Final Project
Biochemistry 218 Computational Molecular Biology
Jong Lee

Recent Advances in Statistical Methods for Genome-Wide Association Studies

- How do we distinguish a needle from a string of hay?

Introduction to Genome-Wide Association Studies

Genome-wide association studies (GWAS) are done to identify relationships between genomic polymorphisms and disease states or traits of an organism. The completions of the Human Genome Project in 2003 and the International HapMap Project in 2007 have enabled finding genomic variations linked to an individual's risk of certain diseases. As of March 12 2010, 507 publications have already reported 2403 single nucleotide polymorphisms (SNPs) linked with human diseases or traits. (1) There are over three million SNPs genotyped by the HapMap Project and, therefore, still much more promising results from theses association studies are expected. (2)

However, the nature of GWAS brings up concerns such as verification of statistical significance when a polymorphism is thought to be linked with a disease or a trait. Genotyping SNPs from a pool of six billion bases from a sample and identifying significant SNPs across several thousands of samples is clearly not a straightforward statistics problem. In fact, the term 'population stratification' describes the common situation where the analyzed population of interest consists of subgroups that have different ancestry. (3) This creates conflict with the "golden hypothesis" of statistics where the individuals in a sample population have resulted from random mating and therefore the associations found are solely from the action of the identified polymorphisms.

Solutions to Population Stratification Problem in Genome-Wide Association Studies

To prevent spurious associations between random alleles and phenotypes, two methods, Genomic Control and Structured Association, can be used to reduce false positives from the association studies. In Genomic Control (GC), average inflation factor of null SNPs are calculated to cancel out the possible misrepresentation of unrelated alleles due to population stratification. On the other hand, Structured Association (SA) method assumes that the sample population is structured i.e. has subpopulation of common ancestry, to a certain degree. The method then identifies the subpopulations and corrects for their overrepresentation in the sample pool. However, both GC and SA methods have their own caveats. While GC has limited applicability to a single SNP analysis, SA is computationally demanding since no definitive number of subpopulations can ever be determined. (3)

In 2006, Yu *et al.* have proposed a unified mixed-model approach (MMA) that can treat multiple levels of relatedness. (4) Their association mapping method "integrates genomic tools to uncover population structure and familial relationships with the traditional mixed-model framework that has long been used by animal geneticists." It is flexible and can adjust to various populations with or without substructures. This approach, also known as mixed linear model (MLM), outperforms others in reducing the number of false positives and false negatives, as seen in the figure below.

