



Rezansoff, A. M. et al. (2019) 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*, 49(11), pp. 847-858. (doi:[10.1016/j.ijpara.2019.05.012](https://doi.org/10.1016/j.ijpara.2019.05.012))

There may be differences between this version and the published version. You are advised to consult the publisher's version if you wish to cite from it.

<http://eprints.gla.ac.uk/201176/>

Deposited on: 22 October 2019

Enlighten – Research publications by members of the University of Glasgow  
<http://eprints.gla.ac.uk>

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

3  
4  
5 Andrew M. Rezansoff<sup>a</sup>, Roz Laing<sup>b</sup>, Axel Martinelli<sup>c</sup>, Susan Stasiuk<sup>a</sup>, Elizabeth Redman<sup>a</sup>, Dave  
6 Bartley<sup>d</sup>, Nancy Holroyd<sup>c</sup>, Eileen Devaney<sup>b</sup>, Neil D. Sargison<sup>e</sup>, Stephen Doyle<sup>c</sup>, James Cotton<sup>c</sup>, John S.  
7 Gilleard<sup>a,\*</sup>

8  
9 <sup>a</sup>*Department of Comparative Biology and Experimental Medicine, Faculty of Veterinary Medicine,*  
10 *University of Calgary, Alberta, Canada*

11 <sup>b</sup>*Institute of Biodiversity, Animal Health and Comparative Medicine, College of Medical, Veterinary*  
12 *and Life Sciences, University of Glasgow, Scotland, United Kingdom*

13 <sup>c</sup>*Wellcome Sanger Institute, Hinxton, Cambridgeshire, CB10 1SA, United Kingdom*

14 <sup>d</sup>*Moredun Research Institute, Pentlands Science Park, Bush Loan, Penicuik EH26 0PZ, United*  
15 *Kingdom*

16 <sup>e</sup>*University of Edinburgh, Royal (Dick) School of Veterinary Studies, Easter Bush Veterinary Centre,*  
17 *Roslin, Midlothian, EH25 9RG, United Kingdom.*

18  
19  
20 \* Corresponding author. Professor John Gilleard, Department of Comparative Biology and  
21 Experimental Medicine, Faculty of Veterinary Medicine, 3330, Hospital Drive, University of Calgary,  
22 Calgary, Alberta, T2N 4N1 Canada

23 Tel.: +1 403 210 6327; Fax: +1 403 210 7882. E-mail address: [jsgillea@ucalgary.ca](mailto:jsgillea@ucalgary.ca)

24

25 **Abstract**

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

47 these must be considered before associating the differential expression of specific genes with  
48 phenotypic differences between strains.

49

50 *Keywords: Haemonchus contortus; Transcriptomics; RNAseq; Differential expression; Ivermectin;*

51 Anthelmintic resistance

52

53

## 54 **1. Introduction**

55

56 RNAseq has become the standard approach for the genome-wide analysis and quantification of  
57 gene expression across the life sciences (Wang et al., 2009; Conesa et al., 2016). Established sequence  
58 aligners used in RNAseq analysis pipelines, such as *TopHat2* and its faster successor *HISAT2*, were  
59 developed and their default mapping parameters set, primarily for use on vertebrate species such as  
60 humans, mice, and zebrafish, which have relatively low levels of both intra- and inter-population  
61 genetic diversity (Wang, 1998; Lindblad-Toh et al., 2000; Guryev et al., 2006; Baruzzo et al., 2017).  
62 Further, until relatively recently, applications of RNAseq to non-vertebrate species were largely  
63 confined to laboratory strains of model organisms such as *Drosophila melanogaster* and  
64 *Caenorhabditis elegans*, which also have relatively low levels of genetic diversity (Andersen et al.,  
65 2012; Cingolani et al., 2012). Consequently, most publications make little or no acknowledgement of  
66 the potentially confounding effects of sequence polymorphism on the mapping efficiency of RNAseq  
67 reads and the calling of differentially expressed genes (Baruzzo et al., 2017). RNAseq analysis  
68 pipelines are generally applied to non-model organisms simply using established default parameters,  
69 with no consideration given the level and distribution of sequence polymorphism within and between  
70 the strains or populations being compared (Edwards et al., 2013; Croken et al., 2014; Fiebig et al.,  
71 2015; Papenfort et al., 2015; Antony et al., 2016). However, many taxa show high levels and complex  
72 patterns of intra-species genetic diversity (Blumenthal and Davis, 2004; Dey et al., 2013; Romiguier et  
73 al., 2014; Redman et al., 2015). This is a concern since standard RNAseq alignment benchmarking  
74 studies have shown that the performance of different sequence aligners varies with the genome  
75 complexity and levels of sequence polymorphism when using simulated sequence data (Baruzzo et al.,

76 2017). However, no published experimental studies directly examine the effects of sequence  
77 polymorphism on differential expression analyses using commonly applied RNAseq analysis pipelines.

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

94

## 95 **2. Materials and methods**

96

### 97 *2.1. Haemonchus contortus* strains, sample preparation and sequencing

98

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

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

116

## 117 *2.2. Sequence quality control and read mapping*

118

119           Raw 100 bp reads were inspected using *FastQC* (Andrews, 2010) for overall dataset integrity and  
120 all reads were trimmed at the 5' end by 10 bases. Fifteen bases were also trimmed from the 3' ends of  
121 all reads to remove low quality sequence characteristic of 3' tail ends. The post-trimmed 75 bp reads

122 were used for mapping to the *H. contortus* MHco3(ISE) reference genome assembly (Laing et al.,  
123 2013) with *TopHat2* (Dobin and Gingeras, 2013). The assembly used is an improved version (N50 of  
124 5.24 MB) of the original published *H. contortus* genome assembly (GenBank ID PRJEB506 - N50 of  
125 83.29 kb (Laing et al., 2013)) and contains an expanded set of annotated gene models  
126 (<https://data.mendeley.com/drafts/4z6xv5j5zf>). Numerical identifiers of these additional gene models  
127 begin with HCOI\_0500, and have not yet been submitted to online genomic resources (e.g.  
128 [Uniprot.org](http://Uniprot.org)).

129 *TopHat2* was executed using the following parameter settings: *TopHat2 -N (#) --read-gap-length*  
130 *(%) --read-edit-dist (# + %) -I 40000 -r 200 -a 6 -g 1 --no-discordant --no-mixed --min-intron 10 --*  
131 *microexon-search --mate-std-dev 50 --library-type fr-unstranded ./reference.fasta*  
132 *trimmed\_forward\_reads.fastq trimmed\_reverse\_reads.fastq*. Only *-N* (specifying the number of SNPs  
133 per mapped read allowed by *TopHat2*), *--read-gap-length* (the allowed base count of any indels), and *--*  
134 *read-edit-dist* (the allowed combined base count of both *-N* and *--read-gap-length*) were adjusted  
135 throughout the experiment. Reads of all triplicates of all three populations were initially mapped with  
136 *TopHat2* using a scale of SNP (polymorphism) allowances from 2 to 10 SNPs (*-N*) per read with indel  
137 allowance (*--read-gap-length*) held constant at 3 bases.

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

145

### 146 2.3. RNAseq processing and analysis

147

148 Reads mapped to each gene model were sorted with *samtools sort*, and counted for each of the  
149 three bioreplicates for each strain at the three different SNP allowances – N2, N5, N10 – using the  
150 following command in *HTseq-count*: *htseq-count -i parent -q -s no -f bam -t cds*  
151 *./sorted\_accepted\_hits.bam ./genome\_annotation\_file.gff3* (Anders et al., 2014). Raw mapped read  
152 counts for each gene model of each bioreplicate of each strain were compiled and used as input for  
153 *DESeq2*.

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

161

### 162 2.4. Categorizing gene models on the basis of SNP rates and SNP rate differences between strains

163

164 SNPs within CDS were called using *samtools mpileup* on whole genome sequence (WGS)  
165 datasets created for each of the strains against the MHco3(ISE) genome assembly (Doyle et al., 2019).  
166 SNPs present at > 40% frequency were totaled per gene model for each of the strains. The SNP rate  
167 was calculated for each gene in each strain by dividing the total number of SNPs in the gene by the

168 respective gene model CDS length. The genes were then categorized in two different ways for  
169 subsequent investigation of the effect of sequence polymorphism on read mapping and RNAseq  
170 analysis. First, they were categorized based on their SNP rates in each strain: categories 0%, 0-0.5%,  
171 0.5-1%, 1-2%, 2-5%, and > 5%. Second, they were categorized based on the difference in SNP rates for  
172 each of the three pairwise strain comparisons (i.e. the SNP rate observed in one strain subtracted by?  
173 the SNP rate observed in the other) categories >5-15%, >2-5%, >0-2%, 0%. Genes with a >15%  
174 difference and were not categorized as they were likely to be due to annotation errors and/or overly  
175 short CDS lengths.

## 176

### 177 2.5. Assessment of genuine transcriptomic variation between the strains

## 178

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

188 Gene ontological classifications were obtained from *UniProt.org* (The UniProt Consortium, 2015)  
189 for *H. contortus* gene models of the originally published annotation (Laing et al., 2013). LPGs with  
190 ontological classifications were used as the reference gene set against which enrichment was assessed.

191 Functional enrichment was called in genes  $> \log_2$  1X fold-change differentially expressed in each  
192 pairwise, and each shared strain, comparison. *FunRich* (Pathan et al., 2015) was used to call enriched  
193 gene ontological classes using a statistical significance threshold of Benjamini-Hochberg corrected  
194 FDR adjusted  $P$  values  $< 0.05$ .

195

### 196 **3. Results**

197

198 *3.1. Coding sequence polymorphism affects RNAseq read mapping against the MHco3(ISE) reference*  
199 *assembly for the three different H. contortus strains*

200

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

208 The percentage of RNAseq reads that mapped to the MHco3(ISE) reference genome assembly,  
209 using the default SNP allowance (N2 – allowing two SNPs or two indels per read) in *TopHat2*, was  
210 60.7%, 44.8% and 47.1% for the MHco3(ISE), MHco4(WRS) and MHco10(CAVR) strains,  
211 respectively (Fig. 1). Increasing the *TopHat2* SNP allowance parameter changed the percentage of  
212 RNAseq reads that mapped (Fig. 1). For the MHco3(ISE) strain, the percentage of RNAseq reads  
213 mapping to the reference genome increased as the polymorphism allowance was increased from N2 to

214 N5 (allowing five SNPs and three indels per read) and then decreased as the allowance was further  
215 increased to N10 (allowing 10 SNPs and six indels per read) (Fig. 1). This pattern was very similar for  
216 the MHco4(WRS) and MHco10(CAVR) strains but the maximum percentage of reads mapping  
217 occurred at the N6 allowance, albeit at rates only 0.1% greater than at N5 (Fig. 1). The percentage of  
218 RNAseq reads that mapped to the reference MHco3(ISE) genome assembly was greater for the  
219 MHco3(ISE) strain than for the other two strains at all polymorphism allowances, although the  
220 magnitude of this difference decreased from the N2 to N10 allowance (Fig. 1).

221 A more detailed analysis was undertaken for the N2, N5 and N10 polymorphism allowances at the  
222 level of gene models. Increasing the polymorphism allowance from N2 to N5 resulted in 12,778,  
223 11,101, and 11,324 gene models having a >1% increase in the number of mapped RNAseq reads for  
224 MHco3(ISE), MHco4(WRS), and MHco10(CAVR), respectively (Fig. 2Aa). In contrast, 591, 1,316,  
225 and 1,563 genes showed a >1% decrease in RNAseq reads mapped (Fig. 2Aa). Further increasing the  
226 mapping allowance from N5 to N10 had the opposite effect, with a greater number of gene models  
227 having a decreased rather than an increased number of RNAseq reads mapped: a change in the  
228 polymorphism allowance from N5 to N10 resulted in 12,529, 8,139, and 8,470 gene models having a  
229 >1% decreased number of RNAseq reads mapped, compared with 1,092, 4,682 and 4,953 genes having  
230 an increased number of RNAseq reads mapped for MHco3(ISE), MHco4(WRS), and MHco10(CAVR)  
231 strains, respectively (Fig. 2Ab).

232

233 *3.2. The SNP allowance has a greater effect on RNAseq read mapping for gene models with higher*  
234 *levels of sequence polymorphism*

235

236 There were large differences in the SNP rates of different gene models, relative to the  
237 MHco3(ISE) reference genome, ranging from those with SNP rates of 0% to those above 5%. The  
238 25,111 gene models were binned into several different SNP rate categories to investigate how the  
239 mapping of RNAseq reads to the reference MHco3(ISE) genome assembly was affected by the coding  
240 region SNP rate (Fig. 2B). The MHco4(WRS) and MHco10(CAVR) strains had a significantly greater  
241 proportion of gene models with SNP rates greater than 0.5% (18,910 (75.3%) and 18,886 (75.2%),  
242 respectively) compared with the MHco3(ISE) strain (11,303 (45.0%)] (Z-stat = 69.3 ( $P < 0.000$ ) and  
243 69.1 ( $P < 0.000$ ), respectively) (Fig. 2B).

