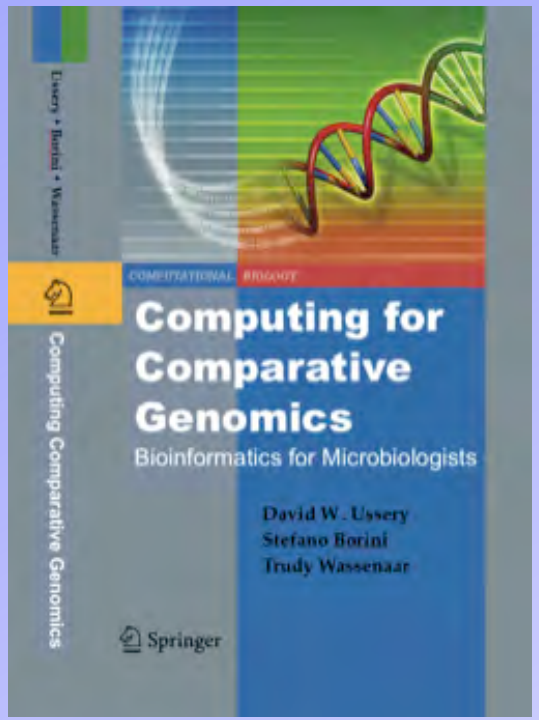
Too Much Data -

Cautionary Tales of Next-generation and Next-next Generation Sequencing



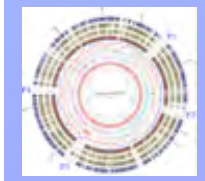
Dave Ussery

Workshop on Comparative Genomics
King Mongkut's University of Technology Thonburi
Bangkok, Thailand

1rst Talk for Tuesday
8 March, 2010

Peppes pizza

www.cbs.dtu.dk

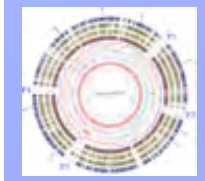


Outline

http://www.cbs.dtu.dk/courses/thaiworkshop2010/

Course outline

	Monday, 8 March	Tuesday, 9 March	Wednesday, 10 March	Thursday, 11 March	Friday, 12 March
9:30-10:30	Introduction to Genomics	Methods to Compare Genomes	Expression of Genes and Proteins	Pan- and core-genomes	An overview and questions for students
10:30-11:00 (Coffee Break)					
11:00-12:00	Sequences as biological information	Genome Annotation	Comparison of Bacterial Proteomes	Microbial Communities	Groups meet to discuss projects
12:00-13:00 (Lunch break)					
13:00-15:00	Downloading Genomes	finding RNA genes	BLAST matrices	Pan- Core-genome plots	Group presentations
15:00-15:30 (Coffee break)					
15:30-17:00	Prediction of genes	Construction of 16S rRNA trees	Genome atlases	BLAST atlases	Group presentations, certificate presentation, closing



Subject: What's new for '"complete genome"' in PubMed
From: My NCBI <efback@mail.nih.gov>
Date: 7 March, 2010 1:20:10 AM GMT+07:00
To: Dave Ussery <dave@cbs.dtu.dk>

Sender's message:
Sent on Saturday, 2010 Mar 06
Search "complete genome"
Click here to view complete results in PubMed. (Results may change over time.)
To unsubscribe from these e-mail updates click here.

PubMed Results
Items 1 - 12 of 12

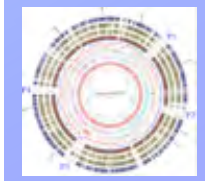
2. Genomic Structure of an Economically Important Cyanobacterium, Arthrospira (Spirulina) platensis NIES-39.

Fujisawa T, Narikawa R, Okamoto S, Ehira S, Yoshimura H, Suzuki I, Masuda T, Mochimaru M, Takaichi S, Awai K, Sekine M, Horikawa H, Yashiro I, Omata S, Takarada H, Katano Y, Kosugi H, Tanikawa S, Ohmori K, Sato N, Ikeuchi M, Fujita N, Ohmori M.

DNA Res. 2010 Mar 4. [Epub ahead of print]

PMID: 20203057 [PubMed - as supplied by publisher]

Comparative Microbial Genomics group



DNA Research Advance Access published March 4, 2010

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

DNA Research pp. 1–19, (2010) doi:10.1093/dnares/dsq004

Genomic Structure of an Economically Important Cyanobacterium, *Arthrospira* (*Spirulina*) *platensis* NIES-39

Takatomo Fujisawa[1], Rei Narikawa[2], Shinobu Okamoto[3], Shigeki Ehira[4], Hidehisa Yoshimura[2], Iwane Suzuki[5], Tatsuru Masuda[2], Mari Mochimaru[6], Shinichi Takaichi[7], Koichiro Awai[8], Mitsuo Sekine[1], Hiroshi Horikawa[1], Isao Yashiro[1], Seiha Omata[1], Hiromi Takarada[1], Yoko Katano[1], Hiroki Kosugi[1], Satoshi Tanikawa[1], Kazuko Ohmori[9], Naoki Sato[2], Masahiko Ikeuchi[2], Nobuyuki Fujita[1,*], and Masayuki Ohmori[4]

Abstract

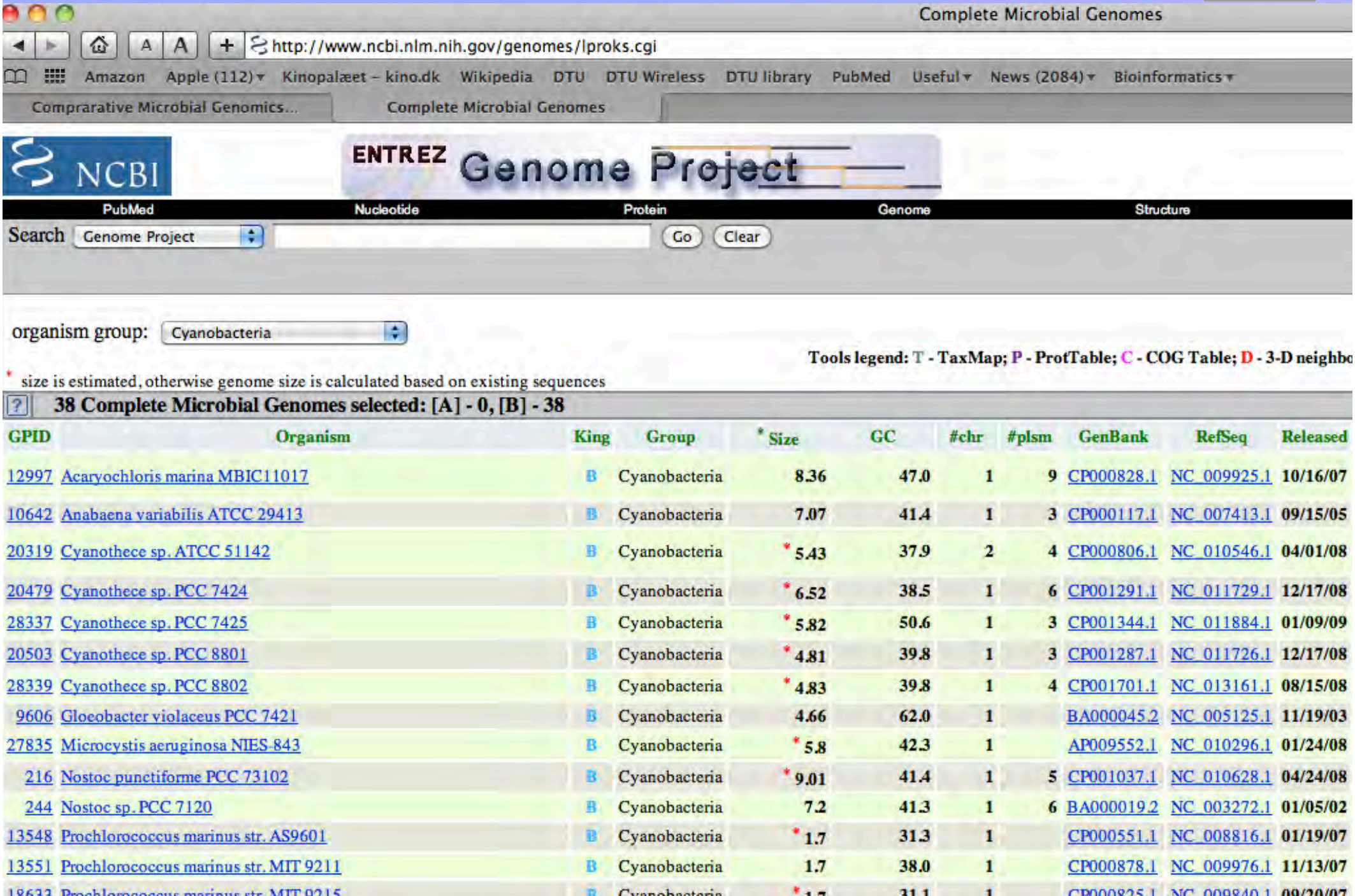
A filamentous non-N_2-fixing cyanobacterium, *Arthrospira* (*Spirulina*) *platensis*, is an important organism for industrial applications and as a food supply. Almost the complete genome of *A. platensis* NIES-39 was determined in this study. The genome structure of *A. platensis* is estimated to be a single, circular chromosome of 6.8 Mb, based on optical mapping. Annotation of this 6.7 Mb sequence yielded 6630 protein-coding genes as well as two sets of rRNA genes and 40 tRNA genes. Of the protein-coding genes, 78% are similar to those of other organisms; the remaining 22% are currently unknown. A total 612 kb of the genome comprise group II introns, insertion sequences and some repetitive elements. Group I introns are located in a protein-coding region. Abundant restriction-modification systems were determined. Unique features in the gene composition were noted, particularly in a large number of genes for adenylate cyclase and haemolysin-like Ca^{2+}-binding proteins and in chemotaxis proteins. Filament-specific genes were highlighted by comparative genomic analysis.

Key words: cyanobacteria; *Arthrospira*; health supplement; genome; cAMP

Center for Biological Sequence analysis | Department of Systems Biology, Technical University of Denmark

DTU

The genome sequence and annotation of *A. platensis* NIES-39 are available at GenBank/EMBL/DDBJ under accession no. AP011615



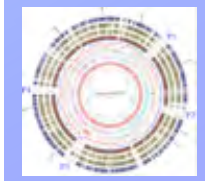
organism group: Cyanobacteria

Tools legend: T - TaxMap; P - ProtTable; C - COG Table; D - 3-D neighbo

* size is estimated, otherwise genome size is calculated based on existing sequences

38 Complete Microbial Genomes selected: [A] - 0, [B] - 38

GPID	Organism	King	Group	* Size	GC	#chr	#plsm	GenBank	RefSeq	Released
12997	Acaryochloris marina MBIC11017	B	Cyanobacteria	8.36	47.0	1	9	CP000828.1	NC_009925.1	10/16/07
10642	Anabaena variabilis ATCC 29413	B	Cyanobacteria	7.07	41.4	1	3	CP000117.1	NC_007413.1	09/15/05
20319	Cyanothece sp. ATCC 51142	B	Cyanobacteria	* 5.43	37.9	2	4	CP000806.1	NC_010546.1	04/01/08
20479	Cyanothece sp. PCC 7424	B	Cyanobacteria	* 6.52	38.5	1	6	CP001291.1	NC_011729.1	12/17/08
28337	Cyanothece sp. PCC 7425	B	Cyanobacteria	* 5.82	50.6	1	3	CP001344.1	NC_011884.1	01/09/09
20503	Cyanothece sp. PCC 8801	B	Cyanobacteria	* 4.81	39.8	1	3	CP001287.1	NC_011726.1	12/17/08
28339	Cyanothece sp. PCC 8802	B	Cyanobacteria	* 4.83	39.8	1	4	CP001701.1	NC_013161.1	08/15/08
9606	Gloeobacter violaceus PCC 7421	B	Cyanobacteria	4.66	62.0	1		BA000045.2	NC_005125.1	11/19/03
27835	Microcystis aeruginosa NIES-843	B	Cyanobacteria	* 5.8	42.3	1		AP009552.1	NC_010296.1	01/24/08
216	Nostoc punctiforme PCC 73102	B	Cyanobacteria	* 9.01	41.4	1	5	CP001037.1	NC_010628.1	04/24/08
244	Nostoc sp. PCC 7120	B	Cyanobacteria	7.2	41.3	1	6	BA000019.2	NC_003272.1	01/05/02
13548	Prochlorococcus marinus str. AS9601	B	Cyanobacteria	* 1.7	31.3	1		CP000551.1	NC_008816.1	01/19/07
13551	Prochlorococcus marinus str. MIT 9211	B	Cyanobacteria	1.7	38.0	1		CP000878.1	NC_009976.1	11/13/07
18633	Prochlorococcus marinus str. MIT 9215	B	Cyanobacteria	* 1.7	31.1	1		CP000825.1	NC_009840.1	09/20/07

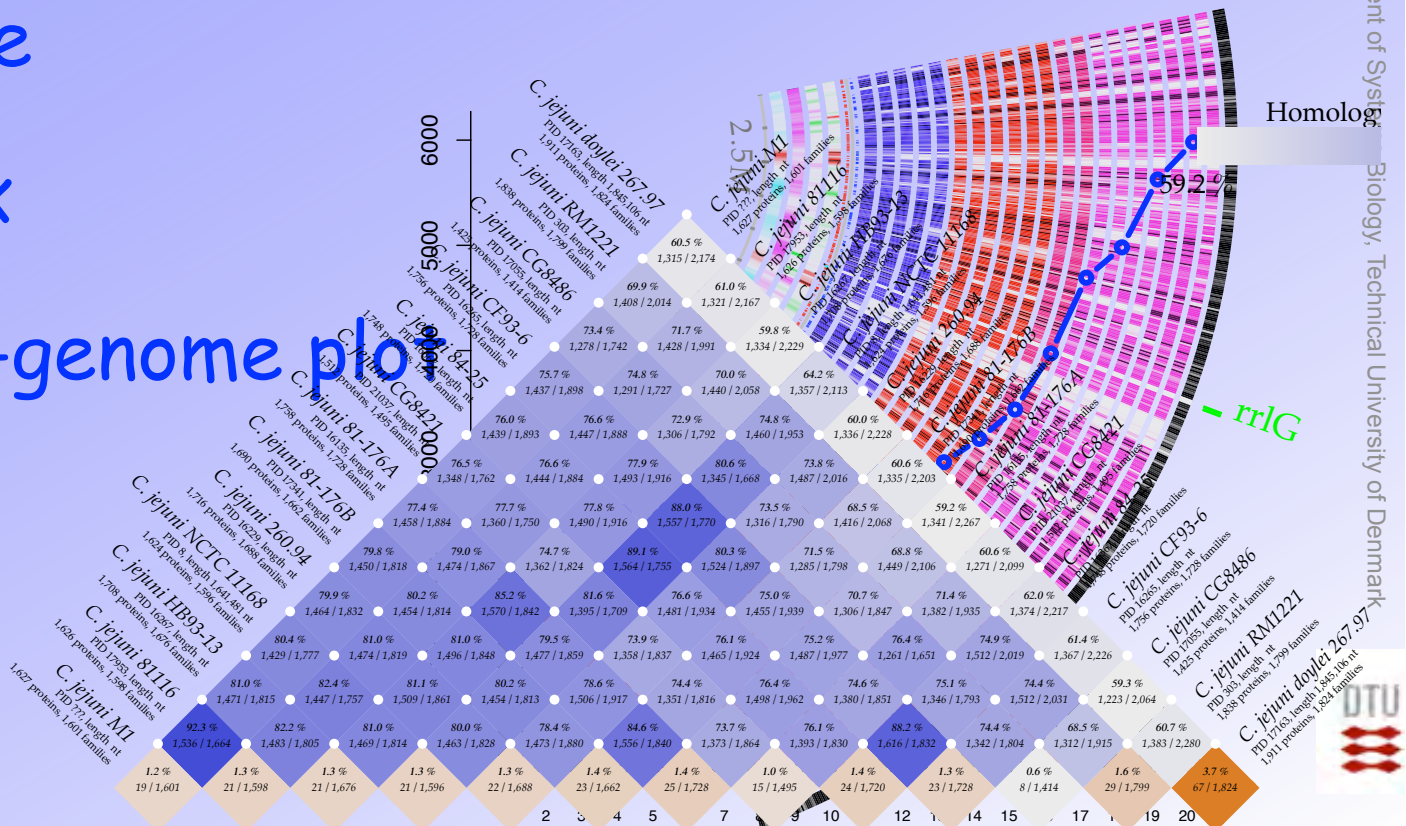


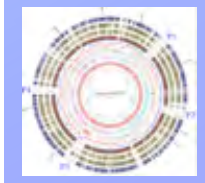
CENTERFOR BIOLOGICAL SEQUENCE ANALYSIS CBS

~40 students, with 7 genomes per student.

- What can be included in the presentations on Friday?

