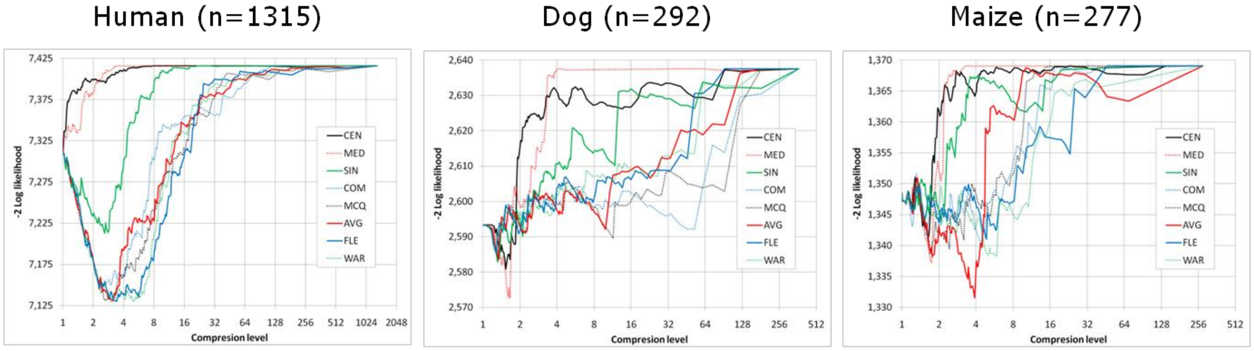
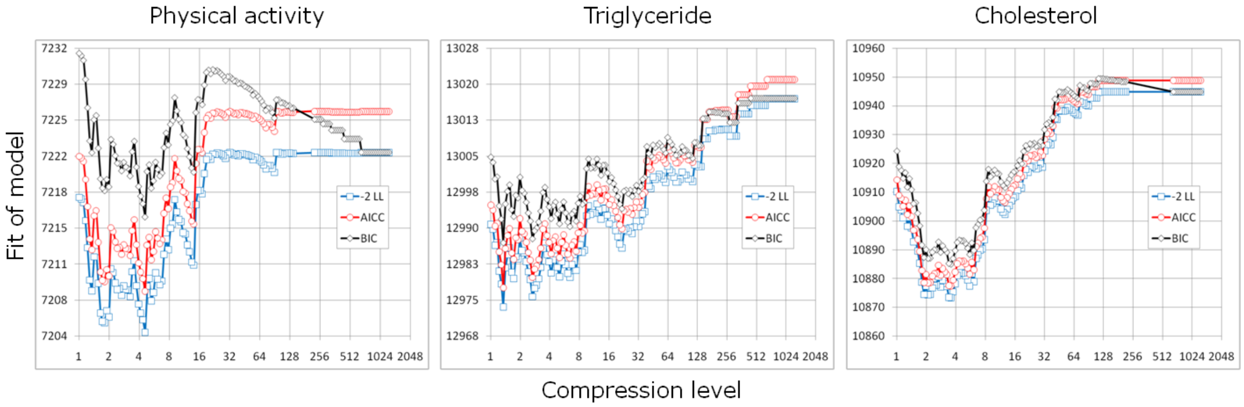
