Peeters et al. Genetics Selection Evolution 2013, 45:27
http://www.gsejournal.org/content/45/1/27

GSE Genetics Selection Evolution

RESEARCH Open Access

Using pooled data to estimate variance components and breeding values for traits affected by social interactions

Katrijn Peeters*, Esther Dorien Ellen and Piter Bijma

Abstract

Background: Through social interactions, individuals affect one another's phenotype. In such cases, an individual's phenotype is affected by the direct (genetic) effect of the individual itself and the indirect (genetic) effects of the group mates. Using data on individual phenotypes, direct and indirect genetic (co)variances can be estimated. Together, they compose the total genetic variance that determines a population's potential to respond to selection. However, it can be difficult or expensive to obtain individual phenotypes. Phenotypes on traits such as egg production and feed intake are, therefore, often collected on group level. In this study, we investigated whether direct, indirect and total genetic variances, and breeding values can be estimated from pooled data (pooled by group). In addition, we determined the optimal group composition, *i.e.* the optimal number of families represented in a group to minimise the standard error of the estimates.

Methods: This study was performed in three steps. First, all research questions were answered by theoretical derivations. Second, a simulation study was conducted to investigate the estimation of variance components and optimal group composition. Third, individual and pooled survival records on 12 944 purebred laying hens were analysed to investigate the estimation of breeding values and response to selection.

Results: Through theoretical derivations and simulations, we showed that the total genetic variance can be estimated from pooled data, but the underlying direct and indirect genetic (co)variances cannot. Moreover, we showed that the most accurate estimates are obtained when group members belong to the same family. Additional theoretical derivations and data analyses on survival records showed that the total genetic variance and breeding values can be estimated from pooled data. Moreover, the correlation between the estimated total breeding values obtained from individual and pooled data was surprisingly close to one. This indicates that, for survival in purebred laying hens, loss in response to selection will be small when using pooled instead of individual data.

Conclusions: Using pooled data, the total genetic variance and breeding values can be estimated, but the underlying genetic components cannot. The most accurate estimates are obtained when group members belong to the same family.

Background

Group housing is common practice in most livestock farming systems. Previous studies have shown that group-housed animals can substantially affect one another's phenotype through social interactions [1-9]. The heritable effect of an individual on its own phenotype is known as the direct genetic effect, while the heritable effect of an individual on the phenotype of a group mate is known as the social, associative or indirect genetic effect [10-14]. Both direct and indirect genetic effects determine a population's potential to respond to selection, *i.e.* the total genetic variance [2,10-14]. Selection experiments in laying hens and quail [1,2,9], and variance component estimates in laying hens, quail, beef cattle and pigs [3-9] have shown that indirect genetic effects can contribute substantially to the total genetic variation in agricultural populations.

Direct, indirect and total genetic variances can be estimated from individual data. However, it can be difficult

* Correspondence: katrijn.peeters@wur.nl
Animal Breeding and Genomics Centre, Wageningen University, P.O. Box 338, 6700 AH Wageningen, The Netherlands

or expensive to obtain individual phenotypes on certain traits, *e.g.* egg production and feed intake. Alternatively, data can be obtained on group level, resulting in pooled records. However, pooling data reduces the number of data points. Moreover, multiple animals influence each data point, increasing the complexity of the data. Although there is an obvious loss of power, previous studies have shown that pooled data can be used to estimate direct genetic variances for traits not affected by social interactions [15-17]. However, with social interactions, indirect genetic effects emerge and the complexity of the data increases further. It is unclear whether pooled data are still informative in these situations. Therefore, the main objective of this study was to determine whether pooled data can be used to estimate direct, indirect and total genetic variances, and breeding values for traits affected by social interactions. In addition, optimal group composition was determined, *i.e.* the optimal number of families represented in a group to minimise the standard error of the estimates.

Methods

This study was performed in three steps. First, all research questions were answered by theoretical derivations. Second, a simulation study was conducted to investigate the estimation of variance components and optimal group composition. Third, individual and pooled survival records on 12 944 purebred laying hens were analysed to investigate the estimation of breeding values and response to selection.

Table 1 lists the main symbols and their meaning.

Theory

Variance components and breeding value estimation

In this section, we examined whether direct, indirect and total genetic variances, and breeding values can be estimated from pooled data.

With social interactions, an individual phenotype consists of the direct genetic (A_D) and environmental (E_D) effects of the individual itself (i), and the indirect genetic (A_I) and environmental (E_I) effects of its group mates (j):

$$P_i = A_{D_i} + E_{D_i} + \sum_{i \neq j}^{n-1} A_{I_j} + \sum_{i \neq j}^{n-1} E_{I_j}, \quad (1)$$

where n is the number of individuals per group [11]. From an animal breeding perspective, the total breeding value (A_T) is of interest because it determines total response to selection. An animal's A_T consists of a direct and indirect component:

$$A_{T_i} = A_{D_i} + (n\text{-}1)A_{I_i}, \quad (2)$$

where A_D is expressed in the phenotype of the animal itself and A_I is expressed in the phenotype of each group mate.

Table 1 Notation key

Symbol	Meaning
i – j	Focal individual - Group mates of the focal individual
A_D	Direct genetic effect \ Direct breeding value
A_I	Indirect genetic effect \ Indirect breeding value
A_T	Total genetic effect \ Total breeding value
E_D	Direct environmental effect
E_I	Indirect environmental effect
$\sigma^2_{A_D}$	Direct genetic variance
$\sigma_{A_{DI}}$	Direct–indirect genetic covariance
$\sigma^2_{A_I}$	Indirect genetic variance
$\sigma^2_{A_T}$	Total genetic variance
σ^2_{Cage}	Cage variance
σ^2_E	Error variance
σ^2_P	Phenotypic variance
$\sigma^2_{E^*}$	Pooled error variance
$\sigma^2_{P^*}$	Pooled phenotypic variance
h^2	Direct genetic variance relative to phenotypic variance \ Heritability
T^2	Total genetic variance relative to phenotypic variance
σ^2_z	Full variance
σ^2_b	Between-family variance
σ^2_w	Within-family variance
r	Relatedness within a family
N	Number of families
m	Number of records per family
o	Family size
n	Group size
^	Hat, denotes estimated values

A pooled record (P^*) consists of the individual phenotypes of all group members (k):

$$P^* = \sum_{k=1}^{n} P_k. \quad (3)$$

It follows from Equations (1) and (3) that, with social interactions, a pooled record consists of the A_D and E_D of each group member, as well as their A_I and E_I that are expressed $n - 1$ times:

$$P^* = \sum_{k=1}^{n} [A_{D_k} + E_{D_k} + (n\text{-}1)(A_{I_k} + E_{I_k})]. \quad (4)$$

Because an animal's A_D and A_I are expressed in the same pooled record, the direct **Z**-matrix that links pooled phenotypes to A_D's and the indirect **Z**-matrix that links pooled phenotypes to A_I's are completely confounded (as shown in Appendix A by using a fictive

example (Table 8)). Consequently, direct and indirect (co)variances, and breeding values cannot be estimated from pooled data.

