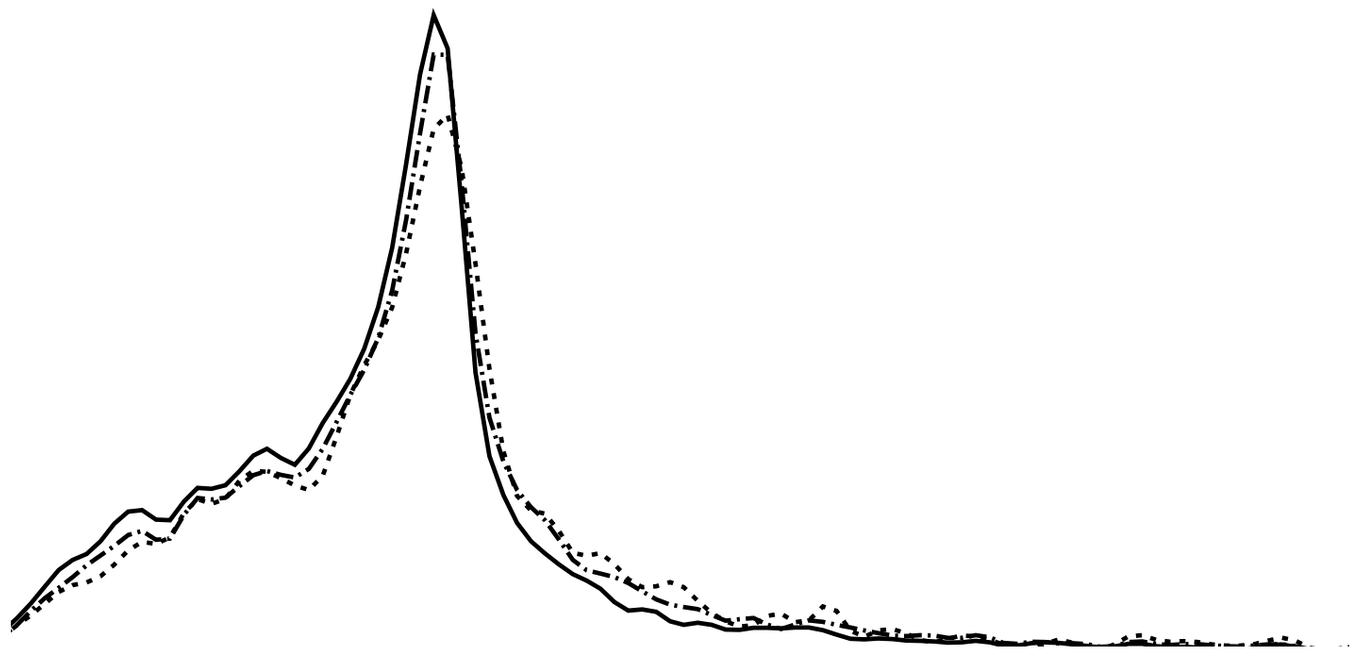
Assessing De-Novo Transcriptome Assemblies

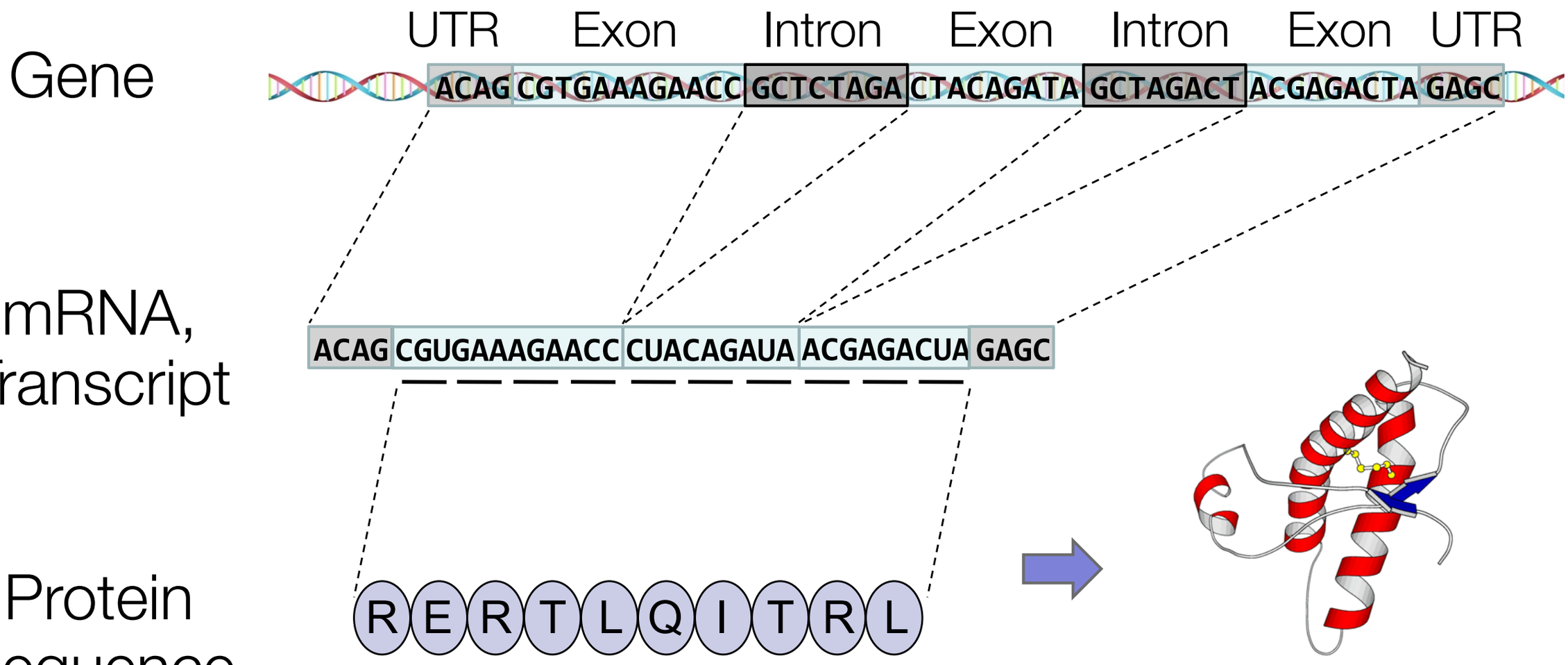
Shawn T. O'Neil
Center for Genome Research and Biocomputing
Oregon State University

Scott J. Emrich
University of Notre Dame

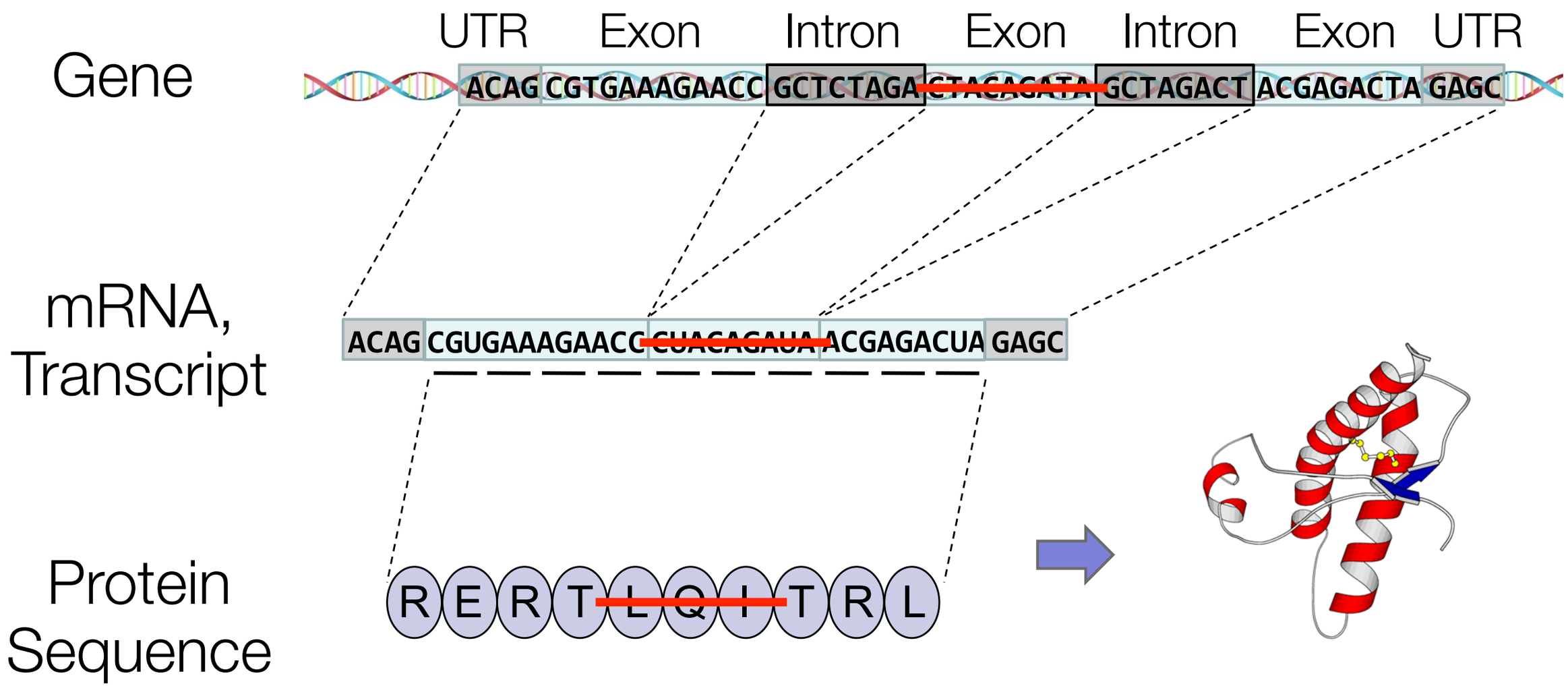
March 2, 2013



Transcripts

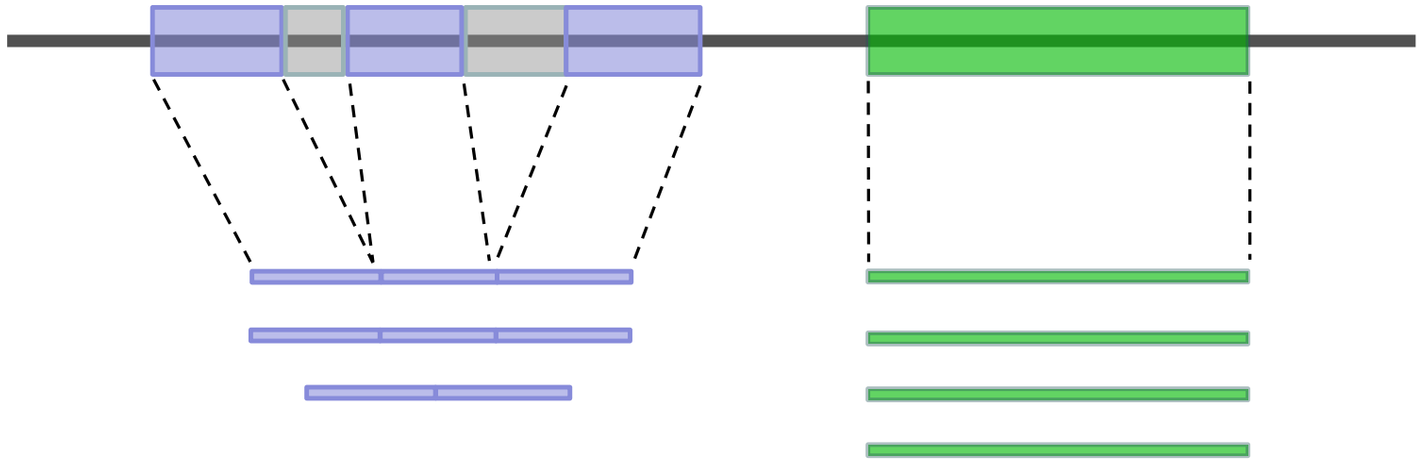


Transcripts



Transcriptome sequencing and assembly

Genome

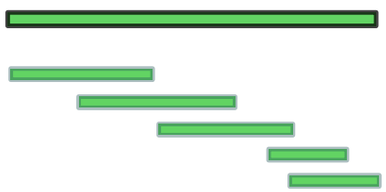


Transcripts
(mRNA in cell)

