Copy number variants in the sheep genome detected using multiple approaches

Gemma M Jenkins [1]*
*corresponding author
Email: gjenkins@abacusbio.co.nz

Michael E Goddard [2]
Email: mike.goddard@ecodev.vic.gov.au

Michael A Black [3]
Email: mik.black@otago.ac.nz

Rudiger Brauning [4]
Email: rudiger.brauning@agresearch.co.nz

Benoit Auvray [3]
Email: bauvray@maths.otago.ac.nz

Ken G Dodds [4]
Email: ken.dodds@agresearch.co.nz

James W Kijas [5]
Email: james.kijas@csiro.au

Noelle Cockett[6]
Email: noelle.cockett@usu.edu

John C McEwan [4]
Email: john.mcewan@agresearch.co.nz

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

[2] Victorian Department of Economic Development, Jobs, Transport and Resources, Bundoora, VIC 3083, AUSTRALIA

[3] Department of Biochemistry, University of Otago, 710 Cumberland St, Dunedin 9054, NEW ZEALAND

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

[5] CSIRO Animal, Food and Health Sciences, Queensland Bioscience Precinct, 306 Carmody Road. St Lucia, QLD 4067, AUSTRALIA

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

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

Abstract

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

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

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

Background

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

There are many examples, particularly in humans, of CNVs influencing traits. These include multiple examples of CNVs associated with cancer susceptibility [10-12], the association of the FCGR3B gene copy number variant with systemic lupus erythematosus (SLE) [13], and CCL3L1 gene copy number, which has been linked to HIV susceptibility [14]. There is also evidence for CNVs influencing traits in other animal and livestock species. A 133kb duplication containing four genes causes hair ridge in Rhodesian and Thai Ridgeback dogs [15]. The chicken Peacomb phenotype is under sexual selection and is caused by a 3.2 kb duplication in an intron of the SOX5 gene [16]. The Peacomb allele contains ~30 copies of the duplication, with variation in copy number present within individuals with the Peacomb phenotype. In pigs, Chen *et al* [17] found seven copy number variable genes that overlapped quantitative trait loci (QTL) for, among other traits, carcass length, backfat thickness, abdominal fat weight, length of scapular, intramuscular fat content of *longissimus* muscle, body weight at 240 days and glycolytic potential of longissimus muscle. Although not an association analysis, Chen *et al* [17] identified one CNV that had previously been associated with skin colour in pigs [18].

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

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

A number of approaches have been used to detect the presence of CNV. The main platforms are comparative genomic hybridisation (CGH) arrays [29-33], SNP arrays [34-37] and depth of coverage metrics applied to whole genome sequence data (e.g., [38-42]). Further, there are a variety of algorithms that can be used to analyse available resultant data. Perhaps the most widely used platform is array CGH, as it represents a cost-effective method to detect CNVs on a genome-wide scale in multiple individuals [43].

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

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

Results

Roche-NimbleGen 2.1M CGH array construction and application

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

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

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

Autosomal CNVR

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

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

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

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

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

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

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

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

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

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

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

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

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

Overlap between autosomal CNVRs and genes

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

Non-autosomal CNVRs

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

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

Discussion

The results reported here provide a genome wide view of the frequency of CNV, an important class of genomic variant that is currently poorly characterised in the sheep genome. Using a custom built Roche-NimbleGen 2.1M CGH array, 9,789 autosomal CNVs were detected in 30 sheep. On average these CNVs covered 0.4% of each animal's genome. This is higher than that reported in the initial sheep survey where, on average, 0.05% of an individual sheep genome comprised CNVs [28]. The difference in estimates is not surprising as this study used a CGH array with 2.1 million probes while Fontanesi *et al* [28] used a CGH array with 385,000 probes. Based on probe spacing in the genome and the filters applied to the data, the earlier study detected CNVs greater than 30kb in length, on average, while this study had a resolution of ~4kb on average. As a result, differences in resolution may have resulted in differences in the number of CNVs detected. This is reflected in the datasets, with the average size of CNVs detected by Fontanesi *et al* [28] being 77.6kb (median 55.9kb) and the average size detected in this study being 30.3kb (median 8.7kb). The individual genome CNV composition estimates are similar to, but slightly lower than, estimates reported in humans (e.g., 0.5%, [48]; 0.78%, [7]; and 1.2%, [8]).

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

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

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

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

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

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

Whole genome sequencing exhibited the highest cross platform verification rate, with 61% of CNVs verified with this platform. The CNVs that were unable to be verified were shorter and had lower absolute $\log_2$ ratios than calls that were able to be verified. Both verified and non-verified CNVs had significantly higher GC content than the genome average, which supports data from the literature reporting that GC-rich regions can be more prone to CNV formation [61, 62]. Non-verified CNVs had significantly higher GC content than verified CNVs. While it is possible that the non-verified CNVs were false negatives in the sequence analysis, it is also possible that they were false positives in the CGH dataset, as false positive CGH calls can be related to regions with high GC content [63, 64]. Future work could involve adjusting CGH intensity data for GC content.

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

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

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

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

Conclusions

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

Methods

Roche-NimbleGen 2.1M CGH array - design overview

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

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

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

Of the 21,310 SNP sequences that could not be mapped to UMD3_OA, 240 were mapped by relaxing 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 designed to cover 218 of these SNP regions.

