Supplementary Information

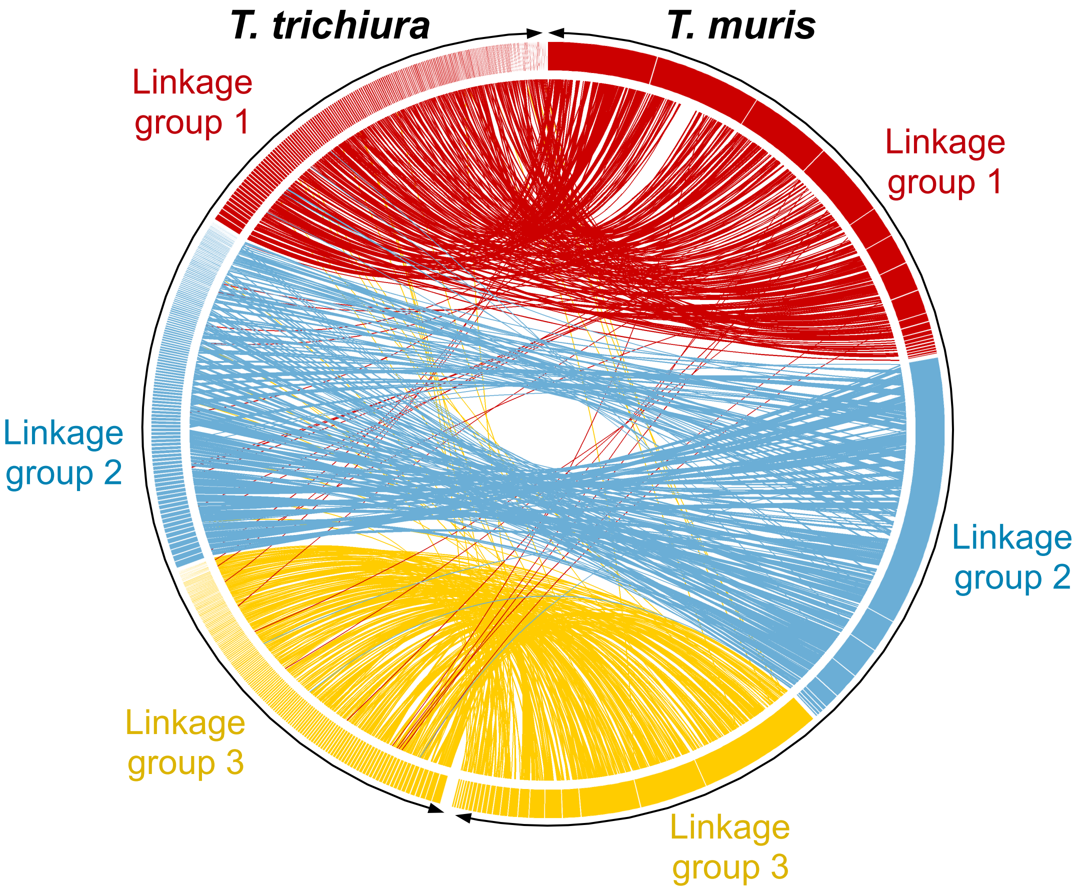
Whipworm genome and dual-species transcriptome analyses provide molecular insights into an intimate host-parasite interaction

Bernardo J. Foth, Isheng J. Tsai, Adam J. Reid, Allison J. Bancroft, Sarah Nichol, Alan Tracey, Nancy Holroyd, James A. Cotton, Eleanor J. Stanley, Magdalena Zarowiecki, Jimmy Z. Liu, Thomas Huckvale, Philip J. Cooper, Richard K. Grencis, Matthew Berriman

Nature Genetics: doi:10.1038/ng.3010

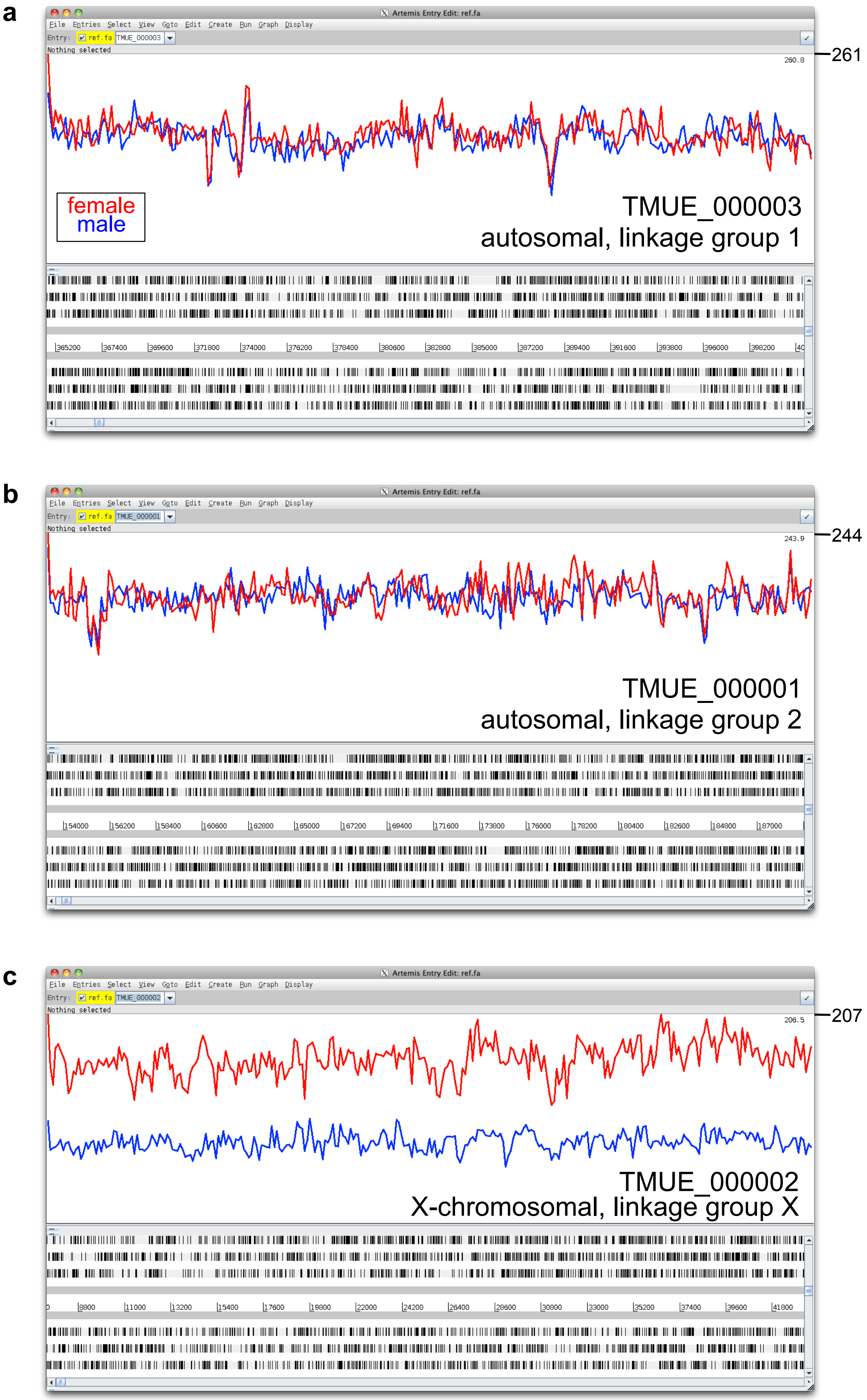
I. SUPPLEMENTARY FIGURES

Supplementary Figure 1

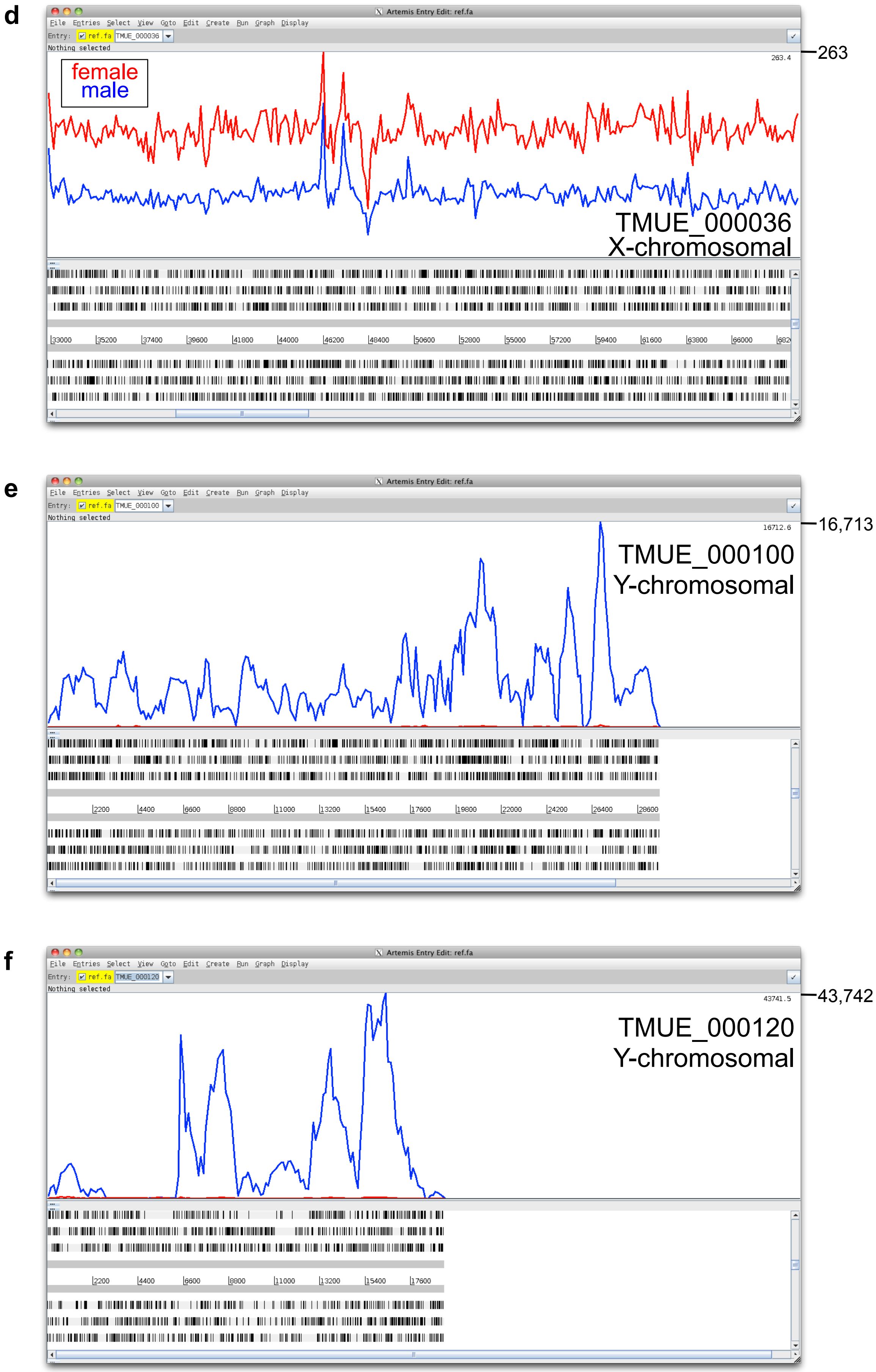


Supplementary Figure 1 High-level synteny between the genomes of *T. trichiura* and *T. muris*. Mapping one-to-one gene orthologs between genome scaffolds of *T. trichiura* (genome assembly v2.1) and *T. muris* (genome assembly v4) followed by clustering of the resulting ortholog pattern identifies three large linkage groups in each genome. Linkage group three (yellow) is putatively identified as sex-specific chromosome(s).

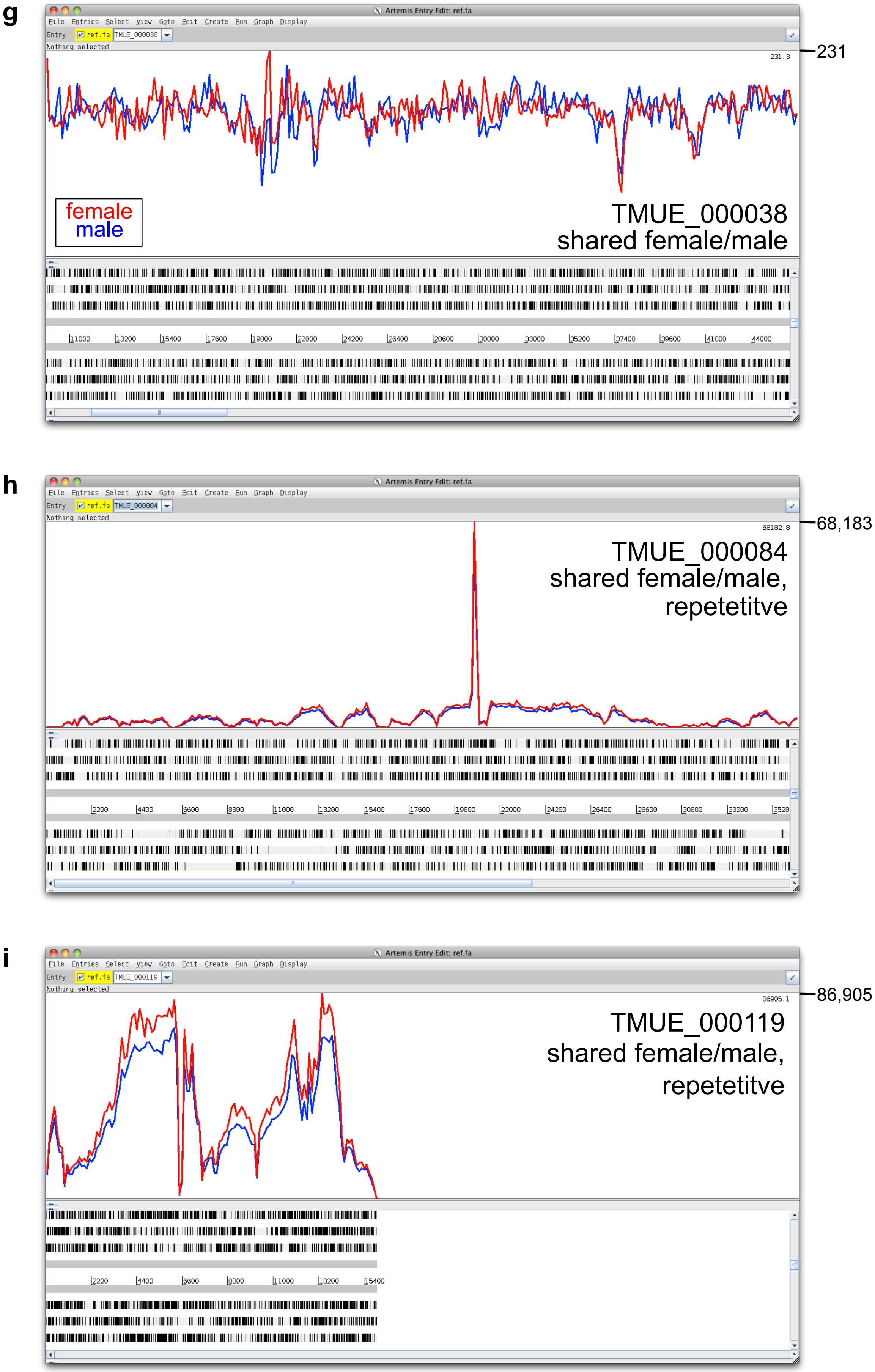
Supplementary Figure 2a-c



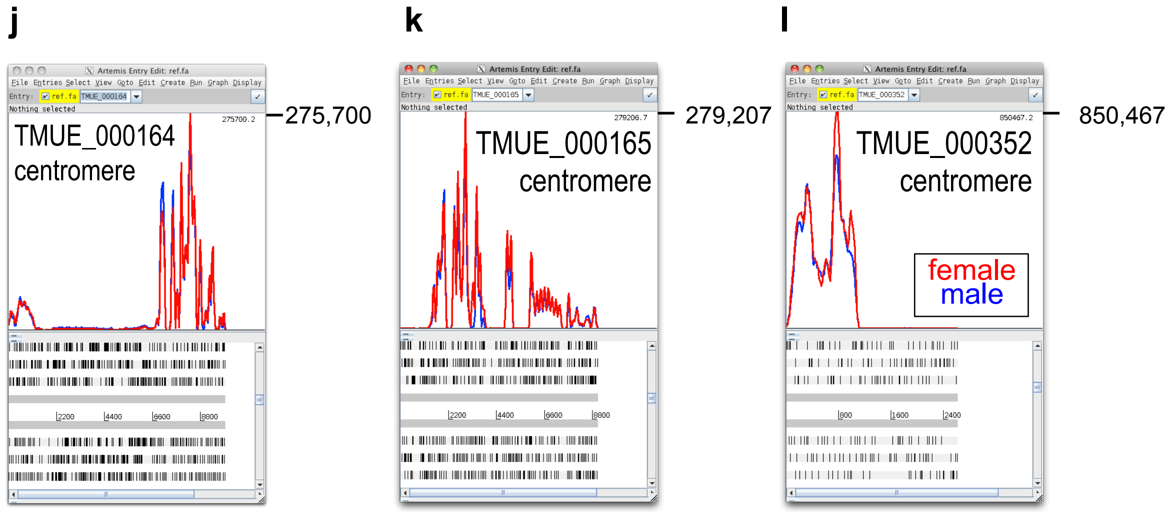
Supplementary Figure 2d-f



Supplementary Figure 2g-i

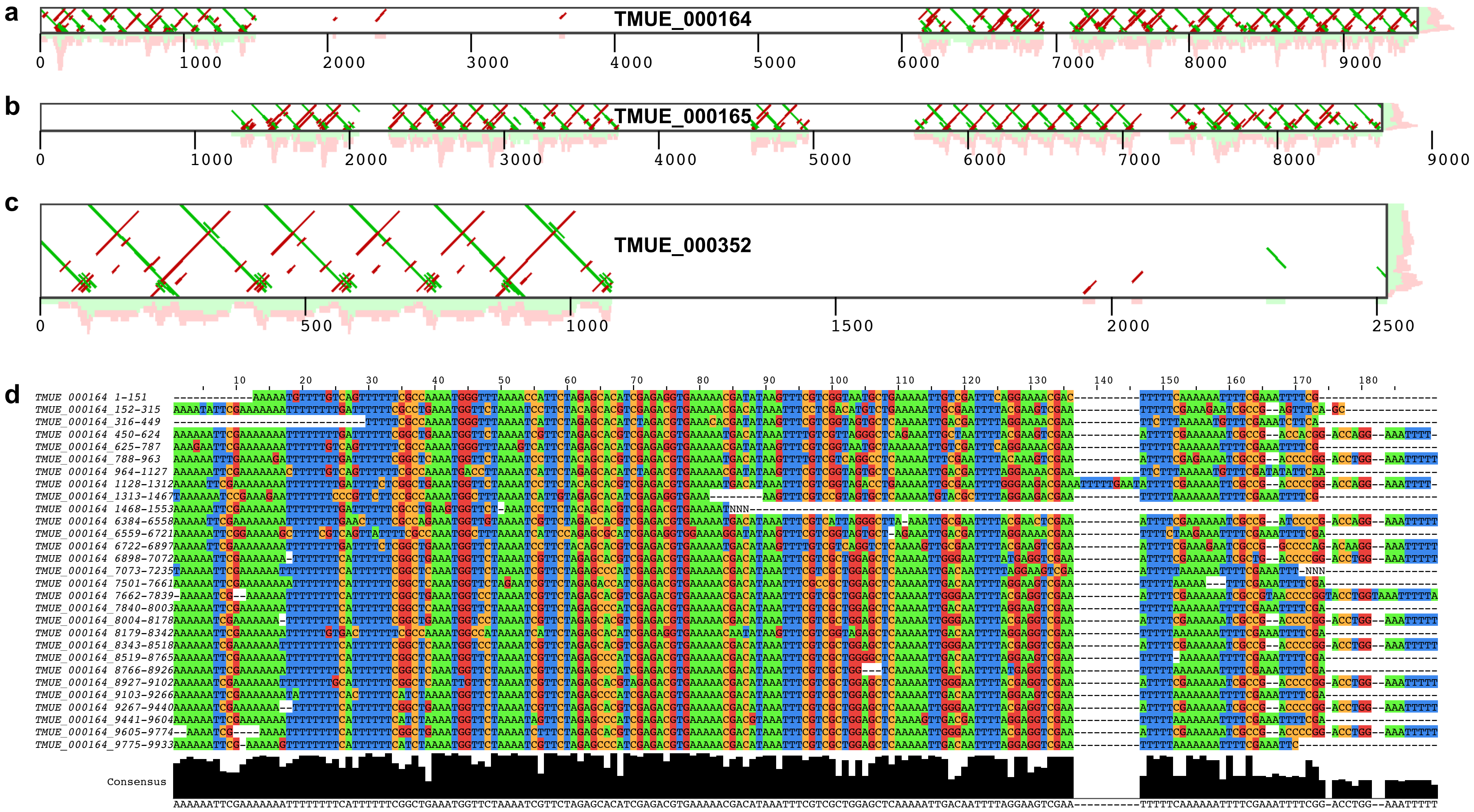


Supplementary Figure 2j-l



Supplementary Figure 2 Illustration of typical high-throughput sequence read coverage over different chromosomal locations. Shown are Artemis screenshots of Illumina data mapped to *T. muris* genome assembly v4. (**a-c**) The three scaffolds have been assigned to three linkage groups based on orthology of genes to *Trichinella spiralis*. Scaffolds TMUE_000003 and TMUE_000001 represent autosomes and show equal read coverage in females (median coverage 152) and males (median coverage 151). In contrast, scaffold TMUE_000002 represents the X chromosome with the coverage in males (blue, median coverage 76) being half that in females (red, median coverage 149). (**d**) Scaffold TMUE_000036 has been inferred to be X-chromosomal based on read coverage. (**e,f**) Scaffolds inferred to represent the Y chromosome based on read coverage, with the very high read coverage indicating highly repetitive sequence content. Taking actual mean read coverage over scaffold TMUE_000120 (on average 58.4x higher than over the disomic autosomes), scaffold length (19.2kb), and the monosomic nature of the Y chromosome (compensated for by multiplication by two) into account suggests a true “uncollapsed” length of this sequence of approximately 2.24 Mb. (**g-i**) Scaffolds that occur in approximately equal proportions in females and males and that have not been assigned to a linkage group based on gene orthology. These scaffolds may represent sequences located on autosomes or shared between the X and the Y chromosomes. Scaffolds TMUE_000084 and TMUE_000119 attract very high read coverage indicating highly repetitive sequence content. Taking actual mean read coverage and scaffold length into account suggests a true “uncollapsed” length of approximately 3.67 Mb for the sequence represented by scaffold TMUE_000119. (**j-l**) Three scaffolds that attract very high read coverage indicating highly repetitive sequence content. Sequence analysis suggests that the highly repetitive sequences of these scaffolds are centromeric (see Supplementary Fig. 3). Taking mean read coverage into account suggests that these three scaffolds together represent approximately 5.33 Mb of “uncollapsed” sequence.

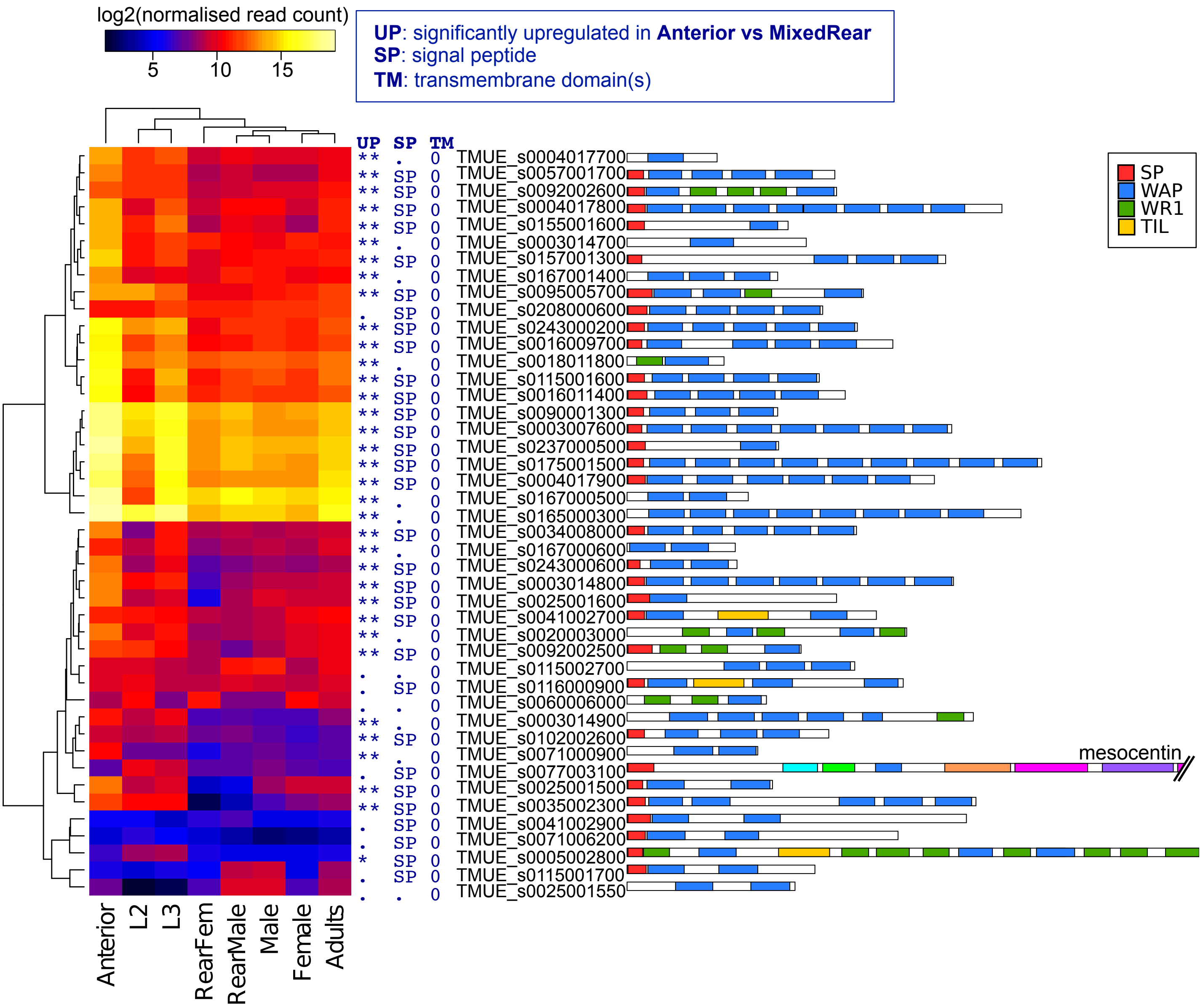
Supplementary Figure 3



Supplementary Figure 3 Sequence analysis putatively identifies centromeric sequences. (**a-c**) Dot-plots comparing the putative centromeric consensus sequence (see bottom of figure panel d) against scaffolds TMUE_000164, TMUE_000165 and TMUE_000352 illustrate the repetitive nature of these sequences. Green indicates forward matches, and red indicates reverse matches. (**d**) Multiple sequence alignment of the putative centromeric monomers present in scaffold TMUE_000164. The repeating monomers are mostly either 164bp or 176bp long, which is comparable to the length of the ~171bp-long monomers of human centromeric alpha-satellite DNA. The monomers found in scaffolds TMUE_000165 and TMUE_000352 are nearly identical to those from scaffold TMUE_000164. See also Supplementary Fig. 2j-l.

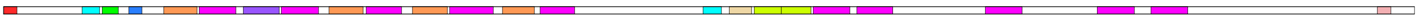
Supplementary Figure 4a

WAP domain (PF00095) containing proteins (most in MEROPS family I17): strongly expressed in the anterior region of adult whipworms and in L3 larvae, and likely secreted (signal peptide present) in most cases.



Note: all WAP domain-containing proteins are annotated as "**WAP domain containing protein, SLPI−like**", except:

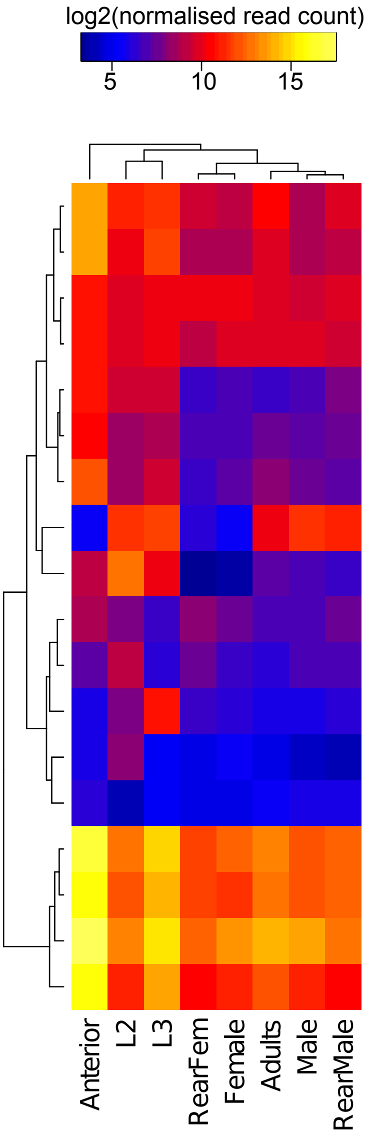
TMUE_s0077003100 **Mesocentin**



Supplementary Figure 4 Transcript-level expression of several gene groups of *T. muris*. Normalised transcript levels of several groups of functionally related genes in *T. muris* comparing parasite tissues, genders, and life cycle stages. The indication of significant transcriptional upregulation in a particular pairwise comparison ("UP") refers to a false discovery rate (FDR) <= 0.01 and FDR > 1E-5 when denoted by one asterisk (*), and to an FDR <= 1E-5 when denoted by two asterisks (**). Supplementary Figs. 4a and 4b are detailed versions of Figs. 3a and 4a in the main text, respectively.

