World Congress on Genetics Applied to Livestock Production, 11. 176

Obtaining variance of gametic diversity with genomic models

D.J.A. Santos[1,4], J.B. Cole[2], P.M. VanRaden[2], T.J. Lawlor[3], H. Tonhati[1] & L. Ma[4]

unesp

USDA

U.S. Registered Holsteins
FOR MAXIMUM PROFIT
HOLSTEIN ASSOCIATION USA, INC.

UNIVERSITY OF MARYLAND
18 56

Central Idea

Breeding value inheritance components:

$$a_i = \frac{1}{2}a_s + \frac{1}{2}a_d + m_i$$

Mendelian sampling

How about future progeny?

Additive model (QTL effect):

Values:
A +5
B +3
C +2

	Sire 1	Heterozygosity Sire 2
	AA bb cc 10 0 0	Aa Bb Cc 5 3 2
PTA expectation/average:	+5	+ 5
Possible values for the gametes	+5	0,2,3,5,5,7,8,10
Probability (binomial distribution)	1	0.125,0.125,0.125,0.25,0.125,0.125,0.125

> 7 (0.25)

Exactly 5

> 0 (0.875)

If linked (phased ABC|abc):
Recombination rates: 0.2 AB|BC

> 0 (0.68)

Exactly 5

0.32,0.08,0.02,0.16,0.02,0.08,0.32

> 7 (0.40)

Statistics Background

Binomial Variances and Covariances || Solutions

$$Var(x) = \sum x^2 - \frac{(\sum x)^2}{N}) = Np_A - \frac{(Np_A)^2}{N} = N(p_A - p_A^2) = \underline{N[p_A(1-p_A)]}$$

$$Var(y) = \sum y^2 - \frac{(\sum y)^2}{N}) = Np_B - \frac{(Np_B)^2}{N} = N(p_B - p_B^2) = \underline{N[p_B(1-p_B)]}$$

$$Cov(x,y) = \sum xy - \frac{\sum x \sum y}{N} = Np_{AB} - \frac{Np_A Np_B}{N} = \underline{N(p_{AB} - p_A\, p_B)}$$

$1*0.5*0.5*S_A{}^2$

$1*0.5*0.5*S_B{}^2$

$1*(p_{AB}-0.5*0.5)*S_AS_B$

Gametic phase

½ centiMorgan (0.01 Morgan)

p_{AB} =0.25 => cov_{ab}=0*S_AS_B

p_{AB} =0 or 0.50 => cov_{ab}= ±0.25*S_AS_B

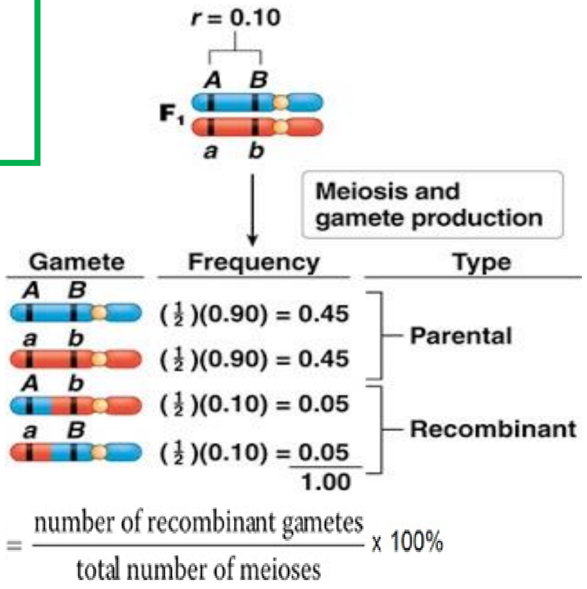
$$\sigma^2{}_{[A+B]} = (\sigma^2{}_A + \sigma^2{}_B + 2\sigma_{AB})$$

$$\sigma^2{}_{[A+B]} = (\sigma^2{}_A + \sigma^2{}_B\ )$$

If independent !!!

$$\boxed{\sigma^2{}_{gamete} = \sigma^2{}_{\Sigma\ Nlocus}}$$

Homozygous loci:
$N*p*(1-p)*S^2 = 0$



Gamete	Frequency	Type
A B	(½)(0.90) = 0.45	Parental
a b	(½)(0.90) = 0.45	
A b	(½)(0.10) = 0.05	Recombinant
a B	(½)(0.10) = 0.05	
	1.00	

$$= \frac{\text{number of recombinant gametes}}{\text{total number of meioses}} \times 100\%$$

Method

Methods for computing

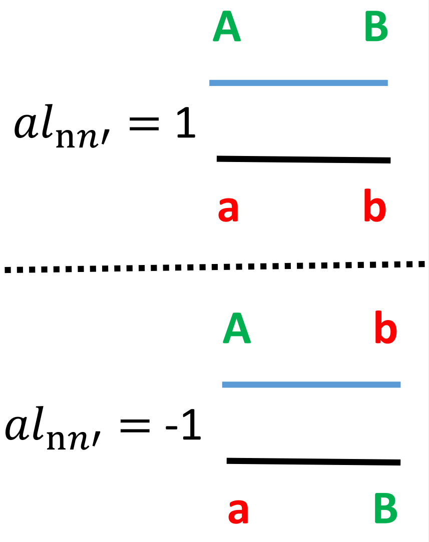
Heterozygote loci

$$\sigma^2_{\Sigma\ \text{Nlocus}} = \begin{bmatrix} S_1 & \dots & S_n \end{bmatrix} \begin{bmatrix} 0.25 & \cdots & al_{n1}\left(-\frac{cM_{1n}}{200}+0.25\right) \\ \vdots & \ddots & \vdots \\ al_{1n}\left(-\frac{cM_{1n}}{200}+0.25\right) & \cdots & 0.25 \end{bmatrix} \begin{bmatrix} S_1 \\ \vdots \\ S_n \end{bmatrix}$$