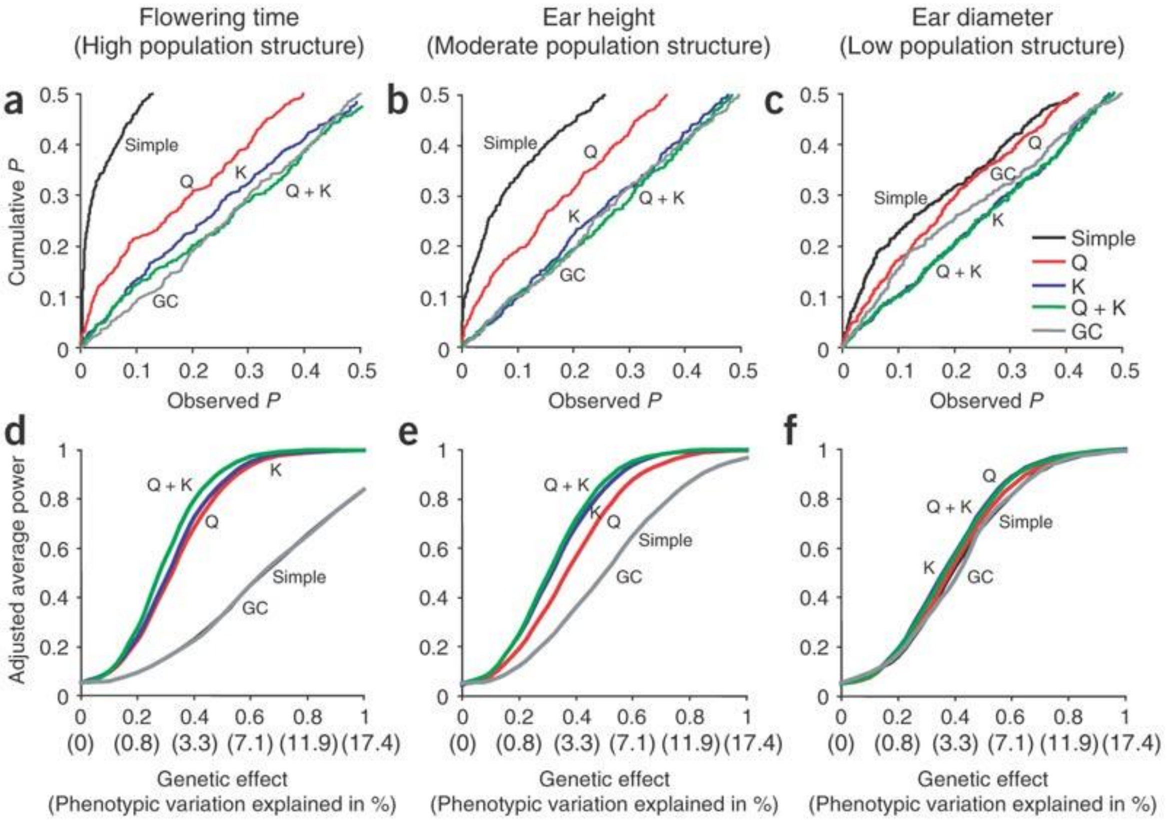


Figure 4. Model comparison with maize quantitative traits. [Reference (4)]
(**a**–**c**) Evaluation of the model type I error rates using random SNPs for flowering time (**a**), ear height (**b**) and ear diameter (**c**). The cumulative distributions of observed *P* values are presented for the simple model, the Q model, the K model, the Q + K model and the simple model with genomic control (GC). Under the expectation that random SNPs are unlinked to the polymorphisms controlling these traits (H_0: no SNP effect), approaches that appropriately control for type I errors should show a uniform distribution of *P* values (a diagonal line in these cumulative plots). The simple model was included only for the purpose of illustrating the effect of ignoring population structure and family relationships, as it is not a standard practice. (**d**–**f**) The adjusted average power of the models for flowering time (**d**), ear height (**e**) and ear diameter (**f**). A genetic effect was added to each random SNP (QTN effect), where k = 0.1, 0.2, 0.5, 0.7, 0.9 and 1.0 times the standard deviation of the phenotypic mean of a trait. Each model was adjusted based on its empirical type I error rate. The adjusted average power for GC is the same as that of the simple model with the empirical threshold *P* value. For convenience of comparison, we list the point value of phenotypic variation explained by a QTN at the allele frequency of p = 0.3.

However, the downside of MMA is that the computational power required increases rapidly as population size increases. According to Zhang *et al.*, the standard MMA involves a O(mpn^3) process, where m is the total number of markers, p is the number of iterations done and n is the number of individuals in a sampled population. (6) (7) The current GWAS on humans may handle about a million markers with a sample size in the order of thousands or, for certain meta-analysis studies, tens of thousands of individuals to identify significant and reliable linkages for certain diseases and traits. (2) (6) Since MMA can analyze a human dataset with 1,315 individuals in about 800 s CPU time for one marker, it may take decades of CPU time if a large-scale GWAS on humans were to be done using the method extensively. (6)

Recent Improvements in Mixed Model Approach

Very recently, two papers were published describing modification to the MMA scheme to increase the computing speed by three to four orders of magnitude. Kang *et al*. showed that, through estimation to the sample structure based on a variance component model, MMA approach gives better performance in association studies than previously reported GC or principal component analysis methods on the identical sample population (Northern Finland Birth Cohort and Wellcome Trust Case Control Consortium). The implementation of this approach was released in publicly available software called EMMAX (Efficient Mixed Model Association eXpedited). (5) (8) On the other hand, Zhang *et al*. have improved MMA by reducing the number of samples by compressing individual data into clustered groups and avoiding reiteration of variance components. They also released implemented software called TASSEL (Trait Analysis by aSSociation, Evolution and Linkage). (6) (9) Remarkably, each approach achieved up to 8000 fold reduction in computing time. The two approaches, however, lose little, if not none, of the statistical significance acquired from original MMA and even sometimes provide results with better significance than previous methods. Two data tables and a figure from the work of Kang *et al*. (5) are shown below to emphasize superior significance of associated SNPs and lower inflation factors by EMMAX by comparison with previous methods.

