# *Prediction of causative genomic relationships using sequence data of five French and Danish dairy cattle breeds*

Irene van den Berg[1,2,3], D. Boichard[2,3] and M. S. Lund[1]

EGS ABG AgroParisTech APIS-GENE AARHUS UNIVERSITY INRA SCIENCE & IMPACT European Commission ERASMUS MUNDUS QGG

[1]Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University, Tjele, Denmark, [2]INRA, UMR1313 Génétique Animale et Biologie Intégrative, Jouy-en-Josas, France, [3]AgroParisTech, UMR1313 Génétique Animale et Biologie Intégrative, Paris, France

# *Introduction*

- Increasing number of sequences individuals
  → possible to use for genomic selection

- Sequence contains causative mutations
  → increase prediction accuracy?

- Across breed: low accuracy using 50K/HD chips → insufficient linkage disequilibrium across breed?

- Low MAF variants not on SNP chips

# *Objective*

To study the potential benefits of sequence data for the prediction of genomic relationships

Different scenarios:

- Within and across breed
- Number of causative mutations
- Distance between causative mutations and prediction markers
- Compare with 50K/HD
- MAF of causative mutations and prediction markers

# Methods

Quantify loss in prediction R$^2$ following de los Campos *et al.* (2013):

R$^2$ if markers are in perfect LD with causative mutations

minimum R$^2$ reduction factor

1-(1-b)$^2$:

$$\bar{R}^2_{n+1,y} \leq R^2_{n+1,y}\left[1-(1-b_{n+1})^2\right]$$

R$^2$ if markers are not in perfect LD with causative mutations

regression coefficient

genomic relationship at causative mutations

b:

$$\bar{G}_{n+1,i} = b_{n+1} G_{n+1,i} + \xi_{n+1,i} \qquad (i = 1, \dots, n)$$

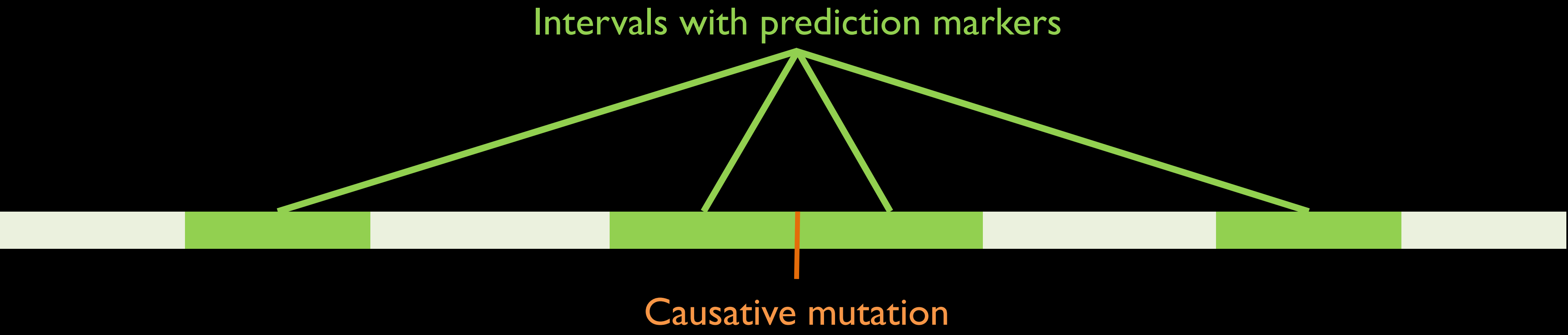
genomic relationship at prediction markers between individual *n* + 1 and individual *i*

residuals

# Methods

- Genomic relationship matrix at causative mutations
  - 10/50/100/250 randomly sampled variants

- Genomic relationship matrix at prediction markers
  - 50K / HD: SNP on 50K / HD chip
  - 50K / HD closest: for each causative mutation, the closest 50K / HD marker
  - Two 1 Kb intervals on both sides of the causative mutations, distance between causative mutations and intervals between 1b and 1Mb

