群体历史推断方法

The Methods for Inferring Population History

刘俊锋 [1,2]，季现超 [1,2,3]，陈华 [1,2,3*]

1. 精准基因组医学重点实验室，中国科学院北京基因组研究所，北京 100101
2. 国家生物信息中心，北京 100101
3. 中国科学院大学，北京 100049

*通讯作者邮箱：chenh@big.ac.cn

引用格式：刘俊锋，季现超，陈华. (2021). 群体历史推断方法. Bio-101 e1010604. Doi: 10.21769/BioProtoc. 1010604.

How to cite: Liu, J. F., Ji, X. C. and Chen, H. (2021). The Methods for Inferring Population History. Bio-101 e1010604. Doi: 10.21769/BioProtoc.1010604. (in Chinese)

摘要：群体历史 (population history) 包括种群瓶颈 (population bottleneck)、扩张 (population expansion)、迁移 (migration) 和混合 (admixture) 等事件，这些事件对群体的遗传多态模式 (polymorphic pattern) 的形成具有重要的影响。目前，已开发了很多基于基因组数据推断群体历史的方法。本实验方案对当前主流方法做了比较分析并给出推断群体历史的基本流程，进而将有助于实验生物学家开展群体遗传领域的相关研究。

关键词: 群体历史；位点频谱；连锁不平衡；扩散过程；溯祖过程

一、概述

群体历史 (population history) 包括种群瓶颈 (population bottleneck)、扩张 (population expansion)、迁移 (migration) 和混合 (admixture) 等事件，这些事件对群体的遗传多态模式 (polymorphic pattern) 的形成具有重要的影响。在较大群体的遗传过程中，对特定基因的随机选择使得群体中相应等位基因的频率相对稳定；而在较小群体的遗传过程中，对特定基因的随机选择则会使得群体中相应等位基因的频率发生明显波动，出现某些等位基因在群体中固定而有些等位基因消失的情形。在遗传过程中，这些造成等位基因频率变化的偶然事件被称为遗传漂变 (genetic drift)。种群瓶颈事件可

视为一种遗传漂变，该事件使得群体数量骤减，进而造成某些等位基因的频率显著升高。在遗传过程中，若特定基因受到自然选择 (natural selection) 作用，则相应等位基因的频率也会明显波动：有些等位基因的频率显著升高，而有些等位基因的频率显著降低，所以很容易混淆遗传漂变和自然选择在基因组上产生的效应。Stajich 和 Hahn (Stajich and Hahn, 2005) 对来自欧洲裔美国人和非洲裔美国人的 151 个位点进行分析，尝试识别出观测到的效应是来自遗传漂变还是源于自然选择。他们发现非洲裔美国人具有更高水平的遗传多态性，并基于溯祖模拟认为这种差异是由于混合事件和种群瓶颈事件造成的。基于混合和种群瓶颈模型，他们发现仍有些位点的相应等位基因的频率明显偏离中性理论 (neutral theory)，故认为该效应源于自然选择。 因此，在分子群体遗传研究中，推断群体历史对理解种群的演化十分重要，有助于分析观测到的群体基因组在自然选择和遗传漂变等进化力 (evolutionary forces) 作用下是如何形成的。

推断群体历史主要包括群体的拓扑结构、分化时间和有效种群大小三部分内容的推断。统计量 F_{st} 表示群体中基因型频率同哈代-温伯格 (Hardy-Weinberg) 平衡下预期基因型频率的偏离程度，用于刻画种群的分化程度。当种群出现分化时，则群体中的基因型频率将偏离哈代-温伯格 (Hardy-Weinberg) 平衡 (即纯合子的频率增加，杂合子的频率降低)，且 F_{st} 的值越大，种群分化的程度就越高。随着种群分化的加剧，子群体逐步形成新群体，通过计算新群体之间的遗传距离 (genetic distance) 可以估计分化时间。群体的有效种群大小 (effective population size, N_e) 是群体中能够有效参与繁殖的个体的数目，该数目通常远小于群体中实际个体数目。群体中的个体数目对群体的演化有着重要影响，有效种群大小越大，群体的遗传多态性 (polymorphism) 就越高，反之遗传多态性越低。因此，群体的遗传多态性信息可以用来推断有效种群大小。Atkinson 等人 (Atkinson *et al.*, 2008) 利用来自全球 357 个人类线粒体 DNA (mitochondrial DNA, mtDNA) 样本对不同区域的群体历史进行了推断。他们发现 14.3 万年前撒哈拉沙漠以南的非洲大陆群体的有效种群大小开始逐步增长，而南亚群体的有效种群大小在 5.2 万年前出现快速增长；欧洲群体的有效种群大小则出现过两次快速增长，分别发生在 4.2 万年前和 1.0 万年前。他们的研究表明人类线粒体 DNA 的多态性蕴含了人类群体历史的相关信息。

目前，已开发了很多基于基因组数据推断群体历史的方法。在早期的方法中，多数方法利用线粒体 DNA 或 Y 染色体的非重组区 (the non-recombining region) 共有单一

基因谱系 (a single gene genealogy) 这一特性构建群体历史。例如，Cann 等人 (Cann *et al.,* 1987) 分析了来自现代人类 5 个群体的 147 个个体的线粒体 DNA，推断现代人类的共同祖先来自大约 20 万年前的一位非洲女性；Underhill 等人 (Underhill *et al.,* 2001) 通过构建 Y 染色体非重组区的谱系关系探讨了现代人类的起源、分化和种群历史。这类方法已被广泛应用于东亚群体历史的研究 (Jin and Su, 2001; Ke *et al.,* 2001; Yao *et al.,* 2002; Kong *et al.,* 2003)。不过，Underhill 和 Kivisild (Underhill and Kivisild, 2007) 指出尽管利用线粒体和 Y 染色体数据较好的分析了人类走出非洲后的群体历史，但是因线粒体和 Y 染色体数据的遗传多态性较低，则无法利用这些数据重构人类走出非洲之前的群体历史。事实上，由于重组的缘故，更多的遗传多态性存储在常染色体区域。因而，当前大多数方法都是通过分析常染色体区域上的遗传多态性推断群体历史，并可大致分为两类。第一类方法基于位点频谱 (site frequency spectrum, SFS)，第二类方法基于连锁不平衡 (linkage disequilibrium, LD)。在群体遗传学研究中，SFS 为基因组中任何随机多态性位点上突变基因频率的采样分布：在一个种群的有限随机样本的所有基因座中，根据衍生等位基因 (derived allele，即突变的等位基因) 的频率对多态性基因座进行分类，并统计落入每个可能频率类别的基因座的比例或数量。如果对基因座上的等位基因无法判别是衍生等位基因还是祖先等位基因 (ancestral allele)，则采用每个基因座上低频等位基因构造 SFS。由衍生等位基因构造的 SFS 称为展开 SFS (unfolded SFS)，由低频等位基因构造的 SFS 则称为折叠 SFS (folded SFS)。SFS 的分布一般近似指数分布，即罕见基因 (rare allele) 的比例很高。若群体近期发生瓶颈事件，则罕见基因的比例将显著下降。因此，基于群体的 SFS 分布信息，通过模拟手段可以推断群体历史。与单个群体的 SFS 相比，联合等位基因频谱可用于推断更复杂的群体历史，包括群体分化和群体混合事件。SFS 的理论和方法是在两类框架下并行发展：扩散近似和溯祖过程。溯祖理论由一套概率模型组成，主要研究采集的样本在回溯过程中找到共同祖先事件的分布。在溯祖理论框架下，可以对观测的遗传数据进行参数估计和假设检验等统计推断。随着测序技术的不断进步，用于统计推断的遗传数据不断增加，进而涌现出许多基于溯祖理论的统计推断方法。溯祖理论广泛应用于基于连锁不平衡推断群体历史的方法中，主要分为两种模型：血源同一 (Identity by Descent, IBD) 模型和溯祖-隐马尔可夫模型 (coalescent-hidden Markov model, coalescent-HMM)。血源同一模型基于溯祖理论通过分析 IBD 的分布推断群体历史；溯祖-隐马尔可夫模型则基于溯祖理论，利用 HMM 框