Extract ⇒ cDNA
Sequence Reads

Assemble

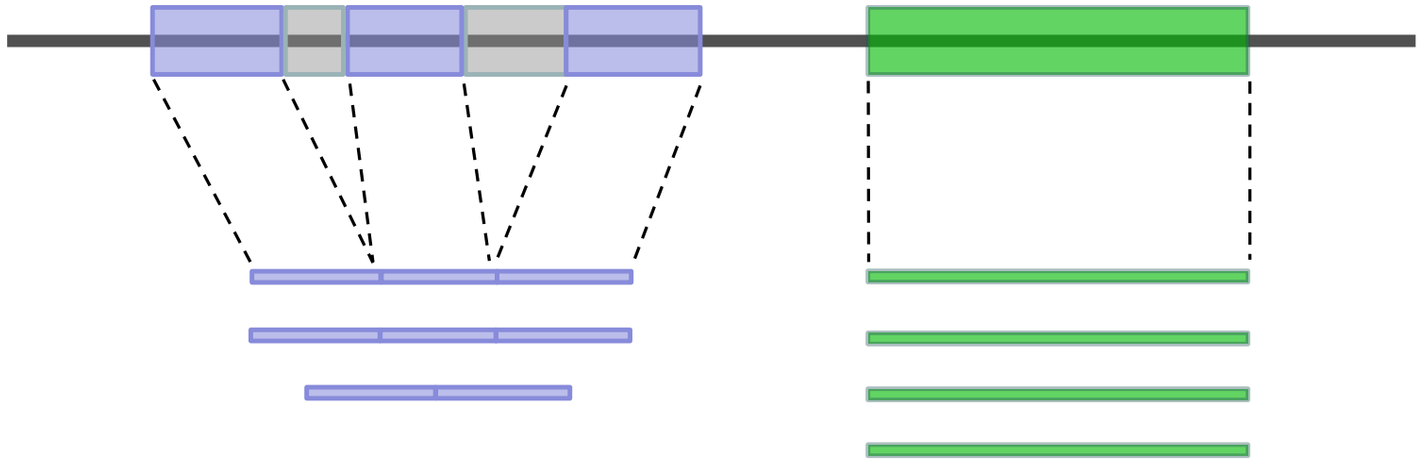
Contigs



Singletons

Transcriptome sequencing and assembly

Genome



Transcripts
(mRNA in cell)

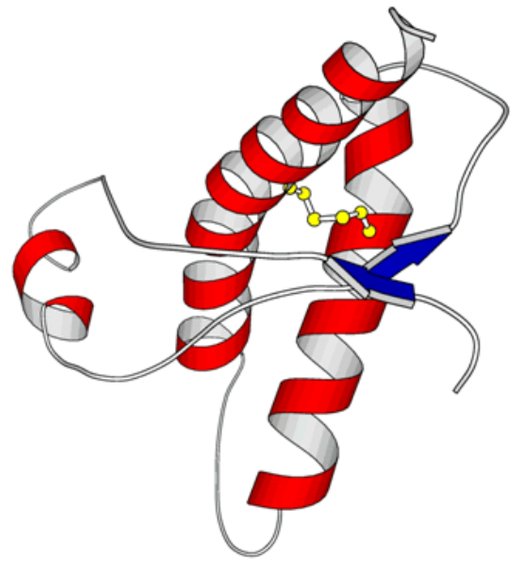
Extract ⇒ cDNA
Sequence Reads

Assemble

Why transcriptomes?

Transcripts isolate the useful:

- Functional considerations
- Expression analysis
- SNP analysis (e.g. Ka/Ks)



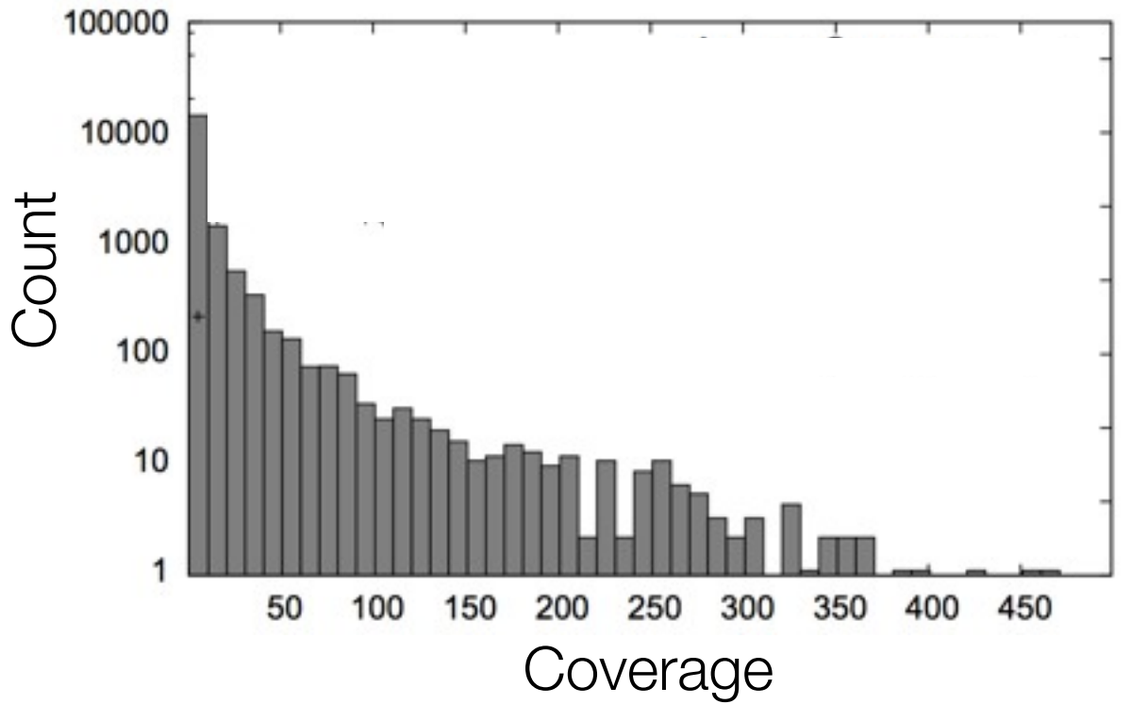
Whole-genome sequencing is (relatively) expensive.

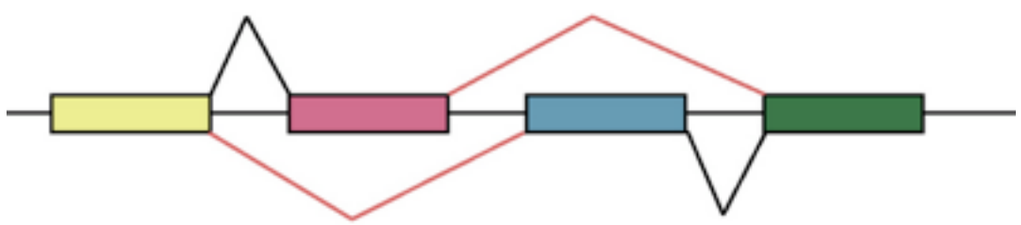
De-novo transcriptome assembly is accessible, particularly for non-model species.

Issues in transcriptome assembly

Unique challenges:

- Highly variant coverage
- Alternative splicing





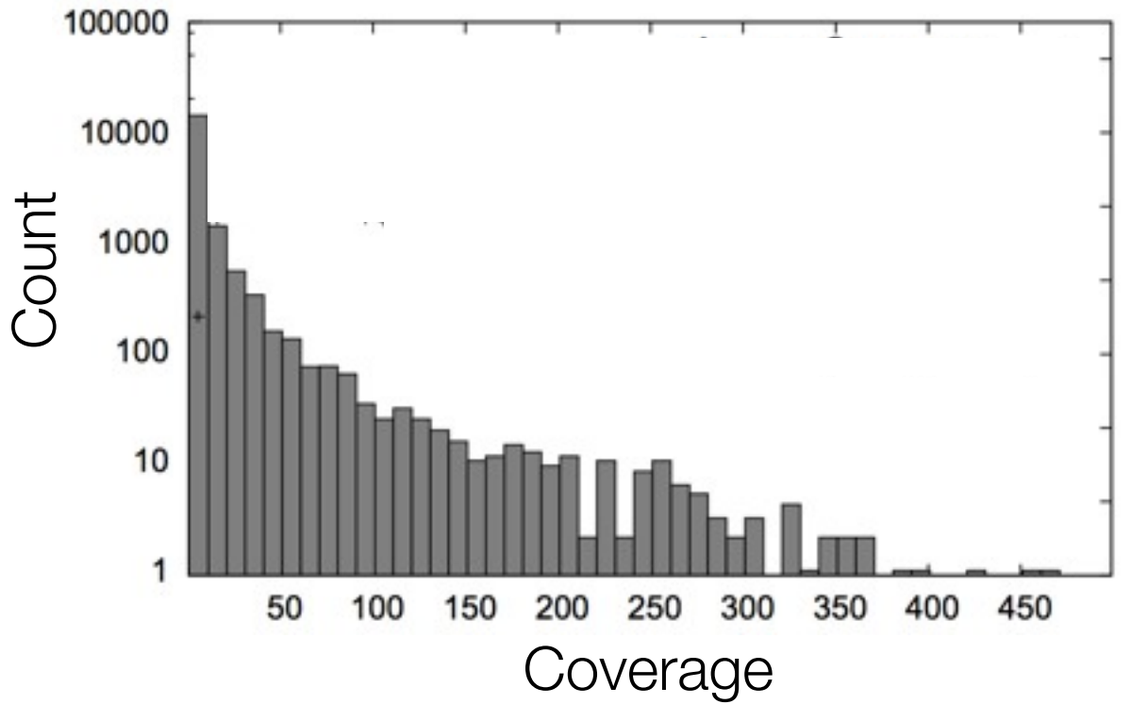
Issues in transcriptome assembly

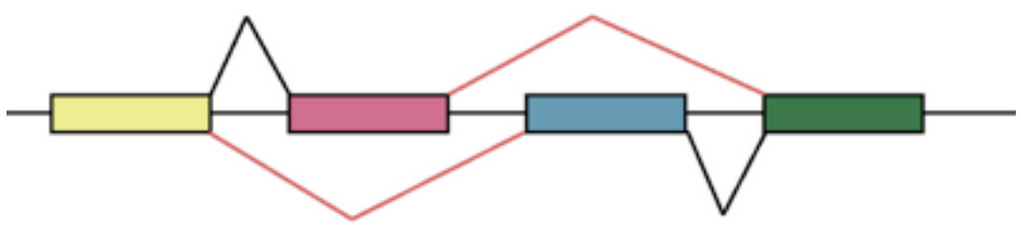
Unique challenges:

- Highly variant coverage
- Alternative splicing

A variety of new tools:

- Newbler (isoforms, 2011)
- Trinity (2011)
- SOAPdenovo-Trans (2011)
- Oases (2012)





Evaluating assemblies/assemblers

- Reuse of Genome Assembly Metrics
  - Contig Count (fewer is better?)
  - Contig N50 Length (longer is better?)
  - Total Assembly Size (more is better?)