It follows from Equations (2) and (4) that, with social interactions, a pooled record contains the total genetic effect of each group member:

$$P^* = \sum_{k=1}^{n} (A_{T_k} + E_k). \quad (5)$$

Equation (5) shows strong similarities with:

$$P^* = \sum_{k=1}^{n} (A_{D_k} + E_k), \quad (6)$$

which shows the content of a pooled record when social interactions do not occur. Previous studies have shown that pooled data can be used to estimate direct genetic variances ($\sigma^2_{A_D}$) and direct breeding values for traits that are not affected by social interactions [15-17]. Similarly, pooled data can be used to estimate total genetic variances ($\sigma^2_{A_T}$) and total breeding values for traits that are affected by social interactions.

Optimal group composition

In this section, the standard error (s.e.) of $\hat{\sigma}^2_{A_T}$ is derived for three experimental designs that differ with respect to group composition, *i.e.* group members belonged to either one, two or *n* families. The s.e. of an estimate of the genetic variance depends on the between- $\left(\sigma^2_b\right)$ and within-family variance $\left(\sigma^2_w\right)$, the relatedness within a family (*r*), the number of families (*N*), and the number of records per family (*m*) [18]:

$$\text{s.e.}\left(\hat{\sigma}^2_A\right) \approx \frac{1}{r}\sqrt{\frac{2}{N\text{-}1}\left[\sigma^4_b + \frac{2\sigma^2_b\sigma^2_w}{m} + \frac{\sigma^4_w}{m(m\text{-}1)}\right]}. \quad (7)$$

Analysis of variance was used to derive σ^2_b and σ^2_w for each design (see Appendix B for derivation).

The s.e. of $\hat{\sigma}^2_{A_T}$ differs between experimental designs because the group composition changes the within-family variance and the number of records per family (Table 2). On the one hand, the within-family variance decreases when the number of families per group decreases, causing a strong decrease in s.e.. On the other hand, the number of records per family decreases when the number of families per group decreases, causing a slight increase in s.e.. Overall, to obtain the most accurate estimate of $\sigma^2_{A_T}$, group members should belong to the same family. The only exception is when family size (*o*) equals group size (*n*). In this case, there is only one record per family and $\sigma^2_{A_T}$ would not be estimable.

Table 2 Within-family variance (σ^2_w) and number of records per family (*m*) for three group compositions

	σ^2_w	*m*
One family	$\frac{1}{n}\left(\sigma^2_{P_D} + 2(n-1)\sigma_{P_{DI}} + (n-1)^2\,\sigma^2_{P_I} + (n-1)r\sigma^2_{A_T}\right) - r\sigma^2_{A_T}$	*o/n*
Two families	$\frac{4}{n}\left(\sigma^2_{P_D} + 2(n-1)\sigma_{P_{DI}} + (n-1)^2\,\sigma^2_{P_I} + \left(\frac{n}{2}-1\right)r\sigma^2_{A_T}\right) - r\sigma^2_{A_T}$	*2o/n*
n families	$n\left(\sigma^2_{P_D} + 2(n-1)\,\sigma_{P_{DI}} + (n-1)^2\sigma^2_{P_I}\right) - r\sigma^2_{A_T}$	*o*

r, *N*, *n*, *o* and σ^2_b do not differ between group compositions.

Ideally, group members should be full sibs rather than half sibs, since an increase in relatedness causes a decrease in the s.e. of $\hat{\sigma}^2_{A_T}$.

Simulation

To validate the theoretical derivations, a simulation study was conducted in R v2.12.2 [19]. A base population of 500 sires and 500 dams was simulated. Each animal in the base population was assigned a direct and indirect breeding value, drawn from $N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\sigma^2_{A_D} & \sigma_{A_{DI}}\\ \sigma_{A_{DI}} & \sigma^2_{A_I}\end{bmatrix}\right)$. The $\sigma^2_{A_D}$ and $\sigma^2_{A_I}$ were set to 1.00, and $\sigma_{A_{DI}}$ was set to –0.50, 0.00 or 0.50. Each sire was randomly mated to a single dam, resulting in 12 offspring per mating for a total of 6000 simulated offspring. For each offspring, direct and indirect breeding values were obtained as: $A_D = \frac{1}{2}A_{D_S} + \frac{1}{2}A_{D_D} + MS_D$ and $A_I = \frac{1}{2}A_{I_S} + \frac{1}{2}A_{I_D} + MS_I$, where the direct and indirect Mendelian sampling terms were drawn from $N\left(\begin{bmatrix}0\\0\end{bmatrix}, \frac{1}{2}\begin{bmatrix}\sigma^2_{A_D} & \sigma_{A_{DI}}\\ \sigma_{A_{DI}} & \sigma^2_{A_I}\end{bmatrix}\right)$. Each offspring was also assigned a direct and indirect environmental value, drawn from $N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\sigma^2_{E_D} & \sigma_{E_{DI}}\\ \sigma_{E_{DI}} & \sigma^2_{E_I}\end{bmatrix}\right)$. The $\sigma^2_{E_D}$ and $\sigma^2_{E_I}$ were set to 2.00, and $\sigma_{E_{DI}}$ was set to –1.00, 0.00 or 1.00. Animals were placed in groups of four. Depending on the scenario, group members belonged to one, two or four families. Individual phenotypes were obtained by summing the direct and indirect genetic and environmental components according to Equation (1). Pooled records were obtained by summing individual phenotypes according to Equation (3). Seven scenarios were simulated, which differed in $\sigma_{A_{DI}}$, $\sigma_{E_{DI}}$ or group composition (Table 3). For each scenario, 100 replicates were produced.

