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

Stephen R. Doyle et al.†

*Haemonchus contortus* is a globally distributed and economically important gastrointestinal pathogen of small ruminants and has become a key nematode model for studying anthelmintic resistance and other parasite-specific traits among a wider group of parasites including major human pathogens. 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 for the MHco3(ISE).N1 isolate. We show a remarkable pattern of conservation of chromosome content with *Caenorhabditis elegans*, but almost no conservation of gene order. Short and long-read transcriptome sequencing allowed us to define coordinated transcriptional regulation throughout the parasite’s life cycle and refine our understanding of cis- and trans-splicing. Finally, we provide a comprehensive picture of chromosome-wide genetic diversity both within a single isolate and globally. These data provide a high-quality comparison for understanding the evolution and genomics of *Caenorhabditis* and other nematodes and extend the experimental tractability of this model parasitic nematode in understanding helminth biology, drug discovery and vaccine development, as well as important adaptive traits such as drug resistance.

†A list of authors and their affiliations appears at the end of the paper.
The recent introduction of long-read DNA sequencing from Pacific Biosciences (PacBio) and Oxford Nanopore has completely changed the quality and contiguity of assemblies of both large and small genomes. The ability to generate highly contiguous and closed prokaryote genomes is becoming routine, even from metagenomic samples containing multiple strains or species, and high-quality chromosome-scale genomes for many eukaryotic species are rapidly becoming available. 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. Ideally, the DNA sequence will be obtained 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, the excessive diversity presents a significant challenge to the assembly process, and consequently, typically results in fragmented assemblies that contain misassembled and excessive haplotypic sequences.

Parasitic helminths are a diverse group of organisms for which significant efforts have been made to develop genomic resources. For some species, high-quality near-complete assemblies have been achieved using multiple complementary sequencing and mapping technologies, and usually with extensive manual improvement. Similar quality assemblies for a few other species are available from WormBase ParaSite, but await formal sequencing and mapping technologies, and usually with extensive manual improvement. For other species, assembly sizes are less than 473 Mb (scaffold N50 = 3.8 Mb), representing only 75.6% of the 369.8 Mb V1 draft assembly length (Table 1); the reduction in assembly size is 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 substantial reduction from the 41,663 gaps present in the V1 assembly (Table 1). We identified 242 of 248 (97.98%) universally conserved orthologs measured by the Core Eukaryotic Genes Mapping Approach (CEGMA; Supplementary Table 1), and 859 of 982 (87.4%) metazoan Benchmarking Universal Single-Copy Orthologs (BUSCOs). The remaining missing orthologs show phylogenetic structure among clade V nematodes (Supplementary Fig. 4), suggesting that many are truly missing from the genomes of these species rather than due to assembly artefacts. The V4 assembly contained 141 more full-length single-copy orthologs and 159 fewer duplicated BUSCOs than identified in the V1 genome (Supplementary Table 1); considering that we identified a similar average number of CEGMA orthologs per core gene in V4 and in the C. elegans genome (1.1 in V4, compared with 1.09 in C. elegans), erroneously-duplicated sequences have been reduced in V4 compared with the draft H. contortus assemblies (V1: 85.89%
Fig. 1 Chromosomal syntenyl between *Haemonchus contortus* and *Caenorhabditis elegans*. a Genome-wide comparison of chromosomal organisation between *H. contortus* V4 and *C. elegans*. 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. Hypothetically, if *C. elegans* and *H. contortus* chromosomes were completely colinear, the genomic positions of orthologs would show a positive linear relationship between the two species; however, this is not the case, whereby an almost complete reshuffling of orthologs was observed.

Table 1 Genome assembly summary statistics.

|                        | *H. contortus* V4 | *H. contortus* V1<sup>a</sup> | *H. contortus* McMaster<sup>b</sup> | *H. contortus* New Zealand<sup>c</sup> | *C. elegans* WB<sup>d</sup> |
|------------------------|-------------------|-------------------------------|-----------------------------------|--------------------------------|-------------------|
| Genome size (bp)       | 283,439,308       | 248,771,548                   | 369,846,877                       | 465,720,674                   | 100,286,401       |
| Scaffolds              | 7                 | 3,907                         | 23,860                           | 14,419                        | 172,892           |
| Scaffold N50 (bp)      | 47,382,676        | 167,270                       | 83,287                           | 56,328                        | 17,493,829        |
| Scaffolds ≥N50         | 3                 | 174                           | 1,151                            | 1,684                         | 3                 |
| Contigs                | 192               | 4,226                         | 65,523                           | 55,322                        | 3                 |
| Contig N50 (bp)        | 3,801,457         | 117,552                       | 20,808                           | 12,900                        | 3                 |
| Contigs ≥N50           | 20                | 439                           | 4,943                            | 6,646                         | 3                 |
| Scaffold N content (bp)| 5,699,711         | 14,755,378                    | 23,804,39                        | 20,495,720                    | 0                 |
| Scaffold gaps          | 185               | 323                           | 41,663                           | 40,903                        | 172,892           |

*Haem* V1<sup>a</sup>: *haemonchus_contortus.PRJEB506.WBPS11.genomic.fa*

*McMaster*<sup>b</sup>: *haemonchus_contortus.PRJNA205202.WBPS11.genomic.fa*

*New Zealand*<sup>c</sup>: ENA accession GCA_007637855.2: GCA_007637855.2_ASM763785v2_genomic.fna

*C. elegans* WB<sup>d</sup>: *caenorhabditis_elegans.PRJNA13758.WBPS9.genomic_softmasked.fa.*

complete and 1.59 average orthologs per CEGMA gene; McMaster: 70.56% complete and 1.43 average orthologs CEGMA gene; Supplementary Table 1).

