

1 Copy number variants in the sheep genome detected using multiple approaches

2

3 Gemma M Jenkins ^{1*}

4 *corresponding author

5 Email: gjenkins@abacusbio.co.nz

6

7 Michael E Goddard ²

8 Email: mike.goddard@ecodev.vic.gov.au

9

10 Michael A Black ³

11 Email: mik.black@otago.ac.nz

12

13 Rudiger Brauning ⁴

14 Email: rudiger.brauning@agresearch.co.nz

15

16 Benoit Auvray ³

17 Email: bauvray@maths.otago.ac.nz

18

19 Ken G Dodds ⁴

20 Email: ken.dodds@agresearch.co.nz

21

22 James W Kijas ⁵

23 Email: james.kijas@csiro.au

24

25 Noelle Cockett ⁶

26 Email: noelle.cockett@usu.edu

27

28 John C McEwan ⁴

29 Email: john.mcewan@agresearch.co.nz

30

31 ¹ AbacusBio Limited, 442 Moray Place, PO Box 5585, Dunedin 9058, NEW ZEALAND

32 ² Victorian Department of Economic Development, Jobs, Transport and Resources, Bundoora, VIC
33 3083, AUSTRALIA

34 ³ Department of Biochemistry, University of Otago, 710 Cumberland St, Dunedin 9054, NEW
35 ZEALAND

36 ⁴ AgResearch, Invermay Agricultural Centre, PB 50034, Mosgiel 9053, NEW ZEALAND

37 ⁵ CSIRO Animal, Food and Health Sciences, Queensland Bioscience Precinct, 306 Carmody Road.

38 St Lucia, QLD 4067, AUSTRALIA

39 ⁶ Utah State University, 1435 Old Main Hill, Logan, UT 84322-1435-1435, USA

40

41 **Keywords:** sheep, copy number variants, array CGH, SNP, sequence

42

43 **Abstract**

44 **Background.** Copy number variants (CNVs) are a type of polymorphism found to underlie phenotypic
45 variation, both in humans and livestock. Most surveys of CNV in livestock have been conducted in
46 the cattle genome, and often utilise only a single approach for the detection of copy number
47 differences. Here we performed a study of CNV in sheep, using multiple methods to identify and
48 characterise copy number changes. Comprehensive information from small pedigrees (trios) was
49 collected using multiple platforms (array CGH, SNP chip and whole genome sequence data), with
50 these data then analysed via multiple approaches to identify and verify CNVs.

51 **Results.** In total, 3,488 autosomal CNV regions (CNVRs) were identified from 30 sheep. The average
52 length of the identified CNVRs was 19kb (range of 1kb to 3.6Mb), with shorter CNVRs being more
53 frequent than longer CNVRs. The total length of all CNVRs was 67.6Mbps, which equates to 2.7% of
54 the sheep autosomes. For individuals this value ranged from 0.24 to 0.55%, and the majority of
55 CNVRs were identified in single animals. Rather than being uniformly distributed throughout the
56 genome, CNVRs tended to be clustered. Application of three independent approaches for CNVR
57 detection facilitated a comparison of validation rates. CNVs identified on the Roche-NimbleGen
58 2.1M CGH array generally had low validation rates, while whole genome sequence data had the
59 highest validation rate.

60 **Conclusions.** This study represents the first comprehensive survey of the distribution, prevalence
61 and characteristics of CNVR in sheep. Multiple approaches were used to detect CNV regions and it
62 appears that the best method for verifying CNVR on a large scale involves using a combination of
63 detection methodologies. The characteristics of the 3,488 autosomal CNV regions identified in this
64 study are comparable to other CNV regions reported in the literature and provide a valuable
65 addition to the small subset of published sheep CNVs.

66

67 **Background**

68 Copy number variants (CNVs) are a type of genomic polymorphism that potentially underlie a
69 significant fraction of phenotypic variation [1]. CNVs are structural variants, defined as stretches of
70 DNA that are greater than 1 kilobase (kb) in size and are duplicated or deleted in the genome of
71 some individuals [2]. Mutation rate estimates for CNVs vary from 1.1×10^{-2} [3] to 1×10^{-8} per locus per
72 generation [4, 5], which reflects the diverse processes by which CNVs are created. They can be over
73 1 megabase (Mb) [6] and are thought to comprise approximately 1% of an individual's genome,
74 which is much higher than the 0.1% thought to comprise SNPs [7, 8]. CNVs can be present in the
75 same or overlapping regions of the genome in multiple individuals, these regions are called copy
76 number variant regions (CNVRs). Copy number variants are distinct from another type of variant,
77 indels (INsertions/DEletionS), in that indels are typically less than 1kb [2]. By definition they are also
78 distinct from segmental duplications (SD). Segmental duplications are defined as being over 1kb in
79 length with at least 90% sequence identity between the duplicated segments and are not
80 polymorphic in the population [9]. In many cases it is likely that segmental duplications were once
81 CNVs that have subsequently become fixed in the population.

82

83 There are many examples, particularly in humans, of CNVs influencing traits. These include multiple
84 examples of CNVs associated with cancer susceptibility [10-12], the association of the FCGR3B gene
85 copy number variant with systemic lupus erythematosus (SLE) [13], and CCL3L1 gene copy number,
86 which has been linked to HIV susceptibility [14]. There is also evidence for CNVs influencing traits in
87 other animal and livestock species. A 133kb duplication containing four genes causes hair ridge in
88 Rhodesian and Thai Ridgeback dogs [15]. The chicken Peacomb phenotype is under sexual selection
89 and is caused by a 3.2 kb duplication in an intron of the SOX5 gene [16]. The Peacomb allele contains
90 ~30 copies of the duplication, with variation in copy number present within individuals with the
91 Peacomb phenotype. In pigs, Chen *et al* [17] found seven copy number variable genes that
92 overlapped quantitative trait loci (QTL) for, among other traits, carcass length, backfat thickness,

93 abdominal fat weight, length of scapular, intramuscular fat content of *longissimus* muscle, body
94 weight at 240 days and glycolytic potential of longissimus muscle. Although not an association
95 analysis, Chen *et al* [17] identified one CNV that had previously been associated with skin colour in
96 pigs [18].

97

98 There have been many CNV studies in cattle, with a range of platforms used to identify CNVs [19-26].
99 Between 51 and 1265 CNVRs [20, 22] have been identified in the various cattle studies, with
100 estimates of the proportion of the cattle genome thought to contain CNVRs ranging from 0.5 to 20%
101 [22, 24]. Although the latter is likely to be an overestimate, the wide range in estimates is likely due
102 to a number of factors, including the technology used to detect CNVs, different CNV calling criteria
103 used, and the number of animals examined

104

105 While there is one notable example of a CNV having a direct effect on a sheep trait – the agouti
106 duplication influencing coat colour [27] - to date, little work has been published on copy number
107 variants in the sheep genome. An initial survey assayed eleven sheep on a cattle Roche-NimbleGen
108 385K oligonucleotide CGH array (oligo aCGH) which included 385,000 probes that were designed
109 based on the cattle genome build btau_4.0 [28]. That study identified 135 CNV regions (CNVR) that
110 covered approximately 0.4% of the sheep genome and ~0.01-0.13% of each individual's genome,
111 which is less than the approximately 1% estimated by Pang *et al* [8] in humans. This suggests many
112 more sheep CNVs remain to be identified.

113

114 A number of approaches have been used to detect the presence of CNV. The main platforms are
115 comparative genomic hybridisation (CGH) arrays [29-33], SNP arrays [34-37] and depth of coverage
116 metrics applied to whole genome sequence data (e.g., [38-42]). Further, there are a variety of
117 algorithms that can be used to analyse available resultant data. Perhaps the most widely used

118 platform is array CGH, as it represents a cost-effective method to detect CNVs on a genome-wide
119 scale in multiple individuals [43].

120

121 Trios have been used in CNV studies to determine the de novo mutation rate and to identify CNVs
122 that represent heritable genetic units [4, 22, 44, 5]. This involves identifying CNVs in a father-
123 mother-progeny trio. CNVs present in progeny and at least one parent are thought of as heritable
124 and CNVs present in progeny but not in either parent indicate either a de novo mutation or an error
125 in CNV identification. Given that CNVs are difficult to detect regardless of the platform or methods
126 used, the best approach appears to be the conservative use of multiple methods to generate a set of
127 high confidence CNV calls.

128

129 The objective of this study was to conduct a survey of sheep CNVRs using a range of detection
130 methods. A Roche-NimbleGen 2.1M CGH array was designed and 36 animals (which included sets of
131 trios) were assayed. Independent detection approaches were used in an attempt to validate the
132 results. Finally, the CNVRs detected in this study were compared to those reported in an earlier
133 survey of the sheep genome [28] and those detected in seven separate cattle studies [19, 20, 25, 28,
134 21-23].

135

136 **Results**

137 *Roche-NimbleGen 2.1M CGH array construction and application*

138 A total of four methodologies were used to detect CNV, with the main approach being the
139 development and application of a 2.1M probe CGH array for the sheep genome. In total, 2,012,210
140 probes were designed with an average spacing across the autosomes of approximately 1.2 Kb. The
141 array was used to assay a total of 36 sheep genomes, consisting of 30 individuals drawn from the
142 International Mapping Flock [45] and a further six from a Reference Panel of International Sheep

143 Genomics Consortium (ISGC) sheep (Supplementary Table 1). The Roche-NimbleGen segMNT
144 algorithm was used to call CNV segments in each animal compared to the reference animal. A
145 logistic regression model was developed using known positives (trio calls) and known false positives
146 (self-self hybridisation calls) to predict true CNVs in the wider dataset, with some further
147 downstream processing. The total number of autosomal segment calls predicted to represent true
148 CNVs by our model, using CGH data from 30 animals, was 12,802. After removing calls based on a
149 series of quality filters, a total of 9,789 autosomal CNV calls remained (Table 1). The mean absolute
150 \log_2 ratio of these calls was 0.54 and the average length was 30kb with a range in length of 1kb-
151 2.5Mb (Table 1).

152 On average, 326 CNVs were detected per individual, with a median of 321 and range of 109 to 643.
153 One animal had notably more CNV calls than the other animals, however, it had the same CNV
154 content on the autosomes (as a percentage of total length in base pairs) as the other animals.

