International Journal for Parasitology xxx (xxxx) xxx

Contents lists available at ScienceDirect

International Journal for Parasitology

journal homepage: www.elsevier.com/locate/ijpara

The confounding effects of high genetic diversity on the determination and interpretation of differential gene expression analysis in the parasitic nematode *Haemonchus contortus*

Andrew M. Rezansoff [a], Roz Laing [b], Axel Martinelli [c], Susan Stasiuk [a], Elizabeth Redman [a], Dave Bartley [d], Nancy Holroyd [c], Eileen Devaney [b], Neil D. Sargison [e], Stephen Doyle [c], James A. Cotton [c], John S. Gilleard [a,*]

[a] Department of Comparative Biology and Experimental Medicine, Faculty of Veterinary Medicine, University of Calgary, Alberta, Canada
[b] Institute of Biodiversity, Animal Health and Comparative Medicine, College of Medical, Veterinary and Life Sciences, University of Glasgow, Scotland, United Kingdom
[c] Wellcome Sanger Institute, Hinxton, Cambridgeshire CB10 1SA, United Kingdom
[d] Moredun Research Institute, Pentlands Science Park, Bush Loan, Penicuik EH26 0PZ, United Kingdom
[e] University of Edinburgh, Royal (Dick) School of Veterinary Studies, Easter Bush Veterinary Centre, Roslin, Midlothian EH25 9RG, United Kingdom

ARTICLE INFO

Article history:
Received 14 April 2019
Received in revised form 17 May 2019
Accepted 22 May 2019
Available online xxxx

Keywords:
Haemonchus contortus
Transcriptomics
RNAseq
Differential expression
Ivermectin
Anthelmintic resistance

ABSTRACT

Differential expression analysis between parasitic nematode strains is commonly used to implicate candidate genes in anthelmintic resistance or other biological functions. We have tested the hypothesis that the high genetic diversity of an organism such as *Haemonchus contortus* could complicate such analyses. First, we investigated the extent to which sequence polymorphism affects the reliability of differential expression analysis between the genetically divergent *H. contortus* strains MHco3(ISE), MHco4(WRS) and MHco10(CAVR). Using triplicates of 20 adult female worms from each population isolated under parallel experimental conditions, we found that high rates of sequence polymorphism in RNAseq reads were associated with lower efficiency read mapping to gene models under default *TopHat2* parameters, leading to biased estimates of inter-strain differential expression. We then showed it is possible to largely compensate for this bias by optimising the read mapping single nucleotide polymorphism (SNP) allowance and filtering out genes with particularly high single nucleotide polymorphism rates. Once the sequence polymorphism biases were removed, we then assessed the genuine transcriptional diversity between the strains, finding ≥824 differentially expressed genes across all three pairwise strain comparisons. This high level of inter-strain transcriptional diversity not only suggests substantive inter-strain phenotypic variation but also highlights the difficulty in reliably associating differential expression of specific genes with phenotypic differences. To provide a practical example, we analysed two gene families of potential relevance to ivermectin drug resistance; the ABC transporters and the ligand-gated ion channels (LGICs). Over half of genes identified as differentially expressed using default *TopHat2* parameters were shown to be an artifact of sequence polymorphism differences. This work illustrates the need to account for sequence polymorphism in differential expression analysis. It also demonstrates that a large number of genuine transcriptional differences can occur between *H. contortus* strains and these must be considered before associating the differential expression of specific genes with phenotypic differences between strains.

1. Introduction

RNAseq has become the standard approach for the genome-wide analysis and quantification of gene expression across the life sciences (Wang et al., 2009; Conesa et al., 2016). Established sequence aligners used in RNAseq analysis pipelines, such as *TopHat2* and its faster successor *HISAT2*, were developed and their default mapping parameters set, primarily for use on vertebrate species such as humans, mice, and zebrafish, which have relatively low levels of both intra- and inter-population genetic diversity (Wang, 1998; Lindblad-Toh et al., 2000; Guryev et al., 2006; Baruzzo et al., 2017). Further, until relatively recently, applications of RNAseq to non-vertebrate species were largely

* Corresponding author at: Department of Comparative Biology and Experimental Medicine, Faculty of Veterinary Medicine, 3330, Hospital Drive, University of Calgary, Calgary, Alberta T2N 4N1, Canada. Fax: +1 403 210 7882.
E-mail address: jsgillea@ucalgary.ca (J.S. Gilleard).

https://doi.org/10.1016/j.ijpara.2019.05.012
0020-7519/

confined to laboratory strains of model organisms such as *Drosophila melanogaster* and *Caenorhabditis elegans*, which also have relatively low levels of genetic diversity (Andersen et al., 2012; Cingolani et al., 2012). Consequently, most publications make little or no acknowledgement of the potentially confounding effects of sequence polymorphism on the mapping efficiency of RNAseq reads and the calling of differentially expressed genes (Baruzzo et al., 2017). RNAseq analysis pipelines are generally applied to non-model organisms simply using established default parameters, with no consideration given the level and distribution of sequence polymorphism within and between the strains or populations being compared (Edwards et al., 2013; Croken et al., 2014; Fiebig et al., 2015; Papenfort et al., 2015; Antony et al., 2016). However, many taxa show high levels and complex patterns of intra-species genetic diversity (Blumenthal and Davis, 2004; Dey et al., 2013; Romiguier et al., 2014; Redman et al., 2015). This is a concern since standard RNAseq alignment benchmarking studies have shown that the performance of different sequence aligners varies with the genome complexity and levels of sequence polymorphism when using simulated sequence data (Baruzzo et al., 2017). However, no published experimental studies directly examine the effects of sequence polymorphism on differential expression analyses using commonly applied RNAseq analysis pipelines.

A good example of the application of RNAseq analysis to non-model organisms is for the investigation of differential expression of candidate genes potentially involved in anthelmintic drug resistance in parasitic nematodes (Xu et al., 1998; Dicker et al., 2011; El-Abdellati et al., 2011; Williamson et al., 2011; Urdaneta-Marquez et al., 2014). *Haemonchus contortus* is arguably the most established parasitic nematode model used for such studies (Gilleard, 2013). It has a good quality reference genome and has extremely high levels of sequence polymorphism (upwards of 5% single nucleotide polymorphism (SNP) rates), both within and between strains or geographical isolates (Laing et al., 2013; Gilleard and Redman, 2016). Consequently, it is an excellent system in which to study the potentially confounding effects of sequence polymorphism on differential gene expression analysis. In this paper, we use three well characterised laboratory passaged strains of *H. contortus* to examine how differences in coding sequence (CDS) polymorphism rates, with respect to the MHco3 (ISE) genome reference strain, affect read mapping and bias differential expression analysis. We show how these confounding effects can be reduced and demonstrate that, even when the effects of sequence polymorphism are minimised, there are still a large number of differentially expressed genes between these three strains. These results have important implications for the application of RNAseq analysis to many non-model organism species with high levels of genetic diversity.

2. Materials and methods

2.1. Haemonchus contortus strains, sample preparation and sequencing

The MHco3(ISE), MHco4(WRS) and MHco10(CAVR) *H. contortus* strains have been previously characterised and are described in detail elsewhere (Redman et al., 2008, 2012; Laing et al., 2013). The MHco3(ISE) strain is susceptible to all main classes of anthelmintic and has been used as the reference genome strain (Laing et al., 2013). The MHco4(WRS) strain is derived from the White River Strain (WRS) that was isolated as an ivermectin-resistant field isolate from South Africa (Van Wyk and Malan, 1988). The MHco10(CAVR) strain is derived from the Chiswick Avermectin Resistant Strain (CAVR) which was originally isolated as an ivermectin-resistant strain as a laboratory contaminant of a field isolate from Australia (Le Jambre et al., 1995).

Three sets of 20 adult female worms were recovered on necropsy at 28 days post experimental infection from the abomasa of three different individual sheep for each *H. contortus* strain; MHco3(ISE), MHco4(WRS) and MHco10(CAVR). Each set of 20 adult females served as one of three biological replicates for RNA-seq analysis for each strain. Adult worms recovered from the abomasum were rinsed and sexed in physiological saline at 37 °C and then immediately snap frozen before total RNA was isolated from each pool of 20 worms using a standard Trizol protocol as described in Laing et al. (2011). RNA samples were assessed on a Bioanalyser 2100 (Agilent) and Illumina transcriptome libraries were prepared as previously described (Laing et al., 2011). Sequencing of transcriptome libraries was performed on an Illumina HiSeq platform to generate 100 bp paired-end reads.

2.2. Sequence quality control and read mapping

Raw 100 bp reads were inspected using *FastQC* (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) for overall dataset integrity and all reads were trimmed at the 5′ end by 10 bases. Fifteen bases were also trimmed from the 3′ ends of all reads to remove low quality sequence characteristic of 3′ tail ends. The post-trimmed 75 bp reads were used for mapping to the *H. contortus* MHco3(ISE) reference genome assembly (Laing et al., 2013) with *TopHat2* (Kim et al., 2013). The assembly used is an improved version (N50 of 5.24 MB) of the original published *H. contortus* genome assembly (GenBank ID PRJEB506 – N50 of 83.29 kb (Laing et al., 2013)) and contains an expanded set of annotated gene models (https://data.mendeley.com/datasets/4z6xv5j5zf/1). Numerical identifiers of these additional gene models begin with HCOI_0500, and have not yet been submitted to online genomic resources (e.g. Uniprot.org).

TopHat2 was executed using the following parameter settings: *TopHat2 -N (#) --read-gap-length (%) --read-edit-dist (# + %) -I 40,000 -r 200 -a 6 -g 1 --no-discordant --no-mixed* --min-*intron 10 --microexon-search --mate-std-dev 50 --library-type fr-unstranded ./reference.fasta trimmed_forward_reads.fastq trimmed_reverse_reads.fastq*. Only *-N* (specifying the number of SNPs per mapped read allowed by *TopHat2*), *--read-gap-length* (the allowed base count of any indels), and *--read-edit-dist* (the allowed combined base count of both *-N* and *--read-gap-length*) were adjusted throughout the experiment. Reads of all triplicates of all three populations were initially mapped with *TopHat2* using a scale of SNP (polymorphism) allowances from 2 to 10 SNPs (*-N*) per read with indel allowance (*--read-gap-length*) held constant at three bases.

Three different allowances for polymorphism were then subsequently chosen for further analysis: low, the *TopHat2* default allowances (denoted N2 – allowing two SNPs or two indels per read), moderate (denoted N5 – allowing five SNPs and three indels per read), and high (denoted N10 – allowing 10 SNPs and six indels per read) allowances for polymorphism, respectively. Varying the indel allowances had very little effect on the percentage of reads mapping to the reference genome (data not shown). *Samtools' flagstat* tool (Li et al., 2009) was used to determine the proportion of reads mapped at each allowance for each strain.