Supplementary Figure 4b

DNase II (PF03265) proteins: tend to be most strongly expressed in the anterior region of adult whipworms and in larvae, and some are likely secreted.

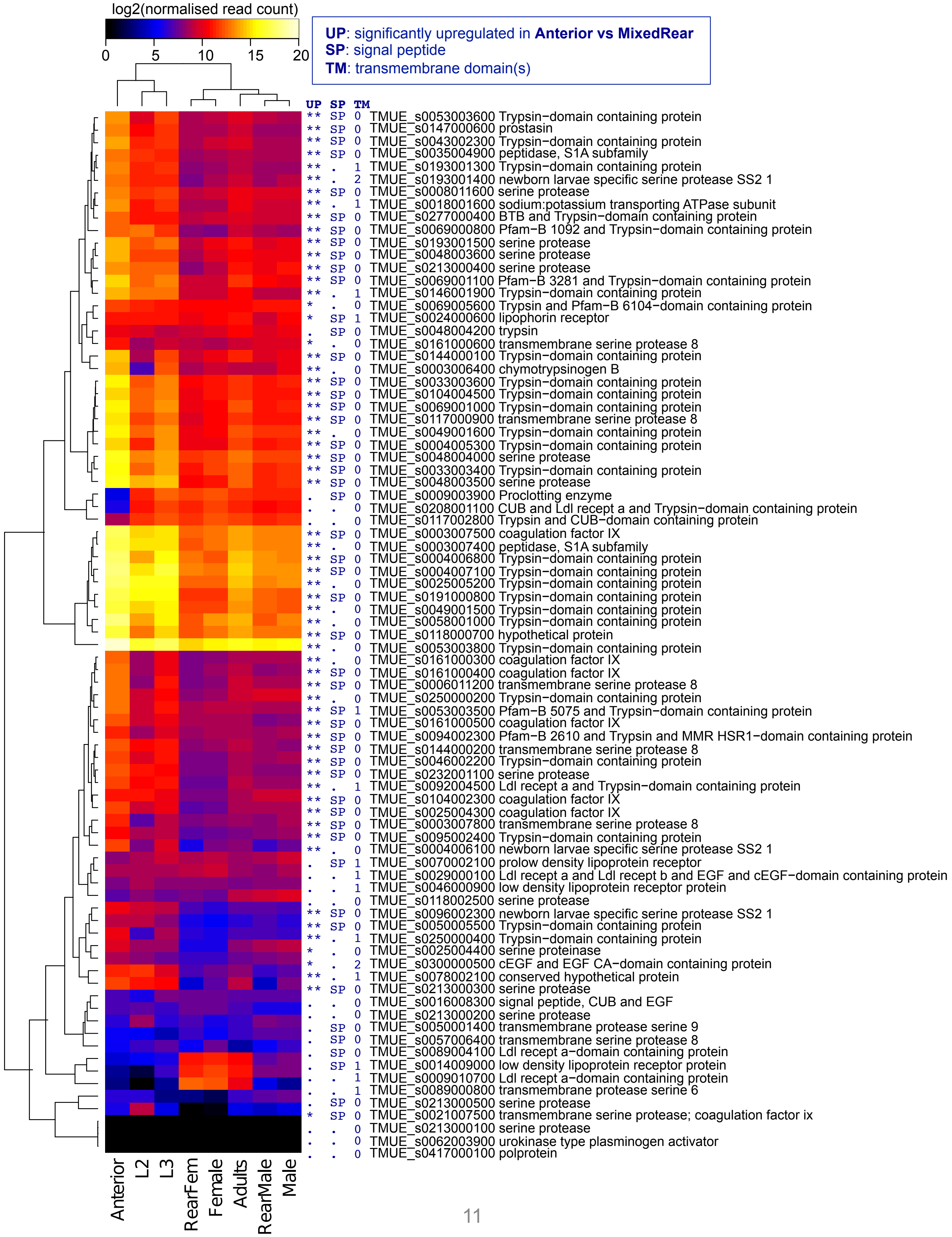


UP: significantly upregulated in **Anterior vs MixedRear**
SP: signal peptide
TM: transmembrane domain(s)

UP	SP	TM	
**	.	0	TMUE_s0241001500 Deoxyribonuclease (DNase) II
**	SP	0	TMUE_s0030004700 Deoxyribonuclease (DNase) II
.	SP	0	TMUE_s0085001500 Deoxyribonuclease (DNase) II
**	SP	0	TMUE_s0083001200 Deoxyribonuclease (DNase) II
**	.	0	TMUE_s0021005000 Deoxyribonuclease (DNase) II
**	.	0	TMUE_s0065000100 Deoxyribonuclease (DNase) II
**	.	0	TMUE_s0011010400 Deoxyribonuclease (DNase) II
.	SP	0	TMUE_s0103003200 Deoxyribonuclease (DNase) II
**	SP	0	TMUE_s0099001700 Deoxyribonuclease (DNase) II
.	.	0	TMUE_s0015002900 Deoxyribonuclease (DNase) II
.	.	0	TMUE_s0065003900 Deoxyribonuclease (DNase) II
.	.	0	TMUE_s0103003300 Deoxyribonuclease (DNase) II
*	.	0	TMUE_s0007004800 Deoxyribonuclease (DNase) II
.	.	0	TMUE_s0044004600 Deoxyribonuclease (DNase) II
**	.	0	TMUE_s0008001500 Deoxyribonuclease (DNase) II
**	SP	0	TMUE_s0030007700 Deoxyribonuclease (DNase) II
**	.	1	TMUE_s0008001600 Deoxyribonuclease (DNase) II
**	SP	0	TMUE_s0044004800 Deoxyribonuclease (DNase) II

Supplementary Figure 4c

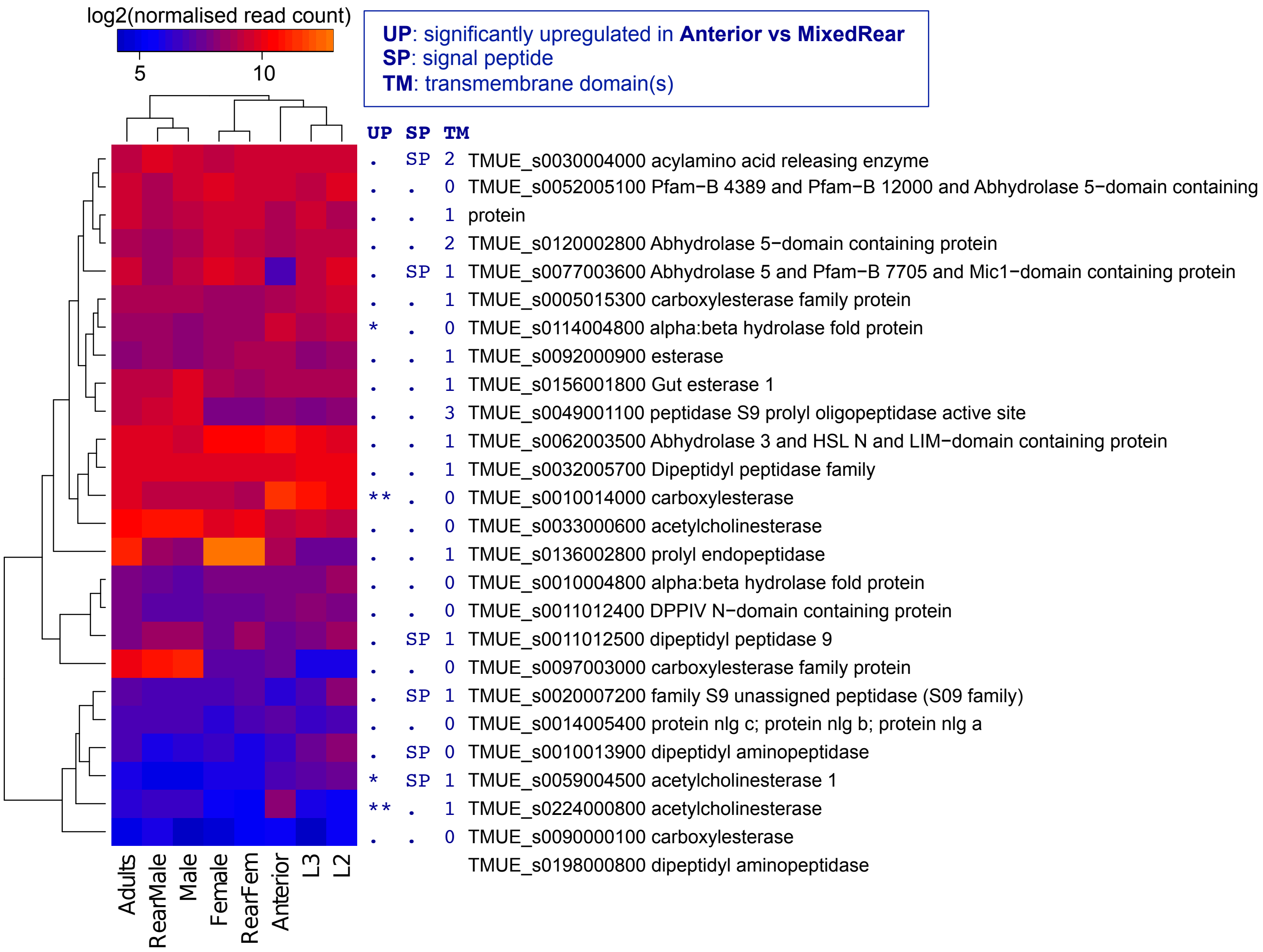
Chymotrypsin-like serine proteases (MEROPS family S1A): tend to be strongly expressed in the anterior region of adult whipworms and in larvae, and most are likely secreted.



UP	SP	TM	
**	SP	0	TMUE_s0053003600 Trypsin−domain containing protein
**	SP	0	TMUE_s0147000600 prostasin
**	SP	0	TMUE_s0043002300 Trypsin−domain containing protein
**	SP	0	TMUE_s0035004900 peptidase, S1A subfamily
**	.	1	TMUE_s0193001300 Trypsin−domain containing protein
**	.	2	TMUE_s0193001400 newborn larvae specific serine protease SS2 1
**	SP	0	TMUE_s0008011600 serine protease
**	.	1	TMUE_s0018001600 sodium:potassium transporting ATPase subunit
**	SP	0	TMUE_s0277000400 BTB and Trypsin−domain containing protein
**	SP	0	TMUE_s0069000800 Pfam−B 1092 and Trypsin−domain containing protein
**	SP	0	TMUE_s0193001500 serine protease
**	SP	0	TMUE_s0048003600 serine protease
**	SP	0	TMUE_s0213000400 serine protease
**	SP	0	TMUE_s0069001100 Pfam−B 3281 and Trypsin−domain containing protein
**	.	1	TMUE_s0146001900 Trypsin−domain containing protein
*	.	0	TMUE_s0069005600 Trypsin and Pfam−B 6104−domain containing protein
*	SP	1	TMUE_s0024000600 lipophorin receptor
.	SP	0	TMUE_s0048004200 trypsin
*	.	0	TMUE_s0161000600 transmembrane serine protease 8
**	SP	0	TMUE_s0144000100 Trypsin−domain containing protein
**	.	0	TMUE_s0003006400 chymotrypsinogen B
**	SP	0	TMUE_s0033003600 Trypsin−domain containing protein
**	SP	0	TMUE_s0104004500 Trypsin−domain containing protein
**	SP	0	TMUE_s0069001000 Trypsin−domain containing protein
**	SP	0	TMUE_s0117000900 transmembrane serine protease 8
**	.	0	TMUE_s0049001600 Trypsin−domain containing protein
**	SP	0	TMUE_s0004005300 Trypsin−domain containing protein
**	SP	0	TMUE_s0048004000 serine protease
**	SP	0	TMUE_s0033003400 Trypsin−domain containing protein
**	SP	0	TMUE_s0048003500 serine protease
.	SP	0	TMUE_s0009003900 Proclotting enzyme
.	.	0	TMUE_s0208001100 CUB and Ldl recept a and Trypsin−domain containing protein
.	.	0	TMUE_s0117002800 Trypsin and CUB−domain containing protein
**	SP	0	TMUE_s0003007500 coagulation factor IX
**	.	0	TMUE_s0003007400 peptidase, S1A subfamily
**	SP	0	TMUE_s0004006800 Trypsin−domain containing protein
**	SP	0	TMUE_s0004007100 Trypsin−domain containing protein
**	.	0	TMUE_s0025005200 Trypsin−domain containing protein
**	SP	0	TMUE_s0191000800 Trypsin−domain containing protein
**	.	0	TMUE_s0049001500 Trypsin−domain containing protein
**	.	0	TMUE_s0058001000 Trypsin−domain containing protein
**	SP	0	TMUE_s0118000700 hypothetical protein
**	.	0	TMUE_s0053003800 Trypsin−domain containing protein
**	.	0	TMUE_s0161000300 coagulation factor IX
**	SP	0	TMUE_s0161000400 coagulation factor IX
**	SP	0	TMUE_s0006011200 transmembrane serine protease 8
**	.	0	TMUE_s0250000200 Trypsin−domain containing protein
**	SP	1	TMUE_s0053003500 Pfam−B 5075 and Trypsin−domain containing protein
**	SP	0	TMUE_s0161000500 coagulation factor IX
**	SP	0	TMUE_s0094002300 Pfam−B 2610 and Trypsin and MMR HSR1−domain containing protein
**	SP	0	TMUE_s0144000200 transmembrane serine protease 8
**	SP	0	TMUE_s0046002200 Trypsin−domain containing protein
**	SP	0	TMUE_s0232001100 serine protease
**	.	1	TMUE_s0092004500 Ldl recept a and Trypsin−domain containing protein
**	SP	0	TMUE_s0104002300 coagulation factor IX
**	SP	0	TMUE_s0025004300 coagulation factor IX
**	SP	0	TMUE_s0003007800 transmembrane serine protease 8
**	SP	0	TMUE_s0095002400 Trypsin−domain containing protein
**	.	0	TMUE_s0004006100 newborn larvae specific serine protease SS2 1
.	SP	1	TMUE_s0070002100 prolow density lipoprotein receptor
.	.	1	TMUE_s0029000100 Ldl recept a and Ldl recept b and EGF and cEGF−domain containing protein
.	.	1	TMUE_s0046000900 low density lipoprotein receptor protein
.	.	0	TMUE_s0118002500 serine protease
**	SP	0	TMUE_s0096002300 newborn larvae specific serine protease SS2 1
**	SP	0	TMUE_s0050005500 Trypsin−domain containing protein
**	.	1	TMUE_s0250000400 Trypsin−domain containing protein
*	.	0	TMUE_s0025004400 serine proteinase
*	.	2	TMUE_s0300000500 cEGF and EGF CA−domain containing protein
**	.	1	TMUE_s0078002100 conserved hypothetical protein
**	SP	0	TMUE_s0213000300 serine protease
.	.	0	TMUE_s0016008300 signal peptide, CUB and EGF
.	.	0	TMUE_s0213000200 serine protease
.	SP	0	TMUE_s0050001400 transmembrane protease serine 9
.	SP	0	TMUE_s0057006400 transmembrane serine protease 8
.	SP	0	TMUE_s0089004100 Ldl recept a−domain containing protein
.	SP	1	TMUE_s0014009000 low density lipoprotein receptor protein
.	.	1	TMUE_s0009010700 Ldl recept a−domain containing protein
.	.	1	TMUE_s0089000800 transmembrane protease serine 6
.	SP	0	TMUE_s0213000500 serine protease
*	SP	0	TMUE_s0021007500 transmembrane serine protease; coagulation factor ix
.	.	0	TMUE_s0213000100 serine protease
.	.	0	TMUE_s0062003900 urokinase type plasminogen activator
.	.	0	TMUE_s0417000100 polprotein

Supplementary Figure 4d

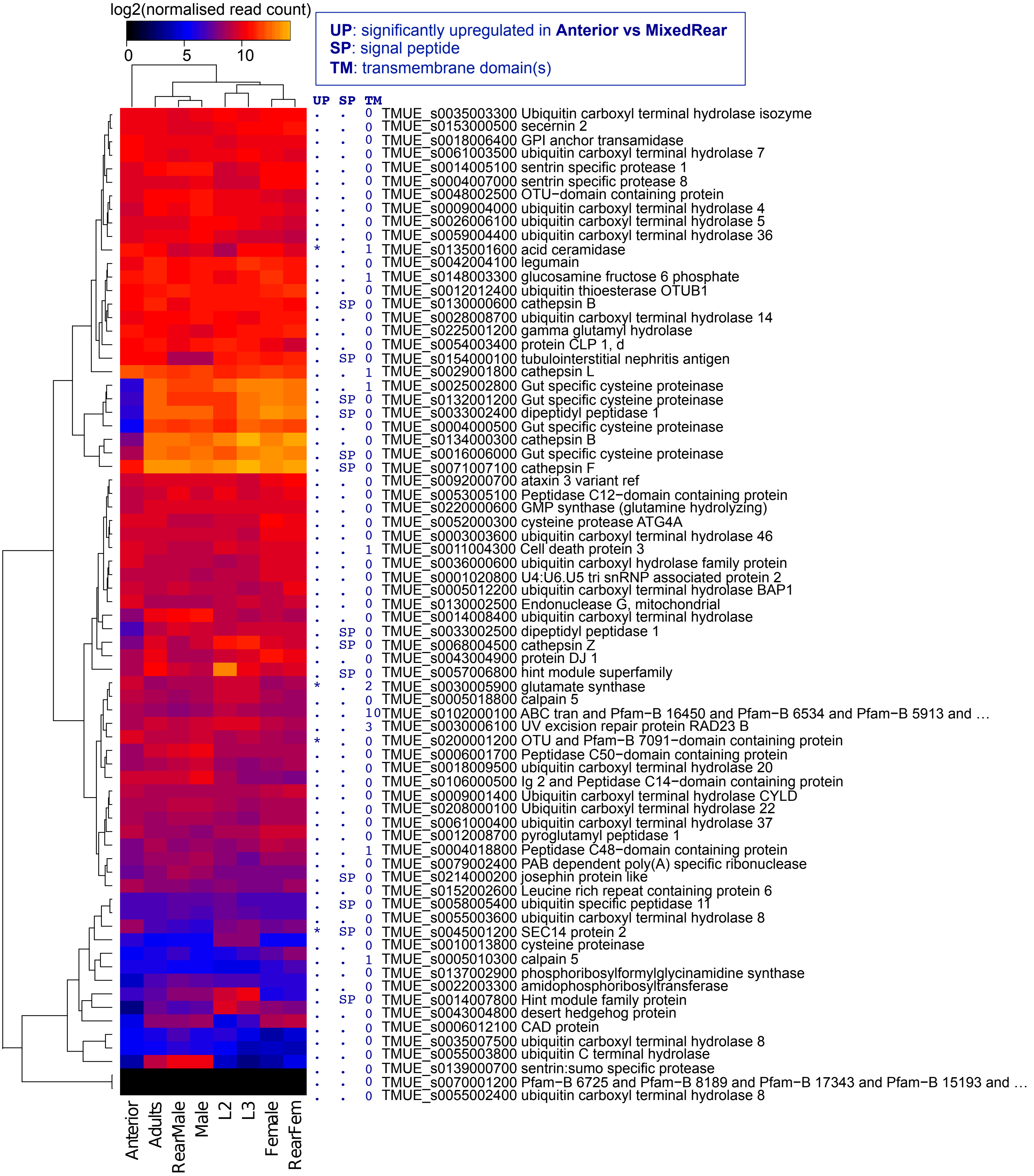
Serine proteases (MEROPS family S9): with no strong group-wide differential expression pattern, with few secreted proteins, and with 1-3 transmembrane domains in the majority of cases.



Nature Genetics: doi:10.1038/ng.3010

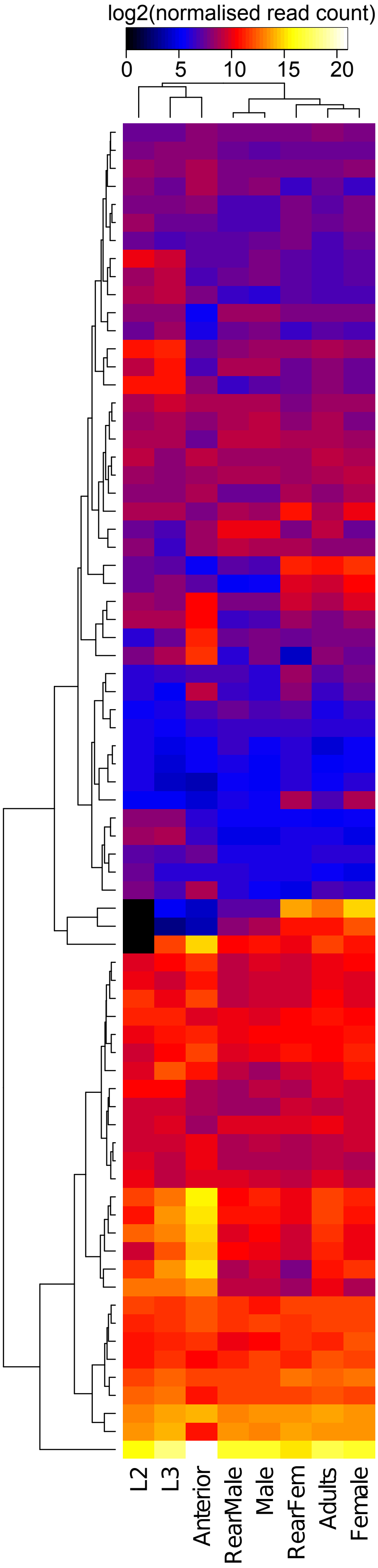
Supplementary Figure 4e

Cysteine proteases: no strong group-wide differential expression pattern and few secreted proteins.



Supplementary Figure 4f

Protease inhibitors (in addition to MEROPS family I17): with no strong group-wide differential expression pattern and some secreted proteins.

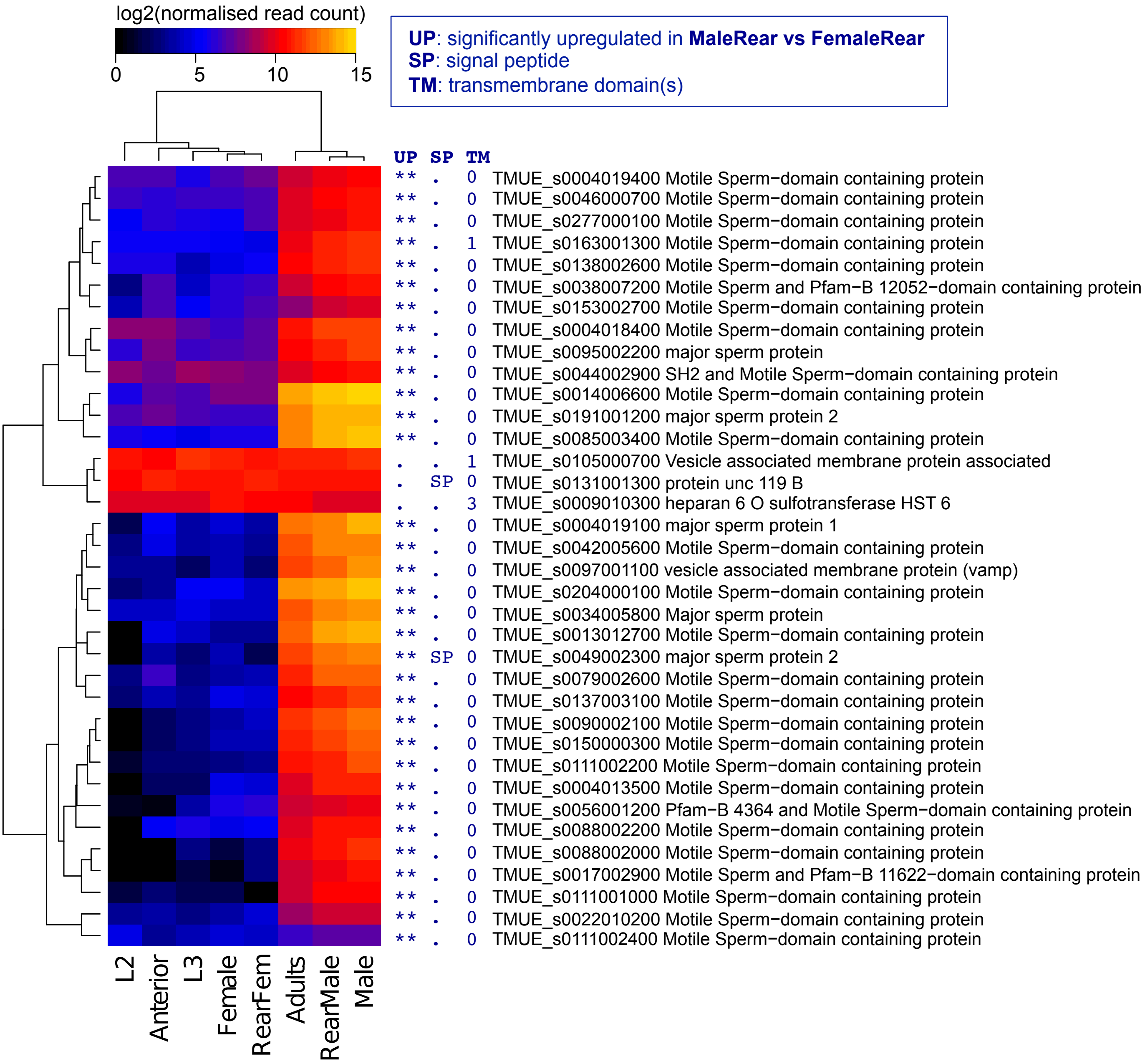


UP: significantly upregulated in **Anterior vs MixedRear**
SP: signal peptide
TM: transmembrane domain(s)
MER: MEROPS family

