
New Gene Evolution Detected by Genomic Computation
Basic Concepts and Examples
Manyuan Long & Liping Wei
VI
I
III
II
BYA
Beijing
Center for Bioinformatics, Peking University

The role bioinformatics can play in biology and medicine

We emphasize a role of common bioinformatics can play in biology and medicine

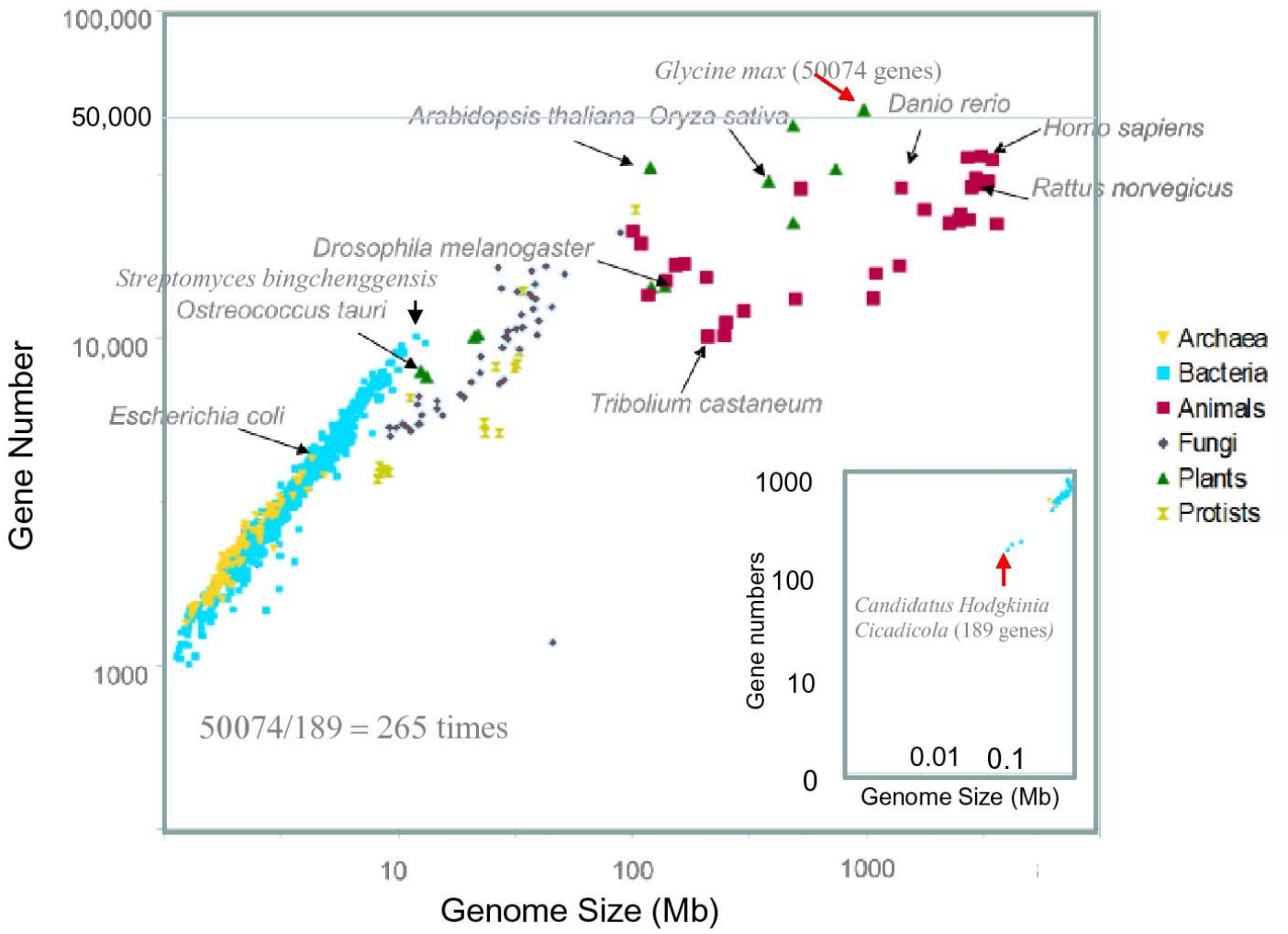
1. Present-day biological and medical studies are in a quick paradigm transition toward genomic analyses in gene identification and expression analysis that created astronomical-scale data.

2. The bioinformatics is a must for data analyses in various levels from preliminary data presentation to advanced interpretation for various scientific problems, with an unprecedented power to detect natural phenomena with the underlying mechanisms.

3. The biological rules and various correlations among the involved factors detected by the bioinformatic analysis from biological and medical studies are illuminating in the progress of understanding basic biological and medical problems.

