Gene mapping in cattle: Lessons learnt from genome-wide variants

Goutam Sahana
Aarhus University, Denmark

Connecting genetic variants to complex phenotypes

1. Identify statistical connections between points (or areas) in the genome and the phenotype
   - Drive hypotheses for biological studies of specific genes/regions in specific context
2. Generate insights on genetic architecture of phenotype
   - No. of loci, effect sizes, MAF, dispersed across the genome etc.
3. Build statistical models to predict phenotype from genotype
   - “Show me your genome and I will tell you what diseases you will get”

Identifying genetic factors: different approaches

1. **Linkage analysis** - largely (if not entirely) unsuccessful because this approach is only adequately powered with realistic sample sizes to identify very large genetic effects
2. **Candidate-gene studies** - suffered from a number of methodological limitations (for example, small number of samples and genetic markers tested and have been largely discontinued
3. **Genome-wide association studies (GWAS)**
   - Development of genotyping arrays (affordable cost)
   - Thousands of individuals genotyped for millions of genetic variants became a reality
   - Method development (imputation, population structure)
   - Became a powerful tool to identify genetic associations

A decade of GWAS - revolutionized complex trait genomics

- Almost any (heritable) complex trait that has been studied, many loci contribute to standing genetic variation
- The mutational target in the genome appears large so that polymorphisms in many genes contribute to genetic variation
- The proportion of variance explained by individual variants is small
- The high rates of replication imply that findings can be trusted
- Larger experimental sample sizes will lead to new discoveries
- We need new visions and methodologies to fully tackle questions about the genetic architecture of complex traits
- The success of GWAS has not translated into an ability to predict phenotypes based on identified associated markers

GWAS: methodology and resource development

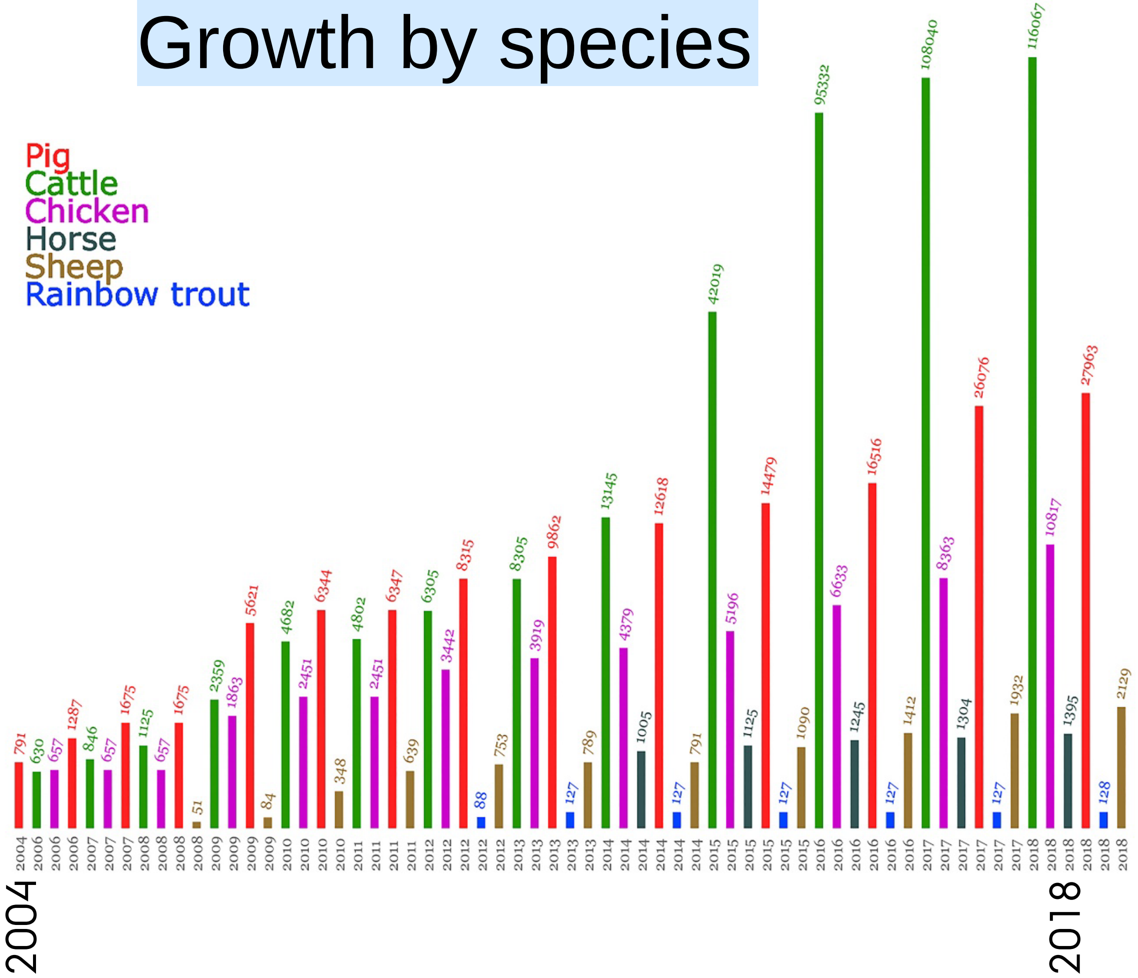
- GWAS data have led to new analysis methods
  - Better modeling population structure and relatedness between individuals in a sample
  - Detecting novel variants on the basis of GWAS summary statistics
  - Estimating and partitioning genetic (co)variance
  - Inferring causality
- GWAS discoveries and interpretation have benefited substantially from improved algorithms in statistical imputation of unobserved genotypes
- Publicly available resources

GWAS and DNA markers

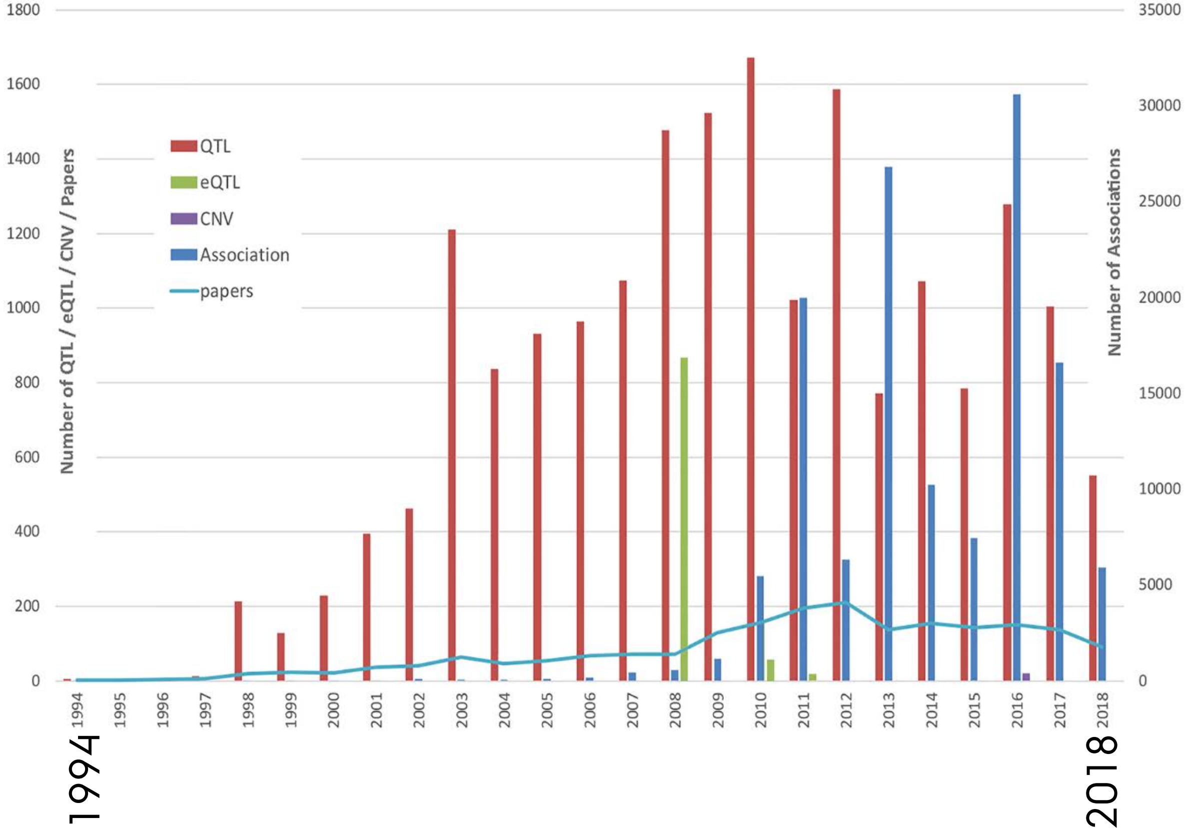
1. **Single nucleotide polymorphism (SNP)**
   I. Common variants (MAF ≥ 5%)
   II. Low-frequency variants (MAF 1-5%)
   III. Rare variants (MAF < 1%)
2. **Indels**: (< 1 kb) are the second most common class of mutation in the genome. They can have far-ranging effects concerning gene expression and genetic disease
3. **Copy number variation (CNV)** are structural variants where the number of copies in the genome varies between individuals