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

252

### 253 *3.3. High levels of sequence polymorphism artificially inflate between-strain RNAseq differential* 254 *expression results*

255

256 We next investigated the influence of CDS polymorphism on the RNAseq differential expression  
257 reported by *DESeq2* between pairwise strain comparisons. We hypothesized that gene models with  
258 large differences in SNP rates (SNPs/bp) between two strains are more likely to be reported as

259 differentially expressed between those strains than gene models with smaller SNP rate differences. To  
260 test this hypothesis, for each gene model we first determined the difference in SNP rates (SNPs/bp)  
261 between each pairwise comparison of the three strains. We then plotted the difference in the SNP rate  
262 between the two strains against the log<sub>2</sub>-fold difference in expression called by *DESeq2* for each gene  
263 model (Fig. 3). Using the MHco4(WRS) and MHco3(ISE) pairwise comparison as an example, for  
264 those gene models with a higher SNP rate in MHco4(WRS) than in MHco3(ISE), a greater number was  
265 reported by *DESeq2* as down-regulated in MHco4(WRS) relative to MHco3(ISE) than as up-regulated  
266 (Fig. 3A). This bias towards down-regulation increased as the SNP rate difference of gene models  
267 between the two strains increased (Fig. 3A). For gene models with a lower SNP rate in MHco4(WRS)  
268 than in MHco3(ISE), the opposite trend was apparent (Fig. 3B). Similar patterns were observed in both  
269 the MHco3(ISE) versus MHco10(CAVR) and MHco4(WRS) versus MHco10(CAVR) pairwise  
270 comparisons (Fig. 3C-F).

271 To further quantify how SNP rate differences between the strains biases reporting of differential  
272 expression, we placed each of the 25,049 gene models with SNP rate data into one of seven “SNP rate  
273 difference” categories for each pairwise strain comparison (data for the MHco3(ISE) versus  
274 MHco4(WRS) pairwise comparison is shown in Fig. 4, and Supplementary Table S2). The percentage  
275 of gene models reported as differentially expressed (with adjusted *P* values < 0.05 and > log<sub>2</sub> 1X fold-  
276 change in expression) was lowest for the 0% SNP rate difference category and increased as the SNP  
277 rate difference category increased (Fig. 4A). This trend was seen at all three SNP mapping allowances  
278 (Fig. 4A). There was also a strong relationship between the directionality of the differential expression  
279 called by *DESeq2* and the directionality of the SNP rate difference between the strains. For SNP rate  
280 difference categories where the SNP rate was greater in MHco4(WRS) than in MHco3(ISE) by at least  
281 2%, the large majority of gene models reported as differentially expressed were down-regulated in

282 MHco4(WRS) relative to MHco3(ISE) (396/425 (93.2%)) (Supplementary Table S2). Conversely, the  
283 large majority of gene models with SNP rates at least 2% lower in MHco4(WRS) than in MHco3(ISE),  
284 were up-regulated in MHco4(WRS) relative to MHco3(ISE) (21/27 (77.8%)) (Supplementary Table  
285 S2).

286

### 287 *3.4. Minimizing the effect of sequence polymorphism differences on differential expression analysis in* 288 *pairwise strain comparisons*

289

290 We next investigated ways to minimize the effect of sequence polymorphism on global  
291 transcriptomic differential expression analysis in pairwise strain comparisons. We first examined the  
292 effect of changing the read mapping polymorphism allowance on the number and bias of the  
293 differentially expressed genes reported by *DESeq2* in pairwise strain comparisons. When the  
294 polymorphism allowance was changed from N2 to N5 or from N5 to N10, there was an overall  
295 decrease in the total number of differentially expressed genes reported in all three pairwise strain  
296 comparisons (Supplementary Table S3). This trend was generally observed for genes in all SNP rate  
297 difference categories (see example of MHco3(ISE) versus MHco4(WRS) pairwise comparison in Fig.  
298 4A). At the default N2 polymorphism allowance, *DESeq2* reported more genes down-regulated than  
299 up-regulated in both MHco4(WRS) and MHco10(CAVR) when each was compared with MHco3(ISE)  
300 (Supplementary Fig. S1; Supplementary Table S3). This bias was reduced as the mapping allowance  
301 was increased to N5 and then N10 (Supplementary Fig. S1; Supplementary Table S3). In contrast, the  
302 MHco4(WRS) and MHco10(CAVR) pairwise comparison showed a relatively equal ratio of down-  
303 regulated and up-regulated gene numbers even at the default N2 polymorphism allowance  
304 (Supplementary Fig. S1; Supplementary Table S3).

305 We then calculated the net (overall mean) differential expression (NDE) of all gene models in  
306 each of the seven “SNP rate difference” categories for each of the pairwise strain comparisons to see if  
307 there was an overall directional bias to the data (data for the MHco4(WRS) and MHco3(ISE) pairwise  
308 strain comparison is shown in Fig. 4B). The NDE in the direction MHco4(WRS) > MHco3(ISE) was  
309 greatest for those gene models in the 5 - 15% MHco4(WRS) > MHco3(ISE) SNP rate difference  
310 category and least for gene models in the 0% SNP rate difference category (Fig. 4B, Supplementary  
311 Table S2A [see comment in table file]). Conversely, the NDE in the direction MHco4(WRS) <  
312 MHco3(ISE) was highest for gene models in the 5 - 15% MHco4(WRS) < MHco3(ISE) SNP rate  
313 difference category and least for the 0% SNP rate difference category (Fig. 4B, Supplementary Table  
314 S2A). The NDE of gene models between strains was highest at the N2 polymorphism mapping  
315 allowance, and least for the N10 polymorphism mapping allowance, in all SNP rate difference  
316 categories (Fig. 4B; Supplementary Table S2A).

317 The NDE of gene models between the strains was relatively close to zero for genes of the three  
318 lowest SNP rate difference categories, particularly at the N5 and N10 polymorphism allowances (Fig.  
319 4B; Supplementary Table S2B). This suggests that gene models with < 2% difference in SNP rate  
320 between strains had a minimal bias in pairwise strain differential expression analyses. We defined these  
321 gene models as “LPG models” in the subsequent differential expression analysis. These represent  
322 17,881 out of the total of 25,111 gene models in the *H. contortus* whole genome annotation (71.2%)  
323 and so represent the majority of gene models (Supplementary Fig. S2).

324

325 *3.5. Investigating genuine transcriptional differences between H. contortus strains*

326

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

337 A total of 1,125 (5.41% of LPGs), 1,498 (7.72% of LPGs), and 824 (3.59% of LPGs) genes were  
338 differentially expressed at  $> 1X \log_2$  fold in the MHco4(WRS) versus MHco3(ISE), MHco10(CAVR)  
339 versus MHco3(ISE), and MHco4(WRS) versus MHco10(CAVR) pairwise comparisons, respectively  
340 (Fig. 5). Of these, 134 genes (41 up-regulated, 93 down-regulated), 259 genes (121 up-regulated, 138  
341 down regulated), and 103 genes (40 up-regulated, 63 down regulated) were  $> 2X \log_2$  fold  
342 differentially expressed, respectively (Fig. 5). The large majority of the most differentially expressed  
343 genes in all strains comparisons were either undescribed or had only broad ontological classifications  
344 (Supplementary Table S4). No previously reported ivermectin resistance candidate LPGs were  
345 observed to be differentially expressed in at  $> 2X \log_2$  fold-change expression in either of the two  
346 ivermectin resistance strains relative to the MHCo3(ISE)-susceptible strain (Supplementary Table S4).

347 We examined the number of genes that were differentially expressed in more than one of the  
348 pairwise strain comparisons to see if a set of genes was common to different pairwise comparisons. The  
349 highest proportion of shared differentially expressed LPGs was between the MHco4(WRS) versus

350 MHco3(ISE) and MHco10(CAVR) versus MHco3(ISE) pairwise strain comparisons (Supplementary  
351 Fig. S4). Of the 2,132 gene models differentially expressed between either MHco4(WRS) and  
352 MHco10(CAVR) versus MHco3(ISE), 491 (23.03%) were differentially expressed with the same  
353 directionality (up- or down- regulated) in both pairwise comparisons at  $>1X$  log<sub>2</sub> fold change (48 gene  
354 models at  $>2X$  log<sub>2</sub> fold change) (Supplementary Fig. S4A). Fewer genes were shared in the other two  
355 strain combinations. Of the 2,025 gene models differentially expressed between either MHco3(ISE)  
356 and MHco4(WRS) strains versus MHco10(CAVR), 297 (14.67%) gene models were differentially  
357 expressed with the same directionality at  $>1$  log<sub>2</sub>-fold change (39 gene models at  $>2$  log<sub>2</sub>-fold  
358 change) in both pairwise comparisons (Supplementary Fig. S4B). Of the 1,794 gene models  
359 differentially expressed between either MHco3(ISE) and MHco10(CAVR) versus MHco4(WRS), only  
360 155 (8.64%) gene models were differentially expressed at  $>1$  log<sub>2</sub>-fold change (eight gene models at  
361  $>2$  log<sub>2</sub> fold change) with the same directionality in both comparisons (Supplementary Fig. S4C). Both  
362 these percentages represent a significantly lower proportion of differentially expressed genes shared  
363 than were observed shared in MHco4(WRS) and MHco10(CAVR) versus MHco3(ISE) (Z-stats = 6.8  
364 ( $P < 0.000$ ), and 12.1 ( $P < 0.000$ ), respectively).

365

366 *3.6. Investigating the effect of sequence polymorphism on differential expression analysis of two*  
367 *gene families of relevance to ivermectin resistance research*

368

369 Sixty-seven ligand-gated chloride channels (LGICs) and 86 ABC transporters identified in the  
370 published *H. contortus* draft genome (Laing et al, 2013) were examined for differential expression  
371 between the MHco4(WRS) and MHco10(CAVR) ivermectin-resistant strains and the susceptible  
372 MHco3(ISE) strain. Three different differential expression analyses were compared to assess the

373 impact of accounting for sequence polymorphisms differences between the strains; using the default N2  
374 SNP allowance on all 25,111 gene models, using the N5 SNP allowance on all 25,111 genes, and using  
375 the N5 SNP allowance on the set of 17,881 LPGs. There was a substantial reduction in the total number  
376 of differentially expressed genes reported using the N5 allowance on the LPG gene set compared with  
377 the N2 default allowance on the full gene set (Table 1). When comparing the two ivermectin-resistant  
378 strains with the ivermectin-sensitive [susceptible?] strain, only three of the LPGs – *Hco-lgc-55*, *Hco-*  
379 *pmp-6*, and *Hco-lgc-44* – showed differential expression at the N5 allowance in both the MHco4(WRS)  
380 and MHco10(CAVR) versus MHco3(ISE) pairwise comparisons. *Hco-lgc-55* had > 2X log<sub>2</sub> fold up-  
381 regulation in both cases (Table 1).

382

#### 383 **4. Discussion**

384

385 Differential expression analysis, either at the single gene or whole transcriptome level, between  
386 parasitic nematode strains and isolates is a common experimental approach. For example, a number of  
387 candidate anthelmintic resistance genes have been identified by differential expression analysis of  
388 drug-resistant and -susceptible isolates (Xu et al., 1998; Dicker et al., 2011; El-Abdellati et al., 2011;  
389 Williamson et al., 2011). In the case of *H. contortus*, we reasoned that the extremely high levels of  
390 sequence polymorphism both within and between laboratory strains and field isolates (reviewed in  
391 Gilleard and Redman (2016)), might confound the validity of such comparisons when using RNAseq,  
392 which is now the central approach to conducting differential gene expression analyses. The majority of  
393 researchers use only the default parameters of RNAseq data analysis pipelines and do not explore the  
394 effect of different parameters on results reported (Baruzzo et al., 2017). It has been shown, using  
395 simulated datasets, that the parameter with the greatest impact on performance is the number of

396 mismatches tolerated by during? read mapping (Baruzzo et al., 2017). Since this seemed likely to be a  
397 particular issue for organisms with high levels of sequence polymorphism, we undertook a detailed  
398 analysis to examine the extent to which this may impact RNAseq-based differential expression analysis  
399 between *H. contortus* strains, and investigated how it could be mitigated to allow genuine  
400 transcriptional differences to be assessed. We used *TopHat2* (Dobin and Gingeras, 2013) as our read  
401 mapping software as this has been the mapping program most commonly used for RNAseq analysis  
402 over a number of years and currently has the most citations in RNAseq literature. There are a number  
403 of alternative mapping tools available whose use is becoming increasingly common, such as *HISAT2*  
404 (Kim et al., 2015), which is the recommended successor of *TopHat2*, but these tools are similarly  
405 sensitive to changes in the mismatch parameter (Baruzzo et al., 2017).

