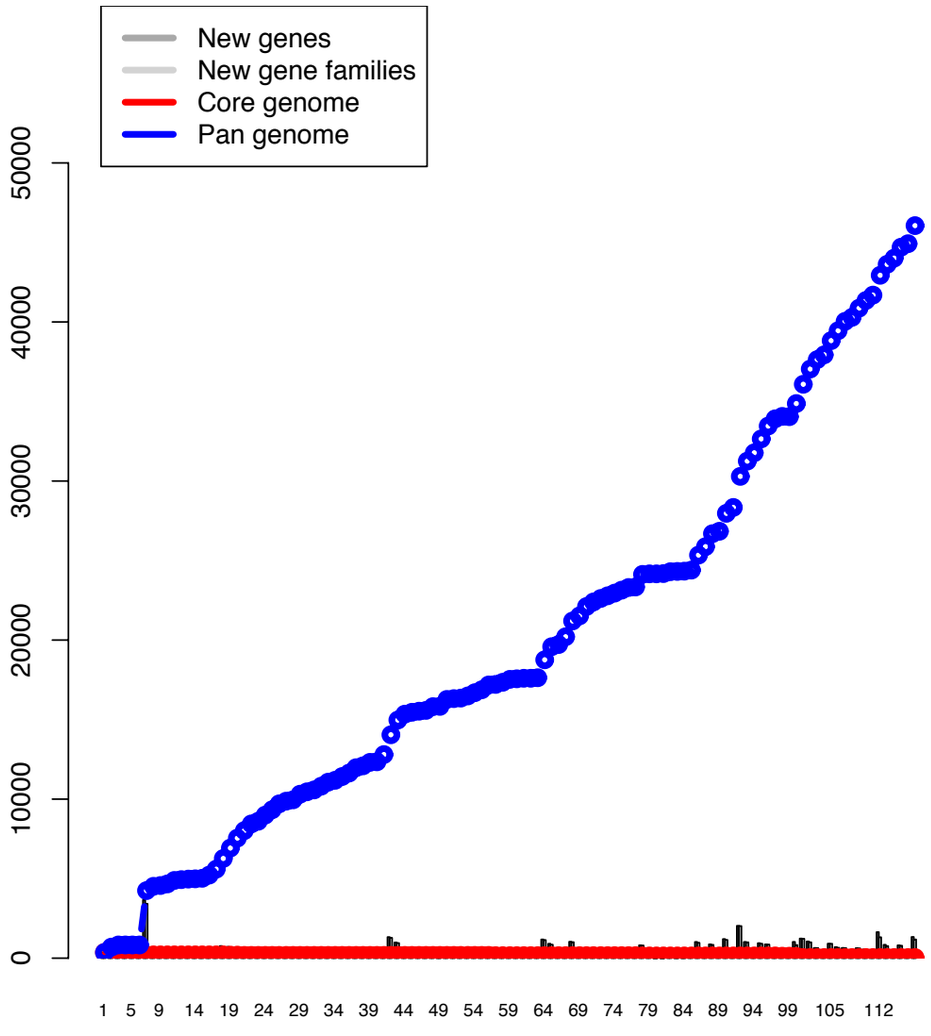
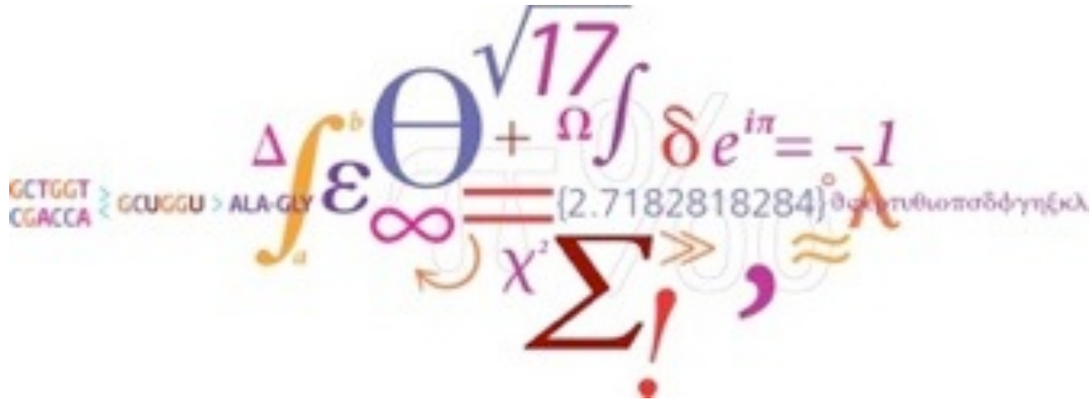
Comparative Genomics

Cautionary Tales of Next-generation Sequencing





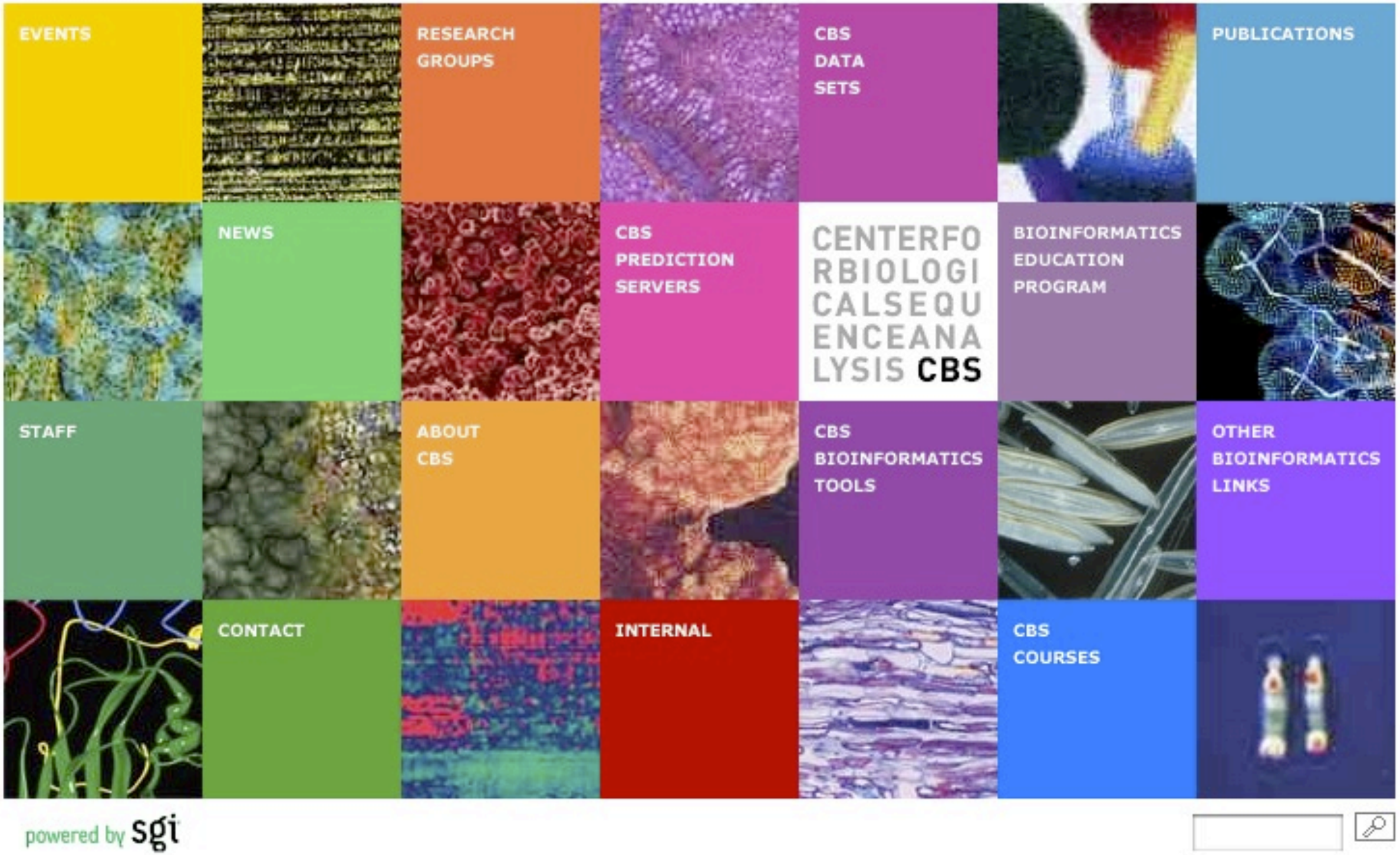
Dave Ussery
UiO course #MBV-INF 4410
Bioinformatics for Molecular Biology

Comparative Genomics lecture
Friday, 10 September, 2010

Center for Biological Sequence Analysis
Department of Systems Biology

Outline

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



www.cbs.dtu.dk

1. The problem - too much data!

Technology

The data deluge

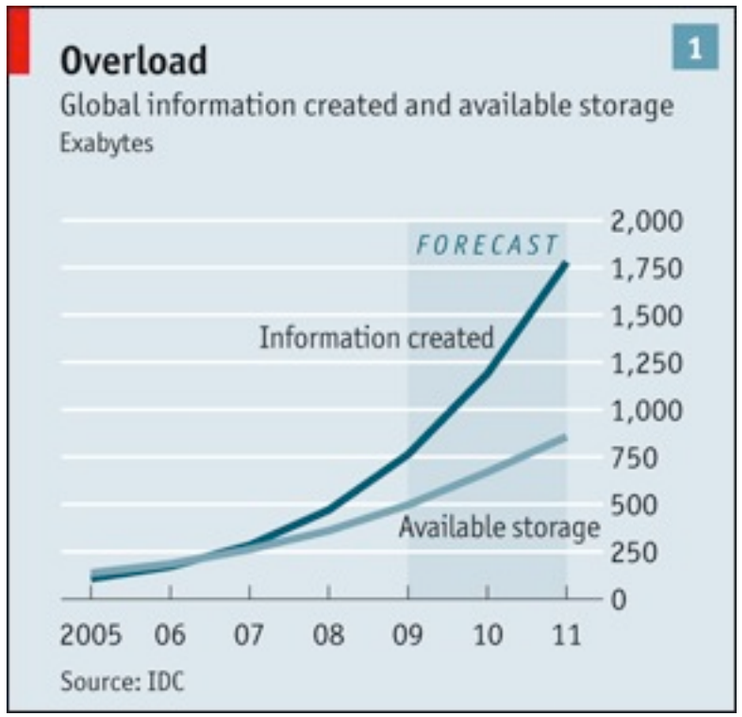
Businesses, governments and society are only starting to tap its vast potential

