

Zhang-Lab 生信小课堂 第廿三期

Applied Bioinformatics Club (ABC)

和趣求真  秉实生信

(张建伟生物信息学课题组 <https://zhang.hzau.edu.cn>)

群体遗传学基础分析： π 值与Tajima's D的计算与解读

2025.10.10 二综C102 15:00 欢迎大家交流学习!

主讲人：丁悦
2025/10/10

目录

- 01 核心概念与理论基础
- 02 π 值计算与解读
- 03 Tajima's D值计算与解读
- 04 综合评估

核心概念与理论基础

- 遗传变异：指一个物种群体内，不同个体在基因组特定位点上的DNA序列差异，是群体遗传分析的基本单元。最常见的类型是单核苷酸多态性。
- 中性理论：是分子进化的标准零假设。该理论认为，在分子水平上，绝大多数观察到的遗传变异是中性或近中性的，其命运主要由随机遗传漂变决定，而非自然选择。
- 自然选择：基于表型适应度的非随机过程，会定向改变等位基因频率，在基因组上留下特征性的统计模式，即选择信号。
- 选择信号：基因组中因自然选择作用而导致其遗传变异模式（如多样性水平、等位基因频率分布等）显著偏离中性理论预期的区域。

群体遗传学通过量化种群内的遗传变异模式，来推断塑造这些模式的进化动力。

π 值计算与解读 - π 值含义

PI值：（ π 值，核苷酸多样性）反映群体中任意两个随机选择的序列间平均核苷酸差异。

计算公式：

$$\pi = \frac{\sum_{i < j} d_{ij}}{\binom{n}{2} \times L}$$

- d_{ij} : 第*i*条和第*j*条序列之间的核苷酸差异数。
- n : 群体中的序列数（样本量）。
- L : 分析的序列长度（排除缺失或不可比位点）。
- $(n/2)$: 序列对的总数，即 $n(n-1)/2$

π 值计算与解读 - π 值计算

PI值计算:

输入数据: 533_fixed.vcf (VCF文件, 包含所有样本的基因型信息)

工具: `vcftools`

代码示例:

```
1 bsub -J vcftools -n 10 -R span[hosts=1] -o %J.out -e %J.err -q normal "vcftools --vcf 533_fixed.vcf --window-pi 100000 --window-pi-step 100 --out 533
```

分亚群计算PI值

输入数据: cA.txt cB.txt GJ-tmp.txt GJ-trop.txt 533_fixed.vcf

```
1 #!/bin
2 sample_files=(../subgroup2/*.txt)
3 for sample in "${sample_files[@]"; do
4     new_name1=$(echo "$sample" | awk -F '.' '{print $3}') # 添加前缀并保持扩展名不变
5     new_name2=$(echo "$new_name1" | awk -F '/' '{print $3}')
6     output="./subgroup2/pi_window_${new_name2}"
7     vcftools --vcf 533_fixed.vcf --window-pi 100000 --window-pi-step 100 --keep "$sample" --out "$output"
8 done
```