Based on the previous section, expectations are that the use of a direct–indirect animal model for pooled data will fail to differentiate between direct and indirect genetic effects, while the use of a traditional animal model for pooled data will yield estimates of $\sigma^2_{A_T}$. To validate these theoretical predictions, both models were run. First, the simulated pooled records were analysed

Table 3 Scenarios used to simulate data

	Scenario[§]	$\sigma_{A_{DI}}$	$\sigma_{E_{DI}}$	Group composition
Reference scenario	1	0.00	0.00	Four families
Different $\sigma_{A_{DI}}$	2	−0.50	0.00	Four families
	3	0.50	0.00	Four families
Different $\sigma_{E_{DI}}$	4	0.00	−1.00	Four families
	5	0.00	1.00	Four families
Different group compositions	6	0.00	0.00	Two families
	7	0.00	0.00	One family

[§] $\sigma^2_{A_D}$ and $\sigma^2_{A_I}$ were set to 1.00; $\sigma^2_{E_D}$ and $\sigma^2_{E_I}$ were set to 2.00.

with the following direct–indirect animal model in ASReml v3.0 [20]:

$$\mathbf{y}^* = \boldsymbol{\mu}^* + \mathbf{Z}^*_{\mathrm{D}}\mathbf{a}_{\mathrm{D}} + \mathbf{Z}^*_{\mathrm{I}}\mathbf{a}_{\mathrm{I}} + \mathbf{e}^*, \quad (8)$$

where $\mathbf{y}^*$ is a vector that contains pooled records (P^*); $\boldsymbol{\mu}^*$ is a vector that contains the pooled mean; $\mathbf{Z}^*_{\mathrm{D}}$ is an incidence matrix linking the pooled records to A_{D}'s (each pooled record was linked to the A_{D}'s of the four group members); $\mathbf{a}_{\mathrm{D}}$ is a vector that contains A_{D}'s; $\mathbf{Z}^*_{\mathrm{I}}$ is an incidence matrix linking the pooled records to A_{I}'s (each pooled record was linked to the A_{I}'s of the four group members); $\mathbf{a}_{\mathrm{I}}$ is a vector that contains A_{I}'s; and $\mathbf{e}^*$ is a vector that contains residuals. Second, the simulated pooled records were analysed with the following traditional animal model in ASReml v3.0 [20]:

$$\mathbf{y}^* = \boldsymbol{\mu}^* + \mathbf{Z}^*\mathbf{a} + \mathbf{e}^*, \quad (9)$$

where $\mathbf{y}^*$, $\boldsymbol{\mu}^*$ and $\mathbf{e}^*$ are as explained above; $\mathbf{Z}^*$ is an incidence matrix linking the pooled records to A's (each pooled record was linked to the A's of the four group members); and $\mathbf{a}$ is a vector that contains A's.

Based on the previous section, expectations are that the most accurate prediction of $\sigma^2_{A_{\mathrm{T}}}$ will be obtained when group members belong to the same family. To validate this theoretical prediction, the predicted s.e. of $\hat{\sigma}^2_{A_{\mathrm{T}}}$ was compared to (i) the standard deviation (s.d.) of 100 estimates of $\sigma^2_{A_{\mathrm{T}}}$ ($\hat{\sigma}^2_{A_{\mathrm{T}}}$'s reported by ASReml) and (ii) the mean of 100 s.e.'s of $\hat{\sigma}^2_{A_{\mathrm{T}}}$ (s.e.'s reported by ASReml) for three group compositions (scenarios 1, 6 and 7 of Table 3).

Data analyses

The dataset was part of the pre-existing database of Hendrix Genetics (The Netherlands) and contained routinely collected data for breeding value estimation. Animal Care and Use Committee approval was therefore not required.

To validate the theoretical derivations and to gain insight into response to selection, individual and pooled data on survival in purebred laying hens (*Gallus gallus*) were analysed. Survival in group-housed laying hens is a well-known example of a trait affected by social interactions, since a bird's chance to survive depends on the feather pecking and cannibalistic behaviour of its group mates. Ellen et al. [5] used individual survival data on three purebred lines to estimate direct and indirect genetic (co)variances. Large and statistically significant indirect genetic effects were found in two out of three purebred lines. In the current study, we used data from the same two lines. Data were provided by the "Institut de Sélection Animale B.V.", the layer breeding division of Hendrix Genetics. Data on 13 192 White Leghorn layers were provided of which 6276 were of line W1 and 6916 were of line WB.

At the age of 17 weeks, the hens were placed in two laying houses. The laying houses consisted of four or five double rows, and each row consisted of three levels. Interaction with neighbours on the back of the cage was possible, but interaction with neighbours on the side was prevented. Four hens of the same purebred line were randomly assigned to each cage. Hens were not beak-trimmed. Further details on housing conditions and management are in Ellen et al. [5].

The individual phenotype was defined as the number of days from the start of the laying period until either death or the end of the experiment, with a maximum of 398 days. The individual phenotypes were summed per cage to obtain pooled records. If one individual phenotype was missing, the entire cage was omitted from the analysis. The final dataset contained records on 6092 W1 and 6852 WB hens.

To obtain the direct, indirect and total genetic parameters for survival time, the individual phenotypes were analysed with the following direct–indirect animal model in ASReml v3.0 [20]:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}_{\mathrm{D}}\mathbf{a}_{\mathrm{D}} + \mathbf{Z}_{\mathrm{I}}\mathbf{a}_{\mathrm{I}} + \mathbf{V}\mathbf{cage} + \mathbf{e}, \quad (10)$$

where $\mathbf{y}$ is a vector that contains individual phenotypes; $\mathbf{X}$ is an incidence matrix linking the individual phenotypes to fixed effects; $\mathbf{b}$ is a vector that contains fixed effects, which included an interaction term for each laying house by row by level combination, an effect for the content of the back cage (full/empty) and a covariate for the average number of survival days in the back cage; $\mathbf{Z}_{\mathrm{D}}$ is an incidence matrix linking the individual phenotypes to A_{D}'s; $\mathbf{a}_{\mathrm{D}}$ is a vector that contains A_{D}'s; $\mathbf{Z}_{\mathrm{I}}$ is an incidence matrix linking the individual phenotypes to A_{I}'s; $\mathbf{a}_{\mathrm{I}}$ is a vector that contains A_{I}'s; $\mathbf{V}$ is an incidence matrix linking the individual phenotypes to random cage effects; $\mathbf{cage}$ is a vector that contains random cage effects (to account for the non-genetic covariance among phenotypes of cage members [21]); and $\mathbf{e}$ is a vector that contains residuals. This model yields estimates of $\sigma^2_{A_{\mathrm{D}}}$,

$\sigma_{A_{DI}}$ and $\sigma^2_{A_I}$, from which $\hat{\sigma}^2_{A_T}$ can be calculated. Similarly, it yields estimates of A_D's and A_I's, from which $\hat{A}_T$'s can be calculated. To improve a trait, animals should be selected based on their $\hat{A}_T$, since $\sigma^2_{A_T}$ determines a population's potential to respond to selection.