UP	SP	TM	MER	
.	.	0	I04	TMUE_s0003002200 serine proteinase inhibitor
.	.	0	I32	TMUE_s0078000100 baculoviral IAP repeat containing protein 5
*	SP	2	I02	TMUE_s0013004600 kunitz:Bovine pancreatic trypsin
*	.	2	I15	TMUE_s0300000500 cEGF and EGF CA−domain containing protein
.	SP	0	I29	TMUE_s0071007000 cathepsin F cysteine proteinase
.	SP	0	I02	TMUE_s0012002900 kunitz:Bovine pancreatic trypsin
.	.	0	I02	TMUE_s0164001300 kunitz:Bovine pancreatic trypsin inhibitor
.	SP	0	I02	TMUE_s0077003100 Mesocentin
.	SP	1	I63	TMUE_s0014009700 EGF CA and Sushi and HYR and GCC2 GCC3 and hEGF and EGF …
*	.	2	I01	TMUE_s0101003500 Ig 2 and Pfam−B 1401 and PAN 1 and Zona pellucida and Kazal 2 …
.	.	0	I32	TMUE_s0230000400 baculoviral IAP repeat containing protein
.	.	0	I02	TMUE_s0164000600 papilin
.	SP	0	I31	TMUE_s0164001200 Thyroglobulin 1−domain containing protein
.	SP	0	I02	TMUE_s0001028500 scaffold protein salvador
*	SP	0	I02	TMUE_s0005007100 protein mig c; protein mig b; protein mig a
.	.	0	I04	TMUE_s0140000600 serine protease inhibitor 1 serpin
.	.	0	I21	TMUE_s0232000200 neuroendocrine protein 7B2
.	.	0	I15	TMUE_s0001017100 Laminin protein epi 1
.	.	0	I04	TMUE_s0003002000 serine proteinase inhibitor
.	.	12	I01	TMUE_s0001028800 sodium independent organic anion transporter
.	.	0	I02	TMUE_s0164001500 kunitz:bovine pancreatic trypsin inhibitor
.	SP	0	I01	TMUE_s0009003100 thyroglobulin type 1 repeat containing
.	.	0	I51	TMUE_s0004006900 phosphatidylethanolamine binding protein
.	.	0	I02	TMUE_s0156000200 kunitz:Bovine pancreatic trypsin inhibitor
.	.	1	I08	TMUE_s0281000600 CBM 14 and TIL−domain containing protein
.	SP	0	I02	TMUE_s0052005700 tissue factor pathway inhibitor
*	.	12	I01	TMUE_s0051006400 Solute carrier organic anion transporter family
**	.	0	I01	TMUE_s0033003300 K tetra and Kazal 1−domain containing protein
**	SP	0	I29	TMUE_s0071006900 Cathepsin F
**	.	12	I01	TMUE_s0051006200 Solute carrier organic anion transporter family
.	.	0	I01	TMUE_s0071005400 kazal type serine protease inhibitor
*	.	11	I01	TMUE_s0051006500 kazal type serine protease inhibitor
.	.	0	I02	TMUE_s0156000100 kunitz:Bovine pancreatic trypsin inhibitor
.	.	0	I01	TMUE_s0048000200 agrin
.	SP	0	I02	TMUE_s0198001800 kunitz:Bovine pancreatic trypsin
.	.	0	I02	TMUE_s0351000100 kunitz:Bovine pancreatic trypsin
.	.	0	I02	TMUE_s0002001600 kunitz:Bovine pancreatic trypsin
.	SP	0	I02	TMUE_s0130000700 kunitz:Bovine pancreatic trypsin
.	.	0	I29	TMUE_s0010013800 cysteine proteinase
*	SP	0	I08	TMUE_s0005002800 WAP domain containing protein, SLPI−like
*	.	0	I25B	TMUE_s0015009000 GTP binding protein yptV4
.	SP	0	I02	TMUE_s0019004700 kunitz:Bovine pancreatic trypsin
**	SP	0	I02	TMUE_s0186000700 papilin
.	SP	0	I02	TMUE_s0005001100 Kunitz protease inhibitor
.	SP	0	I02	TMUE_s0006007600 serine protease inhibitor
**	.	0	I02	TMUE_s0137000400 collagen alpha 3(VI) chain
**	SP	0	I04	TMUE_s0003002400 proteinase inhibitor I4 serpin
**	SP	0	I39	TMUE_s0138000200 A macroglobulin complement component family
**	SP	0	I08	TMUE_s0010011500 cysteine string protein
.	.	0	I51	TMUE_s0223000400 phosphatidylethanolamine binding protein
.	.	1	I15	TMUE_s0197001000 multiple epidermal growth factor domains
*	.	1	I04	TMUE_s0003002300 Serine proteinase inhibitor
*	.	0	I04	TMUE_s0038004600 60S ribosomal protein L38
.	SP	0	I02	TMUE_s0024001300 TSP 1 and Reeler and Spond N and Peptidase M23 and Kunitz …
.	.	12	I01	TMUE_s0051003100 solute carrier organic anion transporter family
.	.	0	I51	TMUE_s0001006300 phosphatidylethanolamine binding protein
**	.	0	I51	TMUE_s0058003600 phosphatidylethanolamine binding protein
**	.	10	I01	TMUE_s0031002100 solute carrier organic anion transporter family
.	SP	0	I08	TMUE_s0116000900 WAP domain containing protein, SLPI−like
**	.	2	I04	TMUE_s0016004100 Serpin protein
**	SP	0	I25C	TMUE_s0158001500 Cystatin−domain containing protein
**	SP	0	I02	TMUE_s0186000600 papilin
**	.	0	I25B	TMUE_s0014011700 Cystatin and Pfam−B 3190−domain containing protein
**	SP	0	I01	TMUE_s0045002900 serine protease inhibitor Kazal type 4
**	SP	0	I02	TMUE_s0241000200 hypothetical protein
.	.	0	I08	TMUE_s0007006700 chymotrypsin inhibitor
.	.	1	I29	TMUE_s0029001800 cathepsin L
.	.	1	I04	TMUE_s0003001900 serine proteinase inhibitor
.	.	0	I01	TMUE_s0086001300 SPARC
.	.	0	I51	TMUE_s0004006600 OV 16 antigen
.	SP	0	I31	TMUE_s0060005500 Thyroglobulin type 1 repeat family protein
.	.	0	I25B	TMUE_s0010008300 onchocystatin
.	SP	0	I25C	TMUE_s0071007100 cathepsin F
**	.	0	I08	TMUE_s0004002000 chymotrypsin inhibitor

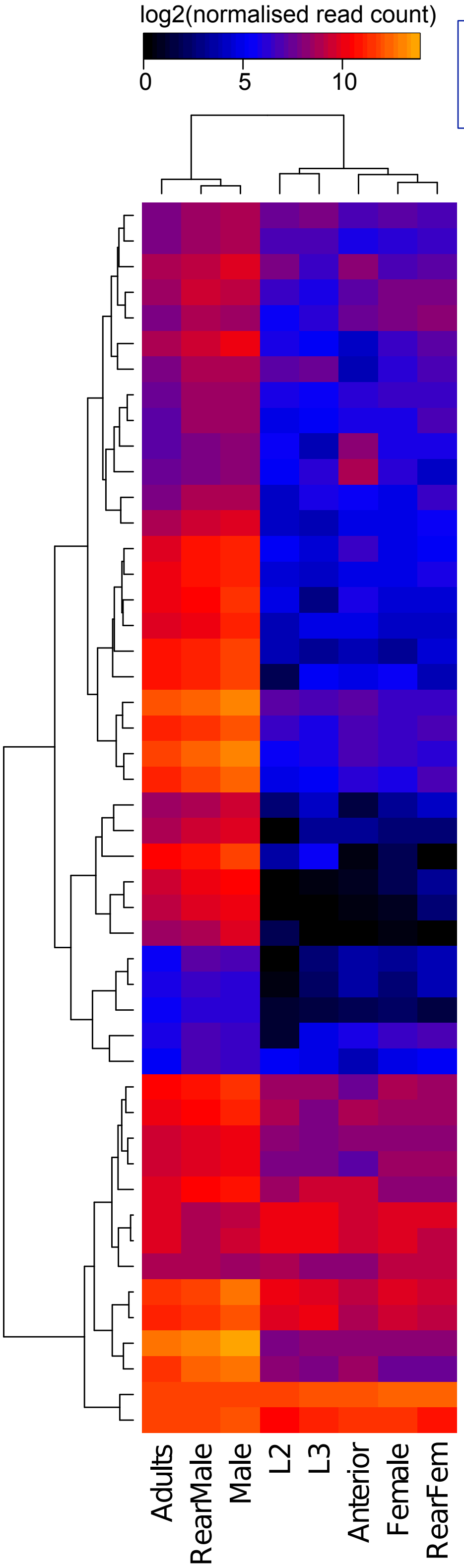
Supplementary Figure 4g

Major Sperm Protein domain (PF00635) containing proteins: with strongly upregulated expression in male whipworms and virtually no secreted proteins.



Supplementary Figure 4h

Casein kinase-related (PTHR11909) proteins: with strongly upregulated expression in male whipworms and no secreted proteins.

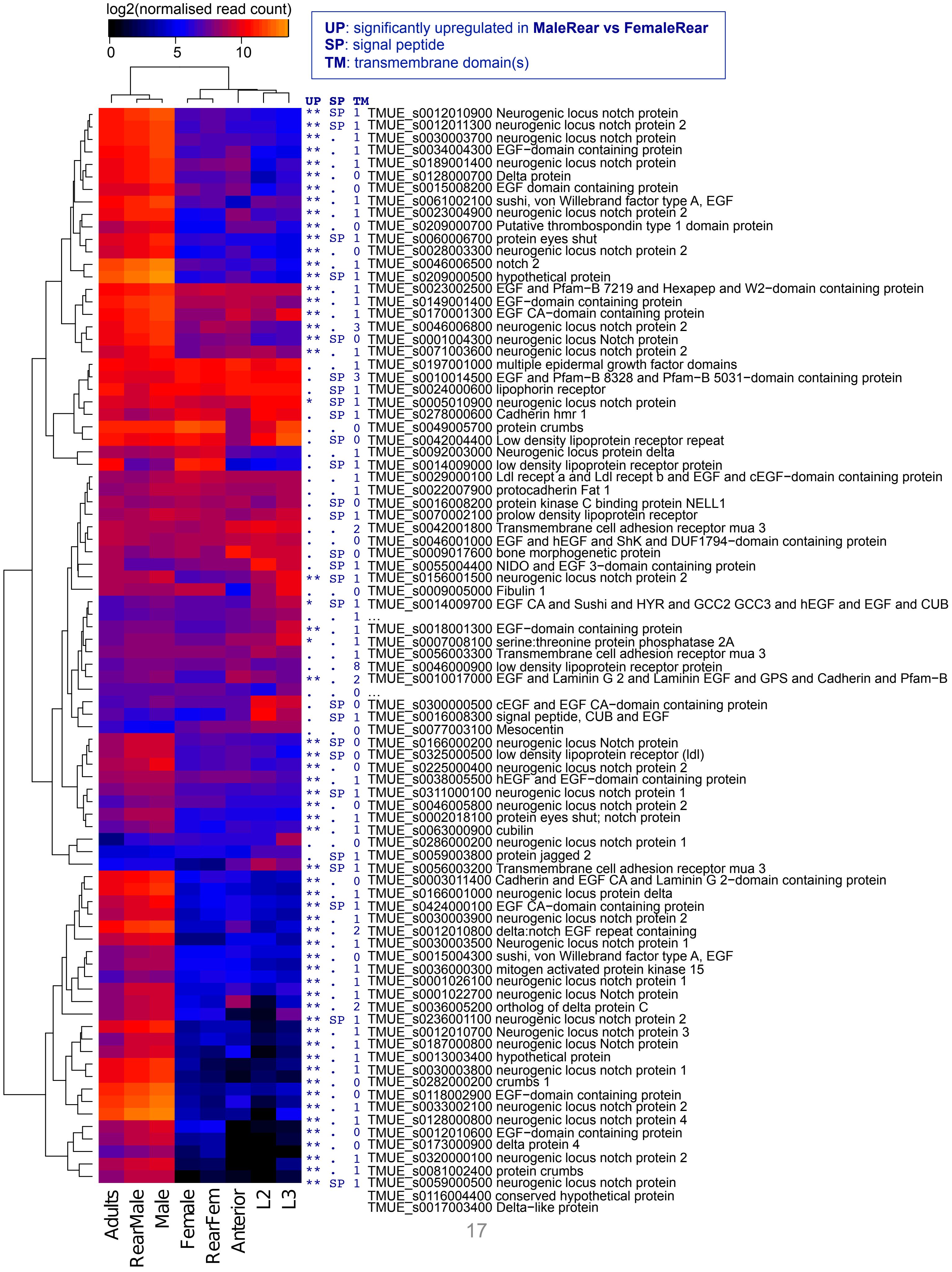


UP: significantly upregulated in **MaleRear vs FemaleRear**
SP: signal peptide
TM: transmembrane domain(s)

UP	SP	TM	
**	.	0	TMUE_s0031000900 Tau tubulin kinase 1
**	.	0	TMUE_s0001001300 Tau tubulin kinase 1
**	.	0	TMUE_s0204001000 tau tubulin kinase 1
**	.	0	TMUE_s0056001900 Pkinase−domain containing protein
*	.	0	TMUE_s0081003600 tau tubulin kinase 2
**	.	0	TMUE_s0032004700 Pfam−B 4134 and Pkinase−domain containing protein
**	.	0	TMUE_s0049000800 cell division cycle 7 protein
**	.	0	TMUE_s0148000800 tau tubulin kinase
**	.	0	TMUE_s0148000700 Pkinase−domain containing protein
**	.	0	TMUE_s0272000200 tau tubulin kinase 1
**	.	0	TMUE_s0236000800 protein vertebrate tau tubulin
**	.	0	TMUE_s0081003500 Pkinase−domain containing protein
**	.	0	TMUE_s0007006600 Tau tubulin kinase 1
**	.	0	TMUE_s0243000300 Tau tubulin kinase 1
**	.	0	TMUE_s0014002700 Tau tubulin kinase 1
**	.	0	TMUE_s0003007300 Tau tubulin kinase 1
**	.	0	TMUE_s0085003000 Tau tubulin kinase 1
**	.	0	TMUE_s0148000500 ttbk family protein kinase; protein serine threonine kinase
**	.	0	TMUE_s0028003500 Pkinase−domain containing protein
**	.	0	TMUE_s0241001300 Pkinase−domain containing protein
**	.	0	TMUE_s0185002100 Tau tubulin kinase 2
**	.	0	TMUE_s0061004100 Tau tubulin kinase 1
**	.	0	TMUE_s0148000600 ttbk family protein kinase; protein serine threonine kinase
**	.	0	TMUE_s0033004200 Pfam−B 4134 and Pkinase−domain containing protein
**	.	0	TMUE_s0151000600 CK1:TTBK protein kinase
**	.	0	TMUE_s0032005200 Tau tubulin kinase 1
**	.	0	TMUE_s0167001300 Tau tubulin kinase 1
**	.	0	TMUE_s0211000600 tau tubulin kinase 1
**	.	0	TMUE_s0044004500 CK1:TTBK protein kinase
**	.	0	TMUE_s0236000500 Pkinase−domain containing protein
**	.	0	TMUE_s0315000300 Pfam−B 4134 and Pkinase−domain containing protein
**	.	0	TMUE_s0195000500 Pfam−B 4134 and Pkinase−domain containing protein
.	.	0	TMUE_s0200001500 Pkinase and Pfam−B 4134−domain containing protein
*	.	0	TMUE_s0013009700 Tau tubulin kinase 1
**	.	0	TMUE_s0001001200 Tau tubulin kinase 1
**	.	0	TMUE_s0016006600 Tau tubulin kinase 1
**	.	6	TMUE_s0008017600 protein R90.1
**	.	0	TMUE_s0025002500 nucleosomal histone kinase 1
**	.	0	TMUE_s0126002200 Tau tubulin kinase 1
.	.	0	TMUE_s0010003500 Tau tubulin kinase 1
.	.	0	TMUE_s0238000100 casein kinase 1 protein, gamma
.	.	1	TMUE_s0076001500 CK1:CK1:CK1 A protein kinase
**	.	0	TMUE_s0146001300 casein kinase i epsilon
**	.	0	TMUE_s0146001200 CK1:CK1:CK1 D protein kinase
**	.	0	TMUE_s0053004600 Tau tubulin kinase 1
**	.	0	TMUE_s0051002900 Tau tubulin kinase 1
.	.	0	TMUE_s0023006300 Casein kinase I alpha
**	.	0	TMUE_s0117002200 casein kinase I alpha

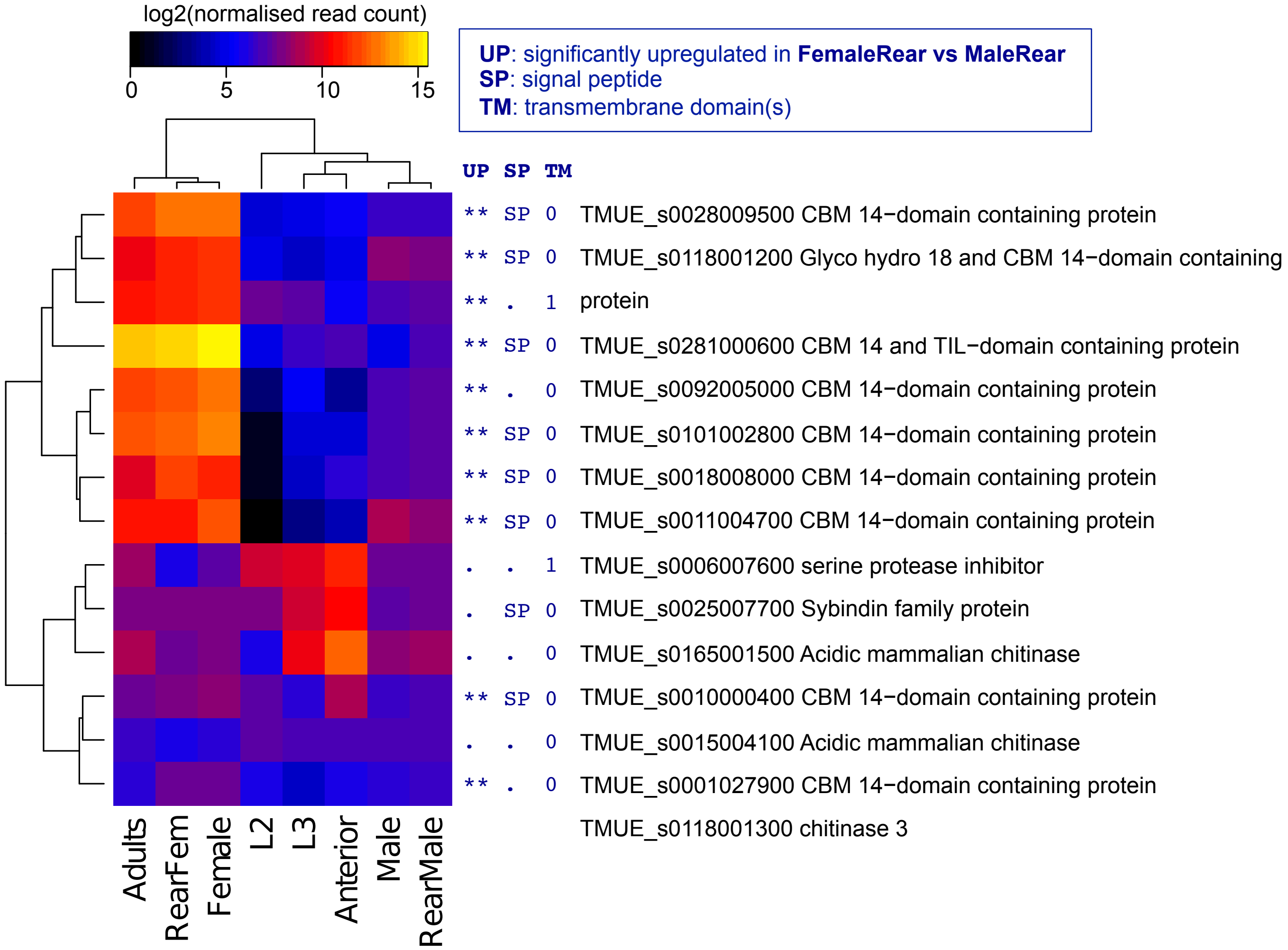
Supplementary Figure 4i

Epidermal Growth Factor-like domain (SM00181) containing proteins: tend to be upregulated in male whipworms and to contain 1-3 transmembrane domains.



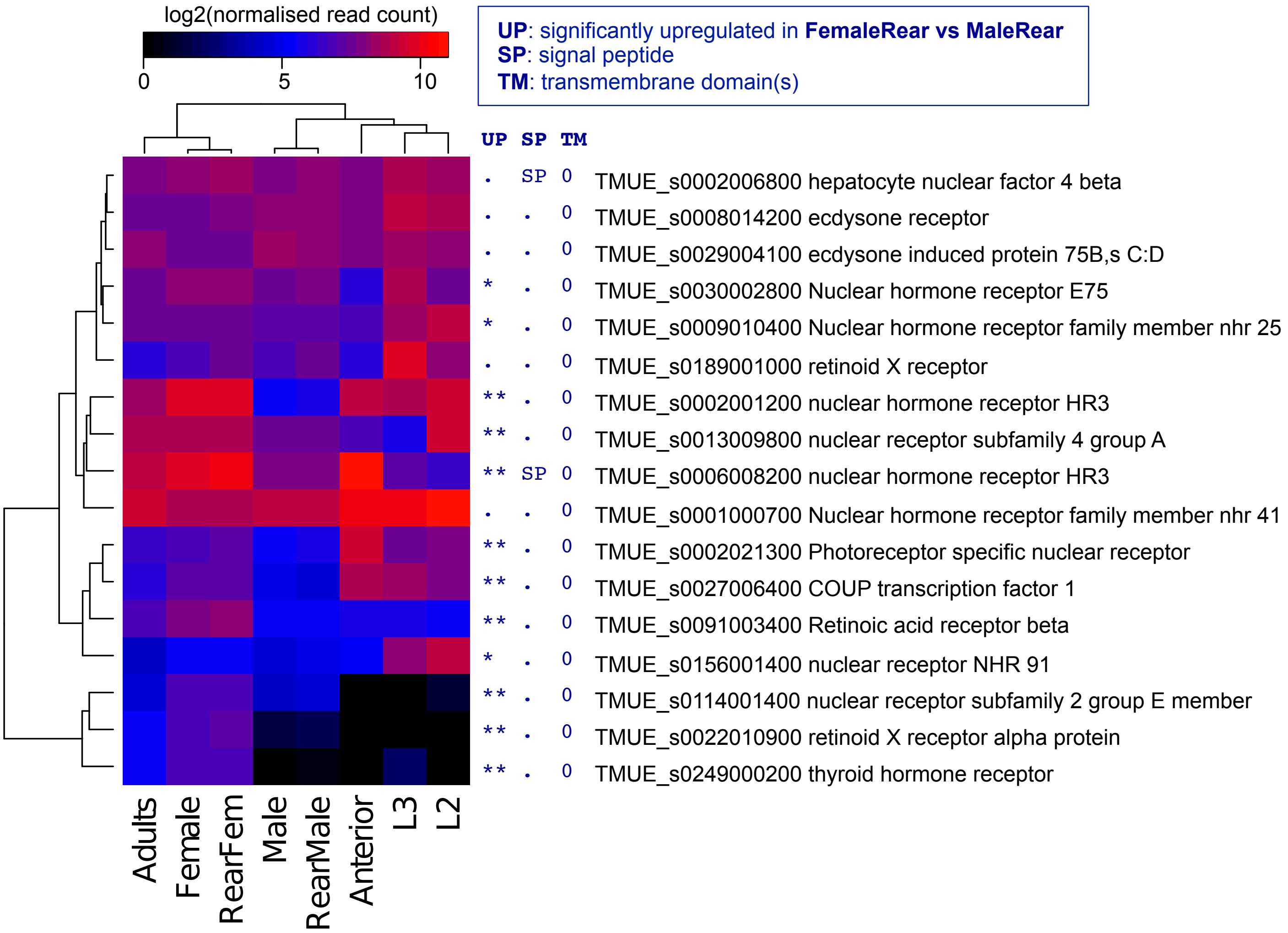
Supplementary Figure 4j

Chitin-binding domain (PF01607) containing proteins: tend to be upregulated in female whipworms or the anterior region and to be secreted.



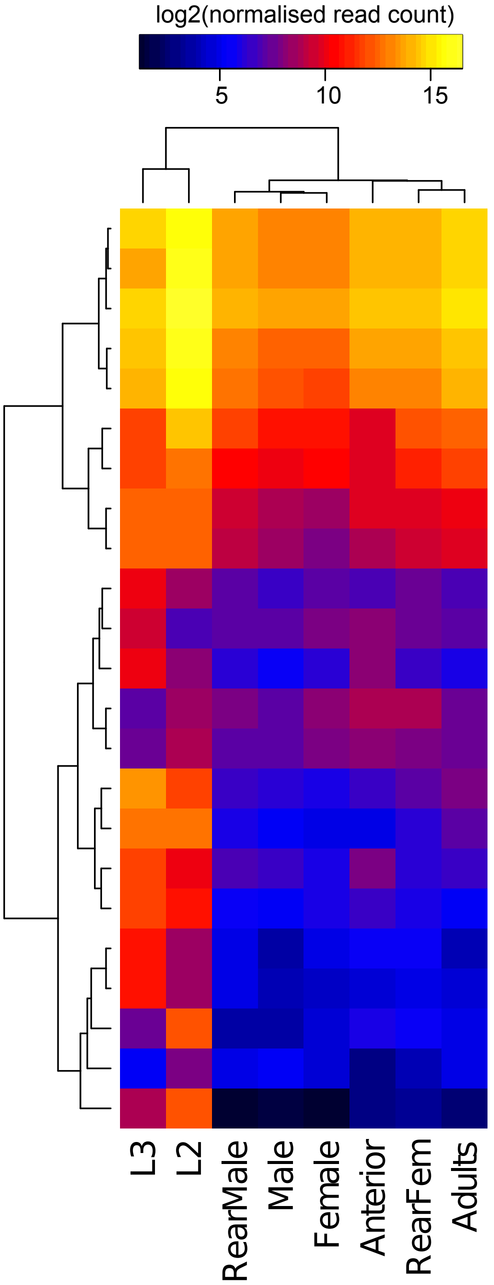
Supplementary Figure 4k

C4 zinc finger in nuclear hormone receptors (SM00399) domain containing proteins: tend to be slightly but significantly upregulated in female whipworms, the anterior region or in larvae.



Supplementary Figure 4I

Nematode cuticle collagen N-terminal domain (PF01484) containing proteins: with strongly upregulated expression in larvae and typically one transmembrane domain.

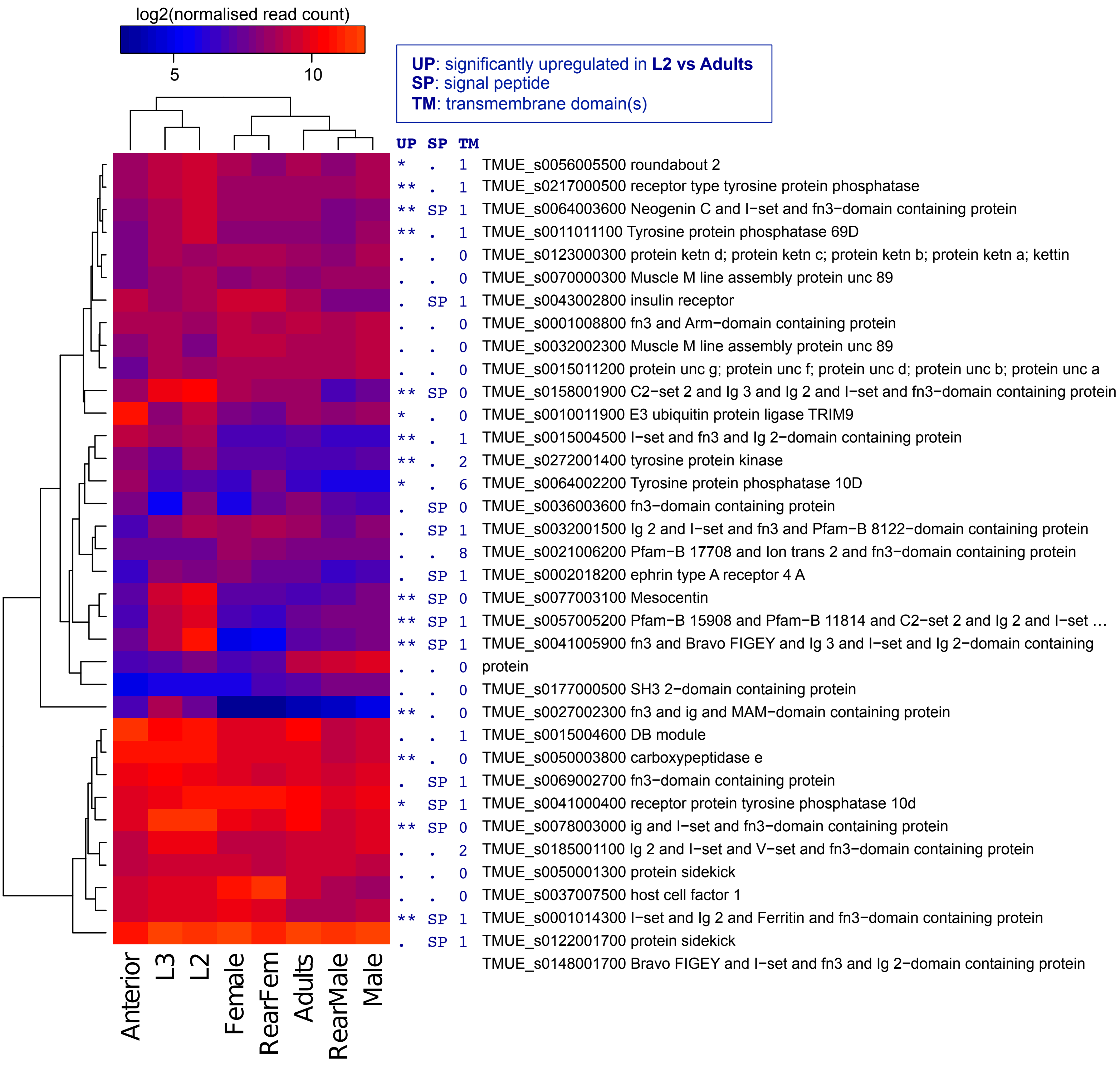


UP: significantly upregulated in **L2 vs Adults**
SP: signal peptide
TM: transmembrane domain(s)

UP	SP	TM	
*	.	1	TMUE_s0032001700 Pfam−B 677 and Collagen and Pfam−B 798 and Col cuticle N−domain …
**	.	1	TMUE_s0072001700 Col cuticle N and Collagen and Pfam−B 677−domain containing protein
*	.	1	TMUE_s0084001000 cuticle collagen 34 protein
**	.	1	TMUE_s0177000600 Cuticle collagen 14
**	.	1	TMUE_s0182001200 collagen protein 48
**	.	1	TMUE_s0084000300 cuticle collagen 7
*	SP	0	TMUE_s0084000500 cuticle collagen 7
**	.	1	TMUE_s0032003200 cuticle collagen rol 6
**	.	1	TMUE_s0032003700 Pfam−B 798 and Collagen and Col cuticle N−domain containing protein
**	.	1	TMUE_s0223000600 Col cuticle N and Collagen and Pfam−B 798 and Pfam−B 677−domain …
.	.	1	TMUE_s0032003500 Collagen and Pfam−B 798 and Col cuticle N−domain containing protein
**	.	1	TMUE_s0042005900 cuticle collagen rol 6
*	.	1	TMUE_s0153002800 cuticle collagen 6 (protein roller 8)
**	.	1	TMUE_s0094003500 Col cuticle N and Collagen and Las1−domain containing protein
**	.	1	TMUE_s0005011800 Col cuticle N−domain containing protein
**	.	1	TMUE_s0281000800 Col cuticle N and Collagen and Pfam−B 798−domain containing protein
**	.	1	TMUE_s0060003000 Collagen and Pfam−B 798 and Col cuticle N−domain containing protein
**	.	1	TMUE_s0032003800 Pfam−B 677 and Collagen and Col cuticle N−domain containing protein
**	.	1	TMUE_s0040004200 cuticle collagen 6 (protein roller 8)
**	.	1	TMUE_s0135002200 Cuticle collagen 34
**	.	1	TMUE_s0050007200 cuticle collagen 39
**	.	1	TMUE_s0079003000 Pfam−B 677 and Collagen and Pfam−B 798 and Col cuticle N−domain …
**	.	1	TMUE_s0209001300 Cuticle collagen lon 3

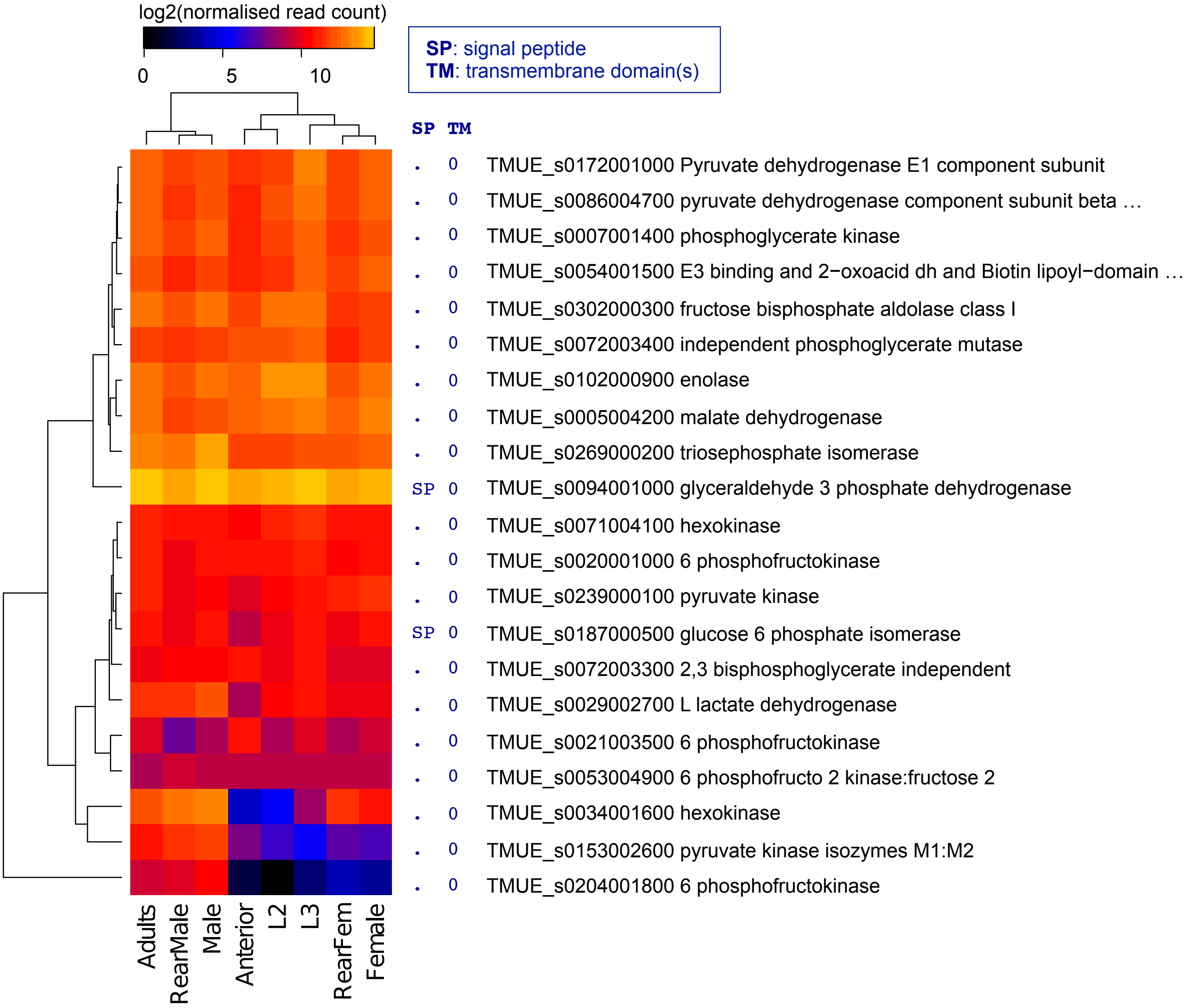
Supplementary Figure 4m

Fibronectin type 3 (SM00060) domain containing proteins: tend to be upregulated in larvae.