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

Roche-NimbleGen 2.1M CGH array design – chromosome unknown

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

Roche-NimbleGen 2.1M CGH array – animals assayed

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

Roche-NimbleGen 2.1M CGH array – segMNT output processing

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

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

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

Validating Roche-NimbleGen 2.1M CGH array segment calls

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

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

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

The crossvalidate procedure in SAS software (*SAS version 9.1*) was used to test model performance. This procedure omits one segment call in turn and re-calculates model coefficients based on all other segment calls per iteration. It then predicts the probability the omitted call represents a true CNV. Threshold values were applied to categorise calls as true or false based on their probabilities – 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. For each probability threshold tested, the number of times the procedure correctly predicted the known segment call status (true or false) was used as a measure of model accuracy. The final probability threshold used was 0.95.

The final model selected was,

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

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

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

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

CNV regions

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

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

Verifying CNVRs across platforms

Three other platforms were used for CNV identification – Roche-NimbleGen 385K CGH array, OvineSNP50 BeadChip, and Illumina HiSeq 2000 sequence data analysis, with each based on a different version of the ovine genome. To perform cross platform validation autosomal CNVRs identified on the Roche-NimbleGen 2.1M CGH array were mapped to genomes BTA_OARv.2 (for use with the 385K CGH array), OARv1 (for use with the OvineSNP50 BeadChip) and OARv3 (for use with sequence data analysis). CNVR sequence and 1,750bps flanking the start and stop of each CNVR were obtained. Sequences were masked with an ovine repeat database isgcandrepbase2 (Supplementary file 1) and BLASTed against each genome, with parameters -F 'm D' -U T -Z 2000 [69]. CNVR start and stop positions on each genome were approximated based on the BLAST alignment. When the predicted CNVR start position was a negative number, it was set to one (i.e. the first base pair of the chromosome).

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

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

Six animals assayed on the 2.1M CGH array were each sequenced to between 9.8X and 14X coverage by paired-end sequencing on the Illumina HiSeq 2000 platform at Baylor College of Medicine. The following analysis was carried out separately for each animal. Sequence reads were positioned on ovine genome OARv3 using the Burrows-Wheeler Alignment (BWA) algorithm [71] and pileup files [72] were used to retrieve read depth information at each base pair position on the autosomes. Reads were portioned into 1kbp overlapping bins, excluding repetitive sequence, using a sliding window of 200bps. Masked repetitive sequence positions were translated to genome build OARv3. As well as excluding repetitive sequence, for each chromosome a maximum read depth was set per chromosome to exclude potentially unmasked repeats from the CNV sequence analysis. The maximum read depth threshold was set based on inspection of the read depth distribution function with the aim of excluding outliers in read depth data. Bins with a maximum read depth exceeding the threshold were deleted from the analysis. The average read depth over all base pairs was determined for each bin after correcting for GC content based on methods presented by Yoon *et al* [73].

Pseudo-Maximum likelihood was used to fit a mixture model to determine if the average read depth for each bin represented a homozygous deletion (copy number, CN=0), heterozygous deletion (CN=1), normal diploid copy number (2), heterozygous duplication (3) or homozygous duplication (4) in the genome. The mixture model used (Table 4) was a mixture of four normal distributions (for modeling CN = 1 to 4) and one half-normal distribution (for CN = 0). Constraints were placed on the parameters of the normal distributions so that the means and variances of the distributions corresponding to CN =1, 3 and 4 were equal to respectively 1/2, 3/2 and 2 times the mean and variance of the distribution corresponding to CN = 2. Model fitting was done on a per chromosome basis, using the R function *nlminb* [74]. Specifically, seven parameters were estimated for each chromosome: μ_2 and σ_2^2, the mean and variance of read depth for a bin corresponding to CN = 2 (the "normal" diploid copy number); σ_0^2, the variance of read depth for a bin corresponding to CN = 0 (homozygous deletion) and four of the five mixture weights (prior probability of a bin falling into each of the five distributions). Where these parameters could not be estimated for a chromosome, average estimates based on all other chromosomes for a given animal were used. Table 5 details the starting values and lower and upper bounds used by *nlminb* for each parameter. Based on those parameter estimates, each bin was assigned to one of the five CNV classes by multiplying the values of each of the five probability density functions for each bin by the corresponding mixture weights (i.e., calculating the posterior probability of a bin being in each of the distributions) and selecting the CNV class with the highest value. For each of the six animals, bins in regions corresponding to autosomal CNVRs identified on the 2.1M CGH array and mapped to OARv3 autosomes were used to determine if the CNVR was verified in the sequence data. In instances where there was conflict between results from the sequence analysis and the 2.1M CGH array, individual animal data were compared to the reference (Poll Dorset) animal. This animal was used as the reference animal in the 2.1M CGH array experiments and therefore results for individual animals may be influenced by the corresponding copy number present in the Poll Dorset.

Comparison of CNVRs to those identified in the literature

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

Overlap between autosomal CNVRs and genes

CNVR sequences were masked (isgcandrepbase2) and BLASTed (parameters -F 'm D' -U T -Z 2000) against OARv3 to obtain their positions on the genome. Positions of the coding sequence of genes on OARv3 were provided by BGI (personal communication, Rudiger Brauning). Overlap between autosomal CNVRs and the coding sequence of genes were determined. CNVRs that overlapped gene coding sequences by 1bp or more were used to derive the proportion of CNVRs overlapping genes. Overlap with the agouti signalling protein and adenosylhomocysteinase genes were used as a positive control, as this locus is observed as duplicated in the sheep genome [27].