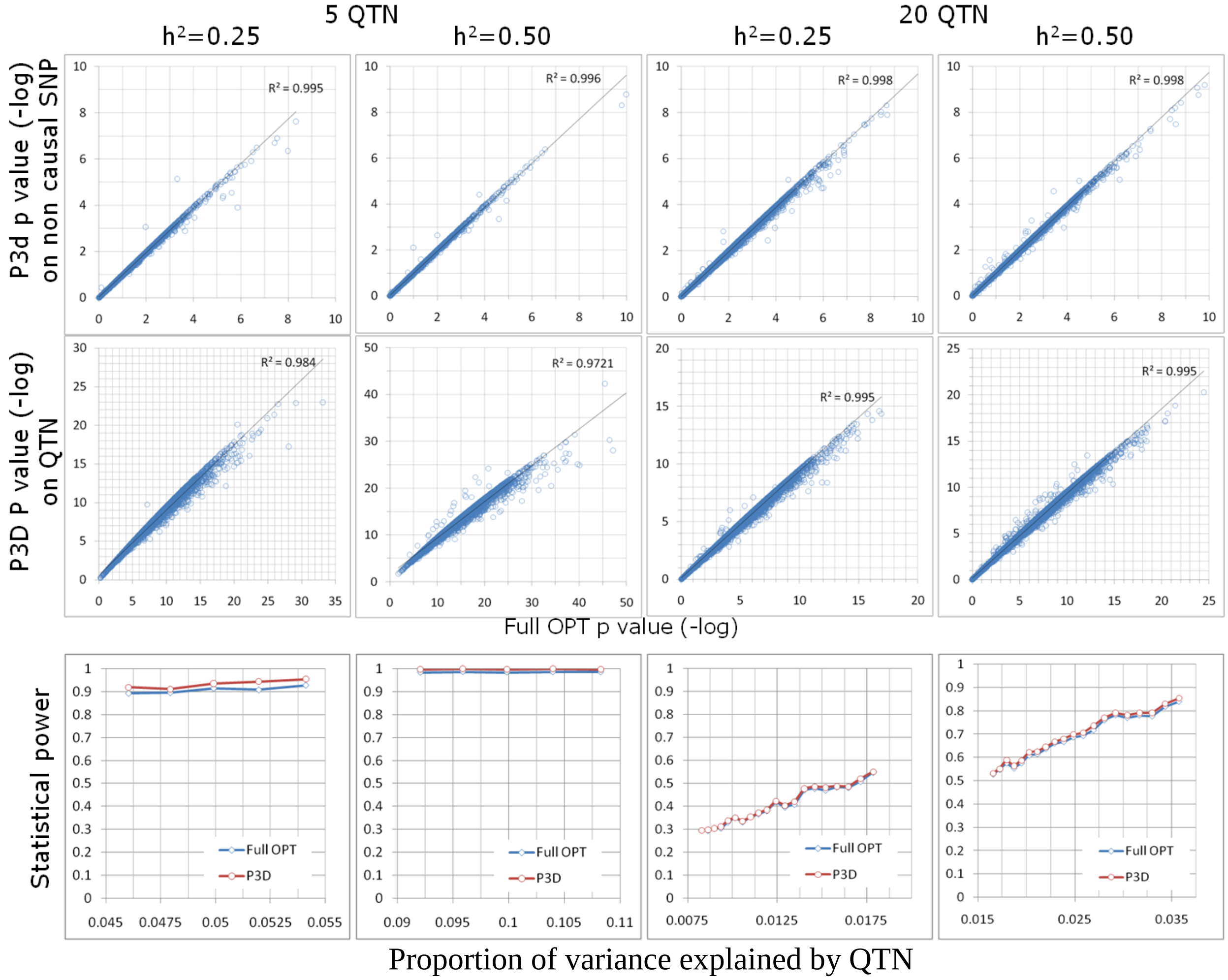


