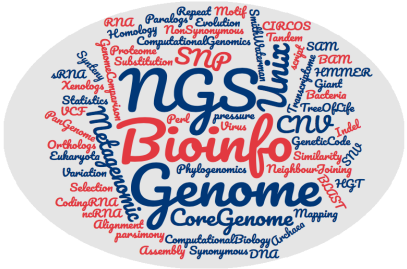
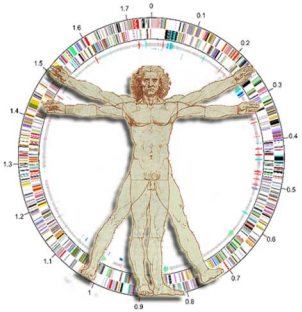


# The rise of Genomes and Bioinformatics

**Fredj Tekaia**

tekaia@pasteur.fr

**Bioinformatics and Genome Analyses**

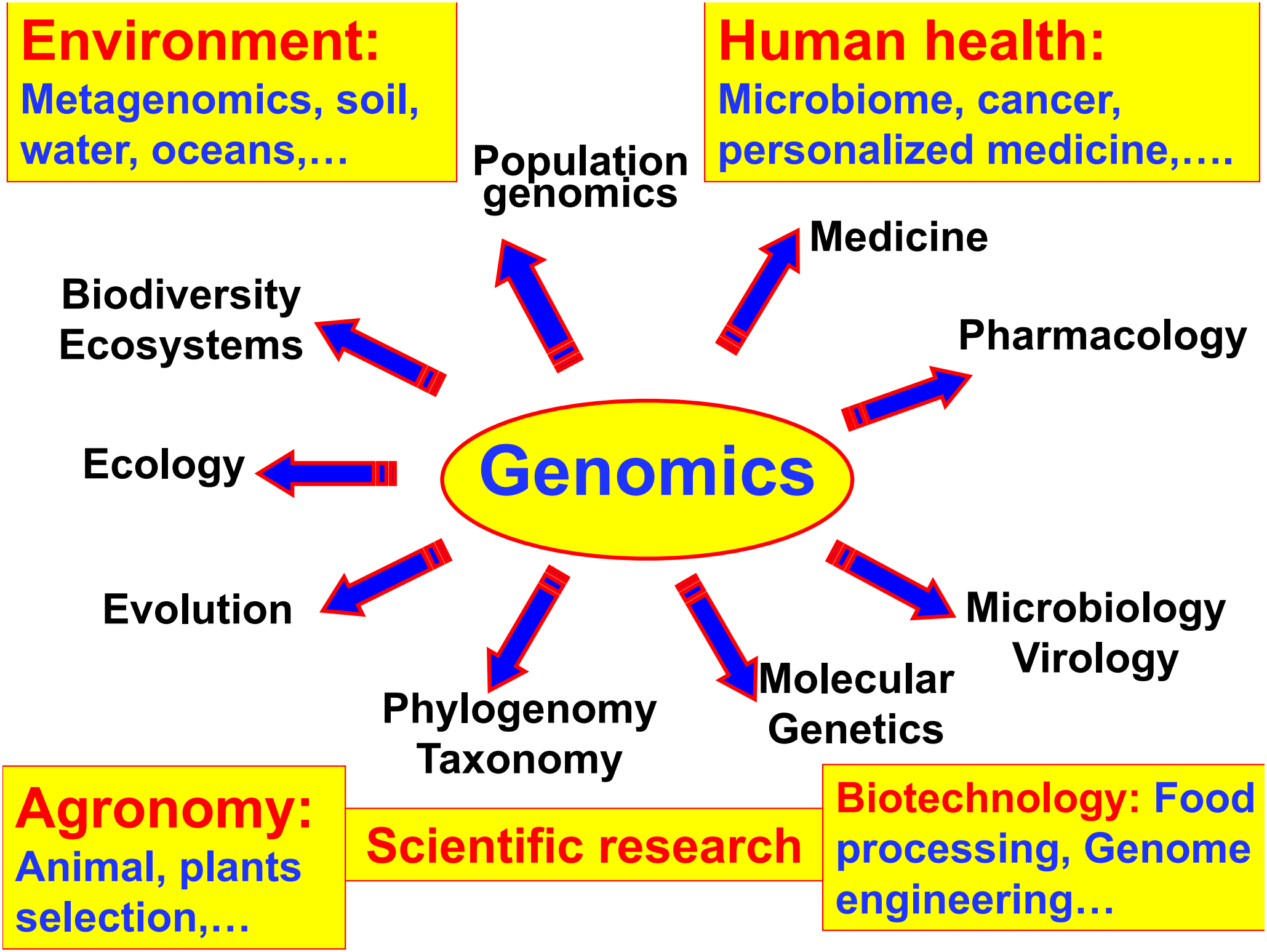
**Institut Pasteur Tunis, Tunisia. September 18 – December 15, 2017.**

# Plan

- Main genomes sequencing projects
- Bioinformatics in the Genome era
- Large-scale genome analyses/comparisons

(duplication, conservation, Orthologs inference, Genome trees, Phylogenomics, Horizontal Gene Transfer,…)

- Next Generation Sequencing & Applications - Metagenomics
- Synthetic Biology
- Concluding notes


Environment:
Metagenomics, soil, water, oceans,…
Human health:
Microbiome, cancer, personalized medicine,….
Population genomics
Medicine
Biodiversity
Ecosystems
Pharmacology
Ecology
Genomics
Evolution
Microbiology
Virology
Phylogenomy
Taxonomy
Molecular
Genetics
Agronomy:
Animal, plants selection,…
Scientific research
Biotechnology: Food processing, Genome engineering…

# The first cellular genomes

1995 → First cellular genome, the bacterium *Haemophilus influenzae*, 1.8 Mb, 1713 genes

**Starting year of Genomics**

1996 → First eukaryotic genome : *Saccharomyces cerevisiae;* 13 Mb, ~6000 genes.

1998 → First metazoan. Nematode *Caenorhabditis elegans*; 97 Mb; ~19,000 genes.

2000 → Fruit fly *Drosophila melanogaster* (137 Mb; ~13,000 genes)

2000 → First plant genome: *Arabidopsis thaliana* (115 Mb; 22670 genes)

2001 → First draft of the human genome (~3000 Mb; 200000 gaps, 28000 genes!).

2004 → First Reference of the human genome (300 gaps)

2010 → First whole-genome sequence of an ancient human (dating 4000 years from Greenland: Inuit land).

# Recently:

- **Gordon D, et al. (2016). Long-read sequence assembly of the gorilla genome. Sciences. 352 (6281). 53-57.**

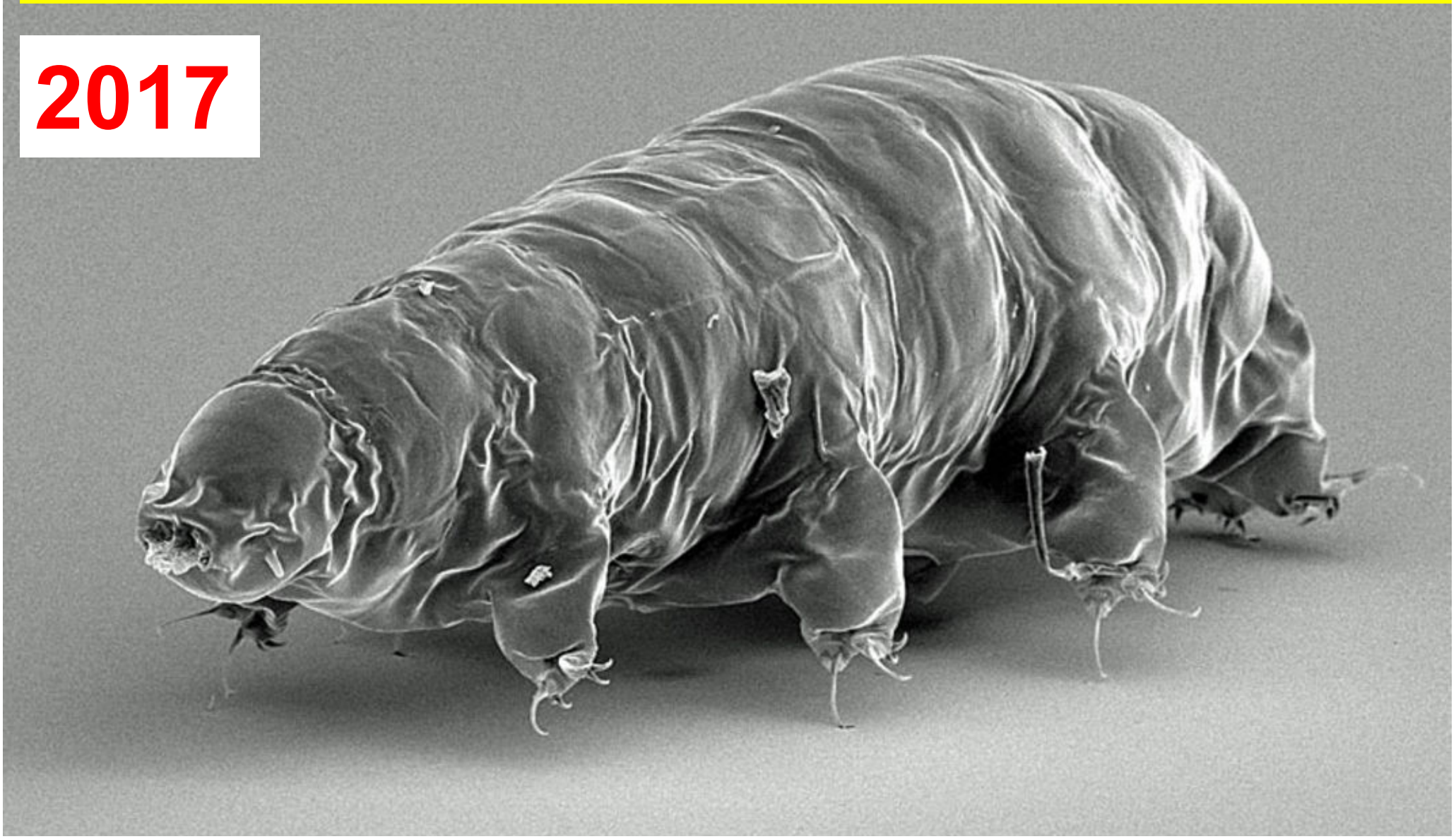
**High-quality assembly of the gorilla genome using single-molecule real-time (SMRT) sequence technology and a string graph de novo assembly algorithm.**

- **Seo Jeong-Sun et al. (2016). Nature doi: 10.1038/nature20098.**

**De novo assembly and phasing of a Korean human genome.**

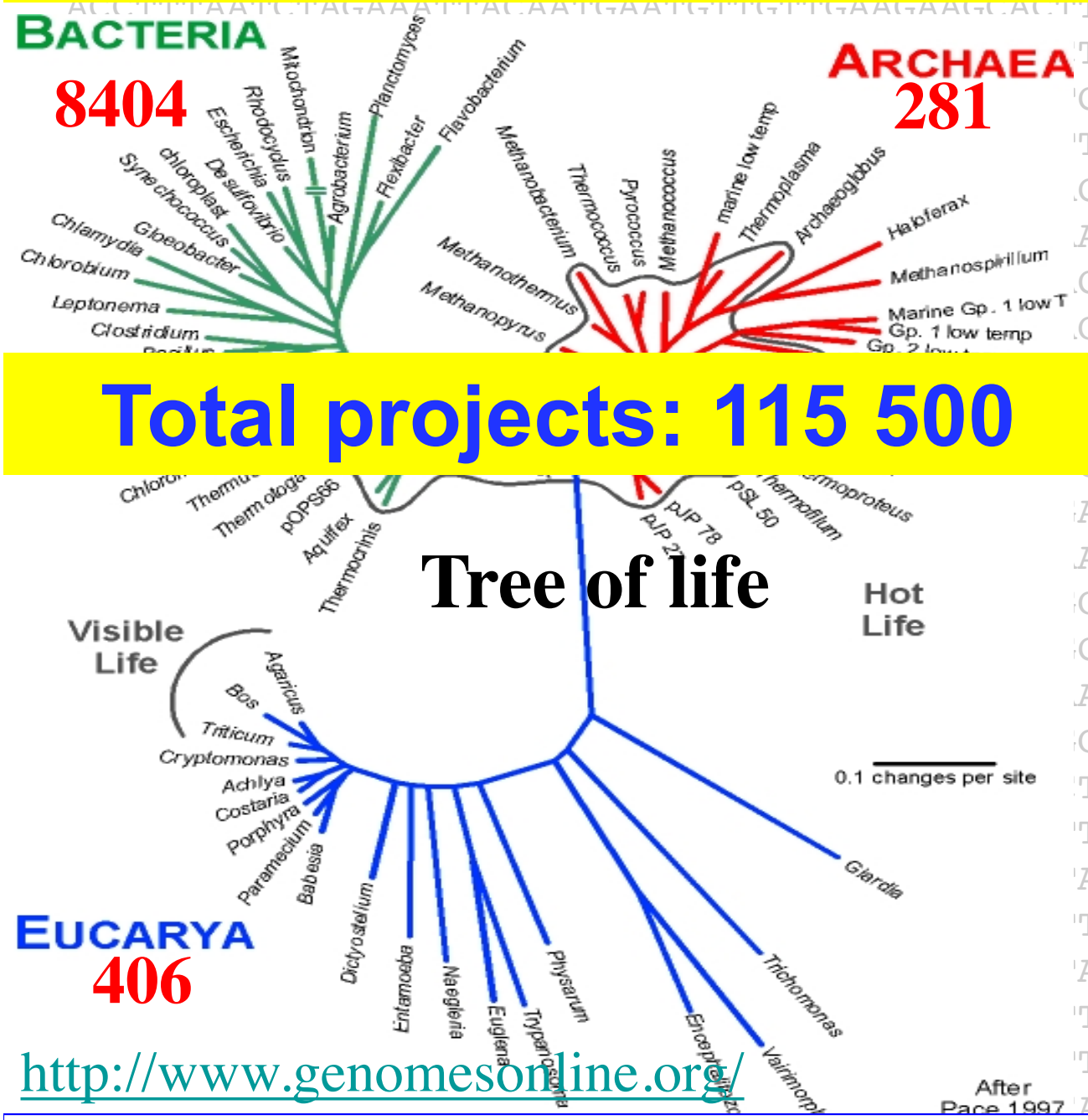
# Tardigrades: supreme resistant!

**2017**



Yoshida et al. 2017. Comparative genomics of the tardigrades *Hypsibius dujardini*and *Ramazzottius varieornatus.* PLoS Biol. 2017 Jul 27;15(7): e2002266. https://doi.org/10.1371/journal.pbio.2002266

# Today Genomes: 22/11/2017




## Total projects: 115 500

http://www.genomesonline.org/

**Viruses:** • Completed: 3504 • Permanent draft: 5008

**Complete sequenced genomes: 9091**

- 8404 Bacteria
- 281 Archaea
- 406 Eukaryotes

**Incomplete genomes: 15525**

- 10740 Bacteria
- 292 Archaea
- 4493 Eukaryotes

**Permanent Draft genomes: 75661**

- 70632 Bacteria
- 781 Archaea
- 4248 Eukaryotes

**Transciptomes: 75/15171**

- 51/1026 Bacteria
- 0/160 Archaea
- 24/13985 Eukaryota

# Many large genome projects

- **Human 1000 genomes project (2500 individuals)**

  **A global reference for human genetic variation**

  *Nature* **526**, 68-74 (01 October 2015) doi:10.1038/nature15393

- **Human Microbiome Project (HMP)**

  **http://commonfund.nih.gov/hmp/**

**Develop research resources to enable the study of the microbial communities that live in and on our bodies and the roles they play in human health and disease.**

- **International Cancer Genome Consortium**

**(https://icgc.org/)**

**Cancer Genome Project (74 project teams) 50 tumor types**

## •Genomic England: The 100K Genomes Project

https://www.genomicsengland.co.uk/the-100000-genomes-project/

**The project will sequence 100,000 genomes from around 70,000 people. Participants are National Health Service patients with rare diseases, plus their families, and patients with cancer.**

## •100K Foodborne Pathogen Genome Project

http://100kgenome.vetmed.ucdavis.edu/index.cfm

**The 100K Pathogen Genome Project aims to address the persistent food safety concerns and create a publicly available genetic database of the most common foodborne disease causing microbes.**

- **1001 genomes project**

**Arabidopsis thaliana Genetic Variation**

- **The 3000 rice genomes project**

**GigaScience Database: http://gigadb.org/dataset/200001**

- **http://1000.fungalgenomes.org/home/**

**More than 800 fungal genomes are already released**

- **http://1000bullgenomes.com**

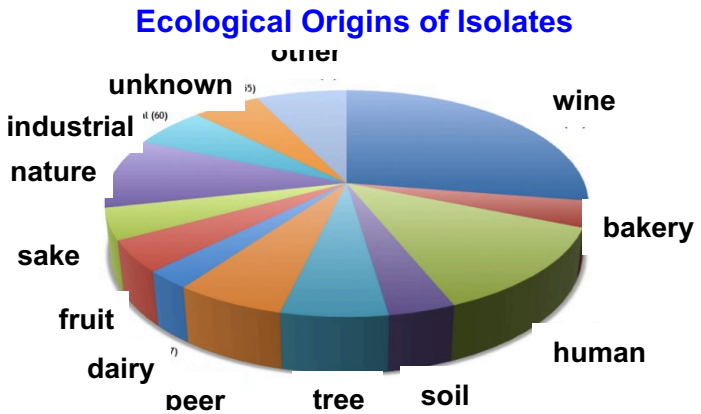
**Toward genomic selection from whole genome sequence data in dairy and beef cattle**

**Earth Microbiome project; ….**

# The 1002 Yeast Genomes Project

## http://1002genomes.u-strasbg.fr






**The goal of this project was to obtain the most comprehensive genomic data, on a single species of yeast, and to provide the most extensive view of the genetic and phenotypic diversity within this model species.**

# Genomic island

*Nature Genetics* 47, 1221 (2015) doi:10.1038/ng.3436

- **Genome of the Netherlands (GoNL)**

(*Nat. Genet.* 46, 818-825, 2014) http://www.nlgenome.nl

Whole Genome of 250 Dutch trio (2 parents +1 child)

- **Iceland** (http://www.nature.com/ng/focus/icelanders/)

Whole genomes of 2,636 Icelanders

- **The UK10K**(*Nature* 526, 82-90, 2015)

http://www.uk10k.org/

The UK10K project identifies rare variants in health and disease

- **The SardiNIA Project** (https://sardinia.irp.nia.nih.gov/)

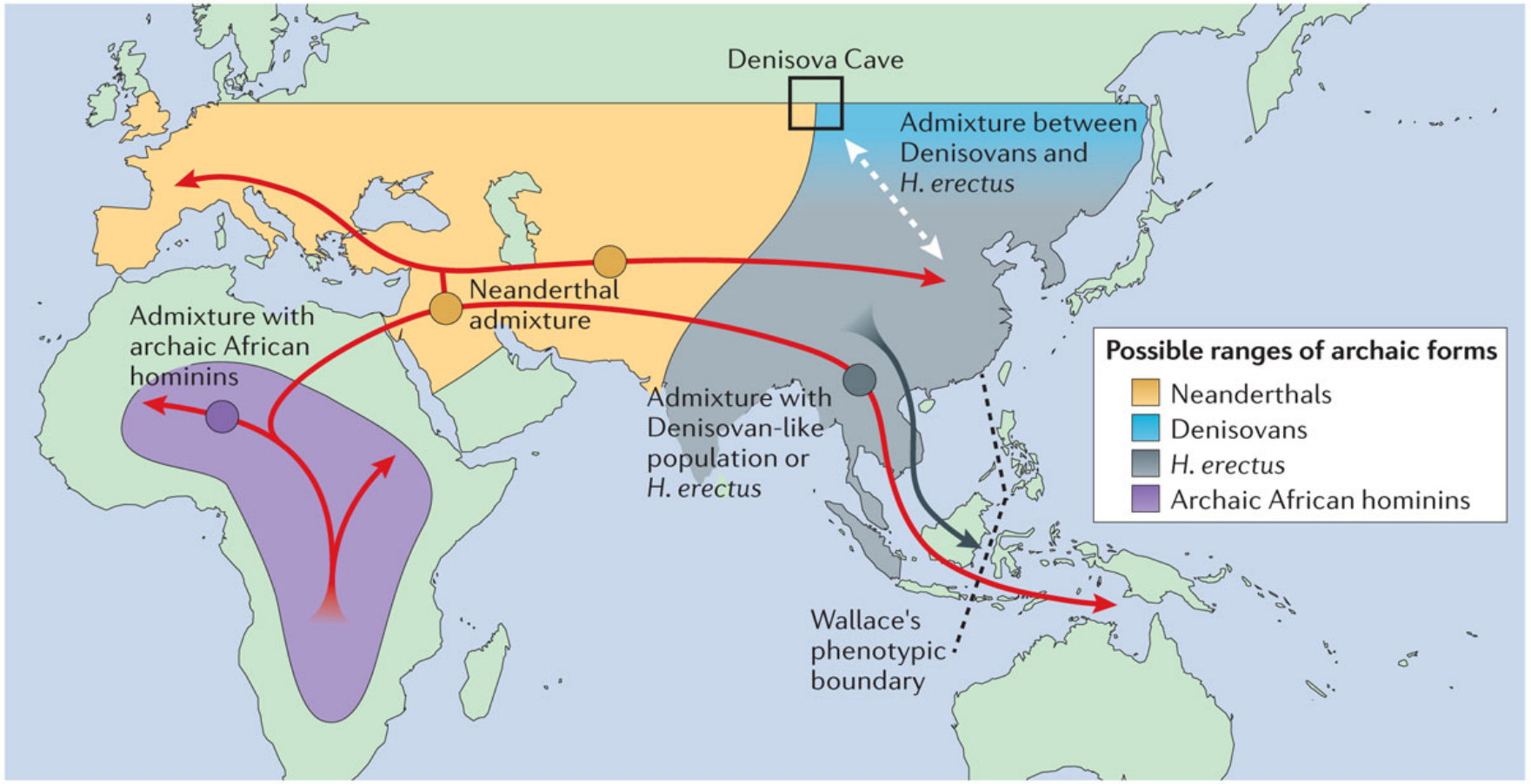
*Identify genetic bases for prominent age-associated changes*

# Ancient DNA sequencing

**Harnessing ancient genomes to study the history of human adaptation**

*Stephanie Marciniak & George H. Perry.* doi:10.1038/nrg.2017.65.
https://www.nature.com/nrg/journal/vaop/ncurrent/pdf/nrg.2017.65.pdf

# The impact of whole-genome sequencing on the reconstruction of human population history




*Krishna R. Veeramah & Michael F. Hammer (2014).* **The impact of whole-genome sequencing on the reconstruction of human population history.** ***Nature Reviews Genetics*** **15, 149-162 (2014)**

# Tracing the peopling of the world through genomics

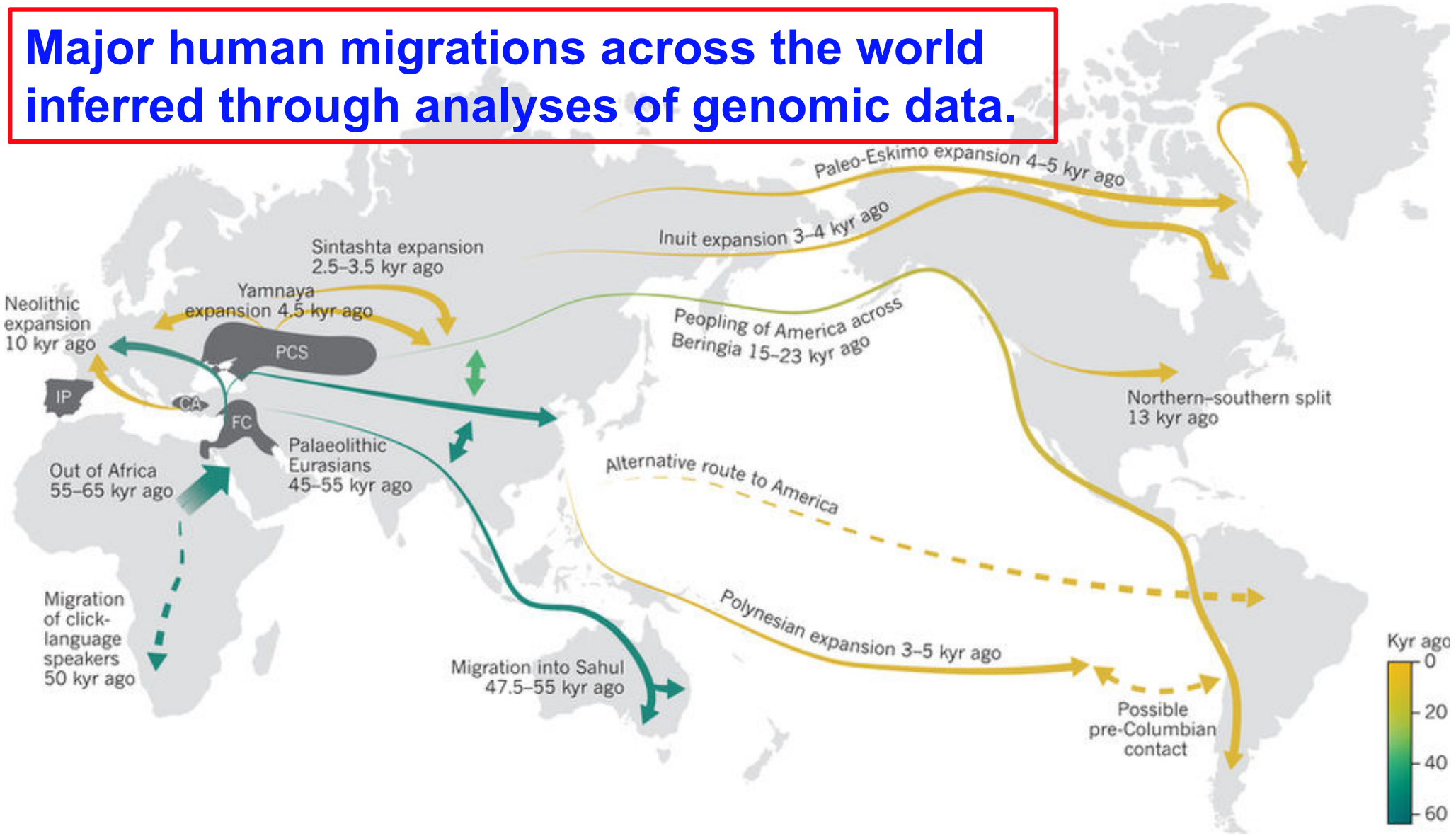
**Advances in the sequencing and the analysis of the genomes of both modern and ancient peoples have facilitated our understanding of human evolutionary history including:**

- **The discovery of interbreeding between anatomically modern humans and extinct hominins.**
- **The development of an increasingly detailed description of the complex dispersal of modern humans out of Africa and their population expansion worldwide.**
- **The characterization of many of the genetic adaptations of humans to local environmental conditions.**

Nielson R et al. (2017). Tracing the peopling of the world through genomics. *Nature* 541, 302–310.

# Tracing the peopling of the world through genomics

**Major human migrations across the world inferred through analyses of genomic data.**



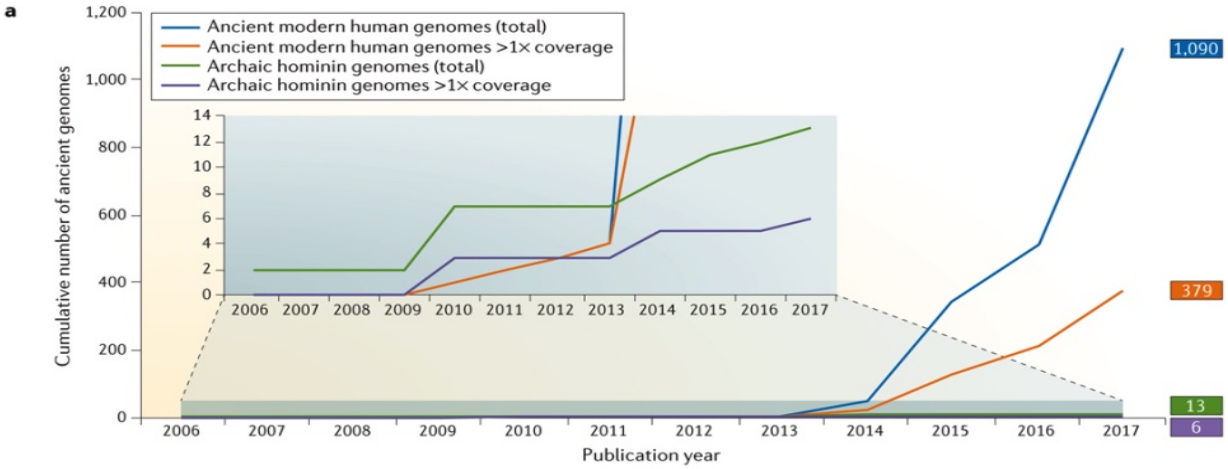

**CA: Central Anatolia; FC: Fertile Crescent; IP: Iberian Peninsula; PCS: Pontic–Caspian steppe.**

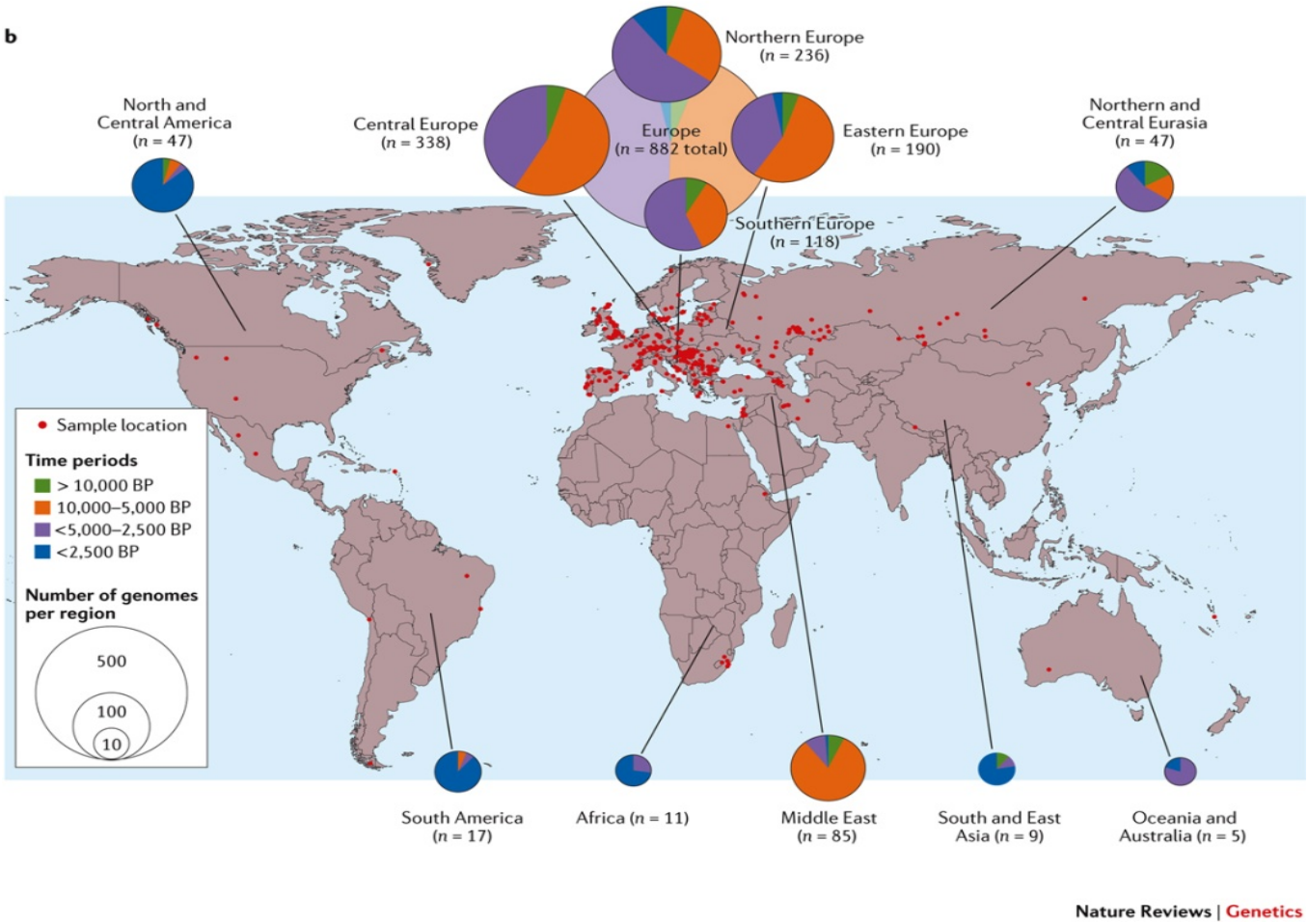
Nielson R et al. (2017). Tracing the peopling of the world through genomics. *Nature* 541, 302–310.

# The recent (and ongoing) ancient genomic explosion

**Genomic-scale ancient DNA data sets are now available for more than 1,100 ancient human and archaic hominin individuals.**

Analyses of these data **aim to identify and track the spatiotemporal trajectories of genetic variants associated with human adaptations to novel and changing environments, agricultural lifestyles, and introduced or co-evolving pathogens.**



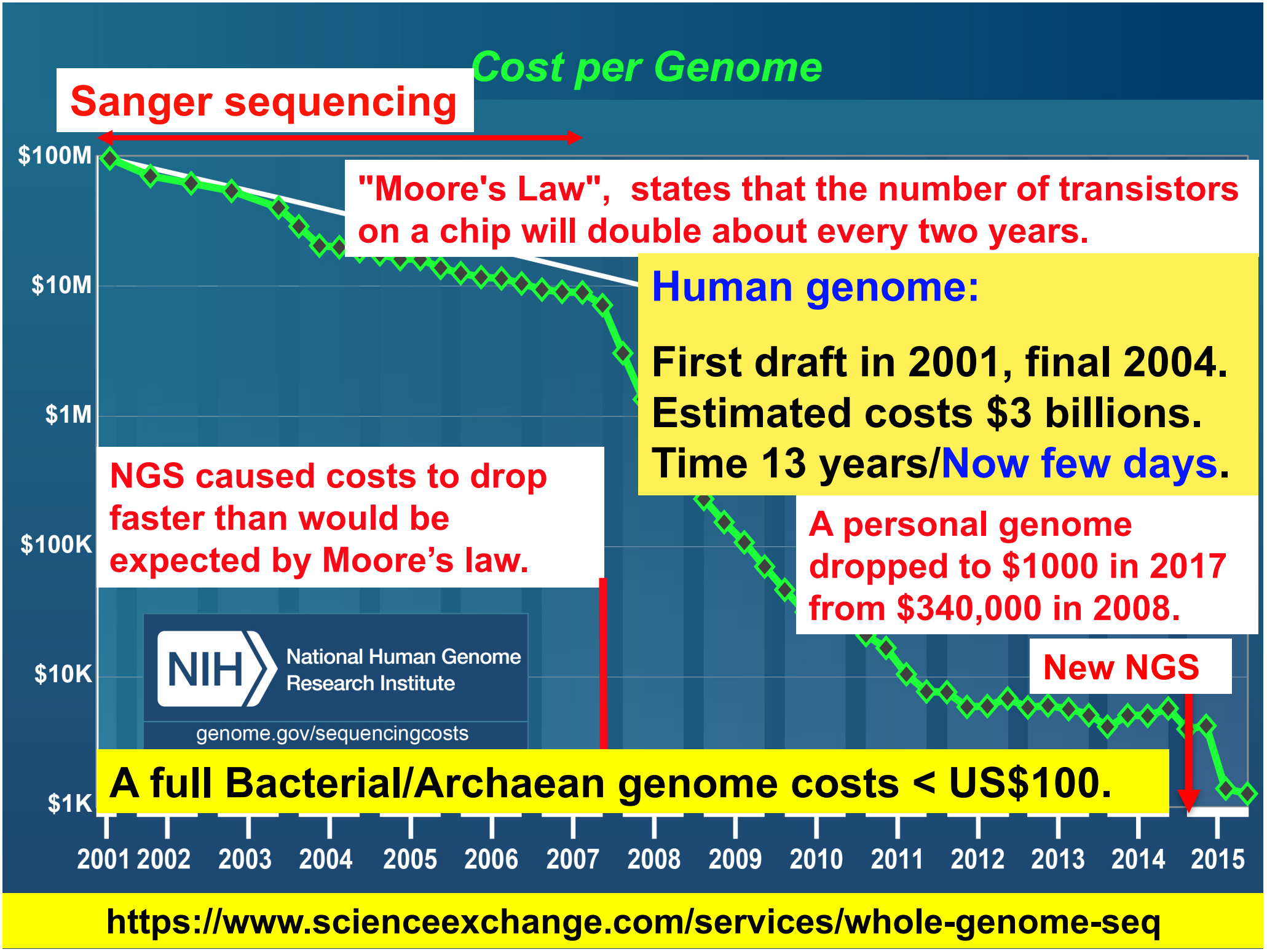




Nature Reviews | Genetics

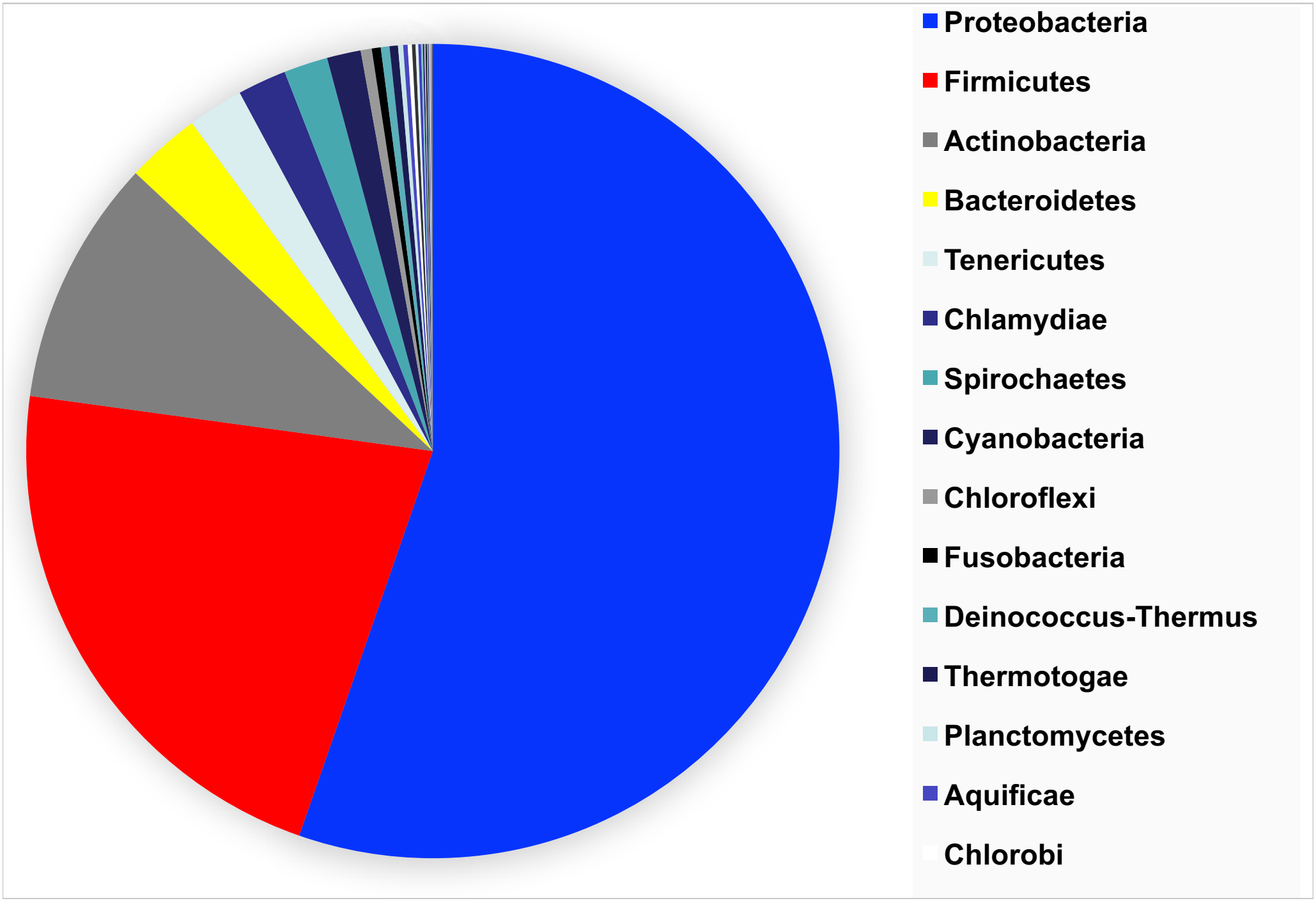
Harnessing ancient genomes to study the history of human adaptation (2017).
Stephanie Marciniak & George H. Perry Nature Reviews Genetics (2017) doi:10.1038/nrg.2017.65.

