IBT 432 Aplikasi Bioinformatika

Analisis komparasi genomika II: Pengenalan Genomika Komparatif

Riza Arief Putranto

Pertemuan V – 6 Oktober 2018
Prodi Bioteknologi
Fakultas Ilmu-Ilmu Kesehatan

Rencana Perkuliahan

1. ~~Kontrak belajar dan pengenalan bioinformatika aplikatif~~
2. ~~Database sekuen dan analisis genomika~~
3. ~~Anotasi sekuen ke genom - Praktik~~
4. ~~Analisis komparasi genomika I~~
5. Analisis komparasi genomika II
6. Analisis komparasi genomika III
7. Analisis komparasi genomika – Praktik
8. Protein modelling I
9. Protein modelling II
10. Protein modelling III
11. Protein modelling - Praktik
12. Visualisasi protein modelling
13. Visualisasi protein modelling - Praktik
14. Presentasi mahasiswa

Contents

□ History of HG Project

□ Definition of CG

□ Concept of synteny

□ Objectives of CG

□ Concept of homology

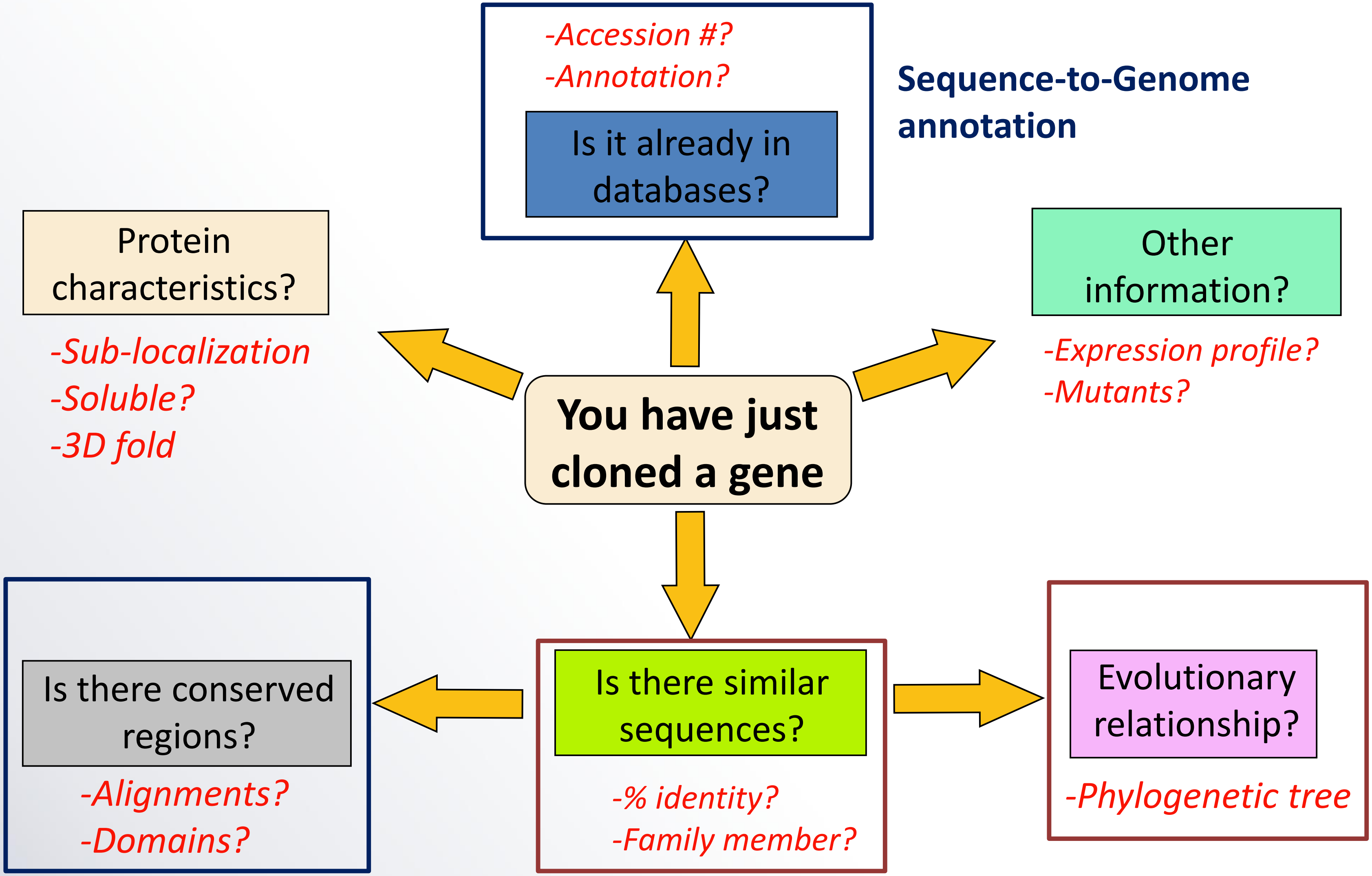
□ Visualization of CG

□ Conclusion

History

- **Human Genome Project (2001-2003)** decided to use smaller genomes as warm-up for human genome
- Resulted in sequencing the following:
  - Many bacteria
  - Model-organism genomes
    - ✓Yeast, ***C. elegans***, ***Arabidopsis***, ***Drosophila***
    - ✓Comparison of these genome sequences provided basis for field of comparative genomics

Remember this: analysis of genomics data



You have just cloned a gene

-Accession #?
-Annotation?

Is it already in databases?

Sequence-to-Genome annotation

Protein characteristics?

-Sub-localization
-Soluble?
-3D fold

Other information?

-Expression profile?
-Mutants?

Is there similar sequences?

-% identity?
-Family member?

Is there conserved regions?

-Alignments?
-Domains?

Evolutionary relationship?

-Phylogenetic tree

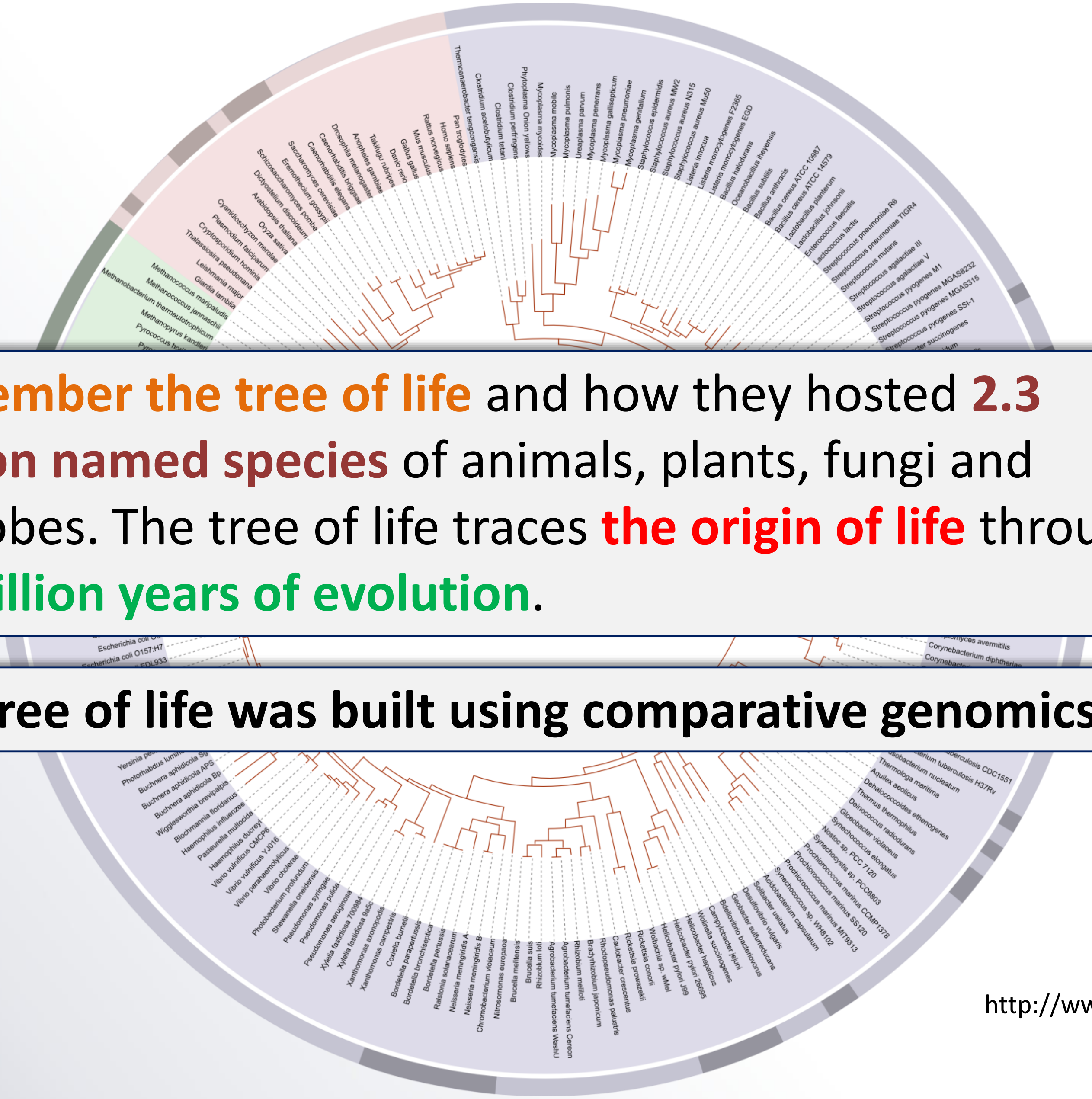
Comparative genome analysis for putative gene identification

Definition of comparative genomics (CG)

- **Comparative genomics** is a field of biological research in which the **genomic features of different organisms are compared**.
- The genomic features may include the **DNA sequence, genes, gene order, regulatory sequences**, and other genomic structural landmarks.
- In this branch of genomics, **whole or large parts** of genomes resulting from genome projects are compared to study **basic biological similarities** and **differences** as well as evolutionary relationships between organisms.

