BIG DATA *BIOINFORMATICS*

or why we all need computers

Wojciech Makałowski
Institute of Bioinformatics
University of Münster, Germany

GRAID - Global Research Alliance for Infectious Disease

GRAID (Global Research Alliance for Infectious Disease) is a platform for collaborative research among scientists around the wo
the fight against infectious disease. According to WHO, infectious diseases are the major cause of death in low-income countri
Combining the resources of the countries worldwide, we have established this platform of scientists and health practitioners fr
the early stage, we are conducting researches primarily in the genetics of infectious disease using state-of-the-art sequencing.
implement sequencing techniques in affected countries by means of portable sequencing developed by Oxford Nanopore Tech
as more bright minds join this platform, we will expand our scope of research and techniques beyond genetics and sequencing
containing and conquering the threat of the humanity.

2018-12-14 15:33

Workshops

MinION Hokkaido, Juli 2019

MinION Manado 2018

MinION Bangkok 2017

MinION Kashiwa 2016

Symposium and Workshop using MinION, July 7-10 2019

Download / Display Flyer

You need a registration for the following 3 days

July 7, 2019

13:00	RNA-extraction
15:00	PCR preparation
16:00	PCR start

July 8, 2019

10:00	Orientation by organizer
10:30	Gel electrophoresis (PCR check)
11:00	What is "Diagnosis-by-Sequencing", introduction of the concept
11:30	gel check
12:00	Lunch
14:45	MinION library prep and Sequencing run
18:00	Evening Seminar by Sysmex
18:30	Welcome party

July 9, 2019

08:45	Introduction to bioinformatic analyses
12:00	Lunch
13:00	Hands-on training (linux, guppy, debarcoding, deindexing, NanoPipe)
18:00	end of the day 3

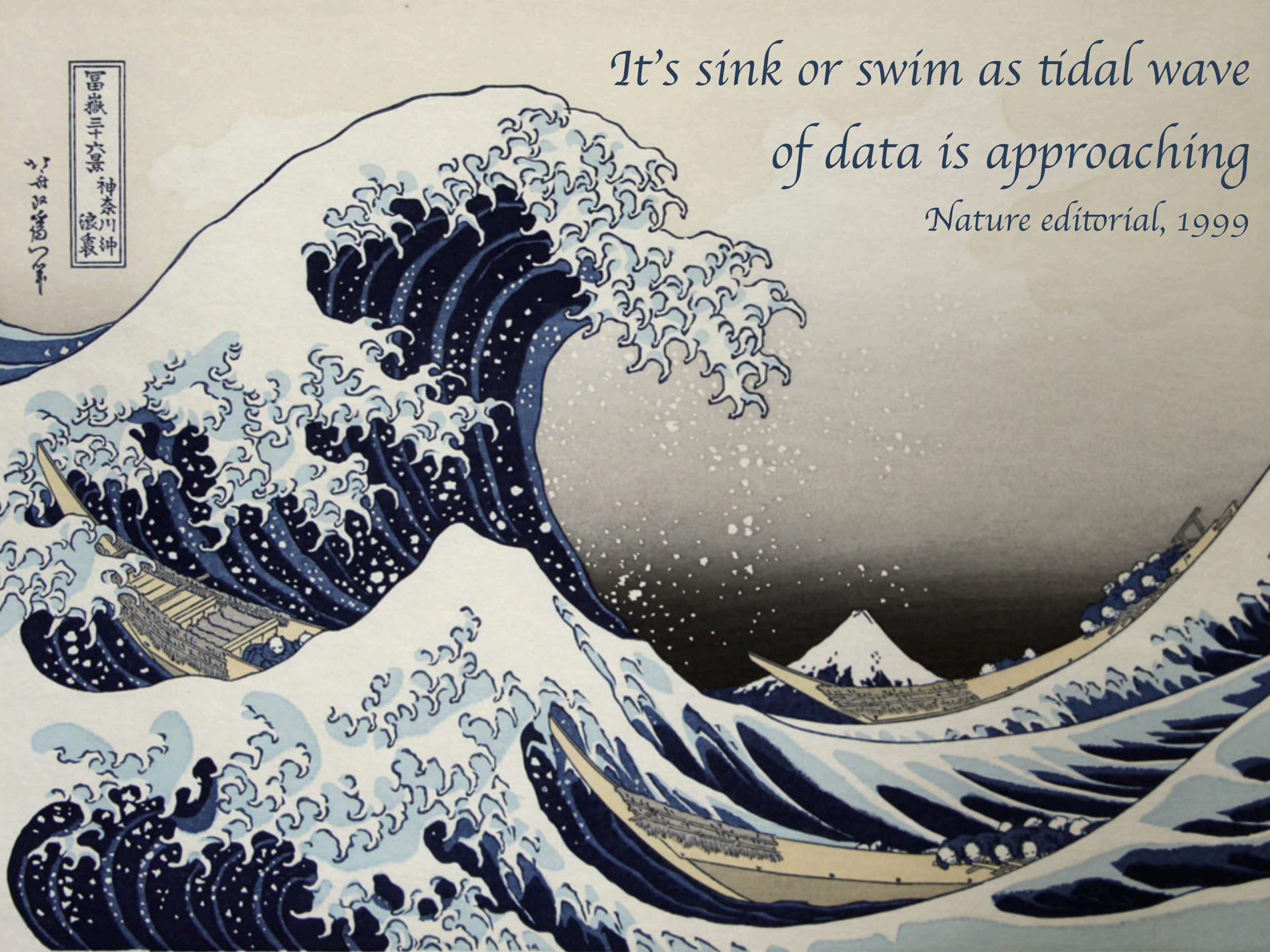
Symposium for Diagnosis-by-Sequencing using MinION (public event; no registration required)

July 10, 2019

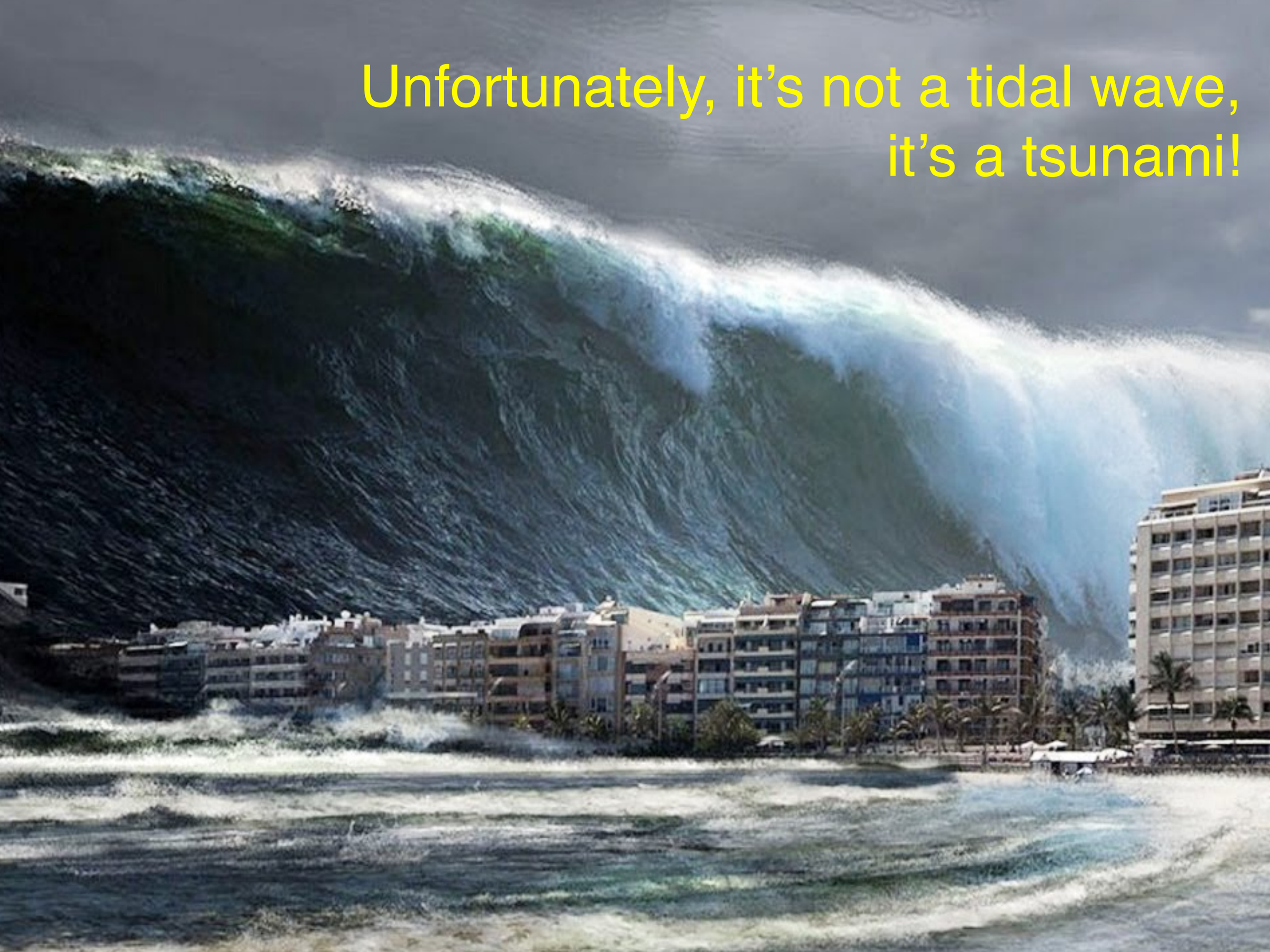
10:00	Opening remarks
10:30	Keynote lecture (Prof. Yutaka Suzuki, University of Tokyo, Japan)

It's sink or swim as tidal wave of data is approaching

Nature editorial, 1999



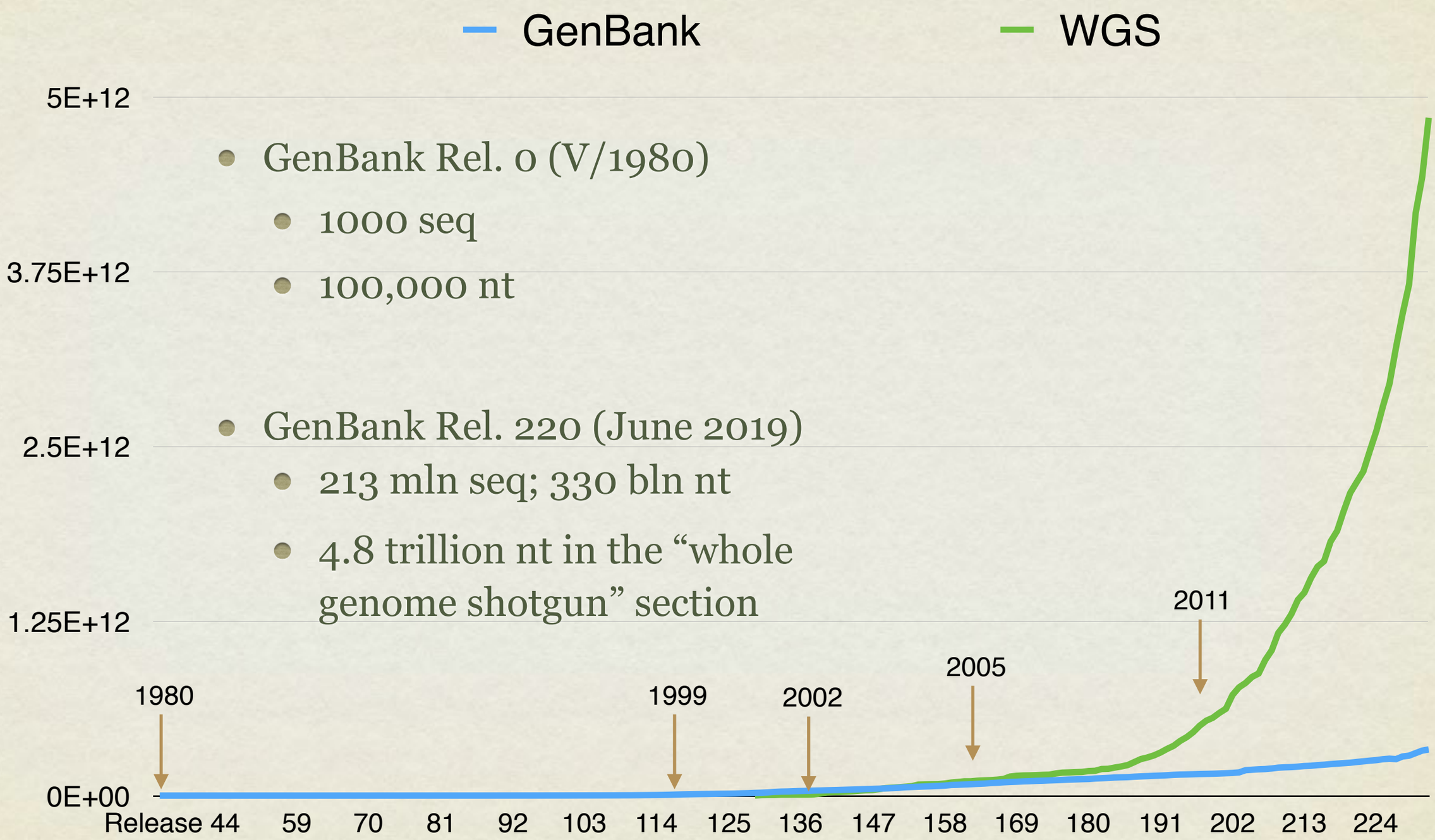
Unfortunately, it's not a tidal wave,
it's a tsunami!



