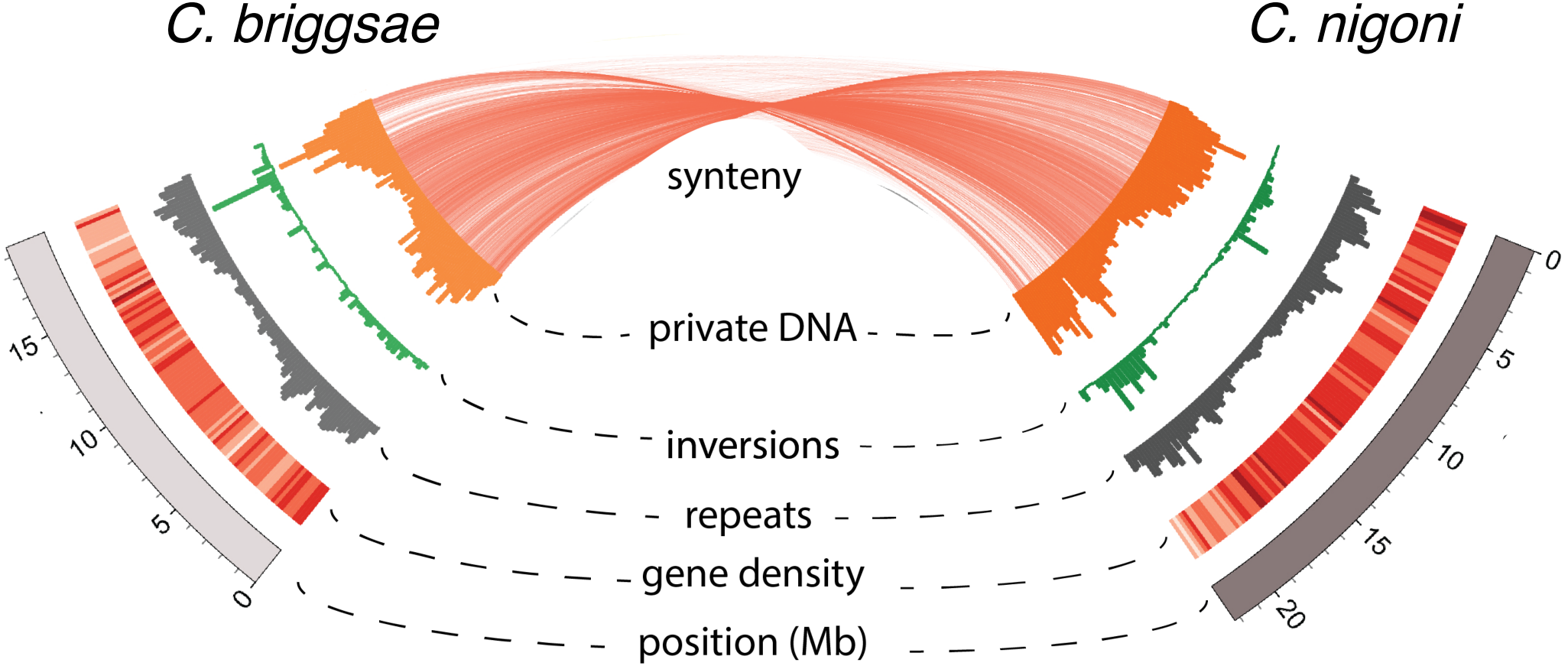
Assembling and biologically interpreting nematode genomes

Erich Schwarz, Cornell



DIBSI, UC Davis, July 2018

Questions

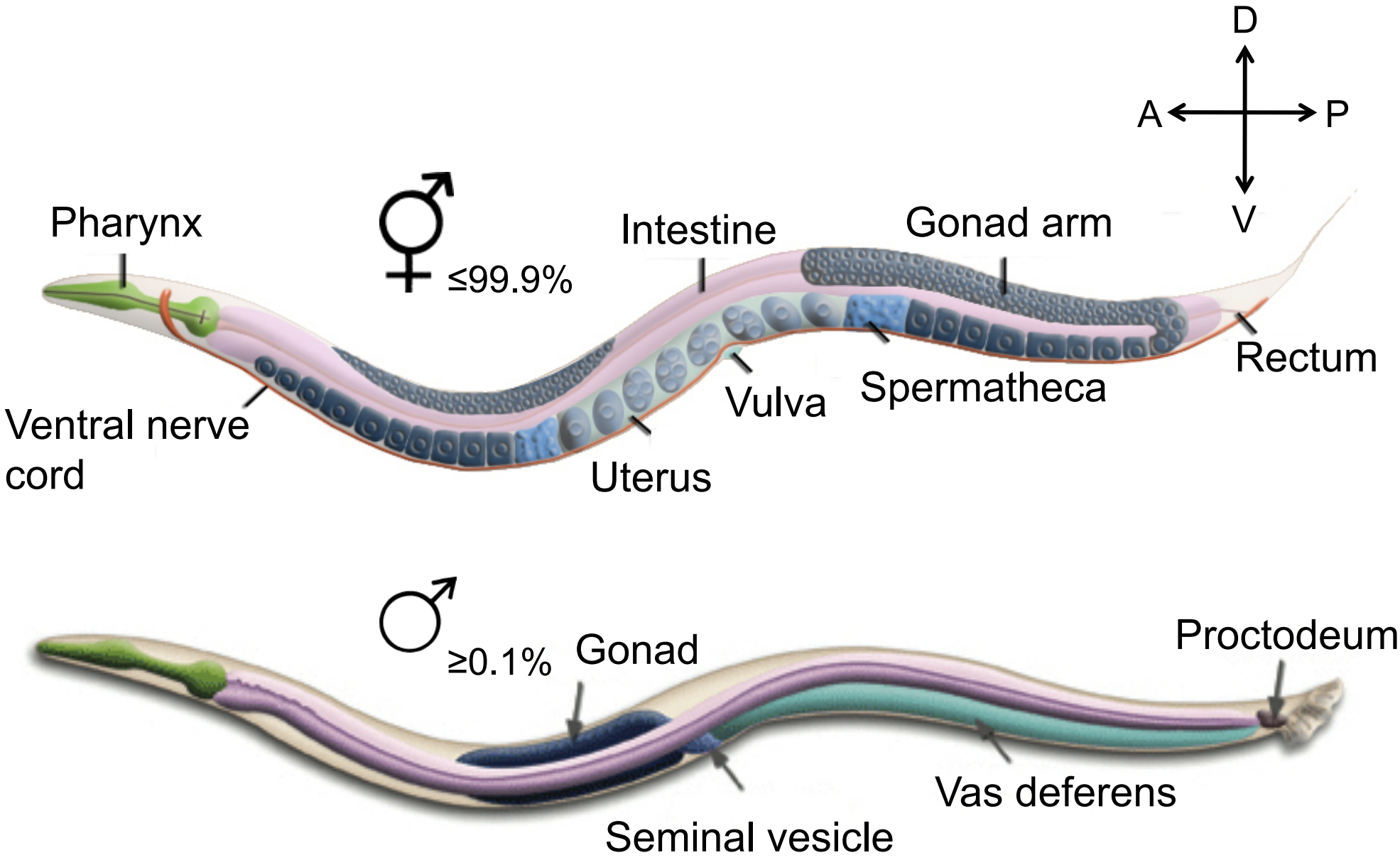
1. Why are nematode genomes interesting?

2. Why is long-read (third-generation) genome assembly a good idea?

3. What parts of a nematode genome are needed for male-female sexuality (versus hermaphroditism)?

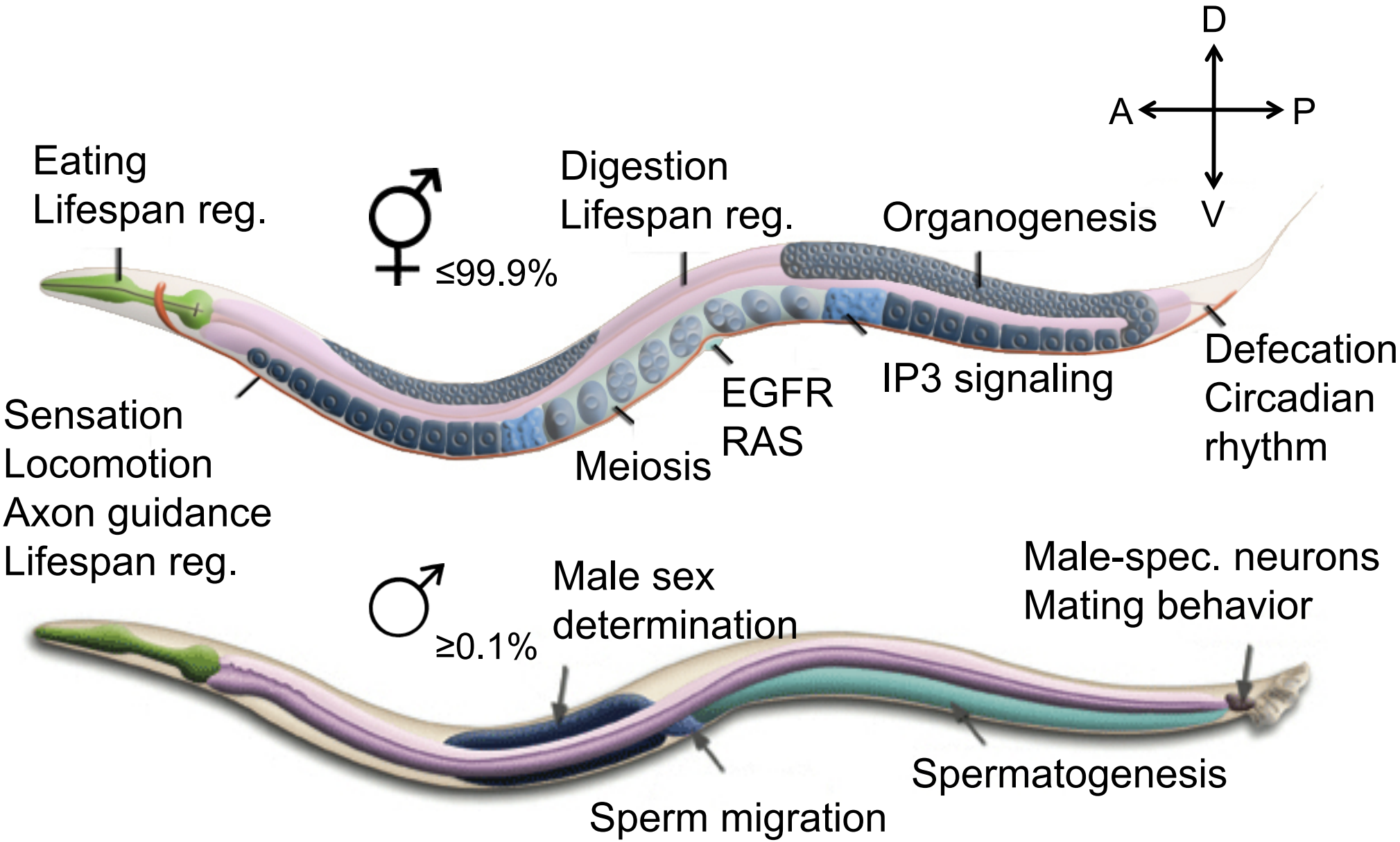
4. What Nth-generation assembly methods might improve our biological analyses?

Caenorhabditis elegans, i.e., "the worm"



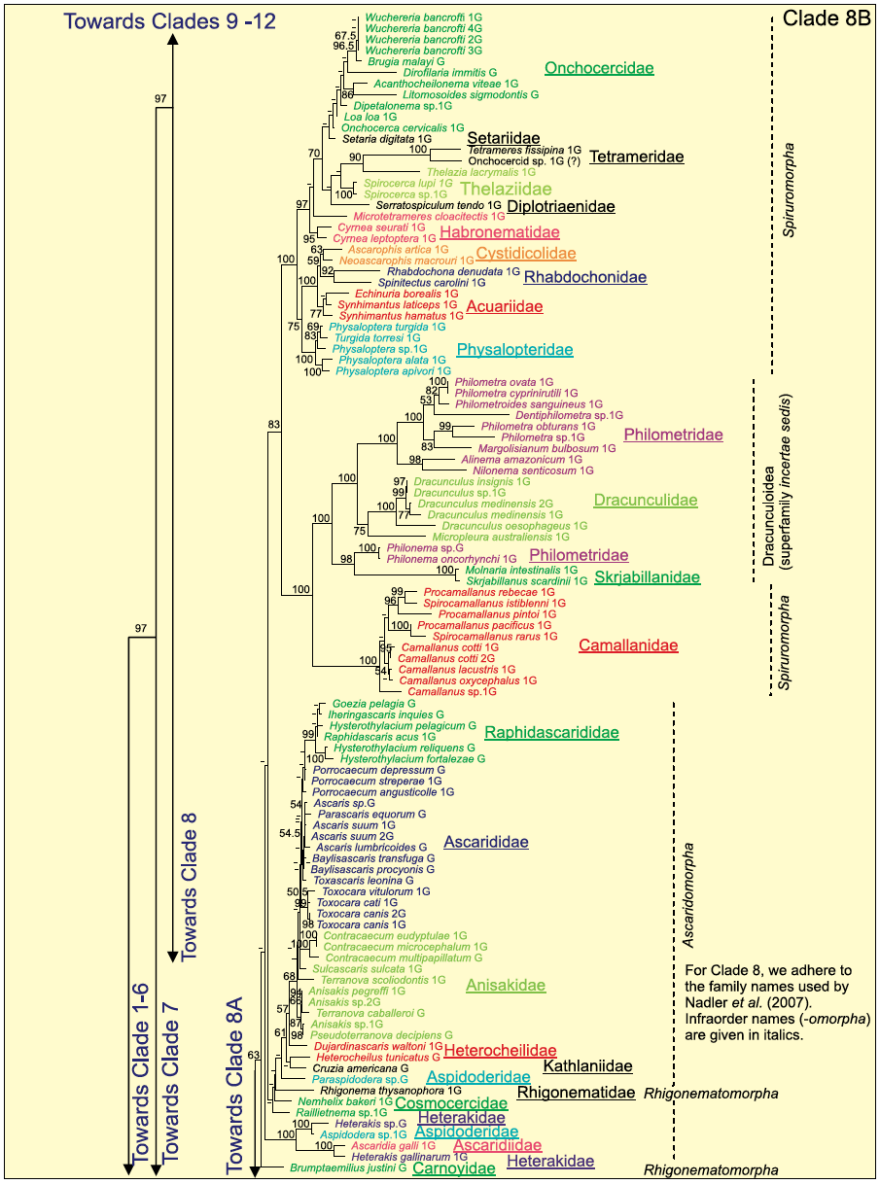
Refs.: Blaxter (2011), PLoS Biol. *9*, e1001050; Hall (2010), *www.wormatlas.org*.

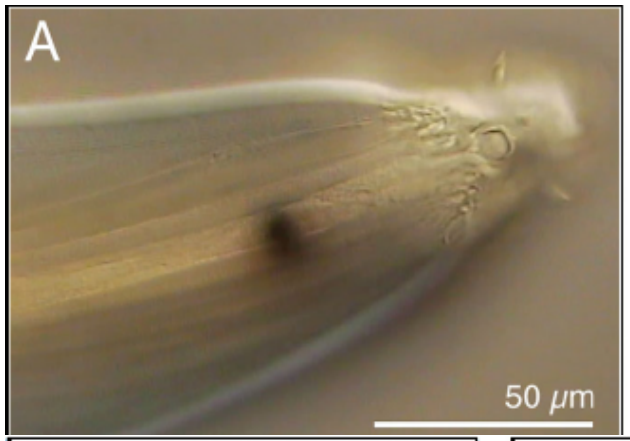
C. elegans is a useful model for animal biology

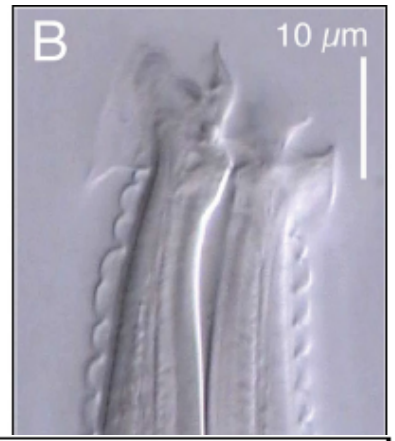


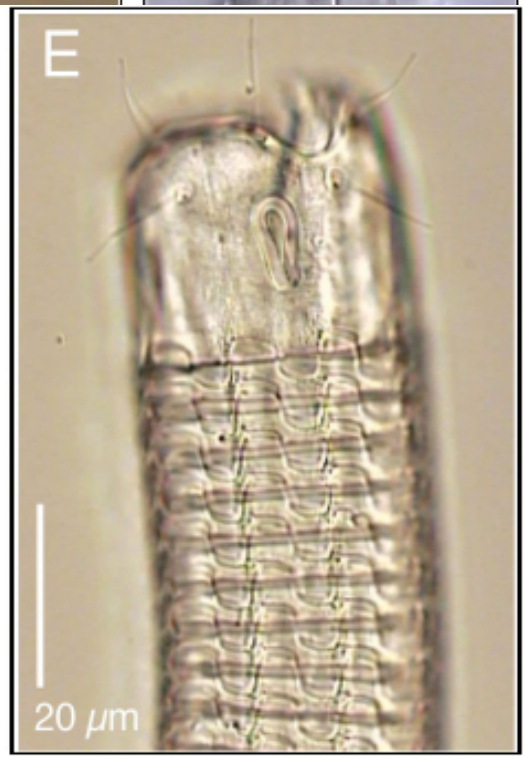
Refs.: Blaxter (2011), PLoS Biol. *9*, e1001050; Hall (2010), *www.wormatlas.org*.

There is a great diversity of other nematodes (~1 M species?), mostly uncharacterized



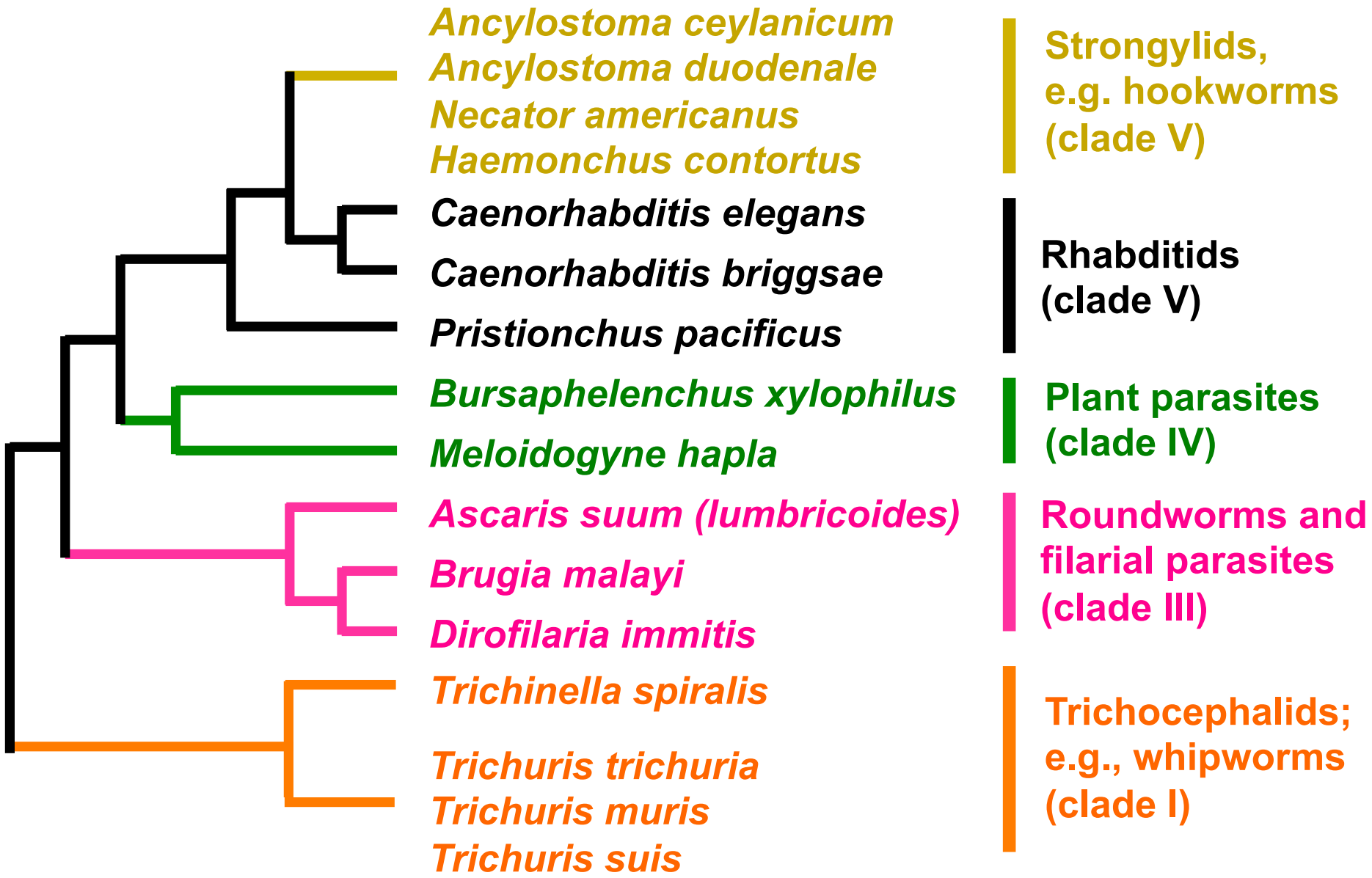






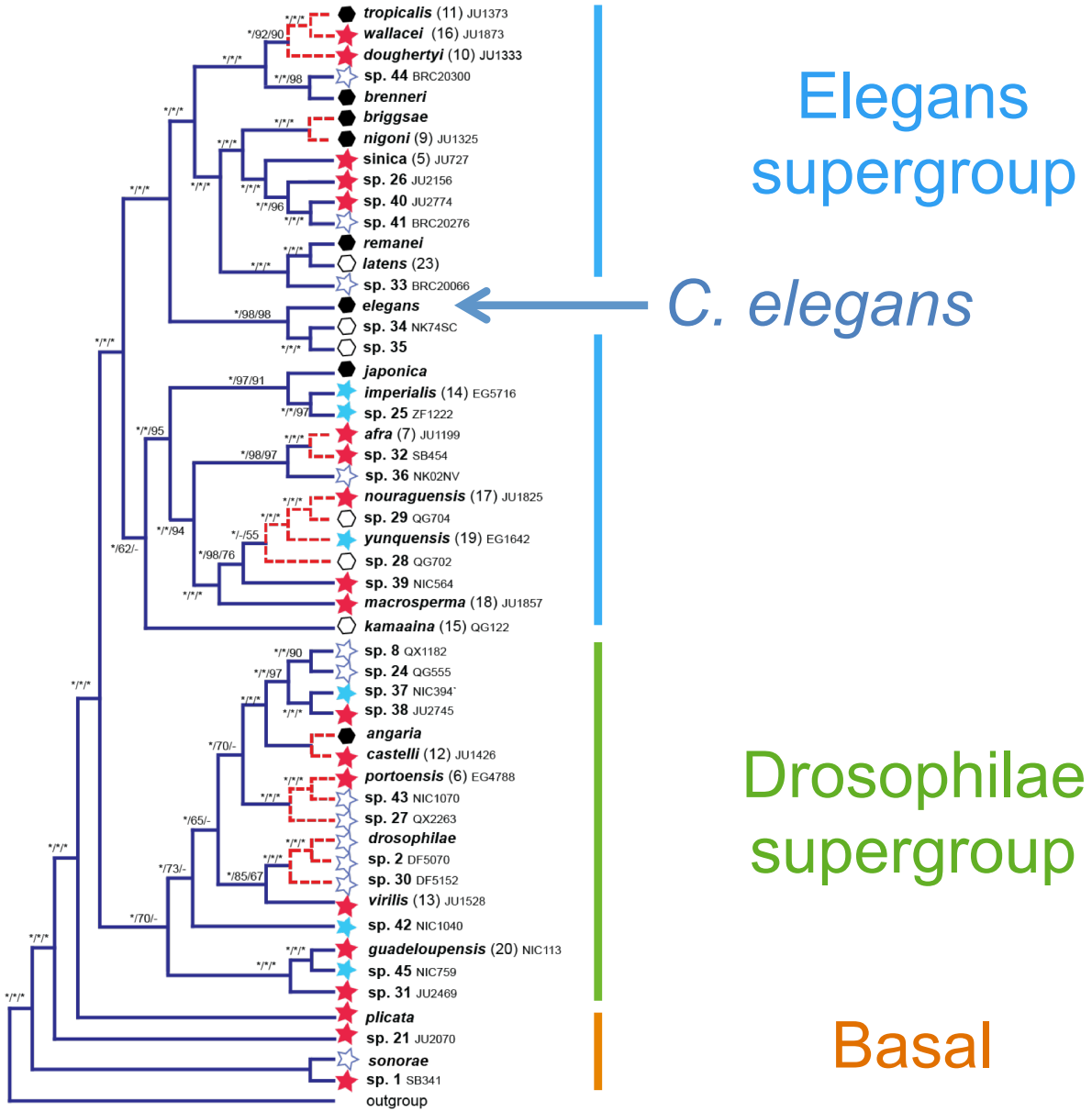
Refs.: Lambshead (1993), Oceanis *19*, 5–24; De Ley (2006), WormBook, *2006 Jan 25*,1-8; van Megen et al. (2009), Nematology *11*, 927-950.