155

156 *Autosomal CNVR*

157 CNV information from all animals was combined to obtain 3,488 CNV regions on the ovine
158 autosomes (Supplementary Table 2). The average length of these CNVRs was 19kb, with a range of
159 1kb to 3.6Mb. Shorter CNVRs were more frequent than longer CNVRs in the genome. The total
160 length of all CNVRs was 67.6Mbps, which equates to 2.7% of the sheep autosomes. For individuals,
161 this value ranged from 0.24 to 0.55%. Most CNVRs were seen in just one animal (Figure 1), however
162 1,424 (41%) were independently called in at least 2 individuals. A small percentage (0.11%) of
163 CNVRs were observed in all animals, which likely indicates the presence of a CNV in the reference
164 animal only - the 'reference effect' [46]. The majority of CNVRs (58%) contained only deletion CNVs,
165 38% of CNVRs contained only duplication CNVs and 4% were compound CNVRs, containing both
166 duplication and deletion CNVs.

167 The number of CNVRs on each chromosome ranged from 76 on chromosome 27 to 185 on
168 chromosome 19 (Figure 2). As can be seen in Figure 2, there was a weak positive linear relationship
169 between chromosome length and number of CNVRs ($R^2=0.27$).

170 The average spacing between CNVRs ranged from one every 347kbp on chromosome 19 to one
171 every 1.2Mb on chromosome 1. The closest CNVRs were approximately 1.5kb apart, while the
172 largest distance separating CNVRs was 8.5Mbps. The two-sample Kolmogorov-Smirnov test showed
173 that the distribution of the CNVRs in the genome (in terms of the inter-CNV distance) was
174 significantly different to that which would be expected should the CNVRs be uniformly distributed
175 ($p\text{-value} = 4.56 \times 10^{-7}$). Specifically, the CNVRs tended to be clustered together in the genome (Figure
176 3).

177 *Cross platform verification of autosomal CNVRs using 385K and SNP50 data*

178 A small subset of animals assayed with the 2.1M CGH array were also used for data generation with
179 either a lower density 385K CGH array (5 individuals) or the OvineSNP50 BeadChip (24 animals;
180 Supplementary Table 1). This facilitated an examination of the proportion of CNVRs independently
181 called across platforms. Using the 2.1 M CGH array, for the five reference animals a total of 935
182 CNVRs (1,268 CNV calls) were identified that could be mapped to genome BTA_OARv.2 for
183 comparison to Roche-NimbleGen 385K CGH array results. Of these, only 13 CNVRs (and 17 CNV
184 calls) had a corresponding segment call in the 385K CGH array dataset (Table 2). The average length
185 of verified CNVR was 387kb, much larger than the average of CNVRs that were not verified using the
186 385K CGH dataset (30kb). Possible explanations for the very low verification rate (1.4%) are provided
187 in the Discussion, but are likely to be caused in part by the differences in the probe density between
188 the two CGH arrays. This prompted the reverse comparison, whereby 52 CNV segment calls made
189 using the 385K CGH array (with absolute \log_2 ratio threshold of 0.25) were examined within the
190 larger 2.1 M CGH array CNV calls. Only 29% (15) of these calls overlapped CNVRs from the 2.1M CGH
191 array.

192

193 A separate comparison was performed against OvineSNP50 BeadChip data. A total of 2,847 CNVRs
194 were observed in the 24 animals common to both platforms (2.1M CGH array and OvineSNP50
195 BeadChip), arising from 7,416 CNVs. Of these, just three CNV calls (two CNVRs) overlapped CNVs
196 called by cnvPartition (Illumina Inc., USA) analysis of the SNP data (Table 2). CNVs predicted by the
197 DNACopy software [47] using Illumina Ovine SNP50 BeadChip data verified more CNVRs than
198 cnvPartition, with 101 CNVs corresponding to 64 CNVRs verified by DNACopy CNV calls (Table 2). The
199 three calls verified with the cnvPartition dataset were not verified by DNACopy.

200

201 *Cross platform verification of autosomal CNVRs using DNA sequence data and 2.1M CGH array data*
202 *in sheep*

203 The final comparison utilised analysis of whole genome sequence from the six reference panel
204 animals. Each individual was sequenced to between 9.8X and 14X genome wide coverage before
205 variation in read depth was used to detect CNVR (see Methods). The same six animals had 852
206 CNVRs arising from 1,164 CNV calls detected using the 2.1M CGH array. Comparing the CNV calls
207 revealed 61% of the Roche-NimbleGen 2.1M CGH array CNV calls were independently identified in
208 the sequence data (Table 2). Two thirds of the CNV calls that were verified were observed as a
209 consistent deletion or duplication CNV across platforms in a specific animal. The remaining verified
210 CNVs were observed as a CNV of the opposite type (deletion versus duplication) in the Poll Dorset
211 animal. This animal was used as the reference animal on the Roche-NimbleGen 2.1M CGH array and
212 therefore CNVs in this animal can be incorrectly observed as CNVs in the test animal when in fact no
213 CNV is present in the test animal. That is, a deletion in the Poll Dorset may be observed as a
214 duplication in the test animal on the 2.1M CGH array, while in the sequence data, the test animal
215 shows no CNV in the region but the Poll Dorset shows a deletion. The same is true for duplications in
216 the Poll Dorset, which will be observed as deletions in the test animal, even if no CNV is present in
217 the test animal in that region.

218 There were instances where the sequence data showed that there was a CNV in the Poll Dorset and
219 the test animal in the same region, but the type (duplication/deletion) of CNV in the test animal was
220 not consistent between the 2.1M CGH array and sequence platforms. For example, a 2.1M CGH
221 deletion that was observed as a duplication in the test and reference animal in the sequence data.
222 These calls were considered to be verified as there were still CNVs present in the sequence data and
223 it is possible that the magnitude of the \log_2 ratio of the CNV call on the 2.1M CGH array was higher in
224 the Poll Dorset than the test animal which could result in inconsistencies between the types of CNVs
225 detected. There were instances in the data where a CNV call of one particular CNV region could be
226 verified in one animal and not in another animal, which indicates that the CNV is likely present in
227 both animals but the sequence analysis failed to identify the CNV in one of the animals.

228

229 Significant differences in absolute \log_2 ratio, length and GC content were observed between the
230 sequence verified and non-verified 2.1M CGH array calls. Verified calls had higher absolute \log_2
231 ratios (0.62 versus 0.50) and were longer (46kb versus 9kb) on average than non-verified calls. This
232 suggests that longer calls with higher absolute \log_2 ratios are either more likely to represent true
233 CNVs or are easier to verify than shorter calls with lower absolute \log_2 ratios. Sequence
234 corresponding to non-verified calls showed significantly higher (two-tailed t-test for proportions) GC
235 content on average compared to verified calls – 44.6 versus 43.0%. Both verified and non-verified
236 calls had significantly higher GC content compared to the genome average (42.6%). More
237 duplications (72.4%) than deletions were verified on the sequence platform - 72.4% versus 54.7%.
238 This is not surprising, as there was less variation in the sequence data in regions with low read
239 depth, which reduces the ability to detect differences in copy number in these regions and hence
240 also CNVs relating to deletions.

241 *Comparison of autosomal CNVRs to those identified in the sheep and cattle literature*

242 In total, we detected 378 (18%) of the 2,154 CNVRs reported in seven other sheep and cattle studies.
243 Of the 2,154 CNVs detected in the seven other studies, 352 were present in more than one study.
244 We detected 132 (38%) of the 352 CNVs observed in multiple studies, whereas we only detected
245 14% of the CNVRs observed in just one other study (Table 3). The more frequently a CNVR was
246 observed in the other studies, the more likely we were to detect the CNVR (Table 3). We were able
247 to detect 31% of the CNVRs identified in the initial sheep study by Fontanesi *et al* [28] and between
248 16-62% of CNVRs detected in the cattle studies.

249 Eleven percent of the 3,336 CNVRs detected in this study and successfully mapped to the btau_4.0
250 genome overlapped CNVRs in these other studies. This is lower than would be expected based on
251 overlap between CNVRs from the other studies with each other, which ranges from 20-77%. By
252 comparison, 28% of the CNVRs from the sheep study by Fontanesi *et al* [28] were observed in at
253 least one of the cattle studies.

254 *Overlap between autosomal CNVRs and genes*

255 Of the 3,335 CNVRs identified on the Roche-NimbleGen 2.1M CGH array that mapped to OARv3
256 autosomes, 1,335 (40%) overlapped the coding sequence of one or more genes; 45% of duplication
257 CNVRs, 36% of deletion CNVRs and 59% of deletion/duplication CNVRs overlapped genes. The
258 proportion of duplications overlapping the coding sequence of genes was significantly different (Chi-
259 squared test, $p < 0.0001$) to the proportion of deletions overlapping genes. Based on permutation
260 analysis, these proportions were significantly greater than that which would be expected if the
261 CNVRs were randomly distributed in the genome ($p=0.01$). Both the agouti signalling protein and
262 adenosylhomocysteinase genes were overlapped by one of our CNVRs, which confirms the presence
263 of the agouti duplication reported by Norris and Whan [27] in this dataset, and thus provides a
264 positive control for the CNVR identification methods presented here. It is important to note that the
265 agouti duplication can be present in multiple copies [27], hence the reason that it shows up even
266 upon comparison to another white fleeced sheep.

267

268 *Non-autosomal CNVRs*

269 The total number of chromosome X Roche-NimbleGen 2.1M CGH array segment calls predicted to be
270 real was 697, however, 308 of these were observed as deletions in males. It is possible some of
271 these are real, particularly if they are present in the pseudo-autosomal region, however, this cannot
272 be confirmed in our analysis as we do not have a clear pseudo-autosomal boundary defined. After
273 filtering all 697 CNV calls based on size and \log_2 ratios, 615 of these were predicted to be real,
274 however, only 317 were either deletions or duplications in females or duplications in males. These
275 317 were used to call CNVRs on chromosome X. In total, we estimate there are at least 114 CNVRs
276 on chromosome X, representing approximately 3.2% of the length of the X chromosome. In addition
277 to chromosome X CNVRs, four CNVRs were identified on UMD3_OA_chrun, observed in one to ten
278 animals. These CNVRs spanned a total length of 19,304bps.

