An improved tree-based statistical method for genome-wide association study

Dwueng-Chwuan Jhwueng[1]

[1]**Department of Statistics, Feng-Chia University, No. 100 Wenhua Rd., Seatwen Taichung Taiwan**

Corresponding author:
Dwueng-Chwuan Jhwueng[1]

Email address: dcjhwueng@fcu.edu.tw

ABSTRACT

In genetic studies, quantitative traits are found possibly associated with genetic data. Due to advanced sequencing technology, many methods have been proposed in genome wide association study (GWAS) to search the single nucleotide polymorphism (SNP) associated with the traits. Currently several methods that account for the evolutionary relatedness among individuals were developed. When comparing with conventional methods without evolutionary relatedness among individuals, tree based methods are found to have better performance when the population structure increases. In this work, we extend a tree based method in previous studies by varying the magnitude of relatedness. The magnitude of relatedness of the evolutionary history is controlled by an Ornstein-Uhlenbeck (OU) process through its parameters. Our method combines a pertinent process and phylogenetic comparative method where the incorporated evolutionary history is built by SNP data. We perform simulation as well as analyze drosophila longevity data set.

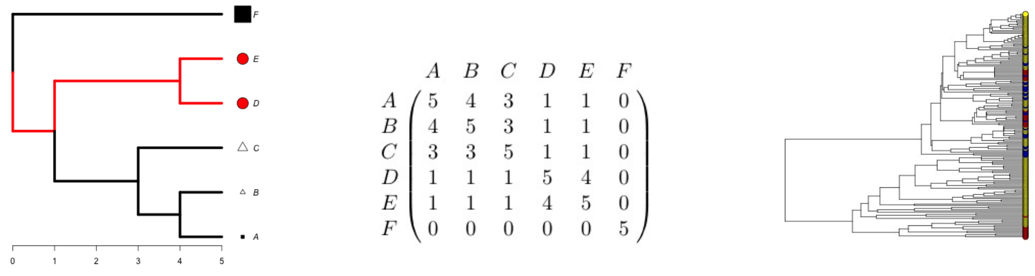
INTRODUCTION

In genetics, searching single nucleotide polymorphism (SNPs) associated with traits helps people to identify and localize the possible origin of disease. In the past, scientists made effort to develop methods that aim to utilize SNPs for seeking connection with relevant trait. As the main goal is to find possible association, SNP from sequencing technology as well as traits measured from various ways are collected from controlled group and case group. However, currently it is still quite challenge to find statistical significance in association study. One difficulty is to successfully link the SNPs with the traits. As those studies were required to meet sufficient rigorous statistical tests during the process. From genetic basis, SNP makers are scanned into analysis using a couple of thousand individuals each with certain long sequence length (around 0.5 million in general).

The statistical methods developed for association studies in literature can be divided into two main categories: the one assumed the evolutionary independence of individuals without relatedness and the other incorporates the evolutionary relatedness of individuals into analysis. For the case of independence assumption among individuals, the observed trait of n individuals with values $y_1, y_2, \cdots, y_n$ are assumed as independent identical distributed random variable from identical statistical distribution. To detect association between trait and SNP datasets, under the evolutionary *independence* assumption, typically a paired t-test is conducted for investigating the significance of the SNPs associated with the trait of interest on the controlled group $y_{i_1}, y_{i_2}, \cdots, y_{i_{n_1}}$, and with the trait on the case group $y_{j_1}, y_{j_2}, \cdots, y_{j_{n_2}}$ where $n_1 + n_2 = n$ (McClurg et al., 2006). The paired t-test method serves a fast and efficient way in association study (Thompson and Fardo, 2016).

In the other category of study for linking SNP and traits, people incorporated evolutionary relatedness represented by a tree for association analysis (Pan et al., 2009; Zhang et al., 2012). A previous work (Thompson and Kubatko, 2013) demonstrated that the tree based method for linking the association between traits and gene can be improved when the covariance structure $\boldsymbol{V}$ among randomly-sampled individuals is estimated from the evolutionary history within each SNP. To initiate the analysis, those methods make use of SNP data to build a phylogenetic tree $\mathbb{T}$ which is a rooted, bifurcated (or multifurcated)

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27118v1 | CC BY 4.0 Open Access | rec: 15 Aug 2018, publ: 15 Aug 2018



$$\begin{array}{c|cccccc} & A & B & C & D & E & F \\ \hline A & 5 & 4 & 3 & 1 & 1 & 0 \\ B & 4 & 5 & 3 & 1 & 1 & 0 \\ C & 3 & 3 & 5 & 1 & 1 & 0 \\ D & 1 & 1 & 1 & 5 & 4 & 0 \\ E & 1 & 1 & 1 & 4 & 5 & 0 \\ F & 0 & 0 & 0 & 0 & 0 & 5 \end{array}$$

(a) (b) (c)

Figure 1. (a) A demonstrated example of evolutionary tree. The horizontal axis represents a pseudo-evolutionary time from past to current. The evolutionary time started with $t = 0$ and stopped at current at $t = 5$. (b) The matrix representation $\boldsymbol{V}$ for the evolutionary tree. (c) A phylogenetic tree built from SNP of 164 diploid observations. The corresponding traits are represented by colored circles. Three colors (red, yellow, and blue) represent the magnitude of the trait value with hypothetical low values of yellow, hight value in red and intermediate values colored blue. Tree and traits were obtained from (Schmitz, 2017).