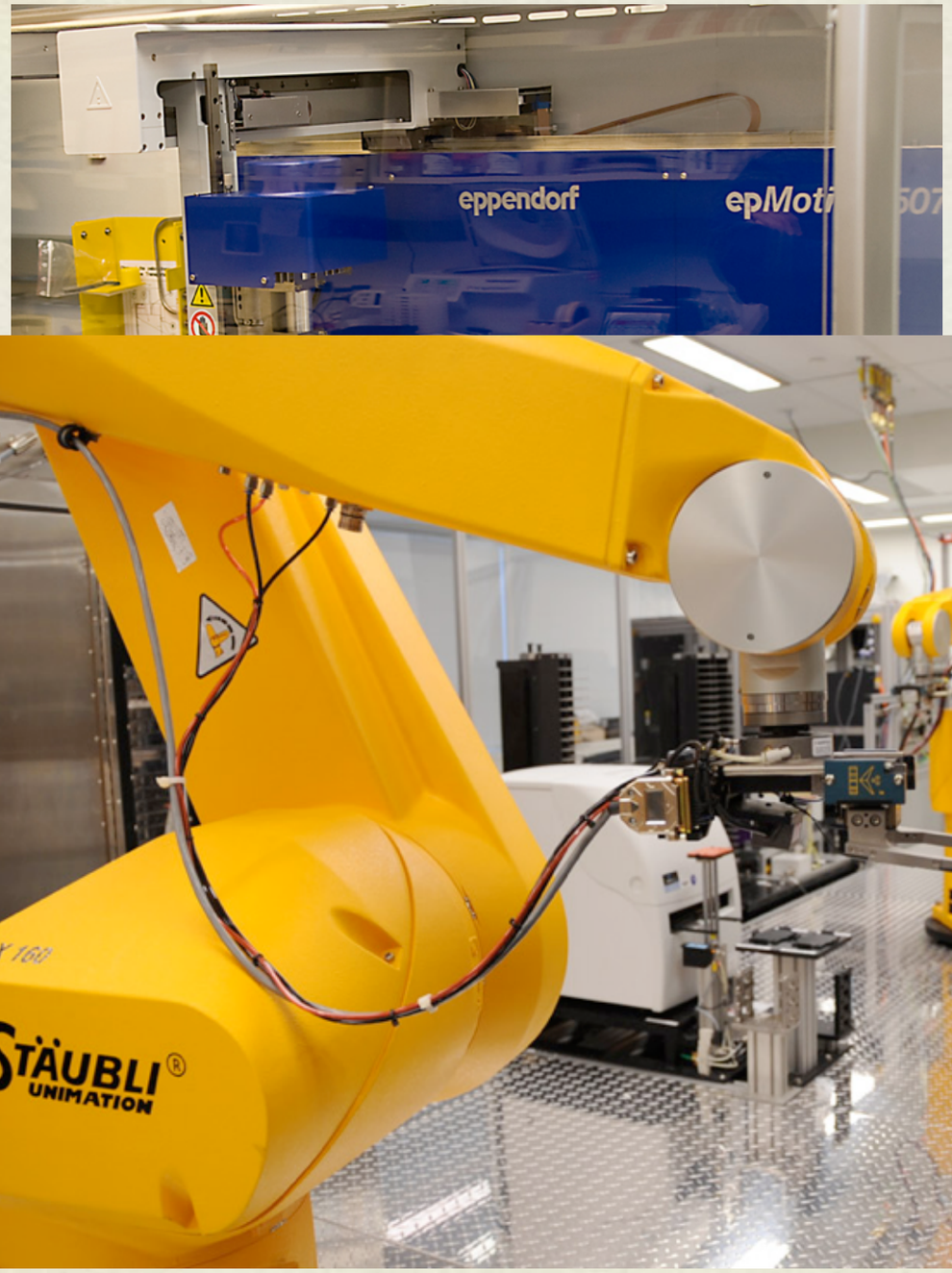
GROWTH OF BIOMEDICAL INFORMATION - GENBANK

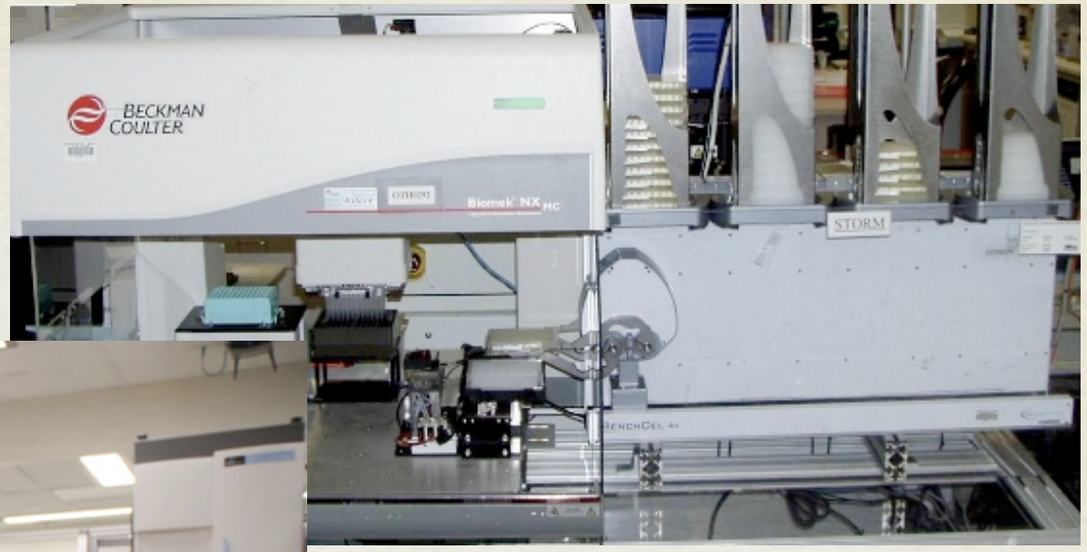
- GenBank Rel. 0 (V/1980)
  - 1000 seq
  - 100,000 nt

- GenBank Rel. 220 (June 2019)
  - 213 mln seq; 330 bln nt
  - 4.8 trillion nt in the “whole genome shotgun” section

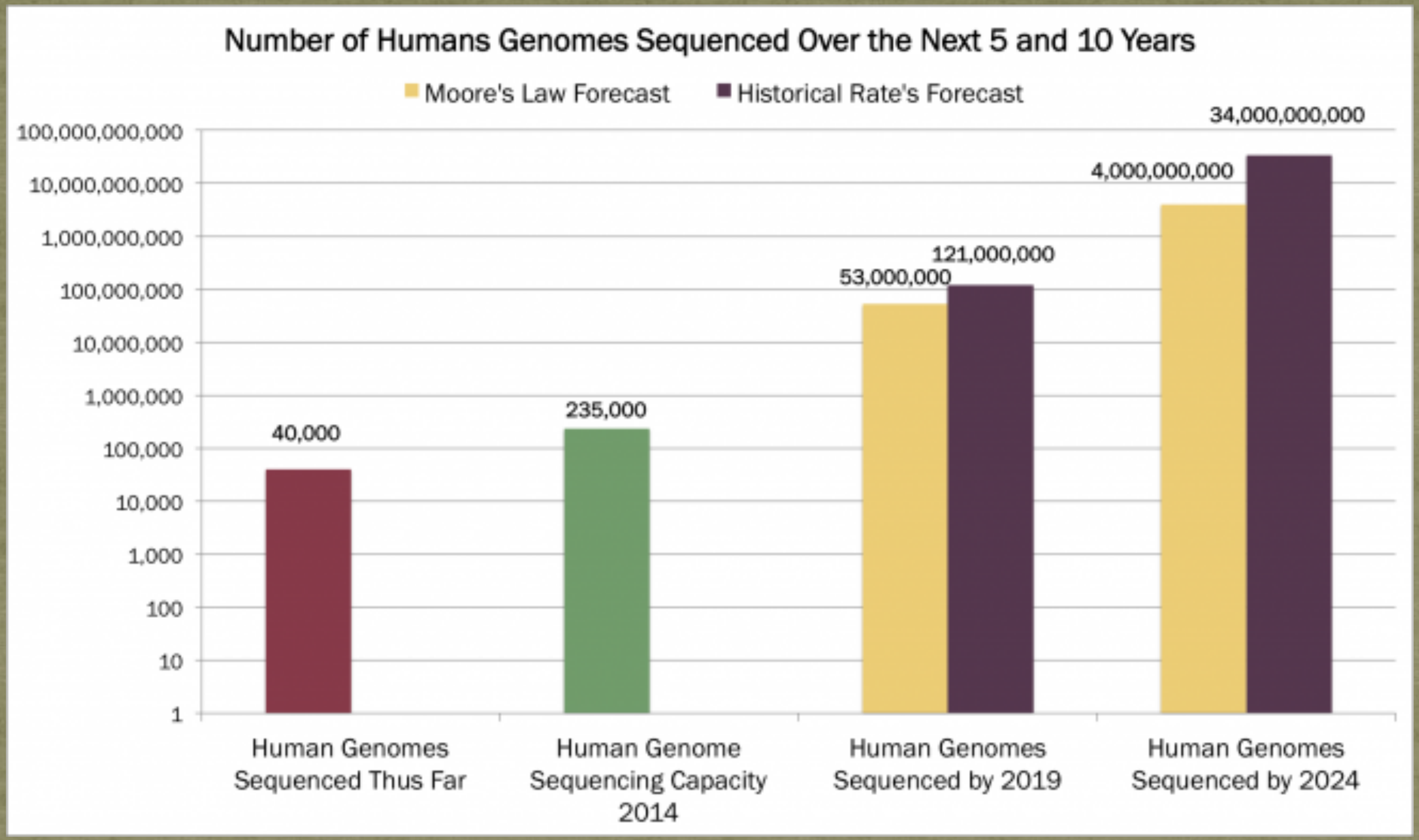


TECHNOLOGY MEETS BIOLOGY





IMPROVING TECHNOLOGY



http://ark-invest.com/genomic-revolution/declining-costs-of-genome-sequencing

CHALLENGE: HOW FROM THIS...

TGCATCGATCGTAGCTAGCTAGCGCATGCTAGCTAGCTAGCTAGCTACGATGCATCG
TGCATCGATCGATGCATGCTAGCTAGCTAGCTAGCATGCTAGCTAGCTAGCTATTGG
CGCTAGCTAGCATGCATGCATGCATCGATGCATCGATTATAAGCGCGATGACGTCAG
CGCGCGCATTATGCCGCGGCATGCTGCGCACACACAGTACTATAGCATTAGTAAAAA
GGCCGCGTATATTTTACACGATAGTGCGGCGCGGCGCGTAGCTAGTGCTAGCTAGTC
TCCGGTTACACAGGTAGCTAGCTAGCTGCTAGCTAGCTGCTGCATGCATGCATTAGT
AGCTAGTGTAGCTAGCTAGCATGCTGCTAGCATGCAGCATGCATCGGGCGCGATGCT
GCTAGCGCTGCTAGCTAGCTAGCTAGCTAGGCGCTAATTATTTATTTTGGGGGGTTA
AAAAAAAAAATTTCGCTGCTTATACCCCCCCCCACATGATGATCGTTAGTAGCTACT
AGCTCTCATCGCGCGGGGGGATGCTTAGCGTGGTGTGTGTGTGTGGTGTGTGTGGTC
CTATAATTAGTGCATCGGCGCATCGATGGCTAGTCGATCGATCGATTTTATATATCT
AAAGACCCCATCTCTCTCTCTTTTCCCTTCTCTCGCTAGCGGGCGGTACGATTTACC
GGCCGCGTATATTTTACACGATAGTGCGGCGCGGCGCGTAGCTAGTGCTAGCTAGTC
AGCTCTCATCGCGCGGGGGGATGCTTAGCGTGGTGTGTGTGTGTGGTGTGTGTGGTC
TGCATCGATCGATGCATGCTAGCTAGCTAGCTAGCATGCTAGCTAGCTAGCTATTGG
CTATAATTAGTGCATCGGCGCATCGATGGCTAGTCGATCGATCGATTTTATATATCT
CGCTAGCTAGCATGCATGCATGCATCGATGCATCGATTATAAGCGCGATGACGTCAG
TCCGGTTACACAGGTAGCTAGCTAGCTGCTAGCTAGCTGCTGCATGCATGCATTAGT

Infer this

HOW TO SOLVE THE PROBLEM - A HUMAN OR A COMPUTER?

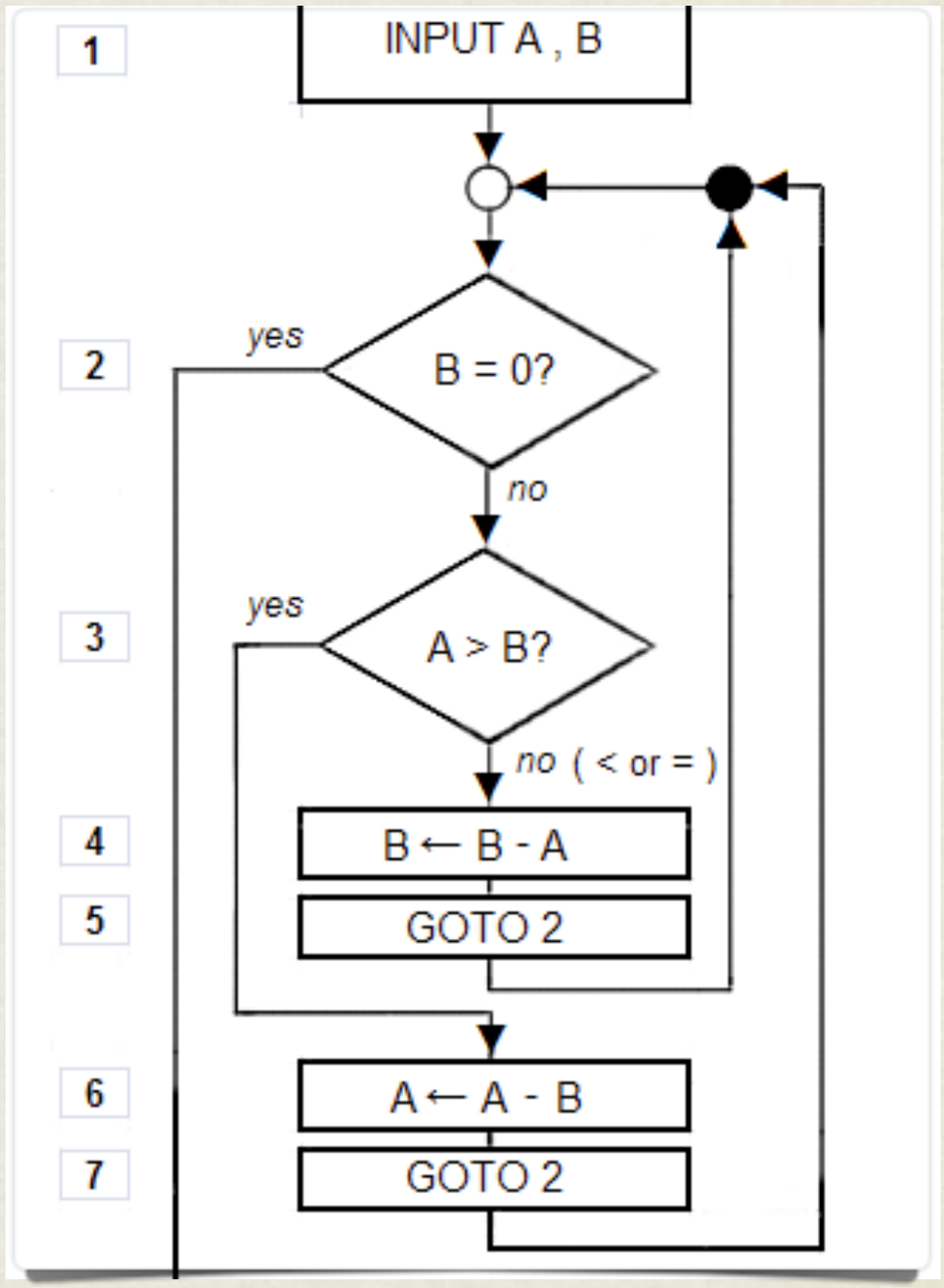
- very smart
- slow
- error prone
- doesn’t like repetitive tasks

- not so smart (stupid)
- extremely fast
- very accurate
- doesn’t understand human languages;

needs instruction provided in a special way

ALGORITHM

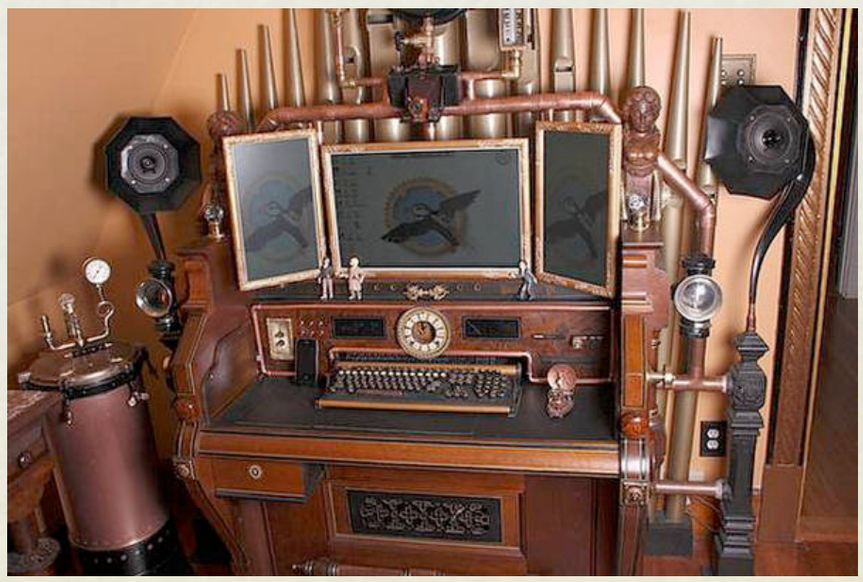
A step-by-step problem-solving procedure, especially an established, recursive computational procedure for solving a problem in a finite number of steps.



