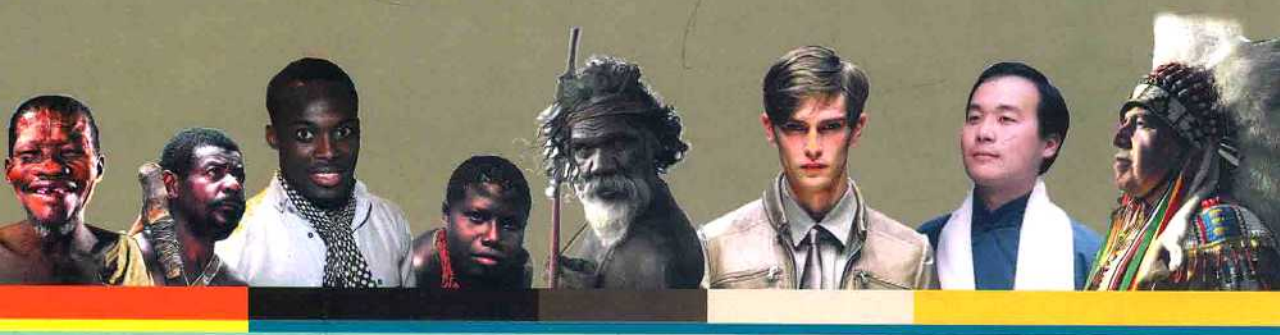
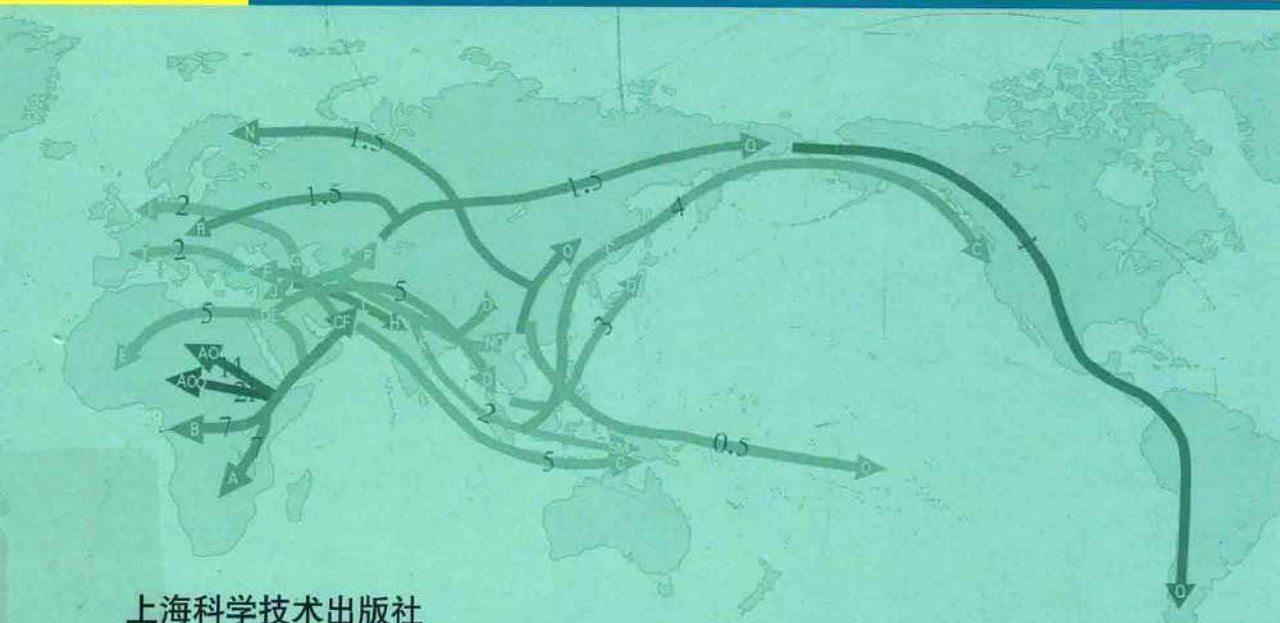
科学专著：生命科学研究

Y染色体与东亚族群演化

李辉 金力 编著





上海科学技术出版社

金力 中国科学院院士。复旦大学副校长、生命科学学院教授、博士生导师。1994年获美国德克萨斯大学生物医学/遗传学博士学位。其后在斯坦福大学从事博士后研究，曾任德克萨斯大学副教授、辛辛那提大学教授。现为德国马普学会外籍会员、国际人类基因组组织理事、上海市遗传学会理事长、上海市人类学会理事长，以及《基因组研究》等7家国际学术期刊的编委。主要研究方向为：医学遗传学及遗传流行病学、计算生物学、人类群体遗传学和基因组学。通过群体遗传学研究证实了东亚现代人的非洲起源，分析了东亚人群的适应性进化与流行病学特征。在《自然》、《科学》、《细胞》等发表论文200余篇。

李辉 复旦大学生命科学学院教授、博士生导师。2005年在复旦大学获得中国首个人类生物学博士学位，师从金力教授。其后在耶鲁大学从事博士后研究。现为亚洲人文与自然研究院副院长、中国肤纹学会副会长、中国青年科技工作者协会会员、上海人类学学会理事等。担任《调查遗传学》等6家国际学术期刊的编委。主要研究方向为：分子人类学、历史人类学、语言人类学、法医人类学。通过学科交叉研究，详细研究了东亚的现代人各地理种、民族系群、家族姓氏和古文明起源的遗传学基础。在《科学》、《自然》等发表论文150多篇。

责任编辑：包惠芳　张　帆　文字编辑：兰明娟　　封面设计：戚永昌

上架建议：人类学/生物学

ISBN 978-7-5478-2418-4

9 787547 824184 >

定价：198.00元

易文网：www.ewen.co

www.sstp.cn

Y染色体与东亚族群演化

李辉　金力 编著

Li Hui　Jin Li

上海科学技术出版社

Shanghai Scientific & Technical Publishers

图书在版编目(CIP)数据

Y染色体与东亚族群演化 / 李辉，金力编著. —上海：上海科学技术出版社，2015.10
（科学专著：生命科学研究）
ISBN 978-7-5478-2418-4

Ⅰ.①Y… Ⅱ.①李… ②金… Ⅲ.①Y染色体—关系—民族人类学—研究—东亚 Ⅳ.①Q982

中国版本图书馆CIP数据核字(2014)第241458号

本书出版受“上海科技专著出版资金”资助

审图号：GS(2015)1431号

责任编辑：包惠芳 张 帆
文字编辑：兰明娟
封面设计：戚永昌

Y染色体与东亚族群演化
李辉 金力 编著

上海世纪出版股份有限公司
上海科学技术出版社 出版
（上海钦州南路71号 邮政编码200235）
上海世纪出版股份有限公司发行中心发行
200001 上海福建中路193号 www.ewen.co
上海中华商务联合印刷有限公司印刷
开本 787×1092 1/16 印张 23.75 插页 4
字数 450千字
2015年10月第1版 2015年10月第1次印刷
ISBN 978-7-5478-2418-4/Q·29
定价：198.00元

内 容 提 要

Y染色体由于其单倍体特性和群体特异性分布，成为分子人类学研究最有力的工具。利用Y染色体可以很好地解析种族的起源、民族的分化、家族的传承。本书共分7章，系统介绍了Y染色体在东亚族群演化过程的分子人类学研究中取得的进展。从Y染色体总单倍群C、D、F、K在东亚的多样性分布，可知东亚人群由不同时期到达的澳大利亚、尼格利陀、蒙古利亚、高加索4个不同种族混合而成。由单倍群O的多样性分布可知东亚的9个民族类群(汉藏、侗傣、苗瑶、南岛、孟高棉、阿尔泰、乌拉尔、叶尼塞、古亚)的分化过程和系属关系。利用高分辨率的Y染色体标记，以曹操家族作为示范，展示了Y染色体在厘清和辨析家族谱系中的强大功能，开创了历史人类学的新领域。

本书是运用遗传学手段解析东亚族群历史的第一部专著，具有开创性价值。可供人类学、人类遗传学、医学等领域的研究人员及相关领域的研究生和高年级本科生参考。

《科学专著》系列丛书序

进入21世纪以来，中国的科学技术发展进入一个重要的跃升期。科学技术自主创新的源头，正是来自科学向未知领域推进的新发现，来自科学前沿探索的新成果。学术著作是研究成果的总结，它的价值也在于其原创性。

著书立说，乃是科学研究工作不可缺少的一个组成部分。著书立说，既是丰富人类知识宝库的需要，也是探索未知领域、开拓人类知识新疆界的需要。特别是在科学各门类的那些基本问题上，一部优秀的学术专著常常成为本学科或相关学科取得突破性进展的基石。

一个国家，一个地区，学术著作出版的水平是这个国家、这个地区科学研究水平的重要标志。科学研究具有系统性和长远性、继承性和连续性等特点，科学发现的取得需要好奇心和想象力，也需要有长期的、系统的研究成果的积累。因此，学术著作的出版也需要有长远的安排和持续的积累，来不得半点虚浮，更不能急功近利。

学术著作的出版，既是为了总结、积累，更是为了交流、传播。交流传播了，总结积累的效果和作用才能发挥出来。为了在中国传播科学而于1915年创办的《科学》杂志，在其自身发展的历程中，一直也在尽力促进中国学者的学术著作的出版。

几十年来，《科学》的编者和出版者，在不同的时期先后推出过好几套中国学者的科学专著。在20世纪三四十年代，出版有《科学丛书》；自20世纪90年代以来，又陆续推出《科学专著丛书》《科学前沿丛书》《科学前沿进展》等，形成了一个以刊物名字样科学为标识的学术专著系列。自1995年起，截至2010年“十一五”结束，在科学标识下，已出版了25部专著，其中有不少佳作，受到了科学界和出版界的欢迎和好评。

为了继续促进中国学者对前沿工作做有创见的系统总结，“十二五”期间，《科学》的编者和出版者决定对科学系列学术著作做新的延伸，将科学专著学术丛书扩展为三个系列品种，即《科学专著：前沿研究》《科学专著：生命科学研究》《科学专著：大科学工程》，继续为中国学者著书立说尽一份力。

随着中国科学研究向世界前列的挺进，我们相信，在科学系列的学术专著之中，一定会有更多中国学者推陈出新、标新立异的佳作问世，也一定会有传世的名著问世！

周光召

（《科学》杂志编委会主编）

2011 年 5 月

序

作为一名语言学家，我当然对语言的演化深感兴趣。语言演化最宏观也最难解的问题是语言怎么涌现，因为关于这个秘密的线索，被埋藏在至少几十万年前。近年来跨学科研究虽然带来不少新知识，可是距离找出一个完整的答案揭开谜底还是极为遥远。只有人类才有语言，同时只有靠语言这样万能的思想工具，人类才可能从猿人演化到现代人。语言和人类演化这两者的关系，几乎像是鸡与蛋的循环那样密切。为了强调这个关系，有人建议把智人（*Homo sapiens*）改成说话人（*Homo loquens*）[①]，因为我们的智慧毕竟有限，而每个人都会说话。

要想得到语言涌现的答案，我们必须从演化的几个不同时间尺度着手[②]，做跨领域的研究。明代末叶的陈第，研究三千年前《诗经》的上古音系，是最早清楚地说明语言总是跟着时间与空间不断演化的学者。但是若论用科学方法来研究大批不同语言间的谱系关系，这个开端必须归功于英国的W. Jones。两百多年前，他就设想欧洲和印度的很多语言都来自同一个源头，也就是古印欧语（Proto-Indo-European）。循着 Jones 的思路，许多语言学家陆续推论出不少和语族分类有关的研究成果。除了印欧语系，世界上目前所有的五六千种语言，很可能都归属于二三十个不同的大语系[③]。

世界上的这些语言，当然都是由不同族群在不同的时间发展出来的。同时他们也在不同的空间里留下生活的痕迹，如石器、陶器等。因此意大利遗传学家 L. L. Cavalli-Sforza 等建议把基因、考古及语言三种数据结合起来，彻底探究人类的演化史。虽然他们 1988 年的文章所能用到的基因材料相对有限[④]，可是这种跨学科的看法，是目前大家都能接受的。

其实达尔文早在 1859 年就说过，语言的谱系应当等于人群的谱系。可是他把话说得太简单了。T. Huxley 在 1865 年的一篇短文里扩充了达尔文

---

① Fry D. Homo Loquens: Man as a Talking Animal. Oxford: Cambridge University Press, 1977.

② 王士元. 语言演化的三个尺度. 科学中国人, 2013, 1: 16—20.

③ Greenberg J H. The methods and purposes of linguistic genetic classification. Language and Linguistics, 2001, **2**: 111－135.

④ Cavalli-Sforza L L, et al. Reconstruction of human evolution: bringing together genetic, archeological and linguistic data. Proceedings of the National Academy of Sciences, 1988, **85**: 6002－6006.

的说法[①]，指出一群人也可能改变自己的语言，而造成语言及族群两种谱系上的差别。说得更准确些，虽然传统的语言谱系树可以代表语言纵向传递的关系，可是事实上我们知道，人群接触时，语言成分的横向传递是经常有的现象。每个语言里都或多或少有借入的词汇或语法结构，这就是横向传递的结果。横向传递越多，语言谱系树所能呈现的语言关系的信息就越不周全。有时候，在人种及语言生态复杂的社会环境下，某个混合语的词汇可能绝大部分来自一种语言，而其语法结构却来自另一种语言。目前语言学家还没有找到理想的方法，能够把纵向和横向这两种传递的过程简单明了地整合起来。

复旦大学做这方面的研究可以说是得天独厚。现代人类学教育部重点实验室成立于 2005 年。金力教授与他 HUGO Pan-Asian SNP Consortium 的研究团队，在 2009 年的 *Science* 杂志上发表了一篇里程碑的报告[②]，取样了将近两千人的基因，详尽分析了整个亚洲人群的谱系。在这个基础上，李辉与金力两位教授又集中用 Y 染色体分析东亚族群的演化。

这本书写得深入浅出，就如李教授在前言里所承诺的，它会把读者带入演化历程的精彩世界。拜读这本书的当下，我已经开始享受书中多彩纷呈的世界了。特别是第 7 章，讨论语音系统的分布与人类扩张的过程，比 Q. Atkinson 发表在 *Science* 杂志里的理论更加精辟[③]，读后真是受益匪浅。此书材料非常丰富，让我总爱不释手地经常展读。尤其是研究民族语言或汉语方言时，这本书也是不可多得的参考书。

Y 染色体是人体细胞核内 46 条染色体的其中一条，是由父系一代一代地传递下去。可是细胞核外的线粒体 DNA，却是由母系传递的。这两种单亲传递的基因所给我们的信息是不一样的。剑桥大学的 P. Forster 跟 C. Renfrew 曾经探讨过[④]，由于男女不同的迁徙历史，会造成人群演化上男性和女性不同的基因结构。如果 A 族征服 B 族后，大批 A 族的军人占领了 B 族的土地，娶 B 族的女人为妻，之后这个 A、B 混合族群的 Y 染色体与线粒

① Huxley T H. On the methods and results of ethnology. Fortnightly Review, 1865, **1**: 257 - 277.

② HUGO Pan-Asian SNP Consortium, et al. Mapping human genetic diversity in Asia. Science, 2009, **326**: 1541 - 1545.

③ Atkinson, Quentin D. Phonemic diversity supports a serial founder effect model of language expansion from Africa. Science, 2011, **332**: 346 - 349.

④ Forster P, Renfrew C. Mother tongue and Y chromosomes. Science, 2011, **333**: 1390 - 1391.

体 DNA 的比率，就会大大改变，因此语言常会跟随外来的征服者或殖民统治而改变。

要认真研究此类的复杂问题，当然要留意很多相关的条件，如人口多寡、文化高低等。例如满族虽征服了汉族，可是反而失去自己的语言而绝大部分汉化了。中国悠久的历史，是由很多不同的族群或分或合而逐渐形成的，目前至少还有一百多种语言。要了解这些语言的演化，就必须参考说这些语言的族群的演化[①]。

总而言之，Y 染色体与东亚族群演化对研究人类演化，特别是大中华许多族群的演化，做出了意义深远的贡献。我衷心地祝贺两位作者巨大的成就，也很乐意用这篇短短的推荐序以飨读者。

“中研院”院士、美国加州大学伯克利分校研究部语言学荣休教授、

香港中文大学伟伦研究教授

王士元

2015 年 7 月于香港马鞍山

① Wang W S Y, Ed. The ancestry of the Chinese language. Journal of Chinese Linguistics, 1995. 李葆嘉，主译. 汉语的祖先. 北京：中华书局，2005.

前 言

本书是一部生物人类学的研究专著，总结了20年来复旦大学人类学学科利用Y染色体基因分型研究东亚族群起源与民族演化的成果。

女娲有体，孰制匠之(《楚辞·天问》)？我们从哪里来，要到哪里去？对于人类的来历，人们孜孜不倦地问了数千年。到今天，人类起源之谜，依然是世界十大科学问题之一。

人类学追踪人类发展的历史，探索人类的未来。在这个学科的100多年发展历程中，古往今来的众多群体被记录、剖析、比较，从而归纳总结出一条条人类形体和文化变迁的路径。

与地球上的所有物种一样，人类也只有一个起源。古老的非洲大陆，是全世界人类的发源地。100多万年前，直立人走出非洲，来到欧亚大陆的各个角落。来到东亚的直立人，演化成了北京猿人。北京猿人的化石在70多年前被发现。之后的60多年中，人们一直认为东亚今天的人群是北京猿人的后代。直到1987年，夏娃学说的提出，发现现代人类是大约数万年前重新走出非洲的人种。学界开始怀疑中国人也是这个数万年前“走出非洲”人群的后代，而与50多万年前的“北京猿人”无关。1998年，复旦大学人类学组通过大规模的基因调查，终于一步步地证实了这个疑问。

复旦大学从20世纪20年代就开始开设人类学课程。当时的课程大纲中，包括“汉族的发源与迁徙”“民族之人类学基础”“中华民族的起源及其混合”等章节。这些问题在本书中的分子人类学研究中得到解答。1952年，通过院系调整，全国的生物人类学教研力量(浙江大学、暨南大学等)都集中到复旦大学，成为我国高等院校中唯一的生物人类学教学科研单位。由吴定良院士担任教研组主任的人类学组是当时生物系九个组中力量最强的，培养了大批我国急需的研究人才，输送到了中国科学院古脊椎动物与古人类研究所等科研单位，为我国人类学研究做出了卓越贡献。周恩来总理非常关心复旦大学人类学组，把它比喻成一颗珍贵的种子。这一颗种子，在中国的土壤中，经历了数十年的风风雨雨，终于成长了起来。

现代人类学教育部重点实验室于2005年年底成立，是目前我国唯一从

事分子人类学研究的实验室。实验室按照学科发展方向和国家重大需求，以“人群的遗传结构研究及其应用”为中心，旨在揭示人类进化过程中人群间和个体间的体质、生理、病理等差异及其形成机制，为疾病的发生和预防研究提供线索，为解决相关人文科学问题提供方法和工具。在金力教授的领导下，现代人类学教育部重点实验室在人类群体遗传学、体质人类学、分子流行病学等方面已达到国际前沿水平，取得了一系列重要成果，主要包括以下方面。

一、现代东亚人群的遗传结构、起源与迁徙：以往的人类迁徙图都把东亚人群的迁徙路线画成由北往南，经过我们的全面研究，全世界的人类迁徙图都把东亚部分改成了由南往北迁徙。

二、汉民族的人口扩张机制：汉族是世界上人口最多的民族，从遗传结构上看，汉族主要是人口扩张形成的，而不是在文化上对不同民族的同化。

三、亚洲古代人群的体质人类学特征：对新疆史前人群的骨骼研究发现，新疆自古是东西方人群混合的最前线。

四、流行病学研究：建立了国内大型分子流行病学研究队列和技术平台，不断收集和分析群体生理病理差异数据，开创医学人类学的新天地。现代人类学教育部重点实验室长期承担着国家重要的科研项目及国际合作项目，已成为在国内外学术界具有重要影响力的人类学研究基地。在国际基因地理人类迁徙研究计划中，本实验室是该项目全球十个中心实验室之一：东亚-东南亚中心。多年来，研究成果不断在 *Science*、*Nature*、*NEJM*、*JCI*、*AJHG* 等国际著名期刊上发表。

人类学的研究，走过了体质测量的时代，经历了分子人类学的阶段，正在谱写基因组人类学的新篇章。经过不断努力，人体和人群的一个个奥秘，终将被陆续解开！就让本书带你进入一个透过微观的 Y 染色体 DNA 分子窥见宏观的东亚族群演化历程的精彩世界吧。

李 辉

2015 年 1 月

目 录

第 1 章 分子人类学与东亚人类历史 …… 1
1.1 写在基因中的历史 …… 2
1.2 分子人类学与中华民族的起源 …… 13
1.3 Y 染色体上的自然选择 …… 25

第 2 章 Y 染色体与种族起源 …… 34
2.1 东亚现代人起源于非洲 …… 35
2.2 用遗传学数据重构人类进化谱系 …… 39
2.3 末次冰期东亚人群由南到北的迁徙 …… 51
2.4 从 Y 染色体看东亚人群的演变 …… 59
2.5 空间分析揭示的中国人群父系和母系遗传结构的差异 …… 72

第 3 章 汉藏族群的 Y 染色体 …… 88
3.1 单倍群 O3 的南方起源 …… 90
3.2 汉文化的扩散源于人口扩张 …… 103
3.3 平话群体是汉族一致性遗传结构的例外 …… 109
3.4 藏族人群的双重起源 …… 122
3.5 藏缅群体南迁过程的性别差异混合 …… 124
3.6 汉藏群体进入喜马拉雅东部的两条迁徙路线 …… 139
3.7 摩梭人的遗传起源 …… 152

第 4 章 南方原住民族的 Y 染色体 …… 166
4.1 苗瑶与孟高棉人群的遗传同源 …… 168
4.2 波利尼西亚人群的起源 …… 183
4.3 南岛西群和侗傣人群的紧密父系遗传关系 …… 190
4.4 长江沿岸史前人群的 Y 染色体 …… 207
4.5 海南原住民父系遗传结构 …… 213
4.6 海南岛卡岱语群体仡隆人的遗传起源 …… 221
4.7 海南岛的占城回辉人世系被本土成分替换 …… 233

第 5 章 北方原住民族的 Y 染色体 …… 243
5.1 中国西北人群的父系遗传结构 …… 244
5.2 阿尔泰语人群没有共同起源 …… 265
5.3 日本列岛人类遗传多样性分布 …… 274

第 6 章 Y 染色体与家族传承 …… 287
6.1 Y 染色体与姓氏 …… 288
6.2 现代 Y 染色体揭示曹操的身世 …… 298
6.3 曹操叔祖的古 DNA 结果与曹操后世子孙相符 …… 303

6.4 鄱阳操姓血缘上并非出自曹操 …… 305
6.5 赛典赤·瞻思丁和郑和的波斯祖源 …… 309

第7章 Y染色体与相关学科发展 …… 312
7.1 东亚人群中 *AZFc* 部分缺失并不提高精子发生障碍风险 …… 313
7.2 *AZFc* 部分缺失可诱发全缺失导致男性不育 …… 324
7.3 世界语音多样性分布格局与人类扩张 …… 338
7.4 古 DNA 分析技术发展的三次革命 …… 347

索引 …… 358

后记 …… 362

第1章　分子人类学与东亚人类历史

分子人类学是一门自然科学和社会科学的交叉学科，主要由遗传学、计算生物学、解剖学、历史学、考古学、民族学、语言学和地理学交叉形成。遗传学是一门自然科学，它不同于人文科学研究，因为自然科学的研究方法是假设驱动。研究者先根据相关背景资料做出假设，然后相应地设计实验，通过实验得出数据，并对这些数据进行分析，从而回答和检验原先的假设。这是自然科学的思维方式。在分子人类学的研究过程中所做的假设，是从人文科学特别是历史学那里找来的根据，但是，一旦假设提出，后面的研究过程实际上不受其他学科的影响，而是遵循遗传学研究自身的路径和规律。结果得出之后，又必须回到原来的相关学科去检验。所以，遗传学仅仅是分子人类学的一个工具。这样一个工具应用于历史研究，就形成了历史人类学，这对于历史研究有很大的作用。

分子人类学研究的材料是人类基因组。人类基因组由细胞核中的染色体和细胞质中的线粒体DNA组成，其物质基础是DNA大分子。DNA由于复制错误，在传代过程中会积累突变。突变主要有单核苷酸多态(SNP)、短串联重复(STR)和拷贝数差异(CNV)等类型。由于遗传漂变、瓶颈效应或者自然选择的作用，突变类型在群体之间会形成一定比例的差异。分子人类学就是用基因组分析人群之间差异来研究人类演化历史的学科。

Y染色体是纯父系遗传的单倍遗传物质，总是由父亲传给儿子，世代相传，在群体间差异最显著，可以很好地追溯人群父系源流，所以被广泛应用于群体遗传结构研究。Y染色体作为分子人类学的一大利器，其本身的特性需要详细地分析，以确保它被用于分子人类学时的可信度。我们对反映群体历史和遗传的Y染色体证据的解读大多基于下述假设，即Y染色体男性特有区段上的标记位点不受自然选择作用。然而，Y染色体的多样性较低，因此很有可能受到自然选择的作用，对于这一问题学界争议了30年之久。近年来，不断发展的测序技术为解析上述问题提供了很大帮助。Y染色体上的X染色体退化基因广泛受到净化选择作用，而扩增区域与睾丸发育相关的基因在进化历程中可能受到正选择影响。这些新发现提示在应用Y染色体进行群体遗传分析时，要注意把可能的自然选择作用也考虑在内。

从Y染色体看，中华民族在三四万年前起源于东亚南方，从布拉马普特拉河与澜沧

江之间进入东亚，大致分两条路线向北扩散，在不同区域演化成不同的文化族群，其后又渐渐融合成一个整体。

1.1　写在基因中的历史

1.1.1　研究人类历史的窗口

有关曹操墓真伪的学术争论的出现，使得生命科学的研究者与历史学者有了深入的交流。本节试图从历史学的视角，讨论遗传学与历史学是如何交汇在一起的。

研究人类历史，最主要的窗口是史籍。史籍记载的是曾经发生过的事件。同历史学相关的还有考古学。考古学研究什么？考古现场有一个墓或一个遗址，是人类活动留下的痕迹。这些痕迹是发生过的事件及其时间、地点、人物所留下的实物证据。通过史籍记载下来的历史事件在今天已无法目验，而考古则让我们回到过去，触摸到那些过去遗留下来的实景实迹。

现在，我们再增加一个考察坐标，那就是基因。实际上，基因虽然不能告诉我们历史事件中的时间、地点和人物，但它可以告诉我们人群间或者人物间的关系。比如，任何两个人之间的生物学关系究竟如何，借助遗传学分析就可以推测；又比如，现在的汉族和藏族两群人之间的关系如何，也可以通过遗传学进行分析和推测。所以，基因研究可以把人们关联在一起。如果说史籍上记载的是一个一个的点，考古也是研究一个一个的点，通过史籍记载，我们可以把这一个个点连在一起，那么，基因研究对于历史研究的贡献，也在于它能把一个个点上的人关联起来。这使我们看到了历史学、考古学和基因研究相互融合、相互交叉的可能。

如果我们试图推测历史上曾经发生过什么，我们主要依靠的是寻找历史的痕迹。这些痕迹可以从多个学科去分析，如历史学、历史语言学、考古学甚至古生物学，还有人类学、气象学、进化遗传学等。但是这些学科告诉我们的时间深度是不一样的（图 1-1）。比如，历史学可以使我们上溯到 4 000 年前，这已经上溯推衍得很远了，像甲骨文就是上溯推衍到 3 000 年之前。历史语言学则通过语言的比较，了解语言是如何进化、怎样分化的，最多的年限可上推到 6 000 年前，如果要推到 1 万年前，则要借助一些猜测了。考古学因为有实物保存，从现在的学科发展来看，至少可以追溯到 250 万年之前。而古生物学、古人类学和进化遗传学可以推得更远，因为人类和黑猩猩作为两个物种在进化上分离是在距今 500 万～

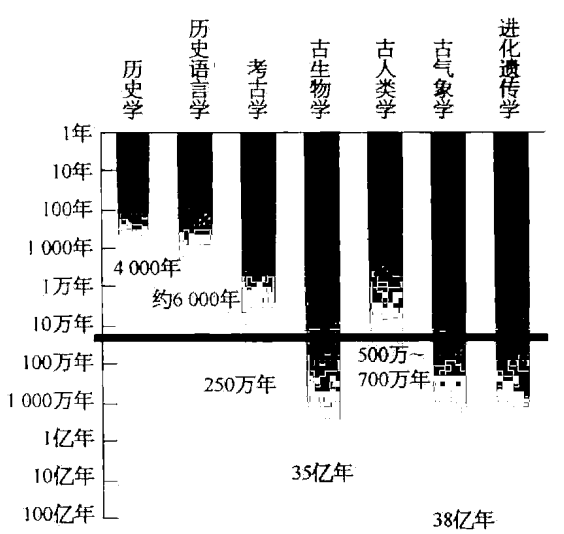


图 1-1　各个学科对历史探索的深度

700万年。历史可以通过史籍研究去推测、了解，但史前史一直是遗传学和考古学的强项，尤其是遗传学，因为它基本上在推测史籍上没有记载的东西。

所以，当历史学和遗传学交汇的时候，遗传学增加了时间的深度。当然，遗传学和历史学最大的差别在于，遗传学是真实的。你可以到处说你是某位历史名人的后代，而遗传学只要获得了那位名人的DNA信息，就可以准确无误地告诉你：你究竟是不是他的后代。遗传学或者分子人类学可以做什么？在笔者看来，它是连接史籍、考古、家谱等的桥梁，其中家谱实际上也是一种历史记载。

从遗传学角度如何看人类历史？简单地说，只有两件事情：人类是如何起源的，然后又是如何迁徙分化的。也可以用两个字来概括，即源和流。遗传学相比考古学、历史学有很大的优势，它的研究对象主要是活着的人。通过研究现代人群的遗传结构去分析、推测人群的进化史。如果结果能同历史事件联系在一起，就可以不叫进化史，而叫人类史。

现在，我们会把分子人类学和遗传学两个名词交互使用，因为当遗传学应用到人类历史研究时，我们把它叫作分子人类学。

遗传学家眼中的人类迁徙史是怎样的呢？它会告诉我们，人类起源于非洲，然后迁徙到世界各地，人类迁徙时间与冰川期的气候相关(图1-2)。最近，美国《科学》杂志发表了两篇文章，作者通过对现代人和尼安德特人的比较，认为人类的祖先和尼安德特人的祖先在基因上有一些交流，但是尼安德特人对于现代人类的基因贡献是微乎其微的[1]。而现代人类的祖先起源于东非，受到了气候波动、环境变迁的驱使，迁徙到世界各地。人类祖先走出非洲基本上分两步：第一步，部分人走出了非洲，到了中东；第二步，中东的群体开始分化成两支，一支沿着印度洋沿岸往东，另一支往北往西迁徙。往东迁徙的成为现代亚洲人的祖先，往西迁徙的则成了现代欧洲人的祖先。于是，我们现在可以看到世界上有黄种人、白种人的差别。

人群迁徙后会发生什么？由于史前人类很少，迁徙出去的人群因为缺乏通信技术，所以就与原来的人群失去了联系。于是他们在信息上没法沟通，在遗传上也没法交流(遗传上有交流的两个人群必须在一起)。这样的地理分散造成了人群的遗传隔离。长时间隔离，导致人群开始分化，从而变得不一样。这种不一样包括人群特有的生物学以及非生物学特征，非生物学特征也叫作文化特征。研究人群分化的动力、过程和结果的学科叫作人类学，注重生物学特征的叫作生物人类学，注重文化特征的叫作文化人类学。

人群的生物学特征是什么呢？例如，遗传学的特征，可以通过基因分析得到；体质形态的特征，即外形长相的差异，比如肤色，至少黑人和白人一眼看上去就不一样；疾病的特征，即不同的人群常患的病不一样。一些非生物学特征则包括社会文化特征、语言特征、宗教特征等，后来还多了一种民族特征。

这些特征为我们提供了分析的途径。人群分隔的时间越短，相似性就越大；分隔的时间越长，相似性就越小。当然这里面有一个隐含的假设，就是人群分离后不交流。我

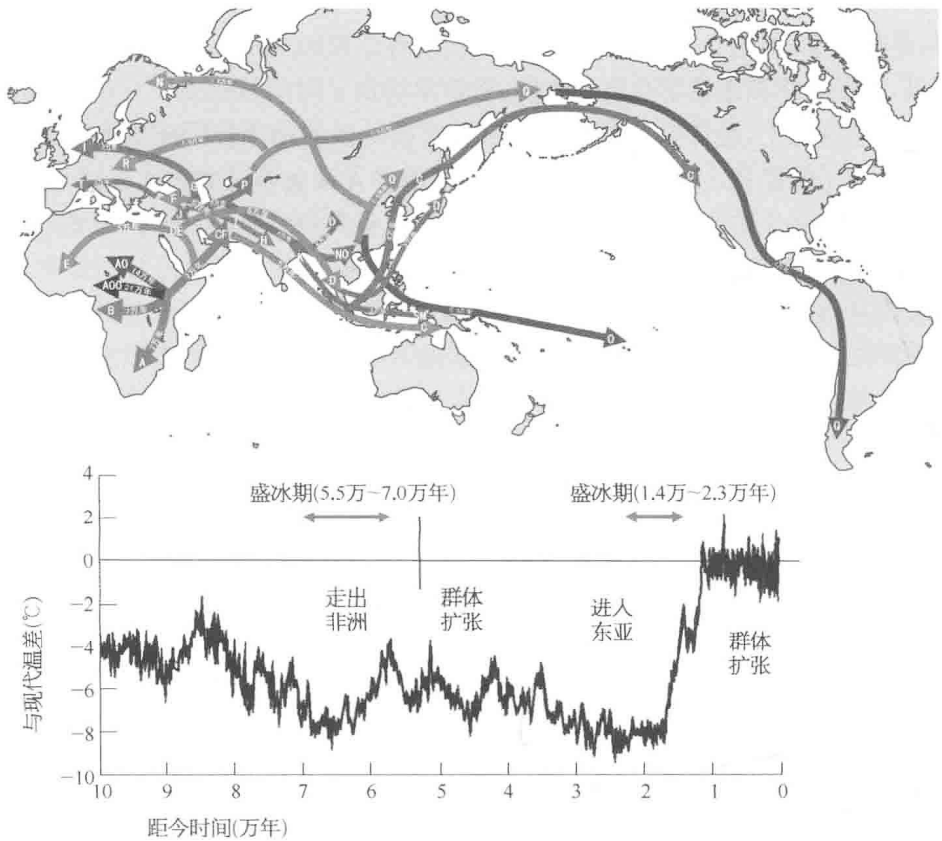


图 1-2　Y 染色体各个主干单倍群的起源和迁徙(数据来自南极洲 Vostoc 冰核)

们可以因此去度量人群之间的关系,包括他们的生物学特征关系,以及非生物学特征关系。因为分隔时间的长短,导致他们分化特征的相似性不一样。就研究不同人群随移动而发生的分化与融合过程而言,最需要追寻的一种可以直接度量,并且可以很精确地加以度量的特征关系,就是遗传上的相似性。

所以,我们尝试通过度量人群间的遗传关系去度量人群间的关系。这样,我们就等于找到了一把尺子。如果说遗传学和人文学科有交融的机会,无非就是说,我们可以把遗传学作为一个很好的测量工具,用来度量人群间的关系。

1.1.2　为什么可以借助遗传学进行分析

生物学可以这么简单地概述:人的细胞有细胞核,细胞核里有 23 对染色体,把染色体拉长之后,就能看到 DNA,它是一个双螺旋结构。DNA 上面有基因,基因能对蛋白质进行编码,表明蛋白质是怎样合成的,蛋白质合成之后就执行某种功能,这些功能的组合就是我们的生物体(图 1-3)。

DNA 的双螺旋结构由很多单位构成,这些单位叫作碱基对。碱基一共有 4 种,分别

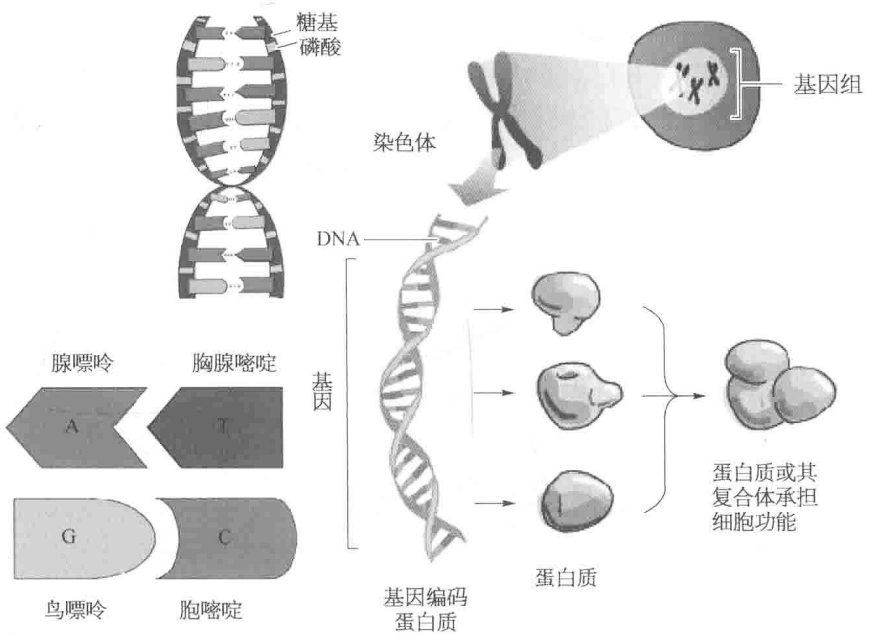


图 1-3 DNA 的结构和功能

是 A、T、G、C，它们把 DNA“缠绕”起来，压缩之后就成了染色体。在整个细胞中，除了一个细胞核，还有细胞器，最重要的细胞器就是线粒体。线粒体是产生能量的地方，它自己也有 DNA，叫线粒体 DNA。所以，我们谈到基因组的时候，不仅指核基因组，还包括线粒体 DNA(图 1-4)。线粒体 DNA 对研究人类迁徙也是一个很好的工具。

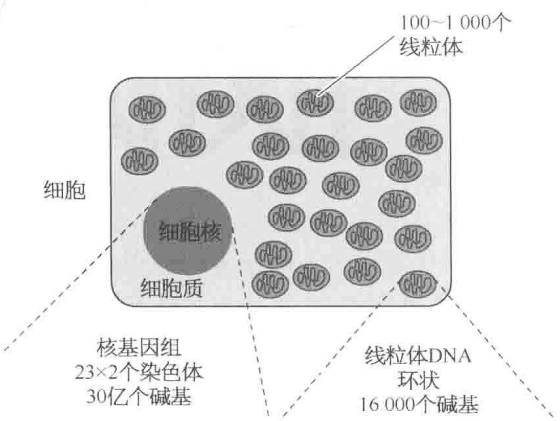


图 1-4 人类基因组包括核基因组和线粒体 DNA

核基因组里面大约有 30 亿个碱基，组成 23 对染色体，所以一个人有 46 条染色体。每一对染色体都有两套，一套来自父亲，一套来自母亲，因此染色体有一个重组问题，父本和母本染色体的相同位置会发生交换而把遗传特征重新组合。但是在基因组里，Y 染色体不会重组。Y 染色体最大的特点是只有男性有，即通过父亲传给儿子，再由儿子传给孙子。父亲的线粒体 DNA 对下一代没有贡献，下一代儿子或者女儿只接受母亲的线粒体 DNA，所以，尽管男性有线粒体 DNA，但不往下传，故线粒体 DNA 只是母系遗传。

通过分别研究 Y 染色体 DNA 和线粒体 DNA，我们可以推测人类的父系历史和母系历史。但是 DNA 这种由 A、T、G、C 4 种碱基组成的序列不安分，它会发生突变。有时原

来的 G 会突变成 T、A 或者 C，诸如此类。这是一个随机现象，主要因为 DNA 在复制的时候，偶尔会发生错误。很多人以为在生物进化过程中，很多结构和规则会变得很完美，其实不然。正是这种复制错误，使得人类在不断进化中会产生变异。当然，有些突变对人体不利，比如体细胞突变会让人得癌症，癌症发生的主要原因就是在人体某些关键的地方，基因组发生了突变。在人群中，突变成了一种普遍的现象，使得人与人之间或人群与人群之间产生出差异。这种差异正好可以被我们用来作为一个标记，去研究人类，它就是遗传标记(图 1－5)。

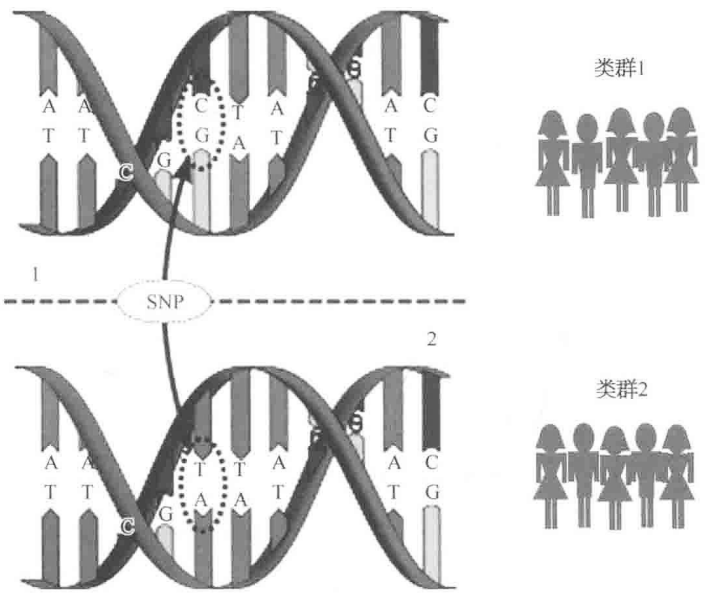


图 1－5　遗传标记的产生和群体差异的形成

人群中任意两个人的基因组差别是多少？平均是千分之一。人大约有 30 亿个基因组，任意两个基因组中，平均有 300 万个不同碱基。所以，如果我们能在整个基因组的水平上进行分析，就会知道人与人的差别非常大，没有完全相同的两个人。正因为出现了这种差别，我们可以回过头来分析这种差别形成的历史。比如，我们最早的祖先是一种基因组，突变以后发生分化，形成了不同版本；再进一步积累突变后又进一步分化；到现在，可以看到各种各样不同版本。这些突变发生在不同位置上，把这一个一个突变点串起来看，就形成个人的特征，成组的突变特征叫作单倍型(图 1－6)。

观察全世界单倍型的分布，可以看到全世界人的单倍型确实不一样。既然不一样，就可以拿来分析。这里，我们强调的是人群间的关系。人的基因不一样，有些会表现在形态上，比如，有些人的耳垂要长一些，有些人的手指能够弯过来；还有一些表现在人的精神状态上，比如，容不容易得精神类疾病，到了高原需不需要氧气瓶等，这些都跟基因有关。

我们之所以能进行遗传学分析，还有一点很重要。尽管人身上有很多类型的细胞，但是，每个细胞里的基因组都一模一样。所以，我们并不需要去分析身体的各个部位来了解基因组，

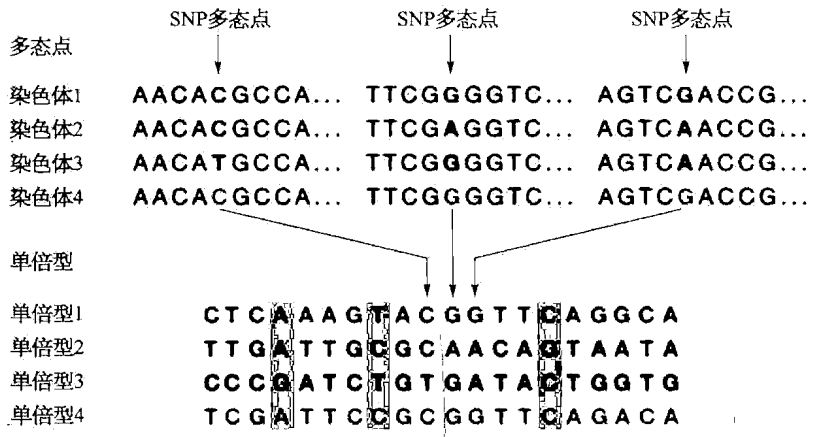


图 1-6 成组的突变特征构成各种单倍型

只要抽少量血，或者搜集一些口腔脱落细胞，或者拔一根头发，就可以进行遗传学分析。

通过分析现代人群，可以把那些特征性的东西挖掘出来。所以，用遗传学研究人类历史，很重要的一点就是提取人群的特征，这些特征最终是用遗传标记表现出来的。

那么，如何提取人群特征？举个例子，黄海海纳百川，它的水一部分来自黄河，一部分来自长江。在黄海搜集一小瓶水，如何告诉人们里面有多少水从黄河流过来，有多少水从长江流过来？要知道，这瓶水就相当于现代人的基因组，它是混合的。有一个道理简单，但操作可能困难的办法：用一缸红墨水、一缸蓝墨水，在黄河源头把一缸红墨水倒进去，到长江源头则把一缸蓝墨水倒进去。过段时间，到黄海取一份水样，如果红墨水的分子有特征、蓝墨水的分子也有特征，那么只要计算红墨水的分子、蓝墨水的分子所占的比例，就知道了黄河水和长江水的比例。这是很重要的道理，因为我们只能分析事物之间有差异的特征性标记。在人群中如何做呢？那就是把不同人群的源流用遗传标记“涂成颜色”。所以说，这个世界应该是一张色彩斑斓的地图(图 1-7)。