directed and ultrametric (each individual has the same height from the root to tip in the tree) graph. To given an illustration, we use a simple tree containing a few individuals. The evolutionary relatedness of five individuals A, B, C, D, E, F is shown by a tree in Fig. 1a. It is expected that the level of relatedness among individuals contain some useful information linking to the trait. For example, as two individuals D and E shared more evolutionary history, it is reasonable to think that their trait are more similar (shown in red circles ●). While individuals A and F are more evolutionary unrelated (independent), hence their characteristics might present more diversity (e.g. the two black squares ■ and ▪in different sizes). The matrix $\boldsymbol{V}$ shown in Fig. 1b is the covariance matrix (an isomorphic transformation) of the tree in Fig. 1a. An element v_{ij} in the matrix $\boldsymbol{V}$ represents relatedness between a pair of individuals i and j. Note that v_{ij} is obtained by measuring the shared evolutionary history from root (scaled at 0) of the tree to their most recent common ancestor.

The tree in Fig. 1c is a larger tree constructed using a SNP data of 164 individuals. It is reasonable to view that trait among individuals are more similar when sharing higher relatedness.

In this paper, we intend to expand model in Thompson and Kubatko (2013). We start by briefly introducing the tree based method as following. Considered a cluster of tree where the trait of n individuals are separated into k clusters. We can use an n by k matrix $D = [d_{ij}]$ to represent the cluster where D is defined by

$$d_{ij} = \begin{cases} 1, & \text{if observation } i \text{ falls in cluster } j; \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

where $i = 1, 2, \cdots, n$ and $j = 1, 2, \cdots, k$.

The matrix D will be useful for the next step analysis of studying trait evolution. Let $Y = (y_1, y_2, \cdots, y_n)^t$ be the trait observed from n individuals, Y can be treated as a random variable with expected value $\mathrm{E}[Y] = D\mu \in \mathbb{R}^n$ where the vector $\mu = (\mu_1, \mu_2, \cdots, \mu_k)^t$ is identified as the mean for the k distinct groups.

We can get cluster trees from setting different number of clusters. The clustered tree can then be transformed into the variance covariance matrix V. We illustrate this by reproducing Fig. 2 in (Thompson and Kubatko, 2013). In Fig. 2, three clustered trees for 6 individuals are shown with different cluster number $k = 0, 2, 3$. The corresponding matrices Vs for the tree of $k = 0$ and $k = 2$ are shown in Table 1.

Here we use the clustered tree to consider the broad-scale phylogenetic relationships among SNPs, this can account for the evolutionary history among genes with using all coalescent relationships where the structure of V is equivalently to the tree topology, and each element in V is an estimate of the covariance structure in the data that is required for estimation of branch lengths along the topology.

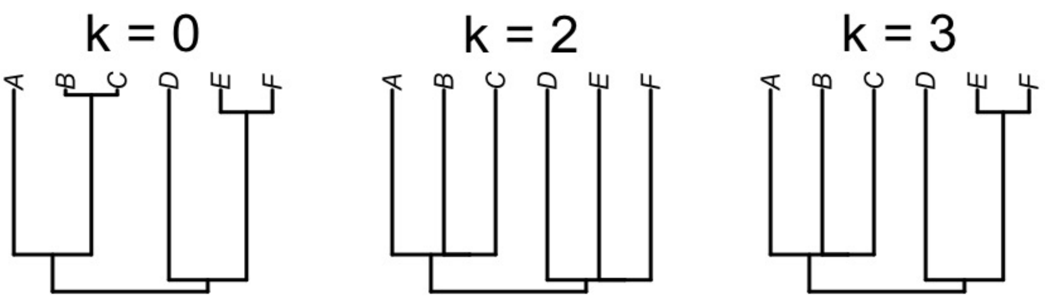


Figure 2. A demonstration of a six-taxon tree with branch lengths. The overall tree ($k = 0$) is shown in the left panel. The corresponding clustered trees for 2 clusters ($k = 2$) and 3 clusters ($k = 3$) are in middle panel and right panel, respectively.

Table 1. The variance covariance matrix V for the tree in Fig. 2. The left matrix is for the left tree $k = 0$, the right matrix is for the middle tree $k = 2$. The numbers in bold in the matrix shows the difference between the two clustering results.

	A	B	C	D	E	F
A	100	18	18	0	0	0
B	18	100	**98**	0	0	0
C	18	**98**	100	0	0	0
D	0	0	0	100	6	6
E	0	0	0	6	100	**89**
F	0	0	0	6	**89**	100

	A	B	C	D	E	F
A	100	18	18	0	0	0
B	18	100	**18**	0	0	0
C	18	**18**	100	0	0	0
D	0	0	0	100	6	6
E	0	0	0	6	100	**6**
F	0	0	0	6	**6**	100

Model for Haploid Data

Haploid of a cell has a single set of unpaired chromosome. For SNPs data of haploid type, the tree can be constructed from sequencing reads as well as from assembled genomes or contigs. Thompson and Kubatko (2013) used a transformation by considering clustering using tree structure that clusters the individual into several subgroups depending on the number of k. In contrast with work in Besenbacher et al. (2008), Thompson and Kubatko (2013) instead assumed that an observation is taken to be a chromosome level which offers an alternative to aggregate information. Next, assume a Brownian motion for trait evolution on the tree (Felsenstein, 1985), the statistical model given a trait Y and a tree $\boldsymbol{T}$ follows a multivariate normal distribution with mean vector $D\mu$ and variance-covariance matrix $\sigma^2 V$

$$Y \sim \mathcal{N}(D\mu, \sigma^2 V) \tag{2}$$

where the parameter σ measures the rate of evolution during the process.