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

419 were increased from N2 to N5, genes with higher levels of polymorphism showed larger proportionate  
420 increases in reads mapped for all three strains (Fig. 2Ca). This further illustrates the impact of sequence  
421 polymorphism on RNAseq read mapping efficiency and how it is greater for more polymorphic genes.

422 Having shown that sequence polymorphism affects RNAseq read mapping to a reference genome  
423 assembly with *TopHat2*, we next investigated how this might bias differential expression analysis using  
424 *DESeq2*, one of the most commonly used bioinformatic tools for RNAseq data analysis (Fig. 3 and Fig.  
425 4A). For each gene model, we plotted the *DESeq2* differential expression results against the difference  
426 in SNP rate (relative to the reference genome assembly) between the two strains being compared (Fig.  
427 3). For each pairwise strain comparison, gene models which had greater differences in the level of  
428 sequence polymorphism between the strains were more likely to be down-regulated than to be up-  
429 regulated in the strain with the highest level of sequence polymorphism (Fig. 3). Further, this bias  
430 increased with the magnitude of difference in polymorphism rate of gene models between the strains  
431 (Fig. 3 and Fig. 4A). This effect was true for all three pairwise strain comparisons, including between  
432 the two “non-reference” MHco4(WRS) and MHco10(CAVR) strains. There is no obvious biological  
433 reason for such differential expression biases, based on differences in SNP polymorphism rates, and so  
434 we concluded this is due to the effect of sequence polymorphism on RNAseq mapping rates.

435 Consequently, biases due to inter-strain differences in SNP polymorphism rates needed to be  
436 minimized before meaningful differential expression analysis could be performed. The first approach to  
437 achieve this was to choose RNAseq read mapping parameters in *TopHat2* to maximize read mapping  
438 efficiency for all the strains. Overall read mapping success peaked at the N5 or N6 SNP mapping  
439 allowances, depending on the strain (with very little difference between these two values (Fig. 1)). At  
440 the level of the gene model, the clear majority of genes had higher numbers of reads mapping at the N5  
441 allowance than at either the N2 or N10 allowances (Fig. 2A). Consequently, the N5 mapping allowance

442 maximized read mapping efficiency. Furthermore, the directional biases in the differential expression  
443 reports between strains were greatly reduced at the N5 mapping allowance (Fig. 4A-B, Supplementary  
444 Fig. S1). Consequently, the N5 mapping allowance was considered optimal to use for further analysis.  
445 However, optimizing the SNP mapping allowance did not completely remove the directional  
446 expression biases. For example, even at the N5 SNP mapping allowance, although the directional  
447 expression bias was close to zero for genes with SNP rate differences between strains of < 2%, it  
448 persisted for genes with a difference in SNP rate of > 2% (Fig. 4B). This led us to conclude that it was  
449 not possible to reliably measure differential expression for those genes with > 2% SNP rate differences  
450 between strains, even at the N5 read mapping allowance. Consequently, we precluded these genes from  
451 subsequent transcriptomic analysis. These results have important implications for differential  
452 expression analysis between different strains/isolates of organisms with high levels of genetic diversity  
453 and suggest that sequence polymorphism needs to be defined and accounted for as part of the analysis.  
454 There are a number of other read mapping tools available for RNAseq analysis, some of which,  
455 although less widely used than *TopHat2*, may be less impacted by high levels of sequence  
456 polymorphism (Baruzzo et al., 2017). *TopHat2* is still widely used but it is noteworthy that the  
457 mapping tool which is increasingly used in place *TopHat2* is *HISAT2*, which is only slightly less  
458 sensitive to changes in mismatch parameters using simulated datasets (Baruzzo et al., 2017). Other read  
459 mapping tools such as *NovoAlign* (<http://www.novocraft.com/products/novoalign/>) or *GSNAP* (Wu and  
460 Nacu, 2010), that may be less impacted by sequence polymorphism, deserve more exploration for use  
461 in RNAseq differential expression pipelines for organisms such as *H. contortus* with high levels of  
462 genetic variation.

463 Pairwise comparisons of three genetically divergent strains of *H. contortus* revealed large numbers  
464 of differentially expressed genes, even after the confounding effects of sequence polymorphism were

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

483 The results of this study also have important implications for anthelmintic resistance research  
484 which, until very recently, has been dominated by candidate gene studies (Gilleard, 2013, 2006;  
485 Rezansoff et al., 2016). In the case of ivermectin resistance, such studies have so far failed to identify  
486 the key loci or genes involved in resistance for any parasitic nematode, including *H. contortus*  
487 (Gilleard, 2013). One common component of candidate gene studies has been to compare the

488 expression levels of specific candidate genes between a small number of ivermectin-resistant and -  
489 susceptible parasite strains (Xu et al., 1998; Dicker et al., 2011; El-Abdellati et al., 2011; Williamson et  
490 al., 2011). It is common for such studies to report differences in expression between resistant and  
491 susceptible strains for candidate genes such as P-glycoproteins (PGPs) or ligand-gated ion channels  
492 (LGICs) [abbreviation previously used for ligand-gated chloride channels]. These differences are  
493 commonly used as circumstantial evidence for a role in resistance. Our results here show the context in  
494 which such studies should be interpreted as a very large number of genes are differentially expressed in  
495 pairwise comparisons of genetically divergent *H. contortus* strains (Fig. 5). LPGs (824 - 1,498) were  
496 differentially expressed between the strains in the study at a level of 2-fold and an adjusted statistical  
497 significance of  $P < 0.05$  (as called by *DESeq2*). This highlights the inherently high levels of  
498 “background” transcriptomic variation that occur between genetically divergent *H. contortus* strains.  
499 Consequently, care must be taken when interpreting a suggested association of differential expression  
500 of a gene with a drug resistance phenotype when a small number of genes are compared between a  
501 small number of drug-resistant and -susceptible strains. This is particularly the case when the degree of  
502 genetic differentiation or the general level of transcriptomic difference that exists between the strains  
503 has not been assessed.

504 Recently, studies analyzing the expression of small numbers of candidate genes are being replaced  
505 with more global transcriptomic studies. The draft *H. contortus* genome and its recent improvement  
506 into a chromosomal level assembly is making such studies increasingly feasible on a genome-wide  
507 scale (Laing et al., 2013; Doyle et al., 2018). The work presented here also has important implications  
508 for global transcriptomic comparisons of drug-resistant and -susceptible strains. Two gene families  
509 often suggested to be involved in ivermectin resistance are the LGICs and ABC transporter genes  
510 (Laing et al., 2013). We used the gene models in the *H. contortus* draft annotation to assess how many

511 members of these gene families were differentially expressed between the MHco4(WRS) and  
512 MHco10(CAVR) ivermectin-resistant strains and the MHco3(ISE) susceptible strain using the default  
513 polymorphism allowance (N2), the optimized polymorphism allowance (N5), and the polymorphism  
514 allowance (N5) but removing the highly polymorphic gene set (Table 1). We found there was a  
515 dramatic reduction in the number of members of these genes families that were determined to be  
516 differentially expressed when polymorphism allowance was increased to the optimal N5 allowance  
517 (Table 1). A further reduction was apparent when the most highly polymorphic genes were discarded  
518 from the analysis (Table 1).

519       These results highlight the fact that a substantial number of differentially expressed genes reported  
520 are likely to be artifacts caused by differences in sequence polymorphism between the strains being  
521 compared which are not accounted for. In the case of our analysis, accounting for sequence  
522 polymorphism reveals a smaller number of differentially expressed candidate genes perhaps worthy of  
523 further investigation. The ABC transporter *Hco-pmp-6*, and two LGICs – *Hco-lgc-55* and *Hco-lgc-44* –  
524 were differentially expressed with the same directionality in both ivermectin-resistant strains relative to  
525 the MHco3(ISE) strain. *Hco-lgc-55* is a tyramine-gated chloride channel whose *C. elegans* homologue,  
526 *Cel-lgc-55*, is expressed in the pharynx and is involved in worm motility (Ringstad et al., 200; Rao et  
527 al., 2010). The ABC transporter *Hco-wht-4*, and the LGICs *Hco-lgc-3*, *Hco-lgc-33*, *Hco-lgc-9*, and  
528 *Hco-acr-24*, were other genes with a > 2X log<sub>2</sub> fold-change differential expression in the  
529 MHco10(CAVR) strain, although these genes were not differentially expressed in the other resistant  
530 strain, MHco4(WRS). *Hco-lgc-3* was the gene with the highest level of up-regulation across both these  
531 gene families, being differentially expressed at greater than 50-fold in MHco3(CAVR) relative to  
532 MHco3(ISE) (Table 1). The gene may be considered of interest given its homology to a paralogous pair  
533 of *C. elegans* proton-gated ion channels, *Cel-pbo-5* and *Cel-pbo-6*, which are required for normal

534 posterior muscle function (Beg et al., 2008). However, further functional and genetic studies are  
535 required before making any inferences of the potential role of these genes in mediating the ivermectin  
536 resistance phenotype of *H. contortus*.

537

### 538 **Acknowledgements**

539 We are grateful to Dr Matt Workentine for comments on the manuscript. We are grateful for  
540 funding from NSERC [full name, country] Discovery Grant, NSERC-CREATE Host-Parasites  
541 Interactions (HPI) program, [country] and the Biotechnology and Biological Sciences Research  
542 Council (BBSRC), [country]. The Moredun Research Institute, United Kingdom (DJB and AAM)  
543 receives funding from the Scottish Government. NS was funded by the Higher Education Funding  
544 Council of England (HEFCE), the Department for Environment, Food and Rural Affairs (DEFRA),  
545 [country] and the Scottish Funding Council (SFC) Veterinary Training Research Initiative (VTRI)  
546 programme VT0102 and supported by Pfizer Animal Health, [country].

547

548

549 **References**

- 550 Anders, S., Pyl, P.T., Huber, W., 2014. HTSeq – A Python framework to work with high-throughput  
551 sequencing data HTSeq – A Python framework to work with high-throughput sequencing data.  
552 Bioinformatics. 31, 0–5. <https://doi.org/10.1093/bioinformatics/btu638>
- 553 Andersen, E.C., Gerke, J.P., Shapiro, J.A., Crissman, J.R., Ghosh, R., Bloom, J.S., Felix, M.-A.,  
554 Kruglyak, L., 2012. Chromosome-scale selective sweeps shape *Caenorhabditis elegans* genomic  
555 diversity. Nat. Genet. 44, 285–295. <https://doi.org/10.1038/ng.1050>
- 556 Antony, H.A., Pathak, V., Parija, S.C., Ghosh, K., Bhattacharjee, A., 2016. Transcriptomic Analysis of  
557 Chloroquine-Sensitive and Chloroquine-Resistant Strains of *Plasmodium falciparum* : Toward  
558 Malaria Diagnostics and Therapeutics for Global Health. Omi. A J. Integr. Biol. 20, 424–432.  
559 <https://doi.org/10.1089/omi.2016.0058>
- 560 Baruzzo, G., Hayer, K.E., Kim, E.J., DI Camillo, B., Fitzgerald, G.A., Grant, G.R., 2017. Simulation-  
561 based comprehensive benchmarking of RNA-seq aligners. Nat. Methods 14, 135–139.  
562 <https://doi.org/10.1038/nmeth.4106>
- 563 Bateman, A., Martin, M.J., O’Donovan, C., Magrane, M., Apweiler, R., Alpi, E., Antunes, R.,  
564 Arganiska, J., Bely, B., Bingley, M., Bonilla, C., Britto, R., Bursteinas, B., Chavali, G., Cibrian-  
565 Uhalte, E., Da Silva, A., De Giorgi, M., Dogan, T., Fazzini, F., Gane, P., Castro, L.G., Garmiri, P.,  
566 Hatton-Ellis, E., Hieta, R., Huntley, R., Legge, D., Liu, W., Luo, J., Macdougall, A., Mutowo, P.,  
567 Nightingale, A., Orchard, S., Pichler, K., Poggioli, D., Pundir, S., Pureza, L., Qi, G., Rosanoff, S.,  
568 Saidi, R., Sawford, T., Shypitsyna, A., Turner, E., Volynkin, V., Wardell, T., Watkins, X., Zellner,  
569 H., Cowley, A., Figueira, L., Li, W., McWilliam, H., Lopez, R., Xenarios, I., Bougueleret, L.,  
570 Bridge, A., Poux, S., Redaschi, N., Aimo, L., Argoud-Puy, G., Auchincloss, A., Axelsen, K.,  
571 Bansal, P., Baratin, D., Blatter, M.C., Boeckmann, B., Bolleman, J., Boutet, E., Breuza, L., Casal-  
572 Casas, C., De Castro, E., Coudert, E., Cucho, B., Doche, M., Dornevil, D., Duvaud, S., Estreicher,  
573 A., Famiglietti, L., Feuermann, M., Gasteiger, E., Gehant, S., Gerritsen, V., Gos, A., Gruaz-  
574 Gumowski, N., Hinz, U., Hulo, C., Jungo, F., Keller, G., Lara, V., Lemercier, P., Lieberherr, D.,  
575 Lombardot, T., Martin, X., Masson, P., Morgat, A., Neto, T., Noupikel, N., Paesano, S.,  
576 Pedruzzi, I., Pilbout, S., Pozzato, M., Pruess, M., Rivoire, C., Roechert, B., Schneider, M., Sigrist,  
577 C., Sonesson, K., Staehli, S., Stutz, A., Sundaram, S., Tognolli, M., Verbregue, L., Veuthey, A.L.,  
578 Wu, C.H., Arighi, C.N., Arminski, L., Chen, C., Chen, Y., Garavelli, J.S., Huang, H., Laiho, K.,

