vit Service & Daten aus einer Quelle

MARTIN-LUTHER-UNIVERSITÄT HALLE-WITTENBERG

Large scale genotype imputing for non-genotyped relatives in German Holsteins

H. Alkhoder, Z. Liu*, F. Reinhardt*, H. H. Swalve**, and R. Reents**

**) Vereinigte Informationssysteme Tierhaltung w. V. (vit), Verden, Germany*
***) Martin-Luther-Universität (MLU), Halle, Germany*

Objectives

- Estimation of genotypes for non-genotyped animals was used in different studies to enhance genomic selection (e.g. Gengler et al., 2007; Pimentel et al., 2013)

- Estimation of genotypes for non-genotyped animals using information of genotyped relatives

- Imputation of genotypes using Fimput software for non-genotyped animals based on estimated SNPs

- Validation of methods and results
  - Accuracy of estimated and imputed genotypes
  - DGV correlations using original and imputed genotypes

Data

- Data base: German Holstein genomic evaluation April 2013

  74,050 genotyped animals

  230,831 animals in pedigree

  45,613 SNPs used for DGV calculation

- Validation study

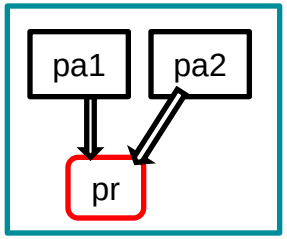
  Genotypes of 2,500 dams were masked

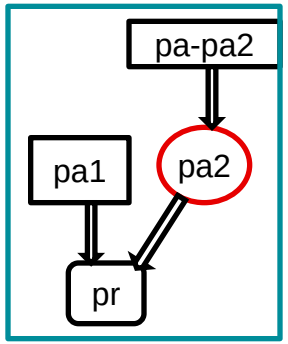
  DGV for all 44 traits

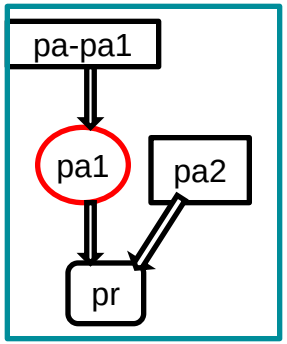
Frequency of non-genotyped relatives

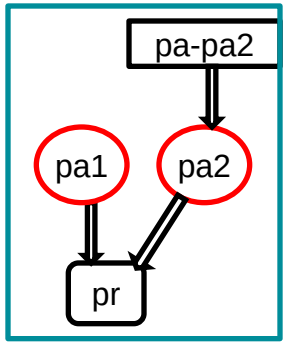
Genotyped relatives	Non-genotyped animals	N records	N unique animals
progeny, mate and sire	cow	45,609	25,000
	bull	161	72
progeny, sire	cow	2,256	1,751
	bull	1,856	461
sire and dam	animal	378	378
one of ancestors	animal	23,185	23,185

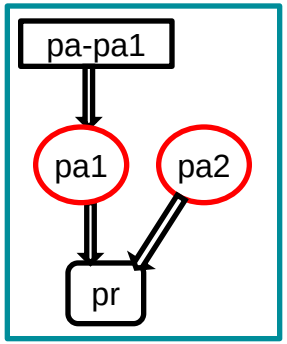
(Non-genotyped animals born > 1997)

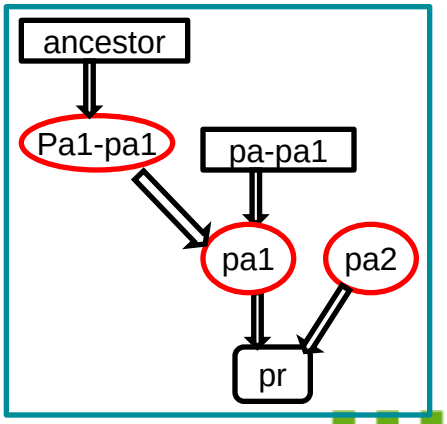






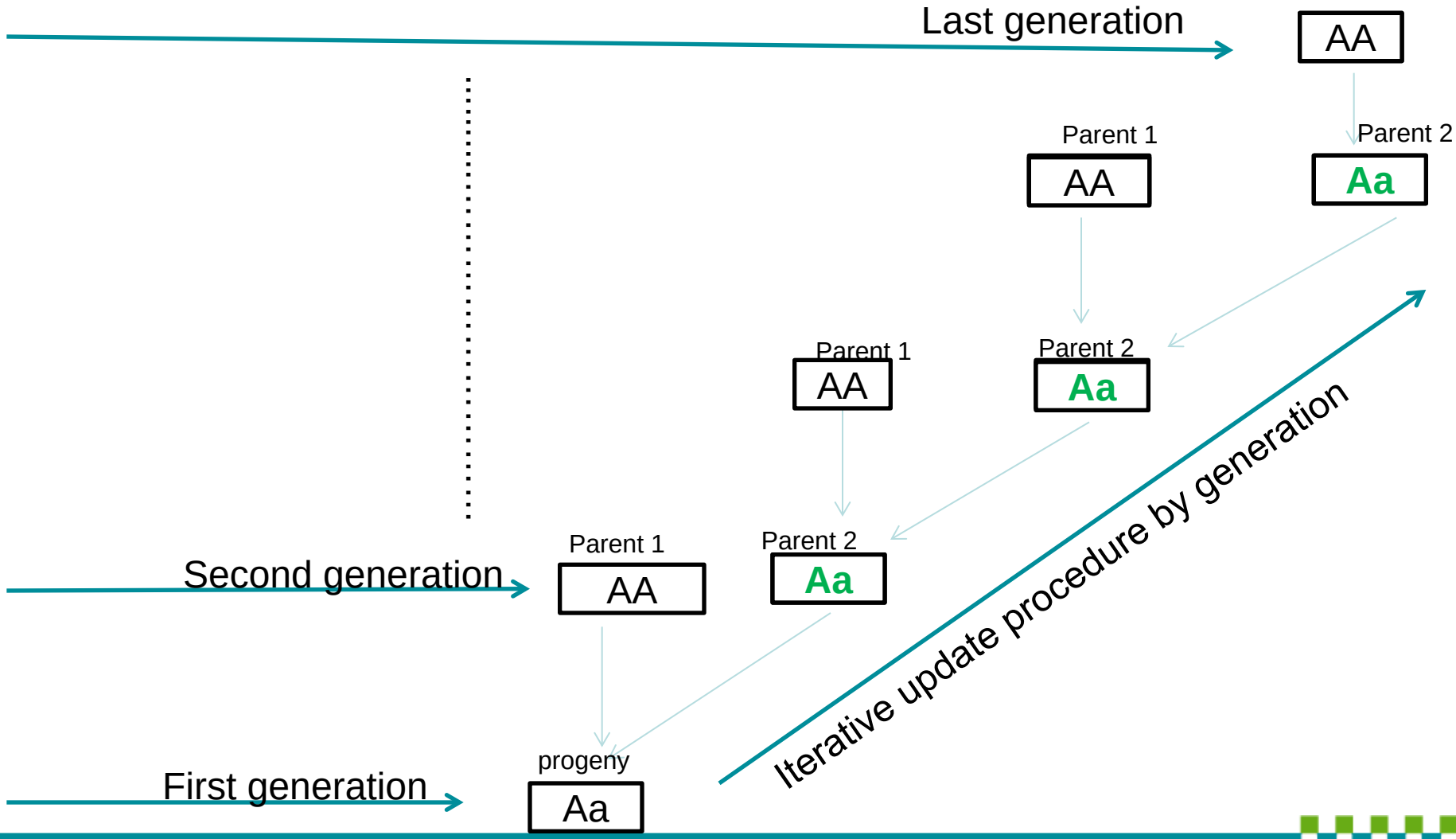






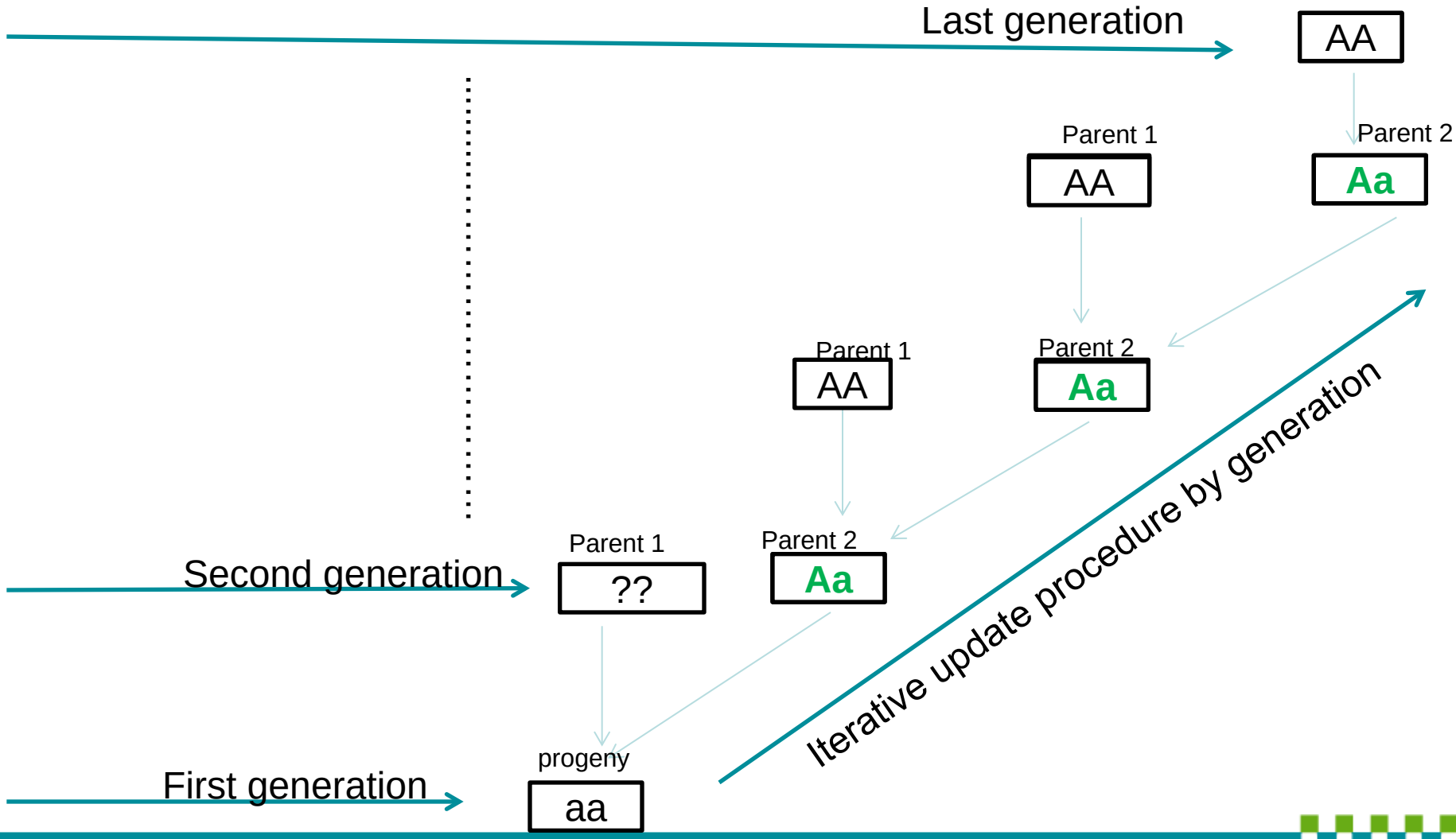
Methods: M1 bottom up

Estimation of SNP genotypes with 100% probability (progeny to ancestors) using three genotyped relatives similar to Pimentel et al., (2013)