279

280 Including the 3,488 CNVRs observed on the autosomes, we estimate there to be approximately
281 3,606 CNVRs in the sheep genome. This includes CNVRs identified on chromosome X and
282 UMD3_OA_chrun. The total length of these 3,606 CNVRs is estimated to be 72.4Mbps, however, it is
283 possible that some of the CNVRs on UMD3_OA_chrun may overlap those identified on the
284 autosomes and therefore this number may be slightly lower.

285

286 **Discussion**

287 The results reported here provide a genome wide view of the frequency of CNV, an important class
288 of genomic variant that is currently poorly characterised in the sheep genome. Using a custom built
289 Roche-NimbleGen 2.1M CGH array, 9,789 autosomal CNVs were detected in 30 sheep. On average
290 these CNVs covered 0.4% of each animal's genome. This is higher than that reported in the initial
291 sheep survey where, on average, 0.05% of an individual sheep genome comprised CNVs [28]. The

292 difference in estimates is not surprising as this study used a CGH array with 2.1 million probes while
293 Fontanesi *et al* [28] used a CGH array with 385,000 probes. Based on probe spacing in the genome
294 and the filters applied to the data, the earlier study detected CNVs greater than 30kb in length, on
295 average, while this study had a resolution of ~4kb on average. As a result, differences in resolution
296 may have resulted in differences in the number of CNVs detected. This is reflected in the datasets,
297 with the average size of CNVs detected by Fontanesi *et al* [28] being 77.6kb (median 55.9kb) and the
298 average size detected in this study being 30.3kb (median 8.7kb). The individual genome CNV
299 composition estimates are similar to, but slightly lower than, estimates reported in humans (e.g.,
300 0.5%, [48]; 0.78%, [7]; and 1.2%, [8]).

301

302 The 9,789 autosomal CNVs reported in this study correspond to 3,488 autosomal CNV regions in the
303 30 animals tested, representing 2.7% of the sheep genome. This is approximately seven times higher
304 than estimated in the initial sheep survey [28], which is to be expected as more animals were
305 assayed in this study. This estimate is similar to the range of estimates in cattle [19, 25, 21, 26, 22,
306 23, 20] and again similar but slightly lower than estimates in humans (3.7%, [7]; 5%, [48]). Estimates
307 in humans are likely to provide a more accurate estimate of CNV composition in the genome, as
308 studies have involved more individuals and used a wider range of technologies, often employed
309 together. As in the Fontanesi *et al* [28] study, this study suffers from the lack of a complete
310 reference sheep genome. We used a sheep genome that was constructed using a cattle reference
311 genome to design probes for inclusion on the 2.1M CGH array. The genome used, UMD3_OA, does
312 not include any regions that are present in the sheep genome but that are not present in the cattle
313 genome. This means that sheep CNVs in regions deleted or of low homology in the cattle genome
314 are likely to have been undetected in this study. Future work will benefit from using a sheep
315 reference genome for CNV analysis.

316

317 There were also 118 CNVRs identified on chromosome X and chromosome unknown. However,
318 these were lower confidence calls and were not considered in further analyses. Of the 3,488
319 autosomal CNVRs identified in this study, 59% were observed in just one animal, which is
320 comparable to results in the literature [7, 35, 22, 23, 37]. One and a half times more deletions than
321 duplications were observed. This imbalance is one that is commonly reported in the literature [49,
322 50, 22] and could be due to ascertainment bias. The ascertainment bias arises because the
323 proportional difference between probe intensity of test and reference animals is greater for copy
324 number losses than gains meaning that deletions are easier to detect than duplications.

325

326 The CNVRs detected in this study tended to be clustered together in the genome. This may be an
327 artefact of the segMNT algorithm and our CNVR calling algorithm, which may have failed to collapse
328 multiple CNVRs originating from one CNVR into one region. However, similar distributions have
329 been reported in other studies [5, 51-53] and also for the closely related segmental duplication
330 variant [9]. If this clustering represents the true underlying distribution in the genome, then it may
331 indicate that the clustered CNVRs are the result of increased mutational activity in repetitive regions
332 of the genome which could facilitate mechanisms such as non-allelic homologous recombination
333 [54]. Determining if the CNVRs are a result of one mutational event or multiple mutational events
334 would require detailed analysis of specific regions, probably using deep sequencing.

335

336 There are reports in the literature that CNVRs are preferentially located outside of gene regions [51,
337 55, 56, 37] and that those CNVs that do overlap genes are more likely to be duplications than
338 deletions [7, 57, 37]. The rationale is that deletions are more disruptive to gene function than
339 duplications and therefore are subject to greater selective pressure. In this study, a significant
340 difference was observed in the proportion of duplications overlapping the coding sequence of genes
341 compared to deletions – 0.45 versus 0.36. However, both of these proportions were significantly
342 higher than would be expected if CNVRs were randomly distributed throughout the genome.

343 Therefore, in this study there is no evidence to suggest that the CNVRs identified in this study are
344 preferentially excluded from genic regions as has been suggested in the literature. Other results
345 reported in the literature have also found an enrichment of CNVs in these regions [30, 53]. Cooper *et*
346 *al* [53] suggest that CNVs that overlap segmental duplications (SDs) are more likely to be enriched in
347 genic regions, while CNVs that do not overlap SDs are enriched in gene poor regions of the genome.
348 As genes and segmental duplications are GC rich [58] and GC rich regions are more prone to CNV
349 formation, then it is possible that certain types of CNVs are enriched in genic regions. While
350 selection against or for CNVs and CNV formation mechanisms are reasonable explanations for the
351 depletion or enrichment of CNVs in genic regions, it is also possible that differences reported in the
352 literature are due to ascertainment bias introduced by using different methods for CNV detection.
353 Again, this illustrates the difficulties associated with CNV identification.

354

355 CNVRs were difficult to verify between CGH arrays and with the OvineSNP50 BeadChip. Partly, this
356 may be due to the fact that the CNVRs identified with the 2.1M CGH array had to be aligned to
357 different genomes for comparison with the 385K CGH array and OvineSNP50 BeadChip. There is also
358 evidence to suggest that SNPs on SNP chips are often biased away from CNV regions [20, 59, 60].
359 Also, SNPs that were included on the SNP array were identified using an earlier version of the sheep
360 genome (OARv1) than was used to design probes for the 2.1M CGH array. Therefore, there may be
361 fewer repetitive regions included in the OARv1 genome, which would add to the paucity of SNPs in
362 CNV regions. The average spacing of SNP probes in the sheep genome is one probe every 60kb,
363 which makes it difficult to detect small CNVs, even in regions where probes are spaced relatively
364 consistently, let alone in regions that have fewer SNPs. In combination, these factors may explain
365 the low cross platform verification rate observed with the Illumina OvineSNP50 BeadChip.

366

367 Whole genome sequencing exhibited the highest cross platform verification rate, with 61% of CNVs
368 verified with this platform. The CNVs that were unable to be verified were shorter and had lower

369 absolute \log_2 ratios than calls that were able to be verified. Both verified and non-verified CNVs had
370 significantly higher GC content than the genome average, which supports data from the literature
371 reporting that GC-rich regions can be more prone to CNV formation [61, 62]. Non-verified CNVs had
372 significantly higher GC content than verified CNVs. While it is possible that the non-verified CNVs
373 were false negatives in the sequence analysis, it is also possible that they were false positives in the
374 CGH dataset, as false positive CGH calls can be related to regions with high GC content [63, 64].
375 Future work could involve adjusting CGH intensity data for GC content.

376

377 This study detected 18% of the CNVRs reported in seven other sheep and cattle studies [19, 20, 25,
378 28, 21-23]. Thirty one percent of the CNVRs that were previously detected in an initial survey of
379 CNVs in the sheep genome [28] were detected in this study. We were able to identify all of the
380 CNVRs that were observed in six of the other studies, but only 14% of CNVRs observed in just one
381 other study. In fact, the more studies a CNVR was detected in, the more likely we were able to
382 identify the CNVR in our analysis. This trend was also reported by Kijas *et al* [22]. This suggests that
383 either these CNVRs are less likely to be false positives or they may be more common than the CNVRs
384 detected in just one study or, alternatively, they may be more likely to occur in both sheep and
385 cattle. Common CNVRs will be present in more individuals in the population and therefore are more
386 likely to be observed in the diverse range of animals tested in the different studies.

387

388 Although it is possible that a CNV has persisted since the common ancestor of sheep and cattle, it is
389 much more likely that repeated mutations at the same site cause a CNVR to be polymorphic in both
390 species. Microsatellites illustrate a similar phenomenon, with approximately 53% of cattle
391 microsatellite markers observed as polymorphic in sheep [45]. This is unlikely to be due to
392 counterbalancing selection maintaining the same alleles in both species, but instead due to repeated
393 and often recurrent mutations at the microsatellite locus maintaining it as polymorphic. The
394 phenomena depends on the high mutation rate of microsatellites - 1.1×10^{-4} per gamete per locus

395 [65] relative to the effective population size of the species, so that additional mutations at the locus
396 are created faster than they can be purged by genetic drift and purifying selection. The high
397 reported CNV mutation rates supports this hypothesis [3]. Another reason that some CNV regions
398 are common to both cattle and sheep could be that they are not necessarily CNVRs persistent since
399 the divergence of sheep and cattle, but are CNVs that have formed independently in sheep and
400 cattle individuals in conserved regions that are more predisposed to CNV formation. An example
401 would be a region containing segmental duplications that was present in the common ancestor
402 between sheep and cattle. It is possible that the CNVRs that are shared between sheep and cattle
403 may indicate that these CNVRs provide a selective advantage and are therefore maintained in both
404 sheep and cattle species. Drawing conclusions about this would require analysing the relevant
405 regions of cattle and sheep genomes to investigate the sequence diversity within the CNV alleles in
406 an attempt to estimate their age.

407

408 Reasons that this study was unable to detect many of the CNVs from the other studies include: CNVs
409 that occur in cattle but not sheep; rare CNVs not seen in our sample of sheep; and false negatives in
410 our study due in part to the different methods used for CNV detection. Similarly, only a small
411 number (11%) of CNVRs identified in this study overlapped CNVs detected in these seven other
412 studies. Again, lack of overlap could be due to the different species or individual animals tested,
413 different methods used for CNV detection, false negatives in other studies and false positives in our
414 dataset. Confirmation rates varied widely across the studies compared to our results. Variation in
415 confirmation rates from different studies has also been reported in the literature for human CNV
416 studies [66, 67].

