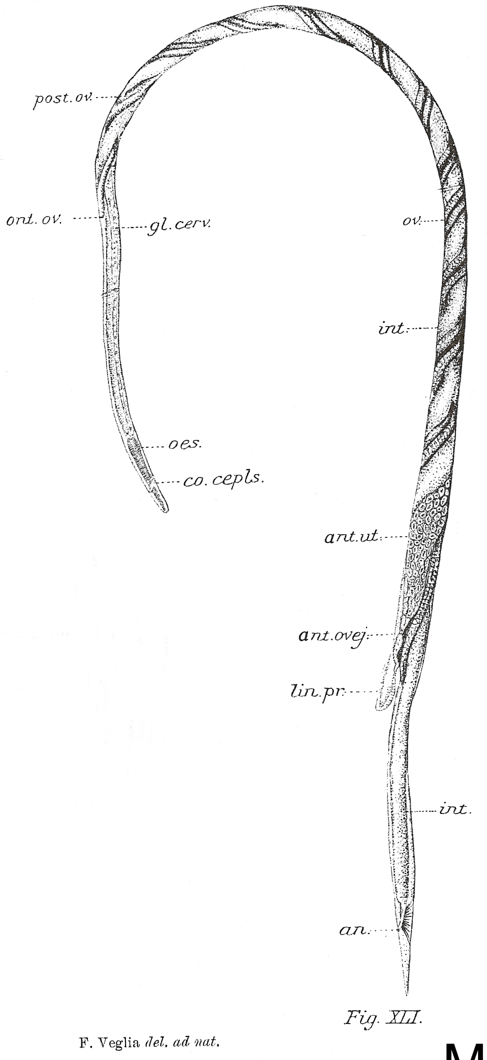
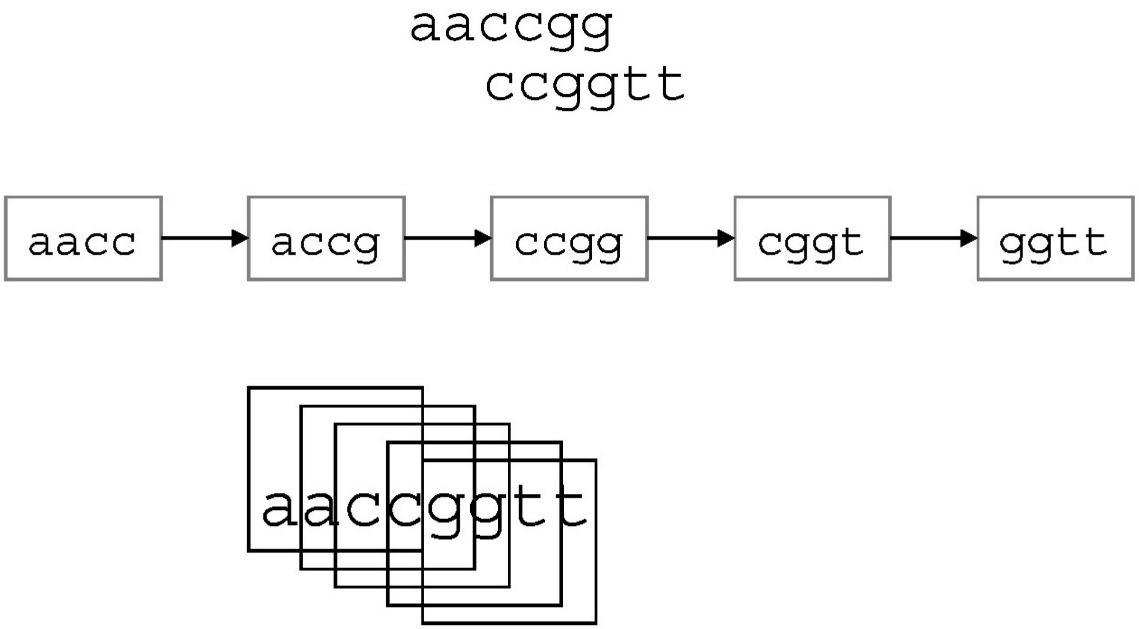
Genome assembly and analysis

Erich Schwarz, Cornell



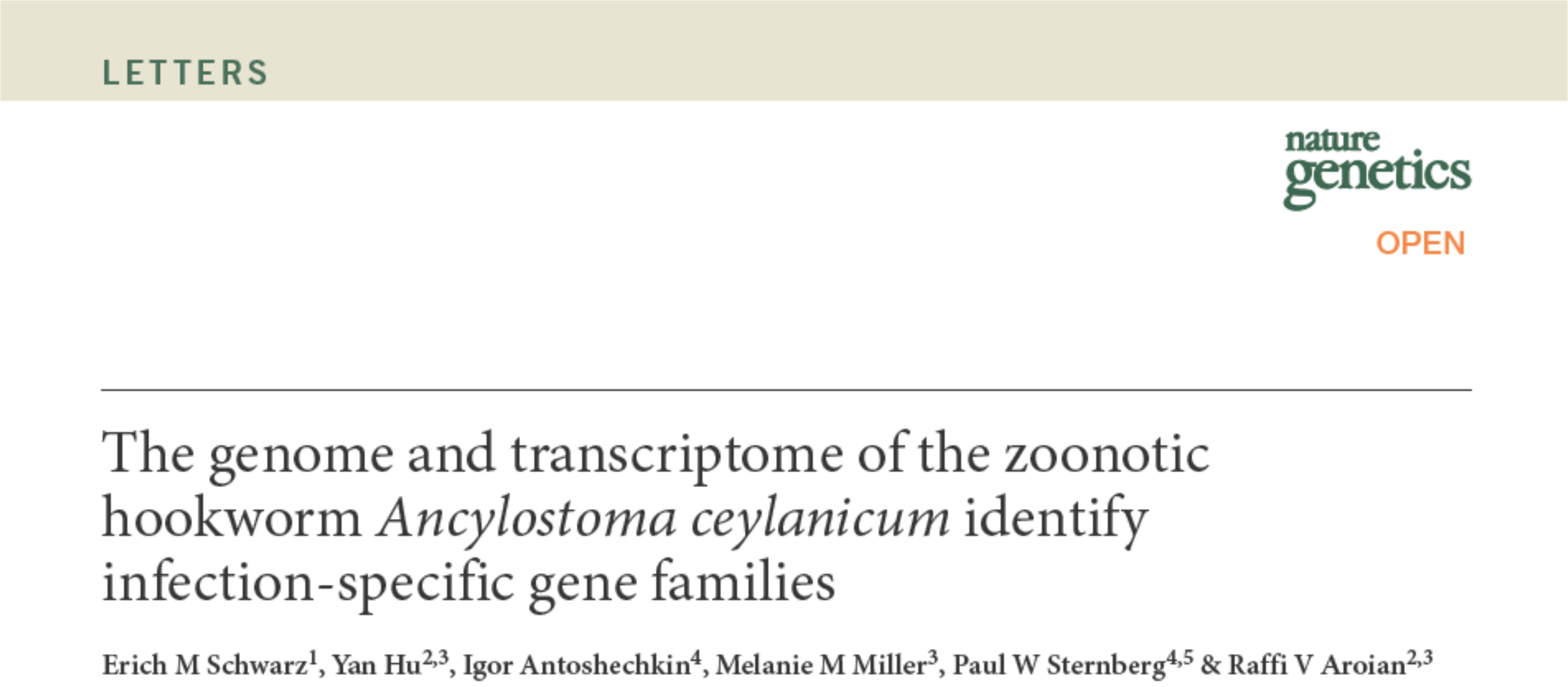

aaccgg
ccggtt

aacc → accg → ccgg → cggt → ggtt

aaccggtt

MSU NGS course, August 2015

Case study: how we characterized a genome



LETTERS

nature genetics

OPEN

The genome and transcriptome of the zoonotic hookworm *Ancylostoma ceylanicum* identify infection-specific gene families

Erich M Schwarz[1], Yan Hu[2,3], Igor Antoshechkin[4], Melanie M Miller[3], Paul W Sternberg[4,5] & Raffi V Aroian[2,3]

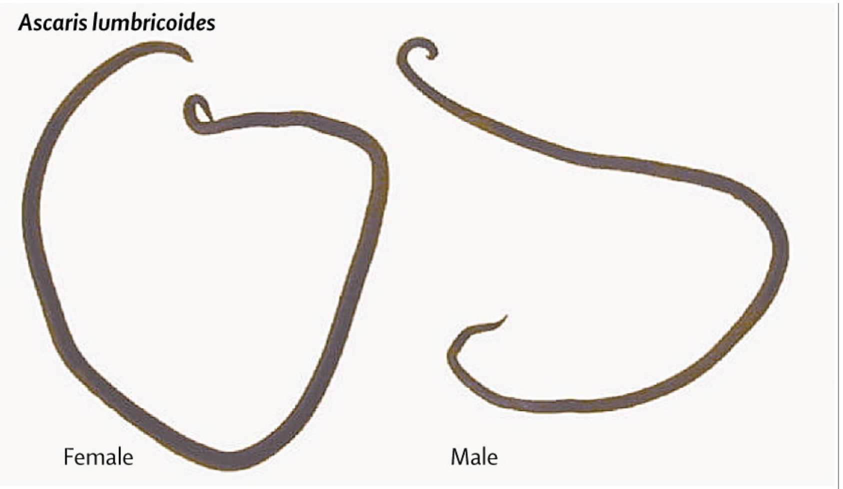
Main text: *http://www.nature.com/ng/journal/v47/n4/full/ng.3237.html*
Supp. text: *http://www.nature.com/ng/journal/v47/n4/extref/ng.3237-S1.pdf*

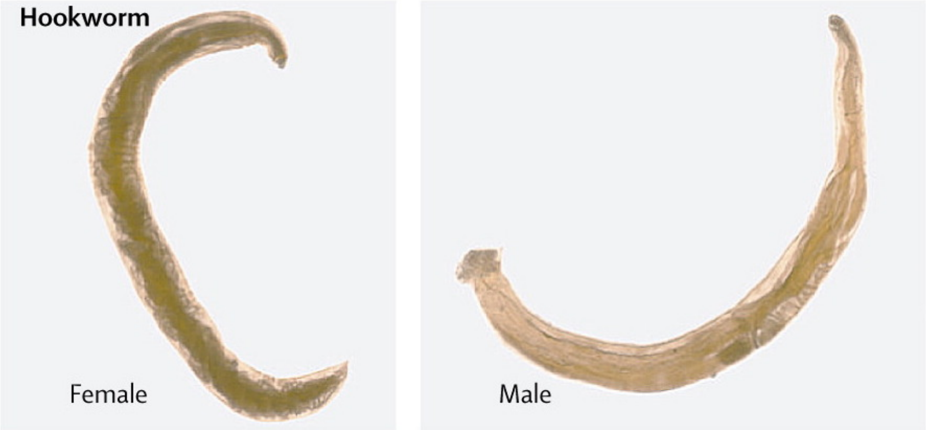
Ref.: Schwarz et al. (2015), Nat. Genet. *47*, 416-422.

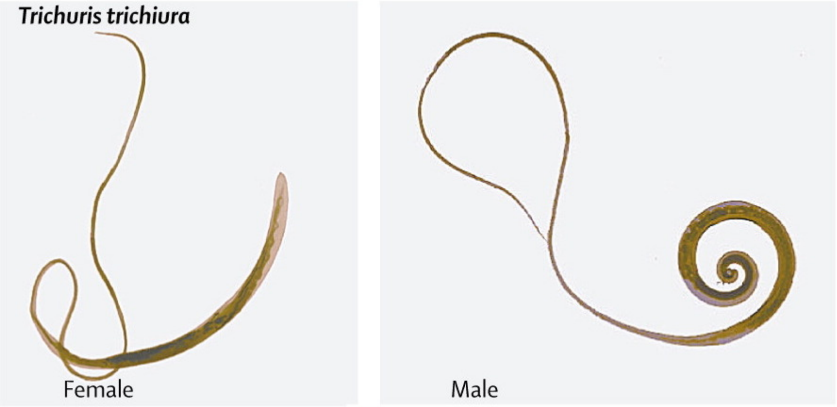
Overview

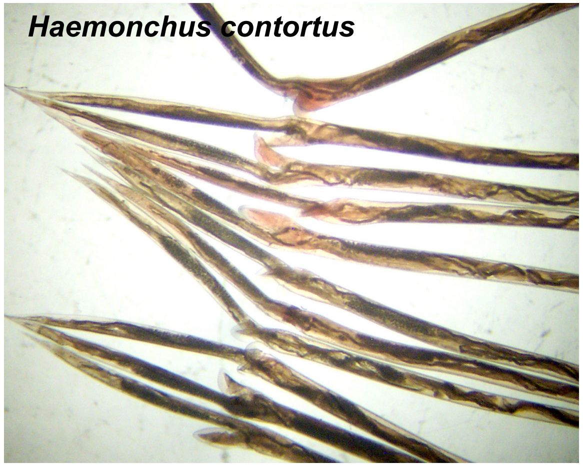
1. Why do we want a hookworm genome?
2. Generating a genome and transcriptome
3. Characterizing the genome
4. Characterizing the transcriptome
5. Predicting drug and vaccine targets
6. Some thoughts on 'descriptive genomics'

Parasitic nematodes infect over one billion human beings, as well as farm animals