579 McGarvey, P., Natale, D.A., Suzek, B.E., Vinayaka, C.R., Wang, Q., Wang, Y., Yeh, L.S.,  
580 Yerramalla, M.S., Zhang, J., 2015. UniProt: A hub for protein information. *Nucleic Acids Res.* 43,  
581 D204–D212. <https://doi.org/10.1093/nar/gku989>

582 Beg, A.A., Ernstrom, G.G., Nix, P., Davis, M.W., Jorgensen, E.M., 2008. Protons Act as a Transmitter  
583 for Muscle Contraction in *C. elegans*. *Cell* 132, 149–160.  
584 <https://doi.org/10.1016/j.cell.2007.10.058>

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

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

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

593 Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., Ruden,  
594 D.M., 2012. A program for annotating and predicting the effects of single nucleotide  
595 polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain. *Fly (Austin)*. 6,  
596 80–92. <https://doi.org/10.4161/fly.19695>

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

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

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

606 Dicker, A.J., Nisbet, A.J., Skuce, P.J., 2011. Gene expression changes in a P-glycoprotein (Tci-pgp-9)  
607 putatively associated with ivermectin resistance in *Teladorsagia circumcincta*. *Int. J. Parasitol.* 41,  
608 935–942. <https://doi.org/10.1016/j.ijpara.2011.03.015>

609 Dobin, A., Gingeras, T.R., 2013. Comment on “TopHat2: accurate alignment of transcriptomes in the

610 presence of insertions, deletions and gene fusions” by Kim et al. 2013. bioRxiv 0–9.  
611 <https://doi.org/10.1101/000851> [move this reference into the text as it has is not a formally  
612 reviewed and published document.]

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

618 Doyle, S.R., Laing, R., Bartley, D.J., Britton, C., Chaudhry, U., Gilleard, J.S., Holroyd, N., Mable,  
619 B.K., Maitland, K., Morrison, A.A., Tait, A., Tracey, A., Berriman, M., Devaney, E., Cotton, J.A.,  
620 Sargison, N.D., 2018. A Genome Resequencing-Based Genetic Map Reveals the Recombination  
621 Landscape of an Outbred Parasitic Nematode in the Presence of Polyploidy and Polyandry.  
622 Genome Biol. Evol. 10, 396–409. <https://doi.org/10.1093/gbe/evx269>

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

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

630 Fiebig, M., Kelly, S., Gluenz, E., 2015. Comparative Life Cycle Transcriptomics Revises *Leishmania*  
631 *mexicana* Genome Annotation and Links a Chromosome Duplication with Parasitism of  
632 Vertebrates. PLoS Pathog. 11, 1–28. <https://doi.org/10.1371/journal.ppat.1005186>

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

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

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

639 Gilleard, J.S., Redman, E., 2016. Genetic Diversity and Population Structure of *Haemonchus contortus*,  
640 in: Advances in Parasitology. pp. 31–68. <https://doi.org/10.1016/bs.apar.2016.02.009> [If a book

641 chapter, include missing information; if a journal article, delete ‘in:’, abbreviate journal title and  
642 include volume.]

643 Guryev, V., Koudijs, M.J., Berezikov, E., Johnson, S.L., Plasterk, R.H.A., van Eeden, F.J.M., Cuppen,  
644 E., 2006. Genetic variation in the zebrafish. *Genome Res.* 16, 491–7.  
645 <https://doi.org/10.1101/gr.4791006>

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

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

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

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

660 Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin,  
661 R., Subgroup, 1000 Genome Project Data Processing, 2009. The Sequence Alignment/Map format  
662 and SAMtools. *Bioinformatics* 25, 2078–2079. <https://doi.org/10.1093/bioinformatics/btp352>

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

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

671 Love, M.I., Huber, W., Anders, S., 2014. Moderated estimation of fold change and dispersion for

672 RNA-seq data with DESeq2. *Genome Biol.* 15, 1–21. <https://doi.org/10.1186/s13059-014-0550-8>

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

676 Pathan, M., Keerthikumar, S., Ang, C.S., Gangoda, L., Quek, C.Y.J., Williamson, N.A., Mouradov, D.,  
677 Sieber, O.M., Simpson, R.J., Salim, A., Bacic, A., Hill, A.F., Stroud, D.A., Ryan, M.T., Agbinya,  
678 J.I., Mariadason, J.M., Burgess, A.W., Mathivanan, S., 2015. FunRich: An open access standalone  
679 functional enrichment and interaction network analysis tool. *Proteomics* 15, 2597–2601.  
680 <https://doi.org/10.1002/pmic.201400515>

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

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

697 Redman, E., Packard, E., Grillo, V., Smith, J., Jackson, F., Gilleard, J.S., 2008. Microsatellite analysis  
698 reveals marked genetic differentiation between *Haemonchus contortus* laboratory isolates and  
699 provides a rapid system of genetic fingerprinting. *Int. J. Parasitol.* 38, 111–22.  
700 <https://doi.org/10.1016/j.ijpara.2007.06.008>

701 Redman, E., Sargison, N., Whitelaw, F., Jackson, F., Morrison, A., Bartley, D.J., Gilleard, J.S., 2012.  
702 Introgression of Ivermectin Resistance Genes into a Susceptible *Haemonchus contortus* Strain by

703 Multiple Backcrossing. PLoS Pathog. 8, e1002534. <https://doi.org/10.1371/journal.ppat.1002534>

704 Redman, E., Whitelaw, F., Tait, A., Burgess, C., Bartley, Y., Skuce, P.J., Jackson, F., Gilleard, J.S.,  
705 2015. The emergence of resistance to the benzimidazole anthelmintics in parasitic nematodes of  
706 livestock is characterised by multiple independent hard and soft selective sweeps. PLoS Negl.  
707 Trop. Dis. 9, e0003494. <https://doi.org/10.1371/journal.pntd.0003494>

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

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

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

719 Sargison, N.D., Redman, E., Morrison, A.A., Bartley, D.J., Jackson, F., Hoberg, E., Gilleard, J.S.,  
720 2019. Mating barriers between genetically divergent strains of the parasitic nematode  
721 *Haemonchus contortus* suggest incipient speciation. Int. J. Parasitol.  
722 <https://doi.org/10.1016/j.ijpara.2019.02.008>

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

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

730 Wang, D.G., 1998. Large-Scale Identification, Mapping, and Genotyping of Single-Nucleotide  
731 Polymorphisms in the Human Genome. Science (80-. ). 280, 1077–1082.  
732 <https://doi.org/10.1126/science.280.5366.1077>

733 Wang, Z., Gerstein, M., Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nat.

734 Rev. Genet. 10, 57–63. <https://doi.org/10.1038/nrg2484>

735 Williamson, S.M., Storey, B., Howell, S., Harper, K.M., Kaplan, R.M., Wolstenholme, A.J., 2011.

736 Candidate anthelmintic resistance-associated gene expression and sequence polymorphisms in a

737 triple-resistant field isolate of *Haemonchus contortus*. Mol. Biochem. Parasitol. 180, 99–105.

738 <https://doi.org/10.1016/j.molbiopara.2011.09.003>

739 Wu, T.D., Nacu, S., 2010. Fast and SNP-tolerant detection of complex variants and splicing in short

740 reads. Bioinformatics 26, 873–881. <https://doi.org/10.1093/bioinformatics/btq057>

741 Xu, M., Molento, M., Blackhall, W., Ribeiro, P., Beech, R., Prichard, R., 1998. Ivermectin resistance in

742 nematodes may be caused by alteration of P-glycoprotein homolog. Mol. Biochem. Parasitol. 91,

743 327–335. [https://doi.org/10.1016/S0166-6851\(97\)00215-6](https://doi.org/10.1016/S0166-6851(97)00215-6)

744

745

746 **Figure Legends**

747

748 [Each figure and its legend, and each supplementary figure and its legend, must ‘stand alone’ ie each  
749 one should be able to be understood without reading the rest of the paper, therefore please ensure  
750 sufficient information is included ege parasite genus/species (preferably in the first summary sentence)  
751 if data refer to a particular parasite. Also include definitions of abbreviations within the figure legend  
752 or at the end of the figure legend for any additional abbreviations used in the figure and insert  
753 descriptive words if appropriate eg if a term is the name of a parasite strain as IJP is a general journal  
754 and papers should be written for non-expert readers.

755 Also check usage of consistent font style and size within each figure eg X and Y axis labels and  
756 legends should match in font style and size. The same font style should be used in all figures.]

757

758 **Fig. 1.** The percentage of RNAseq reads that mapped to the MHco3(ISE) [define/describe] reference  
759 genome assembly at different *TopHat2* SNP define] (polymorphism) allowances (N2 to N10) shown  
760 for each of the three *H. contortus* strains MHco3(ISE), MHco4(WRS), and MHco10(CAVR).

761

762 **Fig. 2.** [Insert a sentence summarizing the figure before describing the panels.] A) The number of genes  
763 which had either a >1% increase (green bars) or >1% decrease (red bars) in the number of RNAseq  
764 reads mapping to them on the reference MHco3(ISE) [define/describe] genome assembly following an  
765 increase in the read mapping polymorphism allowance in *TopHat2* for *H. contortus* strains  
766 MHco3(ISE), MHco4(WRS), and MHco10(CAVR). (a) The data for a change in polymorphism  
767 allowance of N2 to N5 and (b) the data for a change from N5 to N10 are shown. B) The number of  
768 gene models in each SNP [define] rate category for each *H. contortus* strain. The SNP rate for each

769 gene model was calculated by dividing the number of SNPs in each CDS [define] by the total CDS  
770 length for each gene model. C) Ratios of the total number of RNAseq reads mapping to gene models in  
771 each SNP rate category at two different SNP mapping allowances for each *H. contortus* strain. (a) the  
772 N5:N2 ratio and (b) the N10:N5 ratio are shown. Counts of reads mapped were totaled for all genes  
773 within each SNP rate category of each strain (colour coded).

774

775 **Fig. 3.** Scatter plots of the differential expression of gene models, as determined by *DESeq2* (X-axis),  
776 plotted against their difference in SNP [define] rate percentage between the two strains being compared  
777 (Y-axis). Gene model data points in each pairwise comparison are split in two with the left half of each  
778 panel showing the gene models with higher SNP rates in one strain of each pairwise comparison and  
779 the right half of each panel showing the gene models with higher SNP rates in the other pairwise strain.  
780 A and B show the MHco4(WRS) versus MHco3(ISE) comparison, C and D show the MHco10(CAVR)  
781 versus MHco3(ISE) comparison, and E and F show the MHco4(WRS) versus MHco10(CAVR)  
782 comparison. The difference in the SNP rate percentage between the two strains is shown on the Y-axis  
783 and plotted against reported log<sub>2</sub> fold-change differential expression for each gene. The red lines  
784 represent zero differential expression.

