A COMPARISON OF PRINCIPLE COMPONENT ANALYSIS AND FACTOR ANALYSIS FOR QUANTITATIVE PHENOTYPES ON FAMILY DATA

by

Xiaojing Wang

BS, Huanzhong University of Science and Technology, China 1998

MS, College of Life Science, Sichuan University, 2001

Submitted to the Graduate Faculty of

Department of Biostatistics

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2007

UNIVERSITY OF PITTSBURGH

Graduate School of Public Health

This thesis was presented

by

Xiaojing Wang

It was defended on

April 5th, 2007

and approved by

Thesis Advisor:
Eleanor Feingold, Ph.D. Associate Professor
Department of Biostatics and Human Genetics
Graduate School of Public Health
University of Pittsburgh

Committee Member:
Stewart Anderson, Ph.D. Professor
Department of Biostatics
Graduate School of Public Health
University of Pittsburgh

Committee Member:
Candace Kammerer, Ph.D., Associate Professor
Department of Human Genetics
Graduate School of Public Health
University of Pittsburgh

Eleanor Feingold, Ph.D.

A COMPARISON OF PRINCIPLE COMPONENT ANALYSIS AND FACTOR ANALYSIS FOR QUANTITATIVE PHENOTYPES ON FAMILY DATA

Xiaojing Wang, M.S.

Abstract

Background: Multivariate analysis, especially principal component analysis (PCA) and factor analysis (FA) is one of the effective methods by which to uncover the common factors (both genetic and environmental) that contribute to complex disease phenotypes, such as bone mineral density for osteoporosis. Although PCA and FA are widely used for this purpose, a formal evaluation of the performance of these two multivariate methodologies is lacking.

Method: We conducted a comparison analysis using simulated data on 500 individuals from 250 nuclear families. We first simulated 7 underlying (unobserved) genetic and environmentally determined traits. Then we derived two sets of 50 complex (observed) traits using algebraic combinations of the underlying components plus an error term. We next performed PCA and FA on these complex traits and extracted the first factor/principal component. We studied three aspects of the performance of the methods: 1) the ability to detect the underlying genetic/environmental components; 2) whether the methods worked better when applied to raw traits or to residuals (that is, after regressing out potentially significant environmental covariates); and 3) whether heritabilities of composite PCA and FA phenotypes were higher than those of the original complex traits and/or underlying components.

Results: Our results indicated that both multivariate analysis methods behave similarly in most cases, although FA is better able to detect predominant signals from underlying trait, which may improve the downstream QTL analysis. Using residuals (after regressing out potentially significant environmental covariates) in the PCA or FA analyses greatly increases the probability that PCs or factors detect common genetic components instead of common environmental factors, except if there is statistical interaction between genetic and environmental factors. Finally, although there is no predictable relationship between heritabilities obtained from composite phenotypes versus original complex traits, our results indicate that composite trait heritability generally reflects the genetic characteristics of the detectable underlying components.

Public health significance: Understanding the strengths and weaknesses of multivariate analysis methods to detect underlying genetic and environmental factors for complex diseases will improve our identification of such factors. and this information may lead to better methods of treatment and prevention.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

Numerous studies over the past several decades indicate that genes contribute to the development of complex diseases such as osteoporosis, obesity, and diabetes. Many risk factors for these diseases (such as bone mineral density, body fat, glucose levels) have been shown to be moderately to highly heritable. In recent years, many studies suggested that a majority of these highly heritable traits (risk factors) are governed by a set of common genes (i.e. pleiotropy, defined as when two or more phenotypes are co-regulated by a common gene or a common sets of genes) (1-4). The evidence supporting the above hypothesis is that bivariate linkage analyses of some of these traits revealed stronger linkage signals than were obtained from univariate linkage analysis of each trait separately (5-9).

Conventional measurements of these complex disease-related phenotypes produce many intercorrelated phenotypes. For example, bone mineral density (BMD) could be measured by peripheral Quantitative Computed Tomography (pQCT) at distal and shaft sites for both radius and tibia. High phenotypic and genetic correlations were observed from these bone phenotypes due to the common contributions from trabecular and cortical components. Therefore it is possible that there might be a relatively small number of factors (both genetic and environmental) involved in certain metabolic pathways that contribute to variation in an underlying cluster of phenotypes. Identification of these common factors and elucidation of their molecular basis should contribute to a better understanding and possible treatment for some complex diseases.

It is well-known that bivariate and tri-variate genetic analyses are computationally intensive. And genetic analyses of more than three traits are beyond our current methodologies. Therefore, using multivariate analysis (PCA /Factor analysis) might be an alternative yet effective solution to identify common genetic and environmental factors that affect multiple traits. Principal component analysis (PCA) and factor analysis (FA) both involve a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated (PCA) or correlated (FA) variables called principal components or factors. During the PCA/FA extraction, the shared variance of a variable is partitioned from its unique variance and error variance to reveal the underlying factor / PC structure. Only the shared variance appears in the solution. So it is reasonable to believe that these two methods have the potential to classify phenotypic variation into independent / dependent components that may amplify or purify genetic signals and hence be used to dissect genetic networks regulating complex biological systems.

2.0 REVIEW OF THE RELEVANT LITERATURE

Since 2001, ten groups of investigators that we are aware of have published articles that used multivariate analysis (MA) methods in an attempt to dissect the genetic and environmental basis for complex diseases, such as osteoporosis, metabolic syndrome, and asthma. Seven of these groups applied PCA (10-16), while the other three used FA (17-19). In addition, 7 groups used raw phenotypes directly as the input variables, one group used raw traits but performed analysis by gender and generation (15), and the last two groups used residuals (after adjustment for significant covariates) (12,19). The goals of the 10 groups also differed: one group used MA for phenotype clustering/classification, by which it developed composite index scores summarizing characteristics of raw traits from different skeletal sites (18). The remaining 9 groups all focused on exploring the underlying genetic/environmental basis of composite traits (that is, principal components or factors) derived from PCA or FA. Within these 9 groups, two reported genetic or environmental correlations between composite traits and some well-defined real (observed) phenotypes (11,14); two reports focused exclusively on heritability estimation for composite and real traits (12, 19); and three reports concentrated on the association (or linkage) between these composite traits and QTLs (Quantitative Trait Loci) ; (13,16,17) The final two papers did both heritability estimation and association/linkage analysis for composite phenotypes (10,15).

However, many statistical issues remain unaddressed by these reports. First**,** the selection of either PCA or FA seems arbitrary; none of the groups justified why they chose one instead the other. We decided to evaluate the performance of these two approaches. In particular, we wanted to assess which method is better able to detect the underlying environmental or genetic factors. Second, most reports used raw traits as input variable, but a few used residuals after regressing out some important environmental factors. Does analysis of residuals significantly improve the ability of PCA or FA methods to detect underlying genetic components? No direct comparisons to answer this question have been reported. Third, many groups have compared the heritability of composite traits (obtained from PCA or FA) with the original phenotypes. Does higher heritability of the composite trait compared to the original phenotype necessarily imply that the composite trait better reflects the underlying genetic components and thus increase the chance for detecting underlying genes? No literature that we are aware of has addressed this question.

Hence, the goal of this thesis is to explore the answers to the above three questions using simulated data on nuclear families.

3.0 METHODS

3.1 TERMINOLOGY

PCA analysis: Principal component analysis uses the information from the entire correlation coefficient matrix of a set of phenotypes to produce a smaller number of hypothetical ‘‘factors’’ (components) that help explain correlations among the original variables. The analysis was performed in two stages: (a) factor extraction using principal component analysis, and (b) rotation of the principal components using Varimax option. Definition of factors was done by extracting the eigenvalues (which represent the variance explained by each of the principal components) greater than certain threshold (15,17,20).

Factor Analysis: Factor analysis is a data-reduction procedure that uses the information from the entire correlation coefficient matrix of a large number of variables to produce a smaller number of hypothetical factor constructs or components that help explain the correlations between the variables. Factors are initially extracted from the correlation matrix until there is no appreciable variance left. Coefficients, weights, or loadings roughly represent the correlation between the variable item and the factor (17).

Heritability (H2): Heritability measures the proportion of phenotypic variance attributable to genetic variance. Typically, two heritabilities are estimated: narrow sense heritability (h2) and broad sense heritability (H2). Narrow-sense heritability gives the proportion of additive variance in the phenotypic variance, whereas broad-sense heritability measures the proportion of all genetic variance in the phenotypic variance (i.e. including additive, dominance and epistatic effects). (21) Heritability is estimated using information on the theoretical genetic relationships between different relative pairs. (22,23)

3.2 STUDY DESIGN

Our overall study design is illustrated in Figure 1. Three datasets of underlying (unobserved) traits were generated by simulation; 7 underlying traits (E1, E2, G1, G2, G3, S1 and S2) were involved in making these 3 datasets. The differences among these 3 datasets are the variance of the underlying environmental traits and the inclusion of S2 (a gene by sex interaction phenotype). For each of these three datasets of underlying traits, two sets of complex phenotypes were created using arbitrary algebraic functions of the underlying traits. There are 50 complex traits in each of the two function sets. Set 1 involves somewhat simpler algebraic combinations of traits than set 2 (details below). The seven underlying traits represent the unobserved environmental or/and genetic determinants which influence population variation of real traits, which are in turn represented by the sets of 50 complex traits. Using these complex traits, we created three different inputs for further multivariate analysis: raw traits, residuals model 1 (after regressing out E1 and E2), residuals model 2 (after regressing out E1, E2 and Sex). Finally, we

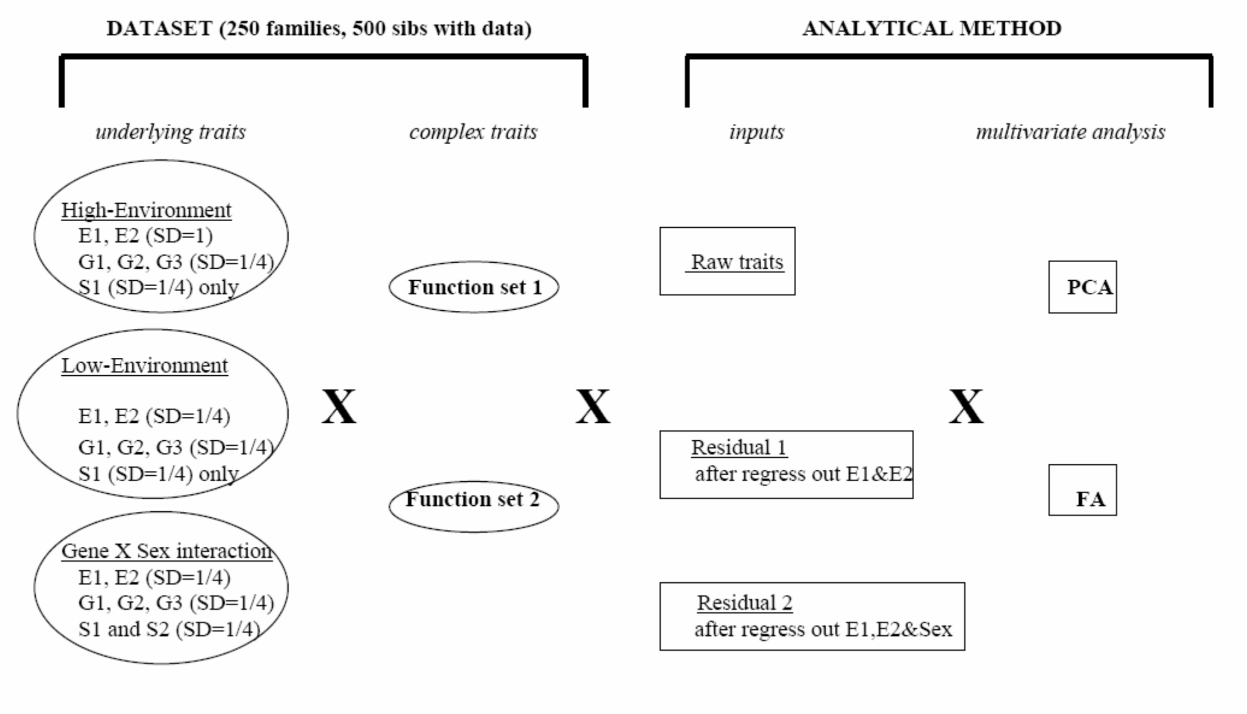


Figure 1 Blueprint for study design

performed both PCA and FA on each dataset x function set x residual combination, for a total of 36 analyses (see figure 1). Each aspect of the study design is described in more detail below.

We evaluated three aspects of the outcomes: 1) the ability to detect the underlying genetic/environmental components; 2) whether the methods worked better when applied to raw traits or to residuals (that is, after regressing out potentially significant environmental covariates).; and 3) heritabilities of composite traits (principal component or factor) comparing to 50 complex traits or 7 underlying traits.

3.3 SIMULATION OF UNDERLYING TRAITS

We first simulated 250 nuclear families with two parents and two offspring within each family. We then simulated 7 underlying original phenotypes (with corresponding genotypes): E1, E2, G1, G2, G3, S1, S2, (see Table 1) for offspring only, for a total of 500 individuals. All of these underlying traits were assumed to be normally distributed conditional on genotype. The simulated “genotypes” for E1 and E2 were not used in the model; these two traits were designed as an environmental model (pure environmental effect, no mean differences between people with different genotypes). Because some environmental factors are likely to be similar between siblings, we also allowed for the effect of a shared common environment for E1 and E2 by simulating these two traits based on bivariate normal distribution with means all equal 1, variance equals 1 or ½ for different dataset and covariance equals to 0.2 for E1 and 0.1 for E2. Another three traits (G1, G2 and G3) are standard simple genetic models (mean differences by genotype). As can be seen in table 1, the genotypic means and error variance for, G1, G2 and G3 are identical (mean 1.5, 2.5 and 3.5 for genotype aa, Aa and AA respectively and all SD=1/4); only the allele frequencies of these traits differ. The trait S1 has different means for males and females, but no interaction between sex and genotype. The trait S2 trait incorporates sex by genotype interaction.

Table 1 Simulation parameters for 7 underlying phenotypes

Parameter	Sex-Specific Genotype	E1	E2	G1	G2	G3	S1	S2
Mean	♂-*aa*	**1**	**1**	**1.5**	**1.5**	**1.5**	**2**	**1.5**
	♂-*Aa*			**2.5**	**2.5**	**2.5**	**3**	**2.5**
	♂-*AA*			**3.5**	**3.5**	**3.5**	**4**	**3.5**
	♀-*aa*			**1.5**	**1.5**	**1.5**	**1**	**1**
	♀-*Aa*			**2.5**	**2.5**	**2.5**	**2**	
	♀-*AA*			**3.5**	**3.5**	**3.5**	**3**	
SD		**1 or 1/4**	**1 or 1/4**	**1/4**	**1/4**	**1/4**	**1/4**	**1/4**
Allele Frequency	**P(a)** **P(A)**	**0.8** **0.2**	**0.5** **0.5**	**0.8** **0.2**	**0.9** **0.1**	**0.95** **0.05**	**0.8** **0.2**	**0.7** **0.3**

3.4 COMPLEX TRAITS

Based on the above underlying “unobserved” traits, we created first set of 50 complex “observed” traits, each of which is an algebraic combination of a subset of the 7 unobserved traits plus the error function (normal distribution with mean 1 and standard deviation 1). The objective of our choices reflects the current genetic/epidemiological assumptions about complex traits regarding the effects of underlying immeasurable genetic / environment factors. For instance, we used additive and multiplicative effects and combinations within and/or between underlying genetic and environmental traits; Moreover, we also included very complicated models in addition because we wanted to assess if PC and factor analysis could recover underlying traits even from extremely sophisticated conditions.

In order to assess even more complex models, we then created another set of 50 complex traits, in which we removed some of the algebraically simpler combinations and substituted more complex ones. All these new 50 functions were similar in format to those complicated ones in the first set of functions (e.g. C41-C50, refer to table 2). When devising our 50 complex traits for each set, we required that each underlying trait have a similar representation across all 50 complex traits. Based on our function summary file for dataset 1 and 2 (Tables 2 and 3), the proportion of times a specific underlying variable (e.g. G1) was included in the definition of a complex trait across all complex traits was as follows: 60% for E1, 54% for E2, 58% for G1, 54% for G2, 56% for G3 and 58% for S1. In dataset 3, we simply substituted S1 for G3 and S2 for S1, so the proportions are the same. Table 2 and 3 is the list of formulas for all 100 complex traits.

These complex traits represent phenotypes that we could observe or directly measure in reality, such as bone mineral density (BMD), body mass index (BMI), glucose level, and blood pressure; whereas the seven original traits represent underlying genetic or environmental components, which contribute to the true variation of measured (complex) trait but are not actually observed or measured.

