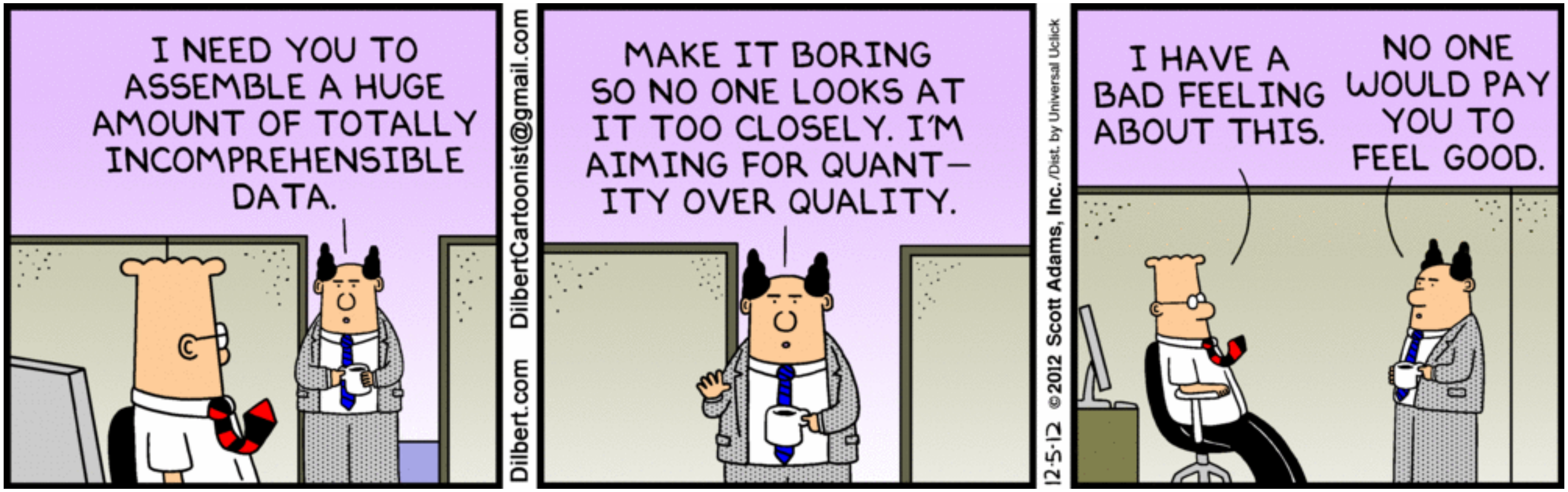
Genome and transcriptome of the zoonotic hookworm *Ancylostoma ceylanicum*

Erich Schwarz, Cornell

MSU NGS course, June 2013

How to avoid *this* situation (hopefully)

Erich Schwarz, Cornell


I NEED YOU TO ASSEMBLE A HUGE AMOUNT OF TOTALLY INCOMPREHENSIBLE DATA.
Dilbert.com DilbertCartoonist@gmail.com
MAKE IT BORING SO NO ONE LOOKS AT IT TOO CLOSELY. I'M AIMING FOR QUANT–ITY OVER QUALITY.
12-5-12 ©2012 Scott Adams, Inc./Dist. by Universal Uclick
I HAVE A BAD FEELING ABOUT THIS.
NO ONE WOULD PAY YOU TO FEEL GOOD.

Scott Adams, *Dilbert*, 5 Dec 2012

Overview

1. What hookworms are and why we care

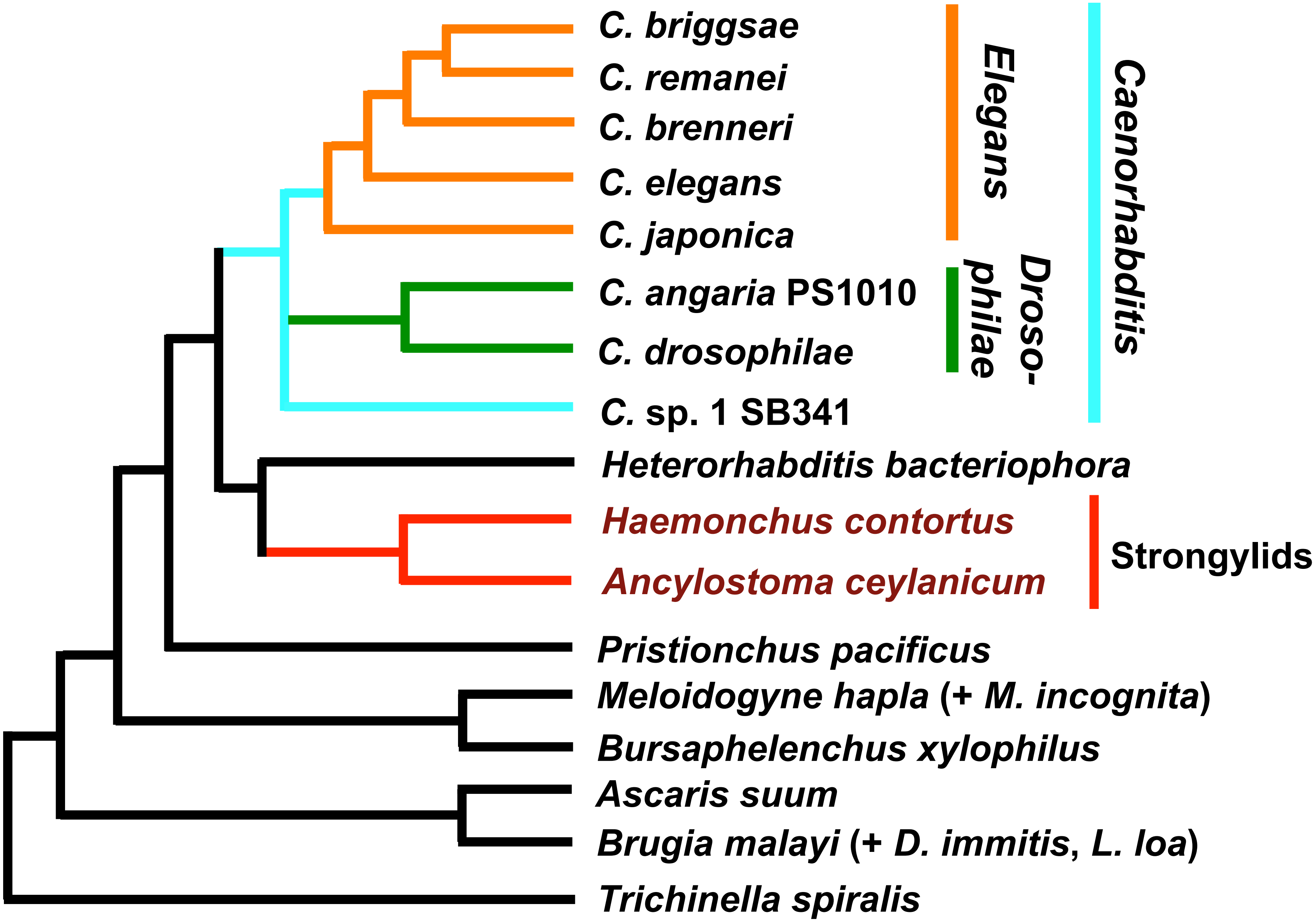
2. *Haemonchus contortus* adventures

3. *Ancylostoma ceylanicum* genome and transcriptome

4. Vaccine targets: ASPRs, and horizontally transmitted genes

5. Some thoughts on 'descriptive genomics'

Nematode phylogeny



C. briggsae

C. remanei

C. brenneri

C. elegans

C. japonica

C. angaria PS1010

C. drosophilae

C. sp. 1 SB341

Heterorhabditis bacteriophora

Haemonchus contortus

Ancylostoma ceylanicum

Pristionchus pacificus

Meloidogyne hapla (+ *M. incognita*)

Bursaphelenchus xylophilus

Ascaris suum

Brugia malayi (+ *D. immitis, L. loa*)

Trichinella spiralis

Elegans

Drosophilae

Caenorhabditis

Strongylids

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

Ancylostoma and *Necator*: worldwide scourges

A. duodenale and *N. americanus* infect
up to 740 million human beings.

Infection can begin in childhood and can last for life.
It is generally not fatal, but can be highly debilitating.

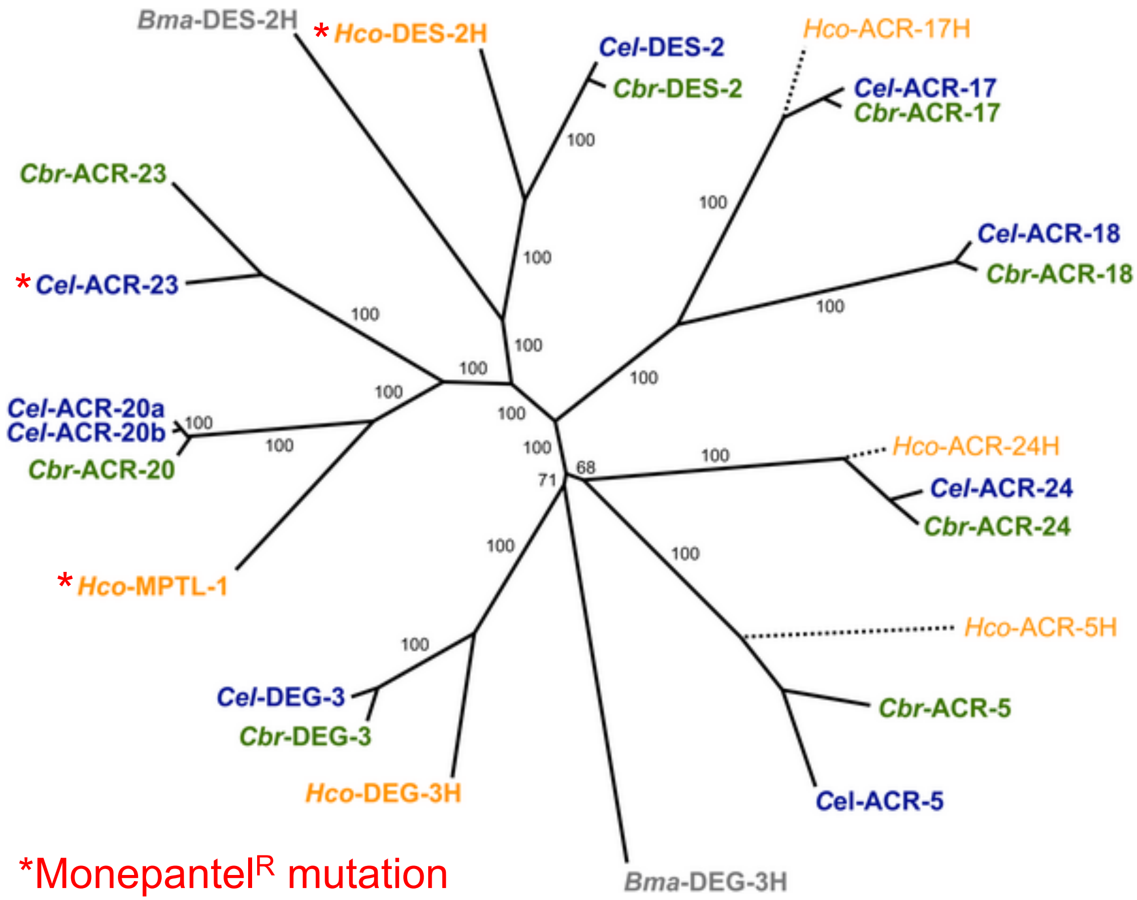
Existing drugs only partially cure hookworm infections.
No vaccines exist.

A. ceylanicum is a relatively minor human parasite, but is able to infect both humans and other animals (e.g., golden hamsters). It is thus the closest thing we have to a model hookworm.

Refs.: Bethony et al. (2006), Lancet *367*, 1521-1532; Brooker et al. (2004), Adv. Parasitol. *58*, 197-288; Conlan et al. (2011), Vet. Parasitol. *182*, 22-40; Hotez et al. (2009), Lancet *373*, 1570-1575; Keiser and Utzinger (2010), Adv. Parasitol. *73*, 197-230; Schneider et al. (2011), Hum. Vaccin. *7*, 1234-1244.

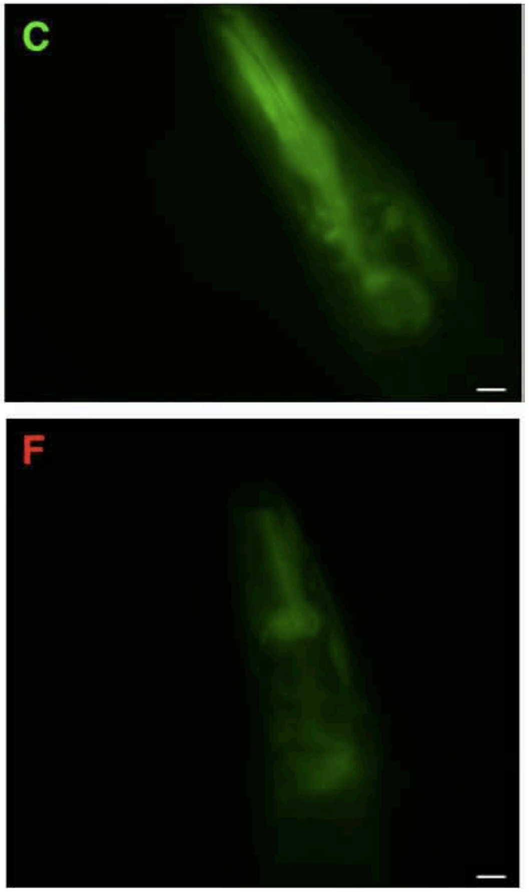
Strongylid genes resemble *C. elegans*

Acetylcholine receptor genes from *C. elegans*, *C. briggsae*, *H. contortus*, and *B. malayi*:



*Monepantel^{R} mutation

Ce-ant-1.1



Hc-ant-1.1

Refs.: Rufener et al. (2009), PLoS Pathog. *5*, e1000380;
Hu et al. (2010), Biotechnol. Adv. *28*, 49-60; Laing et al. (2011), PLoS One *6*, e23216.

