Multiple-trait GWAS based on principal component analysis for residual covariance matrix

STAT8750.03 Bio-informatics

Hengrui Luo & Xiaofei Zhou

Department of Statistics
The Ohio State University

January 23, 2017

What is an multi-trait in genome-wide study? I

- **Genome-wide association study(GWAS)**
  - An examination of a genome-wide set of genetic variants (SNPs) to see if they are associated with a trait (for eaxmple, a certain disease).
- **Multi-trait**
  - Multi-trait means multiple correlated traits in observation/measurement of individuals.
  - A typical problem that can be understood as a multi-trait problem is that: What genes can jointly determine a human disease, say asthma.
- **Super traits**
  - Super traits are linear combination of biological measurable traits
  - The estimate of coefficients of 'super traits' can be transformed back into the original traits estimates, yet it does not produce the same result as that from joint multivariate (multi-trait) analysis.

Statistical challenges and existing methods I

- **Statistical challenges**
  - PCA technique is hard to apply on a data set with a lot of (correlated) covariates(traits).
    - If we consider the correlations between traits, then the computational cost is high since PCA requires us to decompose a matrix of the same dimension as the number of covariates(traits).
    - If we do not consider correlation between traits, then the significance of the test on genetic effects are affected(worsened).
  - PCA technique performed on the response variable lacks biological interpretability.

- There are existing method attempting to solve the problem of analysis of multi-trait problem [1].

  - (1)Adaptive multi-task LASSO [3]
    This is a sparse partial leastsquares regression framework proposed to select trait markers associated with each cluster of genes, which has the ability of incorporating many features biologically.

  - (2)Graph-guided fused LASSO [4]
    This is a network model which represents the correlation among traits using network structure to conduct association analysis, but previous knowledge on genomic locations is not incorporated.

Step 1: Define the quantitative genetic model I

- The following model is adopted by [1] to carry out the inference.

$$y_{il} = \sum_{j=1}^{s} x_{ij}\beta_{lj} + \sum_{j=1}^{m} z_{ij}\alpha_{lj} + e_{il}, i = 1, \cdots, n; l = 1, \cdots, t$$

$$\boldsymbol{y}_i = \begin{pmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{it} \end{pmatrix} = (x_{i1}, x_{i2}, \cdots x_{is}) \begin{pmatrix} \beta_{l1} \\ \beta_{l2} \\ \vdots \\ \beta_{ls} \end{pmatrix} + (z_{i1}, z_{i2}, \cdots z_{im}) \begin{pmatrix} \alpha_{l1} \\ \alpha_{l2} \\ \vdots \\ \alpha_{lm} \end{pmatrix} + \begin{pmatrix} e_{i1} \\ e_{i2} \\ \vdots \\ e_{it} \end{pmatrix}, i = 1, \cdots, n$$

- $y_{i,j}$ is the j-th trait measurement of the i-th individual, where $i = 1, \cdots n; j = 1, \cdots t$. In general gentic models, traits are described by a random variable (discrete or normal) yet the author of [1] focus on the case where these traits are normal random variables as we see.
- $x_{i,j}$ is the incidence value for the i-th subject in the j-th systemic environmental effect, where $i = 1, \cdots s; j = 1, \cdots n$ and the effects are generally measured in advance.

Step 1: Define the quantitative genetic model II

- $\beta_{l,j}$ is the j-th systemic environmental effect for the l-th trait, where $l = 1, \cdots s; j = 1, \cdots t$.
  - Therefore the $\boldsymbol{X}\beta$ part in the model represents the systemic effect the environment have on a certain trait.
- $z_{i,j}$ is the indicator variable of the j-th marker for the i-th subject, simply speaking it represents the genome-type of the i-th subject, where $\begin{cases} -1 & aa \\ 1 & AA \\ 0 & Aa \end{cases}$ corresponding to genome-types and $i = 1, \cdots n; j = 1, \cdots m$.
- $\alpha_{l,j}$ is genetic effect of the j-th marker on the l-th trait, where $l = 1, \cdots m; j = 1, \cdots t$.
  - Therefore the $\boldsymbol{Z}\alpha$ part in the model represents the genetic effect have on a certain trait.
- $e_{l,j}$ is the noise/residual effect where $e_{l,j} \overset{i.i.d.}{\sim} N(0, \sigma_\varepsilon^2)$ and $l = 1, \cdots n; j = 1 \cdots, t$

