Gene Annotation and Orthology Assignment for Newly-sequenced Organisms

Tsai-Ming Lu

Assistant Research Scientist

Bioinformatics Core, ICOB

生物資訊與數據分析

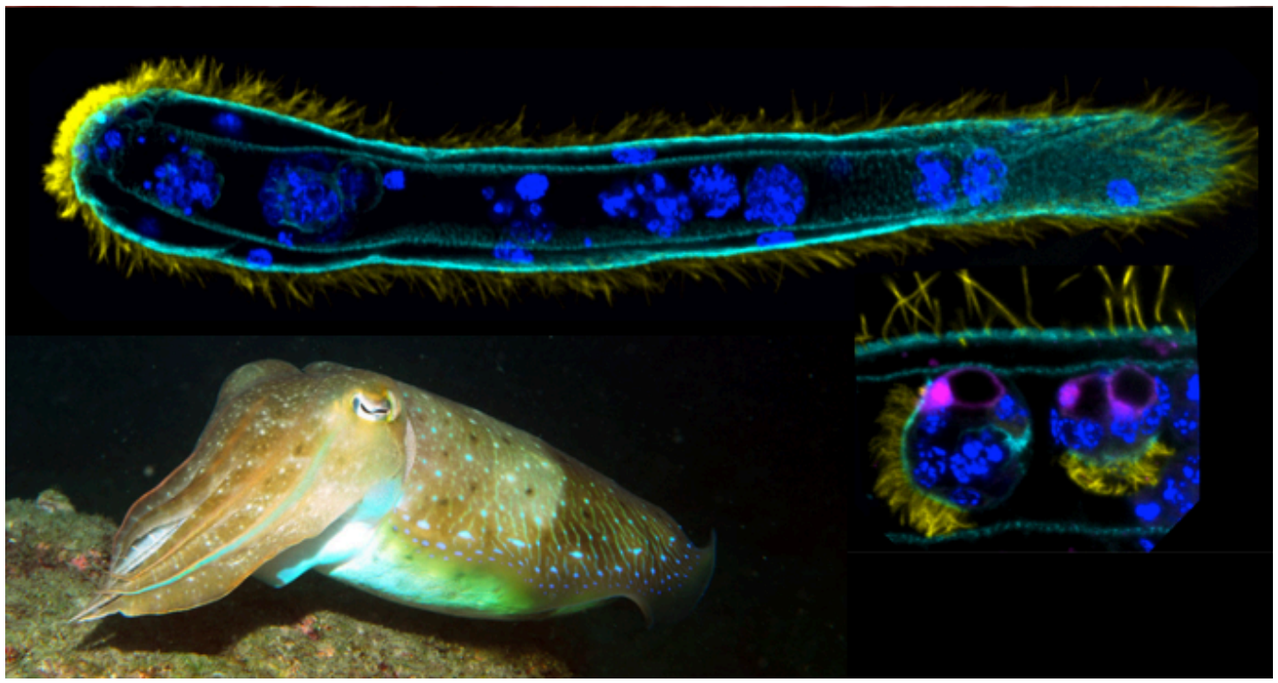
2026/08/25

My background and research topic

ICOB Bioinformatics Core

I am currently employing bioinformatic approaches, combined with my background in evolutionary developmental biology, to study symbiotic relationships in animals.