Overview

1. What hookworms are and why we care

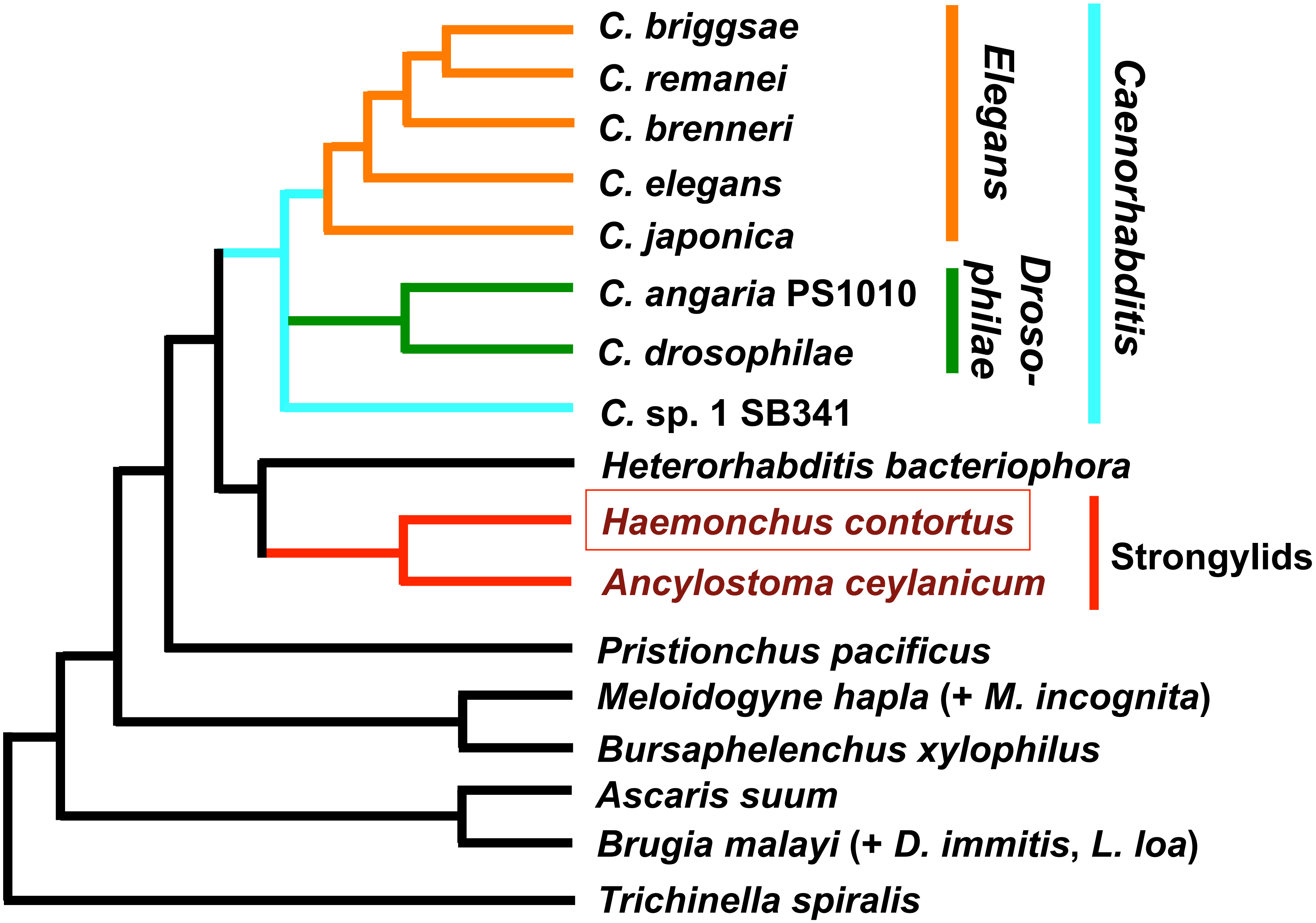
2. *Haemonchus contortus* adventures

3. *Ancylostoma ceylanicum* genome and transcriptome

4. Vaccine targets: ASPRs, and horizontally transmitted genes

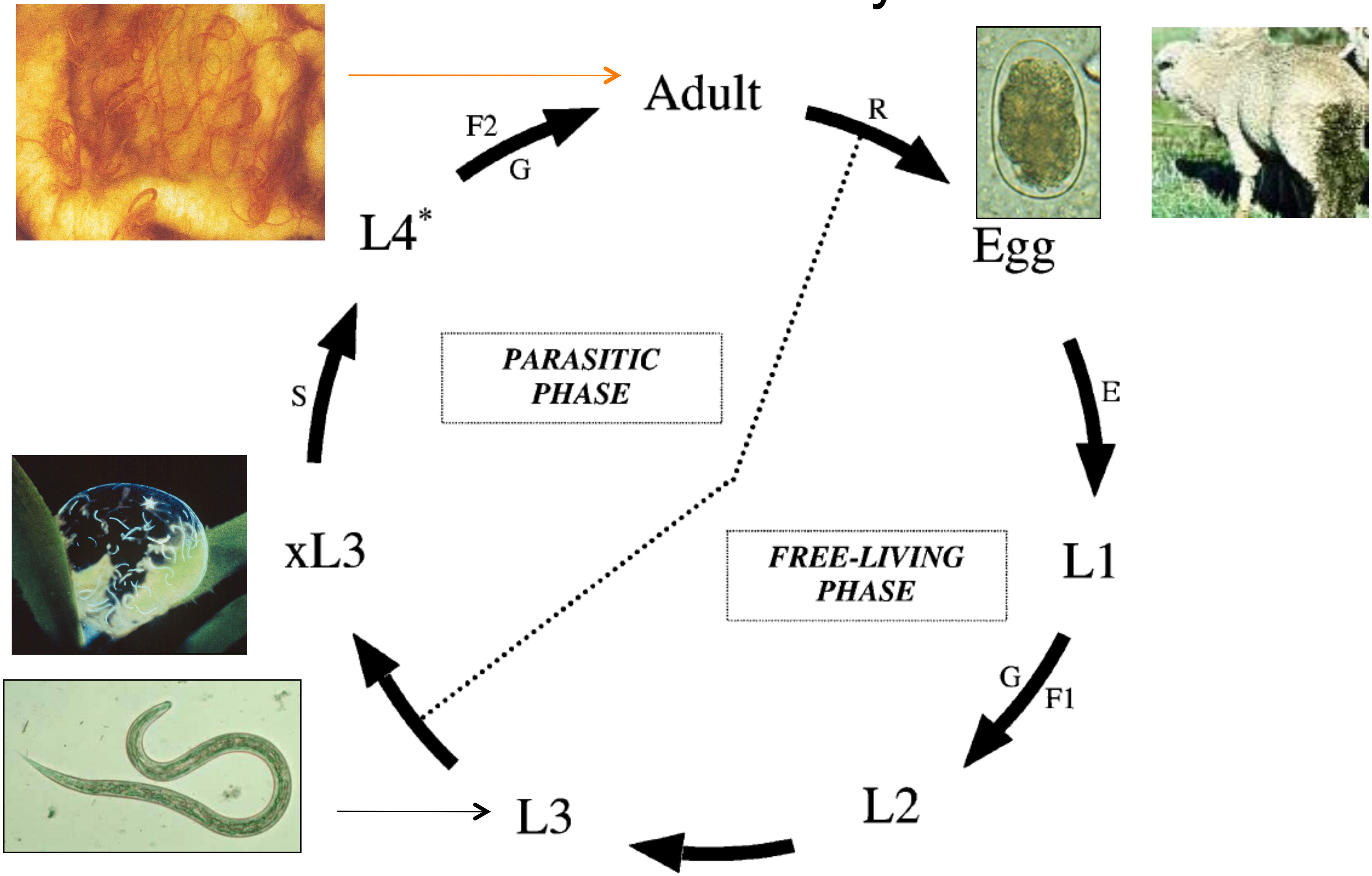
5. Some thoughts on 'descriptive genomics'

Nematode phylogeny



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

H. contortus life cycle



Adult

R

Egg

E

L1

G

F1

L2

L3

xL3

S

$L4^*$

F2

G

PARASITIC PHASE

FREE-LIVING PHASE

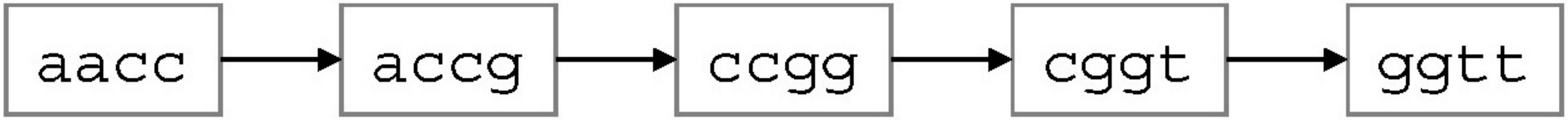
Refs.: Nikolaou and Gasser (2006), Int. J. Parasitol. *36*, 859-868;
Prichard and Geary (2008), Nature *452*, 157-158.

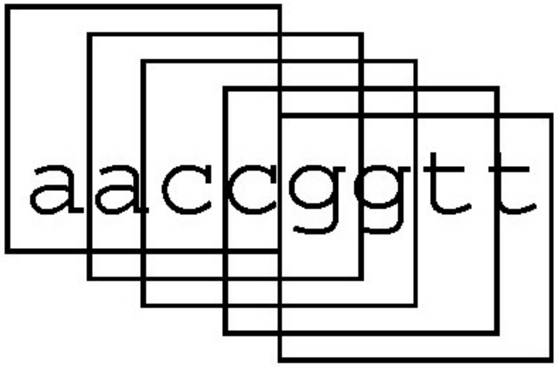
Sequencing *H. contortus* to 186x coverage

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"

```
aaccgg
  ccggtt
```

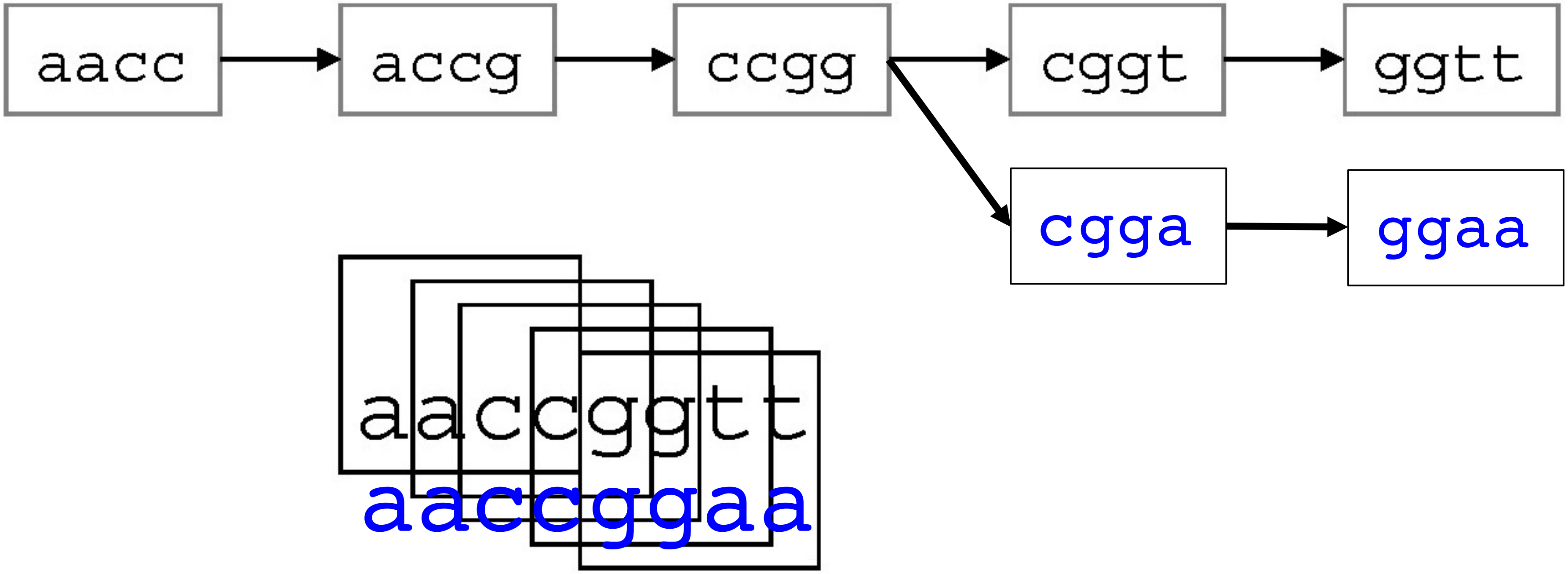




Ref.: Miller et al. (2010), Genomics *95*, 315-327.

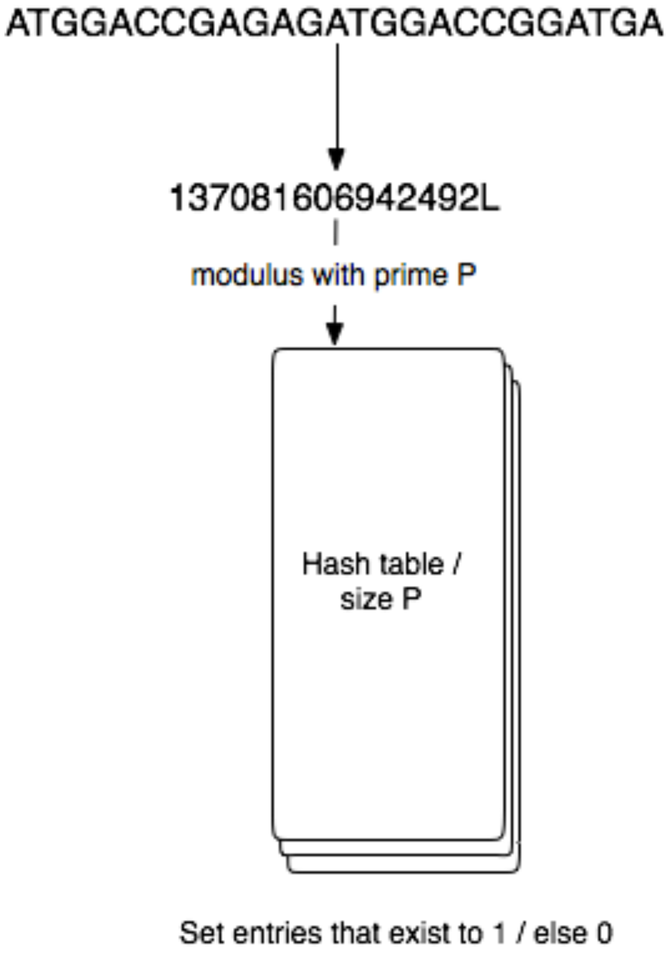
Assembly gets harder when the data get messier

```
  ccggaa
aaccgg
  ccggtt
```



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

khmer: store graph nodes in a Bloom filter

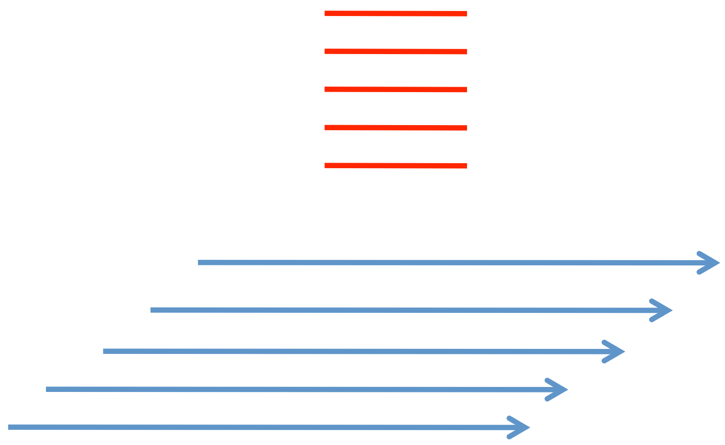

ATGGACCGAGAGATGGACCGGATGA
137081606942492L
modulus with prime P
Hash table / size P
Set entries that exist to 1 / else 0

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).

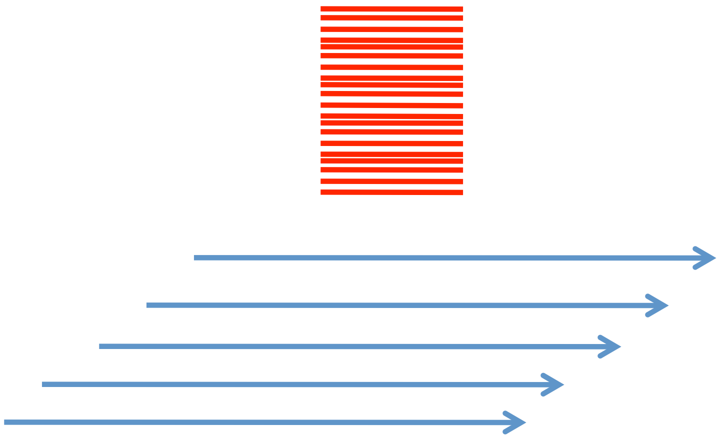
Preprints: *arxiv.org/abs/**1112.4193***, *arxiv.org/abs/**1203.4802***
Source code: *github.com/ctb/**khmer***.
Brown lab at MSU: *ged.msu.edu*

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



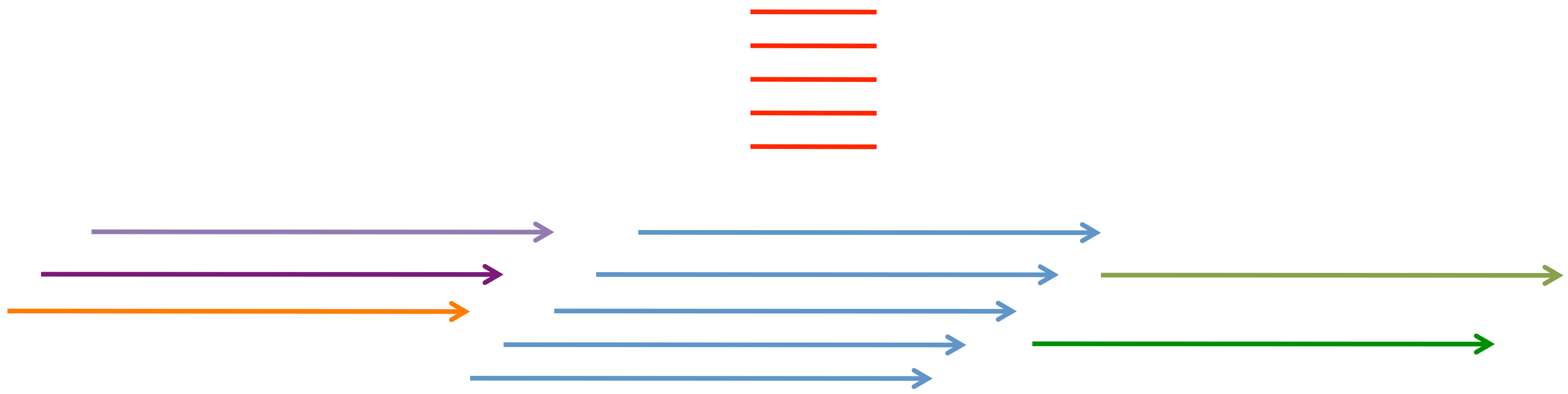
In a perfect world, doing 5x coverage of a genome would mean that each read's k-mers happened 5x.

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



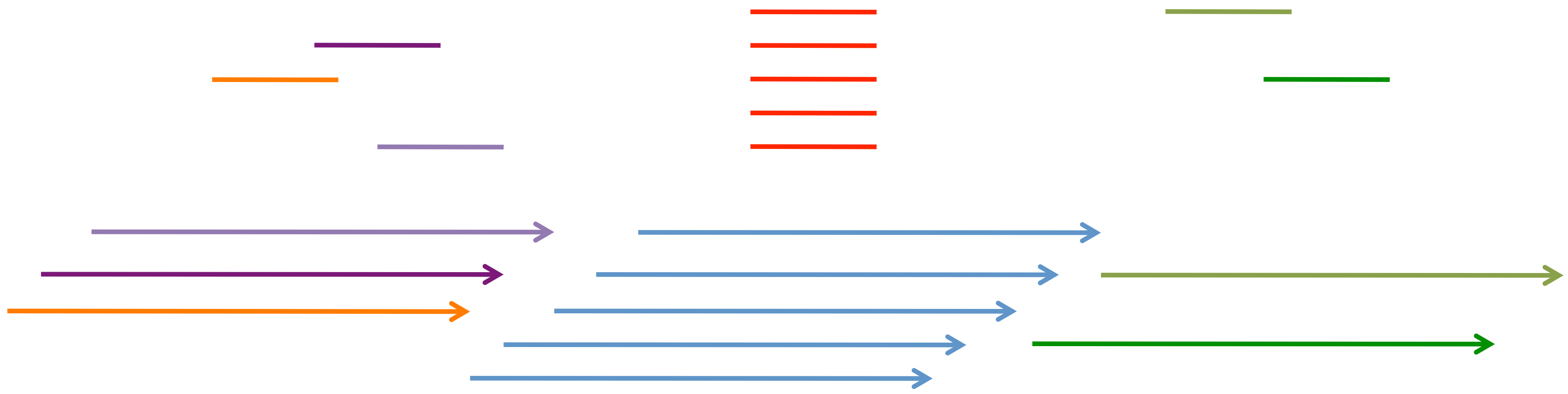
In real life, doing 5x coverage of a repeat means that repeat's k-mers happen 5**N** times, where **N** is much too large.

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



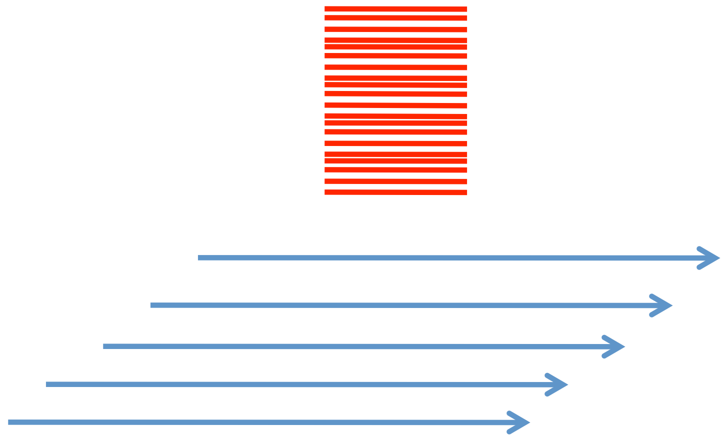
Alternatively, sequencing errors in your reads ...