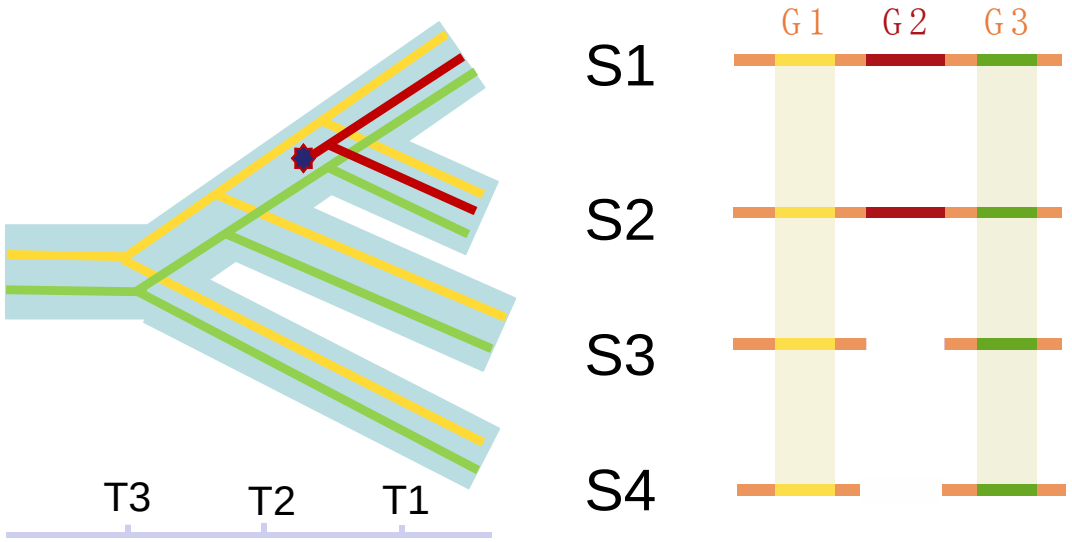
In this class, we are going to apply the bioinformatic analyses to a basic biological problem: the origin and evolution of new genes in a general concept and our understanding of evolution of humans and other mammals. These results are valuable for solving relevant biological and medical problems, exemplified by the case analyses.

New Gene Evolution Added to Genomic Diversity of Organisms



Organisms evolved in number of genes and size of genomes, suggesting a general process of birth and death of genes in evolution

New Gene: Definition for Synten-based Computational Identification



The new gene, G2, that originated in the most recent common ancestor of species S1 and S2 is located between two older genes G1 and G3. Because the divergence time of S1 and S2 is T1, the age of G2 is longer than T1 but shorter than T2, whereas G1 and G3 are older than T3. In general, the units of divergence time are often measured in units of million years ago (MYA).

Long et al., 2003. Annu Rev Genet

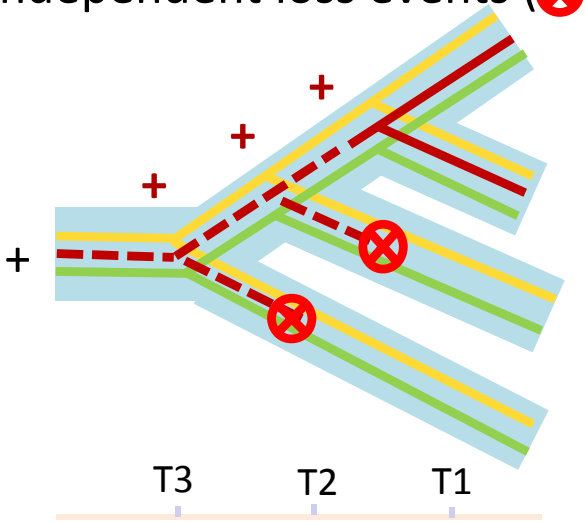
New Gene: Definition for Synten-based Computational Identification

Question: Why do we not define G2 in S3 and S4 as the consequence of gene loss that may have occurred in the ancestor before the divergence of S1 and S2, which may lead to the absence of G2 in S3 and S4?

Solution: the parsimony principle in evolutionary analysis.
The principle of accounting for observations by the hypothesis requiring the fewest or simplest assumptions that lack evidence; in evolution, the principle of invoking the minimal number of evolutionary changes to infer the more likely possibility.

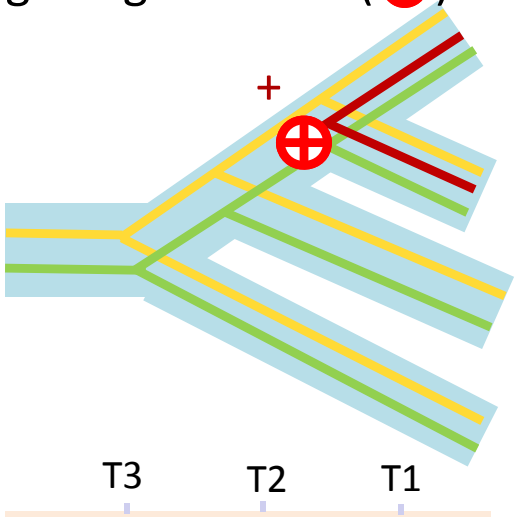
-- Revised from Dauglas J. Futuyma, 2009, Evolution.