Xia X 2013 Comparative Genomics, pp 1-20

Why are we doing CGs?

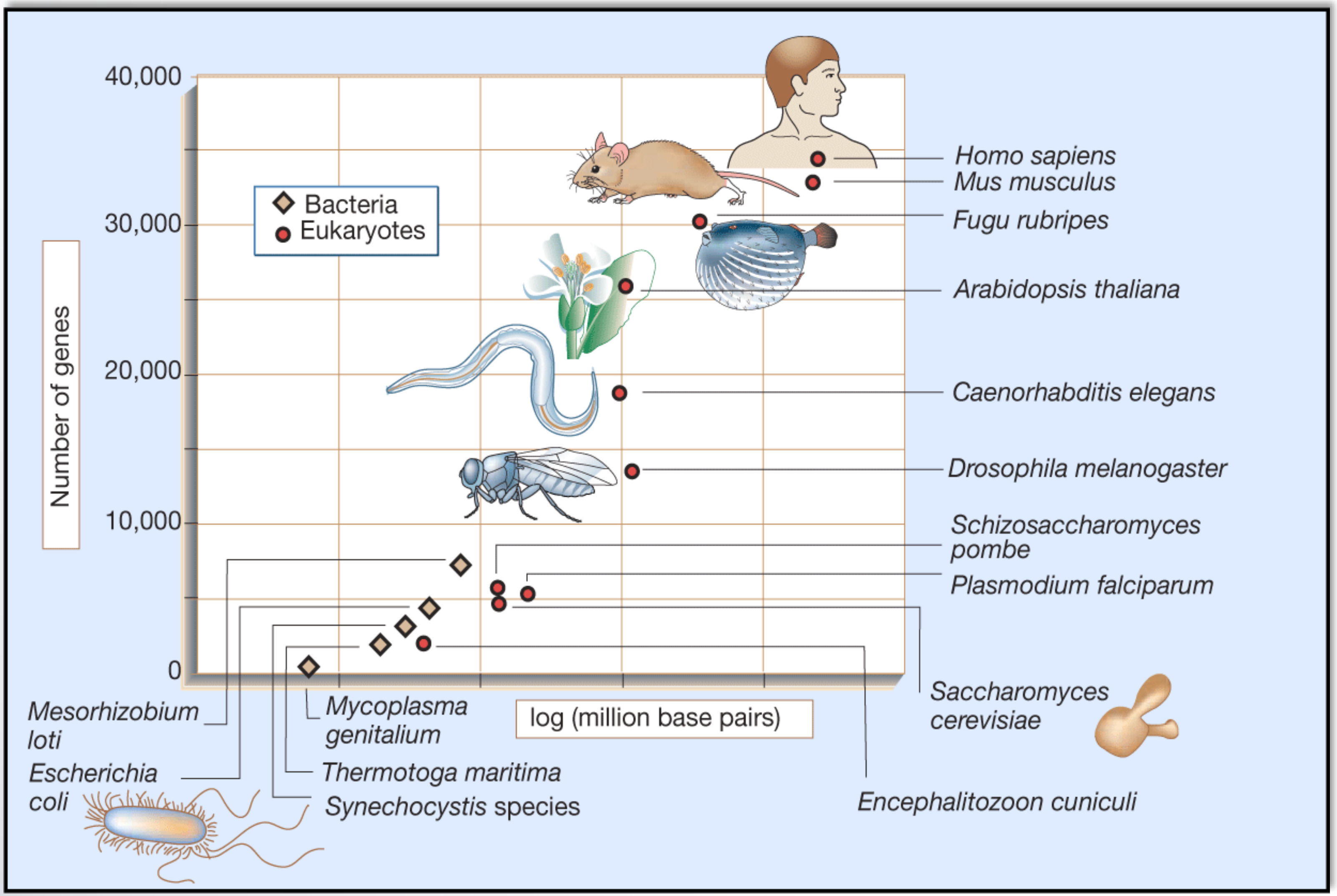


Remember the tree of life and how they hosted **2.3 million named species** of animals, plants, fungi and microbes. The tree of life traces **the origin of life** through **3.5 billion years of evolution**.

The tree of life was built using comparative genomics

http://www.tolweb.org/tree/

Sizes of genomes and numbers of genes

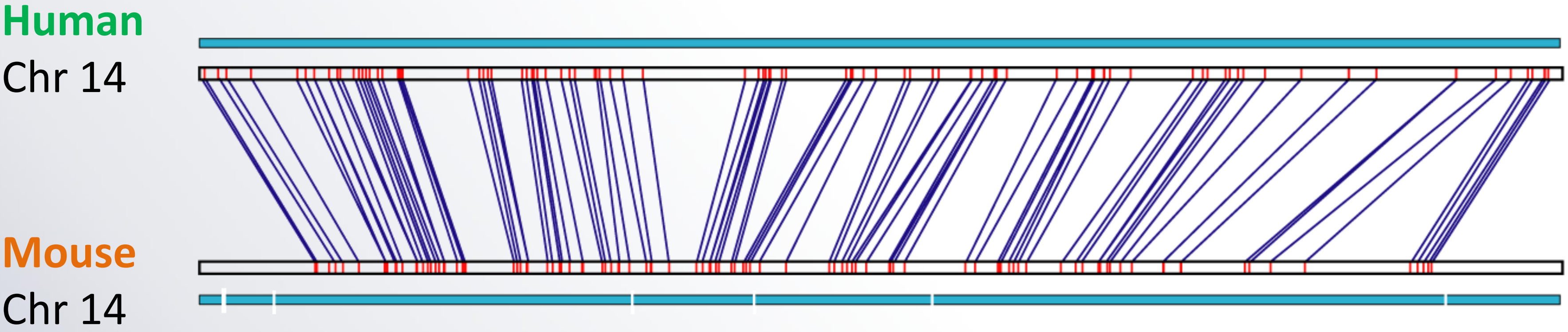


CG built on the concept of synteny

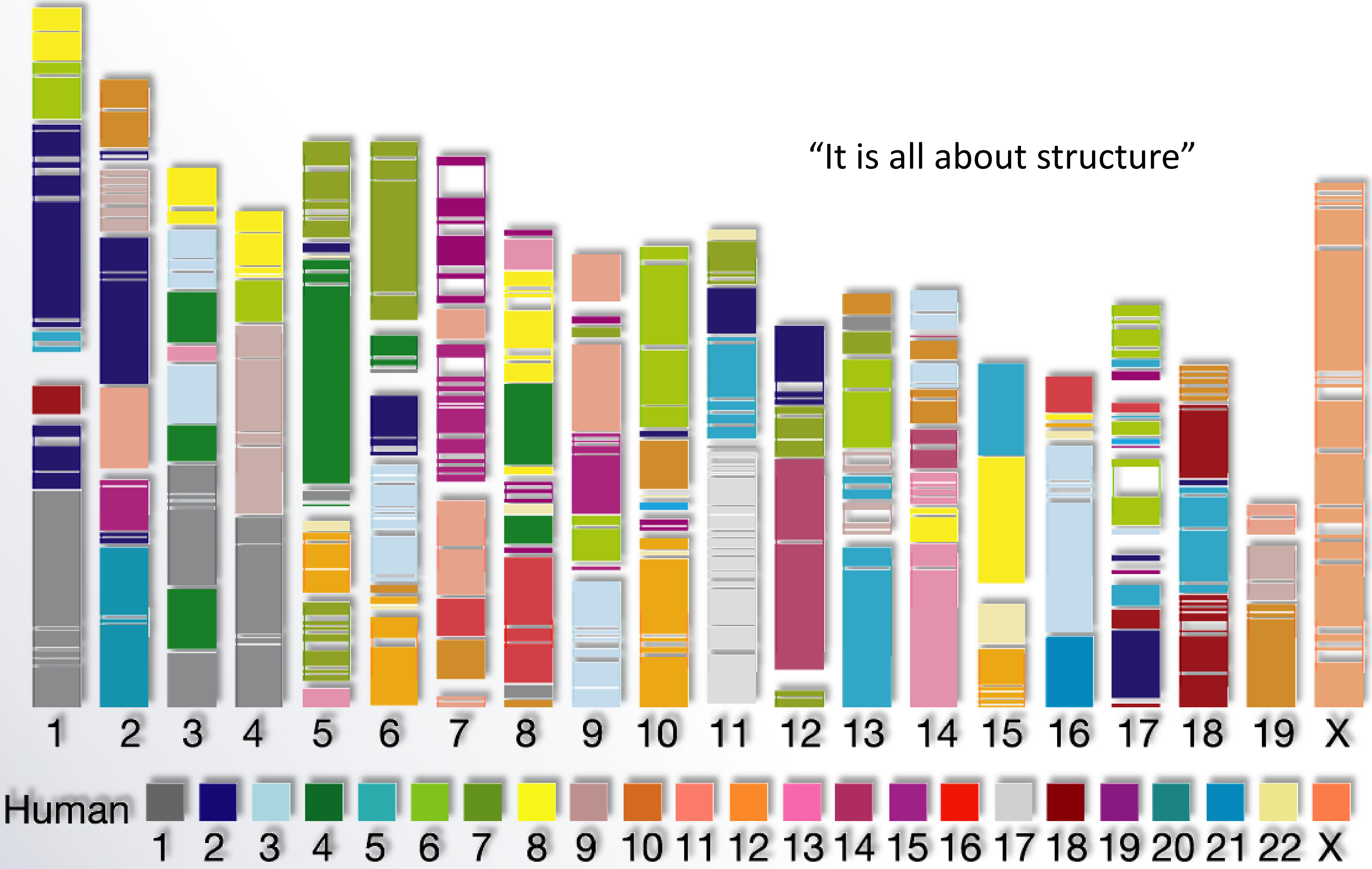
- Synteny: **genes** that are in the **same relative position** on **two different chromosomes**
- Genetic and physical maps compared between species
  - Or between chromosomes of the same species
- **Closely related species** generally have **similar order of genes on chromosomes**
- Synteny can be used to **identify genes in one species** based on **map position in another**