EXAMPLE TASK: PUT SHOES ON!

A human just understands an order and often executes it automatically even without thinking

A computer needs detailed instruction (an algorithm)



PUT SHOES ON!
INSTRUCTION FOR A COMPUTER

1. Find two the same shoes
2. Check if you have left and right shoe
3. Check if they are of the same size
4. Check if this is the right size
5. Put the left shoe on
6. Put the right shoe on
7. Tie the laces

THE ORIGIN OF THE FIELD

Paulien Hogeweg coined the term *bioinformatica* to define “the study of informatic processes in biotic systems”. Hesper B, Hogeweg P (1970) Bioinformatica: een werkconcept. Kameleon 1(6): 28–29. (In Dutch.) Leiden: Leidse Biologen Club.

… but its origin can be tracked back many decades earlier.

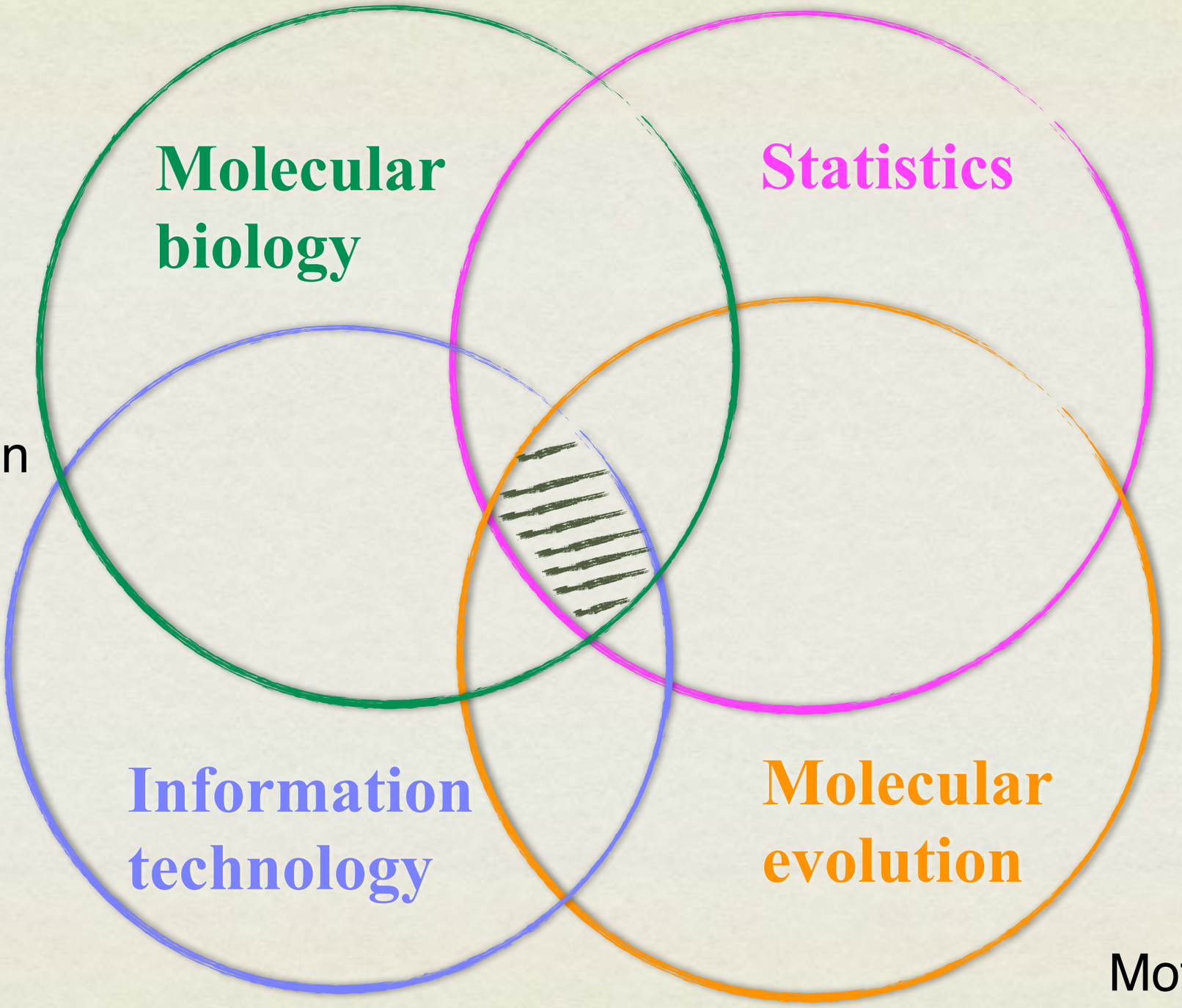
BIOINFORMATICS EMERGED AS AN INTERSECTION BETWEEN DIFFERENT DISCIPLINES

James Watson

Thomas Bayes

Alan Turing

Motoo Kimura

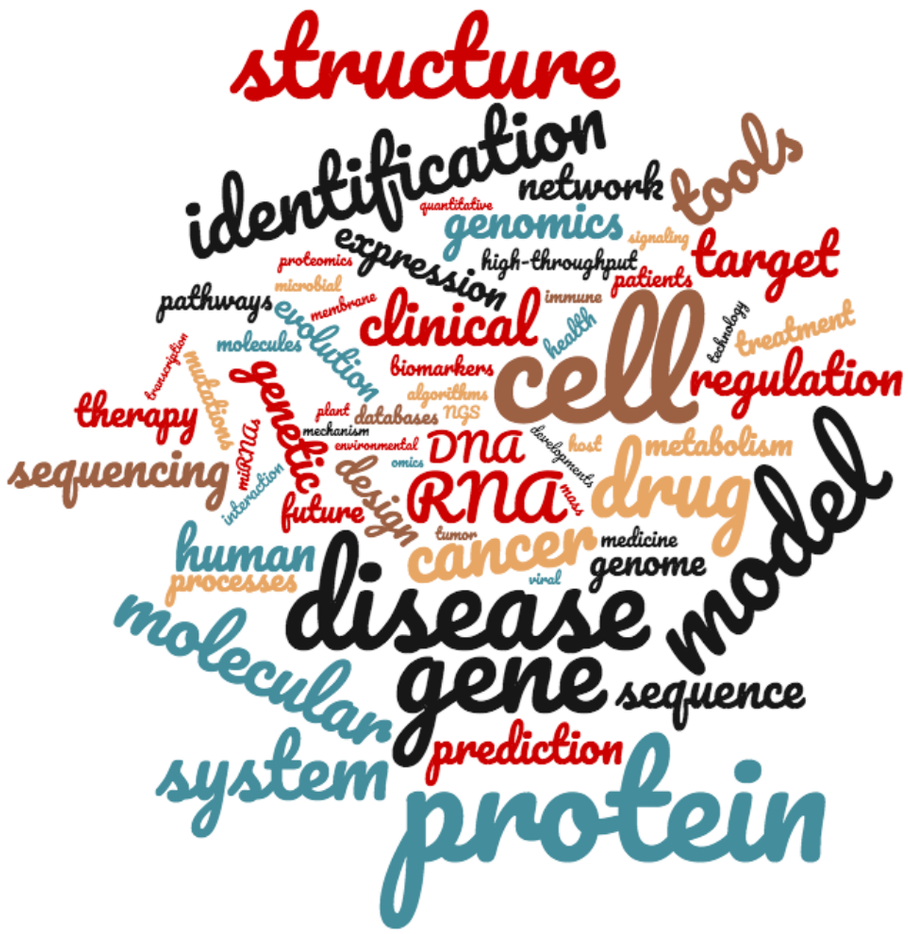


BIOINFORMATICS - DEFINITION

- Research, development, or application of computational tools and approaches for expanding the use of biological data, including those to acquire, store, organize, archive, analyze, or visualize such data.

- Its goal is to enable biological discovery based on existing information or in other words transform biological data into information and eventually into knowledge.

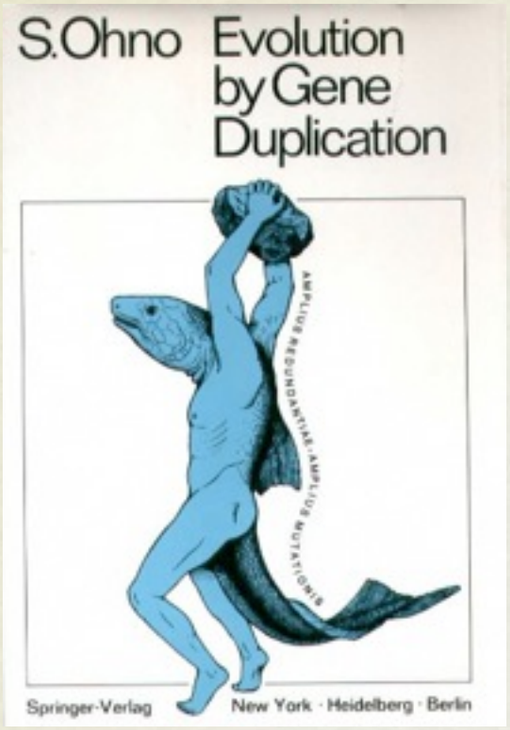
BIOINFORMATICS VERSUS COMPUTATIONAL BIOLOGY

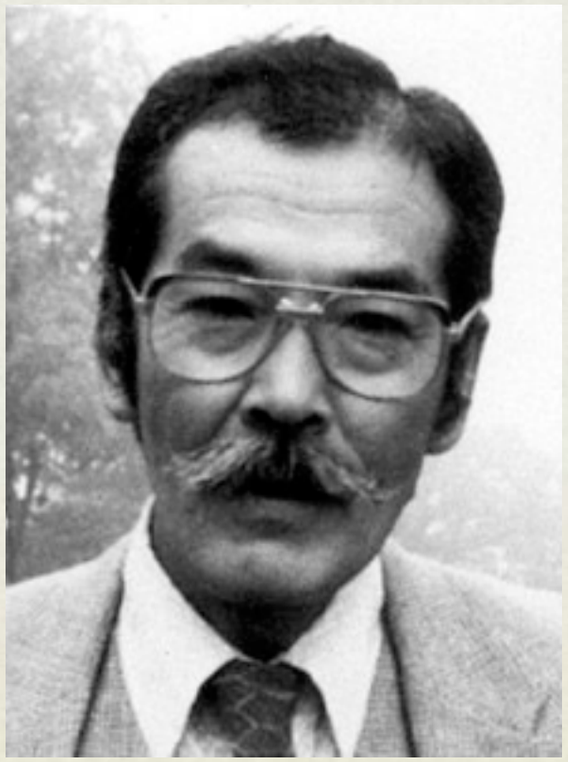


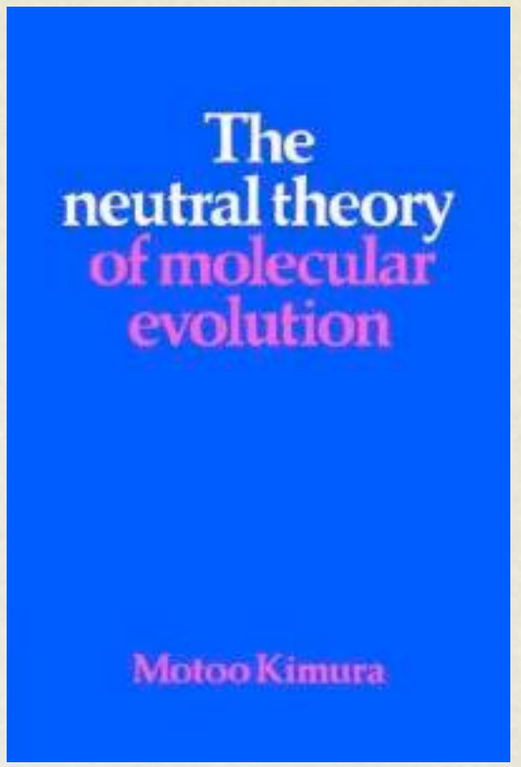
ROLE OF BIOINFORMATICS IN MODERN LIFE SCIENCES

- molecular biology
- molecular evolution
- genomics
- system biology
- protein engineering
- drug design
- human genetics
- personalized medicine