2 independent loss events (⊗)



Gene Loss

1 gene gain event (⊕)

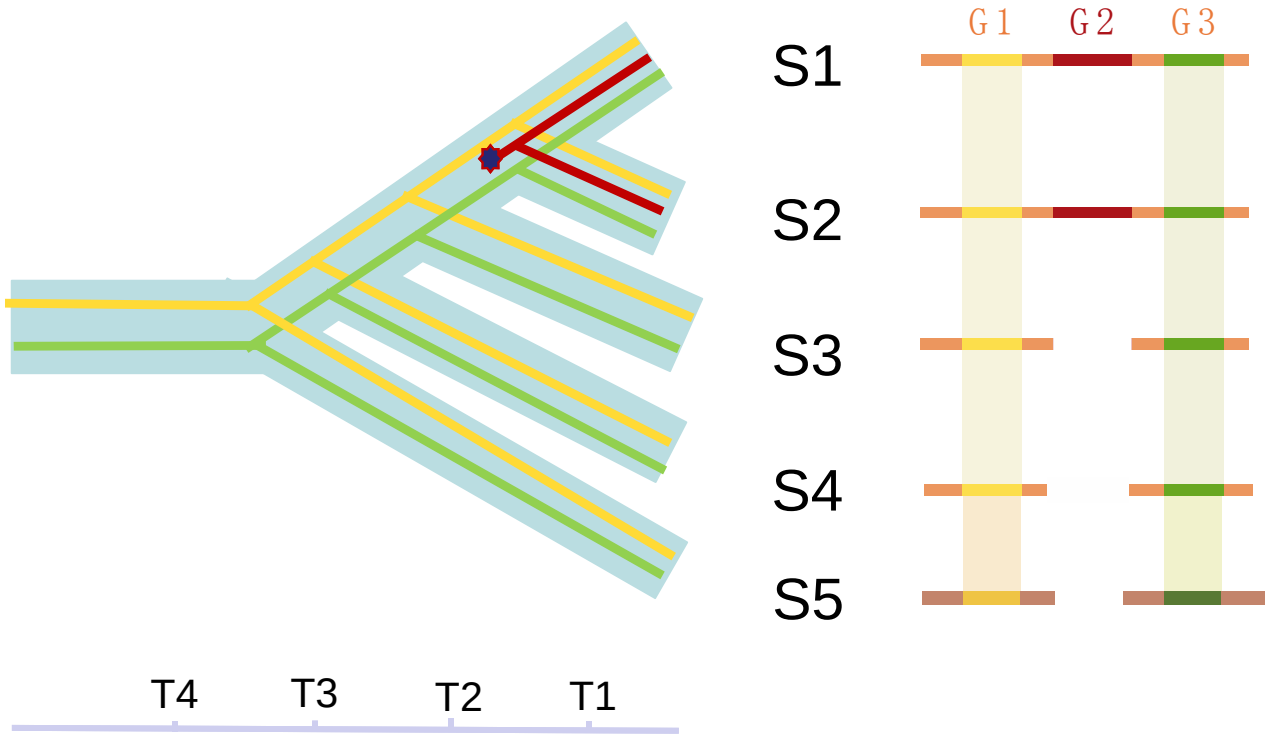


Gene Gain

New Gene: Definition for Synten-based Computational Identification

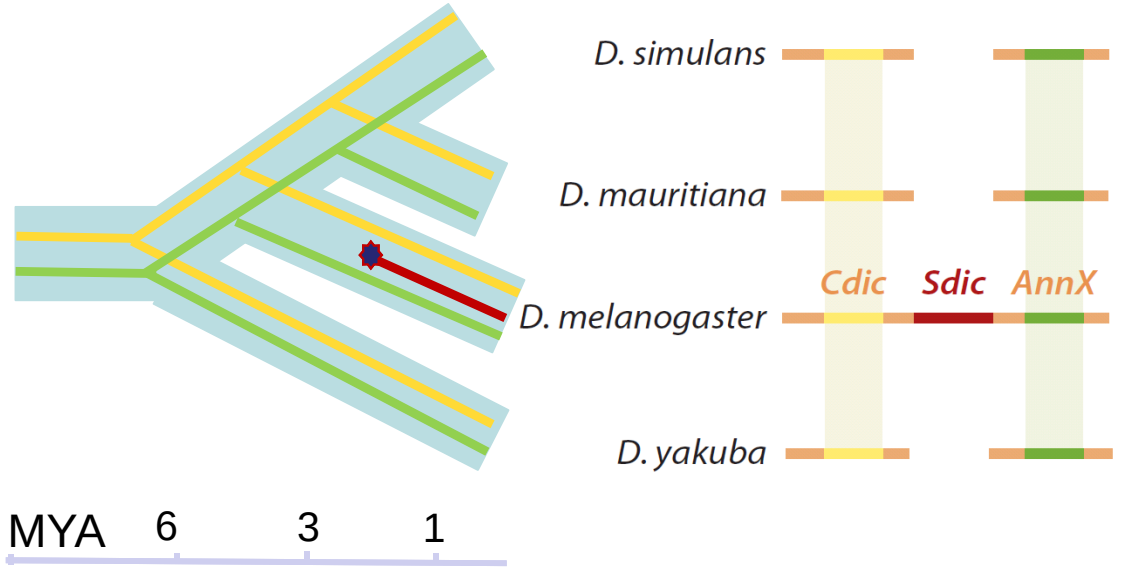
Exercise: Assuming the equal probability of gene gain and loss in each evolutionary change in the process, infer the ancestral state of presence or absence of the gene in T1, T2 and T3 in the two hypotheses of new gene gain or ancestral loss of an old gene. Then, choose the most parsimony hypothesis by calculating the total numbers of evolutionary changes required by the two hypotheses.

