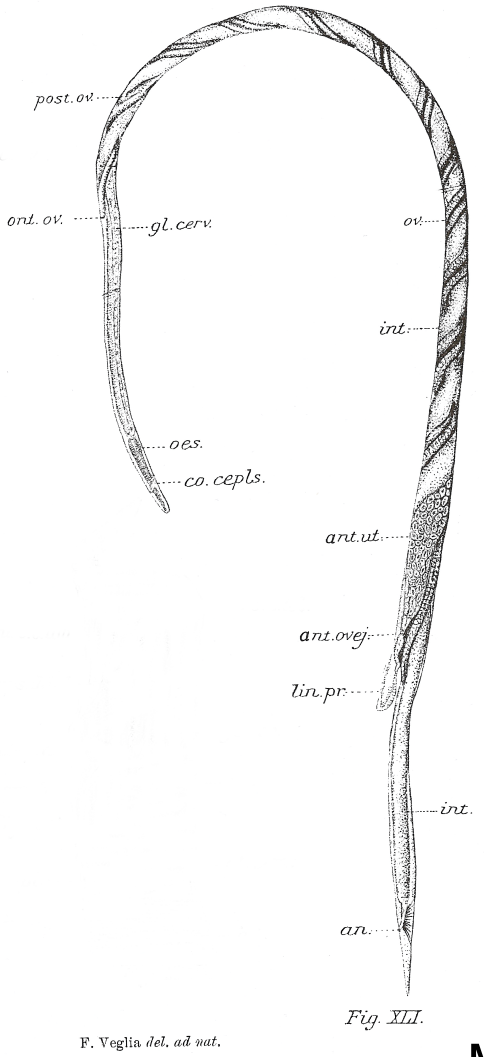
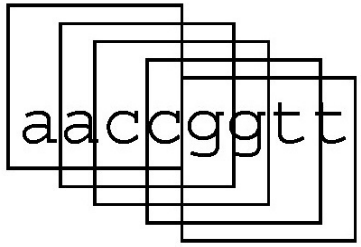
Sequencing nematode genomes usefully

Erich Schwarz, Caltech ➔ Cornell



aaccgg
ccggtt



MSU NGS course, June 2012

Overview

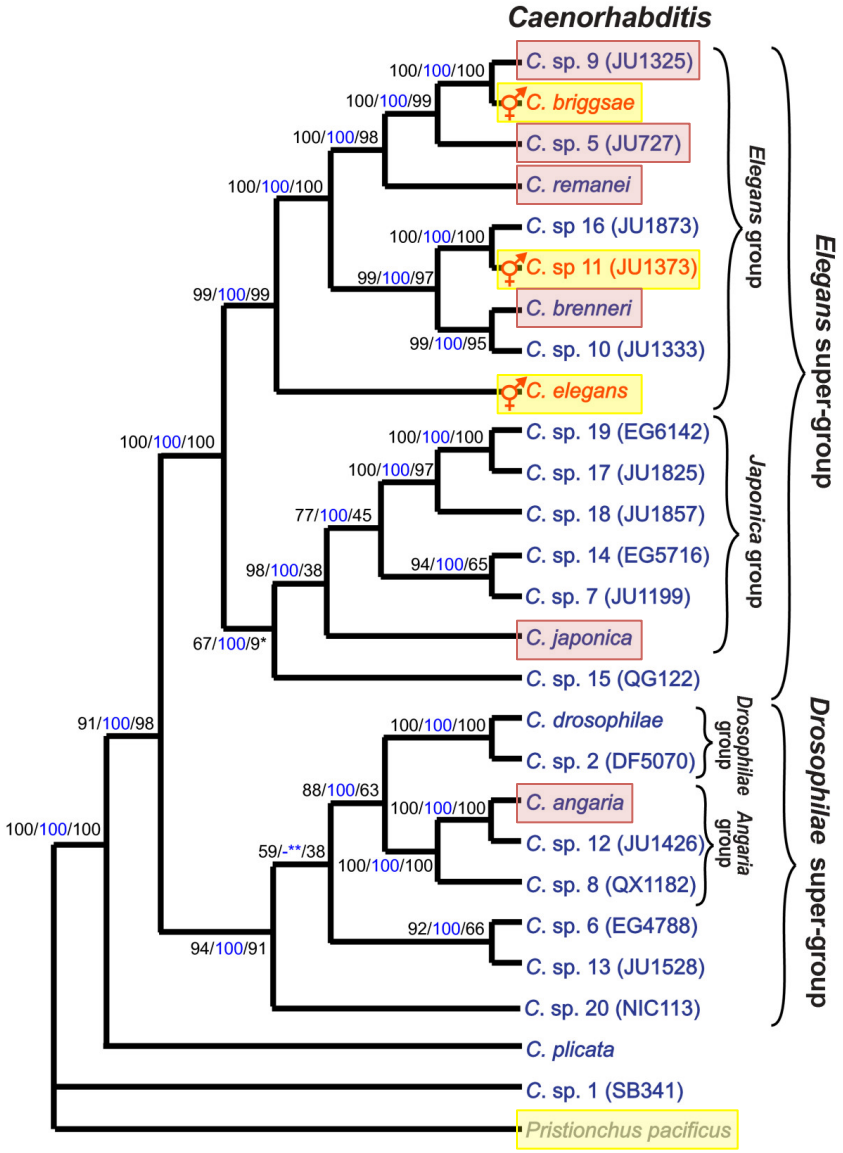
1. *Caenorhabditis angaria* PS1010

2. *Caenorhabditis* spp. 7, 9, 11

3. *Haemonchus contortus*

4. General lessons

Caenorhabditis molecular phylogeny



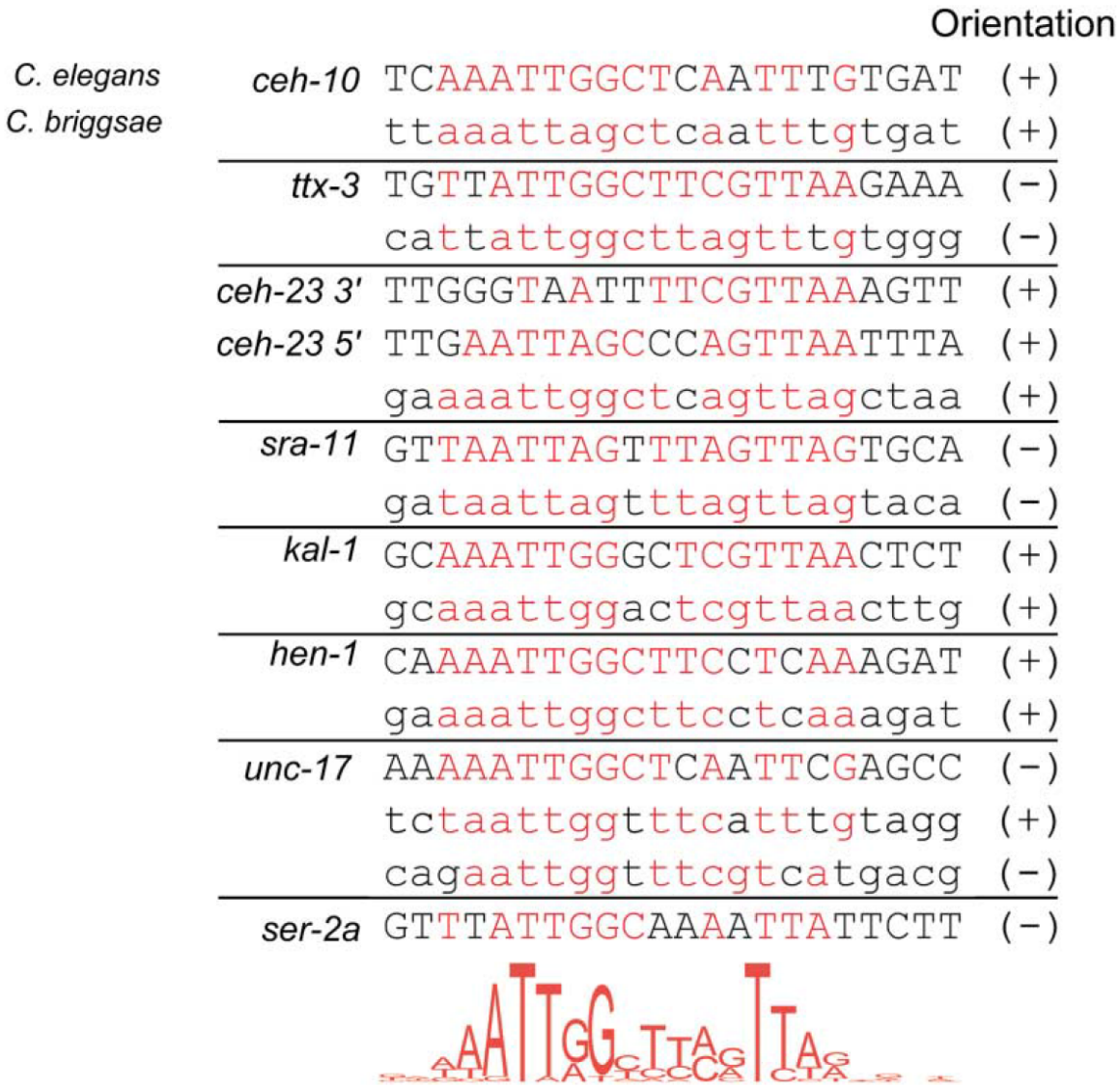
Seq. genomes:

= ♀ ♂

= ♀ ♂

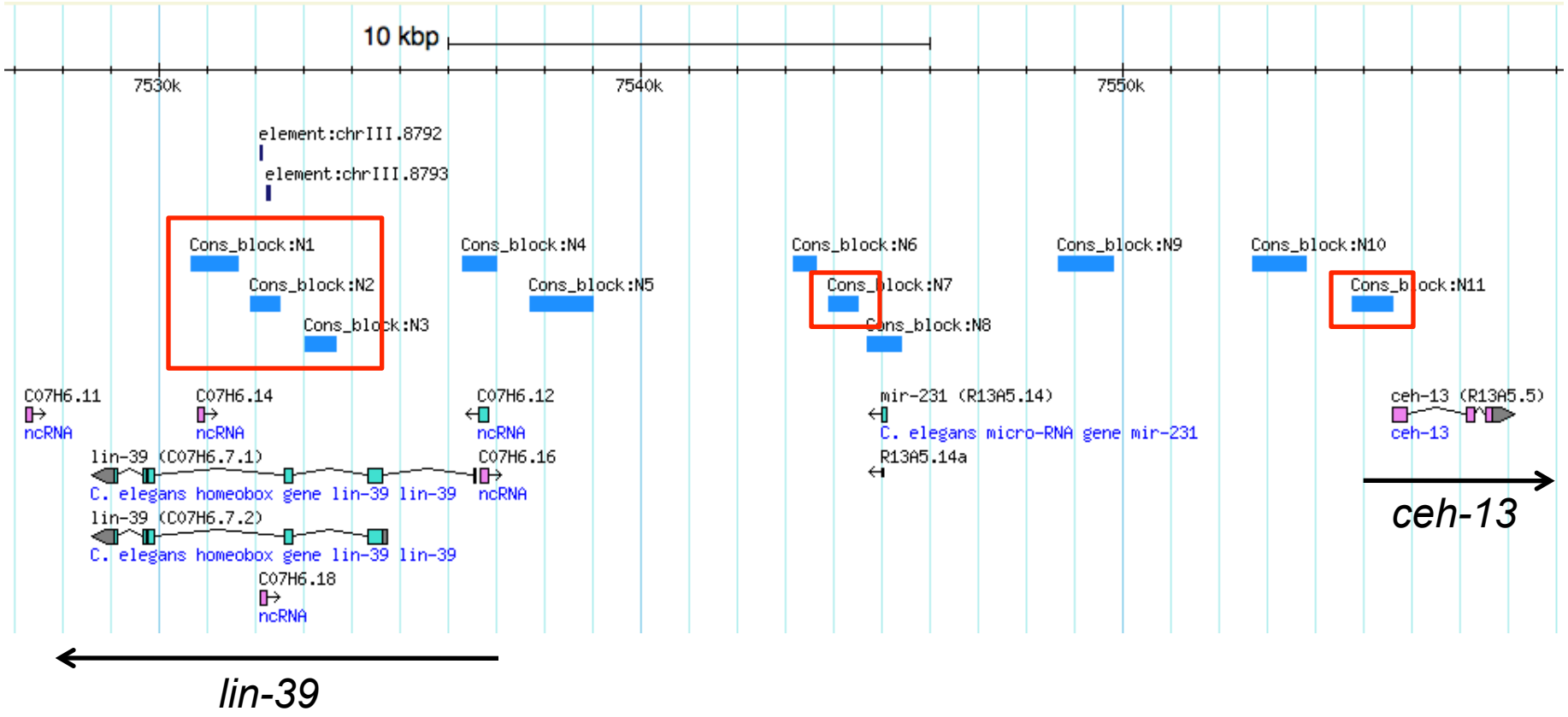
Ref.: Kiontke et al. (2011), BMC Evol. Biol. *11*, 339.

Cis-regulatory DNA tends to be conserved



Ref.: Wenick and Hobert (2004), Dev. Cell *6*, 757-770.

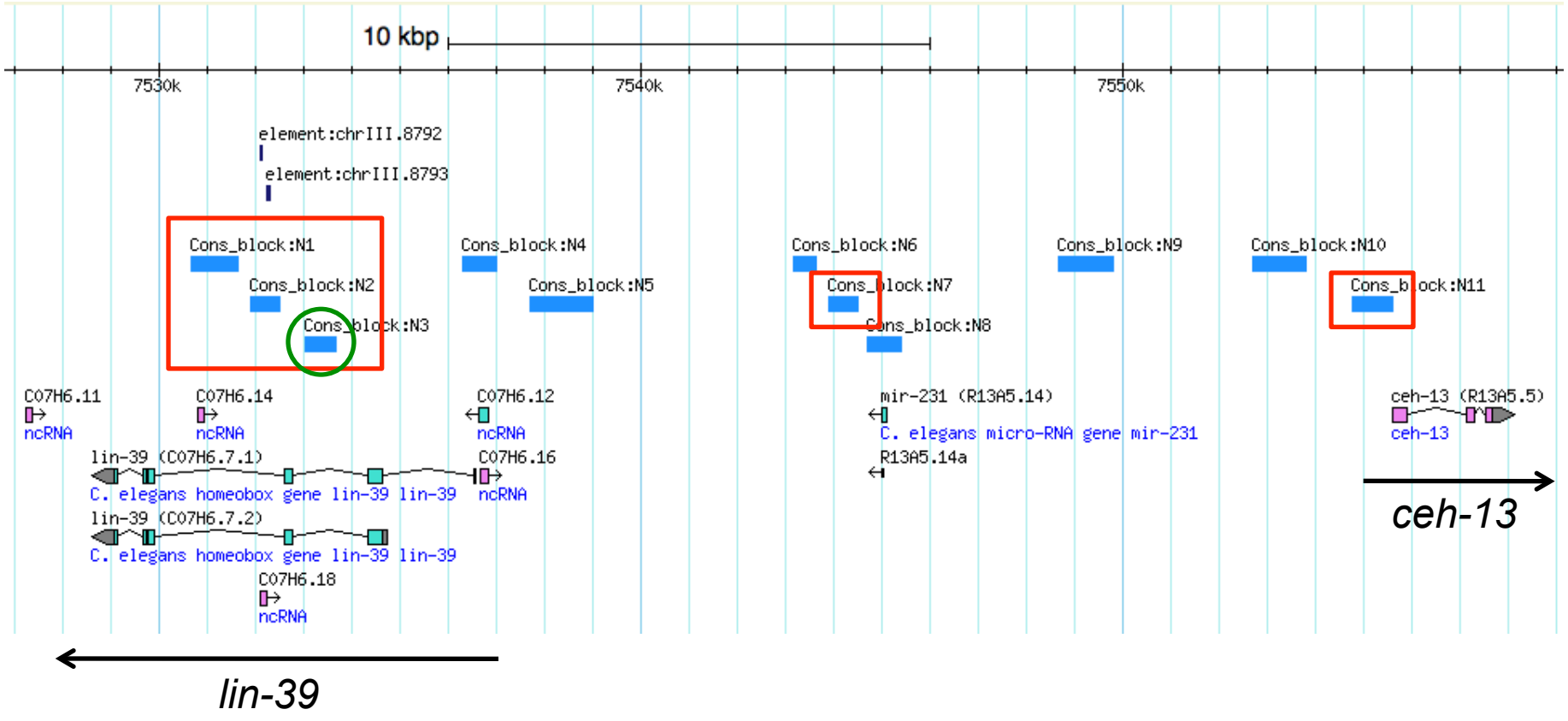
ceh-13/*lin-39* Hox has regulatory elements conserved between *elegans* and *angaria*



= conserved between *elegans* and *angaria*

Ref.: Kuntz, Schwarz, et al. (2008), Genome Res. *18*, 1955-1968.

ceh-13/*lin-39* Hox has regulatory elements conserved between *elegans* and *angaria*



= cons. between *elegans* and *angaria* = cons. between *elegans* and mouse

Ref.: Kuntz, Schwarz, et al. (2008), Genome Res. *18*, 1955-1968.

