Single-step GBLUP for populations under selection and with multibreed data

Zulma Vitezica

UMR 1289, TANDEM, Castanet-Tolosan 31326 France

zulma.vitezica@ensat.fr

ALIMENTATION
AGRICULTURE
ENVIRONNEMENT

ssGBLUP

- ✓ A single-step approach was proposed based on **A** augmented with genomic information
- ✓ **Single step** has advantages such as simplicity, computational time and generality
- ✓ Most genotyped animals have undergone **strong selection**
  - → Single step needs adjustments (Aguilar et al., 2010, Chen et al., 2011, Forni et al. 2011)

Effect of different genomic relationship matrices (**G**) on accuracy (Chen et al., 2011)[1]

280k broiler chicken

4k genotyped with 60k SNPs

Single-step GBLUP (**ssGBLUP**)

G was constructed using equal (0.5, **Geq**) or current (**GC**) allele frequencies

[1] Chen et al., 2011. *J Anim Sci,* 89: 2673-2679.

Bias in genomic predictions

BLUP provides unbiased EBVs

EBVs are biased **upwards**

EBVs are biased **downwards**

	Means of EBV			
	Genotyped		All	
	BW	BM	BW	BM
BLUP	2.77	2.27	1.67	1.38
ssGBLUP				
GEq	3.94	3.21	2.51	2.15
GC	0.45	0.32	-0.05	-0.09

With bias, an accurate comparison of young and older animals is difficult !!

Bias in genomic predictions

A22 and **G** matrices and their inverses were compared with regard to **AvgD** and **AvgOff** elements

	AvgD	AvgOff
$\mathbf{A_{22}}$	1.004	0.014
G		
GC	1.017	0

Unbiased evaluation in the ssGBLUP was obtained by adding a constant to **G** (=0.014) to be compatible with $\mathbf{A_{22}}$

Data of litter size
(Forni et al., 2011)[1]

300k sows

2k genotyped with 60k SNPs

Single-step GBLUP (**ssGBLUP**)

Variance components estimation

[1] Fornin et al., 2011. *Gen Sel Evol, 43*: 1.

Estimates of additive variance

Relationship matrix	Full data set	Genotyped subset
Pedigree (**A**)	1.26 ±0.03	2.3 ±0.5
Genomic (**G**)		
Equal gene freq	similar	3.5 ±0.6
Average min allele freq	similar	3.5 ±0.6
Realized gene freq	similar	2.4 ±0.4
Normalized (**G**/trace(**G**))	similar	2.3 ±0.3

Realistic estimates of additive variance

Parameter estimation may be biased, especially when data sets were small

Statistics of different relationship matrices

Relationship matrix	Avg diagonal	Avg off-diagonal
Pedigree (**A**)	1.0	0.03
Genomic (**G**)		
Equal gene freq	1.25	0.59
Average min allele freq	1.70	1.02
Realized gene freq	0.94	0.00
Normalized (**G**/trace(**G**))	1.00	0.00

Parameter estimation may be biased if **G** is not compatible with **A**

Why are the GEBVs biased ?

- ✓ In traditional BLUP evaluation, all information about selection decisions is included in phenotypes and the relationship matrix (A), and **no bias exists from selection** (Sorensen & Kennedy, 1984)
- ✓ Models for genomic prediction implicitly assume an **unselected genotyped population** (Hayes et al. 2009).
- ✓ In practice, genotyped individuals are highly selected, and genomic prediction models do not take into account this selection

How unbiased GEBVs can be predicted ?

The single-step approach

It is based on the model $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{u} + e$

and $p(\mathbf{u}) \sim N(0, \mathbf{H}\sigma_u^2)$

involves the genetic effect for nongenotyped ($\mathbf{u}_1$) and genotyped ($\mathbf{u}_2$) individuals. Here $\mathbf{H}^{-1}$ is derived as

$$\mathbf{H}^{-1} = \begin{bmatrix} \mathbf{A}^{11} & \mathbf{A}^{12} \\ \mathbf{A}^{21} & \mathbf{A}^{22} + \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix}$$

where **G** is a genomic relationship matrix

Genomic relationship matrix **G**

Assume that **G** is computed according to VanRaden 2008, using observed allelic frequencies

