1.3	16.7						
k6	97	45	18	27	21	1.3	12.5						
k7	96.5	38	12	26	21	1.2	8.3						
k8	96	36	10	26	21	1.2	8.3	27	3	24	19	1.3	0.0
k9	95.5	34	9	25	21	1.2	4.2						
k10	95	34	9	25	21	1.2	4.2						
k11	94.5	34	9	25	20	1.3	4.2						
k12	94	34	9	25	20	1.3	4.2	21	1	20	17	1.2	-16.7
k13	93.5	34	9	25	20	1.3	4.2						
k14	93	34	9	25	20	1.3	4.2						
k12	92.5	34	9	25	20	1.3	4.2						
k13	92	34	9	25	20	1.3	4.2						
k14	91.5	34	9	25	20	1.3	4.2	19	0	19	15	1.3	-20.8

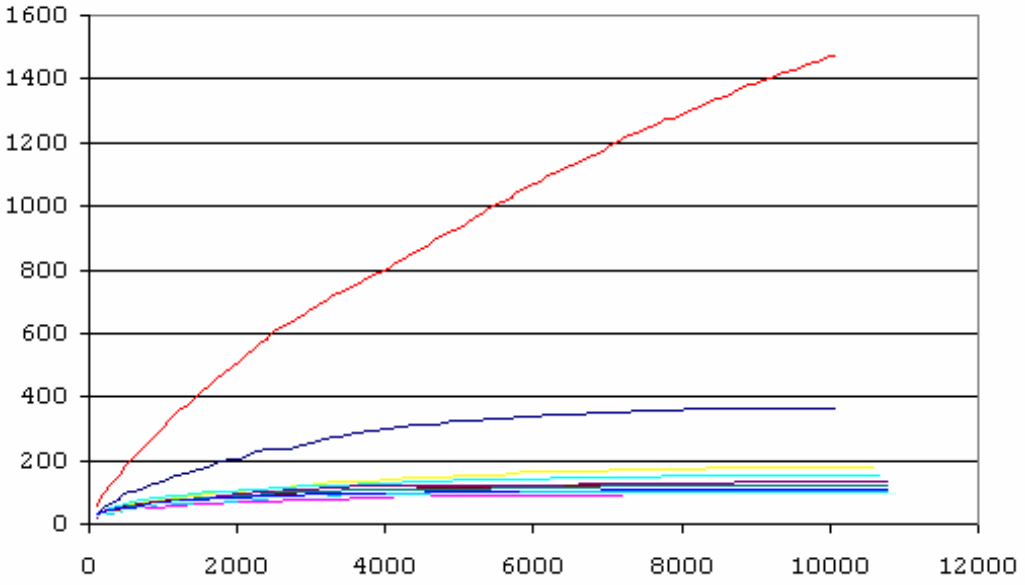
CC no ss k=5


0.02
F3RS00XD1A02H0_116_466
F3RS00XD1ALVL2_1_2
C8|C8|Eukaryota|Alveolata|Ciliophora|C8_
F3RS00XD1A04Q5_203_343
F3RS00XD1AYX8D_1_3
C14|C14|Eukaryota|Alveolata| Ciliophora|C14_
F3RS00XD1A18N2_119_352
F3RS00XD1BCNCT_1_2
C6|C6|Eukaryota|Stramenopiles|Baccillariophyta|C6_
F3RS00XD1A04HN_253_399
E20|E20|Eukaryota|Alveolata|Dinophyceae|E20_
F3RS00XD1A02Q3_205_392
C11|C11|Eukaryota|Alveolata|Ciliophora|C11_
F3RS00XD1A75R9_1_2
F3RS00XD1A037V_245_816
C16|C16|Eukaryota|Alveolata|Ciliophora|C16_
C2|C2|Eukaryota|Alveolata|Ciliophora|C2_
F3RS00XD1A0259_400_462
C4|C4|Eukaryota|Alveolata|Ciliophora|C4_
F3RS00XD1A0IA1_101_332
F3RS00XD1A056K_197_338
C7|C7|Eukaryota|Alveolata|Ciliophora|C7_
C3|C3|Eukaryota|Alveolata|Ciliophora|C3_
F3RS00XD1A09X7_179_327
C1|C1|Eukaryota|Alveolata|Ciliophora|C1_
F3RS00XD1A00ZC_203_383
E18|E18|Eukaryota|Stramenopiles|Bacillariophyta|E18_
F3RS00XD1A04NC_292_399
E13|E13|Eukaryota|Stramenopiles|Bacillariophyta|E13_
F3RS00XD1A02OJ_282_415
E11|E11|Eukaryota|Stramenopiles|Bacillariophyta|E11_
F3RS00XD1A3H71_8_40
C17|C17|Eukaryota|Stramenopiles|Bacillariophyta|C17_
F3RS00XD1A02NG_215_608
C12|C12|Eukaryota|Stramenopiles|Bacillariophyta|C12_
F3RS00XD1A7TIN_1_2
F3RS00XD1A03PH_232_771
C5|C5|Eukaryota|Stramenopiles|Bacillariophyta|C5_
C10|C10|Eukaryota|Stramenopiles|Bacillariophyta|C10_
F3RS00XD1A0253_200_360
E6|E6|Eukaryota|Rhizaria|Cercozoa|E6_