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

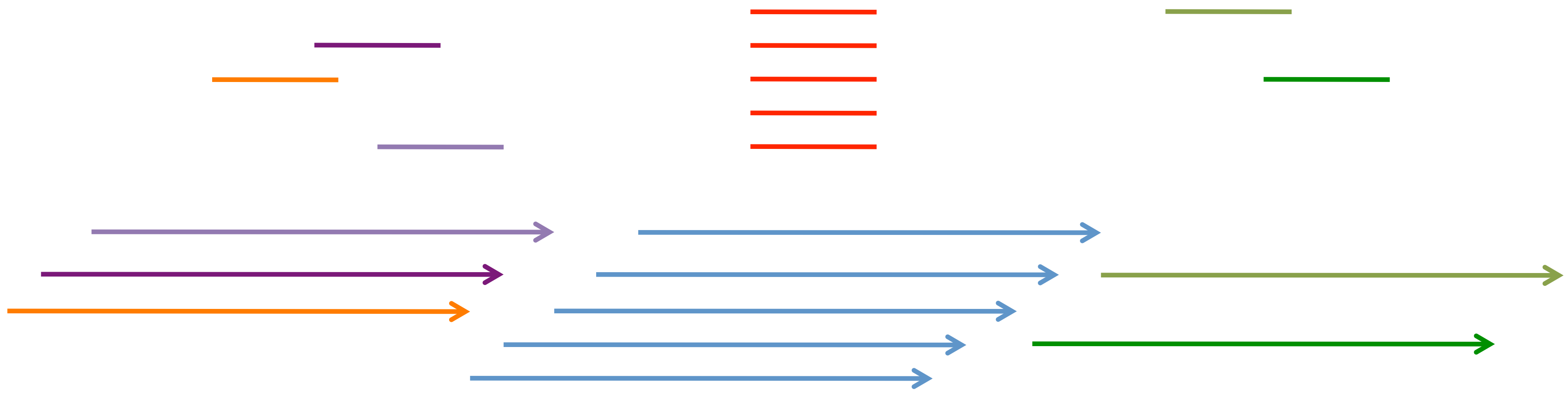


But because we have a way to count k-mers which is economical and fast, we can keep track of when a k-mer becomes too abundant, and start ignoring reads which contain it.

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

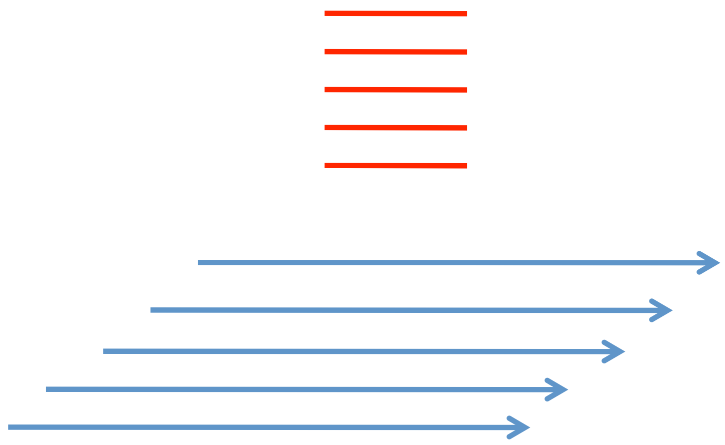
But because we have a way to count k-mers which is economical and fast, we can keep track of when a k-mer becomes too abundant, and start ignoring reads which contain it.
In practice, that censors repeats before assembly.

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 was 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 allowed 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 reduced excess DNA from 130% to 57%.

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

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

So all was well? Not quite.

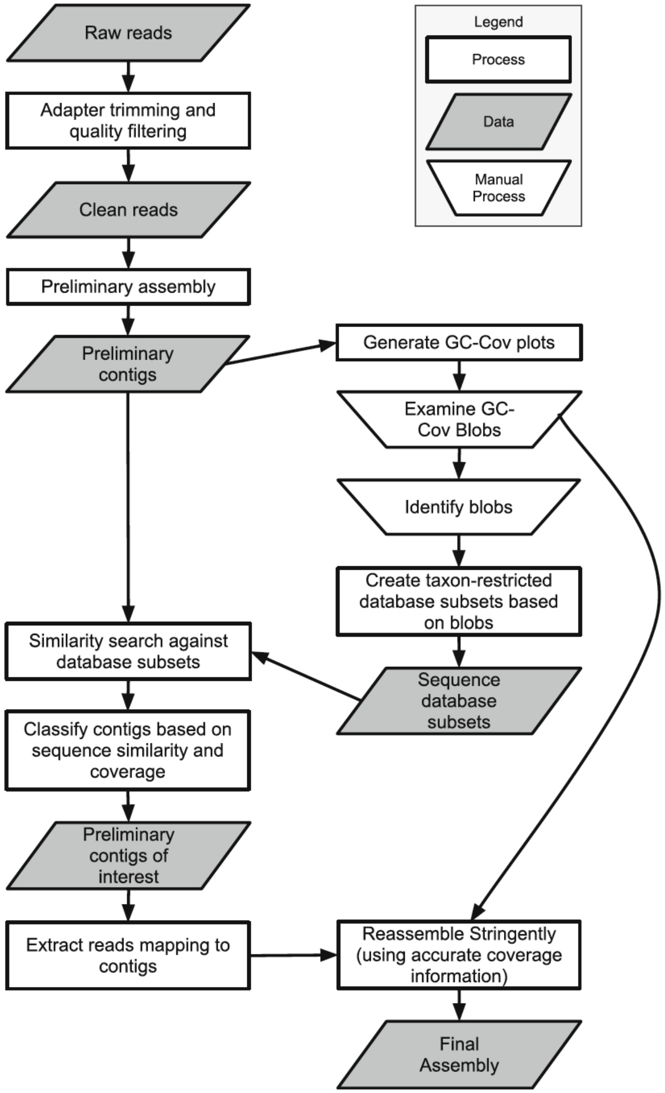
~16,000 protein-coding genes initially predicted

Larger scaffolds gave best BlastP hits
to nematode proteins

But smaller ones gave hits
to *Prevotella ruminicola*, etc. [!]

Meanwhile, assembling *cDNA* from RNA-seq reads
proved unexpectedly difficult, and gave one 'cDNA'
of 66 kb, with bacterial matches in BlastX. [!!]

Decontaminating reads in *Caenorhabditis* sp. 5



Legend

Process

Data

Manual Process

Raw reads

Adapter trimming and quality filtering

Clean reads

Preliminary assembly

Preliminary contigs

Generate GC-Cov plots

Examine GC-Cov Blobs

Identify blobs

Create taxon-restricted database subsets based on blobs

Sequence database subsets

Similarity search against database subsets

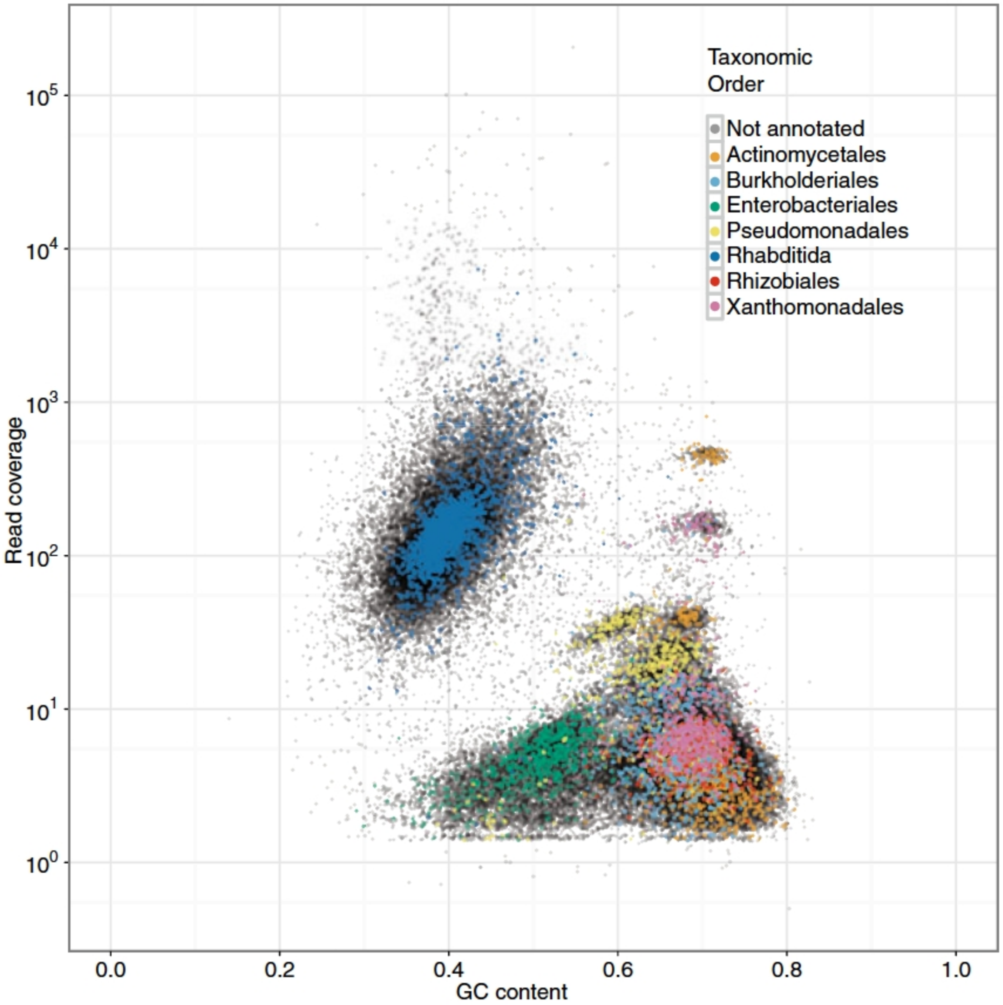
Classify contigs based on sequence similarity and coverage

Preliminary contigs of interest

Extract reads mapping to contigs

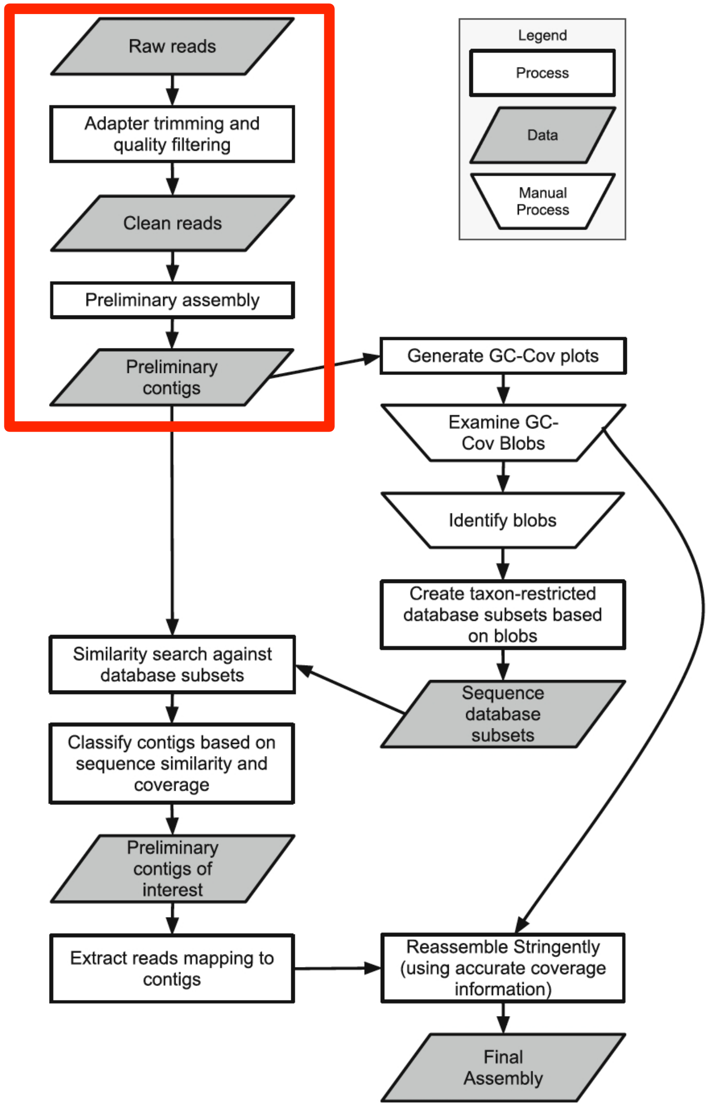
Reassemble Stringently (using accurate coverage information)

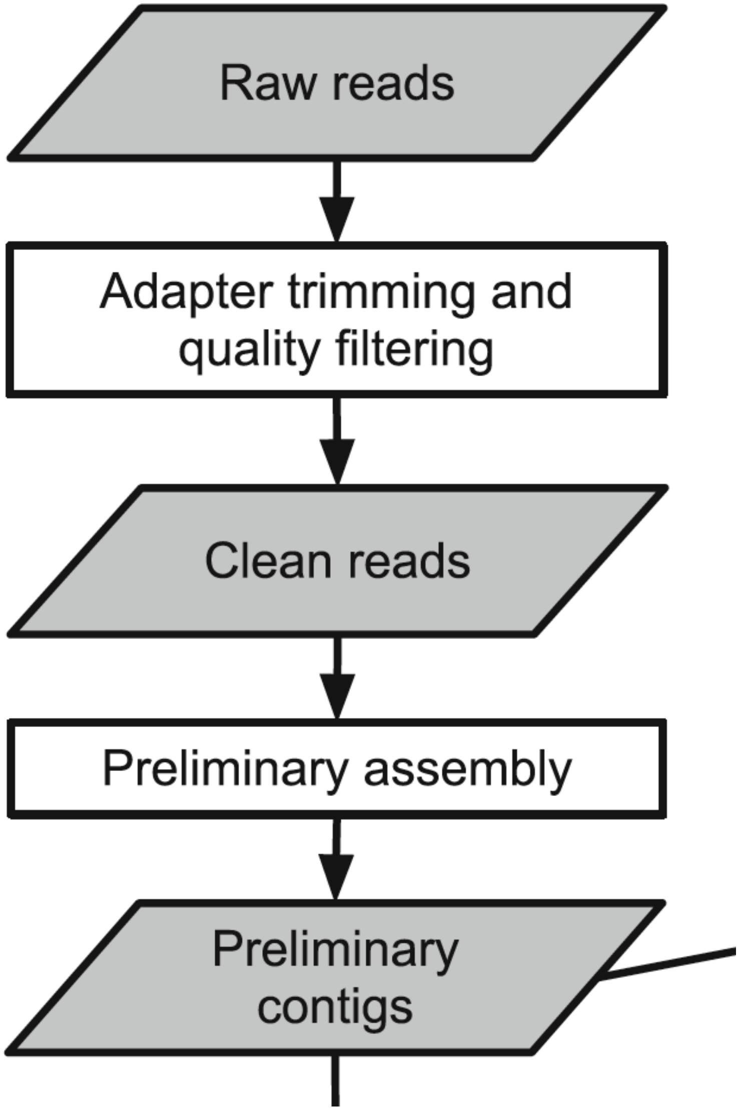
Final Assembly



Ref.: Kumar and Blaxter (2011), Symbiosis *55*, 119-126.
Software documentation: *https://github.com/sujaikumar/assemblage/blob/master/README.md*

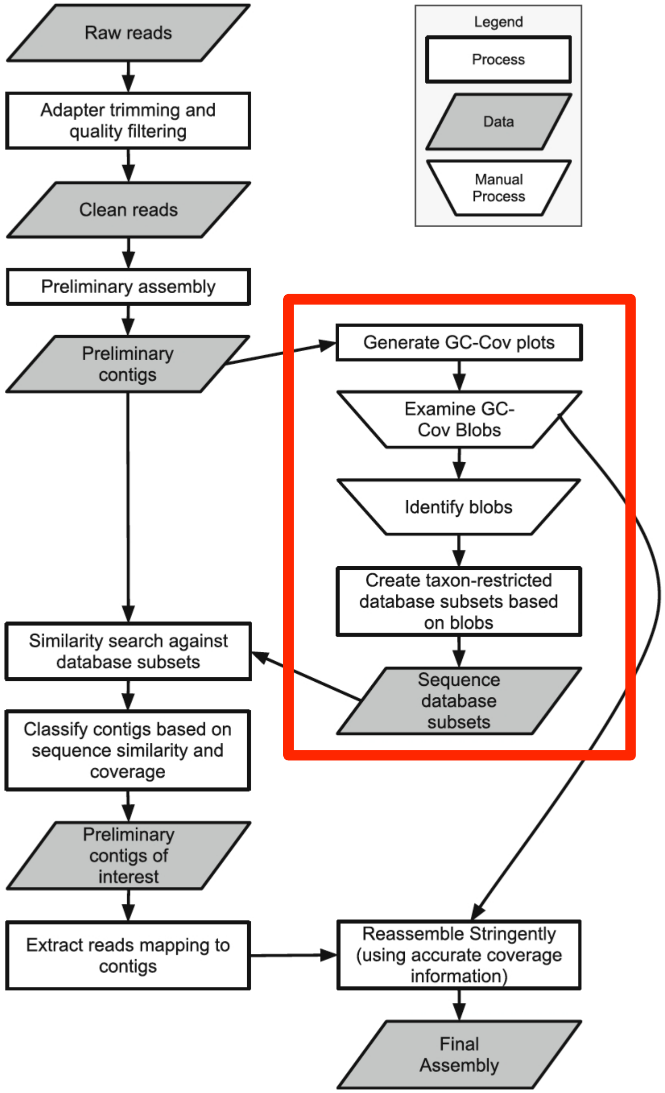
Decontaminating reads in *Caenorhabditis* sp. 5


Raw reads
Adapter trimming and quality filtering
Clean reads
Preliminary assembly
Preliminary contigs
Legend
Process
Data
Manual Process
Generate GC-Cov plots
Examine GC-Cov Blobs
Identify blobs
Create taxon-restricted database subsets based on blobs
Sequence database subsets
Similarity search against database subsets
Classify contigs based on sequence similarity and coverage
Preliminary contigs of interest
Extract reads mapping to contigs
Reassemble Stringently (using accurate coverage information)
Final Assembly



Ref.: Kumar and Blaxter (2011), Symbiosis *55*, 119-126.
Software documentation: *https://github.com/sujaikumar/assemblage/blob/master/README.md*

Decontaminating reads in *Caenorhabditis* sp. 5


Raw reads

Adapter trimming and quality filtering

Clean reads

Preliminary assembly

Preliminary contigs

Generate GC-Cov plots

Examine GC-Cov Blobs

Identify blobs

Create taxon-restricted database subsets based on blobs

Sequence database subsets

Similarity search against database subsets

Classify contigs based on sequence similarity and coverage

Preliminary contigs of interest

Extract reads mapping to contigs

Reassemble Stringently (using accurate coverage information)