This implies that the average BV of genotyped individuals ($\mathbf{u}_2$) is 0

$$p(\mathbf{u}_2) \sim N(0, \mathbf{G}\sigma_u^2)$$

This is possibly NOT the case if there is SELECTION

An improved **G**

Relative to the *pedigree base population*, the average BV of genotyped individuals ($\mathbf{u}_2$), has a value possibly different from 0, say μ

$$p(\mu) \sim N(0, \alpha\sigma_u^2)$$

μ is random because of finite size (drift)

$$p(\mathbf{u}_2|\mu) \sim N(\mathbf{1}\mu, \mathbf{G}\sigma_u^2)$$

$$p(\mathbf{u}_2|\alpha) \sim N(0, (\mathbf{G}+\mathbf{11}'\alpha)\sigma_u^2)$$

$$p(\mathbf{u}|\alpha) \sim N(0, \mathbf{H}^{\dagger}\sigma_u^2)$$

where $\mathbf{H}^{\dagger}$ is as **H**, but substituting **G** for **G** + **11'**α

An improved **G**

The mixed model equations are

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z}+\mathbf{H}^{\dagger -1}\lambda \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix}$$

$$\mathbf{H}^{\dagger -1} = \begin{bmatrix} \mathbf{A}^{11} & \mathbf{A}^{12} \\ \mathbf{A}^{21} & \mathbf{A}^{22}+\left(\mathbf{G}+\mathbf{1}\mathbf{1}'\alpha\right)^{-1}-\mathbf{A}_{22}^{-1} \end{bmatrix}$$

How to find the value for α ?

μ is the average BV of genotyped individuals

$$\mu = \frac{1}{n}\mathbf{1}'\mathbf{u}_2$$

Assume traditional BLUP is unbiased.

μ from either pedigree $\quad \mu_p \sim N\left(\mathbf{0}, \frac{1}{n^2}\mathbf{1}'\mathbf{A}_{22}\mathbf{1}\sigma_u^2\right)$

or single-step $\quad \mu_s \sim N\left(\mathbf{0}, \frac{1}{n^2}\mathbf{1}'\left[\mathbf{G}+\mathbf{1}\mathbf{1}'\alpha\right]\mathbf{1}\right)\sigma_u^2$

How to find the value for α ?

As the **1'1** are simply summations

from either pedigree

$$Var(\mu_p) = \sigma_u^2 \frac{1}{n^2} \sum_i \sum_j \mathbf{A}_{22\ (i,j)}$$

or single-step

$$Var(\mu_s) = \sigma_u^2 \left(\alpha + \frac{1}{n^2} \sum_i \sum_j \mathbf{G}_{i,j} \right)$$

If we equate both variances of μ →

$$\alpha = \frac{1}{n^2} \left[\sum_i \sum_j \mathbf{A}_{22\ (i,j)} - \sum_i \sum_j \mathbf{G}_{i,j} \right]$$

α is simply the difference between means for $\mathbf{A}_{22}$ and **G**

What does α mean ?

$$\alpha = \frac{1}{n^2}\left[\sum_i \sum_j \mathbf{A}_{22\ (i,j)} - \sum_i \sum_j \mathbf{G}_{i,j}\right]$$

α accounts for the fact that $\mathbf{u}_2$ are related through pedigree more than **G** is able to reflect

This is because we do not know base allelic frequencies to construct the 'correct' **G**

What does α mean ?

From Wright's F_{ST}, another interpretation of α is possible...

The F_{ST} can be defined as the mean relationship between gametes in a recent population with respect to an older base population

$$F_{old} = F_{new} + \left(1 - F_{new}\right) F_{ST}$$

Powell et al., 2010

What does α mean ?

$\mathbf{A}_{2\,2}$ involves relationships of genotyped individuals with reference to the base population, and

G corresponds to relationships within the current population.

Consequently, **α** is equal to twice F_{ST}

$$F_{ST} = \frac{1}{2} mean\left(\mathbf{A}_{22} - \mathbf{G}\right) \longrightarrow \mathbf{G}^{*} = \left(1 - \frac{1}{2}\alpha\right)\mathbf{G} + \mathbf{11}'\alpha$$

Which G must we use ?

