Royal Veterinary College
University of London

Using the Variable Effect Predictor with WGS data finds one causal variant for early death in calves

G. E. Pollott, M. Salavati and D. C. Wathes

The issue

- Whole genome sequence (WGS) data has the potential to assist in finding the site of a new deleterious variant causing a novel Mendelian disease
- The challenge is how to find the site of the new variant from ~3 billion base pairs in a mammalian genome (e.g. cattle).
- The search only needs to take place in that proportion of sites showing variation using, say, a Variant Call Format (VCF) file.
- Recent studies in cattle have found this to be 10 to 20 million sites, depending on the number of animals used and other factors e.g. breed
- How can we find the site of a new lethal recessive condition from the single nucleotide variants (SNV) in a VCF file?

RVC

Methods to reduce the search

1. For an autosomal recessive condition, use a suitable ‘runs of homozygosity’ (ROH) method
2. Search for base positions with the ‘correct’ genotype criteria i.e. similarly homozygous cases and heterozygous parental (carrier) controls
3. Use the Variant Effect Predictor to find variants with a ‘high-impact’ SIFT score

- Can we use any 2 of these methods in combination to find a novel recessive condition?

RVC

Dataset

- Irish Moiled calves suffering from an autosomal recessive condition causing postnatal mortality (~10 d of age)
- Parents, relatives and ‘unrelated’ animals
- 71 animals (21 cases; 50 controls) with genomewide SNP-chip data
- 8 WGS animals (3 cases; 5 parental controls)
- Using the 8 WGS animals – generated VCF files for each animal by alignment of the WGS reads to the reference genome (Btau UMD 3.1)
- Variant calling performed on the mapped reads using the Genome Analysis Toolkit (GATK)
- All 8 individual VCF files merged into 1 VCF file

RVC

Method 1 – ROH using the autozygosity by difference (ABD) method – SNP data

- ABD method (Pollott, 2018) looks for regions of the genotype with significantly longer ROH in cases than controls based on permuted probabilities
- Calculates mean length of ROH at each SNP for cases and controls separately, and their difference at each SNP
- Use UMD 3.1 build of Btau genome

G. E. Pollott (2018). Invited review: Bioinformatic methods to discover the likely causal variant of a new autosomal recessive genetic condition using genome-wide data. *Animal.* https://doi.org/10.1017/S1751731118001970