Alternatively, a traditional animal model can be used to analyse individual or pooled data. A traditional animal model on individual data only yields estimates of $\sigma^2_{A_D}$ and A_D's. A traditional model on pooled data is expected to yield estimates of $\sigma^2_{A_T}$ and A_T's, but not of $\sigma^2_{A_D}$ and A_D's. To validate this theoretical prediction, these traditional models were also run. First, the individual phenotypes were analysed with the following traditional (direct) animal model in ASReml v3.0 [20]:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Z}_D\mathbf{a}_D + \mathbf{Vcage} + \mathbf{e}, \quad (11)$$

where **y**, **X**, **b**, $\mathbf{Z}_D$, $\mathbf{a}_D$, **V**, **cage** and **e** are as explained above. Second, the pooled records were analysed with the following traditional animal model in ASReml v3.0 [20]:

$$\mathbf{y}^* = \mathbf{X}^*\mathbf{b}^* + \mathbf{Z}^*\mathbf{a} + \mathbf{e}^*, \quad (12)$$

where $\mathbf{y}^*$ is a vector that contains pooled records (P^*); $\mathbf{X}^*$ is an incidence matrix linking the pooled records to fixed effects; $\mathbf{b}^*$ is a vector that contains fixed effects (the same fixed effects as mentioned above); $\mathbf{Z}^*$ is an incidence matrix linking the pooled records to *A*'s (each pooled record was linked to the *A*'s of the four group members); **a** is a vector that contains *A*'s; and $\mathbf{e}^*$ is a vector that contains residuals.

The estimated variance components and breeding values of all three models were compared. In addition, we calculated the loss in response to selection that would occur when applying a traditional model to individual or pooled data instead of a direct–indirect model to individual data. The direct–indirect model applied to individual data yielded estimates of $\sigma^2_{A_T}$ and A_T's. Based on their $\hat{A}_T$, 250 animals were selected and the corresponding response to selection was calculated. Similarly, for the two traditional animal models, 250 animals were selected based on their $\hat{A}_D$ (obtained from individual data) and $\hat{A}$ (obtained from pooled data). Once the top 250 animals were selected, their $\hat{A}_T$ (obtained from individual data) was used to calculate the total response to selection. Then, the loss in total response to selection was calculated.

Results and discussion

Simulation

The direct–indirect animal model on pooled records failed to converge, confirming that direct and indirect (co)variances cannot be estimated from pooled data. The traditional animal model on pooled records yielded estimates of σ^2_A and $\sigma^2_{E^*}$. These estimates did not differ significantly from the true $\sigma^2_{A_T}$ and $\sigma^2_{E^*}$ (Table 4), where

$$\sigma^2_{A_T} = \sigma^2_{A_D} + 2(n\text{–}1)\,\sigma_{A_{DI}} + (n\text{–}1)^2\,\sigma^2_{A_I} \quad (13)$$

(derived by [14]) and

$$\sigma^2_{E^*} = n\left[\sigma^2_{E_D} + 2(n\text{–}1)\,\sigma_{E_{DI}} + (n\text{–}1)^2\sigma^2_{E_I}\right] \quad (14)$$

(analogous to [17]).

Based on Equation (7), the s.e. of $\hat{\sigma}^2_{A_T}$ was predicted for three scenarios that differed in group composition, *i.e.* group members belonged to one, two or four families. The theoretical s.e. of $\hat{\sigma}^2_{A_T}$ was compared to (i) the s.d. of 100 estimates of $\sigma^2_{A_T}$ ($\hat{\sigma}^2_{A_T}$'s reported by ASReml) and (ii) the mean of 100 s.e.'s of $\hat{\sigma}^2_{A_T}$ (s.e.'s reported by ASReml) (Table 5). The theoretical s.e. of $\hat{\sigma}^2_{A_T}$ did not

Table 4 True and estimated $\sigma^2_{A_T}$ and $\sigma^2_{E^*}$ for five scenarios

	Scenario[§]	$\sigma^2_{A_T}$ [§§]	$\overline{\hat{\sigma}^2_A \pm \text{s.e.}}$	$\sigma^2_{E^*}$ [§§§]	$\overline{\hat{\sigma}^2_{E^*} \pm \text{s.e.}}$
$\sigma_{A_{DI}} = 0.00$ $\sigma_{E_{DI}} = 0.00$	1	10.00	10.10 ± 1.85	80.00	80.56 ± 6.69
$\sigma_{A_{DI}} = -0.50$ $\sigma_{E_{DI}} = 0.00$	2	7.00	7.43 ± 1.59	80.00	79.29 ± 6.08
$\sigma_{A_{DI}} = 0.50$ $\sigma_{E_{DI}} = 0.00$	3	13.00	13.05 ± 2.12	80.00	80.32 ± 7.30
$\sigma_{A_{DI}} = 0.00$ $\sigma_{E_{DI}} = -1.00$	4	10.00	9.70 ± 1.54	56.00	56.54 ± 5.24
$\sigma_{A_{DI}} = 0.00$ $\sigma_{E_{DI}} = 1.00$	5	10.00	9.81 ± 2.10	104.00	104.71 ± 8.03

[§] $\sigma^2_{A_D}$ and $\sigma^2_{A_I}$ were set to 1.00; $\sigma^2_{E_D}$ and $\sigma^2_{E_I}$ were set to 2.00; group members belonged to four different families.
[§§] $\sigma^2_{A_T} = \sigma^2_{A_D} + 2(n\text{–}1)\,\sigma_{A_{DI}} + (n\text{–}1)^2\,\sigma^2_{A_I}$.
[§§§] $\sigma^2_{E^*} = n\left[\sigma^2_{E_D} + 2(n\text{–}1)\,\sigma_{E_{DI}} + (n\text{–}1)^2\,\sigma^2_{E_I}\right]$.

