1 **Cover Page**

2

3 **Title:** Comparison and Characterisation of Mutation Calling from Whole Exome and
4 RNA Sequencing Data for Liver and Muscle Tissue in Lactating Holstein Cows
5 Divergent for Fertility

6

7 **Authors:** Bruce Moran^{*,†}, Stephen T. Butler[‡], Christopher J. Creevey^{†,§}

8

9 **Authors Institutional Affiliations**

10 * Conway Institute of Biomolecular and Biomedical Research, University College
11 Dublin, Belfield, D04 V1W8, Ireland.

12 † Teagasc, Animal and Grassland Research and Innovation Centre, Grange, Meath,
13 C15 PW93, Ireland.

14 ‡ Teagasc, Animal and Grassland Research and Innovation Centre, Moorepark, Cork,
15 P61 C997, Ireland.

16 § Institute of Biological, Environmental and Rural Sciences, Aberystwyth University,
17 SY23 3FL, Wales.

18

19 **Data Accessions:** GSE62159, GSE62160

20

21 **Running Title:** SNVs in WES vs RNASeq in Dairy Cows

22

23 **Summary:** The exome and transcriptome both relate to the same protein-coding
 24 regions of the genome. There has been sparse research on characterising mutations
 25 in RNA and DNA within the same individuals. Here we characterise the similarities in
 26 our Holstein dairy cow animal model. We offer practical and biological results
 27 indicating that RNA sequencing is a useful proxy of exome sequencing, itself shown
 28 to be applicable to this livestock species using a previously untested commercial
 29 application. This potentially unlocks public RNA sequencing data for further analysis,
 30 also indicating that RNA-DNA differences may associate with transcriptomic
 31 divergence.

32

33 **Key Words:** transcriptome, DNA, SNV calling, bovine

34

35 **Corresponding Author**

36 Christopher J. Creevey,

37 F.15, Carwyn James Buidling, Institute of Biological, Environmental and Rural

38 Sciences, Aberystwyth University, SY23 3FL, Wales.

39 Phone: 0044 (0) 1970 621612

40 Email: chc30@aber.ac.uk

41

42 **Abstract**

43 Whole exome sequencing has had low uptake in livestock species, despite allowing
 44 accurate analysis of single nucleotide variant (SNV) mutations. Transcriptomic data
 45 in the form of RNA sequencing has been generated for many livestock species and
 46 also represents a source of mutational information. However, there is little
 47 information on the accuracy of using this data for the identification of SNVs. We
 48 generated a bovine exome capture design and used it to sequence and call
 49 mutations from a lactating dairy cow model genetically divergent for fertility (Fert+,
 50 n=8; Fert-, n=8). We compared mutations called from liver and muscle
 51 transcriptomes from the same animals. Our exome capture demonstrated 99.1%
 52 coverage of the exome design of 56.7MB, whereas transcriptomes covered 55 and
 53 46.5% of the exome, or 24.4 and 20.7MB, in liver and muscle respectively after
 54 filtering. We found that specificity of SNVs in the transcriptome data is
 55 approximately 75% following basic hard-filtering, and could be increased to above
 56 80% by increasing the minimum threshold of reads covering SNVs, but this effect
 57 was negated in more highly covered SNVs. RNA-DNA differences, SNVs found in
 58 transcriptome but not exome, were discovered and shown to have significantly
 59 increased levels of transition mutations in both tissues. Functional annotation of
 60 non-synonymous SNVs specific to the high and low fertility phenotypes identified
 61 immune response-related genes, supporting previous work that has identified
 62 differential expression in the same genes. Publically available RNAseq data may be
 63 analysed in a similar way to further increase the utility of this resource.

64

65 **Introduction**

66 Whole exome sequencing (WES) is a method used to capture DNA that represents
 67 the protein-coding component of a genome. The approach, while only sequencing a
 68 small percentage of a vertebrate genome (typically 1-2%), covers those areas where
 69 a missense or nonsense mutation would have a higher chance of causing a functional
 70 change in the gene product and has most commonly been applied in biomedical
 71 studies (Warr *et al.* 2015). Investigating the protein-coding regions by messenger
 72 RNA sequencing (RNAseq) has been applied across many vertebrate species and has
 73 been widely adopted in livestock species. The specific utility of this kind of research
 74 is in the study of gene expression and alternative transcription but there is also
 75 potential use in variant calling (McGettigan 2013). In livestock species, WES has been
 76 rarely employed, and at the time of writing there were only three bovine exome
 77 studies published (Cosart *et al.* 2011; Hirano *et al.* 2013; McClure *et al.* 2014) despite
 78 there being an evident utility in causative mutation discovery and in the analysis of
 79 ‘extreme phenotypes’ (Emond *et al.* 2012).

80 The difference in primary technical objectives of WES and RNAseq are quite
 81 basic; respectively, the approaches attempt to represent all regions at roughly equal
 82 depth and coverage, and represent the level of expression of transcripts. RNAseq
 83 inherently results in a dataset of unequal depth and coverage which may also be
 84 missing many genes altogether due to not being expressed at the time of sampling. If
 85 RNAseq is to be used for mutation analysis, a key outstanding issue to be resolved is
 86 determining how WES and RNAseq overlap in their abilities to accurately detect
 87 mutations. Several studies using human data have shown that transcriptomic data is
 88 a relevant source of information for variant calling and discovery (Cirulli *et al.* 2010;

89 Quinn *et al.* 2013). However, no information is available, neither specifically for
 90 bovine nor any other livestock species, as to how to most accurately detect SNV
 91 from RNAseq data. Further to this, there has been some controversy surrounding
 92 ‘RNA-DNA differences’ (RDD), the proposed result of ‘RNA editing’ (Li *et al.* 2011).
 93 Several comments published in response to Li and colleagues indicated that RDD are
 94 likely due to issues inherent to RNA sequencing and alignment methods (Pickrell *et*
 95 *al.* 2012; Kleinman and Majewski 2012; Lin *et al.* 2012). With both DNA- and RNA-
 96 derived sequence data for the same individuals it could be possible to determine
 97 instances of RDD, adding an extra dimension to RNAseq data via WES.

