



论文

线粒体DNA基因组揭示的白马人群的遗传结构

毛乙臻, 王飞, 刘雨佳, 陈紫薇, 张金柠, 杨亚军*, 郑鸿翔*

复旦大学生命科学学院, 遗传工程国家重点实验室, 现代人类学教育部重点实验室, 上海 200438

* 联系人, E-mail: yyj229@263.net; zhenghongxiang@fudan.edu.cn

收稿日期: 2024-04-11; 接受日期: 2024-05-21; 网络版发表日期: 2024-10-21

国家自然科学基金(批准号: 32270670)和国家重点研发计划(批准号: 2022YFC3321200, 2022YFC3321201)资助

摘要 白马人是中国最古老的人群之一, 尽管历史和文化等方面的研究提示了白马人与青藏高原相关人群之间的紧密联系, 但是遗传学证据仍然缺乏. 本研究采集49例白马人样本并进行线粒体DNA基因组测序. 同时, 本文额外整合95个人群总计5676个亚洲个体的线粒体DNA单倍群数据进行主成分分析, 发现白马人与青藏高原相关人群最为接近, 提示白马人与青藏高原相关人群相近的遗传关系. 随后, 本研究对白马人群、青藏高原相关人群和汉族共18个人群进行序列多态性分析, 发现白马人群的遗传多样性最低, 且远远低于青藏高原相关人群和汉族. 同时, 本研究挑选出白马人中6个主要的亚单倍群(频率>4%), 分别为A21, D4j1a1, D4b2b+T3398C+A390G+A15613G, M9a1a2, M62b1a和D5b1b. 本文对这6个单倍群构建进化网络, 并估计最近共祖时间. 研究发现白马人6个主要单倍群中有3个单倍群与青藏高原人群特异支系相关, 同时这6个主要单倍群中的4个单倍群都起源于新石器时期及之后时期. 综上所述, 本文推测白马人群与青藏高原相关人群关系密切, 且白马人群的主体形成可能与新石器时期农业发展后青藏高原以及周边区域更为频繁的人群活动相关.

关键词 白马人, 东亚人群, 人类进化遗传学, 线粒体DNA

1 研究背景

白马人是中国最古老的人群之一, 主要居住在中国甘肃省文县和四川省平武、九寨沟等地, 总人口约2万多人. 20世纪50年代, 白马人被划分为藏族的一个分支, 主要依据是他们与松藩、虎牙等藏民生活区毗邻^[1], 故亦称白马藏族. 藏族学者毛儿盖·桑木旦先生也指出白马人在风俗、建筑上与藏族人相似, 如白马藏人头戴毡帽上插羽毛的做法与《白史》中记载的藏族人戴毡帽的习俗相一致; 白马人住房的建筑格局也

与毛尔盖、阿坝等地区的藏族人居住房屋相似^[2]. 同时, 在语言学的研究中, 张济川认为白马语是藏语其中的一个方言, 与藏语中的康方言接近^[3-5]. 但是关于白马人的族源和分类, 学术界还存在诸多争论, 许多学者认为白马人并不属于藏族, 其中羌族学说和氐族学说是两个主要的观点. 研究发现白马人与氐族在分布区域、婚姻制度、取名习惯、农业生产方面十分相似^[6,7], 在语言上白马语也是不同于藏语的独立语言^[8], 从而认为白马人可能是白马氏的后裔. 也有学者认为白马人是古代“宕昌羌”的后裔, 属于羌族, 我国古代

引用格式: 毛乙臻, 王飞, 刘雨佳, 等. 线粒体DNA基因组揭示的白马人群的遗传结构. 中国科学: 生命科学, 2025, 55: 153-167

Mao Y Z, Wang F, Liu Y J, et al. The genetic structure of the Baima population revealed by mitochondrial DNA genomes (in Chinese). Sci Sin Vitae, 2025, 55: 153-167, doi: [10.1360/SSV-2024-0108](https://doi.org/10.1360/SSV-2024-0108)

文献中也有称白马人祖先为白马羌的记载^[9]。虽然上述研究中提出不同的观点,但它们共同强调白马人与藏族、羌族等生活在青藏高原及周边区域的人群之间的联系。

尽管历史和文化等方面的研究提示白马人与青藏高原相关人群之间的紧密联系,但是遗传学证据仍然缺乏。线粒体DNA具有突变率高^[10]、基因组小、母系遗传^[11]、高拷贝数和不发生重组等特点,是遗传学研究中的重要研究对象之一^[12]。这些特性使得线粒体DNA在藏族等少数人群的研究中发挥重要的作用。如Zhao等人^[13]对680个来自西藏及周边地区的线粒体DNA样本进行分析,发现多数藏人中的线粒体DNA谱系可以追溯到东亚北部,揭示汉族和藏族之间存在紧密的遗传关系。Kang等人^[14]对七个高原人群的线粒体DNA进行研究,发现向青藏高原的人群迁徙主要分为三个阶段:末次冰盛期前(距今约2万年前);末次冰盛期后旧石器时代(约距今2万年至1万年前)和新石器时期。并且,在新石器时期的扩张人群形成现代藏人的主体遗传构成。Li等人^[15]通过分析8277个藏族人群线粒体DNA数据,发现新石器时代的小米农民通过采用大麦农业,最终登陆青藏高原高海拔地区(海拔2500米以上)并永久定居。这些研究表明线粒体DNA在中国青藏高原相关人群的人类起源研究中的重要作用,为深入探索白马人及与青藏高原相关人群之间的遗传关系提供可能。

在本研究中,共获得49个白马藏人的线粒体DNA全序数据,此外,通过对已发表的文献和在线数据库进行挖掘,本研究额外整合包含藏族、汉族等95个亚洲人群共计5676个线粒体DNA单倍群数据,另外还整合188例线粒体DNA全序数据。基于这一数据集,本研究旨在从线粒体DNA的角度解析白马人的遗传结构,进一步探究白马人与青藏高原相关人群之间的遗传关系,及白马人的形成和定居过程,以揭示这一独特人群的遗传起源和演化历史。本研究目标为白马人的进化历史提供一些遗传学证据,从而为研究东亚地区的人类遗传多样性和人群演化提供新的线索。

2 材料与方法

2.1 样本来源

本研究共采集49个来源于四川省平武县的白马人

样本进行线粒体DNA全序测序,包括10名男性和39名女性,所选样本均为正常个体,无亲缘关系。受试者均已签署知情同意书,且本研究已获复旦大学伦理委员会批准。此外,为探究白马人的母系遗传结构,本研究从在线数据库和已发表文献中收集188个亚洲人群的线粒体DNA全序样本,其中包括东亚人群(多为中国北方汉族人群)、青藏高原相关人群(基本为藏族,还有少数僮人及生活在印度东北部的藏缅语族人群,其中Ladakh人群为Genbank数据库收录的未发表数据)及其他人群(来自中亚、东南亚及北亚等地区)^[13-36]。

同时,本研究从在线数据库、已发表文献和千人基因组计划中额外整合5676个亚洲个体线粒体DNA单倍群数据,来自于95个人群,根据样本来源分为7个类群,包括青藏高原相关人群(那曲、昌都、日喀则、拉萨、阿里、林芝、山南等地区的藏族、夏尔巴人、僮人及居住在拉达克地区的藏缅语族人群等)、东亚北部人群(中国北方汉族)、东南亚南部人群(中国南方汉族)、东亚东北部人群(日本千叶和东京人群、韩国人群)、中亚人群(中国新疆柯尔克孜人、中国新疆塔吉克人、俄罗斯阿尔泰地区人群、吉尔吉斯斯坦人群和塔吉克斯坦人群)、北亚人群(俄罗斯西伯利亚东部及东北部的楚科奇人、雅库特人、鄂温克人、科里亚克人及中国内蒙古自治区达斡尔族等)和东南亚人群(主要包括泰国的拉瓦族、孟族及中国台湾省的阿美族、泰雅族、布农族、达悟族、马卡道族、排湾族等)。

2.2 线粒体DNA提取和分型

使用DNA提取试剂盒从49个受试白马人的血液样本中提取基因组DNA。提取出的基因组DNA质量由以下两种方法进行质检:(i) 1%琼脂糖凝胶电泳检测DNA完整性和纯度,(ii) 分光光度计检测DNA浓度,选取高品质DNA样本用于构建测序文库。质检后的DNA通过超声波打断,基因组DNA被随机打断成180~280 bp的片段,对DNA片段依次进行修复平末端、添加dA尾和连接Illumina测序接头。连接后的DNA片段通过PCR进行富集。富集后的文库用带生物素标记探针进行液相杂交,然后用链霉素磁珠捕获外显子,通过PCR进行线性富集。最后,对富集到的文库进行磁珠纯化和定量,然后在Illumina测序平台上进行测序。

将原始测序片段导出为Fastq文件,首先采用fastp

软件^[37]对原始Fastq文件进行质控, 去除测序接头和低质量的测序读长. 然后使用bwa将测序读长与修订的剑桥参考序列(revised cambridge reference sequence, rCRS)^[38]进行比对, 生成线粒体DNA基因组的二进制序列比对文件(Binary Alignment/Map, BAM). 其中平均测序深度为 $(19140.35 \pm 5229.36) \times$, 最高为 $27600.20 \times$, 最低为 $6227.30 \times$. 通过Picard中的MarkDuplicates程序(<http://broadinstitute.github.io/picard/>)去除重复的读长片段, 并使用GATK^[39]对线粒体DNA序列进行局部重新比对. 随后通过SAMtools^[40]生成测序片段堆叠文件(Pileup), 根据堆叠文件获得一致序列(consensus sequence), 之后手动检查插入和缺失.