# Neandertal Genome Analysis Consortium Tracks at UCSC

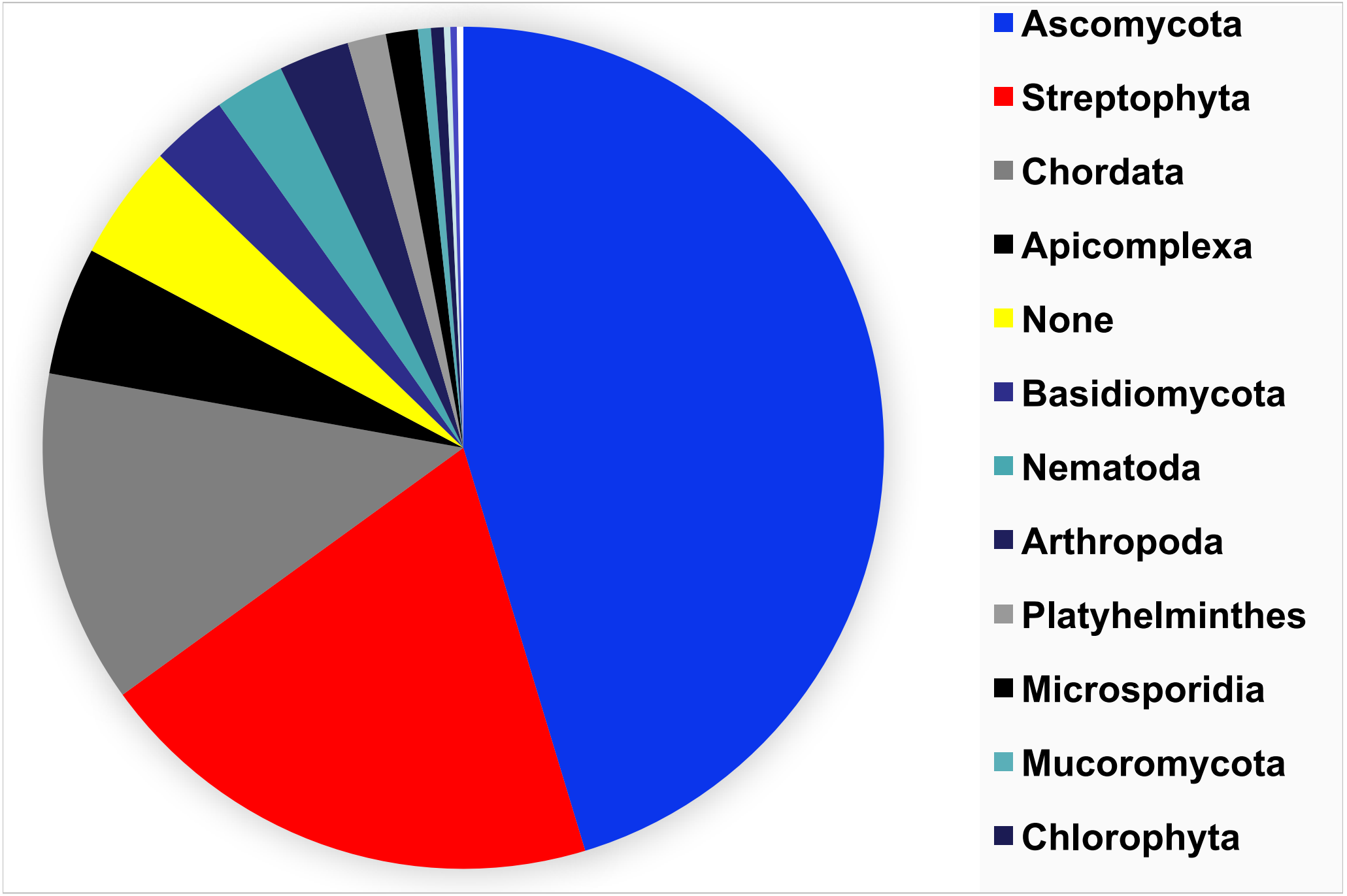
http://genome.ucsc.edu/Neandertal/


Cost per Genome
Sanger sequencing
$100M
$10M
$1M
$100K
$10K
$1K
"Moore's Law", states that the number of transistors on a chip will double about every two years.
Human genome:
First draft in 2001, final 2004.
Estimated costs $3 billions.
Time 13 years/Now few days.
NGS caused costs to drop faster than would be expected by Moore's law.
A personal genome dropped to $1000 in 2017 from $340,000 in 2008.
New NGS
NIH
National Human Genome Research Institute
genome.gov/sequencingcosts
A full Bacterial/Archaean genome costs < US$100.
2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015
https://www.scienceexchange.com/services/whole-genome-seq

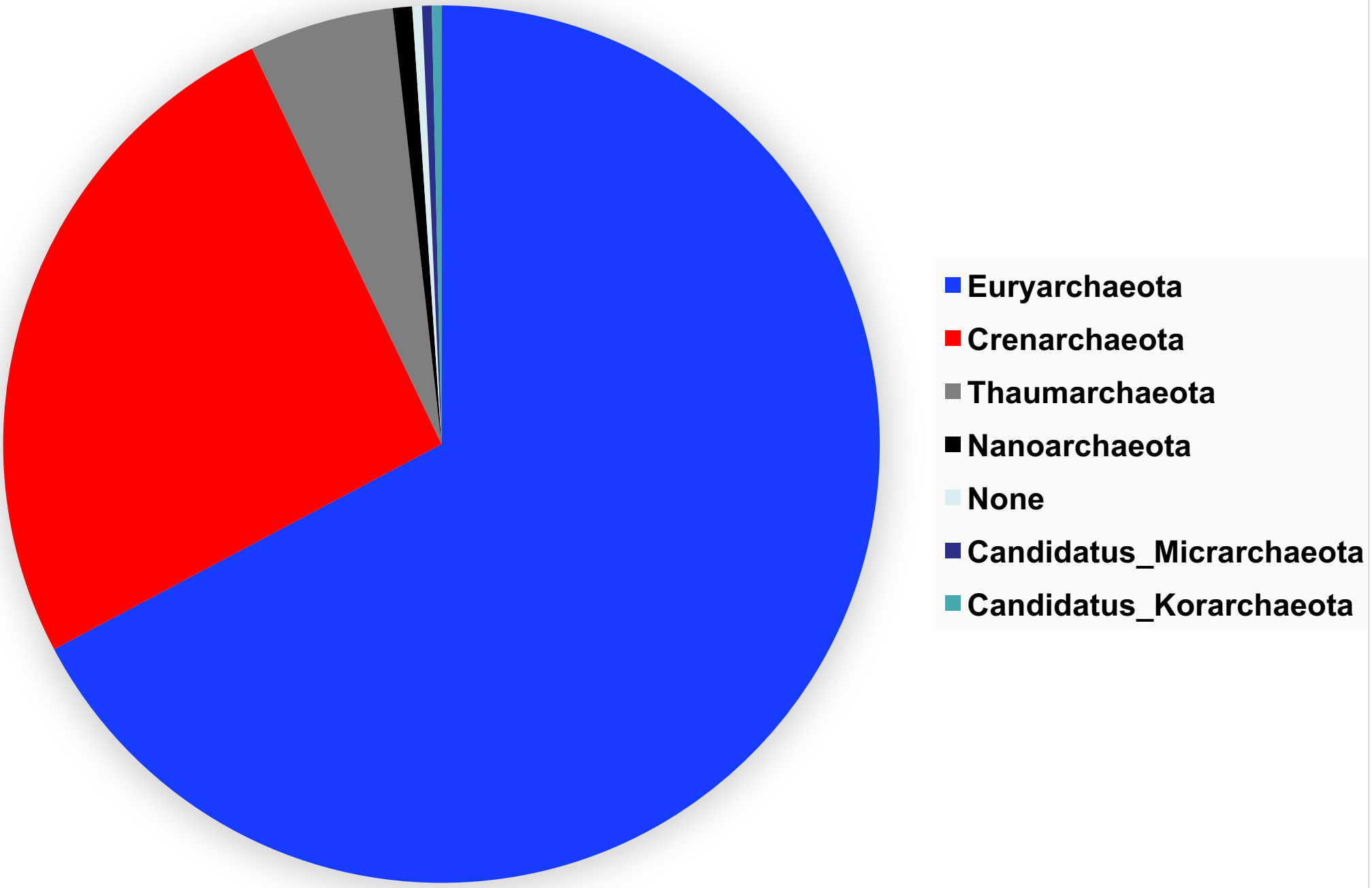
# Bacteria: Phylum distribution of 8404 completely sequenced genomes



# Eukaryotes: Phylum distribution of 406 completely sequenced genomes



# Archaea: Phylum distribution of 281 completely sequenced genomes



# How big are genomes?

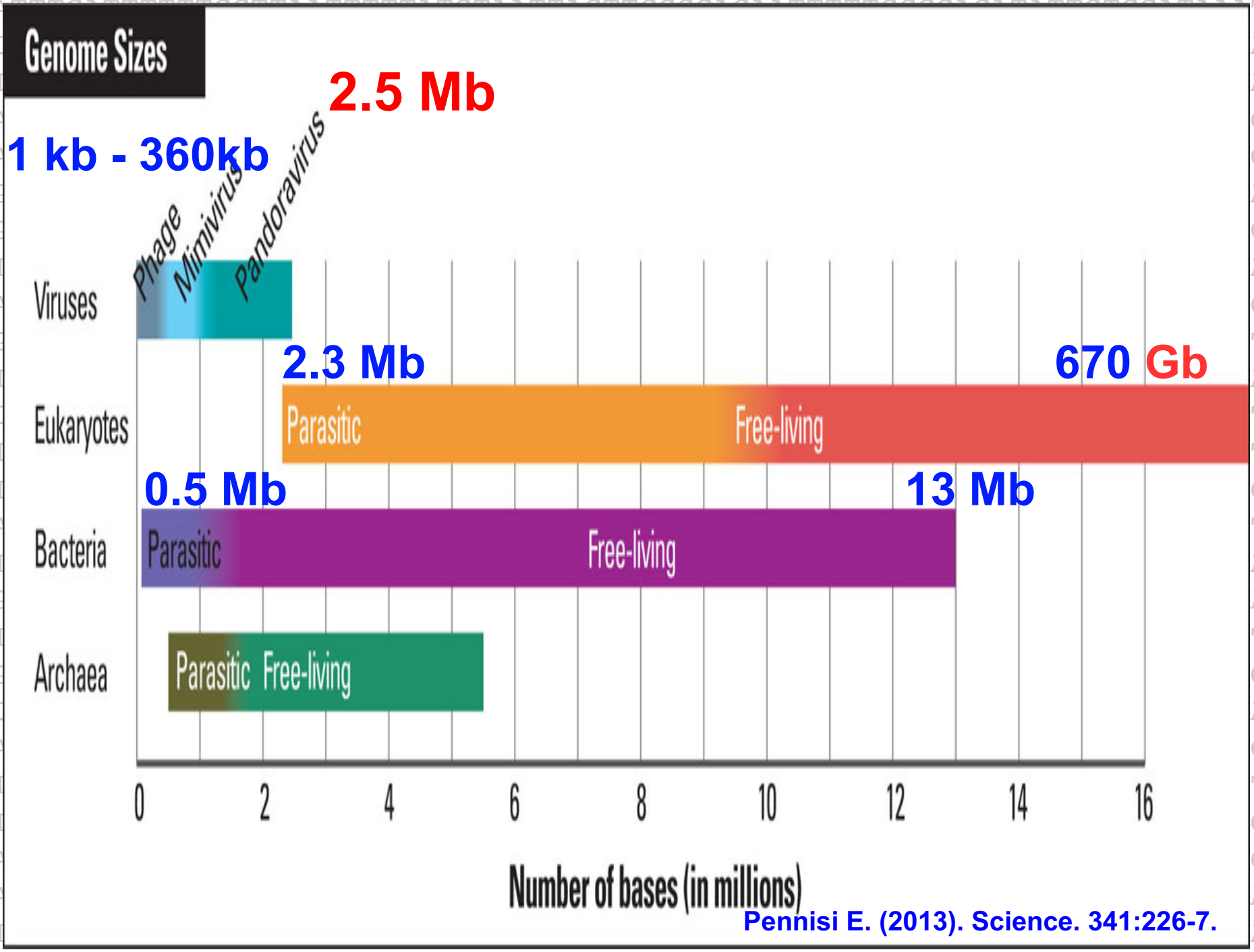
# How big are genome sizes?

- **Viral genomes: 1 kb to 360 kb (*Canarypox virus*)**

**With the discovery of giant viruses:**
**Pandarovirus: 2.5 Mb**
***Larger than Encephalitozoon intestinalis* (2.3 Mbp)**
**and similar to *Encephalitozoon cuniculi* (2.5 Mbp)**

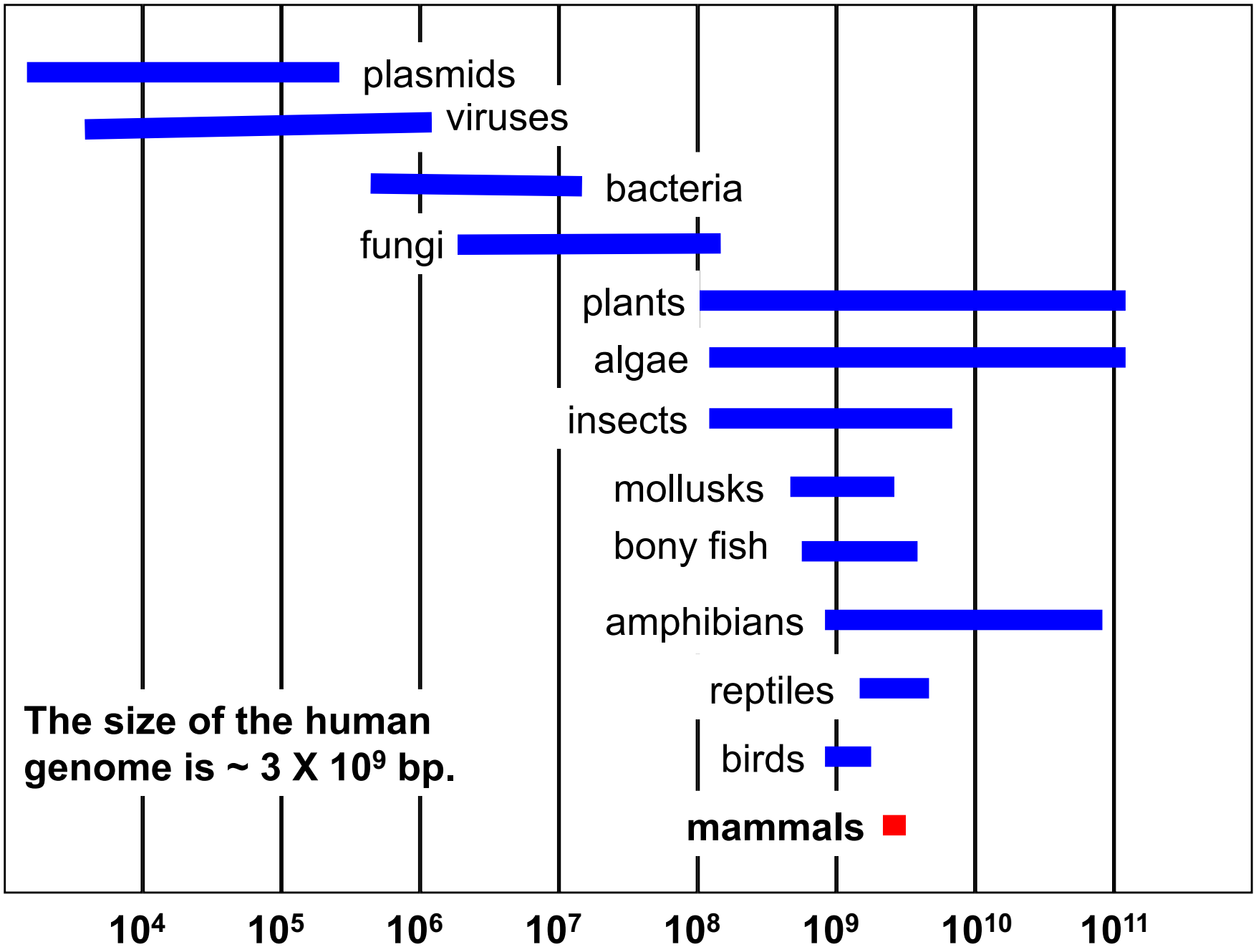
- **Bacterial genomes: 0.5 Mb to 13 Mb;**

- **Eukaryotic genomes: 2.3 Mb to 670 Gb (*Polychaos dubium* (*"Amoeba" dubia*));**



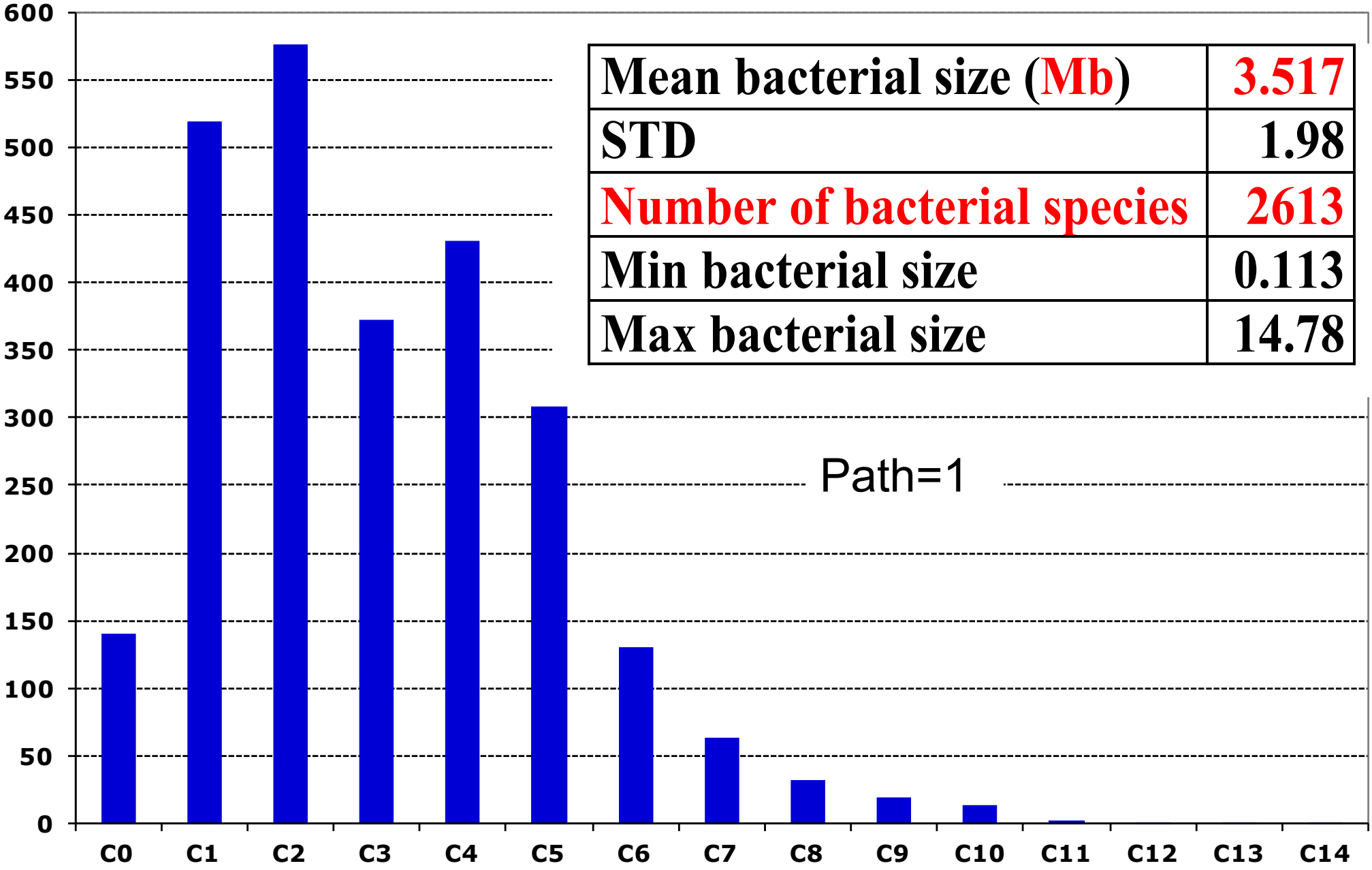

Pennisi E. (2013). Science. 341:226-7.

# Genome sizes (in base pairs)




http://www3.kumc.edu/jcalvet/PowerPoint/bioc801b.ppt

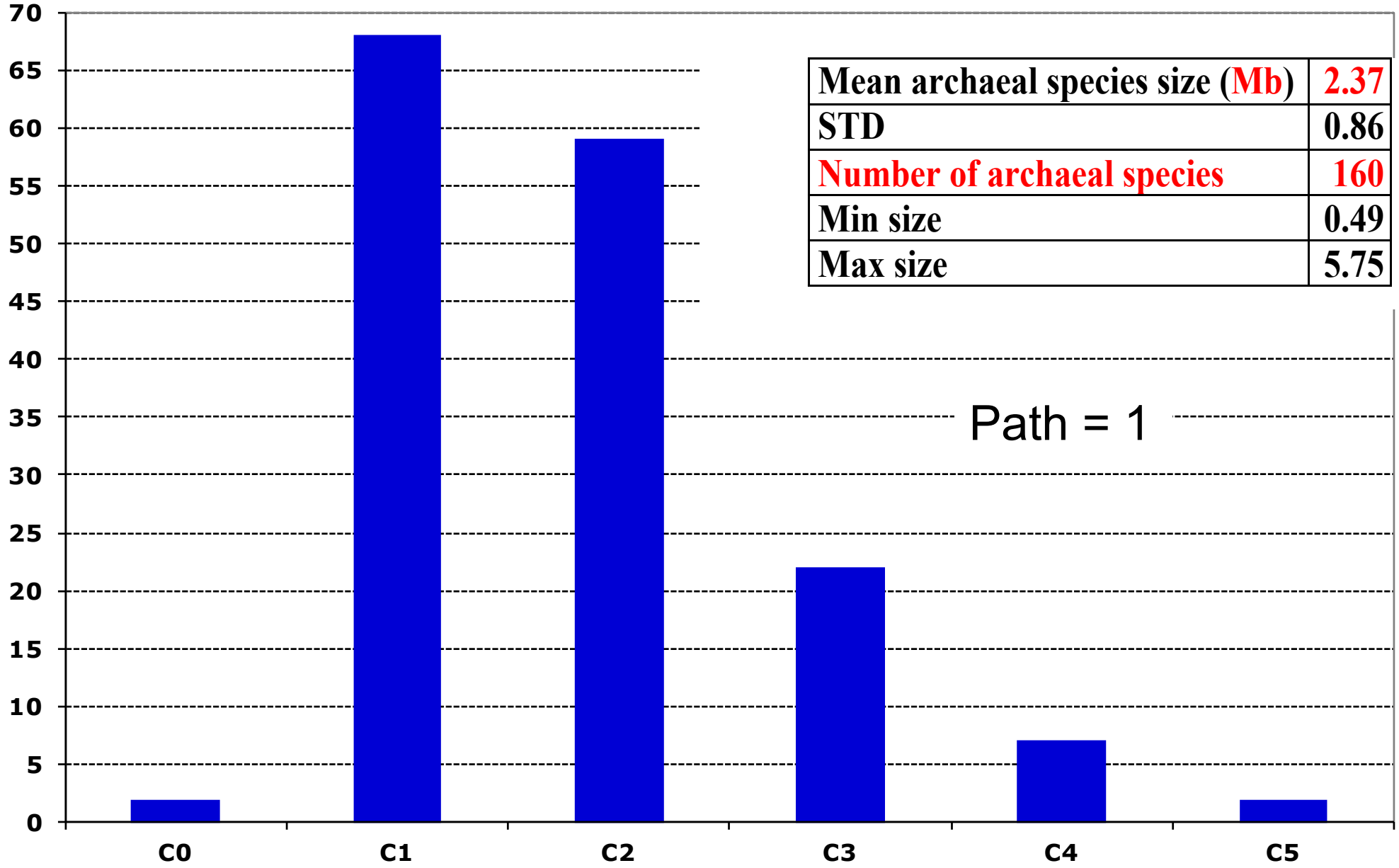
# Size(Mb) distribution of bacterial species




| | |
|---|---|
| Mean bacterial size (Mb) | 3.517 |
| STD | 1.98 |
| Number of bacterial species | 2613 |
| Min bacterial size | 0.113 |
| Max bacterial size | 14.78 |

Tekaia F. 2017. Unpublished results.

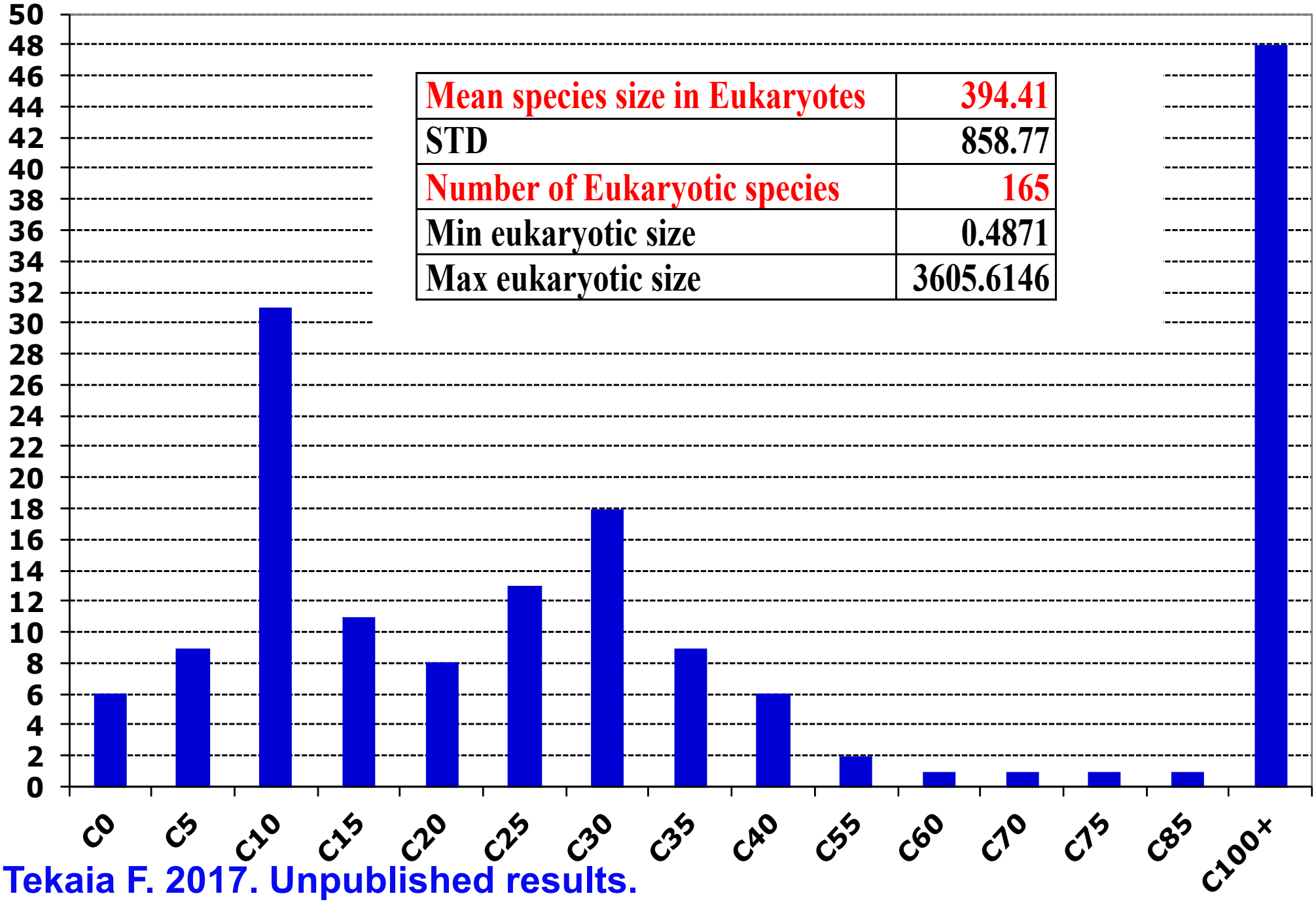
# Size distribution of archaeal species




| | |
|---|---|
| Mean archaeal species size (Mb) | 2.37 |
| STD | 0.86 |
| Number of archaeal species | 160 |
| Min size | 0.49 |
| Max size | 5.75 |

Tekaia F. 2017. Unpublished results.

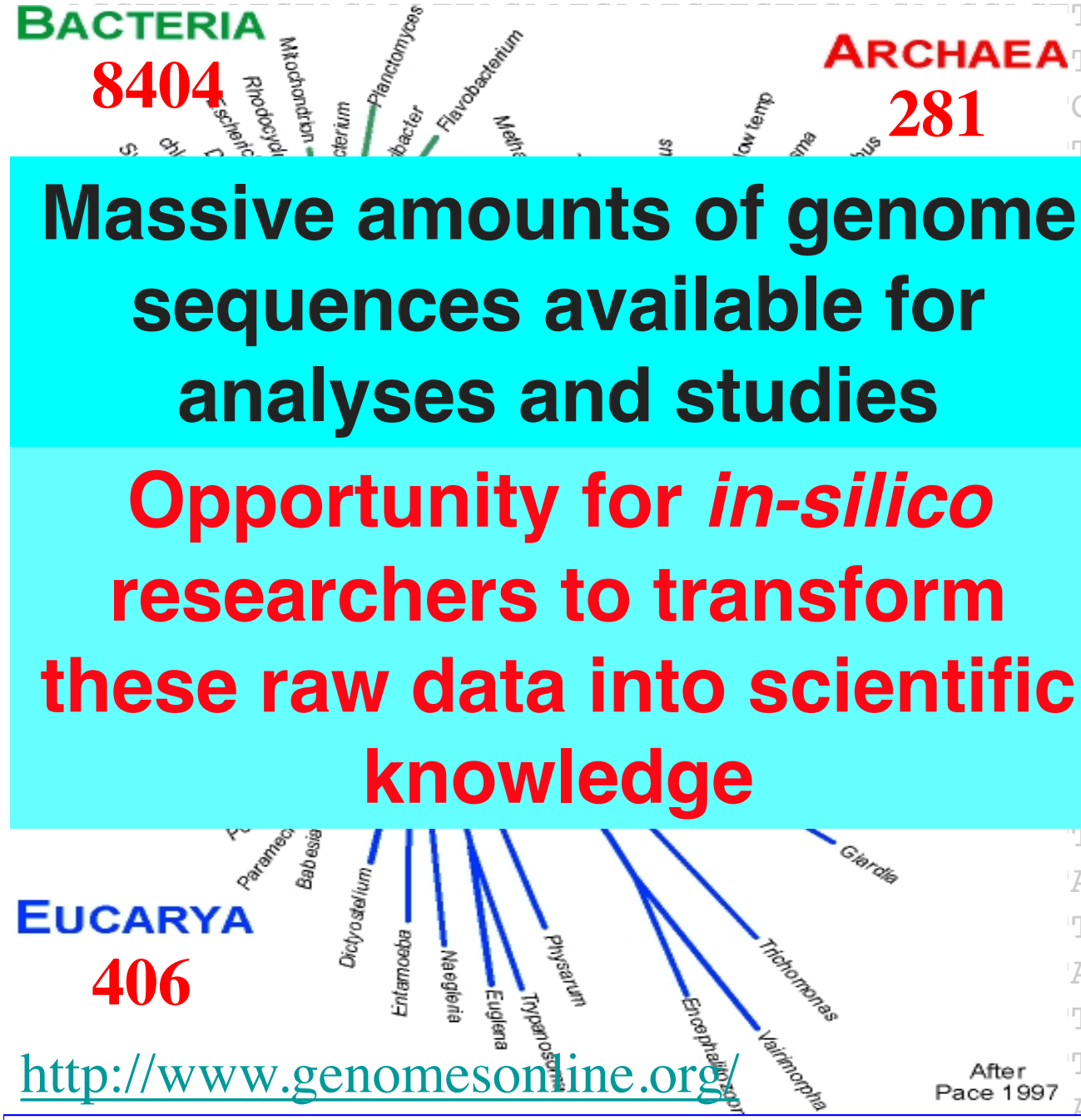
# Size(Mb) distribution of eukaryotic species




| | |
|---|---|
| Mean species size in Eukaryotes | 394.41 |
| STD | 858.77 |
| Number of Eukaryotic species | 165 |
| Min eukaryotic size | 0.4871 |
| Max eukaryotic size | 3605.6146 |

Tekaia F. 2017. Unpublished results.

22/11/2017




# Massive amounts of genome sequences available for analyses and studies

## Opportunity for *in-silico* researchers to transform these raw data into scientific knowledge

http://www.genomesonline.org/

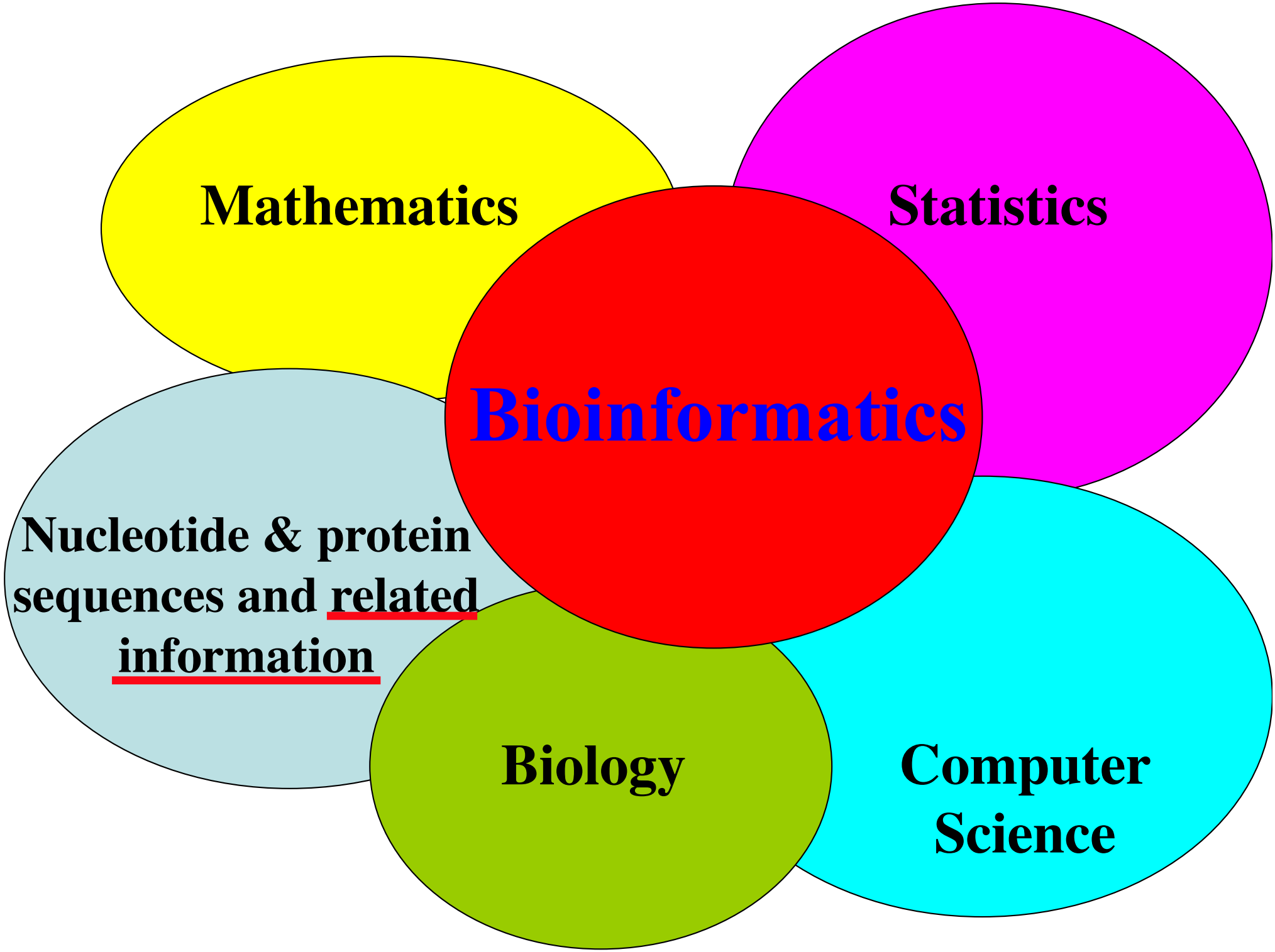
**Complete sequenced genomes: 9091**
- 8404 Bacteria
- 281 Archaea
- 406 Eukaryotes

**Incomplete genomes: 15525**
- 10740 Bacteria
- 292 Archaea
- 4493 Eukaryotes

**Permanent Draft genomes: 75661**
- 70632 Bacteria
- 781 Archaea
- 4248 Eukaryotes

**Transciptomes: 75/15171**
- 51/1026 Bacteria
- 0/160 Archaea
- 24/13985 Eukaryota

**Viruses:** • Completed: 3504 • Permanent draft: 5008


Mathematics
Statistics
Bioinformatics
Nucleotide & protein sequences and related information
Biology
Computer Science

- **Bioinformatics is expected to help in understanding biological processes, by developing and applying computational methods and techniques so that to answer biological questions.**

**Bioinformatics has now firmly established itself as a scientific discipline.**

# Boinformatics: Main directions

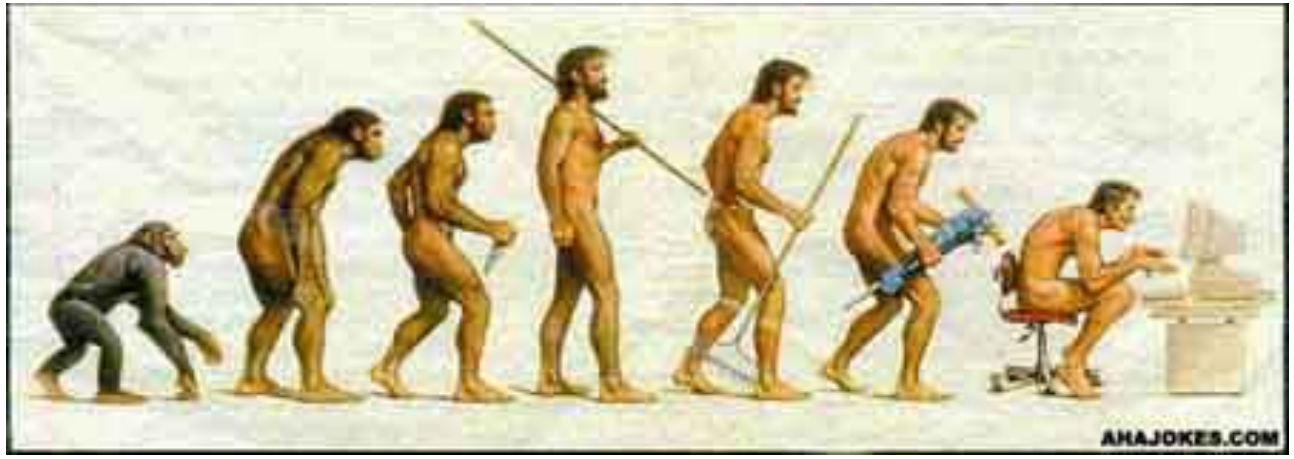
## Specific methods for:

- Genome Assembly and Annotation
- Genome Evolution
- Genomic/Genetic Variation
- Structural Genetics/Genomics
- Next Generation Data Analyses
- Meta-genomic Analyses

# Bioinformatics Before the genome era - at the sequence level

**Main topics include:**

- **Search for similarity**
- **Multiple alignments**
- **Motifs finding**
- **Phylogeny - Evolution**

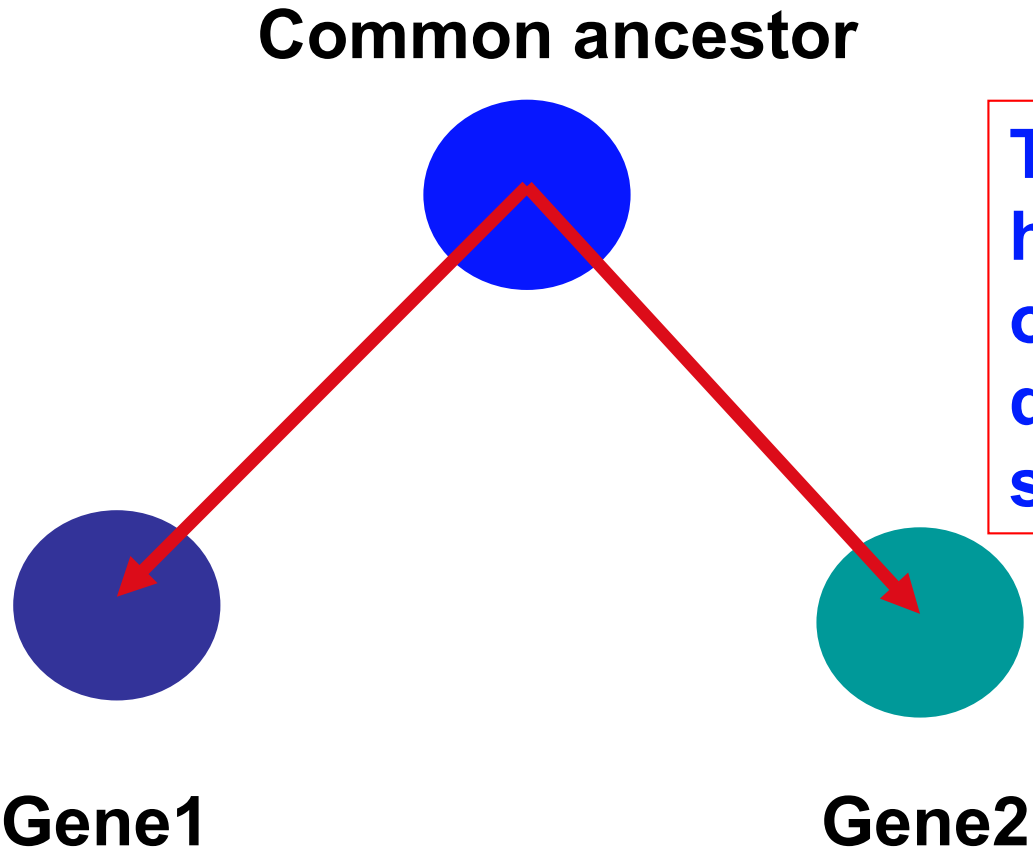



"Nothing in Biology makes sense except in the light of evolution"

Christian Theodosius Dobzhansky, 1973.

