Locating a novel autosomal recessive genetic condition using only WGS data from three cases and six controls; a case study of a new variant in the cattle glucokinase gene

Citation for published version:
Pollott, GE, Piercy, R, Massey, C, Salavati, M, Cheng, Z & Wathes, DC 2022, 'Locating a novel autosomal recessive genetic condition using only WGS data from three cases and six controls; a case study of a new variant in the cattle glucokinase gene', *Frontiers in genetics*, vol. 13, 755693. https://doi.org/10.3389/fgene.2022.755693

Digital Object Identifier (DOI):
10.3389/fgene.2022.755693

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Frontiers in genetics

Locating a novel autosomal recessive genetic disease using only WGS data from three cases and six controls; a case study of a new variant in the cattle glucokinase gene

Geoffrey Pollott[1*], Richard Piercy[2], Claire Massey[2], Mazdak Salavati[3,4] and D. Claire Wathes[3]

[1]Department of Pathobiology and Population Sciences, The Royal Veterinary College, Royal College Street, London, NW1 0TU, UK

[2]Comparative Neuromuscular Diseases Laboratory, Department of Clinical Sciences and Services, Royal Veterinary College, Royal College Street, London NW1 0TU, UK

[3]Department of Pathobiology and Population Sciences, The Royal Veterinary College, Hawkshead Lane, Hatfield, Hertfordshire, AL9 7TA, UK

[4]Now at The Roslin Institute, Easter Bush, Midlothian, EH25 9RG, Scotland, UK

*** Correspondence:**
Corresponding Author
gpollott@rvc.ac.uk

Keywords: Cattle$_1$, Glucokinase gene$_2$, Recessive genetics$_3$, Runs of homozygosity$_4$, WGS$_5$.

Abstract

New Mendelian genetic conditions arise all the time. To manage them effectively methods need to be devised that are quick and accurate. Until recently, finding the causal genomic site of a new autosomal recessive genetic disease has required a two-stage approach using single nucleotide polymorphism (SNP) chip genotyping to locate the region containing the new variant. This region is then explored using fine-mapping methods to locate the actual site of the new variant. This paper explores bioinformatic methods that can be used with just nine whole-genome sequenced animals to simplify and expedite the process to a one-step procedure.

Using whole genome sequencing of only three cases and six controls, the site of a novel variant causing perinatal mortality in Irish Moiled calves was located. Four methods were used to interrogate the variant call format (VCF) datafile of these nine animals; genotype criteria (GCR), autozygosity by difference (ABD), variant prediction scoring and registered SNP information. From more than 8 million variants in the VCF file only one site was identified by all four methods (Chr4: g.77173487A>T (ARS-UCD1.2 (GCF_002263795.1)). This site was a splice acceptor variant located in the glucokinase gene (*GCK*). It was verified on an independent sample of animals from the breed using genotyping by polymerase chain reaction at the candidate site and autozygosity by difference using SNP chips. Both methods confirmed the candidate site. Investigation of the GCR method found that sites meeting the GCR were not evenly spread across the genome but concentrated in regions of long runs of homozygosity. Locating GCR sites was best done using two controls to every case and the controls should be distantly related to the cases, within the breed concerned. Less than 20 animals need to be sequenced when using the GCR and ABD methods together.

The genomic site of novel autosomal recessive Mendelian genetic diseases can be located using less than 20 animals combined with two bioinformatic methods, autozygosity by difference and genotype criteria. In many instances it may also be confirmed with variant prediction scoring. This should speed-up and simplify the management of new genetic diseases to a single-step process.

1 Introduction

A review of 34 papers detailing work to map 38 novel autosomal recessive genetic conditions to their position on the genome (Pollott, 2018) suggested that finding the site of such a condition required the use of at least a two-stage methodology. Firstly the use of a suitable single nucleotide polymorphism (SNP) chip in a case/control study to locate the *region* containing the new variant combined with a method searching for long runs of homozygosity (ROH), the more traditional chi-squared method being shown to be inadequate. The second stage used a range of 'fine-mapping' methods to search within the highlighted region for the *site* of the new variant, many of which resulted in whole genome sequencing (WGS) of a few cases and controls. More recently methods have been developed which use WGS as an initial step but such methods typically require additional resources or sequencing a large number of animals.

The objective of the current paper was to see if it is possible to locate the position of a novel autosomal recessive genetic condition directly using the ideas contained in Pollott (2018) on WGS methodology and using a small number of cases and controls without recourse to more extensive resources which may not always be available. To achieve this, we investigated combining this approach with a range of other bioinformatic tools and genetic ideas which may indicate the site of such a variant by reading the various signals in the WGS data. Using the Variant Call Format (VCF; VCF (2019)) files from a suitable combination of cases and controls it was suggested that typically about 16 animals would need to be sequenced in order to locate a new autosomal recessive variant using a 'genotype criteria' approach (GCR; Pollott (2018)).

Here we test these ideas on a novel genetic disease found in Irish Moiled cattle. The Irish Moiled is an ancient hornless breed native to the island of Ireland. It was popular in the 1800s but by the late 1970s the pedigree herd numbered only 30 breeding females and two bulls. In 1979 the Rare Breeds Survival Trust recognised the Irish Moiled as endangered and placed the breed on its 'critical' list (Irish Moiled Cattle Society, 2020). The population size now numbers about 875 females and 90 bulls. Fortunately, novel fatal genetic diseases are relatively rare but when they do occur it is important to find the cause and implement plans to manage the condition via selective breeding, as soon as possible. A number of Irish Moiled cattle breeders were concerned about the seemingly high occurrence of early calf deaths in their herds. Affected calves (of both sexes) appeared normal at birth, and were initially active, but then deteriorated rapidly, dying within the first week after birth. An initial analysis was undertaken which suggested that there was likely a genetic basis to the disease (see Supplementary File Page S18) and, since it was fatal, that it could only be inherited as a recessive condition.

2 Materials and Methods

Throughout this paper the ARS-UCD1.2 (GCF_002263795.1) build of the cattle genome was used and all chromosome positions quoted relate to it (Ensembl, 2020).

Commented [MS1]: We should really clarify how the coordinates of variants from UMD3.1 were lifted over to ARS-UCD1.2.
I suggest you add this as a step 9 in the supplementary methods section S2-S3.

Without this clarification it looks confusing for the reviewer as why the data wasn't originally mapped to ARS-UCD1.2.

Commented [MS2]: Its beneficial to mention the version of Ensembl. Current version is v104 but I'd assume you have used either 102 or 103 given the timeline of your work.

2.1 Sample Collection

Farmers were contacted via the Irish Moiled breed society and asked to submit hair samples for analysis. Clean tufts of tail hairs were plucked from either live cows/bulls or recently dead calves (within 8 h post mortem) by the animals' owners. These were sealed in a paper envelope for posting to the laboratory. Information recorded included: sample type (bull, dam or dead calf); ear tag numbers of dam and sire; dam herd ID; calving date; calf sex; time of calf death (stillborn/ days after calving). DNA was then extracted from the hair follicles for processing as described in the Supplementary File (Page S2).

2.2 Whole genome sequencing

Nine DNA samples were used in these analyses comprising three dead calves (cases) and six controls which were either parents of the cases or parents of other dead calves (and grandparents of the cases) as illustrated in the pedigree (Supplementary File Figure S3). These nine samples were sequenced on an Illumina NGS platform after sample preparation as described in the Supplementary File (Page S2). Briefly, 1 µg of DNA per sample was processed using a TruSeq Nano DNA LT Library Prep Kit (Illumina, USA) according to the supplied protocol. This produced randomly sheared 350bp inserts. After end repair and adaptor ligation, DNA was amplified via polymerase chain reaction (PCR) and the product purified using AMPure XP (Beckman Coulter, UK). Size-selected DNA from each animal was sequenced on the HiSeq machine to achieve 150bp paired-end reads to cover the bovine genome with an average 30X coverage (> 90 Gbp raw data with > 85% Q30 (Phred scaled)). Alignment, mapping, variant calling and preparation of the final VCF file were carried out on the subsequent reads as described in the Supplementary File (Page S2).

