Development of a Genetic Marker Panel to Predict Reproductive Longevity in Holstein Cattle

Kacper Żukowski[1], Nehil Jain[1], Jyoti Joshi[1], Jeremy E. Koenig[1], Robert G. Beiko[1] and Hein van der Steen[2]

DALHOUSIE UNIVERSITY
Inspiring Minds

[1]Faculty of Computer Science, Dalhousie University, Halifax, NS, Canada
[2]Performance Genomics Inc., Bible Hill, NS, Canada

Table of contents

DALHOUSIE UNIVERSITY
Inspiring Minds

Reproductive Longevity

DALHOUSIE UNIVERSITY
Inspiring Minds

Reproductive Longevity (RL) is the result of a complex of traits involving longevity, ovarian function, fertility, stress resistance (coping with the stress of lactation and reproduction), robustness and health. It is reasonable to assume that a relatively large number of genes and pathways are involved and that the genetic bottlenecks overlap and differ between mouse and cattle.

Reproductive Longevity

RL is affected by a lot of different traits, some of which are low in heritability...

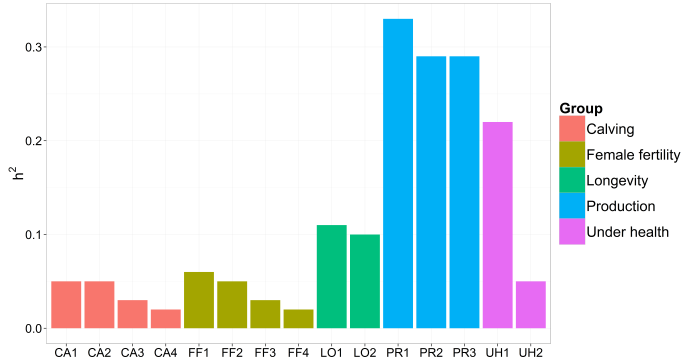
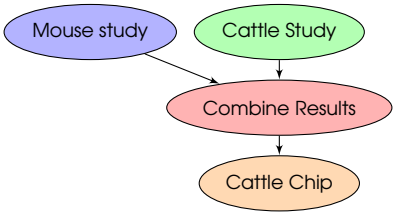


Fig. The average cattle heritability for *Interbull* countries across selected traits (official information, 2015).

The aim of the study

DALHOUSIE UNIVERSITY
Inspiring Minds

- identification of candidate genes associated with RL at different level of mouse and cattle study:
  - low density genome resolution,
  - high density genome resolution,
  - transcriptome,
- commercial DNA marker-based tests of RL for the Holstein cattle industry.



Mouse study

- Material
- Three studies: microarray, poolseq and rnaseq

Material, Unique mouse resource

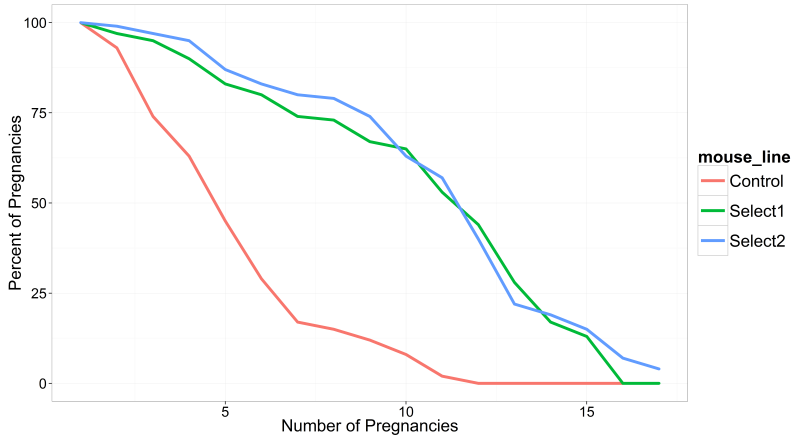
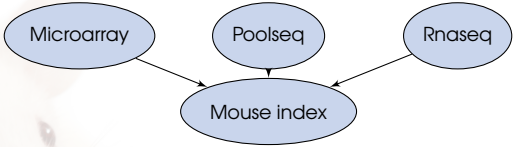


Fig. Effect of selection on number of lifetime pregnancies (Don Crober et al., 2007).

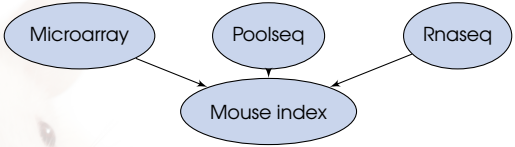
Methods and results



- contrast between select and control lines,

8 pooled samples, mean coverage 50x per pool,
Poolseq analysis based on Cochran-Mantel-Haenszel statistic,
after filtration 6.4M of SNPs,
gene selection based on modified TopQ method described by
580K highly significant SNPs within 13.4K genes,
two additional contrasts: ovary vs. pituitary tissue and reproductive vs.
reproductive females,

Methods and results



- contrast between select and control lines,
- 43k SNPs segregating with MAF > 0.05, LHI and LHGV statistics,

8 pooled samples, mean coverage 50x per pool,

Poolseq analysis based on Cochran-Mantel-Haenszel statistics,

after filtration 6.4M of SNPs,

gene selection based on modified TopQ method described by Le [illegible] 2011,

580K highly significant SNPs within 13.4K genes,

two additional contrasts: ovary vs. pituitary tissue and reproductive vs. [illegible] reproductive females,

Whole genome sequencing approaches (Poolseq)

DALHOUSIE UNIVERSITY
Inspiring Minds

- new method of cost-effectiveness sequencing,

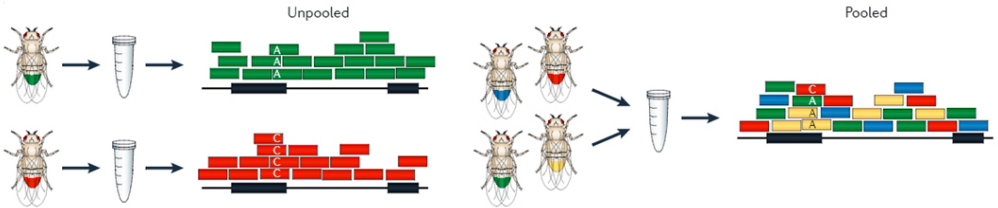


Fig. Comparison of sequencing strategies (Schlötterer et al., 2014).

