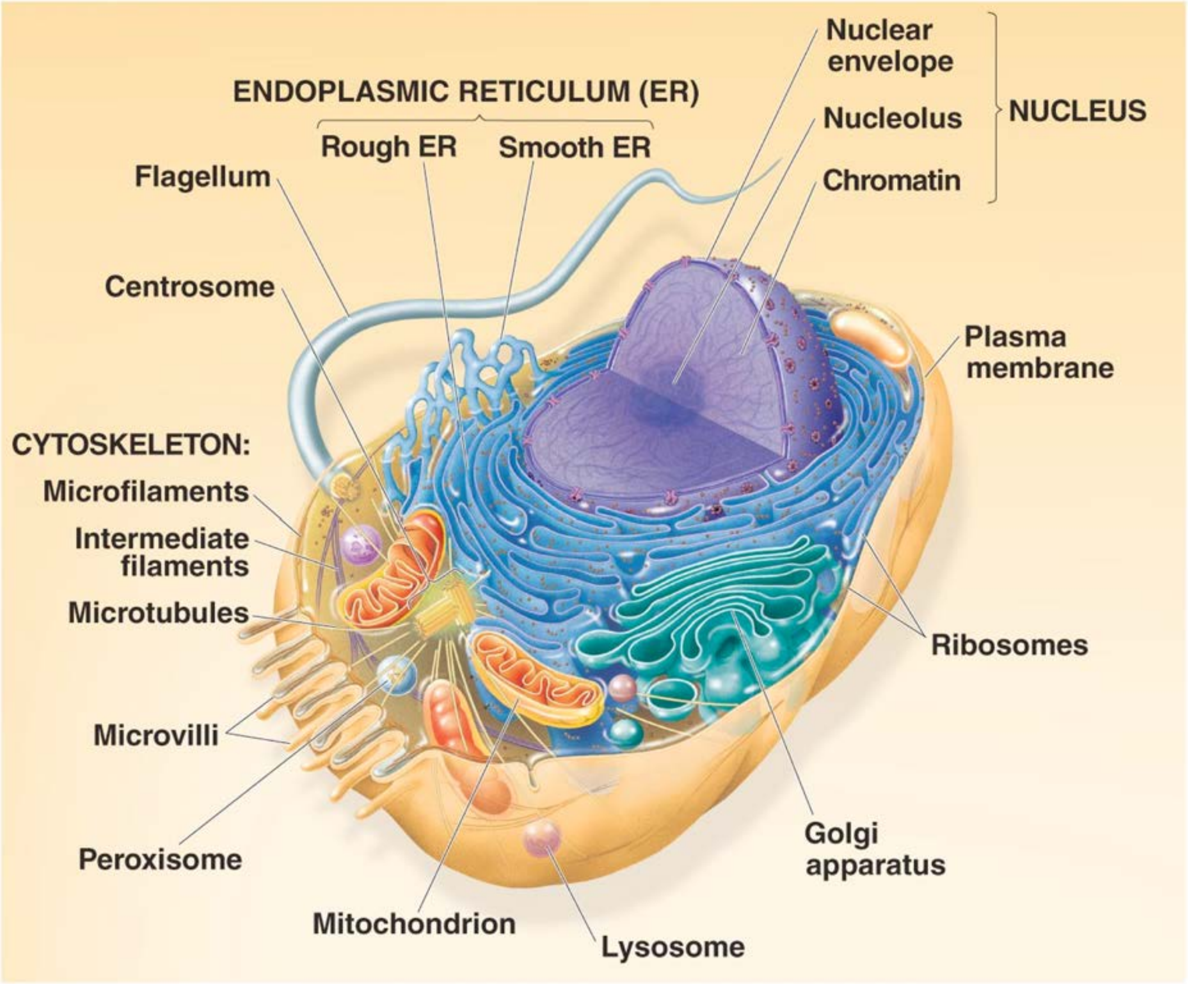
Introduction to Comparative Genomics

Evgeny.Zdobnov@unige.ch

UNIVERSITÉ DE GENÈVE
FACULTÉ DE MÉDECINE

Cell Elements



Cell elemental composition

Cells are 90% water.

The remaining is approximately:

- 50% **protein**
- 15% **carbohydrate**
- 15% **nucleic acid**
- 10% **lipid**
- 10% **miscellaneous**

Cell Elements

- Proteins are the main cellular machinery
- All proteins – proteome
- All DNA – genome
- All RNA – transcriptome
- All lipids – lipidome

Terms

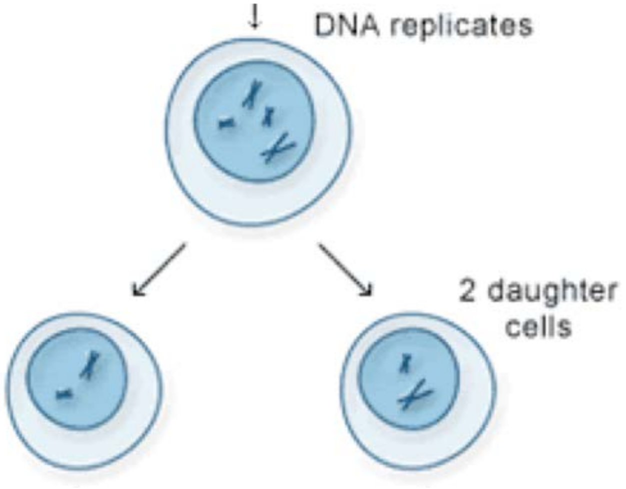
- -omics ⇔ high throughput data acquisition in Molecular Biology

- Bioinformatics ⇔ computational management and analysis of biological data

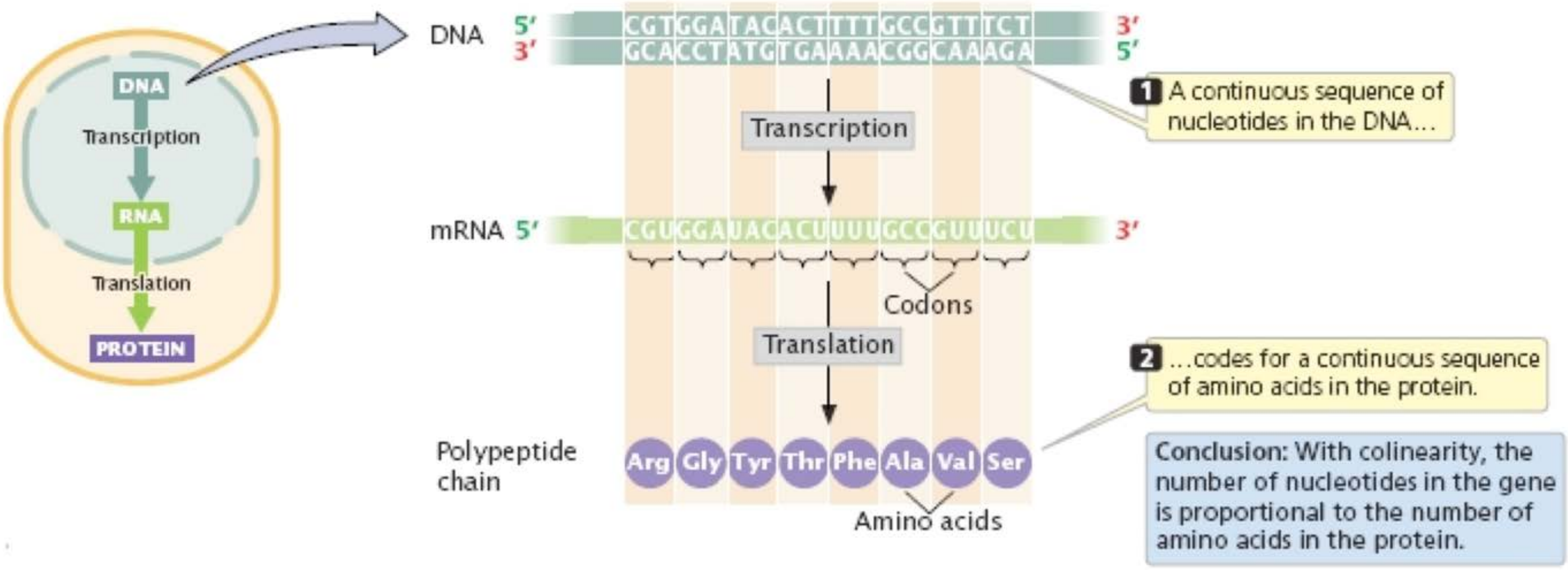
Why Genomics?

UNIVERSITÉ
DE GENÈVE
FACULTÉ DE MÉDECINE

Genome encodes hereditary information



The dogma

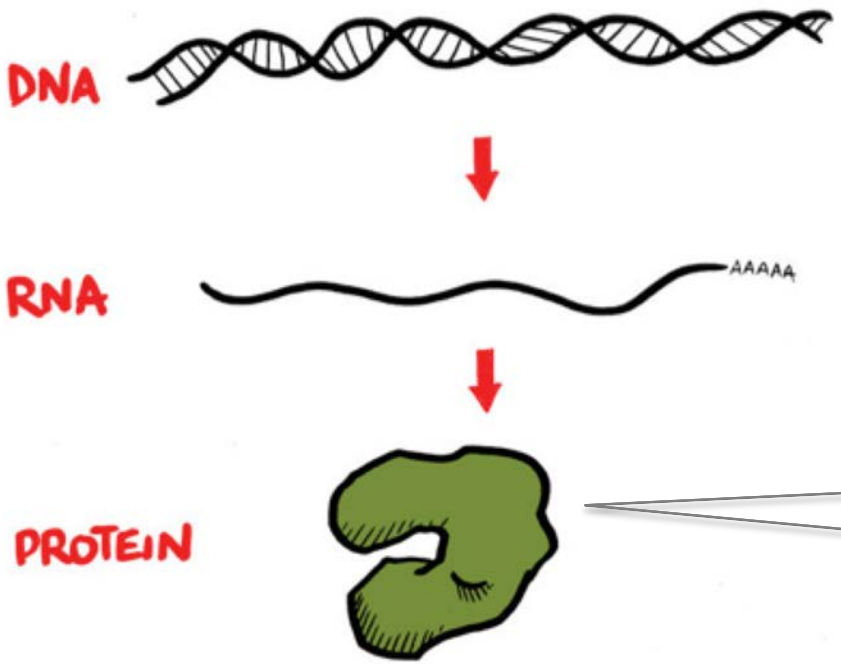

DNA
Transcription
RNA
Translation
PROTEIN
DNA
5′
3′
CGT GGA TAC ACT TTT GCC GTT TCT
GCA CCT ATG TGA AAA CGG CAA AGA
3′
5′
Transcription
mRNA 5′
CGU GGA UAC ACU UUU GCC GUU UCU
3′
Codons
Translation
Polypeptide chain
Arg Gly Tyr Thr Phe Ala Val Ser
Amino acids

1 A continuous sequence of nucleotides in the DNA…

2 …codes for a continuous sequence of amino acids in the protein.

Conclusion: With colinearity, the number of nucleotides in the gene is proportional to the number of amino acids in the protein.

DNA/RNA sequencing is far ahead



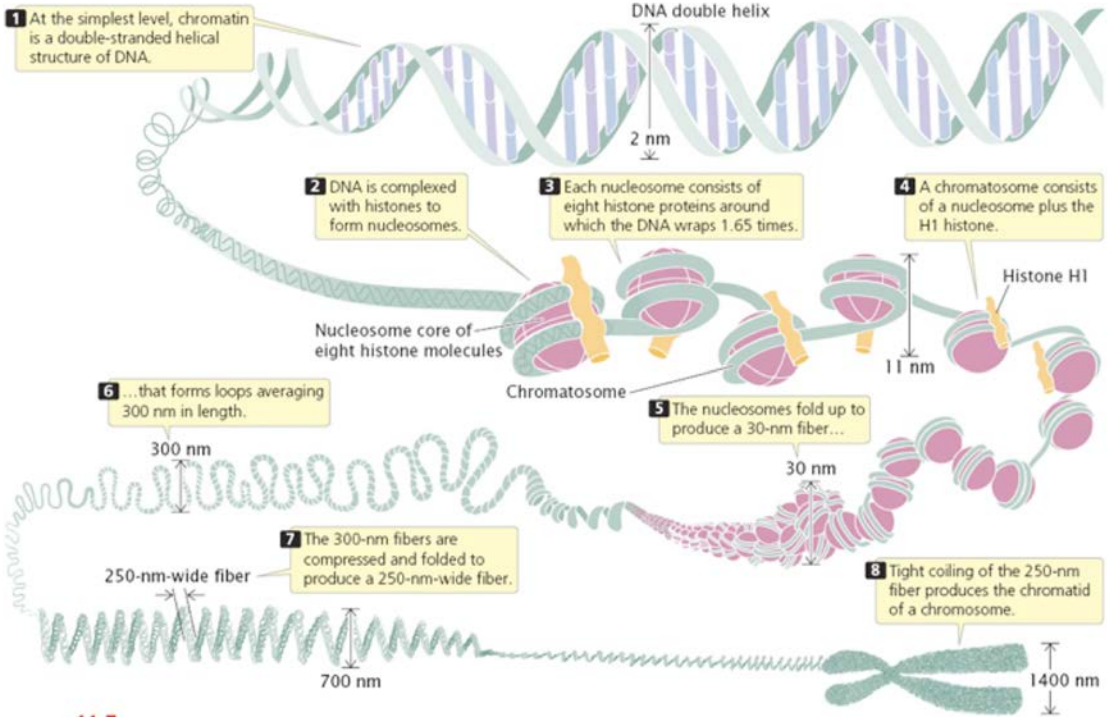
- Heritable
- Only 4 letters
- Sequence matters most

- 21 letters
- Structure matters most

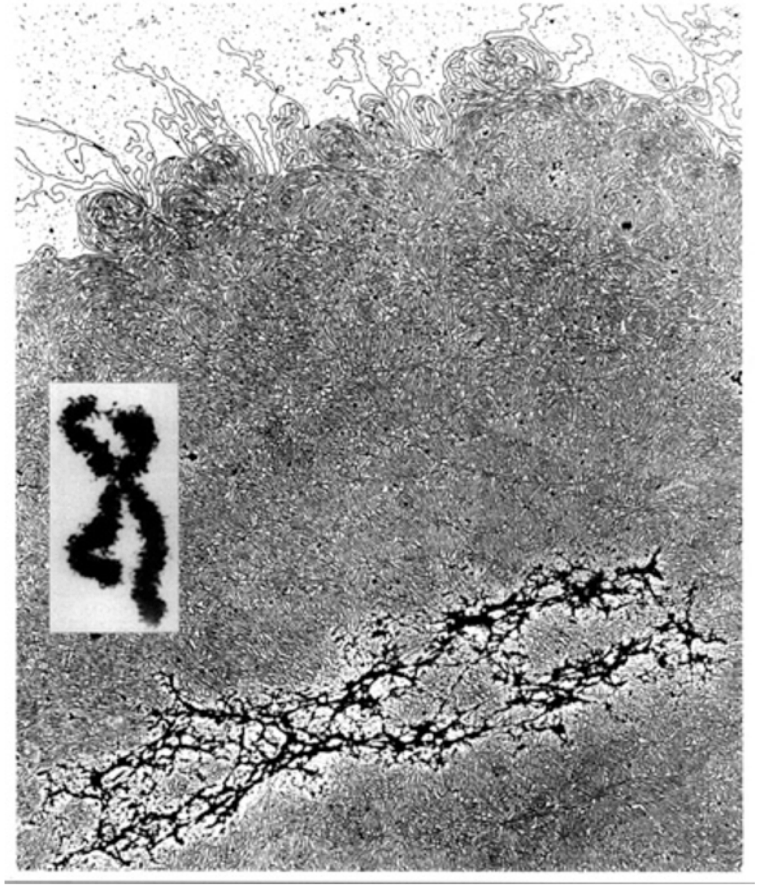
UNIVERSITÉ DE GENÈVE

FACULTÉ DE MÉDECINE

DNA, chromatin, chromosomes



1 At the simplest level, chromatin is a double-stranded helical structure of DNA.
DNA double helix
2 nm
2 DNA is complexed with histones to form nucleosomes.
3 Each nucleosome consists of eight histone proteins around which the DNA wraps 1.65 times.
4 A chromatosome consists of a nucleosome plus the H1 histone.
Histone H1
Nucleosome core of eight histone molecules
11 nm
Chromatosome
6 ...that forms loops averaging 300 nm in length.
5 The nucleosomes fold up to produce a 30-nm fiber...
300 nm
30 nm
7 The 300-nm fibers are compressed and folded to produce a 250-nm-wide fiber.
250-nm-wide fiber
8 Tight coiling of the 250-nm fiber produces the chromatid of a chromosome.
700 nm
1400 nm