Growth of curated data in the Animal QTLdb

Growth by species



Growth by data types



Hu et al. Nucleic Acid Research 2018

Association mapping with common variants

1. Large number of QTL identified
2. Explained a substantial proportion of additive genetic variance
3. Nearly 2,500 QTL-SNP in the LD-chip
4. QTL-SNP increases accuracy in across breed prediction (Aoxing Liu)
5. Sequence variants at QTL peaks from multi-breed GWAS, increase reliability of predictions (Irene van den Berg)

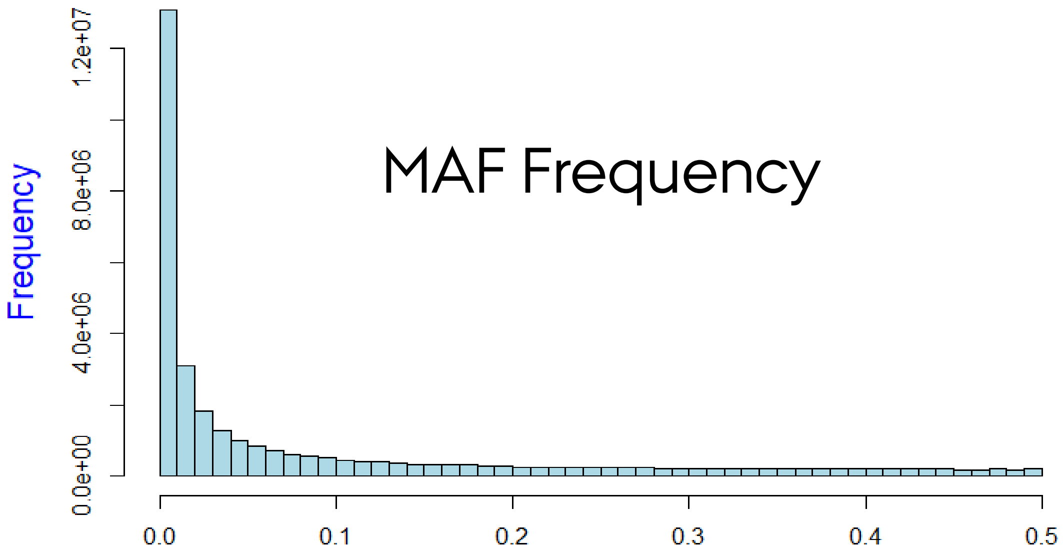
	No. QTL	Variance explained (%)	
		QTLs	Rest of the genome
Fat	23	25.12	60.01
Protein	33	15.34	68.89
Milk	26	21.29	63.97

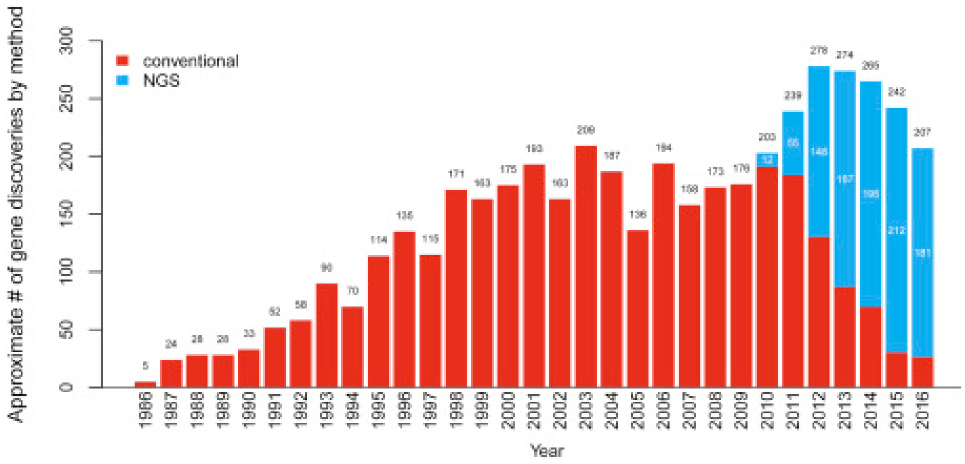
Cai et al. BMC Genetics 2018**19**:30

Rare and low frequency variants

- Large proportion in the genome
- Rare alleles of large effect certainly also make an essential contribution
- Evolutionary and quantitative genetic theory both provide strong expectations for rare variants
- Rare variant can pushes an individual over the disease threshold
- Explain part of the ‘missing heritability’

- Among the gene discoveries in recent years, majority are rare



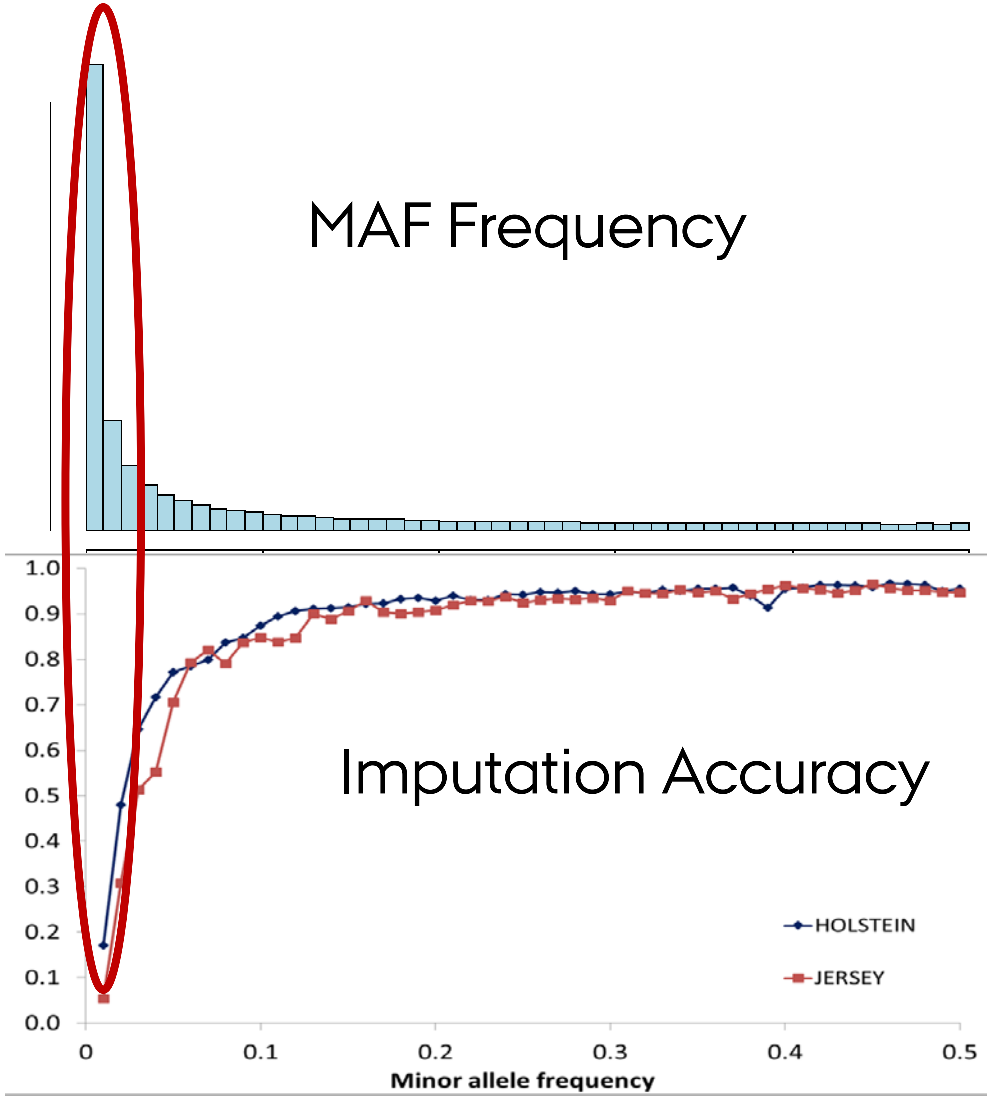


Boycott et al. AJHG 100:695-705 (2017)