1.1.3 遗传标记隐含人类历史痕迹

笔者所在的实验室在过去十几年中，通过提取有群体特征的遗传标记，推测出东亚人群的迁徙方向。下面举几个例子来说明。

第一个例子，先不论东亚人是不是起源于非洲，我们要问的是：他们是怎么来到东亚的？如果从西方直线来，地理上存在着很明显的障碍——喜马拉雅山脉和喀喇昆仑山脉，要翻过这两座山脉几乎不可能。所以，只有其他两条路可走：一条沿着南亚，通过东南亚往北走；另一条通过中亚大草原，从西北往南走。

我们从父系遗传的 Y 染色体角度来看。从全世界各类 Y 染色体上单倍型的关系图中发现，东亚各地的单倍型是多态的，是含群体差异信息的。我们在某些位点上对东亚人群进行研究，同时用一种颜色代表一种单倍型，就会看到，东亚北方人群看上去颜色比较素，南方人群看上去比较多彩，也就是说南方人群单倍型比较多(图 1-8)。而且，北方

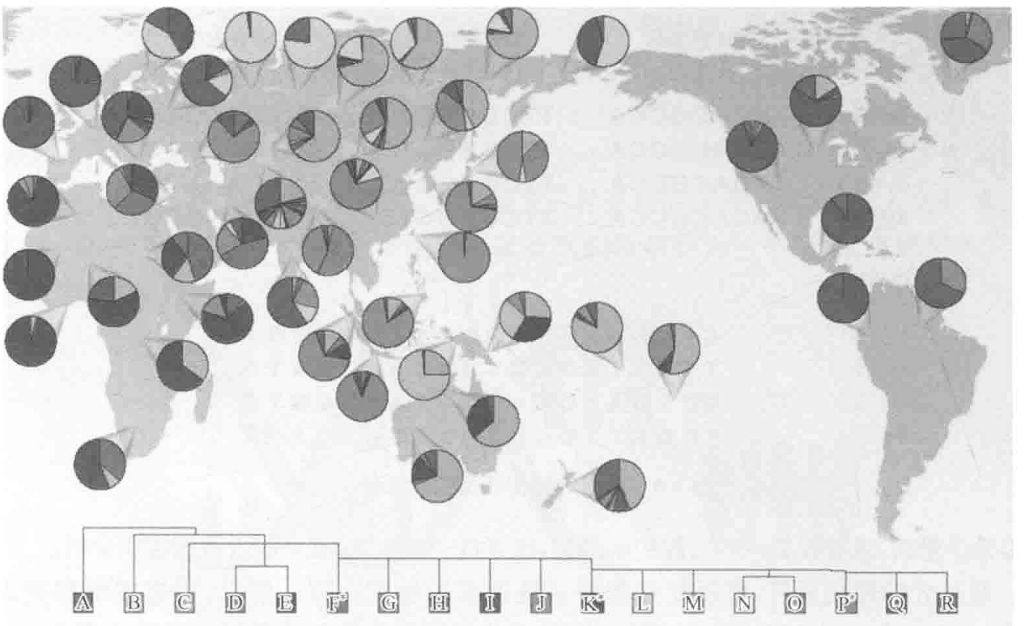


图 1－7　全世界各个人群中的 Y 染色体单倍型的分布

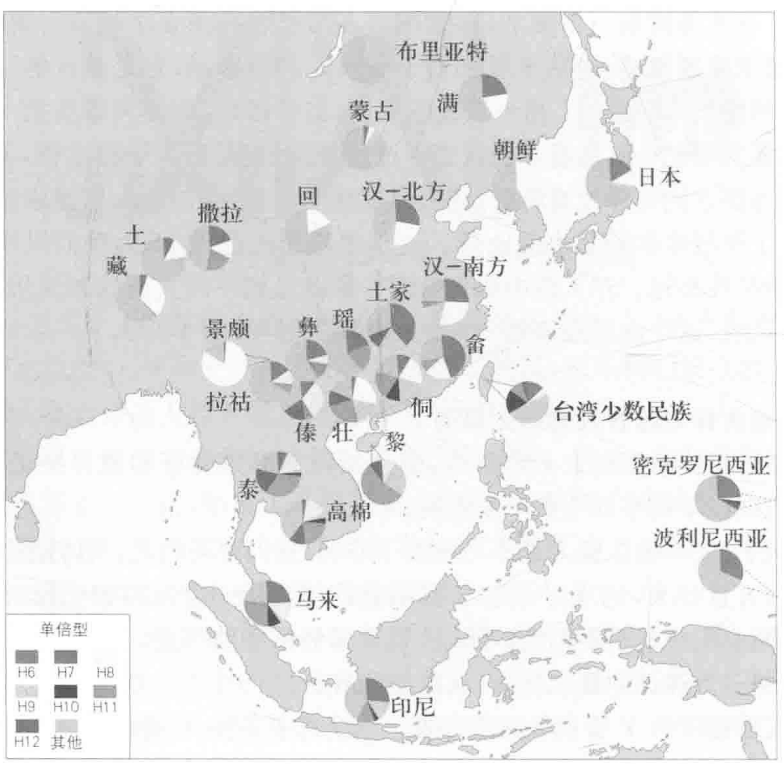


图 1－8　东亚 Y 染色体单倍型的分布

图中区分的单倍型是早期的分类，H6 为 O3－M122，H7 为 O3－M7，H8 为 O3－M134，H9 为 O1－M119，H10 为 O1－M110，H11 为 O2－M95，H12 为 O2－M88。

人群的单倍型在南方人群那里都有。这是第一个观察结果，显示出有特征性的分布规律。

然后，我们任选一种单倍型，来对比南方人群和北方人群。我们可以加上另外一些标记，从突变积累程度中看同一个单倍型的古老程度。我们发现，对同一个单倍型来说，南方的突变积累得比较多，也就是比较古老，北方的比较晚近。

最后，把这些群体的遗传关系通过“主成分”分析方法（图 1-9）进行分析，我们发现，在相似性上，南方人群变异很大，北方人群变异很小。

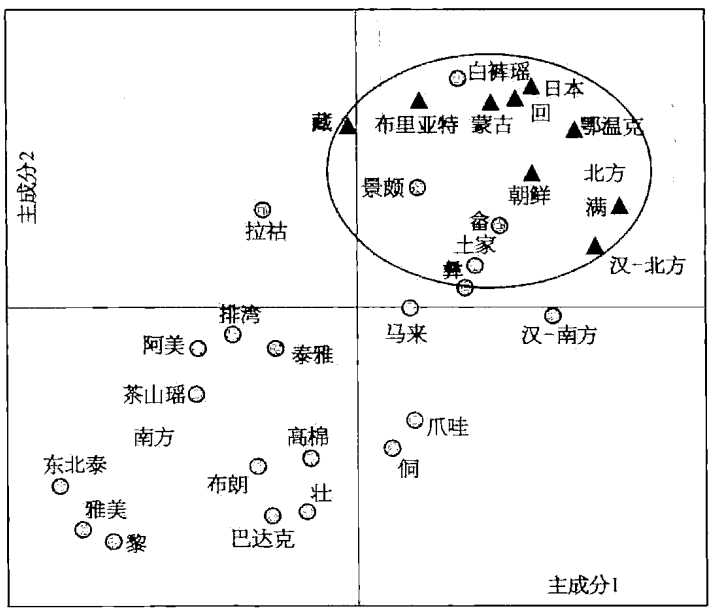


图 1-9 主成分分析揭示东亚南方的多样性大于北方

根据以上 3 个观察结果，我们在 1999 年提出一个假设，东亚人是从南方过来的，因为南方类型比较古老，比较多，也比较分散，差异大[2]。我们又对东亚人从东南亚进入的时间进行估计：距今三四万年前，由南向北迁徙，进入中国和东亚其他地区。当然，对于现代的东亚人群来说，我们不排除其中有北方过来的人群，但是这些人群进入东亚的时间要晚得多，现代考古学和遗传学证据表明，大概最多距今三四千年。

跟遗传标记传递方式最相似的是语言。我们的母语也是从父母那里传下来的。人的基因一半来自父亲，一半来自母亲，所以还比较稳定。而人的语言本身变化则比较快、不稳定。比如，上海城市的历史没有多少年，却出现了独特的上海话。

从语言分类上看，整个东亚大致分成 6 个语系。东亚北边是阿尔泰语系、汉藏语系、苗瑶语系。阿尔泰语系有 3 个代表人群：满族、蒙古族、维吾尔族；汉藏语系有汉族、藏族等；苗瑶语系就是苗族、瑶族和畲族。往南是侗傣语系，包括傣族、壮族等；再往南是南亚语系，包括柬埔寨人、越南人；最南边是南岛语系，包括马来西亚人、印度尼西亚人、菲律宾人以及太平洋大部分岛国人群。研究发现，人类的语言分布同遗传的分布也有高度的

对应关系。

我们估算南方和北方人群对东亚人群的贡献，结论是，南方人群对东亚人群的贡献大于北方人群(图 1－10)。过去，人们都以为东亚人群是从北方过来的。笔者通过研究，认为东亚人群是从南方过来的。现在国际上都接受了东亚人群是从南方过来的这一结论。

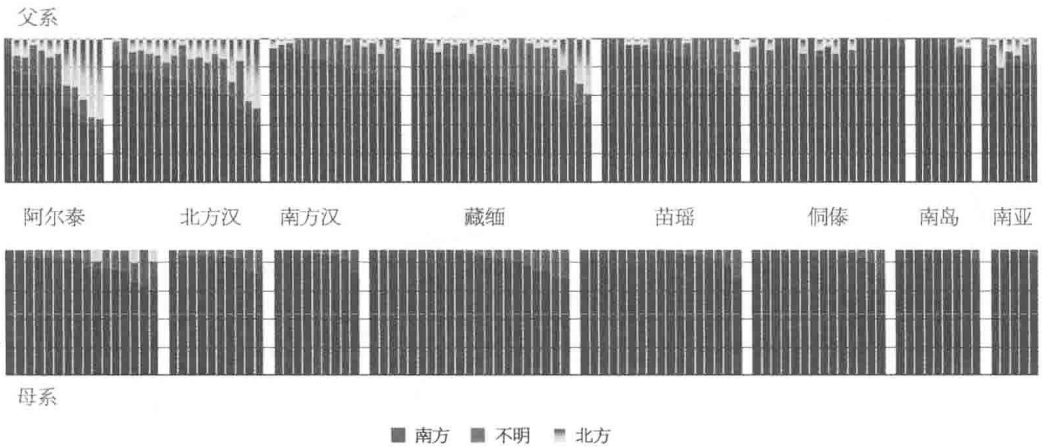


图 1－10　东亚各个群体系统的南北来源成分比例

第二个例子，是中国北方汉族人群的扩张。我们知道，中国在 5 000 年前就形成了几个文明的萌芽(图 1－11)：西北的仰韶文化、东南的河姆渡文化和其后的良渚文化，山东

图 1－11　中国新石器时代的几个主要考古文化系统

有龙山文化，华中地区有大溪文化，其先民有可能是苗瑶族群的祖先。农业主要在仰韶文化和河姆渡文化所在地，粟的种植同仰韶文化相关，而水稻的种植则同河姆渡文化紧密联系在一起。所以说，现在的中国人基本上是这两种文化族群的后代。这样说来，北方的人占据了北方，南方的人占据了南方。北方的人显然是汉族的祖先了。那么，最初南方人讲的是不是汉语呢？为什么现在南方人的方言特别多呢？能不能通过遗传学研究出现在南方都自称汉族的人群究竟是不是遗传意义上的汉族？换句话说，当汉文化向南扩张的时候，它究竟是一种单纯的文化传递或者说“同化”，还是人群带着文化一起往南走？这项工作不难做。实际上，北方汉族的样本很容易采集。尽管南方的现代人群基因库变了，我们还是可以拿现在南方各原住少数民族来同北方的汉族做比较。通过比较，我们来分析究竟是北方的汉族对南方汉族的贡献大，还是南方的少数民族对南方汉族的贡献大。

通过 DNA 分析，我们发现南方汉族的父系和母系结构不太一样。就父系而言，南方汉族的主体基本上是从北方汉族而来的；就母系而言，现在的南方少数民族的母体对南方汉族的贡献更大(图 1－12)。有意思的是，这些贡献还同地理纬度有很大的相关性。这些发现正是借助遗传学得来的。

北方的汉族为什么要来南方？历史学家的研究表明，这同历史上的政局变化相关。在中国历史上有三次大的移民期，即西晋灭亡期、唐朝安史之乱之际以及辽金时期。我们在遗传学上也看到了这样的证据。所以，汉文化的传播主要是由人口迁徙所驱动的，这种迁徙表现出一种强烈的性别偏向性[3]。

1.1.4 历史研究如何用好遗传学工具

遗传学主要研究人群间的关系和个体间的关系。借助遗传学，我们在研读史籍的时候，就可以超越时间、地点限制，获得某些历史人物的个体生命史信息。绝大部分重要历史人物都是有后代的，要研究著名的历史人物，借助遗传学，可以对他们的后代加以分析，做多个相对独立的个体分析。比如，曹操的后代有很多支，每一支都取一些样本，分析他们的 Y 染色体，因为 Y 染色体是跟着父系遗传的，曹操后代的一个共同特征，就是有同样的 Y 染色体单倍型。如果能准确找到这种共同的单倍型，就可以知道曹操的 Y 染色体是什么样的，而无需去检测曹操墓中的骨头。这种工具是有法医学意义的，因为遗传学分析结果可以作为法庭证据。

遗传学分析还可以用来证实或者证伪历史人物或人群之间的遗传关系。这些历史人物之间的关系，历史书说了不全算数，因为历史书的准确度受到写作者认知的影响。如果有条件通过 DNA 判断，那么 DNA“说”的是算数的，因为 DNA 这本书是“自然”书写的。借助遗传学的基因技术，我们还可以对不同地点关联人群的样本进行分析，找出他们的分化路径，来判断历史人物后代或人群的迁徙路线。

遗传学可以用来分析历史人物的民族归属。我们知道，每个民族都有它自己的人群特征，我们可以借助 DNA 技术去分析历史人物的民族特征。当然，前提是我们要对古代

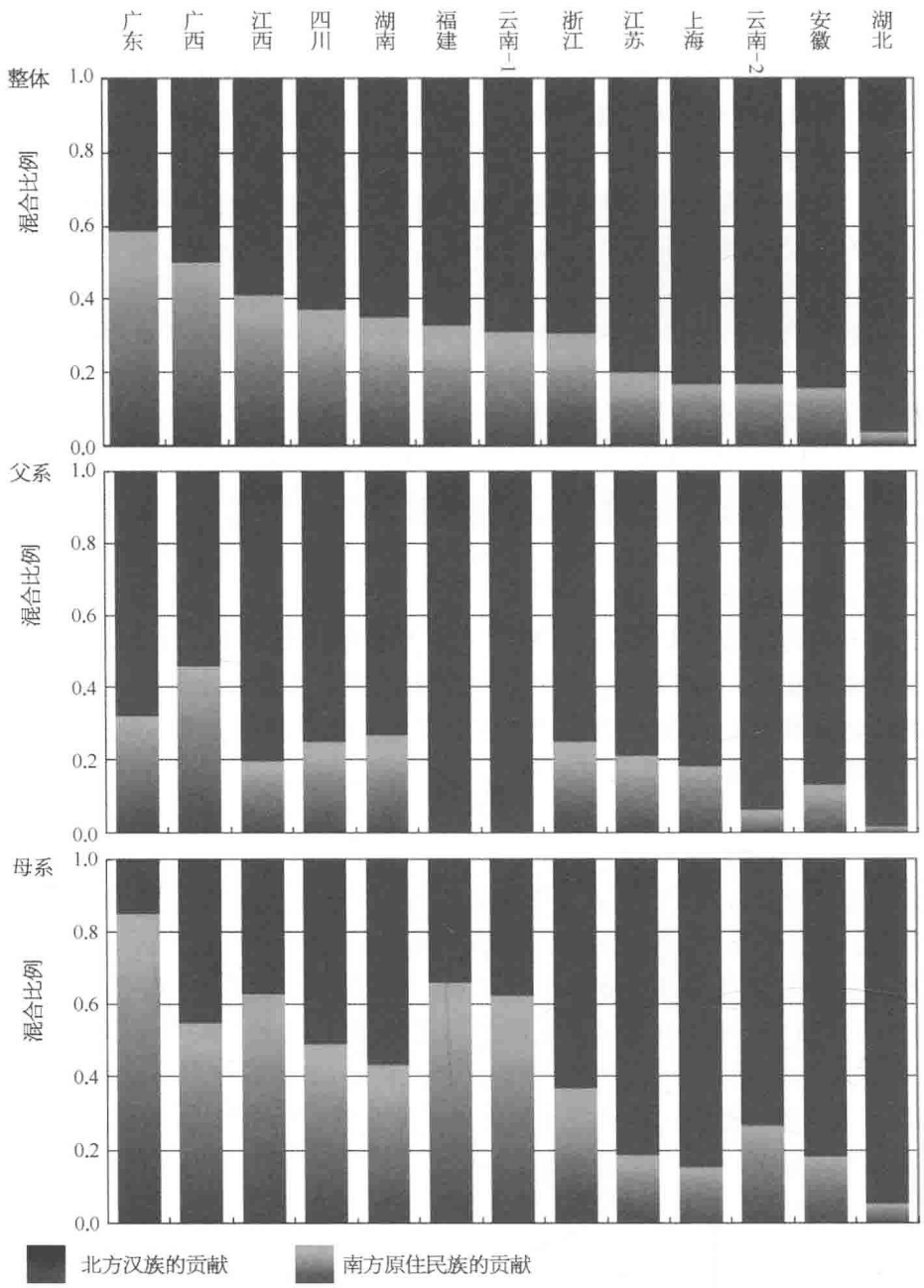


图 1－12　南方汉族的遗传混合比例

民族的 DNA 数据有足够的学术积累。

遗传学还可以用来分析古代人群的迁徙(比如,回鹘人群是从哪里开始,又是什么时候通过哪条路线到达新疆的),研究古代人群的人种特征(比如,匈奴人究竟是黄种人还是白种人),以及古代民族人群间的交流。

分子人类学和遗传学属于自然科学，同人文科学研究有一定差别。遗传学仅仅是一个工具。这样一个工具对于历史研究有很大的作用。至今，笔者所在的研究团队已经获得了200多个民族、近20万个样本可供专业研究。获得这些样本，是以长期艰苦的野外工作为基础的。

参考文献

[1] Green R E, Krause J, Briggs A W, et al. A draft sequence of the Neanderthal genome. Science, 2010, 328: 710-722.

[2] Su B, Xiao J, Underhill P, et al. Y chromosome evidence for a northward migration of modern humans in East Asia during the Last Ice Age. Am J Hum Genet, 1999, 65: 1718-1724.

[3] Wen B, Li H, Lu Du, et al. Genetic evidence supports demic diffusion of Han culture. Nature, 2004, 431: 302-305.

1.2 分子人类学与中华民族的起源

文明的发生、发展、继承过程中，有多少民族辉煌过，又迷失在历史的烟尘中。历史尘封了很多记忆，也遗留下众多可供后人追寻的片断凭据。文物、史书、传说、习俗，透露给我们不少中华民族一路走来的印证。但是深藏在这些印证背后的历史变迁的内在机理，往往像剪不断理还乱的麻线。然而科学给了我们一把剪子，基因组分析这把利剪将认识历史的死结一一剪开，把历史上的一个个人群串联起来，追根溯源，重新书写中华民族来到九州大地并且生生不息的族谱。

1.2.1 人群遗传标记及其变化的因素

人类学是研究人的分类的科学，具体而言就是研究人类的个体和群体之间的“同”和“异”。“同”是人类的起源问题，“异”则是人类的多样化问题。用基因组分析的方法来解读人类群体的起源和多样化历史，成就了一门新兴的学科——分子人类学。这一门学科所倚仗的材料就是人类个体和群体之间的基因组差异，也就是遗传变异。人类的基因组结构大体上是相同的，每个健康人都有23对染色体，每种染色体上的基因排布也都比较固定。每条染色体都是由两条DNA分子长链以双螺旋结构配对组合在一起。DNA分子的基本单位是核苷酸，核苷酸之间的最大差别是其上的碱基部分不同，由A、T、C、G 4种碱基按照特定的顺序呈线性结构排列起来，在人类细胞核中组成23对(46条)DNA分子，细胞质中的线粒体也有一段环形的DNA分子，这就是人类基因组。DNA分子中不同的碱基序列构成了人类基因组的不同部分，有的是编码蛋白质的基因序列，有的是调度基因转录翻译的指挥序列，有的甚至只是使DNA分子形成一定规模和结构的填充序列。虽然现代人个体之间的基因组结构基本相同，但是在细节上，基因内和基因间的碱基序列在人和人之间有较大的差别。不同性质的序列在人类演化过程中的变化特点不同，所以差异程度和差异分布格局也不同。人类基因组的长

度大概是 30 亿个碱基对，这些个体差异可能在整个基因组中所占的比例不到 1%，但总量却依然相当可观，而且正是这些差异促成了人与人、族与族之间那么多外观和生理的差别。

这种个体之间的基因组差异大致分为三大类：短串联重复(STR)、单核苷酸多态(SNP)，以及性质处于这两类之间的拷贝数差异(CNV)。其中 STR 由于变异种类较多，在遗传学研究中应用历史最久。STR 的定义是在基因组的某一区域中一小段片段(往往是若干个碱基)重复了若干次，而且在传代过程中这种重复数可能改变，造成群体中的不同个体在这一片段可能有着不同的重复数。比方说，STR 的重复片段就像火车的车厢，其数目可以调整，整个 STR 区段在个体之间的差别就像每辆火车都可能有不同数目的车厢，所以这种差异的种类可以很多。在传代过程中，重复数目的变化只能是每次增减一节，所以重复数目相近的个体之间的关系往往会比较近。SNP 则是一种完全不同的基因组差异，是在基因组的某一个特定的位置上的一个单一的碱基发生变化。这种变化可能是一种碱基变为另外一种，如 A 变成 G；也可能是缺失或者额外地插入一个碱基；还可能是插入或者缺失一个小小的片段。这就像集装箱卡车一样，可以换不同的集装箱，也可以不装集装箱。大多数 SNP 只有两种形态，但也有少数存在三种或者四种形态。SNP 在基因组中非常常见，也是形成功能差异的主要基因组差别。与 STR 不同的是 SNP 的变化速度非常慢，一般每次传代每个 STR 有大约千分之一的突变概率，而 SNP 每次传代每个碱基只有大约三千万分之一的突变概率。所以每一个 SNP 往往只能出现一次变化过程，更能够体现个体之间稳定的亲缘关系。CNV 是最近广受关注的一种基因组差异[1]，其定义是比 STR 重复片段更复杂的一段片段在基因组中呈现插入、缺失或者重复若干次的差异。CNV 的重复区段比 STR 的功能性强，但是重复数却没有 STR 那么自由，突变的速度也接近 SNP。研究者认为它对人类生理差异的贡献更大。

这些基因组差异是怎么产生的呢？一般认为，这些变化的产生并不受外界环境影响，而是在遗传过程中因为复制错误而突然产生，所以被称为突变。许多突变会对基因结构造成破坏，结果这些突变个体就不会存活下来，这些突变也即消失。没有危害的突变就会在群体中积累起来。群体的历史越久，这种突变的数量就越多。这就是分子人类学计算群体历史年龄的基本概念。

人类基因组中有大约 3 万个基因，有的时候，突变发生在基因结构中，会引起个体生理差异。这些差异在某些环境下没有引起个体生存能力的差异，当环境改变时，其中一种类型的个体可能就不能适应而渐渐在群体中减少，相应的适应者就渐渐增多。这就是自然选择的作用。在同一区域中的个体往往显示出相似的外形和生理特征，那往往是自然选择的结果。一个最典型的例子是热带地区的疟疾高发对地中海贫血基因的正向选择。东南亚和中国南方的人群中地中海贫血基因类型都很常见[2]，带有这种疾病基因类型的个体对疟疾的抗性比正常基因类型的个体要高得多，这就使他们更容易生存下来。结果，不管是南方的原住民，还是从北方来的汉族、缅族移民，在南方瘴气弥漫的地区，这

个基因的疾病类型在群体中的比例都增加了[3]。直观的结果是他们的长相颇相似，嘴唇变厚、鼻翼变宽、眼眶深陷、额头凸起，这都是地中海贫血的个体在胎儿期缺氧的结果。许多人都认为广东人长得像越南人，甚至以为他们有什么关系，其实这种外形和基因的相似性完全不能体现群体之间的遗传亲缘关系。所以在分析群体亲缘关系的研究中，任何可能受到自然选择的基因差异都不是良好的材料。

人类基因组的大部分区段并没有直接生理功能，它们是基因与基因之间的填充序列，基因内部的插入序列。这些序列一般不会受自然选择的影响，可以比较自由地积累突变和形成群体差异。在没有自然选择作用的时候，群体之间的基因组差异又是如何产生的呢？答案是随机的过程。在群体的延续历史中，各种突变都会以一定的概率传递到后代中。当群体越小，各种突变的传承概率就越不均衡，它们的频率就会随着世代传递而波动，这被称为遗传漂变。当群体小到一定程度，遗传漂变的效应非常强烈，有些突变的频率可能突然波动到零，这些突变就丢失了，这种现象在遗传学上叫作瓶颈效应。这就像是一群人通过一个狭小的瓶口，只有非常少的人能过去，过去的种类必然会少于原来群体中的种类，过去以后重新发展起来的群体就与原来很不一样了（图 1－13）。这种瓶颈可能是人类迁徙路径上的一个地理障碍（高山、海峡、沙漠等），也可能是导致历史上人口大规模减少的战争、瘟疫等。总之，人类在瓶颈中经历了人口极少的时期，使得大量的多样性丧失，直到人口再次增长，才重新开始积累基因组的新的多样性。瓶颈效应在人类历史中普遍存在。在人类迁徙过程中，各个区域之间都可能有瓶颈效应的存在，使

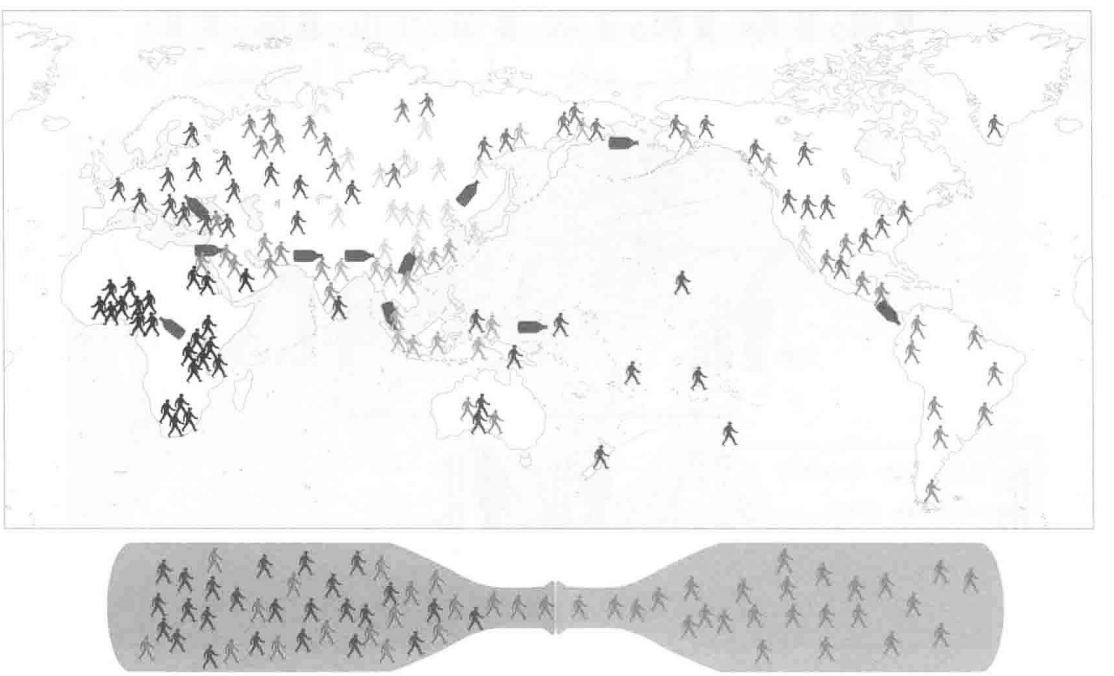


图 1－13 人类迁徙过程中的瓶颈效应

得各个区域的人群基因组多样性或多或少不同。比如美洲大部分地区的印第安人的祖先，在从西伯利亚通过白令陆桥到达阿拉斯加时，人口减少到几十个。后来的北美和南美的主要印第安人群体，都是从这几十个人发展而来的[4]。这是人类迁徙中最著名的瓶颈效应例子。

1.2.2　解析东亚人群形成的过程

依据遗传规则的不同，人类基因组可以分为常染色体、线粒体和 Y 染色体。常染色体是遵循父母双系遗传的，每个人都有两份常染色体，一份来自父亲，一份来自母亲。每一对常染色体称为同源染色体，其中父源染色体和母源染色体的对应部分在传代时可能发生交换，重新组合成一条混合来源的染色体，这一过程叫作同源重组。因为重组的发生，常染色体会受到混血的影响，代表个体的所有祖先的混合历史，分析的难度很大。而线粒体却是单纯的母系遗传，每个人的线粒体基因组都来自其母亲，因为精子的线粒体都在尾柄部，不能进入受精卵，因而不能遗传下去。Y 染色体是男性染色体，只有男性有，每个男性的细胞中都只有一条，自然是父系遗传的。Y 染色体和线粒体较为简单的遗传规则，在细胞内都以单拷贝存在，不会受到混血的影响（图 1－14），可以构建纯系的进化谱，使得我们可以更清晰地辨认其历史轨迹。在分析群体历史的时候，这两类遗传材料更为有效，是目前使用最广的分子人类学材料。

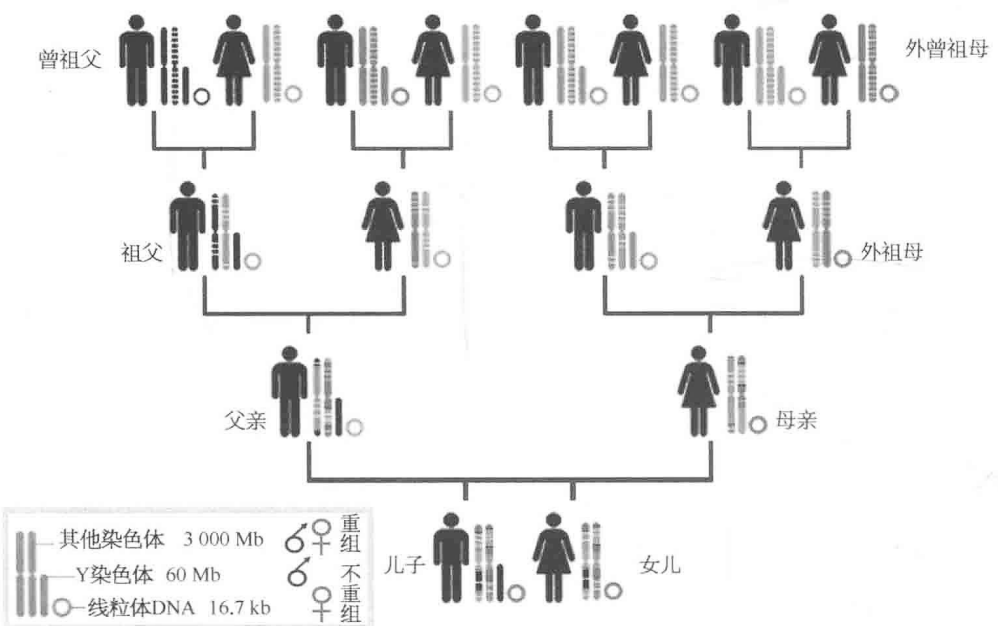


图 1－14　基因组不同部分的遗传方式不同

三类基因组材料的多样性分析[5-7]都指出，全世界的人类起源于东非。因为东非的多样性积累得最高，并且占据了各种遗传材料谱系树的根部结构。从东非到东亚的迁徙

过程中，人类群体经历了许多次瓶颈效应和自然选择的作用，使得人类到达“目的地”以后变得和出发地的人群非常不一样。历史上东非的原住民族群是一个叫作布须曼人的种群，他们操着科伊桑语系的语言，皮肤橙红色，头发蜷曲，个头不高，臀部脂肪特别厚。当一千多年前西非的班图黑人掌握了较高的农业技术（非洲粟的栽培）而开始向东非和南非扩张以后，这一橙色种群的分布范围被压缩到大湖周边和西南非[8]。今天南非的大部分原住民民族都属于布须曼人种，包括曼德拉总统，他的橙色肤色显然与西非来的班图黑人很不一样。而到达东亚的人群，已经变出了笔直的黑发、褐黄到浅黄的皮肤、匀称的身材，迥异于布须曼人。

从非洲向东亚的早期人类迁徙并不是一次完成的[9]。从今天的远东地区人群遗传结构分析看来，至少有两阶段的主要迁徙（图 1－15）构成了现在的远东人群分布格局[10]。印度洋沿岸的顺时针迁徙使得人群至少在 6 万年前就到达了大洋洲[9]，但是今天东亚的大部分人群的历史显然要晚得多[11]。根据到达东亚和太平洋地区的早晚，一般把这两批人分别称为早亚洲人和晚亚洲人。在过去的体质人类学分类上被称为“棕色人种”或者“澳大利亚人种”的澳大利亚、新几内亚和美拉尼西亚原住民显然属于早亚洲人。由于 Y 染色体的特殊人类学性质，对亚太地区的民族关系有最好的辨析度，所以我们主要以 Y 染色体的多样性来说明中华民族的形成过程。国际 Y 染色体命名委员会（YCC）把全世界的 Y 染色体类型分为代号为 A～T 的十几个大类群（单倍群）[12]。大洋洲原住民的 Y 染色体大多属于 C 单倍群，其年代估算非常古老。在全世界范围内搜寻 C 型 Y 染色体的时候，我们发现这一类型从中东的沿海地区出现，沿着印度沿海地区扩展，一直到东南亚和东亚沿海，甚至延伸至美洲（包括南美洲和北美洲，下同）沿海。大洋洲的 C 型也是这一沿海分布的一个分支。这使我们有理由相信，早亚洲人的

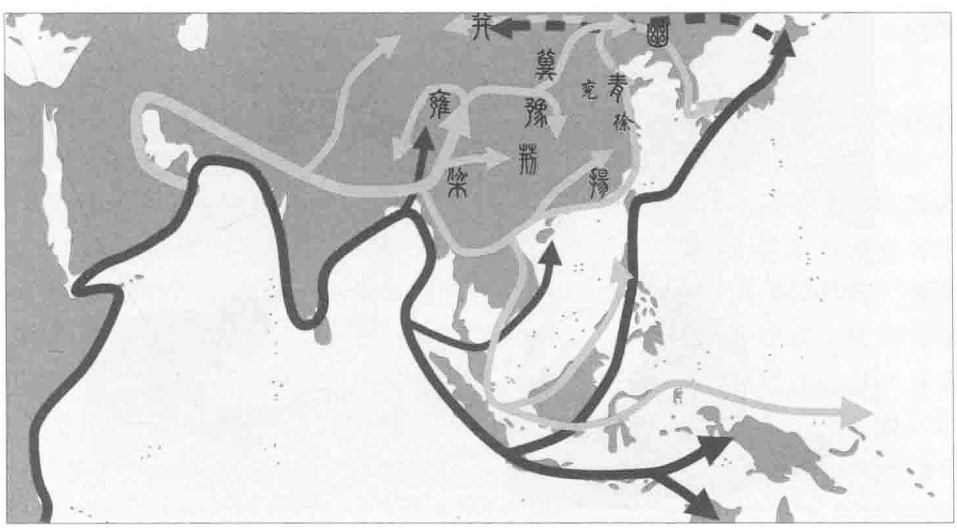


图 1－15 九州大地的人群起源示意图

棕色线条为距今超过 2 万年的迁徙路线，黄色线条为距今 2 万年以内的迁徙路线。

一支是沿着海岸线向东方探索并且扩散开来的。海岸线显然是一条容易通行的快捷方式，而考古发现也已经证实了东南非洲的人类在十几万年前就开始在海边居住并且依靠海洋资源生存[13]。所以这条最早的沿着印度洋海岸顺时针方向的迁徙可能开始于十多万年前的南非。

亚太地区的 Y 染色体类型中，还有一种 D 单倍群，也可能是另一支早亚洲人带来的，虽然他们的携带者可能与 C 型的携带者并不是同一群人，但是到达东亚的时间差不多[14]。这种类型出现在一些更为神秘的人群中（图 1－16）。亚洲的小黑人往往存在于我们的传说中，却也真实地隐匿在偏远的岛屿和深山中。他们也被叫作尼格利陀人，他们的 Y 染色体大多是年龄非常古老的 D 型，线粒体和常染色体也往往是特别古老的类型或者亚洲的底层结构。现在的尼格利陀人分布在菲律宾吕宋岛中部山地、马来半岛北部山地等狭窄地区，还有一支在缅甸东南面海外的安达曼群岛。马来西亚的阿斯利人和菲律宾北部山区艾格达等诸部虽然属于尼格利陀人，但是已经没有 D 型 Y 染色体。安达曼人的 Y 染色体全部都是 D 型，据分析至少在这个群岛上孤独存在了 3 万年以上[15]。D 型 Y 染色体也出现在东亚北部的某些奇特群体中。分布于日本本州岛北部、北海道，以及俄罗斯库页岛南部的虾夷人的 Y 染色体主要是 D 型[16,17]。在中国的调查发现，青藏高原的羌族、藏族系统民族中还有一定比例的 D 型 Y 染色体。奇怪的是，四川和甘肃交界地区的白马氐人，他们的 Y 染色体也全部都是 D 型[10]。可以肯定，这些 D 型 Y 染色体的人群在非常远古的时代中有着共同的来源，但是我们无法得知他们是怎样分布到这些

图 1－16　孑遗的 D 型 Y 染色体的分布

地区的。这些人群现在的外形特征差异已经非常巨大。南方的“小黑人”往往被认为是亚洲地区的“黑种人”，而北海道的虾夷人曾经被认为是“白种人”的一支，白马氏人则与其他典型的“黄种人”没有什么差别。他们都属于我们定义的“早亚洲人”的后代。虾夷人由于历史上经历了澳大利亚人种与尼格利陀人种的交叉混合，或许与“棕种人”更为接近。

晚亚洲人构成了我们现代东亚和太平洋地区人群的主体，在Y染色体类型上主要为O型，还有少量的N型和P(Q和R)型。晚亚洲人可能是追逐猎物从内陆来到远东地区的，直到3万～4万年前他们才到达东南亚地区，在2万年前的盛冰期开始进入中国，随着冰川的消退而北进，从南到北分布到整个九州大地[18]。在晚亚洲人移民九州的过程中，可能与早亚洲人有过生存空间的争夺。最终由于略高的技术和体力，晚亚洲人在大部分地区胜出，早亚洲人只留下很少的比例融合到各地的人群中。这段可能的历史早就被我们遗忘，也有可能改头换面地转化为神话传说中的上古的征战，那些被我们消灭的如“魔兽”一般古怪的族群原形很可能就是早亚洲人。

1.2.3 九州的分化和中华民族的合并

晚亚洲人进入今中国境内以后，迅速扩散开来，并分化成了各个民族。当传说中的大禹王定九州的时候，发现不仅仅九州风土不同，便是黎民也各异。今天流传的九州的划分方式有许多种说法。广为接受的《尚书》里的说法是冀州、兖州、青州、徐州、扬州、荆州、豫州、梁州、雍州，但是兖州、青州、徐州都是山东一带的地方，与其他六州不成比例。在《尔雅》中取消了西南的梁州，增加了东北的幽州，把青州改为营州。《周礼》中也没有梁州，又取消了徐州，增加了东北的幽州和北方的并州。实际上，最均匀的分法是把兖青徐三州合并，加上北方的幽州和并州，这样的九州兼顾了虞夏时代地理区域和风土人情的差异。在中国大一统局面出现之前，九州的文化都是相对独立地起源，并长期保持了各自特色的发展。之后经过不断的交流和融合形成了辉煌的中华文明[19]。

西南的梁州包括秦岭以南直至今缅北野人山的云贵川广阔地区，那里的先民是从孟高棉族群向汉藏族群过渡的人群。虞夏时期以李家村文化为代表的新石器考古文化也体现了从南到北的过渡类型。西方的雍州包括今陕甘青藏一带，这里是汉藏族群的古羌人的家园，也是华夏族的祖先生活的地方，仰韶文化的考古传统在这里发源和传承。南方的荆州在湖广地区，苗瑶族群的祖先荆蛮先民在这里开创了大溪文化传统。扬州的范围从北越到苏南，百越杂处，百越先民创造了大坌坑文化和马家浜-良渚文化等传统。青兖徐州在苏北和山东，传说中的东夷族的家园，是青莲岗文化、大汶口文化以及其后的龙山文化的发源地。豫州在河南，是雍州来的华族与青州的夷族和荆州的蛮族交融的地方，这里的裴李岗文化是九州最古老的文化之一。冀州在河北，是豫州、青州向幽州过渡的区域，也是古代文化交流的前沿之一。幽州在燕山以北，是红山文化传统的发源地，很可能是中国玉石文化的根源。幽州是通古斯族群和古西伯利亚族群的发源地。并州从山西到蒙古及西北，北方的许多民族乃至西伯利亚之西的许多族群都可能源于此地，比

如匈奴族代表的叶尼塞语系族群，从细石器文化传统到陶寺龙山文化，这一区域的文化也随着环境而多变。不同的考古文化代表着不同的人类种群，也同时演化出了不同的语言系群(图 1－17)。

晚亚洲人来到东亚以后，在各地隔离发育形成数个种群，渐渐在遗传结构和体质特征上产生差异，同时也演化出了各类原始的语系。在西南部形成了南亚(孟高棉)-苗瑶原始系群，在东南部形成南岛-侗傣原始系群，在西北部形成汉藏-乌拉尔原始系群，在北方形成叶尼塞-古亚原始系群，在东北形成阿尔泰原始系群。新石器时代晚期，叶尼塞、汉藏、苗瑶、侗傣 4 个系群在中国内地开始共同发育中华文明，其对应的 4 个语系在未知因素的影响下进化出声调等独特形态。而相关联的古亚、乌拉尔、南亚、南岛以及阿尔泰都因为远离原始时代的中华文化圈，没有加入这一演化的早期进程。在其后的历史中，9 个语系人群迁徙、交流、同化、融合，形成了现代错综复杂的分布格局。

九州的文化虽然是各自起源和发展的，但是九州的人民却由共同的祖先分化而成。利用分子人类学对九州各地的古代和现代人群做了详细的调查分析，发现从新石器时代开始，九州的大部分居民是 Y 染色体以 O 型和 N 型为主的“晚亚洲人”。我们都是两三万年前随着冰川的渐渐消退从南方迁来的[18]。到了大约 1 万年前开始的新石器时代，九州各地的人群渐渐在体质上发生了细微的差别。笔者对几个主要考古文化的分析发现，与现在混合如一的中华民族不同，当时几个不同区域的考古文化的人群遗传结构已经有了明显的差异。O 型 Y 染色体下面有若干亚型，在扬州范围内的良渚文化遗址的遗骨中发现了 O1－M119 亚型，在荆州范围内的大溪文化遗址中发现了 O3－M7 亚型，在并州范围内的陶寺遗址中发现了 O3－M134 亚型。这些不同的亚型在各个文化之间形成了明显的遗传分界[20]。在现代民族中，O1 亚型主要分布在侗傣民族和南岛民族中，有些台湾少数民族群体几乎只有这种亚型[21]，这些民族都是古代百越民族的后代，所以扬州就与百越民族、良渚文化、O1 亚型 Y 染色体联系在一起。O3－M7 亚型出现在现在的苗瑶族群和孟高棉族群中，荆州是苗瑶民族的故土，大溪文化也必然是苗瑶的祖先留下的。既然九州的居民都有同一个起源，又如何在各地变成截然不同的民族呢？民族分化的历史可能要追溯到 3 万多年前我们的祖先在东南亚的丛林中探索的时候。