AvgD(**G**)=AvgD(**A22**)

$$\mathbf{G}^{*}=\frac{trace\left(\mathbf{A}_{22}\right)}{trace\left(\mathbf{G}\right)}\mathbf{G}$$

2° moment (variance) of **u2**

sum(**G**)=sum(**A22**)

$$\mathbf{G}^{*}=\mathbf{G}+\mathbf{1}\mathbf{1}'\alpha$$

1° moment (mean) of **u2**

Both,
AvgD(**G**)=AvgD(**A22**)
sum(**G**)=sum(**A22**)

$$\mathbf{G}^{*}=\left(1-\frac{1}{2}\alpha\right)\mathbf{G}+\mathbf{1}\mathbf{1}'\alpha$$

Mean & variance of **u2**
(assumption of random mating)

Which G must we use ?

Both,
AvgD(**G**)=AvgD(**A22**)
sum(**G**)=sum(**A22**)

$$\mathbf{G}^{*}=\left(1-\frac{1}{2}\alpha\right)\mathbf{G}+\mathbf{11}'\alpha$$

Mean & variance of **u2** (assumption of random mating)

$$\mathbf{G}^{*}=a\,\mathbf{G}+\mathbf{11}'b$$

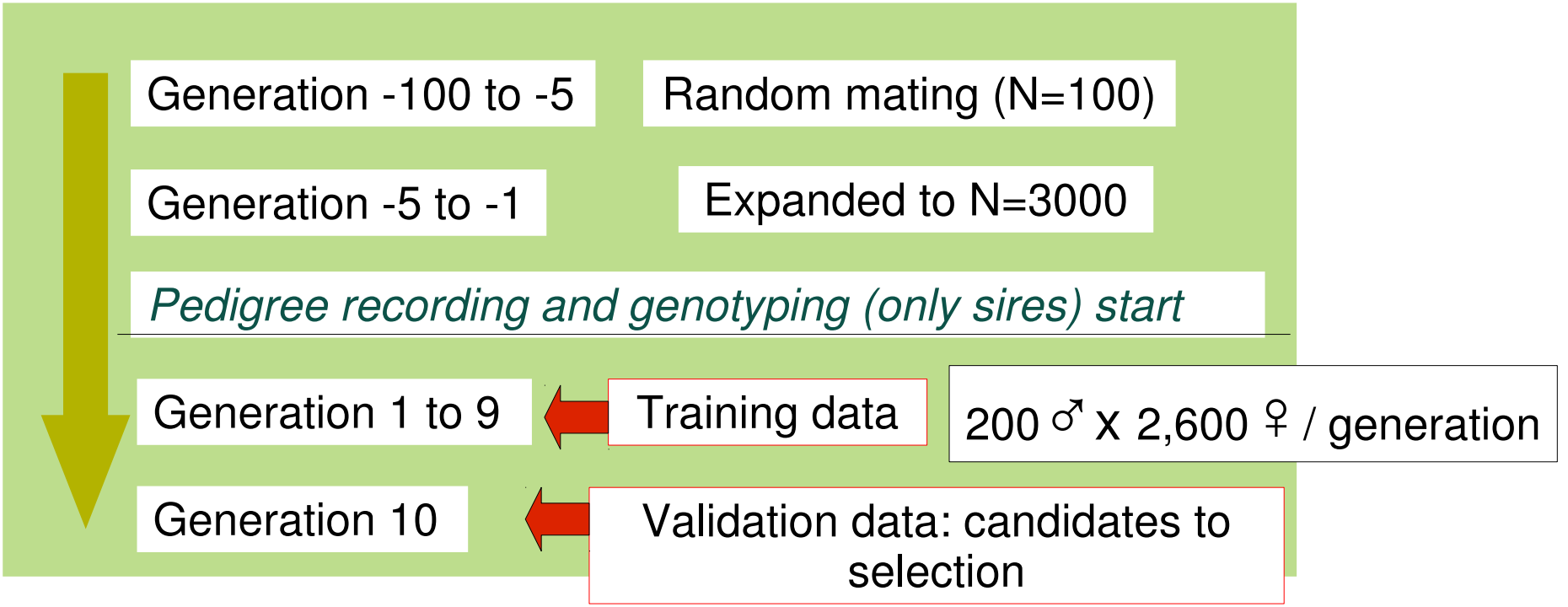
Mean & variance of **u2** (no random mating)
preGSf90

Christensen et al., 2012,

a = 0.859 (Christensen et al., 2012)
a = 0.851 (Vitezica et al., 2011)

