SFU SIMON FRASER UNIVERSITY
ENGAGING THE WORLD

C. briggsae genome annotation and comparative analysis with *C. elegans* using RNA-Seq data

Final oral examination for the degree of

Master of Science

SHINTA THIO

Monday, April 6th, 2020

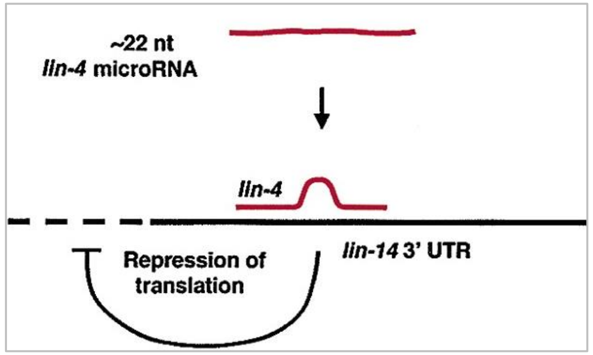
Caenorhabditis elegans as a model organism

C. elegans:

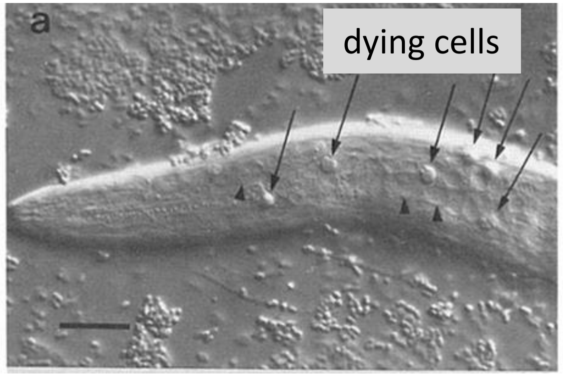
- **Small**: ~1mm long
- **Simple body plan**: 959 cells
- **Compact genome size**: ~100 Mbp, first multicellular genome sequenced[1]

Key discoveries using *C. elegans:*

1. microRNA[2]

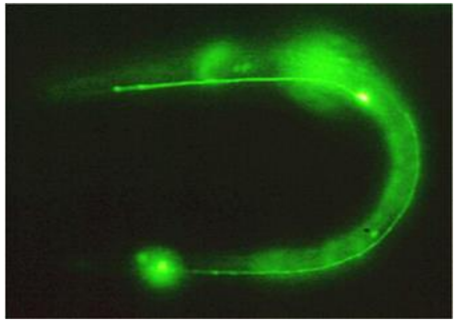


2. Apoptosis pathway[3]



Nomarski photomicrograph, newly hatched *ced-1* larva. Bar=10u

3. Cell visualization in living organism using GFP[5]

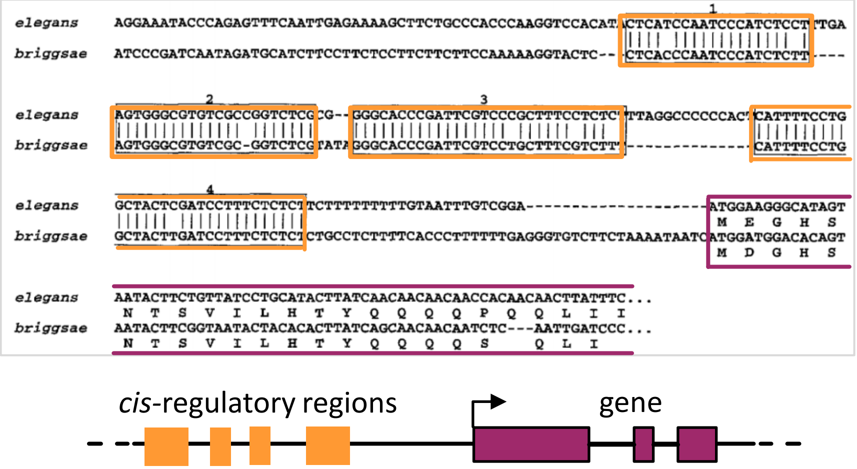


C. elegans is an ideal model organism to investigate biology

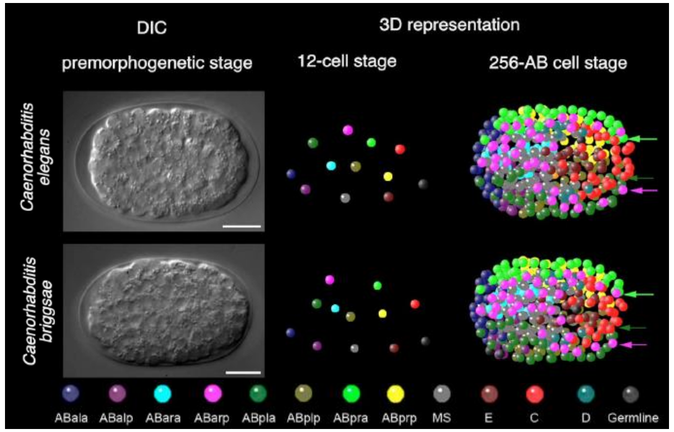
[1] The C. elegans Genome Sequencing Consortium, 1998
[2] Lee et al. , 1993
[3] Ellis & Horvitz, 1986
[4] Chalfie et al., 1994

C. briggsae facilitates *C. elegans* research

1. Improving genome annotation (ortholog-based)[1-7]

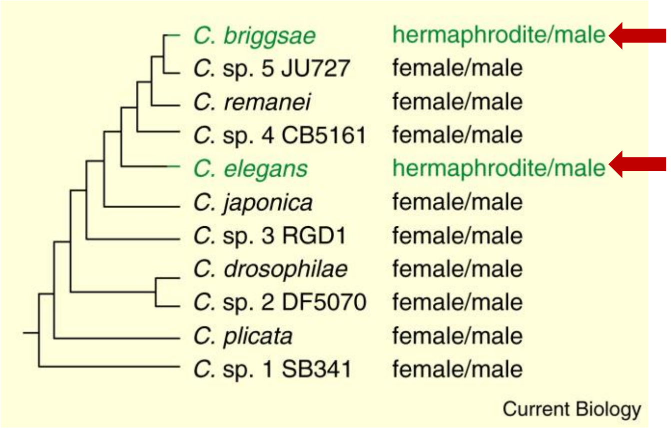

elegans AGGAAATACCCAGAGTTTCAATTGAGAAAAGCTTCTGCCCACCCAAGGTCCACATACTCATCCAATCCCATCTCCTTTGA
briggsae ATCCCGATCAATAGATGCATCTTCCTTCTCCTTCTTCTTCCAAAAAGGTACTC---CTCACCCAATCCCATCTCTT----
1
elegans AGTGGGCGTGTCGCCGGTCTCGCG--GGGCACCCGATTCGTCCCGCTTTCCTCTCTTTAGGCCCCCCACTCATTTTCCTG
briggsae AGTGGGCGTGTCGC-GGTCTCGTATAGGGCACCCGATTCGTCCTGCTTTCGTCTT-------------CATTTTCCTG
2
3
elegans GCTACTCGATCCTTTCTCTCTTCTTTTTTTTTGTAATTTGTCGGA-------------------ATGGAAGGGCATAGT
M E G H S
briggsae GCTACTTGATCCTTTCTCTCTCTGCCTCTTTTCACCCTTTTTTGAGGGTGTCTTCTAAAATAATCATGGATGGACACAGT
M D G H S
4
elegans AATACTTCTGTTATCCTGCATACTTATCAACAACAACAACCACAACAACTTATTTC...
N T S V I L H T Y Q Q Q Q P Q Q L I I
briggsae AATACTTCGGTAATACTACACACTTATCAGCAACAACAATCTC---AATTGATCCC...
N T S V I L H T Y Q Q Q Q S Q L I
cis-regulatory regions
gene

2. Understanding embryonic development[8-10]



wild-type embryos (scale bar 10 μm), DIC = Differential Interference Contrast microscopy

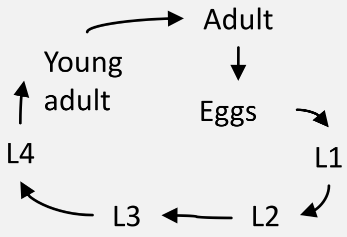
3. Understanding evolution of hermaphroditism[11]

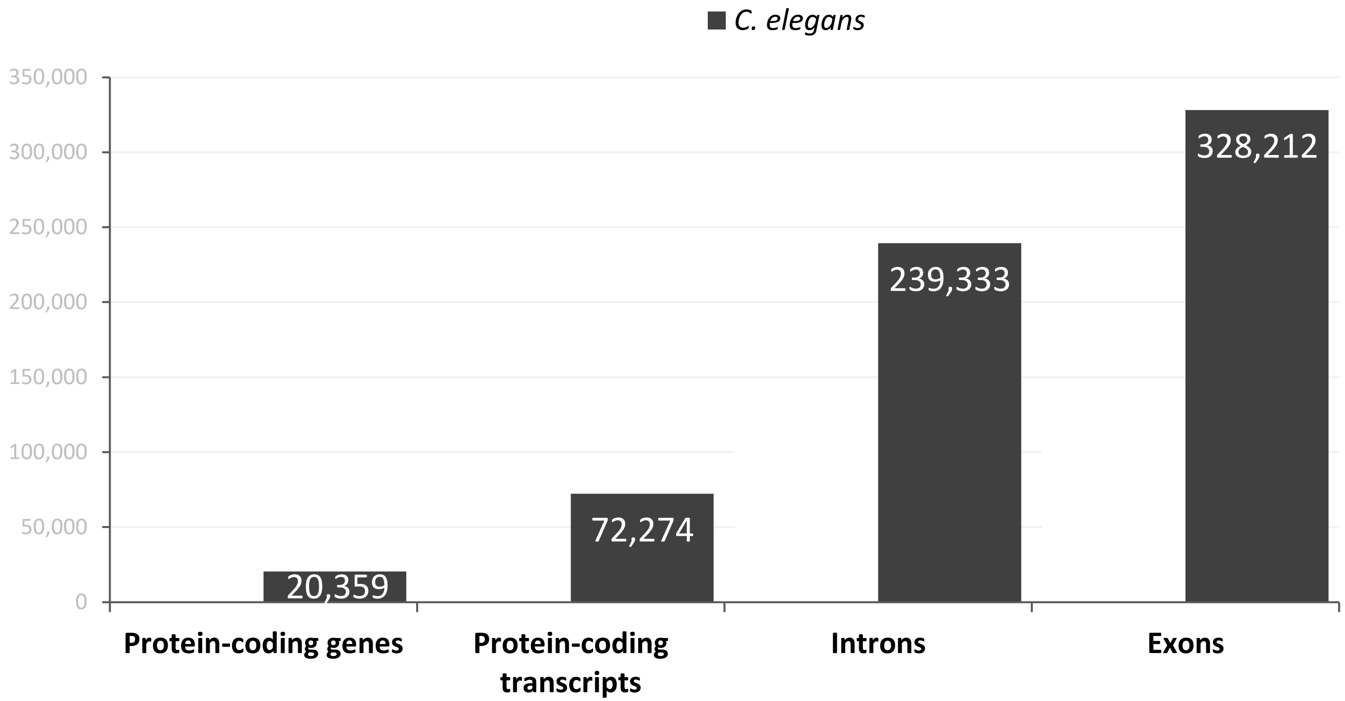

C. briggsae hermaphrodite/male
C. sp. 5 JU727 female/male
C. remanei female/male
C. sp. 4 CB5161 female/male
C. elegans hermaphrodite/male
C. japonica female/male
C. sp. 3 RGD1 female/male
C. drosophilae female/male
C. sp. 2 DF5070 female/male
C. plicata female/male
C. sp. 1 SB341 female/male
Current Biology

C. briggsae is a powerful comparative tool to improve the understanding of *C. elegans*

[1] Clark et al., 1995, [2] The C. elegans Genome Sequencing Consortium, 1998, [3] Chen et al., 2006, [4] Gaudet et al., 2004, [5] Kirouac and Sternberg, 2003, [6] Stein et al, 2003, [7] Uyar et al, 2012, [8] Sulston et al., 1983, [9] Zhao et al., 2008, [10] Memar et al., 2019, [11] Kiontke et al., 2004

Current *C. briggsae* annotation is inadequate

	Life cycle	Reproduction	Number of cells (adult ⚥)	Chromosome	Genome size
C. elegans 	 Adult Eggs L1 L2 L3 L4 Young adult	Hermaphrodite & male (<0.2%)	959 cells	6	100 Mbp
C. briggsae 					108 Mbp

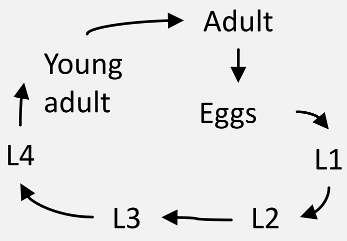


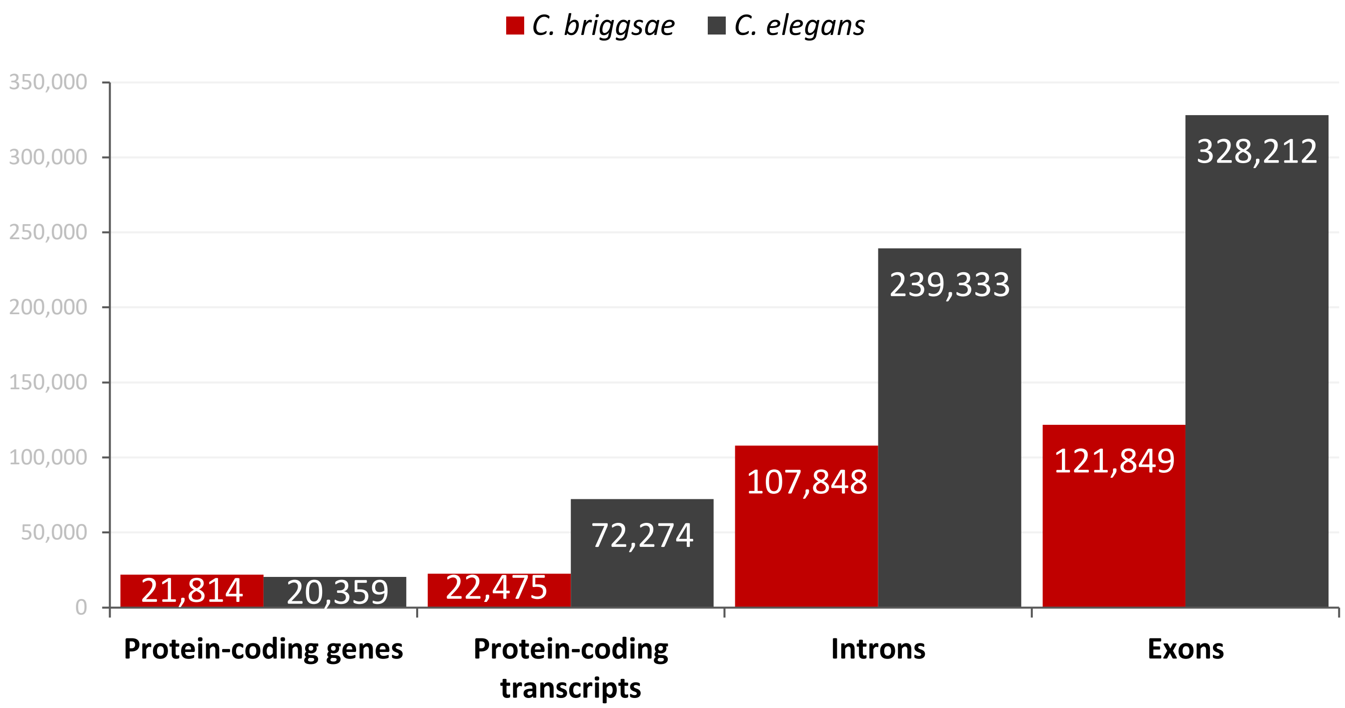
[1] Douglas, M., 2018
[2] WormBase WS250

Current *C. briggsae* annotation is inadequate

C. elegans

C. briggsae

	Life cycle	Reproduction	Number of cells (adult ⚥)	Chromosome	Genome size
C. elegans		Hermaphrodite & male (<0.2%)	959 cells	6	100 Mbp
C. briggsae					108 Mbp



Extensive studies in *C. elegans*, **limited studies** in *C. briggsae*.

[1] Douglas, M., 2018

[2] WormBase WS250, WS254

Current *C. briggsae* annotation is mostly computational

C. briggsae

Computational

- *Ab initio* gene finding
- Sequence conservation

Experimental

- ESTs & Protein-based comparisons
- RNA Sequencing (RNA-Seq, **2 libraries**)

C. elegans

Computational

- *Ab initio* gene finding
- Homology-based gene prediction

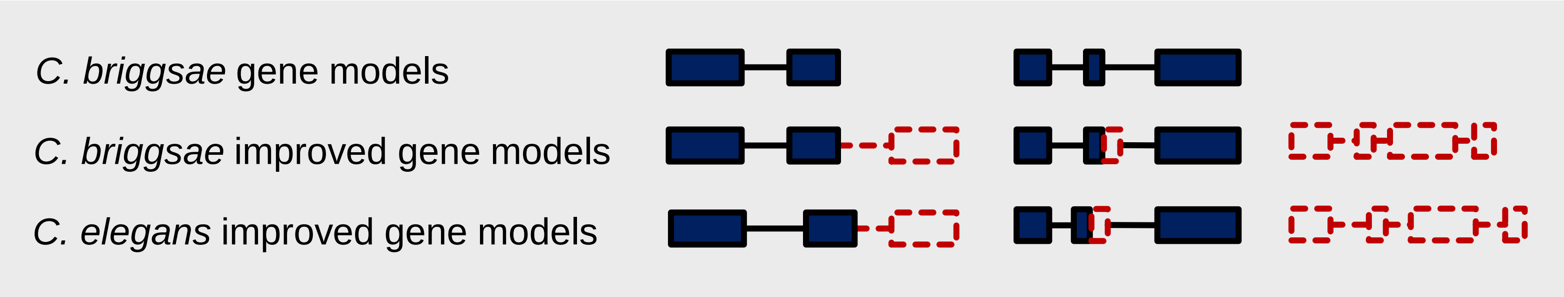
Experimental

- Expressed sequence tags (ESTs)
- Open reading frame sequence tags (OSTs)
- Serial analysis of gene expression (SAGE)
- Rapid Amplification of cDNA ends (RACE)
- Trans-spliced exon coupled RNA end determination (TEC-RED)
- RNA Sequencing (RNA-Seq, **~800 libraries**)

Hypothesis #1: **The *C. briggsae* genome annotation is incomplete and can be improved using additional RNA-Seq data.**

Complete genome annotation is essential for comparative genomics

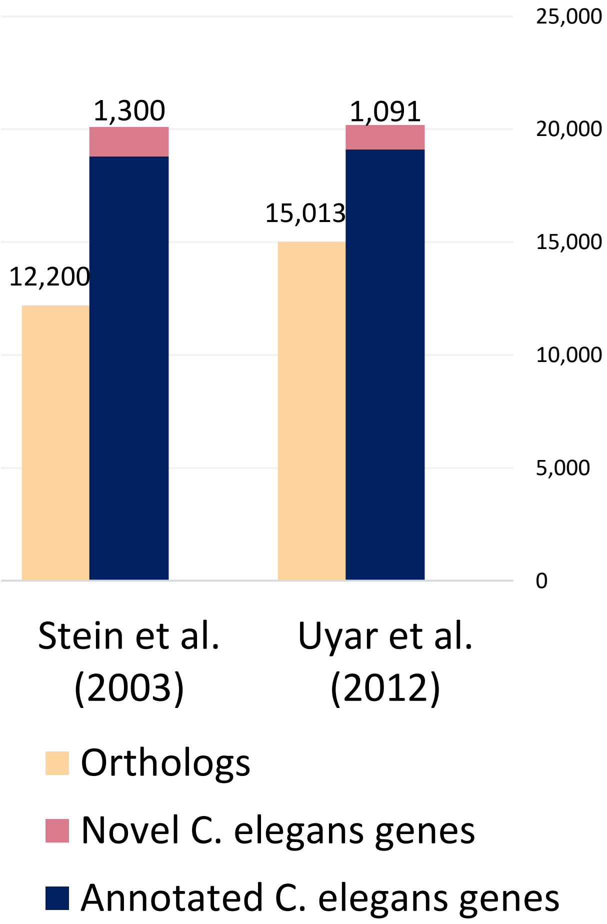
- Protein-coding sequences and *cis*-regulatory regions are usually highly conserved
- A more complete or accurate genome annotation can increase the quality of the annotation of its close relative



C. briggsae gene models

C. briggsae improved gene models

C. elegans improved gene models



<u>Hypothesis #2</u>: **Using the improved *C. briggsae* annotation, we can find additional orthologous relationships with *C. elegans* that were previously missed and additional *C. elegans* gene models.**

Hypotheses

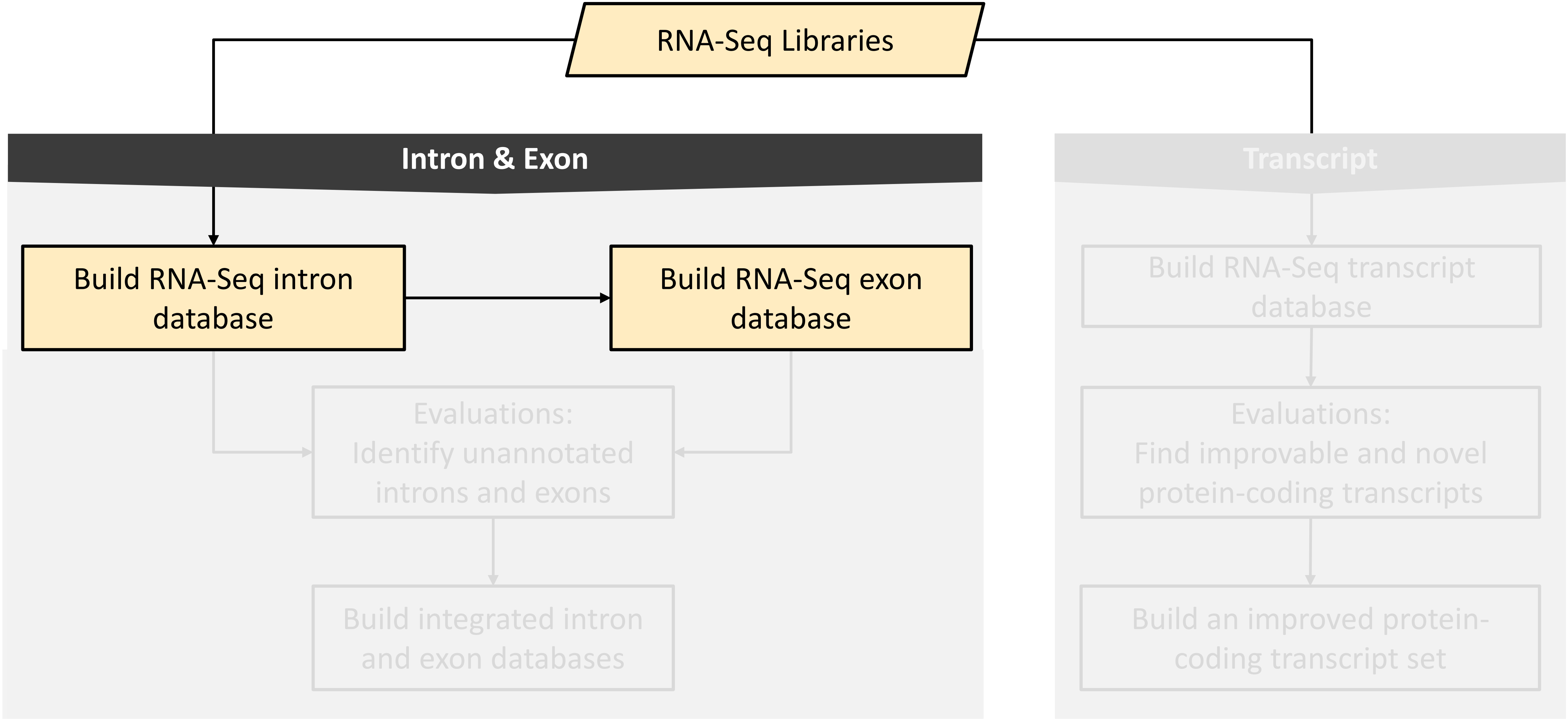
C. briggsae genome annotation is incomplete which limits its utility as a comparative platform for *C. elegans.*

- <u>Hypothesis #1:</u> The different number of molecular features observed between the two species is due to the incomplete annotation of *C. briggsae* genome. The *C. briggsae* annotation can be improved using RNA-Seq data.
- <u>Hypothesis #2:</u> Using the improved *C. briggsae* annotation, we can find additional orthologous relationships with *C. elegans* that were previously missed and additional *C. elegans* gene models.

