

Homología, Evolución, y Bases de Datos

Bioinformática para biotecnología BIT120

10 marzo 2017

Eduardo Castro, PhD

www.castrolab.org

El equipo

- Dr. Fernando Danilo González
(fernando.gonzalez@unab.cl)
- Dr. Eduardo Castro (eduardo.castro@unab.cl)
- Jonathan Canan (jonathancanan@gmail.com)
- Katterinne Méndez (mendez.katterinne@gmail.com)
- Javier Cáceres (ja.caceresmolina@gmail.com)
- Ignacio Ramos (ramostapiai@gmail.com)

Evaluaciones

- La clase se divide en genómica y modelamiento de proteínas (6+6)
- 3 pruebas (60%) & controles e informes (40%) + examen
- Programa <http://tinyurl.com/z5y8hyr>

El programa

Fecha	Planificación de actividades
10 de Marzo	Homología y Evolución + Bases de Datos Biológicas y de Literatura + Búsqueda en Bases de Datos
17 de Marzo	Alineamiento de Pares de Secuencias, Múltiple y Perfiles (HMM's) + Diseño de Partidores + BLAST
24 de Marzo	Ensamblaje de Genomas + Predicción de Genes
31 de Marzo	Modelos de Sustitución Nucleotídica y Proteica + Filogenética Molecular
7 de Abril	Solemne I
21 de Abril	Metagenómica, Metatranscriptómica y Microbioma humano
28 de Abril	Diseño de vectores de clonamiento
5 de Mayo	Visualización, Comparación y Clasificación de Estructura de Proteínas
12 de Mayo	Predicción de Estructura Secundaria y Terciaria de Proteínas
19 de Mayo	Solemne II
26 de Mayo	Búsqueda de proteínas homólogas + Redes de similitud + Modelado por homología
2 de Junio	Bioinformática de enzimas
9 de Junio	Introducción a la Simulación Molecular
16 de Junio	Aplicación de la Simulación Molecular en ingeniería de proteínas
23 de Junio	Solemne III
Última semana de junio según coordinación UNAB	Examen

Bioinformática y Biología Computacional

- **Bioinformática** → organización, almacenaje, clasificación de información biológica. Desarrollo de métodos, algoritmos, y software para analizar información biológica.
- **Biología Computacional** → Aplicación de métodos analíticos y teóricos, modelamiento matemático y simulación computacional para el estudio de sistemas biológicos.



So you want to be a computational biologist?

Nick Loman & Mick Watson

Loman, N., & Watson, M. (2013). So you want to be a computational biologist?. *Nature biotechnology*, 31(11), 996-998.

Propaganda del lab

THE CASTRO LAB

HOME

RESEARCH

PEOPLE

PUBLICATIONS

OPPORTUNITIES

CONTACT

MICROBIAL GENOMICS



Eduardo Castro-Nallar, PhD [✉](mailto:eduardo.castro-nallar@unab.cl) [ORCID](https://orcid.org/0000-0002-1111-1111)
Universidad Andrés Bello
[Center for Bioinformatics and Integrative Biology](#)

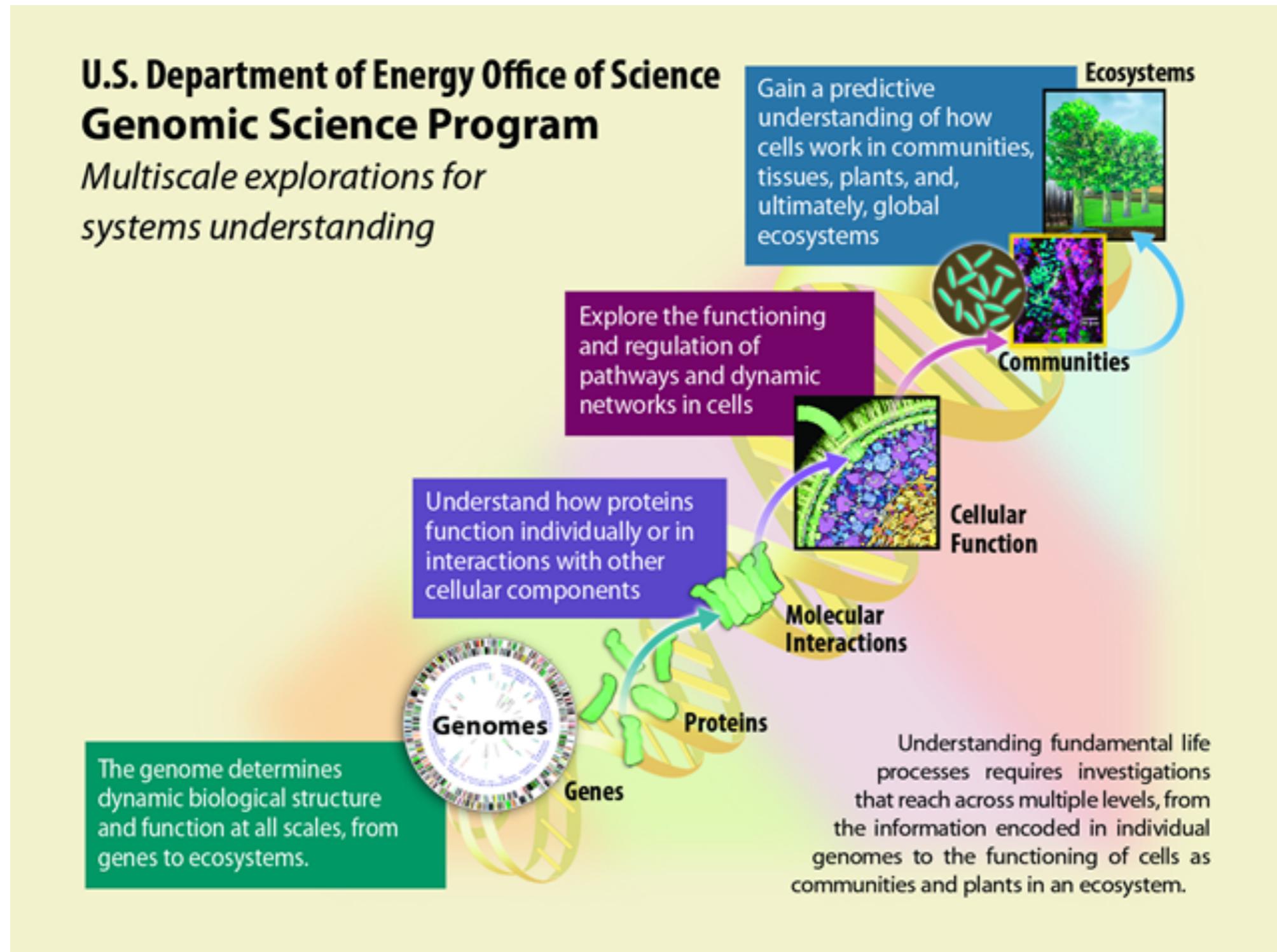


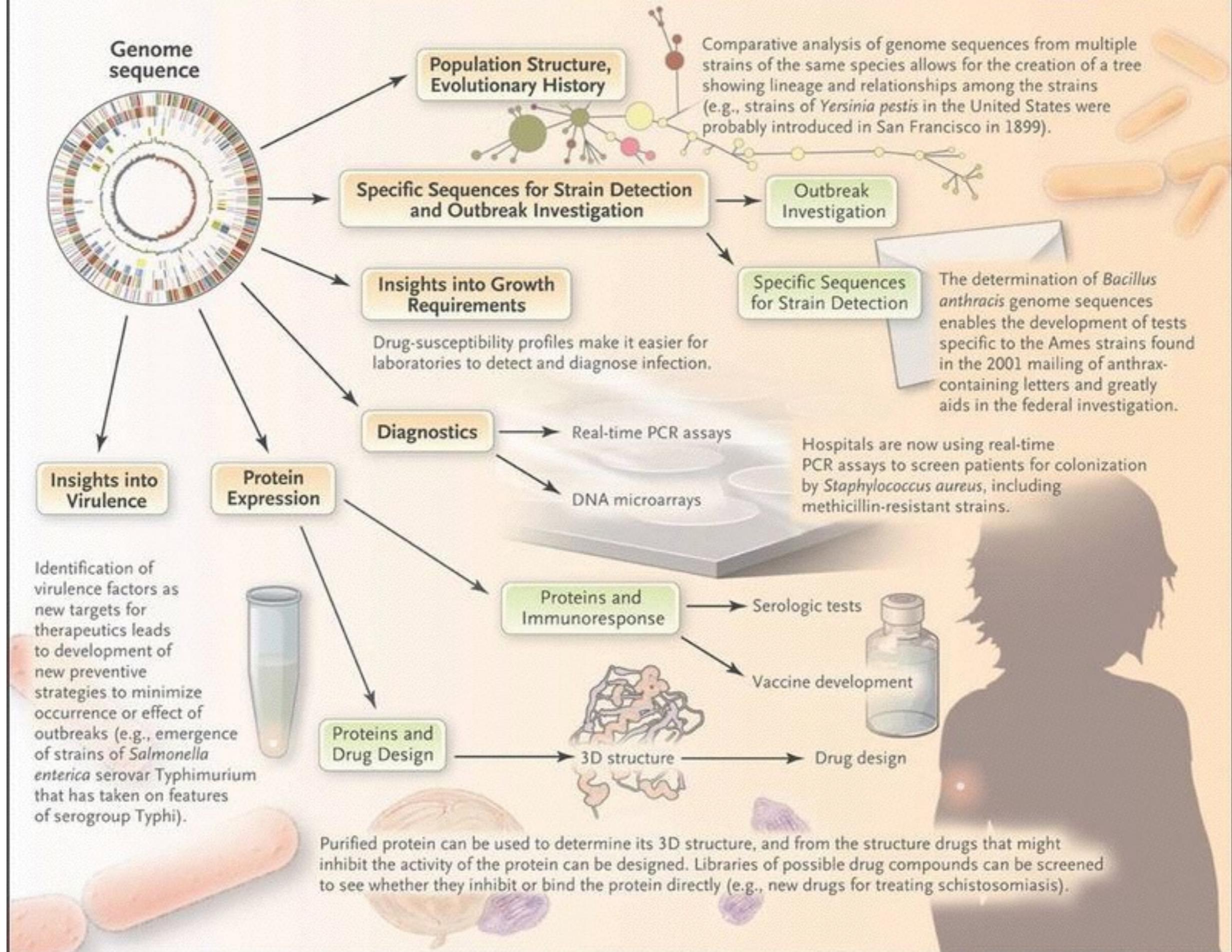
Microorganisms are the most abundant life form on the planet. They are able to colonize and thrive in most environments ranging from the human body to under water thermal vents. In the Castro Lab, we are interested in understanding the causes and consequences of microbial genetic diversity, and how we can apply this knowledge to gain insights into pathogen diagnostics and discovery, microbial distribution and epidemiology, and novel adaptations.

For this, we use molecular and computational biology tools such as high-throughput sequencing, recombinant DNA technology, Bayesian statistics and phylogenetics, transcriptomics and metagenomics.