98 Here we present an analysis of WES of 16 Holstein-Friesian dairy cows. We
 99 have previously published RNAseq analysis (Moran *et al.* 2016) which consisted of 96
 100 transcriptomes, six transcriptomes per animal broken into two tissue types (liver,
 101 muscle) at three time-points relevant to fertility. Multiple studies have
 102 demonstrated that this model of dairy cow fertility has significant differences in
 103 reproductive and metabolic performance (Cummins, Lonergan, Evans, Berry, *et al.*
 104 2012; Cummins, Lonergan, Evans and Butler 2012; Cummins, Waters, *et al.* 2012;
 105 Butler 2014; Moore, Fair, *et al.* 2014; Moore, Scully, *et al.* 2014). We compared WES
 106 and RNAseq derived SNV from this cow model to identify the factors affecting
 107 reliability of variant calls in each, especially due to technical differences including
 108 uniformity and depth/coverage from each approach. We assessed this
 109 independently across the two tissues and also investigated the underlying causes of
 110 any RDD observed between the two types of data. We carried out multi-dimensional
 111 scaling of the population structure using both WES and RNAseq data, and
 112 functionally annotated SNVs from both transcriptome sets.

113

114 **Method Overview**

115 *Exome Capture Design and DNA Extraction*

116 Full details of WES data generation, including DNA extraction and exome capture
117 design, are available as Supplementary Material. Briefly, we used the Roche
118 Nimblegen SeqCap EZ Developer kit and to our knowledge this is the first application
119 of the technology to the bovine genome. Genomic DNA was isolated from the same
120 liver samples acquired and used previously in our RNAseq analysis (see Moran et al
121 2016 for full details of biopsy).

122

123 *Bioinformatic Analysis*

124 Raw fastq files were aligned to the UMD3.1 bovine genome, duplicated reads were
125 removed and BAM files were processed using the Genome Analysis ToolKit (GATK
126 v3.1; McKenna *et al.* 2010; DePristo *et al.* 2011) based on 'best practice' guidelines
127 (Van der Auwera *et al.* 2013). We then investigated the coverage of exome capture
128 and transcriptome (all exons summed) to determine how well the data represented
129 each. Systematic errors were removed using the Syscall tool (Meacham *et al.* 2011)
130 and all non-biallelic SNV were removed. Hard-filters were then applied to the variant
131 calls, requiring 7 reads supporting the alternate allele, and finally SNV not passing
132 the GATK filters were removed. To determine how these filters impacted the
133 transcriptome data we used absolute expression in the form of fragments per
134 kilobase per million reads (FPKM; Mortazavi *et al.* 2008) to plot expression level
135 versus number of reads.

136 We used multi-dimensional scaling to visualize proximity of cows and their
137 shared sireages based on the SNV from WES and transcriptomes. Specificity analysis
138 was employed to determine if we could increase the level of true positives (TP, SNVs
139 found in WES and transcriptome) or reduce the level of false positives (FP, SNVs
140 found in transcriptome but not in WES). For those SNV denoted as FP, we classified
141 their occurrence based on several potentially sequencing-related factors. Following
142 removal based on these classifications, we labelled the resulting SNVs as RNA-DNA
143 differences (RDD). Having created TP, false negative (FN, SNV in WES but not
144 transcriptome) and RDD SNV sets, we then investigated the mutational profiles and
145 transition/transversion ratios of each in both tissues and WES, conducting chi-
146 squared tests to determine if a significant difference in transitions/transversion ratio
147 was apparent. Finally, we annotated the RDD SNV and investigated those non
148 synonymous and found in the differential expression gene set from our previous
149 work for potential functional impact.

150

151 **Results and Discussion**

152 *Exome and Transcriptome Design Coverage*

153 Depth of coverage in individual exomes was greater than 10x at 90% of the exome
154 capture design, with all but one sample having 10x coverage over 93% (Figure 1). The
155 percentage of 10x coverage achieved, a standard indicator of quality (Clark *et al.*
156 2011), was deemed sufficient to proceed with SNV analysis. Coverage across
157 transcriptomes was lower than that of WES (Figure 1), with 10x coverage on average
158 at 85.9% and 83.9% of the expressed regions in liver and muscle respectively. This is
159 likely due to inherently uneven coverage of transcriptomic reads, which yielded very

different coverage profiles than WES. For instance, we observed approximately 30% of transcriptome reads in both tissues were aligned to regions with a depth of 100x or more, while only ~1% of the reads from the exome data aligned to such highly covered regions (Figure 1). While SNVs with lower levels of coverage may be missed in transcriptomes due to the reduced regions covered that we report, there is likely to be an increase in support for SNVs in regions that have a higher depth of coverage. Because of this, a strategy increasing the minimum coverage threshold for SNV discovery in RNAseq data may be prudent.

Sequencing metrics are given in Table 1. Per animal, WES had coverage of 97.1 – 97.6% on-target when compared to the exome capture design, and the combined WES data covered 99.1% of the target design. While this difference between individuals is of interest and may simply reflect the random nature of the exome enrichment process, overall less than one percent of the target design could not be captured/sequenced per sample. Increased sequencing levels might potentially be able to ‘fill in’ those regions missed in individual WES based on being found in other individuals if the design was to be used again. The exome capture design can be refined, with more probes added for regions found to be less well covered. The proportion of bases across the exome capture design available for SNV calling for reads passing HaplotypeCaller filtering were 85.4 – 90.6% for individuals and 94.2% for the combined data. This equates to 48,397,629 – 51,344,557 bp in individual exomes, and a total of 53,384,739 bp in the combined data.

Liver and muscle transcriptome designs covered 44,842,138 and 44,506,796 bp of the bovine genome respectively. This equated to 96.8% of the annotated exonic regions of the UMD 3.1 genome annotation in each transcriptome design, and

184 78.5% of the exome capture design. We therefore concluded that the component of
185 the exome design missing from transcriptomes largely came from the 100bp 5' UTR
186 of each exon, which amounted to over 22.7MB. The two transcriptomes shared a
187 total of 41,815,289 bp (93.4% of the smaller muscle transcriptome) and combined
188 covered the entirety of annotated protein coding regions. The different regions
189 covered by the two tissues is expected because of tissue-specific gene expression;
190 combining two or more tissues has been suggested as a method to increase the
191 potential for calling variants from transcriptome data (Cirulli *et al.* 2010) and our
192 results show this is likely a good approach. We add that time-point based
193 transcriptomic data could increase the potential of SNV calling from RNAseq data.

194 A much lower proportion of reads from the transcriptomes passed GATK
195 filters. Output indicated that only 23.9 – 40.8% and 28.9 – 37% of positions, or
196 10,325,426 – 17,324,585 bp and 10,554,633 – 13,501,253 bp in liver and muscle
197 transcriptomes respectively passed these filters. This eliminated large transcriptomic
198 regions for calling of SNVs. Therefore, the combined transcriptome data which were
199 available for SNV calling represented just 55% and 46.5% of the total liver and
200 muscle transcriptome designs respectively, equating to totals of 24,446,937 and
201 20,685,045 bp. These transcriptome design sizes were much higher versus the
202 individual ranges, likely due to the transcriptomes coming from three quite distinct
203 time-points at which we might expect different physiological processes to be active,
204 and therefore expression of different genes to be occurring. We tested this
205 hypothesis by combining transcriptomes into time-point specific designs, which
206 showed 84 – 91.1% similarity at each time-point between individuals in liver, and

207 85.2 – 91.6% similarity in muscle. Again this reinforces our assertion that using
208 RNAseq from multiple time-points is useful for calling SNVs when available.