Reference allele

locus 1= A

locus 2= B



$al_{nn'} = 1$

$al_{nn'} = \text{-}1$

Independent → cM=0.25 (25% for each gamete)

$$\begin{bmatrix} 0.25 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0.25 \end{bmatrix}$$

$$\boxed{\sigma^2_{\text{gamete}} = \sigma^2_{\Sigma\ \text{Nlocus}}}$$

> 50 cM is considered as independent (that is 50 cM)

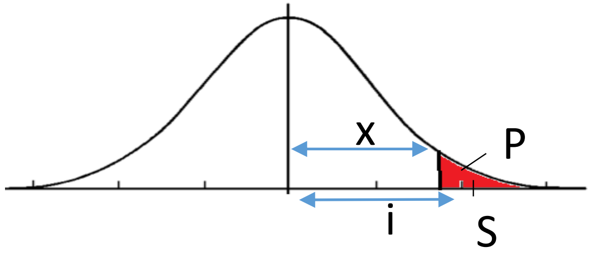
Application

Confidence intervals

Strategies of selection

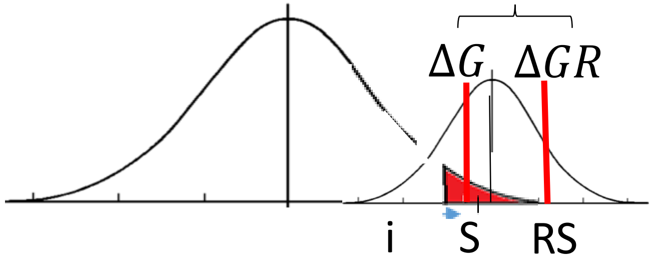
$$\boxed{RPTA_i = PTA_i + \sigma_{gametic\_i} * i_f}$$

Genetic Gain in Future



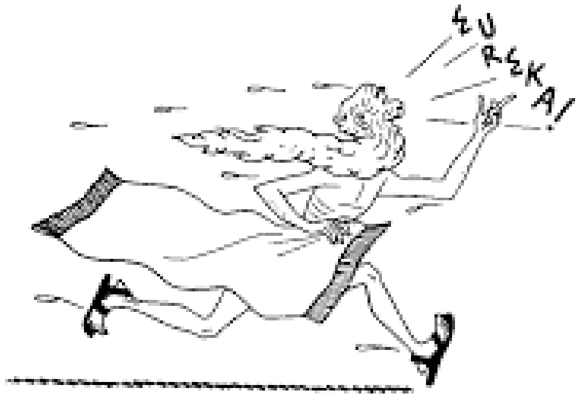
$$\Delta G = r * i * \sigma_a$$

$$\boxed{\sqrt{\sigma^2 a + 4 i_f^2 var(\sigma_{gametic\_i}{}^2)} - \sqrt{\sigma^2 a}}$$



$$\boxed{\Delta GR = r * i * \sqrt{\sigma_a{}^2 + 4 * var(\sigma_{gametic\_i}{}^2) * i_f^2}}$$

Important Questions



In practice, how can we obtain the variance of gametic diversity?

Using marker effects estimated from routine genomic evaluation!!!

Subsequent questions about this approach:

1- Should the recombination rate also be considered (dependence) between the markers?