An assembly of a New Zealand (NZ) *H. contortus* strain was recently made available<sup>33</sup>, which used our V4 chromosomal assembly to scaffold a highly fragmented draft assembly (contigs = 172,899) into seven chromosome-scale scaffolds. The expansion of the NZ 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 the increased proportion of orthologs detected, on average, per partial and complete CEGMA gene (1.42 and 1.71, respectively, vs 1.1 and 1.22 in the V4 assembly; Supplementary Table 1) suggest that contaminating haplotypic sequences have erroneously been incorporated into the NZ assembly. Similarly, 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 insertion and/or deletions from uncorrected PacBio reads and/or assembly gaps (n = 172,892) are affecting coding sequences broadly throughout the NZ assembly. While multiple chromosome-scale assemblies from the same species provide a unique opportunity for comparative analysis, the striking differences between the NZ and the V4 assemblies are unlikely to represent genuine biological variation between strains.

### Resolving haplotypic diversity and repeat distribution within the chromosomes

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 curation. Mapping of the additional haplotypes to the reference-assigned haplotype (Fig. 2a) highlighted both the diversity between the haplotype and cosmoramolocally-placed sequences (Fig. 2b, c second circle) and coverage of the haplotypes, spanning approximately 66.8 ± 24.1% on average of the chromosome sequences (Fig. 2c third circle; range 85.1 ± 12.3% [chromosome I] to 51.8 ± 16.6%
**Fig. 2 Haplotype and repeat diversity within and between chromosomes.**

*a* Genome graph derived from the whole-genome alignment of chromosome and alternate haplotype assemblies. The graph is coloured to reflect alternative haplotype density, distinguishing regions with longer reference only-haplotypes (dark red) from those with shorter, more diverse haplotypes (light red). 

*b* A close-up view of the genome graph structure (box) highlights an example of alternate haplotypes from 7.36 to 7.42 Mbp on Chr X; alternative haplotype blocks are represented by thick red lines and paths between haplotypes as thin black lines. DNA alignment of the 56 kbp alternate haplotype to the genome demonstrates the extent of the genetic diversity between these sequences; in this region, 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. The 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). 

*d* Haplotype switching between the chromosome scaffolds 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.
Whole-genome short-read sequencing of 11 single worms (10 male and 1 female, identified by full- or half-coverage across the X chromosome for XX females and XO males, respectively) followed by competitive read mapping revealed clear evidence of haplotype switching between the chromosomal reference and alternate haplotype assembly based on differences in the sequencing coverage throughout the genome and between samples (Fig. 2d). 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, despite being semi-inbred through a limited number of single-pair matings34, complex haplotypes are segregating in this population, and that the frequency and diversity of these haplotypes are non-randomly distributed throughout the genome.

The use of multiple long-molecule technologies allowed a greater representation of repetitive sequences in the V4 assembly, accounting for approximately 36.43% (compared with 30.34% in V1) of the genome (Supplementary Table 2). Particular repeat classes, including LINEs, LTRs and DNA elements were enriched towards the middle of the chromosomes and the chromosome ends (Fig. 2c inner circle; Supplementary Fig. 5). This distribution pattern is negatively correlated with observed recombination rate domains in both *H. contortus*35 and *C. elegans*36, suggesting that transposons 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 (Fig. 2c inner circle; Supplementary Fig. 5). 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 (Fig. 2c outer circle).

**Generation of a high-quality transcriptome annotation incorporating short and long reads.** Considering the broad-scale changes from V1 to V4 assemblies, we undertook a *de novo* gene finding strategy to annotate the V4 genome. To do so, we used our Illumina short-read RNA-seq collection sampled from seven stages of the parasite life cycle (described in the analysis of the V1 genome23) and supplemented this with PacBio Iso-Seq long-read sequencing (Supplementary Table 3) of full-length cDNA from adult male, female, and juvenile L3 parasites. The overall annotation strategy is presented in Supplementary Fig. 6 and consisted of four key stages: (i) RNA-seq evidence was used to train an initial gene finding using Braker37; (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)38; (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. Together, this strategy achieved a high level of sensitivity and precision, and represented substantial improvement over the V1 annotation (Supplementary Table 4).

The 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 (Fig. 3a), 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 (Fig. 3b; Supplementary Table 5); 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. These longest genes were predominantly located toward the centre of the X chromosome, suggesting that gene length expansion is driven by the accumulation of intronic repetitive elements enriched in this region as described above (Fig. 2c inner circle; Supplementary Fig. 5). Concordantly, both the gene number and gene coverage of 100 kbp windows were significantly lower on the X chromosome relative to the autosomes (pairwise t-test between X and autosomes; *P* = 2.2E−16). 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) (Supplementary Table 6). Surprisingly, a greater number of one-to-one orthologs are identified between V4 and *Haemonchus 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*39, 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.