Table 2 First set of 50 complex traits

Addtion
C1 = e1 + e2 + error[*]
C2 = g1 + g3 + error
C3 = (g1 + g2 + g3)/3+error
C4 = g2 + s1 + error
C5 = e2 + g3 + s1+error

Table 2 continued

C6 = (g1 + g2 +g3+e1+e2+s1)/7+error
C7 = 0.5(e1) + e2 + error
C8 = e1 + 3.5(g3) + error
C9 = 2(g2) + 0.6(g3) + error
C10 = 1/3 * g1 + g2 + 2(g3) + error
C11 = 2(e1) + 1.4(g2) + s1 + error
C12 = e2 + g3 + 3.2(s1)+error
Multiplication and Division
C13 = (e1)(e2) + error
C14 = (e2 +1) / (e1 + 1) + error
C15 = (g2)(g1) + error
C16 = (g1 + 1) / (g3 + 1) + error
C17 = (g3)(g2) + error
C18 = (s1)(e2) + error
C19 = (s1 +1) / (g2 + 1) + error
C20 = (e1)(g3) + error
C21 = (e2)(g1)(s1) + error
C22 = (e1)(g2)(s1)+error
Combination of addition, subtraction, multiplication and division

Table 2 continued

$C23 = (e1 + 1) / ((g3)(s1) + 1) + \text{error}$
$C24 = (g2 + 1) / ((e2)(s1) + 1) + \text{error}$
$C25 = (e1)(e2) + g2 + \text{error}$
$C26 = (g3 + 1) / (g1 + 1) + e2 + \text{error}$
$C27 = (e2 + 1) / (e1 + 1) + s1 + \text{error}$
$C28 = (g3)(g2) + 3.2(s1) + \text{error}$
$C29 = (e2)(s1) + e1 + \text{error}$
$C30 = (s1 + 1) / (g2 + 1) - g3 + \text{error}$
$C31 = (e1)(g1) + 0.5(s1) + \text{error}$
$C32 = 3(g3 + 1) / (s1 + 1) - e2 + \text{error}$
$C33 = (e1)(g3) - (s1 + 1) / (e2 + 1) + \text{error}$
$C34 = (s1 + 1) / (g2 + 1) + (e2 + 1) / (g3 + 1) - 0.7(g1) + \text{error}$
Power, exponentiation, logarithm
$C35 = (g1 + \text{error})^{2}$
$C36 = (s1 + \text{error})^{3}$
$C37 = e^{(g3 + \text{error})}$
$C38 = \sqrt{(s1 + \text{error} + 2)}$
$C39 = \log (g1 + \text{error} + 2)$
Combination of all forms

Table 2 continued

C40 = $(g2)^{2} + (g1+1) / (s1 + 1)$ + error
C41 = $1 / (g3+2)^{3} - 0.7(g2) /$ + error
C42 = $\sqrt{(g3+2)} - s1(e1)$ + error
C43 = $\log((g1)(g2)+2) - \sqrt{(g3+2)}$ + error
C44 = $(g2 + s1 + 1) / ((g1)^{2} + 1)$+ error
C45 = $(g1 + 1) / ((g2)^{2/3} + 1)$ + error
C46 = $(g3 - g1)(s1) + \log(e2+2)$ + error
C47 = $(s1)^{2} + (g1 + g2)(2(e1) - e2)$ + error
C48 = $(e1 + 2(e2) + 1) / ((g1)^{2} + \sqrt{(g3+2)} + 1)$ + error
C49 = $1 / (\log(2(s1) + 2(g3) - e1 + 2) + 1)$ + error
C50 = $(\sqrt{(s1+g2^\wedge 2+e1^\wedge 3+2)+1}) / (e2 - (s1)^{3} + 0.5(g1)(g3) + 1)$ + error

Table 3 Second set of 50 complex traits

All 50 traits Combination of all forms
C1 = $\log(e1+5)(e2) - (s1/g3+2)^{2}$ + error
C2 = $4.4(g1) / \sqrt{(g2-e1+3)}$ + error

Table 3 continued

$C3 = (g1 + e2 + g3)/3 + \text{error}$

$C4 = (e2)^2 + (s1 + 1)/(g2 + 1) + \text{error}$

$C5 = (1.2(e1) + g2 - 3) / ((g3)^2 + \sqrt{(g1+2)} - 2) + \text{error}$

$C6 = (1.1(g1) + 2.7(g2) + (3/5)(g3) + e1 - 1.4(e2) + 2(s1))/7 + \text{error}$

$C7 = \log (e2 + 1.2) - \exp^{(g1 + \text{error})}$

$C8 = \log ((g1)(g2) + 2) - \sqrt{(s1+2)} + \text{error}$

$C9 = 2(g2) + (s1 + 2(e2) + 1) / (e2^2 + \sqrt{(g2+2)} + 1) + \text{error}$

$C10 = 2.3(g2) + e1 + 2(g3) + \text{error}$

$C11 = 0.2(e2) + 1.4(g3) + s1 + \text{error}$

$C12 = (g1/(s1-3)) / (e2^{(2/3)} + 1) + \text{error}$

$C13 = (e1^2 - 3)(\log (g1+3)) + \text{error}$

$C14 = (g2 + (1.2(g2) - 2)^3) / (s1 + 1) + \text{error}$

$C15 = \sqrt{(g2 + \text{error} + 2)} + (g1)^2 (e1)$

$C16 = (s1 + 1)/(g3 + 1) - e1 + \text{error}$

$C17 = (s1)(e2) + 3.3 (e^{(g3 + \text{error})})$

$C18 = \sqrt{(g2+2)} - (g1)(e1) + \text{error}$

Table 3 continued

C19 = (s1 +2.2(g3))/(1.7(g2) - 1.2) + error
$C20 = (s1)(g2) - 2/(g1+4)^{3} - 0.7(g3)/3 + error$
C21 = (s1)(g2)(e2) + error
C22 = g1/ (2.2- g3)(e2)(s1) + error
C23 = (e1 + 1) / ((g1)(g3) + 1) + error
C24 = (2.2(g2) - s1) / ((e2)(e1) + 1) + error
$C25 = (\sqrt{(e1+g2^{2}+g3^{3}+1)}+4) / (s1 - e2^{2} +0.5(g1)(g2) + 1) + error$
C26 = (s1 + 3)/(g3 + 1) - 2.1(e2) + error
C27 = (e2 + 2.1) / (e1 + 1.2) + g3 + error
C28 = (g3-1)(g2) + 3.2(e1)- e1+ error
C29 = (e2)(e1) + g1 + error
C30 = (s1 + 1) / (g2 + 1) - g1+ error
C31 = (e1)(g1) + 0.5(e2) + error
C32 = 3(g3 +1)/(s1 + 1.7) - 2(e1) + error
C33 = (g1)(e2) - (s1 +1) / (g2 + 1) + error
C34 = (s1 +1) / (g2 + 2) + (e2 +1) / (g3 + 1) - 0.7(g1) + error
$C35 = (g1 + error)^{2} + g1 / ((e1)(s1))$

Table 3 continued

$C36 = (s1 + error)^{3} - 2.4(\log (e2 + error + 2))$

$C37 = 3.3(e^{(g3 + error)}) + 1.4(e1 + error)^{2}$

$C38 = \sqrt{(s1 + error + 2)} + (g1)^{2} (e2)$

$C39 = \log (g1 + error + 2) - e^{(e1 + error)}$

$C40 = (g3)^{2} + (g2 + 1)/(e1 + 1) + error$

$C41 = 2 / (e1+5)^{3} - 0.7(g3) / + error$

$C42 = \sqrt{(e2 + 2)} - (s1)(g1) + error$

$C43 = \log (g2 * e1 + 2) - \sqrt{(g3 + 2)} + error$

$C44 = (g1 + e2 + 1) / (g3^{2} + 1) + error$

$C45 = (g1 + 1) / (s1^{(2/3)} + 1) + error$

$C46 = (g1 - e2)(g3) + \log(e1 + 2) + error$

$C47 = (e1)^{2} + (g2 + g3)(2(s1) - e2) + error$

$C48 = (s1 + 2 (g2) + 1) / (e1^{2} + \sqrt{(g3 + 2)} + 1) + error$

$C49 = 4.2 / (\log (2(s1) + 2(g3) - e1 + 2) - 3) + error$

$C50 = (\sqrt{(E2 + E1^{2} + G3^{3} + 2)} + 1) / (g1 - e1^{3} + 0.5(g1)(g2) + 1) + error$

3.5 DATASETS

For each set of functions above, we created three different datasets of underlying traits by simulation to evaluate the performance of the multivariate analysis methods. Datasets 1 and 2 use only 6 out of 7 underlying traits: E1, E2, G1, G2, G3 and S1 (see Tables 2 and 3 and figure 1). The only difference between these two datasets is the standard deviation of E1 and E2: instead of SD=1 in dataset 1, we changed to SD=1/4 (half of the variance) in dataset 2. For the third dataset, we substituted underlying trait S1 for G3 and S2 for S1. However, we kept the functions the same and SD=1/4 for E1 and E2. For example: for trait C49 in second set of functions, we used

$$4.2 / (\log\ (2(s1) + 2(g3) - e1 + 2) -3) + error$$

for dataset 1 and 2, and

$$4.2 / (\log\ (2(s2) + 2(S1) - e1 + 2) -3 + error$$

for dataset 3.

We designed these three datasets to perform the following comparisons: 1) By comparing analyses of dataset 1 and 2, we could compare how two multivariate analysis methods behave when trait variation due to environment decreases; in other words, the proportion of total phenotypic variance due to genetics increases. 2) By comparing analyses of dataset 2 and 3, we could evaluate the behavior of the analysis methods with and without the presence of sex X genotype interaction (by inclusion / exclusion of S2. (Refer to Fig 1.) For simplicity, we will refer to datasets 1, 2 and 3 in the subsequent text as the high-environment dataset, the low-environment dataset, and the gene by sex interaction dataset, respectively.

3.6 STATISTICAL ANALYSIS

The input variables for the multivariate analyses were either 50 complex traits in their original form (raw traits) or residuals of these traits (after removing effects of covariates). Two types of residuals were analyzed: (1) after adjusting for E1 and E2; or (2) after adjusting for E1, E2 and sex. Both residuals were created from 50 continuous traits by multiple regression after the incorporation of corresponding covariates. To mimic analysis methods that would be used in a real study, we only considered the linear form of covariates in the multiple regression, although we recognize there are quadratic or other non-linear effects of E1 and E2 in our arbitrary functions.

The Pearson pairwise correlations among all 50 complex traits (or residuals) were estimated using the R statistical package (V2.4.0 for windows).(24) Principal component and factor analysis were both performed in R using its standard default procedure (varimax rotation, correlation matrix use Pearson) with default option. (Command: *princomp* and *factanal)*.

3.7 EVALUATION

We limited all analyses and evaluations to only first component / factors which in theory account for the greatest proportions of variations from 50 complex traits. Two evaluation strategies were applied. First, we evaluated the ability of each method to detect common underlying environment or genetic components. We performed univariate regression analyses and regressed every underlying trait on the first composite trait (PC or factor). Correlations (R-Squares) between composite trait and underlying original trait were reported respectively. Second, we estimated heritability. For each trial, we estimated heritabilities for all phenotypes, including the 7 underlying traits, the 50 complex traits and the two composite traits (first principal component and first factor). We then compared these heritability estimates within each trial by box plot. The estimated heritability of a trait using data on full-sibs was calculated as: H^2 = 2 times (trait correlation between sibs). (25)

4.0 RESULTS

4.1 ANALYSIS OF CORRELATIONS

Table 4 summarizes all correlations (R-Squares from univariate regression) between composite traits and each underlying phenotype. We generated 3 independent replicates of all our 36 dataset/analysis combinations and obtained very similar results across these three replicates. We just report here the results from one replicate. (Table 4 and 5).

Table 4 Correlations between composite trait and underlying phenotype in function set 1

		Raw Traits		**Residual 1 -regress out E1 and E2**		**Residual 2 -regress out E1, E2&Sex**	
		Correlation **-Factor** *	Correlation **- PC**	Correlation **-Factor**	Correlation **-PC**	Correlation **-Factor**	Correlation **- PC**
High Environment Dataset	E1	0.90	0.45	~ 0	~ 0	~ 0	~ 0
	E2	<0.01	0.23	~ 0	~ 0	~ 0	~ 0
	G1	<0.01	~ 0	<0.01	0.02	<0.01	<0.01
	G2	0.01	0.03	0.01	0.02	0.28	0.22
	G3	<0.01	0.05	0.05	0.10	0.26	0.17
	S1	0.09	0.22	0.88	0.68	0.24	0.34
Low Environment Dataset	E1	0.57	0.24	~ 0	~ 0	~ 0	~ 0
	E2	<0.01	0.10	~ 0	~ 0	~ 0	~ 0
	G1	<0.01	<0.01	0.01	0.03	<0.01	0.01
	G2	0.08	0.18	0.06	0.16	0.34	0.28
	G3	0.02	0.09	0.02	0.07	0.21	0.12
	S1	0.36	0.43	0.90	0.75	0.25	0.32
Gene by Sex Interaction Dataset	E1	0.02	0.08	~ 0	~ 0	~ 0	~ 0
	E2	0.02	0.05	~ 0	~ 0	~ 0	~ 0
	G1	<0.01	<0.01	~ 0	<0.01	<0.01	0.05
	G2	0.06	0.09	0.04	0.07	0.23	0.22
	S1	0.49	0.41	0.41	0.48	0.25	0.14
	S2	0.82	0.70	0.90	0.82	0.12	0.15

Table 5 Correlations between compsite trait and underlying phenotype in function set 2

		Raw Traits		**Residual 1 -regress out E1 and E2**		**Residual 2 -regress out E1, E2&Sex**	
		Correlation **-Factor** *	Correlation **- PC**	Correlation **-Factor**	Correlation **-PC**	Correlation **-Factor**	Correlation **- PC**
High Environment Dataset	E1	0.83	0.76	~ 0	~ 0	~ 0	~ 0
	E2	<0.01	0.03	~ 0	~ 0	~ 0	~ 0
	G1	0.10	0.12	0.89	0.77	0.02	0.84
	G2	<0.01	<0.01	<0.01	0.03	0.29	0.05
	G3	<0.01	<0.01	0.03	0.08	<0.01	<0.01
	S1	<0.01	0.02	<0.01	0.05	0.51	0.04
Low Environment Dataset	E1	0.48	0.32	~ 0	~ 0	~ 0	~ 0
	E2	<0.01	0.04	~ 0	~ 0	~ 0	~ 0
	G1	0.40	0.50	<0.01	0.85	<0.01	0.88
	G2	<0.01	0.02	0.27	0.05	0.27	<0.01
	G3	<0.01	0.02	0.02	0.04	0.03	0.01
	S1	<0.01	0.03	0.70	<0.01	0.45	<0.01
Gene by Sex Interaction Dataset	E1	<0.01	<0.01	~ 0	~ 0	~ 0	~ 0
	E2	0.03	<0.01	~ 0	~ 0	~ 0	~ 0
	G1	<0.01	<0.01	<0.01	<0.01	<0.01	0.56
	G2	<0.01	<0.01	0.01	0.01	0.09	0.01
	S1	0.33	0.58	0.32	0.54	0.58	0.21
	S2	0.93	0.78	0.94	0.80	<0.01	<0.01

From tables above, we derive several conclusions.

First of all, generally speaking, both multivariate analysis methods (FA and PCA) give qualitatively similar results for analyses of all raw traits and most residual models from both sets of functions (Table 4 and 5). In other words, both methods show similar correlations with the underlying traits. However, when the trait models are more complicated (function set 2) and analyses are performed on residuals, these two methods appear to detect different underlying traits. For example, factor analysis was most highly correlated with underlying trait S1, whereas PCA was correlated with trait G1 in the analyses of the second function set, low environment dataset, and using residuals after regressing out E1, E2 or E1, E2 and Sex (Table 5).

Even in those cases in which both methods display qualitatively similar results, we think factor analysis demonstrates higher potency to detect predominant signals from underlying traits than PCA, by which it may benefit the downstream QTL analysis. We found here that when composite traits from both methods show significant correlations to a certain underlying trait, the correlation coefficient (R-Squares) between the first factor and that underlying trait is substantially higher than the corresponding correlations with the first principal component. For example, in the first set of functions, high environment dataset, and residuals after adjustment of E1 and E2 model: correlations between S1 and factor and S1 and PC are 0.88 and 0.68 respectively.