A small but important subset of nematodes parasitize humans, other animals, and plants



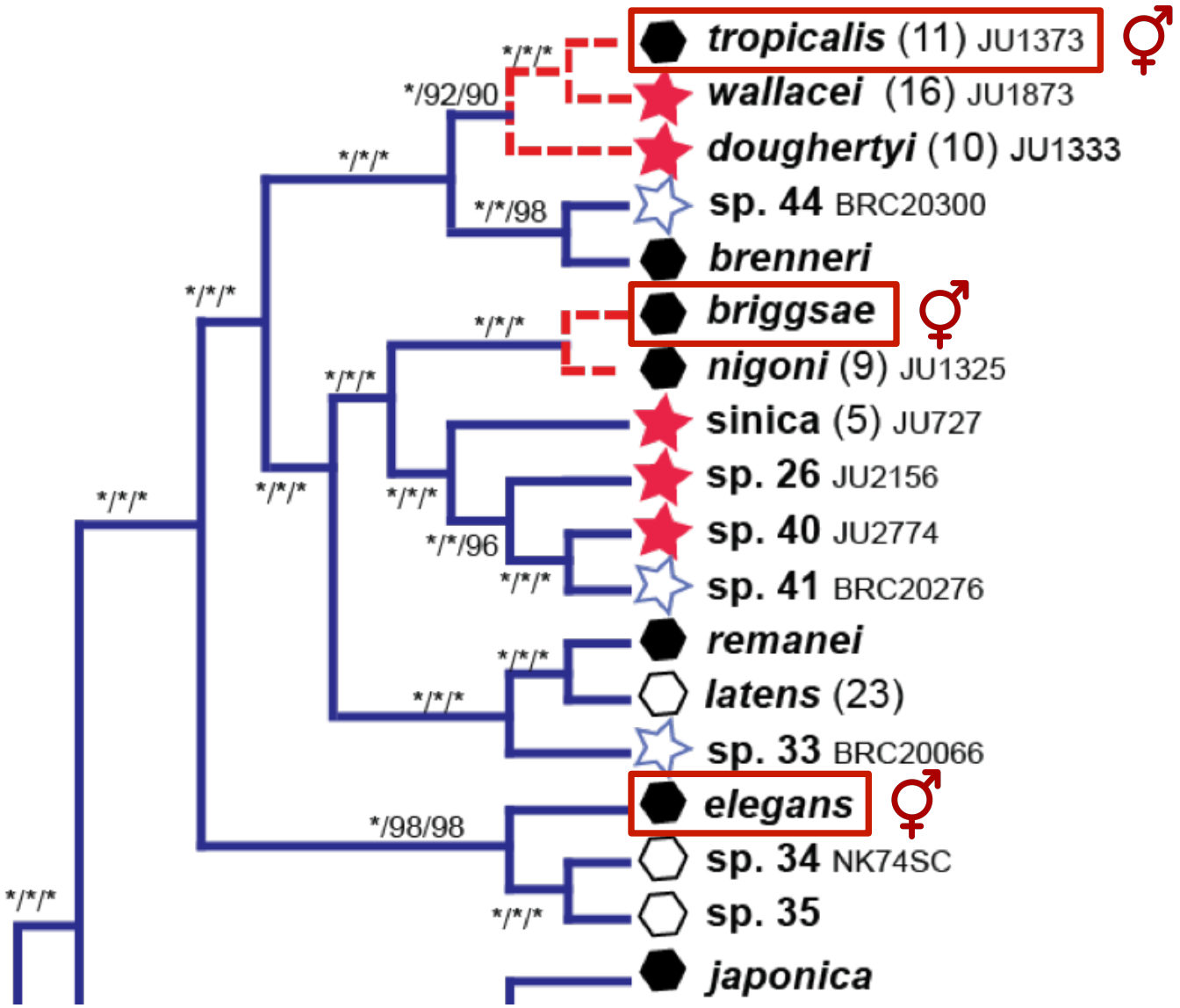
Refs.: Kiontke et al. (2011), BMC Evol. Biol. *11*, 339; van Megen et al. (2009), Nematology *11*, 927-950.

Even within the *Caenorhabditis* genus alone, there are over 50 known species (so far)



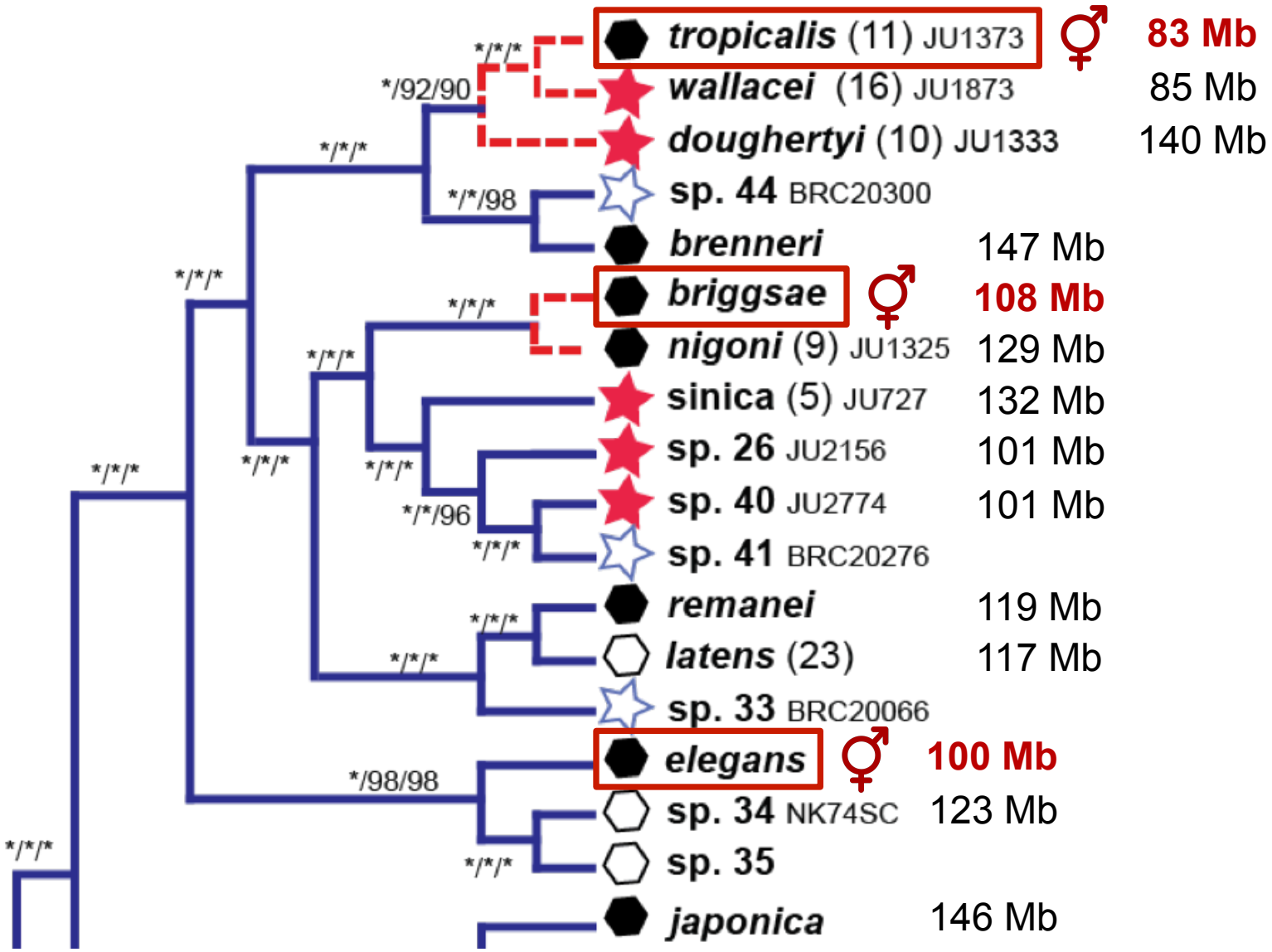
Unpublished data by Kiontke and Fitch, Sep. 2015.

In *Caenorhabditis,* hermaphrodites repeatedly arise from male-female origins. How?



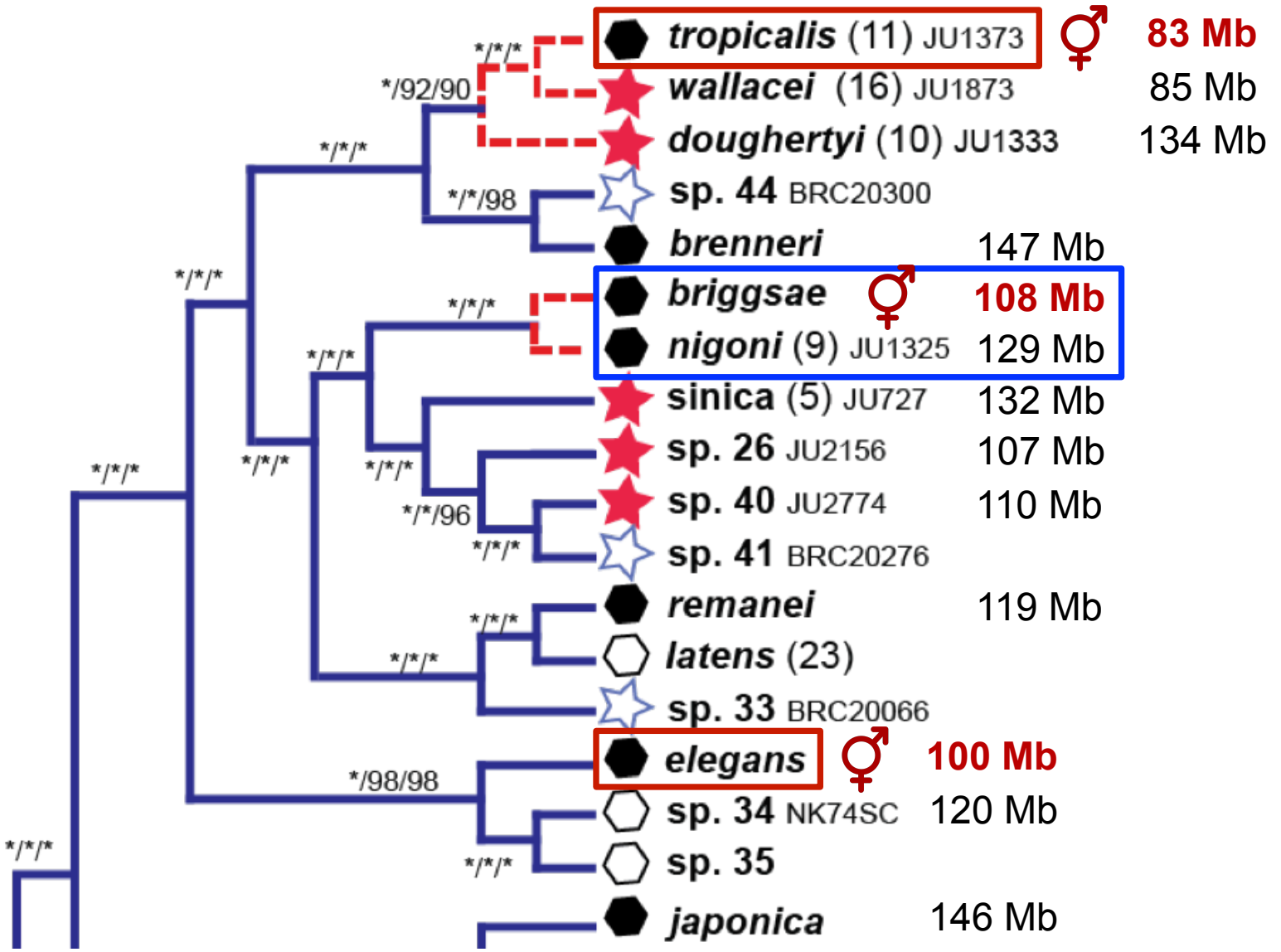
Refs.: Kiontke et al. (2004), PNAS *101*, 9003-9008; Cho et al. (2004), Genome Res. *14*, 1207-1220; unpublished data by Kiontke, Fitch and Blaxter, Sep. 2015.

A clue: genomes of hermaphrodites tend to be smaller than male-female genomes



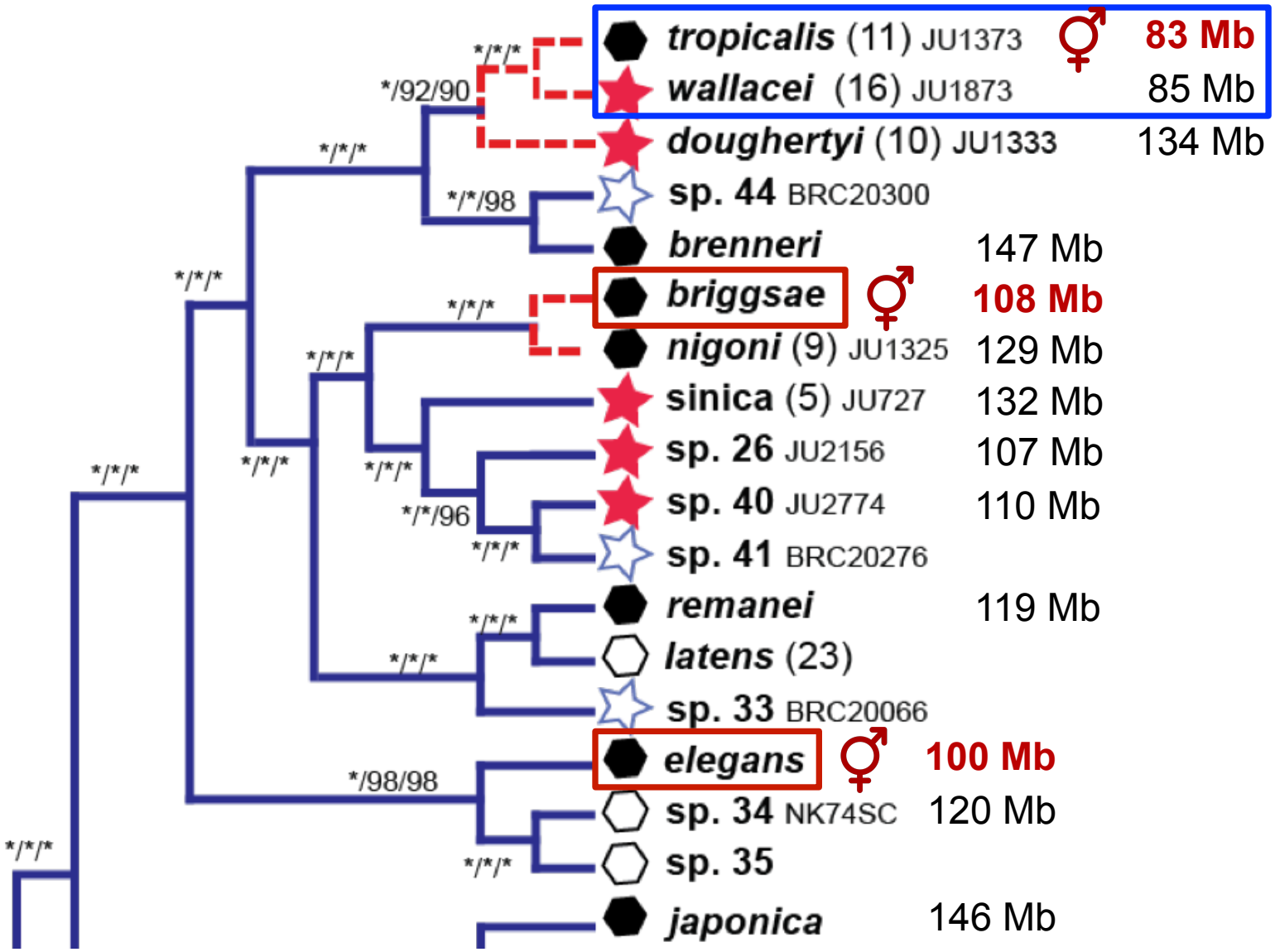
Refs.: *C. elegans* Sequencing Consortium (1998), Science *282*, 2012-2018.; Stein et al. (2003), PLoS Biol. *1*, E45; Fierst et al. (2015), PLoS Genet. *11*, e1005323; unpublished data (Blaxter et al.; Schwarz et al.).

Yet, male-female *C. nigoni* (sp. 9) is partially interfertile with hermaphroditic *C. briggsae*!



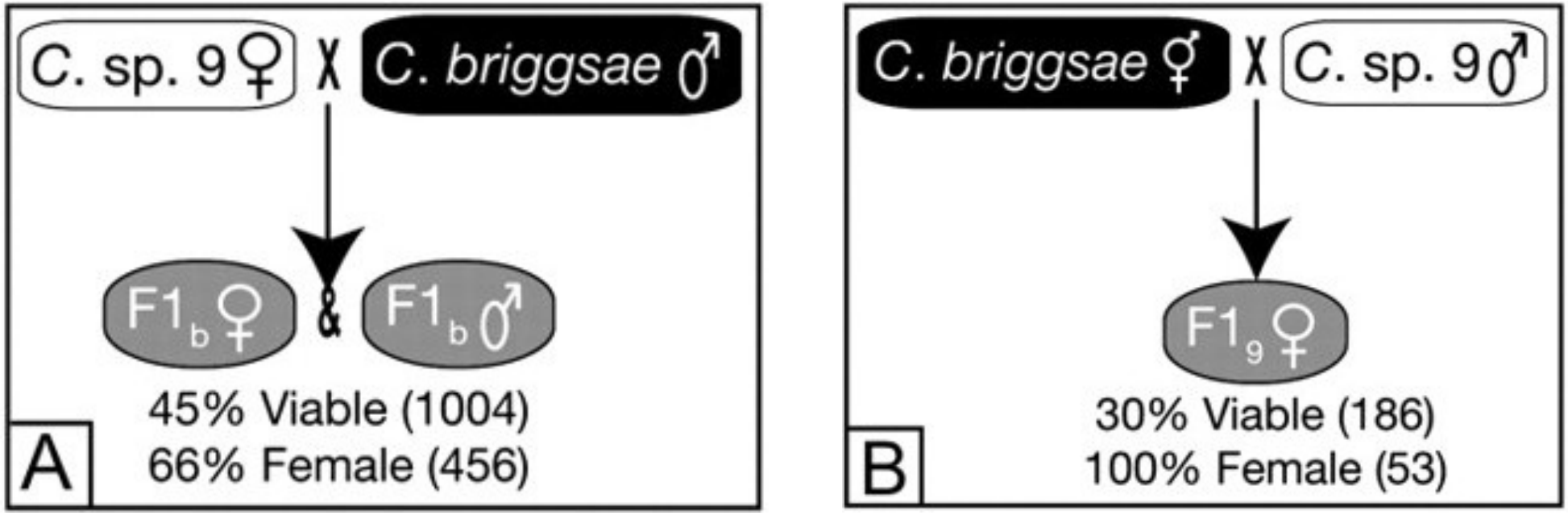
Refs.: *C. elegans* Sequencing Consortium (1998), Science *282*, 2012-2018.; Stein et al. (2003), PLoS Biol. *1*, E45; Fierst et al. (2015), PLoS Genet. *11*, e1005323; unpublished data (Blaxter et al.; Schwarz et al.).

An awkward counterexample: *C. wallacei* versus *C. tropicalis*



Refs.: *C. elegans* Sequencing Consortium (1998), Science *282*, 2012-2018.; Stein et al. (2003), PLoS Biol. *1*, E45; Fierst et al. (2015), PLoS Genet. *11*, e1005323; unpublished data (Blaxter et al.; Schwarz et al.).