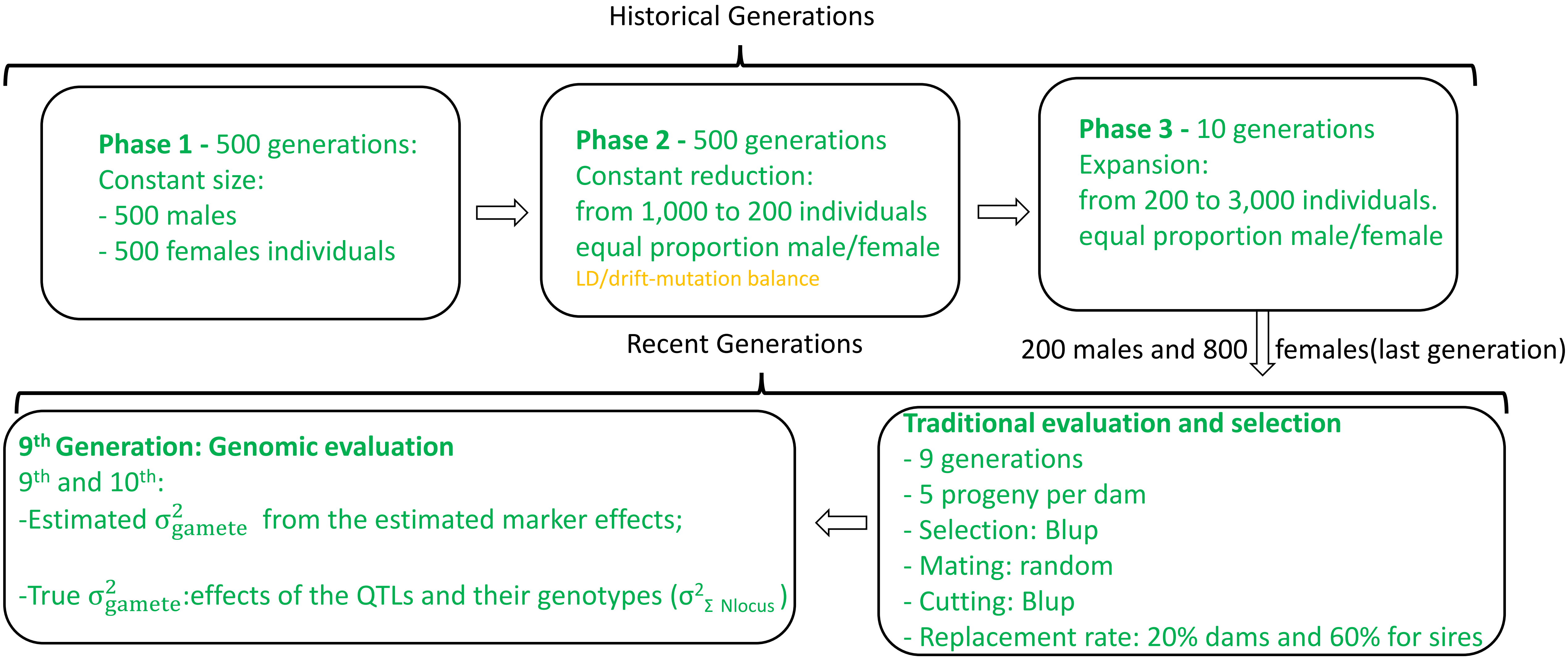
2 - What should the density panel marker be?

3 - Which models to use?

4 - What is the MAF effect?

To answer these questions, simulation study was proposed !!!

Simulation - Population



Historical Generations

Phase 1 - 500 generations:
Constant size:
- 500 males
- 500 females individuals

Phase 2 - 500 generations
Constant reduction:
from 1,000 to 200 individuals
equal proportion male/female
LD/drift-mutation balance

Phase 3 - 10 generations
Expansion:
from 200 to 3,000 individuals.
equal proportion male/female

200 males and 800 females(last generation)

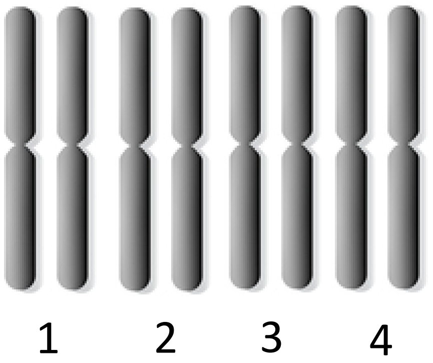
Recent Generations

Traditional evaluation and selection
- 9 generations
- 5 progeny per dam
- Selection: Blup
- Mating: random
- Cutting: Blup
- Replacement rate: 20% dams and 60% for sires

9th Generation: Genomic evaluation
9th and 10th:
-Estimated $\sigma^2_{\mathrm{gamete}}$ from the estimated marker effects;

-True $\sigma^2_{\mathrm{gamete}}$:effects of the QTLs and their genotypes ($\sigma^2_{\Sigma\ \mathrm{Nlocus}}$)

Simulation – Genome and Traits



Genome Size 200 cM

Others Genome Parameters

Mutation Rater QTL	$2.5x10^{-5}$
Mutation Rater Marker	$2.5x10^{-3}$
Marker positions in genome	Evenly spaced
QTL position in genome	Random (uniform distribution)
QTL allele effect	Gamma distribution (β=0.4)

Scenarios: 4 traits (QTLs x h^2) x 2 SNPs panels

Traits:

Nº of QTL:

20 (0.1 QTL/cM) (low density)

200 (1 QTL/cM) (Meuwissen et al., 2001)

h^2:

0.1 and 0.3

$\sigma^2_{\text{phenotypic}} = 1$

4 replicates for each trait

X

Markers and Panels:

200,000 markers were simulated and randomly distributed

HD => 10% of the polymorphic markers sampled each 0.5 cM

SEQ => 20% of the markers also sampled every 0.5 cM and all QTLs

All simulations were performed QMSim version 1.10 (Sargolzaei & Schenkel, 2009)

Genomic Model

Depends on the effects of the markers:

$$y = \mu + Ma + e$$

Residual ~N(0,Iσ_e^2)

Marker

MAF≥ 0.05 (to mimic a conventional genomic evaluation)

1 - Traditional (SNP-BLUP/GBLUP)

$$a \sim N(0,\sigma^2) / u \sim N(0, G\sigma_a^2)$$

2 - Differential shrinkage (Improved LASSO)

$$\Pr(a_i \mid \tau^2) = N(0, \tau_i^2)$$

$$\Pr(\tau_i^2 \mid \lambda) = \lambda^2 \exp(-\lambda^2 \mid \tau_i^2 \mid)$$

The analyses were performed using GS3 v.3 software (Legarra et al., 2015)

Variance components:
- initial values = true values
-interactions: 20,000
-burn-in: 2,000.