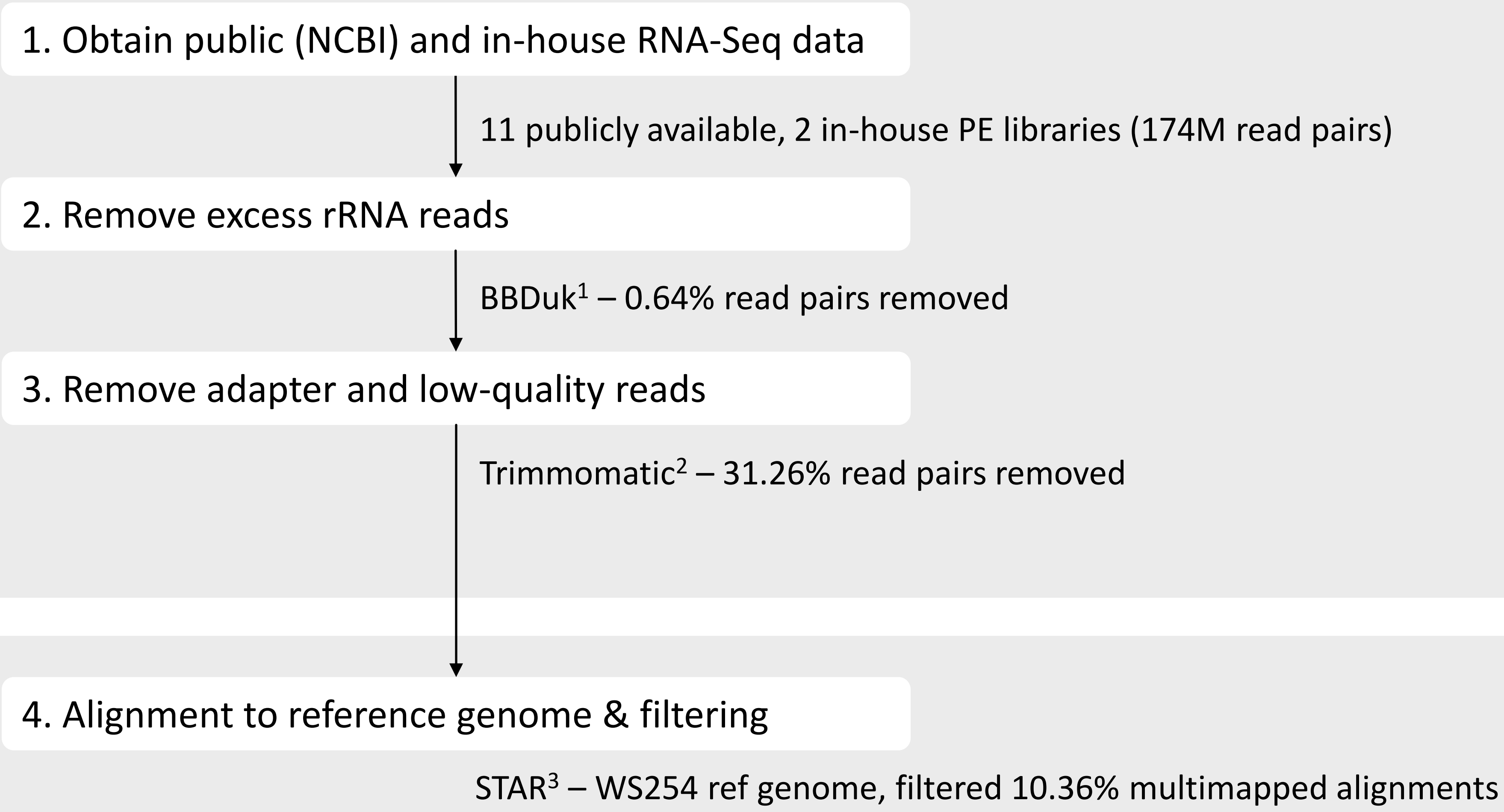
Specific Aims

- Aim #1: Improve the *C. briggsae* genome annotation at the intron, exon, transcript levels.
- Aim #2: Find additional orthologous relationships between *C. briggsae* and *C. elegans* and improve *C. elegans* genome annotation at the transcript level.

Aim 1: Improve *C. briggsae* genome annotation


RNA-Seq Libraries
Intron & Exon
Transcript
Build RNA-Seq intron database
Build RNA-Seq exon database
Build RNA-Seq transcript database
Evaluations: Identify unannotated introns and exons
Evaluations: Find improvable and novel protein-coding transcripts
Build integrated intron and exon databases
Build an improved protein-coding transcript set

Method: Building RNA-Seq intron and exon databases



Data Selection & pre-processing

1. Obtain public (NCBI) and in-house RNA-Seq data

11 publicly available, 2 in-house PE libraries (174M read pairs)

2. Remove excess rRNA reads

BBDuk[1] – 0.64% read pairs removed

3. Remove adapter and low-quality reads

Trimmomatic[2] – 31.26% read pairs removed

Alignment

4. Alignment to reference genome & filtering

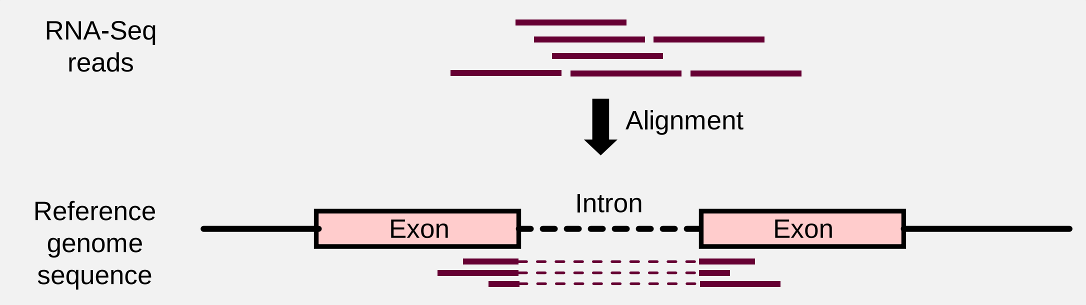
STAR[3] – WS254 ref genome, filtered 10.36% multimapped alignments

[1] https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/bbduk-guide; [2] Bolger et al., 2014; [3] Dobin et al., 2013

Method: Building RNA-Seq intron and exon databases

Build RNA-Seq intron database

5. Identify intron & filter



select introns with at least 5 read supports in 1 library

RNA-Seq intron database
(95,632 introns)

Build RNA-Seq exon database

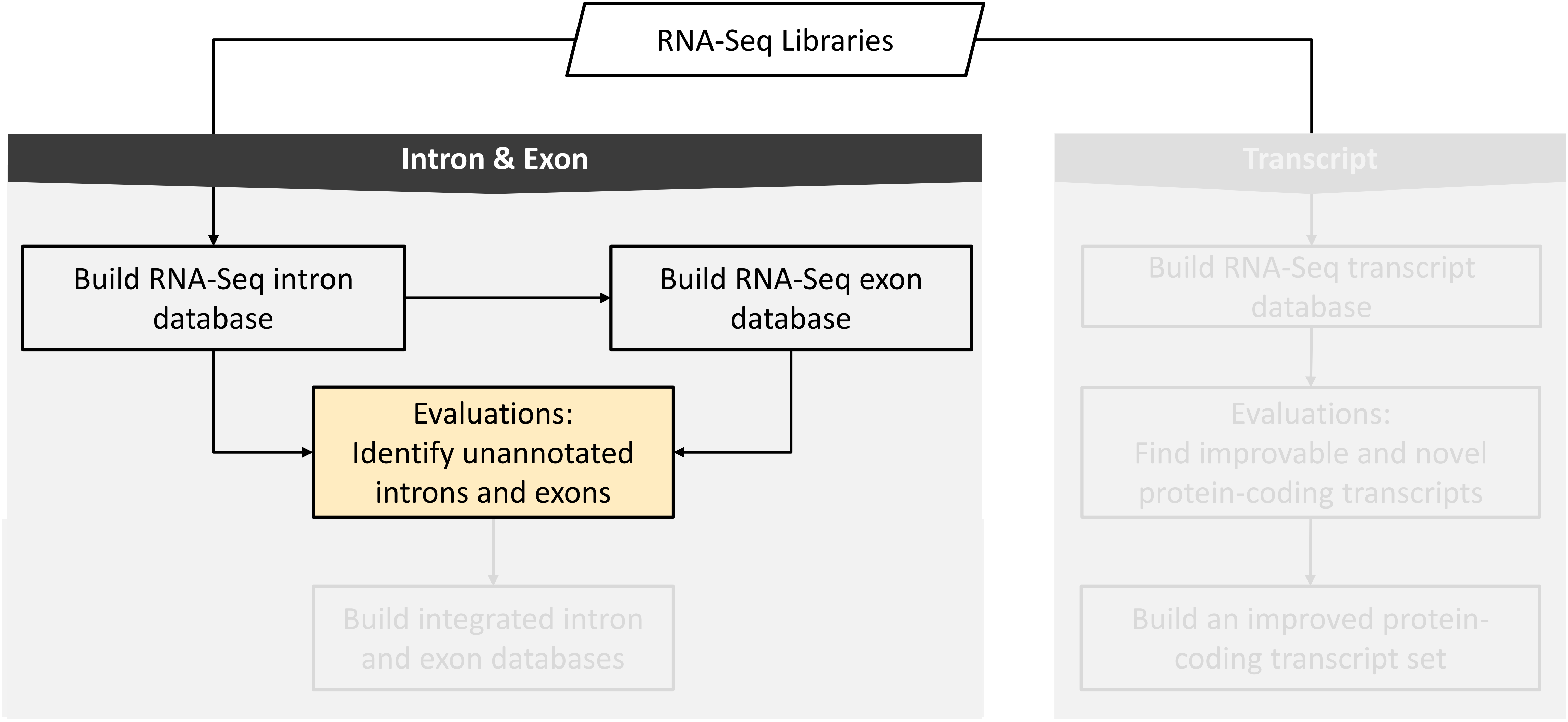
7. Reconstruct exons

ExonTrap[1]

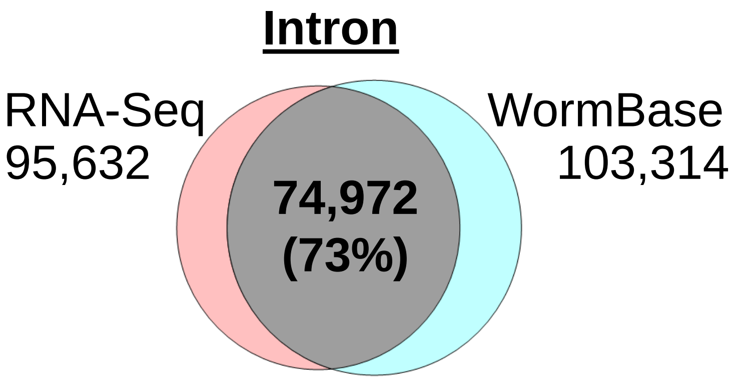
RNA-Seq exon database
(115,689 exons)

[1] Douglas, M., 2018

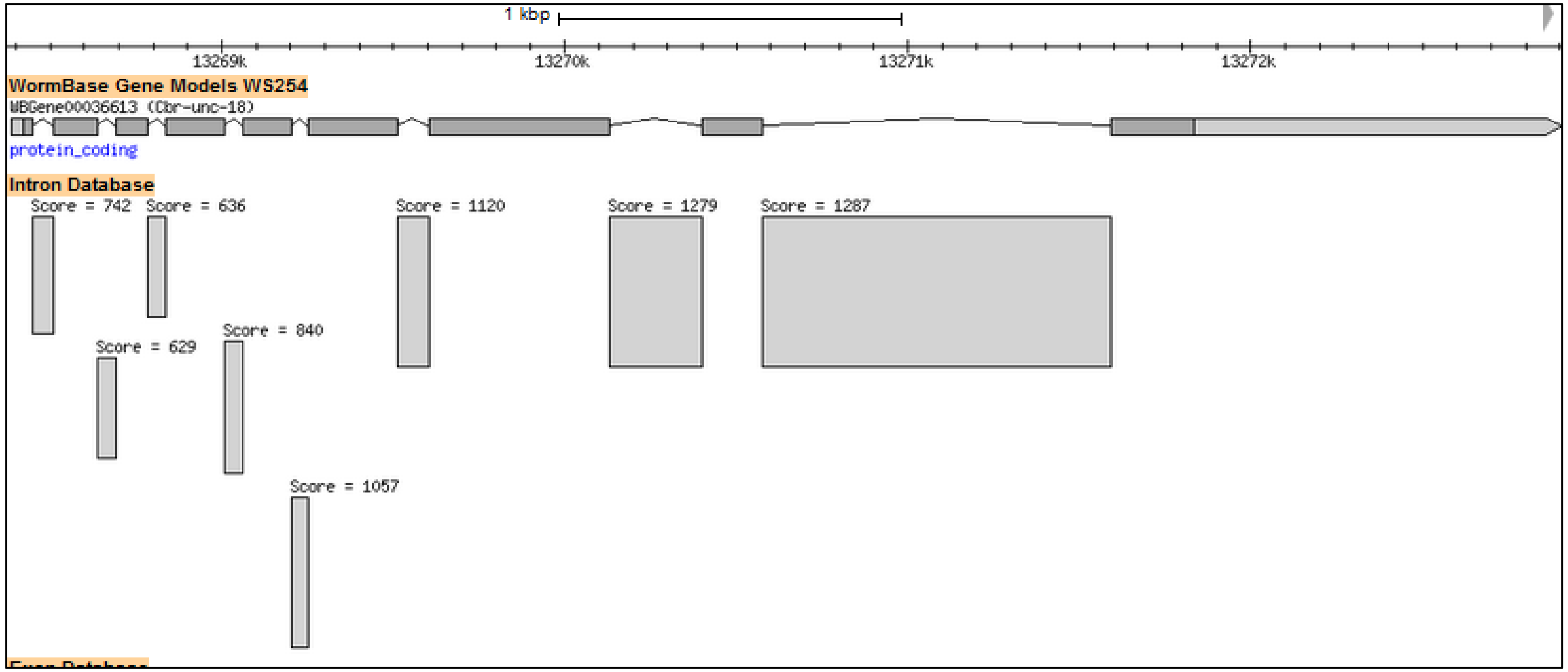
Aim 1: Improve *C. briggsae* genome annotation


RNA-Seq Libraries
Intron & Exon
Build RNA-Seq intron database
Build RNA-Seq exon database
Evaluations: Identify unannotated introns and exons
Build integrated intron and exon databases
Transcript
Build RNA-Seq transcript database
Evaluations: Find improvable and novel protein-coding transcripts
Build an improved protein-coding transcript set

RNA-Seq introns validated 73% of WormBase introns



All WormBase introns are supported by RNA-Seq introns

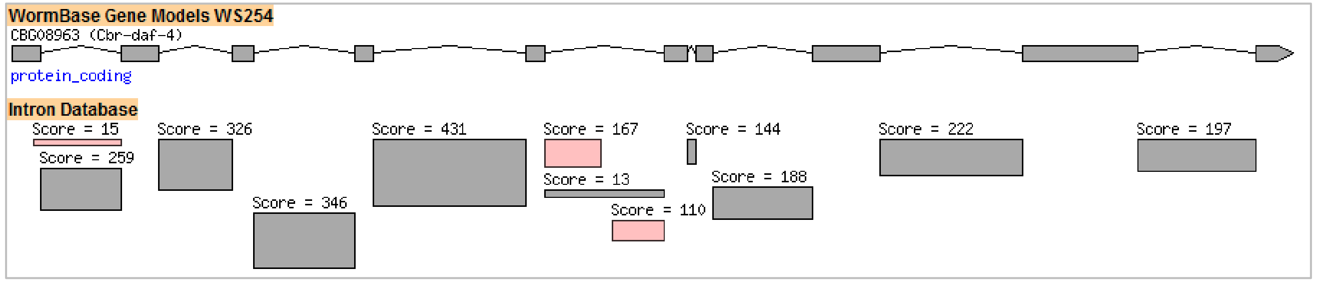


Pictured: *Cbr-unc-18* (uncoordinated), an ortholog of *C. elegans unc-18*

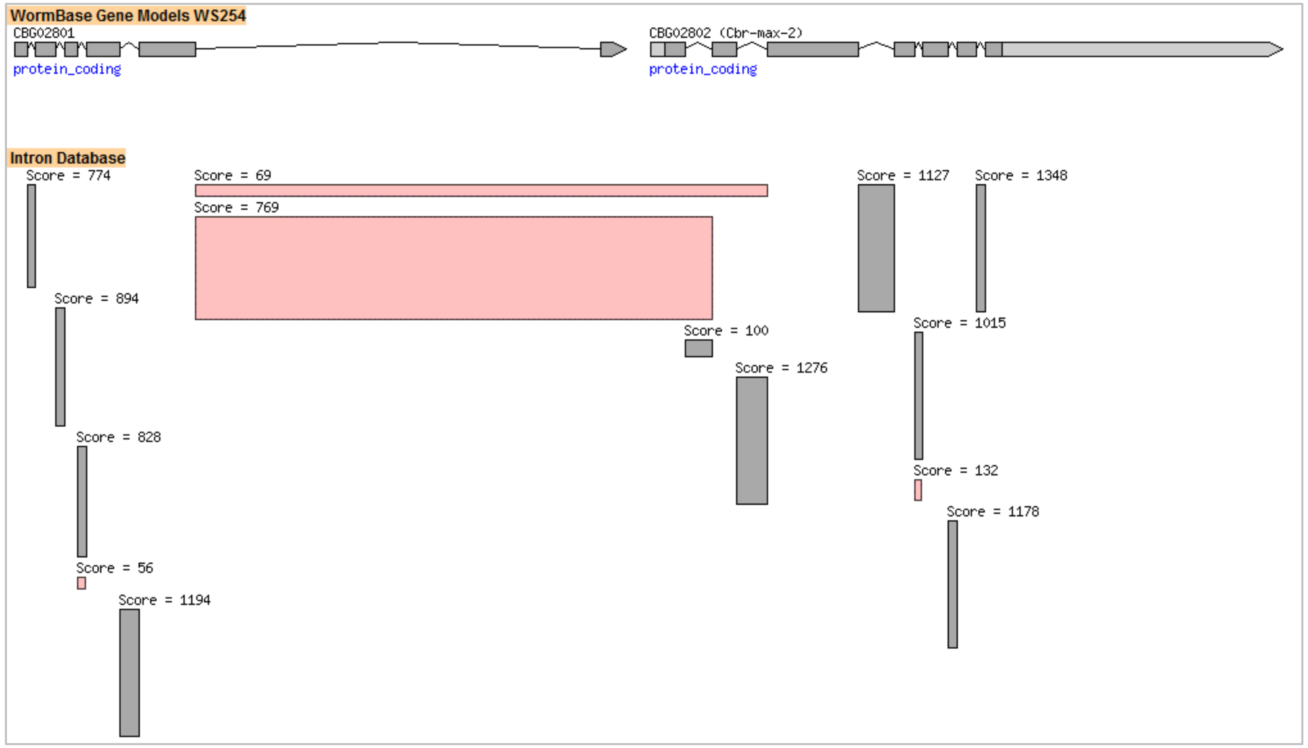
Validating ~¾ of WormBase introns demonstrates the utility of our intron database

WormBase = nematode information resource/database

20,660 novel RNA-Seq introns identified

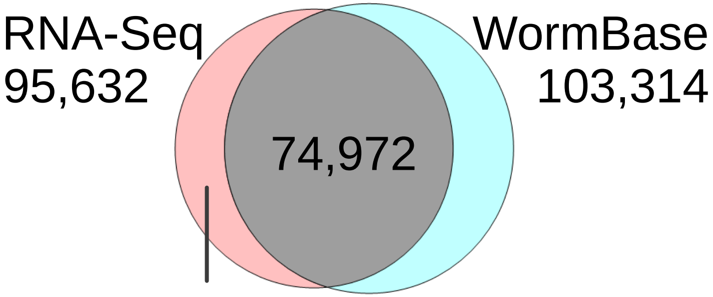


Pictured: *Cbr-daf-4* (abnormal dauer formation), an ortholog of *C. elegans daf-4*



Pictured: *Cbr-max-2* (motor axon guidance), an ortholog of *C. elegans max-2*

Intron



RNA-Seq specific

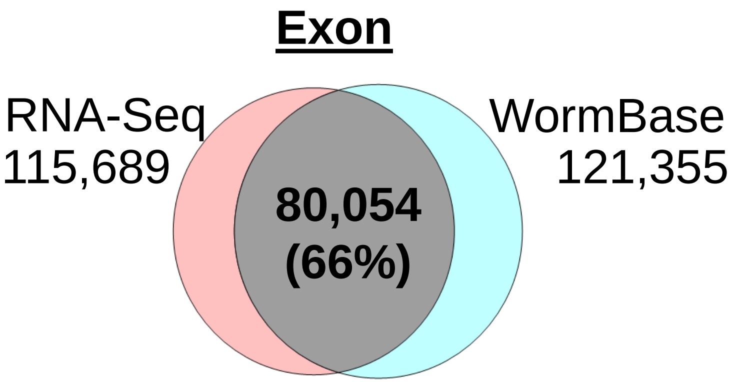
20,660 (22%)

9,516 protein-coding genes affected

Novel introns suggest gene model modifications and novel genes

- 73% located internal of existing genes (top left)
- 12% extending existing genes
- 1% merging existing genes (bottom left)
- 14% of introns did not map to existing genes