209

210 *Variant Filtering*

211 Using gene expression values to filter RNAseq-derived variants was an
212 obvious strategy. We have found that genes with abundance below an FPKM value
213 of approximately 1, in both tissues, were excluded from the SNV calling procedure
214 because of not passing the HaplotypeCaller filters (Figure 2). This result highlights
215 that genes in regions determined as ‘LOW_COVERAGE’ instead of ‘PASS’ (the
216 threshold being 5 reads coverage) were on average at FPKM < 1, and may be useful
217 for other researchers looking for a baseline threshold for removing genes prior to
218 SNV calling.

219 Further filtering of the initial variant calls, beyond the HaplotypeCaller
220 genotyping stage, was carried out by retrospectively removing systematic errors,
221 SNVs with more than a single alternate allele, and SNVs failing our hard-filtering for
222 coverage and occurrence. Initially 219,934 SNVs were called from WES, which was
223 reduced to 135,950 following filtering. This represented a rate of 1 SNV every 417 bp
224 of exome capture design, a rate of 0.2% per base. Of these 88.6% were annotated in
225 dbSNP, the remaining 15,521 being novel. The entire set of SNVs, including those
226 found novel, was used for specificity analysis.

227 Initially 123,295 and 107,406 SNVs were called from the liver and muscle
228 transcriptomes respectively, which following filtering was reduced to 48,076 and
229 37,526 SNVs. Of these 82.5% and 82.3% were found in dbSNP, the remaining 8,436
230 and 6,481 being novel, of which 2,999 and 2,374 (35.5% and 36.6%) were also found

231 in WES. To reduce the possibility of including FP SNVs, technical artefacts of the
232 transcriptome generation process, we only retained those which were also found in
233 WES. This resulted in rates of 1 SNV every 1,043 and 1,334 bp in liver and muscle
234 respectively, rates of 0.095 and 0.075% per base. We also identified 26,067 SNVs
235 common to both the liver and muscle transcriptomes which represented 61.1% and
236 78.1% of the total SNVs in liver and muscle respectively. Of these 1,795, or 6.7%,
237 were novel.

238 To determine if WES and transcriptome could accurately identify samples,
239 and those with shared sireage, and MDS plot was constructed using the combined
240 VCF. Insofar as we can visually assess differences between WES and transcriptome
241 derived SNVs, there is some separation (though not complete) between phenotypes,
242 and we have demonstrated that both approaches are reliable for identifying shared
243 sireage (Figure 3). Further, we have shown that this method can act as a quality
244 control step to assure that samples have not been mislabeled, or suffered other
245 handling errors during the laboratory phase of such an experiment.

246

247 *Specificity Analysis and RNA-DNA Differences*

248 Specificity plateaued at approximately 80% in both transcriptomes, indicating that
249 filtering for depth alone could not improve the FP call rate from RNAseq data (Figure
250 4). However, increasing the minimum read depth supporting the SNV from 8 to 15
251 reads in 4 or more samples increased the specificity to almost 85%. However, after
252 this initial increase in specificity at lower coverage we encountered a decline and
253 plateau between 75 – 80% thereafter (Figure 4). We interpret this as resulting from a
254 greater proportion of FP SNVs being called in the most highly expressed genes, and

255 therefore at higher coverage. This may have occurred with a higher number of
256 aligned reads resulting in a greater chance of observing an artefactual mutation.
257 Therefore, whilst GATK uses a maximum allowed coverage for calling mutations, a
258 further strategy that normalises or down-samples the most highly expressed genes
259 in RNAseq data may be prudent prior to variant calling.

260 Overall, 9,937 and 7,650 FP SNVs in liver and muscle were also annotated in
261 dbSNP, indicating that they have been independently observed in other studies of
262 cows. We investigated why we had only observed these SNVs in the transcriptomes.
263 Filtering of SNVs that were found in regions of WES with low quality only accounted
264 for 196 and 147 in liver and muscle respectively. This reflected the overall high
265 quality of data generated by WES. Filtering for SNVs that were different to the
266 reference genome and fixed in our population removed a further 175 and 122 FP
267 SNVs. We were also able to account for 408 and 237 FP SNVs found in the last 5 bp
268 of reads; 169 and 90 FP SNVs within 5 bp of splice junctions; and 606 and 333 FP
269 SNVs found within 5 bp of one another in liver and muscle respectively. However,
270 due to overlap between SNVs in each of these categories, totals of 762 and 442 FP
271 SNVs were removed, reducing each set by only 11.8% and 8.5% respectively. This
272 resulted in a final set of RNA-DNA differences (RDDs) of 8,869 and 7,004 FP SNVs in
273 liver and muscle. Of these, 2,160 and 1,936 were unique to liver Fert+ and Fert-
274 phenotypes, and 1,743 and 1,899 unique to muscle Fert+ and Fert- phenotypes,
275 respectively.

276 The transition/transversion ratio (Ts/Tv) of true positives (SNVs found in WES
277 and transcriptome; TPs) and RDDs for each tissue is shown in Table S1. In WES, Ts/Tv
278 is expected to be ~3 (Marth *et al.* 2011) and we have found a range from 2.63 to 3.57

279 depending on the SNV set analysed (Table S1). Interestingly, we have found highly
280 significantly different ($p < 0.001$) Ts/Tv ratio between TPs and RDDs. This manifested
281 as an increase in the frequency of transitions in the RDDs. Furthermore, this result
282 was observed in both liver and muscle indicating a general phenomenon of increased
283 transition mutations in RNA versus DNA.