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

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

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

Acknowledgments

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

References

1. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science. 2007;315(5813):848-53. doi:315/5813/848 [pii]

10.1126/science.1136678.

2. Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. Nat Rev Genet. 2006;7(2):85-97. doi:http://www.nature.com/nrg/journal/v7/n2/suppinfo/nrg1767_S1.html.

3. Egan CM, Sridhar S, Wigler M, Hall IM. Recurrent DNA copy number variation in the laboratory mouse. Nat Genet. 2007;39(11):1384-9. doi:10.1038/ng.2007.19.

4. Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA et al. A map of human genome variation from population-scale sequencing. Nature. 2010;467(7319):1061-73. doi:nature09534 [pii]

10.1038/nature09534.

5. Michaelson Jacob J, Shi Y, Gujral M, Zheng H, Malhotra D, Jin X et al. Whole-Genome Sequencing in Autism Identifies Hot Spots for De Novo Germline Mutation. Cell. 2012;151(7):1431-42. doi:http://dx.doi.org/10.1016/j.cell.2012.11.019.

6. Stankiewicz P, Lupski JR. Structural variation in the human genome and its role in disease. Annu Rev Med. 2010;61:437-55. doi:10.1146/annurev-med-100708-204735.

7. Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y et al. Origins and functional impact of copy number variation in the human genome. Nature. 2010;464(7289):704-12. doi:http://www.nature.com/nature/journal/v464/n7289/suppinfo/nature08516_S1.html.

8. Pang A, MacDonald J, Pinto D, Wei J, Rafiq M, Conrad D et al. Towards a comprehensive structural variation map of an individual human genome. Genome Biology. 2010;11(5):R52.

9. Kim PM, Lam HYK, Urban AE, Korbel JO, Affourtit J, Grubert F et al. Analysis of copy number variants and segmental duplications in the human genome: Evidence for a change in the process of formation in recent evolutionary history. Genome Res. 2008;18(12):1865-74. doi:10.1101/gr.081422.108.

10. Long J, Delahanty RJ, Li G, Gao YT, Lu W, Cai Q et al. A Common Deletion in the APOBEC3 Genes and Breast Cancer Risk. J Natl Cancer Inst. 2013;105(8):573-9. doi:djt018 [pii]

10.1093/jnci/djt018.

11. Yang L, Liu B, Huang B, Deng J, Li H, Yu B et al. A functional copy number variation in the WWOX gene is associated with lung cancer risk in Chinese. Hum Mol Genet. 2013;22(9):1886-94. doi:ddt019 [pii]

10.1093/hmg/ddt019.

12. Suehiro Y, Okada T, Shikamoto N, Zhan Y, Sakai K, Okayama N et al. Germline copy number variations associated with breast cancer susceptibility in a Japanese population. Tumour Biol. 2012;34(2):947-52. doi:10.1007/s13277-012-0630-x.
13. Fanciulli M, Norsworthy PJ, Petretto E, Dong R, Harper L, Kamesh L et al. FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. Nat Genet. 2007;39(6):721-3. doi:ng2046 [pii]

10.1038/ng2046.
14. Gonzalez E, Kulkarni H, Bolivar H, Mangano A, Sanchez R, Catano G et al. The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. Science. 2005;307(5714):1434-40. doi:1101160 [pii]

10.1126/science.1101160.
15. Salmon Hillbertz NH, Isaksson M, Karlsson EK, Hellmen E, Pielberg GR, Savolainen P et al. Duplication of FGF3, FGF4, FGF19 and ORAOV1 causes hair ridge and predisposition to dermoid sinus in Ridgeback dogs. Nat Genet. 2007;39(11):1318-20. doi:ng.2007.4 [pii]

10.1038/ng.2007.4.
16. Wright D, Boije H, Meadows JR, Bed'hom B, Gourichon D, Vieaud A et al. Copy number variation in intron 1 of SOX5 causes the Pea-comb phenotype in chickens. PLoS Genet. 2009;5(6):e1000512. doi:10.1371/journal.pgen.1000512.
17. Chen C, Qiao R, Wei R, Guo Y, Ai H, Ma J et al. A comprehensive survey of copy number variation in 18 diverse pig populations and identification of candidate copy number variable genes associated with complex traits. BMC Genomics. 2012;13:733. doi:1471-2164-13-733 [pii]

10.1186/1471-2164-13-733.
18. Johansson Moller M, Chaudhary R, Hellmen E, Hoyheim B, Chowdhary B, Andersson L. Pigs with the dominant white coat color phenotype carry a duplication of the KIT gene encoding the mast/stem cell growth factor receptor. Mamm Genome. 1996;7(11):822-30.
19. Bae J, Cheong H, Kim L, NamGung S, Park T, Chun J-Y et al. Identification of copy number variations and common deletion polymorphisms in cattle. BMC Genomics. 2010;11(1):232.
20. Bickhart DM, Hou Y, Schroeder SG, Alkan C, Cardone MF, Matukumalli LK et al. Copy number variation of individual cattle genomes using next-generation sequencing. Genome Res. 2012;22(4):778-90. doi:10.1101/gr.133967.111.
21. Hou Y, Liu GE, Bickhart DM, Cardone MF, Wang K, Kim ES et al. Genomic characteristics of cattle copy number variations. BMC Genomics. 2011;12:127. doi:1471-2164-12-127 [pii]