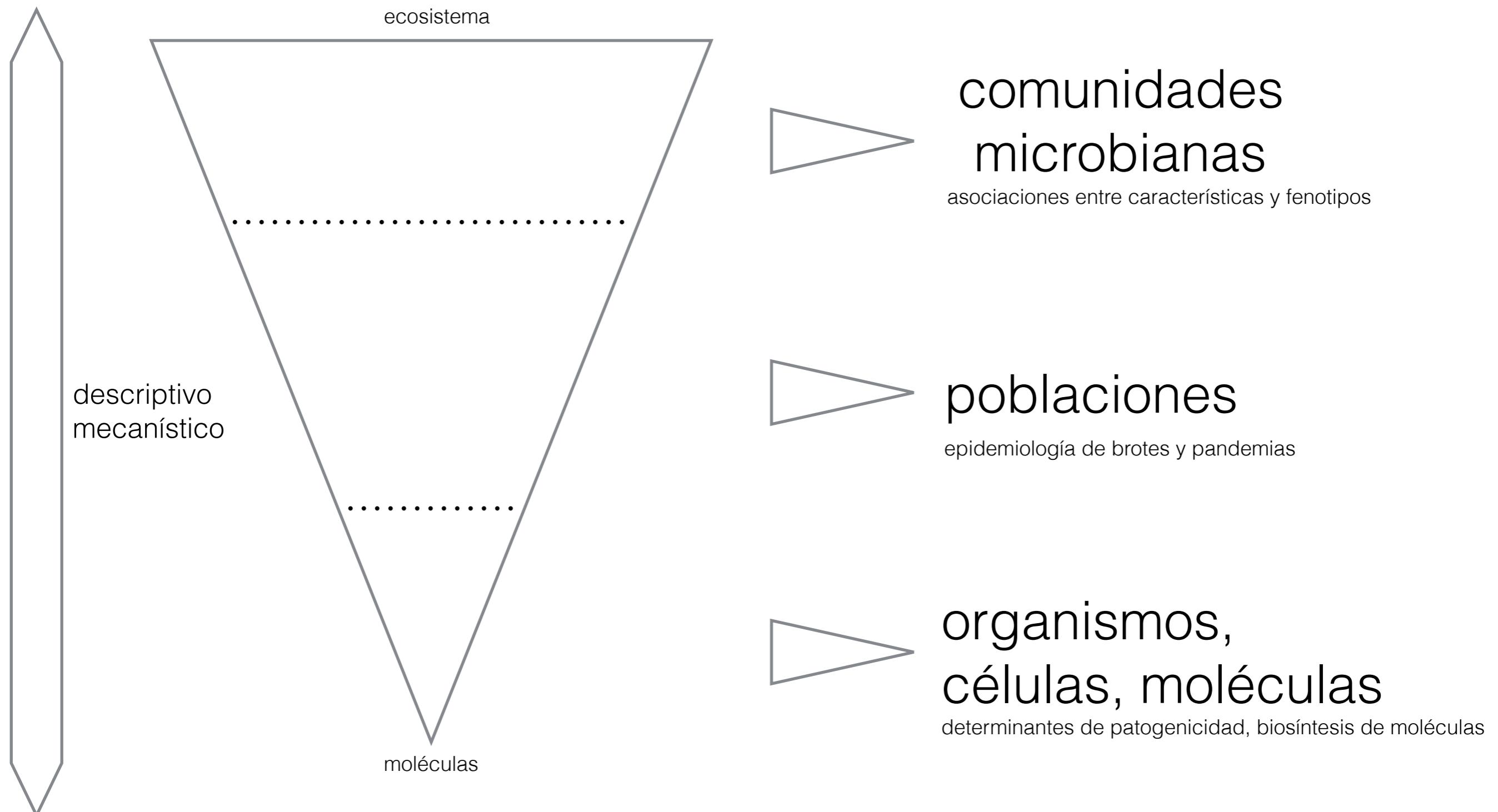
www.castrolab.org

De genes a ecosistemas

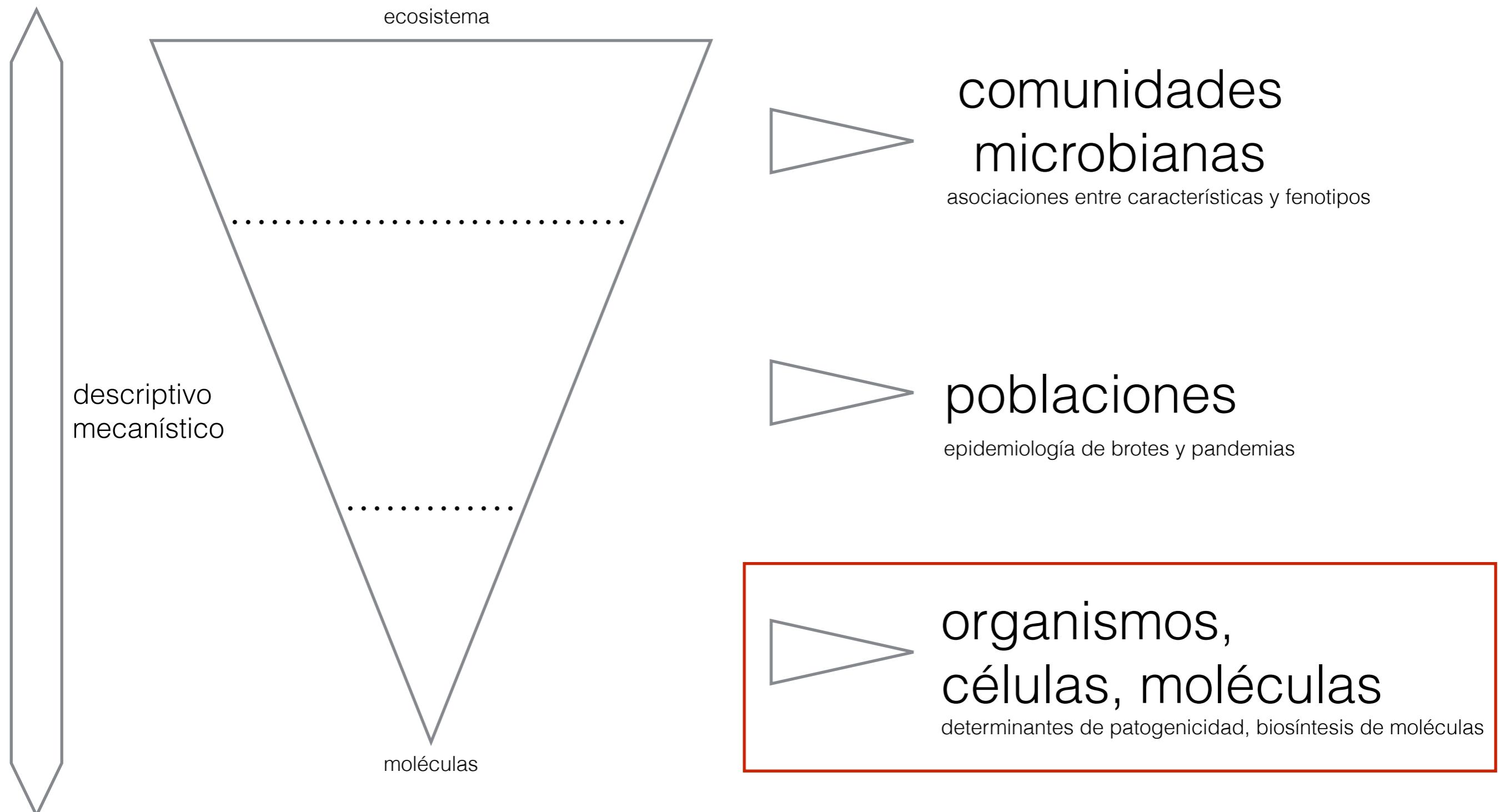




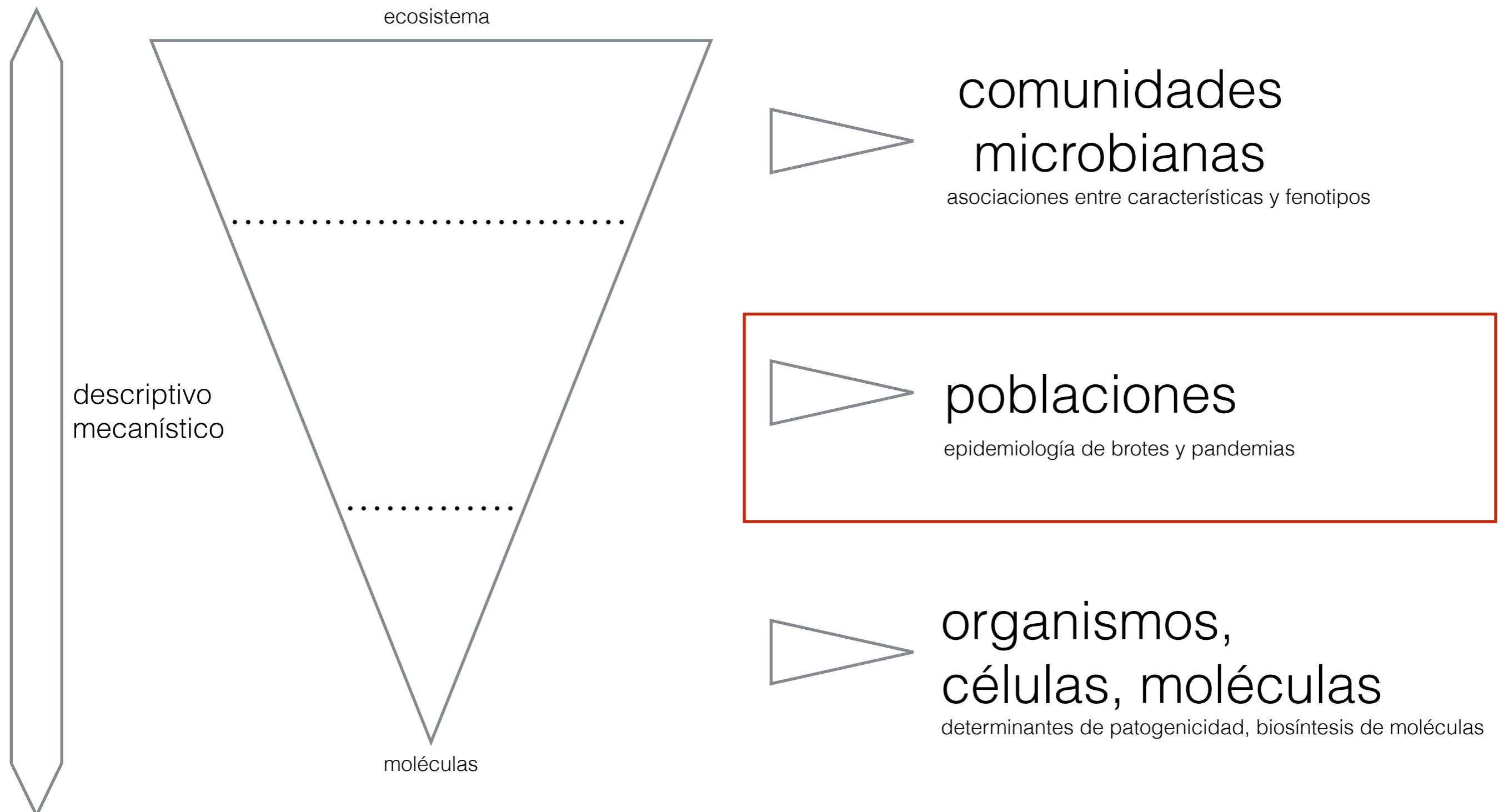
¿Dónde encaja la genómica de microorganismos?



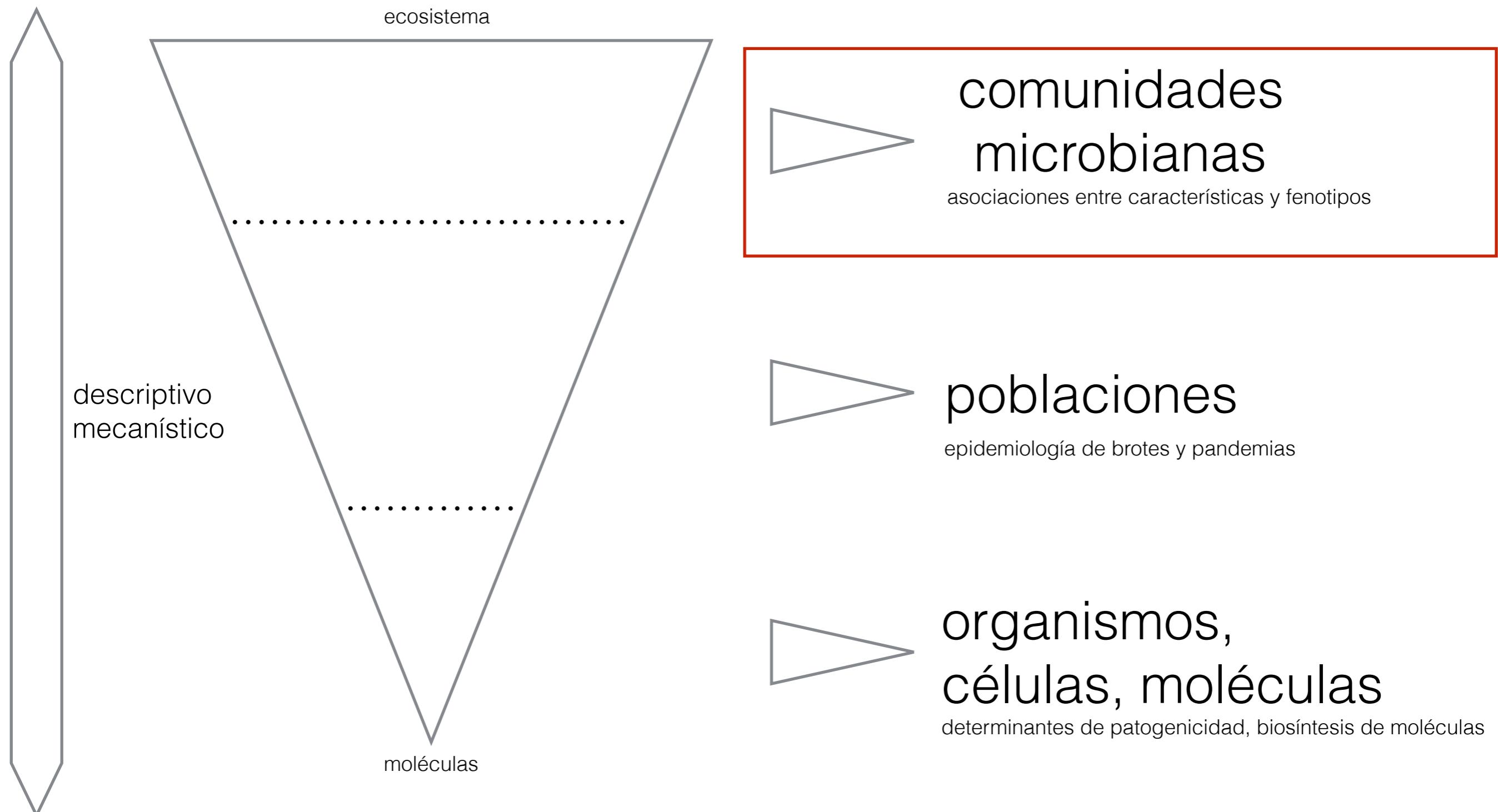
¿Dónde encaja la genómica de microorganismos?



¿Dónde encaja la genómica de microorganismos?

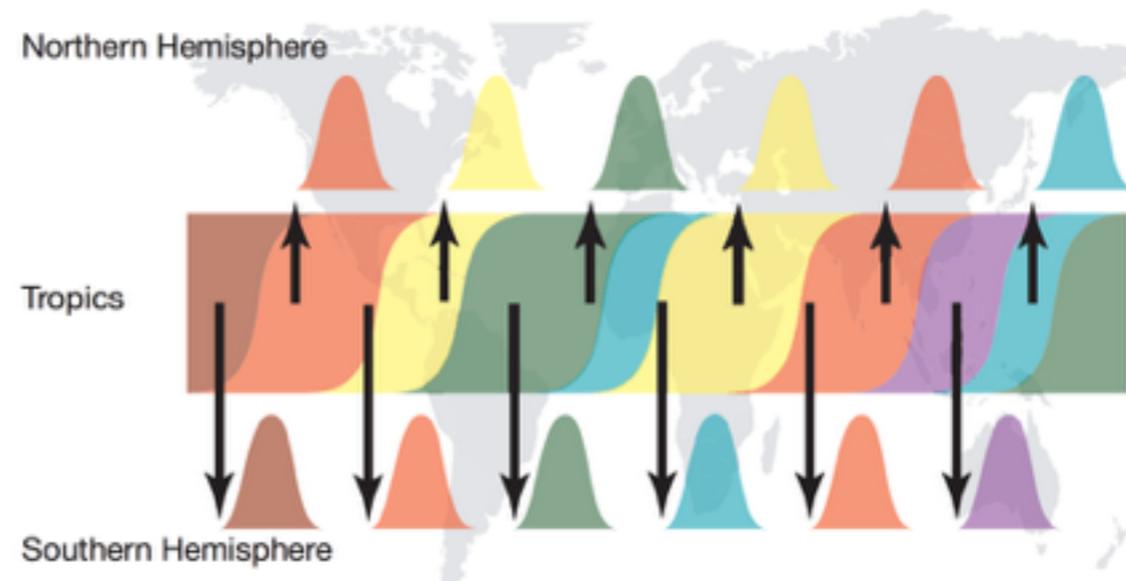
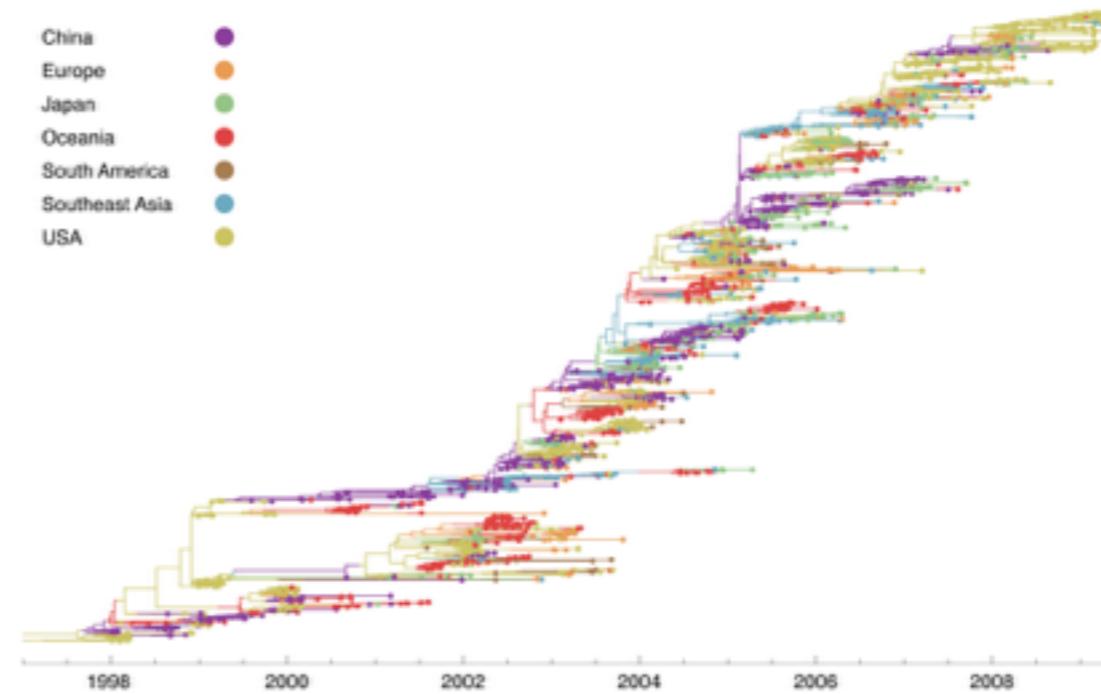


¿Dónde encaja la genómica de microorganismos?



Impacto de la genómica de microorganismos en salud humana y medioambiente

- Epidemiología
- Microbiomas
- Bioprospección



T Bedford, S Cobey, P Beerli, M Pascual, N Ferguson. 2010. Global migration dynamics underlie evolution and persistence of human influenza A (H3N2). PLoS Pathogens. e1000918-e1000918

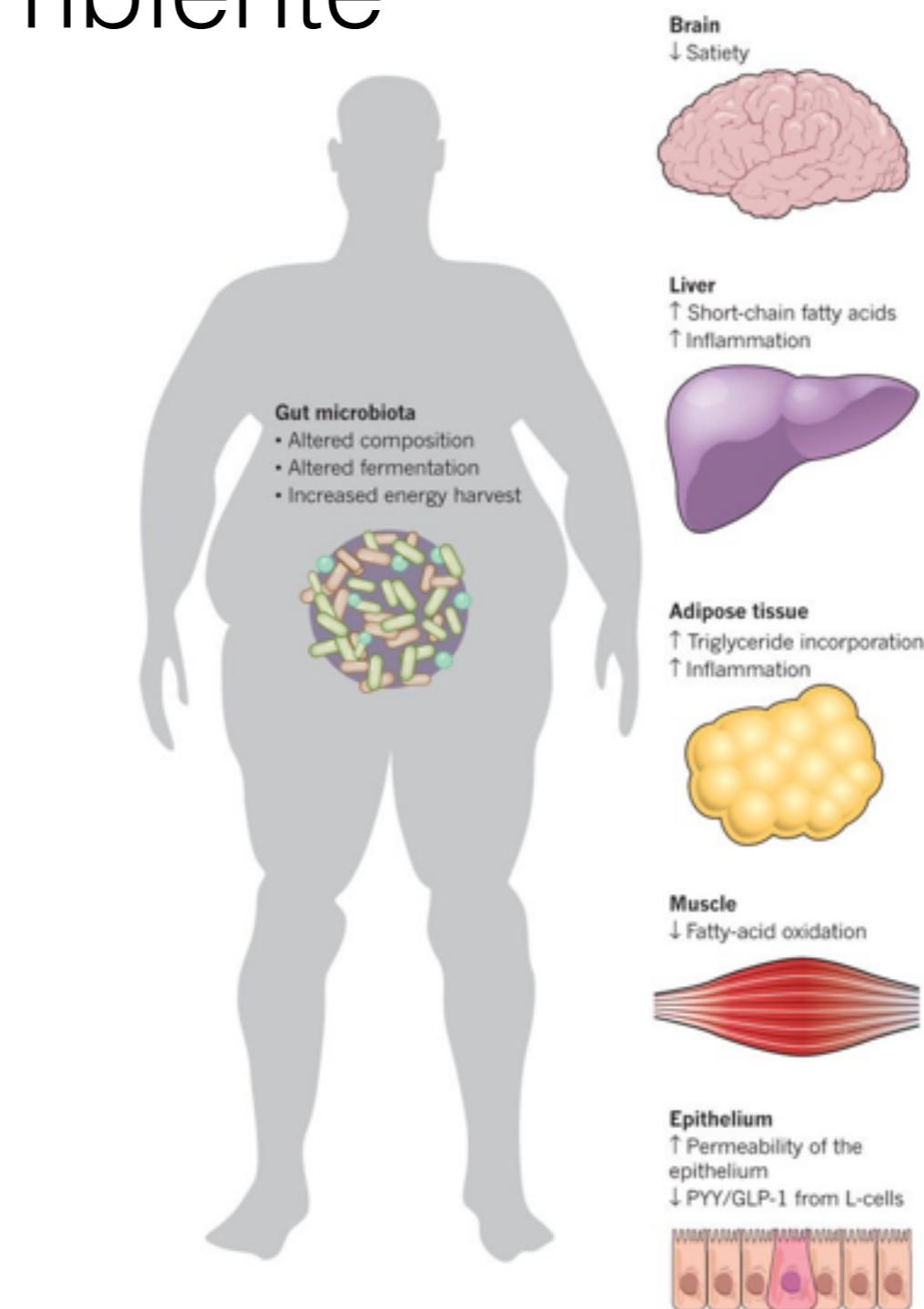
Rambaut, A., Pybus, O. G., Nelson, M. I., Viboud, C., Taubenberger, J. K., & Holmes, E. C. (2008). The genomic and epidemiological dynamics of human influenza A virus. Nature, 453(7195), 615-619.