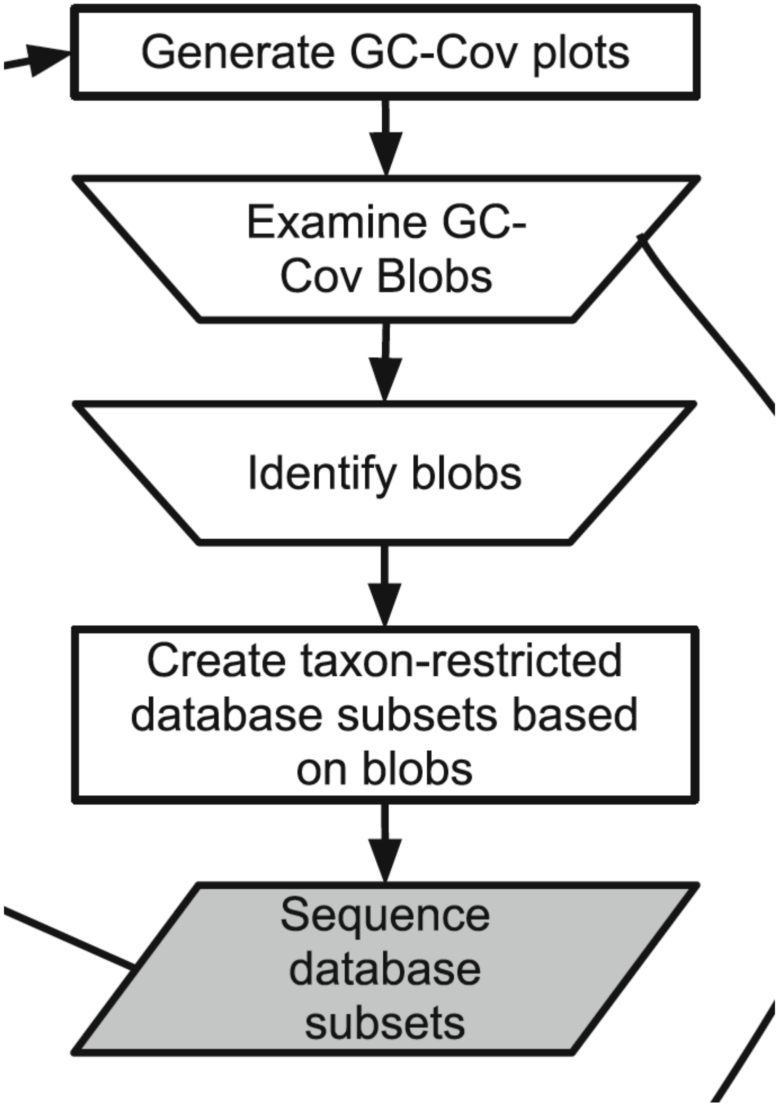
Final Assembly

Legend

Process

Data

Manual Process


Generate GC-Cov plots

Examine GC-Cov Blobs

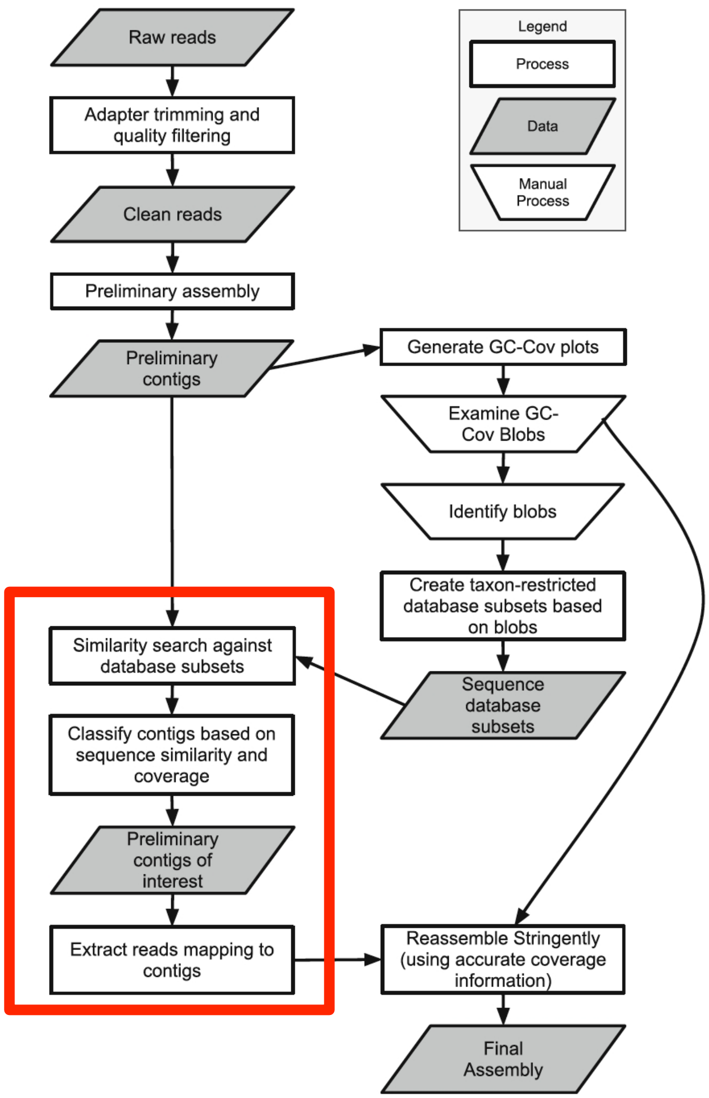
Identify blobs

Create taxon-restricted database subsets based on blobs

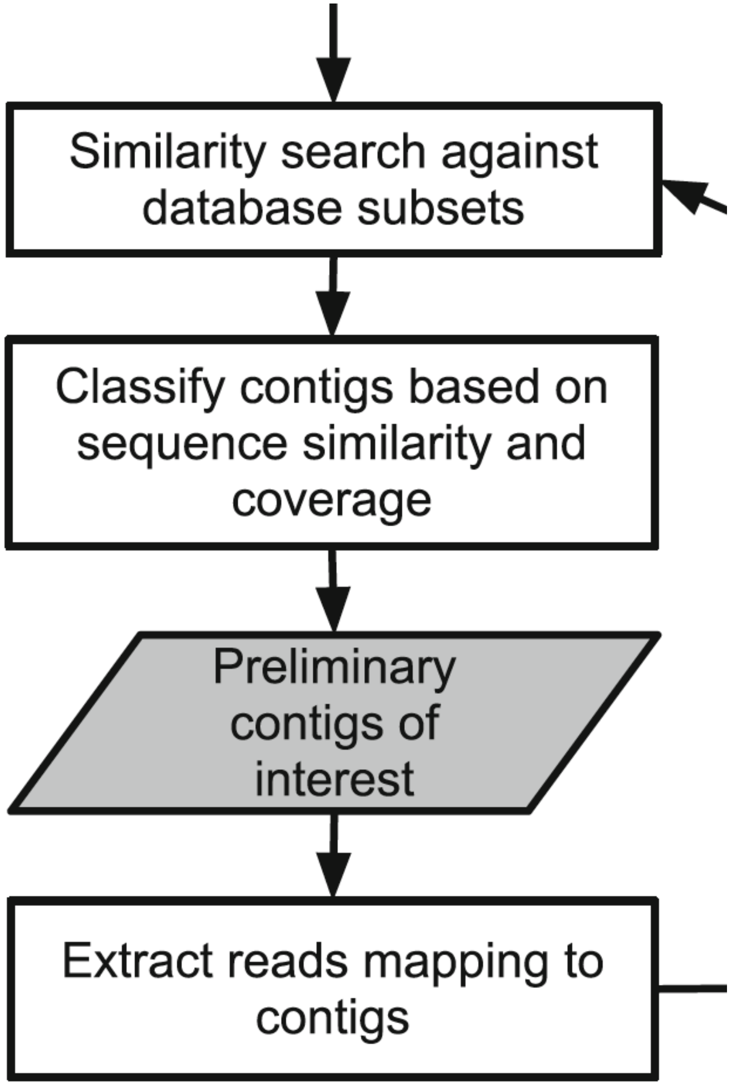
Sequence database subsets

Ref.: Kumar and Blaxter (2011), Symbiosis *55*, 119-126.
Software documentation: *https://github.com/sujaikumar/assemblage/blob/master/README.md*

Decontaminating reads in *Caenorhabditis* sp. 5



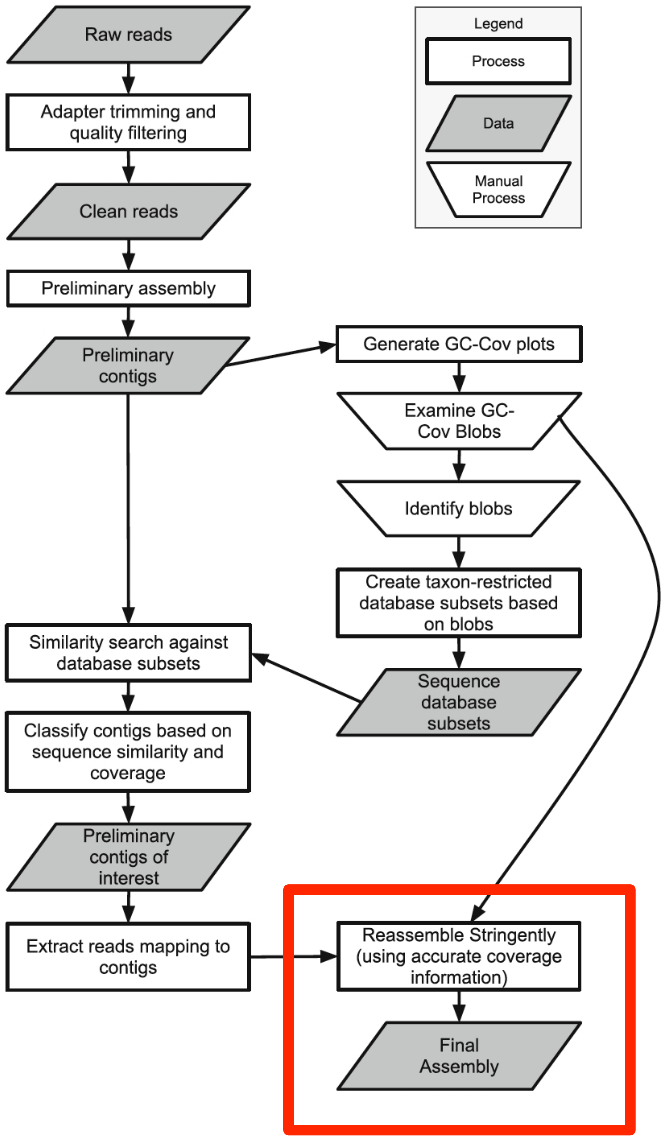
Legend
Process
Data
Manual Process
Raw reads
Adapter trimming and quality filtering
Clean reads
Preliminary assembly
Preliminary contigs
Generate GC-Cov plots
Examine GC-Cov Blobs
Identify blobs
Create taxon-restricted database subsets based on blobs
Sequence database subsets
Similarity search against database subsets
Classify contigs based on sequence similarity and coverage
Preliminary contigs of interest
Extract reads mapping to contigs
Reassemble Stringently (using accurate coverage information)
Final Assembly



Similarity search against database subsets
Classify contigs based on sequence similarity and coverage
Preliminary contigs of interest
Extract reads mapping to contigs

Ref.: Kumar and Blaxter (2011), Symbiosis *55*, 119-126.
Software documentation: *https://github.com/sujaikumar/assemblage/blob/master/README.md*

Decontaminating reads in *Caenorhabditis* sp. 5



Legend

Process

Data

Manual Process

Raw reads

Adapter trimming and quality filtering

Clean reads

Preliminary assembly

Preliminary contigs

Generate GC-Cov plots

Examine GC-Cov Blobs

Identify blobs

Create taxon-restricted database subsets based on blobs

Sequence database subsets

Similarity search against database subsets

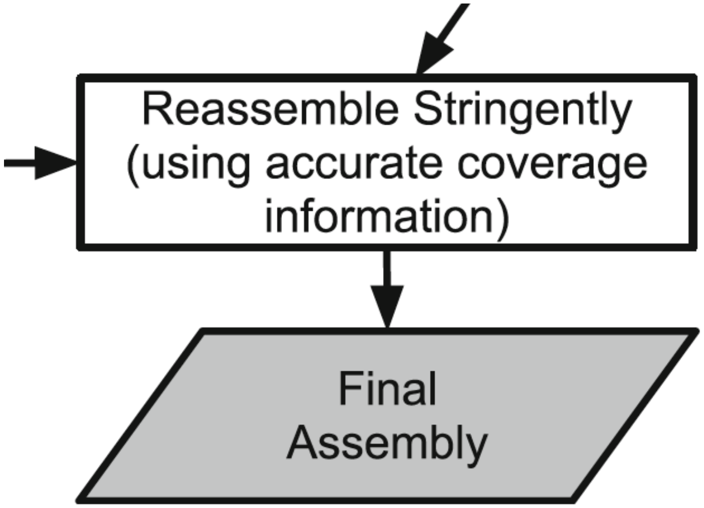
Classify contigs based on sequence similarity and coverage

Preliminary contigs of interest

Extract reads mapping to contigs

Reassemble Stringently (using accurate coverage information)

Final Assembly

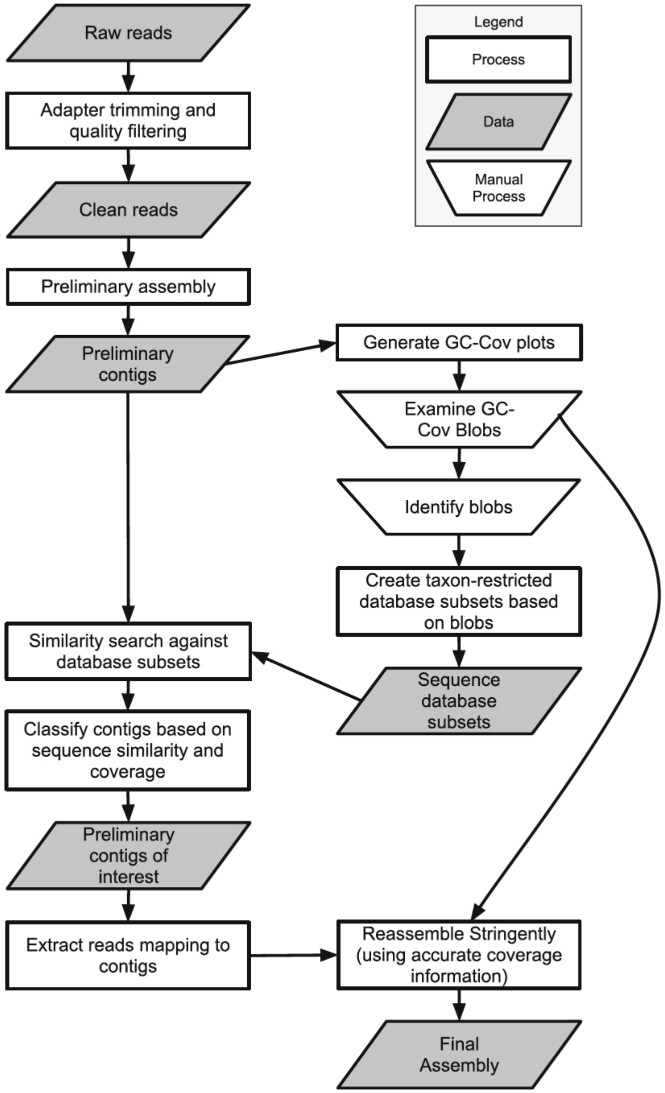


Reassemble Stringently (using accurate coverage information)

Final Assembly

Ref.: Kumar and Blaxter (2011), Symbiosis *55*, 119-126.
Software documentation: *https://github.com/sujaikumar/assemblage/blob/master/README.md*

Ideally, one should get a nice clean split



Legend

Process

Data

Manual Process

Raw reads

Adapter trimming and quality filtering

Clean reads

Preliminary assembly

Preliminary contigs

Generate GC-Cov plots

Examine GC-Cov Blobs

Identify blobs

Create taxon-restricted database subsets based on blobs

Sequence database subsets

Similarity search against database subsets

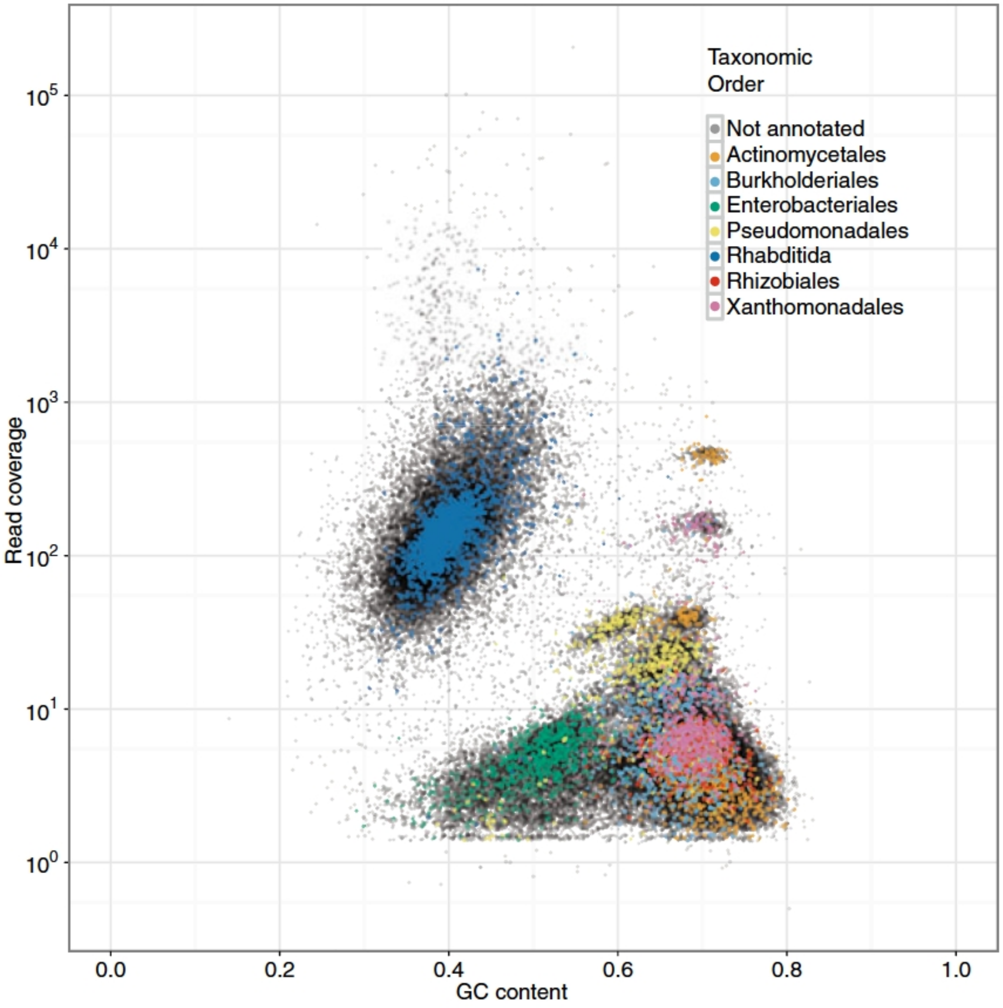
Classify contigs based on sequence similarity and coverage