Dicyemids and cephalopods

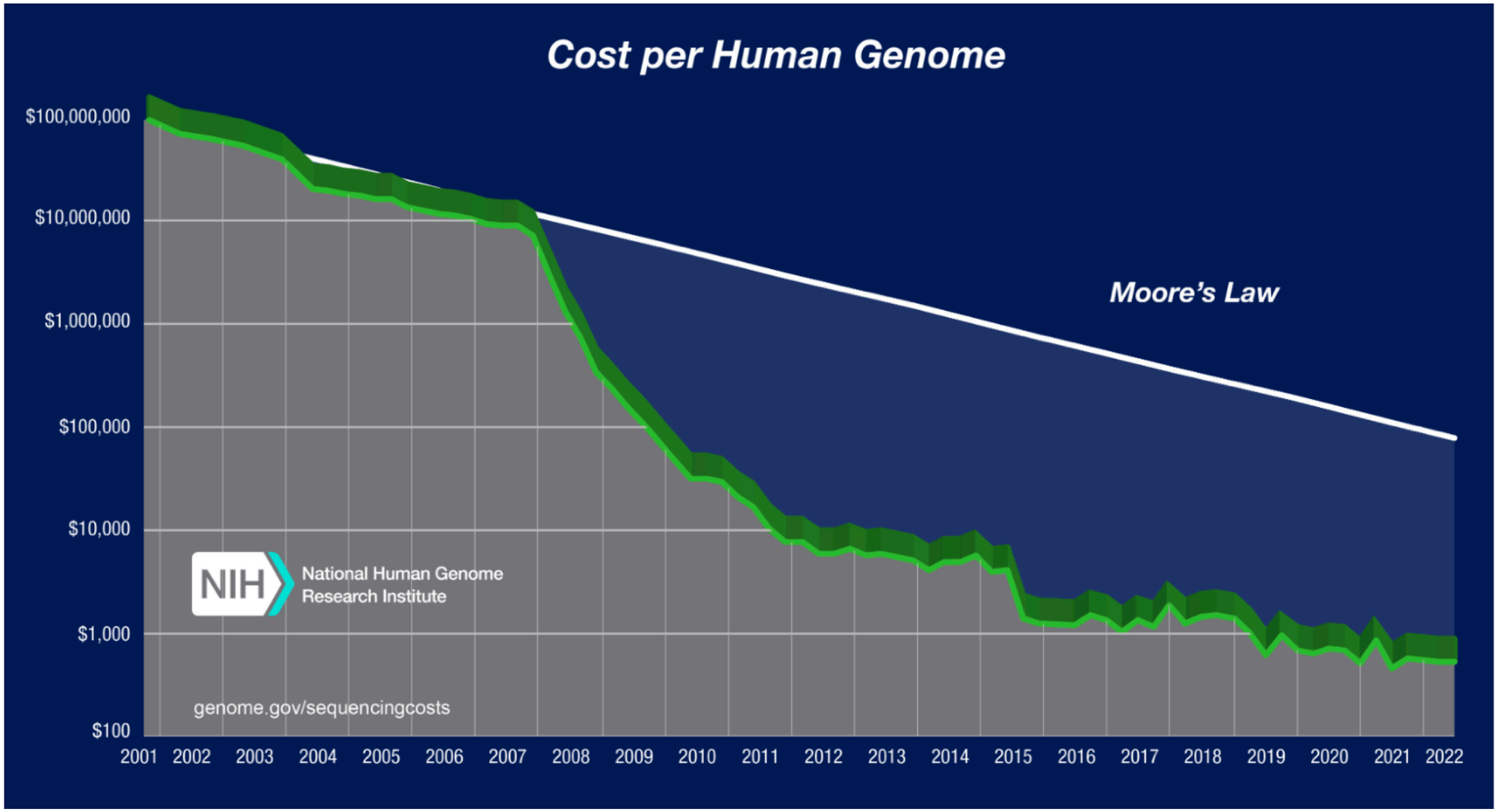


Green algaes and acoels

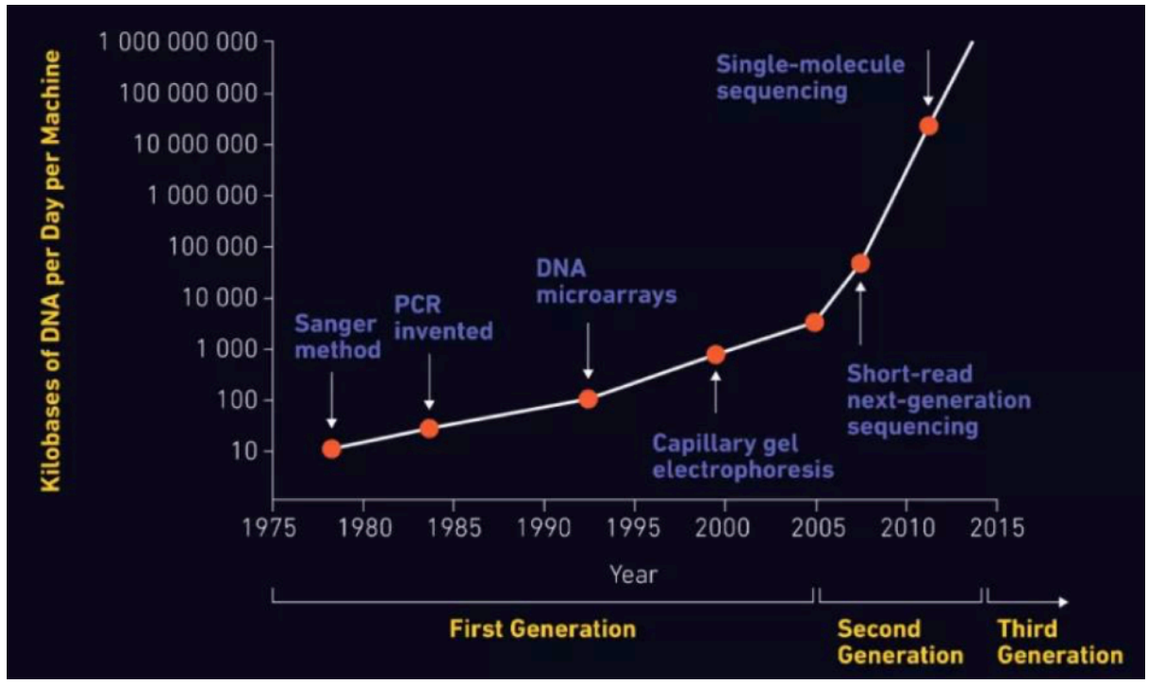
The era of affordable genome projects

Sequencing costs dropped drastically

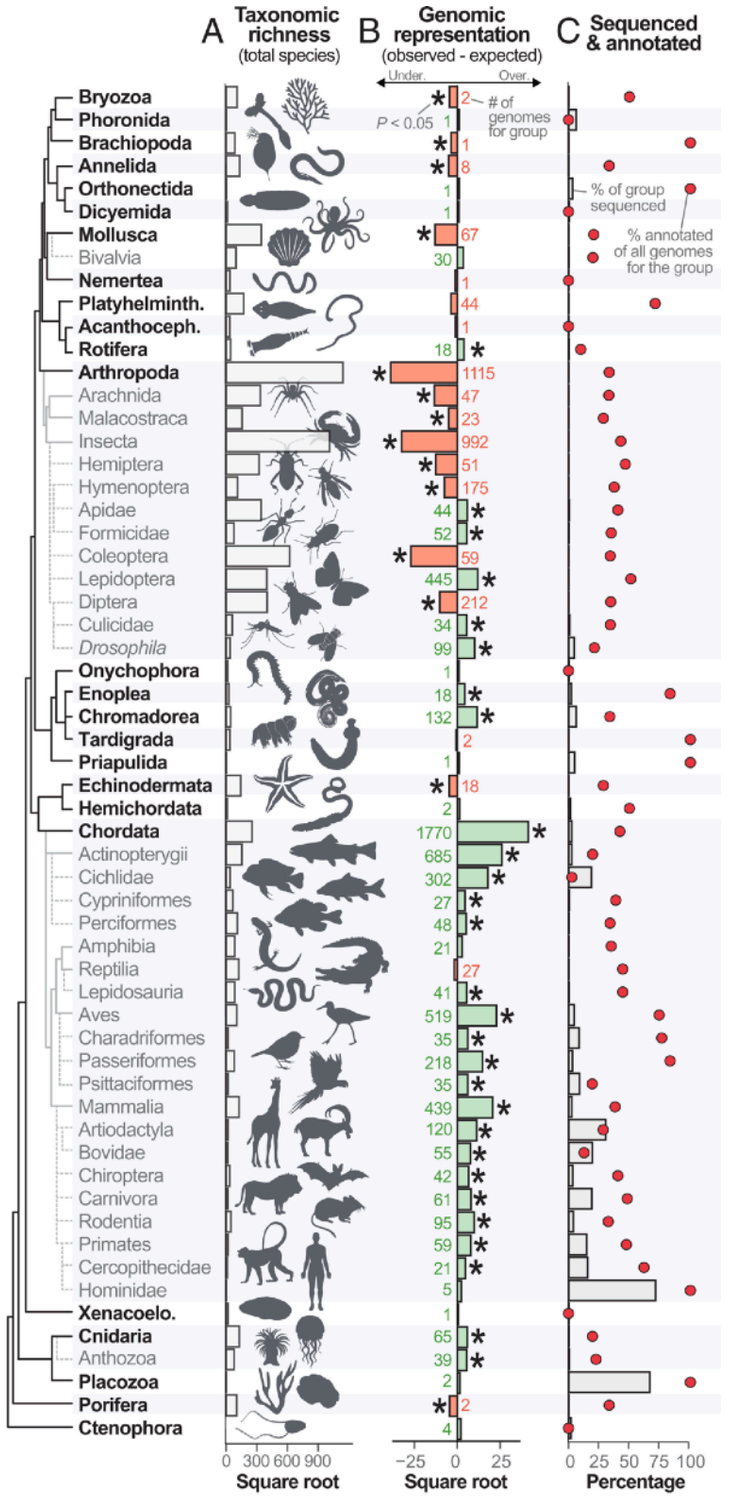
→ Nearly any species can now be sequenced.



From model species to genomic explosion



https://medicaltrend.org/2021/03/16/overview-of-next-generation-sequencing-technology/



PNAS 2021; 118(52): e2109019118.

Why annotate a newly sequenced genomes

Genome

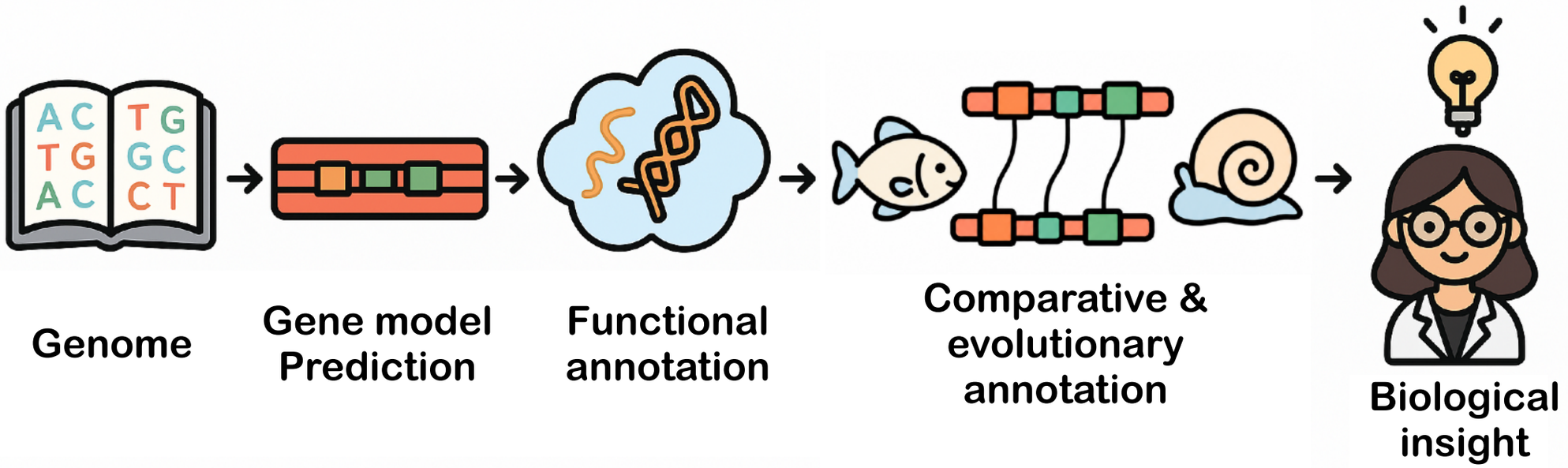
A genome assembly is a long sequence of A, T, G, and C.

Without annotation → readable, but with no guidance to its meaning

A high proportion of genes remain functionally uncharacterized.

In newly sequenced genomes, 30–50% of genes are annotated as “hypothetical proteins.”

The significance of gene annotation

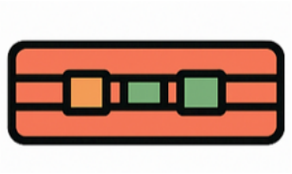


Gene annotation = DNA → gene models + functional descriptions

- Facilitates interpretation of biological features
- Establishes the basis for comparative genomics

Gene annotation workflow

Three major aspects of gene annotation:

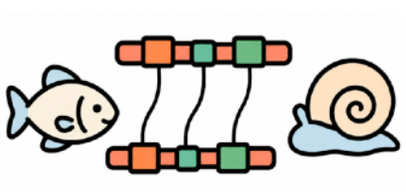


1. Structural annotation →

Where are the genes and exon-intron boundaries?

2. Functional annotation →

What can the predicted proteins probably do?



3. Comparative & evolutionary annotation →

Which genes are orthologs or paralogs?

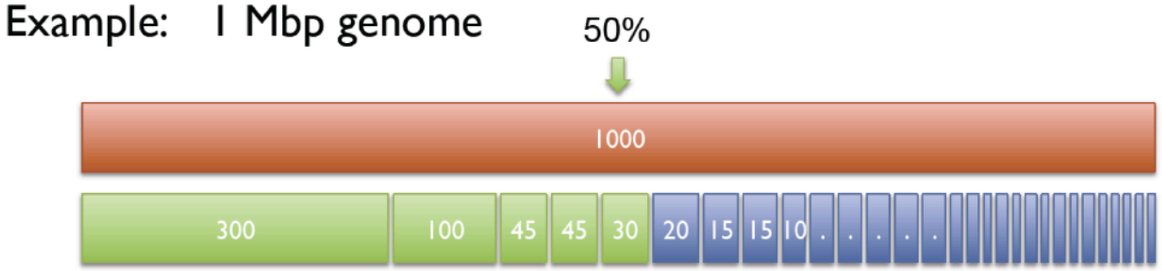
Reliable annotation requires a high-quality genome assembly.

(accurate and complete)

A reliable annotation starts with a reliable assembly

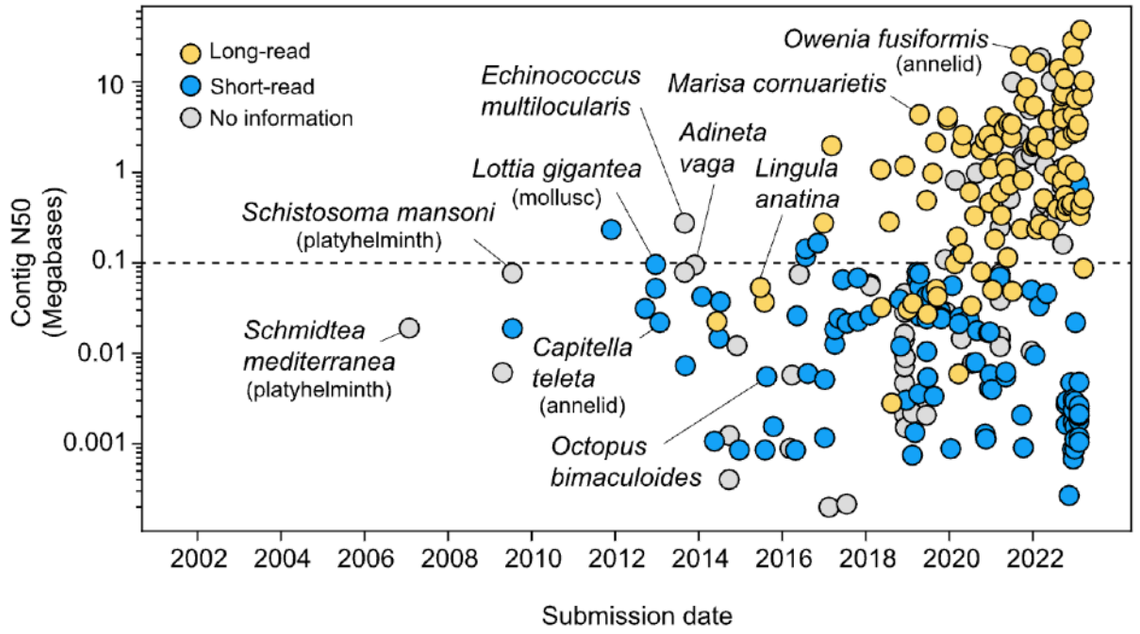
Fragmentation may create annotation errors.

Assembly **contiguity** assessment: N50, contig length



N50 size = 30 kbp
(300k+100k+45k+45k+30k = 520k >= 500kbp)