785

786 **Fig. 4.** [Insert a sentence summarizing the figure before describing the panels.] A) The percentage of  
787 expressed gene models in each SNP [define] rate difference category that are differentially expressed  
788 between MHco3(ISE) [define/describe] and MHco4(WRS) (log<sub>2</sub> fold-change > 1X; adjusted *P* value <  
789 0.05) for each of the three SNP (polymorphism) allowances – N2, N5, and N10 – when mapping. B)  
790 The net log<sub>2</sub> fold differences in expression (NDE) of all expressed genes in each SNP rate difference  
791 category. NDEs are shown for the N2, N5 and N10 SNP allowances when read mapping for the

792 MHco3(ISE) versus MHco4(WRS) pairwise comparison. NDEs are the mean values for all genes in  
793 each SNP rate difference category. Negative NDE values indicate an overall bias towards down-  
794 regulation of genes in the MHco4(WRS) versus MHco3(ISE) strain. Positive values report an overall  
795 bias towards up-regulation of genes.

796

797 **Fig. 5.** The total number of differentially expressed low polymorphic genes (LPGs) observed in each  
798 pairwise strain comparison at the N5 mapping allowance. Gene counts at both  $> 1X \log_2$  fold-change  
799 (orange dots), and  $> 2X \log_2$  fold-change (red dots) thresholds are shown. The blue line on the Y-axis  
800 represents an adjusted  $P$  value of 0.05.

801

802

### 803 **Supplementary Figure Legends**

804

805 **Supplementary Fig. S1.** Volcano plots showing differential expression of gene models at three different  
806 SNP [define] allowances in mapping parameters of *Tophap2* (N2, N5, N10) are shown for each pairwise  
807 strain comparison. The  $\log_2$  fold-change difference in expression from -4 to 4 is represented along the  
808 X-axis of each chart, and *DESeq2*  $-\log_{10}$  adjusted  $P$  values of the differential expression calls from 0 to  
809 30 are represented along the Y axis. Gene positions exceeding a maximum value on either axis are placed  
810 at maximum value on that axis. Red points on the right and left sides of each plot represent genes  
811 differentially expressed at  $> 1X$  and  $< -1X \log_2$  fold-change, respectively, with adjusted  $P$  values  $< 0.05$ .  
812 Blue points represent genes significantly differentially expressed but at less than  $1X \log_2$  fold-change in  
813 either direction.

814

815 **Supplementary Fig. S2.** Venn diagram showing the numbers of gene models qualifying as low  
816 polymorphic genes to be included in the different pairwise strain comparisons. The total number of genes  
817 qualifying as low polymorphic genes in each of the pairwise strain comparisons are shown outside  
818 respective circles (i.e. gene models with differences in SNP [define] rates between the two strains of <  
819 2%). The numbers of these genes shared and not shared among the pairwise strain comparisons are shown  
820 within respective Venn circles.

821

822 **Supplementary Fig. S3.** A PCA [define] plot representing the variance in log gene expression of low  
823 polymorphic genes of each triplicate dataset for each of the three populations when mapped at the N5  
824 mapping allowance.

825

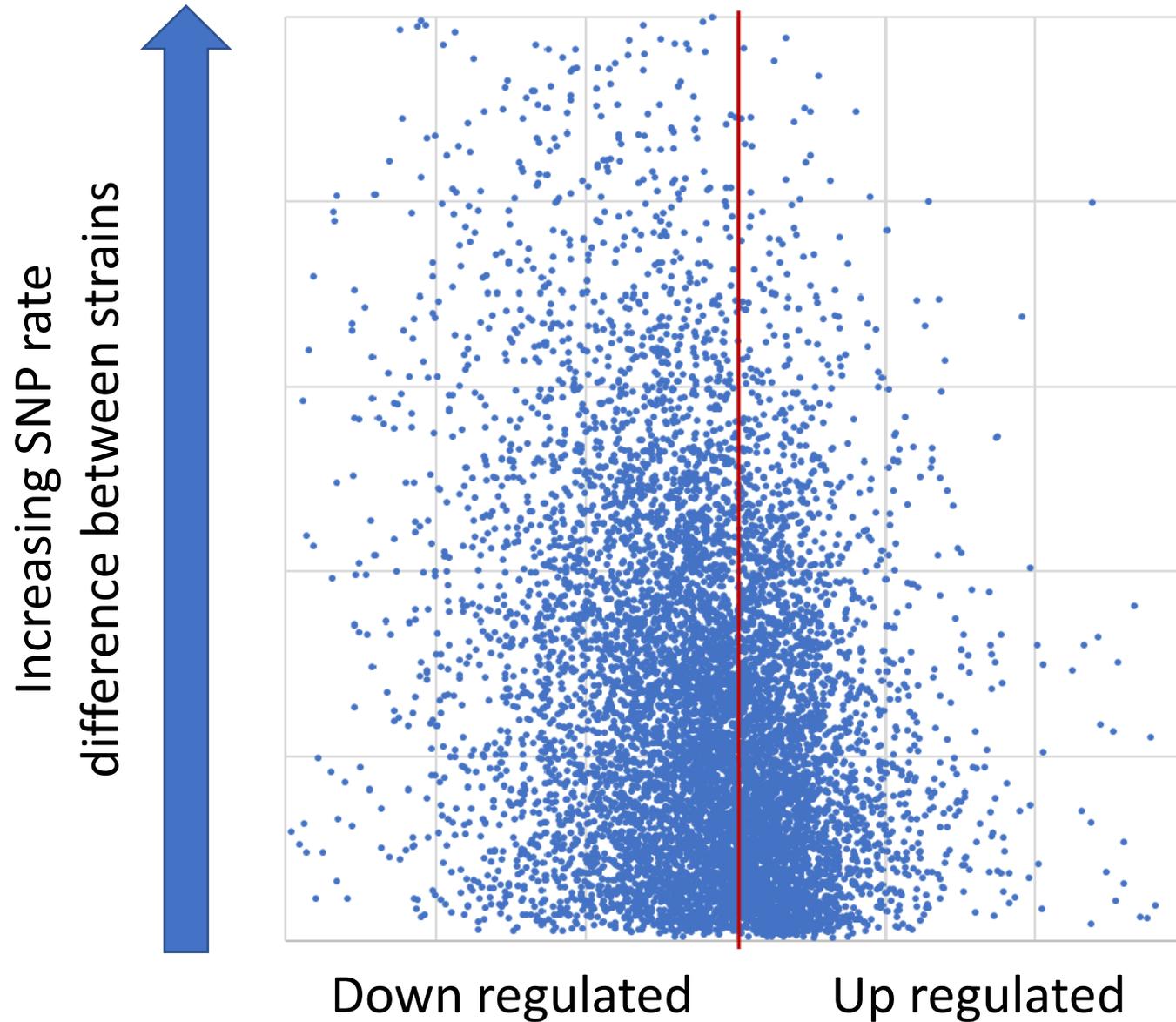
826 **Supplementary Fig. S4.** Venn diagrams showing the numbers of genes differentially expressed in each  
827 pairwise strain comparison, and shared differentially expressed genes between different pairwise strain  
828 comparisons. Venn circles are colour coded by pairwise strain comparison – red represents differentially  
829 expressed gene numbers of the MHco4(WRS) versus MHco3(ISE) comparison, orange represents the  
830 MHco10(CAVR) versus MHco3(ISE) comparison, and green represents the MHco4(WRS) versus  
831 MHco10(CAVR) comparison. Differentially expressed genes were counted and cross-referenced at two  
832 thresholds of differential expression:  $\log_2$  fold-change difference in expression  $> 1$  (*italic*), and  $\log_2$  fold-  
833 change difference in expression  $> 2$  (**bold**).

834

835

836

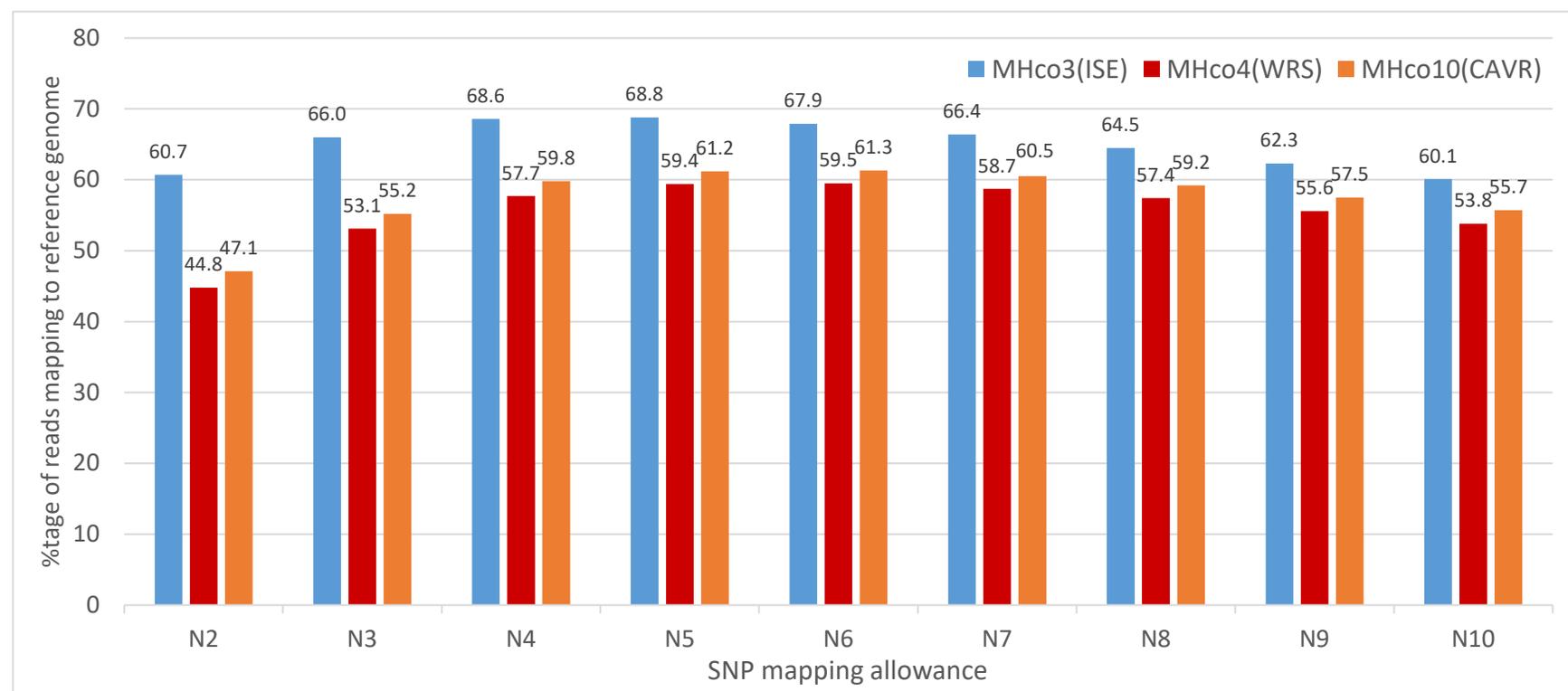
# Artificial bias in differential expression calls due to sequence polymorphism