Methods: M1 bottom up

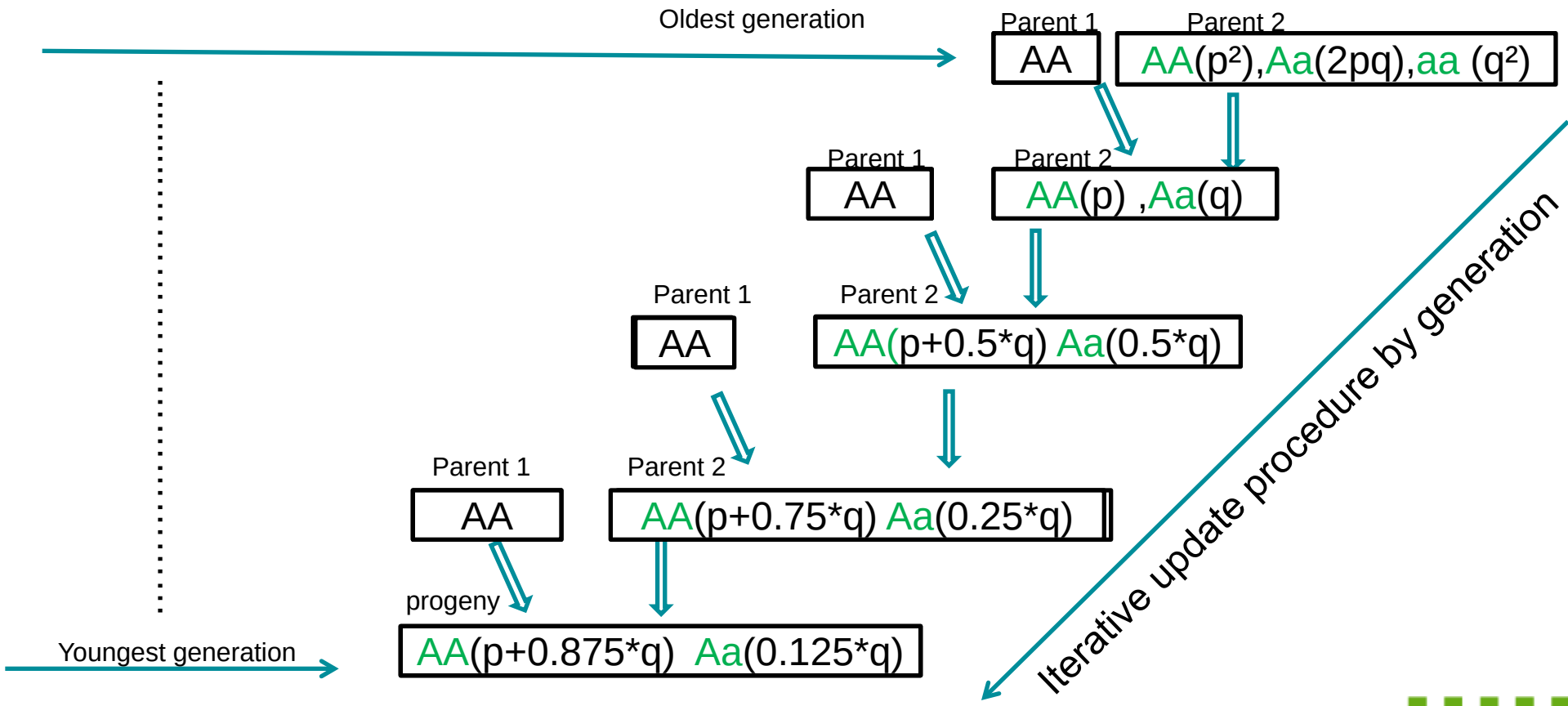
Estimation of SNP genotypes with 100% probability (progeny to ancestors) using two genotyped relatives



Methods: M2 top down

Estimation of SNP genotypes with probability < 1

(ancestors to progeny) similar to Israel and Weller (1998)



Methods: M3 BLUP

Estimation of genotype for non-genotyped animals with a BLUP model (Gene Content method: Gengler et al., 2007)

$Y = \mu I + Zb + e$

- Y is a vector of genotype (0,1 and 2)
- μ is the overall mean
- b is vector of estimated SNP genotype
- Z is incidence matrix for b
- $Z`Z+A^{-1}$ *Æ
- $h^2 = 0.995$

Methods M1, M2 and M3 are conducted on a SNP-by-SNP basis without exploring LD information between SNPs.

Results: Proportion of estimated SNPs and alleles correctly estimated with M1 + M2 + M3

1- All estimated SNPs from M1 (bottom up) were used

2- SNPs from M2 (top down) were selected using different probabilities (e0.99, e0.98, e0.97, e0.95)

3- SNPs from M3 (BLUP) were converted to genotype, if the estimated gene content (EGC_{snp})in the interval (0, 1and 2) ± Diff. where Diff. = 0.05, 0.10, 0.20 and 0.30

M1+M2+M3		Genotype probability level for M2 top down							
		Number of estimated SNPs				Alleles correctly estimated			
		Prob. e0.99	Prob. e0.98	Prob. e0.97	Prob. e0.95	Prob. e0.99	Prob. e0.98	Prob. e0.97	Prob. e0.95
Genotype interval for M3	EGC_{snp} ±0.05	**15,873**	**16,603**	**17,105**	**18,017**	0.988	0.988	0.988	0.987
	EGC_{snp} ±0.10	**19,796**	**20,252**	**20,617**	**21,301**	0.980	0.981	0.980	0.980
	EGC_{snp} ±0.20	**26,319**	**26,638**	**26,866**	**27,277**	0.967	0.967	0.967	0.967
	EGC_{snp} ±0.30	**32,340**	**32,522**	**32,613**	**32,841**	0.953	0.953	0.953	0.953
M1+ M2		**9,898**	**11,540**	**12,452**	**14,049**	0.999	0.998	0.997	0.995

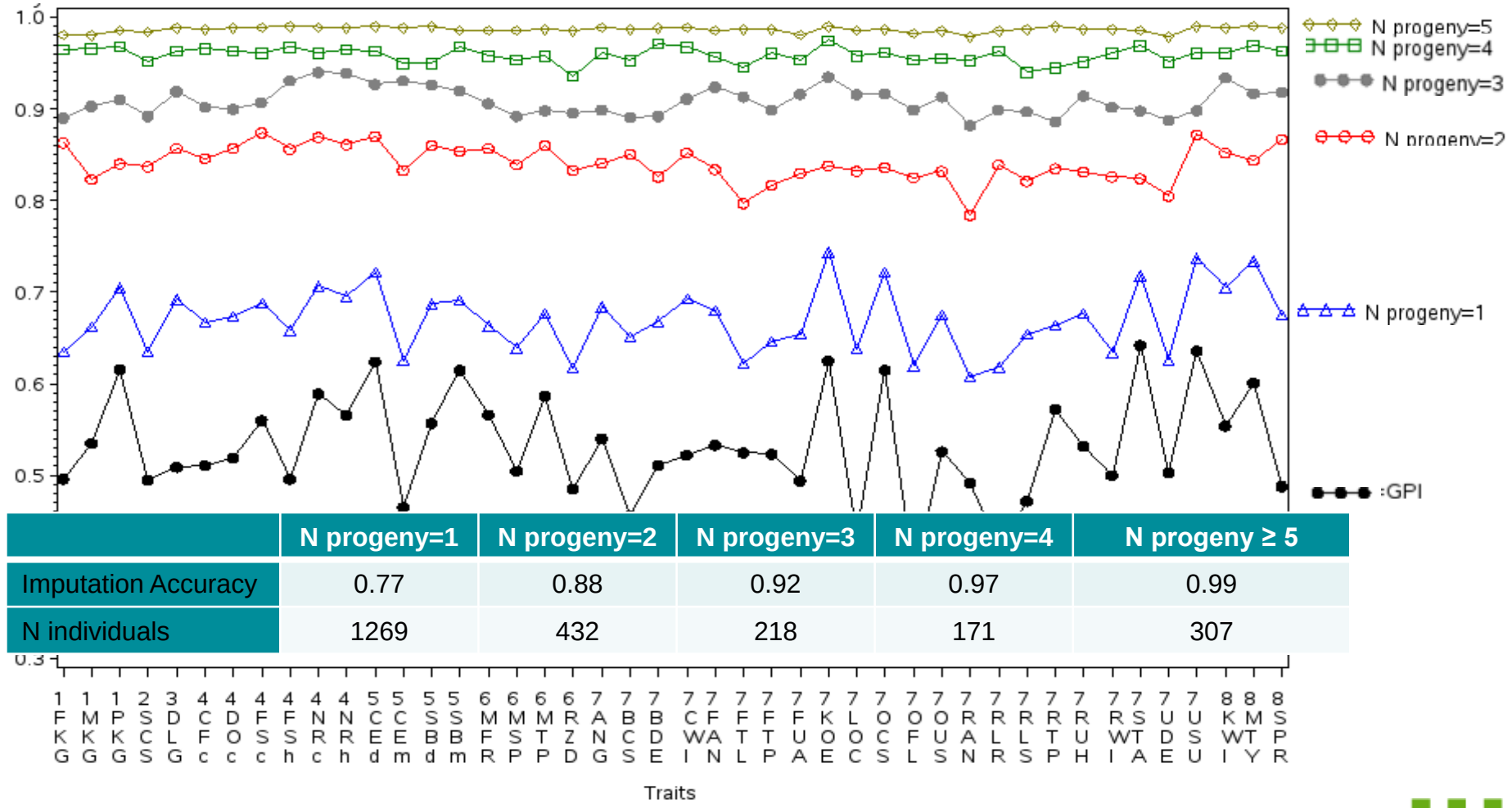
Results: Accuracy of estimated and imputed genotypes for all methods