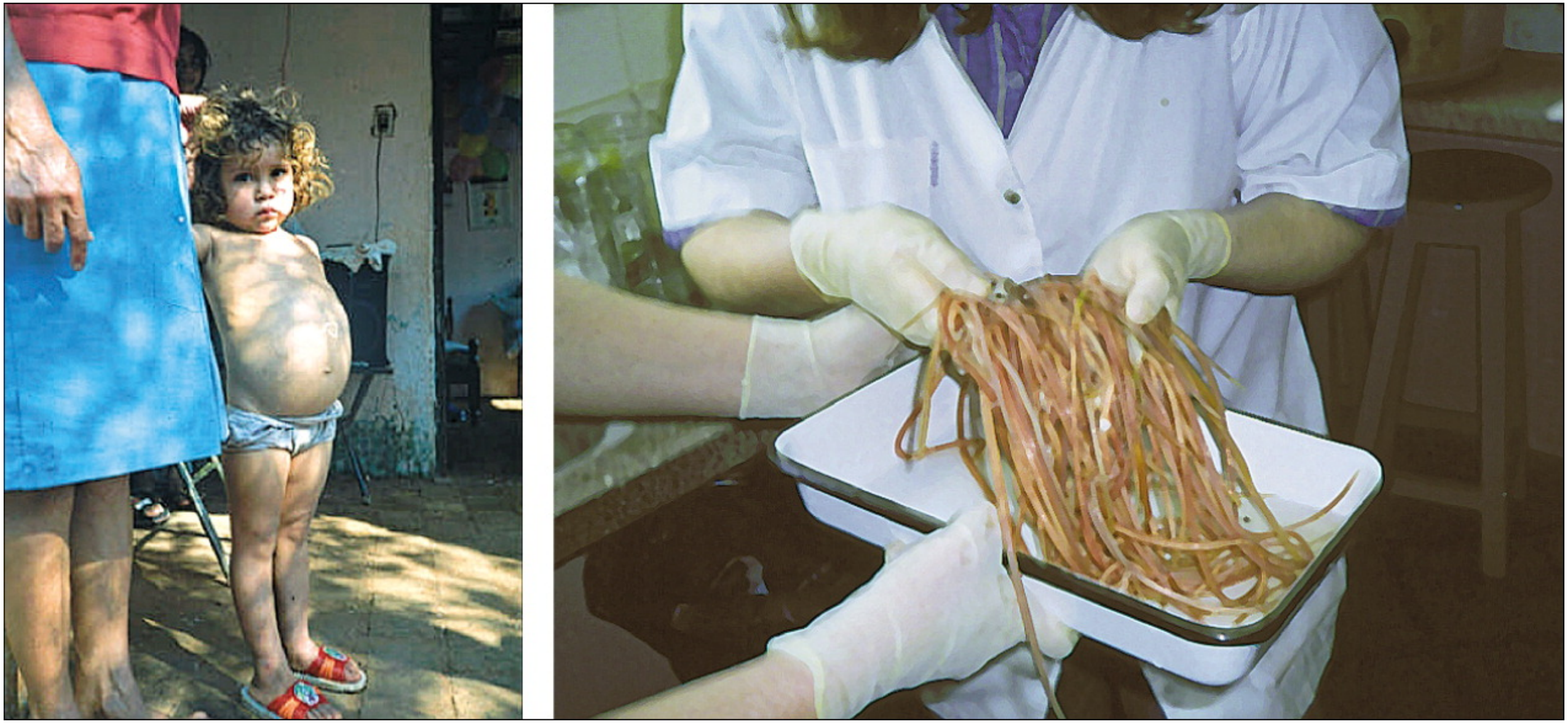






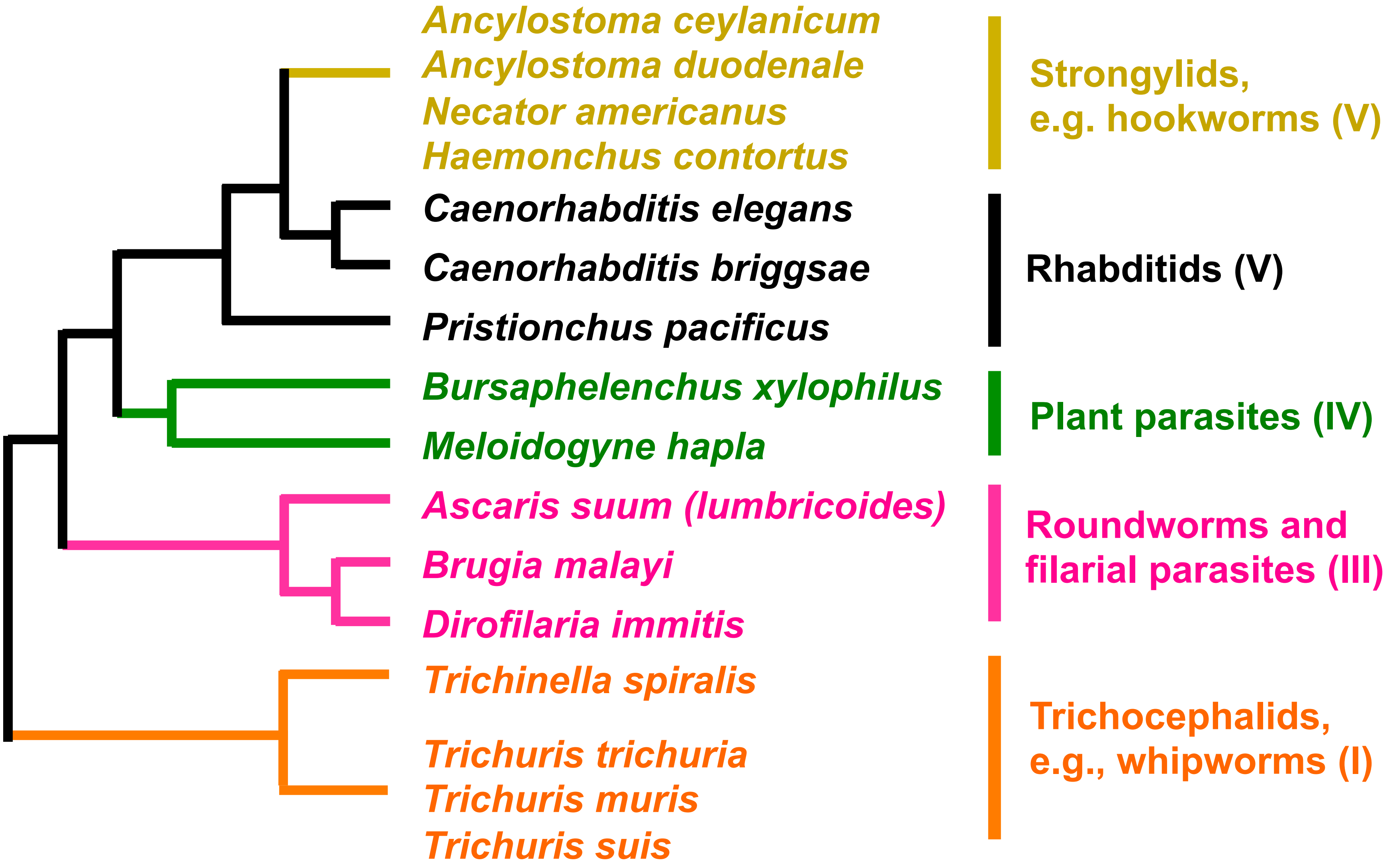
Refs.: CSIRO; Despommier et al. (2005), *Parasitic Diseases* (5th edn.);
Bethony et al. (2006), Lancet *367*, 1521-1532; Vos et al. (2012), Lancet *380*, 2163-2196.

Parasitic nematodes can blind, stunt, or stultify humans; they can kill sheep or goats, and sicken other farm animals



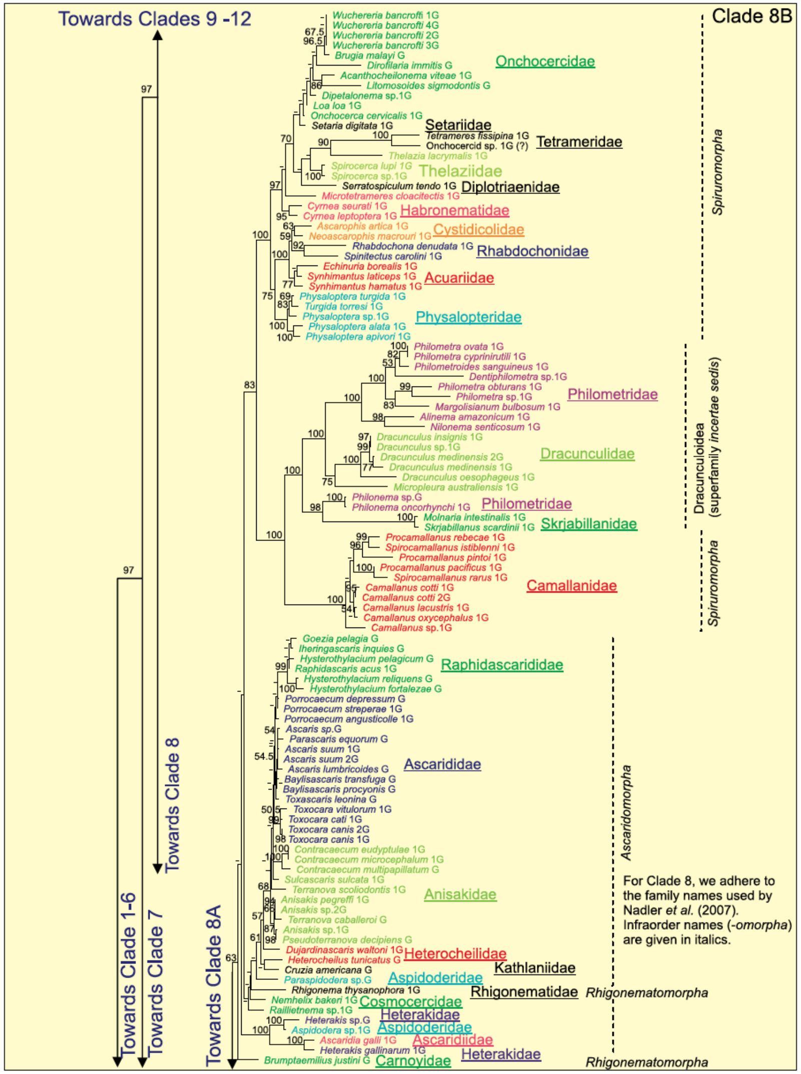
Refs.: Bethony et al. (2006), Lancet *367*, 1521-1532; Vos et al. (2012), Lancet *380*, 2163-2196.

Parasitism has evolved in nematodes several times, independently

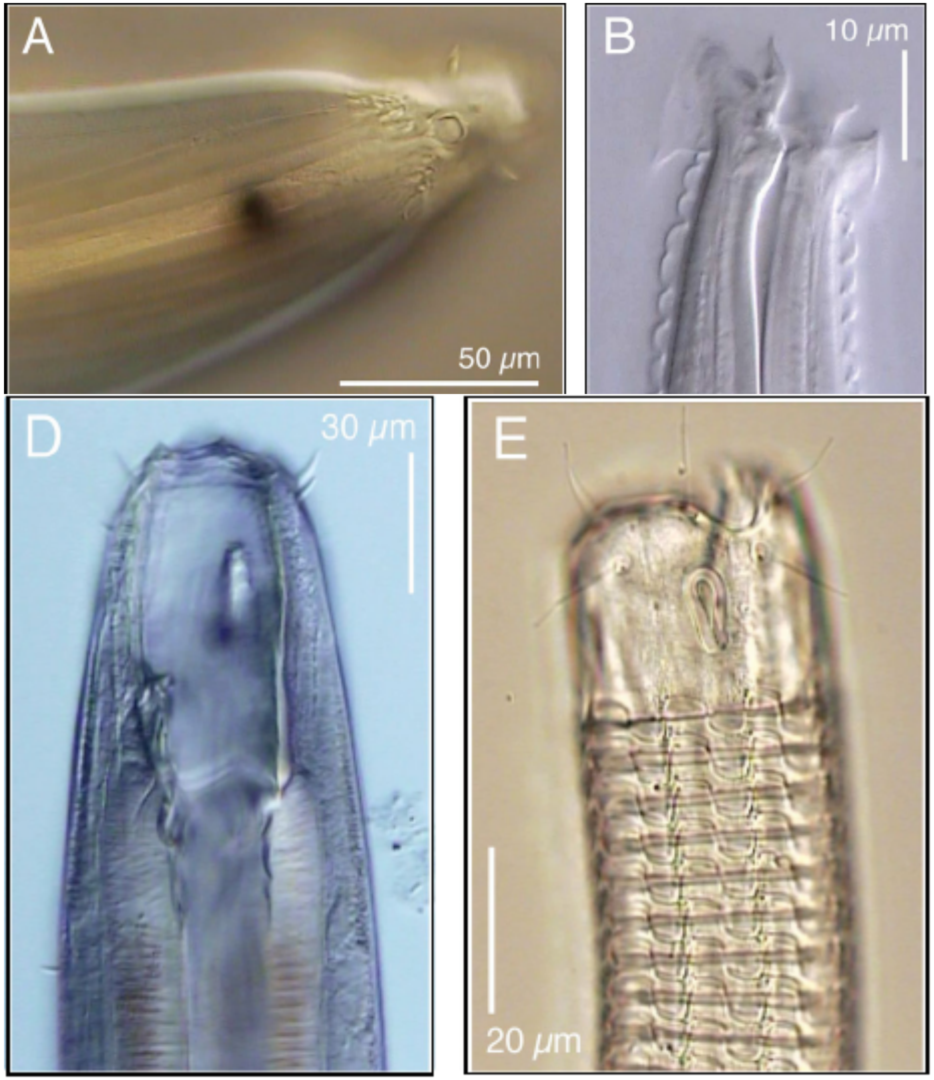


Refs.: Kiontke et al. (2011), BMC Evol. Biol. *11*, 339; van Megen et al. (2009), Nematology *11*, 927-950.

Known parasitic nematodes are a tiny subset of vast species diversity (~1 M species?)

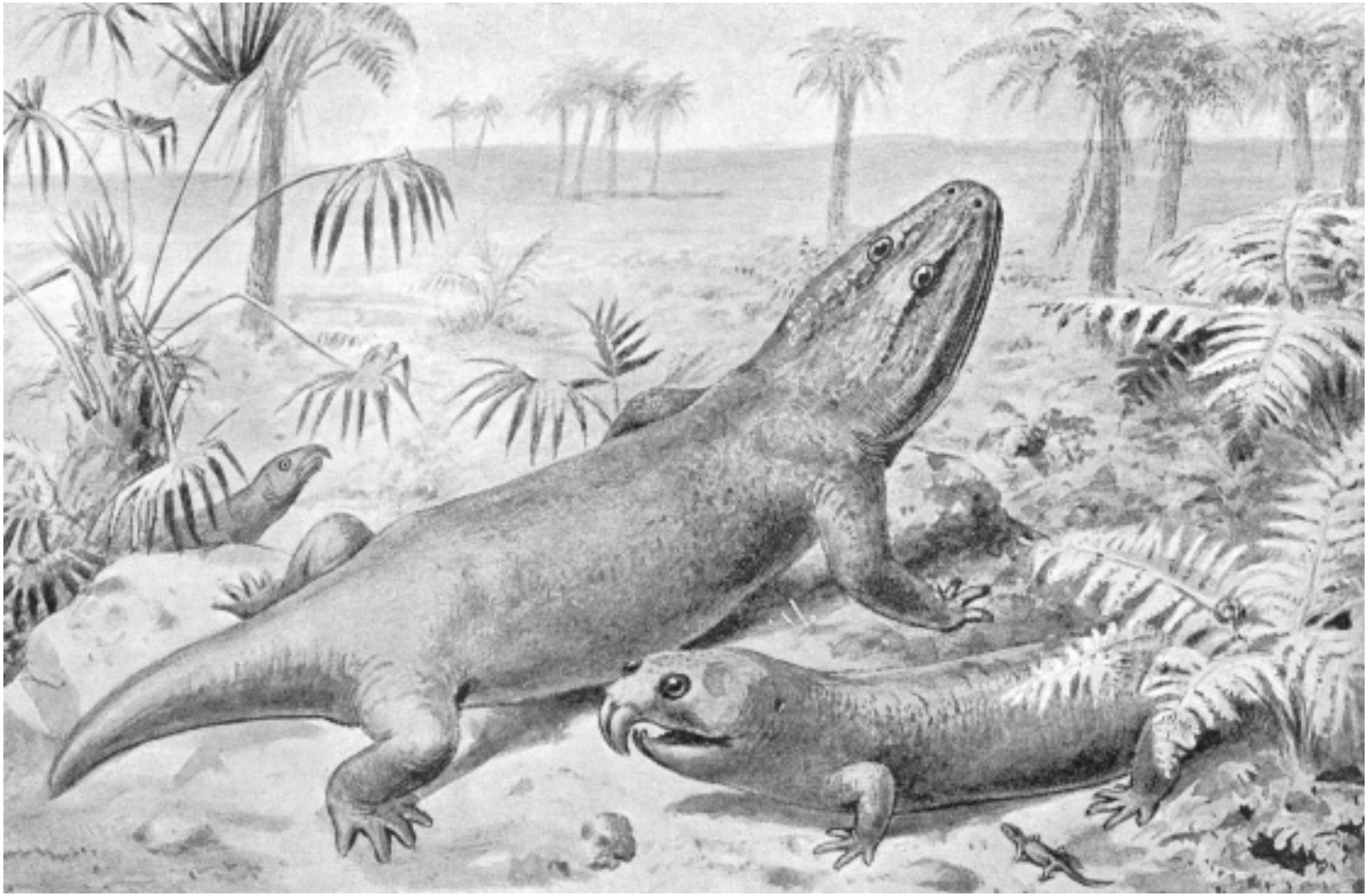


For Clade 8, we adhere to the family names used by Nadler *et al.* (2007). Infraorder names (*-omorpha*) are given in italics.



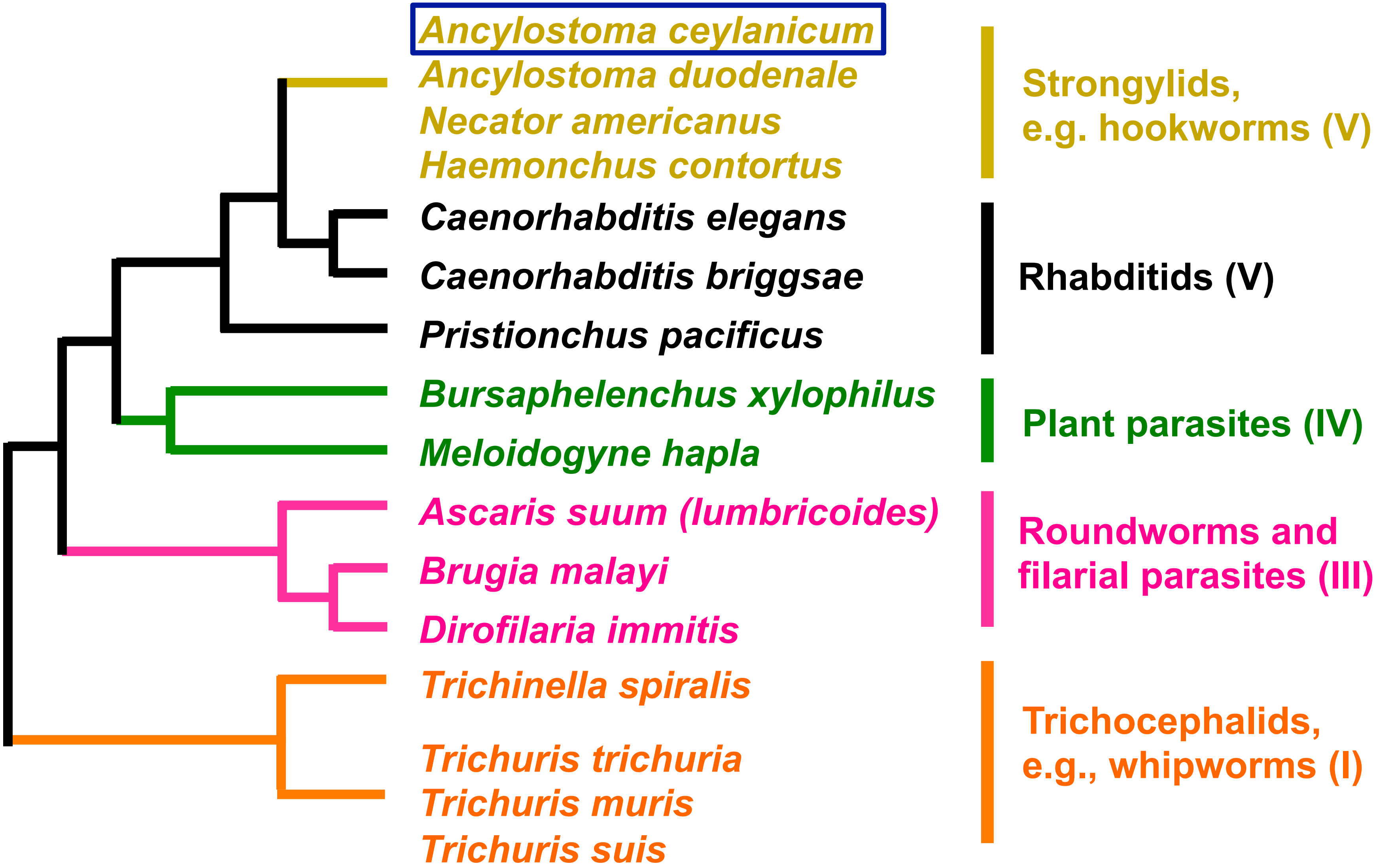
Refs.: Lambshead (1993), Oceanis *19*, 5–24; De Ley (2006), WormBook, *2006 Jan 25*,1-8; van Megen et al. (2009), Nematology *11*, 927-950.