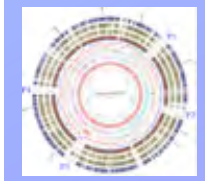
1. Table of related genomes finished and in GenBank
2. 16S rRNA tree
3. BLAST Matrix
4. Core- and pan-genome plots
5. BLAST Atlas





Outline

- The problem - too much data!
- A brief history - The speed of sequencing
- Cautionary tales
- Some approaches to handle this....



1. The problem - too much data!

Technology

The data deluge

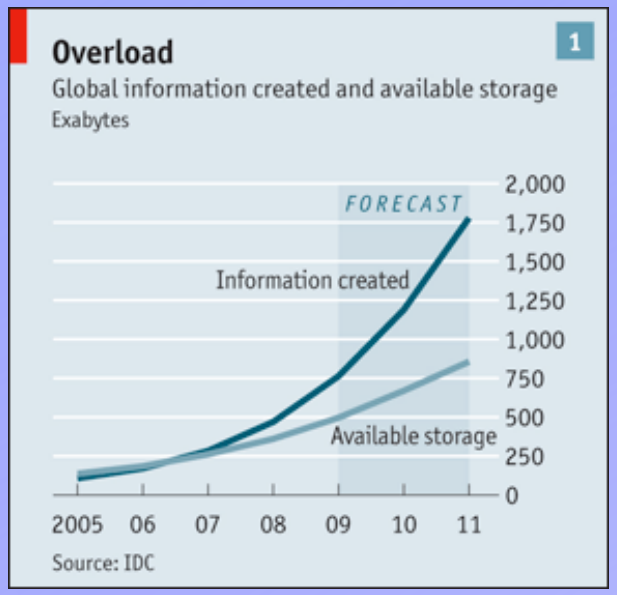
Businesses, governments and society are only starting to tap its vast potential

Feb 25th 2010 | From *The Economist* print edition

EIGHTEEN months ago, Li & Fung, a firm that manages supply chains for retailers, saw 100 gigabytes of information flow through its network each day. Now the amount has increased tenfold. During 2009, American drone aircraft flying over Iraq and Afghanistan sent back around 24 years' worth of video footage. New models being deployed this year will produce ten times as many data streams as their predecessors, and those in 2011 will produce 30 times as many.

Everywhere you look, the quantity of information in the world is soaring. According to one estimate, mankind created 150 exabytes (billion gigabytes) of data in 2005. This year, it will create 1,200 exabytes. Merely keeping up with this flood, and storing the bits that might be useful, is difficult enough. Analysing it, to spot patterns and extract useful information, is harder still. Even so, the data deluge is already starting to transform business, government, science and everyday life (see our special report in this issue). It has great potential for good—as long as consumers, companies and governments make the right choices about when to restrict the flow of data, and when to encourage it.

1. The problem - too much data!



Data inflation

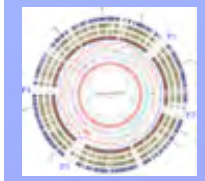
2

Unit	Size	What it means
Bit (b)	1 or 0	Short for "binary digit", after the binary code (1 or 0) computers use to store and process data
Byte (B)	8 bits	Enough information to create an English letter or number in computer code. It is the basic unit of computing
Kilobyte (KB)	1,000, or 2^{10}, bytes	From "thousand" in Greek. One page of typed text is 2KB
Megabyte (MB)	1,000KB; 2^{20} bytes	From "large" in Greek. The complete works of Shakespeare total 5MB. A typical pop song is about 4MB
Gigabyte (GB)	1,000MB; 2^{30} bytes	From "giant" in Greek. A two-hour film can be compressed into 1-2GB
Terabyte (TB)	1,000GB; 2^{40} bytes	From "monster" in Greek. All the catalogued books in America's Library of Congress total 15TB
Petabyte (PB)	1,000TB; 2^{50} bytes	All letters delivered by America's postal service this year will amount to around 5PB. Google processes around 1PB every hour
Exabyte (EB)	1,000PB; 2^{60} bytes	Equivalent to 10 billion copies of *The Economist*
Zettabyte (ZB)	1,000EB; 2^{70} bytes	The total amount of information in existence this year is forecast to be around 1.2ZB
Yottabyte (YB)	1,000ZB; 2^{80} bytes	Currently too big to imagine

Source: *The Economist*

The prefixes are set by an intergovernmental group, the International Bureau of Weights and Measures. Yotta and Zetta were added in 1991; terms for larger amounts have yet to be established.

27 February, 2010 | From *The Economist* print edition



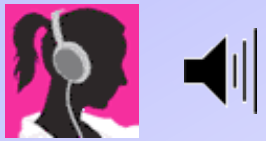
1. The problem - too much data!

How to visualize lots of data....

nature International weekly journal of science

Volume 455 Number 7209 pp1-136

4 September, 2008



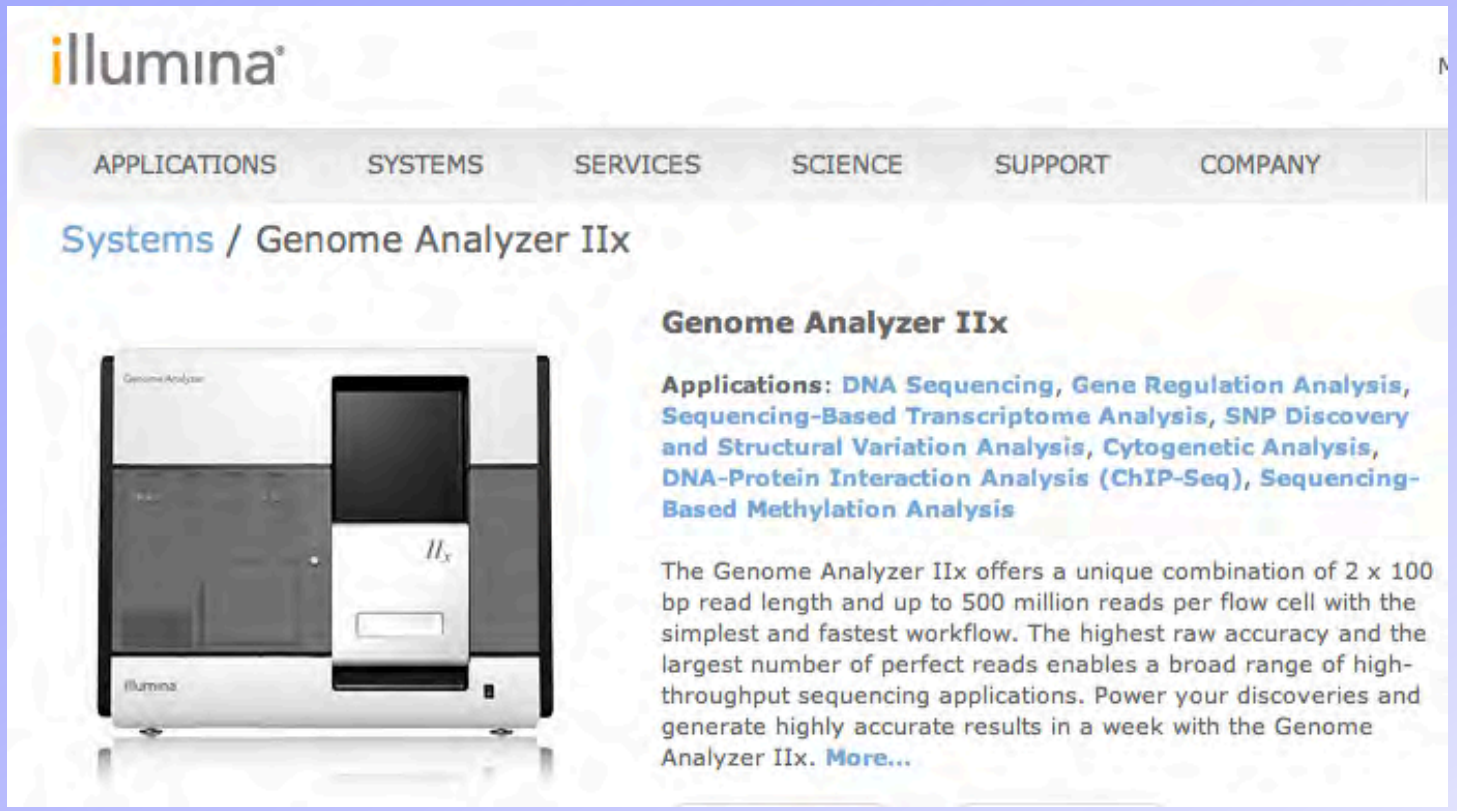
Nature podcast

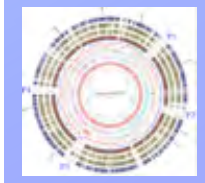
In Nature this week, features and opinion pieces on one of the most daunting challenges facing modern science: how to cope with the flood of data now being generated. A petabyte is a lot of memory, however you say it - a quadrillion, 10^{15}, or tens of thousands of trillions of bytes. But that is the currency of 'big data'. We visited the Sanger Institute's supercomputing centre, and its petabyte of capacity. [News Feature p. 16]

1. The problem - too much data!

Three Current "next-generation" technologies:

1. illumina (aka "Solexa") - 500 million reads (100 bp)

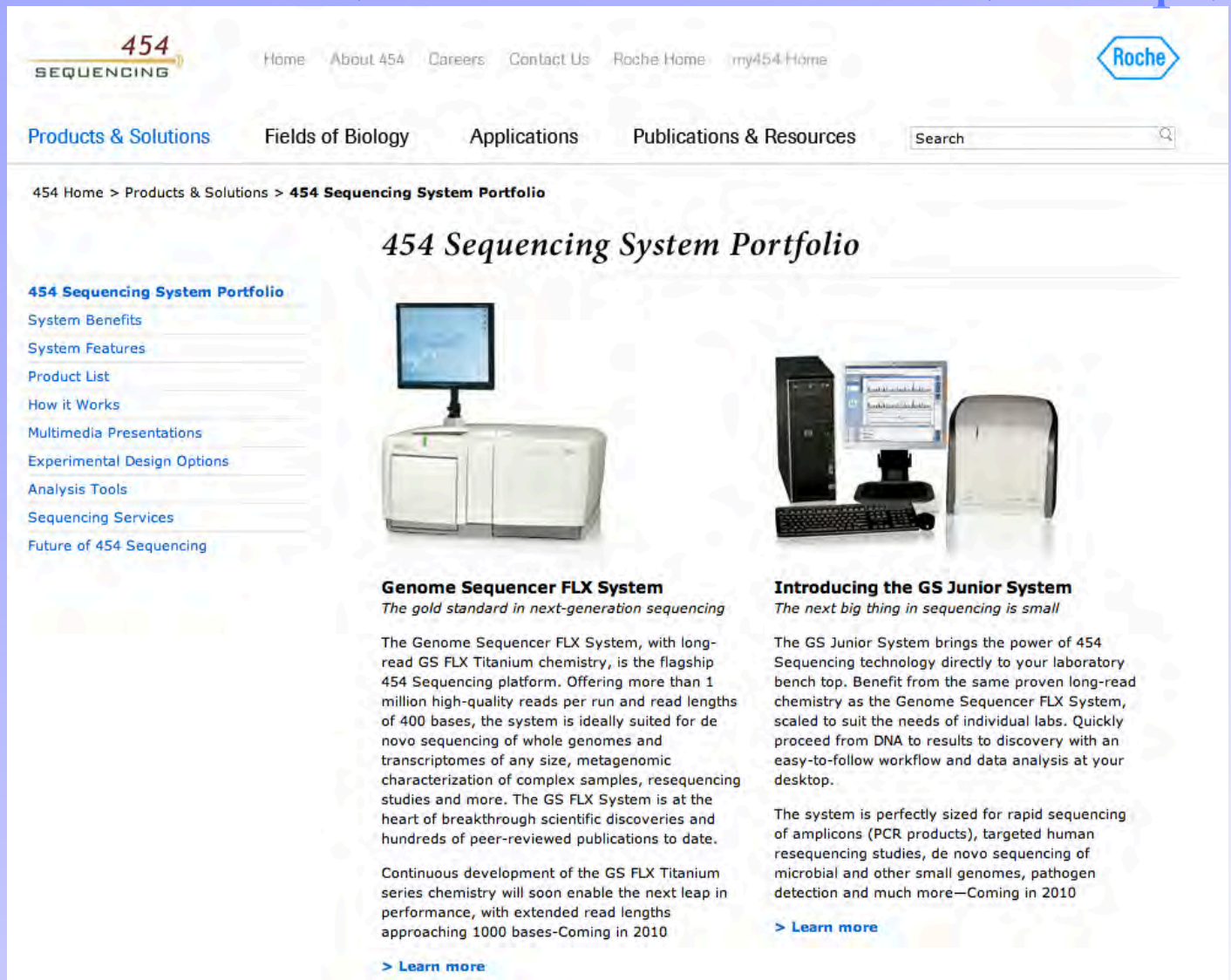




1. The problem - too much data!

Three Current "next-generation" technologies:

1. illumina (aka "Solexa") - 500 million reads (100 bp)
2. Roche 454



454 SEQUENCING

Home About 454 Careers Contact Us Roche Home my454 Home

Roche

Products & Solutions | Fields of Biology | Applications | Publications & Resources | Search

454 Home > Products & Solutions > **454 Sequencing System Portfolio**

454 Sequencing System Portfolio

454 Sequencing System Portfolio
System Benefits
System Features
Product List
How it Works
Multimedia Presentations
Experimental Design Options
Analysis Tools
Sequencing Services
Future of 454 Sequencing

Genome Sequencer FLX System
The gold standard in next-generation sequencing

The Genome Sequencer FLX System, with long-read GS FLX Titanium chemistry, is the flagship 454 Sequencing platform. Offering more than 1 million high-quality reads per run and read lengths of 400 bases, the system is ideally suited for de novo sequencing of whole genomes and transcriptomes of any size, metagenomic characterization of complex samples, resequencing studies and more. The GS FLX System is at the heart of breakthrough scientific discoveries and hundreds of peer-reviewed publications to date.

Continuous development of the GS FLX Titanium series chemistry will soon enable the next leap in performance, with extended read lengths approaching 1000 bases-Coming in 2010

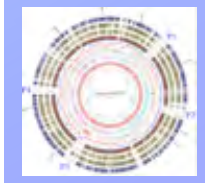
> Learn more

Introducing the GS Junior System
The next big thing in sequencing is small

The GS Junior System brings the power of 454 Sequencing technology directly to your laboratory bench top. Benefit from the same proven long-read chemistry as the Genome Sequencer FLX System, scaled to suit the needs of individual labs. Quickly proceed from DNA to results to discovery with an easy-to-follow workflow and data analysis at your desktop.