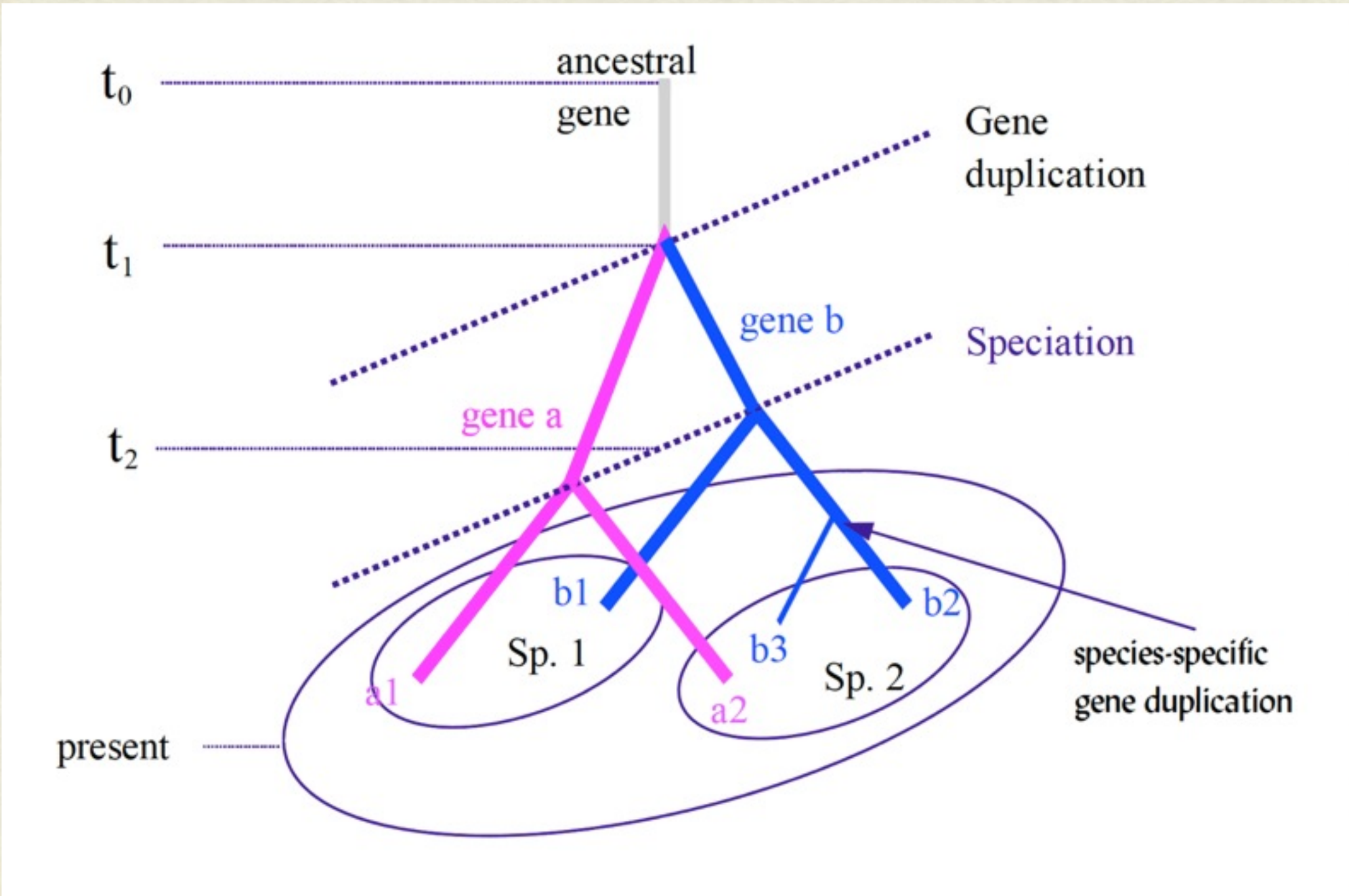
EVOLUTIONARY BASIS OF BIOINFORMATICS



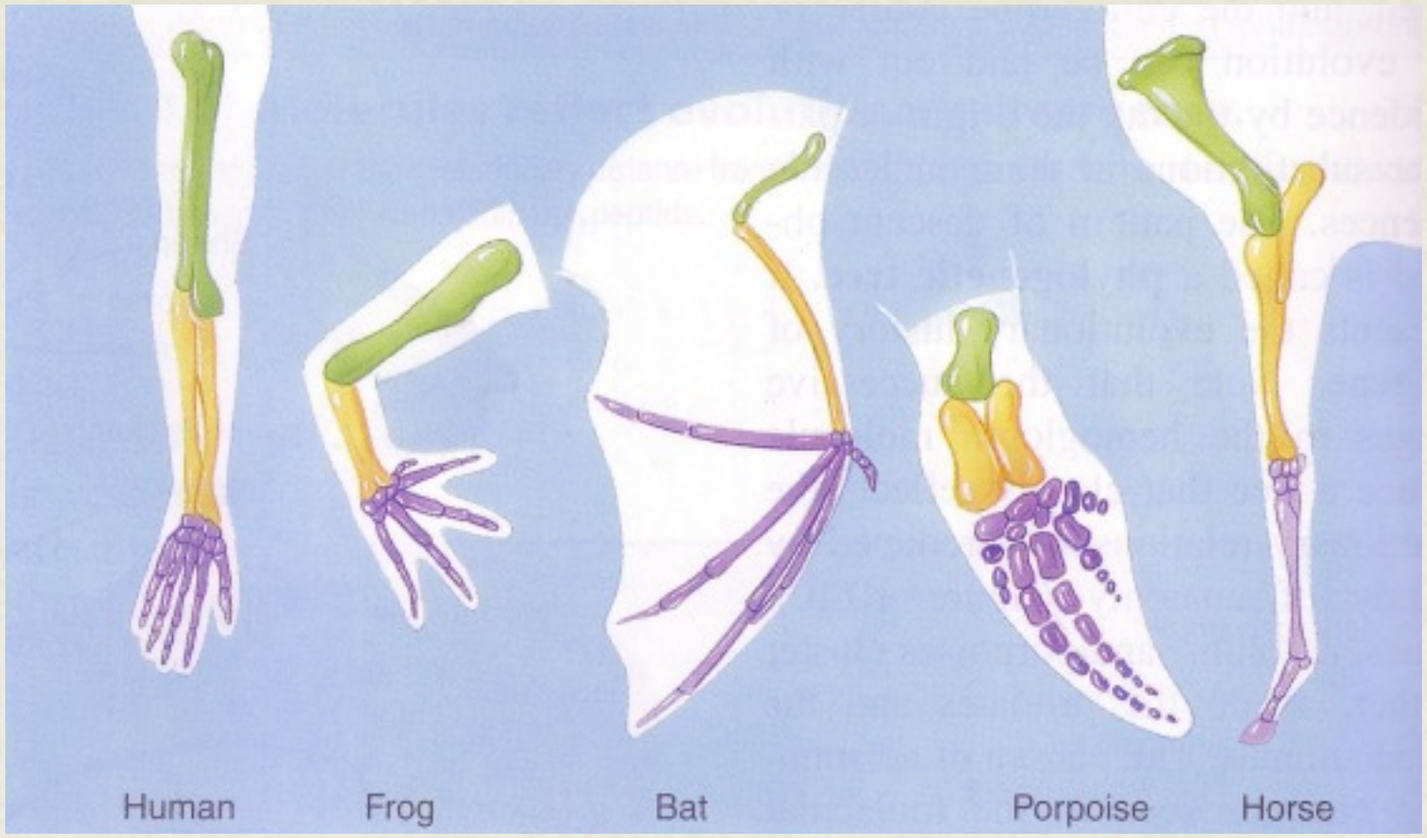




EVOLUTIONARY BASIS OF BIOINFORMATICS



HOMOLOGS



Two anatomical structures or behavioral traits within different organisms which originated from a structure or trait of their common ancestral organism. The structures or traits in their current forms may not necessarily perform the same functions in each organism, nor perform the functions it did in the common ancestor. An example: the wing of a bat, the fin of a whale and the arm of a man are homologous structures.

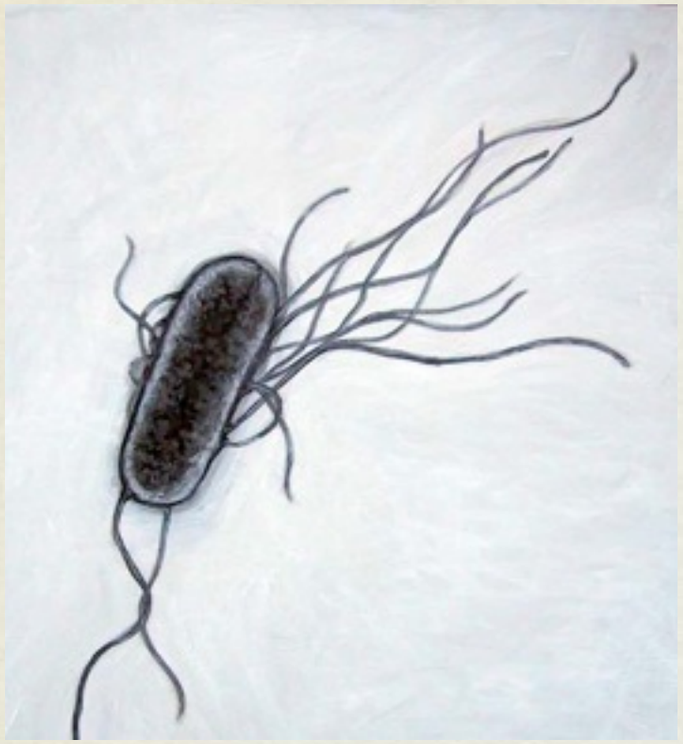
http://www.everythingbio.com/glos/definition.php?ID=3385

HOMOLOGS AT THE MOLECULAR LEVEL

```
cow       ATG---ACTAACATTCGAAAGTCCCACCCACTAATAAAAATTGTAAAC
sheep     ATG---ATCAACATCCGAAAAACCCACCCACTAATAAAAATTGTAAAC
goat      ATG---ACCAACATCCGAAAGACCCACCCATTAATAAAAATTGTAAAC
horse     ATG---ACAAACATCCGGAAATCTCACCCACTAATTAAAATCATCAAT
donkey    ATG---ACAAACATCCGAAAATCCCACCCGCTAATTAAAATCATCAAT
ostrich   ATGGCCCCCAACATTCGAAAATCGCACCCCCTGCTCAAAATTATCAAC
emu       ATGGCCCCTAACATCCGAAAATCCCACCCTCTACTCAAAATCATCAAC
turkey    ATGGCACCCAATATCCGAAAATCACACCCCCTATTAAAAACAATCAAC
```

Two sequences that share common ancestry. Significant sequence similarity usually suggests homology, however sequence similarity may occur also by chance and some homologous sequences may diverge beyond detectable similarity.

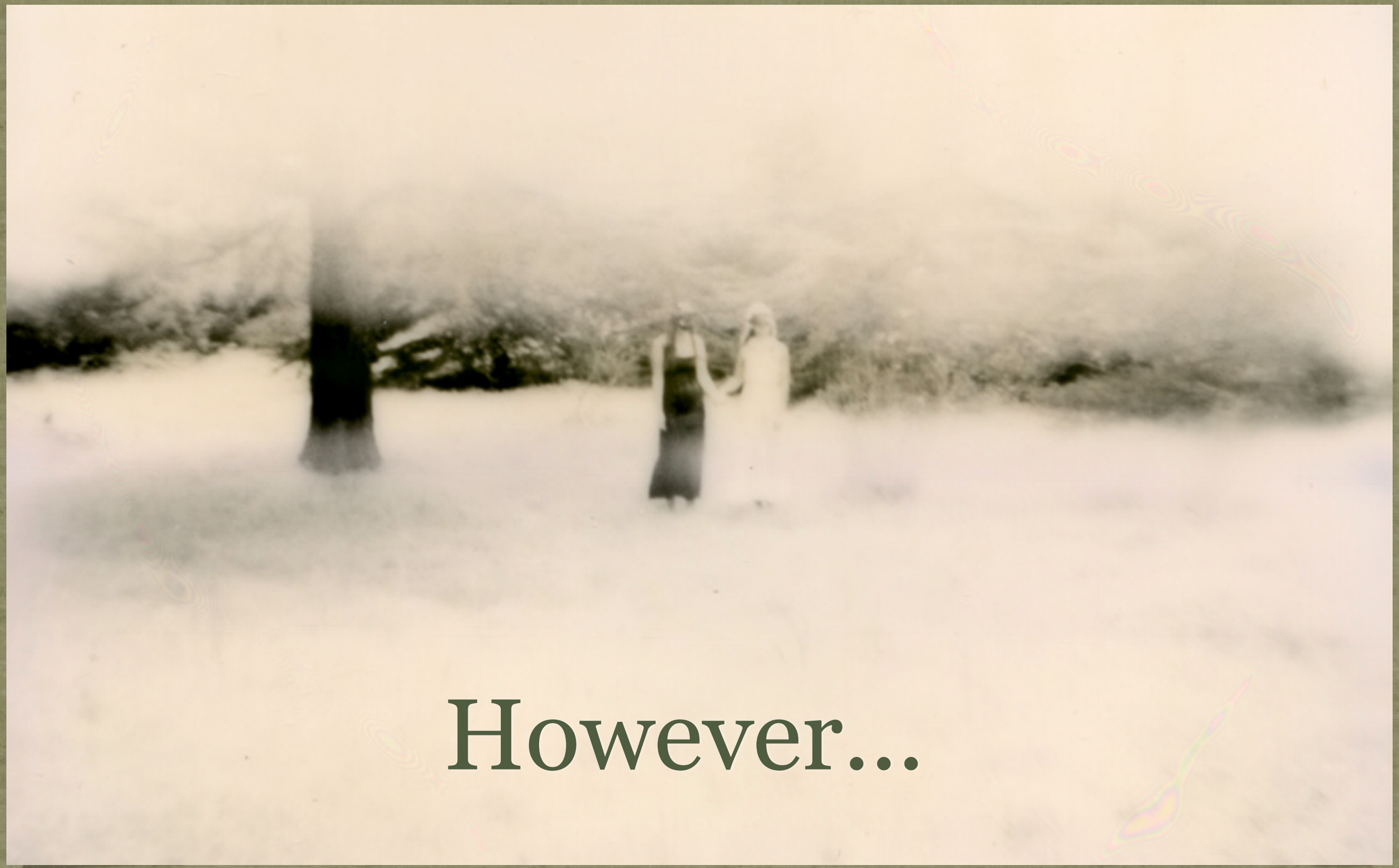
COMPARATIVE GENOMICS



What is true for *E. coli* is also true for elephant.

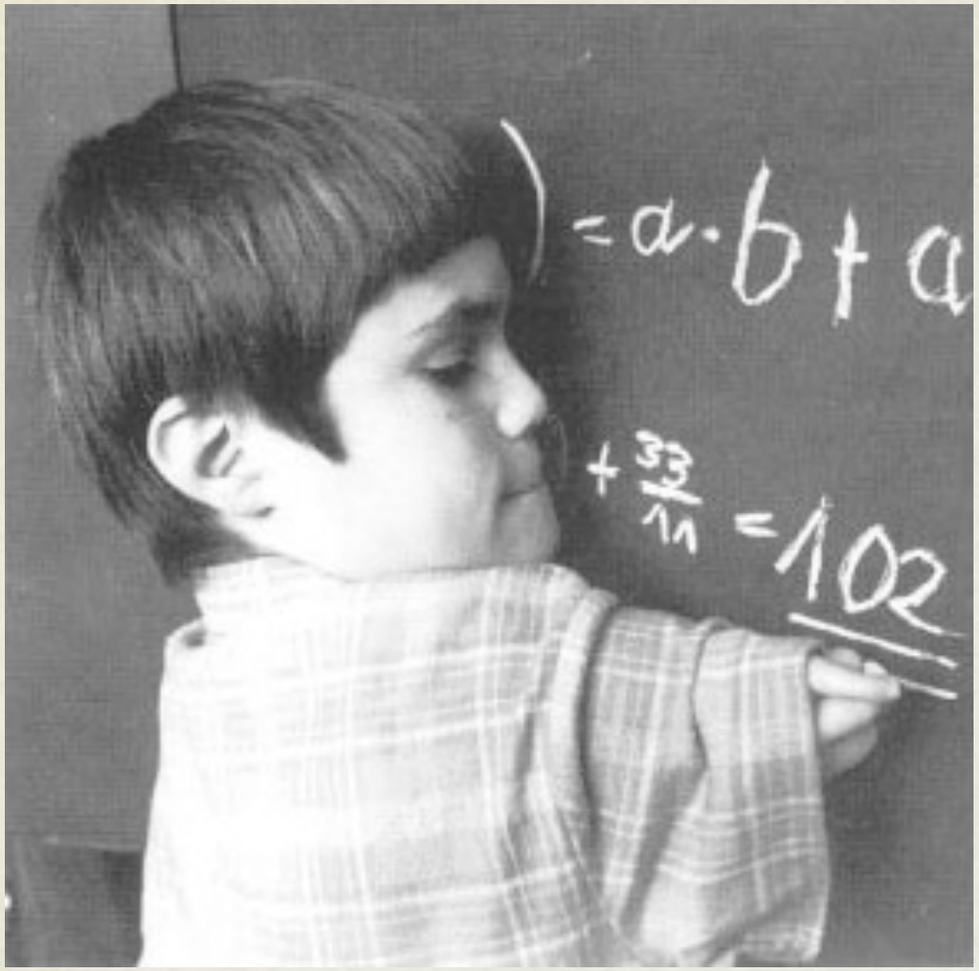
J. Monod, c. 1961

COMPARATIVE GENOMICS



However...