https://blog.csdn.net/u010608296/article/details/102771217



Liao et al. (2023) Briefings in Functional Genomics 1–11

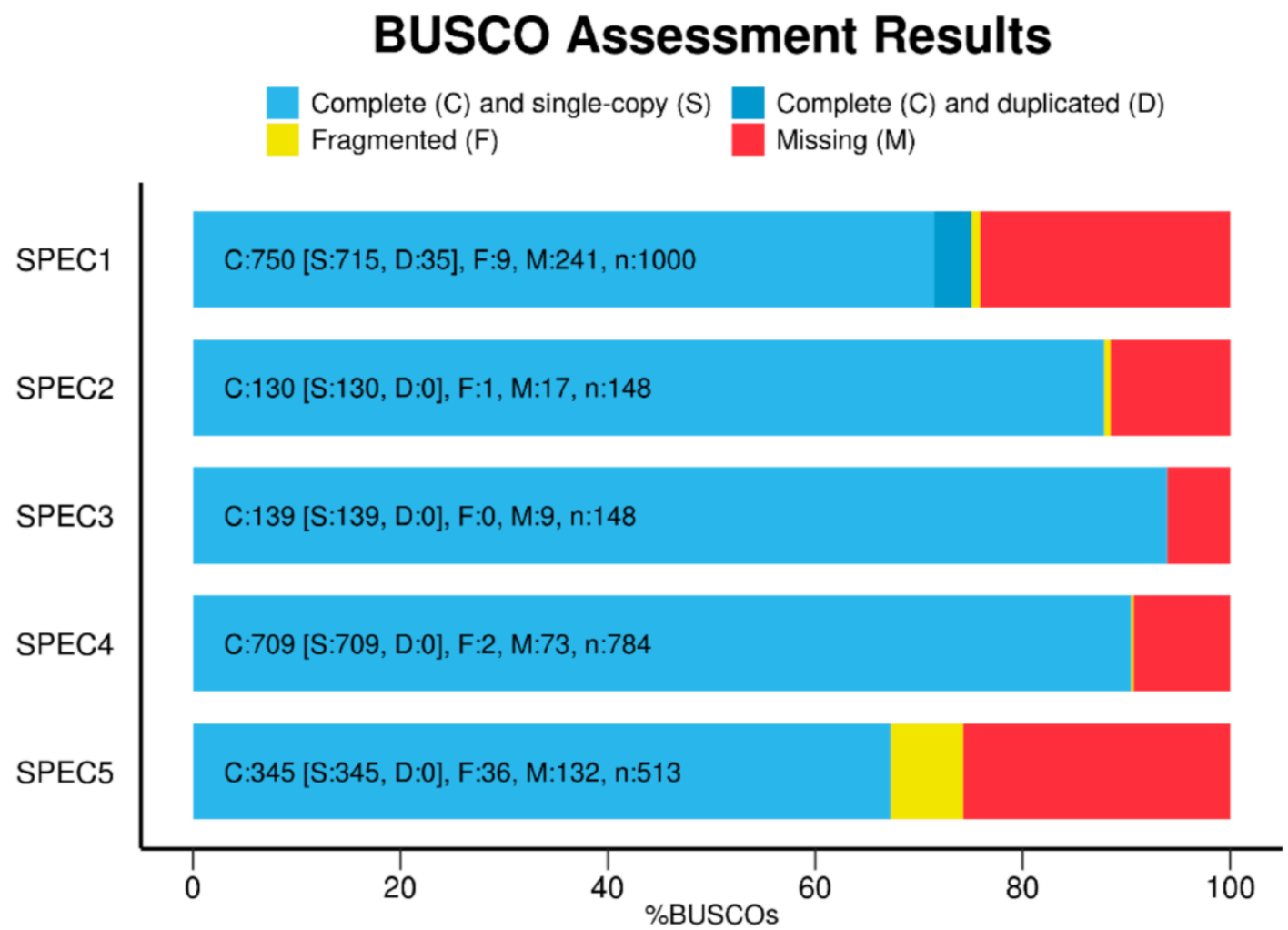
The half of the genome is covered with contigs at least the size of 30 kbp.

Genome quality assessment

BUSCO (**B**enchmarking **U**niversal **S**ingle-**C**opy **O**rthologs)

→ evaluates **completeness** based on conserved orthologs

By checking sets of highly conserved single-copy orthologs



https://busco.ezlab.org/busco_userguide.html

Soft-masked repetitive regions

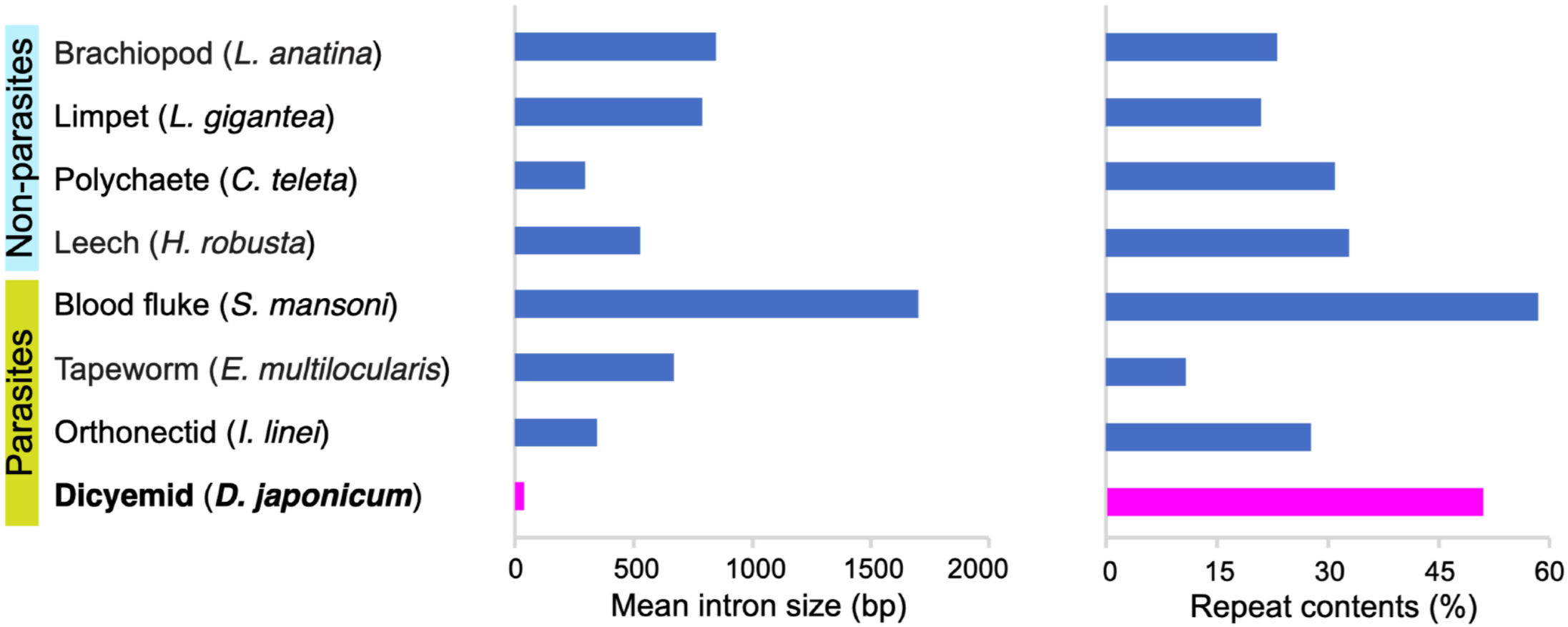
To annotate a novel genome, **repetitive and low complexitiy regions** should be soft-masked to avoid the prediction of false positive gene structures.

TTCCTGATATGTAATTAATCACTAAATGTCTTTAATGGGATCTCTTTCTA
TTGAGATATTTGTAAACTTTCTTCATGTGATTGGTTTACAGATATTCAGG
TTTCTGCAAATGGGTGCTGTCTATATTATAGAATTTTTAGTTGAAATTTT
CAAAATACTCTTTGagtattctcttgtaattatattactttacaaggttt
gtggggcatctctttcatttgtgattacatggttgcagtattcttttgt
tcttagtcagactgtataattgtctgtgaagtccagtaaactttgaaag

Build a **species-specific repeat library** with *RepeatModeler*

Perform soft-masking with *RepeatMasker* using this library

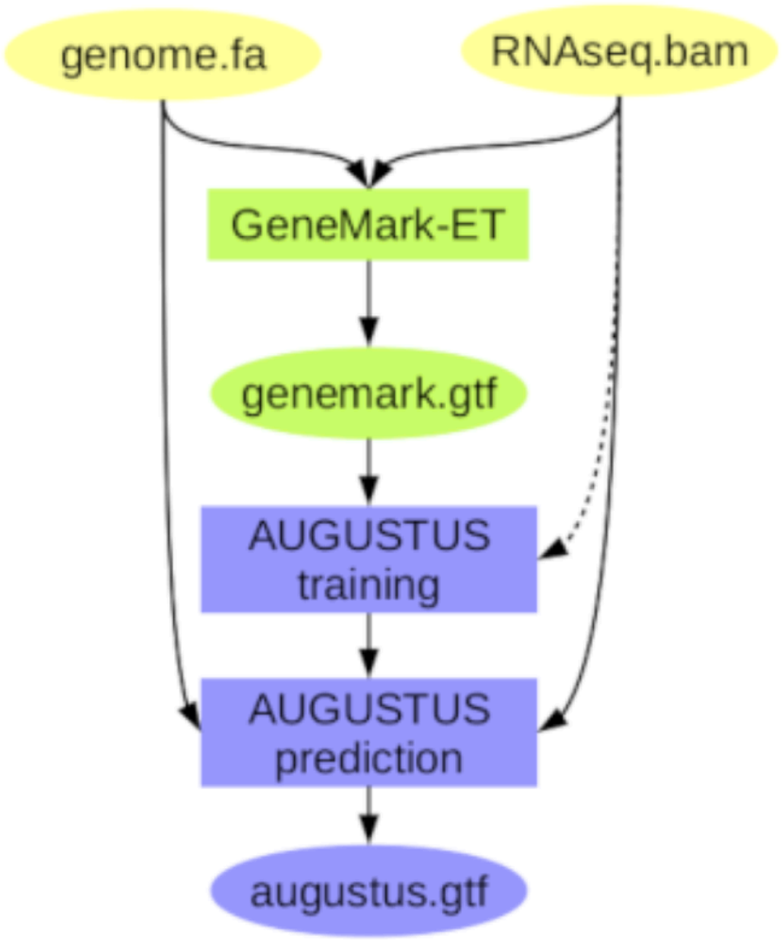
Need lineage-specific models for gene prediction



The genomic features of different species can vary greatly.

Evidence-based annotation using BRAKER3

BRAKER3 is a fully automated pipeline to train and predict highly reliable genes with *GeneMark-ET* and *AUGUSTUS*.



https://github.com/Gaius-Augustus/BRAKER

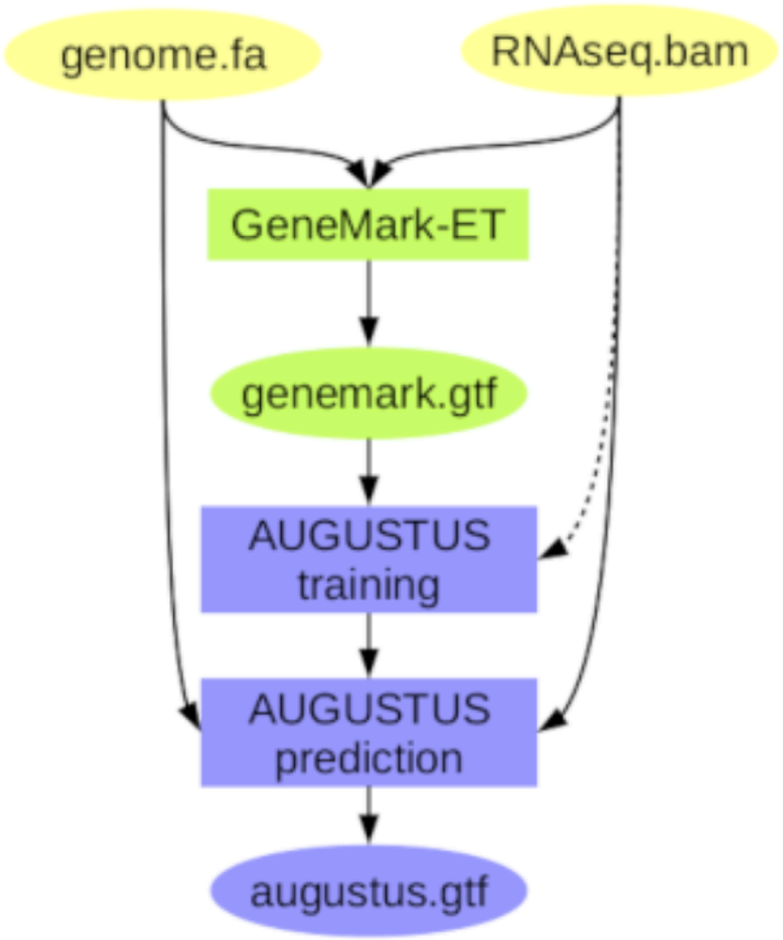
RNA-seq reflects real transcription signals.

Splice-aware mapping aligns RNA-seq reads to the genome to captures splicing junctions.

