Single step evaluations using haplotype segments

M. L. Makgahlela, T. Knürr, G. P. Aamand, I. Strandén

& E. A. Mäntysaari

NAV Nordic Cattle Genetic Evaluation

STØTTET AF
mælkeafgiftsfonden

Introduction

- Genomic evaluations, as originally proposed, were based on haplotype segments, which are;
  - ➢ closely located allele combinations that tend to be jointly inherited
- Many current evaluations however, use large number of SNP markers in models that are;
  - ➢ simplified and less computationally demanding

Introduction

- If the observed reliabilities are low, haplo-block models may improve evaluations

1) They were found to be more reliable than single markers

   - Because ancestral haplotypes may capture greater linkage disequilibrium (LD) with QTL than single markers

2) They could greatly reduce the number of markers for genomic evaluations

3) There are many free haplotyping software available

Objectives

- Examine the reliability of single step with genomic relationship matrix (**G**) constructed using haplotype segments in the Nordic Red dairy cattle (RDC)

- Compare the haplo-block model with standard single-step GBLUP

Data provided by NAV

- **Genotypes**
  - After editing, there were 38,194 informative SNPs available for 4,727 bulls born between 1971-2008
- **Phenotypes**
  - Deregressed Proofs (DRP) of cows for milk, protein and fat
    - Full data (DRP_F) ➔ 3,633,481 cows
    - Reduced data (DRP_R) i.e., discard cows born after > 2005 ➔ 3,146,448 cows
- Full RDC pedigree (n=4,873,703)

For validation

- ApaX in Mix99 program was used for calculating EDCs
- 2 runs of animal model were used to solve deregressed bull EBVs as follows;
  - 1st full run ➔ with DRP_F ➔ generate DRP for 519 validation bulls born between 2002-2008 with EDC>=20
  - 2nd reduced run ➔ with DRP_R ➔ daughters of 4,208 training bulls born between 1971-2005

Construction of Haplotype blocks

1) BayesB fitting joint estimation of SNP effects in multilocus model
2) Rank SNPs by the absolute effect $\widehat{\boldsymbol{\beta}}_g$
3) Haplotype (phase) genotypes using Beagle software
4) Construct **5-SNP** haplotypes (i.e., 2 SNPs before and after the one with the highest $\hat{\beta}_g$)
5) Estimate haplotype variances
6) Number of haplotype segments ➔ 750 and 1500

Single step model

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} \mathbf{G}w^{-1} - \mathbf{A}_{22}{}^{-1} & 0 \\ 0 & 0 \end{bmatrix}, \text{ where}$$

- $\mathbf{A}^{-1}$ includes all animals and $\mathbf{A}_{22}{}^{-1}$ is a sub-matrix for genotyped bulls
- $\mathbf{G}w = (1 - w)\mathbf{G}k + w\mathbf{A}_{22}$

✓ $k = \frac{\text{trace}\mathbf{A}_{ii_{22}}}{\text{trace}\mathbf{G}_{ii}}$; w values were varied at 0.10, 0.20 or 0.40

Single step model

- **Haplo-block G**

$$\mathbf{G} = \mathbf{ZDZ}' \;; \quad \mathbf{Z}_{i,j} \leftarrow (0 - 2p_j); (1 - 2p_j); (2 - 2p_j),$$

  - 0,1 or 2 is the number of 2nd allele
  - p_j is the frequency for the 2nd allele
  - **D** is a diagonal of the estimate of haplotype variances
  - Haplo-block **G** was constructed with segments length 750 (HAP750) and 1500 (HAP1500)

- **Regular SNP-based G:**

$$\mathbf{G} = \mathbf{ZZ}' / \sum 2pq$$

GEBV evaluation

$$\mathrm{DRP}_{\mathrm{R_{cow}}} = 1_{\mathrm{n}}\mu + \mathbf{Z}\mathrm{a} + \mathrm{e},$$