284

285 *Functional Annotation*

286 The numbers of annotated SNVs found uniquely or in both fertility phenotypes in
287 WES and were determined (Figure 5a, b). We have found that of the SNVs unique to
288 a particular phenotype in WES were evenly split, 4,450 and 4,008 were novel, 25 and
289 26% in Fert+ and Fert- respectively. Of those, 1,158 and 1,031 (again approximately
290 25% of each set) were found to be non-synonymous (NS). In RDDs we observed
291 lower proportions being NS, with approximately 15% in both liver and muscle in both
292 Fert+ and Fert-. There was low concordance of differential expression (DE) as found
293 previously (Moran et al 2016) with NS SNVs, with totals of 181 (7.7%) and 142 (7%) in
294 Fert+ and Fert-. In liver, there was a lower proportion of NS RDDs unique to Fert+ or
295 Fert- overlapping the DE gene sets than was found in WES, with 15 in each (5% and
296 6% respectively). In muscle this trend was reversed, and proportionally more NS
297 RDDs in DE genes were found unique to Fert+ or Fert- than in WES, with 22 and 20
298 (9.1% and 8.4%) respectively. In TPs, there was divergence across these proportions,
299 with 47 each in liver (8.1% and 8.7%), and 38 and 41 in muscle (6.3% and 9.5%) for
300 Fert+ and Fert- respectively. All NS SNVs in liver and muscle RDDs and WES,
301 categorized unique to phenotype, are available in Supplementary Table S2.

Several of the DE genes with RDDs are of relevance to the phenotype, including the bovine leukocyte antigen (BOLA) super-family of genes. Major histocompatibility (MHC) genes have been shown to contain increased levels of mutations, a well-studied phenomenon in humans (Penn *et al.* 2002). Our identification of mutations in BOLA genes found uniquely in RNAseq data raises some interesting questions. In liver, Fert- cows had three unique NS SNVs in *BOLA*, and also one in *BOLA-DRA*, both of which were found to be more highly expressed in Fert- cows in liver in our previous study (Moran *et al.* 2016). Fert+ cows had one *BOLA* and two *BOLA-NC* NS SNVs. Muscle transcriptome showed the same effect, with four NS SNVs in *BOLA* and one in *BOLA-NC* in Fert-; two were found in *BOLA* in Fert+, and one in *BOLA-NC* in Fert+. In WES there were higher numbers, which was expected given the higher quality of the data. *BOLA* contained 8 and 15 NS SNPs in Fert+ and Fert- respectively. Interestingly *BOLA-DQA5* had one NS SNV in Fert-, but none in Fert+, whereas the MHC-related *JSP.1* gene was found with 20 NS SNVs in Fert+ cows, but only one in Fert-. We have found this gene up-regulated at all but one of six time-points. This is an indication of the potential of investigating RDDs to determine their influence on gene expression.

Conclusions

We have presented the first full characterisation of a bovine whole exome sequencing experiment, with the added benefit of using both a novel exome-enrichment method and complementary transcriptomic data. Previous studies have confirmed the utility of exome sequencing in cows for specific phenotypic and haplotypic studies (Hirano *et al.* 2013; McClure *et al.* 2014; Newman *et al.* 2015). Our

findings show that the Roche Nimblegen Developer system has very good potential for generation of WES data based on Ensembl annotation, with very high on-target call rates. Unsurprisingly, we have found that SNV calling in WES data is more informative than using RNAseq data for the same purpose. However, we do not feel that this should exclude use of RNAseq from SNV calling. Indeed, WES exists to both reduce costs of DNA sequencing versus whole genome sequencing, but also because mutations in protein-coding regions are expected to be of greater consequence. To this end, RNAseq-called SNVs will inevitably be of value as these genes are by definition expressed. We have shown that some basic steps can be taken to make SNV calling in RNAseq more reliable and we hope these are of interest and might be adopted by others.

Using both data types we have identified a large number of RDDs per transcriptome which significantly deviate in Ts/Tv from our TP SNVs found in both WES and transcriptome. This shows that the results of SNV discovery using RNA sequencing are not identical to WES, and researchers should be aware of this if calling variants without concordant DNA sequencing. However, the ability of transcriptome-defined SNVs to distinguish sireage suggests that taking a conservative approach with SNV analysis resulting from RNAseq can yield biologically relevant and accurate results, and opens up the possibility of mining the vast amount of transcriptomic data available for non-model organisms.

We have also shown that large regions of the transcriptome may not be suitable for variant calling due to low levels of reads passing filters in the gold-standard GATK pipeline. We recommend particular attention be paid to this when using RNAseq data for variant calling. Our key finding that about 20% of SNVs

350 passing all filters were false positives requires further validation. Whilst there is the
 351 intriguing prospect of those SNVs being due to processes such as RNA editing or
 352 other modifications, there is also the prospect that they are due to GC bias,
 353 sequencing errors or because of the evolutionary history of the genes involved. We
 354 stress that whilst the finding of these SNVs is an interesting discussion point little
 355 functional meaning can be implied without further validation and, even if these
 356 RDDs are real, they may be completely benign in terms of gene expression and
 357 function.

358 A stringent approach for identifying SNVs, and particularly the difference in
 359 the transition/transversion ratio observed between the TPs and RDDs, is presented.
 360 There is an obvious need for further protein and Sanger sequencing work to validate
 361 and investigate fully the effect of non-synonymous coding variation found to be
 362 divergent between the phenotypes used here. The example of the BOLA gene family,
 363 known to have high levels of polymorphism in dairy cows (Hou *et al.* 2012), could be
 364 of use in haplotyping animals to further investigate their relationship with fertility.
 365 Whilst we have concentrated on NS RDD SNV annotation, we have identified a large
 366 number of TP SNVs from RNAseq data. This, along with other public data, represents
 367 a wealth of high-quality genetic information that can immediately be harnessed by
 368 researchers for a wide variety of studies.

369

370

371 **Methods**

372

373 *Tissue Biopsy, RNA Extraction and Transcriptome Sequencing Strategy*

374 Tissue biopsies were collected at three time-points relative to parturition (day 0):
 375 late pregnancy (LP), day -18 (sd = 7); early lactation (EL), day 2 (sd = 1; EL); and mid-
 376 lactation (ML), day 147 (sd = 13). Liver tissue was collected by puncture biopsy as
 377 previously described (Cummins, Waters, *et al.* 2012). To collect muscle tissue, a
 378 biopsy site on the semitendinosus muscle was shaved and sanitized with 7.5%
 379 iodinated povidone and methylated spirits. A subcutaneous injection of lidocaine
 380 hydrochloride (2%) was used to anesthetize the area. An incision was made through
 381 the skin, and the biopsy instrument (Biopsy Punch 33-37, Miltex GmbH, Riethein-
 382 Weilheim, Germany) was used to remove a core of muscle tissue. The incision site
 383 was sutured and treated topically with Duphacycline aerosol (3.6% oxytetracycline
 384 hydrochloride: Norbrook Laboratories Ltd., Newry, Northern Ireland). Both liver and
 385 muscle tissue biopsies were immediately rinsed in saline, blotted dry, snap frozen in
 386 liquid nitrogen and stored at -80 °C until RNA extraction.