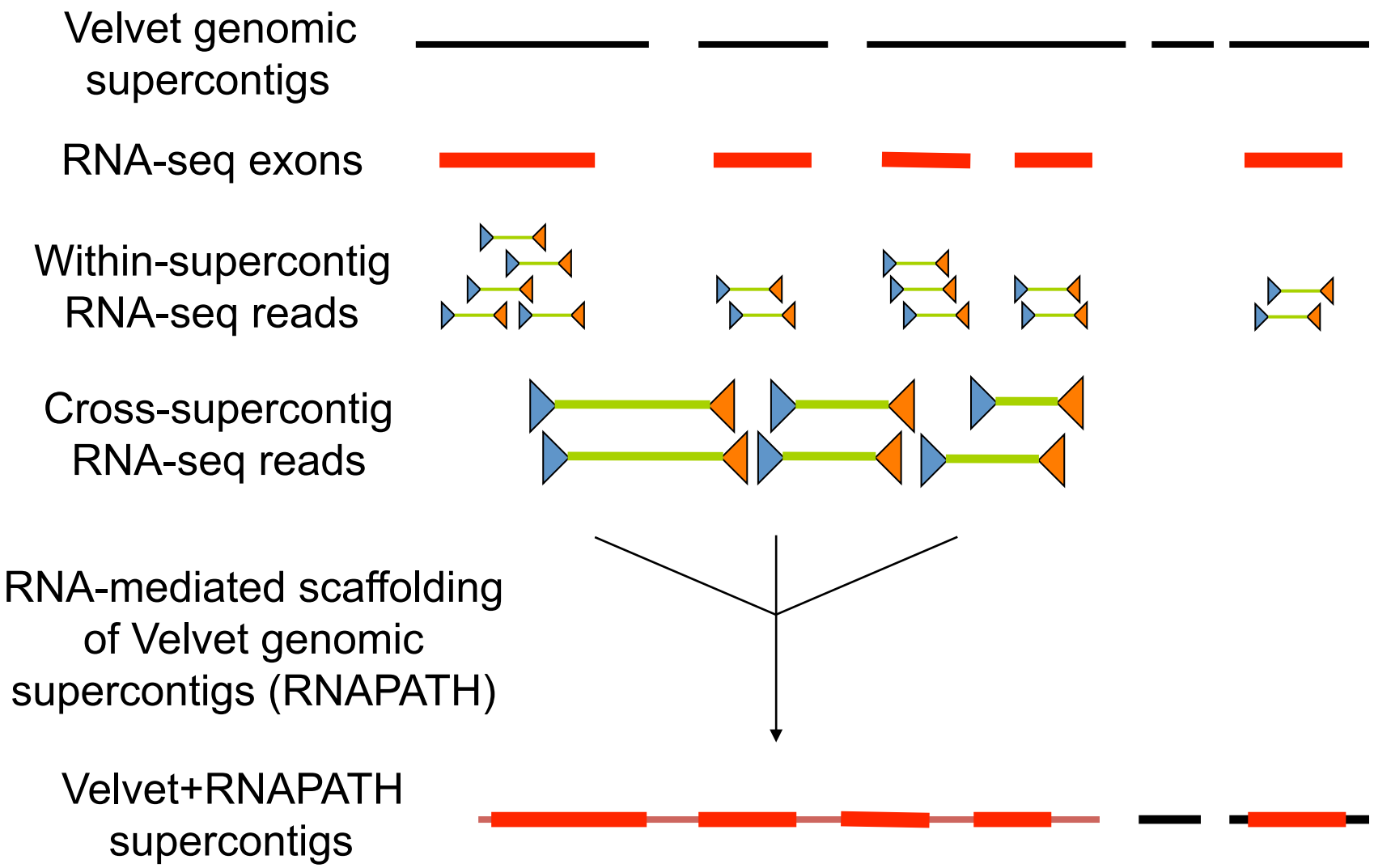
A genome sequence for *C. angaria*

Assembly type	Total (Mb)	Maximum supercontig size (kb)	N50 (kb)	Predicted genes
Genomic	79.8	45.7	5.1	27,741
cDNA	17.8	14.5	1.05	n/a

Est. genome size: 100 Mb.

Ref.: Mortazavi, Schwarz, et al. (2010), Genome Res. *20*, 1740-1747.

RNA scaffolding



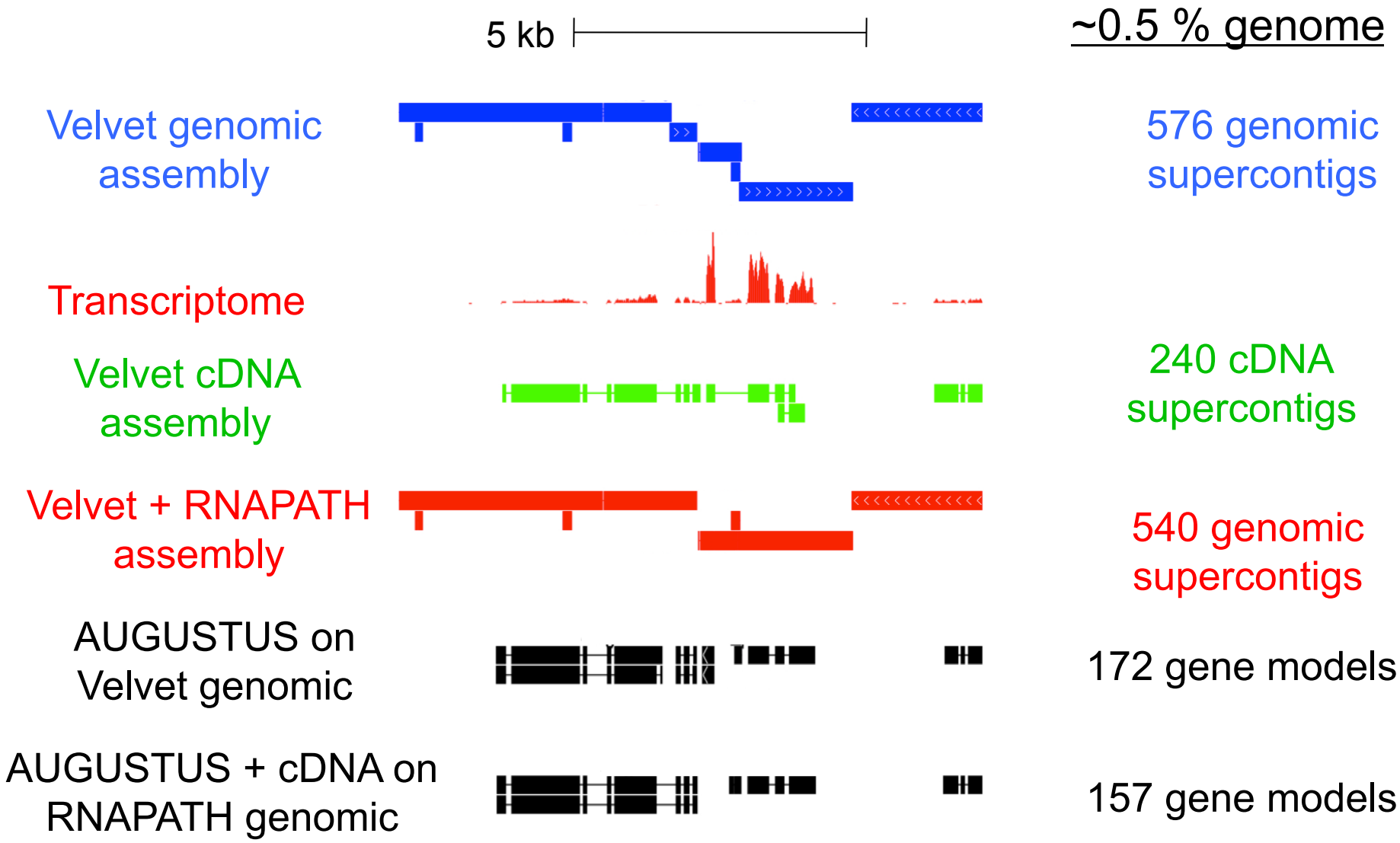
A better genome sequence for *C. angaria*

Assembly type	Total (Mb)	Maximum supercontig size (kb)	N50 (kb)	Predicted genes
Genomic	79.8	45.7	5.1	27,741
Genomic with RNA scaffolding	79.8	96.3	9.4	22,851
cDNA	17.8	14.5	1.05	n/a

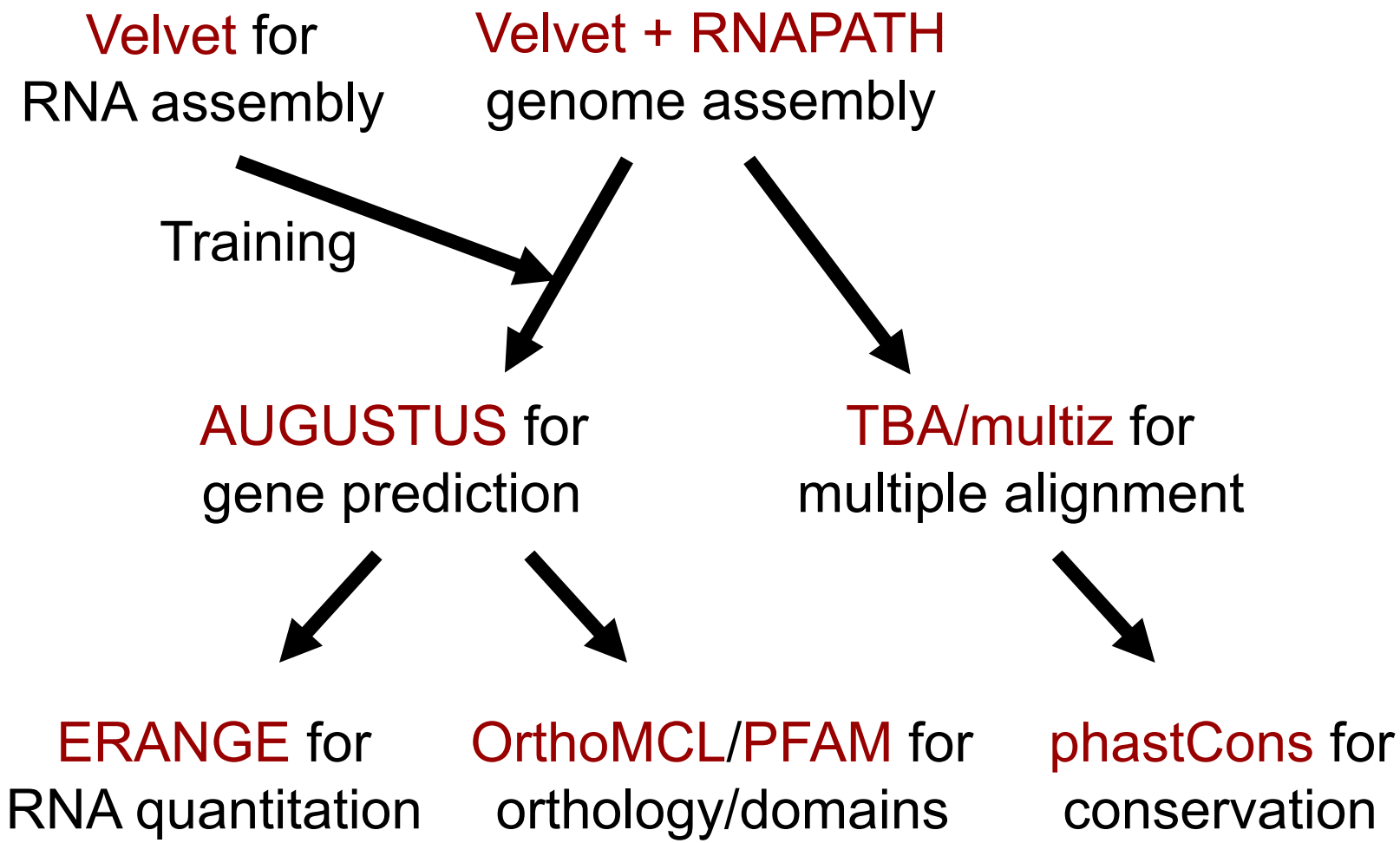
Est. genome size: 100 Mb.

Ref.: Mortazavi, Schwarz, et al. (2010), Genome Res. *20*, 1740-1747.