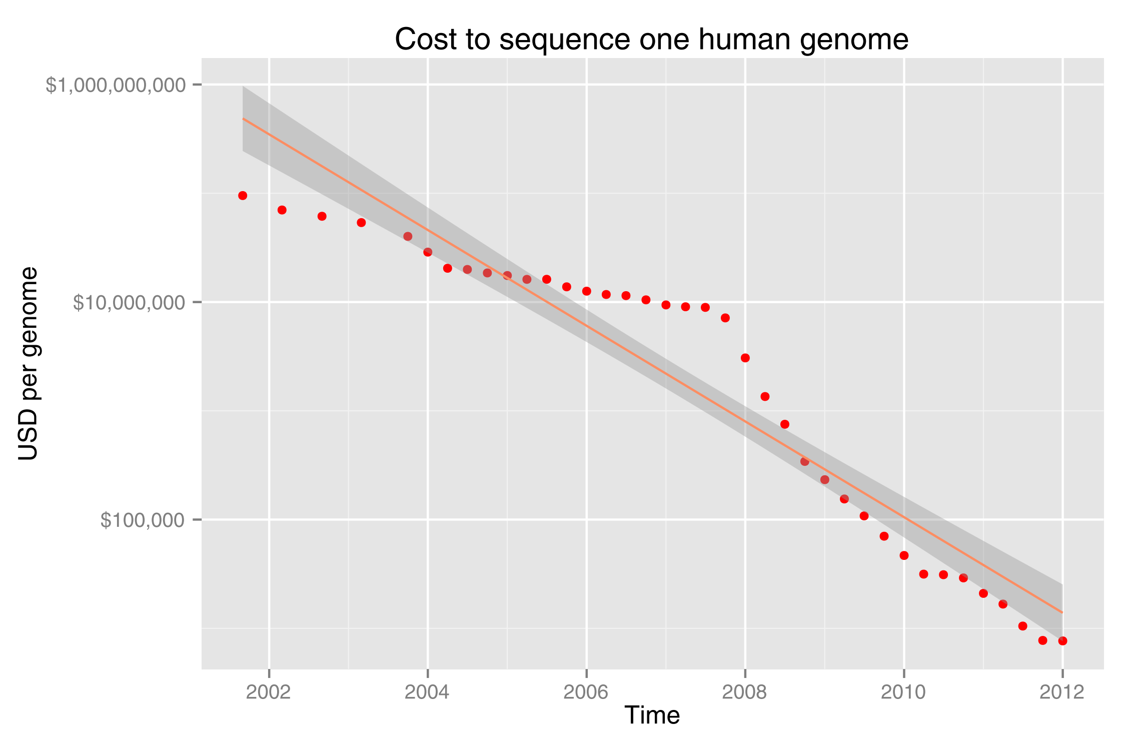


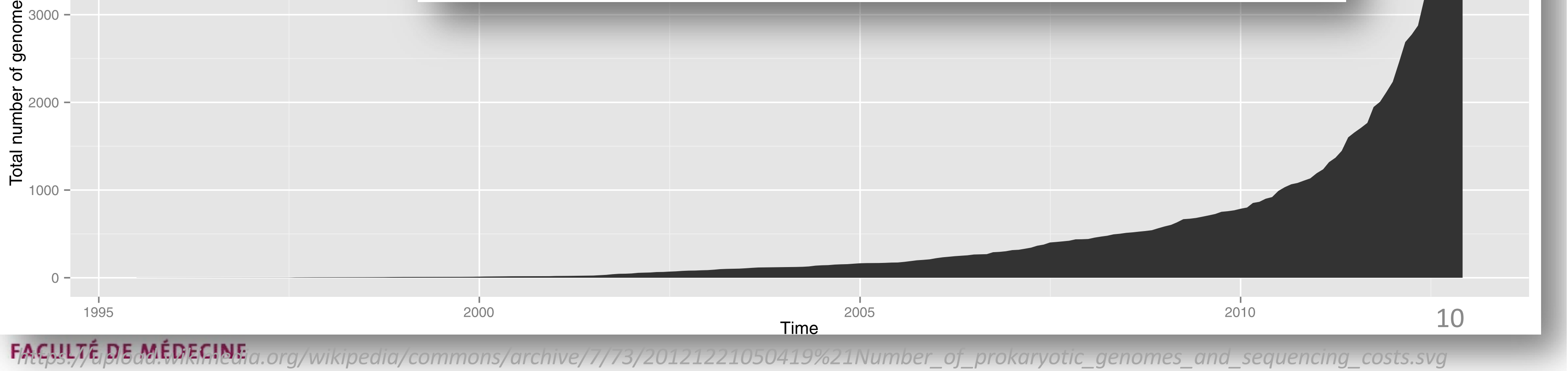
UNIVERSITÉ DE GENÈVE

FACULTÉ DE MÉDECINE

Sequencing cost is decreasing and data are being accumulated fast

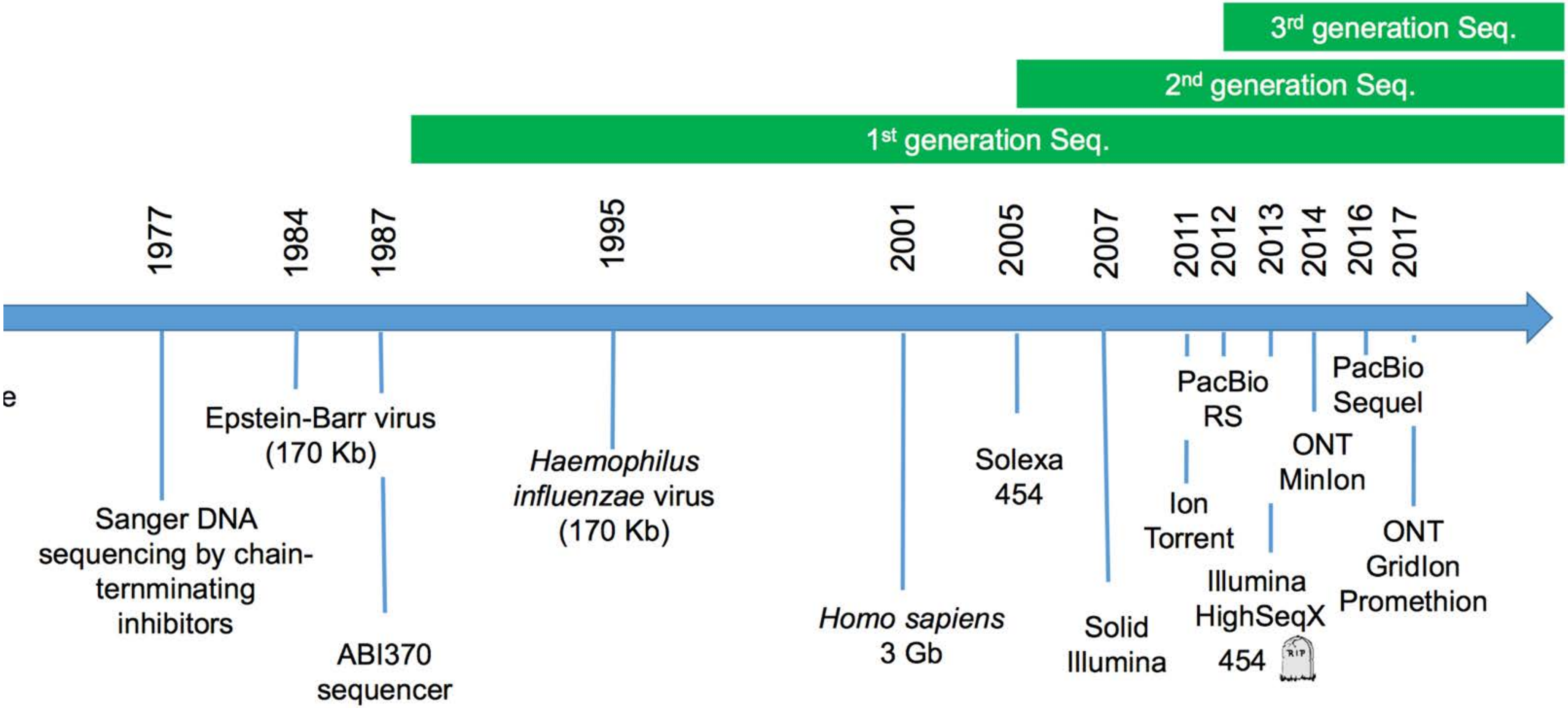
sequencing has been industrialized



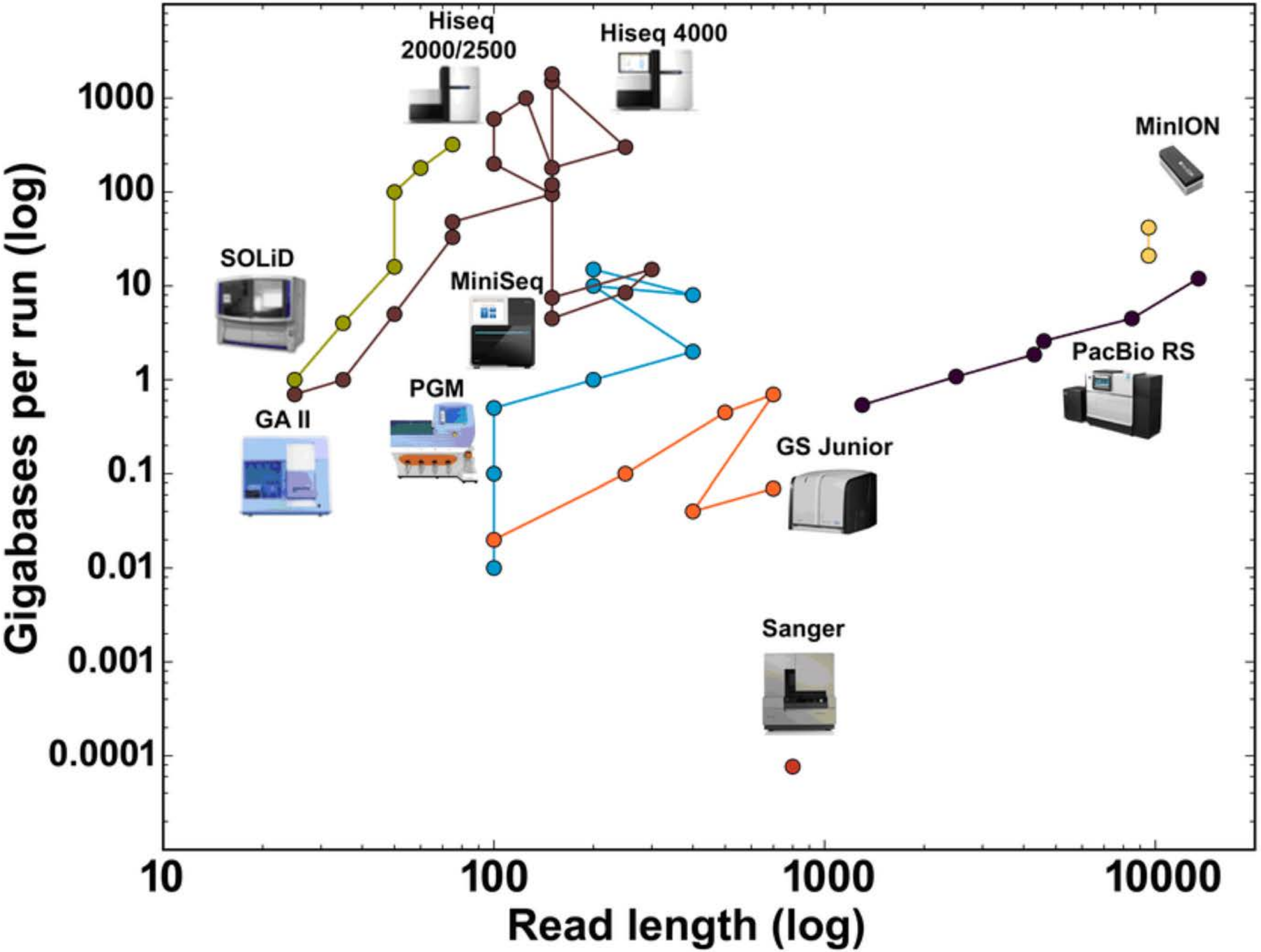


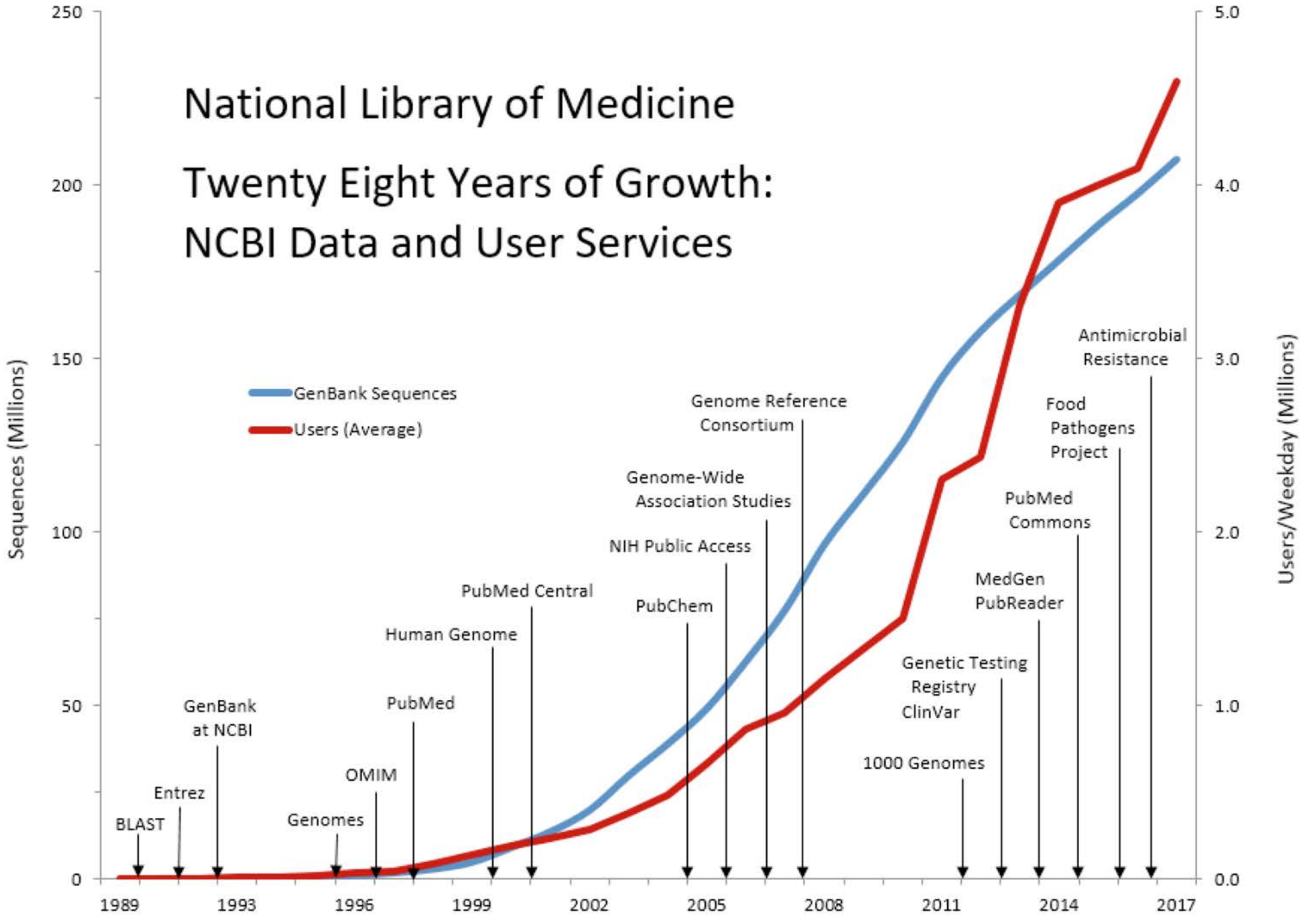
https://upload.wikimedia.org/wikipedia/commons/archive/7/73/20121221050419%21Number_of_prokaryotic_genomes_and_sequencing_costs.svg

Sequencing «Generations»



Sequencers: read length and output





National Library of Medicine

Twenty Eight Years of Growth: NCBI Data and User Services

```
CGCGACCCGAGGCTGCCGCAGGGGGCGGGCTGAGCGCGTGCGAGGCGATTGGTTTGGGGCCAGAGTGGGCGAGGCGCGGAGGTCTGGCCTATAAAGTAG
CGCGGAGACGGGGTGCTGGTTTGCGTCGTAGTCTCCTGCAGCGTCTGGGGTTTCCGTTGCAGTCCTCGGAACCAGGACCTCGGCGTGGCCTAGCGAGTT
TGGCGACGAAGGCCGTGTGCGTGCTGAAGGGCGACGGCCCAGTGCAGGGCATCATCAATTTCGAGCAGAAGGCAAGGGCTGGGACGGAGGCTTGTTTGC
AGGCCGCTCCCACCCGCTCGTCCCCCCGCGCACCTTTGCTAGGAGCGGGTCGCCCGCCAGGCCTCGGGGCCGCCCTGGTCCAGCGCCCGGTCCCGGCCC
TGCCGCCCGGTCGGTGCCTTCGCCCCCAGCGGTGCGGTGCCCAAGTGCTGAGTCACCGGGCGGGCCCGGGCGCGGGGCGTGGGACCGAGGCCGCCGCGG
GCTGGGCCTGCGCGTGGCGGGAGCGCGGGGAGGGATTGCCGCGGGCCGGGGAGGGGCGGGGGCGGGCGTGCTGCCCTCTGTGGTCCTTGGGCCGCCGCC
CGGGTCTGTCGTGGTGCCTGGAGCGGCTGTGCTCGTCCCTTGCTTGGCCGTGTTCTCGTTCCTGAGGGTCCCGCGGACACCGAGTGGCGCAGTGCCAGG
CCAGCCCGGGGATGGCGACTGCGCCTGGGCCCGCCTGGTGTCTTCGCATCCCTCTCCGCTTTCCGGCTTCAGCGCTCTAGGTCAGGGAGTCTTCGCTTT
GTACAGCTCTAAGGCTAGGAATGGTTTTTATATTTTTAAAAGGCTTTGGAAAACAAAAATACGCAACAGAGACCGTTTGTGTGACACTTTGCAGGGAAG
TTGCTGGCCTCTGTTCTAGGTCATGATTGGGCTGCAAGGGCAGAGAAGGTAGCCTTGAACAGAGGTCCTTTTCCTCCTCCTAAGCTCCGGGAGCCAGAG
TTTAACTGACCCTTTTGGGGATTTTTGAGGGCAGTGATCTTAACTTTGGGTGCACAGTTAGCTTATTTGAAGATCTTACTAAAAATACACCAGAGCCCA
CCTCCGACCAATTACATCAAAACCTGTCCTAGTGCAGGGTGAGTATTGCTGTTTTTTGAAAGTTTCCAAAAGTGATTTTGATGTGCACCTACGATTGAG
ACTGTCGTTTGAGGACAGTGGGTGGACTTTCCTATTTGGAAATTAGAAGACCTGGAGTTTCCATTACAGCCAATTGGCACTTAATAACTGTTGTCGGAG
ATTTCTTAAGCCACATTTTCGTAAAG[illegible]CAAACTTCTGAACGACTCTTGTA
AATATACAGAAATATTTTCAGAACTTTTATCAGTAAAATTACAAAACGTGTTGCAAGGAAGGTGCTTGTGATAACACTGTCCCCAGAACCTTAGTGAAG
TACCAACTGGTGGAAAATTTTCTCTTGCACTCGGCTTAAAAATCATGAGGGAATATTTACTATACGAATGAGATTCAGTCTTTAAAGGGGTTTACAGAA
CGTGAGAGGACAGGAACAGTTAGTCTGTGTAAATGTC[illegible]TTAGCCTTTTTCTTTAATAGGTCTCCAGATTTC
GGAAAAGGTTCTTTGGCATTTGACTCCATTTTGCTGTTTCATTTGTCAGACTTCTTTTTGTCCCTCTTTACTTCTCCCCACATAATTCACCAGTACTAG
GTTTTGTTTTTCAGACCAAGTCTCGCTCT[illegible]CAACCTCCGCCTCCCAGGTTCAAGCAAT
CTCCTGCCTCAGCCTCCCGGGTAGCTGGG[illegible]AGTAGAGACGGCGTTTCACCATGTTG
CCAGGATGGTCTCGATCTGTTGACGTCGTGATCCACCCGCCTCGGCCTCCCAAAGTGCTGGGATTACAGGCGTGAGCCACCCCGCCCGGCCACCAGTGC
ATTCTTAAGACGCCTCTGAGGAATCCCTTCTCCCTGGCCATTGAGAATCCATGCATGAACCCAGGTTTTCCACCTTCCCTGAGCAGCTTGCATAGTTCC
TCTTTTAAGCGCCTGACTTCGTTTTGTTTGGTGCCCGTTGTACCTGAGAATGAGCCTTGGATAGTGGAGCATTCCAGCTTTCCAGATATGCAGAGATAA
ACATTGGCTATCAGCTACTTGGCTTGGCCTATTCCGTGTTTAAAATCTTGGACTCTTTGCTAGTTTTTACAGATCAGAATTTTTCACGTATTAATCCAG
TTTCCTAGCTTCTCTTGAAGAATTTTTGGAGATCTCTTCATACTGAGCCTTCATTAGCCCAGGACAGTACTGCTGTAGCAGTTCATATATTTTTTCGCT
CCCAGGCCTGTGTTATTCAC[illegible]ATGTATTACCGGTTAAAT
GAGGTTTCAAAGAACCTGCTGTTTTTGGCCCTGTGCTCTTGATAACAGAGTGTTTGAGGGACAACTTTCACATTTGAGTTTTTCCAAAATTAAAGGTTG
AGAAGAGTCACAGTATCTATTGTCAAAAAGAAAAGAATTTAAAAAGGCAGCAATTGCCAGGATACTTCATTTGAGCAATGATATTTTCCAGTGGAAAGT
ACATCTTAAGGGTTAATGCCCCTTAACTGTTGGCCGTATTTGAAAACAAACCAAGCTAAAAACAAGAGACACTGACATGTTGTATGACGGTGTGGTGTG
ATGTTGTGTTTATTTTAGTCCTGAGATCTAGTTGTAACTTCCTTGATTTCTGTATGTAGCCACGGAGCACCATTACCTGTCACCATTACCTGAATGGCT
TACTGCTTGCTTTCATTTTGGTAGAGTGGAAAGGTTACCTAGGTTTCAGTGCTTGAAAAGATTTCAGAAAGCAGTAGTACGTCTGGTTAGACTAGAATC
GTCCTCTCCTGGGGGCAGTGGAATATAATATTTTCTGACTGCTAATTAAAAATACCTGTGATAGCCGGGCGTGGTGGCTTACGCCTGTAATCCCAGCAC
TTGGGAGGCCGAGACGGGTGGATCACGAGGTCAGCAGATGGAGACCATCCTGGCTAACACGGTGAAACCCCGTCTCTACTAAAAATGCAAAAAAATTAG
CGGGTGTGGTGGTGGGCGCCTGTAGTCCCAGCTACTCAGGAGGCTGAGGCAGGAGAATGGCATGAACCTGGGAGGCGGAGCTTGCAGTGAGCCGAGATC
TGTCACTGCACTCCAGCCTGGGCGACAGAGCGAGACTCGTCTCAAAAAAAAAAAAGAAAAAAACTTATGATGGACACTTAAAAACACTCACTGAGTGGGG
GTGGAGAGCAGGGGTCCCAGGGTAGCCTGTTGGACATTTCCAGGGCGACTTTTTCTTTTTTTTTTTTTAAAGTCAAGTGAGTATGCCATATGGAAAAGG
TGTGCCGTGGAGAAAAAAGCAAGGGGCTCCAGAGTGTAGGATGAGACATACACCTTTTGGGTTAAAAAGGCTGAGGCAGGAGAATGGCGTGAACCCGGGAG
CGGAGCTTGCAGTGAGCTGAGATCATGCCACTGCACTCCAGCCTGGGCGACAGAGCGAGACTCTTGTCTCAAAATAAAAAACGTTTACATGTACATGTA
```

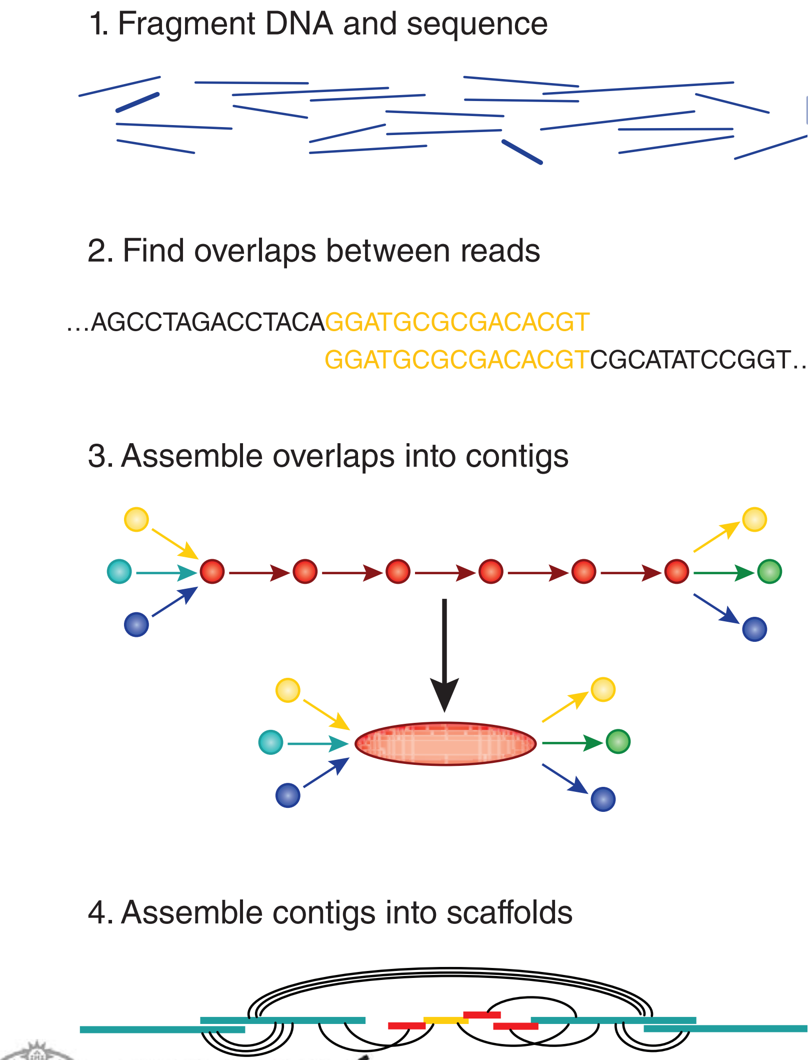
Without interpretation

(by comparisons)

DNA is unintelligible

=> sequence analysis required!