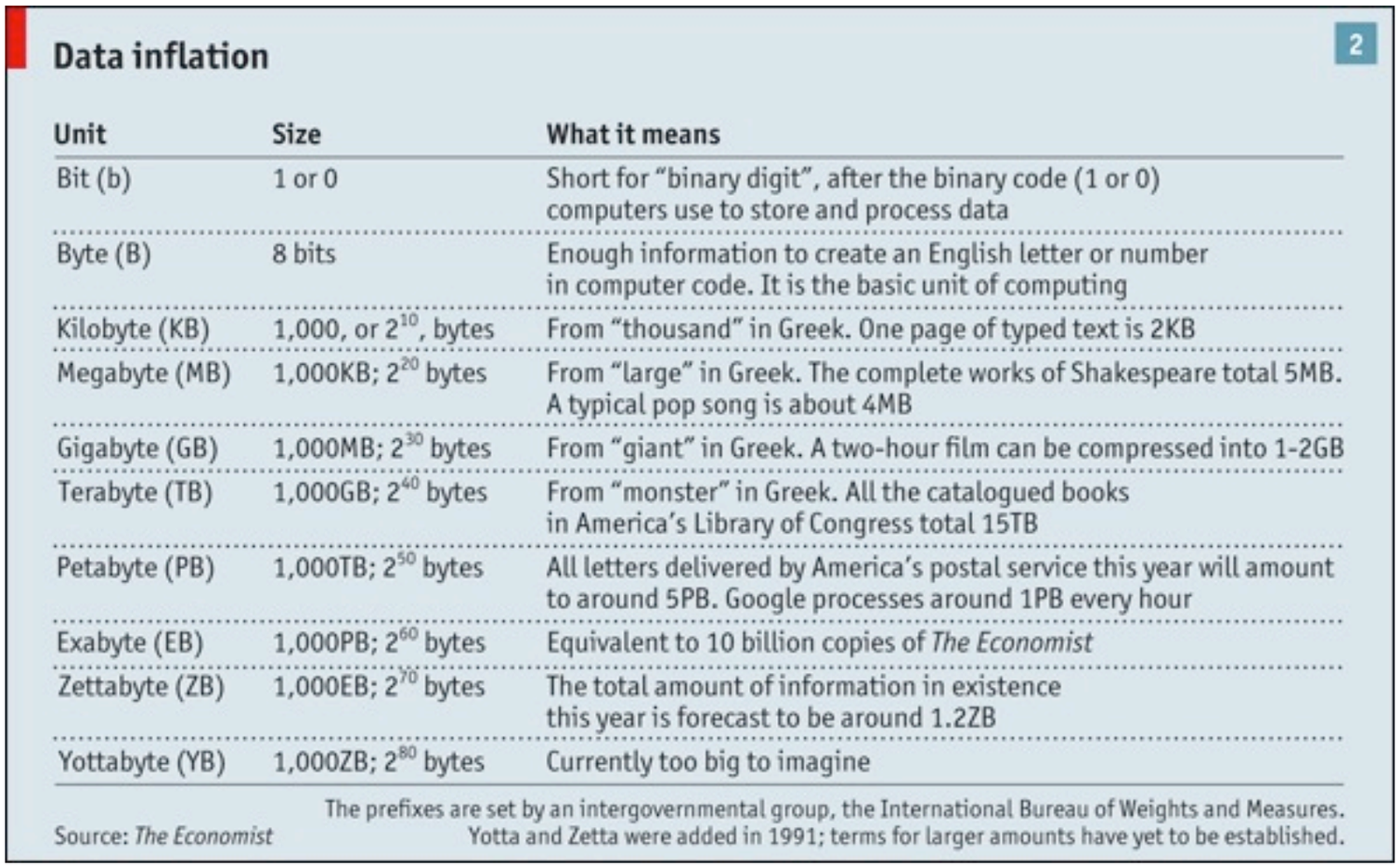
Feb 25th 2010 | From *The Economist* print edition

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!

Is this everybody's future? Probably not. But as the torrent of information increases, it is not surprising that people feel overwhelmed. "There is an immense risk of cognitive overload," explains Carl Pabo, a molecular biologist who studies cognition. The mind can handle seven pieces of information in its short-term memory and can generally deal with only four concepts or relationships at once. If there is more information to process, or it is especially complex, people become confused.

Moreover, knowledge has become so specialised that it is impossible for any individual to grasp the whole picture. A true understanding of climate change, for instance, requires a knowledge of meteorology, chemistry, economics and law, among many other things. And whereas doctors a century ago were expected to keep up with the entire field of medicine, now they would need to be familiar with about 10,000 diseases, 3,000 drugs and more than 1,000 lab tests. A study in 2004 suggested that in epidemiology alone it would take 21 hours of work a day just to stay current. And as more people around the world become more educated, the flow of knowledge will increase even further. The number of peer-reviewed scientific papers in China alone has increased 14-fold since 1990 (see chart 3).

"What information consumes is rather obvious: it consumes the attention of its recipients," wrote Herbert Simon, an economist, in 1971. "Hence a wealth of information creates a poverty of attention." But just as it is machines that are generating most of the data deluge, so they can also be put to work to deal with it. That highlights the role of "information intermediaries". People rarely deal with raw data but consume them in processed form, once they have been aggregated or winnowed by computers. Indeed, many of the technologies described in this report, from business analytics to recursive machine-learning to visualisation software, exist to make data more digestible for humans

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

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)

illumina

APPLICATIONS SYSTEMS SERVICES SCIENCE SUPPORT COMPANY

Systems / Genome Analyzer IIx

Genome Analyzer IIx

Applications: **DNA Sequencing**, **Gene Regulation Analysis**, **Sequencing-Based Transcriptome Analysis**, **SNP Discovery and Structural Variation Analysis**, **Cytogenetic Analysis**, **DNA-Protein Interaction Analysis (ChIP-Seq)**, **Sequencing-Based Methylation Analysis**

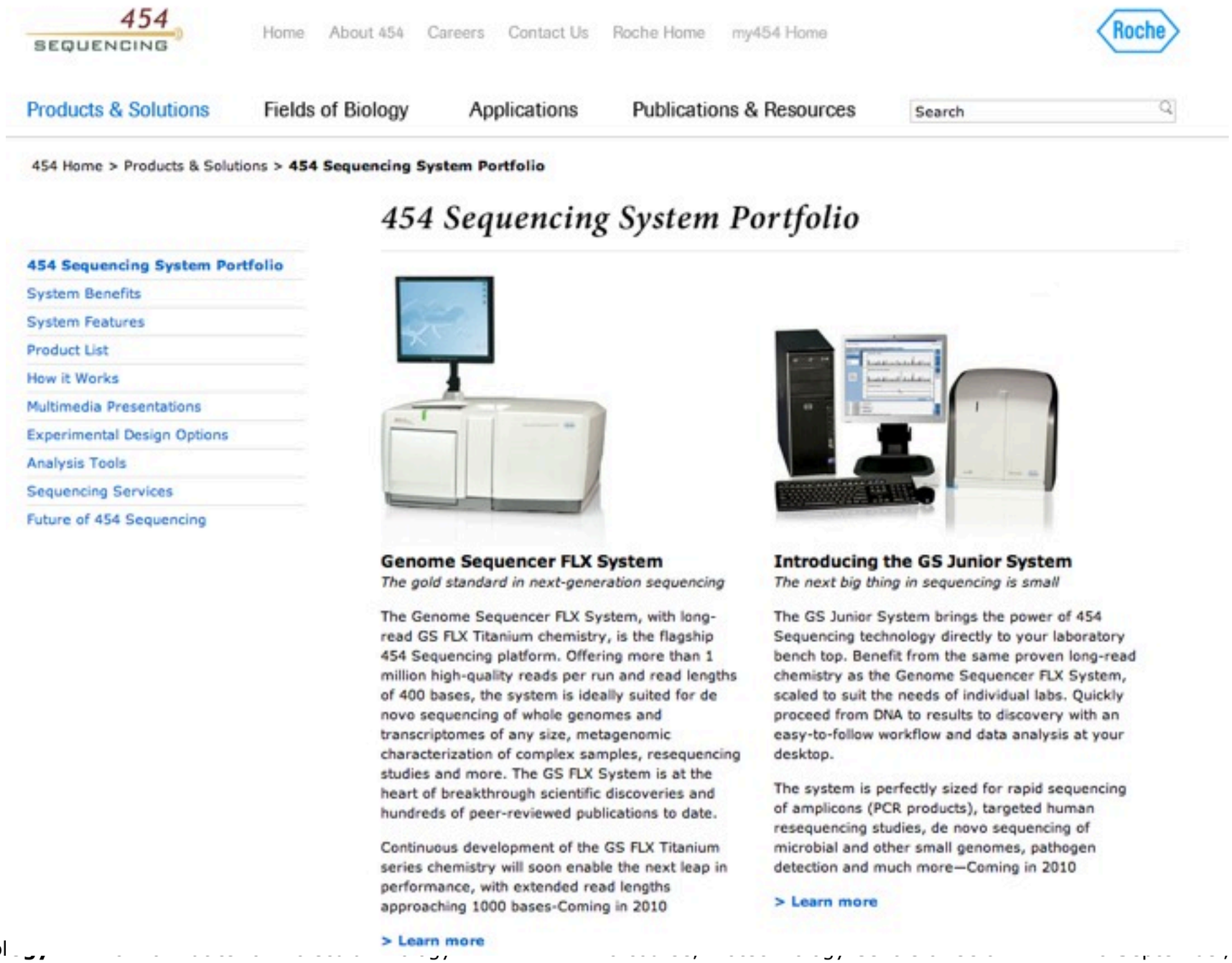
The Genome Analyzer IIx offers a unique combination of 2 x 100 bp read length and up to 500 million reads per flow cell with the simplest and fastest workflow. The highest raw accuracy and the largest number of perfect reads enables a broad range of high-throughput sequencing applications. Power your discoveries and generate highly accurate results in a week with the Genome Analyzer IIx. **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. 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