387 Total RNA was extracted using a standard Trizol-based method. The tissue
 388 sample was weighed and 100 mg cut and homogenized in 3 ml TRI Reagent (Sigma-
 389 Aldrich, Dublin) until fully homogenized using a hand-held device. The homogenate
 390 was removed to sterile Eppendorf tubes (Eppendorf, UK) and incubated at room
 391 temperature (RT) for 5 min; samples were centrifuged at 12,000 × g for 10 min at
 392 4°C, and the supernatant was removed to new sterile tubes. Chloroform was added
 393 at 0.2x the volume of homogenate and incubated at RT for 3 min and samples were
 394 centrifuged at 12,000 × g for 10 min at 4°C. Isopropanol was added at 0.6 times the
 395 volume of supernatant, vortexed and centrifuged at 12,000 × g for 10 min at 4 °C to
 396 pellet the RNA. The supernatant was discarded, the pellet was washed twice in 99%
 397 ethanol (Sigma-Aldrich, Dublin), and centrifuged at 7,500 × g for 5 min at 4°C. The

398 RNA was re-suspended in 50 µl nuclease-free water (Sigma-Aldrich, Dublin). A kit
399 based protocol (RNeasy Plus; Qiagen, UK) was used to clean the total RNA, removing
400 the fraction below 200 bp and any genomic DNA. RNA quality was assessed using the
401 Bioanalyser 2100 (Agilent Technologies, UK) with the RNA Nano chip.

402 Library preparation was conducted using the Truseq v2 kit (Illumina, UK)
403 following the protocol. The Bioanalyser 2100 (Agilent Technologies, UK) was used to
404 visually determine quality of the libraries with a DNA Nano chip. Concentration was
405 determined using the KAPA Library Quantification (KAPA Biosystems, USA) qPCR
406 method to allow equimolar pooling of each library. Four pools of 24 libraries were
407 made, the maximum possible based on the number of barcoded adapters available
408 at the time. Each pool was sequenced for 100 bp using a paired-end strategy across 3
409 separate flowcell lanes on an Illumina HiSeq 2000 (Illumina, USA) to minimize
410 technical variation, resulting in 12 lanes of HiSeq data. RNAseq data are available in
411 the Gene Expression Omnibus (GEO) at accession ID GSE62159.

412

413 *Exome Design, DNA Extraction and Sequencing Strategy*

414 We used the Roche Nimblegen SeqCap EZ Developer kit to perform the experiment
415 and this was, to our knowledge, the first study using this method on the bovine
416 genome. This method requires a user-generated exome sequence from which
417 probes are made; we used the exome regions from the UMD.3.1 Ensembl 70 release
418 genome transfer file (GTF) including 100bp of 5' sequence of all exonic regions of
419 which there are 227,647. Our exome target design covered 202,899 consolidated
420 regions over 56,671,697 bp. Probes were allowed to have up to 5 total matches in
421 the genome and 5 mismatches within those, although 92.5% of probes were unique

422 to a single genomic position and a further 4.5% had only two possible matches
423 within the genome. Probe length was 200bp with approximately 2 million probes
424 covering the exome. Probes targeting mitochondrial DNA were at 1/5th the
425 concentration of nuclear DNA.

426 Genomic DNA (gDNA) was extracted from liver and muscle tissue biopsies
427 using the Zymo Quick-gDNATM MiniPrep kit (Zymo Research Europe, Germany). We
428 processed samples in triplicate as per the protocol with starting volume of 25mg of
429 tissue homogenised with a Precellys bead beater in 500ul of Genomic Lysis Buffer.
430 We eluted in 50ul of DNA Elution Buffer and proceeded to the Zymo Genomic DNA
431 Clean and ConcentratorTM to clean and concentrate the gDNA. We used an input of
432 25ul and output of 10ul. This gDNA was sent to a core facility, which conducted the
433 library preparation, hybridisation with exome capture kit and sequencing on 3 lanes
434 of HiSeq 2000. Our strategy for hybridisation and sequencing was to make one single
435 equimolar pool of all 16 gDNA libraries, each having different adapter barcodes to
436 allow us to de-multiplex sequence data into the correct sample for analysis. We then
437 split the pool in four parts and performed hybridisation captures on each part using
438 2x capture reagents, resulting in 1x capture for every 2 samples. All captures were
439 then re-pooled and then split into three parts for sequencing on three HiSeq lanes.
440 We used pooling strategies to attempt to reduce variation in hybridisation, and
441 because we were ultimately pooling the captured libraries to sequence over three
442 lanes. Thus, the first pooling of library ensured the samples were well mixed prior to
443 capture, and then the capture was well mixed to give as even a distribution of
444 sample to lanes, and previously sample to capture reagents, as possible. Exome data
445 are available in the Gene Expression Omnibus (GEO) at accession ID 62160.

446

447 *Sequence Alignment*

448 WES data were aligned with BWA using the ‘sampe’ algorithm (Li and Durbin 2009).

449 Resulting SAM files were sorted, duplicates removed and read groups added using

450 Picard Tools (available from: www.picard.sourceforge.net). Pre- and post-duplicate

451 removal statistics were generated for SAM files using Samtools Flagstat (Li *et al.*

452 2009) to determine numbers of reads properly paired and aligned. SAM files were

453 converted to BAM files using Samtools View which were subsequently used for

454 variant calling.

455 Transcriptome data were aligned using STAR aligner (Dobin *et al.* 2012) with a

456 ‘two-pass’ strategy for novel splice junction annotation. The aligner was first run

457 using the same gene transfer format (GTF) file that was used in designing the exome

458 from which canonical splice junction boundaries could be determined. This

459 preliminary run resulted in a splice junction database per sample; we collected all

460 such splice junction databases, removed junctions which had less than 20

461 occurrences, and used that set of junctions for our ‘second-pass’ alignment, instead

462 of the GTF file. We then sorted, removed duplicates and added read groups with

463 Picard Tools, generating pre- and post-duplicate removed SAM file statistics using

464 Samtools Flagstat. SAM files were once again converted to BAM files using Samtools

465 View and we proceeded to variant calling.

466

467 *Variant Calling*

468 The Genome Analysis Tool Kit (GATK) version 3.1 (McKenna *et al.* 2010; DePristo *et*

469 *al.* 2011; Van der Auwera *et al.* 2013) ‘walkers’ (algorithms for a variety of data

processing) were used under their ‘best practices’ guidelines for sequence variant calling. For transcriptome BAM files the SplitNCigarReads walker was used to hard clip bases following ‘N’ in the CIGAR string. This removes all “overhangs” into the intronic regions. Having previously defined the splice junctions in STAR we were confident that this approach did not result in elevated sequence removal compared to the overall greater alignment and ultimately potentially greater variant calling accuracy. This walker also reassigns mapping qualities to change MAPQ 255 from STAR, meaning the read aligned uniquely (i.e. to only one position of the genome). The 255 score is determined as ‘unknown’ by GATK walkers, and so all 255 scores were changed to 60 which is used downstream in GATK to mean uniquely aligned reads.