Some assembly required



1. Fragment DNA and sequence

2. Find overlaps between reads

...AGCCTAGACCTACAGGATGCGCGACACGT
GGATGCGCGACACGTCGCATATCCGGT..

3. Assemble overlaps into contigs

4. Assemble contigs into scaffolds

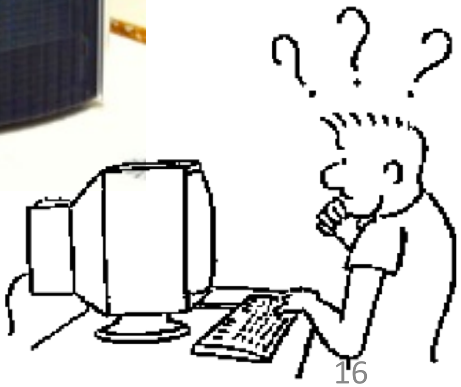
Genome assembly stitches together a genome from short sequenced pieces of DNA.

Genomics is unthinkable without computer data analysis

just one genome

our computer

computers can only execute human intelligence



The promise: i.e. why we are here

Musings

Highly accessed

The $1,000 genome, the $100,000 analysis?

Elaine R Mardis

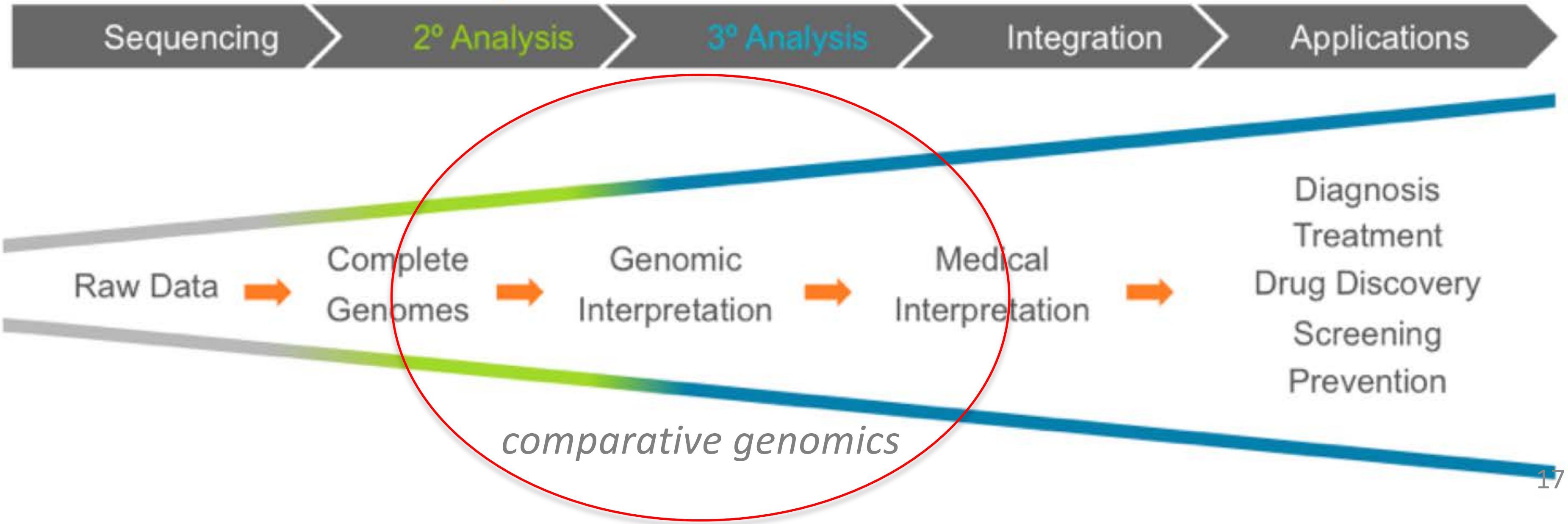
Correspondence: Elaine R Mardis emardis@wustl.edu

Author Affiliations

The Genome Center at Washington University School of Medicine, 4444 Forest Park Blvd, St Louis, MO 63108, USA

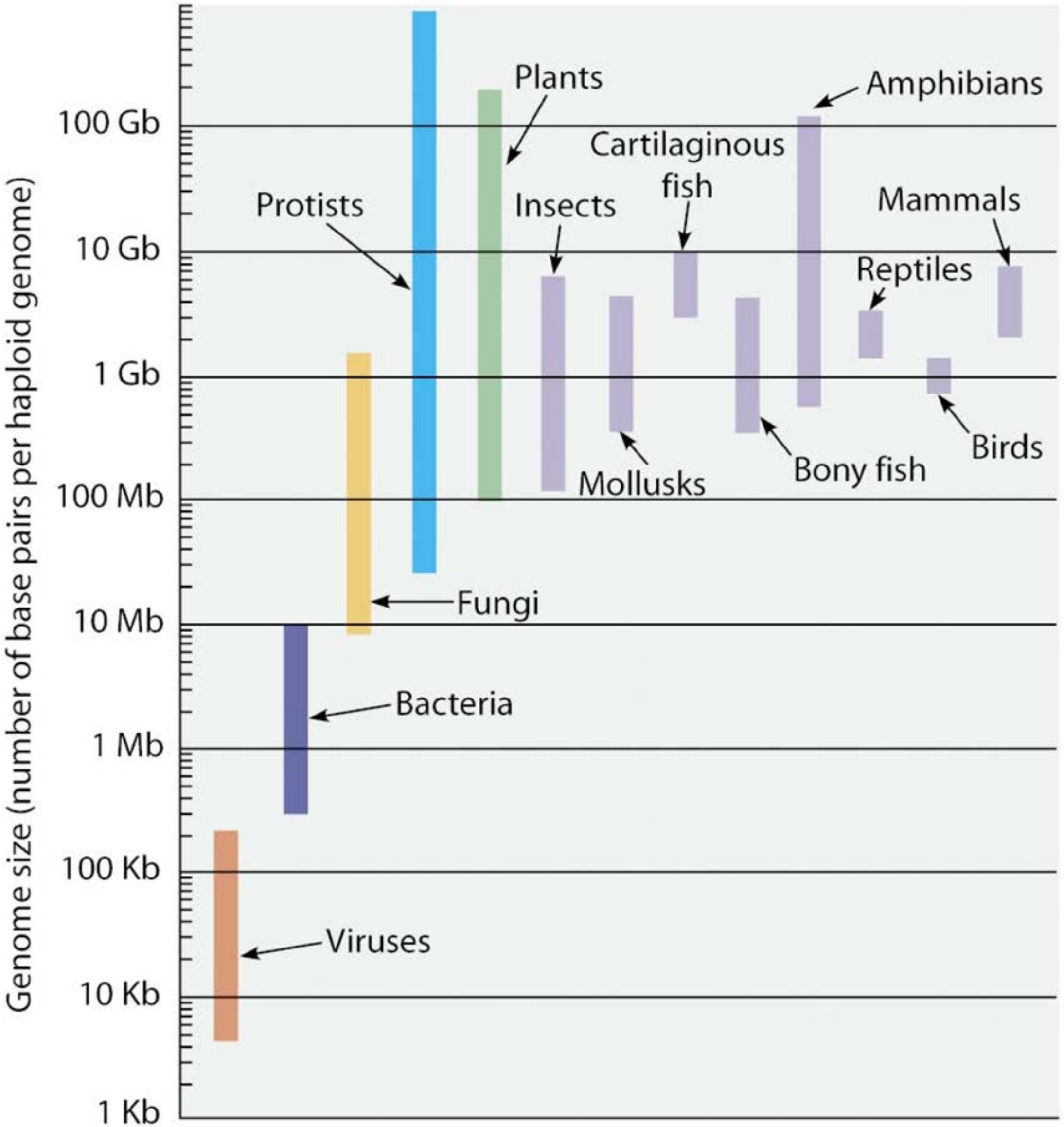
Genome Medicine 2010, **2**:84 doi:10.1186/gm205

Genomics Landscape for future medicine



comparative genomics

Genome sizes



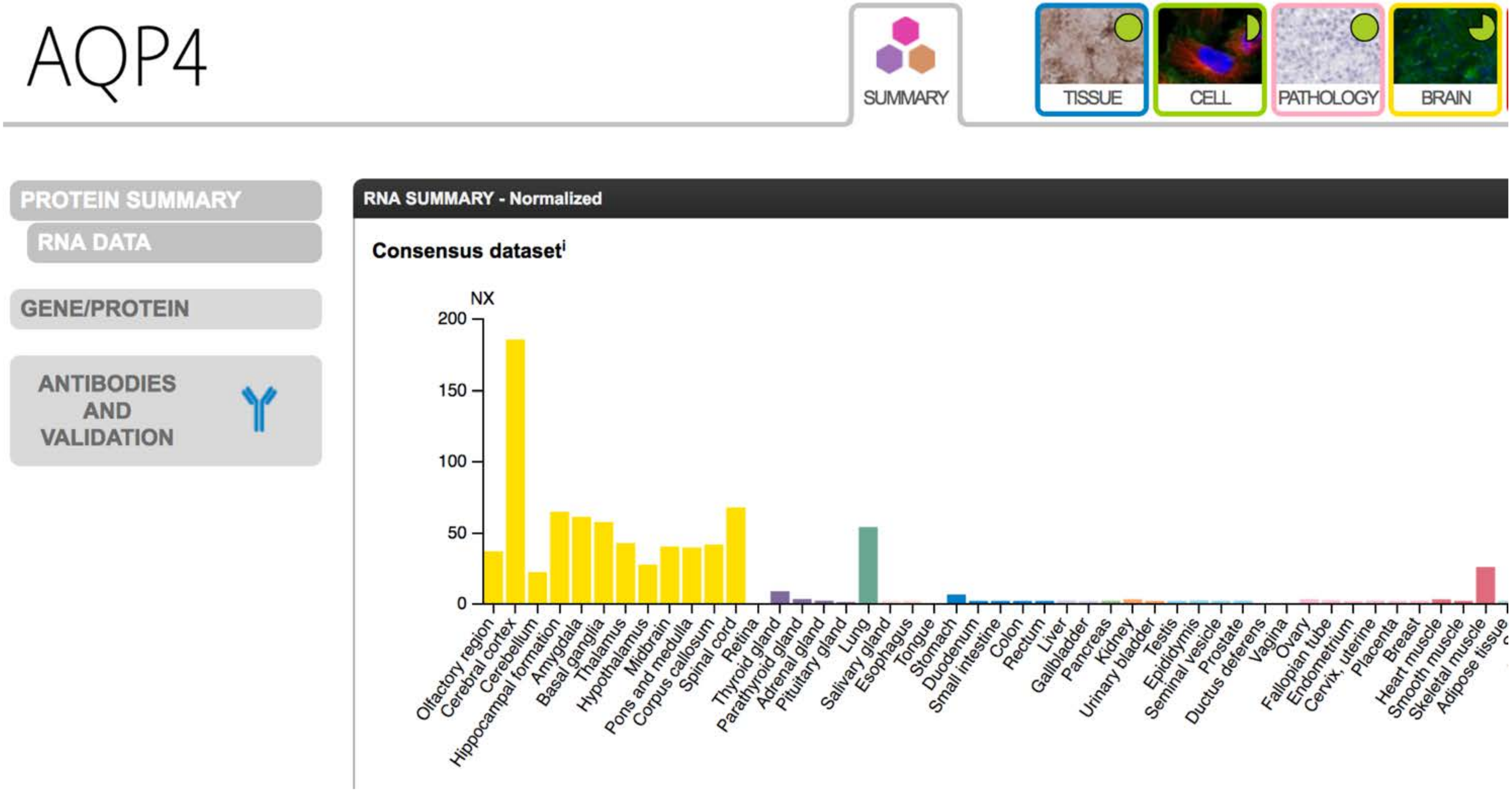
Genomics “Holy Grail”: predicting phenotypic variability from genetic variability

UNIVERSITÉ
DE GENÈVE
FACULTÉ DE MÉDECINE

Gene expression

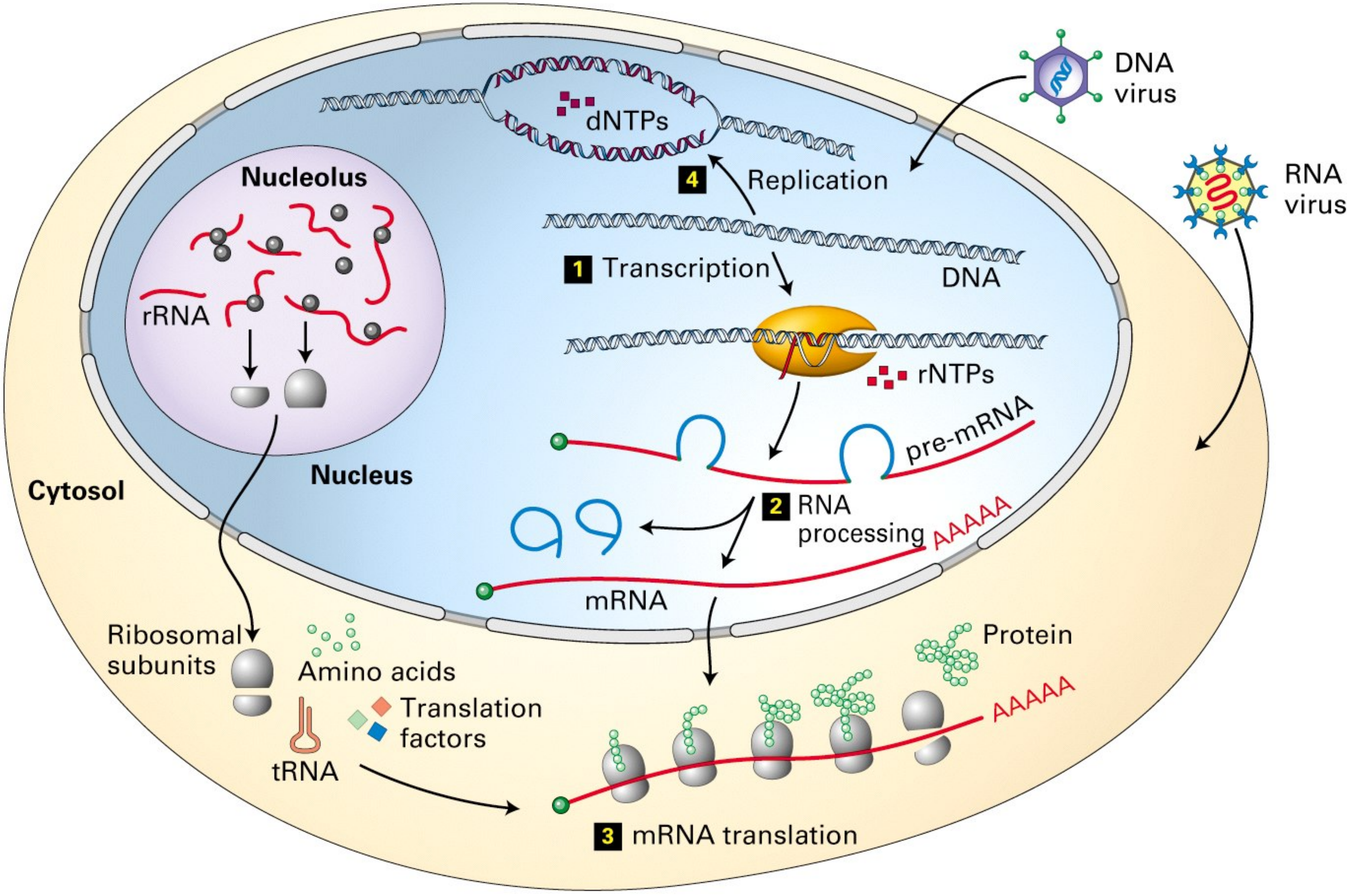
- Proxy to cell functions (via proteins)
- Not all genes expressed
- Highly uneven expression levels

Gene expression



www.proteinatlas.org

Not only genome can be sequenced



Genomics keywords

- **DNA-Seq** is sequencing DNA in the sample
- **RNA-Seq** is sequencing RNA in the sample
- **ChIP-Seq** is sequencing DNA sites interacting with specific protein
- **Meta-genomics** is sequencing many organisms in one sample

Why genomics?

+ “Complete” cellular DNA/RNA snapshot,

+ Protein/NA & NA/NA interactions

+ relative abundance of “reads”

+ **wealth of data**

- it doesn’t tell you about biology;
proteins, interactions, metabolites, etc.;
not even which sequences are meaningful
and which not.

Biological systems

- **Ecosystem**
- **Population**
- Organism
- Organ
- Tissue
- Cell
- Complexes/networks
- Molecules

UNIVERSITÉ
DE GENÈVE
FACULTÉ DE MÉDECINE

Metagenomics: direct sequencing of total DNA/RNA

a mix of intestinal bacteria

www.micronaut.ch

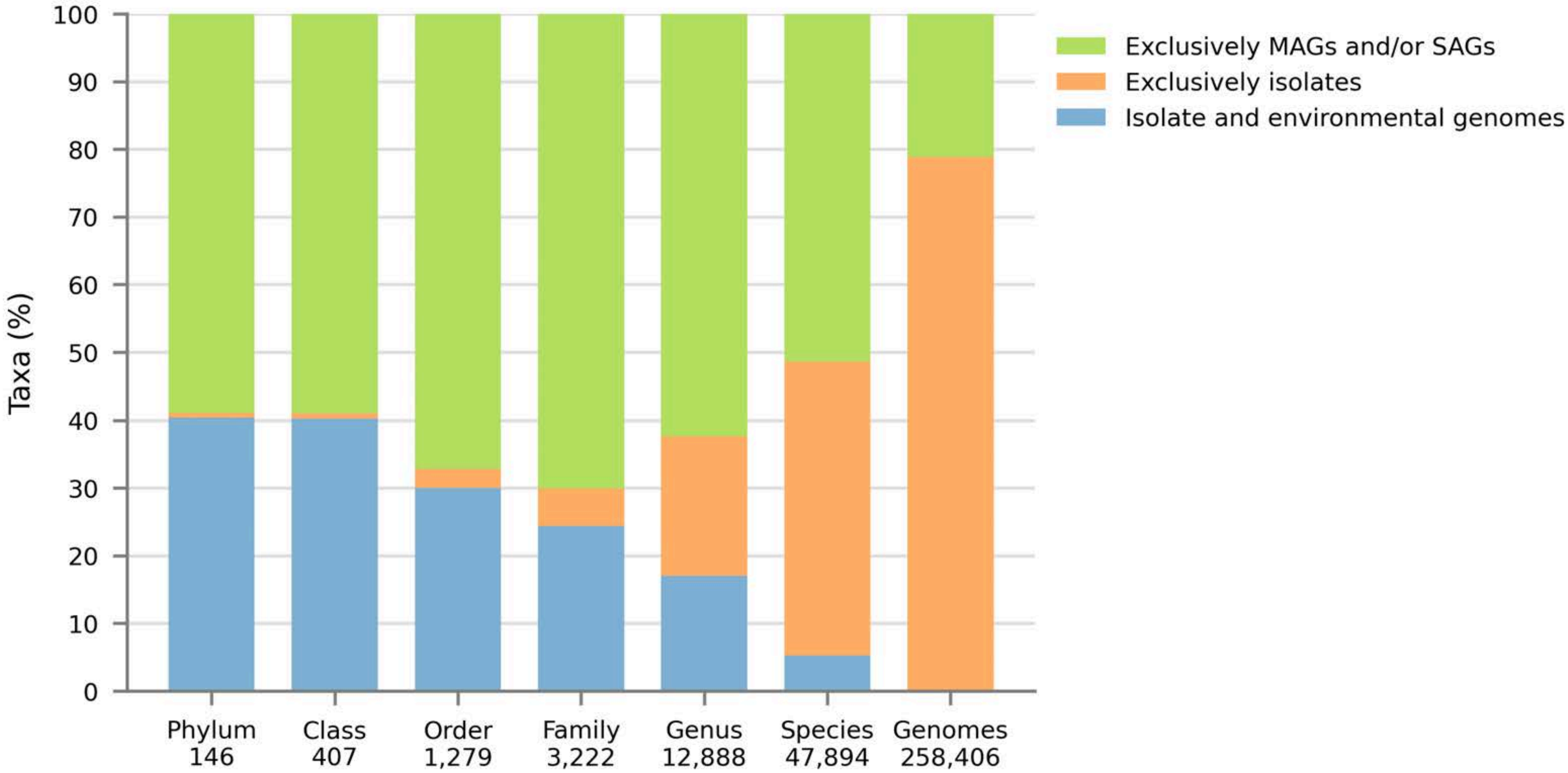
can be sequenced without culturing

Scaling-up and mixing the puzzles

- Who is there?
  or What they can do?
- How many?