→ define exon–intron boundaries

Evidence-based annotation using BRAKER3

BRAKER3 is a fully automated pipeline to train and predict highly reliable genes with *GeneMark-ET* and *AUGUSTUS*.



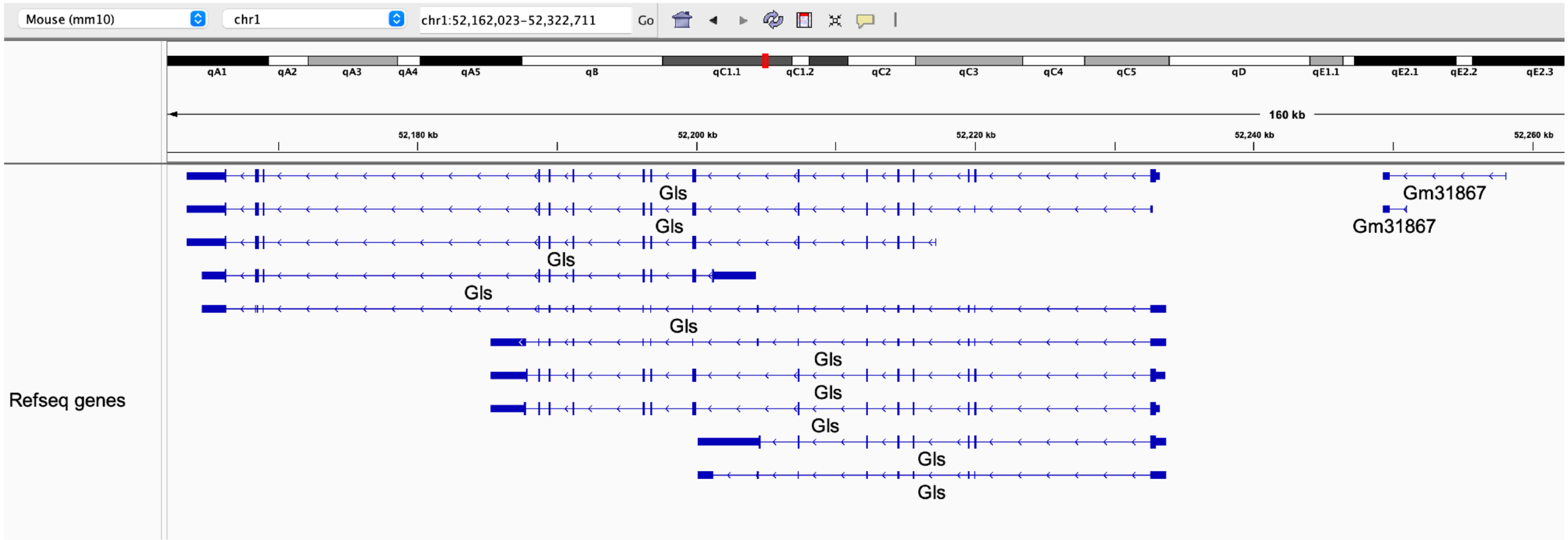
https://github.com/Gaius-Augustus/BRAKER

GeneMark-ET automatically derives an initial training set from RNA-seq data.
Iterative training progressively adapts the algorithm to the target species.

→ model gradually converges.

AUGUSTUS incorporates RNA-seq evidence (splice junctions, exon coverage) as hints to predict high-confidence gene models.

Annotate a novel genome using BRAKER3

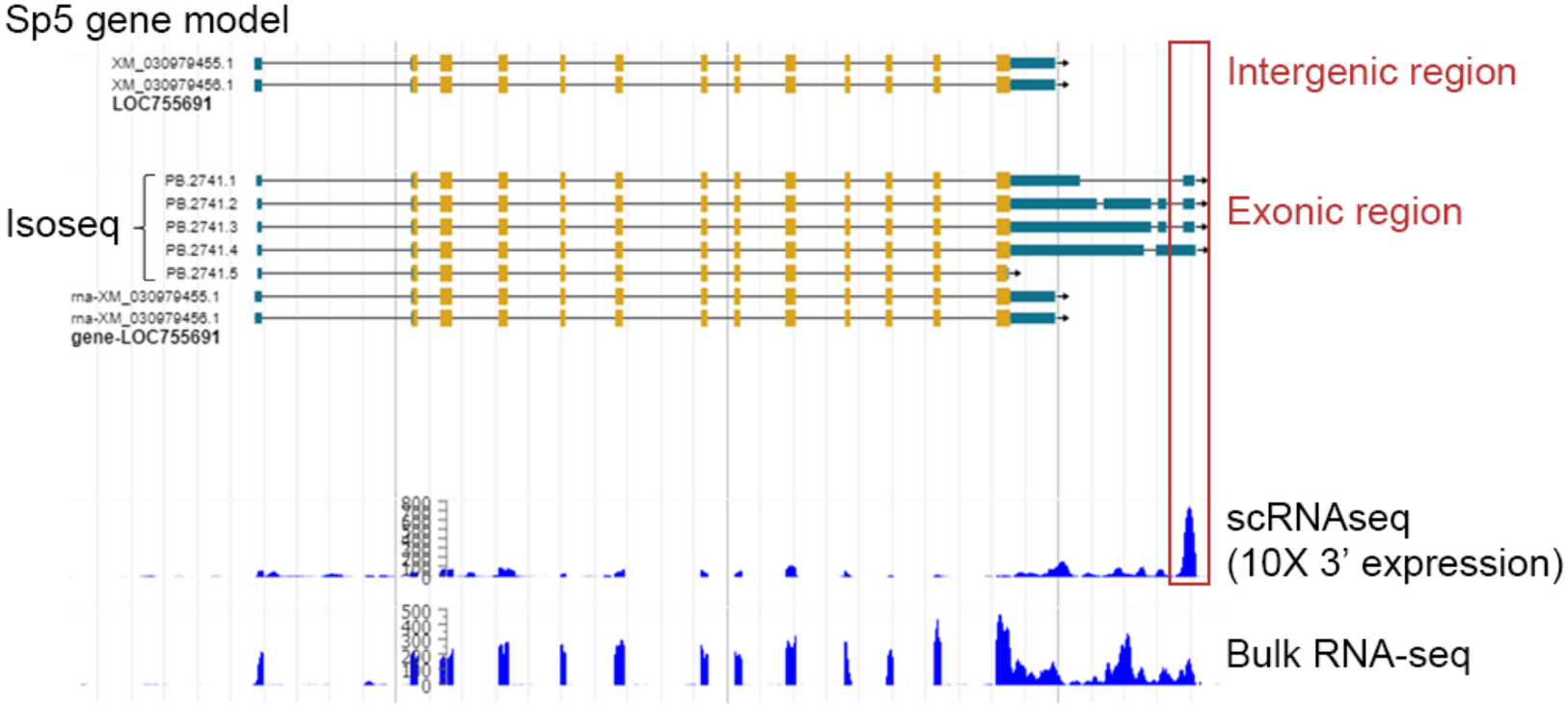


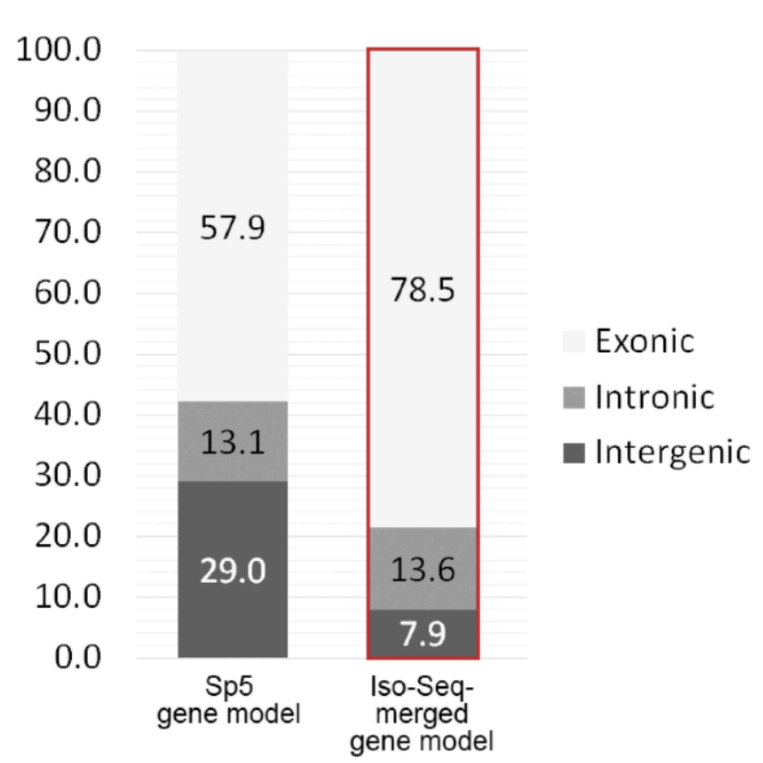
Always check gene prediction results before further usage!

Iso-Seq improves the genome annotation

Full-length reads directly connect exons within the same transcript.

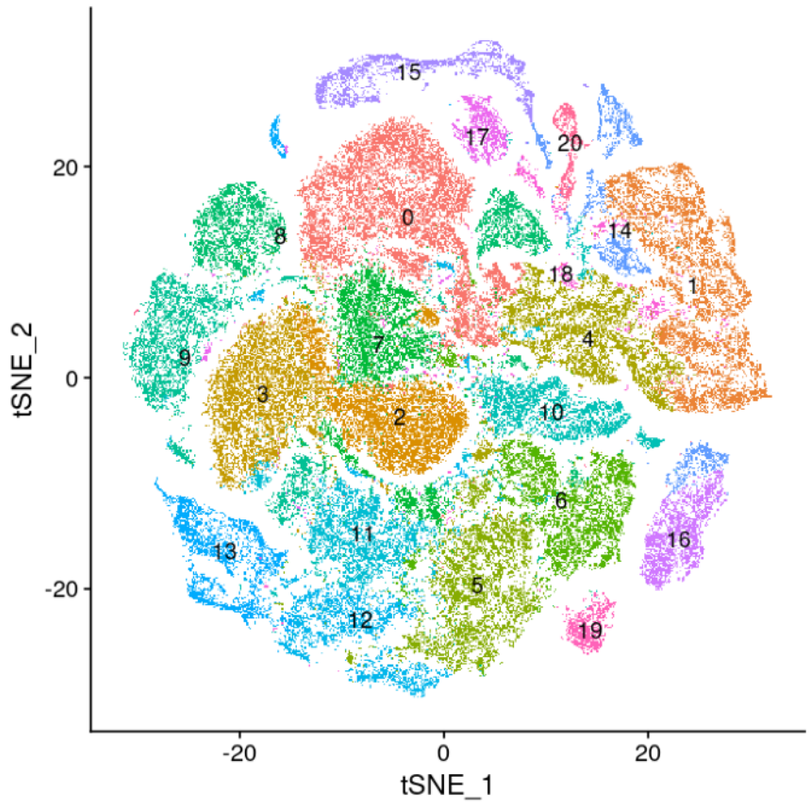
Iso-Seq helps resolve transcript ends, and splice isoforms.