The statistical model in Eq. (2) has analytical formula for the maximum likelihood estimators for the mean $\hat{\mu} = (D^T V^{-1} D^T)^{-1} D^T V^{-1} Y$ and variance $\hat{\sigma}^2 = (Y - D\hat{\mu})^T V^{-1} (Y - D\hat{\mu})/n$, respectively. Therefore, the maximum likelihood can be computed directly once the trait and tree are ready. Thompson and Kubatko (2013) used likelihood score statistics (LSS) score to determine the tree score. LSS is defined as the maximum score over the number of clusters.

$$\text{LSS} = \max_k \{2\ell(\hat{\mu}, \hat{\sigma}^2 | Y, V_k) - k \log n\} \tag{3}$$

where $\ell(\cdot)$ is the log likelihood in Eq. (2). Therefore the hypothesis test for detecting significance of association between SNP and trait of a group of individuals can be carried out using a likelihood framework.

Inference

In order to identify the detection between SNP and trait, Thompson and Kubatko (2013) use the LSS score in Eq. (3) for model in Eq. (2). To access the significance, a permutation test was performed and the LSS score for the tree is calculated according to each permuted trait data set at each locus along a chromosome. The statistical null hypothesis is setting with no linking between the snp and trait. i.e. H_0 : no association.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27118v1 | CC BY 4.0 Open Access | rec: 15 Aug 2018, publ: 15 Aug 2018

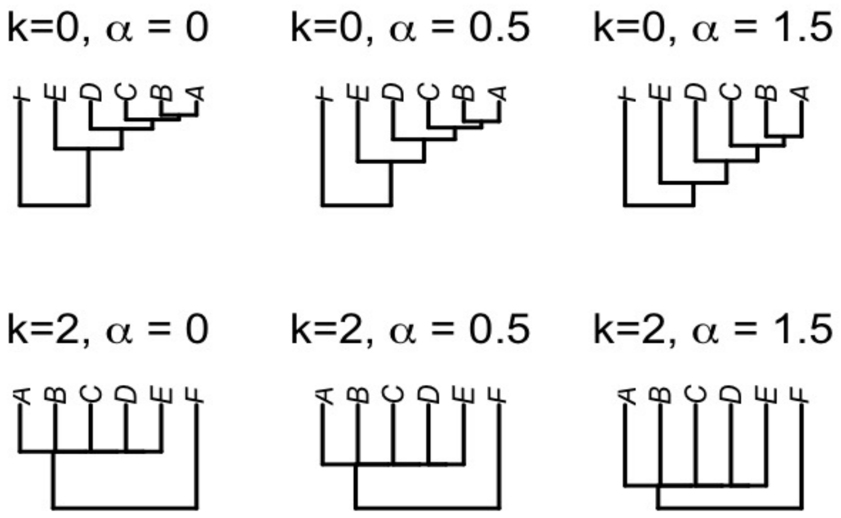


Figure 3. The magnitude of hierarchical clustering is controlled under the OU process using parameter α.

Then a p-value is determined by the ranking of score of the observed data set in an ordered score of the permuted data set. i.e. p-value $= \#\{\text{LSS}_{per} > \text{LSS}_{obs}\}/N$ where LSS_{per} is the score for permuted data set, LSS_{obs} is the score for the observed data set and N is the number of permutations.

In fact, model built in Eq. (2) is under the assumption of Brownian motion for evolution Felsenstein (1985) since the variance covariance matrix is specified by utilizing the tree. Observing that currently there is still a need of methods that include more biologically-realistic situations, we propose a method that extend works in (Thompson and Kubatko, 2013) by introducing Ornstein-Uhlenbeck (OU) process Hansen and Martins (1996) for studying the association between traits and SNP data. Our aim is hope to provide a robust method in GWAS study.

METHODS

OU process for trait evolution

If the trait of the ith individual is assumed to evolve under an Ornstein-Uhlenbeck process (Hansen and Martins, 1996), then the trait value of the individual at time t, denoted as a stochastic variable $y_{i,t}$, is a solution the following stochastic differential equation

$$dy_{i,t} = \alpha(\mu - y_{i,t})dt + \sigma dB_t,\ t > 0. \tag{4}$$

Eq. (4) expresses the dynamic of $y_{i,t}$ with respect to time. On the left hand side of Eq. (4), the term $dy_{i,t}$ is the change in the character $y_{i,t}$ over the infinitesimal interval from time t to $t+dt$. The right hand side of Eq. (4) contains the sum of two terms: the deterministic term $\alpha(\mu - y_t)dt$ and the stochastic terms σdB_t where B_t is a Brownian motion, the real value parameter μ represents the optimal value (an evolutionary niche) of $y_{i,t}$, the positive value parameter σ is the overall rate of evolution, and the positive value parameter α represents the magnitude of force that pulls $y_{i,t}$ back to the optimum μ. When $y_{i,t}$ is far from the optimal μ, the force would have stronger effect (larger value of α) to pull $y_{i,t}$ back to the optimum μ while weaker force (smaller value of α) is presented whenever $y_{i,t}$ is close to the neighborhood of μ.