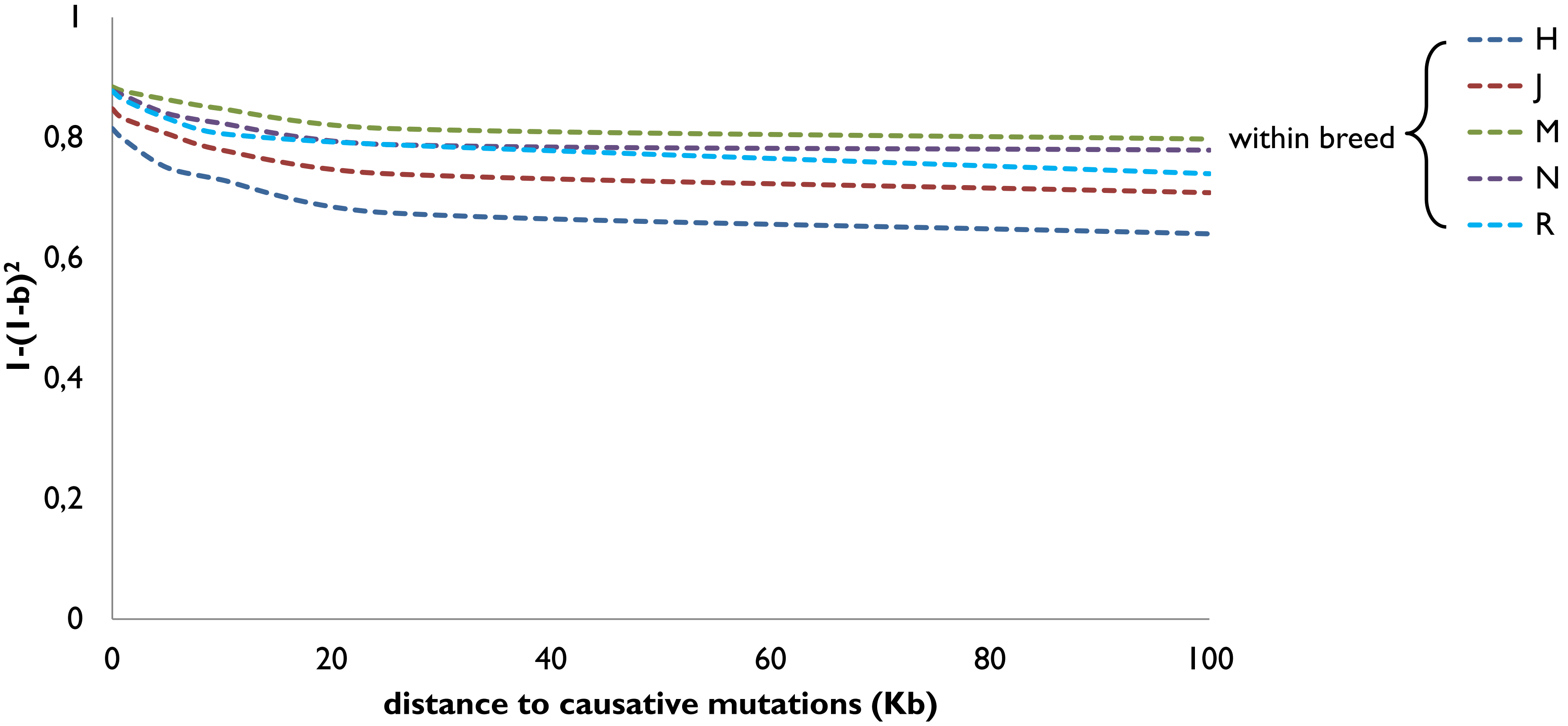


# Data

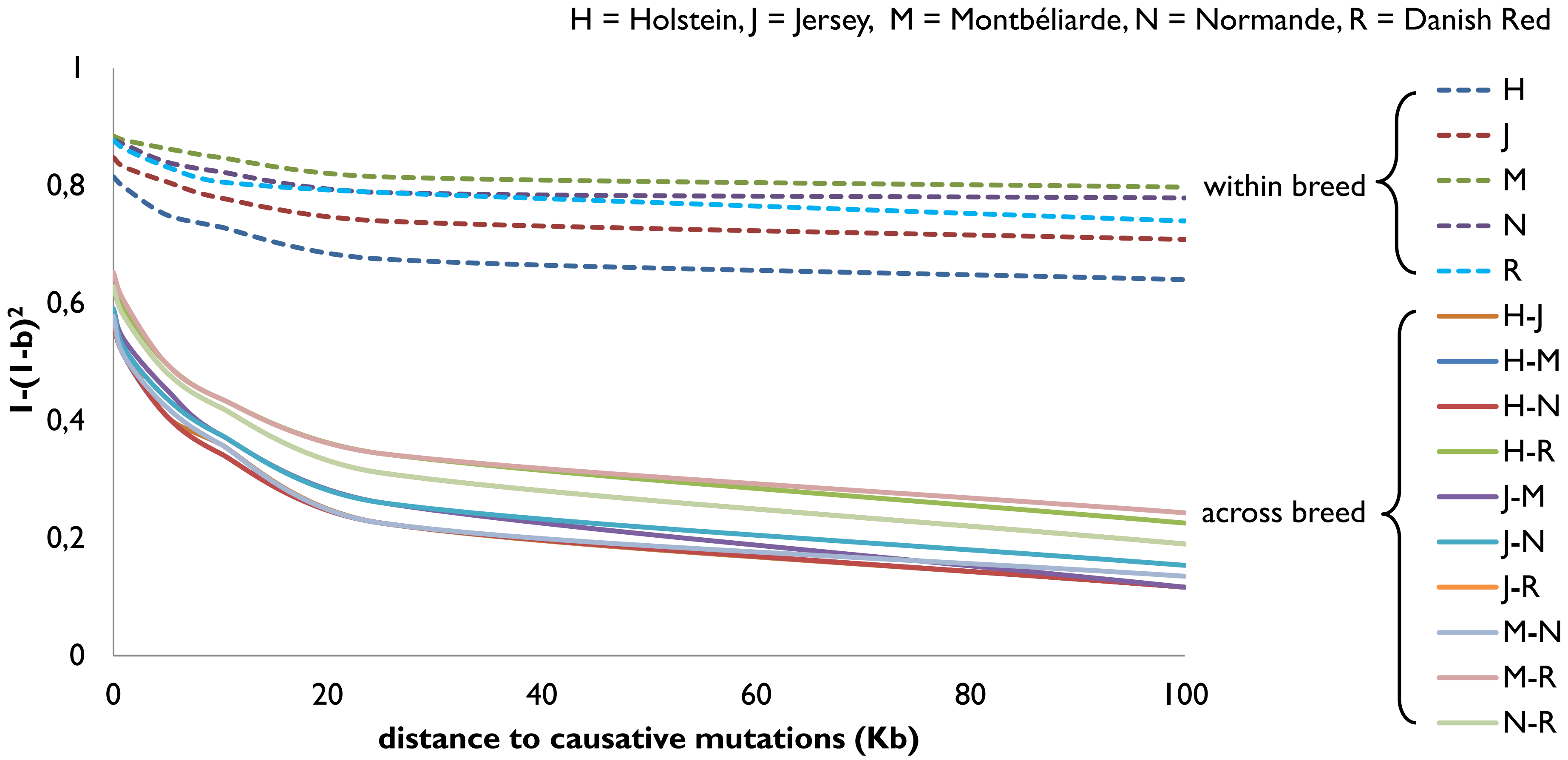
- Sequences, chromosome 1
- 122 Holstein, 27 Jersey, 28 Montbéliarde, 23 Normande and 45 Danish Red
- Causative mutations selected from:
    - All variants segregating in at least one breed
    - Variants with MAF ≤ 0.10
- Prediction markers selected from:
    - All variants segregating in at least one breed
    - Variants with MAF ≥ 0.10
    - Variants present on the 50K/HD chip
- Each scenario was repeated 50 times