2.3. RNAseq processing and analysis

Reads mapped to each gene model were sorted with *samtools sort*, and counted for each of the three bioreplicates for each strain at the three different SNP allowances – N2, N5, N10 – using the following command in *HTseq-count*: *htseq-count -i parent -q -s no -f bam -t cds ./sorted_accepted_hits.bam ./genome_annotation_file.gff3* (Anders et al., 2014). Raw mapped read counts for each gene model

Please cite this article as: A. M. Rezansoff, R. Laing, A. Martinelli 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*, International Journal for Parasitology, https://doi.org/10.1016/j.ijpara.2019.05.012

of each bioreplicate of each strain were compiled and used as input for *DESeq2*.

DESeq2 (Love et al., 2014) was run in *Rstudio* (2015) to identify differential expression between the three strains, at different polymorphism allowances, based on gene model read counts. The *plotPCA* tool in *DESeq2* was used to plot segregation of triplicates based on gene expression of the top 15,000 expressed low polymorphic genes (LPGs) at the moderate N5 allowance. *DESeq2* result tables were exported and manipulated in *Microsoft Excel*. Genes were only called as differentially expressed in this analysis if they (i) showed a greater than 2 fold-change difference in expression between the strains compared, and (ii) yielded adjusted *P* values of less than 0.05.

2.4. Categorising gene models on the basis of SNP rates and SNP rate differences between strains

SNPs within CDS were called using *samtools mpileup* on whole genome sequence (WGS) datasets created for each of the strains against the MHco3(ISE) genome assembly (Doyle et al., 2019). SNPs present at >40% frequency were totaled per gene model for each of the strains. The SNP rate was calculated for each gene in each strain by dividing the total number of SNPs in the gene by the respective gene model CDS length. The genes were then categorised in two different ways for subsequent investigation of the effect of sequence polymorphism on read mapping and RNAseq analysis. First, they were categorised based on their SNP rates in each strain: categories 0%, 0–0.5%, 0.5–1%, 1–2%, 2–5%, and >5%. Second, they were categorised based on the difference in SNP rates for each of the three pairwise strain comparisons (i.e. the SNP rate observed in one strain subtracted from the SNP rate observed in the other) categories >5–15%, >2–5%, >0–2%, 0%. Genes with a >15% difference were not categorised as they were likely to be due to annotation errors and/or overly short CDS lengths.

2.5. Assessment of genuine transcriptomic variation between the strains

Differential expression statistics were called with *DESeq2* for each of the three pairwise strain comparisons at each of the three map allowances. In each pairwise strain comparison at the N5 allowance, genes showing low SNP rate differences (less than 2%) were denoted as LPGs. The number of LPGs up- and down-regulated in each strain comparison at the N5 allowance, and shared up- or down-regulated in two strains versus the third strain, were totaled at both a log2 1× and log2 2× fold-change expression threshold. Candidate anthelmintic resistance gene families, as defined by the published *H. contortus* genome annotation (Laing et al., 2013), were specifically highlighted in that their differential expression was compared at the N2 allowance, the N5 allowance, and the N5 allowance with high polymorphic genes removed.

3. Results

3.1. Coding sequence polymorphism affects RNAseq read mapping against the MHco3(ISE) reference assembly for the three different H. contortus strains

The total combined read counts of the triplicate RNAseq datasets were similar among the three strains at 36,175,121, 36,025,170, and 37,584,775 reads for MHco3(ISE), MHco4(WRS), and MHco10(CAVR), respectively. We determined the total number of CDS SNPs present at >40% frequency, relative to the MHco3(ISE) reference genome assembly, using whole genome sequence datasets independently created for each strain. A total of 701,715, 1,121,242 and 1,143,102 CDS SNPs, representing rates of 2.97%, 4.74% and 4.84% of the 23.63 MB *H. contortus* reference CDS annotation, were present for MHco3(ISE), MHco4(WRS), and MHco10 (CAVR), respectively.

The percentage of RNAseq reads that mapped to the MHco3(ISE) reference genome assembly, using the default SNP allowance (N2 – allowing two SNPs or two indels per read) in *TopHat2*, was 60.7%, 44.8% and 47.1% for the MHco3(ISE), MHco4(WRS) and MHco10 (CAVR) strains, respectively (Fig. 1). Increasing the *TopHat2* SNP allowance parameter changed the percentage of RNAseq reads that mapped (Fig. 1). For the MHco3(ISE) strain, the percentage of RNAseq reads mapping to the reference genome increased as the polymorphism allowance was increased from N2 to N5 (allowing five SNPs and three indels per read) and then decreased as the allowance was further increased to N10 (allowing 10 SNPs and six indels per read) (Fig. 1). This pattern was very similar for the MHco4 (WRS) and MHco10(CAVR) strains but the maximum percentage of reads mapping occurred at the N6 allowance, albeit at rates only 0.1% greater than at N5 (Fig. 1). The percentage of RNAseq reads that mapped to the reference MHco3(ISE) genome assembly was greater for the MHco3(ISE) strain than for the other two strains at all polymorphism allowances, although the magnitude of this difference decreased from the N2 to N10 allowance (Fig. 1).

A more detailed analysis was undertaken for the N2, N5 and N10 polymorphism allowances at the level of gene models. Increasing the polymorphism allowance from N2 to N5 resulted in 12,778, 11,101, and 11,324 gene models having a >1% increase in the number of mapped RNAseq reads for MHco3(ISE), MHco4 (WRS), and MHco10(CAVR), respectively (Fig. 2Aa). In contrast, 591, 1316, and 1563 genes showed a >1% decrease in RNAseq reads mapped (Fig. 2Aa). Further increasing the mapping allowance from N5 to N10 had the opposite effect, with a greater number of gene models having a decreased rather than an increased number of RNAseq reads mapped: a change in the polymorphism allowance from N5 to N10 resulted in 12,529, 8139, and 8470 gene models having a >1% decreased number of RNAseq reads mapped, compared with 1092, 4682 and 4953 genes having an increased number of RNAseq reads mapped for MHco3(ISE), MHco4(WRS), and MHco10(CAVR) strains, respectively (Fig. 2Ab).

3.2. The SNP allowance has a greater effect on RNAseq read mapping for gene models with higher levels of sequence polymorphism

There were large differences in the SNP rates of different gene models, relative to the MHco3(ISE) reference genome, ranging from those with SNP rates of 0% to those above 5%. The 25,111 gene models were binned into several different SNP rate categories to investigate how the mapping of RNAseq reads to the reference MHco3(ISE) genome assembly was affected by the coding region SNP rate (Fig. 2B). The MHco4(WRS) and MHco10(CAVR) strains had a significantly greater proportion of gene models with SNP rates greater than 0.5% (18,910 (75.3%) and 18,886 (75.2%), respectively) compared with the MHco3(ISE) strain (11,303 (45.0%)] (Z-stat = 69.3 ($P < 0.000$) and 69.1 ($P < 0.000$), respectively) (Fig. 2B).

The effect of changing the polymorphism allowance from N2 to N5 on RNAseq read mapping for each of the different SNP rate categories of gene models was examined for each strain (Fig. 2Ca; Supplementary Table S1). The ratio of RNAseq reads mapping to gene models at the N5 compared with the N2 allowance was >1 for all SNP rate categories above 0% for all three strains (Fig. 2Ca). Furthermore, this ratio increased as the SNP rate increased. In contrast, the ratio of RNAseq reads mapping to gene models at the N10 allowance compared with the N5 allowance was <1 except for gene models with a polymorphism frequency of > 5% for strains MHco4(WRS) and MHco10(WRS) (Fig. 2Cb).

Please cite this article as: A. M. Rezansoff, R. Laing, A. Martinelli 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*, International Journal for Parasitology, https://doi.org/10.1016/j.ijpara.2019.05.012

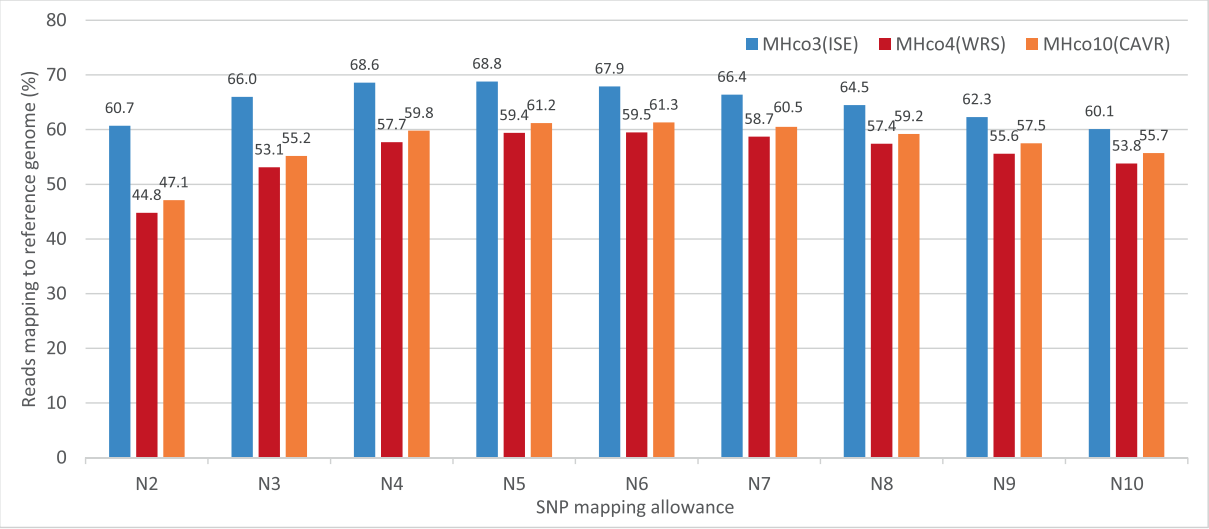

Fig. 1. The percentage of RNAseq reads that mapped to the *Haemonchus contortus* MHco3(ISE) strain reference genome assembly at different *TopHat2* single nucleotide polymorphism (SNP) allowances (N2 to N10) shown for each of the three *H. contortus* strains MHco3(ISE), MHco4(WRS), and MHco10(CAVR).