Gametic Variance

$$\left.\begin{array}{l} 1-\boldsymbol{\sigma}_{\boldsymbol{g}}^{\mathbf{2}} = All\ QTL \\[2em] 2-\boldsymbol{\sigma}_{\mathbf{g\_maf}}^{\mathbf{2}} = QTL\ with\ MAF \geq 0.05 \end{array}\right\} \begin{bmatrix} 0.25 & \dots & al_{\mathrm{n1}}\left(-\frac{cM_{1n}}{200}+0.25\right) \\ \vdots & \ddots & \vdots \\ al_{1n}\left(-\frac{cM_{1n}}{200}+0.25\right) & \dots & 0.25 \end{bmatrix}$$

$$\left.\begin{array}{l} 3-\boldsymbol{\sigma}_{\mathbf{dia}}^{\mathbf{2}} = All\ QTL \\[2em] 4-\boldsymbol{\sigma}_{\mathbf{dia\_maf}}^{\mathbf{2}} = QTL\ with\ MAF \geq 0.05 \end{array}\right\} \begin{bmatrix} 0.25 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0.25 \end{bmatrix}$$

Results

Correlation of True Values

Scenario			QTLs data		
h^2	QTL		$\sigma^2_{\text{g\_maf}}$	σ^2_{dia}	$\sigma^2_{\text{dia\_maf}}$
0.1	20	σ^2_{g}	0.75	0.96	0.69
	200		0.96	0.50	0.48
0.3	20		0.94	0.95	0.90
	200		0.95	0.55	0.52

Medium magnitude

High magnitude !!!

It implies that QTLs with low MAF are important for obtaining accurate estimates of σ^2_{gamete}

σ^2_{gamete} does not depend directly on population allele frequencies but on the individual's heterozygous state (allele carrier).

Correlation between True and Estimated $\sigma^2_{\mathrm{gamete}}$

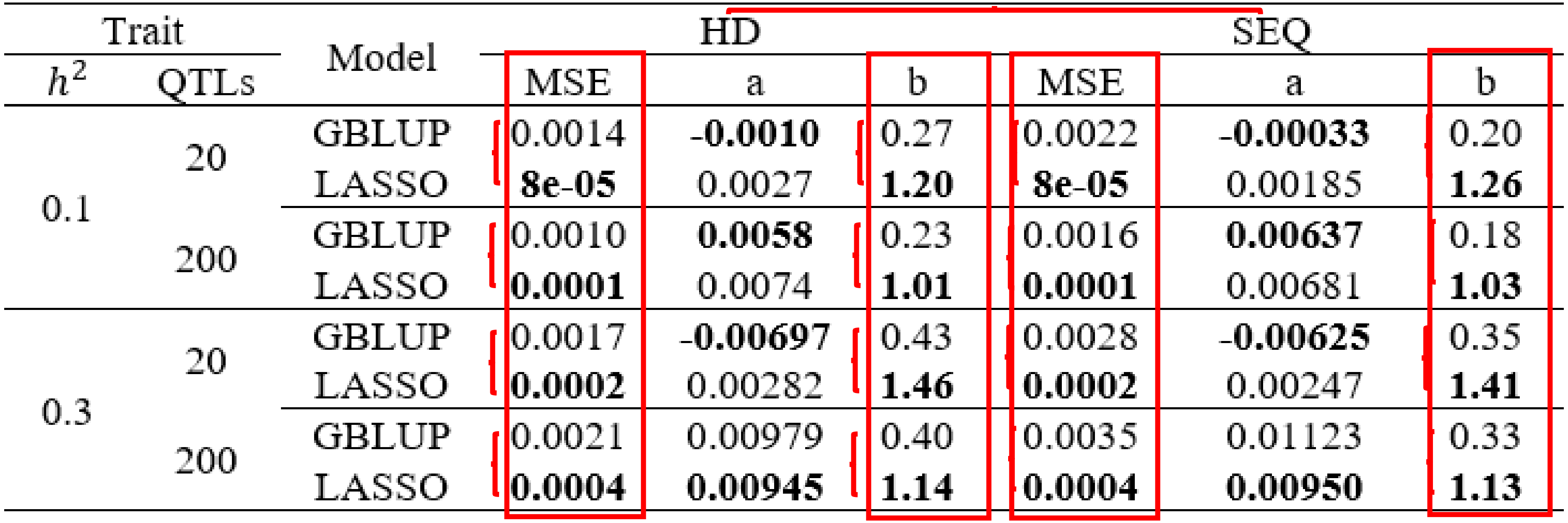
Similar accuracy!

Scenario		High-sensity panel				Sequencing data			
h^2	QTL	σ^2_{gblup}	$\sigma^2_{\mathrm{glasso}}$	$\sigma^2_{\mathrm{dia\_blup}}$	$\sigma^2_{\mathrm{dia\_}lasso}$	σ^2_{gblup}	$\sigma^2_{\mathrm{glasso}}$	$\sigma^2_{\mathrm{dia\_blup}}$	$\sigma^2_{\mathrm{dia\_}lasso}$
0.1	20	0.49	**0.56**	0.17	0.39	0.46	**0.57**	0.20	0.40
	200	0.50	**0.60**	0.29	0.37	0.46	**0.61**	0.29	0.40
0.3	20	0.64	**0.83**	0.28	0.66	0.59	**0.83**	0.07	0.65
	200	0.63	**0.77**	0.25	0.49	0.59	**0.77**	0.29	0.48

Best accuracy

Worst accuracy!