Question: in evolutionary analysis, S4 is called the outgroup species that can be used to help infer the ancestral state of G2 at time T2. Repeat the exercise when you add one more outgroup species that also has no G2 and find if you are more confident for our previous inference that G2 is a new gene that originated between T1 and T2, as is shown below:



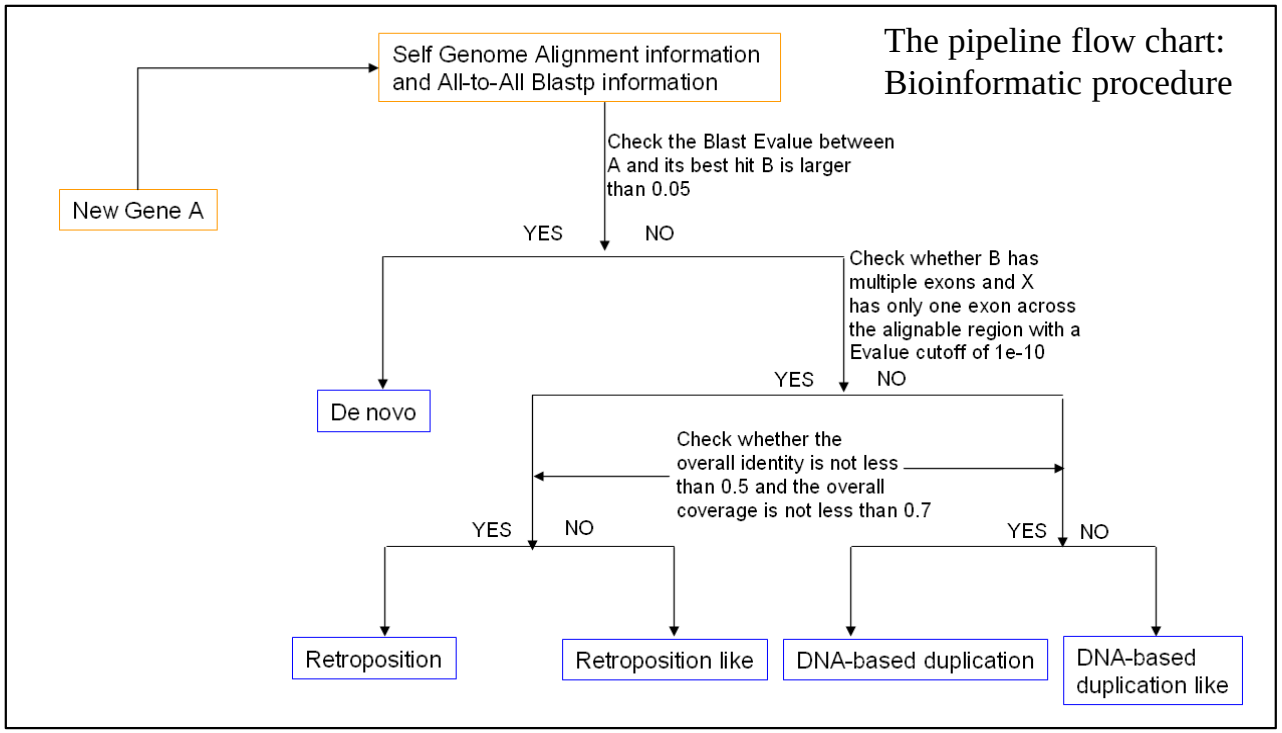
New Gene: An Example for the definition