Evaluating assemblies/assemblers

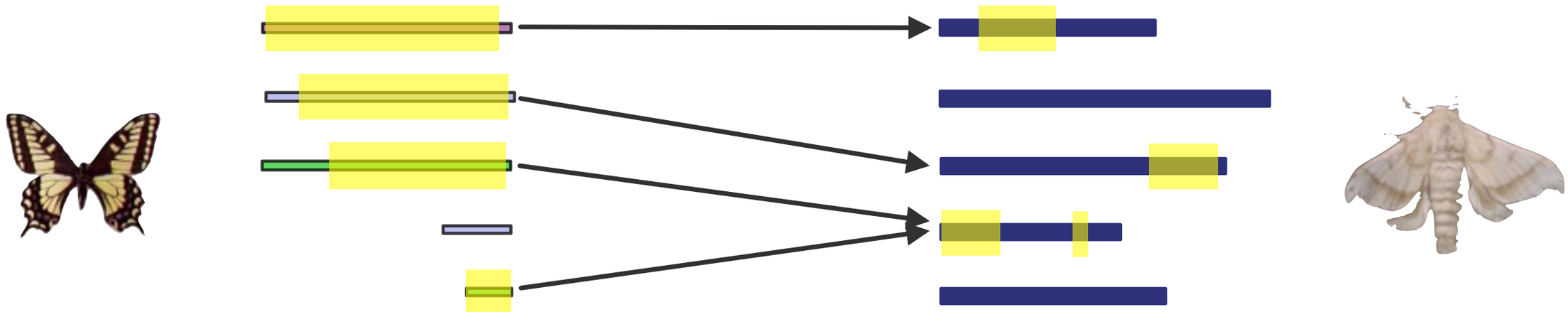
Reuse of Genome Assembly Metrics

- Contig Count (fewer is better?)
- Contig N50 Length (longer is better?)
- Total Assembly Size (more is better?)

Are these appropriate?

- Singletons are often ignored in computing metrics.
- Gene lengths are limited: D. mel. N50 size: 2,616bp
- Paralog collapse is difficult to identify

Annotation-based metrics: BLAST

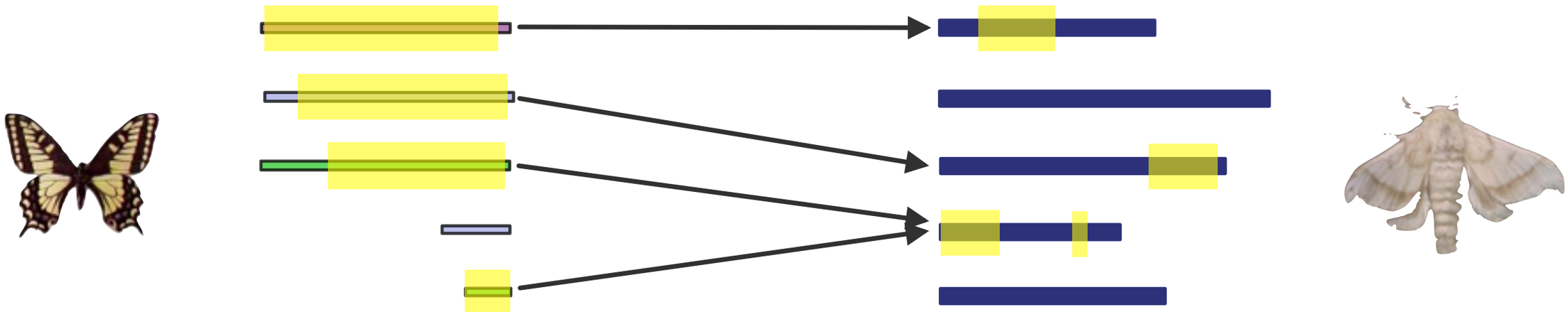


Assembled Transcripts

BLASTX Hits

Reference Protein Sequences.

Annotation-based metrics: BLAST



Assembled Transcripts

BLASTX Hits

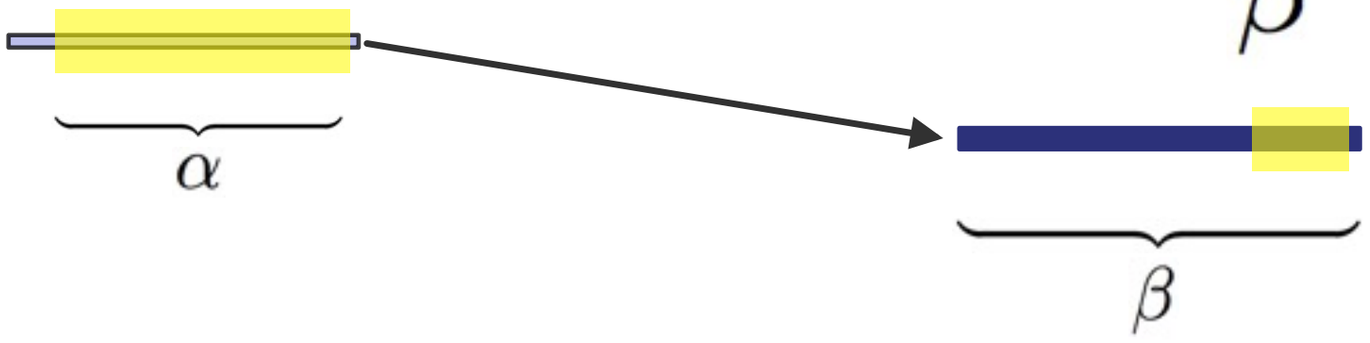
Reference Protein Sequences.

More metrics:

Number of unique hits, percentage of contigs with hits, properties of the hits, etc.

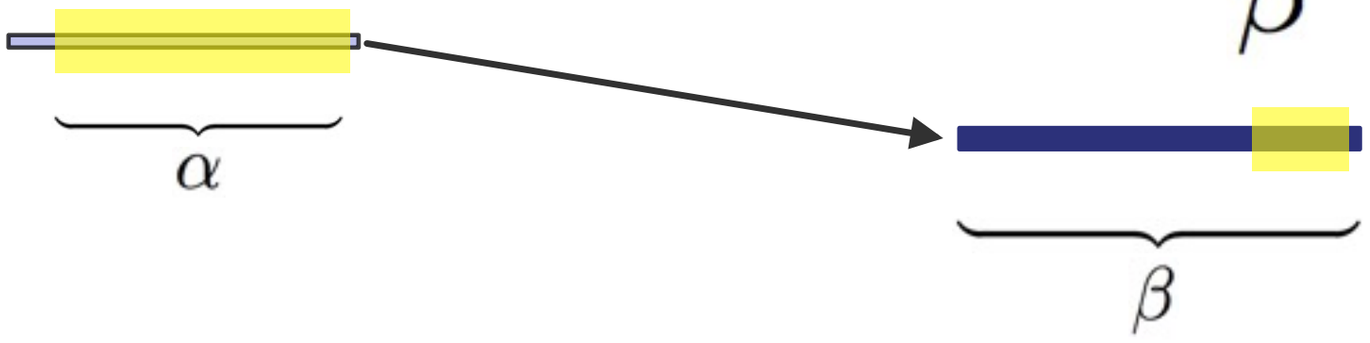
Ortholog Hit Ratio: a unigene completeness metric