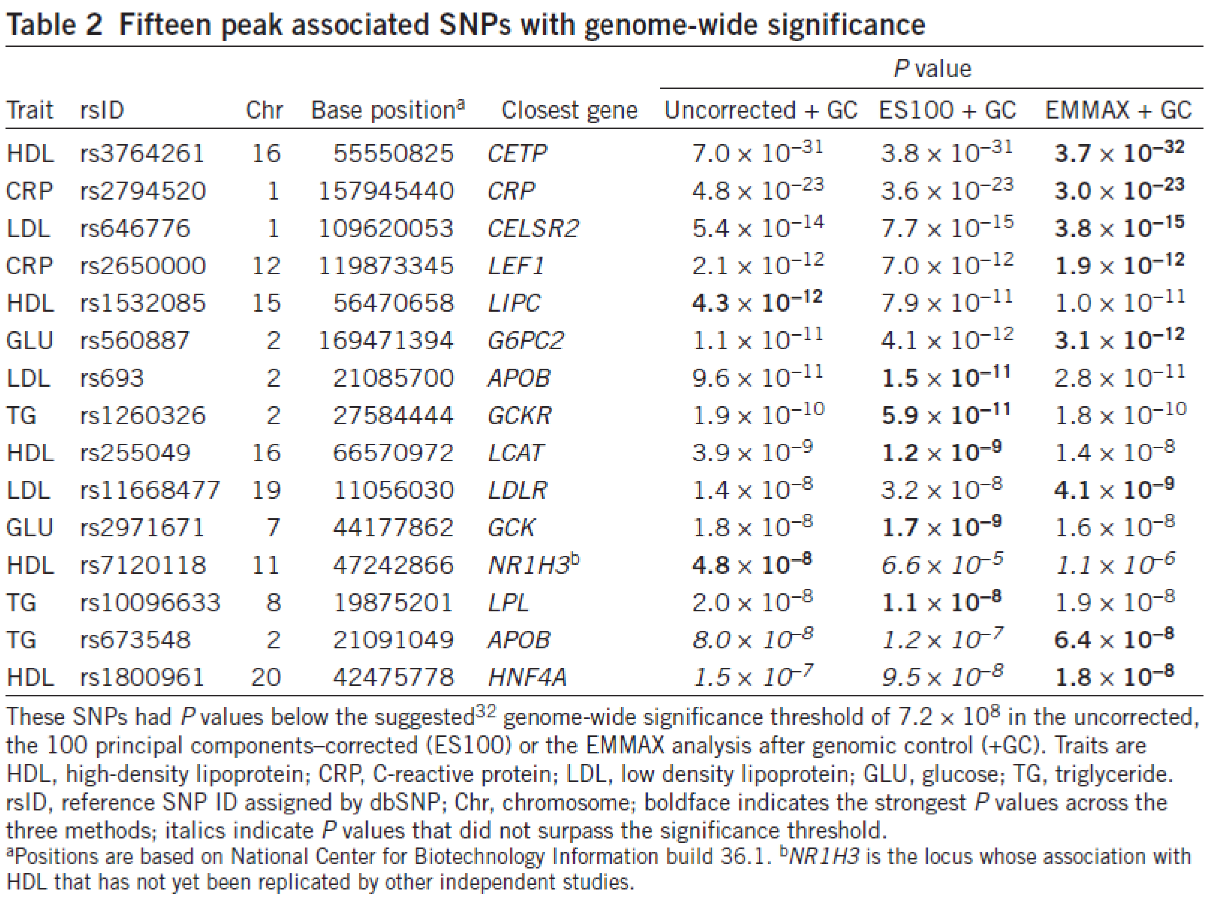


Table 2 Fifteen peak associated SNPs with genome-wide significance

					P value		
Trait	rsID	Chr	Base position[a]	Closest gene	Uncorrected + GC	ES100 + GC	EMMAX + GC
HDL	rs3764261	16	55550825	*CETP*	7.0×10^{-31}	3.8×10^{-31}	**3.7×10^{-32}**
CRP	rs2794520	1	157945440	*CRP*	4.8×10^{-23}	3.6×10^{-23}	**3.0×10^{-23}**
LDL	rs646776	1	109620053	*CELSR2*	5.4×10^{-14}	7.7×10^{-15}	**3.8×10^{-15}**
CRP	rs2650000	12	119873345	*LEF1*	2.1×10^{-12}	7.0×10^{-12}	**1.9×10^{-12}**
HDL	rs1532085	15	56470658	*LIPC*	**4.3×10^{-12}**	7.9×10^{-11}	1.0×10^{-11}
GLU	rs560887	2	169471394	*G6PC2*	1.1×10^{-11}	4.1×10^{-12}	**3.1×10^{-12}**
LDL	rs693	2	21085700	*APOB*	9.6×10^{-11}	**1.5×10^{-11}**	2.8×10^{-11}
TG	rs1260326	2	27584444	*GCKR*	1.9×10^{-10}	**5.9×10^{-11}**	1.8×10^{-10}
HDL	rs255049	16	66570972	*LCAT*	3.9×10^{-9}	**1.2×10^{-9}**	1.4×10^{-8}
LDL	rs11668477	19	11056030	*LDLR*	1.4×10^{-8}	3.2×10^{-8}	**4.1×10^{-9}**
GLU	rs2971671	7	44177862	*GCK*	1.8×10^{-8}	**1.7×10^{-9}**	1.6×10^{-8}
HDL	rs7120118	11	47242866	*NR1H3*[b]	**4.8×10^{-8}**	*6.6×10^{-5}*	*1.1×10^{-6}*
TG	rs10096633	8	19875201	*LPL*	2.0×10^{-8}	**1.1×10^{-8}**	1.9×10^{-8}
TG	rs673548	2	21091049	*APOB*	*8.0×10^{-8}*	*1.2×10^{-7}*	**6.4×10^{-8}**
HDL	rs1800961	20	42475778	*HNF4A*	*1.5×10^{-7}*	*9.5×10^{-8}*	**1.8×10^{-8}**

These SNPs had *P* values below the suggested[32] genome-wide significance threshold of 7.2×10^{8} in the uncorrected, the 100 principal components–corrected (ES100) or the EMMAX analysis after genomic control (+GC). Traits are HDL, high-density lipoprotein; CRP, C-reactive protein; LDL, low density lipoprotein; GLU, glucose; TG, triglyceride. rsID, reference SNP ID assigned by dbSNP; Chr, chromosome; boldface indicates the strongest *P* values across the three methods; italics indicate *P* values that did not surpass the significance threshold.

[a]Positions are based on National Center for Biotechnology Information build 36.1. [b]*NR1H3* is the locus whose association with HDL that has not yet been replicated by other independent studies.