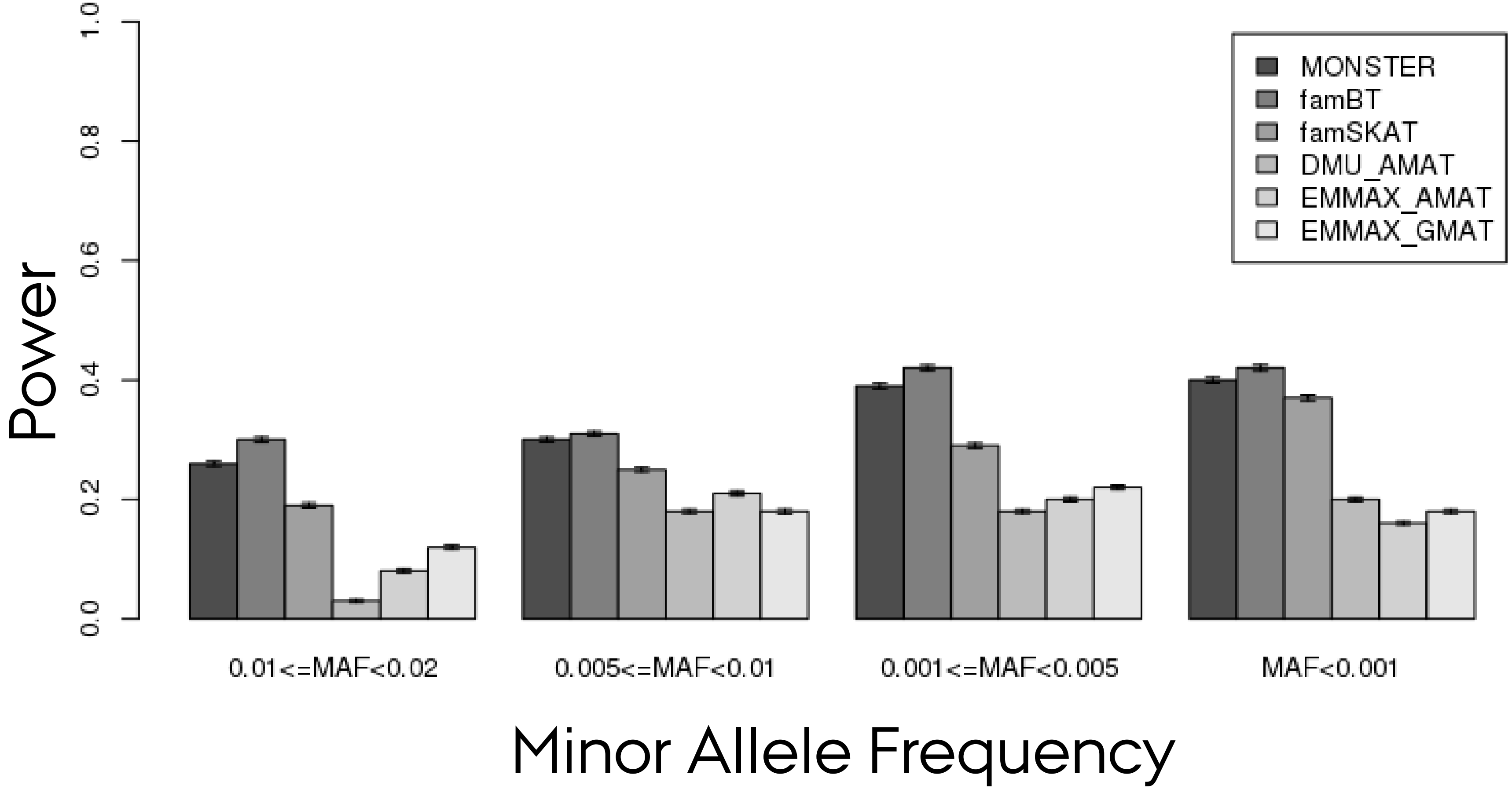
Rare and low frequency variants - limitations

Large proportion in the genome, however,

- largely results in small contribution
  - too rare to contribute to the population variance
  - effect sizes very small
- Poor imputation accuracy



Low power to detect rare variants



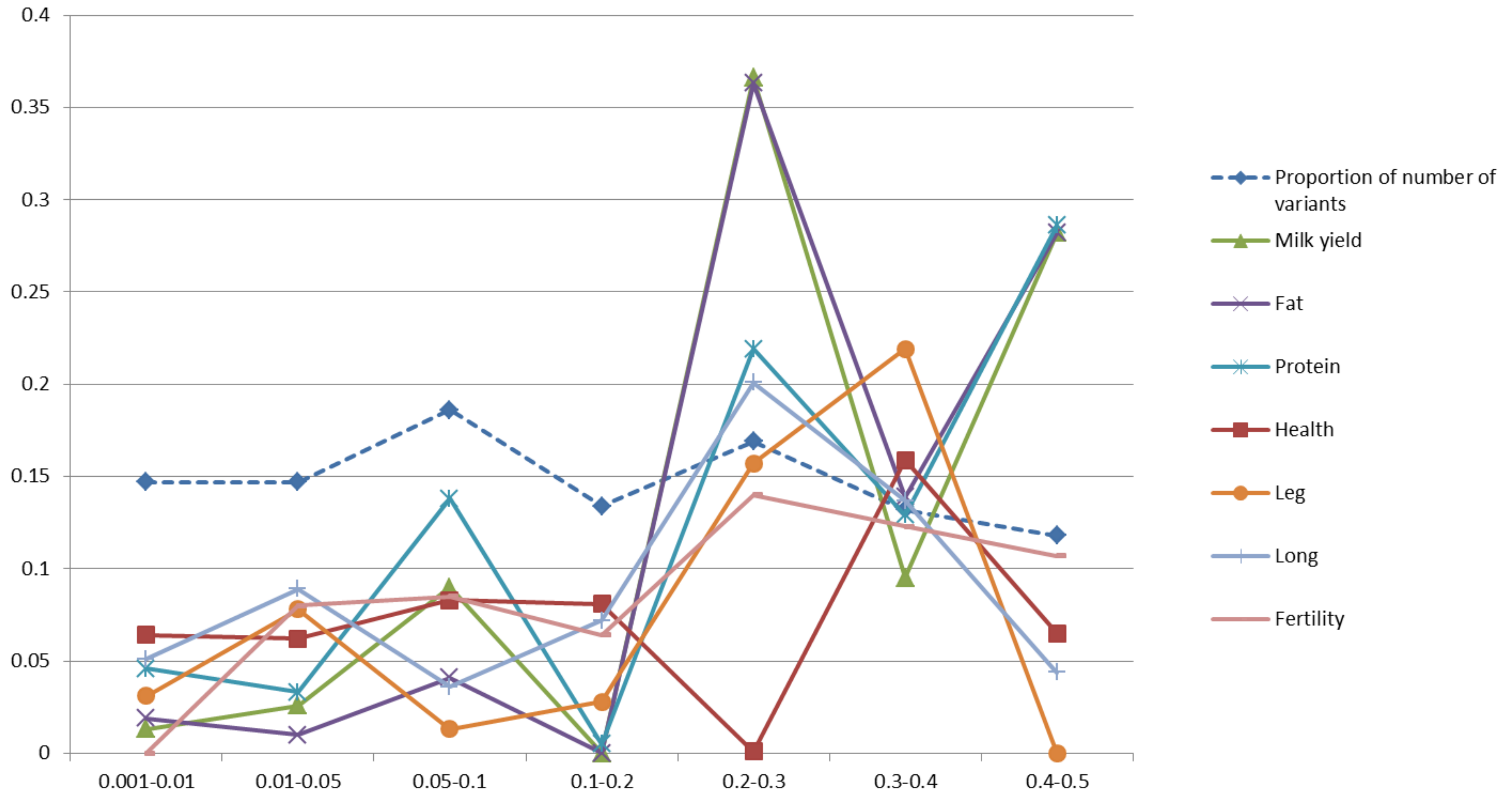
h^2=0.5

h^2_{QTL}=0.01

N=1000

Zhang et al. Genetics Selection Evolution 2016 **48**:60

Relative contribution different MAF-class variants to DRP variance



Zhang et al. Genetics Selection Evolution 2017**49**:60

Rare and low frequency variants: lessons learnt

1. Extremely low power to detect rare variants
2. Method specialized for rare variant mapping performed better compare to commonly applied models for GWAS
3. They explain larger proportion of variance for fitness traits than for production traits
4. No additional improvement in prediction accuracy by including them
5. However, if ‘known’, improves prediction accuracy

Are we looking at ‘wrong’ phenotypes ?