RNA-Seq exons validated 66% of WormBase exons



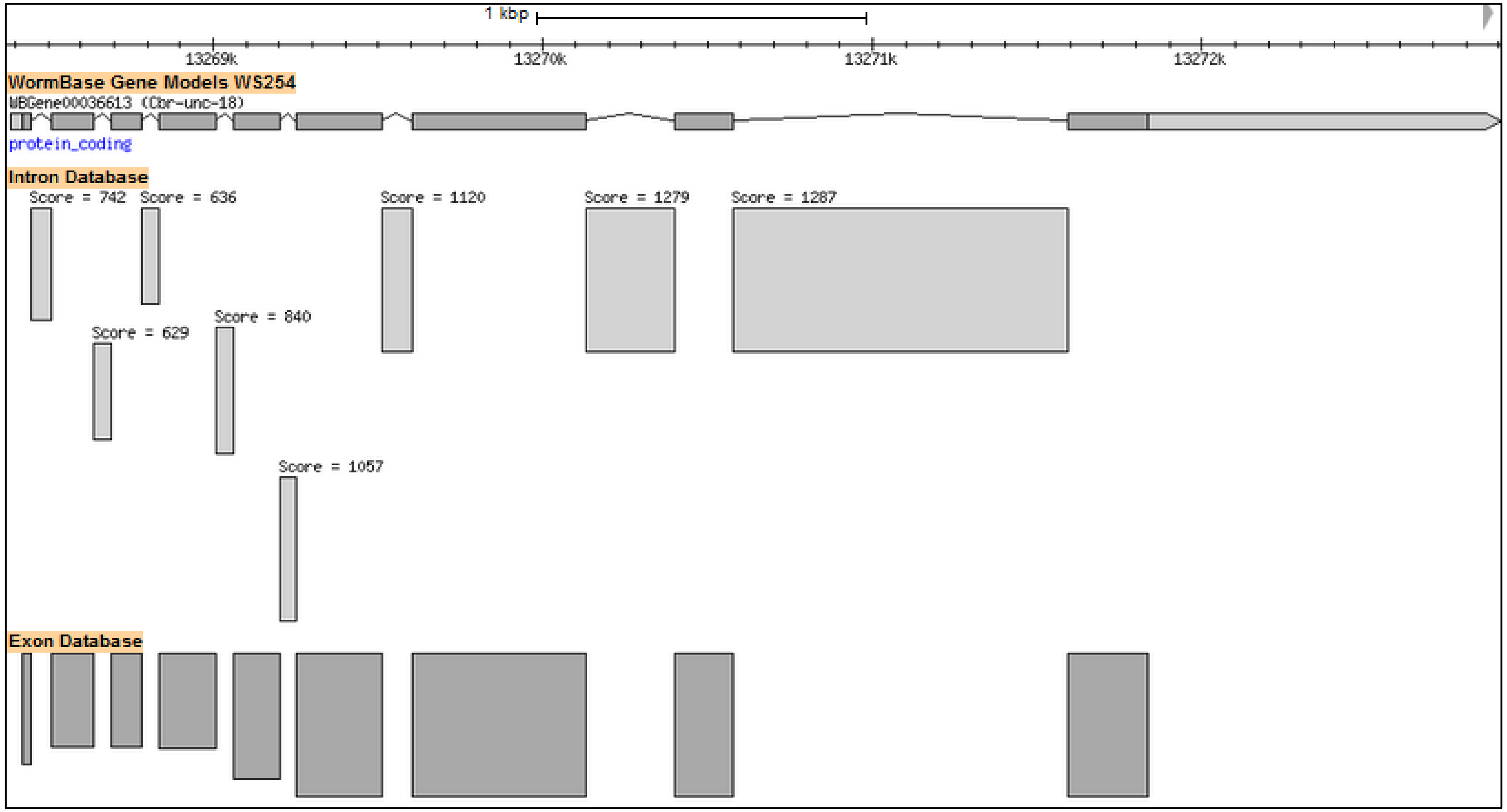
Exon

RNA-Seq
115,689

WormBase
121,355

80,054
(66%)

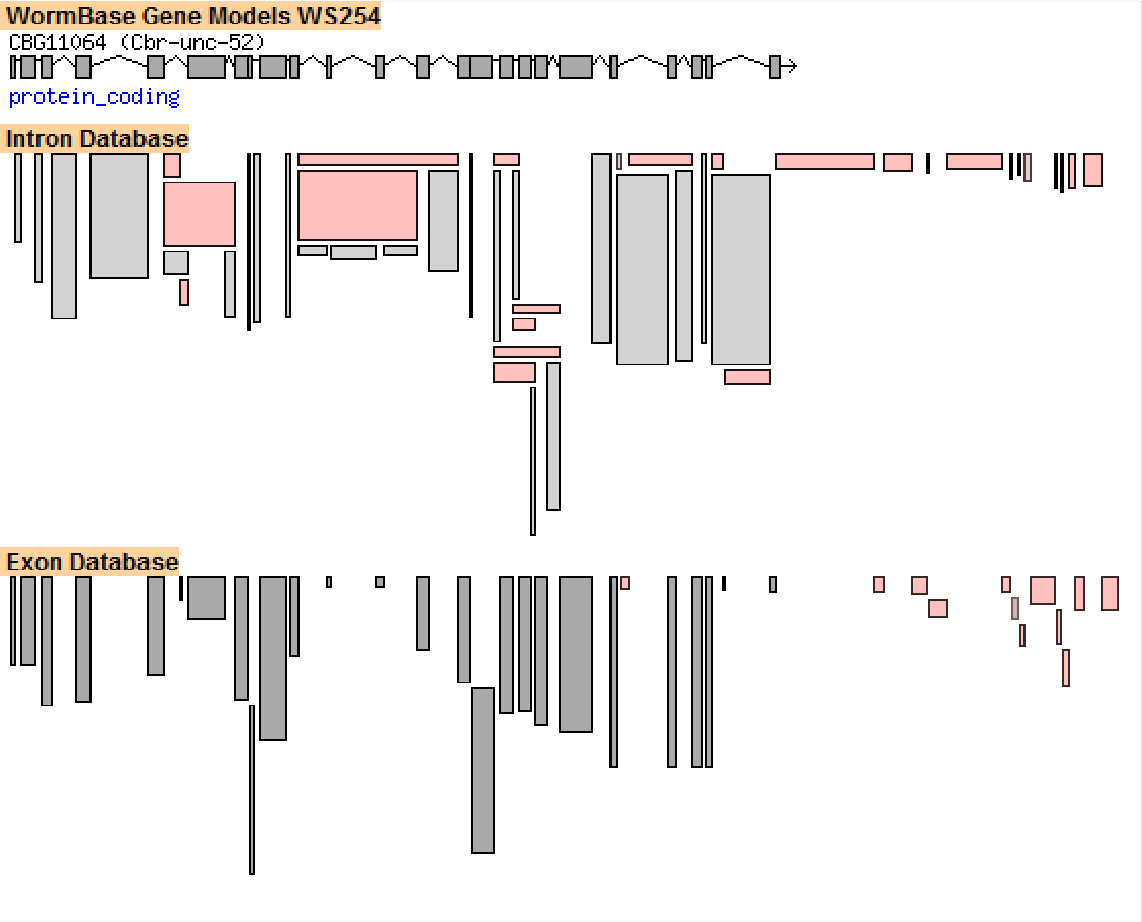
All WormBase exons are supported by RNA-Seq exons



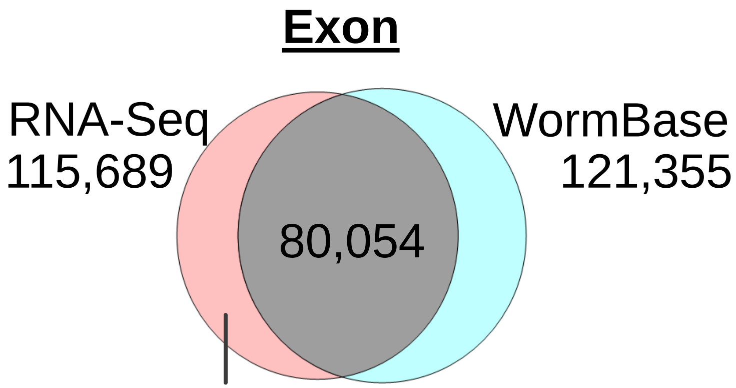
Pictured: *Cbr-unc-18* (uncoordinated), an ortholog of *C. elegans unc-18;* ~46% of WormBase transcripts are completely validated by our RNA-Seq introns and exons

Validating ⅔ of WormBase exons demonstrates the utility of our exon database

35,635 novel RNA-Seq exons identified



Pictured: *Cbr-unc-52* (uncoordinated), an ortholog of *C. elegans unc-52*; 31% of WormBase transcripts are partially validated by our RNA-Seq introns and exons

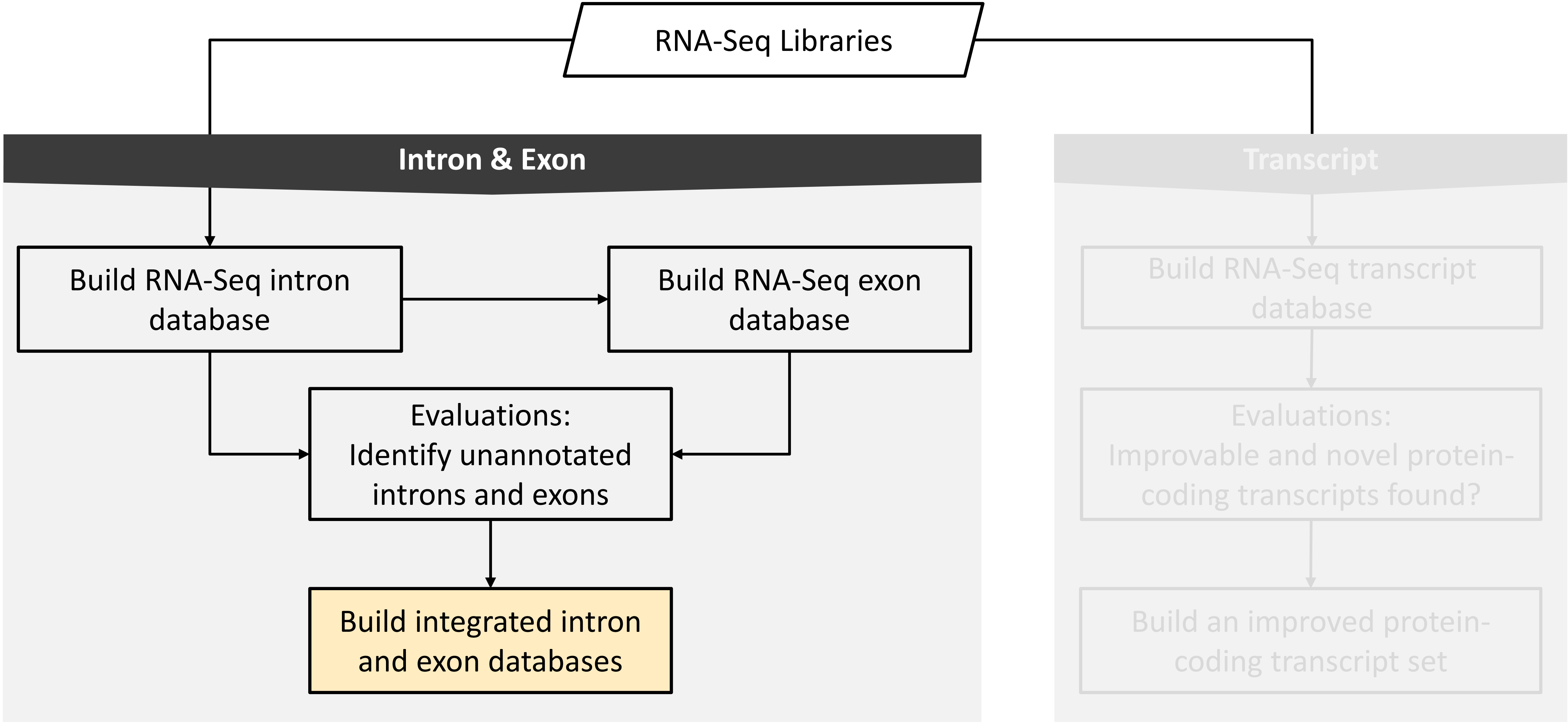


RNA-Seq specific
35,635 (31%)
15,187 protein-coding genes affected)

Novel exons suggest gene model modifications and novel genes

- 76% located internal of existing genes
- 14% extending existing genes (left)
- <1% merging existing genes
- 9% of exons did not map to existing genes

Aim 1: Improve *C. briggsae* genome annotation

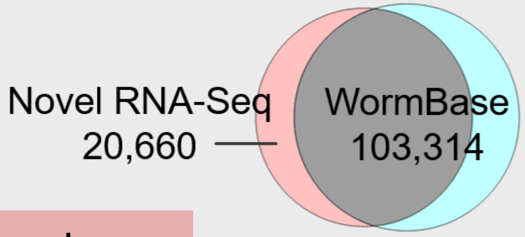


RNA-Seq Libraries
Intron & Exon
Build RNA-Seq intron database
Build RNA-Seq exon database
Evaluations: Identify unannotated introns and exons
Build integrated intron and exon databases
Transcript
Build RNA-Seq transcript database
Evaluations: Improvable and novel protein-coding transcripts found?
Build an improved protein-coding transcript set

Method: Building improved intron and exon databases

Build integrated intron database

Integrate RNA-Seq novel introns into WormBase introns



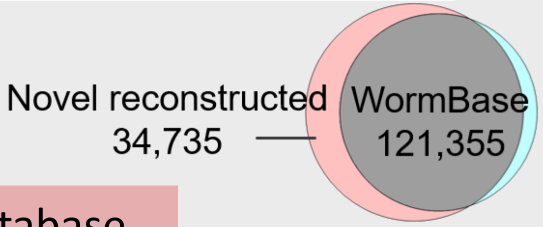
Integrated intron database (123,974 introns)

Build integrated exon database

Reconstruct exon

ExonTrap[1] – reconstructed 147,648 exons

Integrating novel exons into WormBase exons



Integrated exon database (156,090 exons)

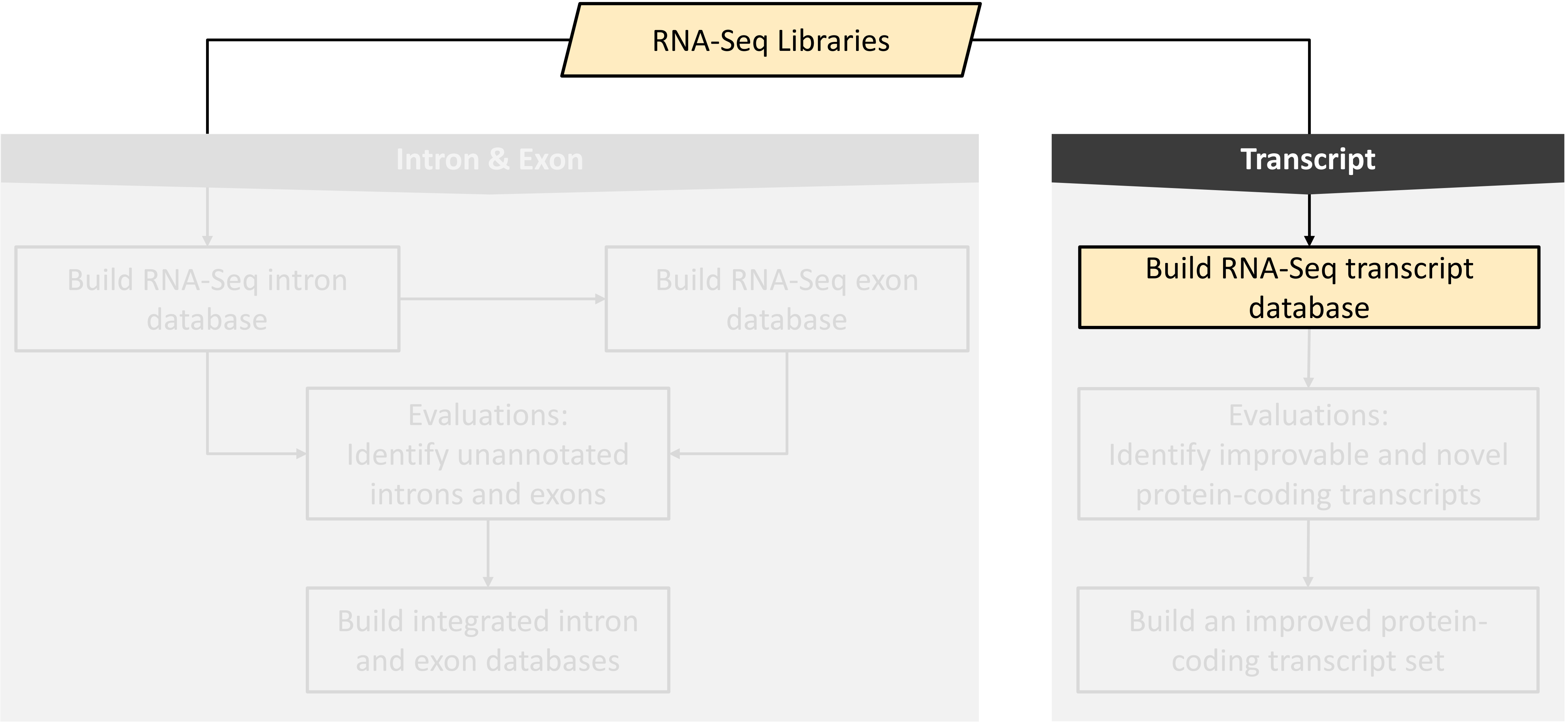
[1] Douglas, M., 2018

Intron and Exon Summary

- **Evidence that the *C. briggsae* annotation is incomplete and can be improved using 13 RNA-Seq libraries**
  - Identified 20,660 novel introns and 35,635 novel exons
  - Validated 73% and 66% WormBase introns and exons

- **Built improved intron and exon databases that serve as a more complete annotation at the intron and exon level**
  - Intron database consisting of 123,974 introns
  - Exon database consisting of 150,690 exons

Aim 1: Improve *C. briggsae* genome annotation


RNA-Seq Libraries
Intron & Exon
Build RNA-Seq intron database
Build RNA-Seq exon database
Evaluations: Identify unannotated introns and exons
Build integrated intron and exon databases
Transcript
Build RNA-Seq transcript database
Evaluations: Identify improvable and novel protein-coding transcripts
Build an improved protein-coding transcript set

Method: Building RNA-Seq transcript database

Build transcript database supported by introns and exons

1. *De-novo* and genome-guided transcript assembly

Cufflinks[1], StringTie[2], TransABySS[3]

2. Select transcripts supported by integrated introns and exons and merge transcripts

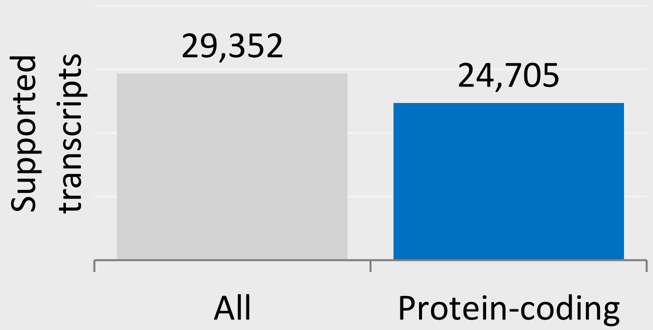
Gffcompare[4] – 29,352 fully supported transcripts

Predict coding regions

3. Identify candidate coding transcripts

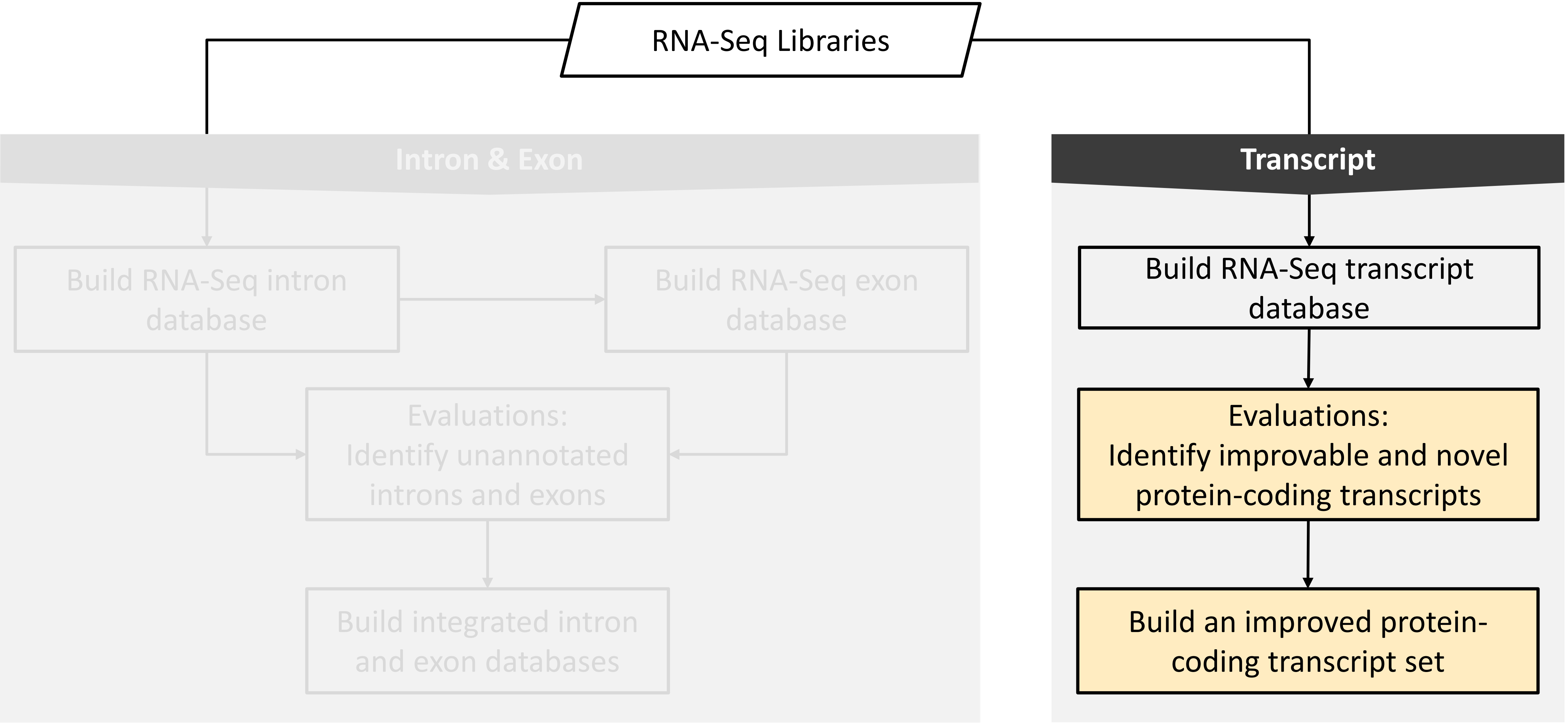
TransDecoder[5]

Protein-coding transcripts **(24,705 transcripts)**



[1] Trapnell et al., 2012; [2] Pertea et al., 2015; [3] Robertson et al., 2010; [4] https://github.com/gpertea/gffcompare; [5] http://github.com/TransDecoder/TransDecoder

Aim 1: Improve *C. briggsae* genome annotation


RNA-Seq Libraries
Intron & Exon
Build RNA-Seq intron database
Build RNA-Seq exon database
Evaluations: Identify unannotated introns and exons
Build integrated intron and exon databases
Transcript
Build RNA-Seq transcript database
Evaluations: Identify improvable and novel protein-coding transcripts
Build an improved protein-coding transcript set

Method: transcript-to-transcript comparison (intron chain)