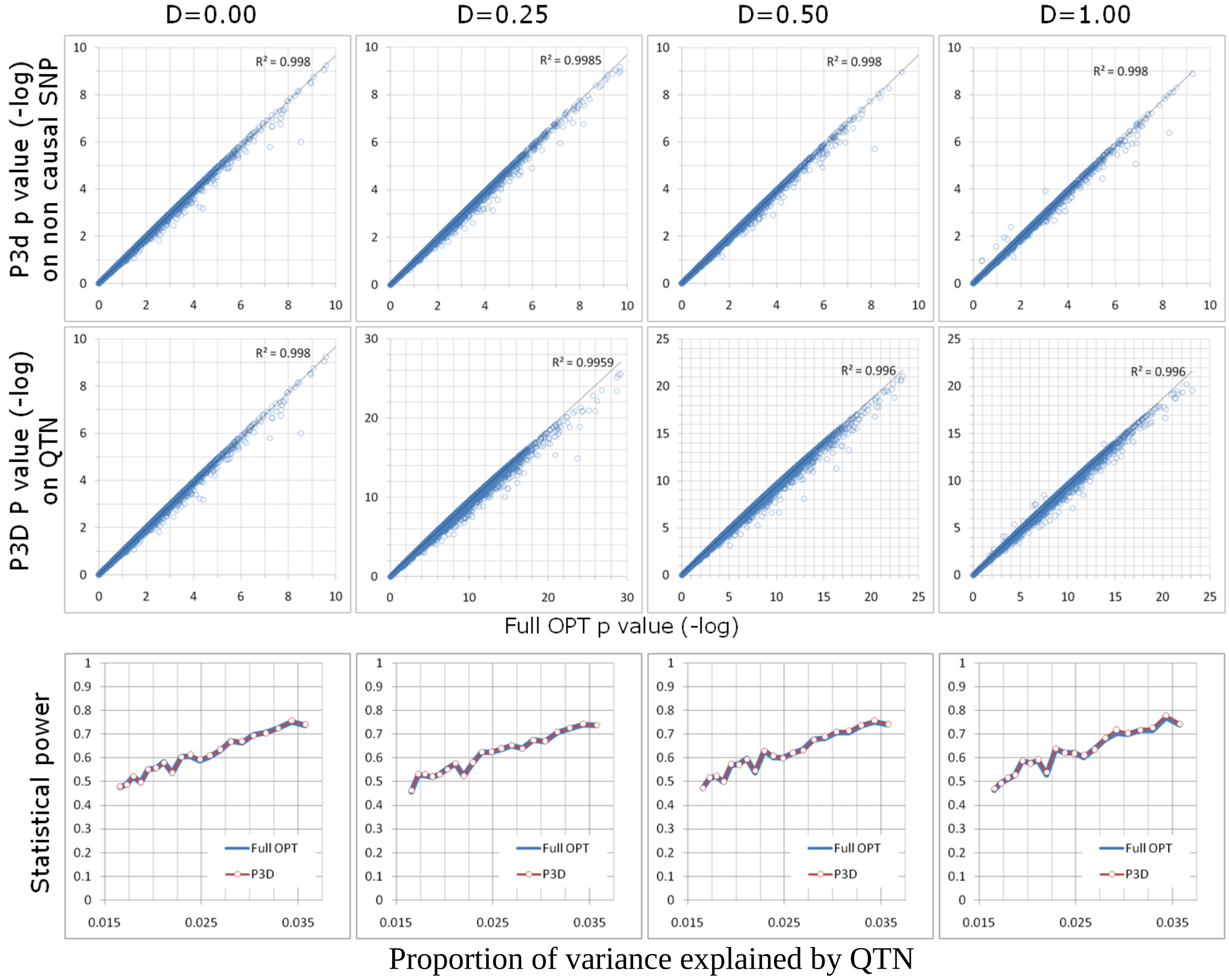
Supplementary Figure 1. The fit of compressed mixed linear model for each combination of eight hierarchical clustering algorithms and compression levels. These clustering algorithms[1] are unweighted pair group method with arithmetic average (AVE), unweighted pair-group centroid (CEN), complete linkage (COM), Lance-Williams flexible-beta method (FLE), McQuitty's similarity analysis (MCQ, also called weighted pair-group method using arithmetic averages), weighted pair-group centroid median (MED), single linkage (SIN) and Ward's method (WAR) implemented by Proc Cluster in SAS[2]. The evaluated phenotypes are height[3], hip dysplasia (Norberg angle)[4] and flowering time (days to pollination)[5] in human, dog and maize, respectively. The model fit is indicated by the negative log likelihood (-LL). Smaller values of –LL indicate better fit.