Step 1: Define the quantitative genetic model III

- The whole model in measurement of traits can be split additively into three parts,
  the trait measurement = the environmental effect + the genetic effect + the residual effect

Step 2: Estimation and inference of model I

- Given the model we define above, its likelihood function is in form of product of normal densities.
  "As phenotypes are correlated with each other but independent among subjects, the likelihood function L is then the product of individual multivariate normal distribution density.[1]"

$$L \propto |\boldsymbol{\Sigma_e}|^{-\frac{1}{2}n} \cdot exp\left[-\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{y}_i-\mu_i)'\boldsymbol{\Sigma_e}^{-1}(\boldsymbol{y}_i-\mu_i)\right]$$

$$\propto |\boldsymbol{\Sigma_e}|^{-\frac{1}{2}n} \cdot exp\left[-\frac{1}{2}\sum_{i=1}^{n}(\tilde{\boldsymbol{y}}_i-\tilde{\mu}_i)'\boldsymbol{\Lambda}^{-1}(\tilde{\boldsymbol{y}}_i-\tilde{\mu}_i)\right]$$

$$\propto |\boldsymbol{\Sigma_e}|^{-\frac{1}{2}n} \cdot exp\left[-\sum_{i=1}^{n}\sum_{l=1}^{t}\frac{1}{2\delta_l}(\tilde{y}_{il}-\tilde{\mu}_{il})^2\right]$$

where

Step 2: Estimation and inference of model II

- $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{e}} = V'\boldsymbol{\Lambda}V$ is the maximum likelihood estimate of the covariance matrix of residual effect $\boldsymbol{e}$.
- $\mu_i = (\mu_{i1}, \cdots \mu_{it})$ is the "known" mean (the environmental effect + the genetic effect) function.
- $\tilde{\boldsymbol{y}}_i = V\boldsymbol{y}_i$ is the transformed response which are no longer correlated; $\tilde{\mu}_i = V\mu_i$
- δ_l is the l-th singular value of the estimated residual matrix $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{e}}$.

- Such a product form will allow us to estimate univariate parameter but still preserving the correlation after transformed back, the procedure goes in following manner
  1. Estimate the expectation $\mu_{\boldsymbol{l}}$ for each trait by solving the objective function, this procedure is like matching procedure used in many statistical methods like image-wrapping [5].

$$\mu_{\boldsymbol{l}} = (\boldsymbol{X}\alpha_{\boldsymbol{l}} + \boldsymbol{Z}\beta_{\boldsymbol{l}}) = argmin \left[\sum_{i=1}^{n} y_{il} - \sum_{j=1}^{s} x_{ij}\beta_{lj} - \sum_{j=1}^{m} z_{ij}\alpha_{lj} \right] + \lambda_2 \sum_{j=1}^{m} |\alpha_{lj}|$$

Step 2: Estimation and inference of model III

2. Calculate estimation of residual covariance matrix $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{e}}$ using empirical covariance and decompose it into eigenvalue decomposition form.