# Results – *Within breed* *(100 causative mutations)*

H = Holstein, J = Jersey, M = Montbéliarde, N = Normande, R = Danish Red

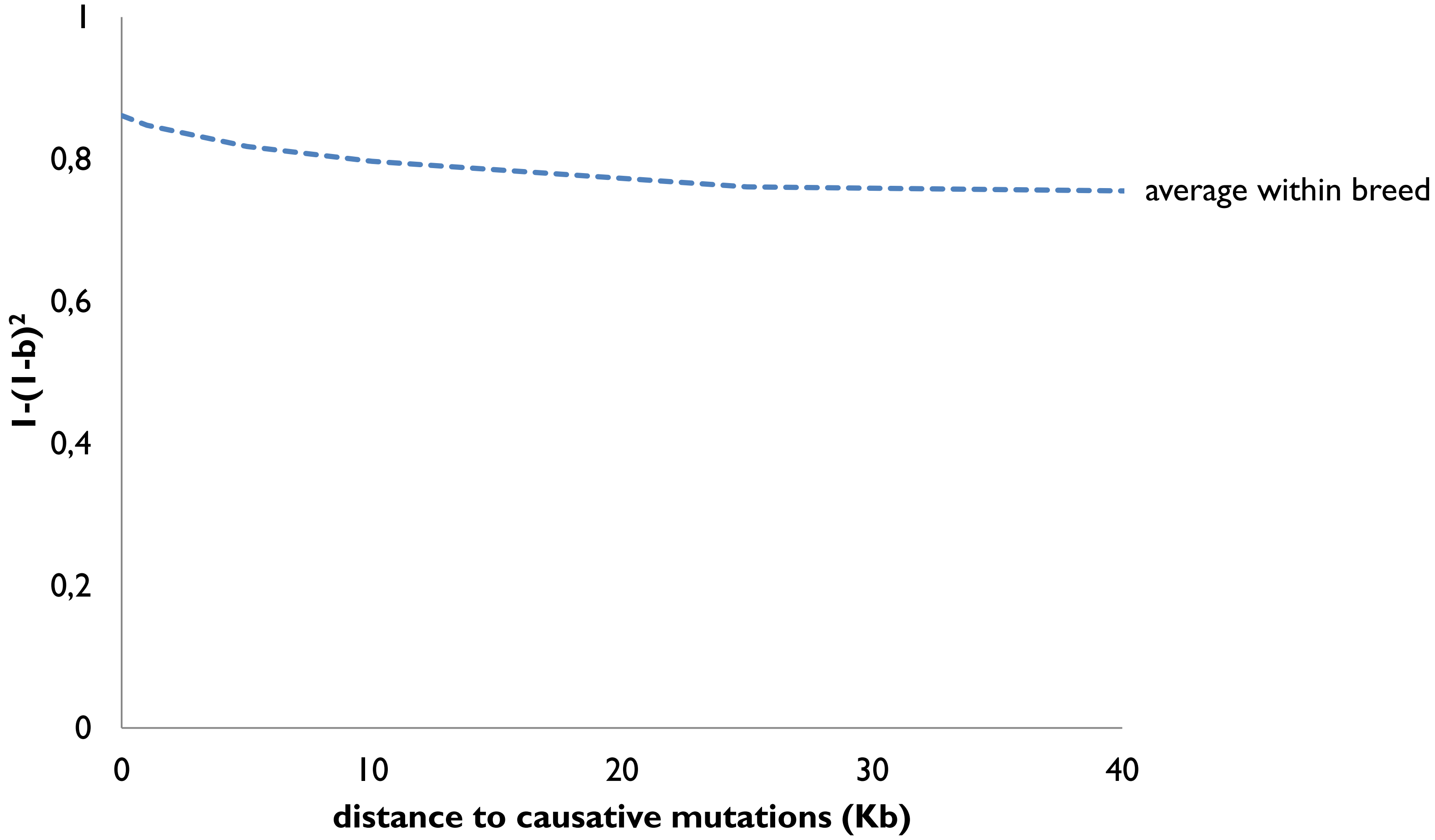


# *Results – Across breed (100 causative mutations)*

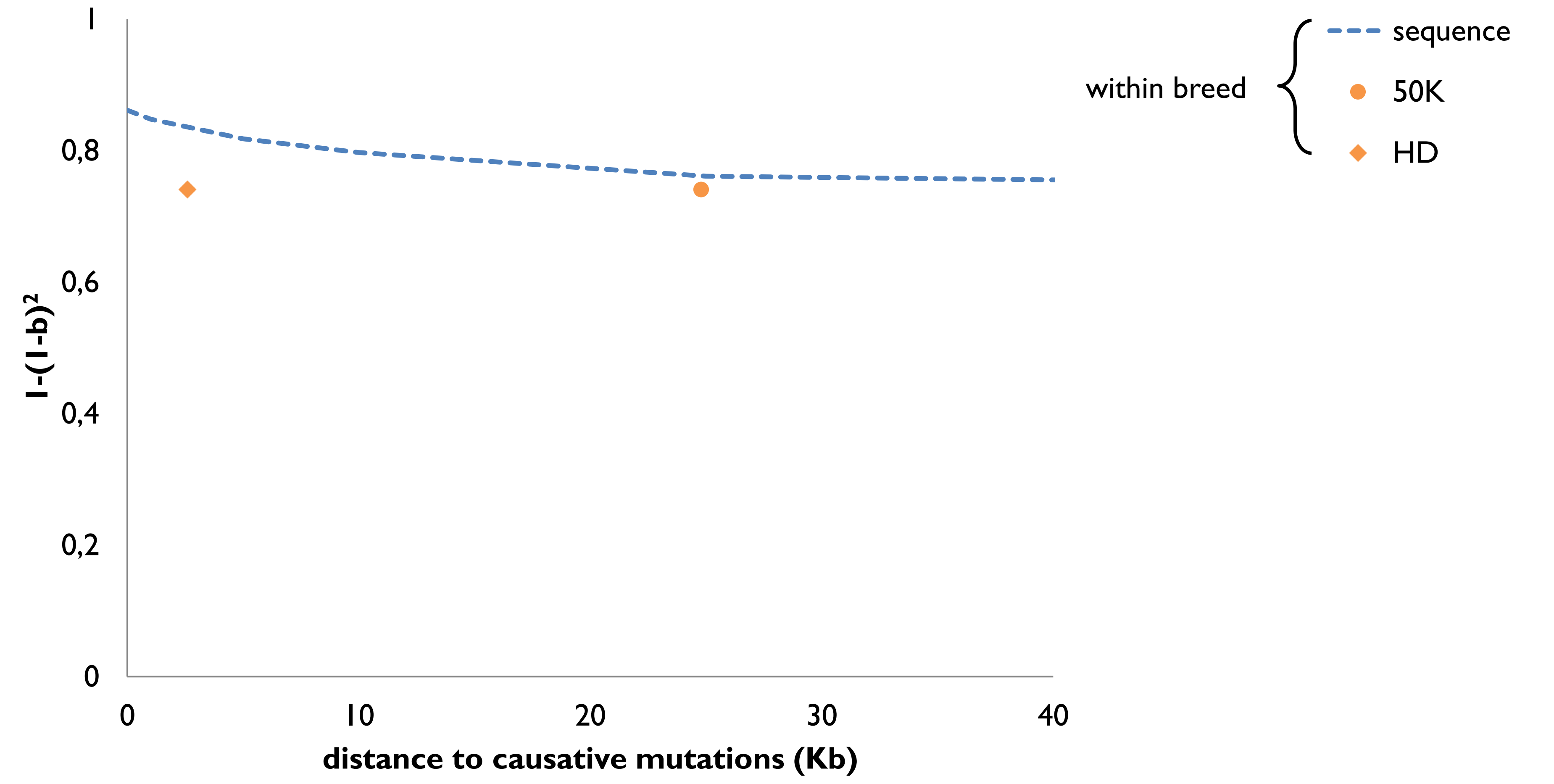


→ $1-(1-b)^2$ decreases when distance between prediction markers and causative mutations increases, faster decrease across breed

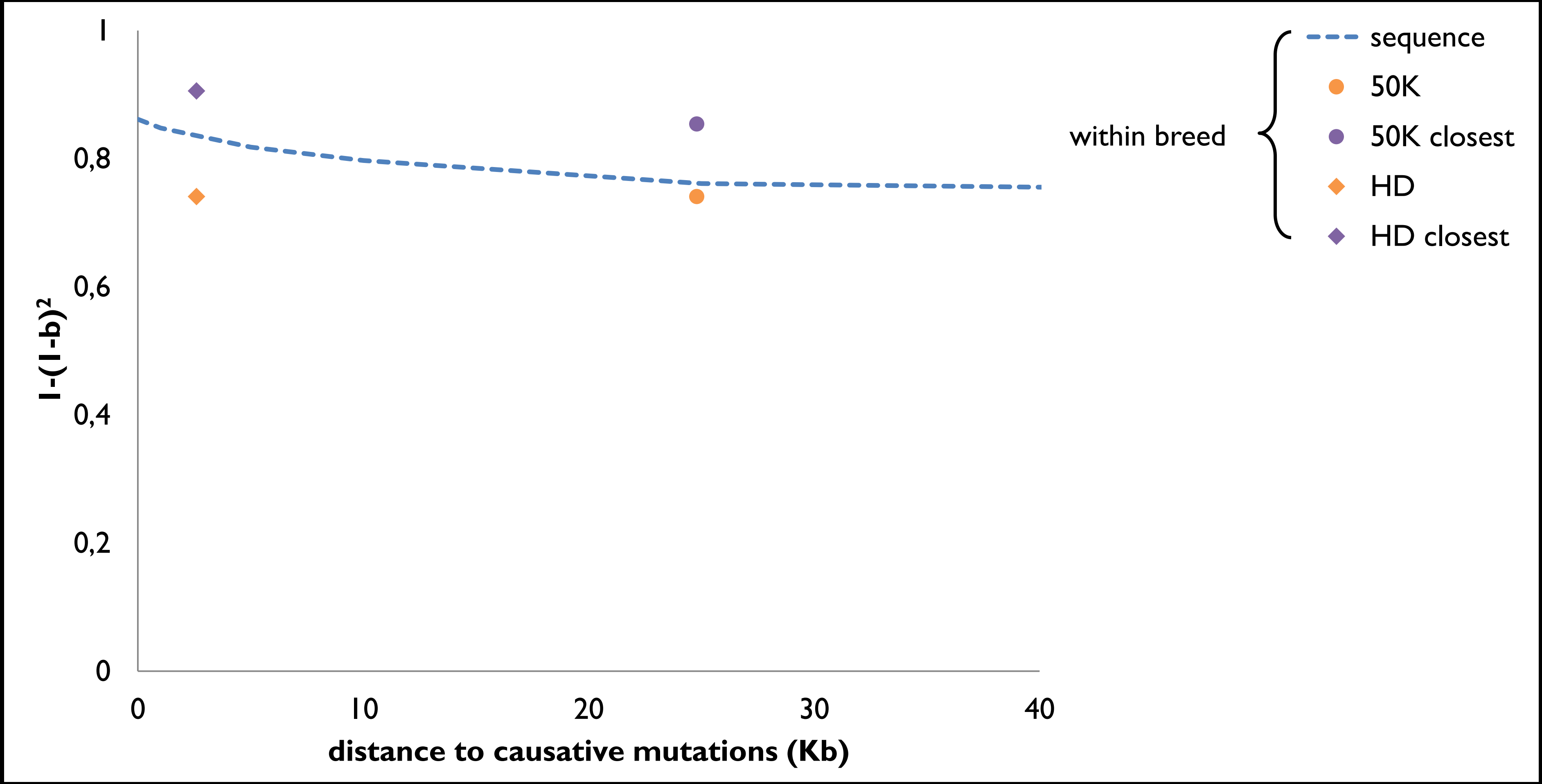
# *Results – Sequence & SNP chips* *(100 causative mutations)*



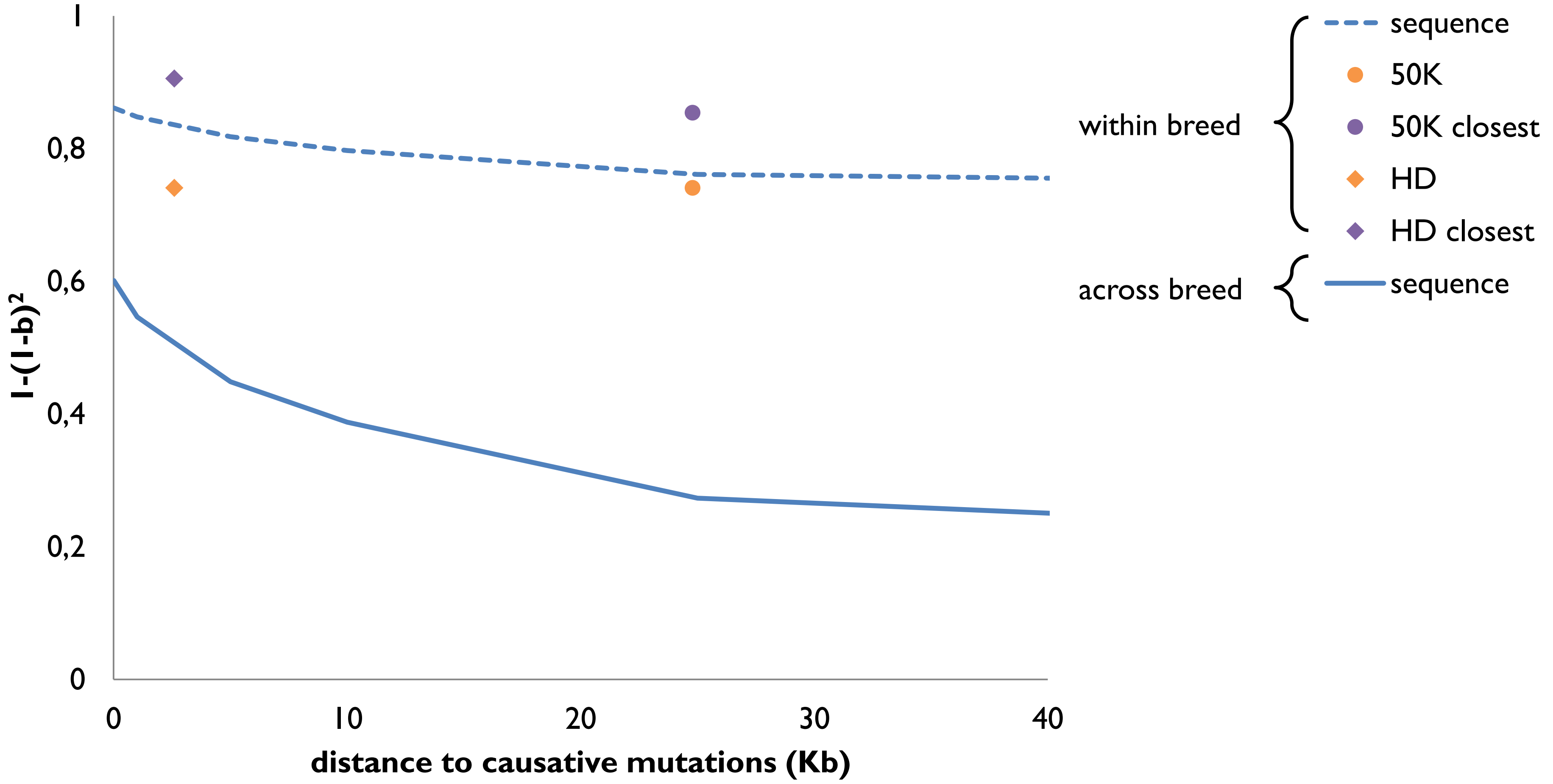
# *Results – Sequence & SNP chips* *(100 causative mutations)*