COMPARATIVE GENOMICS

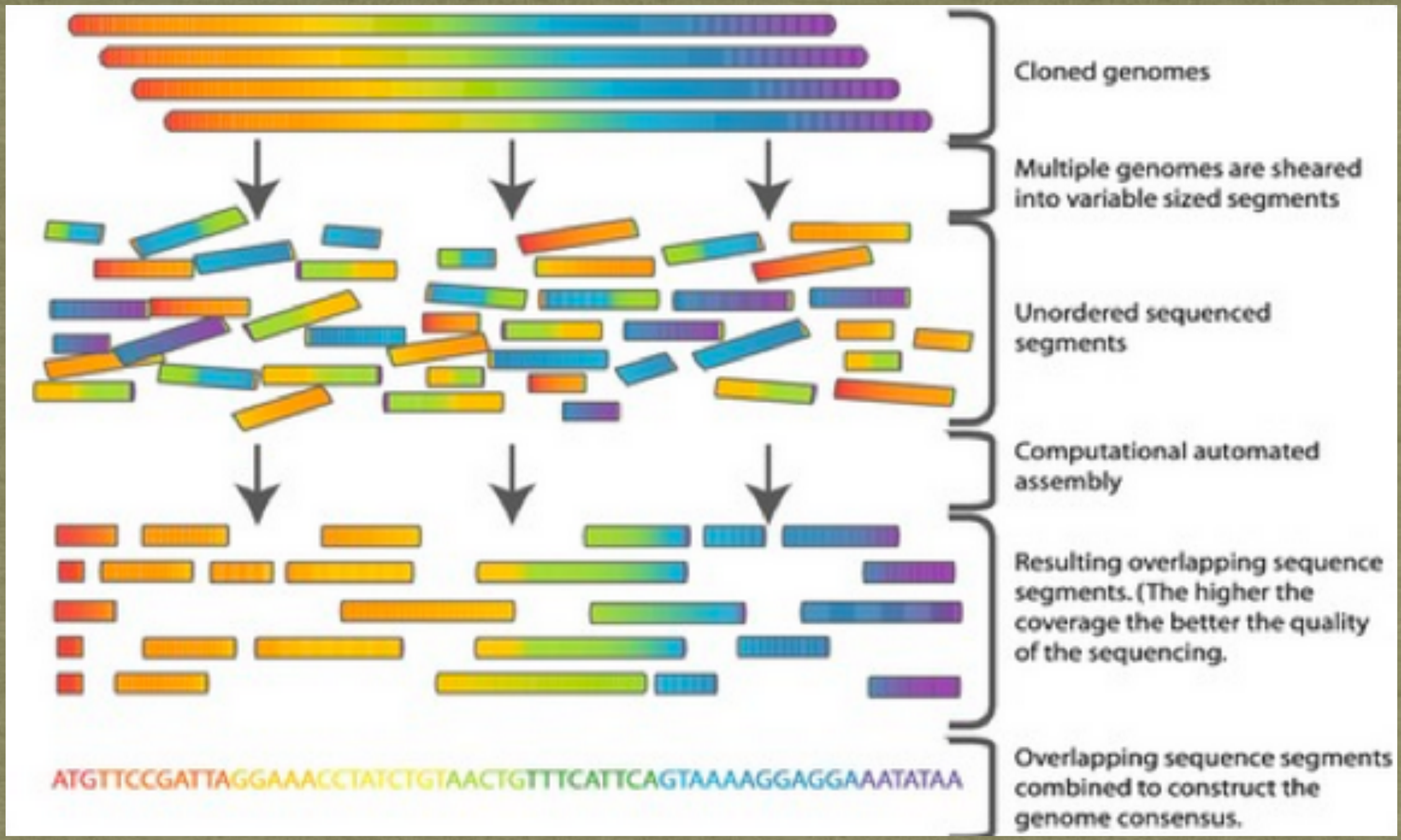


15 000 victims of thalidomide

What is true for mouse is not necessarily true for human…

Nucleotide Sequence Assembly

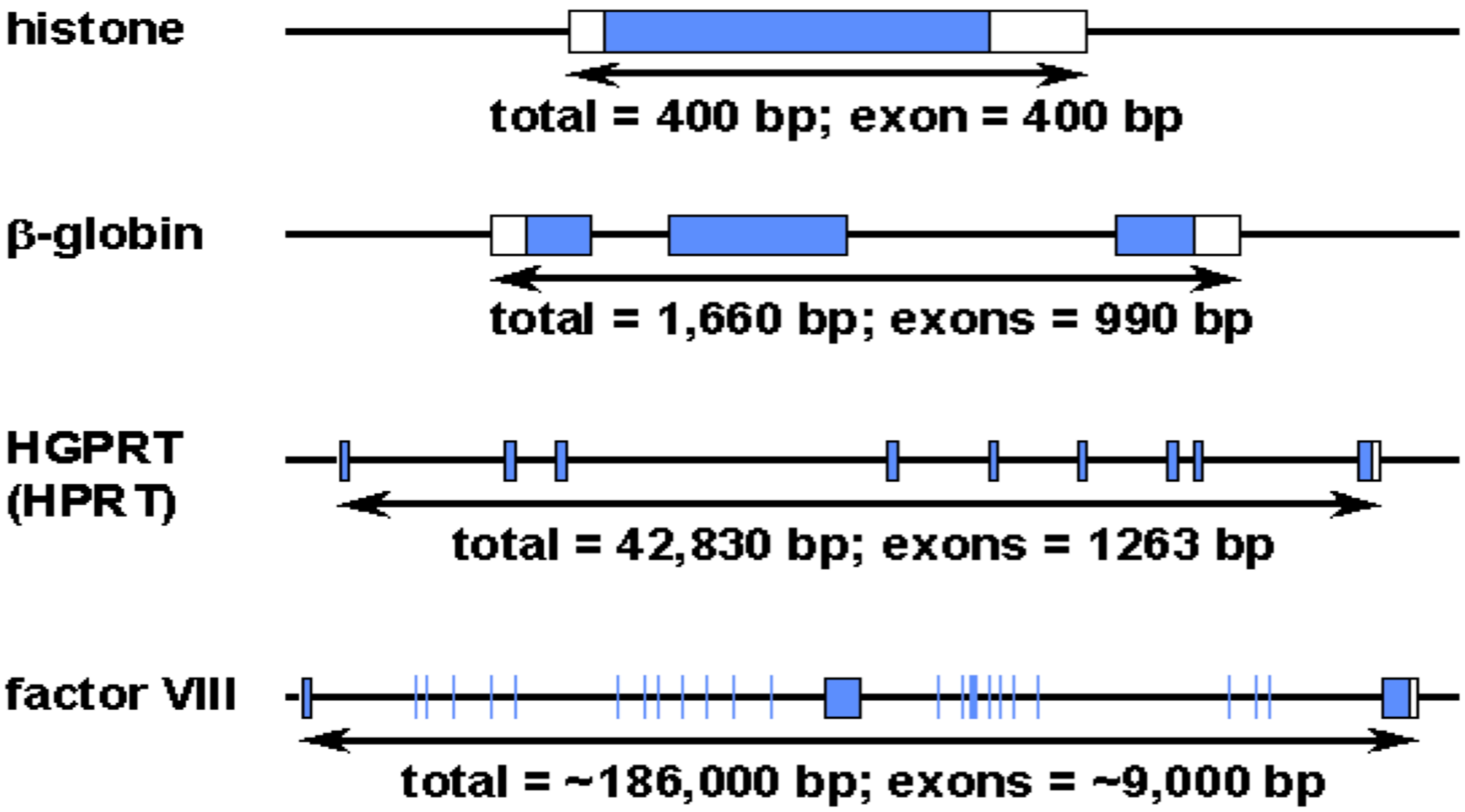
NUCLEOTIDE SEQUENCE ASSEMBLY



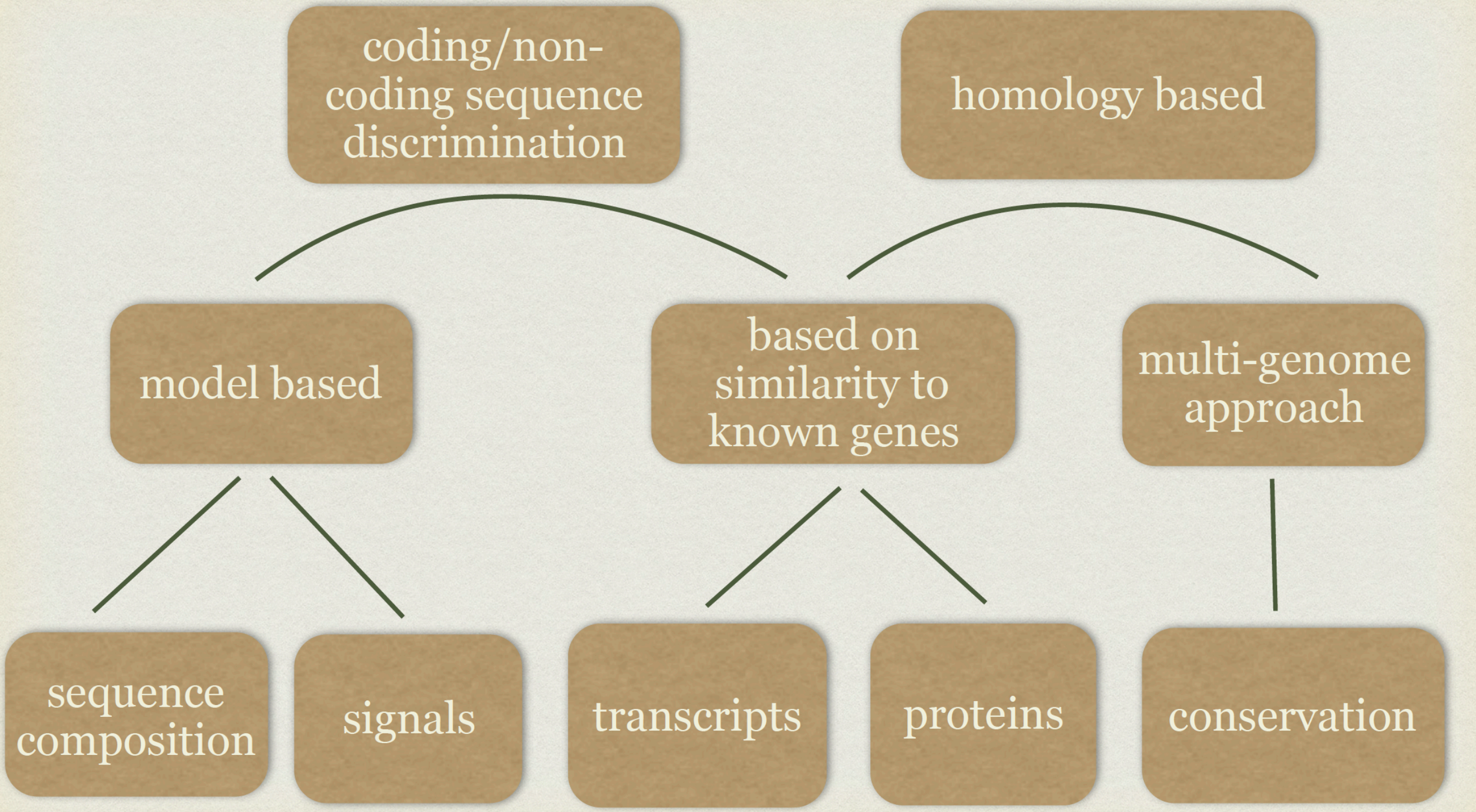
Similarity Search

Gene Prediction

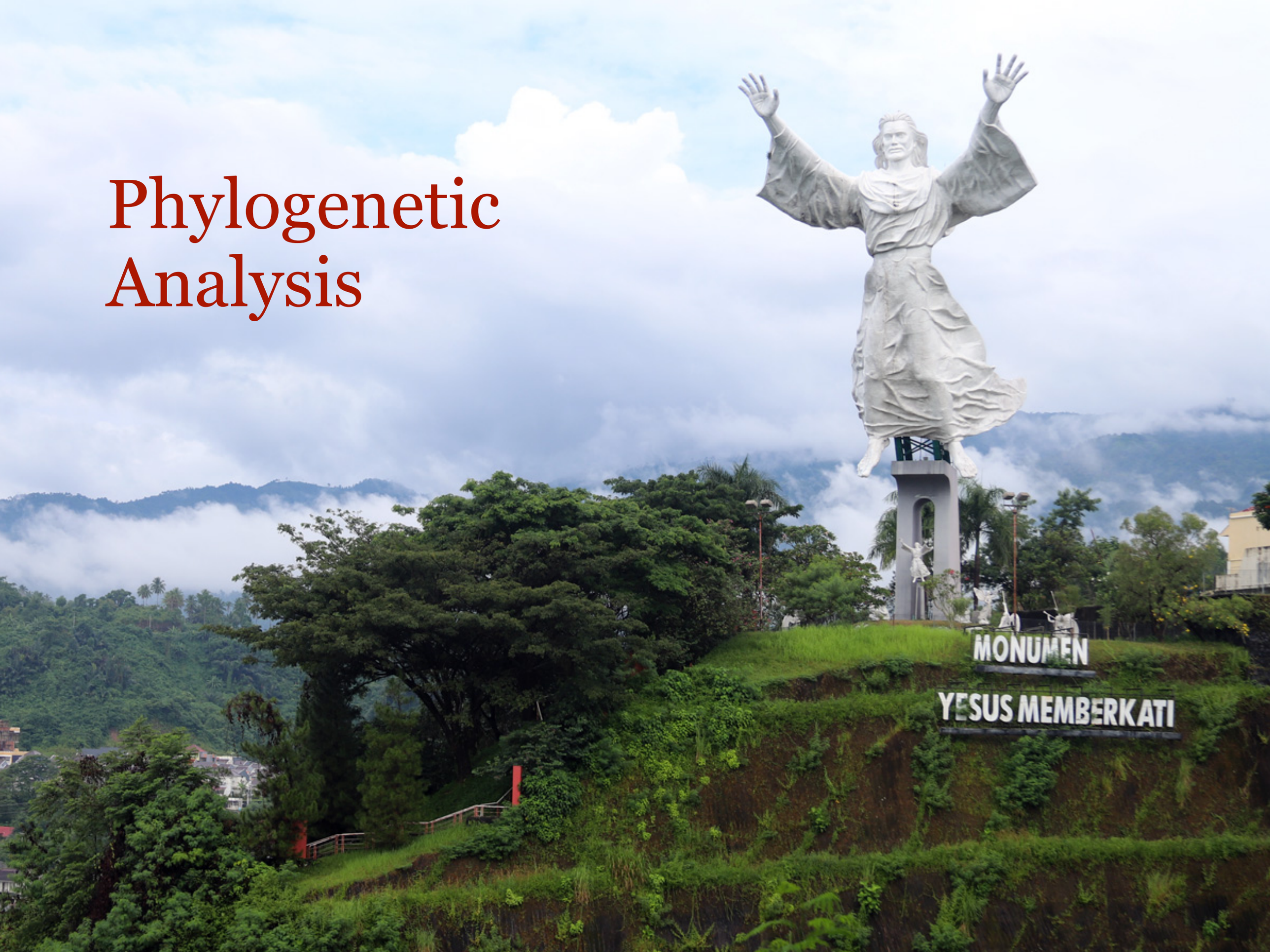
(exon-intron-exon)$_n$ structure of various genes



