Genome Biology

Extensive genomic and transcriptomic variation defines the chromosome-scale assembly of Haemonchus contortus, a model gastrointestinal worm

--Manuscript Draft--

Manuscript Number:	
Full Title:	Extensive genomic and transcriptomic variation defines the chromosome-scale assembly of Haemonchus contortus, a model gastrointestinal worm
Article Type:	Research
Funding Information:	Wellcome Trust (067811) — Not applicable Wellcome Trust (WT206194) — Not applicable Biotechnology and Biological Sciences Research Council (BB/M003949/1) — Not applicable Biotechnology and Biological Sciences Research Council (BB/P024610/1) — Not applicable Biotechnology and Biological Sciences Research Council (BB/K020048/1) — Not applicable
Abstract:	Background Haemonchus contortus is a globally distributed and economically important gastrointestinal pathogen of small ruminants, and has become the key nematode model for studying anthelmintic resistance and other parasite-specific traits among a wider group of parasites including major human pathogens. Two draft genome assemblies for H. contortus were reported in 2013, however, both were highly fragmented, incomplete, and differed from one another in important respects. While the introduction of long-read sequencing has significantly increased the rate of production and contiguity of de novo genome assemblies broadly, achieving high quality genome assemblies for small, genetically diverse, outcrossing eukaryotic organisms such as H. contortus remains a significant challenge. Results Here, we report using PacBio long read and OpGen and 10X Genomics long-molecule methods to generate a highly contiguous 283.4 Mbp chromosome-scale genome assembly including a resolved sex chromosome. We show a remarkable pattern of almost complete conservation of chromosome content (synteny) with Caenorhabditis elegans , but almost no conservation of gene order. Long-read transcriptome sequence data has allowed us to define coordinated transcriptional regulation throughout the life cycle of the parasite, and refine our understanding of cis- and trans -splicing relative to that observed in C. elegans . Finally, we use this assembly to give a comprehensive picture of chromosome-wide genetic diversity both within a single isolate and globally. Conclusions The H. contortus MHco3(ISE).N1 genome assembly presented here represents the most contiguous and resolved nematode assembly outside of the Caenorhabditis genus to date, together with one of the highest-quality set of predicted gene features. These data provide a high-quality comparison for understanding the evolution and genomics of Caenorhabditis and other nematodes, and extends the experimental tractability of this model parasitic nematode in understanding pathogen biology, drug discovery and vaccine development, and important adaptive traits such as drug resistance.
Corresponding Author:	Stephen R Doyle, PhD Wellcome Trust Sanger Institute

	Hinxton, Cambridgeshire UNITED KINGDOM
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Wellcome Trust Sanger Institute
Corresponding Author's Secondary Institution:	
First Author:	Stephen R Doyle, PhD
First Author Secondary Information:	
Order of Authors:	Stephen R Doyle, PhD
	Alan Tracey
	Roz Laing
	Nancy Holroyd
	David Bartley
	Wojtek Bazant
	Helen Beasley
	Robin Beech
	Collette Britton
	Karen Brooks
	Kirsty Maitland
	Axel Martinelli
	Jennifer D Noonan
	Michael Paulini
	Michael A Quail
	Elizabeth Redman
	Faye H Rodgers
	Guillaume Sallé
	Geetha Sankaranarayanan
	Kevin L Howe
	Neil Sargison
	Eileen Devaney
	Matthew Berriman
	John S Gilliard
	James A Cotton
Order of Authors Secondary Information:	
Suggested Reviewers:	Neil Young, PhD Senior Research Fellow, The University of Melbourne, Australia nyoung@unimelb.edu.au Extensive helminth genomics experience/expertise
	Tim Geary, PhD Professor, Queen's University Belfast, UK T.Geary@qub.ac.uk Expert on parasitic nematode biology and anthelminthics
	Erik Andersen, PhD

	Associate Professor, Northwestern University, USA Erik.Andersen@northwestern.edu C. elegans population genomicist
	Christian Rödelsperger, PhD Departmental Project Leader, Max Planck Institute for Developmental Biology, Germany christian.roedelsperger@tuebingen.mpg.de Nematode evolutionary biologist focused on bioinformatics / genomics
	Erich Schwarz, PhD Assistant Research Professor, Cornell University, USA ems394@cornell.edu Functional genomicist, with interests in C. elegans and other nematodes. Lead author on draft McMaster strain Haemonchus contortus genome assembly published in 2013.
	Kirsten Gunsalus, PhD Professor of Biology; NYU-AD Faculty Director of Bioinformatics, New York University, USA kcg1@nyu.edu Uses functional genomics and computational approaches to study development in C. elegans and related nematodes
	Vicky Hunt, PhD Group Leader, The University of Bath, UK bs1vlh@bath.ac.uk Comparative and functional genomics of Caenorhabditids and parasitic nematodes
Additional Information:	
Question	**Response**
Has this manuscript been submitted to this journal before?	No
Experimental design and statistics Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist. Information essential to interpreting the data presented should be made available in the figure legends. Have you included all the information requested in your manuscript?	Yes
Resources A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite Research Resource	Yes

Identifiers (RRIDs) for antibodies, model organisms and tools, where possible. Have you included the information requested as detailed in our Minimum Standards Reporting Checklist " target="_blank">Minimum Standards Reporting Checklist?	

Extensive genomic and transcriptomic variation defines the chromosome-scale assembly of *Haemonchus contortus*, a model gastrointestinal worm

Stephen R. Doyle[1*], Alan Tracey[1], Roz Laing[2], Nancy Holroyd[1], David Bartley[3], Wojtek Bazant[1], Helen Beasley[1], Robin Beech[4], Collette Britton[2], Karen Brooks[1], Kirsty Maitland[2], Axel Martinelli[1], Jennifer D. Noonan[4], Michael Paulini[5], Michael A. Quail[1], Elizabeth Redman[6], Faye H. Rodgers[1], Guillaume Sallé[7], Geetha Sankaranarayanan[1], Kevin L. Howe[5], Neil Sargison[8], Eileen Devaney[2], Matthew Berriman[1], John S. Gilleard[6], James A. Cotton[1*]

*Corresponding authors

1. **Wellcome Sanger Institute, Hinxton, Cambridgeshire, CB10 1SA, United Kingdom**
Stephen R. Doyle: stephen.doyle@sanger.ac.uk; Alan Tracey: alt@sanger.ac.uk; Nancy Holroyd: neh@sanger.ac.uk; Wojtek Bazant: wojtek.bazant@sanger.ac.uk; Helen Beasley: hb521@cam.ac.uk; Karen Brooks: kd1@sanger.ac.uk ; Axel Martinelli: axel.martinelli@gmail.com; Michael A. Quail: mq1@sanger.ac.uk; Faye Rodgers: fr7@sanger.ac.uk; Geetha Sankaranarayanan: gs13@sanger.ac.uk; Matthew Berriman: mb4@sanger.ac.uk; James A. Cotton: jc17@sanger.ac.uk

2. **Institute of Biodiversity Animal Health and Comparative Medicine, College of Medical, Veterinary and Life Sciences, University of Glasgow, Garscube Campus, Glasgow, G61 1QH, United Kingdom**
Roz Laing: Rosalind.Laing@glasgow.ac.uk; Collette Britton: Collette.Britton@glasgow.ac.uk; Kirsty Maitland: Kirsty.Maitland@glasgow.ac.uk; Eileen Devaney: Eileen.Devaney@glasgow.ac.uk

3. **Moredun Research Institute, Pentlands Science Park, Bush Loan, Penicuik EH26 0PZ, United Kingdom**
David Bartley: Dave.Bartley@moredun.ac.uk

4. **Institute of Parasitology, McGill University, 21111 Lakeshore Road, Sainte Anne-de-Bellevue, Québec, H9X3V9 Canada**
Robin Beech: robin.beech@mcgill.ca; Jennifer D. Noonan: jennifer.noonan@mail.mcgill.ca

5. **European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, Cambridgeshire, CB10 1SA, United Kingdom**
Michael Paulini: mh6@ebi.ac.uk; Kevin Howe: klh@ebi.ac.uk

6. **Department of Comparative Biology and Experimental Medicine, Faculty of Veterinary Medicine, University of Calgary, Calgary, Alberta, Canada**
Elizabeth Redman: elmredma@ucalgary.ca; John Gilleard: jsgillea@ucalgary.ca

7. **INRAE - U. Tours, UMR 1282 ISP Infectiologie et Santé Publique, Centre de recherche Val de Loire, Nouzilly, France**
Guillaume Salle: guillaume.salle@inrae.fr

8. **Royal (Dick) School of Veterinary Studies, University of Edinburgh, Edinburgh, EH25 9RG, United Kingdom**
Neil Sargison: Neil.Sargison@ed.ac.uk

Abstract

Background *Haemonchus contortus* is a globally distributed and economically important gastrointestinal pathogen of small ruminants, and has become the key nematode model for studying anthelmintic resistance and other parasite-specific traits among a wider group of parasites including major human pathogens. Two draft genome assemblies for *H. contortus* were reported in 2013, however, both were highly fragmented, incomplete, and differed from one another in important respects. While the introduction of long-read sequencing has significantly increased the rate of production and contiguity of *de novo* genome assemblies broadly, achieving high quality genome assemblies for small, genetically diverse, outcrossing eukaryotic organisms such as *H. contortus* remains a significant challenge.

Results Here, we report using PacBio long read and OpGen and 10X Genomics long-molecule methods to generate a highly contiguous 283.4 Mbp chromosome-scale genome assembly including a resolved sex chromosome. We show a remarkable pattern of almost complete conservation of chromosome content (synteny) with *Caenorhabditis elegans*, but almost no conservation of gene order. Long-read transcriptome sequence data has allowed us to define coordinated transcriptional regulation throughout the life cycle of the parasite, and refine our understanding of *cis*- and *trans*-splicing relative to that observed in *C. elegans*. Finally, we use this assembly to give a comprehensive picture of chromosome-wide genetic diversity both within a single isolate and globally.

Conclusions The *H. contortus* MHco3(ISE).N1 genome assembly presented here represents the most contiguous and resolved nematode assembly outside of the *Caenorhabditis* genus to date, together with one of the highest-quality set of predicted gene features. These data provide a high-quality comparison for understanding the evolution and genomics of *Caenorhabditis* and other nematodes, and extends the experimental tractability of this model parasitic nematode in understanding pathogen biology, drug discovery and vaccine development, and important adaptive traits such as drug resistance.

Keywords

Haemonchus contortus; chromosomal genome assembly; PacBio long-read sequencing; IsoSeq cDNA sequencing; genetic variation; transcriptomics; population genomics

Background

A complete sequence assembly that reflects the biologically defined chromosomal DNA sequence of an organism is the ultimate goal of a *de novo* genome assembly project. Until recently, this goal was unattainable for all but the smallest of genomes or for select organisms, where genome assembly projects have been supported by large consortia and extensive financial and logistical support, for example, the model nematode *Caenorhabditis elegans* [1]. However, high-throughput sequencing has significantly impacted the rate at which draft genomes are produced and, together with the recent introduction of long-read sequencing from Pacific Biosciences (PacBio) and Oxford Nanopore, the quality and contiguity of assemblies of both large and small genomes has completely changed.

Now, the ability to generate highly contiguous and closed prokaryote genomes is becoming routine, even from metagenomic samples containing multiple strains or species [2,3], and high-quality chromosome-scale genomes for a number of eukaryotic species are rapidly becoming available [4–6]. These technologies generally require microgram quantities of high molecular weight DNA, and the success of a genome assembly is highly dependent on the degree of polymorphism present in the DNA that is sequenced. This can be minimised by sequencing DNA from a single individual – so that the only polymorphism is heterozygosity between homologous chromosomes in the case of a diploid or polyploid individual – or from a clonal population, where multiple but genetically identical individuals can be pooled. However, for many small organisms, this is not possible; the only option is to sequence DNA from a pool of genetically distinct organisms and derive a consensus assembly. As modern assembly algorithms are designed and trained on haploid or diploid genomes from single individuals [7–9], the excessive diversity presents a significant challenge to the assembly process, and consequently, typically results in fragmented

assemblies that contain misassembled and haplotypic sequences.

Parasitic helminths are a diverse group of organisms for which significant efforts have been made to develop genomic resources [10]. For some species, including *Onchocerca volvulus* [11], *Strongyloides ratti [12]*, *Taenia multiceps* [13], and *Echinococcus multilocularis* [14], high-quality near-complete assemblies have been achieved using multiple complementary sequencing and mapping technologies, and usually with extensive manual improvement work. Similar quality assemblies for a few other species (S*chistosoma mansoni, Brugia malayi, Hymenolepis microstoma,* and *Trichuris muris*) are available from WormBase ParaSite [15], but await formal description in the peer-reviewed literature. However, for many others, access to sufficient high-quality DNA, small organismal size, and high genetic diversity are significant challenges that need to be overcome to achieve highly contiguous assemblies. The nematode *Haemonchus contortus* is one such species. *H. contortus* is a major pathogen of sheep and goats worldwide and is recognised as a model parasite due to its experimental tractability for drug discovery [16,17], vaccine development [18,19] and anthelmintic resistance research [20]. Draft genome assemblies were published in 2013 for two anthelmintic-sensitive isolates, MHco3(ISE).N1 [21] and McMaster [22]; both assemblies were produced with a combination of short-read, high-throughput sequencing technologies (including both short- and long-insert libraries), resulting in estimated genome sizes of 370 Mbp and 320 Mbp, respectively. However, direct comparison of these two assemblies shows discordance, revealing clear differences in gene family composition [23] and variation in assembly quality and gene content [24]. Although some of these differences reflect true biological variation in this highly polymorphic species [25–27], many differences are technical artefacts, as revealed by comparing fragmented genome assemblies from

sequencing DNA derived from pools of a genetically diverse organism. These fragmented genome assemblies are also less than ideal substrates for many downstream uses of reference genomes, for example, in interpreting the results of genetic experiments in which patterns of genome-wide variation are defined and interpreted [28].

Here we present the chromosome-scale assembly of the MHco3(ISE).N1 isolate of *H. contortus,* representing the first high-quality assembly for any strongylid parasite, an important clade of parasitic nematodes that includes parasitic species of major veterinary and medical importance. Using a hybrid assembly approach incorporating short and long-read sequencing followed by manual finishing, we demonstrate how a highly contiguous assembly overcomes some of the critical limitations imposed by draft assemblies in interpreting large-scale genetic and functional genomics datasets. We provide insight into patterns of transcriptional change and putative co-regulation using a significantly improved genome annotation, derived *de novo* from both short- and long-read cDNA sequencing, and describe within- and between-population genetic diversity that shapes the genome. This manually refined chromosome-scale assembly now offers insight into genome evolution among a broad group of important parasite species and offers a robust scaffold for genome-wide analyses of important parasite traits such as anthelmintic resistance.

Results

Chromosome structure of *Haemonchus contortus*

We have built upon our previous assembly (version 1 (V1); [21]) using a hybrid approach, iteratively incorporating Illumina short-insert and 3 kbp libraries, PacBio long-read, OpGen optical mapping, and 10X Genomics linked-read data (**Additional file 1: Figure S1;** see **Additional file 2: Table S1** for new sequencing data generated) to generate a largely complete, chromosomal-scale genome assembly. Consistent with the karyotype for *H. contortus* [29,30], the core genome assembly consists of five autosomal scaffolds and one sex-linked scaffold (**Additional file 1: Figure S2** shows completion of the X chromosome from assembly V3 to V4, and the relative X-to-autosome genome coverage between a single male and female parasite), each containing terminal telomeric sequences, and a single mitochondrial contig. We assigned chromosome names based on synteny with *C. elegans* chromosomes (**Figure 1 A**); over 80% of the 7,361 one-to-one orthologous genes are shared between syntenic chromosomes between the two species demonstrating the high conservation of genes per chromosome (**Figure 1 B**; top), however, vast rearrangements are evident and very little conservation of gene order remains (**Figure 1 B**; bottom). We determined the extent of microsynteny by comparing conserved, reorientated, and recombined gene pairs between the two species (**Additional file 1: Figure S3 A**); the distance between ortholog pairs in *H. contortus* and *C. elegans* is correlated up to ~ 100 kbp (rho = 0.469, p-value = 2.2E-16) in which pair order is conserved (**Additional file 1: Figure S3 B**), likely representing selective constraint to maintain evolutionary conserved operons, but is lost above 100 kbp (rho = 0.017, p-value = 0.6595) where a greater frequency of recombination between pairs is evident (**Additional file 1: Figure S3 B,C**). Beyond pairs of genes, synteny breaks down rapidly; although almost 50% of shared ortholog pairs are

adjacent to each other, only a single group of 10 orthologs are colinear between the genomes of the two species (**Additional file 1: Figure S3 D**).