# Search for similarity

**One of the most frequent activity in Bioinformatics**




Two genes are homologs if and only if they derive from the same ancestor

**Homology is inferred by sequence similarity**

# Sequence similarity search

http://blast.ncbi.nlm.nih.gov/Blast.cgi

## BLAST

(Basic Local Alignment Search Tool)

### Nucleotide BLAST

- Nucleotide query - nucleotide database [blastn]

### Protein BLAST

- Protein query - protein database [blastp]
- PSI-BLAST Position Specific Iterative BLAST

### Translated BLAST Searches

- Nucleotide query - Protein db [blastx]
- Protein query - Translated db [tblastn]
- Nucleotide query - Translated db [tblastx]

### Seach for conserved domains

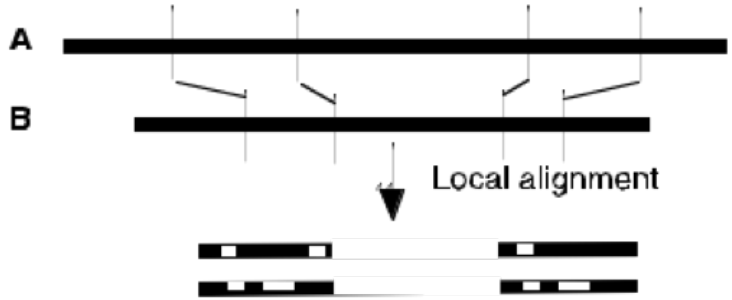
- Search the Conserved Domain Database [RPS-BLAST]

### Pairwise BLAST

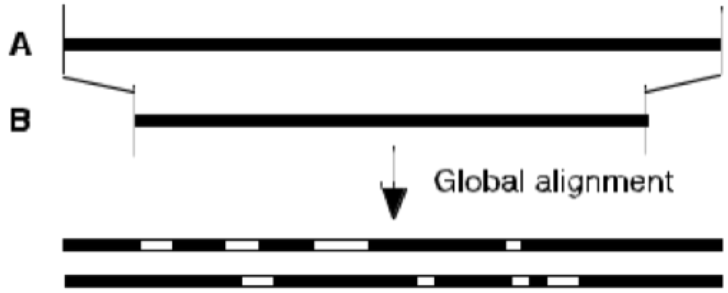
- BLAST 2 Sequences

# Alignments

## Local Alignment




## Global Alignment




- Clustal (w, o..)
- Multalign
- T-coffee
- MUSCLE
- MSA
- Probcons
- MAFFT
- DIALIGN
- Kalign

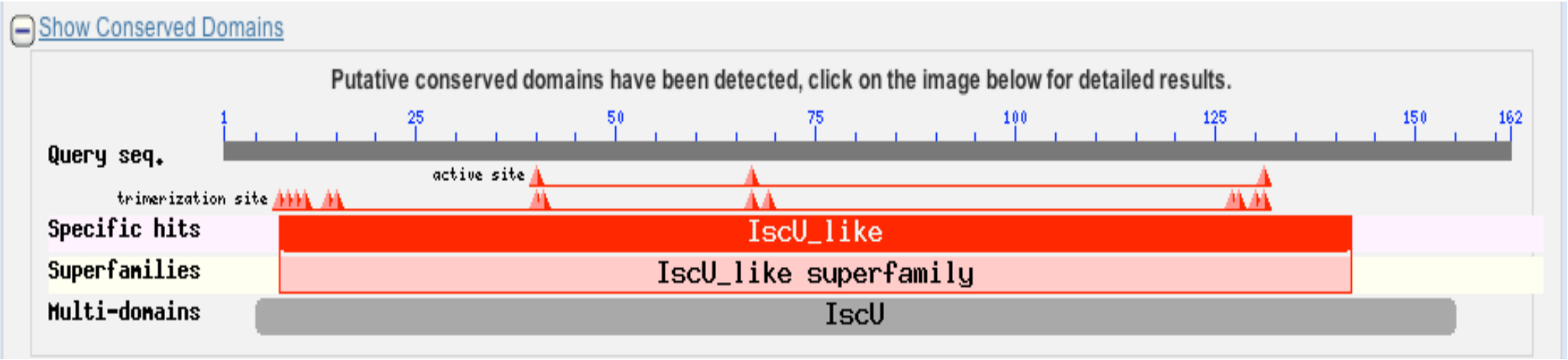
http://www.expasy.org/genomics/sequence_alignment

# Motifs finding and Signals

- **Is there any known motif (domain) in my sequence?**

**rpsblast versus Common Domains Database (CDD/ncbi)**

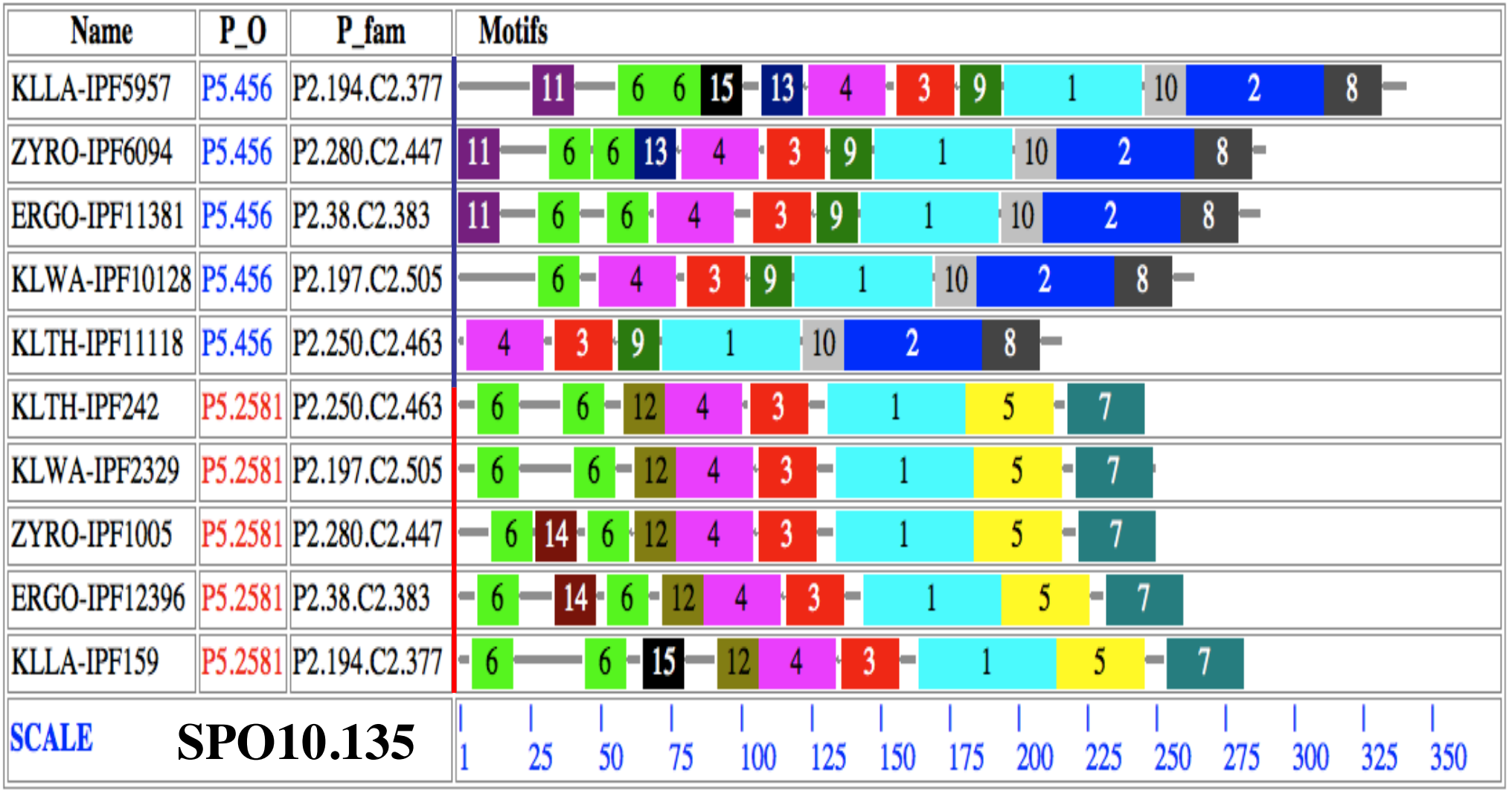
**http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi**



# Motifs finding and Signals

- MEME/MAST find commun motifs in a set of sequences

http://meme-suite.org/



# Phylogeny - Evolution


http://evolution.genetics.washington.edu/phylip/software.html

**Includes a large list of methods and programs**

# Bioinformatics in the genome era

--> Methods used at large-scale

- Genome assembly/Gene prediction resources
- Large-scale genome comparisons

-Paralogs, Orthologs inference and clustering

-Phylogenomics

- Whole genome alignments
- Relationships visualization

New tools and methods for NGS data analyses

# Genome assembly resources

**Given many (millions) of reads, produce a linear or circular genome sequence**

**Difficulties among others:**

- **Coverage**
- **Errors in reads**
- **Reads vary from short (35bp) to long (800bp or recently > 20kb*) and genomes are double-stranded**
- **Non-unique solution**
- **Running time**

*Nakano et al. Human Cell (2017) 30:149–161.

# de novo vs comparative assembly

- **De novo assembly ie assembly is done from scratch**

- **Comparative assembly ie when there is a reference genome**

➔ **Project presentation: Thursday afternoon**

# Gene prediction resources

From newly sequenced genomes, most popular:

## Eukaryotes:

- Genscan

http://genes.mit.edu/GENSCANinfo.html

- Genmark

http://opal.biology.gatech.edu/GeneMark/

## Prokaryotes:

- Glimmer

http://ccb.jhu.edu/software/glimmer/index.shtml

➔ Project presentation: Thursday afternoon

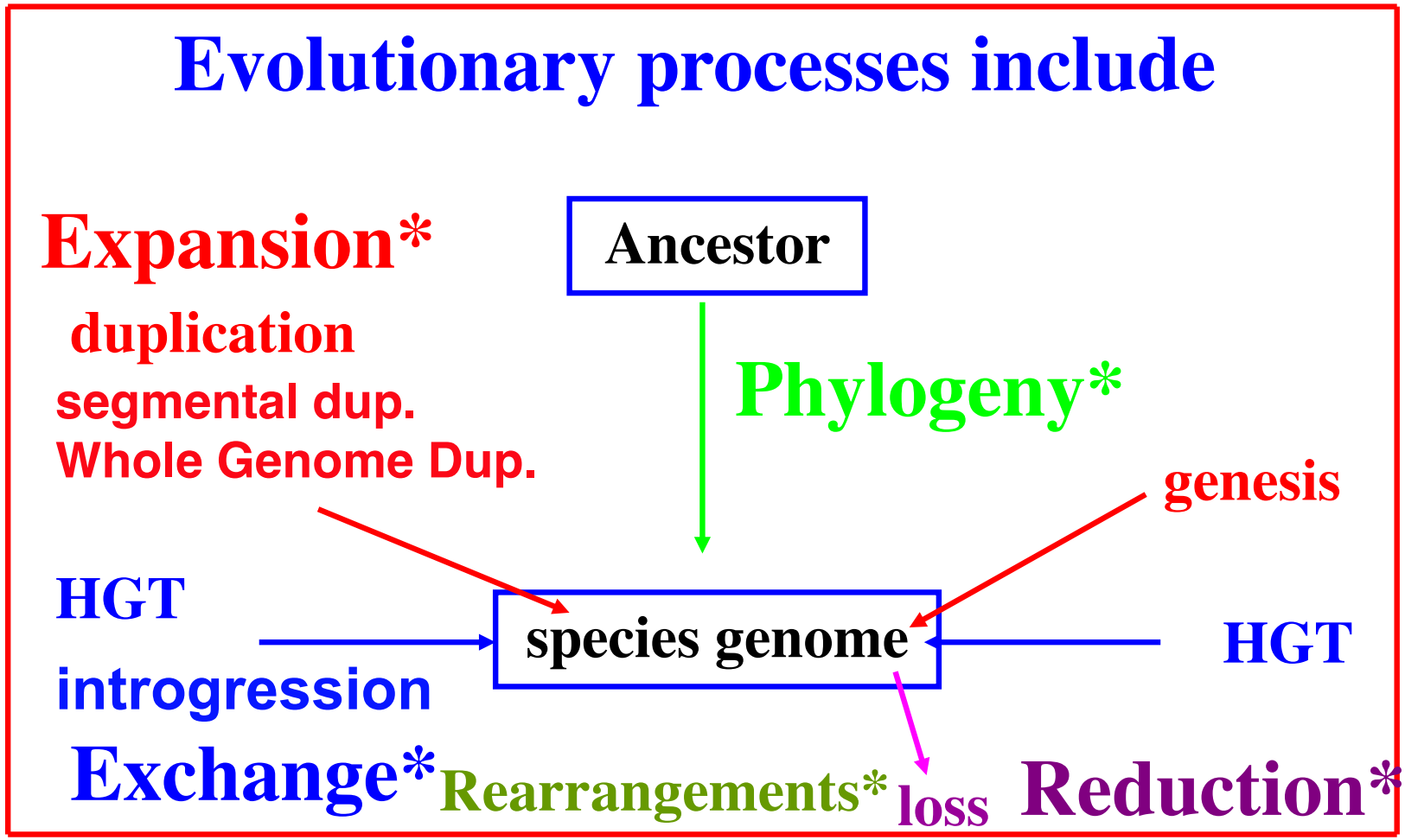
# Large-scale genome comparisons

- Duplication (genes, chromosome segments, whole genome)
- Conservation (genes, chromosomes, segments);
- Specificity (species-specific genes);
- Inferring Paralogs, orthologs and clustering;
- Shared motifs in clusters of paralogs, orthologs;
- Protein conservation profiles;
- Gene Transfer, introgression between species;

➔ Evolution

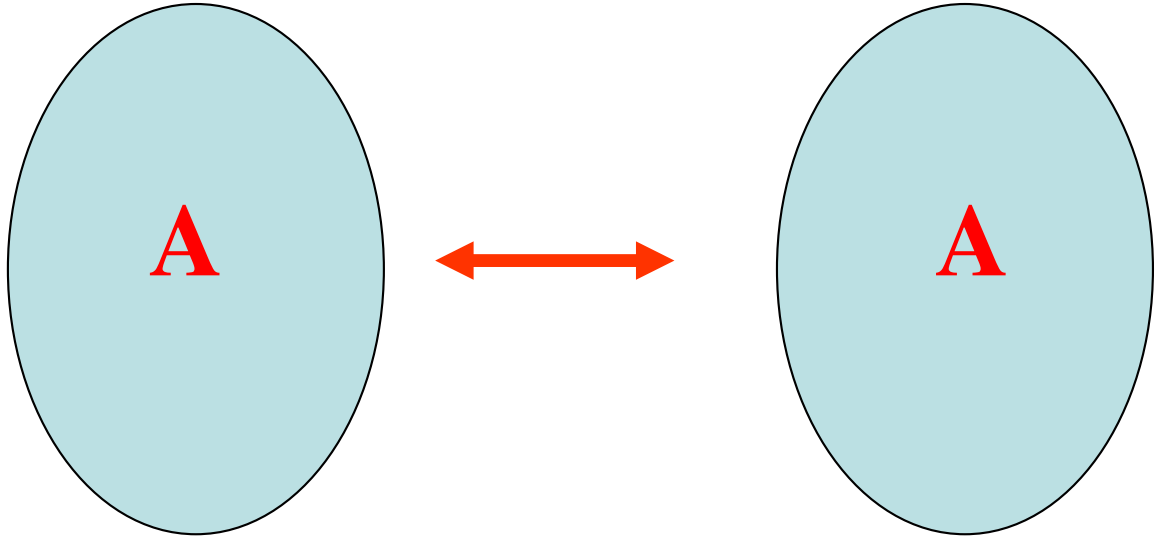
# Large-scale comparative analyses of proteomes revealed significant evolutionary processes:

**Expansion, Exchange and Reduction.**



# Comparing genomes

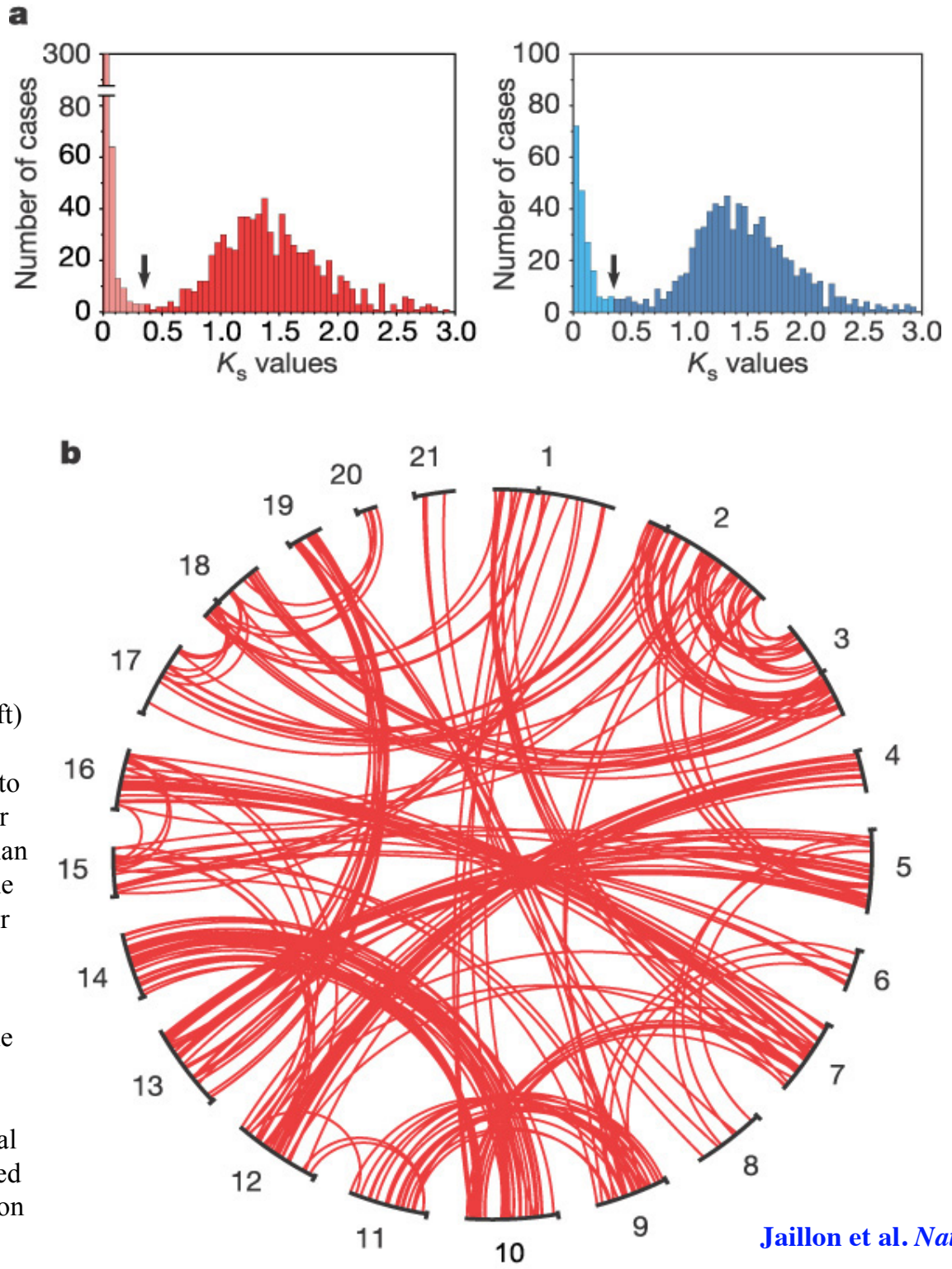
## Intra-species Comparisons




- **Rate of ancestral duplication**

**i.e non-unique genes / total number of genes in A.**

- **Families of genes - Orphans - Organisation**




**Genome duplication.**

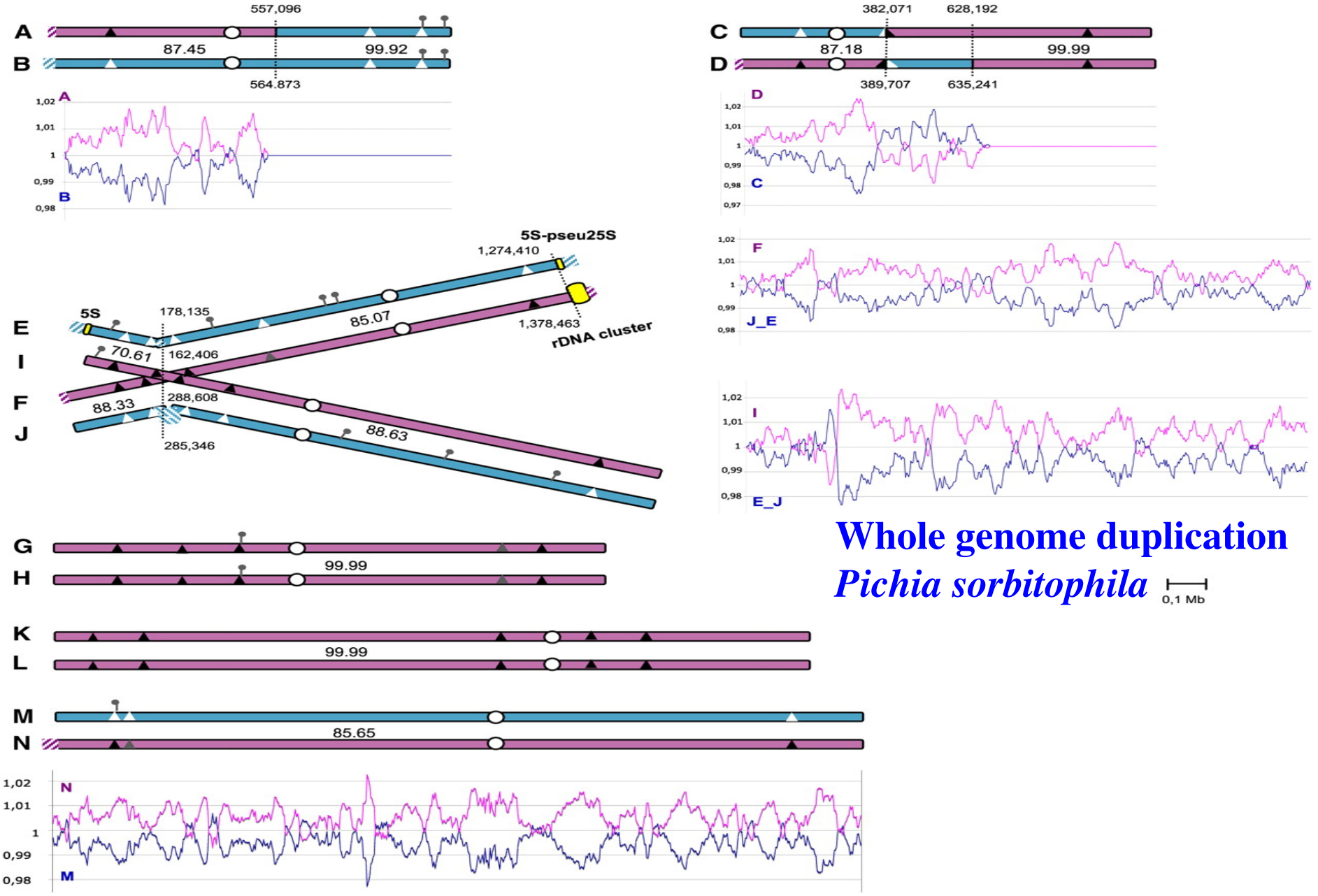
**a**, Distribution of *K*s values of duplicated genes in *Tetraodon* (left) and *Takifugu* (right) genomes. Duplicated genes broadly belong to two categories, depending on their *K*s value being below or higher than 0.35 substitutions per site since the divergence between the two puffer fish (arrows).

**b**, Global distribution of ancient duplicated genes (*K*s > 0.35) in the *Tetraodon* genome. The 21 *Tetraodon* chromosomes are represented in a circle in numerical order and each line joins duplicated genes at their respective position on a given pair of chromosomes.

# Tetraodon genomes

Jaillon et al. *Nature* 431, 946-857. 2004.

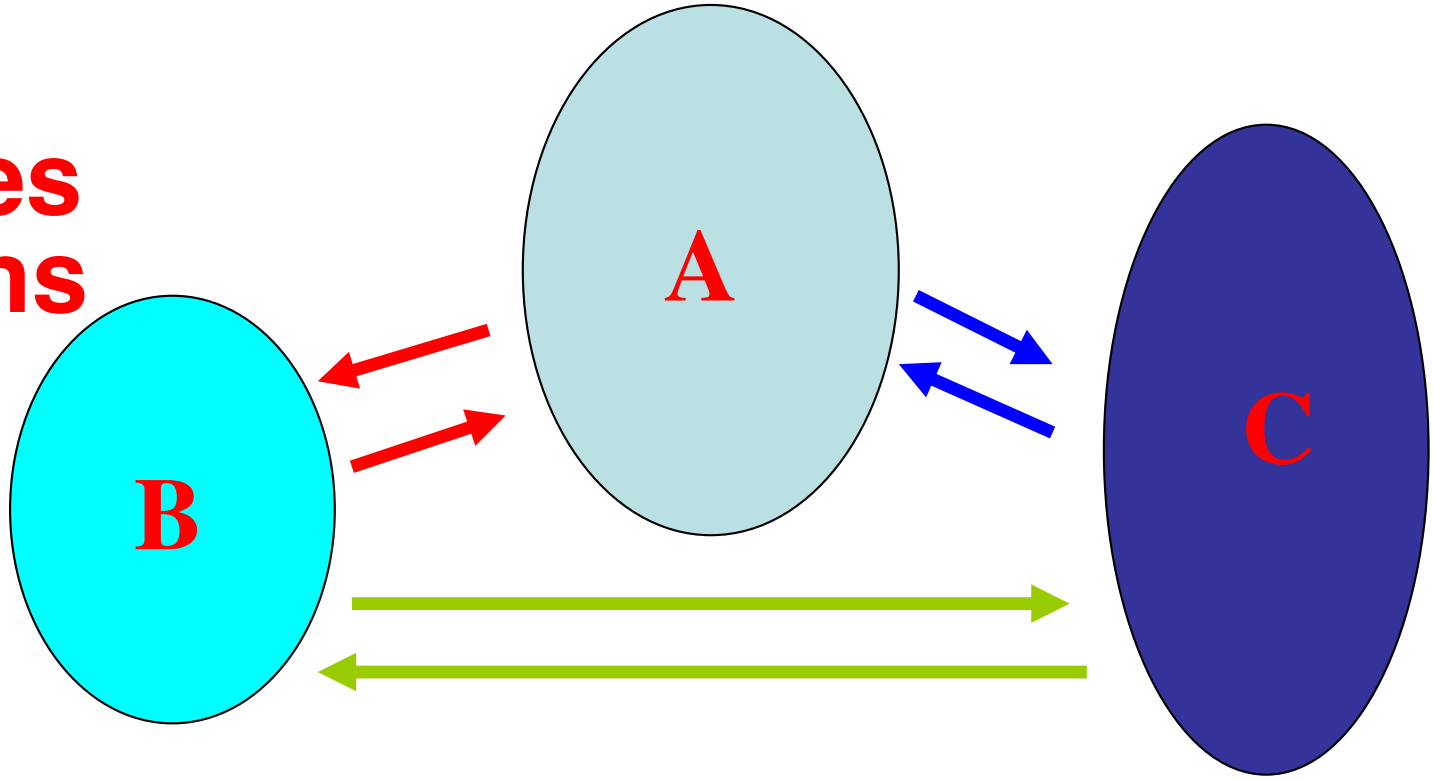
# The hybrid nuclear genome of *P. sorbitophila*




Louis et al. G3. 2;299-311. (2012).

# Comparing genomes

## Inter-species comparisons




- **Rate of ancestral conservation of one species in another species;**

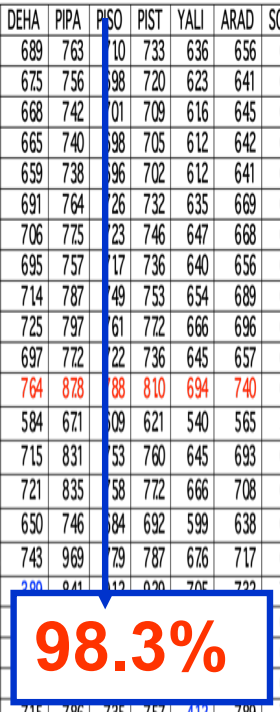
**i.e genes in A that are conserved in B / total number of genes in A; etc…**

- **specific, core-genome, pan-genome, orthologs, synteny, HGT, loss, introgression**

| Spec | MT | MTC | MYBO | MYJL | MYVA | MYSM | MYUL | MYMA | MYLE | MYAV | MYAP | MYAB | MYGI |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MT | 50.8 | 90.3 | 99.5 | 70.3 | 68.0 | 66.5 | 77.3 | 79.5 | 89.2 | 75.7 | 82.2 | 67.5 | 71.1 |
| MTC | 97.0 | | | | 7.5 | 66.2 | 76.9 | 79.2 | 8 | | | 6.8 | 70.6 |
| MYBO | 99.3 | 90.1 | 49.8 | 70.1 | 67.7 | 66.3 | 77.2 | 79.2 | 89.2 | 75.4 | 82.1 | 67.5 | 70.9 |
| MYJL | 78.4 | 72.5 | 78.5 | 59.9 | 86.4 | 81.5 | 75.4 | 79.1 | 84.4 | 82.6 | 82.0 | 77.4 | 87.4 |
| MYVA | 79.6 | 73.6 | 79.7 | 89.4 | 60.1 | 82.4 | 76.1 | 79.4 | 84.0 | 81.8 | 82.3 | 78.0 | 91.0 |
| MYSM | 78.1 | 72.4 | 78.1 | 87.1 | 84.5 | 62.0 | 75.4 | 79.2 | 84.0 | 82.1 | 83.0 | 79.5 | 84.3 |
| MYUL | 84.8 | 79.1 | 85.3 | 76.3 | 73.6 | 70.3 | 56.4 | 90.9 | 88.3 | 80.8 | 83.2 | 73.3 | 73.6 |
| MYMA | 86.5 | 80.0 | 86.4 | 79.1 | 76.5 | 74.0 | 89.0 | 60.6 | 89.0 | 82.6 | 85.9 | 75.2 | 76.9 |
| MYLE | 55.3 | 51.6 | 55.8 | 47.8 | 45.8 | 44.1 | 50.1 | 53.3 | 31.5 | 50.3 | 56.3 | 47.2 | 46.6 |
| MYAV | 81.1 | 75.1 | 81.1 | 78.3 | 74.7 | 71.5 | 78.4 | 81.6 | 87.2 | 57.0 | 96.5 | 73.6 | 75.7 |
| MYAP | 80.8 | 74.8 | 80.9 | 76.9 | 73.5 | 70.6 | 78.0 | 80.6 | 87.2 | 88.2 | 57.1 | 72.5 | 74.8 |
| MYAB | 74.4 | 68.6 | 74.4 | 77.1 | 74.7 | 73.5 | 70.1 | 74.0 | 82.1 | 75.7 | 82.0 | 53.5 | 75.6 |
| MYGI | 78.7 | 72.8 | 78.7 | 88.4 | 88.8 | 80.0 | 74.5 | 78.2 | 87.8 | 80.7 | 86.7 | 76.9 | 58.5 |