架刻画观测到的序列和观测不到的谱系 (genealogy)，进而推断群体历史。

二、基于 SFS 的方法

1. 基于扩散过程 (the diffusion) 推断群体历史

该方法基于的模型是在时间 t 包含$N(t)$个个体的雌雄同体的随机交配二倍体群体，并假设有大量相同且独立的基因座。在每个基因座上，只有两个等位基因 A (突变的等位基因) 和一个 a (祖先等位基因)。在第 t 代，位点的集合使用一个行向量来描述，这个行向量的第 j 个元素$f_j(t)$代表在 j 个染色体上发现 A 的基因座的预期数目，$1 \leq j \leq 2N(t)$。该模型假设为 a 固定的基因座库很大，以致可以假定通过突变创建多态性基因座不会减少该基因座。由遗传漂移和突变引起的$f_j(t)$的变化通过一组差分方程描述：

$$f_j(t+1) = \sum_{i=1}^{2N(t)} f_i(t)p_{ij}(t) + M_j(t), 1 \leq j \leq 2N(t+1)$$

公式右边的第一项表示遗传漂变和自然选择对已经多态的基因座的综合影响，$p_{ij}(t)$是在 t 代中具有 i 个 A 副本的基因座在 t + 1 代中具有 j 个副本的概率。公式右边的第二项表示通过突变和迁移创建新的多态位点。通过假设单态位点的 a 的$2N(t)$拷贝中的每一个以每代概率 μ 突变为 A，来模拟突变的加入，为了有效地模拟等位基因频谱 (allele frequency spectrum, AFS) 而采用了扩散方法。扩散方法是对离散世代中进化的离散个体的群体遗传学的连续近似。一个重要的基本假设是等位基因频率的每代变化很小。因此，当有效群体数量 N_e 大且迁移率和选择系数为$1/N$时，适用扩散近似。如果我们有 P 个群体的样本，则每个群体的抽样数目为$n_1, n_2, \ldots, n_p$ (对于二倍体，n_i通常是从第 i 个群体采样的个体数量的两倍)，用$d_1, d_2, \ldots, d_p$记录在每个群体样本中突变的等位基因的数目 (d_i 表示第 i 个群体样本中突变的等位基因的数目)。对在时间 t，群体1,2, …, P中相对频率$x_1, x_2, \ldots, x_p$的突变的密度函数$\phi(x_1, x_2, \ldots, x_p, t)$进行建模。给定一个无限位点的突变模型并在每一代中进行 Wright-Fisher 复制，任意线性有限种群 P 的动力学由线性扩散方程控制：

$$\frac{\partial}{\partial \tau}\phi = \frac{1}{2}\sum_{i=1,2,\ldots,P}\frac{\partial^2}{\partial x_i^2}\frac{x_i(1-x_i)}{v_i}\phi - \sum_{i=1,2,\ldots,P}\frac{\partial}{\partial x_i}\left(\gamma_i x_i(1-x_i) + \sum_{j=1,2,\ldots,P}\frac{\partial}{\partial x_i}M_{i\leftarrow j}(x_j - x_i)\right)\phi$$

其中公式右侧第一项模拟遗传漂变，第二项模拟选择和迁移。Gutenkunst 等人 (Gutenkunst *et al.*, 2009) 基于上述方法开发了可以分析多个群体历史的软件 (DaDi)，并分析来自三个人类群体的 219 个常染色体基因非编码区 (non-coding) DNA 序列 (长度 5 Mb) 的样本数据，推断了人类走出非洲后的群体历史。由于 DaDi 的计算复杂度随着群体数目的增加而呈指数增长，所以 DaDi 所能分析的群体数目不超过 3 个且每个群体不超过 20 个样本。尽管 DaDi 考虑了自然选择的影响，进而可以分析非同义 SNP 数据；但是，非编码区 DNA 序列更适合群体历史推断。此外，DaDi 没有考虑测序误差和位点间的连锁 (linkage) 效应的影响，并且对祖先等位基因的识别误差敏感。

2. 基于溯祖过程 (coalescent processes) 推断群体历史

根据谱系 (genealogy) 结构估计种群历史，此过程仅取决于溯祖事件的时间，而不取决于序列之间的确切谱系关系。例如，快速连续发生的溯祖事件通常表明群体规模较小。为了重建群体历史，天际线图方法利用了群体大小与溯祖时间期望值之间的相对简单关系。具体而言，每个间隔的平均群体大小可以通过间隔大小γ_i与$i(i-1)/2$的乘积来估算，其中 *i* 是该间隔中的谱系数。因此，这种关系给出了估计谱系中每个溯祖间隔的群体大小估计值，从而产生了群体历史的分段重建。通过谱系结构重建群体历史通常涉及相当大的不确定性，在此称为“溯祖错误”。溯祖是一个随机过程。任何单个谱系都仅表示此过程的单个随机实现。特别是，估计每个溯祖间隔内的群体大小会产生大量误差，并且等效于在仅从分布中给出单个样本的情况下估计指数分布的均值。溯祖错误朝着谱系的根部增加，在谱系的基础上，从较少的世系中重建了种群历史。使用溯祖理论对单种群 SFS 进行的研究，包括研究了群体不变的历史下固定式 SFS 的理论特性，以及具有确定的群体大小变化的非均衡群体大小的 AFS 分析研究。这些基于溯祖理论的方法被用于估计群体增长率并检测群体瓶颈的计算效率。但是，大多数理论和数据分析都是基于单个人群的 AFS，而不是基于

多个人群的联合等位基因频谱 (joint allele frequency spectrum, JAFS)。Excoffier 等人 (Excoffier *et al.*, 2013) 基于溯祖模拟方法开发软件 fastsimcoal2，用于推断更为复杂的群体历史，包括多个群体的分化、瓶颈和迁移。对于简单情形，fastsimcoal2 的分析效果同 DaDi 相近，但比 DaDi 更稳健。虽然简单情形下，fastsimcoal2 运行速度慢于 DaDi，但对于复杂情形，则快于 DaDi。Excoffier 等人采用 fastsimcoal2 分析了四个人类群体的种群历史，分别是非洲裔美国人 (African American, ASW)、欧洲人 (Europeans, CEU)、约鲁巴人 (Yoruba, YRI) 和卢赫雅人 (Luhya, LWK)。尽管 fastsimcoal2 通过溯祖模拟可以分析多个群体 (可以多达 12 个群体)，但是针对同一初始化参数的两次模拟数据会产生不同估计值。fastsimcoal2 同样没有考虑位点间的连锁效应，为了最大限度降低连锁不平衡效应，Excoffier 等人使得所分析的非编码 SNP 间距为 5 kb。