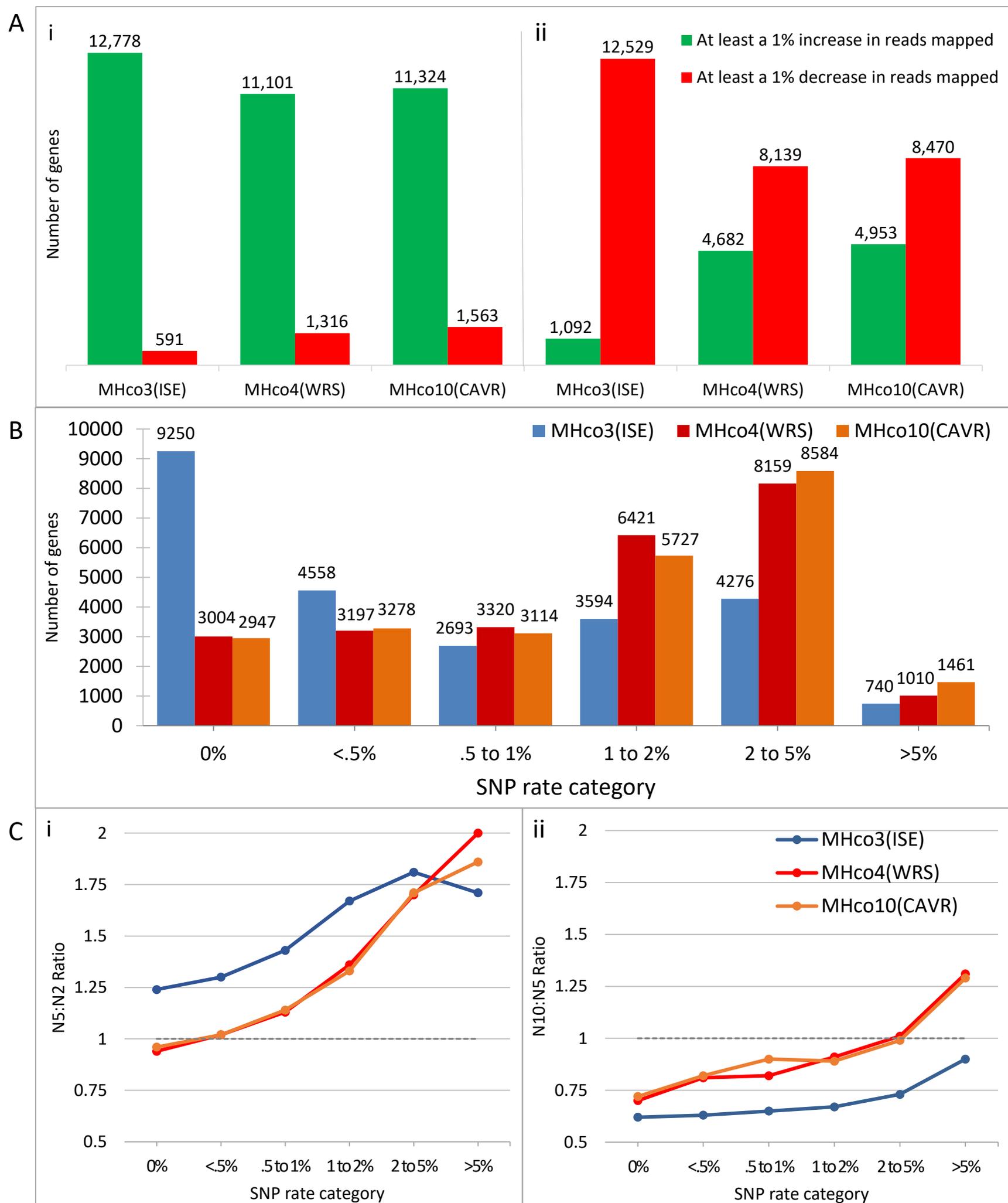
Paper Highlights

- Sequence polymorphism can confound RNAseq analysis in genetically diverse organisms due to read mapping biases
- Optimizing read mapping allowances and excluding highly polymorphic genes reduces differential gene expression analysis biases
- Genetically divergent strains of *H. contortus* have very high levels of inter-strain transcriptional diversity
- Interpretation of inter-strain differential gene expression needs to consider sequence polymorphism and overall transcriptional diversity

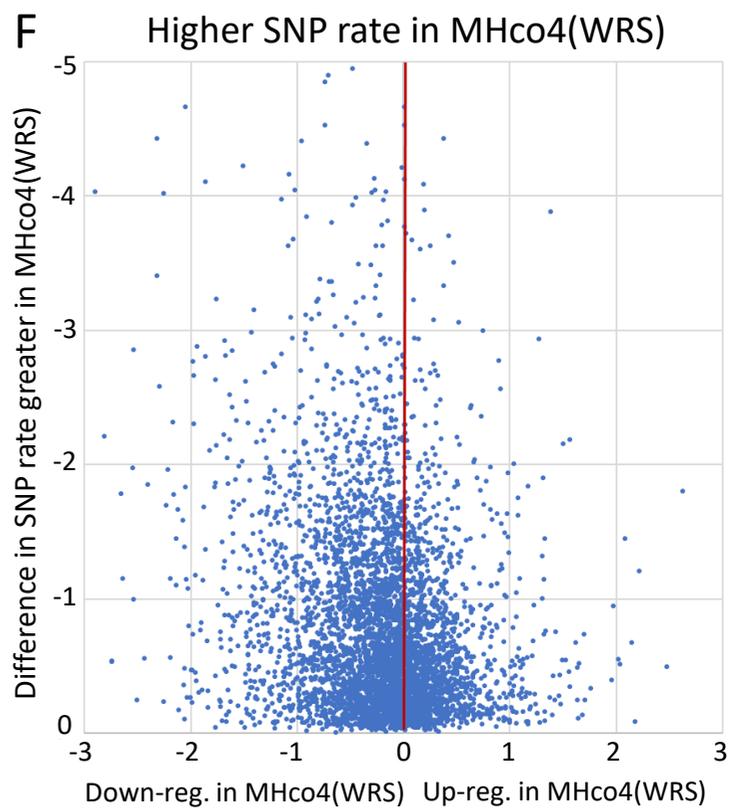
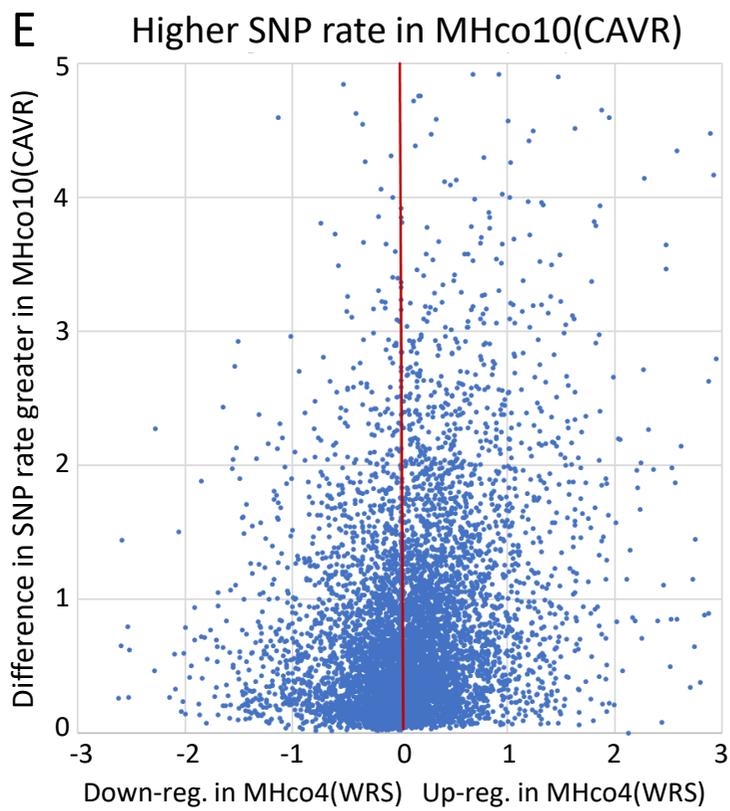
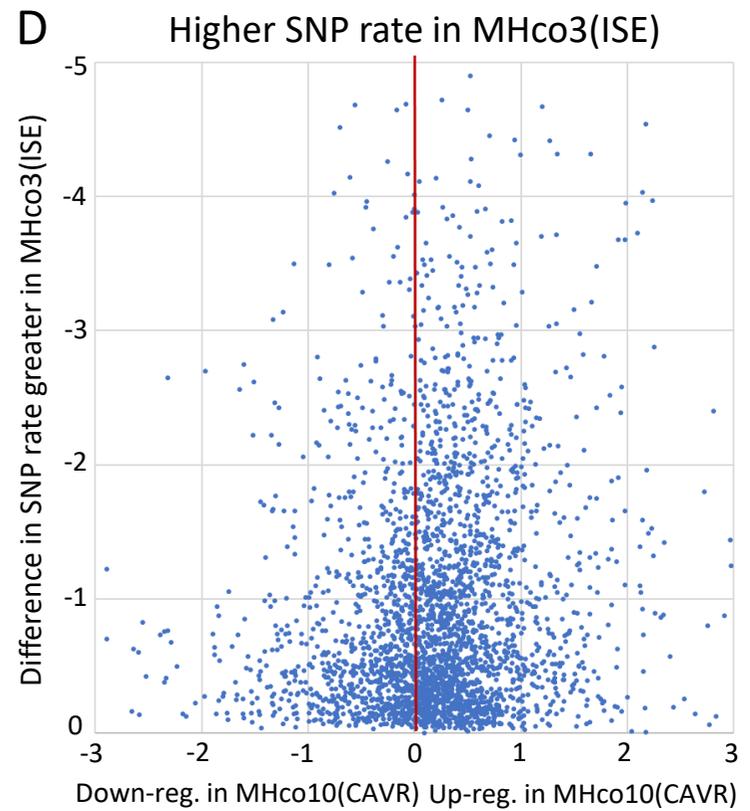
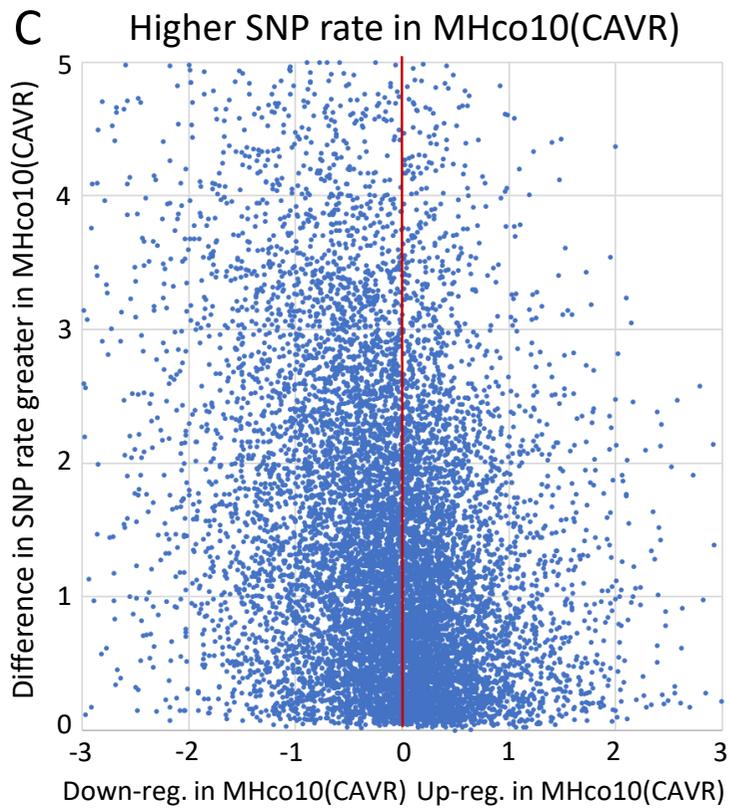
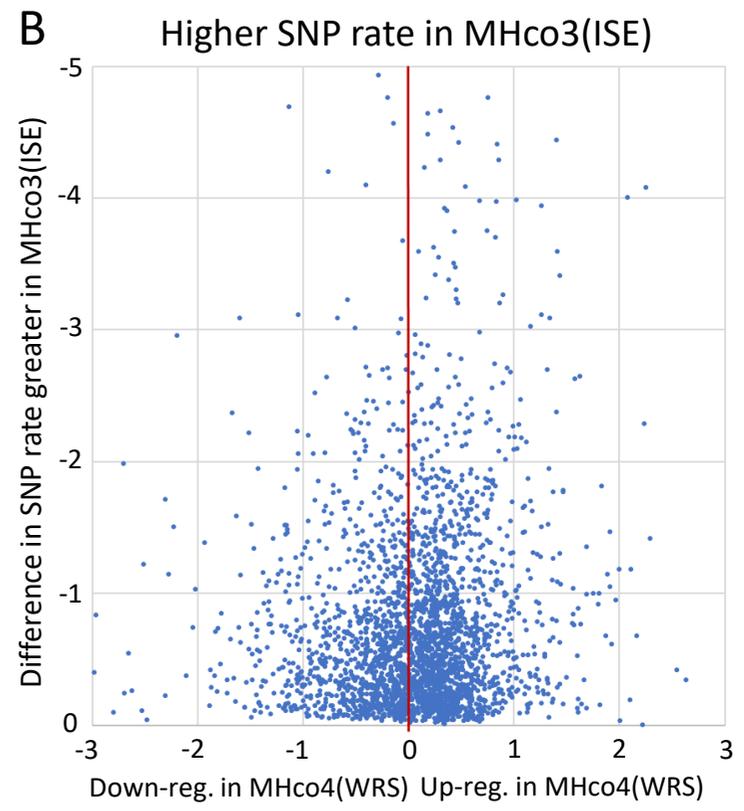
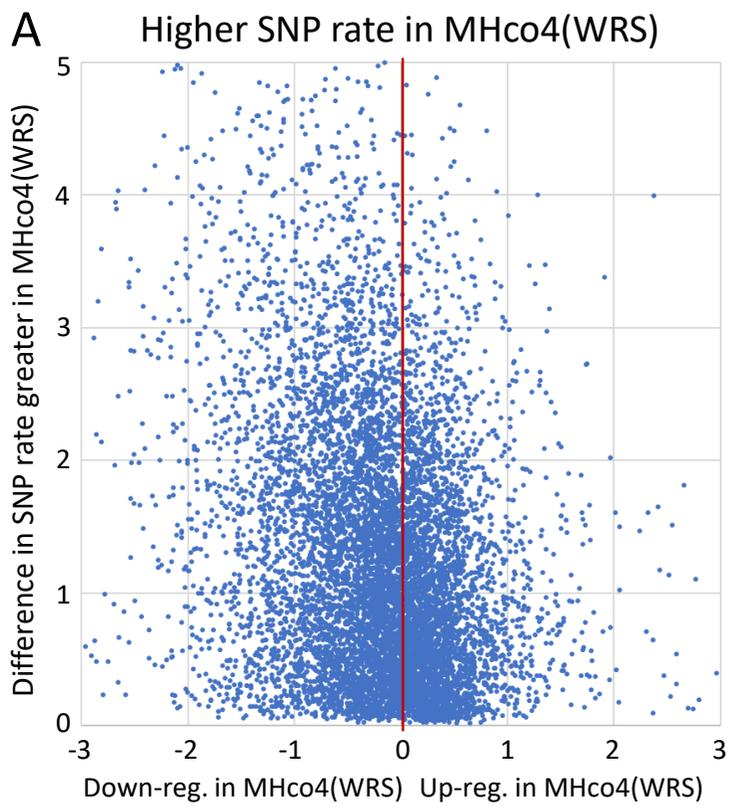
Figure 1.