Bias



Trait		Model	HD			SEQ		
h^2	QTLs		MSE	a	b	MSE	a	b
0.1	20	GBLUP	0.0014	**-0.0010**	0.27	0.0022	**-0.00033**	0.20
		LASSO	**8e-05**	0.0027	**1.20**	**8e-05**	0.00185	**1.26**
	200	GBLUP	0.0010	**0.0058**	0.23	0.0016	**0.00637**	0.18
		LASSO	**0.0001**	0.0074	**1.01**	**0.0001**	0.00681	**1.03**
0.3	20	GBLUP	0.0017	**-0.00697**	0.43	0.0028	**-0.00625**	0.35
		LASSO	**0.0002**	0.00282	**1.46**	**0.0002**	0.00247	**1.41**
	200	GBLUP	0.0021	0.00979	0.40	0.0035	0.01123	0.33
		LASSO	**0.0004**	**0.00945**	**1.14**	**0.0004**	**0.00950**	**1.13**

Mean squared prediction (MSE): ↓values

Coefficient of the linear regression (b): close to one

GBLUP - higher predicted bias (overestimation)

HD X SEQ - Similar Bias

Conclusions

1 - The σ^2_{gamete} can be obtained by GM using HD panels without the need to use sequencing data.

2 - Differential shrinkage models are preferred;

3 - Markers with low MAF should be also used;

4 - The covariance (dependence) among markers should be considered.

For improving the accuracy of the estimations

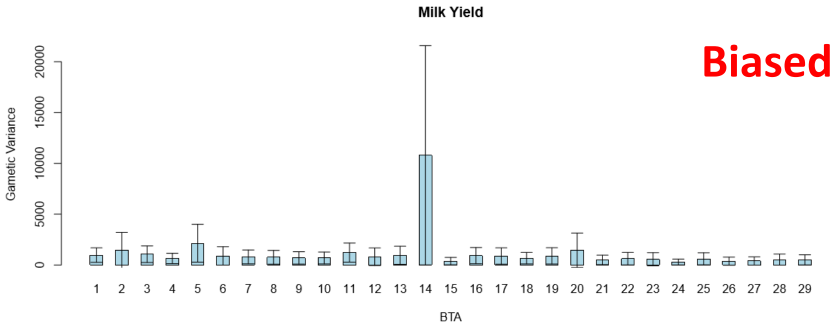
Acknowledgement

Financial Support

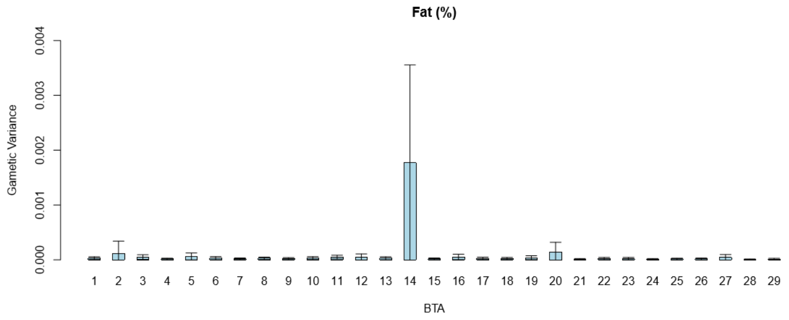
- **BARD Research Project US-4997-17**
- **USDA-NIFA Foundational Grant 2016-67015-24886**
- **FAPESP 2017/00462-5**

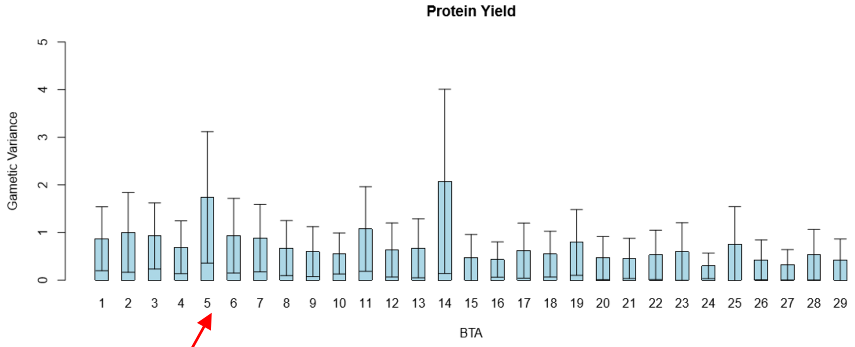
Thank you!!!