Impacto de la genómica de microorganismos en salud humana y medioambiente

- Epidemiología
- Microbiomas
- Bioprospección

Impacto de la genómica de microorganismos en salud humana y medioambiente

- Epidemiología
- Microbiomas
- Bioprospección

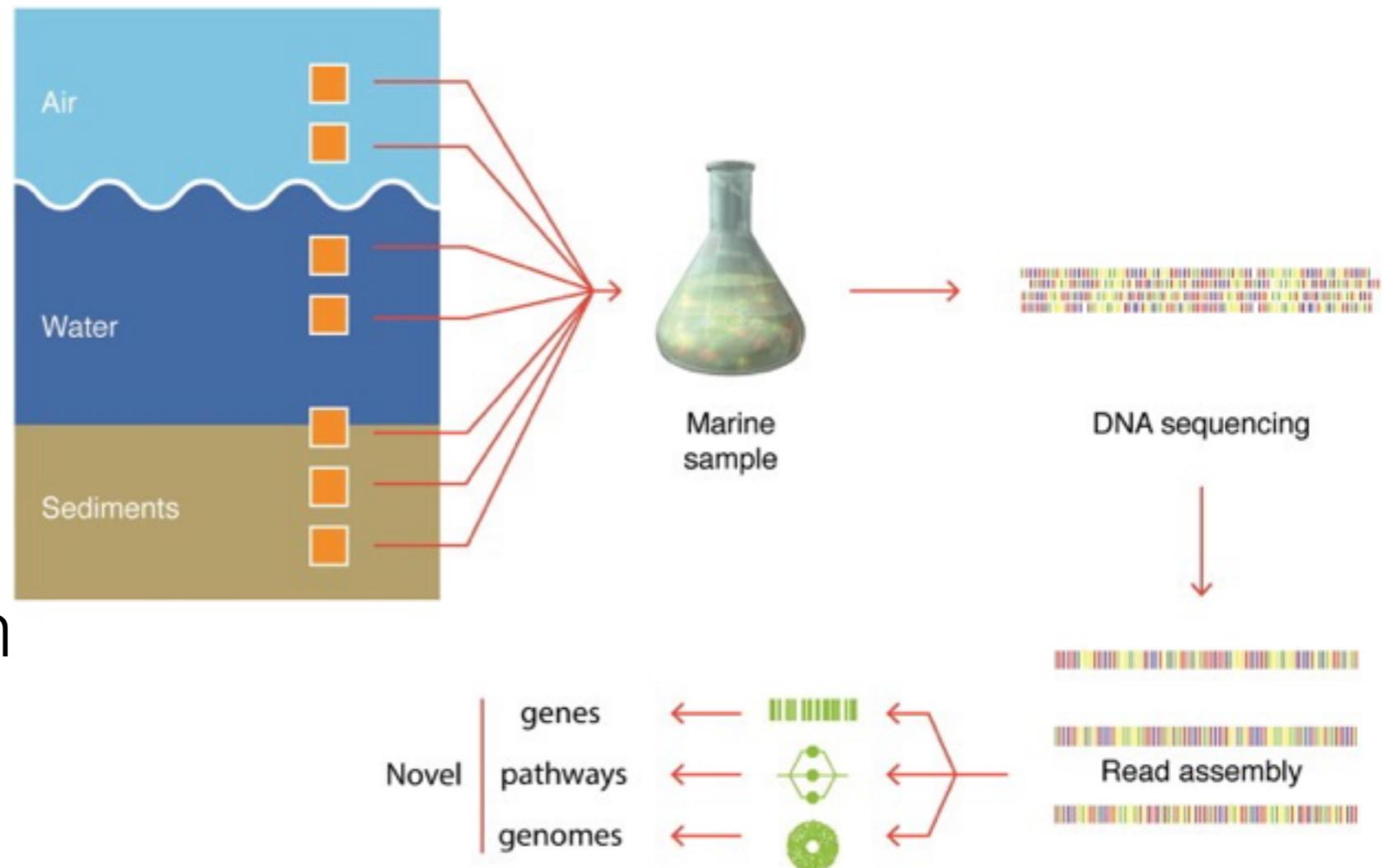


Impacto de la genómica de microorganismos en salud humana y medioambiente

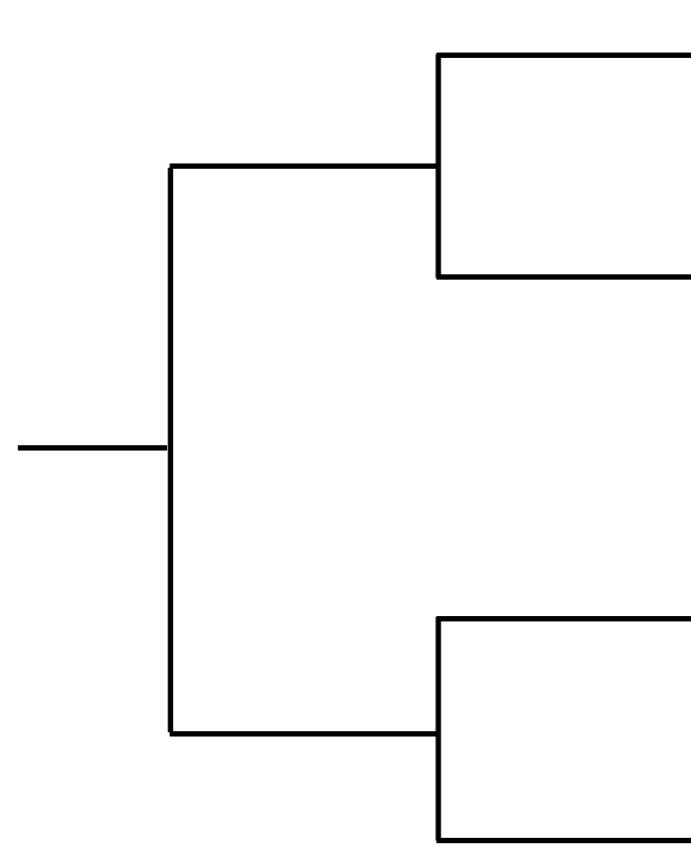
- Epidemiología
- Microbiomas
- Bioprospección

Impacto de la genómica de microorganismos en salud humana y medioambiente

- Epidemiología
- Microbiomas
- Bioprospección

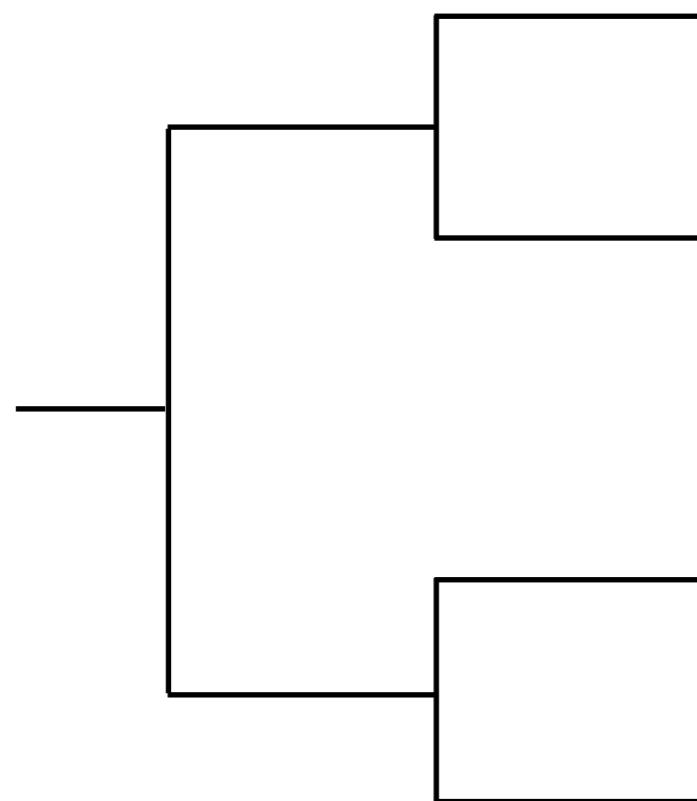


Pilares principales del lab

- 
- Generación de datos genómicos
 - Desarrollo de infraestructura informática y métodos
 - Epidemiología molecular
 - Ecología microbiana

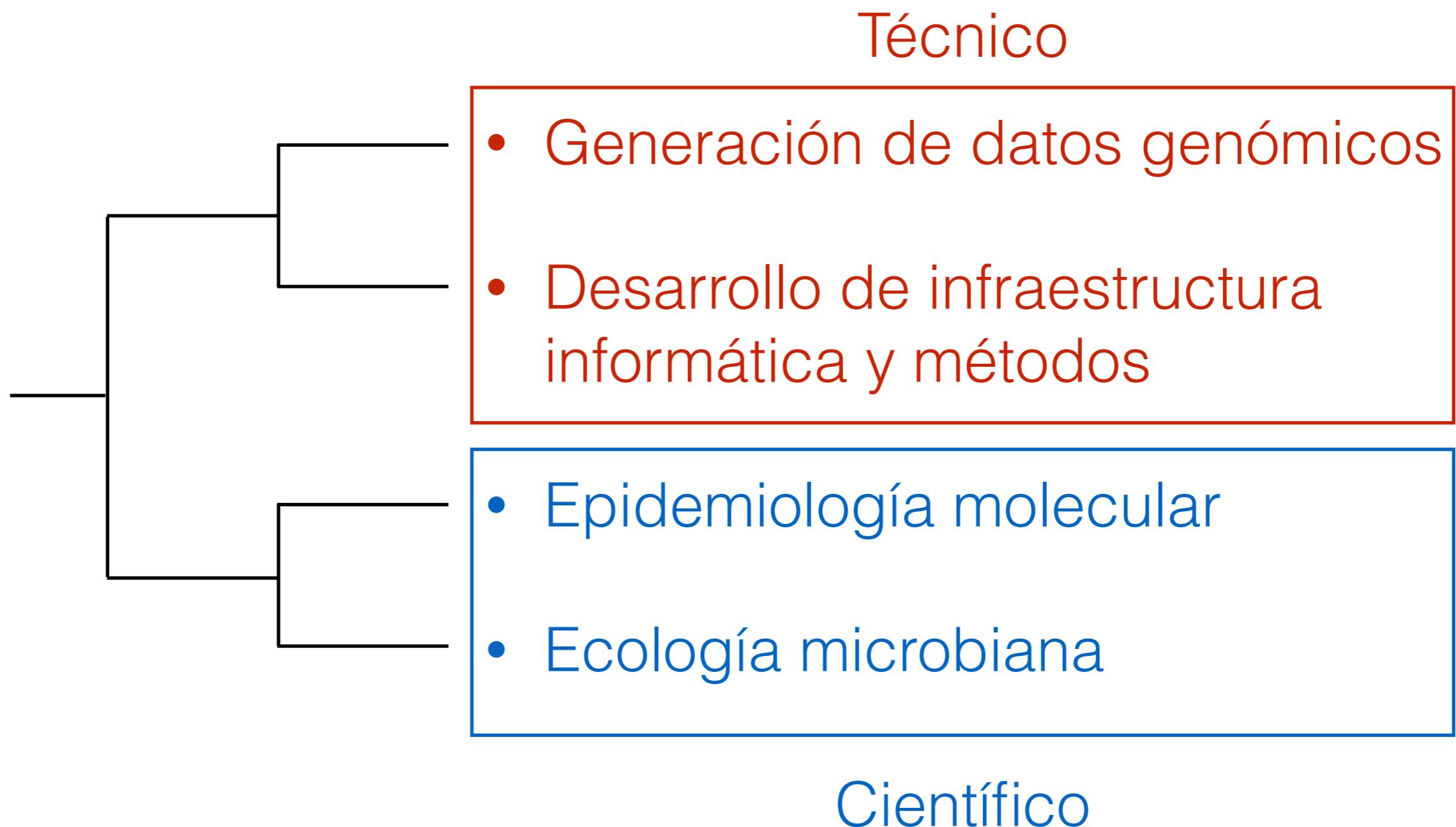
Pilares principales del lab

Técnico



- Generación de datos genómicos
- Desarrollo de infraestructura informática y métodos
- Epidemiología molecular
- Ecología microbiana

Pilares principales del lab



Ahora, la clase...

Objetivos

- Evolución y unidad de las formas de vida
- Homología, paralogía, xenología, etc.
- Organización del conocimiento científico y biológico
- Práctica: uso de bases de datos

Evolución y unidad de las formas de vida

Observación

- Alta biodiversidad
 - Sin embargo todos exhiben elementos que los unifican → organismos comparten ancestros comunes



How to Read the Circle of Life

Primordial life begins at the center and branches out in all directions, leading to the groups of species that exist today (colored rings)

Outer ring: Estimated proportion of all species*

Inner ring: Proportion of the groups named to date

Each black line represents at least 500 descendant species

Dark lines: Many species have been genetically sequenced

Light lines: Few species have been genetically sequenced

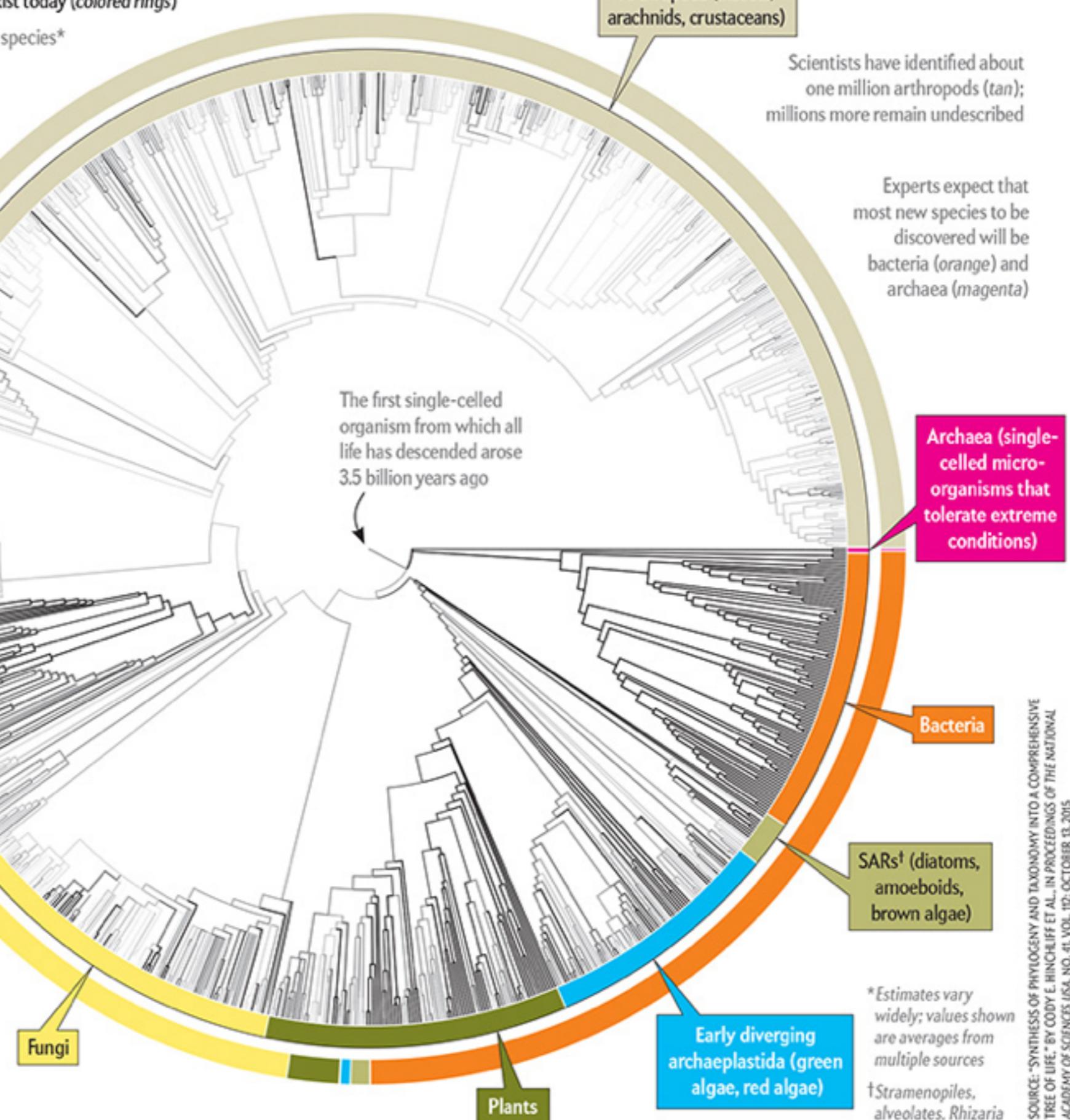
Nematodes (roundworms)

Lophotrochozoa (mollusks, segmented worms, brachiopods)

Deuterostomia (vertebrates, sea stars and urchins, certain worms)

Early diverging metazoa (cnidaria, comb jellies, sponges)

Many deuterostomia (gold) and plants (dark green) are already genetically sequenced (dark lines) because they are culturally or economically important (such as humans!)