Preliminary contigs of interest

Extract reads mapping to contigs

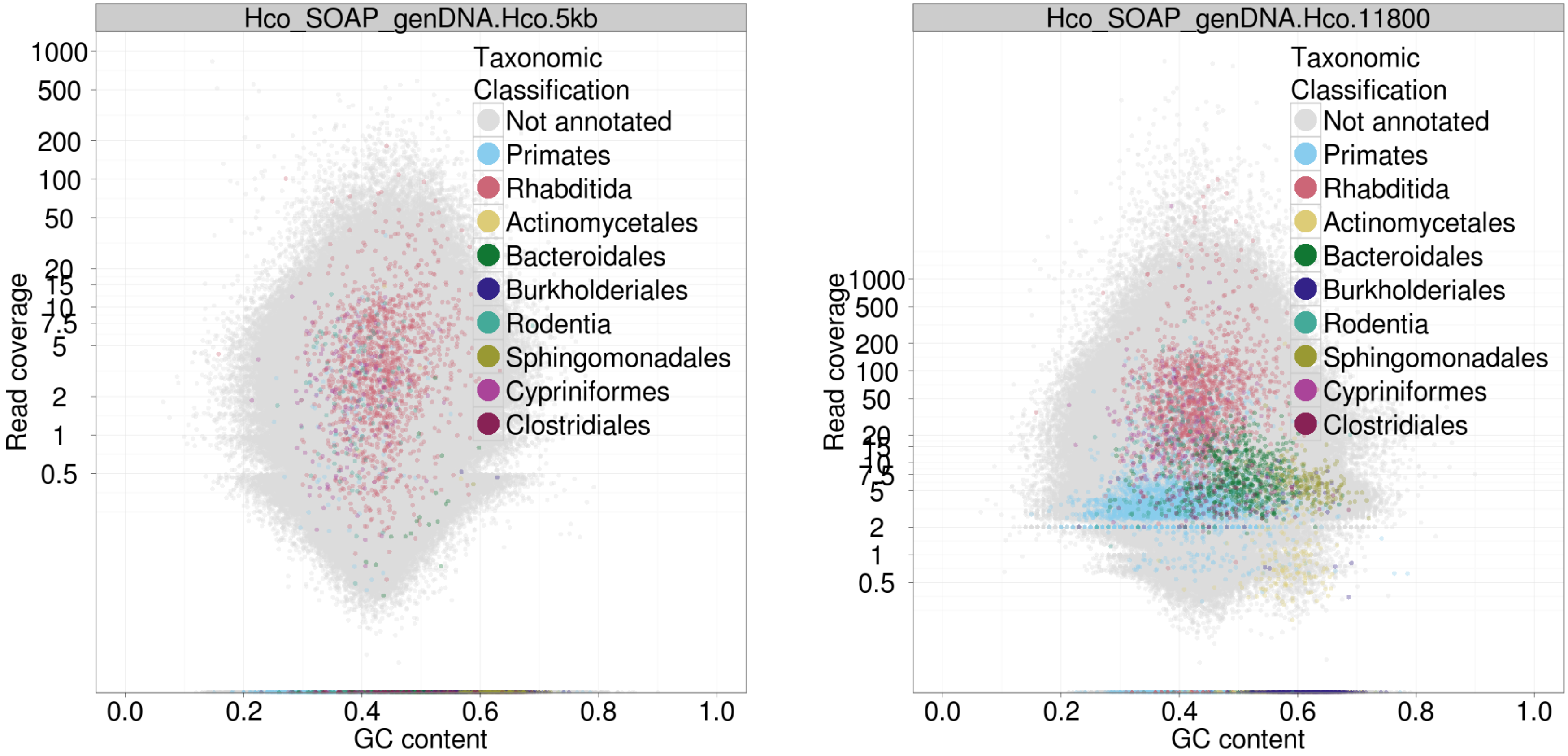
Reassemble Stringently (using accurate coverage information)

Final Assembly



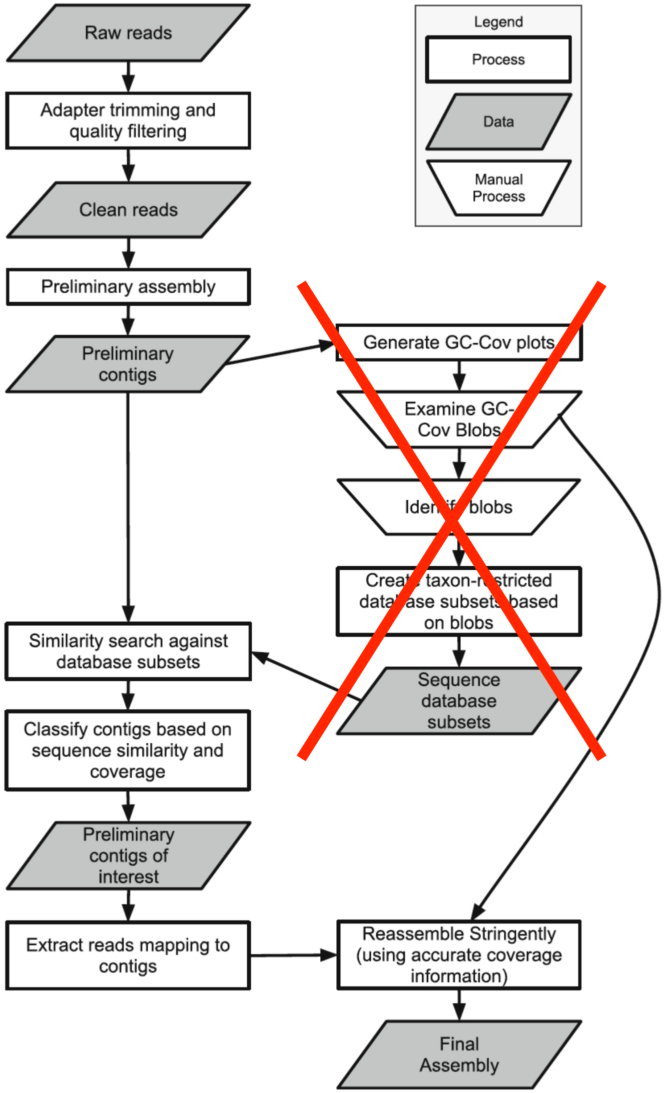
Ref.: Kumar and Blaxter (2011), Symbiosis *55*, 119-126.
Software documentation: *https://github.com/sujaikumar/assemblage/blob/master/README.md*

Haemonchus was not quite that easy



Some libraries were cleaner than others, but no clean split for contaminants existed.

So, just use brute-force MegablastN



Legend

Process

Data

Manual Process

Raw reads

Adapter trimming and quality filtering

Clean reads

Preliminary assembly

Preliminary contigs

Generate GC-Cov plots

Examine GC-Cov Blobs

Identify blobs

Create taxon-restricted database subsets based on blobs

Sequence database subsets

Similarity search against database subsets

Classify contigs based on sequence similarity and coverage

Preliminary contigs of interest

Extract reads mapping to contigs

Reassemble Stringently (using accurate coverage information)

Final Assembly

<u>Nematode database</u>:

Caenorhabditis elegans,
Pristionchus pacificus,
Ascaris suum,
and *Ancylostoma ceylanicum*

<u>Contaminant database</u>:

Bos taurus, *Ovis aries*;
1,991 microbial genomes from EBI;
and cow rumen metagenome

Reassembly and reanalysis

k=21 reassembled to 404 Mb (< 493 Mb)

Scaffold N50 dropped to 18.6 kb [!] from 109 kb

But cDNA no longer had ridiculous 'contigs',
and *Prevotella* etc. were gone. Yay.

Pasi Korhonen at U. Melbourne increased N50 to 33 kb
with a lot of read-editing and gap-filling;
using MAKER2, he predicted **23,610 genes**
(≥ 30 aa, Annot. Edit Dist. <0.4).

MAKER2. Ref.: Holt et al. (2011), BMC Bioinformatics *12*, 491.

Take-home lessons

Even highly recalcitrant genomic data sets can be filtered into serviceable assemblies

Resulting genome good enough for significant biology

Closer comparisons to hookworm possible than with *C. elegans*

Clean DNA/RNA matters!!

Overview

1. What hookworms are and why we care

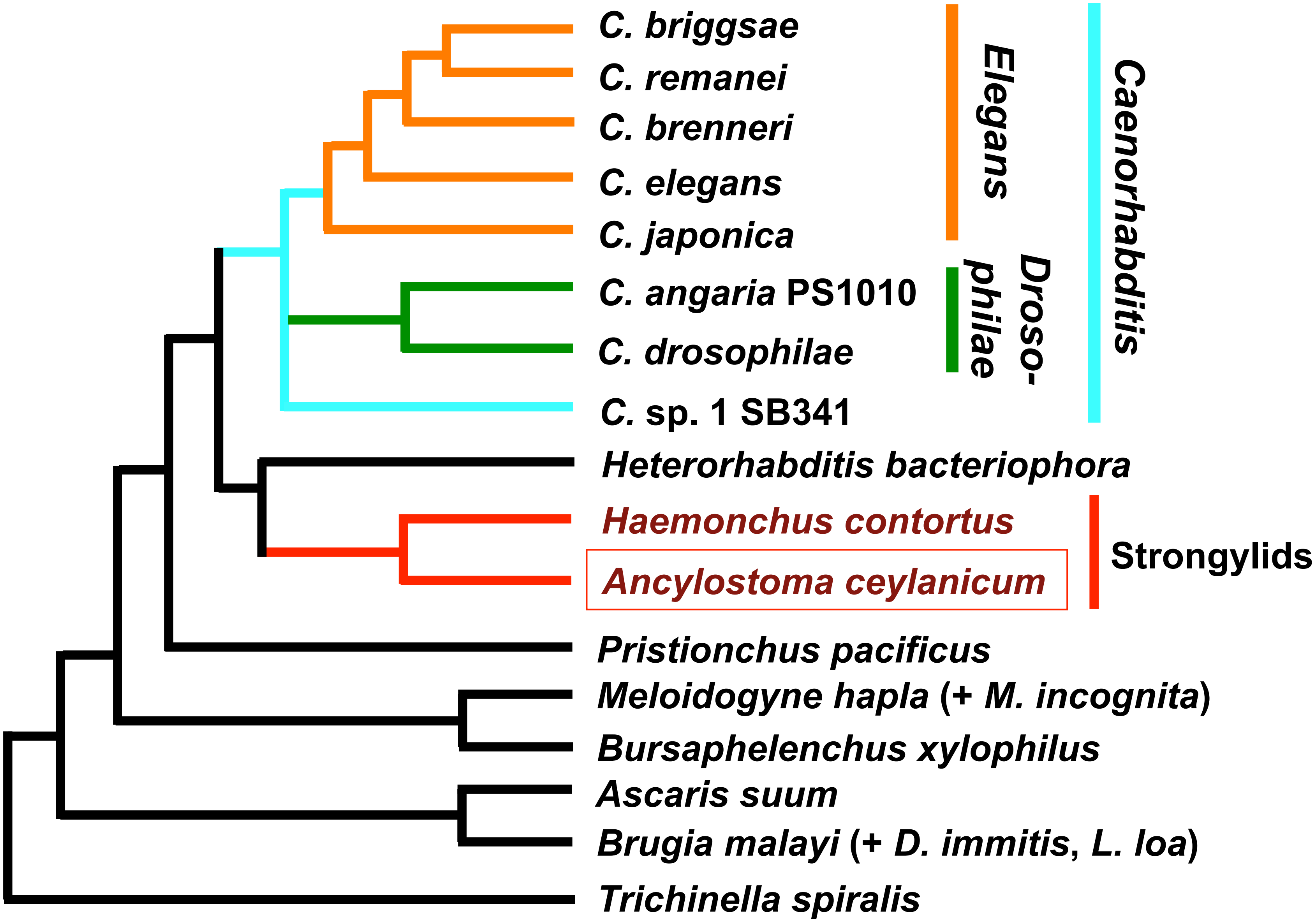
2. *Haemonchus contortus* adventures

3. *Ancylostoma ceylanicum* genome and transcriptome

4. Vaccine targets: ASPRs, and horizontally transmitted genes

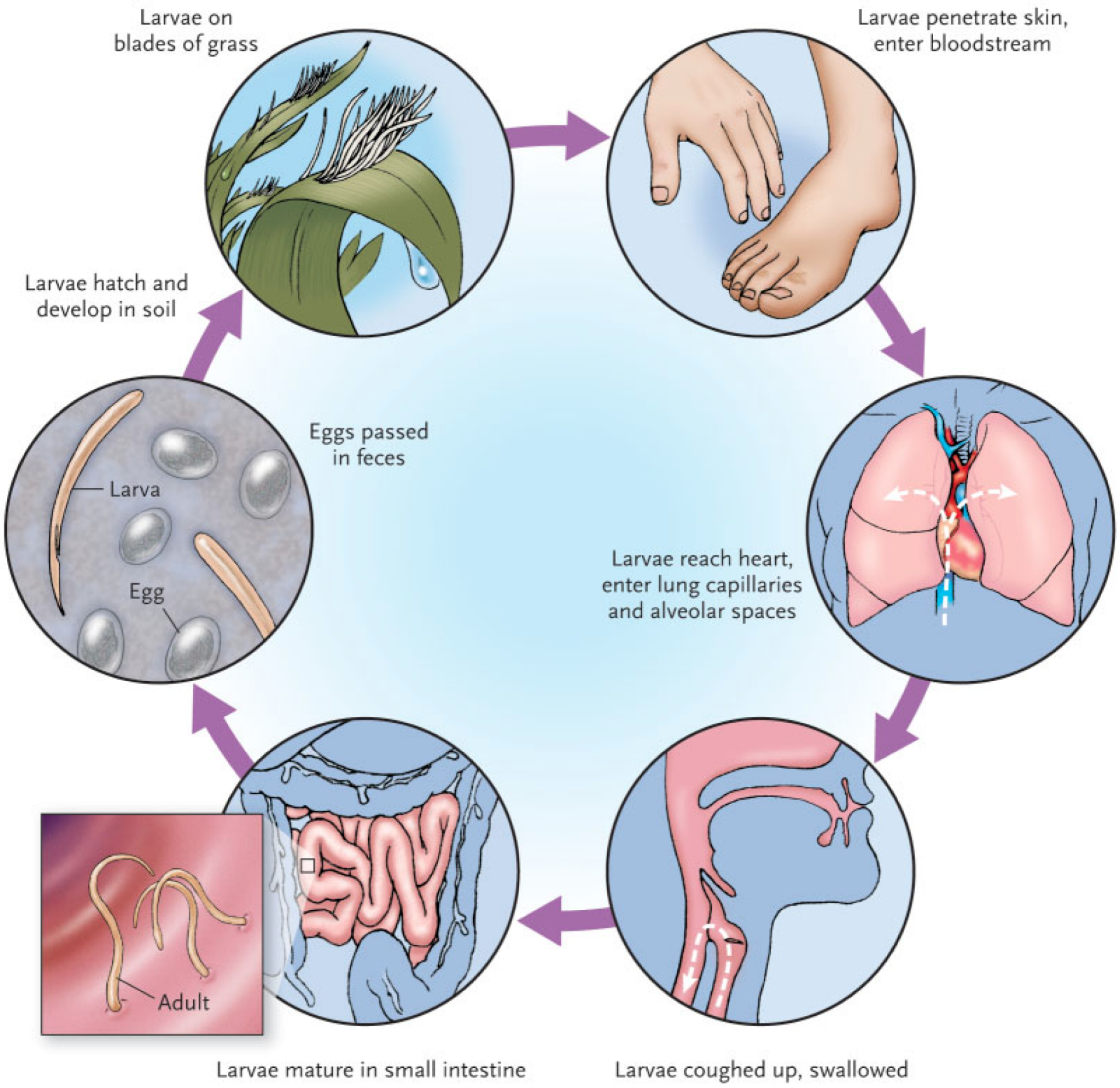
5. Some thoughts on 'descriptive genomics'

Nematode phylogeny



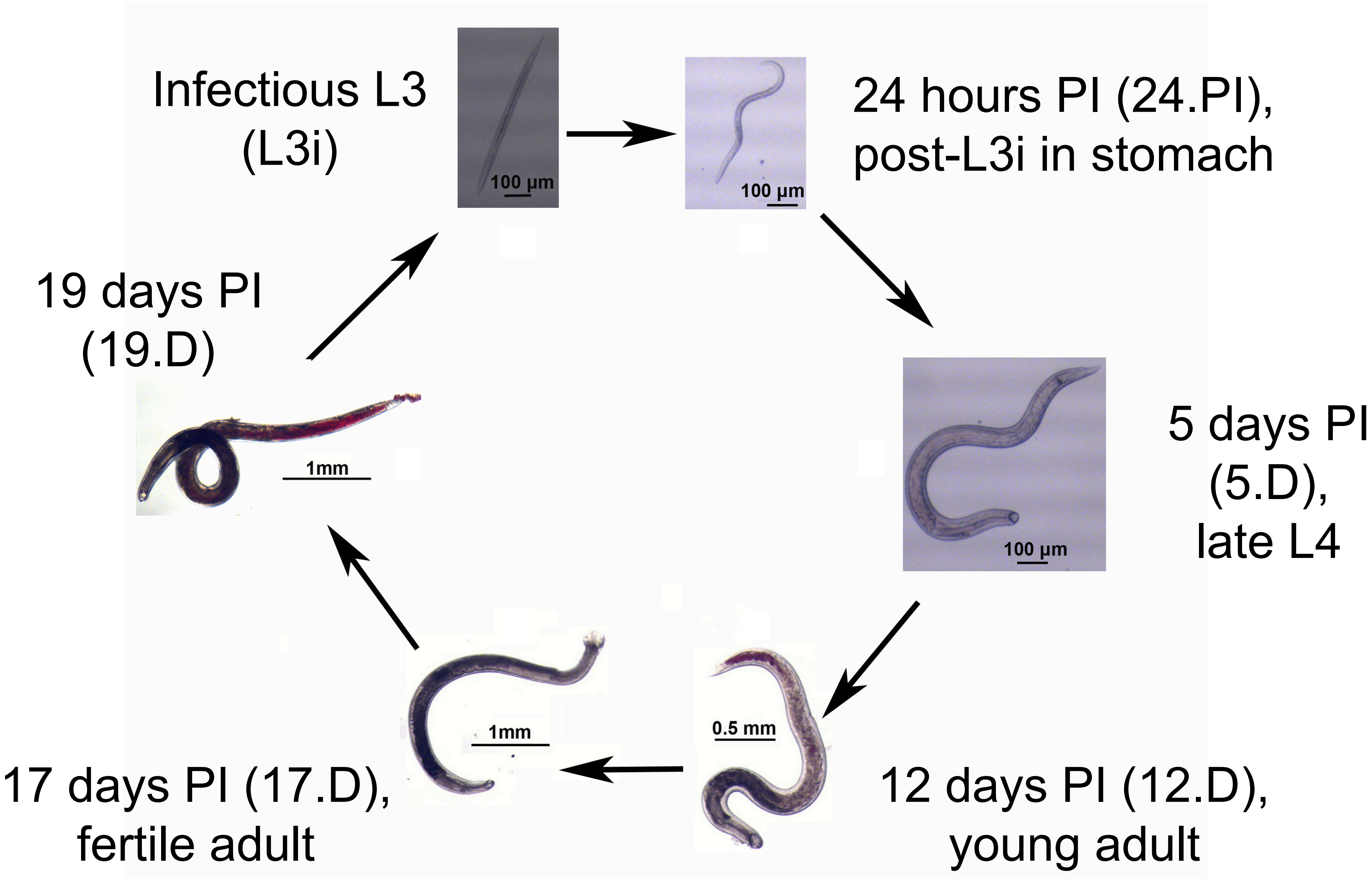
Refs.: Kiontke et al. (2011), BMC Evol. Biol. *11*, 339; van Megen et al. (2009), Nematology *11*, 927-950.