Figure 1. Chromosomal synteny between *Haemonchus contortus* and *Caenorhabditis elegans*

A. Genome-wide comparison of chromosomal organisation between *H. contortus* V4 and *C. elegans,* analysed using *promer* and visualised using *Circos*. **B.** Comparison of shared orthologs, based on assignment to chromosomes (top plot) and their relative position on the chromosome (bottom plot). In the top plot, chromosome assignment of 7,361 one-to-one orthologs between *C. elegans* and *H. contortus* V4 annotations demonstrates that the majority of genes on any given *H. contortus* chromosome are found on the same chromosome in *C. elegans*. Values within each panel represent the percentage of shared orthologs found on the same chromosome of both *H. contortus* and *C. elegans*. In the bottom plot, the relative genomic position of the one-to-one orthologs between *C. elegans* (y-axis) and *H. contortus* V4 chromosomes (x-axis) is shown. Complete within-chromosome synteny would be expected to show a positive linear relationship between the genomic positions of orthologs between *C. elegans* and *H. contortus* chromosomes; this is not observed, with an almost complete reshuffling of orthologs between the two species. Box headers are coloured by *H. contortus* chromosomes, whereas data points within the boxes are coloured by *C. elegans* chromosomes.

The assembly is approximately 283.4 Mbp in length (scaffold N50 = 47.3 Mbp; contig N50 = 3.8 Mbp), representing only 75.6% of the 369.8 Mbp V1 draft assembly length (**Table 1**); the reduction in assembly size is largely due to the identification and separation of redundant haplotypic sequences present in the polymorphic draft V1 and preliminary PacBio assemblies generated. The V4 assembly is highly resolved with only 185 gaps, a significant

reduction from the 41,663 gaps present in the V1 assembly (**Table 1**). As a measure of assembly completeness, we identified 242 of 248 universally conserved orthologs measured by the Core Eukaryotic Genes Mapping Approach (CEGMA; **Additional file 2: Table S2**), and 859 of 982 metazoan Benchmarking Universal Single-Copy Orthologs (BUSCOs). The remaining missing orthologs show phylogenetic structure among clade V nematodes (**Additional file 1: Figure S4 A**), suggesting that many are truly missing from the genomes of these species rather than due to assembly artefacts. These data contained 141 more full-length single-copy orthologs and 159 fewer duplicated BUSCOs than identified in the V1 genome (**Additional file 2: Table S2**); considering that we identified a similar average number of CEGMA orthologs per core gene in V4 relative to the *C. elegans* genome (1.1 in V4, compared with 1.09 in *C. elegans*), these data represent a significant reduction in erroneously-duplicated sequences in V4 compared with the draft *H. contortus* assemblies (V1: 85.89% complete and 1.59 average orthologs per CEGMA gene; McMaster: 70.56% complete and 1.43 average orthologs CEGMA gene; **Additional file 2: Table S2**).

An assembly of a New Zealand (NZ) *H. contortus* strain was recently made available (ENA accession: GCA_007637855.2; [31]), which used our V4 chromosomal assembly to scaffold a highly fragmented draft assembly (contigs = 172,899) into seven chromosome-scale scaffolds. There are striking differences between the NZ and the V4 assemblies that are very unlikely to represent genuine biological variation between strains. The significant expansion of assembly size (465 Mbp vs 283 Mbp in the V4 assembly; **Table 1**) in the context of elevated duplicated BUSCOs (7.3% vs 1.6% in the V4 assembly) and high average orthologs per partial and complete CEGMA genes (1.42 and 1.71, respectively, vs 1.1 and 1.22 in the V4 assembly; **Additional file 2: Table S2**) suggest that contaminating haplotypic sequences

have erroneously been incorporated into the New Zealand assembly. Similarly, although the coding sequences have not yet been made publically available, we suspect that the lower CEGMA statistics together with the reported low average exon number (4 exons per gene vs 9 exons in the V4 assembly) and transcript lengths (666 bp vs 1,237 bp in the V4 assembly; see below) suggest that frameshift errors due to INDELs from uncorrected PacBio reads and/or assembly gaps (n = 172,892) are affecting coding sequences broadly throughout the NZ assembly. When compared against publically available genomes in WormBase ParaSite, the V4 genome assembly represents the most contiguous and resolved nematode assembly outside of the *Caenorhabditis* genus, and second-most complete Clade V parasitic nematode assembly based on the presence of universally conserved orthologs (n = 20; *Ancylostoma ceylanicum* is the most complete based on BUSCO analysis; **Additional file 1: Figure S4 A,B**).

Table 1. Genome assembly summary statistics

	H. contortus V4 Chromosomes	*H. contortus* V4 Haplotypes	*H. contortus* V1 [1]	*H. contortus* McMaster [2]	*H. contortus* New Zealand [3]	*C. elegans* WB [4]
Genome size (bp)	283,439,308	248,771,548	369,846,877	319,640,208	465,720,674	100,286,401
Scaffolds	7	3,907	23,860	14,419	7	7
Scaffold N50 (bp)	47,382,676	167,270	83,287	56,328	83,970,805	17,493,829
Scaffolds ≥ N50	3	174	1,151	1,684	3	3
Contigs	192	4,226	65,523	55,322	172,899	7
Contig N50 (bp)	3,801,457	117,552	20,808	12,900	3,087	17,493,829
Contigs ≥ N50	20	439	4,943	6646	43,632	3
Scaffold N content (bp)	5,699,711	14,755,378	23,804,39	20,495,720	17,288,295	0
Scaffold gaps	185	323	41,663	40,903	172,892	0

1. Haem V1 [21]: haemonchus_contortus.PRJEB506.WBPS10.genomic.fa
2. McMaster [22]: haemonchus_contortus.PRJNA205202.WBPS11.genomic.fa
3. New Zealand [31]: ENA accession GCA_007637855.2: GCA_007637855.2_ASM763785v2_genomic.fna
4. *C. elegans* WBcel235: caenorhabditis_elegans.PRJNA13758.WBPS9.genomic_softmasked.fa

Resolving haplotypic diversity and repeat distribution within the chromosomes

Diverse haplotypic sequences were identified, particularly in our preliminary PacBio assembly (**Additional file 1: Figure S1**; Initial assembly size: 516.6 Mbp), as a result of sequencing DNA derived from a large pool of individuals necessary to obtain sufficient

material. We identified a dominant group of haplotypes in our PacBio assemblies which, together with Illumina read coverage derived from a single worm, was used to partially phase the chromosome assembly sequence by manually curating the haplotypes in the reference assembly. Mapping of the additional haplotypes to the reference-assigned haplotype (**Figure 2 A**) highlighted both the diversity between the haplotype and chromosomal placed sequences (**Figure 2 B**; **Figure 2 C,** second circle) and coverage of the haplotypes, spanning approximately 66.8 ± 24.1% on average of the chromosome sequences (**Figure 2 C,** third circle**;** range 85.1 ± 12.3% for chromosome I to 51.8 ± 16.6% on the X chromosome, measured in 1 Mbp bins)**.** We determined the relative usage of the reference-assigned haplotype and those detected in the alternate haplotype assembly for individuals in the MHco3(ISE).N1 population from which the samples used in the assembly were derived; whole-genome short-read sequencing of 11 single worms (10 male and 1 female, identified by X chromosome coverage; female (XX) = full coverage, male (XO) = half coverage) followed by competitive read mapping to the reference genome and alternate haplotypes revealed clear evidence of haplotype switching based on differences in the coverage profile throughout the genome and between samples (**Figure 2 D**). Haplotypic blocks are evident, including large regions on chromosome 2 and 5 that are devoid of haplotype diversity (full coverage of the reference; blue), as well as regions in which individuals lack any read coverage due to the absence of the reference haplotype in their genome (e.g. red region on chromosome 1, samples 09 & 10). These data suggest that, in spite of being semi-inbred through a limited number of single pair matings [32], complex haplotypes are segregating among individuals in this population, and that the frequency and diversity of these haplotypes are non-randomly distributed throughout the genome.

Despite the reduction in genome size of the V4 assembly, the use of multiple long molecule technologies allowed a greater representation of repetitive sequences, accounting for approximately 36.43% (compared with 30.34% in V1) of the genome (**Additional file 2: Table S3**). Particular repeat classes, including LINEs, LTRs and DNA elements, were enriched towards the middle of the chromosomes and the chromosome ends (**Figure 2 C** , inner circle; **Additional file 1: Figure S5**). This distribution pattern is negatively correlated with observed recombination rate domains in both *H. contortus* [33] and *C. elegans* [34], suggesting that transposon sequences accumulate in regions of low recombination rate throughout the genome. On the X chromosome, this enrichment is striking in the sub-telomeric regions and, in particular, from approximately 8 to 35 Mbp along the chromosome (**Figure 2 C**, inner circle; **Additional file 1: Figure S5**). From a technical perspective, the high frequency of transposon sequences explains the increased frequency of shared sequences between X and the other autosomes, rather than between autosomes alone (**Figure 1 A**). Finally, the pattern of repetitive elements throughout the chromosomes and not the coding sequences (shown below) correlates with and perhaps is driving the nucleotide composition profile throughout the genome (**Figure 2 C**, outer circle).

Figure 2. Haplotype and repeat distribution within and between chromosomes

A. Genome graph derived from the whole-genome alignment of chromosome and alternate haplotype assemblies visualised using *Bandage*. Colour density reflects alternative haplotype density frequency; dark red indicates longer reference only-haplotypes. **B.** Close view of genome graph structure (box), highlighting an example of alternate haplotypes from 7.36 to 7.42 Mbp on Chr X. Aligned regions are represented by thick red lines and paths between regions as thin black lines. DNA alignment of the example 56 kbp alternate

haplotype to the genome using *nucmer* and visualised using *Genome Ribbon* demonstrates the extent of the genetic diversity between these sequences. Two genes are present (red horizontal bars), of which their exons (red ticks) are predominantly found in regions of conservation (ribbons connecting the top and bottom sequences). **C.** Circos plot of genome-wide haplotype and repeat diversity. Outer: mean GC nucleotide frequency in 100 kbp windows. Dashed line represents genome-wide mean GC content (0.429); Second: Distribution (x-axis) and length (y-axis) of mapped haplotype sequences to the chromosome, with sequence similarity between the mapped haplotype and chromosomes indicated by the grey-scale gradient (high similarity = dark, low = light); third: proportion of chromosome covered by a least one haplotype, measured in 1 Mbp windows; inner: density of annotated repeats, including LINEs (green), SINEs (red), DNA (blue), and LTRs (orange). Data is visualised using the R package *circulize*. **D.** Haplotype switching between chromosome and additional haplotypes present in the genome assembly. Whole-genome sequencing was performed on individual worms (n = 11; MHco3(ISE).N1 strain, used in the genome assembly), after which reads were mapped competitively against the chromosomes and haplotype sequences. Normalised read depth coverage is presented, which was calculated per 100 kbp relative to the "maximum genome-wide coverage" set at the 95% quantile to prevent high-coverage outlier regions from skewing coverage estimates. Blue = ~100% of maximum coverage; Yellow = ~50% of maximum coverage; Red = ~ 0% of maximum coverage.

Generation of a high-quality transcriptome annotation incorporating short and long reads

The genome improvements made from the draft V1 assembly to the chromosomal assembly resulted in a reduction of genome size and, at the same time, an increase in the number of core conserved genes. The parsimonious conclusion is that coding sequences were easier to detect in the new assembly and that the reduction of genome size removed previously annotated but redundant genes located on alternate haplotypes in the draft assembly. Considering these broad-scale changes, we undertook a *de novo* gene finding strategy,

rather than relying solely on the previous gene set, to annotate the V4 genome. To do so, we used our comprehensive Illumina short-read RNAseq collection sampled from seven stages of the parasite life cycle (previously described in the analysis of the V1 genome [21]), and supplemented this with PacBio Iso-Seq long-read sequencing of full-length cDNA from adult male, female, and juvenile L3 parasites. To experimentally enhance the recovery of low-abundant transcripts, a duplex-specific nuclease (DSN)-normalisation was performed on the full-length cDNA molecules prior to PacBio sequencing on both RSII and Sequel systems (**Additional file 2: Table S4**), resulting in 23,306 high-quality isoforms after processing.

The overall annotation strategy is presented in **Additional file 1: Figure S6** and consisted of four key stages: (i) RNAseq evidence was used to train *ab initio* gene finding using Braker [35]; (ii) full-length high-quality isoforms generated using Iso-Seq were used to identify candidate gene models using the annotation tool Program to Assemble Spliced Alignments (PASA; [36]; (iii) evidence from steps (i) and (ii), together with 110 manually curated genes and orthologs from the *C. elegans* and *H. contortus* V1 genomes, were incorporated using EvidenceModeller; and finally (iv) UTRs and multi-evidence gene models were updated using the full-length isoforms once again using PASA. We further curated the genome annotation using Apollo [37]; together, this strategy achieved a high level of sensitivity and precision based on comparison with a subset of manually curated genes (**Additional file 2: Table S5**).

Our V4 annotation comprises 19,489 nuclear genes encoding 20,987 transcripts with approximately 56.4% (23,687 of 41,974) of UTRs annotated. The increase in genome contiguity, together with the incorporation of full-length cDNA sequencing (**Figure 3 A**), has

resulted in longer gene (40.5% increase), mRNA (41.9% increase), and exon (30.6% increase) models on average than the previous V1 genome annotation (**Figure 3 B; Additional file 2: Table S6**); our approach allowed better representation of the longest genes (447,146 bp compared with 91,953 bp in V1), and intriguingly, identified 43 of the longest 50 (86%) and 73 of the longest 100 gene models on the X chromosome alone. The reduction in haplotypic sequences together with greatly improved gene models resolved 61% more one-to-one orthologs with *C. elegans* (n = 7,361) than between *C. elegans* and V1 (n = 4,529) (**Additional file 2: Table S7**). Surprisingly, a greater number of one-to-one orthologs are identified between V4 and *H. placei* (n = 9,970; https://parasite.wormbase.org/Haemonchus_placei_prjeb509/Info/Index/) than V4 and V1 (n = 9,595), likely reflecting the higher duplication rate and haplotypic nature of the V1 relative to the *H. placei* assembly, although this also highlights the close relationship between *H. contortus* and *H. placei,* a gastrointestinal pathogen commonly associated with cattle. The increased orthology between close (*H. placei*) and more distantly (*C. elegans*) related species provides support for a more complete and biologically representative gene set in the V4 annotation than the previously described V1 and McMaster annotations.

Figure 3. *De novo* transcriptome annotation incorporating full-length cDNA sequencing results in longer and more complete annotations

A. Example of an improved and now full-length gene model. HCON_00001100, a nuclear hormone receptor family member orthologous to *C. elegans nhr-85*, was incomplete and fragmented in the V1 assembly (multiple colours representing CDSs from three distinct V1 annotations, HCOI00573500, HCOI00573600, and HCOI00573700). The full-length gene model, comprised of six exons (black boxes) flanked by UTR sequences (white boxes) is

supported by full-length cDNA IsoSeq sequences (blue bars). **B.** Comparison of mean length and abundance of annotation features between *H. contortus* V1 (square), *H. contortus* V4 (cross), *H. contortus* McMaster (triangle) and *C. elegans* (circle) transcriptome annotations.

Transcriptional dynamics throughout development and between sexes

Both 2013 publications describing draft assemblies of *H. contortus* provide extensive descriptions of the key developmental transitions between sequential pairs of life stages throughout the life cycle [21,22]. We revisited this in the current annotation, and for visual comparison, we present the top 1,000 most variable transcripts across all life stage samples in the MHco3(ISE).N1 datasets (**Figure 4 A;** see **Additional file 2: Table S8** for all differentially expressed genes from pairwise comparisons of the life stages throughout the life cycle). Most striking was the transcriptional transition between the juvenile free-living life stages and the more mature parasitic stages, reflecting not only the development toward reproductive maturity but also significant changes in the parasite's environment. To explore these developmental transitions more explicitly, we determined the co-expression profiles for all genes throughout the lifecycle, revealing 19 distinct patterns of co-expression containing 8,412 genes (41% of all genes), with an average cluster size of 442.7 genes (**Additional file 1: Figure S7; Additional file 2: Table S9**). Consistent with the pattern of highly variable genes, two dominant clusters of genes were identified: (i) genes with higher expression in free-living stages that subsequently decreased in expression in parasitic stages (cluster 1; n = 1,550 genes; **Figure 4 B**) and (ii) genes that were lowly expressed in free-living stages that increased in expression in parasitic stages (cluster 8; n = 1,542 genes). Functional characterisation of the cluster 1 genes revealed 45 significantly enriched gene ontology (GO) terms (26 molecular function [MF], 13 biological process [BP], & 6 cellular compartment

(CC) terms; https://biit.cs.ut.ee/gplink/l/8jECkGwQQS) that predominantly described ion transport and channel activity, as well as transmembrane signalling receptor activity. For cluster 8, 22 GO terms were significantly enriched (11 MF, 8 BP, 3 CC; https://biit.cs.ut.ee/gplink/l/oJoq6Zh0Re) and included terms describing peptidase activity, as well as protein and organonitrogen metabolism. To provide further resolution, we tested *C. elegans* orthologs of this gene set, which revealed greater resolution on metabolic processes, including terms associated with glycolysis and gluconeogenesis, pyruvate and nitrogen metabolism (https://biit.cs.ut.ee/gplink/l/b3Bx8rhJR5). Finally, we explored whether co-expressed genes contained shared sequence motifs in their 5' UTR that may would be indicative of true transcriptional co-regulation; while broadly poorly represented, we did identify 70 enriched motif sequences of which 17 were associated with gene sets with significantly enriched GO terms (**Additional file 2: Table S10**), for example, cysteine-type peptidase activity in cluster 8 (GATAAGR; motif E-value = 2.90E-02; GO p-value = 4.181E-13), DNA replication in cluster 15 (AAAAATVA; motif E-value = 3.50E-02; GO p-value = 6.269E-4), and molting cycle, collagen and cuticulin-based cuticle in cluster 16 (GTATAAGM; motif E-value = 2.30E-02; GO p-value = 1.926E-6), suggestive of either direct or indirect co-regulation of expression.