Evidencia sobre unidad de la vida

- Código genético
- Registro fósil
- Función y estructura celular —> linea germinal
- Rasgos vestigiales
- Distribución de especies relacionadas

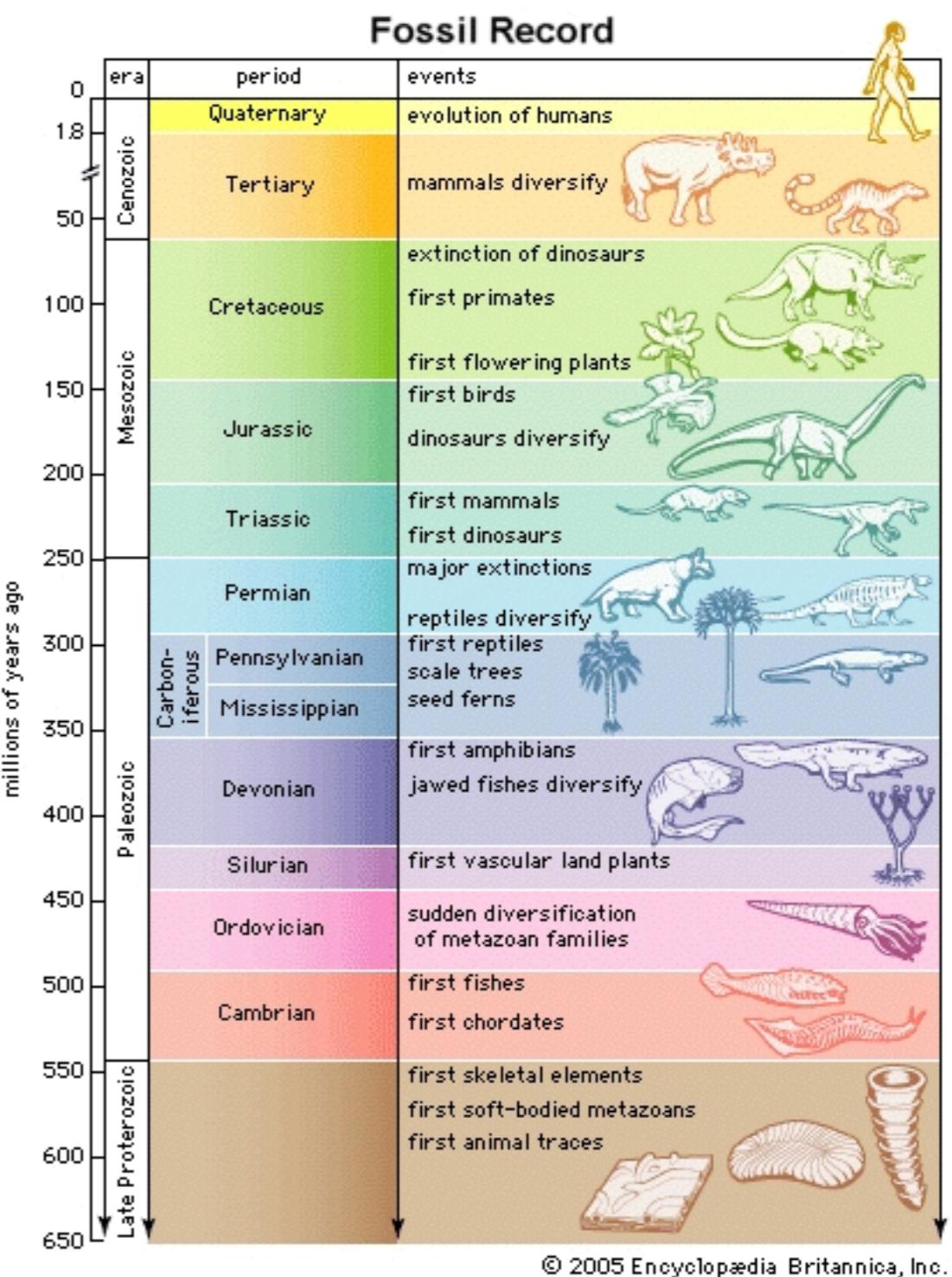
UUU [F] Phe	UCU [S] Ser	UAU [Y] Tyr	UGU [C] Cys
UUC [F] Phe	UCC [S] Ser	UAC [Y] Tyr	UGC [C] Cys
UUA [L] Leu	UCA [S] Ser	UAA [] Ter	UGA [] Ter
UUG [L] Leu	UCG [S] Ser	UAG [] Ter	UGG [W] Trp
CUU [L] Leu	CCU [P] Pro	CAU [H] His	CGU [R] Arg
CUC [L] Leu	CCC [P] Pro	CAC [H] His	CGC [R] Arg
CUA [L] Leu	CCA [P] Pro	CAA [Q] Gln	CGA [R] Arg
CUG [L] Leu	CCG [P] Pro	CAG [Q] Gln	CGG [R] Arg
AUU [I] Ile	ACU [T] Thr	AAU [N] Asn	AGU [S] Ser
AUC [I] Ile	ACC [T] Thr	AAC [N] Asn	AGC [S] Ser
AUA [I] Ile	ACA [T] Thr	AAA [K] Lys	AGA [R] Arg
AUG [M] Met	ACG [T] Thr	AAG [K] Lys	AGG [R] Arg
GUU [V] Val	GCU [A] Ala	GAU [D] Asp	GGU [G] Gly
GUC [V] Val	GCC [A] Ala	GAC [D] Asp	GGC [G] Gly
GUA [V] Val	GCA [A] Ala	GAA [E] Glu	GGA [G] Gly
GUG [V] Val	GCG [A] Ala	GAG [E] Glu	GGG [G] Gly

Código genético

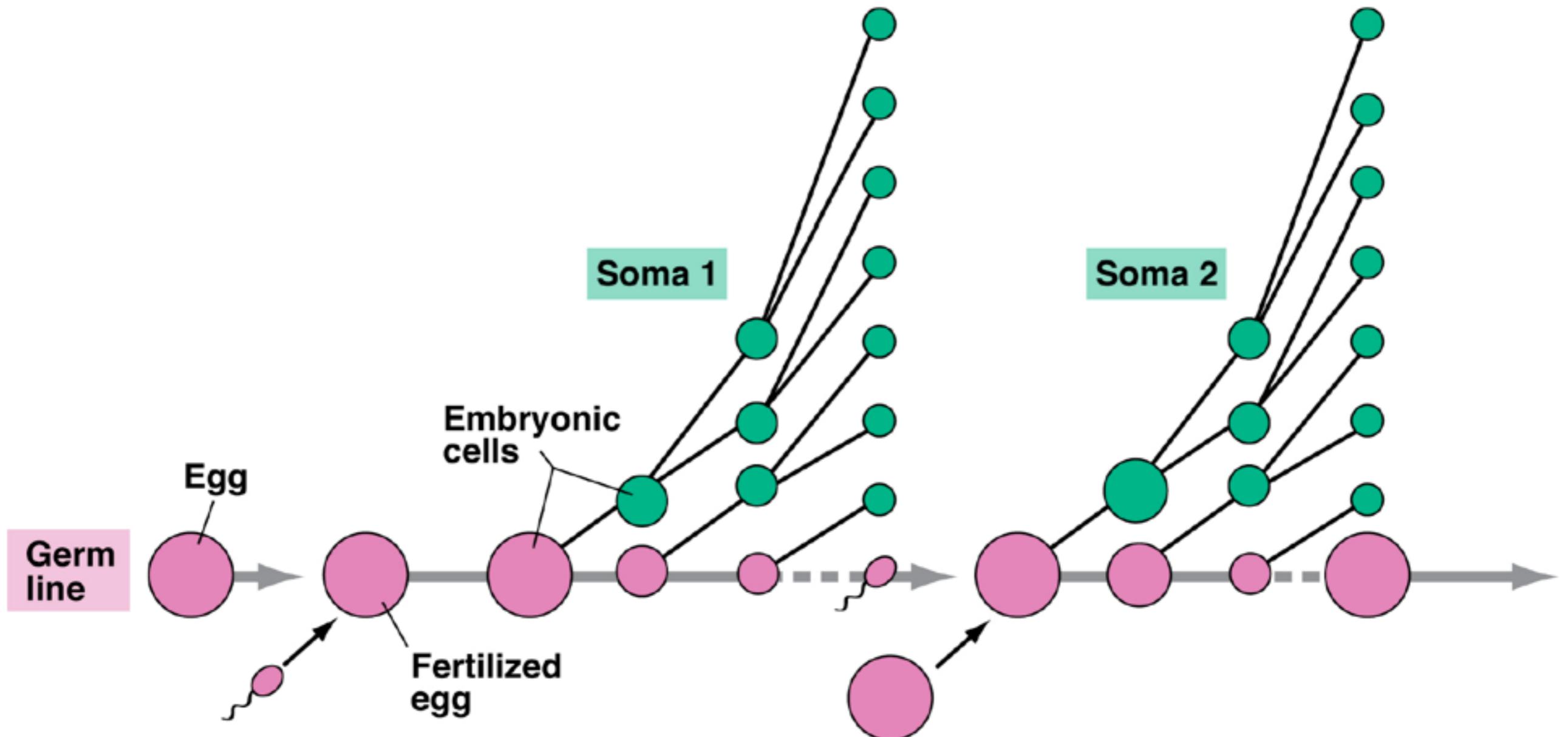
- Sistema común para todas las especies
 - Codones → aa
 - DNA para almacenar información genética
 - Síntesis de proteínas a través de ribosomas
 - Los mismos 20 aminoácidos

Registro fósil

- Registra la historia de la vida a través del tiempo
- 4 mil millones de años
- Documentan relaciones entre ancestros y descendientes

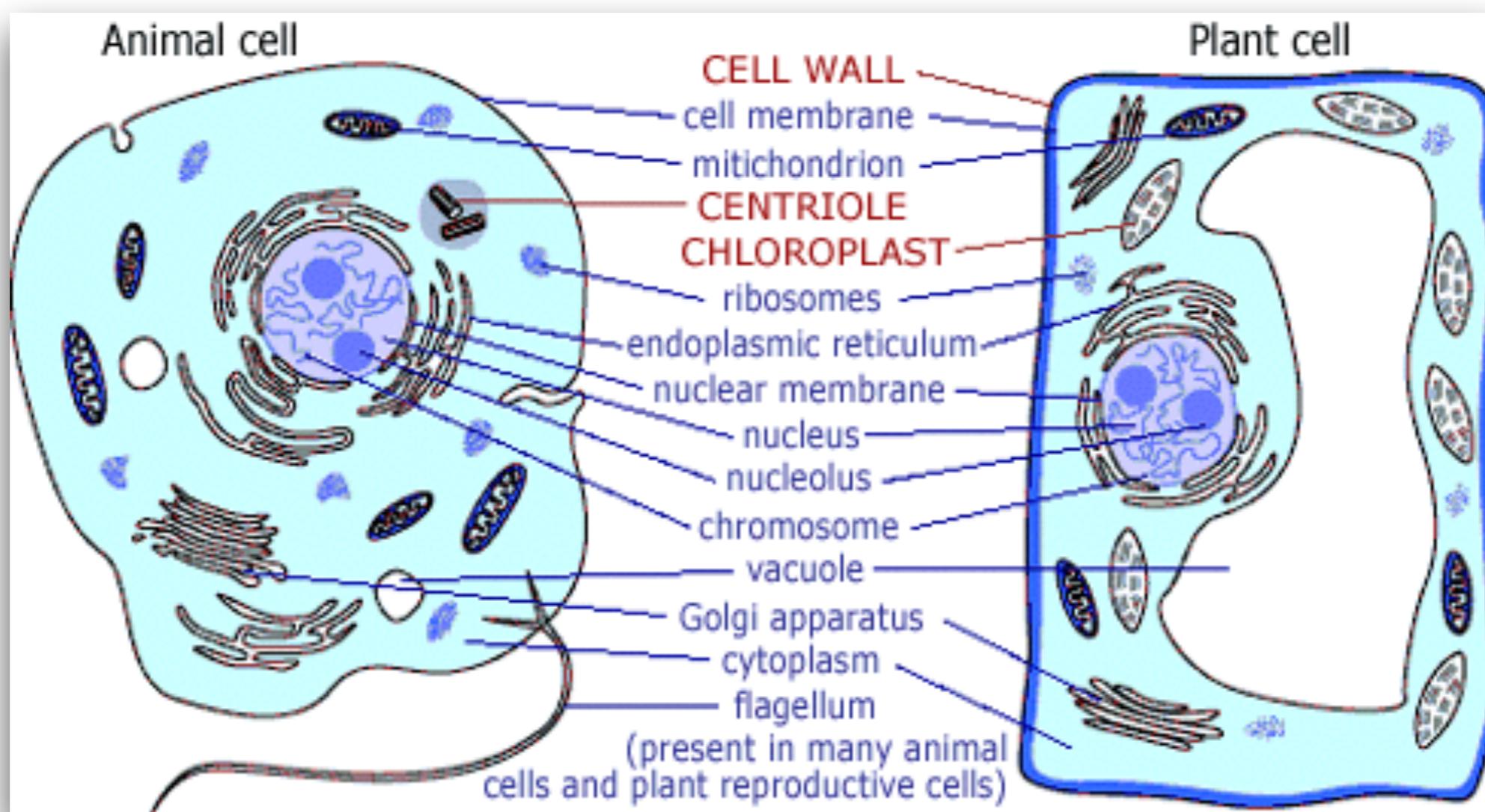


Linea germinal



La linea germinal forma un continuo a través de individuos y generaciones

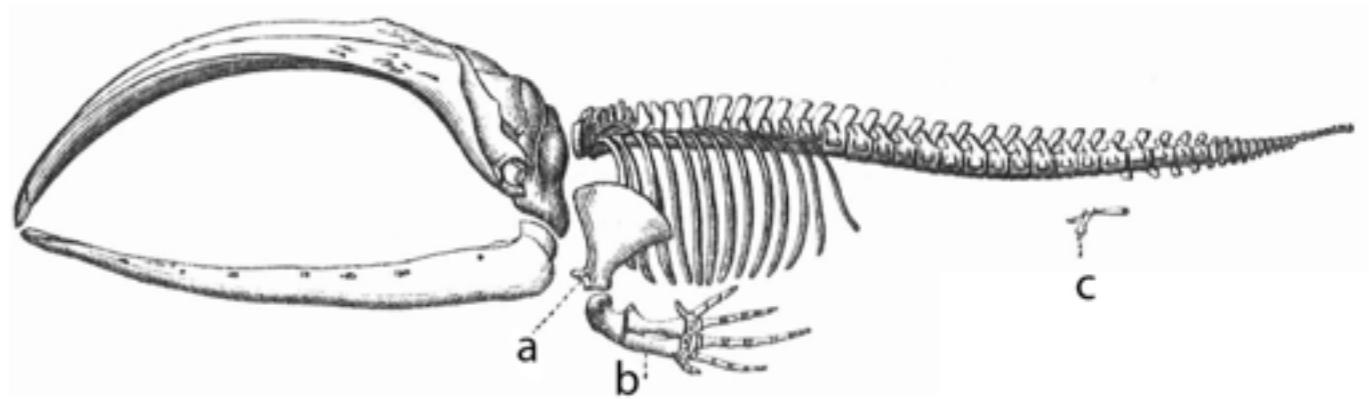
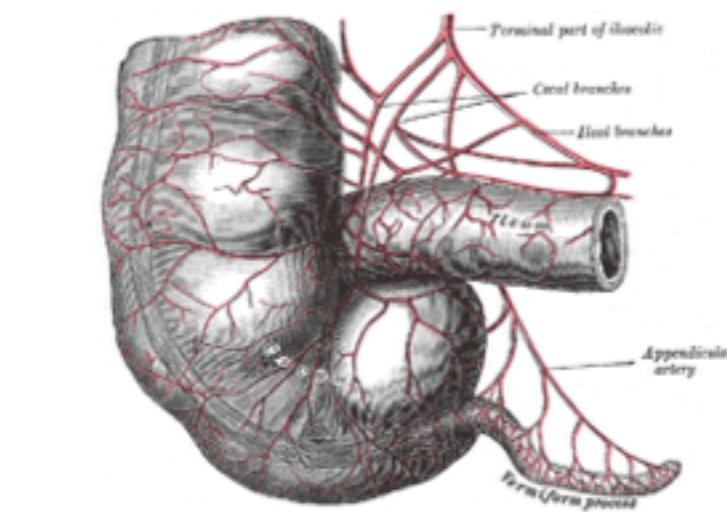
Estructura y función