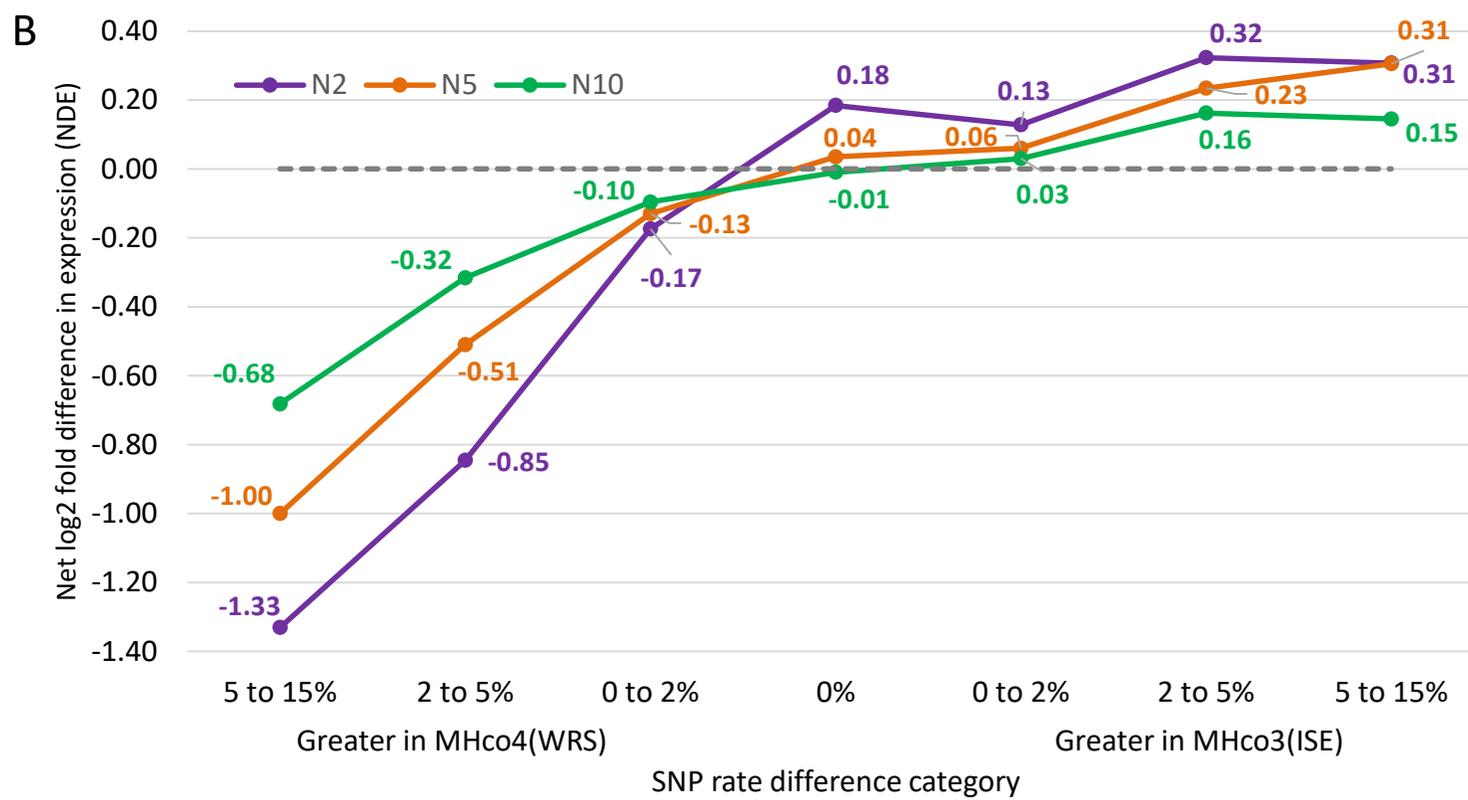
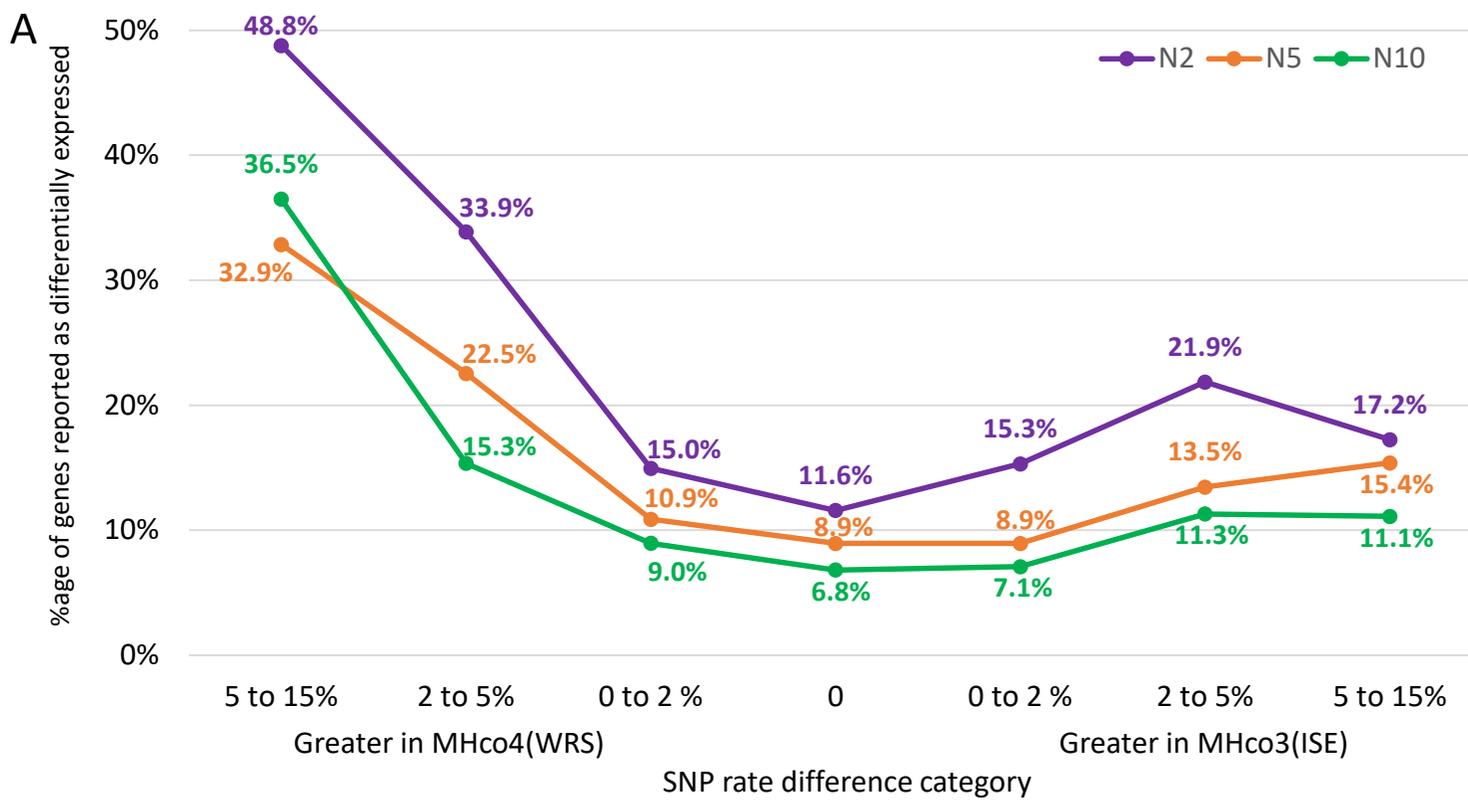
## Figure 2