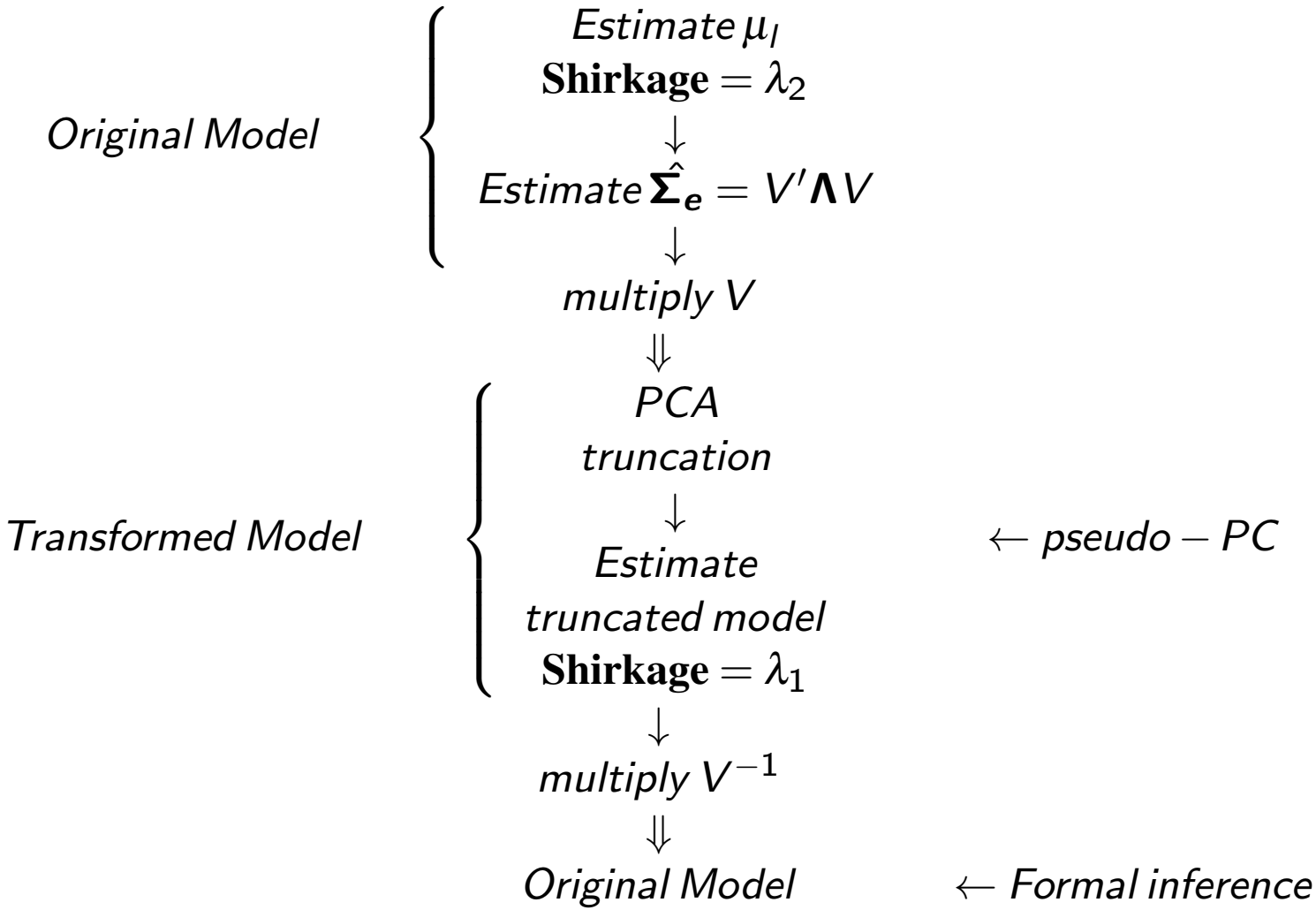
$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{e}} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{y}_i - \mu_i)'(\boldsymbol{y}_i - \mu_i) = V'\boldsymbol{\Lambda}V$$

3. Determine the number of *pseudo principal components*, for example using elbow method according to the cumulative proportion contributed by eigenvalues in matrix $\boldsymbol{\Lambda}$. Generate the pseudo principal components by multiplying a submatrix of V consisting of corresponding eigenvectors.
4. Estimate non-zero genetic effects for each pseudo principal component selected by solving penalized likelihood equation. The solved $\tilde{\alpha}_{lj}$ must be transformed back to $\alpha_{lj} = V^{-1}\tilde{\alpha}_{lj}$ in order to yield a biological interpretation.

$$\tilde{\alpha}_{lj} = argmin\left[\sum_{i=1}^{n}(\tilde{y}_{il} - \tilde{\mu}_{il})^2 + \lambda_1\sum_{j=1}^{m}\left|\tilde{\alpha}_{lj}\right|\right]$$

The whole procedure can be summarized in the flow-chart.