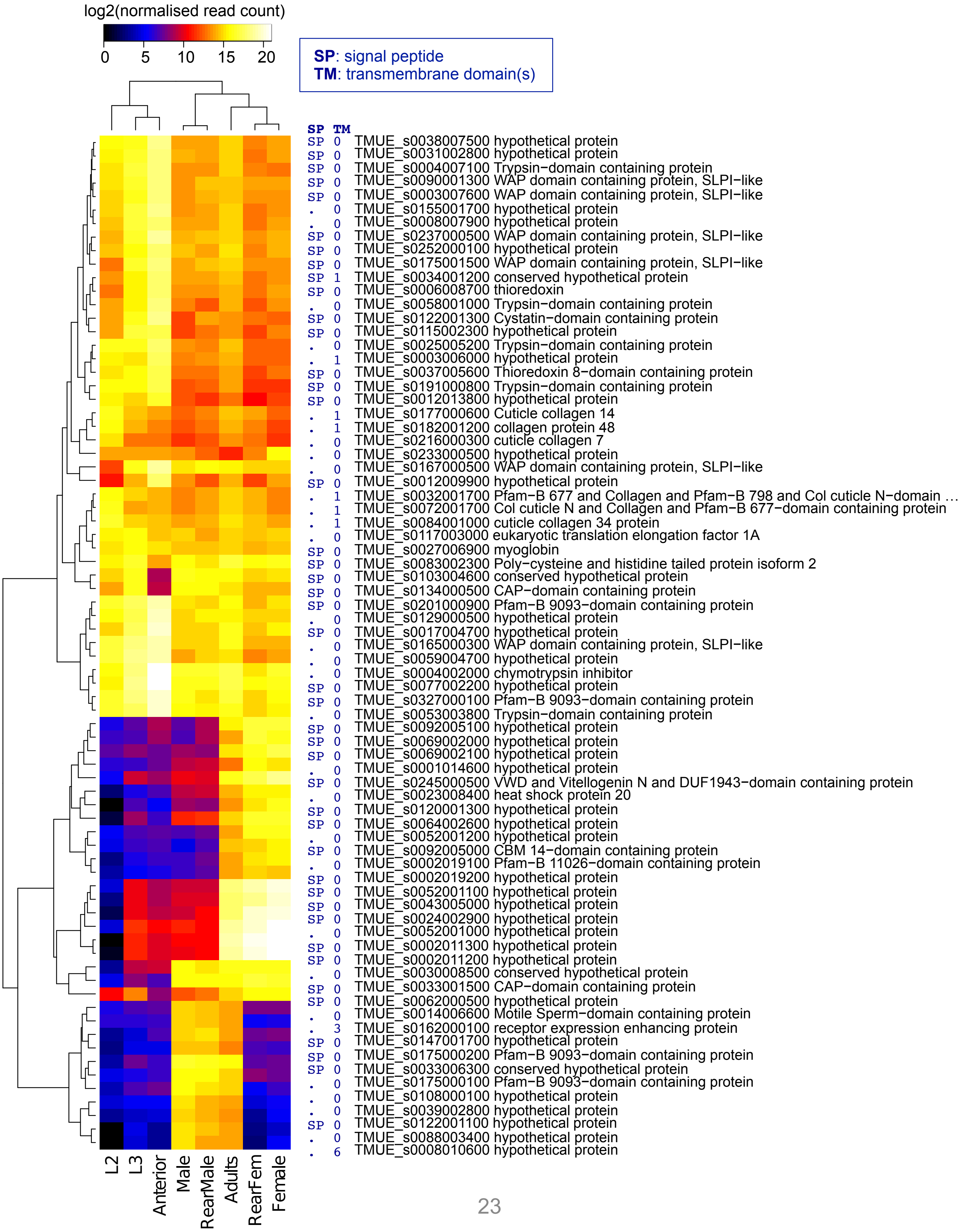
Supplementary Figure 4n

Glycolysis-related proteins (representing a set of “housekeeping genes”): apart from a few apparently gender-specific isoforms of hexokinase, pyruvate kinase, and 6-phosphofructokinase, no strong group-wide differential expression pattern, and few secreted proteins.

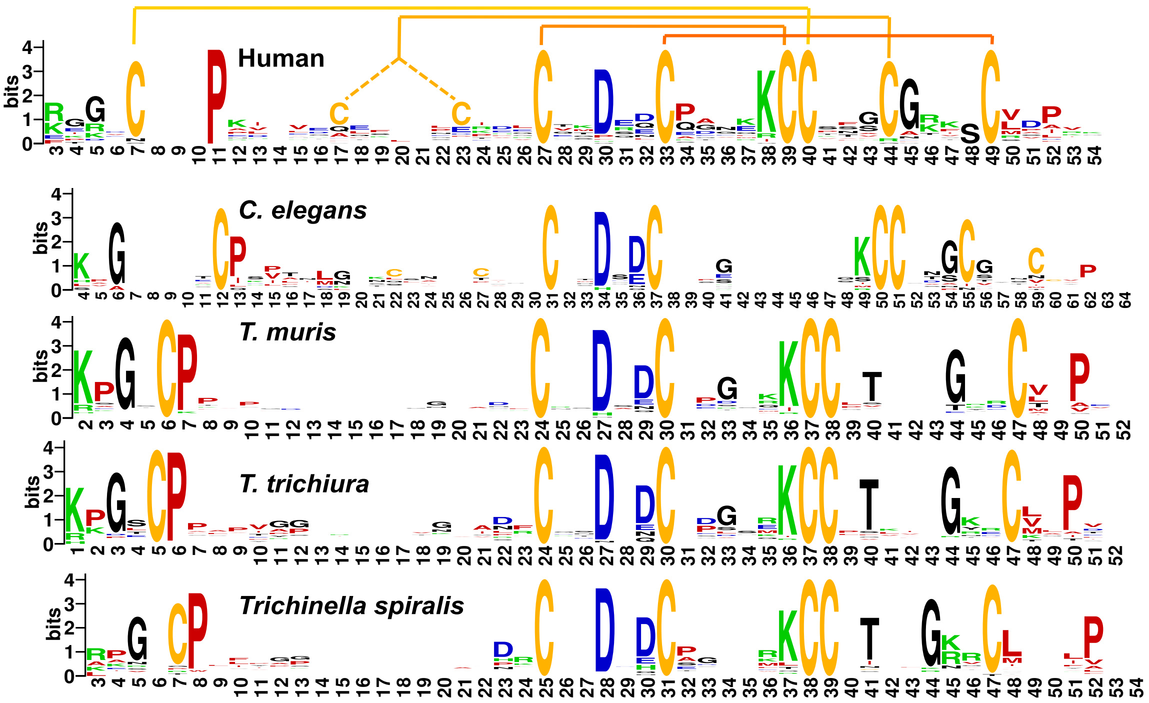


Supplementary Figure 4o

The top 25 most abundant transcripts of each biological sample combined: secreted proteins and those of unknown function (“hypothetical protein”) are overrepresented among the transcriptionally most highly expressed genes of *T. muris*. Fractions of sequences with predicted SP: here 57.3% (43 of 75), proteome-wide 11.5% (1,265 of 11,004). Fractions of “hypothetical proteins”: here 52% (39 of 75), proteome-wide 34.9% (3,837 of 11,004).

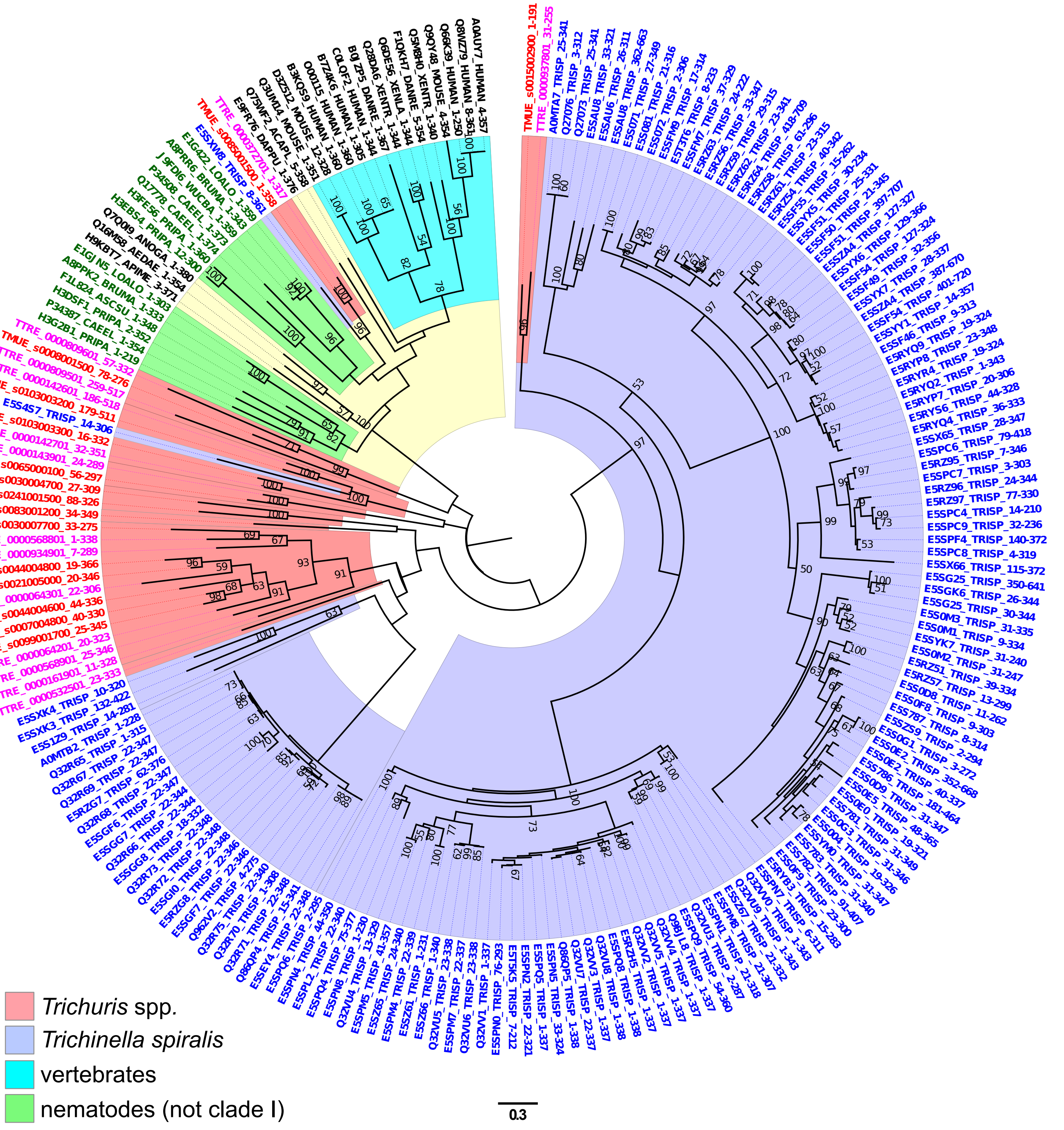


Supplementary Figure 5



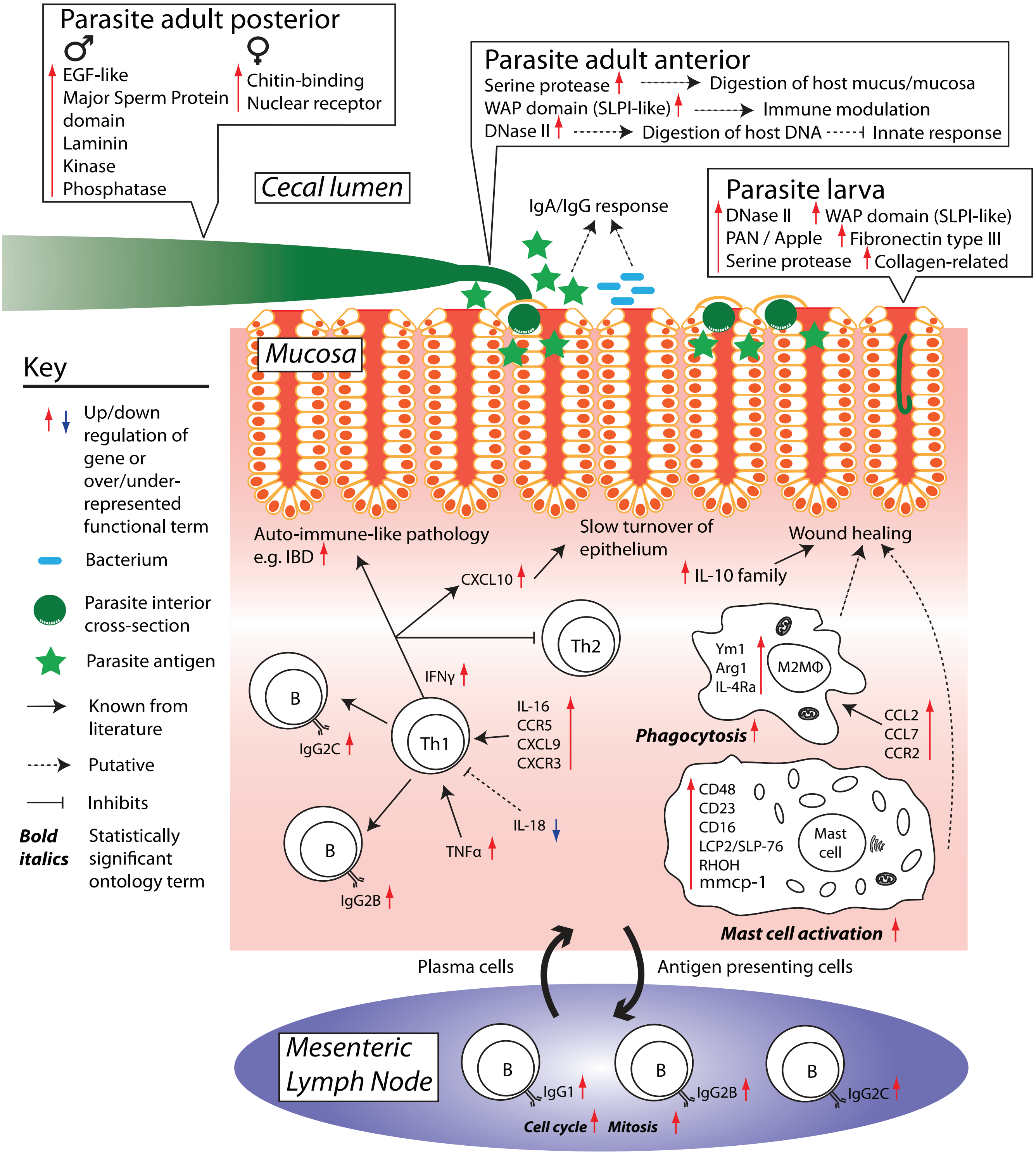
Supplementary Figure 5 Sequence logos illustrating the conserved and distinct sequence characteristics of WAP domains (Interpro IPR008197) found in proteins of *H. sapiens*, *C. elegans*, *T. muris, T. trichiura*, and *Trichinella spiralis*. The canonical four disulfide bonds are highlighted in the sequence logo of the human WAP domains. The sequence logos representing the different species are aligned around the central CxxDxxC motif. This supplementary figure is a detailed version of Fig. 3b in the main text.

Supplementary Figure 6



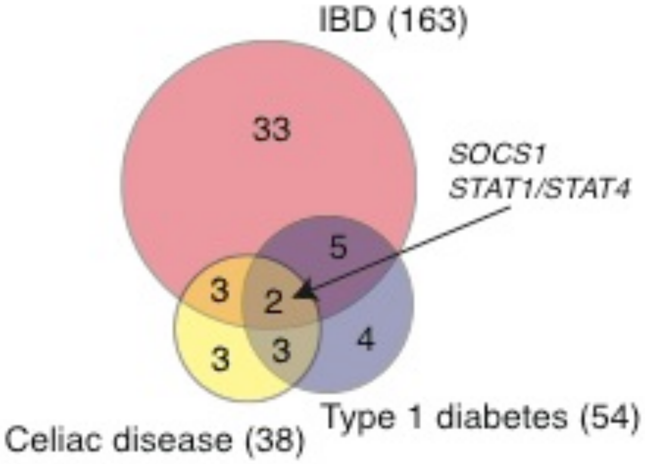
Supplementary Figure 6 Phylogenetic analysis of DNase II-like proteins of *Trichuris* and *Trichinella*. A maximum-likelihood phylogeny of DNase II protein domains (IPR004947) illustrates the relationships between DNase II domains of proteins from *Trichuris* spp, *Trichinella spiralis*, other nematodes, insects/other invertebrates, and vertebrates. The labels include UniProt accession numbers. This supplementary figure is a detailed version of Fig. 4b in the main text.

Supplementary Figure 7



Supplementary Figure 7 A summary of key gene expression changes in *Trichuris muris* and its host. Groups of genes with common function that are significantly enriched amongst upregulated genes for different parasite tissues and life stages were identified (green). For some of these genes, putative roles in host parasite interactions in the adult anterior are proposed (dashed arrows). The transcriptomic response of the host cecum and mesenteric lymph node to *Trichuris* infection is summarized in terms of known (solid black arrows) and hypothesized (dashed black arrows) immunological interactions. Throughout, red/blue arrows indicate an upregulation/downregulation of a gene or enrichment/reduction of a functional category. Changes in transcript abundance were determined using EdgeR and DESeq with a false discovery rate (FDR) of 5%. Functional enrichment was determined by using TopGO with FDR 5% (mouse genes) and by performing a protein domain enrichment analysis with p-value <1% (whipworm genes).

Supplementary Figure 8



Supplementary Figure 8 Overlap in genes involved in *Trichuris* infection and human auto-immune disease. Numbers in brackets are genes identified as associated with the disease by GWAS. Circles represent number of genes that overlap those differentially expressed during *T. muris* infection.

II. SUPPLEMENTARY TABLES

Supplementary Table 2a: Statistically significant overrepresentation on linkage group X of genes with significantly higher (FDR 0.01) transcriptional expression in female *T. muris* parasites

FDR limit for differentially expressed genes 0.01

Comparison: Female vs Male

LG = linkage group

Contingency table

	LG1	**LG2**	**LGX**
genes NOTdifferentiallyExpressed	1510	1370	781
genes UPinFEMALE	776	827	**768**
genes UPinMALE	1021	880	588

Fisher Exact Test P-values

	LG1 vs LG2	**LG1 vs LGX**	**LG2 vs LGX**
NOTdifferentiallyExpressed vs UPinFEMALE	0.0106	**< 2.2e-16**	**4.00E-13**
NOTdifferentiallyExpressed vs UPinMALE	0.3908	0.1171	**0.0233**

Supplementary Table 2b: Genes present on the scaffolds inferred to represent the Y chromosome of T. muris parasites

Gene	Length [bp]	Scaffold (genome v4)	Gene description	Comment
TMUE_s0013006600	267	TMUE_000120	hypothetical_protein	In final gene set v2.3
TMUE_s0268001000	303	TMUE_000341	hypothetical_protein	In final gene set v2.3
TMUE_s0379000200	219	TMUE_000151	hypothetical_protein	In final gene set v2.3
TMUE_s0379000300	441	TMUE_000151	hypothetical_protein	In final gene set v2.3
TMUE_s0410000200	327	TMUE_000100	hypothetical_protein	In final gene set v2.3
TMUE_s0410000300	315	TMUE_000100	hypothetical_protein	In final gene set v2.3
TMUE_s0487000100	342	TMUE_000205	hypothetical_protein	In final gene set v2.3
TMUE_s0487000100	342	TMUE_000466	hypothetical_protein	In final gene set v2.3
TMUE_s0547000200	246	TMUE_000210	hypothetical_protein	In final gene set v2.3
TMUE_s0547000200	246	TMUE_000363	hypothetical_protein	In final gene set v2.3
TMUE_s0624000100	417	TMUE_000294	hypothetical_protein	In final gene set v2.3
TMUE_s0634000100	119	TMUE_000446	hypothetical_protein	In final gene set v2.3
TMUE_s0634000200	516	TMUE_000446	hypothetical_protein	In final gene set v2.3
TMUE_s0720000100	327	TMUE_000161	hypothetical_protein	In final gene set v2.3
TMUE_s0758000200	444	TMUE_000511	hypothetical_protein	In final gene set v2.3
TMUE_s0846000100	303	TMUE_000533	hypothetical_protein	In final gene set v2.3
TMUE_s0866000100	333	TMUE_000350	hypothetical_protein	In final gene set v2.3
TMUE_s0896000100	264	TMUE_000894	hypothetical_protein	In final gene set v2.3
TMUE_s0915000200	516	TMUE_000263	hypothetical_protein	In final gene set v2.3
TMUE_s0928000100	327	TMUE_000422	hypothetical_protein	In final gene set v2.3
TMUE_s1010000200	546	TMUE_000100	hypothetical_protein	In final gene set v2.3
TMUE_s1019000100	321	TMUE_000238	hypothetical_protein	In final gene set v2.3
TMUE_s1070000100	567	TMUE_000259	hypothetical_protein	In final gene set v2.3
TMUE_s1091000100	483	TMUE_000511	hypothetical_protein	In final gene set v2.3
TMUE_s1096000100	300	TMUE_000158	hypothetical_protein	In final gene set v2.3
TMUE_s1122000100	207	TMUE_000394	hypothetical_protein	In final gene set v2.3
TMUE_s1143000100	438	TMUE_000401	hypothetical_protein	In final gene set v2.3
TMUE_s1153000100	291	TMUE_000558	hypothetical_protein	In final gene set v2.3
TMUE_s1216000100	177	TMUE_000422	hypothetical_protein	In final gene set v2.3
TMUE_s1225000100	438	TMUE_000205	hypothetical_protein	In final gene set v2.3
TMUE_s1268000100	420	TMUE_000726	hypothetical_protein	In final gene set v2.3
TMUE_s1327000100	342	TMUE_000899	hypothetical_protein	In final gene set v2.3
TMUE_s1095000200	807	TMUE_000120	hypothetical_protein	In gene set v2.2, but removed from final gene set v2.3 because transposon-related
TMUE_s0303000700	621	TMUE_000146	conserved_hypothetical_protein	In final gene set v2.3
TMUE_s0346000200	735	TMUE_000492	conserved_hypothetical_protein	In final gene set v2.3
TMUE_s0411000100	471	TMUE_000208	conserved_hypothetical_protein	In final gene set v2.3
TMUE_s0470000100	447	TMUE_000275	conserved_hypothetical_protein	In final gene set v2.3
TMUE_s0508000200	300	TMUE_000391	conserved_hypothetical_protein	In final gene set v2.3
TMUE_s0634000300	735	TMUE_000446	conserved_hypothetical_protein	In final gene set v2.3
TMUE_s0722000100	234	TMUE_000180	conserved_hypothetical_protein	In final gene set v2.3
TMUE_s0803000100	657	TMUE_000150	conserved_hypothetical_protein	In final gene set v2.3
TMUE_s0920000100	825	TMUE_000187	conserved_hypothetical_protein	In final gene set v2.3
TMUE_s1097000100	1317	TMUE_000681	conserved_hypothetical_protein	In final gene set v2.3
TMUE_s1141000100	1194	TMUE_000459	conserved_hypothetical_protein	In final gene set v2.3
TMUE_s1151000100	498	TMUE_000240	conserved_hypothetical_protein	In final gene set v2.3

TMUE_s1174000100	735	TMUE_000670	conserved_hypothetical_protein	In final gene set v2.3
TMUE_s1192000100	996	TMUE_000256	conserved_hypothetical_protein	In final gene set v2.3
TMUE_s1310000100	1308	TMUE_000333	conserved_hypothetical_protein	In final gene set v2.3
TMUE_s0323000300	1020	TMUE_000100	conserved_hypothetical_protein	In gene set v2.2, but removed from final gene set v2.3 because transposon-related
TMUE_s1320000100	1356	TMUE_000485	conserved_hypothetical_protein	In gene set v2.2, but removed from final gene set v2.3 because transposon-related
TMUE_s0044000200	2709	TMUE_000310	DUF1759-domain_containing_protein	In gene set v2.2, but removed from final gene set v2.3 because transposon-related
TMUE_s0372000100	1427	TMUE_000277	DUF1759-domain_containing_protein	In gene set v2.2, but removed from final gene set v2.3 because transposon-related
TMUE_s1077000100	369	TMUE_000342	DUF1759-domain_containing_protein	In gene set v2.2, but removed from final gene set v2.3 because transposon-related
TMUE_s0466000100	2232	TMUE_000302	gag_pol_polyprotein	In gene set v2.2, but removed from final gene set v2.3 because transposon-related
TMUE_s1053000100	483	TMUE_000146	gag_pol_polyprotein	In gene set v2.2, but removed from final gene set v2.3 because transposon-related
TMUE_s0803000200	525	TMUE_000150	Pao_retrotransposon_peptidase_family_protein	In gene set v2.2, but removed from final gene set v2.3 because transposon-related
TMUE_s1192000200	525	TMUE_000256	Pao_retrotransposon_peptidase_family_protein	In gene set v2.2, but removed from final gene set v2.3 because transposon-related
TMUE_s0375000100	2805	TMUE_000120	Peptidase_A17-domain_containing_protein	In gene set v2.2, but removed from final gene set v2.3 because transposon-related
TMUE_s0875000100	2052	TMUE_000323	Peptidase_A17-domain_containing_protein	In gene set v2.2, but removed from final gene set v2.3 because transposon-related
TMUE_s1148000100	1590	TMUE_000505	Peptidase_A17-domain_containing_protein	In gene set v2.2, but removed from final gene set v2.3 because transposon-related
TMUE_s0228000100	1047	TMUE_000169	Pfam-B_2707_and_RVT_1-domain_containing_protein	In gene set v2.2, but removed from final gene set v2.3 because transposon-related
TMUE_s0228000200	690	TMUE_000169	Pfam-B_310-domain_containing_protein	In final gene set v2.3
TMUE_s0228000300	438	TMUE_000169	pol_polyprotein	In gene set v2.2, but removed from final gene set v2.3 because transposon-related
TMUE_s0303000800	3873	TMUE_000146	polyprotein	In final gene set v2.3
TMUE_s0642000100	1476	TMUE_000494	rve-domain_containing_protein	In gene set v2.2, but removed from final gene set v2.3 because transposon-related
TMUE_s0444000100	3780	TMUE_000251	RVT_1_and_Peptidase_A17_and_rve-domain_containing_protein	In gene set v2.2, but removed from final gene set v2.3 because transposon-related
TMUE_s0469000200	678	TMUE_000256	Tudor-knot-domain_containing_protein	In final gene set v2.3
TMUE_s0508000100	417	TMUE_000391	Uncharacterized_transposase_protein	In gene set v2.2, but removed from final gene set v2.3 because transposon-related
TMUE_s1113000100	267	TMUE_000242	zinc_finger_protein	In final gene set v2.3

