

群体进化高级分析

生信帮

云南农业大学

云南省生物大数据重点实验室

2023年8月26日

种群历史动态分析



雲南農業大學
Yunnan Agricultural University

云南省生物大数据重点实验室



生信帮

历史上的群体大小对于理解种群的进化历史十分重要。

基本原理：由于重组的存在，基因组被打断成许多小的片段，利用**高深度重测序数据**，我们可以找出这些小片段上的变异位点，进而可以推断这些小片段之间的最近共组时间。而最近共组时间的分布情况包含群体大小随着时间的推移的变化信息，根据最近共组时间分布的比例，可以推算该群体在历史各个时期有效群体大小。

使用上的建议：

1. 测序深度达到20X时，推荐使用MSMC2工具，此时每个群体最多使用4个样本；测序乘数较低时，推荐使用SMC++软件
2. MSMC2可以推断分化时间，前提是先分相 (phase)，这就要求群体的样本量较大，或者有reference panel
3. 获得物种的突变率和代时

种群历史动态分析

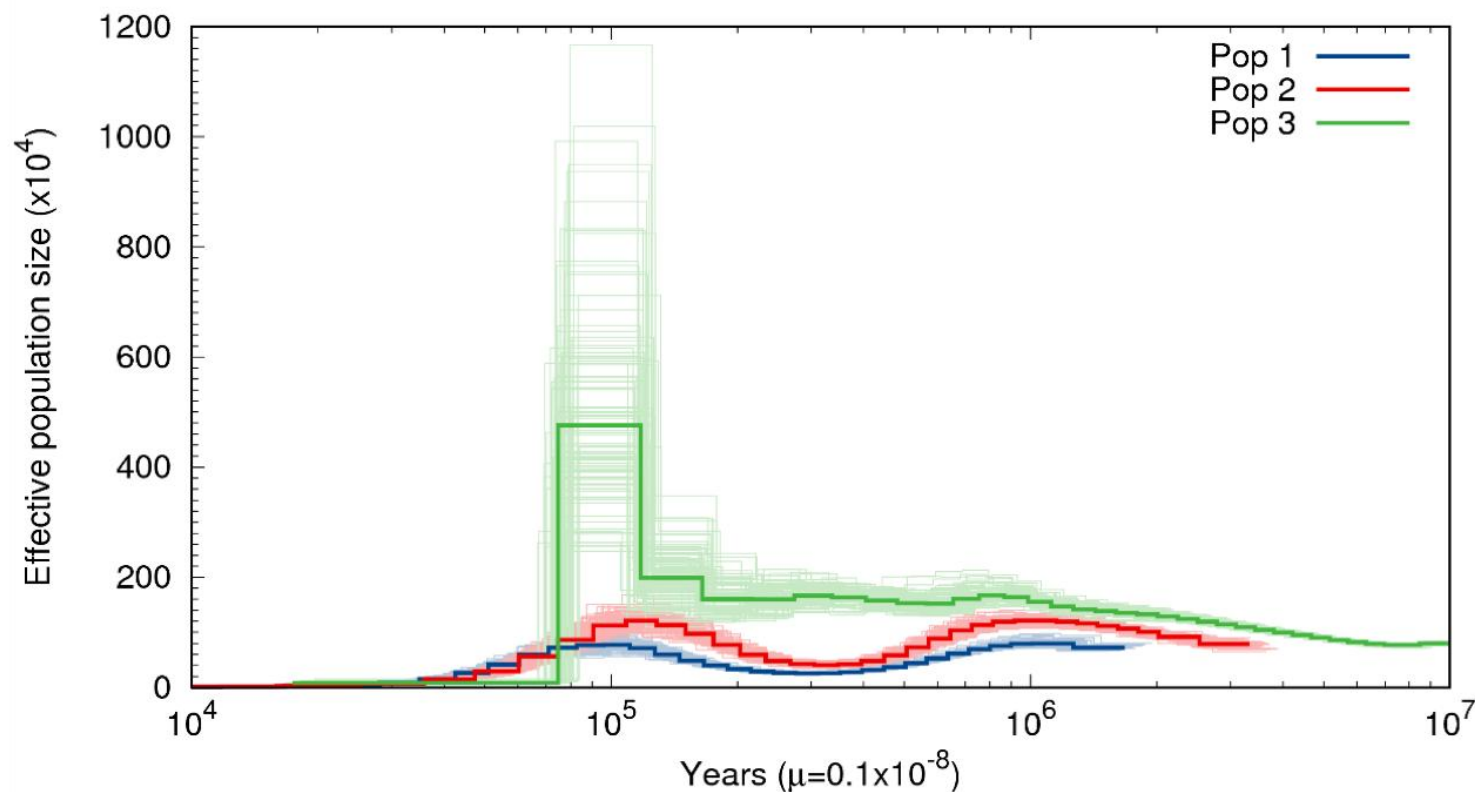


雲南農業大學
Yunnan Agricultural University



生信帮

云南省生物大数据重点实验室



1. 群体在什么时间有效群体大小变化情况 — 环境变化、动植物驯化
2. 不同群体间在什么时间有效群体大小发生了明显差异 — 群体分离

该分析需客户提供如下两个参数：

1. 突变速率；
2. 代时

风险提示： 简化数据做该分析，技术上可实现，但结果很不好，因此非常不推荐使用简化数据做该分析。

种群历史动态分析



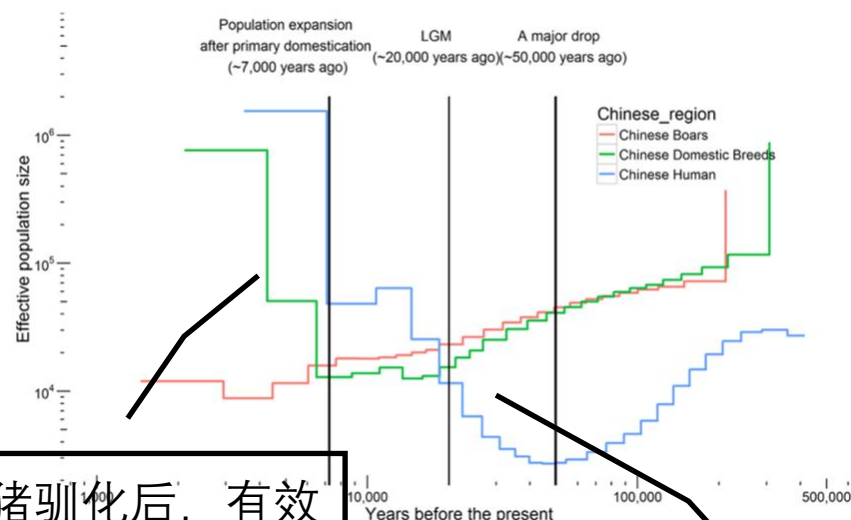
雲南農業大學
Yunnan Agricultural University

云南省生物大数据重点实验室



生信帮

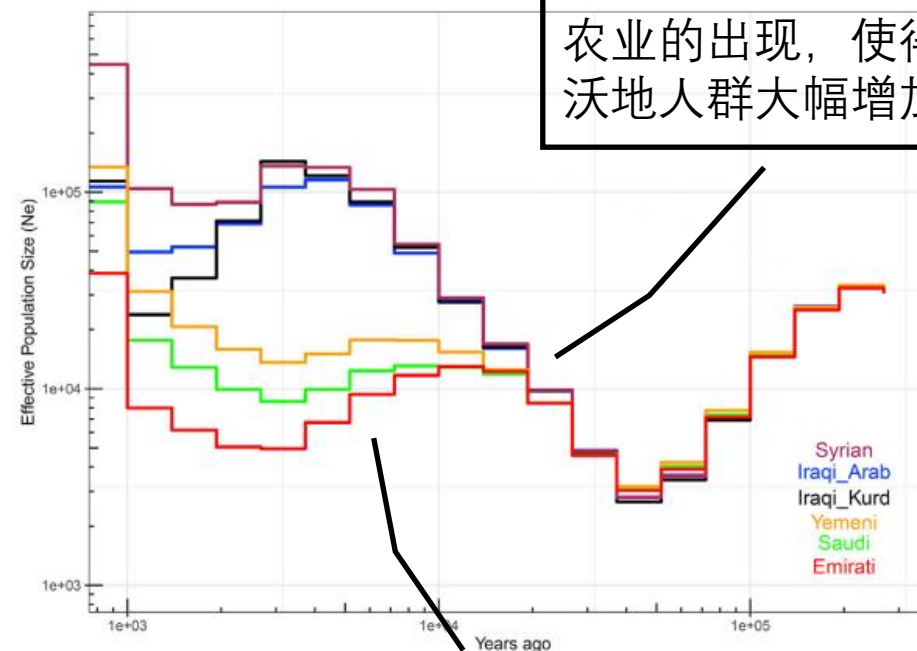
人和驯化动物一起展示，研究驯化问题



家猪驯化后，有效群体大小快速增加

人类有效群体大小逐渐增加的同时，猪的有效群体大小逐渐减少

借助人群变化推断农业起源