Table 5 Theoretically predicted s.e.$\left(\hat{\sigma}^2_{A_T}\right)$, s.d.$\left(\hat{\sigma}^2_{A_T}\right)$ [§] and $\overline{\text{s.e.}\left(\hat{\sigma}^2_{A_T}\right)}$ [§§] for three group compositions

	Scenario[§§§]	s.e.$\left(\hat{\sigma}^2_{A_T}\right)$	s.d.$\left(\hat{\sigma}^2_{A_T}\right)$±s.d.	$\overline{\text{s.e.}\left(\hat{\sigma}^2_{A_T}\right)}$±s.d.
Four families	1	1.88	2.01 ± 0.14	1.85 ± 0.13
Two families	6	1.30	1.23 ± 0.09	1.23 ± 0.08
One family	7	0.92	0.81 ± 0.06	0.92 ± 0.05

[§] s.d.$\left(\hat{\sigma}^2_{A_T}\right)$ based on 100 $\hat{\sigma}^2_{A_T}$'s reported by ASReml.

[§§] $\overline{\text{s.e.}\left(\hat{\sigma}^2_{A_T}\right)}$ based on 100 s.e.'s reported by ASReml.

[§§§] $\sigma^2_{A_D}$ and $\sigma^2_{A_I}$ were set to 1.00; $\sigma_{A_{DI}}$ was set to 0.00; $\sigma^2_{E_D}$ and $\sigma^2_{E_I}$ were set to 2.00; $\sigma_{E_{DI}}$ was set to 0.00.

differ significantly from the values obtained by simulation. Moreover, as predicted, the most accurate estimate of $\sigma^2_{A_T}$ was obtained when group members belonged to the same family. In comparison, the s.e. of $\hat{\sigma}^2_{A_T}$ was twice as large when group members belonged to different families. This indicates that group composition is crucial when aiming to obtain accurate estimates.

Data analyses

Table 6 shows the estimated variance components for individual survival data analysed with a direct–indirect animal model, and the estimated variance components for individual and pooled survival data analysed with a traditional animal model. The direct–indirect animal model on individual data yielded estimates of $\sigma^2_{A_D}$, $\sigma_{A_{DI}}$ and $\sigma^2_{A_I}$. Based on these components, $\hat{\sigma}^2_{A_T}$ was calculated (according to Equation (13)). The traditional animal model on individual data yielded estimates of $\sigma^2_{A_D}$. The traditional animal model on pooled data yielded estimates of σ^2_A that closely resembled the estimates of $\sigma^2_{A_T}$ from individual data. The direct–indirect animal model on individual data also yielded estimates of σ^2_{Cage} and σ^2_E. As derived by Bergsma et al. [21], $\hat{\sigma}^2_{Cage}$ is an estimate of $2\sigma_{E_{DI}} + (n-2)\sigma^2_{E_I}$. As derived by Bijma [22], $\hat{\sigma}^2_E$ is an estimate of $\sigma^2_{E_D} - 2\sigma_{E_{DI}} + \sigma^2_{E_I}$. As shown in Equation (14), $\hat{\sigma}^2_{E^*}$ is an estimate of $n\left[\sigma^2_{E_D} + 2(n-1)\,\sigma_{E_{DI}} + (n-1)^2\,\sigma^2_{E_I}\right]$. Consequently, the $\hat{\sigma}^2_{Cage}$ and $\hat{\sigma}^2_E$ from the direct–indirect animal model on individual data should sum to the $\hat{\sigma}^2_{E^*}$ from the traditional animal model on pooled data. More precisely:

$$\hat{\sigma}^2_{E^*} = n^2\hat{\sigma}^2_{Cage} + n\hat{\sigma}^2_E. \tag{15}$$

The expected $\hat{\sigma}^2_{E^*}$, calculated based on the $\hat{\sigma}^2_{Cage}$ and $\hat{\sigma}^2_E$ from the direct–indirect animal model on individual data, and the $\hat{\sigma}^2_{E^*}$ from the traditional animal model on pooled data closely resembled each other.

Table 6 does not show heritability estimates. Where the classical heritability (h^2) is used to express $\sigma^2_{A_D}$ relative to the phenotypic variance (σ^2_P), T^2 is used to express $\sigma^2_{A_T}$ relative to σ^2_P [21]. Comparing values of T^2 obtained from individual and pooled data would be misleading because they are not expected to be similar. Unlike for a trait that is not affected by social interactions, $\sigma^2_{P^*}$ cannot simply be divided by the number of group members to obtain σ^2_P. When group members are unrelated,

$$\sigma^2_P = \sigma^2_{A_D} + (n-1)\sigma^2_{A_I} + \sigma^2_{E_D} + (n-1)\sigma^2_{E_I} \tag{16}$$

and

$$\begin{aligned}\sigma^2_{P^*} &= n\sigma^2_{A_T} + \sigma^2_{E^*} \\ &= n[\sigma^2_{A_D} + 2(n-1)\sigma_{A_{DI}} + (n-1)^2\sigma^2_{A_I} \\ &\quad + \sigma^2_{E_D} + 2(n-1)\sigma_{E_{DI}} + (n-1)^2\sigma^2_{E_I}].\end{aligned} \tag{17}$$

The non-proportional increase of σ^2_P does not enable a meaningful comparison between values of T^2 obtained from individual and pooled data.

In conclusion, when group members are unrelated, a traditional animal model on individual data yields

Table 6 Estimated variance components (with s.e.) from individual and pooled data on survival in laying hens

	W1	WB
Direct–indirect animal model on individual data		
$\sigma^2_{A_D}$	705 (± 171)	1404 (± 301)
$\sigma_{A_{DI}}$	59 (± 61)	−162 (± 105)
$\sigma^2_{A_I}$	104 (± 41)	232 (± 72)
σ^2_{Cage}	799 (± 166)	1191 (± 238)
σ^2_E	7980 (± 210)	12 675 (± 365)
$\sigma^2_{A_T}$ [§]	1996 (± 640)	2521 (± 842)
Expected $\sigma^2_{E^*}$ [§§]	44 700 (± 2526)	69 752 (± 3513)
Traditional (direct) animal model on individual data		
$\sigma^2_{A_D}$	677 (± 165)	1522 (± 317)
σ^2_{Cage}	1096 (± 127)	1443 (± 186)
σ^2_E	8002 (± 205)	13 008 (± 338)
Traditional animal model on pooled data		
σ^2_A	1979 (± 643)	2521 (± 845)
$\sigma^2_{E^*}$	44 750 (± 2538)	69 750 (± 3519)

[§] In groups of four, $\sigma^2_{A_T}$ equals $\sigma^2_{A_D} + 6\,\sigma_{A_{DI}} + 9\,\sigma^2_{A_I}$.