Imputation using Fimpute was performed using the estimated SNP genotypes with all methods

Methods	Estimated SNPs			Imputation	
	N estimated SNPs	Genotype correctly (%)	N correctly estimated SNPs	Genotype correctly (%)	N correctly estimated and imputed SNPs
M1 bottom up	4,789	99.7	4,774	81.1	36,992
M2 top down (gene probability ≥ 0.99)	5,145	99.2	5,113	73.7	33,616
M3 BLUP (interval 0.05)	10,308	96.7	9,968	81.5	37,174
Joint M1 + M2	9,898	99.6	9,858	81.3	37,083
Joint M1 + M2 + M3	15,873	97.7	15,508	84.3	38,451

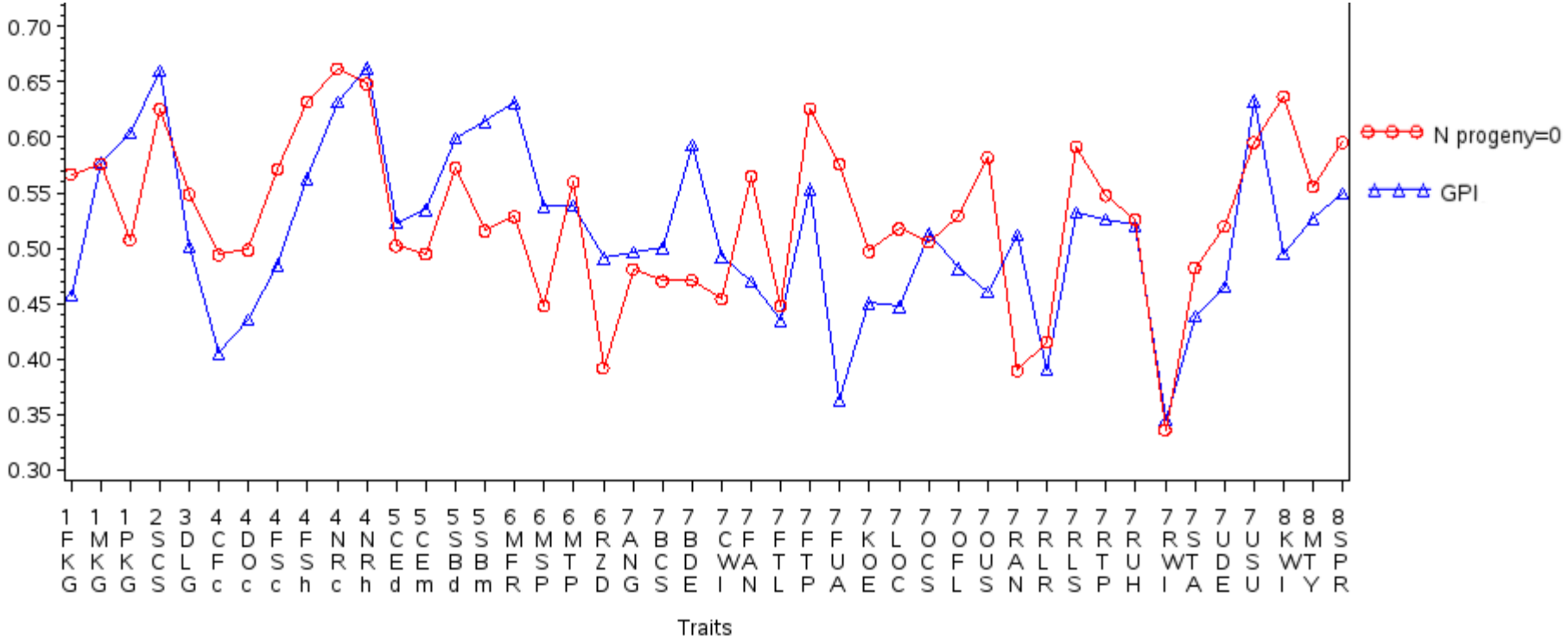
Results: Correlations of DGV using original and imputed genotypes and genomic pedigree index for all 44 traits

Scenario M1+ M2 + M3 + imputation



	N progeny=1	N progeny=2	N progeny=3	N progeny=4	N progeny ≥ 5
Imputation Accuracy	0.77	0.88	0.92	0.97	0.99
N individuals	1269	432	218	171	307

Results: Correlations of DGV using original genotypes and genomic pedigree index for animals without progeny



Average correlations across all traits:

0.52 for genomic pedigree index

0.53 for imputed genotypes

	Animals without progeny
Imputation Accuracy	0.71

Conclusions I

- Genotypes of non-genotyped relatives can be accurately estimated
  - 97% accuracy for dams > 3 genotyped progeny

- Combining the three SNP-by-SNP methods led to highest number of reliable estimated SNPs (34% of all SNPs with 98% accuracy)
  - M3 BLUP (23%) > M2 top down (11%) / M1 bottom up (10%)
  - M2 estimate SNPs with extreme allele frequency
  - M1 estimate SNPs with intermediate allele frequency
  - 23 % SNPs estimated with accuracy of 96.7% using M3 BLUP with a genotype estimate standard error (0, 1 and 2) ± 0.05 ➔ Determining of genotype can be improved by using average gene content at a given population allele frequency as additional information

- Imputing on top of estimated genotypes jointly with the three methods (M1+M2+M3) gave highest accuracy: 97% , >3 progeny
  - M2 + imputing had the lowest accuracy
  - BLUP + imputing H M1 + imputing

Conclusions II

- DGV correlations between original & imputed genotypes higher than 0.95 (for parents with >3 progeny)
- For young animals without progeny the methods did not perform better than using genomic pedigree index
- Our methods are applicable to general pedigree structures
- Saves genotyping costs of reference population for new traits (i.e. health traits)
- Supplies genotypes for non genotyped old animals without DNA samples available

IT-Solutions for Animal Production

Number of estimated SNPs using M1 (bottom up) by groups of non-genotyped relatives

156,781 non-genotyped relatives in pedigree

Some SNPs for 53,727 relative animals were estimated

Genotyped relatives	Non-genotyped animals	N all individual	N estimated individuals	μ N-SNPs	Min. N-SNPs	Max. N-SNPs
progeny, mate and sire	cow	30,604	29,999	4,163	1,463	20,854
	bull	73	73	6,372	2,540	11,058
progeny, sire	cow	2,432	2,363	3,392	2,135	12,481
	bull	719	690	3,433	2,306	20,854
sire and dam	cow	375	375	24,347	20,054	31,852
	bull	3	3	21,736	21,413	22,057
one of ancestors	Cow	50,688	20,099	640	3	13,353
	bull	6,717	2	1,323	1,230	1,417

Proportion of estimated SNPs and allele correctly estimated with different combination of M2 and M3

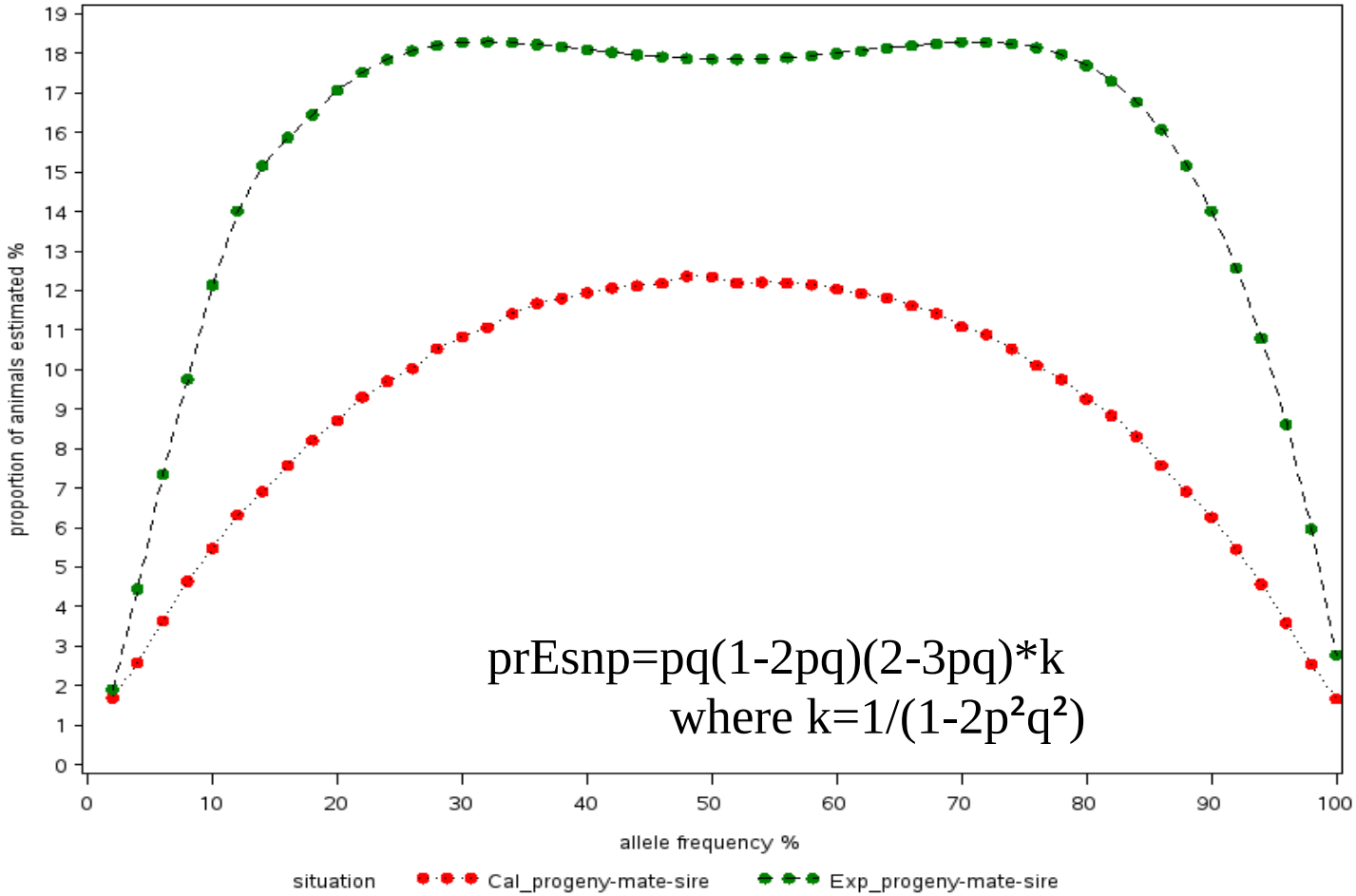
1- All estimated SNPs from M1 (bottom to up) were used

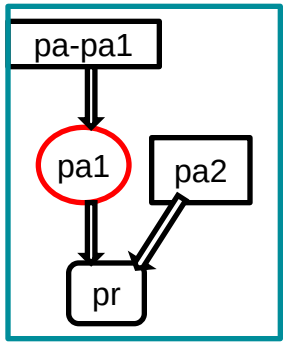
2- The estimated SNPs from M2 (top to down) were selected after different probability (Prob.1e0.99, Prob.1e0.98, Prob.3e0.97 and Prob.4e0.95)

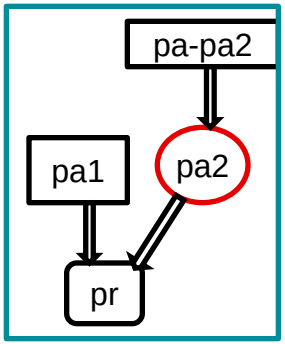
3- The estimated SNPs from M3 (BLUP) were converted to genotype If EBV_{snp} egenotype(0, 1and 2) ± Diff.