得到的完整序列通过MUSCLE v3.8.3.158^[41]与rCRS进行对齐, 并在Mega5^[42]中手动检查305~315, 513~524, 8281~8289, 16183~16193区域的比对情况进行校正, 然后根据PhyloTree Build 17^[43]分配单倍群.

2.3 数据分析

(1) 主成分分析(principal component analysis, PCA). 本研究从在线数据库、已发表文献和千人基因计划中收集包括青藏高原相关人群、东亚北部人群、东亚南部人群、东亚东北部人群、中亚人群、北亚人群和东南亚人群这7个类群的5676个随机人群样本的线粒体DNA单倍群数据. 首先对收集的数据进行质控, 随后, 对所有人群样本按照所属单倍群进行分类, 并将其归并为十八类单倍群, 分别是: A, B4, B5, C, D4, D5', 6, M7, M8, M9, M10, M11, M12'G, N9, R9, R11'B6, Z, M*和N*. 计算各人群中的单倍群频率. 根据各人群的单倍群频率在R(v 4.3.1)中使用prcomp函数进行降维并计算每个主成分的方差解释度, 并使用ggplot^[44]进行可视化.

(2) 序列多态性分析. 利用DnaSP 5.10^[45]进行人群的序列多态性分析, 用来分析的数据集中包括本研究中的白马人群以及随机人群数据集中包括的青藏高原相关人群^[14,23,46]和汉族人群, 对数据集中包括的人群进行序列多态性分析, 分析的指标包括单倍型数、分离位点数、单倍型多样性(H)、核苷酸多样性(π)、群体参数 θ_s 和Tajima's *D*. 其中单倍型数指群体中观察到的不同单倍型的总数; 分离位点指的是群体中至少有两种(或以上)等位基因存在的DNA序列位置, 分离位点数可以衡量群体中遗传变异的多样性程度; 单倍

型多样性(H)指的是单一群体中随机选择两条序列单倍型不同的概率, 是度量群体变异程度的指标^[47]; 核苷酸多样性(π)表示群体内两条DNA序列间的碱基差异的平均值, 用于表征群体的遗传多样性水平^[47]; 群体参数 θ_s 基于分离位点数*S*进行估计, 表示每个DNA序列位点的群体平均突变率^[47,48]; Tajima's *D*检验是由Tajima在1989年提出的统计量, 用于检验DNA序列在进化过程中是否符合中性演化的预期^[49].

(3) 系统发育树的构建. 为更加直观地观察白马人群中的单倍群构成, 本文使用49个白马人群的线粒体DNA全序数据构建白马人群的进化网络, 进化网络采用Network v4.6^[50]软件, 使用中接法(Median-joining)进行构建, 通过整合进化网络和PhyloTree Build 17中的拓扑结构, 本文基于最大简约法最终手动生成49个白马人群的系统发育树, 并进行手动校正并调整.

(4) 进化网络的构建. 本研究选取白马人群中单倍群频率>5%的单倍群作为目标单倍群, 基于此, 筛选出单倍群M9a1a2, D4j1a1, D4b2b+T3398C+A390G+A15613G(D4b2b下游支系, 在PhyloTree Build 17中尚未定义该单倍群, 暂以突变进行命名)、A21和D5b1b, 另外, 在白马人的样本中发现单倍群M62b1a的频率为4.1%, 单倍群M62b1a是单倍群M62的下游支系, M62单倍群被认为起源于末次冰盛期之前的青藏高原, 可能代表青藏高原人群旧石器时期定居的人群, 与高原人群联系紧密^[31], 所以本研究决定将M62b1纳入目标单倍群中. 因此, 本研究最终确定的6个目标单倍群包括M9a1a2, D4j1a1, D4b2b+T3398C+A390G+A15613G, A21, D5b1b以及M62b1a.

进化网络采用Network v4.6软件, 使用中接法(Median-joining)构建, 并进行手动检查. 进化网络中包括216个属于目标单倍群的样本, 其中28个样本来自白马藏族个体, 另外188个样本来自于已收集的线粒体DNA全序数据集, 已收集的样本按照地区进行分类, 分为青藏高原相关人群、东亚人群、东南亚人群、北亚人群和中亚人群.

(5) 最近共祖时间估计. 对6个目标单倍群M9a1a2, D4j1a1, D4b2b+T3398C+A390G+A15613G, A21, D5b1b, M62b1a及其支系进行共祖时间的估计, 使用 ρ 统计量进行估计, 标准偏差(standard deviation, SD)的计算遵循Saillard等人^[51]提出的方法. ρ 和SD的计算在Network v4.6^[50]中完成, 获得的 ρ 和SD值基于完整线粒

体基因组的Soares突变速率对目标单倍群及其支系的最近共祖时间进行估计^[52]。

3 结果与分析

3.1 主成分分析

为探究白马人群与亚洲各人群之间的遗传关系,本研究采集49个白马人的样本并进行线粒体DNA全基因组测序。同时本研究收集5676个亚洲个体总计95个人群的线粒体DNA的单倍群数据,根据样本来源分为7类人群(青藏高原相关人群、东亚北部人群、东亚南部人群、东亚东北部人群、东南亚人群、中亚人群、北亚人群)。对这些人群的单倍群频率进行统计,发现在白马人群中单倍群D4, A和M12'G频率最高,分别为30.61%, 24.49%和12.24%,而单倍群A, D4, M12'G均在青藏高原相关人群、东亚北部人群和北亚人群中以较高频率存在,而在东亚南部和东南亚人群中以低频存在,呈现北高南低的地理变化趋势,表明白马人群在遗传上可能与东亚北部人群联系更加紧密。

进一步根据各人群的单倍群频率进行主成分分析,并利用PC1(33.73%)和PC2(18.55%)绘制散点图,结果如图1所示。PCA图表明,PC1解释地理上从东向西的人群频率差异,而PC2则解释从南到北种群的人群频率差异。本研究观察到白马人群和青藏高原相关人群、北亚人群、东亚北部人族和东亚东北部人群在主成分分析图中的右下方紧密聚集,表明这些人群之间存在较近的遗传距离,反映这些人群间一部分单倍群的频率较为接近;而东亚南部人群和东南亚人群聚集在主成分分析图中的右上方,中亚人群大部分分布于图的左上方,说明这些人群与白马人之间的遗传关系较为疏远。

在主成分分析图中,白马人群与青藏高原相关人群中的门巴人的距离最近,说明白马人群与门巴人群之间存在着最强的遗传联系。通过分析两个人群的单倍群频率分布,本文发现单倍群A在白马人群和门巴人群中的频率分别为24.49%和17.65%,单倍群M12'G的频率分别为12.24%和11.76%,均显示出较高的分布水平。同时,在两个人群中单倍群B4, B5, C, M, M10, M11, M7, M8, N, N9, R9, R11'B6和Z频率都小于5%,显示出较低的分布水平。这些证据表明两个人群在遗传组成上的共性。不同的是,在白马人群中D4单倍群

占30.61%,是该群体中频率最高的单倍群,而在门巴人群中D4的频率仅为5.88%。相反,单倍群D5'6和M9在门巴群体中以23.53%和41.18%的高频率出现,而在白马人群中这两个单倍群的频率较低,仅为6.12%。这显示出两个群体在遗传上也存在一定不同。