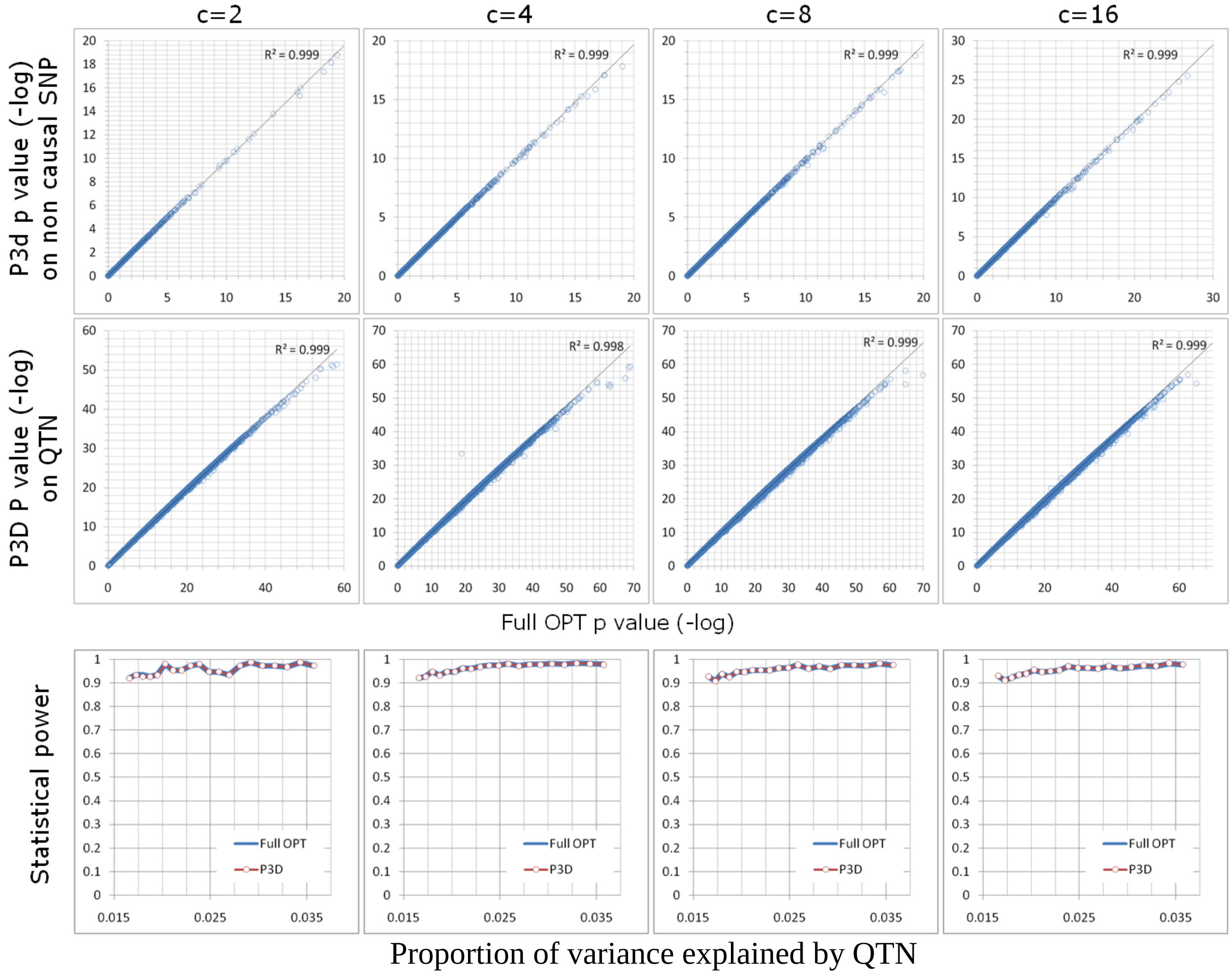
Supplementary Figure 2. Fit of compressed Mixed Linear model (MLM) at different compression levels for three human phenotypes (physical activity, triglyceride, and cholesterol)[3]. A total of 1315 individuals were clustered into groups by implimenting the Unweighted Pair Group Method with arithmetic Average (UPGMA) clustering algorithm using Proc Cluster in SAS[2]. Model fit is indicated by the negative log likelihood (-LL), Adjusted Akaike Information Criterion (AICC), and Bayesian information content (BIC). Smaller values of –LL, AICC and BIC indicate better fit. Compression at a level of one is equivalent to the standard MLM. Compared to the standard MLM, all phenotypes had better fit with compressed MLM at one or more of the tested compression levels. The compression level with the best model fit was variable for each of the three traits.



