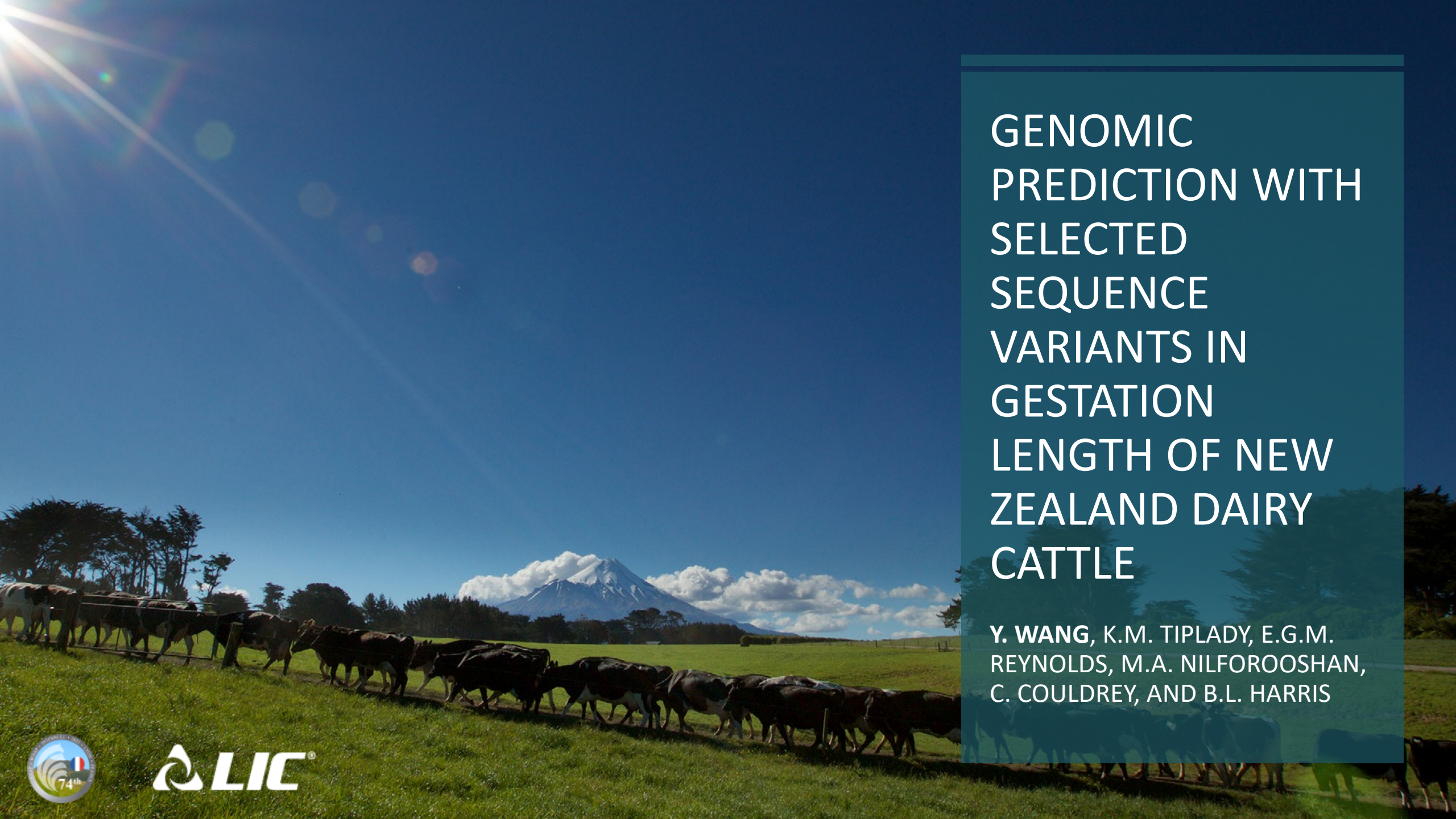
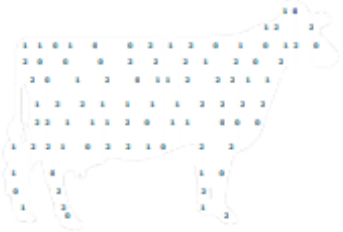


GENOMIC PREDICTION WITH SELECTED SEQUENCE VARIANTS IN GESTATION LENGTH OF NEW ZEALAND DAIRY CATTLE

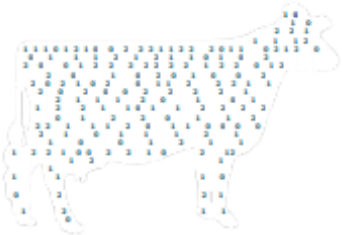
Y. WANG, K.M. TIPLADY, E.G.M. REYNOLDS, M.A. NILFOROOSHAN, C. COULDREY, AND B.L. HARRIS