Nevertheless, can we use *C. nigoni*'s genome to explain how *C. briggsae* evolved from it?



Ref.: Woodruff et al. (2010), Genetics *186*, 997-1012.

Summing up: nematodes

Nematodes include:
a highly tractable model organism (*C. elegans*),
along with ~1M other diverse nematode species
including parasites of one billion humans.

Genomics is a tool that should allow us to dissect
the molecular and genetic basis
of these phenomena,
particularly if we can compare
high-quality genomes to one another.

Questions

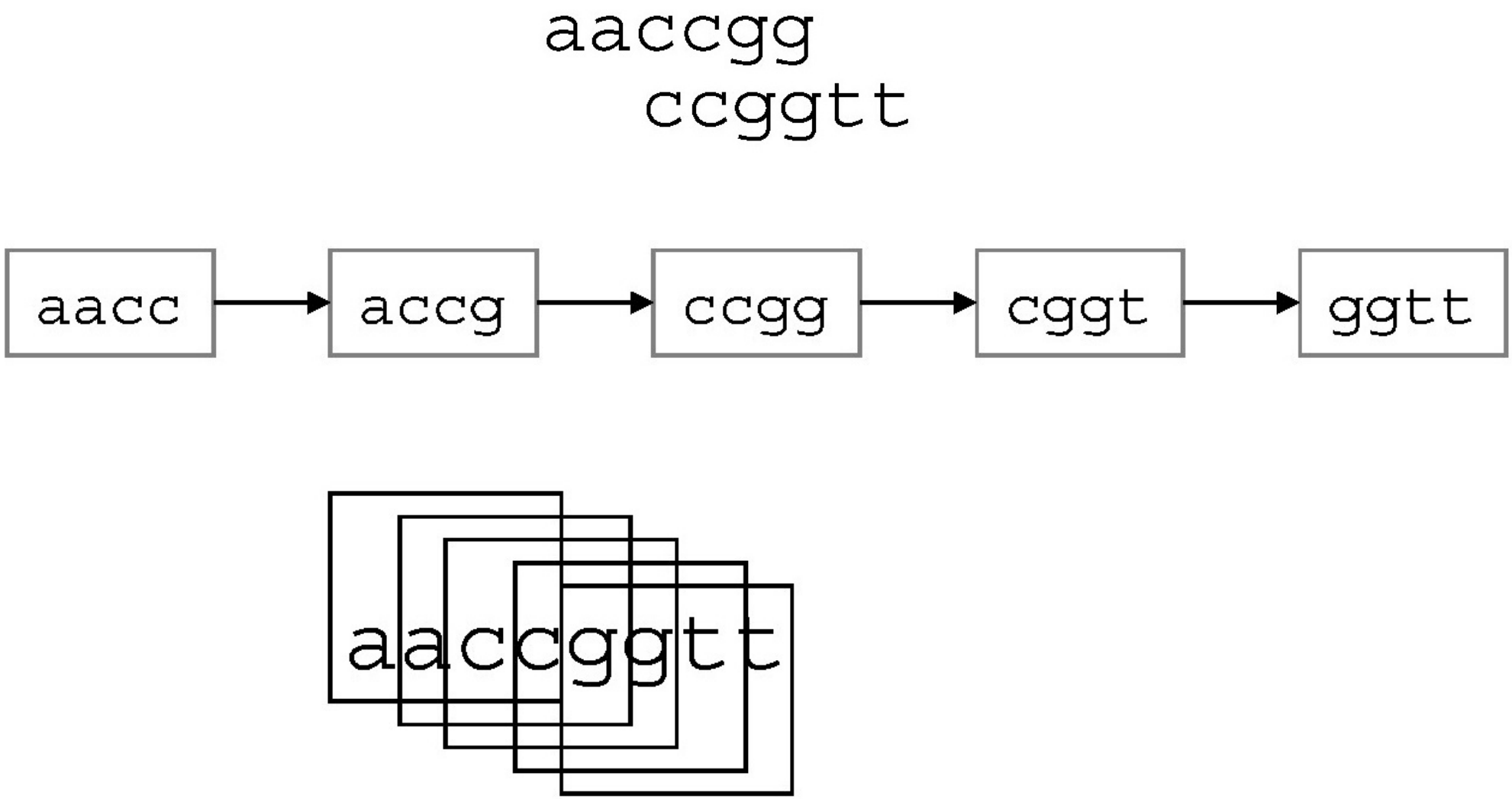
1. Why are nematode genomes interesting?

2. Why is long-read (third-generation) genome assembly a good idea?

3. What parts of a nematode genome are needed for male-female sexuality (versus hermaphroditism)?

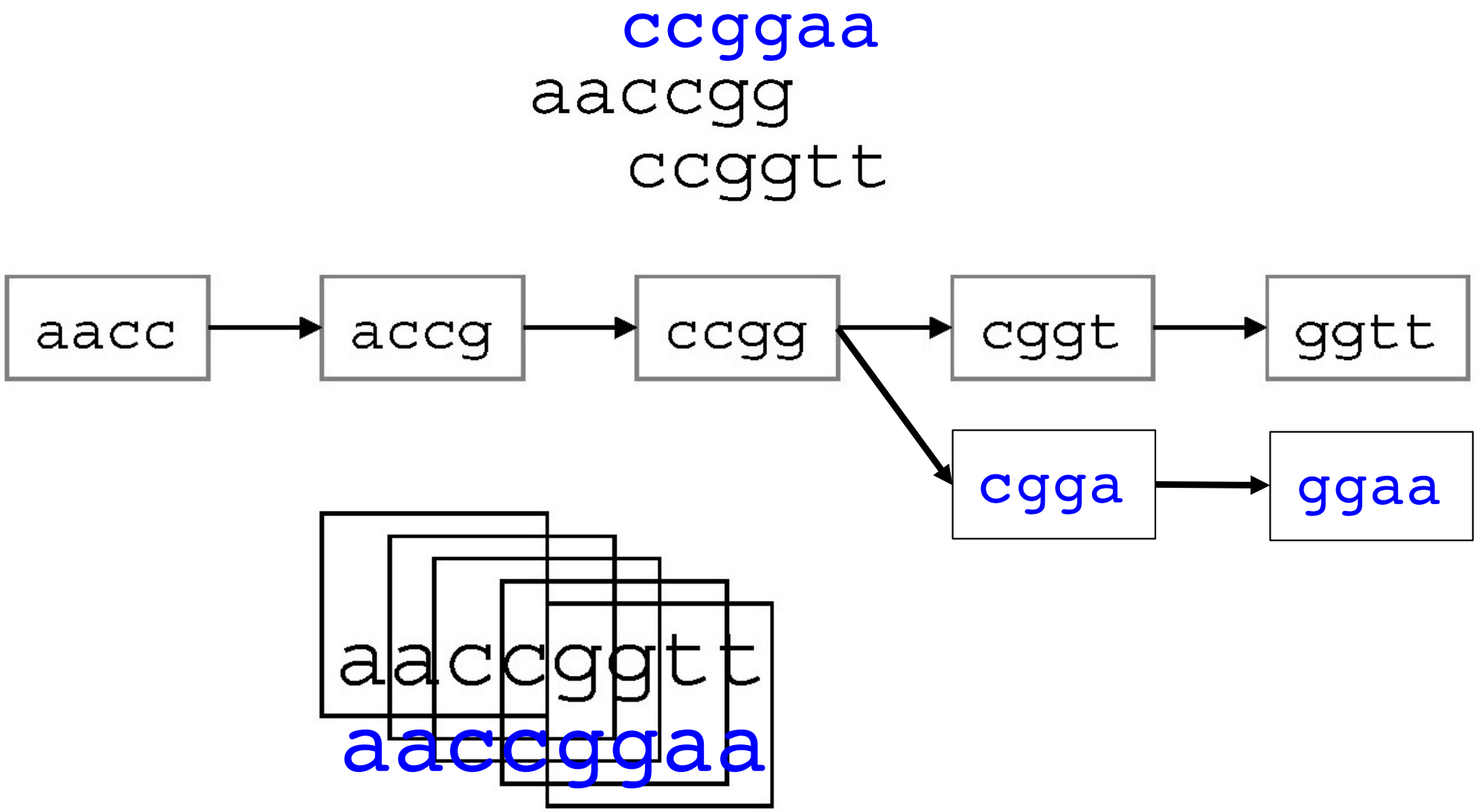
4. What Nth-generation assembly methods might improve our biological analyses?

Second-generation DNA sequencing/assembly



Ref.: Miller et al. (2010), Genomics *95*, 315-327.

Assembly gets harder when the data get messier



Ref.: Miller et al. (2010), Genomics *95*, 315-327.

All real animal genomes are "messy"

Nucleotide bias away from ~50% GC:

Illumina technology was developed with ~50% GC in mind (i.e., human %GC)
C. elegans is more AT-rich (35% GC), as are other *Caenorhabditis* spp.
(Some eukaryotes are even worse: *Plasmodium falciparum* is 19% GC!)
Illumina will systematically under-sequence regions <<50% or >>50% GC

Repetitive DNA elements:

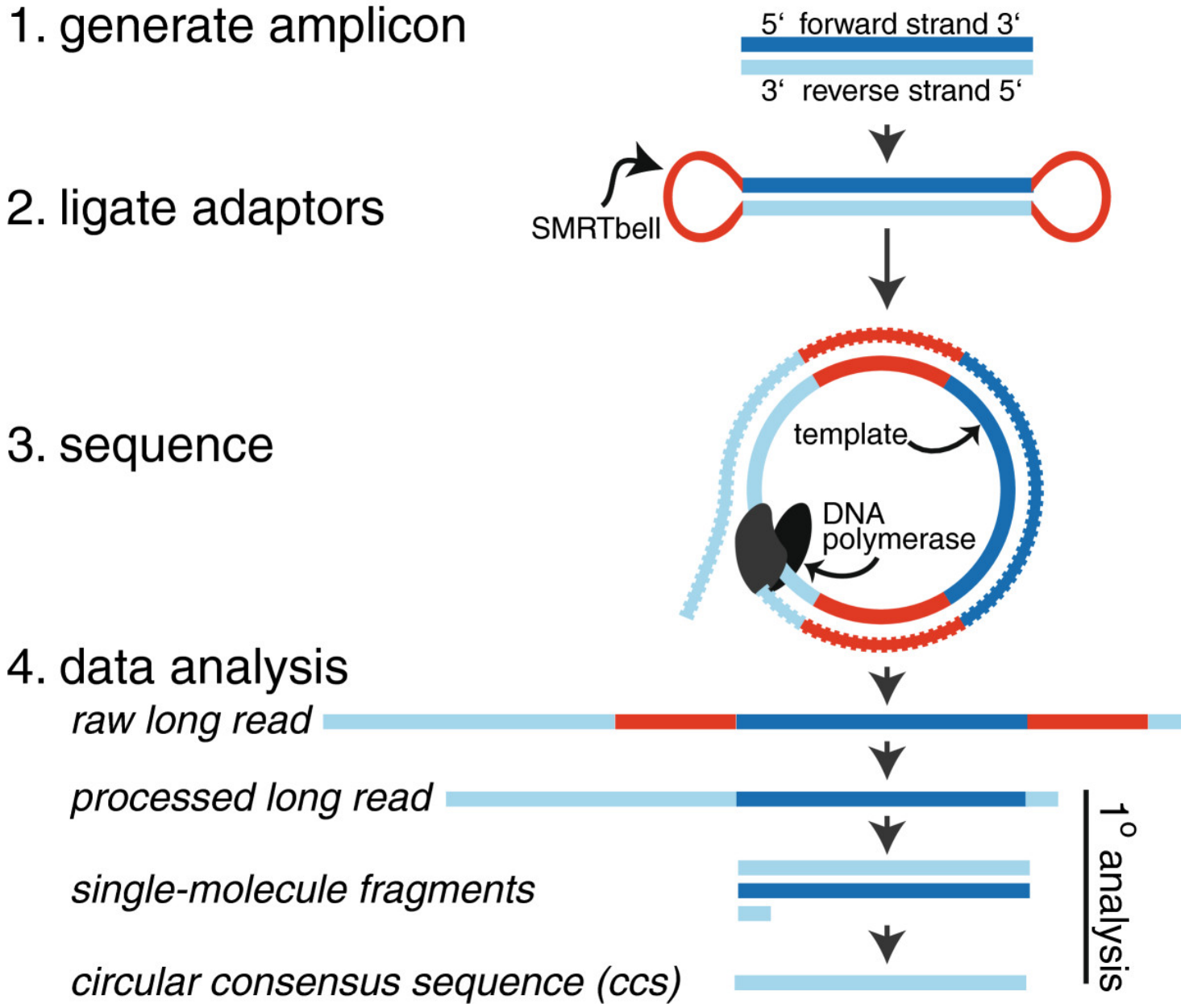
~17% in *C. elegans* genome; ~40% in *A. ceylanicum* (hookworm) genome
For 2cd-gen. genomes, must be jumped over; cannot be directly assembled

Tandem genomic repeats:

These exist in all metazoan genomes examined
They can encode important traits (e.g., species-specific adaptations)
They are systematically lost in Illumina-based sequencing/assembly

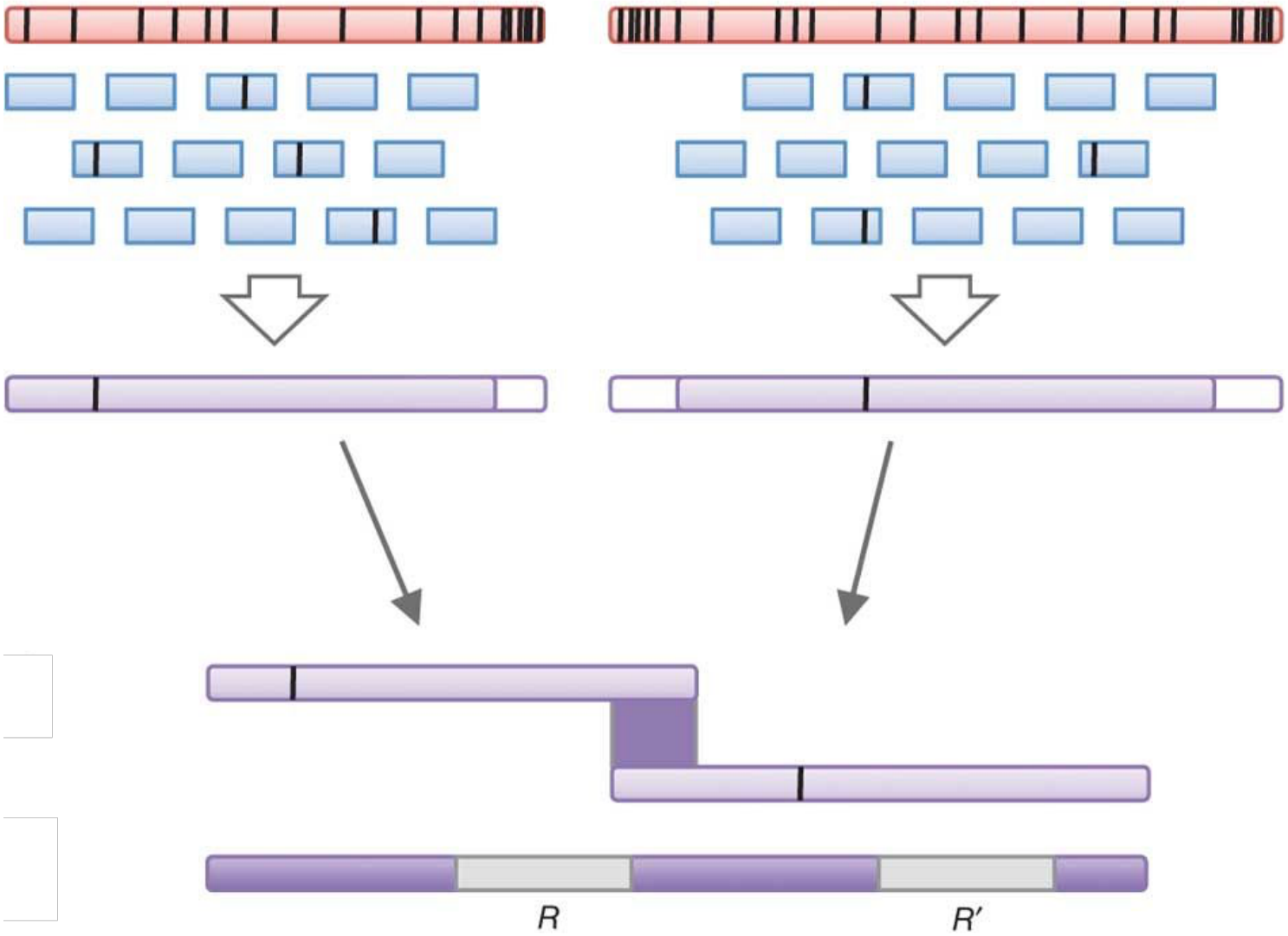
Refs.: Alkan et al. (2011), Nat. Methods *8*, 61-65; Kondrashov (2012), Proc. Biol. Sci. *279*, 5048-5057; Miller et al. (2010), Genomics *95*, 315-327; Qian and Zhang (2014), Genome Res. *24*, 1356-1362.

Solution: long "third-generation" sequence reads



Ref.: Fichot and Norman (2013), Microbiome *1*, 10.

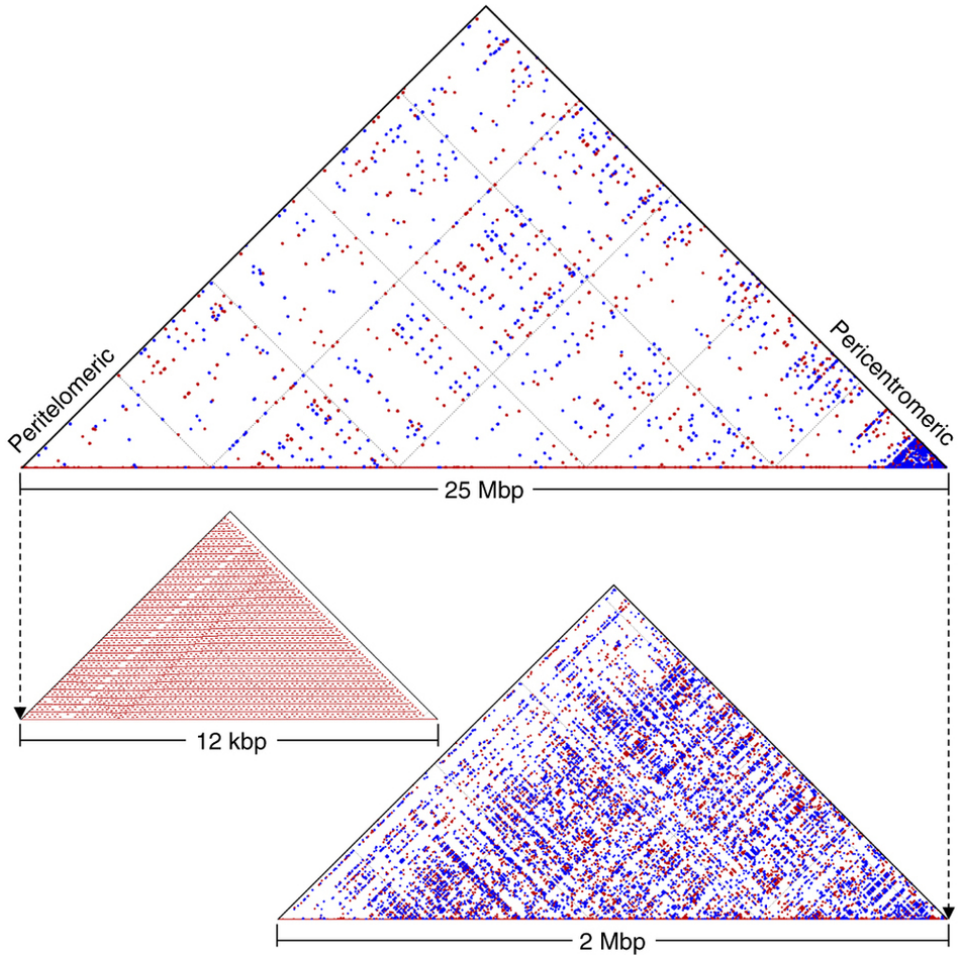
Long reads have a ~15% error rate; but these errors can be reduced to <0.1%



Ref.: Koren et al. (2012), Nat. Biotechnol. *30*, 693-700.

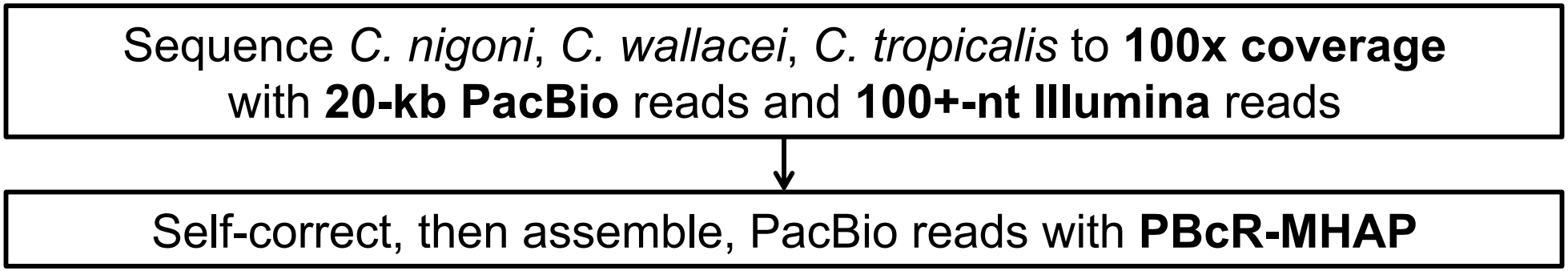
Self-corrected PacBio data (via PBcR-MHAP, Canu, etc.) allows genome assembly

~100x-coverage PacBio data gives 99.99% accuracy, and near-chromosomal blocks of contiguous sequence

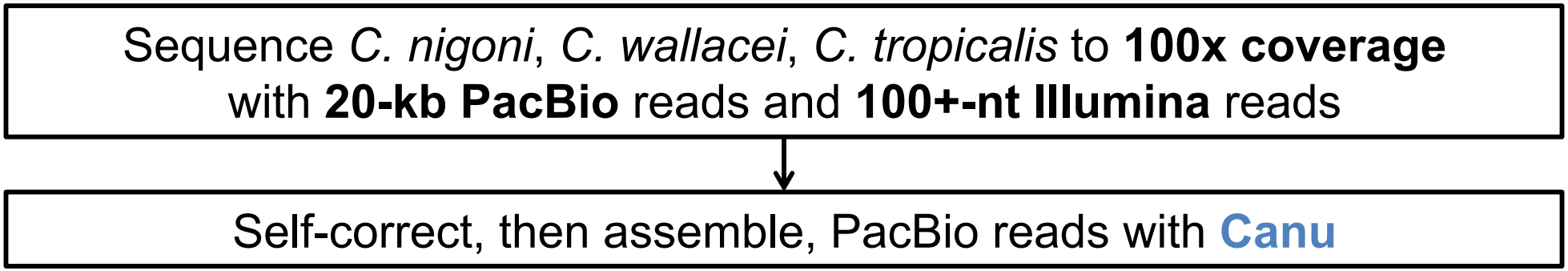


Alignment of assembled *Drosophila* chromosomal arm 3L. Ref.: Berlin et al. (2015), Nat. Biotechnol. *33*, 623-630.