[§§] In groups of four, $\sigma^2_{E^*}$ equals $16\,\sigma^2_{Cage} + 4\,\sigma^2_E$.

estimates of $\sigma^2_{A_D}$, while a traditional animal model on pooled data yields estimates of $\sigma^2_{A_T}$. Moreover, the estimated cage and error variances from a direct–indirect animal model on individual data sum to the pooled error variance from a traditional animal model on pooled data. This result could explain the 'inconsistencies' found by Biscarini et al. [17], who assumed that a traditional animal model on individual and pooled data should yield the same genetic variance. Moreover, Biscarini et al. [17] expected to find a pooled error variance that is four times larger than the individual error variance. For body weight at the age of 19 and 27 weeks, these expectations were met. For body weight at the age of 43 and 51 weeks, however, the genetic variance estimated from pooled data was smaller than expected, while the pooled error variance was larger than expected. Biscarini et al. [17] mentions the emergence of competition effects as a possible cause. We indeed expect to find indirect genetic effects when the individual data on body weight at the age of 43 and 51 weeks were reanalysed with a direct–indirect animal model. Using Equations (13) and (15), the estimated variance components from individual data would resemble the estimated variance components from pooled data.

The regression coefficients of $\hat{A}_D$'s obtained from individual data on the $\hat{A}$'s obtained from pooled data strongly deviated from one (0.363 ± 0.006 for W1; 0.392 ± 0.010 for WB). The regression coefficients of $\hat{A}_T$'s obtained from individual data on the $\hat{A}$'s obtained from pooled data were close to, and not significantly different from, one (1.004 ± 0.003 for W1; 1.001 ± 0.001 for WB). This indicates that the $\hat{A}$'s obtained from pooled data are unbiased estimates of the $\hat{A}_T$'s obtained from individual data.

Table 7 shows Spearman correlation coefficients between $\hat{A}_D$'s and $\hat{A}_T$'s obtained from individual data and the $\hat{A}$'s obtained from pooled data. The Spearman correlation coefficients between the $\hat{A}_T$'s obtained from individual data and the $\hat{A}$'s obtained from pooled data were close to, but significantly different from, one. This indicates only a minor loss in the accuracy of $\hat{A}_T$'s when using pooled instead of individual data, which will be reflected in a minor loss in response to selection when using pooled instead of individual data.

Table 7 Spearman correlation coefficients between $\hat{A}_D$'s and $\hat{A}_T$'s obtained from individual data and $\hat{A}$'s from pooled data on survival in laying hens

	$\hat{A}_D$	$\hat{A}_T$	$\hat{A}$
$\hat{A}_D$		0.513 (± 0.001)	0.412 (± 0.001)
$\hat{A}_T$	0.725 (± 0.001)		0.992 (± 0.001)
$\hat{A}$	0.543 (± 0.001)	0.967 (± 0.001)	

Spearman correlation coefficients for data on W1 hens below the diagonal and for data on WB hens above the diagonal.

To gain more insight, we calculated the loss in response to selection that occurs when applying a traditional model to individual or pooled data instead of a direct–indirect model to individual data. When applying a traditional model to individual data, the loss in total response to selection was 46.9% for W1 (Figure 1A) and 54.9% for WB (Figure 1C). When applying a traditional model to pooled data, the loss in total response to selection was 3.3% for W1 (Figure 1B) and 0.3% for WB (Figure 1D). In conclusion, the loss in total response to selection will be large when using a traditional animal model on individual data, but will be small when using a traditional animal model on pooled data. However, this outcome may be specific to this dataset. Survival in purebred laying hens was recorded in cages with four unrelated birds. Both direct and indirect genetic effects strongly influenced the trait. Group size, group composition, and the relative impact of direct and indirect genetic effects might influence the loss in total response to selection. For example, for body weight at 19 and 27 weeks of age, indirect genetic effects are expected to be small. In that case, an animal's A_T is mainly expressed in the phenotype of the animal itself. Consequently, we expect that more accurate estimated breeding values can be obtained when using individual instead of pooled data. Biscarini et al. [17] found a correlation of ~ 0.75 between the estimated breeding values based on individual and pooled data, resulting in a large loss in response to selection when using pooled instead of individual data. Thus, using pooled data does not always seem to be a proper alternative and requires further research.

Conclusions

Using pooled data, the total genetic variance and breeding values can be estimated, but the underlying direct and indirect genetic (co)variances and breeding values cannot. The most accurate estimates are obtained when group members belong to the same family. While quantifying the direct and indirect genetic effects is interesting from a biological perspective, obtaining the total genetic effect is most important from an animal breeding perspective. When it is too difficult or expensive to obtain individual data, pooled data can be used to improve traits.

Appendix A

This section demonstrates why direct and indirect (co)variances can be estimated from individual data, but cannot be estimated from pooled data.

Consider a situation where four base parents produce six offspring. Animals are kept in groups of two and