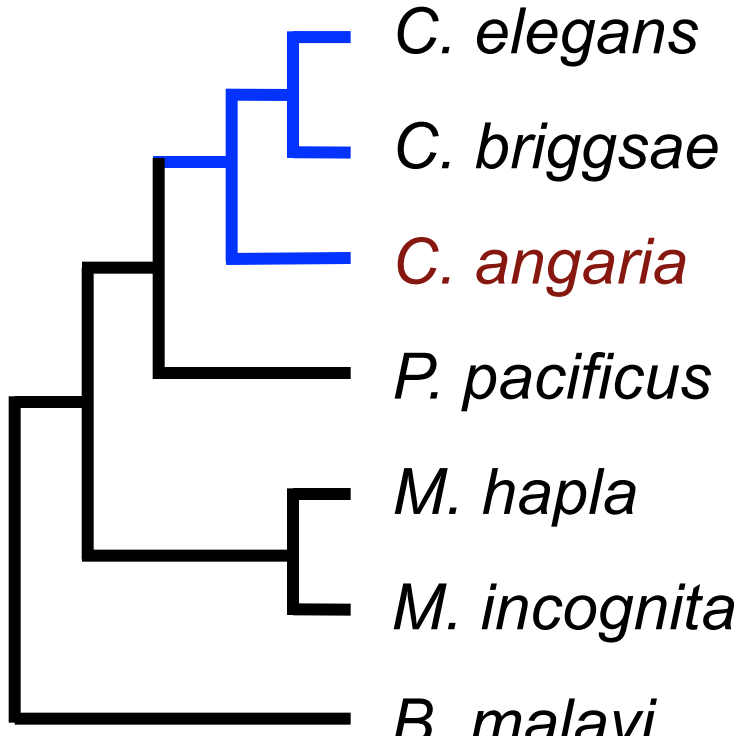
Testing assemblies on 429 kb of Sanger contigs



Annotation strategy used for *C. angaria*



~93% of *elegans* orthologs are found in *angaria*



Strict orthology to:

	elegans	*briggsae*	*angaria*
C. elegans			
C. briggsae	7,960		
C. angaria	5,623	5,553	
P. pacificus	3,466	3,429	3,278
M. hapla	2,214	2,178	2,017
M. incognita	1,734	1,694	1,573
B. malayi	2,794	2,747	2,562

Comparing *angaria* to *elegans* reveals new conserved non-coding DNA elements

Genefinder predictions

WormBase *C. elegans unc-2*

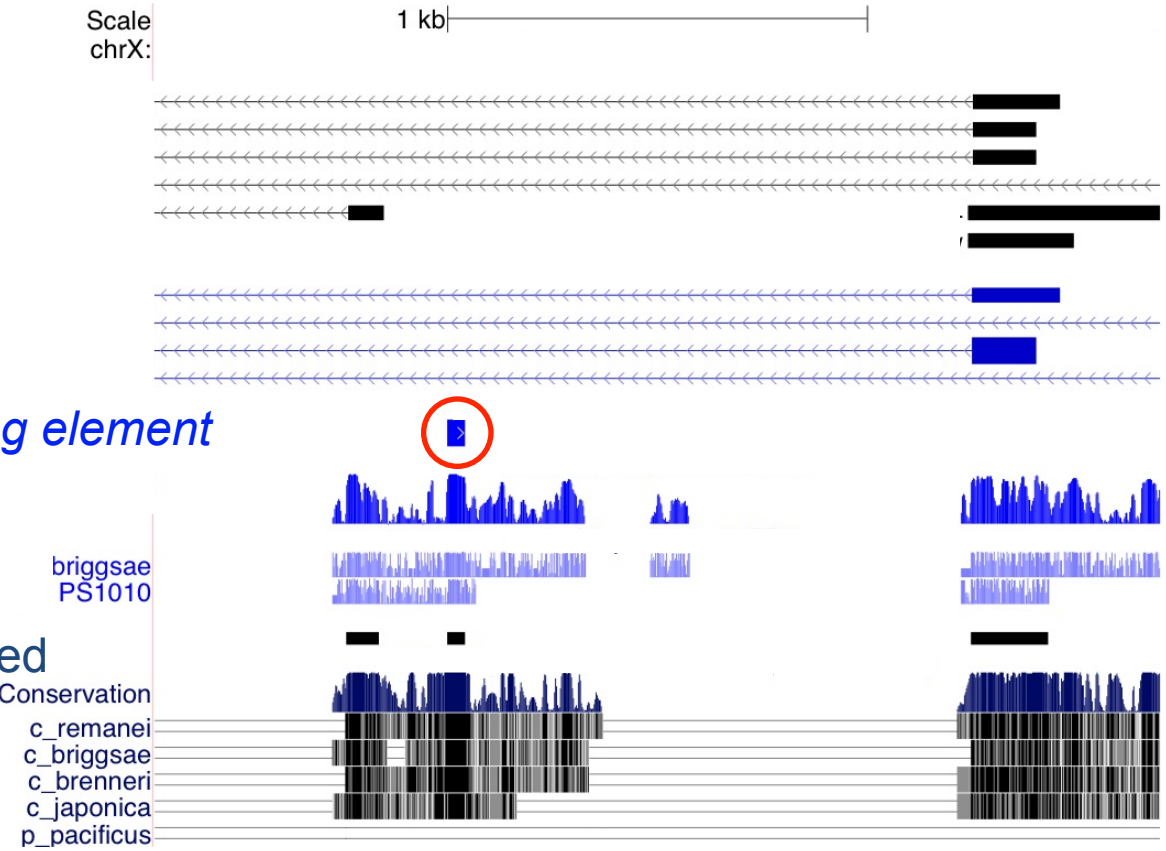
C. angaria conserved non-coding element

3-way PhastCons score

3-way TBA/MultiZ

3-way PhastCons most conserved

UCSC PhastCons and multiz



Finding possible regulatory elements by weeding out everything else

DNA elements	Number	Min. nt	Avg. nt	Max. nt
PhastCons	96 K	3	65	2,514

Finding possible regulatory elements by weeding out everything else

DNA elements	Number	Min. nt	Avg. nt	Max. nt
PhastCons	96 K	3	65	2,514
80% overlap w/ aligns; no repeats	84 K	3	66	2,514

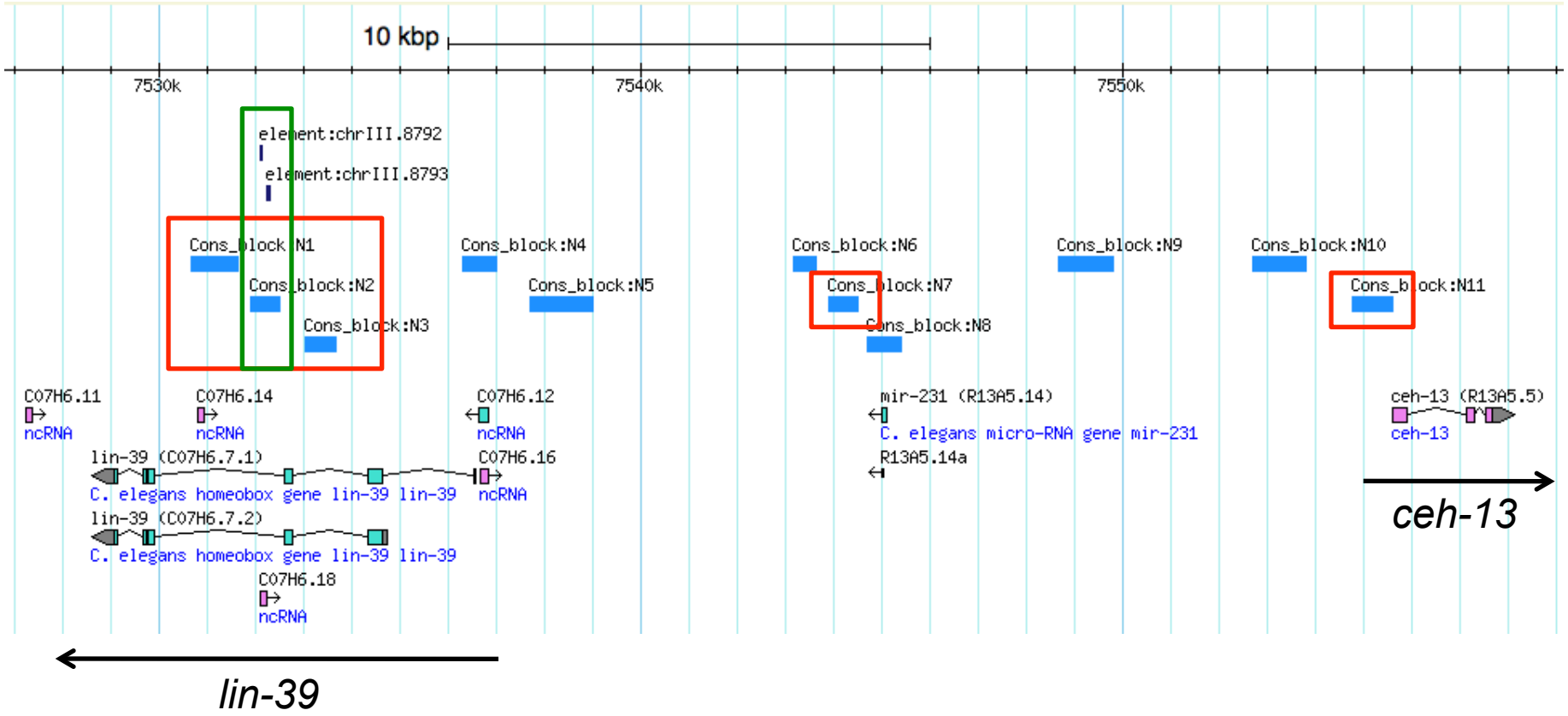
Finding possible regulatory elements by weeding out everything else

DNA elements	Number	Min. nt	Avg. nt	Max. nt
PhastCons	96 K	3	65	2,514
80% overlap w/ aligns; no repeats	84 K	3	66	2,514
No official exons (prot.-cod. or ncRNA)	4,494	7	36	444

Finding possible regulatory elements by weeding out everything else

DNA elements	Number	Min. nt	Avg. nt	Max. nt
PhastCons	96 K	3	65	2,514
80% overlap w/ aligns; no repeats	84 K	3	66	2,514
No official exons (prot.-cod. or ncRNA)	4,494	7	36	444
No unofficial exons (from mGene, etc.)	2,672	7	29	160

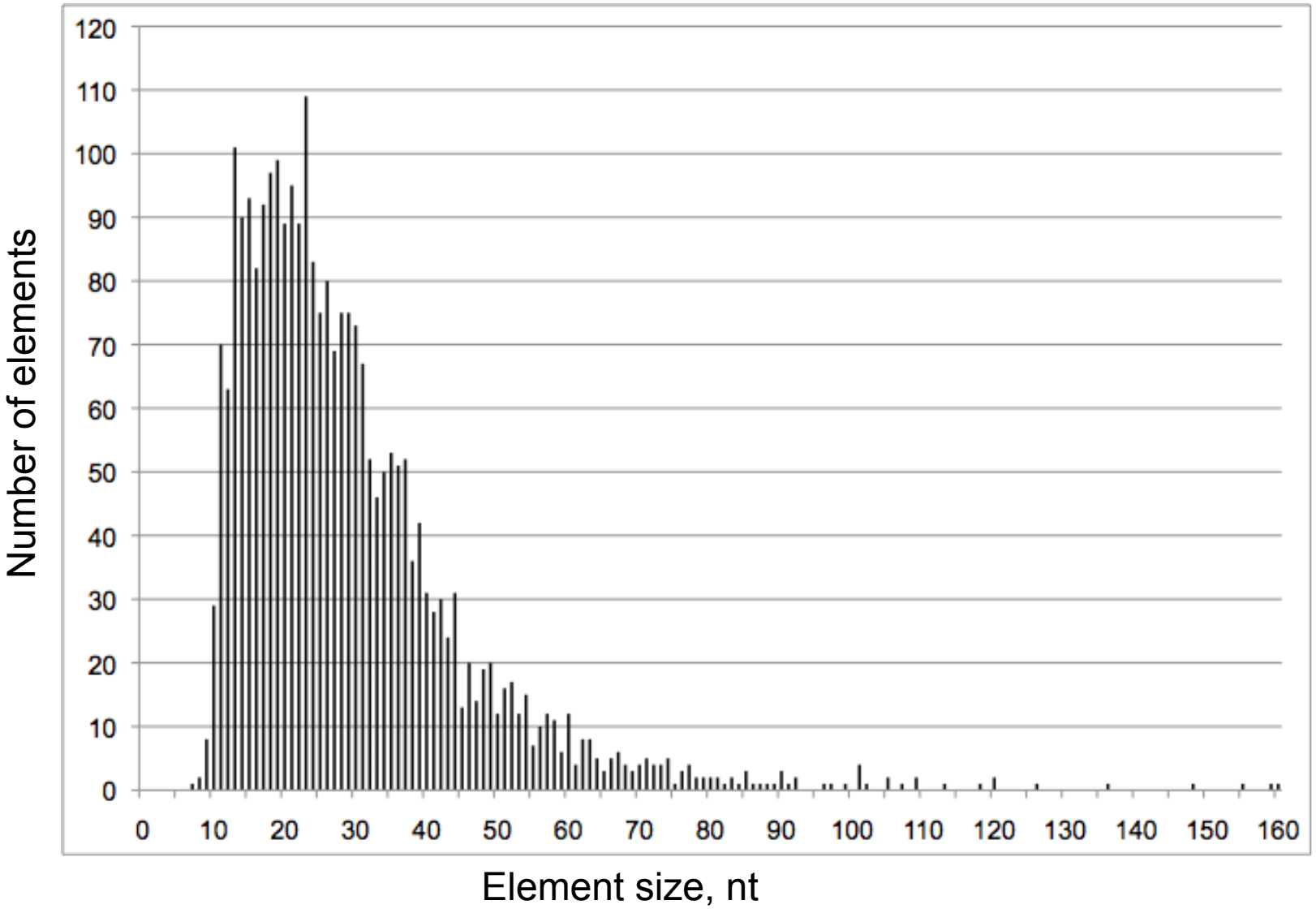
Rediscovering 1/5 Hox regulatory blocks



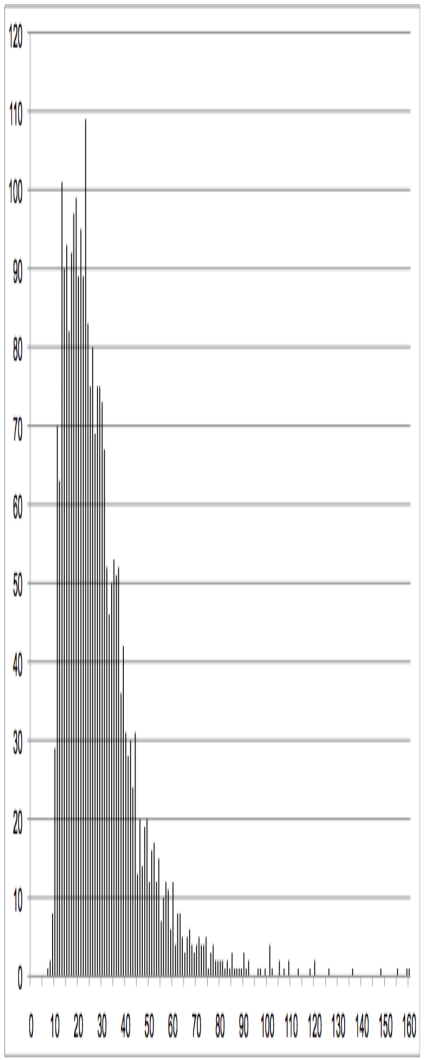
= conserved between *elegans* and *angaria*

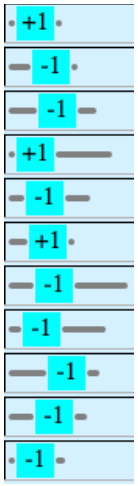
= rediscovered in genome-wide *elegans*/*angaria* comparison

90% of filtered elements are ≤50 nt long

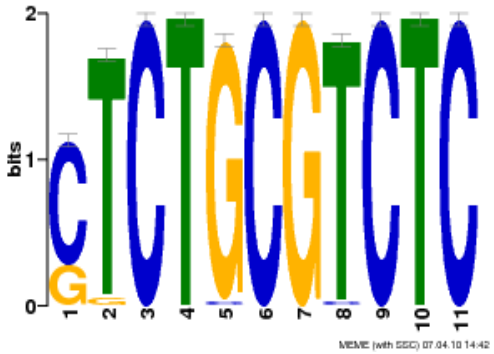


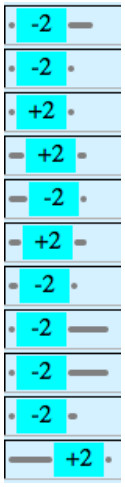
Extracting shorter motifs from the elements



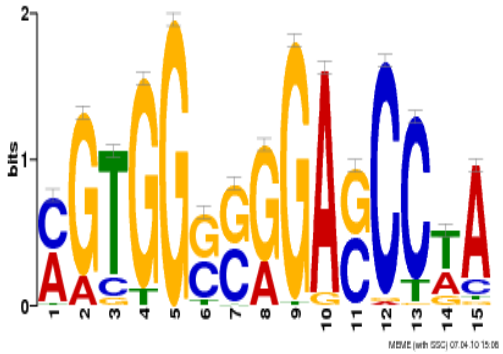



```
          T CTCTGCGTCTC T
          T CTCTGCGTCTC TTCGTCT
     TTTATT CTCTGCGTCTC TCCCTCCCT
          T CTCTGCGTCTC TCACTCTTAC
   TTATTTGT CTCTGCGTCTC TCAAA
     ATTTAT CTCTGCGTCTC
TTCTCTGTTT CTCTGCGTCTC TCATCTCT
CAATTTAACT CTCTGCGTCTC TCAC
       ATTT CTCTGCGTCTC TCGCTCTCTC
       GCTT CTCTGCGTCTC TCGATTTC
        TTT CTCTGCGTCTC
```