F3RS00XD1A0E39_205_457
E16|E16|Eukaryota|Rhizaria|Cercozoa|E16_
F3RS00XD1A038X_153_358
E14|E14|Eukaryota|Stramenopiles|Pseudopirsonia|E14_
F3RS00XD1A0F9Z_149_356
E12|E12|Eukaryota|Rhizaria|Cercozoa|E12_
F3RS00XD1AW56K_1_2
F3RS00XD1A3LA6_38_75
E3|E3|Eukaryota|Rhizaria|Cercozoa|E3_
F3RS00XD1A08MP_221_431
E15|E15|Eukaryota|Rhizaria|unclassified|E15_

Saturation curves

Saturation curves at
at 0,1,2,3,4,5,... differences.





Same, but
discard single singletons

Assign taxonomy

- There is no genome to map on !
- Good quality annotations are in the database of RefSeq ! There is no RefSeq or Uniprot ...
- Which annotation process ?
  - Which algo (blast open, extend...)
  - Which % similarity for which taxonomy (phylum, class, ... Species ?).
    - May depend upon the clade !
    - Depends upon the domain amplified !
      - Bacteria: V6 & V9 (SSU).
      - Eukaryota: V4 & V9 (SSU).
      - Other molecules: LSU rRNA, house keeping genes (single copy).

Assign Taxonomy

p95-a75			95 best	95 best hit
8	Amoebozoa		8	8
590	Archaeplastida		971	971
17705	Chromalveolata		18033	17701
65	Incertae_sedis_Eukaryota		76	65
616	Opisthokonta		616	616
249	Rhizaria		348	249
8	Amoebozoa	Lobosa	8	8
537	Archaeplastida	Chlorophyta	918	918
53	Archaeplastida	Cryptophyta-nucleomorp	53	53
7588	Chromalveolata	Alveolata	7791	7584
615	Chromalveolata	Cryptophyta	660	615
591	Chromalveolata	Haptophyta	600	591
24	Chromalveolata	Katablepharidophyta	24	24
569	Chromalveolata	Picobiliphyta	569	569
8318	Chromalveolata	Stramenopiles	8389	8318
	Incertae_sedis_Eukaryota	Apusomonadidae	3	
47	Incertae_sedis_Eukaryota	Telonemia	47	47
18	Incertae_sedis_Eukaryota	Undetermined_lineage	26	18
198	Opisthokonta	Choanoflagellida	198	198
16	Opisthokonta	Fungi	16	16
402	Opisthokonta	Metazoa	402	402
233	Rhizaria	Cercozoa	325	233
16	Rhizaria	Foraminifera	20	16
	Rhizaria	Radiolaria	3	
19233			20052	19610
28286				