Simulation (QMSim, 2009)

- ✓ 10 K SNPs - 10 chrom. (10 SNPs/cM)
- ✓ Sex-limited trait (h^2 : 0.5 – 0.3 – 0.05)
- ✓ Selection: phenotype / EBV



Genetic evaluation

EBVs of the candidates to selection (last generation) were estimated

1) **BLUP$_{PED}$** pedigree-based evaluation

2) **BLUP$_{DYD}$** 2-step procedure
BLUP$_{PED}$ -> DYD's for sires -> GBLUP

3) **BLUP1$_{STEP}$** single-step (SS) procedure, standard **G**

4) **BLUP$_{ALPHA}$** SS procedure, $\mathbf{G}^{\dagger} = \mathbf{G} + \mathbf{11}' \boldsymbol{\alpha}$

5) **BLUP$_{F_{ST}}$** SS with **G*** proposed by Powell *et al.*, 2010

Results

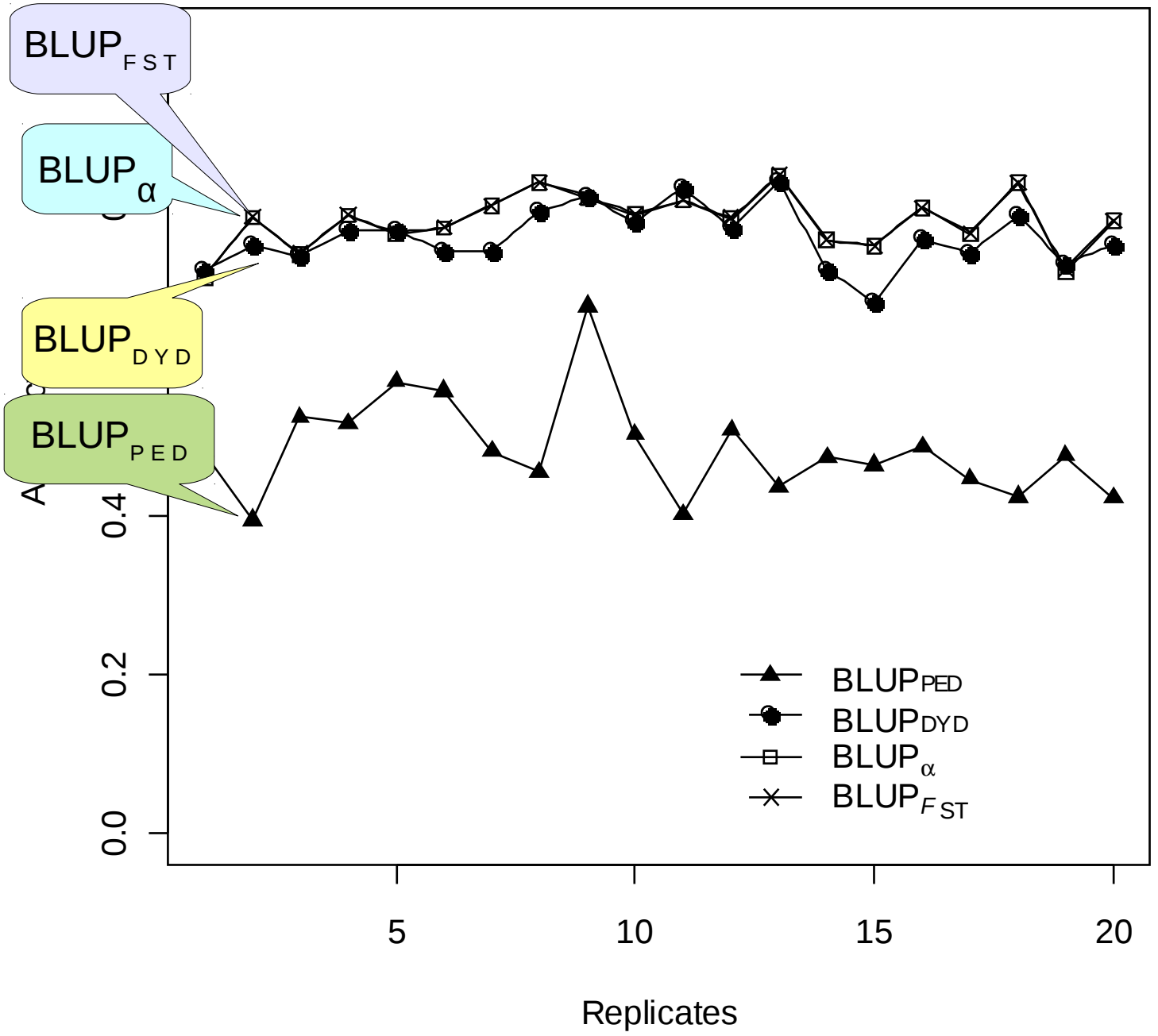
Means of EBV (h^2=0.30) for selection candidates