2.3 Genotype criteria

Information derived from WGS data on a small sample of cases and parental controls may contain a number of signals indicating the site of a novel autosomal recessive condition. 'Across'-animal data should show a typical pattern of homozygous cases and heterozygous parental controls at the candidate site; the 'genotype criteria' approach (Pollott, 2018). Considering a single base position with a reference allele A for the given species and a new variant C which causes a novel autosomal recessive genetic disease, then the expected outcomes from matings between carriers in the population will be offspring with the genotypes AA, AC and CC in the classical Mendelian ratio of 1:2:1. Lethals (CC) would be observed in the dataset if the effect of the new homozygous variant occurred after recording. The term 'genotype criteria' (GCR) was used to mean the particular combination of case and control genotypes required to indicate that a base position could harbour the novel lethal variant (Pollott, 2018). For example, in a dataset comprising five cases and 10 parental controls we would expect to find the novel lethal variant at a position showing CC genotypes in all five cases and a genotype containing the C allele in the 10 parental controls, or all AC in the case of a biallelic position. The probability of occurrence of the GCR for this condition would be $1/3^n$ in cases and $1/3^m$ in parental controls, where n is the number of cases and m is the number of controls that have been whole genome sequenced (Pollott,2018). If the VCF datafile comprised 14 million positions (~0.005 of the cattle genome) then a minimum of 15 animals would probably need to be genotyped in order to find one position with the required genotype criteria i.e. $1/3^{15}$ x 14 million = 0.98 (i.e. ~1), the expected number of sites with the 'correct' genotype criteria from the genome of 15 animals.

A script was written in Perl 5.28 to scan the final VCF file (containing all nine WGS animals) for the expected GCR pattern across cases and controls (i.e. all cases homozygous for the same allele and all

Commented [MS3]: Is it possible to include the code as a supplementary file?

controls heterozygous and containing this allele). In order to qualify for selection a site had to have all genotypes with a Phred-scaled quality score greater than 12 and a depth of coverage more than 11 reads. The identity of these sites was stored along with their relevant VCF record for later scanning and use.

2.4 Autozygosity by difference and runs of homozygosity

'Within'-animal data should show long ROH around the new variant in cases, which are not present in controls. Variants causing a novel autosomal recessive genetic disease are expected to carry with them a very long haplotype originating from the animal in which the variant first arose. When a new case animal is formed then it contains two copies of this long haplotype, only broken up by any recombination events that have occurred since the formation of the original variant, and the new variant will be situated in a long ROH. This idea has been the basis for locating novel variants using SNP chips for a number of years (see Pollott (2018) for a review) and can be used in a number of ways with VCF data. Long ROH throughout the genome could be found and one would expect to find the novel variant in the longest ROH in cases, possibly with adjustment for the situation in controls. Figure 1 shows a hypothetical section of a chromosome containing a novel autosomal recessive variant for a number of cases and controls to illustrate both the ABD and GCR methods.

The autozygosity-by-difference (ABD) method measures runs of homozygosity on each chromosome of each individual in the dataset, cases and controls, using genomewide single-nucleotide variant (SNV) (or SNP) genotypes (Pollott, 2018). Mean ROH length, in Kb, at each SNV positon is calculated for cases and controls separately and then their difference calculated as the ABD score. The likely site of the new variant is in the region with the greatest mean ROH in cases, after taking into account any breed-specific ROH found in controls i.e. the ABD score. The ABD method was programmed in Perl 5.28. The final VCF file was used as the basis to generate a file of sites as input to the ABD method. This method is sensitive to incorrectly called genotypes and so the VCF file was subject to hard filtering as recommended by the Broad Institute (2020) in the absence of suitable databases to use for the recommended variant quality score recalibration (VQSR). A file of SNV were generated from the final VCF file which passed the quality control tests shown in Table S1 following a summary of the quality statistics of the VCF file (see Figures S1 and S2).

The ABD scores were used to look for the potential site of the novel variant causing calf-mortality in the Irish Moiled dataset. The probability of each ABD score was tested using 100 permutations of the data based on the random allocation of animals to phenotypes and recalculation of the ABD scores (Pollott, 2018). Significance at the $P < 0.01$ level was considered as an indicator of a possible site of the new variant.

2.5 Sorting intolerant from tolerant (SIFT) score

In a fatal genetic disease one would expect to find that the products from any change in the sequence would have a drastic effect on the phenotype of the animal so one could not only search for SNV with a potentially drastic effect but could also eliminate those with 'silent' changes. The Variant Effect Predictor (VEP; McLaren et al., 2016)) is a bioinformatic tool which can take a change in a base at a given position on the genome and predict the outcome of that change on the corresponding coding or non-coding genomic feature. Using this method it is possible to model each base position in the VCF file to see the effect of the new variant on the phenotype in the form of a SIFT score.

The final VCF file from all the animals was annotated for variant effect prediction using ~~ENSEMBl~~ Ensembl VEPtools command line v90.5 (McLaren et al., 2016) given the following flags: --tab --fork

8 --offline --species bos_taurus_merged. The VEP was used on all variant sites in the merged VCF file and the results filtered for HIGH SIFT scores using VCFtools (Danecek, 2011). Various outcomes are given in the VEP but here the HIGH outcome was used for the SIFT (Ng and Henikoff, 2002) score since this was a lethal variant. SIFT predicts whether an amino acid substitution affects protein function based on sequence homology and the physical properties of amino acids. The variant impact categories are subjective agreements between VEPtools and SNPEff databases. However, high-impact variants are considered to have protein level disruption or change, while modifier or moderate variants impact non-coding regions of the genome. The SIFT score closer to zero is mostly represented by HIGH or Modifier-impact categories, while tolerated levels (SIFT score of 0.05-1) would show ‘minimal’ to ‘no consequence’ for the function of the genes under said variants. All sites with high-impact scores were captured in a separate file for further processing.

2.6 Novel variants

The dbSNP database of NCBI (NCBI, 2019) contains data on variants already reported by researchers. Novel variants are unlikely to be contained in these datasets and so failure to already have a RS number, an indicator of being in a suitable database, may be another way to reduce the search area along the genome.

A novel variant is unlikely to have been found already in previous studies so is likely to be found at sites that are not already logged in the relevant SNP database. Sites in the final VCF file that did not have an RS number were possible positions for the new variant. The VCF file was scanned for positions not previously allocated a RS number using a script written in Perl 5.28. Such sites were output for further analysis. The sites identified in this way were summarised by the number of genotypes containing the variant allele found at each site. Candidate sites contained a ‘potential’ variant allele in all nine samples.

2.7 Comparing datasets

Putting all these ideas together should make locating the site of the new variant on the genome possible using WGS data from a small number of animals without the need for any other data sources. The above analyses resulted in four independently-derived sets of data, each of which could contain an indication of where the new variant might be found on the genome. These were 1) the sites with the appropriate GCR, 2) sites in long ROH, with a high ABD score significant at $P < 0.01$, 3) sites with a high-impact SIFT score and 4) sites with no RS number. Each dataset was derived from the final VCF file by a method independent of the other three. If a site appears in all four datasets this is likely to be the site of the new variant. The four datasets were compared for overlapping positions by reading them into an Access database and linking on the site position.