WormBase transcript model

complete match

1. Complete match (WB confirmed)

partial match

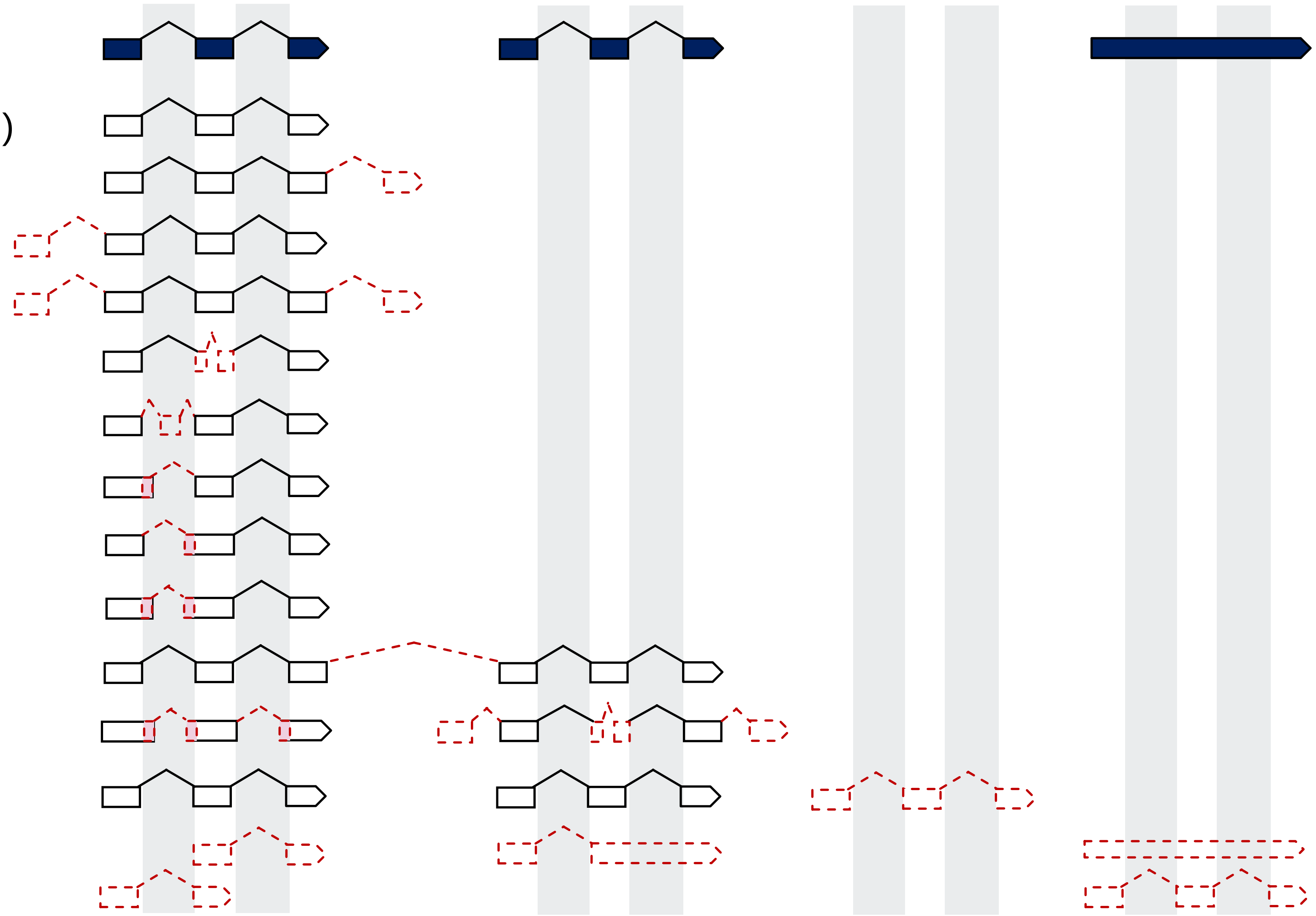
2. 3’ extension
3. 5’ extension
4. 5’ & 3’ extension
5. Intron overlapping internal exon
6. Introns overlapping intron
7. Alternative donor (5’ss)
8. Alternative acceptor (3’ss)
9. Alternative donor & acceptor
10. Merged
11. Complex changes

no match

12. Novel

partial match

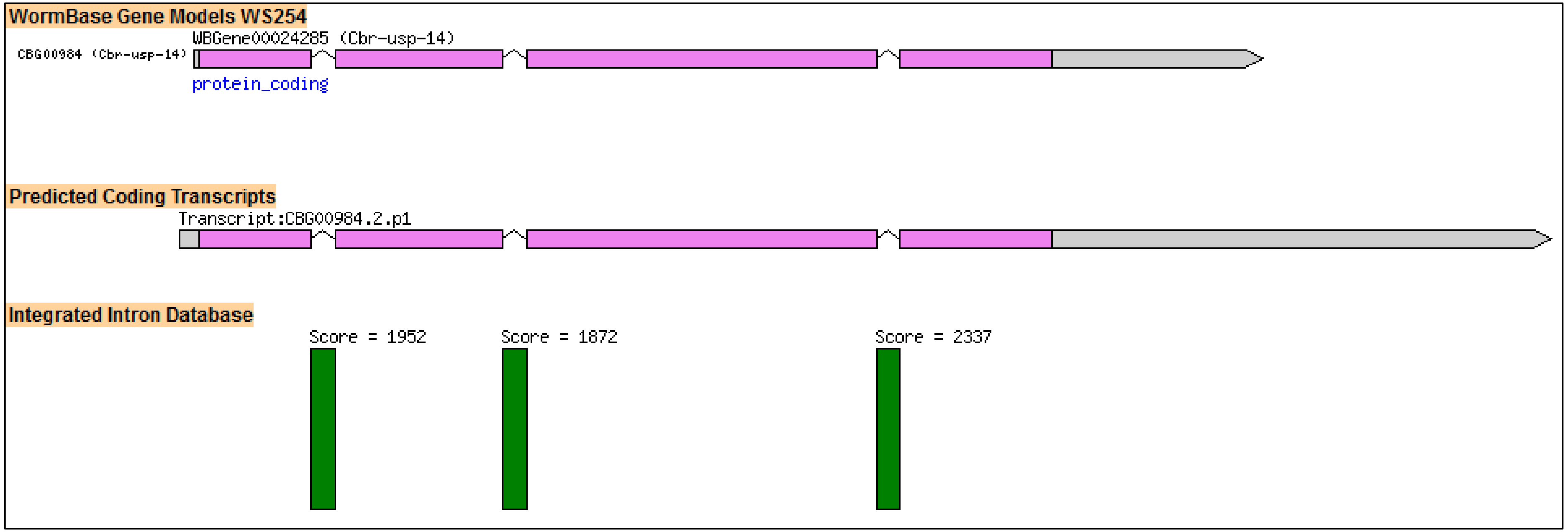
13. Other (single-exon, partial)



Category 1: Complete match

Criteria:

- Both assembled transcript and WB transcript have identical intron chains

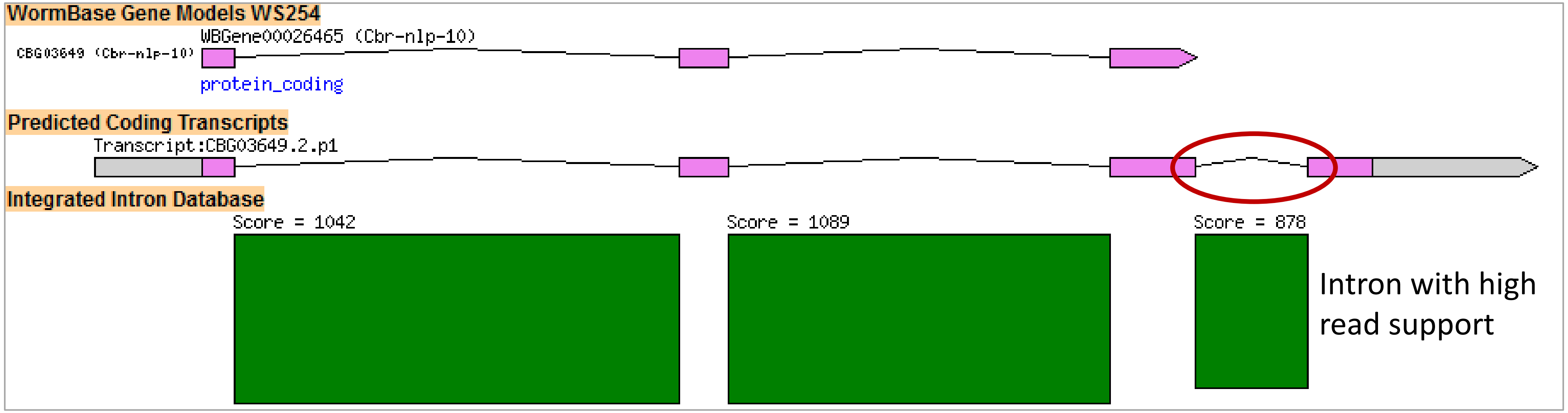

WormBase Gene Models WS254
WBGene00024285 (Cbr-usp-14)
CBG00984 (Cbr-usp-14)
protein_coding
Predicted Coding Transcripts
Transcript:CBG00984.2.p1
Integrated Intron Database
Score = 1952
Score = 1872
Score = 2337

Pictured: transcript CBG00984.2 of *Cbr-usp-14* (ubiquitin specific protease), ortholog of *C. elegans' usp-14.* All introns in WormBase CBG00984 transcript are observed in the assembled transcript

Category 2: partial match - 3’ extension

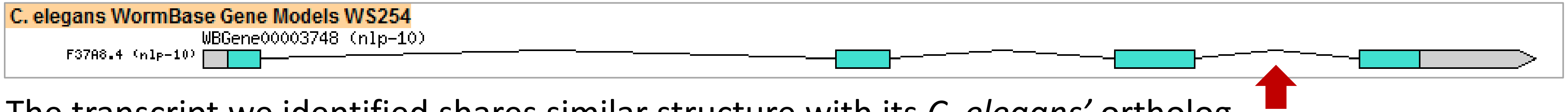
Criteria:

- All introns in the annotated transcript are observed in the assembled transcript
- One or more additional introns are found extending 3’ of the transcript



WormBase Gene Models WS254
WBGene00026465 (Cbr-nlp-10)
CBG03649 (Cbr-nlp-10)
protein_coding
Predicted Coding Transcripts
Transcript:CBG03649.2.p1
Integrated Intron Database
Score = 1042
Score = 1089
Score = 878

Intron with high read support



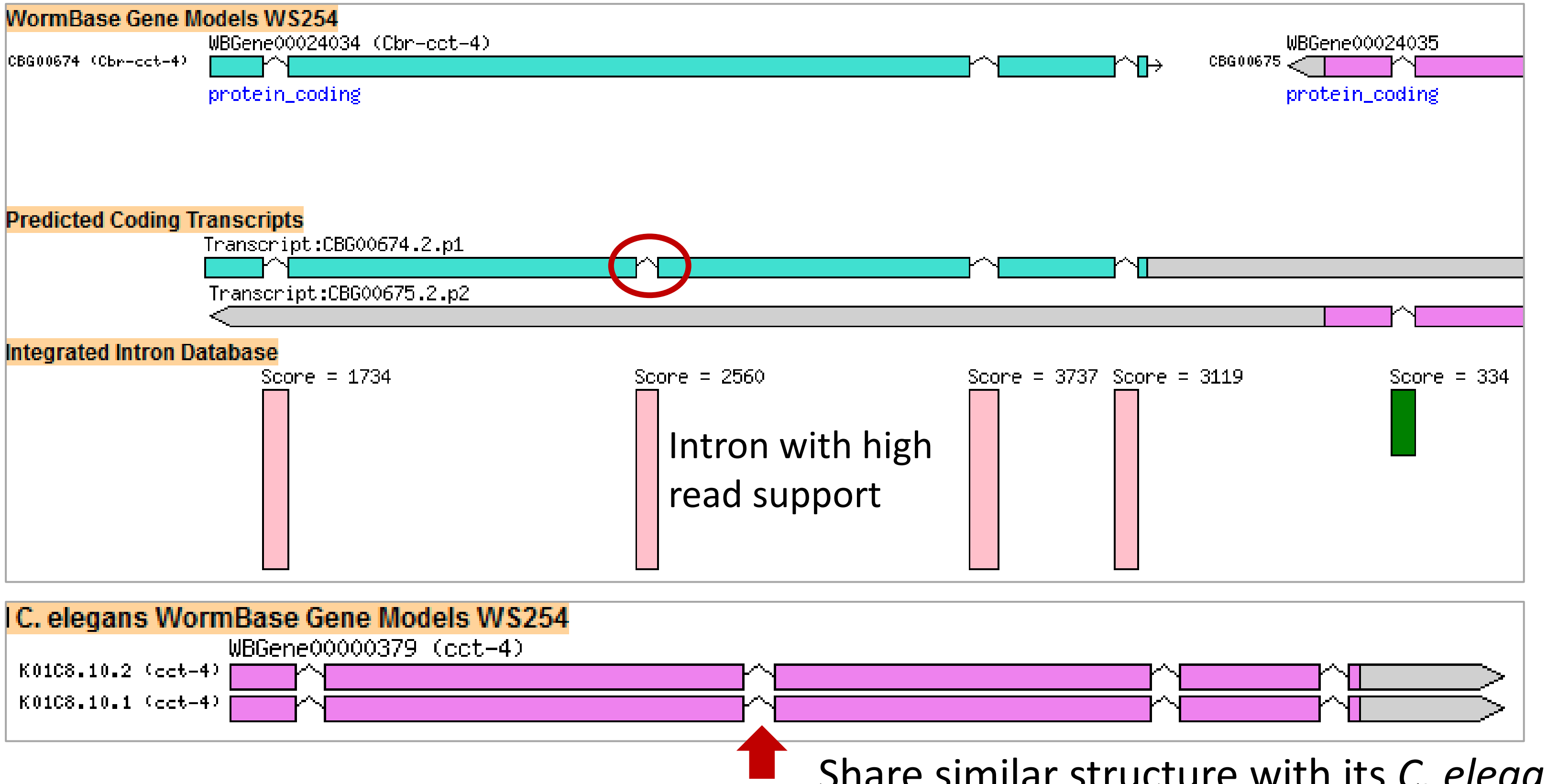
The transcript we identified shares similar structure with its *C. elegans’* ortholog

Pictured: (top) transcript CBG0649.2 of *Cbr-nlp-10* (neuropeptide-like protein), ortholog of *C. elegans’ nlp-10* (bottom)

Category 2: partial match – intron overlapping internal exon

Criteria:

- All introns in the annotated transcripts are observed in the assembled transcripts
- One more additional intron is found overlapping an annotated internal exon of existing transcripts
- Both transcripts have the same leftmost and rightmost intron boundaries



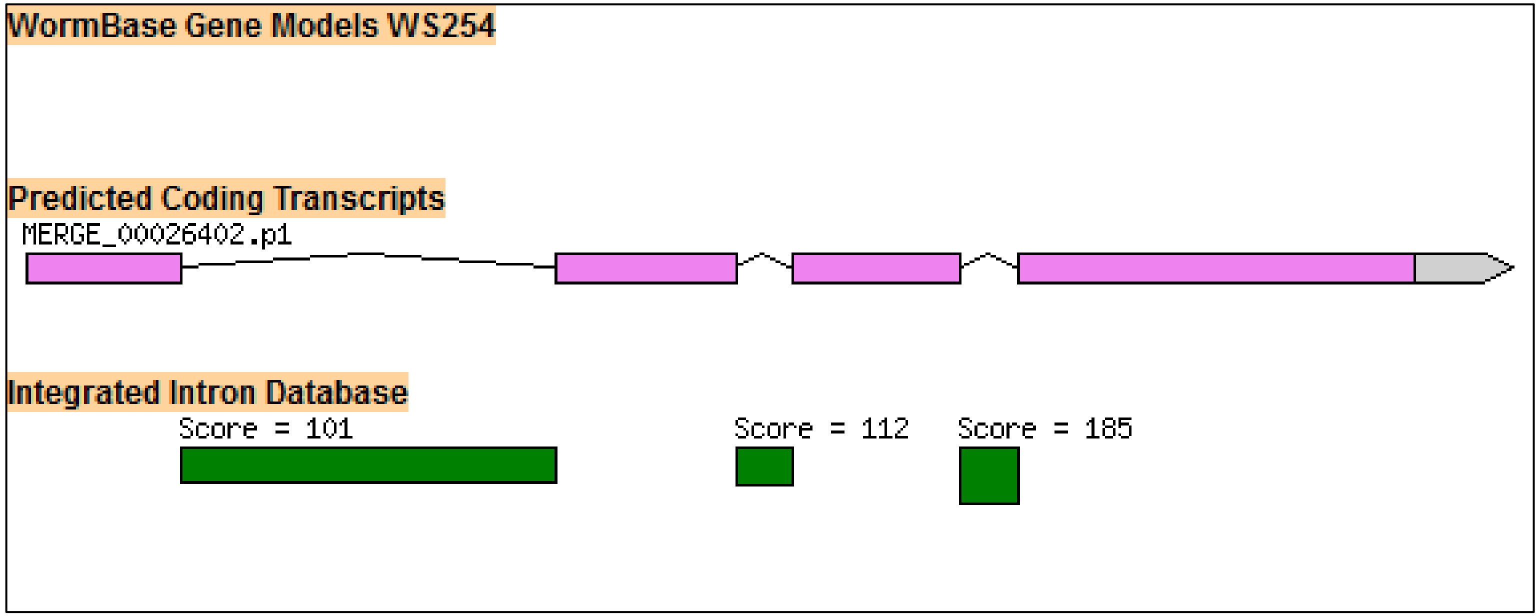
Intron with high read support

Share similar structure with its *C. elegans'* ortholog

Category 12: No match - Novel

Criteria:

- Has not been annotated previously
- No overlapping genes in the genomic regions


WormBase Gene Models WS254
Predicted Coding Transcripts
MERGE_00026402.p1
Integrated Intron Database
Score = 101
Score = 112
Score = 185

Pictured: Transcript MERGE_00026402 (nCBG00109). Previously not annotated by WormBase

Method: transcript-to-transcript comparison (intron chain)

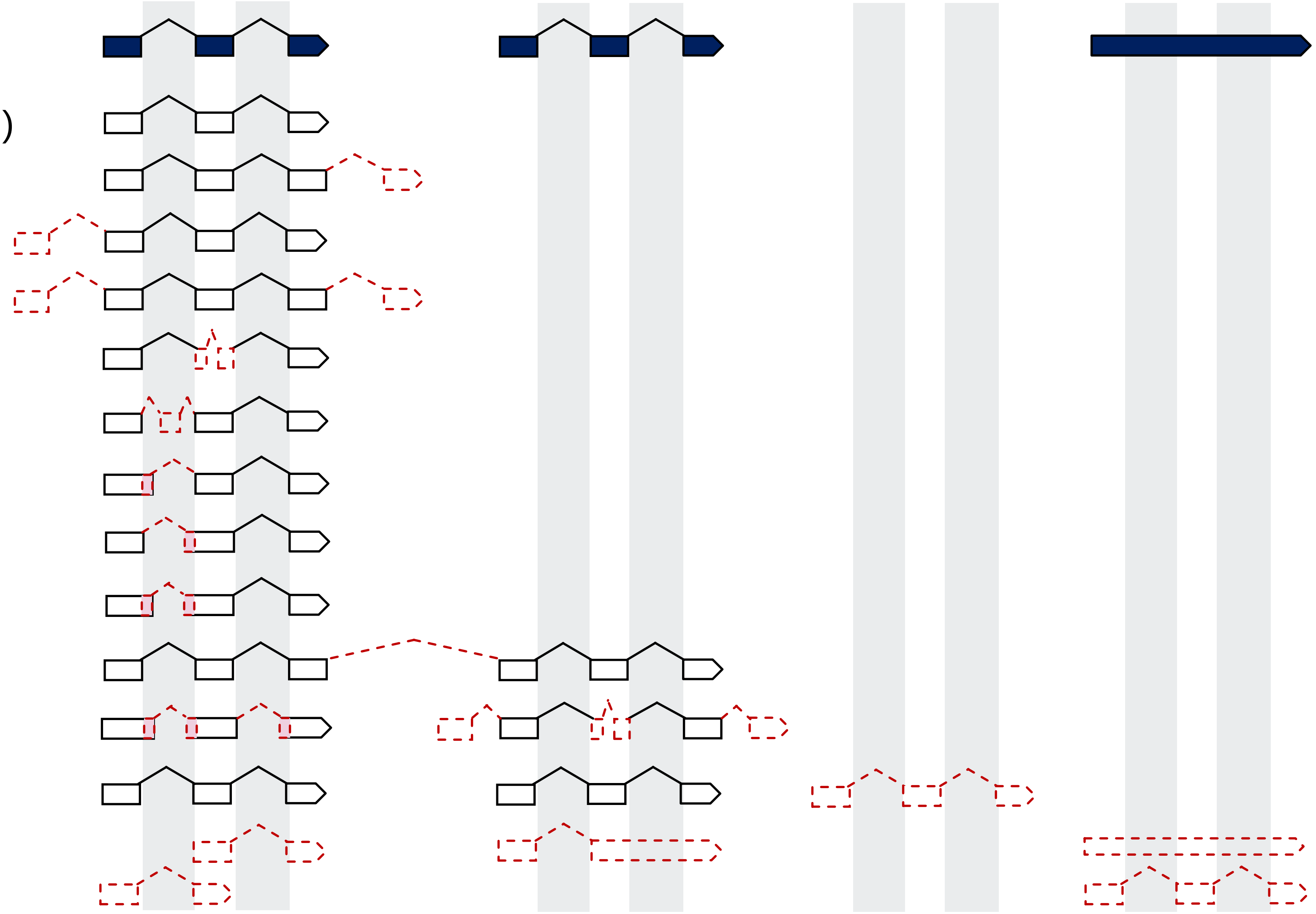
WormBase transcript model

partial match

1. Complete match (WB confirmed)
2. 3’ extension
3. 5’ extension
4. 5’ & 3’ extension
5. Intron overlapping internal exon
6. Introns overlapping intron
7. Alternative donor (5’ss)
8. Alternative acceptor (3’ss)
9. Alternative donor & acceptor
10. Merged
11. Complex changes