In (at least) strongylids, parasitism of vertebrates may have arisen ~350 million years ago



Refs.: Durette-Desset et al. (1994), Int. J. Parasitol. 24, 1139-1165;
image of *Mastodonsaurus* and *Rhynchosauria* from Smit (1894), *Creatures of Other Days*.

Ancylostoma ceylanicum, a model hookworm that infects several mammals



Ancylostoma ceylanicum
Ancylostoma duodenale
Necator americanus
Haemonchus contortus

Strongylids, e.g. hookworms (V)

Caenorhabditis elegans
Caenorhabditis briggsae
Pristionchus pacificus

Rhabditids (V)

Bursaphelenchus xylophilus
Meloidogyne hapla

Plant parasites (IV)

Ascaris suum (lumbricoides)
Brugia malayi
Dirofilaria immitis

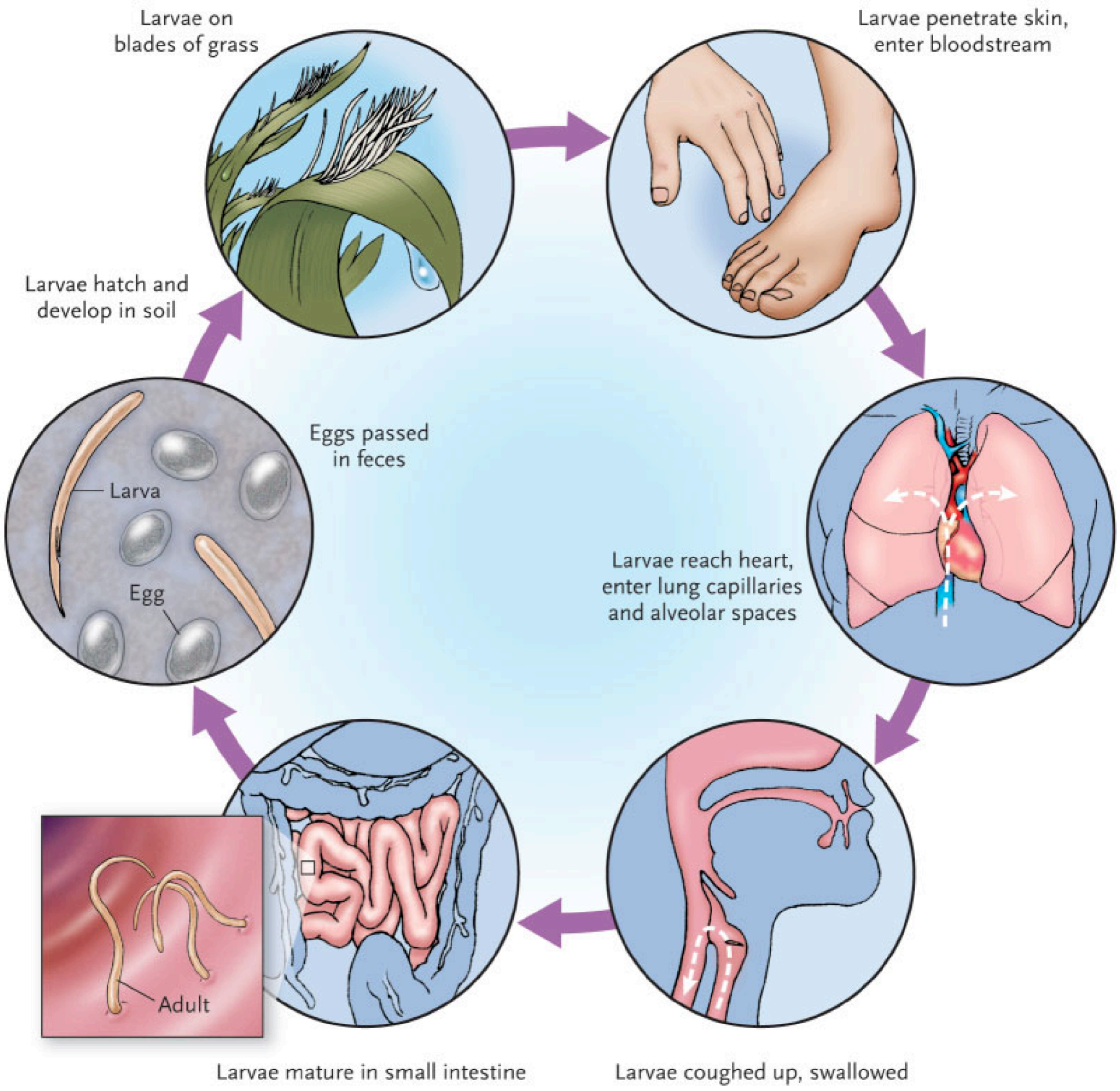
Roundworms and filarial parasites (III)

Trichinella spiralis
Trichuris trichuria
Trichuris muris
Trichuris suis

Trichocephalids, e.g., whipworms (I)

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

Hookworms infect over 400 million human beings


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

Refs.: Hotez et al. (2004), N. Engl. J. Med. *351*, 799-807; Bethony et al. (2006), Lancet *367*, 1521-1532; Vos et al. (2012), Lancet *380*, 2163-2196; Pullan et al. (2014), Parasit. Vectors *7*, 37.

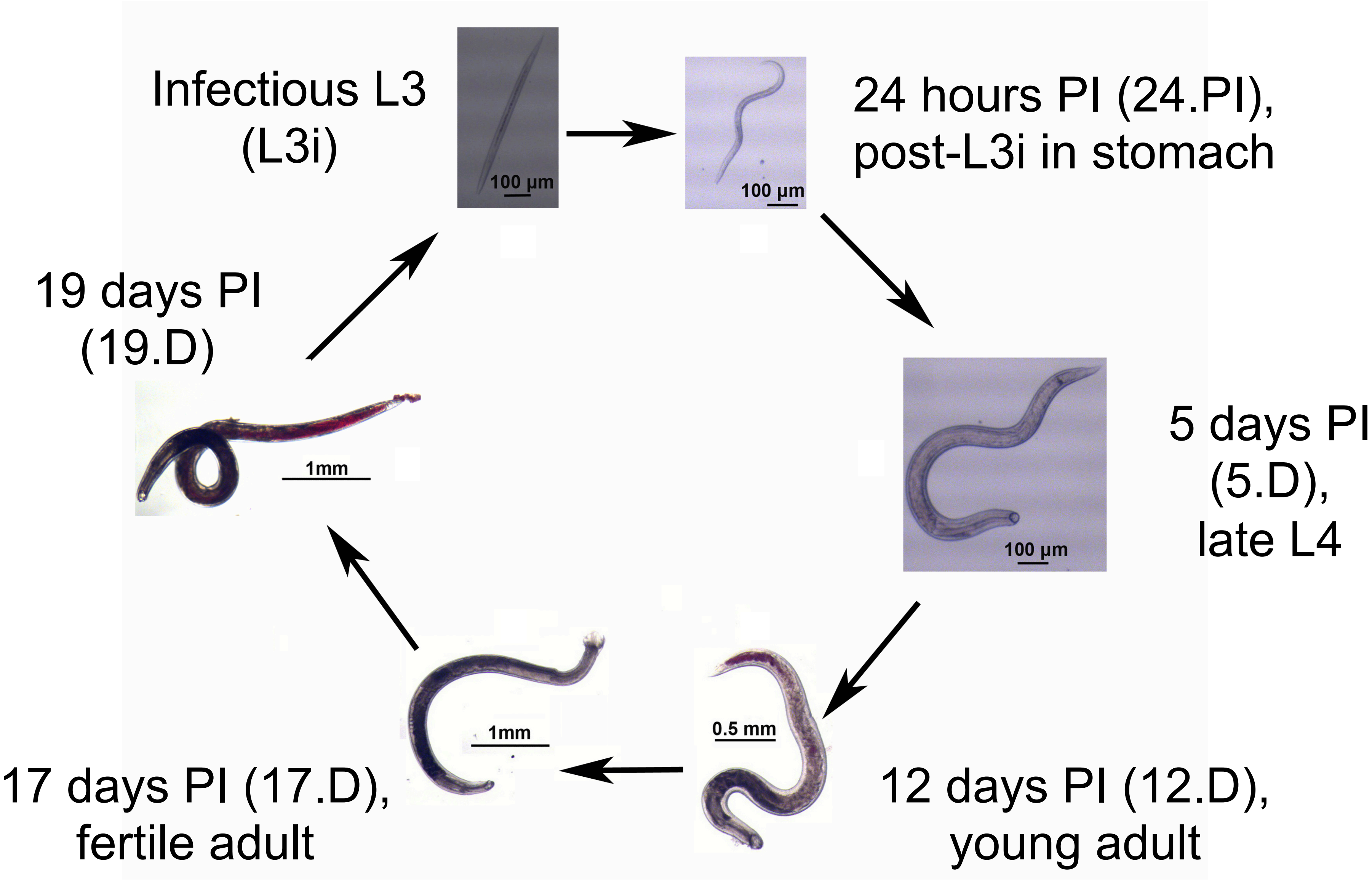
Hookworms treated with one drug, albendazole; no vaccine against them exists (yet)

Refs.: Keiser and Utzinger, (2010), Adv. Parasitol. *73*, 197-230;
Schneider et al. (2011), Hum. Vaccin. *7*, 1234-1244.

Overview

1. Why do we want a hookworm genome?
2. Generating a genome and transcriptome
3. Characterizing the genome
4. Characterizing the transcriptome
5. Predicting drug and vaccine targets
6. Some thoughts on 'descriptive genomics'

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

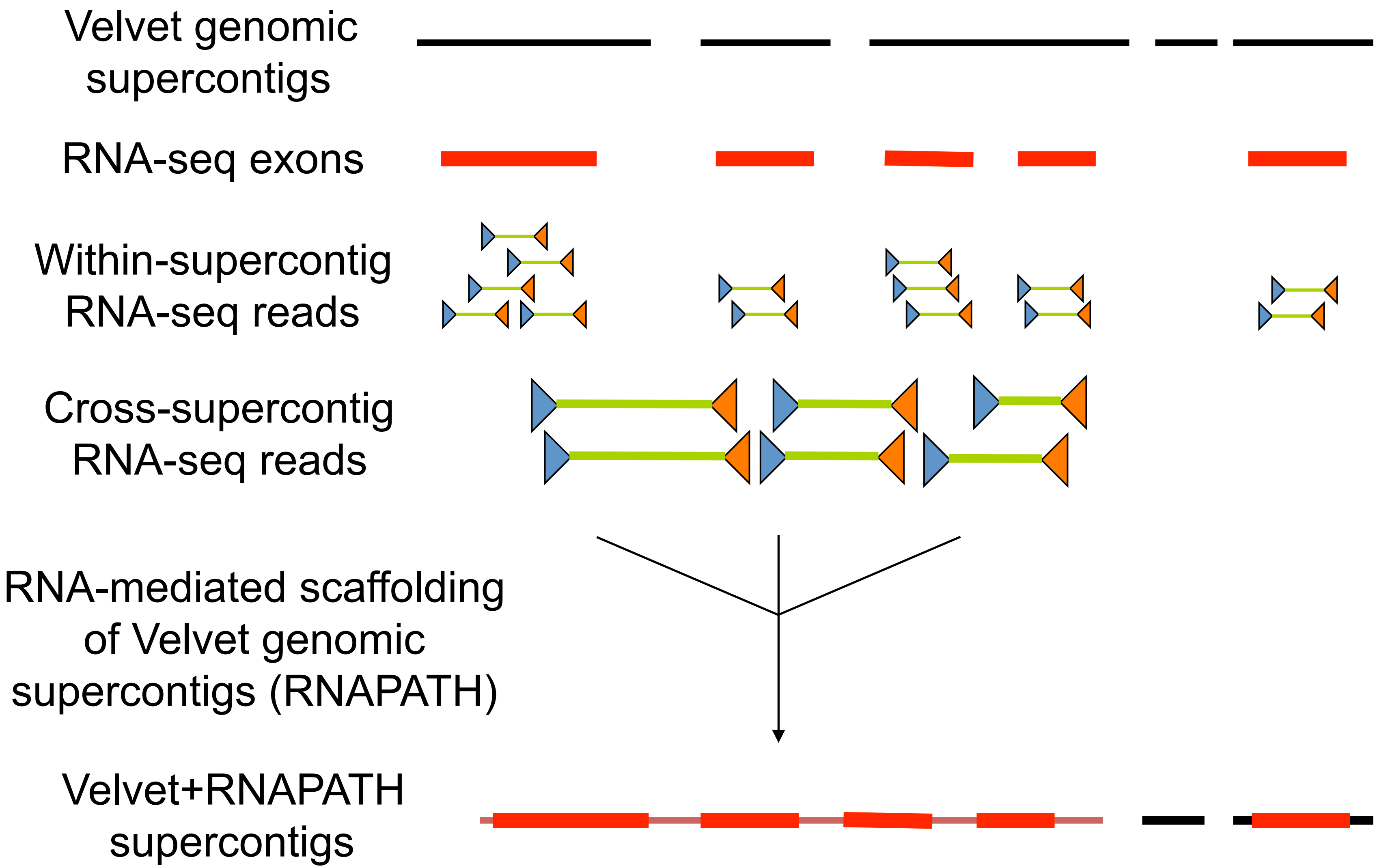
BGI GapCloser 1.12 (release_2011). Ref.: Li et al. (2010). Genome Res. *20*, 265-272.

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

ERANGE 3.2. 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.

Overview

1. Why do we want a hookworm genome?
2. Generating a genome and transcriptome
3. Characterizing the genome
4. Characterizing the transcriptome
5. Predicting drug and vaccine targets
6. Some thoughts on 'descriptive genomics'

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 ~70 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 ~70 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*

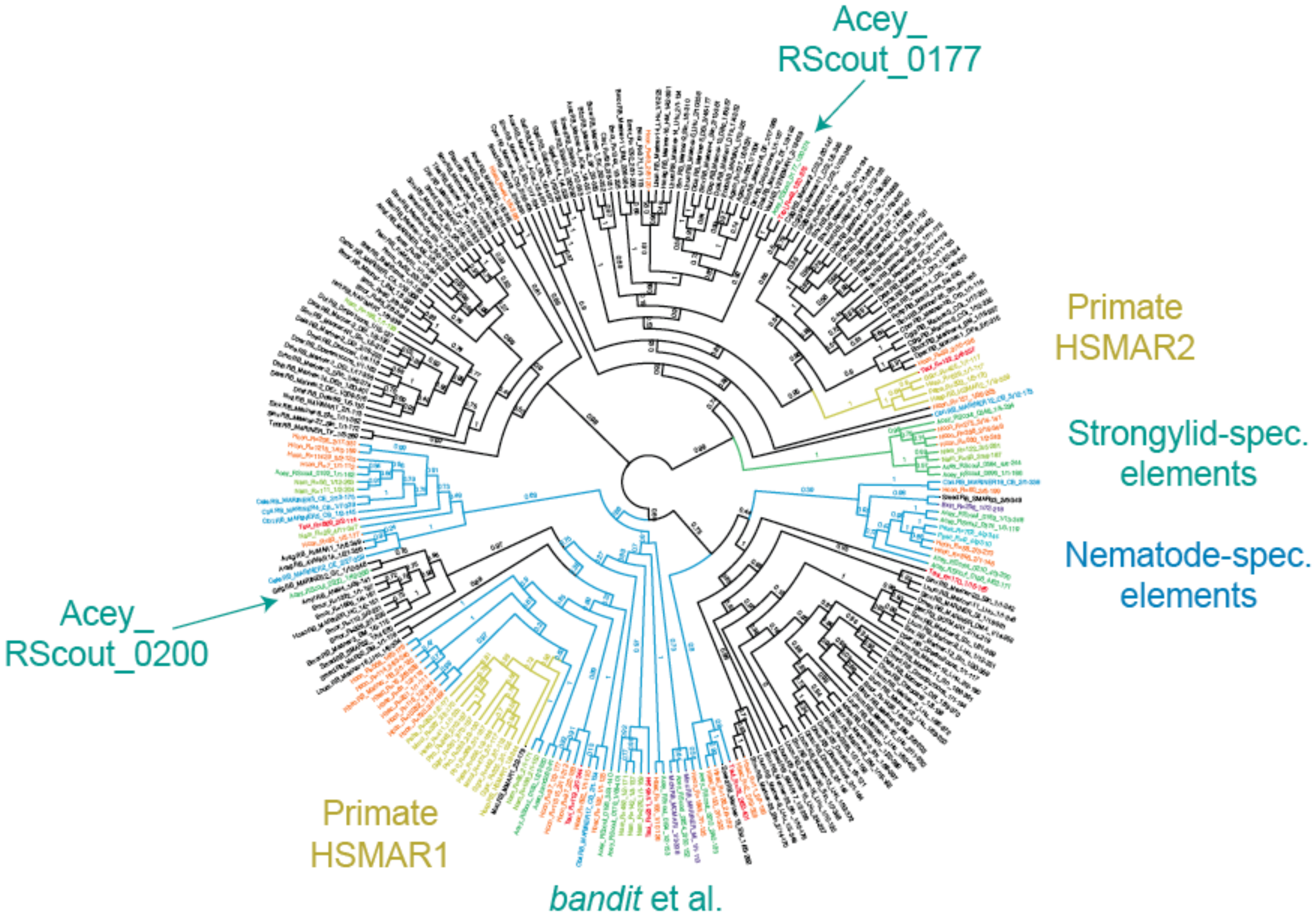
40.5% of genomic DNA is repetitive, over twice the 17% in *C. elegans* or *P. pacificus;* without this difference, *A. ceylanicum*'s genome would be ~70 Mb smaller.