Where Diff. = 0.05, 0.10, 0.12, 0.16, 0.20, 0.25 and 0.30

		Proportion of estimated SNPs				Allele correctly estimated			
		Genotype probability level for M2				Genotype probability level for M2			
		Prob. e0.99	Prob. e0.98	Prob. e0.97	Prob. e0.95	Prob. e0.99	Prob. e0.98	Prob. e0.97	Prob. e0.95
Genotype interval for M3	EBVSNP±0.05	0.348	0.364	0.375	0.395	0.988	0.988	0.988	0.987
	EBVSNP±0.10	0.434	0.444	0.452	0.467	0.980	0.981	0.980	0.980
	EBVSNP±0.12	0.492	0.502	0.508	0.521	0.975	0.975	0.975	0.975
	EBVSNP±0.16	0.521	0.530	0.536	0.547	0.972	0.973	0.973	0.972
	EBVSNP±0.20	0.577	0.584	0.589	0.598	0.967	0.967	0.967	0.967
	EBVSNP±0.25	0.646	0.650	0.654	0.660	0.961	0.961	0.961	0.960
	EBVSNP±0.30	0.709	0.713	0.715	0.720	0.953	0.953	0.953	0.953
M1M2		**0.217**	**0.253**	**0.273**	**0.308**	**0.999**	**0.998**	**0.997**	**0.995**

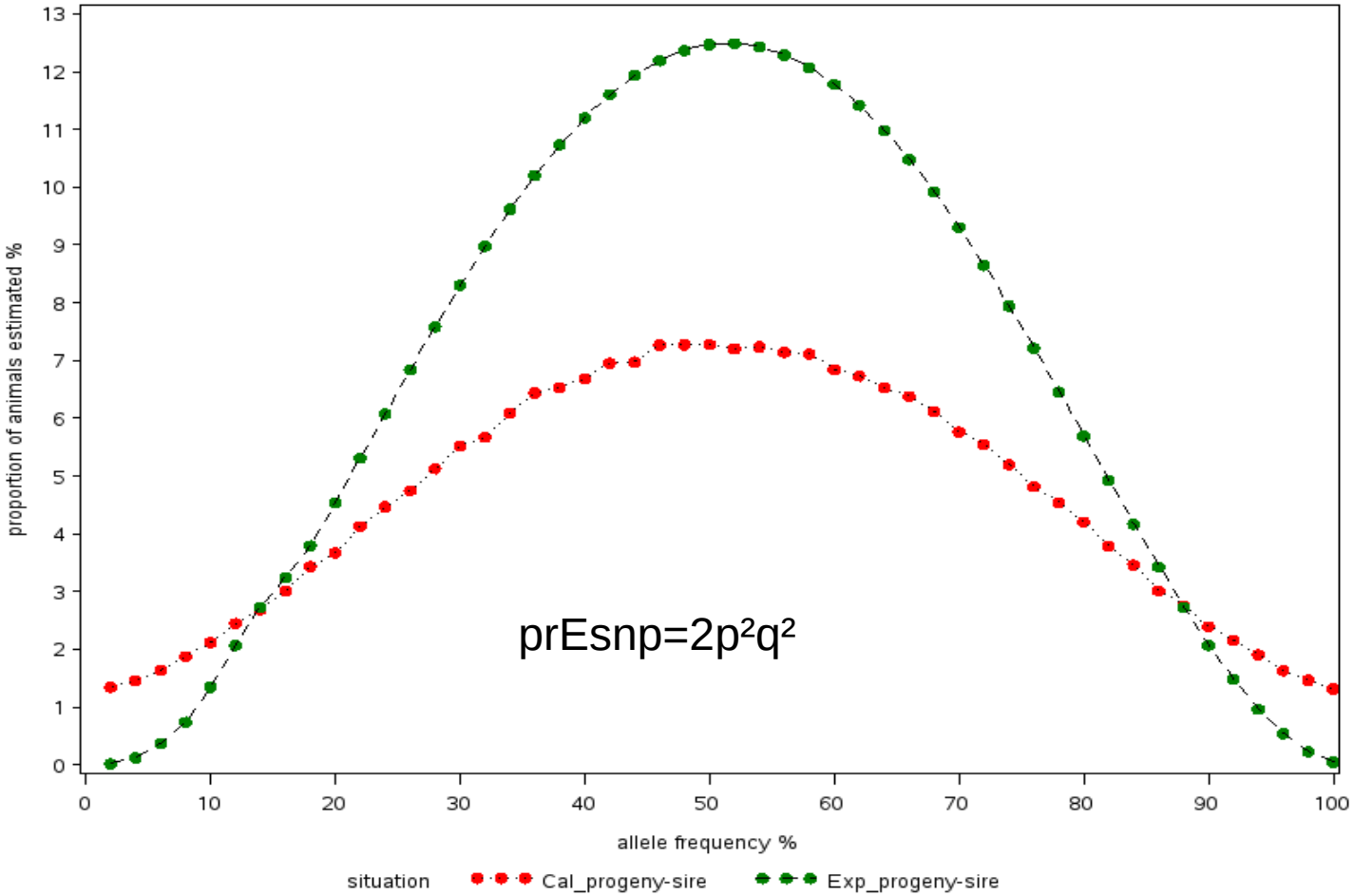


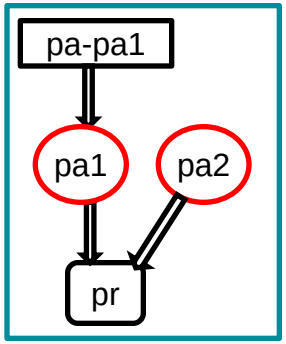




Average of expected proportion for estimated genotype 15%

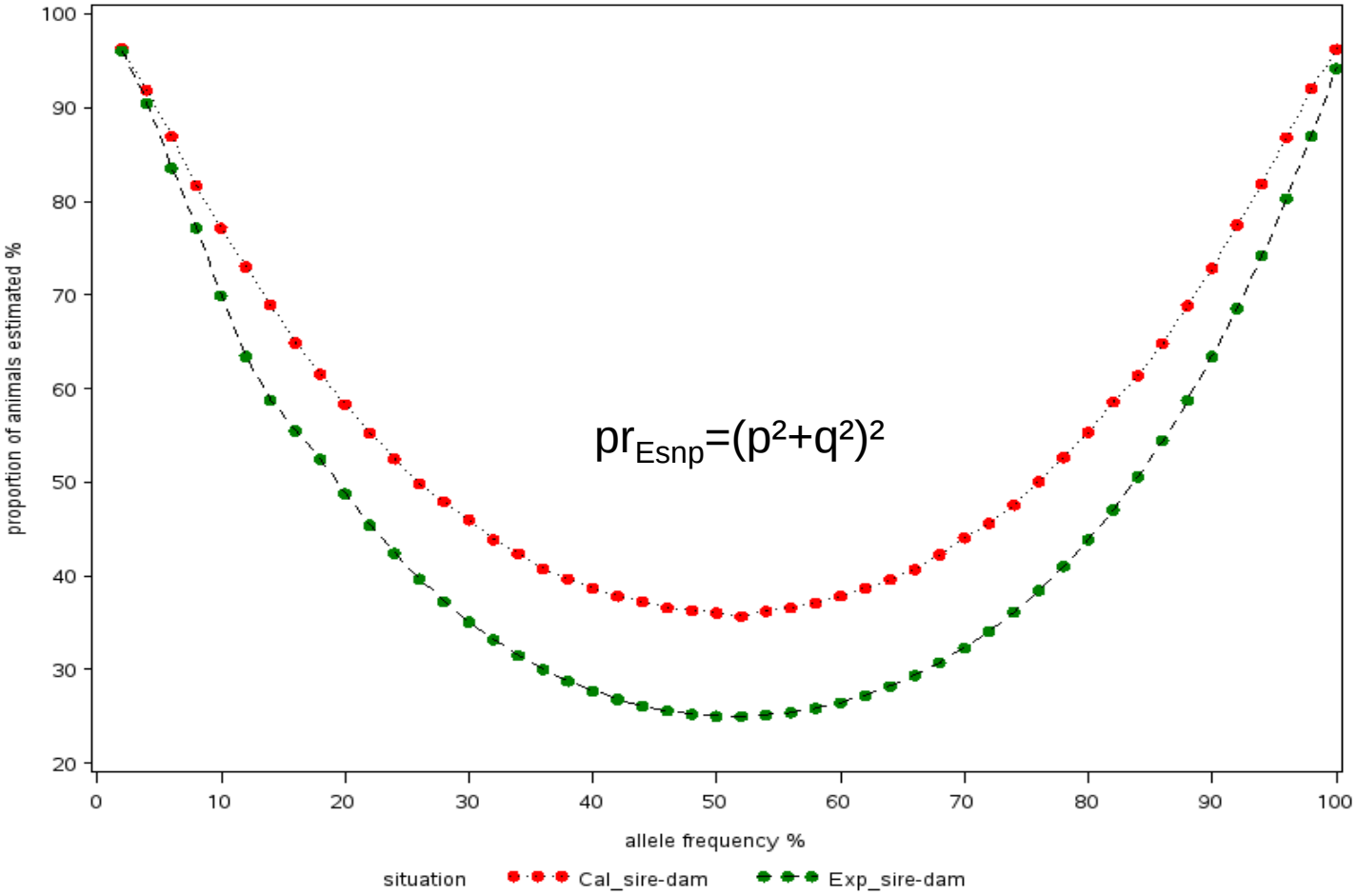
Average of observed proportion for estimated genotype 10%