WES and RNAseq were then processed almost identically. The RealignerTargetCreator and IndelRealigner walkers were run on the BAM file with the Ensembl 74 version dbSNP variants given as known variants. The BaseRecalibrator walker was then used to recalibrate scores around those known variants and PrintReads walker was used to generate a recalibrated BAM file. The HaplotypeCaller walker was used with Phred-scaled emit and call confidences of 30 for WES and 20 for RNAseq in the ‘GVCF’ mode and with the BED file of our exome target design to guide calling. This resulted in a single genomic variant call format (GVCF) file per sample. The entire set of GVCF files for each of WES, and the liver and muscle transcriptomes separately, were then ‘joint genotyped’ using the GenotypeGVCFs walker. This generated a single joint genotyped GVCF file containing information for all 16 exomes at every variant site found in any of the exomes, and 48 samples each in liver and muscle transcriptomes, at each variant site found in that

transcriptome in any of the 48 samples. We call the joint GVCF files the ‘variant call files’ (VCF) henceforth for clarity. We removed all but the SNVs in which a single alternative allele was found in the variant call file. This is because the ENSEMBL version 74 bovine dbSNP file contained ~21.2M positions of which ~19.7M SNPs (~93%) have a single alternate allele and so other variants might be under-represented, making them more difficult to accurately call. We retained a distinction between the two transcriptomes (i.e. liver and muscle tissue) because this might inherently be of interest: (i) to independently validate results; (ii) for functional annotation of SNPs.

503

504 *Systematic Error Removal*

To remove possible systematic sequencing errors we used the Syscall tool (Meacham *et al.* 2011). For WES and each transcriptome separately, we took each realigned, recalibrated BAM file (the file from which SNP calls were made), and also took the set of all heterozygous SNVs from the concordant GVCF file per sample. To save computational time we ran Syscall over the first sample, then took all systematic errors as called by Syscall, removed them from the second samples heterozygous SNVs and proceeded through all samples. A pooled error-file was produced and any SNV positions in that were removed from the VCF. This approach of analysing the BAM files individually in turn was completed in approximately a third of the time estimated for carrying out the same analysis on a single BAM file containing all samples (factoring in the time taken for merging of all exome BAM files, sorting and indexing).

517

518 Hard-filtering of SNVs

519 We had no genotype data from other sources (e.g. whole genome sequence or SNV
520 array) for the animals used in this study and so it was not possible to conduct variant
521 quality score recalibration (VQSR) in GATK. We therefore conducted a series of
522 filtering steps on our variant call files to retain data in which we had higher
523 confidence. We required each SNV to have at least one alternate allele in at least
524 one sample with greater than 7 reads supporting the alternate allele. If the sample
525 was heterozygous the alternate allele also had to be represented by at least 25% of
526 the reads covering the position as well as having at least 7 reads supporting the
527 alternate allele. To test if filtering affected the specificity analysis we increased the
528 hard-filtering thresholds, requiring the SNV to be found in at least 4 different
529 samples, and at depth of at least 15 reads.

530

531 Reads Passing GATK Filters

532 To determine what proportion of WES or transcriptome could be used in SNV calling
533 by the HaplotypeCaller we ran DiagnoseTargets. This walker applies the same filters
534 to the BAM as HaplotypeCaller and outputs a file with intervals tagged as 'PASS' or
535 other indicators of filtering ('NO_READS', 'LOW_COVERAGE'), where 'PASS'
536 represents intervals which pass all the filters for quality which were applied. We
537 merged the intervals not passing filters into BED format and took total bases divided
538 by total size of the exome design to get a percentage of the total target design in
539 which SNVs were not callable. Because transcriptomes were inherently smaller than
540 the entire exome design we used two transcriptome designs, as well as the exome
541 design as above, to get percentage of callable regions per each transcriptome.

542

543 *Coverage Analysis*

544 We used Bedtools genomecov (Quinlan and Hall 2010) to determine coverage across
545 the genome of each of the realigned, recalibrated BAM files, with output in BED
546 format. BED files contained a chromosomal interval between which reads in the
547 BAM were mapped, as well as holding information on the number of reads
548 supporting that interval. Bedtools intersect was used to determine intervals that
549 were common between individual BED files. For exomic BED files, we intersected the
550 individual BED files with our exome design to determine how well the exome capture
551 process had worked, i.e. how much of the entire exome was covered by at least one
552 sequencing read. We also used transcriptome BED files to determine how well
553 transcriptome intersected the exome design. We combined all WES and
554 transcriptome BED files, merged using Bedtools merge to join overlapping regions,
555 and intersected this with the exome design to show what intervals, if any, were
556 missed entirely by WES, and what regions of exome were not covered by the
557 transcriptomes. We used each merged transcriptome separately to calculate
558 percentage of callable regions in each.

559 We expected expression of different genes at the different time-points that
560 were represented in the transcriptomes, but did not expect all genes to be expressed
561 in either tissue. Therefore, we used the absolute abundance of genes in fragments
562 per kilobase per million reads (FPKM; Mortazavi *et al.* 2008), and investigated in
563 which genes variants could be called based on passing filters. This gave a distribution
564 of FPKM prior to and following filtering using the preliminary ‘single alternate allele’

565 and removal of systematic biases, after our hard-filtering, and after the
566 HaplotypeCaller 'PASS' filtering.

567 We also used BED files coverage information to determine the proportion of
568 coverage at depth. We summed all depths (giving the total number of reads covering
569 the transcriptome) and then sequentially took the reads at the next depth minus the
570 reads at the last depth from the total to show what proportion of reads was left at
571 that particular depth. For example for each depth within the window 1 - 10X having
572 5 occurrences in the BED we have 50 total reads (proportion is $50/50 = 1$). Taking
573 away depth=1x leaves 45 reads at depth=2x (proportion of $45/50 = 0.9$) and so on.