三、基于连锁不平衡的方法

1. 基于血源同一模型推断群体历史

依据溯祖理论，如果一个 Wright-Fisher 群体的有效群体大小为 N_e，那么该群体任意两个个体经过一代找到共同祖先的概率为 $1/N_e$，进而找到最近共同祖先 (the most recent common ancestor，TMRCA) 所需要的平均时间为 N_e 代；若在 m 代群体有 $N(m)$ 个个体，则 m 代群体任意两个个体经过一代找到共同祖先的概率为 $1/N(m)$，并且找到最近共同祖先所需要的平均时间为 $N(m)$代。因此，基因组上相应位点找到最近共同祖先所需平均时间的分布可以用来推断不同时期的群体大小。Palamara 等人 (Palamara *et al.*, 2012) 将 IBD 的长度视为连续性随机变量 L，其概率密度函数记为$\mathrm{p}(l|\theta)$；其中，参数集θ包含了与群体历史有关的各种参数。对于种群大小不变的情形 (图 1A)，参数集$\theta = \langle N_e \rangle$，其中$N_e$为有效群体大小；对于种群大小呈指数增长的情形，参数集$\theta = \langle N_a, N_c, G \rangle$，其中$N_a$为祖先群体大小，$N_c$为当前群体大小，$G$ 为指数增长持续的时间。由于给定位点距离最近发生重组位点的长度服从指数分布 (均值为 g_{tmrca}/50 cm，g_{tmrca} 表示找到最近共同祖先所需时间，单位为代)，而 IBD 长度的分布可视为上述独立的指数分布之和，故服从 2 阶埃尔朗分布 (Erlang2 distribution，Erl_2)。因此，Palamara 等人构建了如下概率密度函数用于刻画 IBD 长度分布：

$$\mathrm{p}(l|\theta) = \sum_{g=1}^{\infty} p(g_{tmrca} = g|\theta) \times Erl_2\left(l; \frac{g}{50}\right)$$

当种群大小变化时，$p(g_{tmrca} = g|\theta) = \frac{1}{N(g,\theta)} \prod_{j=1}^{g-1}\left(1 - \frac{1}{N(j,\theta)}\right)$，其中函数 $N(g,\theta)$表示第 g 代时的种群大小，该函数的具体形式依赖于对群体历史所做假设。利用上述关系，Palamara 等人基于三种群体历史模型对 500 位来自耶路撒冷的阿什肯纳兹犹太人基因组进行分析，推断相应群体历史模型下的参数信息。第一种群体历史模型为指数增长模型 (图 1B)，记为 M_E；第二种群体历史模型为瓶颈效应后指数增长模型 (图 1C)，记为 M_{FE}；第三种群体历史模型为伴有瓶颈效应的双指数增长模型 (图 1D)，记为 M_{EFE}。通过采用赤池信息量准则 (Akaike information critation, AIC) 比较三种群体历史模型后，Palamara 等人推断阿什肯纳兹犹太人群体历史如下：200 代前，阿什肯纳兹犹太人祖先群体大小约为 2,300；随后祖先群体进入指数增长阶段，祖先群体大小在 34 代前增至约 45,000；而后祖先群体发生瓶颈效应，祖先群体大小锐减至约 270；最后，经过 33 代的指数增长，当前群体大小约为 4,300,000。Palamara 等人采用 Beagle 软件包 (Browning, S.R. and Browning B.L., 2007) 对样本进行分型 (phasing) 分析并且采用 GERMLINE 软件包 (Gusev *et al.*, 2009) 对共享 IBD 进行估计。为了提高在特定数据集上 IBD 检测的质量，Palamara 等人采用如下方式对 GERMLINE 算法的有关参数做了调整：(1) Palamara 等人采用 GERMLINE 算法的默认参数从真实数据中提取 IBD 片段并推断种群历史相应参数；(2) 基于 (1) 中推断的种群历史参数生成模拟数据，采用 GERMLINE 算法分析模拟数据，找到一组有关参数 "err_hom，err_het，bits" 的取值，使得从模拟数据中提取的 IBD 同 (1) 中提取的相差最小；(3) 基于 (2) 中有关参数 "err_hom，err_het，bits" 的取值重复 (1) 和 (2)，直至从真实数据中提取的 IBD 片段收敛。经过上述步骤，Palamara 等人对 GERMLINE 算法有关参数的调整如下：-min_m 1 -err_hom 0 -err_het 2 -bits 25 -h_extend。