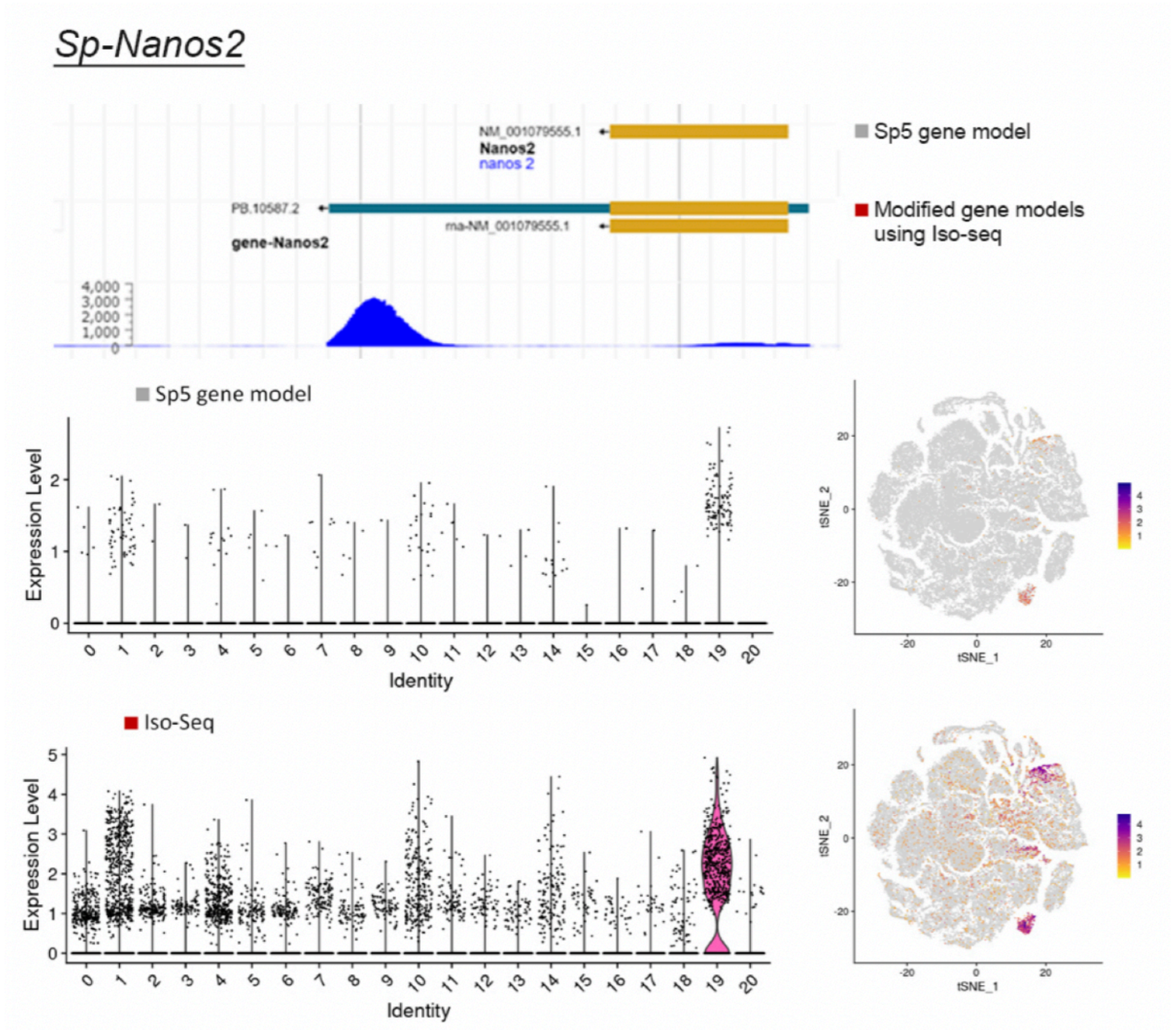




Iso-Seq improves the genome annotation

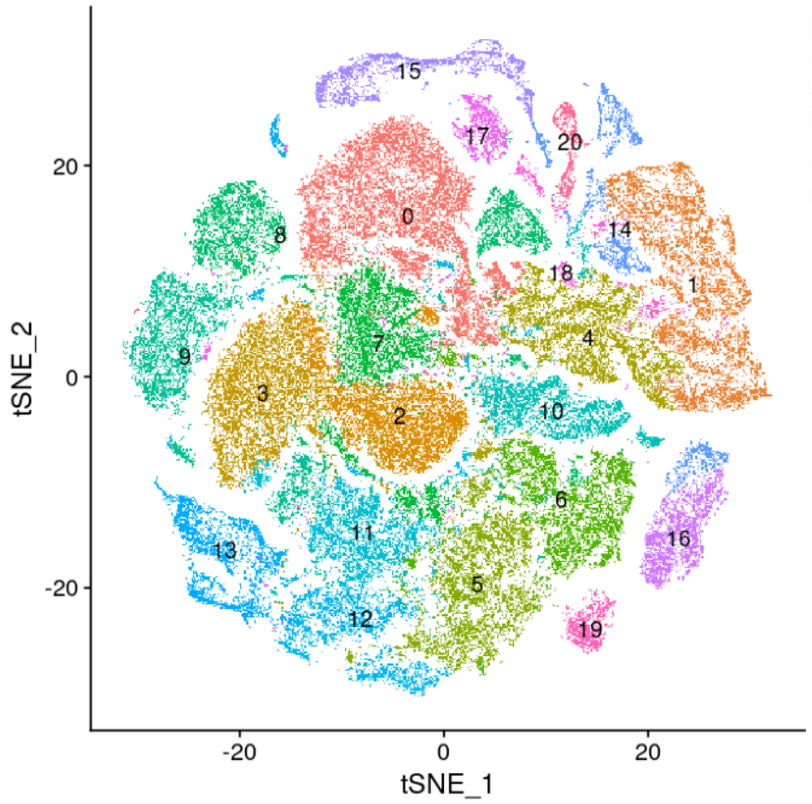
Iso-Seq-refined gene models offer more accurate gene expression quantification.

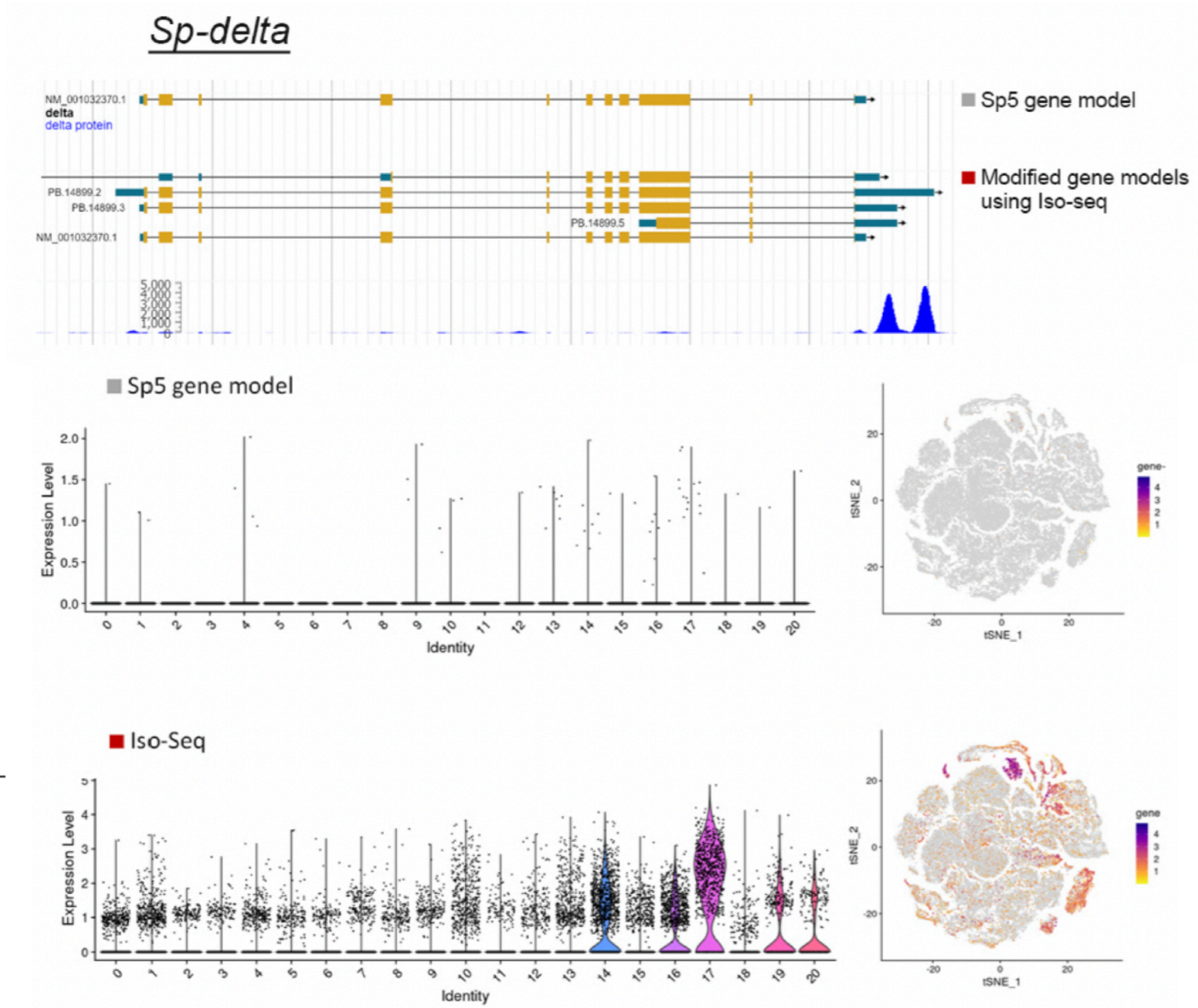




Iso-Seq improves the genome annotation

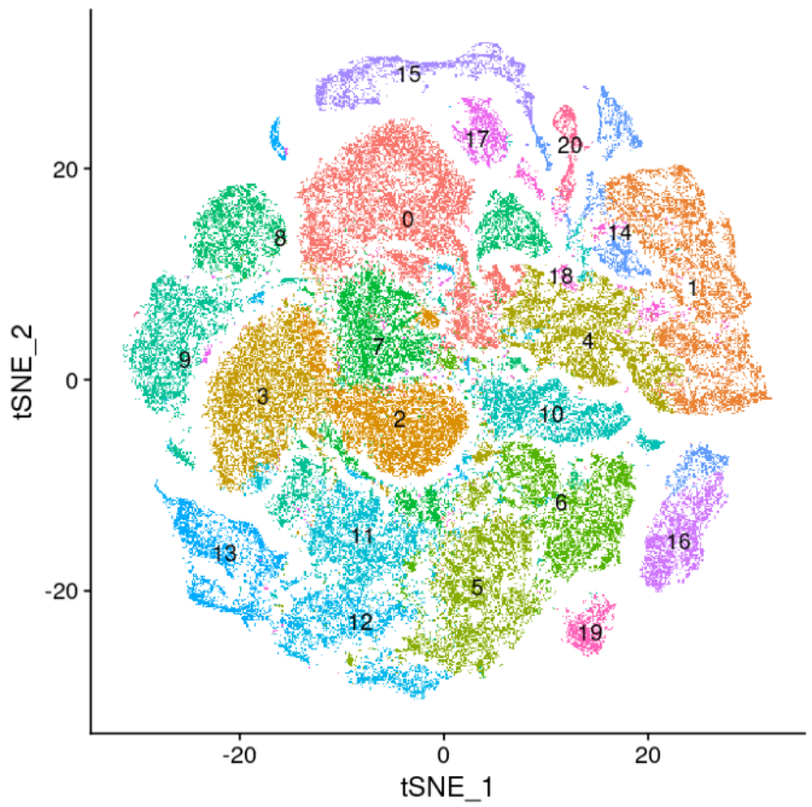
Iso-Seq-refined gene models offer more accurate gene expression quantification.