Mouse vs Human synteny

- When sequences from mouse and human genomes are compared, **regions of remarkable synteny were found**
- Genes are in **almost identical order** for long stretches along the chromosome



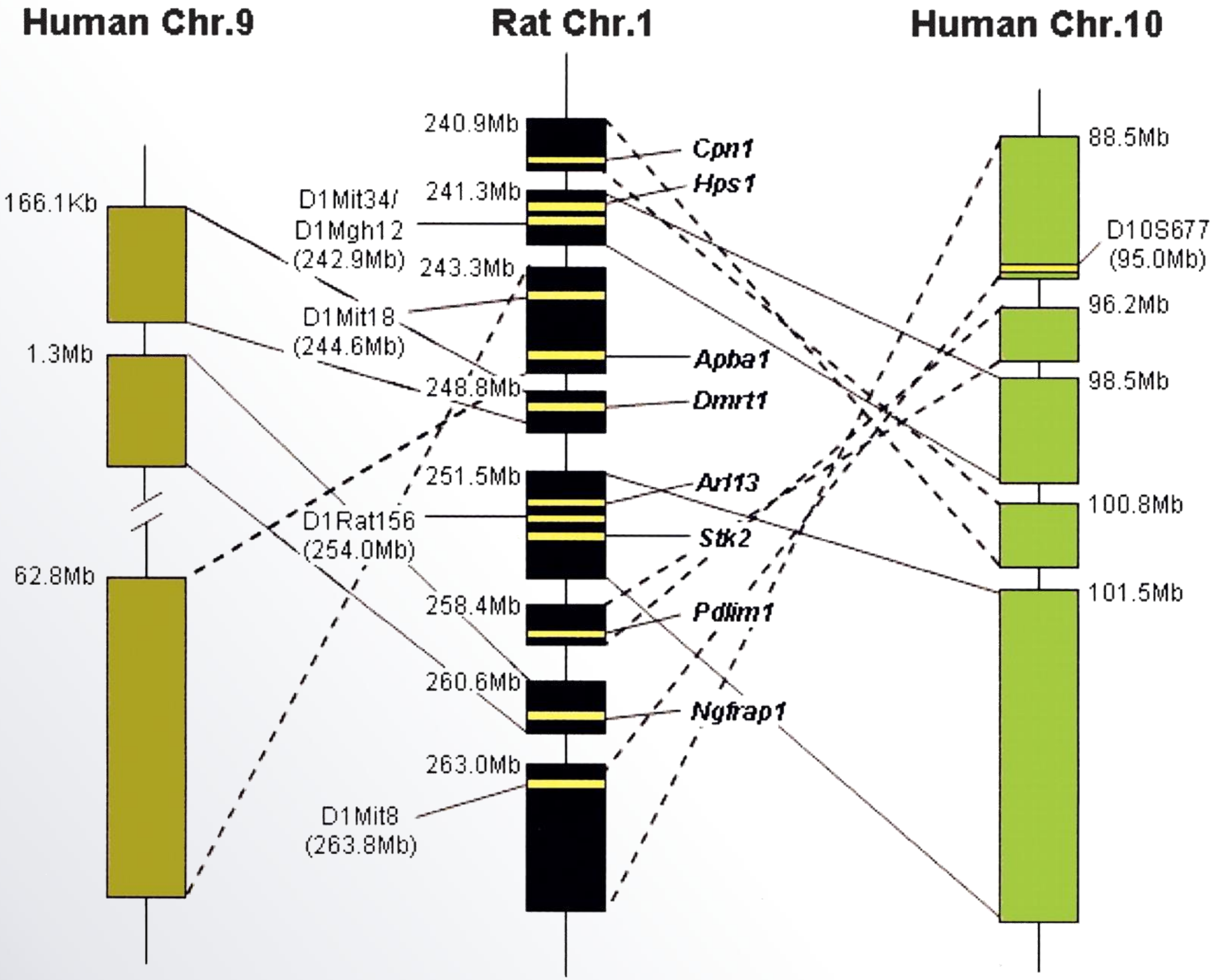
Mouse vs Human synteny



"It is all about structure"