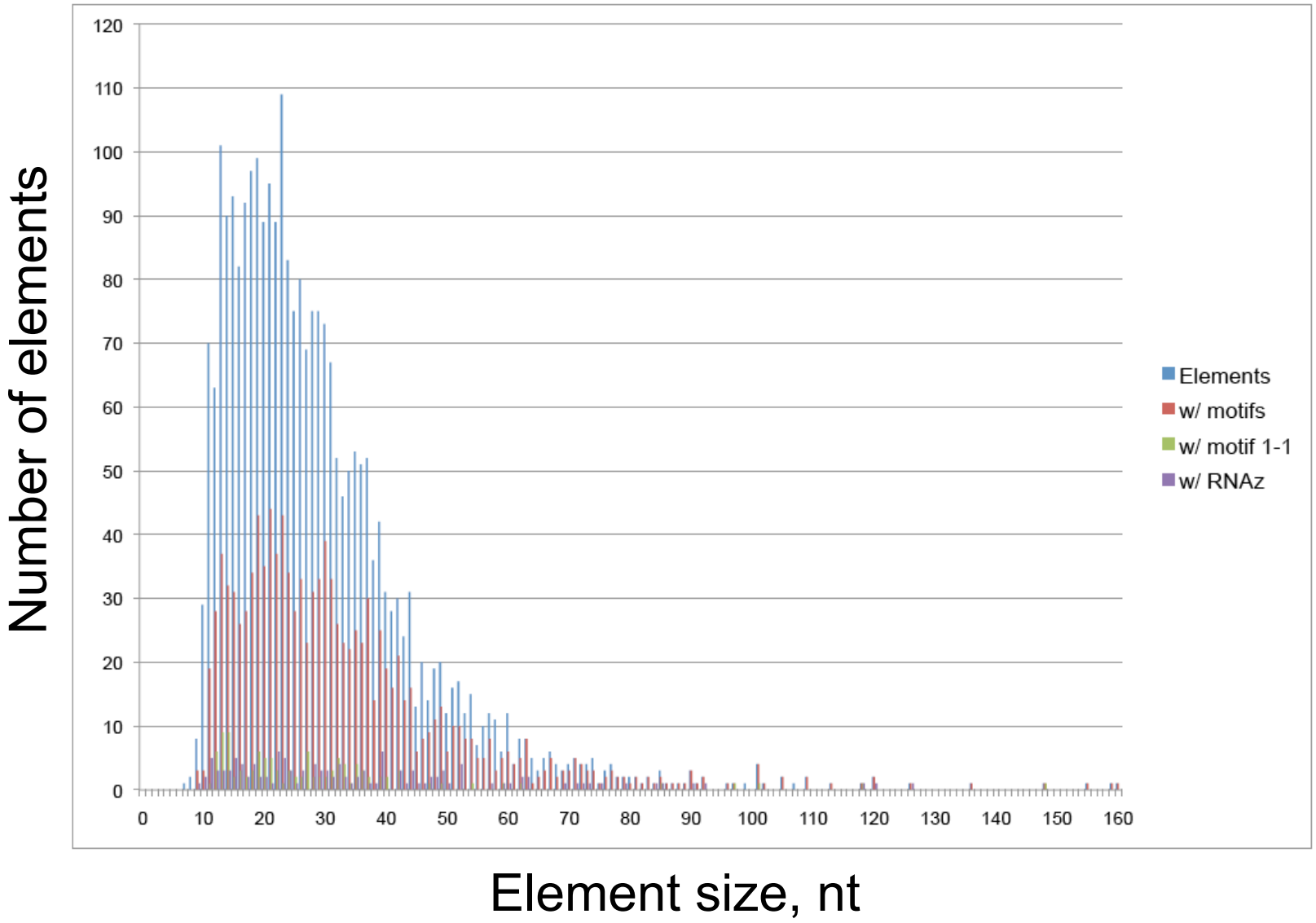



```
  CGGACTTT CGTGGCGGGACCCAA
           CGTGGCGGGACCCAA A
         T CGTGGCGGGACCCAA A
      CAAAA AGTGGGCGGAGCCTA ACA
          A AGTGGGCGGAGCCTA TCCGGA
       AAAA AGTGGGCGGAGCCTA AATC
          A AGTGGGCGGAGCCTA AAA
TTTTTCTAAA AGTGGGCGGAGCCTA
TCAGGCAAAA AGTGGGCGGAGCCTA G
        AAA AGTGGGCGGAGCCTA A
AATTTCCAAA AGTGGGCGGAGCCTA
```



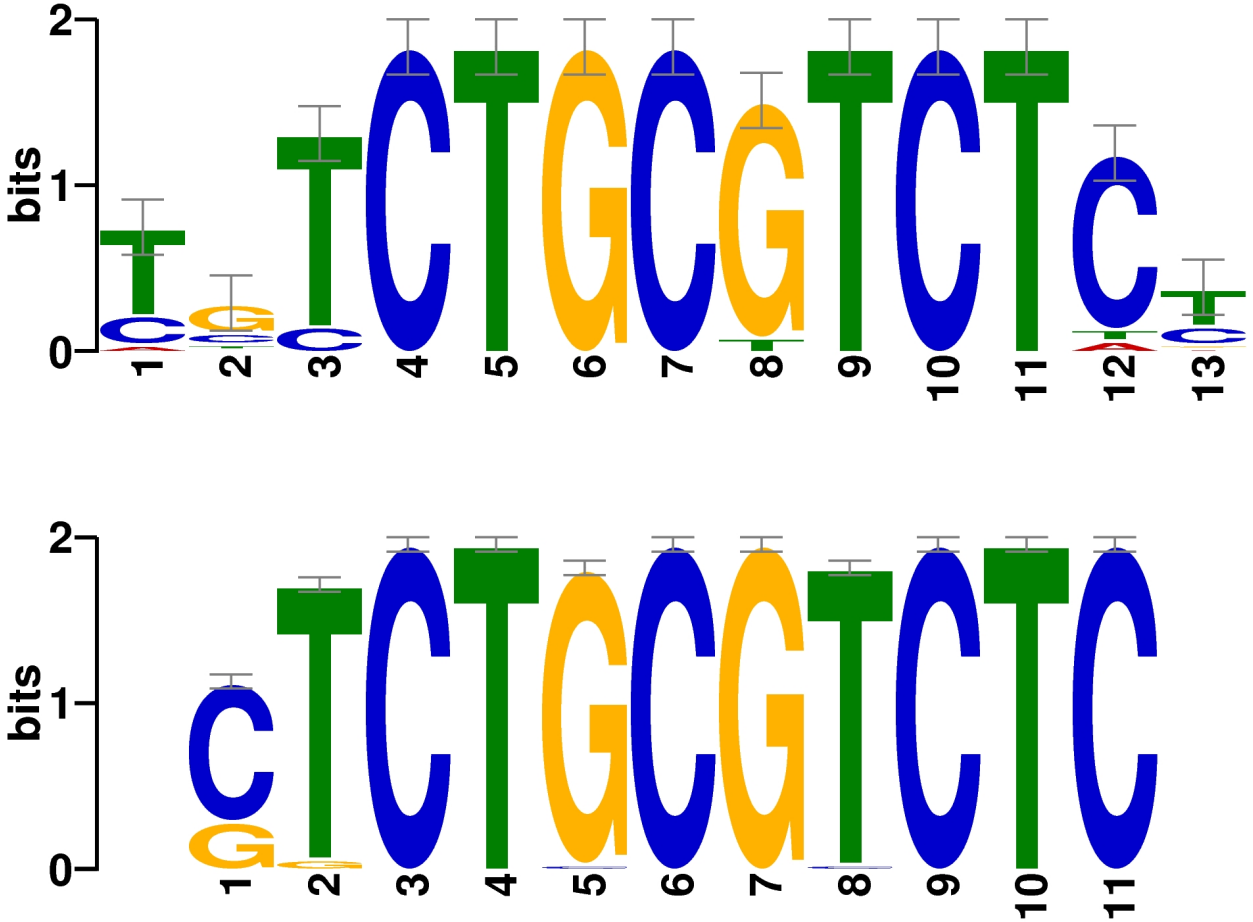
Motifs via MEME. Ref.: Bailey and Elkan (1994), Proc. Int. Conf. Intell. Syst. Mol. Biol. *2*, 28–36.

45% of elements contain 22 shorter motifs



Reprediction of a (newly) known motif

slr-2/jmc-1



Motif 1-1

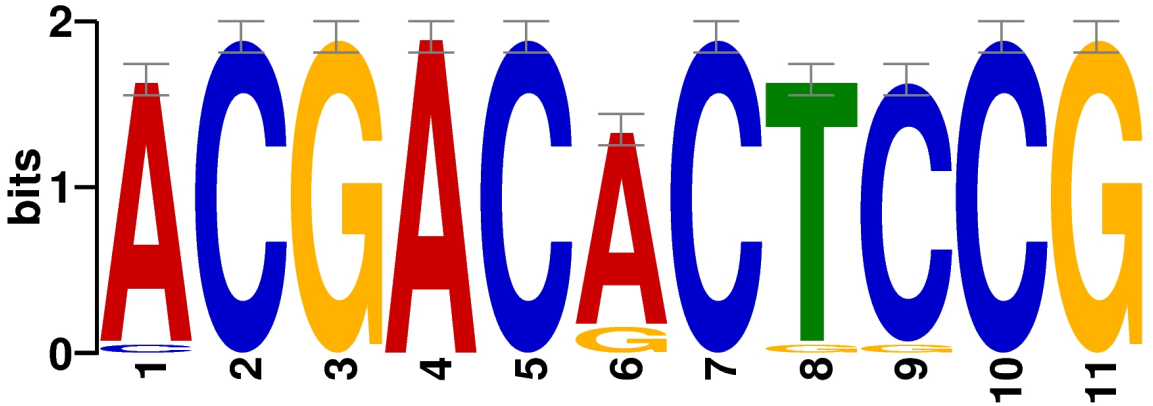
slr-2/jmc-1: Kirienko and Fay (2010), EMBO J. *29*, 727-739.

22 MEME motifs: 7 knowns

Motif	Description	Size (nt)	Consensus sequence	E-value	GO term (P-value < 1e-06)
1-1	*slr-2*/*jmjc-1*	11	STCTGCGTCTC	3.4e-140	
1-5	Muscle 3	15	SMGMSMCSMSMCMSC	1.5e-54	nematode larval development (GO: 0002119; 1.31e-07)
1-6	Muscle 1	15	GASRRAGASASRSAG	4.6e-59	locomotion (GO: 0040011; 2.23e-07); body morphogenesis (GO:0010171; 4.93e-07)
1-9	miRNA 5' flank/Sp1	8	CYCCGCCC	7.2e-42	
1-11	Resembles early *pha-4* / Muscle 4	15	RYGTSWBKGTGTKTG	2.4e-31	
1-14	Muscle 2	15	AGRAGAWGAARAMGA	4.4e-30	locomotion (GO: 0040011; 5.41e-07)
2-30	E2F	8	SGCGCSRA	1.9e-02	locomotion (GO: 0040011; 4.2e-07)

Reprediction of a previous in silico motif

Beer/Tavazoie motif 4



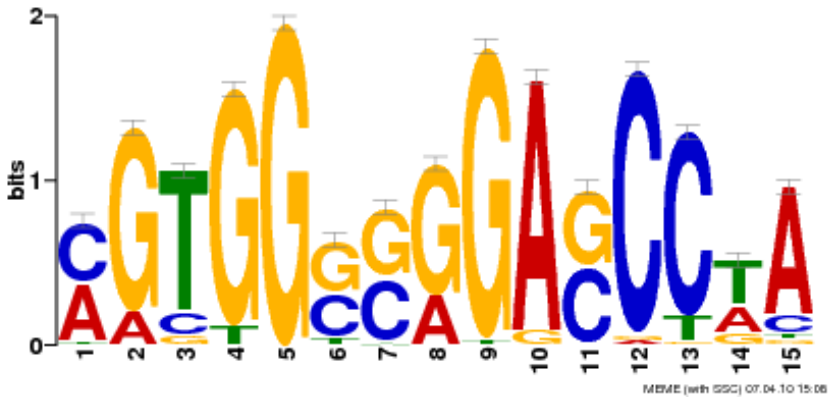
Motif 1-8

375 in silico predictions: Beer and Tavazoie (2004), Cell *117*, 185-198.

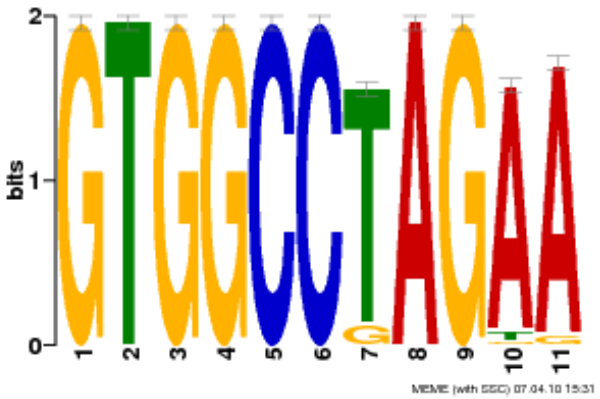
22 MEME motifs: 4 uncharacterized in vivo

Motif	Description	Size (nt)	Consensus sequence	E-value	GO term (P-value < 1e-06)
1-8	Uncharacterized: previously found by Beer and Tavazoie (2004) as highly significant motif (4th out of 375)	11	ACGACACTCCG	6.6e-48	
1-15	Uncharacterized: previously found by Beer and Tavazoie (2004); also has possible mammalian homolog (PF0082.1; Xie et al., 2005)	11	TGCGCCTTTAA	1.9e-26	
1-19	Uncharacterized: previously found in most significant 10% of Beer and Tavazoie (2004) motifs	13	TCGYKKCRAGACC	4.9e-10	
2-18	Uncharacterized: motif 140 of Beer and Tavazoie (2004)	15	BCYCGTAAATCSACA	3.5e-16	

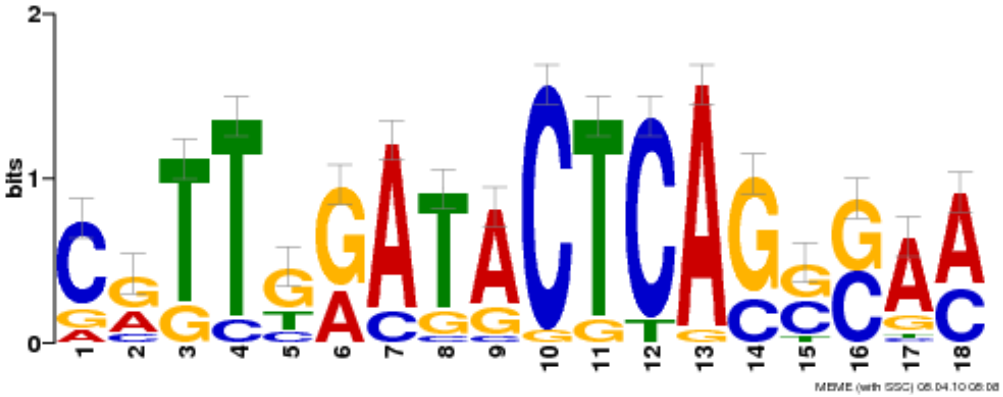
Novel motifs



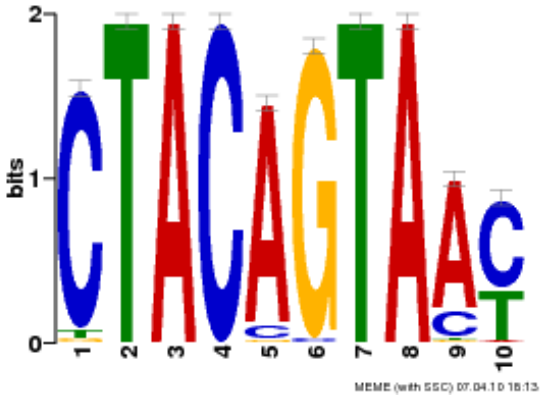
1-2



1-3



2-24



1-10

22 MEME motifs: 11 novel

Motif	Description	Size (nt)	Consensus sequence	E-value	GO term (P-value < 1e-06)
1-2	Novel	15	MGTGGSSRGASCCWA	6.0e-131	
1-3	Novel	11	GTGGCCTAGAA	3.1e-129	
1-4	Novel	14	GCAARYGCGCTCYA	8.7e-135	
1-10	Novel	10	CTACAGTAAY	1.6e-34	
1-17	Novel	15	GTCCKAGAGGASTAC	1.8e-17	
1-18	Novel	11	GGTTCGAHYCC	7.4e-14	
1-20	Novel	15	WTTACWGTTTCAAAA	4.9e-11	
2-22	Novel	15	GACMCCCAWMWYGMC	9.4e-11	
2-24	Novel	18	CRKTKRATRCTCASSSAM	3.8e-11	nematode larval development (GO: 0002119; 4.53e-07)
2-26	Novel	18	ATYWKAWTTGACGMGCAA	8.2e-06	
2-27	Novel	11	RRCTSAAAATB	3.5e-25	

Only ~4% of elements match hypothetical ncRNAs

DNA elements	Number	Min. nt	Avg. nt	Max. nt
Candidate ncRNAs / regulatory sites	2,672	7	29	160
No overlap with any RNAz predictions	2,544	7	29	160
Any overlap with novel RNAz predictions	95	10	39	126

RNAz predictions: Missal et al. (2006), J. Exp. Zool. *306B*, 379-392.