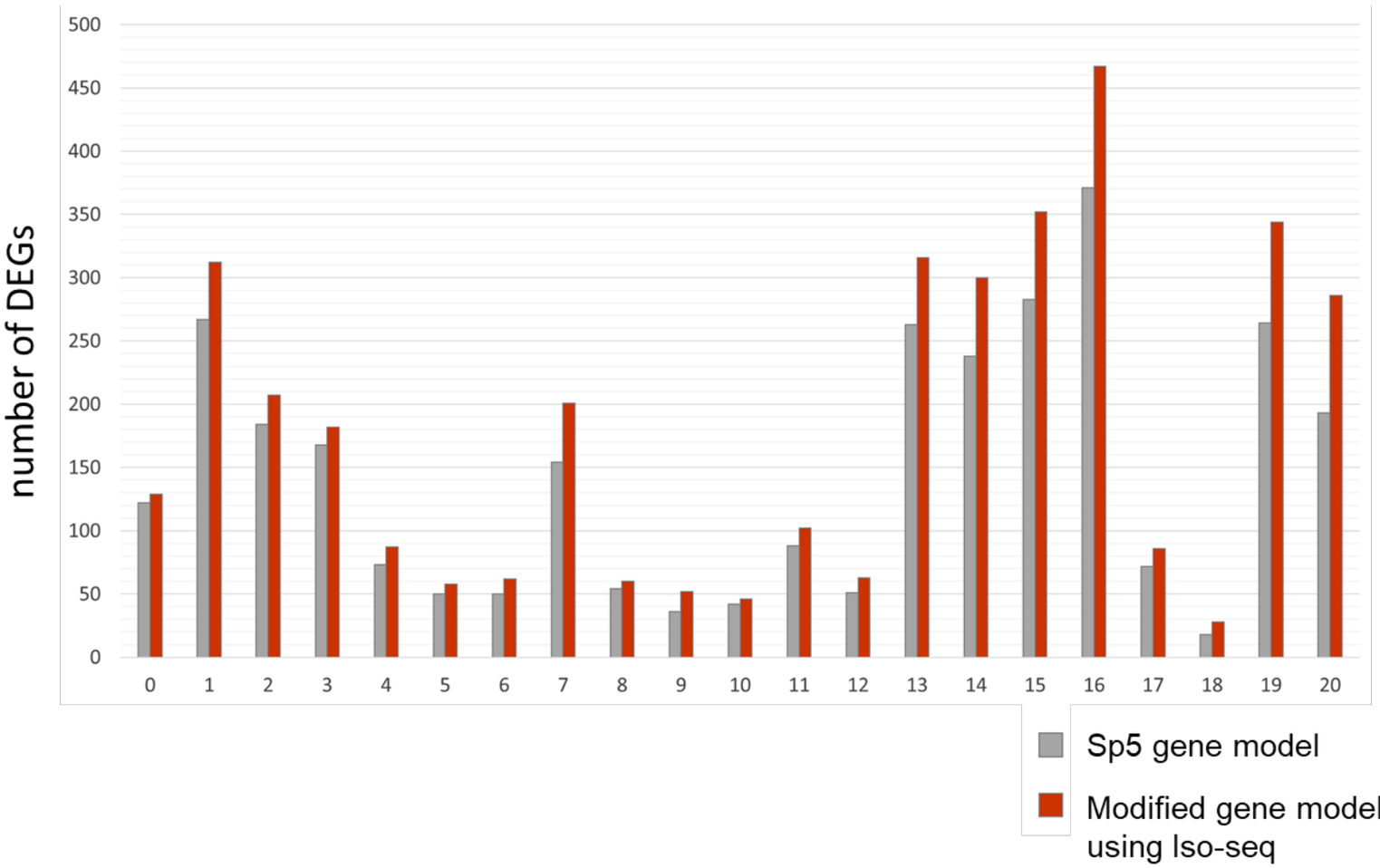




Why improved gene models matter downstream

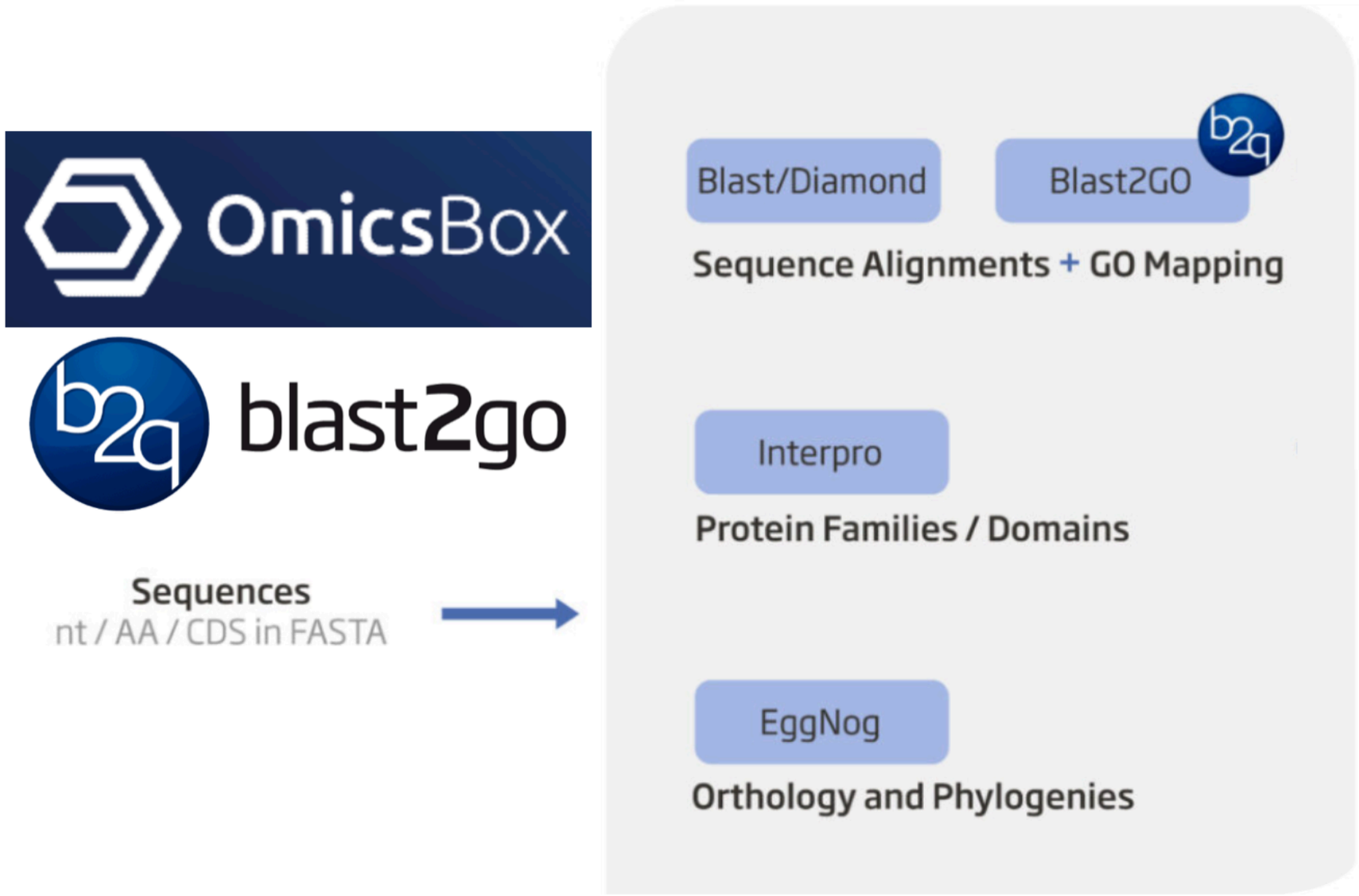
Using Iso-Seq-refined gene models led to an increased number of differentially expressed genes detected.





Functional annotation of gene models

OmicsBox (Blast2GO) functional annotation pipeline



Functional annotation of gene models

Inputs:

Protein FASTA exported from structural annotation (e.g., BRAKER3)

Blast2GO

1) Similarity search: closest characterized proteins
2) Automatic mapping of Gene Ontology (GO) terms from search results

Functional annotation of gene models

InterProScan

1) Integrates multiple databases (Pfam, SMART, ProDom, SUPERFAMILY, CDD, …)
2) Identifies protein families, domains, and motifs.
3) Automatically assigns Gene Ontology (GO) terms

eggNOG-mapper

1) Assign input sequences to orthologous groups in the eggNOG database
2) Provides GO terms, KEGG pathways, COG/eggNOG functional categories, protein names
3) Orthology-based → more accurate functional transfer across species

Functional annotation of gene models

Praesagittifera naikaiensis | 22,128 sequences

	Swiss-Prot (metazoan)	UniProtKB (metazoan)	UniProtKB (6 models)
Reference sequences	107,395	593,423	206,289
GO from Blast2GO	9,374 (42.4%)	12,136 (54.8%)	11,226 (50.7%)
GO from InterProScan		16,606 (75.0%)	
GO from eggNOG-mapper		9,408 (42.5%)	
GO Terms			
Final unique GO Terms	13,992	14,336	14,484

Limits:

1) Sequence similarity does not guarantee identical biological function.
2) Domains can be conserved while expression and regulatory context diverge.
3) Some annotations are not released as a reusable community resource.

From annotated genes to comparative questions

Which genes were retained, lost, duplicated, or rearranged?

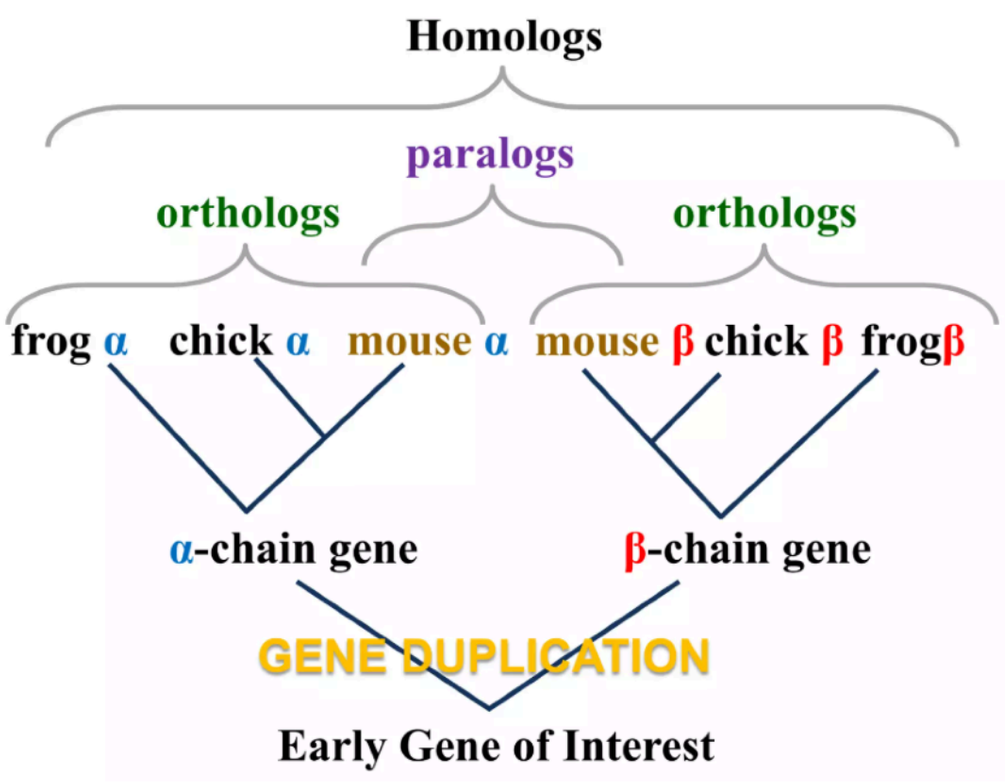
Which expression and functions are comparable across species?

