经典群体遗传学计算方法

Luolintao

giantlinlinlin@gmail.com

代码对应的计算方法

Methods (English)

Nucleotide diversity (π)

For each population, nucleotide diversity was computed as the mean number of pairwise differences per site using allele counts across variant positions:

$$\pi = \frac{1}{L} \sum_{i=1}^{L} 2p_i(1 - p_i)$$

where L is the number of sites and p_i is the alternative allele frequency at site i. This corresponds to `allel.sequence_diversity`.

Watterson's theta (θ_w)

Watterson's theta was computed from the number of segregating sites S and sample size n:

$$heta_w = \frac{S}{a_1}, \quad a_1 = \sum_{k=1}^{n-1} \frac{1}{k}$$

as implemented by `allel.watterson_theta`.

Tajima's D

Tajima's D compares π and θ_w:

$$D = \frac{\pi - \theta_w}{\sqrt{V}},$$

where V is the variance term defined by Tajima (1989) and depends on n and S. We used `allel.tajima_d` with positions to compute D.

Site Frequency Spectrum (SFS)

For each population, segregating biallelic sites were used to compute the folded and unfolded SFS.

- **Unfolded SFS:** using derived allele counts d_i at each site:

$$\mathrm{SFS}(k) = \#\{i : d_i = k\}, \quad k = 1, \ldots, n-1$$

  via `allel.sfs`.
- **Folded SFS:** minor allele counts $m_i = \min(d_i, n - d_i)$:

$$\mathrm{SFS}_{\mathrm{fold}}(k) = \#\{i : m_i = k\}, \quad k = 1, \ldots, \lfloor n/2 \rfloor$$

  via `allel.sfs_folded`.

Genetic differentiation (F_{ST}, Weir & Cockerham)

Pairwise F_{ST} was computed using the Weir & Cockerham estimator. For each pair of populations, per-site components (a, b, c) were computed and aggregated as:

$$F_{ST} = \frac{\sum a}{\sum (a + b + c)}$$

Values were truncated to $[0, 1]$. This corresponds to `allel.weir_cockerham_fst` with the above aggregation.

Sequence divergence (D_{xy})

Between-population sequence divergence was computed as the mean number of pairwise differences per site between two populations:

$$D_{xy} = \frac{1}{L} \sum_{i=1}^{L} (p_{1i}(1 - p_{2i}) + p_{2i}(1 - p_{1i}))$$

using `allel.sequence_divergence`.

Significance testing

When enabled, two resampling-based procedures were applied.

Permutation test for F_{ST} and D_{xy}

For each pair of populations, sample labels were permuted N times. The empirical one-sided p-value was computed as:

$$p = \frac{1 + \sum_{k=1}^{N} \mathbb{I}(T_k \geq T_{\mathrm{obs}})}{N + 1}$$

where T is F_{ST} or D_{xy}.

Bootstrap for π and θ_w

We bootstrapped sites with replacement to generate N resampled datasets and recalculated π and θ_w. For each pair of populations, we computed the difference distribution and derived the

95% confidence interval (2.5–97.5%) and a two-sided bootstrap p-value:

$$p = \frac{1 + \sum_{k=1}^{N} \mathbb{I}(|\Delta_k| \geq |\Delta_{\text{obs}}|)}{N + 1}$$

where Δ is the difference between populations for the metric.

方法（中文）

核苷酸多样性（π）

群体的核苷酸多样性定义为每个位点的平均成对差异数：

$$\pi = \frac{1}{L} \sum_{i=1}^{L} 2p_i(1 - p_i)$$

其中 L 为位点数，p_i 为第 i 位点的替代等位基因频率。对应实现为 `allel.sequence_diversity`。

Watterson's theta（θ_w）

Watterson's theta 根据分离位点数 S 与样本数 n 计算：

$$heta_w = \frac{S}{a_1}, \quad a_1 = \sum_{k=1}^{n-1} \frac{1}{k}$$

对应实现为 `allel.watterson_theta`。

Tajima's D

Tajima's D 用于比较 π 与 θ_w 的差异：

$$D = \frac{\pi - \theta_w}{\sqrt{V}},$$

其中 V 为 Tajima (1989) 所定义的方差项，与 n 与 S 有关。计算采用 `allel.tajima_d`。

频率谱（SFS）

对每个群体，仅保留分离且二等位位点，计算 folded 与 unfolded SFS。

- **Unfolded SFS：** 使用导出等位基因计数 d_i：

  $$\text{SFS}(k) = \#\{i : d_i = k\}, \quad k = 1, \ldots, n-1$$

  通过 `allel.sfs` 实现。
- **Folded SFS：** 使用次等位基因计数 $m_i = \min(d_i, n - d_i)$：

  $$\text{SFS}_{\text{fold}}(k) = \#\{i : m_i = k\}, \quad k = 1, \ldots, \lfloor n/2 \rfloor$$

  通过 `allel.sfs_folded` 实现。

群体分化（F_{ST}，Weir & Cockerham）

对每对群体，使用 Weir & Cockerham 估计量。先计算每个位点的组分 (a, b, c)，再汇总为：

$$F_{ST} = \frac{\sum a}{\sum (a + b + c)}$$

结果截断到 $[0, 1]$。对应实现为 `allel.weir_cockerham_fst`。

群体间序列分歧（D_{xy}）

群体间序列分歧定义为两群体间平均成对差异：

$$D_{xy} = \frac{1}{L} \sum_{i=1}^{L} (p_{1i}(1 - p_{2i}) + p_{2i}(1 - p_{1i}))$$

对应实现为 `allel.sequence_divergence`。

显著性检验

启用显著性分析时，执行两类重采样。

F_{ST} 与 D_{xy} 的置换检验

对每对群体随机置换样本标签 N 次，计算经验单侧 p 值：

$$p = \frac{1 + \sum_{k=1}^{N} \mathbb{I}(T_k \geq T_{\text{obs}})}{N + 1}$$

其中 T 为 F_{ST} 或 D_{xy}。

π 与 θ_w 的 bootstrap

对位点进行有放回抽样，重复 N 次计算 π 与 θ_w。对任意两群体的差异 Δ，计算 95% 置信区间（2.5–97.5% 分位）和双侧 p 值：

$$p = \frac{1 + \sum_{k=1}^{N} \mathbb{I}(|\Delta_k| \geq |\Delta_{\text{obs}}|)}{N + 1}$$