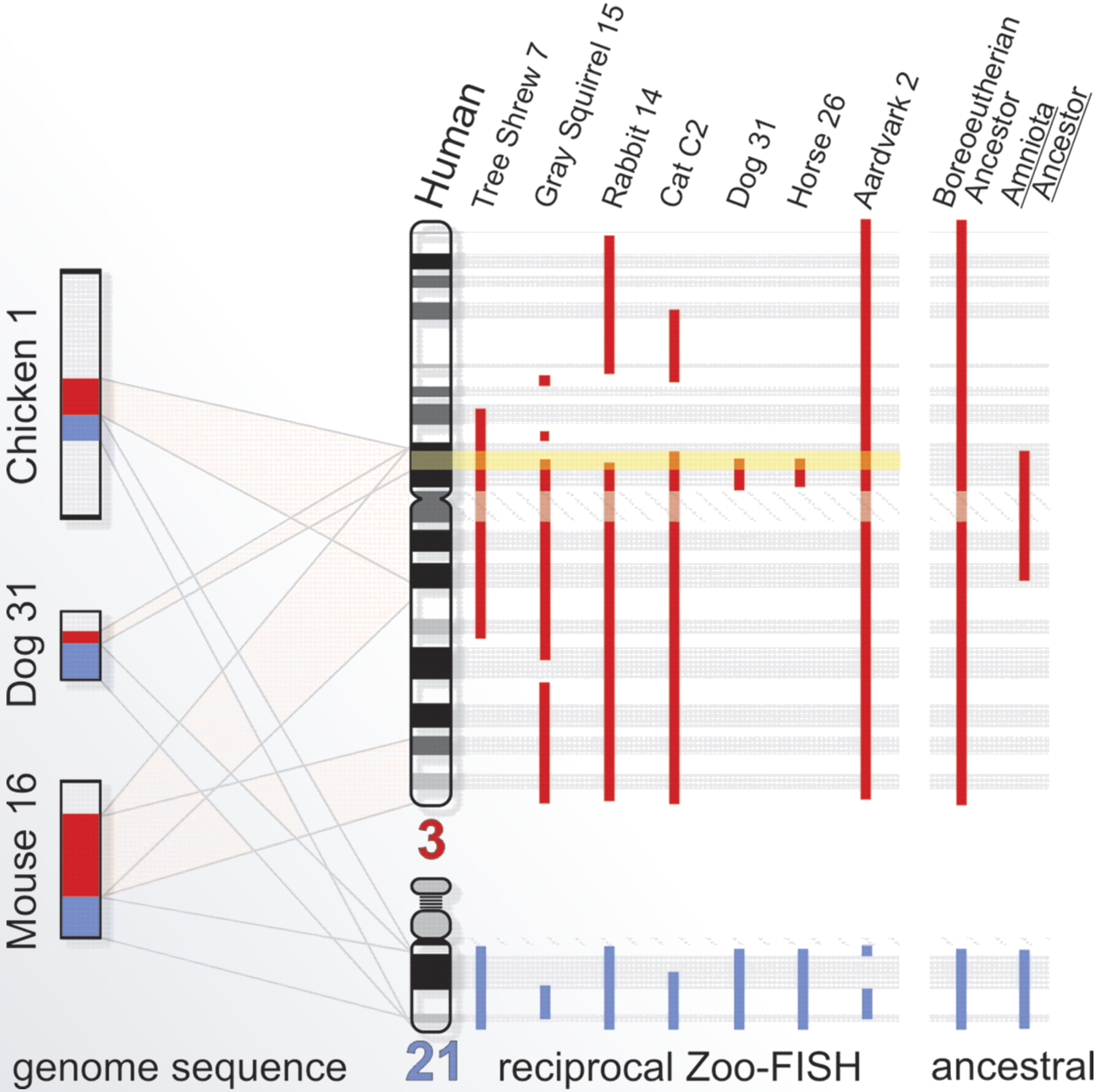
Chromosomes

Human vs Rat synteny



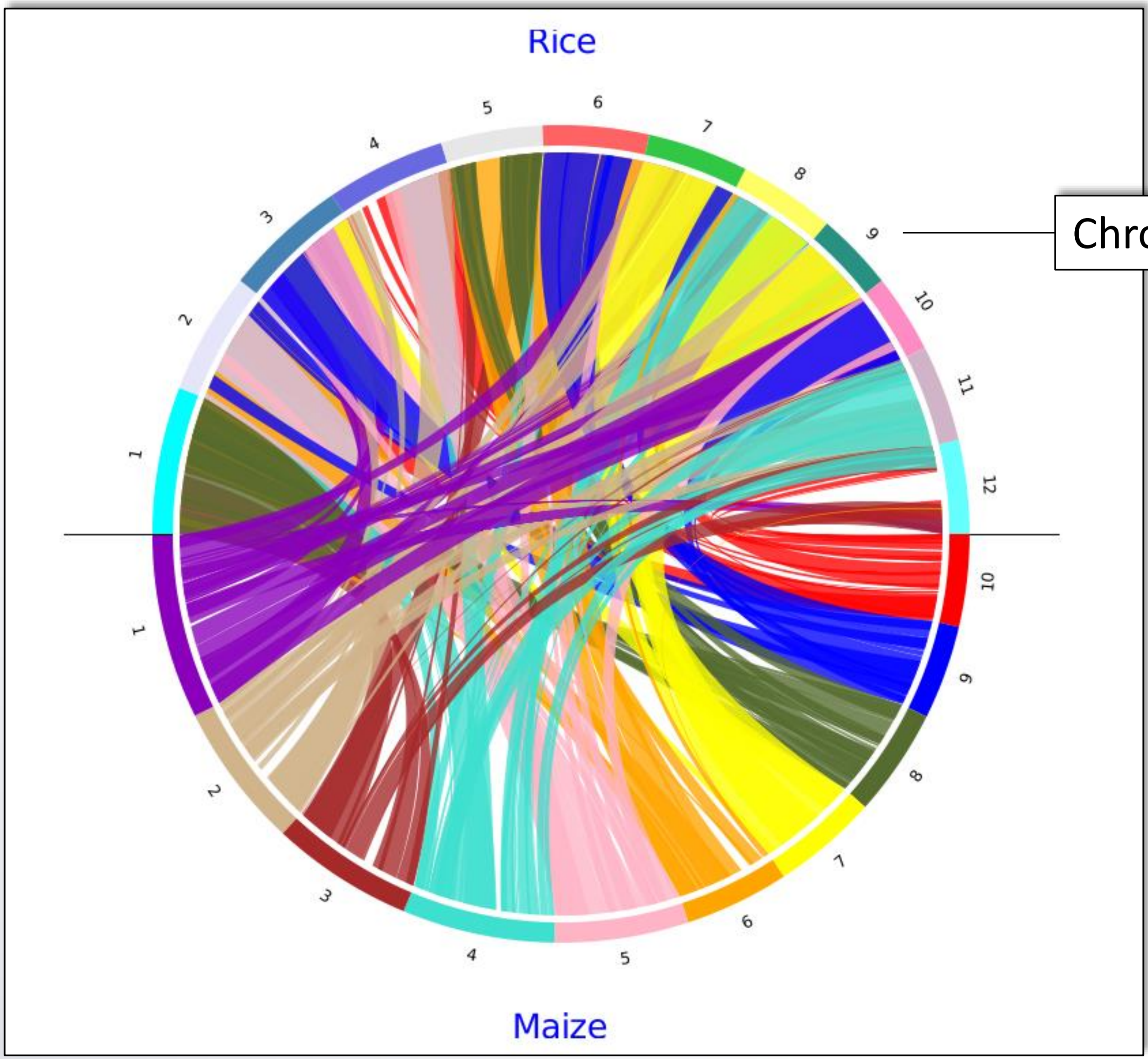
Vitt *et al.* 2004 Genome Research, 14(4), 640-650.

Human vs many species synteny



Froenicke *et al.* 2006 Genome Research, 16(3), 306-310.

Rice vs Maize synteny

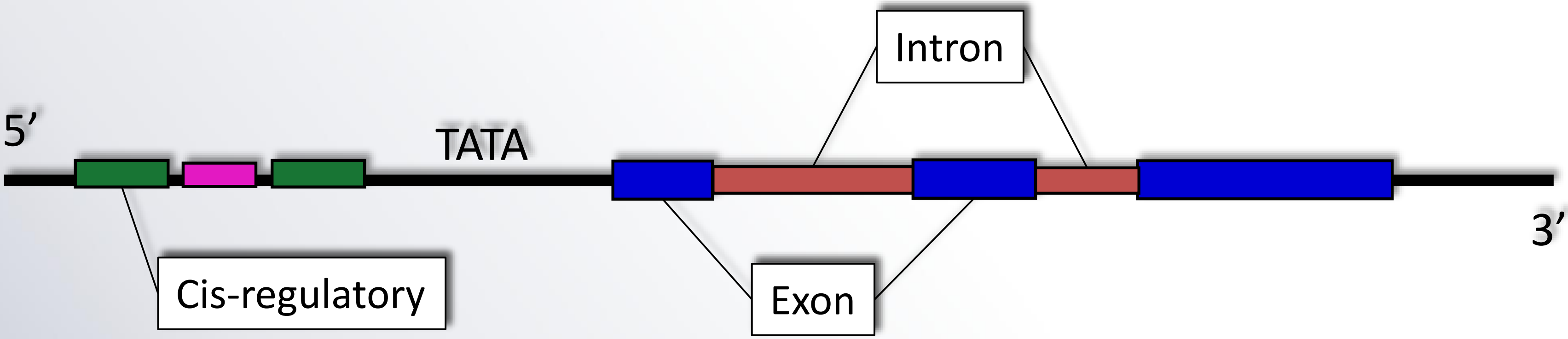


Chromosomes

Soderlund *et al.* 2011 *Nucleic Acids Research*, 39(10), e68-e68.

The objectives of CG

- **Comparison** of **genomic sequences** from **different species** can help identify the following:
  - **Gene structure** (Exon, Intron, 5’UTR, 3’UTR)
  - **Gene function** (Metabolism, Binding, etc)
  - **Regulatory sequences** (Promoters, Enhancers, etc)



How to create a CGs analysis – molecular phylogeny

- The use of **molecular data** to establish the **relationship** between **species**, **organisms** or **gene families**

❑ **Homology**

Sequences that **share common ancestry**

Homologous genes can be similar in sequence, but **similar sequences are not necessarily homologous**

❑ **Orthologs**

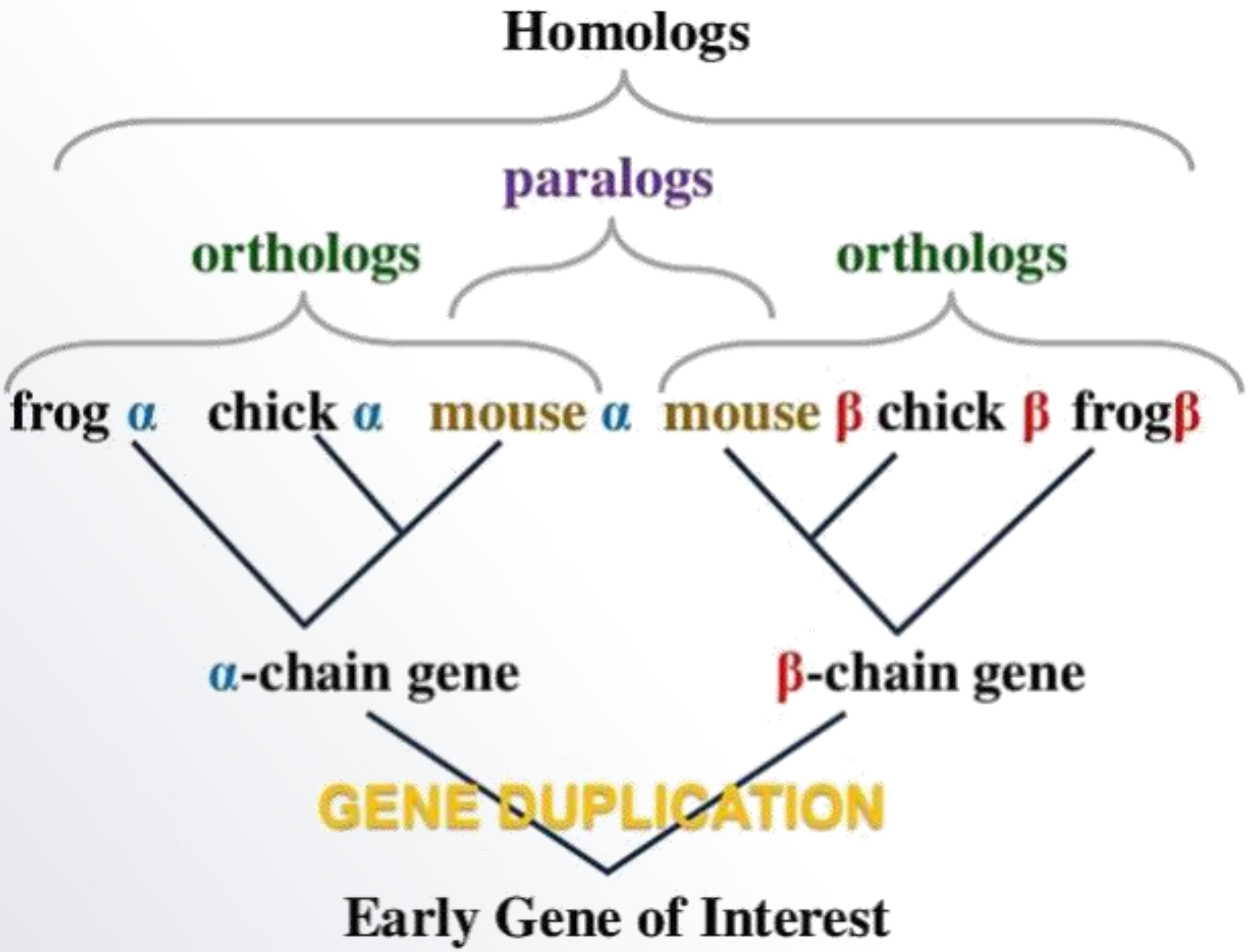
Homologs in **different species** derived by a **speciation** event