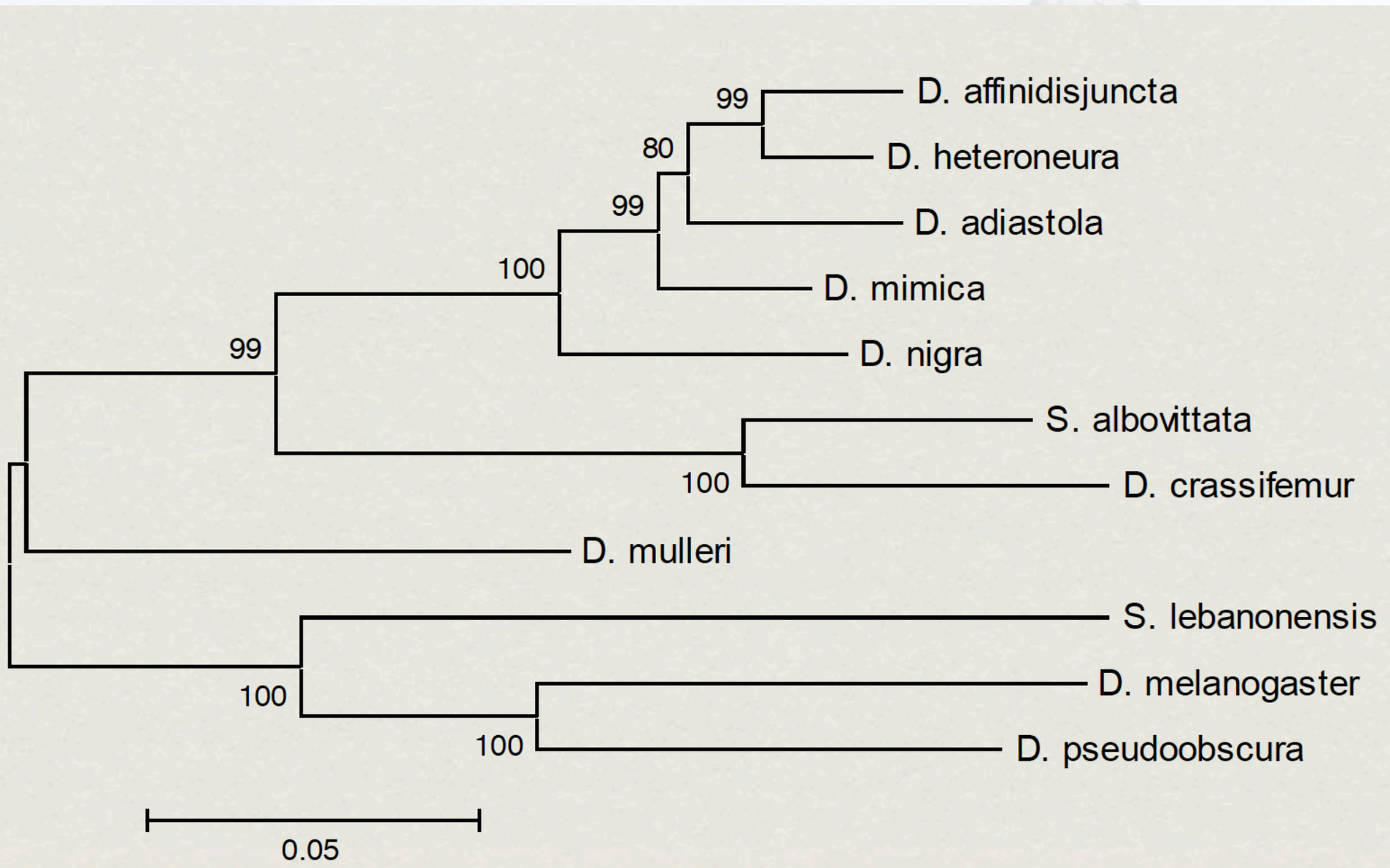
GENE FINDING METHODS



Phylogenetic Analysis



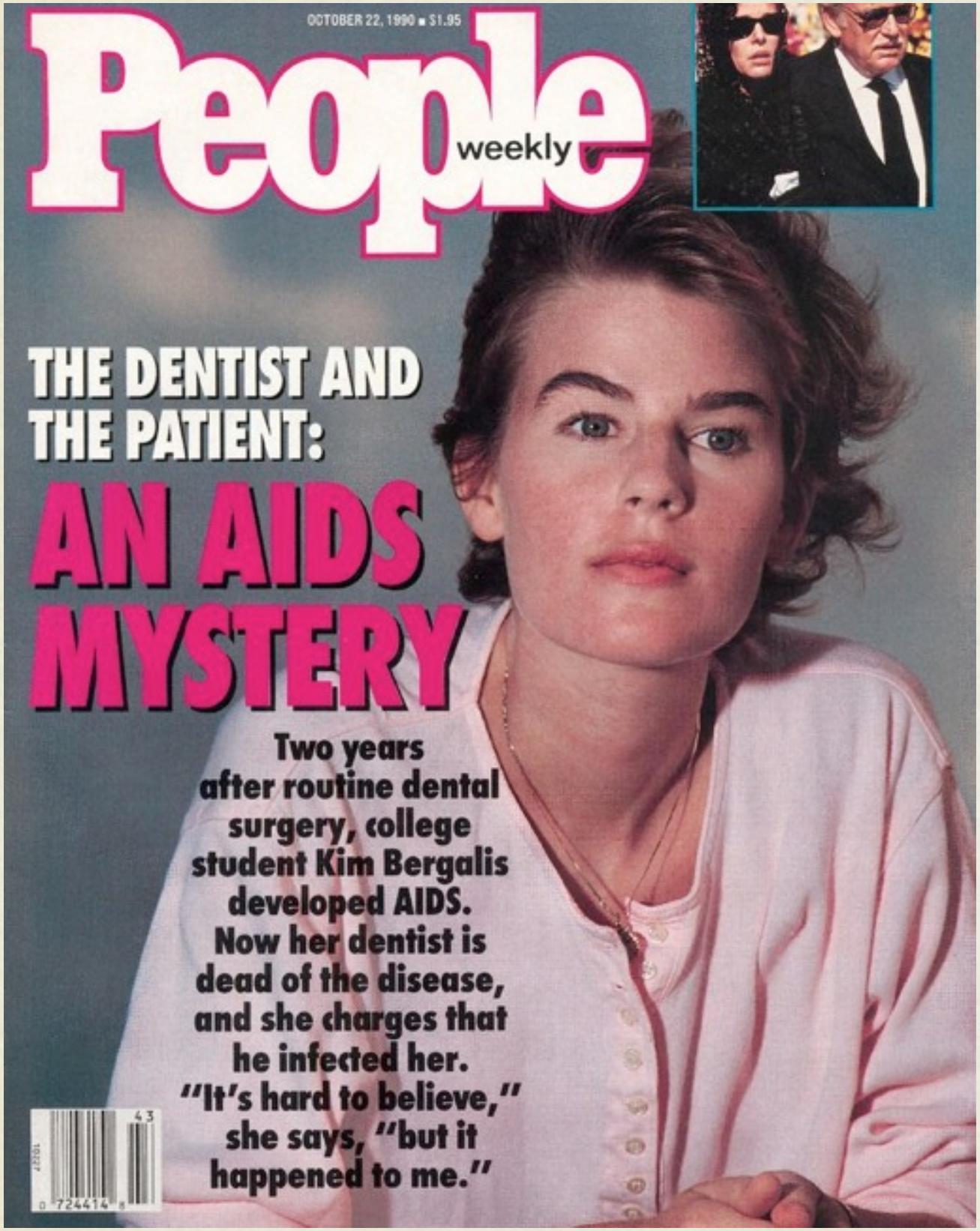
Phylogenetic Analysis



Systems Biology

Translational Bioinformatics

Is Bioinformatics Useful?



Did the Florida Dentist infect his patients with HIV?

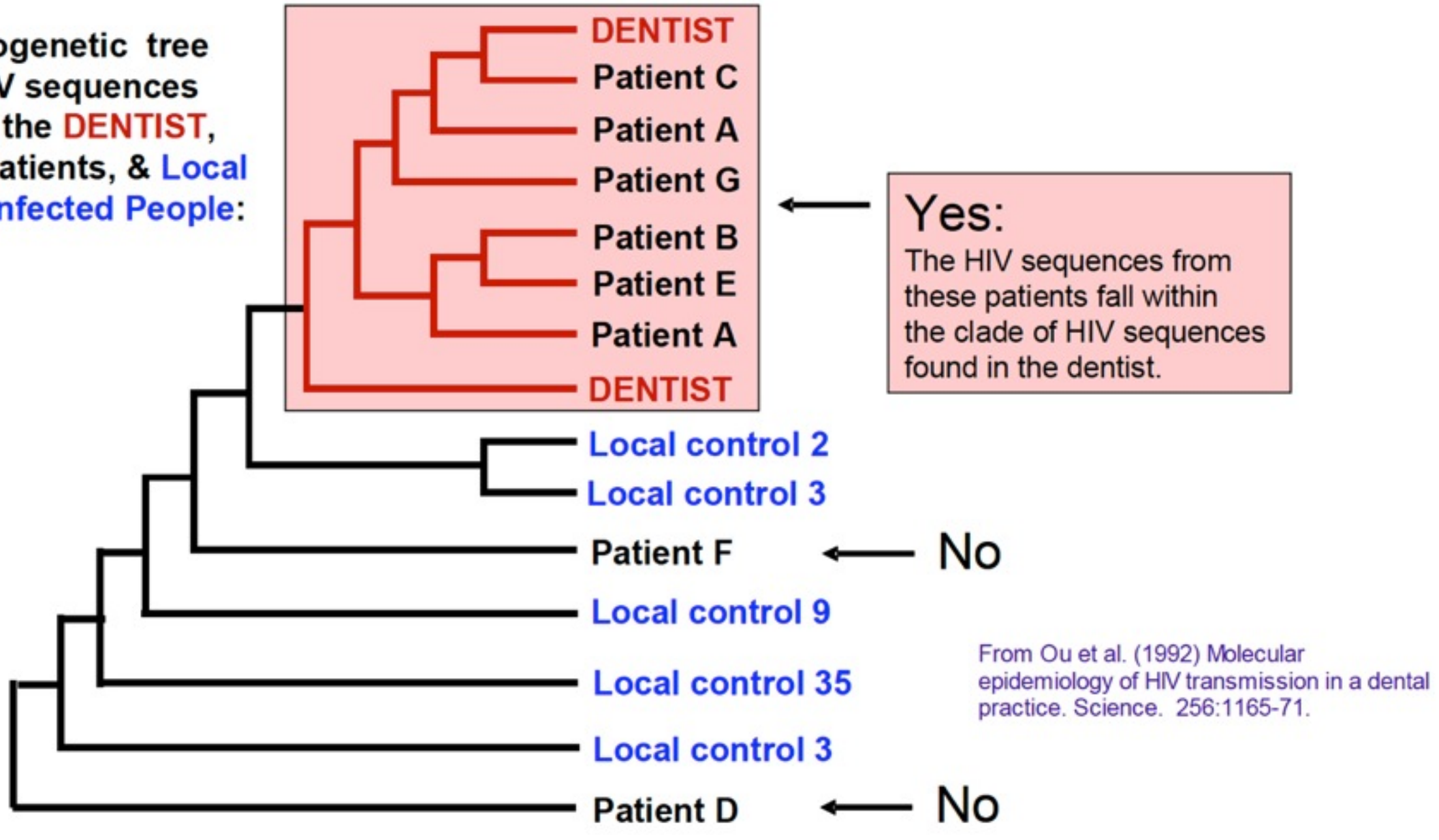
Kimberly Bergalis

(1968-1991)

David J. Acer

(1940-1990)

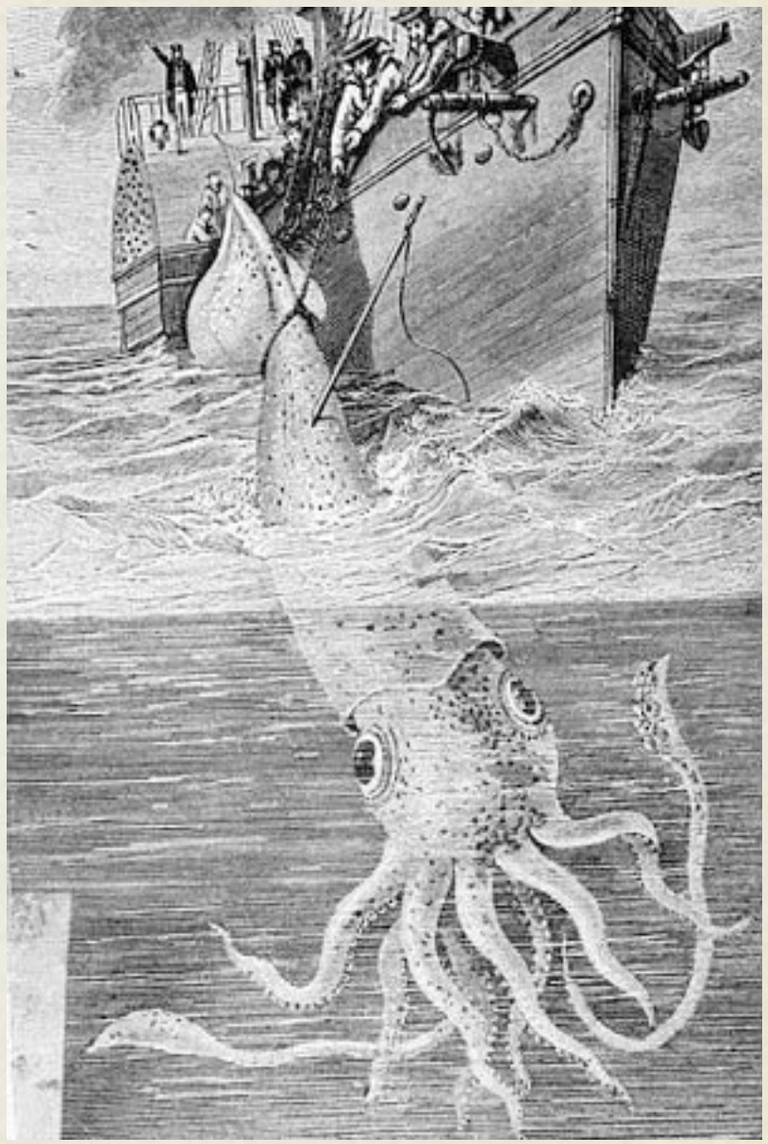
DID THE FLORIDA DENTIST INFECT HIS PATIENTS WITH HIV?