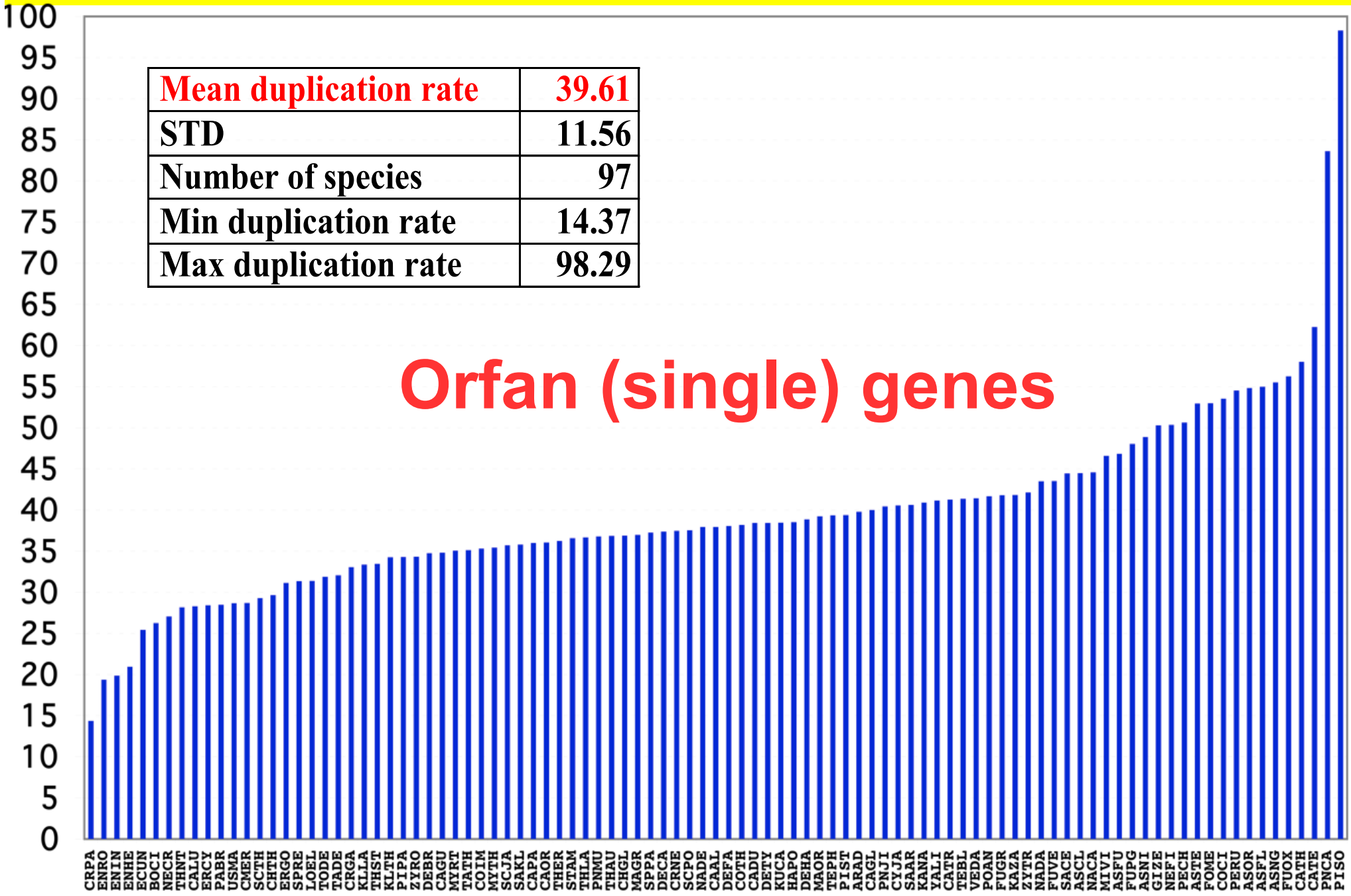
# 43 Yeast & Fungi species

| spec | SACE | CAGL | NACA | NADA | TEPH | TODE | KLLA | KLWA | KLTH | SAKL | ZYRO | KUCA | DEBR | DKBR | OGPO | OGPA | KOPA | DEHA | PIPA | PISO | PIST | YALI | ARAD | SCPO | SCJA | CALU | CAAL | CADU | CATR | CAGU | ECUN | ENIN | ERGO | ERCY | CMER | CRNE | CRGA | USMA | POAN | BOCI | ASFU | NECR | GIZE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SACE | 444 | 928 | 911 | 912 | 915 | 932 | 884 | 667 | 907 | 902 | 929 | 695 | 662 | 680 | 752 | 807 | 751 | 689 | 763 | 710 | 733 | 636 | 656 | 637 | 577 | 652 | 688 | 688 | 686 | 691 | 357 | 427 | 915 | 902 | 347 | 417 | 479 | 420 | 397 | 227 | 472 | 342 | 390 |
| CAGL | 881 | 400 | 889 | 885 | 898 | 902 | 863 | 643 | 877 | 874 | 902 | 679 | 649 | 669 | 734 | 790 | 741 | 675 | 756 | 698 | 720 | 623 | 641 | 628 | 581 | 638 | 679 | 677 | 675 | 675 | 355 | 426 | 896 | 885 | 343 | 411 | 471 | 410 | 381 | 220 | 451 | 334 | 375 |
| NACA | 892 | 917 | 446 | 942 | 906 | 918 | 862 | 651 | 883 | 880 | 909 | 681 | 653 | 673 | 738 | 791 | 742 | 668 | 742 | 701 | 709 | 616 | 645 | 655 | 609 | 661 | 665 | 683 | 663 | 699 | 394 | 428 | 892 | 894 | 401 | 451 | 472 | 456 | 369 | 249 | 442 | 366 | 377 |
| NADA | 886 | 915 | 941 | 435 | 908 | 915 | 859 | 654 | 876 | 874 | 905 | 678 | 650 | 668 | 735 | 793 | 738 | 665 | 740 | 698 | 705 | 612 | 642 | 652 | 604 | 660 | 665 | 683 | 660 | 696 | 396 | 427 | 886 | 890 | 396 | 451 | 468 | 457 | 366 | 246 | 440 | 364 | 375 |
| TEPH | 871 | 905 | 888 | 890 | 394 | 911 | 857 | 644 | 873 | 868 | 902 | 677 | 648 | 665 | 730 | 790 | 737 | 659 | 738 | 696 | 702 | 612 | 641 | 653 | 602 | 657 | 659 | 675 | 655 | 689 | 393 | 429 | 887 | 889 | 402 | 448 | 466 | 457 | 369 | 244 | 439 | 365 | 377 |
| TODE | 899 | 923 | 910 | 910 | 925 | 319 | 891 | 679 | 918 | 913 | 949 | 708 | 674 | 693 | 767 | 819 | 765 | 691 | 764 | 726 | 732 | 635 | 669 | 674 | 623 | 682 | 686 | 702 | 688 | 724 | 402 | 434 | 916 | 919 | 408 | 465 | 486 | 475 | 388 | 260 | 464 | 380 | 398 |
| KLLA | 868 | 895 | 870 | 866 | 882 | 906 | 334 | 666 | 913 | 912 | 909 | 710 | 670 | 692 | 768 | 823 | 763 | 706 | 775 | 723 | 746 | 647 | 668 | 637 | 578 | 662 | 699 | 700 | 699 | 705 | 360 | 436 | 930 | 923 | 349 | 424 | 486 | 424 | 402 | 231 | 484 | 347 | 401 |
| KLWA | 848 | 869 | 862 | 864 | 877 | 903 | 867 | 297 | 937 | 897 | 891 | 701 | 662 | 681 | 760 | 811 | 757 | 695 | 757 | 717 | 736 | 640 | 656 | 630 | 577 | 659 | 689 | 692 | 690 | 693 | 348 | 426 | 904 | 899 | 345 | 421 | 476 | 422 | 380 | 230 | 456 | 342 | 394 |
| KLTH | 890 | 909 | 899 | 895 | 906 | 936 | 913 | 722 | 343 | 934 | 930 | 731 | 690 | 701 | 775 | 838 | 770 | 714 | 787 | 749 | 753 | 654 | 689 | 686 | 620 | 702 | 709 | 722 | 707 | 751 | 412 | 441 | 933 | 926 | 432 | 486 | 508 | 494 | 401 | 270 | 484 | 382 | 402 |
| SAKL | 901 | 924 | 910 | 906 | 918 | 947 | 927 | 699 | 954 | 358 | 941 | 744 | 705 | 713 | 787 | 849 | 784 | 725 | 797 | 761 | 772 | 666 | 696 | 690 | 628 | 713 | 722 | 736 | 720 | 761 | 414 | 439 | 950 | 944 | 433 | 491 | 514 | 497 | 411 | 274 | 493 | 390 | 407 |
| ZYRO | 889 | 915 | 895 | 890 | 910 | 943 | 893 | 669 | 915 | 907 | 343 | 701 | 666 | 687 | 762 | 811 | 758 | 697 | 772 | 722 | 736 | 645 | 657 | 638 | 604 | 679 | 695 | 693 | 691 | 717 | 401 | 430 | 922 | 909 | 410 | 457 | 480 | 484 | 401 | 255 | 474 | 377 | 393 |
| KUCA | 758 | 786 | 768 | 762 | 782 | 800 | 785 | 577 | 789 | 786 | 796 | 38.5 | 777 | 791 | 880 | 917 | 864 | 764 | 878 | 788 | 810 | 694 | 740 | 693 | 640 | 747 | 754 | 751 | 754 | 809 | 417 | 442 | 806 | 788 | 441 | 515 | 540 | 505 | 445 | 298 | 531 | 416 | 445 |
| DEBR | 600 | 619 | 605 | 606 | 617 | 625 | 612 | 451 | 611 | 614 | 624 | 637 | 25.7 | 745 | 720 | 798 | 667 | 584 | 671 | 609 | 621 | 540 | 565 | 561 | 528 | 575 | 577 | 592 | 577 | 614 | 347 | 362 | 630 | 618 | 368 | 410 | 424 | 404 | 335 | 228 | 405 | 320 | 336 |
| DKBR | 730 | 756 | 738 | 737 | 748 | 765 | 758 | 551 | 764 | 765 | 763 | 786 | 949 | 34.6 | 885 | 919 | 816 | 715 | 831 | 753 | 760 | 645 | 693 | 676 | 622 | 706 | 711 | 726 | 708 | 754 | 411 | 430 | 773 | 757 | 430 | 483 | 509 | 476 | 408 | 271 | 488 | 389 | 406 |
| OGPO | 732 | 757 | 740 | 740 | 751 | 770 | 758 | 558 | 770 | 770 | 770 | 802 | 807 | 809 | 38.5 | 978 | 824 | 721 | 835 | 758 | 772 | 666 | 708 | 680 | 620 | 713 | 718 | 732 | 718 | 771 | 413 | 438 | 777 | 758 | 429 | 496 | 520 | 490 | 423 | 286 | 507 | 397 | 428 |
| OGPA | 667 | 688 | 671 | 670 | 682 | 695 | 687 | 512 | 698 | 694 | 697 | 716 | 760 | 724 | 860 | 39.3 | 751 | 650 | 746 | 684 | 692 | 599 | 638 | 624 | 573 | 645 | 647 | 660 | 646 | 698 | 389 | 422 | 700 | 694 | 389 | 451 | 469 | 462 | 379 | 259 | 457 | 369 | 397 |
| KOPA | 750 | 775 | 758 | 756 | 770 | 788 | 776 | 572 | 789 | 789 | 787 | 806 | 761 | 760 | 842 | 896 | 34.3 | 743 | 969 | 779 | 787 | 676 | 717 | 691 | 630 | 732 | 736 | 749 | 732 | 784 | 424 | 444 | 797 | 782 | 439 | 500 | 524 | 494 | 425 | 284 | 510 | 403 | 423 |
| DEHA | 770 | 791 | 758 | 758 | 774 | 799 | 796 | 589 | 811 | 810 | 809 | 789 | 721 | 743 | 827 | 873 | 827 | [illegible] | | | | | | 703 | 599 | 840 | 870 | 874 | 863 | 907 | 364 | 439 | 815 | 793 | 358 | 463 | 523 | 468 | 448 | 266 | 539 | 389 | 454 |
| PIPA | 751 | 776 | 747 | 743 | 758 | 772 | 777 | 560 | 789 | 788 | 786 | 796 | 743 | 760 | 838 | 880 | 961 | [illegible] | | | | | | 690 | 619 | 718 | 735 | 731 | 717 | 766 | 411 | 443 | 798 | 783 | 417 | 482 | 507 | 511 | 425 | 274 | 508 | 402 | 422 |
| PISO | 744 | 773 | 750 | 746 | 761 | 787 | 774 | 568 | 788 | 784 | 784 | 775 | 711 | 734 | 819 | 867 | 817 | [illegible] | | | | | | 689 | 622 | 828 | 835 | 848 | 828 | 892 | 409 | 441 | 790 | 788 | 416 | 487 | 513 | 503 | 414 | 280 | 505 | 409 | 433 |
| PIST | 754 | 777 | 746 | 746 | 758 | 783 | 779 | 574 | 793 | 794 | 790 | 771 | 710 | 732 | 818 | 857 | 814 | [illegible] | | | | | | 654 | 590 | 823 | 870 | 872 | 867 | 883 | 358 | 432 | 798 | 779 | 357 | 450 | 514 | 455 | 437 | 256 | 524 | 379 | 440 |
| YALI | 726 | 744 | 713 | 710 | 725 | 745 | 747 | 549 | 756 | 756 | 758 | 730 | 663 | 695 | 778 | 824 | 771 | 715 | 786 | 735 | 757 | 412 | 789 | 703 | 648 | 673 | 713 | 709 | 709 | 716 | 376 | 458 | 764 | 747 | 383 | 485 | 543 | 494 | 467 | 281 | 556 | 410 | 466 |
| ARAD | 716 | 736 | 724 | 720 | 731 | 754 | 743 | 542 | 752 | 752 | 753 | 744 | 678 | 706 | 790 | 834 | 782 | 713 | 796 | 745 | 756 | 758 | 410 | 740 | 683 | 699 | 704 | 720 | 705 | 754 | 424 | 459 | 753 | 752 | 444 | 538 | 564 | 553 | 472 | 325 | 576 | 460 | 499 |
| SCPO | 615 | 630 | 619 | 614 | 628 | 645 | 623 | 460 | 644 | 633 | 645 | 608 | 558 | 585 | 653 | 710 | 647 | 578 | 647 | 610 | 614 | 592 | 636 | 393 | 845 | 571 | 574 | 585 | 568 | 612 | 427 | 460 | 639 | 636 | 427 | 498 | 517 | 500 | 403 | 272 | 479 | 397 | 411 |
| SCJA | 561 | 575 | 594 | 590 | 604 | 618 | 569 | 414 | 579 | 575 | 586 | 573 | 539 | 559 | 623 | 682 | 621 | 517 | 621 | 581 | 546 | 535 | 606 | 856 | 357 | 512 | 512 | 557 | 513 | 543 | 378 | 451 | 587 | 613 | 359 | 438 | 494 | 435 | 379 | 232 | 447 | 346 | 383 |
| CALU | 704 | 731 | 727 | 728 | 742 | 762 | 735 | 544 | 745 | 746 | 743 | 748 | 692 | 715 | 799 | 846 | 796 | 807 | 796 | 853 | 862 | 639 | 696 | 636 | 587 | 283 | 799 | 827 | 808 | 861 | 351 | 425 | 744 | 763 | 349 | 445 | 500 | 445 | 407 | 249 | 490 | 367 | 422 |
| CAAL | 735 | 759 | 751 | 752 | 763 | 784 | 755 | 577 | 765 | 766 | 768 | 760 | 708 | 725 | 811 | 855 | 806 | 846 | 822 | 871 | 917 | 692 | 707 | 659 | 602 | 815 | 379 | 974 | 919 | 873 | 364 | 435 | 801 | 781 | 358 | 449 | 511 | 456 | 431 | 254 | 513 | 375 | 430 |
| CADU | 746 | 766 | 750 | 750 | 761 | 781 | 768 | 567 | 780 | 778 | 777 | 759 | 705 | 726 | 808 | 854 | 806 | 838 | 807 | 871 | 909 | 679 | 709 | 683 | 631 | 821 | 953 | 384 | 911 | 877 | 399 | 437 | 784 | 781 | 420 | 489 | 511 | 505 | 414 | 279 | 496 | 408 | 428 |
| CATR | 739 | 765 | 734 | 733 | 740 | 767 | 765 | 562 | 778 | 777 | 776 | 749 | 690 | 714 | 792 | 840 | 791 | 827 | 790 | 853 | 902 | 679 | 699 | 641 | 588 | 811 | 913 | 916 | 413 | 868 | 356 | 426 | 781 | 762 | 350 | 444 | 499 | 445 | 408 | 250 | 492 | 365 | 425 |
| CAGU | 713 | 738 | 733 | 729 | 743 | 769 | 739 | 542 | 750 | 748 | 748 | 762 | 698 | 723 | 808 | 855 | 803 | 829 | 802 | 866 | 876 | 644 | 709 | 643 | 580 | 819 | 817 | 835 | 823 | 348 | 355 | 425 | 743 | 762 | 351 | 447 | 509 | 447 | 412 | 253 | 502 | 370 | 435 |
| ECUN | 197 | 205 | 231 | 228 | 234 | 225 | 194 | 140 | 192 | 189 | 198 | 195 | 192 | 201 | 223 | 260 | 232 | 165 | 230 | 207 | 174 | 170 | 206 | 231 | 220 | 166 | 164 | 194 | 1[illegible] | | | | | | | 142 | 178 | 144 | 120 | 6.3 | 138 | 104 | 113 |
| ENIN | 231 | 237 | 238 | 235 | 241 | 233 | 228 | 164 | 228 | 222 | 234 | 202 | 196 | 205 | 227 | 269 | 238 | 193 | 236 | 213 | 208 | 200 | 213 | 268 | 252 | 198 | 194 | 199 | 1[illegible] | | | | | | | 176 | 188 | 187 | 126 | 7.9 | 143 | 129 | 113 |
| ERGO | 849 | 883 | 849 | 848 | 869 | 887 | 882 | 651 | 887 | 887 | 891 | 680 | 657 | 675 | 740 | 797 | 745 | 684 | 760 | 709 | 723 | 626 | 640 | 622 | 582 | 640 | 681 | 678 | 67[illegible] | | | | | | | 410 | 465 | 411 | 382 | 222 | 455 | 335 | 375 |
| ERCY | 823 | 849 | 832 | 832 | 848 | 862 | 854 | 621 | 855 | 857 | 857 | 666 | 643 | 663 | 726 | 779 | 732 | 655 | 733 | 693 | 689 | 610 | 636 | 643 | 601 | 650 | 654 | 668 | 64[illegible] | | | | | | | 449 | 468 | 458 | 368 | 246 | 439 | 366 | 359 |
| CMER | 335 | 344 | 382 | 383 | 391 | 397 | 339 | 250 | 344 | 339 | 351 | 360 | 345 | 360 | 397 | 451 | 405 | 300 | 401 | 368 | 318 | 307 | 378 | 388 | 360 | 304 | 298 | 354 | 292 | 316 | 315 | 386 | 357 | 399 | 287 | 273 | 331 | 286 | 239 | 127 | 283 | 202 | 239 |
| CRNE | 543 | 555 | 585 | 581 | 591 | 606 | 558 | 408 | 564 | 562 | 565 | 597 | 534 | 562 | 635 | 692 | 622 | 533 | 621 | 592 | 563 | 548 | 629 | 637 | 591 | 518 | 521 | 577 | 521 | 559 | 357 | 442 | 558 | 602 | 372 | 375 | 945 | 597 | 427 | 266 | 520 | 384 | 452 |
| CRGA | 581 | 598 | 586 | 582 | 592 | 607 | 598 | 436 | 604 | 603 | 608 | 595 | 537 | 563 | 636 | 695 | 624 | 571 | 622 | 594 | 603 | 583 | 630 | 673 | 625 | 553 | 563 | 577 | 559 | 599 | 404 | 442 | 602 | 601 | 427 | 914 | 331 | 627 | 426 | 292 | 514 | 411 | 446 |
| USMA | 539 | 540 | 596 | 592 | 606 | 620 | 547 | 400 | 558 | 556 | 553 | 604 | 554 | 577 | 649 | 706 | 642 | 527 | 640 | 608 | 556 | 553 | 648 | 686 | 577 | 512 | 515 | 596 | 516 | 549 | 346 | 444 | 554 | 599 | 371 | 581 | 661 | 287 | 465 | 278 | 553 | 393 | 467 |
| POAN | 655 | 666 | 643 | 641 | 653 | 669 | 670 | 481 | 680 | 679 | 681 | 664 | 603 | 629 | 704 | 758 | 686 | 648 | 711 | 661 | 689 | 676 | 716 | 727 | 672 | 622 | 641 | 640 | 623 | 669 | 399 | 435 | 680 | 665 | 427 | 557 | 582 | 594 | 417 | 476 | 760 | 716 | 741 |
| BOCI | 601 | 607 | 637 | 636 | 643 | 663 | 615 | 450 | 630 | 625 | 623 | 661 | 601 | 622 | 705 | 762 | 688 | 606 | 687 | 655 | 636 | 629 | 720 | 693 | 639 | 584 | 587 | 638 | 583 | 633 | 346 | 424 | 619 | 654 | 365 | 517 | 576 | 536 | 709 | 263 | 754 | 626 | 718 |
| ASFU | 678 | 688 | 663 | 661 | 673 | 697 | 696 | 444 | 705 | 707 | 705 | 696 | 623 | 653 | 736 | 786 | 724 | 681 | 739 | 688 | 718 | 697 | 749 | 756 | 611 | 576 | 666 | 663 | 578 | 623 | 333 | 445 | 707 | 690 | 358 | 507 | 600 | 612 | 699 | 444 | 468 | 577 | 709 |
| NECR | 616 | 625 | 665 | 662 | 679 | 695 | 631 | 459 | 639 | 641 | 637 | 683 | 626 | 651 | 730 | 782 | 716 | 615 | 717 | 685 | 649 | 637 | 739 | 702 | 635 | 598 | 599 | 666 | 590 | 639 | 353 | 437 | 636 | 675 | 379 | 524 | 605 | 542 | 783 | 442 | 733 | 271 | 720 |
| GIZE | 666 | 679 | 667 | 663 | 677 | 700 | 685 | 503 | 695 | 694 | 692 | 695 | 623 | 657 | 736 | 785 | 725 | 672 | 726 | 689 | 709 | 688 | 750 | 747 | 678 | 649 | 664 | 673 | 659 | 707 | 414 | 449 | 696 | 688 | 459 | 587 | 615 | 611 | 778 | 501 | 791 | 700 | 516 |




# Rate of Duplication in 97 yeast and fungi species

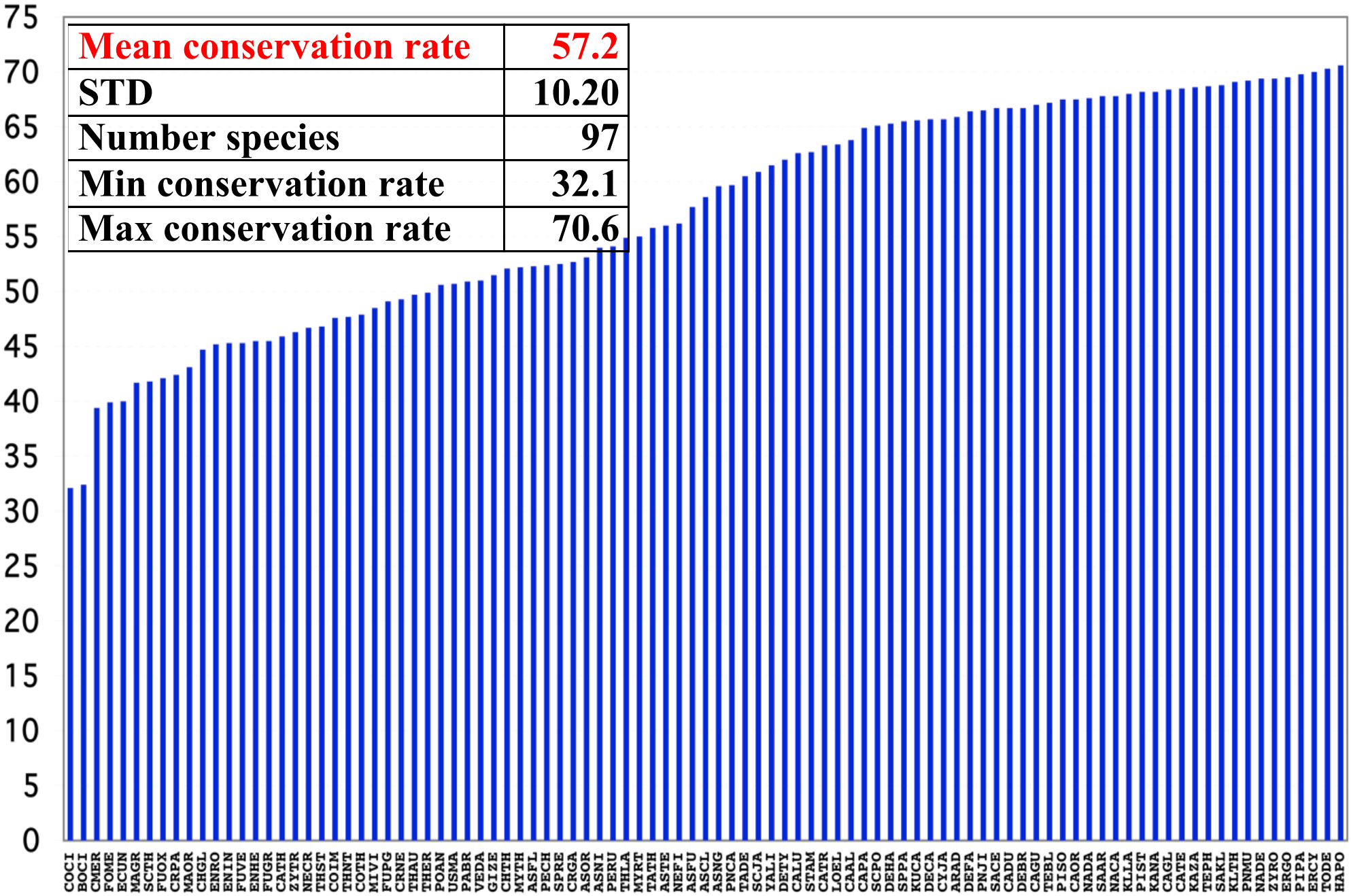
| | |
|---|---|
| Mean duplication rate | 39.61 |
| STD | 11.56 |
| Number of species | 97 |
| Min duplication rate | 14.37 |
| Max duplication rate | 98.29 |




Fredj Tekaia, unpublished

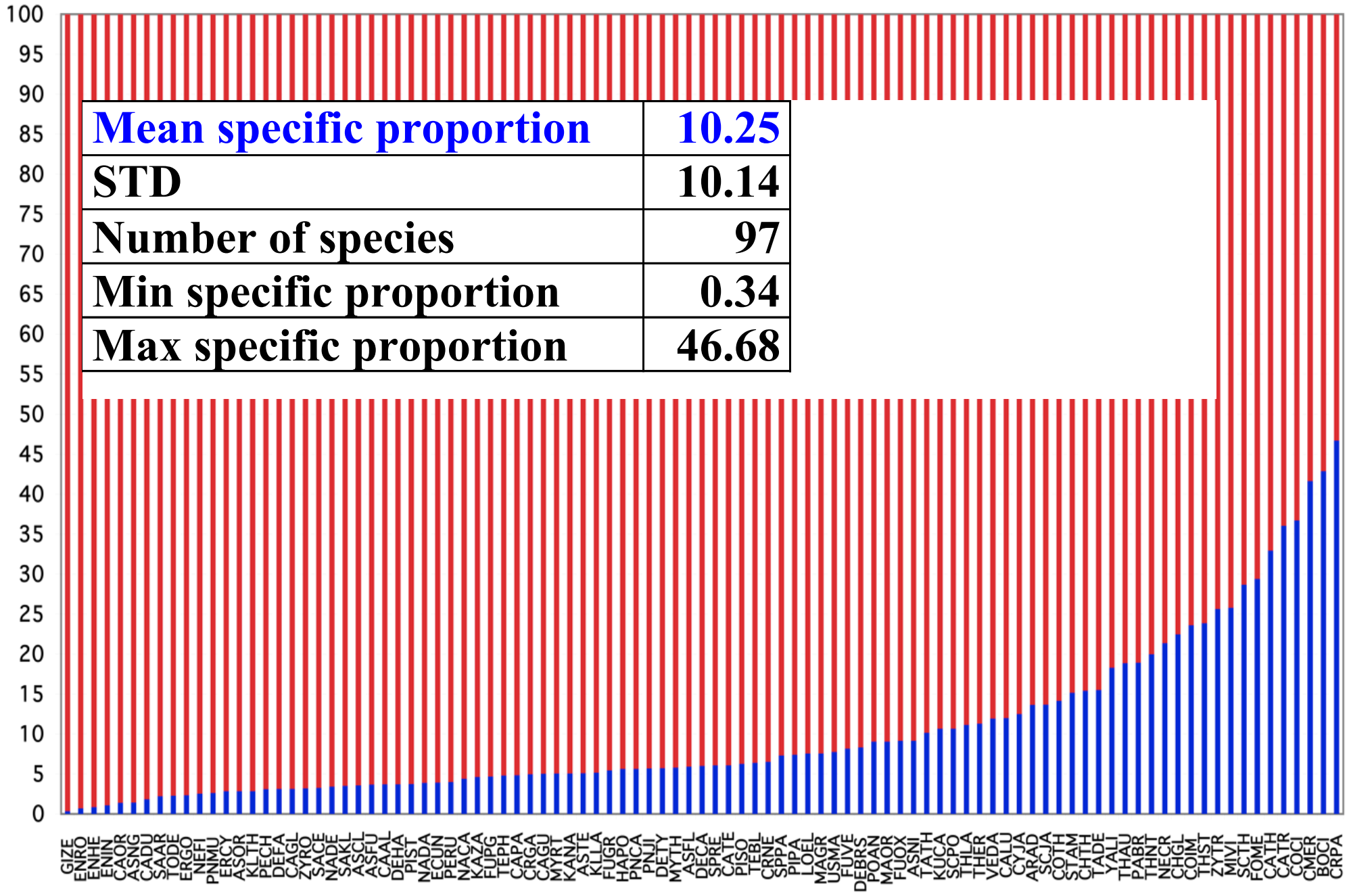
# Rate of Conservation in 97 yeast and fungi species

| | |
|---|---|
| Mean conservation rate | 57.2 |
| STD | 10.20 |
| Number species | 97 |
| Min conservation rate | 32.1 |
| Max conservation rate | 70.6 |




Fredj Tekaia, unpublished

# Specific and Nonspecific genes in 97 yeast and fungi species




| | |
|---|---|
| Mean specific proportion | 10.25 |
| STD | 10.14 |
| Number of species | 97 |
| Min specific proportion | 0.34 |
| Max specific proportion | 46.68 |

Fredj Tekaia, unpublished

# Intra-species polymorphisme:

# Pan-genome/Core-genome/Dispensable genome

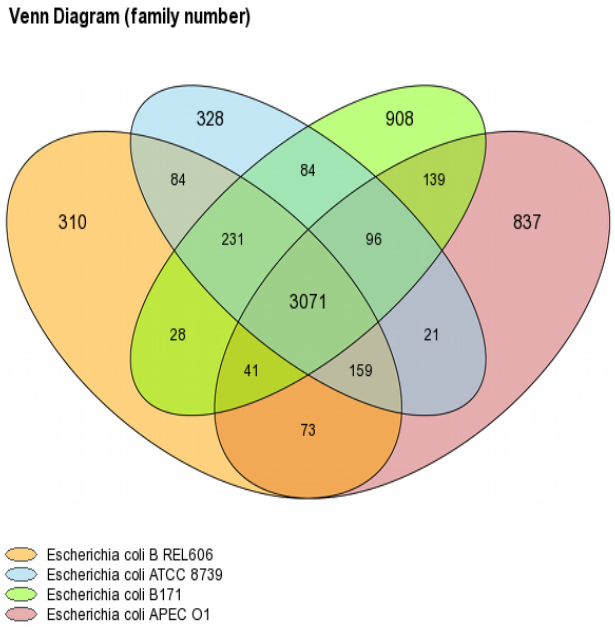
# Pan-genome

- **Availability of large amounts of complete genome sequences allowed genome analyses to be performed on thousands of genomes**

- **Pan-genome analyses provide a framework for estimating the genomic diversity from the dataset at hand and predicting the number of additional whole genomes sequences that would be necessary to fully characterize that diversity.**

# Pan-genome/Core-genome/Dispensable genome

- **Pan-genome:** The pan-genome is the entire gene set of all strains of a species. It includes genes present in all strains (core genome) and genes present only in some strains of a species (dispensable genome).



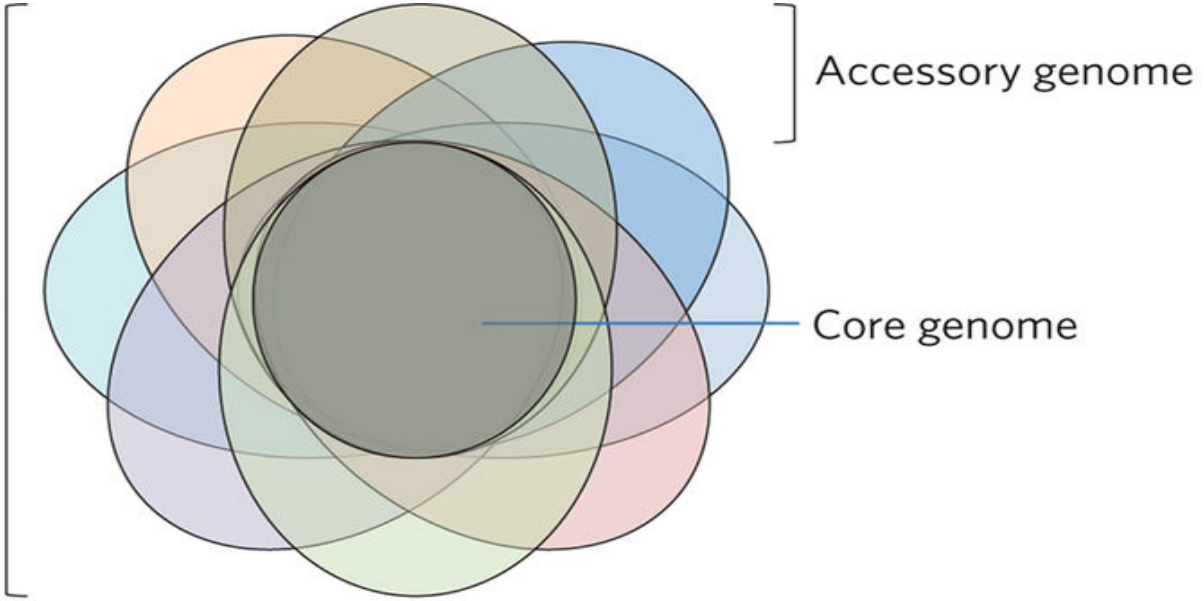

- **Core-genome:**

The core genome is the set of genes present in all strains of a species. (Ancestral part)

- **Dispensable genome/Accessory/variable/flexible:**

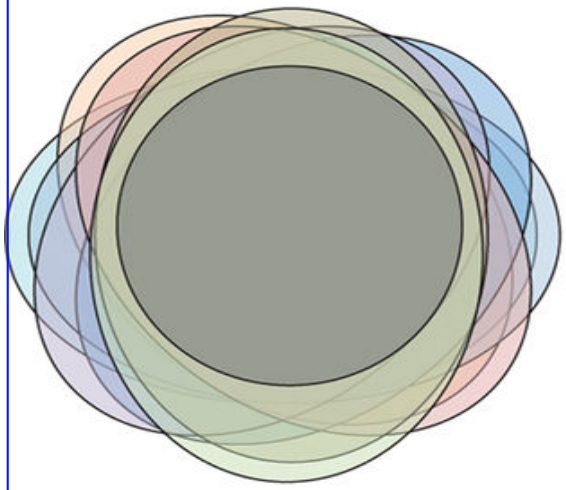
The dispensable genome refers to the set of genes present in at least one strain but not in all of them.

# Schematic representation of pangenomes as Venn diagrams

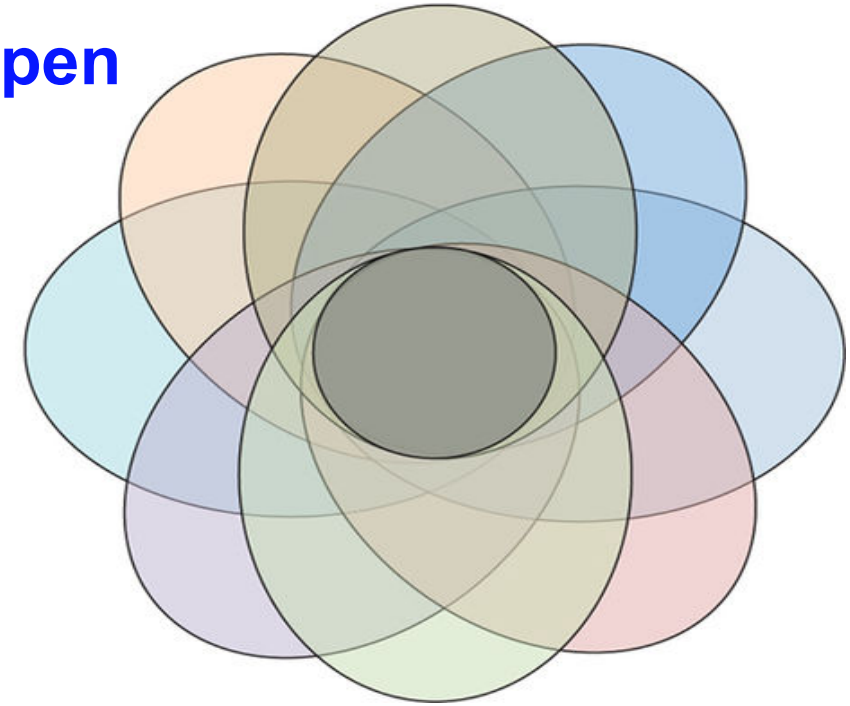



**Species differ in pangenomes sizes, with larger, more open pangenomes correlating with larger long-term effective population sizes and the ability to migrate.**

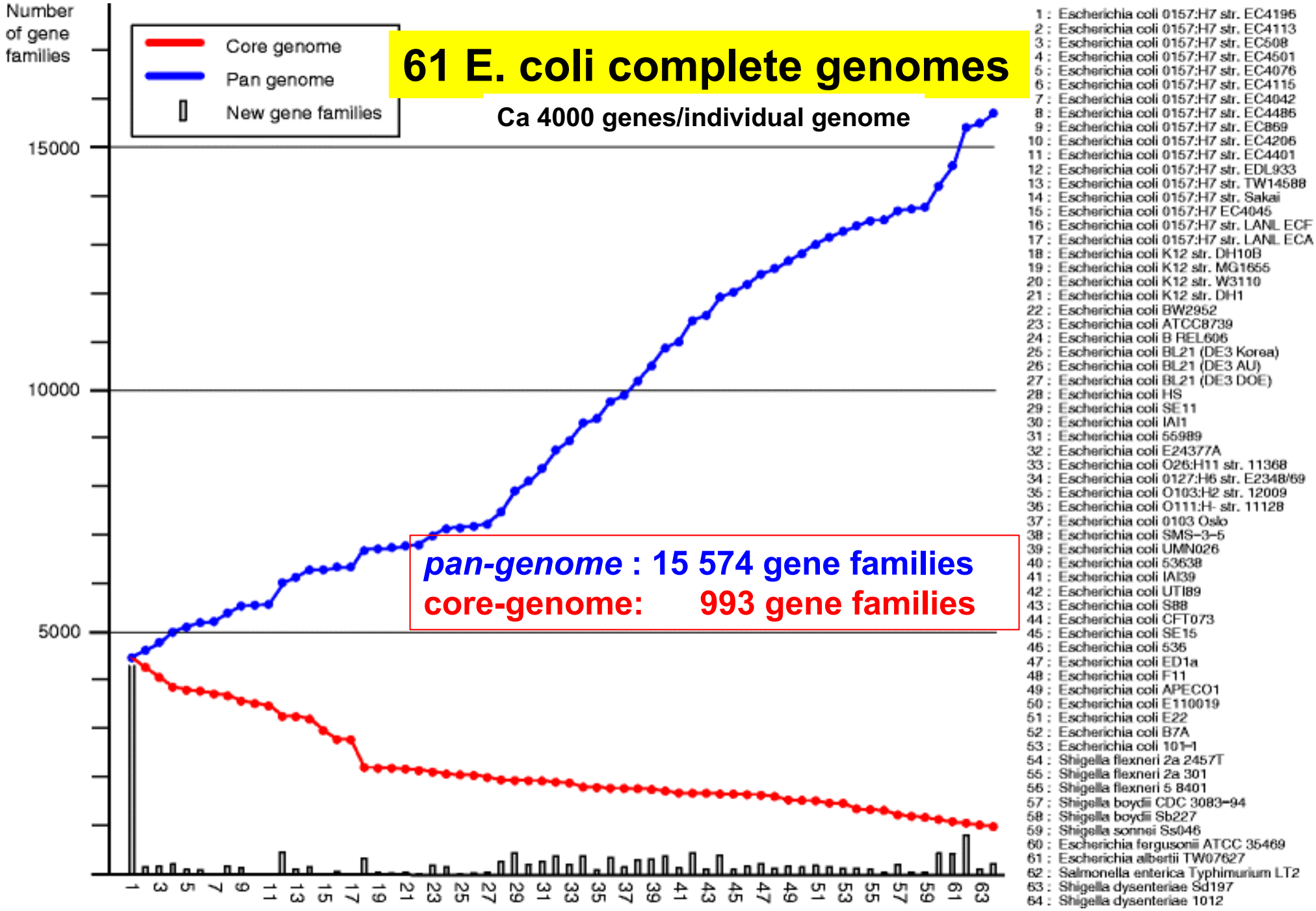
**closed**



**open**

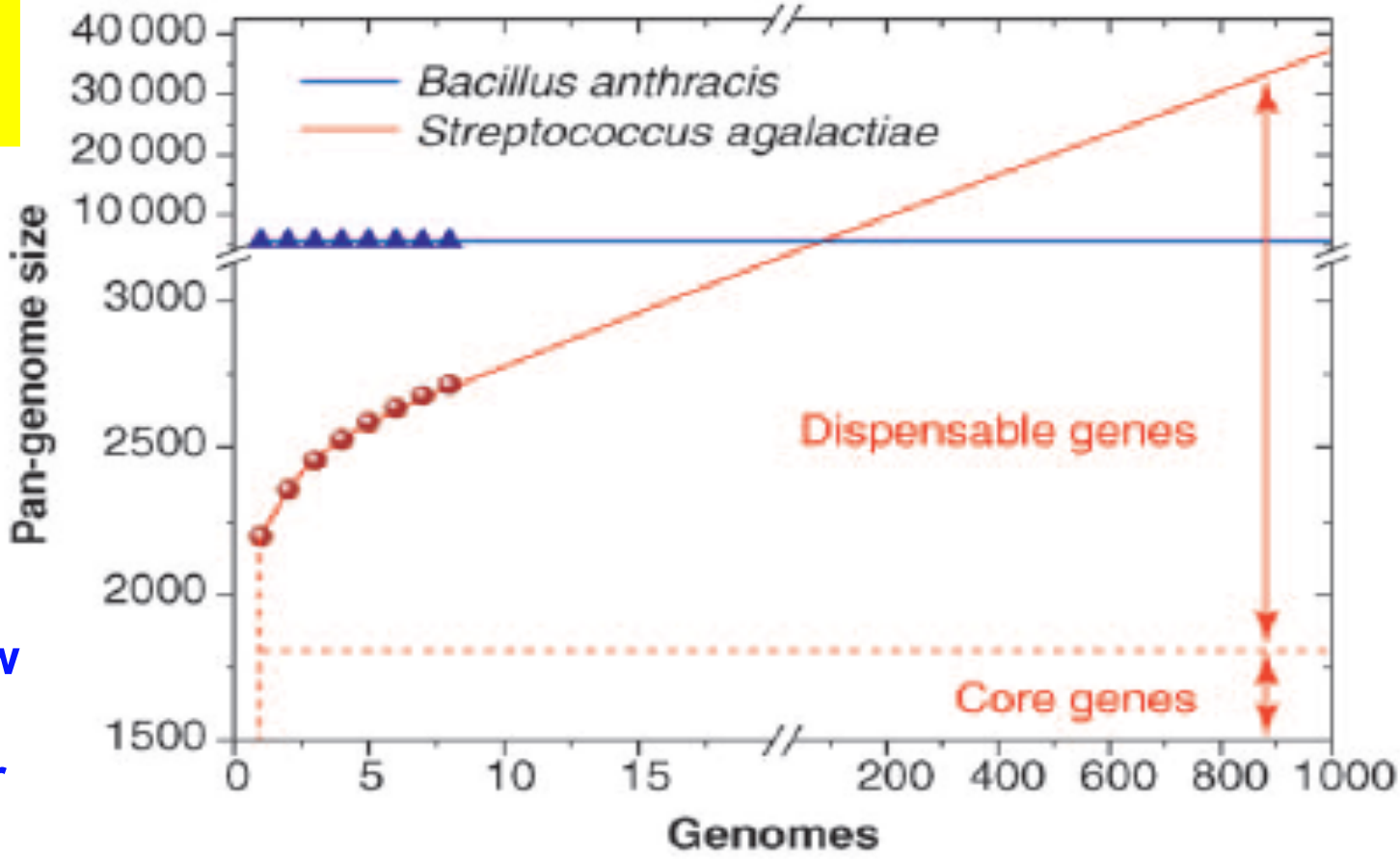


McInerney JO, McNally A, O'Connell MJ. (2017). Why prokaryotes have pangenomes. Nat Microbiol. 2:17040.




Lukjancenko *et al.*, 2010 *Microb. Ecol.* 60: 708-720

# closed/open pan-genomes




The size of a species pan-genome can grow with the number of sequenced strains, or quickly saturate to a limiting value.

**Closed species:** ***e.g. B.* anthracis**

**Open species:** ***e.g. S. agalactiae***

After sequencing a large number of strains, the number of dispensable genes in an open pan-genome is orders of magnitude larger than the size of the core genome, forcing us to reconsider the **definition of a bacterial species.**

Medini *et al.*, 2005 «*The microbial pan-genome*». Curr. Op. Genetics & Development 15: 589-594

# Example: Ratio core/pangenome of several bacterial species

| Species | Genomes | Lifestyle | Intracellular | Niche | % |
|---|---|---|---|---|---|
| *Prochlorococcus marinus* | 12 | Sympatric | no | Marine environment | 5 |
| *Clostridium botulinum* | 14 | Sympatric | no | Soil | 11 |
| *Rhodopseudomonas palustris* | 7 | Sympatric | no | Soil, marine environment | 46 |
| *Sinorhizobium meliloti* | 6 | Sympatric | no | Soil | 49 |
| *Salmonella enterica* | 20 | Sympatric | facultative | Animals | 62 |
| *Acinetobacter baumannii* | 11 | Sympatric | no | ? | 65 |
| *Legionella pneumophila* | 11 | Sympatric | facultative | Amoeba | 69 |
| *Escherichia coli* | 19 | Sympatric | no | Animals | 70 |
| *Bacillus cereus* | 12 | Sympatric | no | Soil | 74 |
| *Campylobacter jejuni* | 14 | Sympatric | facultative | Human, chicken | 76 |
| *Clostridium difficile* | 18 | Sympatric | no | Human gut | 77 |
| *Helicobacter pylori* | 10 | Sympatric | facultative | Human | 78 |
| *Haemophilus influenzae* | 9 | Sympatric | facultative | Human | 80 |
| *Streptococcus pneumoniae* | 10 | Sympatric | no | Human | 82 |
| *Pseudomonas aeruginosa* | 7 | Sympatric | no | Water | 84 |
| *Streptococcus agalactiae* | 5 | Sympatric | no | Human | 84 |
| *Listeria monocytogenes* | 20 | Sympatric | facultative | Amoeba? | 84 |
| *Francisella tularensis* | 13 | Sympatric | facultative | Ticks | 87 |
| *Yersinia pestis* | 12 | Allopatric | facultative | Rodents | 89 |
| *Coxiella burnetii* | 7 | Allopatric | yes | Animals | 90 |
| *Tropheryma whipplei* | 19 | Allopatric | yes | Human | 94 |
| *Mycobacterium tuberculosis* | 20 | Allopatric | yes | Human | 96 |
| *Buchnera aphidicola* | 8 | Allopatric | yes | Aphid | 98 |
| *Bacillus anthracis* | 9 | Allopatric | no | Animals | 99 |
| *Rickettsia rickettsii* | 8 | Allopatric | yes | Ticks | 99 |
| *Chlamydia trachomatis* | 20 | Allopatric | yes | Human | 99 |
| *Rickettsia prowazekii* | 8 | Allopatric | yes | Human | 100 |

Rouli et al. (2015). The bacterial pangenome as a new tool for analysing pathogenic bacteri. New Microbes New Infect. 2015;7:72-85.