Step 2: Estimation and inference of model IV



RESULTS

Simulated Data

- Generate 1,000 SNPs on each of the 6 chromosomes (so 6,000 SNPs in total)
- The 6,000 SNPs are simulated with equal allele frequencies and are evenly distributed
- The correlation between two adjacent SNPs on the same chromosome is constant (=0.1)
- Use multinormal generation and set $x_{ij} = \begin{cases} 1, & \text{if } > 0.675 \\ -1, & \text{if } < -0.675 \\ 0, & \text{otherwise} \end{cases}$
- 10 SNPs (QTLs) across the 6 chromosomes are set to govern the two normally distributed quantitative traits.

RESULTS

Simulated Data

Table 1 Positions and genetic effects of the QTLs simulated

QTL no.	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7	Q_8	Q_9	Q_{10}
Chr. no.	C_1		C_2		C_3	C_4		C_5	C_6	
Position	310	322	296	686	134	64	516	778	344	648
Effect_1	0.00	0.22	0.21	0.12	−0.17	0.07	0.31	0.26	0.14	−0.15
Heritability_1	0.000	0.021	0.019	0.006	0.012	0.002	0.041	0.029	0.008	0.010
Effect_2	−0.25	0.00	0.19	0.20	0.08	−0.26	−0.28	0.19	0.00	0.12
Heritability_2	0.027	0.000	0.015	0.017	0.003	0.029	0.033	0.015	0.000	0.006

Abbreviation: QTL, quantitative trait loci. Effect_*k* and Heritability_*k* (for $k=1,2$) are genetic effect and heritability, respectively, for the *k*th trait.

RESULTS

Simulated Data

The simulated data are analyzed by

- The paper proposed method (Residual PCA)
- Joint analysis based on phenotypic PCA (Phenotypic PCA)
- The conventional multivariate analysis (Multivariate)

RESULTS

Simulated Data

Table 1 Positions and genetic effects of the QTLs simulated

QTL no.	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7	Q_8	Q_9	Q_{10}
Chr. no.	C_1		C_2		C_3	C_4		C_5	C_6	
Position	310	322	296	686	134	64	516	778	344	648
Effect_1	0.00	0.22	0.21	0.12	−0.17	0.07	0.31	0.26	0.14	−0.15
Heritability_1	0.000	0.021	0.019	0.006	0.012	0.002	0.041	0.029	0.008	0.010
Effect_2	−0.25	0.00	0.19	0.20	0.08	−0.26	−0.28	0.19	0.00	0.12
Heritability_2	0.027	0.000	0.015	0.017	0.003	0.029	0.033	0.015	0.000	0.006

Abbreviation: QTL, quantitative trait loci. Effect_*k* and Heritability_*k* (for $k=1,2$) are genetic effect and heritability, respectively, for the *k*th trait.

Table 2 Statistical powers of QTL detection and false positive rates (FPR) obtained with three mapping methods for the simulated data sets with correlation 0.5

Sample size	*Method*	*Statistical power*										*FPR*
		Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7	Q_8	Q_9	Q_{10}	
1000	Residual PCA	78.8	78.0	83.5	70.2	72.2	83.2	99.0	98.0	29.5	85.2	5.5
	Phenotype PCA	55.0	49.0	58.0	56.5	45.0	66.0	94.0	88.8	16.8	49.5	8.5
	Multivariate	78.2	80.8	81.2	67.0	74.2	86.5	96.5	95.5	30.0	85.0	5.0
2000	Residual PCA	88.2	88.0	94.2	78.8	83.2	92.0	99.8	100.0	40.2	91.0	3.8
	Phenotype PCA	67.8	62.5	67.8	65.2	49.0	86.0	93.8	88.8	32.8	63.8	6.8
	Multivariate	88.0	88.5	90.5	79.2	83.8	92.5	100.0	99.5	42.2	91.2	3.3

Abbreviations: PCA, principal component analysis; QTL, quantitative trait loci.