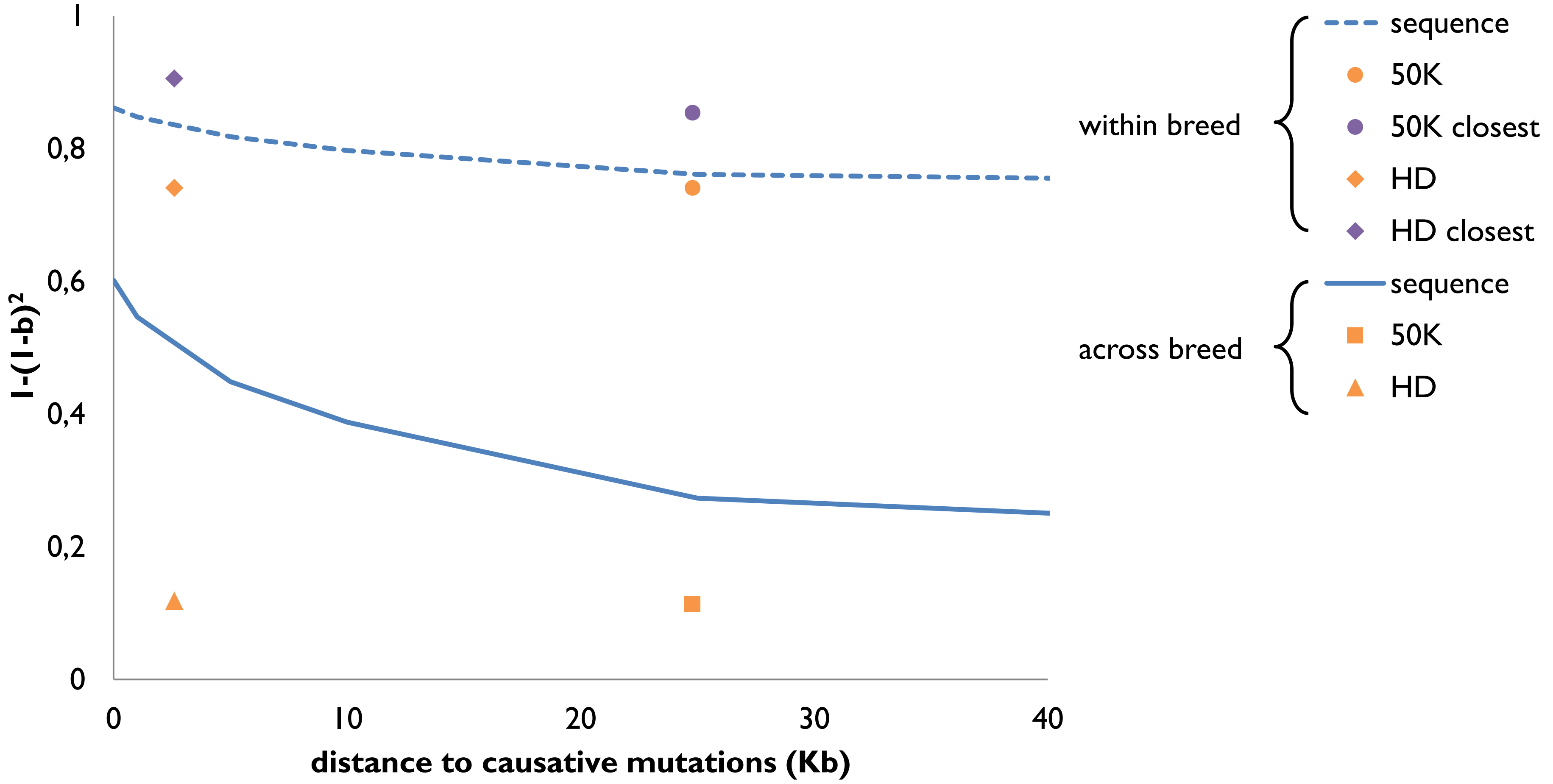
# Results – Sequence & SNP chips *(100 causative mutations)*



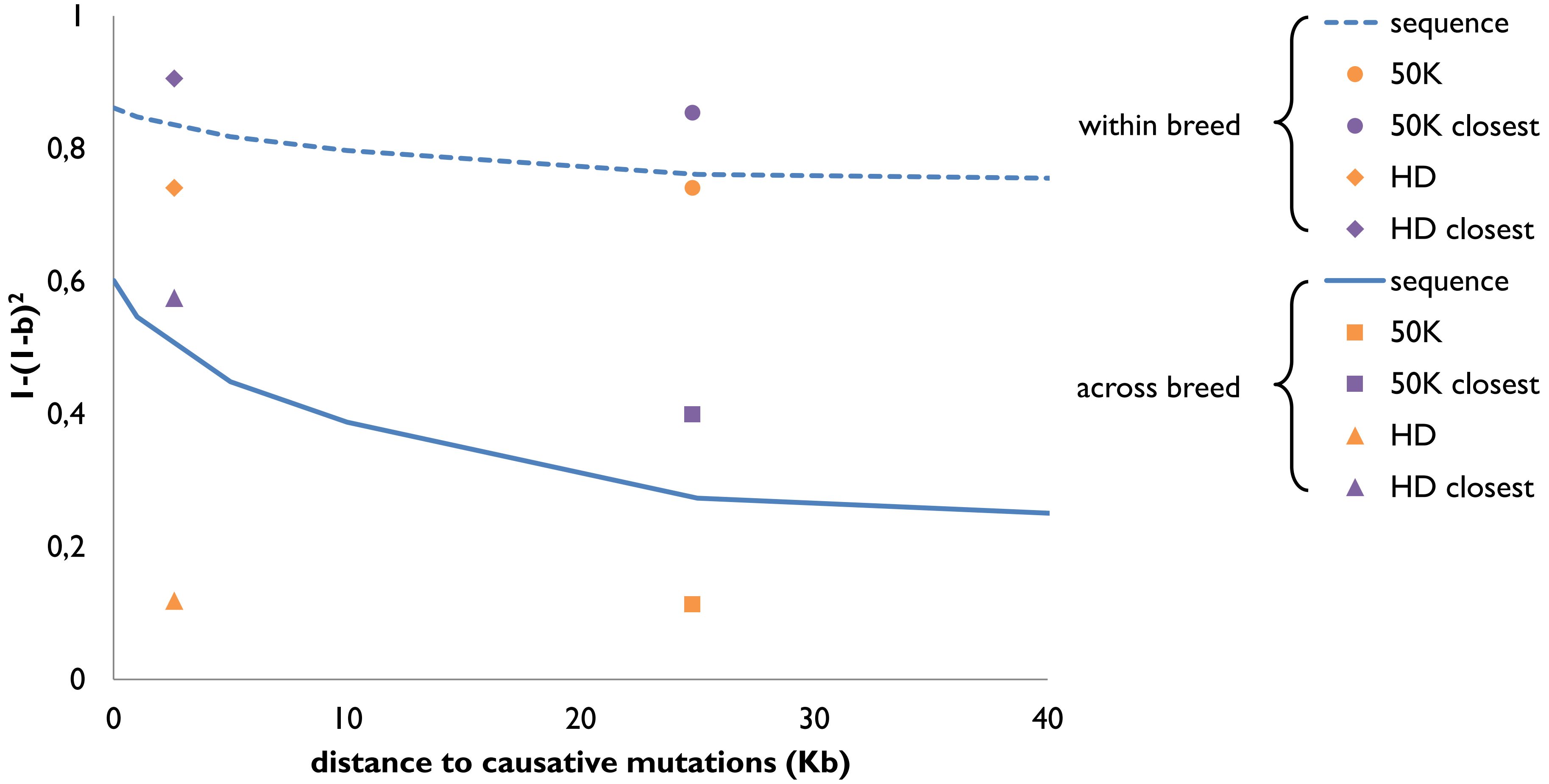
# Results – Sequence & SNP chips *(100 causative mutations)*



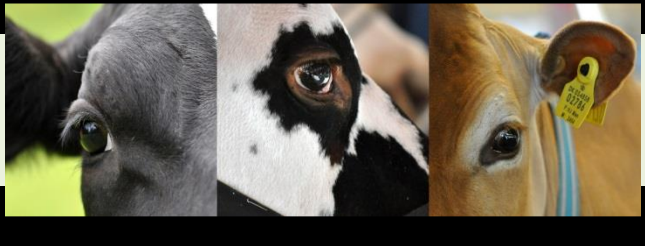
# Results – Sequence & SNP chips *(100 causative mutations)*