THE MYSTERY OF THE CHILEAN BLOB

THE MYSTERY OF THE CHILEAN BLOB

>Chilean_Blob
TAATACTAACTATATCCCTACTCTCCATTCTCATCGGGG
GTTGAGGAGGACTAAACCAGACTCAACTCCGAAAAATTA
TAGCTTACTCATCAATCGCCCACATAGGATGAATAACCA
CAATCCTACCCTACAATACAACCATAACCCTACTAAACC
TACTAATCTATGTCACAATAACCTTCACCATATTCATAC
TATTTATCCAAAACTCAACCACAACCACACTATCTCTGT
CCCAGACATGAAACAAAACACCCATTACCACAACCCTTA
CCATACTTACCCTACTTTCCATAGGGGGCCTCCCACCAC
TCTCGGGCTTTATCCCCAAATGAATAATTATTCAAGAAC
TAACAAAAAACGAAACCCTCATCATACCAACCTTCATAG
CCACCACAGCATTACTCAACCTCTACTTCTATATACGCC
TCACCTACTCAACAGCACTAACCCTATTCCCCTCCACAA
ATAACATAAAAATAAAATGACAATTCTACCCCACAAAAC
GAATAACCCTCCTGCCAACAGCAATTGTAATATCAACAA
TACTCCTACCCCTTACACCAATACTCTCCACCCTATTAT
AG



THE MYSTERY OF THE CHILEAN BLOB

Lineage Report

Cetacea [whales & dolphins]				
. Odontoceti [whales & dolphins]				
. . Physeteridae [whales & dolphins]				
. . . Physeter catodon	1085	3 hits	[whales & dolphins]	Physeter catodon NADH dehydrogenase subunit 2 (nad2) gene,
. . . Kogia breviceps	638	1 hit	[whales & dolphins]	Kogia breviceps complete mitochondrial genome
. . Orcaella brevirostris	593	1 hit	[whales & dolphins]	Orcaella brevirostris isolate 97 mitochondrion, complete ge
. . Grampus griseus	593	1 hit	[whales & dolphins]	Grampus griseus mitochondrion, complete genome
. . Feresa attenuata	592	2 hits	[whales & dolphins]	Feresa attenuata isolate 36 mitochondrion, complete genome
. . Tursiops truncatus (bottle-nosed dolphin)	592	1 hit	[whales & dolphins]	Tursiops truncatus mitochondrion, complete genome
. . Globicephala melas	586	3 hits	[whales & dolphins]	Globicephala melas isolate GlomelG42 mitochondrion, partial
. . Peponocephala electra	580	2 hits	[whales & dolphins]	Peponocephala electra isolate M6 mitochondrion, complete ge
. . Globicephala macrorhynchus	580	4 hits	[whales & dolphins]	Globicephala macrorhynchus isolate Glomac65 mitochondrion,
. . Pseudorca crassidens	577	3 hits	[whales & dolphins]	Pseudorca crassidens mitochondrion, complete genome
. . Orcinus orca (Orca)	569	54 hits	[whales & dolphins]	Orcinus orca isolate ENPTGA2 mitochondrion, complete genome
. . Sotalia fluviatilis	569	2 hits	[whales & dolphins]	Sotalia fluviatilis haplotype 10 NADH dehydrogenase subunit
. . Platanista minor	569	1 hit	[whales & dolphins]	Platanista minor complete mitochondrial genome
. . Steno bredanensis	566	2 hits	[whales & dolphins]	Steno bredanensis isolate StebreS9 mitochondrion, partial g
. Megaptera novaeangliae	636	5 hits	[whales & dolphins]	Megaptera novaeangliae voucher GOM9049 NADH dehydrogenase s
. Balaenoptera bonaerensis	630	1 hit	[whales & dolphins]	Balaenoptera bonaerensis mitochondrial DNA, complete genome
. Eubalaena japonica	619	1 hit	[whales & dolphins]	Eubalaena japonica mitochondrial DNA, complete genome
. Balaenoptera brydei	614	2 hits	[whales & dolphins]	Balaenoptera brydei mitochondrial DNA, complete genome, iso
. Balaena mysticetus (Greenland right whale)	614	2 hits	[whales & dolphins]	Balaena mysticetus mitochondrial DNA, complete genome
. Balaenoptera musculus				
. Balaenoptera edeni				
. Balaenoptera omurai				
. Eschrichtius robustus (California gray whale)				
. Balaenoptera borealis				
. Caperea marginata				
. Balaenoptera physalus (finback whale)				

THE MYSTERY OF THE CHILEAN BLOB

```
>emb|AJ277029.2|  Physeter macrocephalus mitochondrial genome
Length=16428

 Score = 1074 bits (581),  Expect = 0.0
 Identities = 585/587 (99%), Gaps = 0/587 (0%)
 Strand=Plus/Plus

Query  1     TAATACTAACTATATCCCTACTCTCCATTCTCATCGGGGGTTGAGGAGGACTAAACCAGA  60
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  4400  TAATACTAACTATATCCCTACTCTCCATTCTCATCGGGGGTTGAGGAGGACTAAACCAGA  4459

Query  61    CTCAACTCCGAAAAATTATAGCTTACTCATCAATCGCCCACATAGGATGAATAACCACAA  120
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  4460  CTCAACTCCGAAAAATTATAGCTTACTCATCAATCGCCCACATAGGATGAATAACCACAA  4519

Query  121   TCCTACCCTACAATACAACCATAACCCTACTAAACCTACTAATCTATGTCACAATAACCT  180
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  4520  TCCTACCCTACAATACAACCATAACCCTACTAAACCTACTAATCTATGTCACAATAACCT  4579

Query  181   TCACCATATTCATACTATTTATCCAAAACTCAACCACAACCACACTATCTCTGTCCCAGA  240
             |||||||||||| ||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  4580  TCACCATATTCACACTATTTATCCAAAACTCAACCACAACCACACTATCTCTGTCCCAGA  4639

Query  241   CATGAAACAAAACACCCATTACCACAACCCTTACCATACTTACCCTACTTTCCATAGGGG  300
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  4640  CATGAAACAAAACACCCATTACCACAACCCTTACCATACTTACCCTACTTTCCATAGGGG  4699

Query  301   GCCTCCCACCACTCTCGGGCTTTATCCCCAAATGAATAATTATTCAAGAACTAACAAAAA  360
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  4700  GCCTCCCACCACTCTCGGGCTTTATCCCCAAATGAATAATTATTCAAGAACTAACAAAAA  4759

Query  361   ACGAAACCCTCATCATACCAACCTTCATAGCCACCACAGCATTACTCAACCTCTACTTCT  420
             ||||| ||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  4760  ACGAAGCCCTCATCATACCAACCTTCATAGCCACCACAGCATTACTCAACCTCTACTTCT  4819

Query  421   ATATACGCCTCACCTACTCAACAGCACTAACCCTATTCCCCTCCACAAATAACATAAAAA  480
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  4820  ATATACGCCTCACCTACTCAACAGCACTAACCCTATTCCCCTCCACAAATAACATAAAAA  4879

Query  481   TAAAATGACAATTCTACCCCACAAAACGAATAACCCTCCTGCCAACAGCAATTGTAATAT  540
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  4880  TAAAATGACAATTCTACCCCACAAAACGAATAACCCTCCTGCCAACAGCAATTGTAATAT  4939

Query  541   CAACAATACTCCTACCCCTTACACCAATACTCTCCACCCTATTATAG  587
             |||||||||||||||||||||||||||||||||||||||||||||
Sbjct  4940  CAACAATACTCCTACCCCTTACACCAATACTCTCCACCCTATTATAG  4986
```

Data Volume Problem

Type of cancer	Number of whole genome samples	Number of whole exome samples	Data volume (Tb)	Time to download
Colon Adenocarcinoma (COAD)	302	443	33.04	24 days
Lung	134	582	40.95	30 days
Breast	248	1050	69.82	50 days
Prostate Adenocarcinoma (PRAD)	272	1049	26.53	10 days

LET'S TAKE A BREAK

Where to start?

where to Start

ONE OF CENTRAL DEPOSITORIES

NCBI
National Center for Biotechnology Information

https://www.ncbi.nlm.nih.gov

EMBL-EBI
The home for big data in biology

https://www.ebi.ac.uk

DDBJ
DNA Data Bank of Japan

https://www.ddbj.nig.ac.jp/index-e.html

NCBI
National Center for Biotechnology Information

All Databases

search term

NCBI Home
Resource List (A-Z)
All Resources
Chemicals & Bioassays
Data & Software
DNA & RNA
Domains & Structures
Genes & Expression
Genetics & Medicine
Genomes & Maps
Homology
Literature
Proteins
Sequence Analysis
Taxonomy
Training & Tutorials
Variation

Database selection

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

About the NCBI | Mission | Organization | NCBI News & Blog

Submit

Deposit data or manuscripts into NCBI databases

Download

Transfer NCBI data to your computer

Learn

Find help documents, attend a class or watch a tutorial

Develop

Use NCBI APIs and code libraries to build applications

Analyze

Identify an NCBI tool for your data analysis task

Research

Explore NCBI research and collaborative projects

NCBI Databases

Results found in 9 databases for: **makalowski**

Literature

Bookshelf	2
MeSH	0
NLM Catalog	1
PubMed	136
PubMed Central	250

Genes

Gene	0
GEO DataSets	0
GEO Profiles	0
HomoloGene	0
PopSet	1
UniGene	0

Genetics

ClinVar	0
dbGaP	0
dbSNP	0
dbVar	0
GTR	0
MedGen	0
OMIM	2

Proteins

Conserved Domains	0
Identical Protein Groups	0
Protein	325,543
Protein Clusters	0
Sparcle	0
Structure	0

Genomes

Assembly	0
BioCollections	0
BioProject	0
BioSample	0
Genome	0
Nucleotide	361,880

Chemicals

BioSystems	0
PubChem BioAssay	0
PubChem Compound	0
PubChem Substance	0

Let's search for "globin X" sequences

Protein | "globin x" | Search

Create alert | Advanced | Help

Summary | 20 per page | Sort by Default order | Send to:

Filters: Manage Filters

See Gene information for **globin x**
globin in Crassostrea gigas | Danio rerio | Musca domestica | All 10 Gene records
x in Hepatitis B virus | Escherichia virus P2 | Escherichia virus Wphi | All 56 Gene records

See the results of this search (22 items) in our new Identical Protein Groups database.

Items: 1 to 20 of 275

<< First | < Prev | Page 1 of 14 | Next > | Last >>

1. **globin X** [Clonorchis sinensis]
   303 aa protein
   Accession: GAA47520.1 GI: 358339458
   BioProject | Nucleotide | PubMed | Taxonomy
   GenPept | Identical Proteins | FASTA | Graphics

2. **globin X**, partial [Platichthys flesus]
   152 aa protein
   Accession: CCO03031.1 GI: 440575635
   Nucleotide | Taxonomy
   GenPept | Identical Proteins | FASTA | Graphics

Find related data
Database: Select
Find items

Search details
"globin x"[All Fields]
Search
See more...

Recent activity
Turn Off | Clear
"globin x" (275) Protein
Phylogenetic analysis reveals wide distribution of globin X. PubMed

We can make results more specific

Protein "globin x" [title] Search

Create alert Advanced Help

Summary 20 per page Sort by Default order Send to:

See the results of this search (22 items) in our new Identical Protein Groups database.

Items: 1 to 20 of 157

<< First < Prev Page 1 of 8 Next > Last >>

1. **Globin X** [Fasciola hepatica]
   306 aa protein
   Accession: THD28802.1 GI: 1620785640
   BioProject Nucleotide Taxonomy
   GenPept Identical Proteins FASTA Graphics

2. **Globin X**, Uncharacterized protein [Nothobranchius furzeri]
   198 aa protein
   Accession: SBP57577.1 GI: 1077058874
   BioProject Nucleotide Taxonomy
   GenPept Identical Proteins FASTA Graphics

3. **Globin X**, Uncharacterized protein [Nothobranchius kuhntae]
   162 aa protein

Filters: Manage Filters

Results by taxon

Top Organisms [Tree]
- Alvinella pompejana *(134)*
- Tetraodon nigroviridis *(3)*
- Nothobranchius kadleci *(2)*
- Nothobranchius pienaari *(2)*
- Nothobranchius kuhntae *(2)*
- All other taxa *(14)*

More...

Find related data

Database: Select

Find items

Search details

```
"globin x"[title]
```

We can make results *even* more specific

Protein | "globin x" [title] NOT partial [title] | Search

Create alert Advanced Help

Summary 50 per page Sort by Default order Send to:

Filters: Manage Filters

See the results of this search (22 items) in our new Identical Protein Groups database.

Items: 19

1. **Globin X** [Fasciola hepatica]
   306 aa protein
   Accession: THD28802.1 GI: 1620785640
   BioProject Nucleotide Taxonomy
   GenPept Identical Proteins FASTA Graphics

2. **Globin X**, Uncharacterized protein [Nothobranchius furzeri]
   198 aa protein
   Accession: SBP57577.1 GI: 1077058874
   BioProject Nucleotide Taxonomy
   GenPept Identical Proteins FASTA Graphics

3. **Globin X**, Uncharacterized protein [Nothobranchius kuhntae]
   162 aa protein
   Accession: SBR04581.1 GI: 1075786502
   BioProject Nucleotide Taxonomy
   GenPept Identical Proteins FASTA Graphics

Results by taxon

Top Organisms [Tree]
- Tetraodon nigroviridis *(3)*
- Nothobranchius pienaari *(2)*
- Nothobranchius furzeri *(2)*
- Schistosoma mansoni *(2)*
- Nothobranchius kuhntae *(1)*
- All other taxa *(9)*

More...

Analyze these sequences

Run BLAST

Align sequences with COBALT

Identify Conserved Domains with CD-Search

Find in these sequences

Find related data

Database: Select

The cool stuff starts *NOW!*

Protein | "globin x" [title] NOT partial [title] | Search

Create alert | Advanced | Help

Summary | 50 per page | Sort by Default order | Send to: | **Filters:** Manage Filters

See the results of this search (22 items) in our new Identical Protein Groups database.

Items: 19

1. **Globin X** [Fasciola hepatica]
   306 aa protein
   Accession: THD28802.1 GI: 1620785640
   BioProject Nucleotide Taxonomy
   GenPept Identical Proteins FASTA Graphics

2. **Globin X**, Uncharacterized protein [Nothobranchius furzeri]
   198 aa protein
   Accession: SBP57577.1 GI: 1077058874
   BioProject Nucleotide Taxonomy
   GenPept Identical Proteins FASTA Graphics

3. **Globin X**, Uncharacterized protein [Nothobranchius kuhntae]
   162 aa protein
   Accession: SBP04581.1 GI: 1075786502

Results by taxon

Top Organisms [Tree]
- Tetraodon nigroviridis *(3)*
- Nothobranchius pienaari *(2)*
- Nothobranchius furzeri *(2)*
- Schistosoma mansoni *(2)*
- Nothobranchius kuhntae *(1)*
- All other taxa *(9)*

More...

Analyze these sequences

Run BLAST

Align sequences with COBALT

Identify Conserved Domains with CD-Search

Find in these sequences

Find related data

Let's run BLAST

All 19 amino acid sequences will be used as a query

rotein | "globin x" [title] NOT partial [title] | Search

Create alert | Advanced | Help

Summary | 50 per page | Sort by Default order | Send to: | **Filters:** Manage Filters

See the results of this search (22 items) in our new Identical Protein Groups database.

Items: 19

1. **Globin X** [Fasciola hepatica]
   306 aa protein
   Accession: THD28802.1 GI: 1620785640
   BioProject Nucleotide Taxonomy
   GenPept Identical Proteins FASTA Graphics

2. **Globin X**, Uncharacterized protein [Nothobranchius furzeri]
   198 aa protein
   Accession: SBP57577.1 GI: 1077058874
   BioProject Nucleotide Taxonomy
   GenPept Identical Proteins FASTA Graphics

3. **Globin X**, Uncharacterized protein [Nothobranchius kuhntae]
   162 aa protein
   Accession: SBP04581.1 GI: 1075786502

Results by taxon

Top Organisms [Tree]
- Tetraodon nigroviridis *(3)*
- Nothobranchius pienaari *(2)*
- Nothobranchius furzeri *(2)*
- Schistosoma mansoni *(2)*
- Nothobranchius kuhntae *(1)*
- All other taxa *(9)*

More...