3.3. High levels of sequence polymorphism artificially inflate between-strain RNAseq differential expression results

We next investigated the influence of CDS polymorphism on the RNAseq differential expression reported by *DESeq2* between pairwise strain comparisons. We hypothesised that gene models with large differences in SNP rates (SNPs/bp) between two strains are more likely to be reported as differentially expressed between those strains than gene models with smaller SNP rate differences. To test this hypothesis, for each gene model we first determined the difference in SNP rates (SNPs/bp) between each pairwise comparison of the three strains. We then plotted the difference in the SNP rate between the two strains against the log2-fold difference in expression called by *DESeq2* for each gene model (Fig. 3). Using the MHco4(WRS) and MHc03(ISE) pairwise comparison as an example, for those gene models with a higher SNP rate in MHco4 (WRS) than in MHco3(ISE), a greater number was reported by *DESeq2* as down-regulated in MHco4(WRS) relative to MHco3 (ISE) than as up-regulated (Fig. 3A). This bias towards down-regulation increased as the SNP rate difference of gene models between the two strains increased (Fig. 3A). For gene models with a lower SNP rate in MHco4(WRS) than in MHco3(ISE), the opposite trend was apparent (Fig. 3B). Similar patterns were observed in both the MHco3(ISE) versus MHco10(CAVR) and MHco4(WRS) versus MHco10(CAVR) pairwise comparisons (Fig. 3C–F).

To further quantify how SNP rate differences between the strains biases reporting of differential expression, we placed each of the 25,049 gene models with SNP rate data into one of seven "SNP rate difference" categories for each pairwise strain comparison (data for the MHco3(ISE) versus MHco4(WRS) pairwise comparison is shown in Fig. 4, and Supplementary Tables S1, S2). The percentage of gene models reported as differentially expressed (with adjusted *P* values <0.05 and >log2 1× fold-change in expression) was lowest for the 0% SNP rate difference category and increased as the SNP rate difference category increased (Fig. 4A). This trend was seen at all three SNP mapping allowances (Fig. 4A). There was also a strong relationship between the directionality of the differential expression called by *DESeq2* and the directionality of the SNP rate difference between the strains. For SNP rate difference categories where the SNP rate was greater in MHco4(WRS) than in MHco3(ISE) by at least 2%, the large majority of gene models reported as differentially expressed were down-regulated in MHco4(WRS) relative to MHco3(ISE) (396/425 (93.2%)) (Supplementary Table S2). Conversely, the large majority of gene models with SNP rates at least 2% lower in MHco4(WRS) than in MHco3(ISE), were up-regulated in MHco4(WRS) relative to MHco3(ISE) (21/27 (77.8%)) (Supplementary Table S2).

3.4. Minimising the effect of sequence polymorphism differences on differential expression analysis in pairwise strain comparisons

We next investigated ways to minimise the effect of sequence polymorphism on global transcriptomic differential expression analysis in pairwise strain comparisons. We first examined the effect of changing the read mapping polymorphism allowance on the number and bias of the differentially expressed genes reported by *DESeq2* in pairwise strain comparisons. When the polymorphism allowance was changed from N2 to N5 or from N5 to N10, there was an overall decrease in the total number of differentially expressed genes reported in all three pairwise strain comparisons (Supplementary Table S4). This trend was generally observed for genes in all SNP rate difference categories (see example of MHco3(ISE) versus MHco4(WRS) pairwise comparison in Fig. 4A). At the default N2 polymorphism allowance, *DESeq2* reported more genes down-regulated than up-regulated in both MHco4(WRS) and MHco10(CAVR) when each was compared with MHco3(ISE) (Supplementary Fig. S1; Supplementary Table S4). This bias was reduced as the mapping allowance was increased to N5 and then N10 (Supplementary Fig. S1; Supplementary Table S4). In contrast, the MHco4(WRS) and MHco10(CAVR) pairwise comparison showed a relatively equal ratio of down-regulated and up-regulated gene numbers even at the default N2 polymorphism allowance (Supplementary Fig. S1; Supplementary Table S2).

We then calculated the net (overall mean) differential expression (NDE) of all gene models in each of the seven "SNP rate difference" categories for each of the pairwise strain comparisons to see if there was an overall directional bias to the data (data for the MHco4(WRS) and MHco3(ISE) pairwise strain comparison is shown in Fig. 4B). The NDE in the direction MHco4(WRS) > MHco3(ISE) was greatest for those gene models in the 5–15% MHco4(WRS) > MHco3(ISE) SNP rate difference category and least for gene models in the 0% SNP rate difference category (Fig. 4B, Supplementary Table S1). Conversely, the NDE in the direction MHco4(WRS) < MHco3(ISE) was highest for gene models in the

Please cite this article as: A. M. Rezansoff, R. Laing, A. Martinelli 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*, International Journal for Parasitology, https://doi.org/10.1016/j.ijpara.2019.05.012

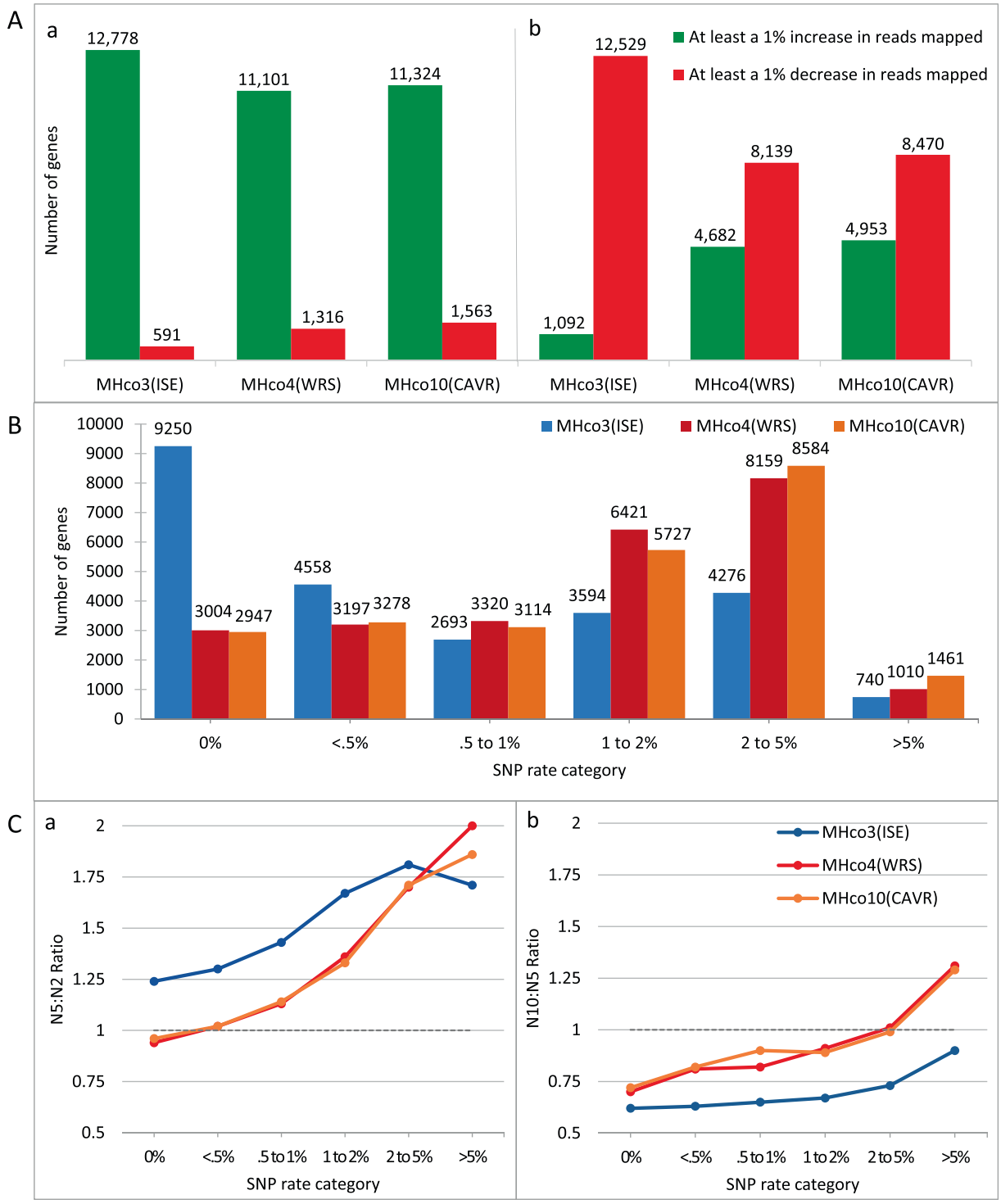

Fig. 2. Changes in single nucleotide polymorphism allowance (N2, N5, N10) when read mapping with *TopHat2* affects the mapping efficiency of reads from each *Haemonchus contortus* strain differently due to variation in the strains' rates of polymorphism. (A) The number of genes which had either a >1% increase (green bars) or >1% decrease (red bars) in the number of RNAseq reads mapping to them on the reference MHco3(ISE) strain genome assembly following an increase in the read mapping polymorphism allowance in *TopHat2* for *H. contortus* strains MHco3(ISE), MHco4(WRS), and MHco10(CAVR). (a) The data for a change in polymorphism allowance from N2 to N5 and (b) the data for a change from N5 to N10 are shown. (B) The number of gene models in each SNP rate category for each *H. contortus* strain. The SNP rate for each gene model was calculated by dividing the number of SNPs in each genes coding sequence by the total coding sequence length of that gene model. (C) Ratios of the total number of RNAseq reads mapping to gene models in each single nucleotide polymorphism rate category at two different single nucleotide polymorphism mapping allowances for each *H. contortus* strain. (a) the N5:N2 ratio and (b) the N10:N5 ratio are shown. Counts of reads mapped were totaled for all genes within each SNP rate category of each strain (colour coded).

5–15% MHco4(WRS) < MHco3(ISE) SNP rate difference category and least for the 0% SNP rate difference category (Fig. 4B, Supplementary Table S2). The NDE of gene models between strains was highest at the N2 polymorphism mapping allowance, and least for the N10 polymorphism mapping allowance, in all SNP rate difference categories (Fig. 4B; Supplementary Table S2).