RESULTS

Simulated Data

Table 1 Positions and genetic effects of the QTLs simulated

QTL no.	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7	Q_8	Q_9	Q_{10}
Chr. no.	C_1		C_2		C_3	C_4		C_5	C_6	
Position	310	322	296	686	134	64	516	778	344	648
Effect_1	0.00	0.22	0.21	0.12	−0.17	0.07	0.31	0.26	0.14	−0.15
Heritability_1	0.000	0.021	0.019	0.006	0.012	0.002	0.041	0.029	0.008	0.010
Effect_2	−0.25	0.00	0.19	0.20	0.08	−0.26	−0.28	0.19	0.00	0.12
Heritability_2	0.027	0.000	0.015	0.017	0.003	0.029	0.033	0.015	0.000	0.006

Abbreviation: QTL, quantitative trait loci. Effect_*k* and Heritability_*k* (for $k=1,2$) are genetic effect and heritability, respectively, for the *k*th trait.

Table 3 Mean estimates and s.ds. (in parentheses) of QTL effects obtained with three mapping methods for the simulated data sets with correlation 0.5

Sample size	*Method*	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7	Q_8	Q_9	Q_{10}
1000	Residual PCA	0.04 (0.01)	0.21 (0.00)	0.21 (0.00)	0.10 (0.01)	−0.21 (0.00)	0.11 (0.01)	0.33 (0.02)	0.25 (0.00)	0.10 (0.00)	−0.14 (0.00)
		−0.23 (0.02)	−0.03 (0.00)	0.21 (0.00)	0.21 (0.01)	0.11 (0.00)	−0.26 (0.01)	−0.31 (0.01)	0.20 (0.00)	−0.02 (0.00)	0.11 (0.01)
	Phenotype PCA	0.12 (0.02)	0.14 (0.00)	0.25 (0.00)	0.15 (NA)	−0.13 (0.01)	0.08 (0.00)	0.28 (0.01)	0.21 (0.01)	0.07 (0.00)	−0.14 (0.00)
		−0.17 (0.01)	−0.10 (0.00)	0.25 (0.00)	0.17 (NA)	0.10 (0.01)	−0.18 (0.00)	−0.27 (0.01)	0.18 (0.01)	−0.07 (0.00)	0.13 (0.00)
	Multivariate	0.03 (0.01)	0.20 (0.01)	0.21 (0.01)	0.09 (0.01)	−0.19 (0.01)	0.09 (0.00)	0.32 (0.01)	0.28 (0.01)	0.14 (0.01)	−0.10 (0.00)
		−0.25 (0.01)	−0.03 (0.01)	0.19 (0.00)	0.21 (0.01)	0.09 (0.00)	−0.29 (0.01)	−0.27 (0.00)	0.17 (0.01)	−0.02 (0.00)	0.09 (0.01)
2000	Residual PCA	0.07 (0.01)	0.22 (0.00)	0.21 (0.00)	0.10 (0.01)	−0.21 (0.00)	0.11 (0.01)	0.31 (0.01)	0.24 (0.00)	0.14 (0.00)	−0.13 (0.00)
		−0.24 (0.01)	−0.03 (0.00)	0.21 (0.00)	0.22 (0.01)	0.12 (0.00)	−0.24 (0.01)	−0.30 (0.01)	0.21 (0.00)	−0.02 (0.00)	0.11 (0.00)
	Phenotype PCA	0.13 (0.02)	0.15 (0.01)	0.22 (0.00)	0.17 (0.00)	−0.13 (0.01)	0.14 (0.01)	0.3 (0.01)	0.23 (0.02)	0.08 (0.01)	−0.18 (0.01)
		−0.18 (0.01)	−0.10 (0.01)	0.22 (0.00)	0.18 (0.00)	0.13 (0.01)	−0.13 (0.01)	−0.29 (0.01)	0.15 (0.02)	−0.08 (0.01)	0.18 (0.01)
	Multivariate	0.02 (0.00)	0.21 (0.01)	0.20 (0.01)	0.13 (0.01)	−0.18 (0.01)	0.10 (0.00)	0.32 (0.01)	0.28 (0.01)	0.12 (0.01)	−0.16 (0.01)
		−0.25 (0.01)	−0.02 (0.01)	0.19 (0.00)	0.20 (0.02)	0.11 (0.01)	−0.26 (0.01)	−0.28 (0.00)	0.19 (0.01)	0.01 (0.00)	0.12 (0.00)

Abbreviations: NA, not available; PCA, principal component analysis; QTL, quantitative trait loci.