Células de distintos organismos poseen estructuras funcionales similares

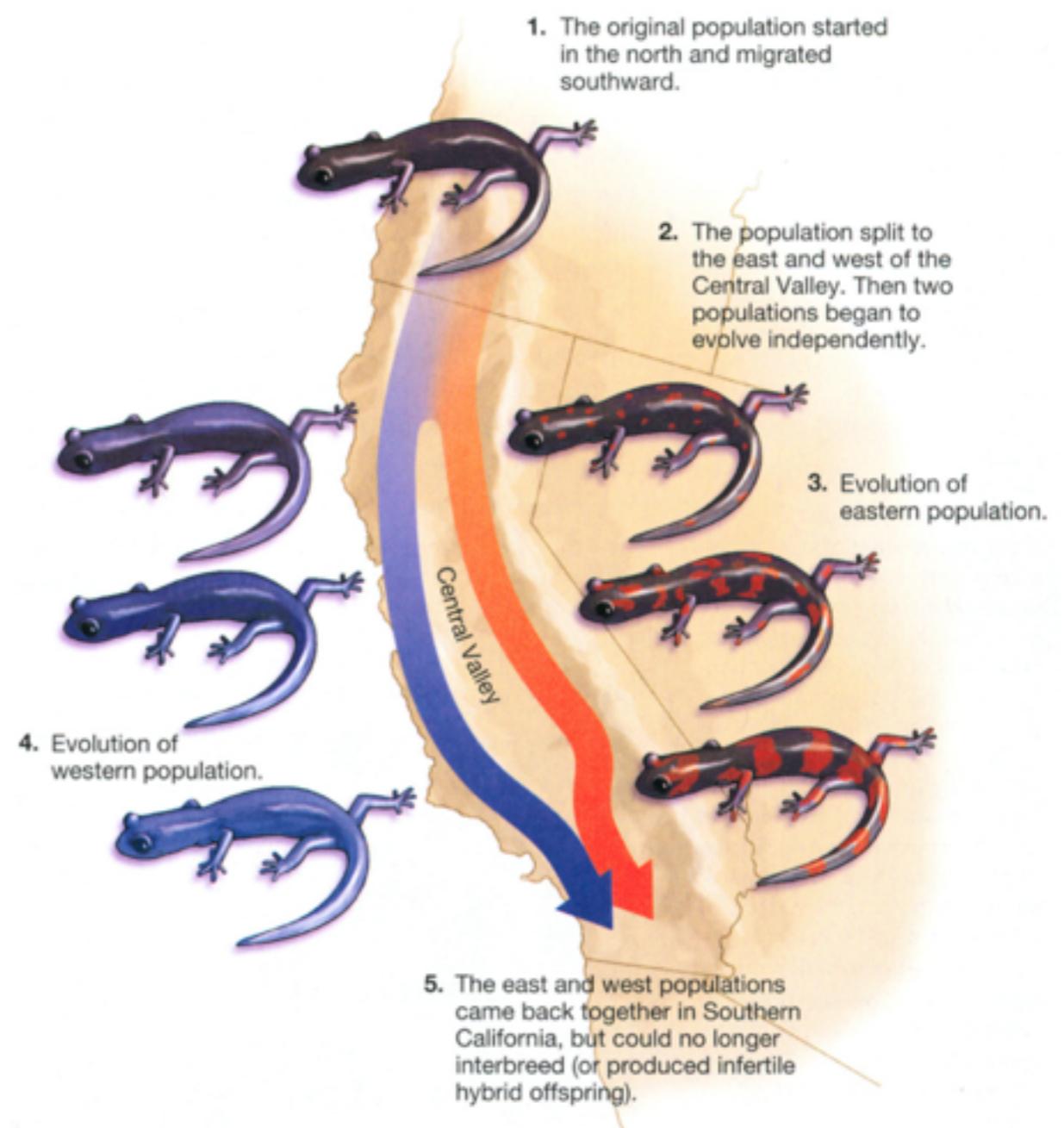
Rasgos vestigiales

- Estructuras determinadas genéticamente que han sido perdido su función ancestral en una especie pero han sido retenidas durante el proceso evolutivo
- Apéndice, cola, reflejo ante frio
- Patas traseras en boa constrictor, ballena



Distribución de especies relacionadas

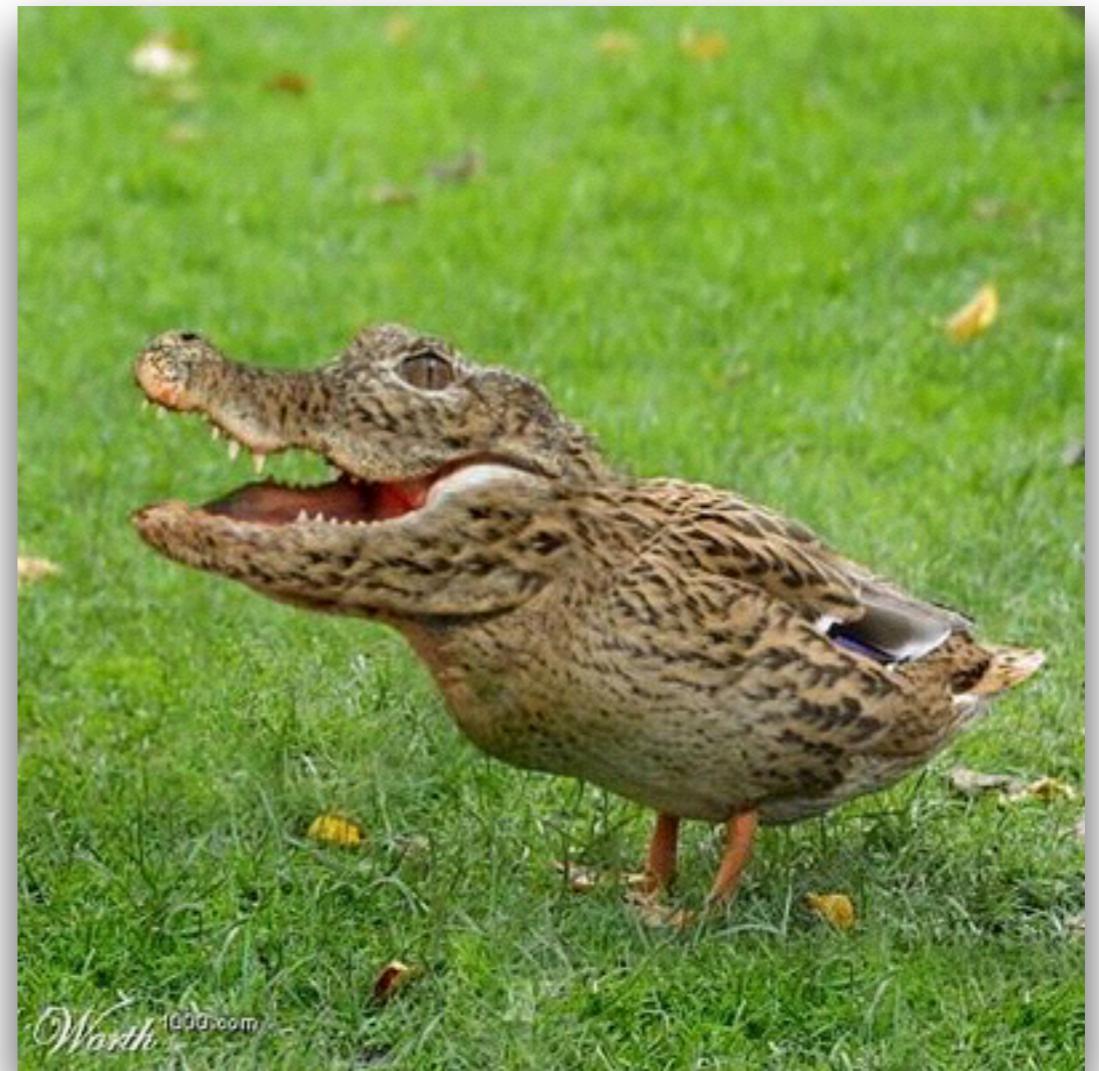
- Geografía aisla poblaciones y genera especies con origen común
- Especies relacionadas tienden a estar correlacionadas geográficamente



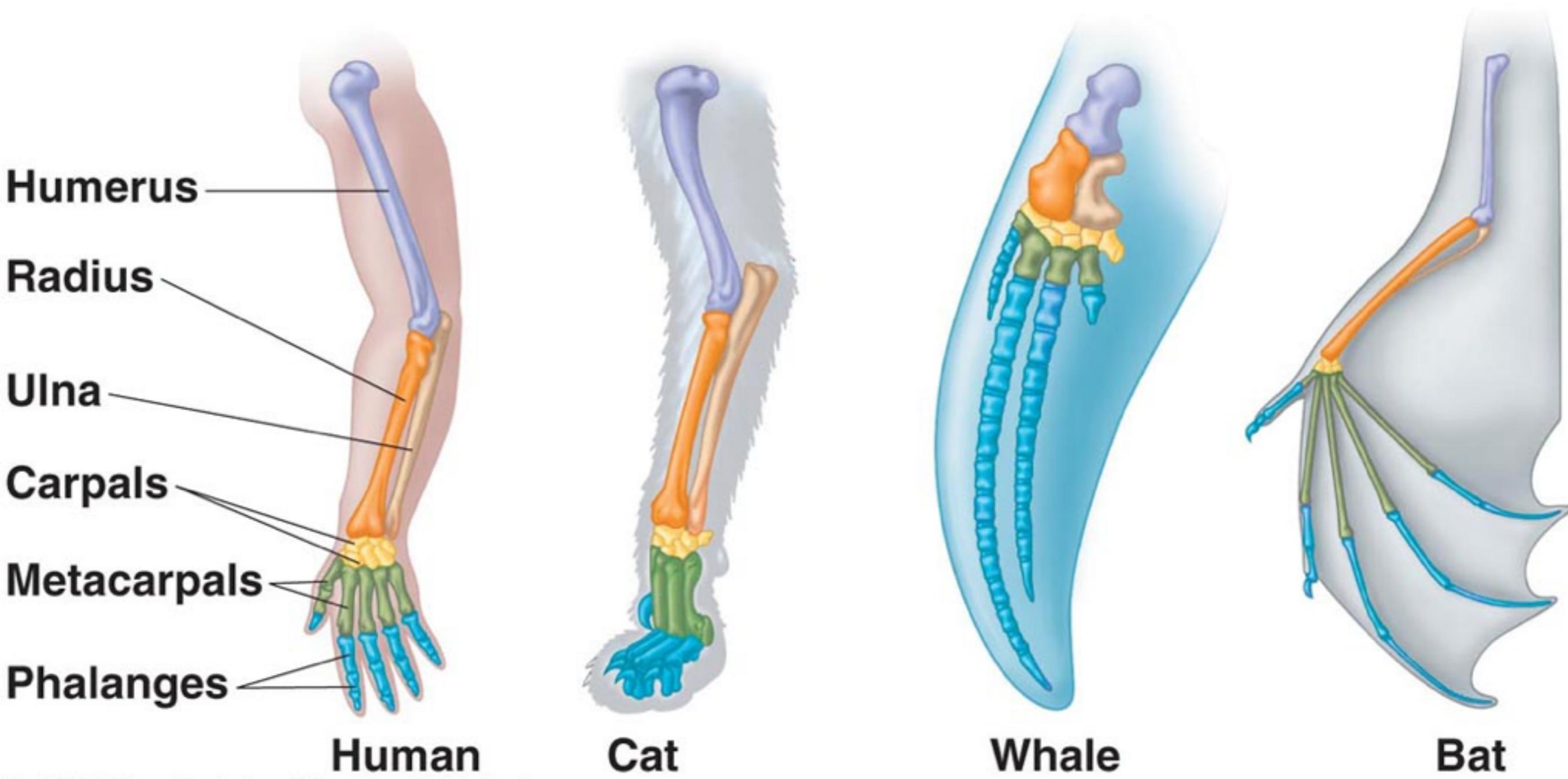
Homología

¿Qué es homología?

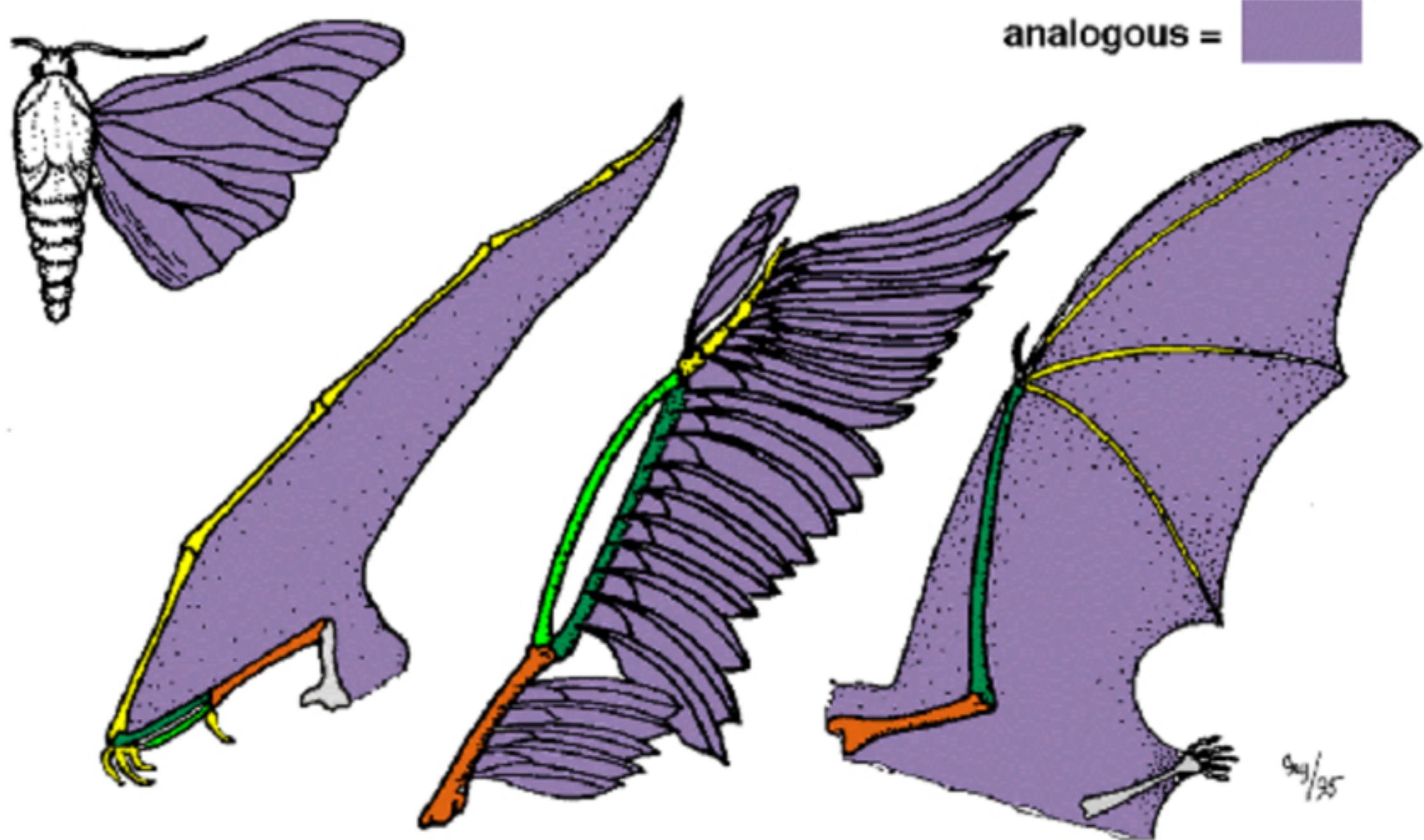
- **homología morfológica**
—> especies que pertenecen al mismo grupo taxonómico muestran similitudes anatómicas debido a que comparten un ancestro común (no por chance)



Estructuras homólogas



Estructuras análogas



¿Qué es homología?

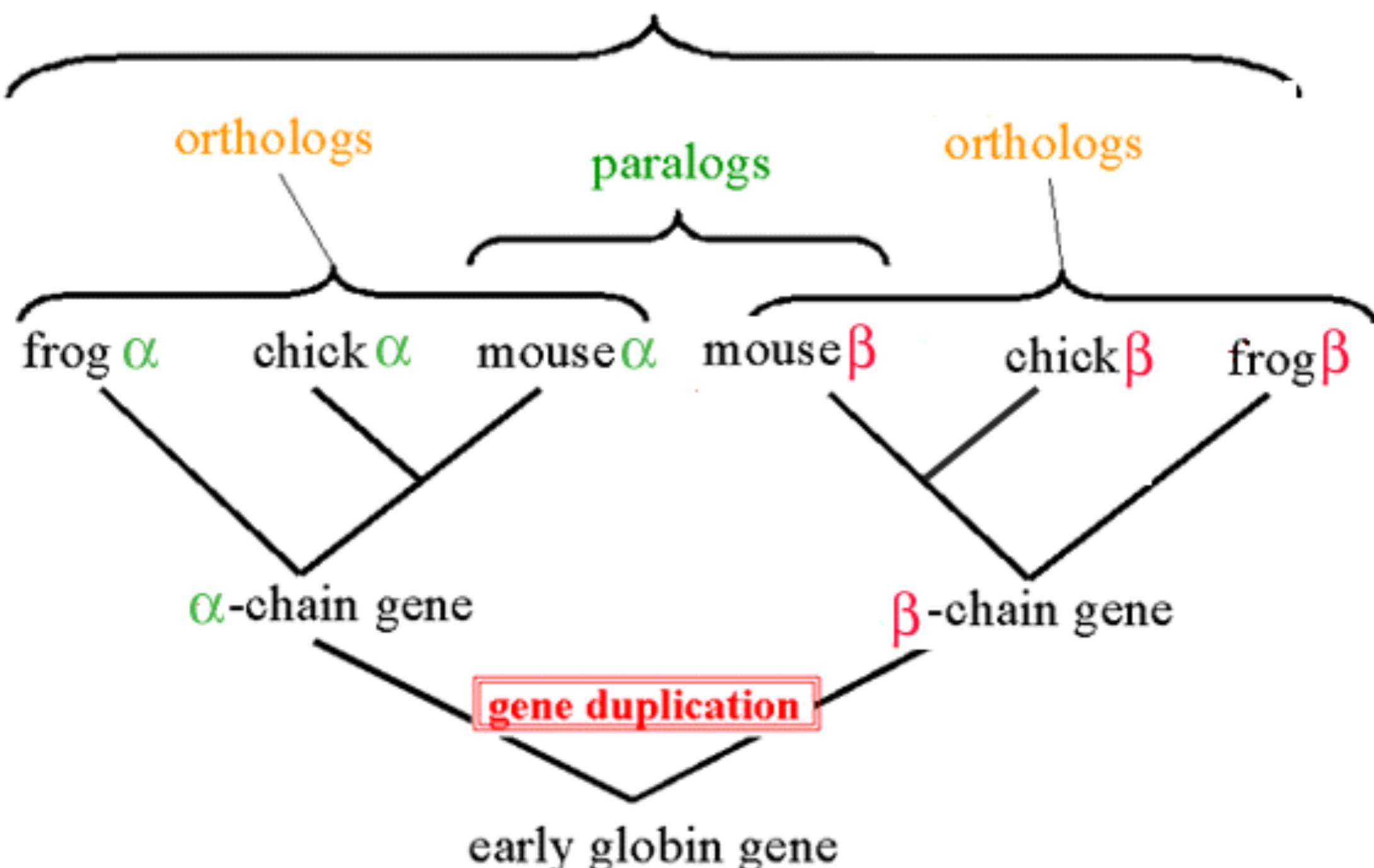
- **homología molecular**—> especies que pertenecen al mismo grupo taxonómico muestran similitudes en DNA, RNA y proteínas debido a que comparten un ancestro común (no por chance)