10.1186/1471-2164-12-127.
22. Kijas JW, Barendse W, Barris W, Harrison B, McCulloch R, McWilliam S et al. Analysis of copy number variants in the cattle genome. Gene. 2011;482(1-2):73-7. doi:S0378-1119(11)00179-X [pii]

10.1016/j.gene.2011.04.011.
23. Liu GE, Hou Y, Zhu B, Cardone MF, Jiang L, Cellamare A et al. Analysis of copy number variations among diverse cattle breeds. Genome Res. 2010;20(5):693-703. doi:10.1101/gr.105403.110.
24. Cicconardi F, Chillemi G, Tramontano A, Marchitelli C, Valentini A, Ajmone-Marsan P et al. Massive screening of copy number population-scale variation in Bos taurus genome. BMC Genomics. 2013;14(1):124. doi:1471-2164-14-124 [pii]

10.1186/1471-2164-14-124.
25. Fadista J, Thomsen B, Holm LE, Bendixen C. Copy number variation in the bovine genome. BMC Genomics. 2010;11:284. doi:1471-2164-11-284 [pii]

10.1186/1471-2164-11-284.

26. Jiang L, Jiang J, Yang J, Liu X, Wang J, Wang H et al. Genome-wide detection of copy number variations using high-density SNP genotyping platforms in Holsteins. BMC Genomics. 2013;14(1):131. doi:1471-2164-14-131 [pii]

10.1186/1471-2164-14-131.

27. Norris BJ, Whan VA. A gene duplication affecting expression of the ovine ASIP gene is responsible for white and black sheep. Genome Res. 2008;18(8):1282-93. doi:gr.072090.107 [pii]

10.1101/gr.072090.107.

28. Fontanesi L, Beretti F, Martelli PL, Colombo M, Dall'Olio S, Occidente M et al. A first comparative map of copy number variations in the sheep genome. Genomics. 2011;97(3):158-65. doi:10.1016/j.ygeno.2010.11.005.

29. Barrett MT, Scheffer A, Ben-Dor A, Sampas N, Lipson D, Kincaid R et al. Comparative genomic hybridization using oligonucleotide microarrays and total genomic DNA. Proc Natl Acad Sci U S A. 2004;101(51):17765-70. doi:0407979101 [pii]

10.1073/pnas.0407979101.

30. Graubert TA, Cahan P, Edwin D, Selzer RR, Richmond TA, Eis PS et al. A high-resolution map of segmental DNA copy number variation in the mouse genome. PLoS Genet. 2007;3(1):e3. doi:06-PLGE-RA-0282R3 [pii]

10.1371/journal.pgen.0030003.

31. Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F et al. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science. 1992;258(5083):818-21.

32. Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D et al. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nat Genet. 1998;20(2):207-11.

33. Yu H, Chao J, Patek D, Mujumdar R, Mujumdar S, Waggoner AS. Cyanine dye dUTP analogs for enzymatic labeling of DNA probes. Nucleic Acids Res. 1994;22(15):3226-32.

34. Jeon JP, Shim SM, Jung JS, Nam HY, Lee HJ, Oh BS et al. A comprehensive profile of DNA copy number variations in a Korean population: identification of copy number invariant regions among Koreans. Exp Mol Med. 2009;41(9):618-28. doi:10.3858/emm.2009.41.9.068.

35. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T et al. Mapping and sequencing of structural variation from eight human genomes. Nature. 2008;453(7191):56-64. doi:http://www.nature.com/nature/journal/v453/n7191/suppinfo/nature06862_S1.html.

36. McCarroll SA, Hadnott TN, Perry GH, Sabeti PC, Zody MC, Barrett JC et al. Common deletion polymorphisms in the human genome. Nat Genet. 2006;38(1):86-92.

37. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD et al. Global variation in copy number in the human genome. Nature. 2006;444(7118):444-54. doi:http://www.nature.com/nature/journal/v444/n7118/suppinfo/nature05329_S1.html.

38. Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. 2011;21(6):974-84. doi:gr.114876.110 [pii]

10.1101/gr.114876.110.

39. Teo SM, Pawitan Y, Ku CS, Chia KS, Salim A. Statistical challenges associated with detecting copy number variations with next-generation sequencing. Bioinformatics. 2012;28(21):2711-8. doi:10.1093/bioinformatics/bts535.

40. Xi R, Hadjipanayis AG, Luquette LJ, Kim TM, Lee E, Zhang J et al. Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. Proc Natl Acad Sci U S A. 2011;108(46):E1128-36. doi:1110574108 [pii]

10.1073/pnas.1110574108.

