Files S1-S10

All files are available for download at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.157032/-/DC1.

File S1: This dataset contains all six types of kinship matrices, each with a dimension 278×278. The type of kinship matrix is indicated by the first column (variable named parm) with 1, 2, 3, 4, 5, 6 indicating a, d, aa, dd, ad and da, respectively. The second column of this dataset gives the row numbers of each type of the kinship matrix. This special format is required by the PROC MIXED program. The data must contain $n+2$ variables with variable names of parm, row, col1, …, coln, where n is the number of lines. Do not mess up the variable names! The type=lin(6) option in the random statement of PROC MIXED means the program is looking for 6 kinship matrices.

File S2: This dataset stores the fixed-effect-adjusted phenotypic values of the four traits, yield, tiller, grain and kgw. The first two columns give the line id and line names. The last column gives the fold id that is used in the five-fold cross validation analysis by the Lasso method implemented via the GlmNet/R program.

File S3: This dataset gives the genotypes of two selected bins, bin1 = 729 and bin2 = 1064. The first column gives the names of the 278 IMF2 crosses. This dataset is used by the PROC HPMIXED program shown in Script S2.

File S4: This dataset list the estimated additive (a) and dominance (d) effects for all the 1619 bins obtained from the weighted mixed model (model incorporating the polygenic covariance structure). The standard errors corresponding to the additive and dominance effects are denoted by a_err and d_err, respectively. Effect specific tests are denoted by f_a and f_d (F test or Wald test). The LRT and p-value are also provided in the dataset. The last column gives the significant status with value 1 for $p < 0.05$ and 0 for $p > 0.05$.

File S5: This dataset list the estimated additive (a) and dominance (d) effects for all the 1619 bins obtained from the unweighted mixed model (model ignoring the polygenic covariance structure). The standard errors corresponding to the additive and dominance effects are denoted by a_err and d_err, respectively. Effect specific tests are denoted by f_a and f_d (F test or Wald test). The LRT and p-value are also provided in the dataset. The last column gives the significant status with value 1 for $p < 0.05$ and 0 for $p > 0.05$.

File S6: This dataset stores all bins selected with LRT > 4.61 (LOD >1) for the main effects (additive and dominance) from the main effect model analysis. The estimated additive and dominance effects along with the standard errors are given in columns headed by a, d, a_err and d_err respectively. The p-values drawn from permutation tests (1000 permuted samples) are also given. The last column indicates the significance status with 1 for $p < 0.05$ and 0 for $p > 0.05$. There are four sheets in the file, one for each trait.

File S7: This excel spread dataset stores all bin pairs selected with LRT > 9.22 (LOD > 2) for the epistatic effects (aa, dd, ad and da) from the epistatic model analysis. The estimated additive × additive, dominance × dominance, additive × dominance and dominance × additive effects along with the standard errors are given in columns headed by aa, dd, ad, da, aa_err, dd_err, ad_err and da_err, respectively. The p-values drawn from permutation tests are also given. The last column indicates the significance status with 1 for $p < 0.05$ and 0 for $p > 0.05$. There are four sheets in the file, one for each trait.

File S8: This dataset contains the estimated non-zero effects from the Lasso analysis for all bins and bin pairs that have survived the preliminary screen. Note that in the preliminary screen, the criteria were LOD > 1 for bins and LOD > 2 for bin pairs. When bin_1 equals bin_2, the effect represents a main effect (a for additive and d for dominance). When bin_1 does not equal bin_2, the effect represents an epistatic effect whose type is indicated by the column headed by type, i.e., aa, dd, ad and da, for the four types of epistatic effects, respectively. The estimated effect and the standard error of the estimate are given in columns headed by estimate and stderr. Wald test and LOD score along with the p-value are also given in the file. The last column shows the significance status from the Lasso analysis with 1 indicating $p < 0.05$ and 0 indicating $p > 0.05$. There are four sheets in the file, one for each trait.