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 a prominent 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.

Hookworms have HSMAR-like repeats with both nematode and mammalian relatives



ML phylogenies via FastTree 2.0. Ref.: Price et al. (2010), PLoS One *5*, e9490.

Genome comparison

	Hookworm: *Ancylostoma ceylanicum*	Roundworm: *Ascaris suum*	Whipworm: *Trichuris muris*	Strongylid: *Haemonchus contortus*	Free-living: *Caenorhabditis elegans*
Total nt:	313 Mb	266 Mb	84.7 Mb	370 Mb	100 Mb
Genes:	27.0* K	15.4 K	11.0 K	21.8 K	20.0 K
Scaffolds:	1.74 K	31.5 K	1.68 K	23.9 K	7
Contigs:	32.2 K	40.6 K	4.38 K	65.5 K	7
% non-N:	96.1	99.2	99.4	93.6	100.0
Scaf. N50 nt:	668 Kb	291 Kb	401 Kb	83.3 Kb	17.5 Mb
Scaf. max. nt:	4.80 Mb	1.46 Mb	1.77 Mb	0.95 Mb	20.9 Mb
Cont. N50 nt:	18.5 kb	46.5 kb	47.8 kb	20.8 kb	[17.5 Mb]
Cont. max. nt:	125 kb	304 kb	304 kb	136 kb	[20.9 Mb]

Refs.: Wang et al. (2012), Dev. Cell *23*, 1072-1080; Laing et al. (2013), Genome Biol. *14*, R88;
Foth et al. (2014), Nat. Genet. *46*, 693-700; Schwarz et al. (2015), Nat. Genet., *47*, 416-422.

Genome comparison

	Hookworm: *Ancylostoma ceylanicum*	Roundworm: *Ascaris suum*	Whipworm: *Trichuris muris*	Strongylid: *Haemonchus contortus*	Free-living: *Caenorhabditis elegans*
Total nt:	313 Mb	266 Mb	84.7 Mb	370 Mb	100 Mb
Genes:	27.0* K	15.4 K	11.0 K	21.8 K	20.0 K
Scaffolds:	1.74 K	31.5 K	1.68 K	23.9 K	7
Contigs:	32.2 K	40.6 K	4.38 K	65.5 K	7
% non-N:	96.1	99.2	99.4	93.6	100.0
Scaf. N50 nt:	[illegible]	[illegible]	[illegible]	[illegible]	17.5 Mb
Scaf. max. nt:	4.80 Mb	1.46 Mb	1.77 Mb	0.95 Mb	20.9 Mb
Cont. N50 nt:	18.5 kb	46.5 kb	47.8 kb	20.8 kb	[17.5 Mb]
Cont. max. nt:	125 kb	304 kb	304 kb	136 kb	[20.9 Mb]

It is generally assumed that parasitism reduces genome sizes.
However, this is not necessarily true for parasitic eukaryotes,
and certainly not true for many sequenced parasitic nematodes.
(Whipworms might be a case where parasitism indeed shrinks the genome.)

Ref.: Raffaele and Kamoun (2012), Nat. Rev. Microbiol. *10*, 417-430.

Genome comparison

	Hookworm: *Ancylostoma ceylanicum*	Roundworm: *Ascaris suum*	Whipworm: *Trichuris muris*	Strongylid: *Haemonchus contortus*	Free-living: *Caenorhabditis elegans*
Total nt:	313 Mb	266 Mb	84.7 Mb	370 Mb	100 Mb
Genes:	27.0* K	15.4 K	11.0 K	21.8 K	20.0 K
Scaffolds:	1.74 K	31.5 K	1.68 K	23.9 K	7
Contigs:	32.2 K	40.6 K	4.38 K	65.5 K	7
% non-N:	96.1	99.2	99.4	93.6	100.0
Scaf. N50 nt:	668 kb	[illegible]	[illegible]	[illegible]	17.5 Mb
Scaf. max. nt:	4.80 Mb	1.46 Mb	1.77 Mb	0.95 Mb	20.9 Mb
Cont. N50 nt:	18.5 kb	46.5 kb	47.8 kb	20.8 kb	[17.5 Mb]
Cont. max. nt:	125 kb	304 kb	304 kb	136 kb	[20.9 Mb]

Most eukaryotic genefinders use an arbitary size minimum of 100 residues for predicted proteins.
This may systematically fail to detect small genes encoding possible effectors of parasitism!

Refs.: Frith et al. (2006), PLoS Genet. *2*, e52; Raffaele and Kamoun (2012), Nat. Rev. Microbiol. *10*, 417-430.

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.

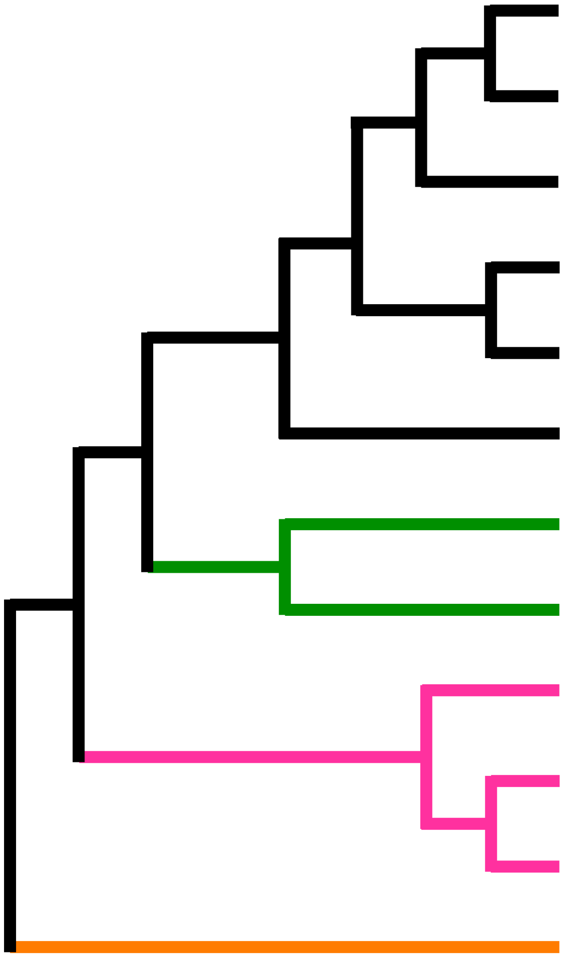
Predict 26,966 protein-coding genes with products of ≥100 a.a.; another 10,050 genes encoding 30-99 a.a.

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.

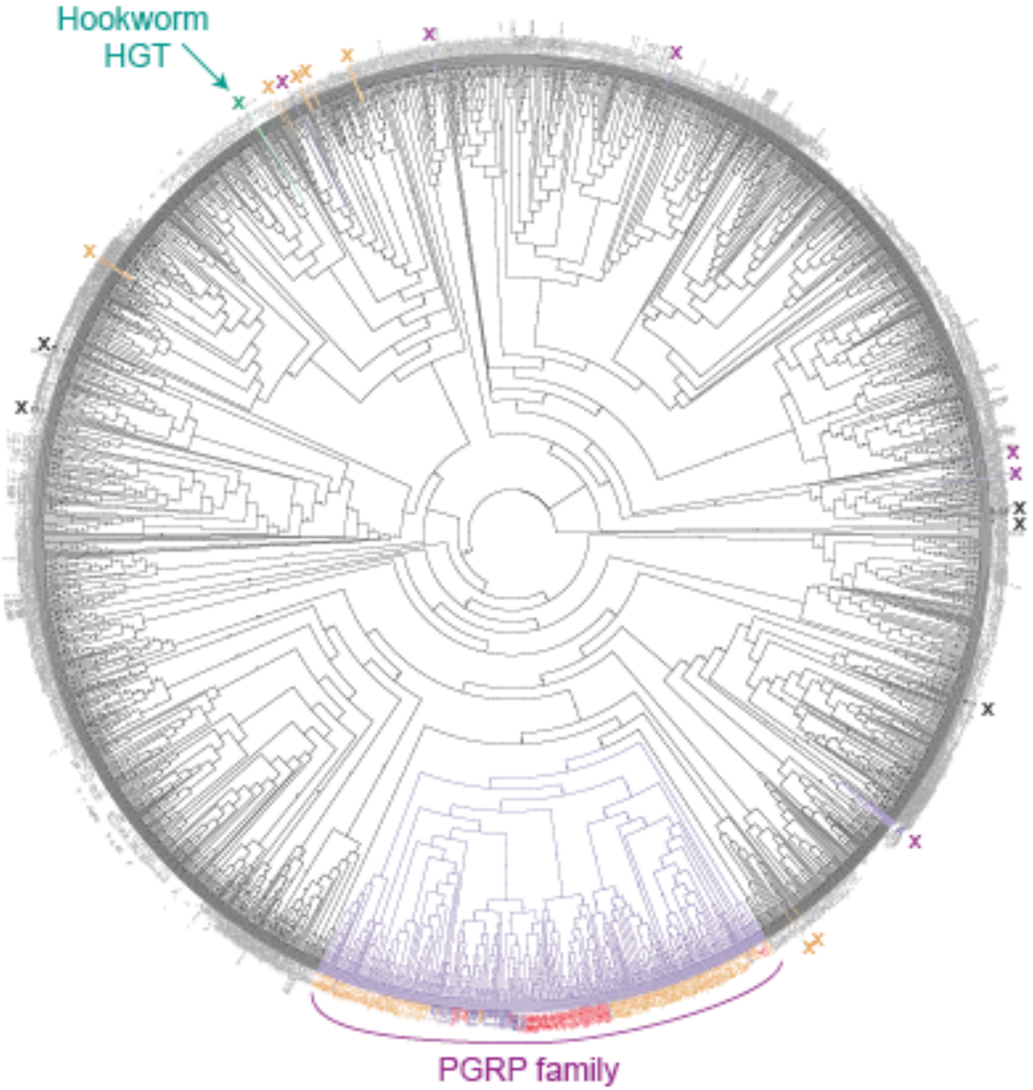
A. ceylanicum and *C. elegans* have similar numbers of genes conserved in other nematodes



	A. cey.	***C. ele.***
A. ceylanicum	*9,860*	7,042
N. americanus	7,506	6,225
H. contortus	5,268	4,576
C. elegans	7,042	*11,323*
C. briggsae	6,933	10,795
P. pacificus	4,833	4,979
B. xylophilus	5,373	5,587
M. hapla	4,161	4,376
A. suum	5,955	6,209
B. malayi	5,312	5,545
D. immitis	5,166	5,308
T. spiralis	3,015	3,186

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

But *amiD* genes in hookworms were transferred horizontally from bacteria!

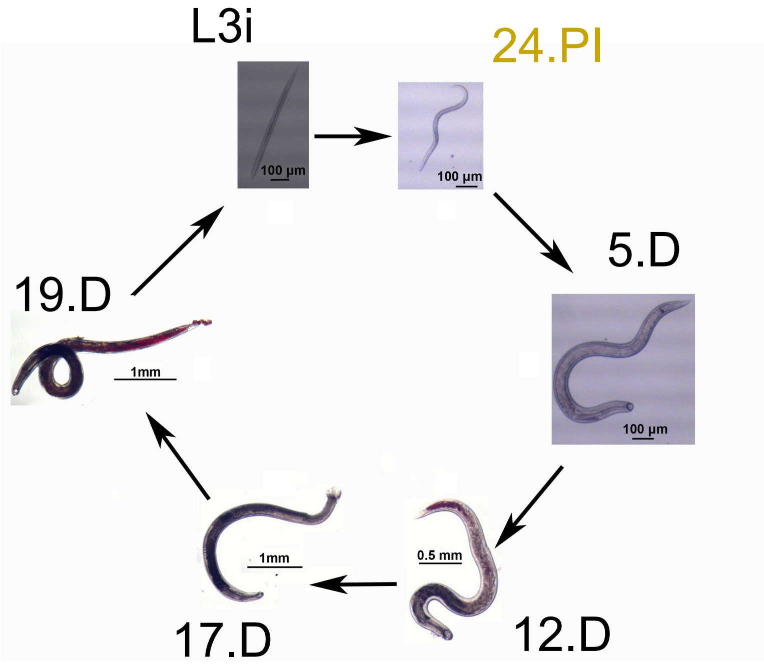


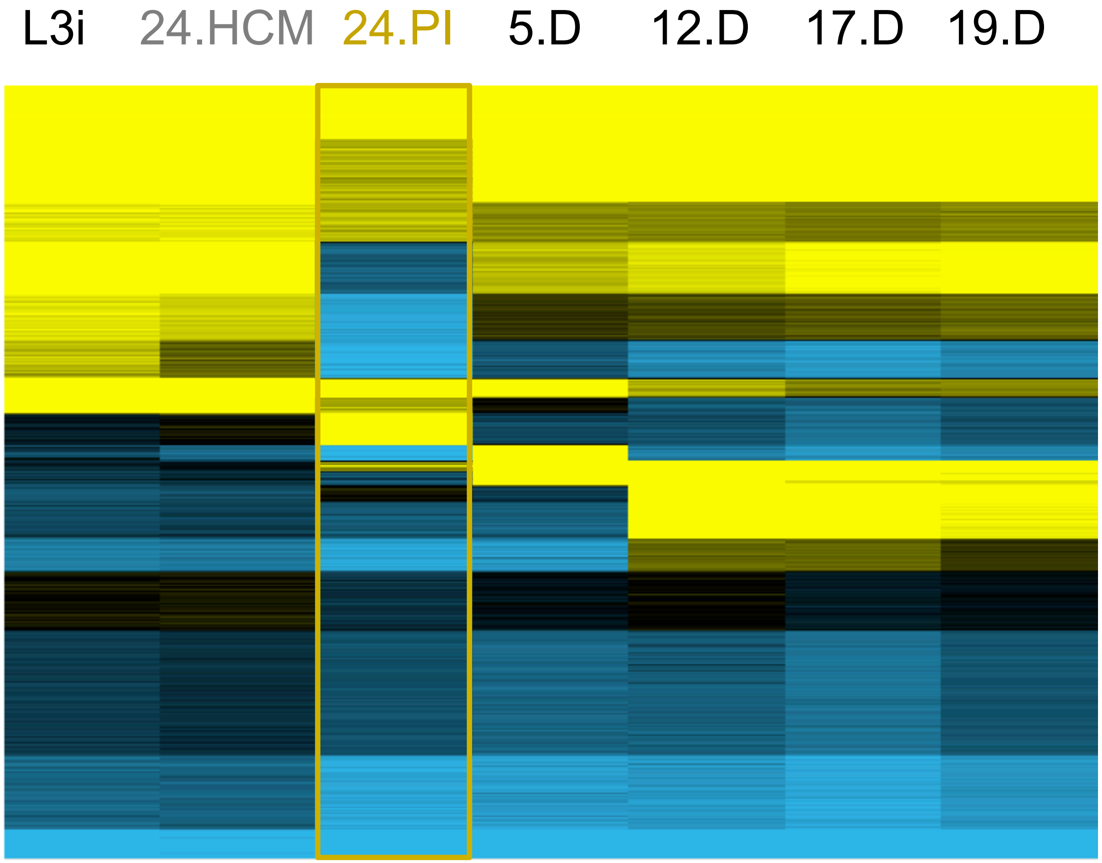
ML phylogenies via FastTree 2.0. Ref.: Price et al. (2010), PLoS One *5*, e9490.

Overview

1. Why do we want a hookworm genome?
2. Generating a genome and transcriptome
3. Characterizing the genome
4. Characterizing the transcriptome
5. Predicting drug and vaccine targets
6. Some thoughts on 'descriptive genomics'

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





L3i to 24.PI: 942 genes up, 1,249 down. L3i to 24.HCM: 240 genes up, 210 down.

RSEM 1.2.0. Ref.: Li and Dewey (2011), BMC Bioinformatics *12*, 323.
NOISeq-sim 2.13, significance ≥0.99. Ref.: Tarazona et al. (2011), Genome Res. *21*, 2213-2223.

Rank-sum statistics 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.

Rank-sum statistics 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, and transcription factors
(N.B.: this is conserved in *N. americanus*, *H. contortus* and *C. elegans*)

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

Rank-sum statistics 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, and transcription factors
(N.B.: this is conserved in *N. americanus*, *H. contortus* and *C. elegans*)

24.PI to late L4 (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.