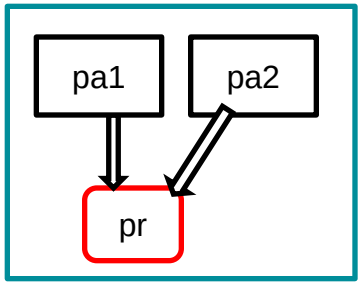




Average of expected proportion for estimated genotype 7%

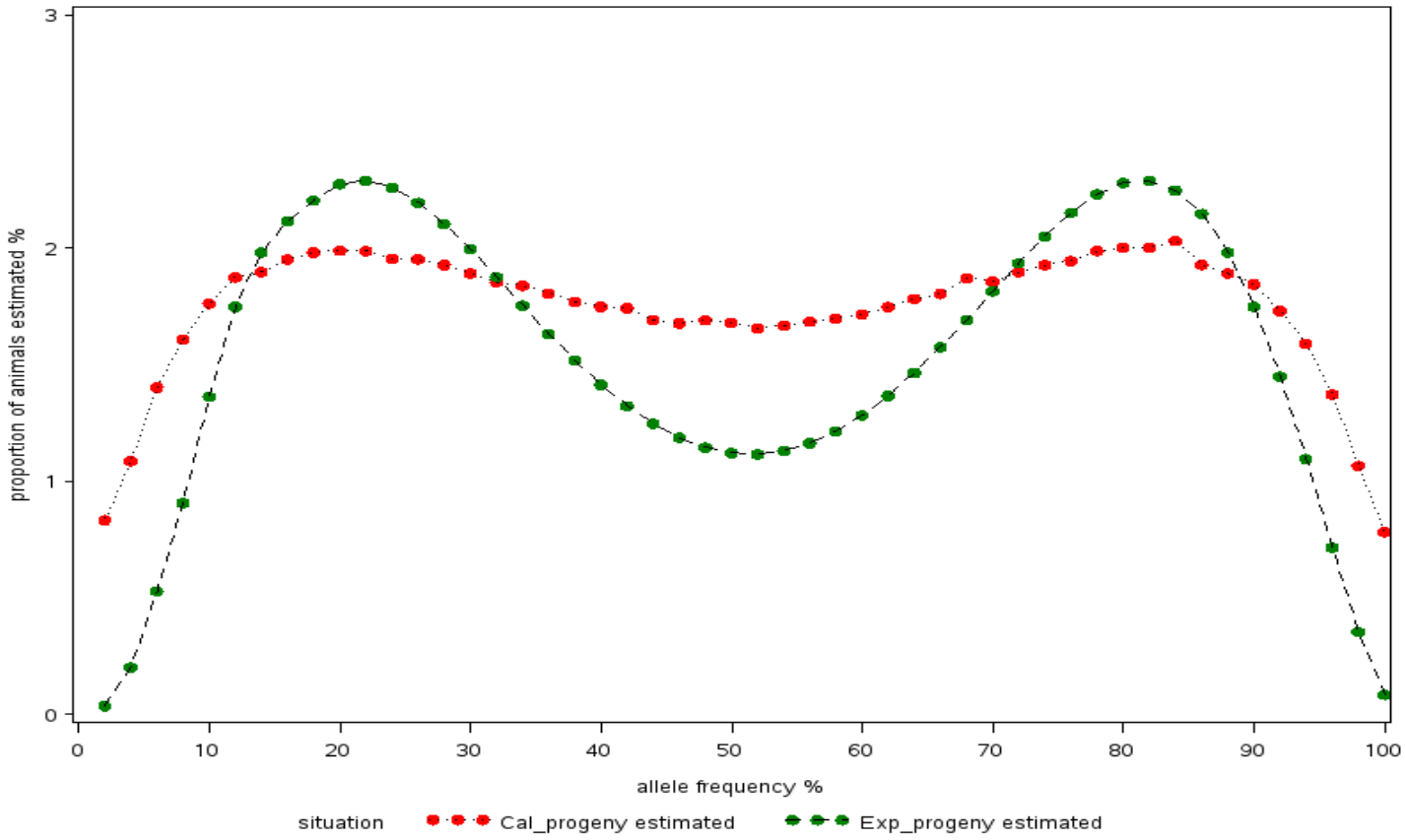
Average of observed proportion for estimated genotype 5%

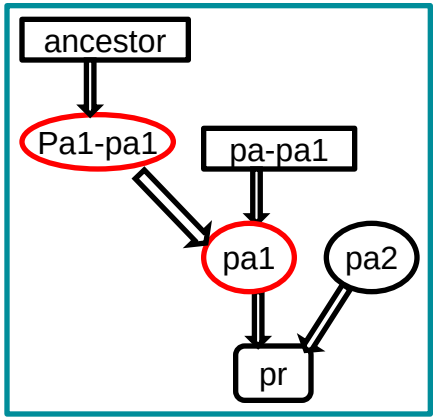




Average of expected proportion for estimated genotype 50%

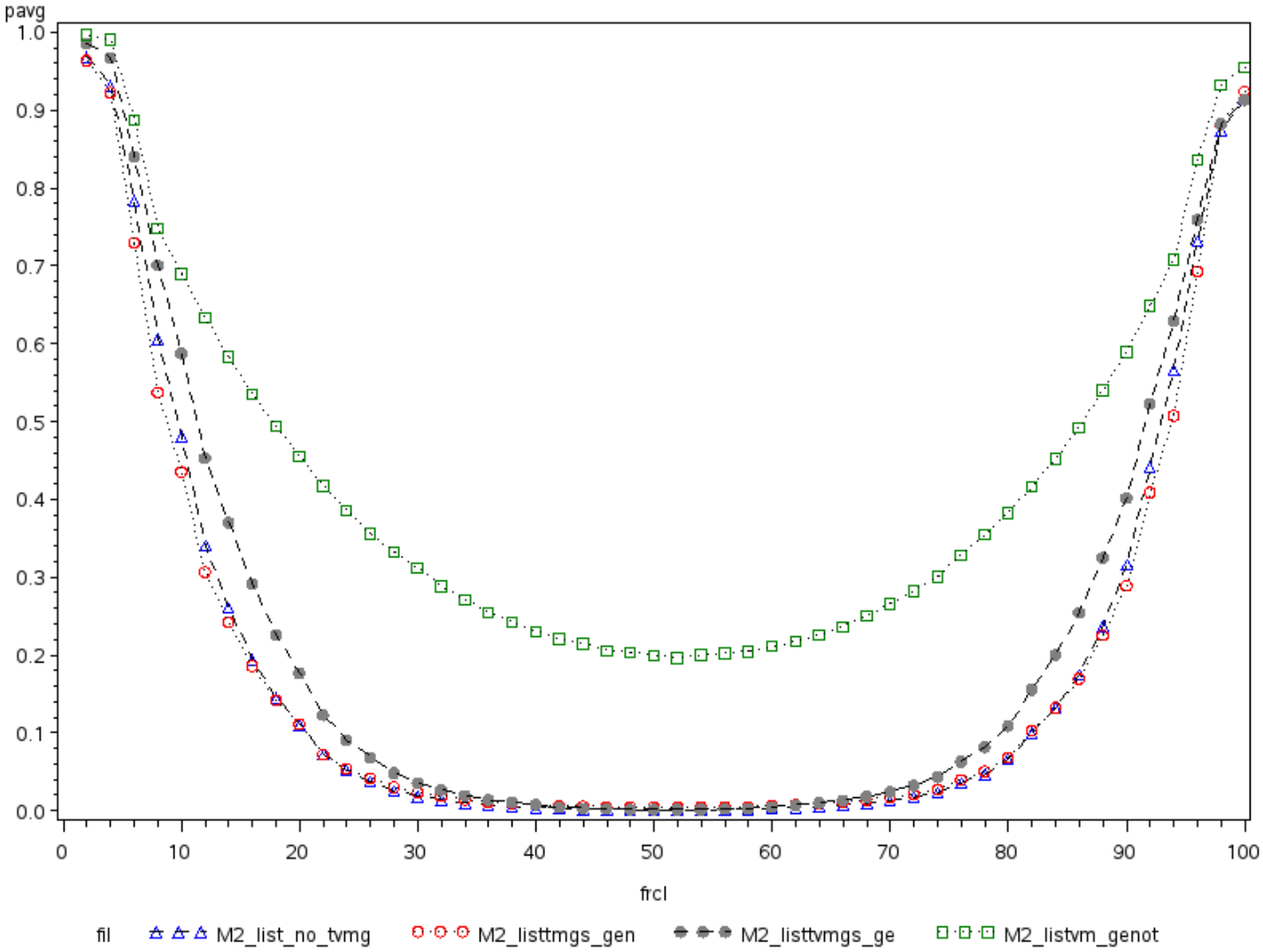
Average of observed proportion for estimated genotype 54%