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

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

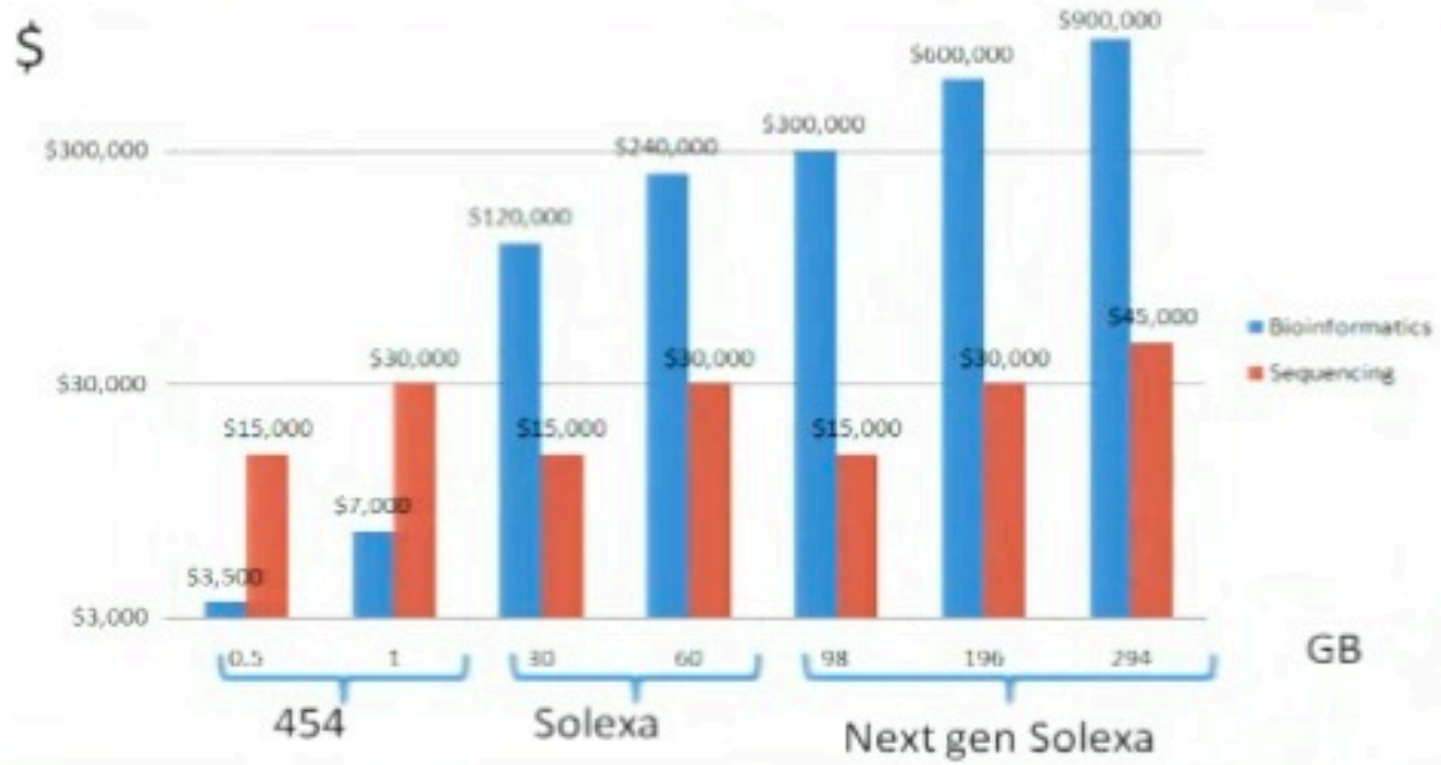
Genome Research, Jan 2009

"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 37 Solexa machines,
8 ABI Solids, 2 Roche 454 machines

>10,000 teraBytes per month!

The problem: Sequencing outpaces Moore's law



- 95GB == 195,600 node hours (on Nehalem 8core, 16GB),
- Illumina HiSeq2000 = 2x100GB/run
- cost is purely BLAST, no storage or transfer cost
- values are in Amazon EC2 (from *Wilkening et al, IEEE Cluster09*)
- note: 10x or 100x improvements over BLASTX will help, but not solve

Screen shot from Foker Meyer's talk at the GSC 9 meeting. (held at the J. Craig Venter Institute, Rockville, Maryland, USA, 28-30 April, 2010).

2. A brief history - The speed of sequencing

What is a genome?

genome d3*i*.noum. *Biol.* Formerly also genom -nom. [a. *G. genom* (H. Winkler *Verbreitung u. Ursache d. Parthenogenesis* (**1920**) iv. 165), irreg. f. *gen* gene[1] + *chromosom* chromosome.] **A haploid set of chromosomes; the sum-total of the genes in such a set.**

1930 *Cytologia* I. 14 Chromosomes from different sets (or genoms) of *Triticum vulgare* show affinity toward each other.

1930 [see allopolyploidy].

1932 *Proc. 6th Int. Congr. Genetics* I. 275 The inviability of deficient genomes in the haploid generation serves to some extent as an alternative distinction between mutation and deficiency.

1932 *Proc. 6th Int. Congr. Genetics* II. 5 There are two species having genoms resembling *C. neglecta*.

1952 C. P. Blacker *Eugenics* x. 243 The appearance of such terms as gene-complex and genome (denoting a set of chromosomes as a working unity) testify to the movement towards holism in genetics.

1965 A. M. Srb et al. *Gen. Genetics* (ed. 2) vii. 190 Among organisms with chromosomes, each species has a characteristic set of genes, or genome. In diploids a genome is found in each normal gamete. It consists of a full set of the different kinds of chromosomes.

1970 *Sci. Amer.* Oct. 19/1 The human genome..consists of perhaps as many as 10 million genes.

THE OXFORD ENGLISH DICTIONARY

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!

year	# genes mapped	#years to sequence human genome
1970	none	not possible
1980	3	~4,000,000 years
1990	12	~1000 years
2000	~25,000	draft
2010	43,887	a few hours (!)

~40 genomes sequenced (so far!)
plans for 1000 genomes

SCIENTIST AT WORK: GEORGE M. CHURCH

On a Mission to Sequence the Genomes of 100,000 People

By DAVID EWING DUNCAN
Published in The New York Times, on 7 June 2010

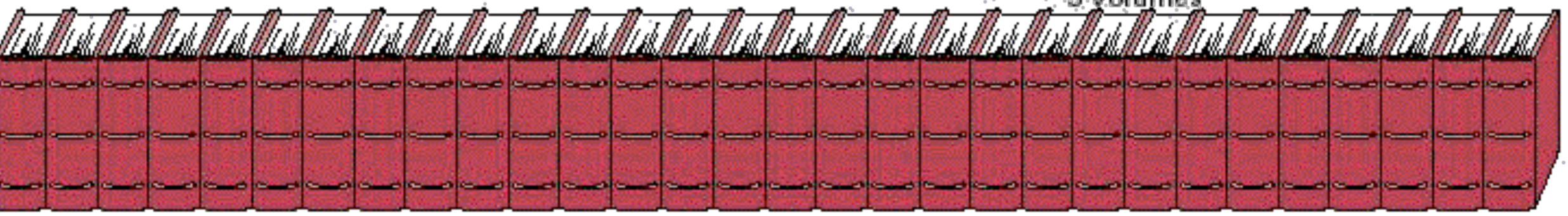
Traditionally, biology is about taking apart things like cells to better understand them. For the geneticist George M. Church, the main objective is to put the pieces back together.

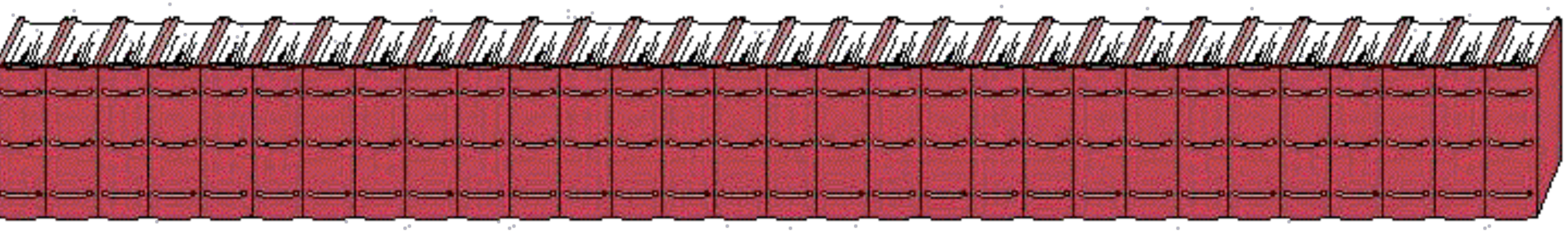
Phage λ
50 kb
2 pages

Escherichia coli
(bacteria)
4.7 Mb
200 pages

Saccharomyces cerevisiae
(yeast)
12.5 Mb
500 pages

Caenorhabditis elegans
(nematode)
Arabidopsis thaliana
(plant)
100 Mb
3 volumes





Human being
3000 Mb
80 volumes

25 kb per page
1500 pages
per volume
(2 inches thick)

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. Jens Jensen'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.

NATURE|Vol 464|1 April 2010 HUMAN GENOME AT TEN NEWS FEATURE

HUMAN GENOME

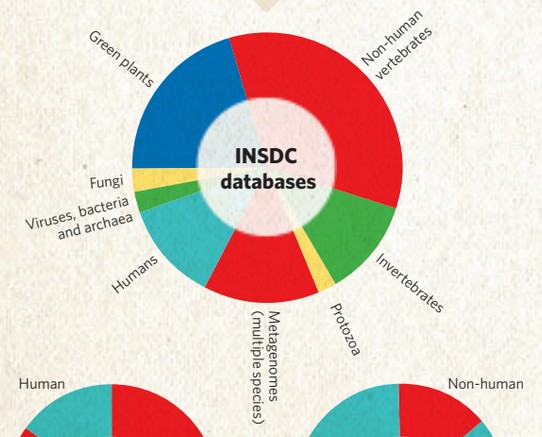
THE SEQUENCE EXPLOSION

At the time of the announcement of the first drafts of the human genome in 2000, there were 8 billion base pairs of sequence in the three main databases for 'finished' sequence: GenBank, run by the US National Center for Biotechnology Information; the DNA Databank of Japan; and the European Molecular Biology Laboratory (EMBL) Nucleotide Sequence Database. The databases share their data regularly as part of the International Nucleotide Sequence Database Collaboration (INSDC). In the subsequent first post-genome decade, they have added another 270 billion bases to the collection of finished sequence, doubling the size of the database roughly every 18 months. But this number is dwarfed by the amount of raw sequence that has been created and stored by researchers around the world in the Trace archive and Sequence Read Archive (SRA).