基于主成分分析,本研究发现青藏高原相关人群与白马人群间的密切联系,尤其是门巴人群与白马人群遗传关系最为接近。为探究人群间单倍群的共性与差异,本文对青藏高原相关人群与白马人群单倍群频率进行进一步分析(表1)。首先,结果发现A, D4在高原人群和白马人群中均显示出高频率,其中白马人群的D4最高,为30.61%,而高原人群的D4频率为4.65%~20.59%,在定日县的藏族中达到最低值。单倍群A在白马人群中频率为24.49%,在青藏高原相关人群中频率为2.20%~26.32%,在夏尔巴人中频率最高,僇人中频率最低,且远低于其他青藏高原相关人群。单倍群A和D4均为东亚北部的常见单倍群^[53],在青藏高原相关人群和白马人群中广泛分布,显示出一定的共性,表明这些群体可能具有相似的遗传成分。

白马人群和青藏高原相关人群中也存在一定差异,如单倍群M9是青藏高原相关人群中频率最高的单倍群(15.48%~41.18%),而在白马人群中M9单倍群的频率仅为6.12%,低于青藏高原相关人群中M9单倍群频率最低的群体(昌都_藏族, 表1)。单倍群的频率差异可能与地理隔离、不同的迁移历史或文化交流有关,反映特定地理或文化因素对遗传构成的影响。

3.2 白马人群序列多态性分析

主成分分析揭示白马人群和青藏高原相关人群存在着紧密的遗传联系,为进一步探究各人群的遗传信息,本研究对白马人群、青藏高原相关人群和部分汉族人群进行序列多态性分析,如表2所示。本研究中49个白马人序列可分为25个单倍型。单倍型多样性指的是单一群体中随机选择两条序列单倍型不同的概率。在白马人群中,单倍型多样性为0.95,青藏高原相关人群的单倍型多样性为0.96~1.00,汉族的单倍型多样性为0.99~1.00,白马人的单倍型多样性远远低于青藏高原相关人群和汉族的单倍型多样性。核苷酸多样性表示群体内两条DNA序列间的碱基差异的平均值,广泛用于表征群体的遗传多样性水平^[54]。白马人的核苷酸多样性为0.0020,仅仅高于林芝僇人群体和定日

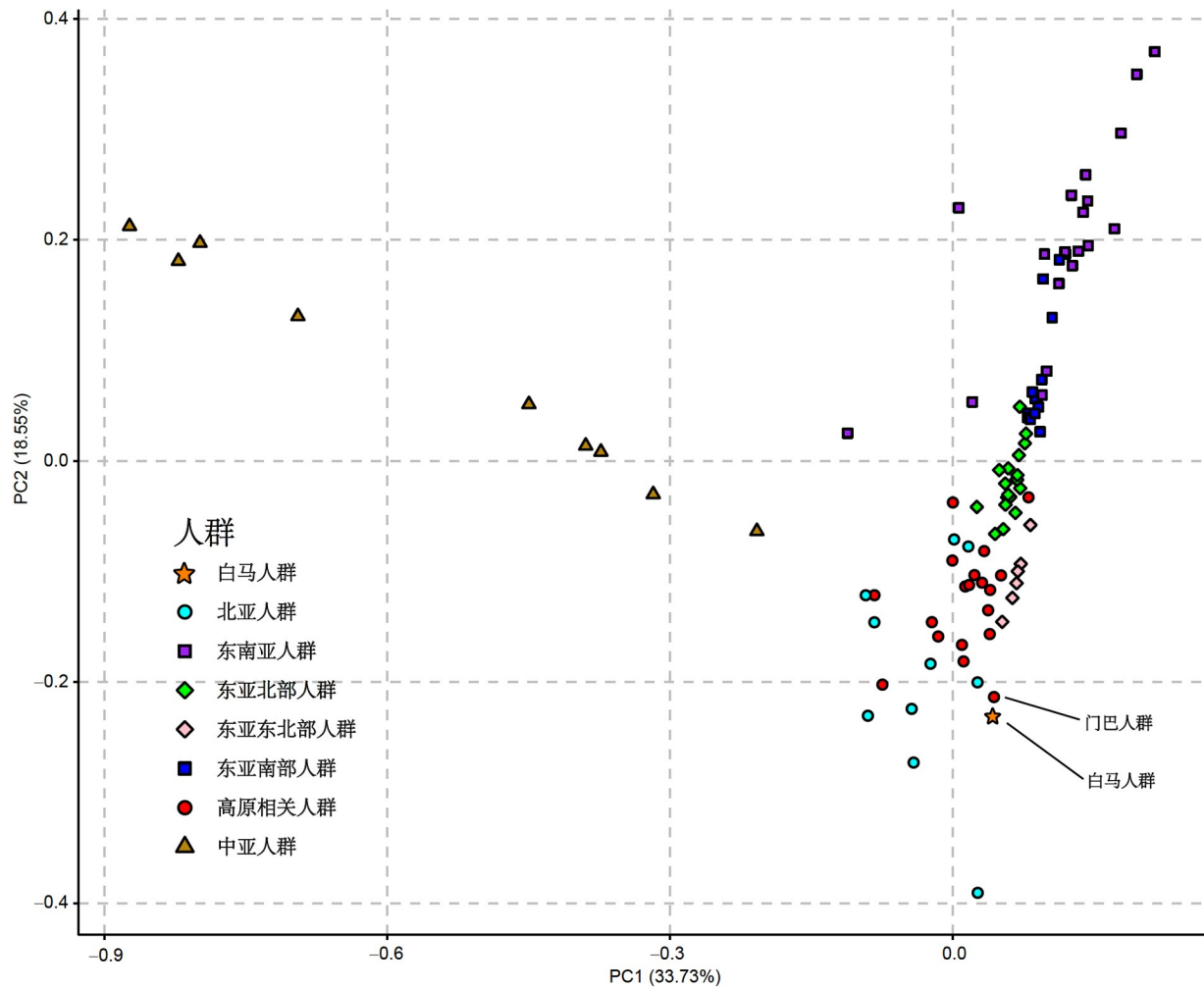


图 1 白马人与亚洲人群的主成分分析图。X轴代表PC1, Y轴代表PC2, 其中PC1解释了数据最大的方差, 占总方差的33.73%, PC2解释了数据的次大方差, 占总方差的18.55%。图例中青藏高原相关人群包括藏族、夏尔巴人、僮人等; 东亚北部人群指中国北方汉族; 东亚南部人群指中国南方汉族; 东亚东北部人群包括日本、韩国人群; 中亚人群包括阿尔泰人、维吾尔族等人群; 北亚人群包括俄罗斯楚科奇人、雅库特人等人群; 东南亚人群主要包括泰国的拉瓦族、孟族及中国台湾省阿美族等人群。图中与白马人群距离最近的人群为门巴人群, 已经标注

Figure 1 Principal Component Analysis of the Baima and Asian Populations. The X-axis represents PC1, and the Y-axis represents PC2, where PC1 explains the largest variance in the data, accounting for 33.73% of the total variance, and PC2 explains the second largest variance, accounting for 18.55% of the total variance. Tibetan Plateau related populations include Tibetans, Sherpas, Deng etc.; the northern East Asian population includes the Han Chinese in northern China; the southern East Asian population refers to the Han Chinese in southern China; the northeastern East Asian populations include Japanese and Korean populations; the Central Asian populations include Altai people, Uyghurs, etc.; the North Asian populations include Chukchi, Yakut populations, etc.; the Southeast Asian populations mainly include the Lawa and Mon of Thailand, as well as the Ami of Taiwan Province of China

藏族群体。汉族的三个群体具有最高的核苷酸多样性, 为0.0022~0.0023。此外, 群体参数 θ_s (表示每个DNA序列位点的群体平均突变率)在白马人群中为0.0036, 是所有分析群体中最低的, 而青藏高原相关人群的 θ_s 为0.0039~0.0057, 汉族则具有最高的 θ_s 值, 范围为0.0074~0.0090。在这些序列多态性分析指标中, 汉族基本表现为最高, 揭示汉族丰富的遗传多样性和较大

的有效群体。其次是青藏高原相关人群, 之前的研究表明, 高原祖先人群约距今8~5千年前从东亚祖先人群中分化出来, 并随后分化为藏人、夏尔巴人等高原人群^[55,56]。同时, 相较于平原地区, 青藏高原具有较为严苛的气候环境和相对隔离的地理区域^[57,58], 这些原因都使得青藏高原相关人群的有效群体较小。而汉族则基本代表东亚的遗传多样性。所以相较汉族而言,