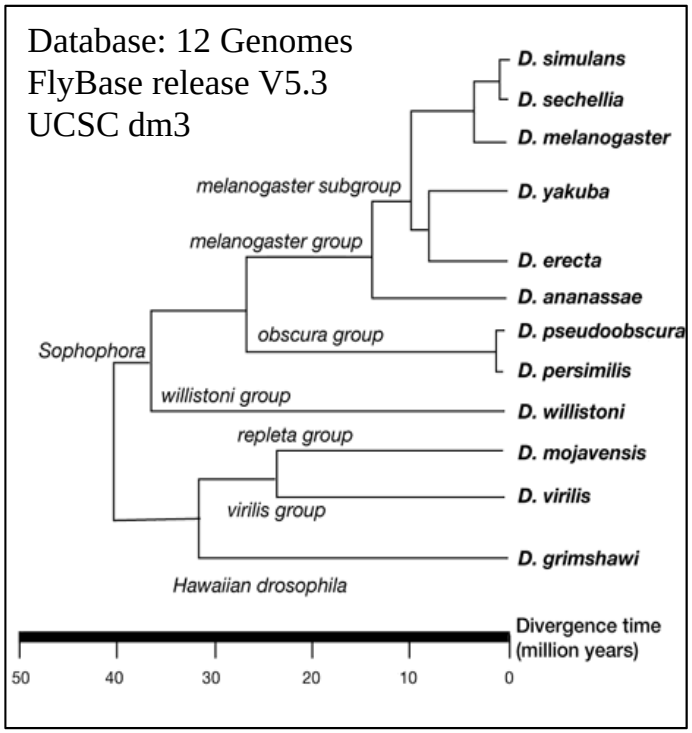
Sdic is a new gene in D. melanogaster that codes for a sperm-specific axonemal dynein subunit, which is immediately flanked by two parental genes, Cdic and Annx.



Long et el, 2003. Annu Rev Genet; Example from: Nurminsky *et al*., 1998, Nature

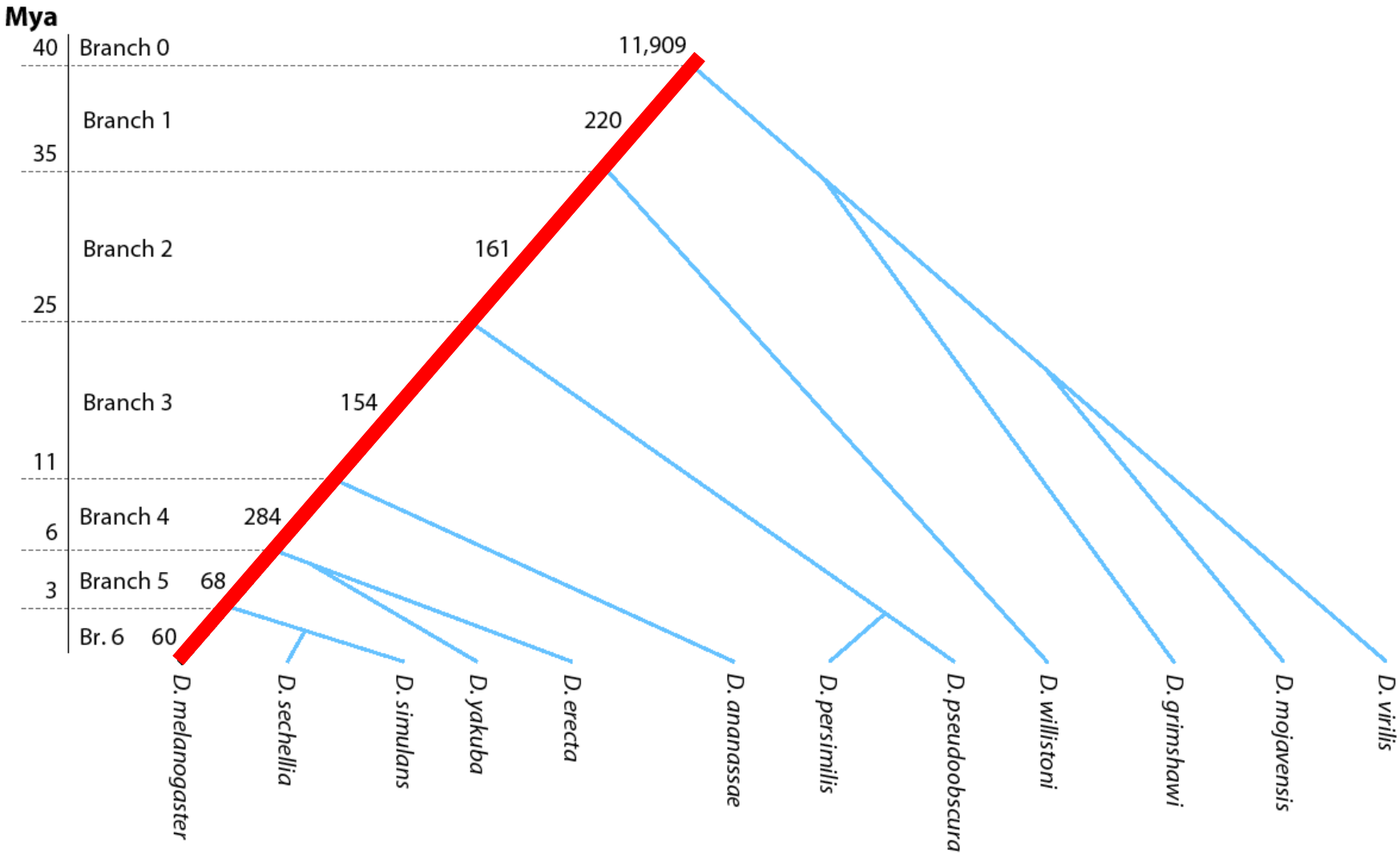
Computational identification of new genes from 12 Drosophila genome sequences