A strategy for high-quality genomes

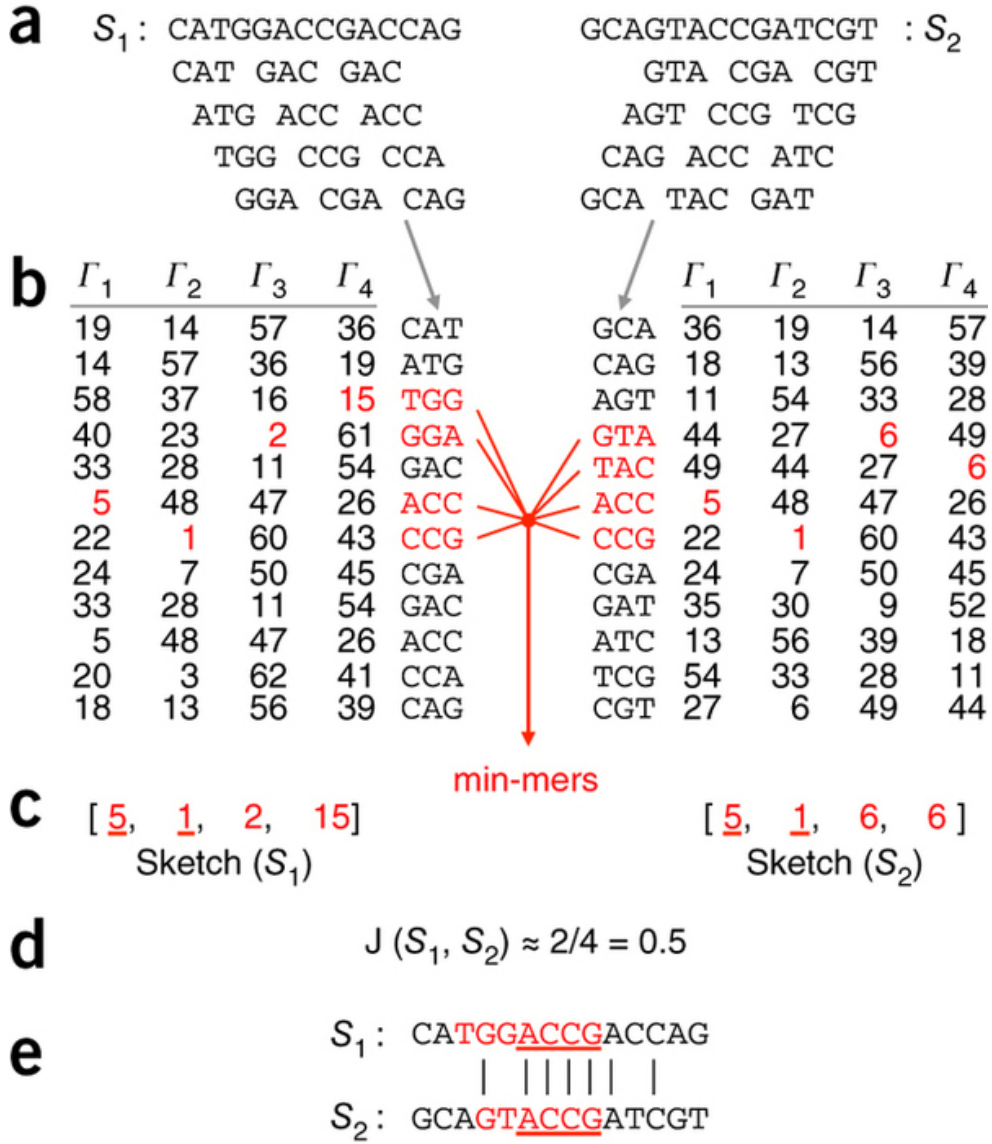


A strategy for high-quality genomes



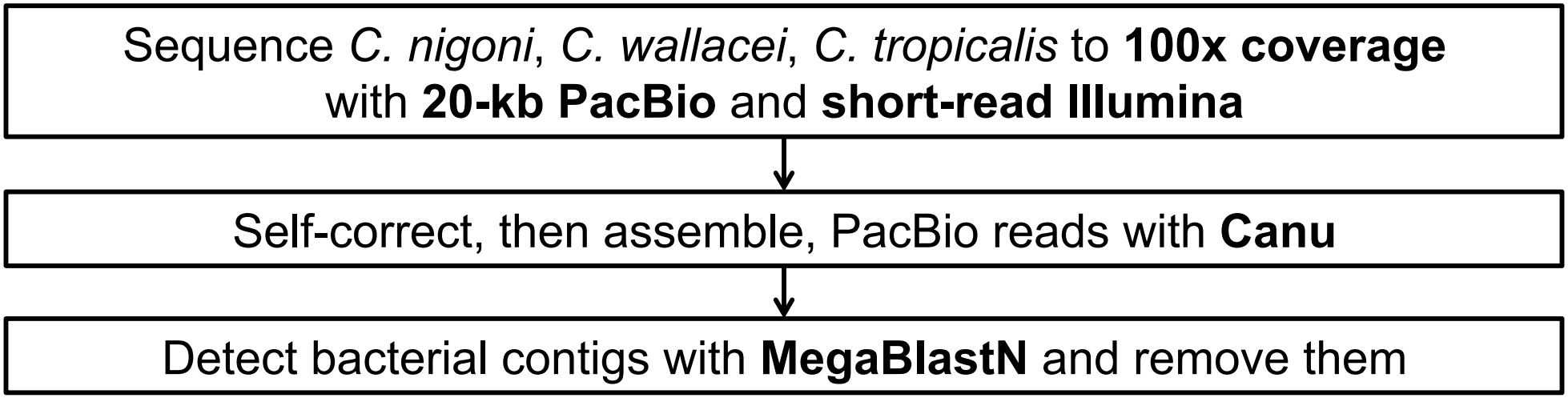
Refs.: Koren et al. (2017), Genome Res. *27*, 722-736; Eccles et al. (2018), BMC Biol. *16*, 6.

De novo genome assembly with self-corrected PacBio data (via PBcR-MHAP or Canu)

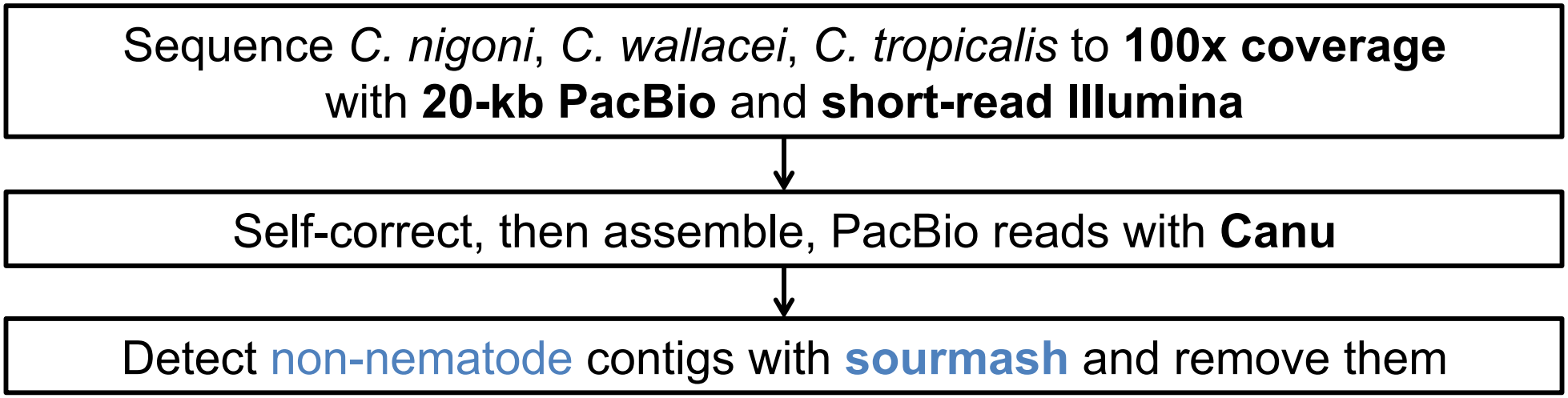


Ref.: Berlin et al. (2015), Nat. Biotechnol. *33*, 623-630.

A strategy for high-quality genomes

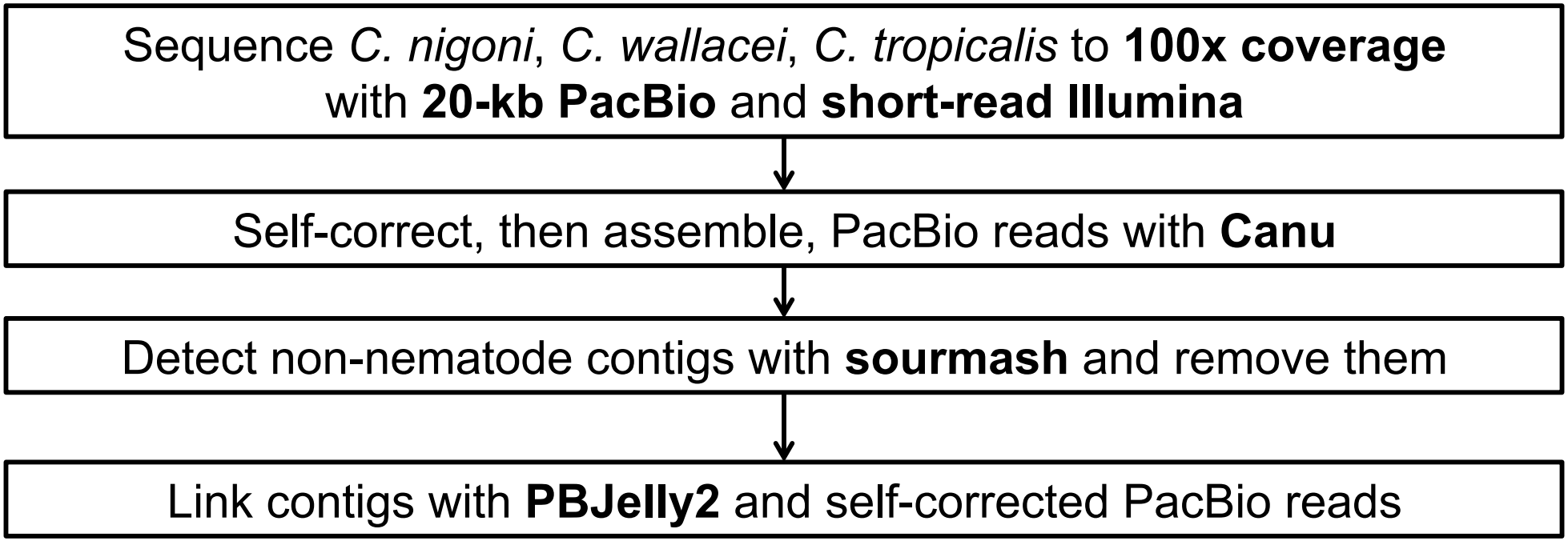


A strategy for high-quality genomes

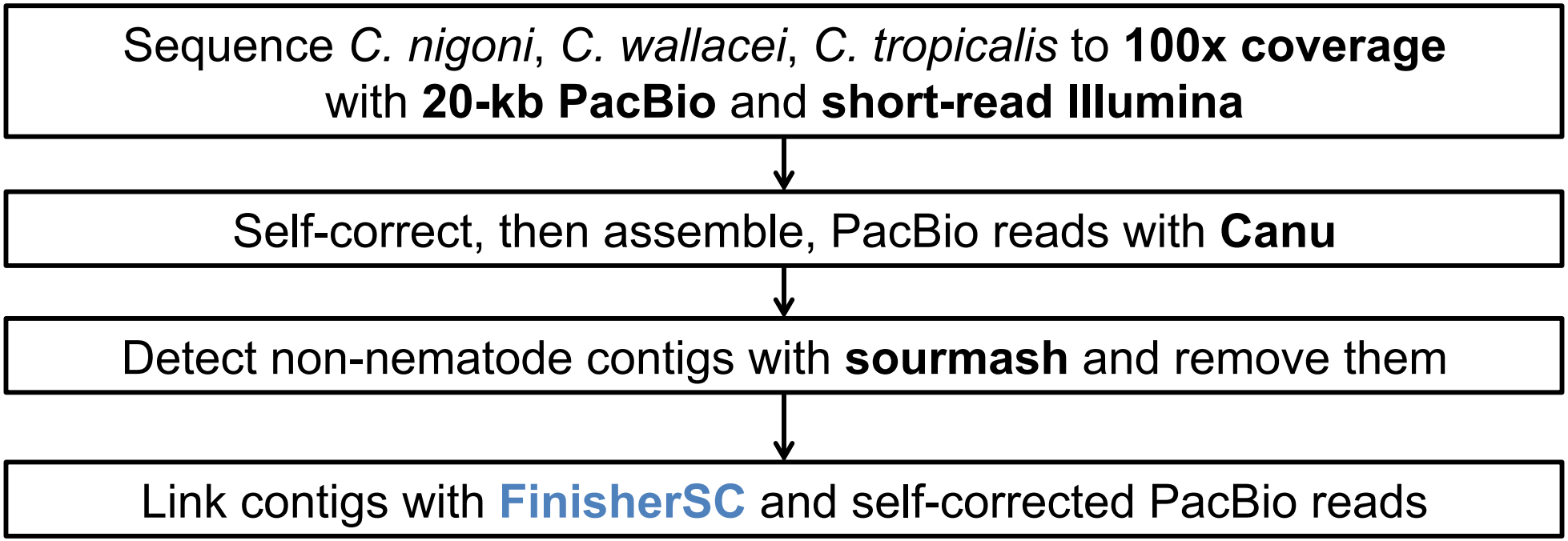


Ref.: *http://ivory.idyll.org/blog/2018-detecting-contamination-in-long-read-assemblies.html*

A strategy for high-quality genomes

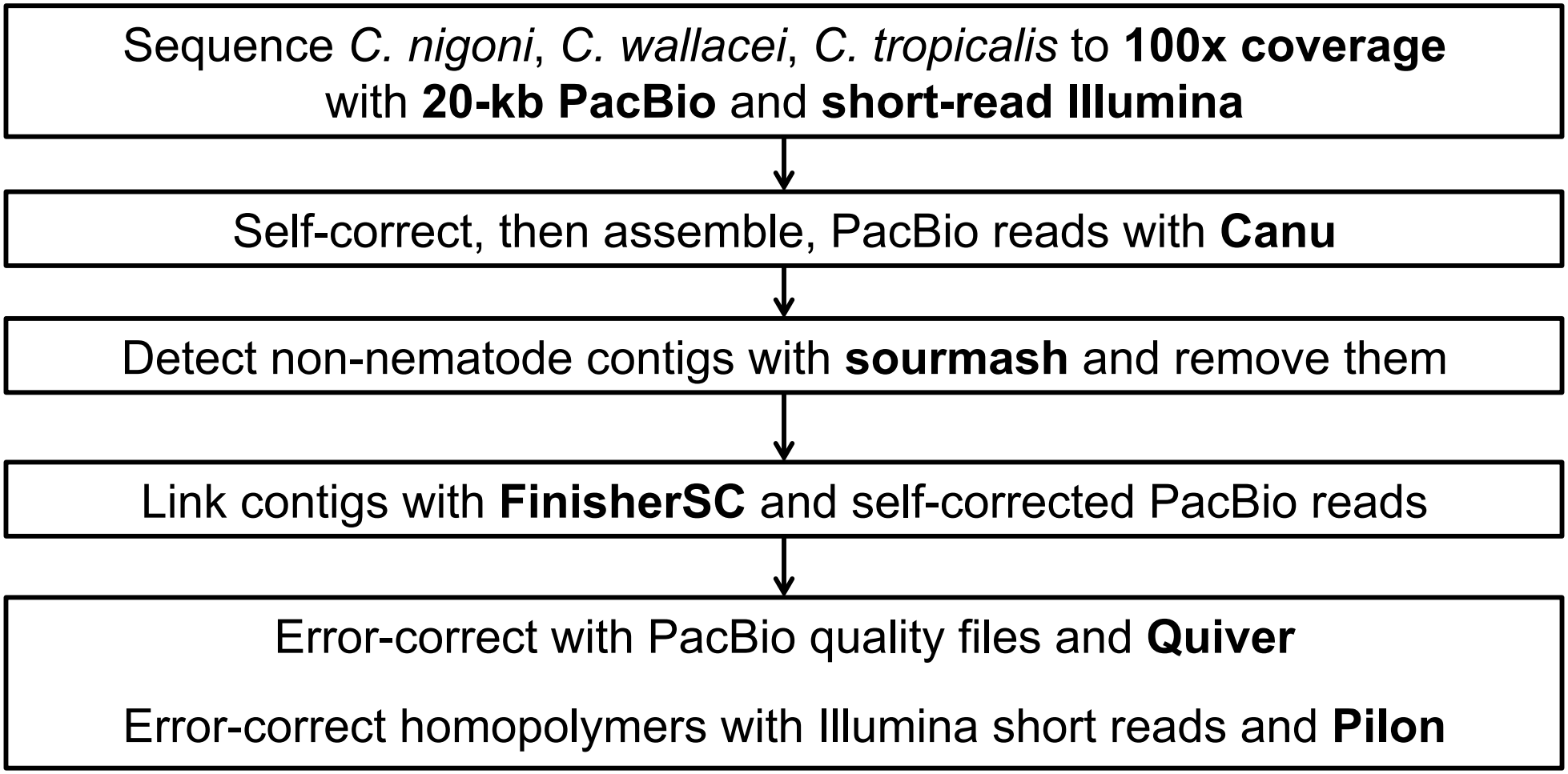


A strategy for high-quality genomes

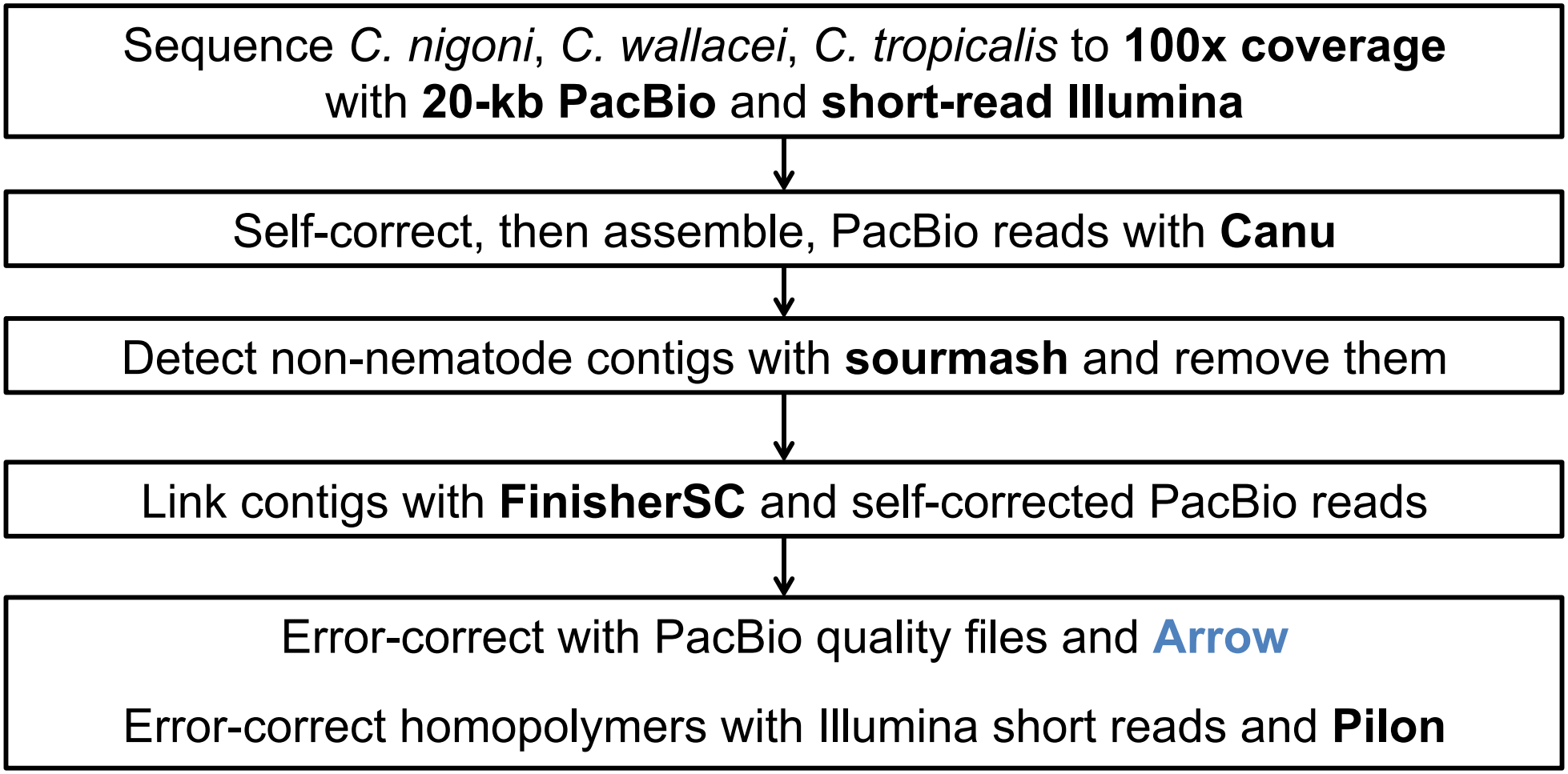


Ref.: Lam et al. (2015), Bioinformatics *31*, 3207-3209.