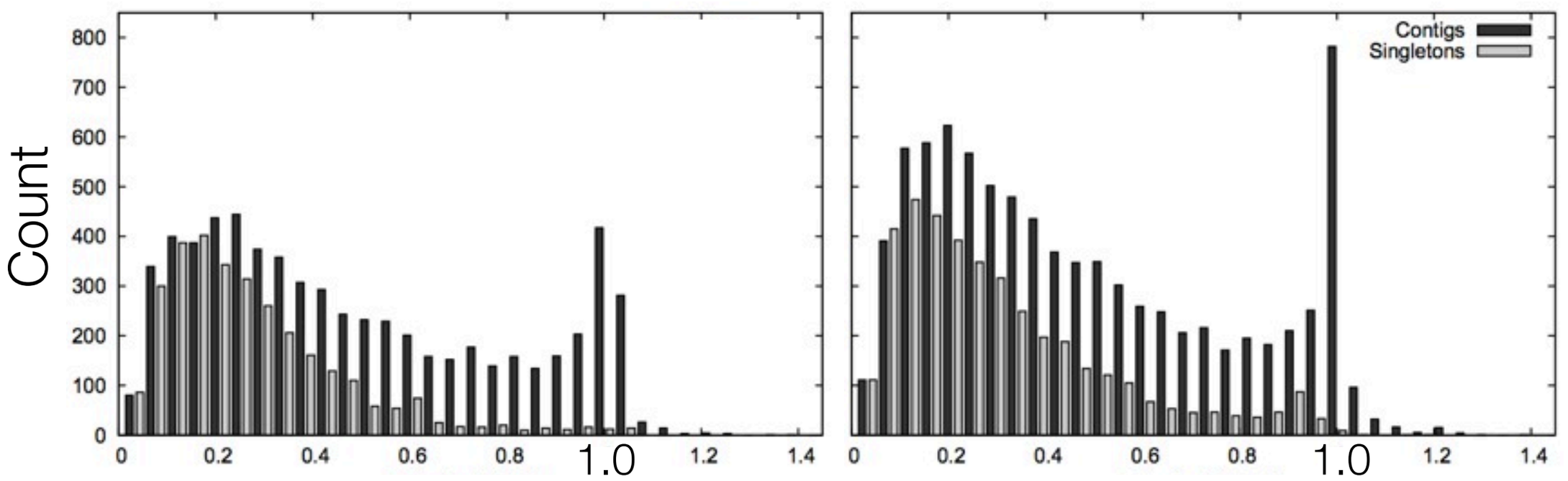
$$OHR = \frac{\alpha}{\beta}$$



Ortholog Hit Ratio: a unigene completeness metric

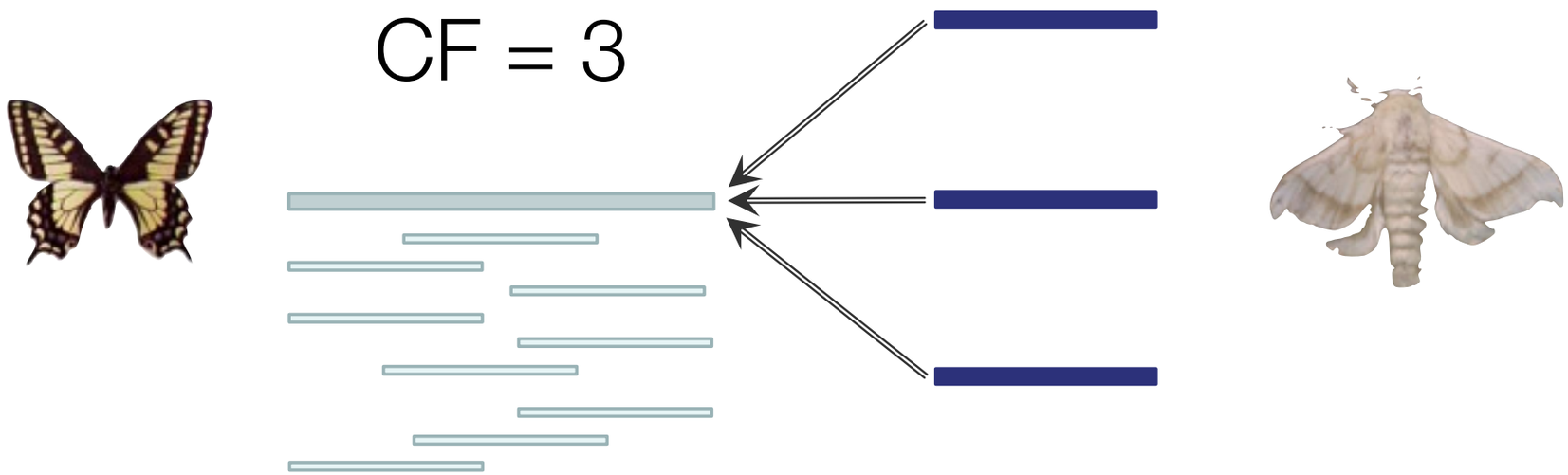
$$OHR = \frac{\alpha}{\beta}$$





Collapse Factor: a unigene over-assembly metric

Reverse annotation: match reference proteins to contigs

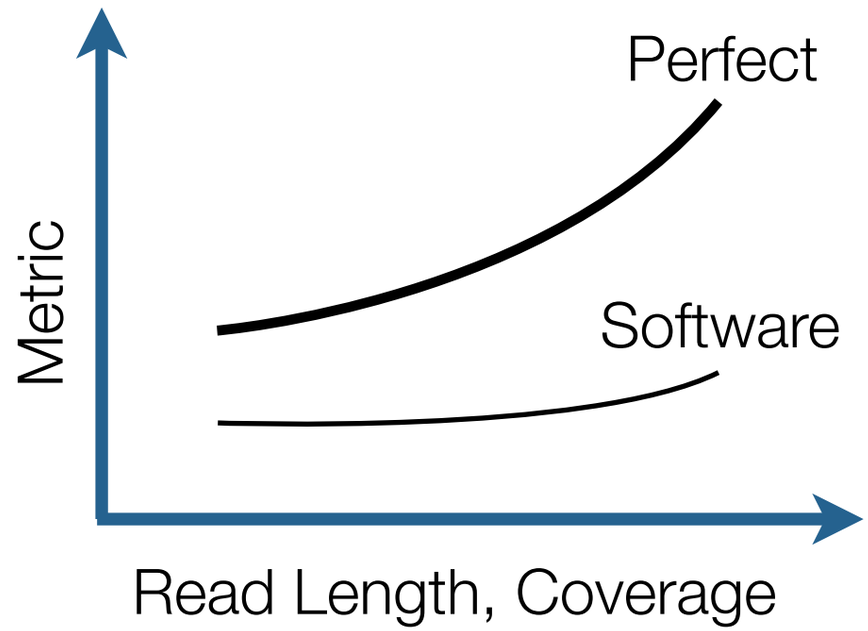


Experimental evaluation of metrics

We simulated 12 datasets
and assembled them
with a “perfect”
assembler and a real
software assembler
(Newbler).

Experimental evaluation of metrics

We simulated 12 datasets and assembled them with a “perfect” assembler and a real software assembler (Newbler).



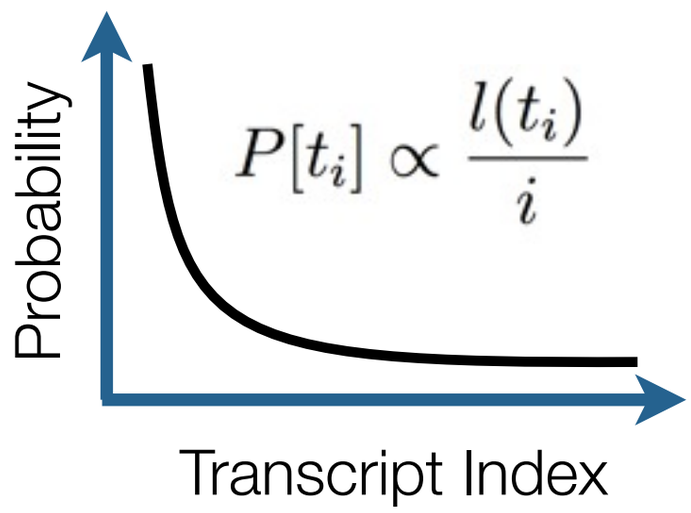
Things we can assume:

1) Perfect assemblies are more accurate.
2) Longer reads and higher coverage are more accurate (for perfect assemblies).

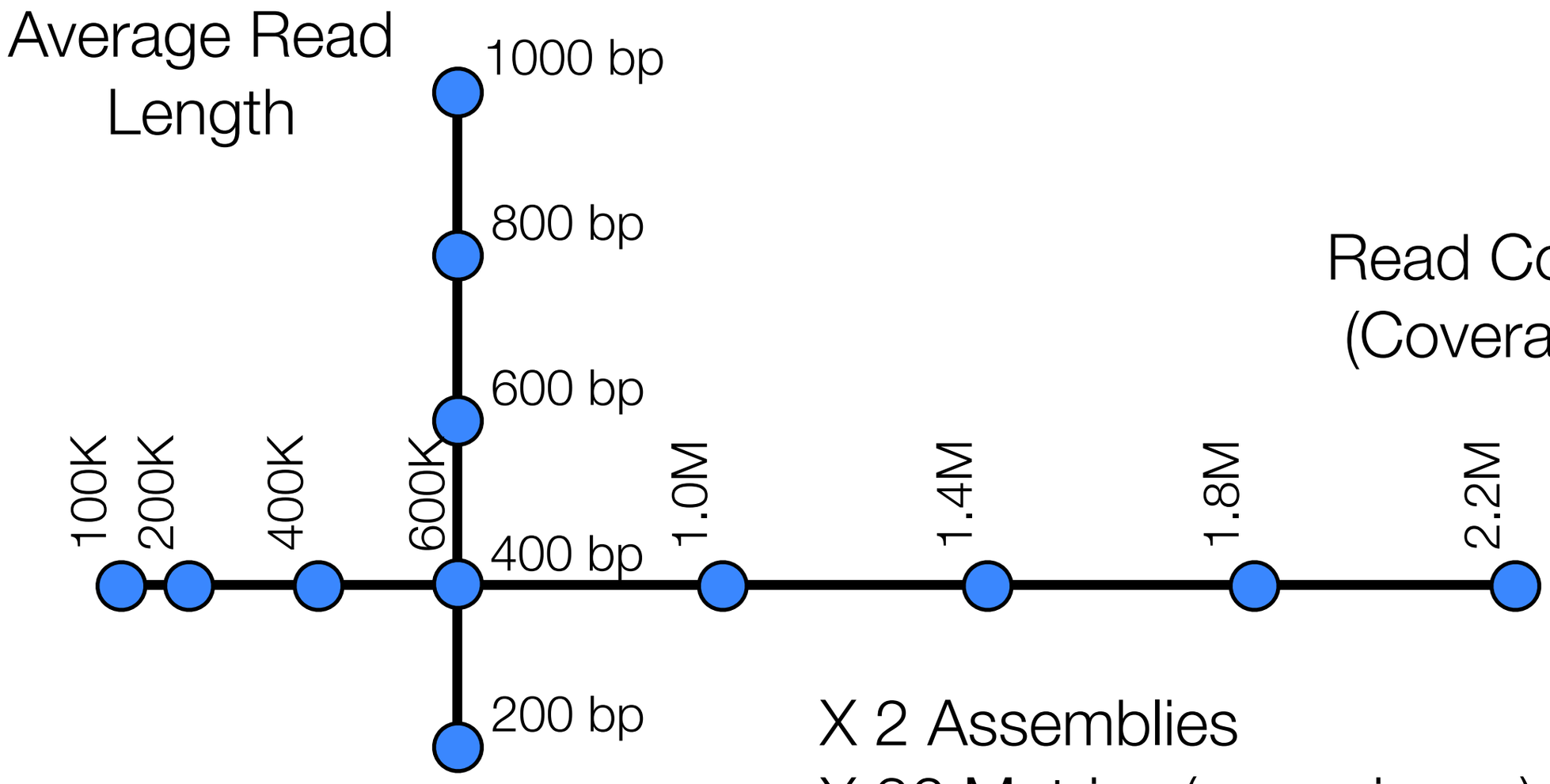
Simulated sequencing

We simulated sequencing from the *Drosophila melanogaster* transcript set, which includes splice variants and untranslated regions.

For each transcript, the probability of sequencing was proportional to a power-law modified by transcript length.

