Contents

Functions and Personnel Involved:

Allison Weber	Quantitative Genetics, Trait Genetics, Association Genetics
Shengchu Wang	Quantitative Genetics, Statistical Genetics
Daolong Wang	Quantitative Genetics, Statistical Genetics

Background

iPlant is an organization founded by the National Science Foundation (NSF)'s Plant Science Cyber-Infrastructure Collaborative (PSCIC). The purpose of iPlant is to build a unified cyber-infrastructure for the plant community at large (public and private) in order to consolidate the most helpful tools currently available in the community, bring them up to industry standard design principles and maintain them long-term. This will facilitate advances in the field by supporting computationally intensive analysis of ever growing datasets (http://www.iplantcollaborative.org/about). In order to accelerate the development of this community resource, Syngenta has agreed to provide industry-scale simulation datasets in order to begin to test the computing capabilities of iPlant, specifically regarding association mapping tools. The goal is to assess the performance of iPlant tools in comparison to other options available to researchers. Syngenta will have the benefits of using the tool earlier than others, and may influence the tool development.

Datasets to be provided

We have agreed to provide two sets of simulated data with known-truth phenotype-genotype associations, based on our real data structures:

- Set #1: Data with no population structure;
- Set #2: Data with population structure.

For each set, 100 replicate datasets were generated.

Methods

In order to provide simulated datasets with representative genome-wide linkage disequilibrium patterns, we generated data based on a real dataset within Syngenta that contains 512 maize inbred lines belonging to eight subgroups, 780K SNP markers across the genome and one phenotypic value (trait) after accounting for all experimental factors (locations, reps and treatments, etc.). Individual line, marker, chromosome, phenotype and subgroup identifiers were masked before simulation.

Simulation models:

Dataset #1: Containing no subgroup effect,

$$y_j = \mu + e_j + \sum_{i=1}^{q} g_{ij} \quad ,$$

where y is the phenotypic value; μ is the general mean; e_j is the random residual; $\sum_{i=1}^{q} g_{ij}$ is the combined genetic effects of a selected set of q associated SNPs. We simulated phenotypic values by using a general mean based on empirical data, plus a random residual from N(0, σ_e^2).

Dataset #2: Containing a subgroup effect,

$$y_{jh} = \mu_h + e_{jh} + \sum_{i=1}^{q} g_{ijh} \quad ,$$

where y is the phenotypic value; μ_h is the subgroup mean; e_{jh} is the random residual; $\sum_{i=1}^{q} g_{ijh}$ is the combined genetic effects of a selected set of q associated SNPs. We simulated phenotypic values by using a different μ_h per group based on empirical data, plus a random residual from N(0, σ_e^2).

For both simulated datasets, we maintained the relationship among markers across the genome for each individual.

Parameter settings:

For each replicate of the two datasets, we generated a set of known-truth associations in the following method.

1. Thirty-five SNPs (minor allele frequency ranging from 0.1 to 0.5) were selected for association with the phenotype. This set of 35 SNPs was fixed for all replicates.
2. Distribution of these SNPs along the 10 chromosomes:

Chromosome	Number
1	0
2	3
3	10
4	4
5	2
6	1
7	5
8	0
9	3
10	7

3. Four levels of genetic effect size were assigned to these SNPs. Their actual phenotypic contributions will be calculated according to heritability and model.

Number	Relative effect size
5	1.00
7	0.50
18	0.25
5	0.10

4. The heritability was set to 0.30, 0.40, and 0.60, by tuning σ_e^2 , each setting led to one simulated trait, resulting in three simulated phenotypes per dataset.

5. Model means were based on empirical phenotypic values.

Group	Mean	Sample
H	-1.15	118
J	-0.65	9
L	0.93	78
O	-4.02	44
S	3.08	74
T	-4.57	23
Y	2.62	67
Z	-1.53	99

For datasets without population structure, an overall mean of 0.057 was used.