# panX: Pan-genome Analysis & Exploration

http://pangenome.tuebingen.mpg.de

# Horizontal Gene Transfer (HGT)

**Not all our genes are from our ancestors - some are "foreign"**

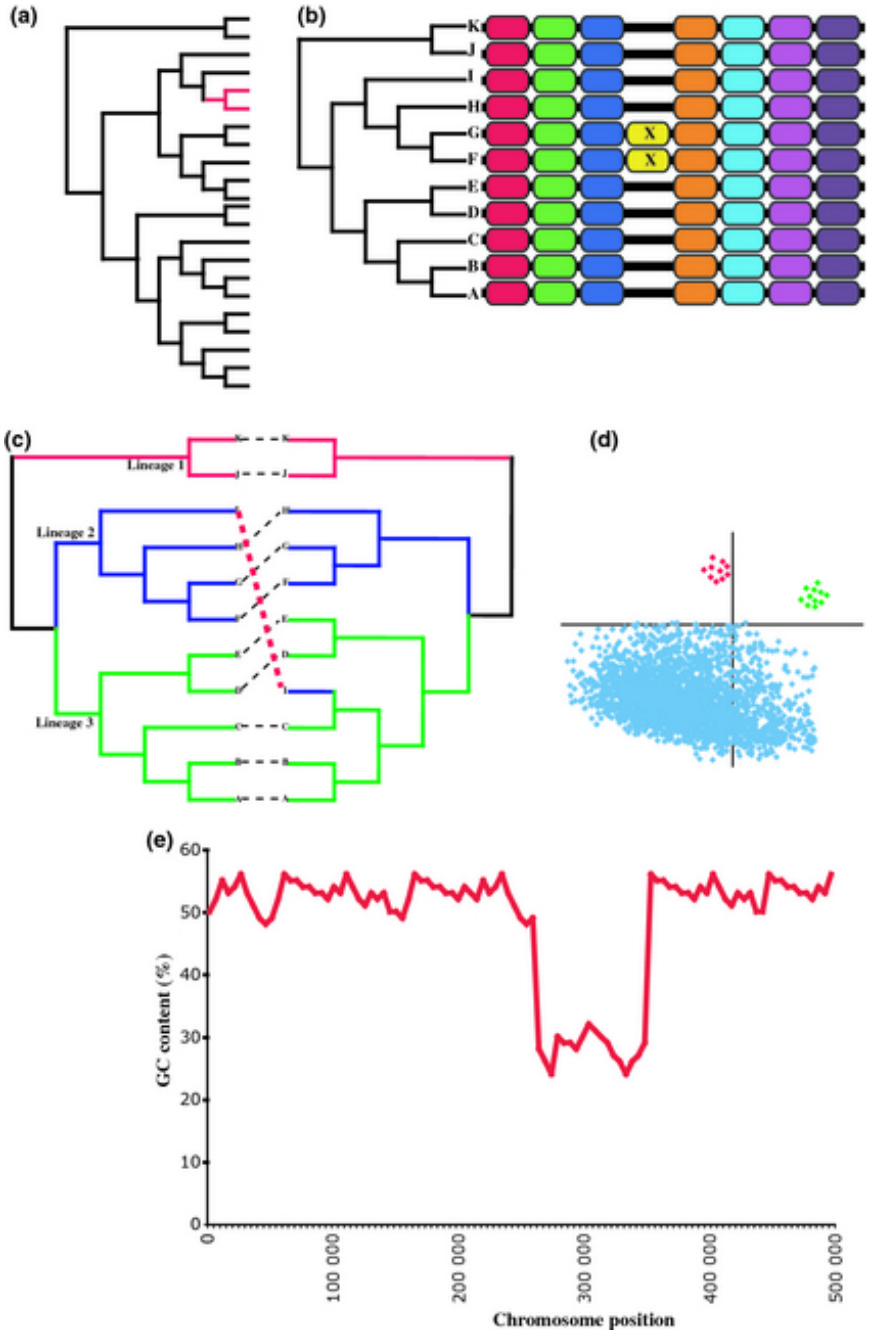
http://www.medicalnewstoday.com/articles/290835.php

# Horizontal Gene Transfer

- A fundamental concept in biology is that heritable material, DNA, is passed from parent to offspring, a process called **vertical gene transfer**.

- An alternative mechanism of gene acquisition is through **horizontal gene transfer (HGT)**, which involves exchange of genetic material between different species.

# *In silico* detection of HGT




**Fitzpatrick.DA. FEMS Microbiology Letters** Volume 329, Issue 1, pages 1-8, 15 DEC 2011 DOI: 10.1111/j.1574-6968.2011.02465.x
http://onlinelibrary.wiley.com/doi/10.1111/j.1574-6968.2011.02465.x/full#fml2465-fig-0001

# HGT in Prokaryotes

# Horizontal Gene Transfer

**is an important mechanism of natural variation among prokaryotes.**

**Gene transfer has been identified as :**

**- a prevalent and pervasive phenomenon**

**- an important source of genomic innovation in bacteria.**

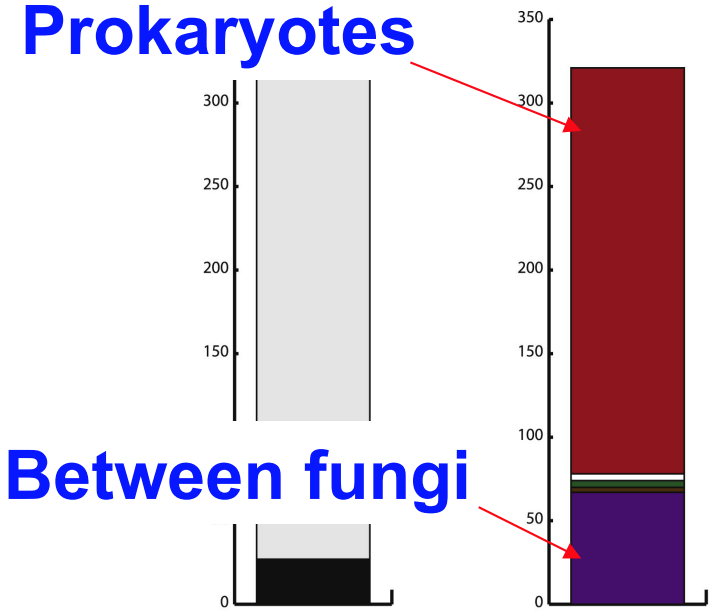
- **Estimates suggest that on average 81% of prokaryotic genes have been involved in HGT at some point of their history.**

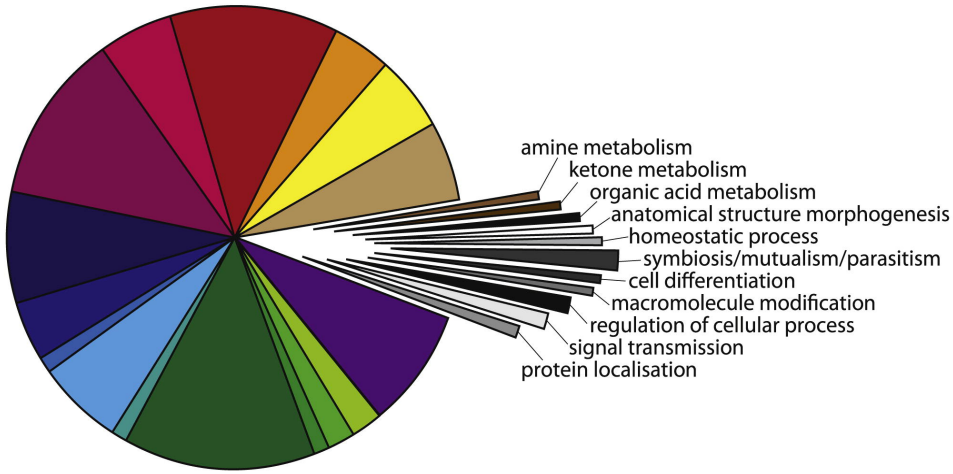
Dagan T, Artzy-Randrup Y, Martin W. (2008). Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. *Proc Natl Acad Sci USA*. 105:10039–44.

# HGT in Eukaryotes

# Ancestry of Horizontal Gene Transfer in Fungi

## 323 transfers into the fungi




**Bar chart: total number of HGTs identified**

- Non secreted HGTs
- HGT predicted to be secreted

**Bar chart: ancestry of HGTS**

- Prokaryotes
- Archaeplastida
- Stramenopiles
- Between fungi
- Virus

**Pie chart: BLAST2GO functional categories**

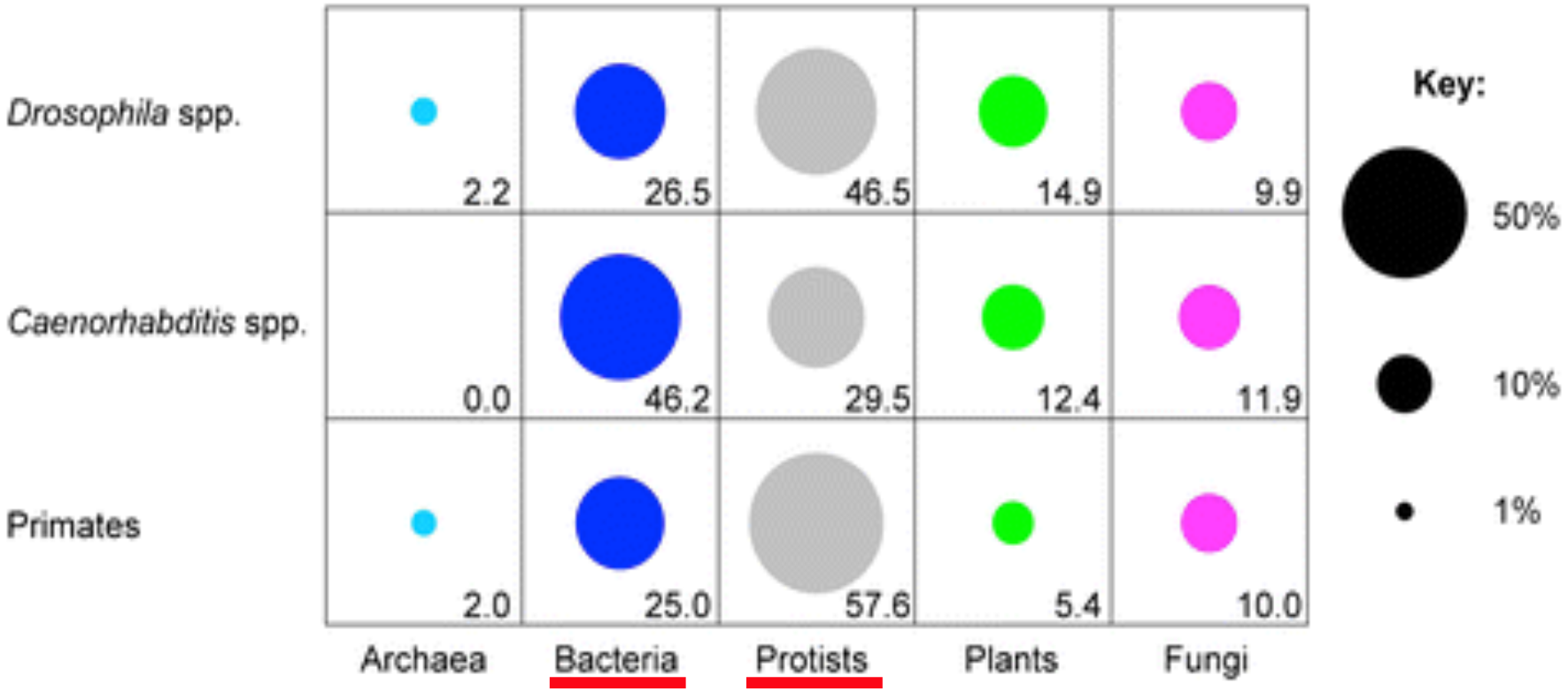
- cellular biosynthetic process
- gene expression
- generation of precursor metabolites and energy
- vitamin metabolism
- macromolecule biosynthetic process
- nucleobase, nucleoside, nucleotide and nucleic acid metabolism
- transport
- cofactor metabolism
- protein metabolism
- cellular macromolecule metabolism
- aromatic compound metabolism
- heterocycle metabolism
- nitrogen compound metabolism
- lipid metabolism
- carbohydrate metabolism
- amino acid and derivative metabolism

Richards TA et al. (2011). Fungal Biology Reviews Volume 25, Issue 2 2011 98 - 110

# HGT in vertebrate and invertebrate genomes

**The most common donors were Bacteria and Protists**

**Other genes came from fungi and viruses.**




Crisp A, Boschetti C, Perry M, Tunnacliffe A, Micklem G. (2015). Expression of multiple horizontally acquired genes is a hallmark of both vertebrate and invertebrate genomes. Genome Biol.16:50.

# Note: Previous results were challenged*

**Our analyses indicate that eukaryotes do not acquire genes through continual LGT like prokaryotes do.**

**We propose a 70 % rule: Coding sequences in eukaryotic genomes that share more than 70 % amino acid sequence identity to prokaryotic homologs are most likely assembly or annotation artifacts.**

*Ku C, Martin WF. (2016). A natural barrier to lateral gene transfer from prokaryotes to eukaryotes revealed from genomes: the 70 % rule. BMC Biol. 2016 Oct 17;14(1):89.

# Horizontal gene transfer is less frequent in eukaryotes than prokaryotes but can be important (retrospective on DOI 10.1002/bies.201300095)

*James O. McInerney*

**So far, in eukaryotes we have not found the diversity of DNA transfer mechanisms that are found in prokaryotes.**
**Future sampling of eukaryotic genomes will undoubtedly shed more light on this controversial area.**

# HGT databases

- HGTree in prokaryotes: http://hgtree.snu.ac.kr

| Type | Number |
|---|---|
| Total non-redundant microbial genomes | 2472 |
| Genomes part of human microbiota | 30 |
| Total protein sequences | 7 748 306 |
| Number of orthologous gene sets | 154 805 |
| Detected putative HGT events | 660 840 |

**2472: 156 Archaea; 2316 Bacteria.**

Jeong H, et al. (2016). HGTree: database of horizontally transferred genes determined by tree reconciliation. Nucl. Acids Res. 2016 44: D610-D619
http://nar.oxfordjournals.org/content/44/D1/D610.abstract?etoc

# Orthologs Inference:
# Context, Methods and Limitations

Tekaia F. Inferring Orthologs: Open Questions and Perspectives (Review).
*Genomics Insights* 2016:9 17–28 doi:10.4137/GEI.S37925.

# Definitions

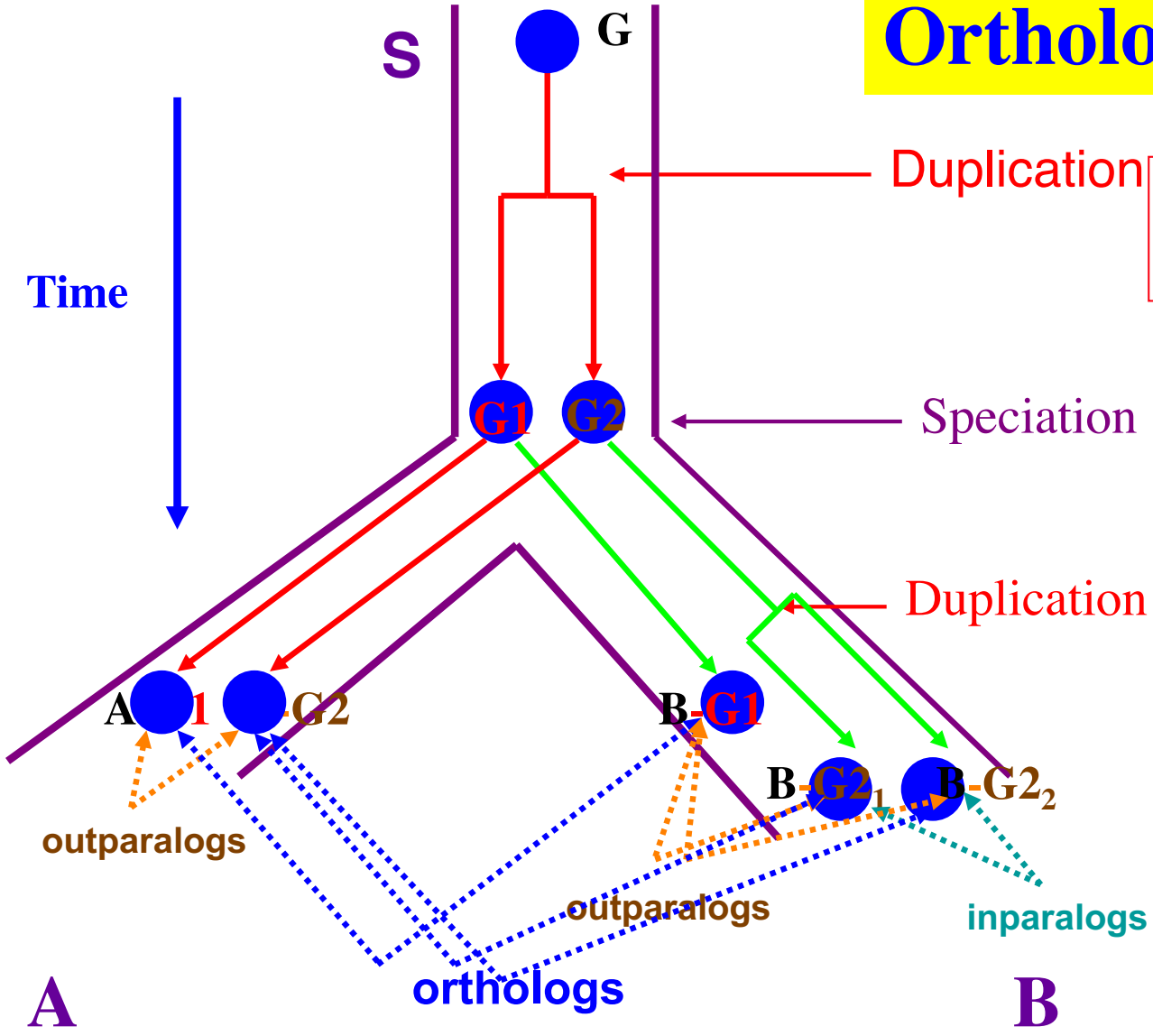
- **Orthologs are homologous genes resulting from a speciation event**

**(Fitch, 1970)**

- **Paralogs are homologous genes resulting from a duplication event**

Fitch WM. Homology a personal view on some of the problems.
*Trends Genet*. 2000;16:227-31. Review.

# Orthologs - Paralogs




Homologs: genes sharing a common origin.

Orthologs: genes in distinct species, originating from a single ancestral gene in the Last Commun Ancestor of the compared genomes.

Paralogs: genes related via duplication.


Koonin EV. 2005. Annu Rev Genet. 39:309-38.

Tekaia F. 2016. Inferring Orthologs: Open questions and perspectives. *Genome Insights*. 9: 17-28.


**Predict these events by comparing genomes?**

# Inferring Orthologs/Paralogs and classification

**Orthologs detection is of fundamental importance in:**

- **Reconstruction of the evolution of species and their genomes (Phylogenomics);**
- **Evolutionary studies of biological systems;**
- **Annotation of newly sequenced organisms;**
- **Functional genomics (transfer of functional annotation);**
- **Gene organization in a given species.**

# Orthologs / Paralogs

## Inferring orthologs is not an easy task:

**Complex combinations of lineage-specific gene duplications, losses, domains fusion, and horizontal gene transfer events often give rise to intricate evolutionary scenarios and complicated relationships when considering more than a pair of genomes.**

Koonin EV. 2005. Annu Rev Genet. 39:309-38.

Gabaldon T, Koonin EV. 2013. Nat Rev Genet. 14:360-6.

# Methods for orthology detection

**Two main approaches to determine orthologous gene classes, based either on:**

- **Sequence similarity**

  **or on**

- **Phylogeny**

# Similarity based Approaches

- **Rely on genome comparisons and clustering of highly similar genes to identify orthologous groups**

| Method | Reference | Algorithm |
|---|---|---|
| COG | Tatusov et al. *Science*. 1997; 278:631-7. | Similarity – Single linkage clustering + Constraints |
| InParanoid | Remm et al. *J Mol Biol*. 2001; 314:1041-52. | Similarity (pair-wise species) |
| MultiParanoid | idem | Extends InParanoid to multiple species |
| OrthoMCL | Li et al. *Genome Res.* 2003; 13:2178-89. | Similarity – MCL clustering algorithm |
| TribeMCL | Enright et al. *Nucleic Acids Res* 2002; 30: 1575–1584. | Similarity – MCL clustering algorithm |
| OMA | Roth et al. *BMC Bioinformatics*. 2008; 9:518. | Similarity - Global sequence alignment |
| eggNOG | Jensen et al. *Nucleic Acids Res*. 2008; 36:D250-4. | Similarity - Detects false RBH due to gene fusion and protein domain shuffling. |
| OrthoFocus | Ivliev et al. *J Bioinform Comput Biol*. 2008; 6:811-24. | Similarity – extended RBH to handle many-to-one and many-to-many relationships |
| SPO | Tekaia & Yeramian. *Gene* 2012. 492: 199-211. | Similarity – RBH – Partition – MCL clustering |

# Phylogeny-based approaches

**use candidate gene families determined by similarity and rely on the reconciliation of the phylogeny of these genes with their corresponding species phylogeny to determine the subset of orthologs**

| Method | Reference | Algorithm |
|---|---|---|
| RIO | Zmasek et al. *BMC Bioinformatics*. 2004; 3:14. | Similarity (HMMER) - bootstrap - Phylogeny |
| Orthostrapper | Storm et al. *Bioinformatics* 2002; 18: 92–99. | Phylogeny - bootstrap |
| PhIGs | Dehal et al. *BMC Bioinformatics*. 2006; 7:201. | Similarity – Multiple sequence alignments – Phylogenetic trees |
| PhyOP | Goodstadt L, Ponting CP. *PLoS Comput Biol.* 2006; 2:e133. | Similarity (overlapping limits) - phylogeny based on $d_S$ (synonymous substitution rates) |
| TreeFam | Li H etal. *Nucleic Acids Research* 2006; 34:D572-80. | Infer orthologs – paralog from the phylogenetic tree of a gene family |
| COCO-CL | Jothi et al. *Bioinformatics* 2006; 22: 779–788. | Similarity - Correlation between sequences – single linkage clustering |
| LOFT | van der Heijden et al. *BMC Bioinformatics*. 2007;8:83. | Assigns hierarchical orthology numbers to genes based on a phylogenetic tree. |
| EnsemblCompara GeneTrees | Gabaldón. *Genome Biol*. 2008; 9:235. | Clustering – multiple alignment – tree generation based on TreeBeST method |
| SYNERGY | Wapinski et al. *Bioinformatics*. 2007;23:i549-58. | Sequence similarity - species phylogeny – reconstruction of underlying gene evolutionary histories |

- **Phylogeny based methods fit best for small sets of genomes**

- **Similarity based methods fit best for large genome-data sets**

# SPO Conservation Profile

## In 8 Aspergilli species

```
SPO8.1  ASFL-ASOR-ASTE-ASNG-NEFI-ASFU-ASCL-ASNI
SPO8.1    1     1     1     1     1     1    1     1
```

1:1

```
SPO4.2  ....-ASOR-....-ASNG-NEFI-ASFU-....-....
SPO4.2    0     1     0     1     1     1    0     0
```

```
SPO15.3 ASFL-ASOR-ASTE-....-NEFI-ASFU-ASCL-ASNI
SPO15.3   1     1     1     0     4     3    3     2
```

n:m

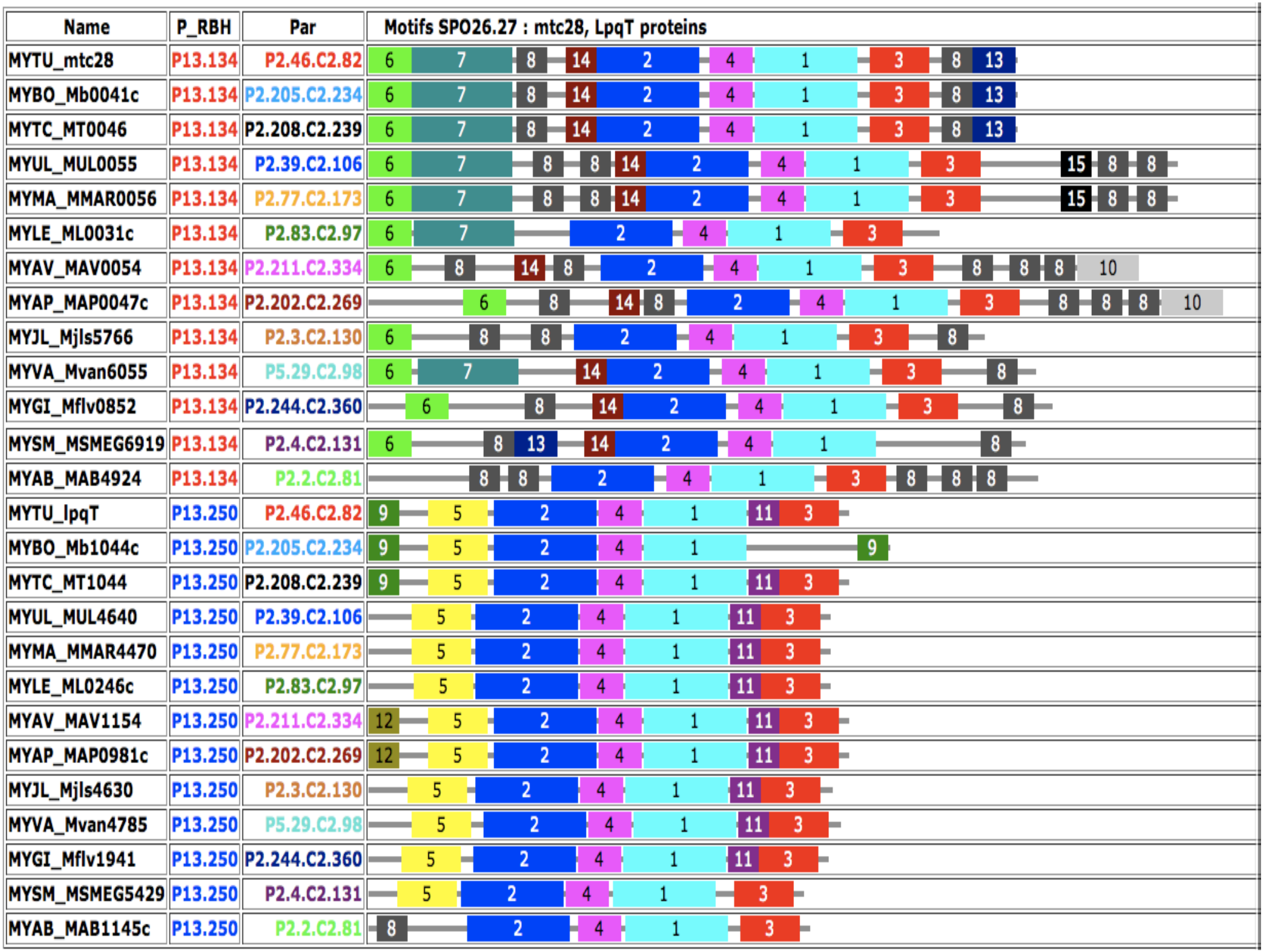
Tekaia F & Yeramian E (2012).Superpartitions: Detection and classification of orthologs. Gene; 492:199-211.

| SPO conservation profile in 13 Mycobacterial species | tot | 1:1 | n:m |
|---|---|---|---|
| Total | 3708 | 2781 | 927 |
| MT-MYBO-MTC-MYUL-MYMA-MYLE-MYAV-MYAP-MYJL-MYVA-MYGI-MYSM-MYAB | 869 | 535 | 334 |
| MT-MYBO-MTC-MYUL-MYMA-MYLE-MYAV-MYAP-MYJL-MYVA-MYGI-MYSM-.... | 21 | 13 | 8 |
| MT-MYBO-MTC-MYUL-MYMA-MYLE-MYAV-MYAP-MYJL-MYVA-MYGI-....-MYAB | 10 | 9 | 1 |
| MT-MYBO-MTC-MYUL-MYMA-MYLE-MYAV-MYAP-MYJL-MYVA-....-MYSM-MYAB | 7 | 4 | 3 |
| MT-MYBO-MTC-MYUL-MYMA-MYLE-MYAV-MYAP-MYJL-....-MYGI-MYSM-MYAB | 1 | 1 | 0 |
| MT-MYBO-MTC-MYUL-MYMA-MYLE-MYAV-MYAP-MYJL-....-....-MYSM-MYAB | 2 | 1 | 1 |
| MT-MYBO-MTC-MYUL-MYMA-MYLE-MYAV-MYAP-MYJL-....-....-MYSM-.... | 4 | 3 | 1 |
| MT-MYBO-MTC-MYUL-MYMA-MYLE-MYAV-MYAP-MYJL-....-....-....-MYAB | 1 | 0 | 1 |
| MT-MYBO-MTC-MYUL-MYMA-MYLE-MYAV-MYAP-....-MYVA-MYGI-MYSM-MYAB | 2 | 2 | 0 |
| MT-MYBO-MTC-MYUL-MYMA-MYLE-MYAV-MYAP-....-MYVA-MYGI-....-.... | 2 | 2 | 0 |
| MT-MYBO-MTC-MYUL-MYMA-MYLE-MYAV-MYAP-....-MYVA-....-....-.... | 1 | 0 | 1 |
| MT-MYBO-MTC-MYUL-MYMA-MYLE-MYAV-MYAP-....-....-MYGI-....-.... | 1 | 0 | 1 |
| MT-MYBO-MTC-MYUL-MYMA-MYLE-MYAV-MYAP-....-....-....-MYSM-MYAB | 1 | 1 | 0 |
| MT-MYBO-MTC-MYUL-MYMA-MYLE-MYAV-MYAP-....-....-....-MYSM-.... | 2 | 0 | 2 |
| MT-MYBO-MTC-MYUL-MYMA-MYLE-MYAV-MYAP-....-....-....-....-MYAB | 1 | 1 | 0 |
| MT-MYBO-MTC-MYUL-MYMA-MYLE-MYAV-MYAP-....-....-....-....-.... | 8 | 5 | 3 |
| MT-MYBO-MTC-MYUL-MYMA-MYLE-MYAV-....-MYJL-MYVA-MYGI-MYSM-MYAB | 10 | 10 | 0 |
| MT-MYBO-MTC-MYUL-MYMA-MYLE-MYAV-....-MYJL-MYVA-MYGI-MYSM-.... | 1 | 1 | 0 |
| MT-MYBO-MTC-MYUL-MYMA-MYLE-MYAV-....-MYJL-MYVA-....-MYSM-MYAB | 1 | 1 | 0 |
| MT-MYBO-MTC-MYUL-MYMA-MYLE-MYAV-....-....-....-....-....-.... | 1 | 0 | 1 |
| MT-MYBO-MTC-MYUL-MYMA-MYLE-....-MYAP-MYJL-MYVA-MYGI-MYSM-MYAB | 7 | 6 | 1 |
| MT-MYBO-MTC-MYUL-MYMA-MYLE-....-MYAP-MYJL-MYVA-MYGI-....-MYAB | 1 | 1 | 0 |
| MT-MYBO-MTC-MYUL-MYMA-MYLE-....-MYAP-....-....-....-....-MYAB | 1 | 1 | 0 |

# SPOs Conservation Profiles

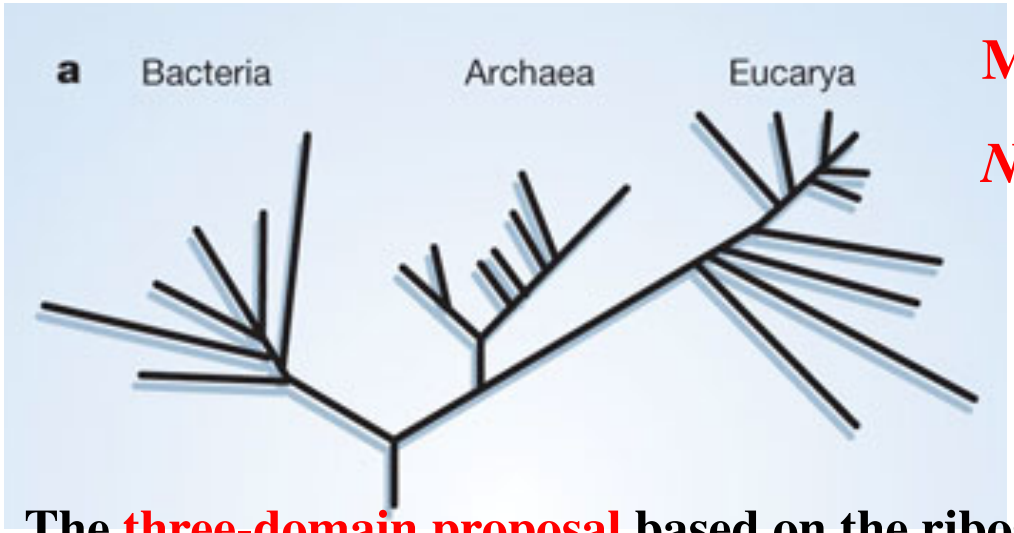
| Sets of species | 13 *Chlamydiae* | 13 *Mycobacteria* | 8 *Aspergilli* |
|---|---|---|---|
| a) Partitions of RBHs | 1202 | 7560 | 11887 |
| b) SPOs | 948 | 3708 | 6192 |
| c) Distinct Cons. Prof. | 73 | 414 | 213 |
| → d) 1:1 Cons. Prof. | 823 (86.8%) | 2780 (75.0%) | 4837 (78.1%) |
| → e) n:m Cons. Prof. | 125 (13.2%) | 928 (25.0%) | 1355 (21.9%) |
| f) SPOs of type n:m contain. in-paralogs | 35 (28.0%) | 535 (57.7%) | 505 (37.3%) |
| g) SPOs with proteins from each species | 522 (55.1%) | 869 (23.4%) | 2474 (40.0%) |
| h) SPOs with exactly one protein from each species | 439 (46.3%) | 535 (14.4%) | 1597 (25.8%) |

**# 1:1 orth clusters >> # n:m orth clusters**


Name
P_RBH
Par
Motifs SPO26.27 : mtc28, LpqT proteins
MYTU_mtc28 P13.134 P2.46.C2.82
MYBO_Mb0041c P13.134 P2.205.C2.234
MYTC_MT0046 P13.134 P2.208.C2.239
MYUL_MUL0055 P13.134 P2.39.C2.106
MYMA_MMAR0056 P13.134 P2.77.C2.173
MYLE_ML0031c P13.134 P2.83.C2.97
MYAV_MAV0054 P13.134 P2.211.C2.334
MYAP_MAP0047c P13.134 P2.202.C2.269
MYJL_Mjls5766 P13.134 P2.3.C2.130
MYVA_Mvan6055 P13.134 P5.29.C2.98
MYGI_Mflv0852 P13.134 P2.244.C2.360
MYSM_MSMEG6919 P13.134 P2.4.C2.131
MYAB_MAB4924 P13.134 P2.2.C2.81
MYTU_lpqT P13.250 P2.46.C2.82
MYBO_Mb1044c P13.250 P2.205.C2.234
MYTC_MT1044 P13.250 P2.208.C2.239
MYUL_MUL4640 P13.250 P2.39.C2.106
MYMA_MMAR4470 P13.250 P2.77.C2.173
MYLE_ML0246c P13.250 P2.83.C2.97
MYAV_MAV1154 P13.250 P2.211.C2.334
MYAP_MAP0981c P13.250 P2.202.C2.269
MYJL_Mjls4630 P13.250 P2.3.C2.130
MYVA_Mvan4785 P13.250 P5.29.C2.98
MYGI_Mflv1941 P13.250 P2.244.C2.360
MYSM_MSMEG5429 P13.250 P2.4.C2.131
MYAB_MAB1145c P13.250 P2.2.C2.81

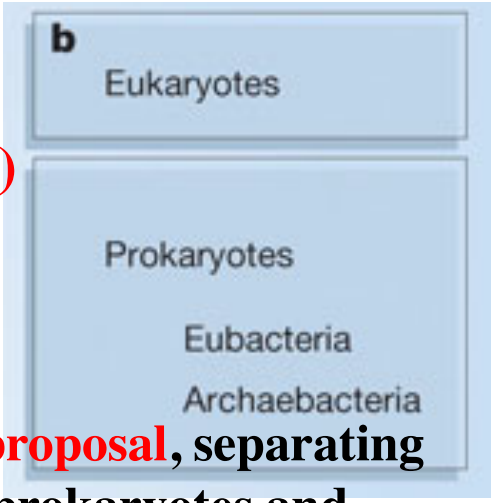
# Genome Trees - Phylogenomics

➔ **Project presentation: Friday afternoon**



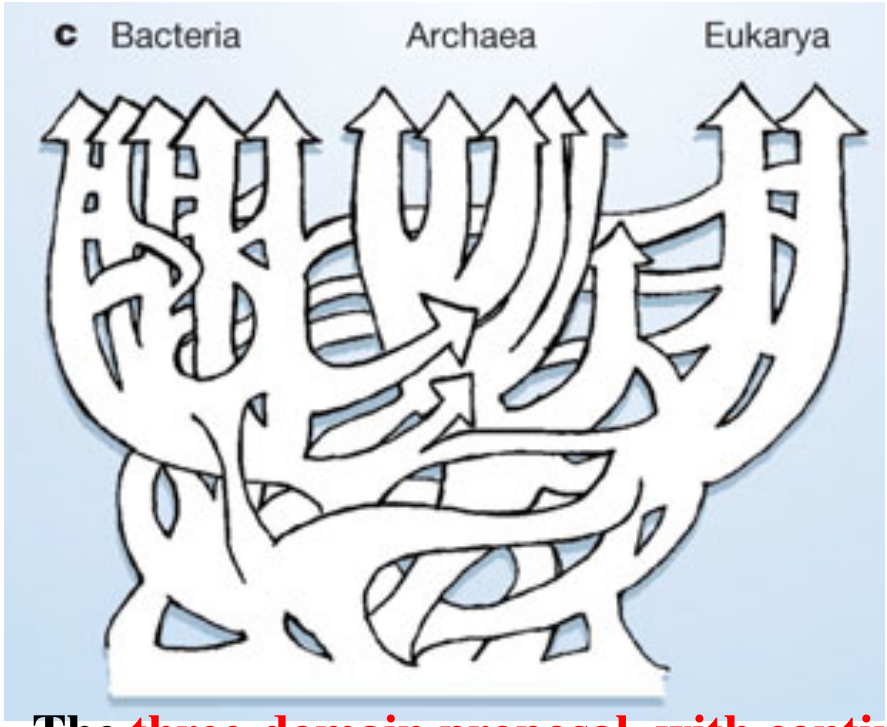

**Martin & Embley**

***Nature* 431:152-5.(2004)**

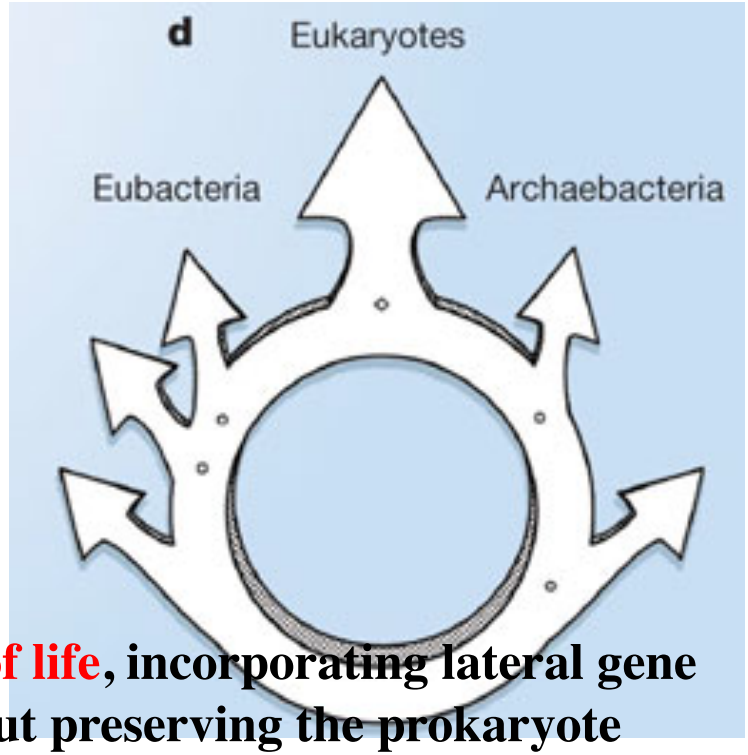



**The three-domain proposal based on the ribosomal RNA tree.** ***Woese et al. PNAS. 87:4576-4579. (1990)***

**The two-empire proposal, separating eukaryotes from prokaryotes and eubacteria from archaebacteria.** ***Mayr, D PNAS 95:9720-23. (1998).***




**The three-domain proposal, with continuous lateral gene transfer among domains.**
***Doolittle. Science 284:2124-8. (1999)***




**The ring of life, incorporating lateral gene transfer but preserving the prokaryote eukaryote divide.**
***Rivera & Lake JA. Nature 431: 152-5. (2004)***