See Editorial, page 649, and human genome special at www.nature.com/humangenome

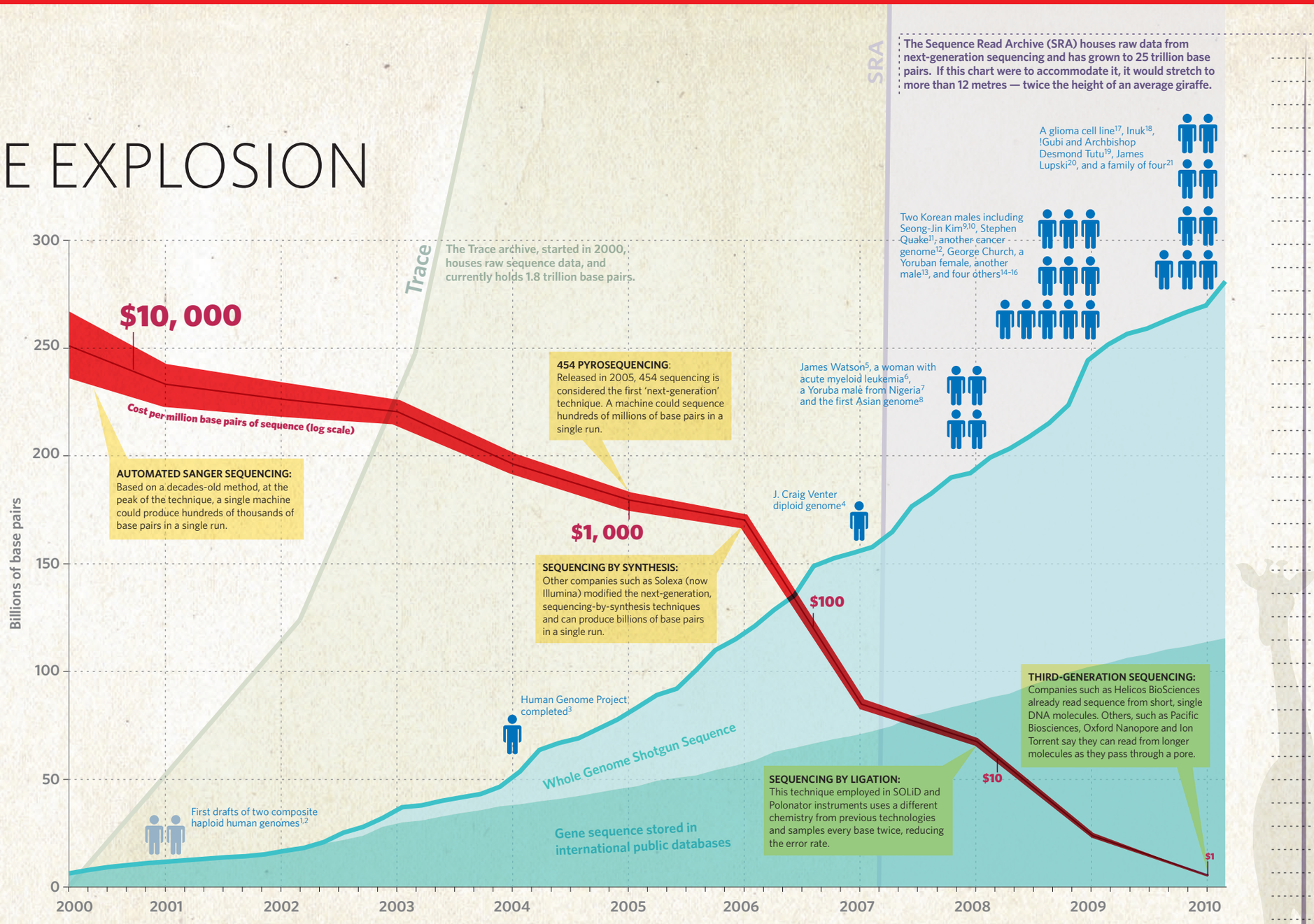
DNA SEQUENCES BY TAXONOMY

International Nucleotide Sequence Database Collaboration: The main repositories of 'finished' sequence span a wide range of organisms, representing the many priorities of scientists worldwide.



Trace Archive: Developed to house the raw output of high-throughput sequencers built in the late 1990s, the trace archive spans a wide range of taxa.

Sequence Read Archive: Houses raw data from next-generation sequencers. Dominated by human sequence, including multiple coverage for more than 170 people.

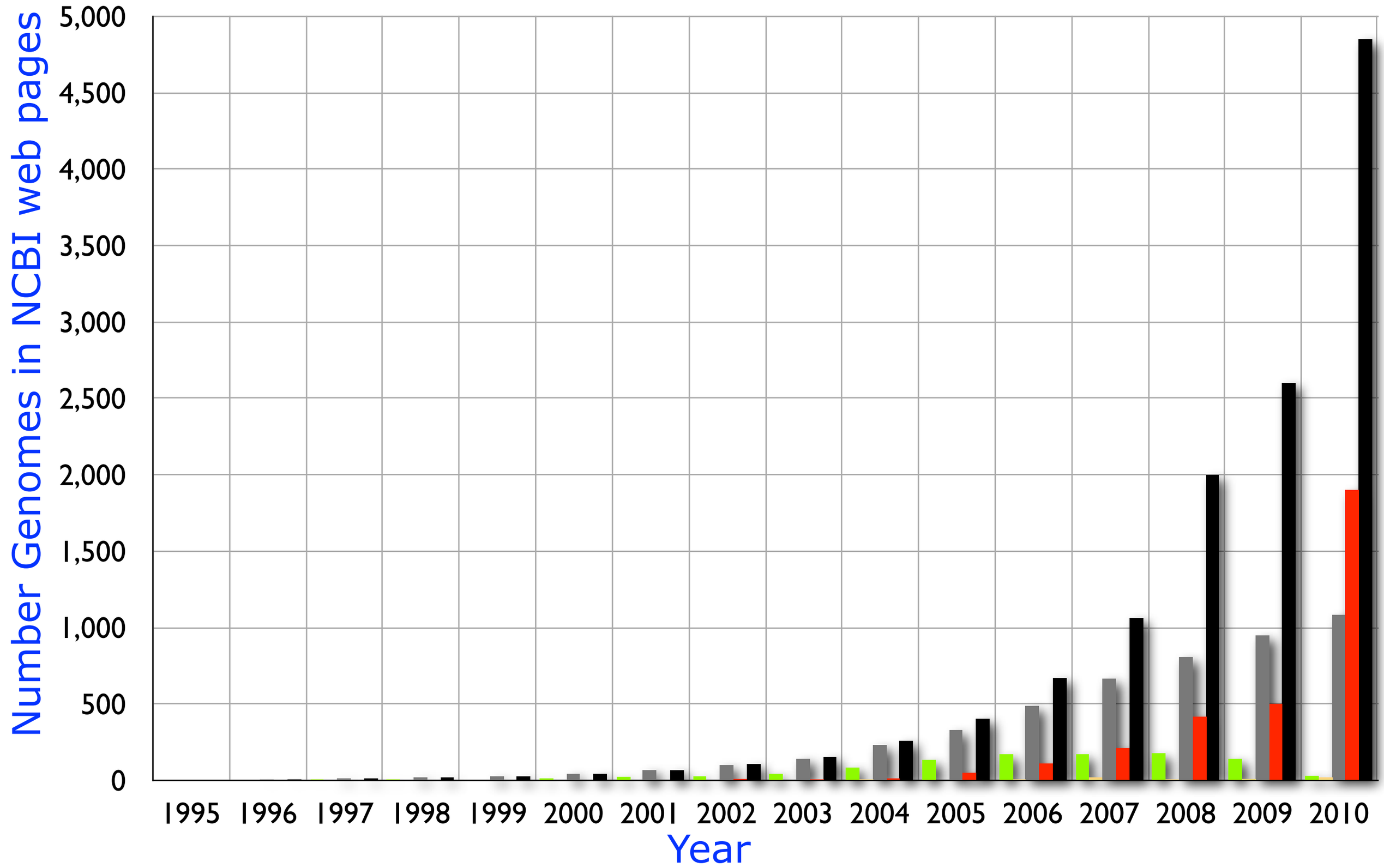


HOW MANY HUMAN GENOMES?

The graphic shows all published, fully sequenced human genomes since 2000, including nine from the first quarter of 2010. Some are resequencing efforts on the same person and the list does not include unpublished completed genomes.

1. Venter, J. C. *et al. Science* **291,** 1304–1351 (2001).
2. International Human Genome Sequencing Consortium *Nature* **409,** 860–921 (2001).
3. International Human Genome Sequencing Consortium *Nature* **431,** 931–945 (2004).
4. Levy, S. *et al. PLoS Biol.* **5,** e254 (2007).
5. Wheeler, D. A. *et al. Nature* **452,** 872–876 (2008).
6. Ley, T. J. *et al. Nature* **456,** 66–72 (2008).
7. Bentley, D. R. *et al. Nature* **456,** 53–59 (2008).
8. Wang, J. *et al. Nature* **456,** 60–65 (2008).
9. Ahn, S.-M. *et al. Genome Res.* **19,** 1622–1629 (2009).
10. Kim, J.-I. *et al. Nature* **460,** 1011–1015 (2009).
11. Pushkarev, D., Neff, N. F. & Quake, S. R. *Nature Biotechnol.* **27,** 847–850 (2009).
12. Mardis, E. R. *et al. N. Engl. J. Med.* **10,** 1058–1066 (2009).
13. Drmanac, R. *et al. Science* **327,** 78–81 (2009).
14. McKernan, K. J. *et al. Genome Res.* **19,** 1527–1541 (2009).
15. Pleasance, E. D. *et al. Nature* **463,** 191–196 (2010).
16. Pleasance, E. D. *et al. Nature* **463,** 184–190 (2010).
17. Clark, M. J. *et al. PLoS Genet.* **6,** e1000832 (2010).
18. Rasmussen, M. *et al. Nature* **463,** 757–762 (2010).
19. Schuster, S. C. *et al. Nature* **463,** 943–947 (2010).
20. Lupski, J. R. *et al. N. Engl. J. Med.* doi:10.1056/NEJMoa0908094 (2010).
21. Roach, J. C. *et al. Science* doi:10.1126/science.1186802 (2010).

Page size by comparison


Bacteria
Archaea
total published
Unfinished
total
Number Genomes in NCBI web pages
5,000
4,500
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

