问题

- Introduction

My name is Ziyuan Wang, a fourth-year undergraduate from Sichuan University, majoring in computational biology.I have been taking plenty of courses related to biology, statistics, and programming during my three-year undergraduate study, during which I got an overall GPA of 3.92/4.00, with the rank in my major as 1/27. My research interests consist of population genetics and deep learning.

- why school

The reputation of PSU all over the world and in China is great.Penn State’s Department of Biology is one of the top-ranked biology departments in the United States. Penn State has a lot of research funding and can produce good research results.
Dr. Huang's research is attracting. He is good at finding scientific problem and using proper computational and deep learning method to solve them.

- why phd
- 群体遗传概念复习
  First of all, I am very interested in research, and I enjoy the process of solving problems. At the same time, I really want to use computational and statistical methods to solve problems related to human diseases. At the same time, my future goal is to pursue an academic career. Obtaining a doctoral degree can improve my competitiveness, and participating in a doctoral program can enable me to learn more professiodnal knowledge under the guidance of advisor.

拟南芥

The main question we want to explore is, what factors influence transposon activity.

Geographic factors, including altitude and geographical location, are known to affect the Arabidopsis genome. As shown here, despite the Arabidopsises are close to each other, genome variation can be large. The root cause is that the genomes of different arabidopsis species are located different growth conditions.

G4 is a short piece of DNA in the genome that, when folded, blocks methylation on CpG island. The role of transposons in genome changes in Arabidopsis thaliana has been reported in many studies.

However, methylation can inactivate transposons leading to genome differences between different species.

LTR is a retrotransposon that can be copied and pasted.LTRs are often the target of epigen- etic regulation, whereas retrotransposons are methylated and inactivated by the host. G4s have been observed in unmethylated regions of genomes of different kingdom before. We speculate that the presence of G4s in LTRs may be related to such inactivating mechanism, probably by interfering with the methylation process. Because G4s formed on one strand would theoretically leave the other strand in a single- stranded state, it is possible they could hinder methylation of the surrounding sequences, even if they were rich in CpG.

It has been speculated that the activity of transposable enzymes is affected by different geographical and climatic conditions. But here, we propose that different geographical and climatic environments will affect the folding of G4, and failure of G4 to fold will lead to methylation of CpG islands on both sides of LTR and inactivation of transposons.

名词

- G-Quadruplex
- retrotransposon
- transposon
- transposase
- Arabidopsis thaliana
- Coalescent theory

群体遗传

Linkage disequilibrium

D

Drift

Random change in the allel frequency from one generation to the next

群体遗传-Drift

Last section

杂合子概率与有效群体数量有关

$$H_t = (1 - \frac{1}{2N_e})^t H_0$$

$$F = \frac{H_e - H_o}{H_e}$$

H_0是初始群体杂合子频率，H_t是经过t代之后群体杂合子频率。F固定指数，实际的杂合子和HW平衡的比较。**衡量种群中基因型实际频率是否偏离遗传平衡理论比例的指标。**

有效群体数量的估计N_e

利用IBD反推N_e
近交有效群体数量 - 在一个群体中两个等位基因是来自共同祖先IBD的概率，和该概率等效的理想群体数就是近交有效群体数。
两种具体计算方式
https://www.isbreeding.net/common/UploadFiles/file/teaching/数量遗传学教学2017/第4章课件.pdf

距离隔离
随着空间距离的增加，交配概率或者配子扩散数量降低。

基因系谱和溯祖模型 Coalescent theory

溯祖：根据当前的群体样本逆推过去群体中发生的事件，直到找到一个共同的祖先。
最近共同祖先 **Most Recent Common Ancester**：在系谱中，对当前样本溯祖，第一个出现的共同的祖先，即最近共同祖先。

基本假设:N_e不变
$2N_e$单倍体群体中，两个单倍体来自于父代服从几何分布，当有效群体数较大时服从指数分布
对于多个系谱的溯祖，系谱越多，两两溯祖发生的概率越大，等待时间越短。
溯祖树高：从当下到k个系谱找到它们的最近共同祖先所需要的时间。树高平均为2N代-4N代，当k=2时，所需时间最短，为2N，随着k的增加，所需时间增长，最长为4N。

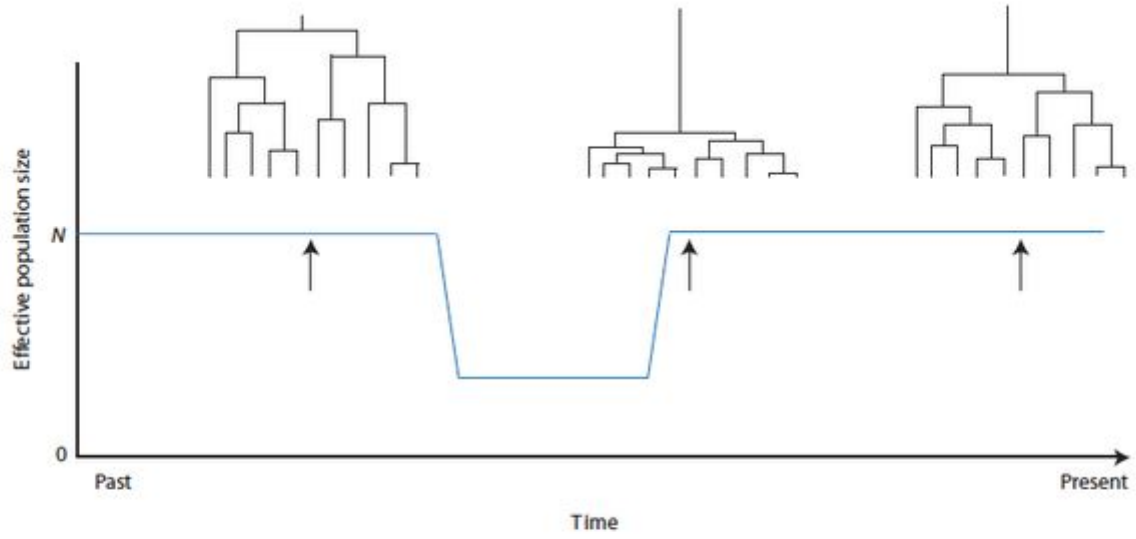

Figure 3.28 The effects of a population bottleneck on gene genealogies. During the bottleneck the chance that two randomly sampled gene copies are derived from one copy in the previous generation $\left(\frac{1}{2N_e}\right)$ increases. This can also be thought of as a reduction in the overall height of a genealogical tree caused by the bottleneck since lineages that find their ancestors during the bottleneck lead to short branches. The overall effect of a bottleneck on coalescence among gene copies sampled in the present depends on the reduction in the effective population size and the duration. The arrows indicate the point in time when gene copies were sampled from the population.

瓶颈时间和溯祖 coalescence

在群体经历了瓶颈事件时，群体中的各个系谱比瓶颈前或后更容易找到共同祖先，所以在瓶颈的溯祖时间变短。
主要原因 N_e 变小

扩张群体：越靠近当下时间，群体数量越大，溯祖时间越长；反之，随着时间回溯，有效群体数量变小，溯祖时间变短。

收缩群体：越靠近当下，群体数量越小，溯祖时间越短；反之，随着时间回溯，有效群体数量扩大，溯祖所需时间越长。

瓶颈事件后种群成为收缩群体

neutual Test

neutual reference locus
HKA use 5' psudogene

MK 同义突变synonynous和非同义突变nonsynonymous

> D<0：群体有过多的低频位点 -> 近期定向选择或者群体扩张；
> D>0：群体有过多的中等频率位点 -> 经历平衡选择或者群体收缩。