❑ **Paralogs**

Homologs in the **same or different species** derived by a **duplication** event

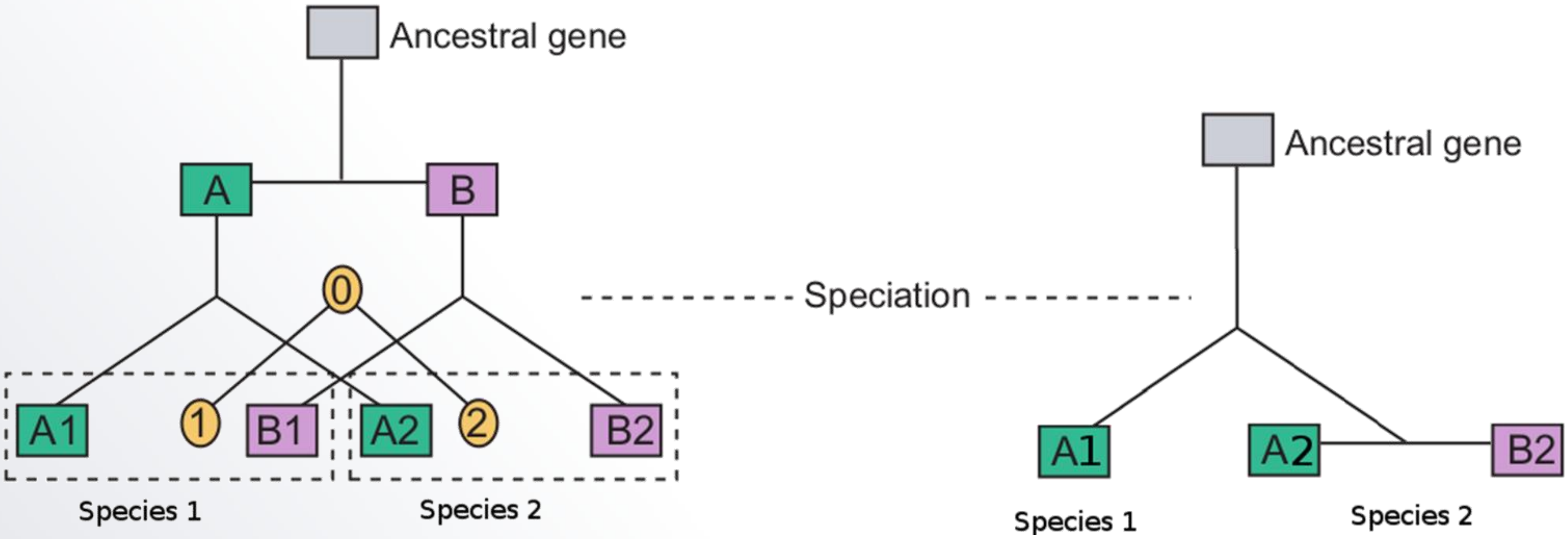
Jensen, R. A. 2001 *Genome Biology,* 2, 1-3