417

418 **Conclusions**

419 In this study, comprehensive information from trios, multiple platforms and different algorithms
420 were used with the aim of verifying CNV segment calls from the Roche-NimbleGen 2.1M CGH array.
421 CNVs are difficult to verify and as is observed in the literature, a combination of approaches appears
422 to be the best way to accurately detect CNVs on a large scale. It is likely that comprehensive
423 sequencing or qPCR would provide clearer information about individual CNV regions and give an
424 indication of the accuracy of the methods used to detect them. Regardless, characteristics of the
425 CNV regions detected in this study are comparable to those reported in the literature, and the CNV
426 regions identified here add to the initial survey of CNVs in the sheep genome by Fontanesi *et al* [28].

427

428 **Methods**

429 *Roche-NimbleGen 2.1M CGH array - design overview*

430 In total, 2,012,210 probes (50-75 base pairs in length) were distributed evenly on non-repetitive
431 regions of the UMD3_OA ovine genome build (an in-house AgResearch comparative sheep genome
432 assembly, built using cattle reference genome UMD3 [68] and accessible at
433 www.sheephapmap.org/CNV/), with an average spacing of approximately one probe per 1,250 base
434 pairs (bps) on the autosomes and one probe per 1700bps on chromosome X. In addition to these
435 probes, a further set of probes was designed around SNPs found on the Illumina OvineSNP50
436 BeadChip, with the aim of increasing cross platform validation between the 2.1M CGH array and
437 OvineSNP50 BeadChip. This involved mapping SNPs and flanking sequence onto UMD3_OA. In
438 some instances, SNP sequences did not map uniquely to the genome, with multiple hits on the same
439 chromosome, suggesting the possibility that multiple copies of the sequence could occur in adjacent
440 duplicated regions (e.g. CNV). As these SNPs may have been in CNV regions, these regions were also
441 used for specific probe design and inclusion on the array. Probes were also designed on
442 chromosome unknown scaffolds. Chromosome unknown scaffolds represent sequence data that
443 cannot be placed on the genome assembly.

444

445 *Roche-NimbleGen 2.1M CGH array design - targeted probe design around OvineSNP50 BeadChip*

446 *SNPs*

447 In total, 28,754 out of 50,064 SNP sequences (either the 50bp OvineSNP50 BeadChip probe or 300bp
448 flanking the SNP) successfully mapped to UMD3_OA (BLAST parameters -U T -F "m D" -e 1e-5, Korf *et*
449 *al* [69]) and met the requirement of having three probes designed to cover them, as selected by one
450 of the following two methods (Figure 4). The first involved designing a probe to cover the SNP base
451 pair position. Flanking probes were designed within 400bp windows 100bp up- or down-stream of
452 the SNP region, where the SNP region consisted of 300bps flanking the SNP position. If three probes
453 were not obtained with this method, then a second method was used. This involved selecting a
454 probe in the SNP region without requiring the probe to cover the SNP position, with flanking probes
455 selected from 400bp windows 100bp up- or down-stream of the SNP region (Figure 4). In total,
456 86,262 probes were designed within or adjacent to 28,754 SNP regions.

457 Of the 21,310 SNP sequences that could not be mapped to UMD3_OA, 240 were mapped by relaxing
458 the BLAST parameters to -W 11 -q -1 -r 1 -s 0 -F "m D" -U T -X 40 [69]. A total of 634 probes were
459 designed to cover 218 of these SNP regions.

460 A subset of 401 SNP sequences mapped to UMD3_OA, but not uniquely - with two top hits on the
461 same chromosome. In total, 879 probes covering 323 of these positions were designed for inclusion
462 on the 2.1M CGH array.

463

464 *Roche-NimbleGen 2.1M CGH array design – chromosome unknown*

465 Chromosome unknown sequences (n=492) were merged into a virtual chromosome,
466 UMD3_chrU_OA, with each sequence separated by 100 N's. Probes were distributed at an average
467 spacing of approximately one every 1,600bps on this chromosome.

468

469 *Roche-NimbleGen 2.1M CGH array – animals assayed*

470 Genomic DNA was extracted from blood samples of 36 animals (Supplementary Table 1), which were
471 assayed on the 2.1M CGH array. Thirty animals were from the International Mapping Flock (IMF)
472 and consisted of families of trios (Figure 5). The IMF animals are crossbreeds of up to five different
473 breeds – Texel, Coopworth, Perendale, Romney and Merino [45]. In addition to the IMF animals, six
474 sheep, sequenced to approximately 10X coverage each, were also assayed on the 2.1M CGH array.
475 These six animals were - Awassi, Merino, Poll Dorset, Romney, Scottish Blackface and Texel
476 purebreds. The Poll Dorset was used as the reference animal for all 2.1M CGH array hybridisations
477 and was also run against itself in a self-self hybridisation to allow characterisation of false positive
478 calls [70, 23].

479

480 *Roche-NimbleGen 2.1M CGH array – segMNT output processing*

481 CNV segments were called in the assayed animals by Roche-NimbleGen using their proprietary
482 segMNT algorithm. This software reports the average \log_2 ratio of a segment (the binary logarithm of
483 the average of the intensity of the test animals probes in a segment call divided by the average of
484 the intensity of the reference animals probes in the same region), the number of datapoints (probes)
485 included in the segment and the length of the segment in base pairs.

486 The variance of normalised \log_2 ratio values over all probes for each animal was obtained. Five
487 animals were deleted from the analysis as their \log_2 ratio data exhibited larger variation than
488 observed in other animals, meaning that they were deemed to be failed CGH hybridisations.

489 Segment calls with absolute \log_2 ratios less than 0.1 were removed from the analysis [7].

490

491 *Validating Roche-NimbleGen 2.1M CGH array segment calls*

492 IMF trios were used to validate segment calls. If a progeny segment call was seen in at least one
493 parent at an identical genomic location (same first and last probe included in the segment call and
494 therefore same genomic start and stop position), the progeny call was considered validated. These
495 calls were deemed to represent “true CNVs” for model building.

496 *Model used to predict CNVs in the wider dataset and downstream filtering*

497 For model building, validated progeny calls were deemed to represent true CNVs and self-self
498 hybridisations were deemed to be false positives. Only autosomal segment calls were used. Forward
499 stepwise logistic regression was used to construct a model, with a binary outcome variable 0 (self-
500 self) or 1 (validated trio segment call). Variables used for model building were: absolute \log_2 ratio
501 (absl2r); whether the call was a deletion or duplication; length, in bps; $\ln(\text{length})$; $\ln(\ln(\text{length}))$;
502 length-squared; number of probes in segment call, datapoints; $\ln(\text{datapoints})$; $\ln(\ln(\text{datapoints}))$;
503 datapoints-squared; and corresponding two- and three-way interactions. If the Wald chi-square
504 statistic for a variable was significant at the 0.3 level it was added to the model. A variable remained
505 in the model if it was significant at the 0.35 level.

506

507 The crossvalidate procedure in SAS software (*SAS version 9.1*) was used to test model performance.
508 This procedure omits one segment call in turn and re-calculates model coefficients based on all
509 other segment calls per iteration. It then predicts the probability the omitted call represents a true
510 CNV. Threshold values were applied to categorise calls as true or false based on their probabilities –
511 true or false. Probability thresholds tested were 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 0.96, 0.97, 0.98 and 0.99.
512 For each probability threshold tested, the number of times the procedure correctly predicted the
513 known segment call status (true or false) was used as a measure of model accuracy. The final
514 probability threshold used was 0.95.

515

516 The final model selected was,

$$\ln\left(\frac{p}{(1-p)}\right) = -0.19 + 29.51absl2r - 4.91(\ln(\ln(length))) + 8.24(\ln(\ln(datapoints)))$$

517 This model was applied to all segment calls not used in model development. Segment calls equal to
518 or greater than the probability threshold of 0.95 were retained. The dataset was further filtered to
519 include only CNVs ≥ 1 kbp in length (so that they conformed to the definition of a CNV, as per [2],
520 only CNVs with ≥ 3 probes in the corresponding segment call and with absolute \log_2 ratio ≥ 0.25 .
521 These filtered segment calls were deemed to represent true CNVs.

522

523 Segment calls on chromosome X were processed through the model and filtered as above. Filtered
524 CNVs on chromosome X were considered to represent true CNVs for female individuals. Duplications
525 on chromosome X in males were considered to represent true CNVs. Deletions on chromosome X in
526 males were assumed to be inconclusive as they could be due to differences in the number of X
527 chromosomes between the male test animal and the female reference animal.

528

529 Segment calls on the virtual chromosome UMD3_chrU_OA were processed differently to segment
530 calls on the autosomes and chromosome X. Chromosome unknown sequences were collated into
531 larger virtual chromosomes, UMD3_chrU_OA, with each sequence separated by 100 N's. Segment
532 calls on this virtual chromosome were discarded if they spanned more than one chromosome
533 unknown sequence or if all probes on one chromosome unknown sequence were included in the
534 segment call. The reason for excluding segment calls where all probes on the chromosome unknown
535 sequence were included in the call was because there was no way to compare the call to nearby
536 sequence to determine if the \log_2 ratio was different to other stretches of DNA in the region. There
537 were two Poll Dorset (self-self hybridisation) segment calls on UMD3_chrU_OA. The \log_2 ratios of
538 these calls were -0.32 and -0.17. Thus calls with absolute \log_2 ratios ≤ 0.32 were removed from the
539 analysis. Segment calls that met these criteria and that contained at least two probes, while
540 excluding at least two probes from the corresponding chromosome unknown sequence, were
541 retained.

542

543 *CNV regions*

544 Across all animals, autosomal and chromosome X CNVs within 1,500bps of one another were
545 collapsed into CNV regions (CNVRs).