UN
DE GENÈVE
FACULTÉ DE MÉDECINE

Bacterial and archaeal genomes



UNIVERSITÉ DE GENÈVE

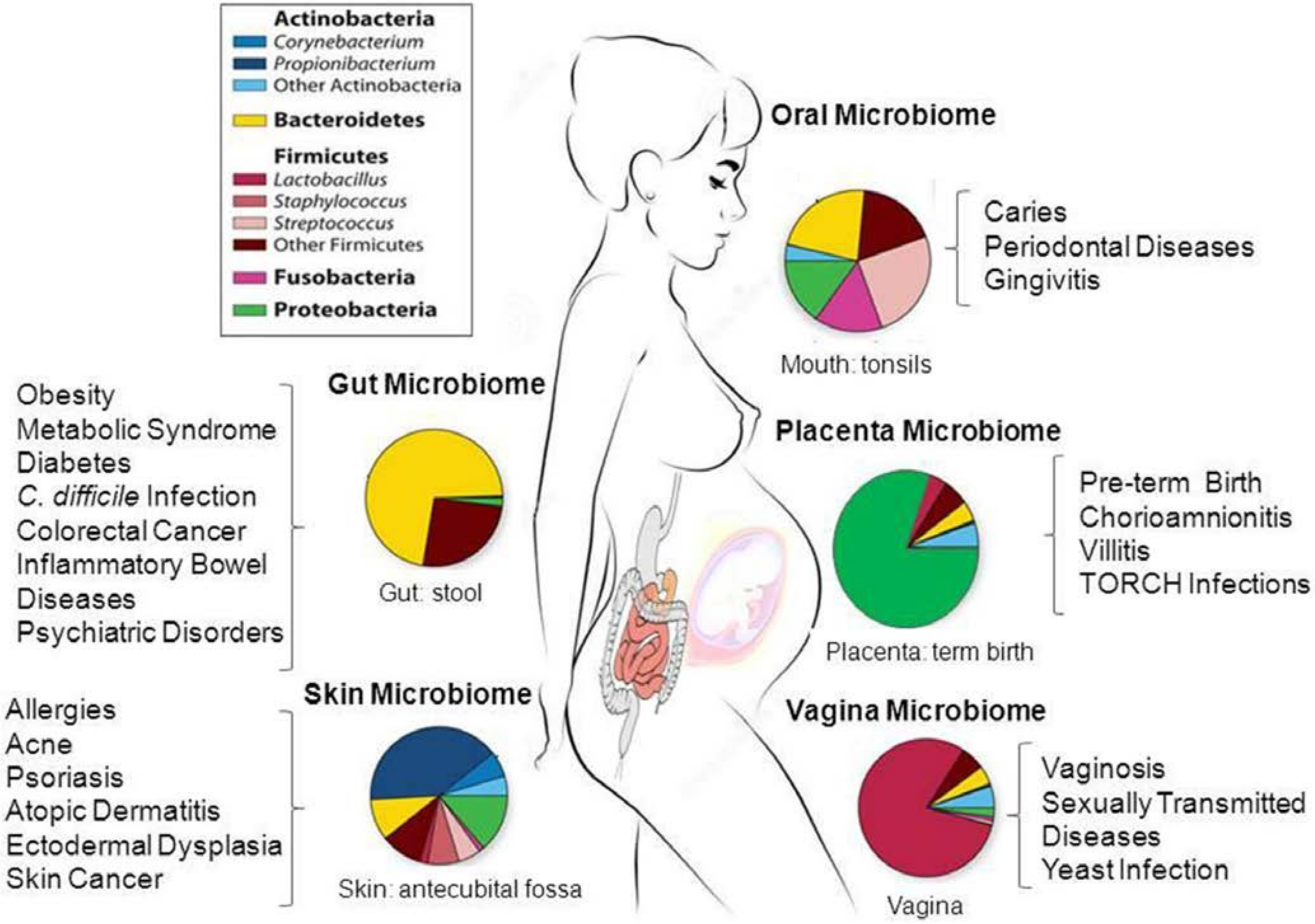
FACULTÉ DE MÉDECINE

gtdb.ecogenomic.org/stats

Recent trends in genomics

- cancer / clinical human variations
- metagenomics
- single-cell and spatial transcriptomics

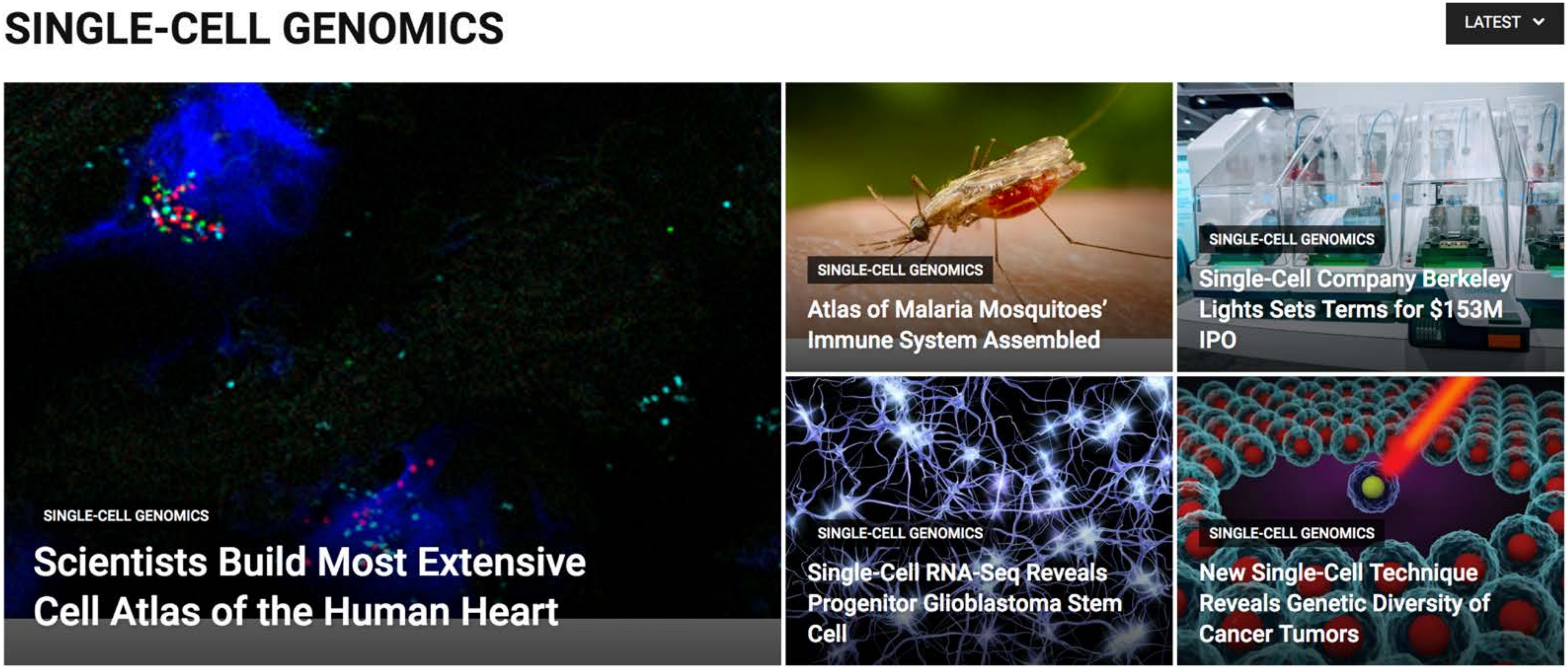
Human microbiomes



doi: 10.3389/fmicb.2015.01050

Earth Microbiome Project: Mapping the microbiome of... everything

by University of California - San Diego



SINGLE-CELL GENOMICS

LATEST

SINGLE-CELL GENOMICS

Scientists Build Most Extensive Cell Atlas of the Human Heart

SINGLE-CELL GENOMICS

Atlas of Malaria Mosquitoes' Immune System Assembled

SINGLE-CELL GENOMICS

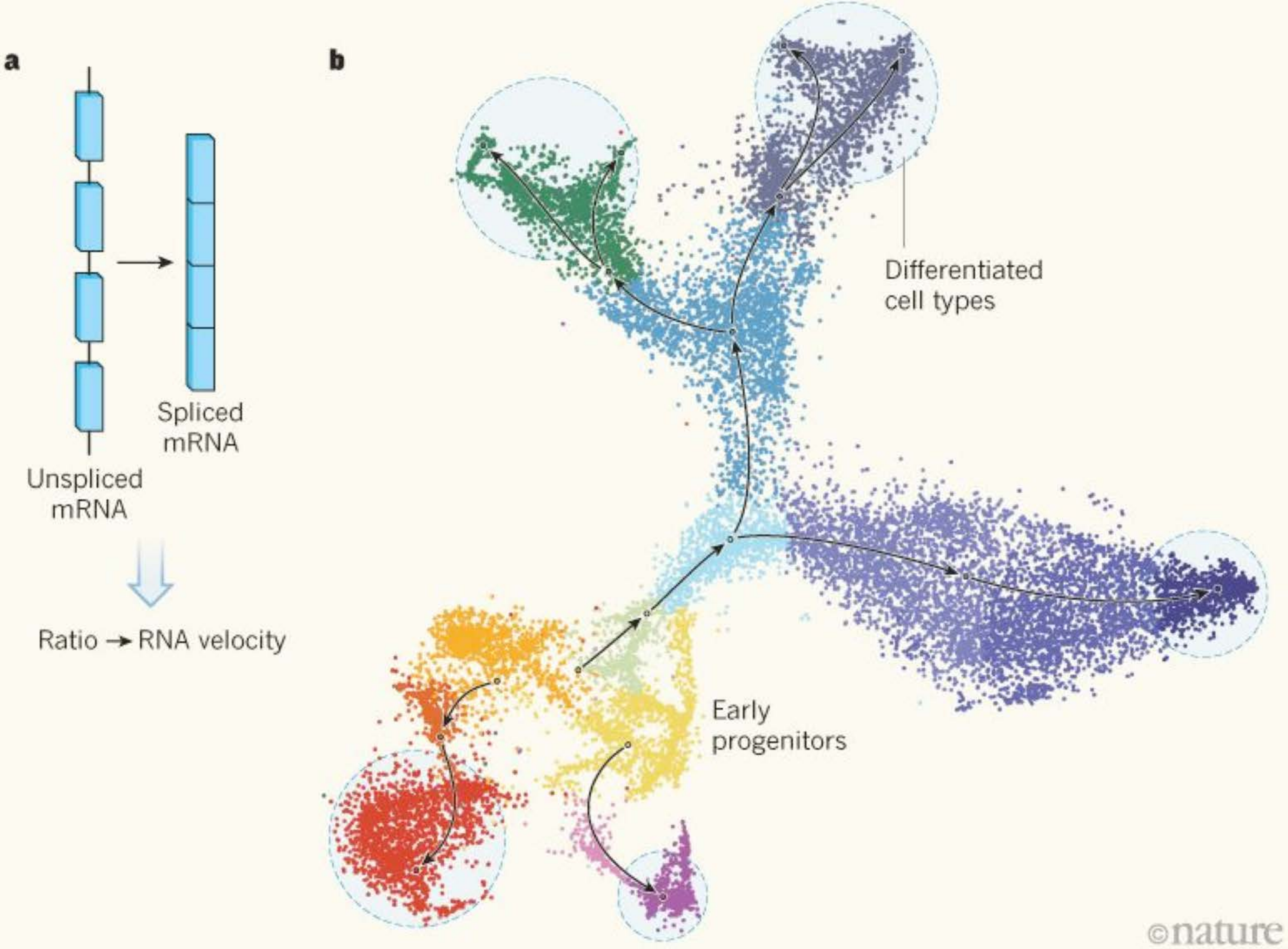
Single-Cell Company Berkeley Lights Sets Terms for $153M IPO

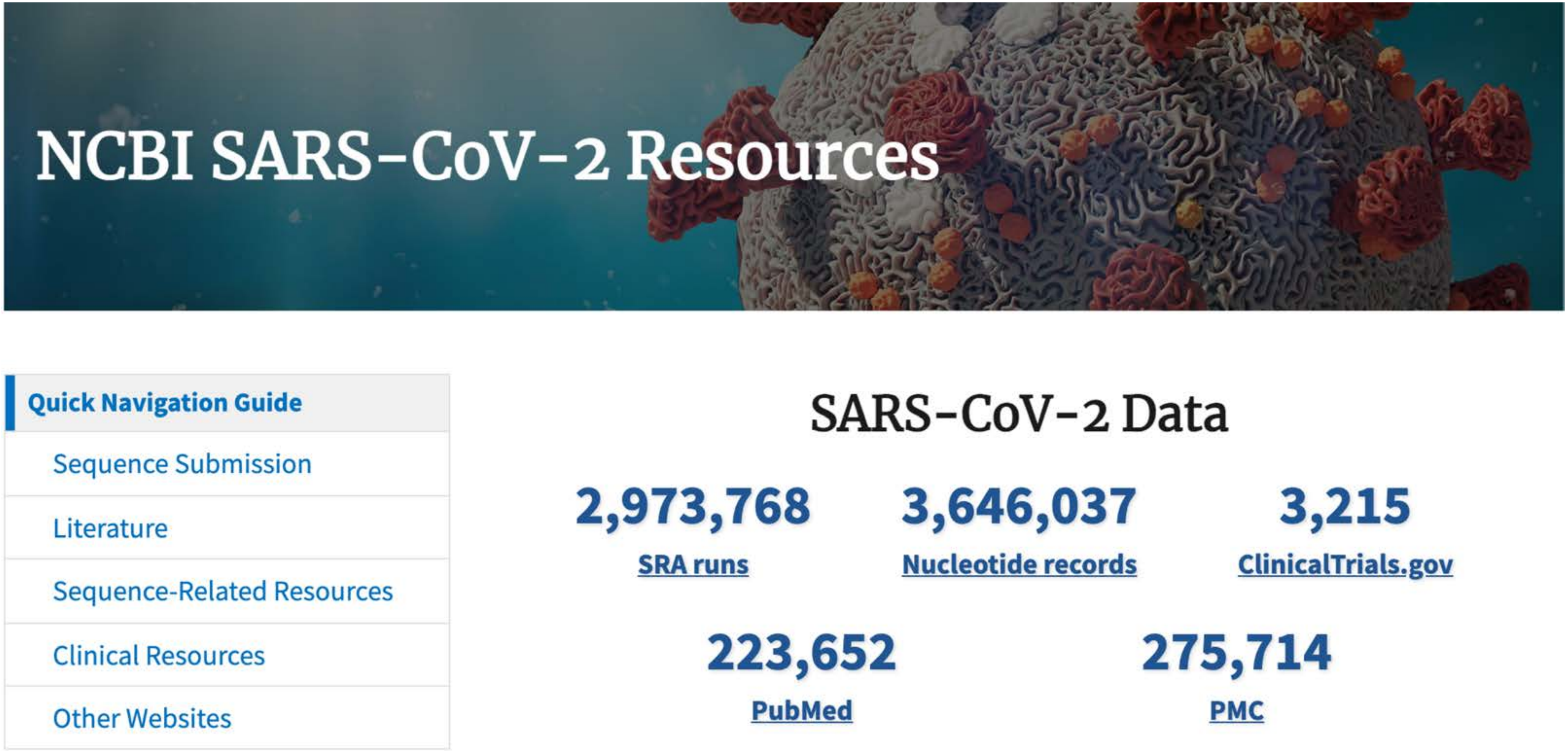
SINGLE-CELL GENOMICS

Single-Cell RNA-Seq Reveals Progenitor Glioblastoma Stem Cell

SINGLE-CELL GENOMICS

New Single-Cell Technique Reveals Genetic Diversity of Cancer Tumors


a
b
Spliced
mRNA
Unspliced
mRNA
Ratio → RNA velocity
Differentiated
cell types
Early
progenitors
©nature



NCBI SARS-CoV-2 Resources

Quick Navigation Guide

- Sequence Submission
- Literature
- Sequence-Related Resources
- Clinical Resources
- Other Websites

SARS-CoV-2 Data

2,973,768
SRA runs

3,646,037
Nucleotide records

3,215
ClinicalTrials.gov

223,652
PubMed

275,714
PMC

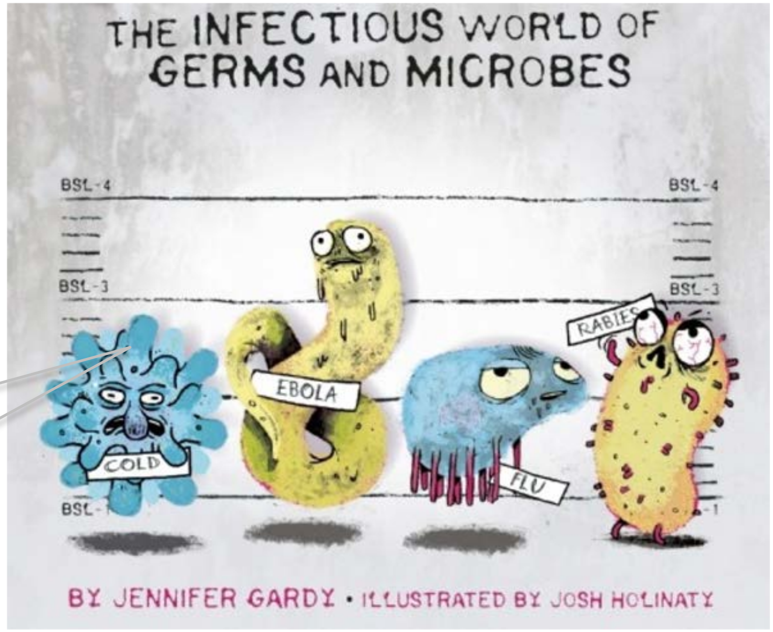
Comparative approaches

It all started with microbes

Research requires

a tool to see, and

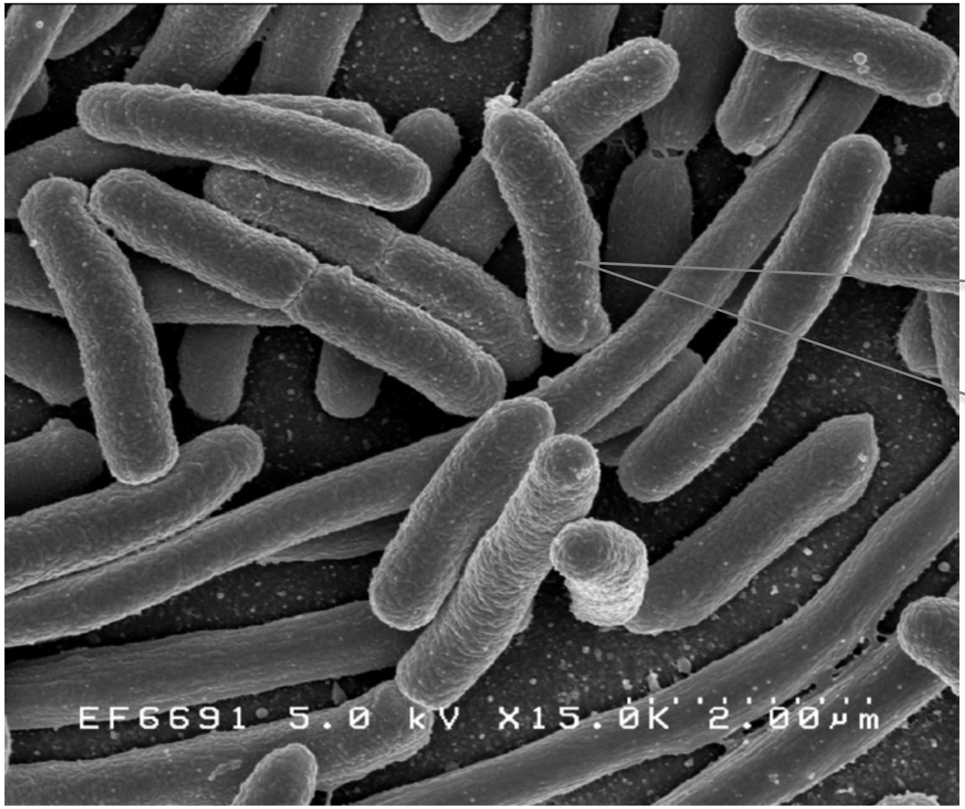
bookkeeping
the comparisons



and model organisms

UNIVERSITÉ
DE GENÈVE
FACULTÉ DE MÉDECINE

Looking same but some are pathogenic, requiring molecular-level investigations



Three strains of *E. coli* can have only 40% genes in common..

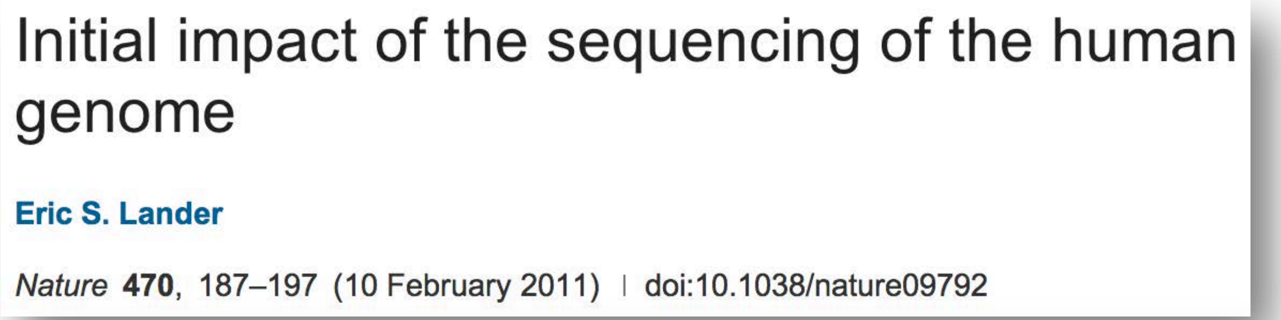
UNIVERSITÉ DE GENÈVE

FACULTÉ DE MÉDECINE

Welch, R.A. (2002). Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. Proceedings of the National Academy of Sciences, 99(26), 17020-17024.

Comparative genomics is about comparing the genomic features of different organisms.

An example



Initial impact of the sequencing of the human genome

Eric S. Lander

Nature **470**, 187–197 (10 February 2011) | doi:10.1038/nature09792

The sequence of the human genome has dramatically accelerated biomedical research.