2.8 Predicting the effect of the novel variant

The SIFT scoring method described above is one method for modelling the effects of a new variant on the phenotype of the animal. PHYRE2 (Kelley et al., 2015) is an alternative approach which searches for homologous sequences in a database of known proteins etc. The reference sequence and its equivalent using the new variant were entered into the PHYRE2 database to see the effect of the site of the proposed new variant on amino acid sequence and protein structure.

2.9 Methods used to confirm the likely base position of the variant

Two independent methods were used in an attempt to confirm the site derived from the methods used on the WGS data described above. A sample of Irish Moiled animals comprising three bulls, 42 cows and 18 dead calves (male and female) were genotyped using SNP-chips. These were then analysed using 1) genotyping by PCR at the suggested site and 2) the ABD method on the SNP-chip data.

2.9.1 SNP processing

DNA samples extracted as described in the Supplementary File (Page S2) were genotyped in the Department of Pathology, University of Cambridge (UK) and Gen-probe (Heron House, Oaks Business Park, Crewe Road, Wythenshawe, Manchester, M23 9HZ, UK) using either: 1) the Illumina BovineSNP50 BeadChip (Version 1, Illumina Inc., San Diego, CA) (50k SNP, n = 17); 2) the BovineHD Genotyping BeadChip (777k SNP, n = 68) or 3) both chips (n = 7).

The SNP genotypes were prepared for all subsequent analyses using PLINK 1.9 (Chang et al., 2015; PLINK, 2017; Purcell et al., 2007). Quality control parameters were used to edit the data. This involved setting a lower limit on both sample and SNP quality at a call rate greater than 90%, and SNPs were retained in the dataset if they were in Hardy-Weinberg Equilibrium. This was determined using Fisher's Exact Test with a probability threshold of 0.05 and using the mid-p adjustment described in (Graffelman and Moreno, 2013). The latest SNP positions were updated to the ARS-UCD1.2 (GCF_002263795.1) build of the bovine genome using SNPchiMp (Nicolazzi et al., 2014; Nicolazzi et al., 2015). In addition a merged set of data was produced using SNPchiMp combining all genotyped animals from both LD and HD datasets with common SNP. This merged dataset was then used in KING (Manichaikul et al., 2010) to generate relatedness coefficients between all genotyped animals, based on whole genome SNP genotypes, to enable pedigree checking. Any animal whose pedigree did not match the relatedness information from the SNP data was discarded. Because nine animals were also used for the WGS analysis they too were excluded from the SNP ABD analysis, in order to produce a dataset of independent animals.

2.9.2 Autozygosity by difference

The ABD method (Pollott, 2018), described above, was used on the merged SNP-chip dataset. The probability of each SNP ABD score (difference between mean ROH length (Kb) from cases and controls at each SNP position) was tested using 1,000 permutations of the dataset based on random allocation of animals to phenotypes and recalculation of the ABD scores. Significance at the $P < 0.001$ level was considered as an indicator of a possible site of the new variant.

2.9.3 Genotyping by PCR analysis

Primers (5' CATGAACCCAGTGTCACAGC 3' and 5' CTCTCCGTGGAAGAGCAGAT 3') were designed using Primer3 (version 4.1.0; http://primer3.ut.ee) to amplify a 218bp product spanning the identified variant locus. Primer design was based on the published sequence for the *Bos taurus* (UMD3.1; GCF_000003055.6) glucokinase gene ENSBTAG00000032288. Exon/intron boundaries were derived from this in combination with mRNA RefSeq NM_001102302. PCR was performed using AmpliTaq Gold polymerase (Applied Biosystems) according to the manufacturer's protocol. Products were purified using the QIAquick PCR purification kit (Qiagen) and sequenced by Sanger sequencing using the forward and reverse primers. Sequence analysis was carried out in CLC Workbench. The candidate site was updated to the ARS-UCD1.2 (GCF_002263795.1) genome build using the UCSC Genome LiftOver facility (UCSC, 2020).

3 Results

The WGS data from all nine animals, three cases and six controls, resulted in a final VCF file comprising 8,234,367 biallelic autosomal single nucleotide variants which were to be used for all subsequent WGS analyses. These are summarised by chromosome in Table 1 along with the length of each chromosome aligned in the ARS-UCD1.2 (GCF_002263795.1) build of the cattle genome.

3.1 Genotype criteria

Searching the final VCF file for sites with the appropriate genotype criteria (all homozygous cases for the same genotype and all controls heterozygous containing one allele forming the homozygote in cases) resulted in 574 sites being identified. These are shown broken down by chromosome in Table 1. Applying the formula $1/3^n$ cases and $1/3^m$ parental controls to over 8 million SNV we would expect to find ~418 sites fitting the genotype criteria. There were clearly more GCR sites than expected in this set of animals. Chromosomes 1, 4, 12 and 23 appeared to have more sites than expected (Table 1).

3.2 Autozygosity-by-difference method

In order to run the ABD method on the final VCF file the hard-filtering criteria shown in Table S1 were used on the extracted biallelic SNVs for the dataset. This resulted in a file of 629,716 SNP for the ABD analysis. The ABD software was used to generate the Manhattan plots shown in Figures 2 and S4. The two plots in S4 show the mean ROH length at each of the base positions in the VCF file for cases and controls respectively whilst Figure 2, the ABD score, shows the difference between them. Long ROH were found on BTAs 4 and 18. Probabilities for the ABD scores were generated from 100 permutations of the dataset and the regions of the genome with $P < 0.01$ are summarised in Table S2. The 0.01 probability level was computed to be at an ABD score of 9,034 Kb. Table S2 shows that the length of BTA4 above the 0.01 probability threshold was 7.804Mb. The highest mean ROH length in cases was 14.197Mb so the long ROH found on BTA4 continued either side of the significant region. Similarly on BTA18, the highest mean ROH score in cases was 22.4Mb long.

3.3 SIFT score

Using the VEP to estimate the effect of each of the SNVs in the final VCF file resulted in 65,961 records of HIGH impact SNVs located at 7,764 different autosomal positions. The distribution of these sites is summarised by chromosome in Table 1.

3.4 Sites with no RS number

The VCF file contained 340,893 sites with no RS number. Only 11% (6,298) sites had genotypes, other than the homozygous reference genome, in all nine cases and controls (NoRS9). The breakdown of these by chromosome is summarised in Table 1. These are likely to contain the novel variant.

3.5 Overlap of GCR, ABD, NoRS9 and SIFT results

So far, four possible datasets were generated that might contain the site of the novel variant causing this new autosomal recessive condition. The overlap between the four datasets is summarised in Table 2. The genotype criteria method resulted in the fewest sites identified (574), with the other methods increasing in the order autozygosity-by-difference, HIGH SIFT score and no RS number with nine genotypes. Combining the GCR method with each of the others in turn allowed the identification of 12 (ABD), 1 (HIGH SIFT) and 4 (NoRS9) sites in common. One site appeared in all four datasets located at position Chr4: g.77173487A>T (ARS-UCD1.2 (GCF_002263795.1). We may tentatively conclude that this is the site of the novel variant causing early calf death in the Irish Moiled breed.

3.6 PHYRE2 pediction

The PHYRE2 prediction (Kelley et al., 2015) of the secondary structures between the reference genome and new variant *GCK* model at the beginning of Exon 8, which contains the possible site of the new variant, Chr4: g.77173487A>T (ARS-UCD1.2 (GCF_002263795.1), predicted the amino acid sequence NPGQQLWY from the reference genome being changed to NPGQQLLY with the new variant.

3.7 Independent confirmation of the results

Over the period of work with Irish Moiled breeders a number of hair samples were collected and these were genotyped on either: 1) the Illumina BovineSNP50 BeadChip (50k SNP, LD, n =17); 2) the BovineHD Genotyping BeadChip (777k SNP, HD, n = 68) or 3) both chips (n = 7). This has allowed us independently to assess the WGS results using two main approaches; Sanger sequencing and the ABD runs of homozygosity method.