Rank-sum statistics 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, and transcription factors
(N.B.: this is conserved in *N. americanus*, *H. contortus* and *C. elegans*)

24.PI to late L4 (5.D), upregulated:
structural components of cuticle, binding cytoskeletal proteins, e.g., actin

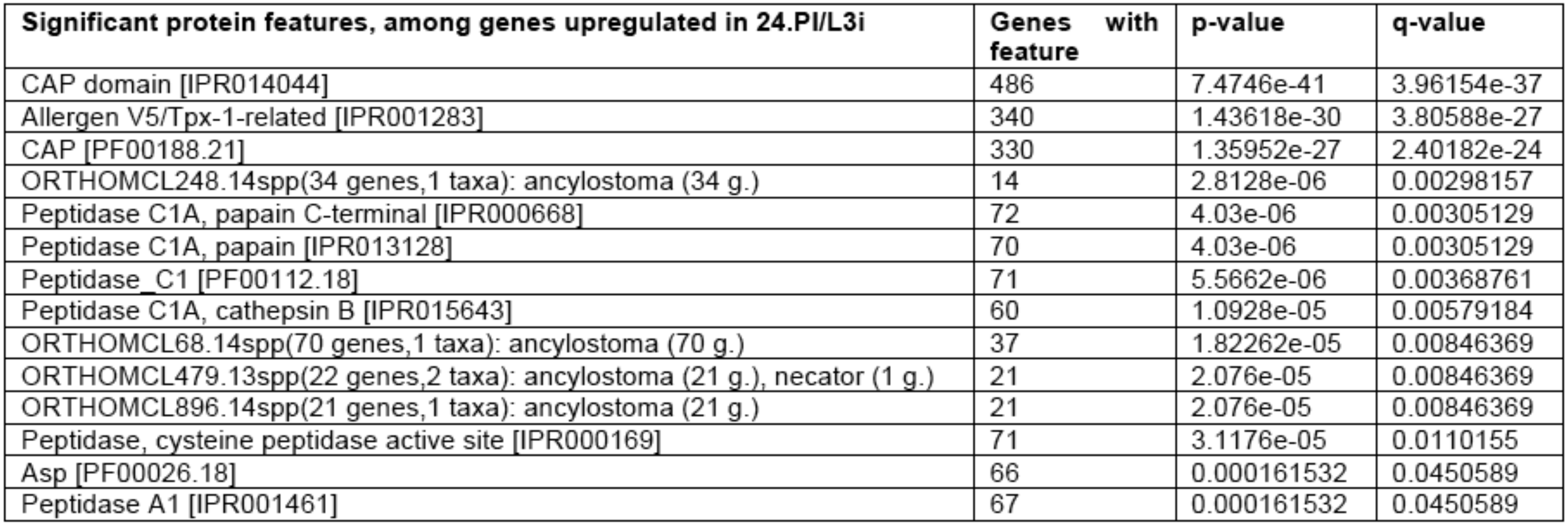
L4 (5.D) to young adult (12.D), upregulated:
protein tyrosine phosphatases and serine/threonine kinases

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

New genes upregulated during early infection

Statistically analyze gene families, not GO terms, for L3i to 24.PI:
i.e., look for protein motifs or orthology groups
overrepresented in genes with high 24.PI/L3i expression ratios.

This mostly gives things that we expect to see:

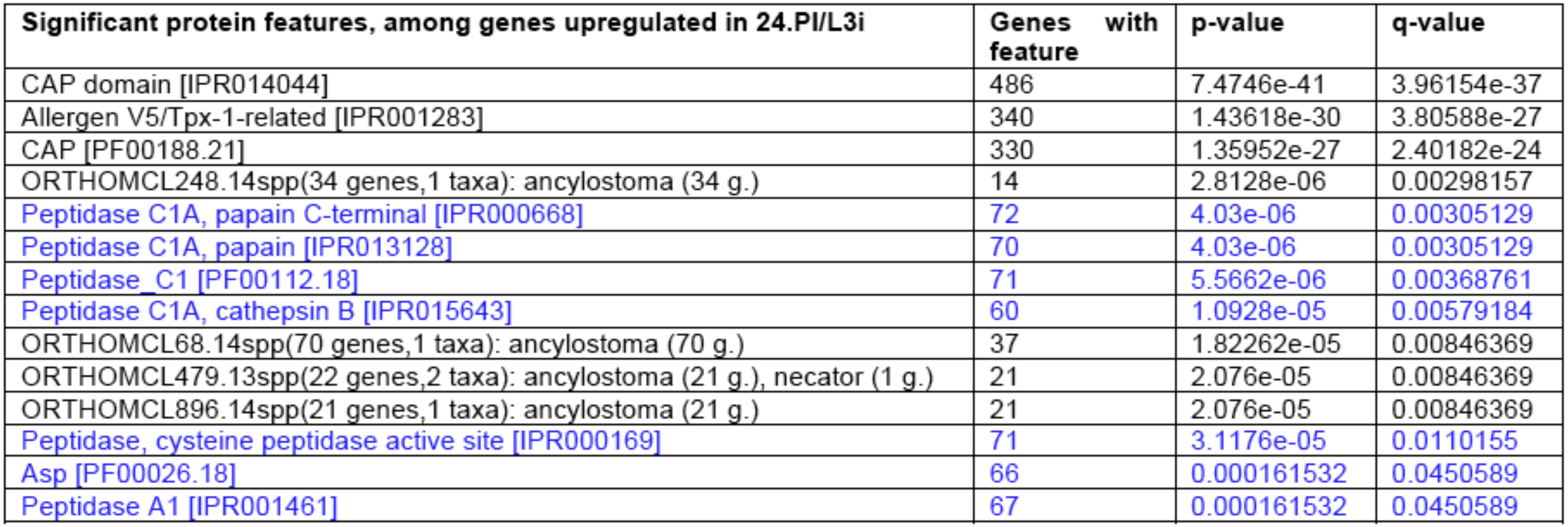


Significant protein features, among genes upregulated in 24.PI/L3i	**Genes with feature**	**p-value**	**q-value**
CAP domain [IPR014044]	486	7.4746e-41	3.96154e-37
Allergen V5/Tpx-1-related [IPR001283]	340	1.43618e-30	3.80588e-27
CAP [PF00188.21]	330	1.35952e-27	2.40182e-24
ORTHOMCL248.14spp(34 genes,1 taxa): ancylostoma (34 g.)	14	2.8128e-06	0.00298157
Peptidase C1A, papain C-terminal [IPR000668]	72	4.03e-06	0.00305129
Peptidase C1A, papain [IPR013128]	70	4.03e-06	0.00305129
Peptidase_C1 [PF00112.18]	71	5.5662e-06	0.00368761
Peptidase C1A, cathepsin B [IPR015643]	60	1.0928e-05	0.00579184
ORTHOMCL68.14spp(70 genes,1 taxa): ancylostoma (70 g.)	37	1.82262e-05	0.00846369
ORTHOMCL479.13spp(22 genes,2 taxa): ancylostoma (21 g.), necator (1 g.)	21	2.076e-05	0.00846369
ORTHOMCL896.14spp(21 genes,1 taxa): ancylostoma (21 g.)	21	2.076e-05	0.00846369
Peptidase, cysteine peptidase active site [IPR000169]	71	3.1176e-05	0.0110155
Asp [PF00026.18]	66	0.000161532	0.0450589
Peptidase A1 [IPR001461]	67	0.000161532	0.0450589

New genes upregulated during early infection

Statistically analyze gene families, not GO terms, for L3i to 24.PI: i.e., look for protein motifs or orthology groups overrepresented in genes with high 24.PI/L3i expression ratios.

E.g., proteases:

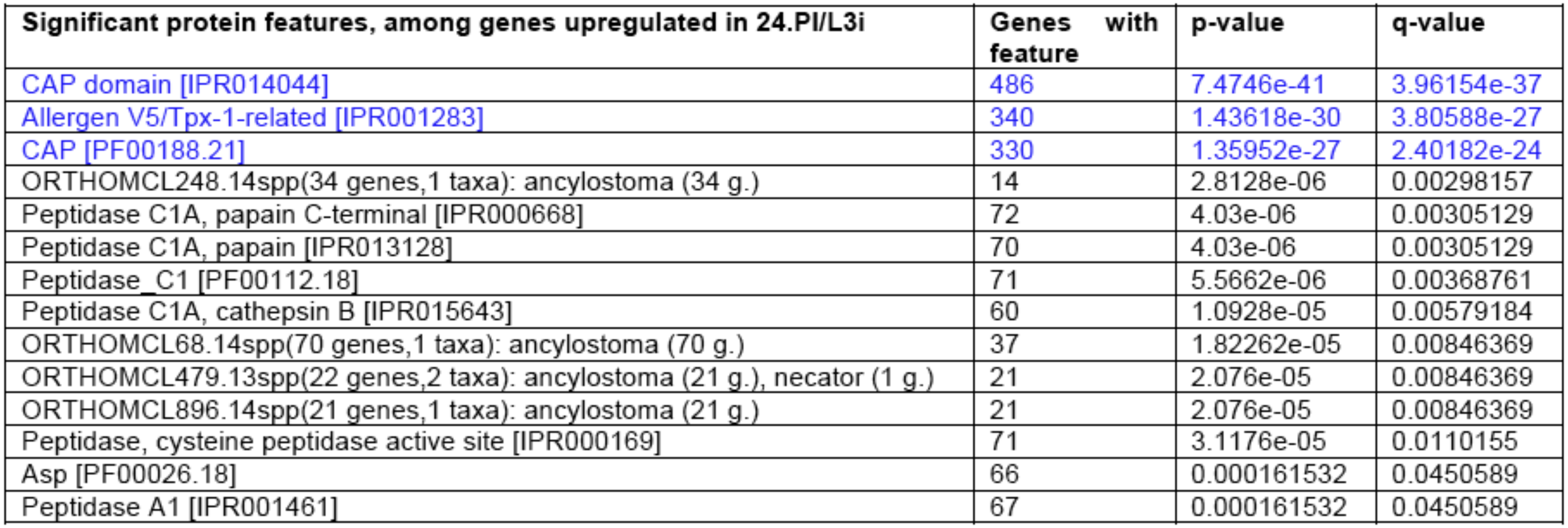


Significant protein features, among genes upregulated in 24.PI/L3i	**Genes with feature**	**p-value**	**q-value**
CAP domain [IPR014044]	486	7.4746e-41	3.96154e-37
Allergen V5/Tpx-1-related [IPR001283]	340	1.43618e-30	3.80588e-27
CAP [PF00188.21]	330	1.35952e-27	2.40182e-24
ORTHOMCL248.14spp(34 genes,1 taxa): ancylostoma (34 g.)	14	2.8128e-06	0.00298157
Peptidase C1A, papain C-terminal [IPR000668]	72	4.03e-06	0.00305129
Peptidase C1A, papain [IPR013128]	70	4.03e-06	0.00305129
Peptidase_C1 [PF00112.18]	71	5.5662e-06	0.00368761
Peptidase C1A, cathepsin B [IPR015643]	60	1.0928e-05	0.00579184
ORTHOMCL68.14spp(70 genes,1 taxa): ancylostoma (70 g.)	37	1.82262e-05	0.00846369
ORTHOMCL479.13spp(22 genes,2 taxa): ancylostoma (21 g.), necator (1 g.)	21	2.076e-05	0.00846369
ORTHOMCL896.14spp(21 genes,1 taxa): ancylostoma (21 g.)	21	2.076e-05	0.00846369
Peptidase, cysteine peptidase active site [IPR000169]	71	3.1176e-05	0.0110155
Asp [PF00026.18]	66	0.000161532	0.0450589
Peptidase A1 [IPR001461]	67	0.000161532	0.0450589

New genes upregulated during early infection

Statistically analyze gene families, not GO terms, for L3i to 24.PI:
i.e., look for protein motifs or orthology groups
overrepresented in genes with high 24.PI/L3i expression ratios.

And, very prominently, Activation-associated Secreted Proteins (ASPs):

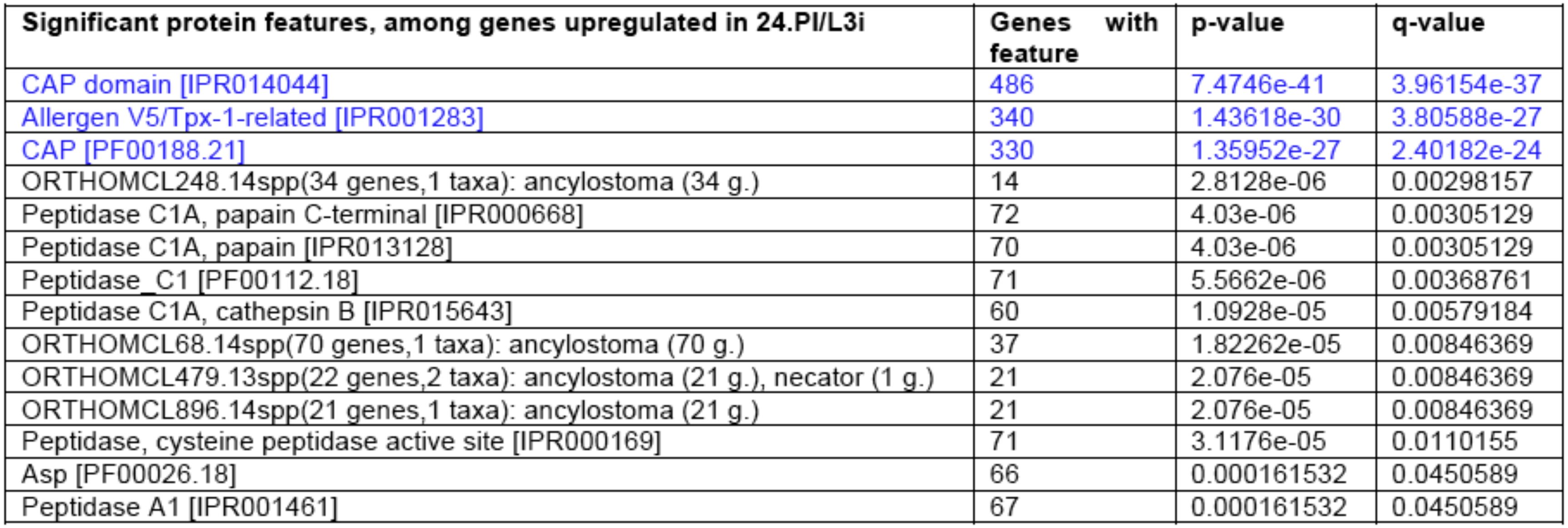


Significant protein features, among genes upregulated in 24.PI/L3i	Genes with feature	p-value	q-value
CAP domain [IPR014044]	486	7.4746e-41	3.96154e-37
Allergen V5/Tpx-1-related [IPR001283]	340	1.43618e-30	3.80588e-27
CAP [PF00188.21]	330	1.35952e-27	2.40182e-24
ORTHOMCL248.14spp(34 genes,1 taxa): ancylostoma (34 g.)	14	2.8128e-06	0.00298157
Peptidase C1A, papain C-terminal [IPR000668]	72	4.03e-06	0.00305129
Peptidase C1A, papain [IPR013128]	70	4.03e-06	0.00305129
Peptidase_C1 [PF00112.18]	71	5.5662e-06	0.00368761
Peptidase C1A, cathepsin B [IPR015643]	60	1.0928e-05	0.00579184
ORTHOMCL68.14spp(70 genes,1 taxa): ancylostoma (70 g.)	37	1.82262e-05	0.00846369
ORTHOMCL479.13spp(22 genes,2 taxa): ancylostoma (21 g.), necator (1 g.)	21	2.076e-05	0.00846369
ORTHOMCL896.14spp(21 genes,1 taxa): ancylostoma (21 g.)	21	2.076e-05	0.00846369
Peptidase, cysteine peptidase active site [IPR000169]	71	3.1176e-05	0.0110155
Asp [PF00026.18]	66	0.000161532	0.0450589
Peptidase A1 [IPR001461]	67	0.000161532	0.0450589

New genes upregulated during early infection

Statistically analyze gene families, not GO terms, for L3i to 24.PI:
i.e., look for protein motifs or orthology groups
overrepresented in genes with high 24.PI/L3i expression ratios.