农业的出现，使得新月沃地人群大幅增加

阿拉伯半岛地区开始荒漠化，人群减少



在遗传学中，基因渗入(introgression)是指两个基因库间的基因流动，通常是经过种间杂交产生。基因渗入是一个长期的过程，它可能需要许多代杂交才能产生。

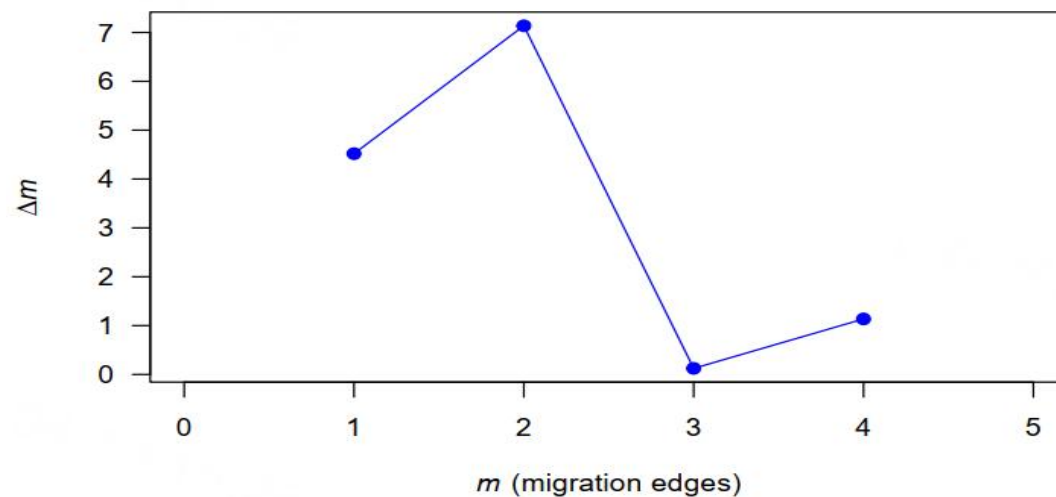
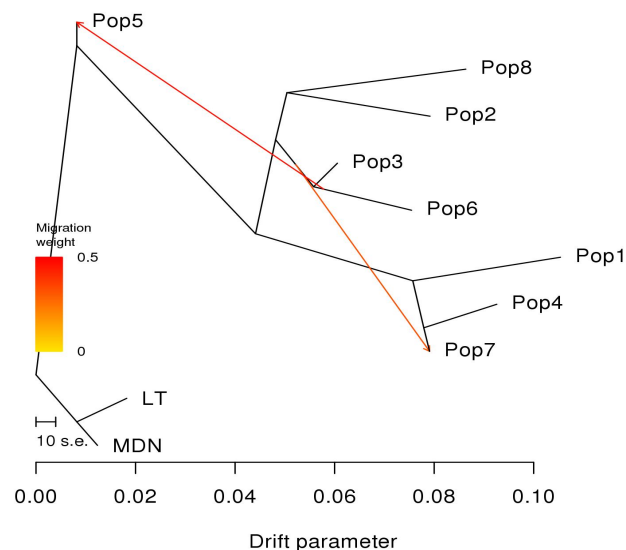
说明：

1. 该类型分析算法都是基于等位基因频率计算的，因此亚群包含的样本量越多，该分析结果越准确，如果样本量过少，技术上可执行，但结果一般不可靠。
2. 该类型分析对老师的相关知识要求较高，分析结果的解读一般需要结合样本材料，物种的进化史等进行合理的故事性描述，这方面分析人员能提供的解释非常有限，也无法评判结果的好坏，仅能客观展示分析结果。



TreeMix

Treemix软件使用全基因组的等位基因频率数据，推断多个群体的分化和混合的模式。该软件输入数据为多个群体的等位基因频率数据，可以生成这些群体的最大似然树，并且可以推测群体混合事件。