Supplementary Table 3a: Gene families showing significant changes (z-score > 1.96) in copy number between clade I species and other nematodes, sorted by total copy number across nematode species.

family	*Trichinella spiralis*	*Bursaphelenchus xylophilus*	*Brugia malayi*	*T. muris*	*C. elegans*	*T. trichiura*	z-score	annotation
FAM47	3	10	8	3	8	3	2.666	Myosin_tail_family_protein,myosin_tail_family_protein,myosin_heavy_chain,_non_muscle
FAM170	4	7	8	2	9	2	2.600	alpha_tubulin,tubulin_alpha_chain
FAM256	0	10	3	1	13	1	2.196	Multidrug_resistance_protein_1
FAM826	0	14	1	0	13	0	2.041	NONE
FAM887	1	8	5	0	13	0	2.384	NONE
FAM174	5	4	1	5	3	5	2.185	amino_acid_permease,protein_kcc,solute_carrier_family_12
FAM1345	14	0	0	4	0	5	2.108	Deoxyribonuclease (DNase) II
FAM366	1	4	11	1	4	1	2.061	17_beta_hydroxysteroid_dehydrogenase_type_6
FAM1503	13	0	0	8	0	1	1.986	hypothetical_protein,Pfam-B_11267_and_gag_pre-integrs_and_rve_and_DUF4219_and_RVT_2_and_Pfam-B_4137_and_zf-CCHC_and_Pfam-B_10563_and_UBN2-domain_containing_protein,zf-CCHC_and_UBN2-domain_containing_protein,pol_polyprotein,Pfam-B_7383_and_zf-CCHC_and_gag_pre-integrs_and_rve_and_Pfam-B_10563_and_UBN2-domain_containing_protein,Pfam-B_7383_and_Pfam-B_10329_and_zf-CCHC_and_UBN2_2-domain_containing_protein,DUF4219_and_Pfam-B_7383_and_zf-CCHC_and_Pfam-B_4137_and_rve_and_Pfam-B_10563_and_UBN2_2_and_RVT_2-domain_containing_protein
FAM1228	8	1	1	5	1	5	2.543	conserved_hypothetical_protein,Pfam-B_310-domain_containing_protein,Pfam-B_2707_and_rve_and_Ribosomal_L50_and_RVT_1-domain_containing_protein,subfamily_M3A_non_peptidase_ue_
FAM1522	1	10	5	0	5	0	2.413	NONE
FAM1515	0	8	3	0	10	0	2.354	NONE
FAM87	2	5	3	2	6	2	2.284	heat_shock_protein_70,78_kDa_glucose_regulated_protein
FAM596	0	10	2	0	8	0	2.221	NONE
FAM367	4	2	2	5	1	5	2.613	solute_carrier_family_12
FAM212	3	3	1	5	2	5	2.185	ABC_tran_domain_containing_protein,ABC_transporter,_ATP_binding_protein,ATP_binding_cassette_sub_family_A
FAM44	4	2	2	4	1	4	2.633	sodium_driven_chloride_bicarbonate_exchanger,anion_exchange_protein,electrogenic_sodium_bicarbonate_cotransporter
FAM144	2	3	4	2	4	2	2.543	32_kDa_beta_galactoside_binding_lectin,galactoside_binding_lectin_family_protein
FAM2306	1	6	3	2	4	1	2.319	tyrosine_protein_kinase_Fps85D,Tyrosine_protein_kinase_Fps85D
FAM345	0	1	7	1	7	1	2.000	histoneo;_histone_h3;_histone_h1
FAM150	3	4	3	2	3	2	1.993	protein_unc_g;_protein_unc_f;_protein_unc_d;_protein_unc_b;_protein_unc_a,Muscle_M_line_assembly_protein_unc_89
FAM346	4	2	3	3	2	3	1.993	Transmembrane_cell_adhesion_receptor_mua_3
FAM2795	5	0	0	7	0	4	2.600	conserved_hypothetical_protein,zf-CCHC_and_RVT_3_and_Pfam-B_695_and_rve-domain_containing_protein,zf-CCHC_and_RVT_3_and_rve-domain_containing_protein,CRAL_TRIO_and_RVT_3_and_rve-domain_containing_protein
FAM1143	1	3	5	2	3	2	2.196	DNA_topoisomerase_2
FAM76	1	5	2	2	5	1	2.148	Tubulin_beta_2C_chain,beta_tubulin
FAM1139	2	3	3	2	3	2	2.739	delta(3'_5')_delta(2,4)_dienoyl_coenzyme_a_isomerase,hydroxysteroid_dehydrogenase_protein_2_like

FAM2840	1	3	4	1	5	1	2.556	kunitz:Bovine_pancreatic_trypsin
FAM114	1	4	3	2	3	2	2.384	metabotropic_glutamate_receptor_7,glutamate_receptor,_metabotropic_5
FAM956	2	3	4	1	3	2	2.384	SWI:SNF_matrix_associated
FAM1394	3	1	2	3	2	4	2.384	Patched_and_Pfam-B_11358-domain_containing_protein,patched_family_protein
FAM2901	0	8	1	0	6	0	2.105	NONE
FAM3606	4	1	0	4	1	4	2.685	Pfam-B_10329_and_Asp_protease_2_and_RVT_1_and_Pfam-B_2707-domain_containing_protein,Pfam-B_10329_and_zf-CCHC_and_RVT_1_and_Pfam-B_2707-domain_containing_protein,Pfam-B_10329_and_zf-CCHC-domain_containing_protein,polyprotein
FAM306	1	4	3	1	4	1	2.657	protein_UNC_32,_a
FAM1533	1	3	4	2	3	1	2.477	solute_carrier_family_25_4,39S_ribosomal_protein_L13
FAM233	1	4	2	1	5	1	2.284	arginine_kinase
FAM365	1	2	4	1	5	1	2.284	protein_CLP_1,_d
FAM2568	0	7	3	1	3	0	2.257	Dehydrogenase:reductase_SDR_family
FAM3273	0	2	4	0	8	0	2.185	NONE
FAM1677	1	2	5	2	3	1	1.993	structural_maintenance_of_chromosomes_protein_3,NADH_dehydrogenase_ubiquinone_Fe_S_protein_2
FAM63	1	4	3	1	3	1	2.633	VAB_10A_protein
FAM2789	1	3	3	1	4	1	2.633	Lactamase_B-domain_containing_protein
FAM95	1	3	3	1	3	2	2.543	Troponin_and_Pfam-B_969-domain_containing_protein
FAM897	1	2	4	1	4	1	2.378	protein_spinster_1
FAM1036	1	4	4	1	2	1	2.378	NDT80_:_PhoGDNA_binding_family_protein
FAM1684	1	2	3	1	5	1	2.185	sideroflexin_1
FAM3757	0	6	1	0	6	0	2.171	NONE
FAM2739	3	2	1	4	1	2	2.138	N_acetyltransferase_10,Trehalase_family_protein
FAM3667	1	2	4	1	3	2	2.138	conserved_hypothetical_protein
FAM3685	0	3	2	0	7	1	2.084	NONE
FAM3669	0	2	3	2	5	1	2.032	DUF290-domain_containing_protein

Supplementary Table 3b: Shared gene families showing significant changes (z-score > 1.96) in copy number between *Trichuris muris* and other nematodes, sorted by absolute z-score (normalised difference in gene copy number).

family	*Trichinella spiralis*	*Bursaphelenchus xylophilus*	*Brugia malayi*	*T. muris*	*C. elegans*	*T. trichiura*	z-score	annotation
FAM139	1	1	1	2	1	1	2.041	HMG_box_family_protein,protein_pangolin,s_A:H:I
FAM502	1	1	1	2	1	1	2.041	DnaJ_protein_subfamily_B_B,DnaJ_domain_containing_protein
FAM1093	1	1	1	2	1	1	2.041	TPR_Domain_containing_protein,tetratricopeptide_repeat_containing

FAM1453	1	1	1	2	1	1	2.041	Glyco_hydro_38_and_Alpha-mann_mid_and_Glyco_hydro_38C-domain_containing_protein,alpha_mannosidase_2x
FAM1626	1	1	1	2	1	1	2.041	zinc_transporter_zip1,metal_cation_transporter,_zinc
FAM1752	1	1	1	2	1	1	2.041	UQ_con-domain_containing_protein
FAM1800	1	1	1	2	1	1	2.041	conserved_hypothetical_protein,DNA_polymerase_lambda
FAM1809	1	1	1	2	1	1	2.041	conserved_hypothetical_protein,Meckel_syndrome_type_1_protein
FAM2357	1	1	1	2	1	1	2.041	NADH_dehydrogenase_(ubiquinone)_iron_sulfur
FAM2751	1	1	1	2	1	1	2.041	28S_ribosomal_protein_S28,_mitochondrial,Mesd-domain_containing_protein
FAM2960	1	1	1	2	1	1	2.041	rho_associated_protein_kinase_2,Pkinase_domain_containing_protein
FAM3155	1	1	1	2	1	1	2.041	NifU_N-domain_containing_protein
FAM3560	1	1	1	2	1	1	2.041	N_alpha_acetyltransferase_15,_NatA_auxiliary
FAM3577	1	1	1	2	1	1	2.041	structural_maintenance_of_chromosomes_protein
FAM3813	1	1	1	2	1	1	2.041	AAA-domain_containing_protein,vesicle_fusing_ATPase_1
FAM3914	1	1	1	2	1	1	2.041	26S_proteasome_non_ATPase_regulatory_subunit_1,26S_proteasome_non_ATPase_regulatory_subunit
FAM3930	1	1	1	2	1	1	2.041	PGM_PMM_II_and_PGM_PMM_III_and_PGM_PMM_I_and_PGM_PMM_IV-domain_containing_protein,glucose_1,6_bisphosphate_synthase
FAM5175	1	1	1	2	1	1	2.041	T_complex_protein_1_theta_subunit,T_complex_protein_1_subunit_theta
FAM5248	1	1	1	2	1	1	2.041	splicing_factor,_arginine:serine_rich
FAM5276	1	1	1	2	1	1	2.041	DNA_polymerase_eta
FAM5676	1	1	1	2	1	1	2.041	conserved_hypothetical_protein,Ank_2_and_Ank-domain_containing_protein
FAM7549	1	1	1	2	1	1	2.041	macrophage_migration_inhibitory_factor
FAM4133	2	2	2	1	2	2	-2.041	thioredoxin_protein_4A_like
FAM217	0	1	1	47	1	0	2.041	conserved_hypothetical_protein,hypothetical_protein
FAM1495	0	1	0	18	0	0	2.038	hypothetical_protein,conserved_hypothetical_protein,jerky_protein
FAM2218	0	0	0	17	0	1	2.038	conserved_hypothetical_protein,Pao_retrotransposon_peptidase_superfamily,zinc_knuckle_protein,Gag_Pol_polyprotein
FAM2502	0	0	0	16	0	1	2.037	hypothetical_protein,conserved_hypothetical_protein
FAM7548	0	0	0	7	0	1	2.020	actin_protein_6A_like,Actin-domain_containing_protein
FAM6042	1	0	0	8	0	1	2.016	conserved_hypothetical_protein,zf-C2H2_4_and_zf-C2H2-domain_containing_protein,zinc_finger_protein,zinc_finger_protein_569,zf-C2H2_4_and_zf-C2H2_and_zf-H2C2_2-domain_containing_protein,zinc_finger_protein_729,pita,_B,zf-H2C2_2_and_zf-C2H2_4_and_zf-C2H2-domain_containing_protein
FAM8363	0	0	0	6	0	1	2.013	hypothetical_protein
FAM8370	0	0	0	6	0	1	2.013	CUG-BP-and_ETR-3-like_factor_3,hypothetical_protein,Calcium_channel_protein,_putative
FAM2217	0	0	0	15	0	3	2.000	transposase,conserved_hypothetical_protein,Putative_tick_transposon,Pfam-B_19879-domain_containing_protein,hypothetical_protein,reverse_transcriptase
FAM9301	0	0	0	5	0	1	2.000	Pfam-B_10329-domain_containing_protein,reverse_transcriptase_family_protein
FAM9306	0	0	0	5	0	1	2.000	conserved_hypothetical_protein,hypothetical_protein

FAM9307	0	0	0	5	0	1	2.000	conserved_hypothetical_protein
FAM7545	1	0	0	6	0	1	1.996	reverse_transcriptase:endonuclease,reverse_transcriptase,Pfam-B_595-domain_containing_protein
FAM5503	1	1	0	6	0	1	1.993	zf-CCHC_and_Pfam-B_2545_and_Asp_protease_2_and_Pfam-B_2707_and_rve-domain_containing_protein,Gap_Pol_polyprotein,rve_and_Pfam-B_10563_and_zf-CCHC_and_Pfam-B_2545_and_Asp_protease_2_and_RVT_1_and_Pfam-B_2707-domain_containing_protein,rve_and_zf-CCHC_and_Pfam-B_10329_and_RVT_1_and_Pfam-B_2707-domain_containing_protein,zf-CCHC_and_Asp_protease_2_and_Pfam-B_10329_and_RVT_1-domain_containing_protein,zf-H2C2_and_rve_and_Pfam-B_10329_and_zf-CCHC_and_Pfam-B_2545_and_RVP_and_RVT_1_and_Pfam-B_2707-domain_containing_protein
FAM5304	0	0	0	9	0	2	1.990	conserved_hypothetical_protein,Pfam-B_19879-domain_containing_protein,ankyrin_2,3:unc44,reverse_transcriptase
FAM6039	2	0	0	8	0	0	1.977	Pfam-B_310-domain_containing_protein,something_about_silencing_protein_10
FAM10557	0	0	0	4	0	1	1.977	CAP_domain_containing_protein,CAP-domain_containing_protein,Pfam-B_19232_and_CAP-domain_containing_protein
FAM10561	0	0	0	4	0	1	1.977	gag_pol_polyprotein,hypothetical_protein

Supplementary Table 3c: Shared gene families showing significant changes (z-score > 1.96) in copy number between *Trichuris trichiura* and other nematodes, sorted by absolute z-score (normalised difference in gene copy number).

family	*Trichinella spiralis*	*Bursaphelenchus xylophilus*	*Brugia malayi*	*T. muris*	*C. elegans*	*T. trichiura*	z-score	annotation
FAM202	1	1	1	1	1	3	2.041	valyl_tRNA_synthetase
FAM1105	1	1	1	1	1	3	2.041	Pfam-B_888-domain_containing_protein
FAM2967	1	1	1	1	1	3	2.041	glutamate_synthase
FAM4025	1	1	1	1	1	3	2.041	ATP_synthase_subunit_alpha,_mitochondrial
FAM4497	1	1	1	1	1	3	2.041	ADP_ribosylation_factor_protein_2_like
FAM5073	1	1	1	1	1	3	2.041	WD_repeat_containing_protein_54
FAM5285	1	1	1	1	1	3	2.041	Pfam-B_5851_and_DUF3677-domain_containing_protein
FAM79	1	1	1	1	1	2	2.041	transcription_factor_E2_alpha
FAM724	1	1	1	1	1	2	2.041	Leucine_rich_repeat_and_calponiny
FAM915	1	1	1	1	1	2	2.041	microtubule_associated_serine:threonine_protein
FAM1450	1	1	1	1	1	2	2.041	protein_slowmo
FAM1755	1	1	1	1	1	2	2.041	phosphorylase_b_kinase_gamma_catalytic_chain
FAM1923	1	1	1	1	1	2	2.041	SYS1_Golgi_localized_integral_membrane_protein
FAM2403	1	1	1	1	1	2	2.041	Pkinase_and_TRAPP-domain_containing_protein
FAM2409	1	1	1	1	1	2	2.041	40S_ribosomal_protein_S15
FAM2480	1	1	1	1	1	2	2.041	WD40_and_CARD_and_NB-ARC_and_Pfam-B_10185-domain_containing_protein
FAM2757	1	1	1	1	1	2	2.041	mitochondrial_tRNA_specific_2_thiouridylase

FAM3201	1	1	1	1	1	2	2.041	2OG-FeII_Oxy_3_and_Ofd1_CTDD_and_Pfam-B_9093-domain_containing_protein
FAM3383	1	1	1	1	1	2	2.041	protein_yippee_5_like
FAM3421	1	1	1	1	1	2	2.041	polypeptide_n_acetylgalactosaminyltransferase_3
FAM3430	1	1	1	1	1	2	2.041	28S_ribosomal_protein_S7,_mitochondrial
FAM3783	1	1	1	1	1	2	2.041	serine:threonine_protein_phosphatase_PGAM5
FAM3803	1	1	1	1	1	2	2.041	MAP_kinase_activating_protein_C22orf5
FAM3836	1	1	1	1	1	2	2.041	ATP_synthase_subunit_beta
FAM3869	1	1	1	1	1	2	2.041	60S_ribosomal_protein_L37a
FAM3880	1	1	1	1	1	2	2.041	ribosomal_protein_L7
FAM3942	1	1	1	1	1	2	2.041	glucose_6_phosphate_isomerase
FAM3967	1	1	1	1	1	2	2.041	RNA_exonuclease_1
FAM4382	1	1	1	1	1	2	2.041	dolichyl_pyrophosphate_Man9GlcNAc2
FAM4412	1	1	1	1	1	2	2.041	tetratricopeptide_repeat_protein_35_B
FAM4530	1	1	1	1	1	2	2.041	mRNA_capping_enzyme
FAM4544	1	1	1	1	1	2	2.041	leucyl_tRNA_synthetase,_cytoplasmic
FAM4594	1	1	1	1	1	2	2.041	methyltransferase_protein_4_like
FAM4937	1	1	1	1	1	2	2.041	acetyl_coenzyme_A_transporter_1
FAM5283	1	1	1	1	1	2	2.041	HCO3_transporter_family_protein
FAM5711	1	1	1	1	1	2	2.041	Exo_endo_phos_and_zf-RING_2-domain_containing_protein
FAM5894	1	1	1	1	1	2	2.041	Apoptosis_linked_gene_2_interacting_protein_X_1
FAM6002	1	1	1	1	1	2	2.041	histone_deacetylase_complex_subunit_SAP18
FAM6953	1	1	1	1	1	2	2.041	Brix-domain_containing_protein
FAM7748	1	1	1	1	1	2	2.041	monocarboxylate_transporter_3
FAM8613	1	1	1	1	1	2	2.041	Pfam-B_1741-domain_containing_protein
FAM1860	2	2	2	2	2	1	-2.041	TM2-domain_containing_protein,mitochondrial_import_inner_membrane_translocase
FAM1966	2	2	2	2	2	1	-2.041	beta_lactamase_protein_2_like,L_aminoadipate_semialdehyde
FAM4665	0	0	0	1	0	11	2.033	neurogenic_locus_notch_protein_2
FAM4666	0	0	1	0	0	11	2.033	NONE
FAM3206	0	0	1	0	1	11	2.028	NONE
FAM8398	0	0	0	1	0	6	2.013	hypothetical_protein
FAM8401	1	0	0	0	0	6	2.013	NONE
FAM1027	1	0	0	4	0	21	2.005	hypothetical_protein,Pfam-B_6031_and_Pfam-B_2655-domain_containing_protein
FAM9344	0	0	0	1	0	5	2.000	Pfam-B_9093-domain_containing_protein

Supplementary Table 3d: Gene families missing from *Trichuris muris* showing significant changes (z-score > 1.96) in copy number across nematodes, sorted by total copy number (NB: all significant families are conserved single-copy genes in other nematode species).

family	*Trichinella spiralis*	*Bursaphelenchus xylophilus*	*Brugia malayi*	*T. muris*	*C. elegans*	*T. trichiura*	z-score	annotation
FAM319	1	1	1	0	1	1	-2.041	NONE
FAM443	1	1	1	0	1	1	-2.041	NONE
FAM677	1	1	1	0	1	1	-2.041	NONE
FAM736	1	1	1	0	1	1	-2.041	NONE
FAM1460	1	1	1	0	1	1	-2.041	NONE
FAM1601	1	1	1	0	1	1	-2.041	NONE
FAM2390	1	1	1	0	1	1	-2.041	NONE
FAM2392	1	1	1	0	1	1	-2.041	NONE
FAM2454	1	1	1	0	1	1	-2.041	NONE
FAM2456	1	1	1	0	1	1	-2.041	NONE
FAM2641	1	1	1	0	1	1	-2.041	NONE
FAM2713	1	1	1	0	1	1	-2.041	NONE
FAM2955	1	1	1	0	1	1	-2.041	NONE
FAM3007	1	1	1	0	1	1	-2.041	NONE
FAM3016	1	1	1	0	1	1	-2.041	NONE
FAM3046	1	1	1	0	1	1	-2.041	NONE
FAM3113	1	1	1	0	1	1	-2.041	NONE
FAM3394	1	1	1	0	1	1	-2.041	NONE
FAM3789	1	1	1	0	1	1	-2.041	NONE
FAM3812	1	1	1	0	1	1	-2.041	NONE
FAM3872	1	1	1	0	1	1	-2.041	NONE
FAM3886	1	1	1	0	1	1	-2.041	NONE
FAM4081	1	1	1	0	1	1	-2.041	NONE
FAM4334	1	1	1	0	1	1	-2.041	NONE
FAM4410	1	1	1	0	1	1	-2.041	NONE
FAM4512	1	1	1	0	1	1	-2.041	NONE
FAM4526	1	1	1	0	1	1	-2.041	NONE

FAM4536	1	1	1	0	1	1	-2.041	NONE
FAM4906	1	1	1	0	1	1	-2.041	NONE
FAM4997	1	1	1	0	1	1	-2.041	NONE
FAM5003	1	1	1	0	1	1	-2.041	NONE
FAM5219	1	1	1	0	1	1	-2.041	NONE
FAM5260	1	1	1	0	1	1	-2.041	NONE
FAM5669	1	1	1	0	1	1	-2.041	NONE
FAM5862	1	1	1	0	1	1	-2.041	NONE
FAM5879	1	1	1	0	1	1	-2.041	NONE
FAM5931	1	1	1	0	1	1	-2.041	NONE
FAM5953	1	1	1	0	1	1	-2.041	NONE
FAM6419	1	1	1	0	1	1	-2.041	NONE
FAM6473	1	1	1	0	1	1	-2.041	NONE
FAM6602	1	1	1	0	1	1	-2.041	NONE
FAM6705	1	1	1	0	1	1	-2.041	NONE
FAM6745	1	1	1	0	1	1	-2.041	NONE
FAM7085	1	1	1	0	1	1	-2.041	NONE
FAM7089	1	1	1	0	1	1	-2.041	NONE
FAM7156	1	1	1	0	1	1	-2.041	NONE
FAM7191	1	1	1	0	1	1	-2.041	NONE
FAM7249	1	1	1	0	1	1	-2.041	NONE
FAM7371	1	1	1	0	1	1	-2.041	NONE
FAM7414	1	1	1	0	1	1	-2.041	NONE

Supplementary Table 3e: Gene families missing from *Trichuris trichiura* showing significant changes (z-score > 1.96) in copy number across nematodes, sorted by total copy number (NB: all significant families are conserved single-copy genes in other nematode species).

family	*Trichinella spiralis*	*Bursaphelenchus xylophilus*	*Brugia malayi*	*T. muris*	*C. elegans*	*T. trichiura*	z-score	annotation
FAM93	1	1	1	1	1	0	-2.041	EGF_CA-domain_containing_protein
FAM194	1	1	1	1	1	0	-2.041	Tweety-domain_containing_protein
FAM441	1	1	1	1	1	0	-2.041	solute_carrier_organic_anion_transporter_family
FAM499	1	1	1	1	1	0	-2.041	scaffold_attachment_factor_B1
FAM730	1	1	1	1	1	0	-2.041	uDENN_and_DENN-domain_containing_protein

FAM1171	1	1	1	1	1	0	−2.041	E3_ubiquitin_protein_ligase_makorin_1
FAM1301	1	1	1	1	1	0	−2.041	Synaptosomal_associated_protein_25
FAM1724	1	1	1	1	1	0	−2.041	calnexin
FAM2673	1	1	1	1	1	0	−2.041	dNK-domain_containing_protein
FAM2758	1	1	1	1	1	0	−2.041	polymerase_(RNA)_II_(DNA_directed)_polypeptide
FAM2923	1	1	1	1	1	0	−2.041	pyroglutamyl_peptidase_1
FAM2990	1	1	1	1	1	0	−2.041	NADH_ubiquinone_oxidoreductase_15_kDa_subunit
FAM3108	1	1	1	1	1	0	−2.041	thiosulfate_sulfurtransferase
FAM3410	1	1	1	1	1	0	−2.041	Dihydropteridine_reductase
FAM3483	1	1	1	1	1	0	−2.041	splicing_factor_45
FAM3524	1	1	1	1	1	0	−2.041	biogenesis_of_lysosome_organelles
FAM3794	1	1	1	1	1	0	−2.041	Activator_of_basal_transcription_1
FAM3800	1	1	1	1	1	0	−2.041	DUF2615-domain_containing_protein
FAM3864	1	1	1	1	1	0	−2.041	vATP-synt_E_and_Urm1-domain_containing_protein
FAM4071	1	1	1	1	1	0	−2.041	Pfam-B_5334-domain_containing_protein
FAM4482	1	1	1	1	1	0	−2.041	peptidyl_prolyl_cis_trans_isomerase
FAM4545	1	1	1	1	1	0	−2.041	fibroblast_growth_factor
FAM4943	1	1	1	1	1	0	−2.041	NEDD8_activating_enzyme_E1_regulatory_subunit
FAM5084	1	1	1	1	1	0	−2.041	DnaJ-domain_containing_protein
FAM5806	1	1	1	1	1	0	−2.041	Pterin_4a_domain_containing_protein
FAM5897	1	1	1	1	1	0	−2.041	cytochrome_c
FAM6400	1	1	1	1	1	0	−2.041	Pfam-B_12879_and_DUF4187_and_G-patch-domain_containing_protein
FAM7128	1	1	1	1	1	0	−2.041	prefoldin_subunit_2
FAM7166	1	1	1	1	1	0	−2.041	FAS_associated_factor_2_B
FAM7222	1	1	1	1	1	0	−2.041	poly_(ADP_ribose)_polymerase_1
FAM7447	1	1	1	1	1	0	−2.041	TAP42-domain_containing_protein
FAM7473	1	1	1	1	1	0	−2.041	Prefoldin_2-domain_containing_protein
FAM7480	1	1	1	1	1	0	−2.041	tRNA_pseudouridine_synthase_2
FAM7483	1	1	1	1	1	0	−2.041	XPA_C-domain_containing_protein
FAM8017	1	1	1	1	1	0	−2.041	conserved_hypothetical_protein
FAM8163	1	1	1	1	1	0	−2.041	heat_shock_70_kDa_protein_13
FAM8601	1	1	1	1	1	0	−2.041	proteasome_subunit_beta_type_1
FAM8606	1	1	1	1	1	0	−2.041	pre_mrna_splicing_factor_spf27

FAM8628	1	1	1	1	1	0	−2.041	nitrogen_permease_regulator_2_protein_like
FAM9771	1	1	1	1	1	0	−2.041	ras_family_small_GTPase
FAM9812	1	1	1	1	1	0	−2.041	L_threonine_3_dehydrogenase,_mitochondrial
FAM11057	1	1	1	1	1	0	−2.041	39S_ribosomal_protein_L20,_mitochondrial
FAM11058	1	1	1	1	1	0	−2.041	I−set−domain_containing_protein
FAM11079	1	1	1	1	1	0	−2.041	Methyltransf_11−domain_containing_protein
FAM11122	1	1	1	1	1	0	−2.041	Leucine_rich_repeat_containing_protein_15
FAM11132	1	1	1	1	1	0	−2.041	Pfam−B_1741−domain_containing_protein
FAM11138	1	1	1	1	1	0	−2.041	polypeptide_n_acetylgalactosaminyltransferase