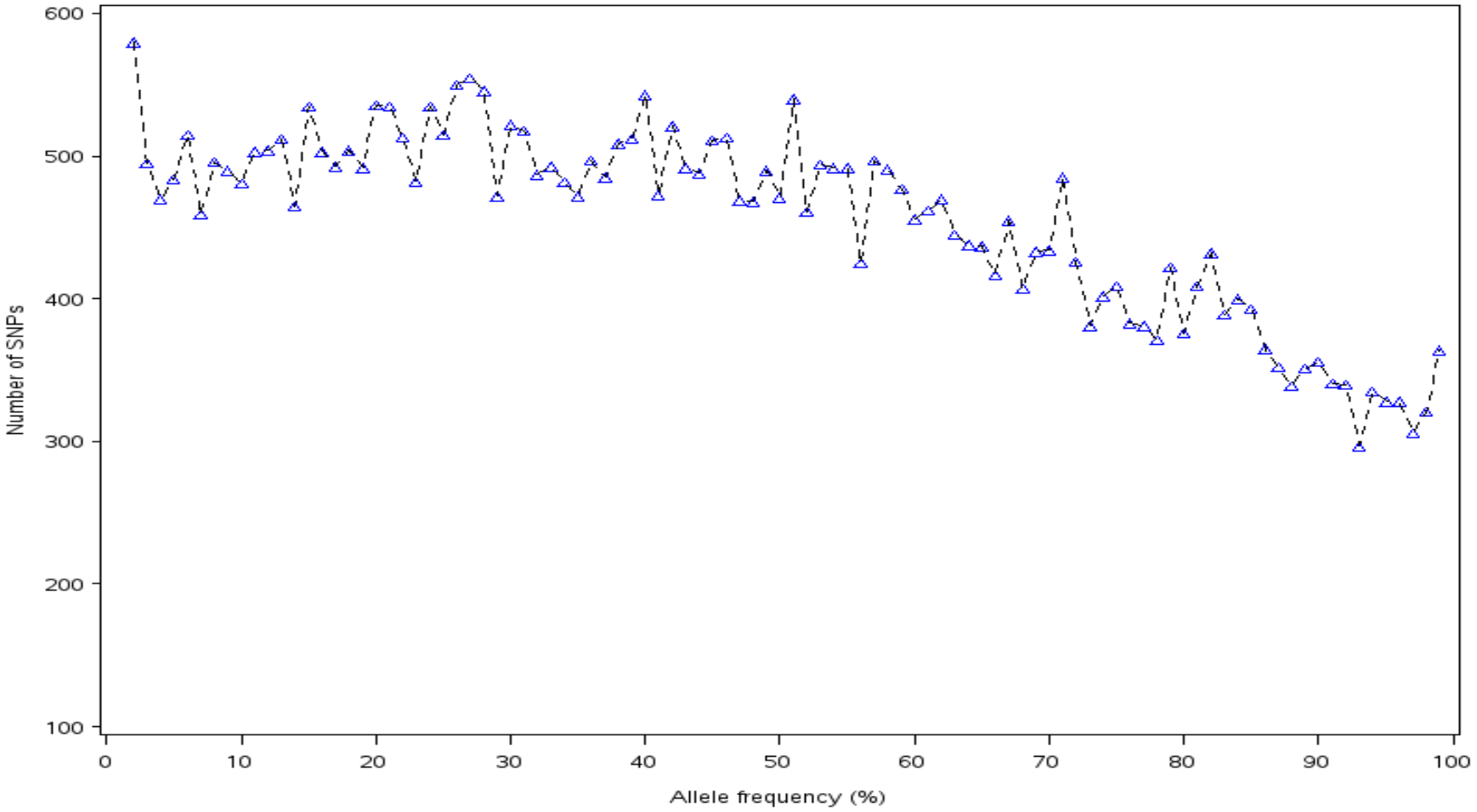


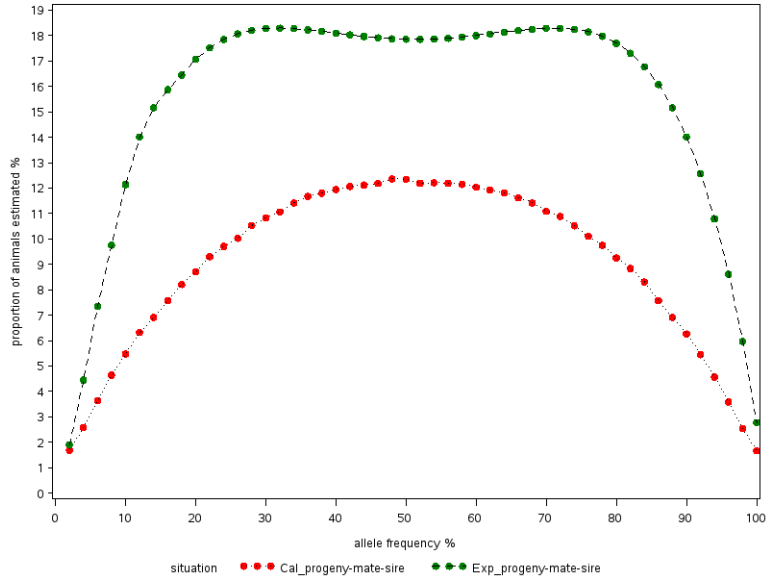


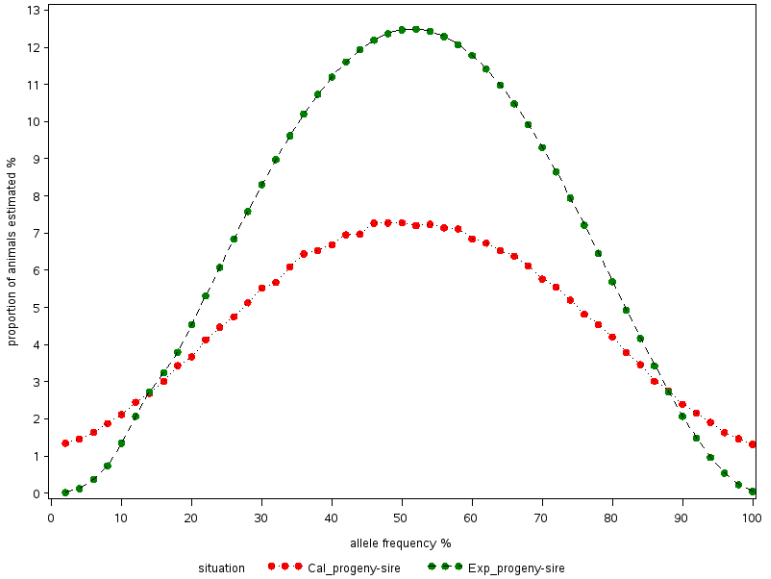
Average of expected proportion for estimated genotype 1.5%

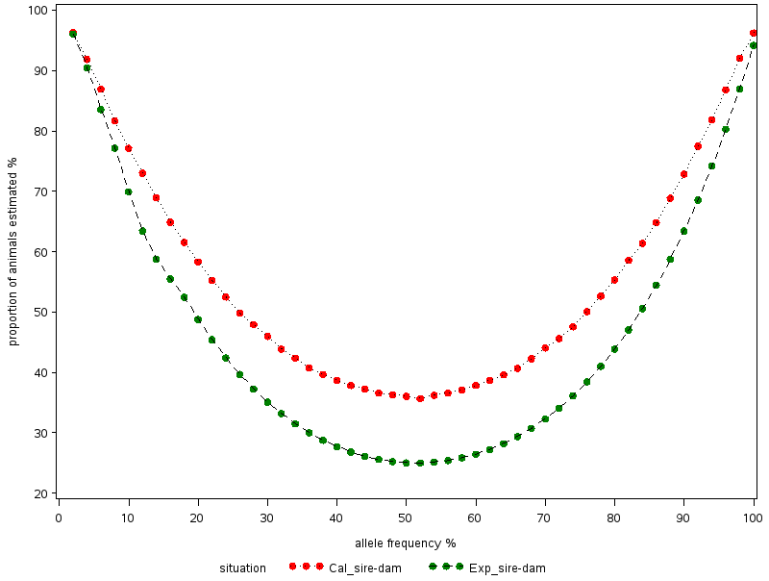
Average of observed proportion for estimated genotype 1.5%

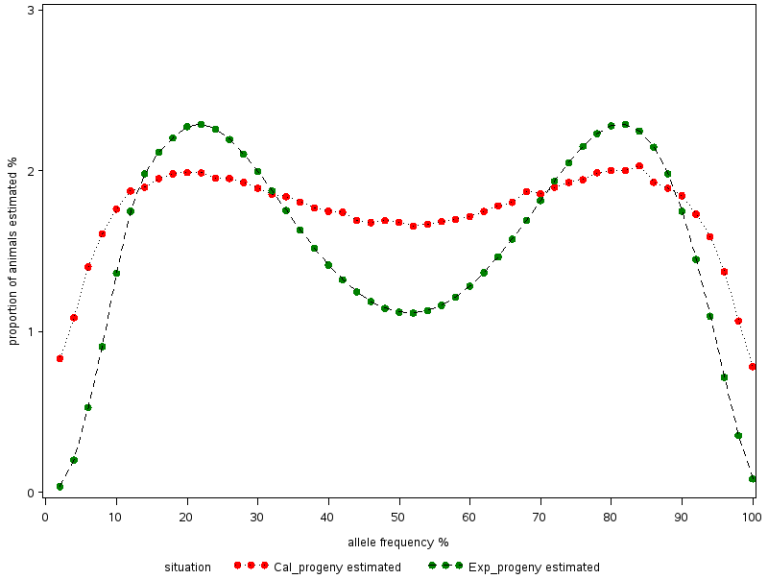

pavg
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
10
20
30
40
50
60
70
80
90
100
frcl
fil
M2_list_no_tvmg
M2_listtmgs_gen
M2_listtvmgs_ge
M2_listvm_genot

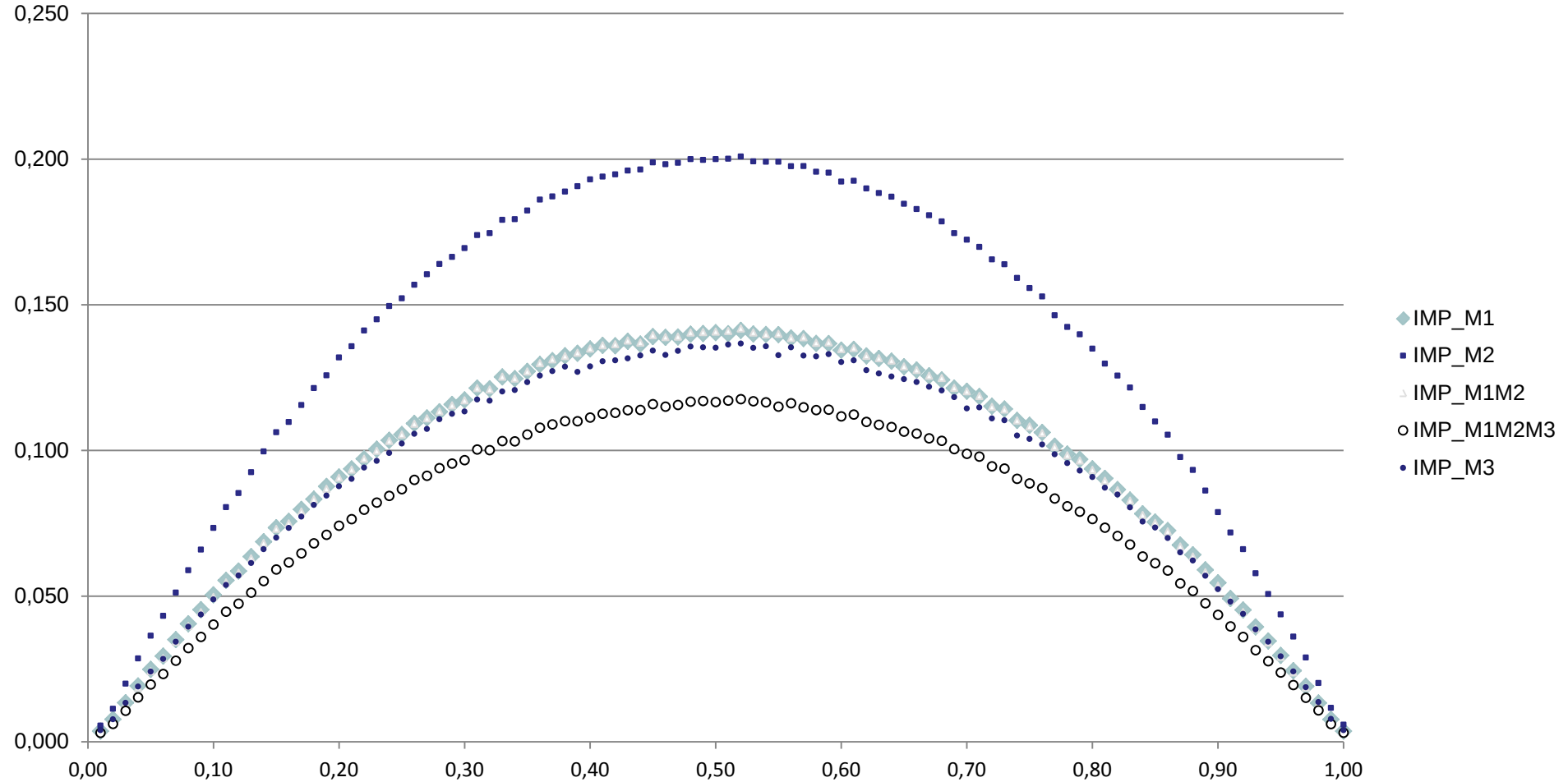

Number of SNPs
600
500
400
300
200
100
0
10
20
30
40
50
60
70
80
90
100
Allele frequency (%)