3.7.1 Sanger sequencing

A sample of 41 animals were Sanger sequenced at position Chr4: g.77173487A>T (ARS-UCD1.2 (GCF_002263795.1)) following PCR of the region surrounding this site. Table 3 shows the Fisher's exact test (Fishre, 1922) results (with the Freeman-Halton extension (1951)) for animals falling into three categories; calves, carriers and live animals of unknown status. The three genotypes are also shown.

Table 3 shows that all TT animals were calves, all of which died of the symptoms described above. All live animals were either AA or AT. The overall results were significant with a probability = 1.868e-05 (0.00001868) for this 3 x 3 table arising by chance, thus indicating a likely association between genotype and health status at this site.

3.7.2 ABD method based on genotypes derived from the PCR analysis

The merged dataset (HD and LD chip data merged using SNPchiMp) comprised 63 animals (18 cases and 45 controls) and 42,453 SNPs after quality control conditions were met. The 42 animals used in the PCR analysis were selected for the ABD analysis, which excluded the WGS animals so that this analysis was independent of the WGS ABD analysis. The animals were allocated to their case/control status based on their PCR genotype at the highlighted location. The results of the ABD analyses are summarised in Figure 3 and S5 and show a 20.8 Mb length of BTA4 with a permuted probability < 0.001, from 1,000 permutations, equivalent to an ABD score greater than 7,023 Kb. This region was from position position Chr4: g. 62872037 to g. 83635054 (ARS-UCD1.2 (GCF_002263795.1)) which includes the site highlighted as the putative causal variant from the WGS analyses.

The results in Figure S5 show a long ROH on BTA21 but this was present in both the case and control animals which negated each other in the ABD score analysis. This is a good example of the benefit of the ABD method. Also, the long ROH found in WGS cases on BTA18 (Figure 2) was not a feature of this larger set of results. There was reduced variability of these results with a higher number of animals compared to those from the WGS dataset analysis with only nine animals.

4 Discussion

This work had two objectives; one general and the other more specific. The generally applied objective was to test the idea that it is possible to find the site of a novel autosomal reccesive variant using just a small number of whole-genome-sequenced animals and appropriate bioinformatic methods, thus circumventing the need for the commonly-used two-stage approach highlighted by the review of Pollott (2018) or the collection and/or use of further data. The specific objective was to find the site of a new autosomal recessive condition thought to exist in Irish Moiled cattle.

4.1 Bioinformatic methods used with WGS data to find the site of a new autosomal recessive variant using a small number of cases and controls

Whole genome sequencing is becoming more widely used to locate single novel variants with major effects and a number of approaches have been used. In a large scale analysis of over Holstein cattle WGSs seven dominant conditions were located using the genome criteria approach (Bourneuf et al, 2017) involving one case for each of the seven conditions and a control population of 1,230 animals. The trio approach has been used by a number of authors (see for example Sayyab et al., 2016). This method takes WGSs of one affected offspring and its two parents and uses the genotype criteria method to find possible sites for the causative variant. The large number of sites identified are further reduced by a range of methods. Using one dog example (Sayyab et al., 2016), a filtering pipeline was established with 7 steps, including genotype criteria and SIFT analysis, Sanger sequencing verification and sequencing of an additional 24 cases/controls. Runs of homozygosity methods have been widely used with SNP-chip data (Pollott, 2018) and Letko et al. (2020) report an example of using this method in Zwartbles sheep to locate a novel autosomal recessive condition associated with Type 1 Primary Hyperoxaluria. Their study relied upon additional data from both the Sheep Genomes project and 79 publicly available genomes of various breeds to provide ’control’ data for the GCR method.

In the current study four bioinformatic methods were tested to try to find the location of the new variant causing early calf death in the nine Irish Moiled animals and relied on the ‘correct’ site appearing in all four methods. No additional data from the Irish Moiled or any other breeds were used. Two of the methods (GCR and ABD) do not require any prior information about genes, SNP or other genomic features but rely on across- and within-animal patterns of information contained in the genotypes at each SNV found in the VCF file. Ideally one would like to use these two methods alone since they are not only independent of any prior knowledge about genome features (except the reference genome for the alignments and generation of the VCF file) but they will also be able to find a new variant causing an autosomal recessive condition anywhere on the genome, even when located outsde a protein coding region: a useful feature of the two methods. As has been seen, using three cases and six controls with the ABD method and GCR combined revealed 12 possible sites in a 7.795 Gb stretch of Chr4 between g.70889821 and g.78684588 (ARS-UCD1.2 (GCF_002263795.1)) involving 18,705 SNVs. The underlying implication of this approach is that with more animals, either cases or controls, we would find fewer sites and so make the search somewhat more straightforward and find just a single causative variant site.

4.1.1 The genotype criteria approach

The method used to find the sites meeting the genotype criteria was based on a number of implied assumptions not stated by Pollott (2018). Firstly that GCR sites would be evenly distribution across the genome. Secondly, the higher the number of animals used the greater the chances of finding the GCR site of the new variant. Thirdly, a GCR site was not dependent on the balance of cases and controls in the samples. Fourthly, the location of a single GCR site was independent of the genetic relationship between cases and controls. Each assumption was tested using the data analysed in this study, either the final VCF file for BTA4 or the SNP-chip data with phenotypes allocted by the PCR results as appropriate. The detail of these investigations is given in the Supplementary File (Pages S10-17).

4.1.1.1 Evenly-spaced GCR sites across the genome

The results in Table 1 show that some chromosomes contain no GCR candidate sites at all (BTAs 3, 7, 8, 9, 10, 13, 15, 16, 18, 19, 20, 21, 22, 25, 27, 28 and 29). In fact these 17 chromosomes comprise 55% of the autosomal genome. Many chromosomes contained far fewer GCR sites than expected whereas others contained a much greater number than expected (BTAs 1 and 23). A GCR site (in this case) comprises two components; the 0/1 in all controls and the 1/1 in all cases (using 0 to mean the reference allele and 1 the new or alternative variant). We might expect the chromosomes containing long ROH in cases potentially to have many more GCR sites than others. Inspection of Figure S4 shows BTAs 4 and 18 as having the longest mean ROH but only BTA1 had a large excess of GCR sites. Using the information shown in the Supplementary File, the original implied assumption from Pollott (2018) of an even distribution of GCR sites across the genome has not been verified here and this has implications for the number of animals required to find a new variant site using WGS data alone. Clearly the long ROH on BTA 4 were linked to a large number of GCR sites but that on BTA18 was not.

4.1.1.2 Using more animals increases the chances of finding a causative GCR site

The second assumption about the use of GCR sites to locate a novel autosomal recessive variant was that the greater the number of animals that are genotyped the better the chance of locating the new variant. It appears, from the work reported in the Supplementary File, that the number of genotyped animals required to find the single candidate sites is 3 to 4 times greater than predicted using the original formula of Pollott (2018); one case and 29 controls appears to require the fewest total animals genotyped to find the single candidate site in the SNP-chip dataset used. However, this ‘long tail’ is due to several GCR sites being close together around the candidate site and always being ‘found’ in the generated datasets. Differentiating between them may require another method or using a different set of controls, perhaps from more distantly related individuals.

4.1.1.3 The balance of cases and controls

The basic calculation of the number of animals required to find a GCR site is independent of the balance between the number of cases and controls used. Figure S10 demonstrates that the lower the number of cases used the fewer the total number of animals required to be genotyped. At first sight these are rather startling results. However, outside the candidate site, it is much more unlikely to find all controls with a heterozygote genotype, whereas there will be many sites with all homozygous genotypes in cases; after all long ROH imply many 1/1 genotypes and so more sites potentially could meet the genotype criteria. The information in the Supplementary File clearly demonstrates that the number of animals required is much closer to the theoretical numbers when large ROH regions are excluded from the analyses. The number of animals required to find the candidate site is still slightly greater than the theoretical figure but this may be due to some other small areas of ROH not removed from the dataset. There are an enduring number of GCR sites in the high-ROH regions which inflate the results in contrast to the theoretical number of sites expected. This illustrates why theory and ’practice’ may differ.