High throughput genotyping

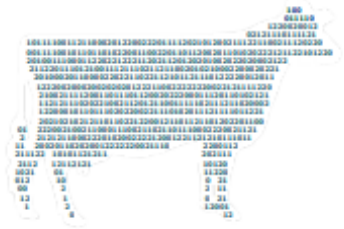
LD (7 K)



MD (54 K)



HD (777 K)



WGS (**21,000** K)

Access to whole-genome sequence data is easier nowadays

In theory, the sequence data should contain causal mutations associated with the genetic variation observed in phenotypic traits

In theory, the use of sequence data is expected to improve genomic evaluation

- Little improvement has been observed with using sequence variants in the prediction for dairy cattle
  - Only causative mutations or variants very close to causative mutations can improve reliability
  - non-causative mutations bring noise
  - Imperfect imputation of sequence

Discovery Population

Step 1:

GWAS analysis

Select the variants which have strong association with the trait

AIM

FIND THE OPTIMAL WAY OF SEPARATING ANIMALS INTO DISCOVERY, TRAINING AND VALIDATION POPULATION

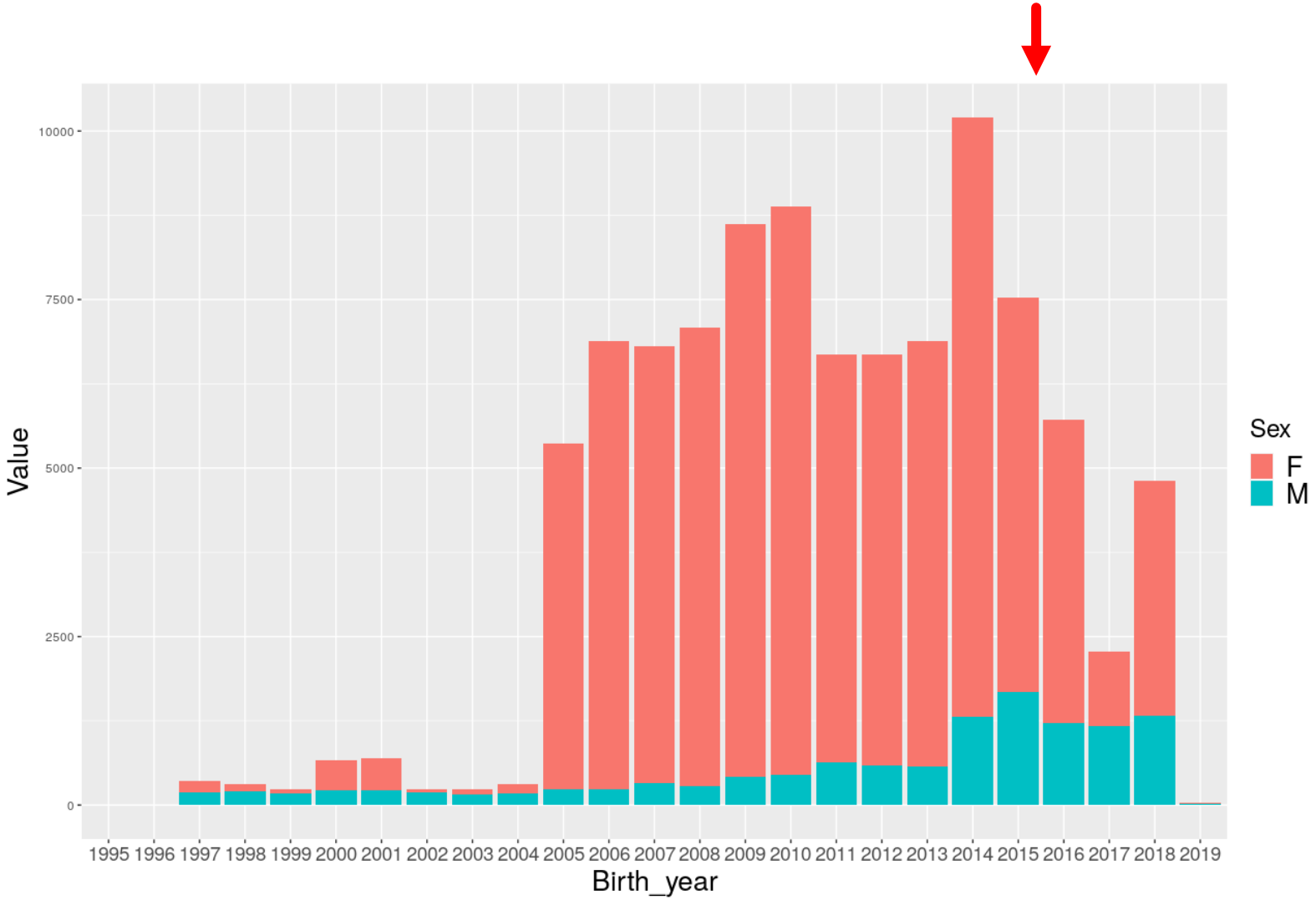
TEST IF ADDING SEQUENCE VARIANTS SELECTED FROM GWAS TO THE FILTERED ILLUMINA50K MARKERS WOULD BENEFIT GENOMIC PREDICTION