The system is perfectly sized for rapid sequencing of amplicons (PCR products), targeted human resequencing studies, de novo sequencing of microbial and other small genomes, pathogen detection and much more—Coming in 2010

> Learn more



1. The problem - too much data!

Three Current "next-generation" technologies:

1. illumina (aka "Solexa") - 500 million reads (100 bp)
2. Roche 454 - > 1 million reads (1000 bp)
3. ABI SOLiD

~100 Gbp per run!
35 bp reads

SPECIFICATION SHEET

Applied Biosystems® SOLiD™ 4 System

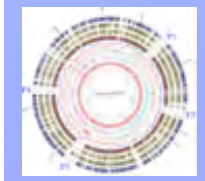
applied biosystems
part of life technologies

SOLiD™ 4
SYSTEM SEQUENCING

Key Benefits

- **Higher accuracy**—detection of causative variation enabled at lower coverage and cost per sample
- **Scalable throughput on a single platform**—80–100 GB of mappable sequence per run
- **Automated workflow**—80% reduction in hands-on time and increased reproducibility in yield allow for significant time and labor savings
- **True paired-end sequencing**—bidirectional sequencing facilitates detection of genetic alterations as well as splice variants and fusion transcripts with lower sample input
- **Robust multiplexing kits**—intelligent barcode strategy enables accurate assignment without introduction of bias

DTU



1. The problem - too much data!

Next-Generation DNA Sequencing/Review

The new paradigm of flow cell sequencing

Robert A. Holt[1] and Steven J.M. Jones

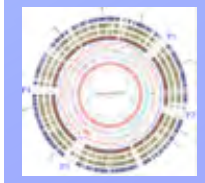
Genome Research, Jan 2009

"Indeed, any of these new machines running at full capacity for a year will generate more sequence than existed in the whole of NCBI at the beginning of 2008. Analysis of the sequence data has rapidly become the limiting step and will likely become the most expensive part. The sheer volume of data will provide challenges in processing, networking, storage, and analysis of the flow-cell images just to provide the initial base calling." after Holt & Jones, 2009

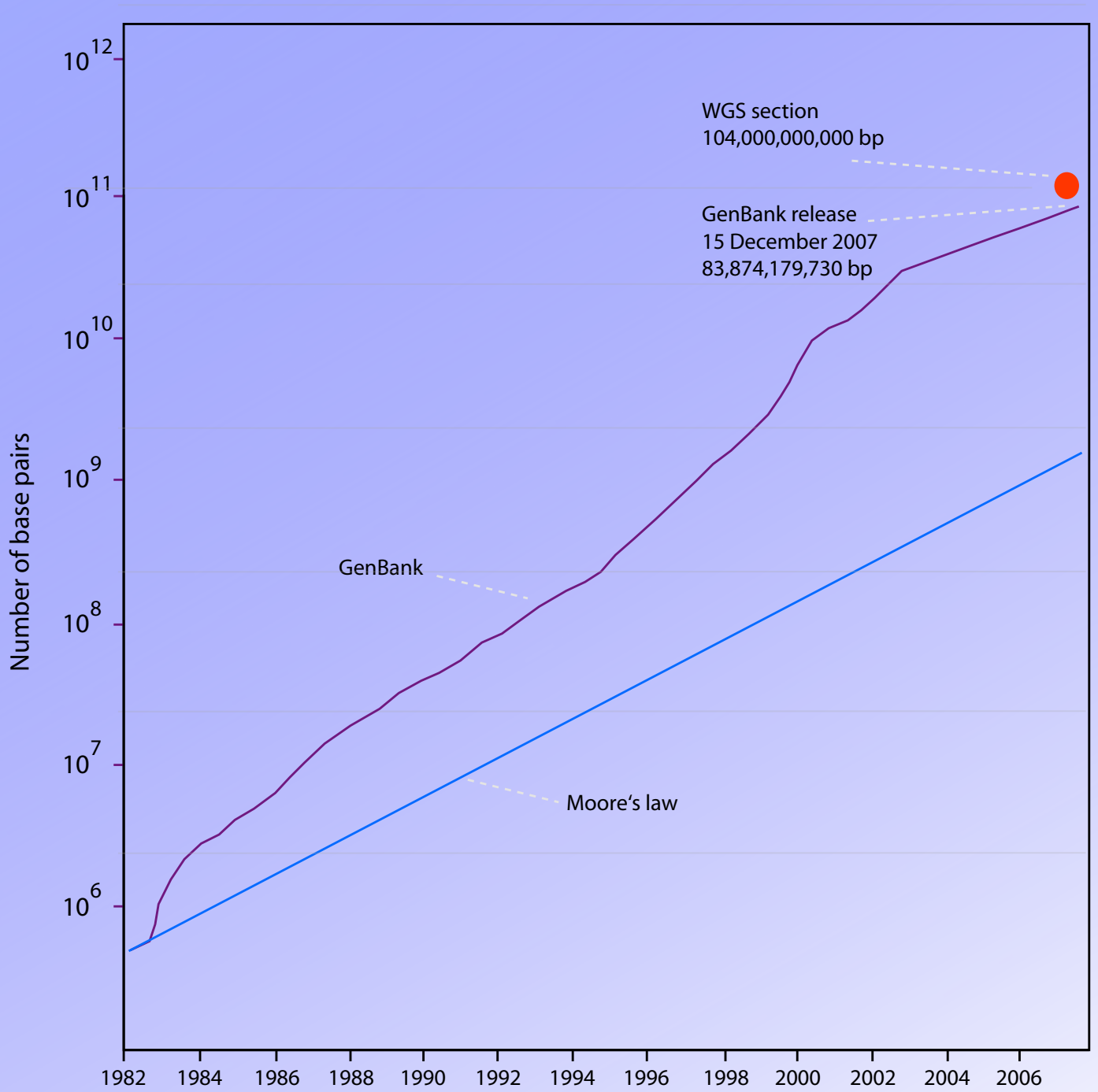
Sanger Center has 28 Solexa machines,
8 ABI Solids, 2 Roche 454 machines

>1000 teraBytes per month!

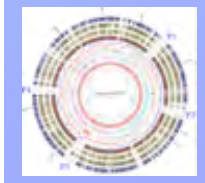
Center for Biological Sequence analysis
Department of Systems Biology, Technical University of Denmark



1. The problem - too much data!



CENTERFORBIOLOGICALSEQUENCEANALYSIS CBS
Center for Biological Sequence analysis
Department of Systems Biology, Technical University of Denmark
DTU



2. A brief history - The speed of sequencing

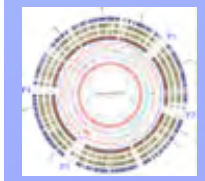
The Human Genome Project

Started more than 20 years ago (~1985)

The U.S. government agreed to invest $200,000,000 U.S. per year for 20 years.

~3,400,000,000 bp per haploid genome
~6,800,000,000 bp per diploid genome

One base per second = 216 years!



2. A brief history - The speed of sequencing

1. “First Human Genome”

$3,000,000,000 + 15 years

2. Celera genome (a.k.a. J. Craig Venter)

$100,000,000 + 0.75 years (9 months)

3. Jim Watson’s genome

$900,000 + 0.17 years (2 months)

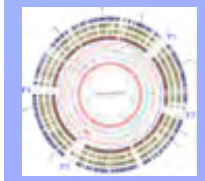
4. John Doe's genome

$1,000 + 0.0002 years (0.1 day)

5. "next next-generation" machines

- Helicos Biosystems machine can sequence human genome in 1 hour (2009).
- Pacific Biosciences machine can sequence human genome in 4 minutes (2010).
- Omni Molecular Recognizer Application - human genome less than $1, <1 minute.

CENTERFORBIOLOGICALSEQUENCEANALYSIS CBS
Center for Biological Sequence analysis
Department of Systems Biology, Technical University of Denmark
DTU



Comparative Microbial Genomics group

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

Center for Biological Sequence analysis

Department of Systems Biology, Technical University of Denmark

Bacteria Archaea total published Unfinished total

Number Genomes in NCBI web pages

4,000
3,500
3,000
2,500
2,000
1,500
1,000
500
0

1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010

Year

DTU

Microbial Genomes Properties

http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi

Apple (125) Useful PubMed Amazon Google Maps News (1638) Wikipedia Bioinformatics

A core gut microbiome i... Complete Microbial Gen...

NCBI ENTREZ Genome Project

connection information discovery

PubMed Nucleotide Protein Genome Structure PopSet Taxonomy OMIM

Search Genome Project Go Clear

Organism info | Complete genomes | Genomes in progress

organism group: -- All --

~~2226 as of 21 Jan, 2009~~

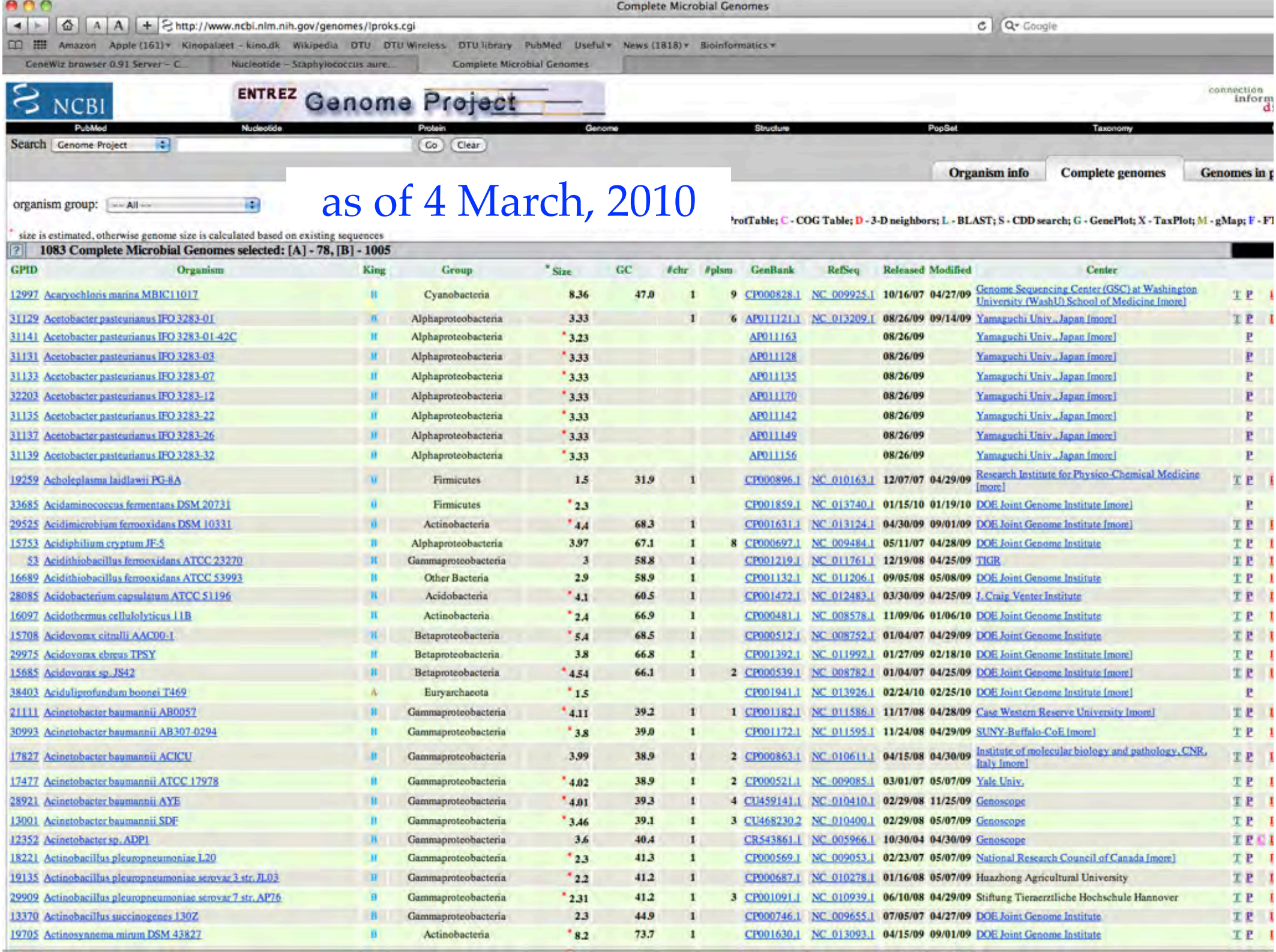
3630 as of 4 March, 2010

Legend: ☐ - complete; ☐ - assembly; ☐ - no sequence.

sequencing status filter: ◉ all; ○ complete; ○ assembly; ○ no sequence.

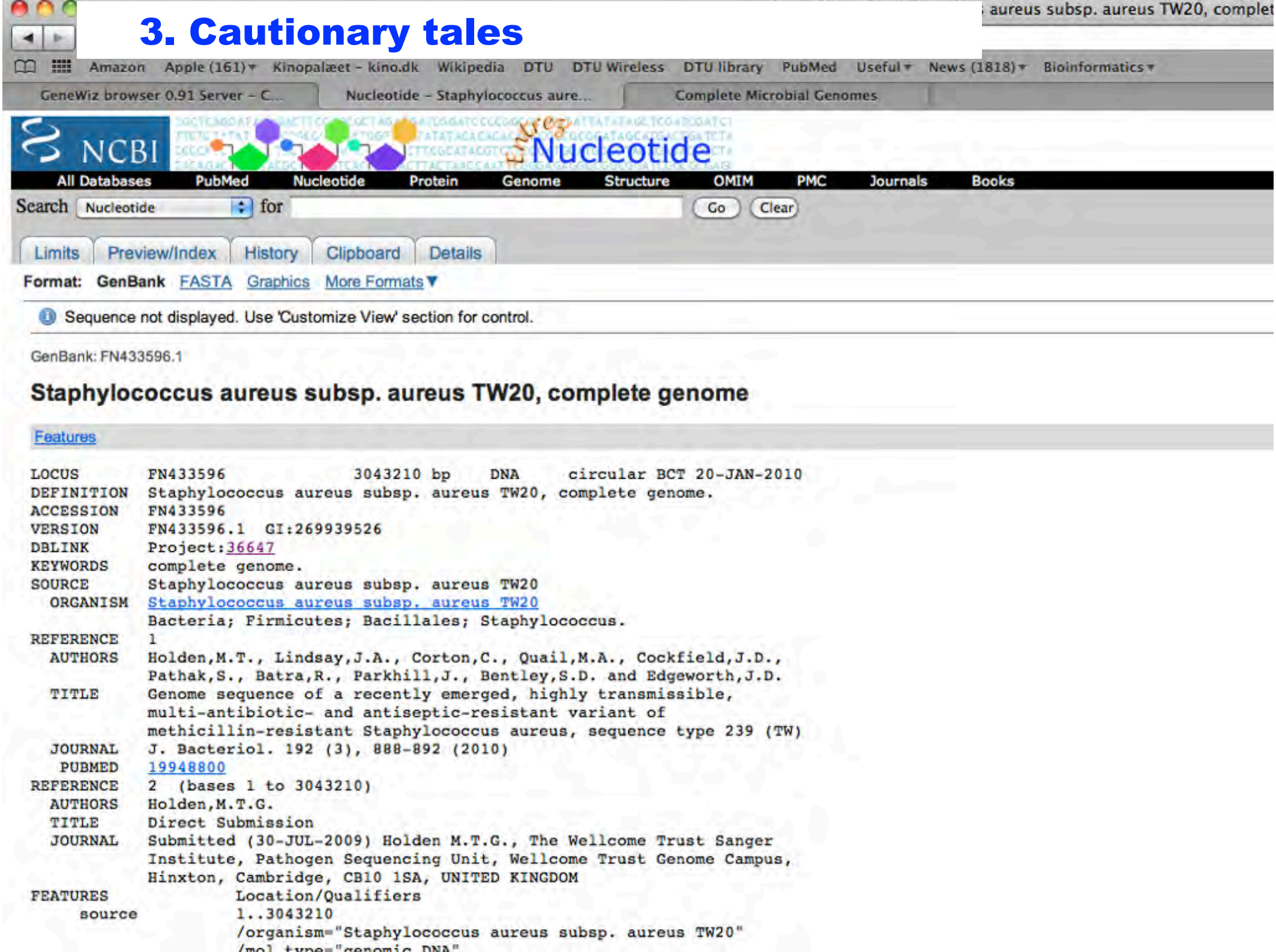
save

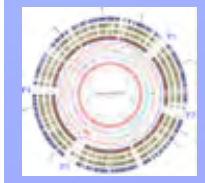
GPID	Organism	King	Size	GC	stain	Shape	Arrangement	Endospores	Motility	Salinity	Oxygen Req.	Habitat	Temp. range	Pathogenic in	Disease
30807	Nostoc azollae' 0708	B				Filaments	Filaments		Yes		Aerobic	Multiple	Mesophilic		
12997	Acaryochloris marina MBIC11017	B	8.36	47.0		Sphere, Ellipse	Singles				Aerobic	Aquatic	Mesophilic	No	None
16707	Acaryochloris sp. CCMEE 5410	B				Sphere, Ellipse	Singles				Aerobic	Aquatic	Mesophilic	No	
32769	Acetohalobium arabaticum DSM 5501	B								Moderate halophilic	Anaerobic	Specialized			
19259	Acholeplasma laidlawii PG-8A	B	1.5	31.9		Pleomorphic				Non-halophilic	Facultative	Specialized	Mesophilic		
32215	Acidaminococcus intestinalis RYC-MR95	B			-	Coccobacillus					Anaerobic	Host-associated	Mesophilic	Human	Opportunistic infections
199	Acidianus brierleyi	A	*1.8			Coccus	Singles	No	No		Facultative		Thermophilic	No	
29525	Acidimicrobium ferrooxidans DSM 10331	B		67-69	+	Rod		No			Anaerobic	Specialized	Thermophilic	No	
15753	Acidiphilium cryptum JF-5	B	3.97	67.1	-	Rod					Aerobic	Multiple	Mesophilic	No	
53	Acidithiobacillus ferrooxidans ATCC 23270	B	3	58.8	-	Rod			Yes		Facultative	Multiple	Mesophilic	No	
16689	Acidithiobacillus ferrooxidans ATCC 53993	B	2.9	58.9	-	Spiral		No	Yes		Aerobic	Specialized	Mesophilic	No	
15771	Acidobacteria bacterium Ellin345	B	5.7	58.4										No	
1	Acidobacterium capsulatum ATCC 51196	B	*4.15		-	Rod	Singles	No	No	Non-halophilic	Aerobic	Multiple	Mesophilic	No	None
28085	Acidobacterium capsulatum ATCC 51196	B			-	Rod	Singles	No	No		Aerobic	Multiple	Mesophilic	No	
16097	Acidothermus cellulolyticus 11B	B	2.4	66.9	+			Yes			Aerobic	Aquatic	Thermophilic	No	
15708	Acidovorax avenae subsp. citrulli AAC00-1	B	5.4	68.5	-	Rod		No	Yes	Non-halophilic	Aerobic	Multiple	Mesophilic	Fruit	Bacterial fruit blotch
32605	Acidovorax delafieldii 2AN	B		65-66	-	Rod		No	Yes		Aerobic	Aquatic	Mesophilic	No	
15685	Acidovorax sp. JS42	B	4.54	66.1	-	Rod		No	Yes		Aerobic	Terrestrial	Mesophilic	No	
19277	Aciduliprofundum boonei	A		39.1		Coccus			Yes	Mesophilic	Anaerobic	Specialized	Thermophilic	No	