To implement OU model in tree based genome wide association study, we use α to control the level of clusters. In Fig. 3, larger values of $\alpha = 0.5$, or $\alpha = 1$ provided more independent relatedness among clusters than the smaller value of $\alpha = 0$ given different number of cluster $k = 0$ (plots in upper panel for the raw tree case) or $k = 2$ (plots in lower panel for 2 cluster case). Implementing OU process could be a potential benefit for detecting the association between snp and trait.

Currently we focus on studying the model in (Hansen and Martins, 1996) with single force, single optimum and single rate as mentioned in Eq. (4) though other more sophiscated models are possible to develope for this purpose (see (OMeara et al., 2006; Butler and King, 2004; Beaulieu et al., 2012)).

OU Model for Haploid Data

The observed trait for the ith individual $y_{i,t}$ under the OU process has normal distributions with the mean

$$E(y_{i,t}|y_0) = y_0 \exp(-\alpha t) + \mu(1 - \exp(-\alpha t)) \tag{5}$$

and variance

$$var[y_{i,t}|y_0] = \frac{\sigma^2}{2\alpha}(1 - \exp(-2\alpha t)) \tag{6}$$

where $y_0 = y_{i,0}$ is the trait value at $t = 0$.

The method used in Thompson and Kubatko (2013) set the parameter $\alpha = 0$ which reduces it to the Brownian Motion with means $E(y_{i,t}|y_0) = y_0$ and $var(y_{i,t}|y_0) = \sigma^2 t$. For OU model with n individuals, the observed trait $Y = (y_1, y_2, \cdots, y_n)^t$ is treated as a random vector that following a multivariate normal distribution with mean vector $\mu = (\mu_1, \mu_2, \cdots, \mu_n)^t_{n\times 1}$ and variance-covariance matrix $V_{n\times n}$ where $V[i,j] = cov[y_{i,t}, y_{j,t}]$ is the covariance between species i and species j of the form

$$cov[y_{i,t}, y_{j,t}] = \sigma^2 V_{\alpha_{ij}} = \frac{\sigma^2}{2\alpha} e^{-\alpha d_{ij}} e^{-2\alpha t_{ij}} \tag{7}$$

where t_{ij} is the branch length shared by the ith and the jth individual and d_{ij} is the distance between the ith and the jth individual on the tree.

Under the OU process, the trait vector observed at the tip denoted as $Y = (y_1, y_2, \cdots, y_n)^t$ would follow a joint multivariate normal distribution

$$Y \sim \boldsymbol{MVN}(D\mu, \sigma^2 V_\alpha) \tag{8}$$

The mean vector and variance can be expressed as a function of α

$$\hat{\mu}(\alpha|\mathbb{T}, Y, D) = (D^T V_\alpha^{-1} D^T)^{-1} D^T V_\alpha^{-1} Y, \tag{9}$$

$$\hat{\sigma}^2(\alpha|\mathbb{T}, Y, D, \hat{\mu}) = \frac{(Y - D\hat{\mu})^T V_\alpha^{-1} (Y - D\hat{\mu})}{n}. \tag{10}$$

By Eq. (9) and Eq. (10), the negative log likelihood function for OU model can be written as a function of α :

$$\ell(\alpha|Y, \mathbb{T}, \hat{\mu}, \hat{\sigma}^2) = \frac{n}{2}\log(2\pi) + \frac{n}{2}\log\hat{\sigma}^2 + \frac{1}{2}\log|V_\alpha| + \frac{1}{2\hat{\sigma}^2}(Y - D\hat{\mu})^t V_\alpha^{-1}(Y - D\hat{\mu}). \tag{11}$$

Inference

From the model in Eq. (8), the hypothesis testing for significance between SNPs and trait can be proceeded through a likelihood framework. To choose the best cluster, we modify the penalized likelihood approach in Thompson and Kubatko (2013) where the likelihood score statistics is calculated as

$$\text{LSS} = \max_{0\le k\le m}\{2\log(\hat{\alpha}, \hat{\mu}, \hat{\sigma}^2|Y, V) - k\log n\} \tag{12}$$

where m is the maximum number of clusters that used for analysis.

To access the statistical significance, we further consider to use an upper bound defined by the maximum of the observed LSS value plus the standard error of the permuted maximum LSS valued multiplied by the $(1-\alpha)$ quantile of t distribution with degree of freedom of $n-1$ where n is the number of individuals. i.e.

$$b = \max_{0\le k\le m} LSS_{\text{obs}} + qt_{\alpha/2, df=n-1}\frac{\text{sd}_{LSS_{per}}}{\sqrt{n}} \tag{13}$$

where $\text{sd}_{LSS_{per}} = \text{sd}(\{\max_{0\le k\le m} LSS_{\text{per}}\}_{i=1}^m)$. And the p-value is caluclated by the number of permuted LSS score greater than this bound b. This provide a more conservative alternative in detecting the significance.

SIMULATION

Haploid Data

In order to assess the performance of the proposed techniques, we simulating the data sets under specific parameter values, the local phylgogenetic tree at each SNP is estimated using SVDquatets (Chifman and Kubatko, 2014). The SVDquatets is currently implemented in PAUP (Swofford, 2011) and computes a score based on singular value decomposition of a matrix of site pattern frequencies corresponding to a split on a phylogenetic tree. These quartet scores can be used to select the best supported topology for quartets of taxa, which in turns can be used to infer the species phylogeny using quartet methods where branch lengths are estimated.