GESTATION LENGTH

- It is a measurement of the **calf** not its dam
- It was calculated as the difference between its dam's calving and mating dates in days
- Both male and female animals have only one gestation length record
- Moderately high heritable trait ($h^2 = 0.44 - 0.52$)
(Winkelman *et al.* 2001)

DATA

- 97,522 animals (33,577 HF <34.4%>, 17,377 J <17.8%>, 46,568 HFJ <47.8%>) have both imputed to sequence and yield deviation of gestation length corrected for contemporary group, sex of the calf, breed and inbreeding.
- Born between 1995 and 2019
- Filtered imputed to sequence data contains ~16million sequence variants (MAF> 0.005, imputation accuracy> 0.9)
- Animals born after 2016 were set as validation population. Parents of the validation animals were removed from both discovery and training population



Design 1
Bias on GWAS

Whole population

97,522 animals

Discovery set

GWAS

Total	**60,000**	**(61.52%)**	
Gender:	♂: 5878 (9.80%)	♀: 54,122 (90.20%)	
Breed:	HF: 20,452 (34.09%)	J: 10,921 (18.20%)	HF*J: 28,627 (47.71%)

Training set

Genomic prediction

Total	**24,690**	**(25.32%)**	
Gender:	♂: 2368 (9.59%)	♀: 22,322 (90.41%)	
Breed:	HF: 8352 (33.83%)	J: 4538 (18.38%)	HF*J: 11,800 (47.79%)

Validation set

Genomic prediction

Total	**12,832**	**(13.16%)**	
Gender:	♂: 3731 (29.08%)	♀: 9101 (70.92%)	
Breed:	HF: 4773 (37.20%)	J: 1918 (14.95%)	HF*J: 6141 (47.86%)

Design 2
Balance both functions

Whole population

97,522 animals

Discovery set

GWAS

Total	**42,345**	**(43.42%)**	
Gender:	♂: 4097 (9.68%)	♀: 38,248 (90.32%)	
Breed:	HF: 14,471 (34.17%)	J: 7701 (18.19%)	HF*J: 20,173 (47.64%)

Training set

Genomic prediction

Total	**42,345**	**(43.42%)**	
Gender:	♂: 4149 (9.80%)	♀: 38,196 (90.20%)	
Breed:	HF: 14,333 (33.85%)	J: 7758 (18.32%)	HF*J: 20,254 (47.83%)

Validation set

Genomic prediction

Total	**12,832**	**(13.16%)**	
Gender:	♂: 3731 (29.08%)	♀: 9101 (70.92%)	
Breed:	HF: 4773 (37.20%)	J: 1918 (14.95%)	HF*J: 6141 (47.86%)

Design 3
Separate by birth year

Whole population

89,738 animals

Discovery set

GWAS

Total	**38,924**	**(46.14%)**	
Gender:	♂: 3152 (8.10%)	♀: 35,772 (91.90%)	
Breed:	HF: 13,433 (34.51%)	J: 8525 (21.90%)	HF*J: 16,966 (43.59%)

Born before 2010

Training set

Genomic prediction

Total	**37,982**	**(40.70%)**	
Gender:	♂: 4774 (12.57%)	♀: 33,208 (87.43%)	
Breed:	HF: 12,569 (33.09%)	J: 5744 (15.12%)	HF*J: 19,669 (51.79%)

Born between 2010 and 2016

Validation set

Genomic prediction

Total	**12,832**	**(13.16%)**	
Gender:	♂: 3731 (29.08%)	♀: 9101 (70.92%)	
Breed:	HF: 4773 (37.20%)	J: 1918 (14.95%)	HF*J: 6141 (47.86%)

Born after 2016

Design 4
Same dataset for discovery and training

Whole population

97,522 animals

Discovery set

GWAS

Total	**84,690**	**(86.84%)**	
Gender:	♂: 8246 (9.74%)	♀: 76,444 (91.90%)	
Breed:	HF: 28,804 (34.01%)	J: 15,459 (18.25%)	HF*J: 40,427 (47.74%)

Training set

Genomic prediction

Total	**84,690**	**(86.84%)**	
Gender:	♂: 8246 (9.74%)	♀: 76,444 (91.90%)	
Breed:	HF: 28,804 (34.01%)	J: 15,459 (18.25%)	HF*J: 40,427 (47.74%)