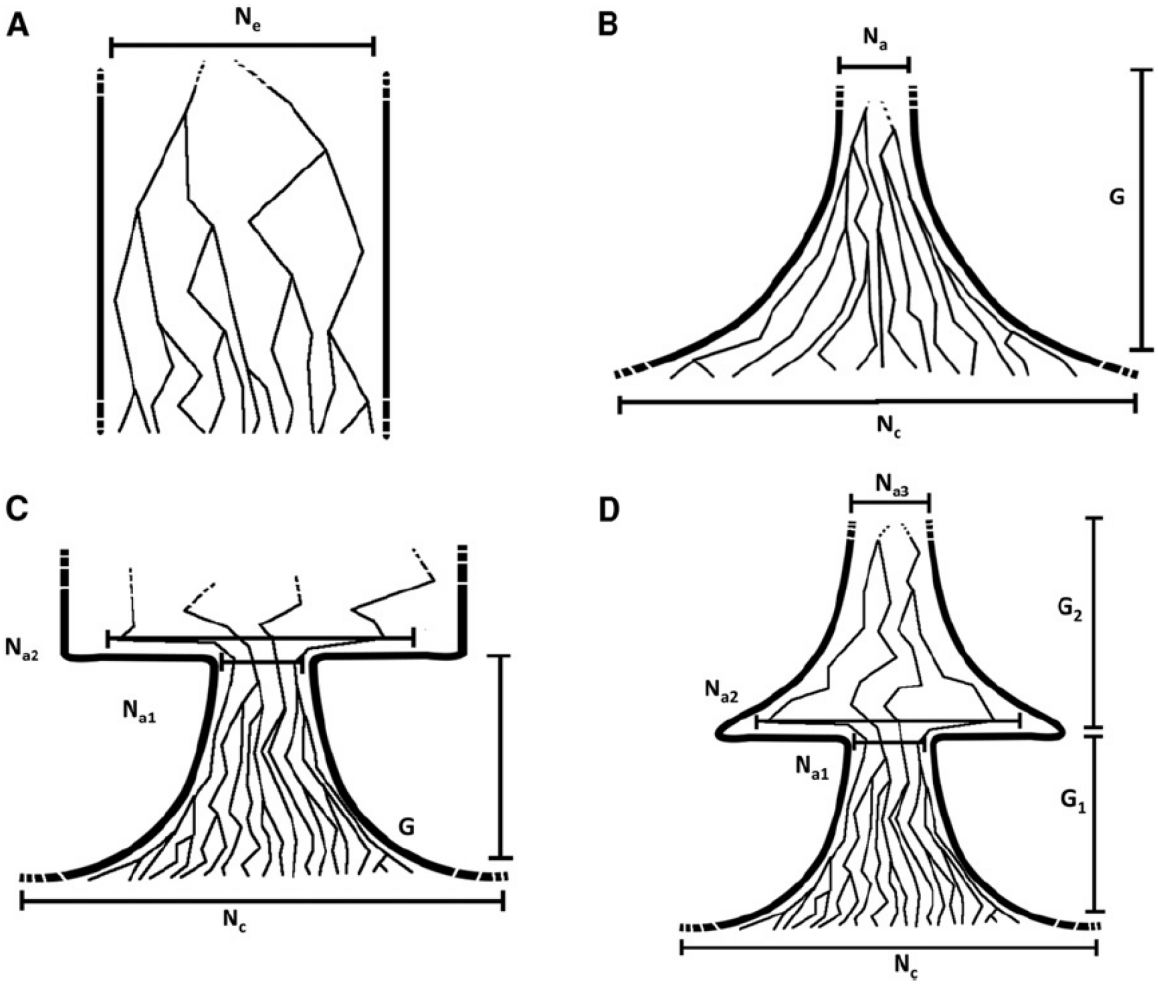

图 1. 群体历史模型 (A) 群体大小不变；(B) 群体大小指数增长；(C) 群体大小发生瓶颈效应后开始指数增长；(D) 群体大小在两次指数增长间发生瓶颈效应。(Palamara *et al.*, 2012)

2. 基于溯祖-隐马尔可夫模型推断群体历史

溯祖过程是一个随机模型，用于刻画时间序列上的样本谱系结构。对于基因组序列上的样本谱系结构，Wiuf 和 Hein (Wiuf and Hein, 1999) 首次构造了一个复杂的、非马氏 (non-Markovian) 过程的随机模型；McVea 和 Gardin (McVean and Gardin, 2005) 则引入了简单的、近似马氏过程的随机模型，称作递次式马氏溯祖 (sequentially Markovian coalescent, SMC) 模型。SMC 模型利用隐马尔可夫框架刻画观测序列，将观测不到的样本谱系结构作为隐变量 (latent variable)。Li 和 Durbin (Li and Durbin, 2011) 认为在一个双倍体个体的基因组中，相应等位基因溯祖时间 (找到最近共同祖先的时间，TMRCA) 的分布蕴含其所在群体的群体历史信

息，即群体大小随时间变化的信息。因此，Li 和 Durbin 基于 SMC 模型开发了成对递次式马氏溯祖模型 (pairwise sequentially Markovian coalescent, PSMC) 用来重构双倍体个体基因组上的 TMRCA 分布，该模型将 SMC 模型中的隐变量由样本谱系结构简化为 TMRCA (图 2a)。在 PSMC 模型中，若观测到的基因型为纯合，则将观测值记为‘0’；否则记为‘1’。当隐变量 TMRCA 的状态为 *t* 时，发射概率 (emission probability) 分别为$\mathrm{e}(0|t) = e^{-\theta t}$和$\mathrm{e}(1|t) = 1 - e^{-\theta t}$；当从状态 *s* 转移到状态 *t* 时，其转移概率如下：

$$\mathrm{p}(t|s) = (1 - e^{-\rho t})q(t|s) + e^{-\rho s}\delta(t - s)$$

其中δ(·)为狄拉克 delta 函数，$q(t|s) = \frac{1}{\lambda(t)}\int_0^{min\{s,t\}}\frac{1}{s} \times e^{-\int_u^t \frac{dv}{\lambda(v)}}du$。PSMC 模型包含如下参数：突变率$\theta$，重组率$\rho$，祖先群体大小$\lambda(t)$。参数$\theta$被定义为$4N_0\mu$，$\mu$表示基因组上单个位点 (site) 经过一代的突变率，其值需要预先设定；N_0为常数，其值依赖于参数θ的估计值。$\lambda(t)$表示 *t* 时刻相对群体大小，其表达式为$\lambda(t) = N_e(t)/N_0$，$N_e(t)$为 *t* 时刻群体大小。为了能够求解 PSMC 模型，Li 和 Durbin 采用如下策略对隐变量 TMRCA 的状态空间进行离散化：设定隐变量 TMRCA 的最大取值为 T_{max}，并将溯祖时间划分为 n 个时间区段，每个区段边界值的表达式为$t_i = 0.1exp[i/nlog(1 + 10T_{max})] - 0.1,\ i = 0,\cdots,n$。对状态空间离散化后，Li 和 Durbin 将函数$\lambda(t)$转换为分段式常函数，即$\lambda(t) = \lambda_i\ \ t \in (t_i, t_{i+1}]\ i = 0,\cdots,n$。经过上述处理，PSMC 模型需要估计如下参数：突变率θ，重组率ρ，祖先群体大小$\lambda_i\ \ i = 0,\cdots,n$。Li 和 Durbin 采用最大期望算法 (Expectation-Maximization algorithm, EM) 对上述参数进行估计，其中参数初始值设置如下：突变率θ依据观测到的杂合度进行计算，重组率ρ为计算所得突变率θ的四分之一，祖先群体大小λ_i 值均为 1。在 EM 算法的最大化步 (Maximization step)，Li 和 Durbin 采用鲍威尔算法 (Powell's algorithm) 求解模型参数。Li 和 Durbin 用 PSMC 模型分析东亚、欧洲和非洲三个群体的历史，其单个位点经过一代的突变率设为 2.5 x 10^{-8}，且一代的时间设为 25 年。分析结果表明，东亚和欧洲两个群体在 2 万年前的群体历史非常接近，都经历严重的瓶颈效应；虽然非洲群体在同时期也经历瓶颈效应，但是相对东亚和欧洲两个群体而言其程度较弱 (图 3)。尽管 PSMC 模型可以推断出不同时间段上的群体大小，但是仿真

结果表明近期 (800 代以内) 和远期 (10 万代以前) 的群体大小的估计效果很差，这是由于在一个双倍体的基因组序列上很难捕获到上述两个时期发生的重组事件。Li 和 Durbin 采用工具 BWA (Li and Durbin, 2009) 进行序列比对，之后采用 SAMtools 工具包 (Li *et al.*, 2009) 对比对结果进行排序、合并。Li 和 Durbin 将 SAMtools 的输出结果做如下处理作为 PSMC 的输入：(1) 每 100 bp 视为一个单位，且不重叠；(2) 在每个单位中，若被 SAMtools 过滤的位点 (site) 数超过 90，则将该单位记为'.'，否则记为'1'或'0'；(3) 在每个单位中，若至少含有一个杂合子，则将该单位记为'1'，否则记为'0'。最后，Li 和 Durbin 对 PSMC 做如下参数设置：-N25 -t15 -r5 -p "4+25*2+4+6"。