表 1 白马人群及青藏高原相关人群的单倍群频率分布^{a)}
Table 1 Haplogroup frequency of the Baima and Tibetan Plateau related populations^{a)}

单倍群 (%)	白马人	巴青 藏族	昌都 藏族	僮人	嘉黎 那曲 藏族	康马 日喀则 藏族	拉达克 地区	拉孜 日喀则 藏族	拉萨 藏族	珞巴族	门巴族	阿里 藏族	林芝市 藏族	山南市 藏族	夏尔巴人	藏族_1 ^[21]	藏族_2 ^[36]	定日县 藏族	仲巴 日喀则 藏族	安多 那曲 藏族
A	24.49	16	10.71	2.2	20.45	19.48	16	14.67	5.88	16.48	17.65	12.5	16.9	8.11	26.32	12.82	8.82	12.79	10.42	9.52
B4	4.08	4	4.76	1.1	0	1.3	2	2.67	2.94	1.1	0	0	1.41	4.05	0	5.13	2.94	1.16	2.08	1.59
B5	4.08	0	1.19	0	0	1.3	0	0	0	0	0	0	2.82	1.35	0	0	2.94	1.16	0	0
C	0	4	3.57	2.2	0	3.9	14	13.33	7.35	3.3	0	0	9.86	4.05	18.42	5.13	2.94	1.16	10.42	0
D4	30.61	20	13.1	15.38	9.09	9.09	2	9.33	7.35	10.99	5.88	18.75	14.08	12.16	6.58	10.26	20.59	4.65	12.5	15.87
D5'6	6.12	2	3.57	8.79	0	2.6	2	1.33	0	12.09	23.53	0	4.23	1.35	3.95	5.13	0	6.98	2.08	3.17
M	4.08	8	10.71	10.99	11.36	9.09	20	9.33	14.71	4.4	0	0	4.23	10.81	14.47	12.82	2.94	10.47	12.5	9.52
M10	0	0	0	0	0	2.6	0	1.33	1.47	0	0	0	0	1.35	0	0	2.94	2.33	0	0
M11	0	0	3.57	1.1	0	1.3	0	1.33	1.47	2.2	0	0	0	0	1.32	2.56	2.94	0	0	3.17
M12'G	12.24	12	15.48	4.4	9.09	6.49	8	9.33	14.71	5.49	11.76	6.25	14.08	9.46	1.32	15.38	20.59	5.81	8.33	3.17
M7	2.04	0	0	0	2.27	0	2	0	0	0	0	6.25	1.41	0	0	0	0	0	0	1.59
M8	0	0	2.38	0	0	1.3	0	0	1.47	0	0	0	0	0	0	0	0	0	0	3.17
M9	6.12	16	15.48	39.56	22.73	22.08	20	22.67	2.5	16.48	41.18	31.25	15.49	31.08	18.42	15.38	23.53	36.05	31.25	22.22
N	0	6	0	0	6.82	2.6	10	5.33	1.47	1.1	0	6.25	1.41	1.35	7.89	2.56	2.94	3.49	2.08	7.94
N9	2.04	0	1.19	0	0	0	0	0	0	0	0	0	0	1.35	0	0	0	1.16	0	0
R11'B6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.16	2.08	0
R9	4.08	8	13.1	4.4	15.91	14.29	4	8	14.71	26.37	0	12.5	14.08	10.81	1.32	10.26	5.88	10.47	6.25	19.05
Z	0	4	1.19	9.89	2.27	2.6	0	1.33	1.47	0	0	6.25	0	2.7	0	2.56	0	1.16	0	0

a) 表头中有两个都为“藏族”，但来源于不同文献中的藏族，已标明来源。其余高原人群的样本来源于Kang等人^[14,23]在2013、2016年的工作和Li等人^[15]在2019年的工作

表 2 白马人群、青藏高原相关人群及汉族的序列多态性分^{a)}
Table 2 Sequence Polymorphism Analysis of the Baima, Tibetan Plateau Related Populations and the Han Chinese^{a)}

群体	样本量	单倍型数	分离位点数	单倍型多样性 (H)	核苷酸多样性 (π)	群体参数 θ_s	Tajima's <i>D</i>
白马人	49	25	266	0.95±0.0190	0.0020	0.0037±0.0010	-1.64
安多县_那曲市_藏族	63	51	396	0.99±0.0050	0.0021	0.0051±0.0014	-2.12*
巴青县_那曲市_藏族	50	49	384	1.00±0.0040	0.0021	0.0052±0.0014	-2.15*
昌都市_藏族	82	80	461	1.00±0.0020	0.0021	0.0056±0.0014	-2.18**
北京市_汉族	142	142	819	1.00±0.0008	0.0022	0.0090±0.0020	-2.50***
丹佛汉族	84	80	605	1.00±0.0020	0.0022	0.0074±0.0018	-2.41**
米林县_珞巴族	86	51	331	0.98±0.0060	0.0021	0.0040±0.0010	-1.62
林芝市_僮人	91	44	326	0.96±0.0100	0.0017	0.0039±0.0010	-1.94*
定日县_藏族	86	74	422	1.00±0.0020	0.0019	0.0051±0.0013	-2.13*
樟木镇_夏尔巴人	76	49	345	0.98±0.0080	0.0021	0.0043±0.0011	-1.74
中国南方汉族	114	105	686	0.99±0.0015	0.0023	0.0079±0.0019	-2.40**
嘉黎县_那曲市_藏族	44	37	306	0.99±0.0070	0.0021	0.0043±0.0012	-1.85*
康马县_日喀则市_藏族	77	72	429	1.00±0.0020	0.0021	0.0053±0.0014	-2.11*
拉孜县_日喀则市_藏族	75	71	455	1.00±0.0030	0.0021	0.0056±0.0014	-2.19**
拉萨_藏族	68	64	385	1.00±0.0030	0.0021	0.0049±0.0013	-2.03*
林芝市_藏族	71	68	436	1.00±0.0030	0.0021	0.0055±0.0014	-2.13**
山南市_藏族	73	71	449	1.00±0.0020	0.0020	0.0056±0.0014	-2.23**
仲巴县_日喀则市_藏族	48	45	319	1.00±0.0050	0.0020	0.0043±0.0012	-1.97*

a) (1) 本研究中参数 θ_s 表示每个DNA序列位点的群体平均突变率。(2) 高原相关人群的样本来源于Kang等人^[14,15]在2013、2016年的工作和Li等人^[23]在2019年的工作, 汉族样本来自千人基因组计划。(3) Tajima's *D*检验中, *, $P<0.05$; **, $P<0.01$; ***, $P<0.001$; 无*代表检验不显著

青藏高原相关人群的遗传多样性较低。白马人群在单倍型多样性、核苷酸多样性以及群体参数 θ_s 中均表现为最低水平, 反映其遗传多样性在分析的群体中是最低的。这提示白马人群可能经历了特定的进化事件, 如群体瓶颈效应或地理隔离, 这不仅使得白马群体的人口较少, 也减少与外部群体的遗传交流, 从而导致白马人群遗传多样性的相对较低。

Tajima's *D*检验分析的结果显示所有人群的Tajima's *D*值均小于0, 除白马人群、米林县的珞巴人群和樟木县的夏尔巴人群外, 其余人群的Tajima's *D*值均显著, 这可能表明汉族和大部分高原群体在近期发生扩张, 而白马人群、珞巴人群和夏尔巴人群人口总数较小, 生活的地区固定, 可能未发生扩张事件。

3.3 目标单倍群进化分析

为探究白马人群中单倍群的组成, 本研究使用49个白马人群样本的线粒体DNA全序数据构建系统发育树, 如图2所示。本研究发现白马人群中同时存在

M、N超级支系的样本, 其中属于M单倍群的样本(30个)多于属于N单倍群的样本(19个)。在M、N的下游, 白马人群主要分布在几个主要的单倍群中, 分别为A、D4、M12'G, 其中单倍群D4频率最高, 达到30.61%。