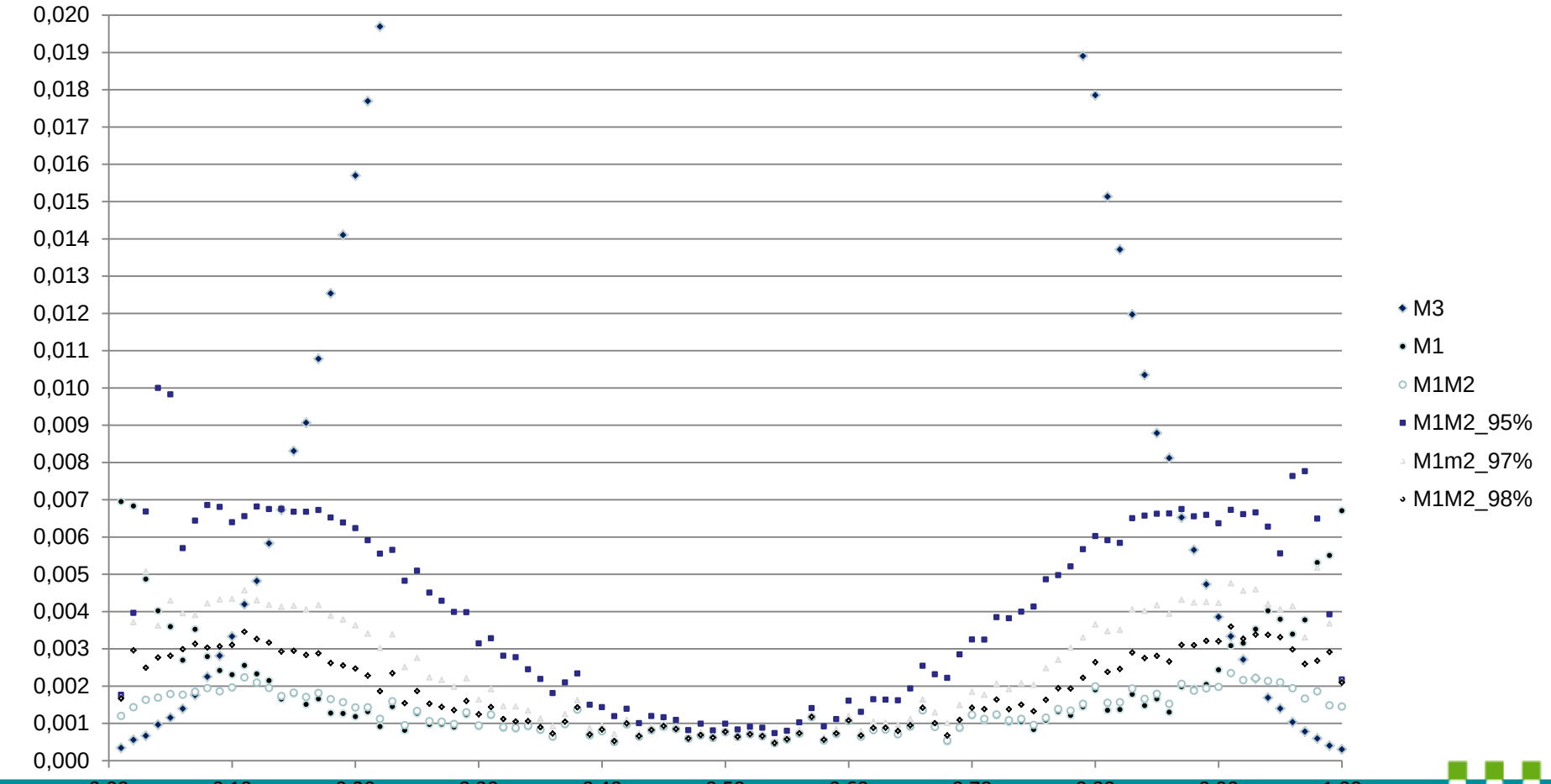

proportion of animals estimated %
allele frequency %
situation
Cal_progeny-mate-sire
Exp_progeny-mate-sire


proportion of animals estimated %
allele frequency %
situation
Cal_progeny-sire
Exp_progeny-sire


proportion of animals estimated %
allele frequency %
situation
Cal_sire-dam
Exp_sire-dam


proportion of animals estimated %
allele frequency %
situation
Cal_progeny estimated
Exp_progeny estimated

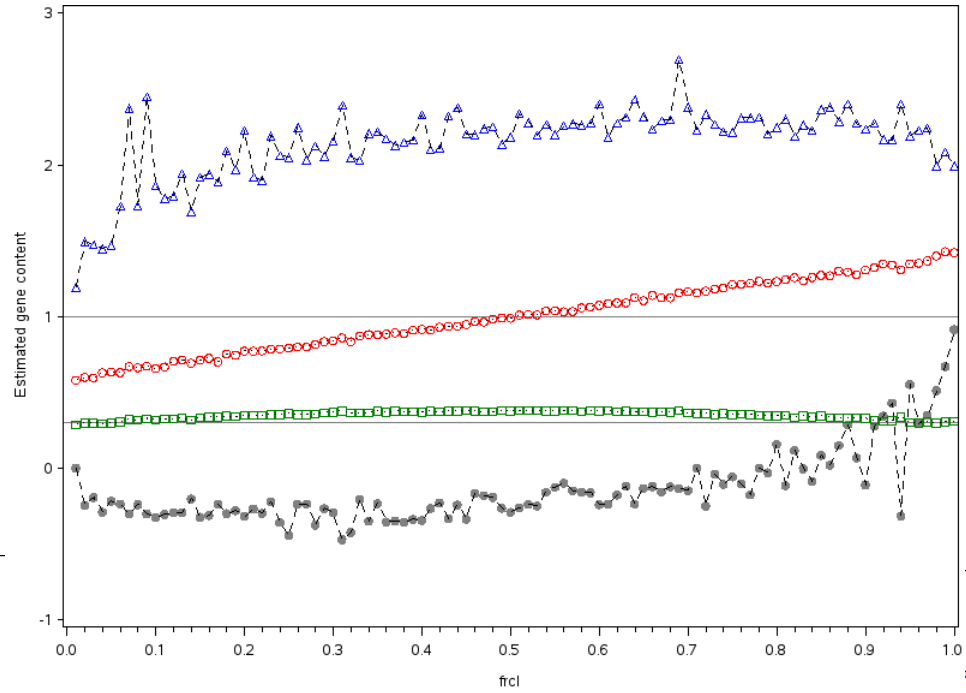

0,250
0,200
0,150
0,100
0,050
0,000
0,00
0,10
0,20
0,30
0,40
0,50
0,60
0,70
0,80
0,90
1,00
IMP_M1
IMP_M2
IMP_M1M2
IMP_M1M2M3
IMP_M3

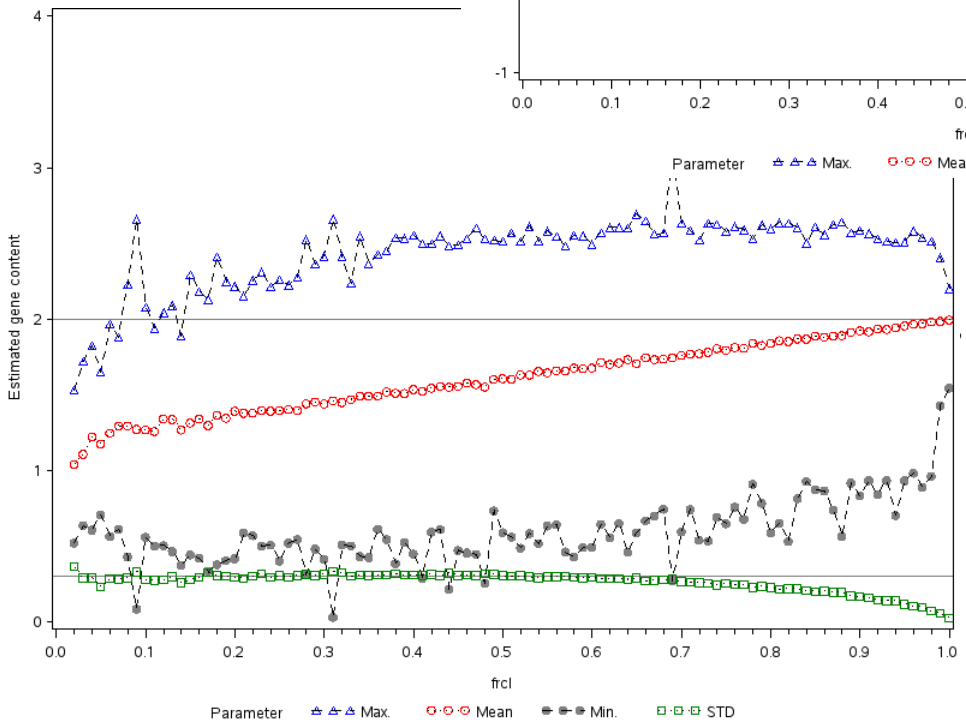

0,020
0,019
0,018
0,017
0,016
0,015
0,014
0,013
0,012
0,011
0,010
0,009
0,008
0,007
0,006
0,005
0,004
0,003
0,002
0,001
0,000
0,00
0,10
0,20
0,30
0,40
0,50
0,60
0,70
0,80
0,90
1,00
M3
M1
M1M2
M1M2_95%
M1m2_97%
M1M2_98%