File S9: This dataset contains the significant effects from the Lasso analysis for all bins and bin pairs that have survived the preliminary screen (LOD > 1 for bins and LOD > 2 for bin pairs). This dataset is a subset of Data S8 that excludes all effects with $p > 0.05$. All effects listed in this table are deemed to be significant. When bin_1 equals bin_2, the effect represents a main effect (a for additive and d for dominance). When bin_1 does not equal bin_2, the effect represents an epistatic effect whose type is indicated by the column headed by type, i.e., aa, dd, ad and da, for the four types of epistatic effects, respectively. The

estimated effect and the standard error of the estimate are given in columns headed by estimate and stderr. Wald test and LOD score along with the p-value are also given in the file. There are four sheets in the file, one for each trait.

File S10: This dataset contains the design matrix for all the significant effects listed in Data S9. The first four columns give the bin and bin pair information along with the effect type information. Variable named typeA shows the type of effect. A numerical version of the type is given in column headed by typeN, with 1, 2, 3, 4, 5 and 6 represent a, d, aa, dd, ad and da, respectively. The numerical type allows sorting by type in the natural order. The remaining $n = 278$ columns store the elements in the design matrix for multiple regression analysis using only the significant effects. This matrix needs to be transposed before fitting a multiple regression model. When bin_1 equals bin_2, the effect represents a main effect (a for additive and d for dominance). When bin_1 does not equal bin_2, the effect represents an epistatic effect whose type is indicated by the column headed by type, i.e., aa, dd, ad and da, for the four types of epistatic effects, respectively. There are four sheets in the file, one for each trait.

File S11

Script S1: SAS PROC MIXED for Polygenic Variance Components Analysis

This is the program code for PROC MIXED in SAS. The code also includes PROC IMPORT used to read the input data. The outputs are directly printed out in the window. In addition, estimated parameters and predicted genomic values are also written in SAS datasets. These SAS datasets can be exported later into physical files using PROC EXPORT (not provided). To make sure that PROC MIXED produces legal estimates of variance components, a lowerb= option is given in the parms statement. The lower bound option of 1e-5 means that each variance component is bounded at 1e-5, i.e., $\sigma_X^2 \geq 1e-5$. There are seven estimated variance components (six polygenic variances and one residual variance). All initial values of the variances are set to 1.0. Users can choose different initial values, depending on the properties of the data. The initial value of one (1) is the default initial value for the variance parameters in PROC MIXED. The trait shown in the code is KGW.

```
/*begin code*/

%let dir=C:\Users\SHXU\Programs;
filename kk "&dir\Data S1.csv" lrecl=20000;
filename phe "&dir\Data S2.csv";

proc import datafile=kk out=kk dbms=csv replace;
proc import datafile=phe out=phe dbms=csv replace;
run;

proc mixed data=phe method=reml;
class line;
model kgw=/solution;
random line/type=lin(6) ldata=kk solution;
parms (1) (1) (1) (1) (1) (1) (1)/lowerb=1e-5 1e-5 1e-5 1e-5 1e-5 1e-5 1e-5;
ods output SolutionR=blup SolutionF=fixed CovParms=covar;
run;

data pred;
   merge phe blup;
run;

proc corr data=pred;
   var kgw estimate;
run;

/*end code*/
```

Comments: The program takes two input files stored in a user defined folder (c:\users\shxu\programs in this example), one file for the kinship matrices (named Data S1.csv in this example) and one for the phenotypic values (named Data S2.csv in this example). The Data S2.csv file must contain a variable for the id number of lines (named line in this example) and a variable for the phenotypic values of the trait in question (named kgw in this example). The program will generate three SAS datasets. One SAS dataset is called blup, which gives the predicted polygenic value for each line, the second SAS dataset is named fixed, which gives the estimated fixed effects and the third SAS dataset named covar gives all the seven estimated variance components, including six polygenic variances and one residual variance. The two input files are provided in Supplemental Data S1 for the kinship matrix and Data S2 for fixed-effect adjusted phenotypic values of 278 lines for four quantitative traits.