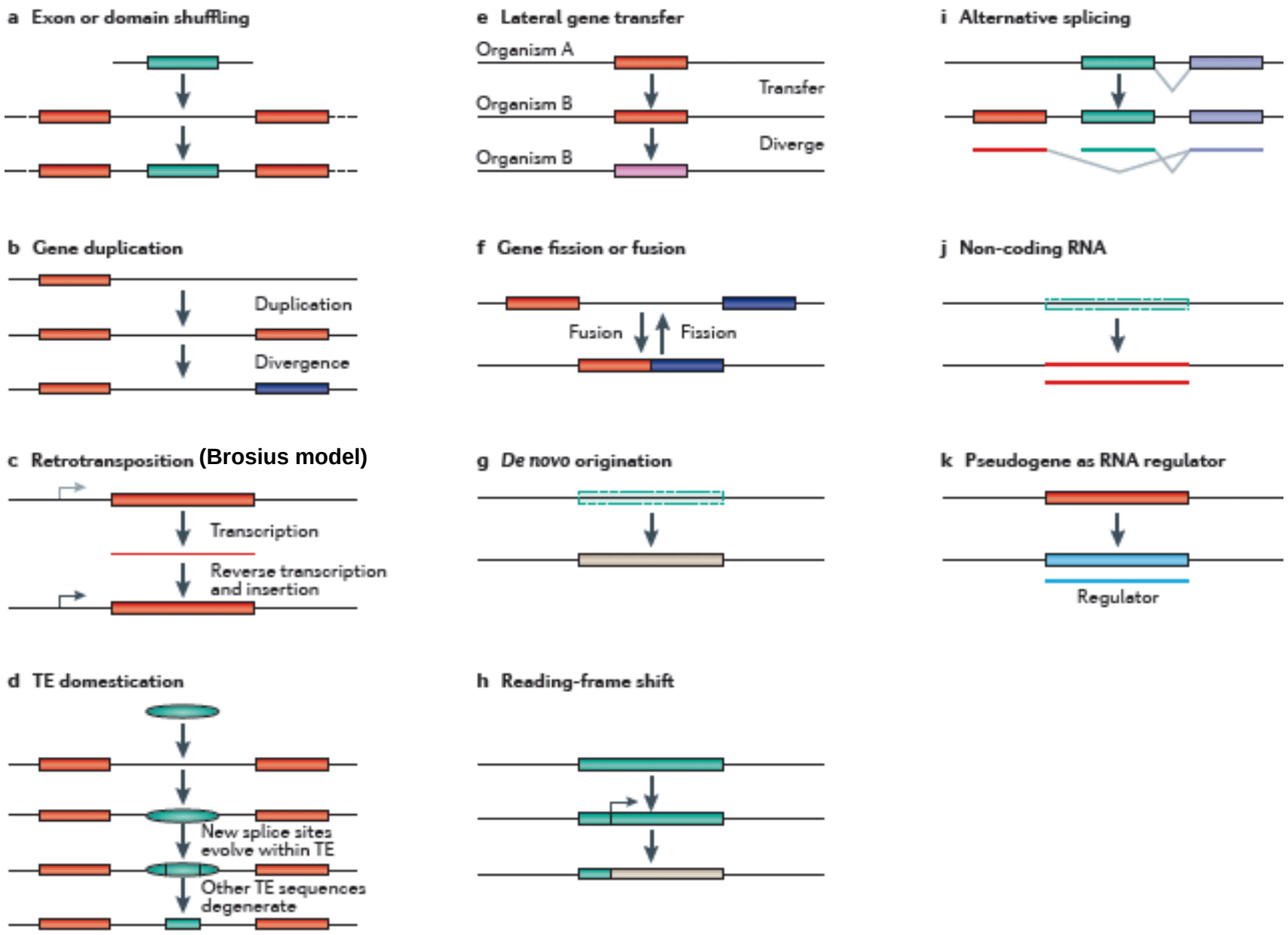
Zhang *et al.*, 2010, Genome Research; *Drosophila* 12 Genomes Consortium, 2007, Nature