And, very prominently, Activation-associated Secreted Proteins (ASPs):



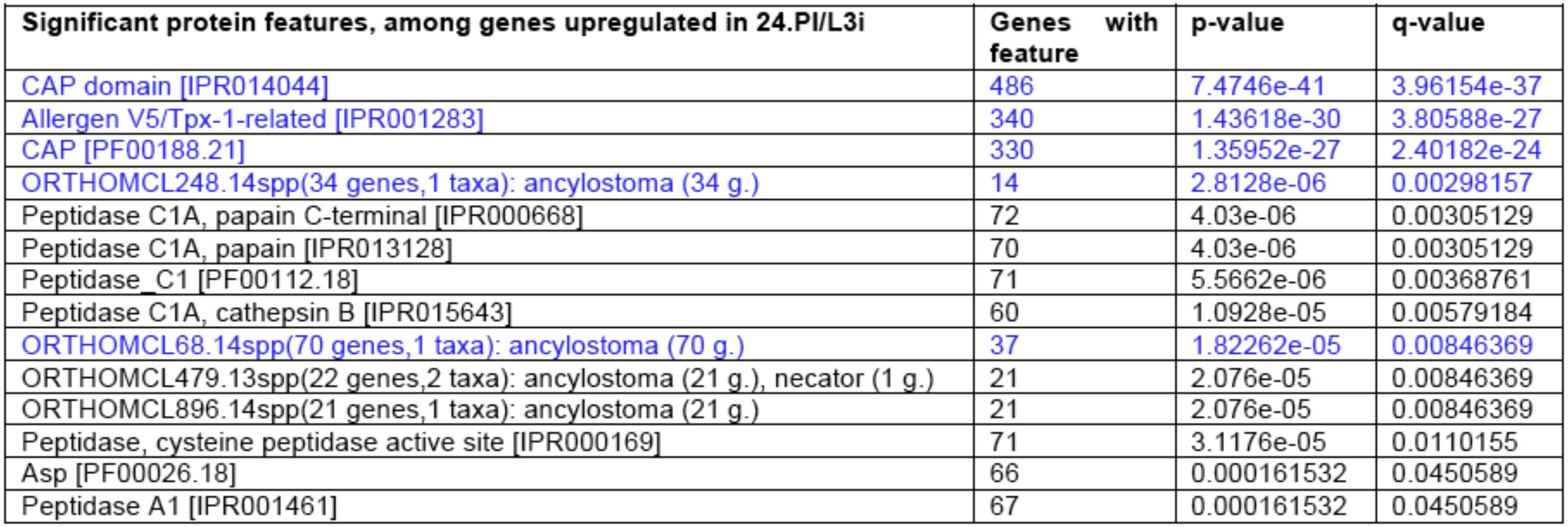
Significant protein features, among genes upregulated in 24.PI/L3i	Genes with feature	p-value	q-value
CAP domain [IPR014044]	486	7.4746e-41	3.96154e-37
Allergen V5/Tpx-1-related [IPR001283]	340	1.43618e-30	3.80588e-27
CAP [PF00188.21]	330	1.35952e-27	2.40182e-24
ORTHOMCL248.14spp(34 genes,1 taxa): ancylostoma (34 g.)	14	2.8128e-06	0.00298157
Peptidase C1A, papain C-terminal [IPR000668]	72	4.03e-06	0.00305129
Peptidase C1A, papain [IPR013128]	70	4.03e-06	0.00305129
Peptidase_C1 [PF00112.18]	71	5.5662e-06	0.00368761
Peptidase C1A, cathepsin B [IPR015643]	60	1.0928e-05	0.00579184
ORTHOMCL68.14spp(70 genes,1 taxa): ancylostoma (70 g.)	37	1.82262e-05	0.00846369
ORTHOMCL479.13spp(22 genes,2 taxa): ancylostoma (21 g.), necator (1 g.)	21	2.076e-05	0.00846369
ORTHOMCL896.14spp(21 genes,1 taxa): ancylostoma (21 g.)	21	2.076e-05	0.00846369
Peptidase, cysteine peptidase active site [IPR000169]	71	3.1176e-05	0.0110155
Asp [PF00026.18]	66	0.000161532	0.0450589
Peptidase A1 [IPR001461]	67	0.000161532	0.0450589

ASPs are a major component of secretions into hosts by hookworms, *H. contortus*, etc.
Many known through cDNA cloning and genomics: ~130 in *N. americanus*.
Bewildering variety of synonyms: CAP, Allergen V5/Tpx-1 related, SCP/TAPS, VAL...

New genes upregulated during early infection

Statistically analyze gene families, not GO terms, for L3i to 24.PI: i.e., look for protein motifs or orthology groups overrepresented in genes with high 24.PI/L3i expression ratios.

ASPs are also incognito members of some orthology groups:

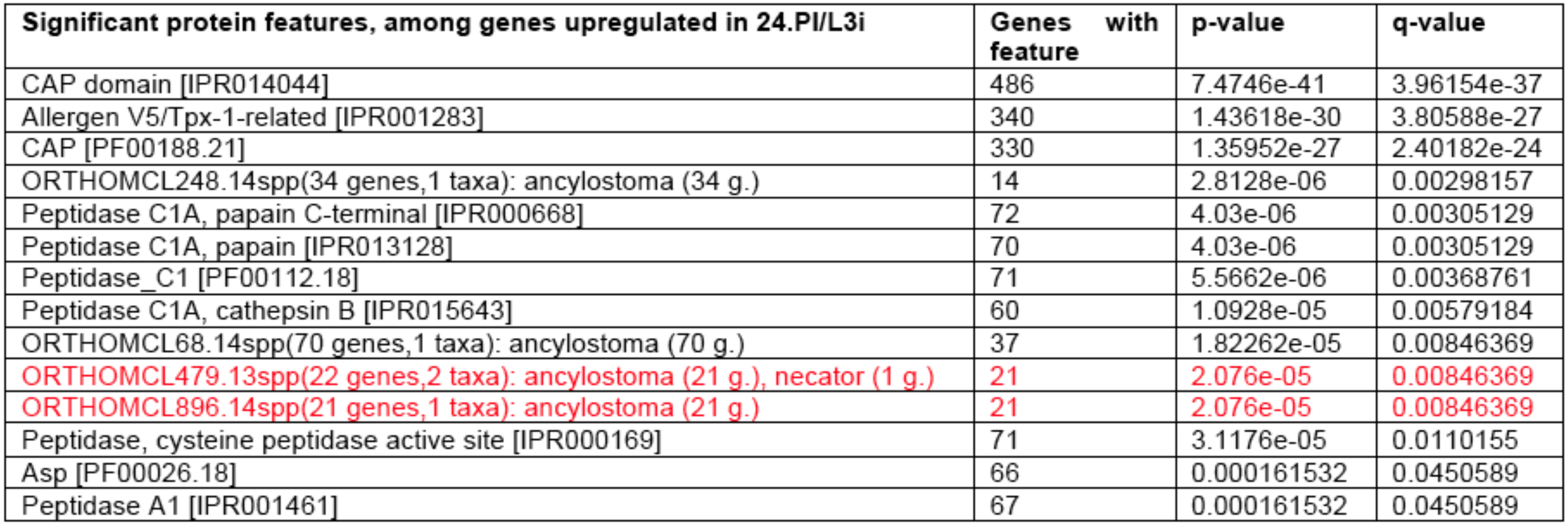


Significant protein features, among genes upregulated in 24.PI/L3i	**Genes with feature**	**p-value**	**q-value**
CAP domain [IPR014044]	486	7.4746e-41	3.96154e-37
Allergen V5/Tpx-1-related [IPR001283]	340	1.43618e-30	3.80588e-27
CAP [PF00188.21]	330	1.35952e-27	2.40182e-24
ORTHOMCL248.14spp(34 genes,1 taxa): ancylostoma (34 g.)	14	2.8128e-06	0.00298157
Peptidase C1A, papain C-terminal [IPR000668]	72	4.03e-06	0.00305129
Peptidase C1A, papain [IPR013128]	70	4.03e-06	0.00305129
Peptidase_C1 [PF00112.18]	71	5.5662e-06	0.00368761
Peptidase C1A, cathepsin B [IPR015643]	60	1.0928e-05	0.00579184
ORTHOMCL68.14spp(70 genes,1 taxa): ancylostoma (70 g.)	37	1.82262e-05	0.00846369
ORTHOMCL479.13spp(22 genes,2 taxa): ancylostoma (21 g.), necator (1 g.)	21	2.076e-05	0.00846369
ORTHOMCL896.14spp(21 genes,1 taxa): ancylostoma (21 g.)	21	2.076e-05	0.00846369
Peptidase, cysteine peptidase active site [IPR000169]	71	3.1176e-05	0.0110155
Asp [PF00026.18]	66	0.000161532	0.0450589
Peptidase A1 [IPR001461]	67	0.000161532	0.0450589

New genes upregulated during early infection

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

However, two (equivalent) orthology groups encode unfamiliar proteins:



Significant protein features, among genes upregulated in 24.PI/L3i	Genes with feature	p-value	q-value
CAP domain [IPR014044]	486	7.4746e-41	3.96154e-37
Allergen V5/Tpx-1-related [IPR001283]	340	1.43618e-30	3.80588e-27
CAP [PF00188.21]	330	1.35952e-27	2.40182e-24
ORTHOMCL248.14spp(34 genes,1 taxa): ancylostoma (34 g.)	14	2.8128e-06	0.00298157
Peptidase C1A, papain C-terminal [IPR000668]	72	4.03e-06	0.00305129
Peptidase C1A, papain [IPR013128]	70	4.03e-06	0.00305129
Peptidase_C1 [PF00112.18]	71	5.5662e-06	0.00368761
Peptidase C1A, cathepsin B [IPR015643]	60	1.0928e-05	0.00579184
ORTHOMCL68.14spp(70 genes,1 taxa): ancylostoma (70 g.)	37	1.82262e-05	0.00846369
ORTHOMCL479.13spp(22 genes,2 taxa): ancylostoma (21 g.), necator (1 g.)	21	2.076e-05	0.00846369
ORTHOMCL896.14spp(21 genes,1 taxa): ancylostoma (21 g.)	21	2.076e-05	0.00846369
Peptidase, cysteine peptidase active site [IPR000169]	71	3.1176e-05	0.0110155
Asp [PF00026.18]	66	0.000161532	0.0450589
Peptidase A1 [IPR001461]	67	0.000161532	0.0450589

One new class of upregulated genes

Proteins in ORTHOMCL896.14spp are
generally secreted, and ~200 a.a. long;
but otherwise non-descript
(neither PFAM nor InterPro classes them as ASPs, etc.).

One new class of upregulated genes

Proteins in ORTHOMCL896.14spp are
generally secreted, and ~200 a.a. long;
but otherwise non-descript
(neither PFAM nor InterPro classes them as ASPs, etc.).

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

One new class of upregulated genes

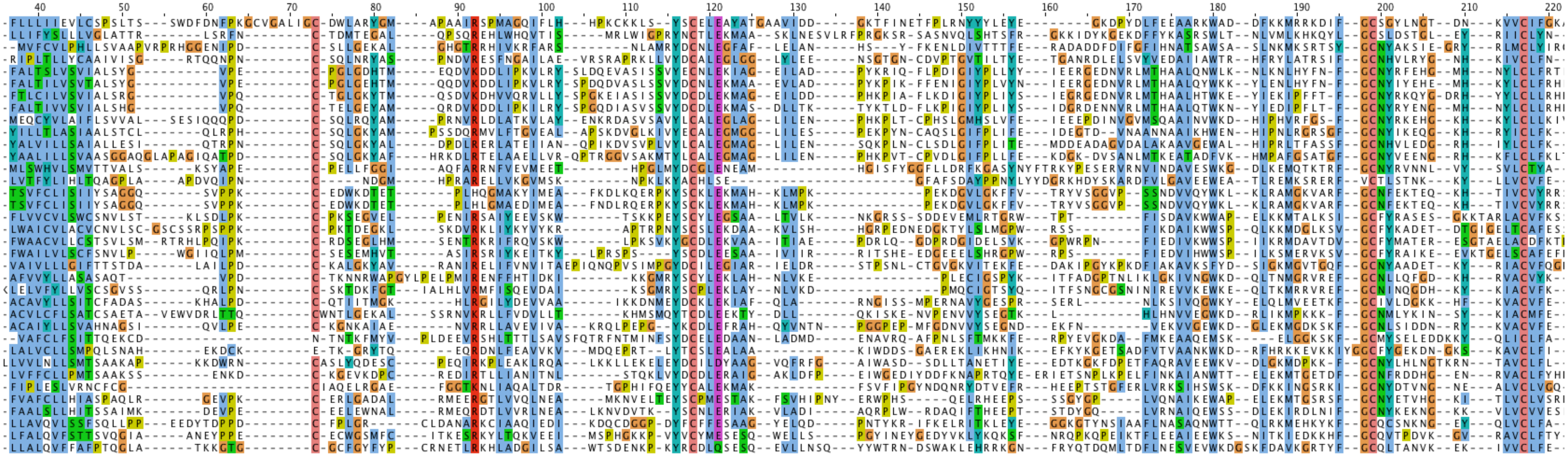
Proteins in ORTHOMCL896.14spp are
generally secreted, and ~200 a.a. long;
but otherwise non-descript
(neither PFAM nor InterPro classes them as ASPs, etc.).

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

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

ASPRs are a diverse subfamily

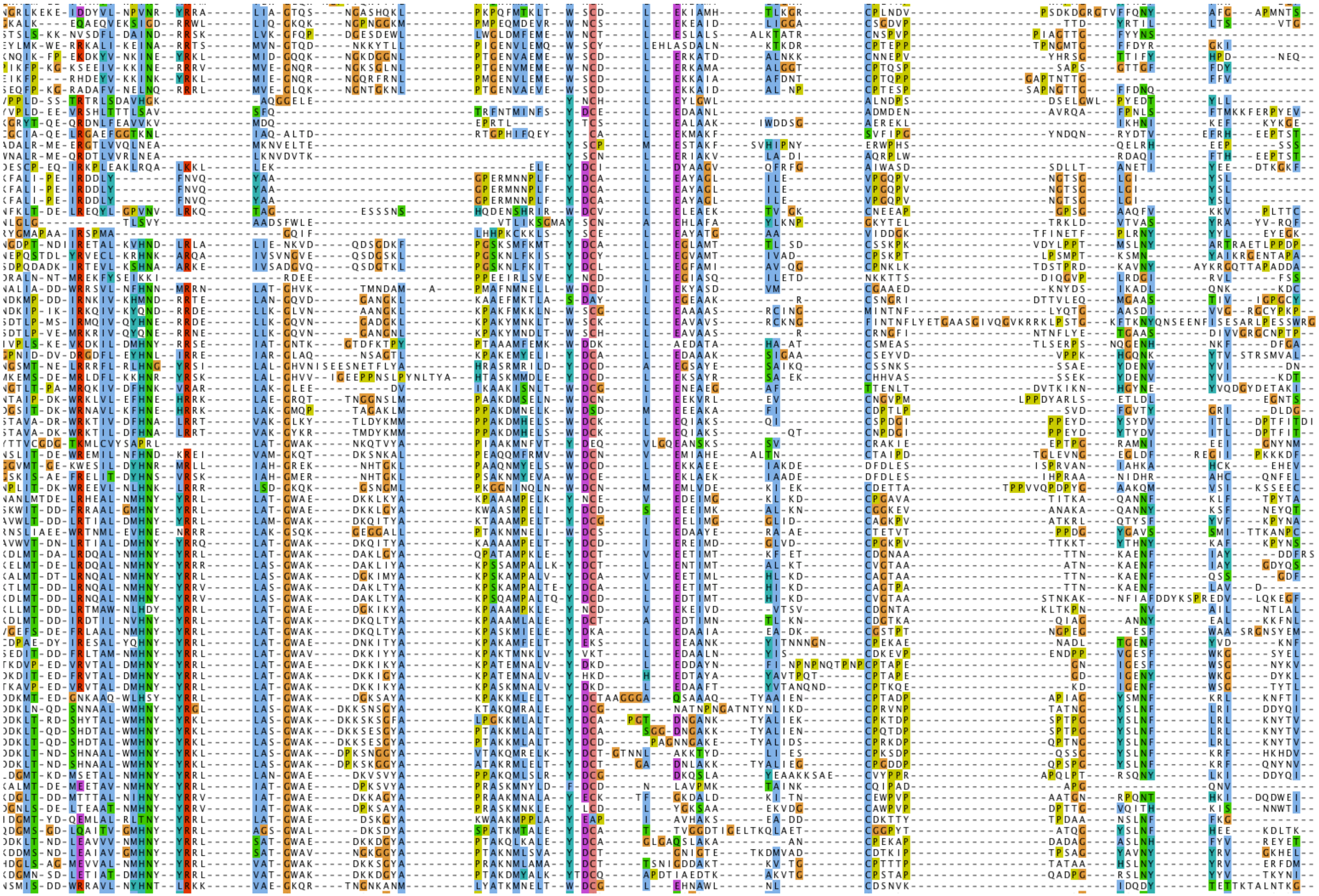
By aligning with MUSCLE, then editing the alignment with JalView, a set of readily alignable ASPRs emerges:



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

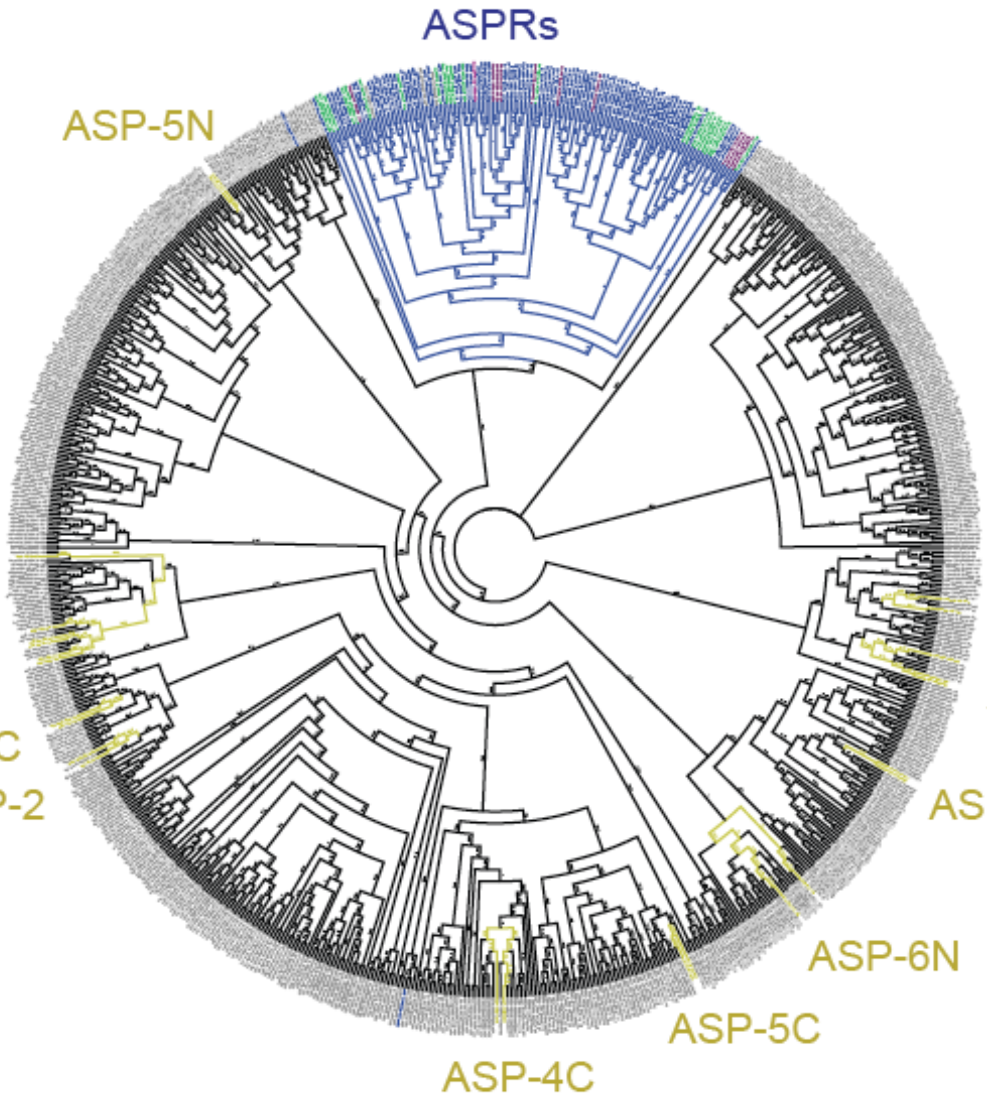
ASPs and ASPRs are a superfamily

These ASPRs can be further aligned with ASP homologs:



ASPs and ASPRs are a superfamily

New ASPs in *A. ceylanicum* and *N. americanus*



ASPRs in *A. ceylanicum*, *N. americanus*, *O. dentatum*, and *H. polygyrus (bakeri)*

Known ASPs in *A. ceylanicum* and *N. americanus* (e.g., ASP-1)

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,

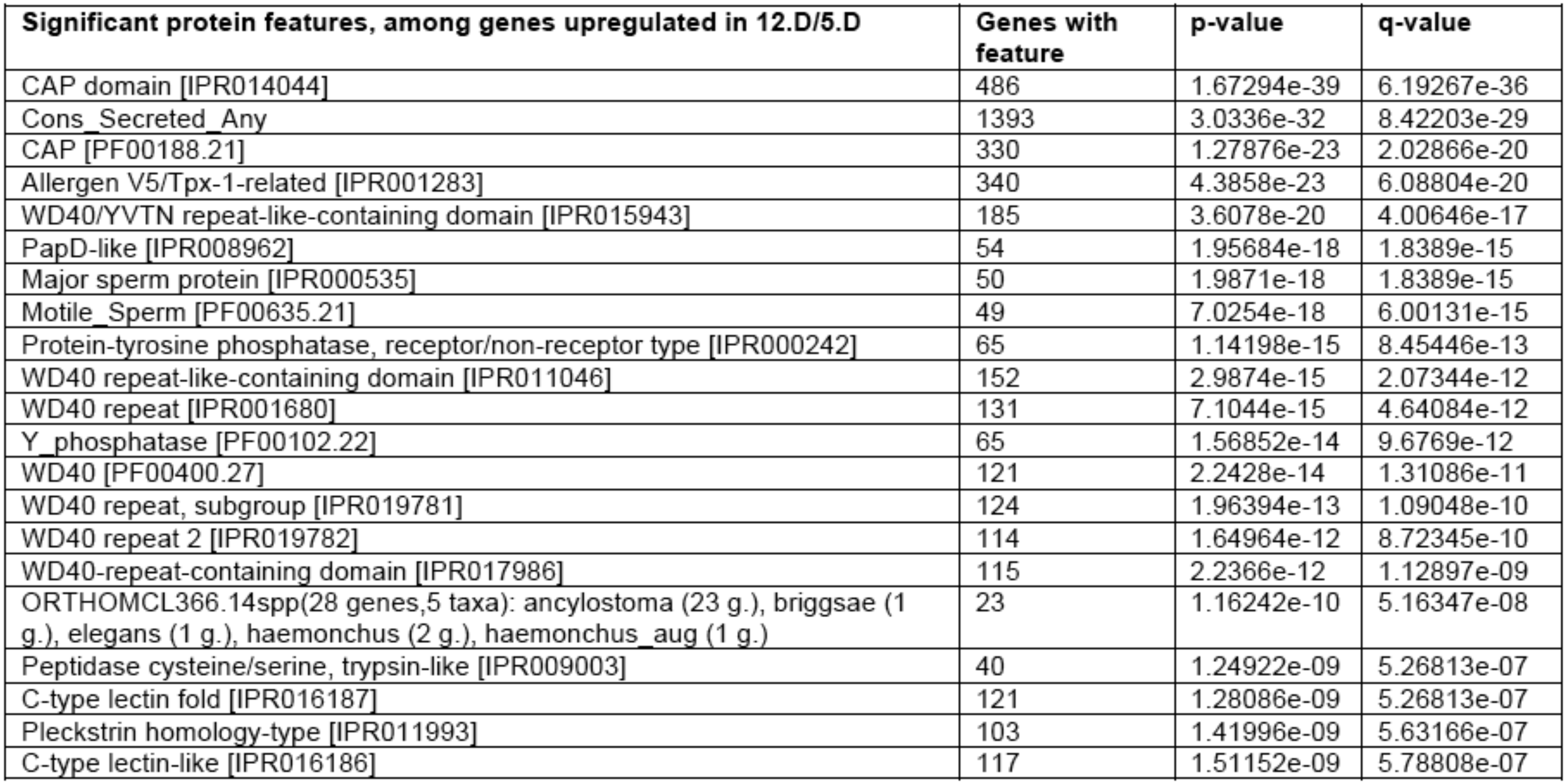
are all consistent with the hypothesis that
ASPRs comprise a new component of hookworm infection.

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

But ASPs are known. Is there anything *new*?

Statistically analyze gene families upregulated
from late L4 larvae (5.D) to young adults (12.D).

Again, most of the upregulated groups are familiar gene families:

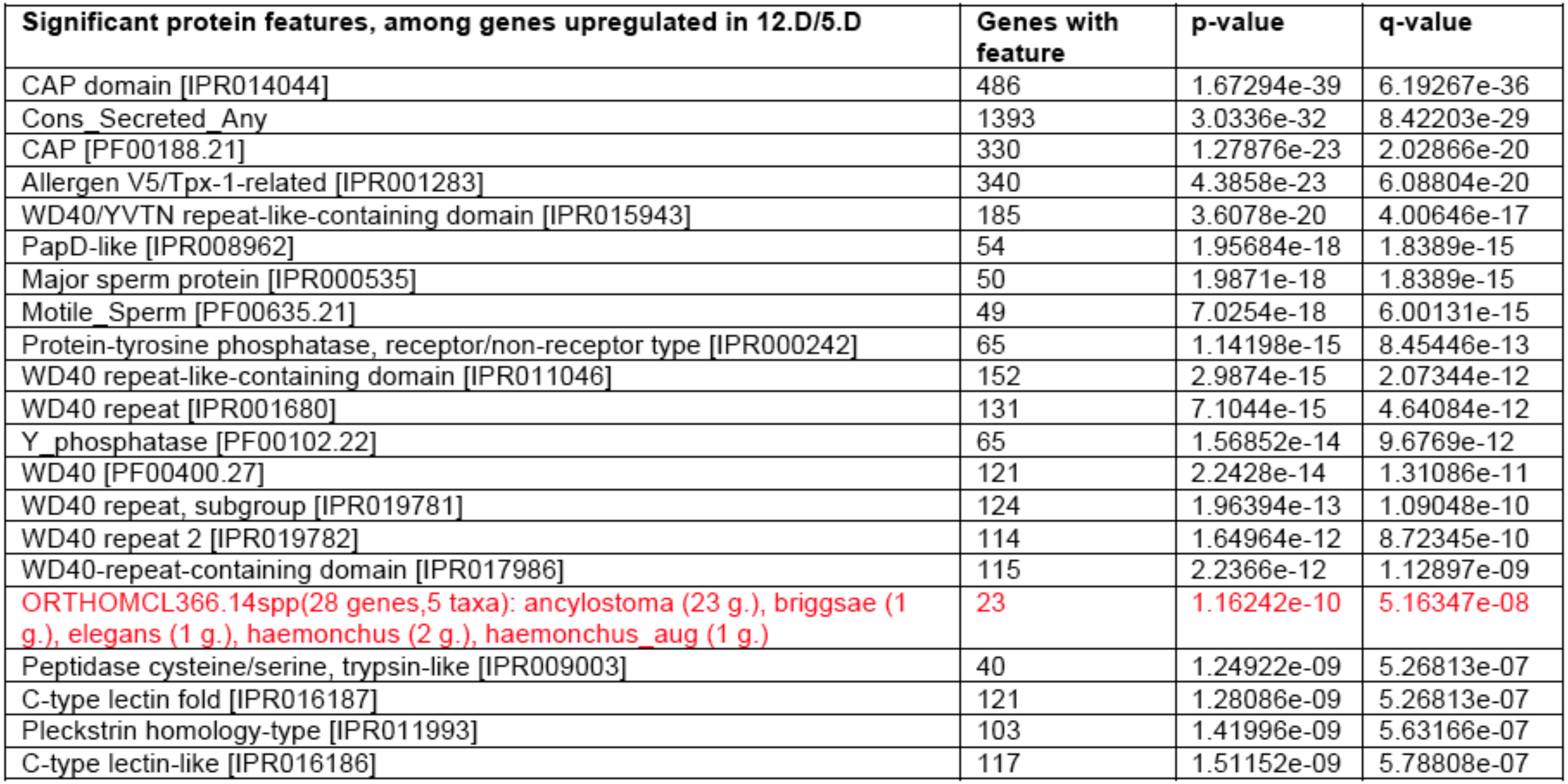


Significant protein features, among genes upregulated in 12.D/5.D	Genes with feature	p-value	q-value
CAP domain [IPR014044]	486	1.67294e-39	6.19267e-36
Cons_Secreted_Any	1393	3.0336e-32	8.42203e-29
CAP [PF00188.21]	330	1.27876e-23	2.02866e-20
Allergen V5/Tpx-1-related [IPR001283]	340	4.3858e-23	6.08804e-20
WD40/YVTN repeat-like-containing domain [IPR015943]	185	3.6078e-20	4.00646e-17
PapD-like [IPR008962]	54	1.95684e-18	1.8389e-15
Major sperm protein [IPR000535]	50	1.9871e-18	1.8389e-15
Motile_Sperm [PF00635.21]	49	7.0254e-18	6.00131e-15
Protein-tyrosine phosphatase, receptor/non-receptor type [IPR000242]	65	1.14198e-15	8.45446e-13
WD40 repeat-like-containing domain [IPR011046]	152	2.9874e-15	2.07344e-12
WD40 repeat [IPR001680]	131	7.1044e-15	4.64084e-12
Y_phosphatase [PF00102.22]	65	1.56852e-14	9.6769e-12
WD40 [PF00400.27]	121	2.2428e-14	1.31086e-11
WD40 repeat, subgroup [IPR019781]	124	1.96394e-13	1.09048e-10
WD40 repeat 2 [IPR019782]	114	1.64964e-12	8.72345e-10
WD40-repeat-containing domain [IPR017986]	115	2.2366e-12	1.12897e-09
ORTHOMCL366.14spp(28 genes,5 taxa): ancylostoma (23 g.), briggsae (1 g.), elegans (1 g.), haemonchus (2 g.), haemonchus_aug (1 g.)	23	1.16242e-10	5.16347e-08
Peptidase cysteine/serine, trypsin-like [IPR009003]	40	1.24922e-09	5.26813e-07
C-type lectin fold [IPR016187]	121	1.28086e-09	5.26813e-07
Pleckstrin homology-type [IPR011993]	103	1.41996e-09	5.63166e-07
C-type lectin-like [IPR016186]	117	1.51152e-09	5.78808e-07

But ASPs are known. Is there anything *new*?

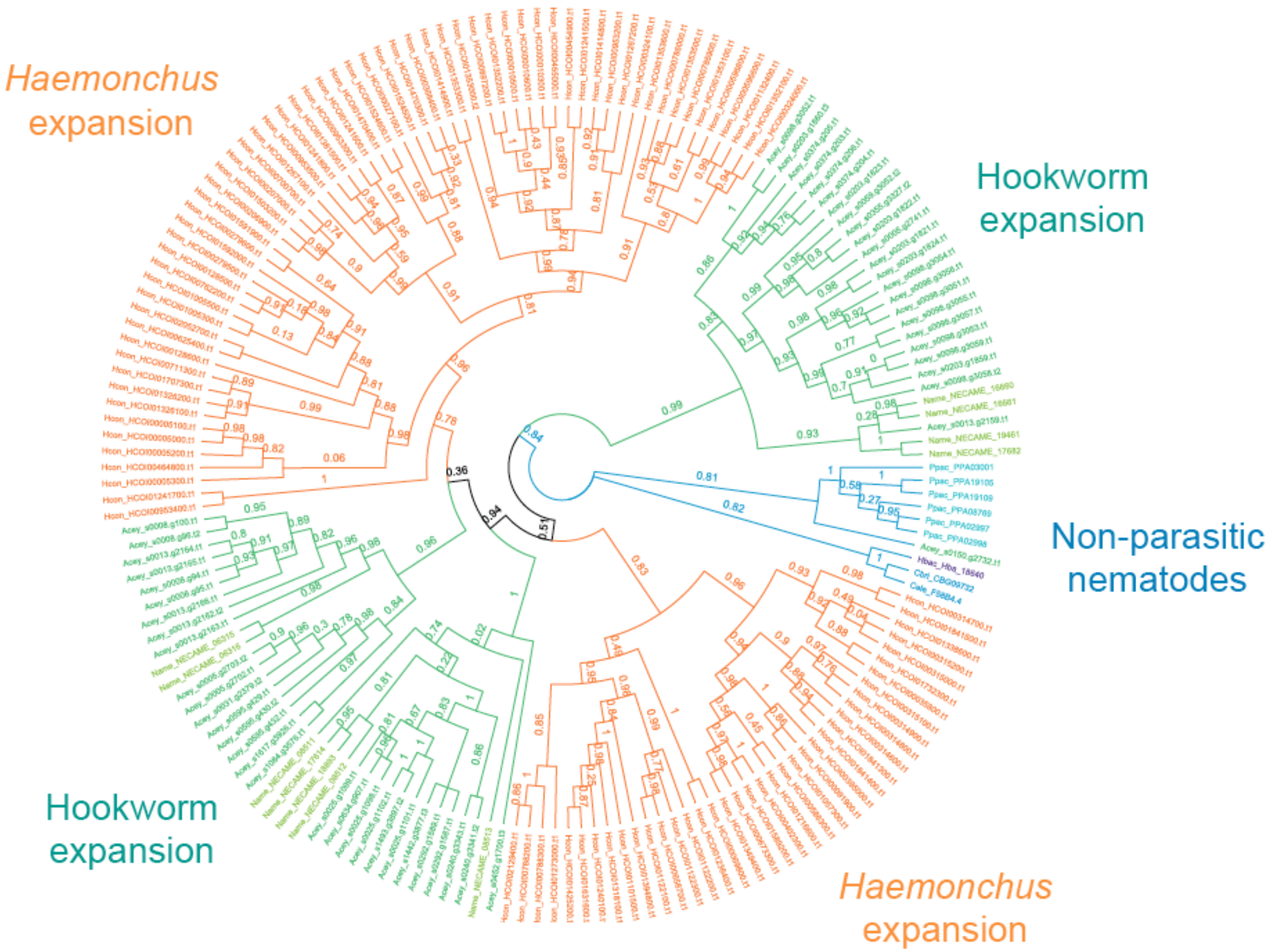
Statistically analyze gene families upregulated
from late L4 larvae (5.D) to young adults (12.D).

Yet, here, also, is a novel upregulated gene family:



Significant protein features, among genes upregulated in 12.D/5.D	**Genes with feature**	**p-value**	**q-value**
CAP domain [IPR014044]	486	1.67294e-39	6.19267e-36
Cons_Secreted_Any	1393	3.0336e-32	8.42203e-29
CAP [PF00188.21]	330	1.27876e-23	2.02866e-20
Allergen V5/Tpx-1-related [IPR001283]	340	4.3858e-23	6.08804e-20
WD40/YVTN repeat-like-containing domain [IPR015943]	185	3.6078e-20	4.00646e-17
PapD-like [IPR008962]	54	1.95684e-18	1.8389e-15
Major sperm protein [IPR000535]	50	1.9871e-18	1.8389e-15
Motile_Sperm [PF00635.21]	49	7.0254e-18	6.00131e-15
Protein-tyrosine phosphatase, receptor/non-receptor type [IPR000242]	65	1.14198e-15	8.45446e-13
WD40 repeat-like-containing domain [IPR011046]	152	2.9874e-15	2.07344e-12
WD40 repeat [IPR001680]	131	7.1044e-15	4.64084e-12
Y_phosphatase [PF00102.22]	65	1.56852e-14	9.6769e-12
WD40 [PF00400.27]	121	2.2428e-14	1.31086e-11
WD40 repeat, subgroup [IPR019781]	124	1.96394e-13	1.09048e-10
WD40 repeat 2 [IPR019782]	114	1.64964e-12	8.72345e-10
WD40-repeat-containing domain [IPR017986]	115	2.2366e-12	1.12897e-09
ORTHOMCL366.14spp(28 genes,5 taxa): ancylostoma (23 g.), briggsae (1 g.), elegans (1 g.), haemonchus (2 g.), haemonchus_aug (1 g.)	23	1.16242e-10	5.16347e-08
Peptidase cysteine/serine, trypsin-like [IPR009003]	40	1.24922e-09	5.26813e-07
C-type lectin fold [IPR016187]	121	1.28086e-09	5.26813e-07
Pleckstrin homology-type [IPR011993]	103	1.41996e-09	5.63166e-07
C-type lectin-like [IPR016186]	117	1.51152e-09	5.78808e-07

Secreted Clade V Proteins (SCVPs): 5.D to 12.D



Secreted proteins of ~150 residues, with no similarities at all to known domains.
Many *Anyclostoma*, *Necator*, and *Haemonchus* genes; few non-parasite ones.