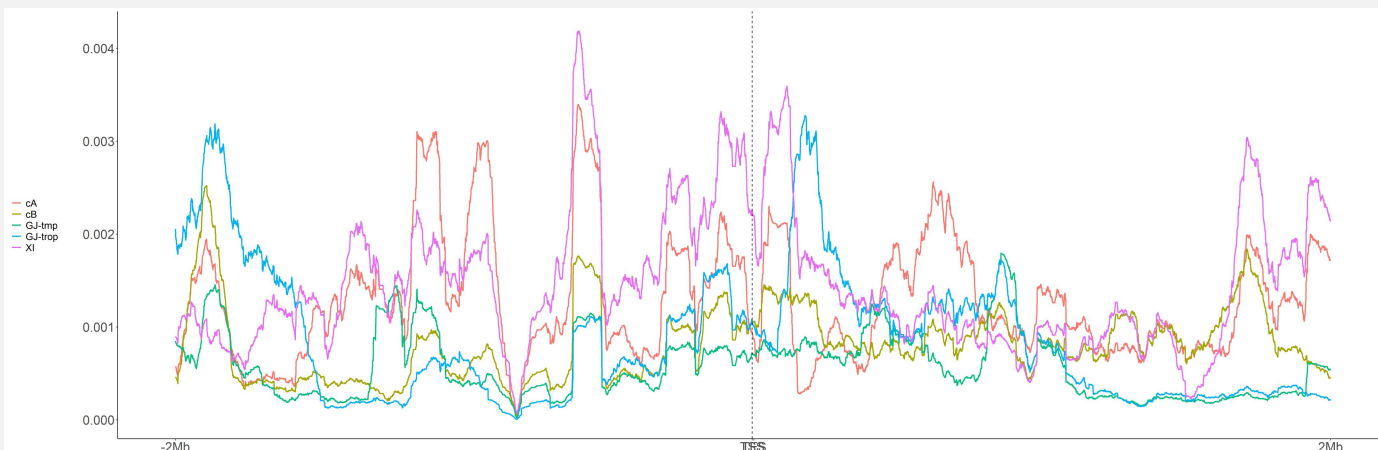
参数控制: Windows设置为100kb; 步长为100

比对结果: pi_window_*.windowed.pi

1	CHROM	BIN_START	BIN_END	N_VARIANTS	PI
2	1	1	100000	917	0.00255472
3	1	101	100100	918	0.00255666
4	1	201	100200	918	0.00255666
5	1	301	100300	918	0.00255666
6	1	401	100400	918	0.00255666
7	1	501	100500	918	0.00255666
8	1	601	100600	918	0.00255666
9	1	701	100700	918	0.00255666
10	1	801	100800	918	0.00255666
11	1	901	100900	918	0.00255666

π 值计算与解读-可视化展示与结果分析

Rscript plot-gene-2Mb.r



*基因上下2Mb受选择情况

首先，关注整个群体或染色体水平的平均 π 值

✓ 高 π 值 (例如 > 0.01): 表明群体遗传多样性高

✓ 低 π 值 (例如 < 0.001): 表明群体遗传多样性低

```
1 library(ggplot2)
2 windowsFonts(A=windowsFont("Times New Roman"), B=windowsFont("Arial"))
3
4 # tajima_d_data <- read.table("rice_core_TajimaD.Tajima.D", header = TRUE, sep = "\t")
5 # pi_data <- read.table("rice_core_pi.windowed.pi", header = TRUE, sep = "\t")
6 # fst_data<-read.table("GJ_XI_fst.txt", header=TRUE, sep = "\t")
7
8 Total_data<-read.table("combined_PI.txt", header=TRUE, sep = "\t")
9
10
11 Total_data <- Total_data[Total_data$CHROM=="2", ] 基因位于染色体号
12
13 #Total_data$BIN_START <- Total_data$BIN_START / 100000
14 Total_data$BIN_START <- Total_data$BIN_START / 100
15 Total_data <- Total_data[Total_data$BIN_START>=323065.14 & Total_data$BIN_START <= 363133.19, ]
16
17 基因的区间
18 ggplot(Total_data, aes(x = BIN_START, y = Value, color = TYPE)) +
19   # geom_rect(aes(xmin=100, xmax=300, ymin=0, ymax=60,col = "lightgrey",fill = "lightgrey")) + #
20   geom_line(size=2.5) +
21   theme_classic() +
22   # theme_ridges(grid = FALSE)+
23   xlab("LOC_0s01g59350") + 基因名称
24
25
26 ylab("")+
27 scale_color_manual(values = c("#ff5722", "#f8b500", "#17b978", "#07689f", "#A079BF")) +
28 #scale_color_manual(values = c("#FD763F", "#238B45")) +
29 # facet_grid(TYPE~., scales = "free_y") +
30
31 theme(legend.position = "left",
32       axis.line = element_line(color = "black", linewidth = 0.5),
33       legend.title=element_blank(), ## 图例位置 (可选: bottom/right/left/top)
34       axis.text = element_text(size = 20),
35       axis.title = element_text(size = 20),
36       strip.text = element_text(size = 20),
37       # axis.ticks.x = element_blank(),
38       legend.text=element_text(size = 15),
39       strip.text.y = element_text(size = 12, face = "bold"),)+
40   # panel.border = element_blank(),
41
42
43 scale_x_continuous(breaks = c(323065.14, 343065.14, 343133.19, 363133.19),
44 #scale_x_continuous(breaks = c(34.306514), # 18414297 18415361
45 labels = c("-2Mb", "TSS", "TES", "2Mb" )) +
46 #labels = c("LOC_0s02g30850")) +
47
48 # ylim(0,1)
49 geom_vline(xintercept = 343065.14, col = "black", lty = 2)
50 geom_vline(xintercept = 343133.19, col = "black", lty = 2)
51 ggsave("Chr2_2.5.png", width = 30, height = 10)
52
```