Structural variants: DNA alternations

1. CNV affecting protein-coding genes contributes substantially to phenotype diversity and disease

2. One human on an average has:
    1. 0.81 deleted gene
    2. 1.75 duplicated gene
    3. 70% ≥1 genic CNV
3. Deletions are potential candidate for loss-of-function
4. Least explored polymorphisms in cattle

Ruderfer et al. Nature Genetics 2016 48:1107–1111

Enrichment of deletions on QTL

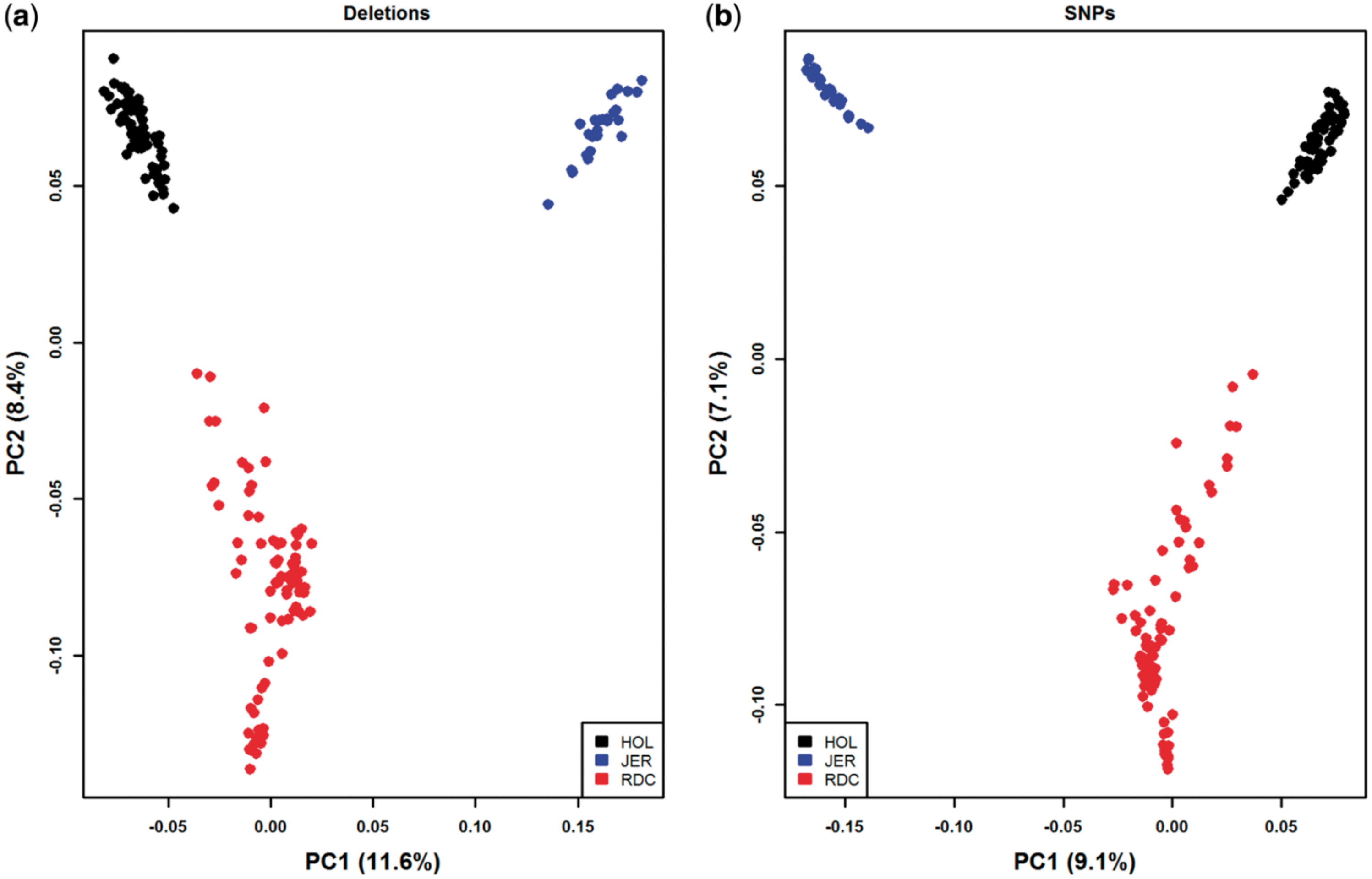
- ❑ 8,480 large deletions (199bp to 773KB)
  - ❑ 82% of which are novel compared with deletions in the dbVar database

Trait Classes[¥]	Fold Enrichment	P value[*]
Health	2	8.91×10^{-10}
Reproduction	1.5	7.4×10^{-11}
Milk	0.8	2.45×10^{-7}
Exterior	0.5	1.85×10^{-4}
Production	0.5	0.002
Meat and Carcass	0.5	0.058

[¥] Trait classes are from cattleQTLdb.