The greatest differences in gene expression between life stages are between L4 (mixed male and females) and either adult males or adult females, or between adult males and adult females, with a strong sex-bias towards increased gene expression in adult males (**Additional file 2: Table S8**). Consistently with *C. elegans*, sex-biased expression is not randomly distributed throughout the genome [38,39]; only 5% (146/2,923) of *H. contortus* male-biased genes are found on the X chromosome (relative to 14.3% of all genes found on

the X chromosome; χ^2= 175.83, df = 1, p = 3.84E-40), whereas more female-biased genes are X linked than expected (24.6% [110/476]; χ^2= 33.59, df = 1, p = 6.78E-9). Comparison of sex-biased genes on the X chromosome revealed that, while almost completely absent in males, X-linked female-biased expression identified *C. elegans* orthologs previously associated with organism development processes, including developmentally upregulated genes (e.g. *smad4*, transcription factor AP2, *vab-3*), genes involved in egg development (e.g. *egg-5, vit-1, cpg-2, egg-1*), as well as genes involved in developmental timing (e.g. *lin-14, kin-20* [40]) that are expressed downstream of the sex-determination regulator *tra-1*. Sex determination in *C. elegans* is mediated by the ratio of X chromosomes to autosomes, whereby organisms with a single X chromosome will develop as a male (X-to-autosome ratio = 0.5:1, i.e., XO) and those with two X chromosomes develop as a hermaphrodite (XX; X-to-autosome ratio = 1:1); similarly, *H. contortus* males are XO and females are XX [29,30]. The detection of the X to autosome copy number initiates a regulatory pathway that first activates a dosage compensation response that down-regulates gene expression in hermaphrodites by one-half, followed by sex determination mediated by the conserved transcription factor *tra-1*; in XX individuals *tra-1* is activated, whereas *tra-1* is repressed in XO individuals. Although the similarities in the XX/XO chromosome structures between *C. elegans* and *H. contortus* suggest that the mechanisms controlling sex determination may be the same, this has not been previously shown. Comparison of *H. contortus* RNAseq data from males and females showed no significant difference in the distribution of expression between autosomes and X-linked genes, and considering only 9.24% of genes (110 + 146 of 2,770 X-linked genes) are differentially expressed on the X chromosome between sexes, these data suggest that X chromosome dosage compensation is active in XX females. To explore this further, we mapped *H. contortus* orthologs of *C. elegans* genes involved in the X

chromosome sensing and dosage compensation response and subsequent sex determination pathway (**Figure 4 C, Additional file 2: Table S11;** Adapted from WormBook [41]). Five of the six *C. elegans* orthologs associated with the dosage compensation complex (DCC) were found, however, almost all genes upstream that initiate the recruitment of the DCC are absent, including *xol-1* and *sdc-1,-2, and -3*. In *C. elegans*, chromosome dosage is determined by measuring the expression of key genes on the X and autosomes; in *H. contortus*, only two of the *C. elegans* X-linked orthologs - *fox-1* and *sex-1* - are present, however, these are located on the autosomes. Thus, while the DCC may still play a role in dosage compensation in *H. contortus*, the mechanism used to determine the X-to-autosome ratio is likely different from *C. elegans*. Identification of most orthologs downstream of *her-1*, which receives X chromosome dosage signals to initiate sex determination, suggests greater conservation of this mechanism between species. Notable is the absence of *H. contortus* orthologs of *tra-2* and *fem-3*, as they are key members of the pathway that receive regulatory signals promoting male or female developmental pathways, respectively, in *C. elegans*.

Figure 4. Transcriptional dynamics throughout development and between sexes

A. Heatmap of the most variable transcripts across the life stages. Transcript abundance measured as transcripts per million (TPM) was determined using Kallisto, from which the TPM variance was calculated across all life stages. The top 1,000 most variable transcripts with a minimum of 1 TPM are presented. Genes differentially expressed in each pairwise comparison throughout the life cycle are presented in **Additional file 2: Table S8**. **B.** Transcript co-expression profiles across life stages of *H. contortus*. Cluster 1, which contains the most co-expressed genes, is presented and broadly reflects the transition between juvenile free-living life stages and mature parasitic stages. The remaining profiles are

presented in **Additional file 1: Figure S8**, with genes by cluster described in **Additional file 2: Table S9**. **C.** Schematic of the chromosome dosage sensing and sexual development pathway of *C. elegans*. Adapted from WormBook [41], we highlight the presence of *H. contortus* orthologs (bold, underlined) of *C. elegans* genes previously demonstrated to promote male (blue) or female (green) development. Missing orthologs from the *H. contortus* genome are translucent. A full list of *C. elegans* genes and *H. contortus* orthologs from the pathway are presented in **Additional file 2: Table S11**.

Transcriptome complexity is generated by extensive *cis*- and *trans*-splicing

The significantly improved genome contiguity and transcriptome representation prompted us to revisit the prevalence and characteristics of RNA splice variation throughout the transcriptome. Spliced leader (SL) *trans*-splicing, which commonly functions to split polycistronic mRNAs into monocistronic units by the addition of a common SL sequence to unrelated pre-mRNAs to form mature mRNAs, has been extensively described in *C. elegans* [42] and has been shown previously to operate in *H. contortus* [43–45] (**Figure 5 A**). To explore this further in the V4 annotation (**Additional file 1: Figure S8 A,B**), we identified 7,293 (37.4%) and 1,139 (5.8%) genes with evidence of SL1 and SL2 *trans*-splicing, respectively, within 200 bp upstream or 50 bp downstream of the first start codon (see **Additional file 2: Table S12** for complete list of genes with SL1 and/or SL2). However, additional evidence of *trans*-splicing was observed; 7,895 and 1,222 annotated transcripts contained SL1/SL2 sequences, and a further 6,883 internal SL cut sites (overlapping internal exons, excluding the 5' and 3' exons) indicated additional non-annotated isoforms were present in the transcriptomes. While our results likely underestimate the frequency of *trans*-splicing due to the modest RNA sequencing depth available, this approach reveals a higher frequency of *trans*-splicing than previously described and presents additional

evidence for the prediction and annotation of novel gene isoforms. Further, we identify degeneracy in the nucleotide sequences of both SL1 and SL2 sequences (**Additional file 1: Figure S8 C**), and sequence conservation in the genome sequence immediately upstream of the splice site, particularly for SL2 (**Additional file 1: Figure S8 D**). In *C. elegans*, SL1 sequences are typically associated with the non-operon genes, or the first coding sequence of an operon, whereas SL2 sequences are used for second and subsequent coding sequences in a polycistronic operon. The intercistronic regions between SL2-spliced genes and their upstream genes in *C. elegans* are typically short, usually approximately 50 to 200 bp on average, with the majority less than 2 kbp [42]. In *H. contortus*, the intergenic distance upstream of genes with SL1 sequences is largely consistent with that of genes without SL1 sequences (**Figure 5 B,** left plot; median = 8,557 bp for SL1-spliced genes, 8,603.5 bp for genes without a SL); however, a small proportion of SL1-spliced genes are found less than 100 bp downstream of the nearest gene. Conversely, while a larger proportion of genes with SL2 sequences are immediately downstream of the nearest gene, as expected (median = 566.5 bp), a substantial proportion (432 [35.4%]) are found with a large intergenic distance (> 2 kbp) to the nearest upstream gene (**Figure 5 B,** right plot). These observations suggest that while broadly consistent with the mechanism by which SL1/SL2 sequences are used in *C. elegans*, (i) *H. contortus* SL1-sequences may be found in downstream genes within a polycistron, (ii) SL2-spliced genes are not necessarily in a polycistron but may function independently, and/or (iii) that recognition of SL2 splicing within an operon remains effective over large physical distances, for example, the 2064 bp region between deg-3 (*HCON_00039370*) and *des*-2 (*HCON_00039380*) operon of *H. contortus* [46]. Although these variations have been described in *C. elegans* [47], they seem to be more of the norm rather than the exception in *H. contortus.*

We extended the splice variation analysis to characterise *cis*-splicing of transcripts, with a particular focus on differential isoform usage between life stages. Using a conservative threshold to detect differential splicing, we identified 1,055 genes with differentially spliced transcripts impacting at least one intron, representing on average 283 genes (FDR < 0.05; range: 43-418) that switch between different isoforms in each pairwise comparison of life stages throughout the lifecycle (**Figure 5 C;** see **Additional file 2: Table S13** for a summary per pairwise comparison between life stages, and **Additional file 2: Table S14** for list of genes that appear differentially spliced between pairwise life stages throughout the life cycle). For example, 231 of 1,068 genes (21.63%; **Additional file 2: Table S13**) show evidence of differential splicing between adult male and female samples. The most differentially spliced intron is located in *HCON_00073480*, a one-to-one ortholog of *C. elegans daz-1* (**Figure 5 D**). daz-1 is an RNA-binding protein involved in oogenesis but not spermatogenesis in *C. elegans [48]*; in *H. contortus*, *HCON_00073480* is predominantly expressed in females (average TPM = 636.22; highest expression in hermaphrodites in *C. elegans*), and is expressed in males, albeit approximately 23-fold less (average TPM = 26.71)(**Figure 5 E**). Only a single transcript is described in *C. elegans*; however, a total of five isoforms are annotated in the *H. contortus* genome, with a clear difference in sex-specific expression of exon 9 that is present in males and absent in females. Further, additional differential splicing was found in the splice donor site of exon 8, and while both alternate splice sites are present in both sexes, the sites are used equally in males but the longer transcript is used in around 60% of transcripts in female worms (**Figure 5 F**).

Figure 5. Extensive *trans*- and *cis*-splicing of gene transcripts throughout the life cycle

A. Example of a gene model, HCON_00096900, visualised using *Apollo* that contains two SL1 trans-spliced sites upstream of the transcriptional start sites of two distinct isoforms, each supported by full-length IsoSeq sequence reads. **B.** Distribution of intergenic distance (log10[bp]) between the start coordinates of genes containing SL1 (red), SL2 (blue) sequences, or with no SL sequence (grey), and the end coordinate of the gene immediately upstream. In the second plot, the relative frequency of SL1 or SL2 sequences is presented. **C.** Heatmap of the top 100 differentially spliced introns identified in pairwise comparisons throughout the life cycle. Data are presented as the proportion of differentially excised intron, ranging from 0 (light) to 1 (dark), for each intron cluster for which differential splicing was identified. A complete list of genes with evidence of differential splicing is presented in **Additional file 2: Table S14**. **D.** Heatmap of the top 100 differentially spliced introns between adult male and adult female worms. The scale is the same as shown in (C). **E.** RNAseq expression data for adult female, adult male, all life stages combined, relative to the annotated isoforms for *HCON_00073480*, a one-to-one ortholog of the *C. elegans daz-1*. **F.** Quantification of differential splicing of *HCON_00073480* between adult female (top) and adult male (bottom) samples, determined using leafcutter. Focused on exons 8, 9, & 10 (black bars), the red lines depict the linking of splice junctions, with the thickness of the line relative to the proportion of reads that support that junction. In this example, two differential splicing events are observed; (i) alternate splice donors of exon 8, and (ii) exon skipping of exon 9.

Distribution of global genetic diversity throughout the chromosomes

H. contortus is a globally distributed parasite, characterised by large population sizes and in turn, high genetic diversity [25–27,49]. We have characterised the distribution of genetic diversity throughout the V4 chromosomes by exploiting a recent large scale analysis of global genetic diversity [25] that we have extended by sequencing an additional 74

individual worms from previously unsampled regions including Pakistan, Switzerland, USA, and the United Kingdom (**Figure 6 A**; https://microreact.org/project/hcontortus_global_diversity). Principal component analysis (PCA) of mtDNA diversity (**Figure 6 B; Additional file 1: Figure S9**) among 338 individual parasites showed that relationships between isolates are congruent with those described by [25]; further, we show that Pakistani worms were genetically positioned between South African and Indonesian individuals, whereas the US and the UK samples were found among globally distributed samples that likely shared a common ancestral population that was distributed globally via modern human movement [25]. Of note, UK parasites were largely genetically distinct from the French worms despite being geographically close, and yet, Swiss parasites were genetically closer to African worms than other European parasites. The US and UK samples consisted of both laboratory-maintained strains and field isolates; notably, apart from MHco3(ISE).N1, the laboratory strains still retained similar levels of genetic diversity to the field samples. MHco3(ISE), from which the reference strain MHco3(ISE).N1 was derived, is believed to have originated in East Africa; although we have no samples from this region, the MHco3(ISE).N1 population was genetically similar to one of the three South African populations (**Additional file 1: Figure S10**; Kokstad population), perhaps reflecting its African origins.

Genome-wide nucleotide diversity (π), calculated in 100 kbp windows throughout the genome, is broadly consistent chromosome-wide between the autosomes (**Figure 6 C-G**). The high frequency of SNPs private to only a single population (relative to variants shared between two or more populations) represents a significant proportion of the overall variation (23.55% of total variation), consistent with previous reports describing high

regional genetic diversity [25–27]. On the X chromosome, nucleotide diversity is much lower than neutral expectation (i.e., $\pi_X/\pi_{autosome} \simeq 0.75$ [50]) at a $\pi_X/\pi_{autosome}$ median ratio of 0.363 (**Figure 6 H**), which suggests a sex-ratio bias in favour of males. This observation, together with the distinct repeat patterns that were negatively correlated with recombination rate, suggests that recombination may be less frequent on the X chromosome than expected. Interestingly, and likely related to these observations, is that variation in X chromosome coverage (**Additional file 1: Figure S11 A**) increases in populations that are more genetically distinct from the reference strain (**Additional file 1: Figure S11 B**); these data suggest that while the sampled nucleotide diversity is observed to be low on the X chromosome, greater biological variation in chromosome content and structure is present among the global sample set that cannot be correctly accounted for by mapping to a linear reference alone.

Finally, we explored evidence of genetic differentiation between populations throughout the genome, revealing multiple discrete regions in each chromosome with high F_{ST} among populations (**Figure 6 C-H**; bottom plots). Consistent with previous reports, strong genetic selection for anthelmintic resistance has impacted genome-wide variation; the largest and most differentiated region is found on chromosome 1 surrounding the beta-tubulin isotype 1 locus (HCON_00005260; 7027492 to 7031447 bp), a target of benzimidazole-class anthelmintics for which known resistant mutations have been identified in these populations [25], as well as a broad region of elevated F_{ST} from approximately 35 to 45 Mbp on chromosome 5 previously shown to be associated with ivermectin resistance in controlled genetic crosses [51]. The top 1% of F_{ST} outlier regions per chromosome (280 10 kbp regions in total) contains 268 genes enriched for 10 GO terms (7 MF & 3 BP; https://biit.cs.ut.ee/gplink/l/HKi5aHzqTa), largely describing proteolytic and cysteine-type

peptidase activity. Analysis of d_N/d_S from alignments of 9,970 one-to-one orthologs from *H. contortus* and the closely related *H. placei* identified 108 genes putatively under positive selection ($d_N/d_S > 1$). Gene set analysis revealed no enrichment for specific GO terms among these genes; closer inspection revealed a large number of genes with no known function, despite most having orthologs in two or more clade V nematodes, highlighting both the challenge of interpreting genetic novelty in non-model organisms and the clear need for functional genomics to provide missing detail in the annotation of organisms like *H. contortus* [52,53]. We did, however, identify a number of genes likely associated with host/parasite interactions (e.g. OV39 ortholog [HCON_00044830], multiple ShKT domain and signal peptide-containing genes [e.g. HCON_00095770, HCON_00124950, HCON_00108210], BAT-2 domain [HCON_00173800]), and broad development and neurological function (e.g. *nck-1, snt-2, flp-5, ned-8, eat-17*), which could reflect adaptation to different host species. We extended this analysis to *H. contortus* variation derived from the global diversity dataset; after filtering, 849 (4.8% of 17,607) genes were identified with a d_N/d_S greater than 2. Similar to the between species analysis, no significant GO terms were enriched in this dataset. However, G-protein coupled receptors (e.g. *aex-2, npr-4, sdr-7, srb-17, npr-27, seb-4, seb-2*), and G-proteins (e.g. *goa-1, gpa-4, gpa-16*), ion and neuropeptide gated transmembrane channels (e.g. *acr-12, acr-19, acr-20, lgc-9, lgc-34, lgc-46, lgc-39, mod-1, ccb-2, kvs-4*) were among the most abundant *pfam* domains identified, highlighting the adaptive potential of sensory and response pathways to external stimuli.