RESULTS

Simulated Data

- The three methods give similar patterns:
  1. statistical power of QTL detection and the precision of parameter estimation increase as the QTL heritability increases
  2. statistical power of QTL detection is higher and false positive rate is lower as the QTL heritability increases
  3. large sample size is beneficial to identify QTL
- The three methods are all able to accurately locate the simulated QTLs.
- Residual PCA is basically identical to Multivariate in terms of statistical power and QTL parameter estimation, but both distinctly outperform Phenotypic PCA
- Residual PCA and Phenotypic PCA require less computational resources than Multivariate

RESULTS

Real Data

Some key aspects of the experiment:

- 986 young Simmental bulls are collected
- 631,396 SNPs are utilized
- 20 carcass and meat quality traits are chosen

Statistical analysis:

- Environmental factors are included in the genetic model
- Population stratification is taken into account
- The first two pseudo traits are analyzed, which together explain $> 85\%$ of the residual covariance matrix variation

RESULTS

Real Data

Table 4 The detected SNPs for the first two pseudo principal components (SPC) of 20 carcass traits and meat quality traits in beef cattle

SPC	*QTL no.*	*SNP*	*Chr.*	*Position*	*−Log(p)*	*Effect*	*Heritability*
First	1	BovineHD0700006504	7	23736205	6.11	0.03	0.01
	2	BovineHD1000023693	10	83167500	4.84	0.03	0.01
	3	BovineHD1500018258	15	63694848	3.69	−0.03	0.01
	4	BovineHD0700005046	7	17994045	3.02	0.10	0.06
	5	BovineHD0900003540	9	13538093	4.40	0.03	0.01
	6	BovineHD2200010203	22	35643205	3.78	0.02	0.00
	7	BovineHD2500007552	25	26940925	4.89	0.12	0.09
	8	BovineHD2700000367	27	1169118	3.67	0.06	0.02
Second	9	BovineHD0500004156	5	13861704	6.47	0.12	0.09
	10	BovineHD0600033075	6	116460483	3.62	−0.04	0.01
	11	BovineHD0700008057	7	28564386	4.33	−0.04	0.01
	12	BovineHD0900016838	9	61346136	4.90	−0.07	0.03
	13	BovineHD1300001192	13	4584757	3.68	0.15	0.13
	14	BovineHD1700021389	17	73171831	3.31	0.11	0.07

Abbreviations: QTL, quantitative trait loci; SNP, single-nucleotide polymorphism.

RESULTS

Real Data

- Genetic effects cannot precisely reflect the impact of the QTLs
- The heritabilities of detected QTLs on 20 analyzed traits can be calculated from the estimated genetic effects and the estimated residual variances, which is estimated by $diag(V^T \Lambda V)$ for original traits
- The heritabilities of detected QTLs can also indicate the extend to which the pleiotropy occurs

RESULTS

Real Data

Table 5 Estimated heritabilities of the detected QTLs for 20 carcass traits and meat quality traits in beef cattle