Profuse gene families encoding secreted proteins

Generally speaking, there has been no "parasitism gene" found by comparing different parasitic nematode genomes.

This is unsurprising, given the multiple origins of parasitism in the nematode phylum.

Refs.: Rogozin (2014) Genet. Res. Int. *2014*, 516508; Zarowiecki and Berriman (2014), Parasitology *8*, 1-13.

Profuse gene families encoding secreted proteins

Generally speaking, there has been no "parasitism gene" found by comparing different parasitic nematode genomes.

This is unsurprising, given the multiple origins of parasitism in the nematode phylum.

However, could there instead be "parasitism expanded multigene families"?

Gene duplication tends to result in subfunctionalization (when it does not just lead to gene loss).

Subfunctionalization allows both higher protein divergence and disparate transcriptional regulation.

Both of these could be useful, by creating decoy antigens that induced the host immune system uselessly.

Refs.: Rogozin (2014) Genet. Res. Int. *2014*, 516508; Zarowiecki and Berriman (2014), Parasitology *8*, 1-13.

Overview

1. Why do we want a hookworm genome?
2. Generating a genome and transcriptome
3. Characterizing the genome
4. Characterizing the transcriptome
5. Predicting drug and vaccine targets
6. Some thoughts on 'descriptive genomics'

Predicted drug targets

Targets were required to be potentially "druggable",
present in multiple parasites but absent from human and mouse,
and required for normal *C. elegans*:

Protein class	**Acey genes**	**Key Cel genes**	**Drug data**
4-coumarate:coenzyme A ligase, class I	10	*acs-10*	n/a
Ammonium/urea transporter	5	*amt-2*	n/a
Cofactor-independent phosphoglycerate mutase	1	*ipgm-1*	Limited druggability
Fumarate reductase	1	F48E8.3	n/a
Glutamate-gated chloride channel	10	*avr-14*, *avr-15*, *glc-2*	*avr-14* observed
Glutamate synthase	1	W07E11.1	n/a
Glutamine-fructose 6-phosphate aminotransferase	3	*gfat-1*, *gfat-2*	n/a
Isocitrate lyase / Malate_synthase	2	*icl-1*	n/a
KH-domain RNA binding	5	*asd-2*, *gld-1*, K07H8.9	n/a
Malate/L-lactate dehydrogenase, YlbC-type	4	F36A2.3	n/a
NADH:flavin oxidoreductase, Oye2/3-type	14	F17A9.4	n/a
Nematode prostaglandin F synthase	3	C35D10.6	n/a
O-acetylserine sulfhydrylase	2	*cysl-1*	n/a
Secreted lipase	6	*lips-8*, *lips-9*	n/a
Trehalose-6-phosphate synthase	5	*gob-1*, *tps-1*, *tps-2*	*gob-1* predicted

avr-14 has been validated experimentally;
gob-1 has been predicted by Berriman and coworkers for *H. contortus*;
ipgm-1 has provoked intense interest (but been difficult to drug).

Refs.: Crowther et al. (2014), PLoS Negl. Trop. Dis. *8*, e2628; Laing et al. (2013), Genome Biol. *14*, R88;
Somvanshi et al. (2014), Mol. Biochem. Parasitol. *193*, 1-8.

Existing vaccine candidates

Activation-associated secreted proteins (ASPs):

Are a major component of secretions into hosts by hookworms, *H. contortus*, etc.
Can be elicited in culture by Hookworm Culture Medium (serum).
Many ASPs known through cDNA cloning and genomics: ~130 in *N. americanus*.
ASPs may suppress clotting and immune responses.

ASP-2 worked as hamster vaccine, but caused hives in humans.

Aspartic proteases (APRs):

Participate in a proteolytic cascade that successively digests hemoglobin.

Glutathione S-transferases (GSTs):

Are thought to counteract the toxicity of globin breakdown products in hookworm gut.

APR-1 and GST-1 both work as vaccines in hamsters;
they are in clinical trials as a mixed vaccine in humans.

Refs.: Xiao et al. (2008), Exp. Parasitol. *118*, 32-40; Diemert et al. (2012), J. Allergy. Clin. Immunol. *130*, 169-176;
Hotez et al. (2013), Vaccine *31 Suppl 2*, B227-B232; Tang et al. (2014), Nat. Genet. *46*, 261-269.

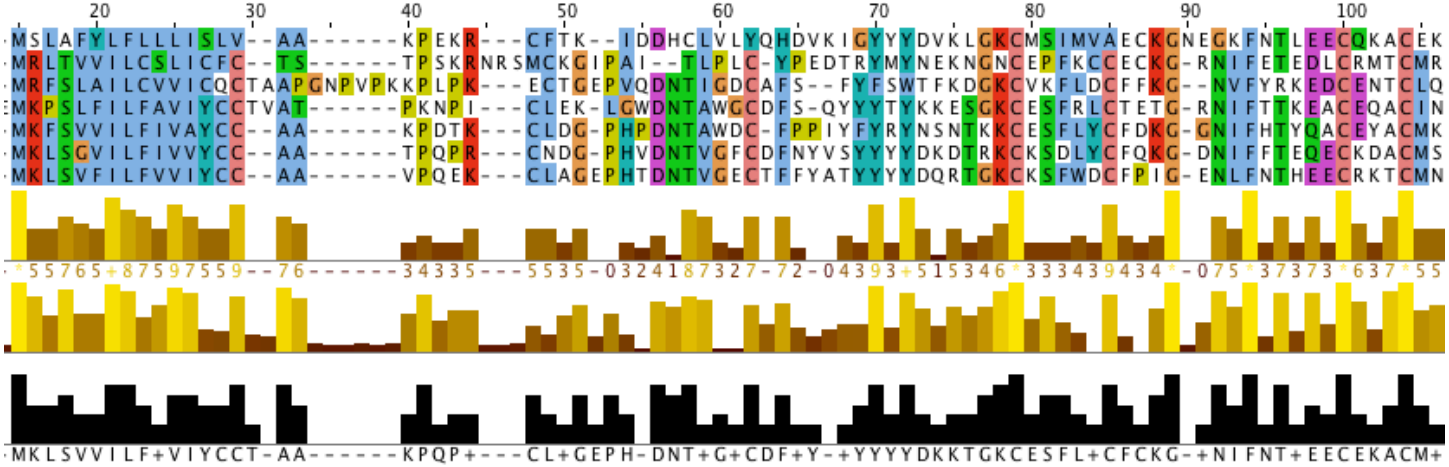
Proteases, and protease inhibitors

Five **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.

Proteases, and protease inhibitors

Five **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.

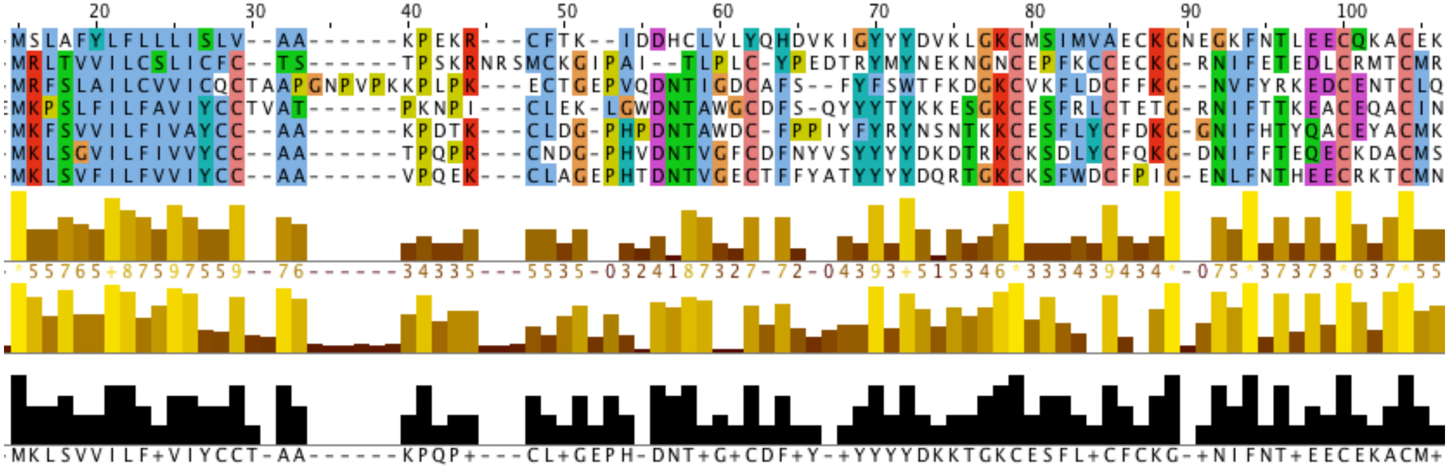
One **small protease inhibitor** (family shown below) is upregulated by 5.D,
has no mammalian homologs,
and does have one *H. contortus* homolog upregulated during infection.


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+

Proteases, and protease inhibitors

Five **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.

One **small protease inhibitor** (family shown below) is upregulated by 5.D,
has no mammalian homologs,
and does have one *H. contortus* homolog upregulated during 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
EMKPSLFILFAVIYCCTVAT-------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---5535-03241873 27-72-04393+515346*333439434*-075*37373*637*55
MKLSVVILF+VIYCCT-AA------KPQP+---CL+GEPH-DNT+G+CDF+Y-+YYYYDKKTGKCESFL+CFCKG-+NIFNT+EECEKACM+

These two gene sets encode ~1% and 0.1% of <u>all</u> adult transcripts.

They are thus likely to encode genes relevant to survival in the host.

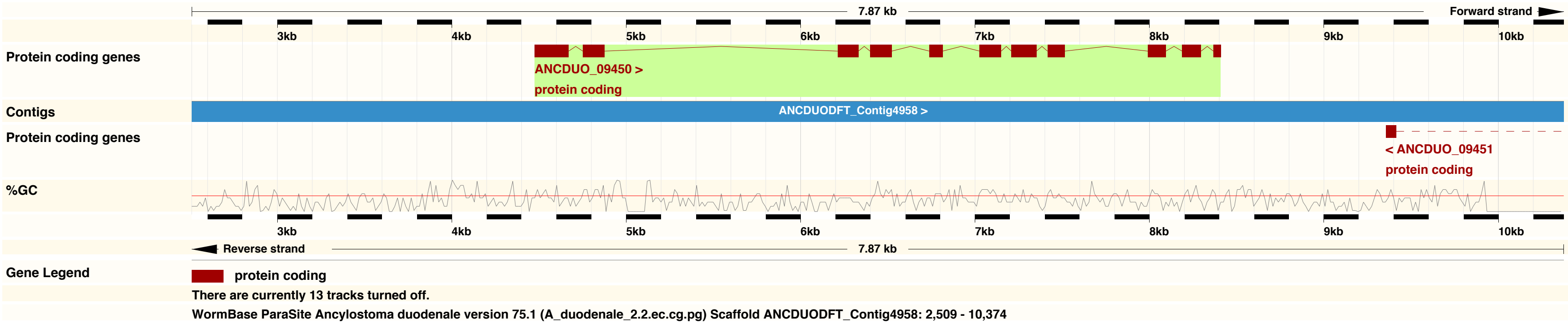
Conclusions:

1. *Ancylostoma ceylanicum* encodes ~31,000 protein-coding genes with detectable expression.

2. Three diverse but conserved gene families are upregulated during succeeding steps of infection. These might be immunological decoys.

3. Other genes are upregulated during infection but are less profuse, and encode functions likely to be required for host survival.
These might be feasible drug or vaccine targets.

61 parasitic nematode genome sequences!
(of which 21 are published, as of Aug. 2015)



There are currently 13 tracks turned off.

WormBase ParaSite Ancylostoma duodenale version 75.1 (A_duodenale_2.2.ec.cg.pg) Scaffold ANCDUODFT_Contig4958: 2,509 - 10,374

Ancylostoma caninum, Ancylostoma ceylanicum, Ancylostoma duodenale,

Acanthocheilonema viteae, Angiostrongylus cantonensis, Angiostrongylus costaricensis, Anisakis simplex, Ascaris lumbricoides, Ascaris suum, Brugia malayi, Brugia pahangi, Brugia timori, Bursaphelenchus xylophilus, Cylicostephanus goldi, Dictyocaulus viviparus, Dirofilaria immitis, Dracunculus medinensis, Elaeophora elaphi, Enterobius vermicularis, Globodera pallida, Gongylonema pulchrum, Haemonchus contortus, Haemonchus placei, Heligmosomoides polygyrus (bakeri), Heterorhabditis bacteriophora, Litomosoides sigmodontis, Loa loa, Meloidogyne floridensis, Meloidogyne hapla, Meloidogyne incognita, Necator americanus, Nippostrongylus brasiliensis, Oesophagostomum dentatum, Onchocerca flexuosa, Onchocerca ochengi, Onchocerca volvulus, Parascaris equorum, Parastrongyloides trichosuri, Rhabditophanes sp. KR3021, Romanomermis culicivorax, Soboliphyme baturini, Steinernema carpocapsae, Steinernema feltiae, Steinernema glaseri, Steinernema monticolum, Steinernema scapterisci, Strongyloides papillosus, Strongyloides ratti, Strongyloides stercoralis, Strongyloides venezuelensis, Strongylus vulgaris, Syphacia muris, Teladorsagia circumcincta, Thelazia callipaeda, Toxocara canis, Trichinella nativa, Trichinella spiralis, Trichuris muris, Trichuris suis, Trichuris trichiura, Wuchereria bancrofti

http://**parasite.wormbase.org**

Overview

1. Why do we want a hookworm genome?

2. Generating a genome and transcriptome

3. Characterizing the genome

4. Characterizing the transcriptome

5. Predicting drug and vaccine targets

6. Some thoughts on 'descriptive genomics'

Begin and end with checks for basic quality

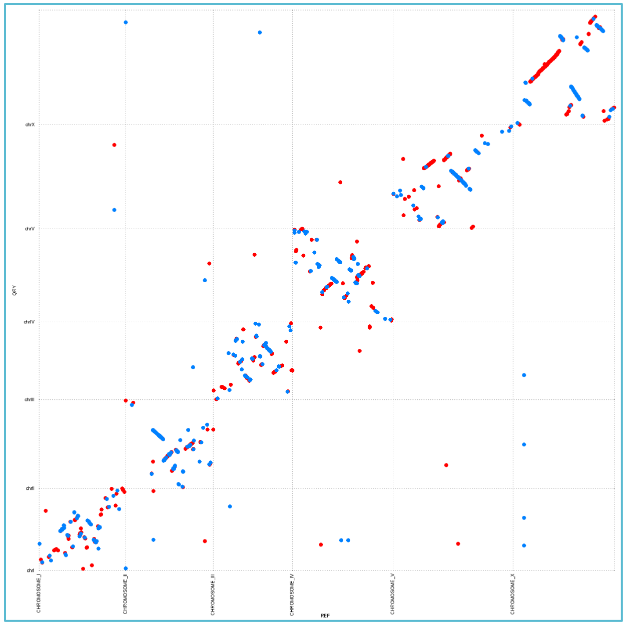
<u>Living organisms sit in a soup of microbes</u>

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

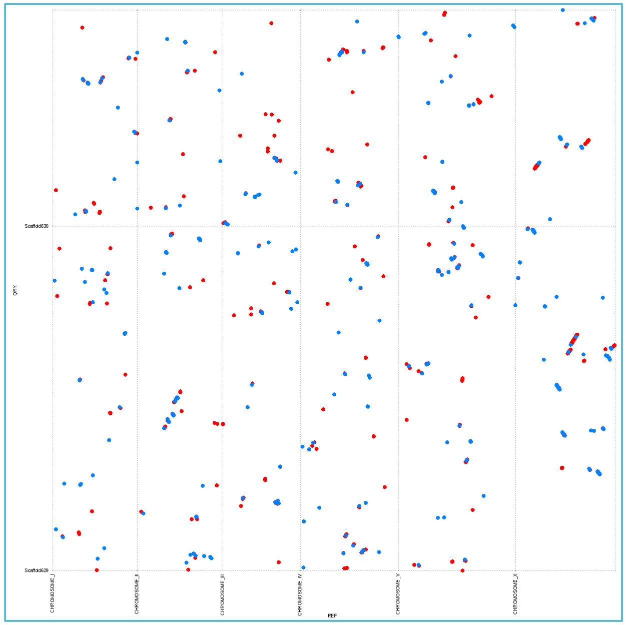
<u>Over-assembly can happen</u>

In case of *C. tropicalis*/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 ~400M 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 ~400M 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 ~400M 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!