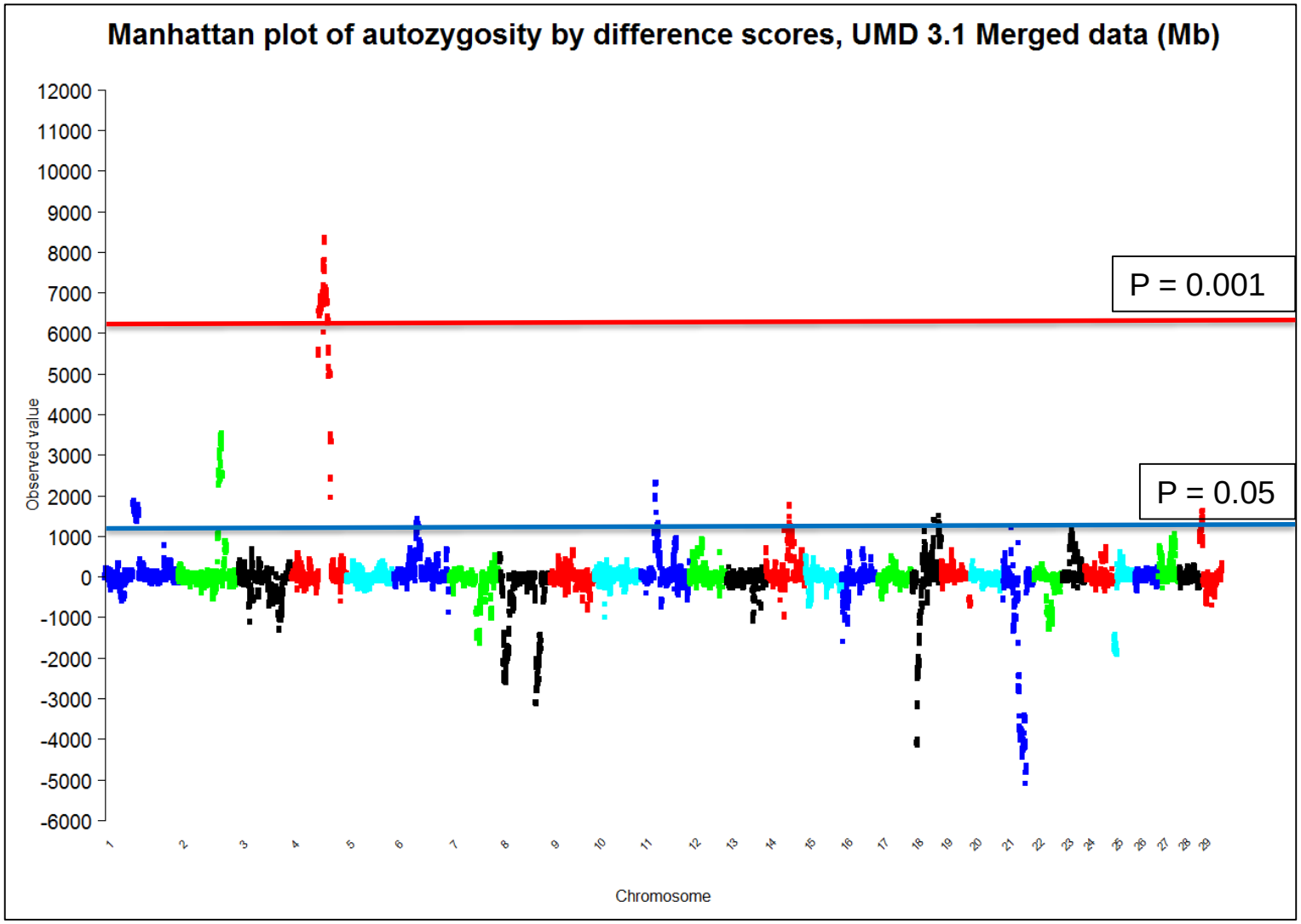
RVC

Method 1 – ROH using the autozygosity by difference (ABD) method – SNP data

- Identified the greatest ABD score to be on BTA4 between 68,658,134 and 79,396,400, a 11Mb length of chromosome
- Only significant region after 1,000 permutations ($P < 0.001$); mean ROH > 6,237Mb

RVC

Method 1 – ROH using the autozygosity by difference method



RVC

Method 2 – Genotype criteria with WGS data - theory

- Autosomal recessive condition genotype criteria
  - All cases homozygous for the same variant
  - All parental controls heterozygous for this variant
- With n cases and m controls then chance of finding a variant with the 'correct' genotype criteria $1/3^{(n+m)}$

- In this dataset with 8 animals and 13.8 million SNV we would expect ~ 2,100 positions
- In a 11 Mb ROH of 37,179 SNV we would expect ~ 6 positions

RVC

Method 2 – Genotype criteria with WGS data - Actual

- Autosomal recessive condition genotype criteria
  - All cases homozygous for the same variant
  - All parental controls heterozygous for this variant
  - 13.8 million SNV in 8 WGS animals

- 1,845 had the ‘correct’ genotype criteria (~ 2,100 predicted)
- 27 in the 11Mb identified by ROH analysis (~ 6 predicted)

RVC

Method 3 – High SIFT score using the Variant Effect Predictor with WGS data

- VCF file further annotated by the Variant Effect Prediction (VEP) tool of the ENSEMBL database
- Indicates the potential severe to moderate effects of a variant within and around coding regions (e.g. 5 kb up or downstream from the transcript start site)
- For all input variants, the VEP returns detailed annotation for effects on transcripts, proteins, and regulatory regions
- One of these is SIFT score (Sorting Intolerant From Tolerant)
- HIGH: The variant is assumed to have high (disruptive) impact on the protein, probably causing protein truncation, loss of function or triggering nonsense mediated decay.