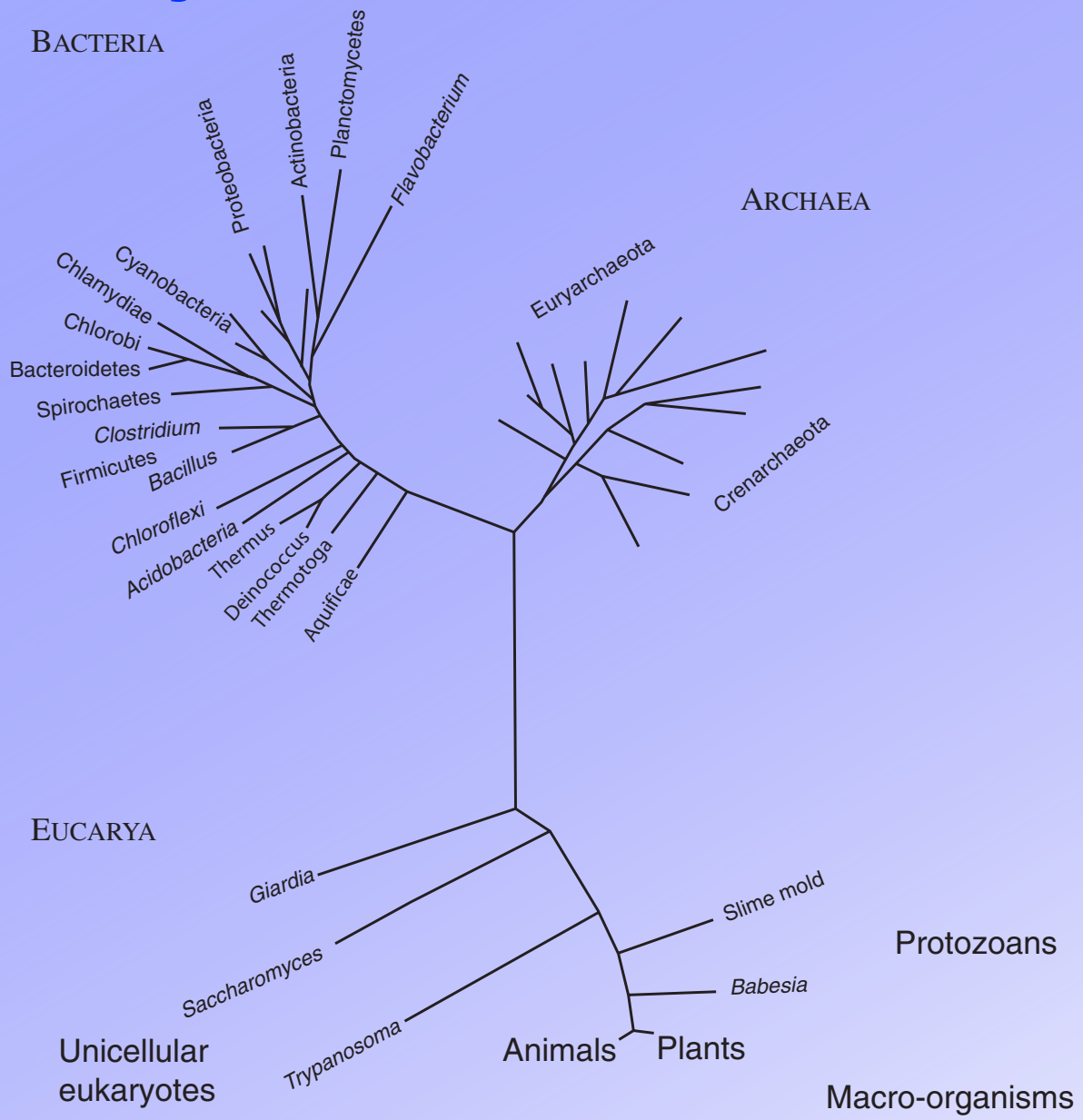
GPID	Organism	King	Group	Size	GC	#chr	#plsm	GenBank	RefSeq	Released	Modified	Center
12997	Acaryochloris marina MBIC11017	B	Cyanobacteria	8.36	47.0	1	9	CP000828.1	NC_009925.1	10/16/07	04/27/09	Genome Sequencing Center (GSC) at Washington University (WashU) School of Medicine [more]
31129	Acetobacter pasteurianus IFO 3283-01	B	Alphaproteobacteria	3.33		1	6	AP011121.1	NC_013209.1	08/26/09	09/14/09	Yamaguchi Univ., Japan [more]
31141	Acetobacter pasteurianus IFO 3283-01-42C	B	Alphaproteobacteria	*3.23				AP011163		08/26/09		Yamaguchi Univ., Japan [more]
31131	Acetobacter pasteurianus IFO 3283-03	B	Alphaproteobacteria	*3.33				AP011128		08/26/09		Yamaguchi Univ., Japan [more]
31133	Acetobacter pasteurianus IFO 3283-07	B	Alphaproteobacteria	*3.33				AP011135		08/26/09		Yamaguchi Univ., Japan [more]
32203	Acetobacter pasteurianus IFO 3283-12	B	Alphaproteobacteria	*3.33				AP011170		08/26/09		Yamaguchi Univ., Japan [more]
31135	Acetobacter pasteurianus IFO 3283-22	B	Alphaproteobacteria	*3.33				AP011142		08/26/09		Yamaguchi Univ., Japan [more]
31137	Acetobacter pasteurianus IFO 3283-26	B	Alphaproteobacteria	*3.33				AP011149		08/26/09		Yamaguchi Univ., Japan [more]
31139	Acetobacter pasteurianus IFO 3283-32	B	Alphaproteobacteria	*3.33				AP011156		08/26/09		Yamaguchi Univ., Japan [more]
19259	Acholeplasma laidlawii PG-8A	B	Firmicutes	1.5	31.9	1		CP000896.1	NC_010163.1	12/07/07	04/29/09	Research Institute for Physico-Chemical Medicine [more]
33685	Acidaminococcus fermentans DSM 20731	B	Firmicutes	*2.3				CP001859.1	NC_013740.1	01/15/10	01/19/10	DOE Joint Genome Institute [more]
29525	Acidimicrobium ferrooxidans DSM 10331	B	Actinobacteria	*4.4	68.3	1		CP001631.1	NC_013124.1	04/30/09	09/01/09	DOE Joint Genome Institute [more]
15753	Acidiphilium cryptum JF-5	B	Alphaproteobacteria	3.97	67.1	1	8	CP000697.1	NC_009484.1	05/11/07	04/28/09	DOE Joint Genome Institute
53	Acidithiobacillus ferrooxidans ATCC 23270	B	Gammaproteobacteria	3	58.8	1		CP001219.1	NC_011761.1	12/19/08	04/25/09	TIGR
16689	Acidithiobacillus ferrooxidans ATCC 53993	B	Other Bacteria	2.9	58.9	1		CP001132.1	NC_011206.1	09/05/08	05/08/09	DOE Joint Genome Institute
28085	Acidobacterium capsulatum ATCC 51196	B	Acidobacteria	*4.1	60.5	1		CP001472.1	NC_012483.1	03/30/09	04/25/09	J. Craig Venter Institute
16097	Acidothermus cellulolyticus 11B	B	Actinobacteria	*2.4	66.9	1		CP000481.1	NC_008578.1	11/09/06	01/06/10	DOE Joint Genome Institute
15708	Acidovorax citrulli AAC00-1	B	Betaproteobacteria	*5.4	68.5	1		CP000512.1	NC_008752.1	01/04/07	04/29/09	DOE Joint Genome Institute
29975	Acidovorax ebreus TPSY	B	Betaproteobacteria	3.8	66.8	1		CP001392.1	NC_011992.1	01/27/09	02/18/10	DOE Joint Genome Institute [more]
15685	Acidovorax sp. JS42	B	Betaproteobacteria	*4.54	66.1	1	2	CP000539.1	NC_008782.1	01/04/07	04/25/09	DOE Joint Genome Institute [more]
38403	Aciduliprofundum boonei T469	A	Euryarchaeota	*1.5				CP001941.1	NC_013926.1	02/24/10	02/25/10	DOE Joint Genome Institute [more]
21111	Acinetobacter baumannii AB0057	B	Gammaproteobacteria	*4.11	39.2	1	1	CP001182.1	NC_011586.1	11/17/08	04/28/09	Case Western Reserve University [more]
30993	Acinetobacter baumannii AB307-0294	B	Gammaproteobacteria	*3.8	39.0	1		CP001172.1	NC_011595.1	11/24/08	04/29/09	SUNY-Buffalo-CoE [more]
17827	Acinetobacter baumannii ACICU	B	Gammaproteobacteria	3.99	38.9	1	2	CP000863.1	NC_010611.1	04/15/08	04/30/09	Institute of molecular biology and pathology, CNR, Italy [more]
17477	Acinetobacter baumannii ATCC 17978	B	Gammaproteobacteria	*4.02	38.9	1	2	CP000521.1	NC_009085.1	03/01/07	05/07/09	Yale Univ.
28921	Acinetobacter baumannii AYE	B	Gammaproteobacteria	*4.01	39.3	1	4	CU459141.1	NC_010410.1	02/29/08	11/25/09	Genoscope
13001	Acinetobacter baumannii SDF	B	Gammaproteobacteria	*3.46	39.1	1	3	CU468230.2	NC_010400.1	02/29/08	05/07/09	Genoscope
12352	Acinetobacter sp. ADP1	B	Gammaproteobacteria	3.6	40.4	1		CR543861.1	NC_005966.1	10/30/04	04/30/09	Genoscope
18221	Actinobacillus pleuropneumoniae L20	B	Gammaproteobacteria	*2.3	41.3	1		CP000569.1	NC_009053.1	02/23/07	05/07/09	National Research Council of Canada [more]
19135	Actinobacillus pleuropneumoniae serovar 3 str. JL03	B	Gammaproteobacteria	*2.2	41.2	1		CP000687.1	NC_010278.1	01/16/08	05/07/09	Huazhong Agricultural University
29909	Actinobacillus pleuropneumoniae serovar 7 str. AP76	B	Gammaproteobacteria	*2.31	41.2	1	3	CP001091.1	NC_010939.1	06/10/08	04/29/09	Stiftung Tieraerztliche Hochschule Hannover
13370	Actinobacillus succinogenes 130Z	B	Gammaproteobacteria	2.3	44.9	1		CP000746.1	NC_009655.1	07/05/07	04/27/09	DOE Joint Genome Institute
19705	Actinosynnema mirum DSM 43827	B	Actinobacteria	*8.2	73.7	1		CP001630.1	NC_013093.1	04/15/09	09/01/09	DOE Joint Genome Institute