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 cycle23,24. We revisited this using the current annotation and, for visual comparison, present the top 1,000 most variable transcripts across all life stage samples in the MHC03 (ISE) N1 datasets (Fig. 4a; Supplementary Data 3). The transcriptional transition between the juvenile free-living life stages and the more mature parasitic stages is striking, reflecting not only the development toward reproductive maturity but also drastic 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 (Supplementary Fig. 7; Supplementary Data 4). 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; Fig. 4b) and (ii) genes with lower expression during free-living stages that subsequently increased in expression in parasitic stages (cluster 8; *n* = 1,542 genes). Functional characterisation of the cluster 1 genes revealed 45 enriched gene ontology (GO) terms (https://biit.cs.ut.ee/gplink/l/8iECKGWQQS) that predominantly described ion transport and channel activity, as well as transmembrane signalling receptor activity. For cluster 8, 22 GO terms were enriched (https://biit.cs.ut.ee/gplink/l/0Iojq62h0Re) 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/b3xB8hrJB). Finally, we explored whether co-expressed genes contained shared sequence motifs in their 5’ UTR that may 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 enriched GO terms (Supplementary Data 5), for example, cysteine-type peptidase activity in cluster 8 (GATAAGR; motif E-value = 2.90E−02; GO P = 4.181E−13), DNA replication in cluster 15 (AAAAATVA; motif E-value = 3.50E−02; GO P = 6.269E−4), and molting cycle, collagen and cuticulin-based cuticle in cluster 16 (GATAAGM; motif E-value = 2.30E−02; GO P = 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 (Supplementary Data 3). Consistent with C. elegans, sex-biased expression is not randomly distributed throughout the genome40,41; 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; $\chi^2 = 175.83$, df = 1, $P = 3.84E−40$), whereas more female-biased genes are X-linked than expected (24.6% [110/476]; $\chi^2 = 33.59$, 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 development, including developmentally upregulated genes (e.g., smad4, transcription factor AP2, vas-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-2042) 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.31,32. 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 similar, this has not been previously shown. Comparison of H. contortus RNA-seq data between males and females showed no significant difference in the distribution of expression between autosomes and X-linked genes; only 9.24% of genes (110 + 146 of 2,770 X-linked genes) are differentially expressed on the X chromosome between sexes suggests 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 (Fig. 4c; Supplementary Data 5; Adapted from WormBook43). In C. elegans, chromosome dosage is determined by measuring the expression of specific 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, whereas the autosomal dosage associated genes sex-1, −2, and −3 are missing. Five of the six C. elegans orthologs associated with the dosage compensation complex (DCC) were found (dpy-26 is missing), however, the genes xol-1 and sdc-1, −2, and −3 that act to initiate the recruitment of the DCC are absent. Thus, while the DCC may still mediate dosage compensation in H. contortus, the mechanism used to determine the X-to-autosome ratio is likely different from C. elegans. Identification of many 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 development, respectively, in C. elegans.

Transcriptome complexity is generated by extensive cis- and trans-splicing. Spliced leader (SL) trans-splicing has been extensively described in C. elegans44 and has been shown previously to operate in H. contortus45–47 (Fig. 5a). To explore this further in the V4 annotation (Supplementary Fig. 8a, b), we identified 7,293 (37.4%) and 1,139 (5.8%) genes with evidence of SL1 and SL2 trans-splicing, respectively (Supplementary Data 7). 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 indicated additional non-annotated isoforms were present in the transcriptomes. While we 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 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 (Supplementary Fig. 8c), and conservation in the genome.
sequence immediately upstream of the splice site, particularly for SL2 (Supplementary Fig. 8d). In *C. elegans*, SL1 sequences are typically associated with 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 region 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. In *H. contortus*, the intergenic distance upstream of genes with SL1 sequences is largely consistent with that of genes without SL1 sequences (Fig. 5b left plot; median = 8,557 bp [SL1-spliced], 8,603.5 bp [no 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 found less than 100 bp downstream of the nearest gene as expected (median distance = 566.5 bp), a substantial proportion (432 [35.4%]) are found with a large intergenic distance (>2 kbp) to the nearest upstream gene (Fig. 5b right plot). These observations suggest that while broadly consistent with *C. elegans*, (i) *H. contortus* SL1-sequences can be found in downstream genes within a polycistron, and may reflect evidence of hybrid operons that contain a second, internal promoter to control the expression within a polycistron, and may reassemble evidence of hybrid operons that contain a second, internal promoter to control the expression within a polycistron, and may function independently, and/or (iii) recognition of SL2 splicing within an operon remains effective over large physical distances, for example, the 2,064 bp region between *deg-3* (HCON_00039370) and *des-2* (HCON_00039380) operon of *H. contortus*. Although these variations have been described in *C. elegans*, they seem to be more the norm rather than the exception in *H. contortus*.

We extended the splice variation analysis to characterise cis-splicing of transcripts and differential isoform usage between life
We identify 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 (Fig. 5c; Supplementary Table 7 [summary], Supplementary Data 8 [gene lists]). For example, 231 of 1,068 genes are differentially spliced between the 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 (Fig. 5d). DAZ-1 is an RNA-binding protein involved in oogenesis but not spermatogenesis in C. elegans; 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)(Fig. 5e). Only a single transcript is described in C. elegans, however, five isoforms are annotated in the H. contortus genome, with sex-specific expression of exon 9 that is present in males and absent in females. Further, 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 approximately 60% of transcripts in female worms (Fig. 5f).