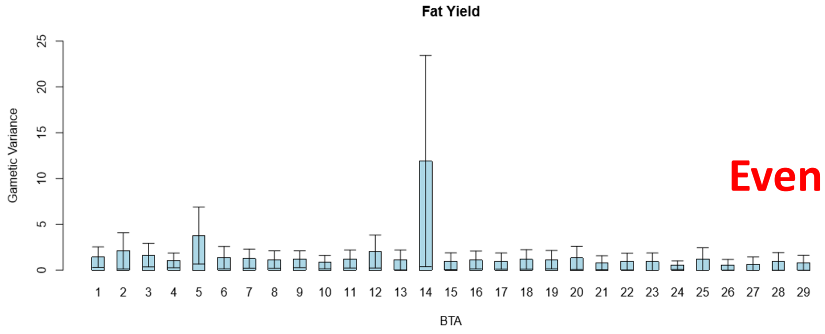
Real Data: USDA/Jersey



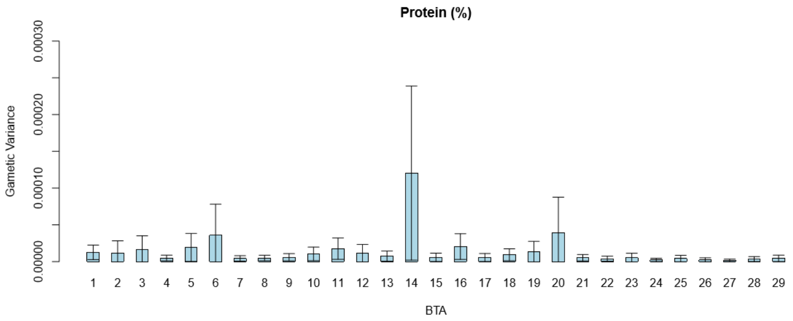
Biased distribution among chromosomes



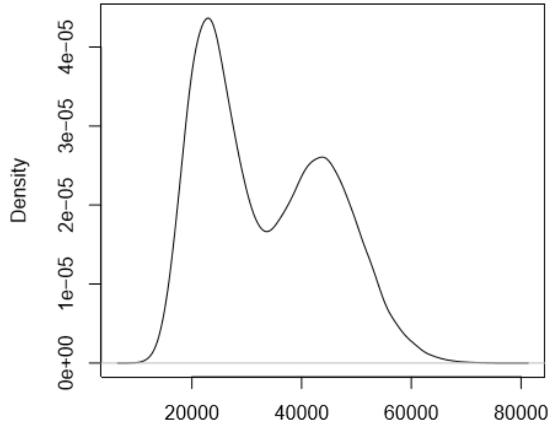




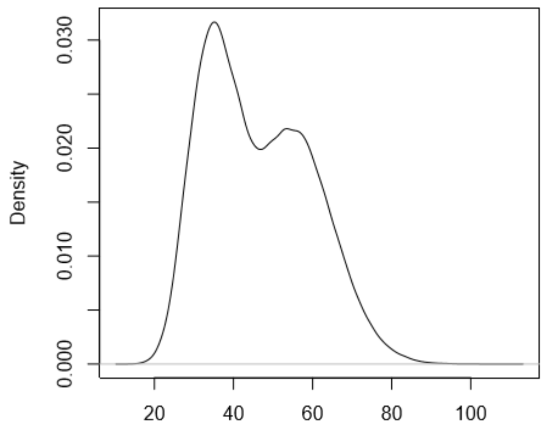
Even distribution among chromosomes



Distribution of σ^2_{gamete} for Production Traits

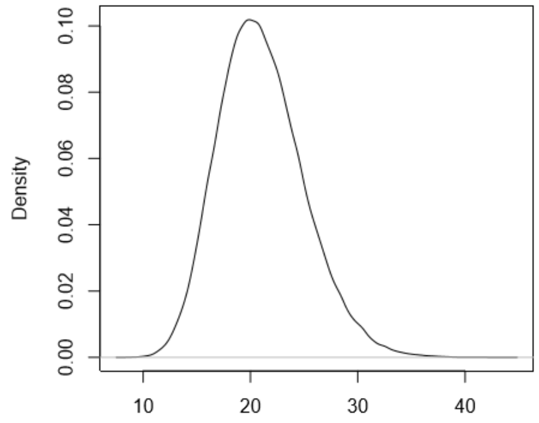


Milk Yield



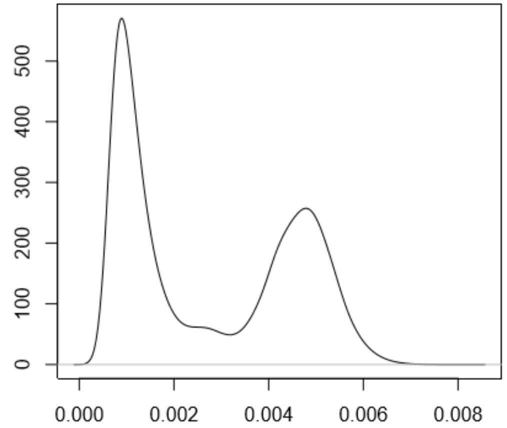
Fat Yield

Atypical Gaussian curve

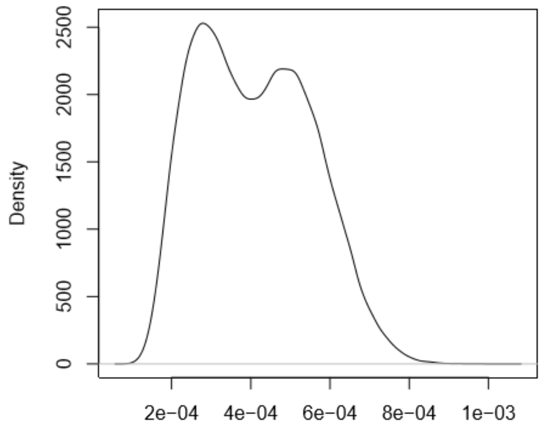


Protein Yield

Close to typical Gaussian curve

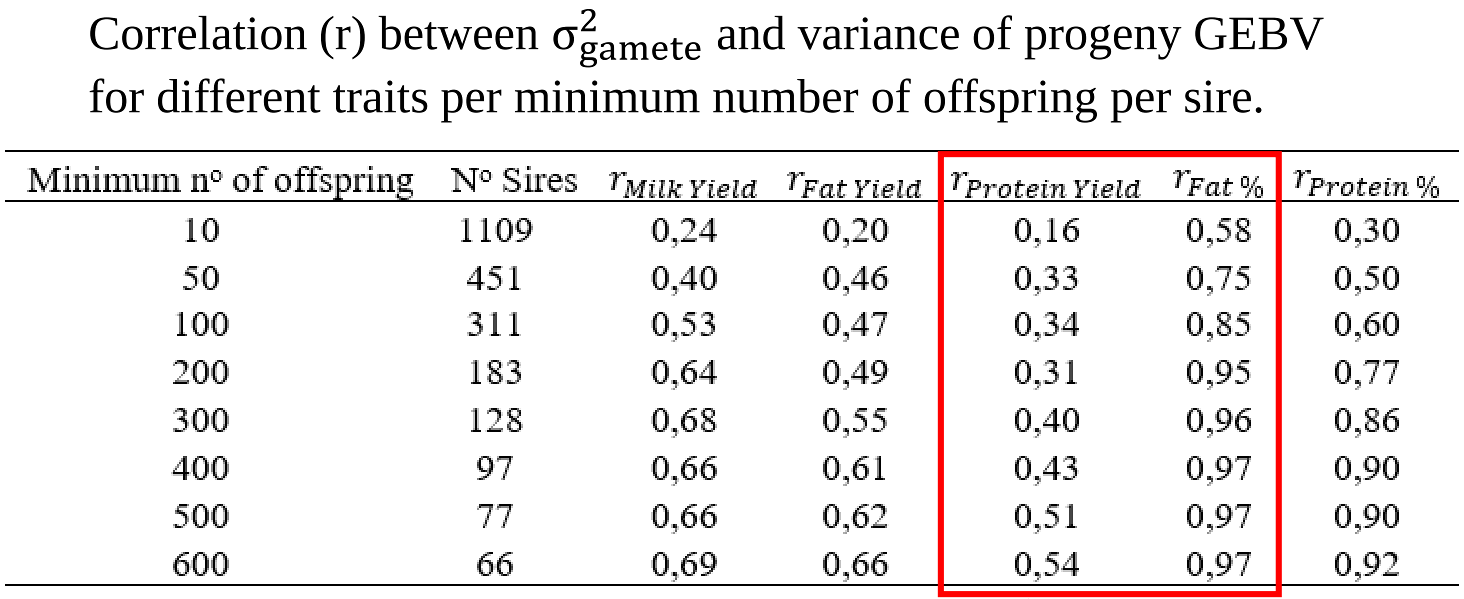


Fat (%)



Protein (%)

Applied example: USDA/Jersey