Figure 6. Genome- and population-wide genetic diversity

A. Global distribution of populations from which genomic data were derived, including 264 samples from Sallé *et al*. (circles) and 74 new samples (triangles) presented. The insert highlights the dense sampling from the United Kingdom (this study) and France [25]. **B.** Principal component analysis of genetic diversity among the globally distributed samples, using 2,401 mtDNA variant sites. The reference strain, MHco3(ISE).N1, is indicated. **C to H.** Chromosome-wide comparison of genetic diversity. Each comparison contains three panels. Top: Estimate of nucleotide diversity (π) calculated in 100 kbp windows, accounting variants found only in a single population (dark), or shared between two or more populations (light). Middle: Estimate of genetic differentiation (F_{ST}) between populations, measured in 100 kbp windows. Bottom: Gene density per chromosome, where each gene is represented is a black line on the sense (upper lines) or antisense (lower lines) DNA strand of the chromosome.

Discussion

The *H. contortus* MHco3(ISE).N1 V4 assembly presented here represents the largest *de novo* chromosome-scale nematode genome to date and provides a critical genomic resource for a broad group of human and veterinary pathogens present among the Clade V nematodes. This assembly is complemented with a high-quality *de novo* genome annotation that incorporates full-length cDNA sequencing to provide a precise transcriptomic resource for within and between species comparison, improved UTR and isoform classification for functional genomics, and should help to improve the annotation of other nematode species. Finally, we provide a comprehensive description of genetic and transcriptomic diversity within individuals and across the species, which informs not only our understanding of key developmental trajectories but also the parasite's capacity to adapt when faced with challenges such as drug exposure.

A key challenge to achieving this chromosomal assembly was the presence of extensive genetic diversity in the sequencing data. Although high-quality genome assemblies are becoming more prevalent through the use of long-read sequencing technologies [4,5,54], a feature of these assemblies is that they are derived from a single, few, or clonal individual(s) in which the genetic diversity is low or absent; most of the genetic diversity present in a single diploid individual can be phased and separated into distinct haplotypes (for example, using software such as *Falcon-Unzip* [8] and *Supernova* [7] for PacBio and 10X Genomics data, respectively), and assembled separately. However, even under these scenarios, heterozygosity remains a key technical limitation towards achieving contiguous genome assemblies. Considering the small physical size of *H. contortus*, it was necessary to pool thousands of individuals to generate sufficient material for long-read sequencing, and while

we did so from a semi-inbred population, there was still significant genetic diversity present in the population of worms sequenced. The generation of such diversity from a single mating pair is likely influenced by reproductive traits including polygamous mating and high fecundity of *H. contortus* female worms. How such haplotypic diversity is maintained in the population or is tolerated within single individuals remains unclear, however, it is likely that the holocentric chromosomes of *H. contortus* likely tolerate significant genetic diversity during homologous pairing of chromosomes at meiosis [55]. In spite of this, we curated a major haplotype in the ISE.N1 population to represent the chromosomes and demonstrated how this haplotype and alternate haplotypes segregated in this semi-inbred population. The absence in the literature to date of such chromosome-scale assemblies for physically small and genetically diverse eukaryotic species likely reflects the significant investment required to improve a draft assembly to chromosomes. However, it is likely that as DNA input requirements for long-read sequencing decrease [56], the ability and ease with which highly contiguous genomes can be generated for small and genetically diverse species will be significantly enhanced.

Completion of the X chromosome represented a key advancement in the current version of the genome. In doing so, we revealed striking differences in: (i) the repeat distribution and content, (ii) an enrichment for the longest genes of the genome, (iii) the lack of apparent genetic diversity relative to the autosomes, and (iv) differences in the gene content that regulates X chromosome dosage and sex determination relative to *C. elegans*. Although we currently have little empirical data on recombination rates on the X chromosome, less recombination is expected on the X chromosome relative to the autosomes due to recombination on X only occurring during female gametogenesis. The domain structure

revealed by the repeat distribution does suggest that recombination patterns may be distinct from the autosomes and warrants further investigation. The difference in nucleotide diversity between the X chromosome and autosomes further emphasises the impact of putative sex biases on the evolution of the X chromosome. *H. contortus* is polyandrous [29,33] and highly fecund, so that reproductive fitness is likely overdispersed between males and females, potentially resulting in different effective population sizes for male and female parasites. These differences may be exacerbated if males and females are differentially affected by anthelmintic exposure as has been demonstrated for some compounds [57]. Further, post-zygotic incompatibility between genetically divergent *H. contortus* strains has been previously demonstrated [58]; however, it is unclear if this occurs within naturally occurring populations and whether this would, in fact, impact effective sex ratios. The dichotomy between the lack of low nucleotide diversity and high variance between samples in the global cohort perhaps reflects a key limitation on the ability of a single linear reference genome to capture genome diversity in a species with high genetic diversity. From a technical perspective, if reads from regions of the genome that contain high diversity do not map to the reference, as demonstrated by differences in sequence read coverage between reference and haplotype sequences among ISE.N1 individuals, only a proportion of the total diversity will be captured. Long-read sequencing data (e.g. PacBio or Oxford Nanopore) from *Haemonchus* populations and bioinformatics approaches designed to capture more complex variation should help reveal the full extent of genetic variation in this species and shed light on the role of structural variation in shaping this genetic variation.

Our high-quality annotation allowed us to extend previous pairwise transcriptomic comparisons and focus on the trajectories of genes with correlated expression throughout

the life cycle. These data identified gene sets that are putatively co-regulated, providing greater insight into coordinated developmental programs than has been described previously. The broad-scale coordinated gene expression associated with ion transport, channel activity, and transmembrane signalling receptor activity throughout the lifecycle was of interest, particularly as ion channels and receptors are the primary targets of a number of anthelmintic drugs used to control *H. contortus* and other helminth species of human and veterinary importance [59]. A better understanding of these changes may explain why some parasites at particular life stages are less sensitive to these drugs [60,61], and may provide the rationale for selective treatment based on the known expression of drug targets. We further explored transcriptome variation by defining and quantifying differential isoform usage throughout the life cycle for the first time. The importance of differential isoform usage is increasingly recognised in the sensitivity of drug targets in nematodes such as *H. contortus*; for example, sex-dependent sensitivity of emodepside is mediated by differential-splicing of the potassium channel *slo-1* [57], and similarly, expression of truncated acetylcholine receptor subunits *acr-8* and *unc-63* isoforms are associated with levamisole resistance [62,63]. Our results almost certainly under-estimate the extent of *cis*-splice variation within and between life stages; stringent filtering resulted in only a subset of genes tested (range: 759-1156 genes per pairwise comparison of life stages), and yet, in each pairwise comparison, approximately 27.32% of genes on average contained evidence of splice variation (**Additional file 2: Table S13**). Further, each pairwise comparison identified cryptic splice sites, indicating that a proportion of splice junctions from novel isoforms undergoing differential splicing were missing from the genome annotation. Ongoing manual curation to the genome annotation to classify isoform variants, together with more fine-grained quantification of transcript usage, will provide additional

insight into the novelty that differential isoform usage provides, particularly for those genes for which isoform switching but not differential expression occurs between key life-stage transitions. Beyond the life cycle, the impact of high genetic diversity on transcriptomic variation within or between wild isolates is not yet clear, although both technical and biological variation between three divergent strains has recently been described [64]. A greater understanding of this genetic-to-transcriptomic variation is required, especially if phenotypic traits associated with transcriptomic differences, e.g. increased gene expression of a drug transporter correlated with resistance, are to be correctly attributed to genetic selection.

Finally, analyses of chromosome-scale assemblies allowed us to explore the distribution of genetic features throughout the genome and to speculate on the evolutionary events that have taken place to explain these distributions. For example, between species analyses have confirmed karyotype structures and/or revealed chromosomal rearrangements among rhabditine nematodes [65,66], filarial nematodes [11], and platyhelminths [67]. The V4 assembly allowed the extent of genome rearrangements between *C. elegans* and *H. contortus* to be described for the first time, and similarly, we revealed numerous gene and repeat family expansions that were missed in the previous assembly and produced a more complete catalogue of coordinated transcription within and between life stages. As further genomes are completed from the Strongylid clade, determining the rates of structural evolution in this important group of parasitic organisms should become possible. Investigating such evolutionary processes will be important for understanding genome-wide variation associated with trait variation, including important adaptive traits such as response to climate change and anthelmintic resistance. *H. contortus* is an established

model for understanding parasite response to anthelmintics and subsequent evolution of resistance to key drugs used to control parasitic nematodes [20]. Our chromosomal assembly has been instrumental in mapping genetic variation associated with ivermectin [51] and monepantel [68] resistance. In the case of ivermectin, the identification of the major genomic region on chromosome 5 associated with resistance would not have been possible without a chromosome-scale assembly, and highlights the potential limitations on the interpretation of such variation using less-contiguous genome assemblies as a reference [28].

Conclusions

Our chromosome-scale reference genome and comprehensive annotation of the MHco3(ISE).N1 isolate provides a significant step forward toward resolving the complete genome composition and transcriptional complexity of *H. contortus* and is an important tool for understanding how genetic variation is distributed throughout the genome in this species. The increasing availability of long-read sequencing together with technological advances in the sequencing of small diverse organisms holds great promise towards the rapid generation of high-quality contiguous genomes for organisms like *H. contortus*; future sequencing, including the generation of new *de novo* genomes and transcriptomes from individuals sampled throughout its geographic range, will provide great insight into how genetic variation is maintained and/or changes in response to persistent challenges, including anthelmintic exposure and climatic change, in this hyper-diverse globally important pathogen.

Materials and Methods

Parasite material

All parasite material used in the genome assembly is derived from the semi-inbred strain of *H. contortus,* MHco3(ISE).N1. This strain was derived from a single round of single-worm mating using the MHco3(ISE) strain (protocol and procedure described in Sargison *et al.* [32]), and maintained as previously described [21].

Iterative improvement of the genome assembly

Our approach to improving the V1 genome to the final V4 chromosomal assembly is outlined in **Additional file 1: Figure S1**. The sequence data generated for this study are described in **Additional file 2: Table S1**.

Initial manual improvement on the V1 genome focused on iterative scaffolding with *SSPACE* [69] and gap-filling with *IMAGE* [70] using Illumina 500 bp and 3 kbp libraries, with additional low coverage data from 3, 8 and 20 kbp libraries generated using Roche 454 sequencing. These improvements were performed alongside breaking of discordant joins using *Reapr* [71], and visual inspection using *Gap5* [72]. Substantial genetic variation was present in the sequence data due to the sequencing of DNA derived from a pool of individuals, resulting in a high frequency of haplotypes that assembled separately and therefore present as multiple copies of unique regions in the assembly. We surmised that much of the assembly fragmentation was due to the scaffolding tools not being able to deal with the changing rates of haplotypic variation so we attempted to solve this manually in gap5. We were aware that we did not have sufficient information to correctly phase these

haplotypes, so instead, we chose the longest scaffold paths available, accepting that our scaffolds would not represent single haplotypes but would rather be an amalgamation of haplotypes representing single chromosomal regions. This approach was initially difficult and time-consuming, and was further confounded by a large number of repetitive sequences present in each haplotype.

Significant improvements in scaffold length were gained by the integration of OpGen (http://www.opgen.com/) optical mapping data. Optical mapping was performed following methods described previously [14] with the following exceptions: high molecular weight DNA was prepared by proteinase K lysis of pools of ~500 *H. contortus* L3 embedded in agarose plugs, after which one of three restriction enzymes (KpnI, AflII and KpnI) were used to generate three separate restriction map datasets. Initial attempts to generate a *de novo* assembly using optical molecules alone was unsuccessful, and therefore, optical contigs were generated using DNA sequence seeds from the genome assembly using *GenomeBuilder* (OpGen) and visualised and manually curated using *AssemblyViewer* (OpGen). Although this approach was successful, it was limited by the quality and integrity of the gap-dense scaffolds and arbitrary nature of the haplotype scaffolding.

Subsequent integration of PacBio long-read data alongside the optical mapping data resulted in major increases in contiguity. PacBio sequencing libraries were prepared and sequenced using the PacBio RSII system. A total of 32.3 Gbp raw subreads (n = 4,085,541, N50 = 10,299 bp) were generated from 33 flow cells, representing approximately 133.8× coverage (based on the estimated genome size of 283 Mbp). A *de novo* PacBio assembly was generated using *Sprai* (v0.9.9.18; http://zombie.cb.k.u-tokyo.ac.jp/sprai/index.html), which

was mapped to the assembly and used to manually resolve many gaps and resolve some of the phasing between haplotypes. Using the *Sprai* PacBio *de novo* assembly, we were also able to incorporate contigs that were previously missing from the working assembly. The increase in contiguity of the PacBio assemblies, further improved using *canu* v1.3 [73], revealed two major but diverse haplotype groups segregating at approximately 65% and 30% frequency in the pooled individuals sequenced; the presence of such diverse haplotypes resulted in a significantly expanded assembly over 500 Mbp. The major haplotype was more contiguous and therefore was chosen as the primary sequence to incorporate and improve the assembly. This approach was supported by competitive mapping of a single worm short read sequencing library (ENA: ERS086777), which was found to preferentially map to different scaffolds and/or different contigs within scaffolds that we inferred were different haplotypes in this single worm. Regions in the chromosome with no ERS086777 coverage, but for which an alternate haplotype was identified with coverage from the PacBio assembly, were manually removed from the main chromosomes and replaced. Further, this selective mapping correlated well with the optical contigs, and once these sequences were removed, much better optical alignments were produced further improving the assembly. Alternate haplotypes from the PacBio assembly, and those removed from the main assembly, were analysed as a separate haplotype assembly.

The increase in contiguity and resolution of shorter repetitive regions using PacBio began to reveal chromosome-specific repetitive units. Although these repeats were highly collapsed in the assembly and were typically not spanned by optical molecules, we were able to iteratively identify and join contigs/scaffolds flanking large tandem repeats that had clear read-pair evidence that they only occurred in a single location in the genome (i.e. read pairs

were all mapped locally once the join was made). These were further supported by local assemblies of subsets of PacBio reads that contained copies of the repeat regions followed by *de novo* assembly using *canu* [73] to reconstruct the flanking unique regions surrounding a repeat. This iterative process resulted in the production of chromosome-scale scaffolds, each terminating with a 6 bp repeat consistent with being a telomeric sequence (sequence motif: TTAGGC).

The V3 assembly, used in the *H. contortus* genetic map [33], consisted of five autosomal chromosomes, with the X chromosome in two major scaffolds. The X chromosome-associated scaffolds were identified by synteny to the *C. elegans* X chromosome using *promer* [74], and by the expected ~0.75× coverage relative to the autosomes in pooled sequencing (due to a mixture of male and females present) and ~0.5× in single male sequencing libraries. We resolved the X chromosome using linked read long-range sequencing using the 10X Genomics Chromium platform (www.10xgenomics.com). DNA prepared from pooled worms was used to generate 10X sequencing libraries, which were subsequently sequenced on a HiSeq 2500 using 250 bp PE reads. Reads were mapped using *bwa mem* (v0.7.17-r1188), followed by scaffolding using *ARCS* [75] and *LINKS* [76]. A single scaffold was generated using this approach, merging the two major X-linked scaffolds as well as two short sequences putatively identified as being X-linked but previously unplaced (**Additional file 1: Figure S2**). The completion of the X chromosome, together with the polishing of the genome first with *arrow* (https://github.com/PacificBiosciences/GenomicConsensus) followed by *Pilon* v1.22 [77], finalised the genome to produce the V4 version presented here.

Genome Completeness

Assembly completeness was determined throughout the improvement process using Core Eukaryotic Genes Mapping Approach (CEGMA; v2.5) [78] and Benchmarking Universal Single-Copy Orthologs (BUSCO; v3.0) [79] gene completeness metrics. For comparison, these metrics were also determined for the complete set of nematode genomes from the WormBase ParaSite version 12 release.

Haplotype analyses

The identification and discrimination of the major and minor haplotypes in our assembly offered an opportunity to characterise the distribution of haplotypes throughout the genome, and the relative diversity between them.

To visualise the broad-scale variation between the chromosome and haplotypes, we generated an approximate genome graph using *minigraph* (https://github.com/lh3/minigraph; parameters: -xggs). The output GFA file was visualised using *Bandage* [80]; a representative set of haplotypes were chosen for further comparison, which were compared using nucmer and visualised using *Genome Ribbon* [81].

To quantify the distribution and diversity of haplotypes, we used *minimap* to align haplotypes per chromosome. Haplotype density was calculated in 1 Mbp windows throughout the genome using *bedtools coverage* [82] and visualised using the R package *circlize* [83].

We determined the relative usage of detected haplotypes in the MHco3(ISE).N1 population by using Illumina sequencing of DNA from 11 individual worms and competitively mapping the sequencing data to both the reference and haplotype assemblies using *bwa mem* [84]. Relative sequence coverage was determined using *bedtools coverage* [82], which was normalised per sample to enable equivalent comparison between samples.