Whole genome sequencing approaches (Poolseq)

- new method of cost-effectiveness sequencing,
- similar to reduced representation library, strategy used to discover large numbers of genome-wide SNP with high MAF (Matukumalli et. al 2009).

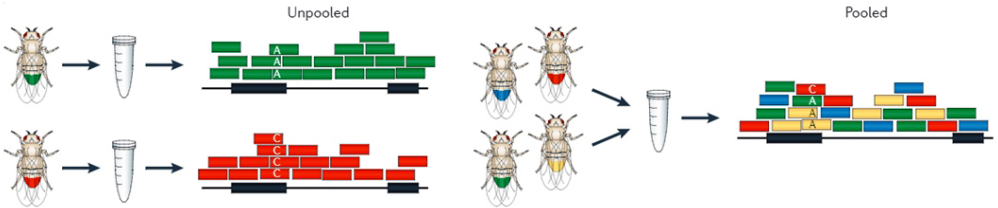
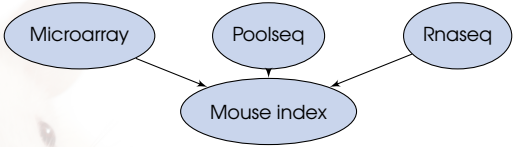


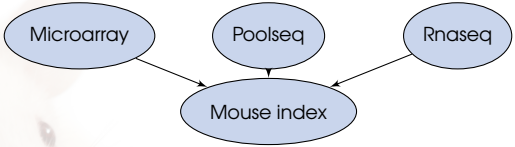
Fig. Comparison of sequencing strategies (Schlötterer et al., 2014).

Methods and results, continue



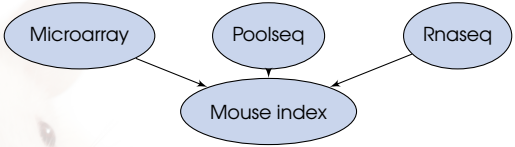
- contrast between select and control lines,
- 43k SNPs segregating with MAF > 0.05, LHI and LHGV statistics,
- 8 pooled samples, mean coverage 50x per pool,

Methods and results, continue



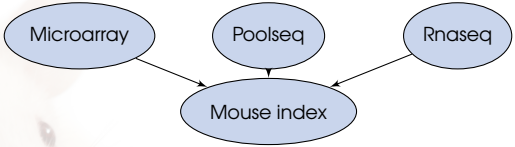
- contrast between select and control lines,
- 43k SNPs segregating with MAF > 0.05, LHI and LHGV statistics,
- 8 pooled samples, mean coverage 50x per pool,
- Poolseq analysis based on Cochran–Mantel–Haenszel statistics (Kofler et al., 2011),

Methods and results, continue



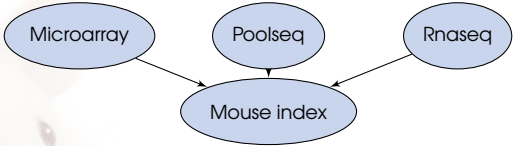
- contrast between select and control lines,
- 43k SNPs segregating with MAF > 0.05, LHI and LHGV statistics,
- 8 pooled samples, mean coverage 50x per pool,
- Poolseq analysis based on Cochran–Mantel–Haenszel statistics (Kofler et al., 2011),
  - after filtration 6.4M of SNPs,

Methods and results, continue



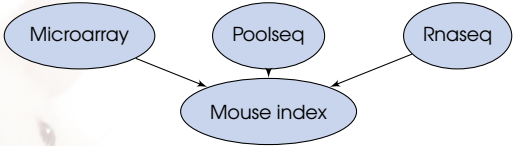
- contrast between select and control lines,
- 43k SNPs segregating with MAF > 0.05, LHI and LHGV statistics,
- 8 pooled samples, mean coverage 50x per pool,
- Poolseq analysis based on Cochran–Mantel–Haenszel statistics (Kofler et al., 2011),
  - after filtration 6.4M of SNPs,
- gene selection based on modified TopQ method described by Lehne et al., 2011,

Methods and results, continue



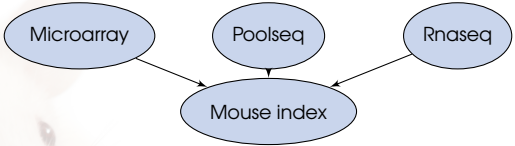
- contrast between select and control lines,
- 43k SNPs segregating with MAF > 0.05, LHI and LHGV statistics,
- 8 pooled samples, mean coverage 50x per pool,
- Poolseq analysis based on Cochran–Mantel–Haenszel statistics (Kofler et al., 2011),
  - after filtration 6.4M of SNPs,
- gene selection based on modified TopQ method described by Lehne et al., 2011,
  - 580K highly significant SNPs within 13.4K genes.