Validation set

Genomic prediction

Total	**12,832**	**(13.16%)**	
Gender:	♂: 3731 (29.08%)	♀: 9101 (70.92%)	
Breed:	HF: 4773 (37.20%)	J: 1918 (14.95%)	HF*J: 6141 (47.86%)

STATISTIC MODELS

Iterative GWAS

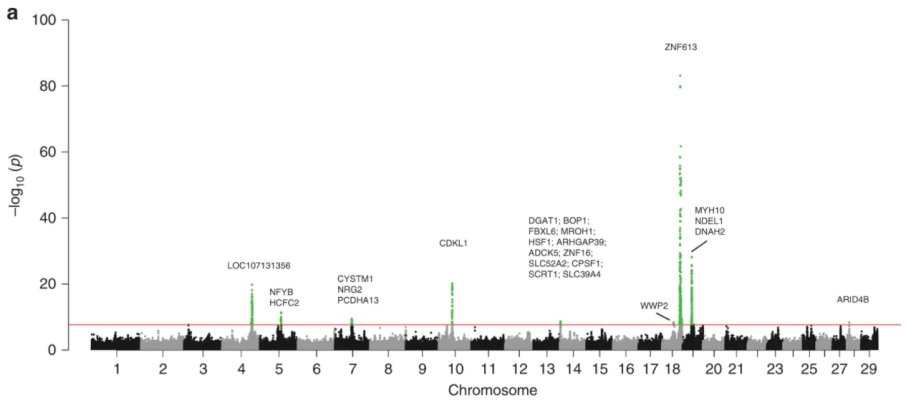
- Leave-one-segment-out strategy using Bolt-LMM (Loh P-R *et al.* 2015)
- Filtered Illumina50k markers were used for capturing population structure
- Significant variants for each chromosome will be set as co-variates for the next GWAS iteration. Iteration will stop once no significant variant shows
- Two p-values were used: $5 \times 10^{-8}$and 1×10^{-5}

Genomic Prediction

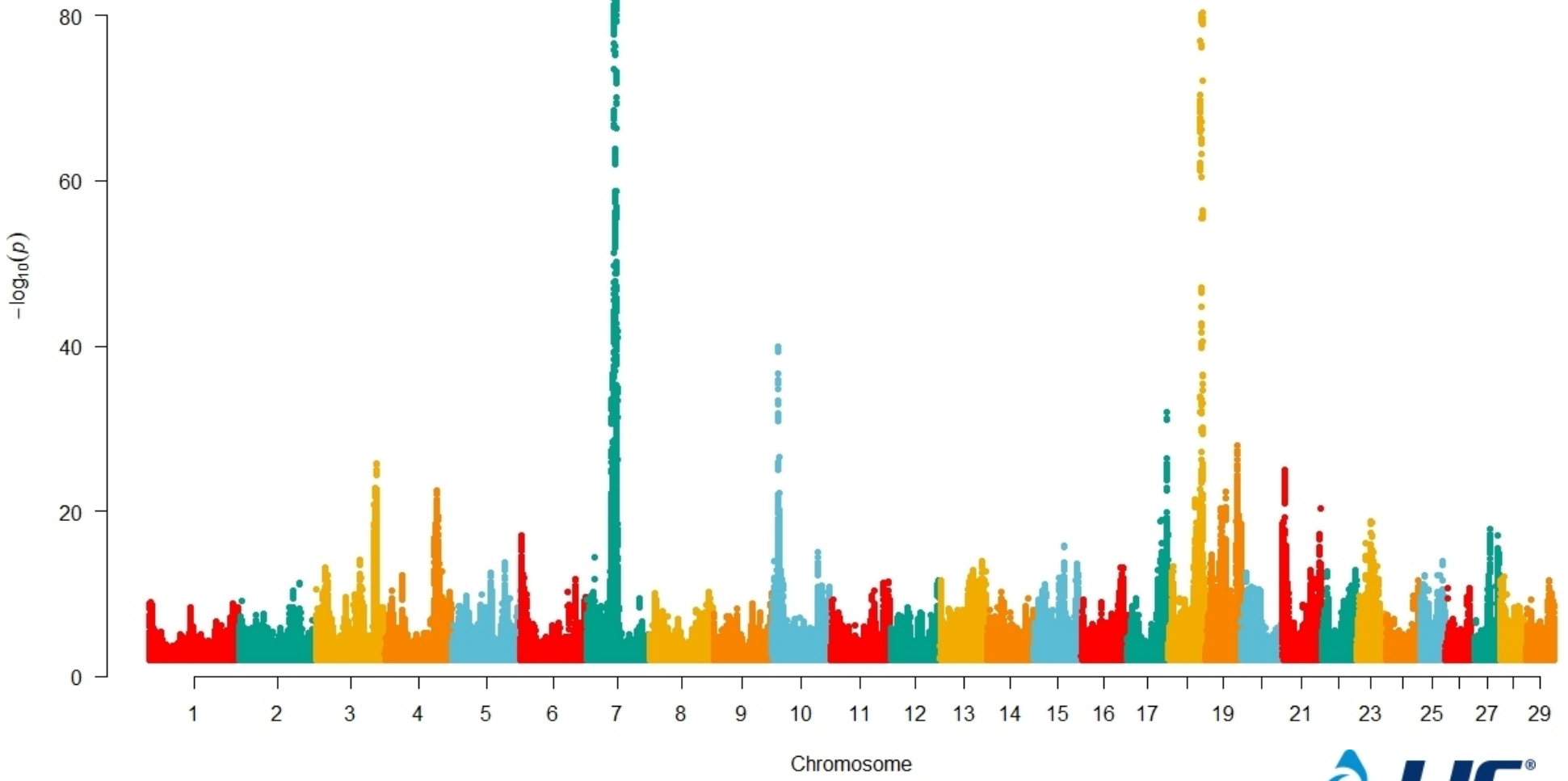
- Univariate model with BayesR implemented in GCTB (Zeng *et al.* 2018 Nature Genetics) with default settings