We also compared results of multivariate analyses performed using raw complex traits versus residuals of the complex traits. As can be seen (Table 4 and 5), PCA or FA analysis of residuals greatly improved detection of common genetic components instead of common environmental factors. For example, instead of picking up E1 for both high and low environment datasets when using raw traits from either function set 1 or 2, factors or PCs detected one of the underlying genetic components. Both PCA and FA obtained the highest correlation with underlying trait S1 for both datasets using residuals after regressing out E1 and E2. Furthermore the correlation between the environmental traits (E1 and E2) and the composite traits derived from the residuals is zero. As stated in the methods, we only regressed out the linear effects of E1 and E2 on the complex traits, even though E1 and E2 were not incorporated in the derivation of complex traits in only a linear fashion. Our limited results might suggest that performing a linear regression of environmental factors can be effective in removing some of the non-linear effects from environmental correlates. However, these results may be dependent on the specific set of non-linear functions we used and thus further evaluations are needed.

Finally, our results indicate that removing the effects of a covariate (i.e., sex, in our example) that has an interaction effect with the genotype on an underlying trait (i.e., trait S2), substantially decreases the potency of PCA or FA for detecting this underlying trait. See the second residual model (after adjustment of E1, E2 and sex) for both sets of functions in Tables 4 and 5.

4.2 ANALYSIS OF HERITABILITIES

We next compared the heritability of the underlying (unobserved) traits, the complex (observed) traits, and the first principal components and factors. Figure 2 shows the boxplot of heritabilities for composite traits compared with heritabilities for 50 complex traits. Table 6 lists mean heritabilties (or twice the sibling resemblance for non-genetic traits like E1 and E2) and the corresponding ranges for underlying traits. All mean heritabilities were calculated after taking the average of heritabilities from three replicates. And the range shows the variations of heritabilities among repeats. As indicated by Figure 2, there is no predictable relationship between heritability of composite traits and heritability of 50 complex traits. In other words, the heritabilities of composite traits are not necessarily higher or lower than original traits. This result is counterintuitive to our expectations, especially for residual models. We expected that the heritabilities of composite traits would be higher than those of the 50 complex traits, because multivariate analysis would incorporate co-variations for multiple traits due to shared genetic factors (pleiotropy), especially after removing environmental factors via regression analysis.

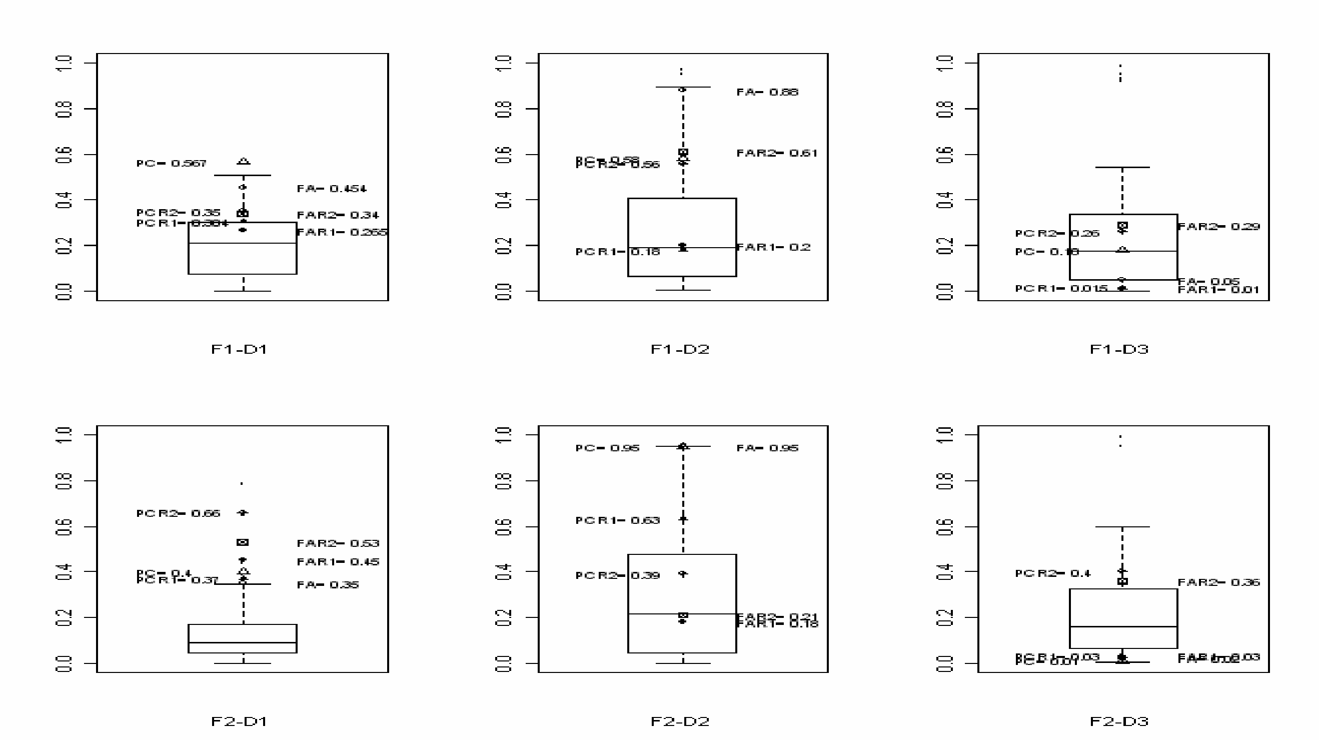
However, further comparisons of the heritabilities for composite traits and underlying phenotypes (table 6), indicates that FA and PCA did the right thing. The genetic/environmental information embedded in the composite trait reflects the genetic/environmental signals from underlying traits which had the highest correlations with the composite traits. For example, in function set 1, high environment dataset, using the raw trait model, the FA composite trait, seems exclusively coming from E1 (correlation =0.90) (Table 4). The heritability (or in this case, twice the sibling correlation) for this composite trait and heritability of E1 are comparable (0.454 vs 0.47). For the same function and dataset, but using the first residual model (adjusting for E1 and E2), the FA composite trait captured information mostly from S1. The corresponding heritabilities of FAR1 and S1 are also comparable (0.265 vs. 0.30).

Table 6 Mean heritability (sibling resemblance) estimation for underlying traits

	E1-1**	**E2-1**	**E1-2**	**E2-2**	**G1**	**G2**	**G3**	**S1**	**S2**
Heritability (H2)	**0.47**	**0.14**	**0.97**	**0.80**	**0.53**	**0.36**	**0.18**	**0.30**	**0.06**
H2 Range	**0.44-0.50**	**0.11-0.22**	**0.94-0.99**	**0.74-0.98**	**0.36-0.69**	**0.11-0.53**	**0.10-0.53**	**0.15-0.40**	**0.05-0.10**

*: numbers in the table indicates the mean heritabilities and its range for each underlying trait from all repeats;

†: E1-1/E2-1: heritability of E1/E2 in high environment dataset (SD=1); E1-2/E2-2: heritability of E1/E2 in low environment and gene by sex interaction dataset (SD=1/4)



* F1-D1/D2/D3: first set of functions, dataset 1, 2 or 3; F2-D1/D2/D3: second set of functions, dataset 1, 2 or 3.

PC/PCR1/PCR2: indicate heritabilities for composite phenotypes from raw trait, residual 1 or residual 2 model respectively;

FA/FA1/FA2: indicate heritabilities for composite phenotypes from raw trait, residual 1 or residual 2 model respectively;

Figure 2 Heritability Estimation of composite and 50 complex traits

5.0 DISCUSSION

There are several interesting and useful conclusions based on our study. The seven underlying traits that we simulated are representative of the unobserved environmental or/and genetic determinants which influence population variation of real traits. Likewise, the sets of 50 complex traits derived from these 7 original phenotypes reflect potentially real phenotypes that could be directly measured. Thus, any statistical analysis that can successfully identify variation attributable to any underlying original trait should theoretically have better power to detect genes when used in a genetic linkage or association analysis.

As indicated by our results, factor analysis seems to have better performance than PC analysis. This conclusion is based on the higher correlation between factors and the most significant underlying traits compared to that of PCs and the underlying traits. In the real world, higher correlations between composite trait and certain underlying phenotype (if it is due to genetics), should increase the probability of detecting and identifying the underlying genes. Hence, we would recommend factor analysis rather than principal component analysis. Another reason we prefer FA, although not shown in this thesis (we only consider 1 composite component), is that PCA assumes orthogonal relationship between its PCs, however FA does not. The assumption of independent extracted components may conflict with the true genetic model. For example, bone scientists hypothesized that genes influencing bone size may differ from genes influencing for bone mineral density (BMD). However these two sets of genes might interact with each other. If we put several bone size and BMD traits together into PCA, it is almost impossible to generate two independent PCs which represent a set of bone size genes and another set of BMD genes respectively.

Another conclusion from these analyses concerns the use of residuals versus raw trait values in multivariate analysis. Our results indicate that regressing out potentially significant environmental covariates should greatly increase the chances for detecting genetic component using both FA and PCA. However, there is an important caveat in the use of this strategy. If the underlying trait exhibits a genotype by environment interaction (see the results of our analyses with sex), removing the linear effects of such environmental covariates may decrease or even remove the genetic signal from the composite trait.

As shown in the results, the heritability for composite traits is not necessarily higher than that of the original complex traits. However, a high heritability does not necessarily predict successful detection of genes by linkage or association analysis. The success of detecting relevant genes depends not only on the number of loci influencing a trait, but also on the relative contribution of each locus, which is not reflected in the magnitude of the heritability estimate.(26) In addition, it is possible that different genotypes cause the same phenotype, an effect known as genetic heterogeneity. Heterogeneity complicates gene mapping and association and similarly is not reflected in the heritability estimate. In our analysis, we observed some examples of this phenomenon. For example, in first set of functions, low environment dataset, using residuals adjusting for E1 and E2, the heritability for both FA and PCA are very low (0.20 and 0.18 respectively). However, both composite traits grab most characteristics from underlying trait S1 (H2 = 0.18) and the heritability for both composite traits corresponds most closely with S1.

These composite traits should be more useful for downstream gene hunting than any of the individual complex traits.

Certain limitations of this study need to be acknowledged. These include 1) lack of further tools which could evaluate the results for PCA and FA when two composite traits were composed of different underlying trait as major components; 2) full consideration of sample size issues. We only simulated 250 families, or 500 sibs with phenotypic data. This sample size might be a little bit small to generate robust estimates for parameters. For example, the range of heritabilities for each of the 7 underlying traits is wide; and 3) Extension of our analysis to other PCs and factors.

In the future, we think follow-up linkage or association analyses might be a high priority in order to evaluate the success of PCA and FA in the final goal of the analyses – detecting disease genes. There are also a number of other extensions that are definitely worthwhile for us to explore: multigenerational family data for complex pedigrees; second or third PCs or factors; modified methodology conditioning on family relationship; using genetic correlation matrix instead of Pearson phenotypic correlation matrix as correlation/covariance matrix in multivariate analysis, etc..