Transcriptome library preparation and sequencing

We made use of the RNAseq data generated for each of the life stages as previously described [21], together with additional full-length cDNA (PacBio Iso-Seq) sequencing performed with the specific aim of generating evidence for the annotation. Three life stages were used in the Iso-Seq analyses: adult female, adult male, and L3. RNA was extracted by homogenising the samples in Trizol in Fast prep (MP bio) followed by chloroform extraction and ethanol precipitation of the RNA, after which 1.2 µg of total RNA was converted to cDNA using SmartSeq2 protocol [85]. Size fractionation of equimolar pooled cDNA of the different stages was performed using a SageELF electrophoresis system (Sage Sciences), from which cDNA size fraction pools spanning 400 to 6,500 bp were collected. To improve the coverage of transcripts that differed in abundance and also in length (and to help minimise library preparation and sequencing biases towards shorter and highly abundant transcripts), we used a cDNA normalisation method followed by size fractionation of the pooled normalised cDNA to recover discrete size fractions of transcripts. The cDNA normalisation, which exploits a duplex-specific nuclease (DSN), was performed using the Trimmer-2 cDNA normalisation kit (Evrogen) following the manufacturer's instructions. Size fractionated cDNA pools were sequenced on five individual flow cells on the RSII instrument. To increase the yield of the sequencing data, a subsequent sequencing run was

performed after pooling the size fractions using a single flow cell on the Sequel instrument. A comparison of the sequences generated by the RSII and Sequel systems is presented in **Additional file 2: Table S4**.

Generation of a high-quality non-redundant Iso-Seq dataset

Full-length Iso-Seq subreads were mapped to the V4 genome using *minimap2* (version 2.6-r639-dirty; [86]) with splice-aware mapping enabled (parameter: *-ax splice)*. The generation of high-quality Iso-Seq reads was performed following the IsoSeq3 pipeline (https://github.com/PacificBiosciences/IsoSeq3), which consisted of four steps: (i) the generation of circular consensus sequencing (CCS) reads from raw subreads (*ccs subreads.bam subreads.ccs.bam --noPolish --minPasses 1*), (ii) the removal of library primers, and reorientation of reads (*lima subreads.ccs.bam primers.fa subreads.demux.bam --isoseq --no-pbi*), (iii) clustering of transcripts, removal of concatemers and trimming of polyA tails (*isoseq3 cluster subreads.demux.primer_5p--primer_3p.bam subreads.isoseq3_unpolished.bam*), and finally, (iv) polishing of the transcripts to generate a high quality set of isoforms (*isoseq3 polish merged_subreads.isoseq3_unpolished.bam merged_subreads.bam merged_subreads.isoseq3_polished.bam*).

Gene model prediction and genome annotation

A schematic of the strategy used for genome annotation is presented in **Additional file 1: Figure S6**. Initial attempts to transfer genome annotations from the V1 assembly to the intermediate V3 assembly using tools such as *Ratt* [87] were unsuccessful, and with the generation of new Iso-Seq data, a new strategy was designed and implemented.

Existing RNAseq data were individually mapped to the V4 genome using *STAR* v2.5.2 [88], before combined into a single bam using *samtools merge*. The merged bam was used as input to *Braker* (version 2.0; [35,89], which can exploit mapped RNAseq information as hints to train the gene prediction tools *GeneMark-EX* and *Augustus*, substantially improving genome annotation. The use of exon, intron, and other hints derived from RNAseq or proteins mapped to the genome can be used to train *Augustus* directly; initial attempts to incorporate mapped Iso-Seq transcripts as an input to *Braker*, or by training *Augustus* using exon and intron hints, were successful and produced superior gene models compared with *Braker* using RNASeq alone. However, as Iso-Seq data encode full-length transcripts that include UTRs in a single sequence, and that in this data there was little discrimination between the depth of coverage between UTR and exon, many UTRs were incorrectly annotated as 5' or 3' exons. As such, only short-read RNAseq data were used as input to *Braker*, resulting in a primary annotation (GFF_1; **Additional file 1: Figure S6**).

To incorporate Iso-Seq data into the genome annotation, we used the *PASA* (v2.2.0) genome annotation tool [90] to map the high-quality full-length transcripts to the genome using both *blat* and *gmap*, akin to the use of a *de novo* transcriptome assembly produced from, for example, *trinity*, to generate a transcript database (https://github.com/PASApipeline/PASApipeline/wiki/PASA_comprehensive_db). The *Braker* GFF_1 was subsequently loaded into *PASA* and integrated with the mapped Iso-Seq transcripts to produce the first round merged annotation (GFF_2; **Additional file 1: Figure S6**).

To incorporate further coding evidence into the annotation, we used *EvidenceModeller* v1.1.1 [36] to apply a differential weighting schema to multiple evidence sources, before

merging them into a single genome annotation. Evidence used included the GFF_1 and GFF_2 annotations from *Braker* and *PASA* as *ab initio* and transcript evidence, respectively, and protein sequences from *C. elegans* and *H. contortus* V1 genome assemblies (obtained from Wormbase ParaSite release 9 caenorhabditis_elegans.PRJNA13758.WBPS9.protein.fa.gz and haemonchus_contortus.PRJEB506.WBPS9.protein.fa.gz, respectively), as well as 110 curated genes; each protein dataset was mapped to the genome using *exonerate* v2.2.0 (parameters: *--model protein2genome --percent 50 --showtargetgff* ; [91]) followed by GFF extraction using *Exonerate_to_evm_gff3.pl* from *EvidenceModeller*. The weighting was applied as follows: PROTEIN (hc_v1 * curated models) = 2; PROTEIN (Ce) = 1; ABINITIO_PREDICTION (GFF_1 from Braker) = 2; TRANSCRIPT (GFF_2 from PASA) = 5.

The merged output of *EvidenceModeller* (GFF_3) was imported back into *PASA* to update gene models with the full-length transcript information, including updates to UTRs and evidence for alternative splicing. Consistent with the author's recommendations, two rounds of iterative update and comparison were performed to maximise the use of the full-length transcript, i.e., Iso-Seq, data (GFF_4; **Additional file 1: Figure S6**). Gene IDs were subsequently renamed with the prefix HCON, and numbered with an 8 digit identifier that incremented by a value of 10 to allow identifiers to be subsequently added for new features if required (GFF_5; **Additional file 1: Figure S6**).

The genome annotation, as well as available evidence used to generate the annotation (including per life stage RNAseq, Iso-Seq [full-length high quality & CCS reads], coverage and repeat tracks), was uploaded to the web portal *Apollo* [37] for further visualisation and manual curation. This platform has been used to provide a means to continually improve

the *H. contortus* genome annotation via manual curation as new information becomes available. These annotation improvements will be subsequently incorporated into WormBase ParaSite.

Annotation statistics were determined using *Genome Annotation Generator (GAG)* [92]. Annotation comparisons to a curated gene set to determine the sensitivity and specificity of the approaches used were performed using *gffcompare* v0.10.1 (http://ccb.jhu.edu/software/stringtie/gffcompare.shtml).

The repetitive content of the genome was analysed using RepeatModeller (v1.0.11) followed by RepeatMasker (v4.0.7). Further analysis of LTRs was performed using LTRharvest and LTRdigest [93] from the GenomeTools v1.5.9 [94] package.

Transcriptome analyses

Quantitative transcriptome analyses were primarily performed using *kallisto* v0.43.1 [95], which quantifies transcripts by pseudo-aligning raw sequencing reads to a transcriptome reference. The transcriptome reference was derived from the reference FASTA and annotation GFF3 data using *gffread* (https://github.com/gpertea/gffread). Raw sequencing reads from each of the life-stages were pseudo-aligned to the reference using *kallisto quant* (parameters: --bias --bootstrap-samples 100 --fusion), before the differential expression between pairwise combinations of life-stages was determined using *sleuth* [96]. For each transcript, differential expression between life-stages was tested using a Wald test, for which a q-value less than 0.05 was deemed significant.

To explore coordinated gene expression throughout the life-stages, we used *clust* v1.8.4 [97] to automatically define and group transcripts by expression level profiles. Transcripts per million (TPM) counts per life-stage were generated from the kallisto analysis described above and used as input to *clust*. To explore putative regulatory sequences that may be shared by co-expressed genes, we searched in each group of clustered transcripts for conserved motifs in the 100 bp region immediately upstream of the start codon for each transcript using *DREME* [98], after which motif-containing sequences were identified using *FIMO* [99] (MEME v5.0.4).

To perform gene set analyses, subsets of *H. contortus* genes together with associated *C. elegans* orthologs were obtained using WormBase ParaSite BioMart [15]. Functional enrichment analyses for both species were performed using g:Profiler (version e97_eg44_p13_d22abce), with a g:SCS multiple testing correction method applying a significance threshold of 0.05 [100].

OrthoFinder v2.2.7 [101] was used to determine orthologous relationships between protein sequences inferred from the *H. contortus* V1 and V4, the McMaster *H. contortus*, as well as *H. placei* and *C. elegans* protein-coding gene annotations.

Trans-splicing: Characterisation of splice leader usage and distribution

We characterised the number and distribution of genes for which mapped RNAseq transcripts contained the canonical *H. contortus* SL1 (GGTTTAATTACCCAAGTTTGAG) and SL2 (GGTTTTAACCCAGTATCTCAAG) sequences. For each splice leader sequence, *cutadapt* v1.13 [102] was used to detect (allowing up to 10% divergence in sequence match [--error-rate =

0.1] and minimum match length of 10 bp [--overlap=10]) and subsequently trim the leader sequences from raw RNAseq reads from each of the life stages, after which all trimmed reads were collated and mapped to the genome using *hisat2* v2.1.0 [103]. The relative coverage of the sequence between the trimmed splice leader site and (i) the start codon (using a window 200 bp upstream and 50 bp downstream of the start codon) and (ii) internal exons (excluding the 1st exon), was determined using *bedtools coverage*. *Weblogo* [104] was used to visualise sequence conservation in the genomic sequence surrounding the splice leader site, and in the splice leader sequences themselves. The spliced leader sequence splice sites for all life stages were combined into a single dataset for plotting and quantitative analyses.

Cis-splicing: Differential isoform usage using Leafcutter

To analyse the extent of differential spliced isoform usage between life-stages, we used *leafcutter* [105](see http://davidaknowles.github.io/leafcutter/ for installation and scripts described below) to quantify the proportion of reads that span introns and then subsequently quantify differential intron usage across samples. First, raw RNAseq reads were mapped to the reference genome using *STAR* [88](parameters: *--twopassMode basic --outSAMstrandField intronMotif --outSAMtype BAM Unsorted*) to generate BAM files per life-stage. Intron junctions (junc files) were generated from the BAM files, which were subsequently clustered using *leafcutter_cluster.py* (parameters: *-m 30 -l 500000*). Differential introns between pairwise life-stages using *leafcutter_ds.R* (parameters: *--min_samples_per_intron=3 --min_samples_per_group=3 --min_coverage=10*), which required the clustered introns and an exon file (derived from the annotation GFF3) as input. Preparation of the data for visualisation of differentially spliced introns was performed using

prepare_results.R; an annotation GTF was required as input, which was generated from the annotation GFF3 using *gffread*. Visualisation was performed using *run_leafviz.R*.

Population genetic analyses

To understand how genetic variation was distributed throughout the chromosomes, we collated sequencing data from our recent global analysis [25] together with in-house archival sequencing data to form the most complete sample cohort of *H. contortus* whole-genome data to date. Reanalysis of the global cohort data was necessary, as the previous analysis was performed using the V3 assembly. In total, 338 samples from 46 populations in 17 countries were analysed; the samples together with geolocation, mitochondrial phylogeny describing genetic connectivity between samples, and ENA references to individual sequencing datasets are available to explore using *Microreact* [106] and can be visualised here: https://microreact.org/project/hcontortus_global_diversity.

Raw sequencing data were mapped to the V4 reference using *bwa mem* [84](parameters: *-Y* and *-M* options were used to use soft clipping for supplementary alignments, and to mark shorter hits as secondary, respectively), followed by identification of duplicate reads using *Picard* (v2.5.0; https://github.com/broadinstitute/picard). Samples for which multiple sequencing lanes were generated were mapped independently and subsequently combined into a single dataset using *samtools merge*. Mapping statistics were generated using *samtools* (v1.3) *flagstat* and *bamtools* (v 2.3.0) *stats*, and compared using *MultiQC* v1.3 [107].

Genetic variation within and between samples was determined using *GATK* (v 4.0.3.0).

GVCFs were first created using *HaplotypeCaller*; to improve efficiency, this process was performed in parallel for each chromosome for each sample separately, and subsequently combined by chromosome using *CombineGVCFs*. Variants were determined per chromosome using *GenotypeGVCFs*, after which the final raw variants were combined using *vcftools concat*. A hard filter was applied using *GATK VariantFiltration* to mark variants using the following criteria: QD < 2.0, FS > 60.0, MQ < 40.0, MQRankSum < -12.5, and ReadPosRankSum < -8.0. Nucleotide diversity and F-stats within and between populations was determined using *vcftools* v0.1.14[108] *--window-pi* and *--weir-fst-pop*, respectively, using a window size of 100 kbp.

To analyse mitochondrial DNA diversity, bcftools view was used to select single nucleotide variants with a read depth > 10, homozygous variant, and with minor allele frequency > 0.01. To generate the phylogeny, variants were used to generate a consensus mitochondrial genome per sample, which were aligned using *mafft* v7.205 [109], and from which maximum likelihood phylogenies were generated using *iqtree* v1.6.5 [110](parameters: *-alrt 1000 -bb 1000 -nt 1*) using automated ModelFinding [111]. Principal component analysis (PCA) of the mitochondrial DNA variation was also performed using the *poppr* [112] package in R.

Sequencing coverage variation between the autosomes and X chromosome for each BAM file was performed using *samtools* [113] *bedcov*, using a BED file of 100 kbp windows generated using *bedtools* [82] *makewindows*.

Calculation of d_N/d_S between *H. contortus* and *H. placei* was performed using *codeml* in

PAML v4.7 [114]. This ratio attempts to quantify selection by comparing the frequency of substitutions at non-silent sites (d_N) with the frequency of substitutions as silent sites (d_S); ratios greater than one indicate positive selection for new protein-coding changes, ratios close to zero indicate neutral evolution, and ratios less than one indicate purifying selection. Note that most genes will have ratios less than one due to the conservation of amino acid sequences. One-to-one orthologs between *H. contortus* and *H. placei* were identified using *Orthofinder* as described above, from which amino acid sequence alignments per orthologous pair were obtained. Codon alignments were generated using *PAL2NAL* v14 [115]. *codeml* was run using m1a and m2a models to examine nearly neutral and positive selection scenarios, from which the log-likelihood (lnL) of each test were compared using a likelihood ratio test. The D test statistic was calculated, where $D = 2*(lnL_{m2a} - lnL_{m1a})$, and compared against a chi-square distribution using *df* = 2 and a significance threshold of 0.05. The resulting gene set was filtered to only include genes with $d_S < 2$, $d_S > 0.02$, and $d_N/d_S > 1$.

To calculate d_N/d_S from the global *H. contortus* dataset, genome-wide variation was filtered using *vcftools* (parameter: --maf 0.01), from which missense and synonymous variants were determined using *SnpEff* v4.3 (parameters: -no-downstream -no-intergenic -no-intron -no-upstream -no-utr). For each gene, the frequency of missense and synonymous variants were each determined from counting variant alleles as a proportion of total genotypes counted for each variant class. The resulting gene set was filtered to only include genes with $d_S < 2$, $d_S > 0.02$, and $d_N/d_S > 2$.

Statistical analysis and visualisation

All statistical analyses, data exploration, and visualisation were performed using R (v3.5.0;

https://www.R-project.org/) unless otherwise stated.

Abbreviations

BUSCO: Benchmarking Universal Single-Copy Orthologs

CCS: circular consensus sequencing

CEGMA: Core Eukaryotic Genes Mapping Approach

DCC: dosage compensation complex

McMaster: draft genome assembly of the *H. contortus* Australian isolate McMaster, published in Schwarz *et al.* 2013

DSN: duplex-specific nuclease

PCA: Principal component analysis

SL: spliced leader

V1: version 1 of the *H. contortus* MHco3(ISE).N1 genome, published in Laing *et al.* 2013

V3: version 3 of the *H. contortus* MHco3(ISE).N1 genome, unpublished

V4: version 4 of the *H. contortus* MHco3(ISE).N1 genome, presented here for the first time

Declarations

Ethics approval and consent to participate

The use of experimental animals to maintain parasite populations for the purposes described in this manuscript was approved by the Moredun Research Institute Experiments and Ethics Committee and were conducted under approved British Home Office licenses in accordance with the Animals (Scientific Procedures) Act of 1986. The Home Office licence numbers were PPL 60/03899 and 60/4421, and the experimental identifiers for these studies were E06/58, E06/75, E09/36 and E14/30.

Consent for publication

Not applicable.