4.1.1.4 The genetic relationship between cases and controls

The nine animals used in the WGS analysis comprised three cases and six parental controls. In order to investigate whether the closely-related controls might inflate the number of GCR sites found, an alternative SNP-chip dataset was derived using controls that were most distantly related to the cases (Supplementary File and Figure S12). In this case the number of animals required to find the new variant site was much closer to the theoretical expectation than with the parental controls.

4.1.1.5 Final genotype criteria summary

In summary, the method of locating the site of a new autosomal recessive variant using genotype criteria has potential. The theoretical expectation of the number of animals needed to be genotyped underestimates the actual number required if the method was the sole way to find the new variant. The 'simulated' datasets in the Supplementary File indicated that the outcome of any *single* set of genotyped animals is likely to vary widely, so needing fewer animals is just as likely as needing more, but of course it is unpredictable for any given set of animals. The best chance of success is likely to come from genotyping slightly more than the theoretical number of animals, using 33-50% from cases and the remainder from more distantly related controls (i.e. any animals from the breed). Relating this back to the 574 GCR sites found from the three cases and six parental controls used in the WGS analyses, here the number of cases used was at the lower bound of this recommendation (33%). The excess of GCR sites found was likely to have been caused by using parental controls, although it was well within the bounds of variability estimated from the SNP data.

4.1.2 The Autozygosity-By-Difference method

The ABD method was developed originally to locate regions of the genome likely to contain a new autosomal recessive variant using SNP data (see for example Posbergh et al., 2018). In the current analysis it has been applied to WGS data for the first time, as well as being used to confirm the results in an independent sample of SNP-genotyped animals. Because the method is sensitive to incorrectly called genotypes, a feature of WGS data, it was necessary to employ hard filtering criteria (see Table S1) of both sites and genotypes in order to get a useable set of data. In this case the VCF file was reduced from ~12 million to ~630,000 sites but this would differ under alternative hard-filtering criteria. Since VQSR methods were not available in the current situation (due to the species and number of animals genotyped) an alternative approach was taken; selecting sites and genotypes to fall with ±2 s.d. of the mean (or peaks in the case of bimodal variables). This allowed the location of several long ROH, one of which was found, by additional methods, to contain the new variant.

As well as using alternative hard-filtering criteria it may be possible to use other approaches, including site sampling or sliding-windows, to locate the region containing the new autosomal recessive variant. Using a site-sampling approach with a VCF file one could randomly select, say, 5-10% of sites evenly spread across the genome with the ABD method. Repeated samples of these SNVs, (say 100), could be randomly drawn and the 100 sets of ABD results averaged at each site. Alternatively, one could use a sliding window of, say, 10,000 base positions and count the number of homozygous variant case and control genotypes in each window. The window would then be moved along the genome at a given interval, say every 1,000 base positions, and the results plotted. One would expect to find the new variant causing the autosomal recessive condition in the region with the highest ABD-type score. Both these methods would overcome the problem of a single incorrectly-called genotype disrupting the long ROH in the ABD results and the need for hard filtering.

Using ABD on the hard-filtered WGS data resulted in the identification of two regions of the genome having an ABD score above the 0.01 probability threshold, and therefore likely to contain the new variant (Figure 2). Two aspects of Figures 2 and S4 are of note. Firstly there were a number of long ROH found in the controls throughout the genome with BTA18 having the largest mean ROH length. This was also found in the cases but the effect of combining the two sets of data in the ABD score was to remove many of these 'breed-specific' ROH and leave those which probably harboured the new variant. This is one of the advantages of the ABD method, particularly in rare breeds, but it has also been shown to be effective in removing long ROH associated with new variants in the myostatin gene in both Texel sheep (Pollott, 2013) and Piedmontese cattle (Biscarini et al., 2013).

The ABD method was also used for another purpose in this work; to confirm the WGS results on an independent set of animals with SNP-chip-derived data (Figure 3). As with the WGS ABD results, there were many long ROH in the SNP-chip dataset in both cases and controls. In this instance there were two very long ROH on BTA4 and BTA21, but interestingly not BTA18 as was found in the WGS data. The smaller number of animals used in the WGS analysis probably resulted in BTA18 having a long ROH due to sampling of closely related individuals. Figure S5 also shows BTA18 to have a long ROH but it was less pronounced in this larger dataset. In Figure 3 the result of subtracting the control ROHs from that of cases at each site reduced the noise considerably and left BTA4 as the only significant peak by a considerable margin. Once again it has been demonstrated that the power of the ABD method to remove noise works effectively.

4.1.3 Combining genotype criteria and ABD results

The approach in this work has been to use several bioinformatic methods on WGS data to see if they can pinpoint the site of a causal variant of a new autosomal recessive condition. Table 2 has highlighted a significant region on BTA4 using the ABD method that contained 12 sites meeting the genotype criteria. The discussion above has suggested that using more animals may have reduced the number of candidate sites by a small amount but a greater use of unrelated controls may have reduced the number of GCR sites in the target area more effectively.

4.1.4 SIFT score

In this set of results the SIFT scores were the crucial factor in determining the site of the new variant. Position Chr4: g.77173487A>T (ARS-UCD1.2 (GCF_002263795.1) was the only one of the 18 GCR sites in the target area to have a high-impact SIFT score (0-0.05). However, only sites in or near a coding region are scored using the SIFT method so it is not always going to find the causative site if it is located outside these regions of the genome.

4.1.5 RS number

The use of the absence of a RS number could be useful but, in this case, did not prove to be the final factor locating the novel variant site.

4.2 Other types of inheritance and effects

This paper reports the search for a new autosomal recessive variant causing a fatal condition in calves using a range of bioinformatic methods. It raises two issues relating to non-fatal autosomal recessive conditions and to other modes of inheritance.

4.2.1 Non-fatal conditions

There is nothing particularly special about the search for the causal variant of new fatal genetic diseases compared to non-fatal conditions. Clearly the potential phenotype is obvious but it may be harder in the initial stages of a new disease to correctly phenotype the dead animals. With a non-fatal condition it is still necessary to correctly phenotype the animals but there are likely to be more opportunities to test them and to return to them for samples, in many instances.

4.2.2 Other modes of inheritance

Having suggested this WGS approach for finding a new autosomal recessive variant the question arises about its usefulness with variants involving other modes of inheritance.

A dominance mode of inheritance can be thought of as the reverse of the recessive mode. One would expect to find cases to be 0/1 or 1/1 and controls to be 0/0 so the genotype criteria would be different compared to the recessive case. However, the number of animals required to use the GCR method would be very similar with cases being $2/(3^n)$ and controls $1/(3^m)$; the numerator having little effect with such a large denominator. The ABD method could only be used if all cases were 1/1 but that is unlikely with a dominant condition due to the large number of heterozygotes likely to be in the population. Alternatively, if there was some way to phenotypically distinguish 1/1 from 1/0 cases this would be useful. The 0/0 controls are unlikely to be situated in long ROH since they are likely to have been subjected to many generations of recombination, so alternative methods may be required. The new dominant variant would be situated in a long haplotype so it may be possible to adapt haplotype discovery methods to this situation. Both the SIFT score and RS number methods would be applicable but they are less powerful than the other two both because they rely on previous knowledge and, in the case of SIFT, it only works for a limited distance around a protein-coding region.