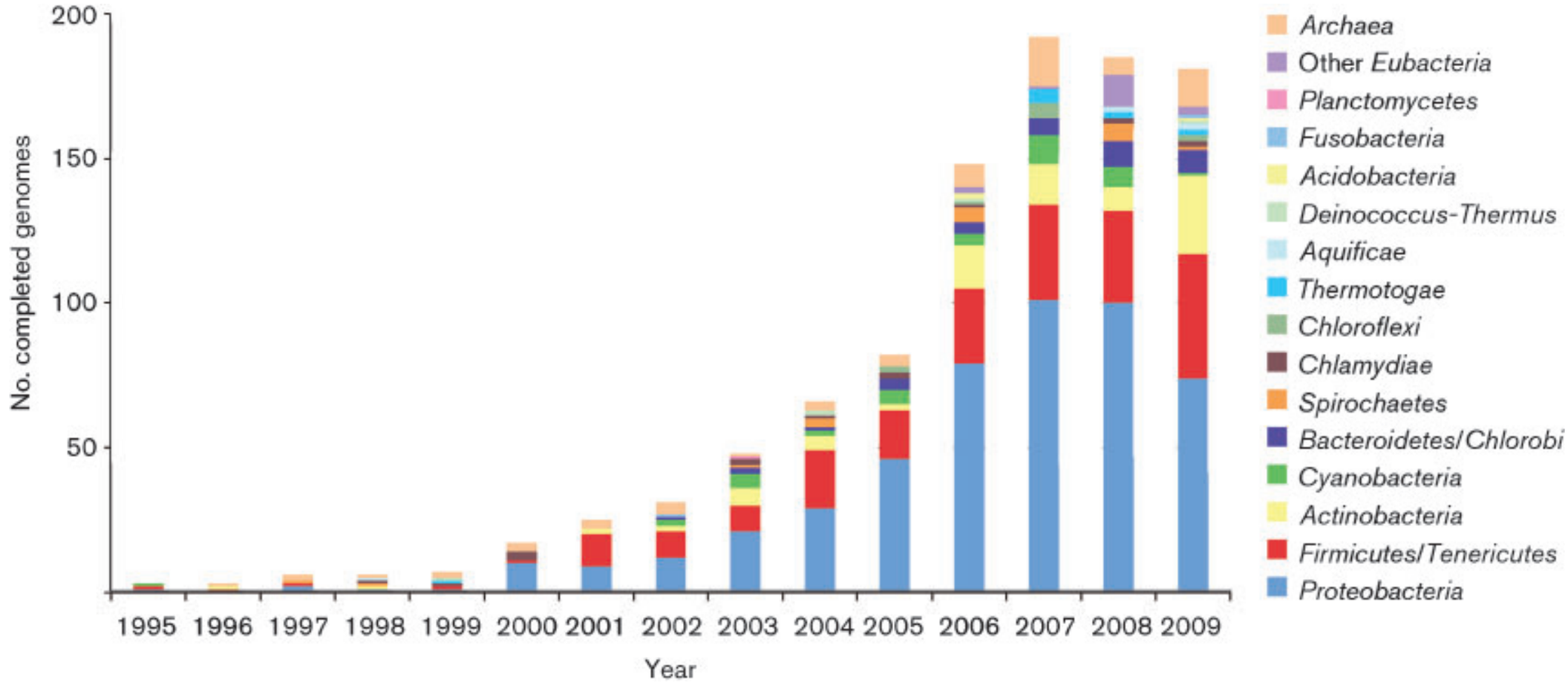
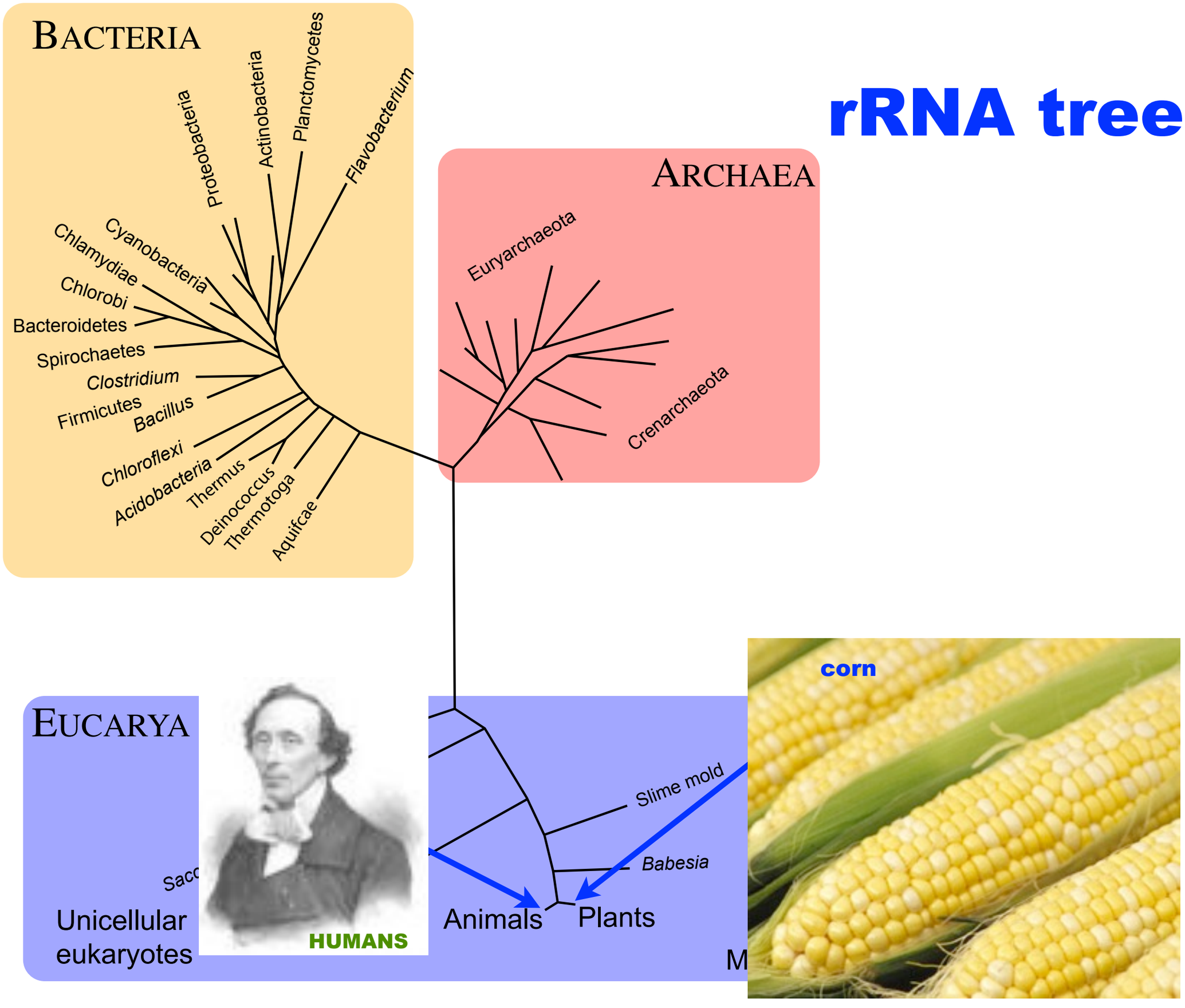


Fig. 1. Increase in the number of genomes completed per year separated by bacterial phylum. Data source: NCBI, complete genomes (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi).

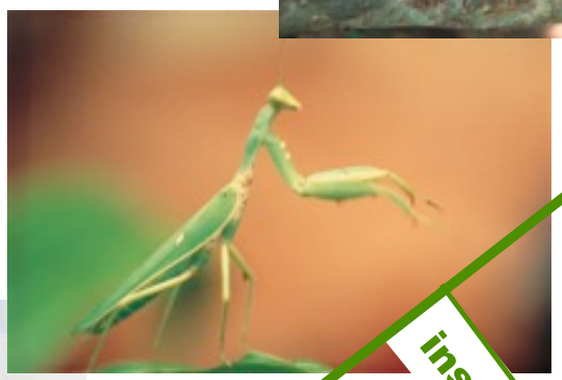
Microbiology, **156**:603-608, (2010).

rRNA tree


BACTERIA
Proteobacteria
Actinobacteria
Planctomycetes
Flavobacterium
Cyanobacteria
Chlamydiae
Chlorobi
Bacteroidetes
Spirochaetes
Clostridium
Firmicutes
Bacillus
Chloroflexi
Acidobacteria
Thermus
Deinococcus
Thermotoga
Aquifcae
ARCHAEA
Euryarchaeota
Crenarchaeota
EUCARYA
Unicellular eukaryotes
HUMANS
Slime mold
Babesia
Animals
Plants
corn