41. Xi R, Lee S, Park PJ. A survey of copy-number variation detection tools based on high-throughput sequencing data. Curr Protoc Hum Genet. 2012;Chapter 7:Unit7 19. doi:10.1002/0471142905.hg0719s75.
42. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics. 2009;25(21):2865-71. doi:btp394 [pii]

10.1093/bioinformatics/btp394.
43. Savarese M, Piluso G, Orteschi D, Di Fruscio G, Dionisi M, Blanco FdV et al. Enhancer Chip: Detecting Human Copy Number Variations in Regulatory Elements. PLoS ONE. 2012;7(12):e52264. doi:10.1371/journal.pone.0052264.
44. Krumm N, Sudmant PH, Ko A, O'Roak BJ, Malig M, Coe BP et al. Copy number variation detection and genotyping from exome sequence data. Genome Res. 2012;22(8):1525-32. doi:10.1101/gr.138115.112.
45. Crawford AM, Dodds KG, Ede AJ, Pierson CA, Montgomery GW, Garmonsway HG et al. An autosomal genetic linkage map of the sheep genome. Genetics. 1995;140(2):703-24.
46. Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Altshuler DM et al. Copy number variation: new insights in genome diversity. Genome Res. 2006;16(8):949-61. doi:gr.3677206 [pii]

10.1101/gr.3677206.
47. Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics. 2004;5(4):557-72. doi:10.1093/biostatistics/kxh008.
48. McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, Wysoker A et al. Integrated detection and population-genetic analysis of SNPs and copy number variation. Nat Genet. 2008;40(10):1166-74. doi:ng.238 [pii]

10.1038/ng.238.
49. Craddock N, Hurles ME, Cardin N, Pearson RD, Plagnol V, Robson S et al. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature. 2010;464(7289):713-20. doi:nature08979 [pii]

10.1038/nature08979.
50. Li Y, Mei S, Zhang X, Peng X, Liu G, Tao H et al. Identification of genome-wide copy number variations among diverse pig breeds by array CGH. BMC Genomics. 2012;13:725. doi:1471-2164-13-725 [pii]

10.1186/1471-2164-13-725.
51. Berglund J, Nevalainen E, Molin A-M, Perloski M, Consortium TL, Andre C et al. Novel origins of copy number variation in the dog genome. Genome Biology. 2012;13(8):R73.
52. She X, Cheng Z, Zollner S, Church DM, Eichler EE. Mouse segmental duplication and copy number variation. Nat Genet. 2008;40(7):909-14. doi:http://www.nature.com/ng/journal/v40/n7/suppinfo/ng.172_S1.html.
53. Cooper GM, Nickerson DA, Eichler EE. Mutational and selective effects on copy-number variants in the human genome. Nat Genet. 2007;39(7 Suppl):S22-9. doi:ng2054 [pii]

10.1038/ng2054.
54. Stankiewicz P, Lupski JR. Genome architecture, rearrangements and genomic disorders. Trends Genet. 2002;18(2):74-82. doi:S0168-9525(02)02592-1 [pii].
55. Zhang F, Gu W, Hurles ME, Lupski JR. Copy number variation in human health, disease, and evolution. Annu Rev Genomics Hum Genet. 2009;10:451-81. doi:10.1146/annurev.genom.9.081307.164217.
56. Conrad DF, Andrews TD, Carter NP, Hurles ME, Pritchard JK. A high-resolution survey of deletion polymorphism in the human genome. Nat Genet. 2006;38(1):75-81. doi:ng1697 [pii]

10.1038/ng1697.
57. Emerson JJ, Cardoso-Moreira M, Borevitz JO, Long M. Natural Selection Shapes Genome-Wide Patterns of Copy-Number Polymorphism in Drosophila melanogaster. Science. 2008;320(5883):1629-31. doi:10.1126/science.1158078.
58. Jurka J, Kohany O, Pavlicek A, Kapitonov VV, Jurka MV. Duplication, coclustering, and selection of human Alu retrotransposons. Proceedings of the National Academy of Sciences of the United States of America. 2004;101(5):1268-72. doi:10.1073/pnas.0308084100.
59. Cooper GM, Zerr T, Kidd JM, Eichler EE, Nickerson DA. Systematic assessment of copy number variant detection via genome-wide SNP genotyping. Nat Genet. 2008;40(10):1199-203. doi:ng.236 [pii]

10.1038/ng.236.
60. Locke DP, Sharp AJ, McCarroll SA, McGrath SD, Newman TL, Cheng Z et al. Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. Am J Hum Genet. 2006;79(2):275-90. doi:S0002-9297(07)63135-8 [pii]

10.1086/505653.
61. Cooper DN, Bacolla A, Férec C, Vasquez KM, Kehrer-Sawatzki H, Chen J-M. On the sequence-directed nature of human gene mutation: The role of genomic architecture and the local DNA sequence environment in mediating gene mutations underlying human inherited disease. Hum Mutat. 2011;32(10):1075-99. doi:10.1002/humu.21557.
62. Nguyen DQ, Webber C, Hehir-Kwa J, Pfundt R, Veltman J, Ponting CP. Reduced purifying selection prevails over positive selection in human copy number variant evolution. Genome Res. 2008;18(11):1711-23. doi:gr.077289.108 [pii]

10.1101/gr.077289.108.
63. Liu G, Hou Y, Robl J, Kuroiwa Y, Wang Z. Assessment of genome integrity with array CGH in cattle transgenic cell lines produced by homologous recombination and somatic cell cloning. Genome Integrity. 2011;2(1):6.
64. Marioni J, Thorne N, Valsesia A, Fitzgerald T, Redon R, Fiegler H et al. Breaking the waves: improved detection of copy number variation from microarray-based comparative genomic hybridization. Genome Biology. 2007;8(10):R228.
65. Crawford AM, Cuthbertson RP. Mutations in sheep microsatellites. Genome Res. 1996;6(9):876-9.
66. Winchester L, Yau C, Ragoussis J. Comparing CNV detection methods for SNP arrays. Brief Funct Genomic Proteomic. 2009;8(5):353-66. doi:elp017 [pii]