# Ancestral duplication and ancestral conservation

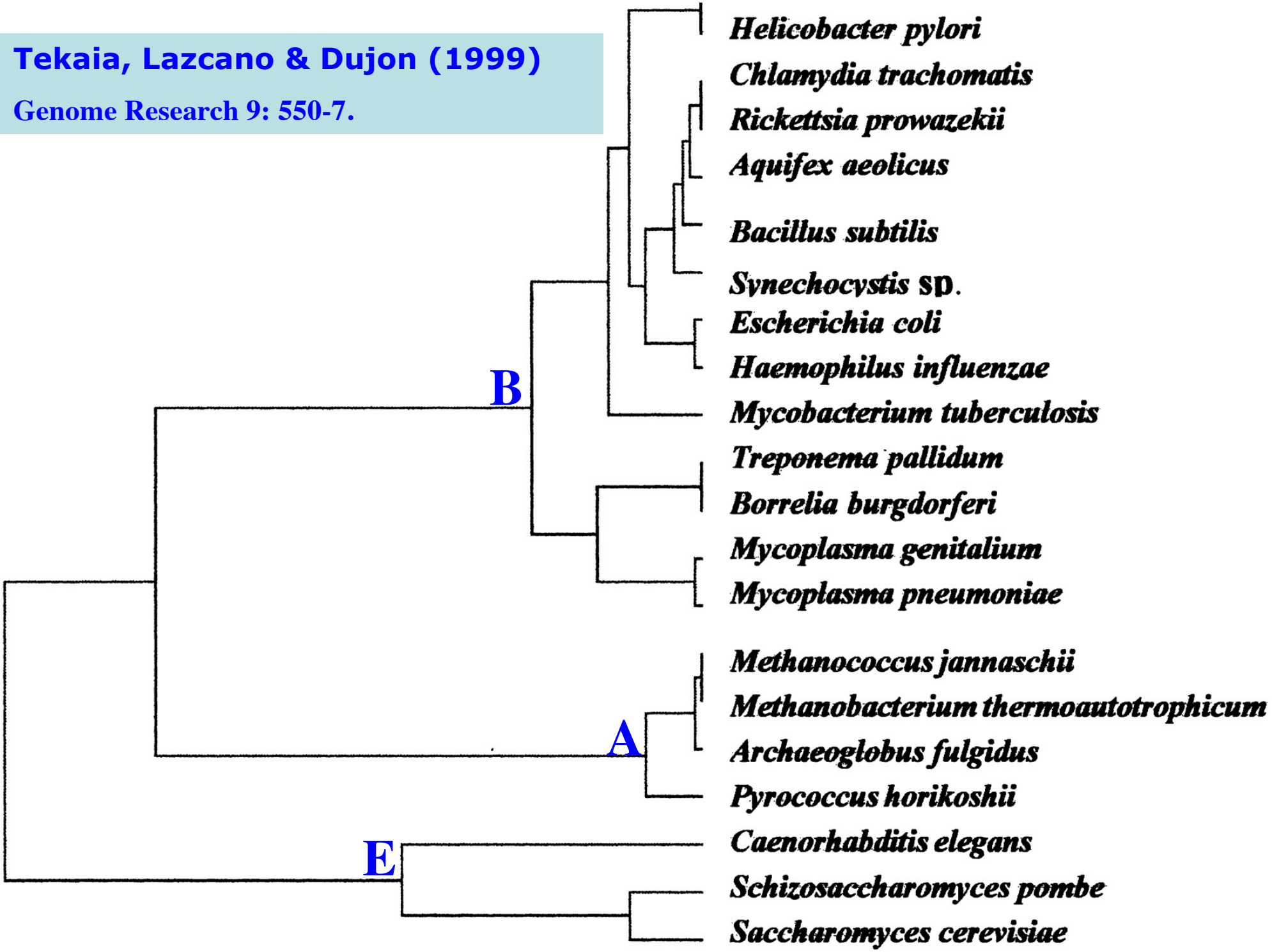
| org | SC | SP | CE | DM | AG | CA | ATH | HS | MUS | FR | PF | ECUN |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SC | 40.5 | 63.9 | 17.5 | 27.1 | 22.3 | 65.9 | 23.4 | 22.9 | 27.3 | 18.0 | 22.5 | 35.8 |
| SP | 58.4 | 37.4 | 18.8 | 29.3 | 26.3 | 54.3 | 25.0 | 25.0 | 29.6 | 20.0 | 24.6 | 38.4 |
| CE | 38.1 | 46.6 | 65.2 | 51.9 | 50.6 | 35.5 | 27.5 | 44.6 | 54.4 | 42.4 | 24.8 | 34.8 |
| DM | 40.5 | 50.2 | 39.2 | 65.8 | 69.9 | 37.5 | 29.5 | 50.3 | 62.7 | 47.9 | 26.5 | 36.3 |
| AG | 40.9 | 50.2 | 39.8 | 73.1 | 59.5 | 38.0 | 30.6 | 50.2 | 60.3 | 48.7 | 26.5 | 36.0 |
| CA | 71.8 | 65.5 | 18.4 | 27.7 | 25.7 | 35.8 | 24.3 | 23.2 | 27.8 | 18.5 | 22.3 | 35.7 |
| ATH | 40.3 | 47.8 | 21.7 | 31.5 | 30.3 | 37.0 | 83.6 | 25.6 | 29.7 | 21.9 | 26.2 | 33.4 |
| HS | 43.0 | 53.3 | 40.0 | 61.3 | 54.5 | 39.7 | 32.1 | 66.7 | 90.8 | 68.8 | 28.2 | 37.7 |
| MUS | 41.7 | 52.5 | 39.5 | 62.1 | 54.7 | 39.1 | 31.5 | 76.8 | 77.8 | 67.7 | 27.6 | 37.2 |
| FR | 42.0 | 52.6 | 40.0 | 60.7 | 59.9 | 39.5 | 32.7 | 68.7 | 81.8 | 63.4 | 27.6 | 37.4 |
| PF | 25.9 | 31.2 | 13.1 | 19.3 | 15.9 | 22.2 | 16.3 | 17.2 | 21.0 | 13.2 | 28.3 | 28.9 |
| ECUN | 19.5 | 23.4 | 8.9 | 13.1 | 10.8 | 16.2 | 11.4 | 12.0 | 15.2 | 9.0 | 13.6 | 26.1 |
| MJ | 11.5 | 13.3 | 4.9 | 6.7 | 6.0 | 10.2 | 6.0 | 4.8 | 5.6 | 3.7 | 8.7 | 15.4 |
| MTH | 13.6 | 16.2 | 4.6 | 7.4 | 7.6 | 11.2 | 8.0 | 5.1 | 6.1 | 4.0 | 8.3 | 15.2 |
| AF | 14.4 | 16.5 | 5.9 | 8.2 | 8.7 | 11.8 | 8.7 | 5.6 | 6.6 | 4.5 | 8.6 | 15.4 |
| PH | 16.3 | 18.7 | 5.0 | 7.1 | 9.2 | 11.1 | 9.7 | 5.2 | 6.0 | 4.1 | 7.9 | 15.3 |
| PA | 14.3 | 15.2 | 5.4 | 7.5 | 7.3 | 11.9 | 7.4 | 5.5 | 6.4 | 4.3 | 8.3 | 15.9 |
| APEM | 15.5 | 20.1 | 4.8 | 7.3 | 10.6 | 10.3 | 9.4 | 5.2 | 5.9 | 3.9 | 7.2 | 14.9 |
| TA | 15.2 | 17.5 | 5.9 | 8.3 | 8.3 | 12.7 | 8.2 | 5.3 | 6.3 | 4.2 | 8.6 | 14.8 |
| TV | 15.4 | 17.8 | 6.2 | 8.3 | 8.7 | 13.3 | 8.3 | 5.6 | 6.8 | 4.4 | 8.7 | 15.0 |
| H | 14.8 | 17.7 | 5.8 | 8.3 | 9.8 | 12.0 | 10.2 | 5.5 | 6.6 | 4.5 | 8.0 | 13.9 |
| SSP2 | 16.7 | 19.4 | 7.1 | 9.1 | 9.4 | 14.2 | 9.5 | 6.2 | 7.4 | 4.9 | 9.5 | 15.9 |
| PFU | 17.0 | 22.8 | 6.5 | 9.3 | 11.1 | 13.3 | 12.3 | 7.0 | 8.0 | 5.6 | 9.1 | 17.1 |
| STO | 18.6 | 23.1 | 6.8 | 8.6 | 11.4 | 13.7 | 11.1 | 5.9 | 7.1 | 4.5 | 9.1 | 15.7 |
| PYAE | 15.6 | 19.5 | 5.3 | 8.2 | 9.9 | 11.8 | 9.5 | 5.8 | 6.9 | 4.5 | 8.1 | 15.0 |
| MA | 16.0 | 18.9 | 7.1 | 10.8 | 12.5 | 14.7 | 9.7 | 7.4 | 8.7 | 6.4 | 9.8 | 17.0 |
| MK | 13.0 | 14.6 | 4.0 | 6.2 | 6.1 | 10.7 | 6.9 | 4.6 | 5.4 | 3.5 | 7.3 | 14.1 |
| MMA | 14.8 | 17.4 | 6.4 | 9.2 | 9.5 | 13.5 | 8.1 | 6.6 | 7.9 | 5.3 | 9.7 | 15.8 |
| HI | 13.0 | 14.3 | 4.8 | 7.3 | 8.5 | 11.1 | 8.7 | 4.4 | 5.4 | 4.0 | 8.2 | 8.7 |
| …. . | | | | | | | | | | | | |

$W_{ij}$

- genome similarity
- shared orthologs

**Tekaia, Lazcano & Dujon (1999)**

**Genome Research 9: 550-7.**



# Conservation profiles

- **99 species (B: 33; A: 19; E:27); 541880 proteins**

```
p 011111100011111111100011011011110100111111011111
```

- **A "conservation profile" is an n-component binary vector describing a protein conservation pattern across n species.**

**Components are 0 and 1, following absence or presence of homologs.**

**Main interesting properties of conservation profiles:**

- **Conservation profiles are signatures of evolutionary relationships;**

- **A conservation profile is the trace of protein evolutionary histories jointly captured in a set of n species (multidimensional feature);**

Tekaia F, Yeramian E. (2005). *PLoS Comput Biol*.1(7):e75

# Protein conservation profiles


**Table : 541880 proteins x 99 species**

- **Different conservation profiles represent different evolutionary histories**

# Genome trees: data matrices

$T = \{T_{ij}$ ; i=1,n; j=1,n; n is the number of surveyed species$\}$

$T_{ij}$ is the overall similarity score between species j and i: conservation; shared orthologs,...

- **Jaccard similarity scores**

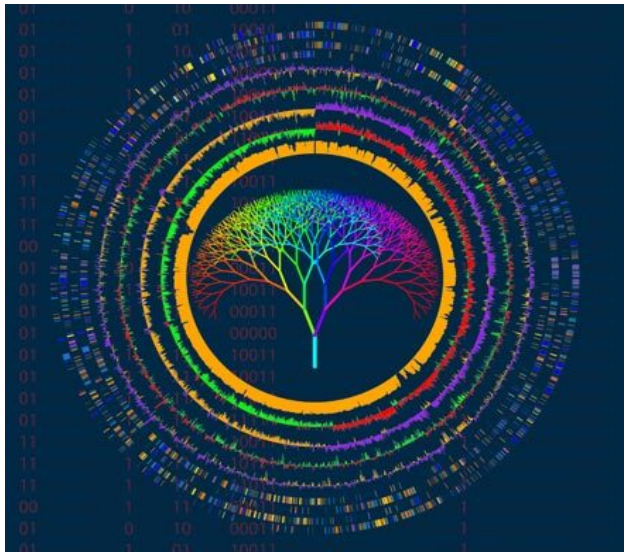
$\{ s_{ij} = N_{11}/(N_{11}+N_{01}+N_{10}); \}$

$\{s_{ij}$ = (Jaccard similarity score between species i and j) $\}$

$T = \{ T_{ij} = 100*s_{ij}$ ; i=1,n; j=1,n$\}$

184130 distinct conservation profiles

# *profiles tree*

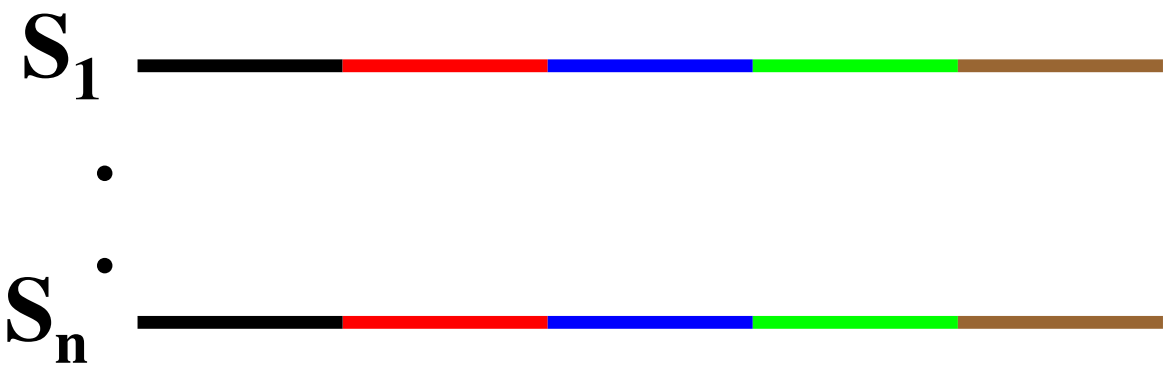





**Tekaia F, Yeramian E. (2005).**
***PLoS Comput Biol*.1(7):e75**

# • Phylogenomic tree

**(based on concatenation of a gene sample common to the considered species)**



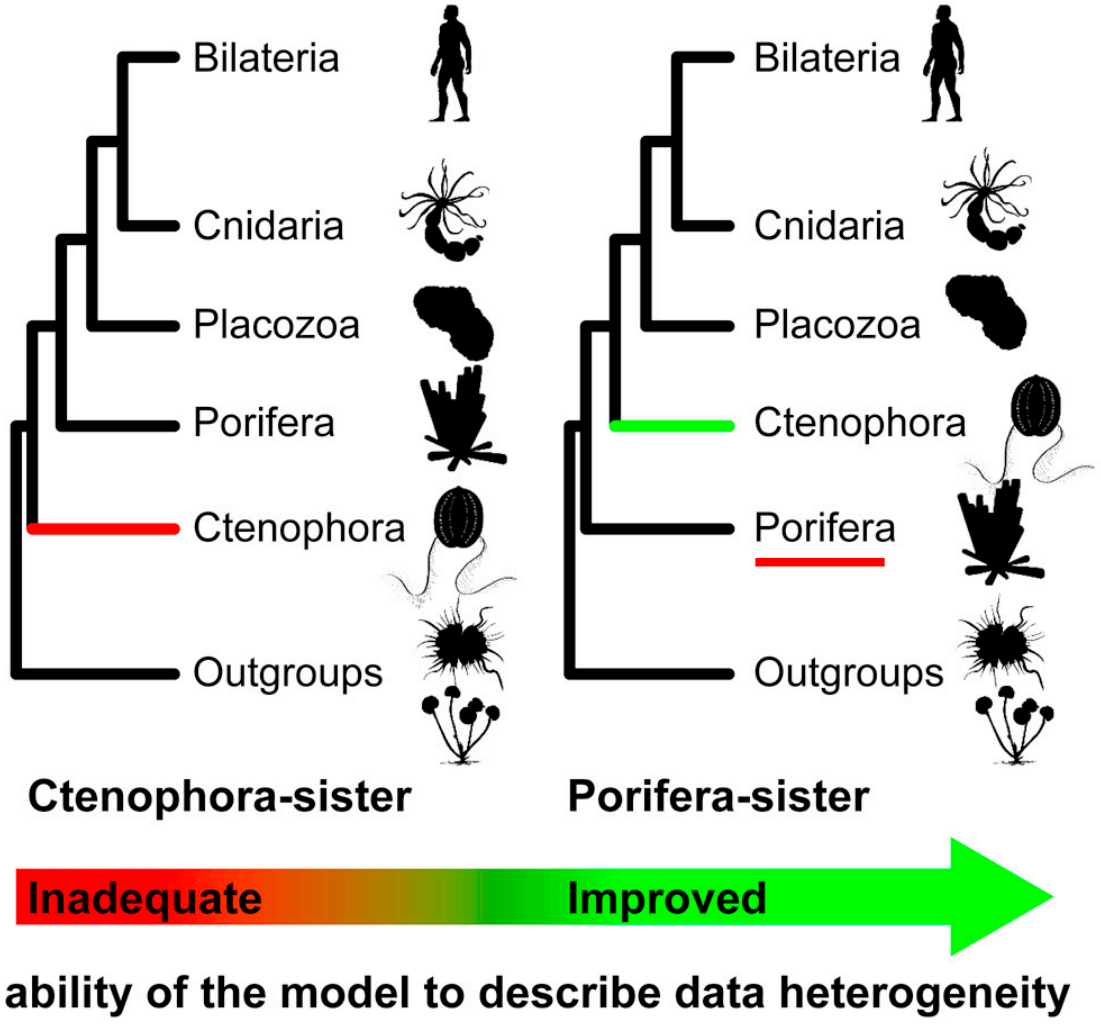

**Difficulties:**

- **genes don't evolve at the same rate nor in the same way;**
- **a limited number of genes are shared among all species;**

**Dagan and Martin (2006). The tree of one percent.** *Genome Biology*, **7:**118.

# Sponges are the sister group of the remaining animals

**Animal relationships depend on the ability of the models to describe the data**



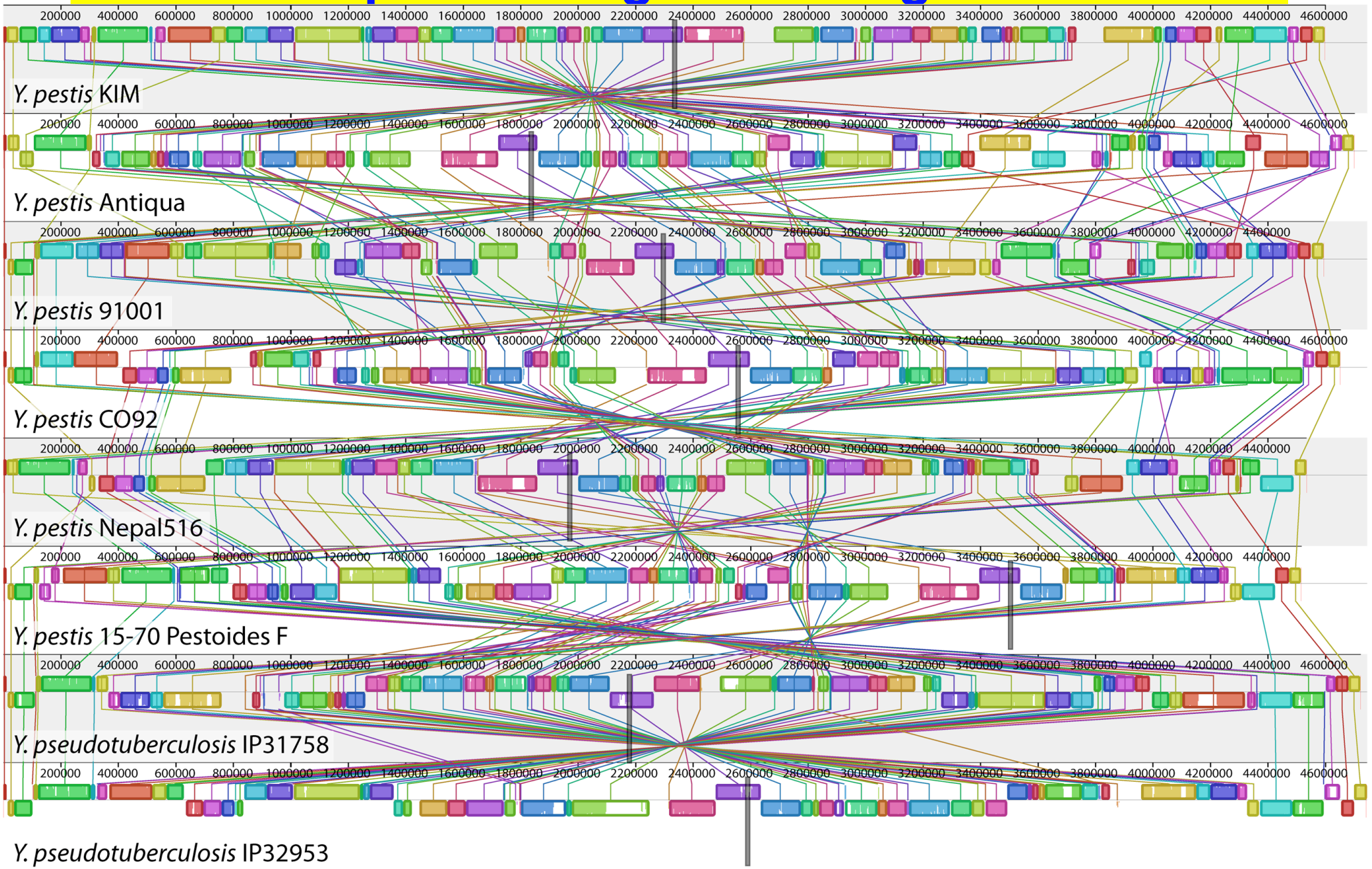

*Current Biology* DOI: (10.1016/j.cub.2017.11.008)

Feuda et al. (2017).Improved Modeling of Compositional Heterogeneity Supports Sponges as Sister to All Other Animals. Current Biology 27, 1–7.

# Whole genome alignments

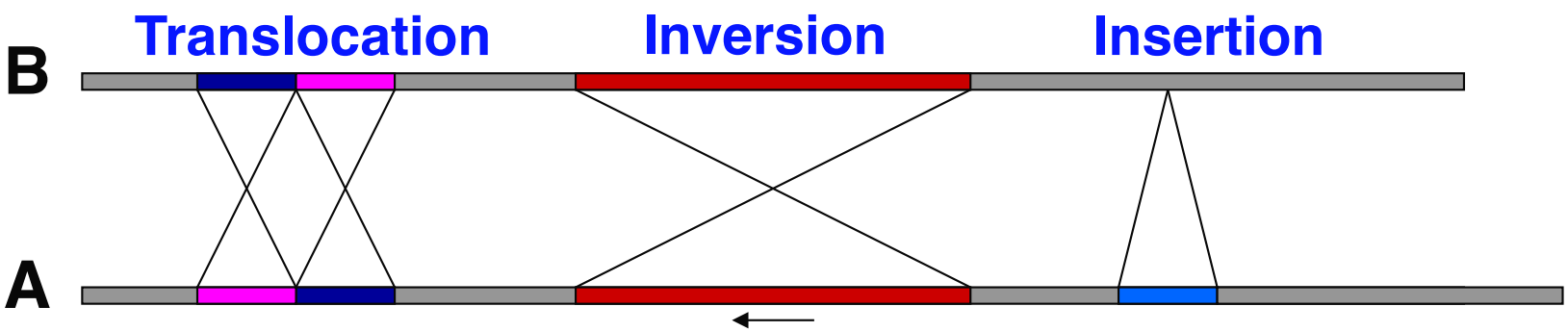
➔ **Project presentation: Thursday afternoon**

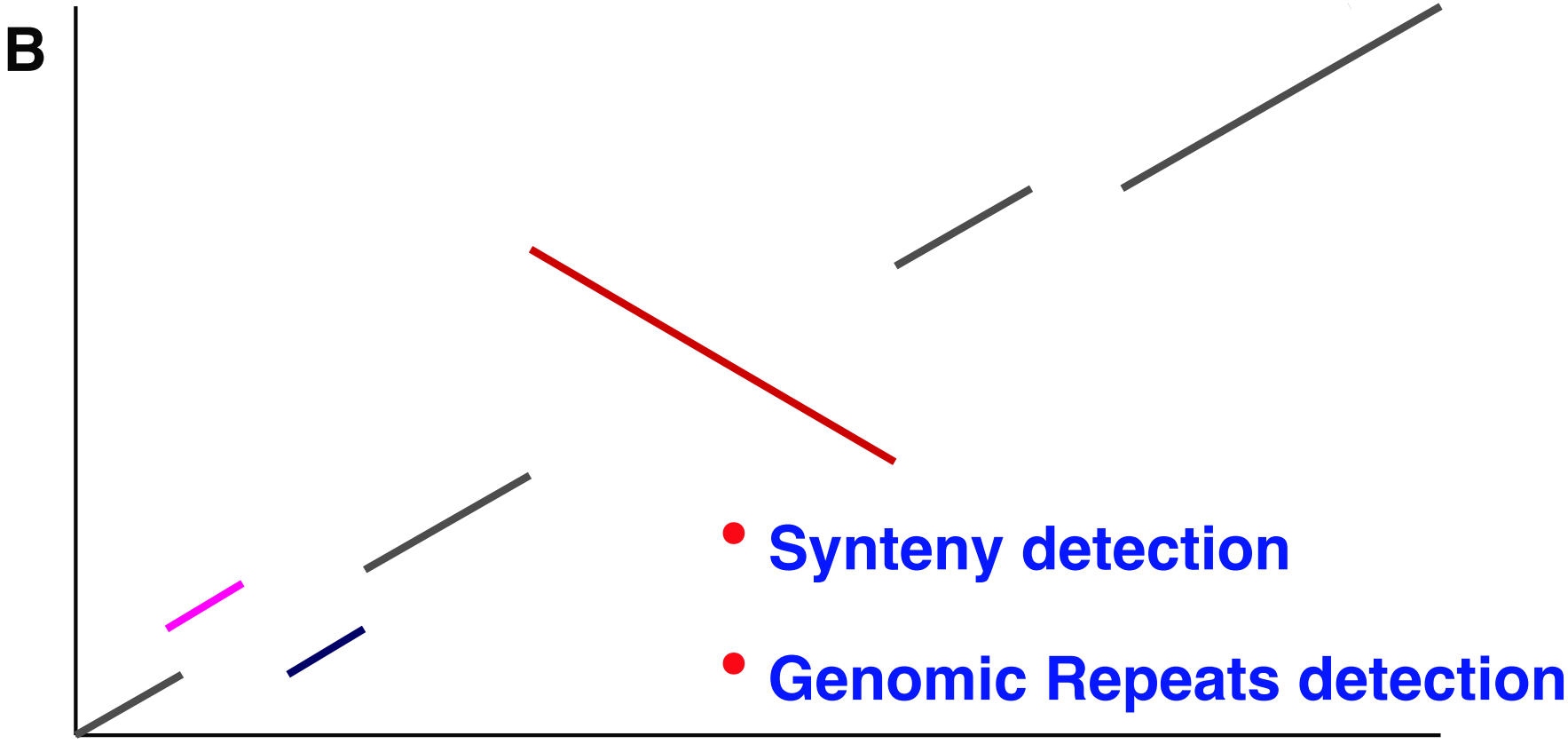
# Multiple whole genome alignments




Aaron E. Darling, Istvan Miklos, Mark A. Ragan (2008). **Dynamics of Genome Rearrangement in Bacterial Populations.** **PLoS Genet. 2008 July; 4(7): e1000128.**

# Whole Alignment of two genomes




- Synteny detection
- Genomic Repeats detection

# Whole genome alignments

- **MUMmer (Maximal Unique Matcher)**

http://mummer.sourceforge.net/

**mauve**

**Multiple Genome Alignment**

http://gel.ahabs.wisc.edu/mauve/

# Whole genome alignments

- LAGAN
- Multi-LAGAN

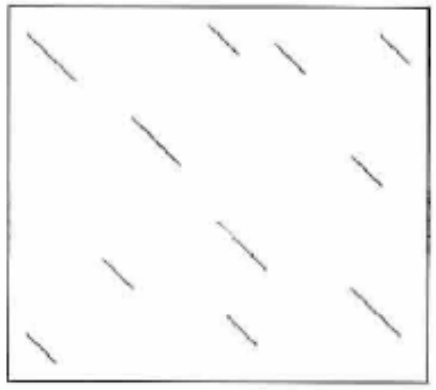
http://lagan.stanford.edu/lagan_web/index.shtml

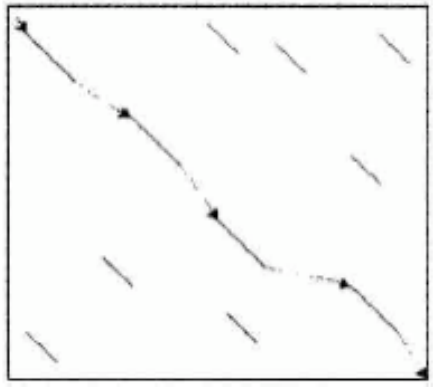
**Three main steps:**

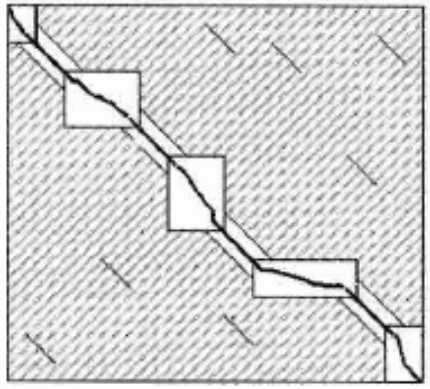
1.Generation of local alignments.

2. Construction of a rough global map.

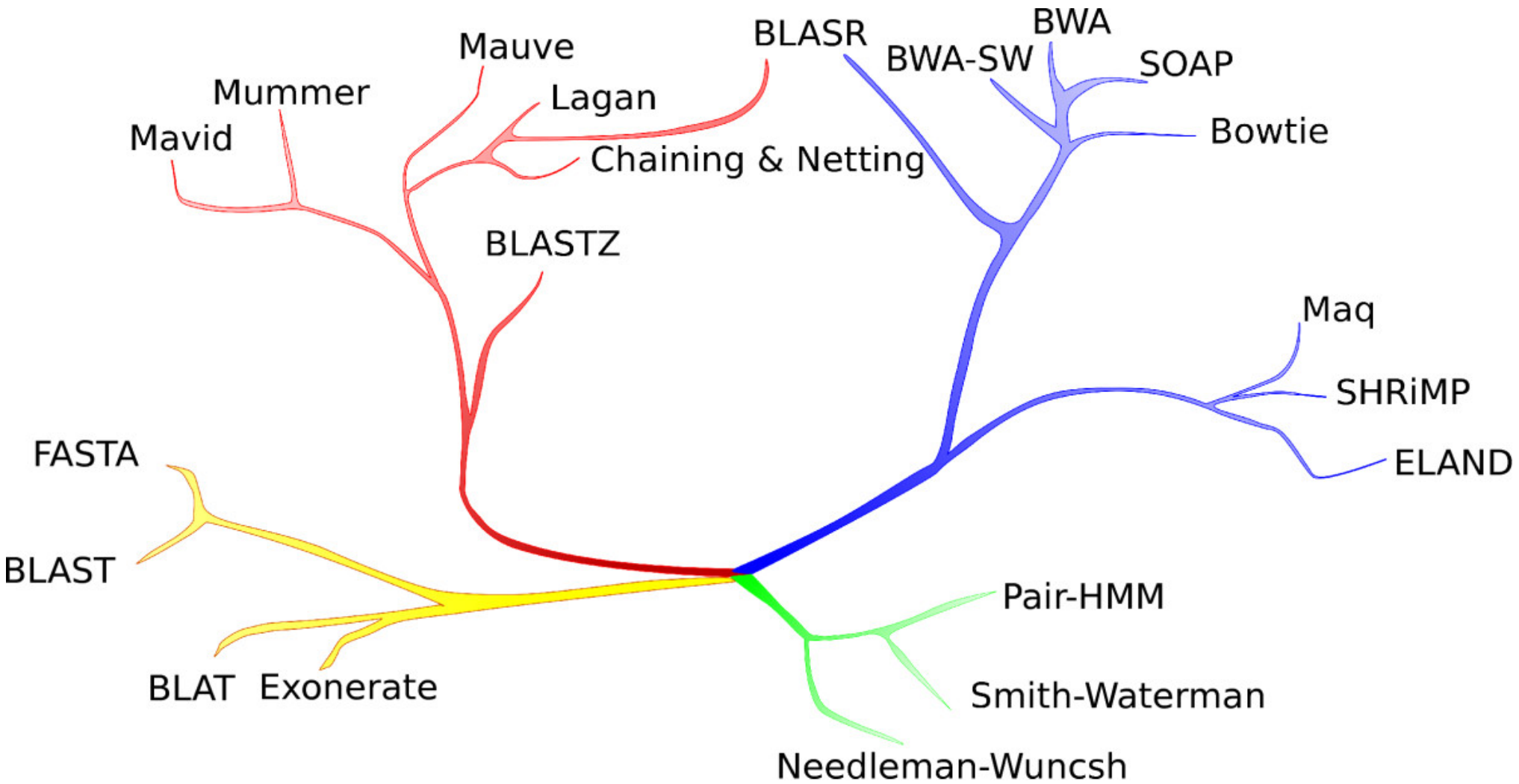
3. Computation of the final global alignment.







# An illustration of relationships between alignment methods




**The applications / corresponding computational restrictions shown are (green) short pairwise alignment / detailed edit model; (yellow) database search / divergent homology detection; (red) whole genome alignment / alignment of long sequences with structural rearrangements; and (blue) short read mapping / rapid alignment of massive numbers of short sequences. Although solely illustrative, methods with more similar data structures or algorithmic approaches are on closer branches. The BLASR method combines data structures from short read alignment with optimization methods from whole genome alignment.**

Chaisson and Tesler (2012). Mapping single molecule sequencing reads using basic local alignment with successive refinement: Application and Theory. *BMC Bioinformatics* 13:238

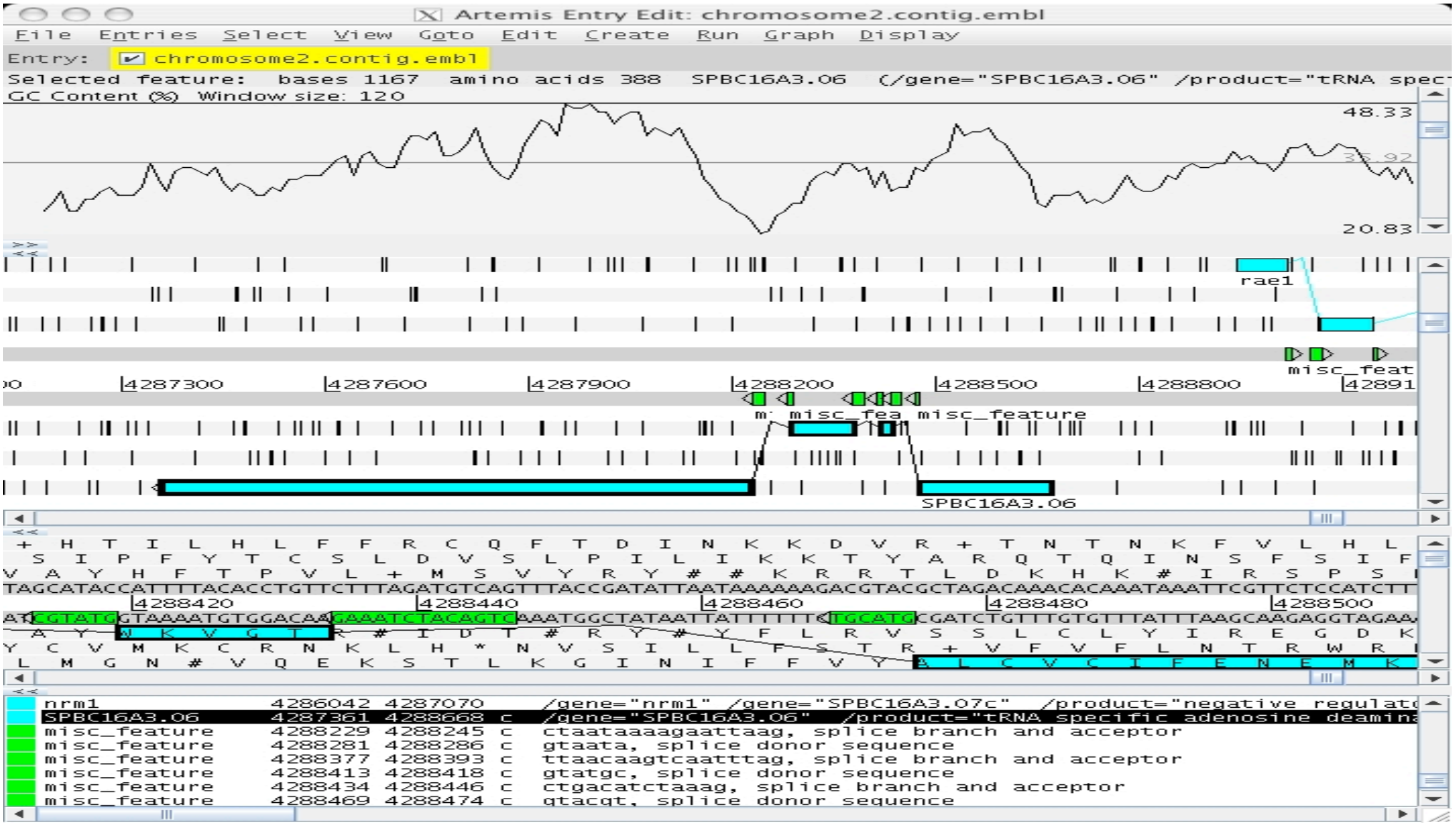
# Visualization tools

# Genomic data visualization with CIRCOS

## http://circos.ca/

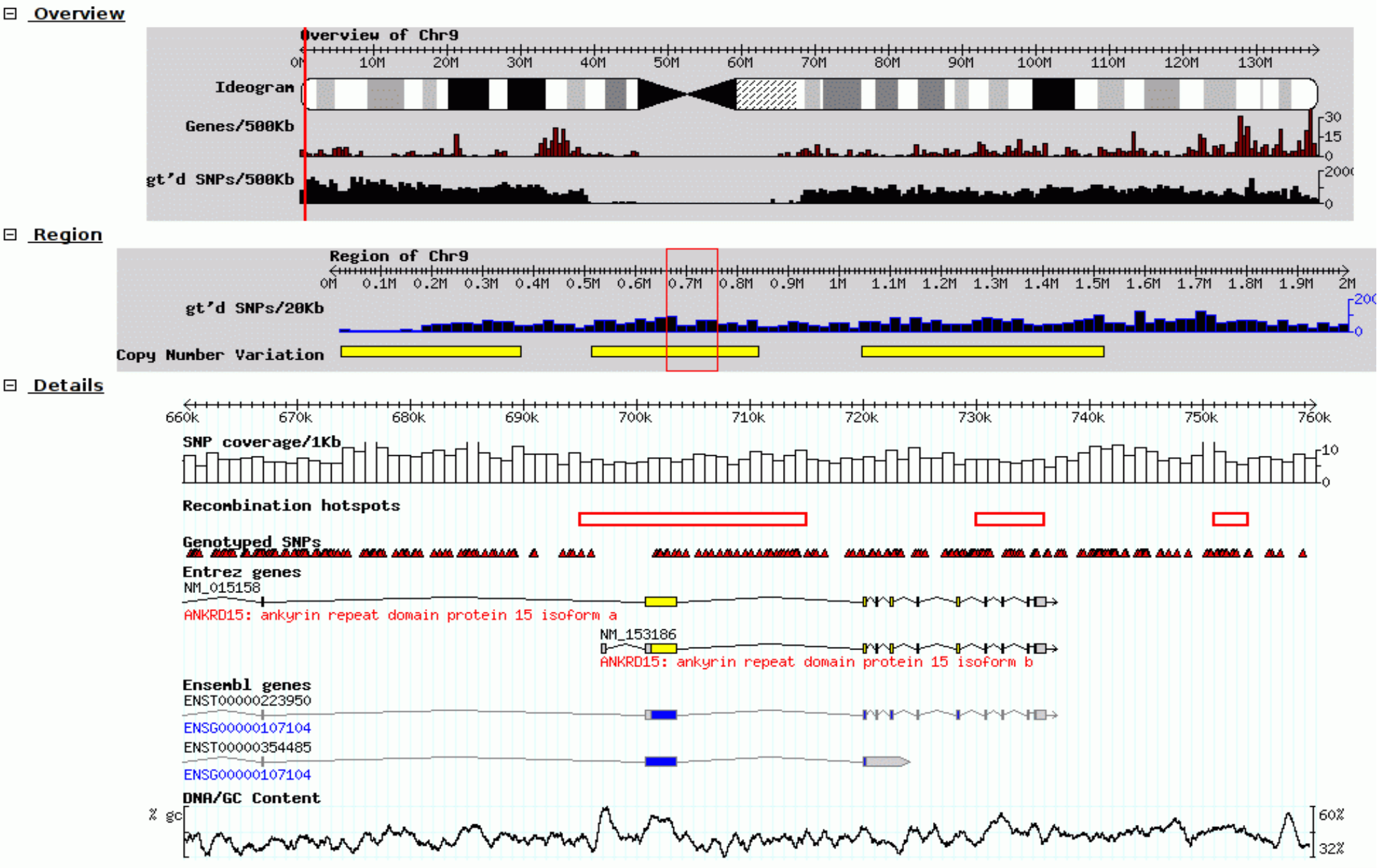
# Artemis: Genome Browser and Annotation Tool

http://www.sanger.ac.uk/resources/software/artemis/



# Gbrowse

# http://gmod.org/wiki/Gbrowse



# Next-generation sequencing technologies

➔ **Project presentation: Monday am**

# Next-generation sequencing technologies

**Allowed a massive increase in available raw sequence data**

**Implying a number of new informatics, computational challenges and difficulties.**

- **New class of bioinformatics methods and tools**

**http://bioinformaticsonline.com/pages/view/26617/list-of-bioinformatics-software-tools-for-next-generation-sequencing**

Curr Opin Biotechnol. 2012 Feb;23(1):9-15. doi: 10.1016/j.copbio.2011.11.013. Epub 2011 Dec 9.**Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis.**Scholz MB, Lo CC, Chain PS.