File S12

Script S2: SAS PROC HPMIXED for Association Studies

This is the program code for PROC HPMIXED in SAS. The code also includes PROC IML for eigenvalue/eigenvector calculation and data manipulation. This program only shows the mixed model association study for two bins, bin1 = 729 and bin2 = 1064. The trait analyzed is KGW because this trait has shown that the two bins have strong interactions in the whole genome analysis. The model contains eight genetic effects (a1, a2, d1, d2, aa, ad, da, and dd). The program requires polygenic variance ratios (lambda values) calculated from the PROC MIXED program. The SAS dataset named lambda contains pre-calculated lambda values for all the four traits and thus the data matrix dimension is 6×4 (six observations and four variables). The program will print all outputs on the screen and also write various output tables into SAS datasets. The most important output is the estimated genetic effects in an output dataset called blup1. In the script, the PROC HPMIXED program is called twice, one for the polygenic model (null model) without fitting the two bins. This call produces a likelihood value ($-2L_0$) under the null model. The second call of this procedure produces a likelihood value ($-2L_1$) under the full model (fitting the two bins). A dataset called lrt is generated by taking the difference between the two likelihood values, lrt = ($-2L_0$) – ($-2L_1$) = $-2(L_0-L_1)$. This likelihood ratio test statistic is testing the significance of the two bins jointly.

```
/*begin code*/

%let dir=C:\Users\SHXU\Programs;
filename kk "&dir\Data S1.csv" lrecl=20000;
filename phe "&dir\Data S2.csv";
filename gen "&dir\Data S3.csv" lrecl=20000;

proc import datafile=kk out=kk dbms=csv replace;
proc import datafile=phe out=phe dbms=csv replace;
proc import datafile=gen out=gen dbms=csv replace;
run;

data lambda;
   input yield tiller grain kgw;
cards;
5.09424E-06   1.190743713   7.00747898    20.04424779
5.09424E-06   2.67523E-05   4.64533E-07   8.84956E-05
3.570809985   1.563937935   3.10387885    2.761061947
2.685277636   2.67523E-05   0.305281739   2.010619469
9.310901681   0.322632424   5.118688159   8.84956E-05
2.826439124   3.024879615   4.64533E-07   1.666371681
;

proc iml;
use lambda;
   read all var{kgw} into lambda;
close;
use phe;
   read all var{kgw} into y;
close;
p=nrow(lambda);
n=nrow(y);
kk=j(n,n,0);
do k=1 to p;
   range=((k-1)*n+1):(k*n);
   use kk;
      read point range into kk0;
   close;
   kk0=kk0[,3:(n+2)];
   kk=kk+kk0*lambda[k];
end;
call eigen(delta,uu,kk);
create delta from delta;
   append from delta;
```

```
close;
create uu from uu;
   append from uu;
close;
x=j(n,1,1);
xu=uu`*x;
yu=uu`*y;
w=1/(delta+1);
yxw=yu||xu||w;
varname={"y" "x" "w"};
create yxw from yxw[colname=varname];
   append from yxw;
close;
quit;

proc hpmixed data=yxw;
   model y = x/noint solution;
   weight w;
   ods output CovParms=parm0 FitStatistics=fit0 ParameterEstimates=fixed0;
   nloptions  maxiter=10000 gconv=1e-8;
run;

proc iml;
use gen;
  read all var{bin729 bin1064} into zz;
close;
k=1;
l=2;
zk=(zz[,k]='A')-(zz[,k]='B');
wk=(zz[,k]='H');
zl=(zz[,l]='A')-(zz[,l]='B');
wl=(zz[,l]='H');
create zz from zz;
append from zz;
close;

use yxw;
  read all into yxw;
close;
use uu;
   read all into uu;
close;
z=zk||zl||wk||wl||(zk#zl)||(zk#wl)||(wk#zl)||(wk#wl);
zu=uu`*z;
yxwz=yxw||zu;
varname={"y" "x" "w" "a1" "a2" "d1" "d2" "aa" "ad" "da" "dd"};
create yxwz from yxwz[colname=varname];
    append from yxwz;
close;
quit;

proc hpmixed data=yxwz;
   effect z=collection(a1 a2 d1 d2 aa ad da dd);
   model y=x/noint solution;
   weight w;
   random z / solution;
   ods output CovParms=parm1 FitStatistics=fit1
              ParameterEstimates=fixed1
              SolutionR=blup1 ConvergenceStatus=conv1;
   nloptions  maxiter=10000 gconv=1e-8;
run;

data lrt;
```

```
   merge fit0(rename=(value=l0)) fit1(rename=(value=l1));
   lrt=l0-l1;
   if _n_=1;
run;

/*end code*/
```

One of the outputs generated from PROC HPMIXED is the estimated genetic effects for the two bins (bin1 = 729 and bin2 = 1064).