10.1093/bfgp/elp017.
67. Pinto D, Darvishi K, Shi X, Rajan D, Rigler D, Fitzgerald T et al. Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. Nat Biotech. 2011;29(6):512-20.
doi:http://www.nature.com/nbt/journal/v29/n6/abs/nbt.1852.html#supplementary-information.
68. Jenkins G. UMD3_OA assembly. 2015. http://www.sheephapmap.org/CNV/UMD3_OA.tgz. Accessed 1 June 2015 2015.
69. Korf I, Yandell M, Bedell J. BLAST. O'Reilly Media; 2003.
70. Fontanesi L, Martelli P, Beretti F, Riggio V, Dall'Olio S, Colombo M et al. An initial comparative map of copy number variations in the goat (Capra hircus) genome. BMC Genomics. 2010;11(1):639.
71. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009;25(14):1754-60. doi:10.1093/bioinformatics/btp324.
72. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9. doi:10.1093/bioinformatics/btp352.
73. Yoon S, Xuan Z, Makarov V, Ye K, Sebat J. Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res. 2009;19(9):1586-92. doi:gr.092981.109 [pii]

10.1101/gr.092981.109.
74. Gay DM. Computing Science Technical Report No.153: Usage summary for selected optimization routines. 1990.

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

*absl2r is the absolute $\log_2$ ratio of the CNV. # number of CGH array probes in the CNV.

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

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)

Table 4. Description of the pseudo-maximum likelihood derived mixture model for estimating copy 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	$\frac{1}{2}\mu_2$	$\frac{1}{2}\sigma_0^2$
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	$\frac{4}{2}\mu_2$	$\frac{4}{2}\sigma_2^2$

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

Variable	Starting value	Lower bound	Upper bound
μ_2	$\bar{\mu}$	-∞	∞
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

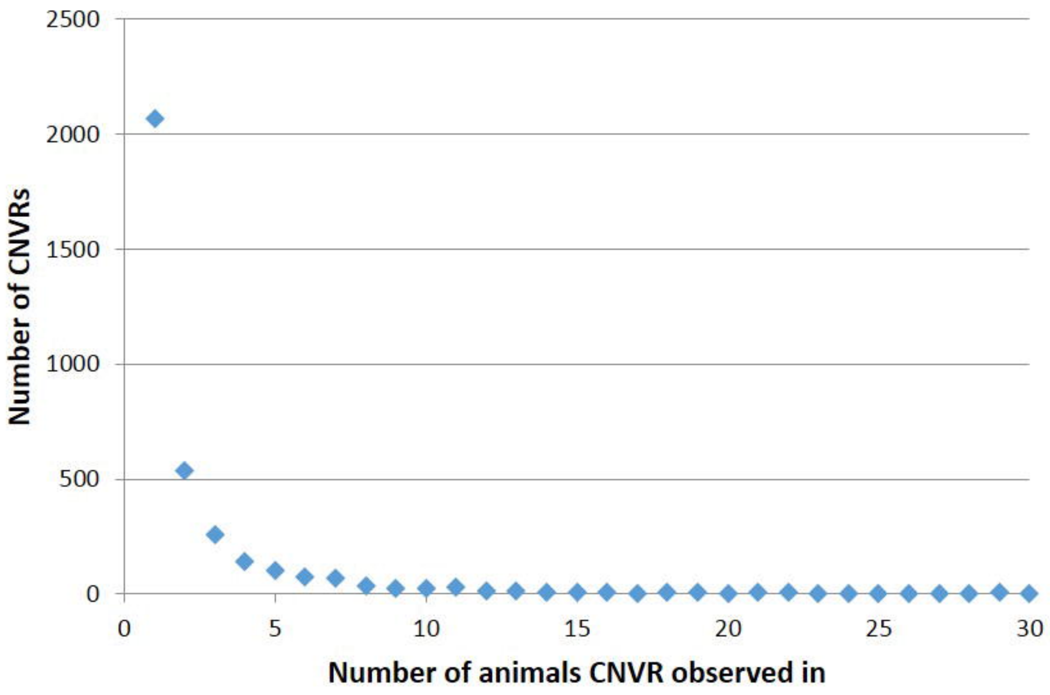
Figure 1. CNVR frequency across animals.