Understanding the concept of homologs



NCBI (1995). Simple NCBI Directory

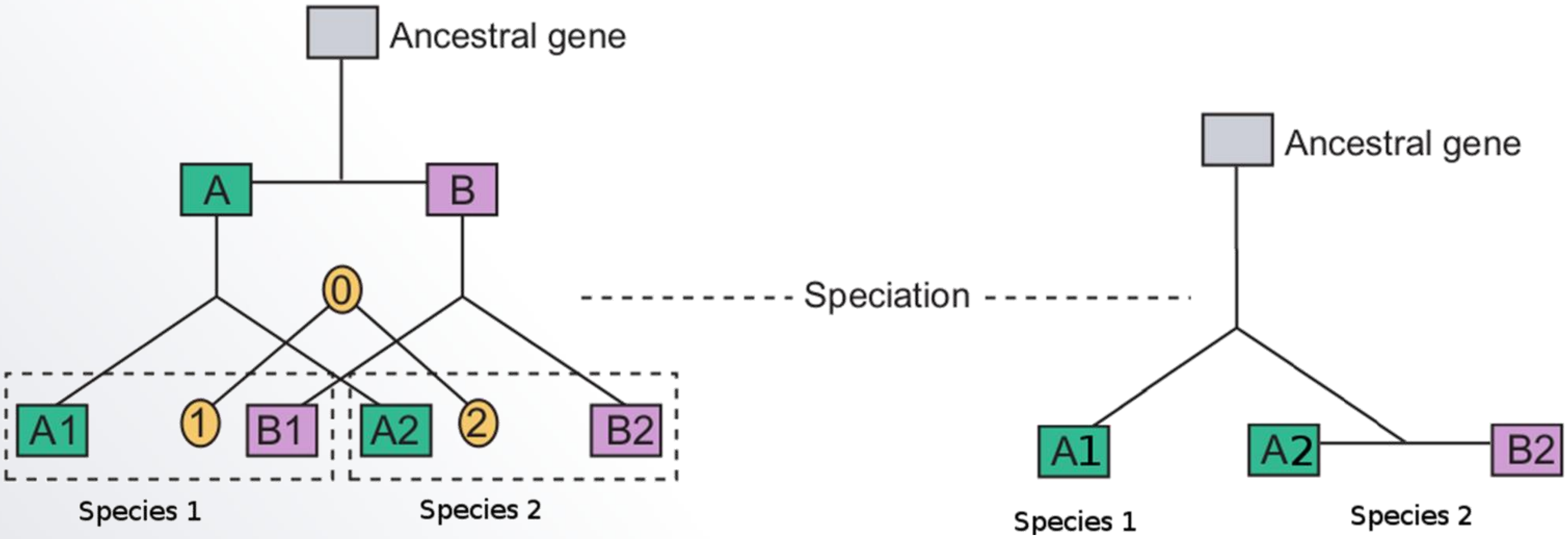
Understanding the concept of homologs



1. A1 and B1 – …
2. A1 and B2 – …
3. A2 and B1 – …
4. A2 and B2 – …
5. A1 and A2 – …
6. B1 and B2 – …

Jensen, R. A. 2001 *Genome Biology,* 2, 1-3

Understanding the concept of homologs

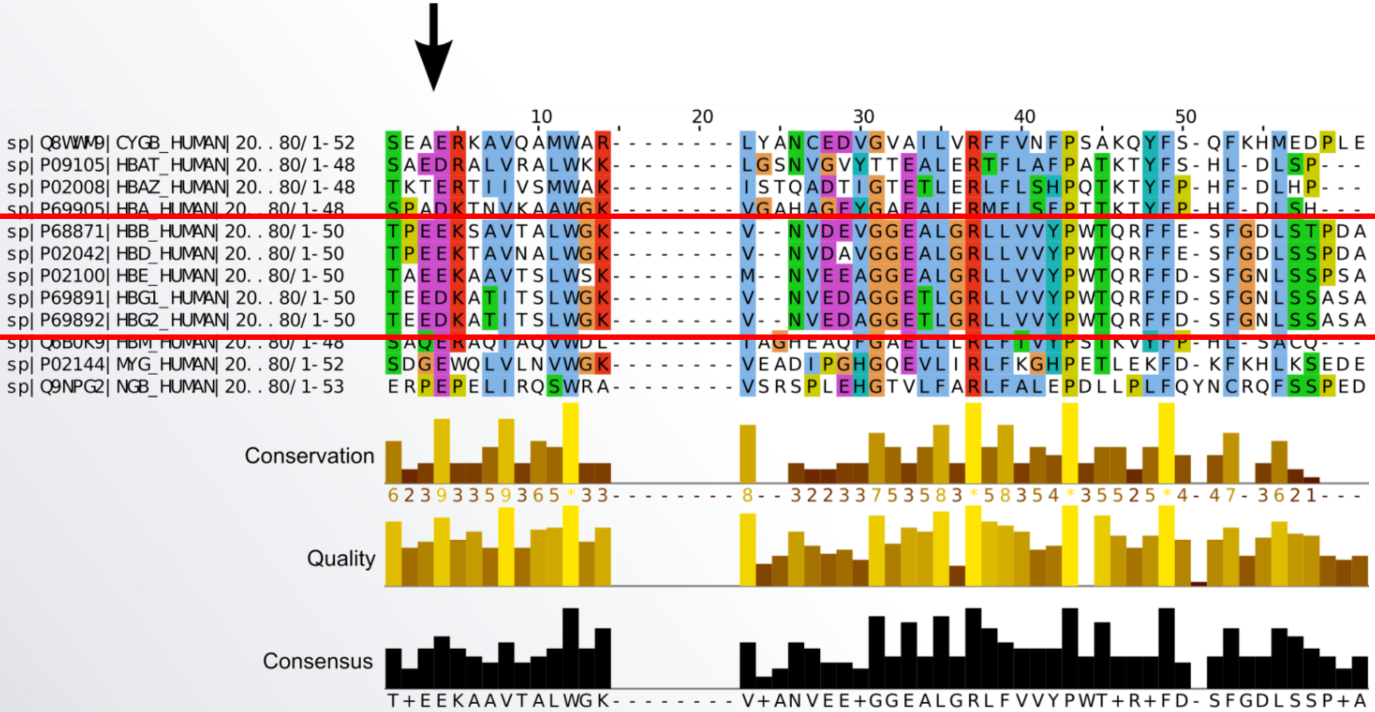


1. A1 and B1 – **paralog**
2. A1 and B2 – **paralog**
3. A2 and B1 – **paralog**
4. A2 and B2 – **paralog**
5. A1 and A2 – **ortholog**
6. B1 and B2 – **ortholog**

Jensen, R. A. 2001 *Genome Biology,* 2, 1-3

Understand the concept of conserved region

Conserved sequences are **similar** or **identical** sequences in nucleic acids (DNA and RNA) or proteins **across species** (**orthologous** sequences) or within a genome (**paralogous** sequences).
Conservation indicates that a sequence has been **maintained by natural selection**.



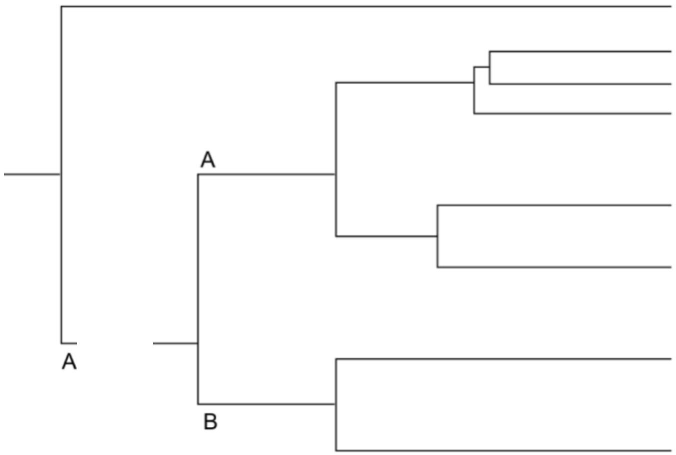
```
                                        10                   20                  30                  40                  50
sp|Q8WWM9|CYGB_HUMAN|20..80/1-52   SEAERKAVQAMWAR--------LYANCEDVGVAILVRFFVNFPSAKQYFS-QFKHMEDPLE
sp|P09105|HBAT_HUMAN|20..80/1-48   SAEDRALVRALWKK--------LGSNVGVYTTEALERTFLAFPATKTYFS-HL-DLSP---
sp|P02008|HBAZ_HUMAN|20..80/1-48   TKTERTIIVSMWAK--------ISTQADTIGTETLERLFLSHPQTKTYFP-HF-DLHP---
sp|P69905|HBA_HUMAN|20..80/1-48    SPADKTNVKAAWGK--------VGAHAGEYGAEALERMFLSFPTTKTYFP-HF-DLSH---
sp|P68871|HBB_HUMAN|20..80/1-50    TPEEKSAVTALWGK--------V--NVDEVGGEALGRLLVVYPWTQRFFE-SFGDLSTPDA
sp|P02042|HBD_HUMAN|20..80/1-50    TPEEKTAVNALWGK--------V--NVDAVGGEALGRLLVVYPWTQRFFE-SFGDLSSPDA
sp|P02100|HBE_HUMAN|20..80/1-50    TAEEKAAVTSLWSK--------M--NVEEAGGEALGRLLVVYPWTQRFFD-SFGNLSSPSA
sp|P69891|HBG1_HUMAN|20..80/1-50   TEEDKATITSLWGK--------V--NVEDAGGETLGRLLVVYPWTQRFFD-SFGNLSSASA
sp|P69892|HBG2_HUMAN|20..80/1-50   TEEDKATITSLWGK--------V--NVEDAGGETLGRLLVVYPWTQRFFD-SFGNLSSASA
sp|Q6B0K9|HBM_HUMAN|20..80/1-48    SAQERAQIAQVWDL--------TAGHEAQFGAELLLRLFTVYPSTKVYFP-HL-SACQ---
sp|P02144|MYG_HUMAN|20..80/1-52    SDGEWQLVLNVWGK--------VEADIPGHGQEVLIRLFKGHPETLEKFD-KFKHLKSEDE
sp|Q9NPG2|NGB_HUMAN|20..80/1-53    ERPEPELIRQSWRA--------VSRSPLEHGTVLFARLFALEPDLLPLFQYNCRQFSSPED

Conservation                       62393359365*33--------8--3223375358 3*58354*35525*4-47-3621---

Quality

Consensus                          T+EEKAAVTALWGK--------V+ANVEE+GGEALGRLFVVYPWT+R+FD-SFGDLSSP+A
```

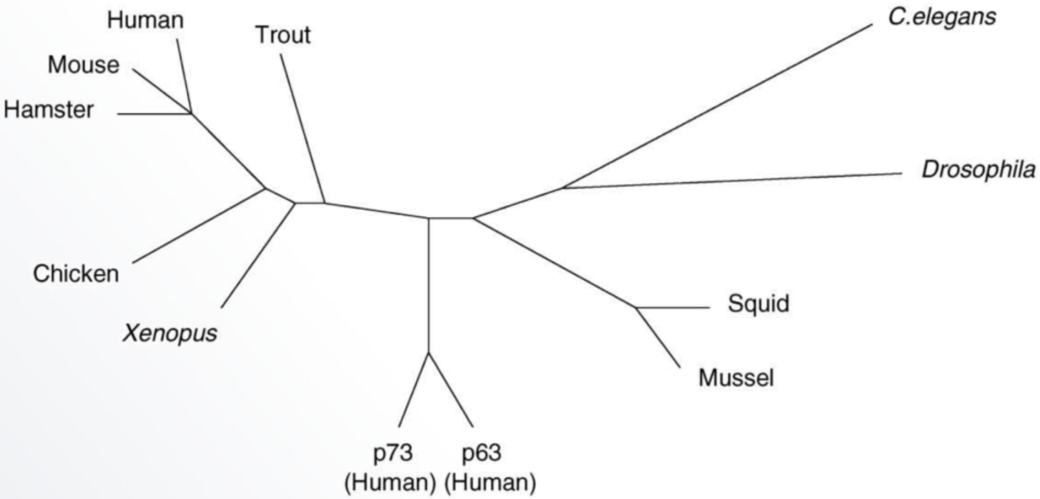
Visualization of CGs – phylogenetic tree

- ❑ Visualize **evolutionary relationships** between **species** and **genes/proteins**

- ❑ **Rooted tree**
  - Order of evolutionary events

- ❑ **Unrooted tree**
  - Evolutionary relationships between descendants





Visualization of CGs – Dot Plot

❑ A **graphical method** for **comparing two biological sequences/genomes** and identifying **regions of close similarity**