APPENDIX A R –CODE FOR SIMULATION

```
# Function which generate family ID, indivudal ID, father ID and mother ID # and gender for
       all the members in one nuclear family #
# Function inputs: family #, usually comes from FOR loop [i]; outputs: family ID, indivudal
       ID, father ID, mother ID and sex #

sim_ID=function(j){

   fa_ped=j;mo_ped=j;child1_ped=j;child2_ped=j
   fa_id=j*10+1;mo_id=j*10+2;child1_id=j*10+3;child2_id=j*10+4
   fa_fa=0;mo_fa=0;child1_fa=fa_id;child2_fa=fa_id
   fa_mo=0;mo_mo=0;child1_mo=mo_id;child2_mo=mo_id

   fa_sex=1;mo_sex=2; # Use SOLAR format: male =1 and female =2 #
   child1_sex = sample(c(1,2),1,replace=T);child2_sex = sample(c(1,2),1,replace=T)


   fainfo=cbind(fa_ped,fa_id,fa_fa,fa_mo,fa_sex,deparse.level = 0)
   moinfo=cbind(mo_ped,mo_id,mo_fa,mo_mo,mo_sex,deparse.level = 0)
   child1info=cbind(child1_ped,child1_id,child1_fa,child1_mo,child1_sex,deparse.level = 0)
```

```
   child2info=cbind(child2_ped,child2_id,child2_fa,child2_mo,child2_sex,deparse.level = 0)

   ped_info=rbind(fainfo,moinfo,child1info,child2info,deparse.level = 0)

   return(ped_info)
}



# Function which generate the genotype and gender for nuclear family: 2 parents and 2 kids #
# Function inputs: Allele frequency for P(A) and P(a), outputs: genotype 0/1/2 for 2 parents
       and 2 kids #

sim_geno=function(p,q){




      ## Father’s genotype and gender ##
           ### variable names: fa and fa_sex ###

         fa.r <- rmultinom(1, size=1, prob=c(pˆ2,2*p*q,qˆ2))
              if (fa.r[1,1]==1) {fa <- 0} else if (fa.r[2,1]==1) {fa <- 1} else if
        (fa.r[3,1]==1) {fa <- 2}
         #print(fa.r)
         #print(fa)
```

```
  fa_sex <- 1 # Use SOLAR format: male =1 and female =2 #
  #print (fa_sex)

## Mother’ s genotype and gender ##
    ### variable names: mo and mo_sex ###

  mo.r <- rmultinom(1, size=1, prob=c(p^2,2*p*q,q^2))
       if  (mo.r[1,1]==1)  {mo  <-  0}  else  if  (mo.r[2,1]==1)  {mo  <-  1}  else  if
 (mo.r[3,1]==1) {mo <- 2}
  #print(mo.r)
  #print(mo)

  mo_sex <- 2 # Use SOLAR format: male =1 and female =2 #
  #print (mo_sex)

## Genotypes and gender for 2 kids##
    ### variable names: child2/2 and child2/2 ###



##### child1's genotype by Mendelian Rules  #####
  {if (fa==0 & mo==0) {child1 <- 0}
  else if ((fa==0 & mo==1)|(fa==1 & mo==0)) {child1 <- rbinom(1, size=1, prob=c(0.5))}
  else if ((fa==0 & mo==2)|(fa==2 & mo==0)) {child1 <- 1}
  else if (fa==1 & mo==1) {
          child.r <- rmultinom(1, size=1, prob=c(p^2,2*p*q,q^2))
```

```
         if (child.r[1,1]==1) {child1<-0}
         else if (child.r[2,1]==1) {child1<-1}
         else if (child.r[3,1]==1) {child1<-2}
         }
 else if ((fa==1 & mo==2)|(fa==2 & mo==1)) {child1 <- 1+rbinom(1, size=1,
prob=c(0.5))}
 else if (fa==2 & mo==2) {child1 <- 2}   }

 #print (child1)




##### child2's genotype by Mendelian Rules  #####
 {if (fa==0 & mo==0) {child2 <- 0}
 else if ((fa==0 & mo==1)|(fa==1 & mo==0)) {child2 <- rbinom(1, size=1, prob=c(0.5))}
 else if ((fa==0 & mo==2)|(fa==2 & mo==0)) {child2 <- 1}
 else if (fa==1 & mo==1) {
         child.r <- rmultinom(1, size=1, prob=c(p^2,2*p*q,q^2))
         if (child.r[1,1]==1) {child2<-0}
         else if (child.r[2,1]==1) {child2<-1}
         else if (child.r[3,1]==1) {child2<-2}
         }
 else if ((fa==1 & mo==2)|(fa==2 & mo==1)) {child2 <- 1+rbinom(1, size=1,
prob=c(0.5))}
```

```
        else if (fa==2 & mo==2) {child2 <- 2}   }

        #print (child2)



        return_geno = rbind(fa,mo,child1,child2,deparse.level = 0) #return all these values
        to the mail function: 4*2 matrix#
        #print (return_geno)

        return (return_geno)
}




# Function which generate the phenotypes ONLY for 2 CHILDREN in each nuclear family #
# Function inputs: type, genotype and gender for 2 children; outputs: phenotypes for these
       two children, phenotypes for parents will be generately as missing #
# There are all together 3 different types of phenotypes #
      #Type 1: environment controlled---no gender or genotype dependence#
      #Type 2: genetic controlled---genotype dependence#
      #Type 3: genetic and controlled with sex difference#
      #Type 4: traits controlled by gene X gender interaction#
      # r indicates the intended trait correlation between sibs. This is only useful for
       environment controlled model. set to 0 for other models #
```

```
sim_pheno=function(type,r,child1,child1_sex,child2,child2_sex){


      # These three are for gene controlled model traits withiout gender difference #
       mu.aa=1.5;sd.aa=1/2
       mu.Aa=2.5;sd.Aa=1/2
       mu.AA=3.5;sd.AA=1/2

      # These six are for gene controlled model traits with gender difference #
       mu.aa.fe=1;sd.aa.fe=1/2
       mu.Aa.fe=2;sd.Aa.fe=1/2
       mu.AA.fe=3;sd.AA.fe=1/2

       mu.aa.ma=2;sd.aa.ma=1/2
       mu.Aa.ma=3;sd.Aa.ma=1/2
       mu.AA.ma=4;sd.AA.ma=1/2

         # These six are for gene X Sex intereaction traits #
       mu.aa.fe.S2=1;sd.aa.fe.S2=1/2
       mu.Aa.fe.S2=1;sd.Aa.fe.S2=1/2
       mu.AA.fe.S2=1;sd.AA.fe.S2=1/2

       mu.aa.ma.S2=1.5;sd.aa.ma.S2=1/2
       mu.Aa.ma.S2=2.5;sd.Aa.ma.S2=1/2
```

```
     mu.AA.ma.S2=3.5;sd.AA.ma.S2=1/2


# environment controlled model #

   if (type==1){

       fa_pheno = NA
       mo_pheno = NA

                                 library(MASS)

                                 cormatrix = matrix(c(1/4,r,r,1/4),2,2)
                                 sibtrait = mvrnorm(n=1,mu=c(1,1),Sigma=cormatrix)
                                 child1_pheno = sibtrait[1]
       child2_pheno = sibtrait[2]
      }

# gene controlled model traits withiout gender difference #

    else if (type==2){


       fa_pheno = NA
       mo_pheno = NA
```

```
       {
        if (child1==0) {child1_pheno<-rnorm(1,mean=mu.aa,sd=sd.aa)}
        else if (child1==1) {child1_pheno<-rnorm(1,mean=mu.Aa,sd=sd.Aa)}
        else if (child1==2) {child1_pheno<-rnorm(1,mean=mu.AA,sd=sd.AA)}
       }

       {
        if (child2==0) {child2_pheno<-rnorm(1,mean=mu.aa,sd=sd.aa)}
        else if (child2==1) {child2_pheno<-rnorm(1,mean=mu.Aa,sd=sd.Aa)}
        else if (child2==2) {child2_pheno<-rnorm(1,mean=mu.AA,sd=sd.AA)}
       }

     }

# gene controlled model traits with gender difference #


   else if (type==3){


        fa_pheno = NA
        mo_pheno = NA
```

```
  {
   if (child1==0 & child1_sex==1) {child1_pheno<-rnorm(1,mean=mu.aa.ma,sd=sd.aa.ma)}
   else if (child1==0 & child1_sex==2) {child1_pheno<-
rnorm(1,mean=mu.aa.fe,sd=sd.aa.fe)}
   else if (child1==1 & child1_sex==1) {child1_pheno<-
rnorm(1,mean=mu.Aa.ma,sd=sd.Aa.ma)}
   else if (child1==1 & child1_sex==2) {child1_pheno<-
rnorm(1,mean=mu.Aa.fe,sd=sd.Aa.fe)}
   else if (child1==2 & child1_sex==1) {child1_pheno<-
rnorm(1,mean=mu.AA.ma,sd=sd.AA.ma)}
   else if (child1==2 & child1_sex==2) {child1_pheno<-
rnorm(1,mean=mu.AA.fe,sd=sd.AA.fe)}
  }


  {
   if (child2==0 & child2_sex==1) {child2_pheno<-rnorm(1,mean=mu.aa.ma,sd=sd.aa.ma)}
   else if (child2==0 & child2_sex==2) {child2_pheno<-
rnorm(1,mean=mu.aa.fe,sd=sd.aa.fe)}
   else if (child2==1 & child2_sex==1) {child2_pheno<-
rnorm(1,mean=mu.Aa.ma,sd=sd.Aa.ma)}
   else if (child2==1 & child2_sex==2) {child2_pheno<-
rnorm(1,mean=mu.Aa.fe,sd=sd.Aa.fe)}
```

```
  else if (child2==2 & child2_sex==1) {child2_pheno<-
 rnorm(1,mean=mu.AA.ma,sd=sd.AA.ma)}
  else if (child2==2 & child2_sex==2) {child2_pheno<-
 rnorm(1,mean=mu.AA.fe,sd=sd.AA.fe)}
 }

}

     else if (type==4){


  fa_pheno = NA
  mo_pheno = NA


 {
  if (child1==0 & child1_sex==1) {child1_pheno<-
 rnorm(1,mean=mu.aa.ma.S2,sd=sd.aa.ma.S2)}
  else if (child1==0 & child1_sex==2) {child1_pheno<-
 rnorm(1,mean=mu.aa.fe.S2,sd=sd.aa.fe.S2)}
  else if (child1==1 & child1_sex==1) {child1_pheno<-
 rnorm(1,mean=mu.Aa.ma.S2,sd=sd.Aa.ma.S2)}
  else if (child1==1 & child1_sex==2) {child1_pheno<-
 rnorm(1,mean=mu.Aa.fe.S2,sd=sd.Aa.fe.S2)}
```

```
   else if (child1==2 & child1_sex==1) {child1_pheno<-
  rnorm(1,mean=mu.AA.ma.S2,sd=sd.AA.ma.S2)}
   else if (child1==2 & child1_sex==2) {child1_pheno<-
  rnorm(1,mean=mu.AA.fe.S2,sd=sd.AA.fe.S2)}
  }



  {
   if (child2==0 & child2_sex==1) {child2_pheno<-
  rnorm(1,mean=mu.aa.ma.S2,sd=sd.aa.ma.S2)}
   else if (child2==0 & child2_sex==2) {child2_pheno<-
  rnorm(1,mean=mu.aa.fe.S2,sd=sd.aa.fe.S2)}
   else if (child2==1 & child2_sex==1) {child2_pheno<-
  rnorm(1,mean=mu.Aa.ma.S2,sd=sd.Aa.ma.S2)}
   else if (child2==1 & child2_sex==2) {child2_pheno<-
  rnorm(1,mean=mu.Aa.fe.S2,sd=sd.Aa.fe.S2)}
   else if (child2==2 & child2_sex==1) {child2_pheno<-
  rnorm(1,mean=mu.AA.ma.S2,sd=sd.AA.ma.S2)}
   else if (child2==2 & child2_sex==2) {child2_pheno<-
  rnorm(1,mean=mu.AA.fe.S2,sd=sd.AA.fe.S2)}
  }


}
```

```
     return_pheno  =  rbind(fa_pheno,mo_pheno,child1_pheno,child2_pheno,deparse.level  =  0)
       #return all these values to the mail function#
     #print (return_pheno)


     return (return_pheno)
}




# Function: Heritability estimation. For full sibs, use the formula H2=2*cor(sibs)#
       H2 = function(sib1_trait,sib2_trait){


              sibcor = cor(sib1_trait,sib2_trait,use="complete.obs")
              heritability = 2*sibcor


              return (heritability)
             }




#  Function:  generate  a  data-matrix  which  contains  the  pedigree  information,  genotype
       information and original phenotype information for 250 families (1000 individuals) #
  #For the 6 original phenotypes: we will simulate 2 environment controlled traits, 3 gene
       controlled traits and 1 gene X gender controlled traits #
  #input: None, output: Data matrix for ped, geno, pheno information #


rawtraits = function(){
```

```
sim_data=matrix(data = c("Family ID","Individual ID","Fa_ID","Mo_ID","Sex",
        "GenoE1","PhenoE1","GenoE2","PhenoE2","GenoG1","PhenoG1","GenoG2","PhenoG2","GenoS1","
        PhenoS1","GenoS2","PhenoS2"),nrow = 1, ncol = 17)

for(i in 1:250){

  pedinfo = sim_ID(i)

  # Generate 6 traits (2E,2G and 2S), each controlled by its own allele respectively #
  genoinfoE1 = sim_geno(0.2,0.8)
  E1pheno = sim_pheno(1,0.20,genoinfoE1[3,1],pedinfo[3,5],genoinfoE1[4,1],pedinfo[4,5])

  genoinfoE2 = sim_geno(0.5,0.5)
  E2pheno = sim_pheno(1,0.10,genoinfoE2[3,1],pedinfo[3,5],genoinfoE2[4,1],pedinfo[4,5])

  genoinfoG1 = sim_geno(0.2,0.8)
  G1pheno = sim_pheno(2,0,genoinfoG1[3,1],pedinfo[3,5],genoinfoG1[4,1],pedinfo[4,5])

  genoinfoG2 = sim_geno(0.1,0.9)
  G2pheno = sim_pheno(2,0,genoinfoG2[3,1],pedinfo[3,5],genoinfoG2[4,1],pedinfo[4,5])

  genoinfoS1 = sim_geno(0.2,0.8)
  S1pheno = sim_pheno(3,0,genoinfoS1[3,1],pedinfo[3,5],genoinfoS1[4,1],pedinfo[4,5])
```

```
      genoinfoS2 = sim_geno(0.3,0.7)
  S2pheno = sim_pheno(4,0,genoinfoS2[3,1],pedinfo[3,5],genoinfoS2[4,1],pedinfo[4,5])




  all_data                                                                                          =
      cbind(pedinfo,genoinfoE1,E1pheno,genoinfoE2,E2pheno,genoinfoG1,G1pheno,genoinfoG2,G2ph
      eno,genoinfoS1,S1pheno,genoinfoS2,S2pheno,deparse.level = 0)
  # print (all_data)


  sim_data=rbind(sim_data,all_data,deparse.level = 0)

}

  write.table(sim_data,file="c:/simdata.csv",sep=",",row.names=F,na="",quote=F,col.names=F)

  sim.data=read.csv("c:/simdata.csv",sep=",",header=T)


# Add one variable to the dataset, which indicates the sib group: all sib1 ==1 and all sib2
      ==2,parents ==NA #

  Nfmly <- length(unique(sim.data$Family.ID))
```

```
  middata <- rep(c(1,2,2,2),Nfmly)
  sim.data$childgrp <- sim.data$Individual.ID - (sim.data$Family.ID*10+middata)

      # names(sim.data)

  write.table(sim.data,file="c:/simdata.csv",sep=",",row.names=F,na="",quote=F)

  sim.data=read.csv("c:/simdata.csv",sep=",",header=T)


  # Calculate the actual allele frequency #

  for(j in c(6,8,10,12,14,16)){

    temp = table(sim.data[,j])
    cat(names(sim.data)[j],"\n\n","                    aa                                     Aa
      AA\n",temp,"\n\n",file="c:/allelecheck.txt",append=T)

    temp_matrix=as.matrix(temp)

      Pa=(temp_matrix[1,1]*2+temp_matrix[2,1])/(2*(temp_matrix[1,1]+temp_matrix[2,1]+temp_ma
      trix[3,1]))
    PA=1-Pa
    cat("P(a)=",Pa,"\n","P(A)=",PA,"\n\n\n",file="c:/allelecheck.txt",append=T)
}
```

```
# Calculate the mean/sd and heritability for each trait (by gender and genotype if necessary) #


#cat("\n\n\nDistribution Parameter\n\n",file="c:/allelecheck.txt",append=T)  #
#for ( k in c(7,9,11,13,15,17)){

#male_data = sim.data[sim.data$Sex==1,];female_data = sim.data[sim.data$Sex==2,]
#aa_male_data = male_data[male_data[,k-1]==2,];Aa_male_data = male_data[male_data[,k-1]==1,];AA_male_data = male_data[male_data[,k-1]==0,]
#aa_female_data = female_data[female_data[,k-1]==2,];Aa_female_data = female_data[female_data[,k-1]==1,];AA_female_data = female_data[female_data[,k-1]==0,]


#all_mean = mean(sim.data[,k],na.rm=T);all_sd=sd(sim.data[,k],na.rm=T)

#male_mean = mean(male_data[,k],na.rm=T);male_sd=sd(male_data[,k],na.rm=T)
#female_mean = mean(female_data[,k],na.rm=T);female_sd=sd(female_data[,k],na.rm=T)

#aa_male_mean = mean(aa_male_data[,k],na.rm=T);aa_male_sd=sd(aa_male_data[,k],na.rm=T)
#Aa_male_mean = mean(Aa_male_data[,k],na.rm=T);Aa_male_sd=sd(Aa_male_data[,k],na.rm=T)
#AA_male_mean = mean(AA_male_data[,k],na.rm=T);AA_male_sd=sd(AA_male_data[,k],na.rm=T)
```

```
#aa_female_mean                                                                  =
   mean(aa_female_data[,k],na.rm=T);aa_female_sd=sd(aa_female_data[,k],na.rm=T)
#Aa_female_mean                                                                  =
   mean(Aa_female_data[,k],na.rm=T);Aa_female_sd=sd(Aa_female_data[,k],na.rm=T)
#AA_female_mean                                                                  =
   mean(AA_female_data[,k],na.rm=T);AA_female_sd=sd(AA_female_data[,k],na.rm=T)

#cat("\n\n",names(sim.data)[k],"\n","all_mean=",all_mean,"
   all_sd=",all_sd,"\n\n",file="c:/allelecheck.txt",append=T)

#cat("male_mean=",male_mean,"  male_sd=",male_sd,"\n",file="c:/allelecheck.txt",append=T)
#cat("female_mean=",female_mean,"
   female_sd=",female_sd,"\n\n",file="c:/allelecheck.txt",append=T)

#cat("aa_male_mean=",aa_male_mean,"
   aa_male_sd=",aa_male_sd,"\n",file="c:/allelecheck.txt",append=T)
#cat("Aa_male_mean=",Aa_male_mean,"
   Aa_male_sd=",Aa_male_sd,"\n",file="c:/allelecheck.txt",append=T)
#cat("AA_male_mean=",AA_male_mean,"
   AA_male_sd=",AA_male_sd,"\n",file="c:/allelecheck.txt",append=T)
#cat("aa_female_mean=",aa_female_mean,"
   aa_female_sd=",aa_female_sd,"\n",file="c:/allelecheck.txt",append=T)
#cat("Aa_female_mean=",Aa_female_mean,"
   Aa_female_sd=",Aa_female_sd,"\n",file="c:/allelecheck.txt",append=T)
```

```
    #cat("AA_female_mean=",AA_female_mean,"
       AA_female_sd=",AA_female_sd,"\n",file="c:/allelecheck.txt",append=T)


   #}

        return(sim.data)

}

#Function: make 50 derived traits using the original 6 phenotypes from the rawtraits
       function#
 # derived traits will be applying transformations on original phenotypes plus error term#
       # error term is a normal distribution with mean 1 and variance 1 #
       # input: none. Will call rawtraits inside the function; output: dataset with ped info
       and 50 derived traits#

alltraits = function(){

              temptraits = rawtraits()

              # make the error matrix, NA for both parents and obs. from normal distribution
       (1,0.25) for two kids#
              Nfmly <- length(unique(temptraits$Family.ID))
              errormatrix = matrix(rnorm(Nfmly*4*50,1,1),nrow=Nfmly*4,ncol=50)
```

```
        count = 0
        for (w in 1:(Nfmly*4)){
                count = count +1
                if (count == 1|count==2){errormatrix[w,]=NA}
                if (count == 4){count = 0}
        }

        # make 50 derived traits#

                    temptraits$C1 = log(temptraits$PhenoE1 +5) * temptraits$PhenoE2 -
(temptraits$PhenoS2 / temptraits$PhenoS1 +2)^2 + errormatrix[,1]
         temptraits$C2 = 4.4*temptraits$PhenoG1 / sqrt (temptraits$PhenoG2 -
temptraits$PhenoE1+3) + errormatrix[,2]
         temptraits$C3 = (temptraits$PhenoG1 + temptraits$PhenoE2 +
temptraits$PhenoS1)/3+errormatrix[,3]
         temptraits$C4 = (temptraits$PhenoE2) ^2 + (temptraits$PhenoS2 +1) /
(temptraits$PhenoG2 + 1) + errormatrix[,4]
         temptraits$C5 = (1.2*temptraits$PhenoE1 + temptraits$PhenoG2 -3) /
(temptraits$PhenoS1^2 +sqrt (temptraits$PhenoG1+2) -2) + errormatrix[,5]
         temptraits$C6 = (1.1*temptraits$PhenoG1 + 2.7*temptraits$PhenoG2 +
(3/5)*temptraits$PhenoS1+temptraits$PhenoE1-
1.4*temptraits$PhenoE2+2*temptraits$PhenoS2)/7+errormatrix[,6]
         temptraits$C7 = log (temptraits$PhenoE2 + errormatrix[,39] + 1.2) - exp
(temptraits$PhenoG1 + errormatrix[,7])
```

```
        temptraits$C8 = log (temptraits$PhenoG1 * temptraits$PhenoG2 +2) - sqrt(temptraits$PhenoS2+2) + errormatrix[,8]

        temptraits$C9 = 2*temptraits$PhenoG2 + (temptraits$PhenoS2 + 2 * temptraits$PhenoE2 + 1) / (temptraits$PhenoE2ˆ2 +sqrt (temptraits$PhenoG2+2) +1) + errormatrix[,9]

        temptraits$C10 = 2.3 * temptraits$PhenoG2 + temptraits$PhenoE1 + 2*temptraits$PhenoS1 + errormatrix[,10]

        temptraits$C11 = 0.2*temptraits$PhenoE2 + 1.4*temptraits$PhenoS1 + temptraits$PhenoS2 + errormatrix[,11]

        temptraits$C12 = (temptraits$PhenoG1 / (temptraits$PhenoS2-3) ) / (temptraits$PhenoE2 ˆ (2/3) + 1) + errormatrix[,12]

        temptraits$C13 = (temptraits$PhenoE1ˆ2 -3) * log (temptraits$PhenoG1+3) + errormatrix[,13]

        temptraits$C14 = (temptraits$PhenoG2 +(1.2* temptraits$PhenoG2 -2)ˆ3) / (temptraits$PhenoS2 + 1) + errormatrix[,14]

        temptraits$C15 = sqrt(temptraits$PhenoG2 + errormatrix[,15] + 2) + (temptraits$PhenoG1) ˆ2 * temptraits$PhenoE1

        temptraits$C16 = (temptraits$PhenoS2 + 1) / (temptraits$PhenoS1 + 1) - temptraits$PhenoE1 + errormatrix[,16]

        temptraits$C17 = temptraits$PhenoS2 * temptraits$PhenoE2 + 3.3*exp (temptraits$PhenoS1 + errormatrix[,17])

        temptraits$C18 = sqrt (temptraits$PhenoG2+2) - temptraits$PhenoG1 * temptraits$PhenoE1  + errormatrix[,18]

        temptraits$C19 = (temptraits$PhenoS2 + 2.2 * temptraits$PhenoS1) / (1.7 * temptraits$PhenoG2) + errormatrix[,19]
```

```
        temptraits$C20 = temptraits$PhenoS2 * temptraits$PhenoG2 - 2 / (temptraits$PhenoG1+4) ^3 - 0.7*(temptraits$PhenoS1) / 3 + errormatrix[,20]

        temptraits$C21 = temptraits$PhenoS2 * temptraits$PhenoG2 * temptraits$PhenoE2 + errormatrix[,21]

        temptraits$C22 = temptraits$PhenoG1/ (2.2- temptraits$PhenoS1 )* temptraits$PhenoE2 * temptraits$PhenoS2+errormatrix[,22]

        temptraits$C23 = (temptraits$PhenoE1 + 1) / (temptraits$PhenoG1 * temptraits$PhenoS1 + 1) + errormatrix[,23]

        temptraits$C24 = (2.2*temptraits$PhenoG2 - temptraits$PhenoS2) / (temptraits$PhenoE2 * temptraits$PhenoE1 + 1) + errormatrix[,24]

        temptraits$C25 = (sqrt (temptraits$PhenoE1 + temptraits$PhenoG2 ^2 + temptraits$PhenoS1 ^3 +1) +4) / (temptraits$PhenoS2 - temptraits$PhenoE2 ^3 +0.5*temptraits$PhenoG1*temptraits$PhenoG2 + 1) + errormatrix[,25]

        temptraits$C26 = (temptraits$PhenoS2 + 3)/(temptraits$PhenoS1) - 2.1 * temptraits$PhenoE2 + errormatrix[,26]

        temptraits$C27 = (temptraits$PhenoE2 + 2.1) / (temptraits$PhenoE1 + 1.2) + temptraits$PhenoS1 + errormatrix[,27]

        temptraits$C28 = (temptraits$PhenoS1-1) * temptraits$PhenoG2 + 3.2*temptraits$PhenoE1- temptraits$PhenoE1+ errormatrix[,28]

        temptraits$C29 = temptraits$PhenoE2 * temptraits$PhenoE1 + temptraits$PhenoG1 + errormatrix[,29]

        temptraits$C30 = (temptraits$PhenoS2 + 1) / (temptraits$PhenoG2 + 1) - temptraits$PhenoG1+errormatrix[,30]

        temptraits$C31 = temptraits$PhenoE1 * temptraits$PhenoG1 + 0.5 * temptraits$PhenoE2 +errormatrix[,31]
```

```
        temptraits$C32 = 3*(temptraits$PhenoS1 +1) / (temptraits$PhenoS2 + 1.7) -
2*temptraits$PhenoE1 + errormatrix[,32]
        temptraits$C33 = temptraits$PhenoG1 * temptraits$PhenoE2 -
(temptraits$PhenoS2 +1) / (temptraits$PhenoG2 + 1) + errormatrix[,33]
        temptraits$C34 = (temptraits$PhenoS2 +1) / (temptraits$PhenoG2 + 2) +
(temptraits$PhenoE2 +1) / (temptraits$PhenoS1 + 1) - 0.7*temptraits$PhenoG1 +
errormatrix[,34]
        temptraits$C35 = (temptraits$PhenoG1 + errormatrix[,35]) ˆ2 +
temptraits$PhenoG1 / (temptraits$PhenoE1* temptraits$PhenoS2)
        temptraits$C36 = (temptraits$PhenoS2 + errormatrix[,36]) ˆ3 - 2.4*
log(temptraits$PhenoE2 + errormatrix[,36] + 2)
        temptraits$C37 = 3.3*exp (temptraits$PhenoS1 + errormatrix[,37]) +
1.4*(temptraits$PhenoE1 + errormatrix[,37]) ˆ2
        temptraits$C38 = sqrt (temptraits$PhenoS2 + errormatrix[,38] + 2) +
(temptraits$PhenoG1) ˆ2 * temptraits$PhenoE2
        temptraits$C39 = log (temptraits$PhenoG1 + errormatrix[,39] + 2) - exp
(temptraits$PhenoE1 + errormatrix[,39])
        temptraits$C40 = (temptraits$PhenoS1) ˆ2 + (temptraits$PhenoG2 +1) /
(temptraits$PhenoE1 + 1) + errormatrix[,40]
        temptraits$C41 = 2 / (temptraits$PhenoE1+5) ˆ3 - 0.7*(temptraits$PhenoS1) /
+ errormatrix[,41]
        temptraits$C42 = sqrt (temptraits$PhenoE2+2) - temptraits$PhenoS2 *
temptraits$PhenoG1 + errormatrix[,42]
        temptraits$C43 = log (temptraits$PhenoG2 * temptraits$PhenoE1 +2) -
sqrt(temptraits$PhenoS1+2) + errormatrix[,43]
```

```
        temptraits$C44 = (temptraits$PhenoG1 + temptraits$PhenoE2 + 1) / (temptraits$PhenoS1 ^2 + 1)+ errormatrix[,44]
        temptraits$C45 = (temptraits$PhenoG1 + 1) / (temptraits$PhenoS2 ^ (2/3) + 1) + errormatrix[,45]
        temptraits$C46 = (temptraits$PhenoG1 - temptraits$PhenoE2)* temptraits$PhenoS1 + log(temptraits$PhenoE1 +2) + errormatrix[,46]
        temptraits$C47 = (temptraits$PhenoE1) ^2 + (temptraits$PhenoG2 + temptraits$PhenoS1) * (2 * temptraits$PhenoS2 - temptraits$PhenoE2) + errormatrix[,47]
        temptraits$C48 = (temptraits$PhenoS2 + 2 * temptraits$PhenoG2 + 1) / (temptraits$PhenoE1^2 +sqrt (temptraits$PhenoS1+2) +1) + errormatrix[,48]
        temptraits$C49 = 4.2 / (log (2*temptraits$PhenoS2 + 2*temptraits$PhenoS1 - temptraits$PhenoE1 + 2) -3) + errormatrix[,49]
        temptraits$C50 = (sqrt (temptraits$PhenoE2 + temptraits$PhenoE1 ^2 + temptraits$PhenoS1 ^3 +2) +1) / (temptraits$PhenoG1 - temptraits$PhenoE1 ^3 +0.5*temptraits$PhenoG1*temptraits$PhenoG2 + 1) + errormatrix[,50]

#output the final dataset#

write.table(temptraits,file="c:/alltraits.csv",sep=",",quote=F,row.names=F,na="")
temptraits = read.csv("c:/alltraits.csv",sep=",",header=T)

    #Plot the histograms of all 56 simulated traits and output it into PDF file #

  pdf(file="c:/alltraits.pdf",title="Histogram for all simulated traits" )
  par(mfrow=c(3,2))
```

```
for ( p in c(7,9,11,13,15,17,19:68)){

   hist(temptraits[,p], main = paste("Histogram of" , names(temptraits)[p]))



              # Calculate H2 for each trait #

              sib1data = temptraits[temptraits$childgrp==1,]

              sib2data = temptraits[temptraits$childgrp==2,]



              h2 = H2(sib1data[,p],sib2data[,p])

              cat("\n",names(temptraits)[p],"                        Narrow            Sense
     Heritability=",h2,"\n",file="c:/H2check.txt",append=T)



 }



dev.off()



     }
```