Please cite this article as: A. M. Rezansoff, R. Laing, A. Martinelli 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*, International Journal for Parasitology, https://doi.org/10.1016/j.ijpara.2019.05.012

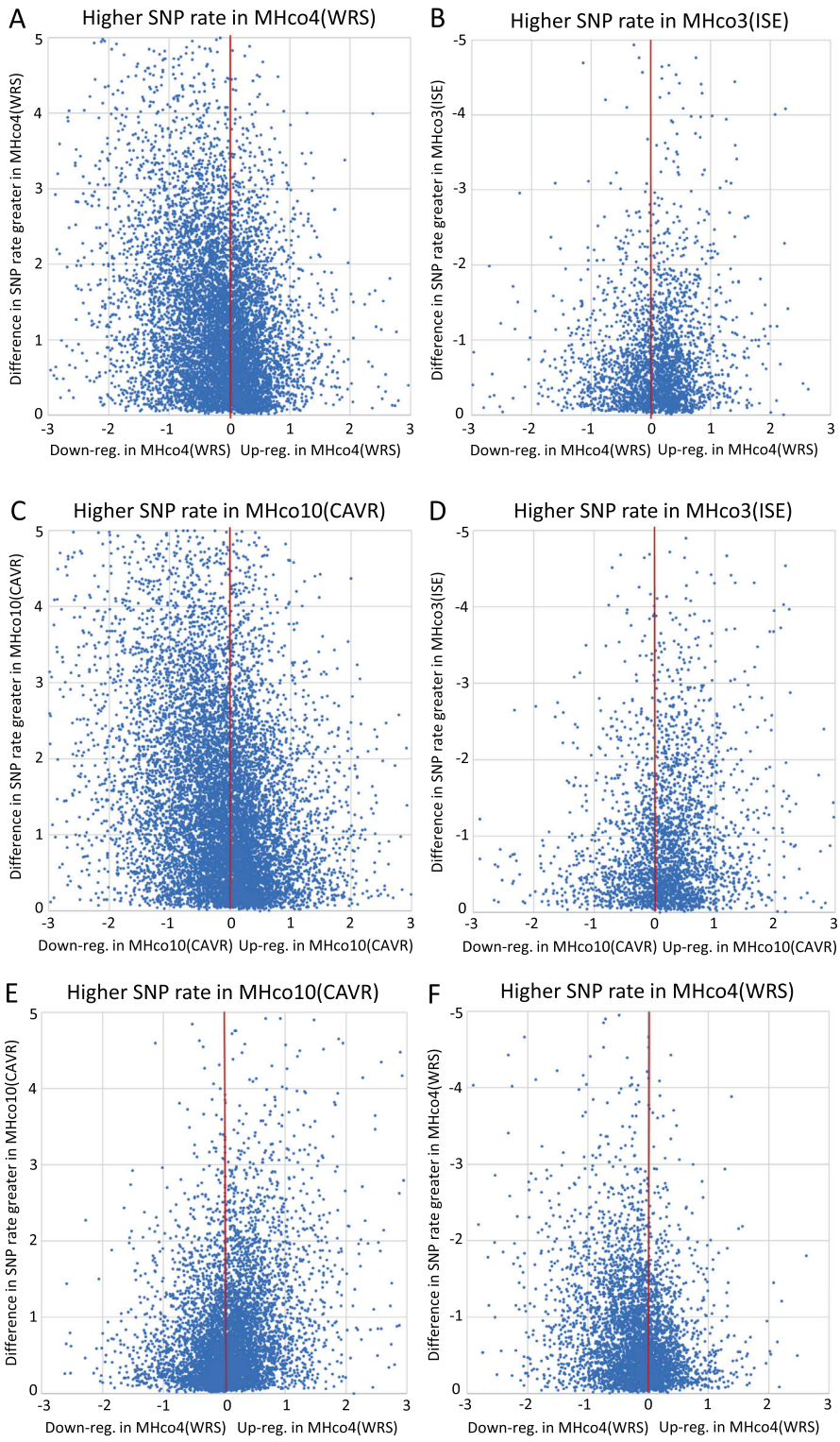

Fig. 3. Scatter plots of the differential expression of gene models, as determined by *DESeq2* (*X*-axis), plotted against their difference in single nucleotide polymorphism rate percentage between the two strains being compared (*Y*-axis). Gene model data points in each pairwise comparison are split into left and right panels with the left panels (A, C, and E) showing the gene models with higher single nucleotide polymorphism rates in one strain of each pairwise comparison and respective right panels (B, D, and F) showing the gene models with higher single nucleotide polymorphism rates in the other pairwise strain. A and B show the MHco4(WRS) versus MHco3(ISE) comparison, C and D show the MHco10(CAVR) versus MHco3(ISE) comparison, and E and F show the MHco4(WRS) versus MHco10(CAVR) comparison. The difference in the single nucleotide polymorphism rate percentage between the two strains is shown on the *Y*-axis and plotted against reported log2 fold-change differential expression for each gene. The red lines represent zero differential expression.

The NDE of gene models between the strains was relatively close to zero for genes of the three lowest SNP rate difference categories, particularly at the N5 and N10 polymorphism allowances (Fig. 4B; Supplementary Table S3). This suggests that gene models with <2% difference in SNP rate between strains had a minimal bias in pairwise strain differential expression analyses. We defined

Please cite this article as: A. M. Rezansoff, R. Laing, A. Martinelli 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*, International Journal for Parasitology, https://doi.org/10.1016/j.ijpara.2019.05.012

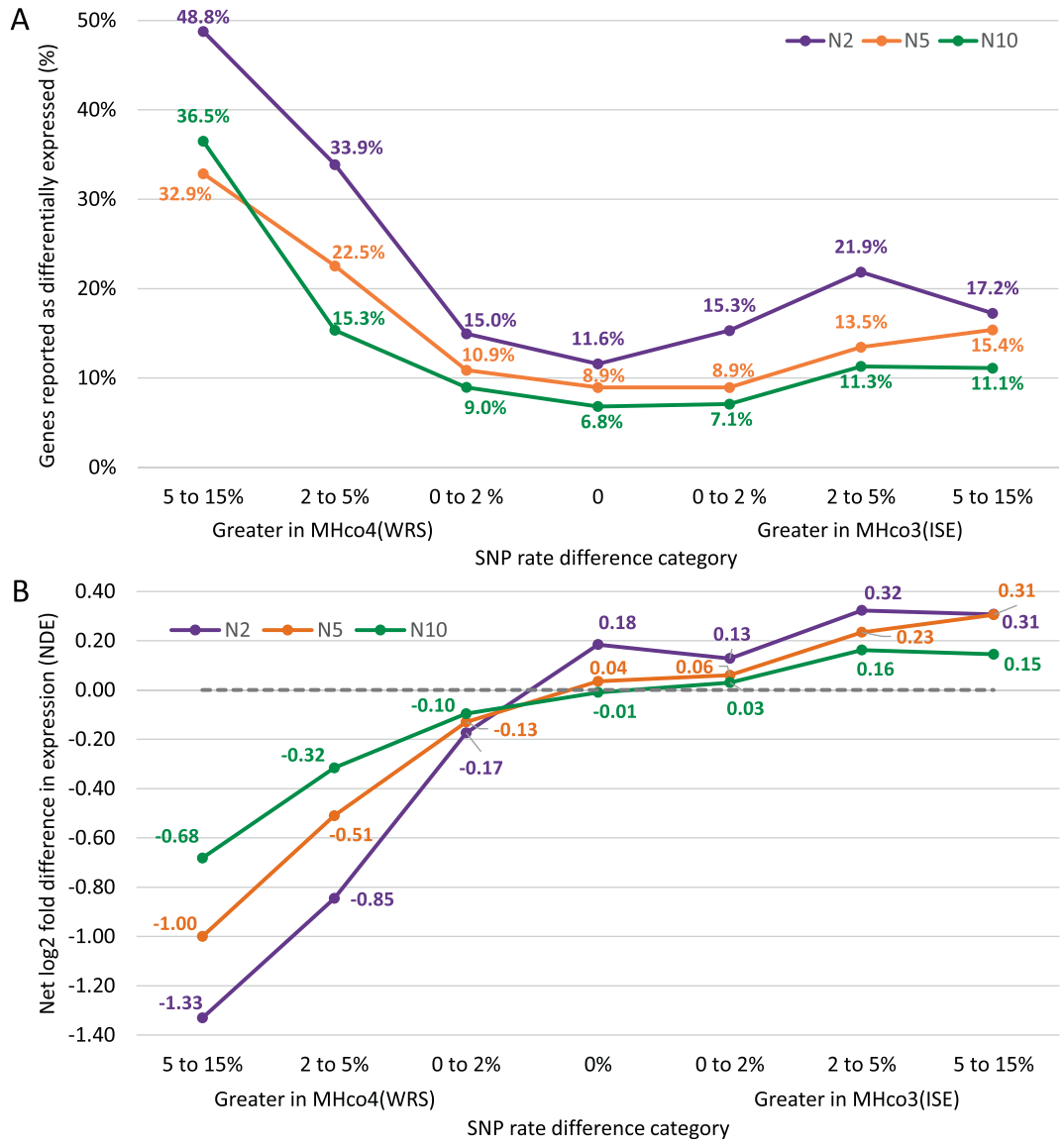

Fig. 4. Gene models with greater single nucleotide polymorphism rate differences between two *Haemonchus contortus* strains show increased rates of differential expression biased towards down-regulation in the strain with higher single nucleotide polymorphism rates. (A) The percentage of expressed gene models in each single nucleotide polymorphism rate difference category that are differentially expressed between the MHco3(ISE) and MHco4(WRS) strains (log2 fold-change >1×; adjusted *P* value <0.05) for each of the three SNP (polymorphism) allowances – N2, N5, and N10 – when mapping. (B) The net log2 fold differences in expression (NDE) of all expressed genes in each SNP rate difference category. NDEs are shown for the N2, N5, and N10 SNP allowances when read mapping for the MHco3(ISE) versus MHco4(WRS) pairwise comparison. NDEs are the mean values for all genes in each single nucleotide polymorphism rate difference category. Negative NDE values indicate an overall bias towards down-regulation of genes in the MHco4(WRS) versus MHco3(ISE) strain. Positive values report an overall bias towards up-regulation of genes.

these gene models as "LPG models" in the subsequent differential expression analysis. These represent 17,881 out of the total of 25,111 gene models in the *H. contortus* whole genome annotation (71.2%) and so represent the majority of gene models (Supplementary Fig. S2).

3.5. Investigating genuine transcriptional differences between H. contortus strains

We restricted the global transcriptomic analysis to the LPG models, as defined in Section 3.4, and used an N5 polymorphism allowance for read mapping to minimise the confounding effect of inter-strain sequence polymorphism. This resulted in the inclusion of 20,781, 19,397, and 22,924 gene models for the MHco4 (WRS) versus MHco3(ISE), MHco10(CAVR) versus MHco3(ISE), and MHco4(WRS) versus MHco10(CAVR) pairwise strain comparisons, respectively (Supplementary Fig. S2). A set of 17,881 genes was common to the analysis set for all three pairwise comparisons (Supplementary Fig. S2). Normalised global expression of each of the nine bioreplicate RNAseq datasets clustered by strain on principal component analysis (PCA) demonstrating that there are transcriptomic differences between the strains, even after the effects of sequence polymorphism on RNAseq mapping are minimised (Supplementary Fig. S3).