Methods and results, continue



- contrast between select and control lines,
- 43k SNPs segregating with MAF > 0.05, LHI and LHGV statistics,
- 8 pooled samples, mean coverage 50x per pool,
- Poolseq analysis based on Cochran–Mantel–Haenszel statistics (Kofler et al., 2011),
  - after filtration 6.4M of SNPs,
- gene selection based on modified TopQ method described by Lehne et al., 2011,
  - 580K highly significant SNPs within 13.4K genes.
- two additional contrasts: ovary vs. pituitary tissue and reproductive vs. non-reproductive females,

Methods and results, continue



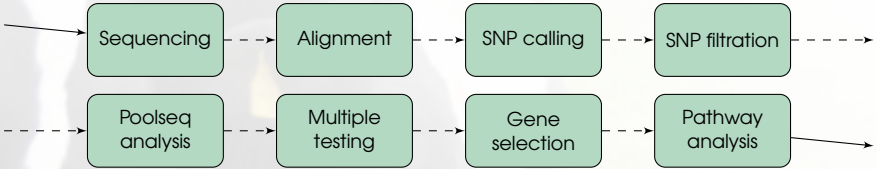
- contrast between select and control lines,
- 43k SNPs segregating with MAF > 0.05, LHI and LHGV statistics,
- 8 pooled samples, mean coverage 50x per pool,
- Poolseq analysis based on Cochran–Mantel–Haenszel statistics (Kofler et al., 2011),
  - after filtration 6.4M of SNPs,
- gene selection based on modified TopQ method described by Lehne et al., 2011,
  - 580K highly significant SNPs within 13.4K genes.
- two additional contrasts: ovary vs. pituitary tissue and reproductive vs. non-reproductive females,
- differential expression analysis.

Cattle study

DALHOUSIE UNIVERSITY
Inspiring Minds

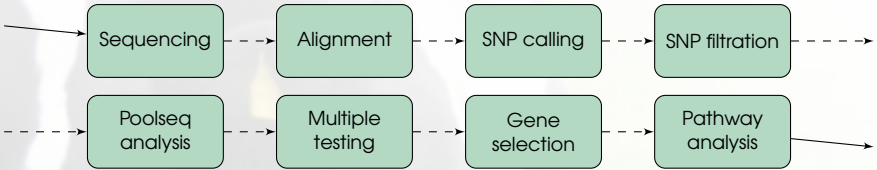
- Poolseq supported by individual sequencing study

Methods, Poolseq



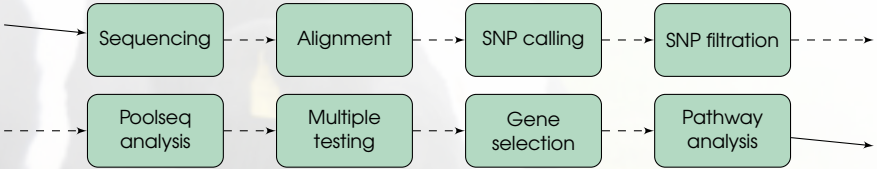
- 16 pools with high and 16 pools with low longevity bulls,

Methods, Poolseq



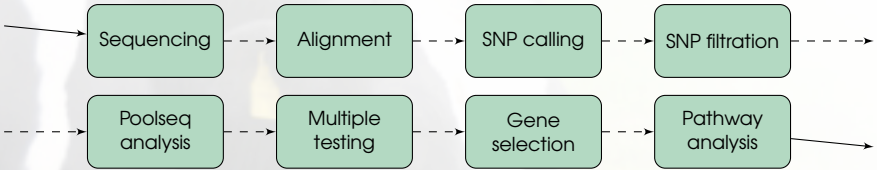
- 16 pools with high and 16 pools with low longevity bulls,
  - mean coverage 12.5x per pool,

Methods, Poolseq



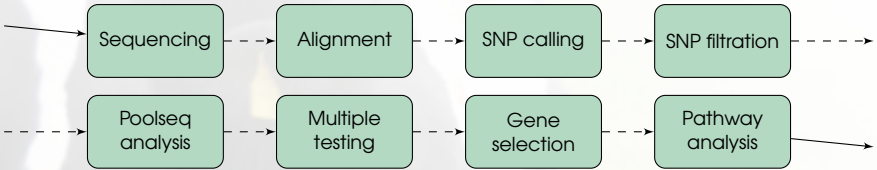
- 16 pools with high and 16 pools with low longevity bulls,
  - mean coverage 12.5x per pool,
  - bulls per pool ranged between 17 and 18,

Methods, Poolseq



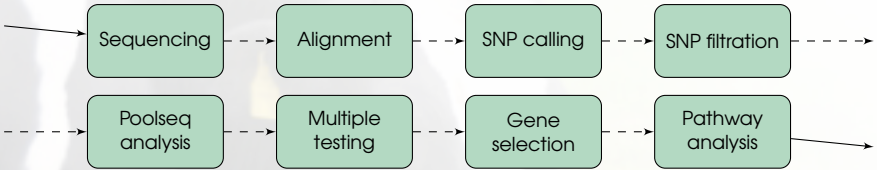
- 16 pools with high and 16 pools with low longevity bulls,
  - mean coverage 12.5x per pool,
  - bulls per pool ranged between 17 and 18,
- Poolseq analysis based on Cochran–Mantel–Haenszel statistics (Kofler et al., 2011),