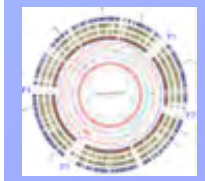
3. Cautionary tales

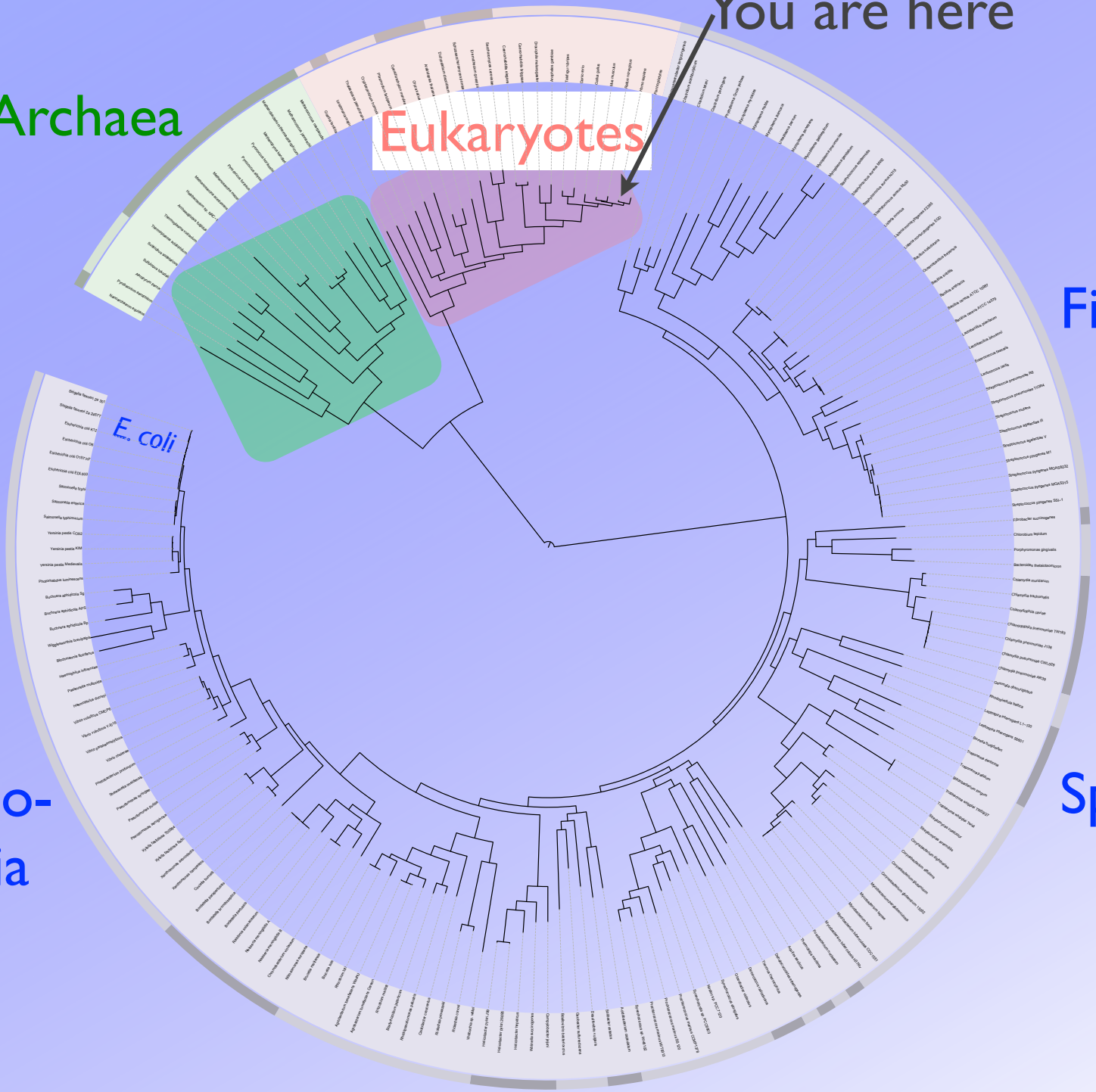




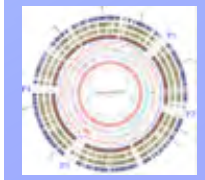
3. Cautionary tales



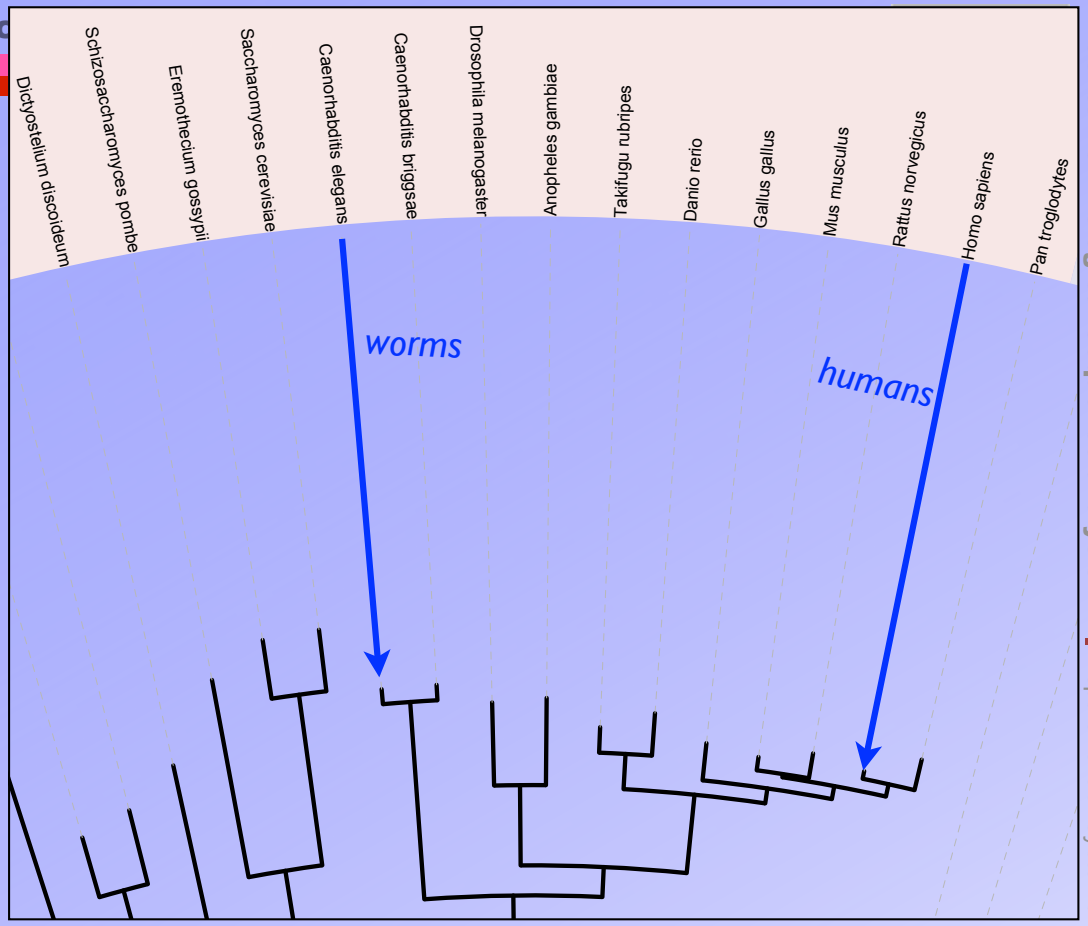


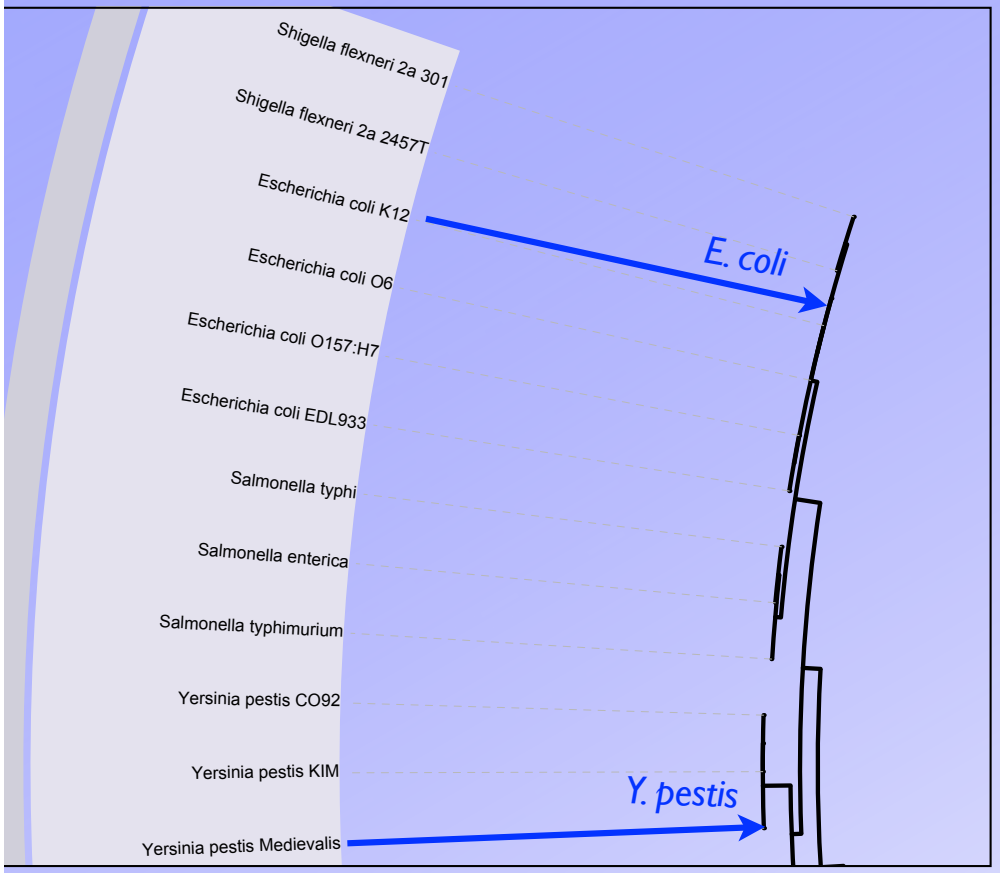


CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS
Center for Biological Sequence analysis
Department of Systems Biology, Technical University of Denmark
DTU

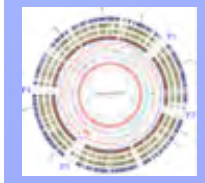


3. Cautionary tales





Center for Biological Sequence analysis
Department of Systems Biology, Technical University of Denmark



3. Cautionary tales

human	chimp	chicken	worm	yeast
	99%	76%	55%	28%
23,621 genes	19,829 genes 19,568 orthologs	18,529 genes 14,000 orthologs	18,000 genes 10,000 orthologs	6,000 genes 1,700 orthologs

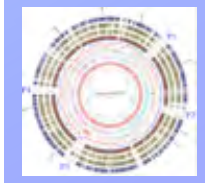
C. botulanium type A, strain ATCC 3502	*C. botulanium* type A, strain ATCC 19397	*C. botulanium* type A, strain Kyoto	*C. botulanium* type C, strain Ecklund	*C. botulanium* type E1, strain BoNT E Beluga
	90%	77%	40%	29%
3606 genes	3553 genes 3202 orthologs	3874 genes 2974 orthologs	2801 genes 1126 orthologs	3760 genes 1092 orthologs