Mesbah-Uddin et al. DNA Research 2018 25:49-59

Large deletions in three cattle breeds

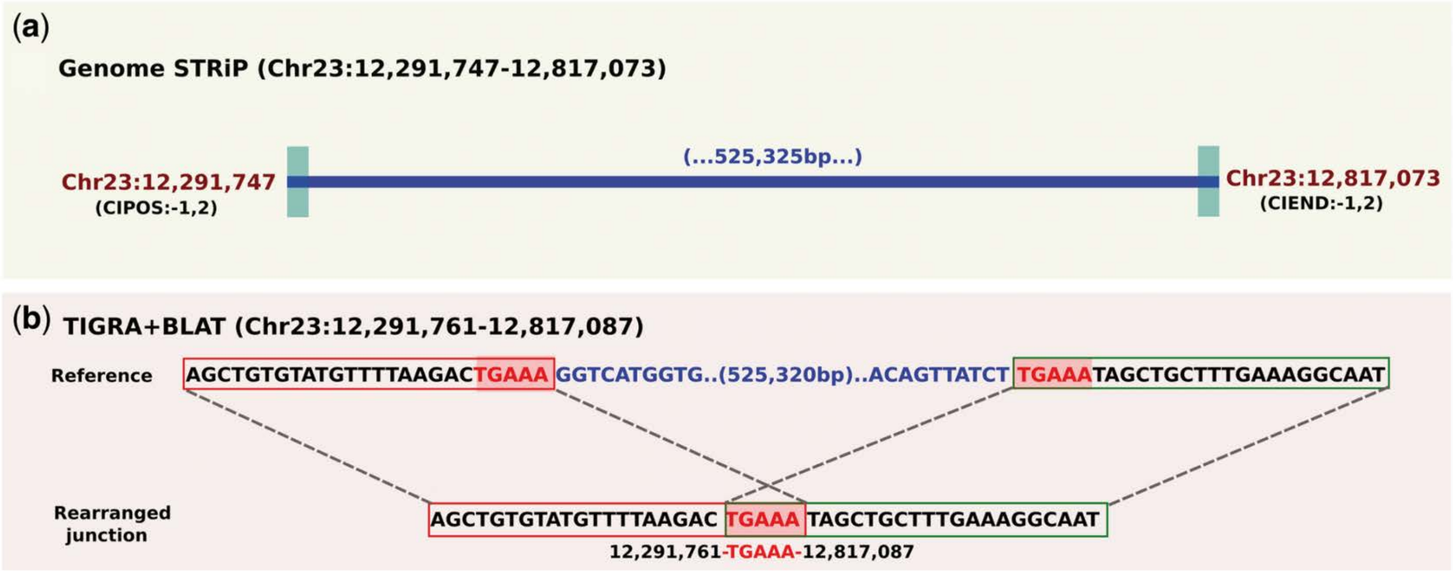


Deletions

SNPs

Mesbah-Uddin et al. DNA Research 2018 25:49-59

Large deletions can lead to causality

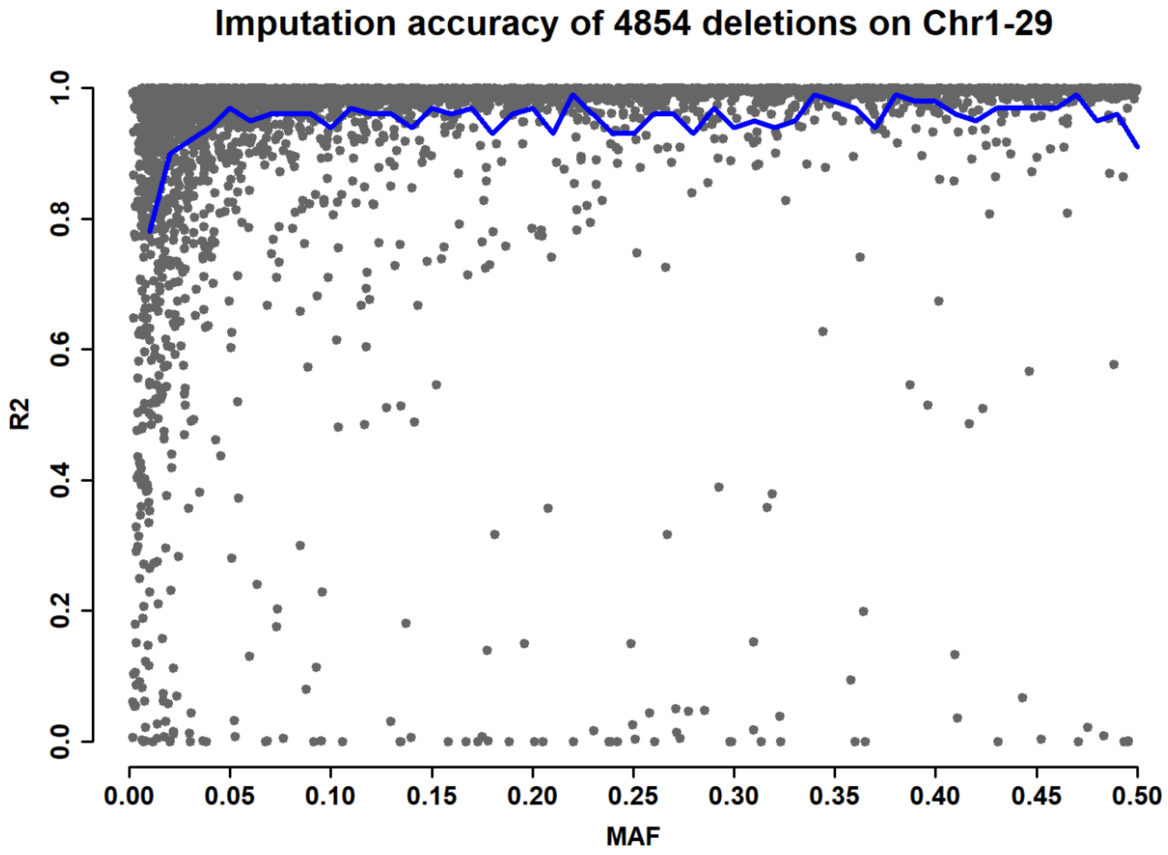


(a)
Genome STRiP (Chr23:12,291,747-12,817,073)
Chr23:12,291,747
(CIPOS:-1,2)
(...525,325bp...)
Chr23:12,817,073
(CIEND:-1,2)
(b) TIGRA+BLAT (Chr23:12,291,761-12,817,087)
Reference
AGCTGTGTATGTTTTAAGACTGAAA GGTCATGGTG..(525,320bp)..ACAGTTATCT TGAAATAGCTGCTTTGAAAGGCAAT
Rearranged junction
AGCTGTGTATGTTTTAAGAC TGAAA TAGCTGCTTTGAAAGGCAAT
12,291,761-TGAAA-12,817,087

A ~525-KB deletion on chromosome 23

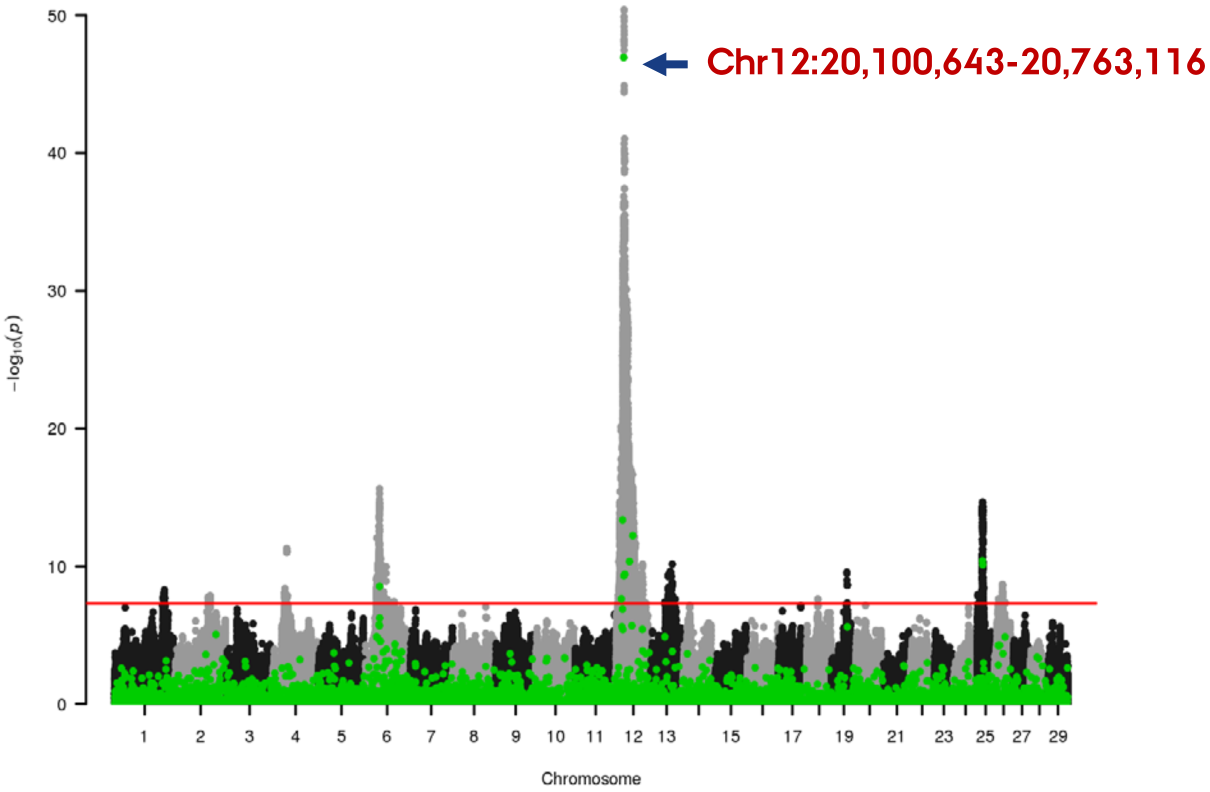
Mesbah-Uddin et al. DNA Research 2018 25:49-59

Large deletions: Imputation and mapping



Imputation accuracy of 4854 deletions on Chr1-29

12970 animals: 6375 Holstein + 4955 Nordic Red cattle + 1640 Jersey



Chr12:20,100,643-20,763,116

Manhattan plot with SNP, indel, and large deletions for fertility index in Nordic Red cattle.

Mesbah-Uddin (unpublished)

Large deletions – Genomic prediction

Method		Proportion of variance explained (V_g/V_P)	Prediction accuracy * (Pearson's correlation)
GCTA[1] – GREML	GRM_DEL	0.467	0.537
	GRM_50K	0.707	**0.628**
	GRM_DEL + 50K	**0.710**	0.627
	GRM_DEL & GRM_50K	0.709	0.627
BayesR[2]	DEL	0.467	0.550
	50k	**0.701**	0.626
	DEL + 50K	0.699	**0.630**

* Random split: 80% training & 20% testing

Mesbah-Uddin (unpublished)

Large deletion study – summary

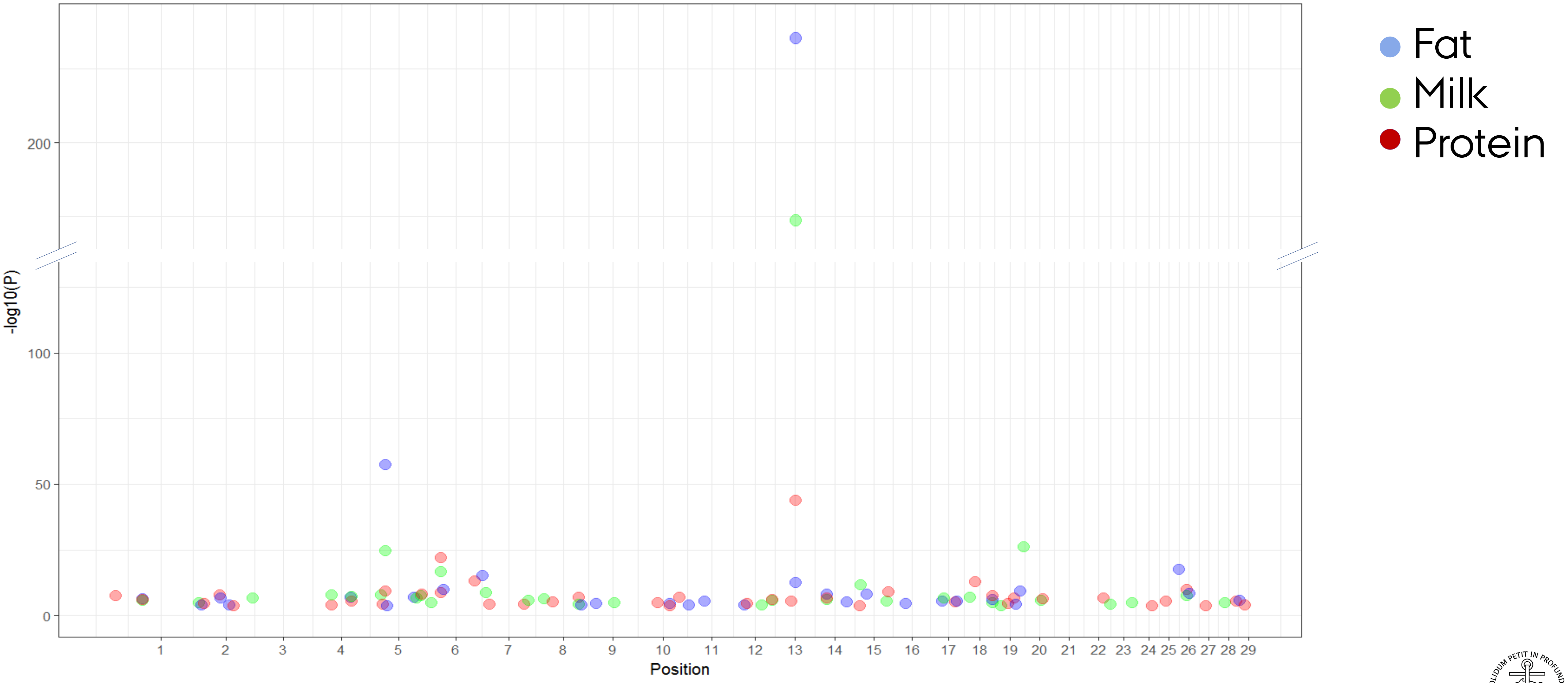
1. Genotype of deletion loci could be inferred from auxiliary read-depth data
2. A high-resolution genetic map of large deletions is provided.
3. Common deletions could be imputed with high accuracy
4. Enrichment of deletions on QTL for health and fertility
5. Causal variant identification (e.g. ~525 KB deletion causing stillbirth in cattle)
   - Managing recessive lethals in a population
6. Potential for inclusion in genomic studies
   - could explain (additional) phenotypic variance
   - could improve prediction accuracy

Pleiotropy is highly prevalent

Complex traits are associated with hundreds to thousands of loci strongly suggests that some of the underlying causal variants are the same.