Methods, Poolseq



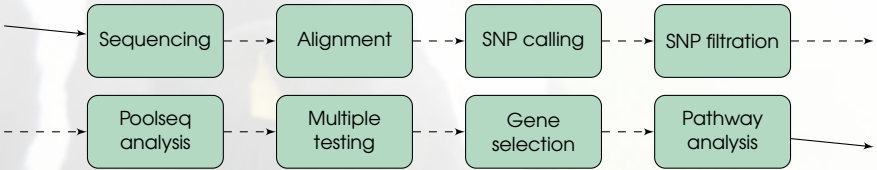
- 16 pools with high and 16 pools with low longevity bulls,
  - mean coverage 12.5x per pool,
  - bulls per pool ranged between 17 and 18,
- Poolseq analysis based on Cochran–Mantel–Haenszel statistics (Kofler et al., 2011),
  - after filtration 10M of SNPs,

Methods, Poolseq



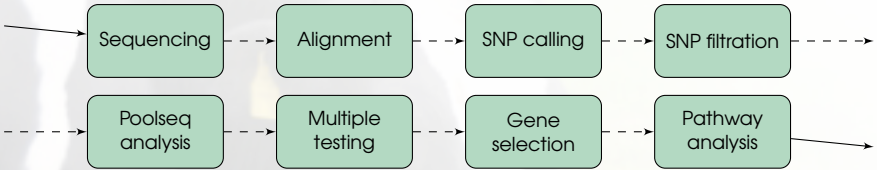
- 16 pools with high and 16 pools with low longevity bulls,
  - mean coverage 12.5x per pool,
  - bulls per pool ranged between 17 and 18,
- Poolseq analysis based on Cochran–Mantel–Haenszel statistics (Kofler et al., 2011),
  - after filtration 10M of SNPs,
  - more than 230K new SNPs were identified,

Methods, Poolseq



- 16 pools with high and 16 pools with low longevity bulls,
  - mean coverage 12.5x per pool,
  - bulls per pool ranged between 17 and 18,
- Poolseq analysis based on Cochran–Mantel–Haenszel statistics (Kofler et al., 2011),
  - after filtration 10M of SNPs,
  - more than 230K new SNPs were identified,
- gene selection based on modified TopQ method described by Lehne et al., 2011,

Methods, Poolseq



- 16 pools with high and 16 pools with low longevity bulls,
  - mean coverage 12.5x per pool,
  - bulls per pool ranged between 17 and 18,
- Poolseq analysis based on Cochran–Mantel–Haenszel statistics (Kofler et al., 2011),
  - after filtration 10M of SNPs,
  - more than 230K new SNPs were identified,
- gene selection based on modified TopQ method described by Lehne et al., 2011,
  - 200K highly significant SNPs within 8.4K genes.

Results, Poolseq

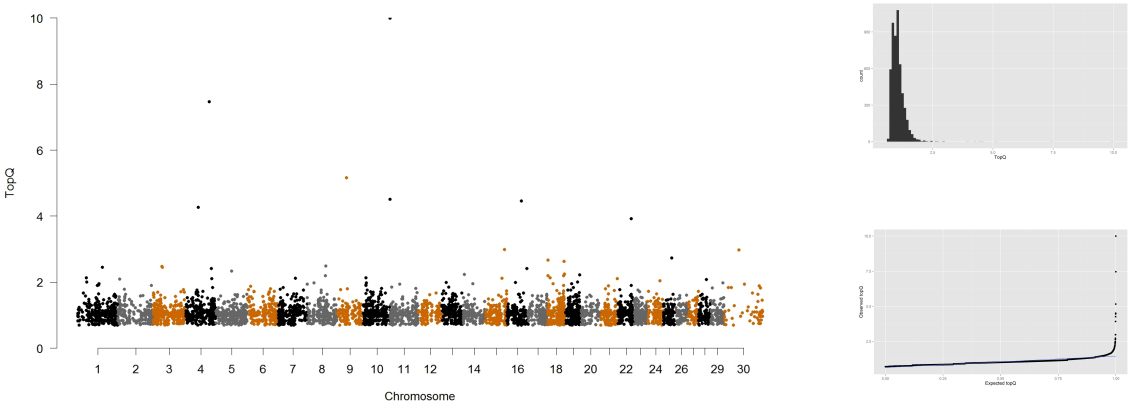


Fig. Manhattan plot for TopQ (left) with histogram (right-top) and qq (right-bottom) plots for all coding genes.

Results, Meta-analysis

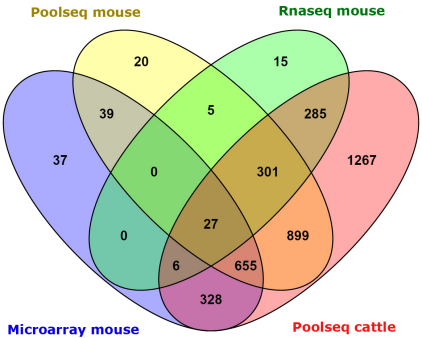


Fig. Venn plot representing significant genes across considered studies based on cattle annotated genes.

- homology issue,
  - 45K for mouse and 24K cattle genes in annotation table,
- more than 3300 significant genes across studies based on cattle annotation table,
- low correlation between studies,
- weighted index used to rank genes,
- more than 1400 significant genes which were useful for cattle,
  - with more than one significant SNP within gene.

SNP panel selection

- SNP selection algorithm based on significance level and linkage disequilibrium,
- testing panel with more than 1400 genes and QTLs (Zhi-Liang et al., 2013),
- testing panel with more than 4500 SNPs,
- SNP classes: within gene, within QTL, intergenomic SNPs.

Conclusions and further plans

DALHOUSIE UNIVERSITY
Inspiring Minds

Conclusions:

- multidisciplinary approach,
- testing panel in production, Illumina approach.