A strategy for high-quality genomes

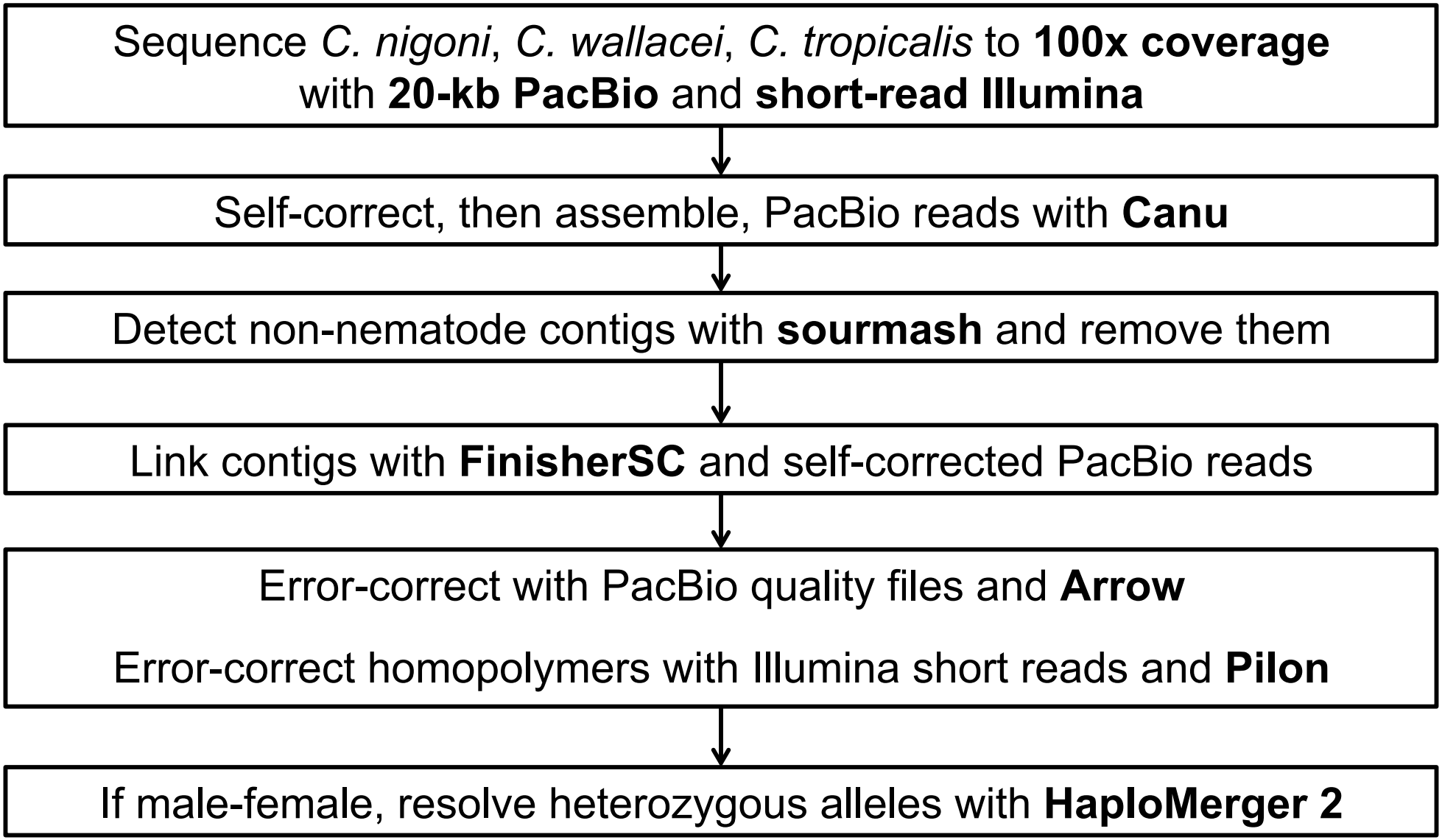


A strategy for high-quality genomes

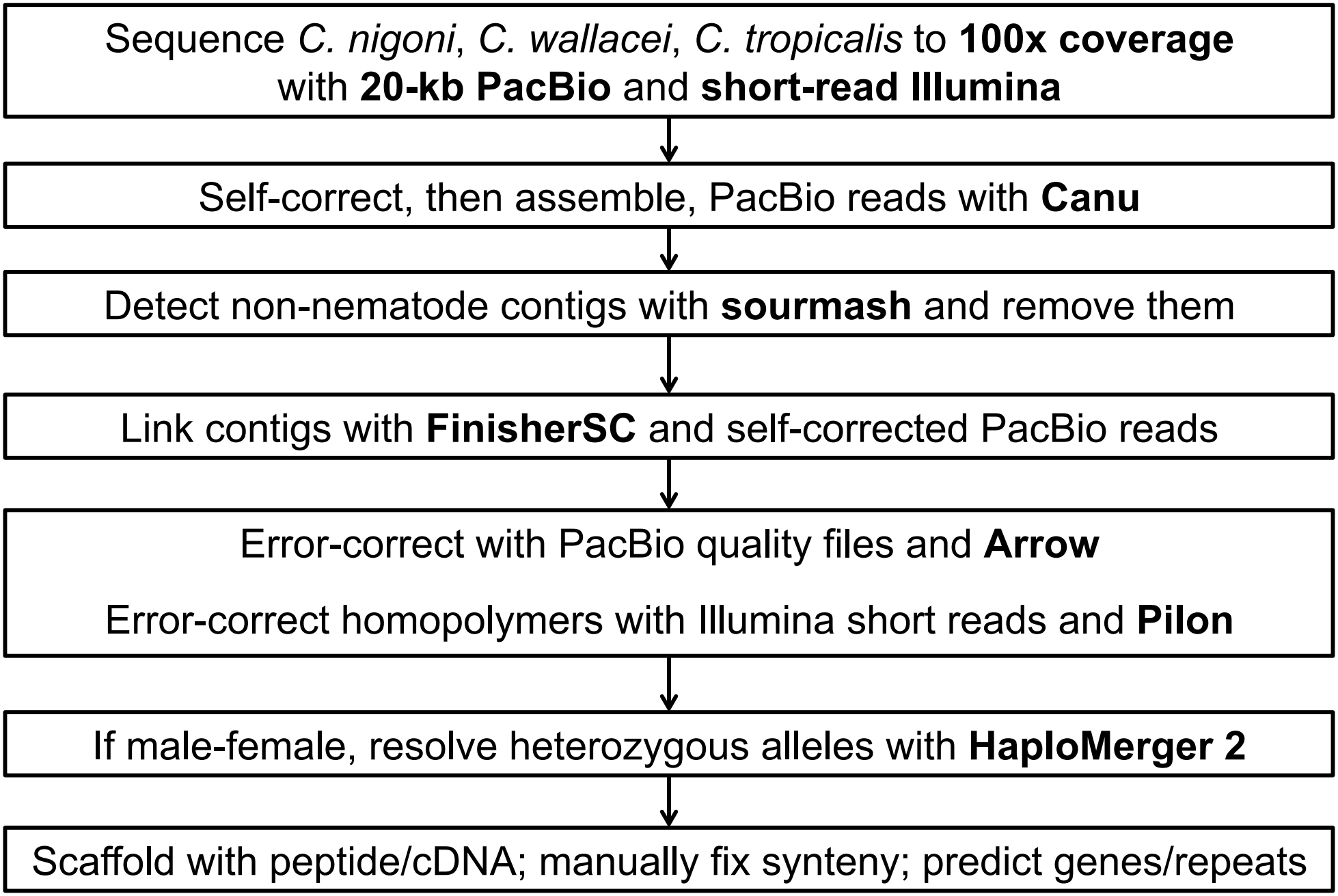


Ref.: *https://github.com/PacificBiosciences/GenomicConsensus*

A strategy for high-quality genomes



A strategy for high-quality genomes



For the full procedure*, read our paper's "exquisitely detailed"† Methods section

Yin et al. (2018), Science *359*, 55-61.

https://doi.org/10.1126/science.aao0827

https://www.ncbi.nlm.nih.gov/pubmed/29302007

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5789457

*But some methods that we used in 2015-2017 can be updated in 2018+.
This is an ongoing issue; genomics is an ever-moving target.

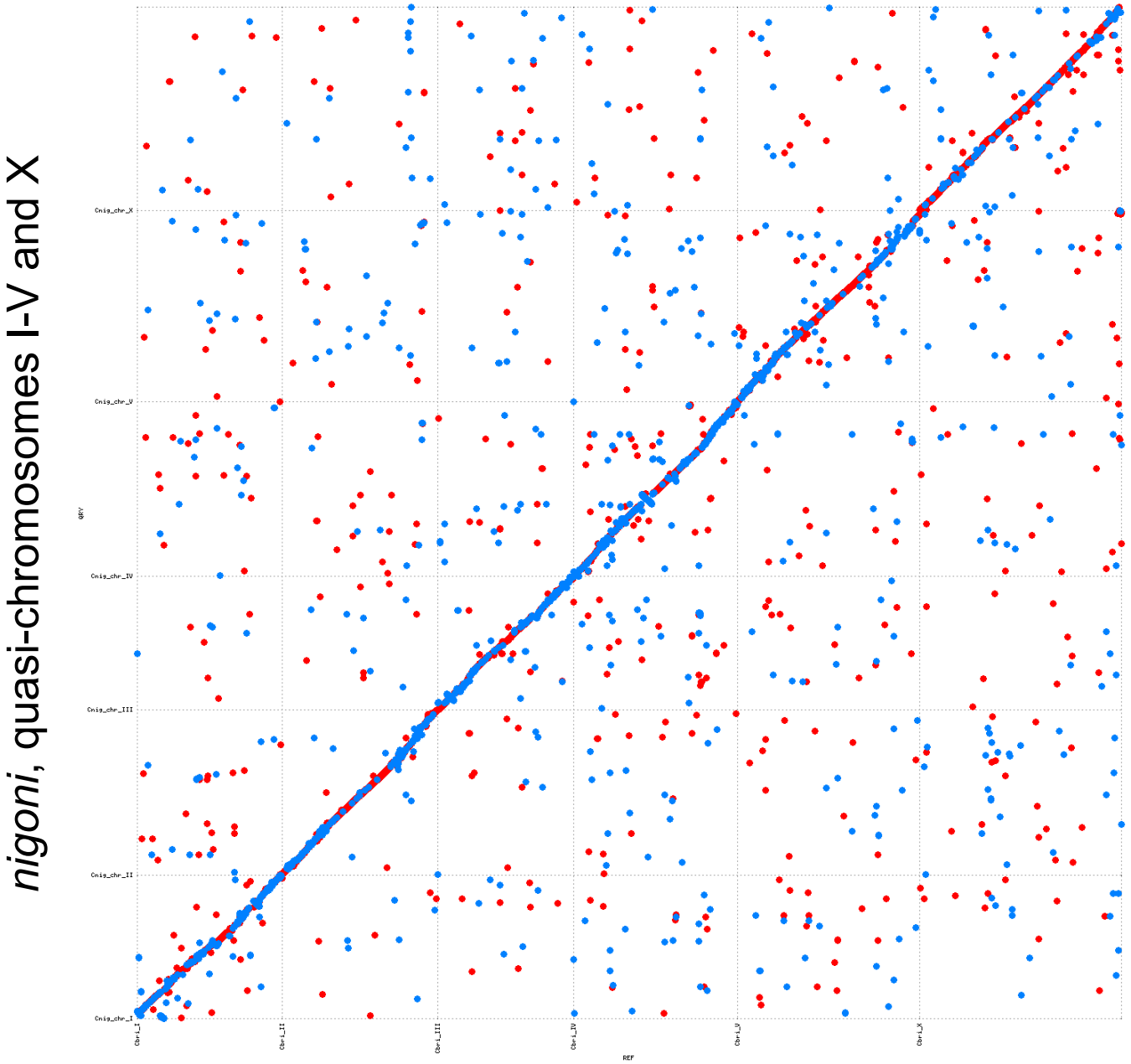
†From an anonymous reviewer:
"Although the methods for genome assembly are exquisitely detailed..."

C. nigoni's PacBio genome is very complete

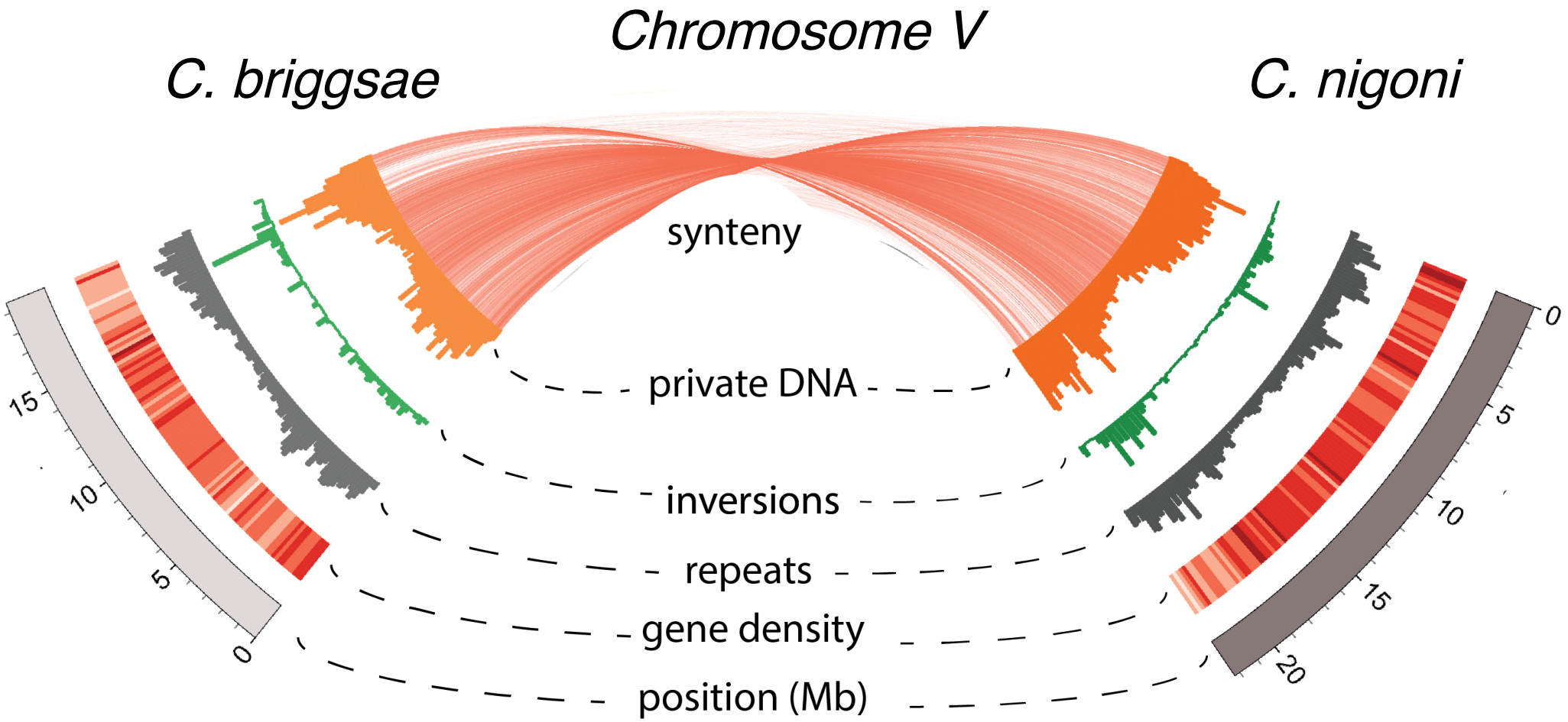
Statistics	***nigoni***	***briggsae***	***elegans***
Total nt	129,488,540	108,384,165	100,286,401
PacBio cov.	96x	n/a	n/a
Contigs	213	6,724	7
Contig N50, nt	3,254,670	41,490	17,493,829
Contig max. nt	9,436,569	516,571	20,924,180
Repetitive	27.3%	26.0%	22.0%
Genes (prot.)	29,167	22,313*	20,257
CEGMA comp.	99.6% (247/248)	99.6% (247/248)	98.4% (244/248)
CEGMA ratio	1.19	1.13	1.12

**C. briggsae* genes were recomputed by the methods used for *C. nigoni*. The official *C. briggsae* gene number is even smaller (21,814).

nigoni aligns cleanly with *briggsae*



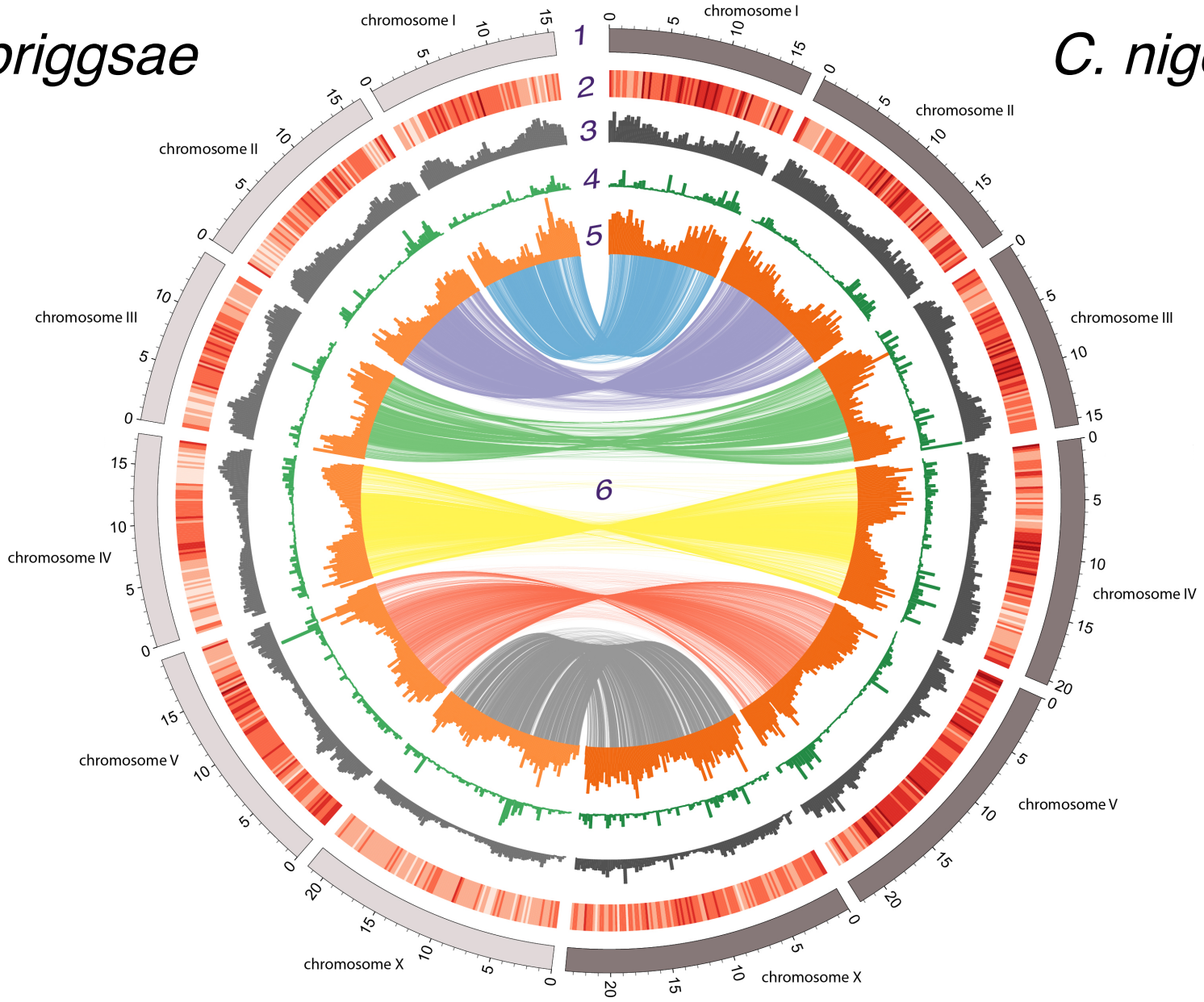
The *C. nigoni* and *C. briggsae* genomes are highly syntenic



The *C. nigoni* and *C. briggsae* genomes are highly syntenic

C. briggsae

C. nigoni



Summing up: third-generation genome assembly

Third-generation genome assembly
allows us to assemble genomes
that come very near perfection.

This allows us to compare the genomes
of related animal species
(such as *C. nigoni* versus *C. briggsae*)
with unprecedented accuracy.

Questions

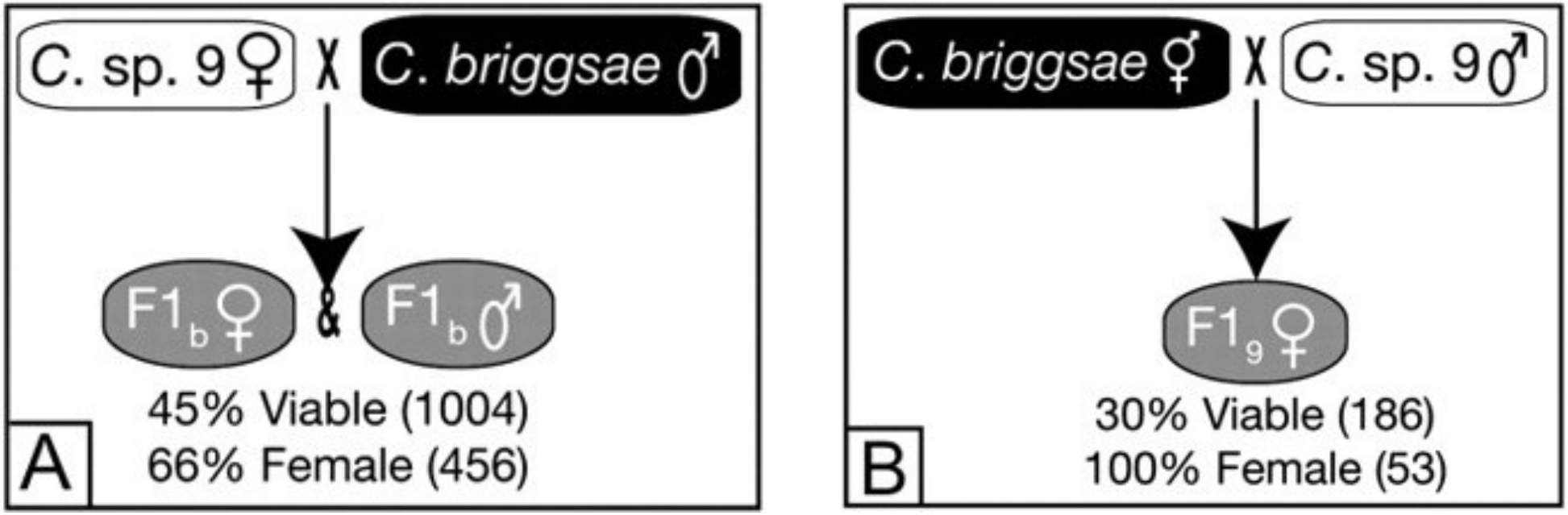
1. Why are nematode genomes interesting?