- Genetic correlations estimates imply that a number of the same variants affect two or more traits in a consistent direction
- The same genetic variants can be significantly associated with multiple diseases and traits in GWAS
- Analytical methods that estimate genetic correlations from GWAS data have provided evidence for widespread pleiotropy
- The true nature of the pleiotropy is currently unknown but, in some cases, could imply an impact of the variants on different tissues, metabolic pathways and/or at different stages

QTLs for Milk, fat and protein in Nordic Holstein

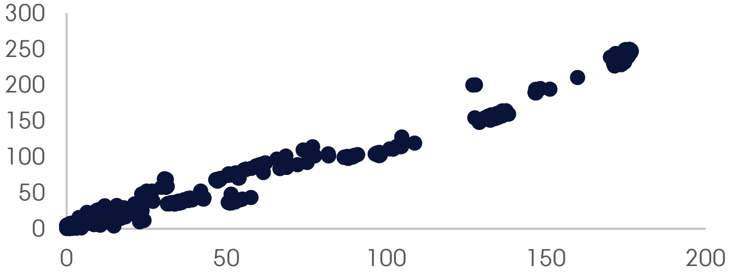


BIG QTL segregating – balancing selection?

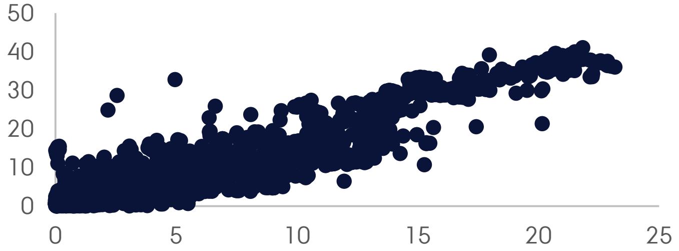
SNP	Gene	Milk	Protein	Fat
Chr5:93945991	MGST1	-2.30	-1.16	+3.07
Chr14:1802266	DGAT1	-5.86	-3.06	+7.15

SNP	Gene	Milk	Mastitis resistance
Chr6:88840407	Intergenic (NPFFR2 & GC)	-1.98	+3.17

BTA14: Milk vs. Fat QTL

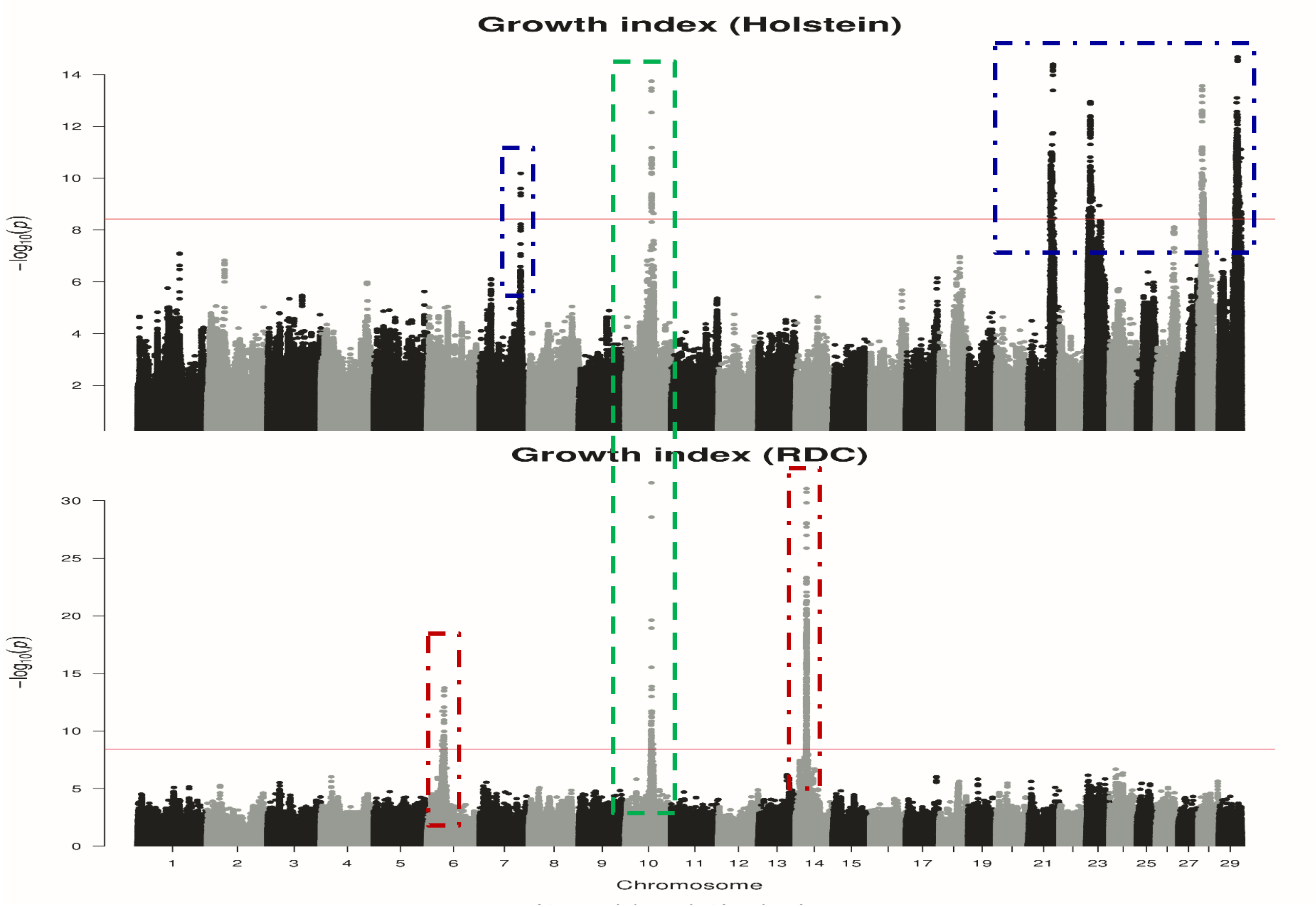


BTA6: Milk vs Mastitis



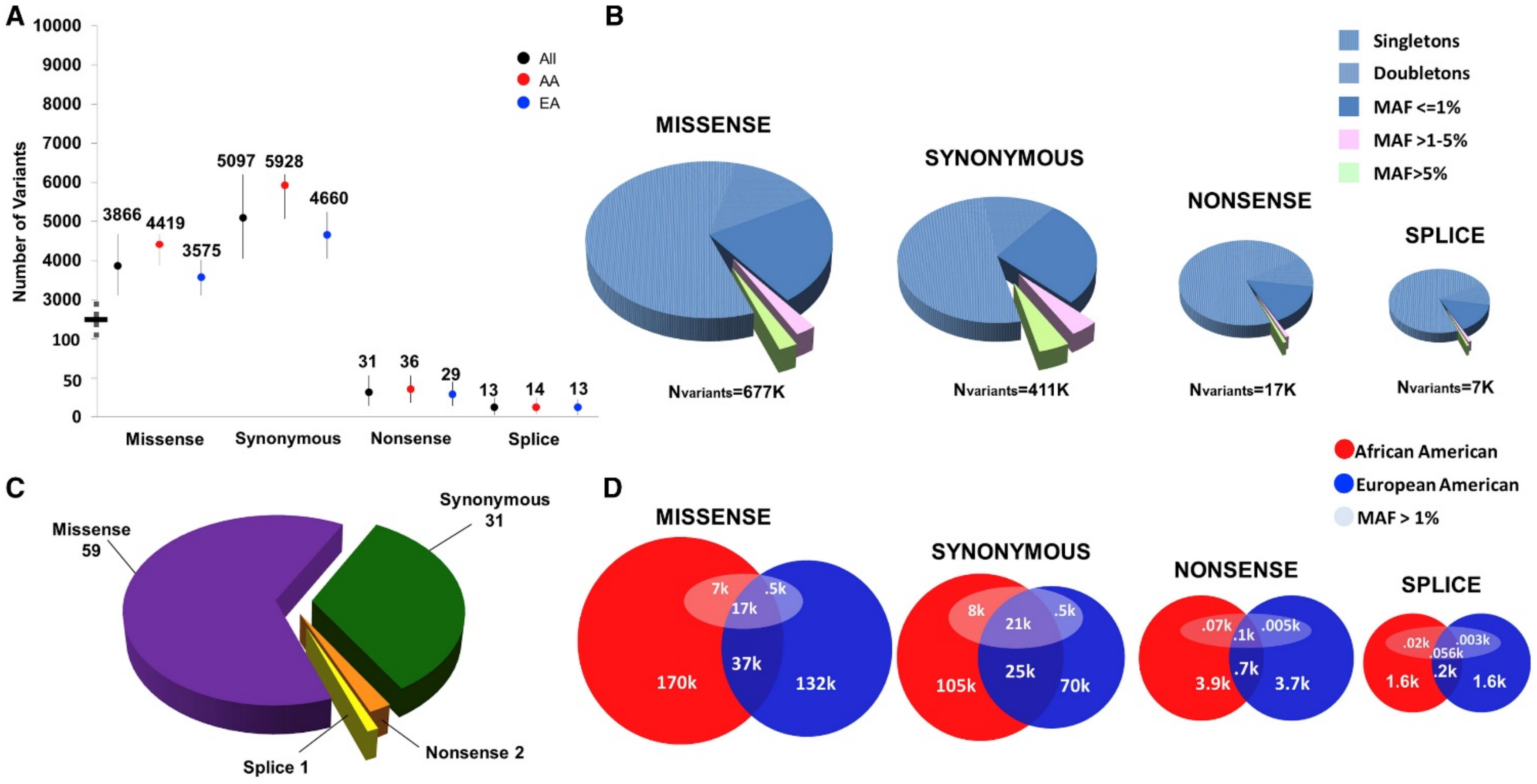
Dusza et al. (unpublished)