$$y = 1\mu + X\beta + e$$

- Prediction accuracy was calculated in the validation set as the correlation between the predicted GEBVs and the yield deviation
- Prediction bias was calculated as the regression coefficient of the yield deviation on the predicted GEBVs



27,214 Holstein bulls,
(Fang *et al.* Communications biology 2019)



COMPUTATIONAL TIME (GWAS)

$p: 5 \times 10^{-8}$

Bias_GWAS (n=60,000): **22** iterations, **283** variants (5-03:45:30)

Balance (n=42,345): **16** iterations, **205** variants (1-17:36:11)

Birth_Year (n=38,924): **13** iterations, **174** variants (1-06:10:15)

Both (n=84,690): **26** iterations, **391** variants (10-16:51:25)

$p: 1 \times 10^{-5}$

Bias_GWAS (n=60,000): **37** iterations, **689** variants (20-07:36:11)

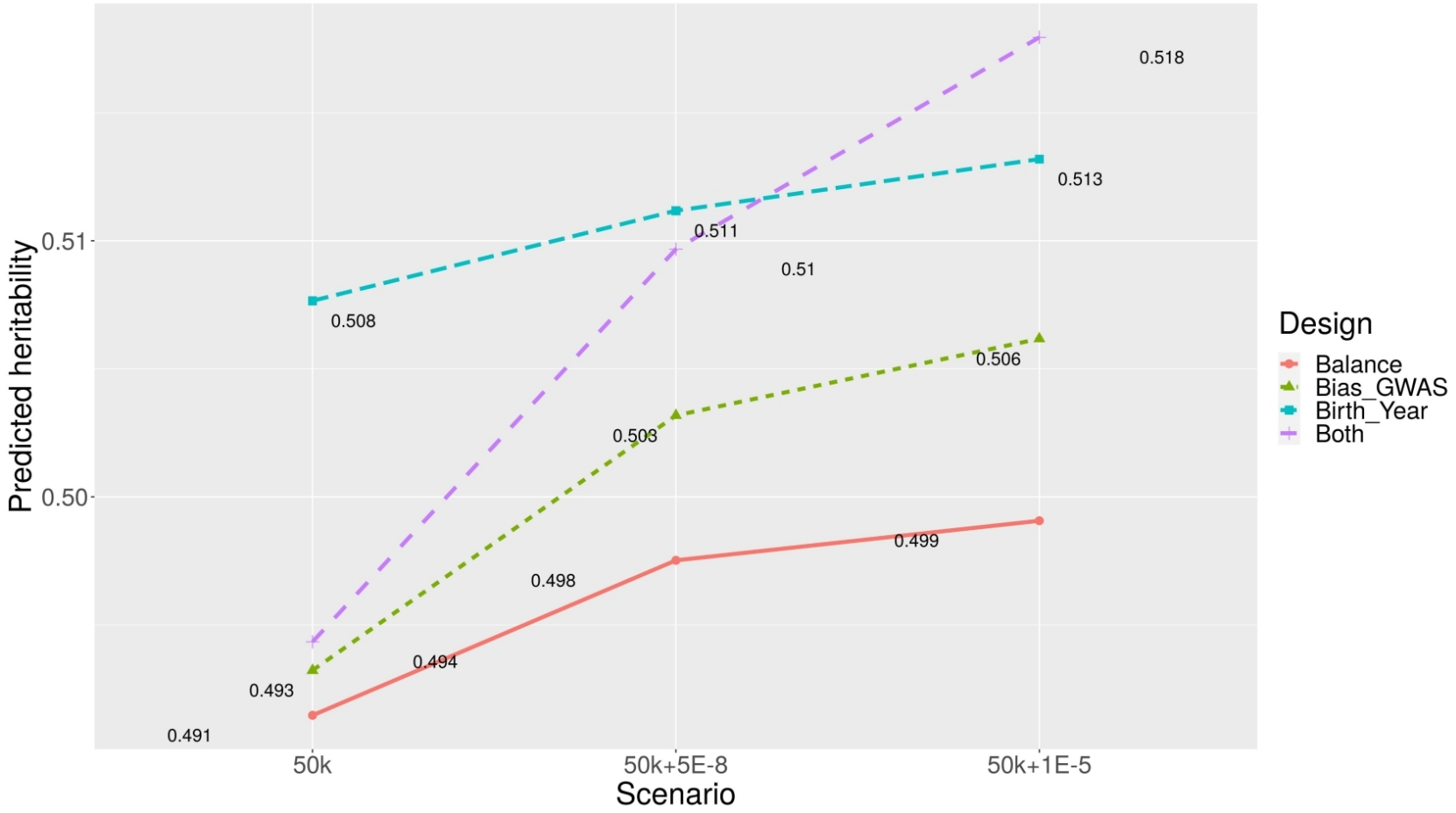
Balance (n=42,345): **30** iterations, **484** variants (7-16:53:48)