❑ **Synteny**

Gene loci are on the same chromosome

❑ **Conserved synteny**

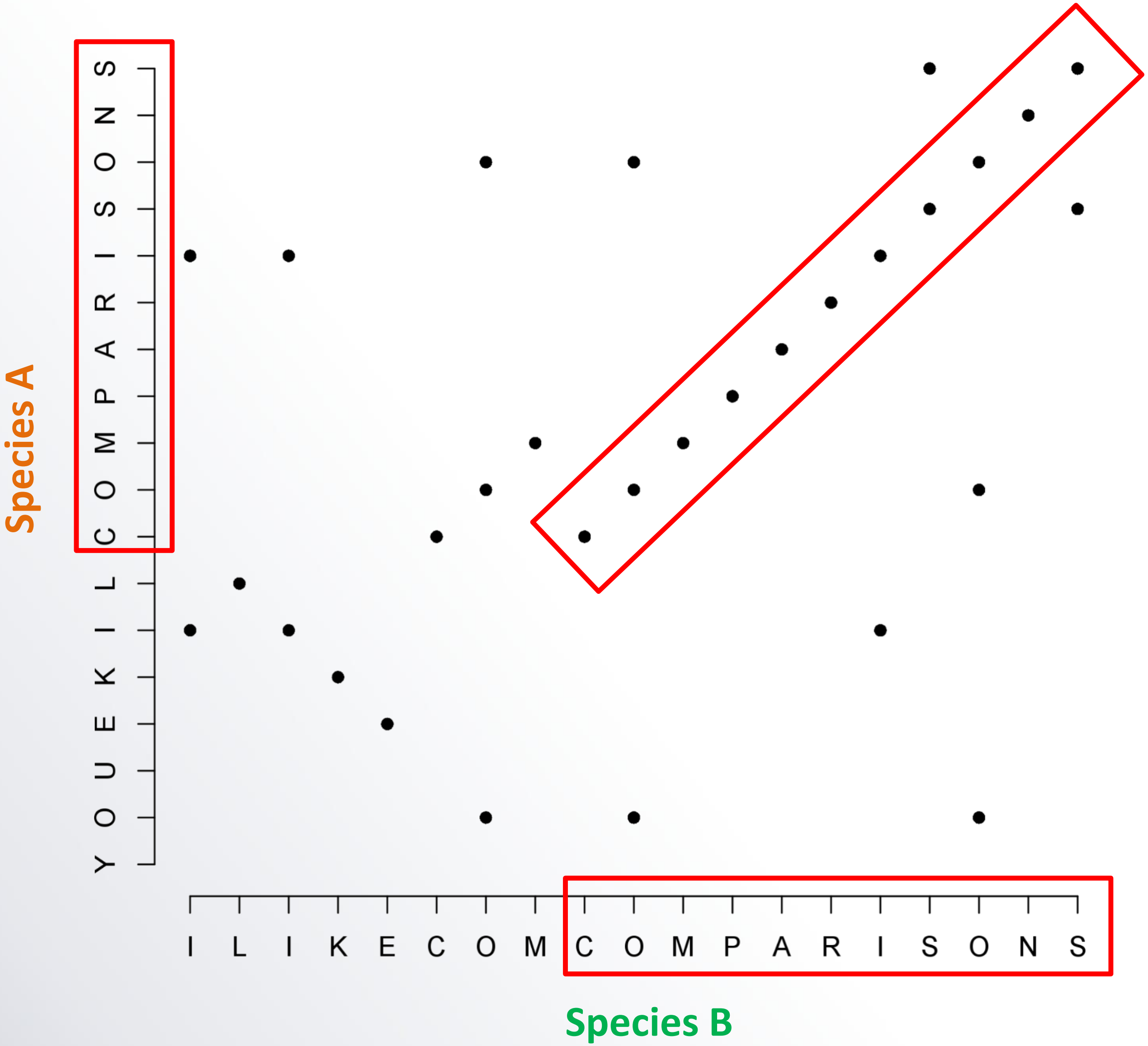
Gene loci are on the same chromosome in different species