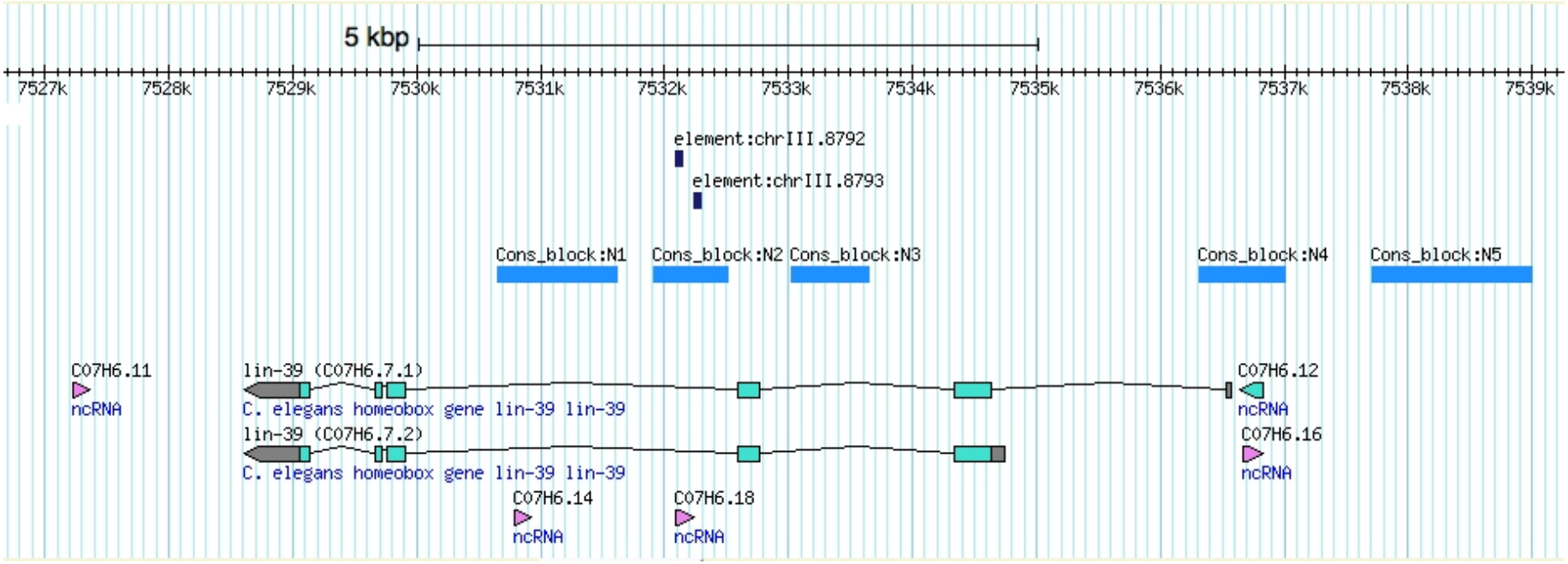
However, 9% match modENCODE ncRNAs

DNA elements	Number	Min. nt	Avg. nt	Max. nt
Candidate ncRNAs / regulatory sites	2,672	7	29	160
No overlap with any RNAz predictions	2,544	7	29	160
Any overlap with novel RNAz predictions	95	10	39	126
Any overlap w/ 7K modENCODE ncRNAs	255	7	39	159

"The 7k-set contains many RNA structural motifs, including some not found in known RNA secondary structure families ... these ncRNA candidates tend to be differentially expressed across development, with many preferentially expressed in the embryo."

Ref.: Gerstein et al. (2010), Science *330*, 1775-1787.

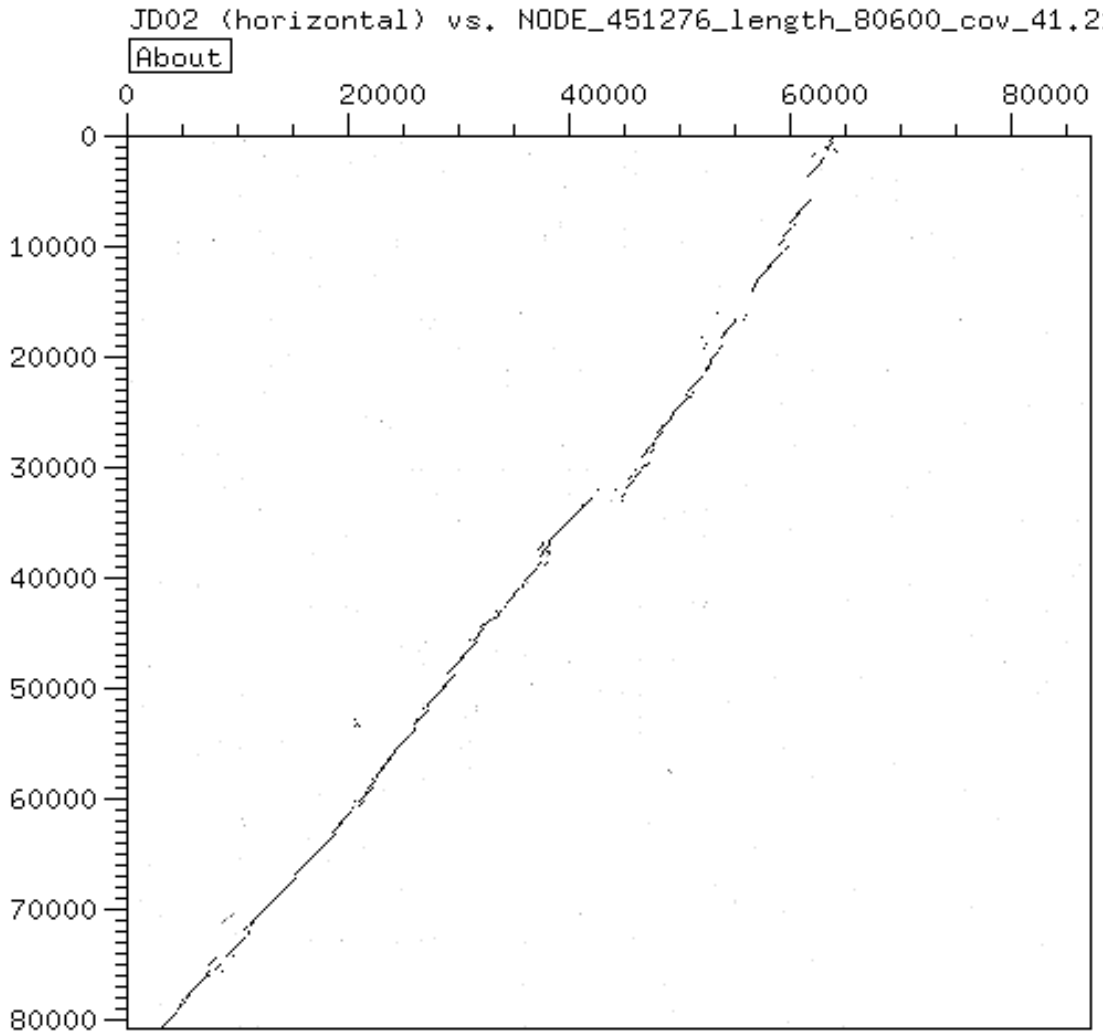
ncRNAs associated with regulatory elements



ncRNA transcribed from enhancers was first found in the *Ultrabithorax* cluster of *Drosophila*, and more recently found in mammalian neurons and macrophages.

Refs.: De Santa et al. (2010), PLoS Biol. *11*, e1000384; Ho et al. (2009), Int. J. Dev. Biol. *53*, 459-468; Kim et al. (2010), Nature *465*, 182-187.

An improved *angaria* genome



Published in 2010:

80 Mb, N50 = 9.4 kb

Added data:

HiSeq of old lib.;
2x75 of new lib.;
2.7 kb jumping lib.

New assembly:

106 Mb, N50 = 80 kb

Comparing the previous assembly to *elegans* and *briggsae* identified ~2,700 ultraconserved ncDNA elements, but it may have missed ~10K more.

Overview

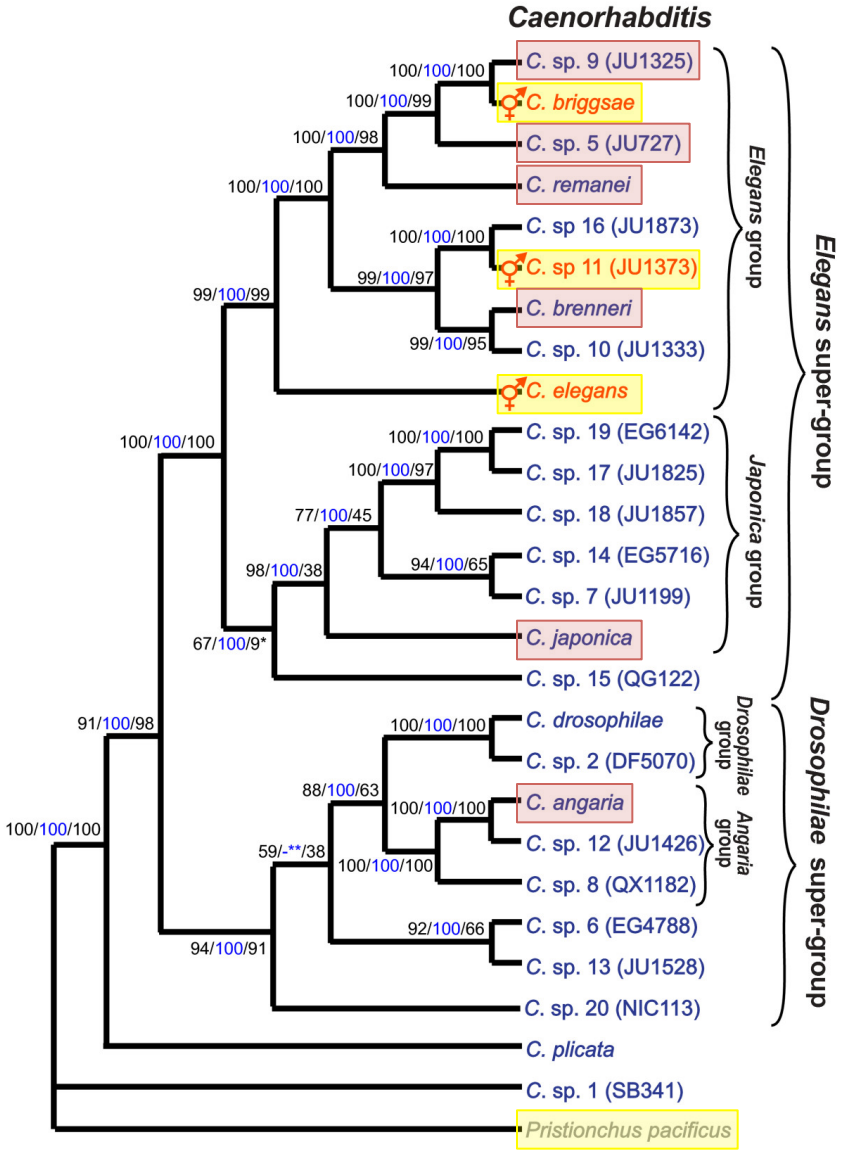
1. *Caenorhabditis angaria* PS1010

2. *Caenorhabditis* spp. 7, 9, 11

3. *Haemonchus contortus*

4. General lessons

Caenorhabditis molecular phylogeny



Seq. genomes:

= ♀ ♂

= ♀ ♂

Ref.: Kiontke et al. (2011), BMC Evol. Biol. *11*, 339.

Conversation at International *C. elegans* meeting, June 2011

Scene: coffee break, right after platform talk
in front of ~200 biologists.

"That was a fine talk about the multiple genomes."

"Thank you!"

"But ... uh ... I have evidence that two of the three new genomes from WashU are cross-contaminated with one another's DNA."

"?!?"

Convincing myself to throw away two genomes

sp. 9 transcriptome produced solely through my hands.

Core enzymes tend to be single-copy/genome.

Core *lipid* enzymes tend to be membrane proteins.

So: in "sp. 7" and "sp. 9" genomes, use TBlastN to find single-copy lipid enzyme genes with strong expression in cDNA:

C33H5.18 (CDP-diglyceride synthase)

pld-1 (Phospholipase D)

ZK669.4 (Lipoamide acyltransferase, mitochon. α-keto acid dehydrog.)

Each case gave <u>both</u> 100% <u>and</u> <100% id. to sp. 9 transcriptome.

Re-sequencing *C.* sp. 9 to 172x coverage

Insert size	Read size	Reads	Total nt	Coverage
335 nt	2x100 nt	372 M	35.6 Gb	178x
Unpaired	100 nt	3.6 M	306 Mb	40x

Used velvet 1.2.03.

Empirically tested k= 75 through 89; k = 85 was theoretical and empirical optimum.
Empirically tested various –exp_cov and –cov_cutoff statistics:
~145 Mb stably appeared; another ≤30 Mb variably appeared

CEGMA and cDNA mapping supported 'auto' options for –exp_cov and –cov_cutoff.

Final assembly:

145 Mb, scaffold N50 = 44.5 kb.

Sequencing: Barbara Meyer et al. (2012).

Second conversation at International *C. elegans* meeting, June 2011

Scene: Q+A session.

"Your scaffold N50 for sp. 11 is 20 Mb.
Isn't that a little ... *large*?"

"Well, sp. 11 is a hermaphroditic species without heterozygosity. So I would assume that WashU was able to assemble it much more thoroughly than the male-female species."

"OK."