These methods could be used for a recessive sex-linked new variant i.e. one found on the X chromosome. Males would provide no useful data in this case so only females would be required. Both the ABD and the GCR methods would work the same way but with a lot fewer sites to search (only the X chromosome data would be needed).

4.3 Finding the causative variant for a perinatal mortality syndrome in Irish Moiled Cattle

The likely site for the causative variant of this fatal perinatal condition in Irish Moiled animals has been successfully located using just six parental control animals and three cases. Perinatal mortality (within 24 h of birth) typically occurs in about 6-10% of calves born (Brickell et al., 2009) with a further 3-4% dying in their first month, mainly from infectious disease (Johnson et al., 2017). The site highlighted at Chr4: g.77173487A>T (ARS-UCD1.2 (GCF_002263795.1) was located in the glucokinase gene (*GCK*) and is a splice acceptor variant. Analysis of the OMIA website (OMIA, 2020; Nicholas and Hobbs, 2012) showed splice acceptor variants to be responsible for ~8% of known variants in non-laboratory animals. There was a clear difference in the PHYRE2 prediction of the secondary structures between the reference genome and new variant *GCK* model. As observed from the SIFT-score results this is expected to have a disruptive effect on the operation of the *GCK* gene.

Glukokinase is a key enzyme found in the liver, pancreas, brain and endocrine cells of the gut. It catalyses the starting point of glycolysis by phosphorylating glucose to form glucose-6-phosphate (Matschinsky et al., 1993). The crystal structure has revealed that glucose binds in a deep cleft between a large and small domain of GCK, resulting in a conformational change and enzyme activation (Kamata et al., 2004). Glucokinase stimulates glucose uptake, glycolysis and glycogen synthesis by hepatocytes, whereas in pancreatic β-cells it plays a crucial role in glucose-stimulated insulin secretion. Glucose homeostasis is essential in mammals and is under tight endocrine control, with insulin acting as the key regulator.

There are currently 922 SNPs listed within the bovine *GCK* gene (NCBI, 2019) but the closest to the new variant site flanking either side were at Chr4: g.77173441 (an intron variant) and Chr4: g.77174392 (ARS-UCD1.2 (GCF_002263795.1)), some 46 and 905 bp away respectively. A segment of the ARS-UCD1.2 (GCF_002263795.1) genome 30 bp either side of the candidate variant was selected and blastn (Altschul et al., 1990) was used with the 61 bp sequence to find any homologous region on the human genome. A 40 bp length of sequence was found with 35 identical bases and a score of 50.9 bits (55) and no gaps. This was located on the reverse strand of human Chr7: g.44146590 to g.44146629 (GRCh38.p13 (GCF_000001405.39), in the *GCK* gene. The location on the human genome equivalent to the candidate variant found in Irish Moiled calves was at Chr7: g.44146620 (GRCh38.p13 (GCF_000001405.39). This was a highly conserved site with 96 out of the 100 vertebrate genomes shown on the UCSC (UCSC, 2020) genome browser all having a T on the forward strand, the remaining 4 being not reported. No SNP was found at this site in the human database but there was a SNP reported at the adjacent position (Chr7: g.44146619; GRCh38.p13 (GCF_000001405.39)) which was catalogued as rs1167675604, a C>T change on the forward strand. This site was also highly conserved in 96 out of the 100 vertebrate genomes on the UCSC genome browser and was also a splice-site acceptor variant. The ClinVar (ClinVar, 2020) record for this variant states that "The variant disrupts a canonical splice site, and is therefore predicted to result in the loss of a functional protein. Found in at least one symptomatic patient, and not found in general population data." It's incidence was estimated to be well below 0.001% of the population. In addition the Varsome (Varsome, 2020) record for this SNP states that the effect of the variant was 'Very Strong' which means " Null variant (intronic within ±2 of splice site) affecting gene *GCK*, which is a known mechanism of disease (gene has 378 known pathogenic variants which is greater than minimum of 3), associated with Diabetes mellitus, permanent neonatal 1, Maturity-onset diabetes of the young, type 2 and Hyperinsulinemic hypoglycemia, familial 3."

The mouse genome was also investigated in the same way but no SNP were found in the candidate region.

Over 600 variants have been reported in the human *GCK* gene, which have varying effects depending on their location (Osbak et al., 2009; OMIM 138079). Heterozygous inactivating variants cause a condition known as maturity onset diabetes of the young, characterised by mild fasting hyperglycaemia. Homozygotes are much more rare in the human population, and neonates present earlier with permanent neonatal diabetes mellitus. In mice, however, pups born with global *GCK* knockout (-/-) are slightly smaller than wildtype animals (+/+), have glucose levels about eightfold higher and die within 3 to 5 days (Grupe et al., 1995). Tissue specific β-cell knockouts die within 4 days of birth whereas hepatic knockout impairs glucose utilization and glycogen synthesis but with only mild hyperglycaemia (Postic et al., 1999).

Pregnancy outcome in women depends on a combination of the genotype of both mother and fetus (Spyer et al., 2001). When the fetus carries a single *GCK* variant this affects glucose homeostasis with reduced insulin secretion, so both placental and birth weight are reduced (Hattersley et al., 1998; Spyer et al., 2008). During pregnancy, the fetal glucose supply is derived almost entirely from the dam across the placenta using facilitated diffusion by glucose transporters. In ruminants this uptake is regulated sequentially by GLUT1 and GLUT3 (SLC2A1 and SLC2A3) (Wooding et al., 2005).

The fetus has a low capacity for endogenous glucose production but this increases in late gestation, in response to the pre-term rise in glucocorticoid production, together with catecholamine and thyroid hormone stimulation. These promote hepatic glycogen synthesis and gluconeogenesis, which are essential in providing the neonatal calf with an adequate glucose supply as milk lactose on its own is insufficient (Hammon et al., 2013). The postnatal maturation in the regulation of energy supply may thus explain why lack of GCK activity is fatal at this stage of life.

4.4 Concluding remarks

The original intention for this work was to locate the site of a novel variant causing perinatal mortality in Irish Moiled calves. This has been achieved, and shown to be located in the *GCK* gene, but in the process it became apparent that there were no straightforward ways to achieve this objective. At best, a two-stage approach was required involving genotyping a group of cases and controls, identifying the genomic region likely to contain the novel variant followed by further work to sequence the identified region and look for appropriate signals in the data. Consequently, a further objective was set in order to simplify the process and investigate whether it would be possible to use a single whole genome sequencing stage with appropriate bioinformatic methodology to find the candidate site. This, too has been achieved by sequencing nine animals, three cases and six parental controls, and applying four methods to the data. In the process it has been possible to investigate some of these methods in more detail and arrive at some general conclusions to aid future such studies.

The VCF file format has proved to be a very practical source of data for this study particularly because it reduced the search ‘area’ from over 2.5 billion base positions down to one involving only 8 million sites. In addition, the VCF file format facilitated finding the novel site when combined with methods to interrogate it for genotype criteria, long runs of homozygosity and the predicted effects of variants on the phenotype of the animal. Using these three methods allowed the identification of a single variant site which was found to have both the genomic and biological properties associated with this novel condition.

In the process of carrying out this work it has been possible to refine the genotype criteria method to demonstrate that in reality only a small number of cases and controls *is* required, controls should outnumber cases by 2:1 and controls should be more distantly related to cases. In addition it has been possible to show that using a runs-of-homozygosity method, previously only used on SNP-chip genotype data, with whole-genome-sequence data it was possible to locate the region of the genome containing the novel variant.