Given an estimated $\mathbb{T}$, the next step to complete the association study is to conduct the phylogenetic comparative analysis to computing the LSS score. Tree from PAUP analysis is a non-ultrametirc tree. Since the expected quantity of trait change (the variance v_{ii} in the variance-covariance matrix V) in comparative analysis under Brownian motion is given by the product of the rate of evolution σ of the trait with branch length and under OU process the $v_{ii}(\alpha)$ is given by the deterministic change inherit from ancestor plus the Brownian motion for random change. So using a non-ultrametric tree is a way to assume different rates of evolution for each branch which leads to a more sophisticated and complex case. To alleviate this, we convert the non-ultrametric tree to an ultrametric tree using the mean path lengths (MPL)(Britton et al., 2002) method where the age of a node is estimated with the mean of the distances from this node to all tips descending from it. Hence we can assume a clock-like trait evolution which means the quantity of change from the root to the tips is the same.

To calculate score in Eq. (12), we current use the number of cluster from $k = 3$ to $k = 5$. Algorithm 1 provides a step-by-step procedure for calculating the p-value.

For each size, we use `ms`(Hudson, 2004) to simulate sequence of length 1000. We use `paup` to analyze the sequence and get the tree by `SVDquartets`. For each haploid size, we simulate 100 replicates of sequence to get 5 trees respectively. To simulate trait, given a tree with known topology and branch length, we consider to use two stages OU model with parameters $\Theta = (\alpha_1, \alpha_2, \sigma_1, \sigma_2, \theta_0, \theta_1, \theta_2)$ where for BM data simulating using $\alpha_1 = \alpha_2 = 1e-6; \theta_0 = 90, \theta_1 = 80, \theta_2 = 100$, we set three different rate evolution $\sigma_1 = \sigma_2 = 1, 5$ and 10, respectively.

We use 100 replicates where for each replicates we simulate traits using the true parameters Θ. We then consider to estimate the parameters using the 100 replicates. Since there are various clusters, we use the parameter estimate from the best selected cluster k^*. For each replicate, we consider to assess the significance of the trait associated with the simulated snp. We use the permutation method in algorithm Thompson and Kubatko (2013) to permute the trait for 500 times. We present our algorithm in Algorithm 1.

RESULT

We present out simulation results for BM model and OU model in the following subsection.

Haploid: OU vs BM

We first simulate snp sequence using `ms`(Hudson, 2002). The `ms` settting

```
ms  10  1 -T -s  1000
```

would generate 10 individuals each is with sequence length 1000. We then use this data and `paup`(Swofford, 2011) to obtain the estimated tree. We simulate traits under BM model using $\sigma^2 = 1$ and treated it as the true data set. To evaluate the p-value for this data set, we use 500 replicates, and a p-value is computed by the ratio of count of the maximum LSS statistics greater than the max LSS of the true trait over 500. For each replicates, we compare the LSS statistics of $k = 3, 4, 5$ cluster to get the maximum LSS statistics. We repeat above procedure 50 times for each tree estimated using of sequence length 1000. The following table is the overall average of parameter estimates using 10 trees. We consider the taxa size of $10, 30, 50$ Table 2 shows the median estimate and the 95% confidence interval of the p-value under BM and OU model

Currently we found for BM model, the overall p-value bandwidth is narrower when compare to OU model. This might indicates that OU model is more conservative to detect the significance than the BM

Algorithm 1 Model Inference

```
1:  simulate snp sequence data set from ms and treat $Y$ as the true data set.
2:  use PAUP and SVDquatets to analyze the data sets and return an estimate tree $\mathbb{T}$ with topology
    and branch length information.
3:  for $j = 1 : k$ do
4:  cluster the tree $\mathbb{T}$ and get $\mathbb{T}_j$ of $j$ cluster and store matrix $D$ and variance covariance matri $V_j$.
5:      if model is BM
6:          simulate trait data $Y$ under mutlinormal distribution.
7:          compute $\log(\hat{\mu}, \hat{\sigma}^2 | Y, \mathbb{T}, V)$;
8:      if model is OU
9:          simulate trait data $Y$ under mutlinormal distribution.
10:         transform $V_j$ into $V_{\alpha,j}$
11:         optimize the log function
12: compute the LSS statistics using formula
13: choose the largest value of $LSS$ and return the best cluster index $j^*$.
14: endfor
15: for $i = 1 : b$ do
16:     obtain sample $Y_i$ by permuting $Y$.
17:     repeat step 2 to step 10 to obtain $LSS_i$.
18: endfor
19: compare $LSS_i$ with $LSS$ and report $p$ value.
```

Table 2. quantile for the σ^2 from simulation, the true value is 1.

Taxa	10	30	50
BM	0.46(0.13,1.16)	0.64(0.35,1.06)	0.63(0.32,1.02)
OU	1.03(0.98,1.17)	1.01(1.00,1.09)	1.01(1.00,1.05)

Table 3. quantile for the α from simulation, the true value is 0.25.

Taxa	10	30	50
OU α	1.07 (0.01,4.65)	0.53(0.01,1.46)	0.57(0.01,1.80)

model. As BM model is a submodel of OU, it is likely that this phenomenon come from data is simulated from OU model with a special case of $\alpha = 0$.