APPENDIX B R-CODE FOR MULTIVARIATE ANALYSIS

```
# option 1: PCA/FA analysis without regression #

alltraits=read.csv("c:/alltraits.csv",sep=",",header=T)
clean=alltraits[,c(2,5,7,9,11,13,15,17,18,19:68)]
temp<-clean[is.na(clean$C1)==F&is.na(clean$C2)==F&is.na(clean$C3)==F&is.na(clean$C4)==F&is.na(clean$C5)==F&is.na(clean$C6)==F&is.na(clean$C7)==F&is.na(clean$C8)==F&is.na(clean$C9)==F&is.na(clean$C10)==F&is.na(clean$C11)==F&is.na(clean$C12)==F&is.na(clean$C13)==F&is.na(clean$C14)==F&is.na(clean$C15)==F&is.na(clean$C16)==F&is.na(clean$C17)==F&is.na(clean$C18)==F&is.na(clean$C19)==F&is.na(clean$C20)==F&is.na(clean$C21)==F&is.na(clean$C22)==F&is.na(clean$C23)==F&is.na(clean$C24)==F&is.na(clean$C25)==F&is.na(clean$C26)==F&is.na(clean$C27)==F&is.na(clean$C28)==F&is.na(clean$C29)==F&is.na(clean$C30)==F&is.na(clean$C31)==F&is.na(clean$C32)==F&is.na(clean$C33)==F&is.na(clean$C34)==F&is.na(clean$C35)==F&is.na(clean$C36)==F&is.na(clean$C37)==F&is.na(clean$C38)==F&is.na(clean$C39)==F&is.na(clean$C40)==F&is.na(clean$C41)==F&is.na(clean$C42)==F&is.na(clean$C43)==F&is.na(clean$C44)==F&is.na(clean$C45)==F&is.na(clean$C46)==F&is.na(clean$C47)==F&is.na(clean$C48)==F&is.na(clean$C49)==F&is.na(clean$C50)==F,]

cor.ma=cov(temp[,c(10:59)],use="pairwise.complete.obs")

# Factor Analysis with orignial traits #

FA=factanal(temp[,c(10:59)],1,covlist=cor.ma,scores="regression")


cat("**********************\n","Factor Analysis\n","**********************","\n\n","loadings","\n",FA$loadings,"\n\n\n","uniquenesses\n",FA$uniquenesses,"\n",file="c:/FA-PCA.txt",append=T)


temp$FA1=FA$scores
write.table(temp,file="c:/FAdata.csv",sep=",",row.names=F,quote=F,na="")
FAdata=read.csv("c:/FAdata.csv",sep=",",header=T)

for (i in 3:8){

reg4 = summary(lm(FAdata$FA1~FAdata[,i]))
```

```
cat("\n\nREGRESS FA1 by E1/E2/G1/G2/G3/S1","\n\n",names(FAdata)[i],"    ","R-Square =
",reg4$r.sq,"\n\n\n",file="c:/FA-PCA.txt",append=T)

}


# PC Analysis with orignial traits #

PCA=princomp(temp[,c(10:59)],cor=T,scores=T)


cat("*********************\n","PC
Analysis\n","*********************","\n\n","loadings","\n",PCA$loadings[,1],"\n\n",file="c:/
FA-PCA.txt",append=T)

temp$PC1=PCA$scores[,1]

write.table(temp,file="c:/FAdata.csv",sep=",",row.names=F,quote=F,na="")

FAdata=read.csv("c:/FAdata.csv",sep=",",header=T)

for (i in 3:8){

reg3 = summary(lm(FAdata$PC1~FAdata[,i]))

cat("\n\nREGRESS PC1 by E1/E2/G1/G2/G3/S1","\n\n",names(FAdata)[i],"    ","R-Square =
",reg3$r.sq,"\n\n\n",file="c:/FA-PCA.txt",append=T)

}



# option 2: PCA/FA analysis with regression of E1 and E2#


pcatry=FAdata
rownum = dim(pcatry)[1]
resmatrix = matrix(rep(NA,rownum*50),ncol=50)

for (i in 10:59){

reg = lm(pcatry[,i]~pcatry$PhenoE1+pcatry$PhenoE2)

resmatrix[,i-9] = reg$res

}

colname=matrix(data =
c("res1","res2","res3","res4","res5","res6","res7","res8","res9","res10","res11","res12","res
```

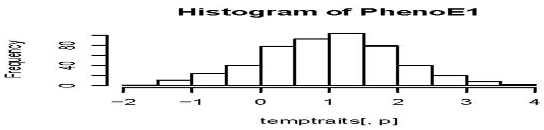
```
13","res14","res15","res16","res17","res18","res19","res20","res21","res22","res23","res24","
res25","res26","res27","res28","res29","res30","res31","res32","res33","res34","res35","res36
","res37","res38","res39","res40","res41","res42","res43","res44","res45","res46","res47","re
s48","res49","res50"),nrow = 1, ncol = 50)

res_data=rbind(colname,resmatrix,deparse.level=0)

write.table(res_data,file="c:/resdata-temp.csv",sep=",",quote=F,row.names=F,na="",col.names=F)

res_data = read.csv("c:/resdata-temp.csv",sep=",",header=T)

temp=cbind(pcatry,res_data,deparse.level=0)  # temp data which contains ID, E1/E2, all 50 C
traits and 50 residuals#


write.table(temp,file="c:/FAdata.csv",sep=",",quote=F,na="",row.names=F)

temp = read.csv("c:/FAdata.csv",sep=",",header=T)


cor.ma=cov(temp[,c(62:111)],use="pairwise.complete.obs")

# Factor Analysis using residuals #

FA=factanal(temp[,c(62:111)],1,covlist=cor.ma,scores="regression")

cat("\n\n***************************\n","Residual Factor
Analysis\n","***************************","\n\n","loadings","\n",FA$loadings,"\n\n\n","unique
nesses\n",FA$uniquenesses,"\n",file="c:/FA-PCA.txt",append=T)

temp$FAR1=FA$scores

write.table(temp,file="c:/FAdata.csv",sep=",",row.names=F,quote=F,na="")

FAdata=read.csv("c:/FAdata.csv",sep=",",header=T)

for (i in 5:8){

reg2 = summary(lm(FAdata$FAR1~FAdata[,i]))

cat("\n\nREGRESS FA1 by G1/G2/G3/S1","\n\n",names(FAdata)[i],"    ","R-Square =
",reg2$r.sq,"\n\n\n",file="c:/FA-PCA.txt",append=T)

}


# PC Analysis using residuals #

PCA=princomp(temp[,c(62:111)],cor=T,scores=T)
```

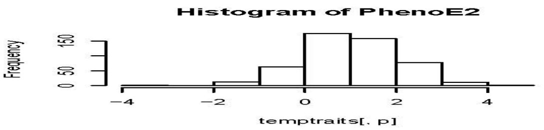
```
cat("********************************\n","Residual PC
Analysis\n","********************************","\n\n","loadings","\n",PCA$loadings[,1],"\n\n"
,file="c:/FA-PCA.txt",append=T)


temp$PCR1=PCA$scores[,1]

write.table(temp,file="c:/FAdata.csv",sep=",",row.names=F,quote=F,na="")

FAdata=read.csv("c:/FAdata.csv",sep=",",header=T)

for (i in 5:8){

reg3 = summary(lm(FAdata$PCR1~FAdata[,i]))

cat("\n\nREGRESS PC1 by G1/G2/G3/S1","\n\n",names(FAdata)[i],"   ","R-Square =
",reg3$r.sq,"\n\n\n",file="c:/FA-PCA.txt",append=T)

}


# Calculate the correlations between FA1/PC1/FAR1/PCR1 and Sex #

cor1=cor(FAdata$Sex,FAdata$FA1,method="spearman")
cor2=cor(FAdata$Sex,FAdata$PC1,method="spearman")
cor3=cor(FAdata$Sex,FAdata$FAR1,method="spearman")
cor4=cor(FAdata$Sex,FAdata$PCR1,method="spearman")


cat("********************************\n","Correlation
Calculation\n","********************************","\n\n",file="c:/FA-PCA.txt",append=T)
cat("Correlation between Sex and FA1 is  ",cor1,"\n",file="c:/FA-PCA.txt",append=T)
cat("Correlation between Sex and PC1 is  ",cor2,"\n",file="c:/FA-PCA.txt",append=T)
cat("Correlation between Sex and FAR1 is  ",cor3,"\n",file="c:/FA-PCA.txt",append=T)
cat("Correlation between Sex and PCR1 is  ",cor4,"\n",file="c:/FA-PCA.txt",append=T)



# remove another corresponding sibs if one sib is missing in the dataset FAdata #
# This step is required by H2 function #

IDmatrix=alltraits[,c(2,19)]
IDmatrix=IDmatrix[is.na(IDmatrix$C1)==F,]
IDmatrix$idindex <- seq(1,500,1)
H2data=merge(IDmatrix,FAdata,by.x="Individual.ID",by.y="Individual.ID",all.x=T,all.y=T)
missingID=H2data$Individual.ID[is.na(H2data$PC1)==T]
misslen=length(missingID)
missingindex <- H2data$idindex[is.na(H2data$PC1)==T]
```