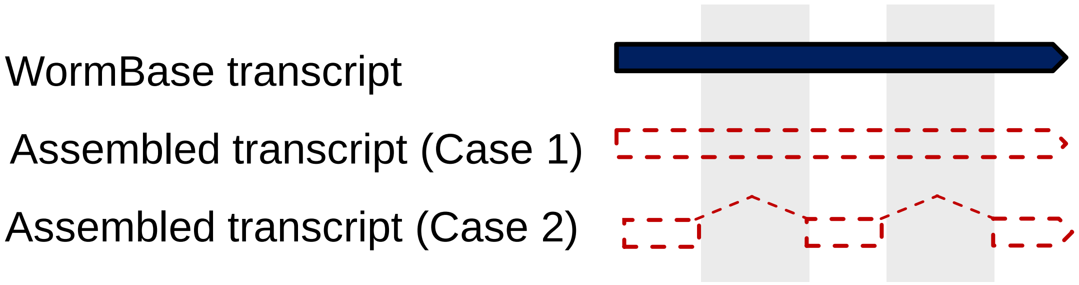
no match

12. Novel
13. Other (single-exon, partial)

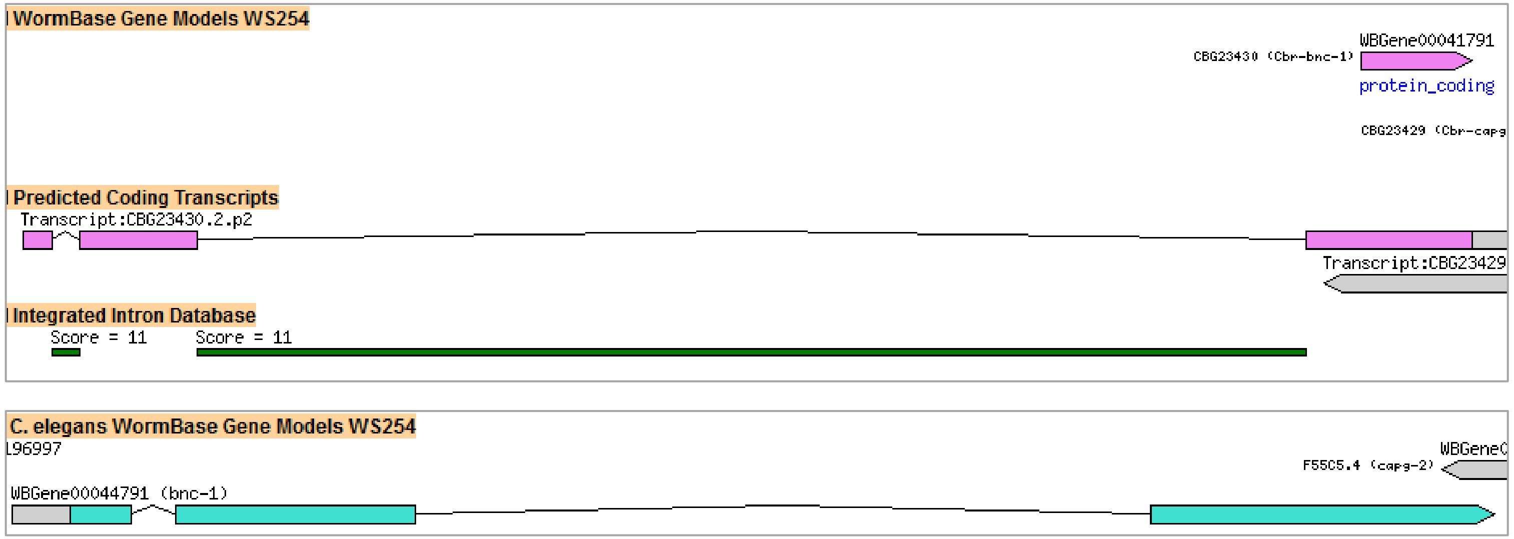


Limitation: single-exon transcripts excluded

- Using intron chain comparison method, a transcript is discarded when either WormBase or assembled transcript is single-exon (i.e., no intron)



Case 2



WormBase Gene Models WS254

WBGene00041791
CBG23430 (Cbr-bnc-1)
protein_coding
CBG23429 (Cbr-capg

Predicted Coding Transcripts
Transcript:CBG23430.2.p2
Transcript:CBG23429

Integrated Intron Database
Score = 11 Score = 11

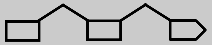
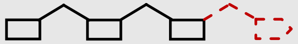
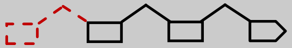
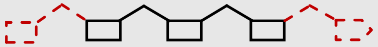
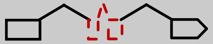
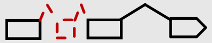
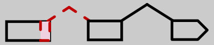
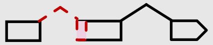
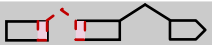
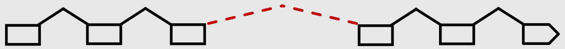
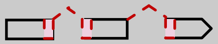
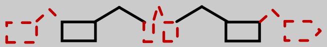
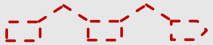
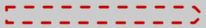
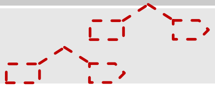
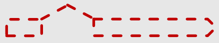
C. elegans WormBase Gene Models WS254
L96997
WBGene0
F55C5.4 (capg-2)
WBGene00044791 (bnc-1)

Single-exon

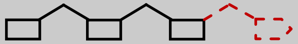
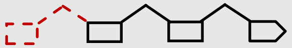
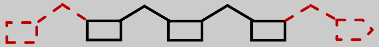
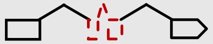
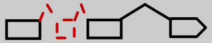
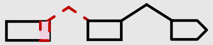
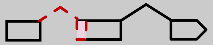
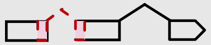
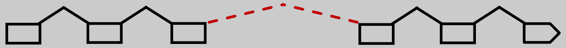
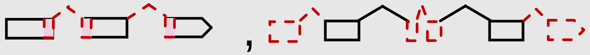
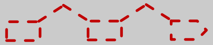
Multi-exon

Pictured: transcript CBG23430.2 of *Cbr-bnc-1* (basonuclin-1 zinc finger protein homolog), ortholog of *C. elegans' bnc-1*

Identified 6,285 candidate protein-coding transcripts

#	Category	Diagram	Protein-coding transcripts	Protein-coding genes
1.	Complete match (WB confirmed)		8,080 (32.71%)	8,055
2.	3’ extension		316 (1.28%)	287
3.	5’ extension		753 (3.05%)	687
4.	5’ & 3’ extension		26 (0.11%)	25
5.	Intron overlapping internal exon		358 (1.45%)	332
6.	Introns overlapping intron		217 (0.88%)	205
7.	Alternative donor (5’ss)		777 (3.15%)	746
8.	Alternative acceptor (3’ss)		882 (3.57%)	810
9.	Alternative donor & acceptor		346 (1.40%)	327
10.	Merging 2 or more genes		206 (0.83%)	116
11.	Complex changes	 , 	2,245 (9.09%)	1,517
12.	Novel		159 (0.64%)	159
13.	Other – Single-exon (no intron)		120 (0.49%)	95
	Other – Partial	 , 	10,220 (41.37%)	5,304

Method: Building an improved transcript database

Category		Diagram	Protein-coding transcripts	With proper start & stop codon
2.	3’ extension		316	284
3.	5’ extension		753	692
4.	5’ & 3’ extension		26	23
5.	Intron overlapping internal exon		358	341
6.	Introns overlapping intron		217	197
7.	Alternative donor (5’ss)		777	717
8.	Alternative acceptor (3’ss)		882	818
9.	Alternative donor & acceptor		346	284
10.	Merged		206	179
11.	Complex changes		2,245	2,015
12.	Novel		159	104
	Total		6,285 (100.00%)	**5,654 (89.96%)**

Integrated protein-coding transcripts
(28,129 transcripts)

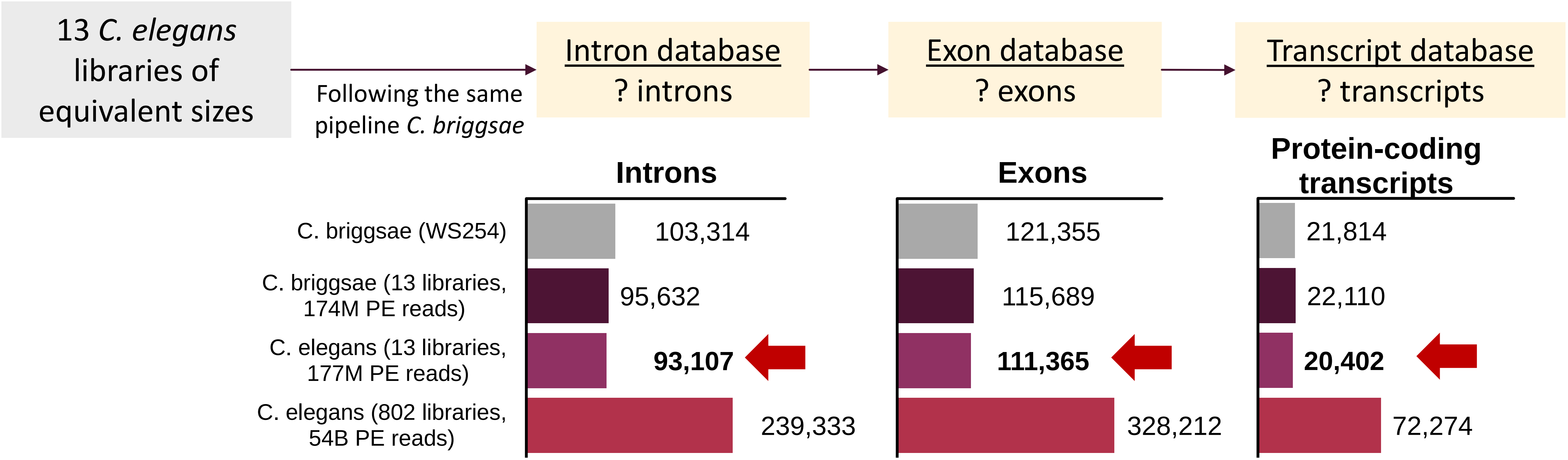
Transcript summary

- **Evidence that the *C. briggsae* annotation is incomplete and can be improved using 13 RNA-Seq libraries**
  - Identified 24,705 protein-coding transcripts, including 5,654 candidate protein-coding transcripts to improve *C. briggsae* annotation

- **Generated an improved *C. briggsae* transcript database consisting of 28,129 transcripts (25% higher than current annotation)**

Discovery power is proportional to RNA-Seq data quantity

- Using 13 *C. briggsae* RNA-Seq libraries, we identified thousands of introns, exons, and transcripts. We hypothesized that data availability does limit the discovery of features.
- Method: Applied the same pipeline on limited *C. elegans* RNA-Seq data



13 *C. elegans* libraries of equivalent sizes

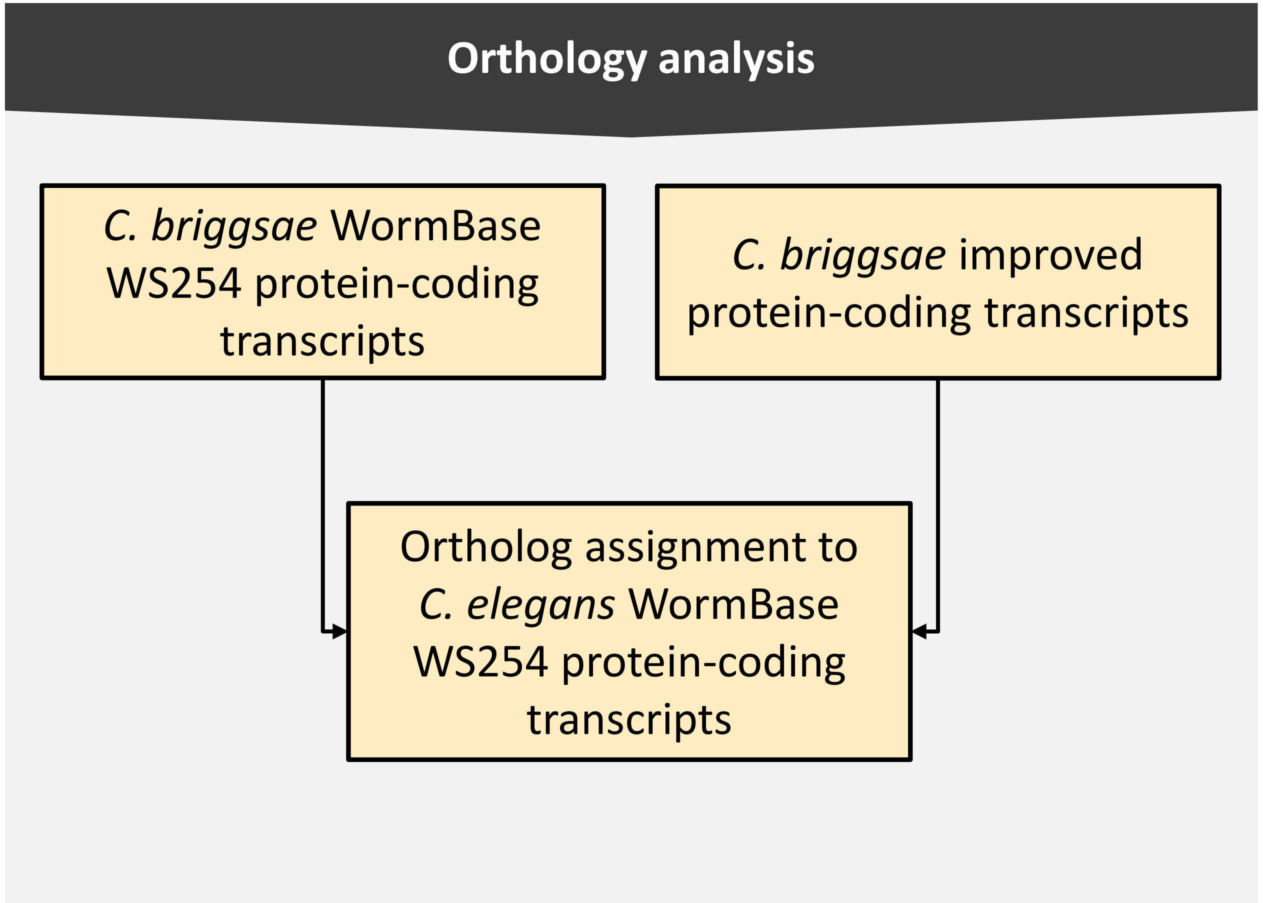
Following the same pipeline *C. briggsae*

Intron database
? introns

Exon database
? exons

Transcript database
? transcripts

Aim 2: Find additional orthologs and improve *C. elegans* annotation

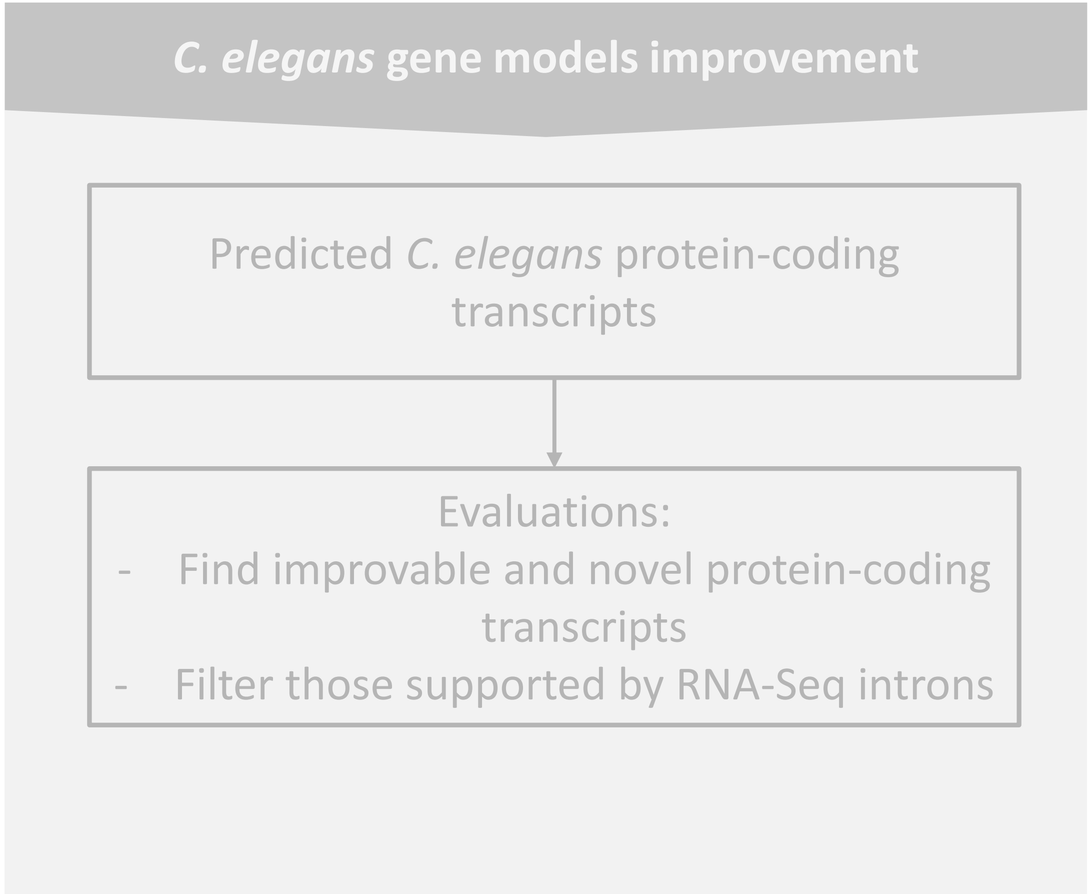


Orthology analysis

C. briggsae WormBase WS254 protein-coding transcripts

C. briggsae improved protein-coding transcripts

Ortholog assignment to *C. elegans* WormBase WS254 protein-coding transcripts



***C. elegans* gene models improvement**

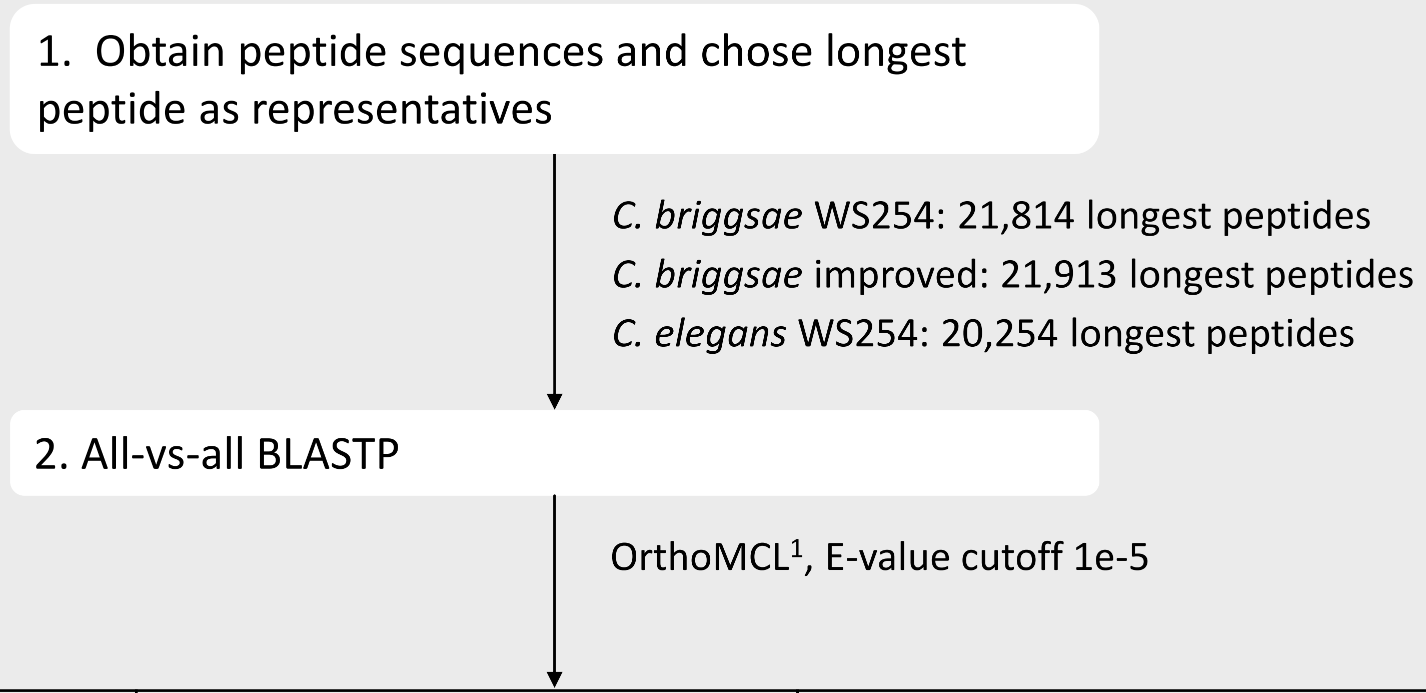
Predicted *C. elegans* protein-coding transcripts

Evaluations:

- Find improvable and novel protein-coding transcripts
- Filter those supported by RNA-Seq introns

32 ortholog pairs were found from novel *C. briggsae* genes

Orthology analysis before and after *C. briggsae* transcript improvement



1. Obtain peptide sequences and chose longest peptide as representatives

C. briggsae WS254: 21,814 longest peptides

C. briggsae improved: 21,913 longest peptides

C. elegans WS254: 20,254 longest peptides

2. All-vs-all BLASTP

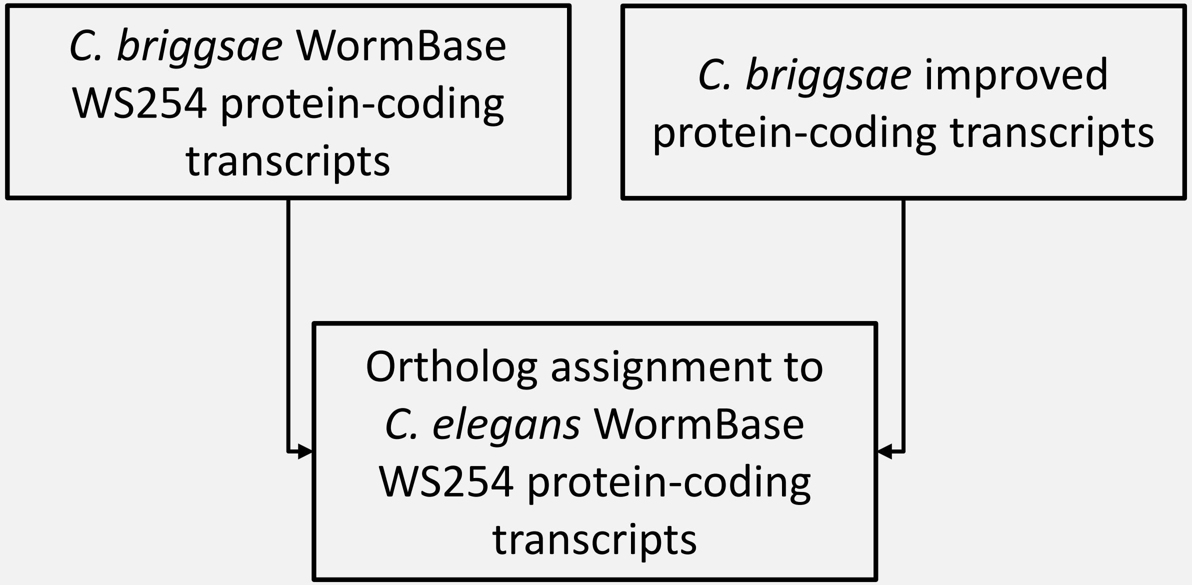
OrthoMCL[1], E-value cutoff 1e-5

	***C. briggsae* WS254 transcripts**	***C. briggsae* improved transcripts**
***C. elegans* WS254 transcripts**	Ortholog pairs = 16,748	Ortholog pairs = 16,880 (**32 pairs belong to novel genes**)