人口不多的中华先民非常容易迷失在东南亚的丛林、山岭和谷地之间。遥远的村落之间渐渐失去联系，随着扩散的距离越来越远，一个个小小的瓶颈效应发生了作用，人群的遗传结构和体质特征开始产生差异。今天，族群不断地迁徙和融合使得大部分群体都有着来自不同地理区域的遗传成分，特别是部分汉族和藏缅民族从北方回到南方[22,23]，他们的 Y 染色体主要是 O3 亚型，其中很多成分融入了南方的本地民族中。这使我们很难再分析人群最初在东南亚向东亚进发时是如何发生群体分化的。但是我们找到了一个很好的群体，在东南亚和东亚之间，中华大地最南端的海南岛是一个群体融合之外的避风港。海南岛的原住民，黎族和仡隆人群体中几乎没有 O3 亚型 Y 染色体，而主要是 O1 和 O2 两种亚型[24]。这说明他们几乎没有受到汉族南下的遗传影响，保留着其中

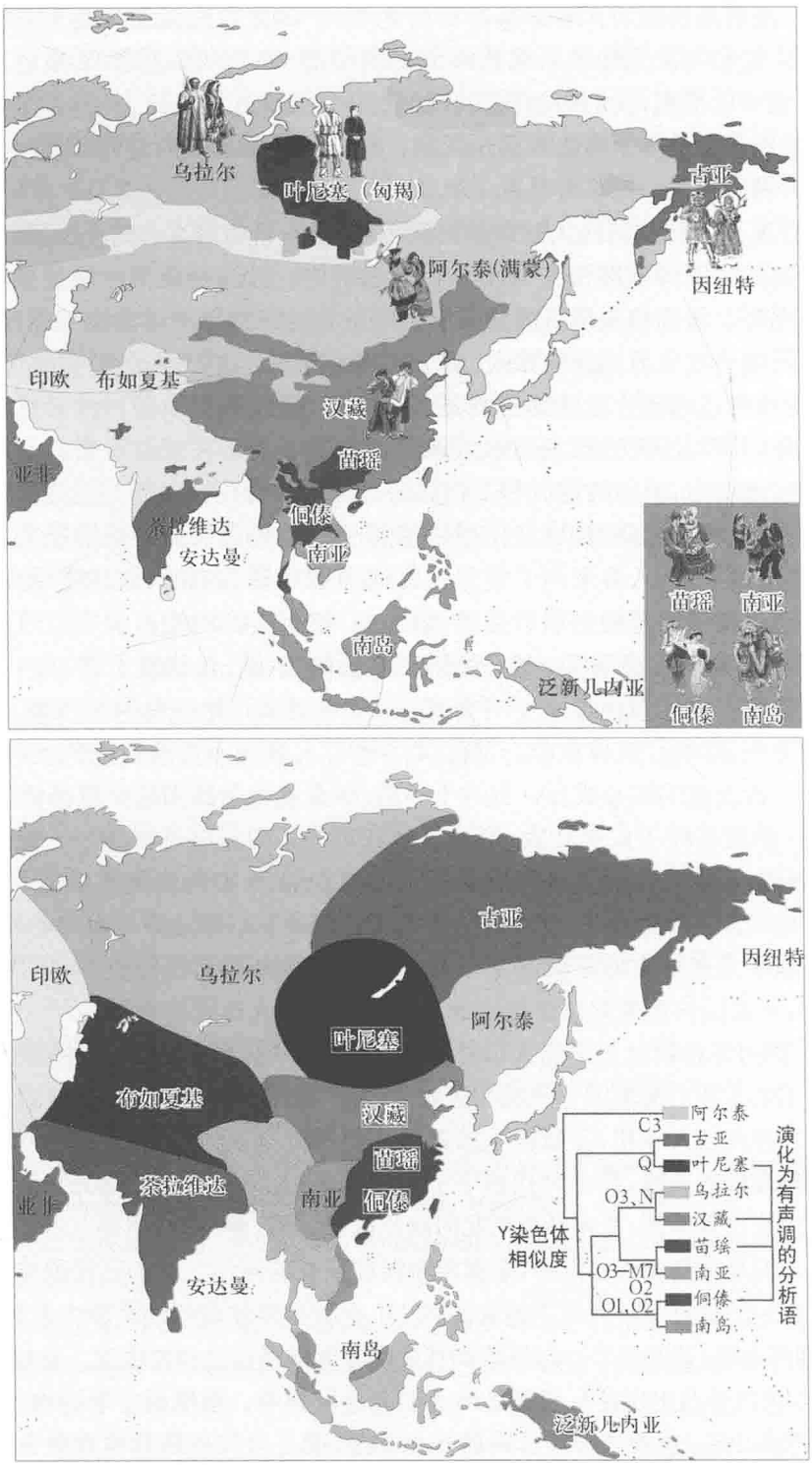


图 1－17 东亚现代的语系分布(上)与可能的上古时期分布(下)

一支先民进入东亚时候的遗传结构。在我们的祖先进入九州大地的时候，由于冰川期海平面的降低，海南岛周围的大陆架露在海面之上，一部分先民就通过这一条快捷通道来到扬州地区。他们与从缅甸进入梁州地区的携带着 O3 和 O2 亚型 Y 染色体的群体不同。所以我们可以推测，我们的祖先至少通过两个入口进入九州大地，一个是从越南进入扬州区域，另一个是从缅甸进入梁州区域。隔离这两路人群的地理障碍很可能是印度支那屋脊——长山山脉。

O2 亚型 Y 染色体在两路人群中都出现了，这是一个非常古老的亚型，其起源可以追溯到印度半岛的孟高棉族群祖先[25]——门达族群中。孟高棉族群可能最接近整个东亚人群的起源系群。孟高棉族群目前分布于印度东北地区到马来半岛北部的东南亚地区，最北是广西云南贵州交界地区的侬人，属于梁州的范围。这个族群在古代的分布可能还更广泛，云南和贵州大部分地区都可能是其分布区，在民族学上被叫作百濮民族。现在这一区域常见的 O2 亚型 Y 染色体很可能是他们留下的。在穿过云贵高原的山谷和丛林来到四川盆地以后，群体的遗传和文化特征都发生了变化。在四川盆地上留下的神奇的营盘山文化、宝墩文化、三星堆文化很可能属于缅彝语支倮倮族群的祖先。大约 1 万年前，从四川继续北上，人群来到了甘青一带的羌塘高原，古羌人在这里孕育成长。“蜀道难”，人群通过蜀道的过程必然产生瓶颈效应。雍州人群的遗传类型只是梁州的一小部分，O2 亚型 Y 染色体就不再出现。所以在以后的羌、藏、汉诸族里面，几乎没有 O2 亚型的踪影，都以 O3 亚型为主。七八千年前，部分人群又开始沿着渭河和黄河往东，来到雍州东部和豫州，发展起粟作农业。陆陆续续地有人群加入这些农业的社群，最终发展成华夏族——古汉族的核心成分。五六千年前，华夏族与羌族彻底分道扬镳[26]。在以后的几千年里，华夏族终于发展壮大，形成了现在世界人口最多的民族——汉族。神奇的是，在父系方面，汉族的主体还是保留着华夏族原初的 Y 染色体遗传结构[22]，只有少数分支受到其他民族的明显融入[27]。这一条沿着青藏高原东部边缘从南向北的迁徙路线，从今天看来似乎是东亚人群来源的主干线。而在当时人类大迁徙途中这可能只是一个很小的分支，从孟高棉祖先向东分化出的百越祖先族群人口可能大得多。

当然，三四万年前跨过长山的人口并不多，不然也不会形成东西如此大的差异。只是到达长山以东的东亚沿海地区——扬州区域内，由于物产丰饶，人口膨胀，新的 Y 染色体突变也产生了。在两三万年里，百越祖先族群的人口持续增长，并扩张到整个扬州地区。良渚文化是百越发展的巅峰，那个时代留下了大量精美的玉器，文明的曙光也隐约出现。有些人群一直向北进入徐州、青州乃至东北的幽州区域。青、徐、兖三州是上古和三代时期的东夷族地区。但是在商朝灭亡以后，东夷族渐渐被华夏族同化。我们已经很难从现在的人群中追寻东夷族的踪迹了。但是百越起源的 O1 亚型在东北幽州的民族中发现，西南方的 O2 亚型在朝鲜出现，这些都是百越族群和其他南方族群通过青州的线索。百越民族特色的 O1 亚型 Y 染色体普遍出现在东南亚岛屿上的南岛民族中。根据语言学的推测，南岛民族起源于中国东南沿海，也就是百越分布的扬州地区，通过台湾岛和菲律宾散布到整个东南亚岛屿和大部分太平洋地区。但是根据 O1 亚型 Y 染色体的 STR 多样性分析，发现台湾少

数民族与马来族群的O1型并没有直接联系，台湾少数民族直接连接在大陆东南的侗傣族群上，而马来族群则是从北部湾地区的侗傣群体中直接起源的[21]。

进入梁州地区的孟高棉族群祖先除了继续向北到达雍州，另一批人向东越过巫山进入云梦大泽所在的荆州。这一批人群就是苗瑶民族的祖先。苗瑶民族除了带有O2和O3亚型Y染色体，还有一种比较特殊的O3-M7小亚型。这种类型还出现在东南亚的某些孟高棉民族中，STR多样性分析证明苗瑶民族确实起源于孟高棉族群。传说中苗族曾经往北方扩张过，但是笔者并没有发现任何证据。苗瑶族群的祖先也被称为荆蛮，他们用极其丰富的想象力构造了绚烂多彩的神灵世界。中国传说中的许多神话来源于荆楚文化。

冀州是南方来的两批人群交汇的地区，西路人群和东路人群带来的许多遗传特征在这里融合，并向幽州地区扩散。人群的交融带来文化的碰撞，九千多年前在辽河流域迸发出了文明初肇的火花，玉器制作技术的发展和宗教思想的产生是这一地区对九州文明最重要的贡献。并州地区的居民主要来源于雍州。在瓶颈效应的影响下，他们的遗传结构继续发生变化，原本从梁州到雍州迁徙人群非常罕见的N型，在这里越来越多，并扩张到西部西伯利亚以及欧洲。在欧洲和亚洲交界处的乌拉尔族群中N型是主要类型，但是这些成分都来自梁雍并州迁徙而来的人群[28]。

从虞夏时期开始，随着文明的产生和发展，距离和地理的阻隔已经不能阻碍族群频繁的交流。四千多年前的龙山文化超越了一州的区域，在各地发展出各种分支文化，并且影响边远地区。它昭示出中华文明将在各地文化交流融合的基础上诞生。通过四千多年的发展和交流，九州大地的人群已经你中有我，我中有你，难分彼此。

参考文献

[1] Jakobsson M, Scholz S W, Scheet P, et al. Genotype, haplotype and copy-number variation in worldwide human populations. Nature, 2008, 451(7181): 998-1003.

[2] 潘尚领. 东亚及东南亚地区G6PD缺陷的地域分布以及疟疾的正选择作用. 现代人类学通讯, 2007,1: 42-52.

[3] 苍铭. 云南边地移民史. 北京: 民族出版社,2004.

[4] Kitchen A, Miyamoto M M, Mulligan C J. A three-stage colonization model for the peopling of the Americas. PLoS One, 2008, 3(2): e1596.

[5] Bowcock A, Ruiz-Linares A, Tomfohrde J, et al. High resolution of human evolutionary trees with polymorphic microsatellites. Nature, 1994, 368(6470): 455-457.

[6] Cann R L, Stoneking M, Wilson A C. Mitochondrial DNA and human evolution. Nature, 1987, 325(6099): 31-36.

[7] Goldstein D B, Ruiz L A, Cavalli-Sforza L L, et al. Genetic absolute dating based on microsatellites and the origin of modern humans. Proc Natl Acad Sci USA, 1995, 92(15): 6723-6727.

[8] Thomas E M. The old way: a story of the first people. New York: Picador, 2006.

[9] Mellars P. Going east: new genetic and archaeological perspectives on the modern human

colonization of Eurasia. Science, 2006, 313: 796 - 800.

[10] 李辉. 走向远东的两个现代人种. "国立"国父纪念馆馆刊, 2004, 14: 164 - 180.

[11] Jin L, Su B. Natives or immigrants: modern human origin in East Asia. Nat Rev Genet, 2000, 1(2): 126 - 133.

[12] The Y Chromosome Consortium. A nomenclature system of the tree of human Y chromosomal binary haplogroup. Genome Res, 2002, 12: 339 - 348.

[13] Marean C W, Bar-Matthews M, Bernatchez J, et al. Early human use of marine resources and pigment in South Africa during the Middle Pleistocene. Nature, 2007, 449(7164): 905 - 908.

[14] Cordaux R, Stoneking M. South Asia, the Andamanese, and the genetic evidence for an "early" human dispersal out of Africa. Am J Hum Genet, 2003, 72(6): 1586 - 1590.

[15] Thangaraj K, Singh L, Reddy A G, et al. Genetic affinities of the Andaman Islanders, a vanishing human population. Curr Biol, 2003, 13: 86 - 93.

[16] Tajima A, Pan I H, Fucharoen G, et al. Three major lineages of Asian Y chromosomes: implications for the peopling of East and Southeast Asia. Hum Genet, 2002, 110(1): 80 - 88.

[17] Tajima A, Hayami M, Tokunaga K, et al. Genetic origins of the Ainu inferred from combined DNA analyses of maternal and paternal lineages. J Hum Genet, 2004, 49(4): 187 - 193.

[18] Su B, Xiao J, Underhill P, et al. Y chromosome evidence for a northward migration of modern humans into Eastern Asia during the Last Ice Age. Am J Hum Genet, 1999, 65(6): 1718 - 1724.

[19] 苏秉琦. 中国文明起源新探. 上海: 三联书店, 2000.

[20] Li H, Huang Y, Mustavich L F, et al. Y chromosomes of prehistoric people along the Yangtze River. Hum Genet, 2007, 122(3 - 4): 383 - 388.

[21] Li H, Wen B, Chen S J, et al. Paternal genetic affinity between western Austronesians and Daic populations. BMC Evol Biol, 2008, 8: 146.

[22] Wen B, Li H, Lu D, et al. Genetic evidence supports demic diffusion of Han culture. Nature, 2004, 431(7006): 302 - 305.

[23] Wen B, Xie X, Gao S, et al. Analyses of genetic structure of Tibeto-Burman populations reveals sex-biased admixture in southern Tibeto-Burmans. Am J Hum Genet, 2004, 74(5): 856 - 865.

[24] Li D, Li H, Ou C, et al. Paternal genetic structure of Hainan aborigines isolated at the entrance to East Asia. PLoS One, 2008, 3(5): e2168.

[25] Kumar V, Reddy A N, Babu J P, et al. Y chromosome evidence suggests a common paternal heritage of Austro-Asiatic populations. BMC Evol Biol, 2007, 7: 47.

[26] Su B, Xiao C, Deka R, et al. Y chromosome haplotypes reveal prehistorical migrations to the Himalayas. Hum Genet, 2000, 107(6): 582 - 590.

[27] Gan R J, Pan S L, Mustavich L F, et al. Pinghua population as an exception of Han Chinese's coherent genetic structure. J Hum Genet, 2008, 53(4): 303 - 313.

[28] Rootsi S, Zhivotovsky L A, Baldovič M, et al. A counter-clockwise northern route of the Y chromosome haplogroup N from Southeast Asia towards Europe. Eur J Hum Genet, 2007, 15: 204 - 211.

1.3 Y染色体上的自然选择

在人类学研究中，父系单向传递的Y染色体有效群体小、突变率低、位点丰富以及有群体特异性的类型分布，因而一直被广泛应用于追溯群体的父系源流、揭示性别间的迁徙及群体遗传差异[1-3]。大部分这样的群体研究都是基于这样的假设：Y染色体非重组区的标记位点是不受选择的[4]。然而，Y染色体并非没有表型效应，例如，男性不育和性腺发育不全等都与Y染色体相关[5,6]。即使Y染色体上用于群体研究的标记位点是没有功能的，但它们可能与其他有益或有害突变及结构变异相关联。因为Y染色体的主干部分不发生重组，任何正向或负向选择作用都将会对整条染色体产生影响[4]。事实上，Y染色体是否经受自然选择已在群体遗传学家中争论了近30年。近年来，随着测序技术的不断发展，越来越多的Y染色体数据使得我们可以一步步解析这一难题。

1.3.1 Y染色体的结构

X染色体和Y染色体是人类性染色体，由一对同源的常染色体历经2亿～3亿年的演化而来[7,8]（图1-18）。人类的Y染色体长约59 Mb，有95%的部分不与X染色体重

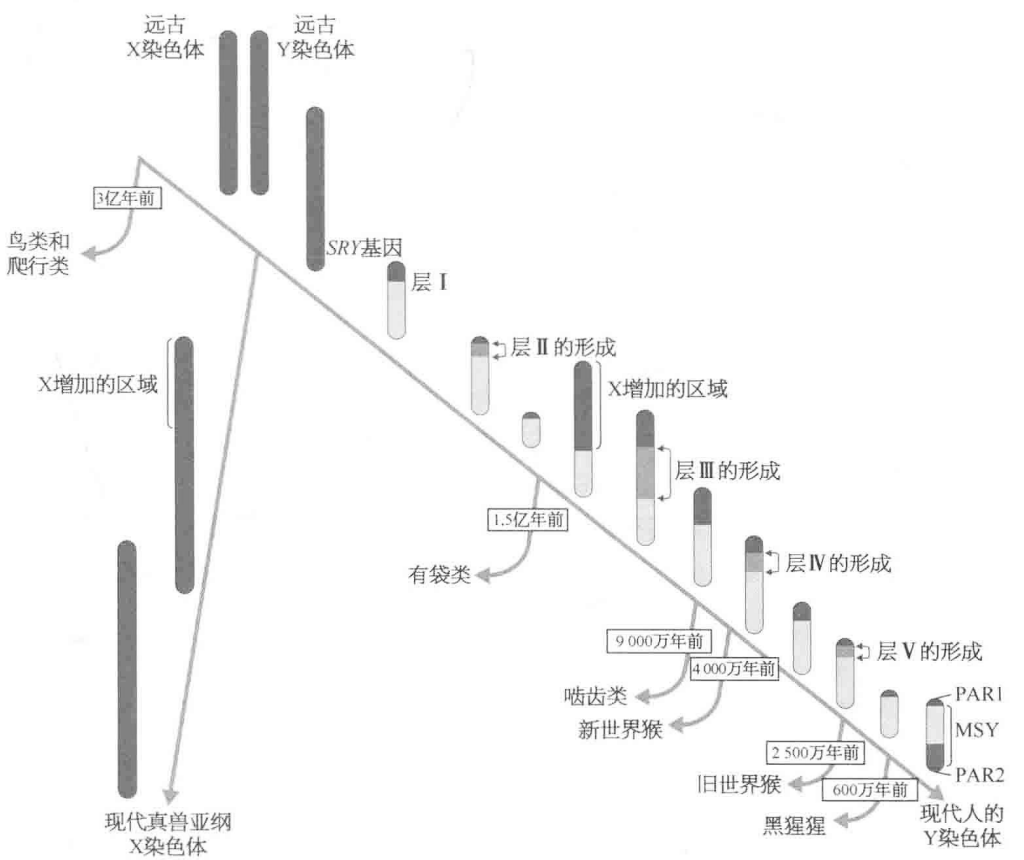


图1-18 Y染色体的起源与演化[12]

组而仅能父子相传[9]。这一部分一直以来被认为是非重组区(NRY),然而新近研究中发现这一区域有大量的基因转换,这使得我们重新命名其为男性特异区(MSY)[10]。拟常染色体区(PAR)位于 MSY 的两端,其中 PAR1 有 2.6 Mb,在 X 染色体和 Y 染色体短臂的两端都存在,而位于长臂的 PAR2 则仅有 320 kb。PAR1 和 PAR2 被认为是 X 染色体和 Y 染色体上古老同源序列的遗存,Y 染色体上 PAR1 和 PAR2 现在仍可与 X 染色体上的 PAR1 和 PAR2 重组[11](图 1-19)。

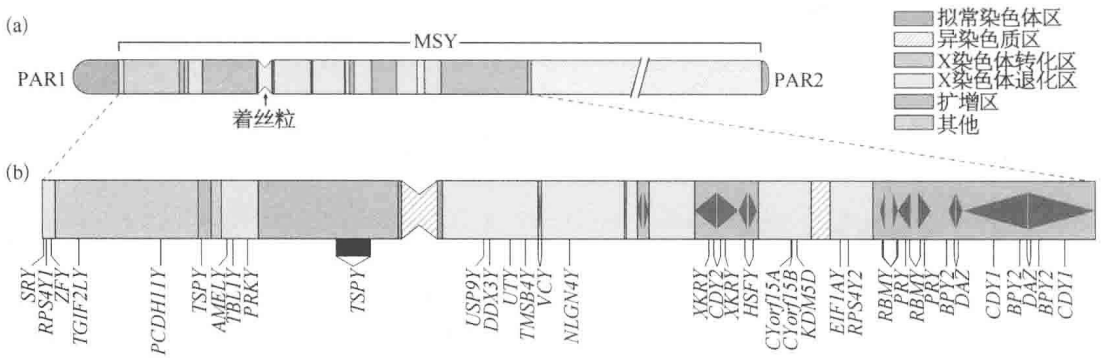


图 1-19　人类 Y 染色体的结构[12]

(a) Y 染色体的整体结构;(b) Y 染色体常染色质区的结构(下部是蛋白质编码基因)

MSY 是高度异质的,由异染色质和常染色质这两种截然不同的序列组成。异染色质部分主要指的是长臂上长达 40 Mb 的异染色质,没有基因分布。常染色质部分则由 X 染色体转化区、X 染色体退化区和扩增区组成[9]。Y 染色体的 X 染色体转化区有 3.4 Mb,是在人与黑猩猩分离之后(300 万~400 万年前)经由一次 X 染色体向 Y 染色体的转化而形成的,其与 X 染色体 q21 区有 99%的序列相似性,但 X 染色体转化区仅有两个基因[12,13]。X 染色体退化区共有 8.6 Mb,被认为是演化成 X 染色体和 Y 染色体的古老常染色体的遗存,其上零星分布着 16 个单拷贝基因和 X 染色体关联基因的假基因。多数的 X 染色体退化基因在人体内广泛表达,但例外的是性别决定基因 *SRY* 则主要在睾丸中表达[9,14]。扩增区约有 10.2 Mb,主要由长的重复序列组成。多数重复序列是回文结构,且形成了 8 个大的回文序列。扩增区的编码基因和非编码基因的密度是最高的,共有 60 个基因分布在 9 个不同的多拷贝基因家族中。与 X 染色体退化区广泛表达的基因相反,扩增区的基因和转录单元主要或仅特异性地在睾丸中表达,因此其对男性特征的演化有重要意义[9,10,12](图 1-19)。

1.3.2　Y 染色体多样性低

自 20 世纪 80 年代中期,Y 染色体上位点的多样性低就开始被广泛报道。那时候,主要靠从质粒文库里分离出 Y 染色体特异性的探针结合限制性酶切来寻找男性特异的限制性片段长度多态(RFLP)[15-19]。例如,Malaspina 等[19]应用 12 个探针和 12 个限制性酶试图在 131 个男性 Y 染色体寻找多态位点。但即使这么系统的研究仍未发现任何

一个多态位点。其他一系列研究也证实了这一结论，与X染色体相比，Y染色体多样性确实很低。当时，学者们提出两个可能的原因来解释这一现象[19]。一是Y染色体除了两端很小的PAR区外都是不会重组的，然而，重组本身是可以引入突变的，这样，严格父系相传的Y染色体因不重组而远离了突变；二是遗传搭车效应，即正选择可作用于有益突变而降低与有益突变相连锁位点的多态性，对Y染色体而言，选择压力作用于任一位点都会降低整个MSY的多样性。

然而，上述RFLP位点反映的都是Y染色体q11区域上的非特异性序列。另外，Y染色体的多样性低是通过与常染色体和X染色体比较得出的，并未使用种间差异来对所研究区域的突变率进行校正。自20世纪90年代中期开始，对人类和其他灵长类Y染色体特定区域的直接测序进一步深化了我们对Y染色体上可能的自然选择这一问题的认识。1995年，Dorit等[20]对38个人类男性样本和黑猩猩、大猩猩以及红毛猩猩等非人灵长类的Y染色体*ZFY*基因内含子的729 bp进行测序，种间比较发现这一内含子区多态位点丰富，然而，在世界范围的人类男性样本中却并未发现突变。Dorit认为这反映了人类男性有着较晚近的共同祖先(27万年，95%置信区间：0～80万年)。尽管时间估算中较大的置信区间可能是由误差导致的，但ZFY区域如此低的多样性也很可能是由晚近的选择性清除造成的。Goodfellow等[21]对5个男性样本和一个黑猩猩的Y染色体上性别决定的SRY区域的18.3 kb进行测序，仅仅发现了3个突变位点。与以上研究不同，Goodfellow将Y染色体多样性与母系遗传的线粒体DNA(mtDNA)进行比较，发现人与人之间Y染色体平均遗传距离是mtDNA的1/250。经过校正Y染色体和mtDNA的突变率，Goodfellow得出Y染色体的共祖时间(3.7万～4.9万年)是mtDNA的共祖时间(12万～47.4万年)的1/10～1/3。与mtDNA相比，Y染色体多样性很低，可能是由于前面提到的搭车效应——由Y染色体上的有益突变引起的选择性清除。另外，人群历史也可能作用于遗传结构，例如，一夫多妻就会使得少部分男性有大量的后代，从而降低了男性的有效群体大小。从此，当学者们试图寻找Y染色体上的选择效应时也开始考虑有效群体大小的可能影响。与Goodfellow的论文同时发表在*Nature*上的还有亚利桑那大学Hammer的论文[22]，但Hammer的实验结果反驳自然选择作用降低Y染色体多样性的见解。Hammer对16个男性和4个黑猩猩的Y染色体2.6 kb的YAP区进行测序，发现任意两个随机选择的YAP区平均有一个SNP位点不一致，后续分析中他选用HKA检验来看是否有选择作用。HKA检验认为，在中性状态下，对于不同的基因或者基因位点而言，即使他们之间的变异程度不同，他们各自的种内多态性(π)与种间分歧度(D)之间的比率还是相同的。Hammer以人与黑猩猩间的差异作为种间分歧度，发现YAP区π：D的比值与mtDNA的*COII*和*ND4－5*基因的π：D值几乎一样。因为Y染色体的有效群体只有常染色体的1/4，中性条件下，Y染色体的多样性也应该是常染色体的1/4，而Hammer的数据中YAP区的π：D值也差不多是β球蛋白的1/4，这些结果均显示Y染色体或许并没有经受选择作用。

上述的直接测序研究都有一个共同的不足之处，即样本量太小。同时，上述研究是

将 Y 染色体与 mtDNA 和常染色体进行比较来分析其多样性，而不同遗传标记系统相比较的方法是否可恰当反映选择效应还处在争议之中。1999 年，Labuda 设计实验巧妙地规避了这些问题，他对 Y 染色体的 *ZFY* 及其在 X 染色体上同源的 *ZFX* 基因的最后一个内含子进行测序，在世界范围的 205 个男性 *ZFY* 内含子的 676 bp 序列里只发现了一个突变，而在 336 条 X 染色体的 ZFX 区的 1 089 bp 里却发现了 10 个突变。ZFX 区内含子的多样性要高于 ZFY 区，但低于其他中性进化的基因组区域。尽管 *ZFY* 和 *ZFX* 的种间遗传距离也降低了，但 HKA 检验和 Tajima 检验都不拒绝中性假设。Labuda 认为 ZFY 区非常低的多样性可能降低了 HKA 和 Tajima 检验的效力，而且自然选择也很难在一个晚近时期扩张的群体里检测到[23]。那么，人群历史是否会影响 Y 染色体的多样性及可能的选择进程？如果是，影响的程度有多大？

1.3.3　群体历史的影响

自 20 世纪 90 年代末，变性高效液相色谱技术(DHPLC)开始用于检测 MSY 单拷贝区域的单核苷酸多态性(SNP)[3,24]。而在刚刚过去的十年间，学者们根据已发现的众多 SNP 已经构建出非常可靠的 Y 染色体谱系树[25,26]，可用于人群历史研究。

男女不同的群体历史，例如，性别偏向的迁徙等，可能会影响对所观察到的父系 Y 染色体、母系 mtDNA 以及常染色体间差异的解读。性别偏向的迁徙指的是人群中女性的迁徙率高于男性[27]。一系列的研究发现，在世界范围内，无论是群体内还是群体间，与 mtDNA 和常染色体相比，MSY 上的 SNP 和 STR 的 F_{ST} 值(遗传距离)较高而多样性较低，这显示 Y 染色体有着更明显的区域性特征[27-32]。更有趣的是著名人类学家 Mark Stoneking 发现[28]，泰国北部的母系群体 mtDNA 的 F_{ST} 值是 MSY 的 2 倍，这说明性别差异很可能是由于父系和母系社会的不同居住方式引起的。现今，大多数人类社会是父系社会，婚后夫妻二人倾向于与丈夫的父母一起居住，这就造成了无论在群体间还是群组间女性总比男性迁徙得更频繁，也就造成了群体间父系 Y 染色体的差异要大于母系 mtDNA 的差异[33]。从父居或许可以解释从局部地区所观察到的遗传上的性别差异，问题是从父居是如何影响洲际或者全球尺度遗传上的性别差异[27]？实际上，将 Y 染色体谱系树上的 SNP 位点用来计算群体多样性是否恰当在学界还是有争议的。因为这些 SNP 位点常常是在少数男性里发现后被用于大群体样本做分型，非随机选择的 SNP 位点可能会造成这样的偏差：SNP 在少数一些群体里最先被确定，应用于后续群体分析中可能会高估这些群体的多样性等[34]。随着测序技术的不断发展，越来越多的数据正在逐渐减小这些偏差。例如，Hammer 通过直接测序全球范围 10 个群体 389 例男性样本 6.7 kb 的 Y 染色体 *Alu* 区以及 mtDNA 770 bp 的 *CO3* 基因，来看全球范围的性别偏向的迁徙是否真正存在。Hammer 发现无论是大洲内还是大洲间 Y 染色体和 mtDNA 的差异都很相近，也就是说大空间尺度的遗传结构并不受居住方式的影响。另外，Y 染色体和 mtDNA 在群体间的多样性呈显著相关关系。这相关性也使 Hammer 觉得群体历史才是造成群体间遗传距离的主要原因而非自然选择[35]。

当我们用群体历史而非自然选择去解释Y染色体和mtDNA间的差异时，需要谨慎处理男女有效群体大小的问题。一夫多妻，男性高死亡率以及男性高繁殖成功率等都会降低男性的有效群体大小，这也使得在与mtDNA和常染色体相比较时Y染色体的多样性低[27,36]。

经过上面的分析，问题变成在讨论自然选择有无作用于Y染色体的时候该如何看待性别偏向的迁徙和有效群体大小的影响。Hammer测序了25个科伊桑人、24个蒙古人和24个巴布亚新几内亚人26.5 kb的MSY非编码区以及782 bp的mtDNA *Cox3*基因，来试图回答这一问题[36]。Hammer使用了Tajima's D和Fu and Li's D*来检验群体参数是否偏离中性，然而只有蒙古样本的Fu and Li's D*是稍显著的负值($0.01<P<0.05$)，科伊桑人的Fu and Li's D*稍显著的正值($0.01<P<0.05$)。但是，这两组数值都没达到统计上的显著($P<0.01$)。在假设男女有效群体大小一致的情况下，Hammer和同事们又在科伊桑人群中进行了瓶颈效应模拟来看是否是选择性清除造成了MSY和mtDNA的谱系差异。然而，无论是强瓶颈效应(100代的时间，群体大小降至1/100)还是弱瓶颈效应(100代的时间，群体大小降至1/10)或者瓶颈效应附加族群衰退等模拟的结果，均与现今在科伊桑群体里观察到的多样性不符。因此，Hammer认为应该是稍高的女性有效群体大小而非自然选择造成了Y染色体和mtDNA的多样性差异。使得Hammer得出上述结论的另一因素是，如果净化选择真的作用于Y染色体和mtDNA，mtDNA应该比MSY承受了更强的应对有害突变的净化选择。这是因为mtDNA上位点的突变率差不多是MSY的10倍，且mtDNA上基因密度大而并不像MSY有那么多的冗余序列。另外，值得一提的是mtDNA的突变可能与男性不育相联系。据报道，即使mtDNA的产能功能稍有损伤就会影响男性精子的活力，但直到mtDNA的功能损伤超过80%才会显出疾病症状[37]。这也就是说，一些mtDNA的突变虽对女性生存影响不大，但却可以造成男性不育。近年来，随着二代测序技术的应用，一批高质量的全基因组数据陆续公布。2013年，加州大学伯克利分校的Sayres等[38]利用16份非洲和欧洲男性的全基因组数据在基因组水平比较来寻找Y染色体上可能的选择效应。若在中性选择下Y染色体和mtDNA的多样性应该是常染色体的1/4，但他们发现Y染色体的多样性极低，甚至只占常染色体多样性的1/40，然而mtDNA的多样性并不见低于中性理论值。在后续的分析中，他们以中性条件下X染色体与常染色体的比值(0.75)、Y染色体和mtDNA与常染色体的比值(0.25)作为标准来检测选择效应。减小男性的有效群体大小确实可以减低Y染色体的多样性，但是当男性有效群体降低到可以解释现在所观察到的Y染色体数据时，常染色体、X染色体和mtDNA的多样性又会与前面设立的标准参考相距甚远。据此，Sayres等认为自然选择降低了Y染色体的多样性。他们进行了进一步的模拟确认了净化选择降低了Y染色体多样性，而且选择作用也很可能作用于Y染色体的扩增区。我们最开始介绍Y染色体结构时已提到扩增区主要负责睾丸的发育，Sayres的预测在这一点上或有意义。

另一棘手的问题是在人口扩张的群体里很难确认Y染色体上有无自然选择作用。

相比自然选择，我们更倾向于将某些 Y 染色体支系的扩张归为社会选择的影响。例如，Y 染色体 C3 星簇的扩张被认为与成吉思汗有关，成吉思汗家族的社会地位增加了其繁衍后代的成功率[39]。类似的，C3c－M48 单倍群可能因清代贵族而扩散开来[40]。另一个例子，O3a1c－002611 单倍群是东亚特有单倍群 O3－M122 下的 3 个主要支系之一，大约占汉族人口的 16%，这一支系在新石器时代晚期人口剧烈扩张则可能与当时农业的繁盛有关[41]。然而，正向选择可能会使带有有益突变的 Y 染色体支系比其他支系扩张得更快，问题是如何将扩张的信号与自然选择的信号区分开。

1.3.4　人类 Y 染色体与其他灵长类的直接比较

群体历史确实给群体遗传研究者们探究 Y 染色体上选择作用带来很大困扰。然而，麻省理工学院 David Page 实验室成功进行了人类、黑猩猩和恒河猴的 Y 染色体全测序，为我们在进化尺度上回答这一问题提供了可能。

人类和恒河猴的 Y 染色体上分为五层，现在普遍认为每一层是由 Y 染色体上的一个序列倒置引起的，也正是这些序列倒置使得 Y 染色体和 X 染色体难以重组。在这五层中，最古老的一层可追溯至 24 000 万年前，最年轻的一层仅形成于 3 000 万年前（图 1－18）。在较古老的四层中，人类和恒河猴一样有 18 个古老基因，即在人与恒河猴分开的 2 500 万年间这四层里没有任何一个基因丢失掉。另外，这 18 个基因中有 17 个基因的非同义突变与同义突变的比值（dN/dS）小于 1，其中 11 个有统计学上的显著意义。这些古老基因进化上如此保守可归结于净化选择作用[8]。

然而，黑猩猩 Y 染色体的 MSY 却在与人分开的 600 万年间丢了 6 个基因，但其扩增区却有 2 倍于人类的回文序列[10]。正如我们一开始就提到的，扩增区的基因主要或仅在睾丸中表达，对精子发生有重要作用。黑猩猩 Y 染色体的扩增区很可能因强烈的精子竞争而经受了很强的正向选择[42]。与此相反，在人类进化的 600 万年里没有任一个 X 染色体退化的基因丢失掉。另外，人与黑猩猩在 X 染色体退化基因编码区的种间遗传距离要显著小于内含子区的种间遗传距离，也说明了在人类进化历程中净化选择在维持 X 染色体退化基因的功能上起了重要作用[43]。

净化选择在更晚近的时期，比如自从 10 万年前人类走出非洲以来，是否还对 Y 染色体有作用？David Page 及其同事还是将目光聚焦到 X 染色体退化基因上来阐述这一问题。他们对 105 个男性 Y 染色体上的 16 个单拷贝 X 染色体退化基因和 5 个单拷贝假基因进行测序，这些男性的 Y 染色体类型涵盖了世界范围的 47 个支系。他们发现非同义位点的核苷酸多样性要显著低于同义位点的核苷酸多样性，也显著低于内含子和假基因的核苷酸多样性。另外，内含子区的核苷酸多样性也比基因组其他区域的多样性要低一个数量级，这表明群体历史因素也可能降低了男性有效群体大小[44]。

1.3.5　结论

近 30 年的争论最终以测序发现了 Y 染色体上自然选择的确切证据而告终。Y 染色

体上古老的X染色体退化基因在进化上非常保守是净化选择在起作用，而扩增区与睾丸发育相关的基因则可能受到正向选择的影响。这一发现提示在用Y染色体数据进行群体遗传学分析时也要将自然选择考虑在内。

参考文献

[1] Jobling M A, Tyler-Smith C. Father and sons: the Y chromosome and human evolution. Trends Genet, 1995, 11: 449 - 456.

[2] Jin L, Su B. Natives or immigrants: origin and migrations of modern humans in East Asia. Nat Rev Genet, 2000, 1: 126 - 133.

[3] Underhill P A, Shen P, Lin A A, et al. Y chromosome sequence variation and the history of human populations. Nat Genet, 2000, 26: 358 - 361.

[4] Jobling M A, Tyler-Smith C. The human Y chromosome: an evolutionary marker comes of age. Nat Rev Genet, 2003, 4: 598 - 612.

[5] Ferlin A, Arredi B, Speltra E, et al. Molecular and clinical characterization of Y chromosome microdeletions in infertile men: a 10-year experience in Italy. J Clin Endocrinol Metab, 2007, 92: 762 - 770.

[6] Hjerrild B E, Mortensen K H, Gravholt CH. Turner syndrome and clinical treatment. Br Med Bull, 2008, 86: 77 - 93.

[7] Lahn B T, Page D C. Four evolutionary strata on the human X chromosome. Science, 1999, 286: 964 - 967.

[8] Hughes J F, Skaletsky H, Brown L G, et al. Strict evolutionary conservation followed rapid gene loss on human and rhesus Y chromosomes. Nature, 2012, 483: 82 - 86.

[9] Skaletsky H, Kuroda-Kawaguchi T, Minx P J, et al. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature, 2003, 423: 825 - 837.

[10] Rozen S, Skaletsky H, Marszalek J D, et al. Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. Nature, 2003, 423: 873 - 876.

[11] Mangs H A, Morris B J. The human pseudoautosomal region (PAR): origin, function and future. Curr Genomics, 2007, 8: 129 - 136.

[12] Hughes J F, Rozen S. Genomics and genetics of human and primate Y chromosomes. Annu Rev Genomics Hum Genet, 2012, 13: 83 - 108.

[13] Page D C, Harper M E, Love J, et al. Occurrence of a transposition from the X chromosome long arm to the Y chromosome short arm during human evolution. Nature, 1984, 311: 119 - 123.

[14] Lahn B T, Page D C. Functional coherence of the human Y chromosome. Science, 1997, 278: 675 - 680.

[15] Lucotte G, Ngo N Y. p49f, a highly polymorphic probe, that detects Taq1 RFLPs on the human Y chromosome. Nucleic Acids Res, 1985, 13(22): 8285.

[16] Ngo K Y, Vergnaud G, Johnsson C, et al. A DNA probe detecting multiple haplotypes of the human Y chromosome. Am J Hum Genet, 1986, 38(4): 407 - 418.

[17] Torroni A, Semino O, Scozzari R, et al. Y chromosome DNA polymorphisms in human populations: differences between Caucasoids and Africans detected by 49a and 49f probes. Ann Hum Genet, 1990, 54(Pt 4): 287 - 296.

[18] Oakey R, Tyler-Smith C. Y chromosome DNA haplotyping suggests that most European and Asian men are descended from one of two males. Genomics, 1990, 7(3): 325 - 330.

[19] Malaspina P, Persichetti F, Novelletto A, et al. The human Y chromosome shows a low level of DNA polymorphism. Ann Hum Genet, 1990, 54(Pt 4): 297 - 305.

[20] Dorit R L, Akashi H, Gilbert W. Absence of polymorphism at the ZFY locus on the human Y chromosome. Science, 1995, 268(5214): 1183 - 1185.

[21] Whitfield L S, Sulston J E, Goodfellow P N. Sequence variation of the human Y chromosome. Nature, 1995, 378(6555): 379 - 380.

[22] Hammer M F. A recent common ancestry for human Y chromosomes. Nature, 1995, 378: 376 - 378.

[23] Jaruzelska J, Zietkiewicz E, Labuda D. Is selection responsible for the low level of variation in the last intron of the ZFY locus? Mol Biol Evol, 1999, 16(11): 1633 - 1640.

[24] Underhill P A, Jin L, Lin A A, et al. Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography. Genome Res, 1997, 7(10): 996 - 1005.

[25] Karafet T M, Mendez F L, Meilerman M B, et al. New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. Genome Res, 2008, 18: 830 - 838.

[26] Yan S, Wang C C, Li H, et al. An updated tree of Y chromosome haplogroup O and revised phylogenetic positions of mutations P164 and PK4. Eur J Hum Genet, 2011, 19 (9): 1013 - 1015.

[27] Seielstad M T, Minch E, Cavalli-Sforza L L. Genetic evidence for a higher female migration rate in humans. Nat Genet, 1998, 20(3): 278 - 280.

[28] Oota H, Settheetham-Ishida W, Tiwawech D, et al. Human mtDNA and Y chromosome variation is correlated with matrilocal versus patrilocal residence. Nat Genet, 2001, 29: 20 - 21.

[29] Nasidze I, Ling E Y, Quinque D, et al. Mitochondrial DNA and Y chromosome variation in the Caucasus. Ann Hum Genet, 2004, 68: 205 - 221.

[30] Destro-Bisol G, Donati F, Coia V, et al. Variation of female and male lineages in sub-Saharan populations: the importance of sociocultural factors. Mol Biol Evol, 2004, 21: 1673 - 1682.

[31] Wen B, Xie X, Gao S, et al. Analyses of genetic structure of Tibeto-Burman populations reveals sex-biased admixture in southern Tibeto-Burmans. Am J Hum Genet, 2004, 74: 856 - 865.

[32] Wood E T, Stover D A, Ehret C, et al. Contrasting patterns of Y chromosome and mtDNA variation in Africa: evidence for sex-biased demographic processes. Eur J Hum Genet, 2005, 13: 867 - 876.

[33] Stoneking M. Women on the move. Nat Genet, 1998, 20(3): 219 - 220.

[34] Jobling M A. The impact of recent events on human genetic diversity. Philos Trans R Soc Lond B

Biol Sci, 2012, 367(1590): 793 - 799.

[35] Wilder J A, Kingan S B, Mobasher Z, et al. Global patterns of human mitochondrial DNA and Y chromosome structure are not influenced by higher migration rates of females versus males. Nat Genet, 2004, 36(10): 1122 - 1125.

[36] Wilder J A, Mobasher Z, Hammer M F. Genetic evidence for unequal effective population sizes of human females and males. Mol Biol Evol, 2004b, 21(11): 2047 - 2057.

[37] Gemmell N J, Sin F Y. Mitochondrial mutations may drive Y chromosome evolution. Bioessays, 2002, 24(3): 275 - 279.

[38] Sayres M A W, Lohmueller K E, Nielsen R. Natural selection reduced diversity on human Y chromosomes. arXiv Preprint arXiv, 2013, 1303.5012.