❑ **Collinearity**

The order of the gene loci is preserved across species

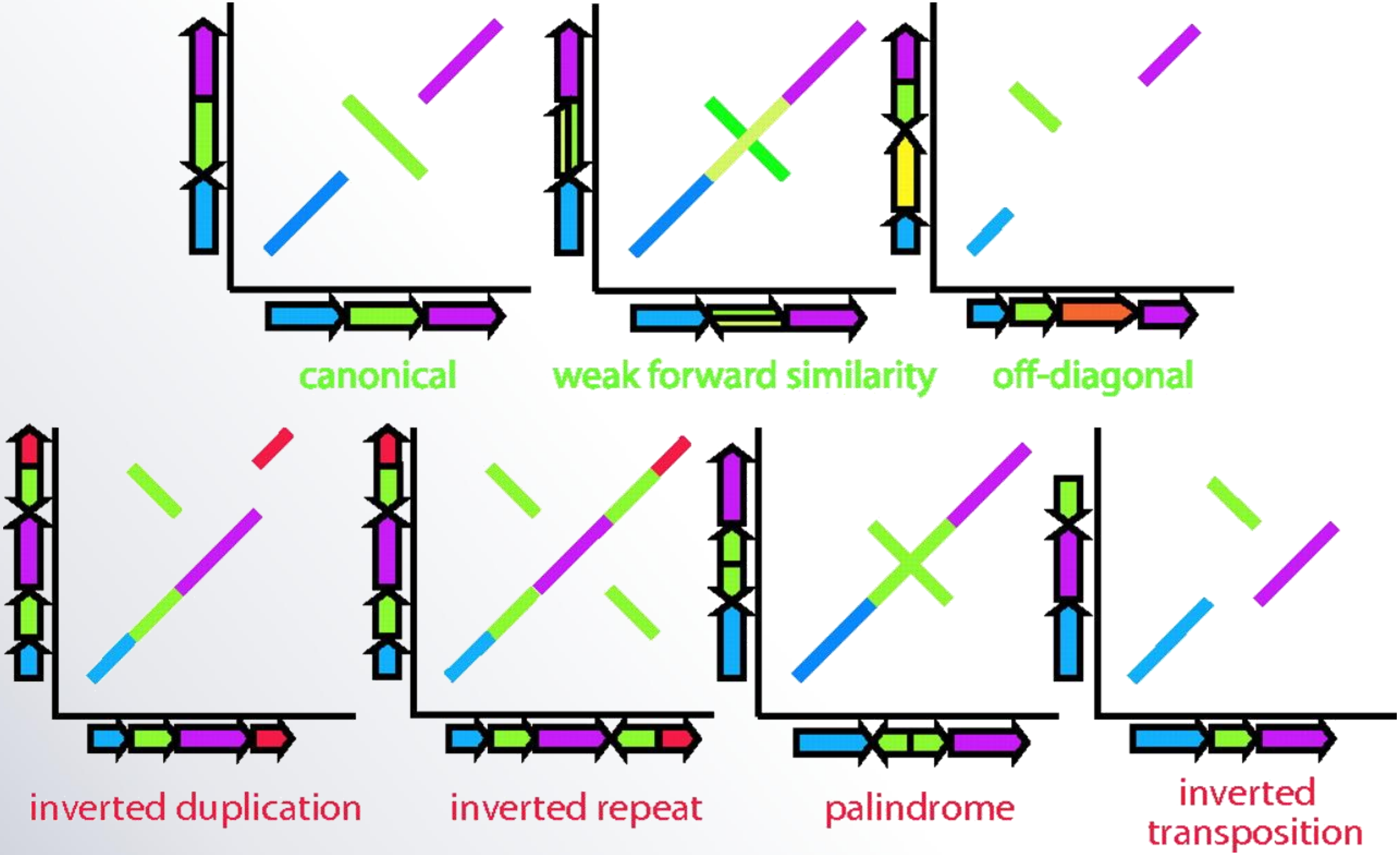
Visualization of CGs – Dot Plot

Match chromosome sequence from species A to species B



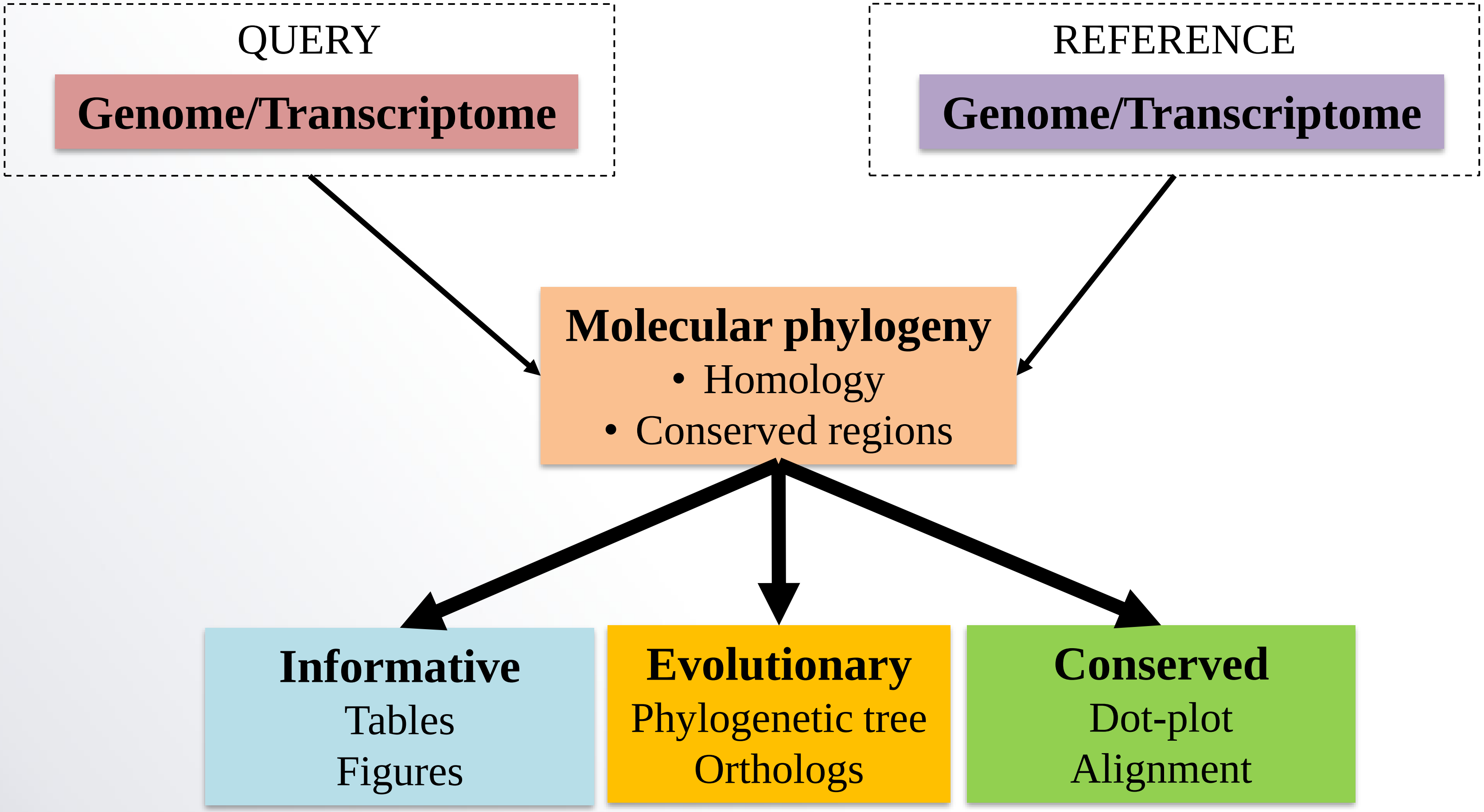
Visualization of CGs – Dot Plot

Different pattern in the genome to genome comparison



Chaisson *et al.* 2006 *PNAS* **103** 19824-9

Principals of CGs



QUERY

Genome/Transcriptome

REFERENCE

Genome/Transcriptome

Molecular phylogeny

- Homology
- Conserved regions

Informative

Tables

Figures

Evolutionary

Phylogenetic tree

Orthologs

Conserved

Dot-plot

Alignment

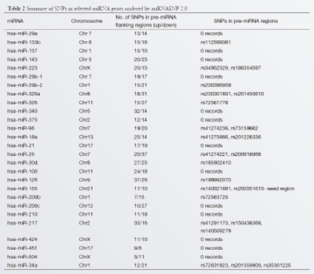
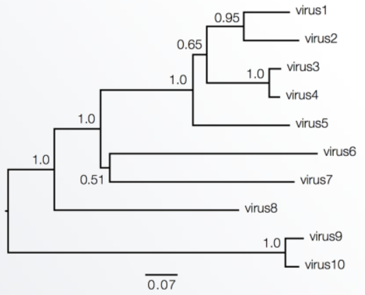
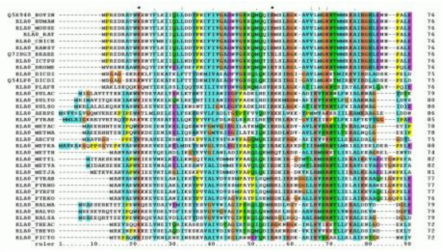


Table 2 Summary of SNPs in selected miRNA genes analyzed by miRNASNP 2.0

miRNA	Chromosome	No. of SNPs in pre-miRNA flanking regions (up/down)	SNPs in pre-miRNA regions
hsa-miR-29a	Chr 7	13/14	0 records
hsa-miR-133b	Chr 6	15/16	rs112599081
hsa-miR-137	Chr 1	15/10	0 records
hsa-miR-143	Chr 5	20/23	0 records
hsa-miR-223	ChrX	20/13	rs34952329, rs186354597
hsa-miR-29b-1	Chr 7	19/17	0 records
hsa-miR-29b-2	Chr1	15/21	rs200396959
hsa-miR-320a	Chr8	18/31	rs200301891, rs201450610
hsa-miR-326	Chr11	15/27	rs72561778
hsa-miR-340	Chr5	32/14	0 records
hsa-miR-375	Chr2	12/14	0 records
hsa-miR-96	Chr7	19/20	rs41274239, rs73159662
hsa-miR-18a	Chr13	25/14	rs41275866, rs201226336
hsa-miR-21	Chr17	17/19	0 records
hsa-miR-25	Chr7	20/37	rs41274221, rs200616968
hsa-miR-30d	Chr8	27/25	rs185802410
hsa-miR-100	Chr11	24/18	0 records
hsa-miR-126	Chr9	37/26	rs199992070
hsa-miR-155	Chr21	17/10	rs140021681, rs200351615- seed region
hsa-miR-200b	Chr1	7/15	rs72563729
hsa-miR-200c	Chr12	10/27	0 records
hsa-miR-210	Chr11	11/18	0 records
hsa-miR-217	Chr2	33/16	rs41291173, rs150436368, rs140509279
hsa-miR-424	ChrX	11/10	0 records
hsa-miR-451	Chr17	9/6	0 records
hsa-miR-504	ChrX	5/11	0 records
hsa-miR-34a	Chr1	12/21	rs72631823, rs201359909, rs35301225

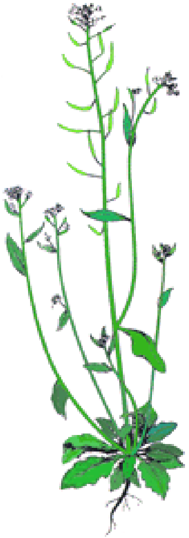




What do we need to do a CGs?

- Genome/transcriptome of **a query/target species**
- Genome/transcriptome of **a reference species**
- **Bioinformatics tools**:
  - ❑ MEGA-BLAST
  - ❑ Multiple Sequence Alignment
  - ❑ Annotator
  - ❑ Tree builder

NCBI



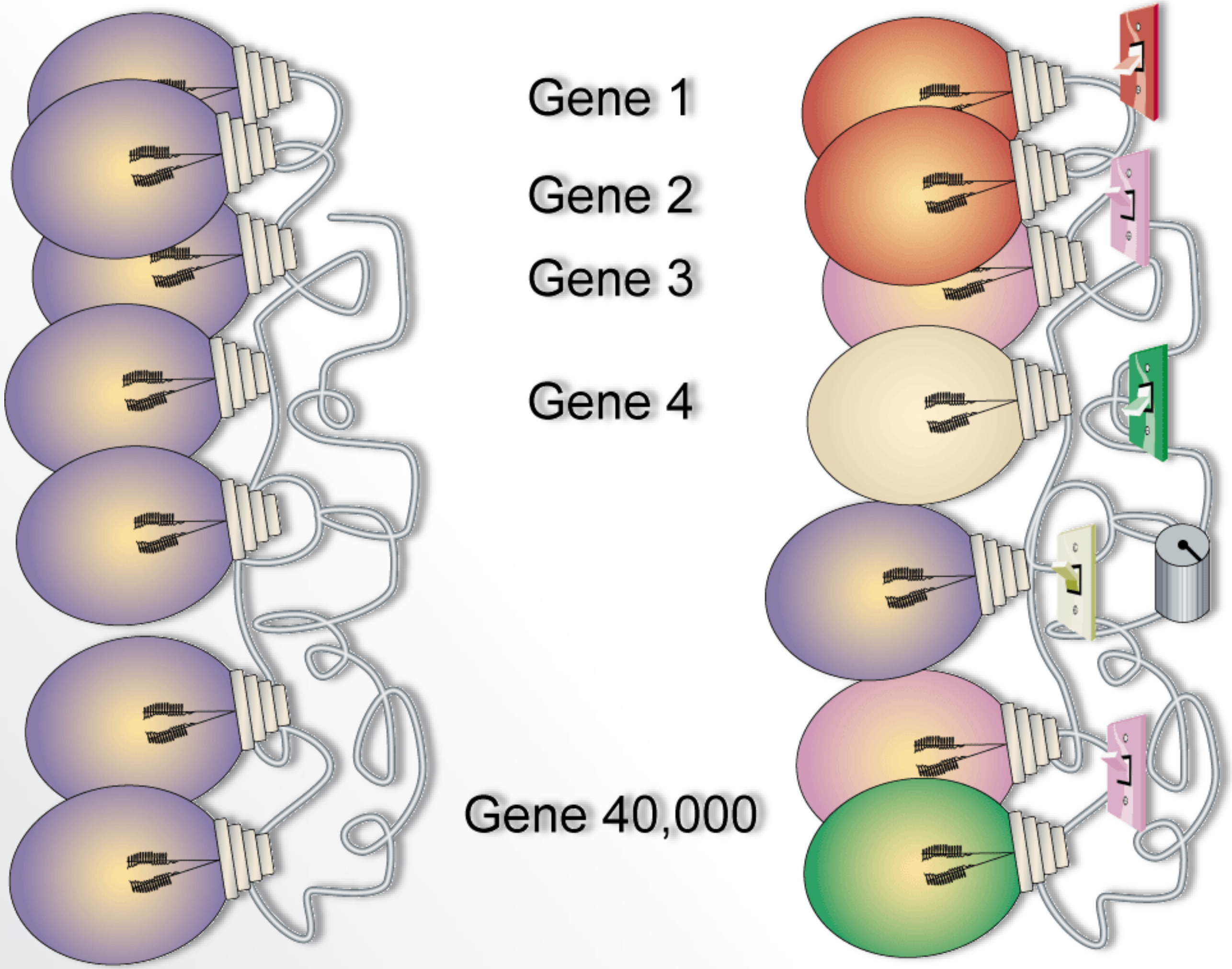
tair

Gramene

JGI

JOINT GENOME INSTITUTE
DEPARTMENT OF ENERGY

Conclusion: the goal of CG



Unknown species

Model species

It was still the fifth course, don’t get dizzy yet