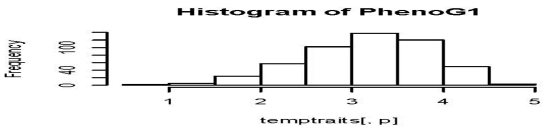
```
IDremainder <- missingID%%2
IDremainder[IDremainder==0]<- -1
missingindex2 <- missingindex+IDremainder
missingrow <- c(missingindex,missingindex2)
nomissing <- H2data[-missingrow,]

# Calculate the Heritabilities for FA1/PC1/FAR1/PCR1 #

              if(misslen!=0){
               sib1data = nomissing[nomissing$childgrp==1,]
               sib2data = nomissing[nomissing$childgrp==2,]
                }

               if (misslen == 0){
               sib1data = FAdata[FAdata$childgrp==1,]
               sib2data = FAdata[FAdata$childgrp==2,]
                }

                 h2FA1 = H2(sib1data$FA1,sib2data$FA1)
                 h2PC1 = H2(sib1data$PC1,sib2data$PC1)
                 h2FAR1 = H2(sib1data$FAR1,sib2data$FAR1)
                 h2PCR1 = H2(sib1data$PCR1,sib2data$PCR1)

                 cat("\n","FA1 Narrow Sense
Heritability=",h2FA1,"\n",file="c:/H2check.txt",append=T)
                 cat("\n","PC1 Narrow Sense
Heritability=",h2PC1,"\n",file="c:/H2check.txt",append=T)
                 cat("\n","FAR1 Narrow Sense
Heritability=",h2FAR1,"\n",file="c:/H2check.txt",append=T)
                 cat("\n","PCR1 Narrow Sense
Heritability=",h2PCR1,"\n",file="c:/H2check.txt",append=T)
```

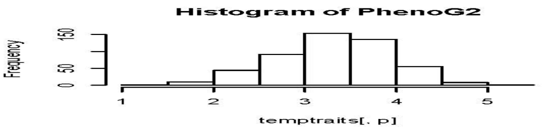
APPENDIX C DISTRIBUTIONS OF COMPLEX TRAITS

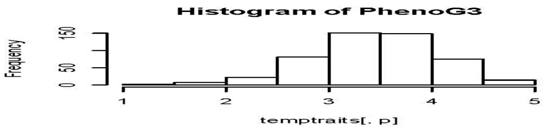
1st set of 50 functions and 6 underlying traits

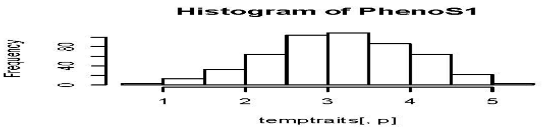


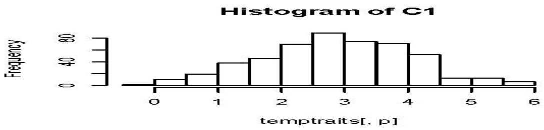


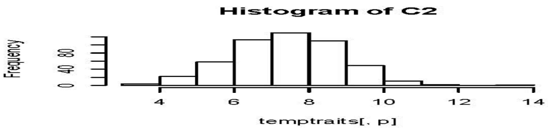


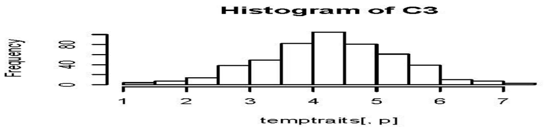


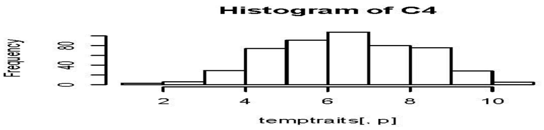


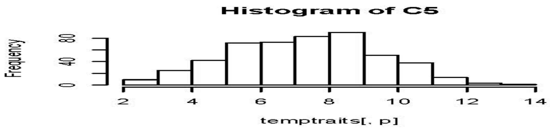


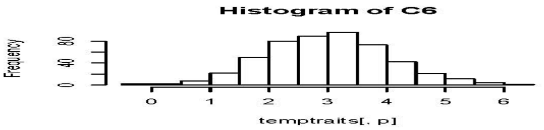


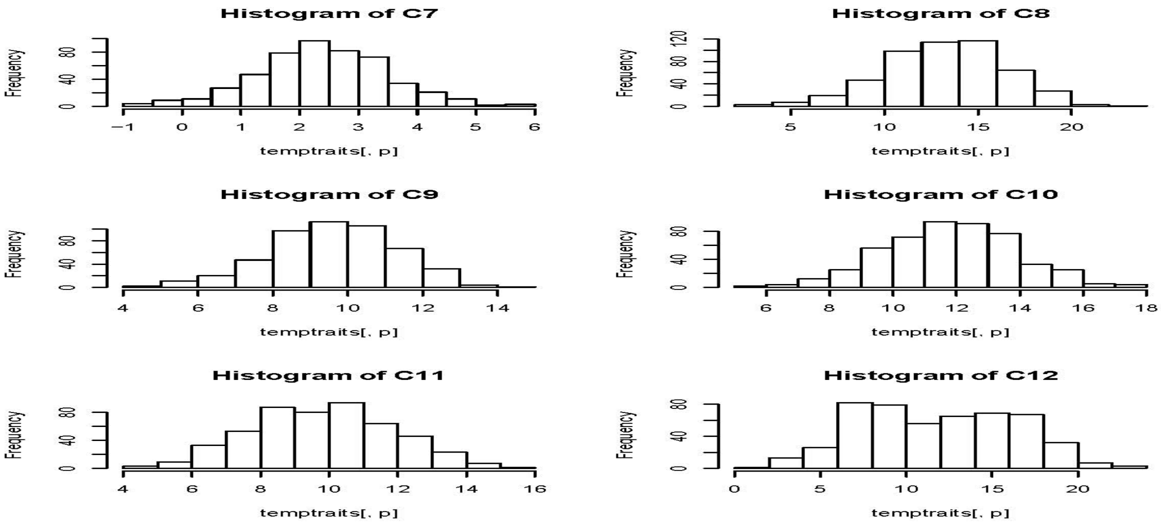


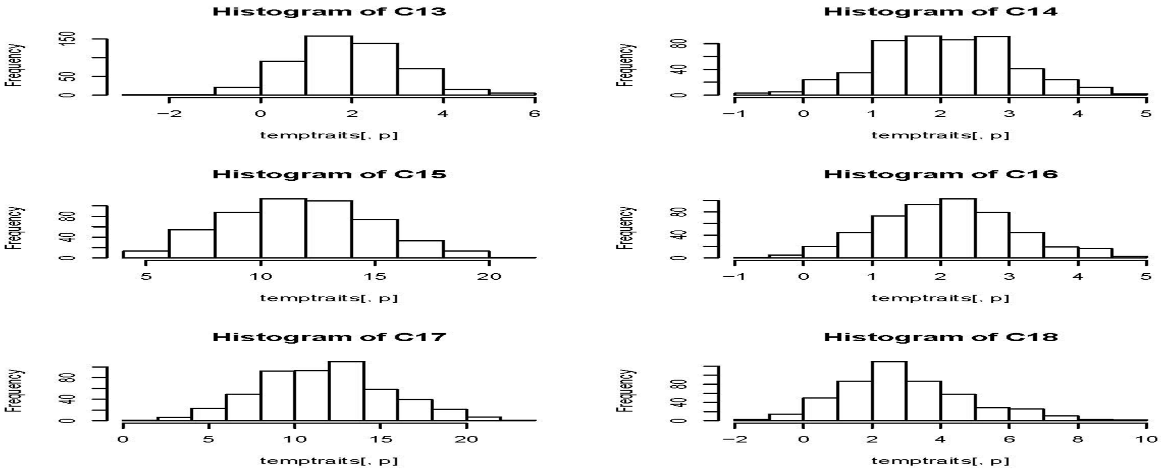


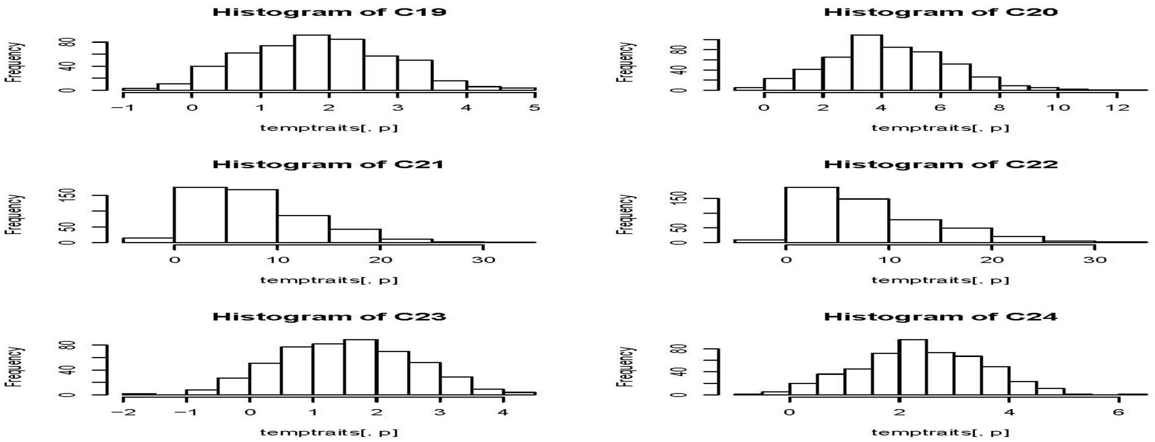


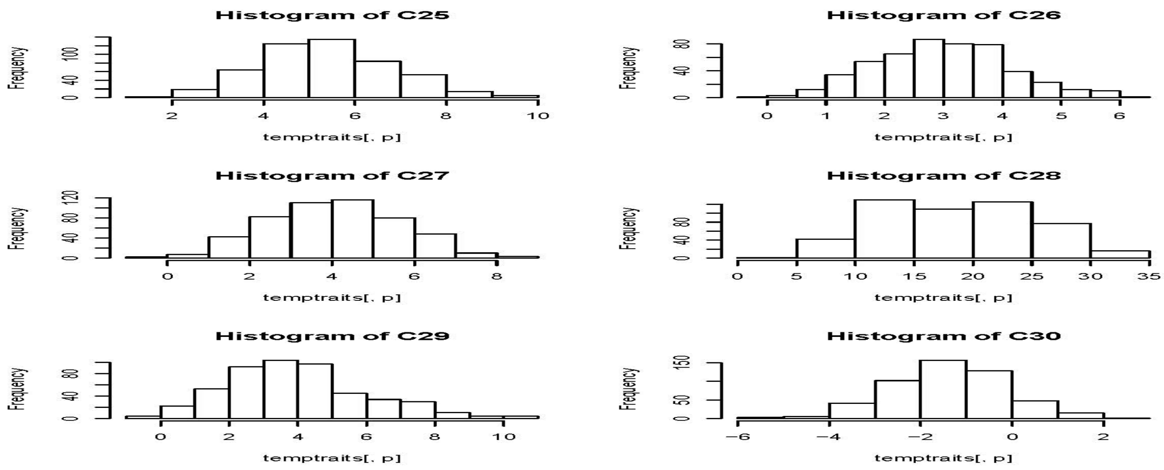




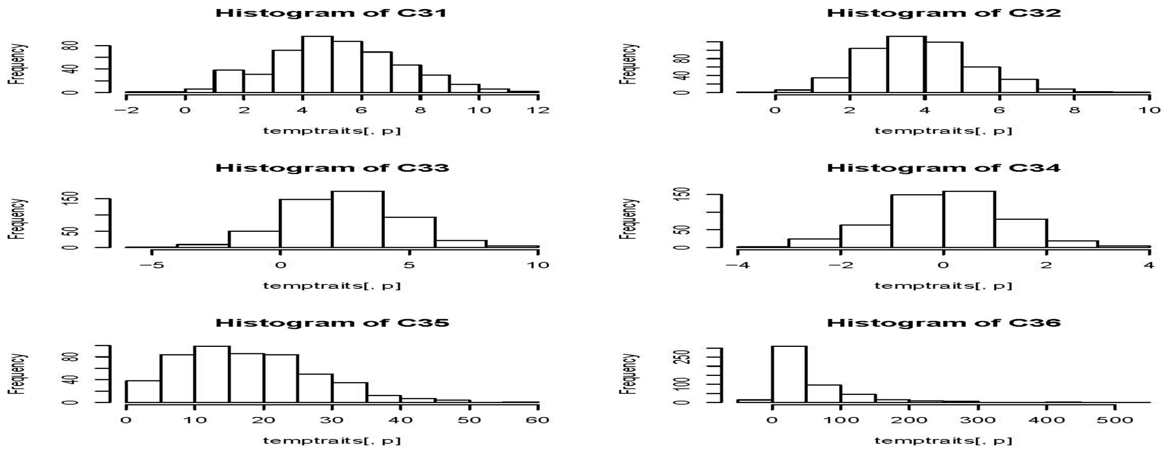

Histogram of C7
Frequency
0 40 80
−1 0 1 2 3 4 5 6
temptraits[, p]
Histogram of C8
Frequency
0 40 80 120
5 10 15 20
temptraits[, p]
Histogram of C9
Frequency
0 40 80
4 6 8 10 12 14
temptraits[, p]
Histogram of C10
Frequency
0 40 80
6 8 10 12 14 16 18
temptraits[, p]
Histogram of C11
Frequency
0 40 80
4 6 8 10 12 14 16
temptraits[, p]
Histogram of C12
Frequency
0 40 80
0 5 10 15 20
temptraits[, p]


Histogram of C13
Frequency
0 50 150
−2 0 2 4 6
temptraits[, p]
Histogram of C14
Frequency
0 40 80
−1 0 1 2 3 4 5
temptraits[, p]
Histogram of C15
Frequency
0 40 80
5 10 15 20
temptraits[, p]
Histogram of C16
Frequency
0 40 80
−1 0 1 2 3 4 5
temptraits[, p]
Histogram of C17
Frequency
0 40 80
0 5 10 15 20
temptraits[, p]
Histogram of C18
Frequency
0 40 80
−2 0 2 4 6 8 10
temptraits[, p]

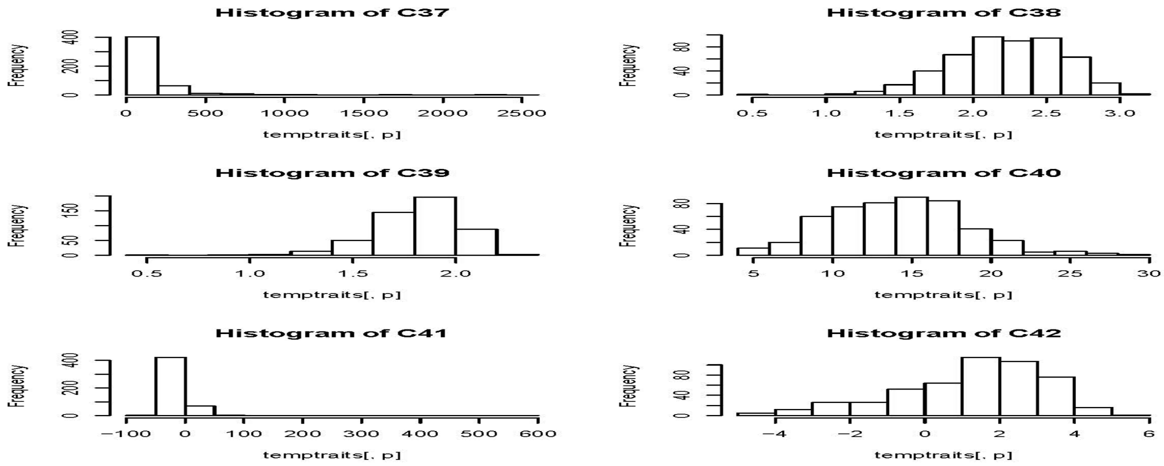

Histogram of C19
Frequency
0 40 80
−1 0 1 2 3 4 5
temptraits[, p]
Histogram of C20
Frequency
0 40 80
0 2 4 6 8 10 12
temptraits[, p]
Histogram of C21
Frequency
0 50 150
0 10 20 30
temptraits[, p]
Histogram of C22
Frequency
0 50 150
0 10 20 30
temptraits[, p]
Histogram of C23
Frequency
0 40 80
−2 −1 0 1 2 3 4
temptraits[, p]
Histogram of C24
Frequency
0 40 80
0 2 4 6
temptraits[, p]