Table 3 shows the median estimate and the 95% confidence interval of the p-value under the OU model for parameter α

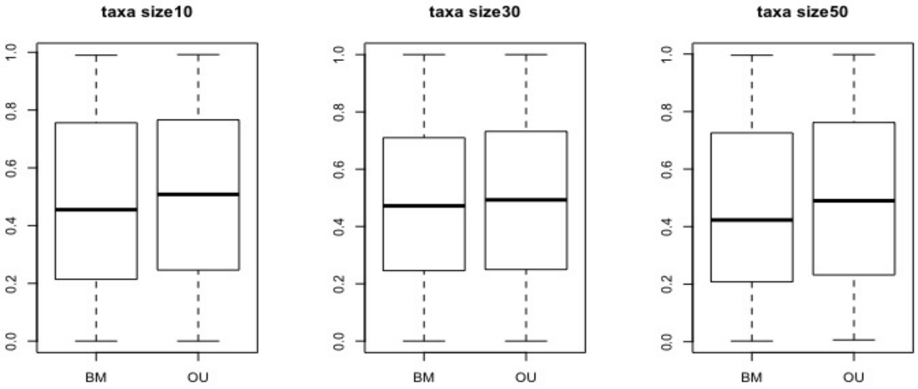


Figure 4. Asseess significance through simulation study under OU model. compare to BM, OU has higher p-value which indicates that OU model is more conservative than BM model.

Figure 4 compare the result of significance under different number of taxa for OU model. Trait data is simulated under OU model, OU has a bit higher p-value than the BM model.

Power Analysis

We access the power of the OU model. Currently we use 100 trials where for each trial a p-value is obtained using algorithm 1. The power is computed by counting the frequencies of p-value smaller than a given significant level (here we set the level to 0.1).

0.0.1 Haploid data from BM model

For the power of OU model, We look at the p-value of OU model when data are simulated from BM model. We show the boxplot in Figure 5

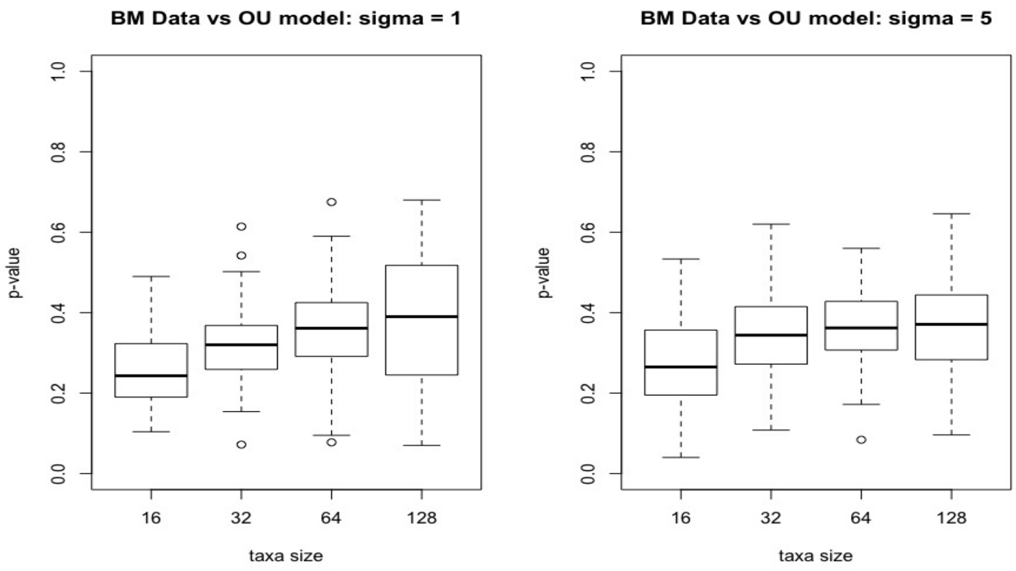


Figure 5. Data are simulated from BM model and analyzed under the OU model. The p-value increases with sample size and hence decrease the power. Overall the p-values does not change in different σ($\sigma = 1, 5$) The results is summarized using 5 trees.

The p-value increases with sample size and hence decrease the power. One possible rationale behind this plot might due to equation (13), when sample size n increases, the bound for determining the p-value shrinks which increases the number of permuted maximized LSS score that exceed this bound, hence increasing the p-value.

Haploid data from OU model

We also look at the power of BM model when data are simulated from OU model We show the box-plot in Figure 6

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27118v1 | CC BY 4.0 Open Access | rec: 15 Aug 2018, publ: 15 Aug 2018