2. Why is long-read (third-generation) genome assembly a good idea?

3. What parts of a nematode genome are needed for male-female sexuality (versus hermaphroditism)?

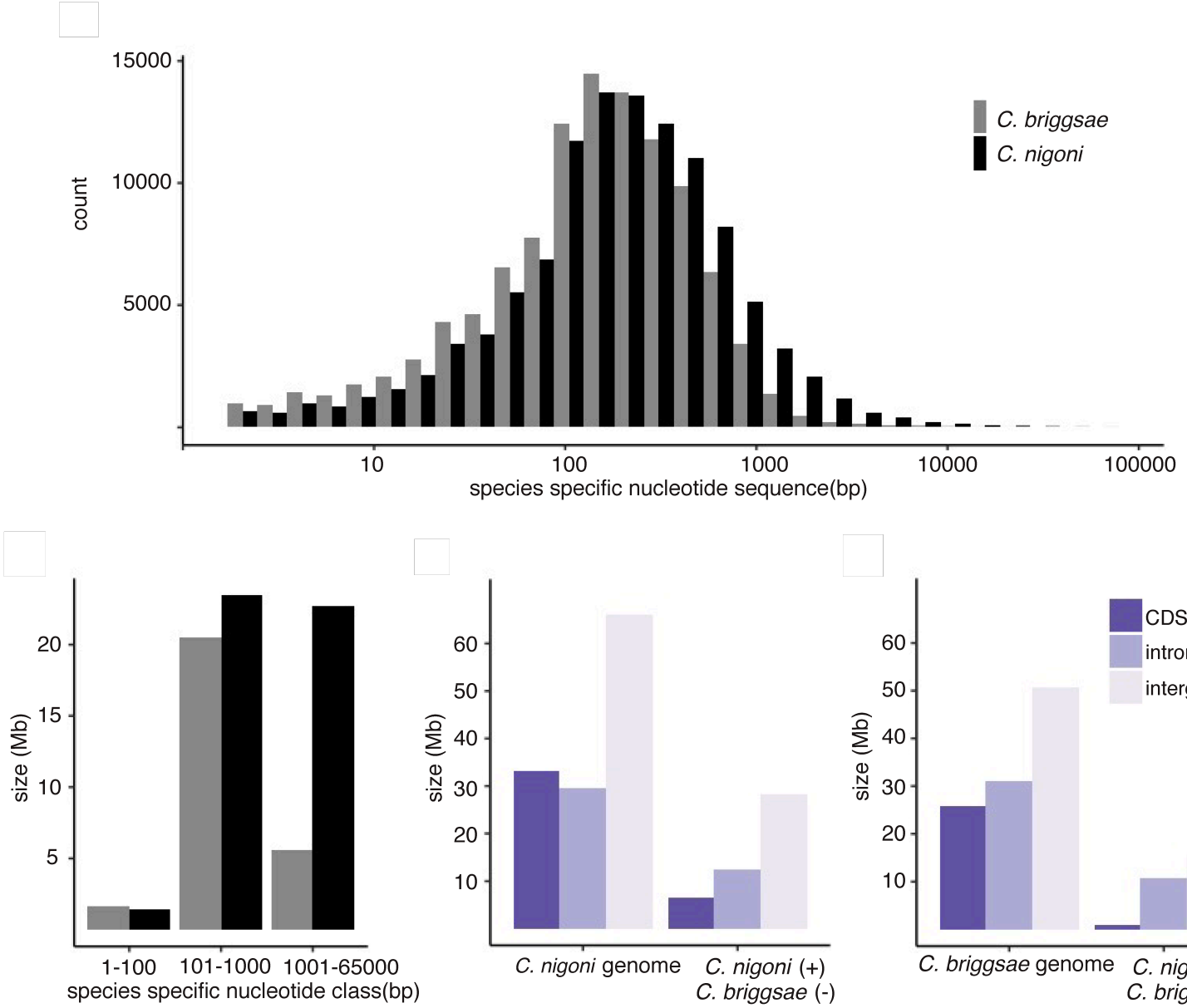
4. What Nth-generation assembly methods might improve our biological analyses?

What can we find in the *C. nigoni* genome that might explain differences from *C. briggsae*?

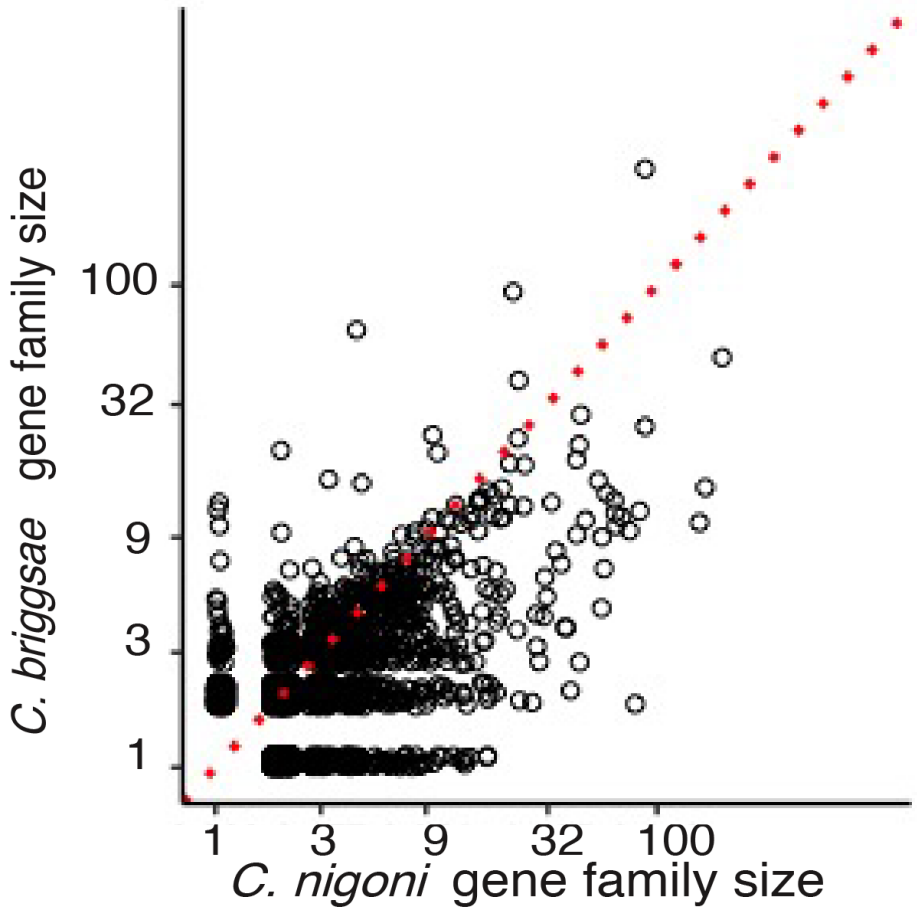


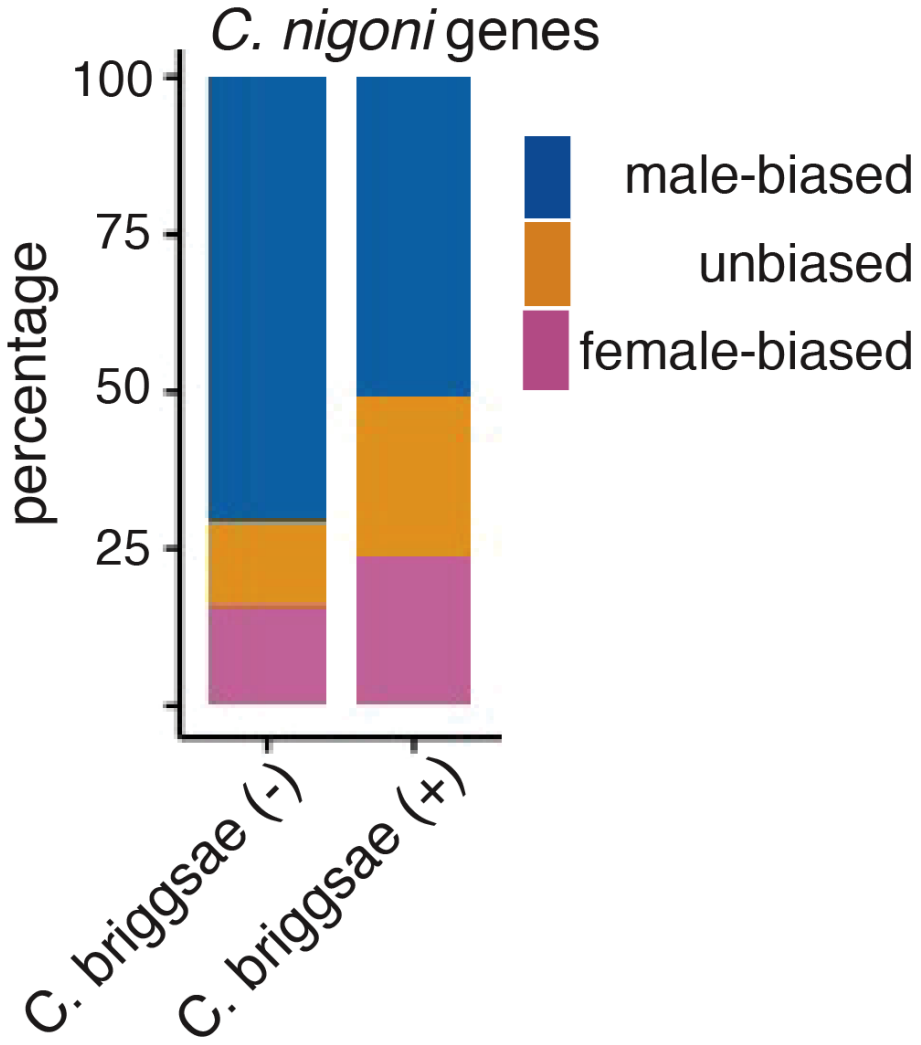
Ref.: Woodruff et al. (2010), Genetics *186*, 997-1012.

Many sequences lost in *C. briggsae* are ≤500 nt, but larger protein-coding genes are lost as well

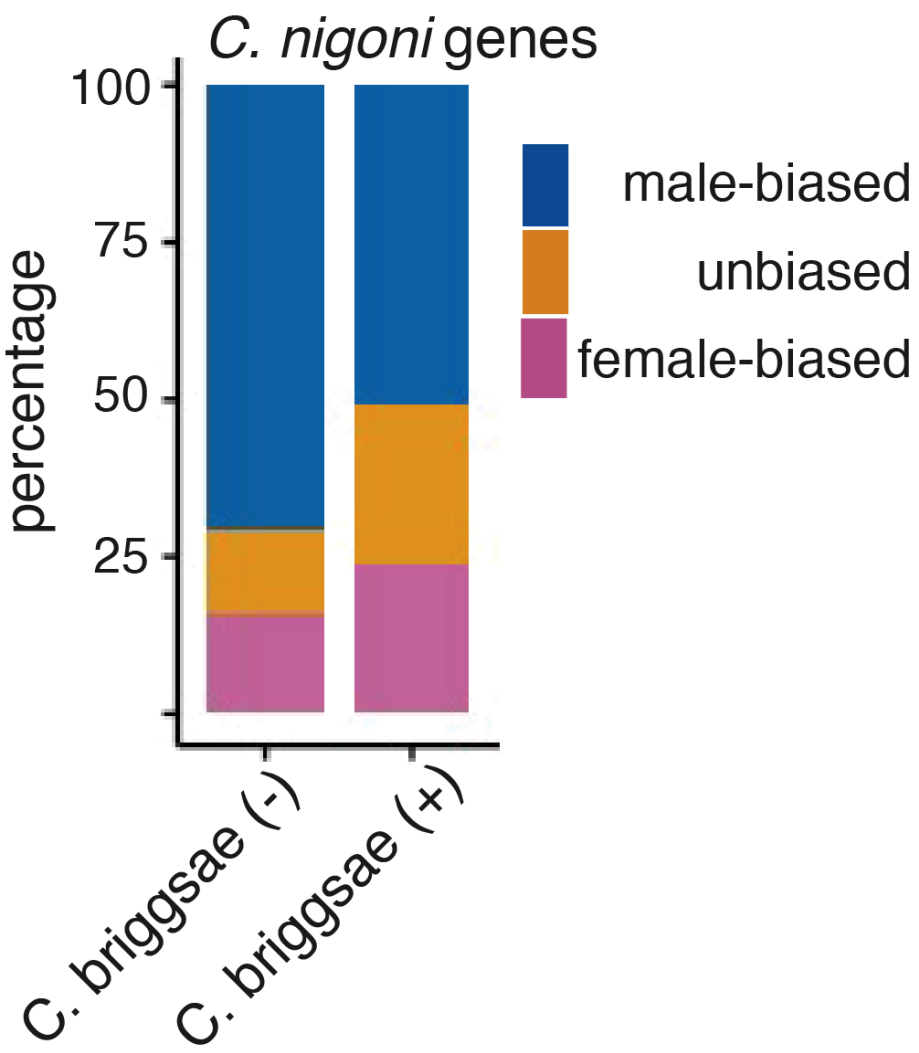


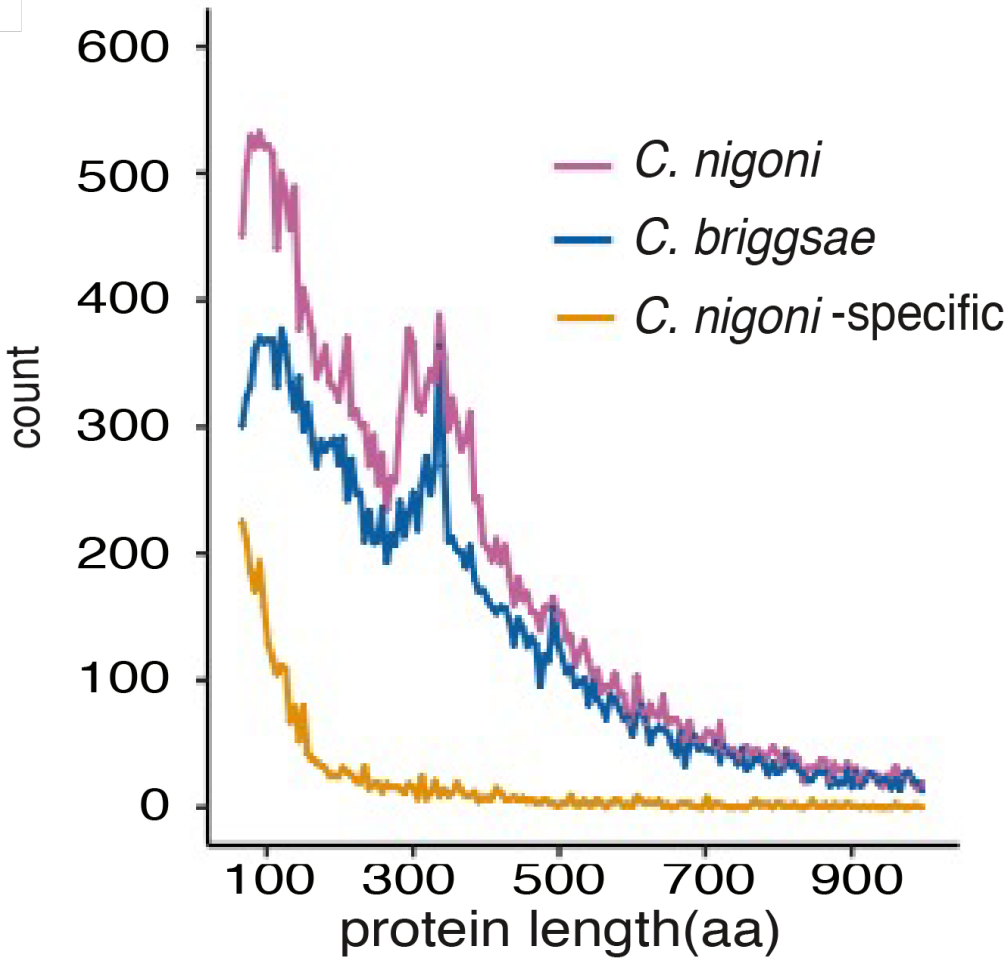
nigoni has more genes per family than *briggsae*, and male-biased genes lack *briggsae* homologs



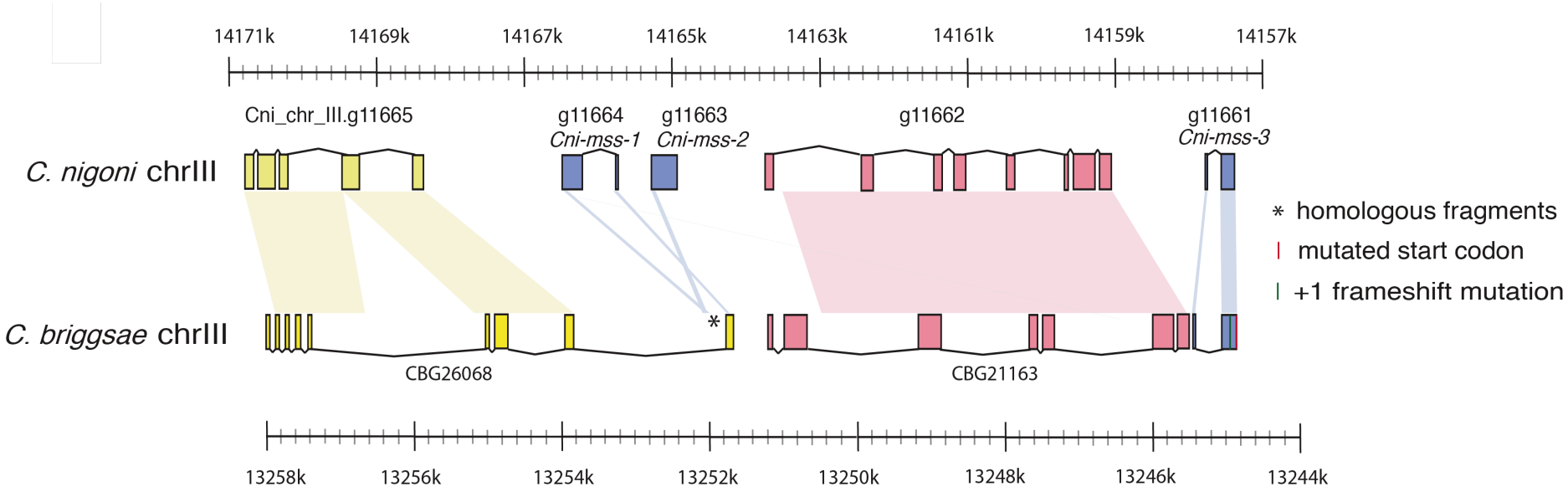


Genes lost in *briggsae* tend to be not only male-biased, but also small





A family of small *male secreted short* (*mss*) genes are consistently lost from *briggsae*