A. duodenale and *N. americanus* in humans


Larvae on blades of grass
Larvae penetrate skin, enter bloodstream
Larvae reach heart, enter lung capillaries and alveolar spaces
Larvae coughed up, swallowed
Larvae mature in small intestine
Adult
Eggs passed in feces
Larva
Egg
Larvae hatch and develop in soil

Ref.: Hotez et al. (2004), N. Engl. J. Med. *351*, 799-807.

Ancylostoma ceylanicum in hamsters



Infectious L3 (L3i)

24 hours PI (24.PI), post-L3i in stomach

5 days PI (5.D), late L4

12 days PI (12.D), young adult

17 days PI (17.D), fertile adult

19 days PI (19.D)

Ref.: Ray et al. (1972), J. Helminthology *46*, 357-362.

RNA-seq of developmental stages

Library	Read type	Paired reads	Paired nt	Single reads	Single nt
L3i	2x 100 nt	49.7 M	4.97 G	1.68 M	168 M
24.HCM	2x 100 nt	50.2 M	5.02 G	1.69 M	169 M
24.PI	1x 50 nt	0	0	22.9 M	1.15 G
5.D	2x 100 nt	60.8 M	6.07 G	93.0 K	9.09 M
12.D	2x 100 nt	65.5 M	6.55 G	97.8 M	9.56 M
17.D	2x 100 nt	92.6 M	9.26 G	135 K	13.2 M
19.D	2x 100 nt	59.5 M	5.95 G	87.4 K	8.52 M
khmer20-2	2x 100 nt	10.6 M	0.957 G	8.82 M	0.556 G

cDNA assembly from 2x100 nt RNA-seq reads

	oases 0.2.07, k = 21-31 (27)
Total nt:	64.3 M
Scaffolds:	333 K
Contigs:	332 K
% non-N:	100
Scaf. N50 nt:	294
Scaf. max. nt:	10,003
Contig N50:	294
Contig max. nt:	10,003

Ref.: Schulz et al. (2012), Bioinformatics *28*, 1086-1092.

Genomic reads

Insert size	Paired reads	Paired nt	Coverage	Single reads	Single nt	Coverage
550 bp	207 M	20.3 G	61.5x	2.44 M	194 M	0.6x
6 kb	43.6 M	4.05 G	12.3x	8.67 M	542 M	1.6x

Libraries were 2x101 and 2x100 nt.
Coverage is based on final genome estimate of 330 Mb.

Stepwise genome assemblies

	velvet k=75
Total nt:	328 M
Scaffolds:	16.5 K
Contigs:	86.0 K
% non-N:	89.6
Scaf. N50 nt:	392 K
Scaf. max. nt:	2.77 M
Cont. N50 nt:	7.77 K
Cont. max. nt:	63.7 K

Assembled with velvet 1.2.05.

Tried k-values from 59 to 81;
picked k=75 as best (vs. k=65).

198 M/261 M reads (75.8%)
used in the k=75 assembly.

Used '*-shortMatePaired2 yes*'
to reject likely jumping chimeras.

N.B.: with k=75, chimeras will have
many anomalous k-mers.

(Did try trimming the jumping reads,
but to no obvious benefit.)

Ref.: Zerbino and Birney (2008), Genome Res. *18*, 821-829.

Stepwise genome assemblies

	velvet k=75	**+GapCloser**
Total nt:	328 M	322 M
Scaffolds:	16.5 K	16.5 K
Contigs:	86.0 K	47.4 K
% non-N:	89.6	96.1
Scaf. N50 nt:	392 K	384 K
Scaf. max. nt:	2.77 M	2.72 M
Cont. N50 nt:	7.77 K	18.0 K
Cont. max. nt:	63.7 K	125 K

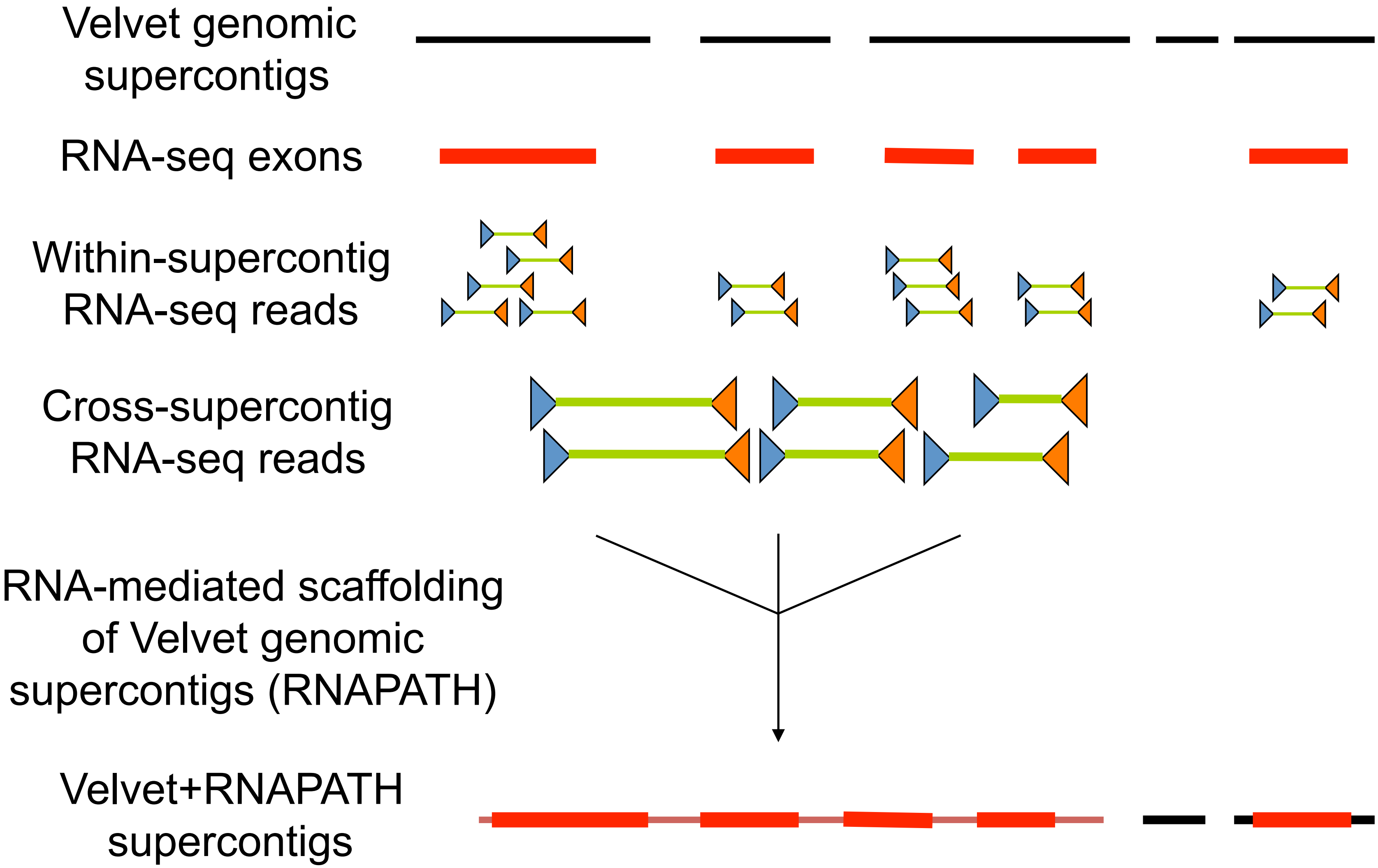
BGI GapCloser 1.12 (release_2011). Ref.: Li et al. (2010). Genome Res. *20*, 265-272.

Stepwise genome assemblies

	velvet k=75	**+GapCloser**	**+HaploMerger**
Total nt:	328 M	322 M	313 M
Scaffolds:	16.5 K	16.5 K	2.14 K
Contigs:	86.0 K	47.4 K	32.2 K
% non-N:	89.6	96.1	96.1
Scaf. N50 nt:	392 K	384 K	393 K
Scaf. max. nt:	2.77 M	2.72 M	2.72 M
Cont. N50 nt:	7.77 K	18.0 K	18.5 K
Cont. max. nt:	63.7 K	125 K	125 K

HaploMerger 20111230. Ref.: Huang et al. (2012), Genome Res. *22*, 1581-1588.

RNA scaffolding can improve genome assemblies



Velvet genomic supercontigs

RNA-seq exons

Within-supercontig RNA-seq reads

Cross-supercontig RNA-seq reads

RNA-mediated scaffolding of Velvet genomic supercontigs (RNAPATH)

Velvet+RNAPATH supercontigs

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

Stepwise genome assemblies

	velvet k=75	+GapCloser	+HaploMerger	Final (+RNA-scaf.)
Total nt:	328 M	322 M	313 M	313 M
Scaffolds:	16.5 K	16.5 K	2.14 K	1.74 K
Contigs:	86.0 K	47.4 K	32.2 K	32.2 K
% non-N:	89.6	96.1	96.1	96.1
Scaf. N50 nt:	392 K	384 K	393 K	668 K
Scaf. max. nt:	2.77 M	2.72 M	2.72 M	4.80 M
Cont. N50 nt:	7.77 K	18.0 K	18.5 K	18.5 K
Cont. max. nt:	63.7 K	125 K	125 K	125 K

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

The genomic assembly is ~95% complete

Counting **31-mer frequencies** (in 197 M reads trimmed to 95 nt) indicates a true genome size of 320 Mb; by this, the 313 Mb assembly is **98% complete**.

Refs.: Abubucker et al. (2008), Mol. Biochem. Parasitol. *157*, 187-192;
Parra et al. (2009), Nucleic Acids Res. *37*, 289-297.

The genomic assembly is ~95% complete

Counting **31-mer frequencies** (in 197 M reads trimmed to 95 nt) indicates a true genome size of 320 Mb; by this, the 313 Mb assembly is **98% complete**.

Given 64 Mb of **cDNA**, 310,647/332,724 sequences could be mapped to the genome with **BLAT**, indicating it to be **93% complete**.

Refs.: Abubucker et al. (2008), Mol. Biochem. Parasitol. *157*, 187-192;
Parra et al. (2009), Nucleic Acids Res. *37*, 289-297.

The genomic assembly is ~95% complete

Counting **31-mer frequencies** (in 197 M reads trimmed to 95 nt) indicates
a true genome size of 320 Mb; by this, the 313 Mb assembly is **98% complete**.

Given 64 Mb of **cDNA**, 310,647/332,724 sequences could be mapped
to the genome with **BLAT**, indicating it to be **93% complete**.

CEGMA indicates that the genome is **91-99% complete**:
91% for complete matches to core eukaryotic genes,
and 99% for partial matches.

Refs.: Abubucker et al. (2008), Mol. Biochem. Parasitol. *157*, 187-192;
Parra et al. (2009), Nucleic Acids Res. *37*, 289-297.

The genomic assembly is ~95% complete

Counting **31-mer frequencies** (in 197 M reads trimmed to 95 nt) indicates
a true genome size of 320 Mb; by this, the 313 Mb assembly is **98% complete**.

Given 64 Mb of **cDNA**, 310,647/332,724 sequences could be mapped
to the genome with **BLAT**, indicating it to be **93% complete**.

CEGMA indicates that the genome is **91-99% complete**:
91% for complete matches to core eukaryotic genes,
and 99% for partial matches.

Consensus of these three assays: true genome size of **~330 Mb**.
By comparison, *A. caninum*'s genome was measured at 347 Mb.

Refs.: Abubucker et al. (2008), Mol. Biochem. Parasitol. *157*, 187-192;
Parra et al. (2009), Nucleic Acids Res. *37*, 289-297.

The genomic assembly is ~95% complete

Counting **31-mer frequencies** (in 197 M reads trimmed to 95 nt) indicates a true genome size of 320 Mb; by this, the 313 Mb assembly is **98% complete**.

Given 64 Mb of **cDNA**, 310,647/332,724 sequences could be mapped to the genome with **BLAT**, indicating it to be **93% complete**.

CEGMA indicates that the genome is **91-99% complete**:
91% for complete matches to core eukaryotic genes,
and 99% for partial matches.

Consensus of these three assays: true genome size of **~330 Mb**.
By comparison, *A. caninum*'s genome was measured at 347 Mb.

N.B.: CEGMA also shows the assembly has 1.13 complete orthologs/genome.
This compares well with *C. elegans*, *C. briggsae*, and *C*. sp. 11,
which have 1.11-1.15 orthologs/genome.
Hence, the level of heterozygosity is probably low.

Refs.: Abubucker et al. (2008), Mol. Biochem. Parasitol. *157*, 187-192;
Parra et al. (2009), Nucleic Acids Res. *37*, 289-297.

A. ceylanicum has a bigger, more repetitive genome than *C. elegans*

40.5% of genomic DNA is repetitive, over twice the 17% in *C. elegans* or *P. pacificus;* without this difference, *A. ceylanicum*'s genome would be ~60 Mb smaller.

Refs.: Price et al. (2005), Bioinformatics *21 Suppl 1*, i351-358; Laha et al. (2007), PLoS Negl. Trop. Dis. *1*, e35.

A. ceylanicum has a bigger, more repetitive genome than *C. elegans*

40.5% of genomic DNA is repetitive, over twice the 17% in *C. elegans* or *P. pacificus;* without this difference, *A. ceylanicum*'s genome would be ~60 Mb smaller.

Expanded genomes may be common among strongylids, versus either *C. elegans* (100 Mb) or *P. pacificus* (~230 Mb).
For instance, *H. contortus* was measured at ~325 Mb.

Refs.: Price et al. (2005), Bioinformatics *21 Suppl 1*, i351-358; Laha et al. (2007), PLoS Negl. Trop. Dis. *1*, e35.

A. ceylanicum has a bigger, more repetitive genome than *C. elegans*

40.5% of genomic DNA is repetitive, over twice the 17% in *C. elegans* or *P. pacificus;* without this difference, *A. ceylanicum*'s genome would be ~60 Mb smaller.

Expanded genomes may be common among strongylids, versus either *C. elegans* (100 Mb) or *P. pacificus* (~230 Mb).
For instance, *H. contortus* was measured at ~325 Mb.

One possible source of the expanded repeats may be horizontal transmission from mammalian hosts.
E.g., *A. caninum* has one Mariner-like element ('bandit') with its highest similarity to human Hsmar1.

Refs.: Price et al. (2005), Bioinformatics *21 Suppl 1*, i351-358; Laha et al. (2007), PLoS Negl. Trop. Dis. *1*, e35.

A. ceylanicum has ≥23,855 genes encoding proteins of ≥100 residues

Make *A. ceylanicum*-specific parameters for the genefinder AUGUSTUS 2.6.1

Run AUGUSTUS with these parameters + BLAT-mapped cDNA

Allow genes down to 30 a.a. max. prod. size, rather than the more typical 100 a.a.

Predict 26,966 protein-coding genes with products of ≥100 a.a.; another 10,050 genes encoding 30-99 a.a.

Refs.: Stanke et al. (2008), Bioinformatics *24*, 637-644; Li and Dewey (2011), BMC Bioinformatics *12*, 323.

A. ceylanicum has ≥23,855 genes encoding proteins of ≥100 residues

Make *A. ceylanicum*-specific parameters for the genefinder AUGUSTUS 2.6.1

Run AUGUSTUS with these parameters + BLAT-mapped cDNA

Allow genes down to 30 a.a. max. prod. size, rather than the more typical 100 a.a.

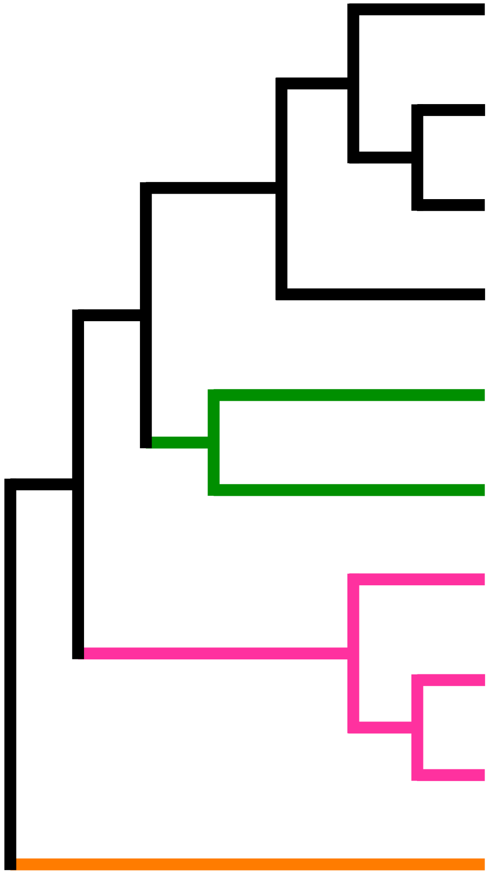
Predict 26,966 protein-coding genes with products of ≥100 a.a.; another 10,050 genes encoding 30-99 a.a.

Using RSEM, map RNA-seq data for *C. elegans* from modENCODE and our own work (on ± albendazole during L4); find that 99.9% of genes in WS230 have ≥5 mapped reads from *some* stage.