Aristotle's ladder of complexity

HUMANS



mammals

birds

reptiles

fish

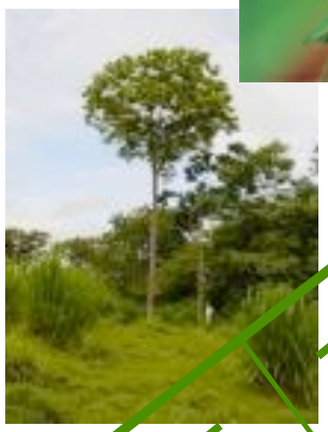
insects

jelly fish

higher plants

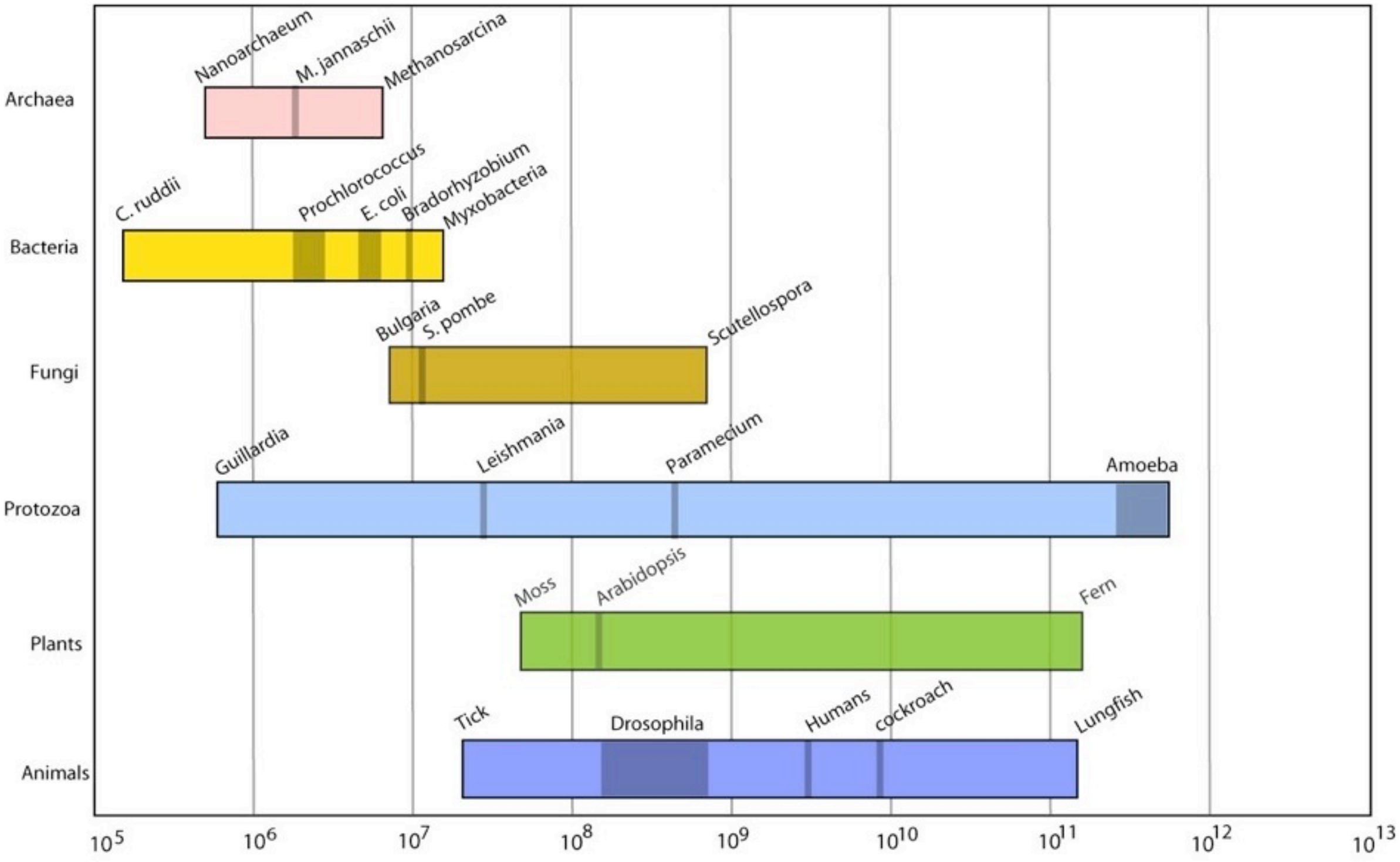
lower plants

minerals



4. Approaches to handle lots of data

Statistics

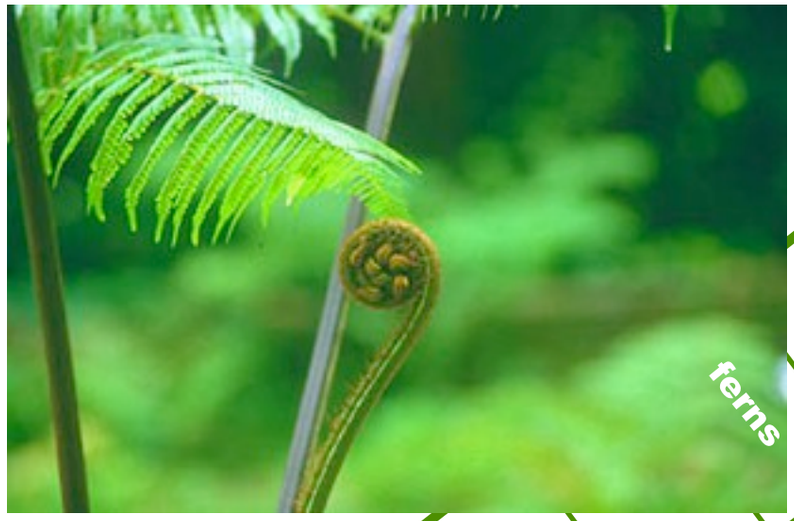


Database of Genome Sizes

(DOGS)

ladder of

complexity



lungfish

cock - roaches

humans

fruit- flies

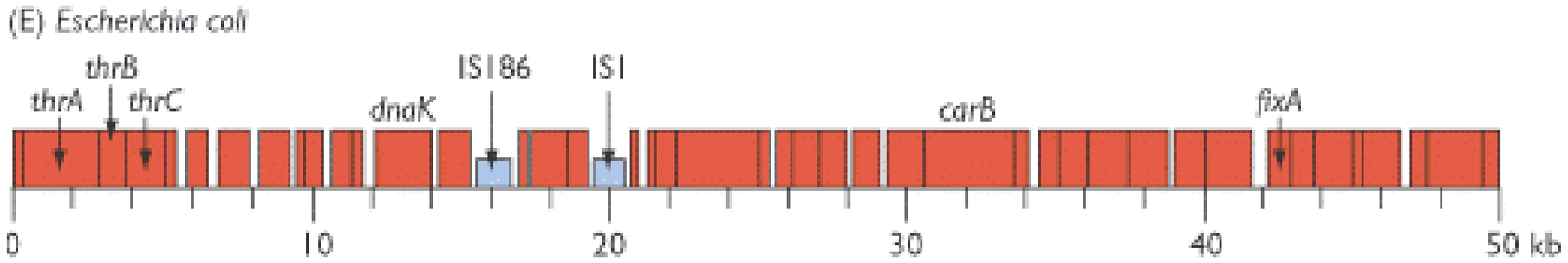
minerals

The “C-value paradox”

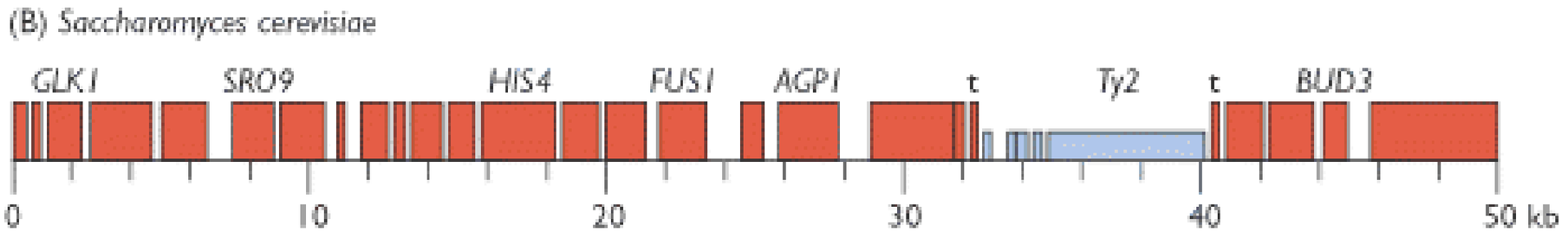
The genome size of an organism is defined as the amount of haploid DNA in a genomic set (e.g., an egg or sperm nucleus). This is also referred to as the "C-value"; the "C" means "constant" or "characteristic", since the size of a genome is usually constant for a given species.

The large difference in genome sizes without any seeming relation to an organism’s complexity, is called the C-value paradox.

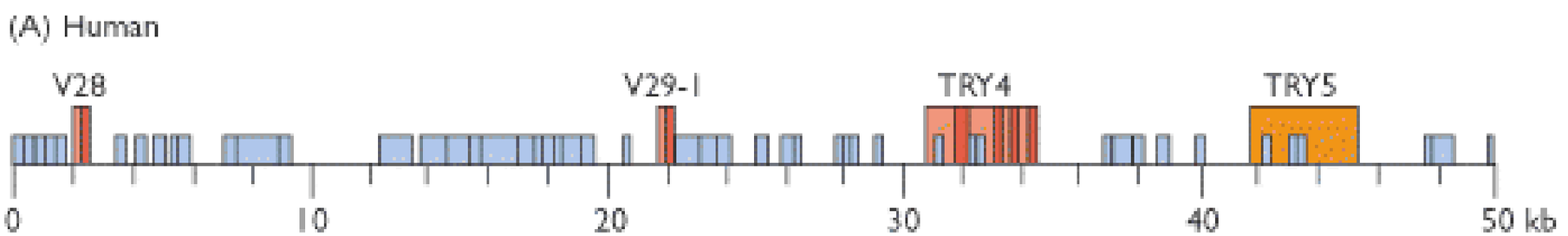
What does all this DNA do?



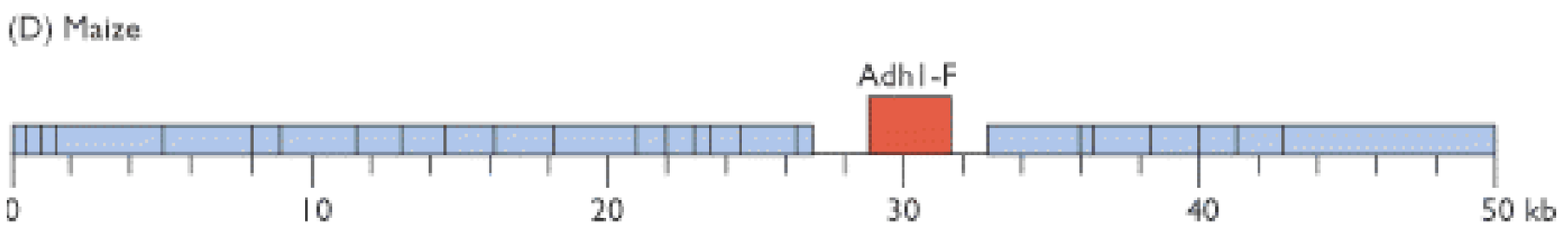
90%



50%



2%



<1%

DNA repeats

The approximate size and characteristics of genomes was characterised in the 1960s, in a classic study of the kinetics of DNA reassociation by Britten and Kohne (1968).

They found that the DNA could be divided into four fractions:

1. foldback DNA
2. highly repetitive DNA
3. middle-repetitive DNA
4. single-copy DNA

The repetitive DNA can either be localised to discrete regions, or dispersed.

Britten,R.J., Kohne,D.E., "Repeated sequences in DNA", Science, **161**:529-540, (1968).

Highly repetitive DNA

Dispersed - e.g., Alu family

- about 300 bp long
- 500,000 copies in humans
- (about 5% of the human genome)
- dispersed throughout the chromosomes

Localised highly repetitive sequences

- about 2-10 bp long
- present in millions of copies, often in large blocks
- (about 6% of the human genome)
- associated with heterochromatin
- usually very high A+T content

Localised repetitive DNA

Often, satellite DNA consists of long tandem arrays of repeated sequences, all localised to one or a few discrete regions in the chromosomes. For example, in the kangaroo rat (*Dipodomys ordii*), more than 50% of the genome consists of three families of repeated sequences:

$(AAG)_n$, where n = ~2.24 x 10^9

$(TTAGGG)_n$, where n = ~2.2 x 10^9

$(ACACAGCGGG)_n$, where n = ~1.2 x 10^9

Middle repetitive DNA

- makes up more than 40% of the human genome
- position varies due to transposable elements
- Includes the following types of sequences:
  - Dinucleotide repeats
  - microsatellite DNA
  - TRInucleotide repeats

- associated with many diseases
- (*e.g.*, Fragile X, muscular distrophy)

Mobile elements create structural variation: Analysis of a complete human genome

Jinchuan Xing,[1] Yuhua Zhang,[1] Kyudong Han,[2] Abdel Halim Salem,[2,3,5] Shurjo K. Sen,[2,6] Chad D. Huff,[1] Qiong Zhou,[1] Ewen F. Kirkness,[4] Samuel Levy,[4] Mark A. Batzer,[2] and Lynn B. Jorde[1,7]

[1]*Department of Human Genetics, Eccles Institute of Human Genetics, University of Utah, Salt Lake City, Utah 84109, USA;* [2]*Department of Biological Sciences, Louisiana State University, Baton Rouge, Louisiana 70803, USA;* [3]*Department of Anatomy, Faculty of Medicine, Suez Canal University, Ismailia 41111, Egypt;* [4]*J. Craig Venter Institute, Rockville, Maryland 20850, USA*

Structural variants (SVs) are common in the human genome. Because approximately half of the human genome consists of repetitive, transposable DNA sequences, it is plausible that these elements play an important role in generating SVs in humans. Sequencing of the diploid genome of one individual human (HuRef) affords us the opportunity to assess, for the first time, the impact of mobile elements on SVs in an individual in a thorough and unbiased fashion. In this study, we systematically evaluated more than 8000 SVs to identify mobile element-associated SVs as small as 100 bp and specific to the HuRef genome. Combining computational and experimental analyses, we identified and validated 706 mobile element insertion events (including *Alu*, L1, SVA elements, and nonclassical insertions), which added more than 305 kb of new DNA sequence to the HuRef genome compared with the Human Genome Project (HGP) reference sequence (hg18). We also identified 140 mobile element-associated deletions, which removed ~126 kb of sequence from the HuRef genome. Overall, ~10% of the HuRef-specific indels larger than 100 bp are caused by mobile element-associated events. More than one-third of the insertion/deletion events occurred in genic regions, and new *Alu* insertions occurred in exons of three human genes. Based on the number of insertions and the estimated time to the most recent common ancestor of HuRef and the HGP reference genome, we estimated the *Alu*, L1, and SVA retrotransposition rates to be one in 21 births, 212 births, and 916 births, respectively. This study presents the first comprehensive analysis of mobile element-related structural variants in the complete DNA sequence of an individual and demonstrates that mobile elements play an important role in generating inter-individual structural variation.