Birth_Year (n=38,924): **21** iterations, **392** variants (5-18:00:34)

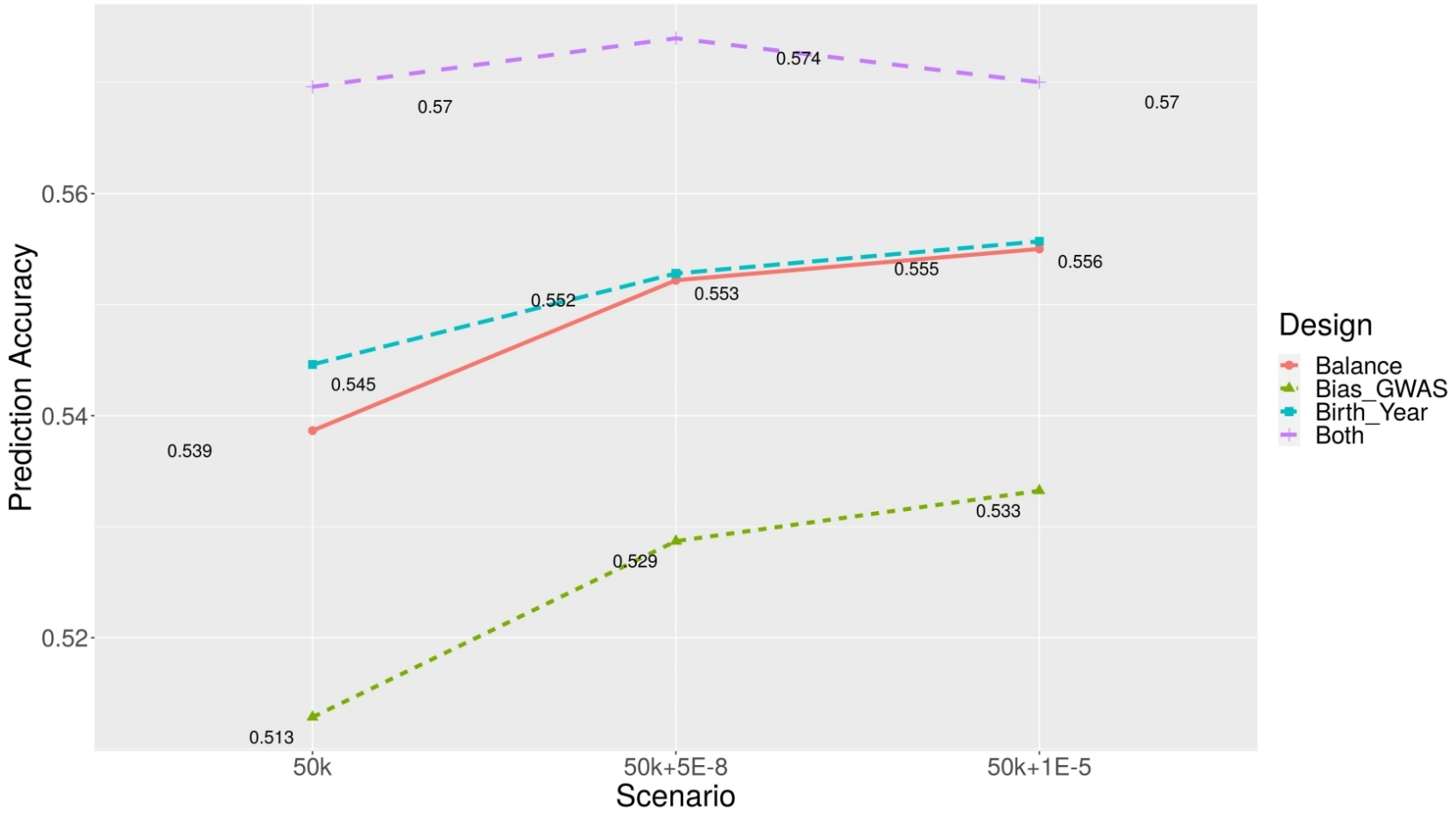
Both (n=84,690): **42** iterations, **783** variants (37-17:53:27)