Supplementary Table 3f: Largest gene families in *Trichuris muris*, sorted by *T. muris* copy number.

family	*Trichinella spiralis*	*Bursaphelenchus xylophilus*	*Brugia malayi*	*T. muris*	*C. elegans*	*T. trichiura*	annotation
FAM217	0	1	1	47	1	0	conserved_hypothetical_protein,hypothetical_protein
FAM806	0	0	0	29	0	0	hypothetical_protein,Pfam−B_7026−domain_containing_protein,Pfam−B_19346−domain_containing_protein,Pfam−B_7026_and_Pfam−B_16276−domain_containing_protein, domain_containing_protein
FAM1124	0	0	0	25	0	0	conserved_hypothetical_protein,hypothetical_protein
FAM2	86	2	793	24	1	14	DUF1759_and_DUF1758−domain_containing_protein,Peptidase_A17_and_rve_and_DUF1759_and_DUF1758−domain_containing_protein,Pao_retrotransposon_peptidase_sup domain_containing_protein,rve_and_DUF1758_and_RVT_1_and_Peptidase_A17−domain_containing_protein,DUF1758_and_RVT_1_and_Peptidase_A17−domain_containing_p domain_containing_protein,RVT_1_and_Peptidase_A17_and_DUF1758−domain_containing_protein,Peptidase_A17_and_DUF1759_and_DUF1758−domain_containing_protein domain_containing_protein,Tas_retrotransposon_peptidase_A16_superfamily,rve_and_DUF1758_and_Peptidase_A17−domain_containing_protein,DUF1759_and_DUF1758_a domain_containing_protein,DUF1759_and_DUF1758_and_Peptidase_A17−domain_containing_protein
FAM1344	0	0	0	23	0	0	RNA_dependent_RNA_polymerase,Mononeg_RNA_pol−domain_containing_protein,large_protein,hypothetical_protein,polymerase
FAM1988	0	0	0	19	0	0	conserved_hypothetical_protein,tigger_transposable_element_derived_protein,hypothetical_protein
FAM1495	0	1	0	18	0	0	hypothetical_protein,conserved_hypothetical_protein,jerky_protein
FAM2219	0	0	0	18	0	0	hypothetical_protein,conserved_hypothetical_protein
FAM2218	0	0	0	17	0	1	conserved_hypothetical_protein,Pao_retrotransposon_peptidase_superfamily,zinc_knuckle_protein,Gag_Pol_polyprotein
FAM2502	0	0	0	16	0	1	hypothetical_protein,conserved_hypothetical_protein
FAM2217	0	0	0	15	0	3	transposase,conserved_hypothetical_protein,Putative_tick_transposon,Pfam−B_19879−domain_containing_protein,hypothetical_protein,reverse_transcriptase
FAM3202	0	0	0	15	0	0	tigger_transposable_element_derived_protein
FAM8	269	3	0	14	0	1	FLYWCH_and_MULE_and_Pfam−B_516−domain_containing_protein,Pfam−B_516_and_FLYWCH_and_MULE−domain_containing_protein,MULE_and_FLYWCH_and_Pfam−B_516− domain_containing_protein,FLYWCH_and_Pfam−B_516_and_MULE−domain_containing_protein
FAM3605	0	0	0	14	0	0	conserved_hypothetical_protein,Pfam−B_516−domain_containing_protein,Pfam−B_516_and_MULE_and_FLYWCH−domain_containing_protein
FAM3607	0	0	0	14	0	0	conserved_hypothetical_protein,tigger_transposable_element_derived_protein,jerky_protein,hypothetical_protein
FAM3608	0	0	0	14	0	0	hypothetical_protein,conserved_hypothetical_protein,Pfam−B_19879−domain_containing_protein

FAM3653	0	0	0	14	0	0	hypothetical_protein
FAM4089	0	0	0	13	0	0	ankyrin_2,3:unc44,DUF1758-domain_containing_protein,polyprotein,conserved_hypothetical_protein,Peptidase_A17_and_DUF667-domain_containing_protein,Pfam-B_198
FAM4090	0	0	0	13	0	0	hypothetical_protein,conserved_hypothetical_protein,Pfam-B_13107-domain_containing_protein,Pfam-B_516-domain_containing_protein,FLYWCH-domain_containing_pro
FAM4659	0	0	0	12	0	0	conserved_hypothetical_protein,hypothetical_protein
FAM4660	0	0	0	12	0	0	Putative_integrase_core_domain_protein,conserved_hypothetical_protein,zf-CCHC-domain_containing_protein,zf-CCHC_and_Pfam-B_10329-domain_containing_protein
FAM5306	0	0	0	11	0	0	hypothetical_protein,Pfam-B_10329-domain_containing_protein,Pfam-B_10329_and_zf-CCHC-domain_containing_protein,zf-CCHC_and_Pfam-B_10329-domain_containir
FAM5309	0	0	0	11	0	0	hypothetical_protein,conserved_hypothetical_protein
FAM6036	0	0	0	10	0	0	hypothetical_protein,conserved_hypothetical_protein
FAM6037	0	0	0	10	0	0	hypothetical_protein,conserved_hypothetical_protein,peripheral_plasma_membrane_protein_CASK
FAM6044	0	0	0	10	0	0	hypothetical_protein
FAM36	91	1	1	9	1	5	conserved_hypothetical_protein,DUF4371-domain_containing_protein
FAM5304	0	0	0	9	0	2	ankyrin_2,3:unc44,conserved_hypothetical_protein,Pfam-B_19879-domain_containing_protein,reverse_transcriptase
FAM78	61	1	1	8	2	7	protein_tag_76,PAZ_and_Piwi-domain_containing_protein
FAM1503	13	0	0	8	0	1	pol_polyprotein,DUF4219_and_Pfam-B_7383_and_zf-CCHC_and_Pfam-B_4137_and_rve_and_Pfam-B_10563_and_UBN2_2_and_RVT_2-domain_containing_protein,hypothet B_4137_and_zf-CCHC_and_Pfam-B_10563_and_UBN2-domain_containing_protein,zf-CCHC_and_UBN2-domain_containing_protein,Pfam-B_7383_and_zf-CCHC_and_gag_p B_7383_and_Pfam-B_10329_and_zf-CCHC_and_UBN2_2-domain_containing_protein
FAM5302	0	0	0	8	0	3	conserved_hypothetical_protein,DDE_Tnp_IS1595_and_Mononeg_RNA_pol-domain_containing_protein,hypothetical_protein,Pfam-B_19346-domain_containing_protein
FAM6039	2	0	0	8	0	0	Pfam-B_310-domain_containing_protein,something_about_silencing_protein_10
FAM6042	1	0	0	8	0	1	conserved_hypothetical_protein,zf-C2H2_4_and_zf-C2H2-domain_containing_protein,zinc_finger_protein,zinc_finger_protein_569,zf-C2H2_4_and_zf-C2H2_and_zf-H2C2_ C2H2-domain_containing_protein
FAM7543	0	0	0	8	0	0	conserved_hypothetical_protein
FAM7544	0	0	0	8	0	0	Pfam-B_10329-domain_containing_protein,RETRotransposon_family_member
FAM7546	0	0	0	8	0	0	Dimer_Tnp_hAT-domain_containing_protein,conserved_hypothetical_protein
FAM7700	0	0	0	8	0	0	hypothetical_protein
FAM2795	5	0	0	7	0	4	conserved_hypothetical_protein,zf-CCHC_and_RVT_3_and_rve-domain_containing_protein,zf-CCHC_and_RVT_3_and_Pfam-B_695_and_rve-domain_containing_protein,CRA
FAM4661	0	0	0	7	0	5	hypothetical_protein,dsrm-domain_containing_protein
FAM6040	2	0	0	7	0	1	gag_pre-integrs_and_rve_and_Pfam-B_10563_and_zf-CCHC_and_Pfam-B_11267-domain_containing_protein,zf-CCHC_and_Pfam-B_11267_and_UBN2-domain_containing_protein,putative_pol_polyprotein;_copia_type_polyprotein_putative,retrovirus_Pol_polyprotein,polyprotein,copia_type_polyprotein,reverse_transcriptase_
FAM7548	0	0	0	7	0	1	actin_protein_6A_like,Actin-domain_containing_protein
FAM8361	0	0	0	7	0	0	conserved_hypothetical_protein,zf-CCHC-domain_containing_protein
FAM8364	0	0	0	7	0	0	protein_crumbs,neurogenic_locus_notch_protein_2,Uncharacterized_transposase_protein,crumbs_1,neurogenic_locus_notch_protein
FAM8365	0	0	0	7	0	0	conserved_hypothetical_protein,zinc_finger_protein_646,hypothetical_protein
FAM8369	0	0	0	7	0	0	conserved_hypothetical_protein,Pfam-B_516-domain_containing_protein
FAM34	4	6	4	6	8	6	multidrug_resistance_associated_protein_4,multidrug_Resistance_protein_family_member,multidrug_resistance_associated_protein_7,multidrug_resistanceprotein_k;_multic
FAM131	52	0	0	6	0	5	neurogenic_locus_notch_protein_2,hypothetical_protein,Neurogenic_locus_protein_delta,Putative_thrombospondin_type_1_domain_protein,EGF-domain_containing_protein
FAM234	4	2	1	6	5	5	glyco_protein_3_alpha_L_fucosyltransferase_A,glycoprotein_3_alpha_L_fucosyltransferase,Alpha_(1,3)_fucosyltransferase_C,glycoprotein_3_alpha_l_fucosyltransferase

FAM2794	0	0	0	6	0	10	Motile_Sperm_and_Pfam-B_11622-domain_containing_protein,Motile_Sperm-domain_containing_protein,Pfam-B_4364_and_Motile_Sperm-domain_containing_protein
FAM5503	1	1	0	6	0	1	zf-CCHC_and_Pfam-B_2545_and_Asp_protease_2_and_Pfam-B_2707_and_rve-domain_containing_protein,Gap_Pol_polyprotein,rve_and_Pfam-B_10563_and_zf-CCHC_and_ domain_containing_protein,rve_and_zf-CCHC_and_Pfam-B_10329_and_RVT_1_and_Pfam-B_2707-domain_containing_protein,zf-CCHC_and_Asp_protease_2_and_Pfam-B_ CCHC_and_Pfam-B_2545_and_RVP_and_RVT_1_and_Pfam-B_2707-domain_containing_protein

Supplementary Table 3g: Largest gene families in *Trichuris trichiura*, sorted by *T. trichiura* copy number.

family	*Trichinella spiralis*	*Bursaphelenchus xylophilus*	*Brugia malayi*	*T. muris*	*C. elegans*	*T. trichiura*	annotation
FAM1027	1	0	0	4	0	21	hypothetical_protein,Pfam-B_6031_and_Pfam-B_2655-domain_containing_protein
FAM2	86	2	793	24	1	14	DUF1759_and_DUF1758-domain_containing_protein,Peptidase_A17_and_rve_and_DUF1759_and_DUF1758-domain_containing_protein,Pao_retrotransposon_peptidase_su domain_containing_protein,rve_and_DUF1758_and_RVT_1_and_Peptidase_A17-domain_containing_protein,DUF1758_and_RVT_1_and_Peptidase_A17-domain_containing_ domain_containing_protein,RVT_1_and_Peptidase_A17_and_DUF1758-domain_containing_protein,Peptidase_A17_and_DUF1759_and_DUF1758-domain_containing_protei domain_containing_protein,Tas_retrotransposon_peptidase_A16_superfamily,rve_and_DUF1758_and_Peptidase_A17-domain_containing_protein,DUF1759_and_DUF1758_ domain_containing_protein,DUF1759_and_DUF1758_and_RVT_1_and_Peptidase_A17-domain_containing_protein
FAM4091	0	0	0	0	0	13	NONE
FAM24	3	1	17	5	7	11	proteinase_inhibitor_I4_serpin,Serine_proteinase_inhibitor,serine_proteinase_inhibitor
FAM3206	0	0	1	0	1	11	NONE
FAM4665	0	0	0	1	0	11	neurogenic_locus_notch_protein_2
FAM4666	0	0	1	0	0	11	NONE
FAM5313	0	0	0	0	0	11	NONE
FAM29	9	0	107	0	0	10	NONE
FAM587	18	0	4	0	0	10	NONE
FAM2794	0	0	0	6	0	10	Motile_Sperm_and_Pfam-B_11622-domain_containing_protein,Motile_Sperm-domain_containing_protein,Pfam-B_4364_and_Motile_Sperm-domain_containing_protein
FAM6050	0	0	0	0	0	10	NONE
FAM126	0	4	2	4	15	9	histone_type;_histone_hc,Histone-domain_containing_protein
FAM5305	0	0	0	3	0	8	histone_h3
FAM78	61	1	1	8	2	7	protein_tag_76,PAZ_and_Piwi-domain_containing_protein
FAM2393	1	1	1	4	1	7	DUF229-domain_containing_protein
FAM5303	1	0	0	3	0	7	Trypsin-domain_containing_protein
FAM5307	0	0	0	4	0	7	hypothetical_protein
FAM8399	0	0	0	0	0	7	NONE
FAM34	4	6	4	6	8	6	multidrug_resistance_associated_protein_4,multidrug_Resistance_protein_family_member,multidrug_resistance_associated_protein_7,multidrug_resistanceprotein_k;_mult
FAM947	20	0	0	1	0	6	Pfam-B_10329-domain_containing_protein
FAM3750	1	2	0	1	3	6	onchocystatin
FAM6828	1	0	0	2	0	6	Kringle-domain_containing_protein,coagulin_factor_II

FAM8398	0	0	0	1	0	6	hypothetical_protein
FAM8401	1	0	0	0	0	6	NONE
FAM9342	0	0	0	0	0	6	NONE
FAM9345	0	0	0	0	0	6	NONE
FAM9394	0	0	0	0	0	6	NONE
FAM36	91	1	1	9	1	5	conserved_hypothetical_protein,DUF4371-domain_containing_protein
FAM67	0	8	3	0	13	5	NONE
FAM71	1	5	5	4	17	5	histone_H2A,histone_H2A_type_2_B
FAM131	52	0	0	6	0	5	hypothetical_protein,Neurogenic_locus_protein_delta,Putative_thrombospondin_type_1_domain_protein,neurogenic_locus_notch_protein_2,EGF-domain_containing_protei
FAM174	5	4	1	5	3	5	amino_acid_permease,protein_kcc,solute_carrier_family_12
FAM212	3	3	1	5	2	5	ABC_tran_domain_containing_protein,ABC_transporter,_ATP_binding_protein,ATP_binding_cassette_sub_family_A
FAM234	4	2	1	6	5	5	glyco_protein_3_alpha_L_fucosyltransferase_A,glycoprotein_3_alpha_L_fucosyltransferase,Alpha_(1,3)_fucosyltransferase_C,glycoprotein_3_alpha_l_fucosyltransferase
FAM367	4	2	2	5	1	5	solute_carrier_family_12
FAM1228	8	1	1	5	1	5	Pfam-B_310-domain_containing_protein,Pfam-B_2707_and_rve_and_Ribosomal_L50_and_RVT_1-domain_containing_protein,subfamily_M3A_non_peptidase_ue_,conserve
FAM1345	14	0	0	4	0	5	Deoxyribonuclease
FAM1642	14	0	0	2	0	5	hypothetical_protein
FAM4279	0	2	0	2	3	5	UDPGT-domain_containing_protein
FAM4661	0	0	0	7	0	5	hypothetical_protein,dsrm-domain_containing_protein
FAM6830	0	0	0	4	0	5	transmembrane_serine_protease_8,BTB_and_Trypsin-domain_containing_protein,Trypsin-domain_containing_protein
FAM9344	0	0	0	1	0	5	Pfam-B_9093-domain_containing_protein
FAM10600	0	0	0	0	0	5	NONE
FAM10604	0	0	0	0	0	5	NONE
FAM10605	0	0	0	0	0	5	NONE
FAM10636	0	0	0	0	0	5	NONE
FAM10658	0	0	0	0	0	5	NONE
FAM10673	0	0	0	0	0	5	NONE
FAM10674	0	0	0	0	0	5	NONE

Supplementary Table 3h: Gene families showing significant changes (z-score > 1.96) in copy number between *Trichuris* species and other nematodes, sorted by total copy number across nematode species.

family	*Trichinella spiralis*	*Bursaphelenchus xylophilus*	*Brugia malayi*	*T. muris*	*C. elegans*	*T. trichiura*	z-score	annotation
FAM170	4	7	8	2	9	2	2.167	alpha_tubulin,tubulin_alpha_chain

FAM1027	1	0	0	4	0	21	1.965	hypothetical_protein,Pfam-B_6031_and_Pfam-B_2655-domain_containing_protein
FAM212	3	3	1	5	2	5	2.289	ABC_tran_domain_containing_protein,ABC_transporter,_ATP_binding_protein,ATP_binding_cassette_sub_family_A
FAM367	4	2	2	5	1	5	2.129	solute_carrier_family_12
FAM2217	0	0	0	15	0	3	2.000	transposase,hypothetical_protein,conserved_hypothetical_protein,Putative_tick_transposon,reverse_transcriptase,Pfam-B_19879-domain_containing_protein
FAM150	3	4	3	2	3	2	2.214	protein_unc_g;_protein_unc_f;_protein_unc_d;_protein_unc_b;_protein_unc_a,Muscle_M_line_assembly_protein_unc_89
FAM2794	0	0	0	6	0	10	2.469	Motile_Sperm_and_Pfam-B_11622-domain_containing_protein,Motile_Sperm-domain_containing_protein,Pfam-B_4364_and_Motile_Sperm-domain_containing_protein
FAM2393	1	1	1	4	1	7	2.390	DUF229-domain_containing_protein
FAM987	2	2	1	3	2	3	2.214	LRR_1_and_LRR_8_and_LRRNT_and_I-set_and_fn3-domain_containing_protein,LRR_6_and_LRR_8_and_Pfam-B_12548-domain_containing_protein,LRR_8_and_Pfam-B_7182-domain_containing_protein
FAM4661	0	0	0	7	0	5	2.530	hypothetical_protein,dsrm-domain_containing_protein
FAM959	3	2	3	1	2	1	2.236	U5_small_nuclear_ribonucleoprotein_40_kDa
FAM1856	2	3	3	1	2	1	2.236	diphthine_synthase
FAM4224	2	3	2	1	3	1	2.236	animal_hem_peroxidase_family_protein;_animal_haem_peroxidase_family_protein
FAM5307	0	0	0	4	0	7	2.449	hypothetical_protein
FAM1043	3	2	4	0	2	0	2.289	NONE
FAM5303	1	0	0	3	0	7	2.273	Trypsin-domain_containing_protein
FAM5302	0	0	0	8	0	3	2.256	hypothetical_protein,conserved_hypothetical_protein,Pfam-B_19346-domain_containing_protein,DDE_Tnp_IS1595_and_Mononeg_RNA_pol-domain_containing_protein
FAM5305	0	0	0	3	0	8	2.256	histone_h3
FAM585	2	2	3	1	2	1	2.214	mitochondrial_ribosomal_protein_L48
FAM2532	2	2	3	1	2	1	2.214	phospholipase,_patatin_family
FAM3089	3	2	2	1	2	1	2.214	sodium_independent_organic_anion_transporter
FAM5304	0	0	0	9	0	2	2.037	ankyrin_2,3:unc44,conserved_hypothetical_protein,reverse_transcriptase,Pfam-B_19879-domain_containing_protein
FAM464	2	2	2	1	2	1	2.582	40S_ribosomal_protein_S10
FAM1182	2	2	2	1	2	1	2.582	zinc_transporter_2
FAM6043	1	0	0	6	1	2	2.073	zf-H2C2_2-domain_containing_protein
FAM5308	2	1	1	3	1	2	2.041	conserved_hypothetical_protein,Ribosomal_L9_N-domain_containing_protein,Pfam-B_13521-domain_containing_protein
FAM6830	0	0	0	4	0	5	2.558	Trypsin-domain_containing_protein,transmembrane_serine_protease_8,BTB_and_Trypsin-domain_containing_protein
FAM930	1	1	1	2	1	3	2.390	transcription_initiation_factor_TFIID_subunit,TFIID-31kDa-domain_containing_protein
FAM6827	1	1	1	3	1	2	2.390	prestin,solute_carrier_family_26
FAM3545	0	1	1	2	1	4	2.176	DSPc-domain_containing_protein,RNA:RNP_complex_1_interacting_phosphatase
FAM6828	1	0	0	2	0	6	2.132	Kringle-domain_containing_protein,coagulin_factor_II
FAM850	1	1	1	2	1	2	2.582	protein_grainyhead

FAM1306	1	1	1	2	1	2	2.582	Orai-1-domain_containing_protein,calcium_release_activated_calcium_channel
FAM2615	1	1	1	2	1	2	2.582	AP_3_complex_subunit_beta_2
FAM2678	1	1	1	2	1	2	2.582	tRNA-synt_His_and_HGTP_anticodon2-domain_containing_protein,RWD_and_Pkinase_and_Pfam-B_6749-domain_containing_protein
FAM2734	1	1	1	2	1	2	2.582	HT004_protein
FAM2787	1	1	1	2	1	2	2.582	phosphoglucomutase
FAM3451	1	1	1	2	1	2	2.582	dentin_matrix_protein_4,Pfam-B_12616_and_DUF1193-domain_containing_protein
FAM4423	1	1	1	2	1	2	2.582	histone_H2A,histone_H2A_variant
FAM5071	1	1	1	2	1	2	2.582	plasma_alpha_l_fucosidase,alpha_L_fucosidase
FAM5126	1	1	1	2	1	2	2.582	26S_proteasome_regulatory_complex_ATPase_RPT2,26S_protease_regulatory_subunit_4
FAM5296	1	1	1	2	1	2	2.582	1,4_alpha_glucan_branching_enzyme
FAM7547	0	0	0	6	0	2	2.202	WAP_domain_containing_protein,_SLPI-like
FAM520	2	2	2	1	1	0	2.041	racgtpase_activating_protein
FAM2820	2	2	2	1	1	0	2.041	Histone_lysine_N_methyltransferase_pr_set7
FAM8362	0	0	0	4	0	3	2.543	hypothetical_protein
FAM8373	0	0	0	3	0	4	2.543	Pfam-B_9093-domain_containing_protein
FAM3245	2	2	2	0	1	0	2.373	NONE
FAM3134	1	1	0	2	1	2	2.214	signal_peptidase_complex_catalytic_subunit,Signal_peptidase_complex_catalytic_subunit
FAM3865	0	1	1	2	1	2	2.214	RNA_polymerase_II_associated_factor_1

Supplementary Table 3i: Gene families showing significant changes (z-score > 1.96) in copy number between *Trichuris muris* and *T. trichiura*, sorted by total copy number in *Trichuris*.

family	*Trichinella spiralis*	*Bursaphelenchus xylophilus*	*Brugia malayi*	*T. muris*	*C. elegans*	*T. trichiura*	z-score	annotation
FAM217	0	1	1	47	1	0	2.480	conserved_hypothetical_protein,hypothetical_protein
FAM806	0	0	0	29	0	0	2.449	hypothetical_protein,Pfam-B_7026-domain_containing_protein,Pfam-B_19346-domain_containing_protein,Pfam-B_7026_and_Pfam-B_16276-domain_containing_ B_19346_and_Pfam-B_7026-domain_containing_protein,Pfam-B_7026_and_Pfam-B_19346-domain_containing_protein
FAM1124	0	0	0	25	0	0	2.449	conserved_hypothetical_protein,hypothetical_protein
FAM1027	1	0	0	4	0	21	2.046	Pfam-B_6031_and_Pfam-B_2655-domain_containing_protein,hypothetical_protein
FAM1344	0	0	0	23	0	0	2.449	RNA_dependent_RNA_polymerase,Mononeg_RNA_pol-domain_containing_protein,large_protein,hypothetical_protein,polymerase
FAM1988	0	0	0	19	0	0	2.449	conserved_hypothetical_protein,tigger_transposable_element_derived_protein,hypothetical_protein
FAM1495	0	1	0	18	0	0	2.473	hypothetical_protein,conserved_hypothetical_protein,jerky_protein
FAM2219	0	0	0	18	0	0	2.449	conserved_hypothetical_protein,hypothetical_protein
FAM2218	0	0	0	17	0	1	2.329	conserved_hypothetical_protein,Pao_retrotransposon_peptidase_superfamily,zinc_knuckle_protein,Gag_Pol_polyprotein
FAM2217	0	0	0	15	0	3	2.000	transposase,conserved_hypothetical_protein,Putative_tick_transposon,Pfam-B_19879-domain_containing_protein,hypothetical_protein,reverse_transcriptase

FAM2502	0	0	0	16	0	1	2.321	hypothetical_protein,conserved_hypothetical_protein
FAM3202	0	0	0	15	0	0	2.449	tigger_transposable_element_derived_protein
FAM3605	0	0	0	14	0	0	2.449	conserved_hypothetical_protein,Pfam-B_516-domain_containing_protein,Pfam-B_516_and_MULE_and_FLYWCH-domain_containing_protein
FAM3607	0	0	0	14	0	0	2.449	conserved_hypothetical_protein,tigger_transposable_element_derived_protein,jerky_protein,hypothetical_protein
FAM3608	0	0	0	14	0	0	2.449	hypothetical_protein,conserved_hypothetical_protein,Pfam-B_19879-domain_containing_protein
FAM3653	0	0	0	14	0	0	2.449	hypothetical_protein
FAM4089	0	0	0	13	0	0	2.449	conserved_hypothetical_protein,Peptidase_A17_and_DUF667-domain_containing_protein,Pfam-B_19879-domain_containing_protein,ankyrin_2,3:unc44,DUF1758-domain_containing_protein,polyprotein
FAM4090	0	0	0	13	0	0	2.449	conserved_hypothetical_protein,Pfam-B_13107-domain_containing_protein,Pfam-B_516-domain_containing_protein,hypothetical_protein,FLYWCH-domain_contai
FAM4091	0	0	0	0	0	13	2.449	NONE
FAM4659	0	0	0	12	0	0	2.449	conserved_hypothetical_protein,hypothetical_protein
FAM4660	0	0	0	12	0	0	2.449	Putative_integrase_core_domain_protein,conserved_hypothetical_protein,zf-CCHC-domain_containing_protein,zf-CCHC_and_Pfam-B_10329-domain_containing_p
FAM4665	0	0	0	1	0	11	2.259	neurogenic_locus_notch_protein_2
FAM3206	0	0	1	0	1	11	2.526	NONE
FAM4666	0	0	1	0	0	11	2.485	NONE
FAM5306	0	0	0	11	0	0	2.449	hypothetical_protein,Pfam-B_10329-domain_containing_protein,Pfam-B_10329_and_zf-CCHC-domain_containing_protein,zf-CCHC_and_Pfam-B_10329-domain_
FAM5309	0	0	0	11	0	0	2.449	hypothetical_protein,conserved_hypothetical_protein
FAM5313	0	0	0	0	0	11	2.449	NONE
FAM6036	0	0	0	10	0	0	2.449	conserved_hypothetical_protein,hypothetical_protein
FAM6037	0	0	0	10	0	0	2.449	hypothetical_protein,conserved_hypothetical_protein,peripheral_plasma_membrane_protein_CASK
FAM6044	0	0	0	10	0	0	2.449	hypothetical_protein
FAM6050	0	0	0	0	0	10	2.449	NONE
FAM6042	1	0	0	8	0	1	2.229	conserved_hypothetical_protein,zf-C2H2_4_and_zf-C2H2-domain_containing_protein,zinc_finger_protein,zinc_finger_protein_569,zf-C2H2_4_and_zf-C2H2_and_ domain_containing_protein,zinc_finger_protein_729,pita,_B,zf-H2C2_2_and_zf-C2H2_4_and_zf-C2H2-domain_containing_protein
FAM6039	2	0	0	8	0	0	2.497	Pfam-B_310-domain_containing_protein,something_about_silencing_protein_10
FAM7543	0	0	0	8	0	0	2.449	conserved_hypothetical_protein
FAM7544	0	0	0	8	0	0	2.449	Pfam-B_10329-domain_containing_protein,RETRotransposon_family_member
FAM7546	0	0	0	8	0	0	2.449	conserved_hypothetical_protein,Dimer_Tnp_hAT-domain_containing_protein
FAM7700	0	0	0	8	0	0	2.449	hypothetical_protein
FAM6040	2	0	0	7	0	1	2.196	gag_pre-integrs_and_rve_and_Pfam-B_10563_and_zf-CCHC_and_Pfam-B_11267-domain_containing_protein,zf-CCHC_and_Pfam-B_11267_and_UBN2-domain_containing_protein,putative_pol_polyprotein;_copia_type_polyprotein_putative,retrovirus_Pol_polyprotein,polyprotein,copia_type_polyprotein,reverse_trans
FAM7548	0	0	0	7	0	1	2.139	actin_protein_6A_like,Actin-domain_containing_protein
FAM8361	0	0	0	7	0	0	2.449	conserved_hypothetical_protein,zf-CCHC-domain_containing_protein
FAM8364	0	0	0	7	0	0	2.449	protein_crumbs,neurogenic_locus_notch_protein_2,Uncharacterized_transposase_protein,neurogenic_locus_notch_protein,crumbs_1
FAM8365	0	0	0	7	0	0	2.449	conserved_hypothetical_protein,zinc_finger_protein_646,hypothetical_protein