[1] Li et al., 2003

Aim 2: Find additional orthologs and improve *C. elegans* annotation

Orthology analysis

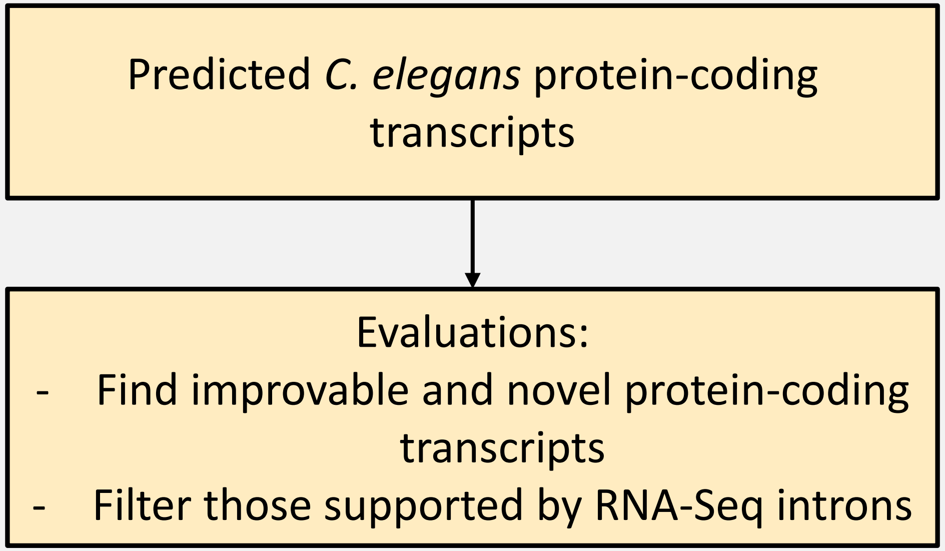


C. briggsae WormBase WS254 protein-coding transcripts

C. briggsae improved protein-coding transcripts

Ortholog assignment to *C. elegans* WormBase WS254 protein-coding transcripts

C. elegans gene models improvement

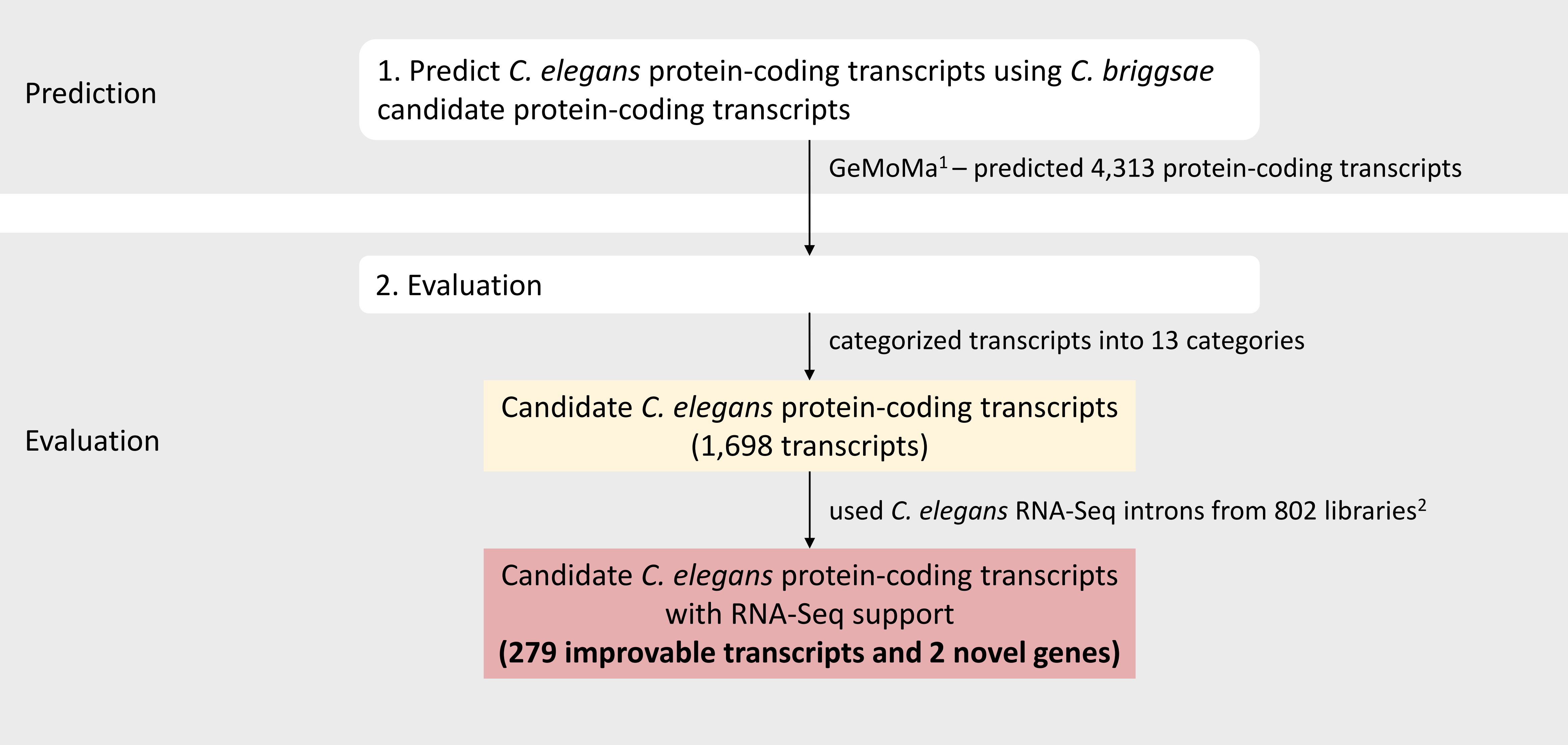


Predicted *C. elegans* protein-coding transcripts

Evaluations:

- Find improvable and novel protein-coding transcripts
- Filter those supported by RNA-Seq introns

Method: *C. elegans* gene model improvement



Prediction

1. Predict *C. elegans* protein-coding transcripts using *C. briggsae* candidate protein-coding transcripts

GeMoMa[1] – predicted 4,313 protein-coding transcripts

Evaluation

2. Evaluation

categorized transcripts into 13 categories

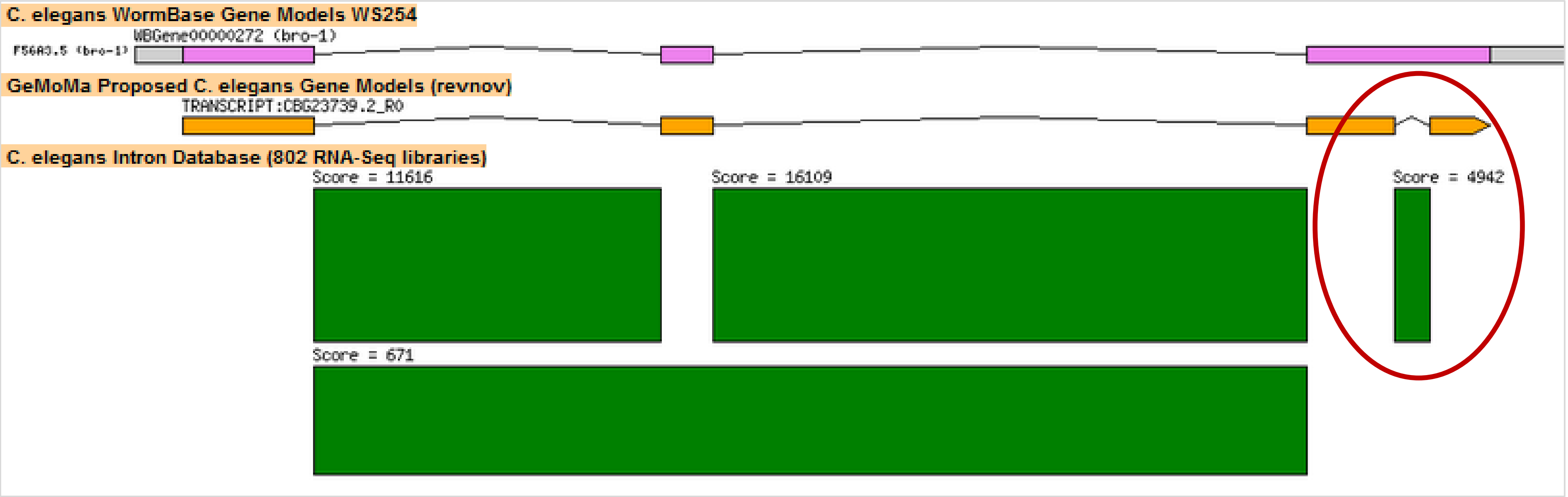
Candidate *C. elegans* protein-coding transcripts (1,698 transcripts)

used *C. elegans* RNA-Seq introns from 802 libraries[2]

Candidate *C. elegans* protein-coding transcripts with RNA-Seq support **(279 improvable transcripts and 2 novel genes)**

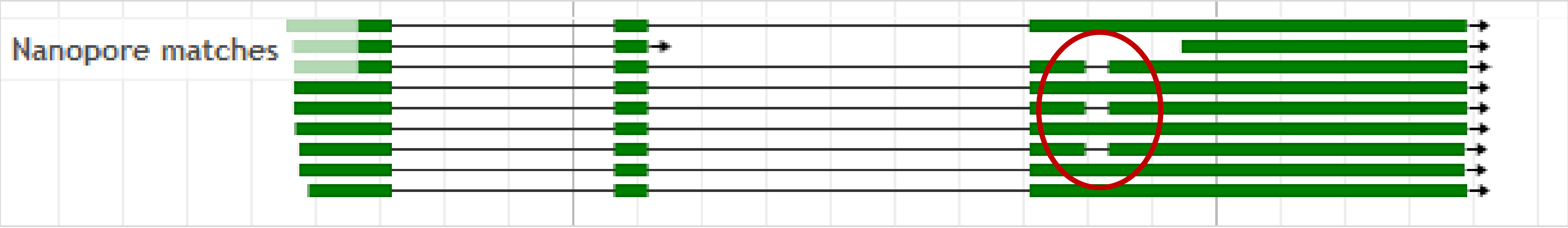
[1] Keilwagen et al., 2018, [2] Douglas, M., 2018

Identified 279 *C. elegans* transcripts supported by RNA-Seq introns


C. elegans WormBase Gene Models WS254
WBGene00000272 (bro-1)
F56A3.5 (bro-1)
GeMoMa Proposed C. elegans Gene Models (revnov)
TRANSCRIPT:CBG23739.2_R0
C. elegans Intron Database (802 RNA-Seq libraries)
Score = 11616
Score = 16109
Score = 4942
Score = 671

C. elegans intron database credit: Matt Douglas

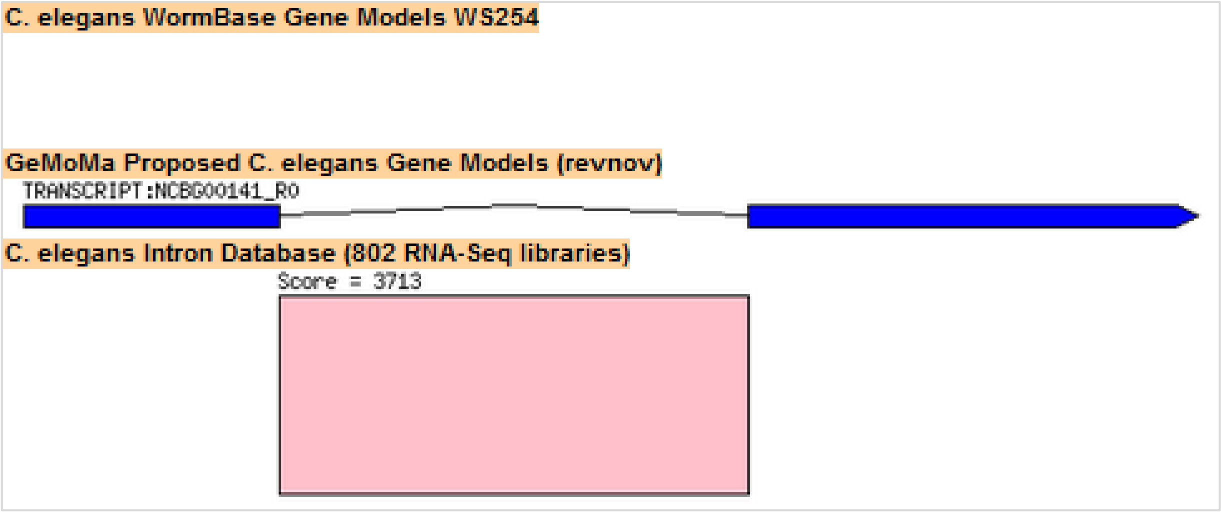
- Random sampling suggested that some are also supported by long-read alignments

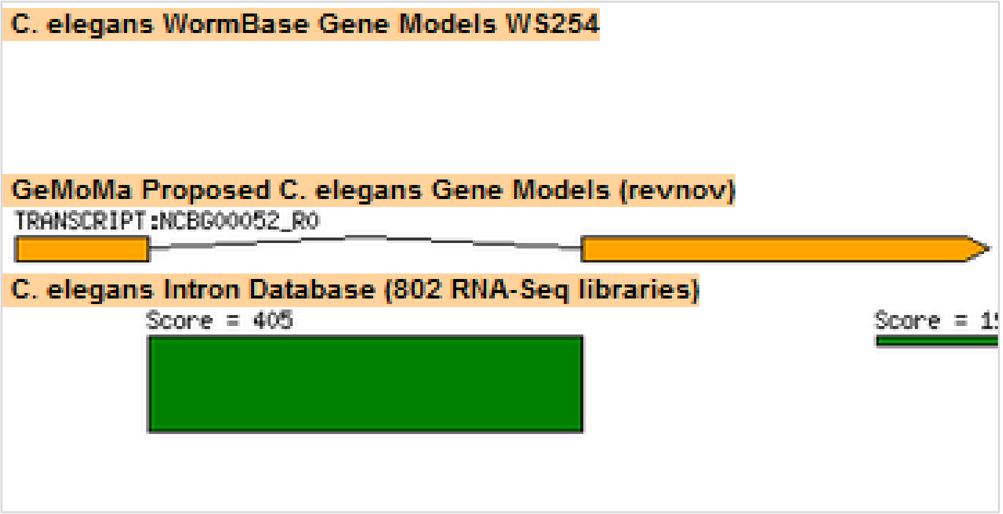


Long-read alignments credit: WormBase JBrowse (as of Dec 20th, 2019)

Pictured: *bro-1* (BROther (Drosophila tx factor partner) homolog), predicted to contribute to sequence-specific DNA binding activity, ortholog of human CBFB (core-binding factor subunit beta). Human ortholog of this gene are implicated in acute myeloid leukemia.

Identified 2 *C. elegans* novel genes supported by RNA-Seq introns


C. elegans WormBase Gene Models WS254
GeMoMa Proposed C. elegans Gene Models (revnov)
TRANSCRIPT:NCBG00141_R0
C. elegans Intron Database (802 RNA-Seq libraries)
Score = 3713


C. elegans WormBase Gene Models WS254
GeMoMa Proposed C. elegans Gene Models (revnov)
TRANSCRIPT:NCBG00052_R0
C. elegans Intron Database (802 RNA-Seq libraries)
Score = 405
Score = 1

- Previously have not been annotated
- No overlapping genes in the genomic regions
- Have more than 400 reads supporting the introns

Comparative analyses summary

- **Homology and RNA-Seq based comparative analysis using the improved *C. briggsae* annotation revealed more ortholog relationships between *C. briggsae* and *C. elegans,* and improved *C. elegans* annotation**
    - Revealed 132 new ortholog pairs, 32 belong to *C. briggsae* novel transcripts
    - Revealed 279 transcripts and 2 novel *C. elegans* genes

Conclusions

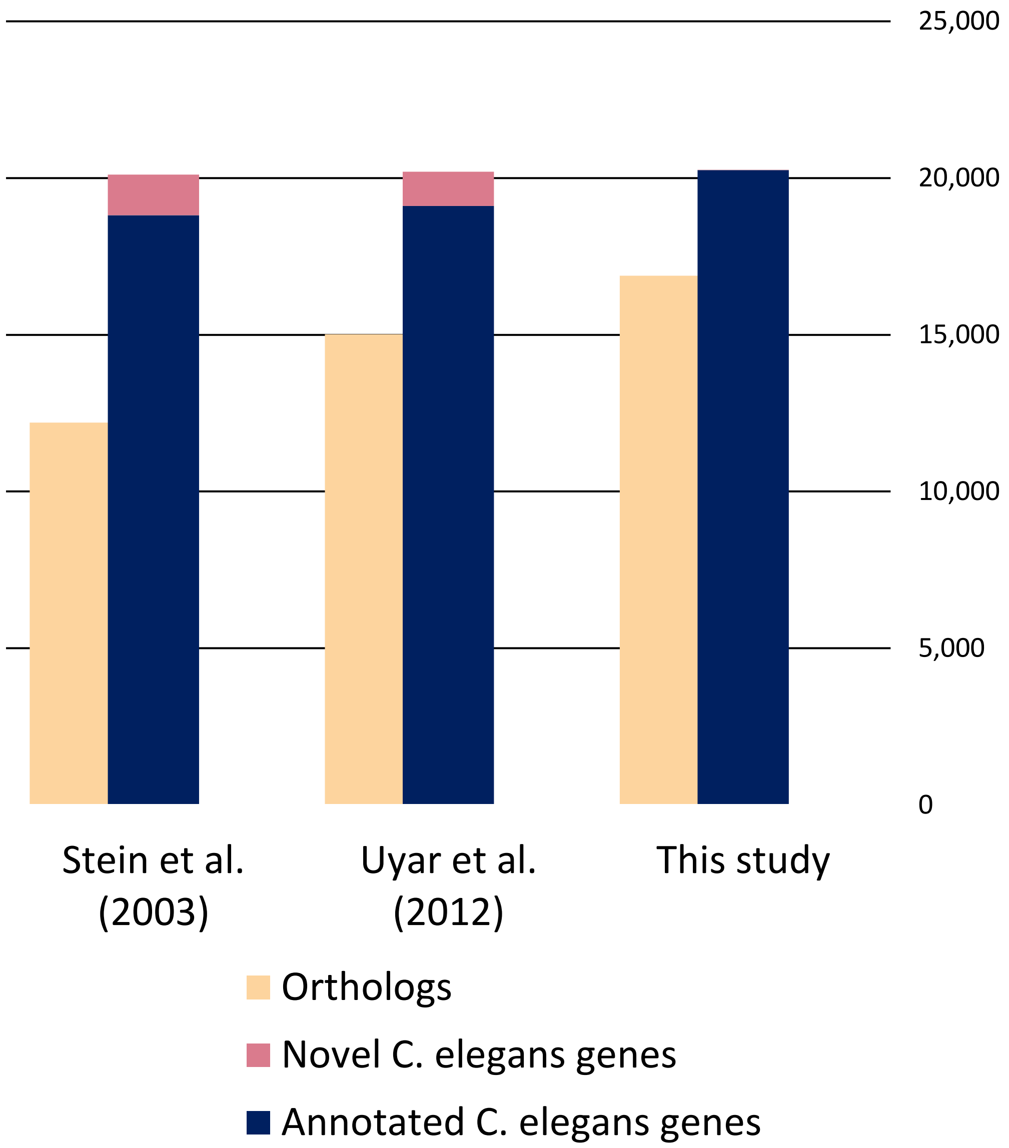
- RNA-Seq provides evidence to improve *C. briggsae* annotation
  - Reveals thousands of novel introns and exons, as well as hundreds of novel protein-coding transcripts

- The improved *C. briggsae* annotation together with comparative analyses reveals novel *C. briggsae*–*C. elegans* ortholog relationships and novel *C. elegans* protein-coding transcripts

- Despite limited data available for *C. briggsae*, the improved annotation has enhanced the utility of *C. briggsae* as a comparative platform for *C. elegans*.

Significance

- As more RNA-Seq data becomes available, this method can be used to further refine not only *C. briggsae* annotation but also *C. elegans* annotation.



Acknowledgements

Senior Supervisor

Dr. Jack Chen

Committee Members & Examining Committee

Dr. Fiona Brinkman

Dr. Ryan Morin

Dr. Christopher Beh

Dr. Mark Paetzel

Chen Lab Members (past&present)

Dr. Jiarui Li

Dr. Michelle Hu

Matthew Douglas

Marija Jovanovic

Kate Gibson

Family and friends

Thio, Adair, Uplands family

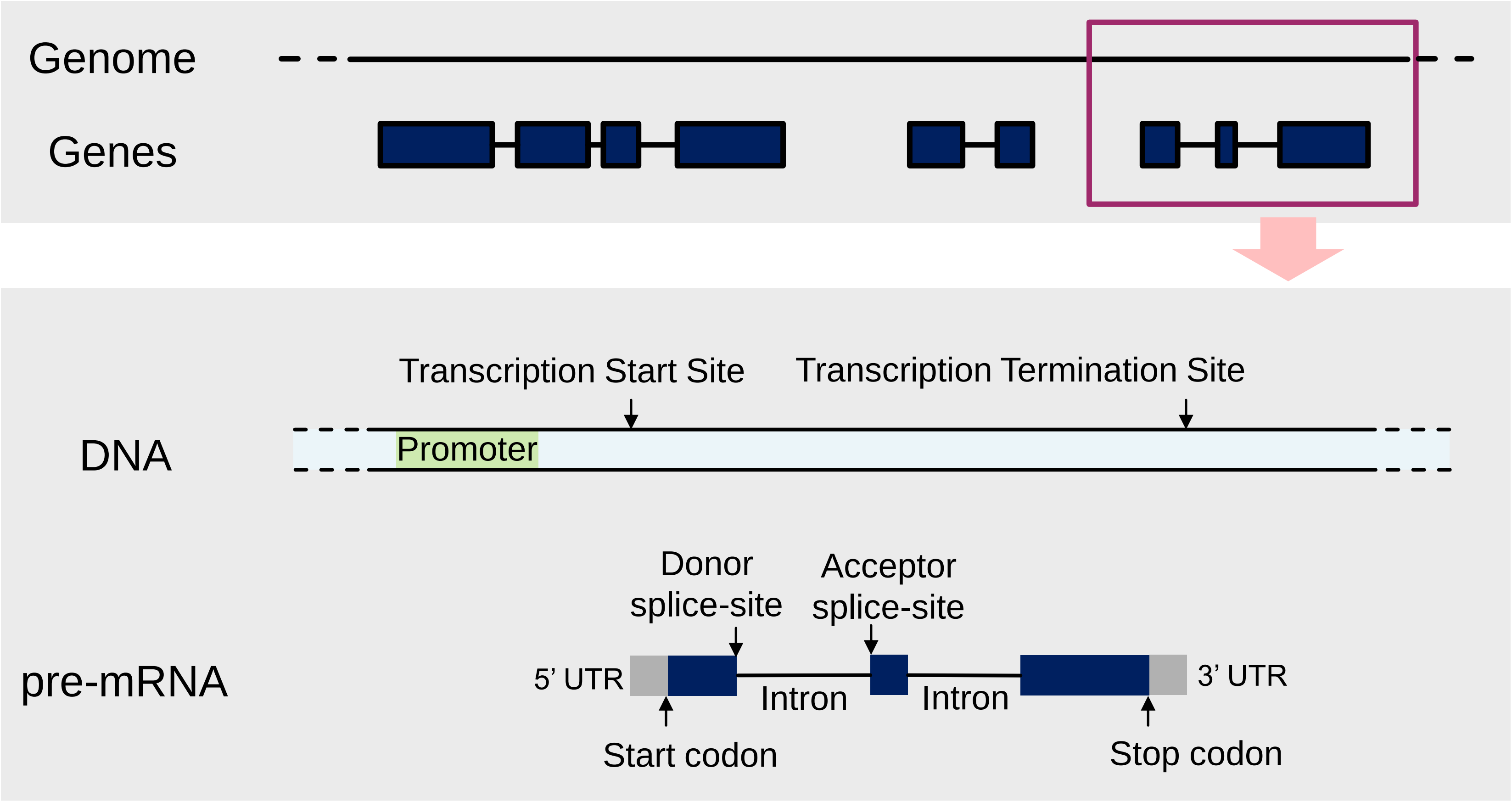
MBB, SFU Omics, Indo friends

Extras

What is genome annotation?