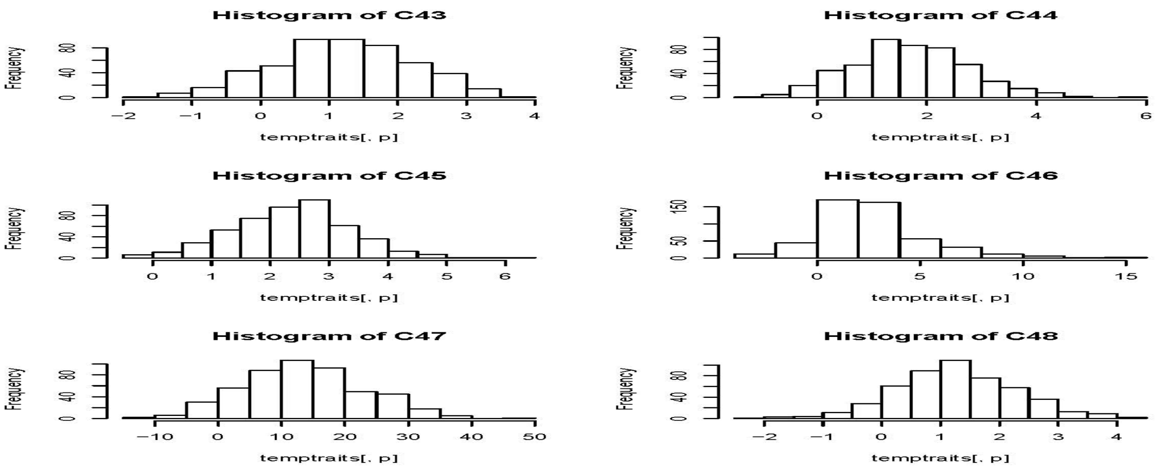
Histogram of C25
Frequency
0 40 100
2 4 6 8 10
temptraits[, p]
Histogram of C26
Frequency
0 40 80
0 1 2 3 4 5 6
temptraits[, p]
Histogram of C27
Frequency
0 40 80 120
0 2 4 6 8
temptraits[, p]
Histogram of C28
Frequency
0 40 80
0 5 10 15 20 25 30 35
temptraits[, p]
Histogram of C29
Frequency
0 40 80
0 2 4 6 8 10
temptraits[, p]
Histogram of C30
Frequency
0 50 150
−6 −4 −2 0 2
temptraits[, p]

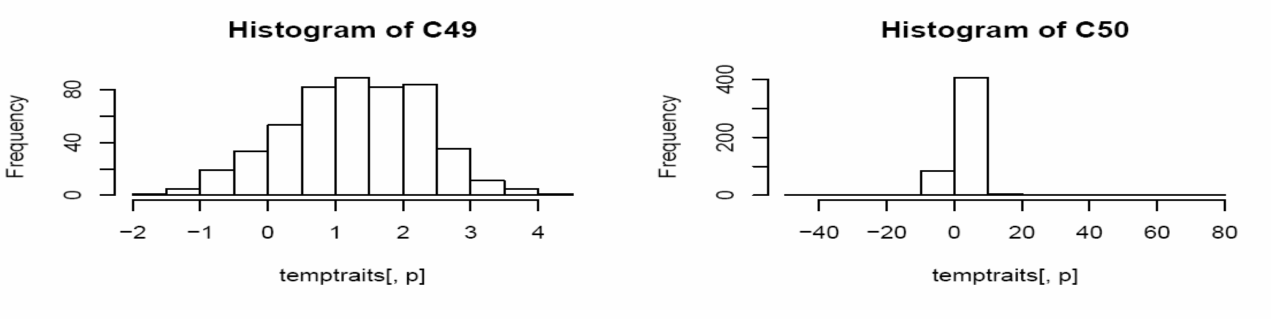


Histogram of C31
Frequency
0 40 80
−2 0 2 4 6 8 10 12
temptraits[, p]
Histogram of C32
Frequency
0 40 80
0 2 4 6 8 10
temptraits[, p]
Histogram of C33
Frequency
0 50 150
−5 0 5 10
temptraits[, p]
Histogram of C34
Frequency
0 50 150
−4 −2 0 2 4
temptraits[, p]
Histogram of C35
Frequency
0 40 80
0 10 20 30 40 50 60
temptraits[, p]
Histogram of C36
Frequency
0 100 250
0 100 200 300 400 500
temptraits[, p]

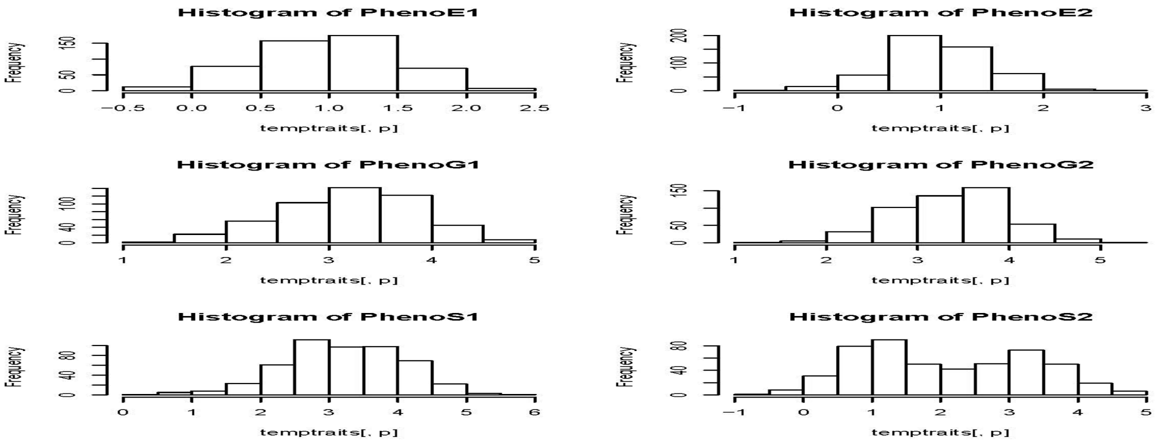


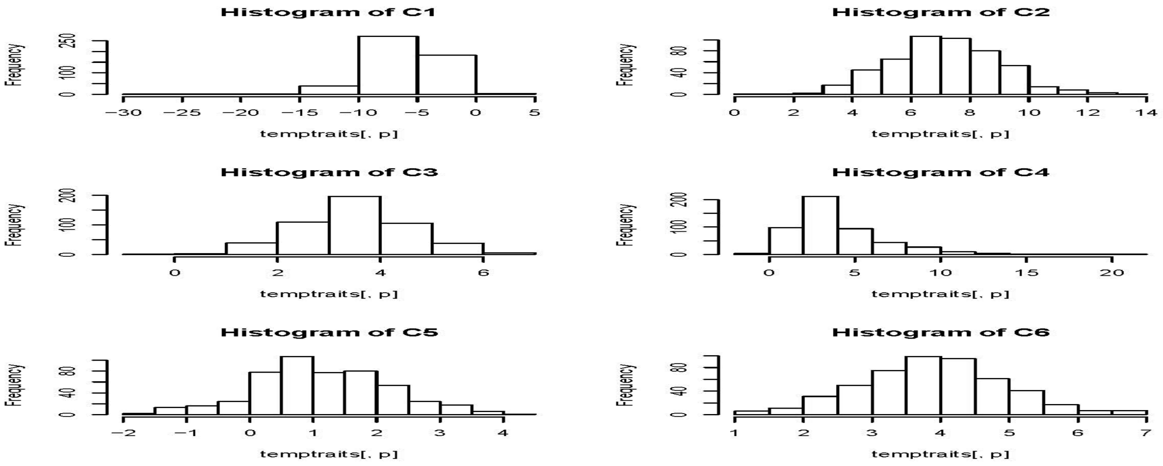
Histogram of C37
Frequency
0 200 400
0 500 1000 1500 2000 2500
temptraits[, p]
Histogram of C38
Frequency
0 40 80
0.5 1.0 1.5 2.0 2.5 3.0
temptraits[, p]
Histogram of C39
Frequency
0 50 150
0.5 1.0 1.5 2.0
temptraits[, p]
Histogram of C40
Frequency
0 40 80
5 10 15 20 25 30
temptraits[, p]
Histogram of C41
Frequency
0 200 400
−100 0 100 200 300 400 500 600
temptraits[, p]
Histogram of C42
Frequency
0 40 80
−4 −2 0 2 4 6
temptraits[, p]

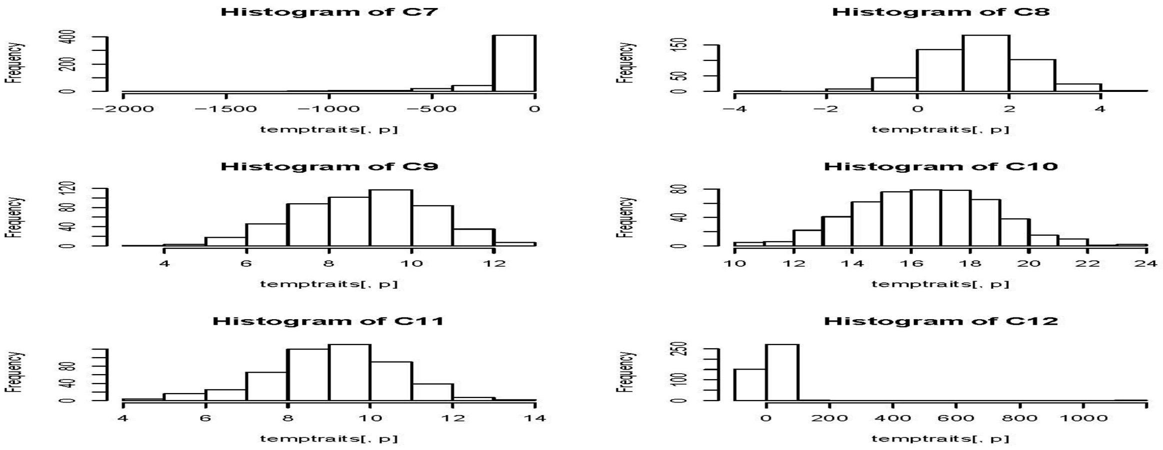


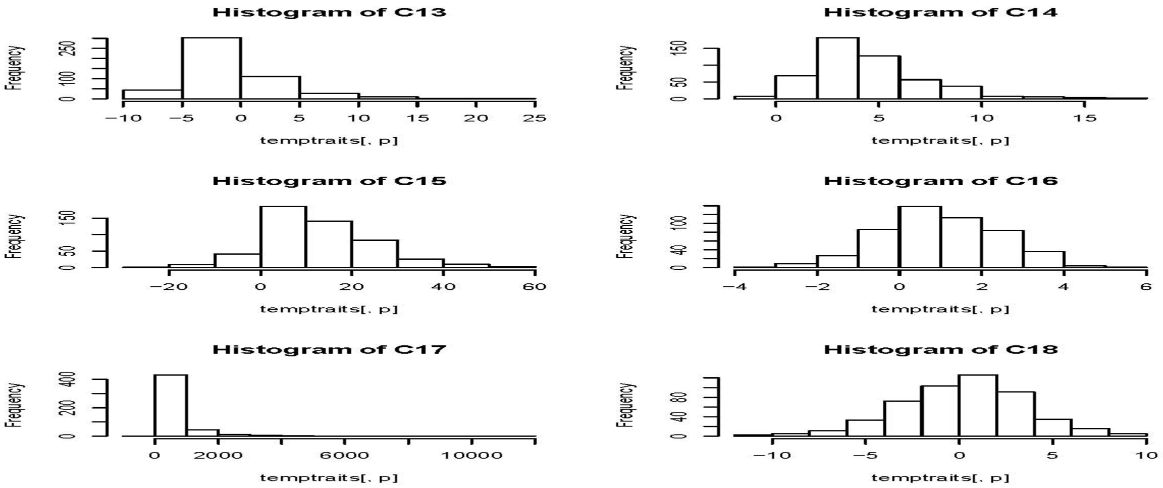


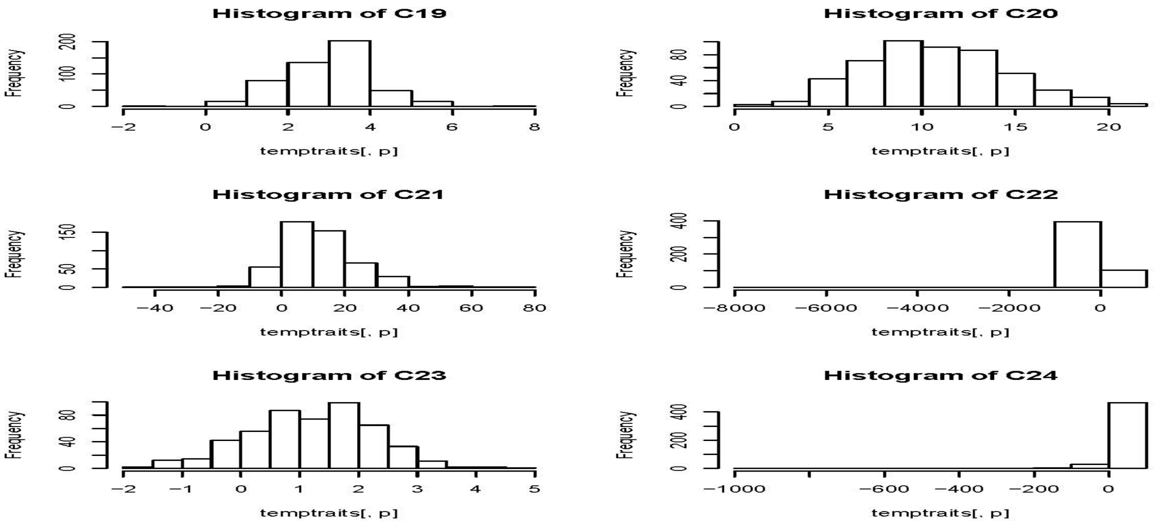
2^{nd} set of 50 functions and 6 underlying traits

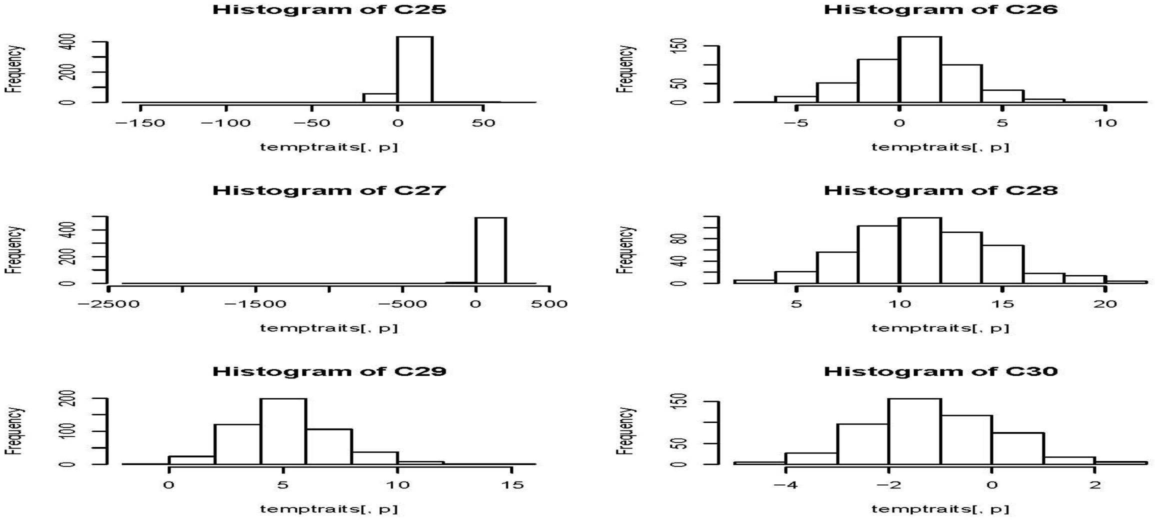


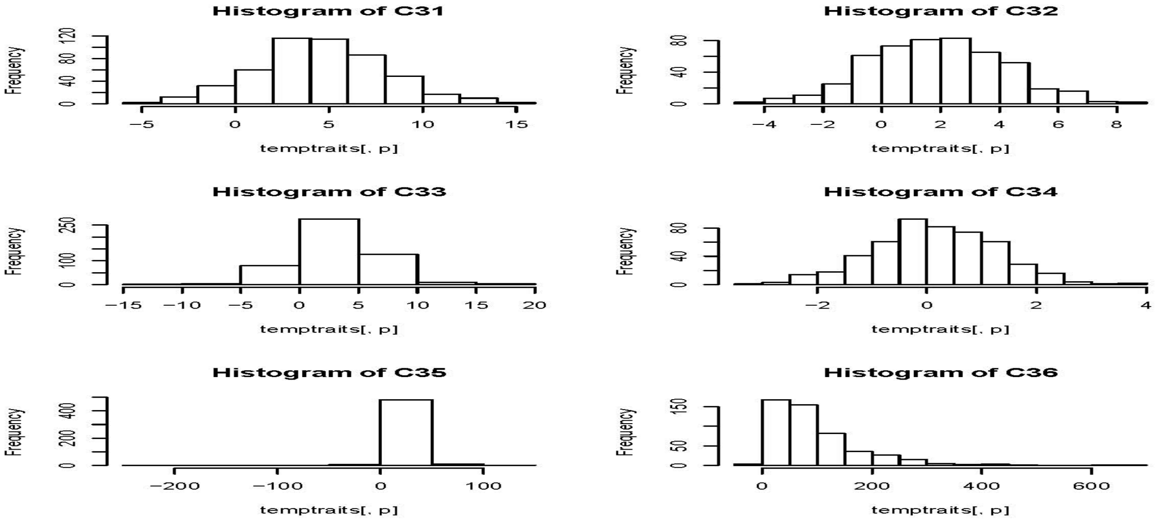

Histogram of C1
Histogram of C2
Histogram of C3
Histogram of C4
Histogram of C5
Histogram of C6
Frequency
temptraits[, p]