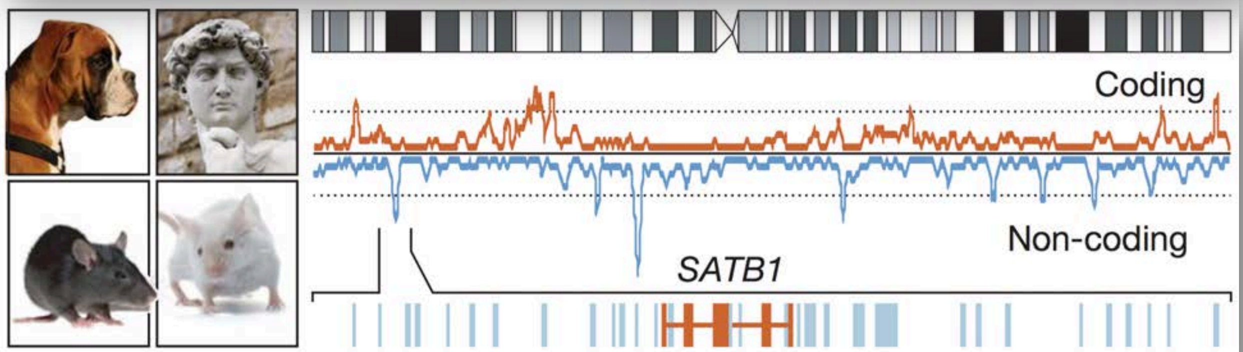


Figure 1 | **Evolutionary conservation maps.** Comparison among the human, mouse, rat and dog genomes helps identify functional elements in the genome.

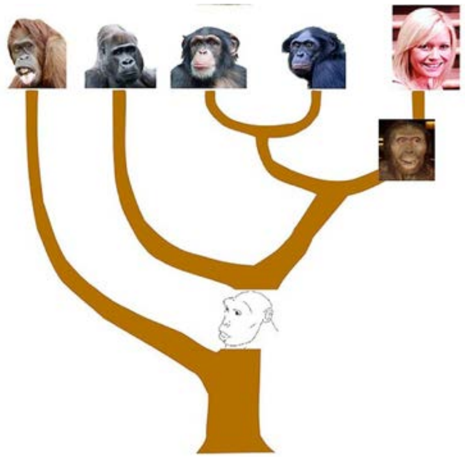
The aims are to:

- *Enable knowledge transfer, e.g. from models to human*
- *Interpret Nature's molecular experimentation*

How genomes evolve

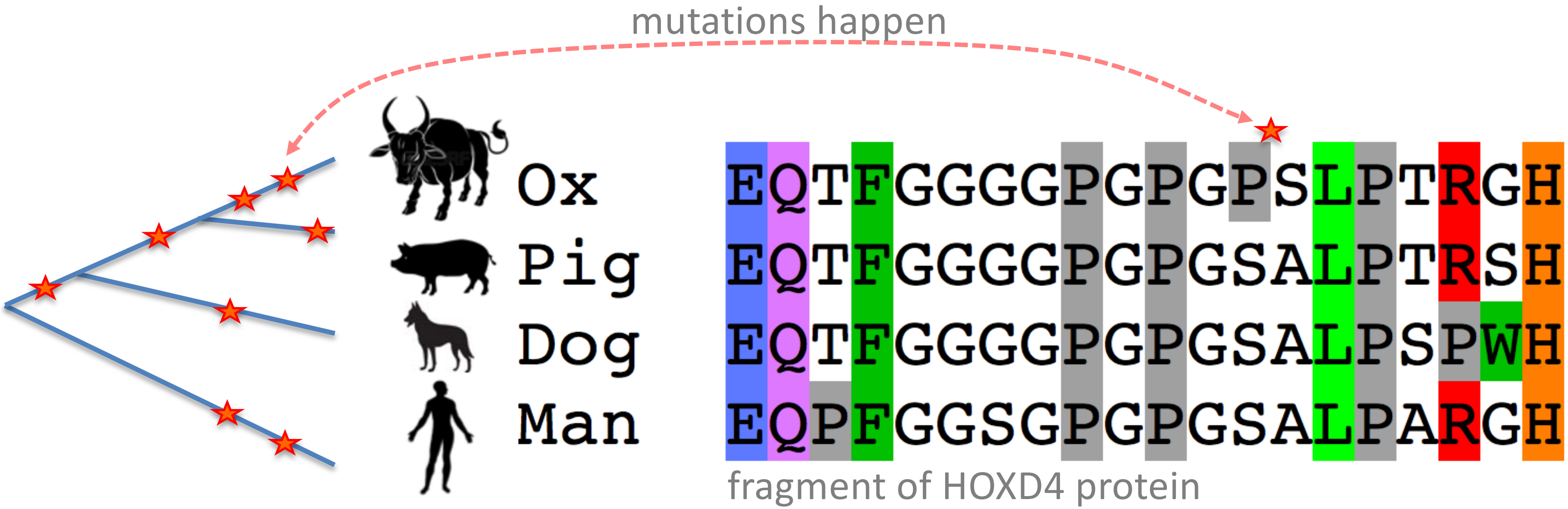
- Accumulation of mutations ⇔ divergence

- Vertical descent by speciation



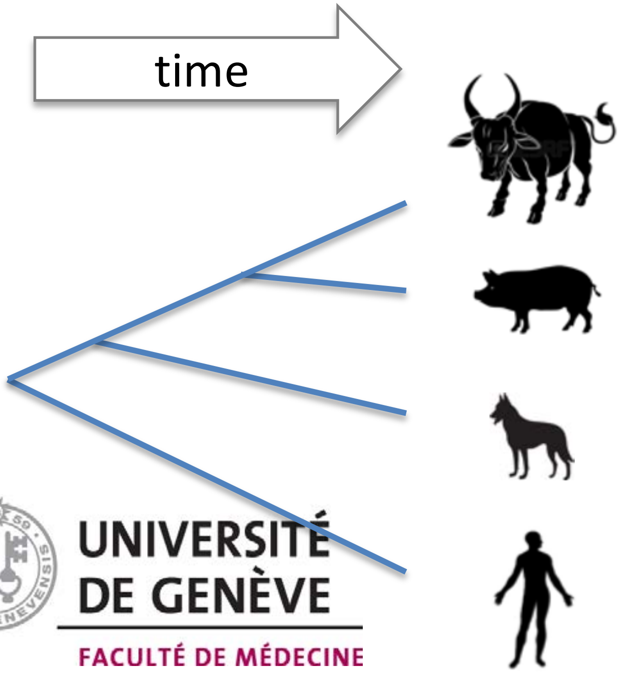
UNIVERSITÉ DE GENÈVE
FACULTÉ DE MÉDECINE

Inheritance of sequence and function

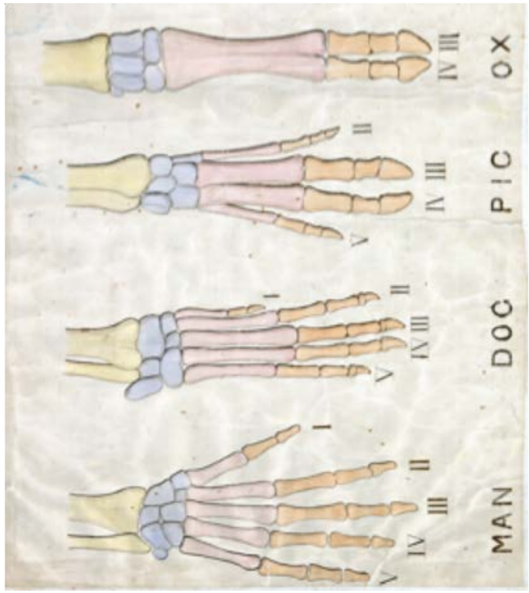

mutations happen

```
Ox   EQTFGGGGPGPGPSLPTRGH
Pig  EQTFGGGGPGPGSALPTRSH
Dog  EQTFGGGGPGPGSALPSPWH
Man  EQPFGGSGPGPGSALPARGH
```

fragment of HOXD4 protein



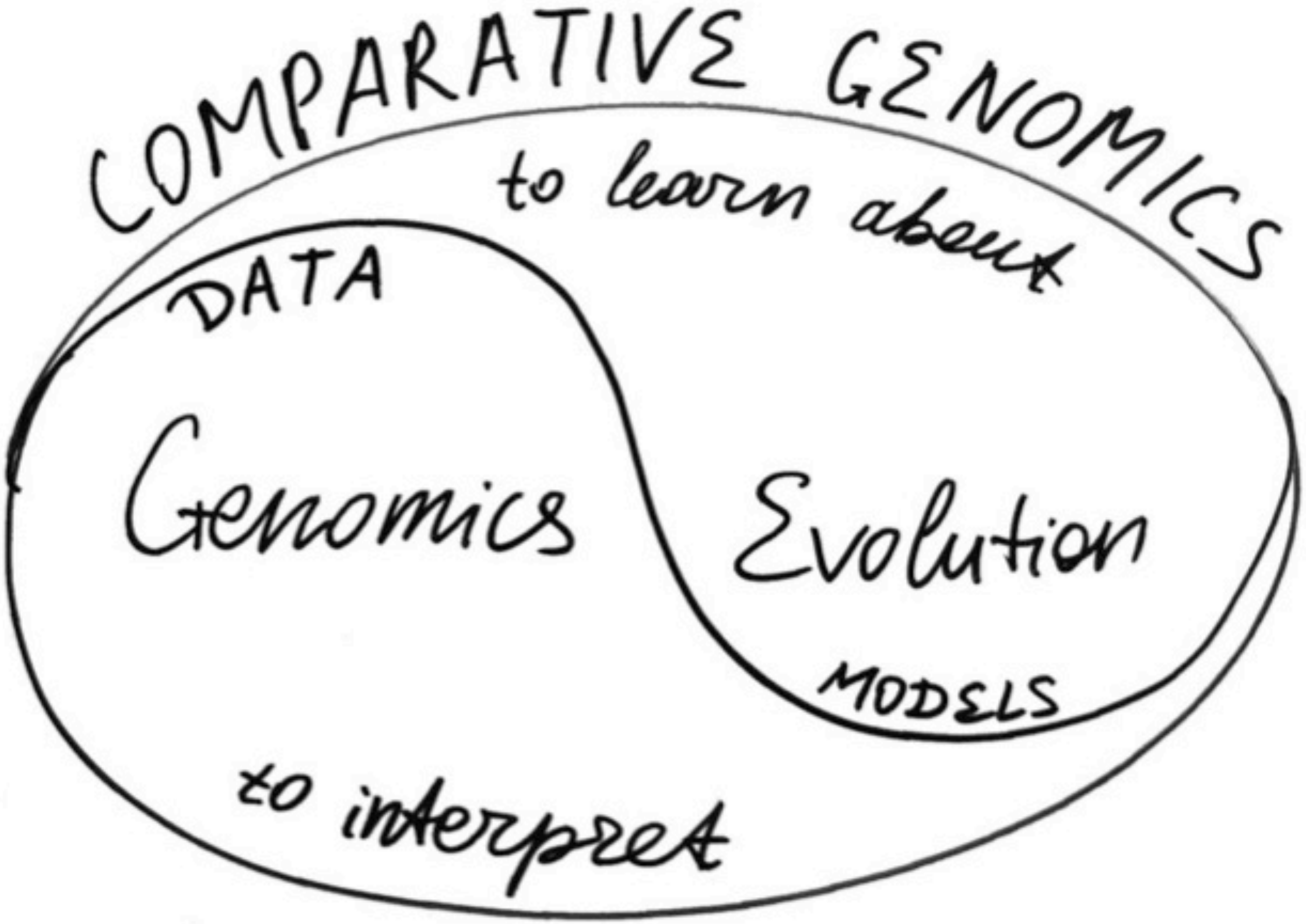
homology of forelimbs



functional selection accepts or rejects mutations

How to get there:

employing knowledge to interpret genomes and using genomes to further our knowledge



General aims

- **Similarities** allow to transfer our knowledge from well studied model organisms to the newly sequenced ones

- **Differences** may shed light on unique species adaptation processes

- Similarity

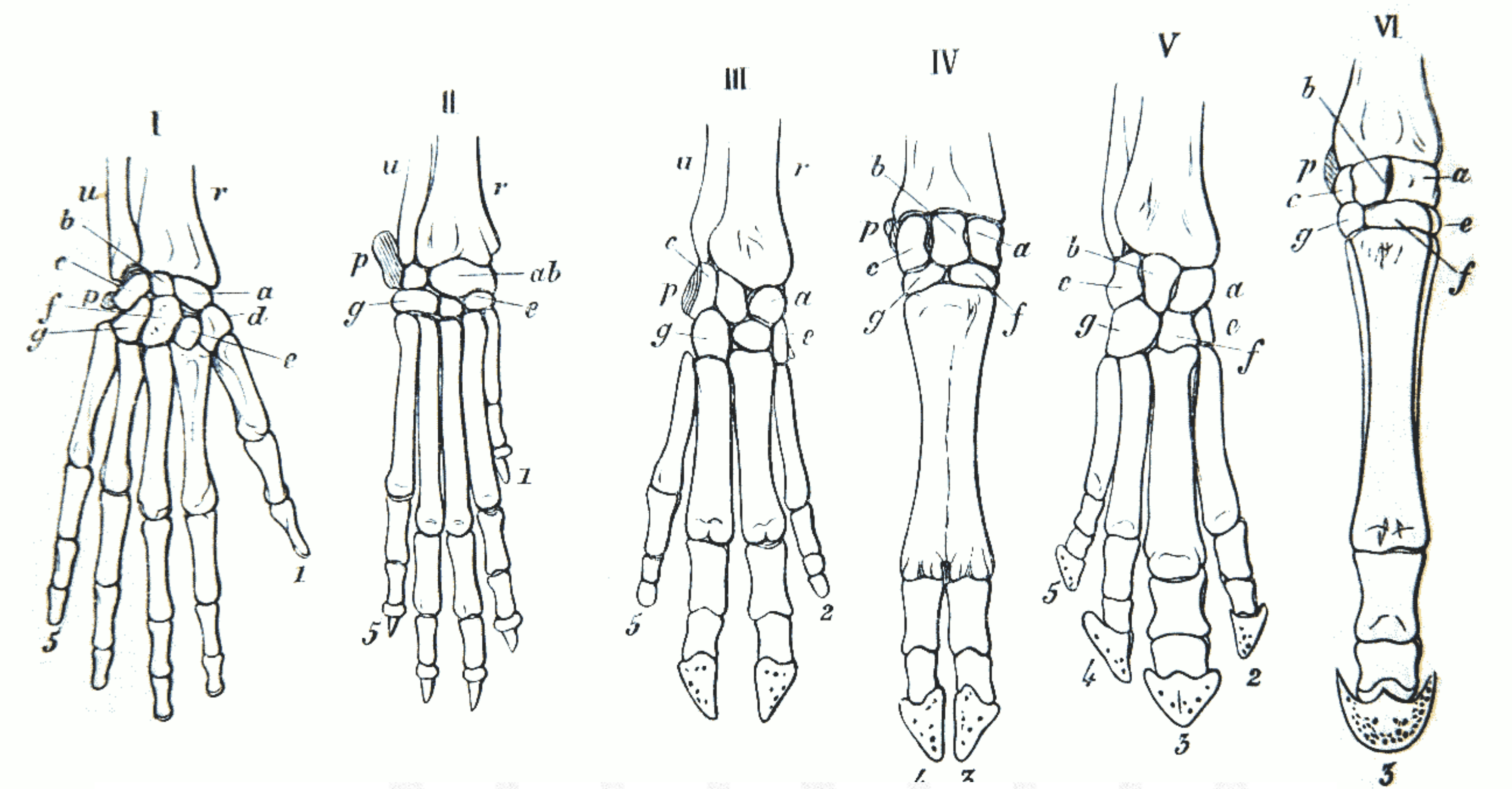
  vs

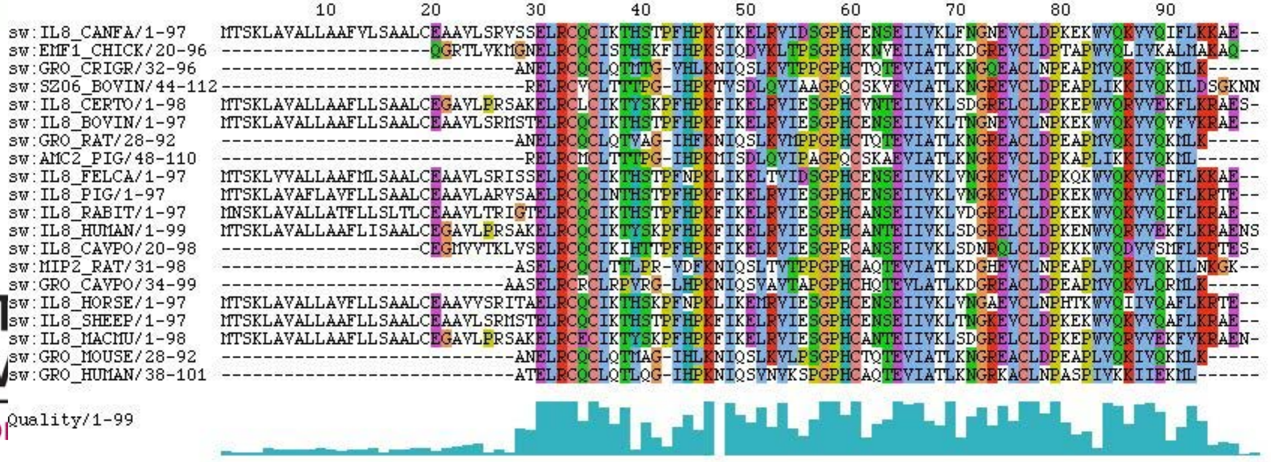
- Homology

  vs

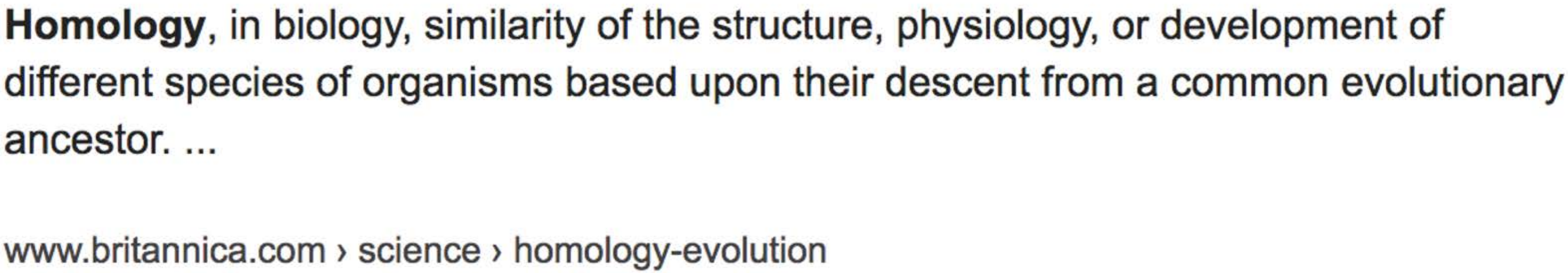
- Orthology

How would you compare?