Further plans:

- many levels of testing: individual SNP, gene, genomic selection approach.
- validation procedure:
  - more than 3000 bulls,
  - stand alone panel,
  - combination with Illumina Bovine50K.

Acknowledgements

Jyoti Joshi, Nehil Jain, Robert Beiko,
Hein van der Steen and people from Lab...

Erik Mullaart and Chris Schrooten
Barry Simpson and Stewart Bauck

NRC-IRAP
ACOA
Mitacs Accelerate

Thank you for your attention

- Questions?

Are you interested in panel utilization?
Hein van der Steen, hein.vandersteen@btinternet.com

genewise R package

DALHOUSIE UNIVERSITY
Inspiring Minds

- high performance R package dedicated for GWAS and Poolseq data,
- utility functions, SNP selection based on multiple testing, mapping SNPs to genes, genewise statistics for gene ranking, annotation,
- SNP selection algorithm to produce Illumina microarray design,
- summary statistics for different steps to give control to the user,
- package is built using `data.table`, `dplyr family`, `qvalue`, `multtest`, `ggplot2`.

https://github.com/nehiljain/genewiseR

Mouse study

DALHOUSIE UNIVERSITY
Inspiring Minds

- Material

Mouse study

DALHOUSIE UNIVERSITY
Inspiring Minds

- Material
- Microarray

Mouse study

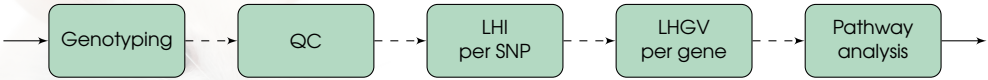
- Material
- Microarray
- Poolseq

Mouse study

- Material

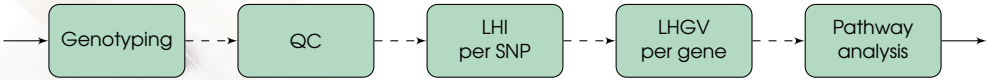
- Microarray
- Poolseq
- Rnaseq

Methods, Microarray



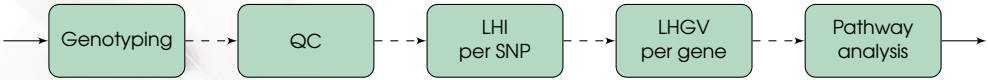
- select (S) and control (C) lines with more than 350 samples,

Methods, Microarray



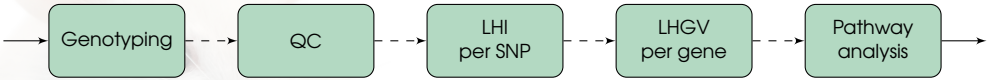
- select (S) and control (C) lines with more than 350 samples,
- 43k SNPs segregating with MAF > 0.05,

Methods, Microarray



- select (S) and control (C) lines with more than 350 samples,
- 43k SNPs segregating with MAF > 0.05,
- Use S-C contrast in allele frequency, association with RL in F2 resource population, LD and location to calculate LHI per SNP,

Methods, Microarray



- select (S) and control (C) lines with more than 350 samples,
- 43k SNPs segregating with MAF > 0.05,
- Use S-C contrast in allele frequency, association with RL in F2 resource population, LD and location to calculate LHI per SNP,
- Used LHI results to calculate LHGV per gene.

Results, Microarray

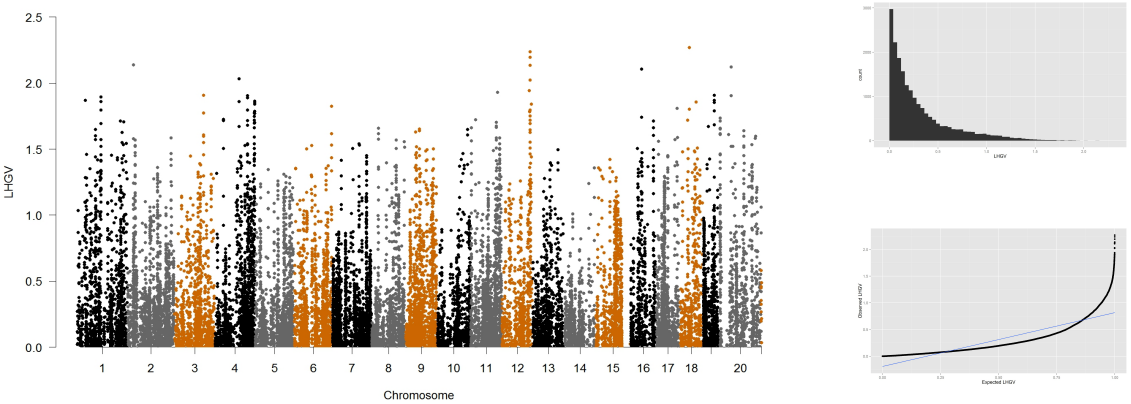
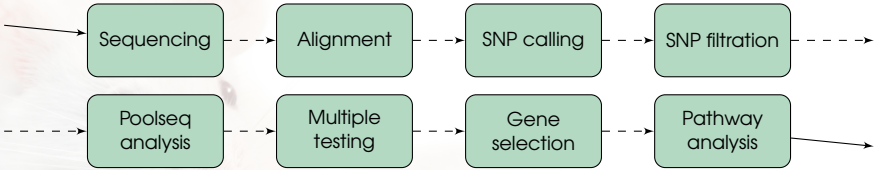


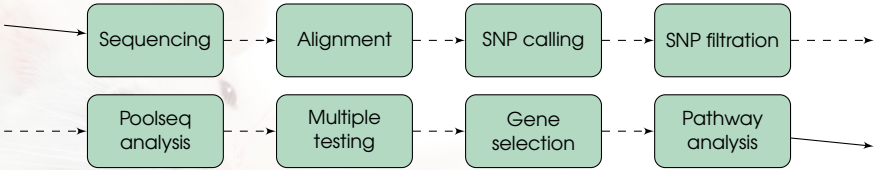
Fig. Manhattan plot for LHGV (left) with histogram (right-top) and qq (right-bottom) plots for protein coding genes.