In future it should be possible to use the combination of genotype criteria and runs-of-homozygosity methods with the appropriate number of cases and controls, suitably distantly related, to locate the site of any new autosomal recessive genetic condition in a relative short time. This should then facilitate a more speedy elimination of the harmful variant from the population by using an appropriate genetic test on available animals.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

GP and DCW designed the study. GP analysed the data, wrote the ABD software, carried out the bioinformatic analyses and wrote the first draft of the paper. MS produced the alignments, VCF files, SIFT and PHYRE2 scores. RP and CM carried out the Sanger sequencing. GP, DCW, RP and MS contributed written material to the final paper.

Funding

We would like to acknowledge the Biotechnology and Biological Sciences Research Council for Stephanie Hill's funding under the Vacation Bursaries scheme in the early stages of this research. The Farm Animal Genetics and Genomics (Genesis) Faraday Partnership - LS1618 is also acknowledged for funding some of the SNP genotyping work used in the confirmatory stage of this research.

Acknowledgments

We would like to acknowledge the help given to us by Irish Moiled cattle breeders who gave us access to their records, and to the breed society for making all the pedigree data etc. available. Dr Zhangrui Cheng and Ms Lauren Hamstead are thanked for their technical assistance with the laboratory analyses.

Data Availability Statement

The data is owned by the Irish Moiled Breeders.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. 215, 403-410. doi: 10.1016/S0022-2836(05)80360-2

Biscarini, F., Corvo, M., del, Stella, A., Albera, A., Ferenčakovic, M., and Pollott, G.E. (2013). Looking for the mutations for arthrogryposis and macroglossia in Piedmontese cattle: preliminary results. J. sobre Prod. Anim., Zaragoza 14 y 15 de mayo de 2013. 538-540.

Bourneuf, E., Otz, P., Pausch, H., et al. (2017). Rapid discovery of de novo deleterious mutations in cattle enhances the value of livestock as model species. Sci. Rep. 7, 11466. doi: 10.1038/s41598-017-11523-3.

Brickell, J.S., McGowan, M.M., Pfeiffer, D.U., and Wathes, D.C. (2009). Mortality in Holstein-Friesian calves and replacement heifers, in relation to body weight and IGF-I concentration, on 19 farms in England. Anim. 3, 1175-82. doi: 10.1017/S175173110900456X.

Broad Institute (2020). https://gatk.broadinstitute.org/hc/en-us. [Accessed July 26, 2021].

Chang, C.C., Chow, C.C., Tellier, L.C.A.M., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience. 4, s13742-015-0047-8. doi: https://doi.org/10.1186/s13742-015-0047-8.

ClinVar. (2020). https://clinvarminer.genetics.utah.edu/ [Accessed October 20, 2020].

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., McVean, G., and Durbin, R. (2011). The variant call format and VCFtools. Bioinformatics (Oxford, England). 27, 2156-8. doi: 10.1093/bioinformatics/btr330. Epub 2011 Jun 7.

Ensembl. (2020) http://oct2018.archive.ensembl.org/Bos_taurus/Info/Annotation [Accessed October 20, 2020].

Fisher, R. A. (1922). On the interpretation of $\chi 2$ from contingency tables, and the calculation of P. J. R. Stat. Soc. 85, 87–94. doi:10.2307/2340521. JSTOR 2340521.

Freeman, G.H., and Halton, J.H. (1951). Note on exact treatment of contingency, goodness of fit and other problems of significance. Biometrika. 38, 141-149.

Graffelman, J., and Moreno, V. (2013). The mid p-value in exact tests for Hardy-Weinberg equilibrium. Stat. Appl. Genet. Mol. Biol. 12, 433-448. doi:10.1515/sagmb-2012-0039

Grupe, A., Hultgren, B., Ryan, A., Ma, Y.H., Bauer, M., and Stewart, T,A. (1995), Transgenic knockouts reveal a critical requirement for pancreatic beta cell glucokinase in maintaining glucose homeostasis. Cell. 83, 69-78. doi: 10.1016/0092-8674(95)90235-x.

Hammon, H.M., Steinhoff-Wagner, J., Flor, J., Schönhusen, U., and Metges, C.C. (2013). Lactation Biology Symposium: role of colostrum and colostrum components on glucose metabolism in neonatal calves. J. Anim. Sci. 91, 685-95. doi: 10.2527/jas.2012-5758.

Hattersley, A.T., Beards, F., Ballantyne, E., Appleton, M., Harvey, R., and Ellard, S. (1998). Mutations in the glucokinase gene of the fetus result in reduced birth weight. Nat. Genet. 19, 268-70. doi: 10.1038/953.

Irish Moiled Cattle Society. (2020). https://www.irishmoiledcattlesociety.com/breed-history/. [Accessed October 20, 2020]

Johnson, K.F., Chancellor, N., Burn, C.C., and Wathes, D.C. (2017). Prospective cohort study to assess rates of contagious disease in pre-weaned UK dairy heifers: management practices, passive transfer of immunity and associated calf health. Vet. Rec. Open. 4, e000226. doi: 10.1136/vetreco-2017-000226.

Kamata, K., Mitsuya, M., Nishimura, T., Eiki, J., and Nagata, Y. (2004). Structural basis for allosteric regulation of the monomeric allosteric enzyme human glucokinase. Structure. 12, 429-38. doi: 10.1016/j.str.2004.02.005.

Kelley, L., Mezulis, S., Yates, C. et al. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc. 10, 845–858. https://doi.org/10.1038/nprot.2015.053

Letko, A., Dijkman, R., Strugnell, B., Häfliger, I.M., Paris, J.M., Henderson, K., Geraghty, T., Orr, H., Scholes, S., and Drögemüller, C. (2020). Deleterious AGXT missense variant associated with Type 1 primary hyperoxaluria (PH1) in Zwartbles sheep. Genes (Basel). 11, 1147. https://doi.org/10.3390/genes11101147.

Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., and Chen, W.M. (2010). Robust relationship inference in genome-wide association studies. Bioinformatics. 26, 2867-2873. doi: 10.1093/bioinformatics/btq559.

Matschinsky, F., Liang, Y., Kesavan, P., Wang, L., Froguel, P., Velho, G., Cohen, D., Permutt, M.A., Tanizawa, Y., Jetton, T.L., *et al*. (1993). Glucokinase as pancreatic beta cell glucose sensor and diabetes gene. J. Clin. Invest. 92, 2092–2098. doi: 10.1172/JCI116809.

McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome Biol. 17, 122. doi:10.1186/s13059-016-0974-4.

NCBI (2019). https://www.ncbi.nlm.nih.gov/snp/. [Accessed October 20, 2020]

Ng, P.C., and Henikoff, S. (2002). Accounting for human polymorphisms predicted to affect protein function. Genome Res. 12: 436-446. doi: 10.1101/gr.212802.

Nicholas, F.W., and Hobbs, M. (2013). Mutation discovery for Mendelian traits in non-laboratory animals: a review of achievements up to 2012. Anim. Genet. 45,157-70. doi: 10.1111/age.12103.

Nicolazzi, E.L., Caprera, A., Nazzicari, N., Cozzi, P., Strozzi, F., Lawley, C., Pirani, A., Soans, C., Brew, F., Jorjani, H., Evans, G., Simpson, B., Tosser-Klopp, G., Brauning, R., Williams, J.L., and Stella A. (2015). SNPchiMp v.3: integrating and standardizing single nucleotide polymorphism data for livestock species. BMC Genomics, 16:283. doi: https://doi.org/10.1186/s12864-015-1497-1

Nicolazzi, E.L., Picciolini, M., Strozzi, F., Schnabel, R.D., Lawley, C., Pirani, A., Brew, F. and Stella A. (2014). SNPchiMp: A database to disentangle the SNPchip jungle in bovine livestock. BMC Genomics. 15, 123. doi: 10.1186/1471-2164-15-123.

OMIA. (2020). Online Mendelian Inheritance in Animals, OMIA. Sydney School of Veterinary Science, https://omia.org/. [Accessed October 20, 2020].