546

547 To determine if CNVRs were uniformly distributed in the genome, a simulated dataset of CNVRs was
548 generated by randomly sampling genomic positions of the identified autosomal CNVRs from a
549 uniform distribution. Spacing was constrained so that CNVRs could not be within 1,500bps of each
550 other. The simulated dataset provided an expected distribution of CNVRs in the genome and
551 corresponding pairwise distances between CNVRs. A Kolmogorov-Smirnov test was performed to
552 determine if the distribution of pairwise distances between CNVRs in the observed dataset was
553 significantly different from that seen in the simulated dataset.

554

555 *Verifying CNVRs across platforms*

556 Three other platforms were used for CNV identification – Roche-NimbleGen 385K CGH array,
557 OvineSNP50 BeadChip, and Illumina HiSeq 2000 sequence data analysis, with each based on a
558 different version of the ovine genome. To perform cross platform validation autosomal CNVRs
559 identified on the Roche-NimbleGen 2.1M CGH array were mapped to genomes BTA_OARv.2 (for use
560 with the 385K CGH array), OARv1 (for use with the OvineSNP50 BeadChip) and OARv3 (for use with
561 sequence data analysis). CNVR sequence and 1,750bps flanking the start and stop of each CNVR
562 were obtained. Sequences were masked with an ovine repeat database isgcandrebase2
563 (Supplementary file 1) and BLASTed against each genome, with parameters -F 'm D' -U T -Z 2000
564 [69]. CNVR start and stop positions on each genome were approximated based on the BLAST

565 alignment. When the predicted CNVR start position was a negative number, it was set to one (i.e.
566 the first base pair of the chromosome).

567

568 The Roche-NimbleGen 385K CGH array is based on the same technology as the Roche-NimbleGen
569 2.1M CGH array; however, it has fewer probes covering the genome, with a probe density of
570 approximately 1 probe per 6,000bps. Twenty animals were run on the 385K CGH array, including
571 five animals (Awassi, Merino, Romney, Scottish blackface and Texel) that were run on the 2.1M CGH
572 array. The Poll Dorset was used as a reference on the 385K CGH array and the 2.1M CGH array.
573 Autosomal CNVRs identified using the 2.1M CGH array were positioned on BTA_OARv.2 as described
574 above. CNVRs positioned on BTA_OARv.2 autosomes were retained for cross platform verification.
575 CNV segments called by the NimbleGen segMNT software in the 385K CGH dataset were processed
576 to include only autosomal segments with absolute \log_2 ratios ≥ 0.25 . Autosomal CNVRs in the five
577 animals were considered verified if there was overlap between their processed 385K CGH segment
578 calls and their 2.1M CGH array CNVR calls mapped to BTA_OARv.2. This comparison was performed
579 separately for each animal.

580

581 Twenty IMF and five sequenced animals had previously been genotyped on the OvineSNP50
582 BeadChip. SNP genotypes for these animals were run through the cnvPartition (Illumina Inc., USA)
583 and DNACopy [47] algorithms. DNACopy results were filtered to include only calls with absolute \log_2
584 ratios ≥ 0.25 . Autosomal CNVRs identified with the 2.1M CGH array and successfully mapped to
585 OARv1 autosomes were considered verified if they overlapped autosomal CNVs predicted by
586 cnvPartition or DNACopy, in the same animal.

587

588 Six animals assayed on the 2.1M CGH array were each sequenced to between 9.8X and 14X coverage
589 by paired-end sequencing on the Illumina HiSeq 2000 platform at Baylor College of Medicine. The

590 following analysis was carried out separately for each animal. Sequence reads were positioned on
591 ovine genome OARv3 using the Burrows-Wheeler Alignment (BWA) algorithm [71] and pileup files
592 [72] were used to retrieve read depth information at each base pair position on the autosomes.
593 Reads were portioned into 1kbp overlapping bins, excluding repetitive sequence, using a sliding
594 window of 200bps. Masked repetitive sequence positions were translated to genome build OARv3.
595 As well as excluding repetitive sequence, for each chromosome a maximum read depth was set per
596 chromosome to exclude potentially unmasked repeats from the CNV sequence analysis. The
597 maximum read depth threshold was set based on inspection of the read depth distribution function
598 with the aim of excluding outliers in read depth data. Bins with a maximum read depth exceeding
599 the threshold were deleted from the analysis. The average read depth over all base pairs was
600 determined for each bin after correcting for GC content based on methods presented by Yoon *et al*
601 [73].

602

603 Pseudo-Maximum likelihood was used to fit a mixture model to determine if the average read depth
604 for each bin represented a homozygous deletion (copy number, CN=0), heterozygous deletion
605 (CN=1), normal diploid copy number (2), heterozygous duplication (3) or homozygous duplication (4)
606 in the genome. The mixture model used (Table 4) was a mixture of four normal distributions (for
607 modeling CN = 1 to 4) and one half-normal distribution (for CN = 0). Constraints were placed on the
608 parameters of the normal distributions so that the means and variances of the distributions
609 corresponding to CN =1, 3 and 4 were equal to respectively 1/2, 3/2 and 2 times the mean and
610 variance of the distribution corresponding to CN = 2. Model fitting was done on a per chromosome
611 basis, using the R function *nlfminb* [74]. Specifically, seven parameters were estimated for each
612 chromosome: μ_2 and σ_2^2 , the mean and variance of read depth for a bin corresponding to CN = 2 (the
613 “normal” diploid copy number); σ_0^2 , the variance of read depth for a bin corresponding to CN = 0
614 (homozygous deletion) and four of the five mixture weights (prior probability of a bin falling into
615 each of the five distributions). Where these parameters could not be estimated for a chromosome,

616 average estimates based on all other chromosomes for a given animal were used. Table 5 details the
617 starting values and lower and upper bounds used by *nlimb* for each parameter. Based on those
618 parameter estimates, each bin was assigned to one of the five CNV classes by multiplying the values
619 of each of the five probability density functions for each bin by the corresponding mixture weights
620 (i.e., calculating the posterior probability of a bin being in each of the distributions) and selecting the
621 CNV class with the highest value. For each of the six animals, bins in regions corresponding to
622 autosomal CNVRs identified on the 2.1M CGH array and mapped to OARv3 autosomes were used to
623 determine if the CNVR was verified in the sequence data. In instances where there was conflict
624 between results from the sequence analysis and the 2.1M CGH array, individual animal data were
625 compared to the reference (Poll Dorset) animal. This animal was used as the reference animal in the
626 2.1M CGH array experiments and therefore results for individual animals may be influenced by the
627 corresponding copy number present in the Poll Dorset.

628 *Comparison of CNVRs to those identified in the literature*

629 CNVR sequences were masked against AgResearch ovine repeat database *isgcandrepbase2* and
630 BLASTed against *btau_4.0* using BLAST parameters -F 'm D' -U T -Z 2000 [69] to obtain their positions
631 on the genome. Genomic positions on *btau_4.0* of CNVs identified from seven other sheep and
632 cattle studies [28, 21-23, 25, 19, 20] were obtained. An overlap of 1bp or more between autosomal
633 CNVRs identified in this study and these seven other studies was used to give an indication as to how
634 many CNVs from other studies we were able to detect and how many of the CNVs detected in this
635 study were also reported in the other studies.

636 *Overlap between autosomal CNVRs and genes*

637 CNVR sequences were masked (*isgcandrepbase2*) and BLASTed (parameters -F 'm D' -U T -Z 2000)
638 against OARv3 to obtain their positions on the genome. Positions of the coding sequence of genes
639 on OARv3 were provided by BGI (personal communication, Rudiger Brauning). Overlap between
640 autosomal CNVRs and the coding sequence of genes were determined. CNVRs that overlapped gene

641 coding sequences by 1bp or more were used to derive the proportion of CNVRs overlapping genes.
642 Overlap with the agouti signalling protein and adenosylhomocysteinase genes were used as a
643 positive control, as this locus is observed as duplicated in the sheep genome [27].

644

645 A Monte Carlo simulation was set up to randomly distribute the CNVRs throughout the sheep
646 genome and to create a distribution of the expected proportion of deletion CNVRs and duplication
647 CNVRs overlapping genes (by at least 1bp). One hundred iterations were run to generate 100
648 expected proportions for both duplications and deletions. For both duplication and deletion CNVRs,
649 the observed proportion was ranked along with the 100 simulated proportions and a two-tailed
650 empirical p-value was calculated.

651

652 **Competing interests**

653 The authors declare that they have no competing interests.

654

655

656 **Authors' contributions**

657 GMJ carried out the research and wrote the manuscript. MEG, MAB and JCM participated in study
658 design, provided input on analysis and revised the manuscript. JWK provided access to data and
659 revised the manuscript. KGD consulted on statistical analysis. BA was involved in sequence analysis.
660 RB carried out and advised on bioinformatic processes and was involved in the design of the Roche
661 NimbleGen 2.1M CGH array. NC revised the manuscript. All authors read and approved the final
662 manuscript.

663

664 Acknowledgments

665 The authors wish to acknowledge the financial contribution made by Ovita Limited and USDA grant
666 AFRI 2009-03305 for providing funding for a large part of this work. The authors also wish to
667 acknowledge James Kijas for providing access to International Sheep Genomics Consortium sheep
668 samples, whole genome sequence from these samples and early access to genome assemblies.