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. We have characterised the chromosomal distribution of genetic diversity by exploiting a recent large scale analysis of global genetic diversity together with data from 74 additional worms from previously unsampled regions including Pakistan, Switzerland, USA, and the United Kingdom (Fig. 6a; https://microreact.org/project/hcontortus_global_diversity). Principal component analysis (PCA) of mtDNA diversity (Fig. 6b; Supplementary Fig. 9) showed that relationships between isolates are congruent with those previously described; 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. The US and UK samples consisted of both laboratory-maintained strains and field isolates; notably, apart from MHco3(ISE).N1, the laboratory strains 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 (Supplementary Fig. 10; Kokstad population), perhaps reflecting its African origins.

Genome-wide nucleotide diversity (π) is broadly consistent chromosome-wide between the autosomes (Fig. 6c–g). The high frequency of SNPs private to only a single population (relative to variants shared between two or more populations) represents a considerable proportion of the overall variation (23.55% of total variation), consistent with previous reports describing high regional genetic diversity. On the X chromosome, nucleotide diversity is much lower than neutral expectation (i.e., \(\pi_x / \pi_{autosome} \approx 0.75\)) at a \(\pi_x / \pi_{autosome}\) median ratio of 0.363 (Fig. 6h), 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, variation in X chromosome coverage (Supplementary Fig. 11a) increases in populations that are more genetically distinct from the reference strain (Supplementary Fig. 11b).

Finally, we reveal multiple discrete regions in each chromosome with high \(F_{ST}\) among populations (Fig. 6b–h; bottom plots). Consistent with previous reports, strong genetic selection for anthelmintic resistance has impacted genome-wide variation; the most differentiated region is found on chromosome 1 surrounding the beta-tubulin isotype 1 locus (HCON_00005260; 7,027,492 to 7,031,447 bp), a target of benzimidazole-class anthelmintics for which known resistant mutations have been identified in these populations, as well as a broad region of elevated \(F_{ST}\) from approximately 35 to 45 Mb on chromosome 5 previously shown to be associated with ivermectin resistance in controlled genetic crosses. The top 1% of \(F_{ST}\) outlier regions per chromosome (280 10 kb regions in total; Supplementary Data 9) contains 269 genes enriched for 10 GO terms (7 MF & 3 BP; https://biit.cs.ut.ee/gplink/l/1EEZzm6USa), largely describing proteolytic and cysteine-type peptidase activity.

Discussion

The H. contortus MHco3(ISE).N1 V4 assembly presented 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 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, these assemblies are typically 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, 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 high 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* tolerate high genetic diversity during homologous pairing of chromosomes at meiosis\(^5\). Despite this diversity, 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 necessary investment required to improve a draft assembly to chromosomes. However, as DNA input requirements for long-read sequencing decrease\(^5\), the ability and ease with which highly contiguous genomes can be generated for small and genetically diverse species will be greatly enhanced.

Completion of the X chromosome represents one of the key advancements in the current genome assembly. In doing so, we...
revealed striking differences between the autosomes and X chromosome in (i) the repeat distribution and content, (ii) an enrichment for the longest genes of the genome, (iii) lower gene density and gene coverage, (iv) the lack of apparent genetic diversity, and (v) 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. Similarly, the distinct repeat domain structure does suggest that recombination patterns may be different 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 polyandrous11,32, and highly fecund, so that reproductive fitness is likely over-dispersed 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 compounds59. The dichotomy between the 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. Long-read sequencing data 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 putatively co-regulated gene sets, providing greater insight into coordinated developmental programmes than has been described previously. The broad-scale coordinated gene expression associated with ion transport, channel activity, and transmembrane signalling receptor activity throughout the life-cycle was of interest, particularly as ion channels and receptors are the primary targets of several anthelmintic drugs used to control *H. contortus* and other helminth species of human and veterinary importance59. The downregulation of sensory and signalling pathway-related genes from pairwise comparisons of juvenile to adult stages has been described previously in both parasitic24,60–63 and free-living64,65. Clade V nematodes, which suggests this is a broadly conserved phenomenon throughout nematode development. However, a better understanding of these changes may explain why some parasites at particular life stages are less sensitive to these drugs66,67, 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, a feature which is increasingly implicated in the sensitivity of drug targets in nematodes such as *H. contortus*; for example, sex-dependent sensitivity of emodepside is mediated by differential-splcing of the potassium channel *slo*11,38, and similarly, expression of truncated acetylatedine receptor subunits *acr-8* and *unc-63* isoforms are associated with levamisole resistance68,69. Our results almost certainly underestimate the extent of cis-splliced variation within and between life stages; stringent filtering resulted in only a subset of genes tested (range: 759–1,156 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 (Supplementary Table 14). Further, each pairwise comparison identified cryptic splice sites, indicating that additional 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 described70. 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, the V4 assembly allowed the extent of genome rearrangements between *C. elegans* and *H. contortus* to be characterised, revealing numerous gene and repeat family expansions that were missed in the previous assembly. As further genomes are completed from the Strongylid clade, determining the rates of structural evolution within and between chromosomes 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 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 nematodes22. Our chromosomal assembly has been instrumental in mapping genetic variation associated with ivermectin54 and monepantel53 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 limitations of interpreting such variation using less-contiguous genome assemblies as a reference30.