574

575 *Multi-Dimensional Scaling Plot of Combined Exome and Transcriptomes*

576 We created a combined VCF containing all SNV data following filtering from WES and
577 both transcriptomes. Where an SNV was found in only one or two datasets, the
578 remaining set(s) had missing values inserted. Because the exome and two
579 transcriptome sets came from the same animals, we visualised the identity-by-state
580 of the animals with the combined VCF. We used PLINK 1.9 (PLINK v1.90b2a) multi-
581 dimensional scaling (MDS) from the cluster method. We plotted the first two
582 dimensions using R statistical software (R Development Core Team 2012) and as all
583 cows except two had shared sireage with one other cow in the study, we
584 superimposed this information onto the plot.

585

586 *Specificity and Transition/Transversion Analysis*

587 We determined what overlap occurred between WES and transcriptomes by
588 specificity analysis. In this context, we defined specificity as the proportion of all

589 SNVs detected that are true positives (Cirulli *et al.* 2010). For this purpose SNVs were
 590 categorised as true positive (TP) when the SNV was found in both WES and
 591 transcriptome VCFs; false positive (FP) when an SNV was found in transcriptome but
 592 not in WES; false negative (FN) when an SNV was found in WES and not in
 593 transcriptome. We used an in-house Perl script available at from the authors on
 594 request) to parse the WES VCF along with one of the transcriptomes' VCF, reporting
 595 back the TP, FP and FN values per mean depth (i.e. average coverage of the SNV in
 596 all samples in which it was found) and also the specificity, which was calculated as
 597 TP/(TP+FP). Our script also produced a new VCF in which each SNV was tagged as
 598 either TP, FP or FN for use in downstream analyses.

599

600 *Detection of RNA-DNA differences*

601 It was expected that WES data should inherently contain more SNVs due to a greater
 602 coverage and range than RNAseq data. We also investigated those SNVs found in
 603 RNAseq but not WES. A variety of possible factors influenced calling these FP SNVs
 604 and so we removed SNVs called as FP when (i) the FP SNP was found in a region not
 605 passing HaplotypeCaller filters in WES, and therefore lacking coverage to be sure the
 606 SNV was not found in WES, and (ii) when the SNV was different from the reference in
 607 the bovine genome but was found fixed in our data. We refer to the remaining FP
 608 SNVs as 'RNA-DNA differences' (RDDs), as coined previously (Li *et al.* 2011). To filter
 609 RDDs we used methods based on further work arising in the field (Pickrell *et al.*
 610 2012; Kleinman and Majewski 2012; Lin *et al.* 2012). Three relevant methods were
 611 used to determine if RDDs were due to sequencing-related factors: (i) RDD was at
 612 the extremity of reads supporting it; to resolve this we took all reads in BAM

613 supporting RDDs at the sample level and if all were within 5 bp of either end of the
614 reads, we removed them; (ii) if the RDD was within 5 bp of any splice junction from
615 our database created in STAR, we removed them; (iii) if the RDD was within 5 bp of
616 any other RDD, we removed them both due to potential indel or other variants
617 causing the SNV.

618 Finally, to address differences between TP and RDD in each transcriptome,
619 we investigated absolute and proportional occurrence of each type of mutation, i.e.
620 G>A or C>G etc. As our transcriptome data was not strand-specific, we could not be
621 certain of the direction of the mutation, and therefore grouped reciprocal mutations
622 i.e G>A and T>C, together. Finally to investigate the properties of RDDs we calculated
623 transition/transversion ratio (Ts/Tv) using total numbers of each reciprocal mutation
624 category and tested whether the difference in the Ts/Tv ratios in RDDs was
625 significantly different to that from the TP dataset using a Chi Squared test ($p < 0.05$).

626

627 *Functional Annotation of SNVs*

628 The cows sequenced were part of a model of fertility. Therefore, it was of interest to
629 identify SNVs found divergent between the two phenotypes (Fert+ and Fert-). We
630 analysed WES, TPs and RDDs individually for each phenotype. Using snpEff (Cingolani
631 *et al.* 2012) we annotated all SNVs and identified those that were non-synonymous
632 (NS). Due to their potential to cause differences in the structure and functions of a
633 gene, we limited our functional annotation to this set of NS SNVs. We also identified
634 those NS SNVs that were unique to one phenotype or the other in each of WES, TPs
635 and RDDs. Finally, we compared these to the sets of genes found to be differentially
636 expressed in the RNAseq analysis from which the transcriptomes were derived.

637

638 **References**

- 639 Butler, S. T., 2014 Genetic control of reproduction in dairy cows. *Reprod. Fertil. Dev.*
640 26: 1.
- 641 Cingolani, P., A. Platts, L. L. Wang, M. Coon, T. Nguyen *et al.*, 2012 A program for
642 annotating and predicting the effects of single nucleotide polymorphisms,
643 SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2;
644 iso-3. *Fly (Austin)* 6: 80–92.
- 645 Cirulli, E. T., A. Singh, K. V. Shianna, D. Ge, J. P. Smith *et al.*, 2010 Screening the
646 human exome: a comparison of whole genome and whole transcriptome
647 sequencing. *Genome Biol.* 11: R57.
- 648 Clark, M. J., R. Chen, H. Y. K. Lam, K. J. Karczewski, R. Chen *et al.*, 2011 Performance
649 comparison of exome DNA sequencing technologies. *Nat. Biotechnol.* 29:
650 908–914.
- 651 Cosart, T., A. Beja-Pereira, S. Chen, S. B. Ng, J. Shendure *et al.*, 2011 Exome-wide
652 DNA capture and next generation sequencing in domestic and wild species.
653 *BMC Genomics* 12: 347.
- 654 Cummins, S. B., P. Lonergan, A. C. O. Evans, D. P. Berry, R. D. Evans *et al.*, 2012
655 Genetic merit for fertility traits in Holstein cows: I. Production characteristics
656 and reproductive efficiency in a pasture-based system. *J. Dairy Sci.* 95: 1310–
657 1322.
- 658 Cummins, S. B., P. Lonergan, A. C. O. Evans, and S. T. Butler, 2012 Genetic merit for
659 fertility traits in Holstein cows: II. Ovarian follicular and corpus luteum
660 dynamics, reproductive hormones, and estrus behavior. *J. Dairy Sci.* 95:
661 3698–3710.
- 662 Cummins, S. B., S. M. Waters, A. C. O. Evans, P. Lonergan, and S. T. Butler, 2012
663 Genetic merit for fertility traits in Holstein cows: III. Hepatic expression of
664 somatotrophic axis genes during pregnancy and lactation. *J. Dairy Sci.* 95:
665 3711–3721.
- 666 DePristo, M. A., E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire *et al.*, 2011 A
667 framework for variation discovery and genotyping using next-generation DNA
668 sequencing data. *Nat. Genet.* 43: 491–498.
- 669 Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski *et al.*, 2012 STAR:
670 ultrafast universal RNA-seq aligner. *Bioinformatics* 29: 15–21.