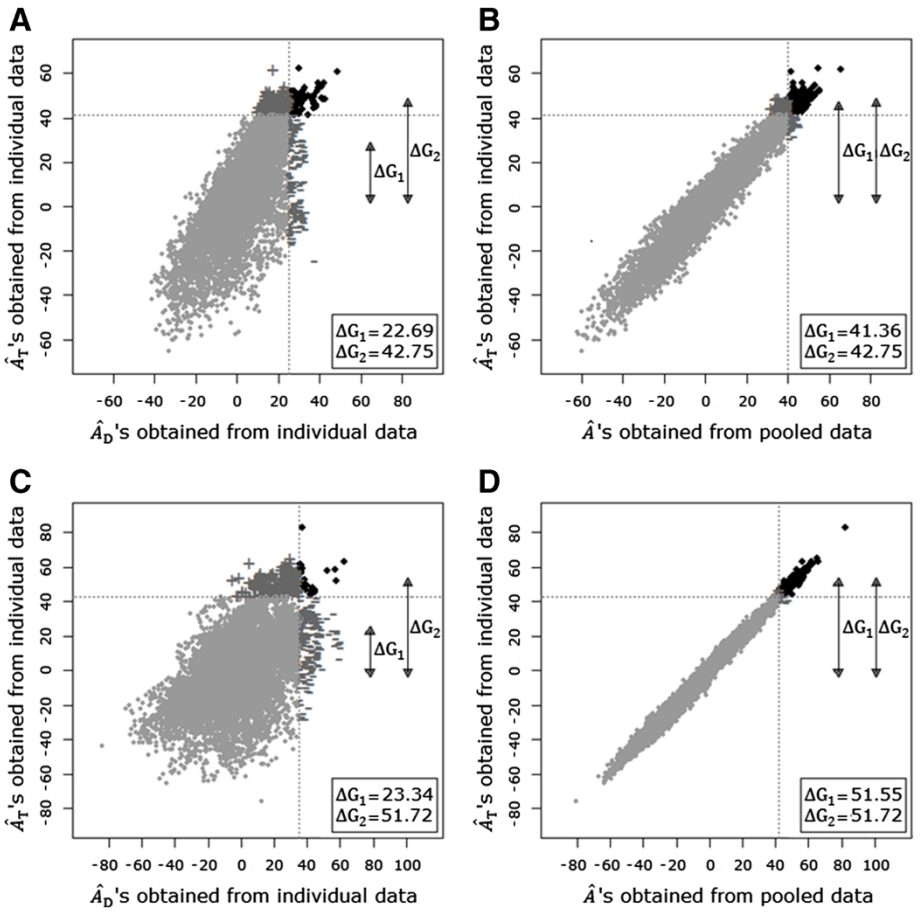

Figure 1 $\hat{A}_T$'s obtained from individual data plotted against $\hat{A}_D$'s obtained from individual data and $\hat{A}$'s obtained from pooled data on survival in laying hens. A and **B** for data on W1 hens. **C** and **D** for data on WB hens. ΔG_1 represents the total response to selection when selecting animals based on their $\hat{A}_D$ obtained from individual data or $\hat{A}$ obtained from pooled data. ΔG_2 represents the total response to selection when selecting animals based on their $\hat{A}_T$ obtained from individual data.

individual phenotypes are recorded on all six offspring (Table 8).

When analysing individual data with a direct–indirect animal model, the **Z**-matrices would be:

$$\mathbf{Z_D} = \begin{bmatrix} 0&0&0&0&1&0&0&0&0&0\\ 0&0&0&0&0&1&0&0&0&0\\ 0&0&0&0&0&0&1&0&0&0\\ 0&0&0&0&0&0&0&1&0&0\\ 0&0&0&0&0&0&0&0&1&0\\ 0&0&0&0&0&0&0&0&0&1 \end{bmatrix},$$

$$\mathbf{Z_I} = \begin{bmatrix} 0&0&0&0&0&1&0&0&0&0\\ 0&0&0&0&1&0&0&0&0&0\\ 0&0&0&0&0&0&0&1&0&0\\ 0&0&0&0&0&0&1&0&0&0\\ 0&0&0&0&0&0&0&0&0&1\\ 0&0&0&0&0&0&0&0&1&0 \end{bmatrix}.$$

Table 8 Example pedigree structure and group composition

Animal	Sire	Dam	Phenotype	Group
1	-	-	-	-
2	-	-	-	-
3	-	-	-	-
4	-	-	-	-
5	1	3	✓	1
6	2	4	✓	1
7	1	4	✓	2
8	2	3	✓	2
9	2	3	✓	3
10	2	4	✓	3

$\mathbf{Z_D}$ and $\mathbf{Z_I}$ are not identical, indicating that the direct and indirect genetic effects are estimated based on different information sources, enabling the model to distinguish between these two effects.

When analysing pooled data with a direct–indirect animal model, the **Z**-matrices would be:

$$\mathbf{Z}^*_{\mathbf{D}} = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix},$$

$$\mathbf{Z}^*_{\mathbf{I}} = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}.$$

$\mathbf{Z}^*_{\mathbf{D}}$ and $\mathbf{Z}^*_{\mathbf{I}}$ are identical, indicating that the direct and indirect genetic effects are estimated based on the same information source, causing complete confounding between direct and indirect genetic effects. The model will not be able to distinguish between these two effects.

Appendix B

Components of variance are determined by analysis of variance, where the full variance $\left(\sigma^2_z\right)$ is partitioned into a between- $\left(\sigma^2_b\right)$ and within-family component (σ^2_w). In this section, the derivation of σ^2_z, σ^2_b and σ^2_w are presented for three group compositions.

(i) When the group is composed of only one family, the A_T of a family is expressed n times in the same pooled record. Therefore, the record of interest is P^*/n.

$$\sigma^2_z = \frac{\sigma^2_{P^*}}{n^2} = \frac{n\left(\sigma^2_{P_D} + 2(n-1)\sigma_{P_{DI}} + (n-1)^2\,\sigma^2_{P_I}\right)}{n^2} + \frac{n(n-1)r\left(\sigma^2_{A_D} + 2(n-1)\,\sigma_{A_{DI}} + (n-1)^2\,\sigma^2_{A_I}\right)}{n^2}$$
$$= \frac{1}{n}\left(\sigma^2_{P_D} + 2(n-1)\sigma_{P_{DI}} + (n-1)^2\,\sigma^2_{P_I} + (n-1)r\,\sigma^2_{A_T}\right)$$
$$\sigma^2_b = r\,\sigma^2_{A_T}$$
$$\sigma^2_w = \frac{1}{n}\left(\sigma^2_{P_D} + 2(n-1)\sigma_{P_{DI}} + (n-1)^2\,\sigma^2_{P_I} + (n-1)r\,\sigma^2_{A_T}\right) - r\,\sigma^2_{A_T}$$

(ii) When the group is composed of two families, the A_T of a family is expressed $n/2$ times in the same pooled record. Therefore, the record of interest is $2P^*/n$.