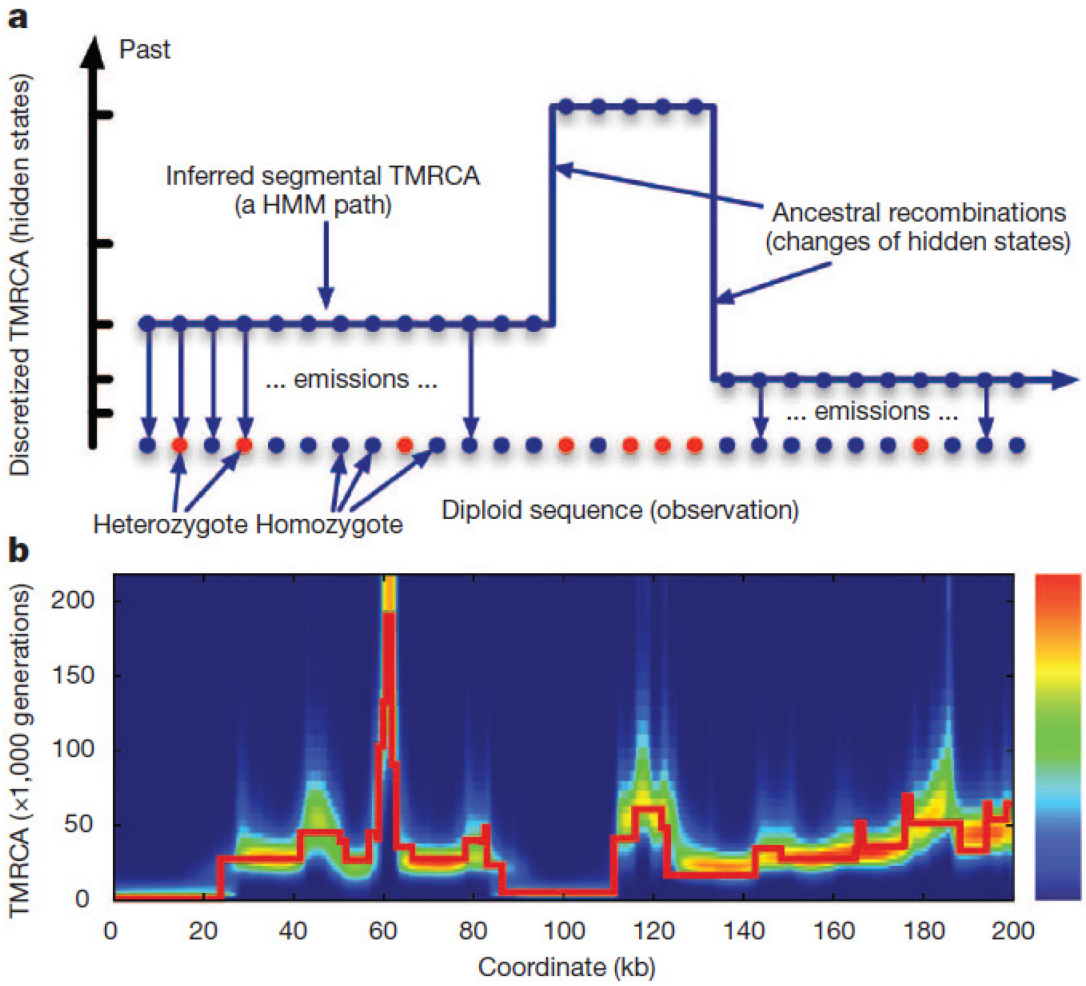

图 2. PSMC 模型示意图及对模拟数据的分析结果 (a) PSMC 模型基于杂合子密度推断 TMRCA；离散 TMRCA 为隐状态，当出现祖先重组事件时隐状态发生转移。

(b) 采用 ms 软件生成 200 kb 区域内的 TMRCA 分布 (加粗红线)；利用 PSMC 模型对模拟数据进行分析，推断的 TMRCA 分布采用热图表示。(Li and Durbin, 2011)

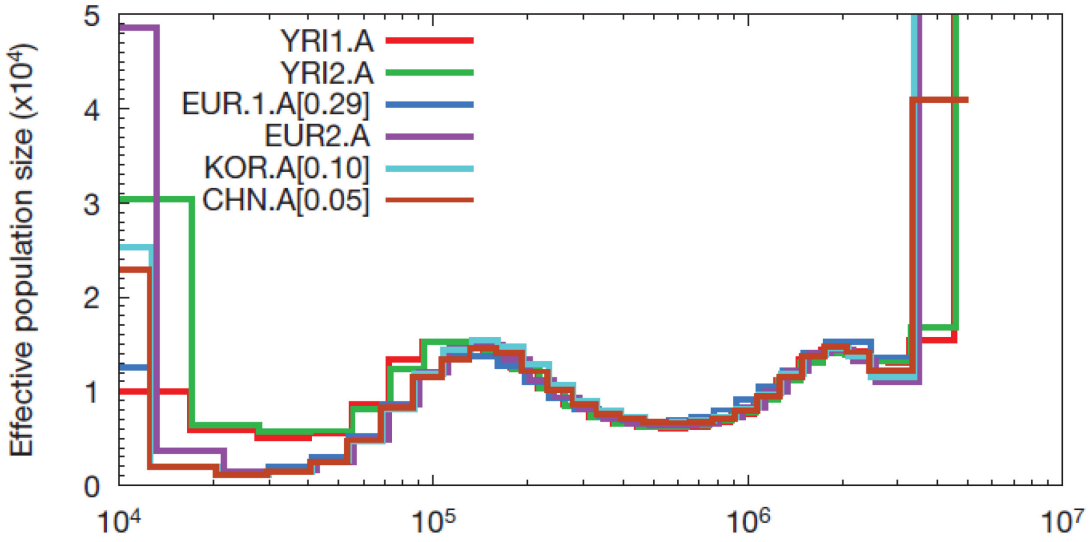

图 3. 基于 6 个个体常染色体的种群历史推断 假设在 CHN.A、KOR.A 和 EUR.1.A 中，分别有 5%、10%和 29%杂合子丢失。(Li and Durbin, 2011)

Schiffels 和 Durbin (Schiffels and Durbin, 2014) 为了能够在基因组序列上捕获近期发生的重组事件，基于 SMC 模型开发了多对递次式马氏溯祖模型 (multiple sequentially Markovian coalescent, MSMC)。MSMC 模型可以分析多个双倍体样本，其隐变量是多样本谱系中首次溯祖时间 *t* 和首次溯祖序列(*i,j*)，重组事件的发生将引起隐变量状态的改变 (图 4)。Schiffels 和 Durbin 基于 SMC 框架给出下述公式描述隐变量从状态(*t,i,j*)转移到状态(*s,k,l*)的概率：

$$q_1(t) = e^{-M\rho t} + (1 - e^{-M\rho t})\frac{1}{t}\frac{1}{M}\int_0^t 1 + (M-3)exp\left(-M\int_u^t \lambda(v)\,dv\right)du \; if \; (t,i,j) = (s,k,l)$$

$$q_2(t|s) = (1 - e^{-M\rho s})\frac{1}{s}\frac{1}{M}2\lambda(t)$$

$$\lambda(t) = \begin{cases} \int_0^t exp\left(-M\int_u^t \lambda(v)dv\right)du & t < s \\ exp\left(-\binom{M}{2}\int_s^t \lambda(v)dv\right)\int_0^s exp\left(-M\int_u^s \lambda(v)dv\right)du & t > s \end{cases}$$