Availability of data and materials

The raw sequence data is available under the ENA accession PRJEB506, with reference to specific sequencing libraries described throughout the text, and/or in Table S3 of Laing et al. 2013. RNAseq data is available from the ENA study ID PRJEB1360. The genome assembly has

been made available at ENA (assembly accession: GCA_000469685.2) and WormBase ParaSite (https://parasite.wormbase.org/Haemonchus_contortus_prjeb506/Info/Index). A static version of the genome annotation used in this paper is available at ftp://ftp.sanger.ac.uk/pub/pathogens/sd21/HCON_V4_GENOME/ (signoff date: 25th Jan 2019), however, the most up-to-date version of the annotation can be accessed at WormBase ParaSite (https://parasite.wormbase.org/Haemonchus_contortus_prjeb506/Info/Index/; [15]). Genome variation data can be visualised using MicroReact (https://microreact.org/project/hcontortus_global_diversity), from which links to ENA accessions for individual sample sequencing data from the global collection can be obtained. Code and workflows used to analyse data are available at https://github.com/stephenrdoyle/hcontortus_genome.

Competing interests

The authors declare that they have no competing interests.

Funding

Work described here was supported by Wellcome's core funding of the Wellcome Sanger Institute (grant WT206194), a Wellcome Trust Project Grant to JSG (grant 067811), and the Biotechnology and Biological Sciences Research Council (BBSRC) grants (BB/M003949/1, BB/P024610/1 and BB/K020048/1). JDN was supported by a Graduate Mobility Award from McGill University.

Acknowledgements

We would like to acknowledge members of the BUG Consortium, and Parasite Genomics group at the Wellcome Sanger Institute, for insightful discussions over the course of this project. We also thank Pathogen Informatics and DNA Pipelines (WSI) for their support and expertise, and the Biosciences Division, Moredun Research Institute, for expert care and assistance with animals.

References

1. C. elegans Sequencing Consortium. Genome sequence of the nematode C. elegans: a platform for investigating biology. Science [Internet]. 1998;282:2012–8. Available from: https://www.ncbi.nlm.nih.gov/pubmed/9851916

2. Nicholls SM, Quick JC, Tang S, Loman NJ. Ultra-deep, long-read nanopore sequencing of mock microbial community standards. Gigascience [Internet]. 2019;8. Available from: http://dx.doi.org/10.1093/gigascience/giz043

3. Stewart RD, Auffret MD, Warr A, Walker AW, Roehe R, Watson M. The genomic and proteomic landscape of the rumen microbiome revealed by comprehensive genome-resolved metagenomics [Internet]. bioRxiv. 2018. Available from: http://dx.doi.org/10.1101/489443

4. Yoshimura J, Ichikawa K, Shoura MJ, Artiles KL, Gabdank I, Wahba L, et al. Recompleting the Caenorhabditis elegans genome [Internet]. Genome Research. 2019. p. 1009–22. Available from: http://dx.doi.org/10.1101/gr.244830.118

5. Low WY, Tearle R, Bickhart DM, Rosen BD, Kingan SB, Swale T, et al. Chromosome-level assembly of the water buffalo genome surpasses human and goat genomes in sequence contiguity. Nat Commun [Internet]. 2019;10:260. Available from: http://dx.doi.org/10.1038/s41467-018-08260-0

6. Belser C, Istace B, Denis E, Dubarry M, Baurens F-C, Falentin C, et al. Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. Nat Plants [Internet]. 2018;4:879–87. Available from: http://dx.doi.org/10.1038/s41477-018-0289-4

7. Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB. Direct determination of diploid genome sequences. Genome Res [Internet]. 2017;27:757–67. Available from: http://dx.doi.org/10.1101/gr.214874.116

8. Chin C-S, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, et al. Phased diploid genome assembly with single-molecule real-time sequencing. Nat Methods [Internet]. 2016;13:1050–4. Available from: http://dx.doi.org/10.1038/nmeth.4035

9. Koren S, Rhie A, Walenz BP, Dilthey AT, Bickhart DM, Kingan SB, et al. De novo assembly of haplotype-resolved genomes with trio binning. Nat Biotechnol [Internet]. 2018;36:1174–82. Available from: http://dx.doi.org/10.1038/nbt.4277

10. International Helminth Genomes Consortium. Comparative genomics of the major parasitic worms. Nat Genet [Internet]. 2018;51:163–74. Available from: http://dx.doi.org/10.1038/s41588-018-0262-1

11. Cotton JA, Bennuru S, Grote A, Harsha B, Tracey A, Beech R, et al. The genome of Onchocerca volvulus, agent of river blindness. Nat Microbiol [Internet]. 2016;2:16216. Available from: http://dx.doi.org/10.1038/nmicrobiol.2016.216

12. Hunt VL, Tsai IJ, Coghlan A, Reid AJ, Holroyd N, Foth BJ, et al. The genomic basis of parasitism in the Strongyloides clade of nematodes. Nat Genet [Internet]. 2016;48:299–307. Available from: http://dx.doi.org/10.1038/ng.3495

13. Li W, Liu B, Yang Y, Ren Y, Wang S, Liu C, et al. The genome of tapeworm Taenia multiceps sheds light on understanding parasitic mechanism and control of coenurosis disease. DNA Res [Internet]. 2018;25:499–510. Available from: http://dx.doi.org/10.1093/dnares/dsy020

14. Tsai IJ, Zarowiecki M, Holroyd N, Garciarrubio A, Sánchez-Flores A, Brooks KL, et al. The genomes of four tapeworm species reveal adaptations to parasitism. Nature [Internet]. 2013;496:57–63. Available from: http://dx.doi.org/10.1038/nature12031

15. Howe KL, Bolt BJ, Shafie M, Kersey P, Berriman M. WormBase ParaSite − a comprehensive resource for helminth genomics. Mol Biochem Parasitol [Internet]. 2017;215:2–10. Available from: http://dx.doi.org/10.1016/j.molbiopara.2016.11.005

16. Kaminsky R, Ducray P, Jung M, Clover R, Rufener L, Bouvier J, et al. A new class of anthelmintics effective against drug-resistant nematodes. Nature [Internet]. 2008;452:176–80. Available from: http://dx.doi.org/10.1038/nature06722

17. Preston S, Jabbar A, Nowell C, Joachim A, Ruttkowski B, Baell J, et al. Low cost whole-organism screening of compounds for anthelmintic activity. Int J Parasitol [Internet]. 2015;45:333–43. Available from: http://dx.doi.org/10.1016/j.ijpara.2015.01.007

18. Bassetto CC, Amarante AFT. Vaccination of sheep and cattle against haemonchosis. J Helminthol [Internet]. 2015;89:517–25. Available from: http://dx.doi.org/10.1017/S0022149X15000279

19. Knox DP, Redmond DL, Newlands GF, Skuce PJ, Pettit D, Smith WD. The nature and prospects for gut membrane proteins as vaccine candidates for Haemonchus contortus and other ruminant trichostrongyloids. Int J Parasitol [Internet]. 2003;33:1129–37. Available from: http://dx.doi.org/10.1016/s0020-7519(03)00167-x

20. Gilleard JS. Haemonchus contortus as a paradigm and model to study anthelmintic drug resistance. Parasitology [Internet]. 2013;140:1506–22. Available from: http://dx.doi.org/10.1017/S0031182013001145

21. Laing R, Kikuchi T, Martinelli A, Tsai IJ, Beech RN, Redman E, et al. The genome and transcriptome of Haemonchus contortus, a key model parasite for drug and vaccine discovery. Genome Biol [Internet]. 2013;14:R88. Available from: http://dx.doi.org/10.1186/gb-2013-14-8-r88

22. Schwarz EM, Korhonen PK, Campbell BE, Young ND, Jex AR, Jabbar A, et al. The genome and developmental transcriptome of the strongylid nematode Haemonchus contortus. Genome Biol [Internet]. 2013;14:R89. Available from: http://dx.doi.org/10.1186/gb-2013-14-8-r89

23. Laing R, Martinelli A, Tracey A, Holroyd N, Gilleard JS, Cotton JA. Haemonchus contortus: Genome Structure, Organization and Comparative Genomics. Adv Parasitol [Internet]. 2016;93:569–98. Available from: http://dx.doi.org/10.1016/bs.apar.2016.02.016

24. Wintersinger J, Mariene GM, Wasmuth J. One species, two genomes: A critical assessment of inter isolate variation and identification of assembly incongruence in Haemonchus contortus [Internet]. bioRxiv. 2018. Available from: http://dx.doi.org/10.1101/384008

25. Sallé G, Doyle SR, Cortet J, Cabaret J, Berriman M, Holroyd N, et al. The global diversity of Haemonchus contortus is shaped by human intervention and climate. Nature Communications [Internet]. 2019;10:4811. Available from:

https://www.nature.com/articles/s41467-019-12695-4

26. Troell K, Engström A, Morrison DA, Mattsson JG, Höglund J. Global patterns reveal strong population structure in Haemonchus contortus, a nematode parasite of domesticated ruminants. Int J Parasitol [Internet]. 2006;36:1305–16. Available from: http://dx.doi.org/10.1016/j.ijpara.2006.06.015

27. Blouin MS, Yowell CA, Courtney CH, Dame JB. Host movement and the genetic structure of populations of parasitic nematodes. Genetics [Internet]. 1995;141:1007–14. Available from: https://www.ncbi.nlm.nih.gov/pubmed/8582607

28. Doyle SR, Cotton JA. Genome-wide Approaches to Investigate Anthelmintic Resistance. Trends Parasitol [Internet]. 2019;35:289–301. Available from: http://dx.doi.org/10.1016/j.pt.2019.01.004

29. Redman E, Grillo V, Saunders G, Packard E, Jackson F, Berriman M, et al. Genetics of mating and sex determination in the parasitic nematode Haemonchus contortus. Genetics [Internet]. 2008;180:1877–87. Available from: http://dx.doi.org/10.1534/genetics.108.094623

30. Bremner KC. Cytological polymorphism in the nematode Haemonchus contortus (Rudolphi 1803) Cobb 1898. Nature [Internet]. 1954;174:704–5. Available from: http://dx.doi.org/10.1038/174704b0

31. Palevich N, Maclean PH, Baten A, Scott RW, Leathwick DM. The Genome Sequence of the Anthelmintic-Susceptible New Zealand Haemonchus contortus. Genome Biol Evol [Internet]. 2019;11:1965–70. Available from: http://dx.doi.org/10.1093/gbe/evz141

32. Sargison ND, Redman E, Morrison AA, Bartley DJ, Jackson F, Naghra-van Gijzel H, et al. A method for single pair mating in an obligate parasitic nematode. Int J Parasitol [Internet]. 2018;48:159–65. Available from: http://dx.doi.org/10.1016/j.ijpara.2017.08.010

33. Doyle SR, Laing R, Bartley DJ, Britton C, Chaudhry U, Gilleard JS, et al. A Genome Resequencing-Based Genetic Map Reveals the Recombination Landscape of an Outbred Parasitic Nematode in the Presence of Polyploidy and Polyandry. Genome Biol Evol [Internet]. 2018;10:396–409. Available from: http://dx.doi.org/10.1093/gbe/evx269

34. Rockman MV, Kruglyak L. Recombinational landscape and population genomics of Caenorhabditis elegans. PLoS Genet [Internet]. 2009;5:e1000419. Available from: http://dx.doi.org/10.1371/journal.pgen.1000419

35. Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. Bioinformatics [Internet]. 2016;32:767–9. Available from: http://dx.doi.org/10.1093/bioinformatics/btv661

36. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol [Internet]. 2008;9:R7. Available from: http://dx.doi.org/10.1186/gb-2008-9-1-r7

37. Dunn NA, Unni DR, Diesh C, Munoz-Torres M, Harris NL, Yao E, et al. Apollo: Democratizing genome annotation. PLoS Comput Biol [Internet]. 2019;15:e1006790. Available from: http://dx.doi.org/10.1371/journal.pcbi.1006790

38. Albritton SE, Kranz A-L, Rao P, Kramer M, Dieterich C, Ercan S. Sex-biased gene expression and evolution of the x chromosome in nematodes. Genetics [Internet]. 2014;197:865–83. Available from: http://dx.doi.org/10.1534/genetics.114.163311

39. Reinke V, Gil IS, Ward S, Kazmer K. Genome-wide germline-enriched and sex-biased expression profiles in Caenorhabditis elegans. Development [Internet]. 2004;131:311–23. Available from: http://dx.doi.org/10.1242/dev.00914

40. Berkseth M, Ikegami K, Arur S, Lieb JD, Zarkower D. TRA-1 ChIP-seq reveals regulators of sexual differentiation and multilevel feedback in nematode sex determination. Proc Natl Acad Sci U S A [Internet]. 2013;110:16033–8. Available from: http://dx.doi.org/10.1073/pnas.1312087110

41. Ellis R. Sex determination in the germ line [Internet]. WormBook. 2006. Available from: http://dx.doi.org/10.1895/wormbook.1.82.1

42. Allen MA, Hillier LW, Waterston RH, Blumenthal T. A global analysis of C. elegans trans-splicing. Genome Res [Internet]. 2011;21:255–64. Available from: http://dx.doi.org/10.1101/gr.113811.110

43. Bektesh S, Van Doren K, Hirsh D. Presence of the Caenorhabditis elegans spliced leader on different mRNAs and in different genera of nematodes [Internet]. Genes & Development. 1988. p. 1277–83. Available from: http://dx.doi.org/10.1101/gad.2.10.1277

44. Redmond DL, Knox DP. Haemonchus contortus SL2 trans-spliced RNA leader sequence. Mol Biochem Parasitol [Internet]. 2001;117:107–10. Available from: https://www.ncbi.nlm.nih.gov/pubmed/11551637

45. Laing R, Hunt M, Protasio AV, Saunders G, Mungall K, Laing S, et al. Annotation of Two Large Contiguous Regions from the Haemonchus contortus Genome Using RNA-seq and Comparative Analysis with Caenorhabditis elegans [Internet]. PLoS ONE. 2011. p. e23216. Available from: http://dx.doi.org/10.1371/journal.pone.0023216

46. Rufener L, Mäser P, Roditi I, Kaminsky R. Haemonchus contortus acetylcholine receptors of the DEG-3 subfamily and their role in sensitivity to monepantel. PLoS Pathog [Internet]. 2009;5:e1000380. Available from: http://dx.doi.org/10.1371/journal.ppat.1000380

47. Blumenthal T. Operon and non-operon gene clusters in the C. elegans genome [Internet]. WormBook. 2014. p. 1–20. Available from: http://dx.doi.org/10.1895/wormbook.1.175.1

48. Maruyama R, Endo S, Sugimoto A, Yamamoto M. Caenorhabditis elegans DAZ-1 is expressed in proliferating germ cells and directs proper nuclear organization and cytoplasmic core formation during oogenesis. Dev Biol [Internet]. 2005;277:142–54. Available from: http://dx.doi.org/10.1016/j.ydbio.2004.08.053

49. Gilleard JS, Redman E. Genetic Diversity and Population Structure of Haemonchus contortus [Internet]. Haemonchus contortus and Haemonchosis – Past, Present and Future Trends. 2016. p. 31–68. Available from: http://dx.doi.org/10.1016/bs.apar.2016.02.009

50. Wilson Sayres MA. Genetic Diversity on the Sex Chromosomes. Genome Biology and Evolution [Internet]. 2018;10:1064–78. Available from: http://dx.doi.org/10.1093/gbe/evy039

51. Doyle SR, Illingworth CJR, Laing R, Bartley DJ, Redman E, Martinelli A, et al. Population genomic and evolutionary modelling analyses reveal a single major QTL for ivermectin drug

resistance in the pathogenic nematode, Haemonchus contortus. BMC Genomics [Internet]. 2019;20:218. Available from: http://dx.doi.org/10.1186/s12864-019-5592-6

52. Viney M. The failure of genomics in biology [Internet]. Trends in Parasitology. 2014. p. 319–21. Available from: http://dx.doi.org/10.1016/j.pt.2014.04.010

53. Wasmuth JD. Realizing the promise of parasite genomics. Trends Parasitol [Internet]. 2014;30:321–3. Available from: http://dx.doi.org/10.1016/j.pt.2014.04.008

54. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol [Internet]. 2018;36:338–45. Available from: http://dx.doi.org/10.1038/nbt.4060

55. Mandrioli M, Manicardi GC. Unlocking holocentric chromosomes: new perspectives from comparative and functional genomics? Curr Genomics [Internet]. 2012;13:343–9. Available from: http://dx.doi.org/10.2174/138920212801619250

56. Kingan SB, Heaton H, Cudini J, Lambert CC, Baybayan P, Galvin BD, et al. A High-Quality De novo Genome Assembly from a Single Mosquito Using PacBio Sequencing. Genes [Internet]. Multidisciplinary Digital Publishing Institute; 2019 [cited 2019 Jan 22];10:62. Available from: https://www.mdpi.com/2073-4425/10/1/62/htm

57. Kashyap SS, Verma S, Voronin D, Lustigman S, Kulke D, Robertson AP, et al. Emodepside has sex-dependent immobilizing effects on adult Brugia malayi due to a differentially spliced binding pocket in the RCK1 region of the SLO-1 K channel. PLoS Pathog [Internet]. 2019;15:e1008041. Available from: http://dx.doi.org/10.1371/journal.ppat.1008041