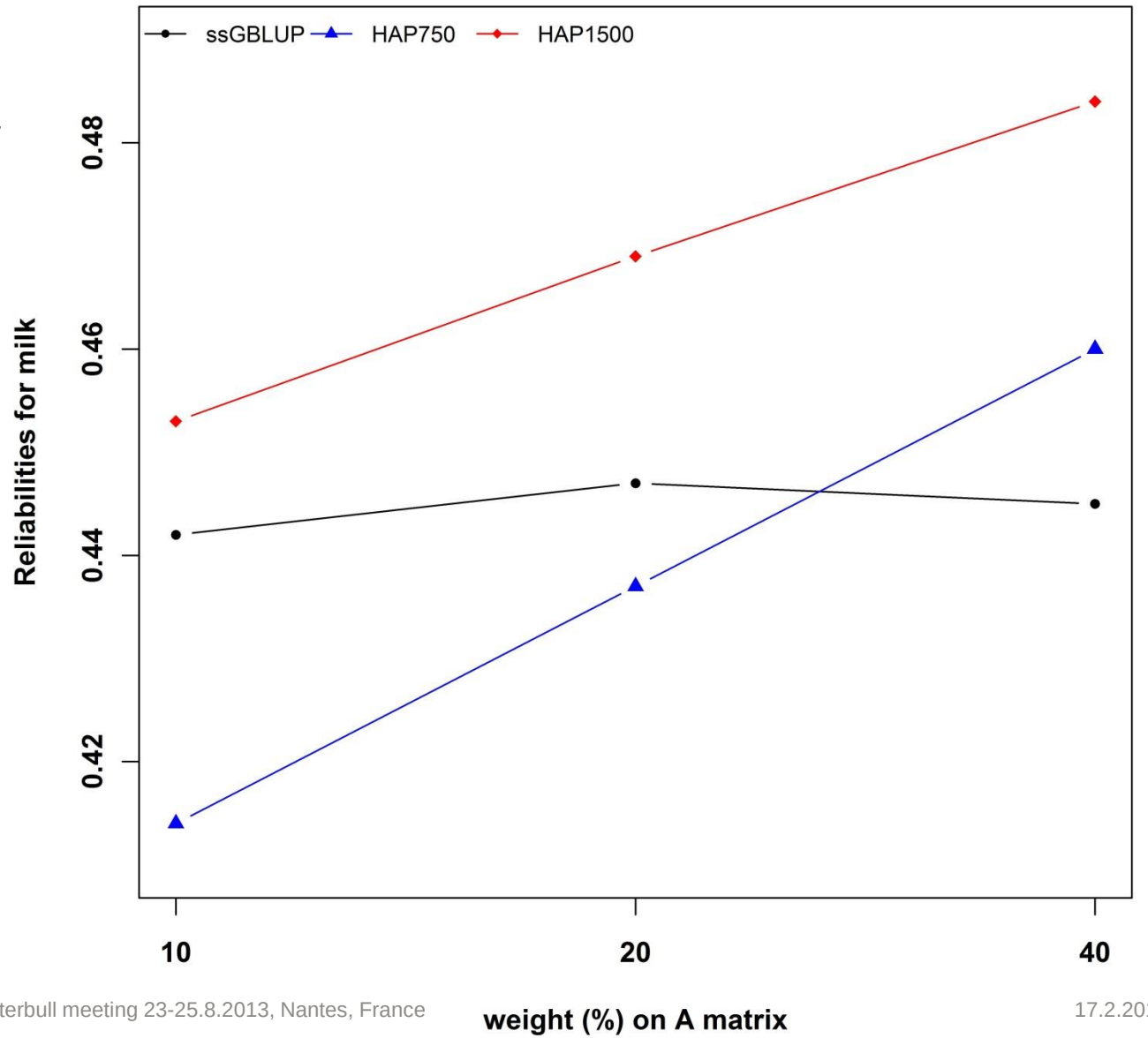
- where:
  - ✓ $\mathrm{var}(\mathbf{a}) = \mathbf{H}\sigma 2_{\mathrm{a}}$ with variances from NAV routine evaluations
  - ✓ $\mathrm{DRP}_{\mathrm{R_{cow}}}$ is the deregressed proof of the daughter of training bulls in the reduced data
  - ✓ Reliability of DRP was used as weight