Before comparing genes across species, we need to know which genes are evolutionary equivalent.

The importance of orthology assignment

Orthology | “who this gene’s relatives are”

Genes in newly sequenced species can be functionally annotated by mapping them to their orthologs in well-studied organisms.



https://www.slideshare.net/slideshow/ncbi-boot-camp-for-beginners-slides/7422486#58

Orthologs = corresponding genes derived from speciation

Paralogs = sibling genes derived from gene duplication

Inferring orthology relationships across species

OrthoFinder defines orthogroups and distinguishes orthologs vs. paralogs

- Perform all-vs-all sequence similarity searches
- Cluster genes into orthogroups.

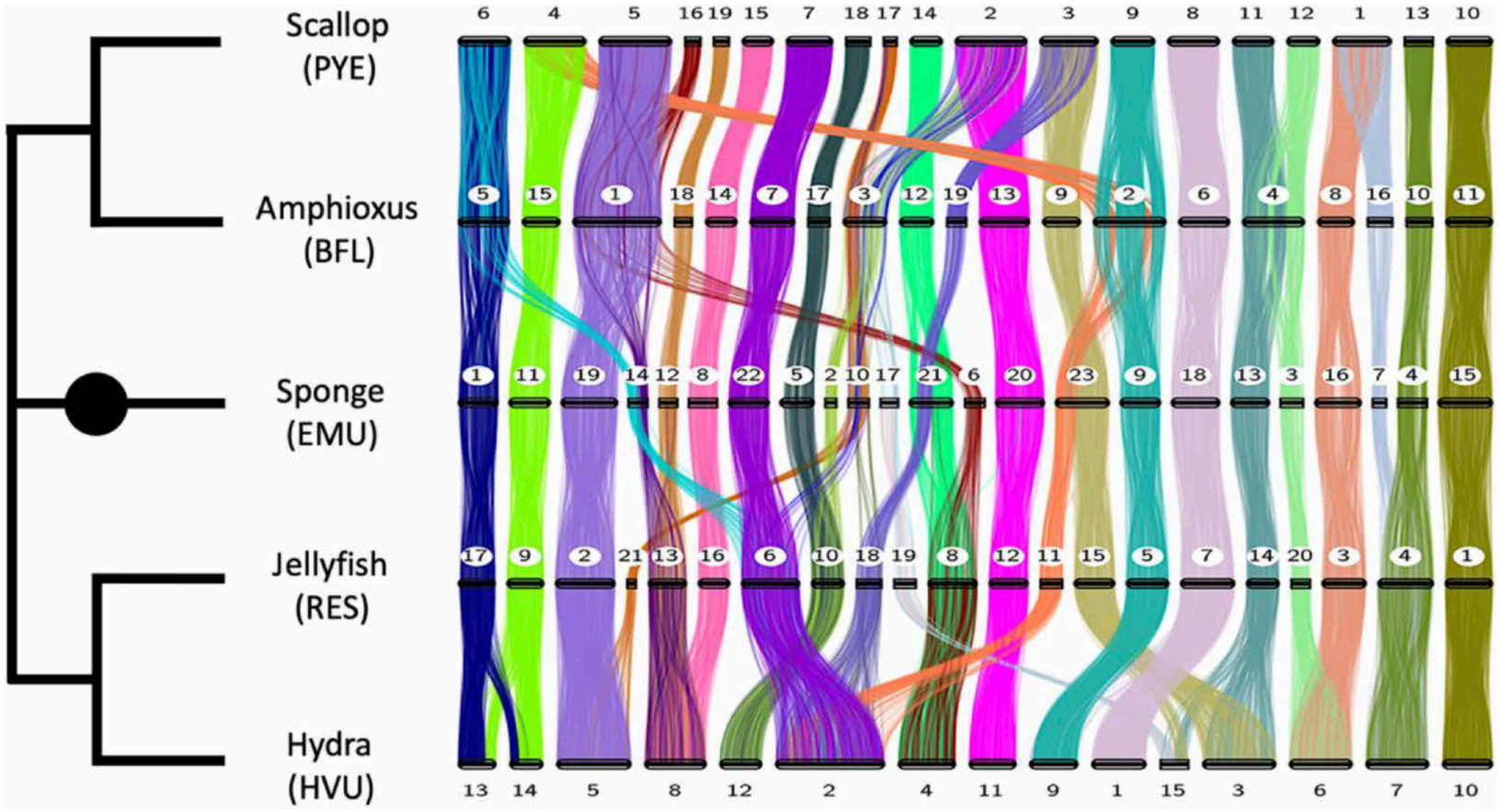
→ insights into gene family expansion and contraction

single-copy orthologs → building phylogenomic trees

Macrosynteny in orthology assignment

Basic concept of synteny

- Synteny → Conservation of gene order across species
- Macrosynteny → Large-scale conservation (chromosomal segments)



Simakov et al. (2022) Sci. Adv. 8: eabi5884

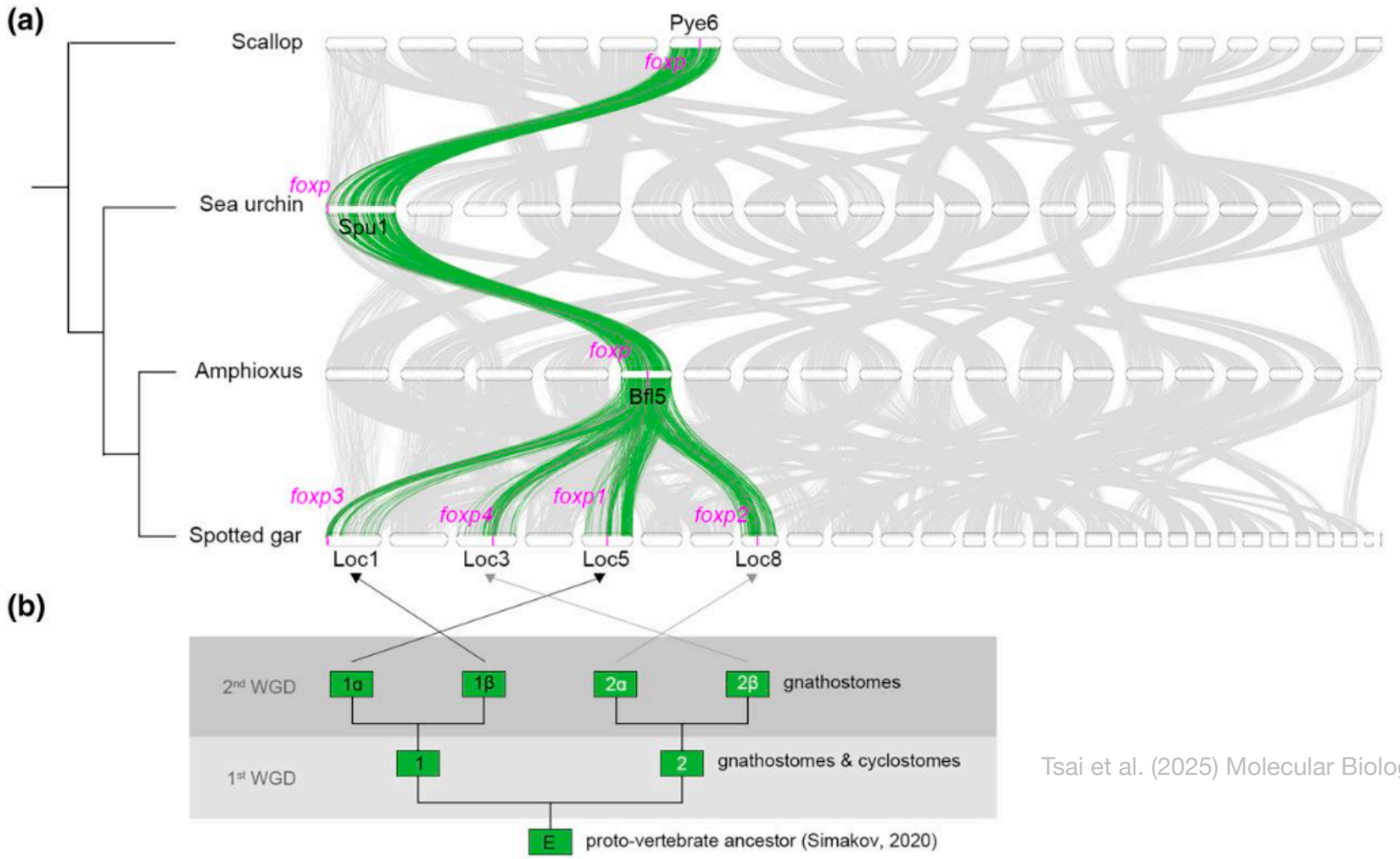
Macrosynteny in orthology assignment

ALG	Amphioxus	Scallop	Bilaterian ancestor (24)	Sponge	Cnidarian ancestor (21)	Jellyfish	Hydra	No. of shared orthologs
G	BFL11	PYE10	G	EMU15	G	RES1	HVU10	137
B1	BFL10	PYE13	B1	EMU04	B1⊗B2	RES4	HVU7	100
B2	BFL16	PYE1	B2	EMU07/24				88
M	BFL8	PYE1	M	EMU16	M	RES3	HVU6	102
C2	BFL3L	PYE17	C2	EMU10	C2	RES21	HVU5	20
A1a	BFL1	PYE5	A1a⊗A1b	EMU19	A1a	RES2	HVU5	209
A1b				EMU14R	A1b⊗B3	RES13	HVU8	52
B3	BFL18	PYE19	B3	EMU12				47
P	BFL14	PYE15	P	EMU08	P	RES16	HVU8	76
L	BFL15	PYE4	L	EMU11	L	RES9	HVU14R	72
L_							HVU13R	31
Ea	BFL5	PYE6	Ea⊗Eb	EMU01	Ea	RES17	HVU13L	87
Ea_							HVU14L	26
Eb				EMU14L	Eb⊗F⊗Qb	RES6	HVU2	45
F	BFL7	PYE7	F	EMU22				142
Qb	BFL3R	PYE2	Qb⊗Qa	EMU10				12
Qa				EMU02	Qa⊗J1	RES10	HVU12	31
J1	BFL17	PYE18	J1	EMU05				55
R	--	PYE12	R	EMU17	R⊗Qc	RES19	HVU15L	32
Qc	BFL3R	PYE2	Qc⊗Qd					13
Qd				EMU10	Qd⊗O2	RES18	HVU2	21
O2	BFL19	PYE3	O2	EMU21mid				46
N	BFL12	PYE14	N	EMU21end	A2⊗N	RES8	HVU4	106
A2	BFL1	PYE16	A2	EMU06				41
H	BFL13	PYE2	H	EMU20	H	RES12	HVU11	136
J2	BFL2eve	PYE4	J2	EMU23L	J2	RES11	HVU2	64
C1	BFL2odd	PYE9	C1	EMU09	C1	RES5	HVU9	142
D	BFL6	PYE8	D	EMU18	D	RES7	HVU1	158
D_							HVU9t	14
K	BFL9	PYE3	K	EMU23R	K	RES15	HVU3	100
K_							HVU15R	17
I	BFL4R	PYE11	I	EMU13	I	RES14	HVU3	79
I_							HVU15R	14
O1	BFL4L	PYE12	O1	EMU03	O1	RES20	HVU6	54