[Supplemental material is available online at http://www.genome.org. The sequence data from this study have been submitted to GenBank (http://www.ncbi.nlm.nih.gov/Genbank/) under accession nos. FI569689–FI569698.]

19:1516–1526 © 2009 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/09; www.genome.org

Conclusion (part 1):

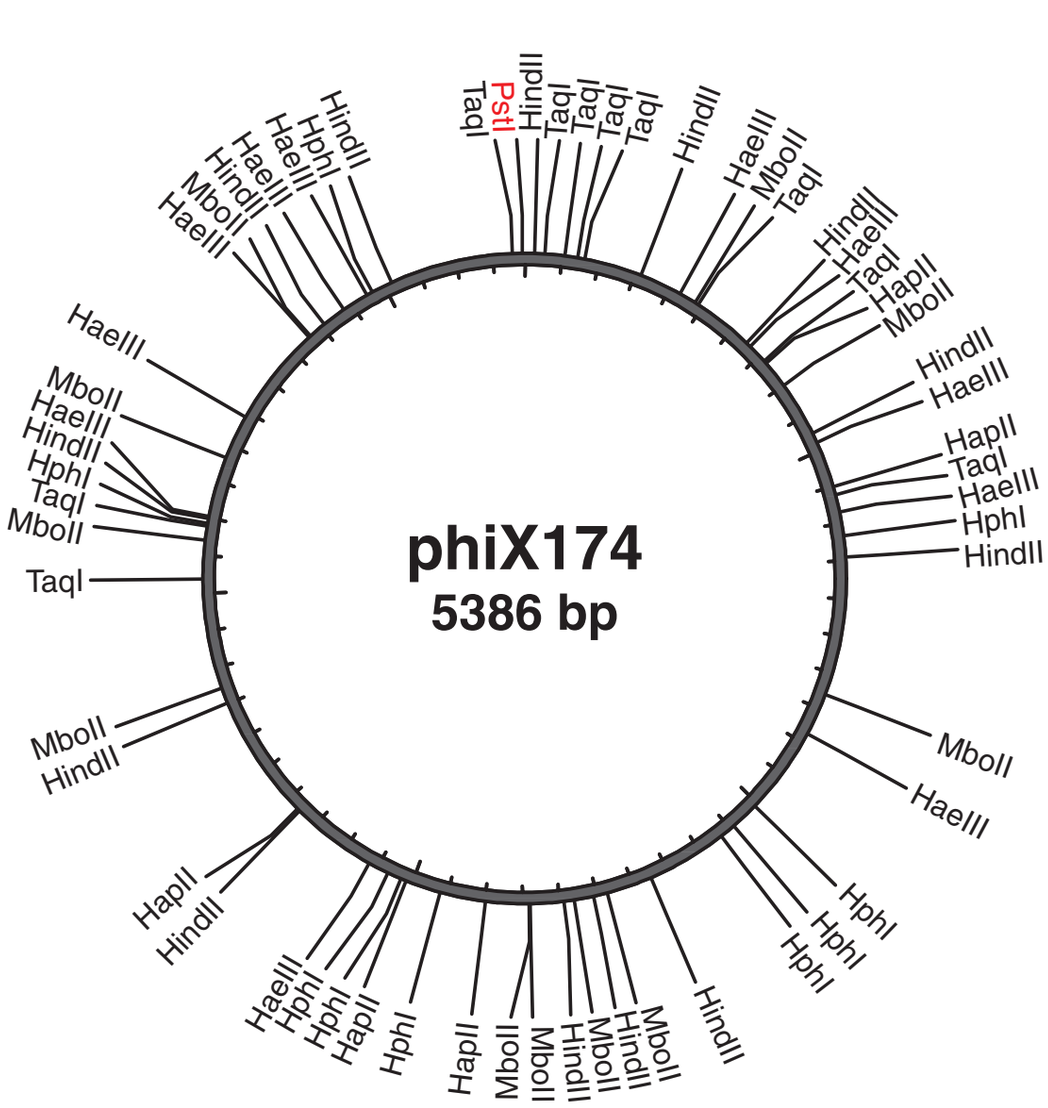
People are different!

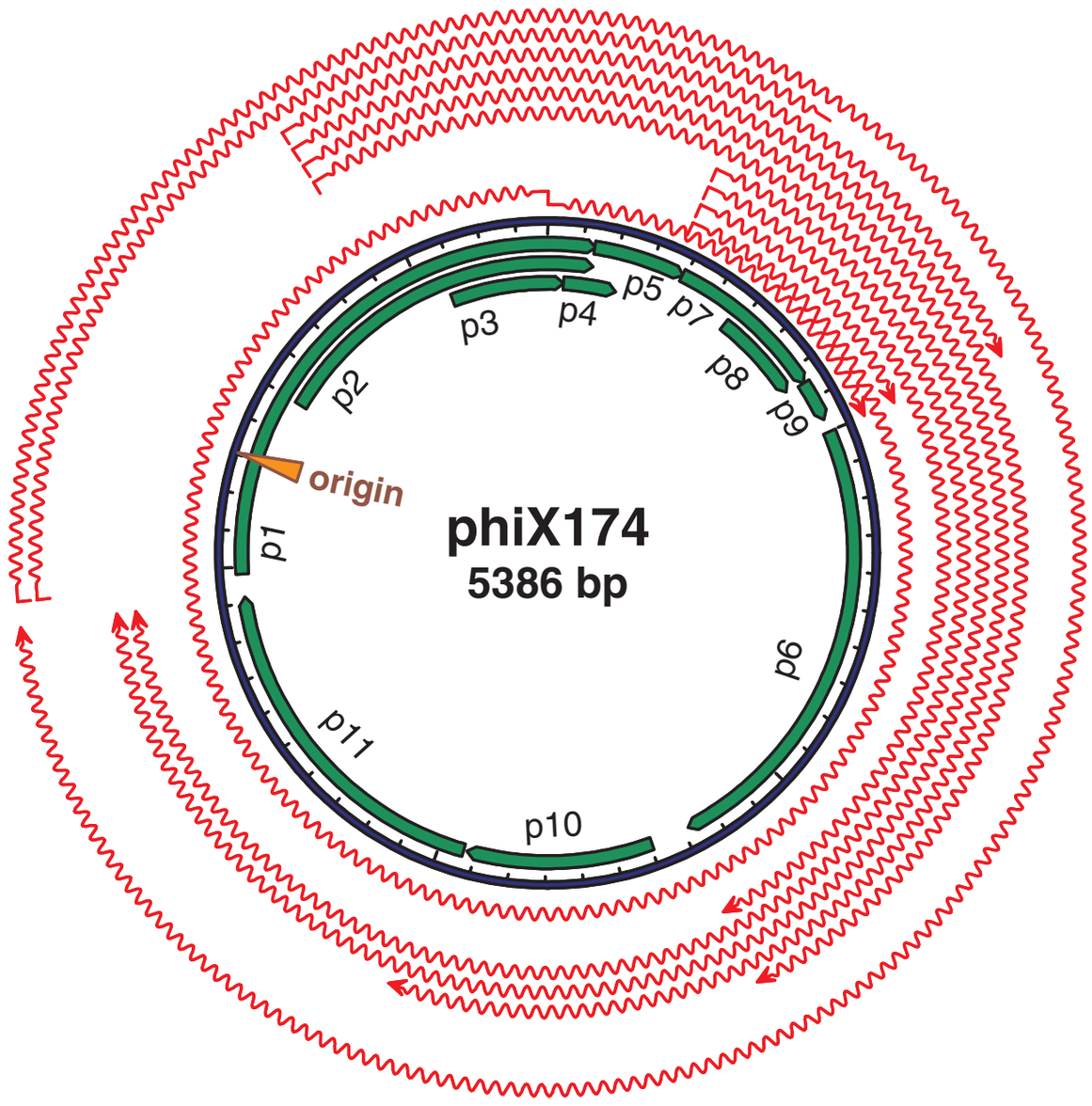
Alu repeats - 1 in 21 births => 6.8 billion people/21 = 323 MILLION variants!

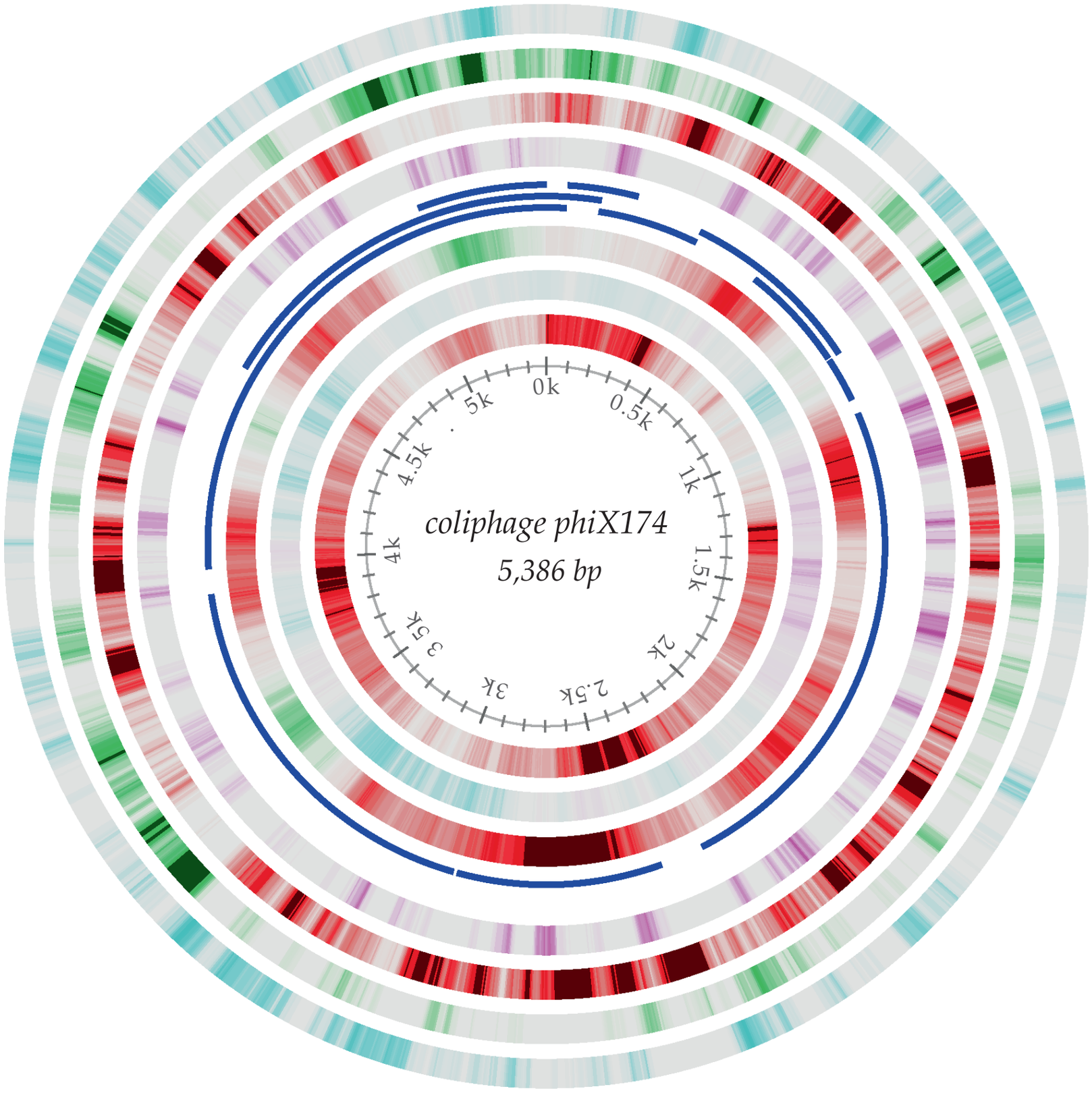
L1 repeats - 1 in 180 births => 6.8 billion people/180 = 37 MILLION variants!