Trait no.	*QTL no.*													
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*
1	0.0027	0.0039	0.0039	0.0354	0.0027	0.0017	0.0437	0.0132	0.1262	0.0157	0.0157	0.0394	0.1925	0.1118
2	0.0100	0.0106	0.0112	0.1110	0.0100	0.0062	0.1397	0.0435	0.0662	0.0085	0.0085	0.0214	0.1039	0.0581
3	0.0110	0.0115	0.0124	0.1215	0.0115	0.0067	0.1558	0.0477	0.0573	0.0074	0.0071	0.0181	0.0902	0.0505
4	0.0091	0.0097	0.0103	0.1002	0.0091	0.0054	0.1271	0.0387	0.0748	0.0097	0.0091	0.0236	0.1164	0.0651
5	0.0014	0.0014	0.0015	0.0150	0.0014	0.0008	0.0192	0.0059	0.1406	0.0178	0.0174	0.0443	0.2196	0.1219
6	0.0008	0.0008	0.0009	0.0085	0.0008	0.0005	0.0108	0.0033	0.1457	0.0185	0.0180	0.0458	0.2276	0.1263
7	0.0004	0.0004	0.0004	0.0044	0.0004	0.0003	0.0054	0.0016	0.1492	0.0186	0.0186	0.0469	0.2332	0.1294
8	0.0003	0.0003	0.0003	0.0033	0.0003	0.0002	0.0042	0.0013	0.1498	0.0190	0.0185	0.0471	0.2338	0.1298
9	0.0001	0.0001	0.0001	0.0012	0.0001	0.0001	0.0016	0.0005	0.1513	0.0192	0.0186	0.0476	0.2367	0.1312
10	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.1523	0.0193	0.0188	0.0479	0.2379	0.1320
11	0.0076	0.0079	0.0085	0.0830	0.0076	0.0046	0.1055	0.0322	0.0881	0.0112	0.0108	0.0277	0.1375	0.0763
12	0.0171	0.0178	0.0191	0.1872	0.0172	0.0103	0.2376	0.0727	0.0075	0.0009	0.0009	0.0023	0.0117	0.0065
13	0.0159	0.0166	0.0178	0.1743	0.0161	0.0096	0.2215	0.0677	0.0173	0.0022	0.0021	0.0055	0.0271	0.0150
14	0.0029	0.0030	0.0032	0.0316	0.0029	0.0017	0.0401	0.0123	0.1279	0.0162	0.0158	0.0402	0.1997	0.1108
15	0.0103	0.0107	0.0115	0.1125	0.0103	0.0062	0.1429	0.0437	0.0652	0.0083	0.0080	0.0205	0.1019	0.0566
16	0.0164	0.0171	0.0183	0.1793	0.0165	0.0099	0.2278	0.0696	0.0134	0.0017	0.0017	0.0042	0.0210	0.0116
17	0.0061	0.0064	0.0068	0.0669	0.0062	0.0037	0.0849	0.0260	0.1005	0.0128	0.0124	0.0316	0.1571	0.0872
18	0.0128	0.0133	0.0143	0.1400	0.0129	0.0077	0.1779	0.0544	0.0439	0.0056	0.0054	0.0138	0.0686	0.0381
19	0.0004	0.0004	0.0004	0.0043	0.0004	0.0002	0.0055	0.0017	0.1489	0.0189	0.0184	0.0468	0.2327	0.1291
20	0.0176	0.0176	0.0213	0.1914	0.0176	0.0112	0.2406	0.0775	0.0044	0.0007	0.0007	0.0016	0.0063	0.0028

Abbreviation: QTL, quantitative trait loci.

References I

[1] Gao, H., et al. "Multiple-trait genome-wide association study based on principal component analysis for residual covariance matrix." Heredity 113.6 (2014): 526-532.

[2] Ghalambor, Cameron K., Jeffrey A. Walker, and David N. Reznick. "Multi-trait selection, adaptation, and constraints on the evolution of burst swimming performance." Integrative and Comparative Biology 43.3 (2003): 431-438.

[3] Zhu, Wensheng, and Heping Zhang. "Why do we test multiple traits in genetic association studies?." Journal of the Korean Statistical Society 38.1 (2009): 1-10.

[4] Kim, Seyoung, and Eric P. Xing. "Statistical estimation of correlated genome associations to a quantitative trait network." PLoS Genet 5.8 (2009): e1000587.

References II

[5] Sampson, Paul D., and Peter Guttorp. "Nonparametric estimation of nonstationary spatial covariance structure." Journal of the American Statistical Association 87.417 (1992): 108-119.