[39] Zerjal T, Xue Y, Bertorelle G, et al. The genetic legacy of the Mongols. Am J Hum Genet, 2003, 72: 717 - 721.

[40] Xue Y, Zerjal T, Bao W, et al. Recent spread of a Y chromosomal lineage in northern China and Mongolia. Am J Hum Genet, 2005, 77: 1112 - 1116.

[41] Wang C C, Yan S, Qin Z D, et al. Late Neolithic expansion of ancient Chinese revealed by Y chromosome haplogroup O3a1c-002611. J Syst Evol, doi: 10.1111/j.1759 - 6831.2012.00244.x.

[42] Hughes J F, Skaletsky H, Pyntikova T, et al. Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content. Nature, 2010, 463: 536 - 539.

[43] Hughes J F, Skaletsky H, Pyntikova T, et al. Conservation of Y-linked genes during human evolution revealed by comparative sequencing in chimpanzee. Nature, 2005, 437(7055): 100 - 103.

[44] Rozen S, Marszalek J D, Alagappan R K, et al. Remarkably little variation in proteins encoded by the Y chromosome's single-copy genes, implying effective purifying selection. Am J Hum Genet, 2009, 85(6): 923 - 928.

第2章　Y染色体与种族起源

人类起源与演化是最受关注的科学问题之一。近年来的遗传学研究成果成为理解人类演化历史的最坚实证据。由于黑猩猩等类人猿与现代人的基因组差异极小，所以猩猩科与人科合并了，而黑猩猩属更属于其中的人族——广义的人类。人族源于大约700万年前，其中，真人属在200多万年前源于南猿属或平脸人属，是狭义的人类。人类前期演化出树居人、能人、卢道夫人、匠人等，后期演化出直立人和智人两大分支。基于对智人中的现代人、尼安德特人、丹尼索瓦人的基因组分析比较，发现3种智人是在60万～80万年前分化的，所以智人可以相应分为南方智人(*Homo sapiens australis*)、北方智人(*H. s. septentrionalis*)和东方智人(*H. s. orientalis*)3支。现代人都属于南方智人，即非洲的罗得西亚人，大约20万年前发生了体质变化，在7万年前走出非洲，扩散到全世界，形成现今的8个人种。Y染色体的谱系演化与种族的形成是同步发生的，因此两者有较好的对应关系。正确认识人类历史与种族差异，反对宣扬种族优劣的种族主义，有助于促进人类社会的和谐，也有助于推进医学等相关科学的发展。

东亚人群有着极其丰富的遗传、体质、文化和语言多样性，但这些多样性的具体分布状况以及相互间的渊源关系仍有待进一步解析。随着东亚及其周边人群的分子人类学数据的不断积累，尤其是父系Y染色体研究的一系列进展，使得东亚人群的多样性结构逐渐明晰。为了对东亚现代人起源的诸多假说进行检测，笔者对来自163个人群12 127例男性样本进行了3种Y染色体双等位基因标记(YAP、M89和M130)的分型。所有的样例都携带这3个突变型中的一个。这3个突变(YAP＋、M89T和M130T)均来自3.5万～8.9万年前起源于非洲的另一个突变(M168T)。因此Y染色体数据并不支持本土远古人种(直立人或丹尼索瓦人)对东亚现代人起源有贡献，即便是极小的贡献。

现代人到达东亚并随后扩张的时间和途径一直存有争议。通过使用Y染色体双等位基因标记，笔者研究了古人类在东亚的迁徙模式。研究结果表明东亚南方人群比北方人群的多态性更加显著，北方人群只拥有南方人群单倍型的子集。这种模式表明，5万～10万年前，第一批在东亚定居的现代人于末次冰川期出现在东南亚大陆，恰好与这一时间之前东亚人类化石的缺失相吻合。在最初的定居之后，大规模向北方的迁移使人类扩张至中国北部和西伯利亚。现有的Y染色体数据揭示，现代人走出非洲后由东南亚经多

次迁徙进入东亚。在旧石器时代，现代人最初定居东亚之后，紧接着不断北迁，这奠定了东亚遗传结构的基础。

对考古学、解剖学、语言学和遗传学的数据分析表明，在中国北方群体和南方群体之间存在着明显的分界线。然而，这个分界线的具体位置和隔离程度目前还有争议。在本章中，笔者使用了91个群体的线粒体DNA数据和143个群体的Y染色体数据，系统地探究了中国人群的空间遗传结构以及北方和南方人群之间分化的分界线。研究结果显示了中国人群的母系和父系遗传结构在空间上存在显著差异。中国人群在母系遗传结构上的特征是：北方和南方群体之间存在明显的遗传分化，该明显不间断的遗传分界线大致是长江以北的秦岭-淮河一线。而在父系遗传结构上，在北方和南方群体之间并不存在明显的遗传分化。

来自中亚的移民与东亚人群的基因交流增大了东亚南北人群间的遗传距离。语言、农业、军事与社会地位等文化因素同样影响东亚的遗传结构。将Y染色体与家谱文献相结合，为遗传学研究人类古代历史提供了可能。

2.1 东亚现代人起源于非洲

非洲起源说认为，解剖学上的现代人在十多万年前起源于非洲，然后向非洲之外扩散，完全取代了各地的远古人群[1,2]。这一命题受到遗传学证据和考古学发现的支持[3-9]。最近古DNA的分析支持了欧洲的人群替代，排除了尼安德特人对现代欧洲人的贡献[10,11]。但是有观点认为，中国和东亚其他地区丰富的人类化石（如北京猿人和爪哇猿人），不仅在形态特征上而且在时空分布上，都展现了一定的连续性[12-16]。本节中，利用Y染色体多态性对亚洲现代人起源的几种不兼容的假说进行检测。

对涵盖东南亚、大洋洲、西伯利亚和中亚的163个人群12 127个男性样本检测了3个Y染色体双等位基因标记（YAP、M89和M130）（图2-1）。采样的163个人群来自中亚（克里米亚鞑靼人、伊朗人、东干人、塔吉克族、土库曼人、卡拉卡尔帕克人、乌兹别克东部人群、辛提罗姆人、花剌子模乌兹别克人、维吾尔人、哈萨克人、阿拉伯布哈拉人和吉尔吉斯人）；西伯利亚中部（图瓦人、图法拉人、叶尼塞鄂温克人、布里亚特人群-1和布里亚特人群-2）；鄂霍次克/阿穆尔河（鄂霍次克鄂温克人、沃且/那乃人、上游涅吉达尔人、下游涅吉达尔人、乌德盖人和尼夫赫人）；堪察加/楚科奇（科里亚克人、伊特勒人、楚科奇海和西伯利亚因纽特人）；东亚北部（鄂温克族、满族-1、满族-2、韩国人、日本人、回族-1、回族-2、景颇族、土族、撒拉族、蒙古族-1、蒙古族-2、青海藏民、西藏藏民、云南藏民、新疆哈萨克族和维吾尔族）；中国北方汉族（黑龙江人、辽宁人、河北人、北京人、天津人、山东人、山西人、甘肃人、新疆人、河南人、内蒙古人、青海人、陕西人和吉林人）；中国南方汉族（安徽人、浙江人、江苏人、上海人、湖北人、四川人、江西人、湖南人、福建人、云南人、广西人、广东人和贵州人）；中国台湾少数民族（布农族、泰雅族、雅美族、排湾族和阿美族）；东南亚（土家族、南丹瑶族、金秀茶山瑶族、壮族、侗族、佤族-1、佤族-2、佤族-3、傈尼人、布朗族-1、布朗族-2、拉祜族-1、拉祜族-2、拉祜族-3、拉祜族-4、德昂族、彝族、畲族、黎

族、柬埔寨人、傣族-1、傣族-2、阿卡族、克伦族、傈僳族、基诺族、苗族、瑶族、京族、芒族、纳西族、阿霍姆人、舍族、北部泰国人、东北部泰国人、白族-1 和白族-2)；印度尼西亚/马来西亚(马来沙巴人、马来亚人、阿斯利人、巴塔克人、马来人-北干巴鲁人、米南加保人、巨港人、邦加人、尼亚斯人、达雅克人、爪哇人、腾格里人、巴厘人、萨沙克人、松巴哇人、松巴人、阿洛尔人、望加锡人、武吉士人、特拉珍人、凯里人、万鸦老人、伊里安人、亚庇人和萨卡依人)；波利尼西亚/密克罗尼西亚(特鲁克人、关岛人、普罗乌人、马朱罗人、基里巴斯人、波纳佩人、瑙鲁人、卡平阿玛朗依人、汤加人、美属萨摩亚人和西萨摩亚人)；巴布亚新几内亚高地(澳大利亚原住民、那西奥依美拉尼西亚人、新几内亚人-1、新几内亚人-2、班克斯和托雷斯人、圣吐人、迈沃人)；我国藏南地区和印度东北部(珞巴人、尼希人、阿萨姆人、阿帕塔尼人、拉巴-阿萨姆人、那加人)。对同民族不同编号的人群分别进行抽样。通过聚合酶链反应(polymerase chain reaction，PCR)——限制性片段长度多态性方法进行基因分型。使用错配引物将限制性酶切位点设计为 M130(Bsl I)和 M89(Nla III)。引物序列是 ACAGAAGGATGCTGCTCAGCTT/GCAACTCAGGCAAAGTGAGACAT(M89)和 TATCTCCTCTTCTATTGCAG/CCACAAGGGGGAAAAAACAC(M130)。YAP 的分型参照以前的报道[5, 9]。反复基因分型以澄清模棱两可的分型结果。

作为单基因座的多位点(单倍型)系统，Y 染色体是追寻人类进化历史最强大的分子工具[5,9,17-19]。以往 Y 染色体的研究表明，全球人群分布极具地域性特征，最古老的分支出现于非洲人群，年轻的分支出现于部分非洲人群和所有非洲以外人群[19]。对全球 1 062 个具有代表性的男性样本的研究表明，Y 染色体上的 M168 突变型(C→T 突变)为非洲之外的人群所共有，且该突变源于非洲[19]。M168 突变的年龄估计在 4.4 万年(95%置信区间为：3.5 万～8.9 万年)，记录了晚近源自非洲的迁徙[19]。在 M168T 的谱系下主要有 3 个亚型：定义位点为 YAP(*Alu* 序列插入)[5]、M89(C→T 突变)和 M130(C→T 突变，也称 RPS4Y)(图 2-1)[19,20]。因此这 3 个突变类型可以用来测试东亚远古人群被非洲起源的现代人替代的彻底性。如果发现没有携带这 3 种突变类型中任何一种类

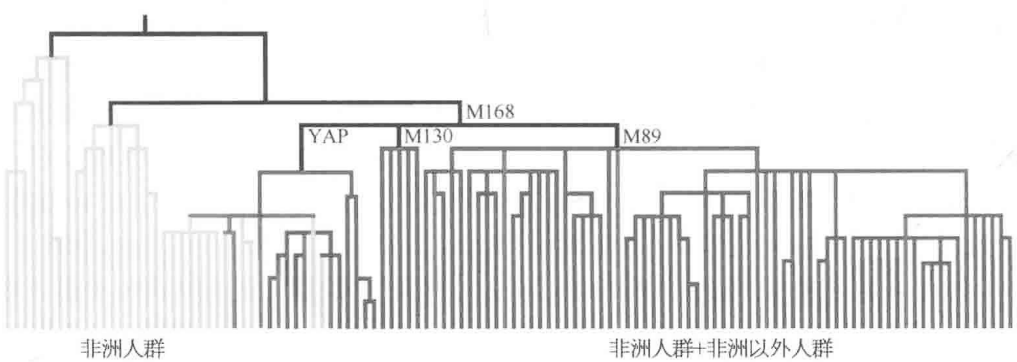


图 2-1　非洲和世界其他人群 Y 染色体单倍型的亲缘关系(据 Underhill 等修改[19])

红色分支是非洲特有的单倍型，蓝色分支是非洲以外特有的单倍型，绿色是非洲人群和非洲以外人群共享的单倍型。

型的男性个体，则表明潜藏着远古人类来源的成分，进而很有可能否定替代的彻底性。

12 127 份样品无一例外都具有这 3 个多态型（YAP +、M89 和 M130）中的一个（表 2-1）。换句话说，他们都属于来自非洲的 M186T 谱系。在现存的东亚人群中没有发现非洲以外的远古 Y 染色体（$P = 5.4 \times 10^{-6}$，假定对现存种群的贡献频率为 1/1 000），因此这意味着不存在非洲之外的独立起源和历时 100 万年之久的全球趋同演化。表明起源于非洲的现代人取代了东亚更早期的人群。

表 2-1 3 个 Y 染色体多态型在亚洲 163 个人群中的频率分布

地　区	群体数	个体数	M89T	M130T	YAP +
中亚	13	173	144	25	4
西伯利亚中部	5	107	70	36	1
鄂霍次克/阿穆尔	6	123	46	77	0
堪察加/楚科奇	4	102	73	29	0
东北亚	17	578	497	42	39
中国北部	14	4 592	4 296	191	105
中国南部	13	5 127	4 984	97	44
中国台湾少数民族	5	58	58	0	0
东南亚	37	620	559	37	24
印度尼西亚/马来西亚	25	355	333	22	0
波利尼西亚/密克罗尼西亚	11	113	89	23	1
新几内亚/美拉尼西亚	7	120	105	17	0
印度东北部	6	59	57	2	0
总　计	163	12 127			

有人认为，支持非洲起源说的大量遗传数据，也可以解释为附带横向杂交模型的多地区起源说[21]。这个模型认为自从 100 多万年前直立人走出非洲，多地区起源的范式在人类到达的各个地区发生，而各大洲之间也有着频繁的基因交流[21]。我们很难分别用线粒体高变区（D-loop 区）和常染色体的标记对横向杂交模型进行测试，因为这些标记分别有着频繁的回复突变或重组[22, 23]。然而，这些问题可以通过大量的 Y 染色体双等位基因标记来规避，因为 Y 染色体不重组，而且这些标记突变率非常低。由此证实，在非洲以外发现的所有 Y 染色体单倍型都来自非洲，并且年龄在 3.5 万～8.9 万年，甚至小于 3.5 万年[19]。此外，如果像多地区起源说者所认为的那样，在过去 100 万年间各大洲人群间发生了广泛的基因交流，而后非洲人群与非洲以外人群产生分化，那么，在非洲人群中观察到的古老 Y 染色体单倍型或者更为古老的单倍型就有可能出现在东亚，但是这在研究结果的数据中没有观察到。但是这一现象不能够排除另一种可能性，即自然选择可

能消除了东亚现代人中的远古 Y 染色体。另一方面，少量来自本土女性的贡献不能被完全排除，可以使用线粒体 DNA 标记进行深入研究。因为 Y 染色体具有相对小的有效群体，远古谱系的消失可能是随机过程引发的，如遗传漂变。然而，笔者的研究包括 163 个来自东亚不同地区的人群，很难想象这 163 个群体会都向同一方向漂变。

使用线粒体 DNA/Y 染色体和常染色体/X 染色体标记，对共同祖先的年代估计不一致，造成了一定的混乱。使用常染色体/X 染色体上的基因对年代进行估计，为 53.5 万～186 万年[24-27]，这远比使用线粒体 DNA 和 Y 染色体更为久远。然而，年代估计的差异可能仅仅反映了 Y 染色体/线粒体 DNA 和 X 染色体/常染色体的有效群体大小存在差异(后者是前者的 3～4 倍)，使得在走出非洲时并不一定伴随发生瓶颈效应，因此后者无法用来鉴定各种假说的真实性[22,28]。

参考文献

[1] Cann R L, Stoneking M, Wilson A C. Mitochondrial DNA and human evolution. Nature, 1987, 325: 31 - 36.

[2] Vigilant L, Stoneking M, Harpending H, et al. African populations and the evolution of human mitochondrial DNA. Science, 1991, 253: 1503 - 1507.

[3] Stringer C B, Andrew P. Genetic and fossil evidence for the origin of modern humans. Science, 1988, 239: 1263 - 1268.

[4] Bowcock A M, Ruiz-Linares A, Tomfohrde J, et al. High resolution of human evolutionary trees with polymorphic microsatellites. Nature, 1994, 368: 455 - 457.

[5] Hammer M F. A recent common ancestry for human Y chromosomes. Nature, 1995, 378: 376 - 378.

[6] Tishkoff S A, Dietzsch E, Speed W, et al. Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. Science, 1996, 271: 1380 - 1387.

[7] Chu J Y, Huang W, Kuang S Q, et al. Genetic relationship of populations in China. Proc Natl Acad Sci USA, 1998, 95: 11763 - 11768.

[8] Quintana-Murci L, Semino O, Bandelt H J, et al. Genetic evidence of an early exit of Homo sapiens sapiens from Africa through Eastern Africa. Nature Genet, 1999, 23: 437 - 441.

[9] Su B, Xiao J, Underhill P, et al. Y chromosome evidence for a northward migration of modern humans into Eastern Asia during the Last Ice Age. Am J Hum Genet, 1999, 65: 1718 - 1724.

[10] Krings M, Stone A, Schmitz R W, et al. Neandertal DNA sequences and the origin of modern humans. Cell, 1997, 90: 19 - 30.

[11] Ovchinnikov I V, Götherström A, Romanova G P, et al. Molecular analysis of Neanderthal DNA from the northern Caucasus. Nature, 2000, 404: 490 - 493.

[12] Brooks A S, Wood B. Palaeoanthropology: The Chinese side-of the story. Nature, 1990, 344: 288 - 289.

[13] Li T, Etler D A. New Middle Pleistocene hominid crania from Yunxian in China. Nature, 1992, 357: 404 - 407.

[14] Wu X Z, Poirier F E. Human evolution in China. Oxford: University Press, 1995.

[15] Etler D A. The fossil evidence for human evolution in Asia. Annu Rev Anthropol, 1996, 25: 275 - 301.

[16] Swisher C C, Rink W J, Antón S C, et al. Latest Homo erectus of Java: potential contemporaneity with Homo sapiens in Southeast Asia. Science, 1996, 274: 1870 - 1874.

[17] Jobling M A, Tyler-Smith C. Fathers and sons: the Y chromosome and human evolution. Trends Genet, 1995, 11: 449 - 456.

[18] Underhill P A, Passarino G, Lin A A, et al. The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. Ann Hum Genet, 2001, 65: 43 - 62.

[19] Underhill P A, Shen P, Lin A A, et al. Y chromosome sequence variation and the history of human populations. Nat Genet, 2000, 26: 358 - 361.

[20] Bergen A W, Wang C Y, Tsai J, et al. An Asian-native American paternal lineage identified by RPS4Y resequencing and by microsatellite haplotyping. Ann Hum Genet, 1999, 63: 63 - 80.

[21] Wolpoff M H, Hawks J, Caspari R. Multiregional, not multiple origins. Am J Phys Anthropol, 2000, 112: 129 - 136.

[22] Jin L, Su B. Natives or immigrants: modern human origin in East Asia. Nature Rev Genet, 2000, 1: 126 - 133.

[23] Stoneking M, Soodyall H. Human evolution and the mitochondrial genome. Curr Opin Genet Dev, 1999, 6: 731 - 736.

[24] Harding R M, Fullerton S M, Griffiths R C, et al. Archaic African and Asian lineages in the genetic ancestry of modern humans. Am J Hum Genet, 1997, 60: 772 - 789.

[25] Kaessmann H, Heissig F, Haeseler A, et al. DNA sequence variation in a non-coding region of low recombination on the human X chromosome. Nature Genet, 1999, 22: 78 - 81.

[26] Harris E, Hey J. X chromosome evidence for ancient human histories. Proc Natl Acad Sci USA, 1999, 96: 3320 - 3324.

[27] Zhao Z, Jin L, Fu Y X, et al. Worldwide DNA sequence variation in a 10-kilobase noncoding region on human chromosome 22. Proc Natl Acad Sci USA, 2000, 97: 11354 - 11358.

[28] Fay J C, Wu C I. A human population bottleneck can account for the discordance between patterns of mitochondrial versus nuclear DNA variation. Mol Biol Evol, 1999, 16: 1003 - 1005.

2.2 用遗传学数据重构人类进化谱系

近来基于基因组学的遗传学研究成果颠覆了以往的古生物学和生物分类法，甚至动摇了传统的人类阶段进化论。笔者将根据最新的遗传学研究成果，从猿类到现代人种来逐步重构人类的进化历程。

2.2.1 人科与猩猩科合并

长期以来，人们认为人这个物种是如此与众不同，应该脱离于动物界，是一个全新的类群。然而，随着系统生物学和进化生物学的建立，生物学家认识到人依然属于灵长类

动物的范畴，与其他猿类有很近的遗传关系。在灵长类中，没有尾巴的物种称为猿。现存猿类有两大类：小猿和大猿。小猿是各种长臂猿，一般单列为一个科，是没有争议的。对于大猿，传统做法是分为猩猩科和人科，猩猩科包括红猩猩、大猩猩和黑猩猩三属，而人科只有人一个属。但是很多进化学家怀疑，把人科从猩猩科划出来完全是人们一厢情愿的做法。近年来不断完善的灵长类基因组学的研究，使我们更深入地认识了猿类的系统发生关系，也确定人并不是一种另类。

因为形态特征的模糊性，传统的形态分类有先天缺陷，不同的进化路线上可能出现类似的形态。而基因组的差异则是明确而且可以量化的，显然是一种更好的进化学研究材料。两个物种之间的基因组差异程度，与它们之间的分化历史长度成正比。所以，通过与地质年代的校正，基因组差异可以转化为分化时间。一般来说，动物界中在大约 1 000 万年以内演化形成的各个物种可以划在一个科内。人与黑猩猩的基因组只有不到 2%的差异，分化历史也不到 600 万年，显然不可能分属两个科。所以，人科与猩猩科就合并了(图 2－2)。目前国际上普遍采用的科名是人科(Hominidae)，其下再分猩猩亚科(红猩猩)和人亚科(大猩猩、黑猩猩、现代人)[1]。但是红猩猩和其他猩猩的分化年代远超过 1 000 万年，所以或许也可以单列一个科。东亚地区发现的早期人科物种，包括腊玛古猿(西瓦古猿)、禄丰古猿、巨猿等，都属于红猩猩类群，而不是人亚科的成员。目前所知的人亚科早期成员是近 1 000 万年前非洲肯尼亚的纳卡里猿，其形态与大猩猩很接近。

图 2－2　类人猿的遗传谱系

在人亚科中，分出了大猩猩族和人族。很多被冠以“人”的物种，其实都包含在人族之中(图 2-3)。根据目前的古生物学发现，最早的人族物种是发现于非洲中部的沙赫人，距今大约 700 万年。这显然已经早于人与黑猩猩的分化年代，所以黑猩猩自然在人族之内，而且从形态上已经比沙赫人更为进化，有更大的脑容量。既然沙赫人都已被称为“人”，或许黑猩猩也应该被证明，不能再称为“猩猩”，至少叫作“黑猿”。实际上中国古代所称的猩猩仅指红猩猩，故又谓其颜色有猩红。

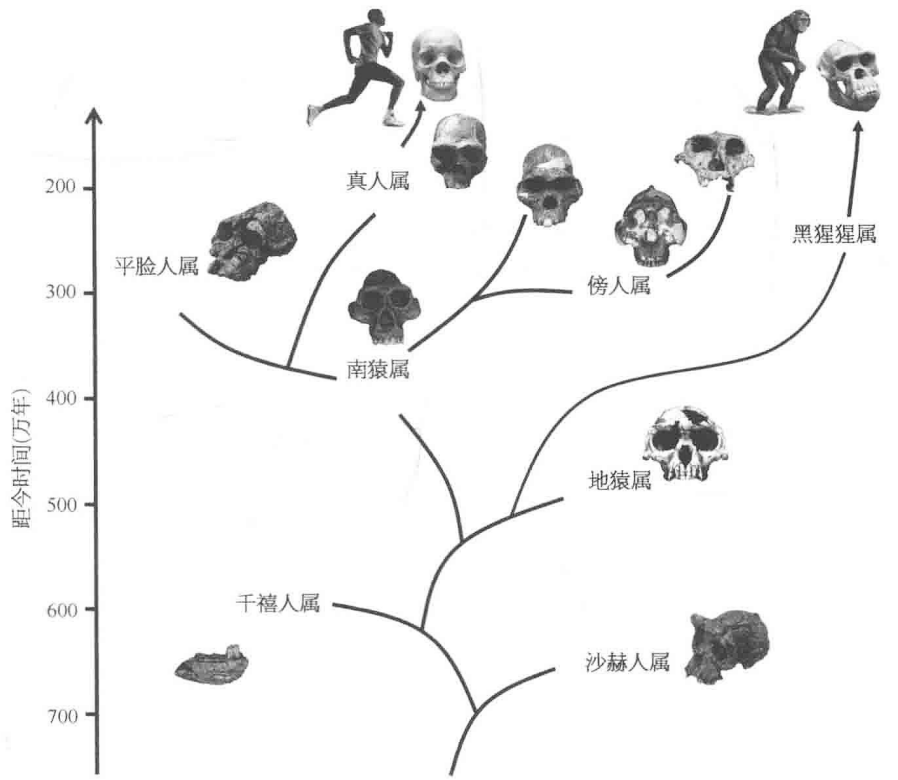


图 2-3 人族各属的系统树

2.2.2 人族分出 8 个属

人族的第二类物种是 2000 年发现于肯尼亚的千禧人，距今有约 600 万年。千禧人的形态与黑猩猩非常接近，而其大腿骨的形态甚至比晚 300 万年的南猿更接近人(真人属)。或许南猿并非我们的直系祖先，人有可能从千禧人直接演化而来。不过由于超过 5 万年的化石几乎无法分析 DNA，所以遗传学在人族演化研究中作用有限。而千禧人的化石也非常少，无法据此做出明确的判断。

地猿发现于埃塞俄比亚，距今约 500 万年。这一类群的形态与黑猩猩更为接近，非常有可能是黑猩猩的祖先。但是它们的牙齿像南猿的牙，所以还是难以判断其属于黑猩

猩还是人的分支。约 400 万年前，南猿出现了，发展成了人族物种中一个兴盛的类群，目前发现的依次有湖畔南猿、阿法南猿、羚羊河南猿、非洲南猿、惊奇南猿、源泉南猿，延续了大约 200 万年。肯尼亚平脸人能否成为一个独立的属，目前还有争议。从南猿演化出了两个进化策略截然相反的类群：傍人和真人。傍人非常粗壮，头顶有发达的矢状嵴，即发达的头部肌肉，后部臼齿是现代人的 2 倍，但是颅腔很小。所以傍人有发达的咀嚼能力，属于四肢发达、头脑简单的类型，很像一种猛兽。但最新研究认为，傍人主要食草。与傍人相反，真人则脑容量不断增大，四肢和牙齿趋向于纤弱。发达的头脑最终使得真人在进化中胜出，繁衍至今。

最有意思的是，距今两三百万年前的非洲，曾经同时生活着好几种人的近亲，有南猿、傍人、真人中的能人和卢道夫人，所以人曾经并不孤单[2]。

2.2.3　真人属的谱系

传统意义上称的人，实际上是狭义的人概念，也就是生物分类学上真人属的各个物种(图 2－4)。真人属起源于 200 多万年前。目前找到的最早的真人化石是非洲东部约

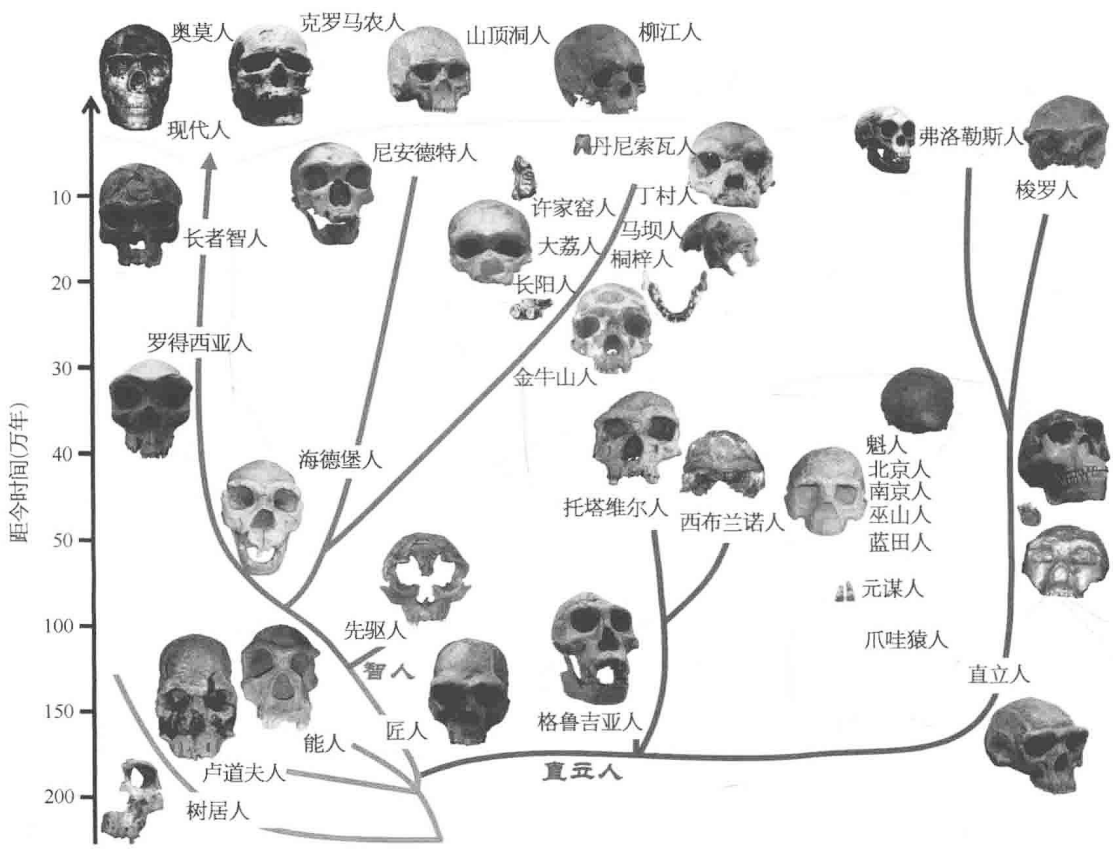


图 2－4　真人属内部的谱系结构

230万年前的能人，这一人种可能延续到了大约140万年前。但是2010年在南非豪登发现的树居人，在形态上比能人更原始，可能是更早出现的人。不过目前找到的树居人化石的时间段是距今60万～190万年，不排除今后还能发现更早的化石。卢道夫人可能是能人的一个分支，发现于肯尼亚，距今大约190万年。

前期的人除了上述3种以外，在130万～180万年前的非洲东部和非洲南部，还演化出了另一种人，即匠人。匠人从脑容量方面看，可能拥有比能人更高的智力，在工具制作方面也比能人更先进。与能人分化以后，匠人成为我们现代人最有可能的直系祖先。由于前期人化石的年代久远，无法进行DNA分析，而四个物种并没有都留下后代供遗传分析，所以分子遗传学对于前期人的谱系分析无法提供帮助。很有可能树居人与能人在200万年前已经分化，而在190万年前卢道夫人和匠人从能人分化出来。

后期的人传统上分为3大类：猿人（直立人）、古人（早期智人）和新人（晚期智人），曾经被认为是人类发展的3个阶段。现在，阶段论早已被古人类学和遗传学的研究结果所抛弃。首先，从古人类学的化石发现看来，直立人走出非洲，从西亚到东亚的扩张早至180万年前。而分子遗传学对现存的各个大洲的现代人分支进行了分析，无论是全基因组分析，还是线粒体DNA分析和Y染色体谱系分析都得到了一致结果，发现所有现代人都是20万年以内重新起源于非洲的。所以现代人不可能是亚洲直立人的后代，直立人和智人是两个不同的分支，而不是两个阶段[3]。

从匠人演化出的直立人分支上，还可能分化出了数个近缘分支，包括法国的托塔维尔人、意大利的西布兰诺人和格鲁吉亚的格鲁吉亚人。180万年前的格鲁吉亚人是迄今发现在非洲之外的最早人类化石。这几种人也往往被认为是直立人的亚种。直立人的标准种是印度尼西亚的爪哇人，50万年前东亚和东南亚的人都属于直立人的各个亚种，其中最著名的有北京猿人、蓝田猿人、元谋猿人等。不过，元谋猿人的化石仅有两颗牙。虽然直立人在东亚和东南亚广泛分布，但种群可能非常小，很多分布点持续时间很短，这些种群已陆续灭亡，其中印度尼西亚爪哇岛的梭罗人一直生存到了14万年前。直立人中最奇特的是印度尼西亚东部弗洛勒斯岛发现的弗洛勒斯人。这个人类物种生存于1.3万～9.4万年前，身材极其矮小，小于110 cm。这是迄今发现的最矮小的人类，可能是因为数万年生存于狭小的海岛，应对贫乏的资源而产生的适应。由于特殊的形态，弗洛勒斯人一般被认为是已经区别于直立人的独立物种[4]。

2.2.4 智人的3个分支

智人的谱系研究最近有了重大进展。成功获得尼安德特人[5]和丹尼索瓦人[6]的全基因组数据可能是近十年内人类进化研究中最重大的成果。欧亚大陆西部的尼安德特人生活到距今大约3万年前，欧亚大陆东部的丹尼索瓦人生活到距今大约4万年前。通过比较尼安德特人、丹尼索瓦人和现代人的全基因组差异，三者之间的演化谱系结构展示得清晰无遗。尼安德特人和丹尼索瓦人之间有大约60万年的分化，而他们与现代人都有大约80万年的分化。所以这3个类型应该代表着智人的3个主要分支。现代人都

是 20 万年以内走出非洲的，其直系祖先可能是非洲早期智人——罗得西亚人。尼安德特人广泛分布于欧洲和西亚，甚至散布到中亚。丹尼索瓦人虽然发现于阿尔泰山区，但是可能代表着整个东亚和东南亚地区的早期智人。所以，早期智人和晚期智人的名称意义并不确切，更好的名称可以是南方智人、北方智人和东方智人(图 2－5)。

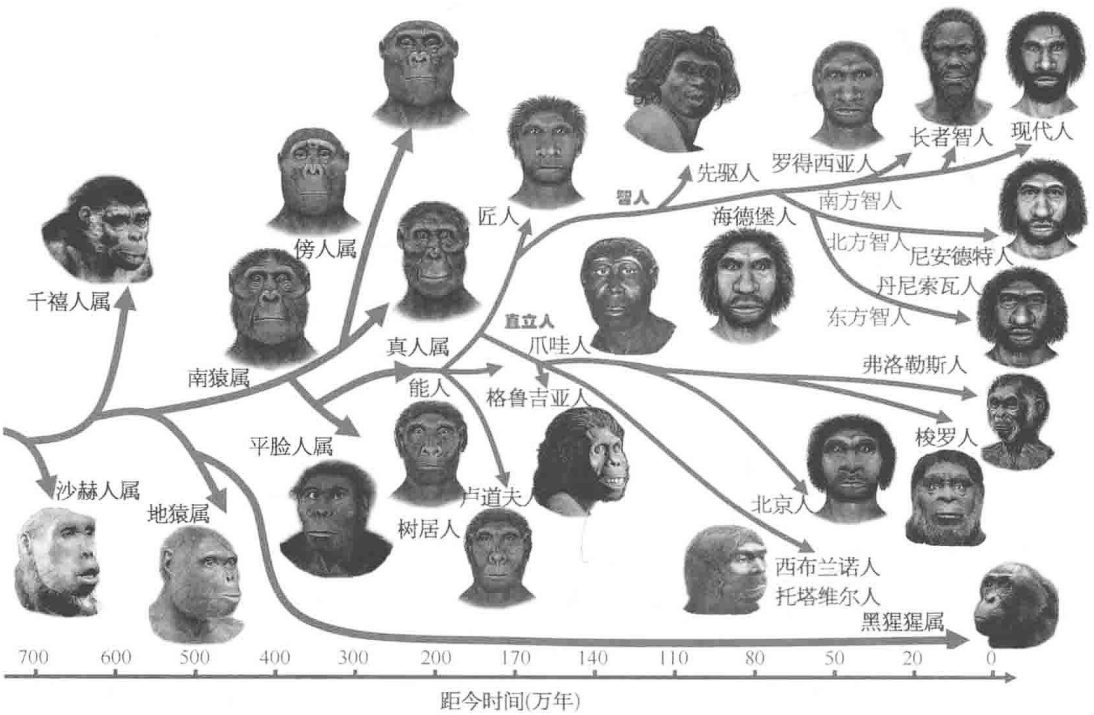


图 2－5　复原像展示人族各属以及真人属各物种的分化历程

不过，母系线粒体 DNA 的谱系分析得出稍有不同的三者间拓扑结构。现代人与尼安德特人分开 40 多万年，两者与丹尼索瓦人分开大约 100 万年[7]。纯母系的结构与全基因组结构的差异，可能暗示着人类迁徙中的复杂故事，一个人群接受其他人群的女性可能是比较容易的。智人分化的年代，与猩猩、大猩猩、黑猩猩 3 个属内各两个物种的分化年代基本一致，原因可能是当时全球气候发生了剧变。

智人的起源时间估计在 120 万年前。迄今发现的最早的欧洲人——西班牙阿塔坡卡发现的先驱人就是那个年代的。先驱人已经具有了很多智人的特征。但由于先驱人只是在西班牙昙花一现，可能不久就灭绝了，成为了人类进化中的旁支，并没有留下后代。最早明确属于智人的人物种是海德堡人。这一类群主要发现于欧洲，生存年代在 40 万～60 万年前。海德堡人的脑容量与现代人基本相当，可能是因为他们身材巨大。欧洲海德堡人的平均身高达到 180 cm。有些学者认为非洲同时期的人类也属于海德堡人，比如南非发现的“巨人”，是人物种中最高大的，达 213 cm。海德堡人可能有了语言，已经开始埋葬死者，很可能是 3 种智人分化之初的阶段，属于尚未形成形态

差异的时期。

对于智人3个分支之间可能发生过的遗传交流，也就是尼安德特人和丹尼索瓦人有没有遗传成分传到现存的现代人中，是人类进化研究中最引人入胜的课题。在尼安德特人和丹尼索瓦人的基因组数据出来之前，对于3种智人之间的遗传交流只能局限于猜想。现在，通过比较3种基因组，我们已经达成比较精确的认识。在2010年之前，通过纯父系的Y染色体和纯母系的线粒体DNA分析，认为在现代人中没有任何尼安德特人或者丹尼索瓦人的成分。但是最近的全基因组分析得到了稍有不同的结果。非洲现代人中，依旧没有发现任何尼安德特人或丹尼索瓦人的遗传成分。但是在非洲之外的现代人群中，都发现有1%～4%的尼安德特人基因组成分。而且，这些基因交流是在大约7万年前现代人刚刚走出非洲的时候发生的，其后就再也没有发生过。现代人与尼安德特人在欧洲共存了数万年，所以走出非洲以后分化形成的世界各地的人群中都保存了相同的尼安德特人基因比例。

丹尼索瓦人虽然发现于北亚地区，但是在亚洲大陆上的现代人群中没有发现任何丹尼索瓦人的遗传成分。反而，在大洋洲的新几内亚原住民人群中发现了大约6%的遗传比例[8]。很有可能是新几内亚原住民的祖先在迁徙途经中南半岛时接触到了丹尼索瓦人群体，发生了基因交流。所以可以确定，丹尼索瓦人的地理分布很广泛，至少从北亚到东南亚都存在，而且人口不少，有机会把可观的遗传基因流传到新几内亚现代人中。丹尼索瓦人生活的时期，与“东亚早期智人”的生活时期大致重合，可以推断所谓“东亚早期智人”与“丹尼索瓦人”就是同一个物种。

东亚现代人为何没有与丹尼索瓦人发生基因交流，这是一个不容易解释的事实。研究者曾经期待早期的东亚现代人会有更多的尼安德特人或者丹尼索瓦人遗传成分。但是，2013年新发布的北京周口店地区4万多年前的田园洞人基因组，却与现代的中国人几乎没有差别，没有更多“早期智人”的遗传成分[9]。看来，3种智人之间的基因交流可能发生过，但是非常有限。

2.2.5 现代人的8个人种

非洲的南方智人在至少16万年前开始发生明显的形态变化，在埃塞俄比亚演化出了长者智人，其形态介于罗得西亚人和现代人之间。但在埃塞俄比亚还发现了几近20万年前的奥莫现代人，其形态特征已经基本属于现代人，而长者智人却还有更多类似罗得西亚人的特征。所以现代人至少20万年前就起源了。如果长者智人是罗得西亚人与现代人之间的过渡类型，说明长者智人可能在比20万年更早时间就形成了，只是有些群体并没有演化成现代人的形态，一直保留到16万年前。但是这些最早的群体并不能全部生存下来，并不能把所有的基因库都流传到现代。因此，从不同遗传方式的基因组区段，可以把现代人的谱系追溯到不同的年代。纯母系的线粒体DNA谱系可以最远追溯到大约20万年前，而纯父系的Y染色体只能追溯到14.2万年前。这说明女性有更公平的生育权，也更容易被其他群体接受。所以14.2万～20万年之间的很多女性都留下了

直系后代至今，而期间的父系只有一个最终留下直系后代至今。

但是，一项新的发现把 Y 染色体的谱系推到 33 万年前[10]。在西非喀麦隆的西北部山区找到了一个亩巴人(Mbo)村子，他们的 Y 染色体与世界其他人群的序列差异极大，可能已经分化了 33 万年。所以研究者认为这可能是罗得西亚人残留的 Y 染色体，定义为单倍群 A00(图 2-6)。而该群体距离最后的罗得西亚人遗址，即 13 000 年前尼日利亚的遗娃来如(Iwo Eleru)遗址不远。研究者认为现代人发生以后，在扩张过程中不断与残存的罗得西亚人群体混血，形成很多混合群体[11]。这样说来，长者智人更可能是混合群体，而不是进化中间步骤。这也更符合现代人-长者智人的先后关系。不过由于该项研究使用的突变率等参数可能并不适用，过高估计了分化年代，重新用 Y 染色体家系突变率估算的结果是 A00 与其他类群的分化年代大约为 20.9 万年，那样，A00 还可能是现代人最早分化出的谱系，Y 染色体与线粒体 DNA 追溯到了同样的年代。

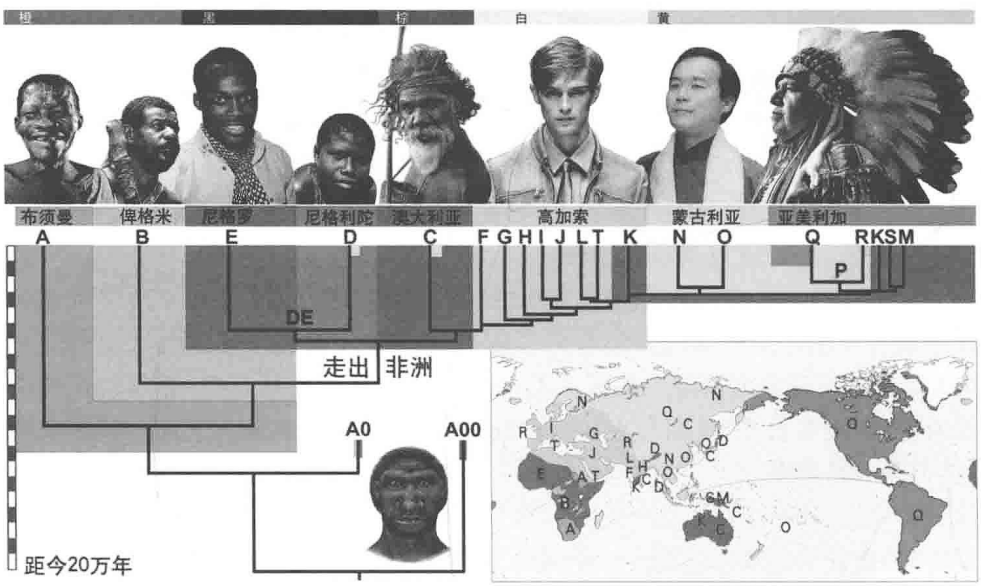


图 2-6　全世界人群的 Y 染色体谱系树

由于男性对族群的主导性，父系的遗传类型(Y 染色体类群)容易变少。所以不同群体之间差异最大的遗传物质是 Y 染色体类群，也叫 Y 染色体单倍群。全世界的 Y 染色体单倍群构成了一个可靠的谱系(图 2-6)。Y 染色体的主要单倍群的形成需要长期的隔离演化，这与现代人种族的隔离演化机制是一致的。所以现代人发展早期，Y 单倍群与人种应该有过很好的对应关系。不过由于近几千年来人群的大规模融合，这种对应关系稍有打乱。