<i>Aquifex aeolicus</i>	MKVRSSVKK---	RCAKCKIIRRKGGRVMVICE-IPSHKQKTG
<i>Bacillus subtilis</i>	MKVRPSVKP---	ICEKCKVIRRKGVKVMVICE-NPKHKQKQG
<i>Campylobacter jejuni</i>	MKVRPSVKK---	MCDKCKVVRRKGGRVRIICE-NPKHKQRQG
<i>Chlamydia trachomatis</i>	MRVSSSIKA---	PSKGDKLVRKGRLYVINKDPRNRKQRQA
<i>Escherichia coli</i>	MKVRASVKK---	LCRNCKIVRKGGRVIRVICSAEPHKQRQG
<i>Helicobacter pylori</i>	MKVRPSVKK---	MCDNCKIIRRGGRVIRVICA-TPKHKQRQG
<i>Lactococcus lactis</i>	MKVRPSVKP---	ICEYCKVIRRNNGRVMVICPANPKHKQRQG
<i>Mycobacterium leprae</i>	MKVNPSPVKP---	MCDKCRVIRRHRRVMVICV-DPRHKQRQG
<i>Mycoplasma genitalium</i>	MKVRASVKP---	ICKDCCKIIRHRHILRVICK-TKKHKQRQG
<i>Rickettsia prowazekii</i>	MKVSSLKSLKKRDKDQIVKRGKIFVINKKNRFRAKQG	
<i>Synechocystis</i> sp.	MKVRASVKK---	MCDKCRVIRRRGRVMVICSANPKHKQRQG
<i>Treponema pallidum</i>	MKIRTSVVKV---	ICDKCKLIKRGIGIIRVICV-NPKHKQRQG
<i>Thermotoga maritima</i>	MKVQASVKK---	RCEHCKIIRRKKRVYVICKVNPKNQKQG
<i>Vibrio cholerae</i>	MKVRASVKK---	ICRNCKVIRRNNGVVRVIC-SEPKHKQRQG
<i>Xylella fastidiosa</i>	MKVSSLKSAKTRHRDCKVIRRGKIFVICKSNPRFKARQR	
Yeast	...	FKVRTSVKK---FCSDCYLVRRKGGRVYIYCKSNKKHKQRQG
Rice	...	MKIRASVRK---ICTKCRLLIRRGRIRVIC-SNPKHQKRQG
Fruit fly	...	FKVKGRLKR---RCKDCYIVVRQERGYVICPTHPRHKQMSM
Mouse	...	FKTKGVVIKK---RCKDCYKVRRGRWFILCKTNPKHKQRQM
Human	...	FKNKTVLKK---RCKDCYLVRRGRWYVYCKTHPRHKQRQM

Tipos de homología

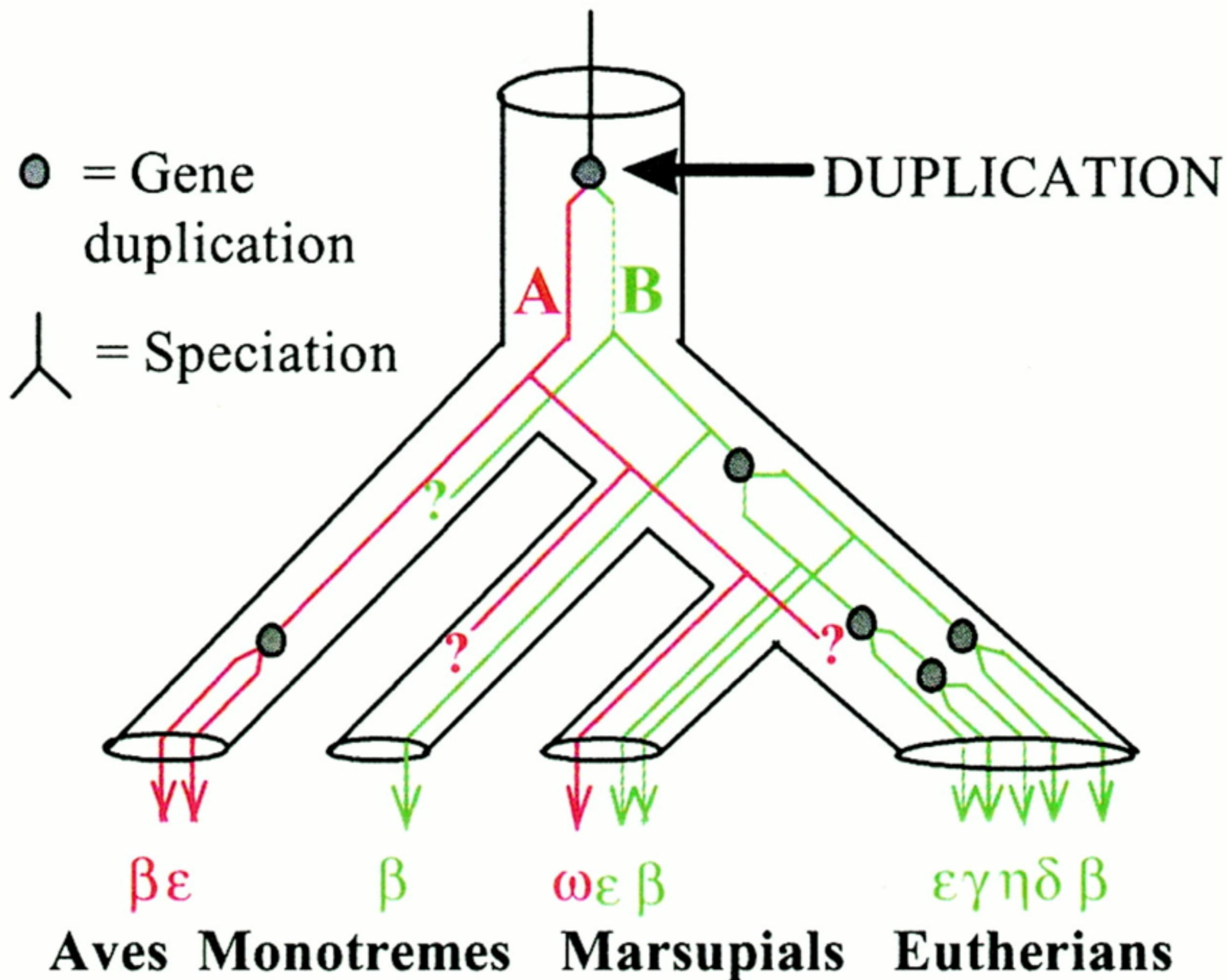
- Ortólogos
- Paralógos
- Xenólogos

Tipos de homología



Ancestral β -globin gene

● = Gene duplication
= Speciation

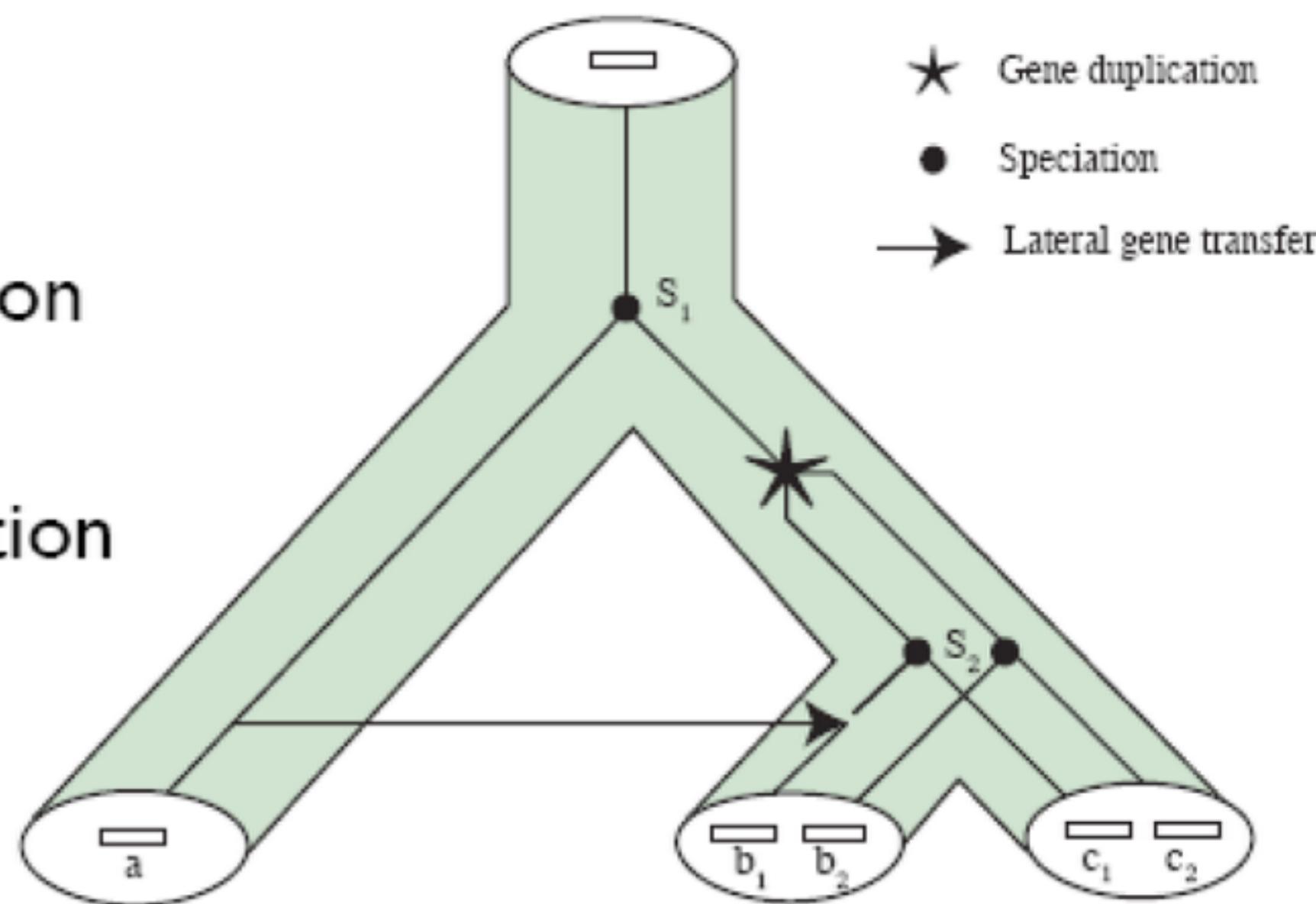


Tipos de homología

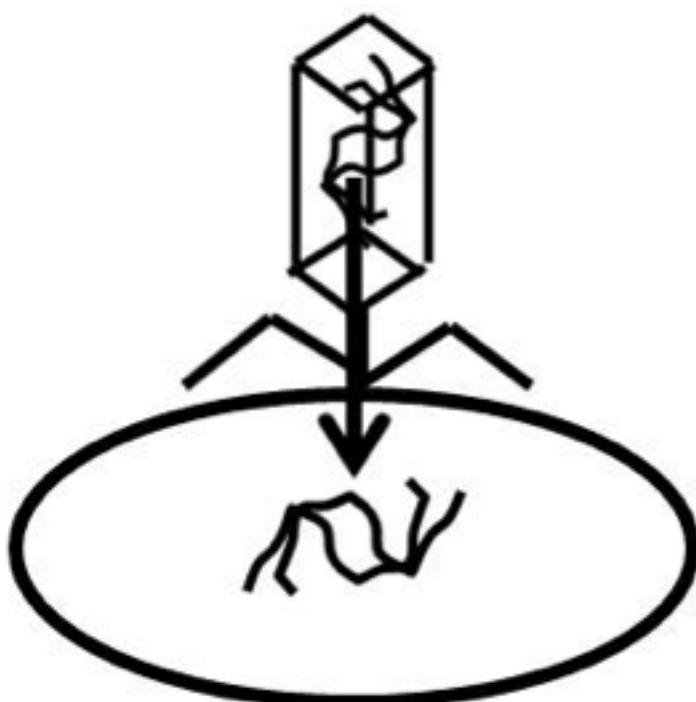
Two genes (or characters) are *homologs* if they have a common ancestor.

Main Subtypes

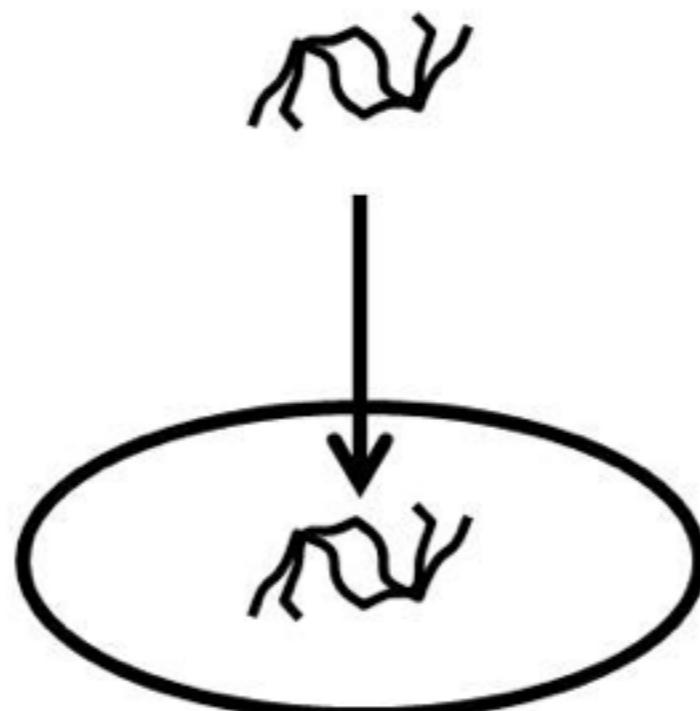
- *Orthologs*: through speciation
- *Paralogs*: through duplication
- *Xenologs*: through lateral transfer



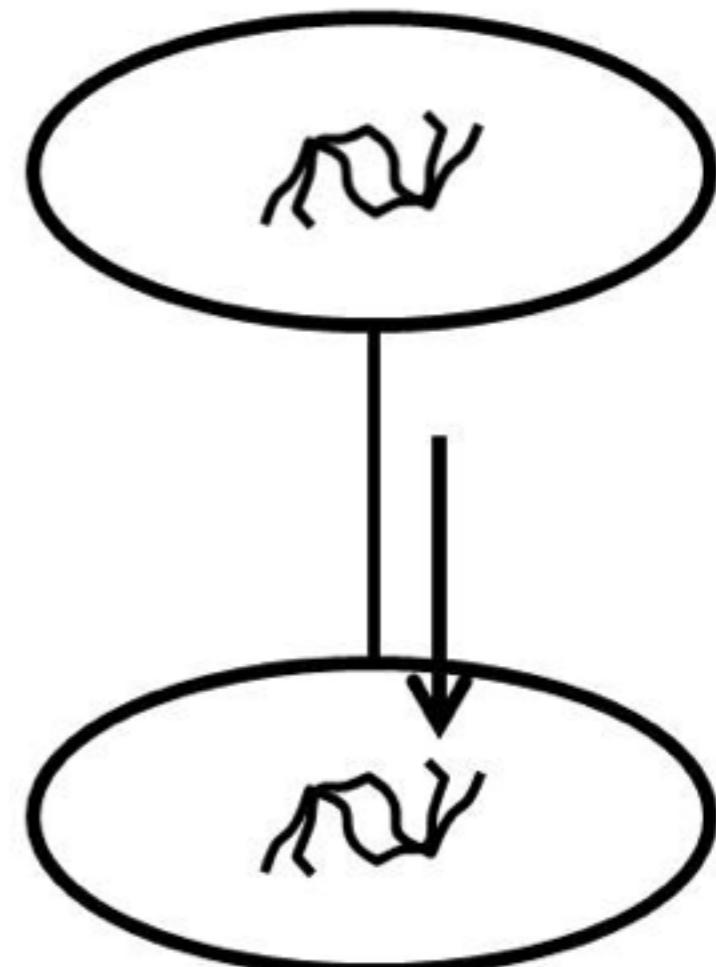
Xenólogos en bacteria



Transduction



Transformation



Conjugation

Bases de datos

¿Por qué necesitamos bases de datos?

- Biología es una ciencia cuantitativa
- Experimentos de alto rendimiento como secuenciamiento masivo producen vastas cantidades de datos
- Necesitamos clasificar, almacenar, buscar información biológica para transformar información en conocimiento

¿Qué pueden hacer las bases de datos?

- Hacer datos disponibles para científicos alrededor del mundo (human-readable)
- Hacer datos disponibles para que puedan ser analizados por algoritmos (machine-readable)

Tipos de bases de datos

- Primarias: Resultados experimentales directos.
Secundarias: Resultados de análisis
- Nucleótidos y proteínas
- Dominios y motivos
- Estructuras 3D
- Expresión génica
- Rutas metabólicas

¿Cómo encuentro la base de datos que necesito?