Solution for Random Effects

Effect	z	Estimate	Std Err Pred	DF	t Value	Pr > |t|
z	a1	0.5388	0.1688	277	3.19	0.0016
z	a2	0.09325	0.1619	277	0.58	0.5651
z	d1	-0.1718	0.1315	277	-1.31	0.1924
z	d2	-0.07827	0.1302	277	-0.60	0.5482
z	aa	0.2817	0.1164	277	2.42	0.0162
z	ad	0.3970	0.1296	277	3.06	0.0024
z	da	0.08289	0.1327	277	0.62	0.5329
z	dd	-0.1170	0.1595	277	-0.73	0.4636

Comments: The program takes three input files stored in a user defined folder (c:\users\shxu\programs in this example), one file for the kinship matrices (named Data S1.csv in this example), one for the phenotypic values (named Data S2.csv in this example) and the third file for the bin genotypes (named Data S3 in this example). The Data S2.csv file must contain a variable for the id number of lines (named line in this example) and a variable for the phenotypic values of the trait in question (named kgw in this example). The program also requires a SAS dataset named lambda to store the six variance ratios obtained from the polygenic analysis via PROC MIXED described in Script S1. The program will generate several SAS datasets. One SAS dataset is called blup1, which gives the estimated genetic effects in the following order: a1, a2, d1, d2, aa, ad, da and dd. The three input files are provided in Supplemental Datas S1, S2 and S3, respectively.

Table S1 Estimated genetic and residual variance components under different models sorted by model size for the yield component traits in rice. This table gives the estimated genetic variance components under different models sorted by model size for the yield component traits obtained from the IMF2 population of rice. Six models were compared and the sizes of the six models range from one to six. The model labeled a is the additive model with only one additive variance component (model size is one). The model labeled a + d is the main effect model with additive and dominance variance components (model size is two). Other models are defined in the same way. The estimated genetic variance components are arranged in lower triangular of a square table for each trait because a particular type of variance component is only available from a model that includes this type of genetic variance.

Trait	Model	a	d	aa	dd	ad	da	Residual
Yield	a	14.4911						23.3308
	a+d	13.1129	9.0443					19.2088
	a+d+aa	11.2609	7.73516	7.38906				14.0566
	a+d+aa+dd	10.7365	0.00001	7.80710	7.23373			12.2930
	a+d+aa+dd+ad	4.32190	0.00001	7.27218	5.442035	16.7565		4.90684
	a+d+aa+dd+ad+da	0.00001	0.00001	7.00964	5.271305	18.2771	5.54845	1.96303
Tiller	a	1.38792						1.39981
	a+d	1.38792	0.00001					1.39981
	a+d+aa	1.05890	0.00001	0.61650				0.97550
	a+d+aa+dd	1.05895	0.00001	0.61645	0.00001			0.97549
	a+d+aa+dd+ad	1.05896	0.00001	0.61650	0.00001	0.00001		0.97545
	a+d+aa+dd+ad+da	0.44516	0.00001	0.58511	0.00001	0.12044	1.13048	0.37376
Grain	a	254.636						124.165
	a+d	245.599	24.3225					113.847
	a+d+aa	193.179	14.9761	69.4831				74.4618
	a+d+aa+dd	192.990	0.00001	68.5104	17.41830			69.7084
	a+d+aa+dd+ad	150.877	0.00001	66.8473	6.55935	110.184		21.5337
	a+d+aa+dd+ad+da	150.913	0.00001	66.8373	6.579376	110.180	0.00001	21.5147
KGW	a	2.82000						0.54720
	a+d	2.69618	0.26283					0.43694
	a+d+aa	2.38064	0.21624	0.32088				0.25818
	a+d+aa+dd	2.34714	0.00001	0.32697	0.245812			0.18642
	a+d+aa+dd+ad	2.34717	0.00001	0.32699	0.245813	0.00001		0.18641
	a+d+aa+dd+ad+da	2.26561	0.00001	0.31201	0.227324	0.00001	0.18803	0.11309