GEBV validation

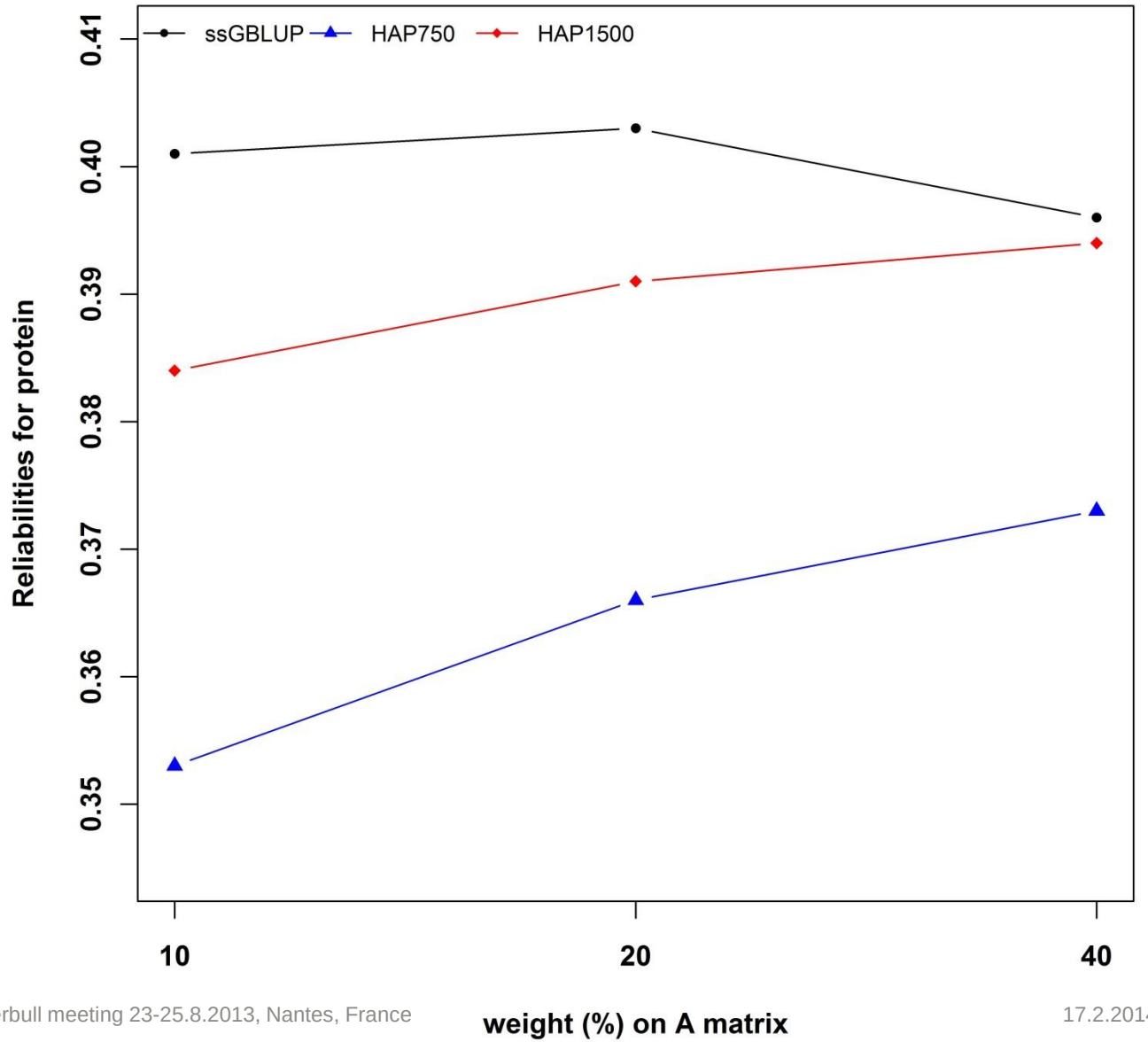
$$DRP_{F_{bull}} = b_0 + b_1 GEBV + e,$$

- where:
  - ✓ $DRP_{F_{cow}}$ is the deregressed proof of the candidate from the full data run
  - ✓ Reliability of DRP was used as weight