A total of 1125 (5.41% of LPGs), 1498 (7.72% of LPGs), and 824 (3.59% of LPGs) genes were differentially expressed at >1× log2 fold in the MHco4(WRS) versus MHco3(ISE), MHco10(CAVR) versus MHco3(ISE), and MHco4(WRS) versus MHco10(CAVR) pairwise comparisons, respectively (Fig. 5). Of these, 134 genes (41 up-regulated, 93 down-regulated), 259 genes (121 up-regulated, 138 down regulated), and 103 genes (40 up-regulated, 63 down regulated) were >2× log2 fold differentially expressed, respectively

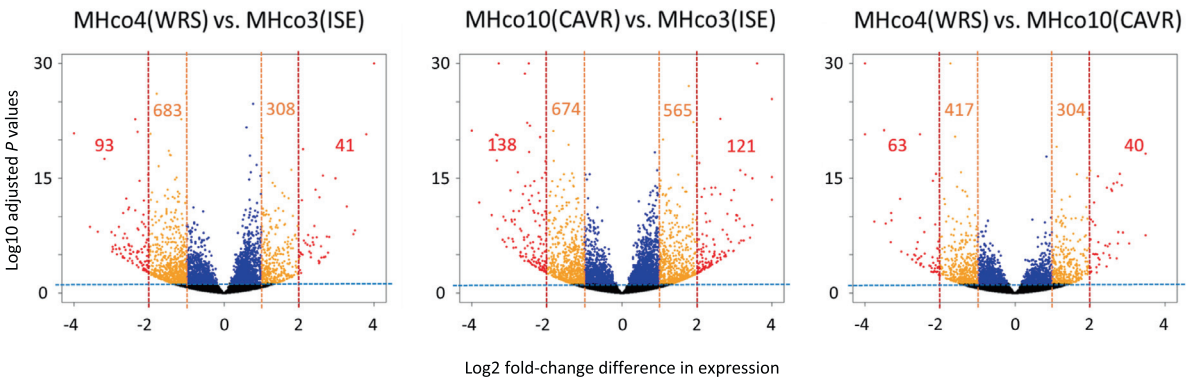

Fig. 5. The total number of differentially expressed low polymorphic genes observed in each pairwise strain comparison at the N5 mapping allowance. Gene counts at both >1× log2 fold-change (orange dots), and >2× log2 fold-change (red dots) thresholds are shown. The blue line on the *Y*-axis represents an adjusted *P* value of 0.05.

(Fig. 5). The large majority of the most differentially expressed genes in all strains comparisons were either undescribed or had only broad ontological classifications (Supplementary Table S5). No previously reported ivermectin resistance candidate LPGs were observed to be differentially expressed in at >2× log2 fold-change expression in either of the two ivermectin resistance strains relative to the MHco3(ISE)-susceptible strain (Supplementary Table S5).

We examined the number of genes that were differentially expressed in more than one of the pairwise strain comparisons to see if a set of genes was common to different pairwise comparisons. The highest proportion of shared differentially expressed LPGs was between the MHco4(WRS) versus MHco3(ISE) and MHco10(CAVR) versus MHco3(ISE) pairwise strain comparisons (Supplementary Fig. S4). Of the 2132 gene models differentially expressed between either MHco4(WRS) and MHco10(CAVR) versus MHco3(ISE), 491 (23.03%) were differentially expressed with the same directionality (up- or down- regulated) in both pairwise comparisons at >1× log2 fold change (48 gene models at >2× log2 fold change) (Supplementary Fig. S4A). Fewer genes were shared in the other two strain combinations. Of the 2025 gene models differentially expressed between either MHco3(ISE) and MHco4(WRS) strains versus MHco10(CAVR), 297 (14.67%) gene models were differentially expressed with the same directionality at >1 log2-fold change (39 gene models at >2 log2-fold change) in both pairwise comparisons (Supplementary Fig. S4B). Of the 1794 gene models differentially expressed between either MHco3(ISE) and MHco10 (CAVR) versus MHco4(WRS), only 155 (8.64%) gene models were differentially expressed at >1 log2-fold change (eight gene models at >2 log2 fold change) with the same directionality in both comparisons (Supplementary Fig. S4C). Both these percentages represent a significantly lower proportion of differentially expressed genes shared than were observed shared in MHco4(WRS) and MHco10(CAVR) versus MHco3(ISE) (Z-stats = 6.8 ($P < 0.000$), and 12.1 ($P < 0.000$), respectively).

3.6. Investigating the effect of sequence polymorphism on differential expression analysis of two gene families of relevance to ivermectin resistance research

Sixty-seven ligand-gated chloride channels (LGICs) and 86 ABC transporters identified in the published *H. contortus* draft genome (Laing et al, 2013) were examined for differential expression between the MHco4(WRS) and MHco10(CAVR) ivermectin-resistant strains and the susceptible MHco3(ISE) strain. Three different differential expression analyses were compared to assess the impact of accounting for sequence polymorphisms differences between the strains; using the default N2 SNP allowance on all 25,111 gene models, using the N5 SNP allowance on all 25,111 genes, and using the N5 SNP allowance on the set of 17,881 LPGs. There was a substantial reduction in the total number of differentially expressed genes reported using the N5 allowance on the LPG gene set compared with the N2 default allowance on the full gene set (Table 1). When comparing the two ivermectin-resistant strains with the ivermectin-susceptible strain, only three of the LPGs – *Hco-lgc-55, Hco-pmp-6,* and *Hco-lgc-44* – showed differential expression at the N5 allowance in both the MHco4(WRS) and MHco10(CAVR) versus MHco3(ISE) pairwise comparisons. *Hco-lgc-55* had >2× log2 fold up-regulation in both cases (Table 1).

4. Discussion

Differential expression analysis, either at the single gene or whole transcriptome level, between parasitic nematode strains and isolates is a common experimental approach. For example, a number of candidate anthelmintic resistance genes have been identified by differential expression analysis of drug-resistant and -susceptible isolates (Xu et al., 1998; Dicker et al., 2011; El-Abdellati et al., 2011; Williamson et al., 2011). In the case of *H. contortus*, we reasoned that the extremely high levels of sequence polymorphism both within and between laboratory strains and field isolates (reviewed in Gilleard and Redman (2016)), might confound the validity of such comparisons when using RNAseq, which is now the central approach to conducting differential gene expression analyses. The majority of researchers use only the default parameters of RNAseq data analysis pipelines and do not explore the effect of different parameters on results reported (Baruzzo et al., 2017). It has been shown, using simulated datasets, that the parameter with the greatest impact on performance is the number of mismatches tolerated while read mapping (Baruzzo et al., 2017). Since this seemed likely to be a particular issue for organisms with high levels of sequence polymorphism, we undertook a detailed analysis to examine the extent to which this may impact RNAseq-based differential expression analysis between *H. contortus* strains, and investigated how it could be mitigated to allow genuine transcriptional differences to be assessed. We used *TopHat2* (Kim et al., 2013) as our read mapping software as this has been the mapping program most commonly used for RNAseq analysis over a number of years and currently has the most citations in RNAseq literature. There are a number of alternative mapping tools available whose use is becoming increasingly common, such as *HISAT2* (Kim et al., 2015), which is the recommended successor of *TopHat2*, but these tools are similarly sensitive to changes in the mismatch parameter (Baruzzo et al., 2017).

Please cite this article as: A. M. Rezansoff, R. Laing, A. Martinelli 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*, International Journal for Parasitology, https://doi.org/10.1016/j.ijpara.2019.05.012

Table 1
Differentially expressed *Haemonchus contortus* genes of ivermectin resistance candidate gene families, the ligand-gated ion channels, and ABC transporters. The respective log2 fold-change differences in expression are shown for both of the ivermectin-resistant strains – MHco4(WRS) and MHco10(CAVR) – relative to the ivermectin-susceptible reference strain – MHco3(ISE). In each panel, all genes shown are differentially expressed at the default (N2) single nucleotide polymorphism allowance (two single nucleotide polymorphisms allowed per read), at the N5 single nucleotide polymorphism allowance, and at the N5 SNP allowance with the highly polymorphic genes removed (i.e. only low-polymorphic genes shown), respectively. Highly polymorphic genes comprise genes with >2% difference in single nucleotide polymorphism rate between the two strains compared.

Ligand-gated ion channels

Up-regulated in MHco4(WRS)						Up-regulated in MHco10(CAVR)					
N2		N5		N5 LPGs		N2		N5		N5 LPGs	
Hco-lgc-55	2.64	*Hco-lgc-55*	2.74	*Hco-lgc-55*	2.74	*Hco-lgc-3*	6.83	*Hco-lgc-3*	5.87	*Hco-lgc-3*	5.87
Hco-acc-2	1.15					*Hco-lgc-55*	2.36	*Hco-lgc-55*	2.33	*Hco-lgc-55*	2.33
Hco-lgc-39	1.11					*Hco-acr-6*	1.73	*Hco-acr-6*	1.53	*Hco-acr-6*	1.53
						Hco-lgc-41	1.58	*Hco-lgc-41*	1.48	*Hco-lgc-41*	1.48
						Hco-acc-1	1.53				
						Hco-mptl-1	1.32	*mtpl-1*	1.32	*Hco-mptl-1*	1.32
						Hco-glc-2	1.06				
Down-regulated in MHco4(WRS)						Down-regulated in MHco10(CAVR)					
Hco-glc-5	−2.56	*Hco-glc-5*	−1.47			*Hco-lgc-7*	−4.48	*Hco-lgc-7*	−4.30		
Hco-lgc-34	−2.41	*Hco-lgc-34*	−1.19	*Hco-lgc-34*	−1.19	*Hco-glc-5*	−3.02	*Hco-glc-5*	−1.88		
Hco-lgc-7	−2.30	*Hco-lgc-7*	−2.55			*Hco-lgc-33*	−2.99	*Hco-lgc-33*	−2.91	*Hco-lgc-33*	−2.91
Hco-acr-17b	−1.78					*Hco-lgc-43*	−2.66	*Hco-lgc-43*	−1.37		
Hco-lgc-43	−1.61					*Hco-lgc-9*	−2.26	*Hco-lgc-9*	−2.15	*Hco-lgc-9*	−2.15
Hco-lgc-44	−1.50	*Hco-lgc-44*	−1.14	*Hco-lgc-44*	−1.14	*Hco-acr-24*	−2.09	*Hco-acr-24*	−2.17	*Hco-acr-24*	−2.17
Hco-acr-24	−1.29	*Hco-acr-24*	−1.44			*Hco-ggr-3*	−2.09	*Hco-ggr-3*	−1.69	*Hco-ggr-3*	−1.69
Hco-des-2	−1.18	*Hco-des-2*	−1.27	*Hco-des-2*	−1.27	*Hco-lgc-44*	−1.50	*Hco-lgc-44*	−1.93	*Hco-lgc-44*	−1.93
Hco-lgc-27	−1.12					*Hco-eat-2X*	−1.80				
Hco-lgc-40	−1.07					*Hco-acr-15*	−1.57				
Hco-lgc-50	−1.06					*Hco-lgc-34*	−1.26				
Hco-acr-8	−1.06	*Hco-acr-8*	−1.30	*Hco-acr-8*	−1.30	*Hco-acc-2*	−1.13				
Hco-acr-7	−1.05					*Hco-acr-11*	−1.13				
Hco-acr-15	−1.02					*Hco-lgc-45*	−1.01				
Hco-lgc-42	−1.01										