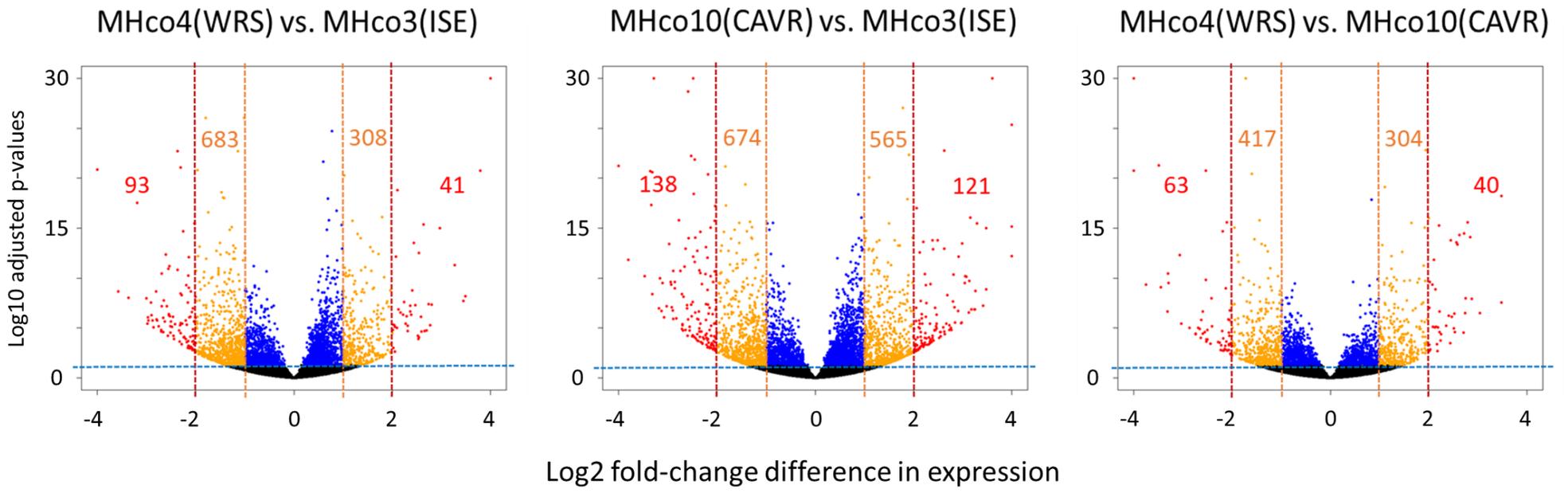
# Figure 3



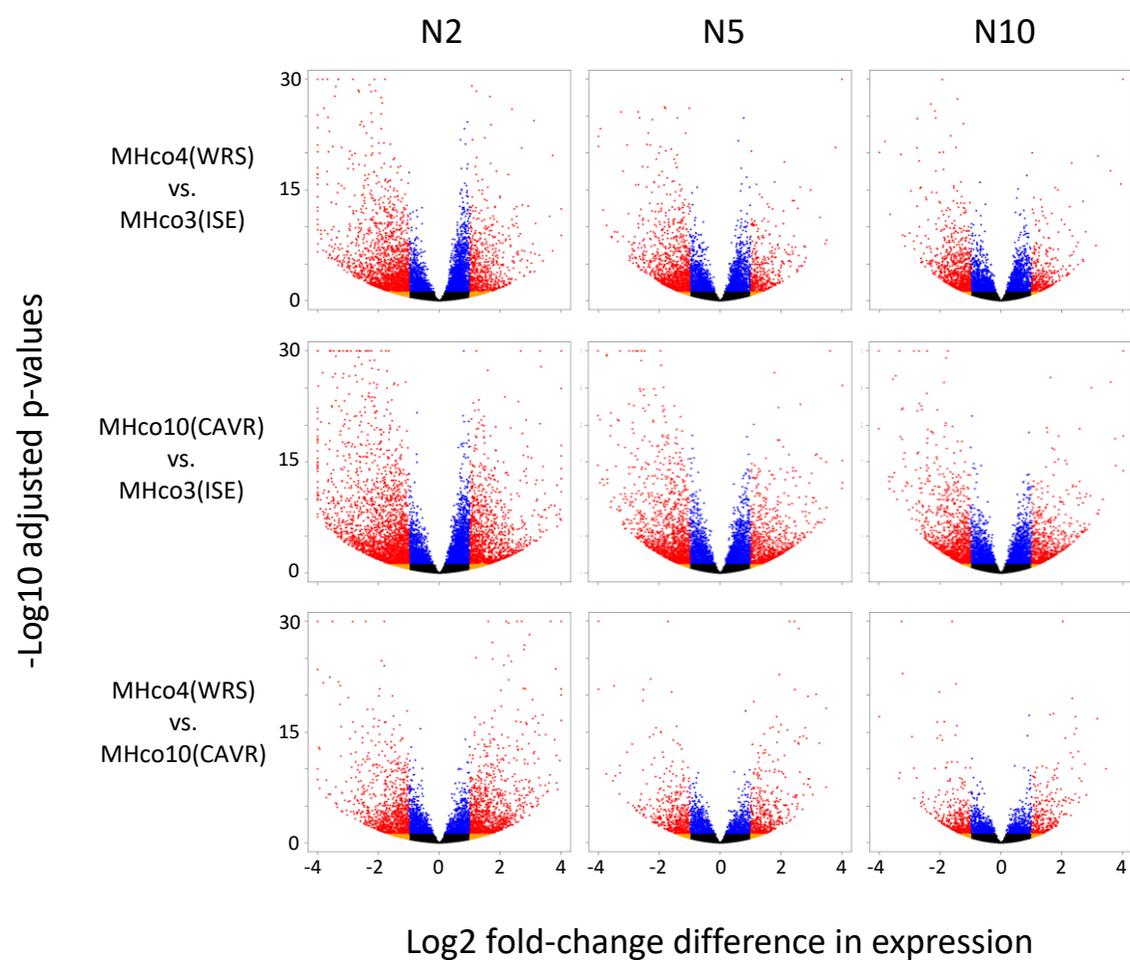
# Figure 4



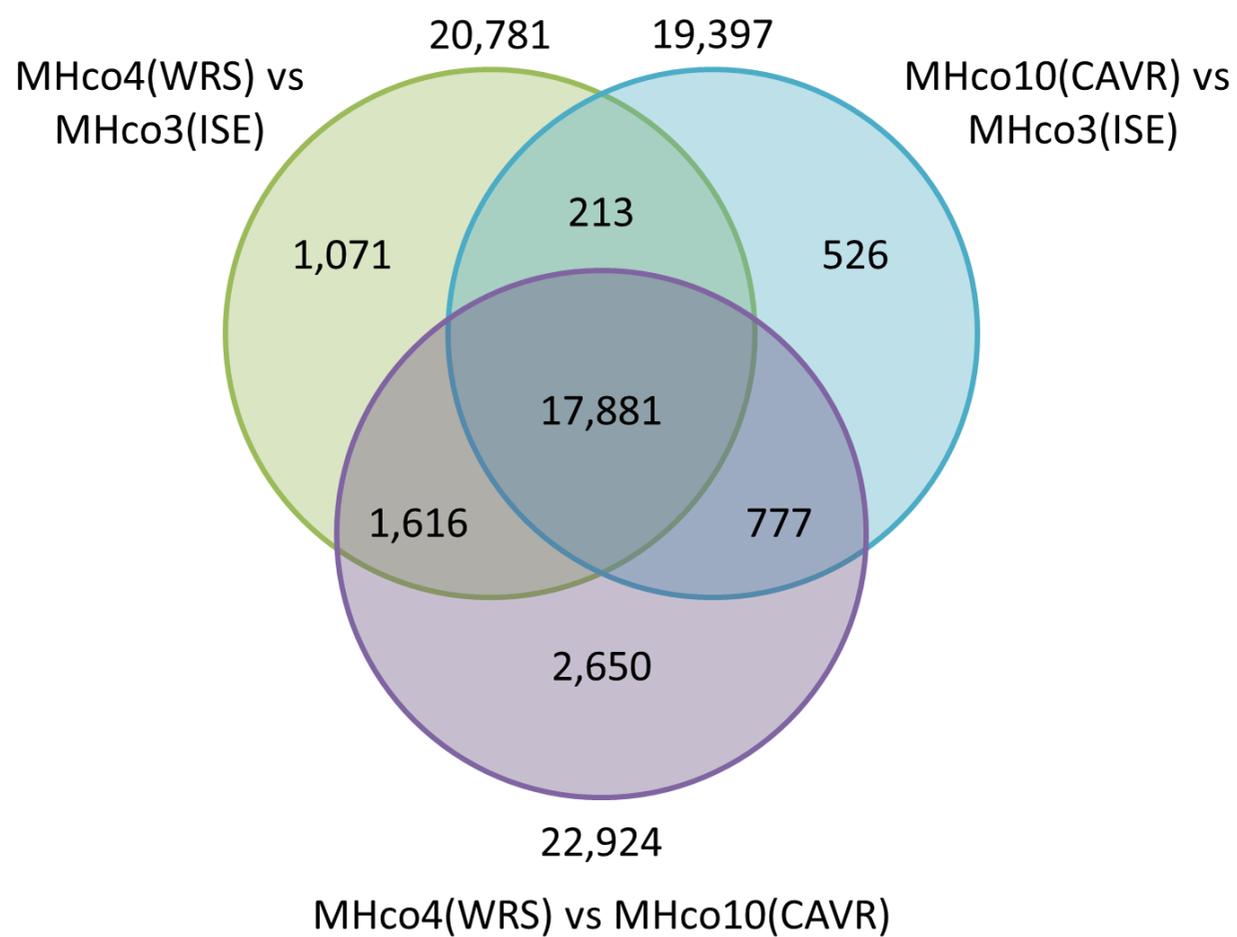
# Figure 5



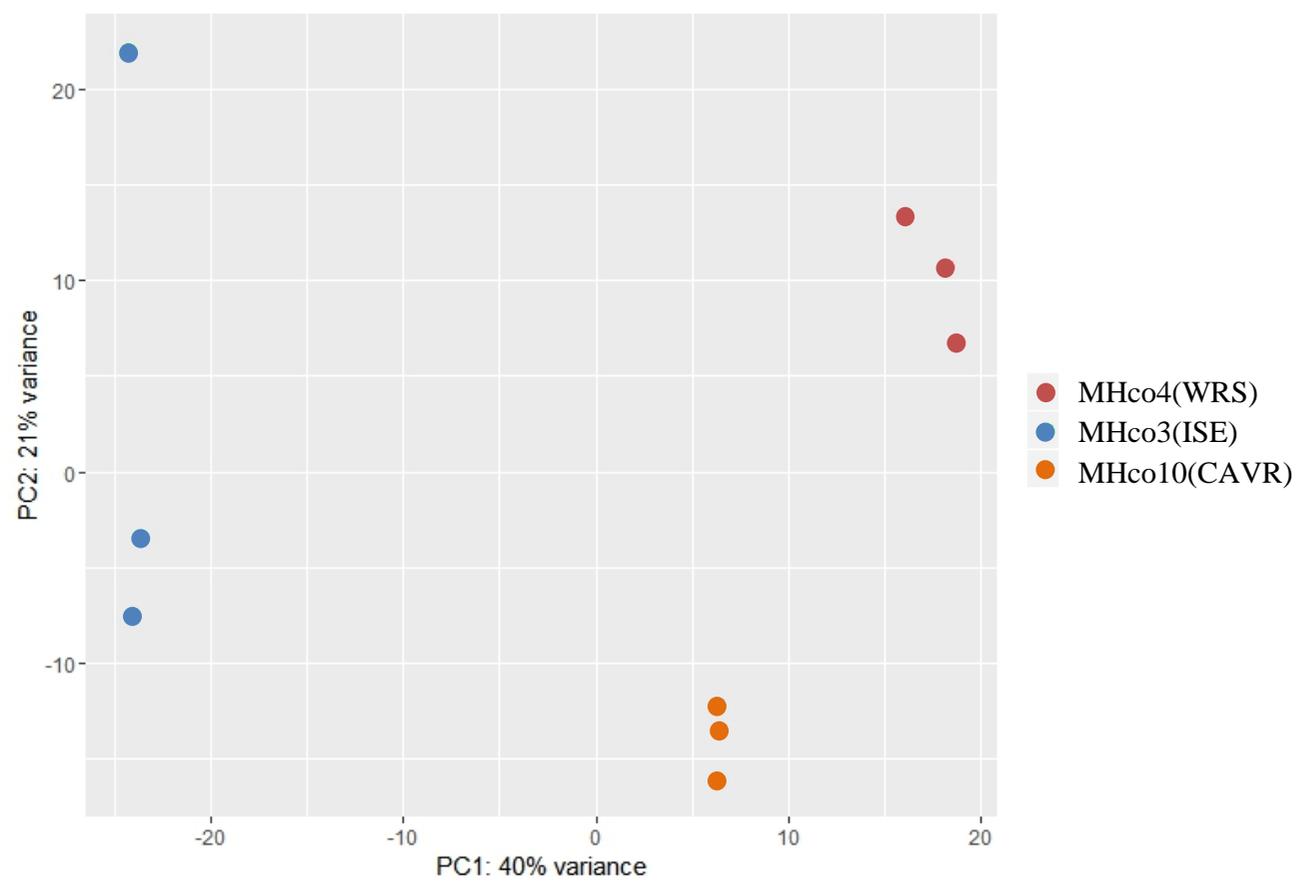
# Supplementary Figure S1



# Supplementary Figure S2

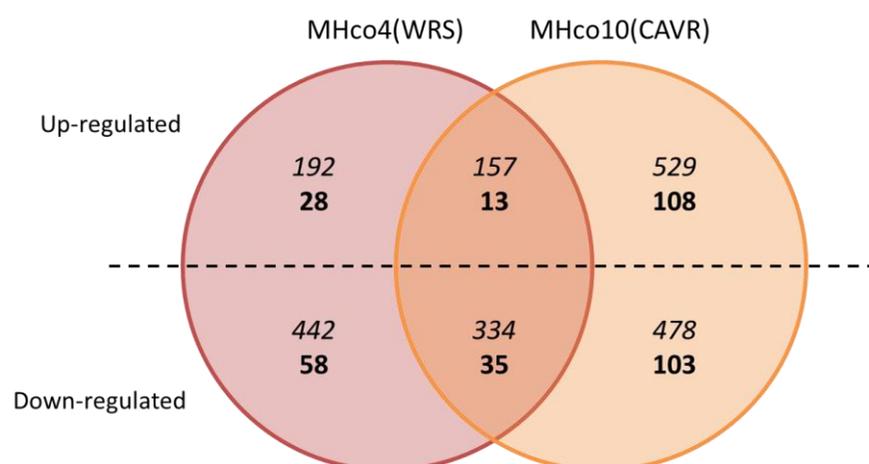


# Supplementary Figure S3

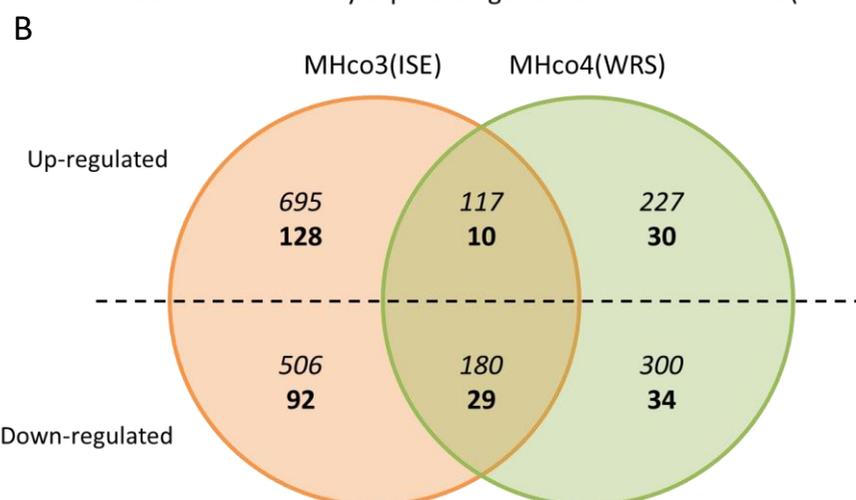


# Supplementary Figure S4

A Number of differentially expressed genes relative to MHco3(ISE)



Number of differentially expressed genes relative to MHco10(CAVR)



C Number of differentially expressed genes relative to MHco4(WRS)