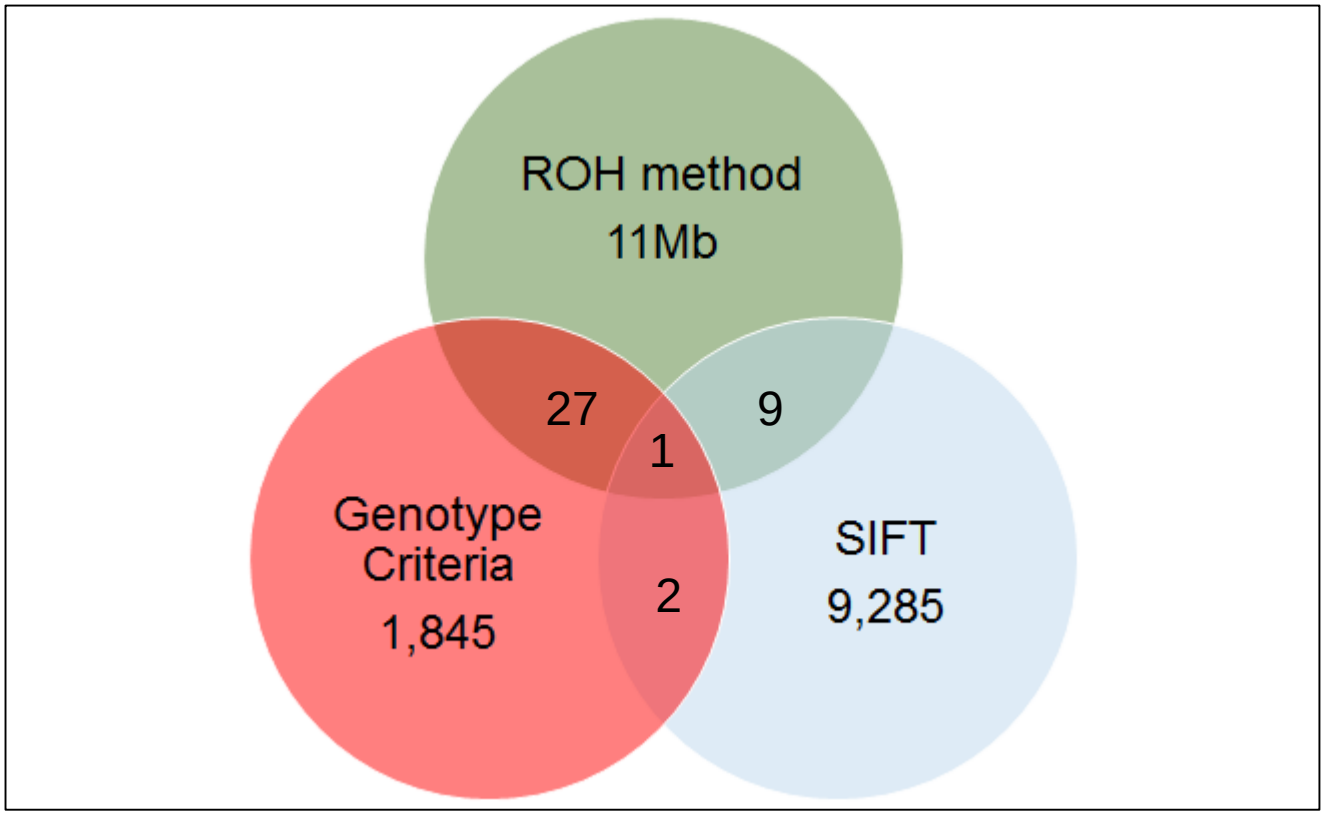
RVC

Method 3 – High SIFT score using the Variant Effect Predictor

- From the VCF files - located 9,285 variant sites on the whole genome with a ‘high-impact’ SIFT score.
- 2 had ‘correct’ genotype criteria for all 9 animals
- 1 in the 11Mb region on Chromosome 4 with the ‘correct’ genotype criteria