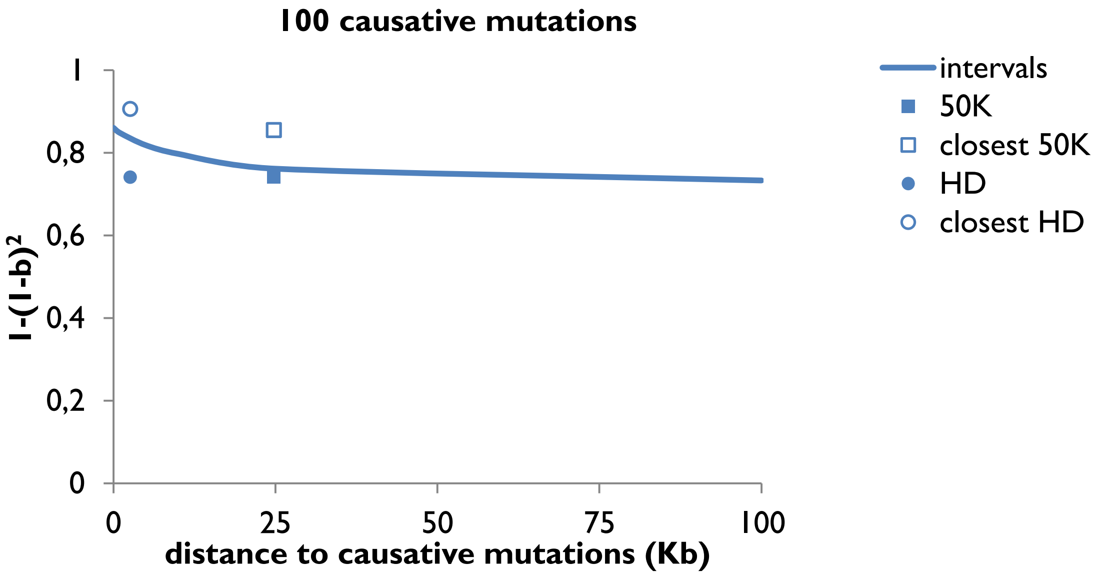
# Results – Sequence & SNP chips *(100 causative mutations)*



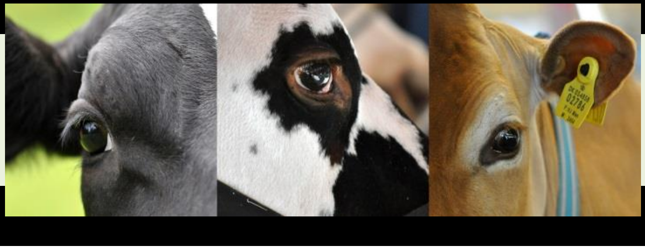
→ Using all 50K/HD markers → lower $1\text{-}(1\text{-}b)^2$ compared to sequence, but higher when only the markers closest to the causative mutations are used



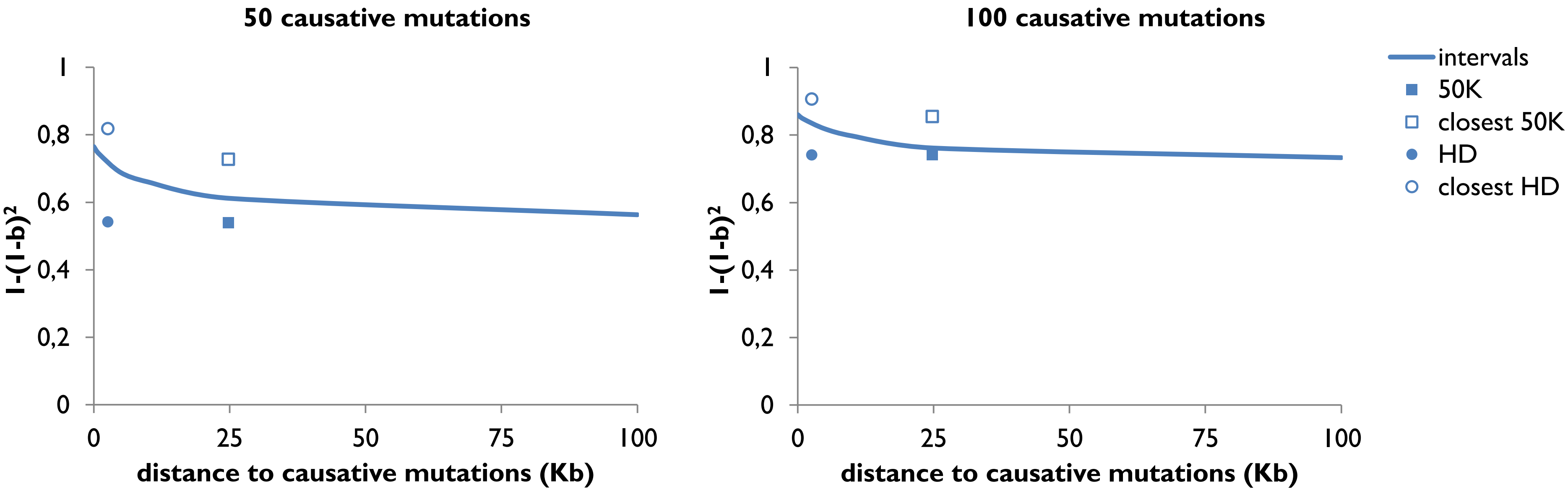
# Results – Number of mutations *(within breed)*