671 Emond, M. J., T. Louie, J. Emerson, W. Zhao, R. A. Mathias *et al.*, 2012 Exome
672 sequencing of extreme phenotypes identifies DCTN4 as a modifier of chronic
673 *Pseudomonas aeruginosa* infection in cystic fibrosis. *Nat. Genet.* 44: 886–889.

674 Hirano, T., N. Kobayashi, T. Matsushashi, D. Watanabe, T. Watanabe *et al.*, 2013
675 Mapping and exome sequencing identifies a mutation in the IARS gene as the
676 cause of hereditary perinatal weak calf syndrome. *PloS One* 8: e64036.

677 Hou, Q., J. Huang, Z. Ju, Q. Li, L. Li *et al.*, 2012 Identification of splice variants,
678 targeted microRNAs and functional single nucleotide polymorphisms of the
679 BOLA-DQA2 gene in dairy cattle. *DNA Cell Biol.* 31: 739–744.

680 Kleinman, C. L., and J. Majewski, 2012 Comment on “Widespread RNA and DNA
681 Sequence Differences in the Human Transcriptome.” *Science* 335: 1302–
682 1302.

683 Li, H., and R. Durbin, 2009 Fast and accurate short read alignment with Burrows-
684 Wheeler transform. *Bioinforma. Oxf. Engl.* 25: 1754–1760.

685 Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*, 2009 The Sequence
686 Alignment/Map format and SAMtools. *Bioinformatics* 25: 2078–2079.

687 Li, M., I. X. Wang, Y. Li, A. Bruzel, A. L. Richards *et al.*, 2011 Widespread RNA and DNA
688 Sequence Differences in the Human Transcriptome. *Science* 333: 53–58.

689 Lin, W., R. Piskol, M. H. Tan, and J. B. Li, 2012 Comment on “Widespread RNA and
690 DNA Sequence Differences in the Human Transcriptome.” *Science* 335: 1302–
691 1302.

692 Marth, G. T., F. Yu, A. R. Indap, K. Garimella, S. Gravel *et al.*, 2011 The functional
693 spectrum of low-frequency coding variation. *Genome Biol.* 12: R84.

694 McClure, M. C., D. Bickhart, D. Null, P. VanRaden, L. Xu *et al.*, 2014 Bovine Exome
695 Sequence Analysis and Targeted SNP Genotyping of Recessive Fertility
696 Defects BH1, HH2, and HH3 Reveal a Putative Causative Mutation in SMC2 for
697 HH3 (S. Moore, Ed.). *PLoS ONE* 9: e92769.

698 McGettigan, P. A., 2013 Transcriptomics in the RNA-seq era. *Curr. Opin. Chem. Biol.*
699 17: 4–11.

700 McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis *et al.*, 2010 The
701 Genome Analysis Toolkit: a MapReduce framework for analyzing next-
702 generation DNA sequencing data. *Genome Res.* 20: 1297–1303.

703 Meacham, F., D. Boffelli, J. Dhahbi, D. I. Martin, M. Singer *et al.*, 2011 Identification
704 and correction of systematic error in high-throughput sequence data. *BMC*
705 *Bioinformatics* 12: 451.

706 Moore, S. G., T. Fair, P. Lonergan, and S. T. Butler, 2014 Genetic merit for fertility
707 traits in Holstein cows: IV. Transition period, uterine health, and resumption
708 of cyclicity. *J. Dairy Sci.*

709 Moore, S. G., S. Scully, J. A. Browne, T. Fair, and S. T. Butler, 2014 Genetic merit for
710 fertility traits in Holstein cows: V. Factors affecting circulating progesterone
711 concentrations. *J. Dairy Sci.* 97: 5543–5557.

712 Moran, B., S. B. Cummins, C. J. Creevey, and S. T. Butler, 2016 Transcriptomics of
713 liver and muscle in Holstein cows genetically divergent for fertility highlight
714 differences in nutrient partitioning and inflammation processes. *BMC*
715 *Genomics* 17:.

716 Mortazavi, A., B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, 2008 Mapping and
717 quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5: 621–
718 628.

719 Newman, J. H., T. N. Holt, J. D. Cogan, B. Womack, J. A. Phillips *et al.*, 2015 Increased
720 prevalence of EPAS1 variant in cattle with high-altitude pulmonary
721 hypertension. *Nat. Commun.* 6: 6863.

722 Penn, D. J., K. Damjanovich, and W. K. Potts, 2002 MHC heterozygosity confers a
723 selective advantage against multiple-strain infections. *Proc. Natl. Acad. Sci. U.*
724 *S. A.* 99: 11260–11264.

725 Pickrell, J. K., Y. Gilad, and J. K. Pritchard, 2012 Comment on “Widespread RNA and
726 DNA Sequence Differences in the Human Transcriptome.” *Science* 335: 1302–
727 1302.

728 Quinlan, A. R., and I. M. Hall, 2010 BEDTools: a flexible suite of utilities for comparing
729 genomic features. *Bioinformatics* 26: 841–842.

730 Quinn, E. M., P. Cormican, E. M. Kenny, M. Hill, R. Anney *et al.*, 2013 Development of
731 Strategies for SNP Detection in RNA-Seq Data: Application to Lymphoblastoid
732 Cell Lines and Evaluation Using 1000 Genomes Data (B. W. Futscher, Ed.).
733 *PLoS ONE* 8: e58815.

734 R Development Core Team, 2012 *R: A Language and Environment for Statistical*
735 *Computing*. Vienna, Austria.

736 Van der Auwera, G. A., M. O. Carneiro, C. Hartl, R. Poplin, G. del Angel *et al.*, 2013
737 From FastQ Data to High-Confidence Variant Calls: The Genome Analysis
738 Toolkit Best Practices Pipeline: The Genome Analysis Toolkit Best Practices
739 Pipeline, pp. 11.10.1–11.10.33 in *Current Protocols in Bioinformatics*, edited
740 by A. Bateman, W. R. Pearson, L. D. Stein, G. D. Stormo, and J. R. Yates. John
741 Wiley & Sons, Inc., Hoboken, NJ, USA.

742 Warr, A., C. Robert, D. Hume, A. Archibald, N. Deeb *et al.*, 2015 Exome Sequencing:
 743 Current and Future Perspectives. *GenomesGenetics* 5: 1543–
 744 1550.

745

746

747

Proportional coverage of Liver, Muscle Transcriptomes on-target

Liver, muscle transcriptome SNVs vs. exome SNVs