Methods, Poolseq



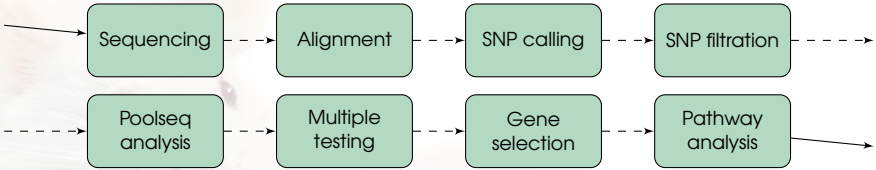
- 4 select and 4 control pools with more than 20 individuals per pool,

Methods, Poolseq



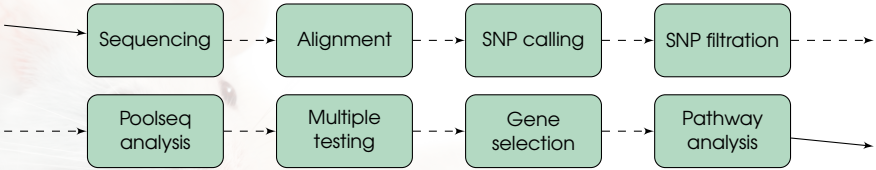
- 4 select and 4 control pools with more than 20 individuals per pool,
  - mean coverage 50x per pool,

Methods, Poolseq



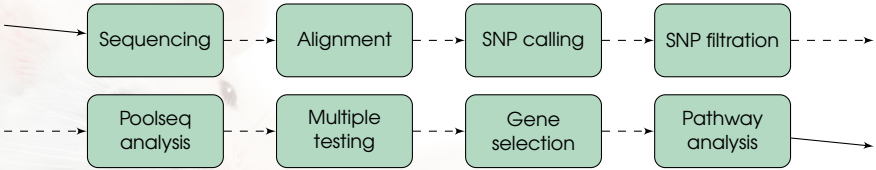
- 4 select and 4 control pools with more than 20 individuals per pool,
  - mean coverage 50x per pool,
- Poolseq analysis based on Cochran–Mantel–Haenszel statistics (Kofler et al., 2011),

Methods, Poolseq



- 4 select and 4 control pools with more than 20 individuals per pool,
  - mean coverage 50x per pool,
- Poolseq analysis based on Cochran–Mantel–Haenszel statistics (Kofler et al., 2011),
  - after filtration 6.4M of SNPs,
- gene selection based on modified TopQ method described by Lehne et al., 2011,

Methods, Poolseq



- 4 select and 4 control pools with more than 20 individuals per pool,
  - mean coverage 50x per pool,
- Poolseq analysis based on Cochran–Mantel–Haenszel statistics (Kofler et al., 2011),
  - after filtration 6.4M of SNPs,
- gene selection based on modified TopQ method described by Lehne et al., 2011,
  - 580K highly significant SNPs within 13.4K genes.

Results, Poolseq

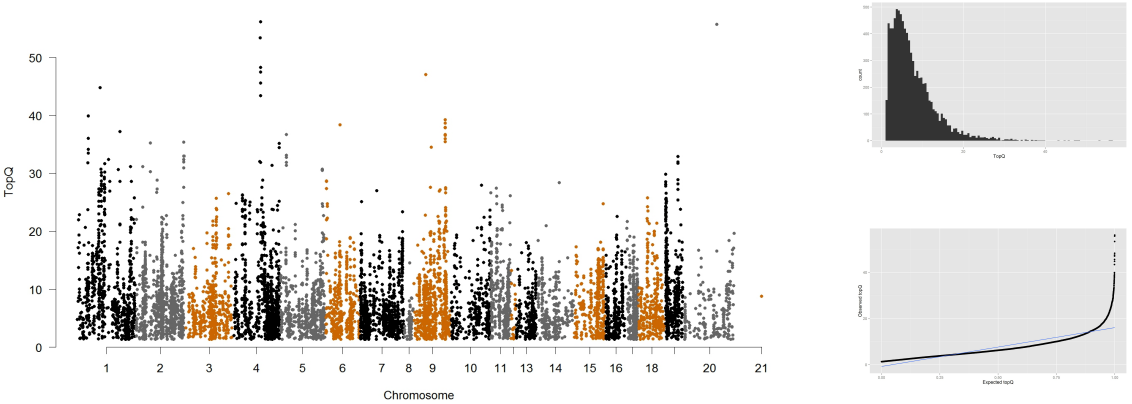
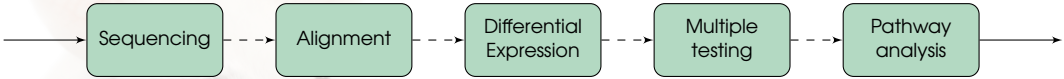


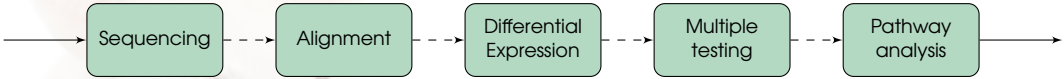
Fig. Manhattan plot for TopQ (left) with histogram (right-top) and qq (right-bottom) plots for protein coding genes.