ABC transporters

Up-regulated in MHco4(WRS)						Up-regulated in MHco10(CAVR)					
Hco-abt-12	1.65					*Hco-wht-4*	2.68	*Hco-wht-4*	2.07	*Hco-wht-4*	2.07
Hco-abt-10	1.48					*Hco-mrp-3*	2.54	*Hco-mrp-3*	1.71		
Hco-mrp-7	1.25					*Hco-pgp-11*	1.94	*Hco-pgp-11*	1.26	*Hco-pgp-11*	1.26
Hco-abt-11	1.11					*Hco-abt-10*	1.85	*Hco-abt-10*	1.07	*Hco-abt-10*	1.07
Hco-abcf-1	1.02	*Hco-abcf-1*	1.02	*Hco-abcf-1*	1.02	*Hco-wht-2*	1.71	*Hco-wht-2*	1.31	*Hco-wht-2*	1.31
Hco-mrp-4	1.01					*Hco-ced-7*	1.50	*Hco-ced-7*	1.11	*Hco-ced-7*	1.11
						Hco-wht-5.2	1.42	*Hco-wht-5.2*	1.13	*Hco-wht-5.2*	1.13
						Hco-haf-9	1.38				
						Hco-pgp-10	1.28				
						Hco-abt-12	1.28				
						Hco-mrp-4	1.27				
						Hco-wht-5.1	1.14				
						Hco-abt-11	1.10				
Down-regulated in MHco4(WRS)						Down-regulated in MHco10(CAVR)					
Hco-pmp-4	−1.89	*Hco-pmp-4*	−1.03			*Hco-pmp-6*	−2.11	*Hco-pmp-6*	−1.77	*Hco-pmp-6*	−1.77
Hco-pmp-6	−1.80	*Hco-pmp-6*	−2.06	*Hco-pmp-6*	−2.06	*Hco-abcf-1X*	−1.90	*Hco-abcf-1X*	−1.36		
Hco-abt-4	−1.59	*Hco-abt-4*	−1.21			*Hco-abt-2*	−1.76	*Hco-abt-2*	−1.36		
Hco-abcf-1X	−1.53	*Hco-abcf-1X*	−1.23	*Hco-abcf-1X*	−1.23	*Hco-pmp-4*	−1.72	*Hco-pmp-4*	−1.00		
Hco-pgp-9	−1.45	*Hco-pgp-9*	−1.13			*Hco-pgp-16*	−1.57				
Hco-abt-2	−1.40					*Hco-pgp-3*	−1.50				
Hco-pgp-16	−1.33	*Hco-pgp-16*	−1.38			*Hco-pgp-9*	−1.24				
Hco-pgp-11	−1.32	*Hco-pgp-11*	−1.30	*Hco-pgp-11*	−1.30	*Hco-pmp-7*	−1.02				
Hco-pmp-2	−1.29	*Hco-pmp-2*	−1.24	*Hco-pmp-2*	−1.24						
Hco-abt-7	−1.07										

A higher percentage of RNAseq reads mapped to the reference genome assembly for MHco3(ISE) than for the MHco4(WRS) and MHco10(CAVR) strains (Fig. 1). This was hypothesised to be due to sequence polymorphism reducing read mapping efficiency and reflecting the higher overall CDS SNP rate in the latter two strains with respect to the MHco3(ISE)-derived reference genome sequence (Fig. 1). This hypothesis was supported by the improvement of overall read mapping efficiency achieved by increasing SNP mapping allowance to N5 (allowing five SNPs and three indels per read) from the default N2 value (allowing two SNPs or two indels per read). This change in SNP mapping allowance resulted in an increase in the number of reads mapped for a large number of gene models (Fig. 2A). This improvement in read mapping efficiency, as a result of increased SNP mapping allowance, was not confined to the MHco4(WRS) and MHco10(CAVR) data, but also occurred with the MHco3(ISE) data. These results suggest that mapping efficiency is affected by both between-strain and within-strain sequence polymorphism. We also investigated the extent to which sequence polymorphism varied among gene models and how this affected read mapping efficiency (Fig. 2B). When SNP allowances were increased from N2 to N5, genes with higher levels of polymorphism showed larger proportionate increases in

Please cite this article as: A. M. Rezansoff, R. Laing, A. Martinelli 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*, International Journal for Parasitology, https://doi.org/10.1016/j.ijpara.2019.05.012

reads mapped for all three strains (Fig. 2Ca). This further illustrates the impact of sequence polymorphism on RNAseq read mapping efficiency and how it is greater for more polymorphic genes.

Having shown that sequence polymorphism affects RNAseq read mapping to a reference genome assembly with *TopHat2*, we next investigated how this might bias differential expression analysis using *DESeq2*, one of the most commonly used bioinformatic tools for RNAseq data analysis (Figs. 3 and 4A). For each gene model, we plotted the *DESeq2* differential expression results against the difference in SNP rate (relative to the reference genome assembly) between the two strains being compared (Fig. 3). For each pairwise strain comparison, gene models which had greater differences in the level of sequence polymorphism between the strains were more likely to be down-regulated in the strain with the higher level of sequence polymorphism (Fig. 3). Further, this bias increased with the magnitude of difference in polymorphism rate of gene models between the strains (Figs. 3 and 4A). This effect was true for all three pairwise strain comparisons, including between the two "non-reference" MHco4(WRS) and MHco10 (CAVR) strains. There is no obvious biological reason for such differential expression biases, based on differences in SNP polymorphism rates, and so we concluded this is due to the effect of sequence polymorphism on RNAseq mapping rates.

Consequently, biases due to inter-strain differences in SNP polymorphism rates needed to be minimised before meaningful differential expression analysis could be performed. The first approach to achieve this was to choose RNAseq read mapping parameters in *TopHat2* to maximise read mapping efficiency for all the strains. Overall read mapping success peaked at the N5 or N6 SNP mapping allowances, depending on the strain (with very little difference between these two values (Fig. 1)). At the level of the gene model, the clear majority of genes had higher numbers of reads mapping at the N5 allowance than at either the N2 or N10 allowances (Fig. 2A). Consequently, the N5 mapping allowance maximised read mapping efficiency. Furthermore, the directional biases in the differential expression reports between strains were greatly reduced at the N5 mapping allowance (Fig. 4A, B, Supplementary Fig. S1). Therefore the N5 mapping allowance was considered optimal to use for further analysis. However, optimising the SNP mapping allowance did not completely remove the directional expression biases. For example, even at the N5 SNP mapping allowance, although the directional expression bias was close to zero for genes with SNP rate differences between strains of <2%, it persisted for genes with a difference in SNP rate of >2% (Fig. 4B). This led us to conclude that it was not possible to reliably measure differential expression for those genes with >2% SNP rate differences between strains, even at the N5 read mapping allowance. Consequently, we precluded these genes from subsequent transcriptomic analysis. These results have important implications for differential expression analysis between different strains/isolates of organisms with high levels of genetic diversity and suggest that sequence polymorphism needs to be defined and accounted for as part of the analysis. There are a number of other read mapping tools available for RNAseq analysis, some of which, although less widely used than *TopHat2*, may be less impacted by high levels of sequence polymorphism (Baruzzo et al., 2017). *TopHat2* is still widely used, but it is noteworthy that the mapping tool which is increasingly used in its place, *HISAT2,* is only slightly less sensitive to changes in mismatch parameters using simulated datasets (Baruzzo et al., 2017). Other read mapping tools such as *NovoAlign* (http://www.novocraft.com/products/novoalign/) or *GSNAP* (Wu and Nacu, 2010), that may be less impacted by sequence polymorphism, deserve more exploration for use in RNAseq differential expression pipelines for organisms such as *H. contortus* with high levels of genetic variation.

Pairwise comparisons of three genetically divergent strains of *H. contortus* revealed large numbers of differentially expressed genes, even after the confounding effects of sequence polymorphism were removed (Fig. 5). The proportion of differentially expressed genes between the *H. contortus* strains far exceed those previously observed in inter-population studies of vertebrate species such as human and mouse (Bottomly et al., 2011; Li et al., 2014), and it is greater than has been reported between different strains of *C. elegans* (N2/Bristol and CB4856/Hawaiian strains) (Capra et al., 2008; Francesconi and Lehner, 2014). This remarkably large number of differentially expressed genes between these *H. contortus* strains may have many different phenotypic traits which could have a variety of implications for their life history traits, epidemiology, pathogenicity and susceptibility to drugs and/or vaccines. This reflects the high genetic diversity of *H. contortus* and of these particular strains. MHco3(ISE), MHco4(WRS) and MHco10(CAVR) are derived from field isolates obtained from different continents and are highly genetically divergent (Redman et al., 2008, 2012; Gilleard and Redman, 2016). For example, the levels of genetic diversity (fixation index (Fst) values) between strains based on microsatellite genotyping ranged from 0.1530 to 0.2696 which is as high or higher than some closely related species in some cases (Redman et al., 2008; Prado-Martinez et al., 2013; Romiguier et al., 2014). Further, although the nematode body plan is superficially simple, a variety of morphological and morphometric traits vary between these three strains, including vulval morphology, oesophagus length, and spicule length in males as well as the extent of the synlophe cuticular ridges in females (Gilleard and Redman, 2016; Sargison et al., 2019). Also, there is evidence of lethality of some hybrid progeny of these strains (Sargison et al., 2019).