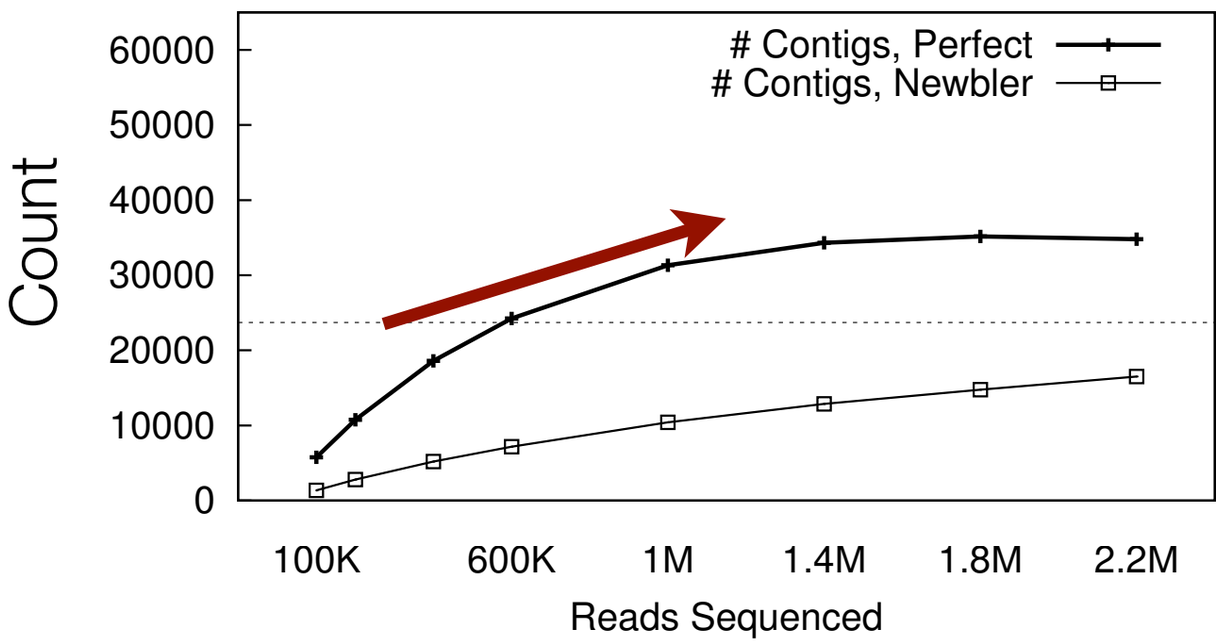


Datasets simulated, metrics evaluated

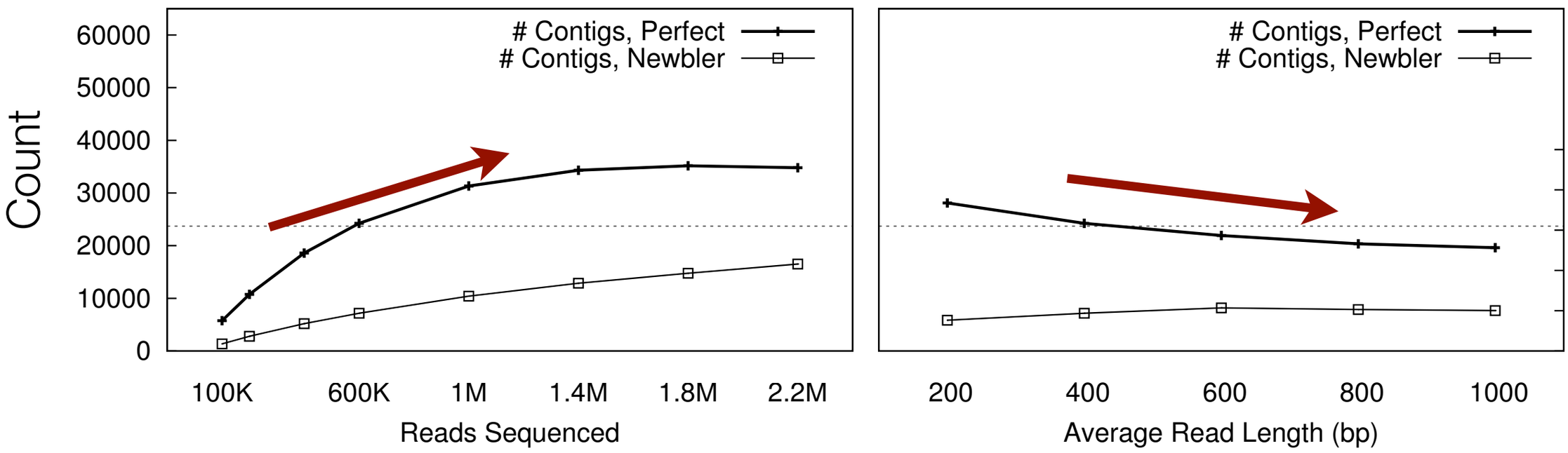


X 2 Assemblies
X 36 Metrics (per unigene)
~ 54.6M numbers and annotations

Basic assembly metrics: contig count

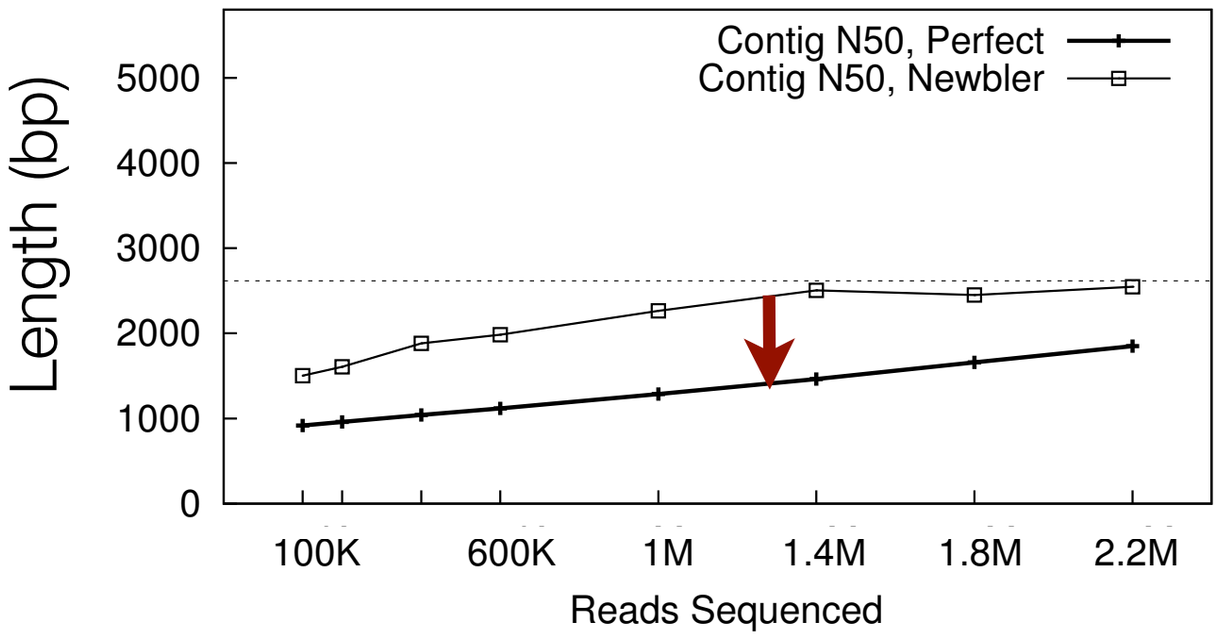


Basic assembly metrics: contig count

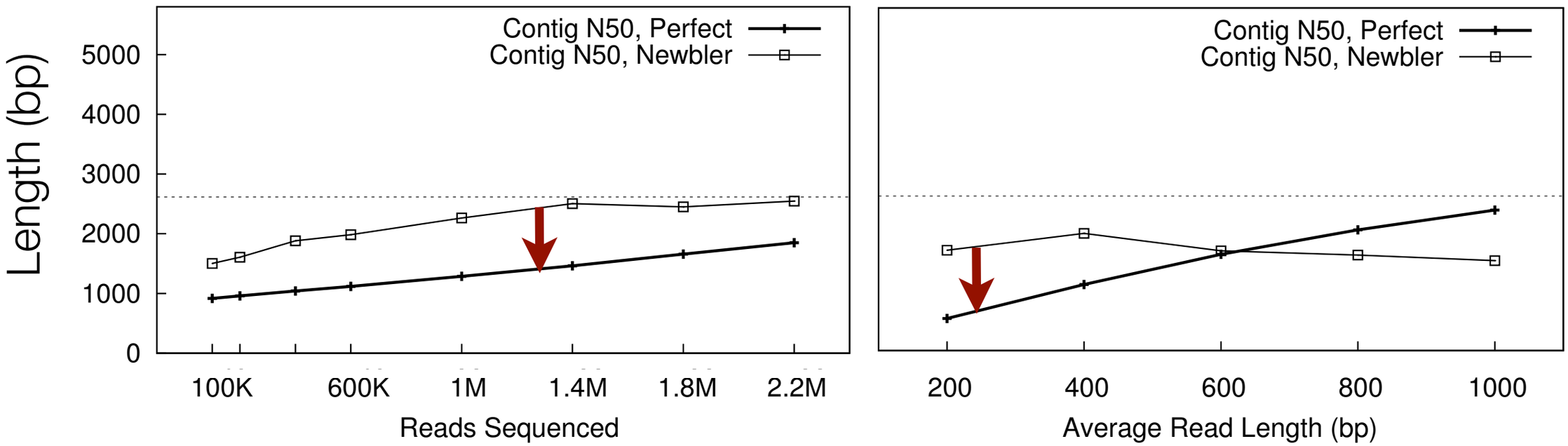


Contig count: larger for perfect assemblies, and not consistent between increasing sequencing depth/ length.

Basic assembly metrics: N50 lengths

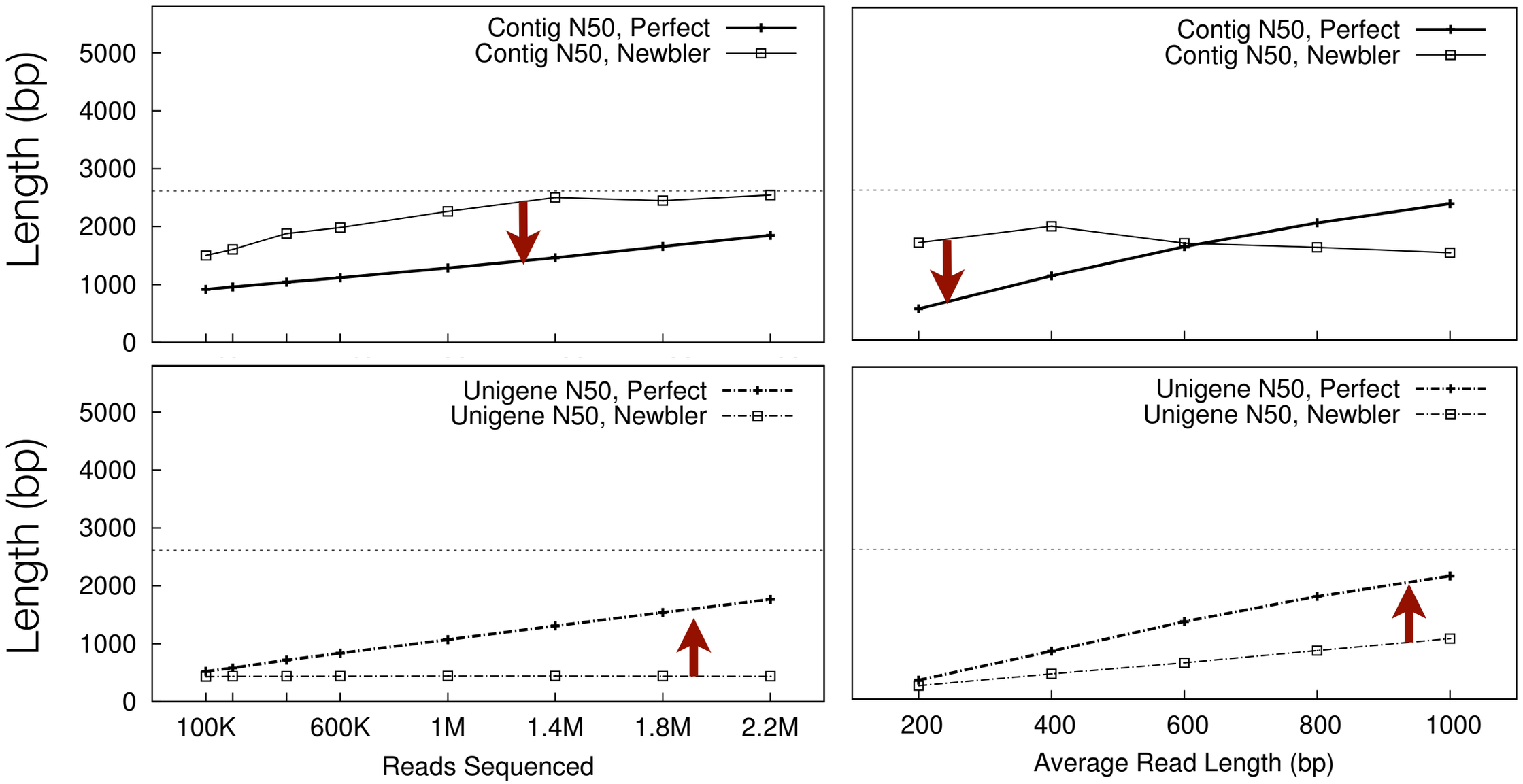


Basic assembly metrics: N50 lengths



Contig N50: Doesn’t quite exceed the true N50 size, but doesn’t show perfect assemblies as better (and is not consistent).