为进一步研究白马人群中的遗传信息及与青藏高原相关人群的联系, 本研究挑选出白马人群中的6个主体单倍群作为本研究重点关注的单倍群, 分别为A21, D4j1a1, D4b2b+T3398C+A390G+A15613G, M9a1a2, D5b1b和M62b1a, 频率都高于4%。其对应的基础单倍群分别为A, D4, M9, D5, 这些单倍群均在东亚北部呈高频, 是当地常见的单倍群^[53,59]。结合已收集的188个来源于亚洲的线粒体DNA全序数据对6个目标单倍群进行进化网络构建, 如图3所示。同时, 对目标单倍群及白马人所在支系进行最近共祖时间估计, 如表3所示。

单倍群A21的进化网络如图3A所示, 单倍群A21呈现扩张状态, 共包括39个样本, 样本都来源于汉族、高原相关人群和白马人群。白马人所在支系与单

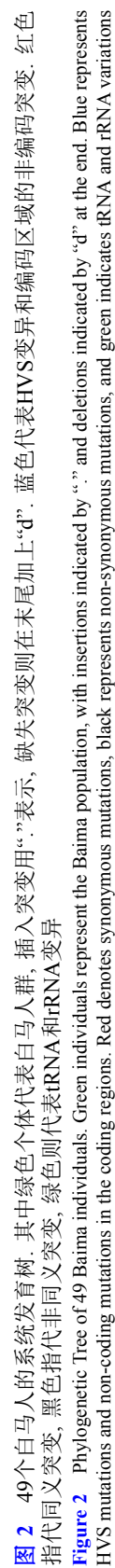


图 2 49个白马人的系统发育树, 其中绿色个体代表白马人群, 插入突变用“.”表示, 缺失突变则在末尾加上“d”, 蓝色代表HVS变异和编码区域的非编码突变, 红色指代同义突变, 黑色指代非同义突变, 绿色则代表tRNA和rRNA变异

Figure 2 Phylogenetic Tree of 49 Baima individuals. Green individuals represent the Baima population, with insertions indicated by “-” and deletions indicated by “d” at the end. Blue represents HVS mutations and non-coding mutations in the coding regions. Red denotes synonymous mutations, black represents non-synonymous mutations, and green indicates rRNA and tRNA variations

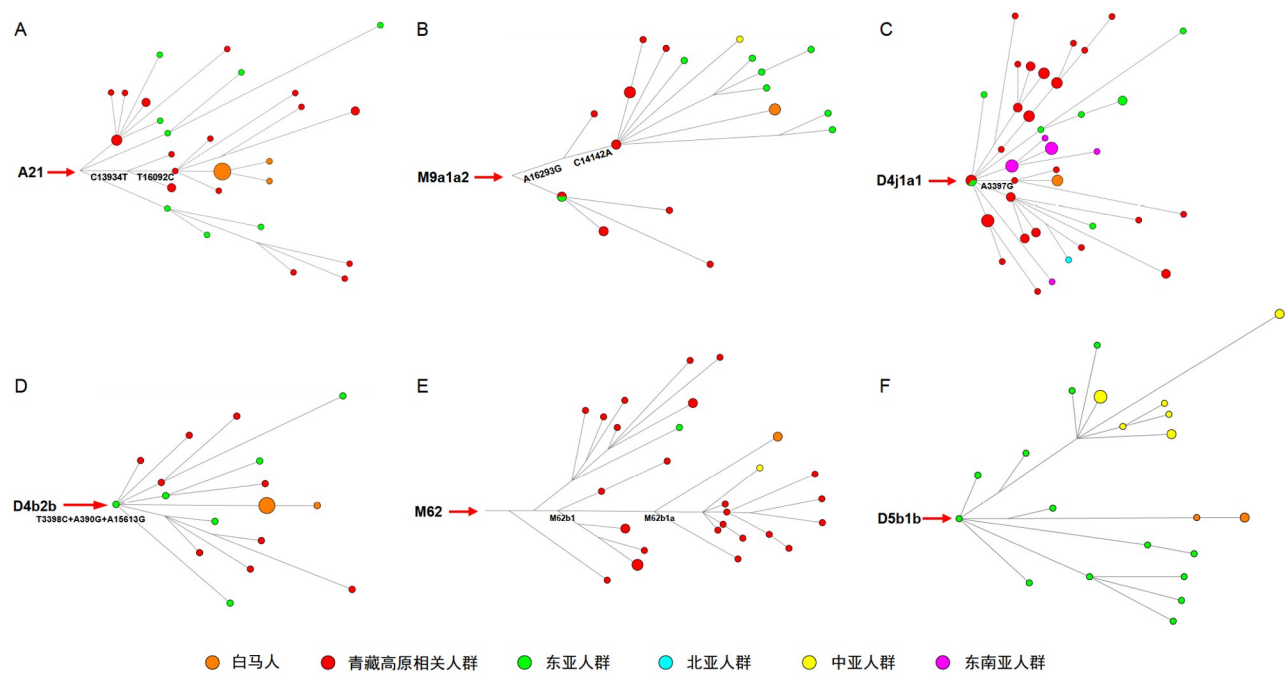


图 3 六个目标单倍群的进化网络图. 东亚样本来源于日本人、中国汉族和回族, 其中多为中国北方汉族人群; 青藏高原人群中包括绝大多数藏族, 还有少数僇人及生活在印度东北部的藏缅语族人群; 中亚人群包括新疆维吾尔族, 乌兹别克斯坦人等; 东南亚人群包括泰国、越南的拉瓦族、孟族等; 北亚样本来自俄罗斯西伯利亚. 箭头处为祖先单倍型

Figure 3 Median-joining networks of six major haplogroups. The East Asian populations include Japanese, Han Chinese and Hui Chinese, most of which are the Han populations in northern China; the Tibetan Plateau related populations include Tibetans, Deng and Tibeto-Burman groups living in the northeastern part of India; the Central Asian populations include the Uyghurs from Xinjiang and Uzbekistan etc.; the Southeast Asian populations include the Lawa and Mon populations from Thailand and Vietnam, etc.; the North Asian populations are from Siberia and Russia. The arrows indicate the ancestral haplotypes

表 3 六个目标单倍群及其亚支系的最近共祖时间估计^{a)}

Table 3 Estimates of the coalescence times of the most recent common ancestors of six major haplogroups and their sublineages^{a)}

单倍群	样本数	最近共祖时间(千年)
A21	39	8.53(4.07~13.11)
A21+C13934T+T16092C	20	4.28(0.85~7.79)
D4b2b+T3398C+A390G+A15613G	23	7.07(3.62~10.60)
D4j1a1	63	6.32(3.92~8.75)
D4j1a1+A3397G	6	3.46(0.28~7.00)
M9a1a2	27	9.84(3.78~16.13)
M9a1a2+A16293G+C14142A	18	6.10(3.17~9.09)
M62	40	26.56(15.46~38.18)
M62b1	22	19.01(10.32~28.07)
M62b1a	14	13.17(6.15~20.47)
D5b1b	27	11.42(6.13~16.88)

a) 最近共祖时间一列中, 括号内代表最近共祖时间的95%置信区间

倍群A21根节点相差2个突变C13934T和T16092C, 该支系中样本都来源于高原相关人群, 为高原人群的特

异支系. 本研究对单倍群A21及其支系进行最近共祖时间估计, 发现单倍群A21的最近共祖时间为距今

8.53(4.07~13.11)千年, 其亚支系A21+C13934T+T16092C的最近共祖时间为距今4.28(0.85~7.79)千年. 本文推测单倍群A21于新石器时期(约距今8.5千年前)在东亚北部起源, 其亚支系A21+C13934T+T16092C于距今四千年扩张并分化, 形成青藏高原相关人群的一支, 而A21+C13934T+T16092C的一个亚支系携带G15106A突变, 为白马人群特异性支系.

单倍群M9a1a2的进化网络如图3B所示. 单倍群M9a1a2的进化网络共包括25个现代样本, 其中有3个白马人样本、13个青藏高原相关人群样本(包括3个傣人和10个藏族人群)、8个东亚人群样本(包括7个中国汉族样本和1个中国回族样本)和1个维吾尔族样本. 该单倍群分为两个支系, 均为东亚人群和青藏高原相关人群混合的支系. 白马人群所在支系与单倍群M9a1a2根节点相差两个突变, 为A16293G和C14142A. 对单倍群M9a1a2及其支系进行共祖时间估计, 发现单倍群M9a1a2的最近共祖时间为9.84(3.78~16.13)千年, 其亚支系M9a1a2+A16293G+C14142A的最近共祖时间为6.10(3.17~9.09)千年. 本文推测M9a1a2于新石器时期(约距今9.8千年前)在东亚北部起源, 其亚支系M9a1a2+A16293G+C14142A于距今6.1千年扩张, 其中M9a1a2+A16293G+C14142A的一个亚支系携带T4454C, A16212G, C16245T突变, 最终分化为白马人的一支.

