



# FastANOVA: an Efficient Algorithm for Genome-Wide Association Study

Department of Computer Science

University of North Carolina at Chapel Hill

March 2012

## Genotype - Phenotype Association Study

Genome-wide association study searches for the genetic factors underlying phenotype variations. The most abundant source of genetic variations are Single Nucleotide Polymorphisms (SNPs), which serve as genetic markers of locations in the genome. Many phenotypes of interests are quantitative variables. These phenotypes are often complex traits and caused by the joint effect of multiple genetic factors. We study the problem of how to speed up the two-locus (SNP-pairs) association study to the genome-wide scale.

**Challenges:** 1. Large number of SNPs – the number of SNPs in public databases ranges from thousands to millions. 2: Statistical significance – multiple test increases type I error. Large permutation test is needed to properly control family-wise error rate. These two challenges impose enormous search space for complete genome-wide association study.

<http://www.bcgsc.ca>



Mouse genome

<http://www.jax.org/>



Phenotypic variations

Figure 1. Left: Twenty mouse chromosomes. Right: Mice showing phenotypic differences. The goal of genotype-phenotype association study is to identify the genetic factors causing phenotypic variations.

## The FastANOVA Algorithm

**Goal:** To scale the two-locus ANOVA test to genome-wide even when large permutation test is required.

**Question:** Do we have to test every SNP-pair and repeat for all phenotype permutations?

**Idea:** Develop an upper bound that can be efficiently calculated and easily incorporated in the algorithm to dramatically prune the search space.

$$SS_B(X_i X_j, Y) \leq SS_B(X_i, Y) + R_1 + R_2$$

↑ constant
↑  $f(n_a)$ 
↑  $f(n_b)$

$$n_a = \min \{ \# X_j = 1, \# X_j = 0 \mid X_i = 0 \}$$

$$n_b = \min \{ \# X_j = 1, \# X_j = 0 \mid X_i = 1 \}$$

Figure 2. The upper bound on  $SS_B(X_i X_j, Y)$ . Given SNP  $X_i$  and phenotype  $Y$ , the right hand side of the inequality only depends on the genotype of  $X_j$ .

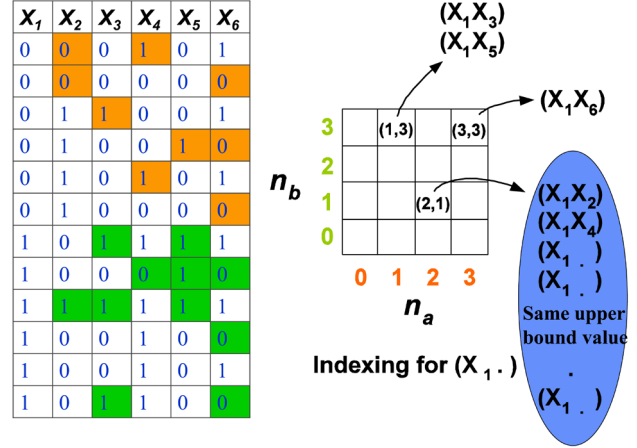


Figure 3. Applying the upper bound to index the SNP-pairs. The SNP-pairs are indexed in the 2D space of  $(n_a, n_b)$ . All SNP-pairs indexed by the same entry share the same upper bound value, which allows to compute the upper bound for a large group of SNPs together.

## Empirical Results

**Dataset:** #SNPs = 44k, #individuals = 26, phenotype: metabolism (water intake). SNP and phenotype data available at <http://www.jax.org>.