其中 M 为单倍型的数目，ρ为重组率，$\lambda(t)$表示 t 时刻相对群体大小。由于 MSMC 模型处理多个样本，其发射概率是单点突变 (singleton mutation) 的概率。若单点突变发生在首次溯祖序列(i,j)上，则发射概率$e(O|t,i,j) = ut$，否则为 1-ut，u为单个位点经过一代的突变率。Schiffels 和 Durbin 运用 MSMC 模型分析了人类 9 个群体的种群历史 (每个群体 2 个样本，4 个单倍型)，结果表明走出非洲后的群体的种群大小持续下降并在 4 万至 6 万年前经历了瓶颈 (图 5)。尽管 MSMC 模型理论上可以分析多个个体的基因组序列，但是由于其模型的复杂性，仿真结果表明 MSMC 模型所能分析的样本数不超过 4 个。采用 MSMC 模型分析真实数据时，不仅需要进行序列比对、排序、合并等处理，还需要做单倍体分型 (haplotype phasing) 处理。Schiffels 和 Durbin 采用工具 Shapeit2 (Delaneau *et al.*, 2013) 进行单倍体分型处理，并提供工具包 "generate_multihetsep.py" 将处理后的数据转换为 MSMC 模型的输入文件。在估计种群大小时，Schiffels 和 Durbin 建议使用参数"-fixedRecombination"。

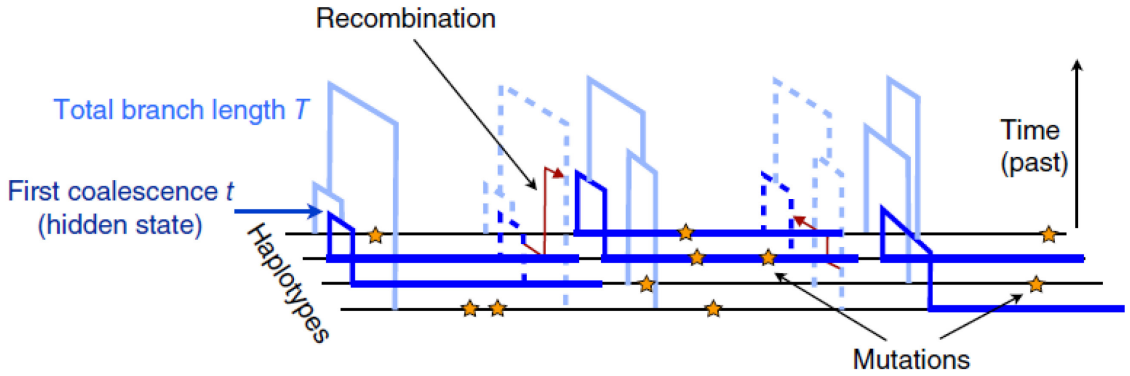

图 4. MSMC 模型示意图 多样本谱系结构因重组事件沿着基因组序列发生改变，其隐变量是多样本谱系中首次溯祖时间 t 和首次溯祖序列 (图中加粗蓝线)。(Schiffels and Durbin, 2014)

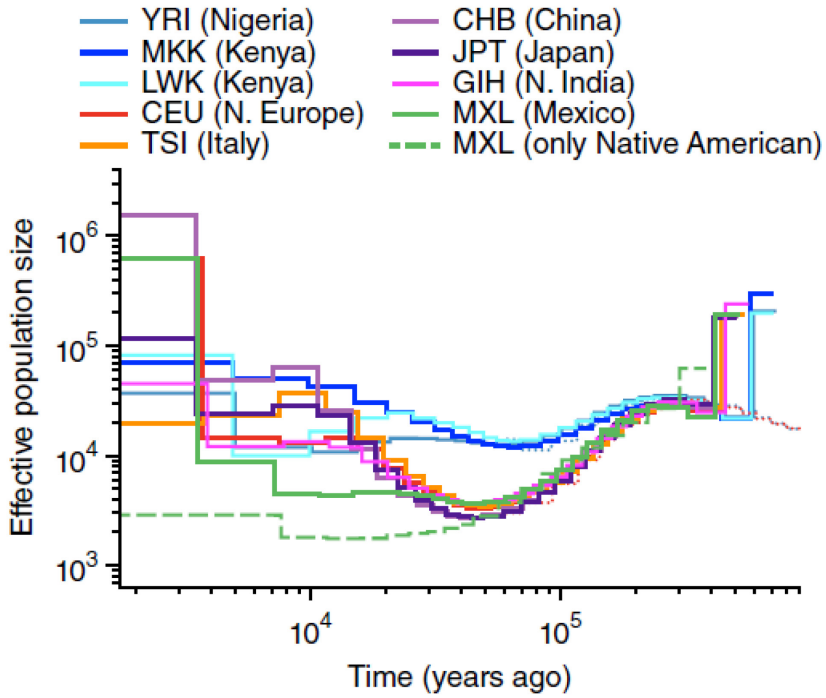

图 5. 9 个群体的种群历史 每个群体 2 个样本，4 个单倍型。(Schiffels and Durbin, 2014)

Terhorst 等人 (Terhorst *et al.*, 2017) 提出了一种称为 SMC++的统计模型，该模型采用如下策略分析多样本数据：(1) 利用基于连锁不平衡的方法分析来自一个已知双倍体个体的基因组序列数据；(2) 利用基于基点频谱的方法分析剩余样本的基因组序列数据。SMC++模型假设在 n+2 个单倍体 (haploids) 样本构成的谱系 (genealogy) 中有 2 个单倍体来自同一个个体并可识别，其所在分支 (lineage) 记为‘1’和‘2’。因此，在任意位点的等位基因状态 (allelic state) 可记为$(a,b) \in \{0,1,2\} \times \{0,1,\cdots,n\}$。若分支‘1’和分支‘2’的溯祖时间为$\tau$ (记为C_{12})，则观察到(a,b)的概率为：

$$\mathrm{P}\big((a,b)|C_{12}=\tau\big)=\frac{\theta}{2}[CSFS(\tau)]_{ab}+O(\theta^2)$$

其中$CSFS(\tau)$为条件基点频谱，θ为突变率。同 MSMC 模型相比，SMC++模型可以分析上百个样本并且不需要做单倍体分型处理。Terhorst 等人运用 SMC++模型推断了长尾草雀 (Long-tailed finch)、斑胸草雀 (Zebra finch) 和黑腹果蝇 (*D.melanogaster*) 的种群历史，其中长尾草雀和斑胸草雀的样本数为 40，单个位点

经过一代的突变率设为 7 x 10^{-10}，且一代的时间设为 3 个月；黑腹果蝇的样本数为 197，单个位点经过一代的突变率设为 3 x 10^{-9}，且一代的时间设为 1 个月。分析结果表明三物种的种群大小约 50 万年前开始下降，长尾草雀和斑胸草雀的种群大小约 6 万年前开始显著上升，而黑腹果蝇的种群大小则约 10 万年前开始上升但不显著 (图 6)。运用 SMC++模型依照下述步骤分析真实数据：(1) 对真实数据进行序列比对、排序、合并等处理；(2) 将处理后的数据通过命令“smc++ vcf2smc”转换为 SMC++模型的输入文件；(3) 通过命令“smc++ estimate”估计种群历史；(4) 通过命令“smc++ plot”可视化分析结果。Terhorst 等人采用三次样条插值 (cubic splines interpolation) 技术对 smc++的分析结果做平滑处理。