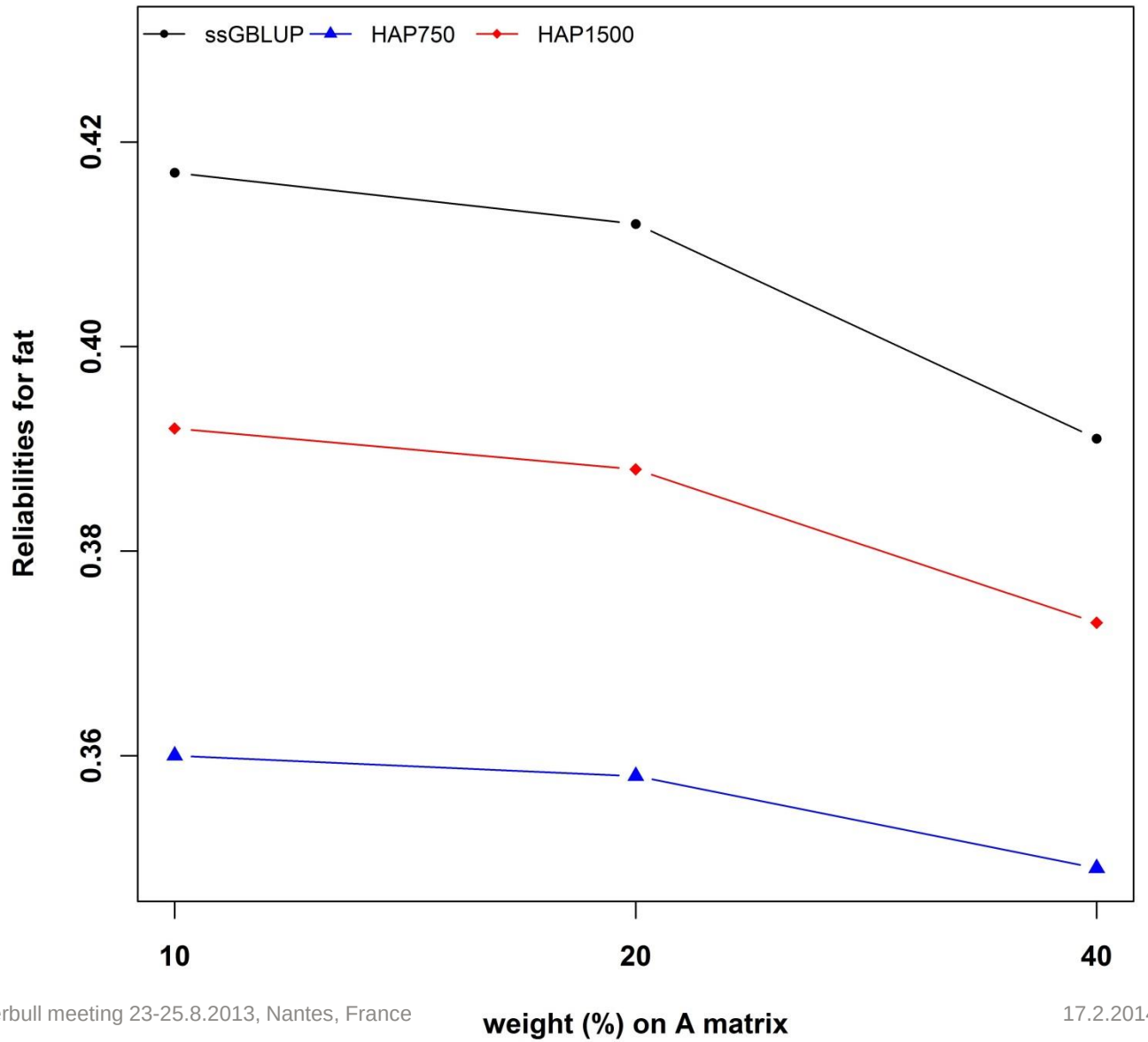
Validation reliabilities for milk



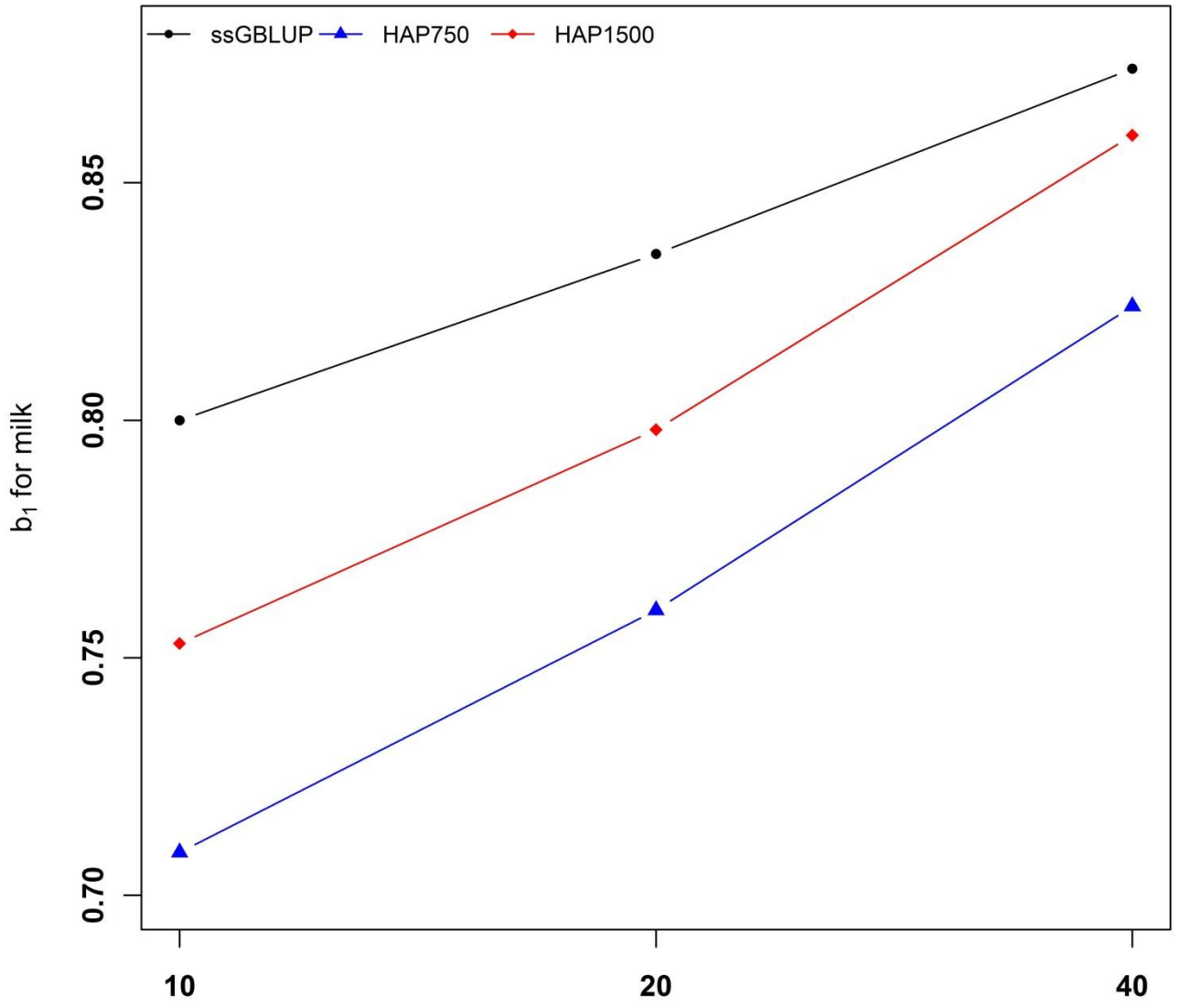
Validation reliabilities for protein



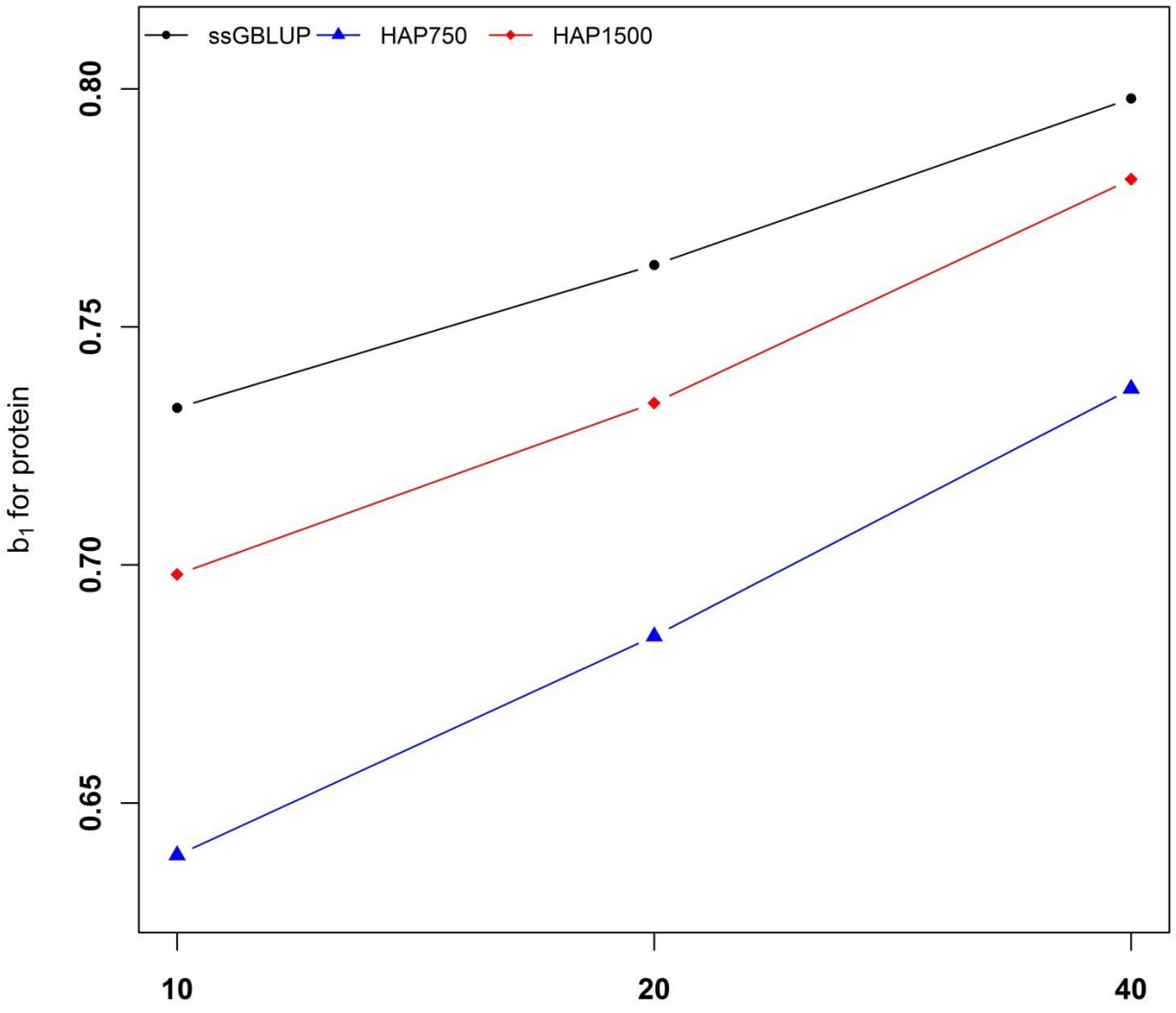
Validation reliabilities for fat



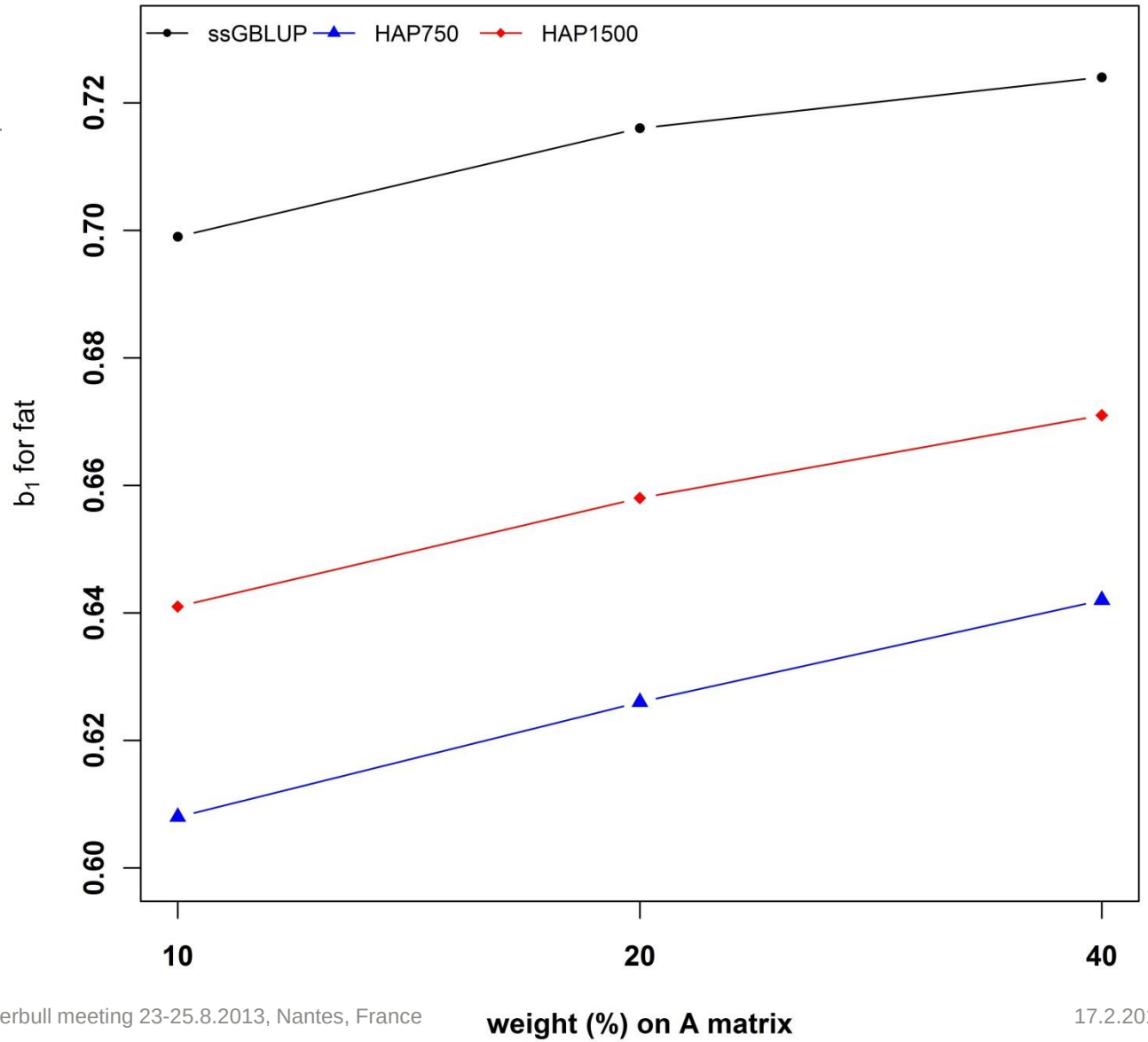
Inflation for milk



Inflation for protein



Inflation for fat



Validation reliabilities of GEBV

Method	Milk	Protein	Fat
w**A**=0.1			
ssGBLUP	0.442	**0.401**	**0.417**
HAP750	0.414	0.353	0.360
HAP1500	**0.453**	0.384	0.392