58. Sargison ND, Redman E, Morrison AA, Bartley DJ, Jackson F, Hoberg E, et al. Mating barriers between genetically divergent strains of the parasitic nematode Haemonchus contortus suggest incipient speciation. Int J Parasitol [Internet]. 2019;49:531–40. Available from: http://dx.doi.org/10.1016/j.ijpara.2019.02.008

59. Wolstenholme AJ. Ion channels and receptor as targets for the control of parasitic nematodes [Internet]. International Journal for Parasitology: Drugs and Drug Resistance. 2011. p. 2–13. Available from: http://dx.doi.org/10.1016/j.ijpddr.2011.09.003

60. Basáñez M-G, Pion SDS, Boakes E, Filipe JAN, Churcher TS, Boussinesq M. Effect of single-dose ivermectin on Onchocerca volvulus: a systematic review and meta-analysis [Internet]. The Lancet Infectious Diseases. 2008. p. 310–22. Available from: http://dx.doi.org/10.1016/s1473-3099(08)70099-9

61. Sabah AA, Fletcher C, Webbe G, Doenhoff MJ. Schistosoma mansoni: Chemotherapy of infections of different ages [Internet]. Experimental Parasitology. 1986. p. 294–303. Available from: http://dx.doi.org/10.1016/0014-4894(86)90184-0

62. Neveu C, Charvet CL, Fauvin A, Cortet J, Beech RN, Cabaret J. Genetic diversity of levamisole receptor subunits in parasitic nematode species and abbreviated transcripts associated with resistance. Pharmacogenet Genomics [Internet]. 2010;20:414–25. Available from: http://dx.doi.org/10.1097/FPC.0b013e328338ac8c

63. Fauvin A, Charvet C, Issouf M, Cortet J, Cabaret J, Neveu C. cDNA-AFLP analysis in levamisole-resistant Haemonchus contortus reveals alternative splicing in a nicotinic acetylcholine receptor subunit [Internet]. Molecular and Biochemical Parasitology. 2010. p. 105–7. Available from: http://dx.doi.org/10.1016/j.molbiopara.2009.11.007

64. Rezansoff AM, Laing R, Martinelli A, Stasiuk S, Redman E, Bartley D, et al. The confounding effects of high genetic diversity on the determination and interpretation of differential gene expression analysis in the parasitic nematode Haemonchus contortus. Int J Parasitol [Internet]. 2019;49:847–58. Available from: http://dx.doi.org/10.1016/j.ijpara.2019.05.012

65. Rödelsperger C, Meyer JM, Prabh N, Lanz C, Bemm F, Sommer RJ. Single-Molecule Sequencing Reveals the Chromosome-Scale Genomic Architecture of the Nematode Model Organism Pristionchus pacificus. Cell Rep [Internet]. 2017;21:834–44. Available from: http://dx.doi.org/10.1016/j.celrep.2017.09.077

66. Tandonnet S, Koutsovoulos GD, Adams S, Cloarec D, Parihar M, Blaxter ML, et al. Chromosome-Wide Evolution and Sex Determination in the Three-Sexed Nematode Auanema rhodensis. G3 [Internet]. 2019;9:1211–30. Available from: http://dx.doi.org/10.1534/g3.119.0011

67. Picard MAL, Cosseau C, Ferré S, Quack T, Grevelding CG, Couté Y, et al. Evolution of gene dosage on the Z-chromosome of schistosome parasites. Elife [Internet]. 2018;7. Available from: http://dx.doi.org/10.7554/eLife.35684

68. Niciura SCM, Tizioto PC, Moraes CV, Cruvinel GG, de Albuquerque ACA, Santana RCM, et al. Extreme-QTL mapping of monepantel resistance in Haemonchus contortus. Parasit Vectors [Internet]. 2019;12:403. Available from: http://dx.doi.org/10.1186/s13071-019-3663-9

69. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics [Internet]. 2011;27:578–9. Available from: http://dx.doi.org/10.1093/bioinformatics/btq683

70. Tsai IJ, Otto TD, Berriman M. Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. Genome Biol [Internet]. 2010;11:R41. Available from: http://dx.doi.org/10.1186/gb-2010-11-4-r41

71. Hunt M, Kikuchi T, Sanders M, Newbold C, Berriman M, Otto TD. REAPR: a universal tool for genome assembly evaluation. Genome Biol [Internet]. 2013;14:R47. Available from: http://dx.doi.org/10.1186/gb-2013-14-5-r47

72. Bonfield JK, Whitwham A. Gap5—editing the billion fragment sequence assembly. Bioinformatics [Internet]. 2010;26:1699–703. Available from: http://dx.doi.org/10.1093/bioinformatics/btq268

73. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive -mer weighting and repeat separation. Genome Res [Internet]. 2017;27:722–36. Available from: http://dx.doi.org/10.1101/gr.215087.116

74. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. Genome Biol [Internet]. 2004;5:R12. Available from: http://dx.doi.org/10.1186/gb-2004-5-2-r12

75. Yeo S, Coombe L, Warren RL, Chu J, Birol I. ARCS: scaffolding genome drafts with linked reads. Bioinformatics [Internet]. 2018;34:725–31. Available from: http://dx.doi.org/10.1093/bioinformatics/btx675

76. Warren RL, Yang C, Vandervalk BP, Behsaz B, Lagman A, Jones SJM, et al. LINKS: Scalable, alignment-free scaffolding of draft genomes with long reads. Gigascience [Internet]. 2015;4:35. Available from: http://dx.doi.org/10.1186/s13742-015-0076-3

77. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One [Internet]. 2014;9:e112963. Available from: http://dx.doi.org/10.1371/journal.pone.0112963

78. Parra G, Bradnam K, Korf I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics [Internet]. 2007;23:1061–7. Available from: http://dx.doi.org/10.1093/bioinformatics/btm071

79. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics [Internet]. 2015;31:3210–2. Available from: http://dx.doi.org/10.1093/bioinformatics/btv351

80. Wick RR, Schultz MB, Zobel J, Holt KE. Bandage: interactive visualization of de novo genome assemblies. Bioinformatics [Internet]. 2015;31:3350–2. Available from: http://dx.doi.org/10.1093/bioinformatics/btv383

81. Nattestad M, Chin C-S, Schatz MC. Ribbon: Visualizing complex genome alignments and structural variation [Internet]. bioRxiv. 2016. Available from: http://dx.doi.org/10.1101/082123

82. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics [Internet]. 2010;26:841–2. Available from: http://dx.doi.org/10.1093/bioinformatics/btq033

83. Gu Z, Gu L, Eils R, Schlesner M, Brors B. circlize implements and enhances circular visualization in R. Bioinformatics [Internet]. 2014;30:2811–2. Available from: http://dx.doi.org/10.1093/bioinformatics/btu393

84. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM [Internet]. arXiv. 2013. Available from: http://arxiv.org/abs/1303.3997

85. Picelli S, Björklund ÅK, Faridani OR, Sagasser S, Winberg G, Sandberg R. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nat Methods [Internet]. 2013;10:1096–8. Available from: http://dx.doi.org/10.1038/nmeth.2639

86. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics [Internet]. 2018;34:3094–100. Available from: http://dx.doi.org/10.1093/bioinformatics/bty191

87. Otto TD, Dillon GP, Degrave WS, Berriman M. RATT: Rapid Annotation Transfer Tool. Nucleic Acids Res [Internet]. 2011;39:e57. Available from: http://dx.doi.org/10.1093/nar/gkq1268

88. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics [Internet]. 2013;29:15–21. Available from: http://dx.doi.org/10.1093/bioinformatics/bts635

89. Lomsadze A, Burns PD, Borodovsky M. Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. Nucleic Acids Res [Internet]. 2014;42:e119. Available from: http://dx.doi.org/10.1093/nar/gku557

90. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK Jr, Hannick LI, et al. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res [Internet]. 2003;31:5654–66. Available from: https://www.ncbi.nlm.nih.gov/pubmed/14500829

91. Slater GSC, Birney E. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics [Internet]. 2005;6:31. Available from: http://dx.doi.org/10.1186/1471-2105-6-31

92. Geib SM, Hall B, Derego T, Bremer FT, Cannoles K, Sim SB. Genome Annotation Generator: a simple tool for generating and correcting WGS annotation tables for NCBI submission. Gigascience [Internet]. 2018;7:1–5. Available from: http://dx.doi.org/10.1093/gigascience/giy018

93. Ellinghaus D, Kurtz S, Willhoeft U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics [Internet]. 2008;9:18. Available from: http://dx.doi.org/10.1186/1471-2105-9-18

94. Gremme G, Steinbiss S, Kurtz S. GenomeTools: a comprehensive software library for efficient processing of structured genome annotations. IEEE/ACM Trans Comput Biol Bioinform [Internet]. 2013;10:645–56. Available from: http://dx.doi.org/10.1109/TCBB.2013.68

95. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol [Internet]. 2016;34:525–7. Available from: http://dx.doi.org/10.1038/nbt.3519

96. Pimentel H, Bray NL, Puente S, Melsted P, Pachter L. Differential analysis of RNA-seq incorporating quantification uncertainty. Nat Methods [Internet]. 2017;14:687–90. Available from: http://dx.doi.org/10.1038/nmeth.4324

97. Abu-Jamous B, Kelly S. Clust: automatic extraction of optimal co-expressed gene clusters from gene expression data. Genome Biol [Internet]. 2018;19:172. Available from: http://dx.doi.org/10.1186/s13059-018-1536-8

98. Bailey TL. DREME: motif discovery in transcription factor ChIP-seq data. Bioinformatics [Internet]. 2011;27:1653–9. Available from: http://dx.doi.org/10.1093/bioinformatics/btr261

99. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics [Internet]. 2011;27:1017–8. Available from: http://dx.doi.org/10.1093/bioinformatics/btr064

100. Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update) [Internet]. Nucleic Acids Research. 2019. p. W191–8. Available from: http://dx.doi.org/10.1093/nar/gkz369

101. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol [Internet]. 2019;20:238. Available from: http://dx.doi.org/10.1186/s13059-019-1832-y

102. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal [Internet]. 2011;17:10. Available from:

http://dx.doi.org/10.14806/ej.17.1.200

103. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods [Internet]. 2015;12:357–60. Available from: http://dx.doi.org/10.1038/nmeth.3317

104. Crooks GE, Hon G, Chandonia J-M, Brenner SE. WebLogo: a sequence logo generator. Genome Res [Internet]. 2004;14:1188–90. Available from: http://dx.doi.org/10.1101/gr.849004

105. Li YI, Knowles DA, Humphrey J, Barbeira AN, Dickinson SP, Im HK, et al. Annotation-free quantification of RNA splicing using LeafCutter. Nat Genet [Internet]. 2017;50:151–8. Available from: http://dx.doi.org/10.1038/s41588-017-0004-9

106. Argimón S, Abudahab K, Goater RJE, Fedosejev A, Bhai J, Glasner C, et al. Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. Microb Genom [Internet]. 2016;2:e000093. Available from: http://dx.doi.org/10.1099/mgen.0.000093

107. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics [Internet]. 2016;32:3047–8. Available from: http://dx.doi.org/10.1093/bioinformatics/btw354

108. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics [Internet]. 2011;27:2156–8. Available from: http://dx.doi.org/10.1093/bioinformatics/btr330

109. Nakamura T, Yamada KD, Tomii K, Katoh K. Parallelization of MAFFT for large-scale multiple sequence alignments. Bioinformatics [Internet]. 2018;34:2490–2. Available from: http://dx.doi.org/10.1093/bioinformatics/bty121

110. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol [Internet]. 2015;32:268–74. Available from: http://dx.doi.org/10.1093/molbev/msu300

111. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods [Internet]. 2017;14:587–

9. Available from: http://dx.doi.org/10.1038/nmeth.4285

112. Kamvar ZN, Tabima JF, Grünwald NJ. Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. PeerJ [Internet]. 2014;2:e281. Available from: http://dx.doi.org/10.7717/peerj.281

113. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics [Internet]. 2009;25:2078–9. Available from: http://dx.doi.org/10.1093/bioinformatics/btp352

114. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol [Internet]. 2007;24:1586–91. Available from: http://dx.doi.org/10.1093/molbev/msm088

115. Suyama M, Torrents D, Bork P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res [Internet]. 2006;34:W609–12. Available from: http://dx.doi.org/10.1093/nar/gkl315