$$\sigma^2_z = \frac{4\sigma^2_{P^*}}{n^2} = \frac{4n\left(\sigma^2_{P_D} + 2(n-1)\,\sigma_{P_{DI}} + (n-1)^2\,\sigma^2_{P_I}\right)}{n^2} + \frac{4n\left(\frac{n}{2}-1\right)r\left(\sigma^2_{A_D} + 2(n-1)\,\sigma_{A_{DI}} + (n-1)^2\,\sigma^2_{A_I}\right)}{n^2}$$
$$= \frac{4}{n}\left(\sigma^2_{P_D} + 2(n-1)\sigma_{P_{DI}} + (n-1)^2\,\sigma^2_{P_I} + \left(\frac{n}{2}-1\right)r\,\sigma^2_{A_T}\right)$$
$$\sigma^2_b = r\,\sigma^2_{A_T}$$
$$\sigma^2_w = \frac{4}{n}\left(\sigma^2_{P_D} + 2(n-1)\sigma_{P_{DI}} + (n-1)^2\,\sigma^2_{P_I} + \left(\frac{n}{2}-1\right)r\,\sigma^2_{A_T}\right) - r\,\sigma^2_{A_T}$$

(iii) When the group composition is random, the A_T of a family is only expressed once per pooled record. Therefore, the record of interest is P^*.

$$\sigma^2_z = \sigma^2_{P^*} = n\left(\sigma^2_{P_D} + 2(n-1)\,\sigma_{P_{DI}} + (n-1)^2\,\sigma^2_{P_I}\right)$$
$$\sigma^2_b = r\,\sigma^2_{A_T}$$
$$\sigma^2_w = n\left(\sigma^2_{P_D} + 2(n-1)\,\sigma_{P_{DI}} + (n-1)^2\,\sigma^2_{P_I}\right) - r\,\sigma^2_{A_T}$$

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

KP, EDE and PB participated in the design of the study. KP conducted the study. KP, EDE and PB wrote the paper. PB was the principal supervisor of the study. All authors read and approved the manuscript.

Acknowledgements

This research was financially supported by the Netherlands Organization for Scientific Research (NWO) and coordinated by the Dutch Technology Foundation (STW). The Institut de Sélection Animale B.V., a Hendrix Genetics Company, provided the data and was closely involved in this research. The authors would like to thank Ewa Sell-Kubiak, Naomi Duijvesteijn and Sophie Eaglen for their valuable input.

Received: 29 October 2012 Accepted: 23 May 2013
Published: 26 July 2013

References

1. Craig DM: **Group selection versus individual selection: an experimental analysis.** *Evolution* 1982, **36:**271–282.
2. Muir WM: **Incorporation of competitive effects in forest tree or animal breeding programs.** *Genetics* 2005, **170:**1247–1259.
3. Van Vleck LD, Cundiff LV, Koch RM: **Effect of competition on gain in feedlot bulls from Hereford selection lines.** *J Anim Sci* 2007, **85:**1625–1633.
4. Chen CY, Kachman SD, Johnson RK, Newman S, Van Vleck LD: **Estimation of genetic parameters for average daily gain using models with competition effects.** *J Anim Sci* 2008, **86:**2525–2530.
5. Ellen ED, Visscher J, van Arendonk JAM, Bijma P: **Survival of laying hens: genetic parameters for direct and associative effects in three purebred layer lines.** *Poult Sci* 2008, **87:**233–239.

6. Chen CY, Johnson RK, Newman S, Kachman SD, Van Vleck LD: **Effects of social interactions on empirical responses to selection for average daily gain of boars.** *J Anim Sci* 2009, **87:**844–849.
7. Duijvesteijn N, Knol EF, Bijma P: **Direct and associative effects for androstenone and genetic correlations with backfat and growth in entire male pigs.** *J Anim Sci* 2012, **90:**2465–2475.
8. Peeters K, Eppink TT, Ellen ED, Visscher J, Bijma P: **Indirect genetic effects for survival in domestic chicken (*Gallus gallus*) are magnified in crossbred genotypes and show a parent-of-origin effect.** *Genetics* 2012, **192:**705–713.
9. Muir WM, Bijma P, Schinckel A: **Multilevel selection with kin and non-kin groups, experimental results with Japanese quail (*Coturnix japonica*).** *Evolution* 2013. In press.
10. Willham RL: **The covariance between relatives for characters composed of components contributed by related individuals.** *Biometrics* 1963, **19:**18–27.
11. Griffing B: **Selection in reference to biological groups I. Individual and group selection applied to populations of unordered groups.** *Aust J Biol Sci* 1967, **20:**127–139.
12. Moore AJ, Brodie ED, Wolf JB: **Interacting phenotypes and the evolutionary process: I. Direct and indirect genetic effects of social interactions.** *Evolution* 1997, **51:**1352–1362.
13. Wolf JB, Brodie ED, Cheverud JM, Moore AJ, Wade MJ: **Evolutionary consequences of indirect genetic effects.** *Trends Ecol Evol* 1998, **13:**64–69.
14. Bijma P, Muir WM, van Arendonk JAM: **Multilevel selection 1: quantitative genetics of inheritance and response to selection.** *Genetics* 2007, **175:**277–288.
15. Olson KM, Garrick DJ, Enns RM: **Predicting breeding values and accuracies from group in comparison to individual observations.** *J Anim Sci* 2006, **84:**88–92.
16. Biscarini F, Bovenhuis H, van Arendonk JAM: **Estimation of variance components and prediction of breeding values using pooled data.** *J Anim Sci* 2008, **86:**2845–2852.
17. Biscarini F, Bovenhuis H, Ellen ED, Addo S, van Arendonk JAM: **Estimation of heritability and breeding values for early egg production in laying hens from pooled data.** *Poult Sci* 2010, **89:**1842–1849.
18. Lynch M, Walsh JB: *Genetics and analysis of quantitative traits.* Sunderland: Sinauer Associates Inc; 1998.
19. Venables WN, Smith DM, the R development core team: *An Introduction to R.* Vienna: R Foundation for Statistical Computing; 2011.
20. Gilmour AR, Gogel BJ, Cullis BR, Welham SJ, Thompson R: *ASReml user guide, Release 3.0.* Hemel Hempstead: VSN International Ltd; 2009.
21. Bergsma R, Kanis E, Knol EF, Bijma P: **The contribution of social effects to heritable variation in finishing traits of domestic pigs (*Sus scrofa*).** *Genetics* 2008, **178:**1559–1570.
22. Bijma P: **Socially affected traits, Inheritance and genetic improvement**. In *Encyclopedia of sustainability science and technology.* Edited by Meyers RA. Springer science and business media LLC. doi:10.1007/978-1-4419-0851-3. In press.

doi:10.1186/1297-9686-45-27
Cite this article as: Peeters *et al.*: Using pooled data to estimate variance components and breeding values for traits affected by social interactions. *Genetics Selection Evolution* 2013 **45**:27.