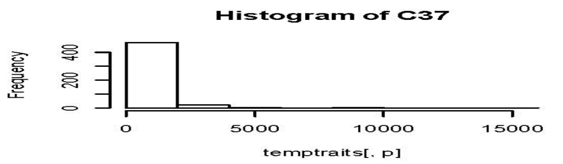

Histogram of C7
Histogram of C8
Histogram of C9
Histogram of C10
Histogram of C11
Histogram of C12
Frequency
temptraits[, p]

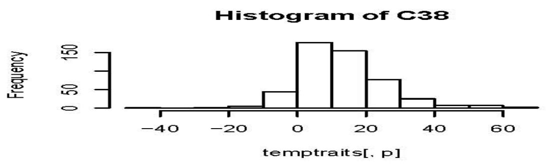

Histogram of C13
Histogram of C14
Histogram of C15
Histogram of C16
Histogram of C17
Histogram of C18
Frequency
temptraits[, p]

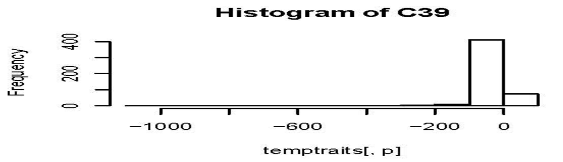

Histogram of C19
Frequency
0
100
200
−2
0
2
4
6
8
temptraits[, p]
Histogram of C20
Frequency
0
40
80
0
5
10
15
20
temptraits[, p]
Histogram of C21
Frequency
0
50
150
−40
−20
0
20
40
60
80
temptraits[, p]
Histogram of C22
Frequency
0
200
400
−8000
−6000
−4000
−2000
0
temptraits[, p]
Histogram of C23
Frequency
0
40
80
−2
−1
0
1
2
3
4
5
temptraits[, p]
Histogram of C24
Frequency
0
200
400
−1000
−600
−400
−200
0
temptraits[, p]

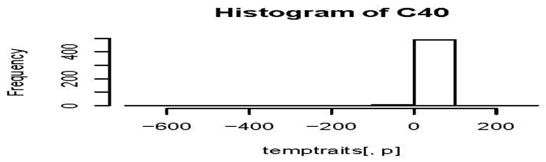

Histogram of C25
Frequency
0
200
400
−150
−100
−50
0
50
temptraits[, p]
Histogram of C26
Frequency
0
50
150
−5
0
5
10
temptraits[, p]
Histogram of C27
Frequency
0
200
400
−2500
−1500
−500
0
500
temptraits[, p]
Histogram of C28
Frequency
0
40
80
5
10
15
20
temptraits[, p]
Histogram of C29
Frequency
0
100
200
0
5
10
15
temptraits[, p]
Histogram of C30
Frequency
0
50
150
−4
−2
0
2
temptraits[, p]

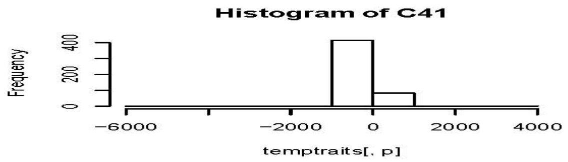

Histogram of C31
Frequency
0
40
80
120
−5
0
5
10
15
temptraits[, p]
Histogram of C32
Frequency
0
40
80
−4
−2
0
2
4
6
8
temptraits[, p]
Histogram of C33
Frequency
0
100
250
−15
−10
−5
0
5
10
15
20
temptraits[, p]
Histogram of C34
Frequency
0
40
80
−2
0
2
4
temptraits[, p]
Histogram of C35
Frequency
0
200
400
−200
−100
0
100
temptraits[, p]
Histogram of C36
Frequency
0
50
150
0
200
400
600
temptraits[, p]

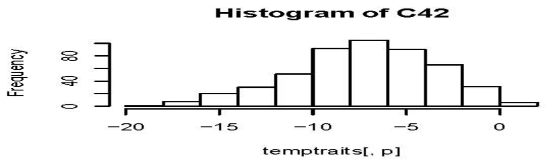

Histogram of C37
Frequency
0
200
400
0
5000
10000
15000
temptraits[, p]

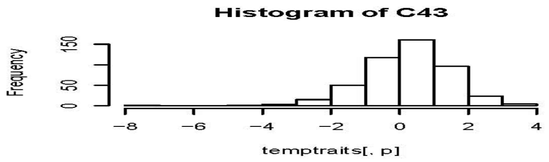

Histogram of C38
Frequency
0
50
150
−40
−20
0
20
40
60
temptraits[, p]

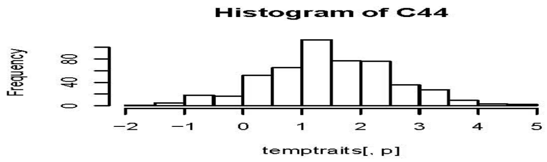

Histogram of C39
Frequency
0
200
400
−1000
−600
−200
0
temptraits[, p]

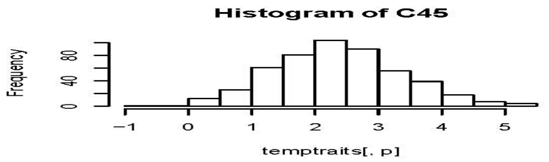

Histogram of C40
Frequency
0
200
400
−600
−400
−200
0
200
temptraits[, p]

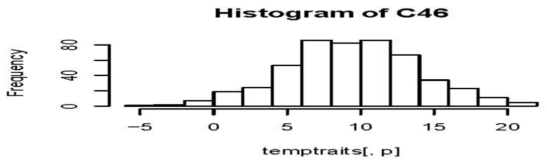

Histogram of C41
Frequency
0
200
400
−6000
−2000
0
2000
4000
temptraits[, p]

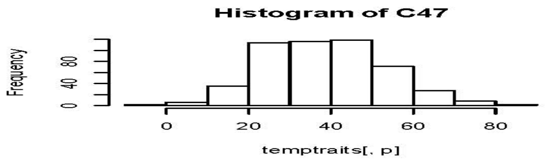

Histogram of C42
Frequency
0
40
80
−20
−15
−10
−5
0
temptraits[, p]

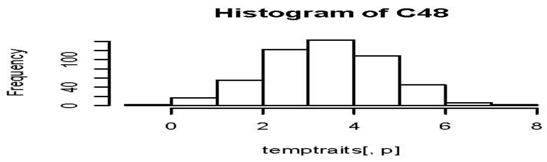

Histogram of C43
Frequency
0
50
150
−8
−6
−4
−2
0
2
4
temptraits[, p]

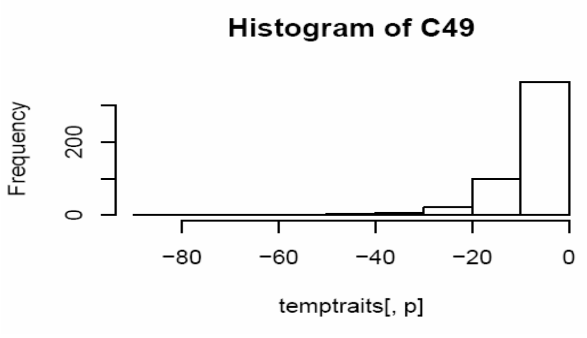

Histogram of C44
Frequency
0
40
80
−2
−1
0
1
2
3
4
5
temptraits[, p]

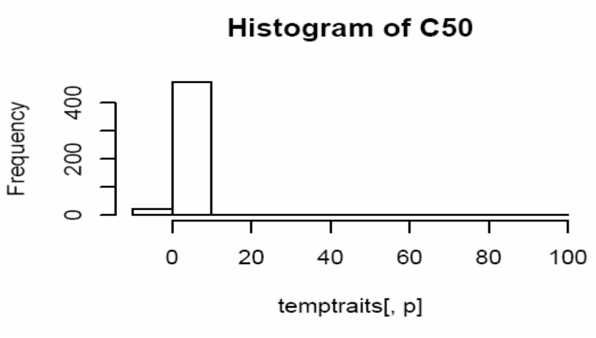

Histogram of C45
Frequency
0
40
80
−1
0
1
2
3
4
5
temptraits[, p]


Histogram of C46
Frequency
0
40
80
−5
0
5
10
15
20
temptraits[, p]


Histogram of C47
Frequency
0
40
80
0
20
40
60
80
temptraits[, p]


Histogram of C48
Frequency
0
40
100
0
2
4
6
8
temptraits[, p]


Histogram of C49
Frequency
0
200
−80
−60
−40
−20
0
temptraits[, p]


Histogram of C50
Frequency
0
200
400
0
20
40
60
80
100
temptraits[, p]

BIBLIOGRAPHY

1. Deng FY, Lei SF, Li MX, Jiang C, Dvornyk V, Deng HW 2006 Genetic determination and correlation of body mass index and bone mineral density at the spine and hip in Chinese Han ethnicity. Osteoporosis International **17**(1)**:**119-24.

2. Hegele RA 1997 Candidate genes, small effects, and the prediction of atherosclerosis. Critical Reviews in Clinical Laboratory Sciences **34**(4)**:**343-67.

3. Li X, Masinde G, Gu W, Wergedal J, Mohan S, Baylink DJ 2002 Genetic dissection of femur breaking strength in a large population (MRL/MpJ x SJL/J) of F2 Mice: single QTL effects, epistasis, and pleiotropy. Genomics **79**(5)**:**734-40.

4. Mitchell BD, Kammerer CM, Mahaney MC, Blangero J, Comuzzie AG, Atwood LD, Haffner SM, Stern MP, MacCluer JW 1996 Genetic analysis of the IRS. Pleiotropic effects of genes influencing insulin levels on lipoprotein and obesity measures. Arteriosclerosis, Thrombosis & Vascular Biology **16**(2)**:**281-8.

5. Martin LJ, Cianflone K, Zakarian R, Nagrani G, Almasy L, Rainwater DL, Cole S, Hixson JE, MacCluer JW, Blangero J, Comuzzie AG 2004 Bivariate linkage between acylation-stimulating protein and BMI and high-density lipoproteins. Obesity Research **12**(4)**:**669-78.

6. Devoto M, Spotila LD, Stabley DL, Wharton GN, Rydbeck H, Korkko J, Kosich R, Prockop D, Tenenhouse A, Sol-Church K 2005 Univariate and bivariate variance component linkage analysis of a whole-genome scan for loci contributing to bone mineral density. European Journal of Human Genetics **13**(6)**:**781-8.

7. Livshits G, Deng HW, Nguyen TV, Yakovenko K, Recker RR, Eisman JA 2004 Genetics of bone mineral density: evidence for a major pleiotropic effect from an intercontinental study. Journal of Bone & Mineral Research **19**(6)**:**914-23.

8. Li X, Quinones MJ, Wang D, Bulnes-Enriquez I, Jimenez X, De La Rosa R, Aurea GL, Taylor KD, Hsueh WA, Rotter JI, Yang H 2006 Genetic effects on obesity assessed by bivariate genome scan: the Mexican-American coronary artery disease study. Obesity **14**(7)**:**1192-200.

9. Lehman DM, Arya R, Blangero J, Almasy L, Puppala S, Dyer TD, Leach RJ, O'Connell P, Stern MP, Duggirala R 2005 Bivariate linkage analysis of the insulin resistance syndrome phenotypes on chromosome 7q. Human Biology **77**(2)**:**231-46.

10. Chase K, Carrier DR, Adler FR, Jarvik T, Ostrander EA, Lorentzen TD, Lark KG 2002 Genetic basis for systems of skeletal quantitative traits: principal component analysis of the canid skeleton. Proceedings of the National Academy of Sciences of the United States of America **99**(15)**:**9930-5.

11. Guo Y, Zhao LJ, Shen H, Guo Y, Deng HW 2005 Genetic and environmental correlations between age at menarche and bone mineral density at different skeletal sites. Calcified Tissue International **77**(6)**:**356-60.

12. Lin HF, Boden-Albala B, Juo SH, Park N, Rundek T, Sacco RL 2005 Heritabilities of the metabolic syndrome and its components in the Northern Manhattan Family Study. Diabetologia **48**(10)**:**2006-12.

13. Musani SK, Huang-Ge Z, Hsu H-C, Yi N-J, Gorman BS, Allison DB 2006 Principal component analysis of quantitative trait loci for immune response to adenovirus in mice. Hereditas **143:**189-197.

14. Hakulinen MA, Day JS, Toyras J, Weinans H, Jurvelin JS 2006 Ultrasonic characterization of human trabecular bone microstructure. Physics in Medicine & Biology **51**(6)**:**1633-48.

15. Karasik D, Cupples LA, Hannan MT, Kiel DP 2004 Genome screen for a combined bone phenotype using principal component analysis: the Framingham study. Bone **34**(3)**:**547-56.

16. Peacock M, Koller DL, Hui S, Johnston CC, Foroud T, Econs MJ 2004 Peak bone mineral density at the hip is linked to chromosomes 14q and 15q. Osteoporosis International **15**(6)**:**489-96.

17. Holberg CJ, Halonen M, Solomon S, Graves PE, Baldini M, Erickson RP, Martinez FD 2001 Factor analysis of asthma and atopy traits shows 2 major components, one of which is linked to markers on chromosome 5q. Journal of Allergy & Clinical Immunology **108**(5)**:**772-80.

18. Lee WT, Cheung AY, Lau J, Lee SK, Qin L, Cheng JC 2004 Bone densitometry: which skeletal sites are best predicted by bone mass determinants? Journal of Bone & Mineral Metabolism **22**(5)**:**447-55.

19. Austin MA, Edwards KL, McNeely MJ, Chandler WL, Leonetti DL, Talmud PJ, Humphries SE, Fujimoto WY 2004 Heritability of multivariate factors of the metabolic syndrome in nondiabetic Japanese americans. Diabetes **53**(4)**:**1166-9.

20. Anna B Costello JWO 2005 Best Practices in Exploratory Factor Analysis: Four Recommendations for Getting the most from your analysis Practical Assessment Research and Evaluation **10**(7)**:**1-9.

21. Pan L, Ober C, Abney M 2007 Heritability Estimation of Sex-Specific Effects on Human Quantitative Traits. Genetic Epidemiology [Epub ahead of print].

22. Almasy L, Blangero J 1998 Multipoint quantitative-trait linkage analysis in general pedigrees. American Journal of Human Genetics **62**(5)**:**1198-211.

23. Hsu FC, Zaccaro DJ, Lange LA, Arnett DK, Langefeld CD, Wagenknecht LE, Herrington DM, Beck SR, Freedman BI, Bowden DW, Rich SS 2005 The impact of pedigree structure on heritability estimates for pulse pressure in three studies. Human Heredity **60**(2)**:**63-72.

24. Becker RA, M. CJ, Wilks AR 1988 The New S Language. Chapman & Hall, New York.

25. Falconer DS, Mackay TFC 1996 Introduction to Quantitative Genetics 4th ed. Prentice Hall, pp 480.

26. van Asselt KM, Kok HS, van der Schouw YT, Peeters PH, Pearson PL, Grobbee DE 2006 Role of genetic analyses in cardiology: part II: heritability estimation for gene searching in multifactorial diseases. Circulation **113**(8)**:**1136-9.