Conclusions

Our chromosome-scale reference genome and comprehensive annotation provides a step forward toward resolving the complete genome composition and transcriptional complexity of *H. contortus*. The increasing availability of long-read sequencing together with advances in the sequencing of small diverse organisms holds great promise for the rapid generation of high-quality contiguous genomes for organisms such as *H. contortus*; future sequencing, including the generation of new genomes and transcriptomes de novo from individuals sampled throughout its geographic range, will provide great insight into how genetic variation is maintained and 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 was isolated from the semi-inbred strain of *H. contortus*, MHco3(ISE).NI. 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.35), and maintained as previously described36.

Iterative improvement of the genome assembly. Our approach to improving the V1 genome to the final V4 chromosomal assembly is outlined in Supplementary Fig. 1. The sequence data generated for this study are described in Supplementary Data 1.

Initial manual improvement on the V1 genome focused on iterative scaffolding with **SSPACE**22 and gap-filling with **IMAGE**73 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 Reaprer, and visual inspection using GapCloser. Substantial genetic variation was present in the sequence data due to the sequencing disparity from a pool of unique molecules. To address this, we sequenced all chromosomes, and the 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 inability of scaffolding tools to deal with the changing rates of haplotypic variation, so we attempted to solve this manually in GapCloser. As we had identified incorrectly phased haplotypes, we found that in each scaffold, 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.opengen.com/) optical mapping data. Optical mapping was performed following methods described previously 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 mapping alone was unsuccessful, and therefore, optical contigs were generated using DNA sequence reads 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 from sequencing the PacBio reads. A total of 32.1 GB of reads were generated from 33 flow cells, representing approximately 133.8× coverage (based on the estimated genome size between haplotypes. Using the OpGen (http://www.opgen.com/) optical mapping data. Optical mapping was performed following methods described previously 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 mapping alone was unsuccessful, and therefore, optical contigs were generated using DNA sequence reads 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 from sequencing the PacBio reads. A total of 32.1 GB of reads were generated from 33 flow cells, representing approximately 133.8× coverage (based on the estimated genome size between haplotypes. Using the OpGen (http://www.opgen.com/) optical mapping data. Optical mapping was performed following methods described previously 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 mapping alone was unsuccessful, and therefore, optical contigs were generated using DNA sequence reads 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.

This iterative process resulted in the production of high-quality genome assembly graphs. We surmised that much of the assembly fragmentation was due to the inability of scaffolding tools to deal with the changing rates of haplotypic variation, so we attempted to solve this manually in GapCloser. As we had identified incorrectly phased haplotypes, we found that in each scaffold, 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.opengen.com/) optical mapping data. Optical mapping was performed following methods described previously 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 mapping alone was unsuccessful, and therefore, optical contigs were generated using DNA sequence reads 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.

This iterative process resulted in the production of high-quality genome assembly graphs. We surmised that much of the assembly fragmentation was due to the inability of scaffolding tools to deal with the changing rates of haplotypic variation, so we attempted to solve this manually in GapCloser. As we had identified incorrectly phased haplotypes, we found that in each scaffold, 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.opengen.com/) optical mapping data. Optical mapping was performed following methods described previously 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 mapping alone was unsuccessful, and therefore, optical contigs were generated using DNA sequence reads 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.

This iterative process resulted in the production of high-quality genome assembly graphs. We surmised that much of the assembly fragmentation was due to the inability of scaffolding tools to deal with the changing rates of haplotypic variation, so we attempted to solve this manually in GapCloser. As we had identified incorrectly phased haplotypes, we found that in each scaffold, 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.opengen.com/) optical mapping data. Optical mapping was performed following methods described previously 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 mapping alone was unsuccessful, and therefore, optical contigs were generated using DNA sequence reads 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.
To incorporate Iso-Seq data into the genome annotation, we used the PASSA (v2.0) genome annotation tool[25] to map the high-quality full-length transcripts to the genome. This was done using a BLAT and group, 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/PASAP_comprehensive.db). The Brazer GFF was subsequently loaded into PASSA and integrated with the mapped Iso-Seq transcripts to produce the first round merged annotation (GFF3; Supplementary Fig. 4).

To incorporate further coding evidence into the annotation, we used EvidenceModeller v1.1.38 to apply a differential weighting schema to multiple evidence sources before merging them into a single genome annotation. EvidenceModeller was applied to the PASSA 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.PRJE506.WBPS9.protein.fa.gz, respectively), as well as 110 curated genes; each protein dataset was mapped to the genome using evidence v2.0.20 (parameters: --model protein2Genome --percent 50 --showtargetGtf) followed by GFF extraction using Exonerate_to_gtf.gff.pl from EvidenceModeller. The weighting was applied as follows: PROTEIN (hc_v1 curated models) = 2; PROTEIN (Ce_v1 curated models) = 2; ABINITIO_PREDICTION (GFF, from Brazer) = 2; TRANSCRIPT (GFF, from PASSA) = 5. The merged output of EvidenceModeller (GFF3) was imported back into PASSA 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 (Supplementary Fig. 6). 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 (GFF3; Supplementary Fig. 6).