Y 染色体的根部类群是 A 型，仅存在于非洲。其次是 B 型，也在非洲。所以从 Y 染色体来看，现代人肯定起源于非洲。C 以后的类群(CT)从 B 分化出来的年代大约是 7 万

年，所以现代人走出非洲的年代不会早于7万年。A、B、C、D、E这五种类群，每一类群内部的亚型都是大约6万年前开始分化形成的。这一时段就是现代人最早的种族形成时期。在距今7万多年前，地球上发生了一次巨大的灾难，苏门答腊岛上的多峇火山发生了超级大爆发，史称多峇巨灾。此后地球进入了冰期，许多动物种群灭亡，人类群体也大量灭亡。留下的少许小群体隔离分布在非洲中部到东北部，形成了数个种族。其后由于冰期的海平面下降，大陆之间出现了很多新的陆地连接，人类群体开始向各大洲迁徙，种族进一步演化。

1863年，德国生物学家海克尔(Ernst Haeckel)绘制了一张人类种族起源图谱(图2-7)。在这张图谱中，全世界的人分成12个人种。现在，我们对全球的人群有了全面的普查，所以发现海克尔遗漏了两个矮人人种——非洲的俾格米人与亚洲的尼格利陀人。对各人种的遗传基因的分析也发现，海格尔列出的某些人种其实是其他人种的混合群，比如奴比人种和卡佛人种是黑人种与侯腾图人种的不同混合群，德拉威达人种是地中海人种与大洋洲人种的混合，马来人种是蒙古人种与尼格利陀人种的混合。而美洲人种与北极人种的差异，以及大洋洲人种与巴标人种的差异，其实并不大。

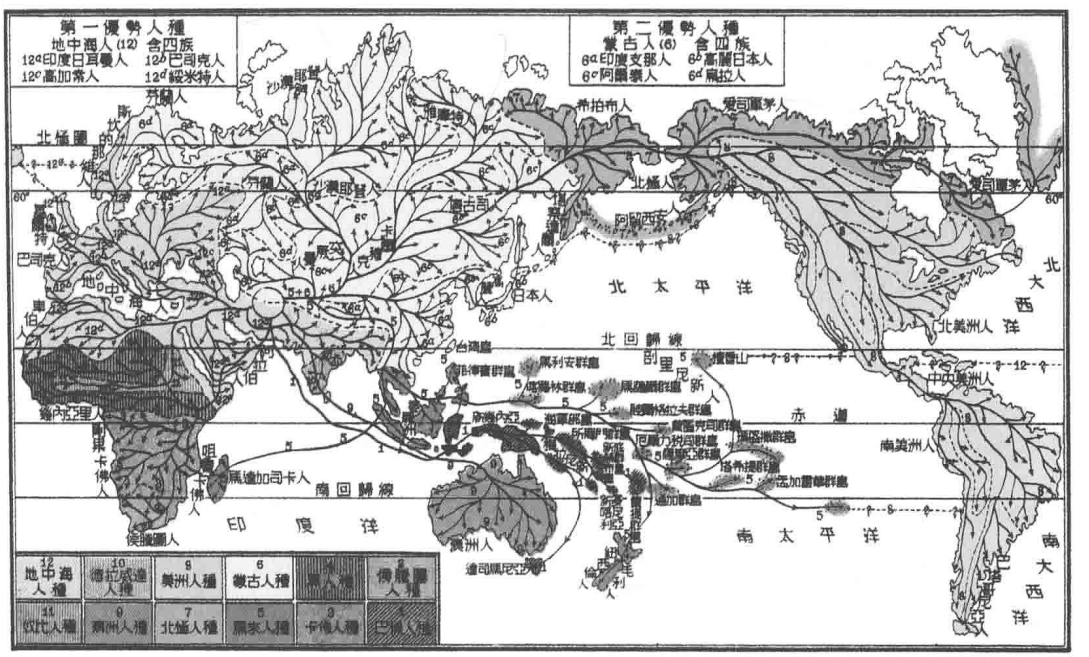


图2-7 海克尔在《自然创造史》中绘制的人类种族起源图谱(马君武译，商务印书馆1936年版)

全世界的人群一共有5种肤色：橙色、黑色、棕色、白色、黄色。从全基因组的分析[12]来看，全世界的人群可以分成8个人种：布须曼、俾格米、尼格罗、尼格利陀、澳大利亚、高加索、蒙古利亚、亚美利加(图2-8)。按照体质形态特征，全世界的现代人也可以分为上述8个人种。近年来，由于政治上反种族主义的需要，西方遗传学界提出

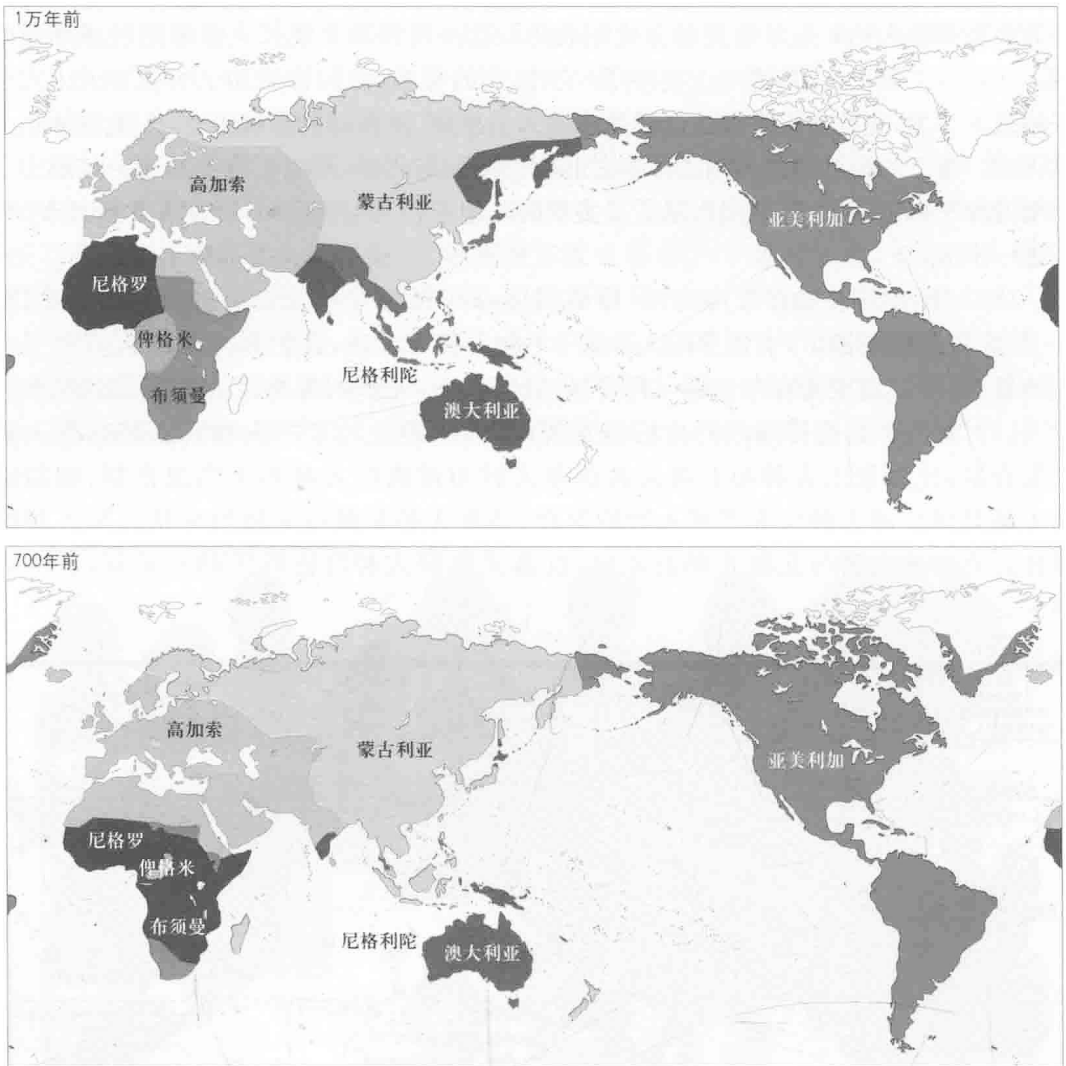


图 2－8　现代人八个人种的历史地理分布示意图

特别的观点，认为种族的概念是没有遗传学根据的。其证据主要是种族之间都存在过渡类型，没有绝对的界线；大多数基因等位型在各个种族内都有一定的频率分布。实际上，种族主义的错误在于认为种族有高低贵贱之分，这导致了人类历史上的多次种族灭绝惨剧。反对种族主义，是要反对种族歧视，反对种族在先天上有优劣之分，而不是否认种族在外形和遗传历史上的客观差异。如果说黑人与白人在生物学上没有差异，这显然不符合客观事实。西方遗传学界提出的种族之间有过渡，其实是近几千年来人群的混合造成的。例如，在加勒比群岛上，还存在美洲印第安人与黑人之间的过渡类型，显然是人群混合形成的，而不是美洲人从非洲渐变而来的过渡类型。等位基因类型在种族之间也大多没有明显差异，毕竟现代人与黑猩猩的基因组也只有 2%以

下的差异。所以种族的基因组之间，只要有少数基因有特异性分布，就足以支持种族的生物学存在。

2.2.6 Y染色体谱系与人种的同步演化

与现代人各个种族对应关系最好的遗传材料是Y染色体的谱系。体质类型的分化与Y染色体单倍群的分化都是在群体地理隔离的条件下同步发生的(图2-8)。根据Y染色体的谱系分析，最古老的类型是A型，集中分布于非洲南部和东北部，也零星分布于中非。相关的人种是非洲南部的布须曼人(旧称开普人种或侯腾图人种)，非洲东北部的尼罗-撒哈拉人(奴比人种)也与之有关。A型下面的有些亚型只出现在埃塞俄比亚的一些群体中。最近的研究指出，A型可以追溯到非洲中部偏东北地区，非洲南部布须曼人的A型也是从北方而来。布须曼人科依桑语系的语音是世界语言中最为特别的，有着复杂的搭嘴音。包括尼罗-撒哈拉人在内的布须曼人种的肤色呈橙红色，而不是常见的非洲人的黝黑色。考古学和遗传学研究都发现，非洲的黑人只是最近一千年来从非洲西部扩张到非洲东部和南部，此前非洲大部分区域的居民都是橙色人种。在黑色人种和橙色人种的接触中，A型Y染色体也流入了非洲中南部的黑人中。

年龄其次的Y染色体类群是B型，大致对应中非、刚果等地热带雨林中的俾格米人。非洲东部坦桑尼亚的哈扎比人Y染色体也多为B型，他们的身高也同样偏矮。俾格米人种非常适应在热带雨林中生活，有些村落完全建造于雨林的树冠上。他们的肤色也偏橙色，不同于西非尼格罗人的黑色，所以也算是一种橙色人种。矮小的俾格米人与高大的尼格罗人在毛发上特征差异也很明显。成年俾格米男人有着浓密的胡须，而尼格罗人的胡须一般很稀疏。

两个橙色人种与其他人群的分化都在7万年以上。其他分支都是7万年之内走出非洲的人群的后代。其中D和E最早是黑人的类群，他们可能是六七万年前在埃塞俄比亚与也门所在的红海口处进入亚洲，而后在红海北部分离(图2-9)。携带E型Y染色体的人群回到非洲，一路向西，成为非洲西部的尼格罗大黑人；而携带D型Y染色体的人群辗转向东迁徙，成为东南亚的尼格利陀小黑人。两种黑人的分布区域相距如此遥远，这是非常不可思议的格局。而在身高上也达到两个极端。尼格罗人非常高大，非洲西部有些种群的成年男子往往超过180 cm，而尼格利陀人成年人一般不会超过150 cm，甚至

图2-9 8个人种及Y染色体根部单倍群的大致迁徙路线

更为矮小。尼格利陀人现在仅存于缅甸以南的安达曼群岛、泰国和马来西亚边境山区、菲律宾中北部山区。但是其对应的 D 型 Y 染色体广泛分布于青藏高原、日本列岛和中南半岛。所以这些区域很可能是尼格利陀人的历史分布区，不过后来在黄色或棕色人种的影响下发生了人群体质变化。很有意思的是，菲律宾的尼格利陀人中没有发现 D 型 Y 染色体，而有着来自新几内亚的棕色人种的 C 型和 K 型。这可能是棕色人种后期的扩张影响。而日本列岛最早的居民绳文人有 D 型 Y 染色体，身高也在 150 cm 以下，应该属于尼格利陀人种，但是面貌特征却是典型的澳大利亚棕色人种。所以，在迁徙路线的末端，人种之间交流的复杂程度远超我们的想象。

携带着 C 型和 F 型 Y 染色体的人群跨过红海以后，继续向东北进发，F 型来到了两河流域，C 型来到印度河流域。在这两个区域中，两个人群演化成了不同的人种。C 型人群形成了棕色人种，在五六万年前扩散到东亚、东南亚和澳大利亚、新几内亚、美拉尼西亚，也被称为澳大利亚人种。而 F 型人群则是白种人和黄种人的祖先。

在三四万年前 F 型开始从两河流域、里海南岸扩张，其下有 G 到 T 等 14 种亚型。G、H、I、J、L、T 在欧亚大陆西部成为高加索人种。高加索人种虽然往往被称为白人，但是肤色不一定很白。大约 2 万年前 O 型和 N 型人群来到东亚形成蒙古人种，取代棕色人种成为东亚的主体人群。大约 1.3 万年前，N 型人群从东亚扩张到北亚和北欧。也是在大约 2 万年前，Q 型和 R 型人群来到了中亚，但是他们并没有在当地形成独特的人种，而是大多融入了周边的人种。大多 Q 型人群向东迁徙加入蒙古人种，部分继续东迁，大约 1.5 万年前跨过白令海峡进入美洲，形成亚美利加人种。R 型是中亚地区的主要类群，但同时大量向西迁徙加入高加索人种，成为南欧人群的主流。

随着 Y 染色体谱系研究的深入，对 Y 染色体各个类群分化时间的分析越来越精确，人类群体演化的历史将越来越明确。客观准确地认识人类的演化历史，了解人种、民族和群体方方面面的异同，使我们更好地理解人群之间、人与自然之间应有的和谐关系，更好地维护人群的身体健康和社会健康。

参考文献

[1] Locke D P, Hillier L W, Warren W C, et al. Comparative and demographic analysis of orangutan genomes. Nature, 2011, 469 (7331): 529 - 533.

[2] Tattersall I. Once we were not alone. Scientific American, 2000, 282 (1): 56 - 62.

[3] Stringer C. Evolution: what makes a modern human. Nature, 2012, 485 (7396): 33 - 35.

[4] Brown P, Sutikna T, Morwood M J, et al. A new small-bodied hominin from the Late Pleistocene of Flores, Indonesia. Nature, 2004, 431 (7012): 1055 - 1061.

[5] Green R E, Krausel J, Briggs A W, et al. A draft sequence of the Neandertal genome. Science, 2010, 328 (5979): 710 - 722.

[6] Reich D, Green R E, Kircher M, et al. Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature, 2010, 468 (7327): 1053 - 1060.

[7] Krause J, Fu Qiaomei, Good J M, et al. The complete mitochondrial DNA genome of an unknown hominin from Southern Siberia. Nature, 2010, 464 (7290): 894 - 897.

[8] Reich D, Patterson N, Kircher M, et al. Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. Am J Hum Genet, 2011, 89 (4): 516 - 528.

[9] Fu Q, Meyer M, Gao X, et al. DNA analysis of an early modern human from Tianyuan Cave, China. PNAS, 2013, 110(6): 2223 - 2227.

[10] Mendez F L, Krahn T, Schrack B, et al. An African American paternal lineage adds an extremely ancient root to the human Y chromosome phylogenetic tree. Am J Hum Genet, 2013, 92(3): 454 - 459.

[11] Hammer M F. Human hybrids. Sci Am, 2013, 308(5): 66 - 71.

[12] The HUGO Pan-Asian SNP Consortium. Mapping human genetic diversity in Asia. Science, 2009, 326: 1541 - 1545.

2.3 末次冰期东亚人群由南到北的迁徙

2.3.1 研究背景

在本项研究之前，我们对现代人扩张进入欧洲、美洲和大洋洲的情况了解得相对清楚，而对现代人从西亚迁徙到东亚的最早路线知之甚少[1]。东亚是有着相对丰富的原始人类化石的少数地区之一，这些化石跨度有几十万年。这些化石证据说明了一种可能性：人属生物自从到达亚洲便可能在当地持续进化[2-6]，这种观点当然是与著名的现代人进化的“走出非洲”假说[7,8]对立的。然而并不是所有的古人类学家都同意亚洲的化石记录清晰显示了从直立人到现代人的连续性[9,10]，因此东亚现代人的亚洲当地起源假说不足以令人信服。

最近一项关于亚洲人群的研究，利用微卫星位点的分析结果，质疑了亚洲当地起源假说，并暗示东亚的现代人起源于非洲[11]。这项研究也证实了过去观察到的蒙古人种北方群体和南方群体之间的遗传和形态巨大差异[12,13]，并指出这种差异可能来自东亚人群向北方的迁徙[11]。然而，这项研究的缺点在于没有提供明确的证据来支持这一假说，原因是它所使用的微卫星标记有很高的突变率[11]。因此，一项基于信息丰富的稳定双等位基因标记的系统性研究，可以促进对史前东亚人群迁移的进一步认识。

近些年来，研究人员认识到了 Y 染色体标记在解决现代人迁徙样式问题上的效力[14]。一种强有力的检测突变的技术——变性高效液相色谱法的引入，使得高效地确认 Y 染色体上的双等位基因标记成为可能[15-18]。双等位基因标记是单碱基突变或小的插入/缺失，每个标记在人 Y 染色体进化中通常只出现一次，所以比起微卫星位点更加稳定。Y 染色体非重组区上的这种标记允许无歧义单倍型的重建，它不会被重组和重复突变所破坏，所以能够提供很多信息用于追踪古人类的迁移。鉴于 Y 染色体的有效群体大小比常染色体小，Y 染色体特异的多态标记很可能是研究早期人类迁徙最好的基因材料，因为它通常伴随着早期人类迁徙的瓶颈事件而变得更加显著了。

在本节的研究中，一系列 Y 染色体双等位基因和微卫星标记被用来检测东亚人群的遗传结构。现存东亚人群的 Y 染色体单倍型分布被用来重建这一区域内的古人类迁徙样式。

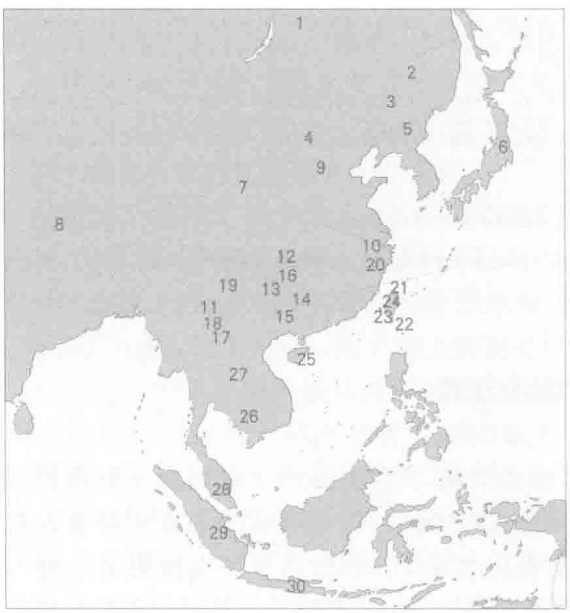


图 2－10　东亚 30 个人群的地理分布图

数字 1～30 与表 2－2 一致。中国汉族人群被分成了北方人群（人群 9）和南方人群（人群 10），这一分类以长江作为分界。

2.3.2　材料和方法

1. DNA 样本

在世界范围内收集了 925 个男性的 DNA 样本，其中 739 个来自东亚人群。在与中国人群基因组多样性计划的合作下，完成了对来自中国 21 个少数民族人群的 DNA 样本的收集。中国汉族样本来自生活在 22 个省的人群，他们的地理来源由他们的祖父母辈 4 个个体的出生地来确定。此外，对过去一些项目中获得的样本也做了分析，包括来自东北亚 3 个人群（布里亚特人、朝鲜人和日本人）和东南亚 5 个人群（柬埔寨人、泰国人、马来西亚人、巴达克人和爪哇人）以及来自亚洲以外的其他 12 个人群（3 个来自非洲、3 个来自美洲、2 个来自欧洲、4 个来自大洋洲）的样本（图 2－10，表 2－2）。

2. 单倍型分型和系统发育树的构建

一共扫描了 19 个 Y 染色体双等位基因位点。它们中有 7 个来自过去的报道[17, 19-21]，包括 M3（C 到 T 的突变）、M5（A 到 G 的突变）、M7（C 到 G 的突变）、M9（C 到 G 的突变）、M15（9 个碱基对的插入）、M17（一个碱基对的缺失）和 DYS287（YAP）。其他 12 个单核苷酸多态性在本研究中第一次报道，包括 M45（G 到 A 的突变）、M50（T 到 C 的突变）、M88（A 到 G 的突变）、M89（C 到 T 的突变）、M95（C 到 T 的突变）、M103（C 到 T 的突变）、M110（T 到 C 的突变）、M111（4 个碱基对的缺失）、M119（A 到 C 的突变）、M120（T 到 C 的突变）、M122（T 到 C 的突变）以及 M134（1 个碱基对的缺失）。本研究所采用的这 19 个标记是从 166 个 Y 染色体双等位基因标记[22]中选出的，依据是它们在东亚人群中具有多态性。一种等位基因特异的聚合酶链反应（PCR）被用来对 Y 染色体双等位基因标记进行基因分型。对于每个 Y 染色体位点，设计两个等位基因特异性配对的引物去识别这个位点上的两个不同的等位基因。PCR 之后，产物通过琼脂糖凝胶电泳观察。此外，3 个 Y 染色体微卫星位点（DYS389、DYS390 和 DYS391）也如 Kayser 等描述的那样[23]进行分类。基于简约原则绘制 Y 染色体单倍型的系统发育树，并且引入多重

表 2-2 东亚和世界人群 Y 染色体单倍型频率分布

区域与人群	频率(%)																
	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11	H12	H13	H14	H15	H16	H17
	ABC	DE	D1	F	K	O3 - M122	O3 - M7	O3 - M134	O1 - M119	O1 - M110	O2 - M95	O2 - M88	Q - M120	P	Q - M3	R - M17	M - M5
东亚																	
北方																	
1. 布里亚特(4)	75			25													
2. 埃文基(8)	50				12.5	12.5		25									
3. 中国东北(18)	16.7			11.1	22.2	27.8		16.7	5.6								
4. 蒙古(24)	58.3	4.2		8.3	12.5	4.2		4.2	4.2							4.2	
5. 朝鲜(7)					57.1			42.9									
6. 日本(29)	20.7	27.6			20.7	17.2		10.3	3.4								
7. 回族(20)	10	5		20	30			20						10		5	
8. 西藏(8)		12.5	25	12.5				50									
9. 北方汉族(82)	8.5			2.4	22.0	29.3		23.2	9.8				4.9				
南方																	
10. 南方汉族(280)	7.9	0.4		1.4	12.9	25.4	1.8	27.9	16.8		3.6	0.7	1.4				
11. 景颇族(5)								100									
12. 土家族(10)	10				20	30	10		20			10					
13. 南丹瑶族(布努)(10)	50				20		30										
14. 金秀瑶族(拉珈)(10)	20		30			10					40						
15. 壮族(28)	3.6		3.6	7.1	3.6	3.6		25	17.9		25	10.7					
16. 侗族(10)	20					10		20	20	10	20						
17. 布朗族(5)	20				20						60						

（续表）

区域与人群	频率(%)																
	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11	H12	H13	H14	H15	H16	H17
	ABC	DE	D1	F	K	O3 - M122	O3 - M7	O3 - M134	O1 - M119	O1 - M110	O2 - M95	O2 - M88	Q - M120	P	Q - M3	R - M17	M - M5
18. 拉祜族(5)	20			60							20						
19. 彝族(14)			14.3		42.9	21.4		7.1			14.3						
20. 畲族(11)	18.2				9.1	18.2	27.3	18.2			9.1						
21. 泰雅族(24)						29.2	4.2	4.2	54.2	8.3							
22. 雅美族(8)									25		75						
23. 排湾族(11)								18.2	54.6	27.3							
24. 阿美族(6)									100								
25. 黎族(11)								9.1	27.3		54.6	9.1					
26. 柬埔寨(26)	3.8		3.8	11.5	11.5	3.8		15.4	3.8	3.8	23.1	11.5		3.8		3.8	
27. 泰国东北(20)				5	5	5	5		5	5	45	20		5			
28. 马来西亚(13)				7.7	7.7	30.8		15.4	7.7	23.1	7.7						
29. 巴达克(18)	5.6			5.6	11.1	11.1	16.7		22.2		27.8						
30. 爪哇(11)	9.1			9.1	27.3	9.1			18.2	9.1	18.2						
亚洲以外																	
非洲人(24)	20.8	79.2															
美洲印第安人(26)														3.8	96.2		
欧洲人(39)	10.3			12.8	25.6									51.3			
大洋洲(100)	16			2	40									4			38

注：东亚人群之前的数字1～30与图2-10中表示这些人群的数字一致。17种单倍型下方标注了目前标准的Y染色体单倍群编号。

分支来适应同样简约的拓扑结构。

3. 统计方法

为了估计中国汉族人 M122C 单倍型的产生时间，使用公式 $t = -N_e \ln(1 - V/N_e m)$。假设群体大小($N_e$是有效群体大小)保持不变，$V$ 是群体中微卫星位点(STR)重复次数的方差，m 是突变率，由单步突变模型可以推导出这个公式。如果群体经历过强烈的瓶颈事件并随即快速扩张，可以证明这个公式依然近似合理。Takahata[23] 提出被广泛接受的现代人有效群体大小是 5 000～10 000。鉴于与非洲相比亚洲人的遗传多样性相对较小，2 000 这个值已经是对东亚男性有效群体大小的高估。考虑到这项研究中观察到的 160 个个体的方差水平，有效群体大小允许的最小整数是 750。

2.3.3 研究结果

在分析了所有 19 个 Y 染色体双等位基因标记的个体中，我们构建了 17 个 Y 染色体单倍型。所有群体中的 Y 单倍型频率分布见表 2-2。在简约假设下，不会观察到重复突变，笔者绘制了 17 个 Y 单倍型的系统发育树，其中 H1 被认为是祖先单倍型，原因是它出现在黑猩猩中(图 2-11)。H1 和 H2 单倍型相对古老，它们在非洲和非洲以外人群中都有出现，说明决定这种单倍型突变的出现早于现代人最初走出非洲。

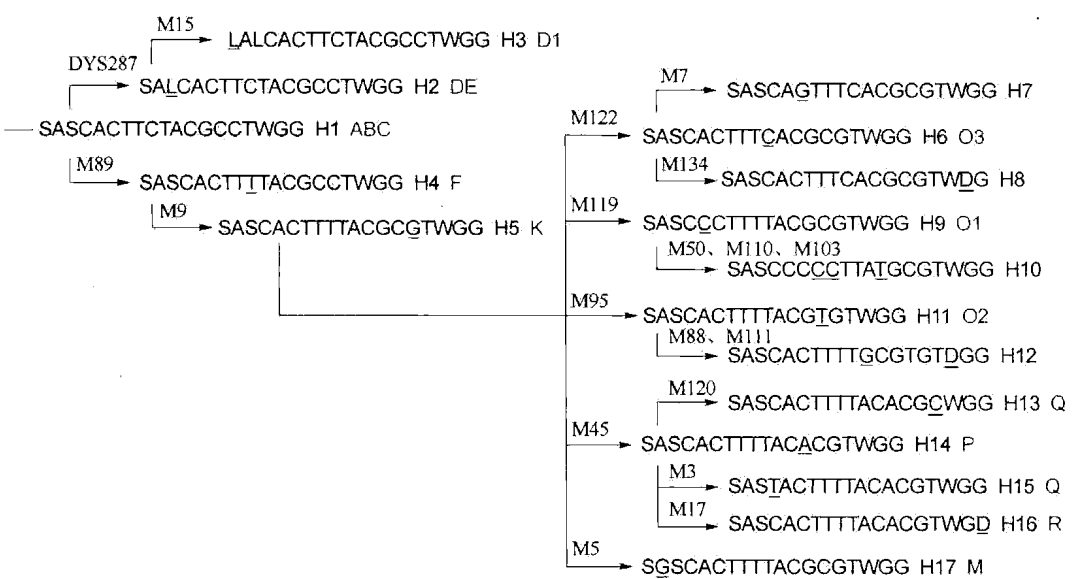


图 2-11 东亚和世界人群 17 个 Y 染色体单倍型的大部分简约树

字母 A、T、G、C 代表多态位点序列。字母 S、L 分别代表小或大的等位基因。字母 W、D 分别代表野生型和缺失等位基因。带下划线的字母表示该突变发生在分支以上的每一个位点(对于突变的详细描述，见材料和方法部分)。红色为国际 Y 染色体命名委员会的标准编号。

有趣的是，H5 看起来是所有分布在各个地区的其他单倍型共同祖先，并且它很可能出现在走出非洲之后。支持这一解读的事实是，H5 和它所有的衍生物都可以通过位点

M9 上的 C 到 G 的突变来辨别，而所有非洲单倍型在这一位点上都是 C[20]。H15 和 H17 分别是美洲印第安人和大洋洲特异的单倍型，这是之前报道过的[17,18]。H14 在欧洲人群中高频出现，但也出现在大洋洲、亚洲和美洲印第安人中。值得注意的是，8 个单倍型（H6～H13）是严格的亚洲人特异的。

2.3.4　讨论

在上述提到的 8 个亚洲人特异的单倍型中，H6、H7 和 H8 都有 M122 位点上的 T 到 C 的突变（图 2－11）。它们一同组成了绝大部分所研究的东亚人群中主流的单倍型，尤其是中国汉族人（平均 54.1%），并且它们不存在于亚洲之外人群中（表 2－2），这表明本研究中东亚人群来自同一个远古群体。此外，当对南方和北方的非汉族亚洲人群比较其 Y 单倍型的频率分布时，笔者发现北方人群中发现的单倍型只是南方人群中发现的单倍型的子集。例如，H7 和 H10～H12 只在南方非汉族人群中发现，不存在于北方非汉族人群中。如果我们假设 H7 和 H10～H12 这些单倍型在南北方人群中以同样的频率出现，那么在北方非汉族人群中观察不到这些南方特异单倍型的概率是 3.998×10^{-10}。这种差异在北方和南方汉族人群中也存在，尽管不像非汉族人群那样显著，原因是汉族在近期历史中频繁向南或向北迁徙。

北方和南方人群的差异通过主成分分析的结果进一步体现出来（图 2－12）。这一分析显示，所有北方人群都聚拢在右上角，同东亚南方人群很好地分隔开，后者比前者多样性更高。可以观察到，包括柬埔寨人和泰国人在内的东南亚人群有着最高的多态性，因为他们展现了几乎所有亚洲特异的单倍型（表 2－2）。鉴于这一观察，可以很合理地下结论：北方人群起源于南方人群，并且远古非洲移民进入东亚的最初定居开始于东南亚大陆，从那里他们向北部扩张至东亚的其他部分。一项由 Ballinger 等完成的研究[24]也说明了东亚人的南方起源。

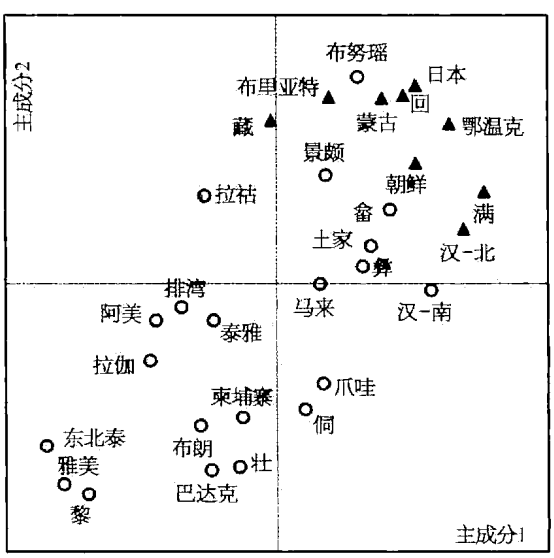


图 2－12　东亚 30 个人群的 Y 染色体单倍型频率的主成分分析

这些人群的地理分布由图 2－10 所示。三角形代表北方人群；圆形代表南方人群。这张地图涵盖了原始遗传变异的 44%。

为了估计现代人进入东亚的时间，笔者对 M122 位点上携带 C 等位基因（亚洲特异单倍型 H6～H8 所具有的等位基因）的个体的 3 个 Y 染色体微卫星位点进行了观测。在 DYS391、DYS390 和 DYS389 上分别总计观察到了 5、8 和 6 个等位基因。在这一估计中使用了单步突变模型和 0.18% 的突变率[25,26]。为了把人群亚结构对估计的可能影响

降到最低，只考虑中国汉族样本（总共 160 个 M122－C 个体）。当假设有效群体大小为 750～2 000（见材料和方法部分）时，对于 DYS390 估计的世代数为 919～3 032，是所有 3 个估计中最古老的。因此，如果假设每个世代的时间是 20 年，M122C 的历史为 18 000～60 000 年。笔者认为这个估计反映了伴随着现代人进入东亚的瓶颈事件的历史，因为东南亚人群中 M122－C 等位基因广泛存在，提示这一突变早于现代人进入东亚。

由于估计男性有效群体大小和突变率所包含的必然误差，准确地确定古人类迁徙或突变的时间很困难。然而，形态学和考古学的知识可以帮助我们缩小估计历史的范围。根据 Turner 等做过的形态学研究[27]，亚洲北部人群所谓的 Sinodont 齿系产生于 18 000～25 000 年前。一种类似的齿系在所有东南亚人群中占支配地位，被认为是 Sinodont 齿系的祖先。因此，这一牙齿进化的结果趋向于排除 18 000 年的殖民时间，这是我们对现代人进入亚洲历史估计值的下限，这一下限来自对有效群体大小的高估。此外，阿尔泰山和西伯利亚东南部贝加尔湖地区的考古发现了 25 000～45 000 年前的现代人石器文化证据[28]。因此，东亚人群的首次到来应该早于北亚的石器文化。最近的考古研究证据表明，巴布亚新几内亚在 35 000～50 000 年前被现代人定居，而澳大利亚原住民可能更早[29,30]。因此，如果我们承认东南亚大陆是所有东亚人群包括西伯利亚和大洋洲人的故乡，M122 谱系的时间深度的上限（60 000 年前）看来是一个对非洲现代人最早进入东亚时间的可能估计（目前看来，O3 亚型并不是最早定居东亚的现代人带来的，而是大约 2 万年前的移民带来的，所以本研究的下限时间估计可能更接近事实）。

末次间冰期发生在 15 000～75 000 年前，尽管在东亚的分布和确切时间尚不清楚[31]。有趣的是，在对东亚原始人类化石仔细检查后，笔者发现古人（远古智人）与现代人化石在时间上有显著的不连续性。所有的古人化石历史都超过 10 万年，而所有的现代人化石历史都小于 5 万年（其中大部分为 1 万～3 万年）。因此，东亚至今尚未发现 5 万～10 万年前的古人类化石。由于这一地区在这一时间段之前和之后都有丰富的古人类化石记录[4,5]，这种不连续性便显得尤为不寻常。中国古人类化石记录如此长时间的不连续性，以及在不连续期前后发现的古人与现代人类化石之间明显的形态区别，这两个事实的存在，使得这一化石记录裂隙完全无法随意地用“尚未发现”来解释。

总而言之，本研究所展现的证据表明，现代人第一次进入东亚南部的时间为大约 60 000 年前，之后人类向北方迁徙，恰好赶上该地区冰川退却。这可能暗示生活在东亚的古人类在末次冰川期之前或在此期间消失了，而非洲现代人的后代来到了东亚的广阔大陆上。一个 H6 和 H8 单倍型占主体地位的子群体此后艰难跋涉到了北方，这些人成为后来的中国北方人，以及再往后的西伯利亚人。

参考文献

[1] Cavalli-Sforza L L, Menozzi P, Piazza A. The history and geography of human genes. Princeton: Princeton University Press, 1994.

[2] Brooks A S, Wood B. Paleoanthropology: the Chinese side of the story. Nature, 1990, 344:

288 - 289.

[3] Li T, Etler D A. New Middle Pleistocene hominid crania from Yunxian in China. Nature, 1992, 357: 404 - 407.

[4] Wu X Z, Poirier F E. Human evolution in China. Oxford: Oxford University Press, 1995.

[5] Etler D A. The fossil evidence for human evolution in Asia. Annu Rev Anthropol, 1996, 25: 275 - 301.

[6] Wolpoff M H. Interpretations of multiregional evolution. Science, 1996, 274: 704 - 707.

[7] Cann R L, Stoneking M, Wilson A C. Mitochondrial DNA and human evolution. Nature, 1987, 325: 31 - 36.

[8] Vigilant L, Stoneking M, Harpending H, et al. African populations and the evolution of human mitochondrial DNA. Science, 1991, 253: 1503 - 1507.

[9] Stringer C B, Andrew P. Genetic and fossil evidence for the origin of modern humans. Science, 1988, 239: 1263 - 1268.

[10] Wilson A C, Cann R L. The recent African genesis of humans. Sci Am, 1992, 266: 68 - 73.

[11] Chu J Y, Huang W, Kuang S Q, et al. Genetic relationship of populations in China. Proc Natl Acad Sci USA, 1998, 95: 11763 - 11768.

[12] Zhao T, Zhang G, Zhu Y, et al. The distribution of immunoglobulin Gm allotypes in forty Chinese populations (in Chinese). Acta Anthropol Sin, 1986, 6: 1 - 8.

[13] Weng Z, Yuan Y, Du R. Analysis of the genetic structure of human populations in China (in Chinese). Acta Anthropol Sin, 1989, 8: 261 - 268.

[14] Jobling M A, Tyler-Smith C. Fathers and sons: the Y chromosome and human evolution. Trends Genet, 1995, 11: 449 - 455.

[15] Oefner P J, Underhill P A. Comparative DNA sequencing by denaturing high-performance liquid chromatography (DHPLC). Am J Hum Genet, 1995, Suppl 57: A266.

[16] Oefner P J, Underhill P A. DNA mutation detection using denaturing high performance liquid chromatography (DHPLC)//Dracopoli N C, Haines J L, Korf B R, et al. Current protocols in human genetics, New York: Wiley & Sons, 1998, Suppl 19. 7.10.1 - 7.10.12.

[17] Underhill P A, Jin L, Zemans R, et al. A pre-Columbian human Y chromosome-specific transition and its implications for human evolution. Proc Natl Acad Sci USA, 1996, 93: 196 - 200.

[18] Underhill P A, Jin L, Lin A A, et al. Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography. Genome Res, 1997b, 7: 996 - 1005.

[19] Vollrath D, Foote S, Hilton A, et al. The human Y chromosome: a 43-interval map based on naturally occurring deletions. Science, 1992, 258: 52 - 59.

[20] Underhill P A, Jin L, Lin A A, et al. An African origin of human Y chromosomes. Am J Hum Genet, 1997a, Suppl 61: A18.

[21] Kayser M, Caglia A, Corach D, et al. Evaluation of Y chromosomal STRs: a multicenter study. Int J Legal Med, 1997, 110: 125 - 133.

[22] Underhill P A, Shen P, Lin A A, et al. Y chromosome sequence variation and the history of human populations. Nat Genet, 2000, 26(3): 358-361.

[23] Takahata N. Allelic genealogy and human evolution. Mol Biol Evol, 1993, 10: 2-22.

[24] Ballinger S W, Schurr T G, Torroni A, et al. Southeast Asian mitochondrial DNA analysis reveals genetic continuity of ancient mongoloid migrations. Genetics, 1992, 130: 139-152.

[25] Heyer E, Puymirat J, Dieltjes P, et al. Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees. Hum Mol Genet, 1997, 6: 799-803.

[26] Bianchi N O, Catanesi C I, Bailliet G, et al. Characterization of ancestral and derived Y chromosome haplotypes of New World native populations. Am J Hum Genet, 1998, 63: 1862-1871.

[27] Turner C G II. Shifting continuity: modern human origin//Brenner S, Hanihara K. The origin and past of modern humans as viewed from DNA. Singapore: World Scientific, 1993: 216-243.

[28] Vasil'ev S A. The upper Paleolithic of Northern Asia. Curr Anthropol, 1993, 34: 82-92.

[29] Brown P. Recent human evolution in East Asia and Australia. Philos Trans R Soc, 1992, 337: 235-242.

[30] Swisher C C III, Rink W J, Anton S C, et al. Latest Homo erectus of Java: potential contemporaneity with *Homo sapiensin* Southeast Asia. Science, 1996, 274: 1870-1874.

[31] Dawson A G. Ice age earth: late quaternary geology and climate. New York: Chapman & Hall, 1992.