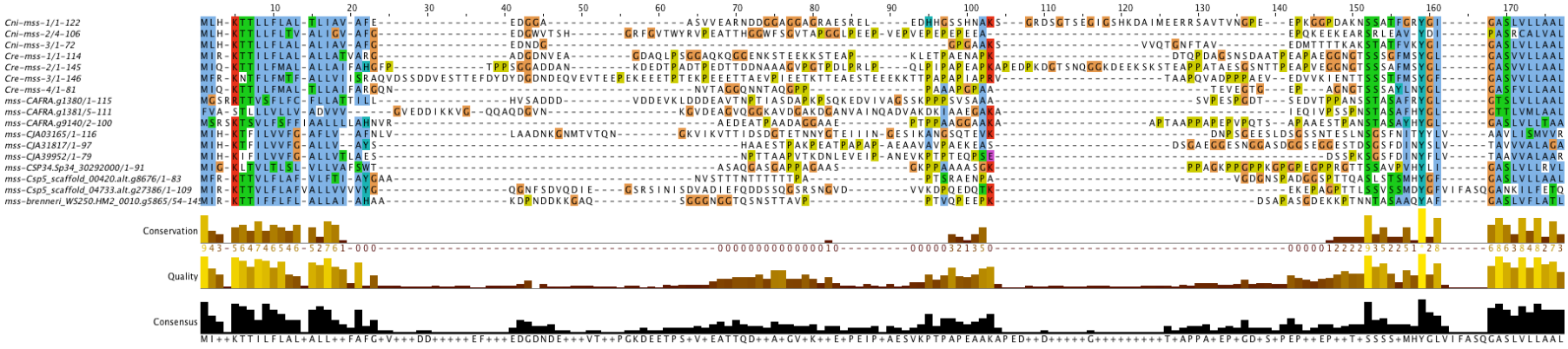


Conserved signal sequence

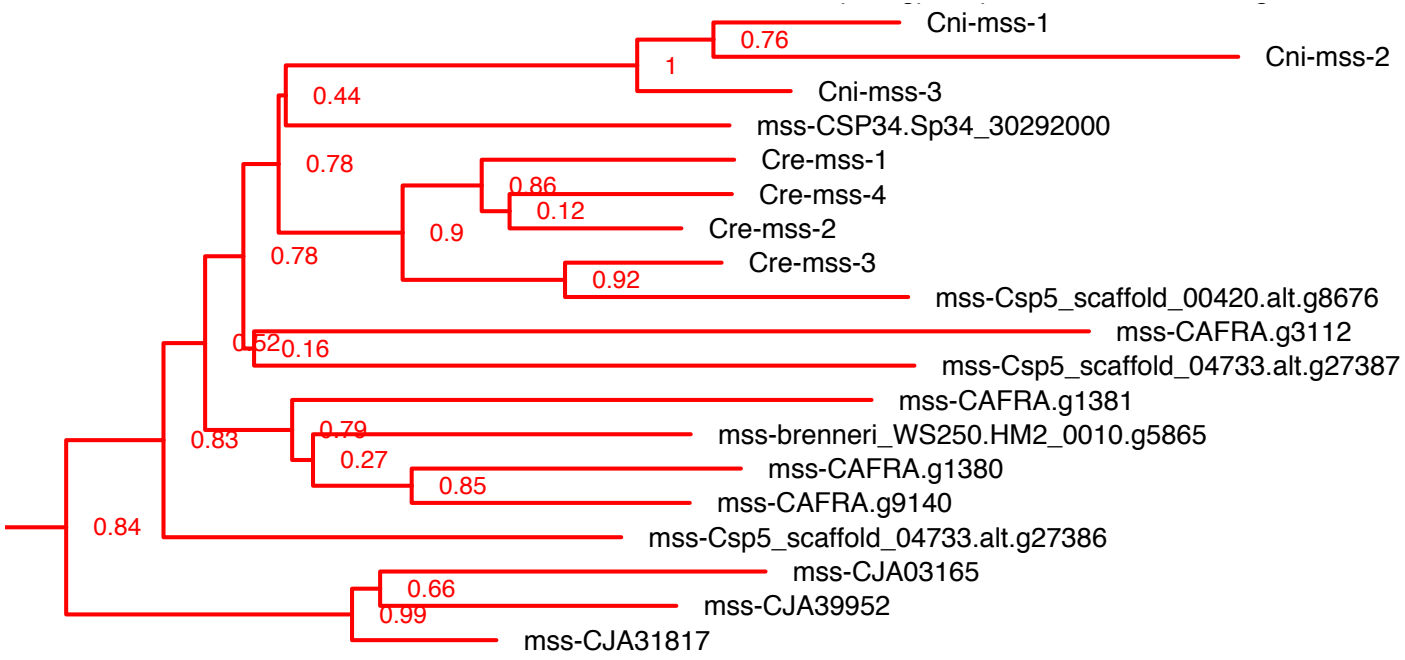
Hypervariable and O-glycosylated region

GPI-anchoring region

Conserved C-terminus



mss genes form a coherent family with genes only in male-female *Caenorhabditis*



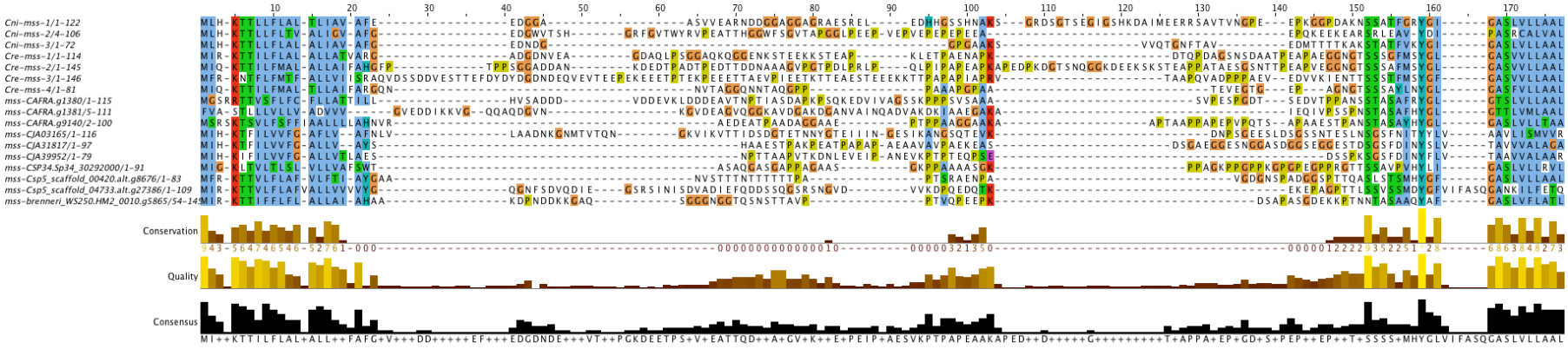
mss

Conserved signal sequence

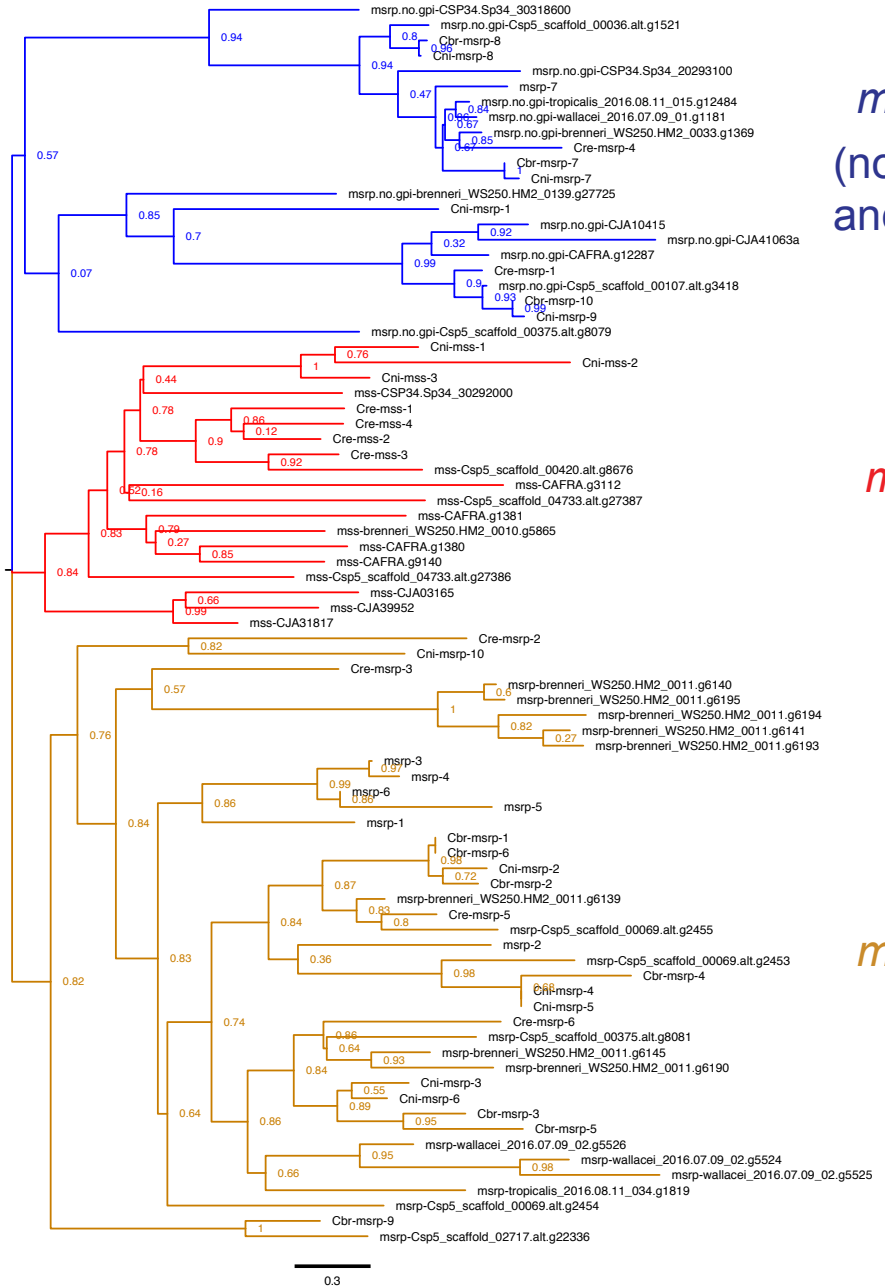
Hypervariable and O-glycosylated region

GPI-anchoring region

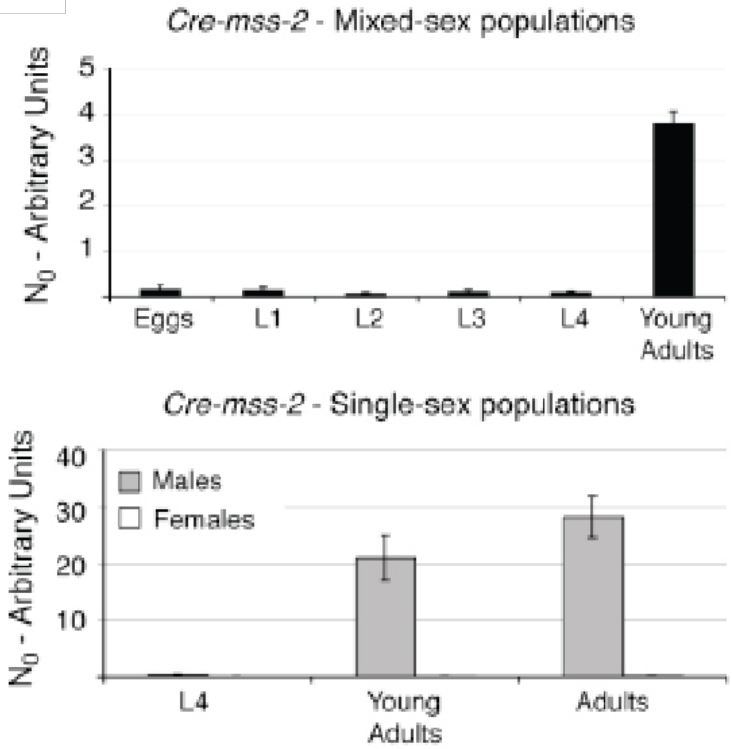
Conserved C-terminus

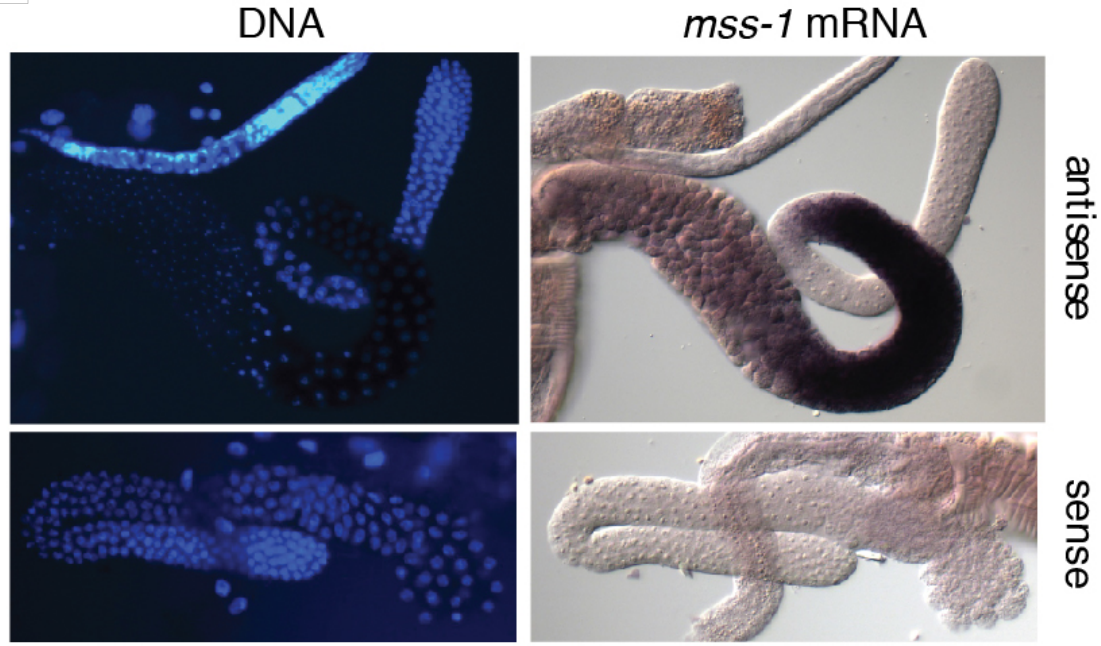


mss belongs to a larger gene superfamily with *msrp* paralogs that are in *C. briggsae*, etc.



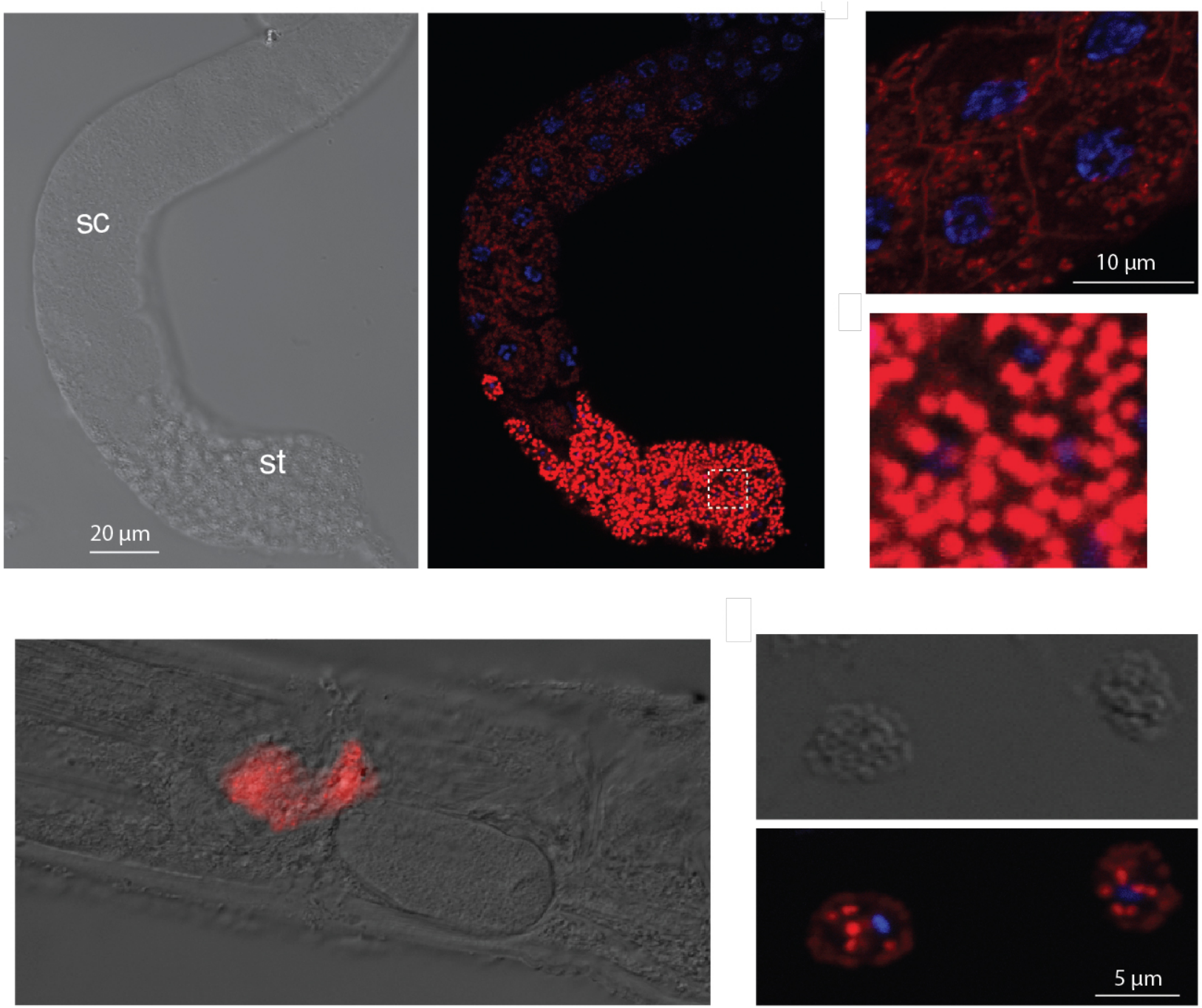
C. remanei mss genes are expressed in pachytene-stage primary spermatocytes; other *mss* genes have male-biased expression



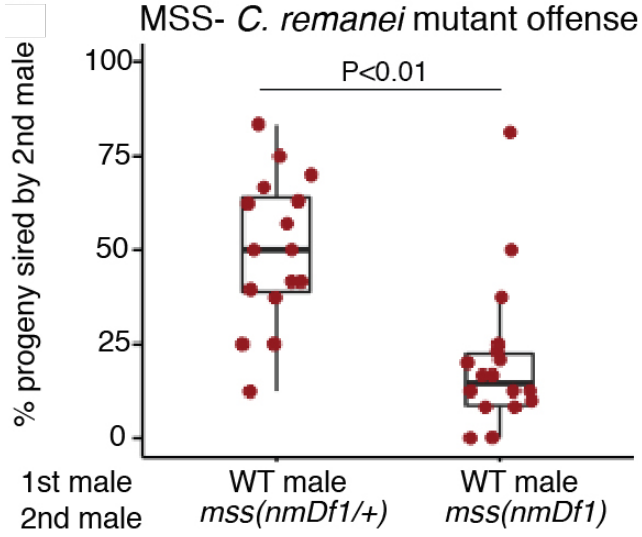


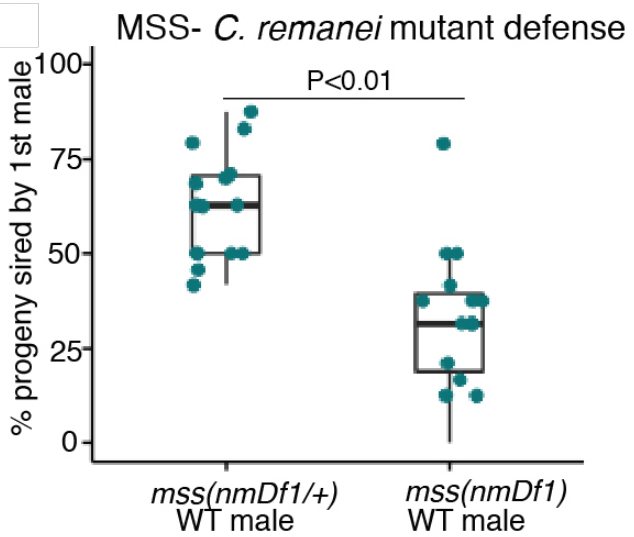
C. brenneri: 1 *mss* gene is 500-fold up in males versus females
C. nigoni: 1/3 *mss* genes are 60-fold up in males versus females
C. remanei: 4/4 *mss* genes are 2,000-fold up in males versus females

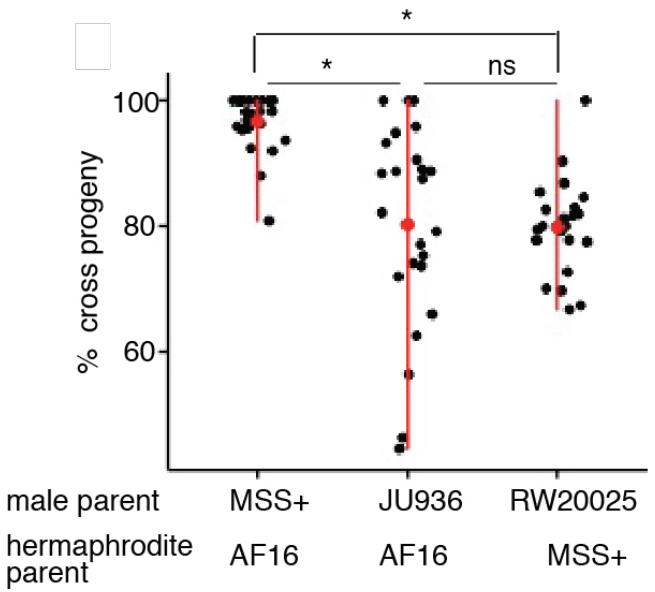
C. remanei MSS is a sperm cell-surface protein

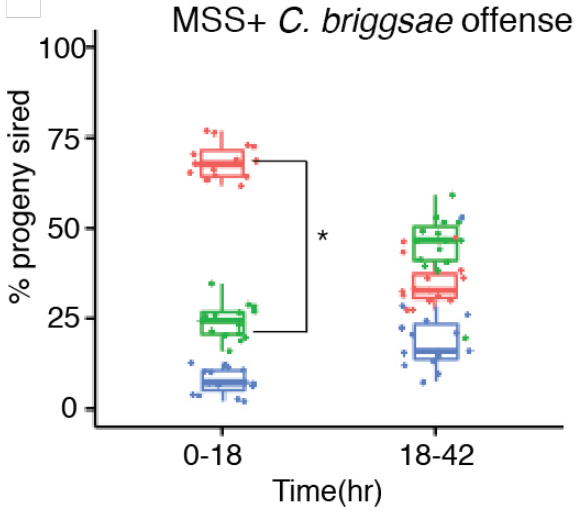


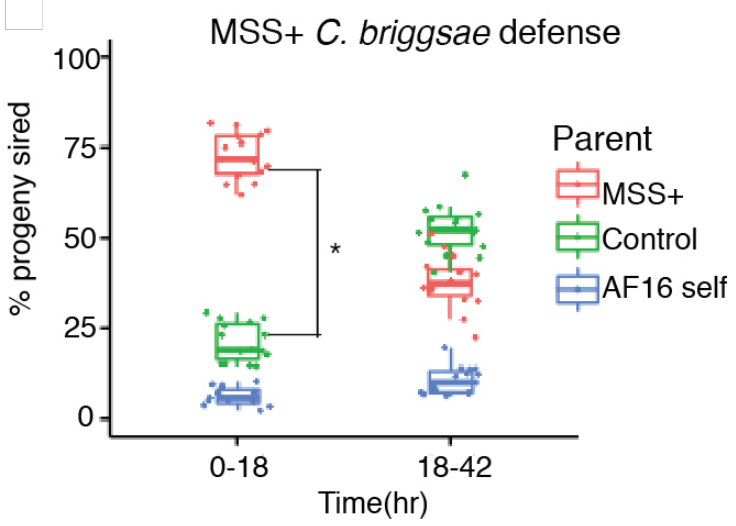
mss genes make males reproductively competitive

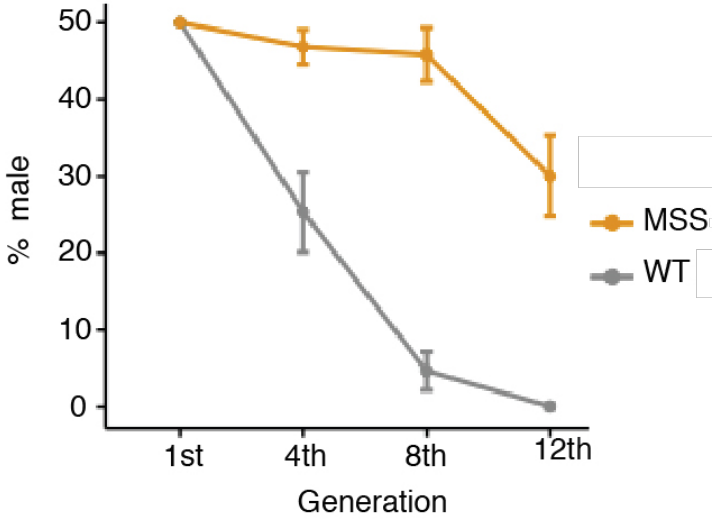












But, *why* does restoring *mss* to *C. briggsae* work?

If you stop and think about it, that experiment should not work.

C. briggsae is thought to have diverged from *C. nigoni* ~1,000,000 years ago, and to have become a hermaphrodite ~100,000 years ago.