风险提示：该分析的群体数量最好大于4，否则可能没有分析结果。



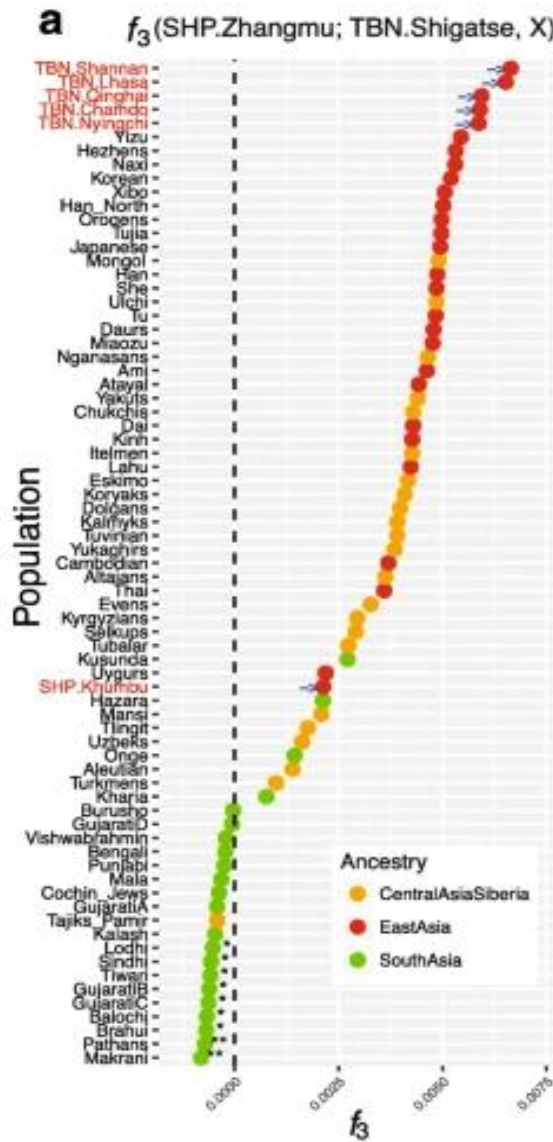
f3统计量

F3中的3表示3个群之间的基因流，该统计量主要是计算一个群和另外两个群之间等位基因频率的相关性。如果一个群（C）与另外两个群（A,B）之间的等位基因频率相关性越高，则说明C群可能混合了A和B的基因。

该分析中，C群为目标群体，A,B为祖源群体，及检测C群是否A群和B群混合而来。（负值表示具有显著的统计学效应）

SourceA	SourceB	TargetC	F3	StdErr	Z
Feral	Wild	Domestic	0.09717	0.003649	26.63
Domestic	Wild	Feral	-0.017973	0.001347	-13.339
Domestic	Feral	Wild	0.767581	0.017444	44.003

风险提示：该分析至少要有3个群体。





RFMix

RFMix软件侧重于检测渗入片段或混合基因。

分析需要设置一个目标群，一个参考群。

该分析必须提供物种的遗传图谱。



D统计量

D统计也叫ABBA BABA检测，更适合检测种间的基因流。

该分析是在4个群体间进行的，因此需要至少有4个群。

Pop1(W)	Pop2(X)	Pop3(Y)	Pop4(Z)	D-stat	Z	BABA	ABBA
pop3	pop8	pop7	pop6	0.0601	7.895	33690	29872
pop3	pop8	pop7	pop1	-0.1060	-18.338	54049	66865
pop3	pop8	pop7	pop5	-0.0307	-3.880	57530	61189
pop3	pop8	pop7	pop4	-0.0299	-3.744	57404	60953

- 显著性阈值 $|Z| > 3$



连续性纯合片段 (runs of homozygosity, ROHs) 是指一类基因组中出现的连续不间断的纯合现象，表现为一段染色体区域缺乏杂合子。

ROH分析对于我们理解群体历史、评估近交水平、解析复杂性状和疾病的遗传机理、以及检测基因与性状之间的关联都有非常重要的作用。已有研究发现ROH与精神分裂症、老年痴呆和多种肿瘤均存在关联。