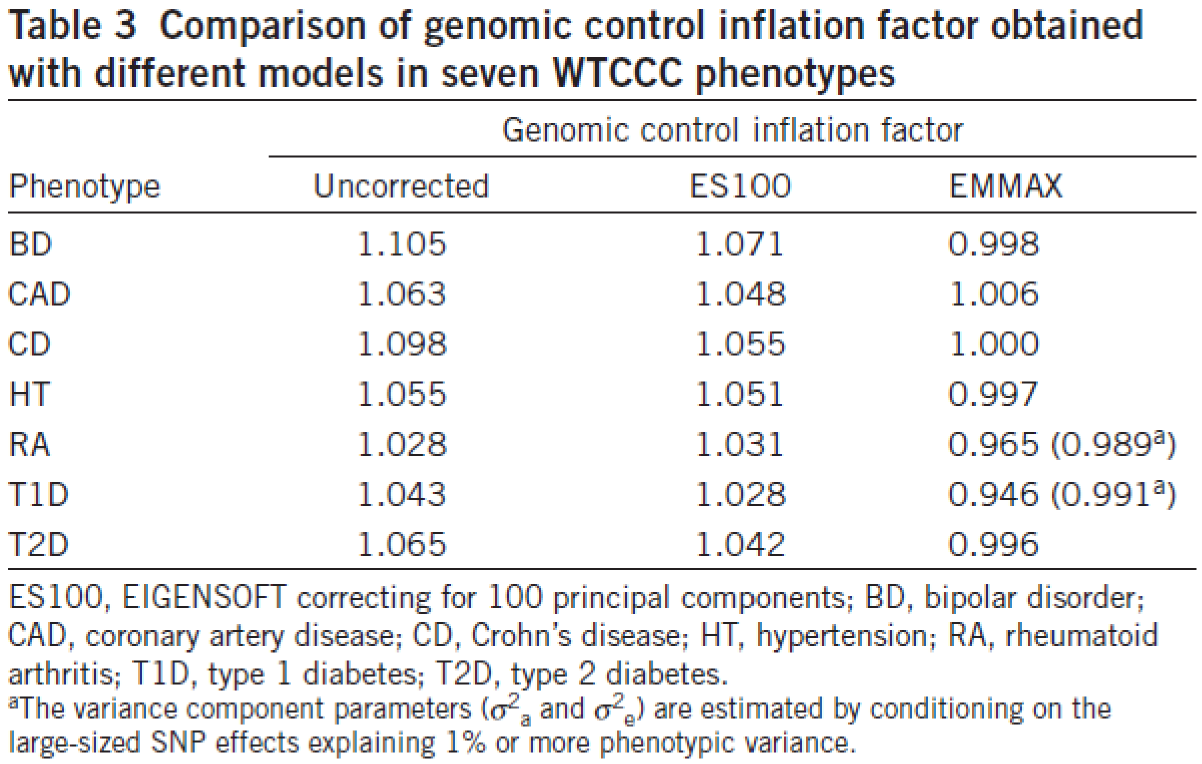