- NAR database issue
- Database Journal

NAR Database Summary Paper Category List

Nucleotide Sequence Databases
RNA sequence databases
Protein sequence databases
Structure Databases
Genomics Databases (non-vertebrate)
Metabolic and Signaling Pathways
Human and other Vertebrate Genomes
Human Genes and Diseases
Microarray Data and other Gene Expression Databases
Proteomics Resources
Other Molecular Biology Databases
Organelle databases
Plant databases
Immunological databases
Cell biology

The screenshot shows the homepage of Nucleic Acids Research. At the top, the journal title 'Nucleic Acids Research' is displayed in white on a red header. Below the header, there are links for 'ABOUT THIS JOURNAL', 'CONTACT THIS JOURNAL', and 'SUBSCRIPTIONS'. A navigation bar at the bottom includes 'Oxford Journals > Science & Mathematics > Nucleic Acids Research > Volume 44, Issue 1'. The main content area features a section titled 'The 2016 database issue of Nucleic Acids Research and an updated molecular biology database collection' by Daniel J. Rigden, Xosé M. Fernández-Suárez, and Michael Y. Galperin. There is also a link to 'Author Affiliations'.

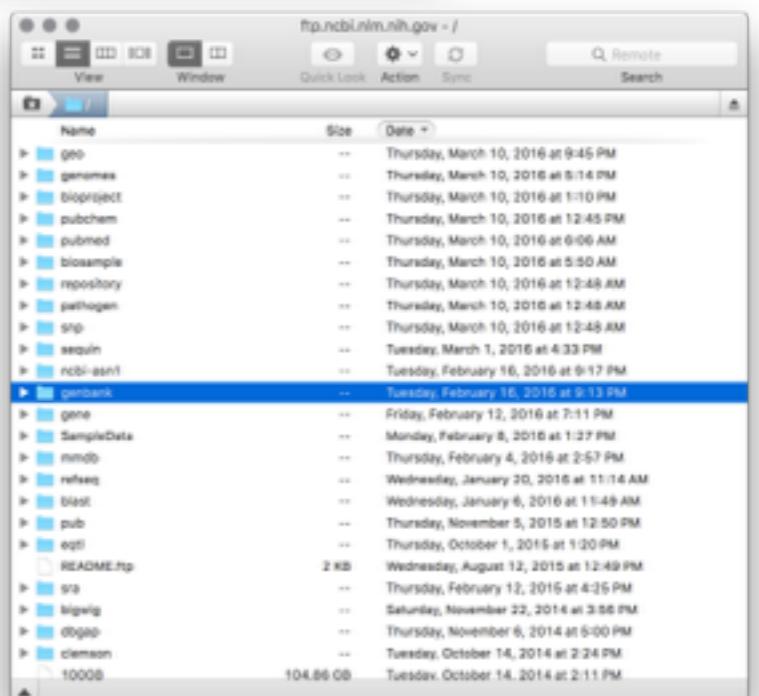
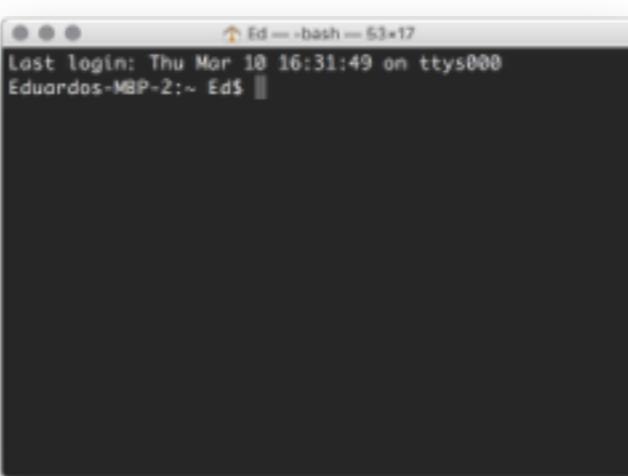
The screenshot shows the homepage of DATABASE: The Journal of Biological Databases and Curation. The title 'DATABASE' is prominently displayed in large blue letters, with the subtitle 'The Journal of Biological Databases and Curation' in smaller text. Below the title, there are links for 'ABOUT THIS JOURNAL', 'CONTACT THIS JOURNAL', 'SUBSCRIPTIONS', and 'CURRENT ISSUE'. A navigation bar at the bottom includes 'Oxford Journals > Science & Mathematics > Database'. A 'READ THIS JOURNAL' button is visible. On the right side, there is a welcome message: 'Welcome to Database: The Journal of Biological Databases and Curation', followed by links for 'A Fully Open Access Journal', 'View Current Content', 'Browse the Archive', 'Biocuration Virtual Issue', 'Biomart Virtual Issue', and 'Now Indexed in PubMed Central'.

¿Cómo accedo a los datos contenidos en bases de datos?

- Interface gráfica →

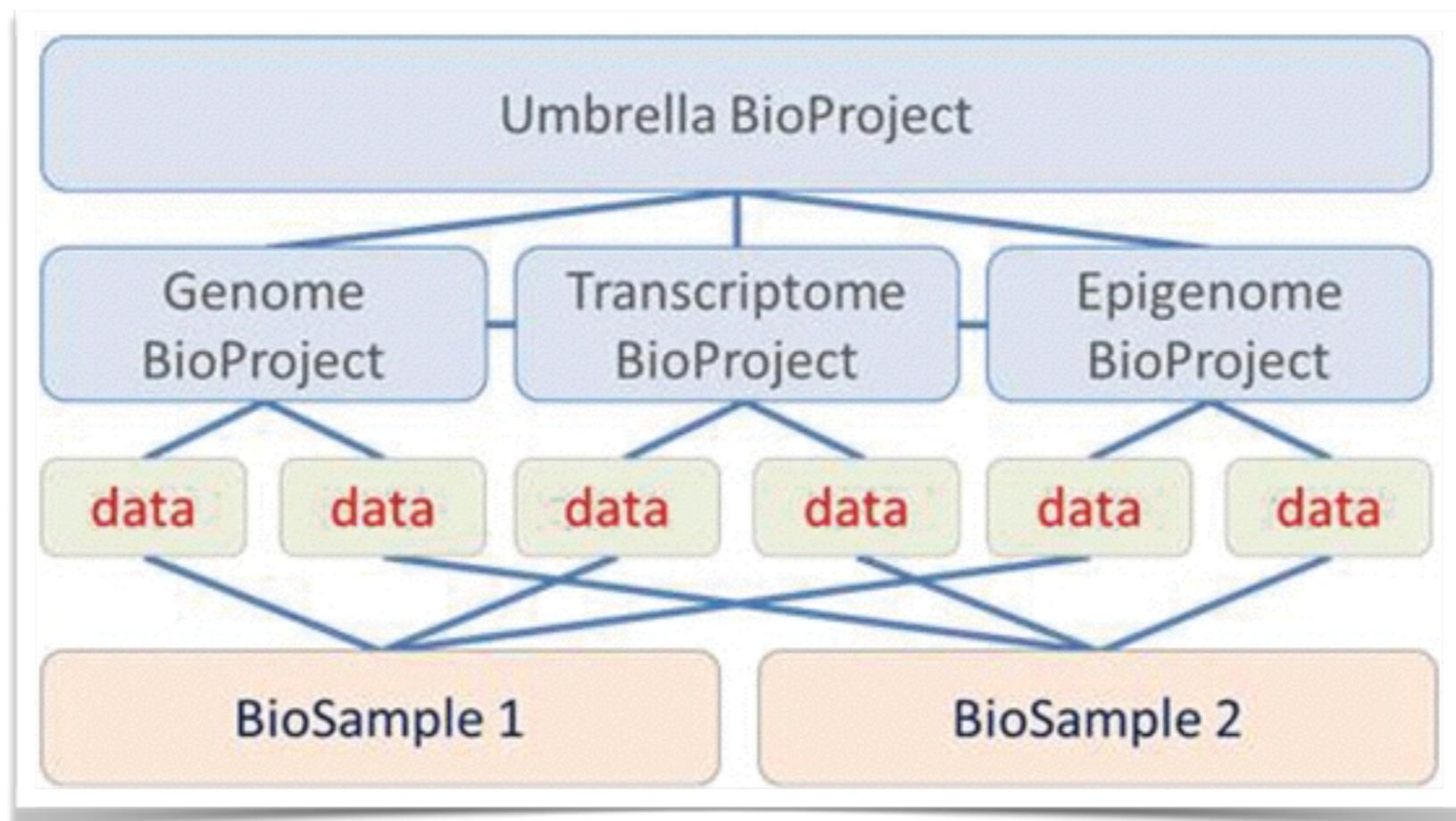


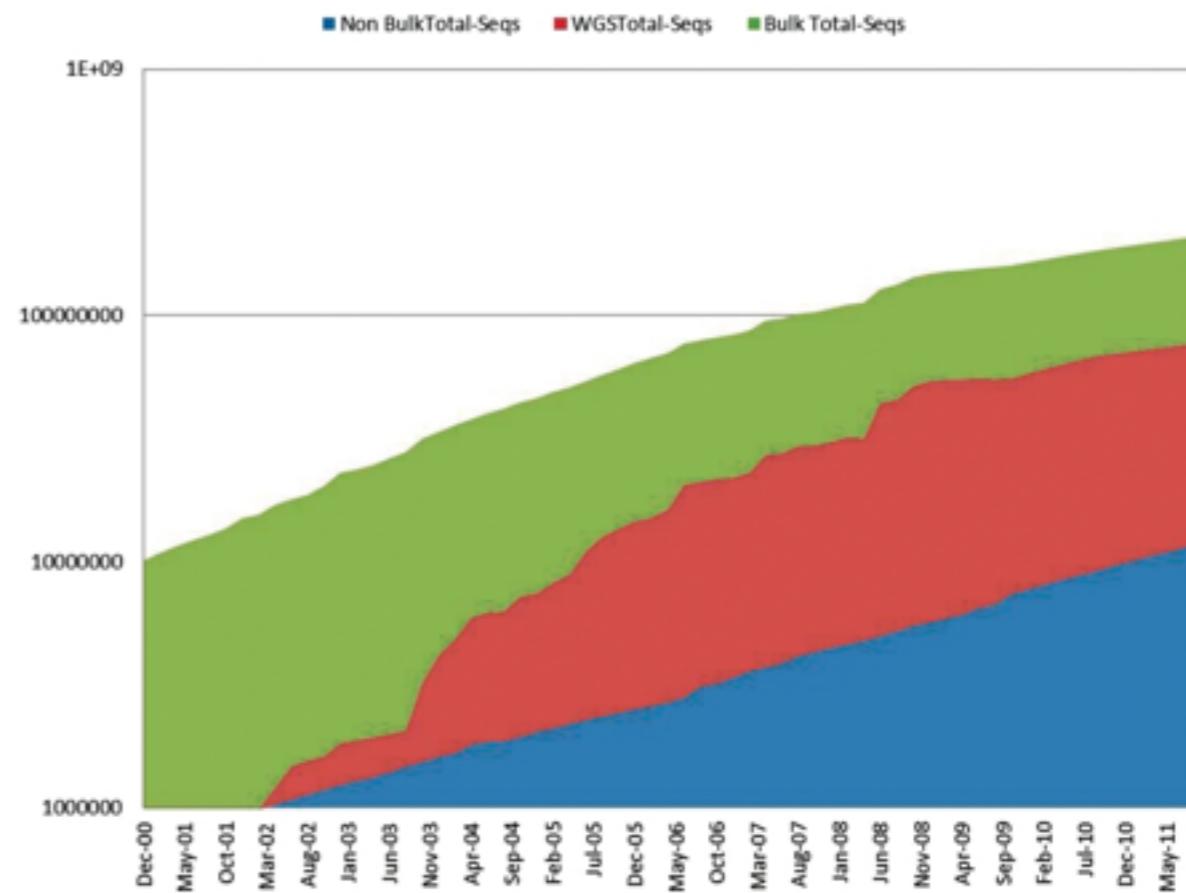
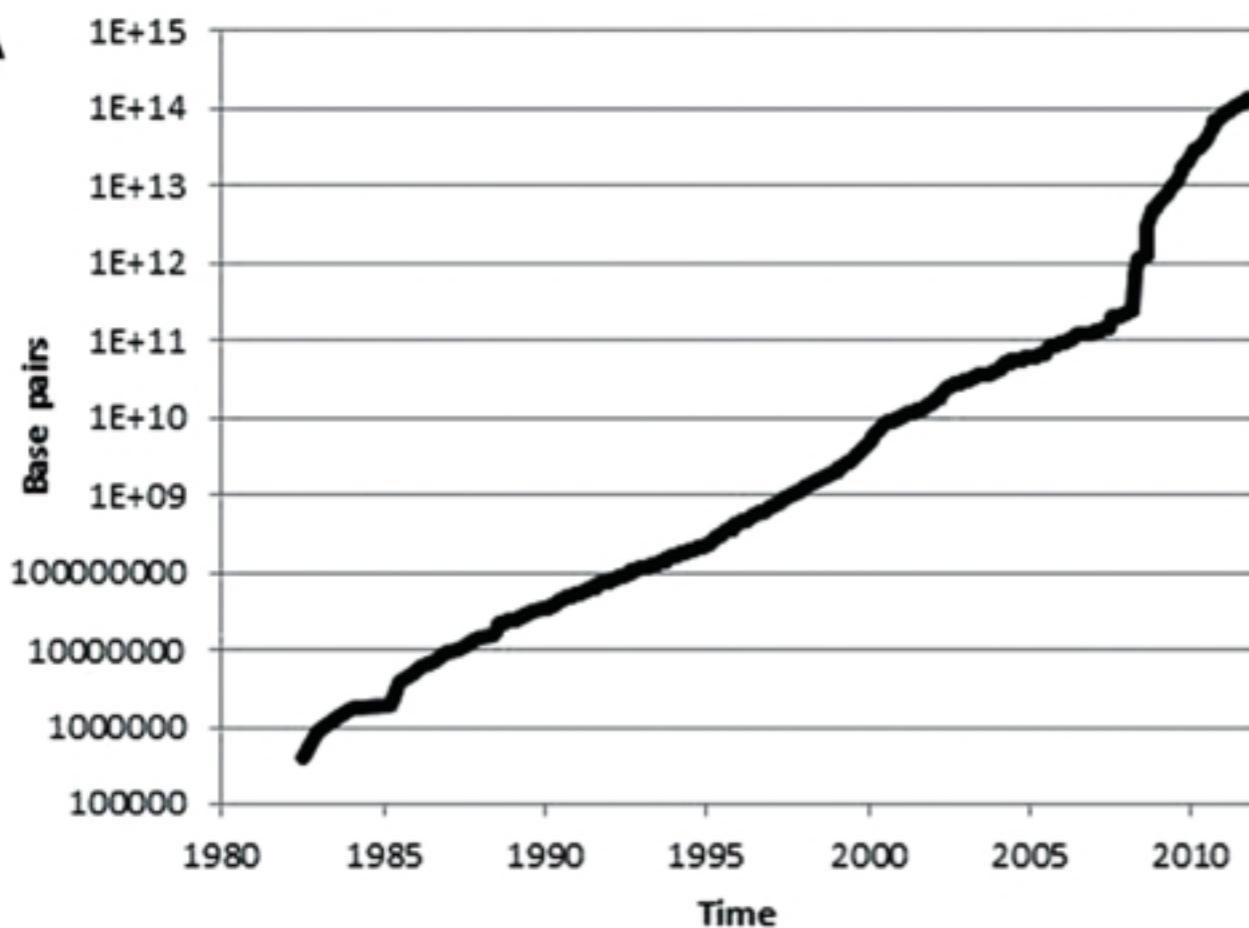
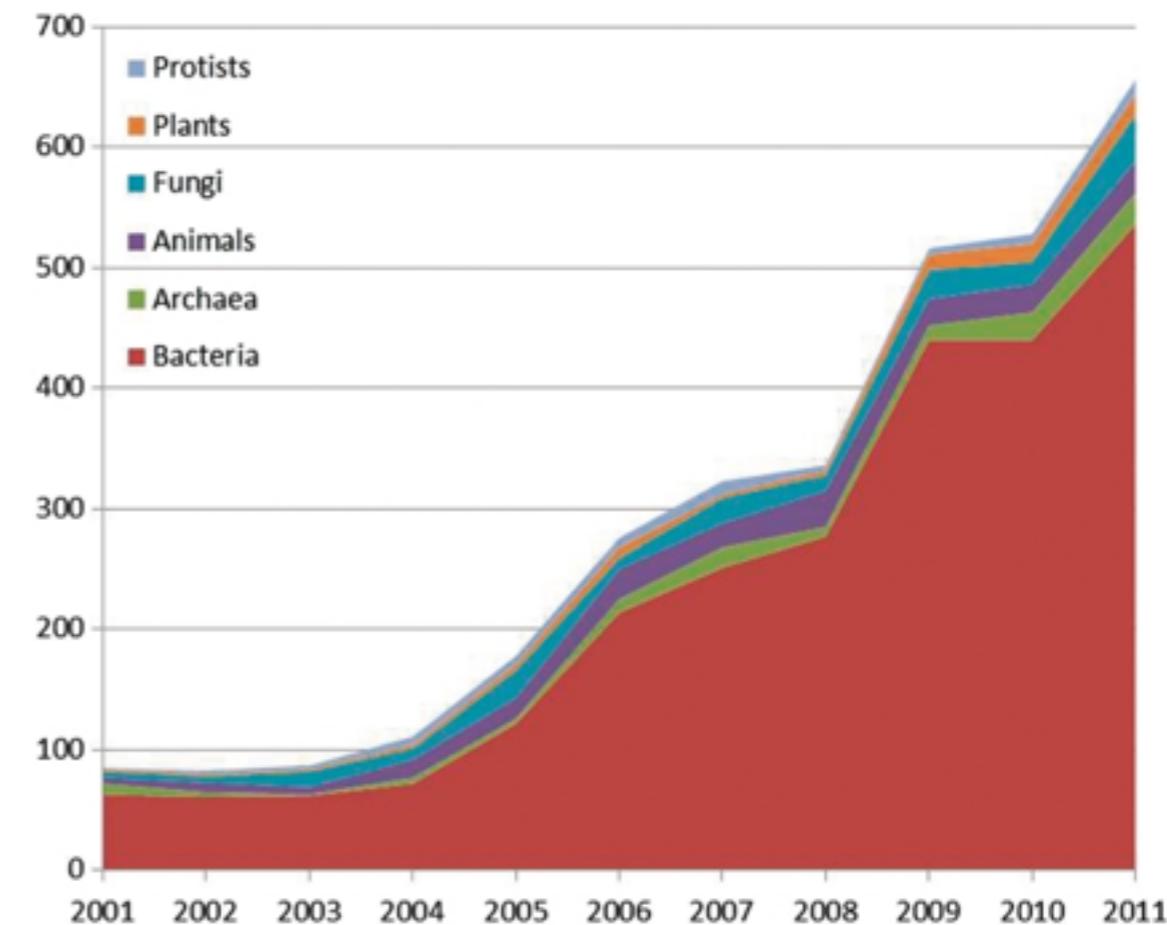
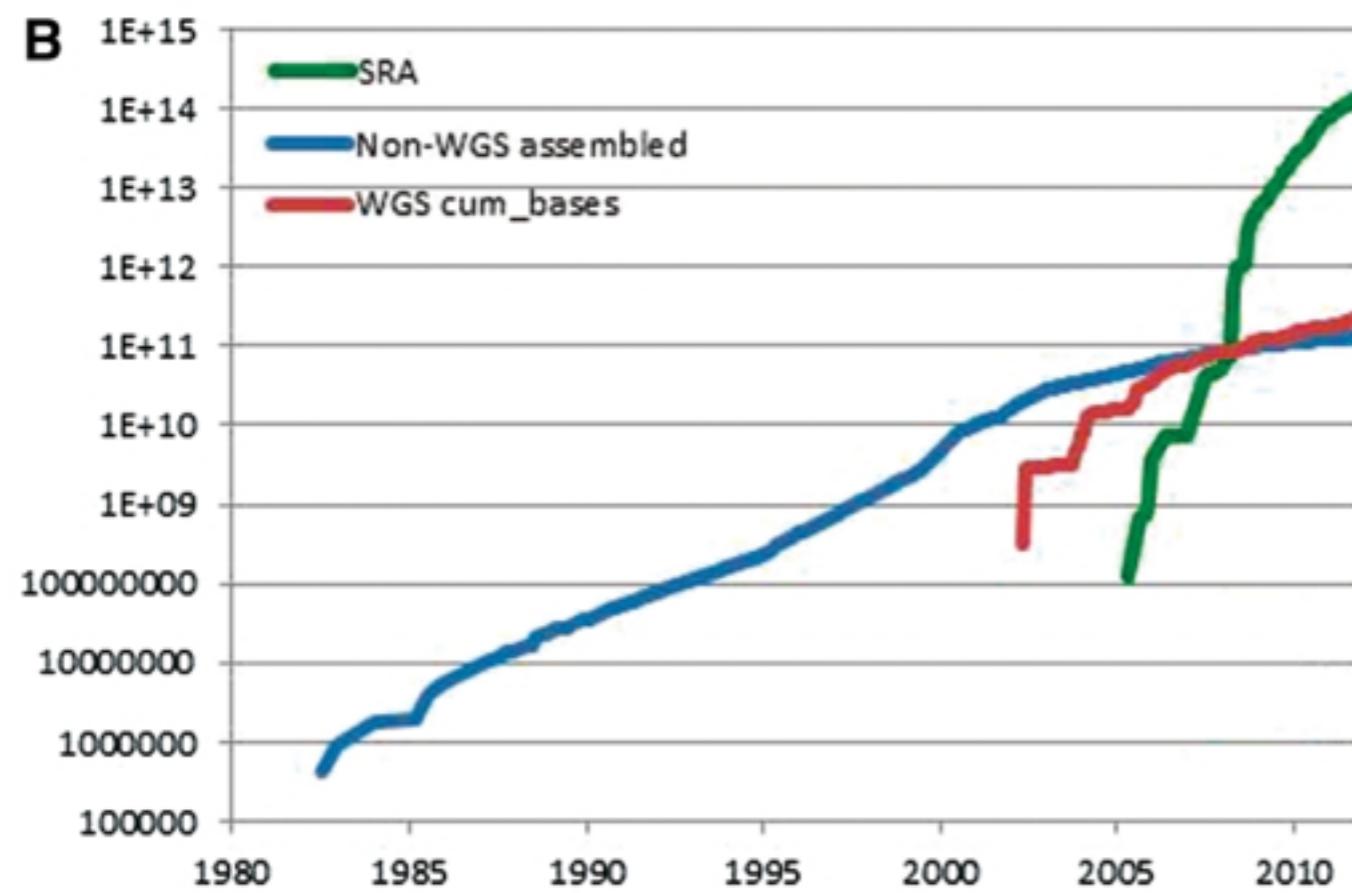
- “Puerta de atrás” FTP, scripting →
ftp.ncbi.nlm.nih.gov