2.4 从Y染色体看东亚人群的演变

东亚是亚洲的一部分，面积广阔、风光旖旎。东亚有着世界22%的人口，主要分为4种体质类型：新石器时代东亚人，即蒙古人种；旧石器时代大洋洲人，即澳大利亚人种；旧石器时代东南亚人，即尼格利陀人种；还有欧洲人，即高加索人种。东亚有阿尔泰、南亚、南岛、侗傣、苗瑶、汉藏和印欧等7个语系200多种语言，这使得东亚成为世界上研究人类进化、遗传多样性和基因与文化相互作用的最重要区域之一[1]。

在过去数年中，分子人类学的研究者们通过使用常染色体和X染色体、父系Y染色体、母系线粒体等各类遗传标记体系来解析东亚人群的遗传多样性。常染色体和X染色体遗传自父母双方，会被重组所打乱，而Y染色体上主干的非重组区呈严格父系遗传，并且Y染色体的"有效群体大小"理论上至多为常染色体的1/4，X染色体的1/3，对漂变非常敏感，容易形成群体特异性多态标记，从而包含更多的关于群体历史的信息。Y染色体的这些特点使其成为研究人类进化和迁徙最强有力的工具之一[2,3]。

Y染色体进入人们的视野，开始于其在追溯现代人起源上的应用。自20世纪90年代以来，人类学界争论最激烈的话题是东亚地区现代人的起源问题。由于东亚出土了大量的古人类化石，一些人类学家认为东亚地区的人类是本土连续进化的，支持全球现代人的多地区起源。然而，1999年宿兵等[4]采用Y染色体非重组区的19个SNP来研究东亚人群，得出东亚地区现代人起源于非洲，并由南方进入东亚，而后向北方迁徙。随后，

2000 年柯越海等[5]对东亚地区 12 127 份男性随机样本的 Y 染色体进行 SNP 分型研究。Y 染色体突变 M168 被认为是约 6.4 万年前现代人走出非洲时所产生的突变，其原始型仅出现在东非人群中，除非洲以外的人群都是突变型。柯越海等的研究虽然没有直接检测 M168 这个突变，但他们检测了 M89、M130 和 YAP 这 3 个 M168 下游的突变。结果显示这一万多份样品无一例外都带有 M89、M130 和 YAP 3 种突变之一，也就是说都是 M168 突变型。现在来看，东亚现代人或许与一些古人种有基因交流，但从父系角度看，现存的东亚人群都是走出非洲的后裔，这是支持现代人非洲单一起源的强有力的遗传学证据。下一问题就要回答早期现代人是如何迁徙来到东亚的。

人群的迁徙和分布与气候的变迁有着密切的关系，为便于从不同角度探索和认识人群演变规律，这里介绍一些近 10 万年来的气象学材料。在距今 1 万～11 万年，也就是考古学上的旧石器时代到中石器时代，地球处于末次冰川期[9]，那段时间，海平面远低于现在，许多现在的岛屿与大陆相连，成为人类迁徙的重要通道。距今 2.65 万年到 1.9 万～2 万年是末次盛冰期，是末次冰川期中气候最寒冷、冰川规模最大的时期，亚洲的绝大部分、北欧和北美都被冰雪覆盖，人类的生存空间也随冰川蔓延而逐渐缩小。大约 1.5 万年前，气温开始转暖，冰川开始退却，现代人才迎来了迁徙的黄金时期[10,11]。

本节主要应用 Y 染色体的数据来分析东亚人群的迁徙历史，并探讨现代人最初定居东亚及其后的迁徙和扩张模式、微进化历程等。

2.4.1　北方路线还是南方路线

现阶段比较一致的看法是东亚的欧洲人种类型来自西北[10,12]，澳大利亚人和尼格利陀人来自东南[4,10]，最具争议的还是蒙古利亚人来自哪里，有 3 种可能的模式：① 蒙古利亚人由北向南迁徙，与东南亚和中国南方的尼格利陀和澳大利亚人种混合；② 蒙古利亚人来自南方；③ 北方人群来自北方，南方人群来自南方，自 1 万多年前的晚更新世以来，蒙古利亚人在南北方共同进化[13]。Y 染色体是解决这一争议的有力工具。

Y 染色体可以分为 20 种主干单倍群，编号从 A～T(P 可能不存在)，其中 O－M175、C－M130、D－M174 和 N－M231 是东亚 4 个主要单倍群，约占东亚全部男性的 93%(图 2－13)。其他单倍群，如 E－SRY4064、G－M201、H－M69、I－M170、J－P209、L－M20、Q－M242、R－M207 和 T－M70 仅占东亚男性的 7%[12]。

O－M175 是东亚最大的单倍群，约 75%的中国人以及超过 50%的日本人都可归到这一类型下，因此有理由认为它代表着蒙古利亚人。O－M175 分出 3 个主要的下游单倍群 O1a－M119、O2－M268 以及 O3－M122，这 3 个单倍群约占东亚男性的 60%[14,15]。O1a－M119 在中国东南沿海、侗傣族群、台湾少数民族中集中分布[16]。O2－M268 在汉族中占 5%以上[14]，O2a1－M95 是 O2 的主要支系，在华南、南方少数民族、中南半岛及印度门哒人群中分布较多[16,17]。O2b－M176 是 O2 下的另一支系，最主要集中于朝鲜半岛、朝鲜族和日本弥生人，越南人和汉族中也有少量分布[18,19]。O3－M122 是中国最常见的单倍群，遍及整个东亚和东南亚，占汉族的 50%～60%。O3a1c－002611、O3a2c1－

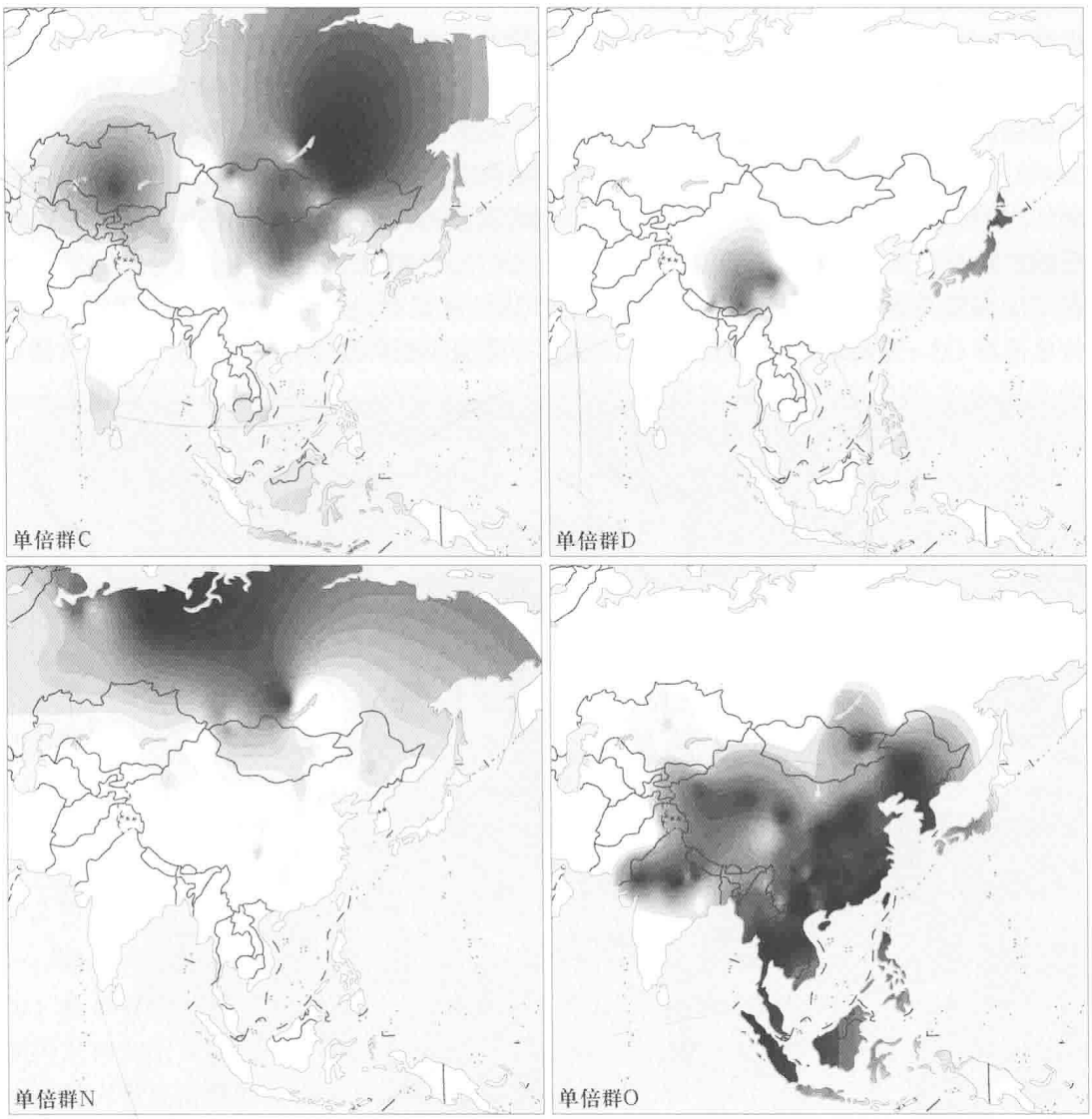


图 2-13 Y染色体总单倍群C、D、N和O在东亚的地理分布

M134和O3a2c1a-M117是O3的3个主要支系，各占汉族的12%～17%。O3a2c1a-M117在藏缅族群中也有较多分布。O3的另一支系O3a2b-M7在苗瑶和孟高棉人群中高频出现，但在汉族中却不足5%[14,15]。

宿兵等[4]在亚洲大范围群体样本中，对包括M119、M95和M122在内的19个Y染色体SNP位点以及3个STR位点进行了检测。在随后的主成分分析中，北方人群紧密聚在一起，且都被包含在南方人群的聚类簇之内，南方人群比北方人群多样性更高。他们认为北方人群来自旧石器时代定居南方的南方人群。他们还使用STR位点的一步突变模式和0.18%突变率估算O3-M122这一单倍群的时间为1.8万～6.0万年，这一时

间可能反映的是最初定居东亚的瓶颈时期。2005 年，石宏等[15]对东亚多个群体的 2 000 多个 O3 样本进行了更系统的研究，他们的研究也发现南方群体中的 O3 - M122 的多样性高于北方，支持 O3 - M122 的南方起源。他们进一步使用均方差（ASD）方法和 STR 的进化突变率（0.000 69 每位点每 25 年）[20,21]估算 O3 支系北迁的时间为 2.5 万～3.0 万年。最近，蔡晓云等[22]对东南亚的孟高棉和苗瑶族群中的 O3a2b - M7 和 O3a2c1a - M117 进行了系统研究，揭示其在 1.9 万年前末次盛冰期时经由东南亚进入东亚的单向瓶颈扩散[22]。O3 的另一主要支系 O3a1c - 002611 的 STR 位点多样性也与其他兄弟支系一样有着大体上的自南向北递减的趋势[23]。总体来看，绝大多数证据都支持 Y 染色体单倍群 O3 - M122 经由南方路线进入东亚并逐渐向北扩散（图 2 - 14）。

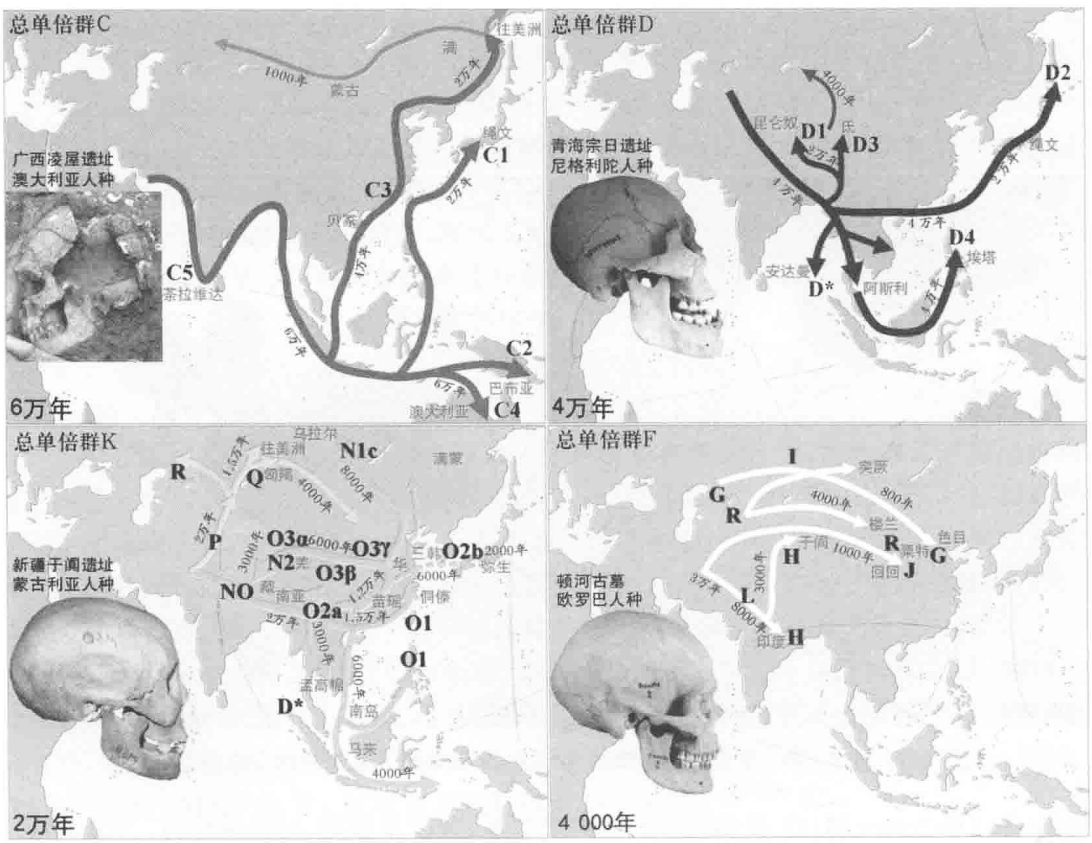


图 2 - 14　Y 染色体总单倍群 C、D、K 和 F 在东亚的迁徙

虚线表示另外可能的迁徙路线。

2.4.2　东亚最早的定居者

东亚的特征单倍群 O - M175 的产生时间，由足够多的 STR 的位点估算很可能不超过 3 万年，因此总单倍群 O 很可能根本不是东亚最早的定居者。单倍群 C - M130 的人群却极可能是最早到达东亚的人群。总单倍群 C 从阿拉伯半岛南部、巴基斯坦、印度、斯

里兰卡、东南亚、东亚、大洋洲到美洲都有分布，尤其在远东和大洋洲高频分布，但在撒哈拉以南的非洲没有被发现(图2-13)。C下游的分支，如C1-M8、C2-M38、C3-M217、C4-M347、C5-M356和C6-P55，都有着区域特异性分布[24]。C3-M217是分布最广的支系，在蒙古和西伯利亚群体中最高频出现。单倍群C1仅在日本人和琉球人中出现，但频率很低，不足5%。单倍群C2出现在从印度尼西亚东部到波利尼西亚的太平洋岛屿人群，尤其是在波利尼西亚的一些群体中，且由于连续的奠基者效应和遗传漂变而成为上述地方的特征单倍群[19,25]。C4几乎仅局限在大洋洲的澳大利亚原住民中。C5在印度及其周边的巴基斯坦和尼泊尔等地低频出现[26,27]。C6则仅出现在新几内亚高地上[28]。总单倍群C的分布模式说明了这个单倍群很可能是在亚洲大陆起源，且那时还没到达东南亚。

为更清楚说明总单倍群C的源流，钟华等[24]以取自东亚和东南亚140多个群体的465个总单倍群C作为样本，检测了C内部的12个SNP和8个STR位点。他们发现C3的STR多样性最高出现在东南亚，且呈自南向北、自东向西递减的趋势，ASD方法估算时间落在3.2万～4.2万年，这表明旧石器时代C3是沿海岸线逐渐向北扩张的(图2-14)。总单倍群C很可能在6万年前就已到达东南亚和澳大利亚，比其向北扩散的时间要早得多，这也就是说总单倍群C在蒙古利亚人(单倍群O)来之前就已在东亚生活了数万年。经过如此长的时间，总单倍群C的人群或已与蒙古利亚人有着不同的体质特征。因为现在总单倍群C的人群多有澳大利亚人的体质特征，如澳大利亚原住民、巴布亚人和一些达罗毗荼人，笔者认为总单倍群C是由澳大利亚人体质特征的人带来的，他们达到远东的时间要早于其他现代人。北京周口店出土的1万年前的人骨就有着澳大利亚人的体质特点，或也支持澳大利亚人是东亚最早定居者。

2.4.3 东亚的黑人遗存

最具神秘色彩的是Y染色体总单倍群D的迁徙历史，迄今为止我们仍对此知之甚少。总单倍群D是从非洲的DE-M1(YAP插入)单倍群衍生出来的，很可能与矮黑体质的尼格利陀人相关联。单倍群E是D的兄弟支系，E随着高黑人西迁非洲，D则可能由矮黑人东迁带到东亚(图2-15)。

单倍群D-M174在安达曼尼格利陀人、北部藏缅群体和日本的阿伊努人中高频分布，在其他东亚、东南亚和中亚群体也有低频分布(图2-13)[17,19,29,30]。D下分D1-M15、D2-M55和D3-P99 3个主要支系，还有许多未明确定位的小支系。D1在藏族、羌语支和彝语支人群中广泛分布，在东亚其他群体中也有低频分布[31,32]。D2仅分布于日本，占日本的40%以上，是上古绳文人的主要成分。D3在青藏高原东部(康区)、白马人及纳西族等群体中高频[31]。D*多在安达曼群岛被发现[30]，且已被隔离了至少2万年。其他一些被包含在D*中的小支系也多分布于西藏周边藏缅语人群、东南亚人群，阿尔泰人中也有少量来源不明的D*。这些D*的内部谱系需要详细调查分析。总单倍群D高频人群的肤色大多较深，包括安达曼人、一些藏缅和孟高棉人等。阿伊努人肤色变白可能是为了吸收更多紫外线以适应高纬度地区生存。

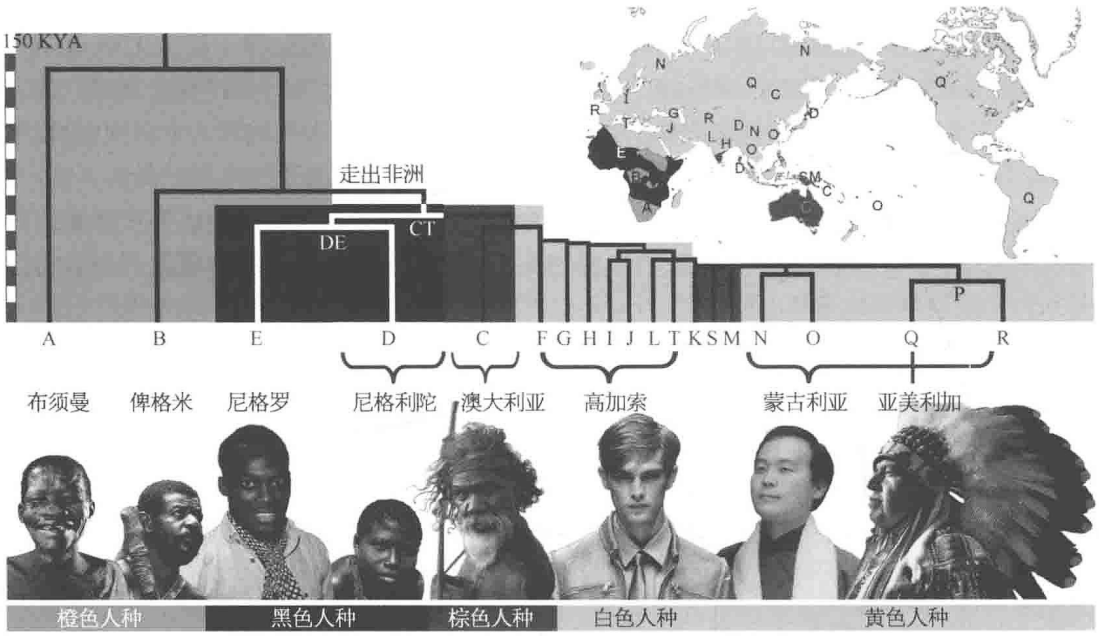


图 2-15　Y 染色体谱系的主干以及其与不同体质特征的现代人可能联系

对于总单倍群 D 的起源，Chandrasekar 等认为 CT-M168 在南亚分出了 YAP 插入和 D-M174 突变，因为他们在印度东北一些族群中发现有 YAP 插入而在安达曼群岛上检测到了 M174 突变[33]。这样来看，同样带有 YAP 插入的总单倍群 E 也很可能是亚洲起源，但没有证据进一步支持。如果总单倍群 D 诞生于非洲，那非常有趣的是它是如何随着总单倍群 CF 的群体来到东亚的？

另一个不可思议的问题是总单倍群 D 是如何由东亚的西南角一路到达日本的。它可能通过东亚大陆北上，还可能是经由巽他大陆，但穿过东亚大陆似乎更近。石宏等推论总单倍群 D 北上扩张到中国西部的时间约在 6 万年前（ASD 方法），要早于东亚其他主要支系的迁徙。随后，这一先头部队可能通过北向路线经由朝鲜半岛到达日本，或者通过南向路线经由中国台湾和琉球所形成的大陆桥到达日本，这一过程中他们可能与澳大利亚人相遇过。后因总单倍群 O 的北上以及新石器时代汉族扩张，总单倍群 D 的主体人群可能就被挤出了中国东部[31]。但是无论是遗传学上还是考古学上都没有任何证据表明 D2 或尼格利陀人曾到过中国大陆东部。相反，从马来半岛到波利尼西亚的巽他大陆至今仍有大量的尼格利陀人。尼格利陀人或许在旧石器晚期占据了整个巽他大陆。那么，这些人群可能直接从菲律宾到中国台湾和琉球。唯一难以解释的是在菲律宾的尼格利陀人中从未发现过 D 的存在，他们的父系或许已在约 1.8 万年前（BATWING 方法）被来自巴布亚岛的 C2 和 K 的扩张所取代[34]，当然也可能被非常晚近时期来自东亚大陆的总单倍群 O 所替换[35]。因为相关数据的不足，东亚的黑人遗存——总单倍群 D 的源流还远未揭开。

2.4.4 往来西北的迁徙

总单倍群O的兄弟支系是单倍群N-M231，总单倍群N在欧亚大陆北部，尤其是包括芬兰、乌戈尔、萨摩耶德和尤卡吉尔等分支的乌拉尔语人群，以及阿尔泰语人群和因纽特人中高频分布，它还低频出现在东亚内陆(图2-13)[29,36]。对于总单倍群N的详细分析显示，N在东欧的高频是缘于很晚近的迁徙，这次迁徙从1.2万～1.4万年前(ASD方法)开始，由内亚/南西伯利亚出发，走一条逆时针的北部路线[36]。N的下游分支N1a-M128低频分布于中国北部一些群体，例如满族、锡伯族、鄂温克族和朝鲜族，以及中亚的一些突厥语族群中。另一分支N1b-P43在北部的萨莫耶德人中广泛分布，也在一些乌拉尔和阿尔泰人群中低频或中频分布，N1b在6万～8万年前诞生于西伯利亚[37,38]。频率最高的下游单倍群N1c-Tat，可能在1.4万年前起源于中国西部地区，然后在西伯利亚经历多次瓶颈效应，最后扩散到东欧和北欧[36]。这些研究把总单倍群N的起源追溯到中国西南或东南亚，总单倍群N的人群艰苦跋涉由东南亚穿越大陆一直到北欧，谱写了壮丽的迁徙史诗。

总单倍群N的迁徙史为东亚人群南方起源提供了又一项强有力的证据。然而仍有一些研究在质疑南方起源。Karafet等对来自东亚和中亚地区的25个群体的1 300多份样本进行Y染色体分型研究，他们发现各单倍群间的两两差异在东亚南部是非常小的，且东亚南北群体之间并未发现遗传分化[29]。薛雅丽等[39]使用贝叶斯全似然法来分析取自中国、蒙古、韩国和日本的27个群体近1 000份样本的45个Y染色体SNP和16个STR位点。他们发现东亚北方群体的Y染色体的STR多样性要高于南方，北方群体的扩散要早于南方群体[39]。但随后石宏指出Karafet所观察到的北方群体的高多样性应是由近期的人群混合造成的，薛雅丽等分析结果也存在这一问题，蒙古族、维吾尔族和满族的基因多样性高应是由他们与西方人群以及汉族大规模混合的结果。另外，薛雅丽等所选取的南方群体代表性不够，长期地理隔离所造成的群体内部的瓶颈效应或对基因多样性的估算有较大影响[31]。

后续的争论集中在如何辨析中亚和欧亚西部人群对东亚的基因贡献。钟华等[13]通过对117个群体的近4 000份样本的Y染色体进行高分辨率的分型判断，试图阐明这一问题。在钟华等的研究中，单倍群O-M175、C-M130、D-M174和N-M231仍显出了南方路线较大基因贡献。然而，与中亚和欧亚西部相关的单倍群，例如单倍群R-M207和Q-M242，多在东亚西北地区出现，且它们的频率自西向东有递减的趋势。另外，单倍群R-M207和Q-M242的Y染色体STR多样性也提示了北方路线存在的可能性，也就是1.8万年前由中亚到北亚的迁徙和3 000年前沿丝绸之路的人群混合。

2.4.5 母语还是父语

人群的遗传模式常会被其居住习俗和生存方式等社会文化因素所影响。东亚人群的Y染色体可以很好地反映这些文化烙印。例如从父居的群体之间应表现出族群关系与父系的Y染色体，而不是与母系的线粒体较强的关联性。东亚的语言确实与父系的Y

染色体[15,22,24,40,41]和整个基因组多样性[42]有较强的对应关系，但不与母系线粒体相关。例如，Y 染色体单倍群 O3 - M134 与汉藏语系人群相关[15,40]，O2 - M95 与南亚语系相关[41]。而且语言学家们所提出的语系之间的系统发生关系也可反映在 Y 染色体上，但不与全基因组多样性相关。例如，苗瑶和南亚语系间的近缘性由单倍群 O3 - M7 所反映[22]，侗傣语系和南岛语系间的近缘性由 O1 - M119 所反映[43]。

另一有趣的话题是关于语言扩张和 Y 染色体的分布模式。现代语言在扩张过程中是否也经历了一系列的奠基者效应？Atkinson 发现音素(元音、辅音和声调)的多样性有着出非洲的递减趋势，因而推论现代语言也是源自非洲[44]。但是这一论断并未得到语言学界的广泛认可。王传超等指出，Atkinson 的结论仅在音素多样性被简单分为 3～5 类之后才成立，而应用没有简化的音素原始数值进行分析却得出了由亚洲中部向外的多样性递减趋势[45]。当然这不是关于语言源流的最终结论[46]。音素多样性的分布模式反映的或许不是现代人的最初起源过程，而是现代人在亚洲中南部的二次扩张。Y 染色体的数据也支持亚洲扩张。Y 染色体谱系树的根部在非洲，但仅是总单倍群 A 和 B 是非洲本土类型，这两个单倍群或许从未离开非洲。其他单倍群都在 CF 和 DE 之下，均源自一个 5 万～7 万年前走出非洲的古老突变 M168。他们或在 3 万～4 万年前从西亚扩张开来，并衍生出了由 C 到 T 的所有单倍群[3]。因此，在非洲高频出现的单倍群 E 很可能来自亚洲扩张后的回流。在非洲，单倍群 A 集中出现在科依桑人和撒哈拉人中，单倍群 B 则主要出现在俾格米人及刚果周围的其他群体中。非洲的主要群体，也就是班图人或尼日尔-刚果人，或许都是亚洲回流到非洲的。单倍群 DE - M1(YAP +)经由亚洲返回非洲已被一些研究者提出[47-49]，但这推论也遭到了其他人质疑[50,51]。此外，在喀麦隆高频出现的 R1 - M173 更明确地支持由亚洲向撒哈拉以南非洲的人群回迁[52]。

语言与 Y 染色体相关但不与线粒体相关，或许反映的是由从父居所引起的性别偏向性迁徙过程。从父居指的是夫妻婚后与丈夫的父母住在一起或住在他们附近。Forster 等指出如果父母双方语言不同，那么多是父亲的语言在家庭中占主导地位[53]。然而，因为全基因组多样性也与语系相关，语系形成之时父系和母系应该已完好保存下来。因此，线粒体与语言间的不相关或不能简单地用从父居来解释。有可能是因为狩猎或战争使得最初群体中女性的有效群体大于男性，那么同一语言群体中线粒体受到遗传漂变的影响就小。当然其他解释，例如优选男性、男性生育周期长、后代的数目以及突变率的不同等也都有可能。

其他文化因素，例如农业、军事和社会地位等，都有可能影响遗传模式。文化的传播模式有两种：一是人口流动驱动文化传播；二是单纯的文化传播，不涉及人群的流动[54]。举例而言，近东到欧洲的农业传播是否伴随着大量的新石器农民不断迁徙一直是争论的焦点[55]。Chikhi 等应用家系似然法去分析包含 22 个 Y 染色体 SNP 的大批量数据，发现近东的农业进入欧洲伴随着大量的基因流动，这支持了人口流动驱动文化传播的模式[56]。另一个例子，文波等研究了 28 个汉族群体的 Y 染色体和线粒体的遗传结构，结果显示南北方汉族有着相近的 Y 染色体结构，但是南北汉族在线粒体上却有着较大差

别。或许是由于战乱或饥荒，大量的北方移民来到南方而改变了中国南方的遗传结构。他们认为汉族人口扩张和汉文化的南下是符合人口驱动的文化传播模式的，而且在这次扩张中男性占主体地位。

2.4.6 历史名人的Y染色体

人群扩张还可以与特殊的社会地位相联系，比如成吉思汗家族。成吉思汗(1162—1227)南征北战，建立了历史上疆域最辽阔的国家。他和他的父系亲属因其很高的政治地位而有不计其数的后代，这无疑增高了他们的Y染色体在群体中所占的频率。结果，可能是成吉思汗或者其近亲宗族的Y染色体类型(C3 * xC3c，星簇)出现在了从太平洋一直延伸到里海的广阔地域，占全世界男性的0.5%[57]。有趣的是，C3 * 星簇最高频地出现在哈萨克斯坦的克烈部[58]。克烈部高频的C3 * 星簇难以归因于成吉思汗，成吉思汗家族的Y染色体类型或许不是星簇。无论如何，社会选择确实在C3 * 星簇的扩张中起到了重要作用。同样，Y染色体单倍群C3c－M48被推断为清代(1644—1912)满族皇室的类型，占东亚男性的3.3%[59]。

与研究成吉思汗的谱系一样，Y染色体可以用来追溯历史名人。人们的姓氏大多继承自父亲，而Y染色体是严格的父子相传的基因组片段。所以姓氏与Y染色体的遗传应该是平行的，有共同姓氏的男性可能有相同或相近的Y染色体类型。那么，结合家谱材料，通过研究历史人物现存后代的Y染色体，可以揭示历史人物之间的父系关系[60]。王传超等用Y染色体分型比对的方法确认了若干有1 800多年历史、延续70～100代的大跨度家系，这些家系宣称是魏武帝曹操后裔。曹操后裔的Y染色体类型为O2 * －M268，与西汉丞相曹参后裔的Y染色体O3－002611并不一致。所以，曹操一直自称的源自曹参的贵族血统并不被遗传学所支持[61]。

将遗传学应用到古代史研究将会不断增多。举例而言，在家谱学上，同一姓氏的不同家族可以通过遗传学检测来填补谱牒材料的缺环[62]。深度家系对于研究Y染色体的进化也有很大价值，例如薛雅丽等[63]通过测序相隔13代的两个体的Y染色体得出Y染色体上的碱基突变率为3.0×10^{-8}/(位点·世代)。更深度的家系将会是更准确估算突变率的更好的材料。

2.4.7 总结展望

Y染色体在解析东亚现代人源流史中起到了重要作用。尽管有许多问题仍有待探索，但史前迁徙过程的基本框架已经明晰。占东亚男性90%以上的C、D、N和O 4个总单倍群很可能起源于东南亚，随澳大利亚人、尼格利陀人和蒙古利亚人这3种不同体质特征的现代人经历了3次大的迁徙浪潮。欧亚中西部特征的Y染色体单倍群E、G、H、I、J、L、Q、R和T在中国西北的分布模式反映出了来自西方的近期基因交流和可能的北部路线的影响，这些单倍群自西向东的递减趋势也可以清晰地观察到。

然而，现阶段东亚的Y染色体研究遇到了两个瓶颈。一是东亚特异单倍群O－M175

的解析度太低。虽然，总单倍群O人口众多，但O下的位点却比R和E都少。例如，002611、M134和M117这3个位点代表了东亚近2.6亿人，但没有更下游的位点可以用来更精细解析这些群体的遗传结构。另一个瓶颈是支系和群体分化时间的估算。现在绝大部分的时间估算用的是Y染色体的STR位点，尽管这在理论上说得通，但对于最恰当的STR估算时间的方式还一直有争议。尤其值得提出的是这里有两种经常用到的Y染色体STR突变率，即进化突变率[20,21]和家系突变率[64]。如何选用这两种突变率争议很大，因两者估算出的时间甚至可相差3倍。而且STR位点的相似性及多变性也使得时间估算的准确度大打折扣。因此，文中提到一些时间点也仅仅是作为某些单倍群或人群分化的粗略参考。

二代测序技术的不断发展使得全测序大样本量和深度家系的Y染色体成为可能。例如，千人基因组计划在其低覆盖项目中已经以1.83的平均深度测序了77个男性的Y染色体，以15.23的深度测序了两个连续三代的男性家系[65]。更进一步的深度测序不但可以细化Y染色体谱系树，又可以为进化研究提供较精确的生物钟校准。

参考文献

[1] Cavalli-Sforza L L. The Chinese human genome diversity project. Proc Natl Acad Sci USA, 1998, 95: 11501-11503.

[2] Jobling M A, Tyler-Smith C. Father and sons: the Y chromosome and human evolution. Trends Genet, 1995, 11: 449-456.

[3] Underhill P A, Shen P, Lin A A, et al. Y chromosome sequence variation and the history of human populations. Nat Genet, 2000, 26: 358-361.

[4] Su B, Xiao J, Underhill P, et al. Y chromosome evidence for a northward migration of modern human into East Asia during the Last Ice Age. Am J Hum Genet, 1999, 65: 1718-1724.

[5] Ke Y, Su B, Song X, et al. African origin of modern humans in East Asia: a tale of 12 000 Y chromosomes. Science, 2001, 292: 1151-1153.

[6] Green R E, Krause J, Briggs A W, et al. A draft sequence of the Neandertal genome. Science, 2010, 328: 710-722.

[7] Reich D, Green R E, Kircher M, et al. Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature, 2010, 468: 1053-1060.

[8] Wang C C, Farina S E, Li H. Neanderthal DNA and modern human origins. Quatern Int, 2012, doi: 10.1016/j.quaint.02.027.

[9] Shi Y F, Cui Z J, Li J J. Quaternary glacier in eastern China and the climate fluctuation. Beijing: Science Press, 1989.

[10] Jobling M A, Hurles M, Tyler-Smith C. Human evolutionary genetics (origins, peoples and disease). New York: Garland Science, 2004.

[11] Clark P U, Dyke A S, Shakun J D, et al. The Last Glacial Maximum. Science, 2009, 325: 710-714.

[12] Zhong H, Shi H, Qi X B, et al. Extended Y chromosome investigation suggests postglacial

migrations of modern humans into East Asia via the northern route. Mol Biol Evol, 2011, 28(1): 717 - 727.

[13] Piazza A. Towards a genetic history of China. Nature, 1998, 395: 636 - 639.

[14] Yan S, Wang C C, Li H, et al. Genographic consortium: an updated tree of Y chromosome haplogroup O and revised phylogenetic positions of mutations P164 and PK4. Eur J Hum Genet, 2011, 19(9): 1013 - 1015.

[15] Shi H, Dong Y L, Wen B, et al. Y chromosome evidence of southern origin of the East Asian-specific haplogroup O3 - M122. Am J Hum Genet, 2005, 77(3): 408 - 419.

[16] Kayser M, Choi Y, van Oven M, et al. The impact of the Austronesian expansion: evidence from mtDNA and Y chromosome diversity in the Admiralty Islands of Melanesia. Mol Biol Evol, 2008, 25(7): 1362 - 1374.

[17] Su B, Jin L, Underhill P, et al. Polynesian origins: insights from the Y chromosome. Proc Natl Acad Sci USA, 2000, 97(15): 8225 - 8228.

[18] Ding Q L, Wang C C, Farina S E, et al. Mapping human genetic diversity on the Japanese Archipelago. Advances in Anthropology, 2011, 1(2): 19 - 25.

[19] Hammer M F, Karafet T M, Park H, et al. Dual origins of the Japanese: common ground for hunter-gatherer and farmer Y chromosomes. J Hum Genet, 2006, 51: 47 - 58.

[20] Zhivotovsky L A. Estimating divergence time with the use of microsatellite genetic distances: impacts of population growth and gene flow. Mol Biol Evol, 2001, 18: 700 - 709.

[21] Zhivotovsky L A, Underhill P A, Cinnioğlu C, et al. The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. Am J Hum Genet, 2004, 74: 50 - 61.

[22] Cai X, Qin Z, Wen B, et al. Genographic consortium: human migration through bottlenecks from southeast Asia into East Asia during Last Glacial Maximum revealed by Y chromosomes. PLoS One, 2011, 6(8): e24282.

[23] Wang C C, Yan S, Qin Z D, et al. Late Neolithic expansion of ancient Chinese revealed by Y chromosome haplogroup O3a1c - 002611. J Syst Evol, 2012, doi: 10.1111/j.1759 - 6831.2012.00244.x.

[24] Zhong H, Shi H, Qi X B, et al. Global distribution of Y chromosome haplogroup C reveals the prehistoric migration routes of African exodus and early settlement in East Asia. J Hum Genet, 2010, 55(7): 428 - 435.

[25] Kayser M, Brauer S, Cordaux R, et al. Melanesian and Asian origins of polynesians: mtDNA and Y chromosome gradients across the Pacific. Mol Biol Evol, 2006, 23: 2234 - 2244.

[26] Sengupta S, Zhivotovsky L A, King R, et al. Polarity and temporality of high-resolution Y chromosome distributions in India identify both indigenous and exogenous expansions and reveal minor genetic influence of Central Asian pastoralists. Am J Hum Genet, 2006, 78: 202 - 221.

[27] Gayden T, Cadenas A M, Regueiro M, et al. The Himalayas as a directional barrier to gene flow. Am J Hum Genet, 2007, 80: 884 - 894.

[28] Karafet T M, Mendez F L, Meilerman M B, et al. New binary polymorphisms reshape and

increase resolution of the human Y chromosomal haplogroup tree. Genome Res, 2008, 18: 830 - 838.

[29] Karafet T M, Xu L, Du R, et al. Paternal population history of East Asia: sources, patterns, and microevolutionary processes. Am J Hum Genet, 2001, 69(3): 615 - 628.

[30] Thangaraj K, Singh L, Reddy A G, et al. Genetic affinities of the Andaman Islanders, a vanishing human population. Curr Biol, 2003, 13(2): 86 - 93.

[31] Shi H, Zhong H, Peng Y, et al. Y chromosome evidence of earliest modern human settlement in East Asia and multiple origins of Tibetan and Japanese populations. BMC Biol, 2008, 6: 45.

[32] Wen B, Xie X, Gao S, et al. Analyses of genetic structure of Tibeto-Burman populations reveals sex-biased admixture in southern Tibeto-Burmans. Am J Hum Genet, 2004, 74(5): 856 - 865.

[33] Chandrasekar A, Saheb S Y, Gangopadyaya P, et al. YAP insertion signature in South Asia. Ann Hum Biol, 2007, 34: 582 - 586.

[34] Delfin F, Salvador J M, Calacal G C, et al. The Y chromosome landscape of the Philippines: extensive heterogeneity and varying genetic affinities of Negrito and non-Negrito groups. Eur J Hum Genet, 2011, 19: 224 - 230.

[35] Scholes C, Siddle K, Ducourneau A, et al. Genetic diversity and evidence for population admixture in Batak Negritos from Palawan. Am J Phys Anthropol, 2011, 146: 62 - 72.

[36] Rootsi S, Zhivotovsky L A, Baldovic M, et al. A counter-clockwise northern route of the Y chromosome haplogroup N from Southeast Asia towards Europe. Eur J Hum Genet, 2007, 15: 204 - 211.

[37] Derenko M, Malyarchuk B, Denisova G, et al. Y chromosome haplogroup N dispersals from south Siberia to Europe. J Hum Genet, 2007, 52(9): 763 - 770.

[38] Mirabal S, Regueiro M, Cadenas A M, et al. Y chromosome distribution within the geo-linguistic landscape of northwestern Russia. Eur J Hum Genet, 2009, 17(10): 1260 - 1273.

[39] Xue Y, Zerjal T, Bao W, et al. Male demography in East Asia: a north-south contrast in human population expansion times. Genetics, 2006, 172(4): 2431 - 2439.

[40] Wen B, Li H, Lu D, et al. Genetic evidence supports demic diffusion of Han culture. Nature, 2004, 431: 302 - 305.

[41] Kumar V, Reddy A N, Babu J P, et al. Y chromosome evidence suggests a common paternal heritage of Austro-Asiatic populations. BMC Evol Biol, 2007, 7: 47.

[42] The HUGO Pan-Asian SNP Consortium. Mapping human genetic diversity in Asia. Science, 2009, 326: 1541.

[43] Li H, Wen B, Chen S J, et al. Paternal genetic affinity between Western Austronesians and Daic populations. BMC Evol Biol, 2008, 8: 146.

[44] Atkinson Q D. Phonemic diversity supports a serial founder effect model of language expansion from Africa. Science, 2011, 332: 346 - 349.

[45] Wang C C, Ding Q L, Tao H, et al. Comment on "Phonemic diversity supports a serial founder effect model of language expansion from Africa". Science, 2012, 335: 657.

[46] Atkinson Q D. Response to comment on "phonemic diversity supports a serial founder effect

model of language expansion from Africa". Science, 2012, 335: 657.

[47] Hammer M F, Spurdle A B, Karafet T, et al. The geographic distribution of human Y chromosome variation. Genetics, 1997, 145: 787 - 805.

[48] Hammer M F, Karafet T, Rasanayagam A, et al. Out of Africa and back again: nested cladistic analysis of human Y chromosome variation. Mol Biol Evol, 1998, 15: 427 - 441.

[49] Hammer M F, Karafet T M, Redd A J, et al. Hierarchical patterns of global human Y chromosome diversity. Mol Biol Evol, 2001, 18: 1189 - 1203.

[50] Underhill P A, Passarino G, Lin A A, et al. The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. Ann Hum Genet, 2001, 65: 43 - 62.

[51] Underhill P A, Roseman C C. The case for an African rather than an Asian origin of the human Y chromosome YAP insertion//Jin L, Seielstad M, Xiao C. Recent advances in human biology, Vol. 8: genetic, linguistic and archaeological perspectives on human diversity in Southeast Asia. New Jersey: World Scientific Publishing, 2001: 43 - 56.

[52] Cruciani F, Santolamazza P, Shen P D, et al. A back migration from Asia to sub-Saharan Africa is supported by high-resolution analysis of human Y chromosome haplotypes. Am J Hum Genet, 2002, 70: 1197 - 1214.

[53] Forster P, Colin R. Mother tongue and Y chromosomes. Science, 2011, 333: 1390 - 1391.

[54] Cavalli-Sforza L L, Menozzi P, Piazza A. The history and geography of human genes. Princeton: Princeton Univ, Press, 1994.

[55] Sokal R, Oden N L, Wilson C. Genetic evidence for the spread of agriculture in Europe by demic diffusion. Nature, 1991, 351: 143 - 145.

[56] Chikhi L, Nichols R A, Barbujani G, et al. Y genetic data support the Neolithic demic diffusion model. Proc Natl Acad Sci USA, 2002, 99: 11008 - 11013.

[57] Zerjal T, Xue Y, Bertorelle G, et al. The genetic legacy of the Mongols. Am J Hum Genet, 2003, 72: 717 - 721.

[58] Abilev S, Malyarchuk B, Derenko M, et al. The Y chromosome C3 * star-cluster attributed to Genghis Khan's descendants is present at high frequency in the Kerey clan from Kazakhstan. Hum Biol, 2012, 84(1): 79 - 89.

[59] Xue Y, Zerjal T, Bao W, et al. Recent spread of a Y chromosomal lineage in northern China and Mongolia. Am J Hum Genet, 2005, 77: 1112 - 1116.

[60] Wang C C, Yan S, Li H. Surnames and the Y chromosomes. Commun Contemp Anthropol, 2010, 4: e5/27 - 34.

[61] Wang C, Yan S, Hou Z, et al. Present Y chromosomes reveal the ancestry of Emperor CAO Cao of 1 800 years ago. J Hum Genet, 2012, 57(3): 216 - 218.

[62] Sykes B, Irven C. Surname and the Y chromosome. Am J Hum Genet, 2000, 66: 1417 - 1419.

[63] Xue Y, Wang Q, Long Q, et al. Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree. Curr Biol, 2009, 19(17): 1453 - 1457.

[64] Gusmão L, Sánchez-Diz P, Calafell F, et al. Mutation rates at Y chromosome specific microsatellites. Hum Mutat, 2005, 26: 520 - 528.

[65] 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature, 2010, 467: 1061 - 1073.