NeSI: requested for 16G memory 16CPU for each analysis

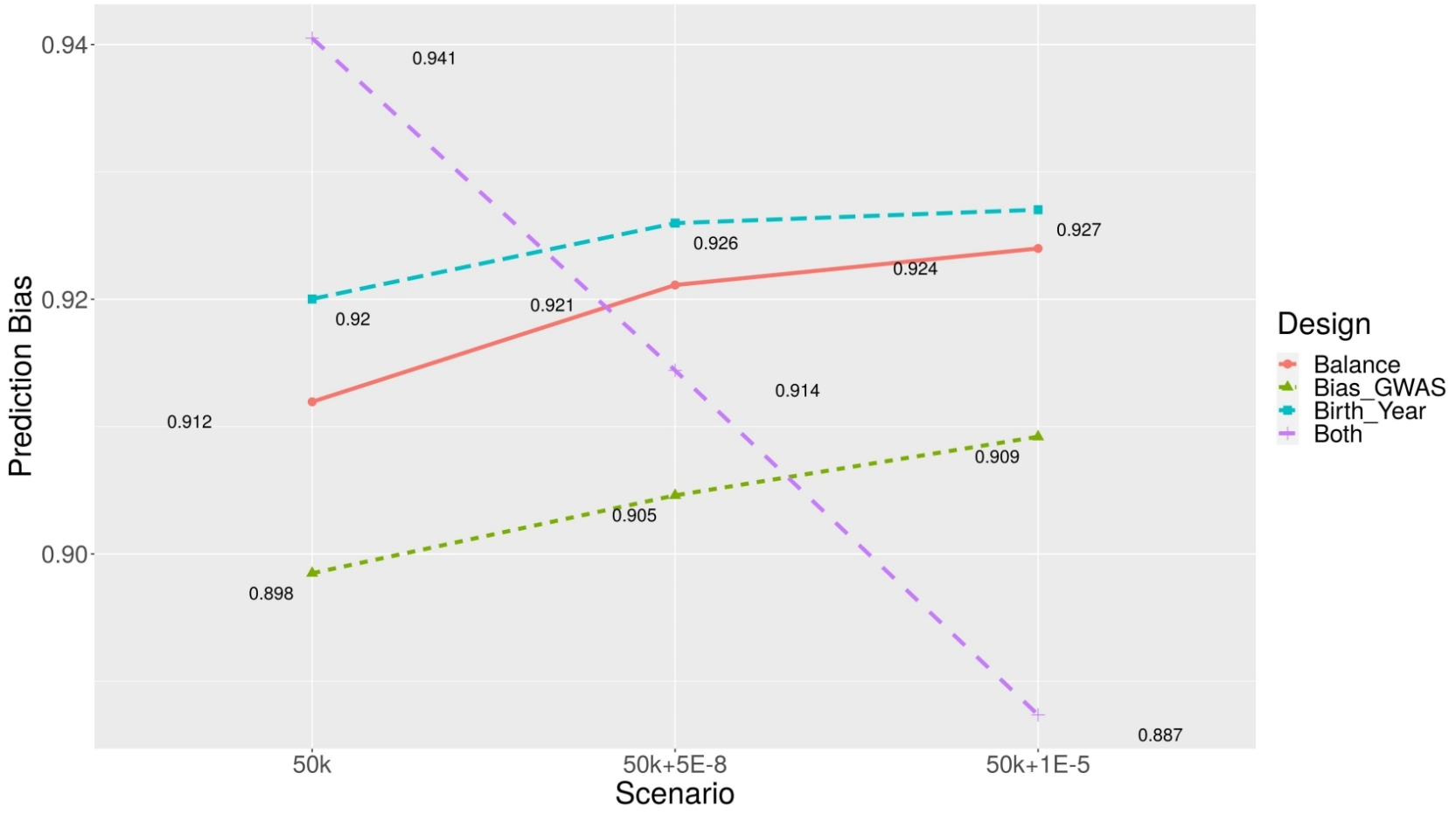
PREDICTED HERITABILITY



PREDICTION ACCURACY

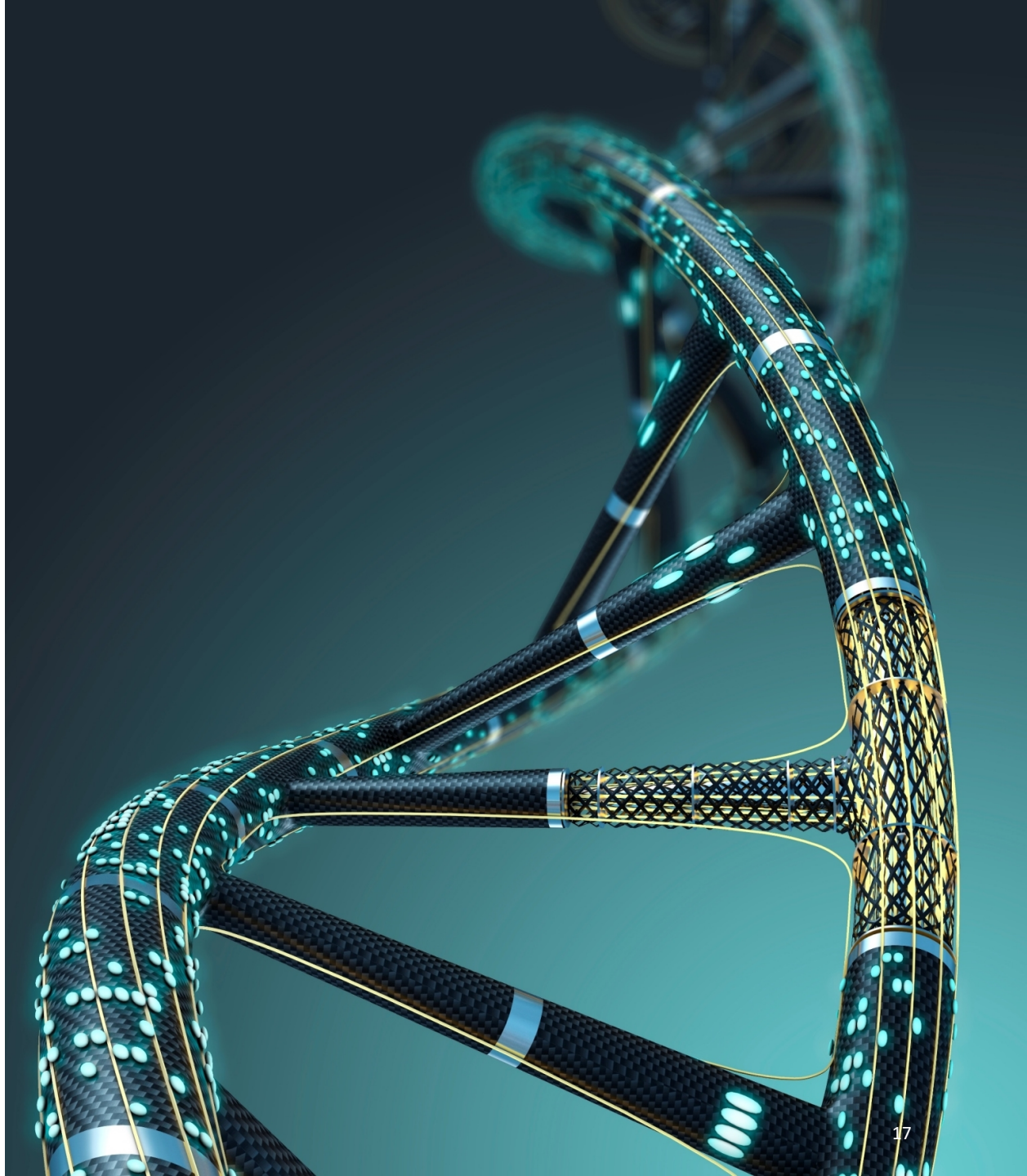


PREDICTION BIAS



TAKE HOME MESSAGE

- More variants were selected when more animals were added to the discovery set. However, the benefit of adding more SNPs in the prediction model did not exceed the benefit of adding more animals to the training population.
- Same population used as the discovery and training population achieved the highest prediction accuracy along with the highest bias, which is not desirable.
- Based on birth year, separation is the best option. A less stringent p-value leads to more iterations and more sequence variants selected, increasing the prediction accuracy. However, it takes much more time.



ACKNOWLEDGEMENT

This study was supported by Genomics Aotearoa Better Breeding Values project, MPI Sustainable Food and Fibre Futures (SFFF) and MPI Resilient Dairy Research Program. Computational resources were provided by New Zealand eScience Infrastructure (NeSI).

Thank you