Are QTLs population specific ?



Mao et al. J Anim Sci 94:1426-1437 (2016)

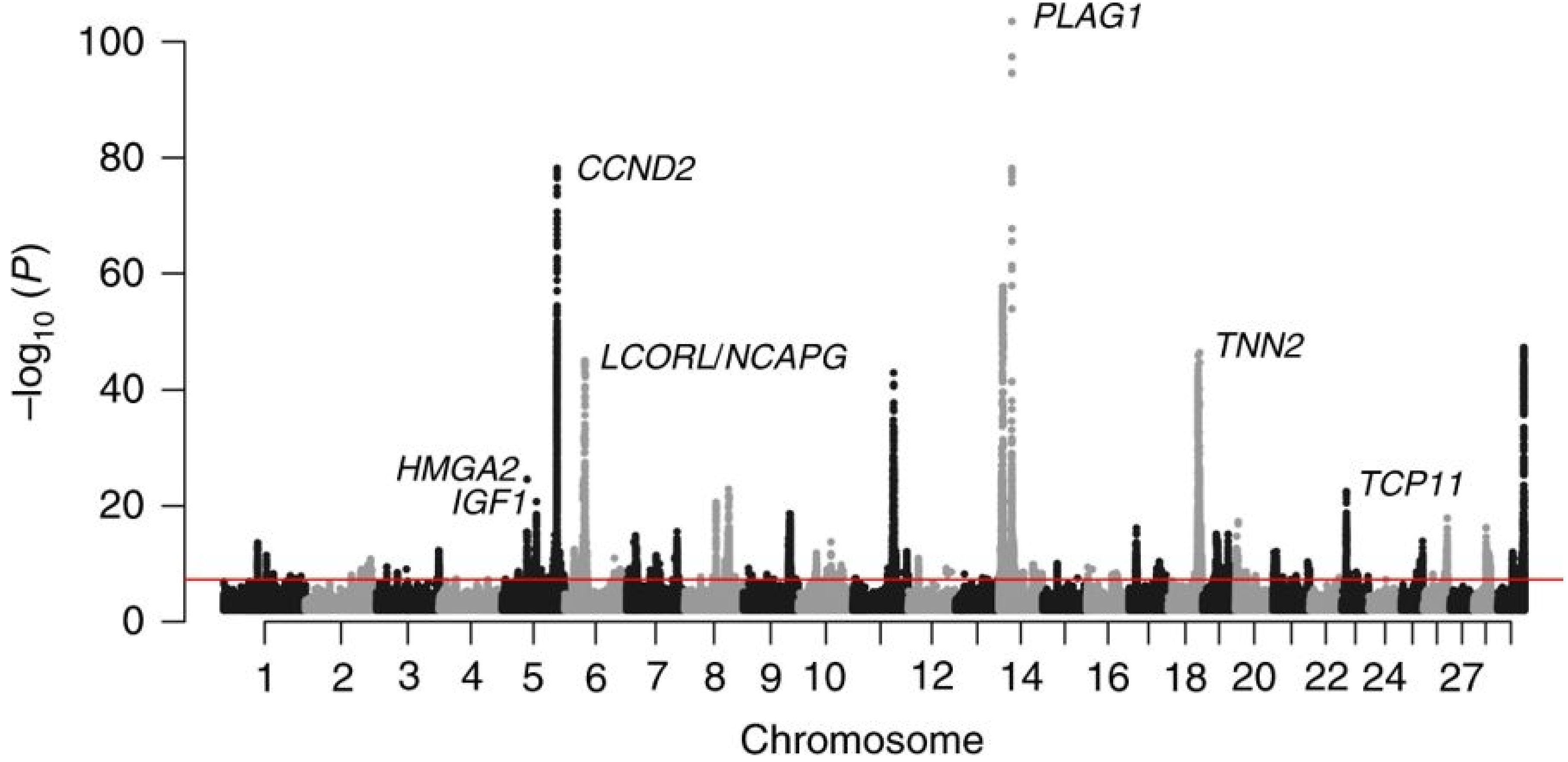
High impact variants – largly population specific



Auer et al. The American Journal of Human Genetics 2016 **99**:791–801

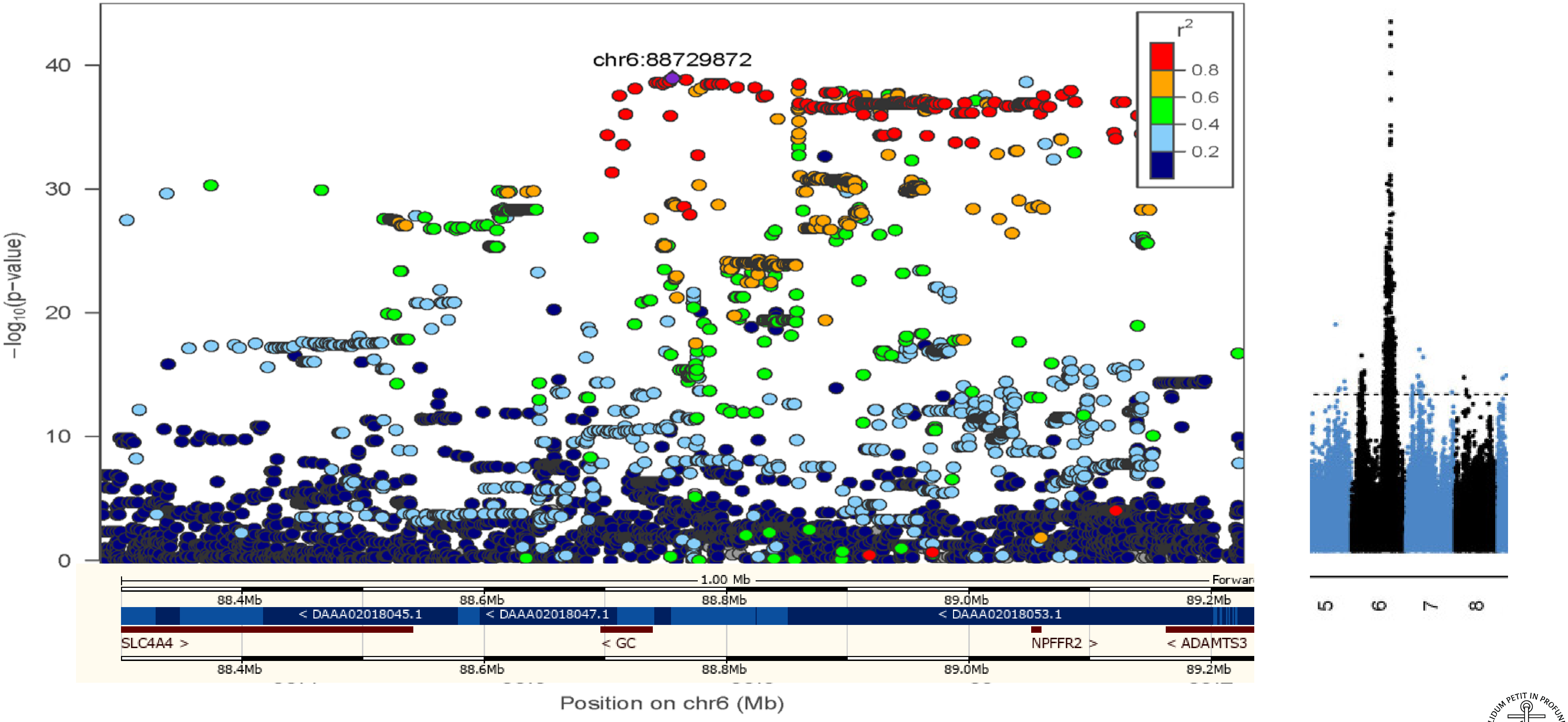
Collaboration is essential

Manhattan plot for the meta-analysis of bovine stature with $n = 58{,}265$ animals