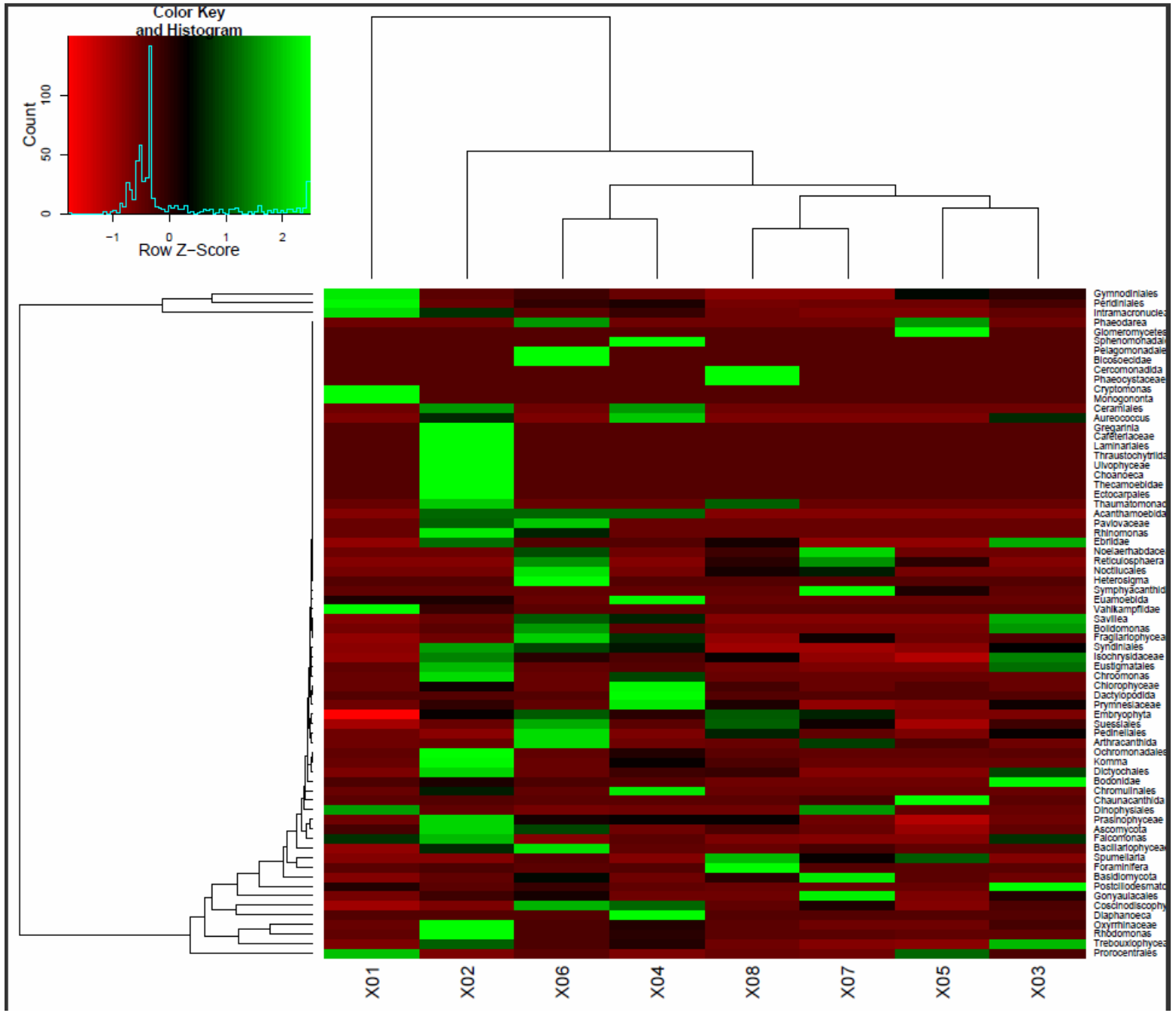
Numbers of 16S rRNA sequences per species

	>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.


Color Key
and Histogram
Count
0
50
100
−1
0
1
2
Row Z−Score
X01
X02
X06
X04
X08
X07
X05
X03

Current Problems

- Choose domain to amplify
- Choose primers
- Cluster tags
- Assign taxonomy
- Compare samples

- Raw data (images are lost).
- Store tags, taxonomy and metadata in a secure manner.
  - SRA takes only .sff files.
  - Project in development with INIST.
- Query “analyzed” datasets.
  - By similarity.
  - By taxonomy.
  - By metadata (pH, ℃, salinity, ...).
- Cloud computing, GPU computing, blades of CPU ?
- Dedicated algorithms !
- Bandwith transfert problems ?
- Build RefSeq database of good, well annotated sequences
  - Silva for Bacteria.
  - In progress for Eukaryota (col. Laure Guillou, Roscoff).
- A dedicated ontology is now required.