2.5　空间分析揭示的中国人群父系和母系遗传结构的差异

2.5.1　研究背景

对考古学、解剖学、语言学和遗传学的数据分析表明，在中国北方人群和南方人群之间存在明显的分界线[1]。在经典的遗传标记[2-5]、STR 标记[6]、线粒体 DNA(mtDNA)[7-9]和 Y 染色体 SNP 标记上[10,11]都可以观察到中国南方人群和北方人群之间的遗传分化。然而，这个遗传分界线的具体位置和强度目前还有争议[10,12]。使用经典的遗传标记，肖春杰等[5]提出这个遗传分界线大约位于长江沿线。文波等[7]发现，线粒体 DNA 单倍群的分布显示了中国汉族的北方人群和南方人群之间有着相当显著的分化，而北方汉族和南方汉族在 Y 染色体单倍群上的分化却难以辨识。此外，丁远春等[12]使用了 3 种不同的人类遗传标记系统(线粒体 DNA、Y 染色体和常染色体 STR)，发现中国北方-南方的梯度实际上是连续渐变的，由此他们得出了这样的结论：用一个简单的距离隔离模型可以更好地描述中国南方人群和北方人群之间的遗传分界线(中国南方人群和北方人群之间的遗传分化是由距离隔离造成的)。因此，在不同的研究中不一致甚至是有些冲突的观察结果，需要科学家对东亚人群的空间遗传结构进行更贴近、更细致的研究，这样细致的研究结果可以让我们更好地理解人群的起源与进化，也可以有助于分子流行病学的研究设计。在本节中，我们系统地探究了中国人群的空间遗传结构以及北方人群和南方人群之间分化的分界线。

为了描述北方人群和南方人群之间的分化，特别是他们之间的分界线，我们必须使用统计学的技术来处理人群的地理位置和他们的高维度遗传数据。在前面的研究中都使用了常见的统计方法[2,3,7-12]，如主成分分析(PCA)和聚类分析，但是这些方法并不适合用于反映空间和地理的信息[13]。在本节中，我们结合主成分分析与逆向距离加权(IDW)插值法来使群体的遗传模式可视化，并且发现在线粒体 DNA 与 Y 染色体数据中的地理遗传梯度[14]。同时，我们使用了改进的 Monmomier 算法来辨识空间遗传的分界线[13,15]，遗传距离直方图被用来辨识空间自相关性的统计显著性[16]。地理信息系统(GIS)是一个非常有效的管理、分析和展示地理信息的工具，在本节中我们使用这个系统在地图上展示空间遗传模式。该系统整合了常用数据库的操作和统计分析，并且可以用地图以独特的方式来展示地理信息。

在人类群体遗传学中，关于线粒体 DNA 和 Y 染色体多样性的研究非常普遍[1,17]。这类研究为分析遗传结构的空间模式提供了充足的数据。在本节中，我们在 143 个中国人群中分别构建了 36 个线粒体 DNA 单倍群和 91 个 Y 染色体单倍群的空间数据库。这些数据可以用来分析和描述中国人群中的空间遗传结构和遗传分化的分界线，主要用于比较母系和父系之间的遗传结构。

2.5.2 材料和方法

1. 样本和样本的空间数据库

本节收集了以前研究中来自中国各地80个操不同方言的人群中3 193个无血缘关系的个体Y染色体和线粒体DNA数据[7,18-20]。还有一些数据是从文献中获得，然后加入到研究中来的。最终数据的样本总量是91个中国人群3 435个个体的线粒体DNA数据和143个中国人群5 790个个体的Y染色体数据。这些数据囊括了来自中国所有省份的样本。

2. 线粒体DNA和Y染色体多倍型和单倍群

在空间数据库中，根据线粒体高变区段(HVS)的基序模式和编码区的变异区段定义36个单倍群(A、B*、B4、B4a、B4b1、B5*、B5a、B5b、C、D*、D5、D5a、F*、F1a、F1b、F1c、F2a、G、M*、M7*、M7a、M7a1、M7b*、M7b1、M7b2、M7c、M8a、M9、N*、N9a、R*、R9a、R9b、R9c、Y和Z)，随后针对这些单倍群用东亚的线粒体DNA构建的系统发生树来确定进化关系。13个双等位基因的Y染色体标记YAP、M15、M130、M89、M9、M122、M134、M119、M110、M95、M88、M45和M120被用来定义9个单倍群(C、D、F*、K*、O3*、O3e、O1、O2a和P*)，并使用了国际Y染色体命名委员会的命名法来命名。

3. 地理遗传梯度的探测

为了量化线粒体DNA和Y染色体的空间方差，结合主成分分析与逆向距离加权(IDW)插值法来展示群体的遗传模式和探测地理遗传梯度[14,22]。主成分分析先是用来获得每个群体的主要成分的得分(PC1和PC2)。逆向距离加权(IDW)插值法用来获得PC1和PC2的合成地图。IDW插值法用来产生基因频率分布的等高线图，在等高线图中的每一个点的插值估算是根据地点附近的值加权它们到该点的距离后得到的。在本节中，自然断开(Jenks)法被用于地理遗传梯度的分类[23]。

4. 空间遗传分界线的识别

空间分界线表示在某个地点可以观测到突然的变化。在目前的研究中，“改进的Monmomier算法”模型(BARRIER，版本2.2)[13,15,24]常被用来识别空间分界线。Monmomier算法的目标是通过发现相邻配对样本(人群)之间最大的差异，在一个地理地图上显示遗传距离矩阵中所包含的数据。我们用F_{ST}统计结果作为距离尺度。

5. 空间遗传自相关性的探测

为了同时描述多个单倍群的空间模式，使用F_{ST}作为距离尺度，应用空间遗传软件(SGS，版本1.0d)中的遗传距离直方图分析来探测空间遗传的自相关性[16]。遗传距离直方图表示的是：如果所有配对群体的平均遗传距离(F_{ST})都属于一个空间距离类别，就用这个平均遗传距离对空间距离类别作图，模拟过程被用来测试空间遗传自相关性的统计学显著程度。对于描述一个单一的单倍群的空间模式，使用Crime-Stat III软件(版本3.0)中的Moran correlogram功能加上Moran's I统计量[25,26]，来探测每个单倍群空间遗传的自相关性。蒙特卡罗模拟过程被用来探测空间遗传的自相关性的显著性。

在本节中所有地图都由ArcGIS9.0(Environmental Systems Research Institute Inc.)产生。

2.5.3　研究结果

1. 母系和父系地理遗传的梯度

图 2－16 显示使用了线粒体 DNA 和 Y 染色体单倍群的主成分 1 和主成分 2 的得分插值产生的地理遗传的梯度。图 2－16a 是线粒体主成分 1 地图(贡献率为 19.83%)，图 2－16b 是线粒体主成分 2 地图(贡献率为 14.84%)，这 2 个图是使用 91 个群体中的 36 个线粒体 DNA 单倍群得到的。主成分 1 地图揭示了一个显而易见的北方-南方地理遗传梯度，而主成分 2 地图则显示了一个东西向的梯度。图 2－16c 是 Y 染色体主成分 1 地图(贡献率为 33.07%)，图 2－16d 是 Y 染色体主成分 2 地图(贡献率为 17.07%)。这两个图是使用 143 个群体中的 9 个 Y 染色体单倍群得到的。Y 染色体的北方-南方梯度不

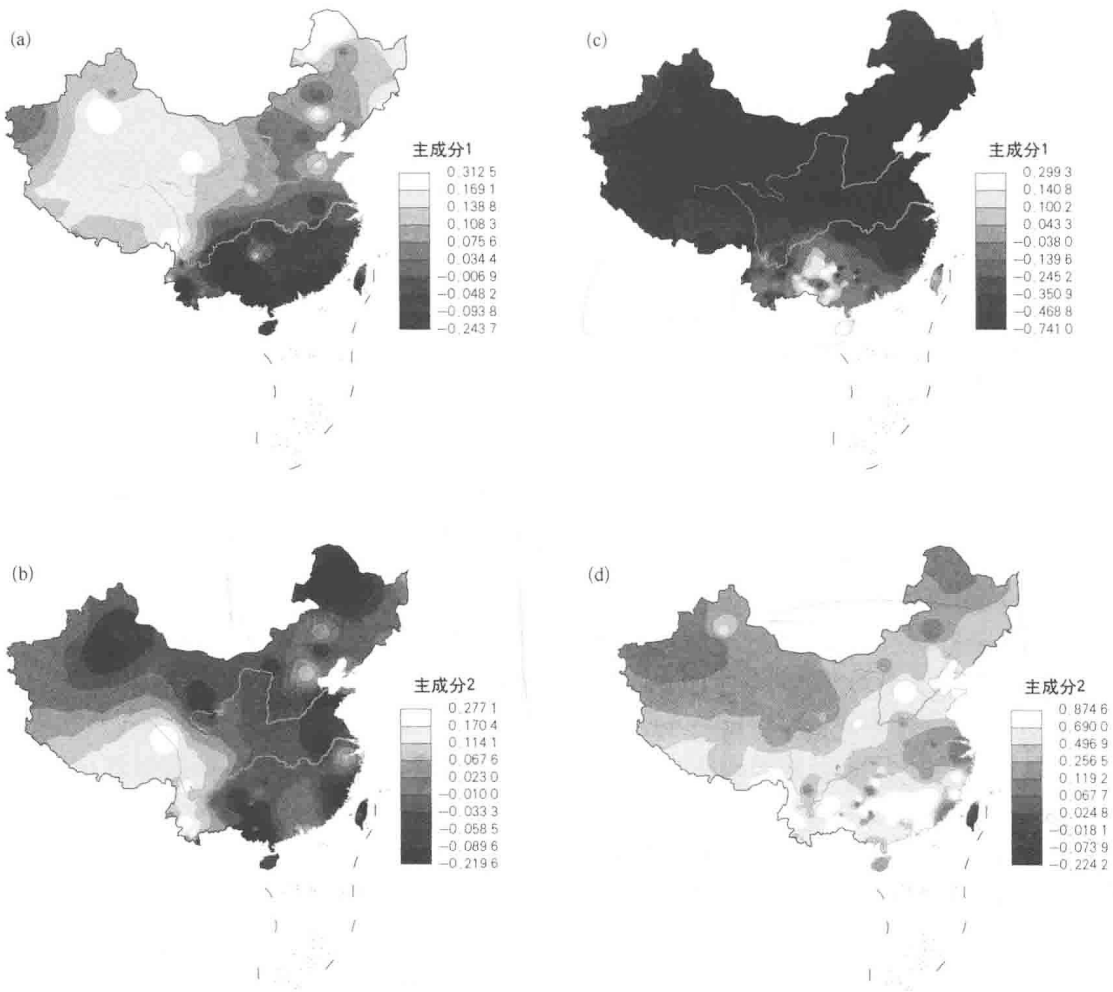


图 2－16　线粒体 DNA 和 Y 染色体单倍群的主成分 1 和主成分 2 得分插值产生的地理遗传梯度

使用 91 个人群中的 36 个线粒体 DNA 单倍群的(a)主成分 1(19.83%)和(b)主成分 2(14.84%)的分值产生的地理遗传的梯度图。(c)和(d)使用 143 个人群中的 9 个 Y 染色体单倍群的主成分 1(33.07%)和主成分 2(17.07%)分值产生的地理遗传的梯度图。

是那么明显。因此，在母系和父系上的地理遗传梯度的模式是不同的。

2. 母系和父系空间遗传的分界线

当把汉族人群和非汉族人群的线粒体 DNA 和 Y 染色体数据都包括在内时，我们发现遗传分界线主要位于外围的山区(图 2-17a、c)，没有观测到在北方和南方之间有统计学上显著的遗传分界线。

然而，当我们只观测汉族人群的数据，北方人群和南方人群之间的遗传分界线开始显现(图 2-17b、d)。这个分界线在母系上是统计学显著的，但是在父系上则要弱化得多。在母系上，在北方人群和南方人群之间具有显著不间断的遗传分界线，最明显的遗传分界线大致是长江以北的淮河-秦岭一线(图 2-17b)。我们还可以观测到另外两条分界线，一条在长江以南，另一条在黄河以北，虽然从自展值上来看，它们统计学上的重要性不如中间的那条分界线(图 2-17b)。在父系上，显现的遗传分界线显示了一个和母系分界线完全不同的模式。父系的遗传分界线可以被观测到，不过是一种更碎片化的模式。在北方和南方之间也存在不间断的遗传分界线，但是这些界线在统计学上并不显著(图 2-17d)，显示了北方人群和南方人群在父系上的分化比在母系上要小很多。

3. 母系和父系遗传结果的空间自相关性

我们使用了遗传距离直方图分析来检查多个单倍群的空间自相关性的统计显著性。图 2-18 分别显示了基于线粒体 DNA 和 Y 染色体单倍群频率计算后得到的 14 个空间距离类群的平均 F_{ST} 的遗传距离直方图。在这些图中，平均遗传距离在地理距离上作图。平均遗传距离在图中是通过 1 000 次模拟得到 95% 的置信区间的线条。遗传距离直方图再次显示了母系和父系之间不同的空间结构。

在母系上，在 91 个汉族和非汉族人群中都找不到空间的自相关性(图 2-18a)，在 19 个汉族人群中也找不到空间的自相关性(图 2-18b)。这个结果表明在中国人群中母系遗传亚结构的分化是随机分布的。在父系上，遗传距离随着地理距离的增加而增加(图 2-18e、f)。遗传距离在地理间隔为 1 800～2 100 km 的类群中显著增加，表明在中国北方到南方的地形上存在着一个显著的父系空间自相关性。

为了进一步测试每个单倍群的线粒体 DNA 和 Y 染色体的空间自相关性，笔者使用了 Moran's I 统计量和 Moran coreelogram 软件来探测每个单倍群的线粒体 DNA 和 Y 染色体的空间自相关性。每一个单倍群的空间遗传自相关性的统计显著性由 1 000 次蒙特卡罗模拟过程测试后得到。类似的，Moran coreelogram 软件中的 Moran's I 统计量也显示了母系和父系之间不同的空间结构。

在母系上，当把所有的汉族和非汉族人群都包括在内，其中 21 个单倍群(A、D*、D5a、G、M9、Z、M*、B4a、B5a、F*、F1a、M7*、M7b*、M7b1、B*、B5*、F1b、F2a、N9a、R9c 和 Y)显示了空间自相关性，而其他 15 个单倍群(C、D5、M7c、M8a、N*、B4b、B4b1、R9a、R9b、B4、B5b、F1c、M7a、M7b2 和 R*)中并不存在空间自相关性。这一结果表明了有些线粒体 DNA 单倍群在它们所在的地理区域内保存了丰富的空间自相关性，而其他

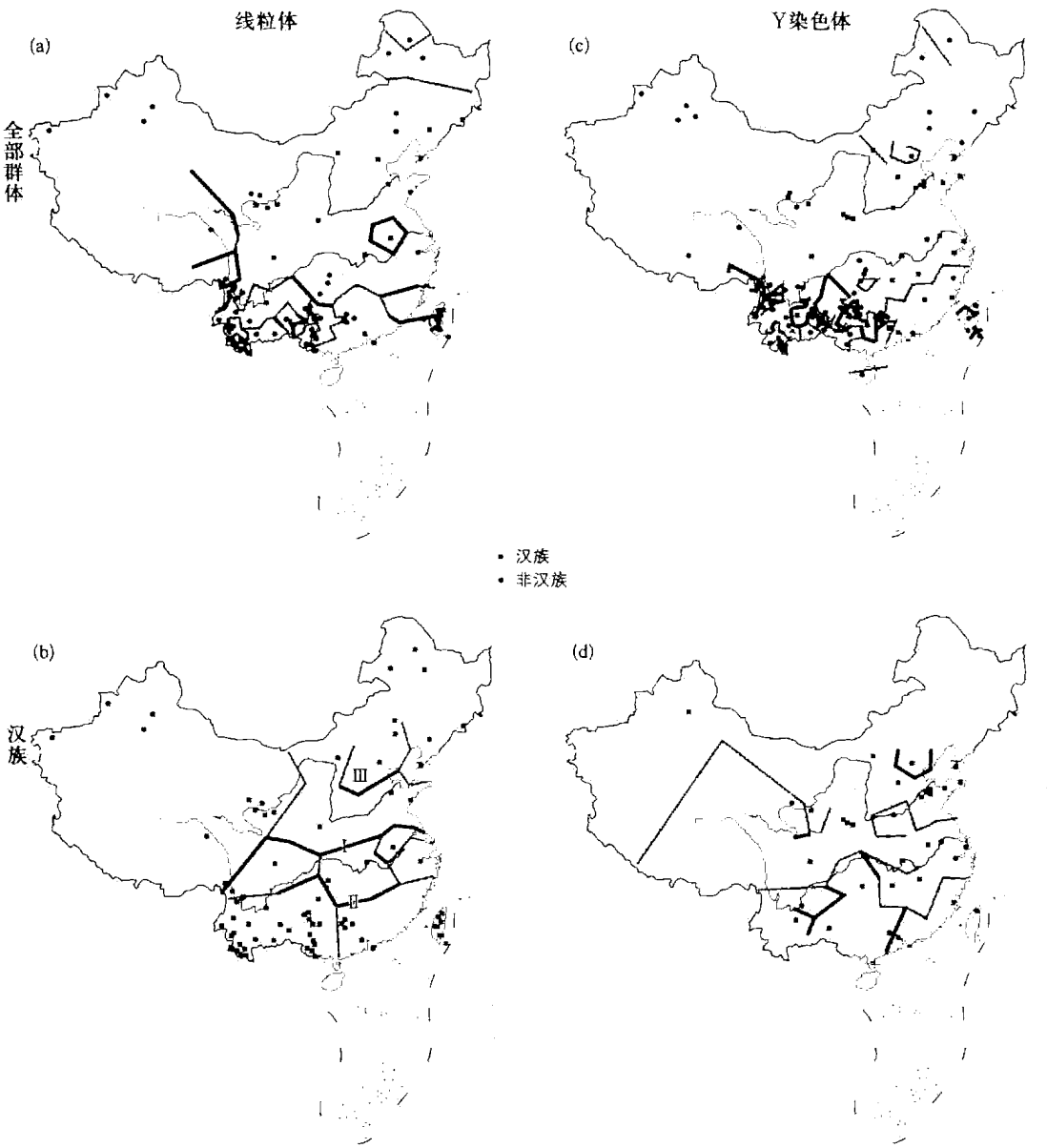


图 2－17　中国人群中的线粒体 DNA 和 Y 染色体的空间遗传分界线

在(a)中的分界线是用汉族和非汉族 91 个人群的 36 个线粒体 DNA 的单倍群计算得到的，而在(b)中的分界线是用汉族 19 个人群的 36 个线粒体 DNA 的单倍群计算得到的。在(c)中的分界线是用汉族和非汉族 143 个人群的 9 个 Y 染色体的单倍群计算得到的，而在(d)中的分界线是用汉族 35 个人群的 9 个 Y 染色体的单倍群计算得到的。每个边缘线条的粗细和它的自展值成正比。连续线条说明自展值大于 80%，而间断线条说明自展值小于 80%。

的单倍群在整个中国区域的水平上是随机分布的。然而，当把多个线粒体 DNA 单倍群的母系空间模式同时整合起来后，笔者发现在整个中国区域的水平上并不存在空间自相关性，因此表明在整个地理区域上一些母系遗传亚结构的分化是随机分布的。当只包括汉族人群后，除了 D＊、N＊、F1a 和 M7b＊单倍群外，绝大多数线粒体 DNA 单倍群都不

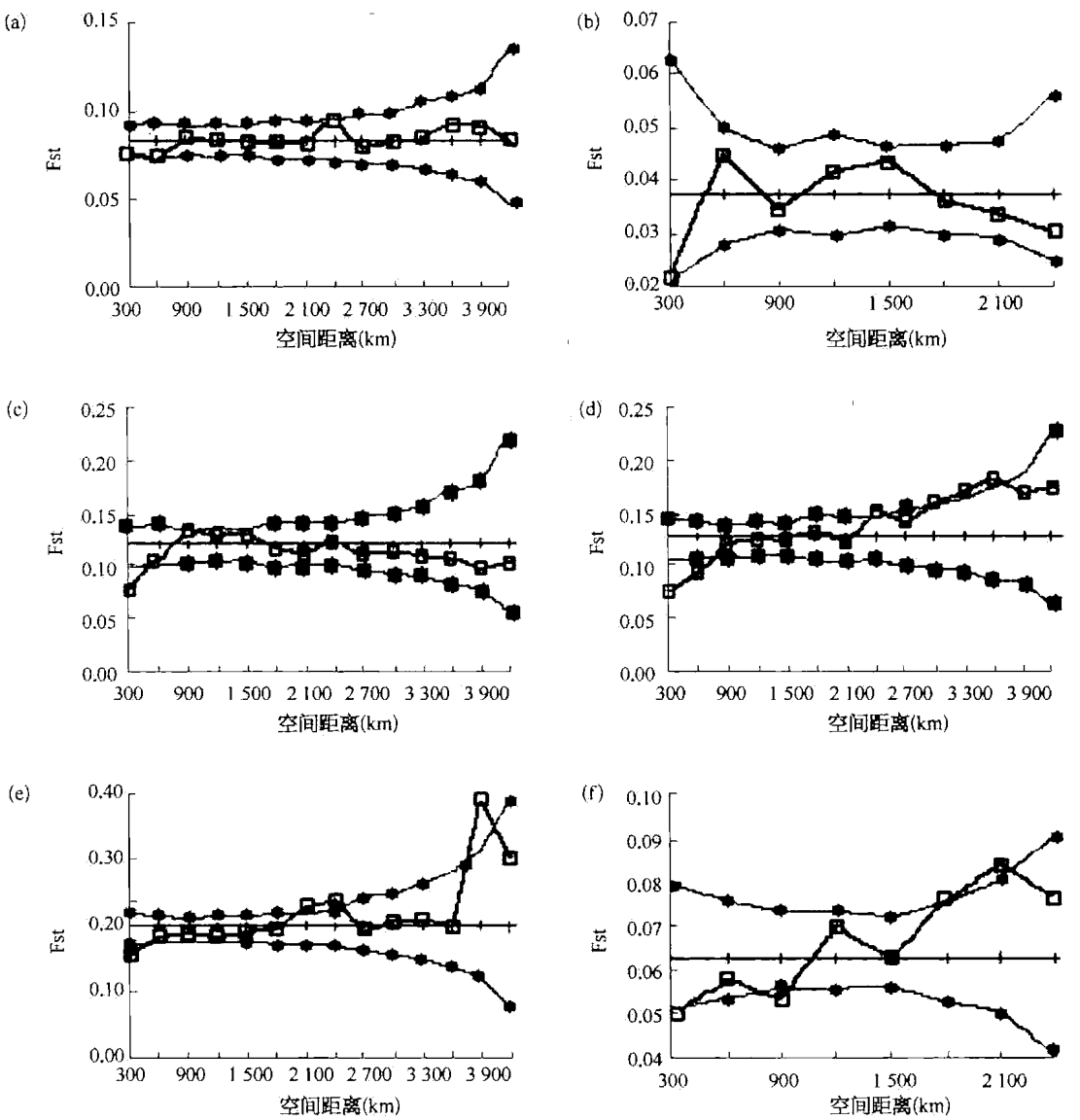


图 2-18 基于线粒体 DNA 和 Y 染色体单倍群频率计算后得到的 14 个空间距离类群的平均 F_{ST} 的遗传距离直方图

(a) 由汉族和非汉族 91 个群体的 36 个线粒体 DNA 单倍群计算得到遗传距离直方图;(b) 由汉族 19 个群体的 36 个线粒体 DNA 单倍群计算得到遗传距离直方图;(c)和(d)分别是用汉族和非汉族 91 个人群的线粒体 DNA 计算得到的 10 个北方主流的单倍群和 23 个南方主流的单倍群;(e) 由汉族和非汉族 143 个人群中的 9 个 Y 染色体单倍群计算得到的遗传距离直方图;(f) 由汉族 35 个人群的 9 个 Y 染色体单倍群计算得到的遗传距离直方图。线条包括了 1 000 次模拟的 95%的置信区间(空心方块:观测值;加号:参照/均值;实心方块:95%的置信区间上限或下限)。

存在空间自相关性,显示了整个中国从北方到南方汉族人群中绝大多数的线粒体 DNA 单倍群是随机分布的。

在父系上,当把所有的汉族和非汉族人群都包括在内时,除 O3 * 单倍群外,绝大多数

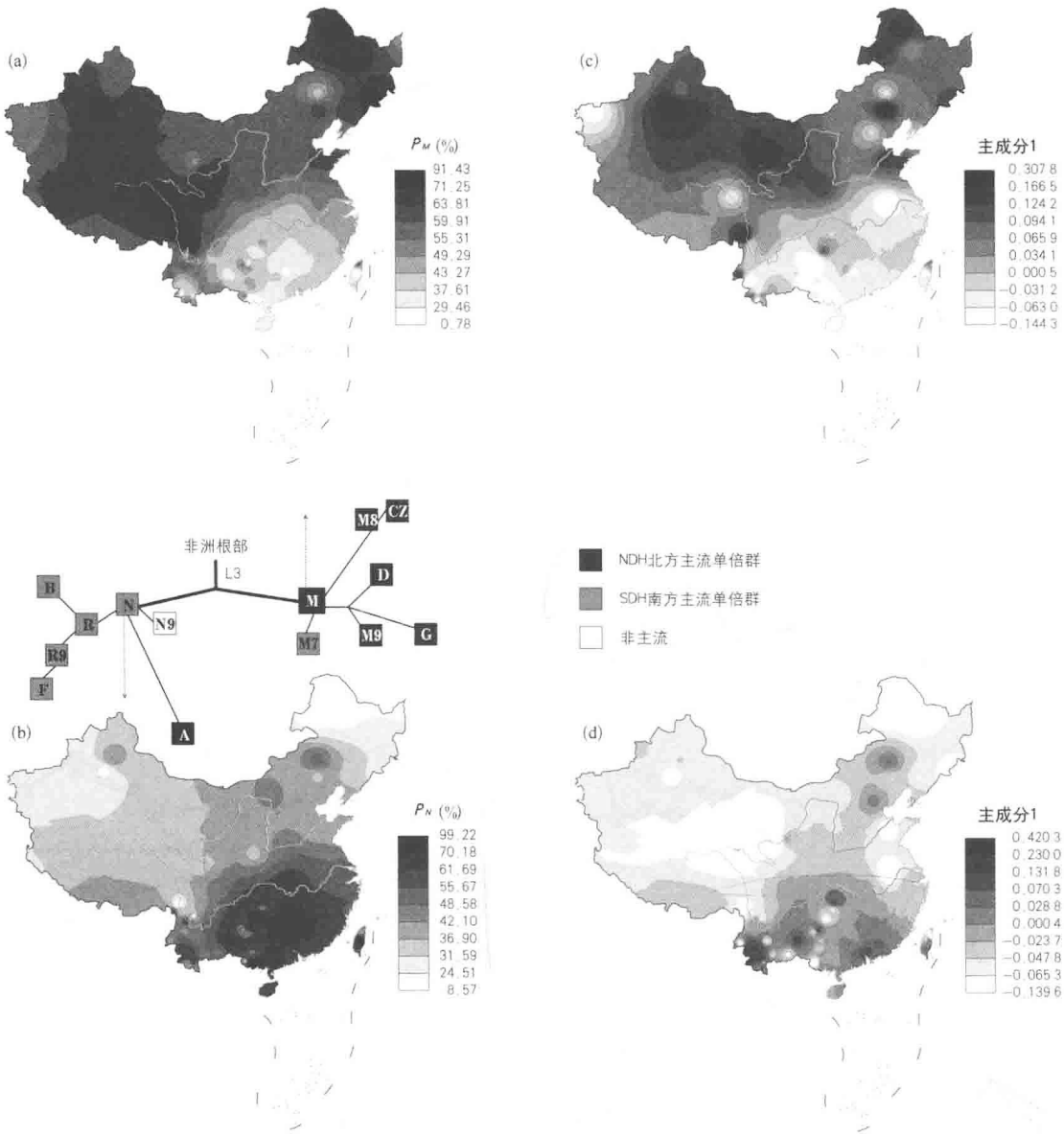


图 2－19　主流线粒体 DNA 单倍群的频率地图

(a) 单倍群 M(包括了 16 个亚单倍群)的频率地图,(b) 单倍群 N(包括了 20 个亚单倍群)的频率地图。每个单倍群地图是根据它在 91 个群体中的频率制作的。进化树由单倍群 L3 定根,树有两个主要分支(M 和 N)。(c)和(d)分别是在 91 个人群中的 10 个北方主流单倍群的线粒体 DNA 的主成分 1(30.33%)合成地图和 23 个南方主流单倍群的线粒体 DNA 的主成分 1(22.50%)合成地图。

Y染色体单倍群(C、D、F*、K*、O3e、O1、O2a和P*)都显现了空间自相关性,显示了中国从北方到南方汉族人群中绝大多数Y染色体单倍群是非随机分布的。当只包括汉族人群后,又可以发现除了D和P*单倍群以外,绝大多数Y染色体单倍群存在空间自相关性。

4. 母系的空间遗传分布

Kivissild等构建了东亚线粒体DNA进化树的主干部分[21]。他们的结果证实了东亚的线粒体DNA谱系是有区域特异性的,而且谱系完全由两个超单倍群M和N覆盖。这个基于全线粒体DNA序列的系统分类结果证实了现存的基于RFLP的东亚线粒体类型的分类结果,也支持了北方和南方人群之间的差别。

图2-19分别显示了单倍群M(包括16个亚单倍群)和单倍群N(包括20个亚单倍群)的频率分布地图。每个单倍群的地图是根据它在91个人群里的频率创建的(图2-19a、b)。笔者用单倍群L3给系统进化树定根,该树有两个主要的枝干(M和N)。地图显示了在中国线粒体DNA单倍群的分布呈现了一个独特的南北之间的分化。单倍群M的频率在北方比较高,包括了北方的汉族、北方的阿尔泰语系和北方的藏缅语族人群(图2-19a),而单倍群N的频率在南方比较高,包括了南方的汉族,南方的侗傣语系、苗瑶语系、南亚语系、南岛语系和南方的藏缅语族人群(图2-19a)。使用图2-17b中的分界线I,根据每个单倍群的频率,可以把绝大多数的单倍群要么分到南方主流的单倍群(南方主流的单倍群包括R、B、R9、F和M7),要么分到北方主流的单倍群(北方主流的单倍群包括A、N、M9、D、G、M8和CZ),虽然有几个单倍群无法被分到这两个主流单倍群中。

表2-3显示了北方和南方主流的线粒体DNA单倍群的分布情况。单倍群A、C、D*、D5、D5a、G、M7c、M8a、M9、N*和Z是北方主流的单倍群,在北方比南方有显著高的频率。单倍群M*、B*、B4、B4a、B4b1、B5*、B5a、F*、F1a、F1b、F1c、F2a、M7*、M7a、M7b*、M7b1、M7b2、R*、R9a、R9b和R9c则是南方主流的单倍群,在南方比北方有显著高的频率。除了M7以外,绝大多数源自M谱系中的单倍群都是北方主流的单倍群;而除了N9和A以外,绝大多数源自N谱系中的单倍群都是南方主流的单倍群。

表2-3 线粒体DNA的北方和南方主流单倍群的分布情况

单倍群	在南方的频率		在北方的频率		Fisher精确检验(*P*值)	主流类型
	个数	频率(95%置信区间,%)	个数	频率(95%置信区间,%)		
A	128	5.48 (4.56~6.41)	90	8.17 (6.56~9.79)	0.003 4	北方
C	90	3.86 (3.07~4.64)	72	6.54 (5.08~8.00)	0.000 7	北方
D*	284	12.17 (10.84~13.49)	225	20.44 (18.05~22.82)	0.000 0	北方
D5	59	2.53 (1.89~3.16)	47	4.27 (3.07~5.46)	0.007 9	北方
D5a	25	1.07 (0.65~1.49)	25	2.27 (1.39~3.15)	0.008 8	北方

（续表）

单倍群	在南方的频率		在北方的频率		Fisher 精确检验（P 值）	主流类型
	个数	频率（95% 置信区间，%）	个数	频率（95% 置信区间，%）		
G	58	2.49（1.85～3.12）	65	5.90（4.51～7.30）	0.000 0	北方
M7c	25	1.07（0.65～1.49）	23	2.09（1.33～3.12）	0.027 9	北方
M8a	153	6.56（5.55～7.56）	137	12.44（10.49～14.39）	0.000 0	北方
M9	21	0.90（0.52～1.28）	25	2.27（1.39～3.15）	0.002 1	北方
N*	28	1.20（0.76～1.64）	39	3.54（2.45～4.63）	0.000 0	北方
Z	26	1.11（0.69～1.54）	39	3.54（2.45～4.63）	0.000 0	北方
M*	240	10.28（9.05～11.51）	84	7.63（6.06～9.19）	0.012 4	南方
B4a	153	6.56（5.55～7.56）	23	2.09（1.33～3.12）	0.000 0	南方
B4b1	62	2.66（2.00～3.31）	11	1.00（0.50～1.78）	0.001 4	南方
B5a	142	6.08（5.11～7.05）	9	0.82（0.37～1.55）	0.000 0	南方
F*	80	3.43（2.69～4.17）	19	1.73（1.04～2.68）	0.004 4	南方
F1a	250	10.71（9.46～11.97）	40	3.63（2.61～4.91）	0.000 0	南方
M7*	24	1.03（0.62～1.44）	3	0.27（0.06～0.79）	0.021 0	南方
M7b*	85	3.64（2.88～4.40）	12	1.09（0.56～1.90）	0.000 0	南方
M7b1	101	4.33（3.50～5.15）	25	2.27（1.47～3.33）	0.002 5	南方
R9a	43	1.84（1.30～2.39）	4	0.36（0.01～0.72）	0.000 2	南方
B*	16	0.69（0.35～1.02）	2	0.18（0.02～0.65）	0.074 0	南方
B4	94	4.03（3.23～4.83）	31	2.82（1.92～3.97）	0.079 4	南方
B5*	10	0.43（0.16～0.69）	3	0.27（0.06～0.79）	0.568 4	南方
B5b	15	0.64（0.32～0.97）	16	1.45（0.83～2.35）	0.051 3	
F1b	54	2.31（1.70～2.93）	25	2.27（1.47～3.33）	1.000 0	南方
F1c	18	0.77（0.42～1.13）	11	1.00（0.50～1.78）	0.549 6	南方
F2a	30	1.29（0.83～1.74）	11	1.00（0.50～1.78）	0.613 7	南方
M7a	3	0.13（0.03～0.38）	1	0.09（00.00～0.51）	1.000 0	南方
M7b2	9	0.39（0.13～0.64）	9	0.82（0.37～1.55）	0.127 4	南方
N9a	50	2.14（1.55～2.73）	29	2.63（1.77～3.76）	0.393 6	
R*	20	0.86（0.48～1.23）	5	0.45（0.15～1.06）	0.281 4	南方

（续表）

单倍群	在南方的频率		在北方的频率		Fisher 精确检验（P 值）	主流类型
	个数	频率（95% 置信区间，%）	个数	频率（95% 置信区间，%）		
R9c	9	0.39（0.13～0.64）	0	0.00（0.00～0.27）	0.065 6	南方
Y	9	0.39（0.13～0.64）	8	0.73（0.31～1.43）	0.197 8	

注：统计学显著性的定义为 $P<0.05$。虽然单倍群 B*、B4、B5*、F1b、F1c、F2a、M7a、M7b2、R* 和 R9c 在南方和北方之间在统计学上并不显著，但是它们实际上在南方的语言人群中占主流地位。因此，它们被识别为南方主流的单倍群。

为了进一步研究北方和南方主流的单倍群的空间遗传结构，我们分别制作了它们的合成地图（图 2-19c、d）和它们的距离直方图（图 2-18c、d）。图 2-19c、d 分别是在 91 个人群中的 10 个北方主流的单倍群的主成分 1（30.33%）的线粒体 DNA 合成地图，和 23 个南方主流的单倍群的主成分 1（22.50%）的线粒体 DNA 合成地图。图 2-18c、d 是 91 个人群中用 10 个北方主流的单倍群和 23 个南方主流的单倍群计算的线粒体 DNA 的距离直方图。在北方和南方主流的单倍群地图中，北方的遗传结构和南方的遗传结构是不同的，北方和南方人群的差异是非常明显的，特别是北方汉族和南方汉族之间（图 2-19c、d）。当分别用 10 个北方主流的单倍群（图 2-18c）和 23 个南方主流的单倍群（图 2-18d）计算距离直方图，可以发现显著的空间自相关性。对于北方主流的单倍群，在第一个距离类群中（到 300 km 为止）的遗传距离要显著低于随机预测的结果，而 900 km 的距离类群中的遗传距离要显著高于随机预测的结果。对于南方主流的单倍群，遗传距离随着地理距离的增加而增加。在第一和第二个距离类群中（到 600 km 为止）的遗传距离要显著低于随机预测的结果，而 2 400～3 600 km 的距离类群中的遗传距离要显著高于随机预测的结果。这表明在北方和南方有着显著的母系的空间自相关性。

5. 父系的空间遗传分布

东亚人群的 Y 染色体单倍群的系统进化树采用了笔者实验室[1]的研究结果。我们使用了 M168 单倍群为该进化树定根，该树可以分为 3 个主要的枝干（M89、M1 和 M130）。图 2-20 显示了 Y 染色体系统发生树主要枝干 M89 和单倍群 K* 的频率地图，使用了 143 个人群中这两个单倍群的频率来建立它们的频率地图。正如所预料的那样，Y 染色体单倍群的空间分布模式与线粒体 DNA 单倍群的空间分布模式是非常不同的。

主要枝干 M89 在所有采样的人群中都是普遍存在的，而且没有北方和南方的差异（图 2-20a）。此外，M89 的亚分支 M9 及其 M9 的子分支群（M95、M119 和 M122）在除了阿尔泰语系人群和西藏人群外所有的人群中都是普遍存在的，而且在绝大多数人群中频率还非常高。单倍群 K* 的分布划分出了一直延伸到云南的藏缅走廊（图 2-20b）。因此，绝大多数 Y 染色体单倍群无法被归类为北方主流群或是南方主流群。然而，在不同语系的人群中，我们确实可以观测到单倍群频率的差异，这显示语言分类和 Y 染色体单倍群之间有着相关性。例如，单倍群 O3*、O3e 和 K* 在除了南岛语系外的绝大多数

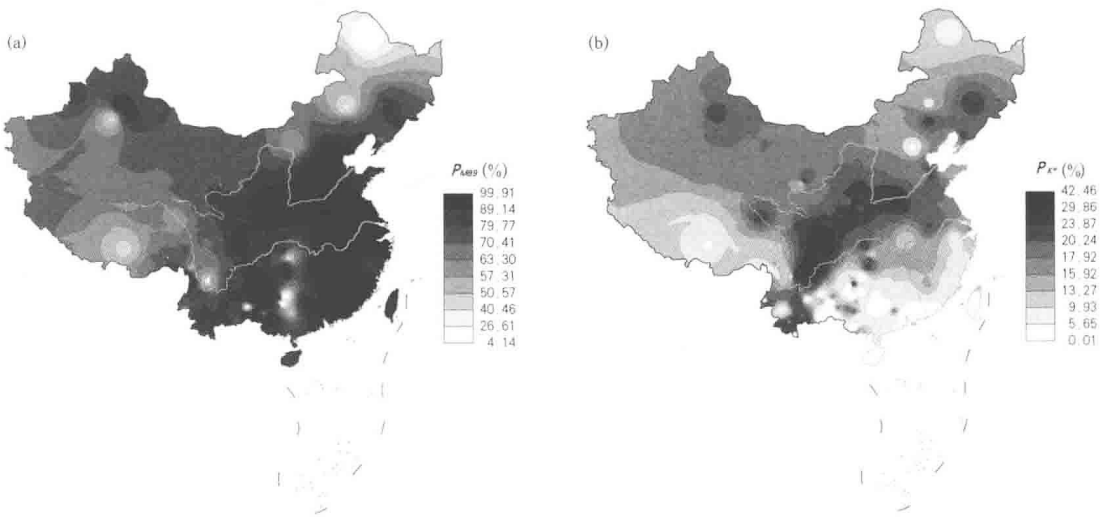


图 2-20　143 个汉族和非汉族人群中的两个 Y 染色体单倍群的频率地图

(a) 包含 7 个单倍群的主要枝于 M89 的地图;(b) 单倍群 K* 的地图

人群中都有着很高的频率,而单倍群 C、P* 和 F* 在阿尔泰语系人群中,单倍群 O1 在南岛语系人群中,单倍群 O2a 在侗傣语系人群中,单倍群 D 在藏缅语族人群中,都分别占主流地位(表 2-4)。

表 2-4　在不同人群中 Y 染色体单倍群的频率

单　倍　群	阿尔泰语系 (%)	南亚语系 (%)	南岛语系 (%)	侗傣语系 (%)	苗瑶语系 (%)	汉族 (%)	藏缅语族 (%)
C	25.36	9.33	0.00	5.12	5.21	7.30	6.60
D	6.30	0.00	0.00	4.12	3.87	1.81	12.69
F*	11.75	3.30	0.00	3.50	0.89	5.98	5.09
K*	17.19	15.94	0.95	8.24	10.52	16.24	19.25
O3*	10.89	20.33	5.21	10.49	35.86	30.66	21.74
O3e	12.18	23.63	1.42	12.36	17.56	23.81	22.16
O1	0.21	0.00	82.94	9.12	5.70	6.91	3.53
O2a	0.29	27.84	9.48	46.82	20.39	3.51	7.65
P*	14.04	0.00	0.00	0.25	0.00	3.79	1.28
样本大小(人)	698	182	211	801	672	1 823	1 403

为了识别在北方人群和南方人群中父系的遗传均一性,笔者使用了汉族人群线粒体 DNA 的分界线 I 把汉族人群的 Y 染色体分为北方群和南方群(图 2-17b),然后测试出

每个Y染色体单倍群在北方群和南方群之间的差异。表2-5显示了Y染色体单倍群在北方群和南方群中的分布。该表指明了北方群和南方群共享相似的Y染色体单倍群的频率，这些相似度的主要特征是在几乎所有的汉族人群中都普遍携带M89、O3*、O3e和K*突变($P>0.05$)。单倍群C和D的频率分布在绝大多数汉族人群中并不普遍，在南方汉族和北方汉族中的分布也没有显著性的差异($P>0.05$)。虽然在北方汉族和南方汉族中单倍群F*、O1、O2a和P*的差异是显著的($P<0.05$)，但是除了O1外它们在汉族人群中是不常见的，O1在南方汉族中的频率比较高(14.09%，95%置信区间为11.40%～17.15%)。因此，父系不同于母系，绝大多数Y染色体单倍群无法被归类为北方主流群或是南方主流群。

表2-5 在汉族北方群和南方群中Y染色体单倍群的频率分布

单倍群	在南方的频率		在北方的频率		Fisher精确检验(P值)
	个数	频率(95%置信区间，%)	个数	频率(95%置信区间，%)	
F(M89)	548	91.95 (89.46～94.00)	879	89.24 (87.13～91.11)	0.080 7
C (M130)	40	6.71 (4.84～9.02)	84	8.53 (6.86～10.45)	0.210 2
D (M1)	8	1.34 (0.58～2.63)	22	2.23 (1.40～3.36)	0.255 4
F*	8	1.34 (0.58～2.63)	75	7.61 (6.04～9.45)	0.000 0
K*	93	15.60 (12.78～18.77)	174	17.67 (15.33～20.19)	0.299 6
O3*	166	27.58 (24.28～31.64)	291	29.54 (26.71～32.50)	0.492 4
O3e	144	24.16 (20.77～27.80)	252	25.58 (22.88～28.43)	0.549 5
O1 (M119)	84	14.09 (11.40～17.15)	31	3.15 (2.15～4.44)	0.000 0
O2a (M95)	45	7.55 (9.56～9.97)	14	1.42 (0.77～2.37)	0.000 0
P* (M45)	8	1.34 (0.58～2.63)	42	4.26 (3.09～5.72)	0.001 0
总 数	596		985		

注：显著性水平定义为$P<0.05$。

2.5.4 讨论

1. 空间遗传结构：母系对比父系

对于母系：① 存在一个独特的北-南地理遗传梯度(图2-16a)；② 北方和南方人群之间有显著的遗传分化(图2-16a)；③ 在北方和南方人群之间存在着一个可以辨认出来的分界线(图2-17b)。必须注意到的是：这个在南北之间可辨认的分界线只有把非汉族人群去除以后才会显现(图2-17b)。当把所有的人群都放在一起分析后，这些分界线就只分布于中国少数民族生活的边缘地区，而且这些分界线并不明显(图2-17a)。最明显的分界线大致是长江以北的淮河-秦岭一线(图2-17b)，这个结果和肖春杰等[5]用经

典标记得到的结果不一致。

为了标示出线粒体DNA单倍群的地理分布，笔者使用了91个人群中36个单倍群的空间数据来制作它们的频率地图，并用了东亚线粒体DNA进化树的主干部分作为指导。分支M主要分布在北方，而分支N主要分布在南方，其中也有一些很重要的例外(图2-19a、b)。在源自M的5个重要的谱系中，有4个(M8、M9、G和D)主要分布在北方，而M7及其他的亚分支主要分布在侗傣人群中(图2-19)。侗傣人群是位于东南亚的一个南方本土的人群，东南亚是现代人类进入东亚的入口[1,6,9,10,28]。在源自N的3个重要谱系中，分支R及其他的亚分支主要分布在北方，但分支A则主要分布在北方的藏缅人群中，而分支N9则分布在整个东亚地区(图2-19，表2-3)。

根据每个线粒体单倍群的频率分布，它可以被归类为南方主流群或是北方主流群。在使用南方主流或是北方主流的单倍群得到的空间遗传结构中，南方和北方的遗传结构是非常不同的，特别是在汉族北方人群和南方人群之间区别更是显著(图2-18c、d，图2-19c、d)。在汉族人群中绝大多数的线粒体DNA单倍群是随机分布的，而且在汉族人群中有着遗传分化的一些母系亚结构也是随机分布的(图2-18b)。

父系的空间遗传结构与母系的结果相比，揭示了一个完全不同的模式。与母系不同，在父系上北方人群和南方人群之间并不存在显著的遗传分化，即使只分析汉族人群，这个结论也是成立的(表2-5)。当我们把所有的人群都囊括起来分析，在那些汉族和少数民族邻接的边缘地区中可以观测到显著而不间断的分界线。当我们只分析汉族人群，在北方汉族和南方汉族之间就缺乏显著而不间断的父系遗传分界线(图2-17c、d)；汉族人群之间绝大多数Y染色体单倍群有着显著的空间自相关性；在中国从北到南，存在着一个显著而广泛的父系空间自相关性(图2-18f)。

在过去的2 000年中，主要的人口移动都是向中国南方的移动[8,29-32]。文波等[7]显示了这种人口移动是有性别偏向性的，移动人口中男性比女性多得多。因此，这些有性别偏向性的基因流，对现存人群的遗传结构有着巨大的影响，正如本研究显示的那样，这些基因流导致了人群中母系和父系之间不同的结构分化。

2. 遗传分界线的空间模式：汉族对比全体民族

对母系和父系来说，只有在仅分析汉族人群时，北方人群和南方人群的遗传分界线才会出现(图2-17b、d)。这表明空间遗传分界线的模式与研究尺度息息相关。汉族人群和中国西南人群之间的F_{ST}值要比汉族之间高。当把所有的非汉族人群从研究中去除后，南方汉族和北方汉族之间的F_{ST}值就开始变得非常显著了，南方人群和北方人群之间的遗传分界线也显现出来了(图2-17b、d)。这种与尺度相关的效应，也出现在日本人群之间的进化关系研究中，该研究使用了105个短串联重复序列的多态位点[33]。

3. 空间数据库和统计学方法

虽然我们的空间数据库包括了91个中国人群3 435个个体的线粒体DNA数据和143个中国人群5 790个个体的Y染色体数据，就东亚人群的遗传结构复杂性而言，样本点在人群中的分布和密度还远未达到令人满意的程度。而从另一方面来说，本研究中线

粒体 DNA 给出的数据解析度要高于 Y 染色体的解析度(研究中只分析了 9 个广泛分布的 Y 染色体单倍群,而分析了 36 个线粒体 DNA 单倍群)。在母系和父系之间的结果差异可能是有偏差的。本研究的另外一个缺陷是,没有把东亚和东南亚其他重要地区的数据包括进来,可能降低研究结果的精确性,而这种数据缺乏是由于(某些政府或科研机构)缺乏对这些人群进行研究的努力。

很多方法可以用于遗传变量的插值轮廓地图的绘制:如 Genography 中的 Cavalli-sforza 方法[4]、IDW 插值法[14],以及 Kriging 技术等[34]。本研究选择了 IDW 插值法来展示空间遗传模式和探测地理遗传梯度,该算法和其他算法相比,产生的结果类似或者稍微好一点。这个结论基于笔者用自己的数据测试了不同的算法,然后比较了不同的算法产生的结果。

还有几种方法可以用来探测空间遗传分界线,如 Wombling 方法、分子差异空间分析(SAMOVA)和改进的 Monmomier 算法[13,15]。本研究选择了改进的 Monmomier 算法[13,15]来识别空间遗传分界线。该算法避免了 Wombling 方法的地形插值中,潜在的人造连续性或间断性的假象,而和 SAMOVA 相比在寻找空间遗传分界线时表现得稍微好一些。

空间遗传自相关性可以用不同的统计量来探测和发现。最常用的测量值是 Moran 指数和 Geary 指数[25,26]。近来一个新的统计值被称为距离直方图,是基于遗传距离的空间自相关性的多位点测量值,被用来探测空间遗传自相关性,也用来测试空间自相关性的统计显著性[36]。本研究选择空间遗传软件(SGS,版本 1.0d)中的遗传距离直方图来探测空间遗传自相关性,其中 F_{ST} 统计量作为遗传距离的测量值。构建遗传距离直方图至少有两个优势[16]:① 它可以同时描述多个变量的空间模式;② 它使用了遗传距离中已完善的概念来测量不相同的地方。在 SGS 中的遗传距离直方图的另一个优点是:空间遗传自相关性的统计学显著性可以用模拟过程来测试。

参考文献

[1] Jin L, Su B. Natives or immigrants: origin and migrations of modern humans in East Asia. Nat Rev Genet, 2000, 1: 126 - 133.