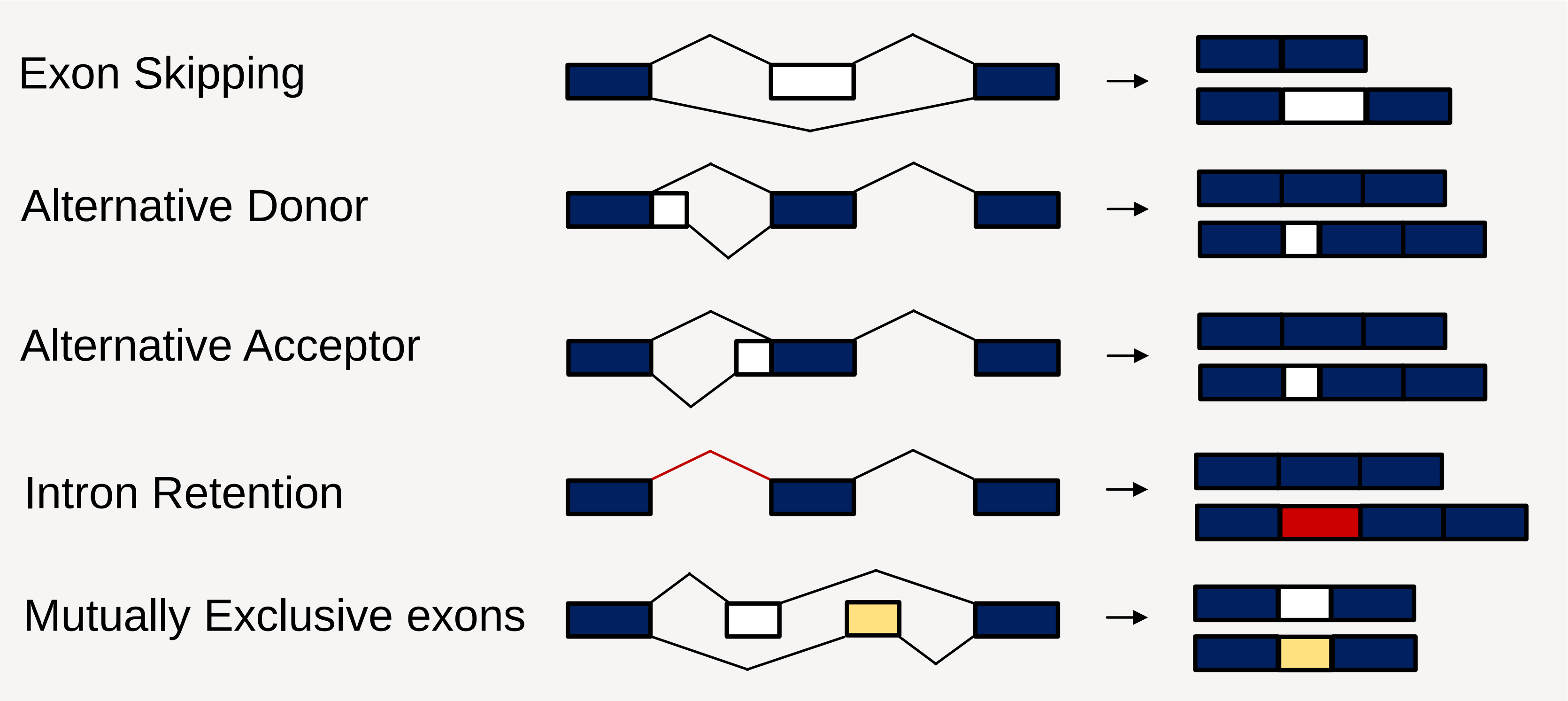
The process of finding location and structure of all genes in the genome

- Introns
- Exons
- Splice sites
- Coding regions
- Promoter sequences
- TF binding sites
- and many more..

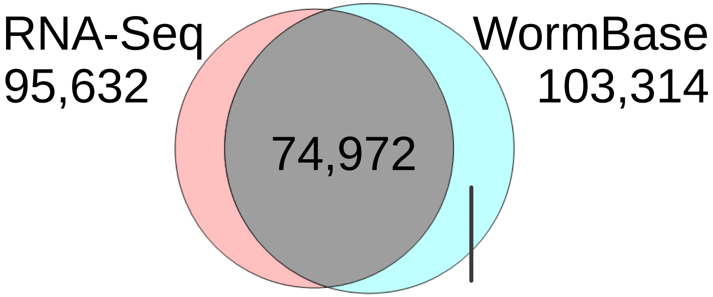


Challenge in genome annotation: alternative splicing

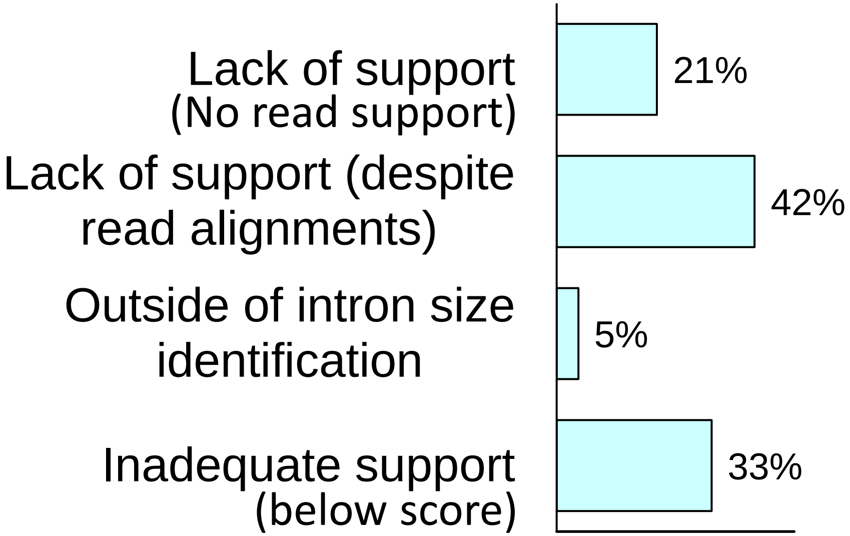
- Genes can have alternative splicing pathways to process pre-mRNAs into two or more mature transcripts that encode different proteins.
- This can contribute to the completeness of the annotation.

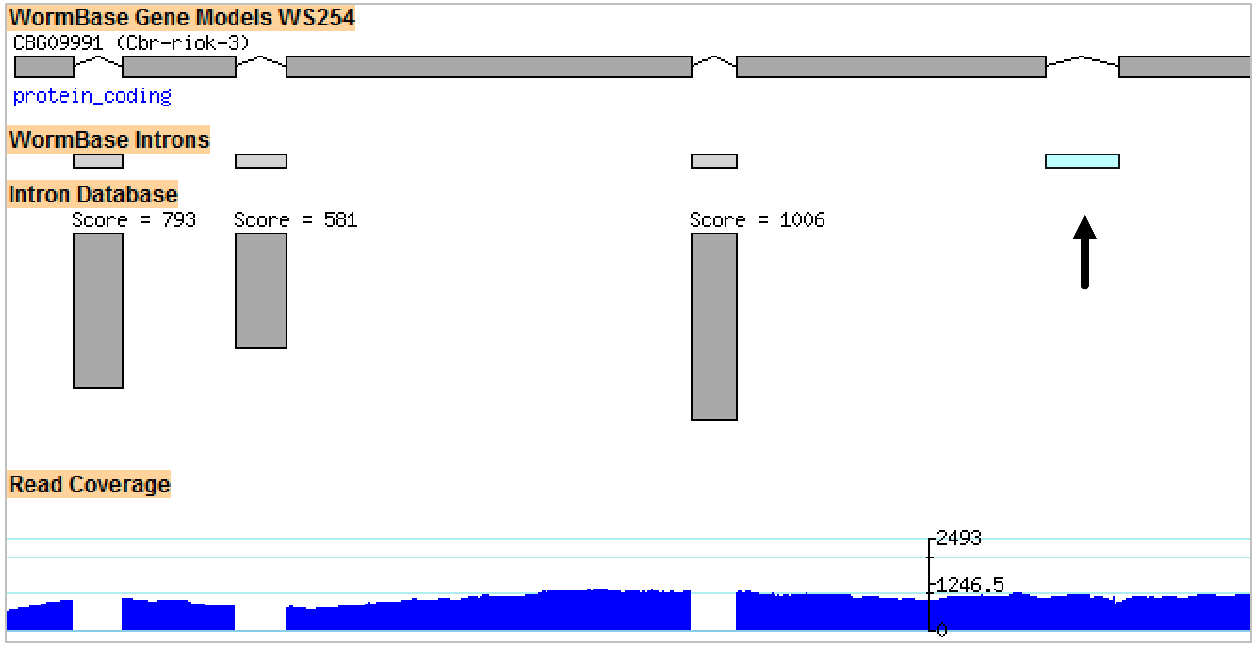


WormBase introns not supported by RNA-Seq

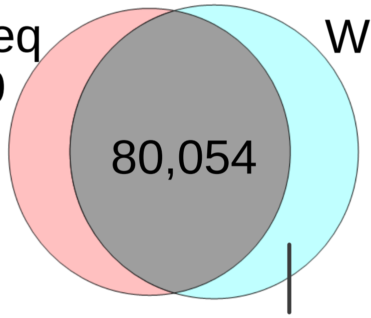


<u>WormBase specific</u>
28,341 (27%)





WormBase exons not supported by RNA-Seq

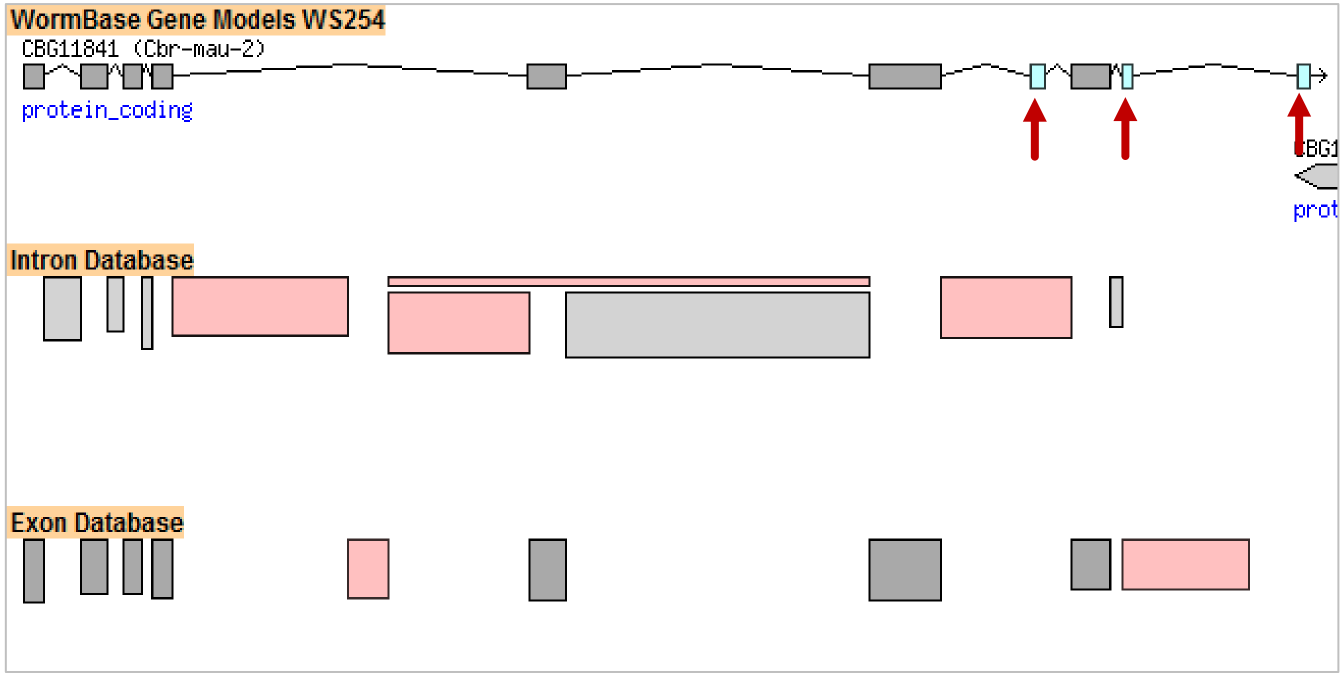


RNA-Seq
115,689

WormBase
121,355

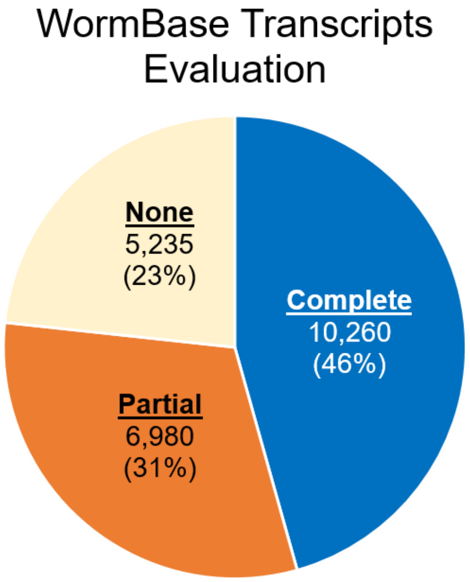
80,054

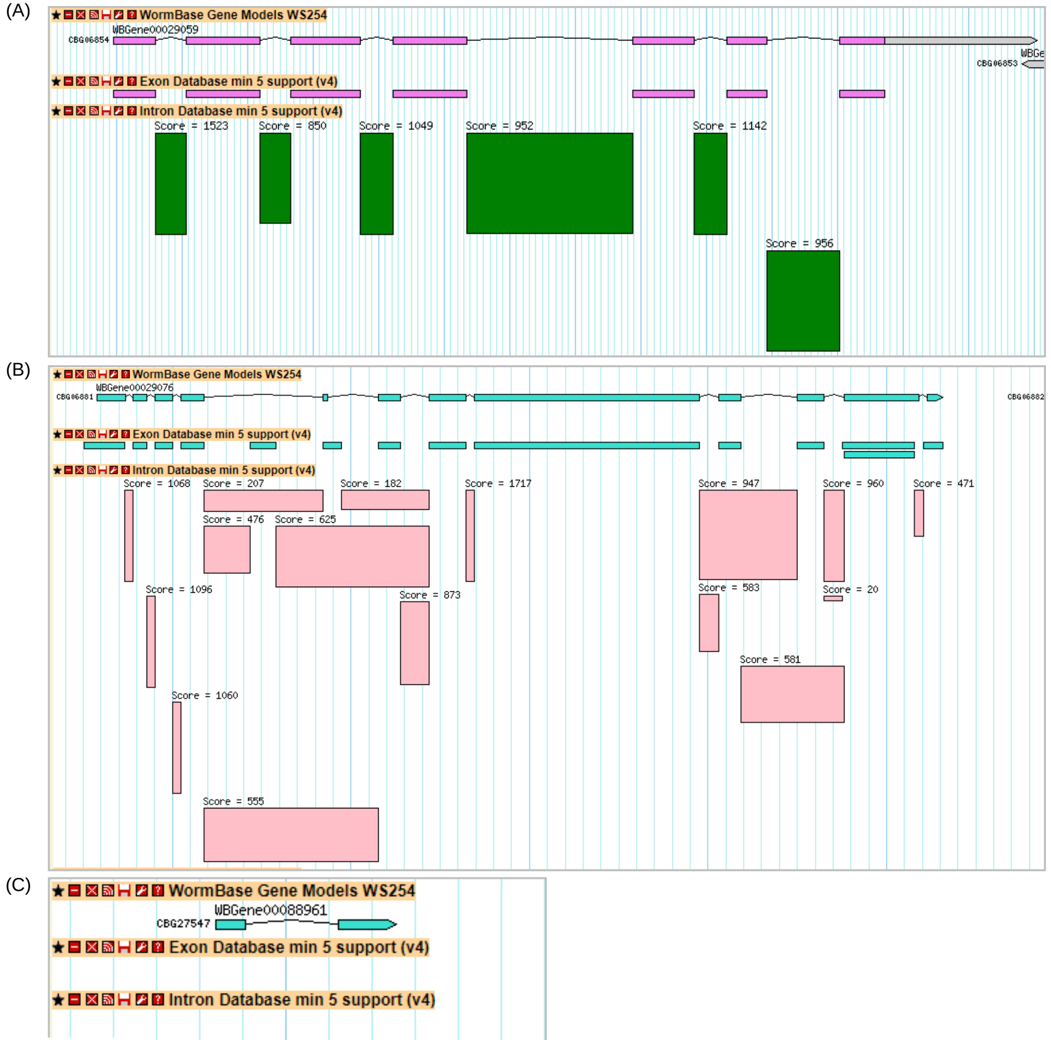
WormBase specific
41,301 (34%)

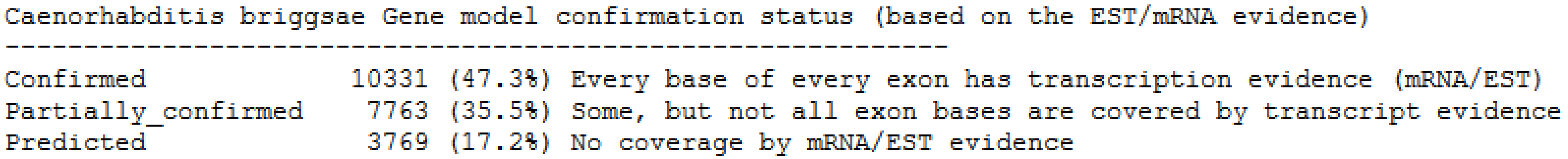


WormBase transcripts evaluation

- Less than 50% of *C. briggsae* transcript models are validated by our intron and exon databases
- 31% are partially validated, which maybe due to mispredicted genes and lowly expressed transcripts
- 23% of transcripts are not validated at all.