FAM8369	0	0	0	7	0	0	2.449	conserved_hypothetical_protein,Pfam-B_516-domain_containing_protein
FAM8399	0	0	0	0	0	7	2.449	NONE
FAM3750	1	2	0	1	3	6	2.340	onchocystatin
FAM5503	1	1	0	6	0	1	2.214	zf-CCHC_and_Pfam-B_2545_and_Asp_protease_2_and_Pfam-B_2707_and_rve-domain_containing_protein,Gap_Pol_polyprotein,rve_and_Pfam-B_10563_and_zf-CC B_2545_and_Asp_protease_2_and_RVT_1_and_Pfam-B_2707-domain_containing_protein,rve_and_zf-CCHC_and_Pfam-B_10329_and_RVT_1_and_Pfam-B_2707-domain_containing_protein,zf-CCHC_and_Asp_protease_2_and_Pfam-B_10329_and_RVT_1-domain_containing_protein,zf-H2C2_and_rve_and_Pfam-B_10329_and B_2545_and_RVP_and_RVT_1_and_Pfam-B_2707-domain_containing_protein
FAM7545	1	0	0	6	0	1	2.138	reverse_transcriptase:endonuclease,reverse_transcriptase,Pfam-B_595-domain_containing_protein
FAM8363	0	0	0	6	0	1	2.082	hypothetical_protein
FAM8370	0	0	0	6	0	1	2.082	CUG-BP-and_ETR-3-like_factor_3,hypothetical_protein,Calcium_channel_protein,_putative
FAM8398	0	0	0	1	0	6	2.082	hypothetical_protein

Supplementary Table 11: Differential expression of mouse genes of interest between naïve and *T. muris*-infected cecum

Cytokine	Ensembl gene id	Known role	Direction	Fold change	FDR
IFN-gamma	ENSMUSG0000005517	Inflammation	Up	26	<5%
IL-1 Beta	ENSMUSG00000027398		Up	23	<5%
IL-4	ENSMUSG00000000869		-	-	-
IL-5	ENSMUSG00000036117		-	-	-
IL-6	ENSMUSG00000025746		Up	17	<5%
IL-9	ENSMUSG00000021538		-	-	-
IL-10	ENSMUSG00000016529		Up	20	5-10%
IL-13	ENSMUSG00000020383		-	-	-
IL-15	ENSMUSG00000031712		Down	4	<5%
IL-16	ENSMUSG00000001741		Up	3	<5%
IL-18	ENSMUSG00000039217		Down	10	<5%
IL-21	ENSMUSG00000027718		Up	Inf	5-10%
IL-22	ENSMUSG0000007469	Therapeutic T. trichiura infection (Broadhurst et al 2010)	Up	10	5-10%
IL-27	ENSMUSG00000044701		Up	17	5-10%
IL-33	ENSMUSG00000024810		Up	4	5-10%
IL-34	ENSMUSG00000031750		Up	5	<5%
TNF	ENSMUSG0000002440	Inflammation	Up	12	<5%

Cytokine receptor	Ensembl gene id	Known role	Direction	Fold change	FDR
IL-1R antagonist	ENSMUSG00000026981		Up	23	<5%
IL-1R type II	ENSMUSG00000026073		Up	6	<5%
IL-2R beta	ENSMUSG00000068227		Up	8	<5%
IL-2R gamma	ENSMUSG00000031304		Up	8	<5%
IL-3R alpha	ENSMUSG00000068758		Up	4	<5%
IL-4R alpha	ENSMUSG00000030748		Up	5	<5%
IL-10R alpha	ENSMUSG00000032089		Up	7	<5%
IL-12R beta 1	ENSMUSG00000000791		Up	17	<5%
IL-12R beta 2	ENSMUSG00000018341		Up	13	<5%
IL-17R E	ENSMUSG00000043088		Down	6	<5%
IL-21R	ENSMUSG00000030745		Up	5	<5%
IL-27R alpha	ENSMUSG0000000546	Colitis (Villarino et al. 2008)	Up	3	5-10%

Chemokine	Ensembl gene id	Known role	Direction	Fold change	FDR
CCL2	ENSMUSG0000003538	Macrophage chemoattraction	Up	42	<5%
CCL5	ENSMUSG0000003504	nonocyte/T cell/eosinophil chemoattractior	Up	21	<5%
CCL7	ENSMUSG0000003537	Macrophage chemoattraction	Up	13	<5%
CXCL5	ENSMUSG0000002937	Neutrophil regulation/activation	Up	19	<5%
CXCL9	ENSMUSG0000002941	IFN-g induced chemoattractant for T cells	Up	97	<5%
CXCL10	ENSMUSG0000003485	Regulation of epithelial cell turnover	Up	20	<5%
CXCL16	ENSMUSG0000001892	NK T cell migration	Up	7	<5%

Chemokine receptor	Ensembl gene id	Known role	Direction	Fold change	FDR
CCR2	ENSMUSG0000004910	Macrophage chemoattraction	Up	9	<5%
CCR5	ENSMUSG0000007922	nonocyte/T cell/eosinophil chemoattractan	Up	22	<5%
CXCR2	ENSMUSG0000002618	Neutrophil regulation/activation	Up	70	<5%
CXCR3	ENSMUSG0000005023	Regulation of epithelial cell turnover	Up	14	<5%
CXCR6	ENSMUSG0000004852	-	Up	8	<5%

Immunoglobulin constant chains	Ensembl gene id	Known role	Direction	Fold change	FDR
IgA	ENSMUSG00000095079		Up	4	<10%
IgE	ENSMUSG00000087642		-	-	-
IgG1	ENSMUSG00000076614		Up	42	<5%
IgG2B	ENSMUSG00000076613		Up	15	<5%
IgG2C	ENSMUSG00000076612		Up	57	<5%
IgM	ENSMUSG00000076617		Up	9	<5%

Other	Ensembl gene id	Known role	Direction	Fold change	FDR
Arg1	ENSMUSG00000019987		Up	522	<5%
CD4	ENSMUSG00000023274		Up	4	<5%
CD16	ENSMUSG00000059089		Up	44	<5%
CD23	ENSMUSG00000058715		Up	6	<5%
CD25	ENSMUSG00000026770		-	-	-
CD48	ENSMUSG00000015355		Up	4	<5%
Foxp3	ENSMUSG00000039521		-	-	-
LCP2/SLP-76	ENSMUSG00000002699		Up	8	<5%
mcpt-1	ENSMUSG00000022227		Up	109	<5%
RHOH	ENSMUSG00000029204		Up	4	<5%
Ym1	ENSMUSG00000040809		Up	94	<5%

Supplementary Table 15: Genomic libraries of *T. muris* and *T. trichiura*

Illumina paired end sequence data for genome assembly

Organism	median insert size (bases)	read length (bases)	total yield (kilobases)	ENA sample accession number	library type
T. muris	482	76	1,189,822	ERS016744	standard
T. muris	237	76	1,065,280	ERS016965	PCR-free
T. muris	241	100	19,432,799	ERS016965	PCR-free
T. trichiura	455	100	21,278,650	ERS056020	PCR-free

Shotgun and paired end 454 sequence data for *T. muris* genome assembly

library type	mean insert size (bases)	total yield (megabases)	ENA sample accession number
shotgun	n/a	1376.7	ERS016965
paired	3000	1468.1	ERS016965
paired	8000	829.4	ERS244696

Illumina paired end sequence data for separate male- and female-specific chromosomal analyses in *T. muris*

Sex	number of individuals	median insert size (bases)	read length (bases)	total yield (kilobases)	ENA sample accession number
male	*1*	350	100	17,856,062	ERS326152
female	*11*	346	100	18,913,337	ERS326151

Supplementary Table 16a: RNA-seq libraries of *Trichuris muris* life cycle stages, morphological regions, and genders

stage	organism part	number of reads	ENA sample accession number	ENA Study accession number	Array Express Study Accession Number	Approximate no. of worms
female	whole	57,353,866	ERS092077	ERP002000	E-ERAD-125	20
male	whole	63,267,958	ERS092078	ERP002000	E-ERAD-125	40
L3	whole	97,219,262	ERS092416	ERP002000	E-ERAD-125	80
L3	whole	95,404,266	ERS092417	ERP002000	E-ERAD-125	80
L3	whole	77,412,612	ERS092418	ERP002000	E-ERAD-125	80
Adults	whole	78,034,450	ERS092413	ERP002000	E-ERAD-125	7
Adults	whole	76,443,982	ERS092414	ERP002000	E-ERAD-125	7
Adults	whole	97,388,714	ERS092415	ERP002000	E-ERAD-125	7
Adult male	whole	73,271,100	ERS092411	ERP002000	E-ERAD-125	15-20
Adult male	whole	93,014,642	ERS092412	ERP002000	E-ERAD-125	15-20
Adult female	whole	99,154,882	ERS092410	ERP002000	E-ERAD-125	15-20
Adult female	whole	95,489,468	ERS092419	ERP002000	E-ERAD-125	15-20
Adult (mixed) intestinal phase	anterior	19,042,100	ERS092566	ERP002000	E-ERAD-125	100
Adult (mixed) intestinal phase	anterior	20,480,010	ERS092567	ERP002000	E-ERAD-125	100
Adult (mixed) intestinal phase	anterior	39,431,492	ERS092568	ERP002000	E-ERAD-125	100
Adult female lumenal phase	rear end	25,216,324	ERS092569	ERP002000	E-ERAD-125	65
Adult female lumenal phase	rear end	22,233,346	ERS092570	ERP002000	E-ERAD-125	65
Adult female lumenal phase	rear end	25,753,804	ERS092571	ERP002000	E-ERAD-125	65
Adult male lumenal phase	rear end	19,998,130	ERS092572	ERP002000	E-ERAD-125	35
Adult male lumenal phase	rear end	28,575,550	ERS092573	ERP002000	E-ERAD-125	35
Adult male lumenal phase	rear end	23,479,798	ERS092574	ERP002000	E-ERAD-125	35
L2	whole	26,943,510	ERS195817	ERP002000	E-ERAD-125	150-450

All data are 100 bp paired end reads.

Supplementary Table 16b: RNA-seq of *Mus musculus* response to *Trichuris muris* infection

sample description	mouse identifier	number of reads	ENA sample accession number	ENA Study accession number	Array Express Study Accession Number
infected wormy cecum, worms left in cecum	1	56,772,210	ERS167948	ERP002560	E-ERAD-181
infected non-wormy cecum	1	21,861,330	ERS167949	ERP002560	E-ERAD-181
naïve MLN	1	38,681,928	ERS167950	ERP002560	E-ERAD-181
infected wormy cecum, worms removed from cecum	5	25,917,952	ERS167951	ERP002560	E-ERAD-181
infected non-wormy cecum	5	102,339,636	ERS167952	ERP002560	E-ERAD-181
infected MLN	5	37,839,610	ERS167953	ERP002560	E-ERAD-181
naïve wormy cecum	9	30,878,770	ERS167954	ERP002560	E-ERAD-181
naive non-wormy cecum	9	58,081,706	ERS167955	ERP002560	E-ERAD-181
naïve MLN	9	12,367,976	ERS167956	ERP002560	E-ERAD-181
naïve wormy cecum	12	26,882,810	ERS167957	ERP002560	E-ERAD-181
naive non-wormy cecum	12	35,551,572	ERS167958	ERP002560	E-ERAD-181
naïve MLN	12	45,406,130	ERS167959	ERP002560	E-ERAD-181
infected wormy cecum, worms left in cecum	2	38,508,088	ERS167960	ERP002560	E-ERAD-181
infected non-wormy cecum	2	16,629,034	ERS167961	ERP002560	E-ERAD-181
infected MLN	2	13,393,252	ERS167962	ERP002560	E-ERAD-181
infected wormy cecum, worms removed from cecum	6	66,298,346	ERS167963	ERP002560	E-ERAD-181
infected non-wormy cecum	6	19,937,558	ERS167964	ERP002560	E-ERAD-181
infected MLN	6	26,012,152	ERS167965	ERP002560	E-ERAD-181
naïve wormy cecum	10	122,959,842	ERS167966	ERP002560	E-ERAD-181
naive non-wormy cecum	10	29,089,220	ERS167967	ERP002560	E-ERAD-181
naïve MLN	10	33,272,308	ERS167968	ERP002560	E-ERAD-181
naïve wormy cecum	13	77,912,600	ERS167969	ERP002560	E-ERAD-181
naive non-wormy cecum	13	31,619,046	ERS167970	ERP002560	E-ERAD-181
naïve MLN	13	29,053,184	ERS167971	ERP002560	E-ERAD-181
infected wormy cecum, worms left in cecum	3	38,260,350	ERS167972	ERP002560	E-ERAD-181
infected non-wormy cecum	3	41,858,106	ERS167973	ERP002560	E-ERAD-181
infected MLN	3	28,546,718	ERS167974	ERP002560	E-ERAD-181
infected wormy cecum, worms removed from cecum	7	17,482,126	ERS167975	ERP002560	E-ERAD-181
infected non-wormy cecum	7	25,491,194	ERS167976	ERP002560	E-ERAD-181
infected MLN	7	25,223,414	ERS167977	ERP002560	E-ERAD-181
naïve wormy cecum	11	23,179,374	ERS167978	ERP002560	E-ERAD-181
naive non-wormy cecum	11	18,646,504	ERS167979	ERP002560	E-ERAD-181
naïve MLN	11	18,568,222	ERS167980	ERP002560	E-ERAD-181
infected wormy cecum, worms left in cecum	4	13,766,456	ERS167981	ERP002560	E-ERAD-181
infected non-wormy cecum	4	35,879,338	ERS167982	ERP002560	E-ERAD-181
infected MLN	4	31,194,586	ERS167983	ERP002560	E-ERAD-181
infected wormy cecum, worms removed from cecum	8	37,283,248	ERS167984	ERP002560	E-ERAD-181
infected non-wormy cecum	8	45,572,986	ERS167985	ERP002560	E-ERAD-181
infected MLN	8	33,801,694	ERS167986	ERP002560	E-ERAD-181
naïve wormy cecum	14	29,559,234	ERS167987	ERP002560	E-ERAD-181
naive non-wormy cecum	14	30,191,428	ERS167988	ERP002560	E-ERAD-181
naïve MLN	14	29,644,302	ERS167989	ERP002560	E-ERAD-181

All data are 100 bp paired end reads. Parasite-containing samples contained approximately 10 worms.

"Naïve" = from uninfected mouse; "wormy cecum" = a section of cecum where the worms reside; "non-wormy cecum" = a section of cecum without worms.

III. SUPPLEMENTARY NOTE

Trichuris muris genome sequencing

Illumina

Illumina libraries originated from *T. muris* parasites, Edinburgh strain, grown in male athymic nude mice that were 6-12 weeks of age and bred in the Biological Services Unit at the University of Manchester. All mice were housed in isolator cages, and all animal experiments were performed under the auspices of the University of Manchester ethical review committee and under the Home Office Scientific Procedures Act (1986). The parasites were removed from the ceca of nude mice and cleaned to remove as much host contaminating material as possible. They were then cultured in single wells at 37°C, 5% CO_2 for up to 7 days in RPMI,100U/ml penicillin,100µg/ml of streptomycin. Unembryonated eggs were collected from the individually cultured females and embryonated in tissue culture flasks in milliQ water for at least 8 weeks in the dark to develop. Eggs from a single female worm were then used to re infect nude mice and passaged sequentially as above 5-7 times prior to selecting worms for DNA extraction to decrease genetic variability. Genomic DNA (gDNA) was obtained using DNA extraction buffer, briefly male worms were incubated overnight at 56°C in 0.1M Tris-HCL, pH 8.5, 0.05M EDTA, 0.2M NaCl, 1% w/v SDS and 200 µg/ml of proteinase K. Phenol, chloroform extraction was used to purify the gDNA and incubation at 37 °C for 30 min with 1.25U of RNAse to remove contaminating RNA. The DNA was precipitated using 100% ethanol and glycogen. After centrifugation at 12000g, the pellet was resuspended in nuclease free water and concentration obtained with an Invitrogen Qubit fluorometer. Using protocols previously described gDNA was used directly for preparation of one amplification-free[78] and one standard Illumina library[79,80]. DNA was eluted after each enzymatic stage using a Qiagen QIAquick PCR purification kit. Size selection of the adapter ligated DNA was performed on a 2% agarose gel and the DNA extracted. The standard library was then amplified using 16 cycles of PCR. Further details are in Supplementary Table 15.

Libraries were denatured using 0.1M sodium hydroxide and diluted in a hybridisation buffer to allow the template strands to hybridise to adapters attached to the flowcell surface. Cluster amplification was performed on the Illumina cluster station or cBOT using the v4 cluster generation kit following the manufacturer's protocol and then a SYBRGreen QC was performed to measure cluster density and determine whether to pass or fail the flowcell for sequencing, followed by linearization, blocking and hybridization of the R1 sequencing primer. The hybridized flow cells were loaded onto the Illumina Genome Analyser IIX for 76 cycles or the Illumina HiSeq for 100 cycles of sequencing-by-synthesis using the v4 or v5 SBS sequencing kit. *In situ*, the linearization, blocking and hybridization step was repeated to regenerate clusters, release the second strand for sequencing and to hybridise the R2 sequencing primer followed by another 76 or 100 cycles of sequencing to produce paired end reads. These steps were performed using proprietary reagents according to manufacturer's recommended protocol (http://www.illumina.com/). Data was analysed from the Illumina sequencing machines using the RTA1.6 or RTA1.8 analysis pipelines.

Male and female samples

Adult male and female *T. muris* worms were collected from ceca of nude mice and cleaned to remove host tissue. Male and female worms were separated based on size and morphology. The posterior (reproductive) ends of eleven individual females were removed (and discarded) and the anterior ends were pooled together. Genomic DNA was extracted using the Promega Wizard DNA Purification Kit. Amplification-free Illumina libraries were produced using methods described above and sequenced on an Illumina HiSeq to produce 100bp paired-end reads.

454

Genomic DNA, prepared using the method described above for Illumina sequencing, was used to produce paired-end (3 kb and 8 kb) and shotgun 454 libraries (Supplementary Table 15) using standard Roche protocols (http://www.454.com) and sequenced using the 454 Life Sciences GS-20 and GS-FLX sequencer (Roche).

Optical map

High molecular weight *T. muris* genomic DNA was prepared by proteinase K lysis of trypsin digested adults mixed with molten agarose set in plugs. Briefly, male worms or mixed sexed adult worms were removed from the cecum of an athymic nude mouse and thoroughly cleaned to remove as much contaminating host material as possible. Agarose plugs were prepared using a Bio Rad CHEF genomic DNA plug kit, for mammalian DNA. The worms were chopped using a scalpel blade and re suspended in cell lysis buffer. This was mixed in a 2:1 ratio with 2% CleanCut® agarose and set on ice. The plugs were incubated overnight at 56 °C in a proteinase k buffer and washed 4 times the following day. The penultimate wash had 1mM PMSF added to inhibit further proteinase k digestion. The agarose plugs were stored at 4°C in kit wash buffer until processed. DNA molecules were stretched and immobilized along microfluidic channels before digestion with the restriction endonuclease SpeI, yielding a set of ordered restriction fragments in the order that they occur within the genome. The fragments were fluorescently stained and visualised to determine the fragment sizes. Assembling overlapping fragment patterns of single molecule restriction maps produced an optical map of the genome. The *T. muris* optical map consists of 46 contigs, an assembled size of 85.24Mb and approximately 70 x genome coverage of optical data. The optical data were generated and analysed using the Argus Optical Mapping System from OpGen and analysed with associated MapManager and MapSolver software tools (http://www.opgen.com/products-services/argus-system).

T. muris genome assembly and improvement

For *T. muris*, 3,181 Mb of 454 reads (Supplementary Table 15) were assembled into contigs with Celera assembler[54] v7.0. In addition, 2.7 Gb of 76bp paired end Illumina reads were used to close 451 gaps and correct the sequence using 15 iterations of IMAGE[57] producing *T. muris* assembly v1.

Gap5 [81] was used to interrogate and edit the version 1 assembly (un-softclipping reads and manually joining gaps based on read pairs and sequence matches). A separate *de novo* Illumina assembly using Velvet[53] v1.2.03 was performed and remapped to the Celera assembly, allowing 143 gaps in the Celera assembly to be joined using consensus sequence from contigs in the Velvet assembly. Contamination was removed by identifying contigs with high homology to non-invertebrate genes. Inverted repeats were identified by (i) observing incorrectly orientated read pairs at sequence gaps between contigs and (ii) sequence searches hitting to two or more places either side of the gap. These were resolved by correct placement of spanning paired end reads within gap5. Following the manual correction of misassemblies as described above, re-scaffolding was performed using SSPACE[56] and gaps were subsequently closed using 20 iterations of IMAGE[57], resulting in **assembly v2.1**.

Assembly v2.1 was further improved by using Reapr[55] to identify potential misassembled regions, which were then manually fixed in Gap5 [81]. Additional scaffolding was performed in Gap5 using paired end read information. These improved scaffolds were then aligned against the optical map contigs in MapSolver (OpGen), allowing incorrect joins in the sequence assembly to be resolved and new joins to be made. Typically, sequence scaffolds and optical contigs did not begin or end at the same point; many optical scaffolds extended beyond sequence scaffolds and *vice versa*. Therefore, a combination of the two data types (optical scaffolds and assembly sequence scaffolds) was used to correct assembly errors, resulting in genome **assembly v3.0**. Finally, sequence scaffolds that were confidently spanned by optical map contigs were joined even in cases where the corresponding gaps were larger than 10kb. Such sequence scaffolds were joined by a stretch of the letter "N" of appropriate length to indicate the approximate gap length as evidenced by the optical map. This produced large linkage groups ("superscaffolds") and resulted in the final genome **assembly v4.0**.

T. trichiura genome sequencing

Adult *T. trichiura* worms were obtained from young children with *T. trichiura* ova in stool samples. The study protocol was approved by the ethics committees of the Hospital Pedro Vicente Maldonado, Pichincha Province, and Pontificia Universidad Catolica del Ecuador, Quito, Ecuador. Informed written consent was obtained from the mother or primary carer of each child. Children were treated with a single dose of 5 mg/kg of Combantrin (oxantel and pyrantel pamoate, Pfizer) and stool samples were collected for 24 hours after treatment. Expelled worms were washed thoroughly in sterile saline and stored in liquid nitrogen before being shipped on dry ice. Genomic DNA (234 ng) isolated from a single male adult worm, using Qiagen Genomic-tip-20, was used to make a PCR-free short fragment Illumina library using the same protocol as used for *T. muris* above.

T. trichiura genome assembly and improvement

First, 450 bp fragment paired end Illumina reads were corrected and assembled using SGA[52] v0.9.17. This draft assembly was then used to calculate the highest occurrence of unique k-mers between odd values from 41 to 81 using GenomeTools (http://genometools.org). This kmer along with the corrected sequence reads was subsequently used to generate a second assembly using Velvet[53] v1.2.03. Scaffolds less than 500bp from this assembly were discarded. A hybrid assembly was created by merging the SGA assembly with scaffolds greater than 15 kb from the Velvet assembly. 11 iterations of IMAGE gap closure[57] were run on the hybrid assembly and further gaps were closed manually within Gap5 [81]. Illumina paired end reads that failed to map to this improved hybrid assembly were assembled with Velvet[53] to create a 'bin' assembly, which was merged with the improved hybrid assembly. Contamination was removed by identifying contigs with high homology to non-invertebrate genes. Further scaffolding was performed with SSPACE and gap closure undertaken using IMAGE[57] (2 iterations) followed by Gapfiller[58]. Manual improvement was then carried out in Gap5 [81] by using Reapr[55] to target misassemblies, and gaps were closed manually to overcome the fragmentation within the assembly caused by haplotypic differences. Following manual improvement, ICORN[82] v2 was run to resolve any errors created by the process of merging haplotypic sequence, ensuring that a true representation of one or other allele was achieved over at least the length of the Illumina library fragment size (~450 bp) and to reduce the amount of phasing between haplotypes. Finally, IMAGE[57] (3 iterations) and Gapfiller[58] were run once more, resulting in v2.0 of the *T. trichiura* genome assembly. Finally, contamination screening was performed again, yielding assembly v2.1.

Transcriptome sequencing - *T. muris*

Adult worms or larval stages were prepared using TRIZOL® and lysing matrix D, (1.4 mm ceramic spheres) and a Fastprep24 both from MP biomedicals. The tubes containing worms, TRIZOL and matrix were subjected to 3 x 20 second cycles and placed on ice in between each cycle. RNA extraction was carried out according to the manufacturers protocol and RNA was resuspended in water and quantified using an Agilent 2100 Bioanalyzer microfluidics platform.

Paired end Illumina transcriptome libraries were prepared from total RNA. Two protocols (TruSeq and Illumina mRNA-seq kit) were used for preparing the Illumina transcriptome libraries listed in Supplementary Table 16. Polyadenylated mRNA was purified from total RNA using oligo-dT dynabead selection. In the TruSeq protocol, enzymatic fragmentation was used, and in the Illumina mRNA-seq kit protocol fragmentation was performed using metal ion hydrolysis with the Ambion RNA fragmentation kit. First strand synthesis, primed using random oligonucleotides, was followed by 2nd strand synthesis with RNaseH and DNApolI to produce double-stranded cDNA using the Illumina mRNA Seq kit or the TruSeq Illumina kit. Template

DNA fragments were end-repaired with T4 and Klenow DNA polymerases and blunt-ended with T4 polynucleotide kinase. A single 3′ adenosine was added to the repaired ends using Klenow exo- and dATP to reduce template concatemerization and adapter dimer formation, and to increase the efficiency of adapter ligation. Adapters (containing primer sites for sequencing, and index sequences when using the TruSeq protocol) were then ligated. Libraries made with the TruSeq protocol were amplified by PCR (to enrich for properly ligated template strands, to generate enough DNA, and to add primers for flowcell surface annealing). AMPure SPRI beads were used to purify amplified templates before pooling based on quantification using an Agilent Bioanalyser chip. Pooled TruSeq libraries were then pooled and size selected (300-400bp fragments) using the Caliper. After adaptor ligation, individual libraries made with the Illumina mRNA-seq kit were size selected using the caliper before PCR amplification followed by AMPure SPRI bead clean up and removal of adaptors with a second Caliper run. Kapa Illumina SYBR Fast qPCR kit was used to quantify the Illumina mRNA-seq libraries before pooling. All transcriptome libraries were run on Illumina HiSeq 2000 sequencing machines as described for genome Illumina sequencing above.

Gene predictions

T. muris

CEGMA predictions[6] were first used to train the *T. muris*-specific parameters in Augustus[59] v2.4. Split RNAseq reads suggestive of intron boundaries were converted into intron hints and were used to create a first set of *de novo* gene predictions with Augustus. Based on these, a set of 469 manually curated gene predictions were created and used to re-train Augustus. The re-trained Augustus v2.5.5 predictor was then used in conjunction with RNAseq hints to predict **gene set v2.1** with 12,126 gene models, based on genome assembly v2.1. A set of 172 genes that were either of high interest (e.g. WAP domain-containing proteins) or appeared to be incorrect based on semi-automatic screens and manual inspection (e.g. extremely uneven RNAseq coverage across a gene) were then manually curated to yield **gene set v2.2** with 12,145 genes. In addition, 1,141 likely transposon-related genes were identified and removed to yield the final **gene set v2.3** containing 11,004 genes. Likely transposon-related genes were identified based on RepeatRunner[83] (run as part of Maker, see below for *T. trichiura*) and on the presence in the predicted protein sequences of any of the following Pfam domains: PF00077, PF00078, PF00665, PF00680, PF01541, PF03184, PF03221, PF03564, PF03732, PF05380, PF07727, PF10551, PF12762, PF13456, PF13961, and PF14227 (Pfam[84] v27; predicted as part of Interproscan, see below). Potentially transposon-related genes were retained in the gene set if they were predicted to also contain another, not transposon-related Pfam domain.

T. trichiura

Gene predictions for *T. trichiura* were conducted by various methods available in MAKER[60] v2.2.28. The MAKER annotation pipeline consists of 4 general steps to generate high-quality annotations by taking into account evidence from multiple sources. First, assembled contigs (genome assembly v2.0) were filtered against RepeatRunner[83] and a species-specific repeat library (generated by Repeat Modeler[85] [www.repeatmasker.org]) using RepeatMasker[86] (http://www.repeatmasker.org), to identify and mask repetitive elements in the genome. Second, gene predictors Augustus[59] v2.5.5, GeneMark-ES[61] v2.3a (self-trained), and SNAP[62] 2013-02-16 were employed to generate *ab initio* gene predictions that can use evidence within Maker. Further species-specific gene models were provided to Maker using comparative algorithms against the *T. trichiura* genome: genBlastG[87] output of *Caenorhabditis elegans* gene models (WormBase[88]) and RATT[89] (Rapid Annotation Transfer Tool) output based on *T. muris* gene models. These models cannot be influenced by Maker evidence as they were provided by GFF file. Next, species-specific cDNAs and proteins from related organisms were aligned against the genome using BLASTN and BLASTX[90], and these alignments were further refined with respect to splice sites using Exonerate[91]. As there are no publicly available species-specific expressed sequence tags (ESTs) and just five cDNAs available from INSDC[92] (International Nucleotide Sequence Database Consortium), the contribution of these data as evidence is minimal. Finally, the protein homology alignments, comparative gene models and *ab initio* gene predictions were integrated and filtered by MAKER and in-house scripts to produce a set of evidence-informed gene annotations.