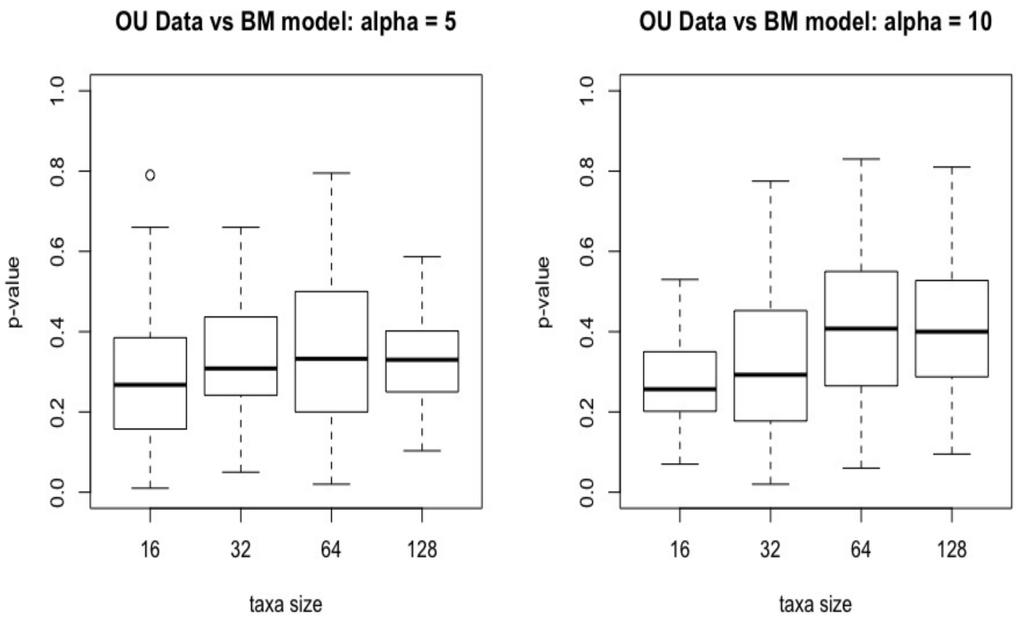


Figure 6. Data are simulated from OU model and analyzed under the BM model. The p-value increases with sample size and hence decrease the power. Overall the p-values does not change in different α ($\alpha = 5, 10$) The results is summarized using 1 tree. Result using summarizing 5 trees has similar pattern with 5 tree but slightly lower in the 128 taxa case.

Drosophila melanogaster

Fruit flies (Drosophila melanogaster) have haploid cell. In liteature, there are studies about the logenvity of fruit flies. Durham et al. (2014) identified that the senescence (a decline in physiological function in age) trait is related to the longevity of fly. They provided evidences that individuals alles influence fecundity in an age specific manner and so the genetic basis of natural variation in fecundity chanes dramatically with age. They complete a genome-wdie assicuation to indentify single-nucleotide polymorphism (SNPs) affecting lifespan and age-specific fecundity using the Drosophila melanogaster genertic Reference panel. They identified 1,031 SNPs affecting fecundity and 52 influcing lifespan. Only one SNP is aoosicated with both early and late-age fecudity. The age-speciefic effect of candidate genes on fecunity is validated using RNA interence. Their result provides support for the mutation accumulation theory of aging.

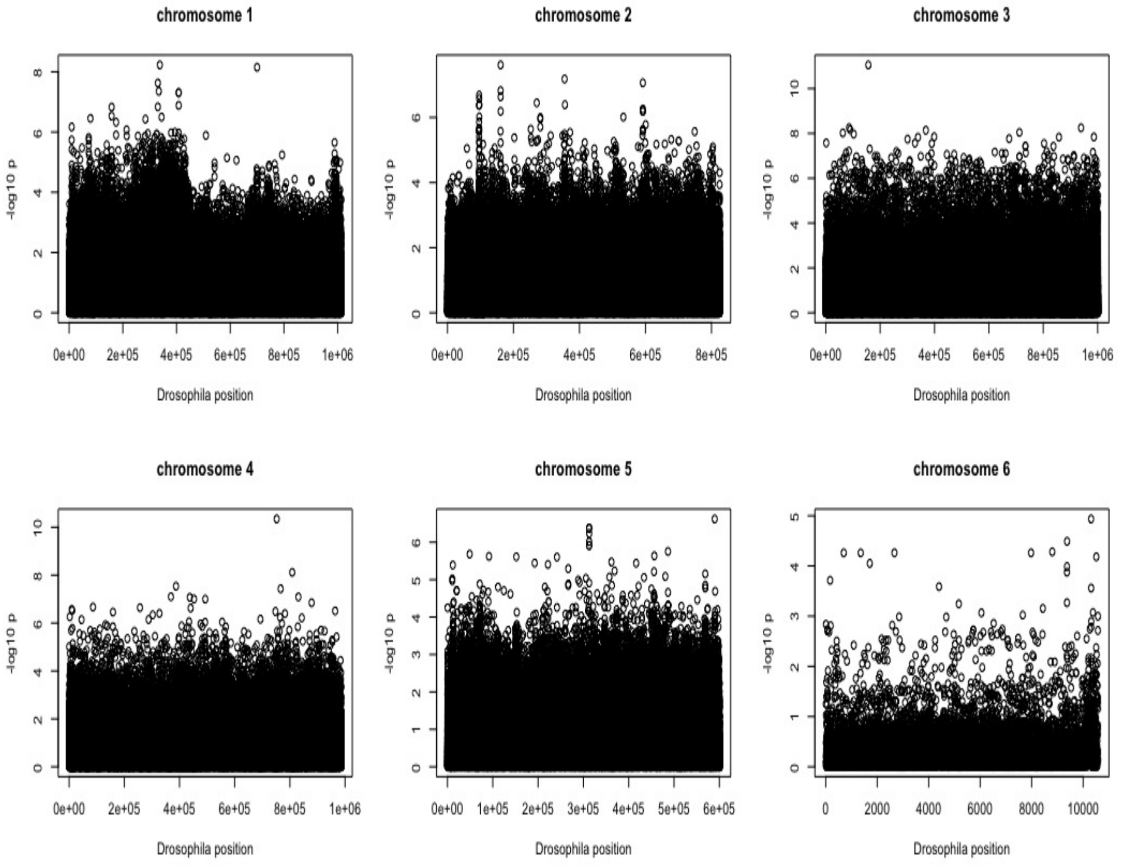


Figure 7. Genotype-phenotype association for 6 chromosomes in 205 drosophila and calculated using single-SNP linear regression, while controlling for genetic structure(tree was built under SNP dataset). $-\log_{10}(p-value) > 4$ or p-value < 0.0001 are regarded as SNPs significant.