起始终止以及基因的位置

Tajima's D值计算与解读 - Tajima's D值含义

Tajima's D: 一种用于检测群体遗传数据是否符合中性进化假说的统计量，通过比较两种基于突变率的遗传多样性估计值（ π 值和Watterson's θ ）的差异，揭示自然选择或群体历史事件（如扩张、瓶颈）的信号。

计算公式: $Tajima\ D = \theta\pi - \theta_s$

$\theta\pi$: 群体序列两两比较差异位点数累加/总两两比较对数，对中等频率变异敏感

θ_s : 是群体总变异位点数/序列数的倒数累加，基于分离位点数，对低频变异敏感

Tajima's D值计算与解读 - Tajima's D值计算

输入数据: 533_fixed.vcf (VCF文件, 包含所有样本的基因型信息)

工具: `vcftools`

命令:

```
1 bsub -J vcftools -n 10 -R span[hosts=1] -o %J.out -e %J.err -q normal "vcftools --vcf 533_fixed.vcf --TajimaD 100000 --out TajimaD_output"
```

分亚群计算Tajima's D值

输入数据: cA.txt cB.txt GJ-tmp.txt GJ-trop.txt 533_fixed.vcf

```
1 #!/bin
2 sample_files=(../subgroup2/*.txt)
3 for sample in "${sample_files[@]"; do
4     new_name1=$(echo "$sample" | awk -F '.' '{print $3}') # 添加前缀并保持扩展名不变
5     new_name2=$(echo "$new_name1" | awk -F '/' '{print $3}')
6     output="../subgroup2/TajimaD_${new_name2}"
7     vcftools --vcf 533_fixed.vcf --TajimaD 100000 --keep "$sample" --out "$output" done
```

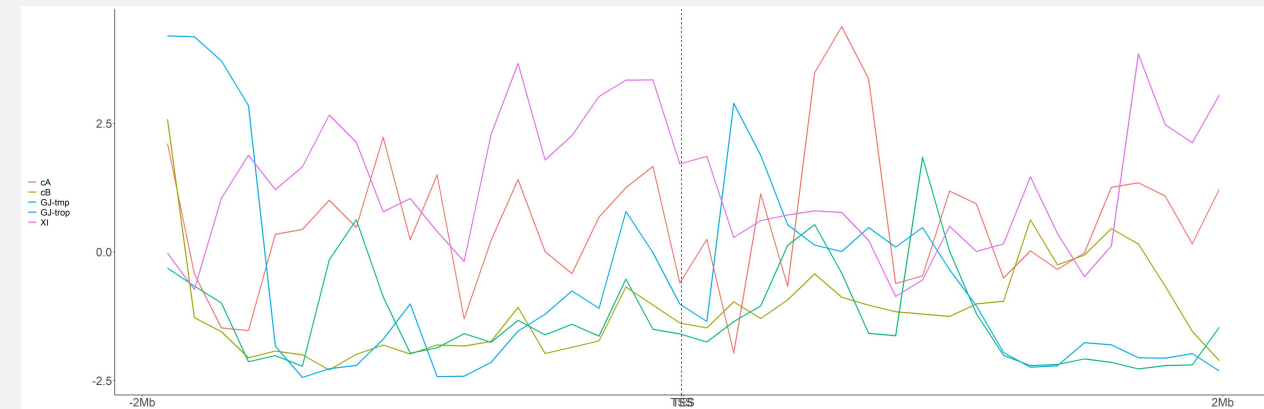
参数控制: Windows设置为100kb; 步长为100

比对结果: TajimaD_*.D

	CHROM	BIN	START	N_SNPS	TajimaD
1	1	0	804	2.84306	
2	1	100000	660	2.95579	
3	1	200000	789	3.13731	
4	1	300000	674	3.74606	
5	1	400000	995	3.0541	
6	1	500000	893	0.121218	
7	1	600000	916	0.187686	
8	1	700000	738	2.31137	
9	1	800000	533	1.45462	
10	1	900000	786	2.02726	
11	1	1000000	1118	1.89235	
12	1	1100000	824	0.467203	
13	1	1200000	443	-0.0817554	