Basic assembly metrics: N50 lengths



Including singletons improves the metric.

Contigs are easy: using all the data is hard

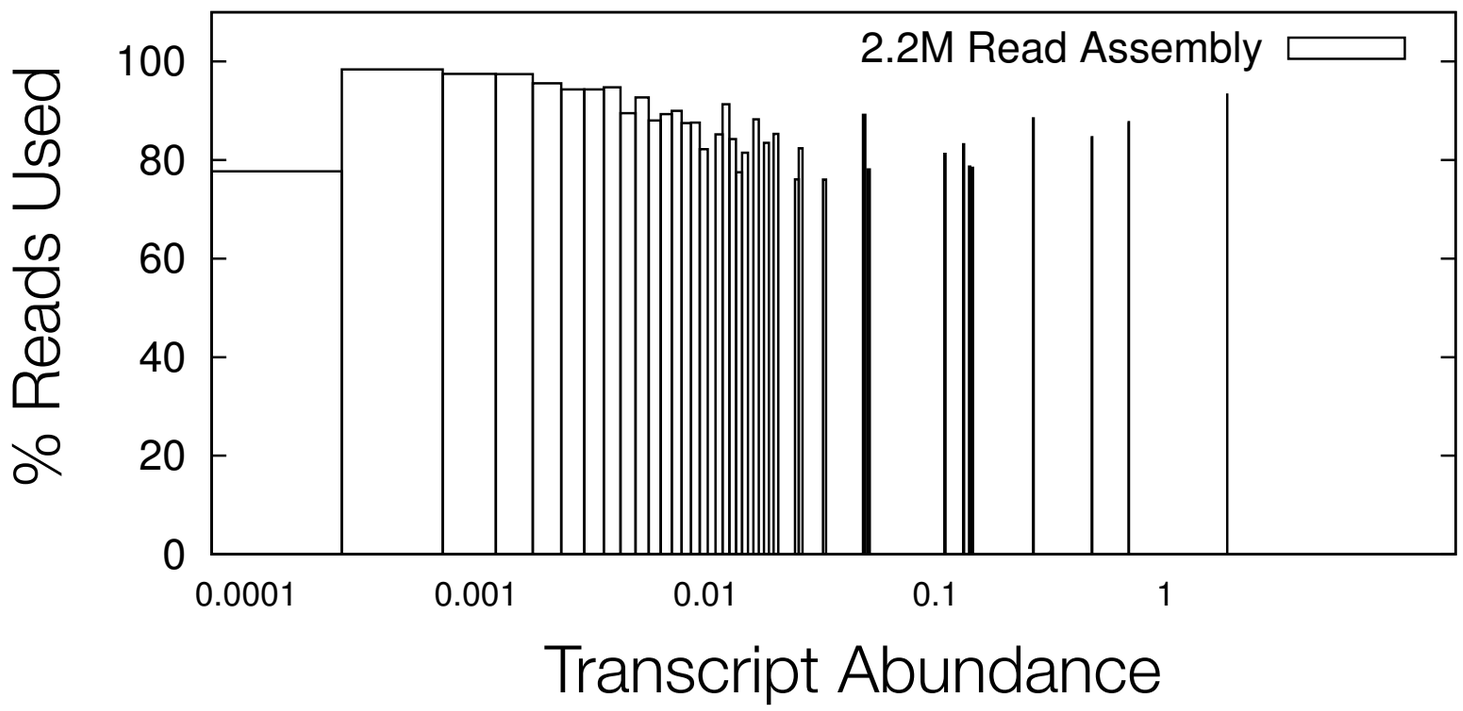
Often, contig-only assembly metrics made the software assembler look good, but weren't accurate.

It's easy to assemble a few highly covered contigs!

Contigs are easy: using all the data is hard

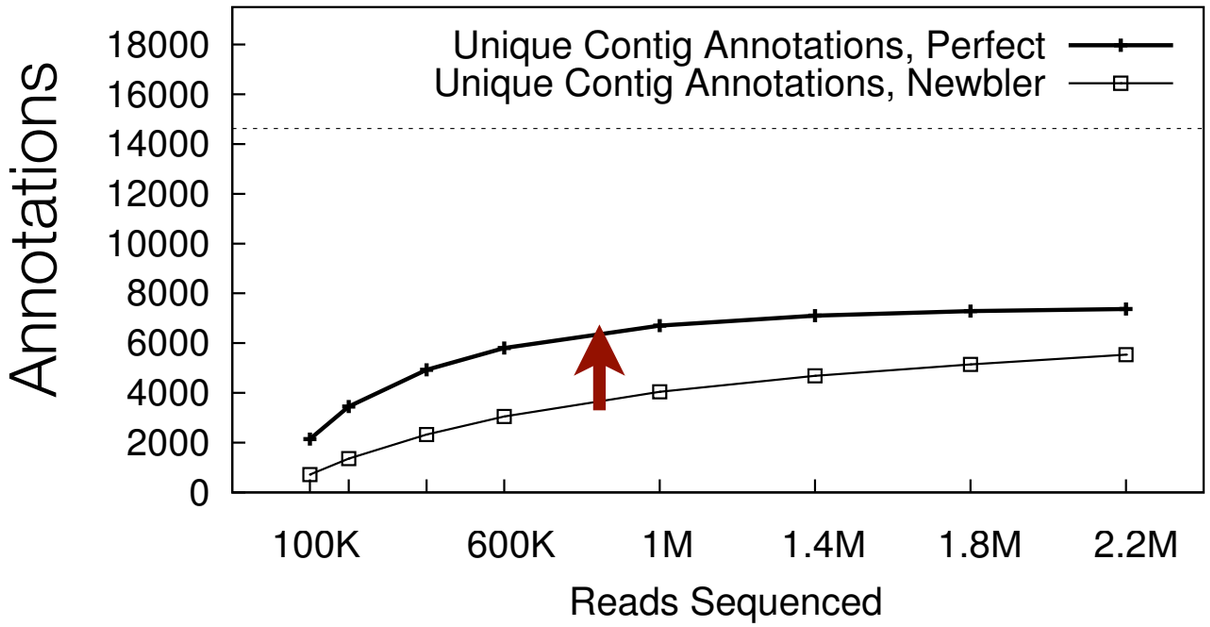
Often, contig-only assembly metrics made the software assembler look good, but weren't accurate.

It's easy to assemble a few highly covered contigs!



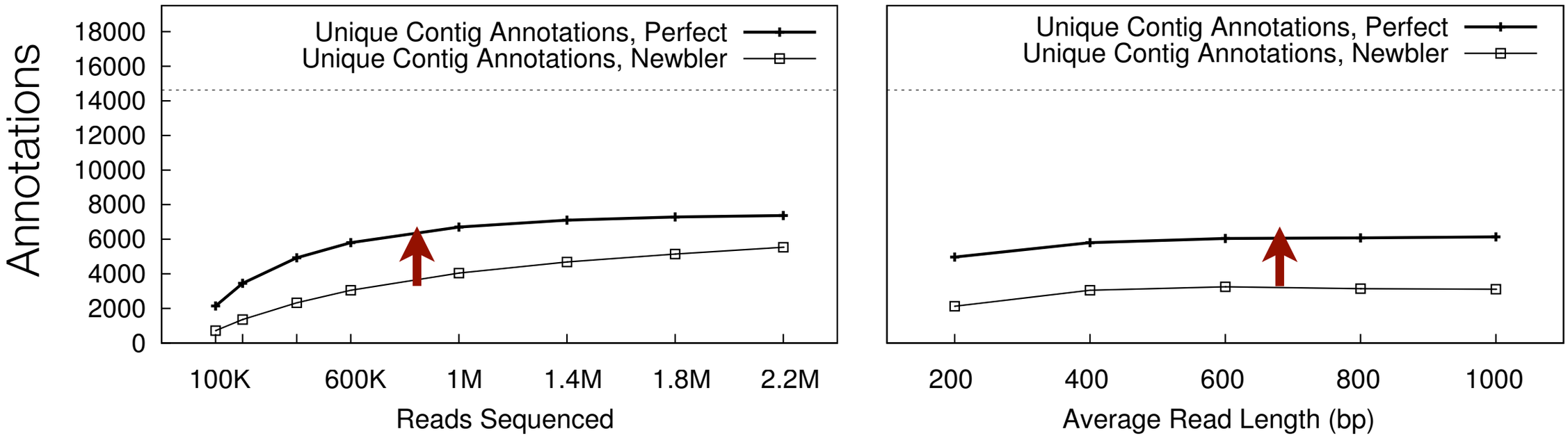
Annotation metrics: reference protein hits

Annotation against related reference: *B. mori*



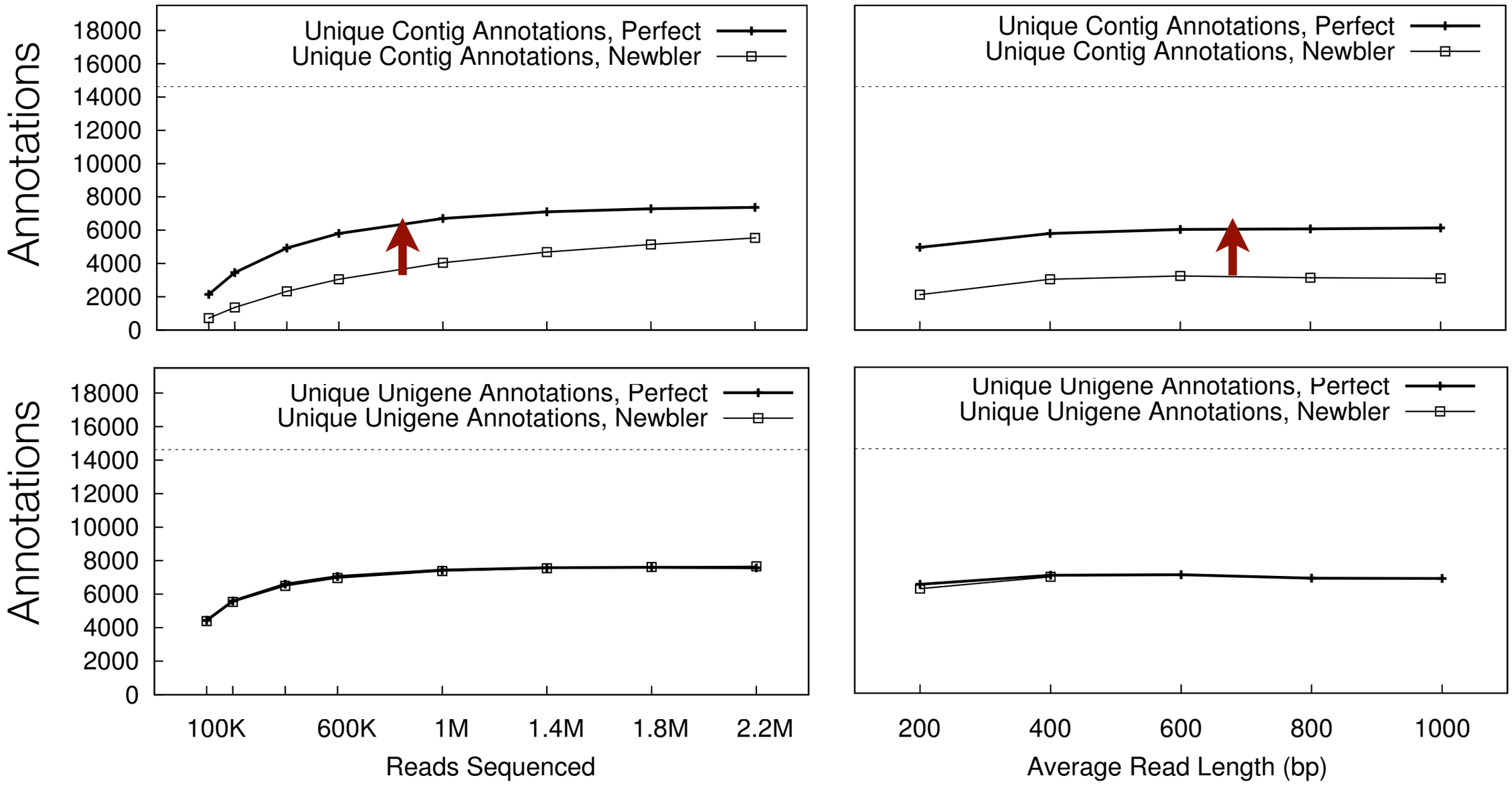
Annotation metrics: reference protein hits