669

670 References

- 671 1. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N et al. Relative impact of
672 nucleotide and copy number variation on gene expression phenotypes. *Science*.
673 2007;315(5813):848-53. doi:315/5813/848 [pii]
674 10.1126/science.1136678.
- 675 2. Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. *Nat Rev Genet*.
676 2006;7(2):85-97. doi:http://www.nature.com/nrg/journal/v7/n2/supinfo/nrg1767_S1.html.
- 677 3. Egan CM, Sridhar S, Wigler M, Hall IM. Recurrent DNA copy number variation in the laboratory
678 mouse. *Nat Genet*. 2007;39(11):1384-9. doi:10.1038/ng.2007.19.
- 679 4. Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA et al. A map of human genome
680 variation from population-scale sequencing. *Nature*. 2010;467(7319):1061-73. doi:nature09534 [pii]
681 10.1038/nature09534.
- 682 5. Michaelson Jacob J, Shi Y, Gujral M, Zheng H, Malhotra D, Jin X et al. Whole-Genome Sequencing
683 in Autism Identifies Hot Spots for De Novo Germline Mutation. *Cell*. 2012;151(7):1431-42.
684 doi:<http://dx.doi.org/10.1016/j.cell.2012.11.019>.
- 685 6. Stankiewicz P, Lupski JR. Structural variation in the human genome and its role in disease. *Annu*
686 *Rev Med*. 2010;61:437-55. doi:10.1146/annurev-med-100708-204735.
- 687 7. Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y et al. Origins and functional impact of
688 copy number variation in the human genome. *Nature*. 2010;464(7289):704-12.
689 doi:http://www.nature.com/nature/journal/v464/n7289/supinfo/nature08516_S1.html.
- 690 8. Pang A, MacDonald J, Pinto D, Wei J, Rafiq M, Conrad D et al. Towards a comprehensive structural
691 variation map of an individual human genome. *Genome Biology*. 2010;11(5):R52.
- 692 9. Kim PM, Lam HYK, Urban AE, Korbelt JO, Affourtit J, Grubert F et al. Analysis of copy number
693 variants and segmental duplications in the human genome: Evidence for a change in the process of
694 formation in recent evolutionary history. *Genome Res*. 2008;18(12):1865-74.
695 doi:10.1101/gr.081422.108.
- 696 10. Long J, Delahanty RJ, Li G, Gao YT, Lu W, Cai Q et al. A Common Deletion in the APOBEC3 Genes
697 and Breast Cancer Risk. *J Natl Cancer Inst*. 2013;105(8):573-9. doi:djt018 [pii]
698 10.1093/jnci/djt018.
- 699 11. Yang L, Liu B, Huang B, Deng J, Li H, Yu B et al. A functional copy number variation in the WWOX
700 gene is associated with lung cancer risk in Chinese. *Hum Mol Genet*. 2013;22(9):1886-94. doi:ddt019
701 [pii]
702 10.1093/hmg/ddt019.

- 703 12. Suehiro Y, Okada T, Shikamoto N, Zhan Y, Sakai K, Okayama N et al. Germline copy number
704 variations associated with breast cancer susceptibility in a Japanese population. *Tumour Biol.*
705 2012;34(2):947-52. doi:10.1007/s13277-012-0630-x.
- 706 13. Fanciulli M, Norsworthy PJ, Petretto E, Dong R, Harper L, Kamesh L et al. FCGR3B copy number
707 variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. *Nat*
708 *Genet.* 2007;39(6):721-3. doi:ng2046 [pii]
709 10.1038/ng2046.
- 710 14. Gonzalez E, Kulkarni H, Bolivar H, Mangano A, Sanchez R, Catano G et al. The influence of CCL3L1
711 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science.* 2005;307(5714):1434-
712 40. doi:1101160 [pii]
713 10.1126/science.1101160.
- 714 15. Salmon Hillbertz NH, Isaksson M, Karlsson EK, Hellmen E, Pielberg GR, Savolainen P et al.
715 Duplication of FGF3, FGF4, FGF19 and ORAOV1 causes hair ridge and predisposition to dermoid sinus
716 in Ridgeback dogs. *Nat Genet.* 2007;39(11):1318-20. doi:ng.2007.4 [pii]
717 10.1038/ng.2007.4.
- 718 16. Wright D, Boije H, Meadows JR, Bed'hom B, Gourichon D, Vieaud A et al. Copy number variation
719 in intron 1 of SOX5 causes the Pea-comb phenotype in chickens. *PLoS Genet.* 2009;5(6):e1000512.
720 doi:10.1371/journal.pgen.1000512.
- 721 17. Chen C, Qiao R, Wei R, Guo Y, Ai H, Ma J et al. A comprehensive survey of copy number variation
722 in 18 diverse pig populations and identification of candidate copy number variable genes associated
723 with complex traits. *BMC Genomics.* 2012;13:733. doi:1471-2164-13-733 [pii]
724 10.1186/1471-2164-13-733.
- 725 18. Johansson Moller M, Chaudhary R, Hellmen E, Hoyheim B, Chowdhary B, Andersson L. Pigs with
726 the dominant white coat color phenotype carry a duplication of the KIT gene encoding the
727 mast/stem cell growth factor receptor. *Mamm Genome.* 1996;7(11):822-30.
- 728 19. Bae J, Cheong H, Kim L, NamGung S, Park T, Chun J-Y et al. Identification of copy number
729 variations and common deletion polymorphisms in cattle. *BMC Genomics.* 2010;11(1):232.
- 730 20. Bickhart DM, Hou Y, Schroeder SG, Alkan C, Cardone MF, Matukumalli LK et al. Copy number
731 variation of individual cattle genomes using next-generation sequencing. *Genome Res.*
732 2012;22(4):778-90. doi:10.1101/gr.133967.111.
- 733 21. Hou Y, Liu GE, Bickhart DM, Cardone MF, Wang K, Kim ES et al. Genomic characteristics of cattle
734 copy number variations. *BMC Genomics.* 2011;12:127. doi:1471-2164-12-127 [pii]
735 10.1186/1471-2164-12-127.
- 736 22. Kijas JW, Barendse W, Barris W, Harrison B, McCulloch R, McWilliam S et al. Analysis of copy
737 number variants in the cattle genome. *Gene.* 2011;482(1-2):73-7. doi:S0378-1119(11)00179-X [pii]
738 10.1016/j.gene.2011.04.011.
- 739 23. Liu GE, Hou Y, Zhu B, Cardone MF, Jiang L, Cellamare A et al. Analysis of copy number variations
740 among diverse cattle breeds. *Genome Res.* 2010;20(5):693-703. doi:10.1101/gr.105403.110.
- 741 24. Cicconardi F, Chillemi G, Tramontano A, Marchitelli C, Valentini A, Ajmone-Marsan P et al.
742 Massive screening of copy number population-scale variation in *Bos taurus* genome. *BMC Genomics.*
743 2013;14(1):124. doi:1471-2164-14-124 [pii]
744 10.1186/1471-2164-14-124.
- 745 25. Fadista J, Thomsen B, Holm LE, Bendixen C. Copy number variation in the bovine genome. *BMC*
746 *Genomics.* 2010;11:284. doi:1471-2164-11-284 [pii]
747 10.1186/1471-2164-11-284.

- 748 26. Jiang L, Jiang J, Yang J, Liu X, Wang J, Wang H et al. Genome-wide detection of copy number
749 variations using high-density SNP genotyping platforms in Holsteins. *BMC Genomics*. 2013;14(1):131.
750 doi:1471-2164-14-131 [pii]
751 10.1186/1471-2164-14-131.
- 752 27. Norris BJ, Whan VA. A gene duplication affecting expression of the ovine ASIP gene is responsible
753 for white and black sheep. *Genome Res*. 2008;18(8):1282-93. doi:gr.072090.107 [pii]
754 10.1101/gr.072090.107.
- 755 28. Fontanesi L, Beretti F, Martelli PL, Colombo M, Dall'Olio S, Occidente M et al. A first comparative
756 map of copy number variations in the sheep genome. *Genomics*. 2011;97(3):158-65.
757 doi:10.1016/j.ygeno.2010.11.005.
- 758 29. Barrett MT, Scheffer A, Ben-Dor A, Sampas N, Lipson D, Kincaid R et al. Comparative genomic
759 hybridization using oligonucleotide microarrays and total genomic DNA. *Proc Natl Acad Sci U S A*.
760 2004;101(51):17765-70. doi:0407979101 [pii]
761 10.1073/pnas.0407979101.
- 762 30. Graubert TA, Cahan P, Edwin D, Selzer RR, Richmond TA, Eis PS et al. A high-resolution map of
763 segmental DNA copy number variation in the mouse genome. *PLoS Genet*. 2007;3(1):e3. doi:06-
764 PLGE-RA-0282R3 [pii]
765 10.1371/journal.pgen.0030003.
- 766 31. Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F et al. Comparative
767 genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science*.
768 1992;258(5083):818-21.
- 769 32. Pinkel D, Seagraves R, Sudar D, Clark S, Poole I, Kowbel D et al. High resolution analysis of DNA
770 copy number variation using comparative genomic hybridization to microarrays. *Nat Genet*.
771 1998;20(2):207-11.
- 772 33. Yu H, Chao J, Patek D, Mujumdar R, Mujumdar S, Waggoner AS. Cyanine dye dUTP analogs for
773 enzymatic labeling of DNA probes. *Nucleic Acids Res*. 1994;22(15):3226-32.
- 774 34. Jeon JP, Shim SM, Jung JS, Nam HY, Lee HJ, Oh BS et al. A comprehensive profile of DNA copy
775 number variations in a Korean population: identification of copy number invariant regions among
776 Koreans. *Exp Mol Med*. 2009;41(9):618-28. doi:10.3858/emm.2009.41.9.068.
- 777 35. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T et al. Mapping and
778 sequencing of structural variation from eight human genomes. *Nature*. 2008;453(7191):56-64.
779 doi:http://www.nature.com/nature/journal/v453/n7191/supinfo/nature06862_S1.html.
- 780 36. McCarroll SA, Hadnott TN, Perry GH, Sabeti PC, Zody MC, Barrett JC et al. Common deletion
781 polymorphisms in the human genome. *Nat Genet*. 2006;38(1):86-92.
- 782 37. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD et al. Global variation in copy
783 number in the human genome. *Nature*. 2006;444(7118):444-54.
784 doi:http://www.nature.com/nature/journal/v444/n7118/supinfo/nature05329_S1.html.
- 785 38. Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: an approach to discover, genotype, and
786 characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res*.
787 2011;21(6):974-84. doi:gr.114876.110 [pii]
788 10.1101/gr.114876.110.
- 789 39. Teo SM, Pawitan Y, Ku CS, Chia KS, Salim A. Statistical challenges associated with detecting copy
790 number variations with next-generation sequencing. *Bioinformatics*. 2012;28(21):2711-8.
791 doi:10.1093/bioinformatics/bts535.
- 792 40. Xi R, Hadjipanayis AG, Luquette LJ, Kim TM, Lee E, Zhang J et al. Copy number variation detection
793 in whole-genome sequencing data using the Bayesian information criterion. *Proc Natl Acad Sci U S A*.
794 2011;108(46):E1128-36. doi:1110574108 [pii]
795 10.1073/pnas.1110574108.