单倍群D4j1a1的进化网络如图3C所示, 单倍群D4j1a1的进化网络中共包括63个样本, 其中有3个白马人样本, 40个青藏高原相关人群样本, 8个中国汉族样本, 11个东南亚样本和1个北亚样本. 单倍群D4j1a1呈现出星状扩张, 白马人群集中在下游的一个青藏高原人群的特异支系中, 该支系与单倍群D4j1a1的根节点相差一个突变, 为A3397G. 支系D4j1a1+A3397G为高原人群特异支系, 其中包括3个藏族样本和3个白马样本. 1个来自于日喀则的藏族样本在该支系D4j1a1+A3397G的根节点上, 3个白马样本在其下游, 共享同一个单倍型, 与根节点相差一个突变C9533T. 对单倍群D4j1a1及其支系进行共祖时间估计, 发现单倍群D4j1a1的最近共祖时间为6.32(3.92~8.75)千年, 其亚支系D4j1a1+A3397G的最近共祖时间为3.46(0.28~7.00)千年. 本文推测D4j1a1起源于新石器时期(约距今6.3千年前), 从东亚北部扩张, 其亚支系D4j1a1+A3397G于距今三至四千年分化并形成青藏高原相关人群, 其中D4j1a1+A3397G的一个亚支系携带C9533T

突变, 为白马人群特异支系.

本研究在单倍群D4b2b的下游发现一个新的支系, 与D4b2b根节点相差三个突变, 为T3398C, A390G和A15613G. 在PhyloTree Build 17中, 该支系尚未被命名, 根据突变将该支系暂称为D4b2b+T3398C+A390G + A15613G. 本研究对这个支系构建进化网络, 如图3D所示. 单倍群D4b2b+T3398C+A390G+A15613G的进化网络共包括23个样本, 其中包括7个白马人、9个藏族人群和7个汉族样本. 该单倍群是一个汉族、藏族人群混合的支系, 有一汉族样本位于根节点. 所有白马人群聚集于同一支系, 该支系与单倍群根节点相差三个突变, 为A8308G, A8347C和T16519C. 对单倍群D4b2b+T3398C+A390G+A15613G进行最近共祖时间估计, 发现单倍群D4b2b+T3398C+A390G+A15613G的最近共祖时间为7.07(3.62~10.60)千年. 本研究推测单倍群D4b2b+T3398C+A390G + A15613G于新石器时期(约距今7千年前)在东亚北部起源. 随后该支系向高原扩张, 其中部分亚支系分化为藏人并最终定居于高原, 还有一亚支系携带A8308G, A8347C和T16519C突变, 形成白马人群特异性支系.

单倍群M62于末次冰盛期前在青藏高原起源, 是藏民中存在旧石器时代晚期定居者的证据^[13,31], 单倍群M62的进化网络如图3E所示. 单倍群M62的进化网络共包括40个样本, 基本来源于青藏高原相关人群, 是高原相关人群的特异单倍群. 2个白马人样本在支系M62b1a下游, 共享同一单倍型, 与M62b1a的根节点相差5个突变. 对单倍群M62及其支系进行共祖时间估计, 发现单倍群M62的最近共祖时间为26.56(15.46~38.18)千年, 提示该单倍群可能在末次冰盛期前起源于青藏高原人群中. 支系M62b1a的最近共祖时间为13.17(6.15~20.47)千年. 本研究推测单倍群M62b1a于新石器时期前(约距今13千年前)的起源于青藏高原相关人群中, 其亚支系携带G1719A, T11204C, A14274G, C16292T和C16429T突变, 为白马人群特异性支系, 同时表明白马人的形成与新石器时期之前已迁徙至高原的人群存在一定关联.

单倍群D5b1b的进化网络图如图3F所示. 单倍群D5b1b的进化网络共包括27个样本, 样本主要来源为东亚人群, 且东亚人群与中亚、白马人群相比, 距离单倍群D5b1b根节点的突变更少, 说明单倍群D5b1b最早在东亚地区起源与发展. 白马人群在单倍群D5b1b

的下游, 相差3个突变, 为C5899T, A8033G和T8793C. 对单倍群D5b1b进行共祖时间估计, 发现单倍群D5b1b的最近共祖时间为11.42(6.13~16.88)千年, 表明其起源于新石器时期之前, 其中一支随后分化形成白马人. 值得注意的是, 在单倍群D5b1b的样本中, 本文尚未发现青藏高原相关人群的样本, 说明该支系可能与青藏高原人群关系较为疏远.

对目标单倍群的进化网络进行分析, 本研究观察到这些单倍群样本主要来自于青藏高原相关人群(主要为藏族)和东亚人群(主要为汉族), 从而揭示白马人与青藏高原相关人群存在着紧密的遗传联系. 深入分析白马人群所在支系, 发现在单倍群A21, D4j1a1和M62b1a中, 白马人的样本都位于青藏高原相关人群的特异支系中; 在单倍群D4b2b+T3398C+A390G+A15613C和M9a1a2中, 白马人的样本位于青藏高原相关人群和汉族人群的混合支系中. 这一发现进一步证明白马人群与青藏高原相关人群在遗传上存在着密切相关性. 而在单倍群D5b1b中, 样本来源基本为汉族, 没有青藏高原相关人群样本, 提示单倍群D5b1b可能与青藏高原人群不相关. 研究还发现, 白马人通常具有独特的单倍型, 即白马人群之间共享一系列特异性突变, 而这些单倍型在其他人群中并未观察到, 表明白马人群在形成之后, 可能与外界存在着较少的基因交流.

对这6个单倍群中白马人样本分化的单倍群进行最近共祖时间估计, 本文发现单倍群A21+C13934T+T16092C的共祖时间为距今4.28(0.85~7.79)千年; 单倍群D4j1a1+A3397G的共祖时间为距今3.46(0.28~7.00)千年; 单倍群D462b+T3398C+A390G+A15613G的共祖时间为距今7.07(3.62~10.60)千年; 单倍群M9a1a2+A16293G+C14142A的共祖时间为距今6.10(3.17~9.09)千年; 单倍群M62b1a的共祖时间为距今 13.17(6.15~20.47)千年; 单倍群 D5b1b 的共祖时间为距今 11.42(6.13~16.88)千年, 提示6个主要单倍群中的4个单倍群都起源于新石器时期开始之后. 据此, 本文推测白马人群的主体形成可能在新石器时期开始之后, 与新石器时期的人群活动密切相关.

4 讨论

本研究首先对49个白马人的线粒体DNA进行单

倍群频率分析, 发现D4, A和M12'G在白马人群中显示出较高频率, 分别为30.61%, 24.49%和12.24%, 而这些单倍群在青藏高原相关人群、东亚北部人群和北亚人群中同样呈现较高频率. 为进一步探究白马人群与青藏高原相关人群之间的联系, 本文额外整合5676个亚洲个体的线粒体DNA单倍群数据进行主成分分析. 本研究发现白马人群与青藏高原相关人群遗传关系较为紧密, 而与中亚人群、东南亚人群及中国南方人群的遗传关系较疏远.

在主成分分析中, 结果发现白马人与青藏高原相关人群之间的遗传联系紧密, 为进一步探究这些人群间的关系, 本研究挑选青藏高原相关人群和汉族, 与白马人群共同进行序列多态性分析, 观察到白马人群的单倍型多样性、核苷酸多样性和群体参数都低于青藏高原相关人群和汉族人群, 证明白马人群存在着较低的遗传多样性, 可能是因为地理隔离与外界人群基因交流较少.

基于白马人群的单倍型频率, 本研究挑选出6个白马人群中主要的亚单倍群进行进化网络的构建, 分别为A21, D4i1a1, D462b+T3398C+A390G+A15613G, M9a1a2, M62b1a和D5b1b. 本文发现在单倍群A21, D4j1a1和M62b1a中, 白马人的样本都定位于青藏高原相关人群的特异支系内, 从青藏高原相关人群的特异支系中分化; 在单倍群D4b2b+T3398C+A390G+A15613G和M9a1a2中, 白马人样本分布于青藏高原人群和汉族人群共同构成的支系中; 而单倍群D5b1b中未发现青藏高原相关人群样本. 白马人群的6个主要支系中有3个支系与青藏高原人群特异支系相关, 这个结果从支系层面进一步揭示白马人群与青藏高原相关人群关系密切, 这与主成分分析中得出的结果一致.