Supplementary Figure 3. The *P* values and statistical power of association tests by using the one step mixed linear model with full optimization (Full OPT) for all unknown parameters and "Population Parameters Previously Determined" (P3D) on a maize phenotype simulated with two heritability level and either five or twenty QTNs (Quantitative Trait Nucleotides). The heritability was defined as the proportion of additive genetic variance over the total variance, which was the sum of additive genetic variance and residual variance. The additive genetic effect was simulated from the QTNs assigned to the SNPs (Single Nucleotide Polymorphisms) in the maize dataset[5]. As all individuals were inbred lines, no dominant effect was included. The experiment was repeated 1000 times. For each replicate, the number of non causal SNPs that were randomly sampled was the same as the number of causal QTNs. The top two panels display the *P* values for the full OPT (X axis) and P3D (Y axis). Each dot represents a test on a non causal SNP (the top panel) and a causal QTN (the middle panel). The *P* values from P3D are highly correlated with the ones from the full OPT for the non causal SNPs and causal QTNs (R^2>97%). The empirical statistical power for detecting the causal QTNs was displayed (the bottom panel) as function of the proportion of the total variation explained (X axis). The P3D approach and the full OPT had approximately the same statistical power for detecting the causal QTNs.



Supplementary Figure 4. The *P* values and statistical power of association tests by using the one step mixed linear model with full optimization (Full OPT) for all unknown parameters and "Population Parameters Previously Determined" (P3D) on a dog phenotype simulated with different dominant effects (D). The phenotype was controlled by 20 QTNs (Quantitative Trait Nucleotides) which were randomly assigned to the SNPs (Single Nucleotide Polymorphisms) in the dog dataset[4]. A heritability of 0.5 was defined as the proportion of additive genetic variance over the total variance, which was the sum of additive genetic, dominance and residual variances. It was assumed that there was no epistatic effect. The experiment was repeated 1000 times. For each replicate, the number of non causal SNPs that were randomly sampled was the same as the number of causal QTNs. The top two panels display the *P* values of for the full OPT (X axis) and P3D (Y axis). Each dot represents a test on a non causal SNP (the top panel) and a causal QTN (the middle panel). The *P* values from P3D are highly correlated with the ones from the full OPT for the non causal SNPs and causal QTNs (R^2>99%). The empirical statistical power for detecting the causal QTNs was displayed (the bottom panel) as function of the proportion of the total variation explained (X axis). The P3D approach and the full OPT had approximately the same statistical power for detecting the causal QTNs.



Supplementary Figure 5. The *P* values and statistical power of association tests by using the one step mixed linear model with full optimization (Full OPT) for all unknown parameters and “Population Parameters Previously Determined” (P3D) on a human phenotype simulated with different compression levels. The phenotype was controlled by 20 QTNs (Quantitative Trait Nucleotides) which were randomly assigned to the SNPs (Single Nucleotide Polymorphisms) from the human dataset[3]. A heritability of 0.5 was defined as the proportion of additive genetic variance over the total variance, which was the sum of additive genetic variance and residual variance. It was assumed there were no dominance and epistatic effects. The experiment was repeated 1000 times. For each replicate, the number of non causal SNPs that were randomly sampled was the same as the number of causal QTNs. The top two panels display the *P* values for the full OPT (X axis) and P3D (Y axis). Each dot represents a test on a non causal SNP (the top panel) and a causal QTN (the middle panel). The *P* values from P3D are highly correlated with the ones from the full OPT for the non causal SNPs and causal QTNs (R^2>99%). The empirical statistical power for detecting the causal QTNs was displayed (the bottom panel) as function of the proportion of the total variation explained (X axis). The P3D approach and the full OPT had approximately the same statistical power for detecting the causal QTNs.