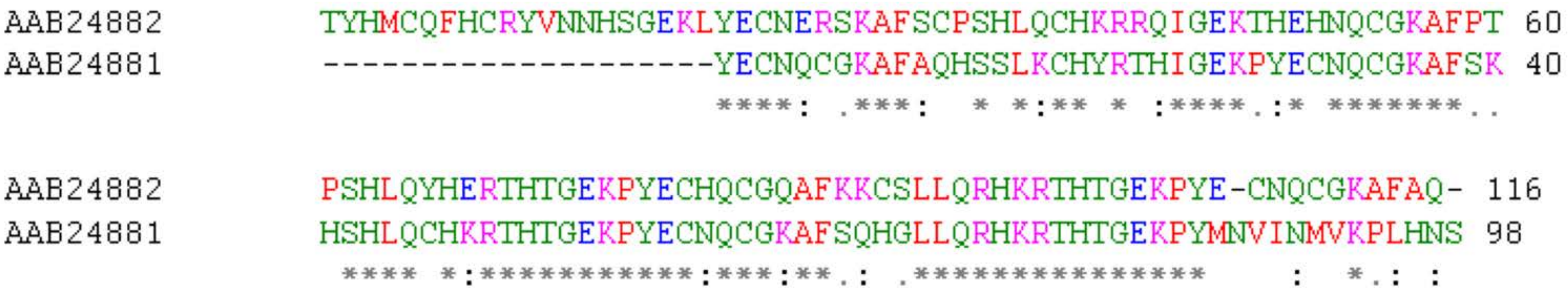
```
                            10        20        30        40        50        60        70        80        90
sw:IL8_CANFA/1-97   MTSKLAVALLAAFVLSAALCEAAVLSRVSSELRCQCIKTHSTPFHPKYIKELRVIDSGPHCENSEIIVKLFNGNEVCLDPKEKWVQKVVQIFLKKAE--
sw:EMF1_CHICK/20-96 --------------------QGRTLVKMGNELRCQCISTHSKFIHPKSIQDVKLTPSGPHCKNVEIIATLKDGREVCLDPTAPWVQLIVKALMAKAQ--
sw:GRO_CRIGR/32-96  ----------------------------ANELRCQCLQTMTG-VHLKNIQSLKVTPPGPHCTQTEVIATLKNGQEACLNPEAPMVQKIVQKMLK-----
sw:SZ06_BOVIN/44-112-----------------------------RELRCVCLTTTPG-IHPKTVSDLQVIAAGPQCSKVEVIATLKNGREVCLDPEAPLIKKIVQKILDSGKNN
sw:IL8_CERTO/1-98   MTSKLAVALLAAFLLSAALCEGAVLPRSAKELRCLCIKTYSKPFHPKFIKELRVIESGPHCVNTEIIVKLSDGRELCLDPKEPWVQRVVEKFLKRAES-
sw:IL8_BOVIN/1-97   MTSKLAVALLAAFLLSAALCEAAVLSRMSTELRCQCIKTHSTPFHPKFIKELRVIESGPHCENSEIIVKLTNGNEVCLNPKEKWVQKVVQVFVKRAE--
sw:GRO_RAT/28-92    ----------------------------ANELRCQCLQTVAG-IHFKNIQSLKVMPPGPHCTQTEVIATLKNGREACLDPEAPMVQKIVQKMLK-----
sw:AMC2_PIG/48-110  -----------------------------RELRCMCLTTTPG-IHPKMISDLQVIPAGPQCSKAEVIATLKNGKEVCLDPKAPLIKKIVQKML------
sw:IL8_FELCA/1-97   MTSKLVVALLAAFMLSAALCEAAVLSRISSELRCQCIKTHSTPFNPKLIKELTVIDSGPHCENSEIIVKLVNGKEVCLDPKQKWVQKVVEIFLKKAE--
sw:IL8_PIG/1-97     MTSKLAVAFLAVFLLSAALCEAAVLARVSAELRCQCINTHSTPFHPKFIKELRVIESGPHCENSEIIVKLVNGKEVCLDPKEKWVQKVVQIFLKRTE--
sw:IL8_RABIT/1-97   MNSKLAVALLATFLLSLTLCEAAVLTRIGTELRCQCIKTHSTPFHPKFIKELRVIESGPHCANSEIIVKLVDGRELCLDPKEKWVQKVVQIFLKRAE--
sw:IL8_HUMAN/1-99   MTSKLAVALLAAFLISAALCEGAVLPRSAKELRCQCIKTYSKPFHPKFIKELRVIESGPHCANTEIIVKLSDGRELCLDPKENWVQRVVEKFLKRAENS
sw:IL8_CAVPO/20-98  -------------------CEGMVVTKLVSELRCQCIKIHTTPFHPKFIKELKVIESGPRCANSEIIVKLSDNRQLCLDPKKKWVQDVVSMFLKRTES-
sw:MIP2_RAT/31-98   ----------------------------ASELRCQCLTTLPR-VDFKNIQSLTVTPPGPHCAQTEVIATLKDGHEVCLNPEAPLVQRIVQKILNKGK--
sw:GRO_CAVPO/34-99  ---------------------------AASELRCRCLRPVRG-LHPKNIQSVAVTAPGPHCHQTEVLATLKDGREACLDPEAPMVQKVLQRMLK-----
sw:IL8_HORSE/1-97   MTSKLAVALLAVFLLSAALCEAAVVSRITAELRCQCIKTHSKPFNPKLIKEMRVIESGPHCENSEIIVKLVNGAEVCLNPHTKWVQIIVQAFLKRTE--
sw:IL8_SHEEP/1-97   MTSKLAVALLAAFLLSAALCEAAVLSRMSTELRCQCIKTHSTPFHPKFIKELRVIESGPHCENSEIIVKLTNGKEVCLDPKEKWVQKVVQAFLKRAE--
sw:IL8_MACMU/1-98   MTSKLAVALLAAFLLSAALCEGAVLPRSAKELRCECIKTYSKPFHPKFIKELRVIESGPHCANTEIIVKLSDGRELCLDPKEPWVQRVVEKFVKRAEN-
sw:GRO_MOUSE/28-92  ----------------------------ANELRCQCLQTMAG-IHLKNIQSLKVLPSGPHCTQTEVIATLKNGREACLDPEAPLVQKIVQKMLK-----
sw:GRO_HUMAN/38-101 ----------------------------ATELRCQCLQTLQG-IHPKNIQSVNVKSPGPHCAQTEVIATLKNGRKACLNPASPIVKKIIEKML------

Quality/1-99
```



Homology, in biology, similarity of the structure, physiology, or development of different species of organisms based upon their descent from a common evolutionary ancestor. ...

www.britannica.com › science › homology-evolution

Sequence alignment

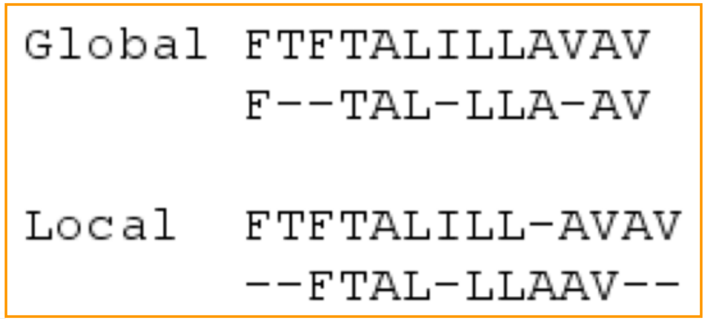


```
AAB24882        TYHMCQFHCRYVNNHSGEKLYECNERSKAFSCPSHLQCHKRRQIGEKTHEHNQCGKAFPT 60
AAB24881        --------------------YECNQCGKAFAQHSSLKCHYRTHIGEKPYECNQCGKAFSK 40
                                    ****: .***:  * *:** * :****.:* *******..

AAB24882        PSHLQYHERTHTGEKPYECHQCGQAFKKCSLLQRHKRTHTGEKPYE-CNQCGKAFAQ- 116
AAB24881        HSHLQCHKRTHTGEKPYECNQCGKAFSQHGLLQRHKRTHTGEKPYMNVINMVKPLHNS 98
                 **** *:***********:***:**.: .***************    :  *.: :
```

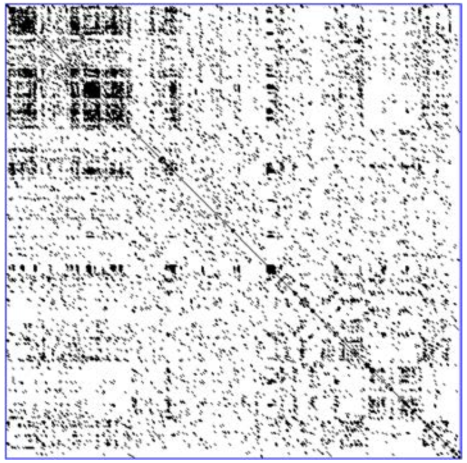
* - identical

: - conserved substitutions (same colour group)

. - semi-conserved substitution (similar shapes).

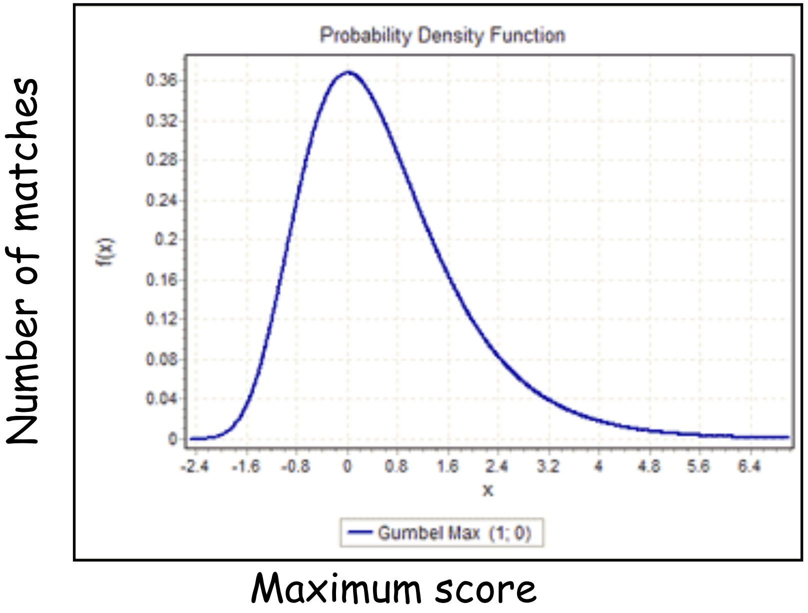


```
Global  FTFTALILLAVAV
        F--TAL-LLA-AV

Local   FTFTALILL-AVAV
        --FTAL-LLAAV--
```



The significance of similarity scores

Extreme-value distribution of <u>random sequence</u> alignment scores



E-value (expected value): the expected number of *<u>random hits</u>* with the same score expected by chance in *<u>this</u>* database.

Seq. similarity identification tools:

- SSAHA
- Blat
- BLAST
- Smith-Waterman (Paralign)
- PSI-Blast / RPS-Blast
- CS-Blast
- HHMER
- HHsearch

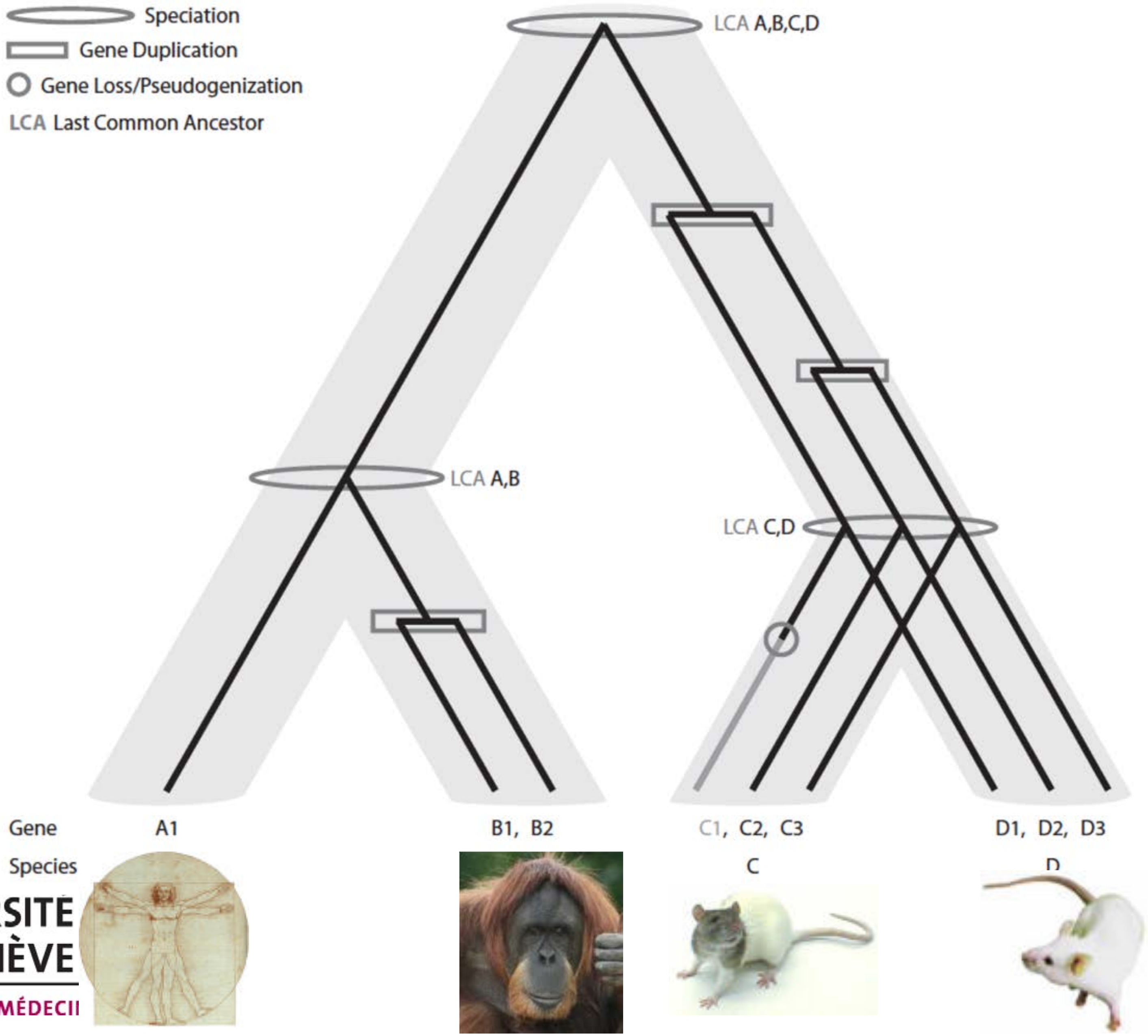
Orthology definition

Originally the term was introduced in 1970 by Walter Fitch

Two homologous genes in two different species that derive from a single gene in the last common ancestor of the species

and better rephrased by Koonin in 2005:
Genes originating from a single ancestral gene in the last common ancestor of the compared genomes.

Orthologs


Speciation
Gene Duplication
○ Gene Loss/Pseudogenization
LCA Last Common Ancestor
LCA A,B,C,D
LCA A,B
LCA C,D
Gene
A1
B1, B2
C1, C2, C3
D1, D2, D3
Species
C
D

Please note

- Similarity

  could be

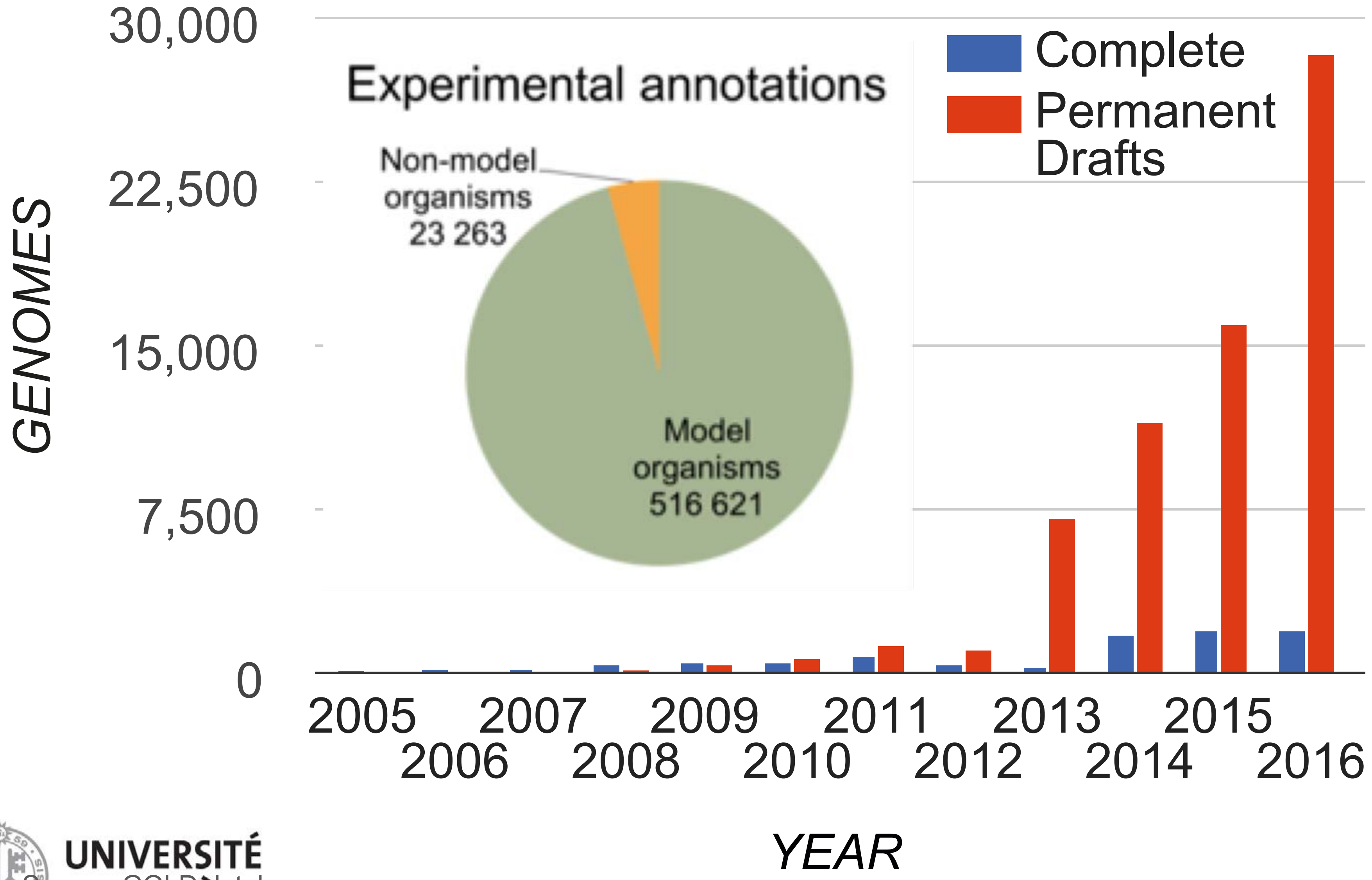
- Homology

  could be

- Orthology

i.e. all orthologs are homologs and look similar;
not all similar looking sequences are homologs,
and not all homologs are orthologs;
and there is no 'function' in these definitions.

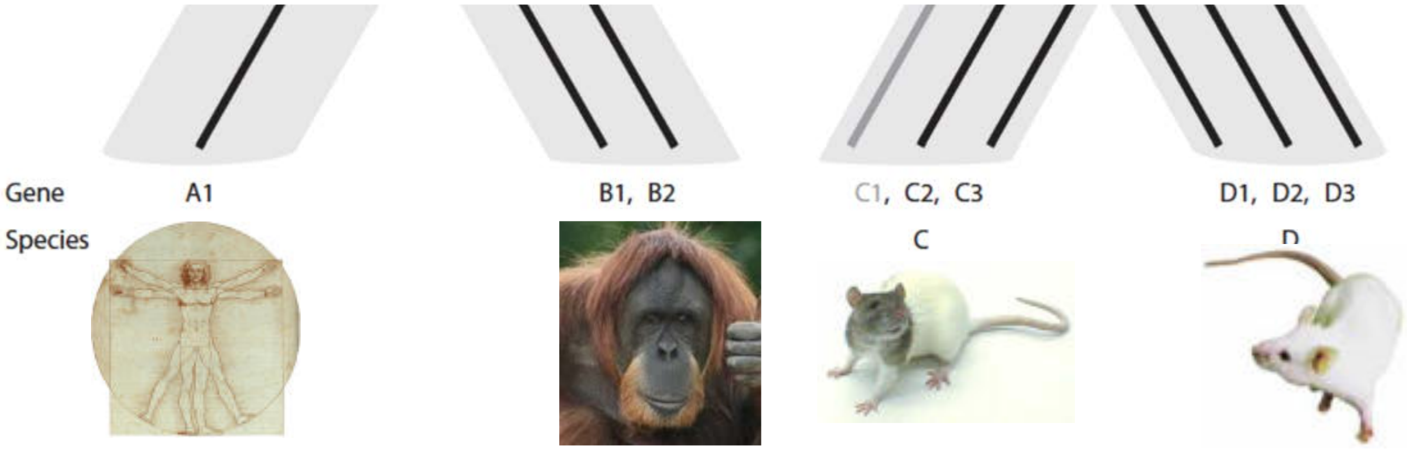
The growth of the need: linking genomics data to gene function knowledge



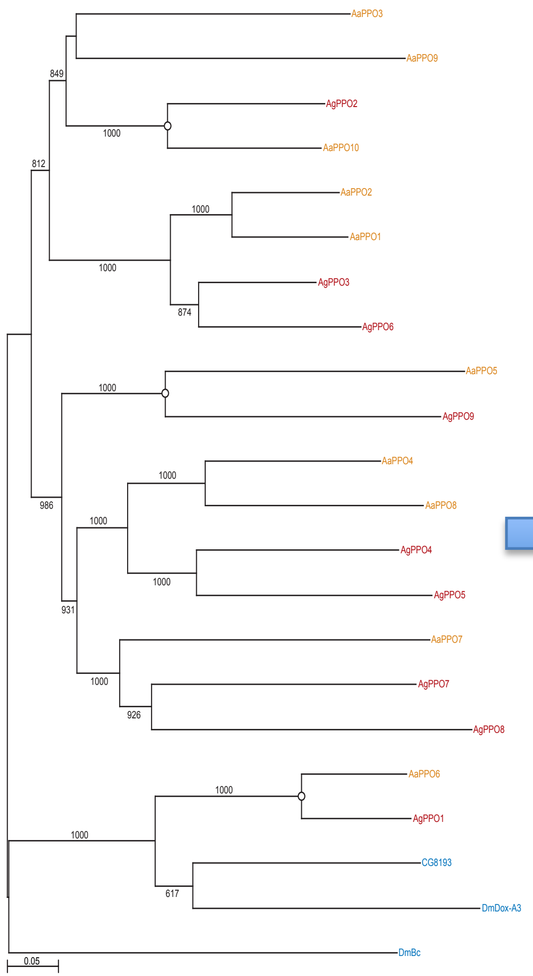
Source: GOLD database

How to identify orthologs

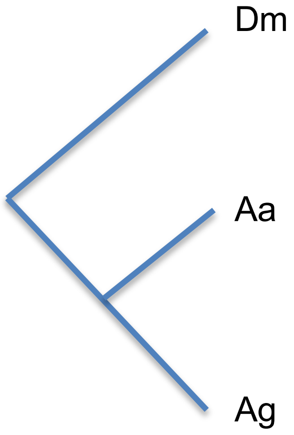
?