The genome annotation, as well as all available evidence used to generate the annotation (including per life-stage RNA-seq, Iso-Seq, full-length high quality & CCS reads, coverage and repeat tracks), was uploaded to the web portal Apollo[94] 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 were subsequently incorporated into Wormbase ParaSite.

Annotation statistics were determined using Genome Annotation Generation (GAG)[95]. Annotation comparison was only made between the previously published V1 draft assembly[25], the final V4 chromosomal assembly presented here, the McMaster[96] and NZ isolate[97] draft assemblies, and the C. elegans reference; no comparable annotation was generated for the intermediate assemblies generated before V4 was finalised, including assembly versions V2 and V3. Annotation comparisons to a curated gene set to determine the sensitivity and specificity of the approaches used were performed using gffcompare v0.10.1 (https://ccb.jhu.edu/software/stringtastic/gffcompare.shtml), and were further analysed for completeness using BUSCO (mode proteins –lineage metazoa odv_089). The repetitive content of the genome was analysed using RepeatModeler (v0.11) followed by RepeatMasker (v4.0.7). Further analysis of LTRs was performed using LTRharvest and LTRAligner[98] from the GenomeTools v1.5.9[27] package.

Transcriptome analyses. Quantitative transcriptome analyses were primarily performed using kallisto v0.43.119, which quantifies transcripts by pseudo-aligning raw sequencing reads to a transcriptome reference. The transcriptome reference was derived from the reference Fasta and annotation GFF data using gffread (https://github.com/gpirt/eftread.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[100]. 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 leafcutter to analyse differential spliced isoform usage between life-stages, we used leafcutter[94] (see https://github.com/leafcutter/leafcutter/tree/v2.2.0) genome annotation tool to map the high-quality full-length transcripts to the genome using leafcutter_cluster.py (parameters: –min_samples_per_intron=3 –min_coverage=10), which required the clustered introns and an exon file (derived from the annotation GFF3) as input. Preparation of the exon file of differentially spliced introns was performed using prepare_results.R, an annotation GTF file was required as input, which was generated from the annotation GFF3 using gffread. The visualisation was performed using run_leafc chiropr.

Population genetic analyses. To understand how genetic variation was distributed throughout the chromosomes, we collated sequencing data from one of the most complete global analyses[101] 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 phylodency describing genetic connectivity between samples, and ENA references to individual sequencing datasets are available to explore using scripts (see below) to quantify the proportions of reads that span introns and then subsequently quantify differential intron usage across samples. First, raw RNA-seq reads were mapped to the reference genome using STAR[102] (parameters: –twopassMode basic –outSAMstrandField intronMotif –outSAMtype BAM Unsorted) to generate BAM files. Intron junctions (in files) were then extracted using bam files, which were subsequently clustered using leafcutter_cluster.py (parameters: –min_coverage 00000). Differential introns between pairwise life-stages were identified using leafcutter_dR (parameters: –min_samples_intron=3 –min_coverage 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 exon file of differentially spliced introns was performed using prepare_results.R, an annotation GTF file was required as input, which was generated from the annotation GFF3 using gffread. The visualisation was performed using run_leafc chiropr.

To analyse mitochondrial DNA diversity, bedtools was used to select single nucleotide variants with a read depth > 10, homoyzygous 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[11], and from which maximum likelihood phylogenies were generated using iqtree v1.6.5[111] (parameters: –alrt 1000 –bb 1000 –nt 1) using automated ModelFinding[112]. Principal component analysis (PCA) of the mitochondrial DNA variation was performed using the poppy[113] package in R.

Sequencing coverage variation between the autosomes and X chromosome for each BAM file was performed using leafc chiropr tool. 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, ReadPosRankSum < 8.0. Nucleotide diversity and F-statistics within and between populations was determined using vcfstats v0.1.14[114] –window-pi and –weir-fst-pop, respectively, using a window size of 100 bp.
Ethics approval. 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.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Code availability
Code and workflows used to analyse data and reproduce figures are available at https://github.com/stephenrdoyle/hcontortus_genome and Zenodo https://doi.org/10.5281/zenodo.4069270.

Data availability
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 Liang et al. 2013. RNA-seq data is available from the ENA study ID PRJEB1360. The genome assembly has been made available at ENA (assembly accession: GCA_000496855.2) and WormBase ParaSite (https://parasite.wormbase.org/Ehamacontus_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/sd2I/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/Ehamacontus_contortus/prjeb506/Info/Index/17).

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. All relevant data are available from the corresponding authors upon request.

Received: 5 March 2020; Accepted: 14 October 2020; Published online: 09 November 2020

References
1. Nicholls, S. M., Quick, J. C., Tang, S. & Loman, N. J. Ultra-deep, long-read nanopore sequencing of mock microbial community standards. Gigascience 8, giz013 (2019).
2. Stewart, R. D. et al. Compendium of 4,941 ruminant metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. Nat. Biotechnol. 37, 953–961 (2019).
3. Yoshimura, J. et al. Recompleting the genome of parasitic nematode Haemonchus contortus. Nat. Commun. 10, 4811 (2019), however, the most up-to-date version of the annotation can be accessed at WormBase ParaSite (https://parasite.wormbase.org/Ehamacontus_contortus/prjeb506/Info/Index/17). A static version of the genome annotation used in this paper is available at ftp://ftp.sanger.ac.uk/pub/pathogens/sd2I/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/Ehamacontus_contortus/prjeb506/Info/Index/17).