Osbak, K.K., Colclough, K., Saint-Martin, C., Beer, N.L., Bellanné-Chantelot, C., Ellard, S., and Gloyn, A,L. (2009), Update on mutations in glucokinase (GCK), which cause maturity-onset diabetes of the young, permanent neonatal diabetes, and hyperinsulinemic hypoglycemia. Hum. Mutat. *30*, 1512-26. doi: 10.1002/humu.21110.

PLINK. (2017) https://zzz.bwh.harvard.edu/plink/contact.shtml. [Accessed 20 October, 2020].

Pollott, G.E. (2013). Do selective sweeps in sheep breeds indicate the genomic sites of breed characteristics? Book of abstracts of the 64th European Association for Animal Production Annual Meeting, Nantes, France. p627. Wageningen Academic Publishers; Netherlands.

Pollott, G. E. (2018). Invited review: Bioinformatic methods to discover the likely causal variant of a new autosomal recessive genetic condition using genome-wide data. Anim. 12, 2221–2234. doi: 10.1017/S1751731118001970.

Posbergh, C.J., Pollott, G.E., Southard, T.L., Divers, T.J., and Brooks, S.A. (2018). A non-synonymous change in adhesion G protein-coupled receptor L3 associated with risk for Equine Degenerative Myeloencephalopathy in the Caspian Horse. J. Equine. Vet. Sci. 70, 96-100.

Postic, C., Shiota, M., Niswender, K.D., Jetton, T.L., Chen, Y., Moates, J.M., Shelton, K.D., Lindner, J., Cherrington, A.D., and Magnuson, M.A. (1999). Dual roles for glucokinase in glucose homeostasis as determined by liver and pancreatic beta cell-specific gene knock-outs using Cre recombinase. J. Biol. Chem. 274, 305-15. doi: 10.1074/jbc.274.1.305.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., and Sham, P.C. (2007). PLINK: a toolset for whole-genome association and population-based linkage analysis. Am. J. Hum. Genet. 81, 559-75. doi: 10.1086/519795.

Sayyab, S., Viluma, A., Bergvall, K., Brunberg, E., Jagannathan, V., Leeb, T., Andersson, G., and Bergström, T.F. (2016). Whole-Genome sequencing of a canine family trio reveals a FAM83G variant associated with hereditary footpad hyperkeratosis. G3 (Bethesda). 6, 521-7. doi: 10.1534/g3.115.025643.

Spyer, G., Hattersley, A.T., Sykes, J.E., Sturley, R.H., MacLeod, and K.M. (2001). Influence of maternal and fetal glucokinase mutations in gestational diabetes. Am. J. Obstet. Gynecol. 185, 240-1. doi: 10.1067/mob.2001.113127.

Spyer, G., Slingerland, A.S., Knight, B.A., Ellard, S., Clark, P.M., Hauguel-de Mouzon, S., and Hattersley, A,T. (2008), Mutations in the glucokinase gene of the fetus result in reduced placental weight. Diabetes Care. 31, 753-7. doi: 10.2337/dc07-1750.

UCSC. (2020) https://genome.ucsc.edu/. [Accessed October 20, 2020].

Varsome. (2020) https://varsome.com/variant/hg19/rs1167675604. [Accessed October 20, 2020].

Wooding, F.B., Fowden, A.L., Bell, A.W., Ehrhardt, R.A., Limesand, S.W., and Hay, W.W. (2005). Localisation of glucose transport in the ruminant placenta: implications for sequential use of transporter isoforms. Placenta. 26, 626-40. doi: 10.1016/j.placenta.2004.09.013.

Table 1 A summary of results by chromosome

BTA	Length (bp)	Number of SNV in VCF file	Estimated number of GCR	Actual number of GCR	Number of SIFT sites	Number of NoRS9 sites
1	158,534,110	535,135	27	489	251	648
2	136,231,102	452,534	23	2	279	164
3	121,005,158	394,128	20	0	383	271
4	120,000,601	373,759	19	13	485	342
5	120,089,316	401,056	20	1	458	173
6	117,806,340	362,930	18	2	174	220
7	110,682,743	358,947	18	0	491	249
8	113,319,770	319,261	16	0	204	485
9	105,454,467	328,559	17	0	179	184
10	103,308,737	353,791	18	0	293	161
11	106,982,474	324,386	16	6	337	132
12	87,216,183	335,302	17	15	139	145
13	83,472,345	232,616	12	0	249	188
14	82,403,003	262,990	13	1	108	135
15	85,007,780	285,748	15	0	374	158
16	81,013,979	268,622	14	0	225	388
17	73,167,244	278,406	14	1	198	106
18	65,820,629	194,623	10	0	522	384
19	63,449,741	227,124	12	0	449	110
20	71,974,595	230,092	12	0	110	84
21	69,862,954	202,249	10	0	205	260
22	60,773,035	180,922	9	0	147	164
23	52,498,615	255,467	13	43	566	333
24	62,317,253	228,628	12	0	83	107
25	42,350,435	142,293	7	0	225	43
26	51,992,305	177,147	9	1	133	120
27	45,612,108	175,471	9	0	99	87
28	45,940,150	179,521	9	0	92	99
29	51,098,607	172,660	9	0	306	358

Total	2,489,385,779	8,234,367	418	574	7,764	6,298

Legend: BTA = chromosome number. GCR = genotype criteria sites. NoRS9 = number of site with no RS number and with at least one alaternate allele in all 9 genotypes.

Table 2 The overlap between the four methods for locating a likely novel variant site

Method	Genotype criteria (GCR)	Autozygosity by difference (ABD)	High-impact SIFT score (SIFT)	No registered SNP number (NoRS9)
GCR	**574 (12)**			
ABD	12 (12)	**838 (575)**		
SIFT	1 (1)	12 (12)	**7,764 (12)**	
NoRS9	4 (4)	37 (36)	8 (1)	**6,298 (36)**
GCR+ABD			1 (1)	1 (1)
GCR+SIFT				1 (1)
ABD+SIFT				1 (1)
GCR+ABD+SIFT				1 (1)

Legend: The table shows the number of SNV in the final VCF file identified by each method. (numbers in the BTA4 high-ABD region shown in parentheses).

Table 3 Animal status by genotype for the 41 Sanger-sequenced animals at Chr4: g.77173487A>T (ARS-UCD1.2 (GCF_002263795.1)).

Animal status	AA	AT	TT	Total
Calves	4	2	7	13
Known adult carriers (live)	0	6	0	6
Status unknown adults (live)	13	9	0	22
Total	17	17	7	41

Figure 1 A chromosome containing a novel autosomal recessive variant demonstrating both the GCR and ABD methods

Legend. Heat map of 28 controls (left side of the figure) and 7 cases (right side of the figure) for a hypothetical chromosome containing a novel autosomal recessive variant. Chromosomes for the individuals run from top to bottom of the figure. The colors represent homozygous major allele genotypes (red), homozygous minor allele genotypes (green) and heterozygote genotypes (yellow). Biallelic variants are assumed. The solid black bar across the controls represents the position of the new variant with a GCR of all heterozygotes in controls and all the same homozygote in cases. The

ROH around the candidate position for cases are shown as a black rectangle running up and down the chromosome until a heterozygote is found. The ABD method takes the mean length, in Kb, of ROH in cases minus that in controls at each position to calculate the ABD score.

Figure 2 Manhattan plots of the ABD analysis of nine WGS animals (Kb)

Legend: P < 0.01 at ABD score = 9,034 Kb).

Figure 3 Manhattan plot of the ABD analysis of the SNP-chip analysis based on the genotypes found in the PCR analysis (Kb)

Legend: These results were based on animals with phenotyping informed by the PCR results. (P < 0.001 at ABD score = 7,023 Kb).