Simakov et al. (2022) Sci. Adv. 8: eabi5884

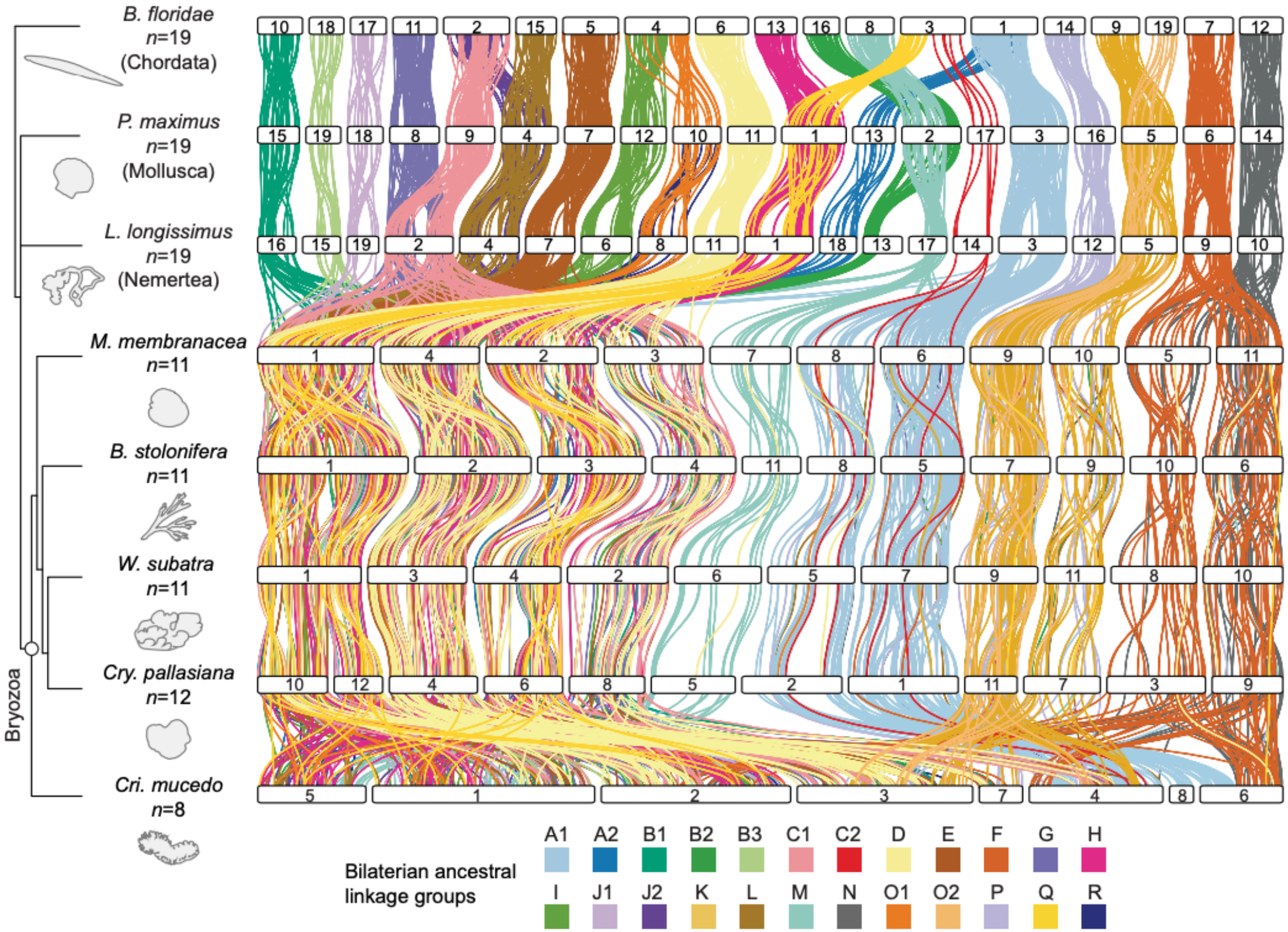
Macrosynteny resolves an unstable deep tree

1) Different sequence models recovered different deep FoxP topologies.
2) The critical early duplication nodes had low support.
3) The chromosome history independently supported the 2R-WGD scenario.



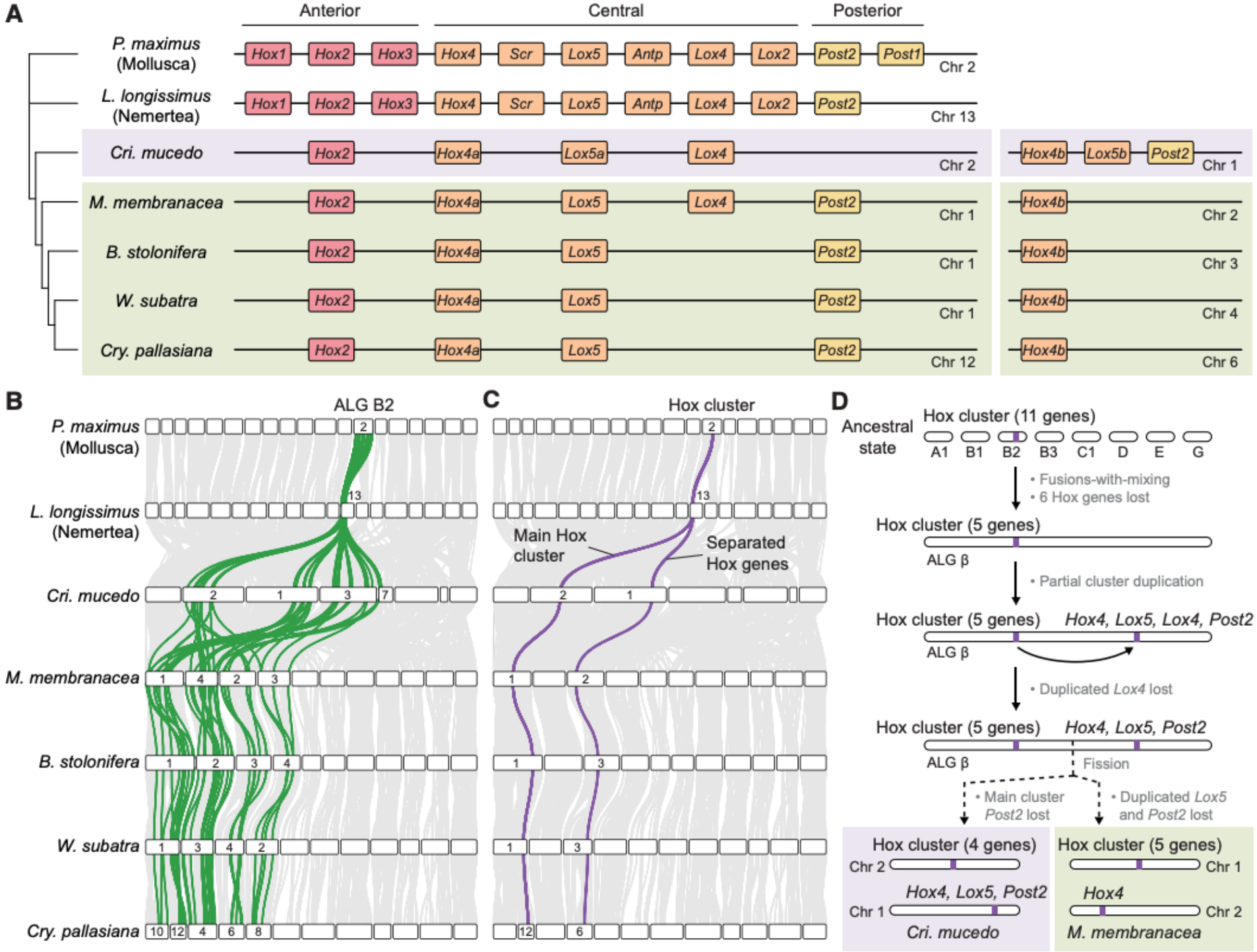
Tsai et al. (2025) Molecular Biology and Evolution. 42:1-17.

Macrosynteny in orthology assignment



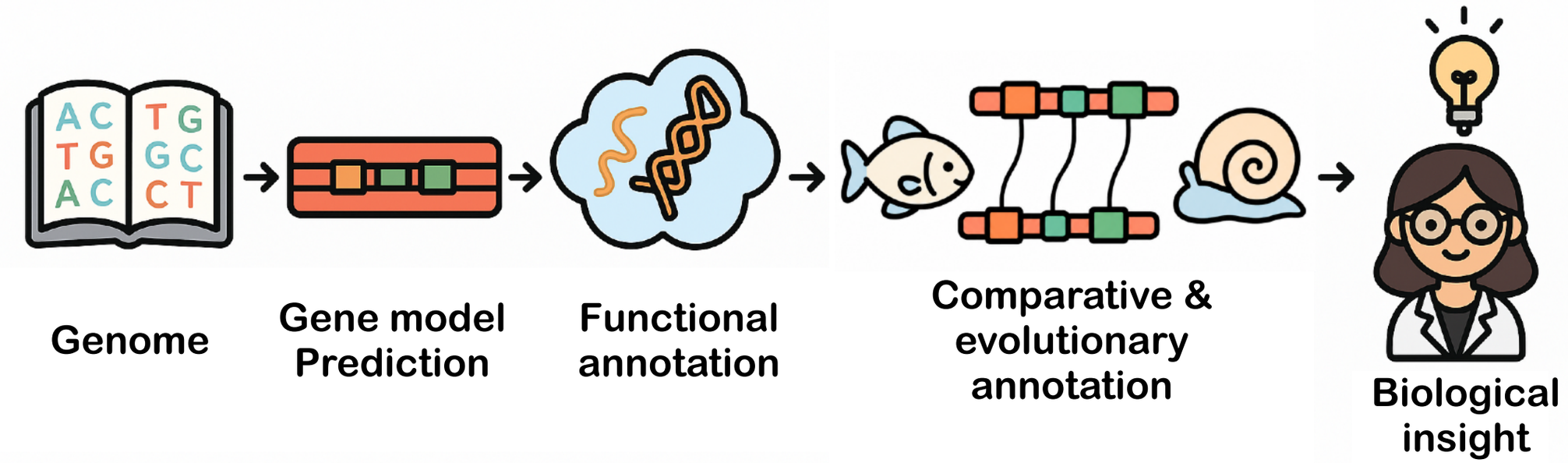
Lewin et al. (2025) Genome Res. 35(1):78-92.

Macrosynteny in orthology assignment



Lewin et al. (2025) Genome Res. 35(1):78-92.

Take-home messages



A genome assembly becomes a useful reference through iterative annotation.

Structural, functional, and comparative annotation answer different questions and should be integrated.

ALG and synteny preserving chromosome-history information complement gene trees.

Thank you for your attention!

tmlu@as.edu.tw