Tajima's D值计算与解读 - 可视化展示与结果解读

Rscript plot-gene-2Mb.r



基因上下2Mb受选择情况

计算值分析:

D=0: 数据符合中性进化模型 (π 与 θW 无显著差异)。

D>0: 可能表明平衡选择或群体收缩 (低频等位基因较少)。

D<0: 可能提示定向选择 (正选择) 或群体扩张 (近期种群增长)。

```
1 library(ggplot2)
2 windowsFonts(A=windowsFont("Times New Roman"), B=windowsFont("Arial"))
3
4 # tajima_d_data <- read.table("rice_core_TajimaD.Tajima.D", header = TRUE, sep = "\t")
5 # pi_data <- read.table("rice_core_pi.windowed.pi", header = TRUE, sep = "\t")
6 # fst_data<-read.table("GJ_XI_fst.txt", header=TRUE, sep = "\t")
7
8 Total_data<-read.table("combined_TajimaD.txt", header=TRUE, sep = "\t")
9
10
11 Total_data <- Total_data[Total_data$CHROM=="2", ] # 基因位于染色体号
12
13 Total_data$BIN_START <- Total_data$BIN_START / 100
14 Total_data <- Total_data[Total_data$BIN_START>=164142.97 & Total_data$BIN_START <= 204153.61, ] # 基因的区域
15
16 ggplot(Total_data, aes(x = BIN_START, y = Value, color = TYPE)) + # 基因名称
17   # geom_rect(aes(xmin=100, xmax=300, ymin=0, ymax=60,col = "lightgrey",fill = "lightgrey")) + #
18   geom_line(size=2.5) +
19   theme_classic() +
20   # theme_ridges(grid = FALSE)+
21   # xlab("LOC_Os01g59350") +
22   xlab("LOC_Os02g30850") + # 绘制线的颜色
23
24   vlab("")+
25   scale_color_manual(values = c("#ff5722", "#f8b500", "#17b978", "#07689f", "#A079BF")) + # 起始终止以及基因的位置
26   #scale_color_manual(values = c("#FD/63F", "#23BAC5")) +
27   # facet_grid(TYPE~., scales = "free_y") +
28
29   theme(legend.position = "left",
30         axis.line = element_line(color = "black", linewidth = 0.5),
31         legend.title=element_blank(), ## 图例位置 (可选: bottom/right/left/top)
32         axis.text = element_text(size = 20),
33         axis.title = element_text(size = 20),
34         strip.text = element_text(size = 20),
35         # axis.ticks.x = element_blank(),
36         legend.text=element_text(size = 15),
37         strip.text.y = element_text(size = 12, face = "bold"),)+
38   # panel.border = element_blank(),
39
40   scale_x_continuous(breaks = c(164142.97,184142.97,184153.61,204153.61),
41
42                       labels = c("-2Mb", "TSS", "TES", "2Mb" )) +
43                       #labels = c("LOC_Os02g30850")) +
44
45   # ylim(0,1)
46   geom_vline(xintercept = 184142.97, col = "black", lty = 2)
47   geom_vline(xintercept = 184153.61, col = "black", lty = 2)
48   ggsave("Chr01.png", width = 30, height = 10)
49
```

核苷酸多样性 (π) 定量评估了群体内的遗传变异数量，而 Tajima's D 则描述了其等位基因频率频谱的特征。将这两种统计量相结合，是识别偏离中性进化模式、从而推断自然选择作用的关键策略。

- 低 π 值 + 显著负的 Tajima's D: 该区域极有可能包含一个经历**强烈正选择的适应性基因**。
- 正常/高 π 值 + 显著正的 Tajima's D: **平衡选择**，该基因可能参与诸如病原体防御、自交不亲和等过程，其多样性本身具有适应意义。
- 低 π 值 + Tajima's D ≈ 0 或 轻微为正: **种群规模小或经历过瓶颈效应，无强烈局部选择**。
- 正常 π 值 + Tajima's D ≈ 0 : **中性进化——基因组的大部分常态**。
- 低 π 值 + 显著正的 Tajima's D (罕见但存在): **长期平衡选择下的局部变异清除**，在长期维持两个古老等位基因的同时，基因转换等机制清除了它们之间的重组区域变异，导致局部 π 值降低，但两个主效等位基因本身仍保持中等频率，从而产生正的D值。

Q&A

汇报人：丁悦