Methods, Rnaseq



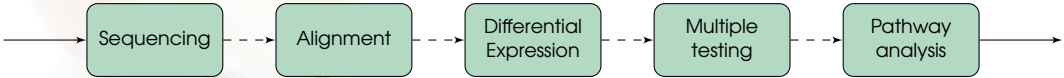
- 20 pools with more than 20 individuals per pool,

Methods, Rnaseq



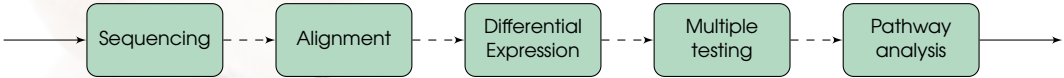
- 20 pools with more than 20 individuals per pool,
  - more than 163M reads per pool (MAPQ30),

Methods, Rnaseq



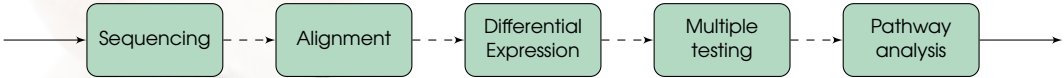
- 20 pools with more than 20 individuals per pool,
  - more than 163M reads per pool (MAPQ30),
- 3 different contrasts:

Methods, Rnaseq



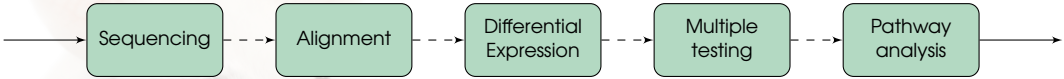
- 20 pools with more than 20 individuals per pool,
  - more than 163M reads per pool (MAPQ30),
- 3 different contrasts:
  - select and control,

Methods, Rnaseq



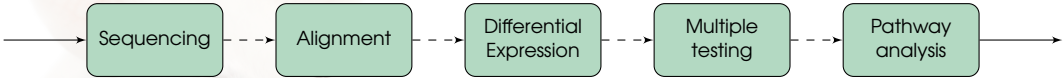
- 20 pools with more than 20 individuals per pool,
  - more than 163M reads per pool (MAPQ30),
- 3 different contrasts:
  - select and control,
  - ovary and pituitary tissue,

Methods, Rnaseq



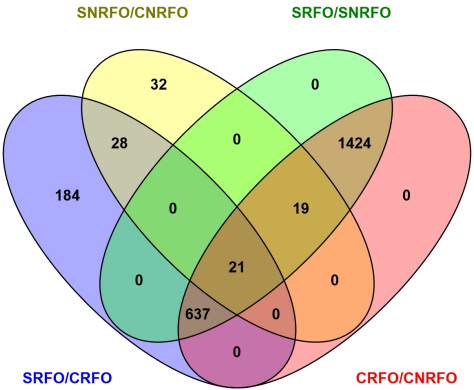
- 20 pools with more than 20 individuals per pool,
  - more than 163M reads per pool (MAPQ30),
- 3 different contrasts:
  - select and control,
  - ovary and pituitary tissue,
  - reproductive and non-reproductive females,

Methods, Rnaseq



- 20 pools with more than 20 individuals per pool,
  - more than 163M reads per pool (MAPQ30),
- 3 different contrasts:
  - select and control,
  - ovary and pituitary tissue,
  - reproductive and non-reproductive females,
- differential expression analysis with utilization DESeq and edgeR Bioconductor R packages.

Results, Rnasea



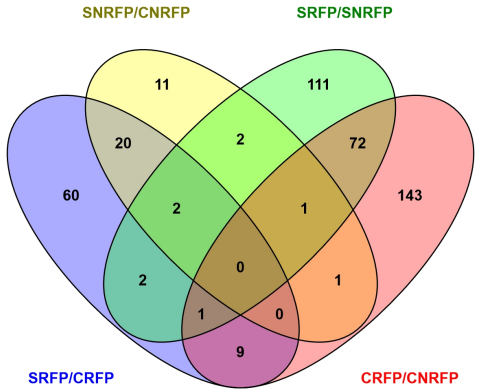


Fig. Venn plots for highly significant genes representing different contrasts for ovary (left) pituitary (right) tissue.

Results, Rnaseq

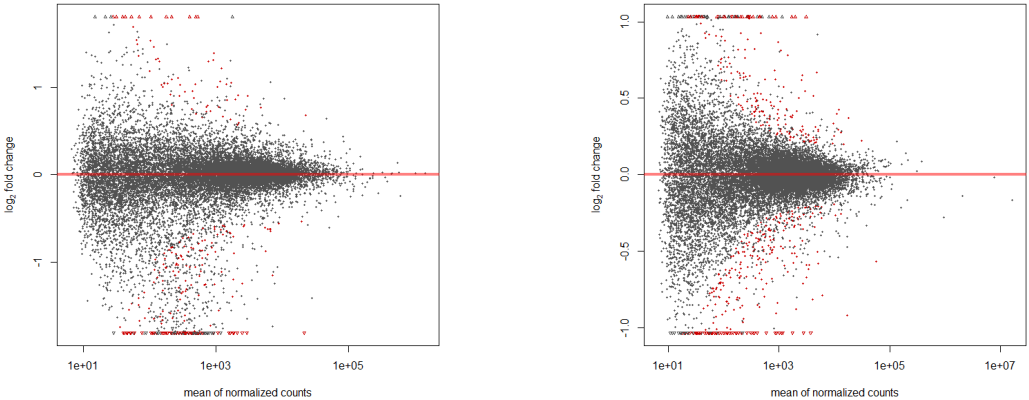
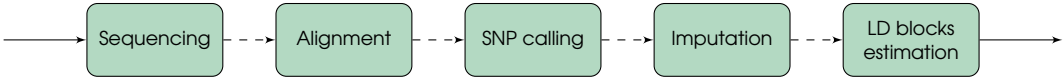


Fig. Plot of normalized mean versus $\log_2$ fold change for the contrast S vs. C for ovary (left) pituitary (right) tissue.