DISCUSSION AND CONCLUSION

In this work, we extend the tree-based methods described in Thompson and Kubatko (2013) for genome-wide association study(GWAS) for the haploid case. Our method considers incorporating phylogenetic tree built under the SNP dataset and then use the tree as a dependent evidence among individuals. We then use clustering technique in order to identify any possible associations between a trait and SNP maker. To cluster tree, we consider to alter the strengths of affinity among individuals but not change the topology. To do this, we apply a Gaussian process called Ornstein-Uhlenbeck process to stretch/lengthen/shrink the branch lengths in the tree.

We evaluate the performance of our model as well as compare the existing tree-based model via through accessing their statistical power. Currently, we found that the overall statistical performance for our model is with lower powers when true data are simulated from the alternative models (data simulated from BM model). This might due to the tree is estimated from the SNP data. However, the major issue that contributes to this lower power of OU model could be the clustering procedure which changes the structure of the affinity among the individuals. Hence true data loses some information inherited from the model. In particular, this might due to the clustering k index and the matrix D transform the mean and variance among individual V which might cause the different result of estimation from the true value. For OU model, we find that α and σ^2 cannot be estimated well simultaneously.

It is possible to report the false discovery rate for both BM and OU model, in that case we can compare both models. We also can compare the model by determining the sample size at a threshold power level. Smaller size would report a higher power of the model. Finally, we hope that our model can benefit the research community in GWAS research area. While our model is planned to analyze the haploid dataset, we also wish to extend it to apply to association study in primate or human.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27118v1 | CC BY 4.0 Open Access | rec: 15 Aug 2018, publ: 15 Aug 2018

R script as well as other analysis result can be accessed at `https://github.com/djhwueng/OUsnp`.

ACKNOWLEDGMENTS

We thank Brian O'Meara, Shin-Fu Tsai, and Yuchung Wei for their helpful suggestions on improving this manuscript.

REFERENCES

Beaulieu, J., Jhwueng, D.-C., Boettiger, C., and O'Meara, B. (2012). Modeling stabilizing selection: expanding the Ornstein-Uhlenbeck model of adaptive evolution. *Evolution*, 66(8):2369–2383.

Besenbacher, S., Mailund, T., and Schierup, M. H. (2008). Local phylogeny mapping of quantitative traits: Higher accuracy and better ranking than single marker association in genomewide scans. *Genetics*.

Britton, T., Oxelman, B., Vinnersten, A., and Bremer, K. (2002). Phylogenetic dating with confidence intervals using mean path lengths. *Molecular Phylogenetics and Evolution*, 24(1):58 – 65.

Butler, M. and King, A. (2004). Phylogenetic comparative analysis: a modeling approach for adaptive evolution. *The American Naturalist*, 164:683–695.

Chifman, J. and Kubatko, L. (2014). Quartet inference from snp data under the coalescent model. *Bioinformatics*, 30(23):3317–3324.

Durham, M. F., Magwire, M. M., Stone, E. A., and Leips, J. (2014). Genome-wide analysis in drosophila reveals age-specific effects of snps on fitness traits. *Nature communications*, 5:4338.

Felsenstein, J. (1985). Phylogeny and the comparative method. *America Naturalist*, 125(1):1–15.

Hansen, T. and Martins, E. (1996). Translating between microevolutionary process and macroevolutionary patterns: the correlation structure of interspecific data. *Evolution*, 50:1404–1417.

Hudson, R. R. (2002). Generating samples under a wright fisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–338.

Hudson, R. R. (2004). ms a program for generating samples under neutral models.

McClurg, P., Pletcher, M. T., Wiltshire, T., and Su, A. I. (2006). Comparative analysis of haplotype association mapping algorithms. *BMC Bioinformatics*, 7:61.

OMeara, B., Ané, C., Sanderson, M., and Wainwright, P. (2006). Testing different rates of continuous trait evolution using likelihood. *Evolution*, 60:922–933.

Pan, F., McMillan, L., de Villena, F. P.-M., Threadgill, D., and Wang, W. (2009). Treeqa: Quantitative genome wide association mapping using local perfect phylogeny trees. *Pac Symp Biocomput*, pages 415–426. 19209719[pmid].

Schmitz, L. (2017). Visualizing comparative data in a phylogenetic framework.

Swofford, D. L. (2011). PAUP*: phylogenetic analysis using parsimony, version 4.0b10.

Thompson, K. L. and Fardo, D. W. (2016). Comparing performance of non–tree-based and tree-based association mapping methods. *BMC Proceedings*, 10(7):43.

Thompson, K. L. and Kubatko, L. S. (2013). Using ancestral information to detect and localize quantitative trait loci in genome-wide association studies. *BMC Bioinformatics*, 14(1):200.

Zhang, Z., Zhang, X., and Wang, W. (2012). Htreeqa: Using semi-perfect phylogeny trees in quantitative trait loci study on genotype data. *G3: Genes, Genomes, Genetics*, 2(2):175–189.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27118v1 | CC BY 4.0 Open Access | rec: 15 Aug 2018, publ: 15 Aug 2018