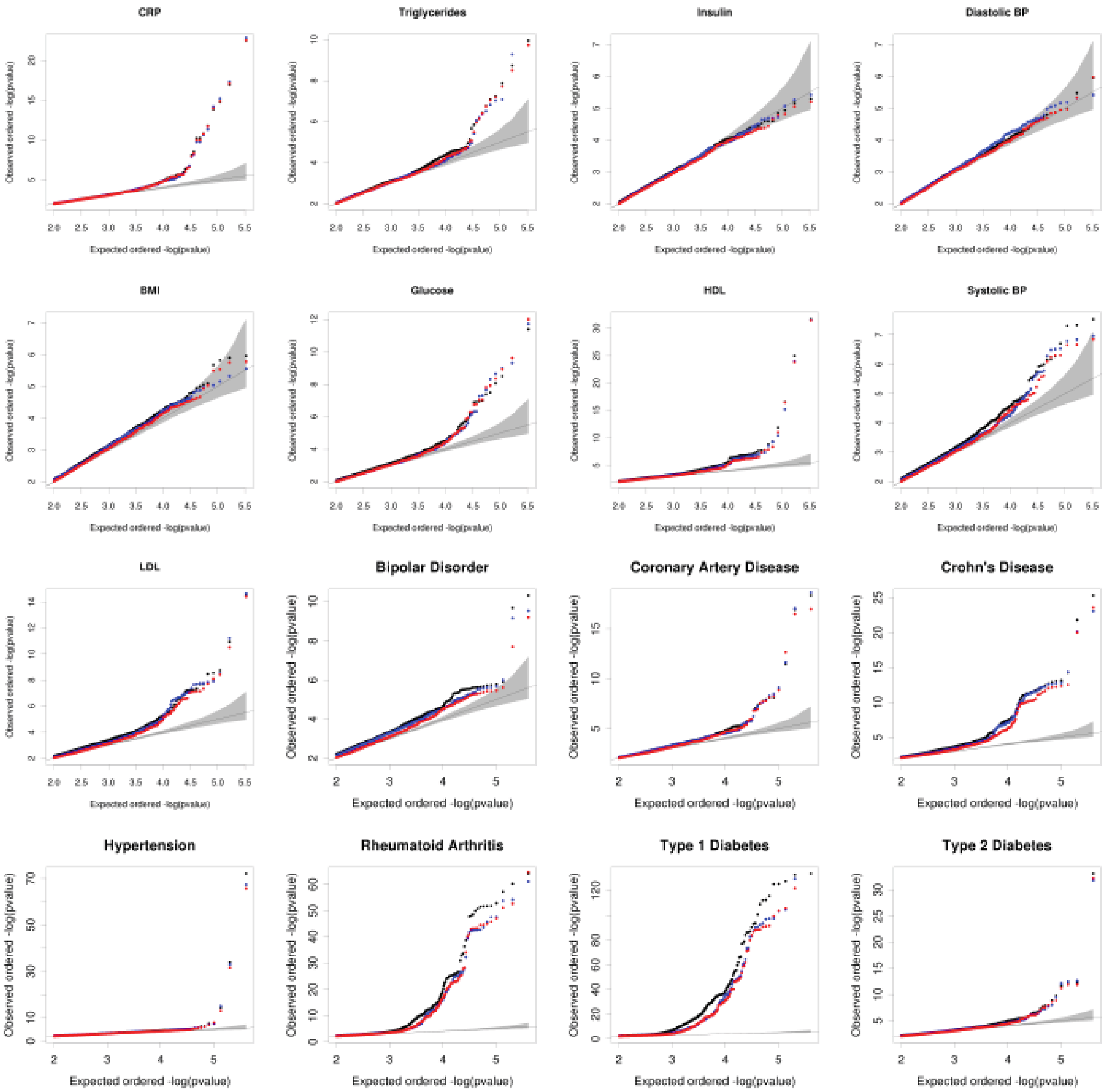