Correlation (r) between σ^2_{gamete} and variance of progeny GEBV for different traits per minimum number of offspring per sire.

Minimum nº of offspring	Nº Sires	$r_{Milk\ Yield}$	$r_{Fat\ Yield}$	$r_{Protein\ Yield}$	$r_{Fat\ \%}$	$r_{Protein\ \%}$
10	1109	0,24	0,20	0,16	0,58	0,30
50	451	0,40	0,46	0,33	0,75	0,50
100	311	0,53	0,47	0,34	0,85	0,60
200	183	0,64	0,49	0,31	0,95	0,77
300	128	0,68	0,55	0,40	0,96	0,86
400	97	0,66	0,61	0,43	0,97	0,90
500	77	0,66	0,62	0,51	0,97	0,90
600	66	0,69	0,66	0,54	0,97	0,92

Incresing

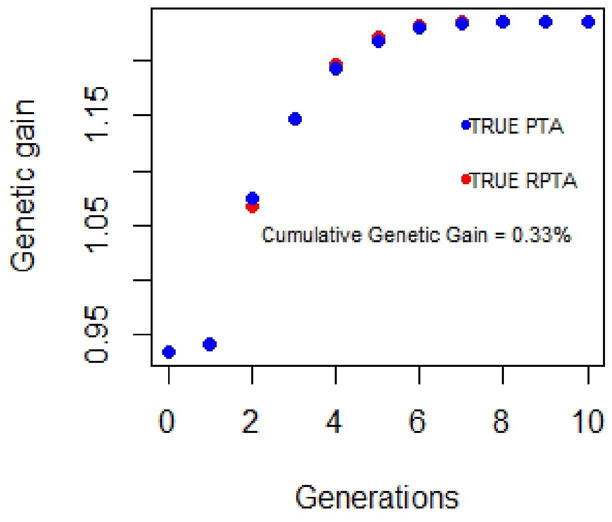
Lowest Protein Yield

Greatest Fat %

Motivating Results – TRUE RPTA / PTA

Simulation: Future generations; sires (i=1.75) and Dam (i=0.97).