By this same criterion, find evidence for expression in 23,855 *A. ceylanicum* genes with ≥100 a.a. (89% total); 3,111 *A. ceylanicum* genes with 30-99 a.a. (31% total).

Refs.: Stanke et al. (2008), Bioinformatics *24*, 637-644; Li and Dewey (2011), BMC Bioinformatics *12*, 323.

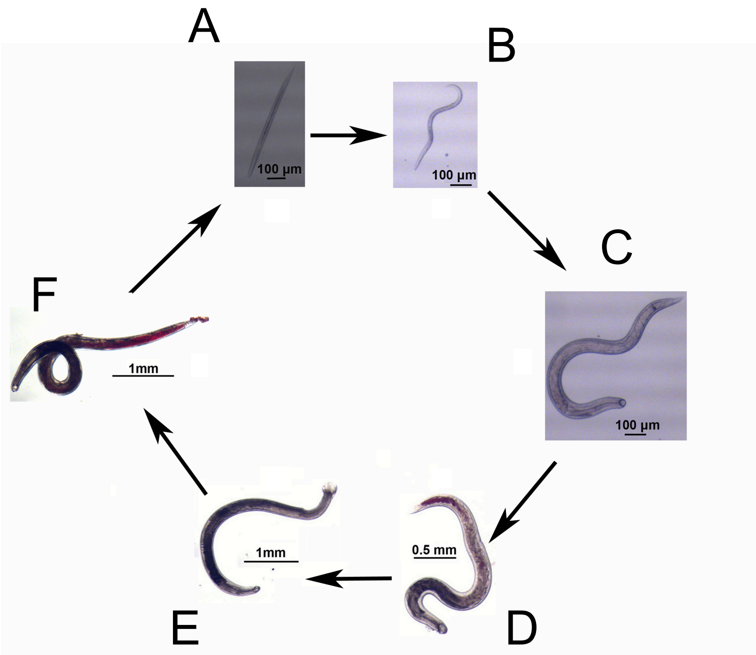
But, strict gene orthologies in *Ancylostoma* closely resemble those seen for *C. elegans*

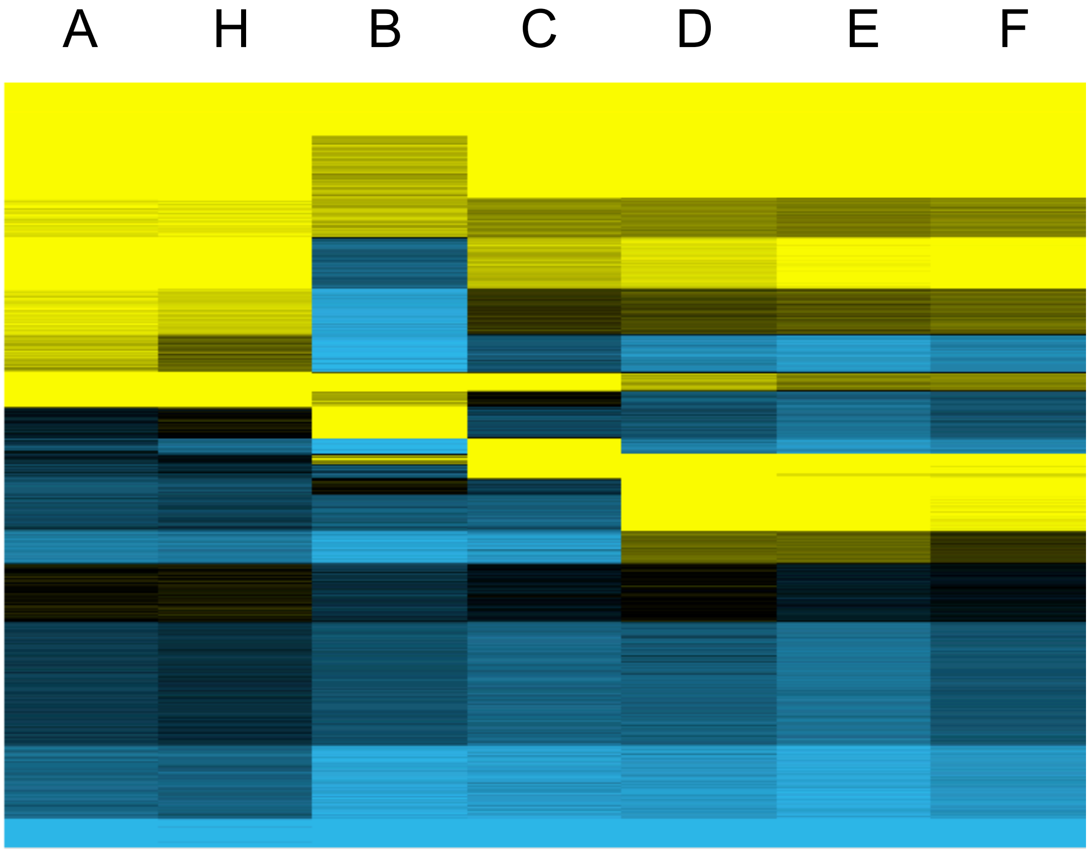


	A. cey.	***C. ele.***
A. ceylanicum	*8,283*	7,026
C. elegans	7,026	*11,213*
C. briggsae	6,951	10,758
P. pacificus	4,880	4,987
B. xylophilus	5,411	5,558
M. hapla	3,840	3,986
A. suum	5,943	6,072
B. malayi	3,778	3,780
D. immitis	5,217	5,254
T. spiralis	3,022	3,122

OrthoMCL 1.3. Ref.: Li et al. (2003), Genome Res. *13*, 2178-2189.

In vivo infection has much stronger effects on gene expression than its in vitro model





RSEM 1.2.0. Ref.: Li and Dewey (2011), BMC Bioinformatics *12*, 323.

Some changes in gene regulation are highly significant; others less so

In vivo infection has a far stronger effect than its in vitro model:

L3i to 24.PI: 1,146 upregulated; 1,352 downregulated.
In contrast, L3i to 24.HCM: 108 upregulated, 50 downregulated.

Two subsequent transistions are equally significant:

24PI to 5.D: 1,798 upregulated, 846 downregulated.
5.D to 12.D: 1,781 upregulated, 676 downregulated.

Later changes are minor:

12.D to 17.D: 0 upregulated, 2 downregulated.
17.D. to 19.D: 0 upregulated, 0 downregulated

edgeR X, with q-value ≤10^{-3} as significant. Ref.: Robinson et al. (2010), Bioinformatics *26*, 139-140.

Gene Ontology shows up- *and* down-regulated functions during infection

L3i to 24.PI, upregulated:
proteases, protease inhibitors, nucleases, and protein synthesis

FUNC 0.4.5. Ref.: Prüfer et al. (2007), BMC Bioinformatics *8*, 41.

Gene Ontology shows up- *and* down-regulated functions during infection

L3i to 24.PI, upregulated:
proteases, protease inhibitors, nucleases, and protein synthesis

L3i to 24.PI, downregulated:
GPCRs, receptor-gated ion channels, other neurotransmission-related
(N.B.: this is conserved in *H. contortus* and *C. elegans*)

FUNC 0.4.5. Ref.: Prüfer et al. (2007), BMC Bioinformatics *8*, 41.

Gene Ontology shows up- *and* down-regulated functions during infection

L3i to 24.PI, upregulated:
proteases, protease inhibitors, nucleases, and protein synthesis

L3i to 24.PI, downregulated:
GPCRs, receptor-gated ion channels, other neurotransmission-related
(N.B.: this is conserved in *H. contortus* and *C. elegans*)

24.PI to 5.D, upregulated:
structural components of cuticle, binding cytoskeletal proteins, e.g., actin

FUNC 0.4.5. Ref.: Prüfer et al. (2007), BMC Bioinformatics *8*, 41.

Gene Ontology shows up- *and* down-regulated functions during infection

L3i to 24.PI, upregulated:
proteases, protease inhibitors, nucleases, and protein synthesis

L3i to 24.PI, downregulated:
GPCRs, receptor-gated ion channels, other neurotransmission-related
(N.B.: this is conserved in *H. contortus* and *C. elegans*)

24.PI to 5.D, upregulated:
structural components of cuticle, binding cytoskeletal proteins, e.g., actin

5.D to 12.D, upregulated:
protein tyrosine phosphatases and serine/threonine kinases

FUNC 0.4.5. Ref.: Prüfer et al. (2007), BMC Bioinformatics *8*, 41.

Overview

1. What hookworms are and why we care
2. *Haemonchus contortus* adventures
3. *Ancylostoma ceylanicum* genome and transcriptome
4. Vaccine targets: ASPRs, and horizontally transmitted genes
5. Some thoughts on 'descriptive genomics'

Ugh, how do you look for an *antigen*?

1. Look for new components of early infection.
If the organism thinks something is worth upregulating and spitting into the host, maybe it's worth preempting immunologically.

2. Look for genes which have some evidence of being extraordinarily relevant to the worm's survival in the host.

3. Or, since it is not particularly obvious how to do (2), just look for Interesting Stuff and hopefully get lucky.

Look for genes upregulated during infection

Run rank-sum statistics on proteins, for L3i to 24PI:
i.e., look for protein motifs or orthology groups
disproportionately represented in genes with high 24PI/L3i expression ratios.

This mostly gives things which we expect to see:

```
AP domain [IPR014044]                                       2.915e-58
Allergen V5/Tpx-1-related [IPR001283]                       3.0354e-42
CAP [PF00188.21]                                            8.6906e-39
Peptidase C1A, papain C-terminal [IPR000668]                7.934e-09
Peptidase C1A, papain [IPR013128]                           7.934e-09
Peptidase_C1 [PF00112.18]                                   1.27468e-08
Peptidase C1A, cathepsin B [IPR015643]                      3.7652e-08
Peptidase, cysteine peptidase active site [IPR000169]       1.30596e-07
ORTHOMCL896.14spp(21 genes,1 taxa): ancylostoma (21 g.)     1.73942e-06
Transthyretin-like [IPR001534]                              2.5116e-06
Asp [PF00026.18]                                            4.5258e-06
Peptidase A1 [IPR001461]                                    4.5258e-06
Galectin, carbohydrate recognition domain [IPR001079]       0.00123752
Gal-bind_lectin [PF00337.17]                                0.0031598
Apyrase [IPR009283]                                         0.0047722
Apyrase [PF06079.6]                                         0.0047722
Pfam-B_14321 [PB014321]                                     0.005368
Peptidase S28 [IPR008758]                                   0.0054872
D-amino-acid oxidase [IPR023209]                            0.0061606
DNase_II [PF03265.10]                                       0.010297
```

Look for genes upregulated during infection

Run rank-sum statistics on proteins, for L3i to 24PI:
i.e., look for protein motifs or orthology groups
disproportionately represented in genes with high 24PI/L3i expression ratios.

But it shows one set of proteins which aren't obviously a known group:

AP domain [IPR014044]	2.915e-58
Allergen V5/Tpx-1-related [IPR001283]	3.0354e-42
CAP [PF00188.21]	8.6906e-39
Peptidase C1A, papain C-terminal [IPR000668]	7.934e-09
Peptidase C1A, papain [IPR013128]	7.934e-09
Peptidase_C1 [PF00112.18]	1.27468e-08
Peptidase C1A, cathepsin B [IPR015643]	3.7652e-08
Peptidase, cysteine peptidase active site [IPR000169]	1.30596e-07
ORTHOMCL896.14spp(21 genes,1 taxa): ancylostoma (21 g.)	**1.73942e-06**
Transthyretin-like [IPR001534]	2.5116e-06
Asp [PF00026.18]	4.5258e-06
Peptidase A1 [IPR001461]	4.5258e-06
Galectin, carbohydrate recognition domain [IPR001079]	0.00123752
Gal-bind_lectin [PF00337.17]	0.0031598
Apyrase [IPR009283]	0.0047722
Apyrase [PF06079.6]	0.0047722
Pfam-B_14321 [PB014321]	0.005368
Peptidase S28 [IPR008758]	0.0054872
D-amino-acid oxidase [IPR023209]	0.0061606
DNase_II [PF03265.10]	0.010297

One new class of upregulated genes

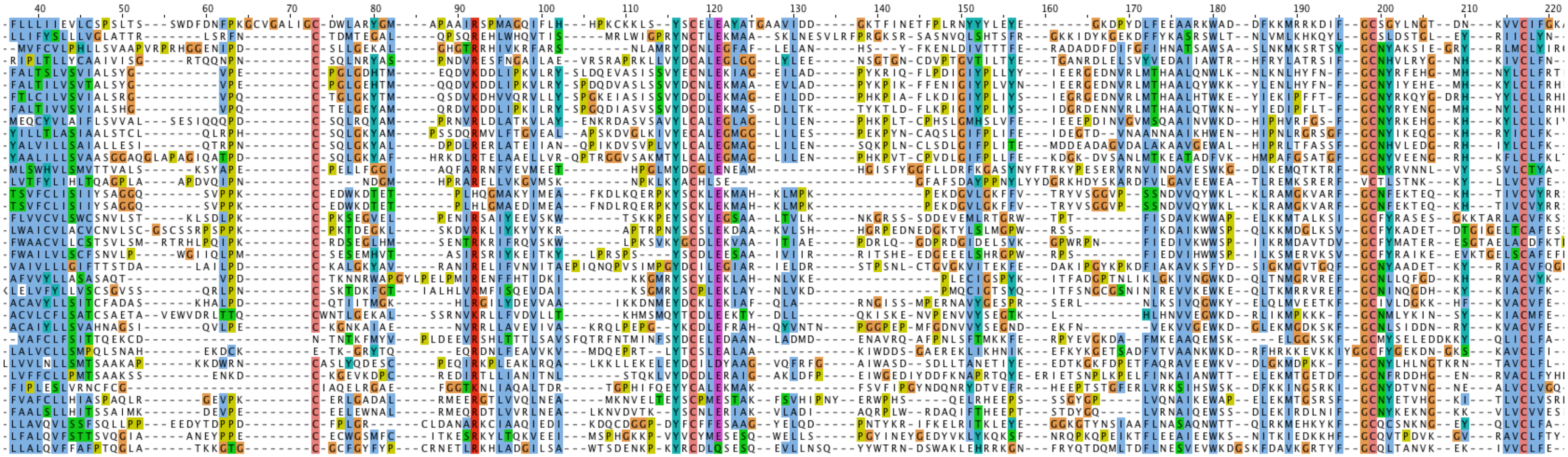
The proteins in ORTHOMCL896.14spp are
generally predicted to be secreted, and ~200 a.a. long;
but are otherwise non-descript (neither HMMER3 nor
InterProScan classes them as ASPs, etc.).

So, look at them with iterative psi-BLAST against a compendium
of nematode proteins:

With a threshold of E ≤ 10-12, closed set, no obvious homologies.
With a threshold of E ≤ 10-9, still closed, but one ASP.
With a threshold of E ≤ 10-6, many ASPs.
Thus, this is a cryptic ASP-like subfamily!
So call them: **ASPRs.**

ASPRs are a diverse subfamily

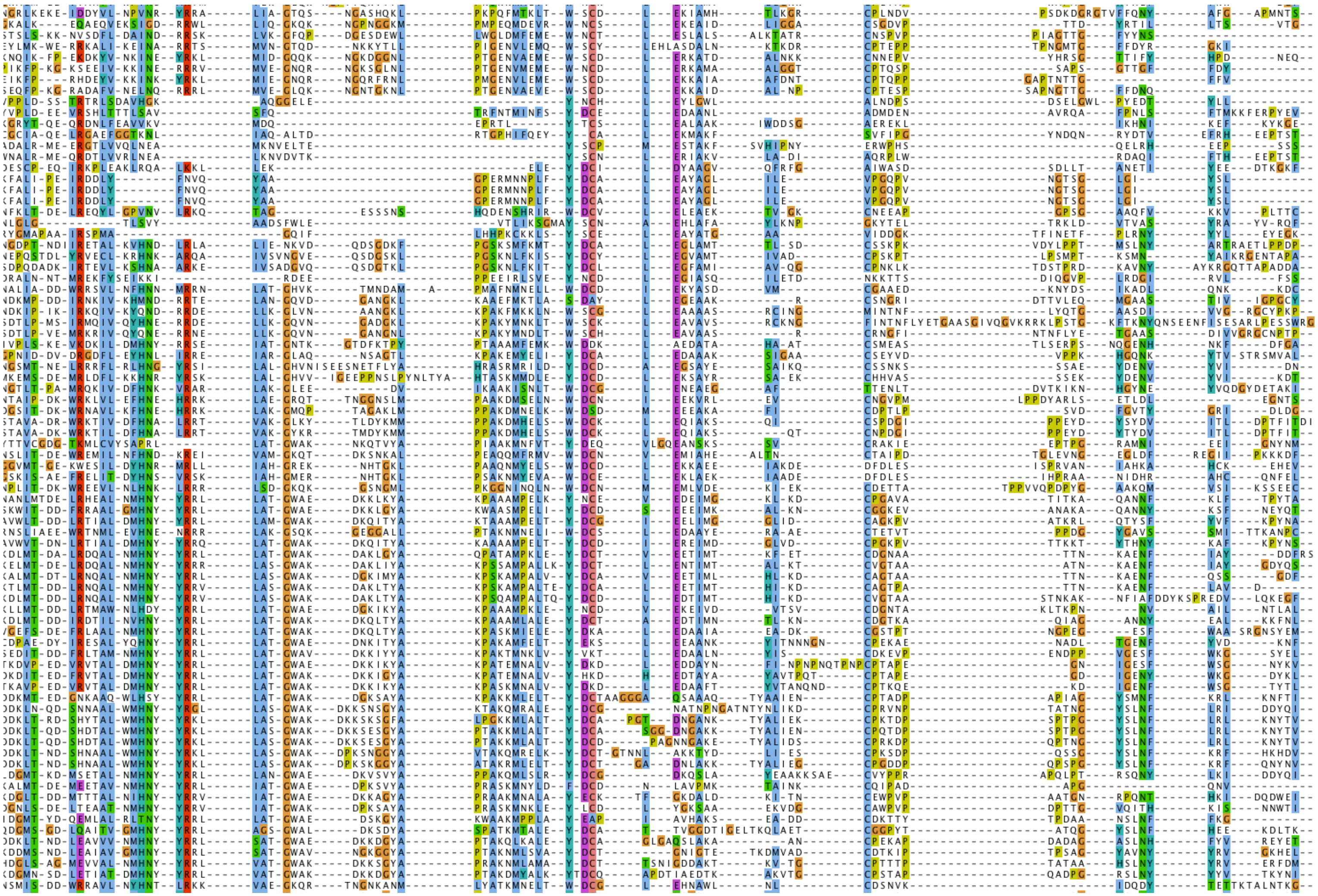
By aligning with MUSCLE, then editing the alignment with JalView, a set of 36 readily alignable ASPRs emerges:



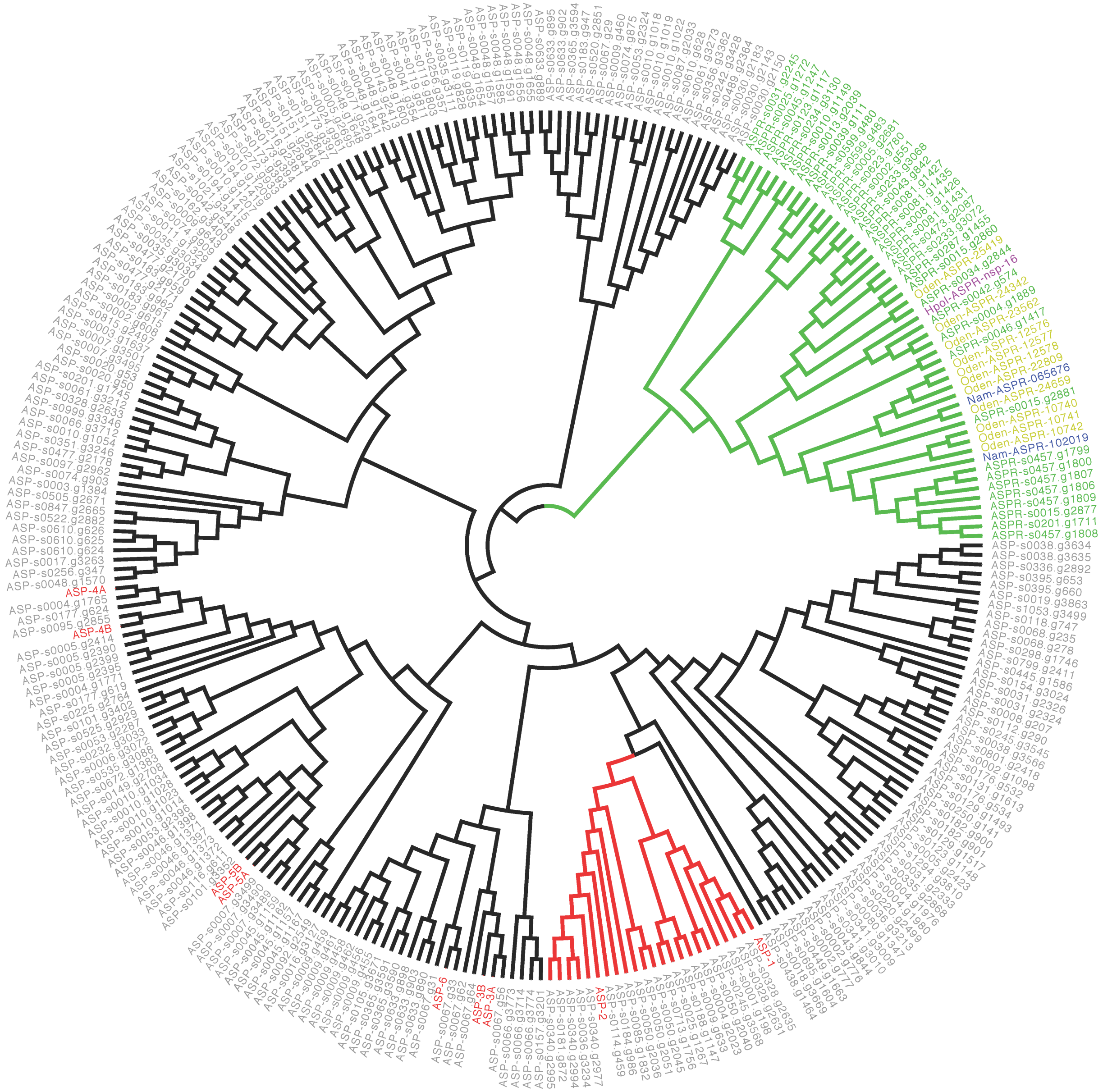
Refs.: Edgar (2004), BMC Bioinformatics 5, 113; Waterhouse et al. (2009), Bioinformatics *25*, 1189-1191.

ASPs and ASPRs are a superfamily

These 36 ASPRs can be further aligned with 235 ASP homologs:



ASPs and ASPRs are a superfamily



ASPRs include one known excretory-secretory (ES) protein from the parasitic nematode *Heligmosomoides polygyrus bakeri*

This ASPR was published by Hewitson and coworkers as
a completely unclassifiable protein, "novel secreted protein 16",
identified by ES proteomics.

General prediction of secretion for ASPRs,
obvious similarity to a known ES component,
subtle similarity to ES components ASP-1 and ASP-2,
and strong upregulation during early infection,

are all consistent with the hypothesis that
ASPRs comprise a new component of hookworm infection.
They are therefore candidates for vaccines.

Ref.: Hewitson et al. (2011), J Proteomics *74*, 1573-1594.

Look for something Interesting: phylogenetic groups with evidence of HGT

HGT has been seen extensively for bacterial genes in plant parasites; also seen for bacterial and beetle genes in *Pristionchus pacificus*

So look for genes with mammalian but not *C. elegans* orthologies or protein motifs; check instances by hand with BlastP

Look for something Interesting: phylogenetic groups with evidence of HGT

HGT has been seen extensively for bacterial genes
in plant parasites; also seen for bacterial and beetle genes in
Pristionchus pacificus

So look for genes with mammalian but not *C. elegans* orthologies
or protein motifs; check instances by hand with BlastP

Find the following *A. ceylanicum* genes:

Two genes =~ macrophage mannose receptor
Two genes =~ asialoglycoprotein receptor 2
Three genes =~ neurocans and other C-lectins

Look for something Interesting: phylogenetic groups with evidence of HGT

HGT has been seen extensively for bacterial genes
in plant parasites; also seen for bacterial and beetle genes in
Pristionchus pacificus

So look for genes with mammalian but not *C. elegans* orthologies
or protein motifs; check instances by hand with BlastP

Find the following *A. ceylanicum* genes:

Two genes =~ macrophage mannose receptor
Two genes =~ asialoglycoprotein receptor 2
Three genes =~ neurocans and other C-lectins
Six of these are upregulated from 5.D to 12.D

Look for something Interesting: phylogenetic groups with evidence of HGT

HGT has been seen extensively for bacterial genes in plant parasites; also seen for bacterial and beetle genes in *Pristionchus pacificus*

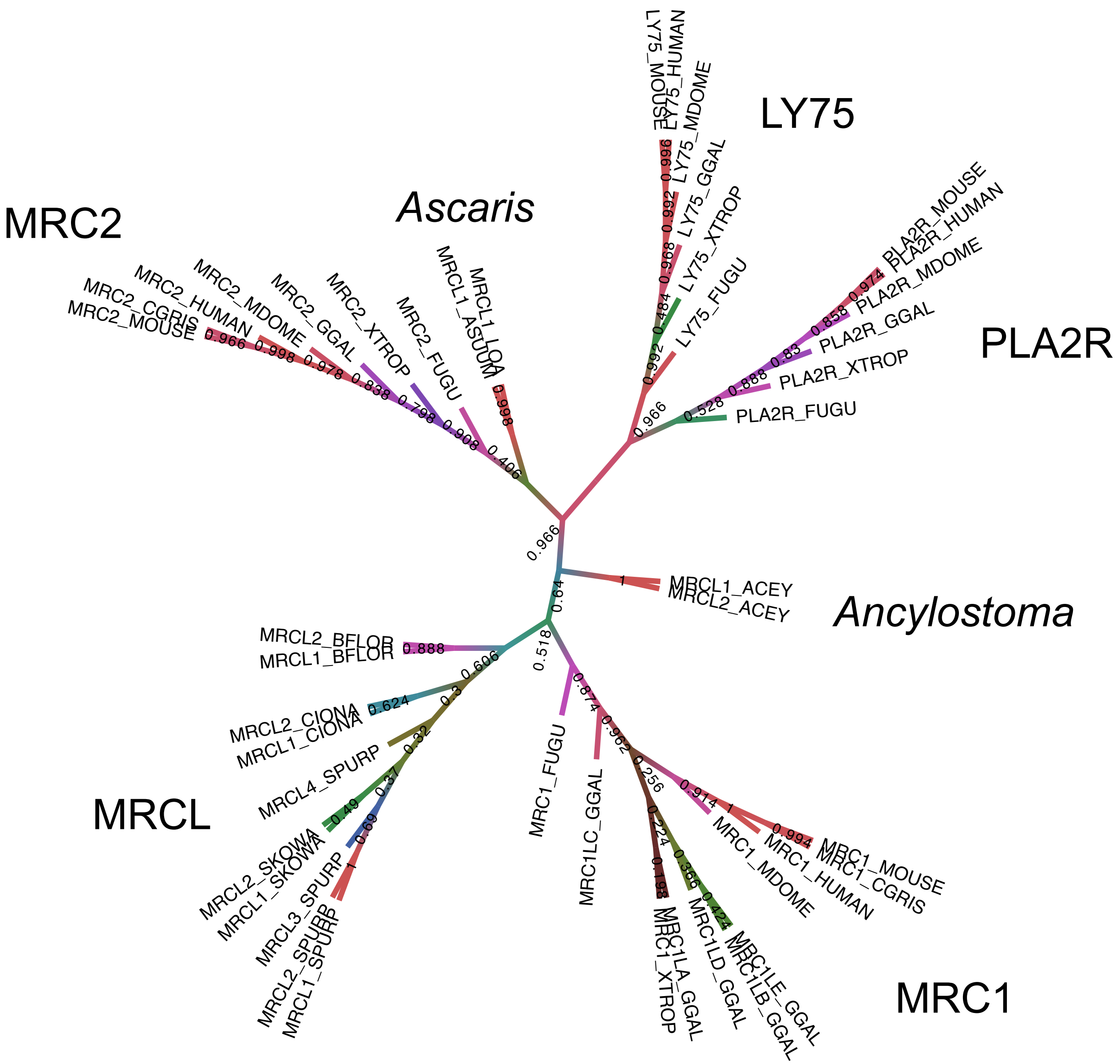
So look for genes with mammalian but not *C. elegans* orthologies or protein motifs; check instances by hand with BlastP

Find the following *A. ceylanicum* genes:

Two genes =~ macrophage mannose receptor
Two genes =~ asialoglycoprotein receptor 2
Three genes =~ neurocans and other C-lectins
Six of these are upregulated from 5.D to 12.D

(And one more =~ N-acetylmuramoyl-L-alanine amidase *amiD*!)

Nematode MR xenologs are distinct

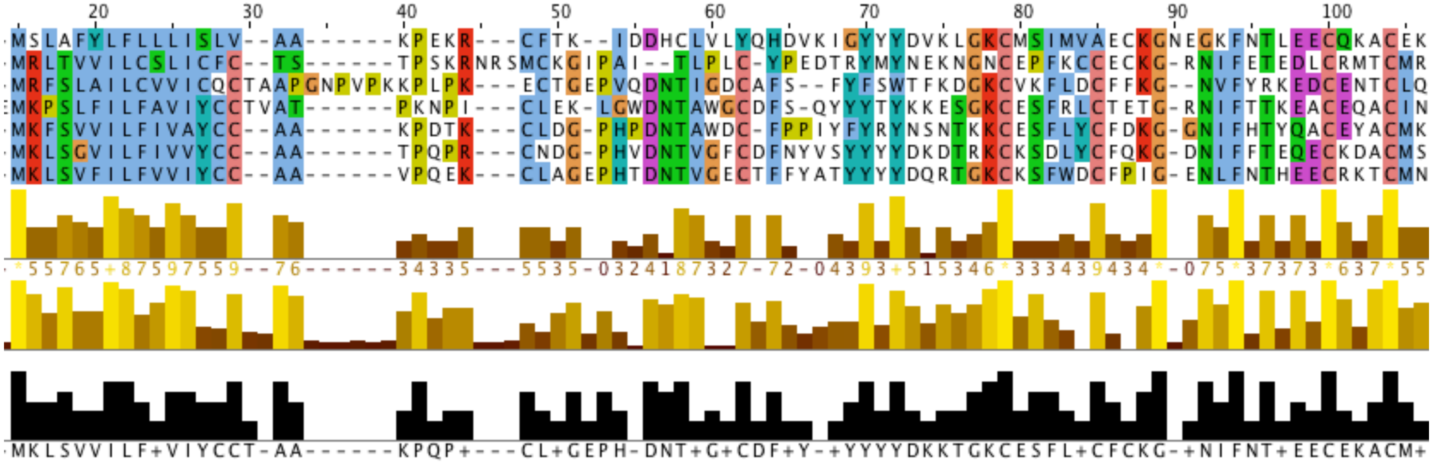


Proteases, and protease inhibitors

12 cathepsin B-like proteases are significantly upregulated by 5.D,
have no obvious mammalian homologs,
do have four homologs in *H. contortus* significantly upregulated during infection,
and may be required for digestion of host proteins or immunosuppression.

7 small protease inhibitors (shown below) are upregulated by 5.D,
have no mammalian homologs,
and do have one *H. contortus* homolog upregulated during infection.

Not exotic xenologs, but still well worth vaccinating against, given that
both proteases and protease inhibitors are likely to be crucial for infection.


20
30
40
50
60
70
80
90
100
MSLAFYLFLLLISLV--AA------KPEKR---CFTK--IDDHCLVLYQHDVKIGYYYDVKLGKCMSIMVAECKGNEGKFNTLEECQKACEK
MRLTVVILCSLICFC--TS------TPSKRNRSMCKGIPAI--TLPLC-YPEDTRYMYNEKNGNCEPFKCCECKG-RNIFETEDLCRMTCMR
MRFSLAILCVVICQCTAAPGNPVPKKPLPK---ECTGEPVQDNTIGDCAFS--FYFSWTFKDGKCVKFLDCFFKG--NVFYRKEDCENTCLQ
MKPSLFILFAVIYCCTVAT------PKNPI---CLEK-LGWDNTAWGCDFS-QYYYTYKKESGKCESFRLCTETG-RNIFTTKEACEQACIN
MKFSVVILFIVAYCC--AA------KPDTK---CLDG-PHPDNTAWDC-FPPIYFYRYNSNTKKCESFLYCFDKG-GNIFHTYQACEYACMK
MKLSGVILFIVVYCC--AA------TPQPR---CNDG-PHVDNTVGFCDFNYVSYYYYDKDTRKCKSDLYCFQKG-DNIFFTEQECKDACMS
MKLSVFILFVVIYCC--AA------VPQEK---CLAGEPHTDNTVGECTFFYATYYYYDQRTGKCKSFWDCFPIG-ENLFNTHEECRKTCMN
*55765+87597559--76------34335---5535-0324187327-72-04393+515346*333439434*-075*37373*637*55
MKLSVVILF+VIYCCT-AA------KPQP+---CL+GEPH-DNT+G+CDF+Y-+YYYYDKKTGKCESFL+CFCKG-+NIFNT+EECEKACM+

Overview

1. What hookworms are and why we care

2. *Haemonchus contortus* adventures

3. *Ancylostoma ceylanicum* genome and transcriptome

4. Vaccine targets: ASPRs, and horizontally transmitted genes

5. Some thoughts on 'descriptive genomics'

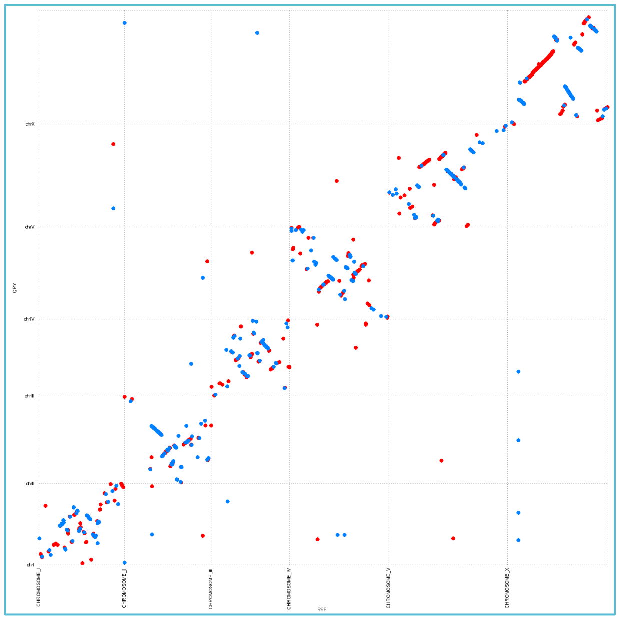
Begin and end with checks for basic quality

Living organisms sit in a soup of microbes

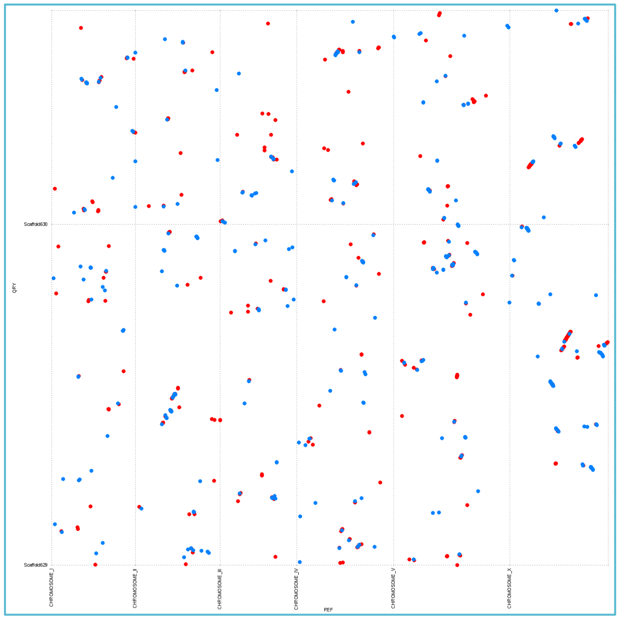
Microbial contamination slowed both *C. angaria* and *H. contortus*

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

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/vaccine targets for ~700M sick humans (hookworms)

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/vaccine targets for ~700M sick humans (hookworms)

"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.

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/vaccine targets for ~700M sick humans (hookworms)

"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:

Yan Hu, Raffi Aroian	*Ancylostoma ceylanicum*
Pasi Korhonen, Bronwyn Campbell, Neil Young, Ross Hall and Robin Gasser	*Haemonchus contortus*
Paul Sternberg	*A. ceylanicum* and *H. contortus*
Titus Brown, Adina Howe, Jason Pell	*khmer* software and filtering
Brian Williams	cDNA libraries, *H. contortus*
Igor Antoshechkin	Optimized sequencing protocols
Jacobs Genome Center	Illumina sequencing
NIH, HHMI, and Cornell	Funding!