SCA repeats - 1 in 916 births => 6.8 billion people/916 = 7 MILLION variants!

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 ttttactttt
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 cagtaccttа 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


phiX174
5386 bp
TaqI
PstI
HindII
TaqI
TaqI
TaqI
TaqI
HindII
HaeIII
MboII
TaqI
HindII
HaeIII
TaqI
HapII
MboII
HindII
HaeIII
HapII
TaqI
HaeIII
HphI
HindII
MboII
HaeIII
HaeIII
HphI
HindII
MboII
HaeIII
HaeIII
MboII
HaeIII
HindII
HphI
TaqI
MboII
TaqI
HaeIII
MboII
HindII
HapII
HindII
HaeIII
HphI
HphI
HapII
HphI
HapII
MboII
MboII
HindII
MboII
HindII
MboII
HindII
HphI
HphI
HphI
HaeIII


phiX174
5386 bp
origin
p1
p2
p3
p4
p5
p7
p8
p9
p6
p10
p11



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 +

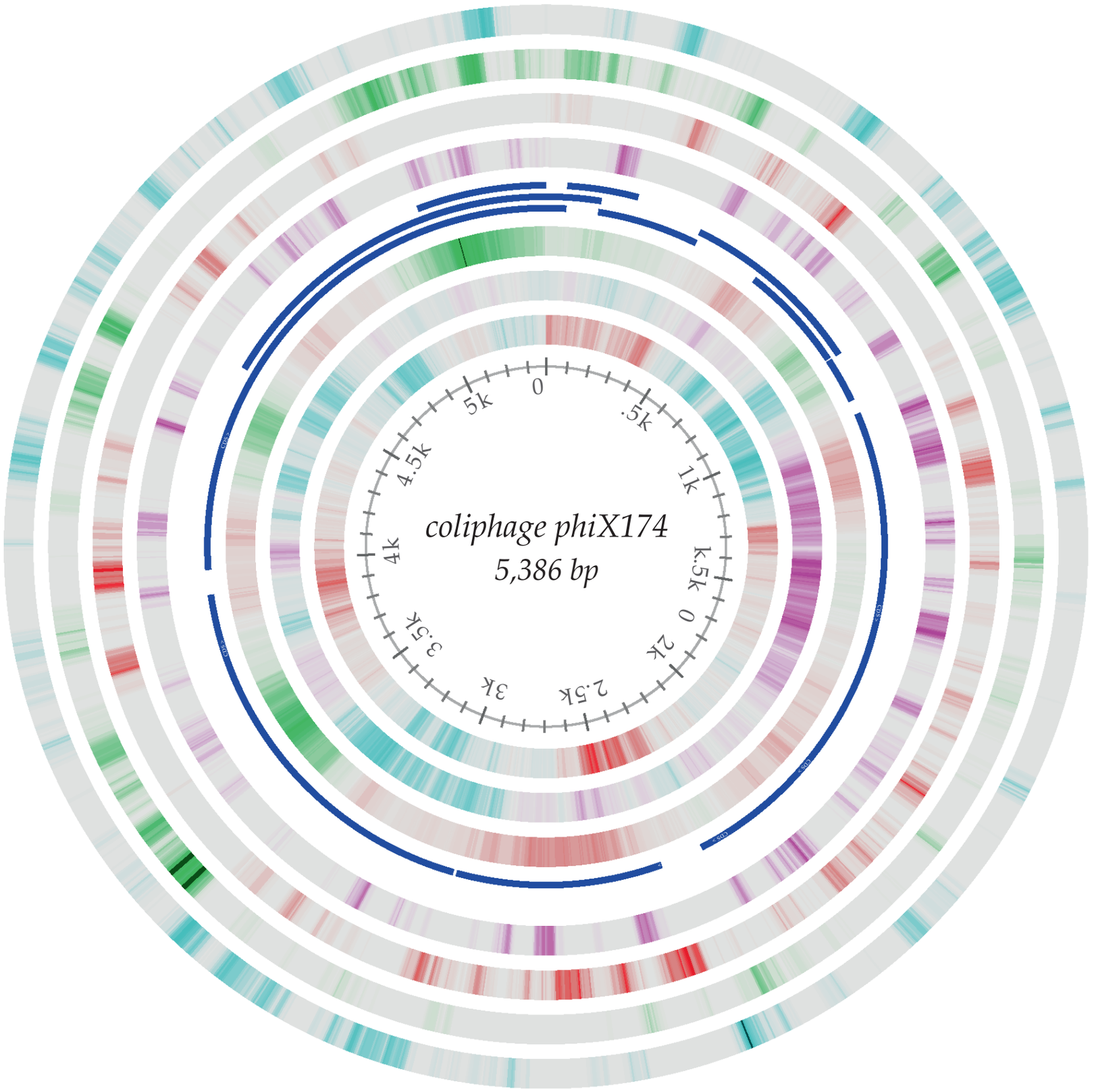
AT Skew
–0.20 fix avg 0.20

GC Skew
–0.20 fix avg 0.20

Percent AT
0.40 fix avg 0.60

Resolution: 3

BASE ATLAS



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

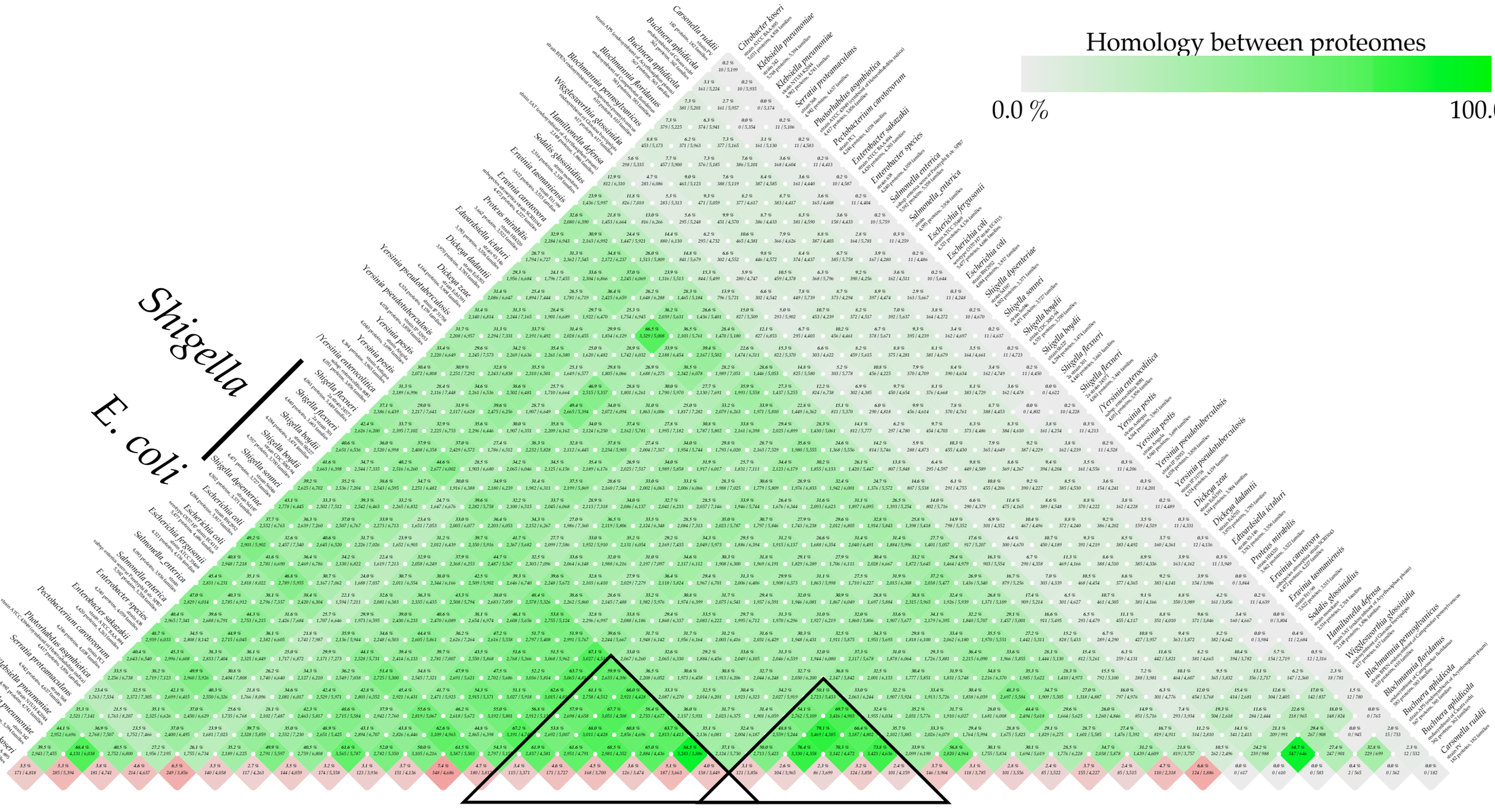
dev avg

0.47 0.63

Resolution: 3

BASE ATLAS

40 *enteric* genomes



Conclusion (part 2):

People are different!

Alu repeats - 1 in 21 births => 6.8 billion people/21 = 323 MILLION variants!

L1 repeats - 1 in 180 births => 6.8 billion people/180 = 37 MILLION variants!

SCA repeats - 1 in 916 births => 6.8 billion people/916 = 7 MILLION variants!

Bacteria are incredibly diverse!