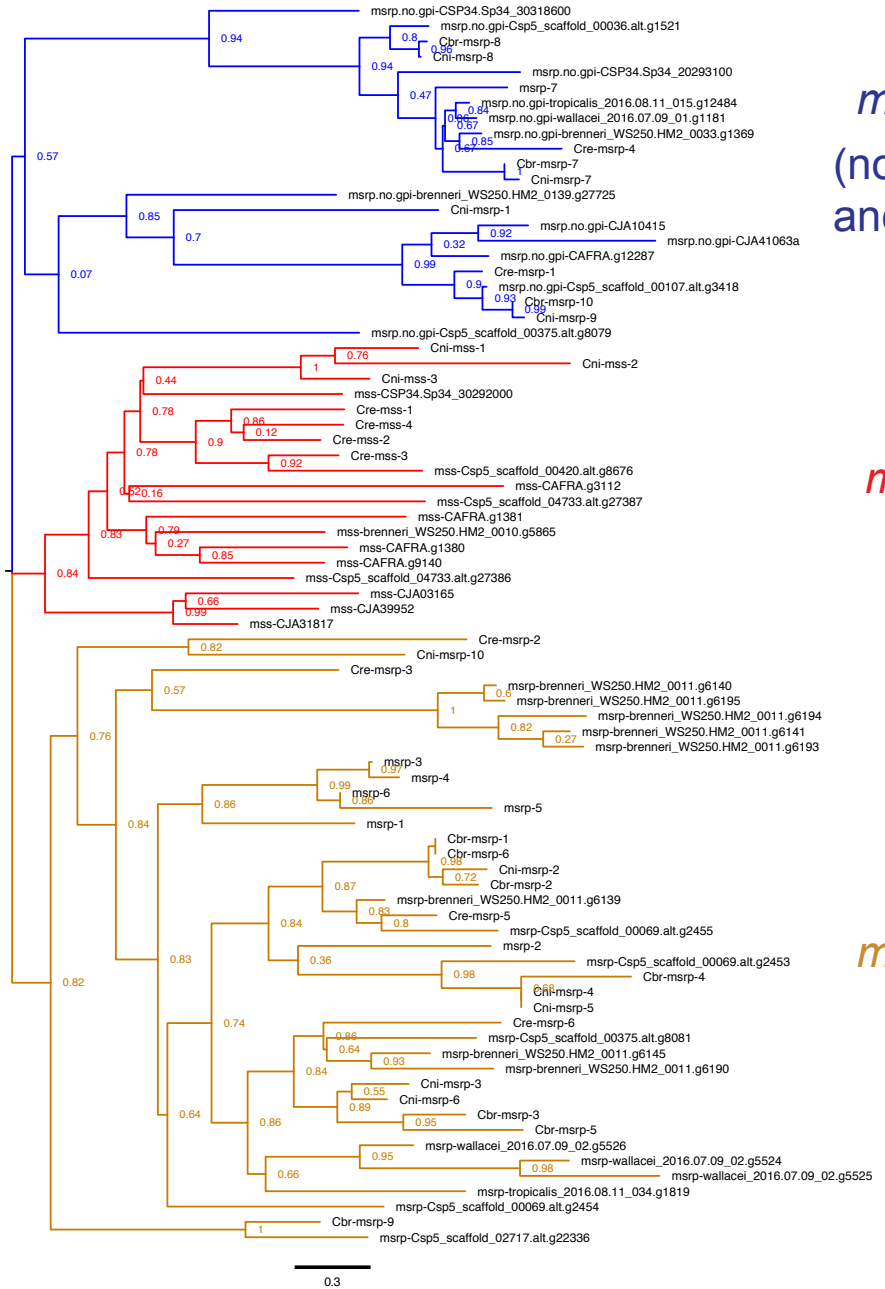
Since *mss* is absent from 11 different wild isolates of *C. briggsae*, it seems likely to have lost all of its *mss* genes around the time that it became hermaphroditic.

So, we are putting a gene into *C. briggsae* that it has not seen for 0.1 million years. Yet that gene elicited a powerful response.

What is going on?

Ref.: Cutter et al. (2010), Mol. Ecol. *19*, 798-809..

Hypothesis: MSS proteins act on the same (still unknown) target as MSRP proteins



Summing up: male-female genomics

C. nigoni versus *C. briggsae*
is almost a laboratory experiment
in the genomic basis of male-female
versus hermaphroditic sexual reproduction.

C. briggsae has smaller gene families
and fewer male-expressed genes
that encode small proteins.

mss is a strikingly powerful instance of
such a lost gene.

Questions

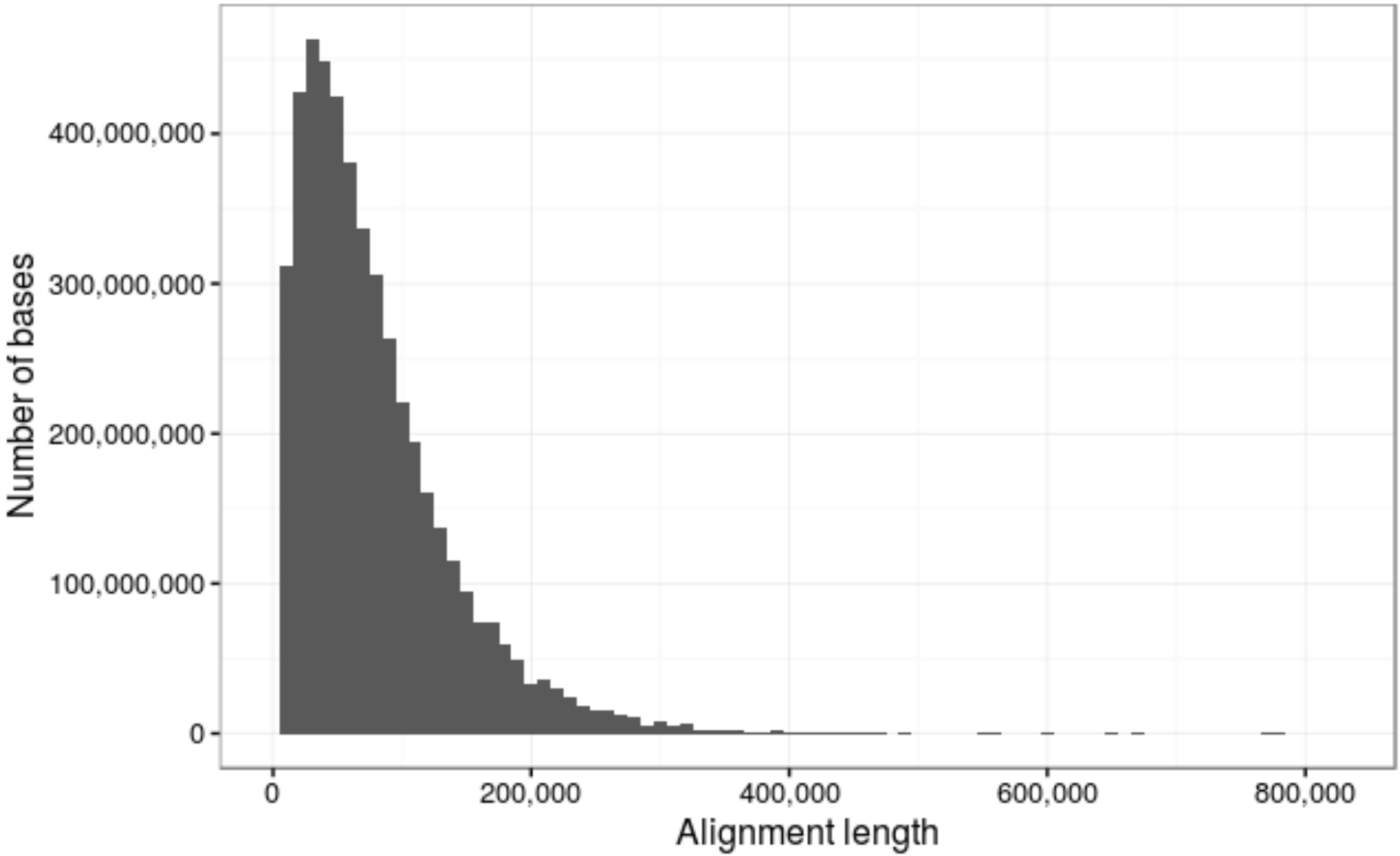
1. Why are nematode genomes interesting?

2. Why is long-read (third-generation) genome assembly a good idea?

3. What parts of a nematode genome are needed for male-female sexuality (versus hermaphroditism)?

4. What Nth-generation assembly methods might improve our biological analyses?

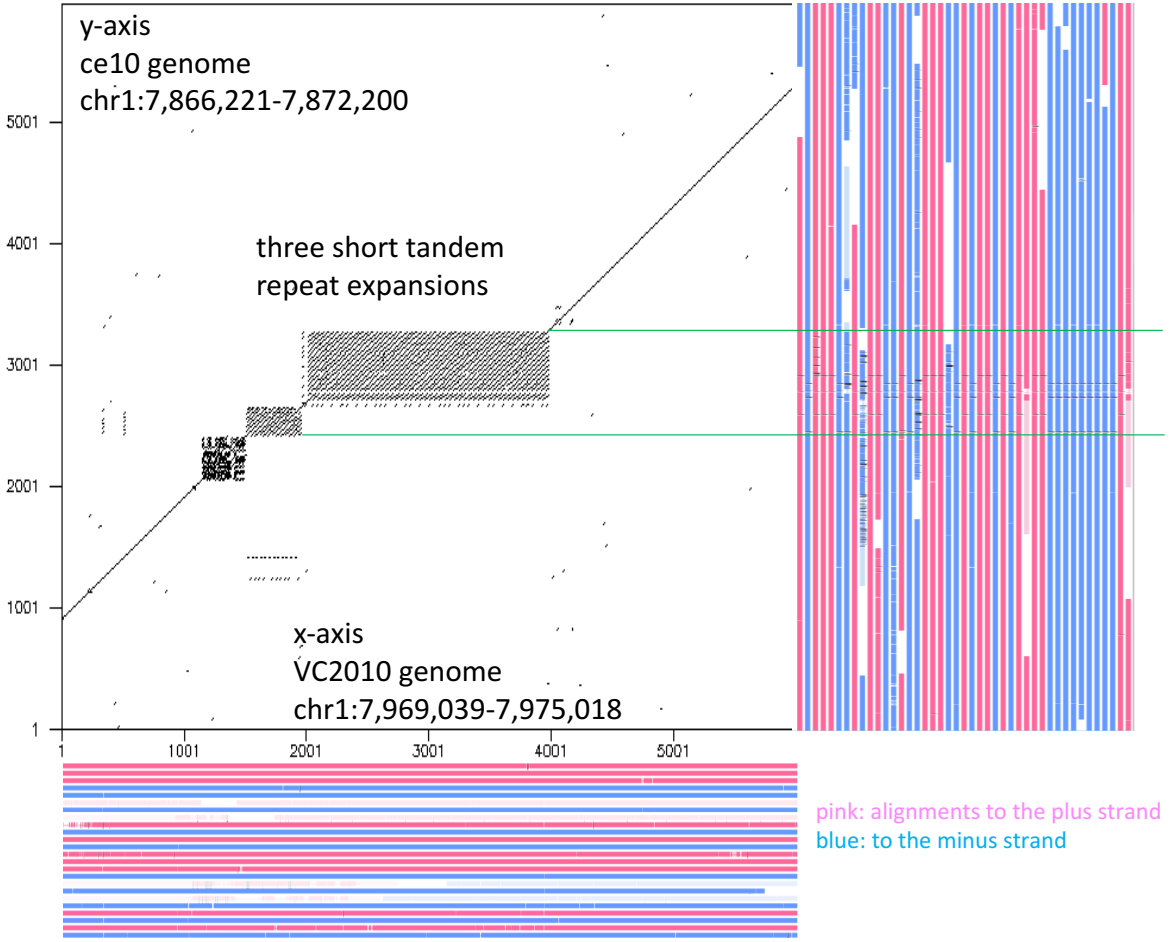
Beyond PacBio [1]: Oxford Nanopore is generating single genomic reads of 80-800 kb



Ref.: *http://lab.loman.net/2017/03/09/ultrareads-for-nanopore*

Nanopore sequencing has allowed gap-closure in resequencing of *C. elegans*

Confirmation of short tandem repeat expansion in VC2010



Alignments of PacBio raw reads from VC2010 to the ce10 reference genome.

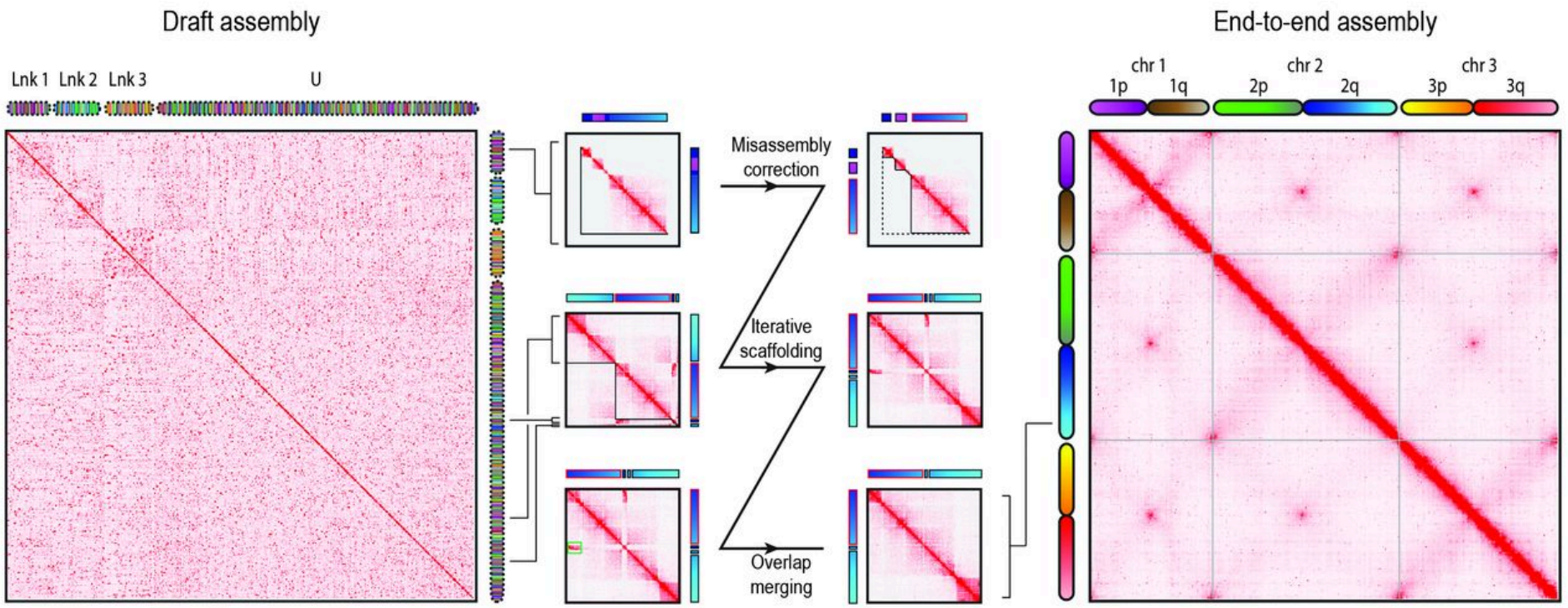
All alignments have small black bars that represent deletions of short tandem repeats that could not be aligned to the ce10 reference because the repeat expansions are underestimated in the ce10 genome.

pink: alignments to the plus strand
blue: to the minus strand

These alignments of PacBio raw reads from VC2010 support the three short tandem repeat expansions in the VC2010 genome assembly. Alignments that do not cover the expansions are masked by white boxes.

C. elegans Resequencing Consortium, unpublished.

Beyond PacBio [2]: efficient Hi-C scaffolding of the *Aedes aegypti* genome (1.26 Gb!)

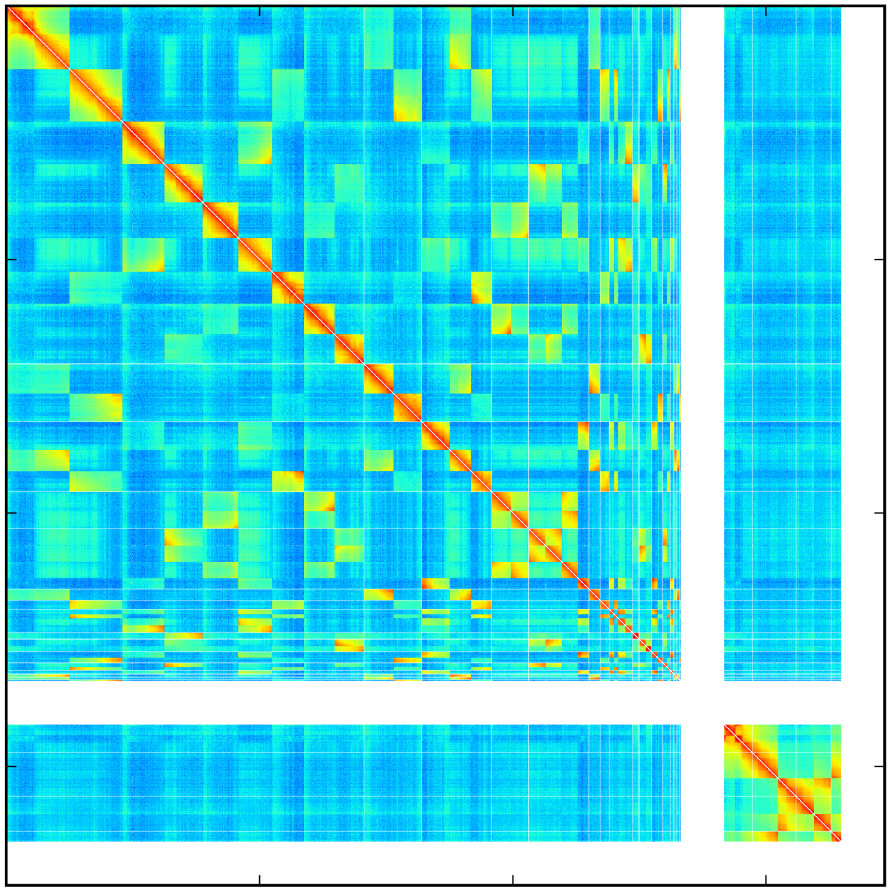


"The resulting assembly ... contained three huge scaffolds (307, 472, and 404 Mb in length) comprising 93.6% of the input sequence... The three huge scaffolds correspond to chromosomes 1, 2, and 3..."

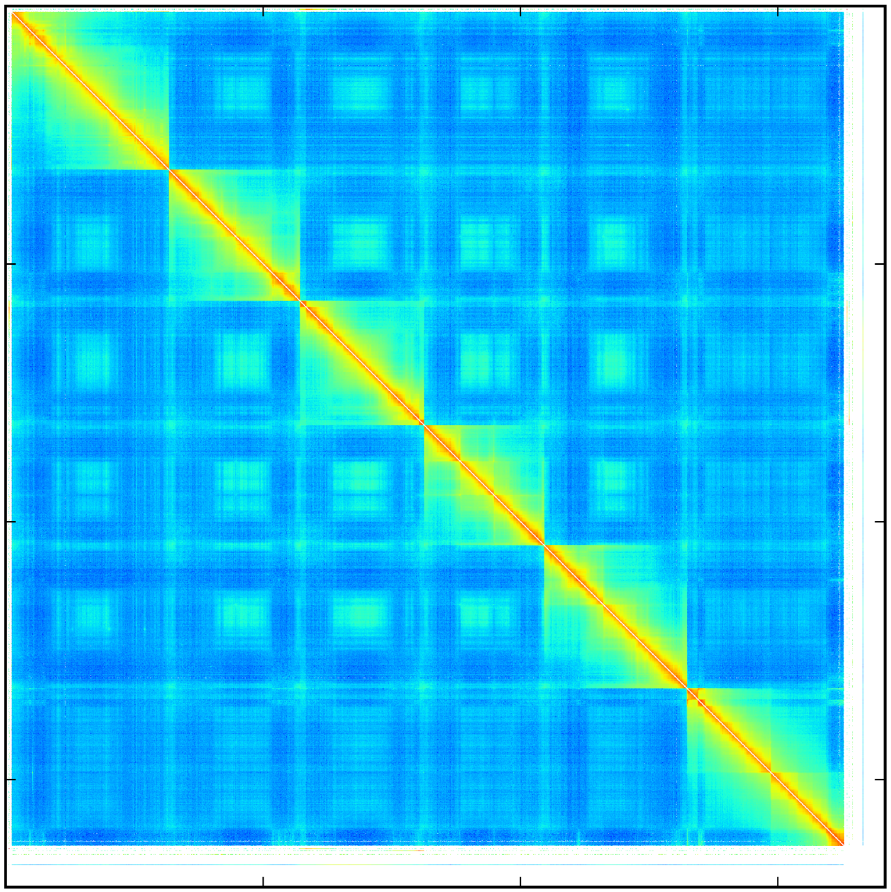
Ref.: Dudchenko et al. (2017), Science *356*, 92-95; PubMed 28336562.

Hi-C scaffolding is working on *C. tropicalis* and *C. wallacei*

C. tropicalis, 140 contigs



C. tropicalis, scaffolded



Erika Anderson and Barbara Meyer, unpublished.

Summing up: possible near-future of genomics

In the immediate future, it should be practical to make
near-perfect animal genome assemblies
that have one scaffold per chromosome
and one to a few contigs per chromosome.

Given intelligent analysis of these genomes,
one should also have
near-perfect genomic comparisons
of diverse animal species.

Biological insight will still be crucial:
there need to be functional validations of
well-chosen subsets of the genomic data.
And there will still be surprises!

Thanks:

U. Maryland:
Da Yin
Rebecca Felde
Eric Haag

U. Toronto:
Cristel Thomas
Asher Cutter

UC Berkeley:
Caitlin Schartner
Ed Ralston
Erika Anderson
Barbara Meyer

UC Davis:
Ian Korf

U. Tokyo:
Jun Yoshimura
Kazuki Ichikawa
Shinichi Morishita

Stanford:
Massa Shoura
Karen L. Artiles
Idan Gabdank
Lamia Wahba
Cheryl Smith
Andrew Fire

U. Minnesota/CGC:
Ann Rougvie

U. Edinburgh:
Georgios Koutsovoulos
Lewis Stevens
Sinduja Chandrasekar
Mark Blaxter

U. British Columbia:
Mark Edgley

U. Washington:
Bob Waterston

CSHL:
Eric Antoniou

Cal. Inst. of Tech.:
Paul Sternberg