The MAKER genome annotation pipeline was run three consecutive times. 1) In the absence of a species-specific trained gene predictor, Augustus and SNAP were trained using CEGMA[6] protein evidence gained from the default eukaryotic orthologous groups (KOGs) and HMM profiles of nematode orthologous groups[11] (NOGs). 2) The first run of MAKER was performed using the est2genome and protein2genome option with the handful of taxon-specific cDNAs and nematode protein sequences, respectively. Gene models obtained from the first run were used to retrain SNAP and models from the second run were used to retrain Augustus. 3) With the trained models, MAKER was run a third time using a taxonomically broader protein set that included metazoan proteins from the UniProt Complete proteome database[93] and a subset of helminth proteomes from GeneDB[94], yielding **gene set v2.0** (based on genome assembly v2.0). Removing genes located on genome scaffolds identified as contamination (assembly v2.1) resulted in **gene set v2.1** with 9,856 genes. Finally, 206 likely transposon-related genes were identified (as described for the *T. muris* gene set) and removed to yield the final gene set **v2.2** containing 9,650 genes.

Functional gene annotation

Gene product descriptions for *T. muris* were determined as follows: gene models were searched against the UniProt database using BLASTP[90], and the top 10 hits for each gene were retrieved. The product descriptions of the BLAST hits were filtered using custom-built scripts. For genes without a good or informative BLAST hit (e-value < 0.0001 or many annotations as “hypothetical”), gene product descriptions were - if possible - based on Pfam protein domains predicted with Interproscan[77]. Genes lacking both Pfam domain and BLAST hit were labelled "hypothetical protein". For *T. trichiura* genes with a one-to-one ortholog in *T. muris* (as determined by OMA[65] v0.99t) product names were transferred between the orthologs. For all other *T. trichiura* genes, gene product descriptions were determined as described above for *T. muris* genes.

For further functional characterisation of the proteins, Interproscan v5.0.7 was run on both *Trichuris* proteomes. InterproScan conducted searches against Phobius[95] v1.01 to detect signal peptides and transmembrane domains, and against the following databases to identify further functional protein domains: Pfam[84] v27.0, SMART[96] v6.2, Gene3D v3.5.0, PANTHER v7.2, SUPERFAMILY v1.75, PRINTS v42.0, ProSiteProfiles v20.89, and ProSitePatterns. *T. muris* has 1300 genes with predicted signal peptides, and *T. trichiura* has 966. GO terms were assigned to proteins based on the Interpro[97] results, collecting GO term assignments associated with hits to Pfam and an E-value <= 0.01. Illustrations of protein domain architectures ("gene cartoons") were based on Interproscan v5.0.7 results for searches against the databases Pfam, SMART, and Phobius. Sequence logos for WAP domain-containing proteins were based on WAP domains as identified by Interpro (IPR008197). The following putative WAP domains were excluded from this analysis: WAP domains predicted to be shorter than 35 or longer than 55 amino acids, predicted to not contain any or to contain two "CC" dipeptides, and domains derived from the *Trichuris* and *Trichinella* homologs of mesocentin (TMUE_s0077003100, TTRE_0000351901, EFV57447). As a result, the sequence logos were based on 137 WAP domains from 42 proteins of *T. muris*, 61 WAP domains from 20 proteins of *T. trichiura*, 58 WAP domains from 23 proteins of *T. spiralis*, 38 WAP domains from 26 proteins of *H. sapiens*, and 19 WAP domains from 8 proteins of *C. elegans*. WAP domain alignments were produced with Mafft[66] v6.857 with the –auto parameter, and the sequence logos were created with the weblogo[98] server at http://weblogo.berkeley.edu/logo.cgi. The cysteine disulfide bonds highlighted in the sequence logo are based on the structure of the human proteins[99]. Proteases and protease inhibitors were detected using the MEROPS batch BLAST server[100]. KEGG orthology (KO) identifiers were based on bi-directional best hits using the KAAS webserver[101].

Gene family clustering and phylogenetic analysis

Gene family clusters were predicted using OrthoMCL[63] v2.0, and orthologs predicted using Inparanoid[64] v4.1. Identity between orthologs was calculated based on unfiltered global alignments of these pairs using Mafft[66] v6.857 with the –auto parameter. The phylogenetic tree in Supplementary Fig. 4 was constructed from the 236 gene families that contain only a single gene per genome and are present in at least 6 of the 8 species included in the comparison. Predicted amino acid sequence for each of these proteins were aligned using Mafft as above, and these alignments trimmed using GBlocks[67] v0.91b with options '-t=p -s=y -p=s'. The best-fitting empirical model of amino acid substitution for each alignment was found under the minimum AICc criterion from those implemented in RAxMLHPC[68] v7.2.8, and the maximum-likelihood phylogeny found using the best model for each alignment as a partitioned analysis in RAxMLHPC, using the default rapid heuristic algorithm, and with clade support estimated from 1,000 bootstrap samples. The pattern of gene family gains and losses on this tree was inferred under the dollo parsimony algorithm using the dollop program from the v3.69 of the Phylip package[102].

The phylogenetic tree for DNase II-like proteins was created based on Interpro-predicted DNase II protein domains (IPR004947). For *Trichuris* spp. such protein domains were identified in our *Trichuris* gene sets, and for select other taxa relevant proteins were identified and downloaded from the Interpro website (http://www.ebi.ac.uk/interpro/entry/IPR004947) and the DNase II domains extracted according to the predicted domain boundaries. DNase II domains shorter than 50 amino acids or longer than 400 amino acids were discarded. The tree was thus based on the DNase II domains of 18 *T. muris* proteins (20 domains), 15 *T. trichiura* proteins, 166 *T. spiralis* proteins (178 domains), 17 proteins of other nematodes, 5 proteins of other invertebrates, 8 human proteins, 5 mouse proteins, and 6 other vertebrate proteins. A multiple protein sequence alignment was created with MAFFT[66] v6.857 employing the –auto parameter. Alignment positions with more than 50% gaps were discarded, followed by removal of DNase II domain sequences that contained more than 50% gaps. A maximum-likelihood tree was calculated with RAxMLHPC[68] v7.7.2 using the WAG amino acid matrix, optimization of substitution rates, a GAMMA model of rate heterogeneity, and an estimation of the proportion of invariable sites. Clade support was estimated based on 500 bootstrap replicates. The resulting tree was visualized using FigTree v1.4.0 (http://tree.bio.ed.ac.uk/software/figtree/).

Chromosome-level analysis

Assigning chromosomal linkage groups by gene orthology

RATT[89] was used to transfer gene models from *T. muris* genome assembly v2.1 to assembly v4. For the 17 largest scaffolds of *T. muris* (assembly v4) and the 11 largest scaffolds of *T. spiralis*, the numbers of one-to-one orthologs - as identified by Inparanoid[64] v4.1 - between the *T. muris* and the *T. spiralis* scaffolds were counted and the resulting data subjected to clustering by hclust in R, which identified 3 chromosomal linkage groups for each species (Fig. 1a). The linkage group assignments made for *T. spiralis* were then used to assign to linkage groups all scaffolds of *T. muris* that were linked by at least 3 one-to-one orthologs to the corresponding *T. spiralis* genome scaffold (Supplementary Table 1a). As a result, 48 *T. muris* scaffolds representing 90.6% of the assembly could be assigned to one of the three linkage groups. Similarly, 483 scaffolds representing 61.0% of the *T. trichiura* genome assembly v2.1 could be assigned to one of the three chromosomal linkage groups based on one-to-one orthologs with *T. muris* (Supplementary Table 1b).

Calculating read coverage and heterozygosity

Relative read coverage and heterozygosity per chromosomal linkage group (Fig. 1b) were determined by first mapping Illumina data against the relevant genome assembly using SMALT v0.7.4 (http://www.sanger.ac.uk/resources/software/smalt/) employing an exhaustive search (-x) and repetitive mapping (-r) with parameters wordlen=13 (-k), skipstep=2 (-s), minid=0.75 (-y), and insertmax=1000 (-i). Absolute read coverage was determined by running the genomecov command of BEDTools[69] v2.17.0 over a BAM alignment file, followed by calculating both median and mean read coverage per genomic scaffold. Relative read coverage was determined by dividing the absolute median or mean read coverage of a scaffold by the absolute median or mean read coverage of all scaffolds assigned to linkage groups 1 and 2 (which represent the two autosomes). For the plots in Fig. 1b, binned read coverage was calculated for windows of 10kb, discarding genomic windows shorter than 5kb. A pileup including base and variant calling was generated from the read alignment by SAMtools mpileup. The number of heterozygous sites was determined per 10kb window of genomic sequence based on the genotype tag (GT) in the VCF file, filtering for a genotype quality (GQ) of >90 (phred-scaled). Genomic windows shorter than 5kb (e.g. at the end of scaffolds) were not included in the analysis.

Read coverage to infer chromosomal location and estimate sequence lengths

Read coverage, and in particular gender-specific read coverage, was used to help infer the chromosomal location (i.e. autosomal, X chromosomal, Y chromosomal, shared female/male, or centromeric) that genomic scaffolds represent (Supplementary Table 1c). To do so we analysed both median and mean read coverage over genomic scaffolds. **Median read coverage** over a scaffold is a measure that is more resistant to occasional outliers such as low coverage in SNP-dense regions and high coverage in regions of the assembly containing **"collapsed repeats"** (i.e. regions that occur as multiple near-identical repeat units in the actual parasite genome but are represented by fewer or even only a single such repeat unit in the genome assembly, leading to an artifactual pileup of sequence reads over such collapsed regions of the genome assembly). Median read coverage is therefore an appropriate and accurate metric for scaffolds whose sequence is well resolved and corresponds correctly to the actual parasite genome. In contrast, **mean read coverage** is a more appropriate metric for the analysis of scaffolds that contain a high proportion of collapsed repeats, which also often leads to highly uneven read coverage across a scaffold. In such cases, the median read coverage may lead to spurious results. In addition, the read coverage mean (as opposed to the median) is the more adequate metric to quantitatively reflect the actual DNA content in the parasite represented by a genomic scaffold, and may therefore be used to estimate the true, "uncollapsed" length of genomic sequence. A summary of such estimated chromosomal lengths is provided in Supplementary Table 1d, resulting in an estimated haploid female genome size for *T. muris* of 106.01 Mb. Read coverage was visualized by directly showing mapped Illumina read coverage over the genome sequence using Artemis[103] (see Supplementary Fig. 2). Scaffolds with particularly low read coverage (i.e. if the sum of relative median read coverage in females and males - which would be expected to be 200% for the autosomes, 150% for the X chromosome, and 100% for the Y chromosome - was <75%) were excluded from further analysis.

X chromosome

In males, the scaffolds of one genomic linkage group (as defined by gene orthology, see above) consistently showed a 50% reduced median read coverage compared to the median read coverage observed in scaffolds assigned to the other two linkage groups in both *T. muris* (absolute male median read coverage of 79 vs 159, Supplementary Table 1a) and *T. trichiura* (absolute male median read coverage of 97 vs 193, Supplementary Table 1b), whereas in *T. muris* females the median read coverage remained essentially the same (absolute female median read coverage of 160 vs 161, Supplementary Table 1a) - thereby allowing this linkage group to be identified as representing the X chromosome. Scaffolds that could not be assigned to a linkage group based on gene orthology were also inferred to belong to the X chromosome if (1) their median read coverage in females was the same as the median autosomal read coverage (i.e. if the relative median coverage was between 0.8 and 1.2), and if (2) their relative median read coverage in males was reduced by about 50% (i.e. if the ratio of relative median read coverage in male / females was between 0.4 and 0.6) (Supplementary Table 1c).

Shared female/male

Other scaffolds that could not be assigned to a linkage group based on gene orthology showed approximately equal read coverage in both females and males of *T. muris*. Such sequences could either derive from one of the autosomes or be located on both the X and the Y chromosome. Scaffolds that were not already assigned to a linkage group based on gene orthology or assigned to the X chromosome based on coverage (see above) were therefore labeled "shared female/male" if they showed a ratio of both relative median and relative mean read coverage for male vs. females of between 0.8 and 1.2 (Supplementary Table 1c).

Y chromosome

Scaffolds for which the relative mean read coverage in males was greater than 3 times the relative mean read coverage in females were inferred to represent the Y chromosome (Supplementary Table 1c). This rule assigns 179 genomic scaffolds of *T. muris* to the Y chromosome, with 166 of these scaffolds (92.7%) comprising significant amounts of collapsed repetitive content (defined as scaffolds where the sum of relative mean read coverage in females and males - which would be expected to be 200% for the autosomes, 150% for the X chromosome, and 100% for the Y chromosome - was >300%). Due to this extraordinarily high proportion of repetitive genomic content, the estimated (based on read coverage) "uncollapsed" length of the Y chromosome of 24.42 Mb (Supplementary Table 1d) is significantly larger than the simple sum of the corresponding 179 scaffolds of the assembly (0.64 Mb). In fact, this estimated length of the Y chromosome is nearly as large as the estimated lengths of the X chromosome (27.00 Mb) and the two autosomes (31.96 Mb and 25.70 Mb), which is in good agreement with published data[7].

Centromeres

Three genomic scaffolds of *T. muris* with particularly high read coverage (TMUE_000352, TMUE_000164, TMUE_000165) exhibited a repeat structure suggestive of centromeric sequences (Supplementary Fig. 3), with repeat units of both 164 bp and 176 bp length, which is very similar to the length of the ~171 bp-long monomers of human centromeric alpha-satellite DNA[104]. Together, these putative centromeric sequences are estimated to comprise approximately 5.33 Mb, which corresponds to 5.0% of the *T. muris* genome (Supplementary Table 1d). The repeat structure of these genomic scaffolds is illustrated by dot-plots that were created with YASS[105] at http://bioinfo.lifl.fr/yass/ and by a multiple sequence aligment that was generated with MAFFT[66] v6.85 at http://www.ebi.ac.uk/Tools/msa/mafft/ and visualised using Jalview[106] v2.8 at http://www.jalview.org/ (Supplementary Fig. 3).

Gene expression analysis - *T. muris*

Paired-end Illumina reads derived from transcriptome sequencing (Supplementary Table 16a) were mapped to the *T. muris* genome sequence using TopHat[71] v1.4.1. The number of reads per gene was determined with BEDTools[69] v2.10.1 and calculated by summing the raw reads over all exons of a gene (*T. muris* gene set v2.2). Differential gene expression analysis between different biological samples was carried out using the Bioconductor package edgeR[73] v3.2.4 and by calculating tagwise dispersion followed by the exact test for differential expression. Genes represented by fewer than two counts per million (CPM) in at least three of the 20 samples were discarded, yielding expression data for 9,858 of 11,004 *T. muris* genes. For the differential expression analysis for the pairwise comparison of *T. muris* anterior versus mixed rear, "mixed rear" was defined as comprising samples "female rear" and "male rear". After having provided edgeR with the data of all parasite RNAseq samples at once and having carried out normalization for library size (calcNormFactors()), the library size-adjusted read counts ("pseudo-counts" in edgeR) were divided by gene length to yield "normalised counts per kb of gene length", which were used to compare transcript-level expression of genes between samples.

GO term enrichment analysis was performed with the Bioconductor package topGO[75] v2.12.0, selecting a minimum node size of 5 and the "classic" algorithm under the Fisher statistic. Protein domain enrichment analysis was based on the Interproscan v5.0.7 results (see above). Protein domain predictions were included if the E-value was smaller than 0.01 (Pfam and Gene3D) or 0.05 (SMART), while for other database searches E-values were not available. The number of proteins with a given protein domain (i.e. a protein with multiple copies of the same domain was counted only once) were then counted in the results of the differential expression analysis carried out with edgeR, accepting a transcript as differentially expressed in a given pairwise comparison if the following conditions were met: FDR <= 1E-5 and fold-change >= 2.0. P-values for the statistical enrichment of a given protein domain in a given differential expression comparison were calculated in R using Fisher's exact test (two-sided) and were then multiplied by the number of tests carried out for the results of a given protein domain database to yield an adjusted P-value (Supplementary Table 8).

Identification of novel drug targets

In order to do a bioinformatics ranking of all proteins in *T. muris* (gene set v2.3) for their suitability as a drug target (Supplementary Tables 13, 14), we used the following information:

1. ***Trichuris* expression** data, i.e. edgeR-normalised expression values per kb of gene length as described above.

2. **Orthology** of *T. muris* protein sequences to those in *T. trichiura* was identified by Inparanoid (see above). Orthology to *C. elegans*, mouse, human and drug targets was determined using a stand-alone version of OMA[65] v0.99t.

3. Protein homolog **essentiality** in mouse and *C. elegans*: mouse phenotype data was retrieved from Mouse Genome Informatics (http://www.informatics.jax.org) downloaded through the "Genes and Markers Query Form", and *C. elegans* data were retrieved from Wormbase using WormMart (Wormbase.org). Only lethal phenotypes were selected.

4. Whether a protein is predicted to be an **enzyme**, as evidenced by a KEGG orthology (KO) identifier.

5. **Druggability** information from ChEMBL[107]. Druggability of potential targets was assessed using ChEMBL Ensemble scores which were determined as follows: all Interproscan predictions for *T. muris* proteins with E-value <= 0.001 and associated Interpro accession number were collected, representing predictions by Pfam, SMART, TIGRFAM, Gene3D, and PIRSF. All Protein Data Bank (PDB) entries associated with the Interpro domains were downloaded from Interpro (e.g. http://www.ebi.ac.uk/interpro/entry/IPR001031/structures) and the maximum Ensemble score per PDB accession was extracted from table "domain_drugebility.txt" v2.0 from ChEMBL (ftp://ftp.ebi.ac.uk/pub/databases/chembl/DrugEBIlity/releases/2.0/). The maximum ChEMBL Ensemble score of all PDB entries associated with an Interpro domain was used as this Interpro domain's Ensemble score. Finally, the Ensemble score of a protein was determined as the maximum Ensemble score of any of its predicted protein domains.

6. DrugBank[51] (http://www.drugbank.ca/downloads): drug target sequences for "All drug targets" (n=3985) and "Approved drug targets" (n =1479). The drugs associated with drug targets were extracted from Drugbank (e.g. http://www.drugbank.ca/molecules/1295) and filtered for "approved" drugs while nutraceuticals were excluded. Orthologs were extracted between DrugBank and the *Trichuris* species using OMA[65] v0.99t.

7. Therapeutic Targets Database[108] sequence data: for all targets (n=1502), and sequence data for successful targets (n=334) and TTD targets information (http://bidd.nus.edu.sg/group/cjttd/TTD_Download.asp). Orthologs were extracted between TTD and the *Trichuris* species using OMA[65] v0.99t.

8. The results were manually inspected, nutraceutical targets filtered out, and further investigated using literature searches.

Transcriptome sequencing - mouse

Fourteen male C57BL/6 mice 6-8 weeks of age (purchased from Harlan Olac) were infected with approximately 25 *T. muris* eggs by oral gavage and were killed 42 days post infection alongside uninfected controls. C57BL/6 mice are susceptible to a low dose *T. muris* infection, resulting in chronic infection. The samples taken from the mice were as follows. Mesenteric lymph node, MLN, a section of cecum where the worms reside and a section of cecum where there were no worms. These were termed “wormy” and “non wormy cecum”. In half of the “wormy cecum” samples the worms were left “in situ” (n=4), and in half the worms were removed (n=4). The same samples were generated from uninfected controls resulting in samples termed, for example, uninfected wormy cecum. Serum was also taken from these mice to confirm infection status of the mice by ELISA. We examined tissues from the cecum and mesenteric lymph node (MLN) in naïve and infected mice to determine the response to infection. Specifically we looked at “wormy” cecum, where worms preferentially bind in a low dose infection and “non-wormy” cecum where they do not. Each of these was considered in both naïve and infected mice. As an additional control we sequenced both infected wormy cecum with the worm removed, as well as with the worm in.

Tissue samples were removed from RNAlater (Invitrogen) and washed once in 1X PBS. The material was mechanically homogenized in 1ml TRIzol (Invitrogen). After the addition of 200 ul chloroform with isoamyl alcohol (24:1), the aqueous phase was aspirated and 1 volume of 70% ethanol was added. The sample was added to an RNeasy Mini spin column (Qiagen), washed and eluted according to the manufacturer's instructions. Quantity and quality of the RNA was assessed using the Agilent Bioanalyzer. Transcriptome libraries (Supplementary Table 16b) were made using the TruSeq kit and sequencing method as described above ('Transcriptome sequencing – *T. muris*').

Gene expression analysis - mouse

A combined reference of mouse (mm10) and *T. muris* transcripts (*T. muris* gene set v2.2) was prepared by extracting transcript sequences from the genome annotation. Paired-end Illumina reads were mapped to the reference using Bowtie2 [109] and the effective numbers of reads per transcript were enumerated using eXpress[72]. The number of reads per gene was then calculated by summing over all transcripts related to each gene. DESeq[74] was used to determine the reliability of our replicates and found that one infected MLN sample was an outlier (ERS167950). This was removed from further analysis. DESeq was then used to determine genes differentially expressed between different conditions. A false discovery rate of 5% was applied except where stated. GO terms enriched in differentially expressed genes were determined using innateDB[76] and TopGO[75]. Differences between naïve and infected cecum were determined using naïve samples (ERS167954, ERS167966, ERS167978, ERS167957, ERS167969 and ERS167987) and wormy samples with the worm left in (ERS167948, ERS167960, ERS167972, ERS167981). Reads mapping to *T. muris* transcripts and those mapping ambiguously were removed prior to analysis. Differences between naïve and infected MLN were determined using samples ERS167956, ERS167968, ERS167980, ERS167959, ERS167971, ERS167989 and ERS167962, ERS167974, ERS167983, ERS167953, ERS167965, ERS167977, ERS167986 respectively. A comparison was done between non-wormy infected and wormy infected cecum to show that the host immune response is not localized to the site of infection using non-wormy samples ERS167949, ERS167961, ERS167973, ERS167982, ERS167952, ERS167964, ERS167976, ERS167985 and wormy samples ERS167948, ERS167960, ERS167972 and ERS167981. In each case sequences from multiple lanes were combined for individual samples prior to analysis.

GWAS analysis

We looked for whether genes that are associated with immune-mediated diseases are also enriched for those that were differentially expressed in the cecum of infected vs. uninfected mice. Lists of associated loci from published genome-wide association studies (GWAS) were extracted for four immune-mediated complex diseases: Crohn's disease, ulcerative colitis, celiac disease and type 1 diabetes, as well as two complex traits: height and body mass index, where immune-related genes are unlikely to play a major role. Testing for enrichment was performed using a Monte Carlo simulation approach adapted from Raine et al.[110], which accounts for linkage disequilibrium between associated SNPs and non-random arrangement of functionally related genes within the genome.

We filtered the list of all genes that were tested for differential expression down to those with a unique human ortholog using Ensembl. We also excluded those not labeled as protein-coding in Gencode[111] v17, and not located on human autosomes (GWAS often do not include the sex chromosomes). Of the 15,278 genes that remained, 574 were differentially expressed. For each disease/trait, autosomal SNPs that exceeded genome-wide significance ($p < 5e-8$) in the GWAS with the largest sample size using a European population were extracted[112-116]. For each differentially expressed gene, we defined a gene-region spanning +/-50kb from the transcript start/stop site. To account for the non-random clustering of genes with similar expression patterns and function[117], groups of differentially expressed genes that have overlapping +/-50kb windows were combined into single gene windows. In total, 454 gene windows were constructed from the original list of 574 differentially expressed genes. For each disease/trait, an associated locus was defined as the genomic region spanning a 0.1cM window either side of the associated SNP. Recombination rates were obtained using data from the 1000 Genomes Project[118]. Where multiple SNPs showed overlapping windows, only the window assigned to the SNP with the most significant p-value was considered. We then counted the number of associated loci that overlap at least one differentially expressed gene-window. To assess the statistical significance of this overlap, we randomly sampled 454 genes from the full list of 15,278 genes, while ensuring that if a sampled gene has a +/-50kb window overlapping that of another previously sampled gene, then the windows are merged and these genes are only counted once. We then calculated the number of associated loci that overlap at least one of these randomly sampled lists of genes. The sampling process was repeated 100,000 times for each disease/trait, and the empirical p-value was the number times the overlap with the randomly sampled genes exceeds the overlap with the observed differentially expressed genes, divided by 100,000.

References

78. Kozarewa, I. *et al.* Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat Methods* 6, 291-5 (2009).

79. Bentley, D.R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456, 53-9 (2008).

80. Quail, M.A. *et al.* A large genome center's improvements to the Illumina sequencing system. *Nat Methods* 5, 1005-10 (2008).

81. Bonfield, J.K. & Whitwham, A. Gap5--editing the billion fragment sequence assembly. *Bioinformatics* 26, 1699-703 (2010).

82. Otto, T.D., Sanders, M., Berriman, M. & Newbold, C. Iterative Correction of Reference Nucleotides (iCORN) using second generation sequencing technology. *Bioinformatics* 26, 1704-7 (2010).

83. Smith, C.D. *et al.* Improved repeat identification and masking in Dipterans. *Gene* 389, 1-9 (2007).

84. Punta, M. *et al.* The Pfam protein families database. *Nucleic Acids Research* 40, D290-D301 (2012).

85. Smit, A.F.A. & Hubley, R. RepeatModeler Open-1.0. (2008-2010).

86. Smit, A.F.A., Hubley, R. & Green, P. RepeatMasker Open-3.0. (1996-2010).

87. She, R. *et al.* genBlastG: using BLAST searches to build homologous gene models. *Bioinformatics* 27, 2141-3 (2011).

88. Yook, K. *et al.* WormBase 2012: more genomes, more data, new website. *Nucleic Acids Res* 40, D735-41 (2012).

89. Otto, T.D., Dillon, G.P., Degrave, W.S. & Berriman, M. RATT: Rapid Annotation Transfer Tool. *Nucleic Acids Res* 39, e57 (2011).

90. Altschul, S.F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25, 3389-402 (1997).

91. Slater, G.S. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* 6, 31 (2005).

92. Nakamura, Y., Cochrane, G. & Karsch-Mizrachi, I. The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Res* 41, D21-4 (2013).

93. UniProtConsortium. Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Res* 41, D43-7 (2013).

94. Logan-Klumpler, F.J. *et al.* GeneDB--an annotation database for pathogens. *Nucleic Acids Res* 40, D98-108 (2012).

95. Kall, L., Krogh, A. & Sonnhammer, E.L. A combined transmembrane topology

and signal peptide prediction method. *J Mol Biol* 338, 1027-36 (2004).

96. Letunic, I., Doerks, T. & Bork, P. SMART 6: recent updates and new developments. *Nucleic Acids Res* 37, D229-32 (2009).

97. Hunter, S. *et al.* InterPro: the integrative protein signature database. *Nucleic Acids Res* 37, D211-5 (2009).

98. Crooks, G.E., Hon, G., Chandonia, J.M. & Brenner, S.E. WebLogo: a sequence logo generator. *Genome Res* 14, 1188-90 (2004).

99. Grutter, M.G., Fendrich, G., Huber, R. & Bode, W. The 2.5 A X-ray crystal structure of the acid-stable proteinase inhibitor from human mucous secretions analysed in its complex with bovine alpha-chymotrypsin. *EMBO J* 7, 345-51 (1988).

100. Rawlings, N.D. & Morton, F.R. The MEROPS batch BLAST: a tool to detect peptidases and their non-peptidase homologues in a genome. *Biochimie* 90, 243-59 (2008).

101. Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A.C. & Kanehisa, M. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res* 35, W182-5 (2007).

102. Felsenstein, J. PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics* 5(1989).

103. Rutherford, K. *et al.* Artemis: sequence visualization and annotation. *Bioinformatics* 16, 944-5 (2000).

104. Alkan, C. *et al.* Organization and evolution of primate centromeric DNA from whole-genome shotgun sequence data. *PLoS Comput Biol* 3, 1807-18 (2007).

105. Noe, L. & Kucherov, G. YASS: enhancing the sensitivity of DNA similarity search. *Nucleic Acids Res* 33, W540-3 (2005).

106. Waterhouse, A.M., Procter, J.B., Martin, D.M., Clamp, M. & Barton, G.J. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25, 1189-91 (2009).

107. Gaulton, A. *et al.* ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res* 40, D1100-7 (2012).

108. Liu, X. *et al.* The Therapeutic Target Database: an internet resource for the primary targets of approved, clinical trial and experimental drugs. *Expert Opin Ther Targets* 15, 903-12 (2011).

109. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9, 357-9 (2012).

110. Raine, T., Liu, J.Z., Anderson, C.A., Parkes, M., & Kaser, A. Generation of primary human intestinal T cell transcriptomes reveals differential expression at genetic risk loci for immune-mediated disease. *Gut* Online First: 5 May

2014. doi:10.1136/gutjnl-2013-306657

111. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* 22, 1760-74 (2012).

112. Jostins, L. *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* 491, 119-24 (2012).

113. Trynka, G. *et al.* Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nat Genet* 43, 1193-201 (2011).

114. Bradfield, J.P. *et al.* A genome-wide meta-analysis of six type 1 diabetes cohorts identifies multiple associated loci. *PLoS Genet* 7, e1002293 (2011).

115. Lango Allen, H. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* 467, 832-8 (2010).

116. Speliotes, E.K. *et al.* Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat Genet* 42, 937-48 (2010).

117. Hurst, L.D., Pal, C. & Lercher, M.J. The evolutionary dynamics of eukaryotic gene order. *Nat Rev Genet* 5, 299-310 (2004).

118. Genomes Project, C. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* 491, 56-65 (2012).