New genes distribution mapped in the evolutionary tree of Drosophila



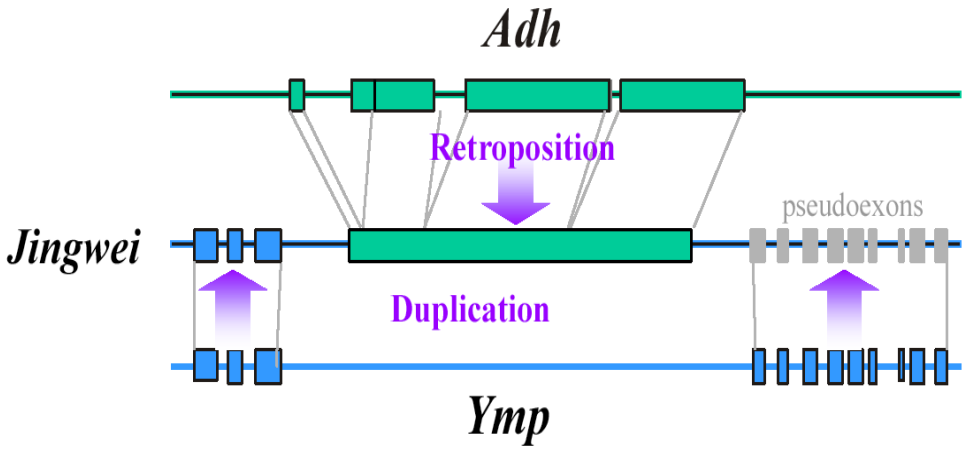
Zhang *et al.*, 2010, Genome Res

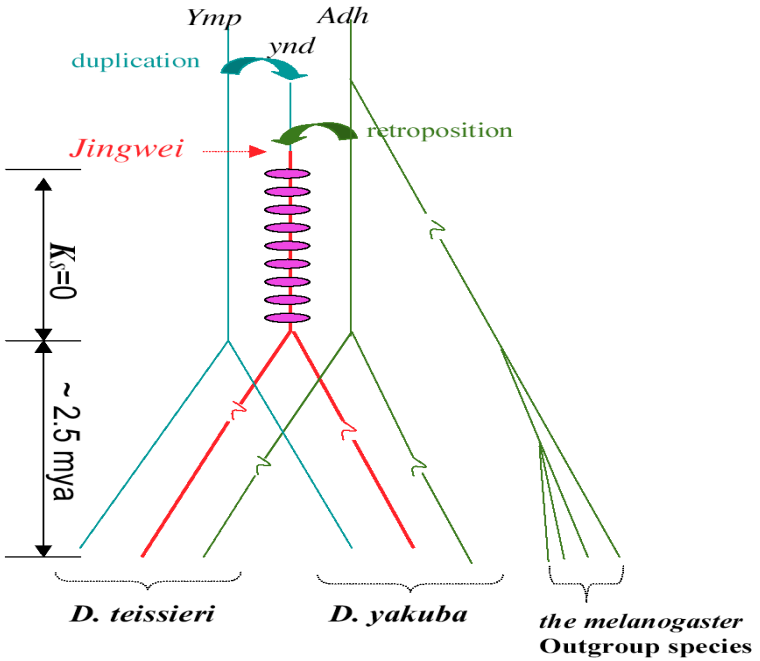
Mechanisms of New Gene Origination



Chen *et al.*, 2013. Nature Rev Genet

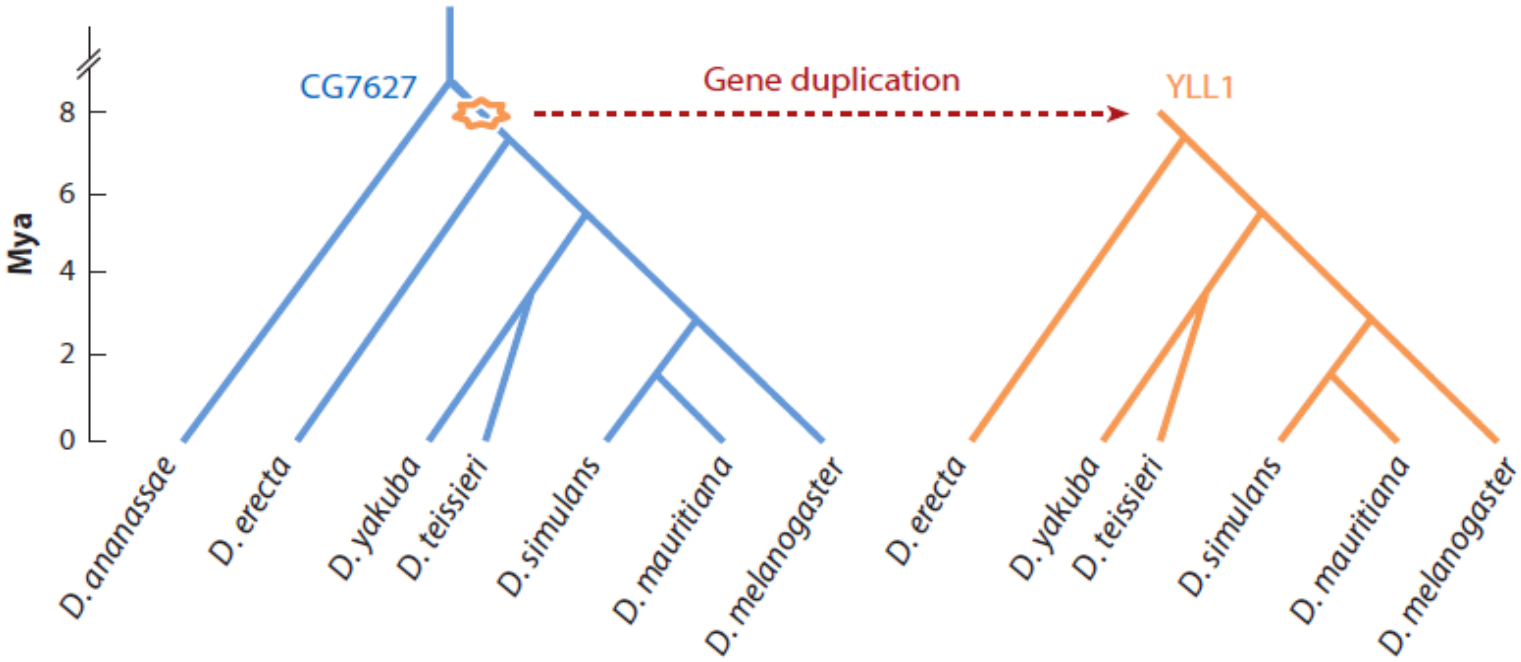
Example from the first observed new gene: the Jingwei in Drosophila that reveal several mechanisms can be involved to generate a new gene





Long and Langley, 1993, Science; Wang *et al.*, 2000, Mol Biol Evol; Zhang *et al.*, 2004, Pro Natl Acad Sci USA

Biological importance of new genes



LINES	METHOD/MUTATION	PHENOTYPES
11343	P-element insertion	Lethal, pupal stage
SH2-0995	EMS/G717S	Lethal, pupal stage
SH2-1101	EMS/T765I	Lethal, pupal stage
SH2-0504	EMS/synonymous	Viable
V39539	RNAi/constitutive Gal4	Lethal, pupal stage
V39540	RNAi/constitutive Gal4	Lethal, pupal stage

Chen *et al.*, 2010, Science; Long *et al.*, 2013, Annu Rev Genet

Biological importance of new genes: Examples for Published New Genes

New genes	Age (million years)	Origination mechanism	Expression	Phenotype	Function	Refs
Drosophila spp.						
Sdic family	0–3	DNA duplication	Testis	Sperm competition	Cytoskeleton	69–71
sphinx	0–3	Retrotransposition	Neuronal and reproductive tissue	Male courtship	ncRNA	77,111
jingwei	0–3	Retrotransposition	Testis	Recruitment pheromone and hormone	Alcohol metabolism	7,17
p24-2	0–3	DNA duplication	Multiple stages and tissues	Development, male reproduction	Protein trafficking	46,47