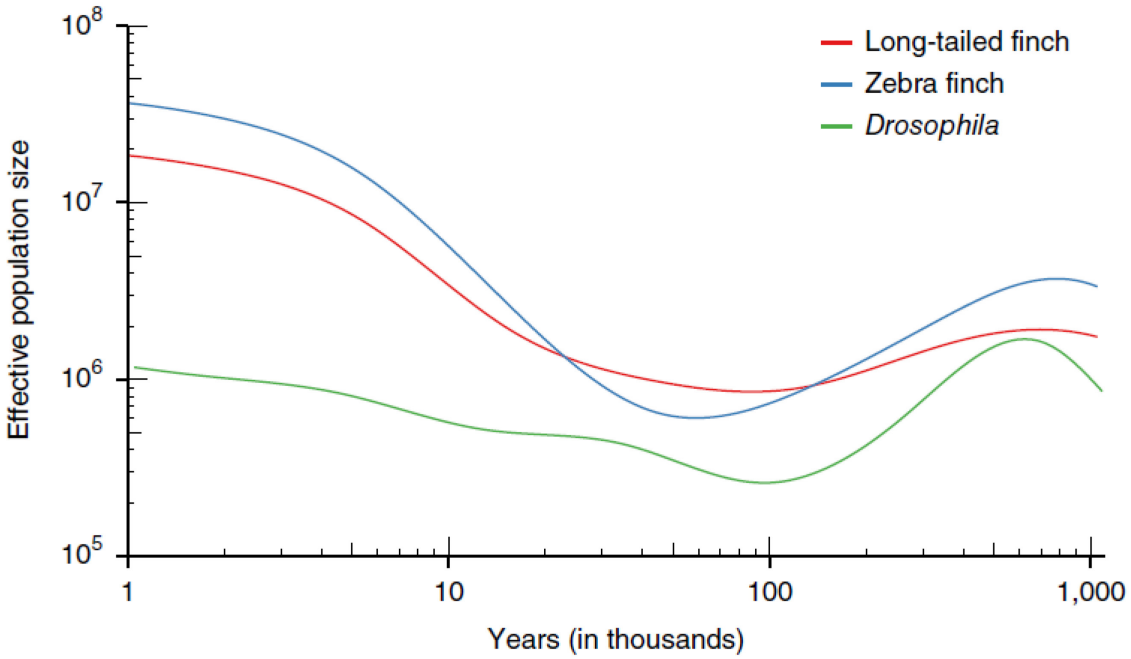

图 6. 长尾草雀 (Long-tailed finch)、斑胸草雀 (Zebra finch) 和黑腹果蝇 (D.melanogaster) 的种群历史 (Terhorst *et al.*, 2017)

四、推断群体历史的基本流程

虽然群体历史的推断方法有很多种，但是不管采用何种方法，其推断群体历史的基本流程 (图 7) 大体一致。下面以采用 psmc (版本 0.6.5) 在 linux 平台上分析来自 NCBI 数据库的阿德利企鹅样本数据 (SRR1145007.sra) 为例 (分析结果见图 8)，介绍群体历史推断的基本流程。 (1) 下载原始数据：从 NCBI 数据库下载原始数据可采用工具包 sra

toolkit (版本 2.9.6) 中的 prefetch 命令，其命令格式为：prefetch SRR1145007； (2) 过滤原始数据：首先将下载的原始数据由 sra 格式转为 fastq 格式，可采用工具包 sra toolkit (版本 2.9.6) 中的 fastq-dump 命令，其命令格式为：fastq-dump --split-files SRR1145007.sra。之后对生成的 fastq 格式的原始数据进行过滤，可采用工具包 ngsqc toolkit (版本 2.3.3) 中的 IlluQC.pl 命令，其命令格式为：perl IlluQC.pl -pe SRR1145007_1.fastq SRR1145007_2.fastq 2 A -z g -o SRR1145007；(3) 将过滤后的原始数据同参考序列进行比对：目前用于序列比对的软件多达 60 余种，其中 Bowtie2、BWA、MAQ 和 SOAP2 是比较常用的四款软件，适用于二代高通量短序列数据。若采用 Bowtie2 (版本 2.3.5.1) 进行比对，则首先需要为参考序列 adelie.fa (下载链接为 http://gigadb.org/dataset/100006) 建立索引，其命令格式为：bowtie2-build adelie.fa adelie。之后进行比对并采用工具 samtools (版本 1.9) 生成 bam 文件，其命令格式为：bowtie2 -p 5 -x adelie -1 SRR1145007_1_filter.fq -2 SRR1145007_2_filter.fq | samtools sort -O bam -o SRR1145007.bam --threads 3。(4) 进行变异分析：可采用 bcftools (版本 1.9) 工具包中的相应命令分析比对序列，检测 snp 和 indel 等变异信息，其命令格式为： bcftools mpileup -C50 -f adeliae.fa SRR1145007.bam | bcftools call -c -V indels | vcfutils.pl vcf2fq -d 10 -D 100 | gzip > SRR1145007.fq.gz。(5) 推断群体历史：采用群体历史推断软件所提供的转换工具将第 (4) 步中生成的数据转成输入数据，之后进行群体历史推断。psmc (版本 0.6.5) 的转换工具是 fq2psmcfa，其命令格式为：fq2psmcfa -q20 SRR1145007.fq.gz > adelie.psmcfa，推断群体历史的命令为：psmc –N30 -t15 -r5 -p "4+25*2+4+6" -o adelie.psmc adelie.psmcfa。