100 causative mutations

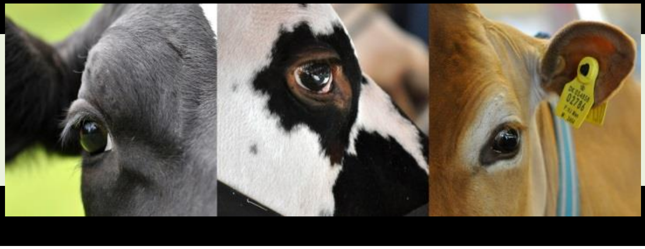


# Results – Number of mutations *(within breed)*

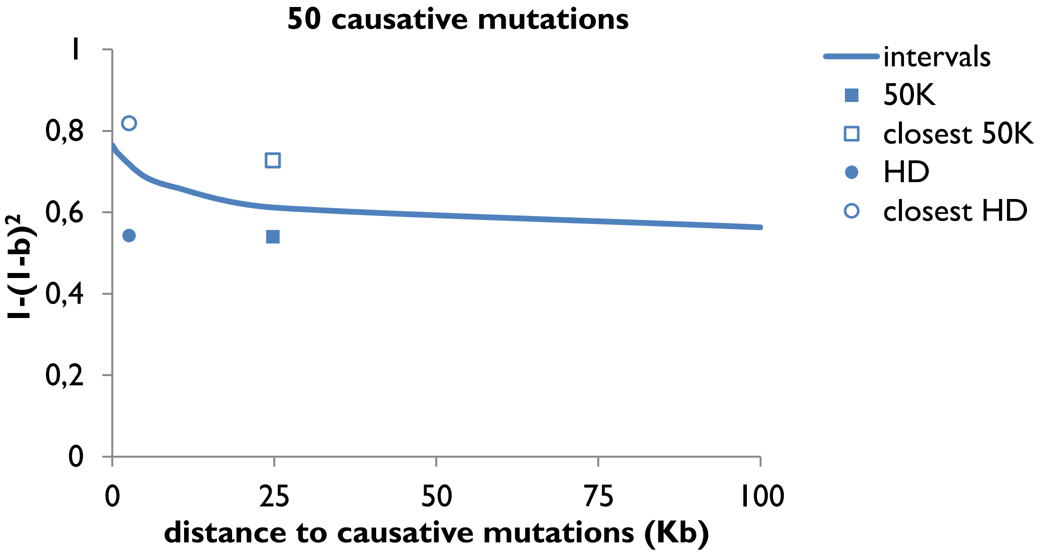


50 causative mutations

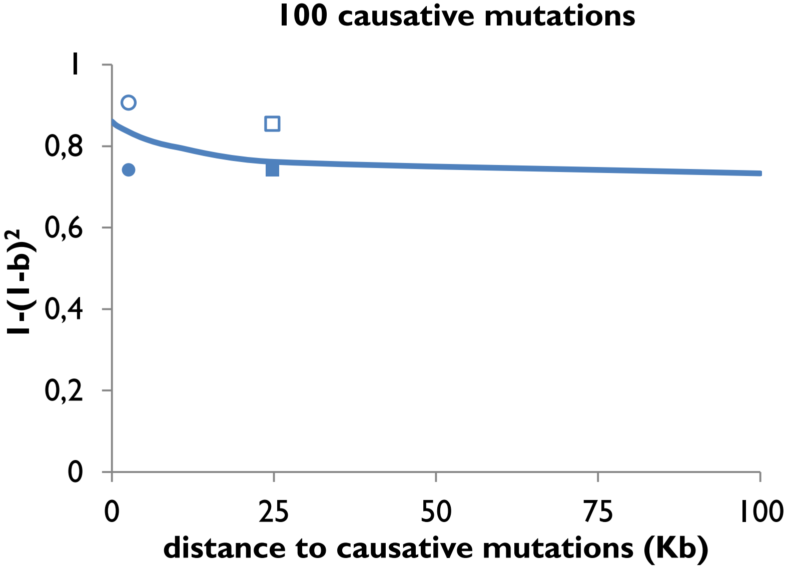
100 causative mutations



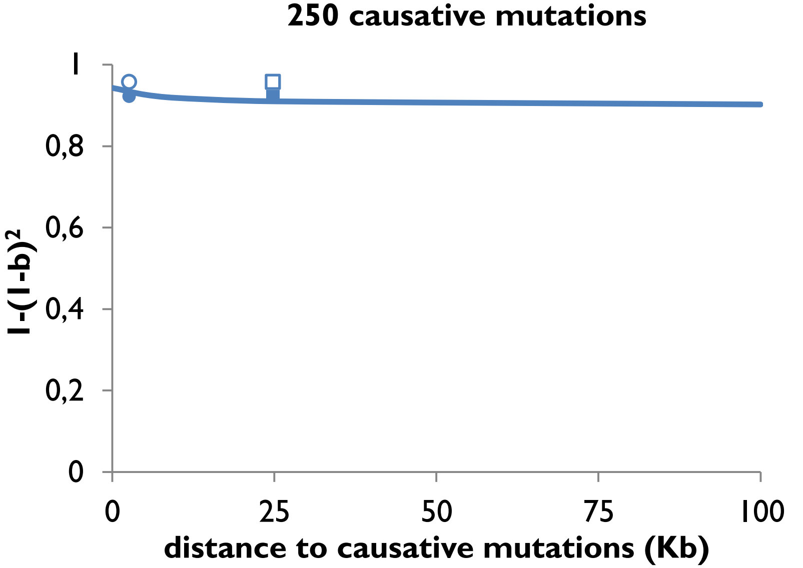
# *Results – Number of mutations (within breed)*