Bouwman et al. Nature Genetics 50: 362-367(2018)

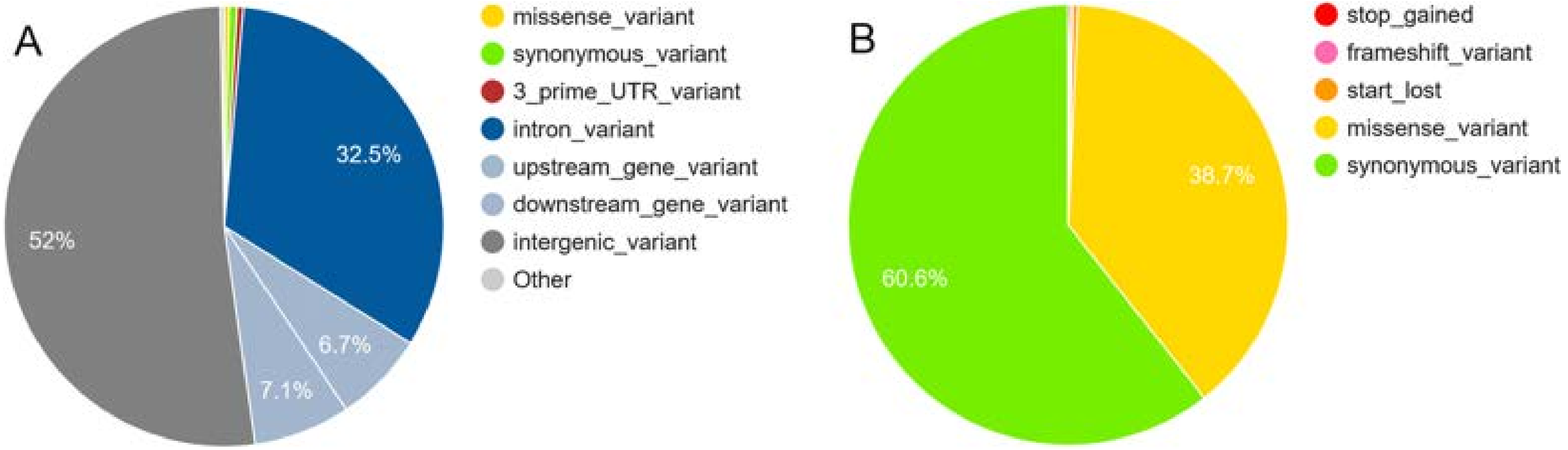
Linkage disequilibrium concealing causative locus



Cai et al. BMC Genomics 2018 **19**:656

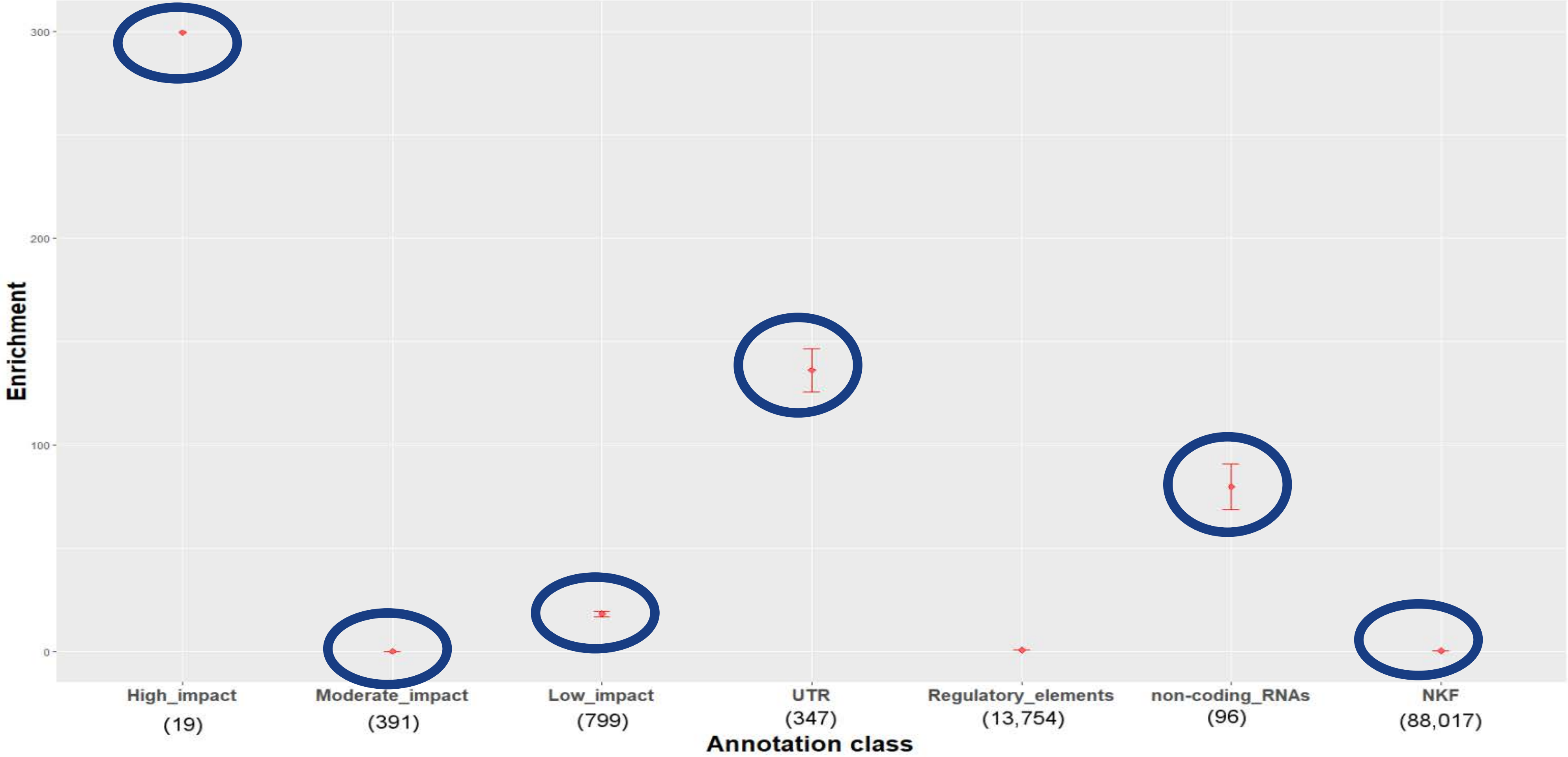
Path from GWAS to biology

- ➢ An association between a genetic variant at a genomic locus and a trait is not directly informative with respect to the target gene
- ➢ The mechanism whereby the variant is associated with phenotypic differences is not known
- ➢ New types of data have provided opportunities to bridge the knowledge gap from sequence to consequence.



Cai et al. BMC Genomics 2018 **19**:656

Enrichment of Variant Effect Predictor (VEP) annotations



Conclusions

1. The decade of GWAS constitutes a clear improvement in the recent history of reproducibility in genetic research; findings can be trusted.
2. Requirement for large sample sizes; a culture of data sharing
3. QTL-SNP and sequence variants at QTL peaks increase reliability of predictions; opportunity of utilizing across breed information
4. Can also be deployed to map molecular traits like gene expression, proteomic, and metabolomics measures; intermediate phenotype
5. RNA-based studies (eQTL) studies can identify variants that influence the gene's expression – may guide to establish causality
6. Functional annotation of cattle genome is incomplete; work (FAANG) in progress
7. The issue of establishing causality is a challenging one - plenty of biology to pursue

“The more we find, the more we see, the more we come to learn.”

Sir Tim Rice, Aida, 2000

Acknowledgements

Zexi Cai
Qianqian Zhang
Md Mesbah-Uddin
Xiaowei Mao
Thu Hong Le
Magdalena Dusza
Julia Gazzoni Jardim
Xiaoping Wu
Bernt Guldbrandtsen
Mogens S. Lund

AARHUS UNIVERSITY

GenSAP

Ministeriet for Fødevarer, Landbrug og Fiskeri
NaturErhvervstyrelsen
gudp

NAV Nordisk Avlsværdi Vurdering
Nordic Cattle Genetic Evaluation

InnovationsFonden
FORSKNING, TEKNOLOGI & VÆKST I DANMARK

mælkeafgiftsfonden

VIKING