The results of this study also have important implications for anthelmintic resistance research which, until very recently, has been dominated by candidate gene studies (Gilleard, 2013, 2006; Rezansoff et al., 2016). In the case of ivermectin resistance, such studies have so far failed to identify the key loci or genes involved in resistance for any parasitic nematode, including *H. contortus* (Gilleard, 2013). One common component of candidate gene studies has been to compare the expression levels of specific candidate genes between a small number of ivermectin-resistant and -susceptible parasite strains (Xu et al., 1998; Dicker et al., 2011; El-Abdellati et al., 2011; Williamson et al., 2011). It is common for such studies to report differences in expression between resistant and susceptible strains for candidate genes such as P-glycoproteins (PGPs) or LGICs. These differences are commonly used as circumstantial evidence for a role in resistance. Our results here show the context in which such studies should be interpreted as a very large number of genes are differentially expressed in pairwise comparisons of genetically divergent *H. contortus* strains (Fig. 5). LPGs (824–1498) were differentially expressed between the strains in the study at a level of 2-fold and an adjusted statistical significance of $P < 0.05$ (as called by *DESeq2*). This highlights the inherently high levels of "background" transcriptomic variation that occur between genetically divergent *H. contortus* strains. Consequently, care must be taken when interpreting a suggested association of differential expression of a gene with a drug resistance phenotype when a small number of genes are compared between a small number of drug-resistant and -susceptible strains. This is particularly the case when the degree of genetic differentiation or the general level of transcriptomic difference that exists between the strains has not been assessed.

Recently, studies analysing the expression of small numbers of candidate genes are being replaced with more global transcriptomic studies. The draft *H. contortus* genome and its recent improvement into a chromosomal level assembly is making such studies increasingly feasible on a genome-wide scale (Laing et al., 2013; Doyle et al., 2018). The work presented here also has important implications for global transcriptomic comparisons of

Please cite this article as: A. M. Rezansoff, R. Laing, A. Martinelli 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*, International Journal for Parasitology, https://doi.org/10.1016/j.ijpara.2019.05.012

drug-resistant and -susceptible strains. Two gene families often suggested to be involved in ivermectin resistance are the LGICs and ABC transporter genes (Laing et al., 2013). We used the gene models in the *H. contortus* draft annotation to assess how many members of these gene families were differentially expressed between the MHco4(WRS) and MHco10(CAVR) ivermectin-resistant strains and the MHco3(ISE) susceptible strain using the default polymorphism allowance (N2), the optimised polymorphism allowance (N5), and the polymorphism allowance (N5) but removing the highly polymorphic gene set (Table 1). We found there was a dramatic reduction in the number of members of these genes families that were determined to be differentially expressed when polymorphism allowance was increased to the optimal N5 allowance (Table 1). A further reduction was apparent when the most highly polymorphic genes were discarded from the analysis (Table 1).

These results highlight the fact that a substantial number of differentially expressed genes reported are likely to be artifacts caused by differences in sequence polymorphism between the strains being compared which are not accounted for. In the case of our analysis, accounting for sequence polymorphism reveals a smaller number of differentially expressed candidate genes perhaps worthy of further investigation. The ABC transporter *Hco-pmp-6*, and two LGICs – *Hco-lgc-55* and *Hco-lgc-44* – were differentially expressed with the same directionality in both ivermectin-resistant strains relative to the MHco3(ISE) strain. *Hco-lgc-55* is a tyramine-gated chloride channel whose *C. elegans* homologue, *Cel-lgc-55,* is expressed in the pharynx and is involved in worm motility (Ringstad et al., 2009; Rao et al., 2010). The ABC transporter *Hco-wht-4*, and the LGICs *Hco-lgc-3*, *Hco-lgc-33*, *Hco-lgc-9*, and *Hco-acr-24*, were other genes with a >2× log2 fold-change differential expression in the MHco10(CAVR) strain, although these genes were not differentially expressed in the other resistant strain, MHco4(WRS). *Hco-lgc-3* was the gene with the highest level of up-regulation across both these gene families, being differentially expressed at greater than 50-fold in MHco3(CAVR) relative to MHco3(ISE) (Table 1). The gene may be considered of interest given its homology to a paralogous pair of *C. elegans* proton-gated ion channels, *Cel-pbo-5* and *Cel-pbo-6*, which are required for normal posterior muscle function (Beg et al., 2008). However, further functional and genetic studies are required before making any inferences of the potential role of these genes in mediating the ivermectin resistance phenotype of *H. contortus*.

Acknowledgements

We are grateful to Dr Matt Workentine for comments on the manuscript. We are grateful for funding from Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant (grant number RGPIN-2015-03976), NSERC-CREATE Host-Parasites Interactions (HPI) program and the Biotechnology and Biological Sciences Research Council (BBSRC), UK. The Moredun Research Institute, United Kingdom (DJB and AAM) receives funding from the Scottish Government. NS was funded by the Higher Education Funding Council of England (HEFCE), the Department for Environment, Food and Rural Affairs (DEFRA), UK and the Scottish Funding Council (SFC) Veterinary Training Research Initiative (VTRI) programme VT0102 and supported by Pfizer Animal Health. NH, SD and JAC contributions were supported by Wellcome Trust grant WT206194. RL and ED contributions were supported by BBSRC grant BB/M003949.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijpara.2019.05.012.

References

Anders, S., Pyl, P.T., Huber, W., 2014. HTSeq – A Python framework to work with high-throughput sequencing data HTSeq – A Python framework to work with high-throughput sequencing data. Bioinformatics 31. https://doi.org/10.1093/bioinformatics/btu638.

Andersen, E.C., Gerke, J.P., Shapiro, J.A., Crissman, J.R., Ghosh, R., Bloom, J.S., Felix, M.-A., Kruglyak, L., 2012. Chromosome-scale selective sweeps shape *Caenorhabditis elegans* genomic diversity. Nat. Genet. 44, 285–295. https://doi.org/10.1038/ng.1050.

Antony, H.A., Pathak, V., Parija, S.C., Ghosh, K., Bhattacherjee, A., 2016. Transcriptomic analysis of chloroquine-sensitive and chloroquine-resistant strains of *Plasmodium falciparum*: toward malaria diagnostics and therapeutics for global health. Omi. J. Integr. Biol. 20, 424–432. https://doi.org/10.1089/omi.2016.0058.

Baruzzo, G., Hayer, K.E., Kim, E.J., Camillo B, D.I., Fitzgerald, G.A., Grant, G.R., 2017. Simulation-based comprehensive benchmarking of RNA-seq aligners. Nat. Methods 14, 135–139. https://doi.org/10.1038/nmeth.4106.

Beg, A.A., Ernstrom, G.G., Nix, P., Davis, M.W., Jorgensen, E.M., 2008. Protons act as a transmitter for muscle contraction in *C. elegans*. Cell 132, 149–160. https://doi.org/10.1016/j.cell.2007.10.058.

Blumenthal, T., Davis, R.E., 2004. Exploring nematode diversity. Nat. Genet. https://doi.org/10.1038/ng1204-1246.

Bottomly, D., Walter, N.A.R., Hunter, J.E., Darakjian, P., Kawane, S., Buck, K.J., Searles, R.P., Mooney, M., McWeeney, S.K., Hitzemann, R., 2011. Evaluating gene expression in C57BL/6J and DBA/2J mouse striatum using RNA-Seq and microarrays. PLoS One 6. https://doi.org/10.1371/journal.pone.0017820.

Capra, E.J., Skrovanek, S.M., Kruglyak, L., 2008. Comparative developmental expression profiling of two *C. elegans* isolates. PLoS One 3. https://doi.org/10.1371/journal.pone.0004055.

Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., Ruden, D.M., 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain. Fly (Austin) 6, 80–92. https://doi.org/10.4161/fly.19695.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M.W., Gaffney, D.J., Elo, L.L., Zhang, X., Mortazavi, A., 2016. A survey of best practices for RNA-seq data analysis. Genome Biol. https://doi.org/10.1186/s13059-016-0881-8.

Croken, M.M.K., Ma, Y., Markillie, L.M., Taylor, R.C., Orr, G., Weiss, L.M., Kim, K., 2014. Distinct strains of *Toxoplasma gondii* feature divergent transcriptomes regardless of developmental stage. PLoS One 9, 1–10. https://doi.org/10.1371/journal.pone.0111297.

Dey, A., Chan, C.K.W., Thomas, C.G., Cutter, A.D., 2013. Molecular hyperdiversity defines populations of the nematode *Caenorhabditis brenneri*. Proc. Natl. Acad. Sci. 110, 11056–11060. https://doi.org/10.1073/pnas.1303057110.

Dicker, A.J., Nisbet, A.J., Skuce, P.J., 2011. Gene expression changes in a P-glycoprotein (Tci-pgp-9) putatively associated with ivermectin resistance in *Teladorsagia circumcincta*. Int. J. Parasitol. 41, 935–942. https://doi.org/10.1016/j.ijpara.2011.03.015.

Doyle, S.R., Laing, R., Bartley, D.J., Britton, C., Chaudhry, U., Gilleard, J.S., Holroyd, N., Mable, B.K., Maitland, K., Morrison, A.A., Tait, A., Tracey, A., Berriman, M., Devaney, E., Cotton, J.A., Sargison, N.D., 2018. 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. 10, 396–409. https://doi.org/10.1093/gbe/evx269.

Doyle, S.R., Illingworth, C.J.R., Laing, R., Bartley, D.J., Redman, E., Martinelli, A., Holroyd, N., Morrison, A.A., Rezansoff, A., Tracey, A., Devaney, E., Berriman, M., Sargison, N., Cotton, J.A., Gilleard, J.S., 2019. Population genomic and evolutionary modelling analyses reveal a single major QTL for ivermectin drug resistance in the pathogenic nematode, *Haemonchus contortus*. BMC Genomics 20, 218. https://doi.org/10.1186/s12864-019-5592-6.

Edwards, J.A., Chen, C., Kemski, M.M., Hu, J., Mitchell, T.K., Rappleye, C.A., 2013. Histoplasma yeast and mycelial transcriptomes reveal pathogenic-phase and lineage-specific gene expression profiles. BMC Genomics 14, 695. https://doi.org/10.1186/1471-2164-14-695.

El-Abdellati, A., De Graef, J., Van Zeveren, A., Donnan, A., Skuce, P., Walsh, T., Wolstenholme, A., Tait, A., Vercruysse, J., Claerebout, E., Geldhof, P., 2011. Altered avr-14B gene transcription patterns in ivermectin-resistant isolates of the cattle parasites, *Cooperia oncophora* and *Ostertagia ostertagi*. Int. J. Parasitol. 41, 951–957. https://doi.org/10.1016/j.ijpara.2011.04.003.

Fiebig, M., Kelly, S., Gluenz, E., 2015. Comparative life cycle transcriptomics revises *Leishmania mexicana* genome annotation and links a chromosome duplication with parasitism of vertebrates. PLoS Pathog. 11, 1–28. https://doi.org/10.1371/journal.ppat.1005186.

Francesconi, M., Lehner, B., 2014. The effects of genetic variation on gene expression dynamics during development. Nature 505, 208–211. https://doi.org/10.1038/nature12772.