TRUE RPTAs were corrected for number of offspring;

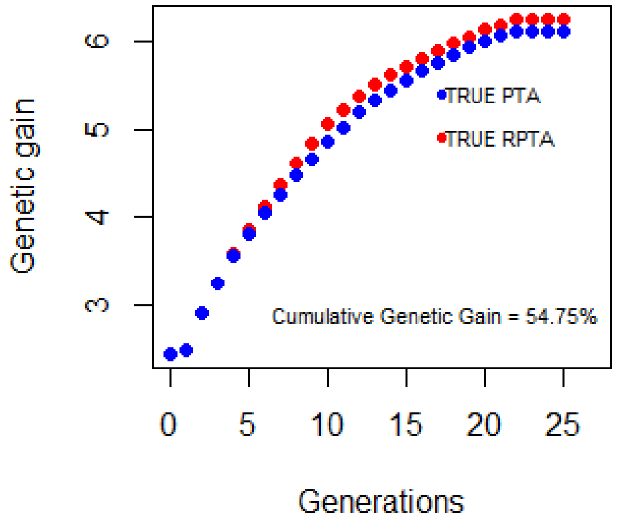


0.1 QTL/cM
σ^2_a=0.3(h^2=0.3)

ΔG=0.33%
7 generations

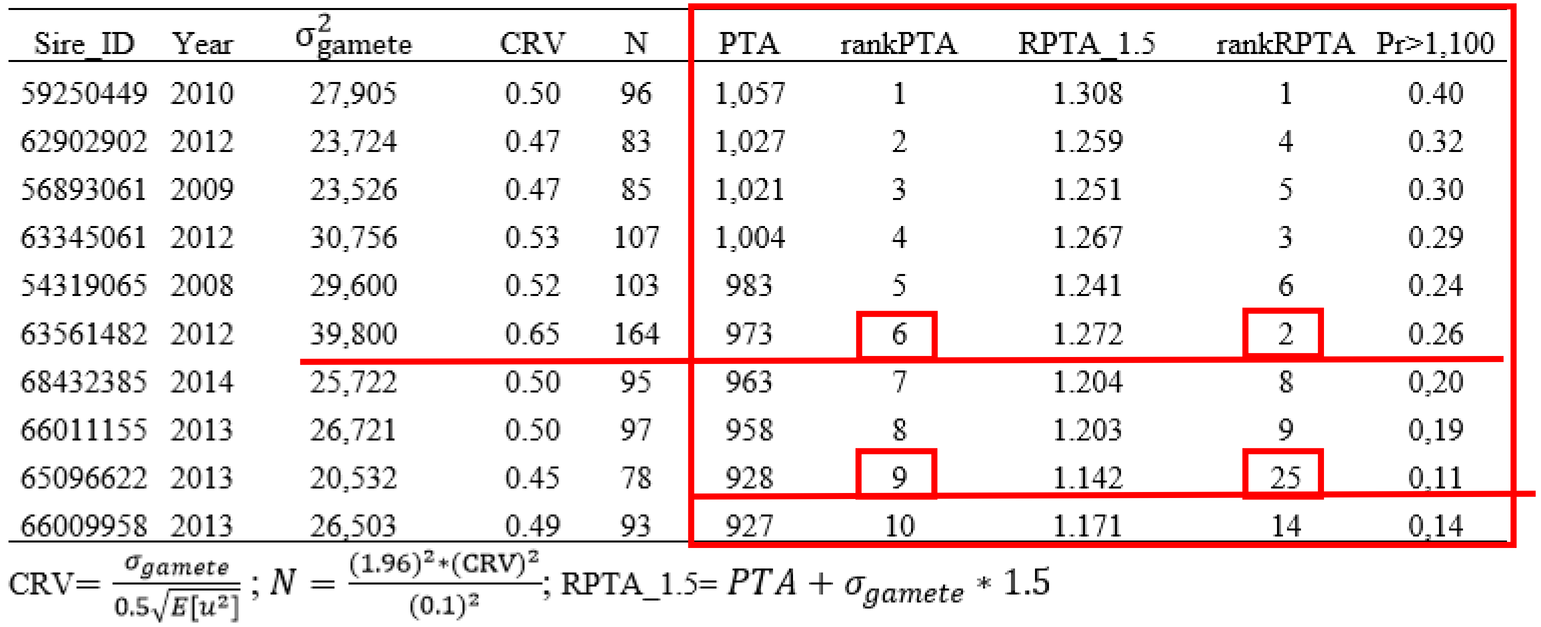
1 QTL/cM
σ^2_a=0.3(h^2=0.3)

ΔG=54.75%
25 generations



Applied example: USDA/Jersey

Genetic summary for Top 10 Sires for Milk Yield.



Sire_ID	Year	σ^2_{gamete}	CRV	N	PTA	rankPTA	RPTA_1.5	rankRPTA	Pr>1,100
59250449	2010	27,905	0.50	96	1,057	1	1.308	1	0.40
62902902	2012	23,724	0.47	83	1,027	2	1.259	4	0.32
56893061	2009	23,526	0.47	85	1,021	3	1.251	5	0.30
63345061	2012	30,756	0.53	107	1,004	4	1.267	3	0.29
54319065	2008	29,600	0.52	103	983	5	1.241	6	0.24
63561482	2012	39,800	0.65	164	973	6	1.272	2	0.26
68432385	2014	25,722	0.50	95	963	7	1.204	8	0,20
66011155	2013	26,721	0.50	97	958	8	1.203	9	0,19
65096622	2013	20,532	0.45	78	928	9	1.142	25	0,11
66009958	2013	26,503	0.49	93	927	10	1.171	14	0,14

CRV= $\frac{\sigma_{gamete}}{0.5\sqrt{E[u^2]}}$; $N = \frac{(1.96)^2 * (\text{CRV})^2}{(0.1)^2}$; RPTA_1.5= $PTA + \sigma_{gamete} * 1.5$