# Next-generation sequencing technologies

- **Next-Generation Sequencing (NGS) has evolved into a popular strategy for genotyping.**

- **The advantage of the NGS approach also include higher coverage and resolution, more accurate estimation of copy numbers, more precise detection of breakpoints, and higher capabilities to identify novel CNVs.**

- **High coverage, read pairs, long reads**

# A variety of applications:

- **Genomes: *de novo* or re-sequencing;**
- **Transcriptome sequencing (RNA-Seq);**
- **Cancer genomics;**
- **Human genetic variation analyses:**

**-Single Nucleotide Polymorphism (SNP), - Small Insertion/Deletion (Indels), Structural variations (SV), large structural variants;**

- **Disease-targeted sequencing;**
- **Ancient genome;**
- **Metagenomics;**
- **Epidemiology;**
- **Clinical applications (therapeutic decisions);...**

*****Papers on Application of Next-Generation Sequencing:
http://www.nature.com/nrg/series/nextgeneration/index.html ******

# Where can we find Genome data?

## Genomes Resources

**Nucleic Acids Research Table of Contents Alert Database issue Vol. 45, No. D1, 4 January 2017**

http://nar.oxfordjournals.org/content/45/D1?etoc

# Resources for genomes

**There are two main archive resources for genomes:**

**ncbi** **National Center for Biotechnology Information**
http://www.ncbi.nlm.nih.gov/sites/genome

**ebi** **European Bioinformatics Institute**
http://www.ebi.ac.uk/genomes/

**Many others resources are available from sequencing Institutions including:**

**Sanger** **The welcome Trust Sanger Institut**
http://www.sanger.ac.uk/resources/downloads/

**Ensembl** **http://www.ensembl.org/index.html**

**Broad Institut** http://www.broadinstitute.org/science/data

**Genomes News Network** http://www.genomenewsnetwork.org/

**…and many other sites for specific projects**

# Genomic Variation

## Sequence versus Structural Variation

## http://www.ensembl.org/info/genome/variation/

# Genomic Variation

**Includes:**

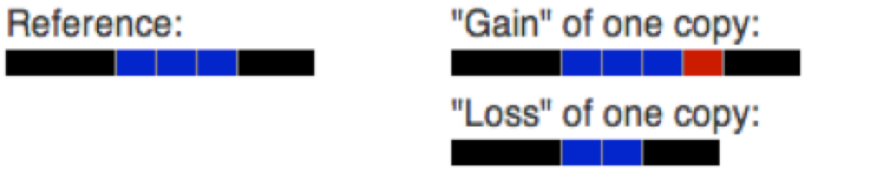
- **Single nucleotide variants**
- **Large structural variants**

➔ **Project Presentation: Monday am**

# Sequence variants:

| Type | Description | Example (Reference / Alternative) |
|---|---|---|
| SNP | Single Nucleotide Polymorphism | Ref: ...TTGACGTA... Alt: ...TTGGCGTA... |
| Insertion | Insertion of one or several nucleotides | Ref: ...TTGACGTA... Alt: ...TTGATGCGTA... |
| Deletion | Deletion of one or several nucleotides | Ref: ...TTGACGTA... Alt: ...TTGGTA... |
| Indel | An insertion and a deletion, affecting 2 or more nucleotides | Ref: ...TTGACGTA... Alt: ...TTGGCTCGTA... |
| Substitution | A sequence alteration where the length of the change in the variant is the same as that of the reference. | Ref: ...TTGACGTA... Alt: ...TTGTAGTA... |

# Structural variants (>= 50 bases):

| Type | Description | Example (Reference / Alternative) |
|---|---|---|
| CNV | Copy Number Variation: increases or decreases the copy number of a given region |   |
| Inversion | A continuous nucleotide sequence is inverted in the same position |   |
| Translocation | A region of nucleotide sequence that has translocated to a new position |   |

- The full list of variants in Ensembl can be found at this link:

http://www.ensembl.org/info/genome/variation/data_description.html#classes

# 1000 Genomes Project

**2015**

**1 October**


## A global reference for human genetic variation

The 1000 Genomes Project Consortium*

**An integrated map of structural variation in 2504 human genomes from 26 global populations**

**Revealed:**

- **84.7 million single-nucleotide variants (SNVs);**
- **3.6 million insertion/deletion (indels) variants;**
- **more than 60000 structural variants (SVs) (ie based on event size >= 15bp )**

*Sudmant et al. Nature* 526, 75-81 (01 October 2015) doi:10.1038/nature15394

# Resources for mining Variants

## Cancer:

| Database | Link |
| --- | --- |
| International Cancer Genome Consortium (ICGC) | icgc.org |
| Catalogue of Somatic Mutation in Cancer (COSMIC) | cancer.sanger.ac.uk |
| cBioPortal for Cancer Genomics | cbioportal.org |
| Cancer Cell Line Encyclopedia (CCLE) | broadinstitute.org/ccle |

## Plants:

- 1001genomes.org (A. thaliana)
- www.onekp.org (1000 plants genomes)

# Metagenomics

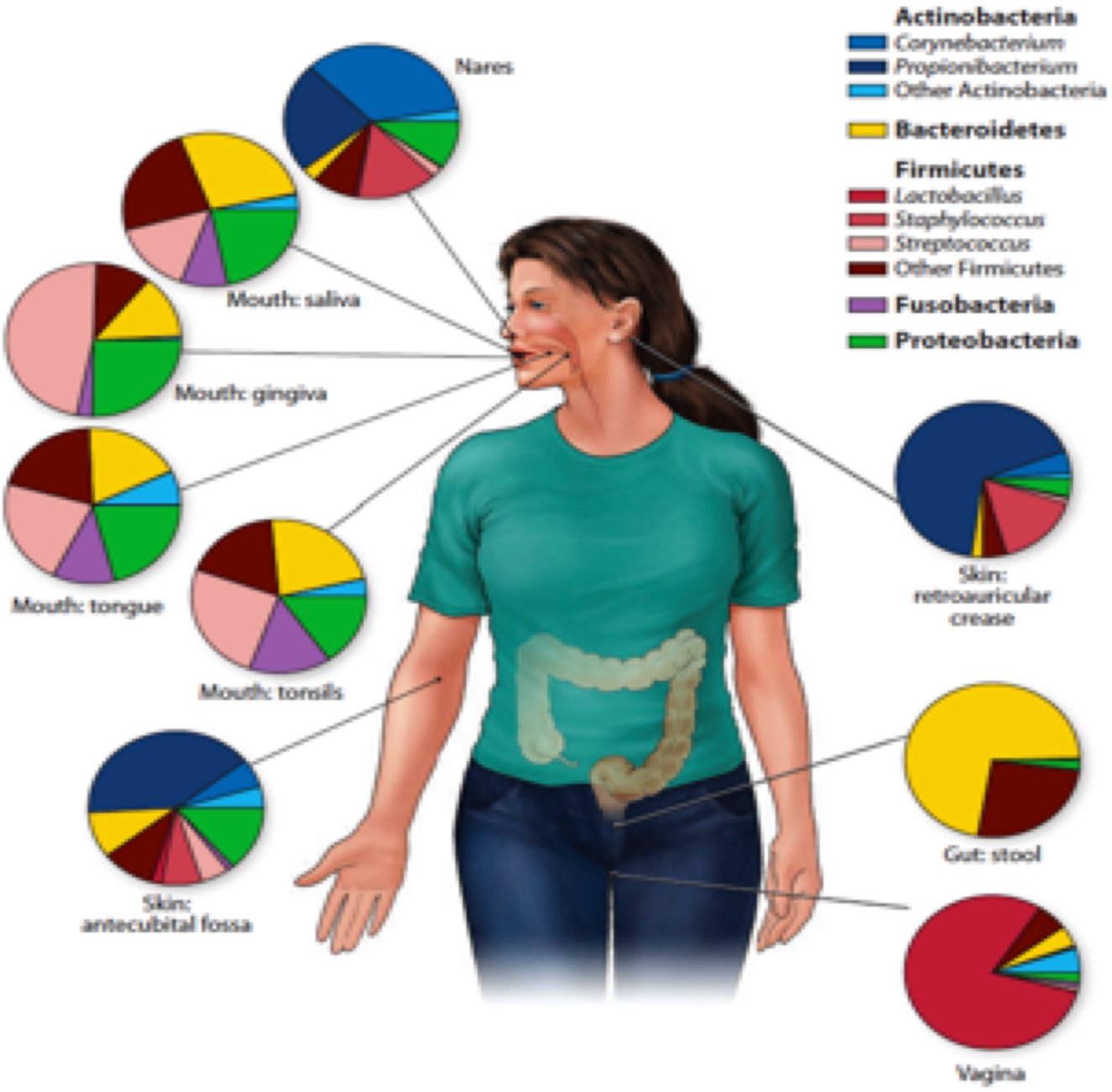
From samples (human, sea, polluted area, soils,… ):

- Who is there and abundance
- What are their functions

(what strains and what genes are present)

- Is it possible to recover genomes
- Are there novel pathogenic organisms

# Human Microbiome





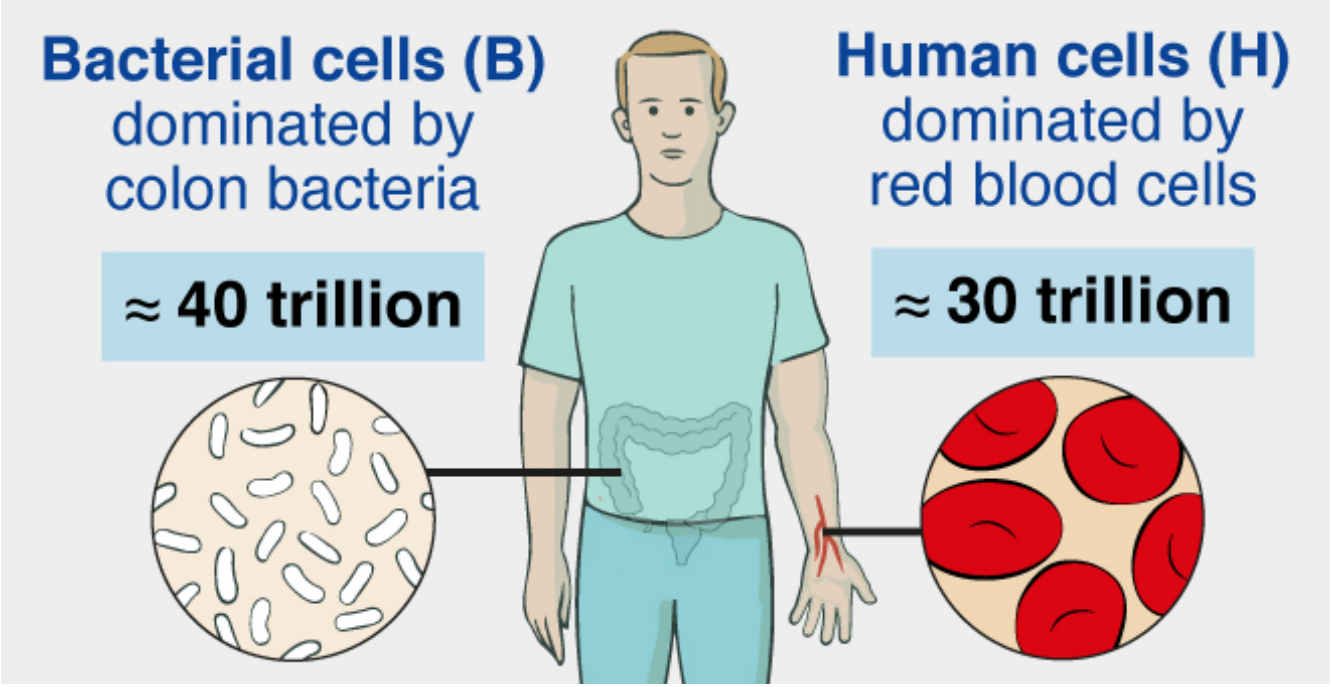

Grice EA and Segre JA.(2012). The Human Microbiome: Our Second Genome. *Annual Review Hum. Genet.* 13:151-70.

# Microbes and Microbiome

## The Ratio of Bacteria to Human Cells in the Adult Body Revised

With the revised estimates for the number of human ($3.0.10^{13}$) and bacterial cells ($3.8.10^{13}$) in the body, we can give an updated estimate of **B/H=1.3**, with an uncertainty of 25% and a variation of 53% over the population of standard 70 kg males.

**This B/H value of about 1:1 (with the associated uncertainty range) should replace the 10:1 or 100:1 values that are stated in the literature until more accurate measurements become available.**




Sender R, Fuchs S, Milo R (2016) Revised Estimates for the Number of Human and Bacteria Cells in the Body. PLoS Biol 14(8): e1002533

# A human gut microbial gene catalogue established by metagenomic sequencing

**Metagenomic sequencing, assembly and characterization of 3.3 million non-redundant microbial genes, derived from 576.7 gigabases of sequence, from faecal samples of 124 European individuals.**

**The gene set, c.a 150 times larger than the human gene complement, contains an overwhelming majority of microbial genes of the cohort.**

**Over 99% of the genes are bacterial, indicating that the entire cohort harbours between 1,000 and 1,150 prevalent bacterial species and each individual at least 160 such species.**

➔ **Project presentation: Friday afternoon**

Qin J et al. (2016). Nature 4: 59-67.

# Synthetic Biology

# Synthetic Biology

## Designing and building genomes

**A core theme in synthetic biology, "understanding by creating", inspired the efforts to generate synthetic cell.**

**https://www.nature.com/nature/focus/synbio/index.html**

**http://syntheticyeast.org**

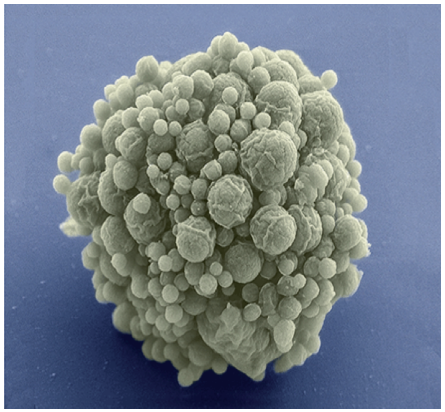
**Genome Project-write: GP-write**

Kannan K and Gibson DG. 2017. Yeast genome, by design. Science, 2017;35pp.1024-1025.

# 2016 Synthetic Biology

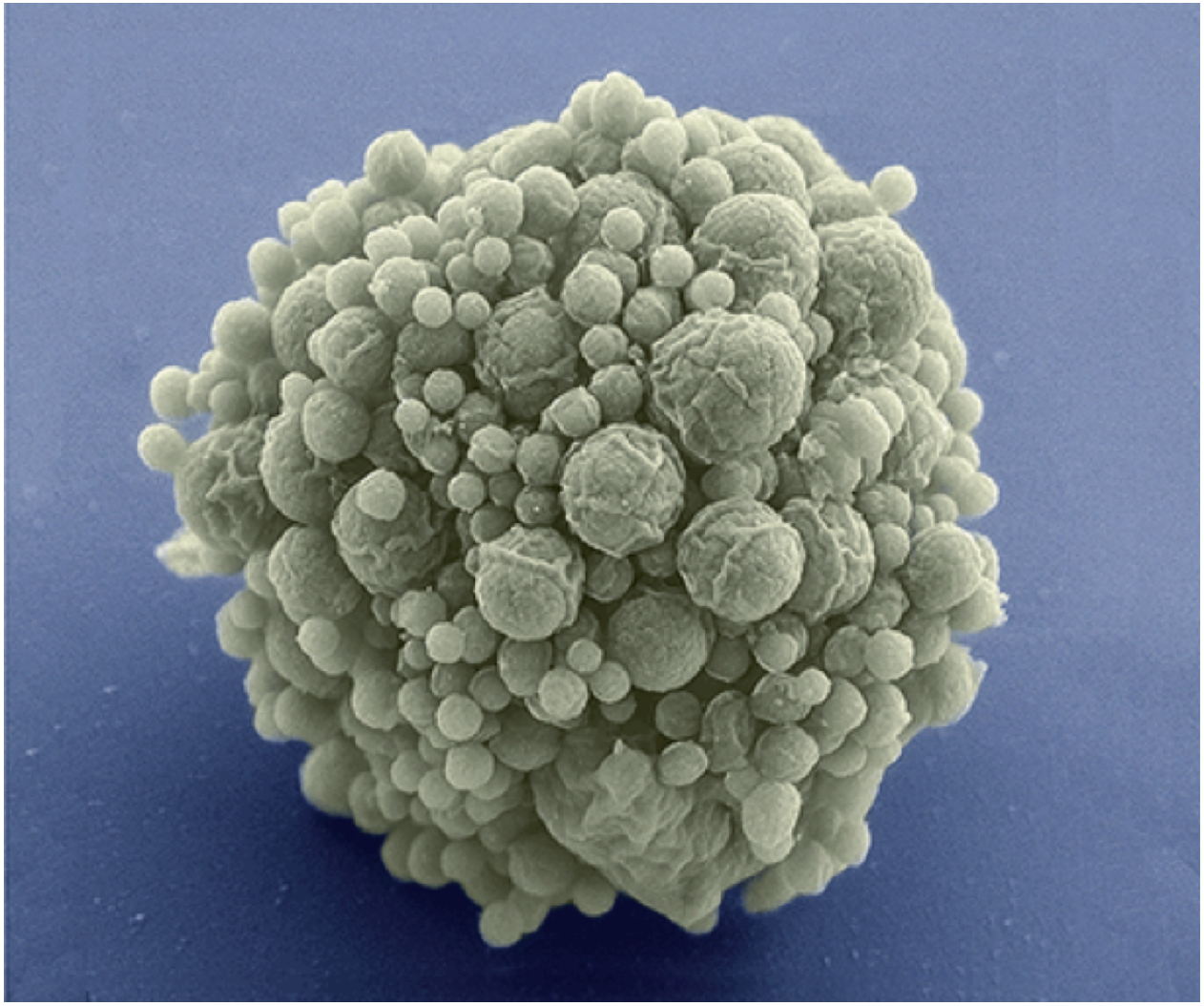
## Designing and building a minimal bacterial genome

**Three cycles of design, synthesis, and testing, with retention of quasi-essential genes, produced JCVI-syn3.0 (531 kilobase pairs, 473 genes), which has a genome smaller than that of any autonomously replicating cell found in nature.**



**Hutchison CA 3rd, et al**.Science. 2016 Mar 25;351(6280). **Design and synthesis of a minimal bacterial genome.**

# Synthetic microbe has fewest genes, but many mysteries. Robert F. Service



2016

## Meet Syn 3.0 and its record-setting small number of genes, 473

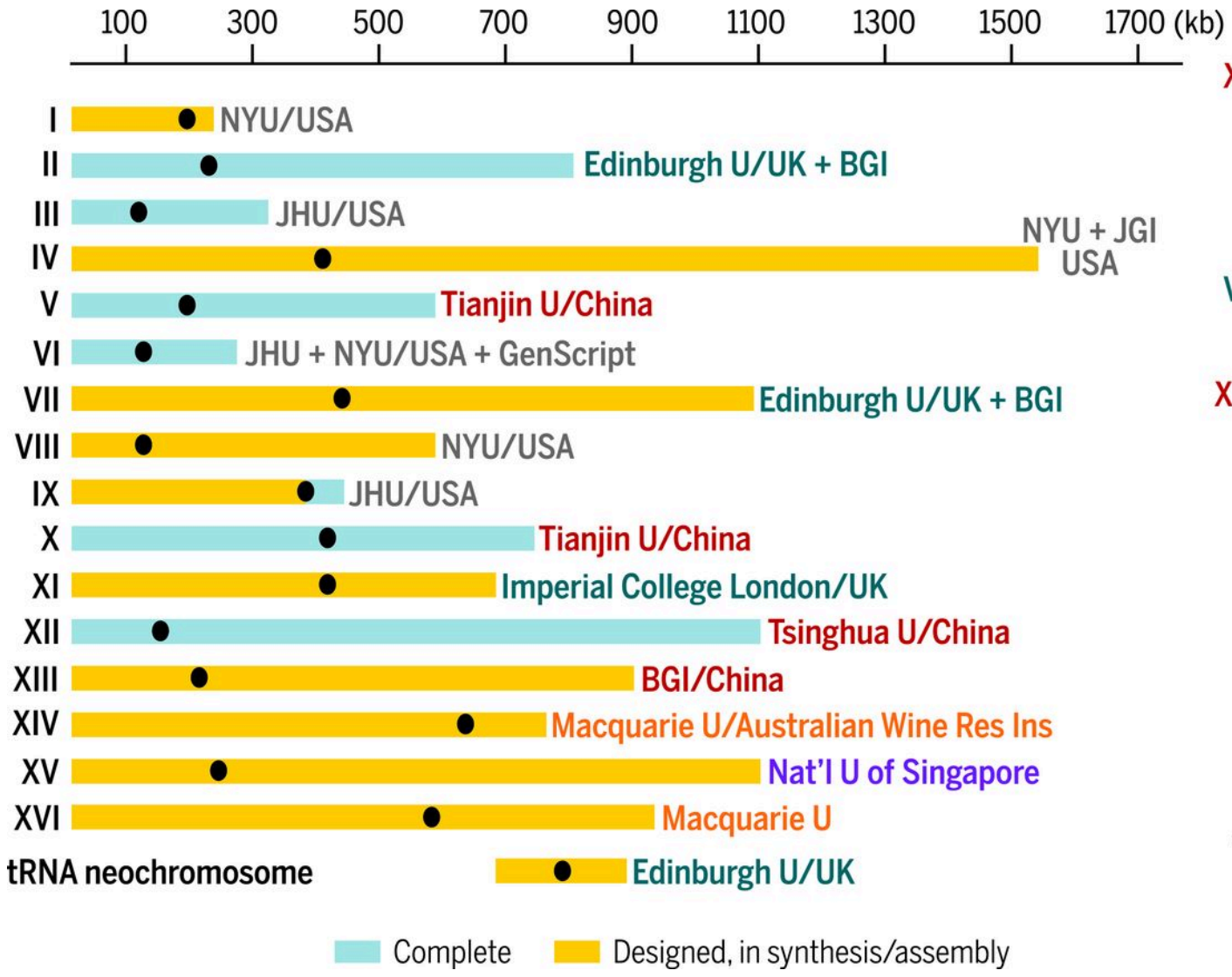
*Science.* 25 Mar 2016:Vol. 351, Issue 6280, pp. 1380-1381DOI: 10.1126/science.351.6280.1380.

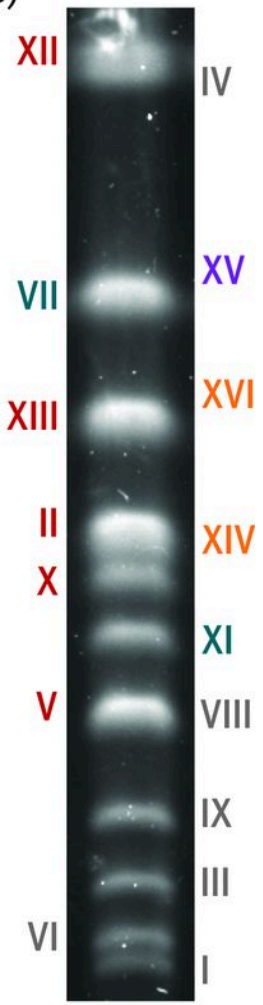

# 2017

**Chemically synthesized genomes like Sc2.0 are fully customizable and allow experimentalists to ask otherwise intractable questions about chromosome structure, function, and evolution with a bottom-up design strategy.**

Illustration of a hypothetical yeast genome structure model encompassing all synthetic chromosomes (gold) completed to date (white, native chromosomes). One-third of the chromosomes in the yeast genome have now been designed and synthesized by the Synthetic Yeast Genome Project (Sc2.0). A three-dimensional model of the chromosomes was generated with the Hi-C method. An "envelope" (thick tube shapes) represents the population of chromosome-interacting molecules; translucent tips reveal a 30-nanometer fiber modeled inside.

# Sc2.0 Consortium chromosome assignments



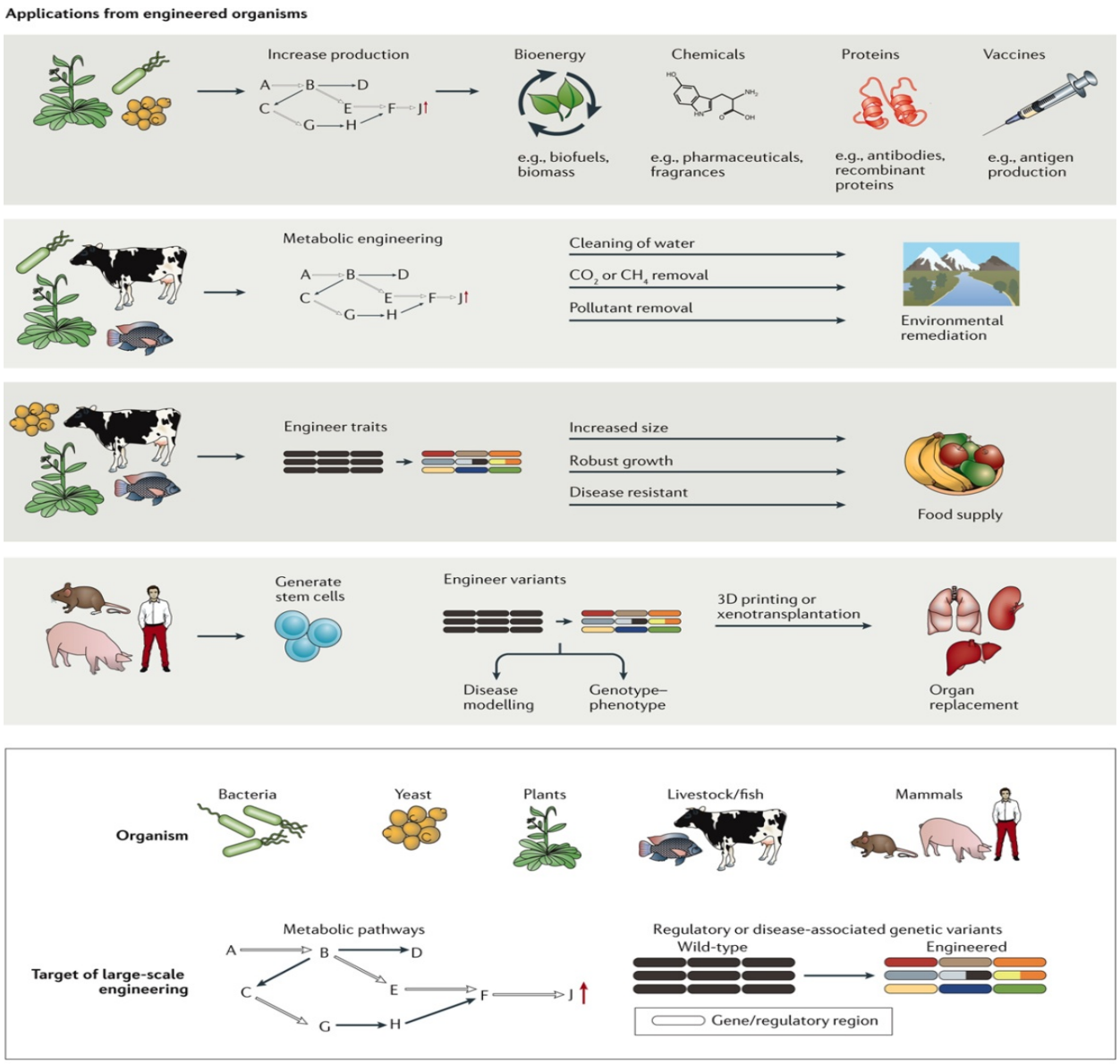

When the researchers put chunks of synthetic DNA into yeast cells, the cells swapped out parts of their original DNA for the matching engineered snippets.

Sarah M. Richardson et al. Science 2017;355:1040-1044

# Design of a synthetic yeast genome

Science AAAS

# Applications from Engineered Organisms



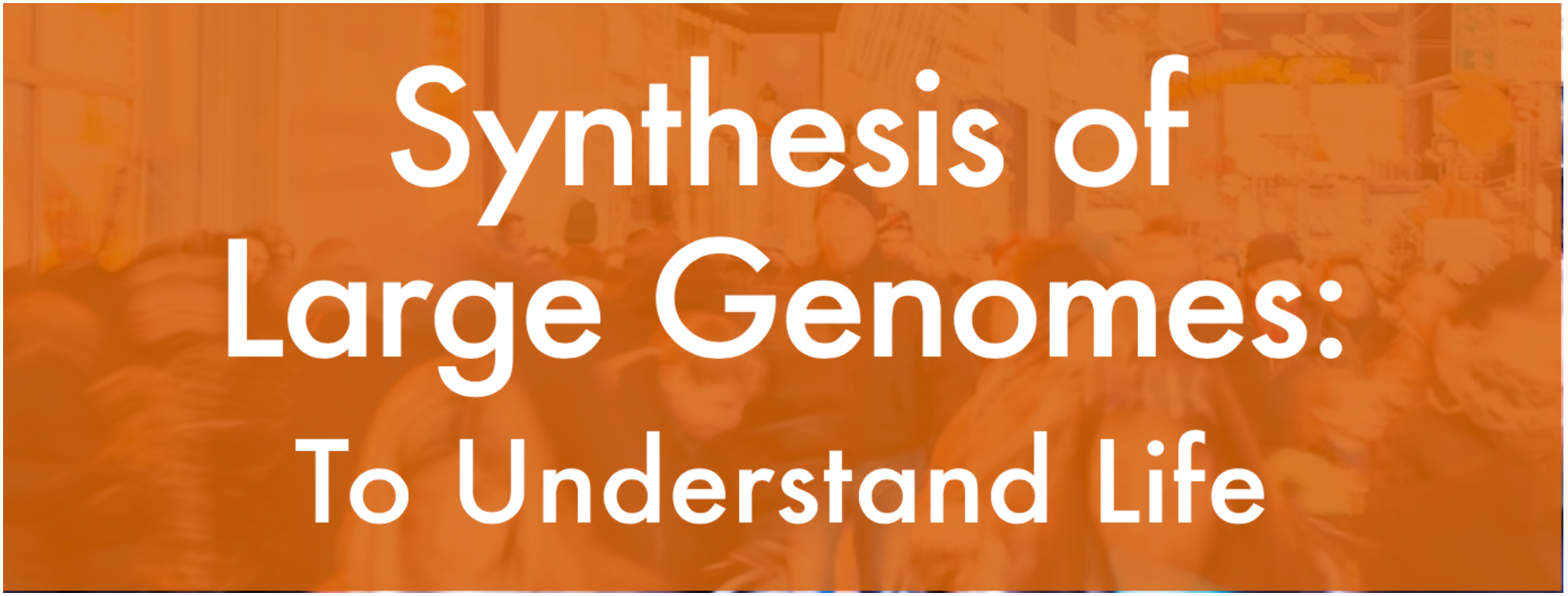

Exponential advances in genome sequencing and engineering technologies have enabled **interrogations into the impact of DNA variation (genotype) on cellular function (phenotype).** **These advances have also prompted realistic discussion of writing and radically re-writing complex genomes.** In this Perspective, we detail the **motivation for large-scale engineering,** discuss the progress made from such projects in bacteria and yeast and describe how various genome-engineering technologies will contribute to this effort. Finally, **we describe the features of an ideal platform and provide a roadmap to facilitate the efficient writing of large genomes**.

➔ **Project presentation: Monday am**

Raj Chari & George M. Church. (2017). Beyond editing to writing large genomes. 2017. Nature Reviews Genetics (2017) doi:10.1038/nrg.2017.59. Published online 30 August 2017.

# Genome Project-write: GP-write

http://engineeringbiologycenter.org



# Genome Project-write: GP-write

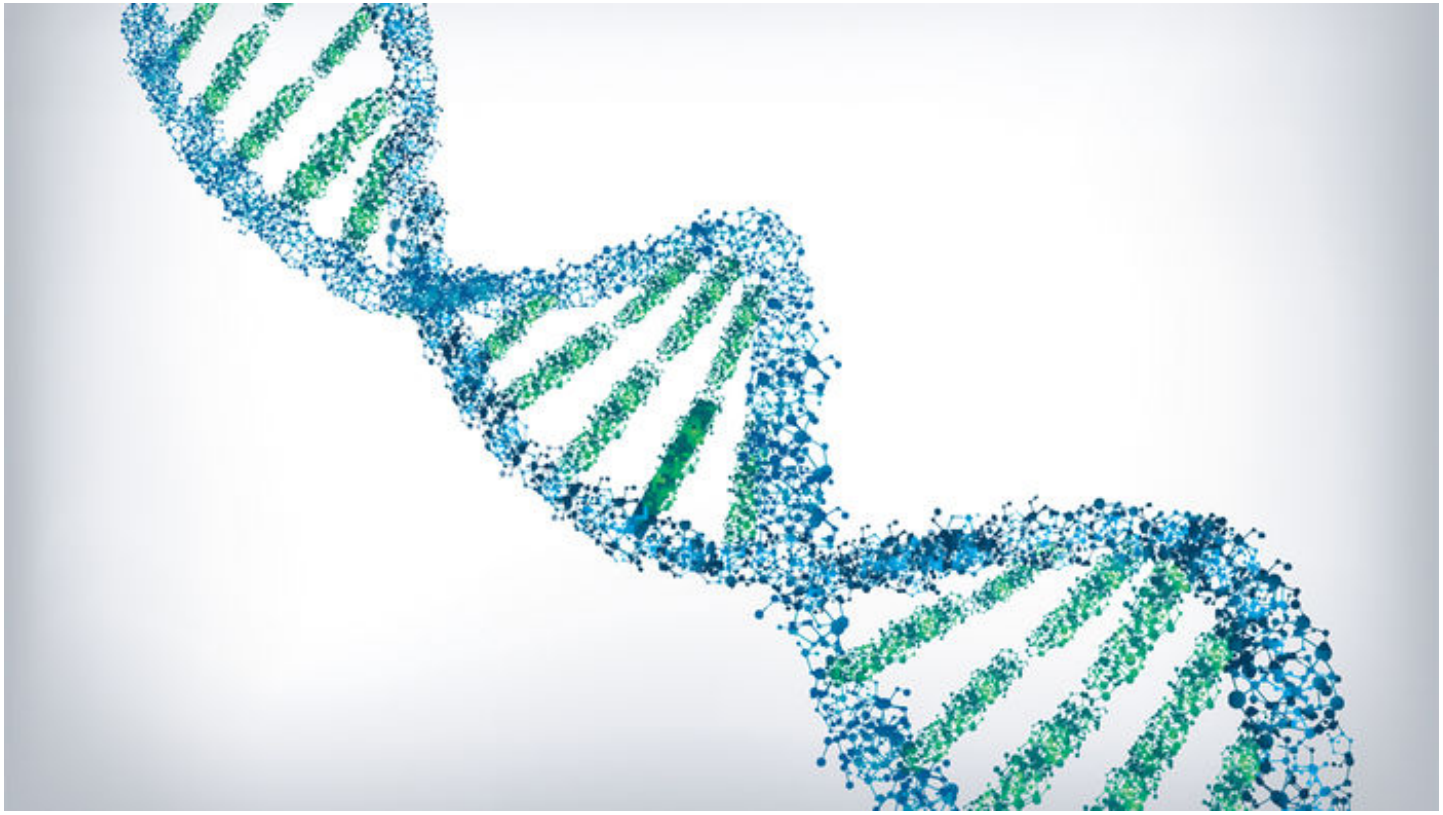
http://engineeringbiologycenter.org

## From Observation to Action

GP-write will enable scientists to move beyond observation to action.

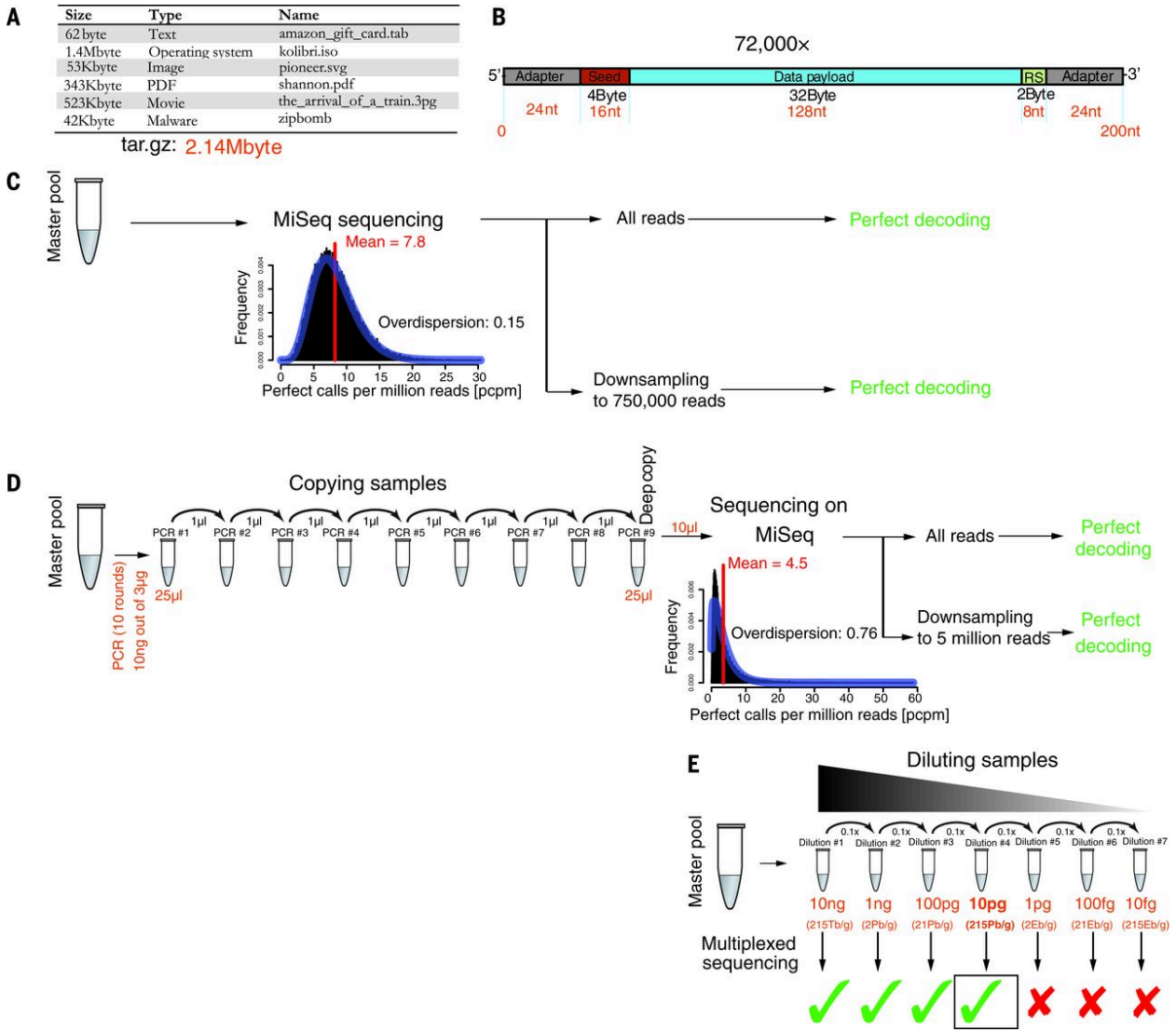
Writing DNA is the future of science and medicine.

# DNA: Medium for Storing Data



**A new method of storing data in the nucleotide bases of DNA is the highest-density storage scheme ever invented.**

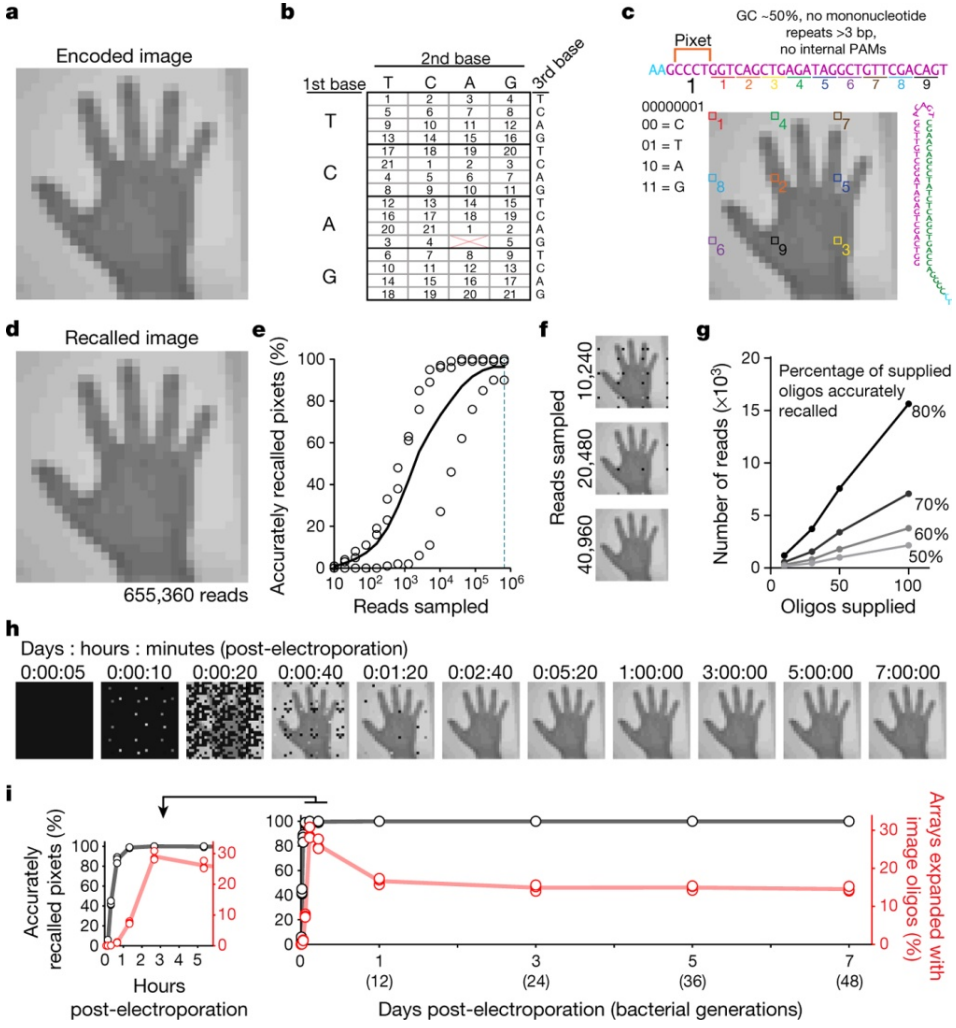
# Experimental setting and results for storing data on DNA.