Gilleard, J.S., 2013. Haemonchus contortus as a paradigm and model to study anthelmintic drug resistance. Parasitology 140, 1506–1522. https://doi.org/10.1017/S0031182013001145.

Gilleard, J.S., Redman, E., 2016. Genetic diversity and population structure of *Haemonchus contortus*. Adv. Parasitol. 93, 31–68. https://doi.org/10.1016/bs.apar.2016.02.009.

Gilleard, J.S., 2006. Understanding anthelmintic resistance: the need for genomics and genetics. J. Parasitol. Int. https://doi.org/10.1016/j.ijpara.2006.06.010.

Please cite this article as: A. M. Rezansoff, R. Laing, A. Martinelli 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*, International Journal for Parasitology, https://doi.org/10.1016/j.ijpara.2019.05.012

Guryev, V., Koudijs, M.J., Berezikov, E., Johnson, S.L., Plasterk, R.H.A., van Eeden, F.J. M., Cuppen, E., 2006. Genetic variation in the zebrafish. Genome Res. 16, 491–497. https://doi.org/10.1101/gr.4791006.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., Salzberg, S.L., 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 14, R36. https://doi.org/10.1186/gb-2013-14-4-r36.

Kim, D., Langmead, B., Salzberg, S.L., 2015. HISAT: a fast spliced aligner with low memory requirements. Nat. Methods 12, 357–360. https://doi.org/10.1038/nmeth.3317.

Laing, R., Hunt, M., Protasio, A.V., Saunders, G., Mungall, K., Laing, S., Jackson, F., Quail, M., Beech, R., Berriman, M., Gilleard, J.S., 2011. Annotation of two large contiguous regions from the *Haemonchus contortus* genome using RNA-seq and comparative analysis with *Caenorhabditis elegans*. PLoS One 6,. https://doi.org/10.1371/journal.pone.0023216 e23216.

Laing, R., Kikuchi, T., Martinelli, A., Tsai, I., Beech, R., Redman, E., Holroyd, N., Bartley, D., Beasley, H., Britton, C., Curran, D., Devaney, E., Gilabert, A., Hunt, M., Jackson, F., Johnston, S., Kryukov, I., Li, K., Morrison, A., Reid, A., Sargison, N., Saunders, G., Wasmuth, J., Wolstenholme, A., Berriman, M., Gilleard, J., Cotton, J., 2013. The genome and transcriptome of *Haemonchus contortus*, a key model parasite for drug and vaccine discovery. Genome Biol. 14, R88. https://doi.org/10.1186/gb-2013-14-8-r88.

Le Jambre, L., Gill, J., Lenane, I., Lacey, E., 1995. Characterization of an avermectin resistant strain of Australian *Haemonchus contortus*. Int. J. Parasitol. 25, 691–698.

Li, J., Lai, K., Ching, A.K.K., Chan, T., 2014. Genomics Transcriptome sequencing of Chinese and Caucasian population identifies ethnic-associated differential transcript abundance of heterogeneous nuclear ribonucleoprotein K (hnRNPK). Genomics 103, 56–64. https://doi.org/10.1016/j.ygeno.2013.12.005.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 1000 Genome Project Data Processing Subgroup, 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

Lindblad-Toh, K., Lander, E.S., Winchester, E., Daly, M.J., Wang, D.G., Hirschhorn, J.N., Laviolette, J.-P., Ardlie, K., Reich, D.E., Robinson, E., Sklar, P., Shah, N., Thomas, D., Fan, J.-B., Gingeras, T., Warrington, J., Patil, N., Hudson, T.J., 2000. Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse. Nat. Genet. 24, 381–386. https://doi.org/10.1038/74215.

Love, M.I., Huber, W., Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15, 1–21. https://doi.org/10.1186/s13059-014-0550-8.

Papenfort, K., Förstner, K.U., Cong, J., Sharma, C.M., Bassler, B.L., 2015. Differential RNA-seq of Vibrio cholerae identifies the VqmR small RNA as a regulator of biofilm formation. Proc. Natl. Acad. Sci. 112, E766–E775. https://doi.org/10.1073/pnas.1500203112.

Prado-Martinez, J., Sudmant, P.H., Kidd, J.M., Li, H., Kelley, J.L., Lorente-Galdos, B., Veeramah, K.R., Woerner, A.E., O'Connor, T.D., Santpere, G., Cagan, A., Theunert, C., Casals, F., Laayouni, H., Munch, K., Hobolth, A., Halager, A.E., Malig, M., Hernandez-Rodriguez, J., Hernando-Herraez, I., Prüfer, K., Pybus, M., Johnstone, L., Lachmann, M., Alkan, C., Twigg, D., Petit, N., Baker, C., Hormozdiari, F., Fernandez-Callejo, M., Dabad, M., Wilson, M.L., Stevison, L., Camprubí, C., Carvalho, T., Ruiz-Herrera, A., Vives, L., Mele, M., Abello, T., Kondova, I., Bontrop, R.E., Pusey, A., Lankester, F., Kiyang, J.A., Bergl, R.A., Lonsdorf, E., Myers, S., Ventura, M., Gagneux, P., Comas, D., Siegismund, H., Blanc, J., Agueda-Calpena, L., Gut, M., Fulton, L., Tishkoff, S.A., Mullikin, J.C., Wilson, R.K., Gut, I.G., Gonder, M.K., Ryder, O.A., Hahn, B.H., Navarro, A., Akey, J.M., Bertranpetit, J., Reich, D., Mailund, T., Schierup, M.H., Hvilsom, C., Andrés, A.M., Wall, J.D., Bustamante, C. D., Hammer, M.F., Eichler, E.E., Marques-Bonet, T., 2013. Great ape genetic diversity and population history. Nature 499, 471–475. https://doi.org/10.1038/nature12228.

Rao, V.T.S., Accardi, M.V., Siddiqui, S.Z., Beech, R.N., Prichard, R.K., Forrester, S.G., 2010. Characterization of a novel tyramine-gated chloride channel from *Haemonchus contortus*. Mol. Biochem. Parasitol. 173, 64–68. https://doi.org/10.1016/j.molbiopara.2010.05.005.

Redman, E., Packard, E., Grillo, V., Smith, J., Jackson, F., Gilleard, J.S., 2008. Microsatellite analysis reveals marked genetic differentiation between *Haemonchus contortus* laboratory isolates and provides a rapid system of genetic fingerprinting. Int. J. Parasitol. 38, 111–122. https://doi.org/10.1016/j.ijpara.2007.06.008.

Redman, E., Sargison, N., Whitelaw, F., Jackson, F., Morrison, A., Bartley, D.J., Gilleard, J.S., 2012. Introgression of ivermectin resistance genes into a susceptible *Haemonchus contortus* strain by multiple backcrossing. PLoS Pathog. 8,. https://doi.org/10.1371/journal.ppat.1002534 e1002534.

Redman, E., Whitelaw, F., Tait, A., Burgess, C., Bartley, Y., Skuce, P.J., Jackson, F., Gilleard, J.S., 2015. The emergence of resistance to the benzimidazole anthlemintics in parasitic nematodes of livestock is characterised by multiple independent hard and soft selective sweeps. PLoS Negl. Trop. Dis. 9,. https://doi.org/10.1371/journal.pntd.0003494 e0003494.

Rezansoff, A.M., Laing, R., Gilleard, J.S., 2016. Evidence from two independent backcross experiments supports genetic linkage of microsatellite Hcms8a20, but not other candidate loci, to a major ivermectin resistance locus in *Haemonchus contortus*. Int. J. Parasitol. 46, 653–661. https://doi.org/10.1016/j.ijpara.2016.04.007.

Ringstad, N., Abe, N., Horvitz, H.R., 2009. Ligand-gated chloride channels are receptors for biogenic amines in *C. elegans*. Science 325, 96–100. https://doi.org/10.1126/science.1169243.

Romiguier, J., Gayral, P., Ballenghien, M., Bernard, A., Cahais, V., Chenuil, A., Chiari, Y., Dernat, R., Duret, L., Faivre, N., Loire, E., Lourenco, J.M., Nabholz, B., Roux, C., Tsagkogeorga, G., Weber, A.A.T., Weinert, L.A., Belkhir, K., Bierne, N., Glémin, S., Galtier, N., 2014. Comparative population genomics in animals uncovers the determinants of genetic diversity. Nature 515, 261–263. https://doi.org/10.1038/nature13685.

Sargison, N.D., Redman, E., Morrison, A.A., Bartley, D.J., Jackson, F., Hoberg, E., Gilleard, J.S., 2019. Mating barriers between genetically divergent strains of the parasitic nematode Haemonchus contortus suggest incipient speciation. Int. J. Parasitol. https://doi.org/10.1016/j.ijpara.2019.02.008.

Urdaneta-Marquez, L., Bae, S.H., Janukavicius, P., Beech, R., Dent, J., Prichard, R., 2014. A dyf-7 haplotype causes sensory neuron defects and is associated with macrocyclic lactone resistance worldwide in the nematode parasite *Haemonchus contortus*. Int. J. Parasitol. 44, 1063–1071. https://doi.org/10.1016/j.ijpara.2014.08.005.

Van Wyk, J.A., Malan, F.S., 1988. Resistance of field strains of *Haemonchus contortus* to ivermectin, closantel, rafoxanide and the benzimidazoles in South Africa. Vet. Rec. 123, 226–228. https://doi.org/10.1136/vr.123.9.226.

Wang, D.G., 1998. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science (80-) 280, 1077–1082. https://doi.org/10.1126/science.280.5366.1077.

Wang, Z., Gerstein, M., Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. 10, 57–63. https://doi.org/10.1038/nrg2484.

Williamson, S.M., Storey, B., Howell, S., Harper, K.M., Kaplan, R.M., Wolstenholme, A. J., 2011. Candidate anthelmintic resistance-associated gene expression and sequence polymorphisms in a triple-resistant field isolate of *Haemonchus contortus*. Mol. Biochem. Parasitol. 180, 99–105. https://doi.org/10.1016/j.molbiopara.2011.09.003.

Wu, T.D., Nacu, S., 2010. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics 26, 873–881. https://doi.org/10.1093/bioinformatics/btq057.

Xu, M., Molento, M., Blackhall, W., Ribeiro, P., Beech, R., Prichard, R., 1998. Ivermectin resistance in nematodes may be caused by alteration of P-glycoprotein homolog. Mol. Biochem. Parasitol. 91, 327–335. https://doi.org/10.1016/S0166-6851(97)00215-6.

Please cite this article as: A. M. Rezansoff, R. Laing, A. Martinelli 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*, International Journal for Parasitology, https://doi.org/10.1016/j.ijpara.2019.05.012