Annotation against related reference: *B. mori*

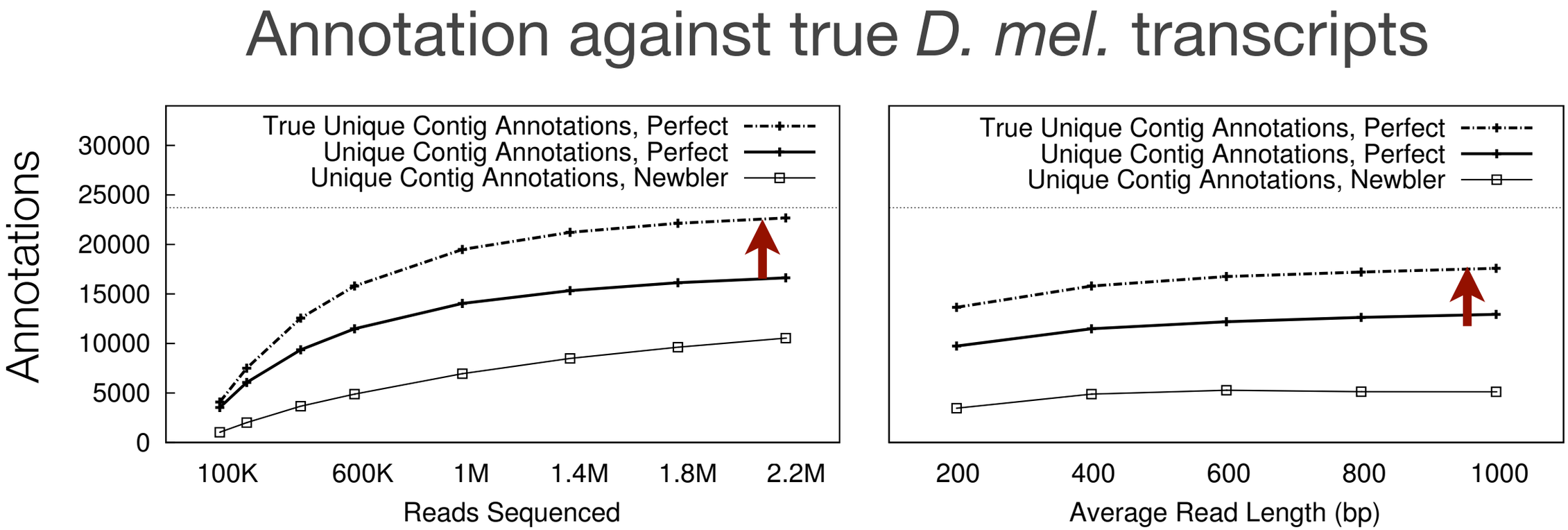


Annotation metrics: reference protein hits

Annotation against related reference: *B. mori*



Annotation metrics: protein hits—ideal



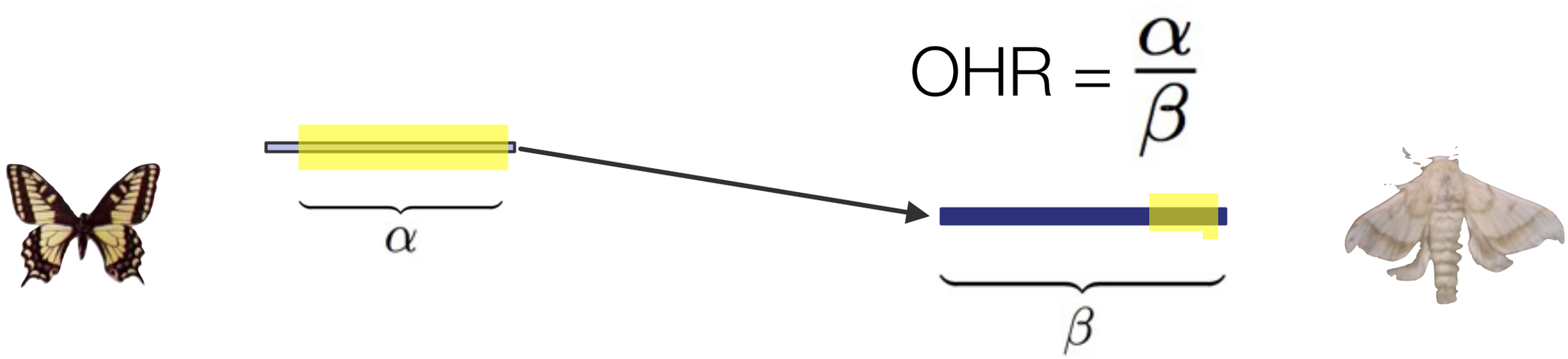
Annotation metrics: reads are static

In general, annotation-based metrics performed well when considering only contigs, less well on unigenes.

Annotation metrics: reads are static

In general, annotation-based metrics performed well when considering only contigs, less well on unigenes.

This wasn't the case for the average Ortholog Hit Ratio, which may be better considered an assembly metric.



Comparing OHRs

D. mel. Contig

B. mori OHR = 0.28

D. mel. OHR = 0.48

Comparing OHRs

D. mel. Contig

B. mori OHR = 0.78

D. mel. OHR = 0.48

$$\text{OHR Error} = \frac{\textit{B. mori}\ \text{OHR}}{\textit{D. mel.}\ \text{OHR}}$$

Comparing OHRs

D. mel. Contig

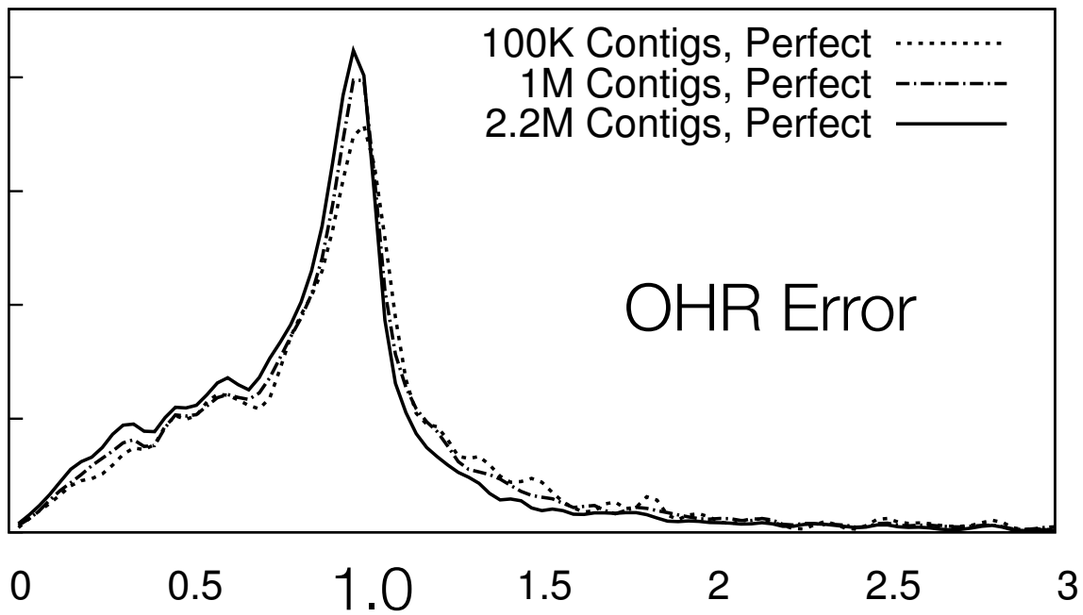
B. mori OHR = 0.78

D. mel. OHR = 0.48

$$\text{Normalized OHR Error} = \frac{\textit{B. mori}\ \text{OHR}}{\textit{D. mel.}\ \text{OHR}} \bullet \frac{\textit{B. mori}\ \text{Gene Length}}{\textit{D. mel.}\ \text{Gene Length}}$$

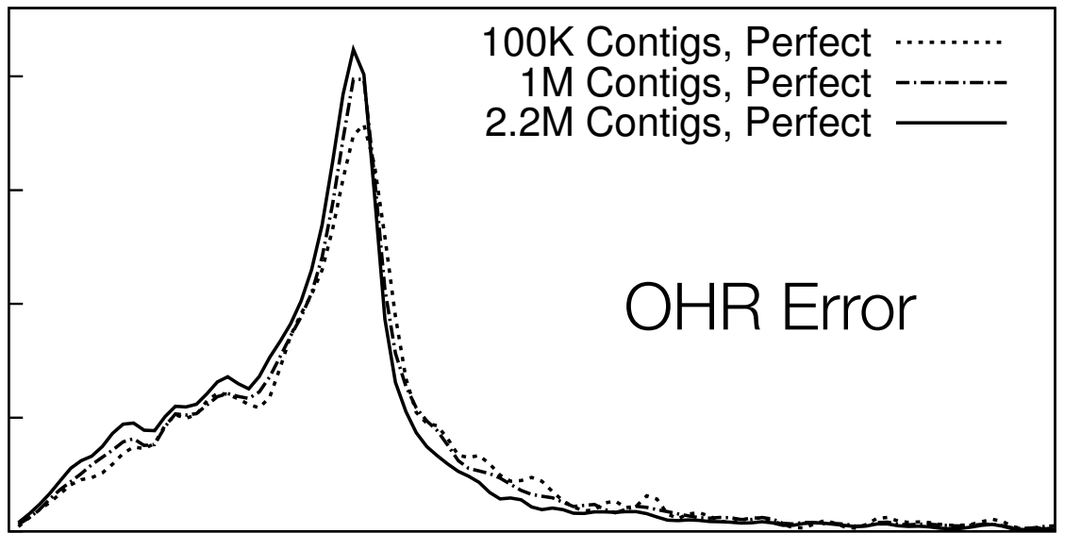
OHR error distributions

Ortholog hit ratios are generally conservative, tending to underestimate assembly completeness.