Methods, WGS



- individual WGS,

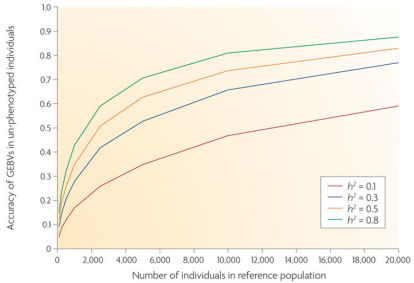
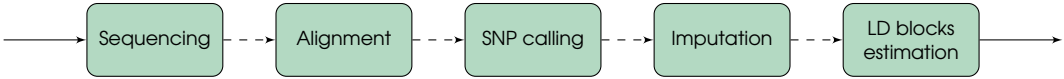


Fig. Accuracy of GEBVs of un-phenotyped individuals (Goddard and Hayes, 2009).

Methods, WGS



- individual WGS,
- 126 bulls with high and low EBVs for longevity,

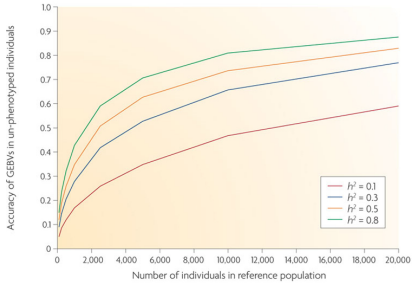
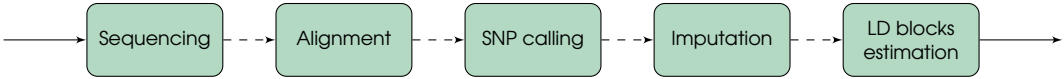


Fig. Accuracy of GEBVs of un-phenotyped individuals (Goddard and Hayes, 2009).

Methods, WGS



- individual WGS,
- 126 bulls with high and low EBVs for longevity,
- mean coverage 1.5x per individual bull,

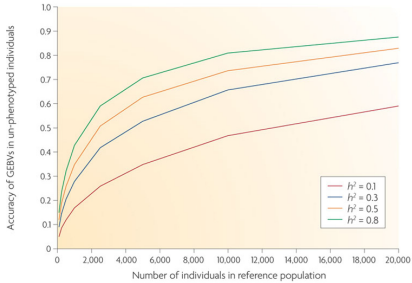
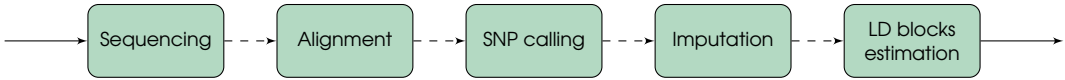


Fig. Accuracy of GEBVs of un-phenotyped individuals (Goddard and Hayes, 2009).

Methods, WGS



- individual WGS,
- 126 bulls with high and low EBVs for longevity,
- mean coverage 1.5x per individual bull,
- to estimate linkage disequilibrium between SNPs and estimate LD blocks,

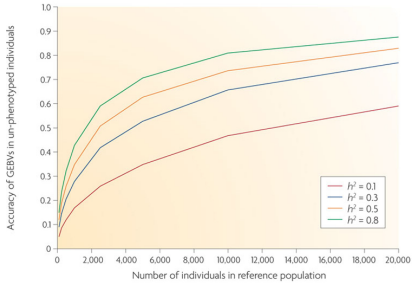
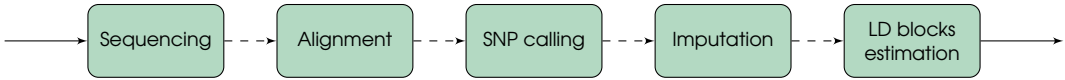


Fig. Accuracy of GEBVs of un-phenotyped individuals (Goddard and Hayes, 2009).

Methods, WGS



- individual WGS,
- 126 bulls with high and low EBVs for longevity,
- mean coverage 1.5x per individual bull,
- to estimate linkage disequilibrium between SNPs and estimate LD blocks,
- the lack of power...

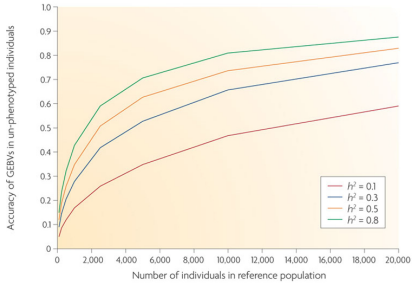


Fig. Accuracy of GEBVs of un-phenotyped individuals (Goddard and Hayes, 2009).

Results, Pathway analysis

Fig. Overall representation of Gene Ontology terms in more than 1400 cattle genes.

- selection for RL is selection for Fertility, Longevity, Ability to combine reproduction and lactation till late in life, Health and Robustness

Results, Pathway analysis

Fig. Overall representation of Gene Ontology terms in more than 1400 cattle genes.

- selection for RL is selection for Fertility, Longevity, Ability to combine reproduction and lactation till late in life, Health and Robustness
- results of pathway analysis support the above