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

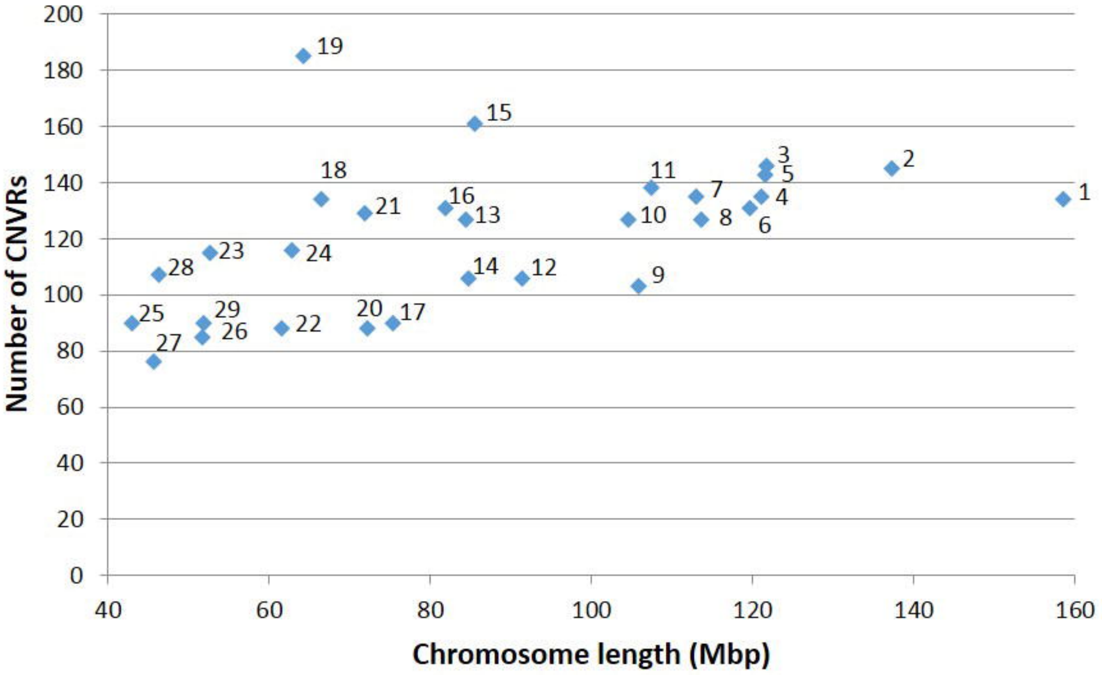
Figure 3. Cumulative density plot of the distances separating CNVRs. The red line reflects the observed pairwise distances between CNVRs, while the blue line reflects the simulated (expected if CNVRs are uniformly distributed in the genome) distances separating CNVRs.

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

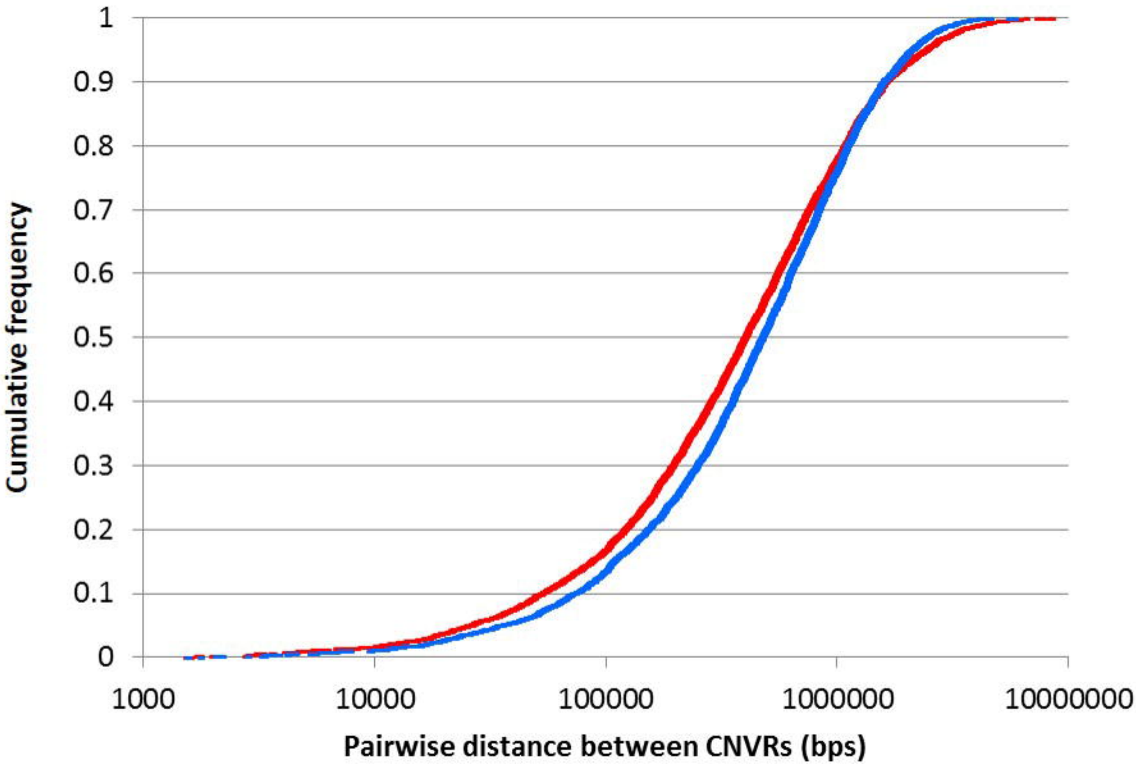
Figure 5. Pedigree of International Mapping Flock (IMF) animals assayed on the Roche NimbleGen 2.1M CGH array. Some animals (green) appear in more than one pedigree. Segment calls from animals IMF66, IMF91, IMF95, IMF108 and IMF112 (red) were removed from the analysis due to failed 2.1M CGH arrays.