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

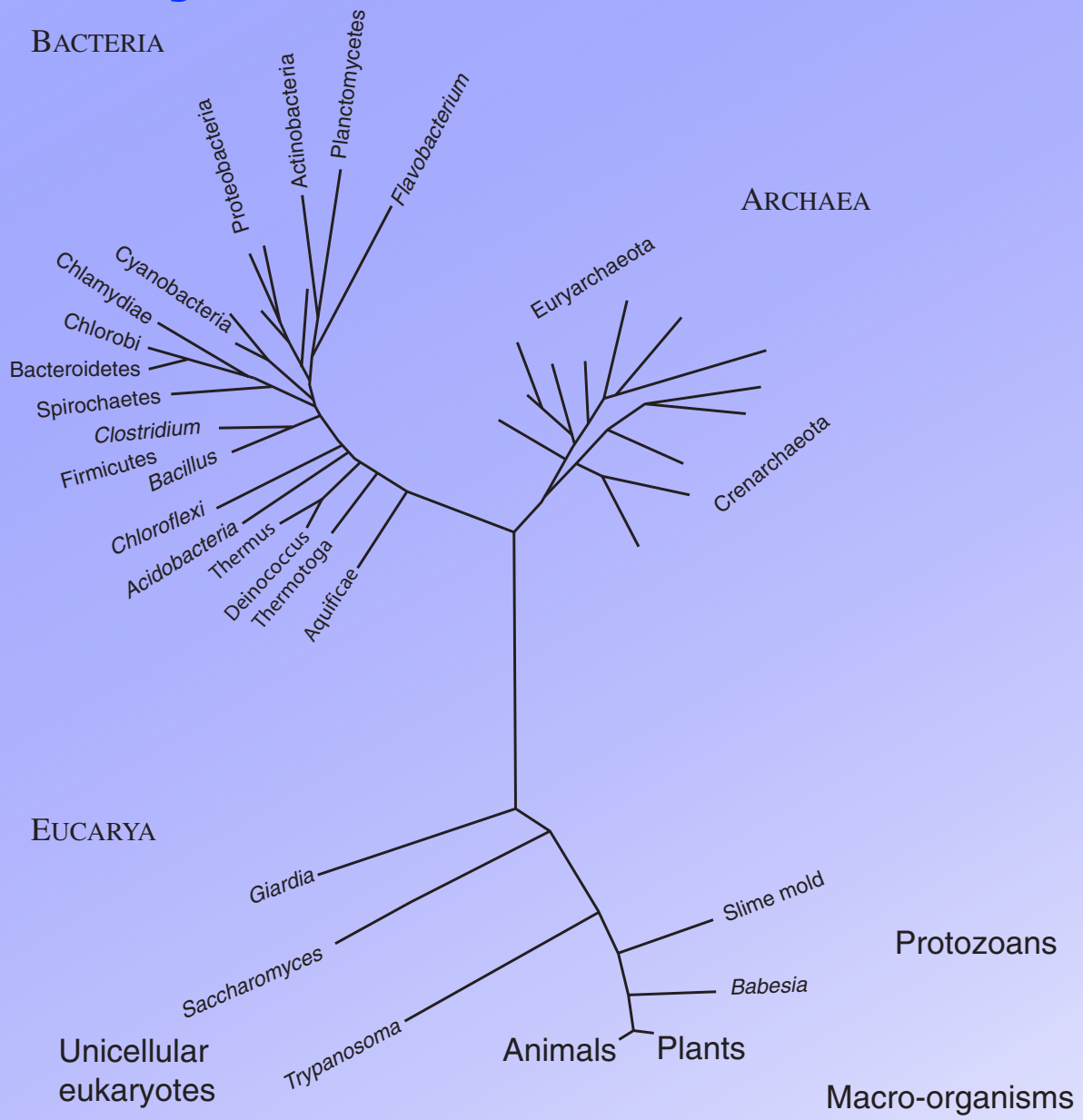
Center for Biological Sequence analysis

Department of Systems Biology, Technical University of Denmark

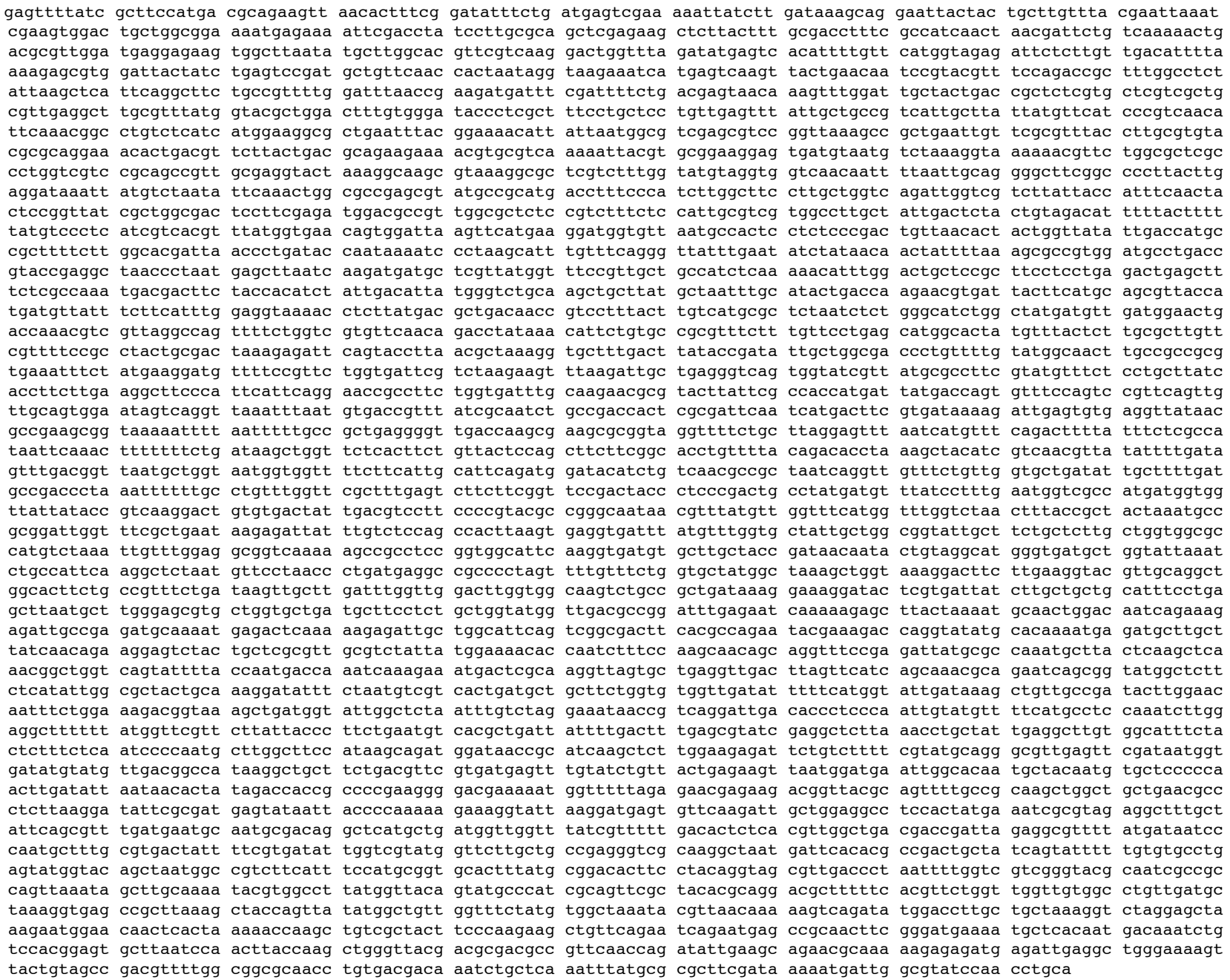
DTU

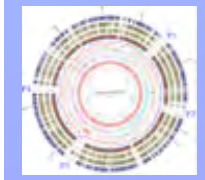


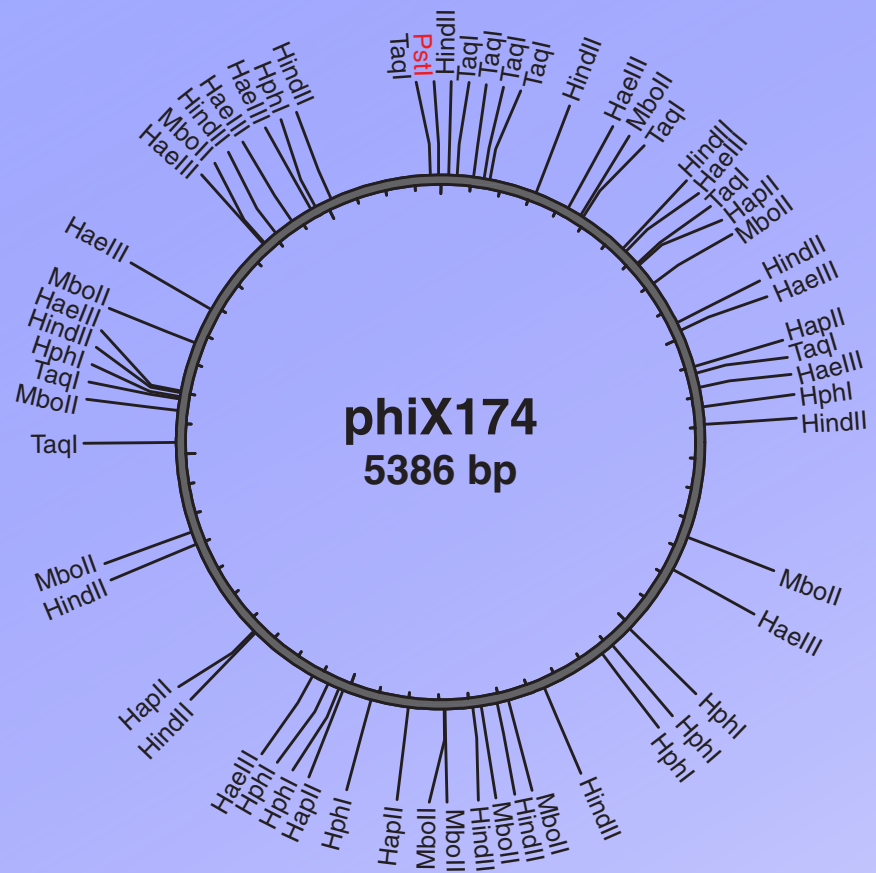
3. Cautionary tales

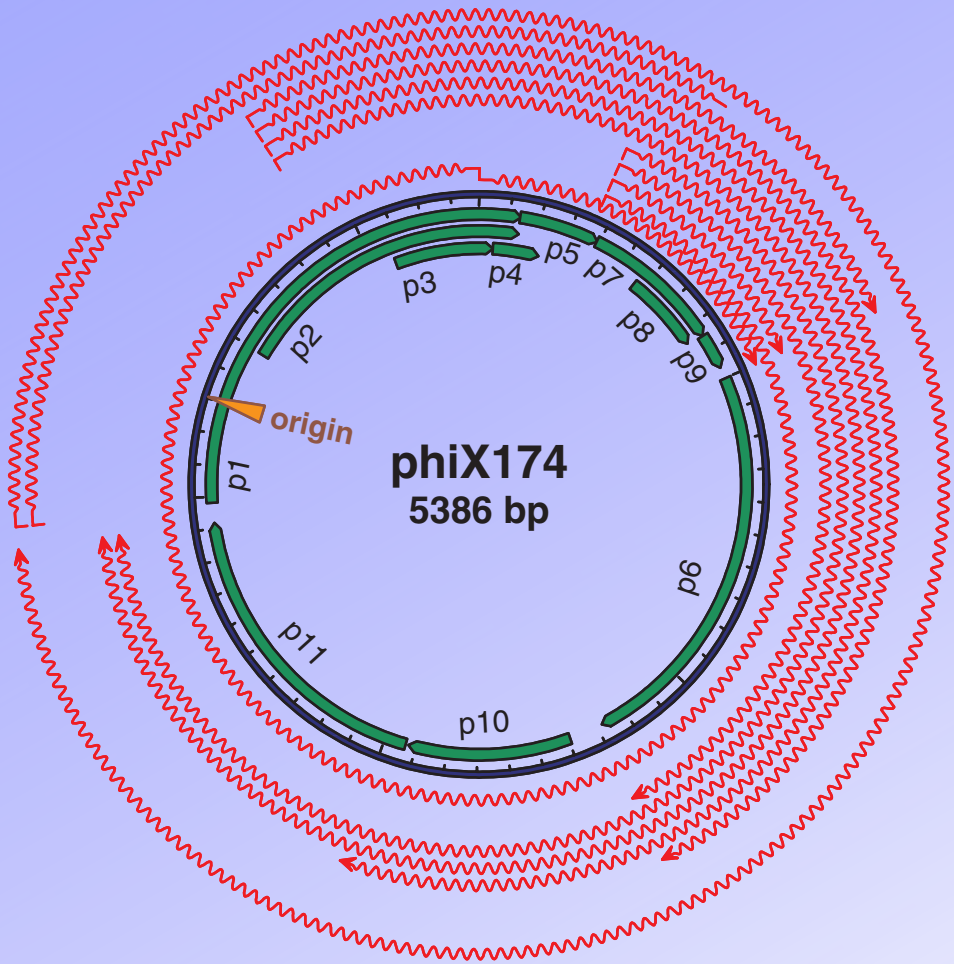


4. Approaches to handle lots of data - Visualization

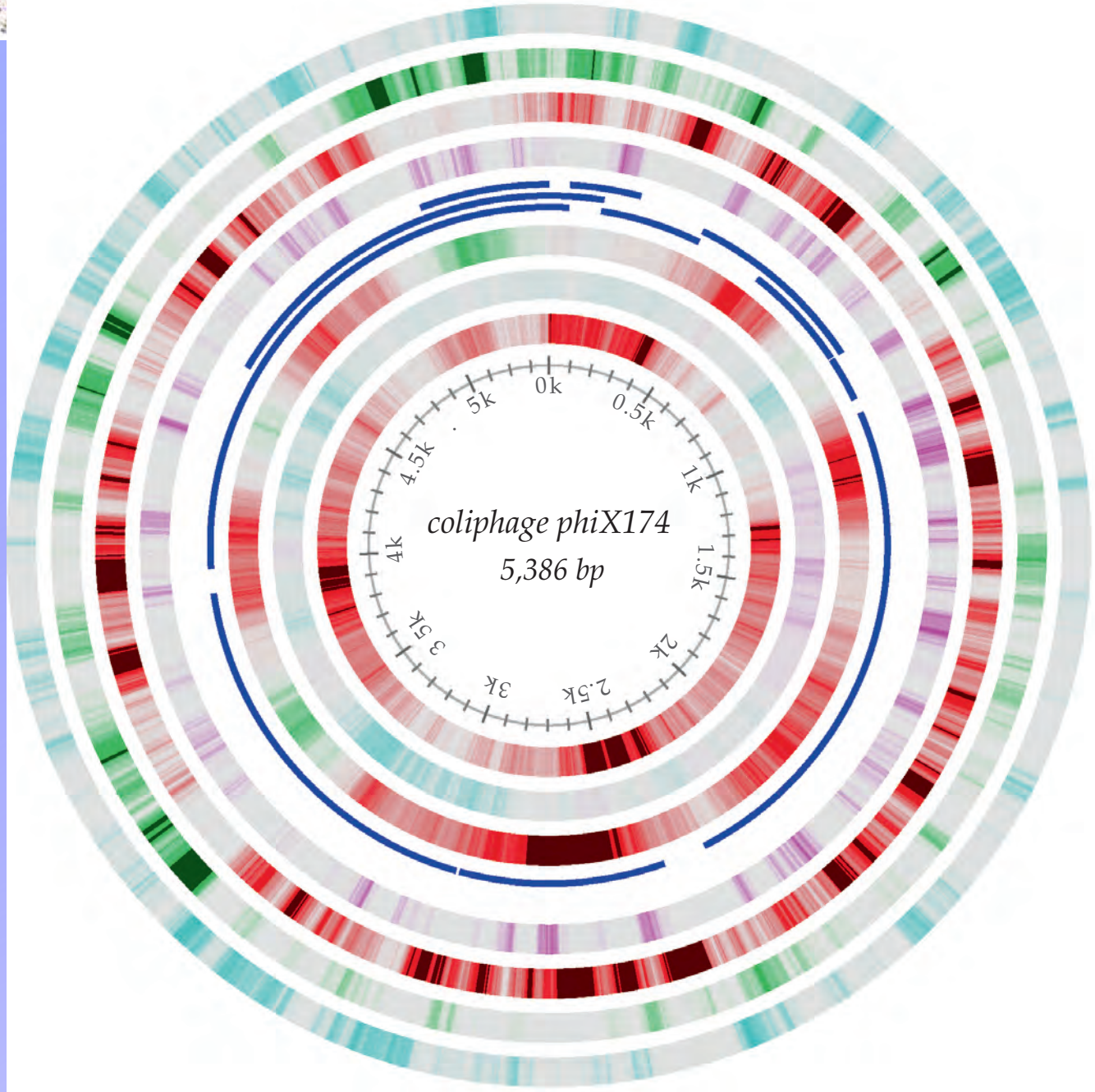








Center for Biological Sequence analysis
Department of Systems Biology, Technical University of Denmark



G Content
0.00 fix avg 0.40

A Content
0.00 fix avg 0.40

T Content
0.00 fix avg 0.40

C Content
0.00 fix avg 0.40

Annotations:
CDS +

AT Skew
–0.20 fix avg 0.20

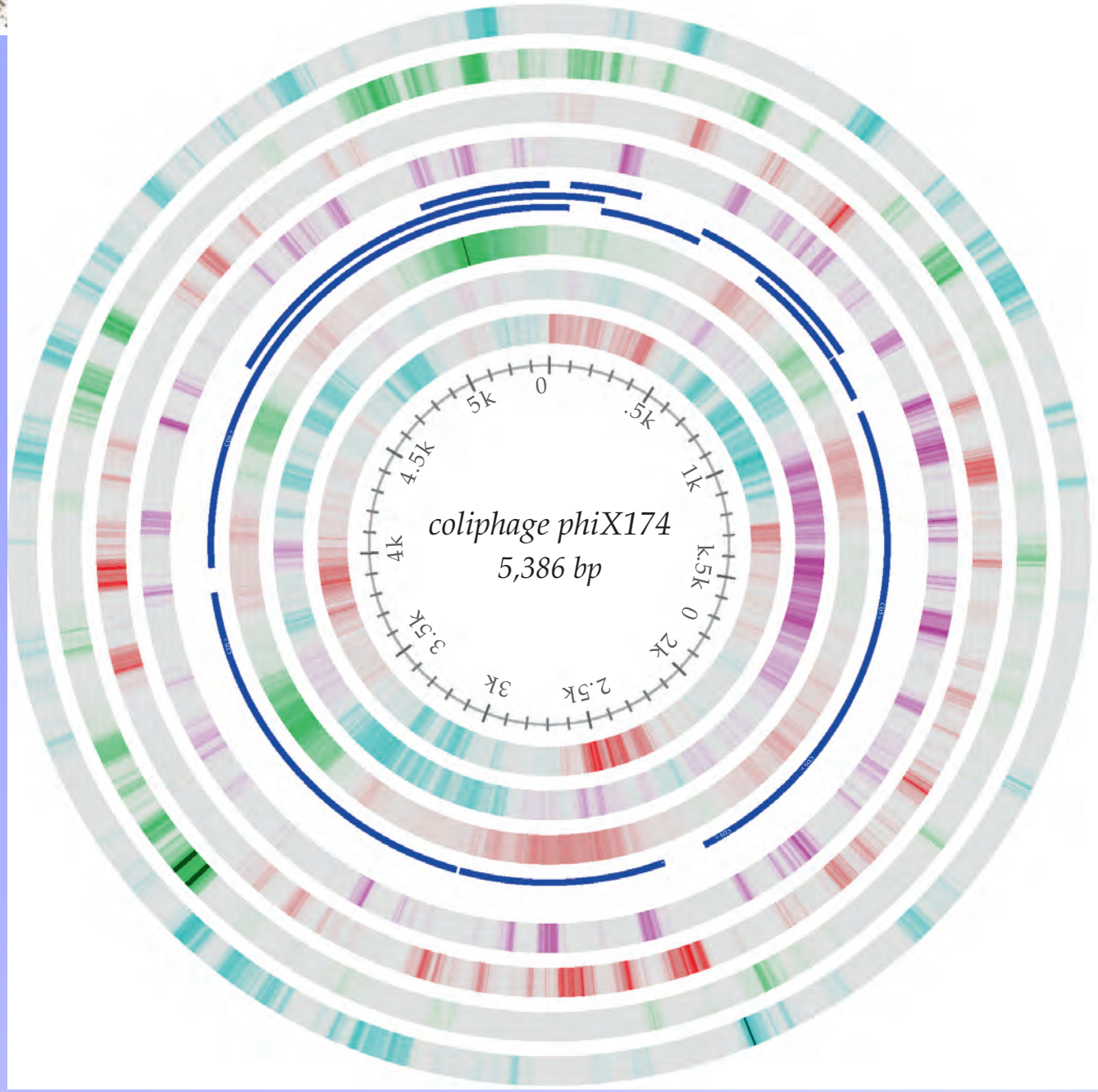
GC Skew
–0.20 fix avg 0.20

Percent AT
0.40 fix avg 0.60

Resolution: 3

BASE ATLAS

Center for Biological Sequence analysis
Department of Systems Biology, Technical University of Denmark



G Content
dev avg
0.07 0.39

A Content
dev avg
0.01 0.47

T Content
dev avg
0.10 0.53

C Content
dev avg
0.04 0.39

Annotations:
CDS +

AT Skew
dev avg
–0.33 0.18

GC Skew
dev avg
–0.10 0.14

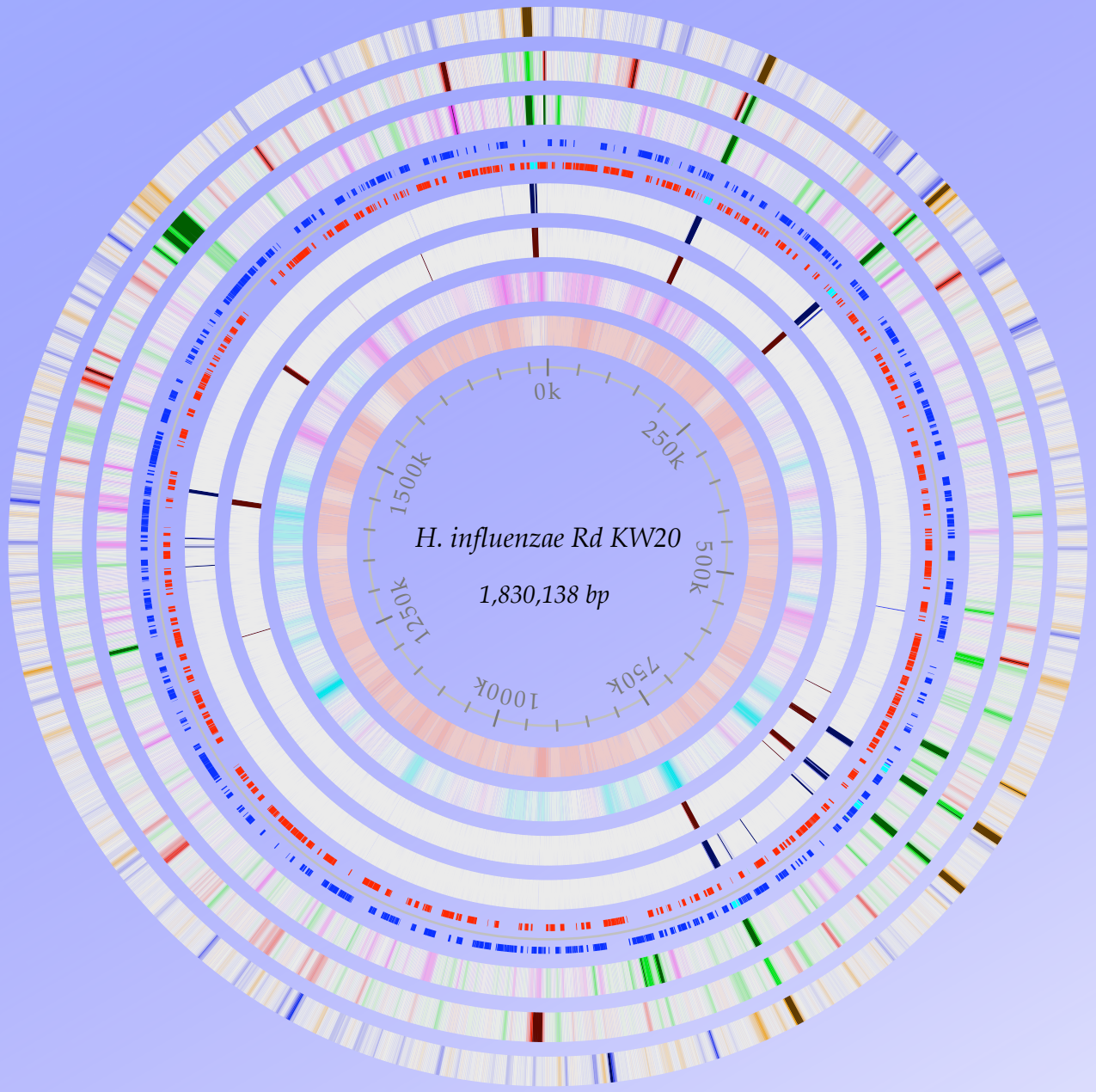
Percent AT
dev avg
0.47 0.63

Resolution: 3

BASE ATLAS

Center for Biological Sequence analysis
Department of Systems Biology, Technical University of Denmark