Prediction method	Py	PEBV
	TBV=0.53 (0.03)	TBV=2.01 (0.15)
BLUP$_{PED}$	0.54	2.05
BLUP$_{DYD}$	0.17	0.61
BLUP$_{1STEP}$	0.29	1.41
BLUP$_{ALPHA}$	0.52	2.10
BLUP$_{Fst}$	0.52	2.10

Sd err ~0.04

Sd err ~0.15

Results



Accuracy of different genomic prediction methods across 20 replicates

h^2=0.30

P_{EBV} selection

Conclusions

- ✔ Single-step method with correction (either BLUP$_{FST}$ or BLUP$_{Chistensen\ et\ al.\ 2012}$) is the recommended method for accounting for bias in genomic predictions
- ✔ Bias is an important concern with 2-step procedure (BLUP$_{DYD}$)

What would happen with GEBV in a joint evaluation of multiple lines or breeds ?

Evaluation of a multi-line broiler chicken population using ssGBLUP

Simeone et al., 2011. JABG.

Lines were combined to create a multi-line pop

Table 1 Descriptions of the phenotypic data for body weight for all animals and genotyped animals in lines A and B, and the multi-line population[1]

Line	No. of records	Mean (SD)
A		
All animals	183 695	24.50 (3.22)
Genotyped animals[2]	3195	25.12 (2.97)
B		
All animals	164 149	23.53 (3.17)
Genotyped animals	3001	23.39 (2.63)
Multi-Line[3]		
All animals	347 844	24.04 (3.24)
Genotyped animals	6196	24.28 (2.94)

The purpose was to examine changes of GEBVs in analysis of two lines when **G** was constructed with different allele frequencies

Table 2 Statistics for GEBVs and EBVs estimated for the multi-line population[1]

Population	Mean (G)EBV (SD)	Minimum	Maximum
Line A, traditional	−0.14 (0.47)	−1.32	1.01
Line A, SSP[2] (A)	0.07 (0.59)	−1.71	1.70
Multi (A)	0.07 (0.60)	−1.75	1.76
Multi (B)	−0.12 (0.60)	−1.96	1.54
Multi (AB)[3]	0.13 (0.58)	−1.64	1.70
Multi (0.5)	−0.18 (0.53)	−1.81	1.29
Line B, traditional	−0.28 (0.34)	−1.31	0.69
Line B, SSP (B)	0.00 (0.47)	−1.51	1.31
Multi (A)	−0.25 (0.47)	−1.74	0.93
Multi (B)	0.02 (0.49)	−1.54	1.25
Multi (AB)	−0.06 (0.47)	−1.49	1.11
Multi (0.5)	−0.34 (0.42)	−1.64	0.76

* Frequencies used to scale G are in parentheses

GEBV predicted using the correct allele frequencies for each line

The means of GEBV changed with different **G**, but animals were ranked appropriately within each line

In ssGBLUP, **G** may be used to evaluate simultaneously multiple lines or breeds, but should be adjusted to obtain proper GEBVs

Multi-breed evaluation with ssGBLUP in New Zealand

Harris et al., 2012.

Base Adjustment Method

Computing $\mathbf{G}$

- Extended the correction suggested by Vitezica et al. (2011) for multi-breed genomic relationship matrices
- Similar to multi-breed mean adjustments in Harris and Johnson (2010)
- Computationally easier than current adjustments

$$\alpha_{ij} = \mu + \alpha^H p_i^H p_j^H + \alpha^J p_i^J p_j^J + \alpha^X (p_i^H p_j^J + p_j^H p_i^J)$$

Interbull bulletin, n°46, Cork Ireland, May 28 – 31, 2012.