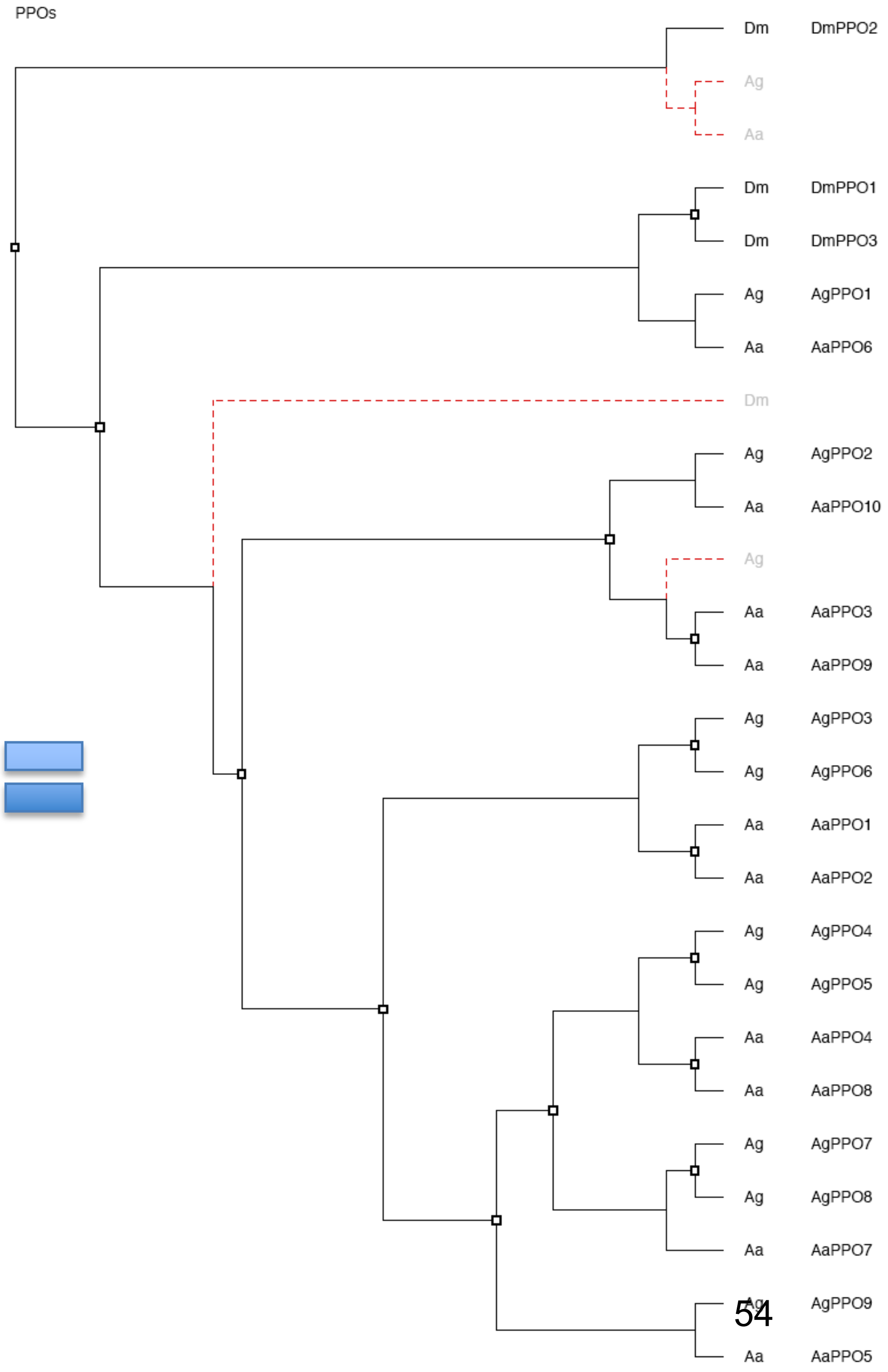
1. Tree-reconciliation (complicated)



+



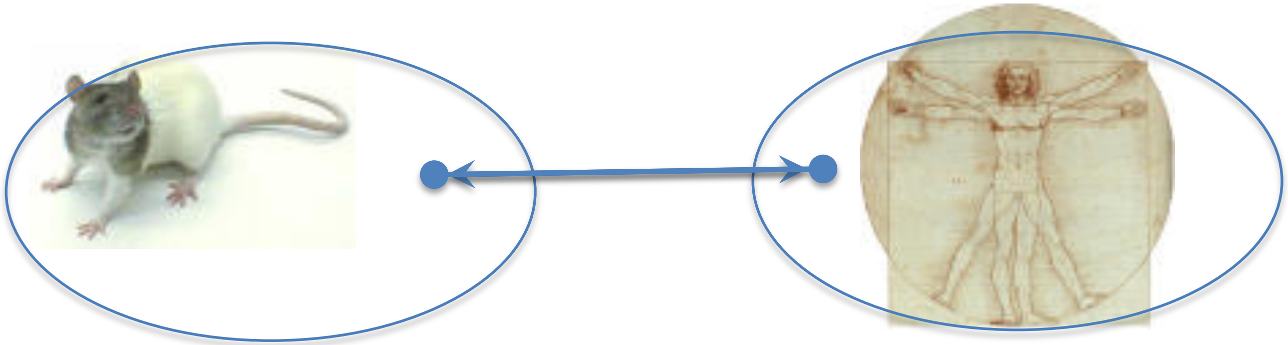
=



UNIVERSITÉ
DE GENÈVE
FACULTÉ DE MÉDECINE

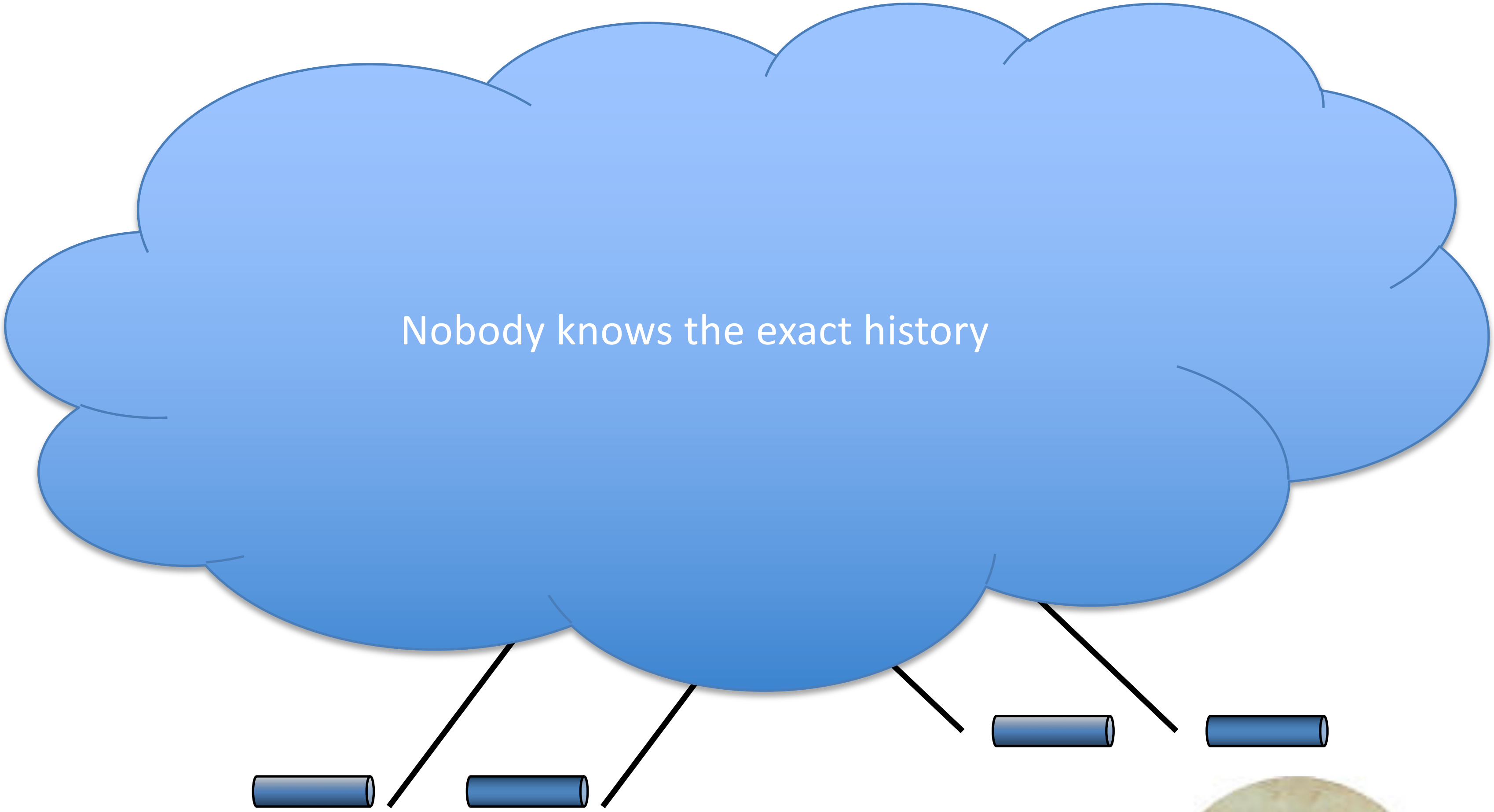
2. Best-Reciprocal-Hits (BRH)

- *what are these?*



- *why do they work?*

#1 Why BRH is indicative of orthology?



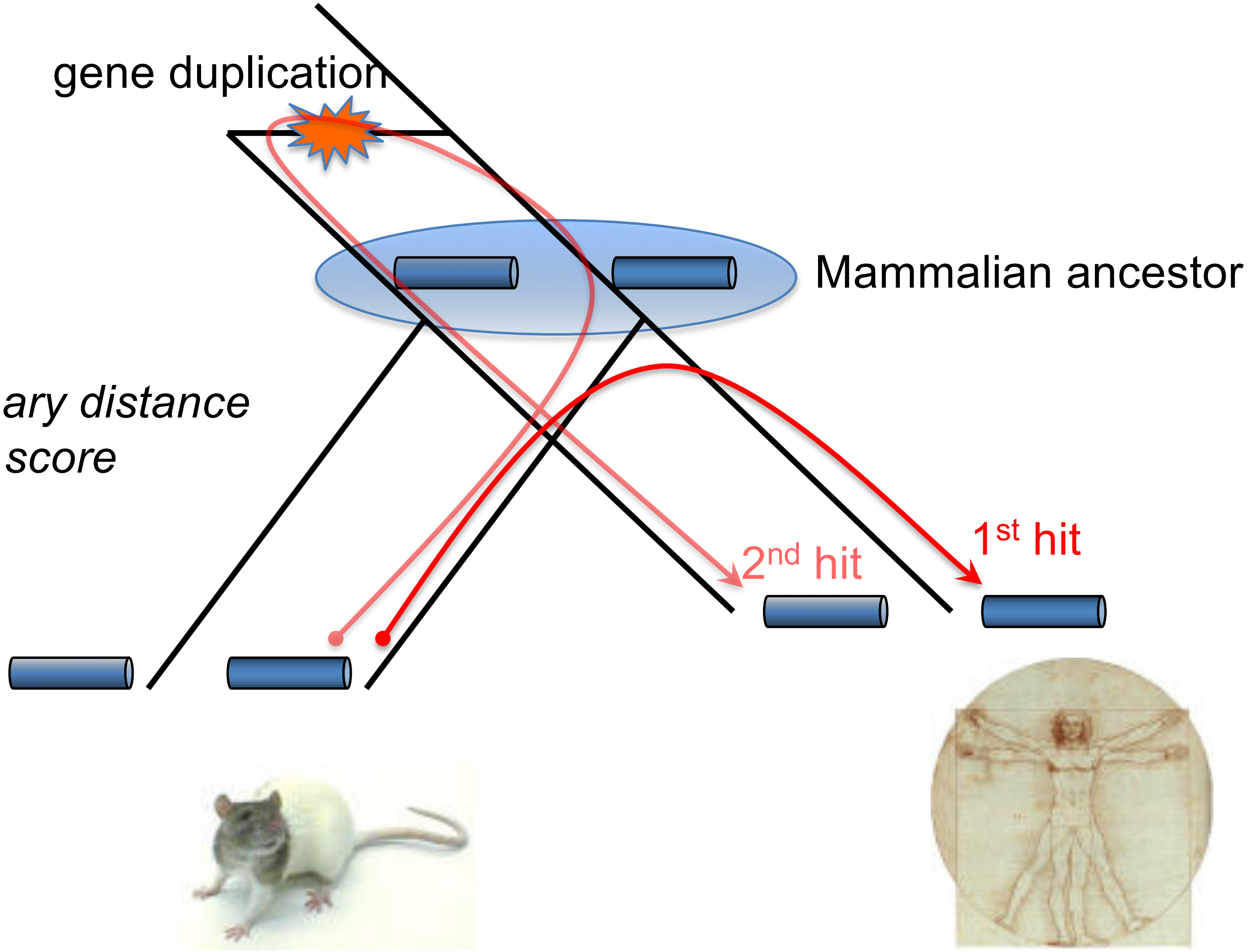
rat genes

human genes

UNIVERSITÉ DE GENÈVE

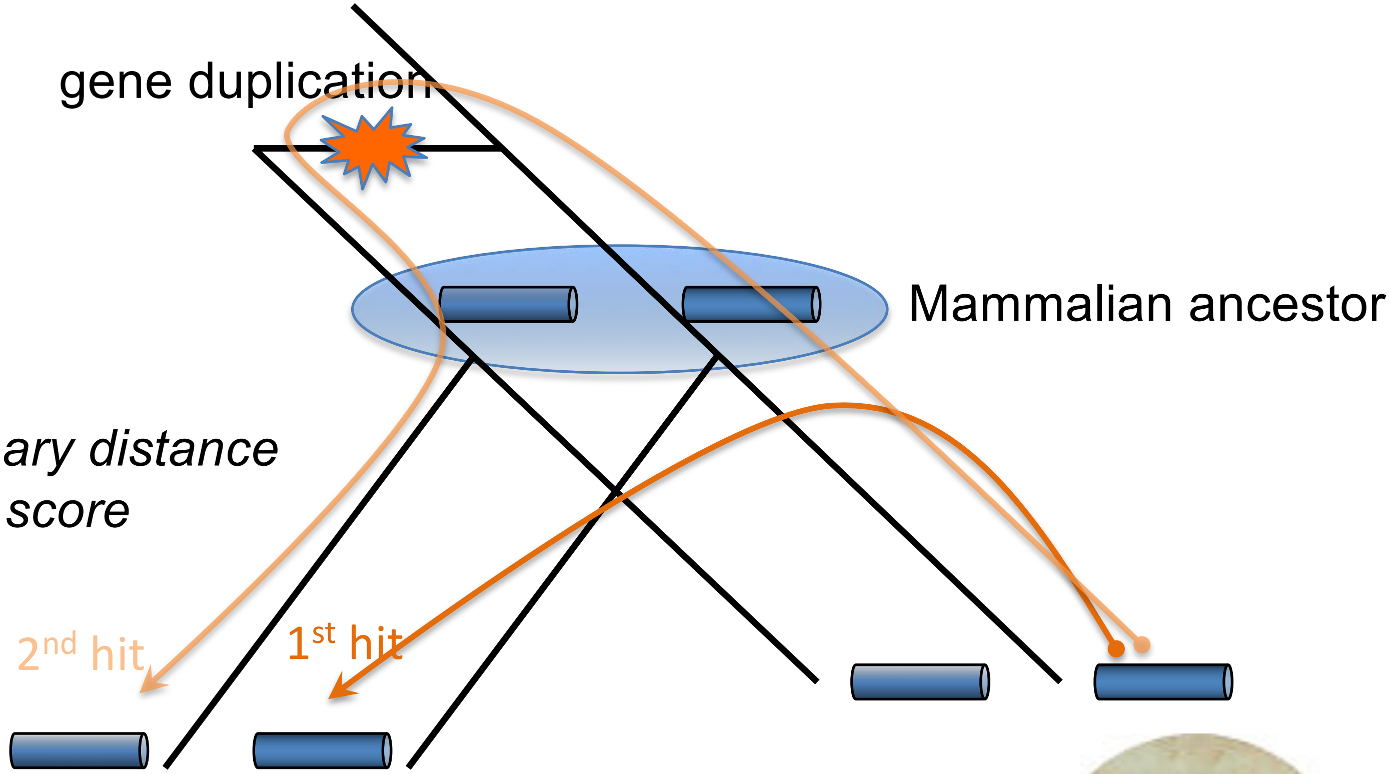
FACULTÉ DE MÉDECINE

#2 lets BLAST
a rat gene to all human genes



the longer evolutionary distance
the worse similarity score

#3 lets BLAST in reverse
the best human gene hit to all rat genes



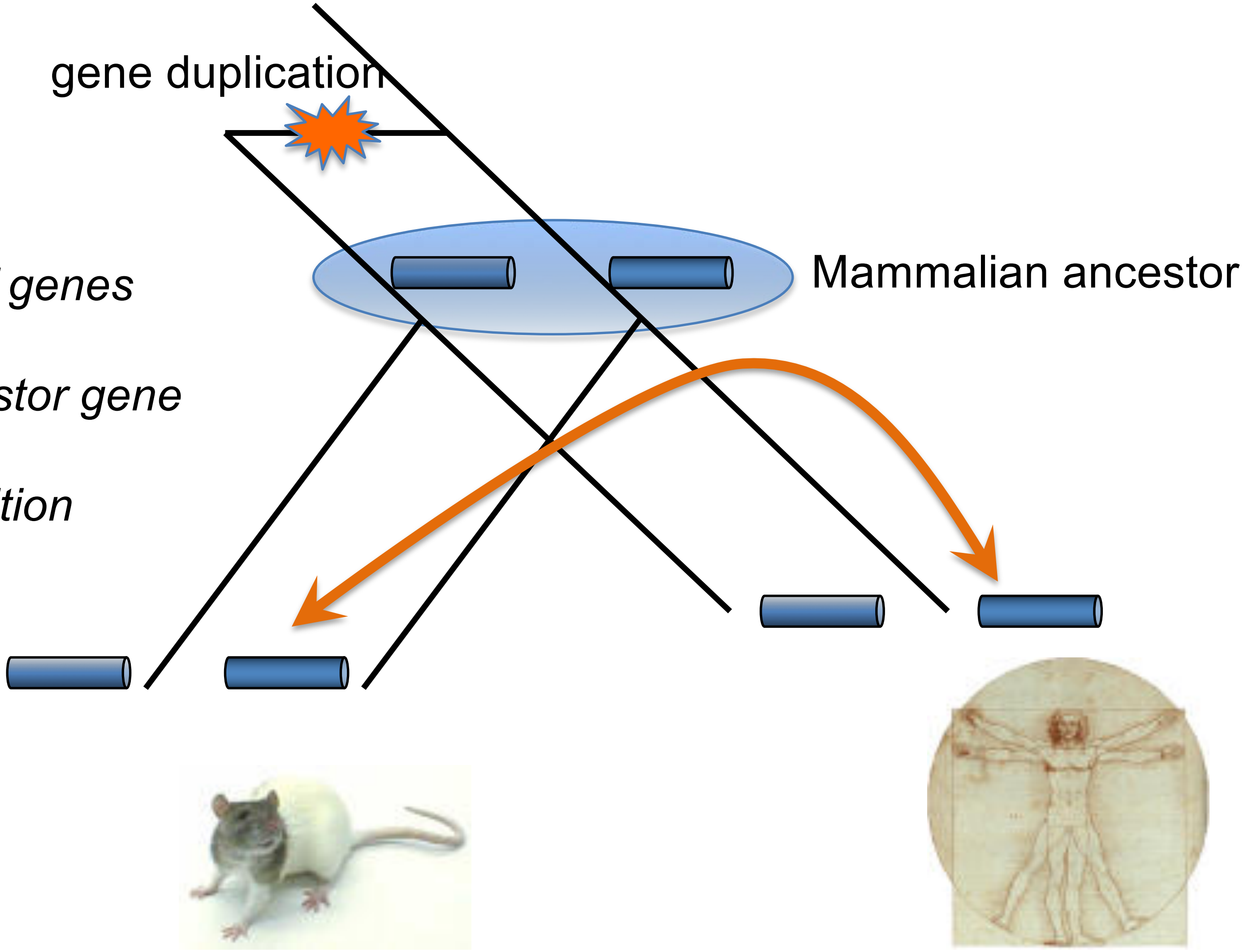
the longer evolutionary distance
the worse similarity score

UNIVERSITÉ
DE GENÈVE
FACULTÉ DE MÉDECINE

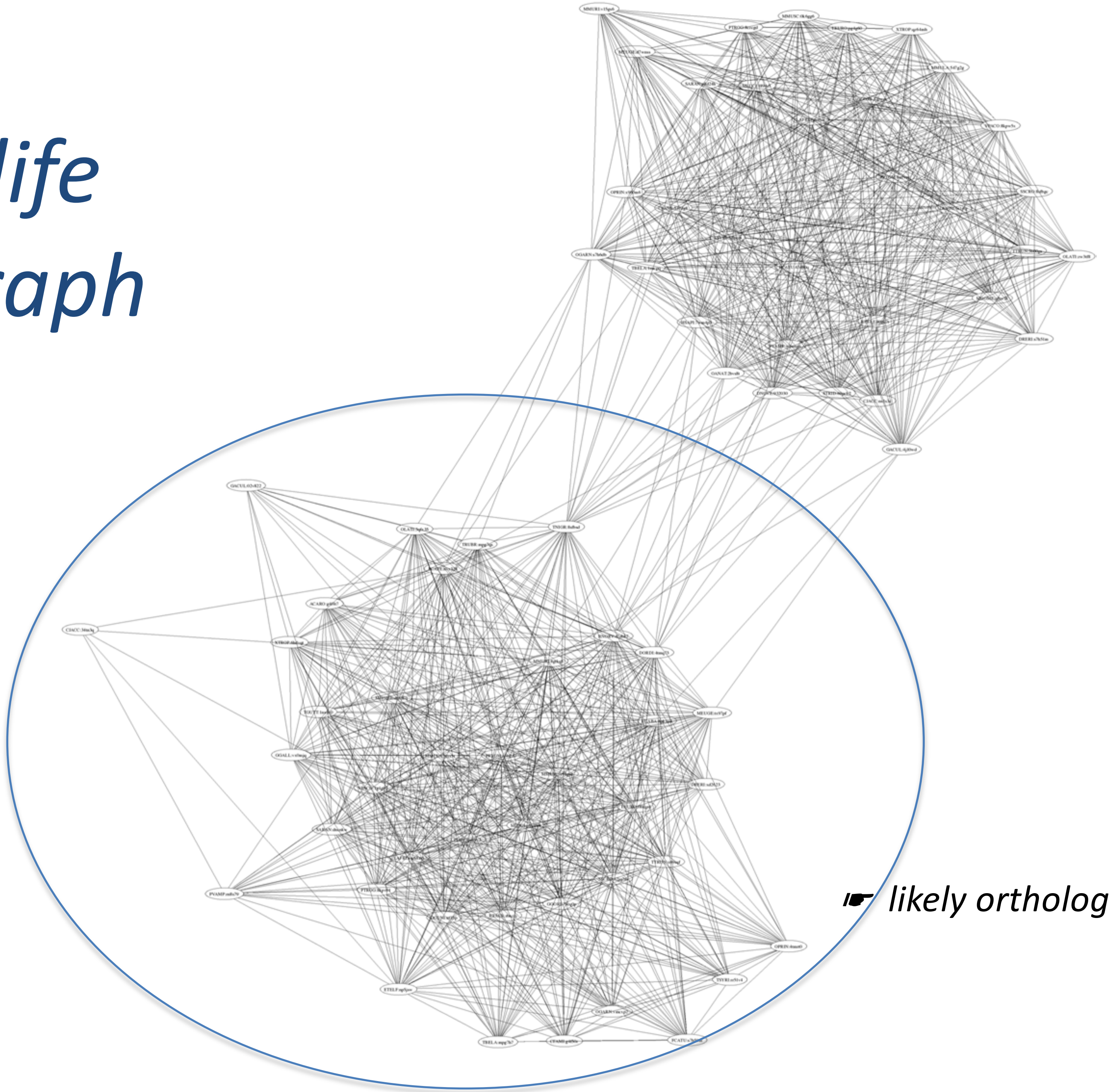
#4 there is a reciprocally best matching pair of genes between rat and human