Analyze these sequences

Run BLAST

Align sequences with COBALT

Identify Conserved Domains with CD-Search

Find in these sequences

Find related data

Standard Protein BLAST

blastn | **blastp** | blastx | tblastn | tblastx

BLASTP programs search protein databases using a protein query. more...

Reset page | Bookmark

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

THD28802.1
SBP57577.1
SBR04581.1
SBR92617.1
SBQ48984.1

List of proteins from our search

Query subrange

From

To

We are beta testing a New Results page

Click here if you would like to see your results in the new format. You can always switch back to the Traditional Results page.

Or, upload file Choose File no file selected

Job Title

Enter a descriptive title for your BLAST search

Align two or more sequences

Choose Search Set

Database Non-redundant protein sequences (nr)

Organism Optional

Enter organism name or id--completions will be suggested exclude +

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude Optional

Models (XM/XP) Non-redundant RefSeq proteins (WP) Uncultured/environmental sample sequences

Database choice and results limitation options

Program Selection

Algorithm

- Quick BLASTP (Accelerated protein-protein BLAST)
- blastp (protein-protein BLAST)
- PSI-BLAST (Position-Specific Iterated BLAST)
- PHI-BLAST (Pattern Hit Initiated BLAST)
- DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm

Different types (algorithms) of BLAST

BLAST Search **database nr** using **Blastp (protein-protein BLAST)**

Show results in a new window

Klick here to start BLAST

BLAST results

BLAST® » blastp suite » results for RID-J1D9YG74015 Home Recent Results Saved Strategies Help

Edit Search | Save Search | Search Summary | How to read this report? | BLAST Help Videos | Back to Traditional Results Page | BETA

Job Title	gb|THD28802.1|
RID	J1D9YG74015 *Search expires on 07-07 22:50 pm* Download All
Results for	1:gb|THD28802.1 Globin X [Fasciola hepatica](306aa)
Program	BLASTP Citation
Database	nr See details
Query ID	THD28802.1
Description	Globin X [Fasciola hepatica]
Molecule type	amino acid
Query Length	306
Other reports	Distance tree of results Multiple alignment

Select query

Filter Results

Filter results

Organism *only top 20 will appear* exclude

Type common name, binomial, taxid or group name

Add organism

Percent Identity to

E value to

Filter Reset

Descriptions | Graphic Summary | Alignments | Taxonomy

Sequences producing significant alignments Download Manage Columns Show 100

select all *100 sequences selected* GenPept Graphics Distance tree of results Multiple alignment

Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
Globin X [Fasciola hepatica]	635	635	100%	0.0	100.00%	THD28802.1
Globin X [Fasciola gigantica]	614	614	99%	0.0	97.05%	TPP63302.1

Type of results

BLAST results: descriptions

Descriptions | Graphic Summary | Alignments | Taxonomy

Sequences producing significant alignments — Download — Manage Columns — Show 100

select all *100 sequences selected* — GenPept — Graphics — Distance tree of results — Multiple alignment

Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
Globin X [Fasciola hepatica]	635	635	100%	0.0	100.00%	THD28802.1
Globin X [Fasciola gigantica]	614	614	99%	0.0	97.05%	TPP63302.1
hypothetical protein CRM22_002376 [Opisthorchis felineus]	304	304	85%	1e-99	56.49%	TGZ71927.1
unnamed protein product [Echinostoma caproni]	297	297	69%	3e-98	68.20%	VDP67930.1
hypothetical protein T265_04650 [Opisthorchis viverrini]	300	300	93%	7e-98	52.08%	XP_009167705.1
hypothetical protein CSKR_5917s [Clonorchis sinensis]	298	298	87%	2e-97	54.78%	RJW73736.1
globin [Opisthorchis viverrini]	295	295	92%	3e-96	52.11%	OON21854.1
Leghemoglobin-1 isoform 1 [Schistosoma japonicum]	218	218	83%	2e-65	42.15%	TNN15233.1
SJCHGC09035 protein [Schistosoma japonicum]	204	204	58%	1e-61	50.00%	AAW24922.1
unnamed protein product [Schistosoma margrebowiei]	205	205	77%	3e-60	44.03%	VDP16629.1
uncharacterized protein DC041_0005585 [Schistosoma bovis]	201	201	77%	9e-59	43.22%	RTG83182.1
unnamed protein product [Schistosoma mattheei]	168	168	61%	1e-46	43.08%	VDP64096.1
PREDICTED: cytoglobin-2-like [Biomphalaria glabrata]	115	115	48%	3e-27	35.81%	XP_013084131.1
PREDICTED: cytoglobin-1-like [Callorhinchus milii]	112	112	54%	4e-26	35.71%	XP_007891388.1
PREDICTED: neuroglobin-like [Haplochromis burtoni]	111	111	52%	4e-26	37.89%	XP_005952972.2
PREDICTED: neuroglobin-like [Gekko japonicus]	112	112	48%	5e-26	37.58%	XP_015274271.1
neuroglobin isoform X2 [Oreochromis niloticus]	110	110	48%	9e-26	37.33%	XP_019201382.1
PREDICTED: neuroglobin-like [Neolamprologus brichardi]	110	110	48%	2e-25	37.33%	XP_006793470.1
PREDICTED: neuroglobin-like isoform X2 [Pundamilia nyererei]	110	110	48%	2e-25	37.33%	XP_005754405.1
neuroglobin-like isoform X1 [Erpetoichthys calabaricus]	110	110	48%	3e-25	36.49%	XP_028653231.1
GbX2 [Callorhinchus milii]	108	108	45%	3e-24	36.69%	AKU74647.1

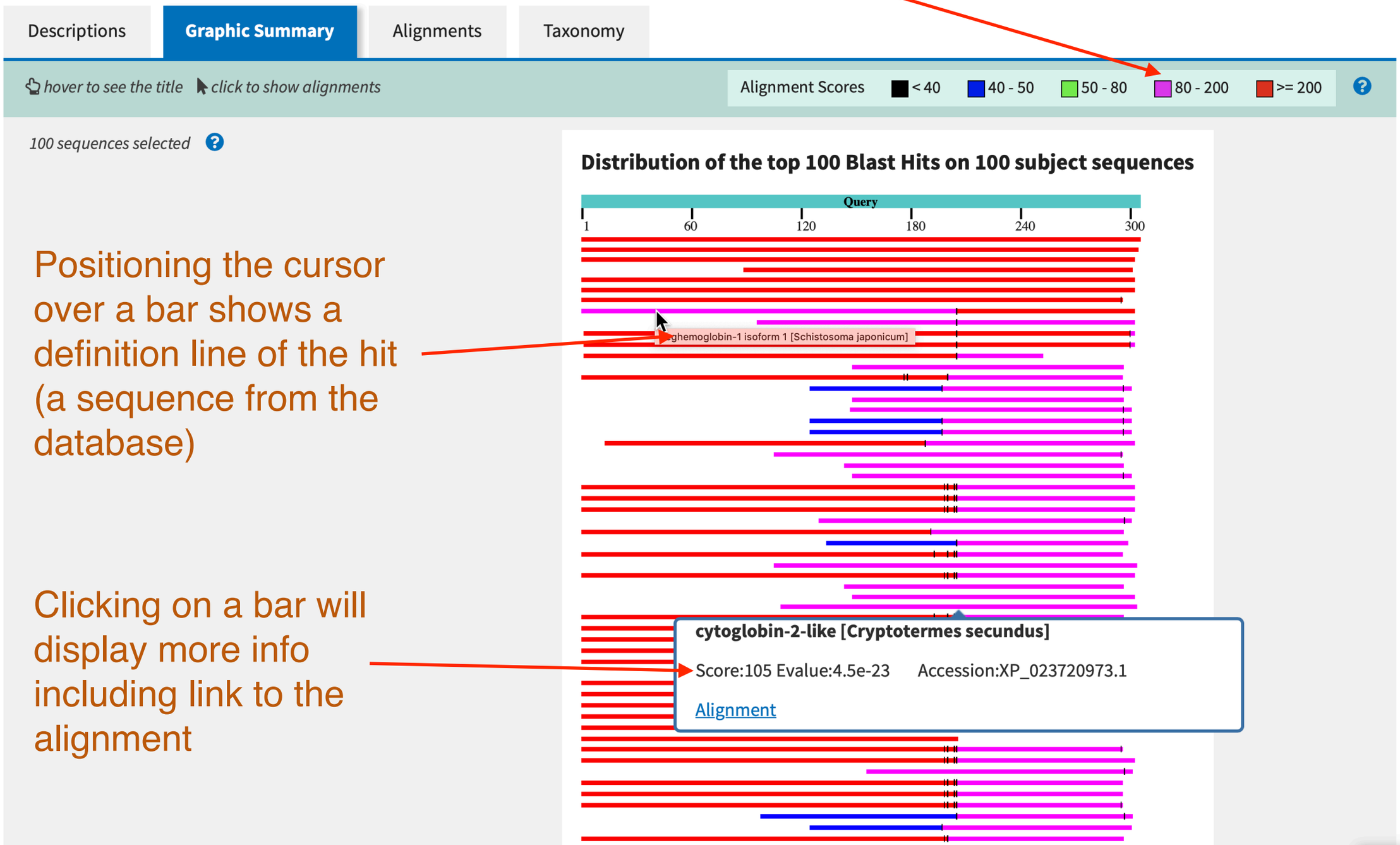
Definition from the source database

Select number of results to see.

Warning: if you want to see more than 100 results, you need to specify it on the first screen under “Algorithm parameters”

BLAST results: descriptions

Alignment score: higher is better 👍🏽



Positioning the cursor over a bar shows a definition line of the hit (a sequence from the database)

Clicking on a bar will display more info including link to the alignment

BLAST results: alignments

Clicking on Sequence ID will get you to the original record

Download GenPept Graphics ▼Next ▲Previous ◀◀Descriptions

hypothetical protein CRM22_002376 [Opisthorchis felineus]

Sequence ID: TGZ71927.1 Length: 303 Number of Matches: 1

Range 1: 45 to 299 GenPept Graphics ▼Next Match ▲Previous Match

Score	Expect	Method	Identities	Positives	Gaps
304 bits(779)	1e-99	Compositional matrix adjust.	148/262(56%)	190/262(72%)	7/262(2%)

```
Query  39   IEPDRETEEDNTSLSPDPNLQVQGNKILISIRKRMRRFLGAPTESSVLHKSMQDLSLDSA  98
            I PD     +NT ++ D N+  QG++IL+ +++ +RRF G    SS L KS QDL+L +
Sbjct  45   ISPDT----NNTQIAQD-NVSAQGSRILLCLKRHIRRFFGTNAGSSGLRKSQQDLTLANT  99

Query  99   GYEARKSSSTGNGMQNIASKMTRDSYITNDVPDDIQSIKREYEKALITLTSLSDGEIRAV  158
              RK+S T     N   K  +D  IT  VP+D++S K  Y  AL+ L SL+D ++  V
Sbjct  100  TVVERKASPTEKNSGNQFDK--KDISITCGVPEDVESFKTNYNAALLRLDSLTDAQVNGV  157

Query  159  RTSWMMLKTHIEKIGVIVFLGLFEEHSDFRDAFARFRGKQLMEITRDPALQAHGLRVLNI  218
            ++SWM+LK HIEKIGVIVFLGLFEEHSDFRDAFARFR KQL  +TRDPA QAHGLRVLN+
Sbjct  158  QSSWMILKIHIEKIGVIVFLGLFEEHSDFRDAFARFRQKQLSNLTRDPAFQAHGLRVLNV  217

Query  219  VDKLVSRLQKVETIQDFILSLGCRHCKYVPSIKLIPCVGEQLLEAFHPVLEEQGVWTKDT  278
            VDK++SRL +++TIQDF+LSLG +HC+YVP+I+L+P VGEQLLEA  PVLEEQG+W  DT
Sbjct  218  VDKIISRLHRIDTIQDFLLSLGSKHCRYVPNIELVPAVGEQLLEAVRPVLEEQGLWDDDT  277

Query  279  ETGWTILLDFLTKAMRYGLART  300
               GW  +L +L  AMRYGL R+
Sbjct  278  AVGWDAVLAYLNCAMRYGLVRS  299
```

A little bit of alignment statistics

The middle line shows matches and mismatches. The mismatches with a positive score are shown as “+” and mismatches with the negative scores are shown as blanks.

BLAST results: alignments

Descriptions | Graphic Summary | Alignments | **Taxonomy**

Reports | **Lineage** | **Organism** | **Taxonomy**

Different levels of taxonomy

100 sequences selected

Organism	Blast Name	Score	Number of Hits	Description
Bilateria	animals		122	
. Digenea	flatworms		16	
. . Echinostomatoidea	flatworms		3	
. . . Fasciola	flatworms		2	
. . . . Fasciola hepatica	flatworms	635	1	Fasciola hepatica hits
. . . . Fasciola gigantica	flatworms	614	1	Fasciola gigantica hits
. . . Echinostoma caproni	flatworms	297	1	Echinostoma caproni hits
. . Opisthorchis felineus	flatworms	304	1	Opisthorchis felineus hits
. . Opisthorchis viverrini	flatworms	300	3	Opisthorchis viverrini hits
. . Clonorchis sinensis	flatworms	298	2	Clonorchis sinensis hits
. . Schistosoma japonicum	flatworms	218	2	Schistosoma japonicum hits
. . Schistosoma margrebowiei	flatworms	205	1	Schistosoma margrebowiei hits
. . Schistosoma bovis	flatworms	201	1	Schistosoma bovis hits
. . Schistosoma mattheei	flatworms	168	1	Schistosoma mattheei hits
. . Schistosoma mansoni	flatworms	103	2	Schistosoma mansoni hits
. Biomphalaria glabrata	gastropods	115	1	Biomphalaria glabrata hits
. Callorhinchus milii	chimaeras	112	3	Callorhinchus milii hits

Clicking on organism name will take you to NCBI taxonomy browser

Clicking here will take you to the list of hits sorted by taxonomy

Tons of materials to learn from 👻

Learn

NCBI creates a variety of educational products including courses, workshops, webinars, training materials and documentation. NCBI educational events are free and open to everyone. All NCBI educational materials are available for anyone to re-use and distribute.

Webinars & Courses

In-person courses, live webinars and webinar recordings

Conferences & Presentations

Booth exhibits and workshops at scientific conferences

Tutorials

Tutorials: Training materials in HTML, PDF and video formats

Documentation

Online manuals, handbooks, fact sheets and FAQs

Follow Us

NCBI News & Blog

New human genome annotation release with MANE Select and other improvements!

03 Jul 2019

There's a new RefSeq annotation available for the human genome, and it's quite an update! About the release Annotation release 109.20190607 is the first release of our new

Microbial Virulence in the Cloud hackathon August 13 – 15 2019

02 Jul 2019

From August 13 – 15 2019, the NCBI will run a bioinformatics hackathon on the NIH campus! We're specifically looking for folks who have experience in working with computational

GenBank release 232

01 Jul 2019

GenBank release 232.0 (6/20/2019) is now

http://bioinformatics.uni-muenster.de