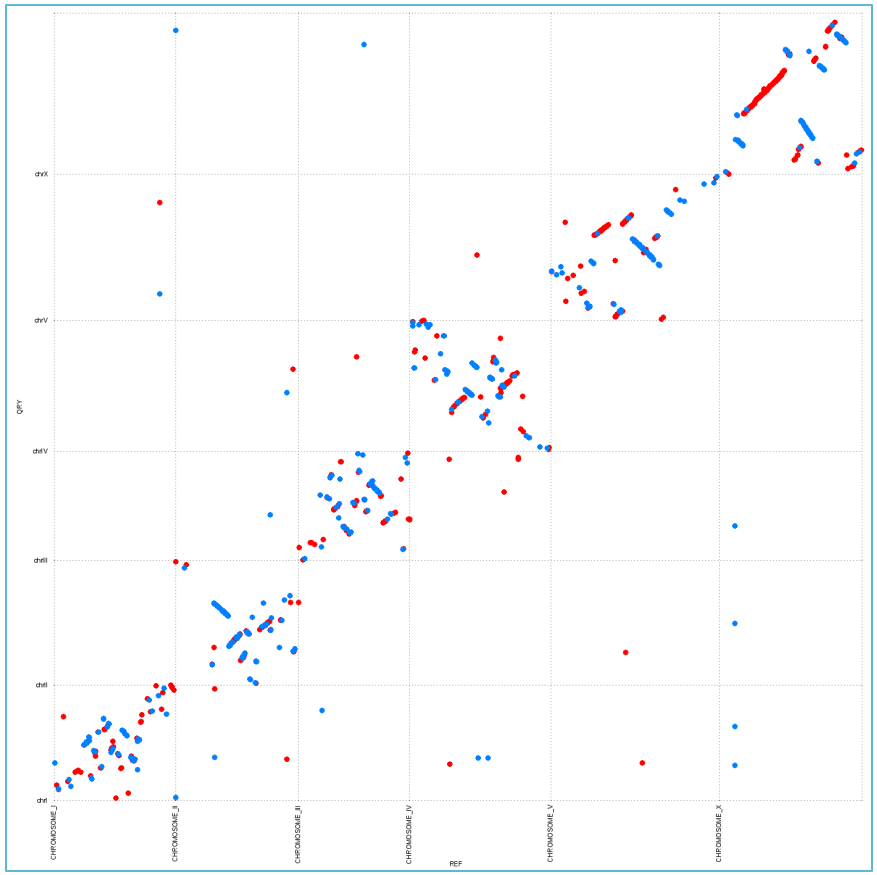
Third conversation at CSHL meeting on Evolution of *Caenorhabditis*, April 2012

Scene: Q+A session.

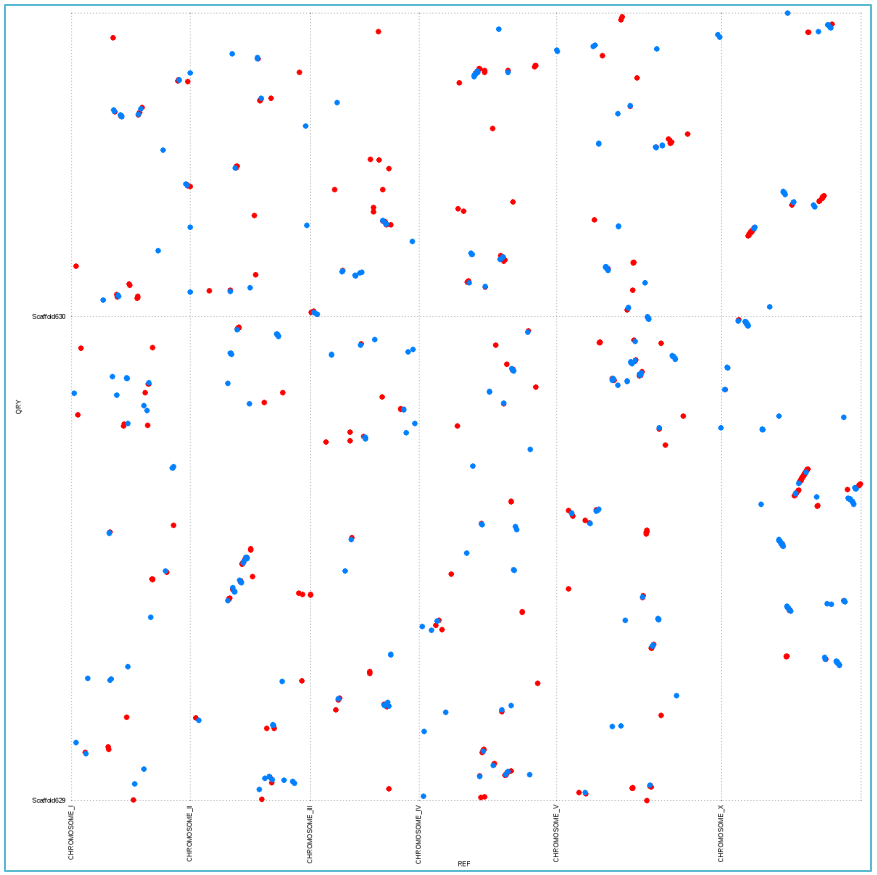
"Your scaffold N50 for sp. 11 is still 20 Mb.
Are you *sure* you believe that?"

"Hm. Let me think about it."

Synteny revealed sp. 11 over-assembly



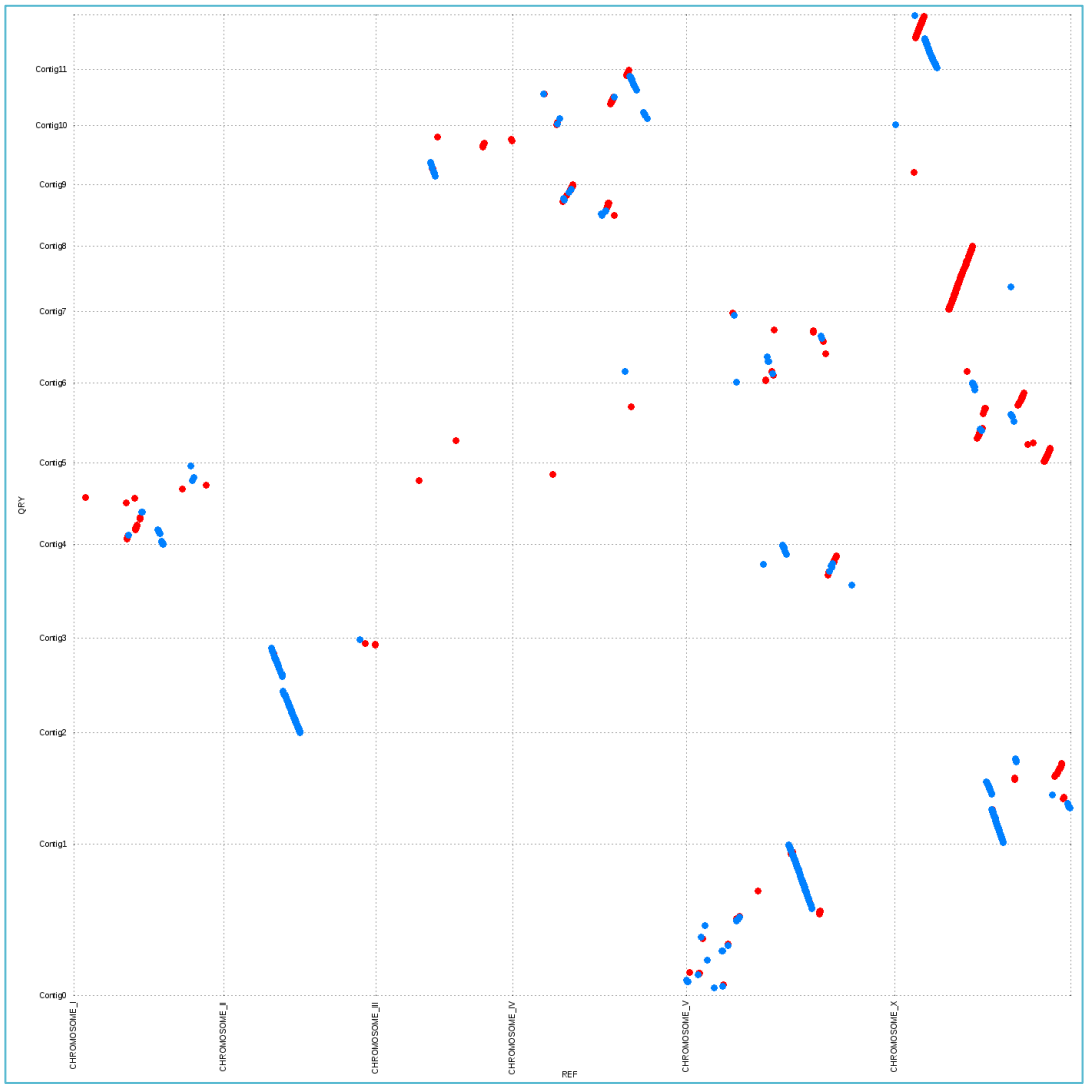
elegans vs. *briggsae*



elegans vs. sp. 11

Ref.: Ross et al. (2011), PLoS Genet. 7, e1002174..

An earlier sp. 11 assembly works far better



elegans vs. previous sp. 11

Assembly source: Patrick Minx et al. (2012).

Current *Caenorhabditis* genomes

Species	Group	Sexes	Size (Mb)	Scaf.	N50	Release
elegans	Elegans	Herm.	100	7	17.5 Mb	[ongoing]
briggsae	Elegans	Herm.	108	12	17.5 Mb	April 2011
remanei	Elegans	M./Fem.	145	3.7 K	436 kb	July 2007
brenneri	Elegans	M./Fem.	190	3.3 K	382 kb	May 2008
japonica	Japonica	M./Fem.	166	19 K	94 kb	Aug. 2011
sp. 5	Elegans	M./Fem.	132	15 K	25 kb	March 2012
sp. 9	Elegans	M./Fem.	145	18 K	44.5 kb	April 2012
sp. 11	Elegans	Herm.	88	11 K	950 kb	June 2012
angaria	Angaria	M./Fem.	106	34 K	80 kb	June 2012

Predicted protein-coding genes

Species	Sexes	Genes	Prediction method	Release
elegans	Herm.	20,520	Manual curation	[ongoing]
briggsae	Herm.	21,936	nGasp + some manual	April 2011+
remanei	M./Fem.	31,471	nGasp + some manual	May 2008+
brenneri	M./Fem.	30,667	nGasp + AU. + manual	May 2008+
japonica	M./Fem.	29,962	AUGUSTUS + RNA-seq	Aug. 2011+
sp. 5	M./Fem.	34,696	AUGUSTUS	March 2012
sp. 9	M./Fem.	41,938	AUGUSTUS + RNA-seq	April 2012
sp. 11	Herm.	26,212	AUGUSTUS	June 2012
angaria	M./Fem.	29,760	AUGUSTUS + RNA-seq	June 2012