Table 3 Comparison of genomic control inflation factor obtained with different models in seven WTCCC phenotypes

	Genomic control inflation factor		
Phenotype	Uncorrected	ES100	EMMAX
BD	1.105	1.071	0.998
CAD	1.063	1.048	1.006
CD	1.098	1.055	1.000
HT	1.055	1.051	0.997
RA	1.028	1.031	0.965 (0.989[a])
T1D	1.043	1.028	0.946 (0.991[a])
T2D	1.065	1.042	0.996

ES100, EIGENSOFT correcting for 100 principal components; BD, bipolar disorder; CAD, coronary artery disease; CD, Crohn's disease; HT, hypertension; RA, rheumatoid arthritis; T1D, type 1 diabetes; T2D, type 2 diabetes.

[a]The variance component parameters (σ^2_a and σ^2_e) are estimated by conditioning on the large-sized SNP effects explaining 1% or more phenotypic variance.



Supplementary Figure 2: QQ-plots on the log10 scale of the association p-values obtained for nine traits according to three different models for 9 NFBC66 metabolic trais and 7 WTCCC disease phenotypes. In black, results from the unadjusted analysis; in blue results from the analysis conducted using 100 PC, and in red results from EMMAX.

Conclusion

Genome-wide association studies are powerful tools in identifying novel connections between genome and various diseases and traits. Nevertheless, high occurrences of false positives and false negatives in these association studies provide a statistical challenge in distinguishing the positives from the negatives. The situation is analogous to finding a needle in a hay stack. If one were to ignore that a well-dried string of hay may be as straight and as sharp as a needle, he or she might end up with something completely irrelevant when inappropriate criteria are applied. Mixed model approach (MMA) takes the population structure into account in analyzing associations between phenotypes and alleles. Moreover, two studies have recently reported evidences that MMA can be tractable in large-scale studies by applying estimations between the steps of analyses with improved precision without compromising the analytical power. These accomplishments enable MMA to be used more extensively in GWAS on humans, with greater number of markers and population size and complete the analyses in hours or days instead of years. Future studies may investigate the possible synergistic or additive effects on the combination of the two approaches to provide further improved performance. Although we have been already seeing the explosion of remarkable results from genome-wide association studies, we may be expecting to get most of the “needles” out of a haystack.

Reference

(1) Hindorff LA, Junkins HA, Mehta JP, and Manolio TA. A Catalog of Published Genome-Wide Association Studies. Available at: www.genome.gov/gwastudies. Accessed 03/12/2010

(2) The International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs, *Nature*, *449*, 851-861

(3) Balding, D. J. (2006) A tutorial on statistical methods for population association studies, *Nature Reviews Genetics*, *7*, 781-791

(4) Yu, J. *et al*. (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness, *Nature Genetics, 38(2)*, 203-208

(5) Kang, H. M. *et al*. (2010) Variance component model to account for sample structure in genome-wide association studies, *Nature Genetics, advance online publication*, 7 March 2010 (doi:10.1038/ng.548)

(6) Zhang, Z. *et al*. (2010) Mixed linear model approach adapted for genome-wide association studies, *Nature Genetics, advance online publication*, 7 March 2010 (doi:10.1038/ng.546)

(7) Zhang, Z., Buckler, E. S., Casstevens, T. M. & Bradbury, P. J. Software engineering the mixed model for genome-wide association studies on large samples, *Briefings in Bioinformatics*, *10 (6)*, 664-675

(8) http://genetics.cs.ucla.edu/emmax/

(9) http://www.maizegenetics.net/tassel/