Yaniv Erlich, and Dina Zielinski Science 2017;355:950-954

Published by AAAS

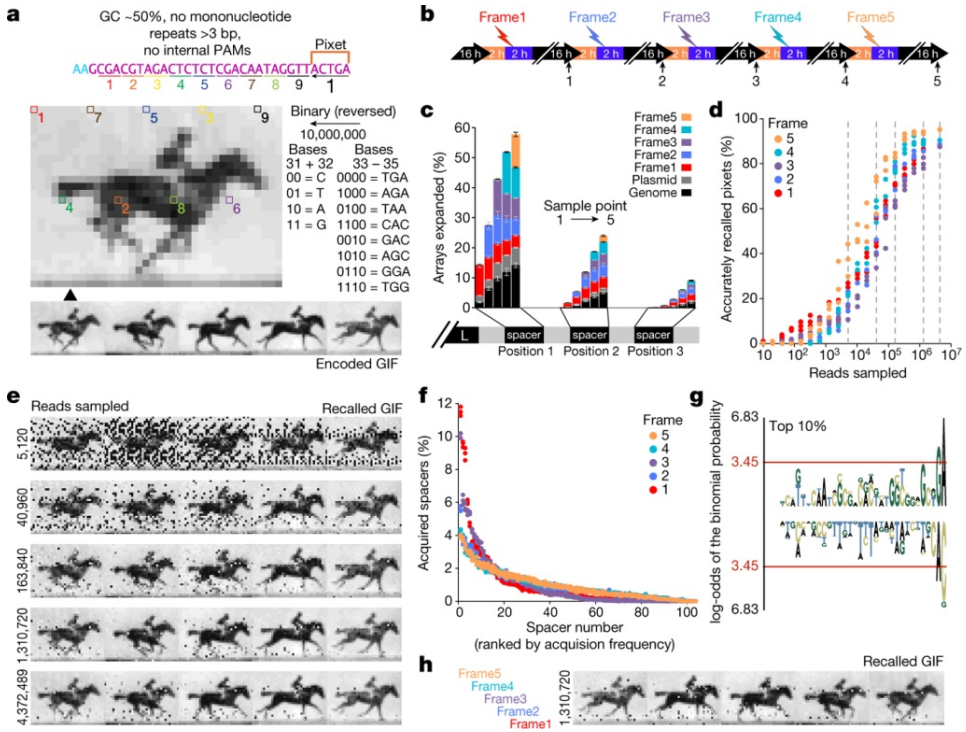
# An image into the genome




**S L Shipman *et al. Nature* 1–5 (2017) doi:10.1038/nature23017**

nature

# Encoding a GIF in bacteria




**S L Shipman *et al.* *Nature* 1–5 (2017) doi:10.1038/nature23017**

nature

# Artificial intelligence tools are helping to reveal the genetic components of autism.

**Identifying functional effects of noncoding variants is a major challenge in human genetics.**

**DeepSEA, a deep learning-based algorithmic framework, that directly learns a regulatory sequence code from large-scale chromatin-profiling data, enabling prediction of chromatin effects of sequence alterations with single-nucleotide sensitivity.**

**Machine learning were used to build a map of gene interactions. The few well-established autism risk genes were compared to thousands of other unknown genes, looking for similarities. They flagged another 2500 genes likely to be involved in autism among the 25000 human genes.**

Zhou J, Troyanskaya OG.(2015). Predicting effects of noncoding variants with deep learning-based sequence model. Nat Methods. 2015 Oct;12(10):931-4.

Krishnan A, et al. (2016). Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. Nat Neurosci. 2016 Nov;19(11):1454-1462.

# Wrap up

# Genomes are too big and highly variable in size

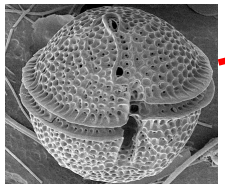
*Fritillaria assyriaca*
100 000 Mb

*Zea mais*
*3 000 Mb*

*Vitis vinifera*
487 Mb

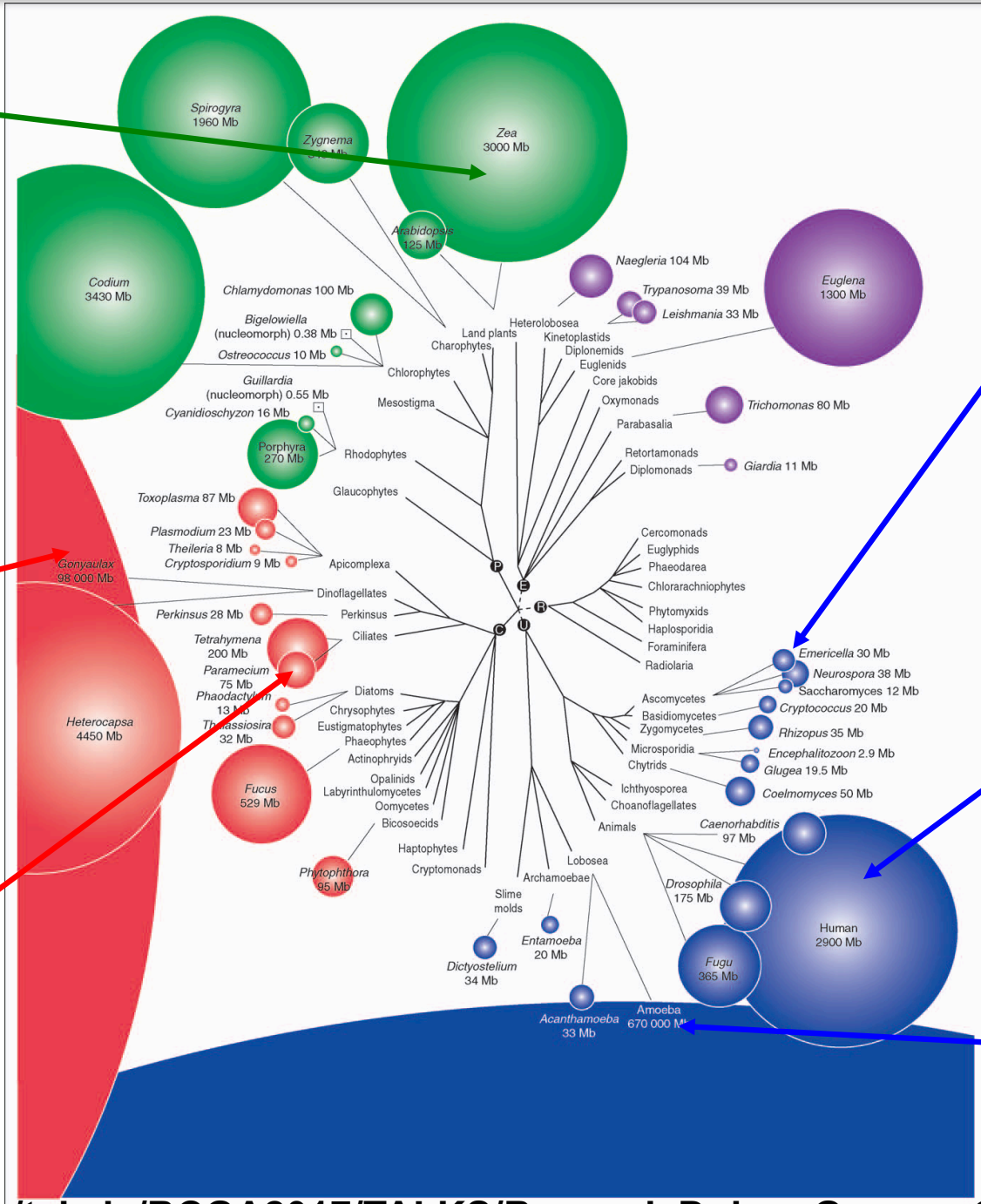
1000 X

1000 X

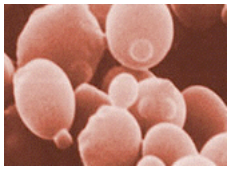


*Gonyaulax grindleyi*
98 000 Mb

*Paramecium tetraurelia*
100 Mb

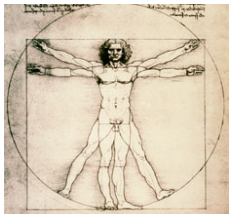



55 000 X

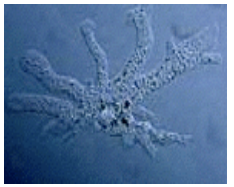


*Saccharomyces cerevisiae* 12 Mb

*Necturus lewisi*
100 000 Mb



*Homo sapiens*
3 100 Mb



*Amoeba dubia*
670 000 Mb

https://webext.pasteur.fr/tekaia/BCGA2017/TALKS/Bernard_Dujon_GenomeComparisons.pdf

# Three main lessons from genomics

## 1: Genomes are (much) too big

The C-value paradox (Swift, 1950)

**C-value paradox is the complex puzzle surrounding the extensive variation in nuclear genome size among eukaryotic species. At the center of the C-value paradox is the observation that genome size does not correlate with organismal complexity; for example, some single-celled protists have genomes much larger than that of humans.**

## 2: There are too many genes in genomes

## 3: There are alien and orphan genes in every genome

https://en.wikipedia.org/wiki/C-value

https://webext.pasteur.fr/tekaia/BCGA2017/TALKS/Bernard_Dujon_GenomeComparisons.pdf

## Gene gains

Duplications
Fission / fusion
*De novo* formation
Horizontal acquisitions

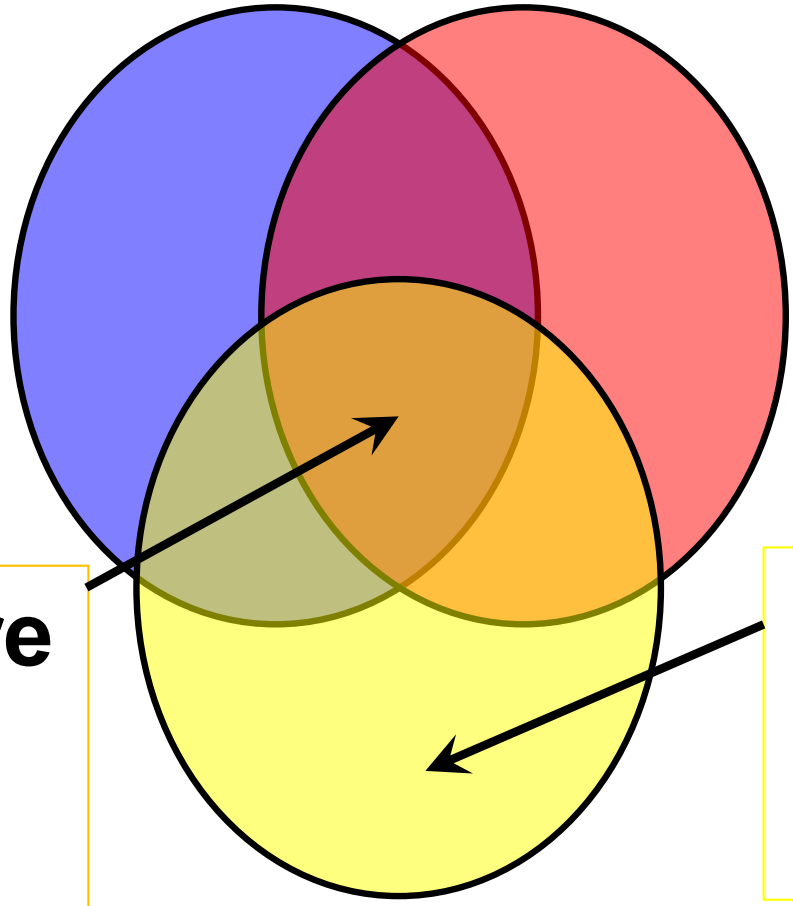

## Gene losses

Pseudogenes
Deletions

➔ **Each genome is only a snapshot in time within continual changes, not an optimized structure**

https://webext.pasteur.fr/tekaia/BCGA2017/TALKS/Bernard_Dujon_GenomeComparisons.pdf

# Universal results of comparative genomics

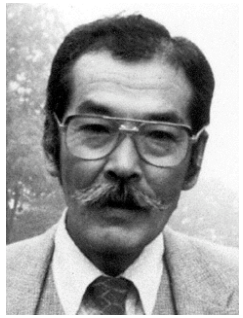



**Common core of genes (ancestral origin)**

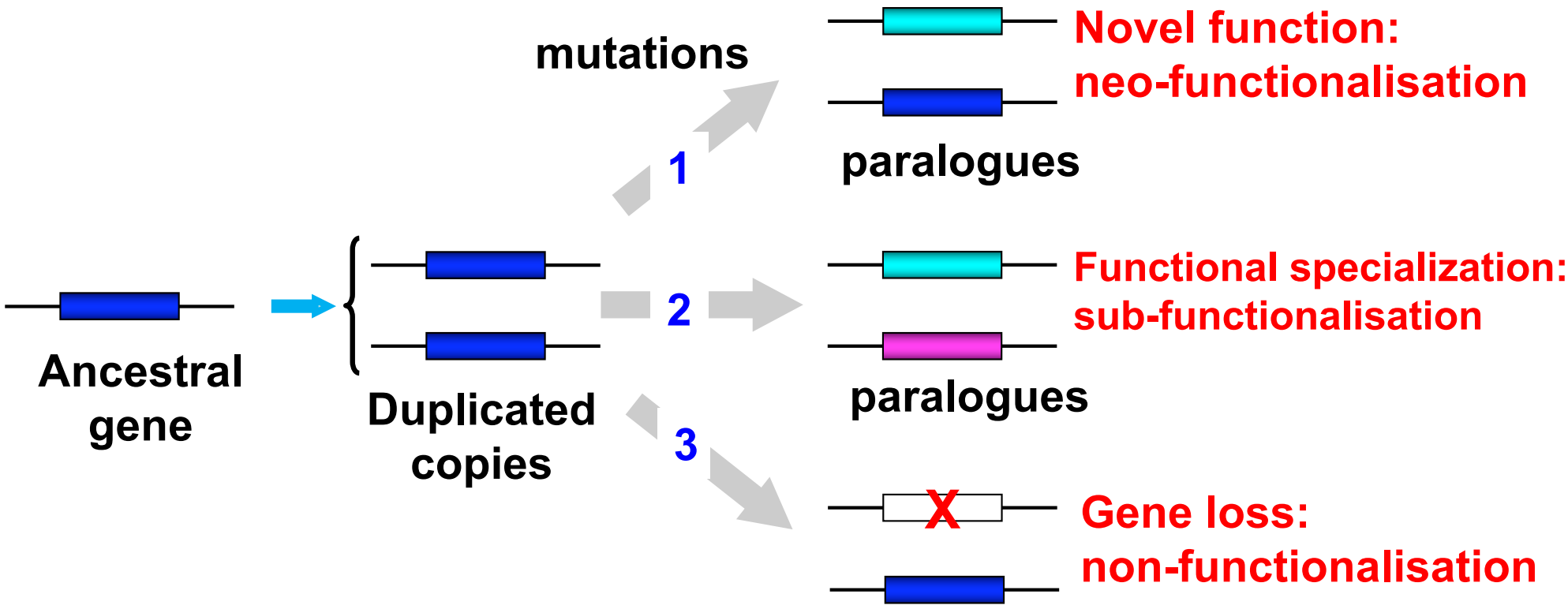
**Restricted set of genes (origin ?)**

https://webext.pasteur.fr/tekaia/BCGA2017/TALKS/Bernard_Dujon_GenomeComparisons.pdf

# Gene duplications and consequences

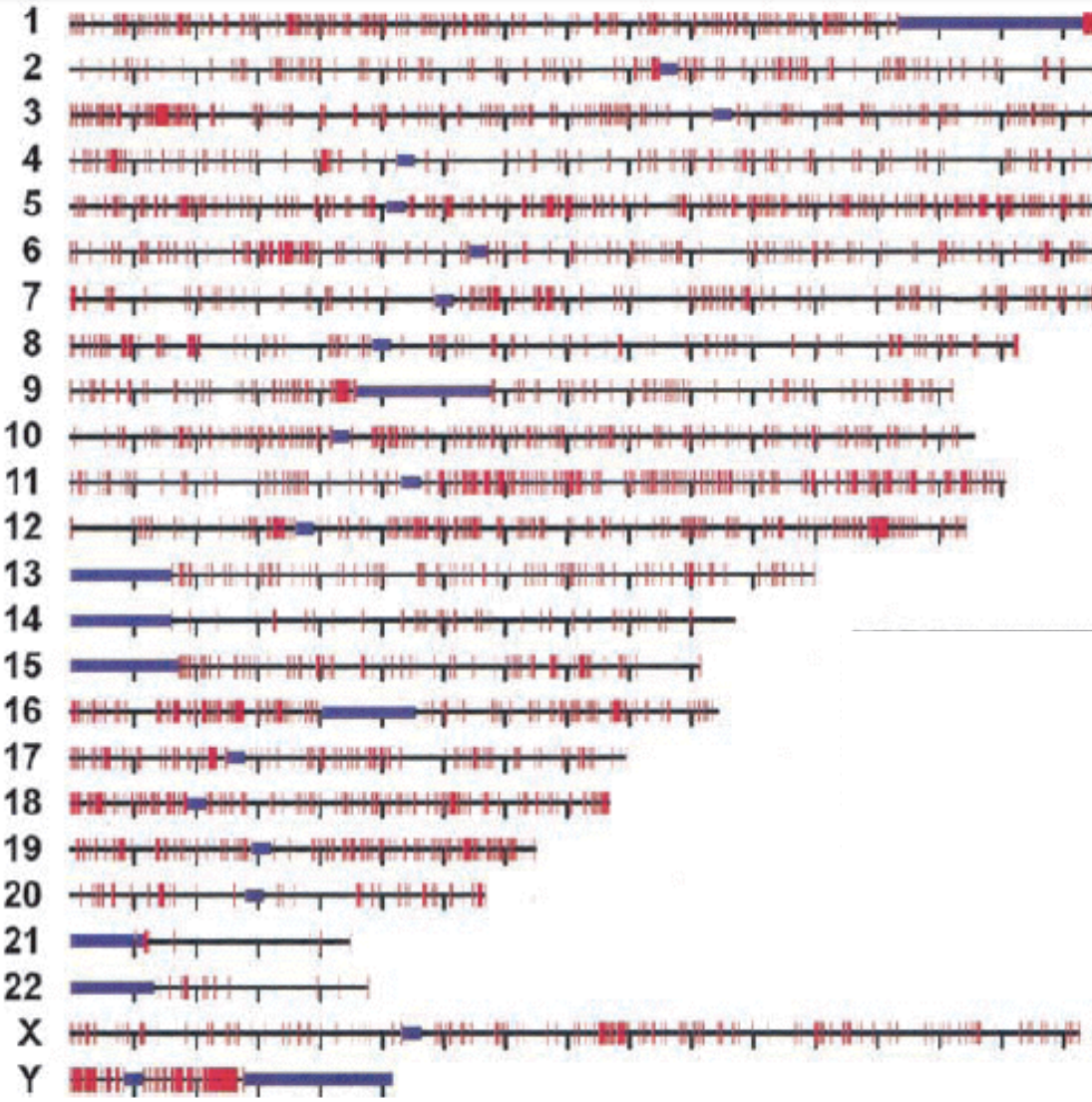


Susumu Ohno, 1970



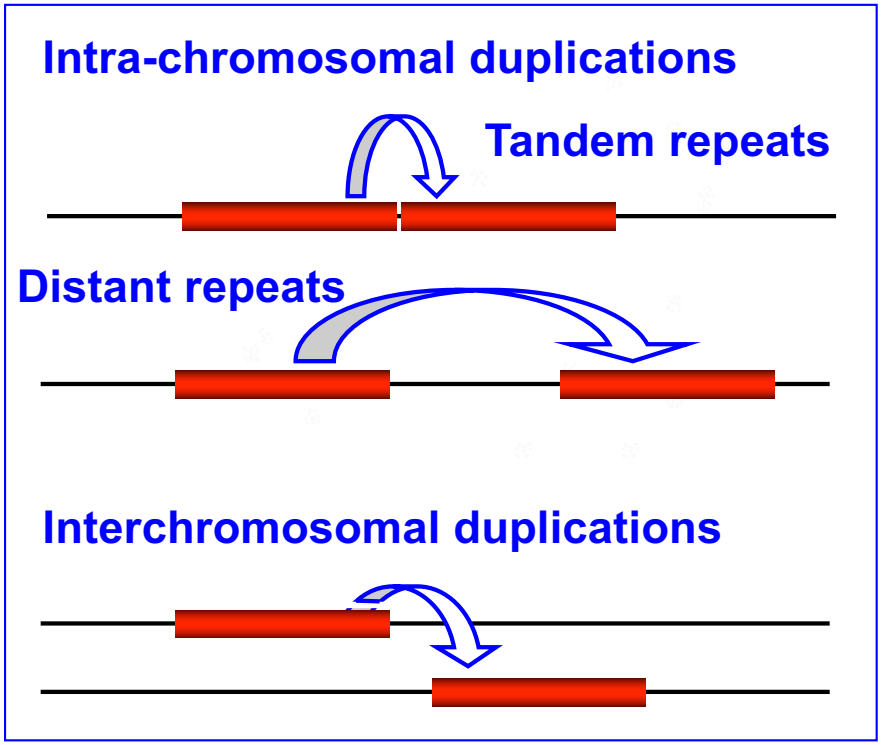

https://webext.pasteur.fr/tekaia/BCGA2017/TALKS/Bernard_Dujon_GenomeComparisons.pdf

# Segmental duplications in the human genome

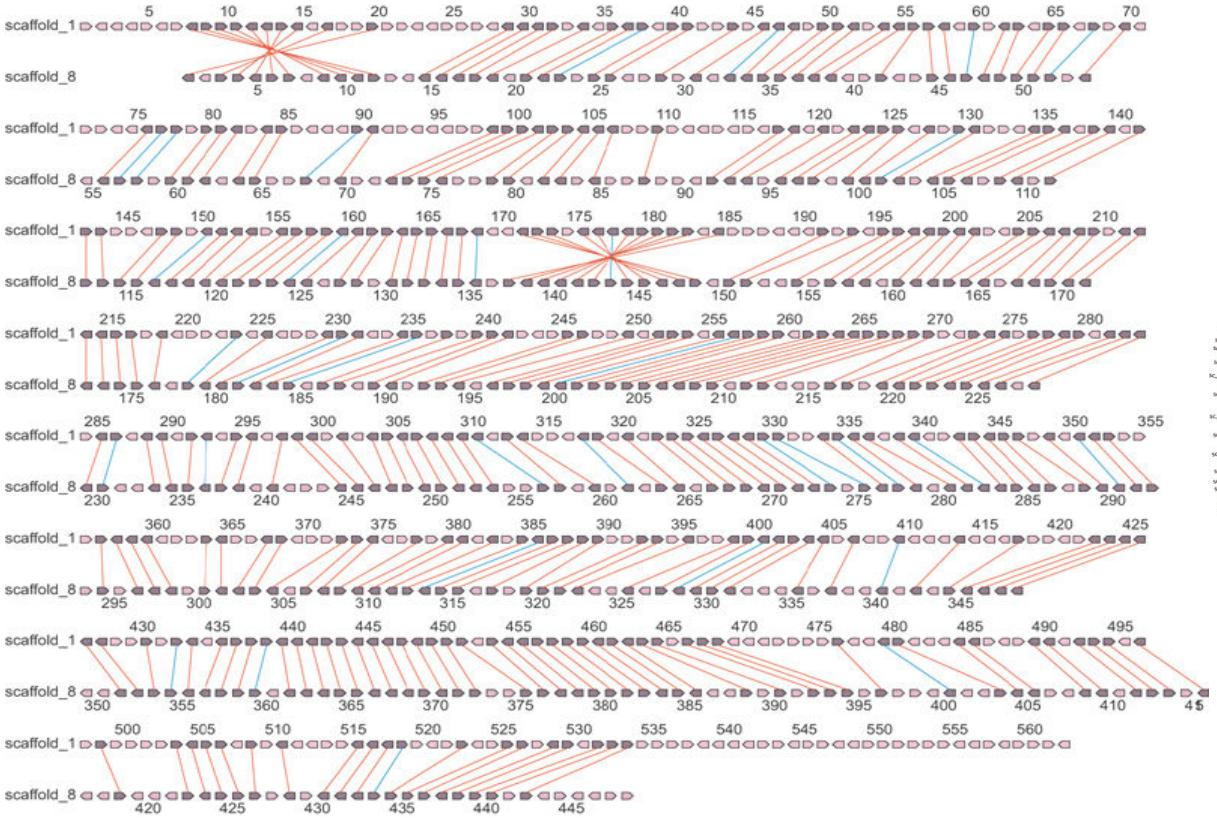


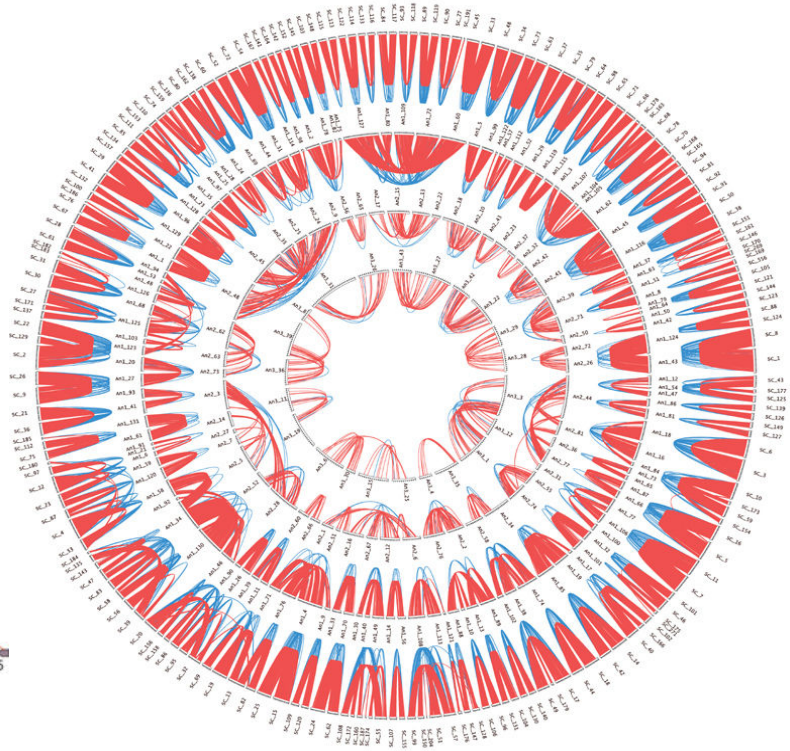

➢~ 5% of human genome is made of segmental duplications
(98-100 % identical sequences > 1 kb )

Duplicated sequences
Centromeres
Scale 10 Mb

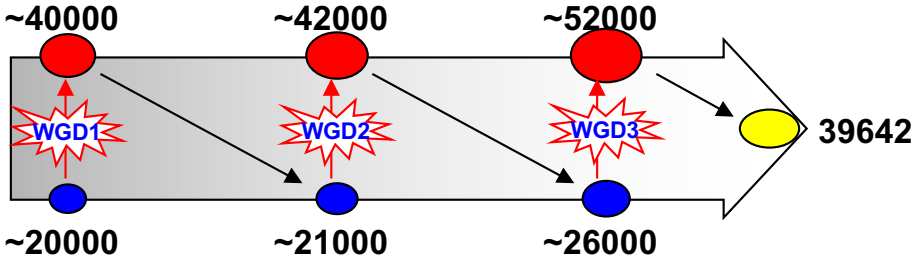


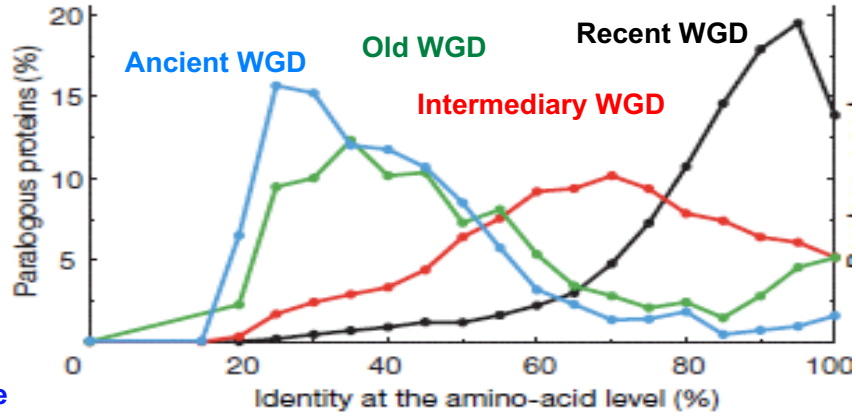
# Several successive whole-genome duplications in *Paramecium*



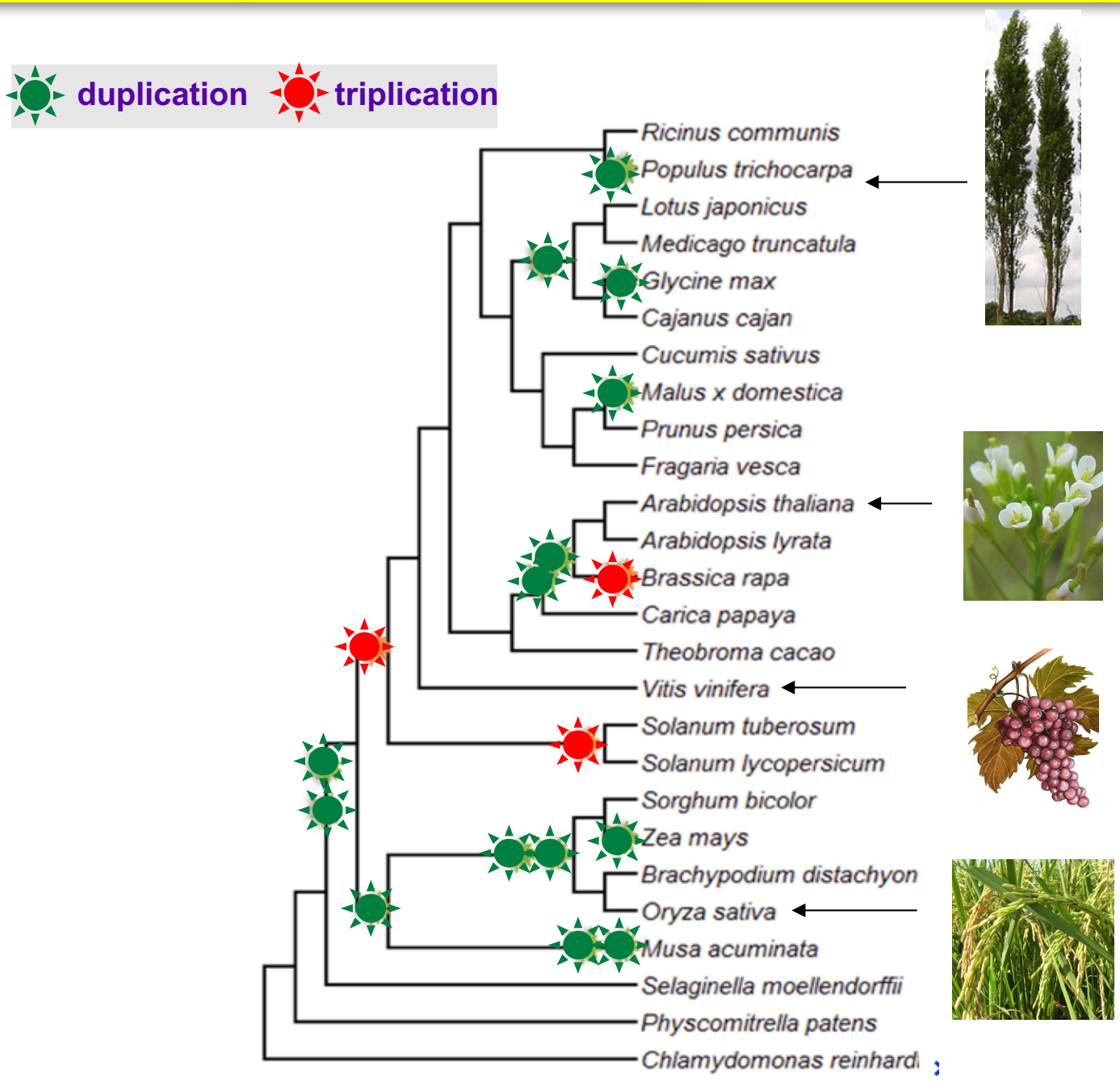
**Comparison of two scaffolds originating from a common ancestor after a recent WGD in P*aramecium tetraure*"**




**Aury et al. (2006). Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia.** Nature. 2006 Nov 9;444(7116):171-8.

# Multiple genome duplications and triplication in *Streptophyta*




https://webext.pasteur.fr/tekaia/BCGA2017/TALKS/Bernard_Dujon_GenomeComparisons.pdf

# Concluding notes

# Biology has changed

- Data explosion
- New types of data
- High-throughput biology
- Emphasis on systems (not single parameters)
- Increasing interest in applied biology
    - molecular medicine
    - agriculture
    - environment;….

## Genome studies show that:

- **Access to entire genome sequences has revolutionized our understanding of how genetic information is stored and organized in DNA, and how it has evolved over time.**

- **The sequence of a genome provides significant details of the gene catalogue within a species.**

- **Comparisons of complete genome sequences show the acceleration in the understanding of species organisation, links between genes, functions of the genes, evolution of genes, genomes and species.**

## So far Genome research has succeeded in:

- **Understanding the genome structure**
- **Understanding the biology of genomes**

## In Human, advances are to be expected in:

- **Understanding the biology of disease**
- **Advancing the Science of Medicine**
- **Improving the Effectiveness of Healthcare**

Green ED et al. 2011.

Charting a course for genomic medicine from base pairs to bedside. *Nature*. 470:204-13.

# Bioinformatics Challenges and Perspectives

- **Bioinformatics is a fascinating research domain offering many opportunities for Biologists, Mathematicians, Statisticians, Computer scientists and soon for Clinicians,…**

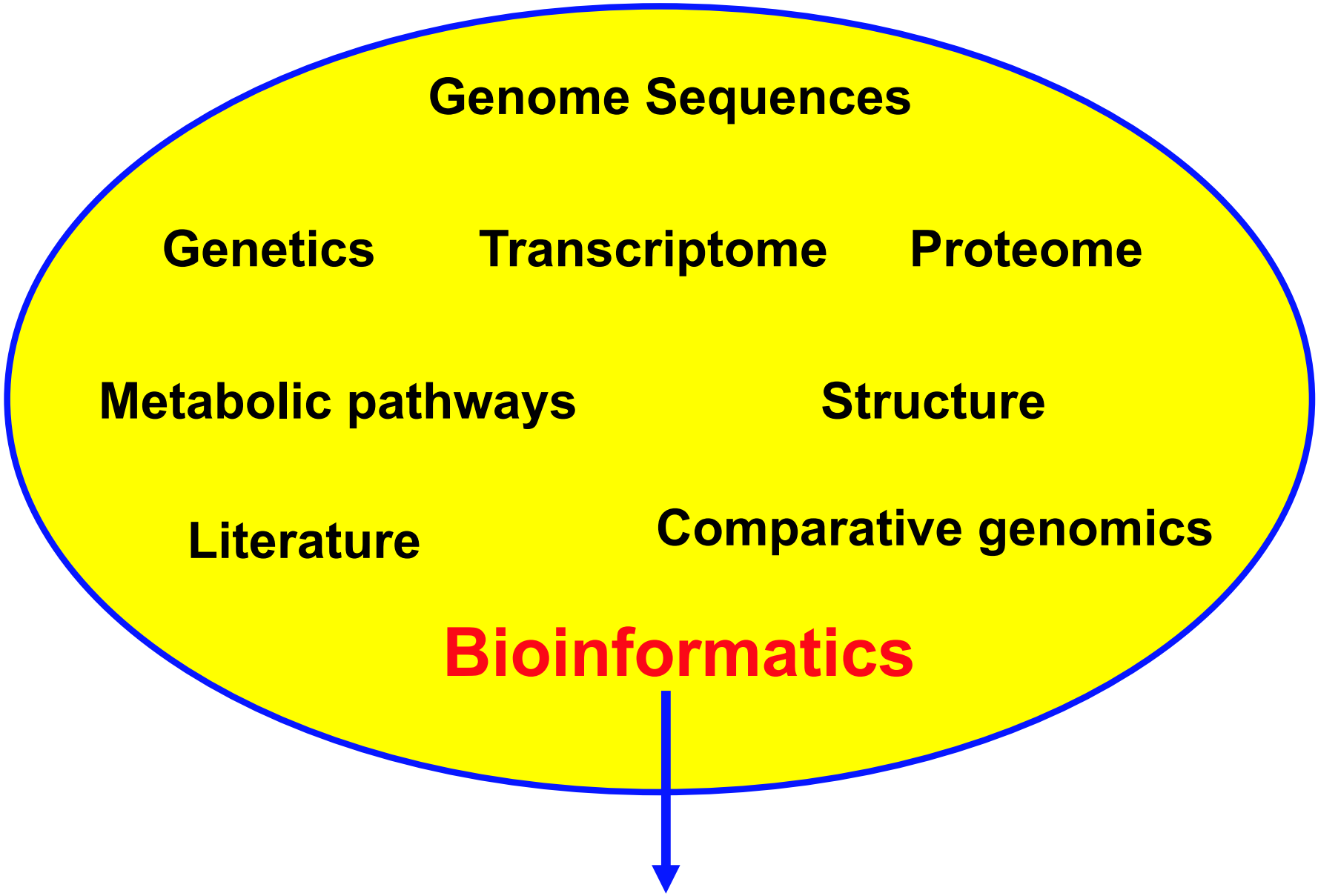
**==> Soon: Biology *in silico* with discovery applications in wet Labs.**

# Bioinformatics challenges

- Bioinformatics research tools are now common practice in bioscience laboratories and the impact of bioinformatics on medicine, ecology, and cell biology is expected to increase continuously.

- Need for computational infrastructure that supports data processing, sharing and providing training resources that will allow researchers to better understand, handle, and compare large datasets

- Technologies continue to improve, there will be an ever continuous need to develop algorithms capable of delivering computing improvements




***in silico* models / wet lab experiments**

# All Biology is Computational Biology

**Markowetz F (2017). PLoS Biol 15(3): e2002050.**

- **Computational thinking and techniques are so central to the quest of understanding life that today all biology is computational biology**

- **The next modern synthesis in biology will be driven by mathematical, statistical, and computational methods being absorbed into mainstream biological training, turning biology into a quantitative science.**

**Computational biology brings order into our understanding of life**

**Computational biology lets you see the big picture**

**Computational biology provides an atlas of life**

**Computational biology turns ideas into hypotheses**

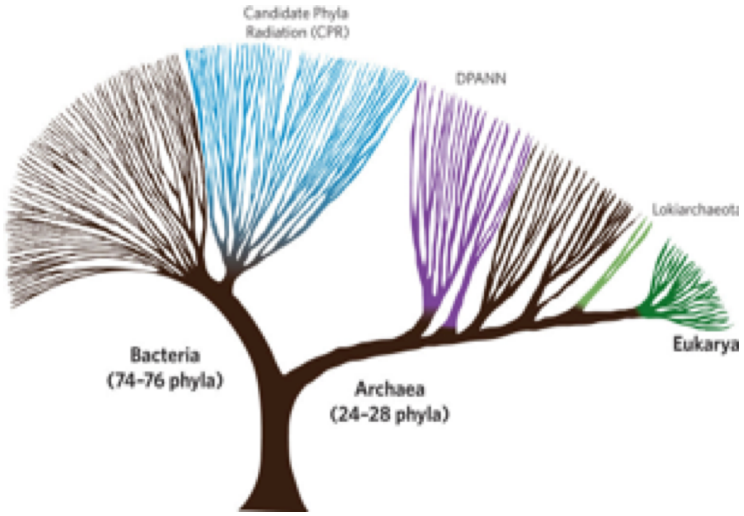
# Challenging concepts

## What Is Speciation?

B. Jesse Shapiro, Jean-Baptiste Leducq, James Mallet

Plos Genetic, 2016.

http://dx.doi.org/10.1371/journal.pgen.1005860




## What Is the Tree of Life?

W. Ford Doolittle, Tyler D. P. Brunet

Plos Genetic, 2016.

# Stay up to date with the scientific publications

# Few bibliographic resources

- PubMed: http://www.ncbi.nlm.nih.gov/pubmed/
- Nature: http://www.nature.com/
- Nature Review Genetics
http://www.nature.com/nrg/index.html
- Sciences: http://www.sciencemag.org/magazine
- Genome Research: http://genome.cshlp.org/
- NAR: http://nar.oxfordjournals.org/
- Bioinformatics:
http://bioinformatics.oxfordjournals.org/
- Computational Molecular Biology:
http://www.liebertpub.com/CMB
- Current Biology:
http://www.cell.com/current-biology/home

# http://www.nature.com/naturejobs/science/

# References

- **Bioinformatics and Genomes Analyses courses:**
**http://www.pasteur.fr/~tekaia/BCGA_WProgs.html**

- **On-courses: list of available courses (short/long)**
**http://www.on-course.eu/**

- **Tekaia F. 2016. Inferring Orthologs: Open questions and perspectives. *Genome Insights*. 9: 17-28.**

- **Tekaia F. 2016. Genome Data Exploration Using Correspondence Analysis. *Bioinformatics and Biology Insights*. 10: 1-14.**

- **Tattini L, D'Aurizio R, Magi A. 2015. Detection of Genomic Structural Variants from Next-Generation Sequencing Data. *Front Bioeng Biotechnol.*;3:92.**

- **Zhao M, et al. 2013. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics*(Suppl 11):S1DOI: 10.1186/1471-2105-14-S11-S1**

- **Software tools: http://omictools.com/sequencing-category**

# Thank You