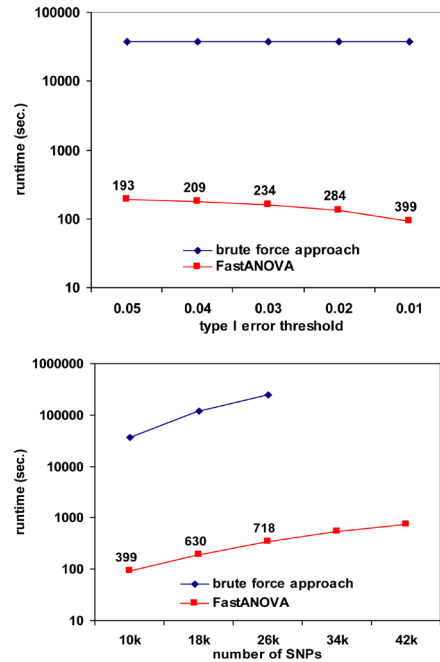


Figure 4. Runtime comparison between FastANOVA and the brute force approach. FastANOVA is two-order of magnitude faster than the alternative approach.

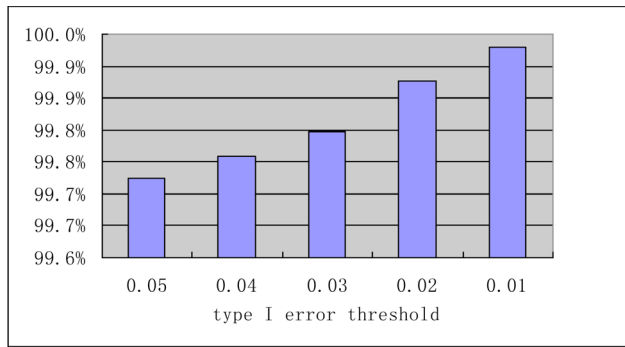


Figure 5. The percentage of the SNP-pairs pruned by applying the upper bound. Most of the SNP-pairs are pruned without performing any test.

### Current Project Members

Wei Wang, faculty member  
 Xiang Zhang, graduate research assistant  
 Fei Zou, graduate research assistant

### For More Information

Wei Wang  
 Professor  
 919-962-1744  
[weiwang@cs.unc.edu](mailto:weiwang@cs.unc.edu)

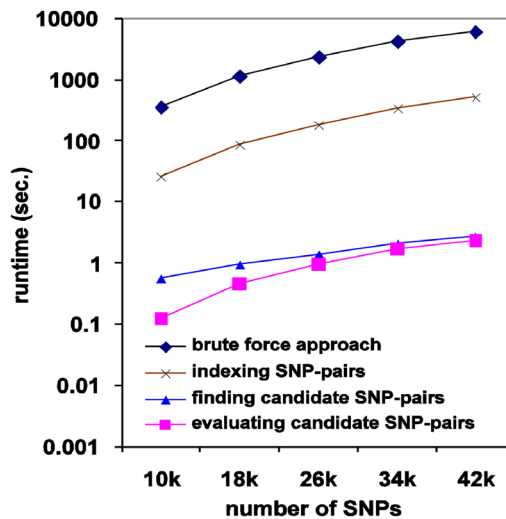
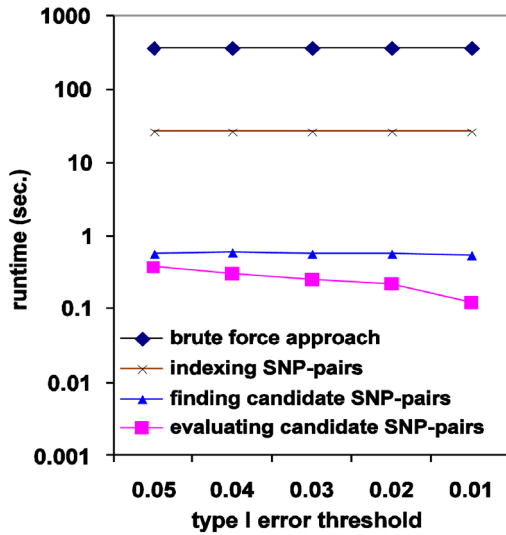


Figure 6. The runtime of each component of FastANOVA, including indexing the SNP-pairs, retrieving candidates, and evaluating candidate test values.