Number of CNVRs
2500
2000
1500
1000
500
0
0
5
10
15
20
25
30
Number of animals CNVR observed in

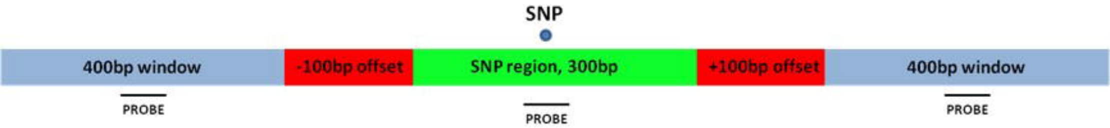


Number of CNVRs
Chromosome length (Mbp)
200
180
160
140
120
100
80
60
40
20
0
40
60
80
100
120
140
160
19
15
18
21
16
13
24
23
28
14
12
25
29
26
22
20
17
27
11
10
9
7
8
3
5
4
6
2
1

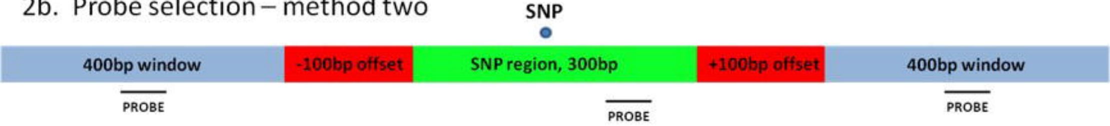


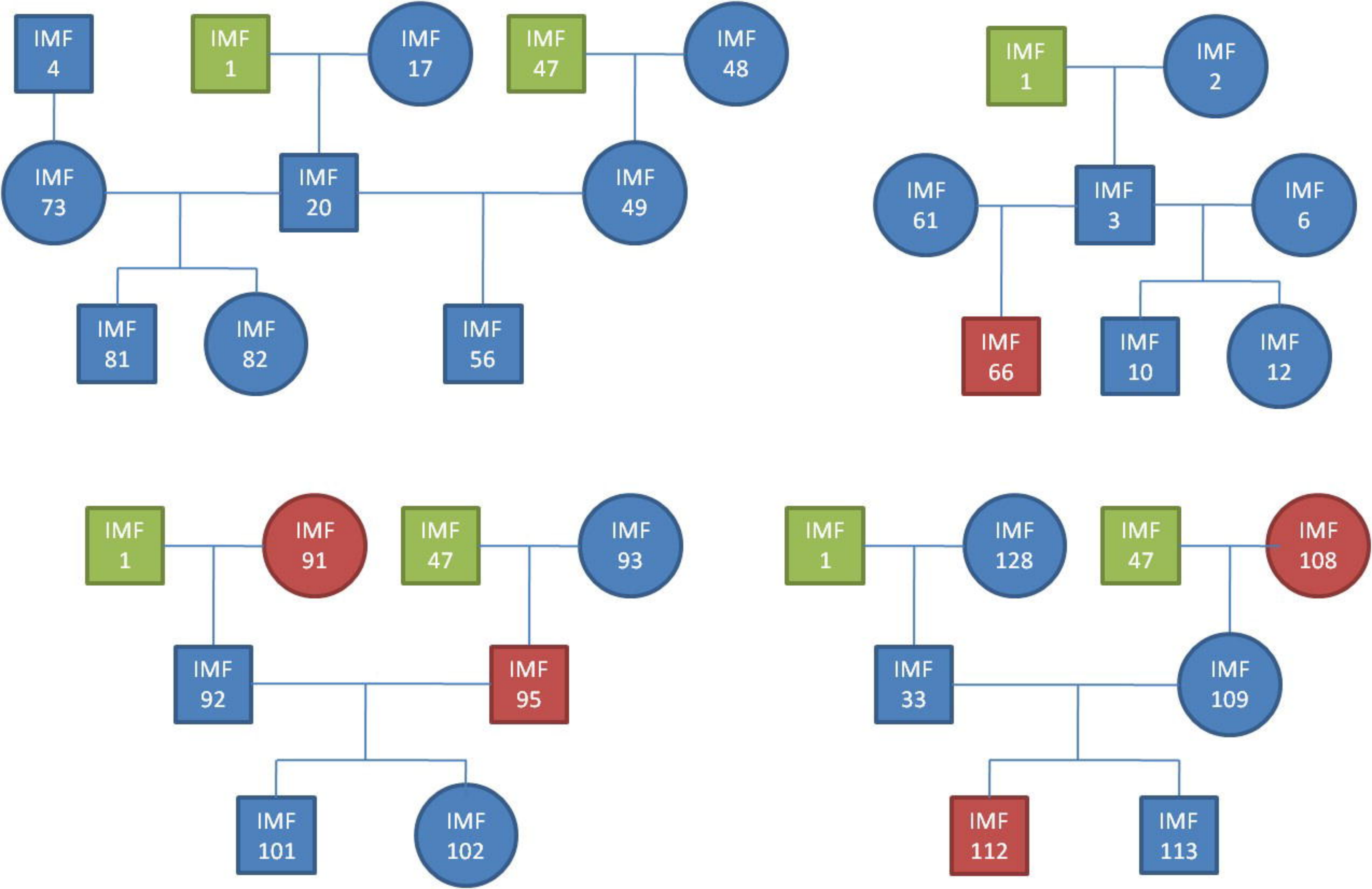
Cumulative frequency
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1000
10000
100000
1000000
10000000
Pairwise distance between CNVRs (bps)

2a. Probe selection – method one



2b. Probe selection – method two





IMF 4
IMF 1
IMF 17
IMF 47
IMF 48
IMF 73
IMF 20
IMF 49
IMF 81
IMF 82
IMF 56
IMF 1
IMF 2
IMF 61
IMF 3
IMF 6
IMF 66
IMF 10
IMF 12
IMF 1
IMF 91
IMF 47
IMF 93
IMF 92
IMF 95
IMF 101
IMF 102
IMF 1
IMF 128
IMF 47
IMF 108
IMF 33
IMF 109
IMF 112
IMF 113