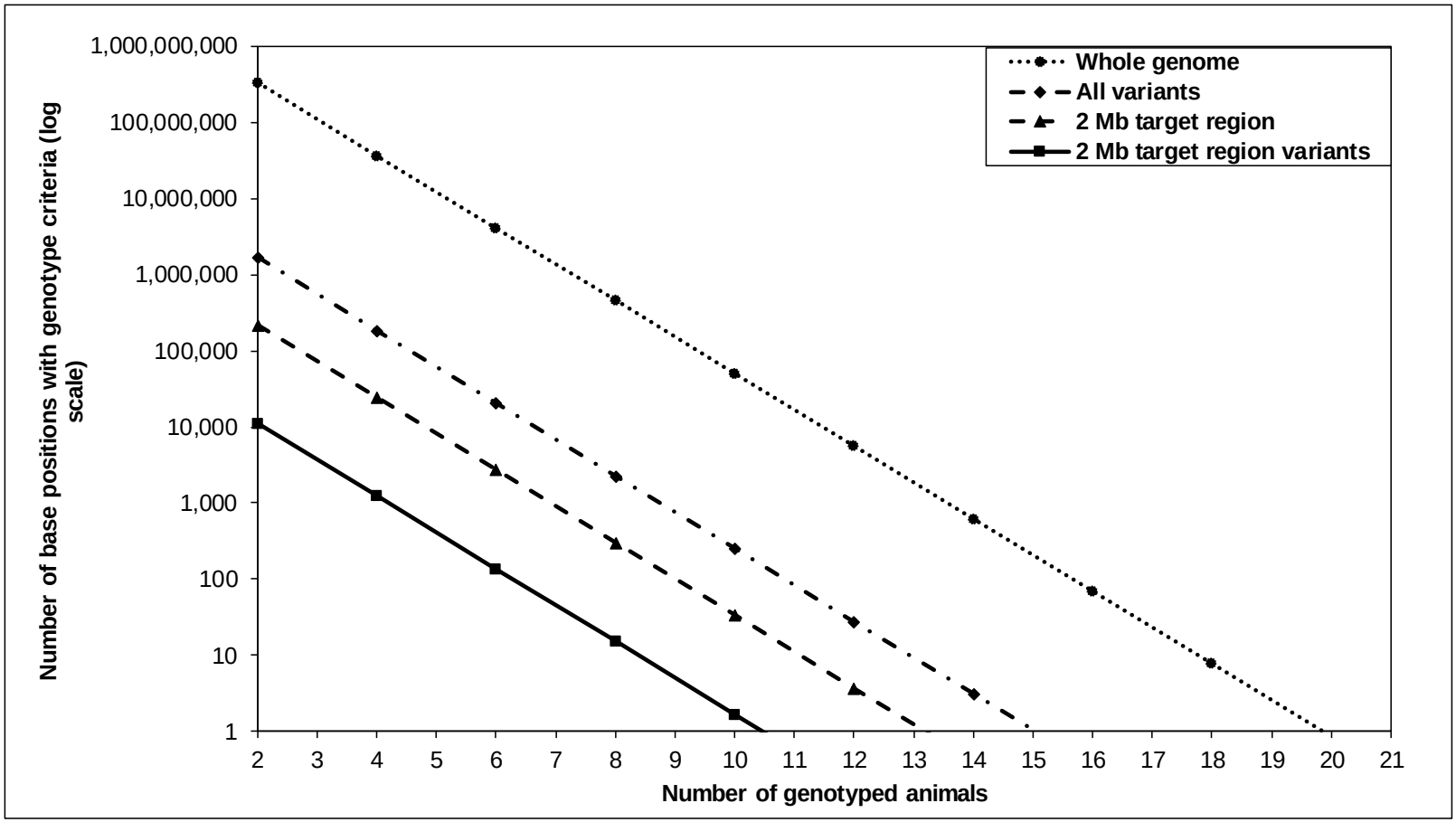
RVC

Methods results summary – likely sites



RVC

How many animals do we need?



RVC

Likely causal site

- Single-base change splice acceptor variant in the glucokinase gene (GKN) and is likely to have drastic protein folding changes (PHYRE2 prediction)
- Glucokinase plays a key role in glucose uptake and regulation of insulin secretion
- Variant not previously reported in cattle or human homolog
- Human mutations in GKN are associated with early-onset diabetes
- Future work will be undertaken in the breed to investigate these findings and implement a suitable breeding programme for controlling the condition

RVC

Final comments

- Small number of WGS samples required to find site of the causal variant of a new autosomal recessive condition
- Trade off between number of samples and number of methods required
- At least two independent methods needed
- Probably don't need the SNP-based methods if samples are limited

- Only need genotype criteria method if publically available 1000-bull genomes data available (or similar in other species; subject to permissions and good reference genome)

RVC

Acknowledgements

- Breeders for supplying data and DNA
- Genesis-Faraday and Rare Breeds Survival Trust for funding some SNP genotyping

RVC