Overview

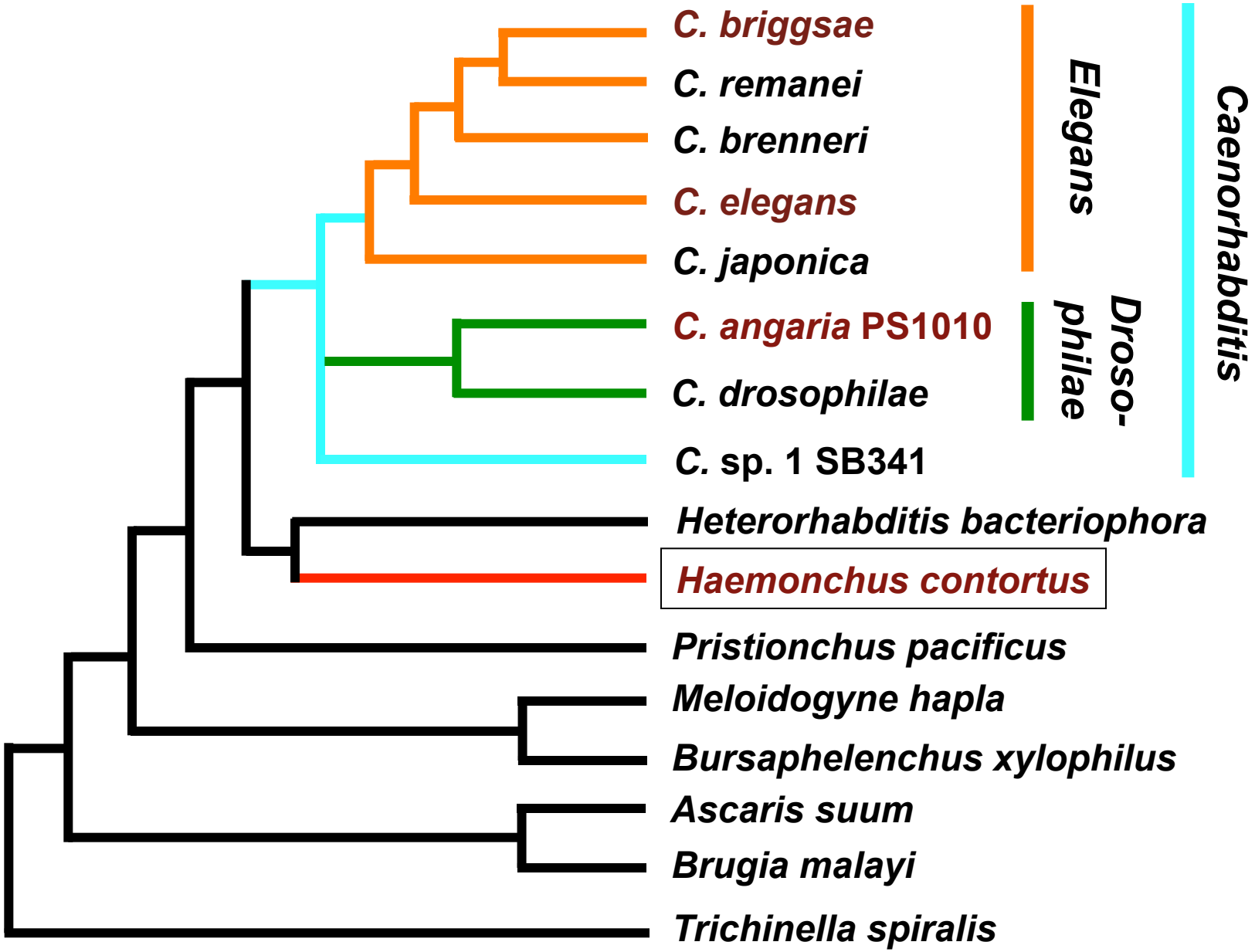
1. *Caenorhabditis angaria* PS1010

2. *Caenorhabditis* spp. 7, 9, 11

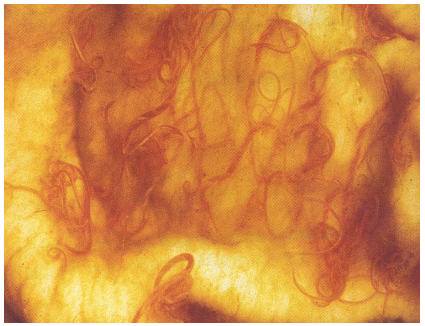
3. *Haemonchus contortus*

4. General lessons

Nematodes outside of *Caenorhabditis*



H. contortus life cycle

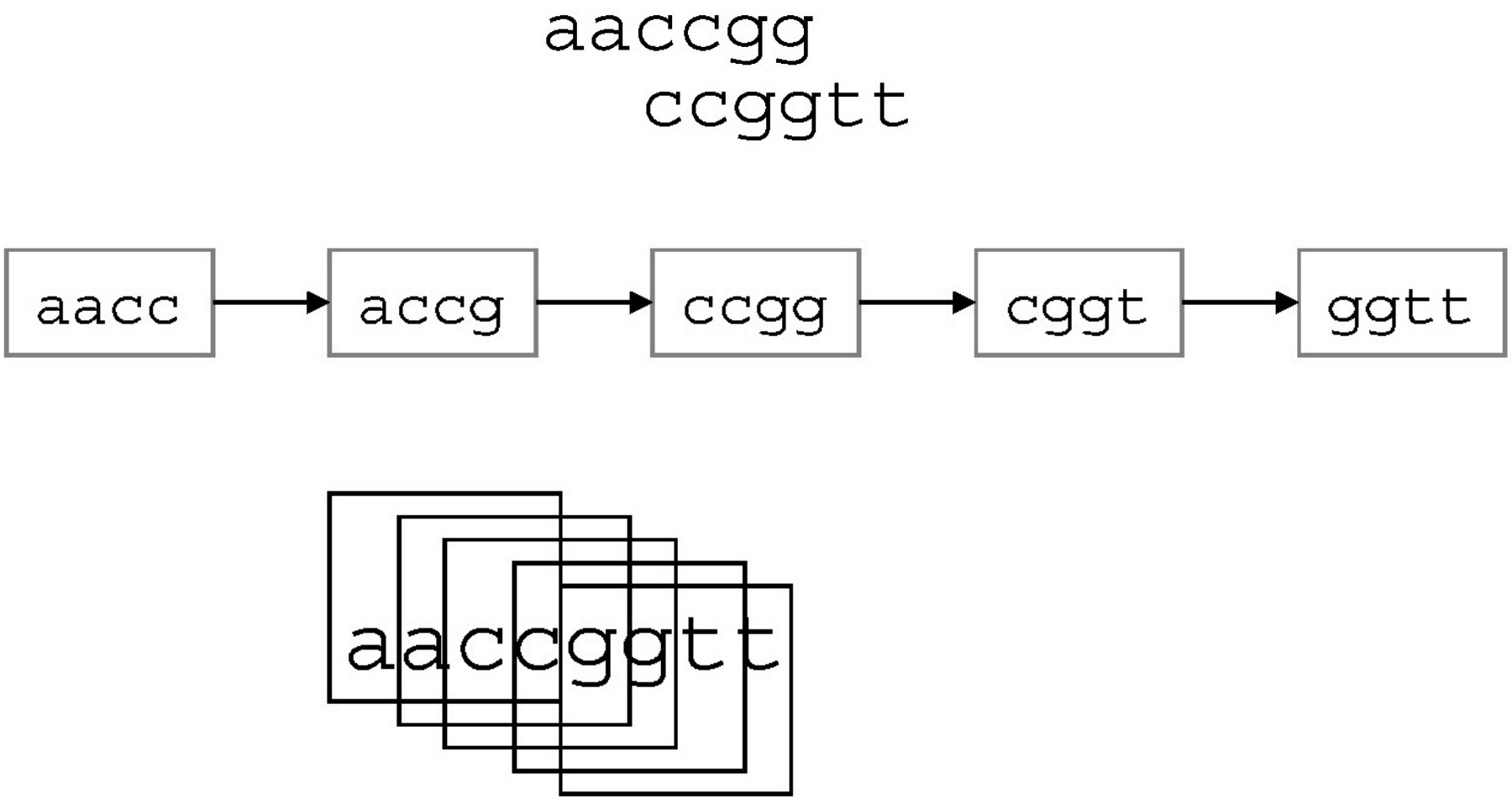


Refs.: Nikolaou and Gasser (2006), Int. J. Parasitol. *36*, 859-868;
Prichard and Geary (2008), Nature *452*, 157-158.

Sequencing *H. contortus* to 186x coverage

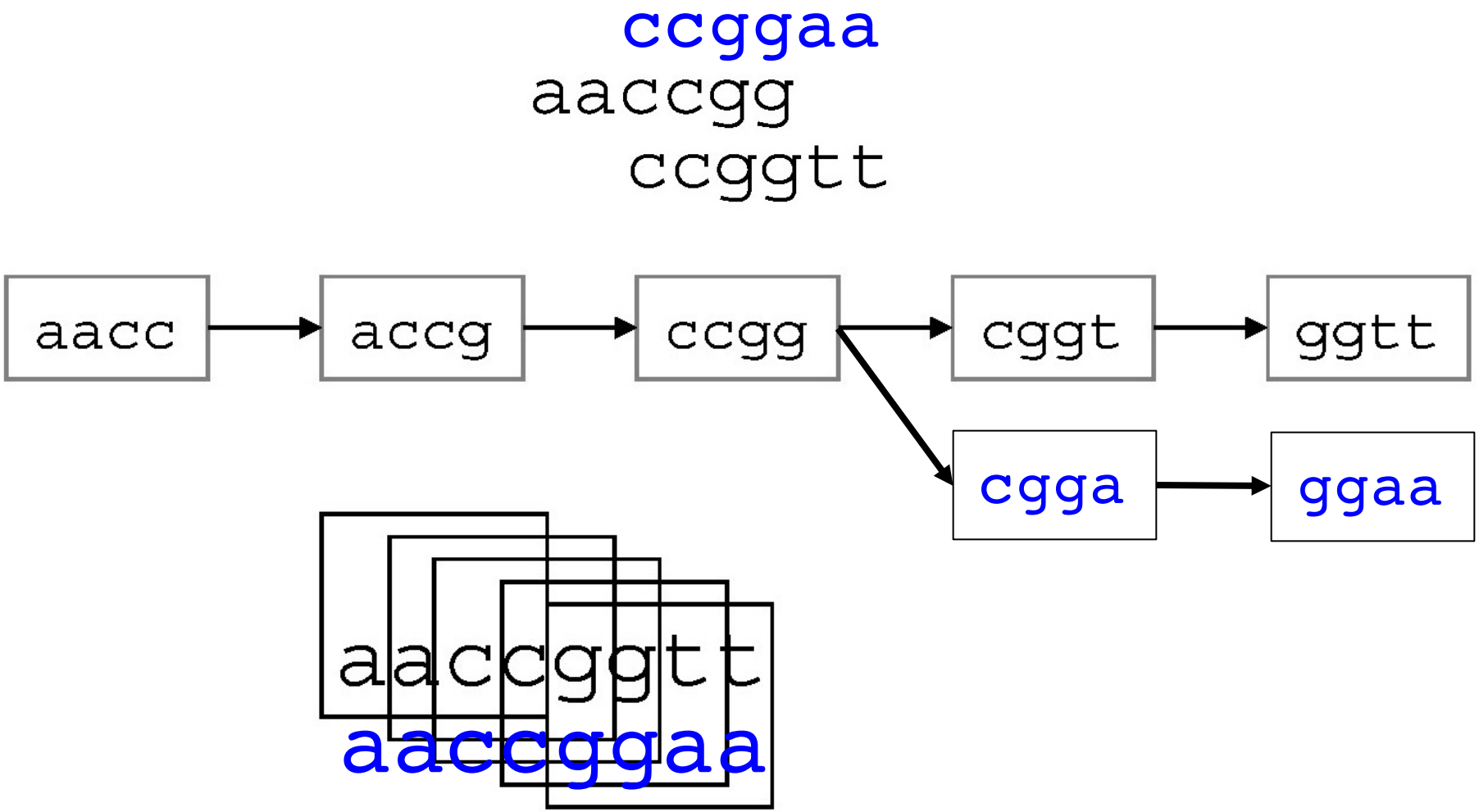
Insert size	Read size	Reads	Total nt	Coverage
300 nt	2x75 nt	107 M	8.0 Gb	25.3x
500 nt	2x75 nt	170 M	12.7 Gb	40.3x
500 nt	2x100 nt	235 M	22.9 Gb	72.6x
2 kb	2x49 nt	87 M	4.2 Gb	13.5x
5 kb	2x49 nt	45 M	2.2 Gb	6.9x
10 kb	2x49 nt	38 M	1.9 Gb	6.0x
Unpaired	48-100 nt	94 M	6.8 Gb	21.7x

Next-gen. DNA sequencing uses small "words"



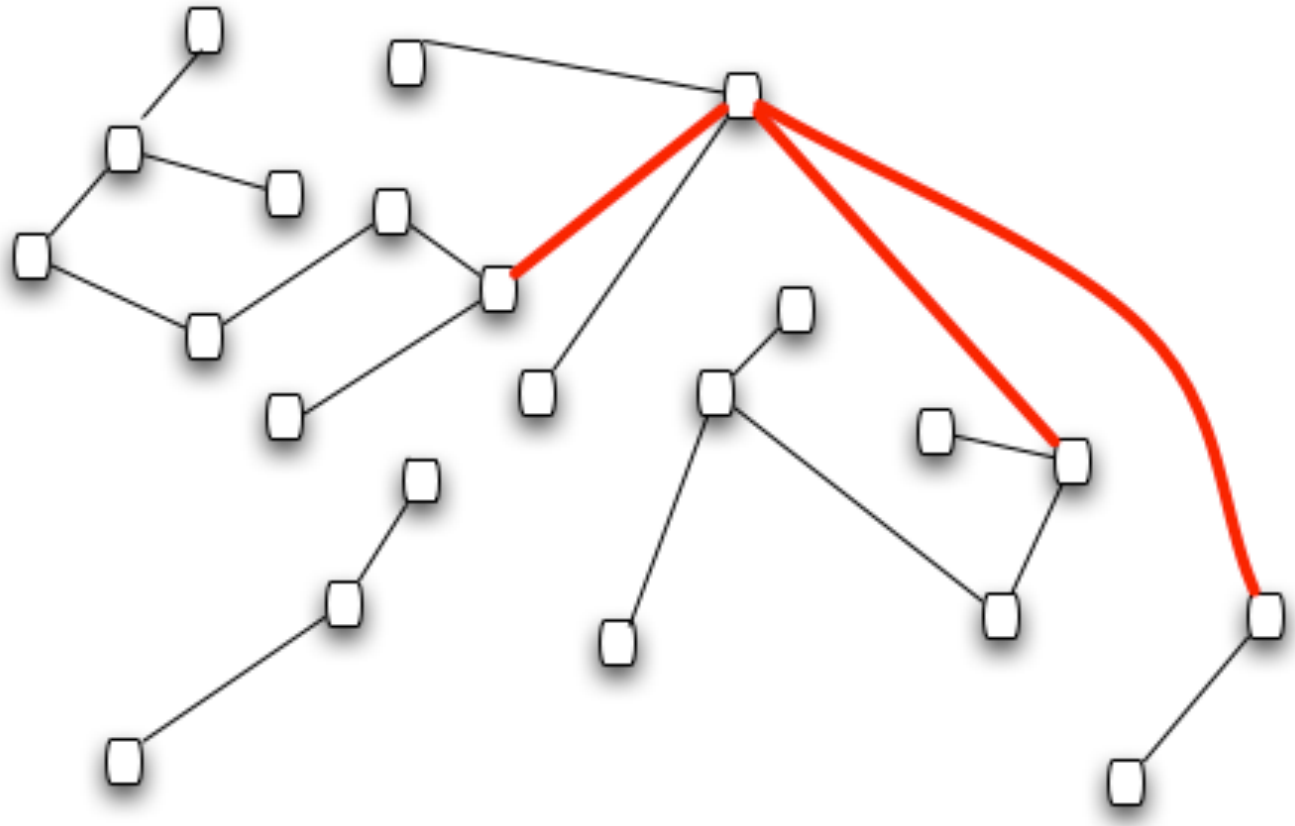
Ref.: Miller et al. (2010), Genomics *95*, 315-327.

Assembly gets harder when the data get messier

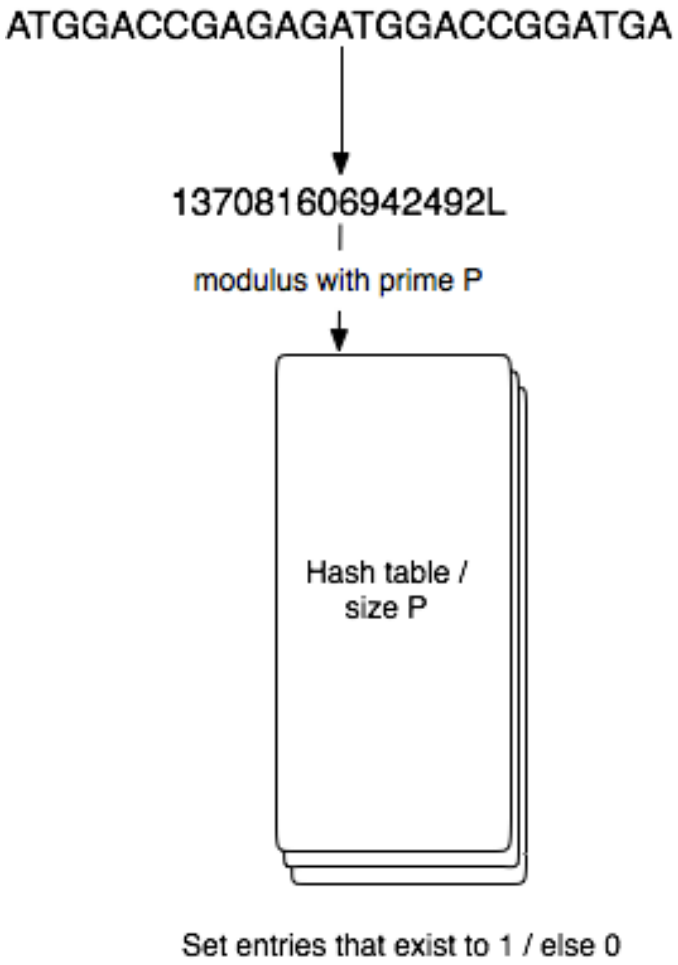


Ref.: Miller et al. (2010), Genomics *95*, 315-327.

One overconnected read kills a whole assembly



khmer: store graph nodes in a Bloom filter



Draw graphs normally (in space of adjacent k-mers).

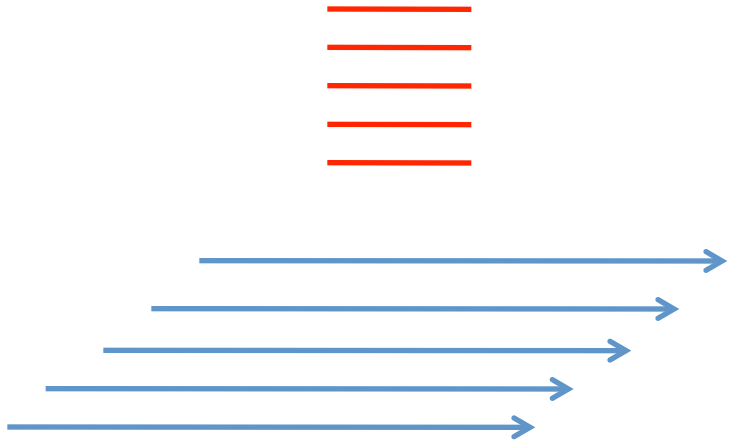
But, track the presence or absence of individual nodes with a Bloom filter (a modulus-based hash table without collision tracking).

Preprints: *arxiv.org/abs/**1112.4193***, *arxiv.org/abs/**1203.4802***

Source code: *github.com/ctb/**khmer***.