ROH片段的长短可以推断近交历史。长的ROH片段反映最近世代发生过近交，而短的ROH说明较远世代产生近交，因为世代数越短，ROH片段被重组打断的可能性就越小。

ROH检测方法：

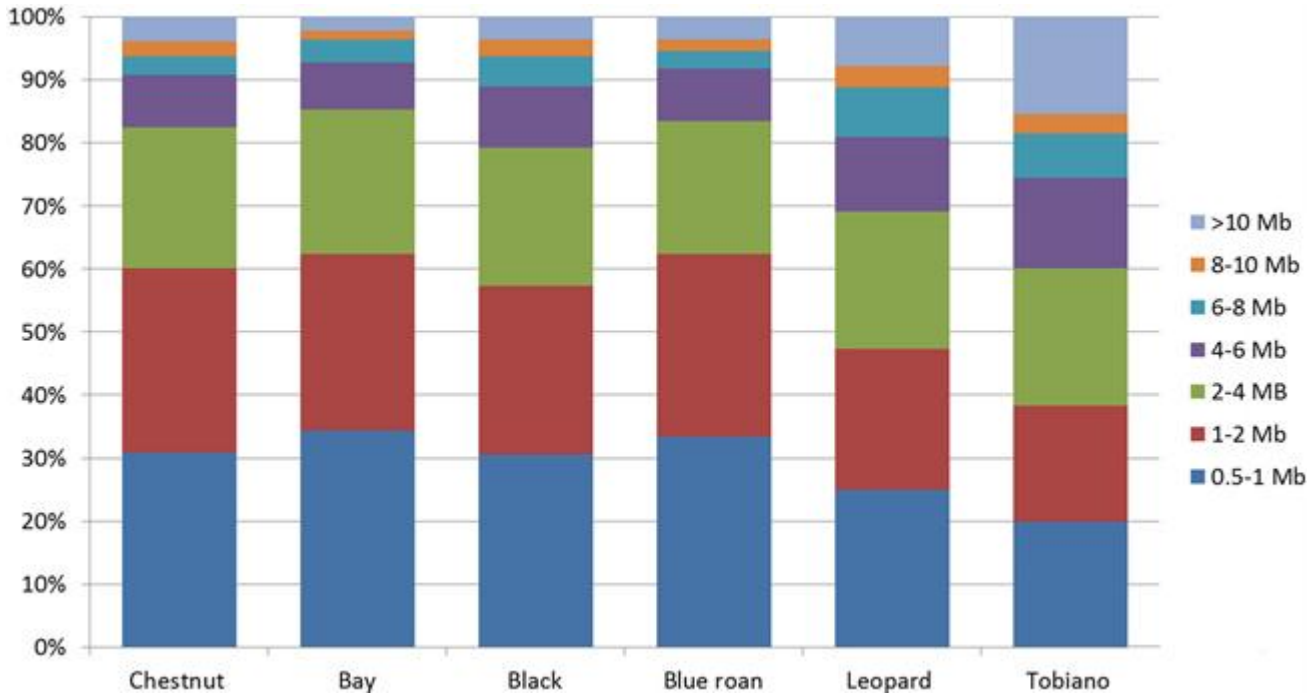
在基因组某一段区域内，当一定数量或一定密度的SNPs均表现为纯合时，可以判断该区域存在ROHs现象，该区域即为ROH片段。

风险提示： 简化数据检测ROHs很不可靠。

计算纯合片段 (ROH)



id	group	sum	Froh_genome
00494633	Feral	461251	0.00176111108538123
00495003	Feral	464666	0.00177414996086676
00495010	Feral	404745	0.00154536446805021
00495014	Feral	291047	0.00111125200393485
10685	Feral	427908	0.00163380355234636
10691	Feral	844008	0.00322252275865081
10692	Feral	378152	0.00144382923648747
51-16	Domestic	1090795	0.00416478482730319





总结：哪些分析简化数据可以做，哪些分析不能或风险很大？

定性类分析（比较不同组的大小、看不同组间的相对关系）**可以用简化数据做：**

例如：群体分层分析、基因流/渗入、LDdecay分析判定不同组间的衰减速度快慢、计算 F_{st} 、计算 π 比较不同组的多样性高低。

定量类分析（计算数值类、基于高密度的准确snp的算法类分析）**最好不要用简化数据做：**

例如：LDdecay分析计算每个群体的基因组衰减距离、计算 π 获得每个群体的多样性数值、种群动态历史分析、ROH分析、选择性清除分析等。



雲南農業大學
Yunnan Agricultural University

云南省生物大数据重点实验室

生信帮，答你所问