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. All relevant data are available from the corresponding authors upon request.

Received: 5 March 2020; Accepted: 14 October 2020; Published online: 09 November 2020

References
1. Nicholls, S. M., Quick, J. C., Tang, S. & Loman, N. J. Ultra-deep, long-read nanopore sequencing of mock microbial community standards. Gigascience 8, giz013 (2019).
2. Stewart, R. D. et al. Compendium of 4,941 ruminant metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. Nat. Biotechnol. 37, 953–961 (2019).
3. Yoshimura, J. et al. Recompleting the genome of parasitic nematode Haemonchus contortus. Nat. Commun. 10, 4811 (2019), however, the most up-to-date version of the annotation can be accessed at WormBase ParaSite (https://parasite.wormbase.org/Ehamacontus_contortus/prjeb506/Info/Index/17). A static version of the genome annotation used in this paper is available at ftp://ftp.sanger.ac.uk/pub/pathogens/sd2I/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/Ehamacontus_contortus/prjeb506/Info/Index/17).

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. All relevant data are available from the corresponding authors upon request.

Received: 5 March 2020; Accepted: 14 October 2020; Published online: 09 November 2020

References
1. Nicholls, S. M., Quick, J. C., Tang, S. & Loman, N. J. Ultra-deep, long-read nanopore sequencing of mock microbial community standards. Gigascience 8, giz013 (2019).
2. Stewart, R. D. et al. Compendium of 4,941 ruminant metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. Nat. Biotechnol. 37, 953–961 (2019).
3. Yoshimura, J. et al. Recompleting the genome of parasitic nematode Haemonchus contortus. Nat. Commun. 10, 4811 (2019), however, the most up-to-date version of the annotation can be accessed at WormBase ParaSite (https://parasite.wormbase.org/Ehamacontus_contortus/prjeb506/Info/Index/17). A static version of the genome annotation used in this paper is available at ftp://ftp.sanger.ac.uk/pub/pathogens/sd2I/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/Ehamacontus_contortus/prjeb506/Info/Index/17).