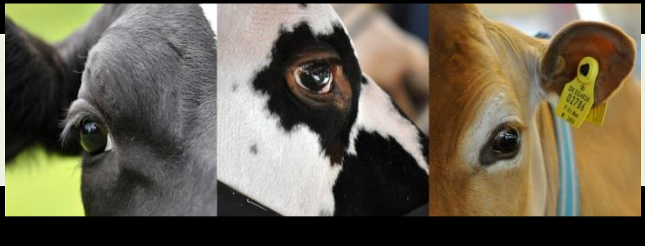
50 causative mutations



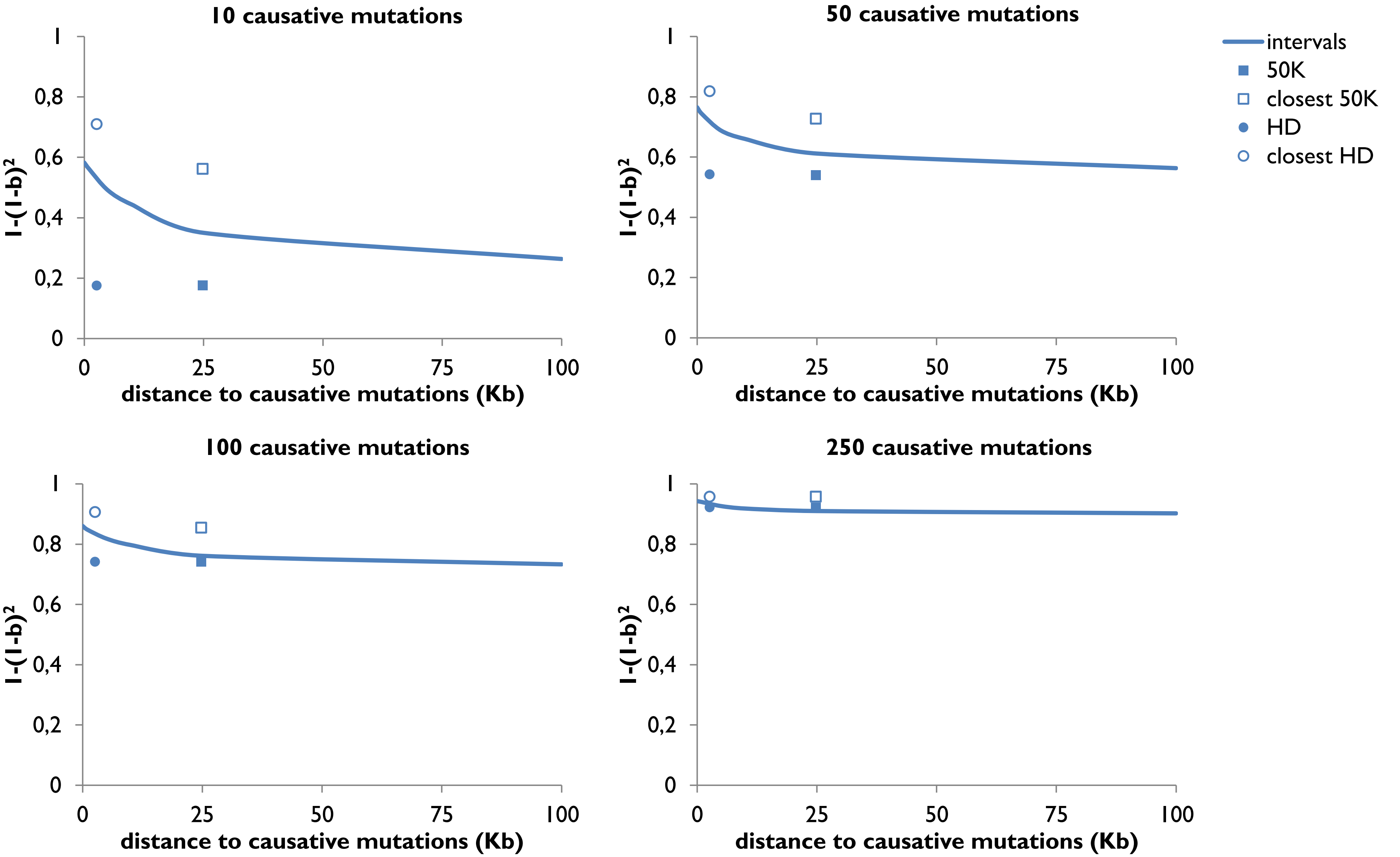
100 causative mutations

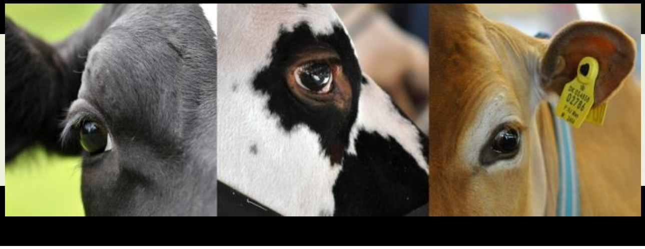


250 causative mutations

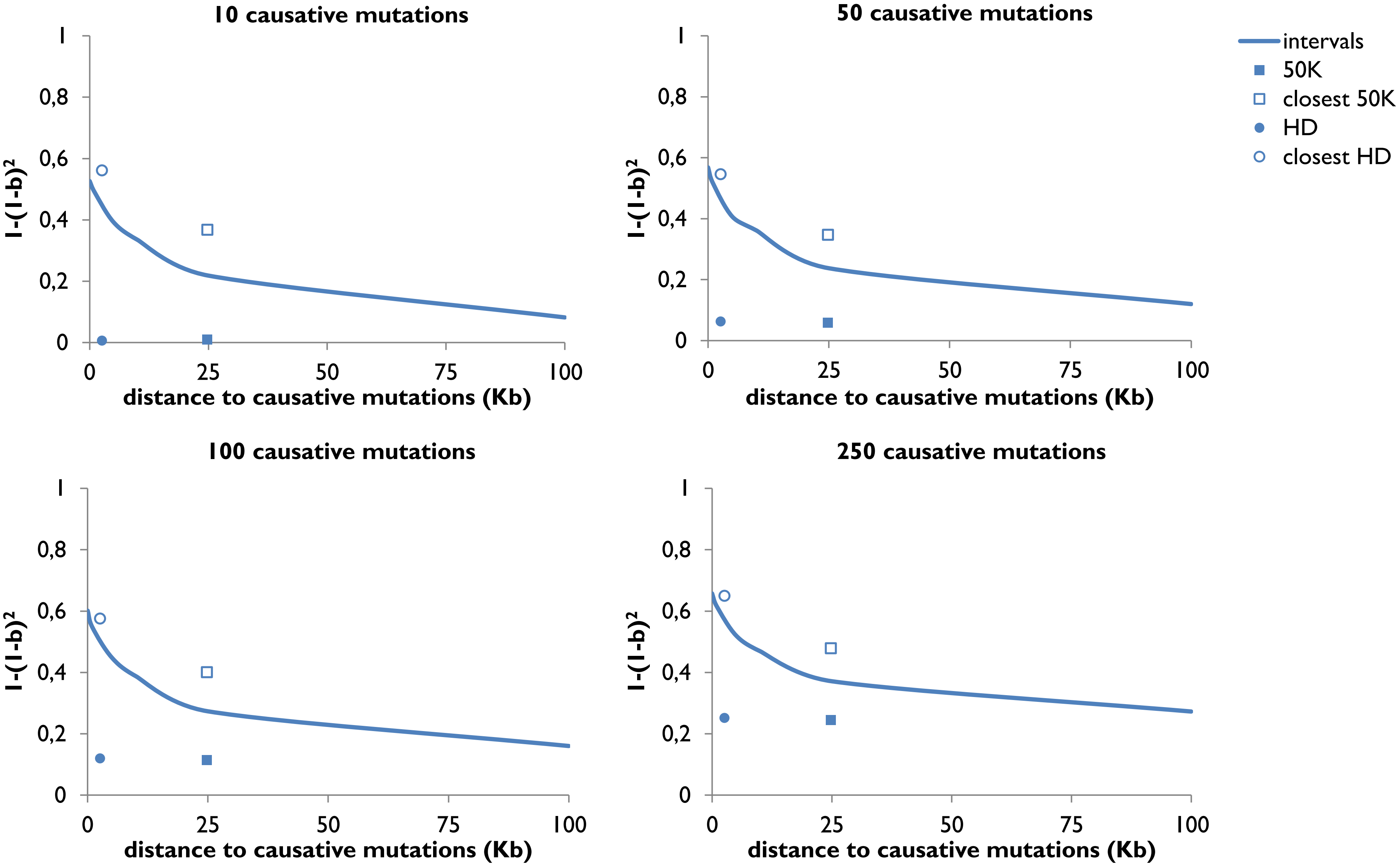


# *Results – Number of mutations (within breed)*





# Results – Number of mutations *(across breed)*



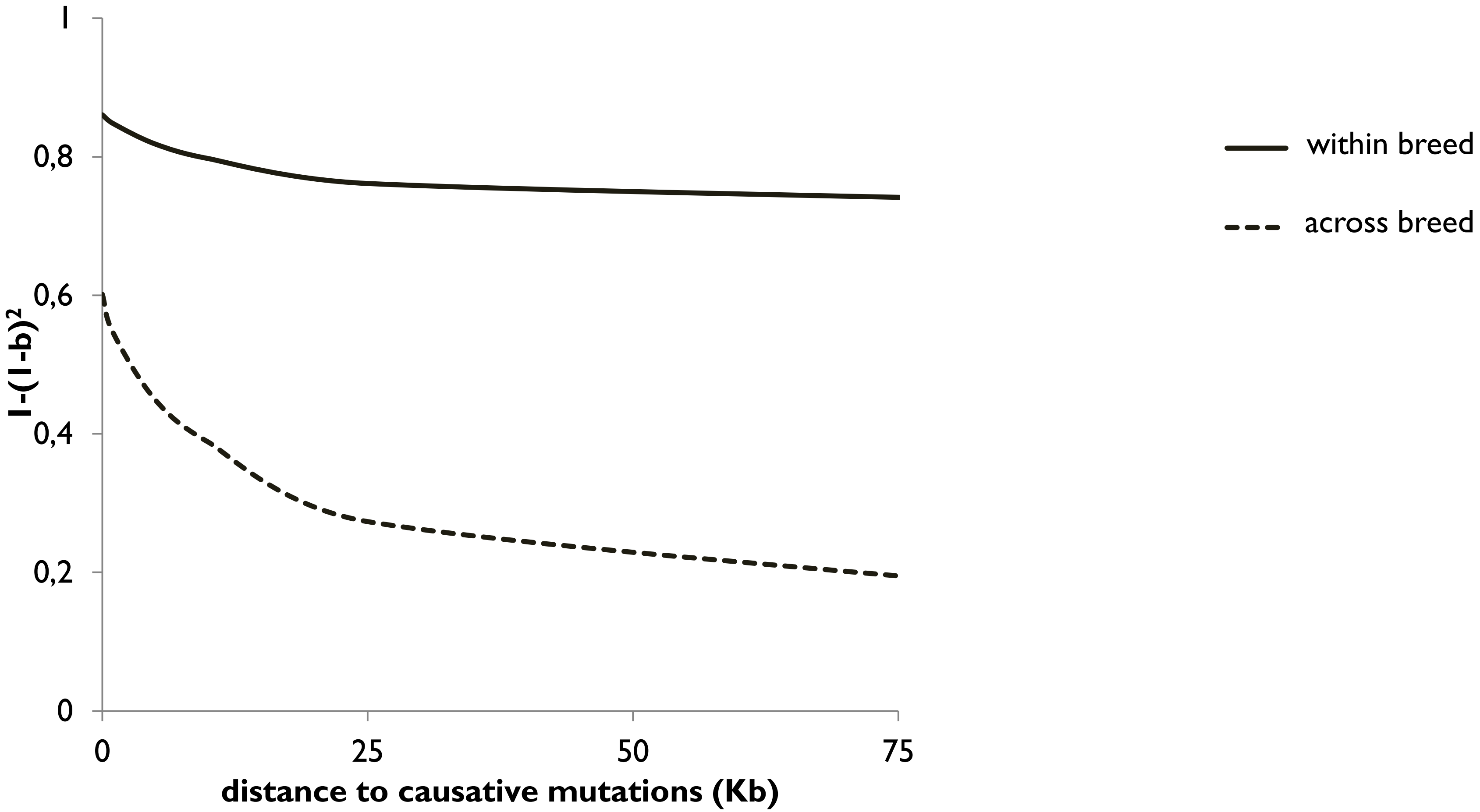
10 causative mutations

50 causative mutations

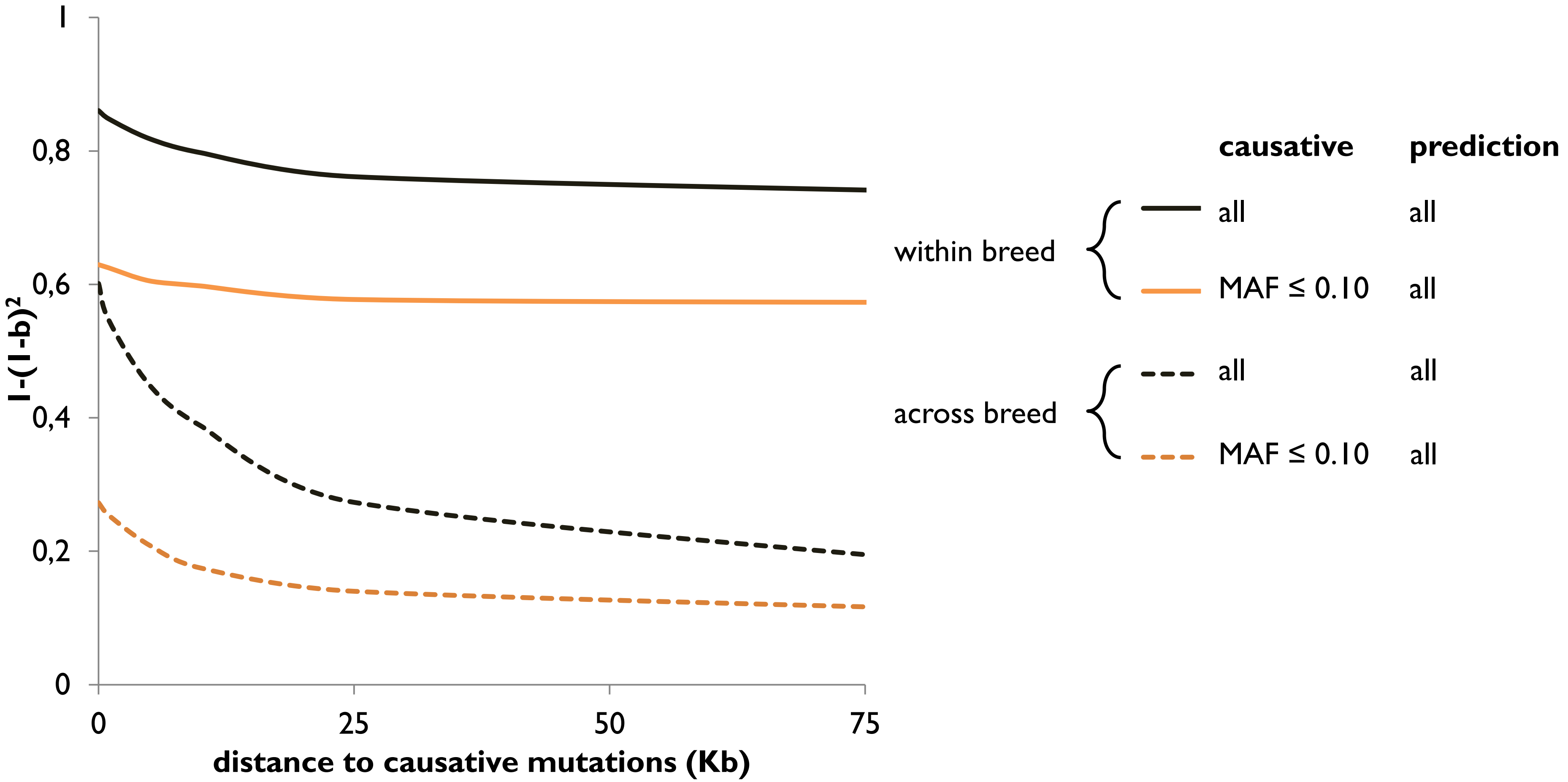
100 causative mutations

250 causative mutations

# *Results – MAF (100 causative mutations)*

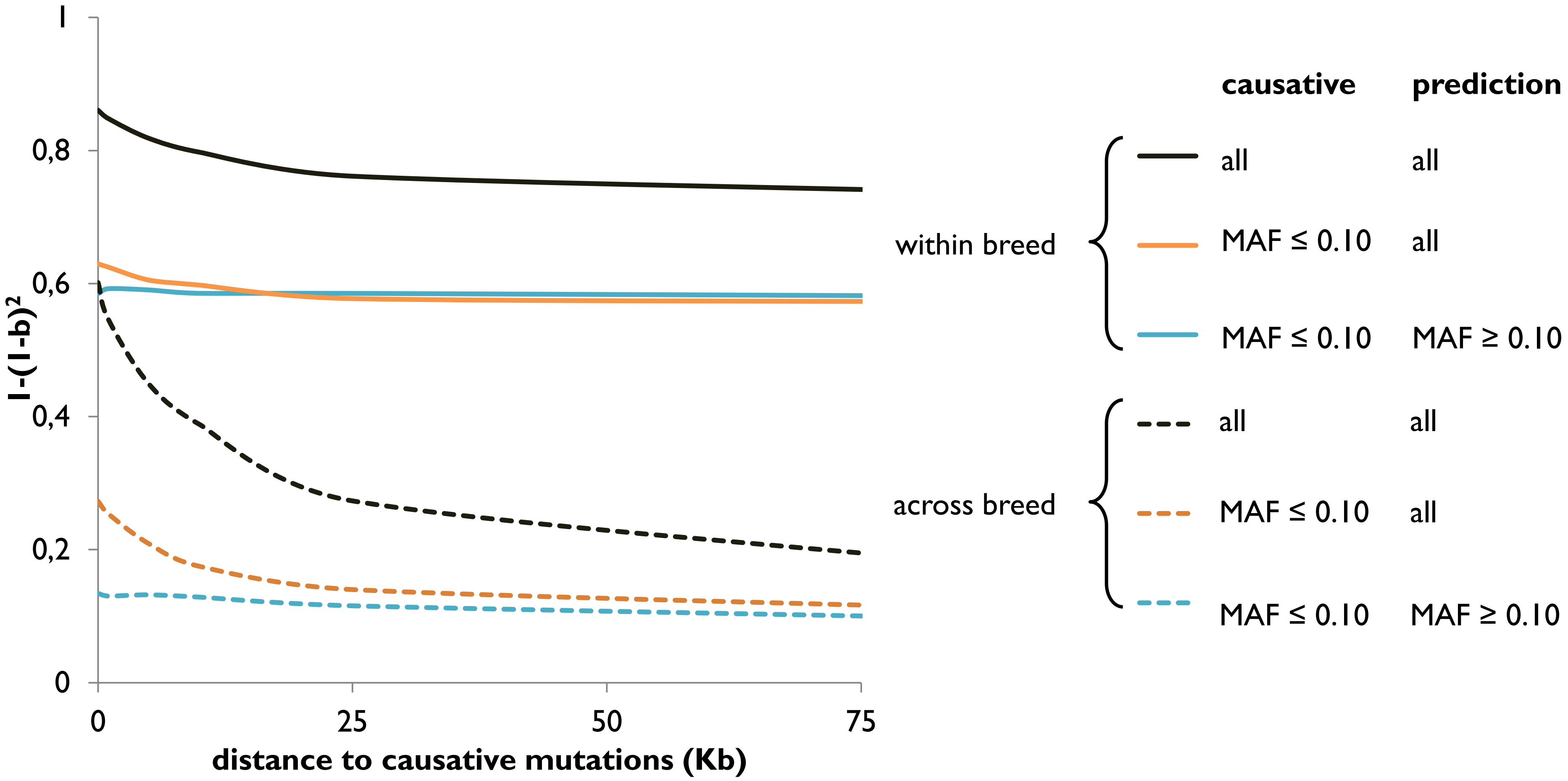