gagttttatc gcttccatga cgcagaagtt aacactttcg gatatttctg atgagtcgaa aaattatctt gataaagcag gaattactac tgcttgttta cgaattaaat
cgaagtggac tgctggcgga aaatgagaaa attcgaccta tccttgcgca gctcgagaag ctcttacttt gcgacctttc gccatcaact aacgattctg tcaaaaactg
acgcgttgga tgaggagaag tggcttaata tgcttggcac gttcgtcaag gactggttta gatatgagtc acattttgtt catggtagag attctcttgt tgacatttta
aaagagcgtg gattactatc tgagtccgat gctgttcaac cactaatagg taagaaatca tgagtcaagt tactgaacaa tccgtacgtt tccagaccgc tttggcctct
attaagctca ttcaggcttc tgccgttttg gatttaaccg aagatgattt cgattttctg acgagtaaca aagtttggat tgctactgac cgctctcgtg ctcgtcgctg
cgttgaggct tgcgtttatg gtacgctgga ctttgtggga taccctcgct ttcctgctcc tgttgagttt attgctgccg tcattgctta ttatgttcat cccgtcaaca
ttcaaacggc ctgtctcatc atggaaggcg ctgaatttac ggaaaacatt attaatggcg tcgagcgtcc ggttaaagcc gctgaattgt tcgcgtttac cttgcgtgta
cgcgcaggaa acactgacgt tcttactgac gcagaagaaa acgtgcgtca aaaattacgt gcggaaggag tgatgtaatg tctaaaggta aaaaacgttc tggcgctcgc
cctggtcgtc cgcagccgtt gcgaggtact aaaggcaagc gtaaaggcgc tcgtctttgg tatgtaggtg gtcaacaatt ttaattgcag gggcttcggc cccttacttg
aggataaatt atgtctaata ttcaaactgg cgccgagcgt atgccgcatg acctttccca tcttggcttc cttgctggtc agattggtcg tcttattacc atttcaacta
ctccggttat cgctggcgac tccttcgaga tggacgccgt tggcgctctc cgtctttctc cattgcgtcg tggccttgct attgactcta ctgtagacat ttttacttttt
tatgtccctc atcgtcacgt ttatggtgaa cagtggatta agttcatgaa ggatggtgtt aatgccactc ctctcccgac tgttaacact actggttata ttgaccatgc
cgcttttctt ggcacgatta accctgatac caataaaatc cctaagcatt tgtttcaggg ttatttgaat atctataaca actattttaa agcgccgtgg atgcctgacc
gtaccgaggc taaccctaat gagcttaatc aagatgatgc tcgttatggt ttccgttgct gccatctcaa aaacatttgg actgctccgc ttcctcctga gactgagctt
tctcgccaaa tgacgacttc taccacatct attgacatta tgggtctgca agctgcttat gctaatttgc atactgacca agaacgtgat tacttcatgc agcgttacca
tgatgttatt tcttcatttg gaggtaaaac ctcttatgac gctgacaacc gtcctttact tgtcatgcgc tctaatctct gggcatctgg ctatgatgtt gatggaactg
accaaacgtc gttaggccag ttttctggtc gtgttcaaca gacctataaa cattctgtgc cgcgtttctt tgttcctgag catggcacta tgtttactct tgcgcttgtt
cgttttccgc ctactgcgac taaagagatt cagtacctta acgctaaagg tgctttgact tataccgata ttgctggcga ccctgttttg tatggcaact tgccgccgcg
tgaaatttct atgaaggatg ttttccgttc tggtgattcg tctaagaagt ttaagattgc tgagggtcag tggtatcgtt atgcgccttc gtatgtttct cctgcttatc
accttcttga aggcttccca ttcattcagg aaccgccttc tggtgatttg caagaacgcg tacttattcg ccaccatgat tatgaccagt gtttccagtc cgttcagttg
ttgcagtgga atagtcaggt taaatttaat gtgaccgttt atcgcaatct gccgaccact cgcgattcaa tcatgacttc gtgataaaag attgagtgtg aggttataac
gccgaagcgg taaaaatttt aatttttgcc gctgaggggt tgaccaagcg aagcgcggta ggttttctgc ttaggagttt aatcatgttt cagactttta tttctcgcca
taattcaaac tttttttctg ataagctggt tctcacttct gttactccag cttcttcggc acctgtttta cagacaccta aagctacatc gtcaacgtta tattttgata
gtttgacggt taatgctggt aatggtggtt ttcttcattg cattcagatg gatacatctg tcaacgccgc taatcaggtt gtttctgttg gtgctgatat tgcttttgat
gccgacccta aatttttgc ctgtttggtt cgctttgagt cttcttcggt tccgactacc ctcccgactg cctatgatgt ttatcctttg aatggtcgcc atgatggtgg
ttattatacc gtcaaggact gtgtgactat tgacgtcctt ccccgtacgc cgggcaataa cgtttatgtt ggtttcatgg tttggtctaa ctttaccgct actaaatgcc
gcggattggt ttcgctgaat aagagattat ttgtctccag ccacttaagt gaggtgattt atgtttggtg ctattgctgg cggtattgct tctgctcttg ctggtggcgc
catgtctaaa ttgtttggag gcggtcaaaa agccgcctcc ggtggcattc aaggtgatgt gcttgctacc gataacaata ctgtaggcat gggtgatgct ggtattaaat
ctgccattca aggctctaat gttcctaacc ctgatgaggc cgcccctagt tttgtttctg gtgctatggc taaagctggt aaaggacttc ttgaaggtac gttgcaggct
ggcacttctg ccgtttctga taagttgctt gatttggttg gacttggtgg caagtctgcc gctgataaag gaaaggatac tcgtgattat cttgctgctg catttcctga
gcttaatgct tgggagcgtg ctggtgctga tgcttcctct gctggtatgg ttgacgccgg atttgagaat caaaaagagc ttactaaaat gcaactggac aatcagaaag
agattgccga gatgcaaaat gagactcaaa aagagattgc tggcattcag tcggcgactt cacgccagaa tacgaaagac caggtatatg cacaaaatga gatgcttgct
tatcaacaga aggagtctac tgctcgcgtt gcgtctatta tggaaaacac caatctttcc aagcaacagc aggtttccga gattatgcgc caaatgctta ctcaagctca
aacggctggt cagtatttta ccaatgacca aatcaaagaa atgactcgca aggttagtgc tgaggttgac ttagttcatc agcaaacgca gaatcagcgg tatggctctt
ctcatattgg cgctactgca aaggatattt ctaatgtcgt cactgatgct gcttctggtg tggttgatat ttttcatggt attgataaag ctgttgccga tacttggaac
aatttctgga aagacggtaa agctgatggt attggctcta atttgtctag gaaataaccg tcaggattga caccctccca attgtatgtt ttcatgcctc caaatcttgg
aggctttttt atggttcgtt cttattaccc ttctgaatgt cacgctgatt attttgactt tgagcgtatc gaggctctta aacctgctat tgaggcttgt ggcatttcta
ctctttctca atccccaatg cttggcttcc ataagcagat ggataaccgc atcaagctct tggaagagat tctgtctttt cgtatgcagg gcgttgagtt cgataatggt
gatatgtatg ttgacggcca taaggctgct tctgacgttc gtgatgagtt tgtatctgtt actgagaagt taatggatga attggcacaa tgctacaatg tgctccccca
acttgatatt aataacacta tagaccaccg ccccgaaggg gacgaaaaat ggtttttaga gaacgagaag acggttacgc agttttgccg caagctggct gctgaacgcc
ctcttaagga tattcgcgat gagtataatt accccaaaaa gaaaggtatt aaggatgagt gttcaagatt gctggaggcc tccactatga aatcgcgtag aggctttgct
attcagcgtt tgatgaatgc aatgcgacag gctcatgctg atggttggtt tatcgttttt gacactctca cgttggctga cgaccgatta gaggcgtttt atgataatcc
caatgctttg cgtgactatt ttcgtgatat tggtcgtatg gttcttgctg ccgagggtcg caaggctaat gattcacacg ccgactgcta tcagtatttt tgtgtgcctg
agtatggtac agctaatggc cgtcttcatt tccatgcggt gcactttatg cggacacttc ctacaggtag cgttgaccct aattttggtc gtcgggtacg caatcgccgc
cagttaaata gcttgcaaaa tacgtggcct tatggttaca gtatgcccat cgcagttcgc tacacgcagg acgctttttc acgttctggt tggttgtggc ctgttgatgc
taaaggtgag ccgcttaaag ctaccagtta tatggctgtt ggtttctatg tggctaaata cgttaacaaa aagtcagata tggaccttgc tgctaaaggt ctaggagcta
aagaatggaa caactcacta aaaaccaagc tgtcgctact tcccaagaag ctgttcagaa tcagaatgag ccgcaacttc gggatgaaaa tgctcacaat gacaaatctg
tccacggagt gcttaatcca acttaccaag ctgggttacg acgcgacgcc gttcaaccag atattgaagc agaacgcaaa aagagagatg agattgaggc tgggaaaagt
tactgtagcc gacgttttgg cggcgcaacc tgtgacgaca aatctgctca aatttatgcg cgcttcgata aaaatgattg gcgtatccaa cctgca



Intrinsic Curvature
dev avg
0.19 0.26

Stacking Energy
dev avg
-7.91 -7.09

Position Preference
dev avg
0.14 0.17

Annotations:
CDS +
CDS -
rRNA
tRNA

Global Direct Repeats
fix avg
5.00 7.50

Global Inverted Repeats
fix avg
5.00 7.50

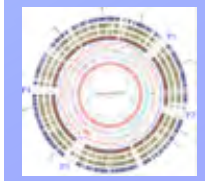
GC Skew
dev avg
-0.07 0.07

Percent AT
fix avg
0.20 0.80

Resolution: 733

GENOME ATLAS

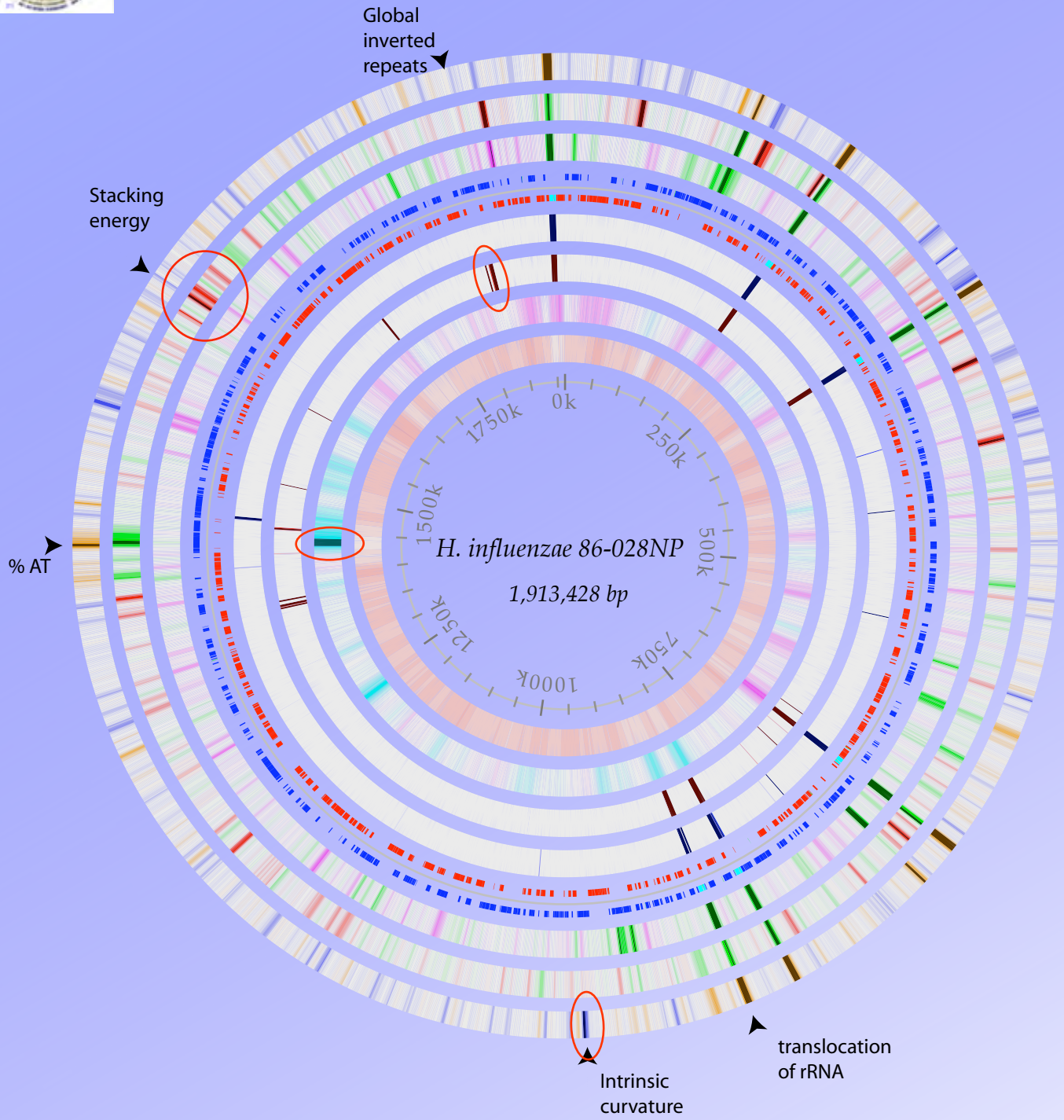
Center for Biological Sequence analysis
Department of Systems Biology, Technical University of Denmark



CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

Center for Biological Sequence analysis

Department of Systems Biology, Technical University of Denmark



Intrinsic Curvature
dev avg
0.19 0.26

Stacking Energy
dev avg
-7.89 -7.11

Position Preference
dev avg
0.14 0.17

Annotations:
CDS +
CDS -
rRNA
tRNA

Global Direct Repeats
fix avg
5.00 7.50

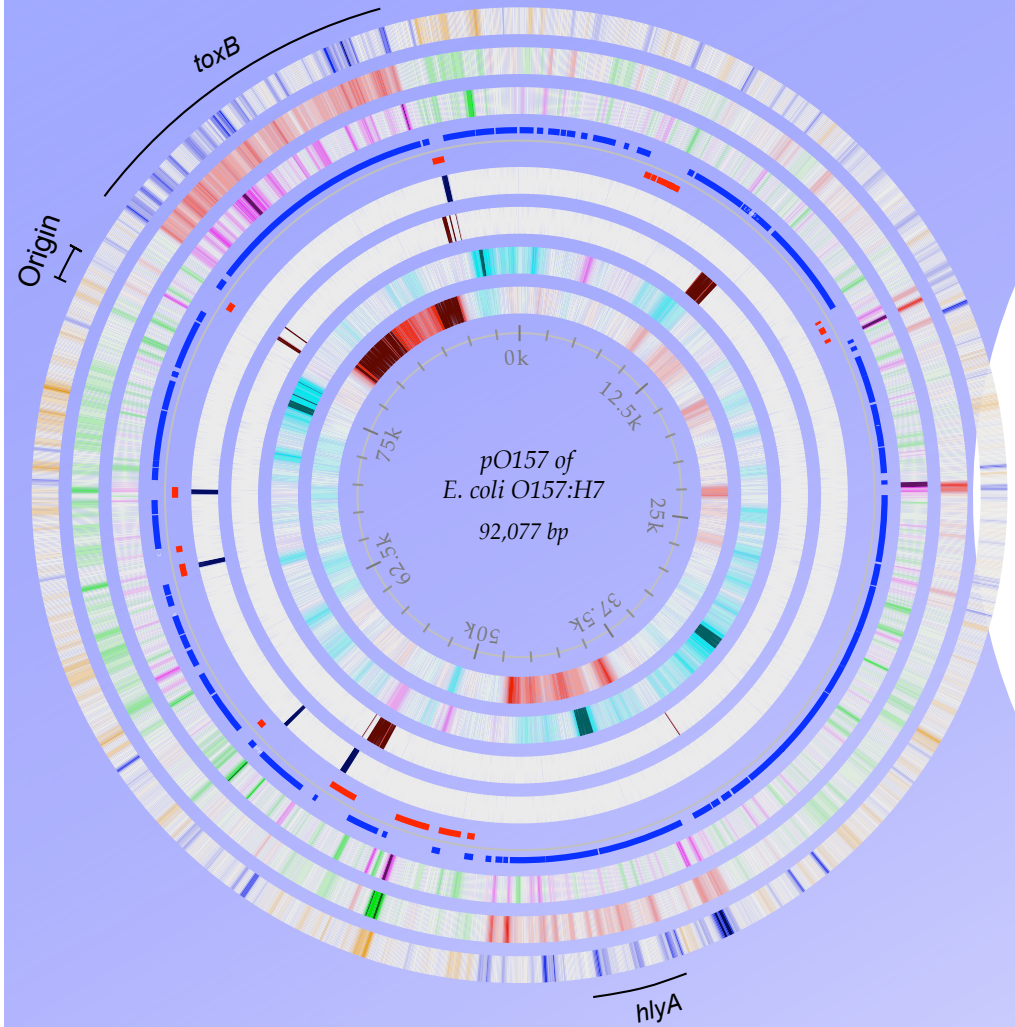
Global Inverted Repeats
fix avg
5.00 7.50

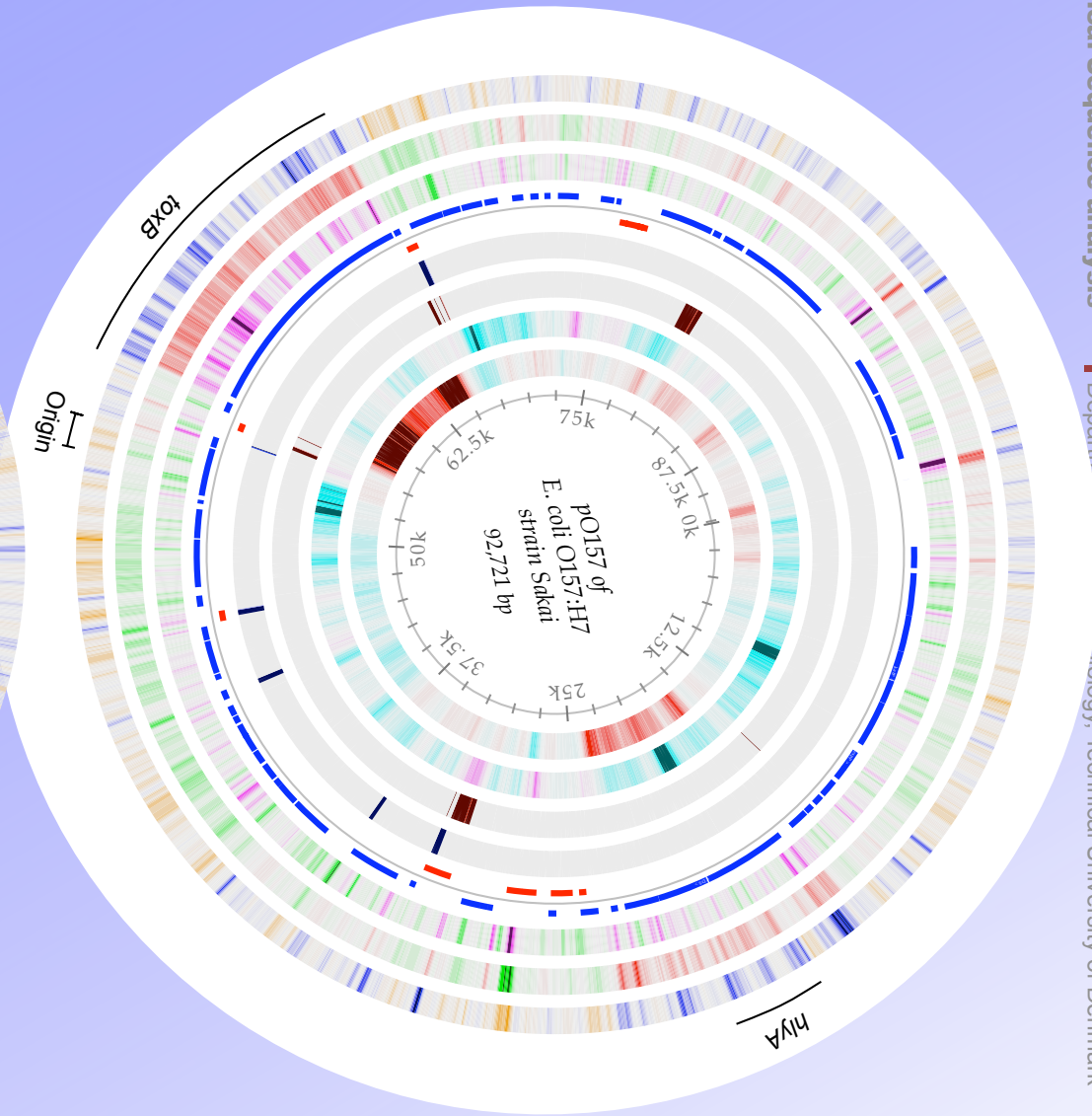
GC Skew
dev avg
-0.07 0.07

Percent AT
fix avg
0.20 0.80

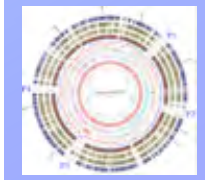
Resolution: 766

GENOME ATLAS

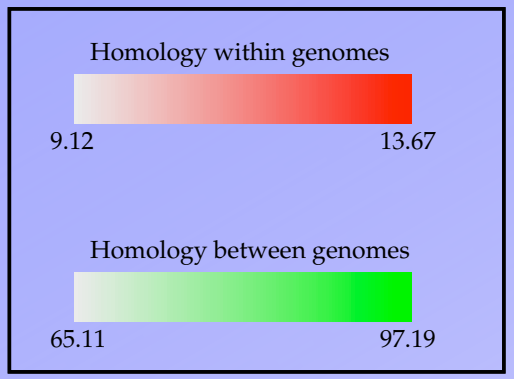




Center for Biological Sequence analysis
Department of Systems Biology, Technical University of Denmark



CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS



Escherichia coli strain CFT073 *5379 genes, bp*	*Escherichia coli* strain K-12 substr. DH10B *4126 genes, bp*	*Escherichia coli* strain K-12 isolate W3110 *4387 genes, bp*	*Escherichia coli* strain K-12 isolate MG1655 *4289 genes, bp*
3625 / 5379 67.4%	4010 / 4126 97.2%	4217 / 4387 96.1%	391 / 4289 9.1%
3637 / 5379 67.6%	3994 / 4126 96.8%	464 / 4387 10.6%	4165 / 4289 97.1%
3502 / 5379 65.1%	564 / 4126 13.7%	3939 / 4387 89.8%	3920 / 4289 91.4%
680 / 5379 12.6%	3539 / 4126 85.8%	3616 / 4387 82.4%	3567 / 4289 83.2%

Escherichia coli
strain K-12 isolate MG1655
4289 genes, 4,639,221 bp

Escherichia coli
strain K-12 isolate W3110
4387 genes, 4,641,433 bp

Escherichia coli
strain K-12 substr. DH10B
4126 genes, 4,686,137 bp

Escherichia coli
strain CFT073
5379 genes, 5,231,428 bp

DTU

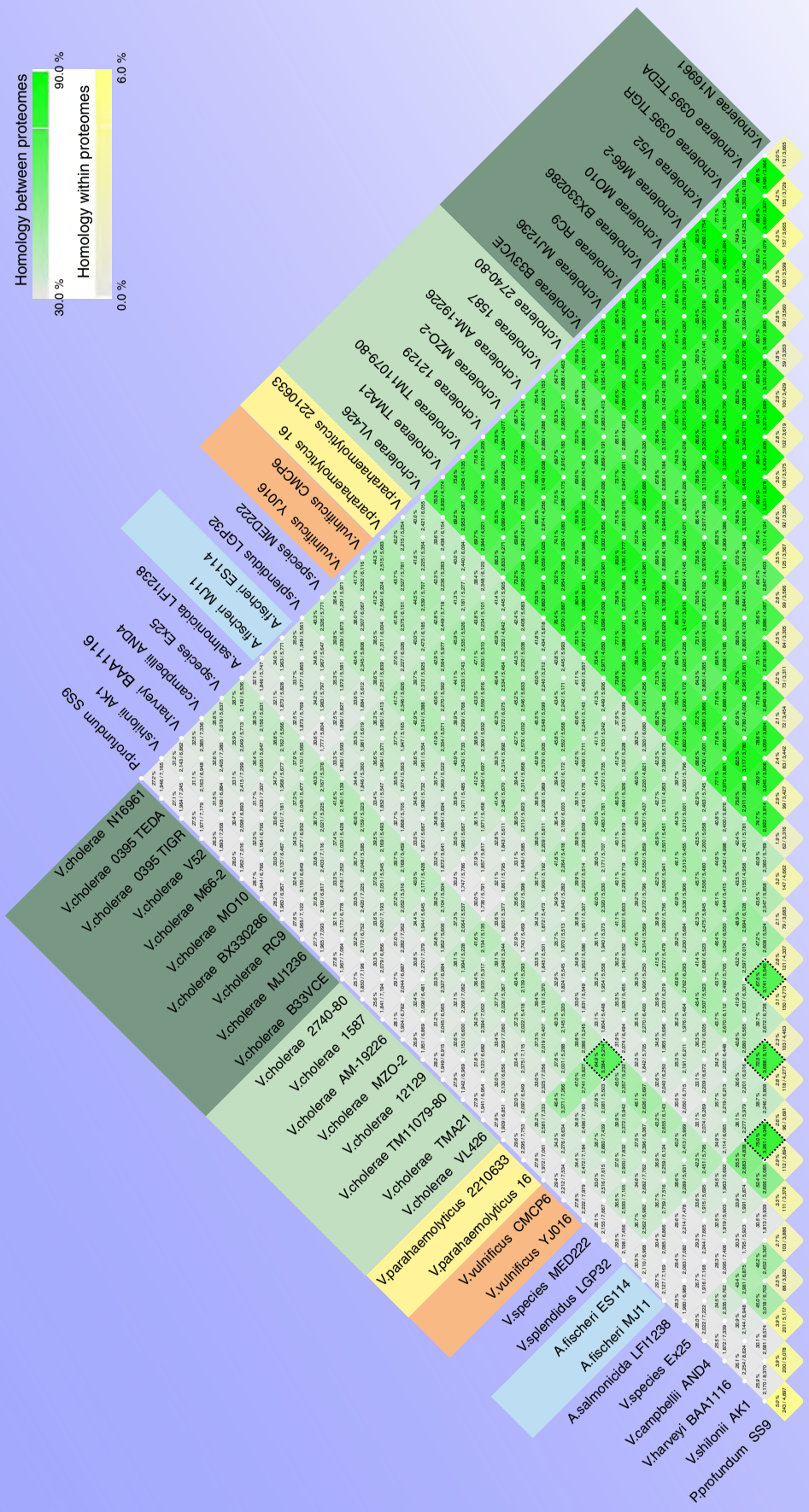
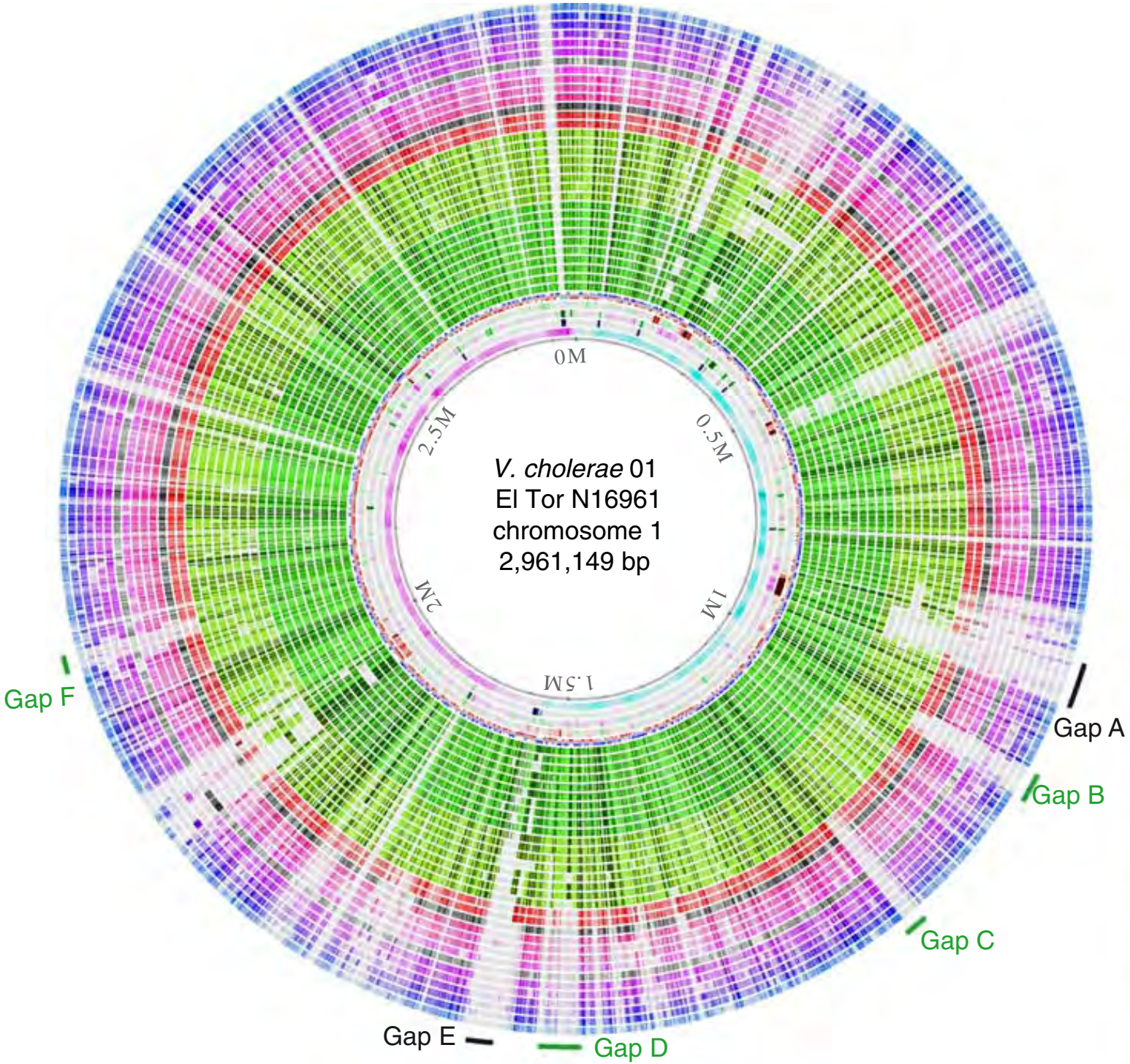
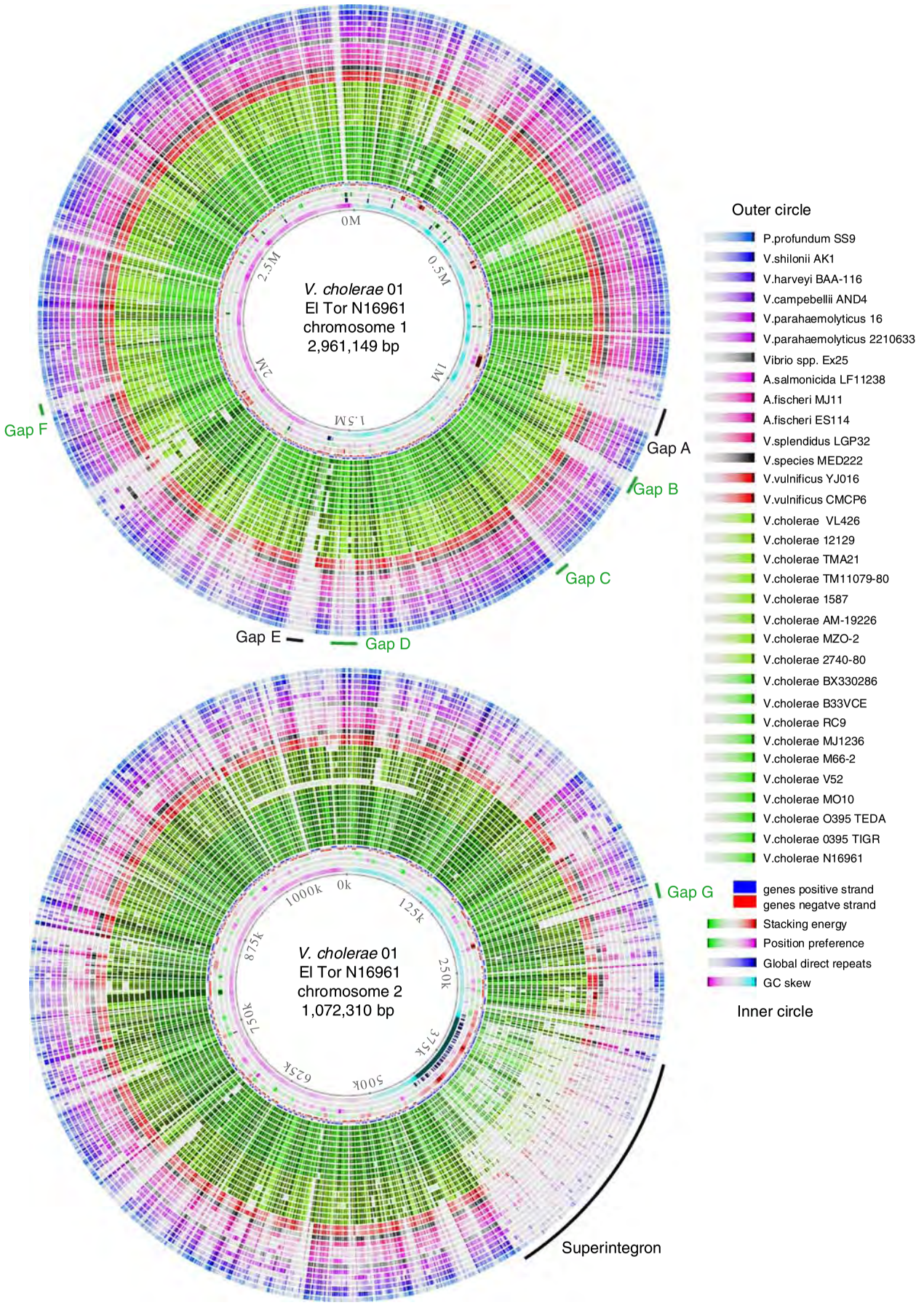


Figure 4 BLAST matrix of the 32 *Vibrionaceae* genomes. The colours highlighting the species are the same as in Fig. 1. Since the reciprocal similarity (reported as percent) is not readable at this resolution, every matrix cell is coloured using the scales as indicated. The *bottom row* identifies hits (other than hits-to-self) found within a genome. Four matrix cells reporting high pairwise similarities are *outlined*; their numbers are specified in the text


0M
0.5M
1M
1.5M
2M
2.5M
V. cholerae 01
El Tor N16961
chromosome 1
2,961,149 bp
Gap A
Gap B
Gap C
Gap D
Gap E
Gap F


V. cholerae 01
El Tor N16961
chromosome 1
2,961,149 bp
0M
0.5M
1M
1.5M
2M
2.5M
Gap A
Gap B
Gap C
Gap D
Gap E
Gap F
V. cholerae 01
El Tor N16961
chromosome 2
1,072,310 bp
0k
125k
250k
375k
500k
625k
750k
875k
1000k
Gap G
Superintegron
Outer circle
P.profundum SS9
V.shilonii AK1
V.harveyi BAA-116
V.campebellii AND4
V.parahaemolyticus 16
V.parahaemolyticus 2210633
Vibrio spp. Ex25
A.salmonicida LF11238
A.fischeri MJ11
A.fischeri ES114
V.splendidus LGP32
V.species MED222
V.vulnificus YJ016
V.vulnificus CMCP6
V.cholerae VL426
V.cholerae 12129
V.cholerae TMA21
V.cholerae TM11079-80
V.cholerae 1587
V.cholerae AM-19226
V.cholerae MZO-2
V.cholerae 2740-80
V.cholerae BX330286
V.cholerae B33VCE
V.cholerae RC9
V.cholerae MJ1236
V.cholerae M66-2
V.cholerae V52
V.cholerae MO10
V.cholerae O395 TEDA
V.cholerae 0395 TIGR
V.cholerae N16961
genes positive strand
genes negatve strand
Stacking energy
Position preference
Global direct repeats
GC skew
Inner circle