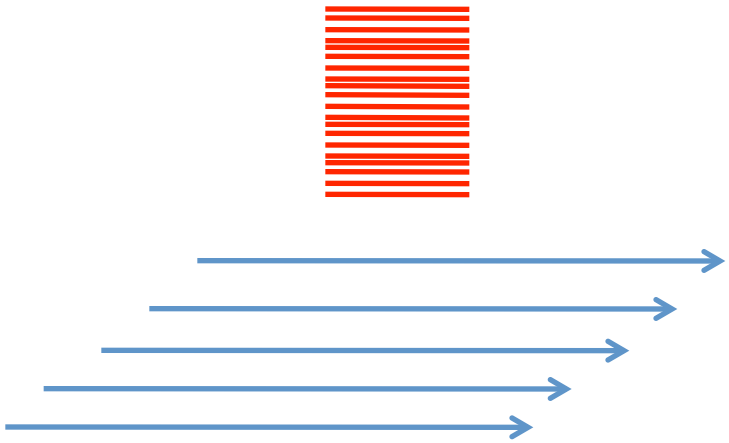
Brown lab at MSU: *ged.msu.edu*

Efficient k-mer counting allows "digital normalization" of reads



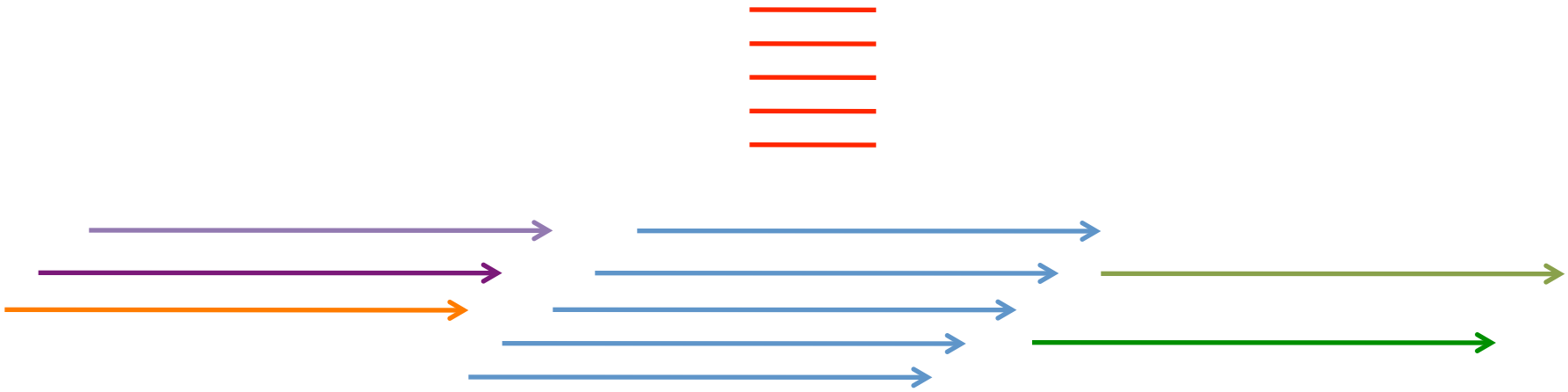
In a perfect world, doing 5x coverage of a genome would mean that each read's k-mers happened 5x.

Efficient k-mer counting allows "digital normalization" of reads



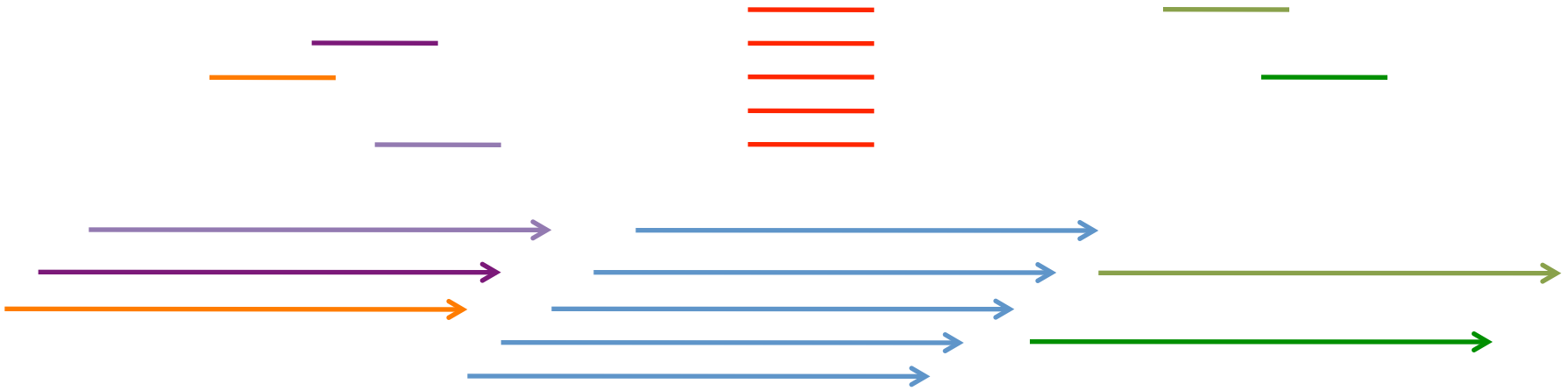
In real life, doing 5x coverage of a repeat means that repeat's k-mers happen 5**N** times, where **N** is much too large.

Efficient k-mer counting allows "digital normalization" of reads



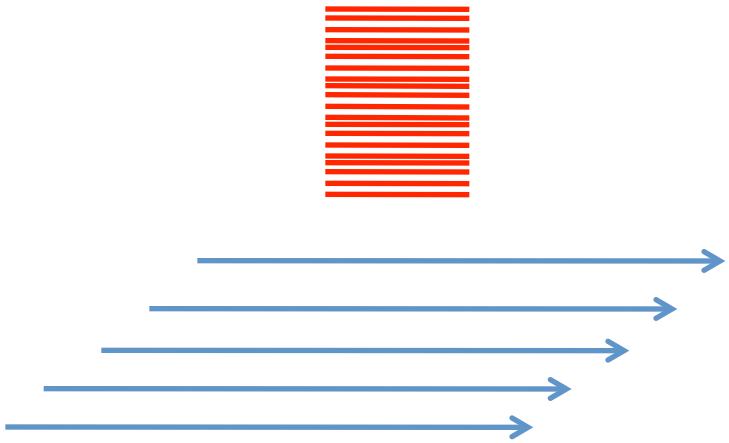
Alternatively, sequencing errors in your reads ...

Efficient k-mer counting allows "digital normalization" of reads



Alternatively, sequencing errors in your reads ...
... can create unique (erroneous) k-mers.
These exist nowhere on planet Earth except in your data, and in your CPU cycles, and your RAM...

Efficient k-mer counting allows "digital normalization" of reads

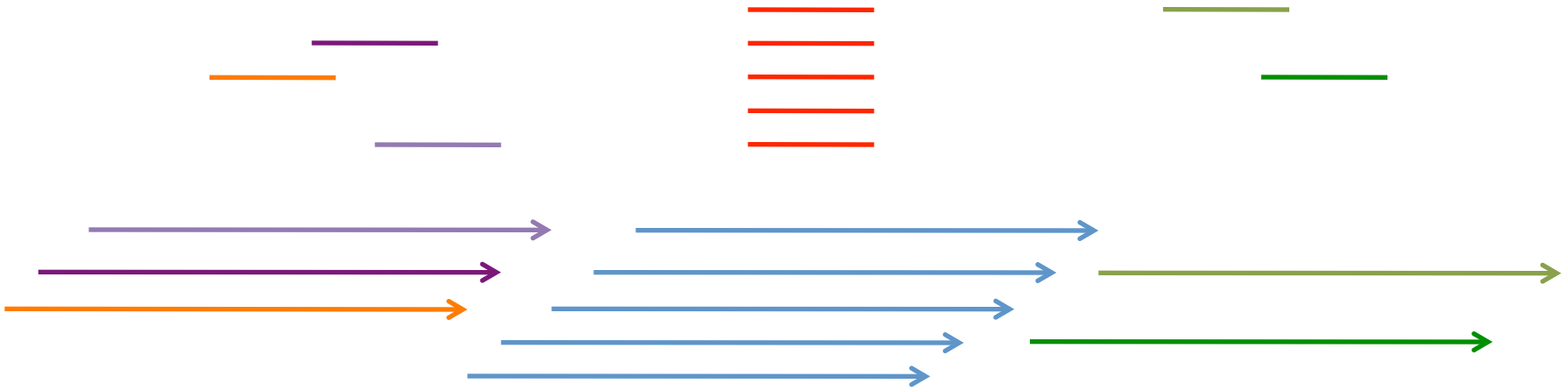


But because we have a way to count k-mers which is economical and fast, we can keep track of when a k-mer becomes too abundant, and start ignoring reads which contain it.

Efficient k-mer counting allows "digital normalization" of reads

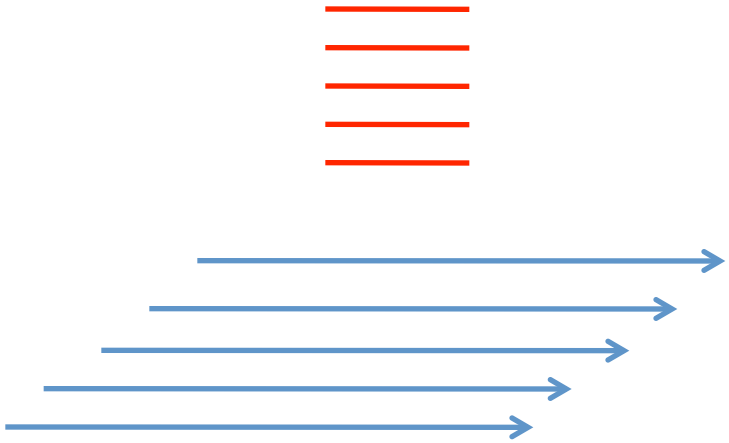
But because we have a way to count k-mers which is economical and fast, we can keep track of when a k-mer becomes too abundant, and start ignoring reads which contain it.
In practice, that censors repeats before assembly.

Efficient k-mer counting allows "digital normalization" of reads



We can also require that a read's k-mers must all have been observed at least 2x. If any reads have unique k-mers ...

Efficient k-mer counting allows "digital normalization" of reads



We can also require that a read's k-mers must all have been observed at least 2x. If any reads have unique k-mers ...
... we consider them likely to be noise, and discard them.
The assembler never wastes its time on them.

A usable genome assembly for *H. contortus*!

Assembly type	Total (Mb)	Scaffolds	Max. scaffold size	N50 (kb)
Genomic, k=41	725	284 K	1.24 Mb	121
Top fraction	315	1.2 K	[1.24 Mb]	265

Assembly time, 4.5 hours, rather than ∞ hours.

But, the actual genome size is 315 ± 25 Mb.

So, what about that ~2-fold excess size?

Digital normalization allows very small DNA 'words', with k=21 instead of k=41

Assembly type	Total (Mb)	Scaffolds	Max. scaffold size	N50 (kb)
Genomic, k=41	725	284 K	1.24 Mb	121
Top fraction	315	1.2 K	[1.24 Mb]	265
Genomic, k=21	493	195 K	815 kb	109
Top fraction	315	2.0 K	[815 kb]	181

k=21 reduces excess DNA from 130% to 57%.

92% of egg, L4 cDNA maps (vs. 99% for k=41).

k=21 assembly time was 9 days rather than 4.5 hours.

The (highly provisional) genome contents

~16,000 protein-coding genes currently predicted
(noticeably below free-living *Caenorhabditis*, *Pristionchus*;
if real, perhaps due to parasitism?)

~5,400 orthology groups between *H. contortus*
and published nematodes

Larger scaffolds give best BlastP hits to nematode proteins

Smaller ones give hits to *Prevotella ruminicola*, etc. [!]

Systematic filtering will clearly reduce more excess;
not clear yet by how much

Overview

1. *Caenorhabditis angaria* PS1010

2. *Caenorhabditis* spp. 7, 9, 11

3. *Haemonchus contortus*

4. General lessons

Begin and end with checks for basic quality

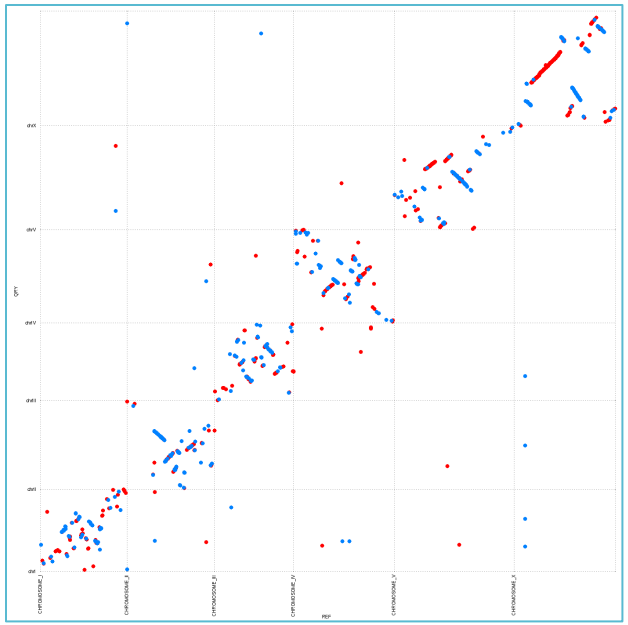
Living organisms sit in a soup of microbes

Microbial contamination slowed both *C. angaria* and *H. contortus*

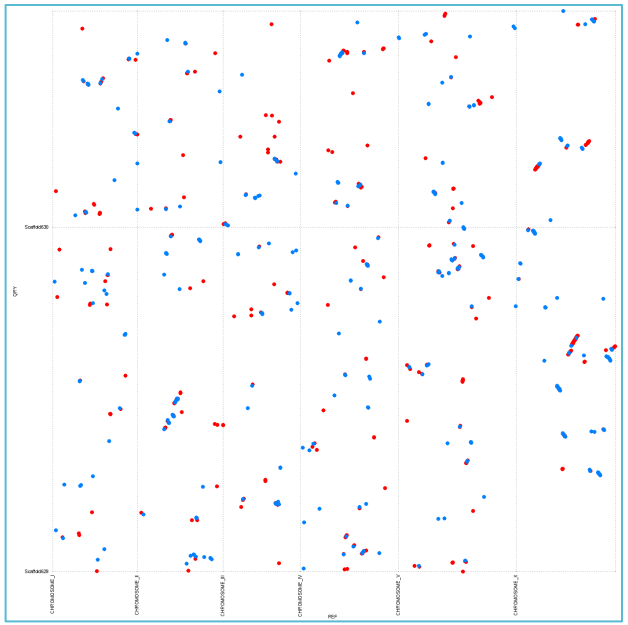
Over-assembly can happen

In recent case of *C*. sp. 11, detected with chromosomal synteny

cDNA from RNA-seq might be another reality check



elegans vs. *briggsae*



elegans vs. sp. 11

Sequence data are a moving target

Barbara Meyer (Berkeley) wants shotgun genomes with *chromosome-sized* scaffolds

Proactively, she arranged Pacific Biosciences re-sequencing of seven *Caenorhabditis* genomes

At any moment, could get handed new data for *C. angaria* and six other species, to make third-gen genome assemblies

Good problem to have, but it's emblematic of why genomics requires a tight loop of plan-execute-analyze

How do you get biology out of your genome?

"Begin with the end in mind." --Stephen Covey

ultraconserved cis-regulatory DNA (*C. angaria*)
hermaphrodite-specific DNA (*Caenorhabditis* spp.)
drug targets for one billion sick humans (*Haemonchus* et al.)

"Given sufficient eyes, all bugs are shallow." --Eric Raymond

Give talks to intelligent critics well before you publish.
Be eclectic in what you use. "Naive" or "obsolete" tools or data can be surprisingly useful.

"There is no perfectly shaped part of the motorcycle and never will be, but when you come as close as these instruments take you, remarkable things happen, and you go flying across the countryside under a power that would be called magic if it were not so completely rational in every way." –Robert Pirsig

Persistent attention to quality pays off.

Thanks:

Ali Mortazavi	*C. angaria* DNA assembly and RNAseq RNAPATH; TBA/multiz align.
Bronwyn Campbell, Neil Young, Ross Hall	*H. contortus* RNA extraction and analysis
Robin Gasser	*H. contortus* genomic DNA extraction
Titus Brown, Adina Howe, Jason Pell	*khmer* software and filtering
Brian Williams	cDNA library construction
Igor Antoshechkin	Optimized sequencing protocols
Jacobs GC and BGI	Illumina sequencing
Barbara Meyer	*C.* sp. 9 genomic sequencing
Pat Minx	*C.* sp. 11 re-assembly

Good luck!