# *Results – MAF (100 causative mutations)*

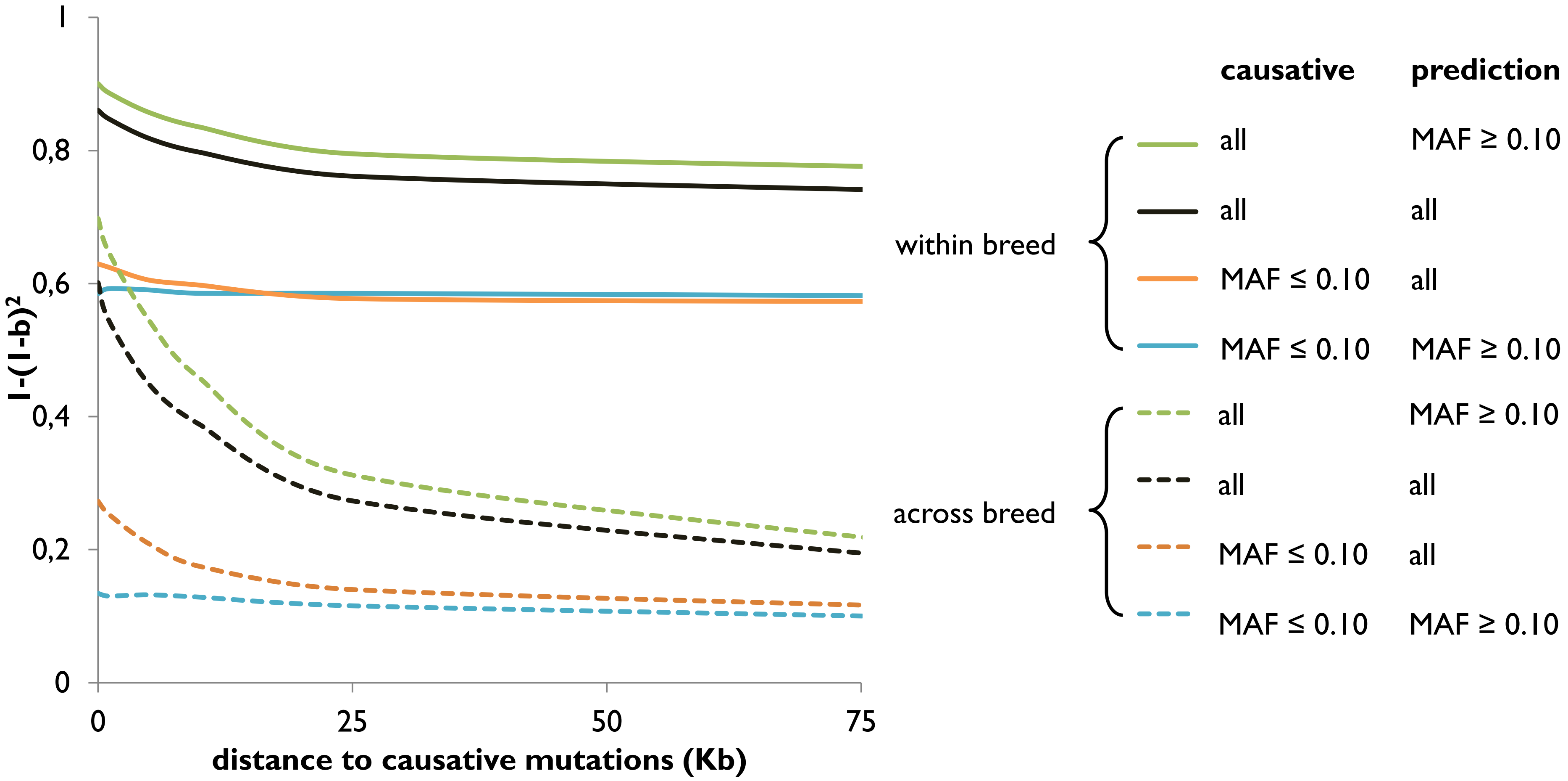


| | | causative | prediction |
|---|---|---|---|
| within breed | ━━ | all | all |
| | ━━ | MAF ≤ 0.10 | all |
| across breed | ╍╍ | all | all |
| | ╍╍ | MAF ≤ 0.10 | all |

# Results – MAF *(100 causative mutations)*



# *Results – MAF (100 causative mutations)*



# Conclusions

- Use of sequence data can improve prediction $R^2$
- Not by increasing density, but by selecting the right variants
- Larger improvement across breed than within breed
- More improvement with lower number of causative mutations
- Inclusion of rare variants only improves prediction if they are (in high LD with) causative mutations