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. All relevant data are available from the corresponding authors upon request.
46. Redmond, D. L. & Knox, D. P. Haemochinus contortus SL2 trans-spliced RNA leader sequence. Mol. Biochem. Parasitol. 117, 107–110 (2001).
47. Leng, R. et al. Annotation of two large contiguous regions from the haemochinus contortus genome using RNA-seq and comparative analysis with Caenorhabditis elegans. PLoS ONE 6, e23216 (2011).
48. Huang, P. et al. Identification and analysis of internal promoters in Caenorhabditis elegans operons. Genome Res. 17, 1478–1485 (2007).
49. Bonfield, J. K. & Whitworth, H. Acetylecolline receptors with long read contigs from the Haemochinus contortus genome project. Mol. Biochem. Parasitol. 195, R5 (2015).
50. Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS ONE 9, e112963 (2014).
51. Blumthaler, T. Operon and non-operon gene clusters in the C. elegans genome. WormBook 1–20, https://doi.org/10.1895/wormbook.1.175.1 (2014).
52. Maruyama, R., Endo, S., Sugimoto, A. & Yamamoto, M. Caenorhabditis elegans daf-1 is expressed in proliferating germ cells and directs proper nuclear organization and cytoplastic core formation during oogenesis. Dev. Biol. 277, 142–154 (2005).
53. Gilleard, J. S. & Redman, E. Genetic Diversity and Population Structure of Haemochuns contortus. Haemochuns contortus and Haemonchosis – Past, Present and Future Trends 31–68, https://doi.org/10.1016/bs.apar.2016.02.009 (2016).
54. Wilson Sayres, M. A. Genetic diversity on the sex chromosomes. Genome Biol. Evol. 10, 1064–1078 (2018).
55. Doyle, S. R. et al. Population genomic and evolutionary modelling analyses reveal a major QTL for invermectin drug resistance in the pathogenic nematode Haemonchus contortus. BMC Genomics 20, 218 (2019).
56. Jain, M. et al. Nanopore sequencing and assembly of a human genome with single-molecule real-time sequencing reads. Nat. Biotechnol. 36, 338–345 (2018).
57. Mandrioli, M. & Manicardi, G. C. Unlocking holocentric chromosomes: new perspectives from comparative and functional genomics? Curr. Genomics 13, 343–349 (2012).
58. Kingan, S. B. et al. A high-quality de novo genome assembly from a single mosquito using PacBio sequencing. Genes 10, 62 (2019).
59. Kashyap, S. S. et al. Emodepside has sex-dependent immunoblasting effects on adult Brugia malayi due to a differentially spliced binding pocket in the RCK1 region of the SLO-1 K channel. PLoS Pathog. 15, e1008041 (2019).
60. Wolstenholme, A. J. Ion channels and receptor as targets for the control of parasitic nematodes. Int. J. Parasitol. 41, 2–13 (2011).
61. Schwarz, E. M. et al. The genome and transcriptome of the zoonotic hookworm Ancylostoma ceylanicum identify infection-specific gene families. Nat. Genet. 47, 416–422 (2015).
62. Tang, Y. T. et al. Genome of the human hookworm Necator americanus. Nat. Genet. 46, 261–269 (2014).
63. Wang, Z. et al. Characterizing Ancylostoma caninum transcriptome and expressed gene clusters from gene expression data. Int. J. Parasitol. 18, 243–257 (2018).
64. Choi, Y.-J. et al. A deep sequencing approach to comparatively analyze the transcriptome of lifecycle stages of the parasitic nematode Haemonchus contortus. Mol. Biochem. Parasitol. 195, R47 (2015).
65. Kim, D. H., Grün, D. & van Oudenaarden, A. Dampening of expression oscillations by synchronous regulation of a microRNA and its target. PLoS Biol. 13, e1002154 (2015).
66. Neveu, C. et al. Genetic diversity of levamisole receptor subunits in parasitic nematode species and abbreviated transcripts associated with resistance. Exp. Parasitol. 162, 102–112 (2016).
67. Niciura, S. C. M. et al. Extreme-QTL mapping of monepantel resistance in Haemonchus contortus. PLoS Negl. Trop. Dis. 6, e1000380 (2012).
68. Walker, B. J. et al. Identification and analysis of internal promoters in Caenorhabditis elegans operons. Genome Res. 17, 1478–1485 (2007).
69. Bonfield, J. K. & Whitworth, H. Acetylecolline receptors with long read contigs from the Haemochinus contortus genome project. Mol. Biochem. Parasitol. 195, R5 (2015).
70. Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS ONE 9, e112963 (2014).
71. Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 23, 1061–1067 (2007).
72. Haemonchus contortus. Genomes 2, 1409 (2009).
73. Kim, D. H., Grün, D. & van Oudenaarden, A. Dampening of expression oscillations by synchronous regulation of a microRNA and its target. PLoS Biol. 13, e1002154 (2015).
74. Dunn, N. A. et al. Apollo: democratizing genome annotation. PLoS Comput. Biol. 15, e1006790 (2019).
75. Geil, S. M. et al. Genome annotation generator: a simple tool for generating and correcting WGS annotation tables for NCBI submission. Gigascience 7, 1–5 (2018).
76. Ellingham, D. et al. Haemonchus contortus SL2 trans-spliced RNA leader sequence. Mol. Biochem. Parasitol. 117, 107–110 (2001).
77. Redmond, D. L. & Knox, D. P. Haemochinus contortus SL2 trans-spliced RNA leader sequence. Mol. Biochem. Parasitol. 117, 107–110 (2001).
78. Gardner, M. B. et al. Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. Science 330, 1775–1778 (2010).
79. Kim, D. H., Grün, D. & van Oudenaarden, A. Dampening of expression oscillations by synchronous regulation of a microRNA and its target. Nat. Genet. 45, 1357–1344 (2013).
80. Saéz, J.-P. et al. Effect of single-dose invermectin on Onchocerca volvulus: a systematic review and meta-analysis. Lancet Infect. Dis. 8, 310–322 (2008).
81. Sabah, A. A., Fletcher, C., Webb, G. & Doenhoff, M. J. Schistosoma mansoni: chemotherapy of infections of different ages. Exp. Parasitol. 61, 294–303 (1986).
82. Neves, C. et al. Genetic diversity of levamisole receptor subunits in parasitic nematode species and abbreviated transcripts associated with resistance. Pharmacogenet. Genomics 20, 414–420 (2010).
83. Fauvin, A. et al. cDNA-AFLP analysis in levamisole-resistant Haemonchus contortus reveals alternative splicing in a nicotinic acetylcholine receptor subunit. Mol. Biochem. Parasitol. 170, 105–110 (2010).
84. Ghosh, M. 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. 49, 847–858 (2019).
85. Nicuira, S. C. M. et al. Extreme-QTL mapping of monepantel resistance in Haemonchus contortus. Vet. Parasitol. 265, 103621 (2019).
86. Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SPSPACE. Bioinformatics 27, 578–579 (2011).
87. Tsai, I. J., Otto, T. D. & Berriman, M. Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. Genome Biol. 11, R41 (2010).
88. Hunt, M. et al. REAP: a universal tool for genome assembly evaluation. Genome Biol. 14, R47 (2013).
89. Bonfield, J. K. & Whitworth, H. Gap5—editing the billion fragment sequence assembly. Bioinformatics 26, 1699–1703 (2010).
Author contributions
J.A.C., I.S.G., M.B., E.D. and S.R.D designed the study. S.R.D. led the analysis and interpretation of the data, A.T., A.M., H.B., K.B., R.B. and J.D.N., analysed data. W.B., J.A.C., J.S.G., M.B., E.D. and S.R.D designed the project. We also thank Pathogen Informatics and DNA Pipelines (WSI) for their support and expertise, and the Biosciences Division at the Moredun Research Institute for expert care and assistance with animals. Work described here was supported by Wellcome and coordinated samples and sequencing. S.R.D. and J.A.C. wrote the manuscript, with contribution and review of the final manuscript by all authors.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s42003-020-01377-3.

Correspondence and requests for materials should be addressed to S.R.D. or J.A.C.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher’s note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.