**FastEpistasis: a high performance computing solution for quantitative trait epistasis**

Thierry Schüpbach¹, Ioannis Xenarios¹, Sven Bergmann²,³ and Karen Kapur²,³,∗

¹Vital-IT Group, ²Molecular Modeling Group, Swiss Institute of Bioinformatics, Lausanne and ³Department of Medical Genetics, University of Lausanne, Rue du Bugnon 27, CH-1005 Lausanne, Switzerland

Associate Editor: David Posada

**ABSTRACT**

**Motivation:** Genome-wide association studies have become widely used tools to study effects of genetic variants on complex diseases. While it is of great interest to extend existing analysis methods by considering interaction effects between pairs of loci, the large number of possible tests presents a significant computational challenge. The number of computations is further multiplied in the study of gene expression quantitative trait mapping, in which tests are performed for thousands of gene phenotypes simultaneously.

**Results:** We present FastEpistasis, an efficient parallel solution extending the PLINK epistasis module, designed to test for epistasis effects when analyzing continuous phenotypes. Our results show that the algorithm scales with the number of processors and offers a reduction in computation time when several phenotypes are analyzed simultaneously. FastEpistasis is capable of testing the association of a continuous trait with all single nucleotide polymorphism (SNP) pairs from 500,000 SNPs, totaling 125 billion tests, in a population of 5000 individuals in 29, 4 or 0.5 days using 8, 64 or 512 processors.

**Availability:** FastEpistasis is open source and available free of charge only for non-commercial users from http://www.vital-it.ch/software/FastEpistasis.

**Contact:** karen.kapur@unil.ch

**Supplementary information:** Supplementary data are available at Bioinformatics online.

Received on February 24, 2010; revised on March 31, 2010; accepted on March 31, 2010

1 INTRODUCTION

Genome-wide association studies (GWASs) have been instrumental in identifying genetic variants associated with complex traits such as human disease or gene expression phenotypes (Hirschhorn et al., 2005). While many GWAS results have been reported analyzing single nucleotide polymorphisms (SNPs) one-at-a-time, only recently have studies begun to extend analysis methods to analyzing single nucleotide polymorphisms (SNPs) one-at-a-time. The test for epistasis reduces to testing whether the interaction term is significantly different from zero.

FastEpistasis methods are briefly outlined, with further details provided in the Supplementary Material. The computations are optimized by splitting the analysis tasks into three separate applications: pre-, core- and post-computation. The pre-computation phase loads PLINK binary format data files, reformats the data for faster computations and reduces the number of conditions to check in the core phase. The core computational phase is designed to embarrassingly parallelize the computations, iterating through SNP pairs and efficiently carrying out the tests for epistasis. The computations are based on applying the QR decomposition to derive least squares estimates of the interaction coefficient and its standard error. The core computation software comes in several versions to take advantage of different high-performance architectures—a Symmetric Multiprocessing (SMP) version and a clustered Message Passing Interface (MPI) version. An optional post-computation phase is provided to aggregate results from each processor or core, include detailed SNP information, compute P-values from each test, and convert to test files.

We assessed the performance of our software using International HapMap Project genotypes (Frazer et al., 2007) and random phenotypes (see supplementary material for details). Unless stated otherwise, results from all SNP pair epistasis tests are output.

3 RESULTS

We compared the performance of FastEpistasis and PLINK epistasis tests for several sets of SNP pairs, using a single core to enable
Averages are taken over 10 runs with SDs in parentheses. SNP pairs are derived from Epistasis tests per second completed by FastEpistasis core Table 1.

|                  | 10^3 tests (s) | Individuals | 10^3 tests (s) | 500 | 538.59 (3.9) | 5000 | 45.56 (0.4) |
|------------------|----------------|-------------|----------------|-----|--------------|------|-------------|
| 60               | 1393.14 (82.7) | 1000        | 289.44 (3.7)   | 100 | 1214.15 (38.4) | 3000 | 81.00 (0.7) |
|                  |                |             |                | 500 | 538.59 (3.9) |      |             |