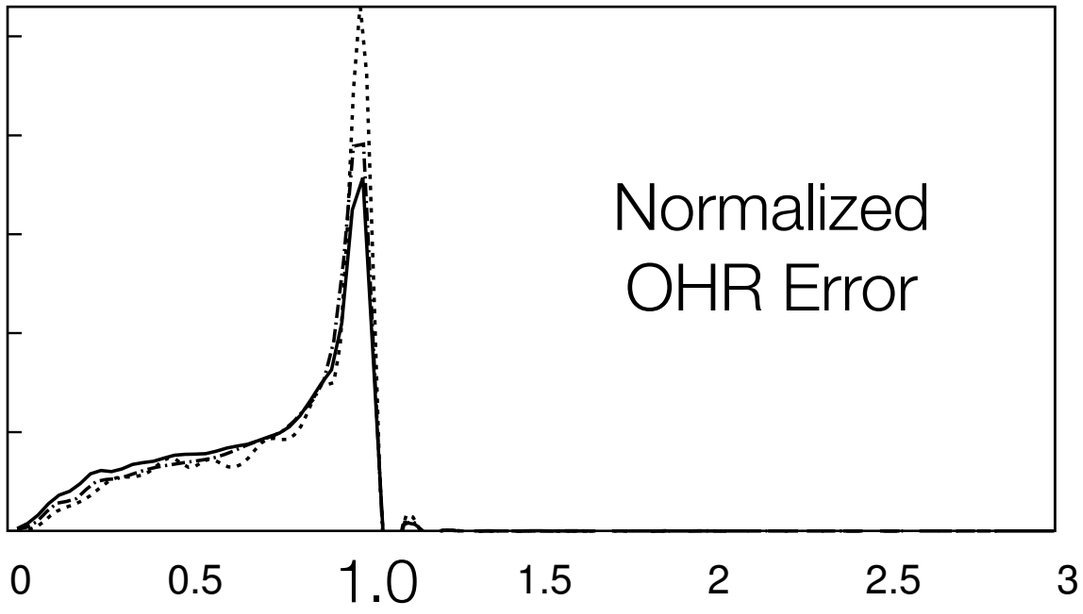


OHR error distributions

Ortholog hit ratios are generally conservative, tending to underestimate assembly completeness.



Overestimates of transcript discovery are usually due to indels in the species annotated against.



Overview - assembly metrics

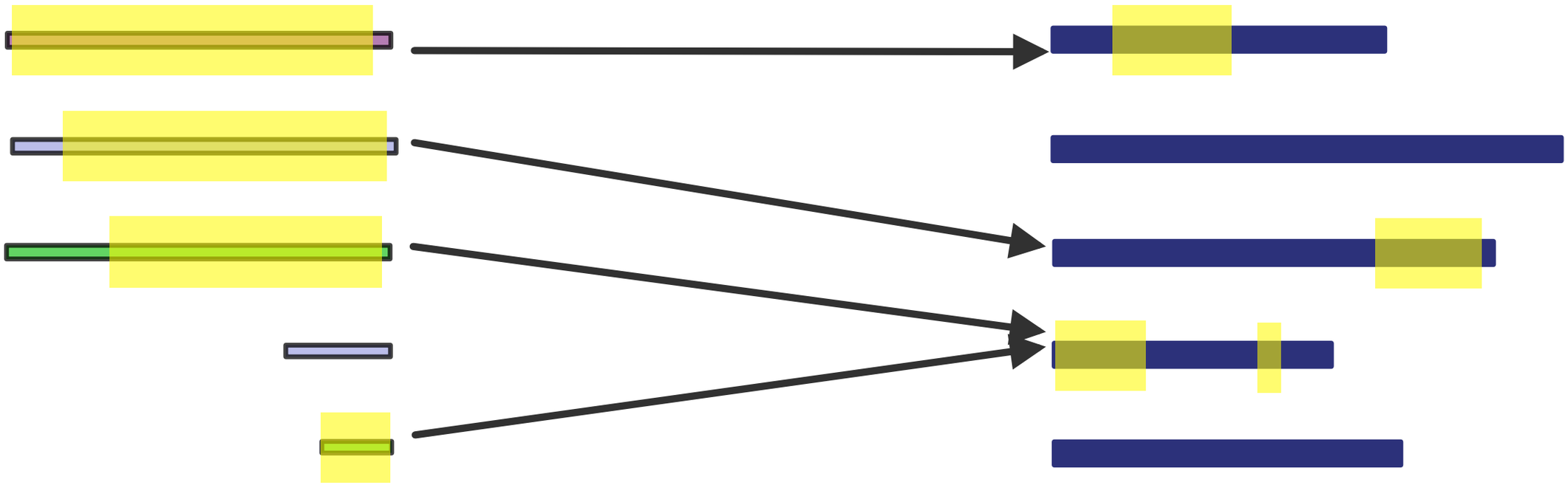
Metric	Consistent
Count of Contigs	No
Average Coverage, Contigs	No
Average Length/N50, Contigs	No
Average OHR, Contigs	No
% of Reads Assembled	Yes
Average Coverage, Unigenes	Yes
Average Length/N50, Unigenes	Yes
Average OHR, Unigenes	Yes

Overview - annotation metrics

Metric	Consistent
Unique Annotations, Contigs	Yes
Average Collapse Factor, Contigs	Yes
Unique Annotations, Unigenes	No
Average Collapse Factor, Unigenes	No

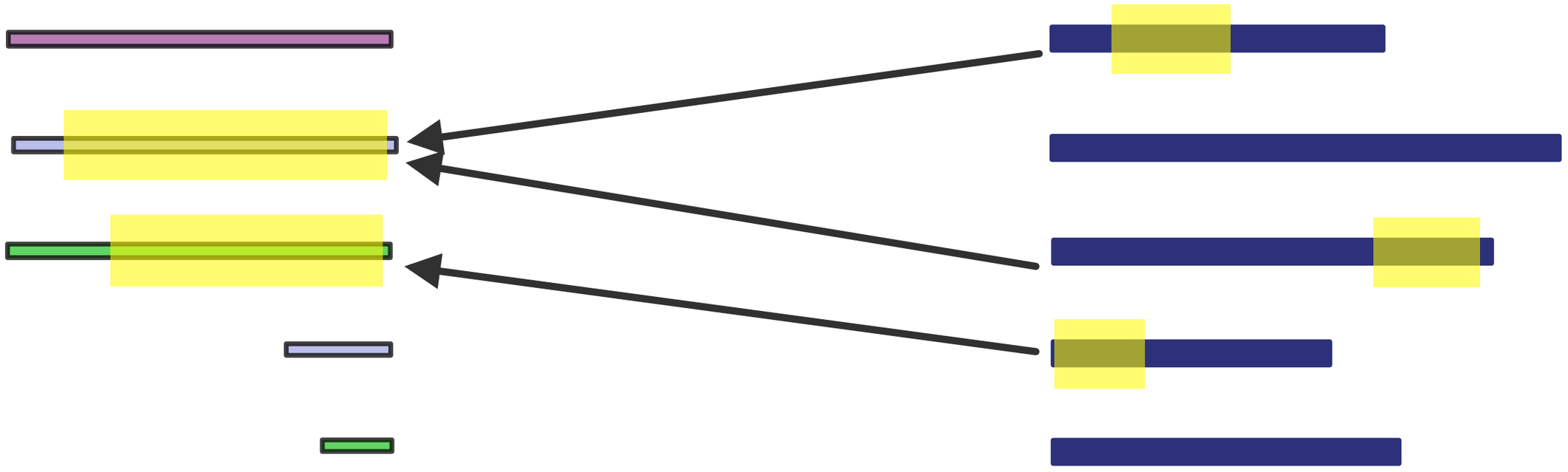
Other interesting annotation metics

Protein match count (rarefaction)



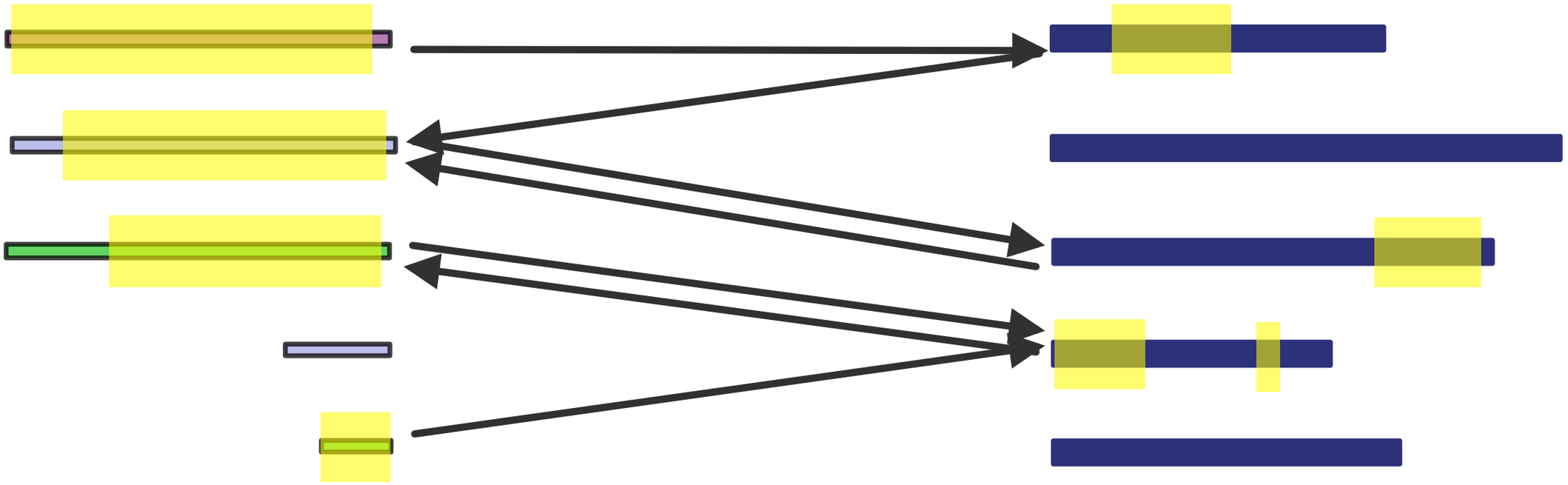
Other interesting annotation metics

Reverse protein match count (reverse rarefaction)



Other interesting annotation metics

Reciprocal best hit (RBH) count



Overview - annotation metrics

Metric	Consistent
Unique Annotations, Contigs	Yes
Average Collapse Factor, Contigs	Yes
Unique Annotations, Unigenes	No
Average Collapse Factor, Unigenes	No
Reverse Annotations, Contigs	Yes
Reverse Annotations, Unigenes	Yes
Reciprocal Best Hits, Contigs	Yes
Reciprocal Best Hits, Unigenes	Yes

Conclusion

We need metrics specific to problems at hand, and we should validate those metrics.

Unassembled reads are an important part of a de-novo transcriptome assembly.

Combine assembly and annotation-based metrics; also, consider matching proteins to your assembly (TBLASTN) in addition to vice-versa (BLASTN).