Supplementary Note

The implementation of compression and P3D in SAS involved the use of two SAS procedures, five SAS macros and one main SAS program. The two SAS procedures are Proc Mixed and Proc Cluster. Proc Mixed is used to solve a mixed linear model (MLM) to estimate variance components and perform statistical test on marker effects (F test). Proc Cluster is used to cluster individuals into groups with specific clustering algorithm and number of groups.

The five SAS macros and their functions are as follows:

1. SAS macro getAvgKin calculates kinship among groups based on kinship among individuals.
2. SAS macro LORG converts kinship into the format required by Proc Mixed.
3. SAS macro Matrixconvert converts a kinship matrix in a square format to three-column (row, column and value) format or reverse the conversion.
4. SAS macro setPars saves the variance estimates from the reduced model as SAS macro variables, which are used by the full model to test genetic marker effects.

The main SAS program is named FastQK.sas. The demonstration data is the maize data used in this study and can be downloaded from the Tutorial data set of TASSEL software package[6].

Optimization of model fit can be achieved by varying the compression level and changing the method option in Proc Cluster. Optimization of clustering algorithm and compression level can be achieved by iteratively changing the cluster algorithm and compression level specified by the SAS macro variables CA and CL, respectively.

Association tests can be performed with five options as follows:

1. Full optimization for variance components: For each testing marker, variance components are optimized for the combination of cluster algorithm and compression level specified earlier.
2. P3D: The estimates of variance components from the reduced model are used instead of re-estimating them. The statistical power is approximately the same as the first option (full optimization).
3. The residual approach: The residual from the reduced model is used as the dependent variable[7] in a General Linear Model without including the random genetic effect. Consequently no iteration is involved to estimate variance components. This approach has the same statistical power as the full optimization only for a trait with low heritability.
4. The BLUP approach: The best linear unbiased prediction (BLUP) of the random genetic effect from the reduced model is used as the dependent variable[8,9]. This approach has the same statistical power as the full optimization only for a trait with high heritability.
5. Regression on BLUP: Original phenotypes are still used as dependent variables. The BLUP from the reduced model is used as covariates. This option was proposed by an anonymous reviewer of this paper.

The first two options include the random genetic effect in the MLM. The last three options do not include the random genetic effect, consequently, they use GLM. As the SAS Proc Mixed and Proc GLM have different output, Proc Mixed are actually used to solve the GLM problem in the last three options for the convenience of saving the results, although there is no random statement for the Proc Mixed in the last three options.

References

1. Romesberg HC (2004) Cluster Analysis for Researchers. LULU Press, North Carolina, USA.
2. SAS II (2002.) SAS. Statistical Analysis Software for Windows, 9.0 ed. Cary, NC USA.
3. Lai CQ, Arnett DK, Corella D, Straka Rj, Tsai MY, et al. (2007) Fenofibrate effect on Triglyceride and Postprandial Response of Apolipoprotein A5 variants: The GOLDN study. Arterioscler thromb Vasc Biol: 1417-1425.
4. Zhang Z, Zhu L, Sandler J, Friedenberg SS, Egelhoff J, et al. (2009) Estimation of heritabilities, genetic correlations, and breeding values of four traits that collectively define hip dysplasia in dogs. American Journal of Veterinary Research 70: 483-492.
5. Yu JM, Pressoir G, Briggs WH, Bi IV, Yamasaki M, et al. (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature Genetics 38: 203-208.
6. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, et al. (2007) TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23: 2633-2635.
7. Aulchenko YS, de Koning D-J, Haley C (2007) Genomewide Rapid Association Using Mixed Model and Regression: A Fast and Simple Method For Genomewide Pedigree-Based Quantitative Trait Loci Association Analysis. Genetics 177: 577-585.
8. Calvo JH, Marcos S, Jurado JJ, Serrano M (2004) Association of the heart fatty acid-binding protein (FABP3) gene with milk traits in Manchega breed sheep. Anim Genet 35: 347-349.
9. Buckler ES, Holland JB, Bradbury PJ, Acharya CB, Brown PJ, et al. (2009) The genetic architecture of maize flowering time. Science 325: 714-718.