Xcbp1	3–6	Retrotransposition	Neuronal tissue	Foraging behaviour	Chaperone	76
Mammals						
FGF4	~0.01	Retrotransposition	Distal humerus	Humerus development	FGF signalling	66
SRGAP2C	1.0–3.4	Partial DNA duplication	Brain	Predicted to affect cortex development in a mouse model	Unknown	109, 110
CDC14C	7–12	Retrotransposition	Brain and testis	Unknown	Cell cycle	21
CYPA	<10	Retrotransposition	Unknown	Viral infection	HIV-1 resistance	29
POLDI	2.5–3.5	*De novo* origination	Testis	Knockout reduced testis weight and sperm motility	Unknown	44
TBC1D3	<35	Segmental	Prostate	Insulin modulation	IGF signalling	128
Plants						
CYP98A8	<28	Retrotransposition	Vascular tissue, pistilis, root tip, etc.	Pollen development	Phenolic synthesis	19
CYP84A4	<8	Gene duplication	Stem and seedling	Unknown	Arabiodopyrone biosynthesis	20
CYP98A9	<28	Retrotransposition	Vascular tissue, pistilis, root tip, etc.	Pollen development	Phenolic synthesis	19

Chen *et al.*, 2013, Nature Rev Genet

SUMMARY

1. A new gene is a gene that originated recently in a genome and can be identified by syntenic alignment of genomic sequences from a group of closely species.

2. A number of molecular mechanisms can generate new genes and more than one mechanism can be involved in making one new gene.

3. New genes can be biologically important as old or ancient genes. In fruitflies, essential functions can evolve rapidly any time in evolution.