Ejemplo más popular

- International Nucleotide Sequence Database Collaboration (INSDC)
- 30 años de antigüedad



A**B**

Clases especiales de bases de datos

- No curadas
 - NCBI nt, TrEMBL
- Curadas (con diferentes niveles de evidencia)
 - Swiss-Prot, PIR
- Especializadas
 - PeptidesAtlas —> espectros de masas de péptidos

... de enzimas

- KEGG
- BRENDA
- CAZymes

The image displays three separate web browser windows side-by-side, each showing a different biological database interface.

- KEGG:** The first window shows the KEGG homepage. It features a sidebar with links like "KEGG Home", "KEGG Database", "KEGG Objects", "KEGG Software", and "KEGG Labs". The main content area is titled "KEGG: Kyoto Encyclopedia of Genes and Genomes" and describes KEGG as a database resource for understanding high-level functions and utilities of the biological system.
- BRENDA:** The second window shows the BRENDA homepage. It features a prominent logo for "BRENDA - The Comprehensive Enzyme Information System". Below the logo, it says "New BRENDA release online since January, 6th 2016". It includes links for "Word Maps" and "Pathway Maps".
- CAZy:** The third window shows the CAZy homepage. It features a logo for "CAZy - Carbohydrate-active enzymes". The main content area is titled "Welcome to the Carbohydrate-Active enZYmes Database". It includes a sidebar with links for "Sequence Search", "Genome Explorer", "Ontology Explorer", and "Functional Enzyme Parameters".

... de proteínas

- Protein Data Bank

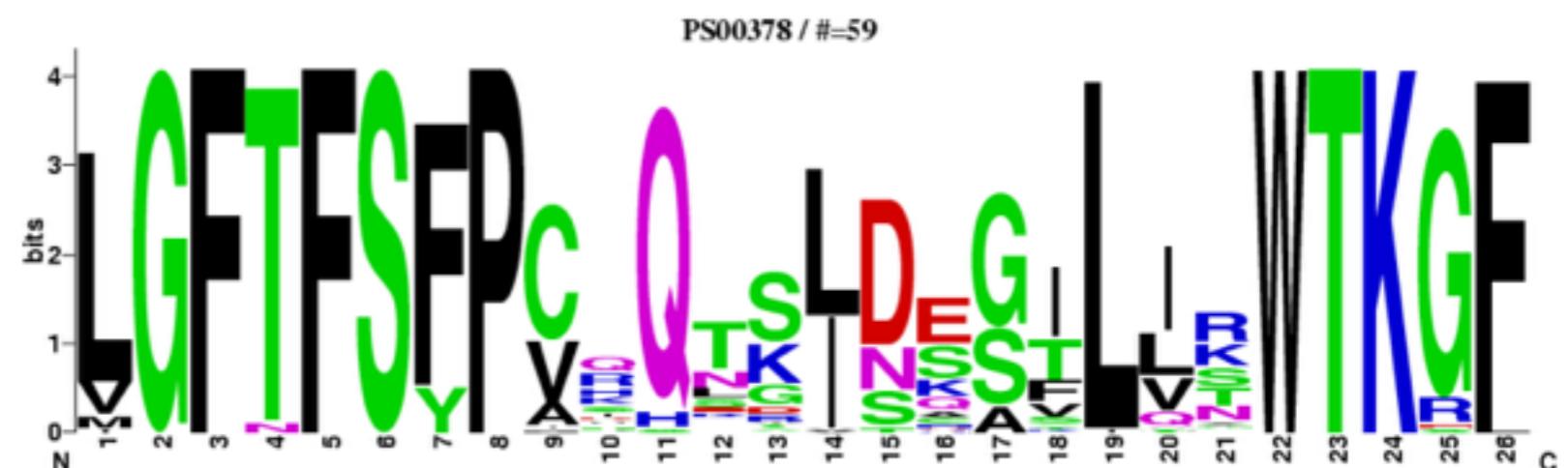


Bases de datos secundarias

- PROSITE
- Pfams
- Rfams
- PRINTS

PROSITE [HEXOKINASES PS00378]

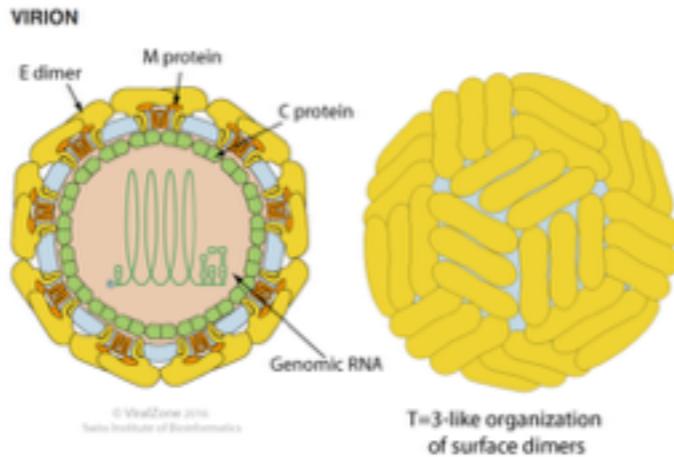
- Database of protein domains, families and functional sites
- Hexokinases signature: Pattern [LIVM]-G-F-[TN]-F-S-[FY]-P-x(5)-[LIVM]-[DNST]-x(3)-[LIVM]-x(2)-W-T-K-x-[LF].



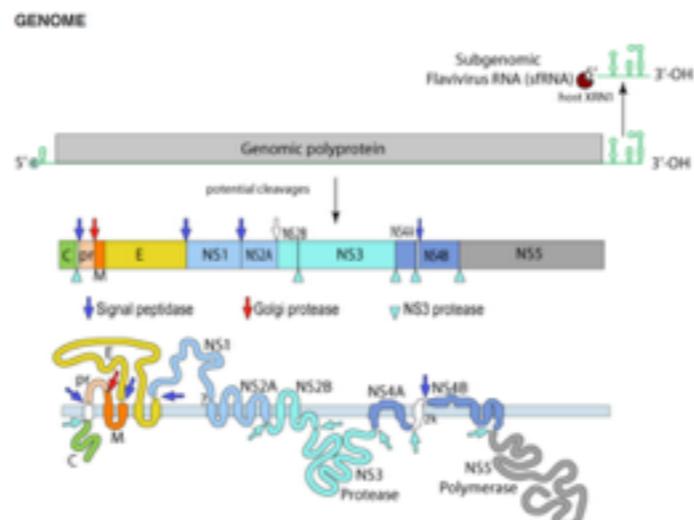
Zika virus y taxonomía

Zika virus (strain Mr 766)

- UniProt Taxonomy
- NCBI Taxonomy
- ViralZone



Enveloped, spherical, about 50 nm in diameter. The surface proteins are arranged in an icosahedral-like symmetry.



Monopartite, linear, ssRNA(+) genome of 10,794 bp. The genome 5' end has a methylated nucleotide cap for canonical cellular translation. The 3' terminus is not polyadenylated but forms a loop structure. This secondary structure leads to the formation of a subgenomic flavivirus RNA (sfRNA) through genomic RNA degradation by host XRN1. sfRNA is essential for pathogenicity, and may play a role in inhibiting host RIG-I antiviral activity as shown for Dengue virus.

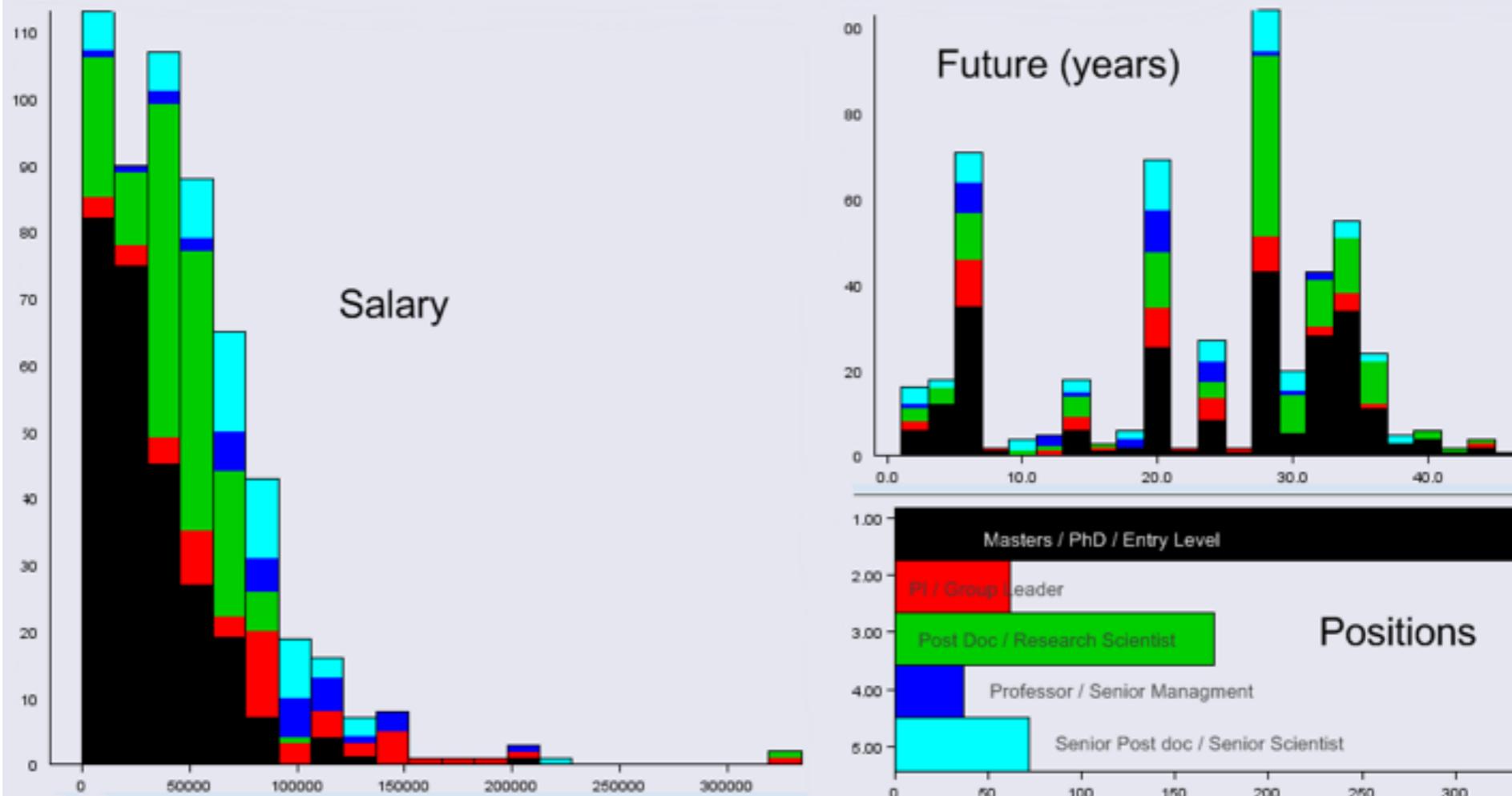
GENE EXPRESSION

The virion RNA is infectious and serves as both the genome and the viral messenger RNA. The whole genome is translated in a polyprotein 3,419 aa long, which is processed co- and post-translationally by host and viral proteases.

REPLICATION

CYTOPLASMIC in mammals, **NUCLEAR** in insects?

1. **Attachement** of the viral envelope protein E to **host receptors** mediates internalization into the host cell by **apoptotic mimicry**.
2. **Fusion of virus membrane with host endosomal membrane**. RNA genome is released into the cytoplasm.
3. The positive-sense genomic ssRNA is translated into a polyprotein, which is cleaved into all structural and non structural proteins (to yield the replication proteins).
4. Replication takes place at the surface of endoplasmic reticulum in **cytoplasmic viral factories**. A dsRNA genome is synthesized from the genomic ssRNA(+).
5. The dsRNA genome is **transcribed/replicated** thereby providing viral mRNAs/new ssRNA(+) genomes.
6. Virus **assembly** occurs at the endoplasmic reticulum. The virion **buds via the host ESCRT complexes** at the endoplasmic reticulum, is transported to the Golgi apparatus.
7. The prM protein is cleaved in the Golgi, thereby maturing the virion which is fusion competent.
8. Release of new virions by **exocytosis**.



Ahora, el lab...

[https://github.com/
bioinf-biotec](https://github.com/bioinf-biotec)