- 796 41. Xi R, Lee S, Park PJ. A survey of copy-number variation detection tools based on high-throughput
797 sequencing data. *Curr Protoc Hum Genet.* 2012;Chapter 7:Unit7 19.
798 doi:10.1002/0471142905.hg0719s75.
- 799 42. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: a pattern growth approach to detect break
800 points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics.*
801 2009;25(21):2865-71. doi:btp394 [pii]
802 10.1093/bioinformatics/btp394.
- 803 43. Savarese M, Piluso G, Orteschi D, Di Fruscio G, Dionisi M, Blanco FdV et al. Enhancer Chip:
804 Detecting Human Copy Number Variations in Regulatory Elements. *PLoS ONE.* 2012;7(12):e52264.
805 doi:10.1371/journal.pone.0052264.
- 806 44. Krumm N, Sudmant PH, Ko A, O'Roak BJ, Malig M, Coe BP et al. Copy number variation detection
807 and genotyping from exome sequence data. *Genome Res.* 2012;22(8):1525-32.
808 doi:10.1101/gr.138115.112.
- 809 45. Crawford AM, Dodds KG, Ede AJ, Pierson CA, Montgomery GW, Garmonsway HG et al. An
810 autosomal genetic linkage map of the sheep genome. *Genetics.* 1995;140(2):703-24.
- 811 46. Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Altshuler DM et al. Copy number variation:
812 new insights in genome diversity. *Genome Res.* 2006;16(8):949-61. doi:gr.3677206 [pii]
813 10.1101/gr.3677206.
- 814 47. Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of
815 array-based DNA copy number data. *Biostatistics.* 2004;5(4):557-72.
816 doi:10.1093/biostatistics/kxh008.
- 817 48. McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, Wysoker A et al. Integrated detection
818 and population-genetic analysis of SNPs and copy number variation. *Nat Genet.* 2008;40(10):1166-
819 74. doi:ng.238 [pii]
820 10.1038/ng.238.
- 821 49. Craddock N, Hurles ME, Cardin N, Pearson RD, Plagnol V, Robson S et al. Genome-wide
822 association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls.
823 *Nature.* 2010;464(7289):713-20. doi:nature08979 [pii]
824 10.1038/nature08979.
- 825 50. Li Y, Mei S, Zhang X, Peng X, Liu G, Tao H et al. Identification of genome-wide copy number
826 variations among diverse pig breeds by array CGH. *BMC Genomics.* 2012;13:725. doi:1471-2164-13-
827 725 [pii]
828 10.1186/1471-2164-13-725.
- 829 51. Berglund J, Nevalainen E, Molin A-M, Perloski M, Consortium TL, Andre C et al. Novel origins of
830 copy number variation in the dog genome. *Genome Biology.* 2012;13(8):R73.
- 831 52. She X, Cheng Z, Zollner S, Church DM, Eichler EE. Mouse segmental duplication and copy number
832 variation. *Nat Genet.* 2008;40(7):909-14.
833 doi:http://www.nature.com/ng/journal/v40/n7/supplinfo/ng.172_S1.html.
- 834 53. Cooper GM, Nickerson DA, Eichler EE. Mutational and selective effects on copy-number variants
835 in the human genome. *Nat Genet.* 2007;39(7 Suppl):S22-9. doi:ng2054 [pii]
836 10.1038/ng2054.
- 837 54. Stankiewicz P, Lupski JR. Genome architecture, rearrangements and genomic disorders. *Trends*
838 *Genet.* 2002;18(2):74-82. doi:S0168-9525(02)02592-1 [pii].
- 839 55. Zhang F, Gu W, Hurles ME, Lupski JR. Copy number variation in human health, disease, and
840 evolution. *Annu Rev Genomics Hum Genet.* 2009;10:451-81.
841 doi:10.1146/annurev.genom.9.081307.164217.
- 842 56. Conrad DF, Andrews TD, Carter NP, Hurles ME, Pritchard JK. A high-resolution survey of deletion
843 polymorphism in the human genome. *Nat Genet.* 2006;38(1):75-81. doi:ng1697 [pii]

- 844 10.1038/ng1697.
845 57. Emerson JJ, Cardoso-Moreira M, Borevitz JO, Long M. Natural Selection Shapes Genome-Wide
846 Patterns of Copy-Number Polymorphism in *Drosophila melanogaster*. *Science*. 2008;320(5883):1629-
847 31. doi:10.1126/science.1158078.
848 58. Jurka J, Kohany O, Pavlicek A, Kapitonov VV, Jurka MV. Duplication, coclustering, and selection of
849 human Alu retrotransposons. *Proceedings of the National Academy of Sciences of the United States*
850 *of America*. 2004;101(5):1268-72. doi:10.1073/pnas.0308084100.
851 59. Cooper GM, Zerr T, Kidd JM, Eichler EE, Nickerson DA. Systematic assessment of copy number
852 variant detection via genome-wide SNP genotyping. *Nat Genet*. 2008;40(10):1199-203. doi:ng.236
853 [pii]
854 10.1038/ng.236.
855 60. Locke DP, Sharp AJ, McCarroll SA, McGrath SD, Newman TL, Cheng Z et al. Linkage disequilibrium
856 and heritability of copy-number polymorphisms within duplicated regions of the human genome.
857 *Am J Hum Genet*. 2006;79(2):275-90. doi:S0002-9297(07)63135-8 [pii]
858 10.1086/505653.
859 61. Cooper DN, Bacolla A, Férec C, Vasquez KM, Kehrer-Sawatzki H, Chen J-M. On the sequence-
860 directed nature of human gene mutation: The role of genomic architecture and the local DNA
861 sequence environment in mediating gene mutations underlying human inherited disease. *Hum*
862 *Mutat*. 2011;32(10):1075-99. doi:10.1002/humu.21557.
863 62. Nguyen DQ, Webber C, Hehir-Kwa J, Pfundt R, Veltman J, Ponting CP. Reduced purifying selection
864 prevails over positive selection in human copy number variant evolution. *Genome Res*.
865 2008;18(11):1711-23. doi:gr.077289.108 [pii]
866 10.1101/gr.077289.108.
867 63. Liu G, Hou Y, Robl J, Kuroiwa Y, Wang Z. Assessment of genome integrity with array CGH in cattle
868 transgenic cell lines produced by homologous recombination and somatic cell cloning. *Genome*
869 *Integrity*. 2011;2(1):6.
870 64. Marioni J, Thorne N, Valsesia A, Fitzgerald T, Redon R, Fiegler H et al. Breaking the waves:
871 improved detection of copy number variation from microarray-based comparative genomic
872 hybridization. *Genome Biology*. 2007;8(10):R228.
873 65. Crawford AM, Cuthbertson RP. Mutations in sheep microsatellites. *Genome Res*. 1996;6(9):876-
874 9.
875 66. Winchester L, Yau C, Ragoussis J. Comparing CNV detection methods for SNP arrays. *Brief Funct*
876 *Genomic Proteomic*. 2009;8(5):353-66. doi:elp017 [pii]
877 10.1093/bfgp/elp017.
878 67. Pinto D, Darvishi K, Shi X, Rajan D, Rigler D, Fitzgerald T et al. Comprehensive assessment of
879 array-based platforms and calling algorithms for detection of copy number variants. *Nat Biotech*.
880 2011;29(6):512-20.
881 doi:<http://www.nature.com/nbt/journal/v29/n6/abs/nbt.1852.html#supplementary-information>.
882 68. Jenkins G. UMD3_OA assembly. 2015. http://www.sheepmap.org/CNV/UMD3_OA.tgz.
883 Accessed 1 June 2015 2015.
884 69. Korf I, Yandell M, Bedell J. BLAST. O'Reilly Media; 2003.
885 70. Fontanesi L, Martelli P, Beretti F, Riggio V, Dall'Olio S, Colombo M et al. An initial comparative
886 map of copy number variations in the goat (*Capra hircus*) genome. *BMC Genomics*. 2010;11(1):639.
887 71. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform.
888 *Bioinformatics*. 2009;25(14):1754-60. doi:10.1093/bioinformatics/btp324.
889 72. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N et al. The Sequence Alignment/Map
890 format and SAMtools. *Bioinformatics*. 2009;25(16):2078-9. doi:10.1093/bioinformatics/btp352.
891 73. Yoon S, Xuan Z, Makarov V, Ye K, Sebat J. Sensitive and accurate detection of copy number
892 variants using read depth of coverage. *Genome Res*. 2009;19(9):1586-92. doi:gr.092981.109 [pii]

893 10.1101/gr.092981.109.

894 74. Gay DM. Computing Science Technical Report No.153: Usage summary for selected optimization
895 routines. 1990.

896

897

898

899 **Table 1.** Characteristics of CNVs predicted true by the model (n=9,789) and filtered to remove
 900 artefacts.

Variable	Mean	Median	Std dev	Min	Max
absl2r*	0.54	0.43	0.36	0.25	3.47
length (bp)	30,332.02	8,706	107,369.37	1,003	2,522,449
Datapoints [#]	14.99	9	23.91	3	446

901 *absl2r is the absolute log₂ ratio of the CNV. # number of CGH array probes in the CNV.

902 **Table 2.** Cross platform verification results. Number of CNV calls that were verified and not verified.

	Verification platform			
	385K CGH array	Illumina OvineSNP50 BeadChip - cnvPartition	Illumina OvineSNP50 BeadChip - DNACopy	Sequence analysis ~ 10X coverage
Verified	17 (1.34%)	3 (0.04%)	101 (1.36%)	714 (61.34%)
Not verified	1,251	7,413	7,315	450
Total	1,268	7,416	7,416	1,164

903

904 **Table 3.** Comparison between CNVRs observed in this study and CNVRs observed in the literature.

Number of studies CNVR observed in	Number of CNVR	Number of these CNVR identified in this study (%)
1	1,802	246 (13.7)
2	255	82 (32.2)
3	66	24 (36.4)
4	20	16 (80.0)
5	7	6 (85.7)
6	4	4 (100)

905

906 **Table 4.** Description of the pseudo-maximum likelihood derived mixture model for estimating copy

907 number in sequence data.