- *BRH joins a pair of genes via a single last-common-ancestor gene*

☛ *orthologs by definition*



a real-life BRH graph

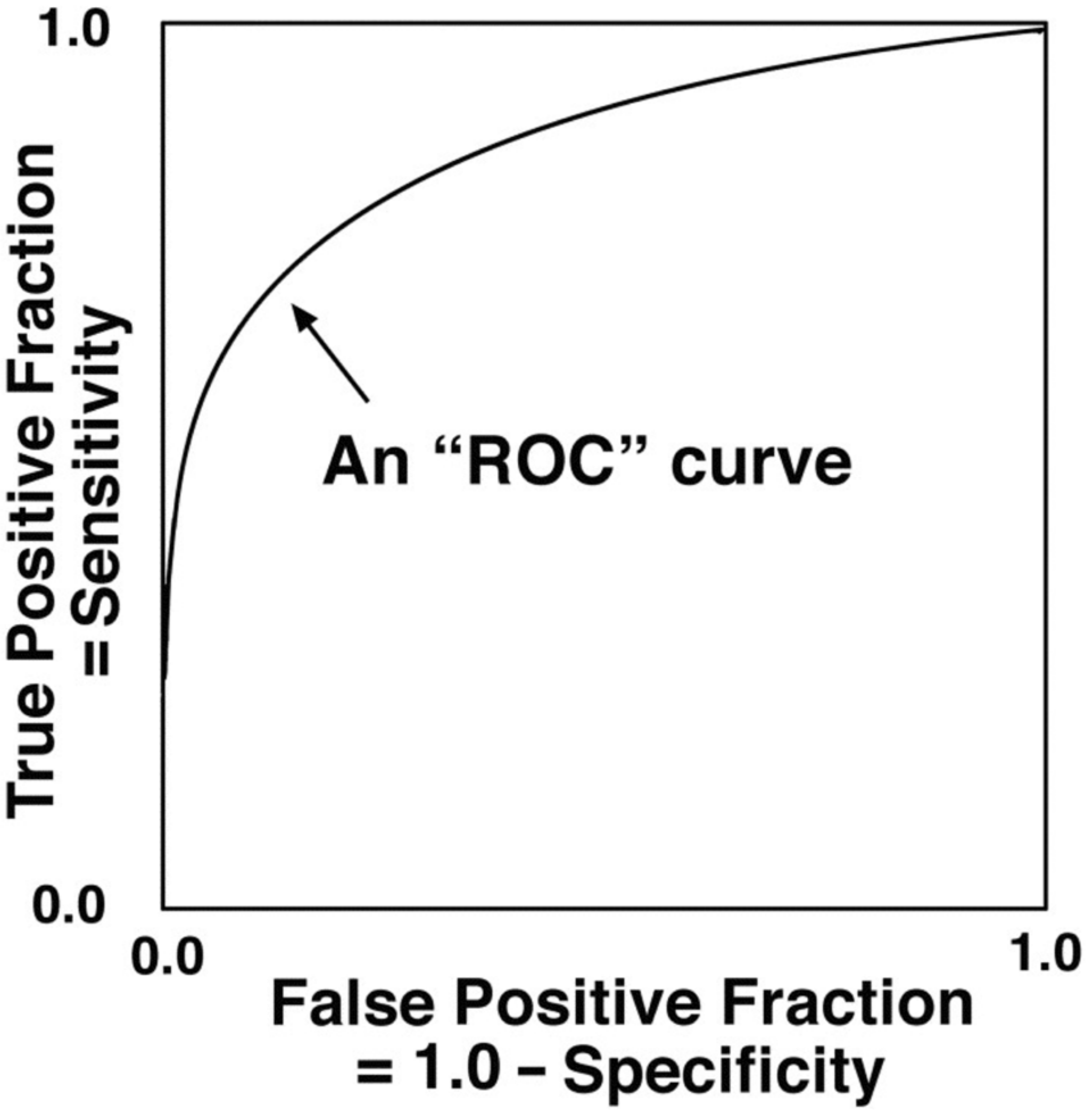


☛ *likely orthologs*

Orthology identification: DIY is more error-prone

- COG/KOG
- InParanoid
- eggNOG
- OrthoMCL
- OrthoFinder
- OrthoDB (Orthologer)

Sensitivity vs. Specificity tradeoff



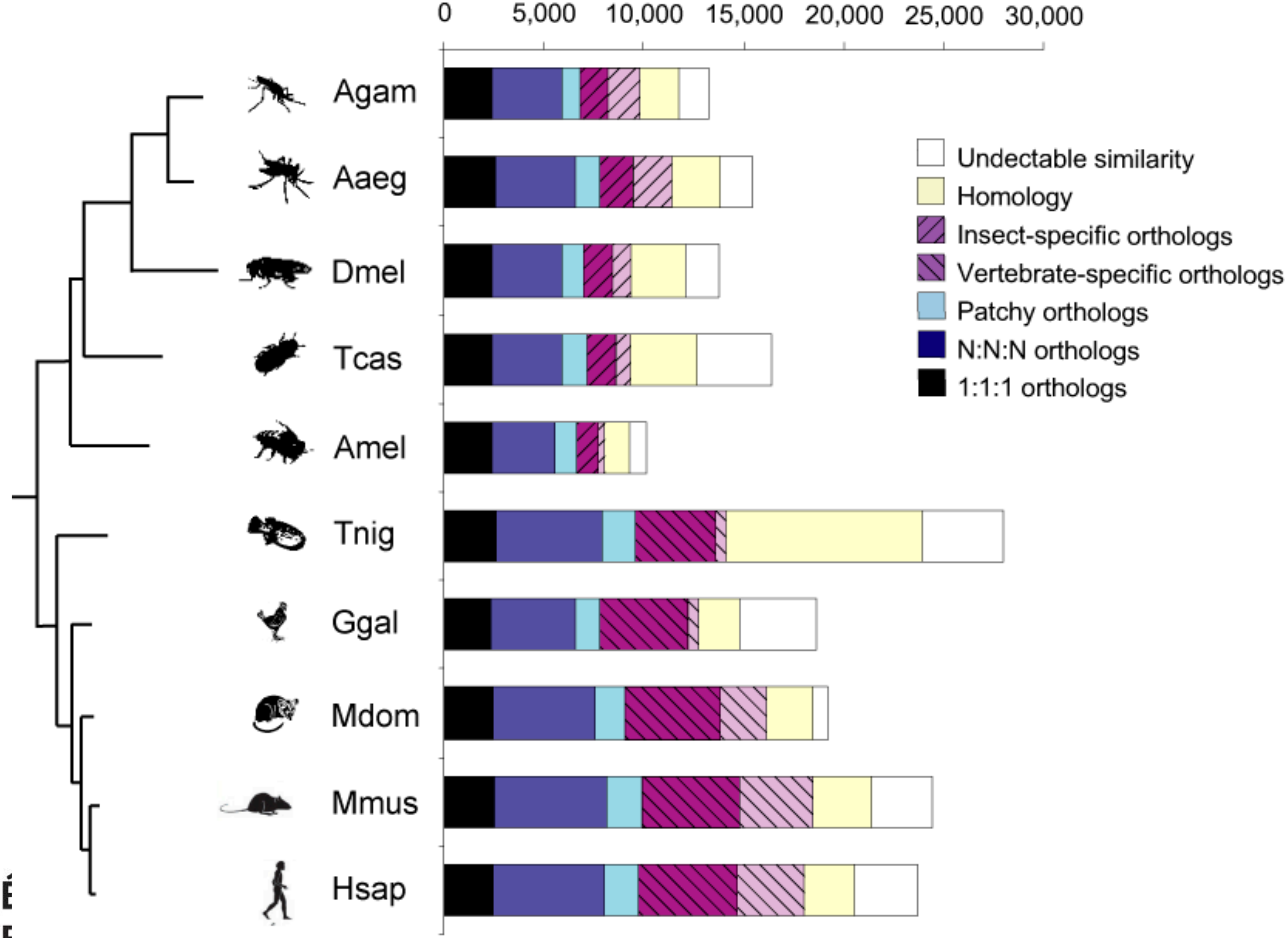
To keep in mind

- Sensitivity *vs.* Specificity tradeoff
- Cost ($ or time) *vs* Accuracy tradeoff

☛ find or do benchmarking

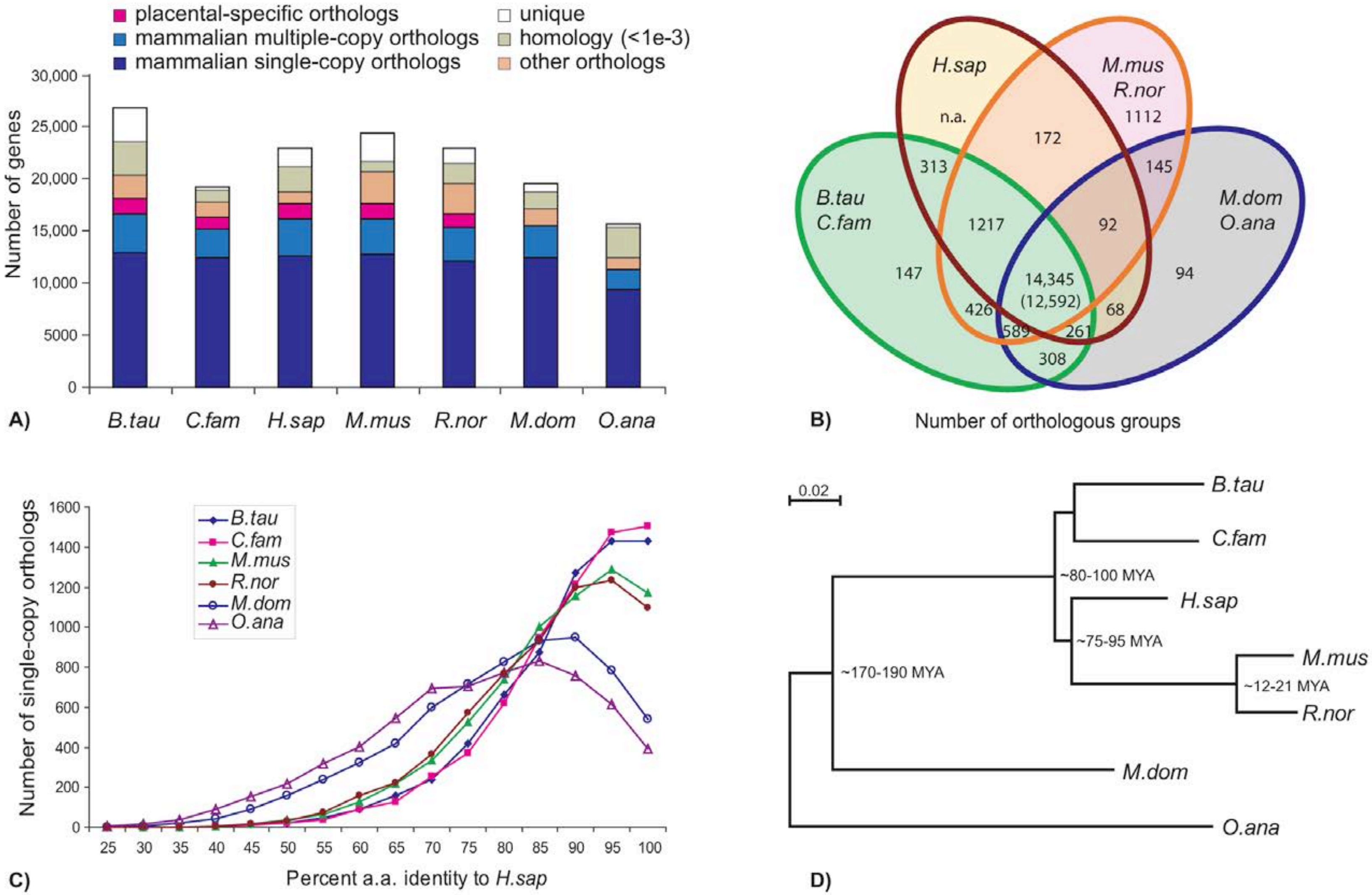
Examples of comparative study

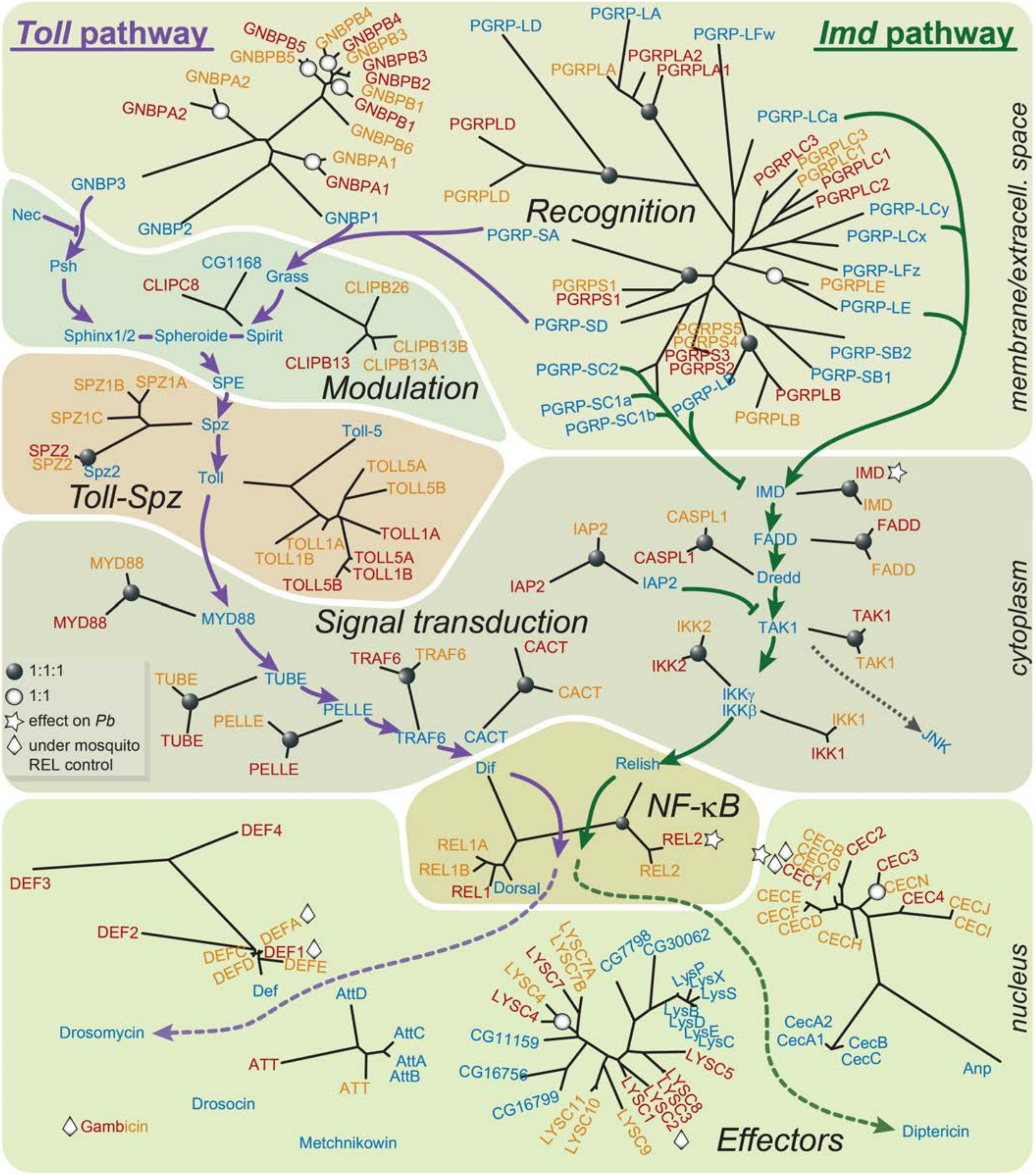
Counting genes



UNIVERSITÉ DE GENÈVE
FACULTÉ DE MÉDECINE

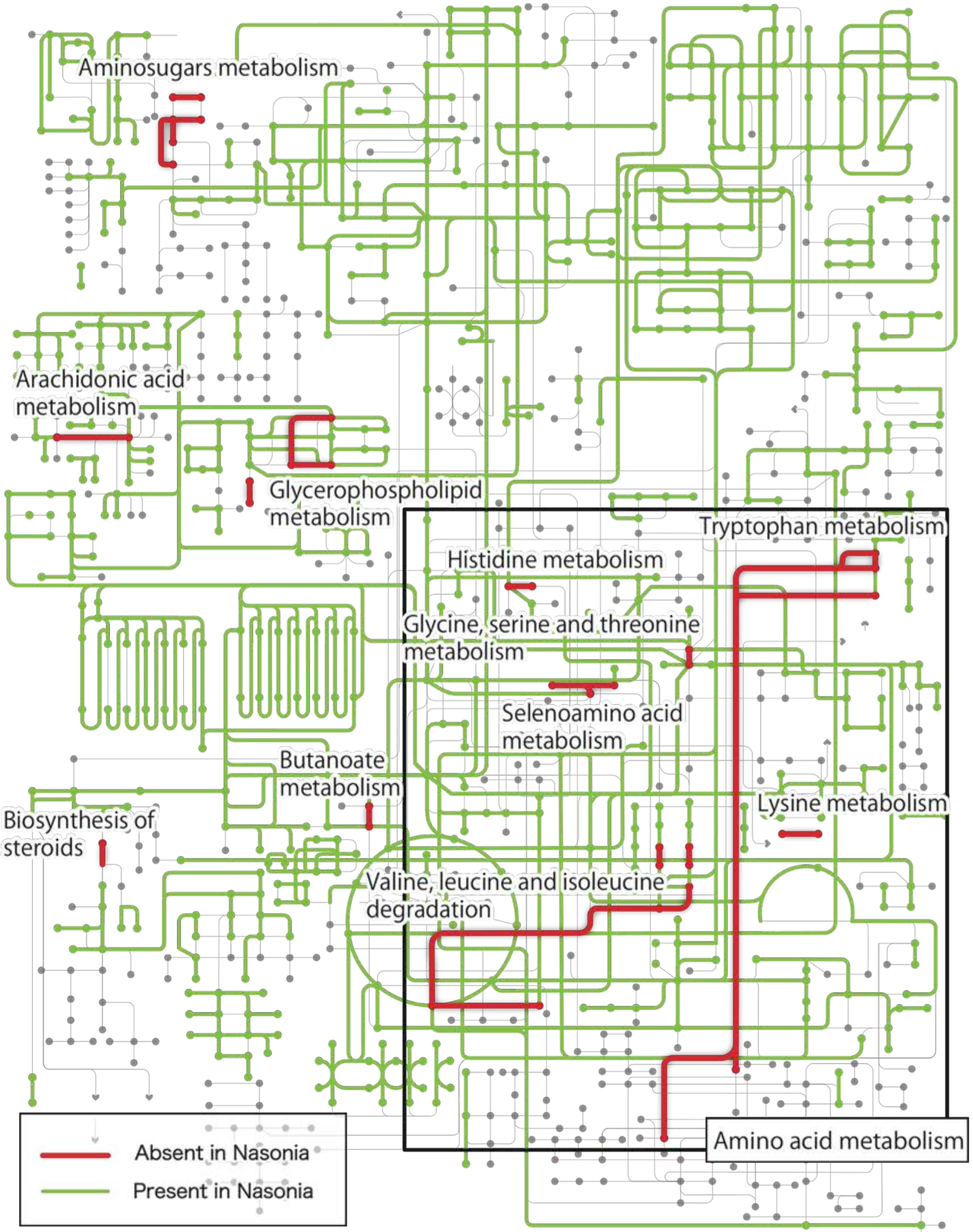
Cow is molecularly closer to human than mouse





Pathway perspective

Nasonia problem with amino acid metabolism



UNIVERSITÉ DE GENÈVE
FACULTÉ DE MÉDECINE