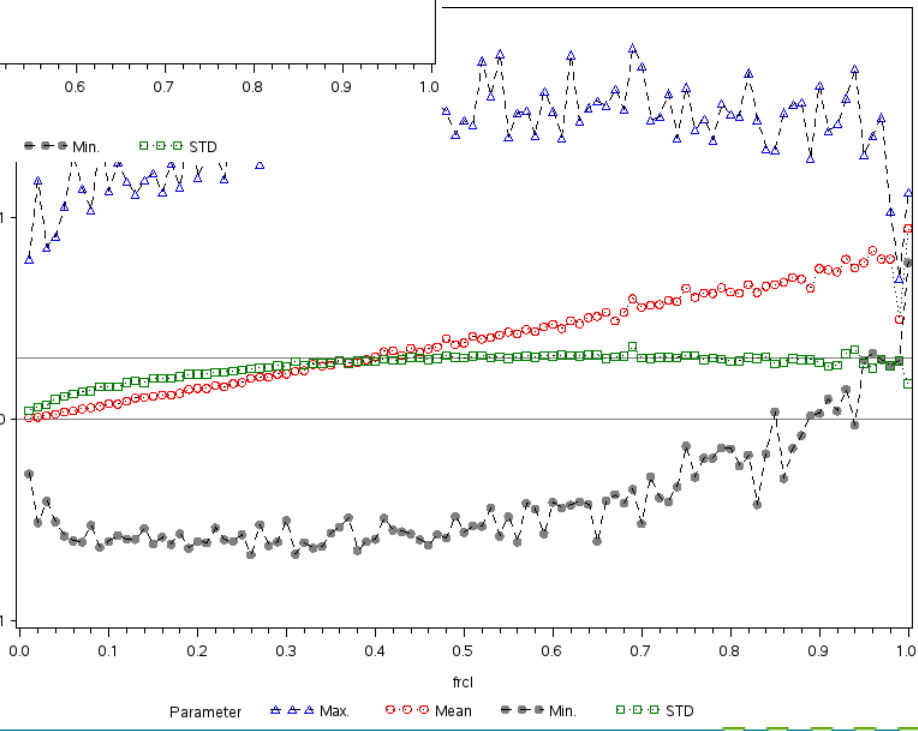
Results

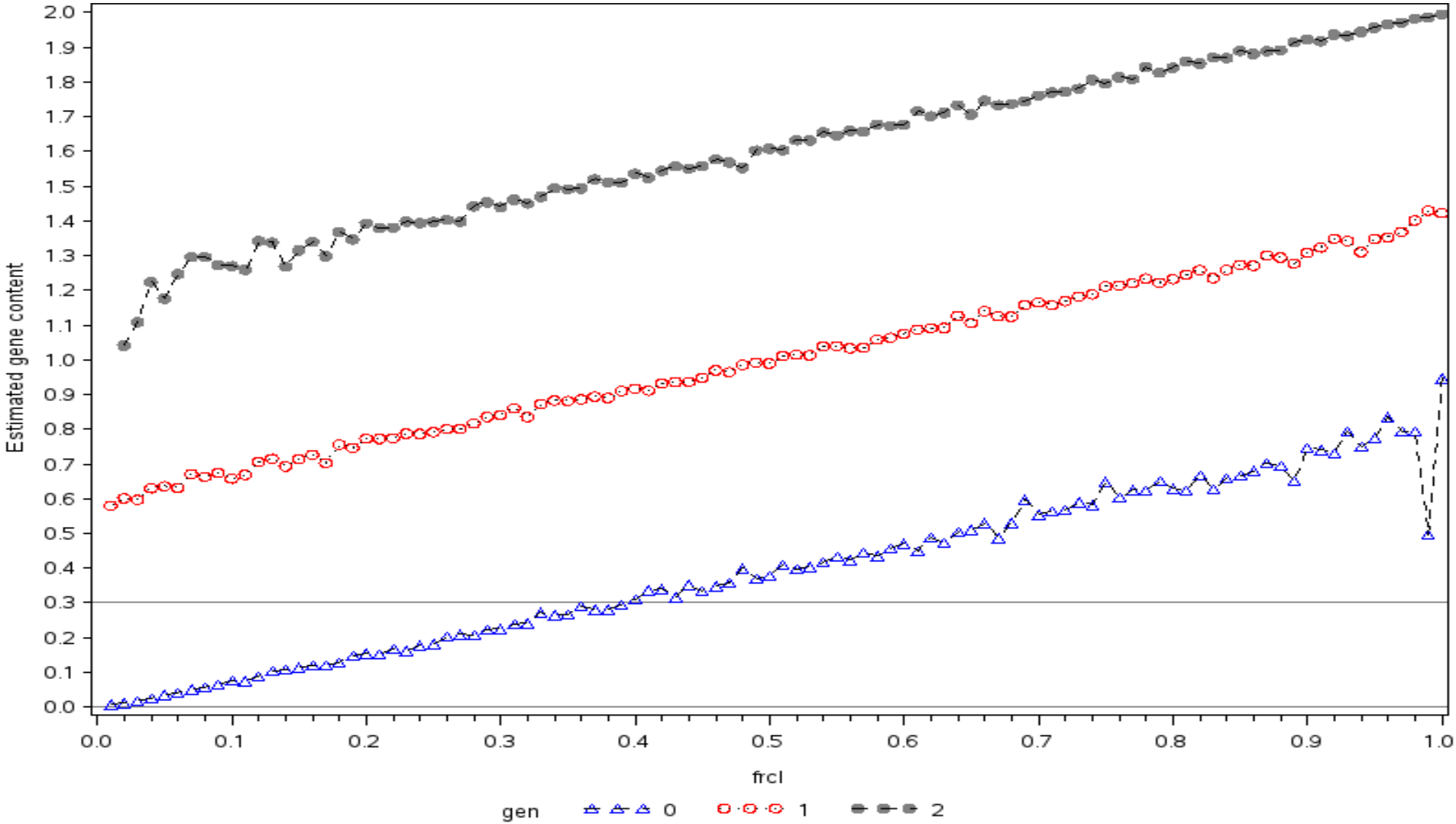
Estimated and imputed SNPs with all methods

	Estimated SNPs			Imputation		
	N estimated SNPs	Genotype correctly estimated (%)	Allele correctly estimated (%)	N imputed SNPs	Genotype correctly estimated (%)	Allele correctly estimated (%)
M1 (progeny to ancestor)	10.5	99.7	99.9	100	81.1	90.2
M2 (ancestor to progeny)	11.3	99.2	99.6		73.7	85.9
M3 (BLUP)	22.6	96.7	98.4		81.5	90.4
Joined M1M2	21.7	99.6	99.9		81.3	90.2
M1M2M3	34.8	97.7	98.9		84.3	92.0


Estimated gene content
frcl


Estimated gene content
frcl
Parameter
Max.
Mean
Min.
STD


frcl
Parameter
Max.
Mean
Min.
STD


Estimated gene content
2.0
1.9
1.8
1.7
1.6
1.5
1.4
1.3
1.2
1.1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
frcl
gen
0
1
2

N	progeny	Sire	Non-geno.
1	AA	AA	
2	AA	Aa	
1	AA	aa	$p^2 * q^2$
2	Aa	AA	
4	Aa	Aa	
2	Aa	aa	
1	aa	AA	$p^2 * q^2$
2	aa	Aa	
1	aa	aa	

	progeny	mate	sire	Non-geno.
1	AA	AA	AA	
2	AA	AA	Aa	
1	AA	AA	aa	p2*p2*q2
2	AA	Aa	AA	
4	AA	Aa	Aa	
2	AA	Aa	aa	p2*2pq*q2
1	AA	aa	AA	
2	AA	aa	Aa	
1	AA	aa	aa	
2	Aa	AA	AA	2pq*p2*p2
4	Aa	AA	Aa	
2	Aa	AA	aa	
4	Aa	Aa	AA	
8	Aa	Aa	Aa	
4	Aa	Aa	aa	
2	Aa	aa	AA	
4	Aa	aa	Aa	
2	Aa	aa	aa	2pq*q2*q2
1	aa	AA	AA	
2	aa	AA	Aa	
2	aa	AA	aa	
2	aa	Aa	AA	2pq*p2*q2
4	aa	Aa	Aa	
2	aa	Aa	aa	
1	aa	aa	AA	p2*q2*q2
2	aa	aa	Aa	

Methods: M3 BLUP gene content

Without laboratory error
To solve the equation we must estimate μ

$$q_x=(1 \quad A_{xy}A_y^{-1})=\begin{pmatrix}\mu\\ q_y-1\mu\end{pmatrix}$$

Laboratory error exist, BLUP can be used to solve the equation to estimate μ and predicted gene content d_x

$$\begin{pmatrix}1'1 & 1'M\\ M'1 & M'M+A^{-1}\varepsilon\end{pmatrix}\begin{pmatrix}\mu\\ d_y\\ d_x\end{pmatrix}=\begin{pmatrix}1'd_y\\ M'd_y\end{pmatrix}$$

Where d_y is a vector of gene content deviations for animals with genotype records.

d_x is a vector of gene content deviations for unobserved animals.

M is incidence matrix linking q_y to $\begin{pmatrix}d_y\\ d_x\end{pmatrix}$ that can be rewritten as (I_y 0_x).

A is the additive relationship matrix of the structure $A=\begin{pmatrix}A_{yy} & A_{yx}\\ A_{xy} & A_{xx}\end{pmatrix}$, $\varepsilon=\sigma_e^2/\sigma_d^2$.

A some small error variance (e. g. in this study 0.005) is required to solve the system of equations using BLUP methodology.

	Genotype probability level for M2 top down							
	Number of estimated SNPs				Alleles correctly estimated			
	Prob. e0.99	Prob. e0.98	Prob. e0.97	Prob. e0.95	Prob. e0.99	Prob. e0.98	Prob. e0.97	Prob. e0.95
EGC_{snp} ±0.05	**15,873**	**16,603**	**17,105**	**18,017**	0.988	0.988	0.988	0.987
EGC_{snp} ±0.10	**19,796**	**20,252**	**20,617**	**21,301**	0.980	0.981	0.980	0.980
EGC_{snp} ±0.12	**22,442**	**22,898**	**23,171**	**23,764**	0.975	0.975	0.975	0.975
EGC_{snp} ±0.16	**23,764**	**24,175**	**24,449**	**24,950**	0.972	0.973	0.973	0.972
EGC_{snp} ±0.20	**26,319**	**26,638**	**26,866**	**27,277**	0.967	0.967	0.967	0.967
EGC_{snp} ±0.25	**29,466**	**29,648**	**29,831**	**30,105**	0.961	0.961	0.961	0.960
EGC_{snp} ±0.30	**32,340**	**32,522**	**32,613**	**32,841**	0.953	0.953	0.953	0.953