w**A**=0.2			
ssGBLUP	0.447	**0.403**	**0.412**
HAP750	0.437	0.366	0.358
HAP1500	**0.469**	0.391	0.388
w**A**=0.2			
ssGBLUP	0.445	**0.396**	**0.391**
HAP750	0.460	0.373	0.349
HAP1500	**0.484**	**0.394**	0.373

Inflation of GEBV

Method	Milk	Protein	Fat
w**A**=0.1			
ssGBLUP	**0.800**	**0.733**	**0.699**
HAP750	0.709	0.639	0.608
HAP1500	0.753	0.698	0.641
w**A**=0.2			
ssGBLUP	0.835	0.763	0.716
HAP750	0.760	0.685	0.626
HAP1500	0.798	0.734	0.658
w**A**=0.2			
ssGBLUP	**0.874**	**0.798**	**0.724**
HAP750	0.824	0.737	0.642
HAP1500	**0.860**	0.781	0.671

Conclusions

- The validation reliability for milk was clearly increased when using more haplotype segments ➔ HAP1500
  - ➢ 1, 2 and 4 % when the weight on **A** was 0.1, 0.2 and 0.4, respectively
- Reliability for milk with HAP750 was increased by 2% when the weight on **A** was 40%
- These improvements however, were not achieved for protein and fat as reliabilities were low
- Reliabilities of haplo-block models for milk and protein tended to increase with increasing weight on **A** but the opposite was true for fat

Conclusions

- For all traits, the inflation levels of GEBV were greater with haplo-block models
  - ➢ In all cases, inflation intervals with standard single step reduced as the amount of pedigree increased
- The use of haplotype segments appeared to be very promising provided there is balance between the number of haplotypes and optimal scaling with pedigree information

THANK YOU !!!