在时间尺度上, 目标单倍群的时间估计提示白马人可能由两个阶段向青藏高原地区迁徙的人群构成. 第一个阶段是新石器时期之前, 对应的单倍群主要为M62b1a和D5b1b, 单倍群M62起源于末次冰盛期前的高原上, 此时高原上只有零星的人类活动痕迹, 如在曲桑遗址发现的距今约两万一千年的类人脚印^[60], 此时的人类属于原始的狩猎采集者^[13,31], 活动方式为季节性游猎. 支系M62b1a在新石器时期前发展, 该支系中的一支亚支系最终形成白马人. 单倍群D5b1b可能起源于新石器时期前的东亚平原人群中, 随后一支

系分化为白马人。然而,新石器时期前的人群对白马人的基因贡献较为有限,主要的贡献来源于第二阶段,即新石器时期来临之后,对应的单倍群有A21, D4j1a1, D4b2b+T3398C+A390G+A15613G和M9a1a2。这些单倍群均起源、发展于新石器时期以及之后时期,其祖先节点A、D4、M9都是存在于中国北方的常见单倍群^[53]。中国北方是小米农业的起源地,大约在一万至六千年前,小米农业迅速发展并成为当地农民的主要作物^[61]。同时,之前的研究已经表明,中国北方的小米农民在新石器时期向西进行扩张,到达高原东北缘^[14,15,62],催生独特的高原史前文化,这些史前文明伴随着人类扩张,在人群流动中,部分支系可能最

终分化为白马人。但是由于样本较少,只能粗略估计这些单倍群的共祖时间在新石器时期内,证明其与新石器时期人群活动相关,而无法准确确定这些单倍群是否代表新石器时期小米农民向高原扩张的遗传痕迹。

综上所述,本研究从线粒体DNA的角度对白马人的遗传结构进行分析,不仅揭示白马人群与青藏高原相关人群之间存在紧密的遗传联系,而且还发现白马人群的主体形成于新石器时期后,可能与新石器时期农业发展后青藏高原以及周边区域更为频繁的人群活动相关,为理解白马人群的遗传结构和历史演变提供重要线索。

参考文献

- 1 Zhang X H, Yu K L, Yang Y J, et al. A study of physical characteristics of the Baima people in China (in Chinese). *Acta Anthropol Sin*, 2020, 39: 9 [张兴华, 宇克莉, 杨亚军, 等. 中国白马人的体质特征. *人类学学报*, 2020, 39: 9]
- 2 Zhang R F. A review of the ethnic affiliation of the Baima (in Chinese). *Ethnic Hist Stud*, 2010, 1: 13 [张瑞丰. “白马藏人”族属问题研究综述. *民族史研究*, 2010, 1: 13]
- 3 Zhang J C. Baima language and Tibetan (Volume I) (in Chinese). *Minor Lang China*, 1994, 2: 14 [张济川. 白马话与藏语(上). *民族语文*, 1994, 2: 14]
- 4 Zhang J C. Baima language and Tibetan (Volume II) (in Chinese). *Minor. Lang. China*, 1994, 3: 15 [张济川. 白马话和藏语(下). *民族语文*, 1994, 3: 15]
- 5 Qi K J. Evidential categories in Baima language and Tibetan dialects (in Chinese). *Minor Lang China*, 2008, 3: 8 [齐卡佳. 白马语与藏语方言的示范范畴. *民族语文*, 2008, 3: 8]
- 6 Sun H K. The Di Population in history and the Baima in the Sichuan-Gansu region: a preliminary study on the ethnic affiliation of the Baima (in Chinese). *Ethno Natl. Stud*, 1980, 3: 11 [孙宏开. 历史上的氐族和川甘地区的白马人——白马人族属初探. *民族研究*, 1980, 3: 11]
- 7 Huang Y. An analysis of the ethnic origins of the “Baima Tibetans” (in Chinese). *J Lanzhou Univ (Soc. Sci.)*, 2002, 30: 8 [黄英. “白马藏人”族源探析. *兰州大学学报: 社会科学版*, 2002, 30: 8]
- 8 Sun H K. Is Baima language a dialect or a vernacular of Tibetan? (in Chinese). *Linguist Sci*, 2003, 1: 11 [孙宏开. 白马语是藏语的一个方言或土语吗? *语言科学*, 2003, 1: 11]
- 9 Ren N Q. A New Investigation Of Ancient History In Sichuan (in Chinese). Chengdu: Sichuan People's Publishing House, 1986 [任乃强. *四川上古史新探*. 成都: 四川人民出版社. 1986]
- 10 Forster P, Harding R, Torroni A, et al. Origin and evolution of native American mtDNA variation: a reappraisal. *Am J Hum Genet*, 1996, 59: 935
- 11 Wallace D C, Brown M D, Lott M T. Mitochondrial DNA variation in human evolution and disease. *Gene*, 1999, 238: 211–230
- 12 Yuan J J, Ye Z, Bu W J. Phylogeography of widespread species in Eurasia: current progress and future prospects. *Sci Sin Vitae*, 2019, 49: 1155–1164
- 13 Zhao M, Kong Q P, Wang H W, et al. Mitochondrial genome evidence reveals successful Late Paleolithic settlement on the Tibetan Plateau. *Proc Natl Acad Sci USA*, 2009, 106: 21230–21235
- 14 Kang L, Zheng H X, Zhang M, et al. MtDNA analysis reveals enriched pathogenic mutations in Tibetan highlanders. *Sci Rep*, 2016, 6: 31083
- 15 Li Y C, Tian J Y, Liu F W, et al. Neolithic millet farmers contributed to the permanent settlement of the Tibetan Plateau by adopting barley agriculture. *Natl Sci Rev*, 2019, 6: 1005–1013
- 16 Chandrasekar A, Kumar S, Sreenath J, et al. Updating phylogeny of mitochondrial DNA macrohaplogroup m in India: dispersal of modern human in South Asian Corridor. *PLoS ONE*, 2009, 4: e7447