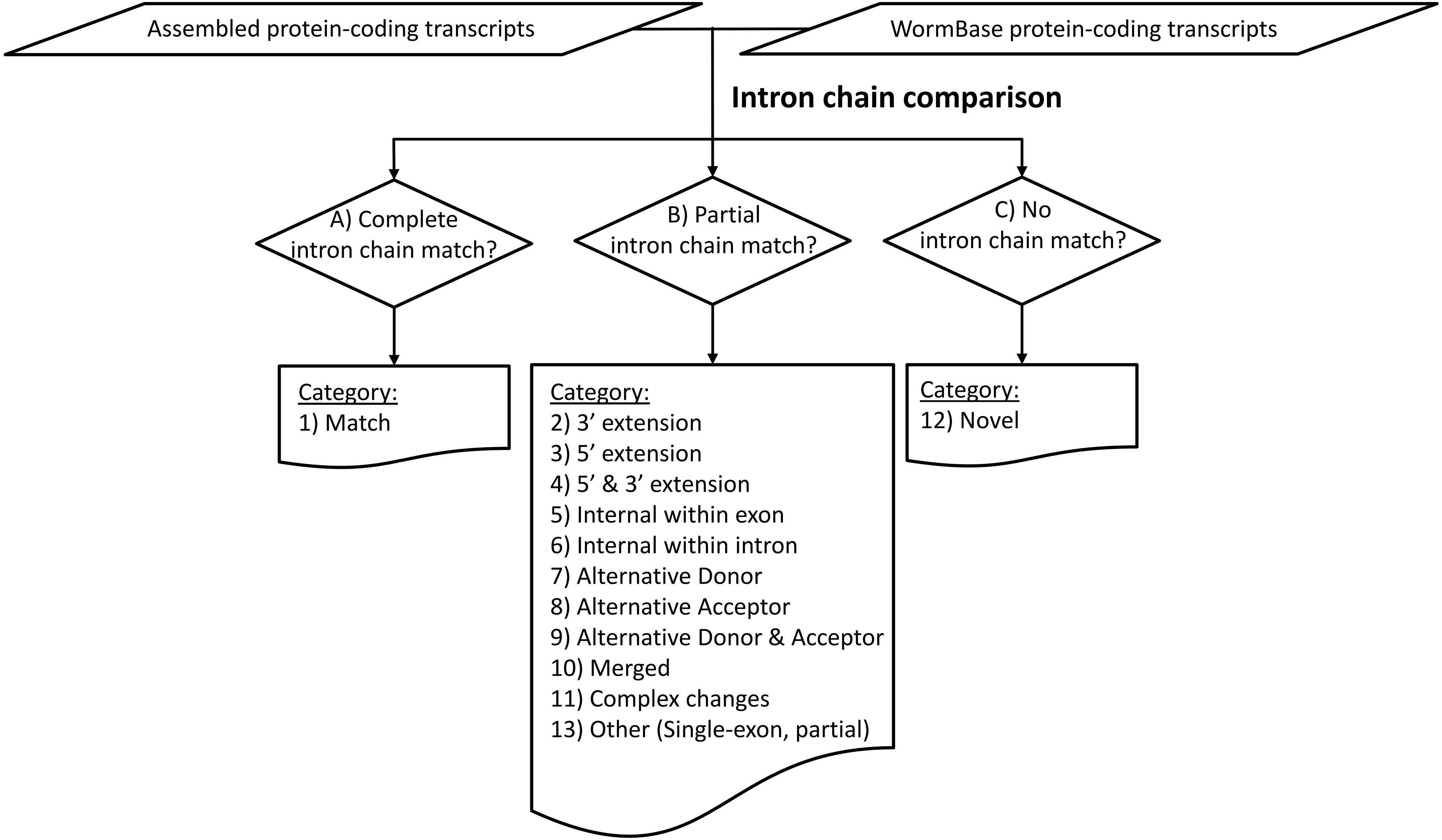


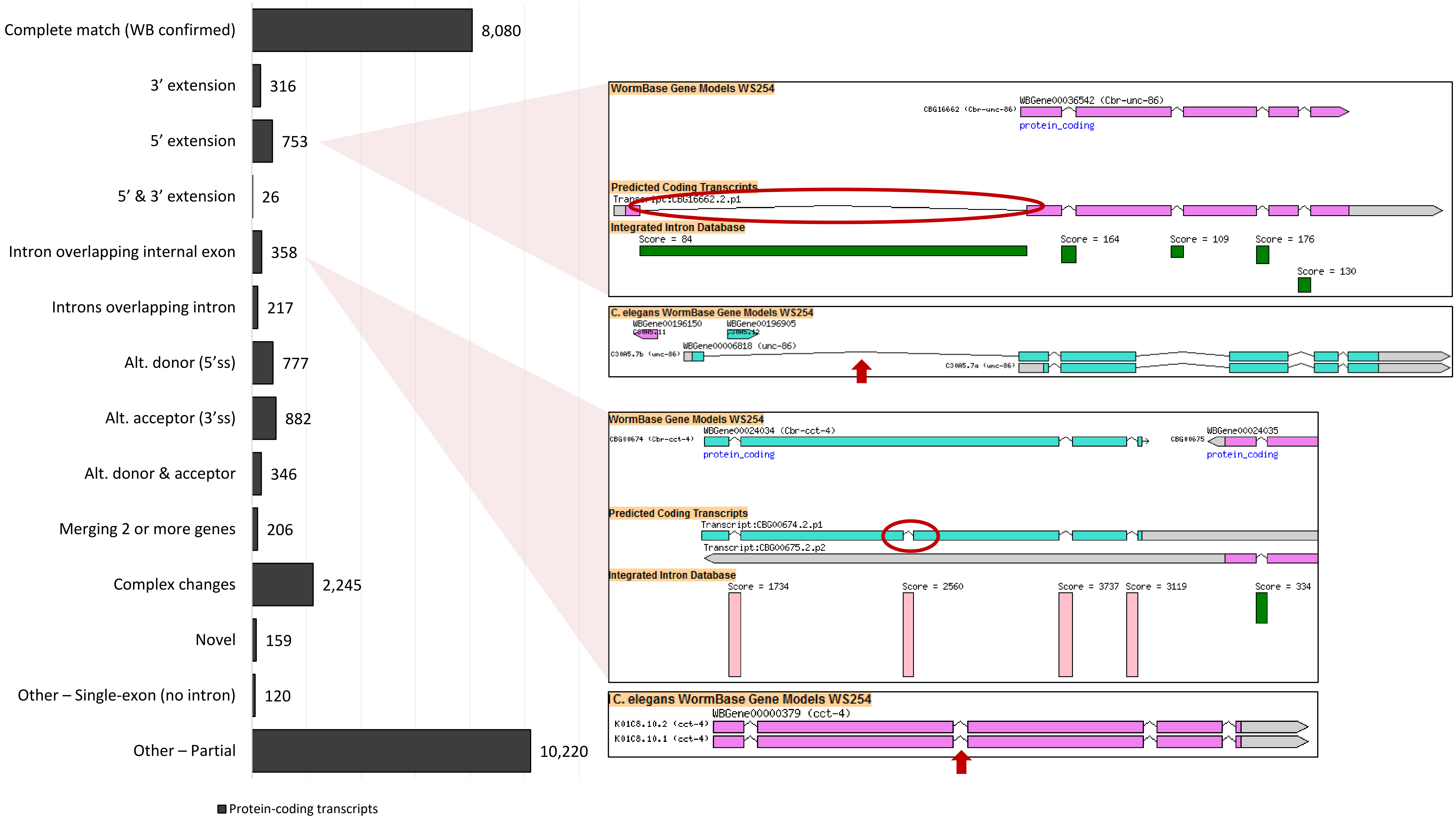




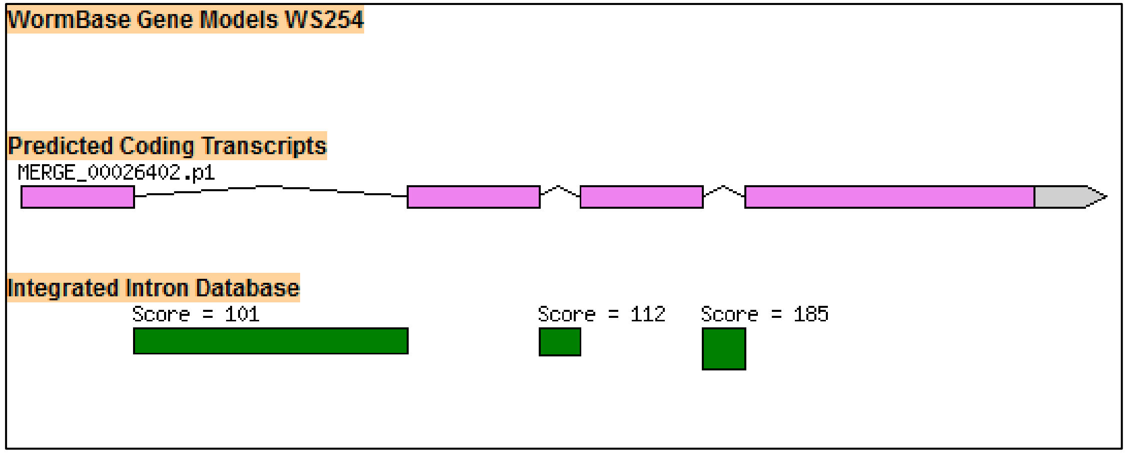
```
Caenorhabditis briggsae Gene model confirmation status (based on the EST/mRNA evidence)
-------------------------------------------------------------
Confirmed             10331 (47.3%) Every base of every exon has transcription evidence (mRNA/EST)
Partially_confirmed    7763 (35.5%) Some, but not all exon bases are covered by transcript evidence
Predicted              3769 (17.2%) No coverage by mRNA/EST evidence
```

Method: transcript-to-transcript comparison (intron chain)


Assembled protein-coding transcripts
WormBase protein-coding transcripts
Intron chain comparison
A) Complete intron chain match?
B) Partial intron chain match?
C) No intron chain match?
Category:
1) Match
Category:
2) 3’ extension
3) 5’ extension
4) 5’ & 3’ extension
5) Internal within exon
6) Internal within intron
7) Alternative Donor
8) Alternative Acceptor
9) Alternative Donor & Acceptor
10) Merged
11) Complex changes
13) Other (Single-exon, partial)
Category:
12) Novel

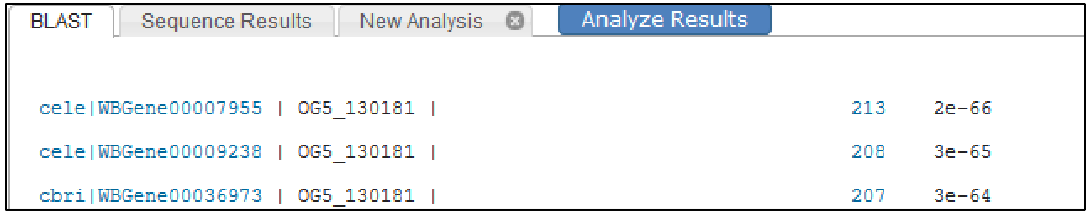

Complete match (WB confirmed)
8,080
3’ extension
316
5’ extension
753
5’ & 3’ extension
26
Intron overlapping internal exon
358
Introns overlapping intron
217
Alt. donor (5’ss)
777
Alt. acceptor (3’ss)
882
Alt. donor & acceptor
346
Merging 2 or more genes
206
Complex changes
2,245
Novel
159
Other – Single-exon (no intron)
120
Other – Partial
10,220
Protein-coding transcripts
WormBase Gene Models WS254
WBGene00036542 (Cbr-unc-86)
CBG16662 (Cbr-unc-86)
protein_coding
Predicted Coding Transcripts
Transcript:CBG16662.2.p1
Integrated Intron Database
Score = 84
Score = 164
Score = 109
Score = 176
Score = 130
C. elegans WormBase Gene Models WS254
WBGene00196150
C30A5.11
WBGene00196905
C30A5.12
WBGene00006818 (unc-86)
C30A5.7b (unc-86)
C30A5.7a (unc-86)
WormBase Gene Models WS254
WBGene00024034 (Cbr-cct-4)
CBG00674 (Cbr-cct-4)
protein_coding
WBGene00024035
CBG00675
protein_coding
Predicted Coding Transcripts
Transcript:CBG00674.2.p1
Transcript:CBG00675.2.p2
Integrated Intron Database
Score = 1734
Score = 2560
Score = 3737
Score = 3119
Score = 334
C. elegans WormBase Gene Models WS254
WBGene00000379 (cct-4)
K01C8.10.2 (cct-4)
K01C8.10.1 (cct-4)

RNA-Seq suggests 159 novel transcript/gene

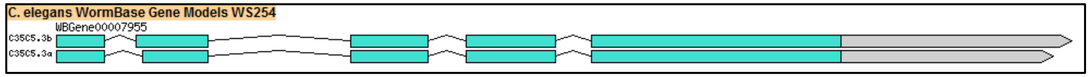


WormBase Gene Models WS254
Predicted Coding Transcripts
MERGE_00026402.p1
Integrated Intron Database
Score = 101
Score = 112
Score = 185

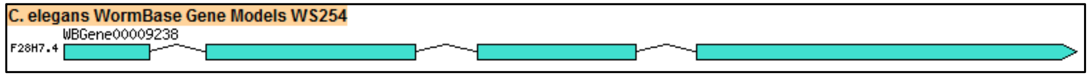
V:10,295,768..10,297,767



```
cele|WBGene00007955 | OG5_130181 |        213   2e-66
cele|WBGene00009238 | OG5_130181 |        208   3e-65
cbri|WBGene00036973 | OG5_130181 |        207   3e-64
```



X:11,545,054..11,546,597



V:10,746,307..10,747,406

- Transcript MERGE_00026402 (nCBG00109) is previously not annotated by WormBase
- Using RNA-Seq data, we found introns and assembled a transcript suggesting putative novel gene in this genomic region
- First and second hits from BLAST result shown that the protein sequence is the most similar to *C. elegans* proteins (right)

https://orthomcl.org/orthomcl/showApplication.do

34% of candidate transcripts do not start with ATG

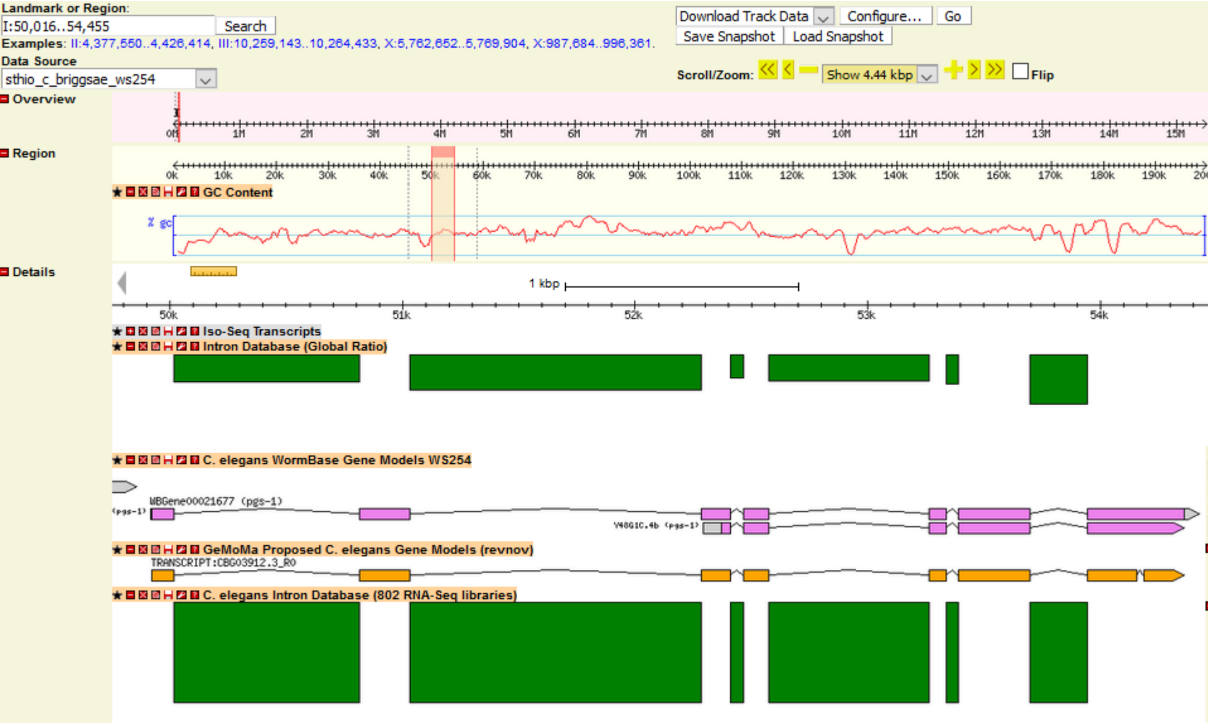
- A functional protein-coding transcript should contain an Open Reading Frame (ORF) that begins with start codon and ends with stop codon[1].
- Limitation of TransDecoder: does not have a start-codon finding function and will include transcript from the beginning if there is no upstream in-frame stop codon at the beginning of the transcript[2,3].

Category		Protein-coding transcripts	Starts with ATG	Does not start with ATG
2.	3’ extension	316	216	100
3.	5’ extension	753	528	225
4.	5’ & 3’ extension	26	15	11
5.	Intron overlapping internal exon	358	246	112
6.	Introns overlapping intron	217	144	73
7.	Alternative donor (5’ss)	777	515	262
8.	Alternative acceptor (3’ss)	882	559	323
9.	Alternative donor & acceptor	346	218	128
10.	Merged	206	119	87
11.	Complex changes	2,245	1496	749
12.	Novel	159	78	81
	Total	6,285	4,134 (66%)	**2,151 (34%)**

[1] Majoros et al., 2014, [2,3] Haas, 2014, 2018

Ortholog pairs breakdown

- 16,880:
  - 1894: modified, same cel, revised cbriggsae transcript (extension etc)
  - 100: new, cel exist in original, contains 4 nCBG new cbriggsae transcripts
  - 32: new from nCBG (4 redundant with one point above)
  - 80: need further analysis, maybe new new (so total new would be plus nCBG)

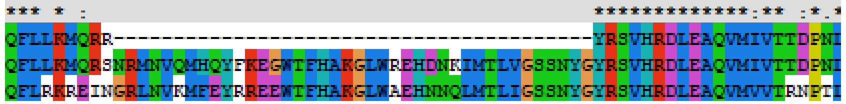


Landmark or Region: I:50,016..54,455 Search

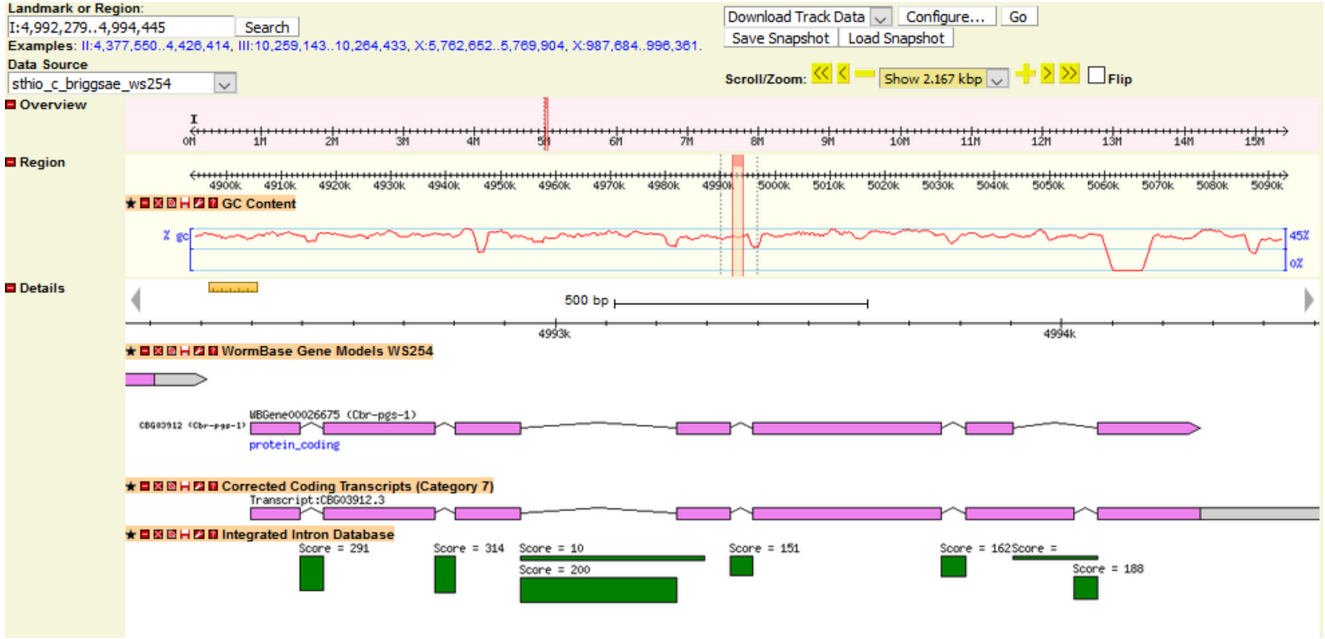
Examples: II:4,377,550..4,426,414, III:10,259,143..10,264,433, X:5,762,652..5,769,904, X:987,684..996,361.

Data Source: sthio_c_briggsae_ws254

Scroll/Zoom: Show 4.44 kbp □ Flip



```
*** *  :                                            *************:**  :*.*
QFLLKMQRR-----------------------------------------YRSVHRDLEAQVMIVTTDPNI
QFLLKMQRSNRMNVQMHQYFKEGWTFHAKGLWREHDNKIMTLVGSSNYGYRSVHRDLEAQVMIVTTDPNI
QFLRKREINGRLNVKMFEYRREEWTFHAKGLWAEHNNQLMTLIGSSNYGYRSVHRDLEAQVMVVTRNPTI
```



Landmark or Region: I:4,992,279..4,994,445 Search

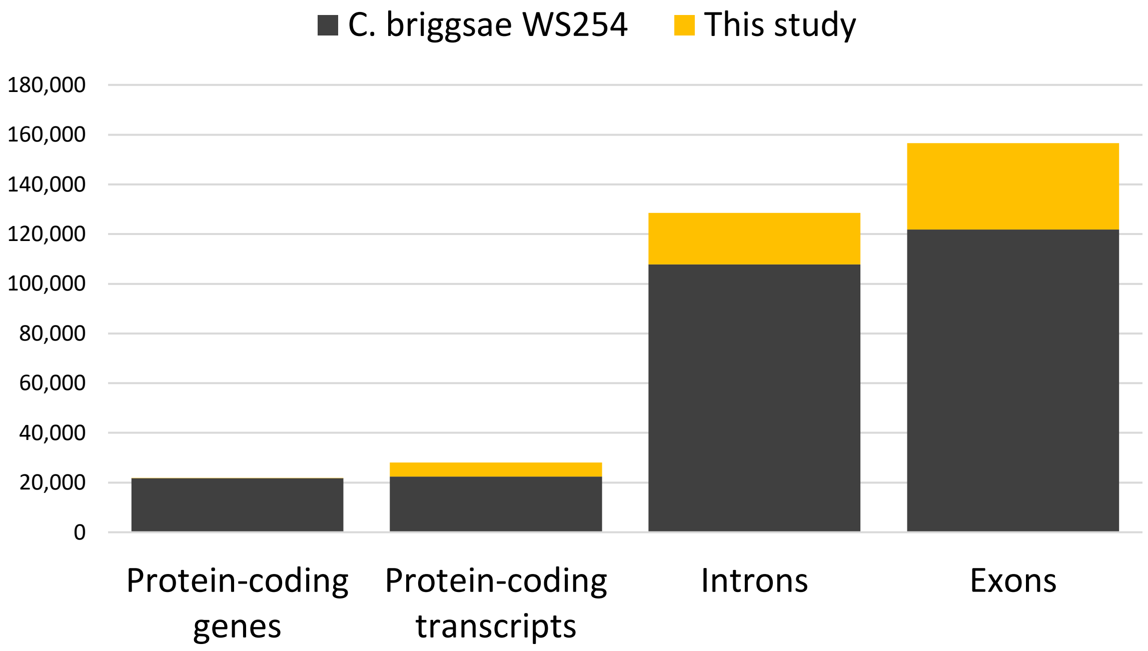
Examples: II:4,377,550..4,426,414, III:10,259,143..10,264,433, X:5,762,652..5,769,904, X:987,684..996,361.

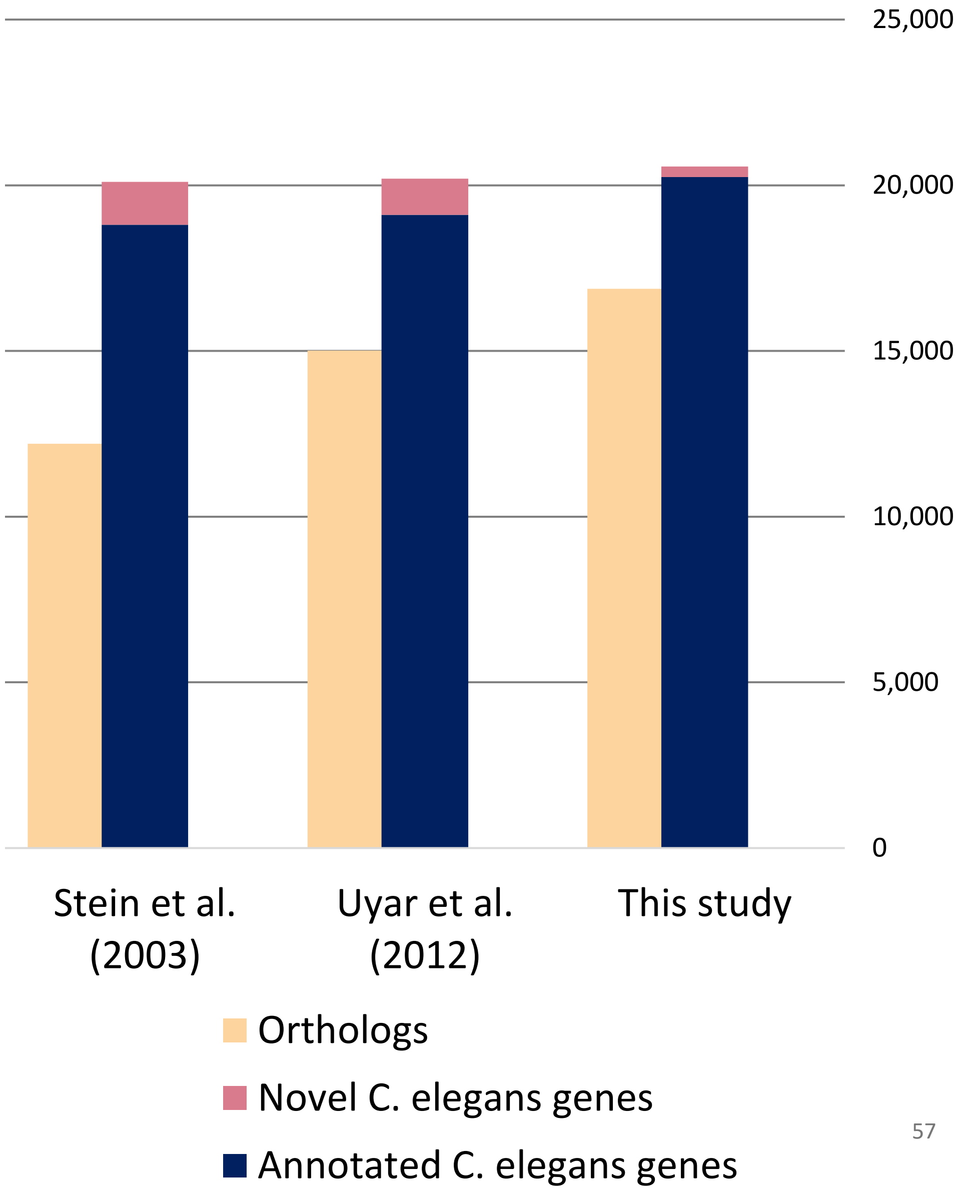
Data Source: sthio_c_briggsae_ws254

Scroll/Zoom: Show 2.167 kbp □ Flip

Significance

- As more RNA-Seq data becomes available, this method can be used to further refine not only *C. briggsae* annotation but also *C. elegans* annotation.


■ C. briggsae WS254
■ This study
Protein-coding genes
Protein-coding transcripts
Introns
Exons


Stein et al. (2003)
Uyar et al. (2012)
This study
■ Orthologs
■ Novel C. elegans genes
■ Annotated C. elegans genes