Figure 1

Click here to access/download;Figure;Figure_1.ai

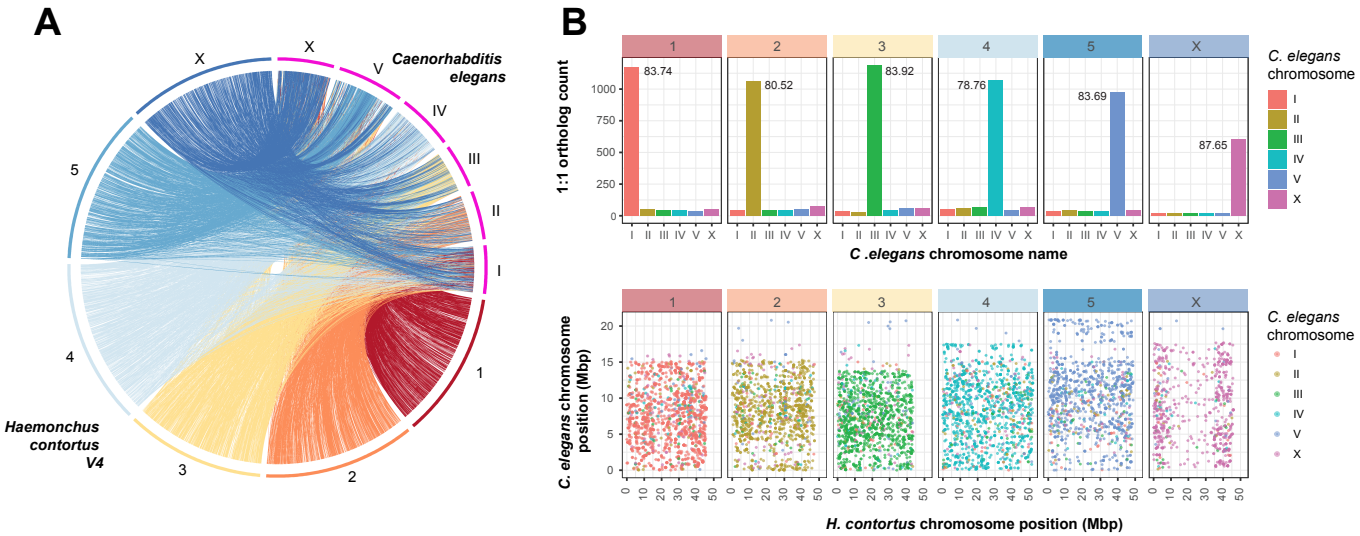


Figure 2

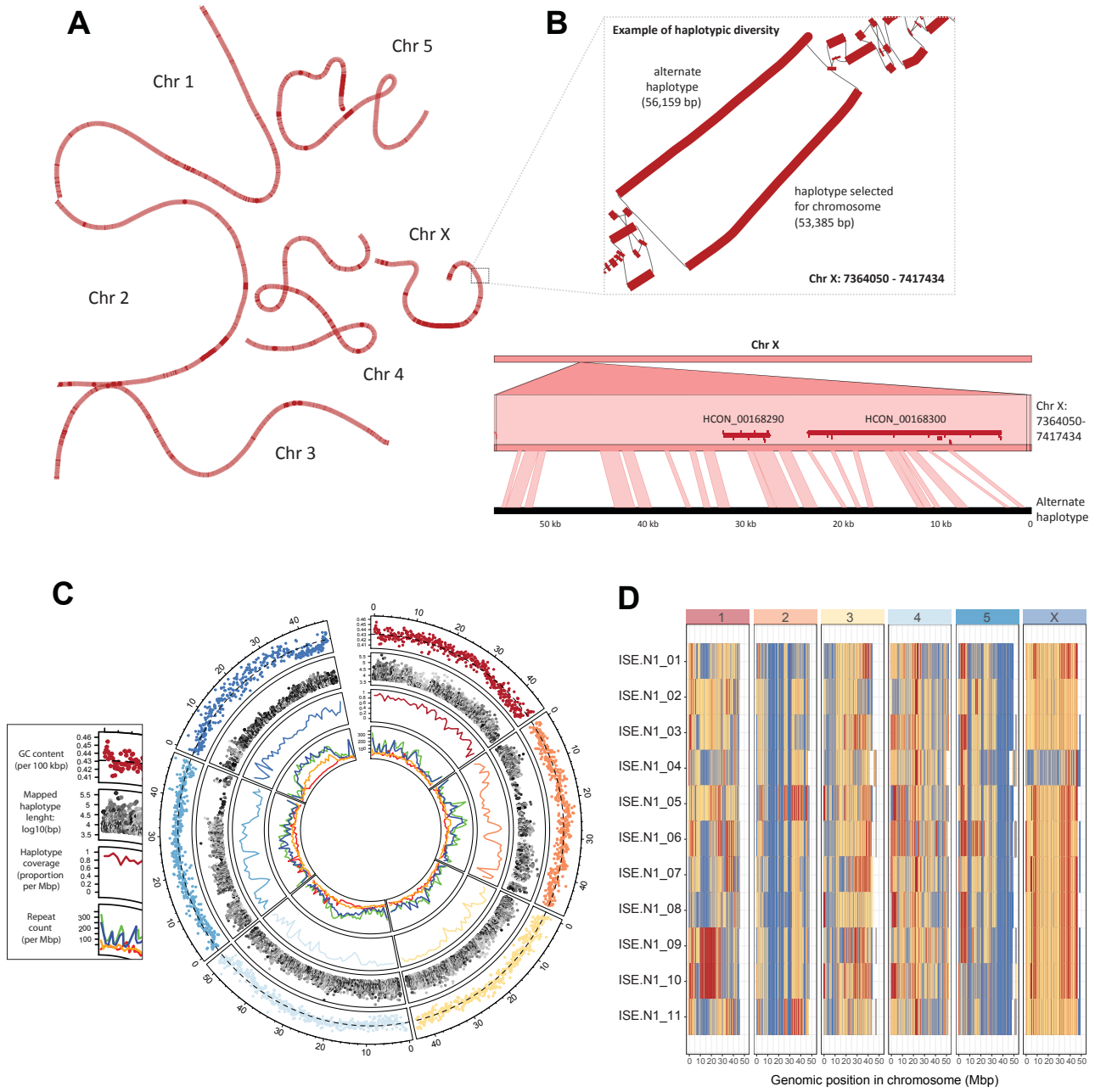


Figure 3

Click here to access/download;Figure;Figure_3.ai

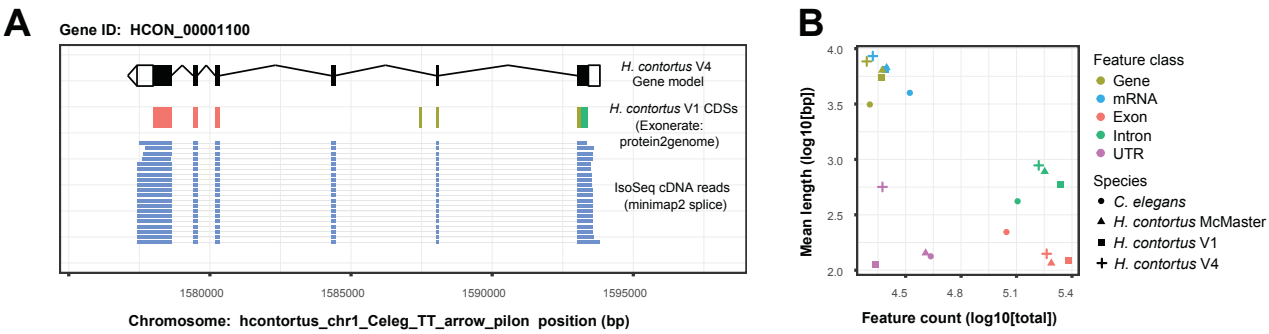


Figure 4

Click here to access/download;Figure;Figure_4.ai

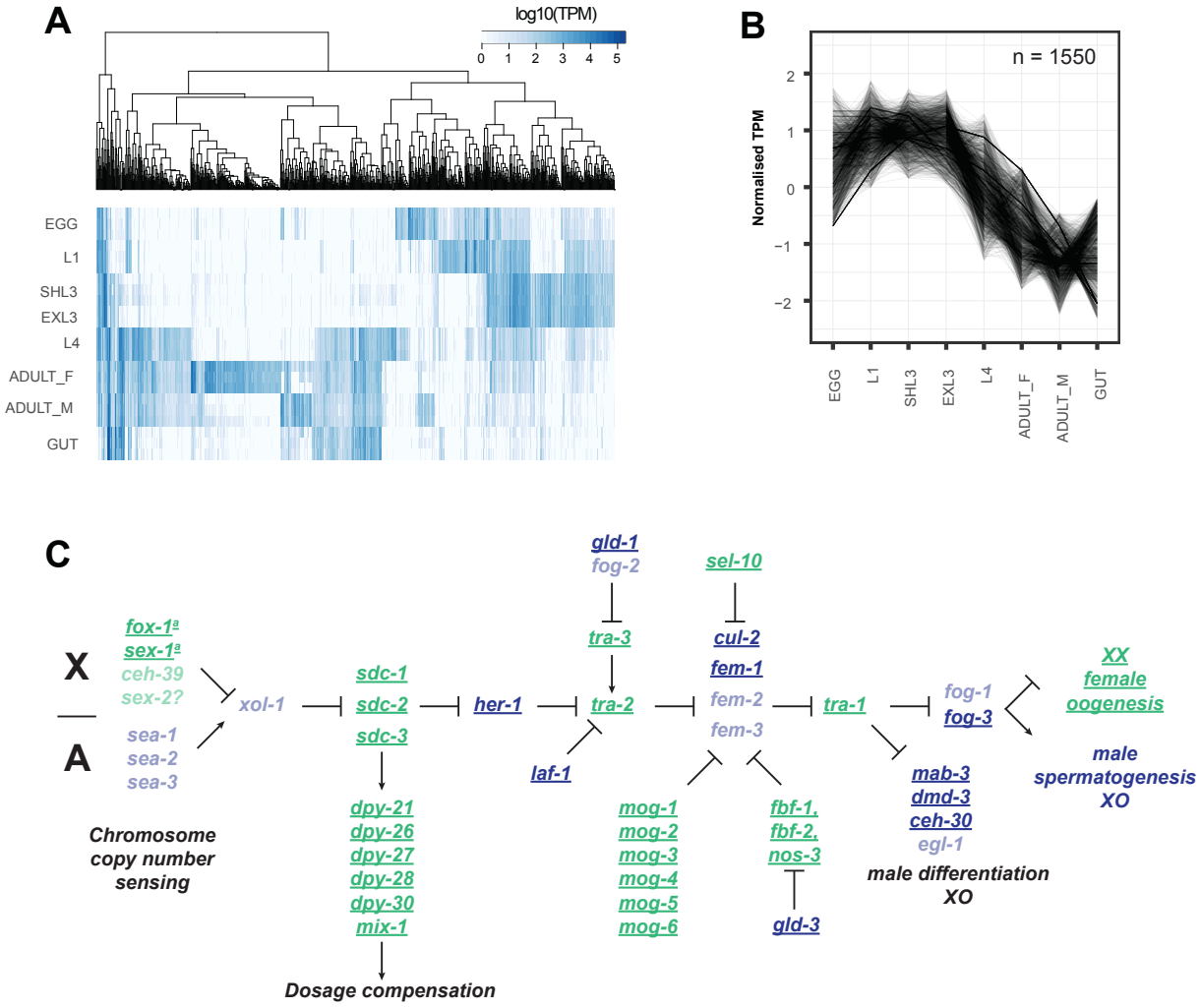


Figure 5

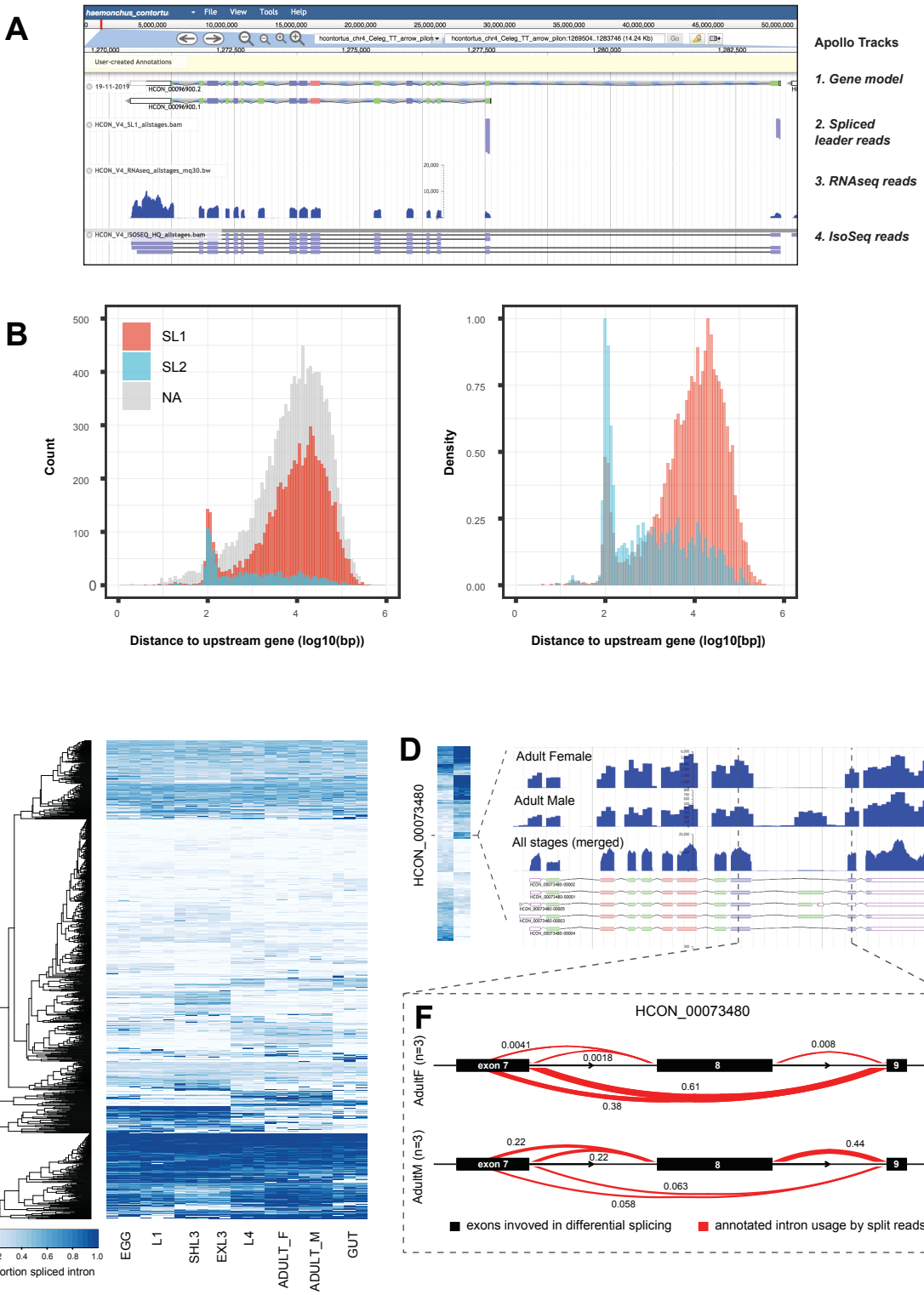
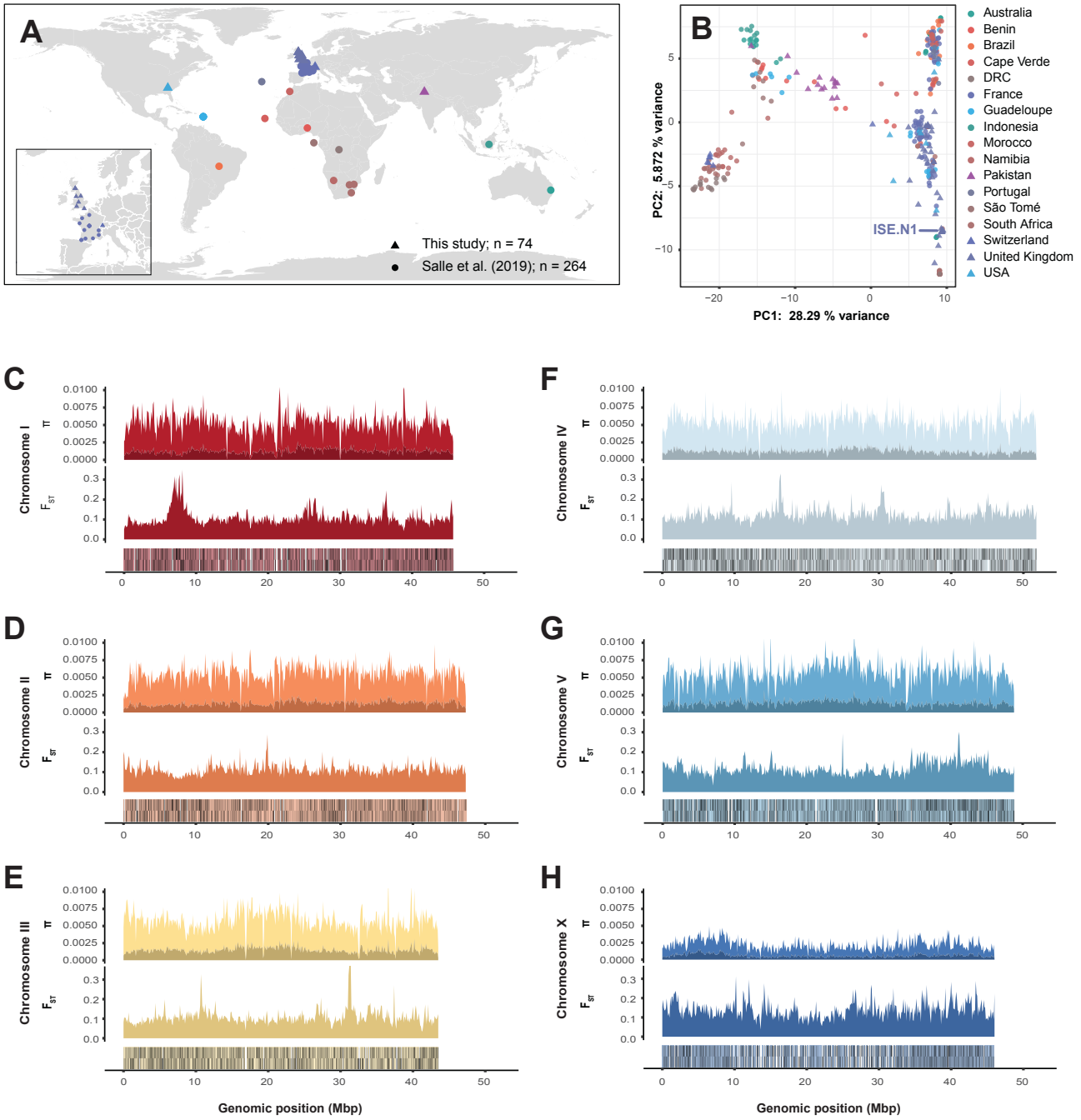


Figure 6



Additional file 1

Additional file 2

wellcome sanger institute

18th February 2020

Dear Editor

We would like to submit the manuscript "**Extensive genomic and transcriptomic variation defines the chromosome-scale assembly of *Haemonchus contortus*, a model gastrointestinal worm**" for publication as a research article in *Genome Biology*. In our study, we report a highly resolved chromosomal genome assembly for *Haemonchus contortus*, a major gastrointestinal pathogen responsible for significant animal health and economic burdens worldwide. We describe chromosome-scale structural variation, patterns of transcriptional change and putative co-regulation throughout its life cycle using a comprehensive genome annotation derived from short- and long-read cDNA sequencing, and within- and between-population genetic diversity from a globally sampled cohort.

Our work represents a step-change in the genetic resources for *H. contortus*. A 370 Mbp highly fragmented and haplotypic draft assembly [n = 23,860; N50 = 0.083 Mbp] for this parasite strain was published in *Genome Biology* in 2013 (Laing et al.; gb-2013-14-8-r88). The utility and impact of this resource is highlighted by the fact that the publication describing the draft genome has been accessed over 20,000 times and cited at least 160 times in the last 7 years. Since completion in early 2019, we have made our chromosomal assembly [283 Mbp; n = 7; N50 = 47.4 Mbp] and all associated resources publicly available before publication via the main community portal for nematode research, WormBase Parasite, and since that time, has been used and referred to in at least 14 publications.

We believe that this work will be of great interest to communities that work on parasitic nematodes. *H. contortus* is the key model representing a group of parasites that includes most of the nematode species of relevance to veterinary health and the hookworms, which are important and neglected human pathogens that infect more than a billion people worldwide. However, we feel our work is particularly suited to *Genome Biology,* rather than a more specialised parasitology journal, as this assembly represents the most complete and accurate reference genome for any nematode species outside *Caenorhabditis*. Our analysis of resolved chromosomes reveals a remarkable pattern of almost complete conservation of chromosome content between the evolutionarily related *H. contortus* and *C. elegans,* but almost no conversation of gene order, and yet, important contrasts with this model, for example, in the patterns of *cis*- and *trans*-splicing, and chromosome dosage compensation components. Further, this work will be broadly applicable to readers of *Genome Biology* with an interest in chromosome-scale biology and potential insights gained from genetic and transcriptomic analyses derived from long molecule sequencing, especially for species with high genetic diversity.

We look forward to hearing from you.

Yours faithfully,

Stephen Doyle & James Cotton
Corresponding authors, on behalf of all authors

Wellcome Sanger Institute

Wellcome Sanger Institute
Wellcome Genome Campus
Hinxton, Cambridge
CB10 1SA

Genome Research Limited
Registered Office:
215 Euston Road
London NW1 2BE

A company registered
in England (No. 2742969)
and a charity registered
in England (No. 1021457)