图 7. 群体历史推断的基本流程

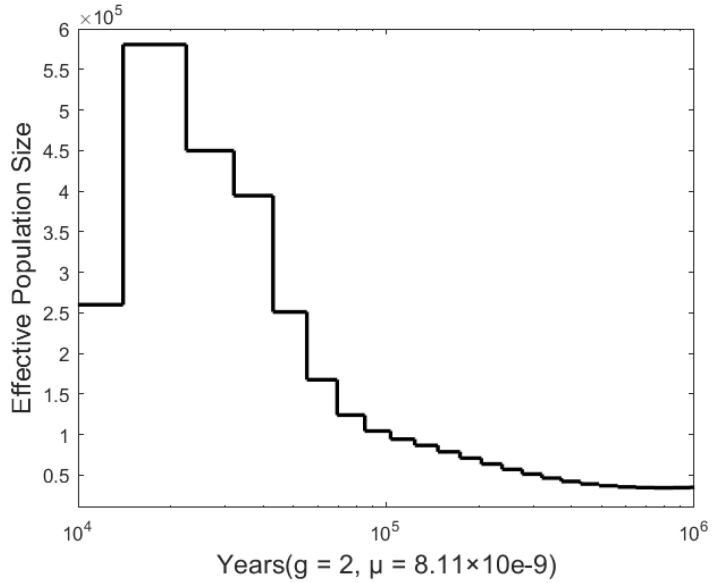

图 8. 阿德利企鹅群体历史 g 表示年/代；μ 表示位点突变率，其单位为代。

五、总结与展望

目前，基于基因组序列数据推断群体历史的方法主要采用两种统计模型：第一种统计模型基于位点频谱信息，第二种统计模型基于连锁不平衡信息。在基于位点频谱信息的统

计模型中，通常采用扩散过程和溯祖过程两种框架构建位点频谱信息同群体历史信息之间的关系。在扩散过程框架下，推断群体历史需要预先给定群体大小变化模型；而在溯祖过程框架下则蕴含一种假设条件：群体大小在每一个溯祖时间段内都是固定不变的。基于位点频谱信息的方法适用于推断近期群体历史，采用的数据通常为非编码区的 DNA 序列。但是，基于位点频谱信息的方法没有考虑群体内部结构和近交个体的影响。因此，采用基于位点频谱信息的方法推断群体历史时，可考虑先进行血源同一和群体结构分析，去除近交样本并根据群体内部结构重新采集样本以消除群体内部结构的影响。

在基于连锁不平衡信息的统计模型中，通常采用血源同一和溯祖-隐马尔可夫两种模型构建连锁不平衡同群体历史信息之间的关系。基于血源同一模型方法的优点在于能够较为准确推断群体近期历史，但其缺点是需要预先给定群体大小变化模型。为了克服这一缺点，该类方法通常给定多种群体大小变化模型并采用统计方法 (如赤池信息量准则) 进行比较，进而选择相对适合观测数据的群体大小变化模型。尽管理论上基于溯祖-隐马尔可夫模型方法可以精细刻画群体历史且无需对群体大小变化做任何假设，但在实际求解过程中存在两种障碍需要克服：(1) 隐变量的状态空间无穷大；(2) 基因树拓扑结构的数目随着样本数的增加而指数增长。通常采用下述方法克服相应障碍：(1) 对连续性隐变量做离散化处理；(2) 用谱系结构的个别特征替代谱系结构，例如 MSMC 模型用样本谱系中首次溯祖时间 t_{first} 代替谱系结构。现有基于溯祖-隐马尔可夫模型方法在离散化隐变量过程中假设群体大小在每一个离散化时间段内都是固定不变的，这使得推断的群体历史不再精细。虽然足够大的隐变量状态空间可以使推断的群体历史变得精细，但是在计算上却无法处理并且让离散化隐变量的处理失去意义。基于连锁不平衡信息的方法可以只分析一个样本的全基因组 DNA 序列 (包括编码区和非编码区) 推断群体历史，不过仿真结果表明对近期群体历史 (800 代以内) 的推断效果较差。尽管部分基于连锁不平衡信息的方法 (例如 MSMC 和 SMC++) 通过增加群体样本试图改善对近期群体历史的推断，但推断效果未有明显改善。

综上所述，现有方法各具优点，但都需要对群体大小变化做一定程度的假设，并且无法精细刻画群体历史。因此，未来推断群体历史的统计模型应具备以下特点：(1) 能够分析多种遗传信息；(2) 无需对群体大小变化做任何假设；(3) 可实现对群体历史的精细刻画。

致谢

感谢葛德燕博士在投稿期间给予的帮助，以及审稿人提出的宝贵意见。

参考文献

1. Atkinson, Q. D., Gray, R. D. and Drummond, A. J. (2008). MtDNA variation predicts population size in humans and reveals a major southern Asian chapter in human prehistory. *Mol Biol Evol* 25: 468-474.
2. Browning, S. R. and Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet* 81: 1084-1097.
3. Cann, R. L., Stoneking, M. and Wilson, A. C. (1987). Mitochondrial DNA and human evolution. *Nature* 325: 31-6.
4. Excoffier, L., Dupanloupet, L., Huerta-Sánchez, E., Sousa, V. C. and Foll, M. (2013). Robust demographic inference from genomic and SNP data. *PLoS Genet* 9(10): e1003905.
5. Gusev, A., Lowe, J. K., Stoffel, M., Daly, M. J., Altshuler, D., Breslow, J. L., Friedman, J. M. and Pe'er, I. (2009). Whole population, genome-wide mapping for hidden relateness. *Genome Res* 19: 318-26.
6. Gutenkunst R. N., Hernandez, R. D., Williamson, S. H., Bustamante, C. D. and Mcvean, G. (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet* 5.
7. Jin, L. and Su, B. (2000). Natives or immigrants: modern human origin in East Asia. *Nat Rev Genet* 1(2): 126-133.
8. Ke, Y. H., Su, B., Song, X. F., Lu, D. R., Chen, L. F., Li, H. Y., Qi, C., Marzuki, S., Deka, R., Underhill, P., Xiao, C., Shriver, M., Lell, J., Wallace, D., Wells, R. S., Seielstad, M., Oefner, P., Zhu, D., Jin, J., Huang, W., Chakraborty, R., Chen, Z. and J, L. (2001). African origin of modern humans in East Asia: a tale of 12,000 Y chromosomes. *Science* 292: 1151-3.
9. Kong, Q. P., Yao, Y. G., Sun, C., Bandelt, H. J., Zhu, C. L. and Zhang, Y. P. (2003). Phylogeny of East Asian mitochondrial DNA lineages inferred from complete sequences. *Am J Hum Genet* 73: 671-6.
10. Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-

Wheeler transform. *Bioinformatics* 25: 1754.

11. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. (2009). The sequence alignment/map format and samtools. *Bioinformatics* 25(16): 2078-2079.
12. Li, H. and Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature* 475: 493–496.
13. McVean, G. A. T. and Cardin, N. J. (2005). Approximating the coalescent with recombination. *Philos Trans R Soc B-Biol Sci* 360: 1387-93.
14. Palamara, P. F., Lencz, T., Darvasi, A. and Pe'er, I. (2012). Length distributions of identity by descent reveal fine-scale demographic history. *Am J Hum Genet* 91: 809-22.
15. Schiffels, S. and Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. *Nature Genet* 46: 919-925.
16. Stajich, J. E. and Hahn, M. W. (2005). Disentangling the effects of demography and selection in human history. *Mol Biol Evol* 22: 63-73.
17. Terhorst, J., Kamm, J. A. and Song, Y. S. (2017). Robust and scalable inference of population history froth hundreds of unphased whole genomes. *Nature Genet* 49: 303-9.
18. Underhill, P. A., Passarino, G., Lin, A. A., Shen, P., Lahr, M. M., Foley, R. A., Oefner, P. J. and Cavalli-Sforza, L. L. (2001). The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. *Ann Hum Genet* 65: 43-62.
19. Underhill, P. A. and Kivisild, T. (2007). Use of Y chromosome and mitochondrial DNA population structure in tracing human migrations. *Annu Rev Genet* 41: 539-564.
20. Wiuf, C. and Hein, J. (1999). Recombination as a point process along sequences. *Theor Popul Biol* 55: 248-259.
21. Yao, Y. G., Kong, Q. P., Bandelt, H. J., Kivisild, T. and Zhang, Y. P. (2002). Phylogeographic differentiation of mitochondrial DNA in Han Chinese. *Am J Hum Genet* 70: 635-651.