Copy number	Distribution	Mixture weights	Mean	Variance
0	Half normal, centered on zero	π_0	$\sqrt{2\pi}\sigma_0$	σ_0^2
1	Normal	π_1	$\frac{1}{2}\mu_2$	$\frac{1}{2}\sigma_2^2$
2	Normal	$1 - \pi_0 - \pi_1 - \pi_3 - \pi_4$	μ_2	σ_2^2
3	Normal	π_3	$\frac{3}{2}\mu_2$	$\frac{3}{2}\sigma_2^2$
4	Normal	π_4	$\frac{4}{2}\mu_2$	$\frac{4}{2}\sigma_2^2$

908

909

910 **Table 5.** Starting values of parameters estimated by pseudo-maximum likelihood.

Variable	Starting value	Lower bound	Upper bound
μ_2	$\bar{\mu}$	$-\infty$	∞
σ_2^2	$\bar{\sigma}_2^2$	0	∞
σ_0^2	1.5	0.01	∞
π_0	0.01	0	0.05
π_1	0.025	0	0.2
π_3	0.001	0	0.2
π_4	0.001	0	0.05

911

912

913

914 **Figure 1. CNVR frequency across animals.**

915 **Figure 2. Number of CNVRs by chromosome length.** Labels correspond to chromosome number.

916 **Figure 3. Cumulative density plot of the distances separating CNVRs.** The red line reflects the
917 observed pairwise distances between CNVRs, while the blue line reflects the simulated (expected if
918 CNVRs are uniformly distributed in the genome) distances separating CNVRs.

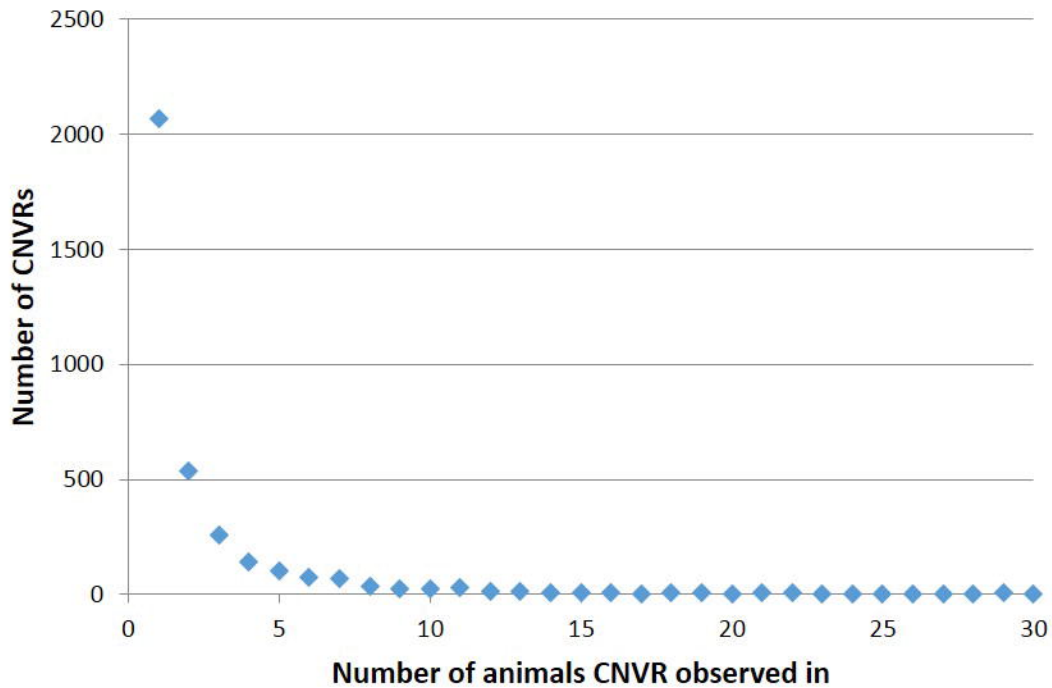
919 **Figure 4. Selection of CGH array probes to cover OvineSNP50 BeadChip SNP positions.** Two
920 methods were used to select probe sets to cover SNPs. The first method (a) involved designing at
921 least one probe to cover the SNP position, with two probes in flanking regions. The second method
922 (b) involved designing a probe within the 300bp region surrounding the SNP and two probes in
923 flanking regions.

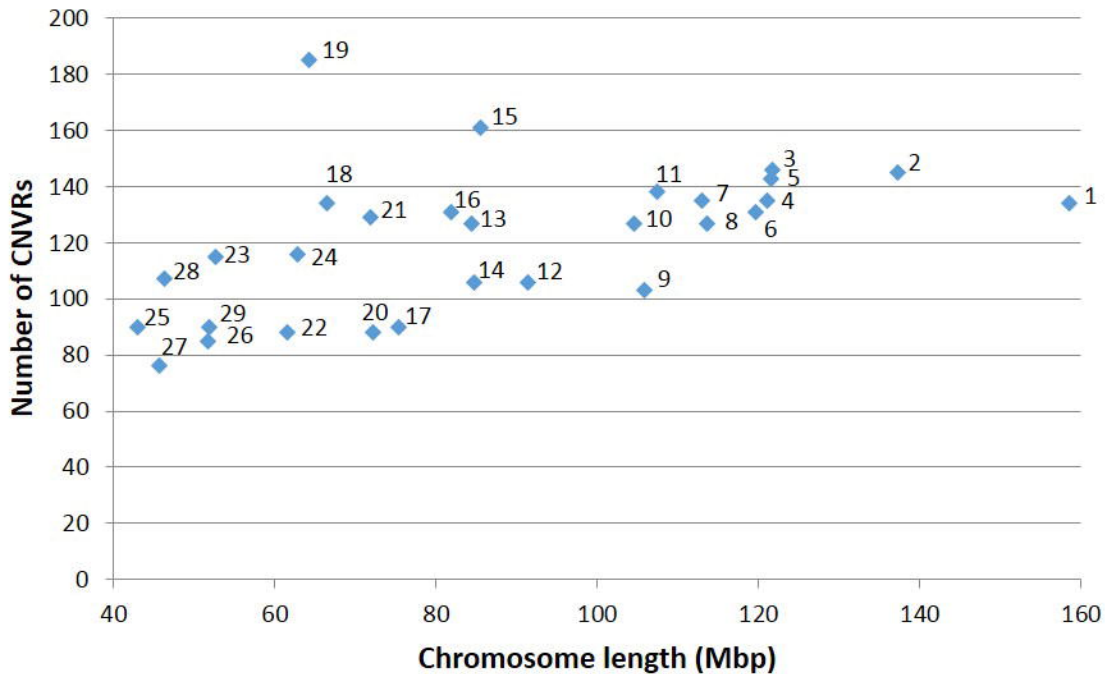
924

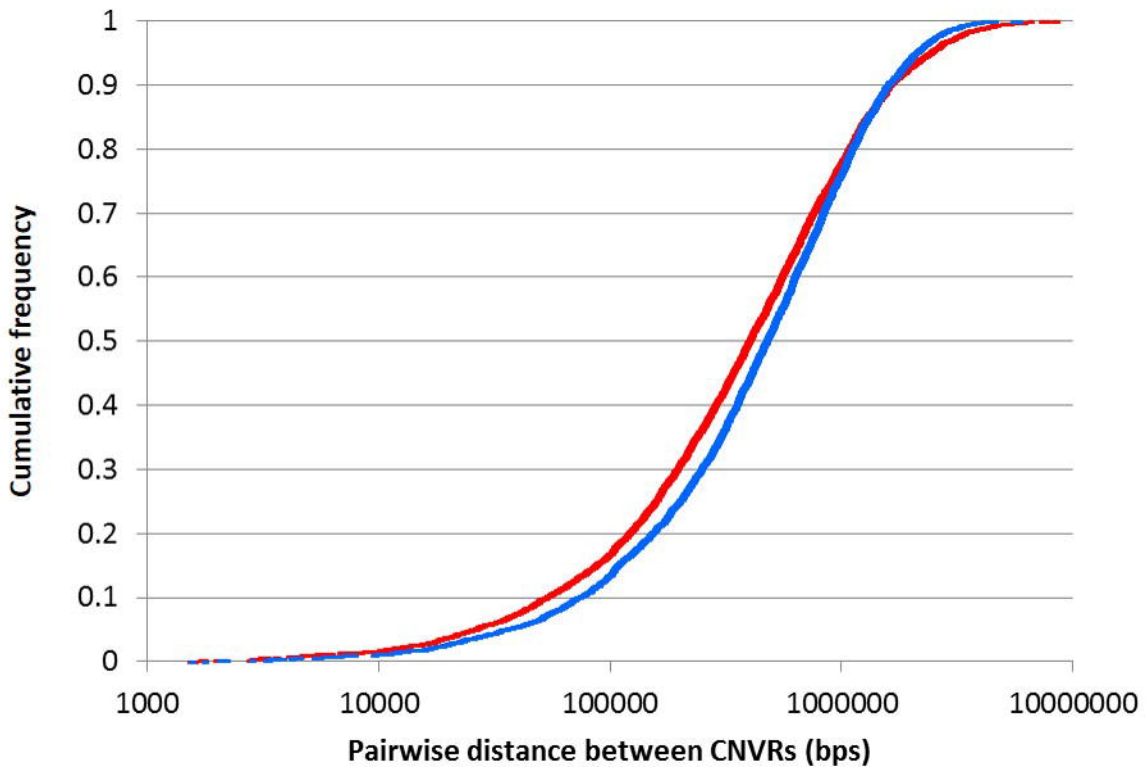
925 **Figure 5. Pedigree of International Mapping Flock (IMF) animals assayed on the Roche NimbleGen**
926 **2.1M CGH array.** Some animals (green) appear in more than one pedigree. Segment calls from
927 animals IMF66, IMF91, IMF95, IMF108 and IMF112 (red) were removed from the analysis due to
928 failed 2.1M CGH arrays.

929

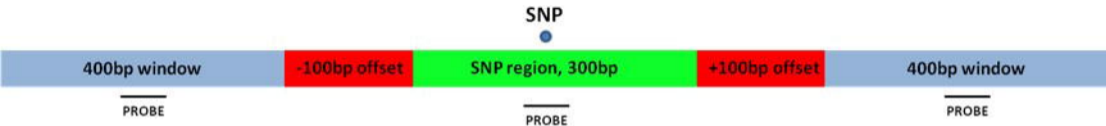
930







2a. Probe selection – method one



2b. Probe selection – method two