- 17 Derenko M, Malyarchuk B, Grzybowski T, et al. Origin and post-glacial dispersal of mitochondrial DNA haplogroups C and D in Northern Asia. *PLoS ONE*, 2010, 5: e15214
- 18 Chiang C W, Mangul S, Robles C, et al. A comprehensive map of genetic variation in the world's largest ethnic group—Han Chinese. *Mol Biol Evol*, 2018, 35: 2736–2750
- 19 Duong N T, Macholdt E, Ton N D, et al. Complete human mtDNA genome sequences from Vietnam and the phylogeography of Mainland Southeast Asia. *Sci Rep*, 2018, 8: 11651
- 20 Auton A, Abecasis G R, Altshuler (Co-Chair) D M, et al. A global reference for human genetic variation. *Nature*, 2015, 526: 68–74
- 21 Hao M, Pu W, Li Y, et al. The HuaBiao project: whole-exome sequencing of 5000 Han Chinese individuals. *J Genet Genomics*, 2021, 48: 1032–1035
- 22 Ji F, Sharpley M S, Derbeneva O, et al. Mitochondrial DNA variant associated with Leber hereditary optic neuropathy and high-altitude Tibetans. *Proc Natl Acad Sci USA*, 2012, 109: 7391–7396
- 23 Kang L, Zheng H X, Chen F, et al. mtDNA lineage expansions in Sherpa population suggest adaptive evolution in Tibetan highlands. *Mol Biol Evol*, 2013, 30: 2579–2587
- 24 Ko C H, Yen J Y, Chen S H, et al. Evaluation of the diagnostic criteria of Internet gaming disorder in the DSM-5 among young adults in Taiwan. *J Psychiatric Res*, 2014, 53: 103–110
- 25 Kutanan W, Kampuansai J, Srikummool M, et al. Complete mitochondrial genomes of Thai and Lao populations indicate an ancient origin of Austroasiatic groups and demic diffusion in the spread of Tai-Kadai languages. *Hum Genet*, 2017, 136: 85–98
- 26 Kutanan W, Kampuansai J, Brunelli A, et al. New insights from Thailand into the maternal genetic history of Mainland Southeast Asia. *Eur J Hum Genet*, 2018, 26: 898–911
- 27 Zheng H X, Li L, Jiang X Y, et al. MtDNA genomes reveal a relaxation of selective constraints in low-BMI individuals in a Uyghur population. *Hum Genet*, 2017, 136: 1353–1362
- 28 Liu J, Wang L D, Sun Y B, et al. Deciphering the signature of selective constraints on cancerous mitochondrial genome. *Mol Biol Evol*, 2012, 29: 1255–1261
- 29 Marchi N, Hegay T, Menecier P, et al. Sex-specific genetic diversity is shaped by cultural factors in Inner Asian human populations. *Am J Phys Anthropol*, 2017, 162: 627–640
- 30 Mikami E, Fuku N, Kong Q P, et al. Comprehensive analysis of common and rare mitochondrial DNA variants in elite Japanese athletes: a case-control study. *J Hum Genet*, 2013, 58: 780–787
- 31 Qin Z, Yang Y, Kang L, et al. A mitochondrial revelation of early human migrations to the Tibetan Plateau before and after the last glacial maximum. *Am J Phys Anthropol*, 2010, 143: 555–569
- 32 Benson D A, Cavanaugh M, Clark K, et al. GenBank. *Nucleic Acids Res*. 2013, 41: D36–42
- 33 Sharma I, Sharma V, Khan A, et al. Ancient human migrations to and through Jammu Kashmir-India were not of males exclusively. *Sci Rep*, 2018, 8: 851
- 34 Tanaka M, Cabrera V M, González A M, et al. Mitochondrial genome variation in Eastern Asia and the peopling of Japan. *Genome Res*, 2004, 14: 1832–1850
- 35 Ueno H, Nishigaki Y, Kong Q P, et al. Analysis of mitochondrial DNA variants in Japanese patients with schizophrenia. *Mitochondrion*, 2009, 9: 385–393
- 36 Yi X, Liang Y, Huerta-Sanchez E, et al. Sequencing of 50 human exomes reveals adaptation to high altitude. *Science*, 2010, 329: 75–78
- 37 Chen S, Zhou Y, Chen Y, et al. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, 2018, 34: i884–i890
- 38 Andrews R M, Kubacka I, Chinnery P F, et al. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat Genet*, 1999, 23: 147
- 39 McKenna A, Hanna M, Banks E, et al. The genome analysis toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data. *Genome Res*, 2010, 20: 1297–1303
- 40 Li H, Handsaker B, Wysoker A, et al. The sequence alignment/map format and SAMtools. *Bioinformatics*, 2009, 25: 2078–2079
- 41 Edgar R C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, 2004, 32: 1792–1797
- 42 Tamura K, Peterson D, Peterson N, et al. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol*, 2011, 28: 2731–2739

- 43 van Oven M, Kayser M. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. [Hum Mutat](#), 2009, 30: E386–E394
- 44 Wickham H, Chang W, Wickham M H. Package ‘ggplot2’. Create elegant data visualisations using the grammar of graphics. R package Ver. 2016, 2: 1–189
- 45 Rozas J, Sánchez-DelBarrio J C, Messeguer X, et al. DnaSP, DNA polymorphism analyses by the coalescent and other methods. [Bioinformatics](#), 2003, 19: 2496–2497
- 46 Li Y C, Tian J Y, Liu F W, et al. Neolithic millet farmers contributed to the permanent settlement of the Tibetan Plateau by adopting barley agriculture. [Natl Sci Rev](#), 2019, 6: 1005–1013
- 47 Nei M. Molecular Evolutionary Genetics. New York: Columbia University Press, 1987
- 48 Watterson G A. On the number of segregating sites in genetical models without recombination. [Theor Population Biol](#), 1975, 7: 256–276
- 49 Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. [Genetics](#), 1989, 123: 585–595
- 50 Bandelt H J, Forster P, Rohl A. Median-joining networks for inferring intraspecific phylogenies. [Mol Biol Evol](#), 1999, 16: 37–48
- 51 Saillard J, Forster P, Lynnerup N, et al. mtDNA variation among greenland eskimos: the edge of the beringian expansion. [Am J Hum Genet](#), 2000, 67: 718–726
- 52 Soares P, Ermini L, Thomson N, et al. Correcting for purifying selection: an improved human mitochondrial molecular clock. [Am J Hum Genet](#), 2009, 84: 740–759
- 53 Xue F, Wang Y, Xu S, et al. A spatial analysis of genetic structure of human populations in China reveals distinct difference between maternal and paternal lineages. [Eur J Hum Genet](#), 2008, 16: 705–717
- 54 Nei M, Li W H. Mathematical model for studying genetic variation in terms of restriction endonucleases. [Proc Natl Acad Sci USA](#), 1979, 76: 5269–5273
- 55 Lu D, Lou H, Yuan K, et al. Ancestral origins and genetic history of Tibetan highlanders. [Am J Hum Genet](#), 2016, 99: 580–594
- 56 Zhang C, Lu Y, Feng Q, et al. Differentiated demographic histories and local adaptations between Sherpas and Tibetans. [Genome Biol](#), 2017, 18: 115
- 57 Bigam A W, Lee F S. Human high-altitude adaptation: forward genetics meets the HIF pathway. [Genes Dev](#), 2014, 28: 2189–2204
- 58 Beall C M. Adaptation to high altitude: phenotypes and genotypes. [Annu Rev Anthropol](#), 2014, 43: 251–272
- 59 Kong Q P, Yao Y G, Sun C, et al. Phylogeny of East Asian mitochondrial DNA lineages inferred from complete sequences. [Am J Hum Genet](#), 2003, 73: 671–676
- 60 Hu Z D. Atlas of Chinese Cultural Relics. Yunnan Volume (in Chinese). Kunming: Yunnan Science and Technology Press, 2001 [胡振东. 中国文物地图集. 云南分册. 昆明: 云南科学技术出版社. 2001]
- 61 Dong G H, Zhang S J, Yang Y S, et al. Agricultural intensification and its impact on environment during Neolithic Age in northern China. [Chin Sci Bull](#), 2016, 61: 2913–2925 [董广辉, 张山佳, 杨谊时, 等. 中国北方新石器时代农业强化及对环境的影响. 科学通报, 2016, 61: 2913–2925]
- 62 Chen F H, Dong G H, Zhang D J, et al. Agriculture facilitated permanent human occupation of the Tibetan Plateau after 3600 B.P. [Science](#), 2015, 347: 248–250

The genetic structure of the Baima population revealed by mitochondrial DNA genomes

MAO YiZhen, WANG Fei, LIU YuJia, CHEN ZiWei, ZHANG JinNing, YANG YaJun^{*} & ZHENG HongXiang^{*}

Ministry of Education Key Laboratory of Contemporary Anthropology, State Key Laboratory of Genetic Engineering, School of Life Sciences, Fudan University, Shanghai 200438, China

^{} Corresponding authors, E-mail: yyj229@263.net; zhenghongxiang@fudan.edu.cn*

The Baima is one of the oldest Chinese ethnic populations. Lines of historical and cultural evidence indicated close relationships between the Baima population and Tibetan Plateau related populations, however, there was little investigation on genetics. In this study, we collected 49 individuals from the Baima population and sequenced the mitochondrial DNA genomes. We conducted principal component analysis of the mitochondrial DNA haplogroup data of a total of 5676 individuals from 95 Asian populations, including the Baima population. As a result, the Baima population was closest to the Tibetan Plateau related populations, indicating a close genetic relationship between the Baima population and the Tibetan Plateau related populations. Subsequent analysis of sequence polymorphism across 18 populations (including Baima, Tibetan Plateau related populations and Han Chinese) showed the lowest genetic diversity in the Baima population, which is significantly lower than those of Tibetan Plateau related populations and Han Chinese. In addition, we identified six major sub-haplogroups (frequency >4%) among the Baima people (A21, D4j1a1, D4b2b+T3398C+A390G+A15613G, M9a1a2, M62b1a, and D5b1b). We constructed median-joining networks and estimated the coalescence times of the most recent common ancestors of these six haplogroups. We found that three out of six principal haplogroups of the Baima showed a close genetic relationship with specific lineages of Tibetan Plateau related populations, and four of the six major haplogroups originated after the advent of the Neolithic. Thus, we concluded that the Baima population has a close relationship with the Tibetan Plateau related populations and the formation of the Baima population may be associated with the development of agriculture after the advent of the Neolithic due to the increased human activity on the Tibetan Plateau and in the surrounding areas.

East Asian populations, Baima population, human evolutionary genetics, mitochondrial DNA

doi: [10.1360/SSV-2024-0108](https://doi.org/10.1360/SSV-2024-0108)