[2] Chen R, Ye G, Geng Z, et al. Revelations of the origin of Chinese nation from clustering analysis and frequency distribution of HLA polymorphism in major minority nationalities in Mainland China (in Chinese). Yi Chuan Xue Bao, 1993, 205: 389 - 398.

[3] Du R, Xiao C J, Cavalli-Sforza L L. Genetic distances between Chinese groups calculated on gene frequencies of 38 loci. Sci China C Life Sci, 1998, 28: 83 - 89.

[4] Cavalli-Sforza L L, Menozzi P, Piazza A. The history and geography of human genes. Princeton: Princeton University Press, 1994.

[5] Xiao C J, Du R F, Cavalli-Sforza L L, et al. Principal component analysis of gene frequencies of Chinese populations. Sci China C Life Sci, 2000, 43: 472 - 481.

[6] Chu J Y, Huang W, Kuang S Q, et al. Genetic relationship of populations in China. Proc Natl

Acad Sci USA, 1998, 95: 11763 - 11768.

[7] Wen B, Li H, Lu D, et al. Genetic evidence supports demic diffusion of Han culture. Nature, 2004, 431: 302 - 305.

[8] Yao Y G, Nie L, Harpending H, et al. Genetic relationship of Chinese ethnic populations revealed by mtDNA sequence diversity. Am J Phys Anthropol, 2002, 118: 63 - 76.

[9] Yao Y G, Kong Q P, Bandelt H J, et al. Phylogeographic differentiation of mitochondrial DNA in Han Chinese. Am J Hum Genet, 2000, 70: 635 - 651.

[10] Su B, Xiao J, Underhill P, et al. Y chromosome evidence for a northward migration of modern humans into Eastern Asia during the Last Ice Age. Am J Hum Genet, 1999, 65: 1718 - 1724.

[11] Karafet T, Xu L, Du R, et al. Paternal population history of East Asia: sources, patterns, and microevolutionary processes. Am J Hum Genet, 2001, 69: 615 - 628.

[12] Ding Y C, Wooding S, Harpending H C, et al. Population structure and history in East Asia. Proc Natl Acad Sci USA, 2000, 97: 14003 - 14006.

[13] Manni F, Guerard E, Heyer E. Geographic patterns of (genetic, morphology, linguistic) variation: how barriers can be detected by 'Monmonier's algorithm'. Hum Biol, 2004, 76: 173 - 190.

[14] Sokal R R, Thomson A B. Spatial genetic structure of human populations in Japan. Hum Biol, 1998, 70: 1 - 22.

[15] Manni F, Guerard E. Barrier vs. 2. 2. manual of the user: population genetics team. Paris: museum of mankind (Musee de l'Homme), Publication distributed by the authors, 2004.

[16] Degen B, Petit R, Kremer A. SGS-spatial genetic software: a computer program for analysis of spatial genetic and phenotypic structures of individuals and populations. J Hered, 2001, 92: 447 - 449.

[17] Jorde L B, Bamshad M, Rogers A R. Using mitochondrial and nuclear DNA markers to reconstruct human evolution. Bioessays, 1998, 20: 126 - 136.

[18] Wen B, Xie X, Gao S, et al. Analyses of genetic structure of Tibeto-Burman populations reveals sex-biased admixture in southern Tibeto-Burmans. Am J Hum Genet, 2004, 74: 856 - 865.

[19] Wen B, Hong S, Ling R, et al. The origin of Mosuo people as revealed by mtDNA and Y chromosome variation. Sci China C Life Sci, 2004, 47: 1 - 10.

[20] Wen B, Li H, Gao S, et al. Genetic structure of Hmong-Mien speaking populations in East Asia as revealed by mtDNA lineages. Mol Biol Evol, 2005, 22: 725 - 734.

[21] Kivissild T, Tolk H V, Parik J, et al. The emerging limbs and twigs of the East Asian mtDNA tree. Mol Biol Evol, 2002, 19: 1737 - 1751.

[22] Barbujani G. Geographic patterns: how to identify them and why. Hum Biol, 2000, 72: 133 - 153.

[23] Jenks, George F. The data model concept in statistical mapping. International Yearbook of Cartography, 1967, 7: 186 - 190.

[24] Monmonier M S. Maximum-difference barriers: an alternative numerical regionalization method. Geogr Anal, 1973, 3: 245 - 261.

[25] Cliff A D, Ord J K. Spatial autocorrelation. London: Pion Limited, 1973.

[26] Sokal R R, Oden N L. Spatial autocorrelation in biology. 1. methodology. Biol J Linnean Soc, 1978, 10: 199 - 228.

[27] Ned L. CrimeStat III a spatial statistics program for the analysis of crime incident locations (version 3.0). Houston, TX: Ned Levine & Associates/Washington, DC, USA: National Institute of Justice, 2004.

[28] Su B, Xiao C, Deka R, et al. Y chromosome haplotypes reveal prehistorical migrations to the Himalayas. Hum Genet, 2000, 107: 582 - 590.

[29] 杜若甫,叶傅升. 中国的民族. 北京: 科学出版社,1993.

[30] 尤中. 云南民族史. 昆明: 云南大学出版社,1994.

[31] 王钟翰. 中国民族史. 北京: 中国社会科学出版社,1994.

[32] 葛剑雄,吴松弟,曹树基. 中国移民史. 福州: 福建人民出版社,1997.

[33] Li S L, Yamamoto T, Yoshimoto T, et al. Phylogenetic relationship of the populations within and around Japan using 105 short tandem repeat polymorphic loci. Hum Genet, 2006, 118: 695 - 707.

[34] Hoffmann M H, Glass A S, Tomiuk J, et al. Analysis of molecular data of *Arabidopsis thaliana* (L.) Heynh. (Brassicaceae) with geographical information systems (GIS). Mol Ecol, 2003, 12: 1007 - 1019.

[35] Dupanloup I, Schneider S, Excoffier L. A simulated annealing approach to define the genetic structure of populations. Mol Ecol, 2002, 11: 2571 - 2581.

[36] Degen B, Scholz F. Spatial genetic differentiation among populations of European beech (*Fagus sylvatica L.*) in Western Germany as identified by geostatistical analysis. Forest Genet, 1998, 5: 191 - 199.

第3章　汉藏族群的Y染色体

在东亚人种的九大语系中，汉藏语系的人群无疑是东亚影响力最大的主体人群。语言学上对汉藏语系的内部分类争议颇多，甚至连汉语族和藏缅语族的分类都有争议。对于藏缅语族内的语支分类，在国际主流观点上也不断变化着，目前的分类是羌白语支（含羌、氐、纳西、白、土家）、藏语支、山南语支（含勒墨、怒、独龙、珞巴等）、景颇语支、缅彝语支、克伦语支等。但是国内语言学界对此说接受度较低。汉语族内可分秦语支（官话、晋语、粤语、平话）、齐语支（闽南、闽北、闽东、闽中、兴化）、楚语支（湘语、赣语、客家话、徽语、吴语）。虽然语言学界还没得到统一结论，分子人类学对于汉藏族群的起源与迁徙已经通过大量遗传结构调查得到了基本结论。汉藏族群的Y染色体单倍群主要是O3类型，也包括少部分N、D、C、O1、O2、R类型等。

为了分析汉藏主流O3类型的来源，本章描述了对东亚现代人进行系统性抽样分析的研究，调查了来自不同东亚群体的2 332个个体，并对东亚特有的Y染色体单倍群（O3－M122）进行遗传分析。结果表明，单倍群O3－M122在东亚人群中占主流，平均频率为44.3%。微卫星数据显示，东亚南部的O3－M122单倍型多样性高于东亚北部，这表明O3－M122突变起源于南部。据估计，O3－M122支系往东亚北方的早期迁徙发生在2.5万～3万年前，这与东亚的现代人化石记录一致。

有着共同文化和语言的汉族，人口超过了116 000万（2000年人口统计），无疑是全世界最大的民族。因此汉文化的扩散过程广受各领域研究者的关注。语言和文化在人群间的扩散有两种不同的模式：一种是人口扩张、人群迁徙模式；另一种是文化传播模式，人群之间有文化传播，而基因交流却很有限。印欧语系欧洲人群的形成机制争议颇多，争论的焦点在于来自近东的农业文明和语言的扩散是否伴随大量农业人口的迁徙。对于汉族的形成也有着同样的争议。通过系统地对汉族群体的Y染色体和线粒体DNA多态性进行分析，发现汉文化向南扩散的格局符合人口扩张模式，而且在扩张过程中男性占主导地位。

目前，对汉族群体遗传结构的研究已日趋完善和深入，各个分支人群的Y染色体和线粒体DNA的遗传结构都已经有了相关研究和报道，除平话人外的汉族九大支系人群无一例外地都显示了汉族在遗传结构上的高度一致性。平话人作为汉族的一个古老支

系，杂居于侗台、苗瑶等南方少数民族人口占局部优势的广西及其周边地区，其遗传结构的类型问题越来越受到关注。本章对中国广西壮族自治区北部贺州、富川县、罗城县、金秀县和武宣县五个地区的平话汉族人群及其周边部分壮族、侗族、仫佬族、拉伽人和瑶族共470份个体(包括195个男性)分别进行了线粒体DNA和Y染色体单倍群的分型，结果发现桂北平话人群的Y染色体单倍群具有高频的南方原住民族单倍型类型K、O*、O2a*，仅罗城和金秀县平话人保留了汉族常见的O3a5。在线粒体DNA单倍群的分型方面，平话人群高频单倍型是B4a、B5a、M*、F1a、M7b1和N*，与其周边原住民族和南方少数民族尤其是侗台语系高频单倍型很接近。根据平话人群以及汉族其他支系和东亚其他民族群体的Y染色体和线粒体DNA单倍群数据所绘制的树型聚类图、主成分分析图，以及网络结构图分析结果显示，平话人在遗传结构上更靠近南方原住民族，而与汉族距离甚远。通过计算汉族群体和南方原住居民对平话人的遗传贡献率，更加证实了这种差距。因此得出结论：平话人在遗传结构上并非汉族的后裔，他们的遗传成分主要源于当地少数民族，是被汉族在语言文化、自身认同感上同化了的广西原住居民。

藏族是汉藏族群中又一重要群体。在通过Y染色体遗传标记对藏族的起源进行研究时，共有3个人群被纳入检测范围，其中两个群体来自西藏中部的卫藏语区，还有一个来自云南的康语区。在这3个群体中识别出两类主要的父系遗传谱系(>80%)，一类很可能来自东南亚(YAP+)，另一类则来自东亚(M122C)。考虑到事实上中亚罕有M122C类群以及东亚罕有YAP+类群，并且遗传漂变机制不太可能形成这种结果，笔者认为藏族人群的Y染色体很可能来源于两组不同的基因库群体。

在东亚历史上，藏缅群体源于中国西北的古老部族，随后向南部迁徙，并在过去的2 600年中与南方原住居民混合。目前，这一族群广泛分布在中国以及东南亚，他们迁徙到南方以后与原住民族之间发生了什么样的基因流动，是藏缅族群遗传研究的关键问题。在过去几个世纪中，美洲的欧洲裔男性对非洲裔和拉美裔的基因流动是定向的，所以在美洲大陆，不同亲本群体的男性和女性世系对混合群体的不均等贡献并不罕见。然而在东亚，尽管存在着大量的人群迁徙记录，但几乎没发现有性别差异的混合。本章分析了中国20多个藏缅群体965份Y染色体样本和754份线粒体DNA样本的多样性。通过考察Y染色体和线粒体DNA标记的单倍群分布及其主成分，揭示了现存南部藏缅群体的遗传结构主要由两个亲本群体形成：北方移民和南方原住民。在混合中，男性和女性世系是有差异的，在现存南部藏缅群体中北方移民对男性世系影响力更强(约62%)，南方原住民对女性世系做出的贡献更多(约56%)。这是揭示南部藏缅群体混合的性别差异的第一个遗传学证据，在遗传学、历史学和人类学方面都具有影响力。

喜马拉雅山区的东部接近早期人类进入东亚的南方入口。这一地区的遗传结构对于东亚人群起源研究尤为重要。但是这一地区的汉藏群体(珞巴族和僜人)几乎没有遗传学调查。本章分析了这两个群体的Y染色体多样性。珞巴族的单倍群包括D、N、O、J、Q和R，说明存在来自藏族的遗传成分(D、N、O)，还有西亚(J)和北亚(Q、R)的成分。僜人的单倍群包括O、D、N和C，与其东面的大多数汉藏族群相似。汉藏群体主要单倍

群O3内部的短串联重复(STR)多样性显示,珞巴族遗传上接近藏族,而僜人接近羌族。在汉藏群体中,羌族的多样性最高,证实其为该族群中最古老的群体。而多样性最低的群体就在喜马拉雅东部地区,说明这一地区是汉藏人群扩张的一个终点。因此,我们认为单倍群O3进入喜马拉雅东部至少有两条路线:北部的藏区和东部的横断山区。

汉藏族群中社会文化最独特的族群是摩梭人。摩梭人居住在中国云南省西北部与四川省接壤的泸沽湖区域,是中国内地唯一的母系社会群体。摩梭人被四川省政府认定属于蒙古族,被云南省政府认定属于纳西族,而他们与纳西族的关系仍然存在争议。通过测量线粒体DNA的HVS-1区段和Y染色体的位点(包括13个SNP和8个STR位点)的基因型多样性,研究了摩梭人与其他5个生活在云南西北部的族群的关系,包括纳西族、藏族、白族、彝族和普米族。结果显示,摩梭人的母系支系整体频率上与纳西族最相似,但主要世系来自普米族。其父系支系来源众多,但整体频率与云南藏族支系更接近。父系和母系比较明显的差别可能是由遗传历史、母系社会结构和走婚习俗造成的。

3.1　单倍群O3的南方起源

3.1.1　研究背景

Y染色体是重建人类群体历史的有效工具[1]。在晚近人群中,后期的定居、分化和迁徙留下的信息都叠加在一起,使得Y染色体遗传变异的地理分布能够为揭示人群史前迁徙提供线索[2]。位于Y染色体非重组区的双等位基因(SNP,低突变率)和微卫星(STR,高突变率)标记结合在一起,已被广泛地应用于推断早期人群历史[1-6]。已有研究表明,在Y染色体SNP背景下基于Y染色体STR(Y-STR)变异的分歧时间估计,比仅仅依据STR的估计要准确得多[7]。该方法最近刚被用于追溯和测定欧洲人群的史前迁徙[8-10]。

由于存在大量的本土人群和复杂的人类居住历史,东亚的人群迁徙,尤其是早期人群的迁徙模式,仍然是不清楚的。相关研究发现东亚北方和南方人群之间存在遗传差异,而研究者之间的争论主要集中于对造成这种差异的早期迁徙历史的不同解释[11-22]。对于Y染色体变异的研究表明,东亚南方的多态性高于东亚北方[16,20]。但是,Karafet等[22]使用一组类似的Y染色体双等位基因标记研究认为,东亚南方和东亚北方之间并不存在遗传差异,因而反对东亚人群早期由南向北的迁徙假说。来自线粒体DNA的数据在这个问题上也并不一致。两个早期的线粒体DNA研究支持东亚现代人的南方起源[23,24],而另一个研究声称东亚南方和东亚北方之间的遗传差异可能源于地理隔离,因而北方起源仍然是可能的[19]。

应当指出的是,当我们讨论人群起源和迁徙时,应该明确界定所研究的时期,即涉及的具体是人类历史的哪一部分。近期人群流动和混合会掩盖或者很大程度上削弱早期人群迁徙的遗传信号。因此,为了获得现代人起源的信息和了解新石器时代之前的早期人群迁徙,人群特异性标记(如Y染色体上的SNP标记)对于研究区域人群迁徙就显得

十分有用[1]。同时，在基于人群特异性的分析中，远缘人群间的近期基因交流能够被鉴别出并被剔除。因此在这个意义上，研究某大陆人群的起源和早期迁徙时，应该格外谨慎地选择遗传标记。因为之前的两个研究[19,22]已表明，通过近期基因交流引入的基因变异会导致错误的解读。同样的逻辑也适用于人群的选择，在推断早期人群历史时，应该倾向于选择在某一地区长期定居的民族人群。

在东亚人群中，M175 支系(图 3-1)下有 3 个区域局限分布(东亚人特有)的 Y 染色体单倍群，即 O3-M122、O2-M95 和 O1-M119，共占东亚人群 Y 染色体总数的 57%(表 3-1)。O3-M122 在东亚人中的频率最高(平均频率 41.8%)(图 3-2)，尤其是在中国汉族中(北方汉族中为 52.06%，南方汉族中为 53.72%)(表 3-1)，而在东亚之外并未观察到。研究表明，O2-M95 和 O1-M119 在东亚南方普遍存在，而且很可能起源于南方[16,17,25,26](表 3-1)。因此，追寻 O3-M122 的起源成了完整了解现代东亚人起源和早期迁徙的关键。

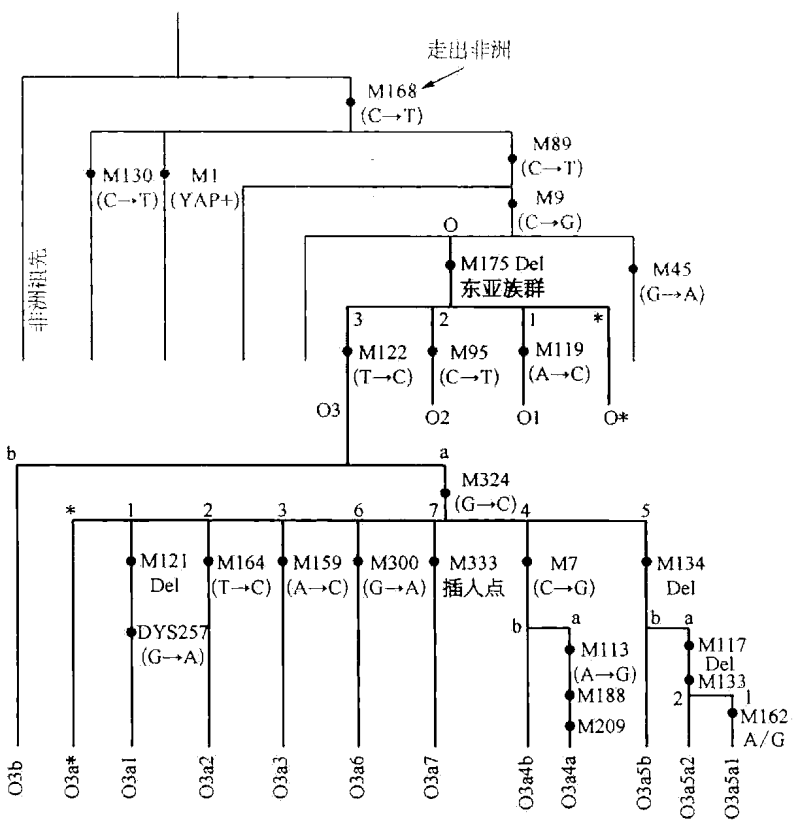


图 3-1 O3-M122 单倍群内 SNP 位点和单倍型的谱系关系

在本节中，通过对东亚人群的广泛抽样，并对东亚特有 Y 染色体单倍群 O3-M122 的 Y 染色体标记(SNP 和 STR)进行基因分型，我们试图阐释清楚单倍群 O3-M122 的起源，进而揭示现代人在东亚的史前迁徙过程。

表3-1　东亚特有Y染色体单倍群在世界范围人群内的分布

人　群	样本量	各人群相关单倍型频率(%)			
		M122	M119	M95	总和
非洲	329				
美洲印第安	221				
欧洲	1 046				
高加索	147				
俄罗斯	243				
西伯利亚	328		3.05		3.05
中亚	1 231	4.55	0.57	0.61	5.73
印度南部	259	0.39		1.16	1.55
东亚北部					34.65
阿尔泰	561	15.15	1.43	1.60	18.18
朝鲜	81	38.27	2.47		40.74
日本	29	27.59	3.45		31.04
北方汉族	413	52.06	4.36	0.72	57.14
回族	74	22.97	1.35		24.32
藏族	129	36.43	0.78		37.21
东亚南部					72.31
南方汉族	1 102	53.72	15.34	6.17	75.23
藏缅	293	48.81	4.20	6.74	59.75
侗傣	178	29.78	10.67	40.45	80.90
苗瑶	249	51.41	2.41	16.10	69.92
南亚	63	11.11	3.18	50.80	65.09
南岛	555	26.31	22.34	16.94	65.59
美拉尼西亚	113	2.65			2.65

注：单倍型频率数据引自发表文献[16-18,22,25,27-31]。

3.1.2　材料和方法

1. 样本

本研究中，样本来自东亚40个人群2 332个无关男性。人群选择的标准依据O3-M122单倍群的分布。大多数人群样本来自中国南部和西南部，那里生活着约80%的中国民族群体，而且大多数定居历史在3 000年以上[33]。中国北方的大多数民族(如回族、

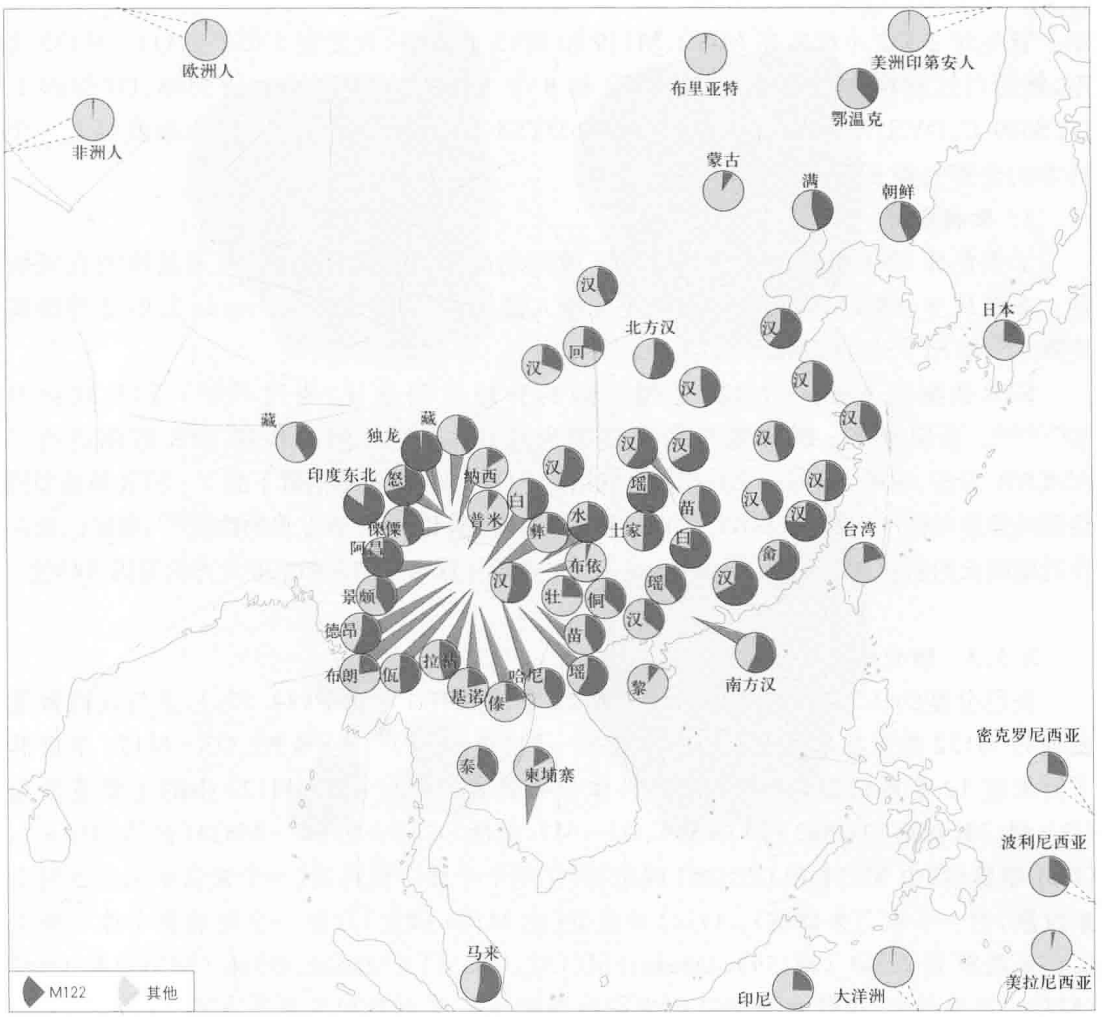


图 3-2 O3-M122 单倍群在东亚和其他大陆人群中的频率分布
(数据引自发表文献[16-18,22,25,27,28,30-32])

维吾尔族和蒙古族)是较晚近时期形成的(晚于 1 000 年前),而且与高加索和中亚人群存在大量混合[33]。因此本研究不涉及这些人群。所研究的人群依据他们的地理分布分别归类为东亚南部人群和东亚北部人群,长江作为划分东亚南方和东亚北方的地理界线。在东亚南方中,有 14 个藏缅语人群有历史记载,表明其在约 3 000 年前从中国北方迁徙而来[33]。云南(中国西南)的 3 个阿尔泰语人群源于近 1 000 年以内的中国北方迁徙[33]。其他人群的数据引自已发表文献[16-18,22,25,27,28,30-32]。

2. Y 染色体标记和基因分型

本研究检测了 15 个 Y 染色体 SNP 标记,包括 M122、M119 和 M95 三个东亚特有的主要支系,以及 M122 下游的衍生 SNP:M134、M117、M162、M7、M113、M159、M121、M164[2,5]、M324、M300、M333 和 DYS257[34]。单倍型命名原则遵循 Y 染色体委员会的

标准[5]。基因分型方法包括PCR-RFLP和测序[18]。Y-SNP的谱系关系如图3-1所示。首先对2 332个样本在M122、M119和M95上分型，共发现1 032个O3-M122类型，然后对这些样本在其他12个SNP和8个STR(DYS19/394、DYS388、DYS389 I、DYS389 II、DYS390、DYS391、DYS392和DYS393)上进一步分型。其中共获得854个样本的全套完整SNP和STR信息。

3. 数据分析

Y染色体SNP数据用来分析地理区域间的标示人群频率分布，以等高线图直观表示。多维尺度分析(MDS)用SNP数据分析人群关系，借助软件Arlequin 2.0(遗传距离矩阵)[35]和SPSS11.0(MDS图)实现。

STR数据用于O3-M122亚型之间的分歧时间估计，使用SNP-STR联合方法[9,10,36]。所用的Y-STR的平均突变率为0.000 69[37]。同时也用STR数据进行了AMOVA分析，使用Arlequin 2.0剖析人群结构。O3-M122亚单倍群下的Y-STR单倍型网络图构建使用程序NETWORK4.1.0.6(Fluxus)。根据Karafet等发表的数据[22]，剔除已知存在近期混合的东亚北方人群，用Arlequin 2.0[35]重新计算东亚南方和东亚北方的基因多样性。

3.1.3 研究结果

在已分型的2 332个样本中，O3-M122个体共有1 032个(44.3%)，这与以往报道发现的M122类型在东亚人群中占主流相一致(表3-1)[16-18,22]。在O3-M122单倍群下游发现12个单倍型，其中7个的具体分布见表3-2。O3-M122中的主要亚型为O3-M134(包括O3a5a2和O3a5b)、O3-M7(包括O3a4b)和O3-M324(包括O3a*)。O3a1单倍型(由M121和DYS257确定)只在两个个体中观察到(一个来自中国北方的山东汉族，另一个来自柬埔寨)。O3a2单倍型(由M164确定)只在一个柬埔寨个体中观察到。未观察到O3a3(M159)、O3a4a(M113)、O3a5a1(M162)、O3a6(M300)和O3a7(M333)等单倍型，这些类型在其他报道中最初用东亚南方和东亚北方筛选板发现于东亚南方样本中[28,38]。等高线图(图3-3)显示了O3-M122主要的亚单倍群在东亚的地理分布。总的来说，O3-M122单倍型的分布在南北人群间并没有表现出明显隔离，所有的主要亚单倍群，除了O3-M7(这一支只在南方人群中发现)，在南北方均有分布，这也暗示着O3-M122谱系在东亚有近期的共同祖先起源。用STR数据计算基因多样性，在东亚南方和东亚北方或者其他不同语言人群间，并未发现显著差异。分子方差分析(AMOVA)在人群间也未发现显著差异。但是，MDS显示，东亚北方群体内部较紧密地聚类在一起，而东亚南方群体内部关系相对松散，方差较大，这表明东亚南方相比于东亚北方多样性更高(图3-4)。值得注意的是，东亚北方和东亚南方间的遗传方差差异也可能源于抽样密度差异。之前的研究表明，北方汉族群体相对更加同质，Y染色体单倍型分布较相似[39,40]。而且，本研究中抽样的4个北方汉族群体涵盖了中国北方的不同地理区域。因此，观察到的遗传方差大致能够反映中国北方人群的真正遗传背景。在MDS图中，苗瑶群体和汉族群体聚在一起，这和有记载的人群混合历史是吻合的[33]。

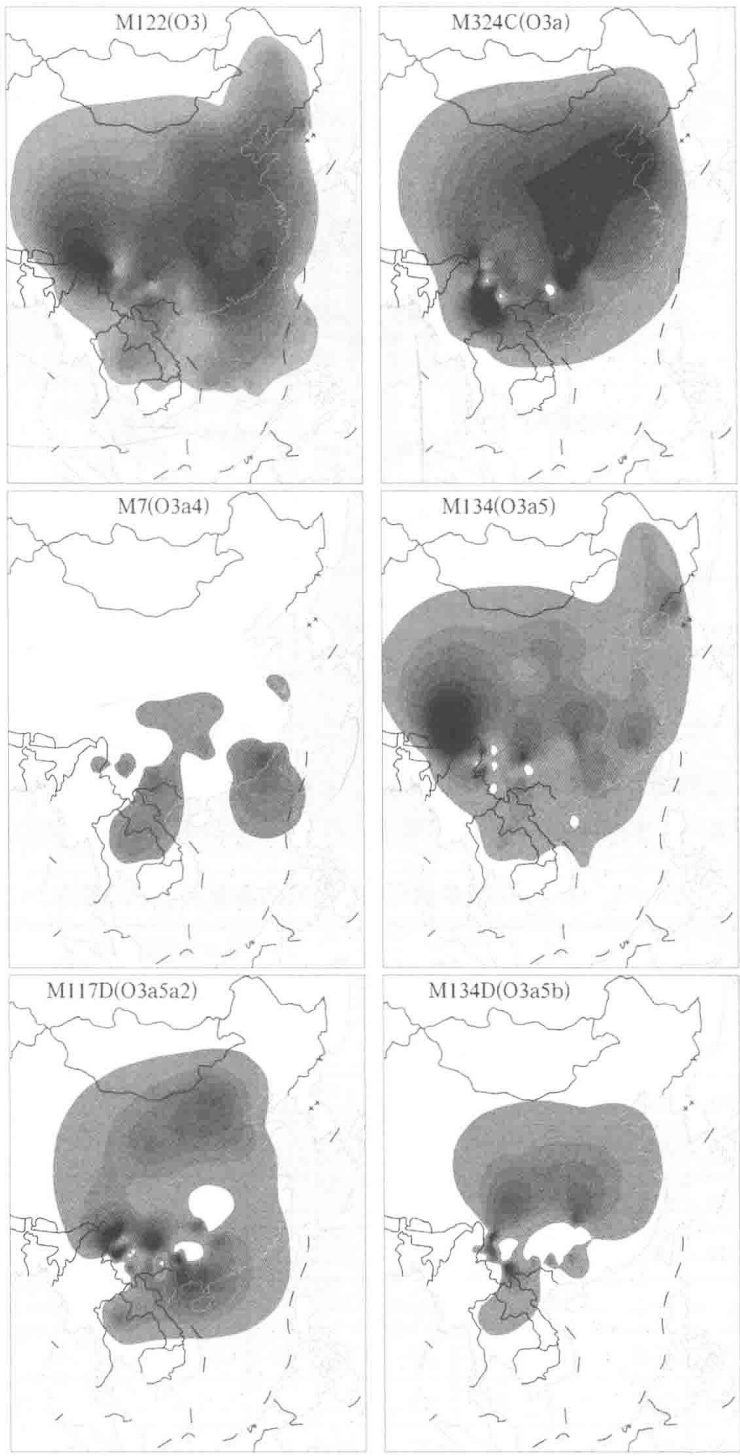


图 3-3 Y 染色体单倍型频率分布等高线图

构图所用的单倍型 O3-M122、O3a-M324、O3a5-M134 和 O3a4-M7 数据源自发表文献[16-18,25,32]和本研究，单倍型 O3a* 和 O3a5a2(M117)的数据源自本研究。

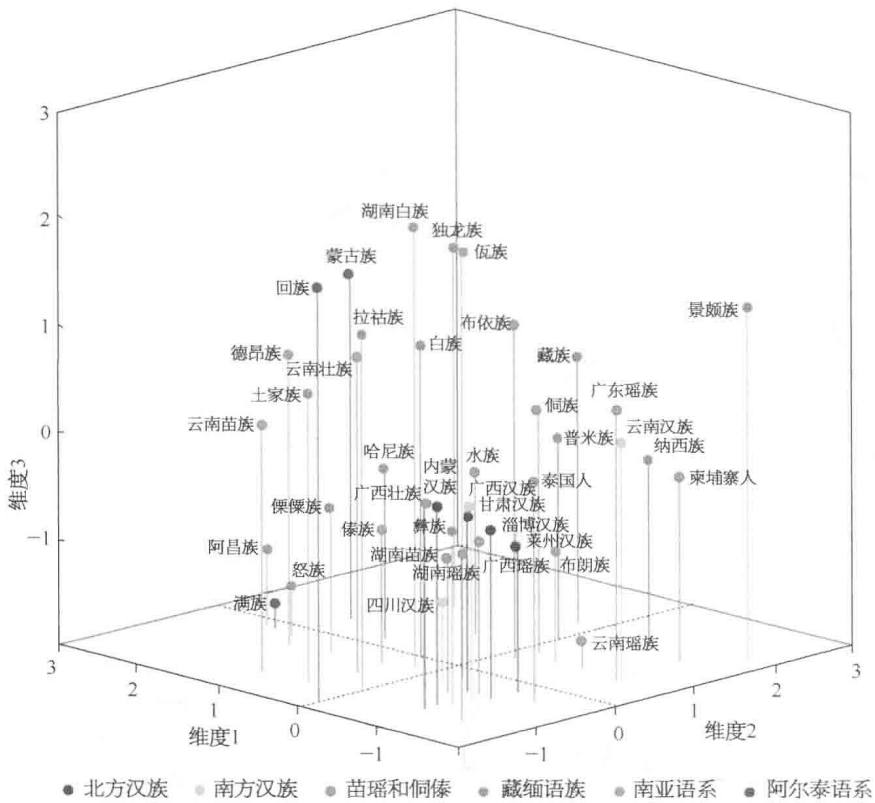


图 3－4　多维尺度分析图基于 40 个人群 O3－M122 的 SNP 单倍型分布

表 3－2　O3－M122 单倍型在所研究的东亚人群中的分布

人　群	标示	语言	样本量	单倍型样本数量							
				M122	O3b	O3a *	O3a1	O3a2	O3a4b	O3a5b	O3a5a2
东亚北部											
内蒙古汉族	H1	汉语	60	29	1	11				2	10
甘肃汉族	H2	汉语	60	22		8				3	5
莱州汉族	H3	汉语	86	49	1	31				5	2
淄博汉族	H4	汉语	98	61	2	18	1			6	8
东亚南部											
四川汉族	H5	汉语	64	38	3	10			2	13	3
广西汉族	H6	汉语	39	15		4				1	3
云南汉族	H7	汉语	81	37	1	16			3	1	16
阿昌族	T1	藏缅语族	40	33		30			3		
白族	T2	藏缅语族	80	38	2	12			11	3	10

（续表）

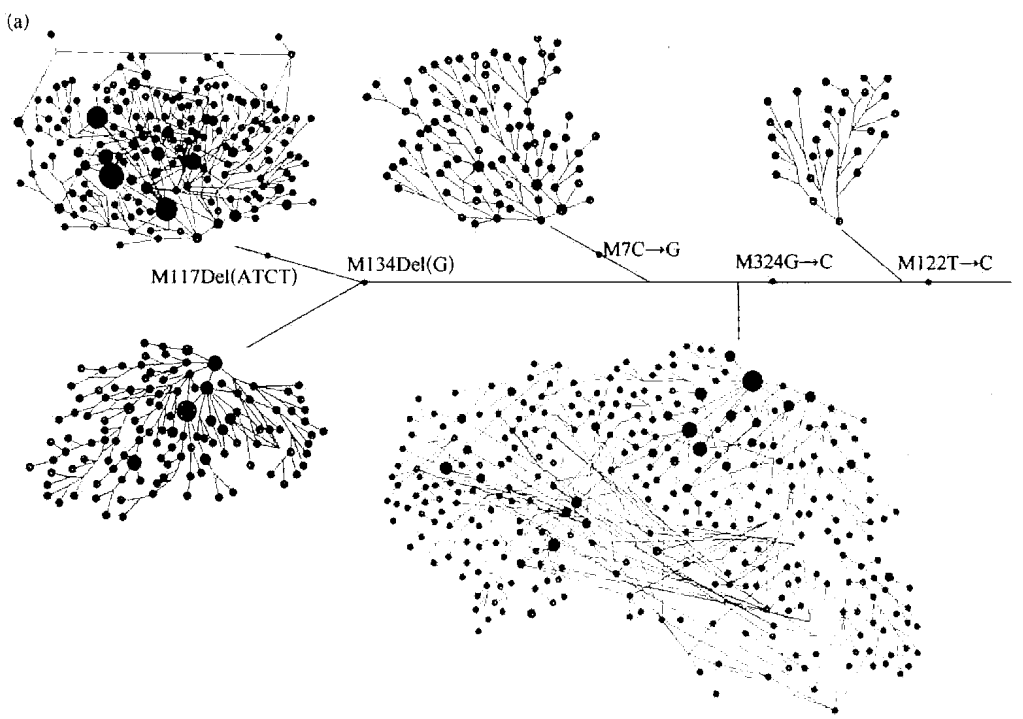
人 群	标示	语言	样本量	单倍型样本数量							
				M122	O3b	O3a *	O3a1	O3a2	O3a4b	O3a5b	O3a5a2
湖南白族	T3	藏缅语族	38	31		12			2	16	1
土家族	T4	藏缅语族	100	54		14			2	4	7
怒族	T5	藏缅语族	50	35		2				2	31
哈尼族	T6	藏缅语族	41	17	1	6				3	7
拉祜族	T7	藏缅语族	88	38	2	19			8	5	4
傈僳族	T8	藏缅语族	49	23	2	17				1	3
彝族	T9	藏缅语族	47	12	1	4			6	1	
景颇族	T10	藏缅语族	17	4	1	1			1		
普米族	T11	藏缅语族	47	4		1					3
纳西族	T12	藏缅语族	87	12		5				4	3
云南藏族	T13	藏缅语族	50	22		5				6	11
独龙族	T14	藏缅语族	28	28		19					9
云南壮族	D1	侗傣语系	47	15	2	6			1	5	1
广西壮族	D2	侗傣语系	39	5	1	2				1	
布依族	D3	侗傣语系	48	3		1			1		1
水族	D4	侗傣语系	40	28		6				3	19
傣族	D5	侗傣语系	132	29	2	13			5	1	1
泰族	D6	侗傣语系	60	21		3			7	7	4
云南苗族	M1	苗瑶语系	48	21	1	8			8	2	2
湖南苗族	M2	苗瑶语系	105	48	1	9			2	1	3
云南瑶族	M3	苗瑶语系	90	50		9			10	20	6
广西瑶族	M4	苗瑶语系	225	104	2	30			6	7	25
湖南瑶族	M5	苗瑶语系	20	15		3			1	1	4
广东瑶族	M6	苗瑶语系	37	24		5			9	1	5
佤族	A1	南亚语系	31	15	1	6				3	5
布朗族	A2	南亚语系	28	6		2				3	1
德昂族	A3	南亚语系	16	9		2				2	5
柬埔寨人	A4	南亚语系	14	2			1	1			
云南满族	C1	阿尔泰语系	41	24		1				2	21

（续表）

人　群	标示	语言	样本量	单倍型样本数量							
				M122	O3b	O3a *	O3a1	O3a2	O3a4b	O3a5b	O3a5a2
云南蒙古族	C2	阿尔泰语系	46	8		4				2	2
云南回族	C3	汉语	15	3						2	1

注：共获得 854 个样本的全套完整 SNP 和 STR 数据，因此一些人群中的亚型总数小于所采集的 O3 - M122 样本总数。

既然东亚特有单倍群 O3 - M122 极有可能是共同起源的，那么该单倍群最初起源于何处？我们对 SNP 和 STR 数据构建的网络图进行了详细的分析。图 3 - 5a 显示，在 O3 - M122 出现之后发生了大量近似 STR 的进化事件，在南北方人群中观察到许多共享的 STR 单倍型，这再一次证明了东亚 M122 支系的近期共同祖先起源。据可查资料记载，现今生活在中国西南的藏缅语人群最早在新石器时代晚期从北方迁徙而来，然后经受了大量南方人群的影响，包括侗傣和南亚语人群[26,33]。此外，尽管苗瑶人群通常被视作南方人群，但事实上他们与汉族人群之间存在可查的人群混合历史[33,41]。南方汉族人群来源于晚近的北方移民，因为汉文化的扩张就发生在过去几千年间[33]。为了消除相对近期人群混合的影响，在剔除藏缅、阿尔泰、苗瑶和南方汉族之后，重新构建了 STR 网络图。重建的网络图（图 3 - 5b）显示，大多数主要 STR 单倍型出现在南方人群中（侗傣和南亚）。因此认为，MDS 分析和 STR 网络图均支持东亚 O3 - M122 谱系的南方起源（图 3 - 5b）。



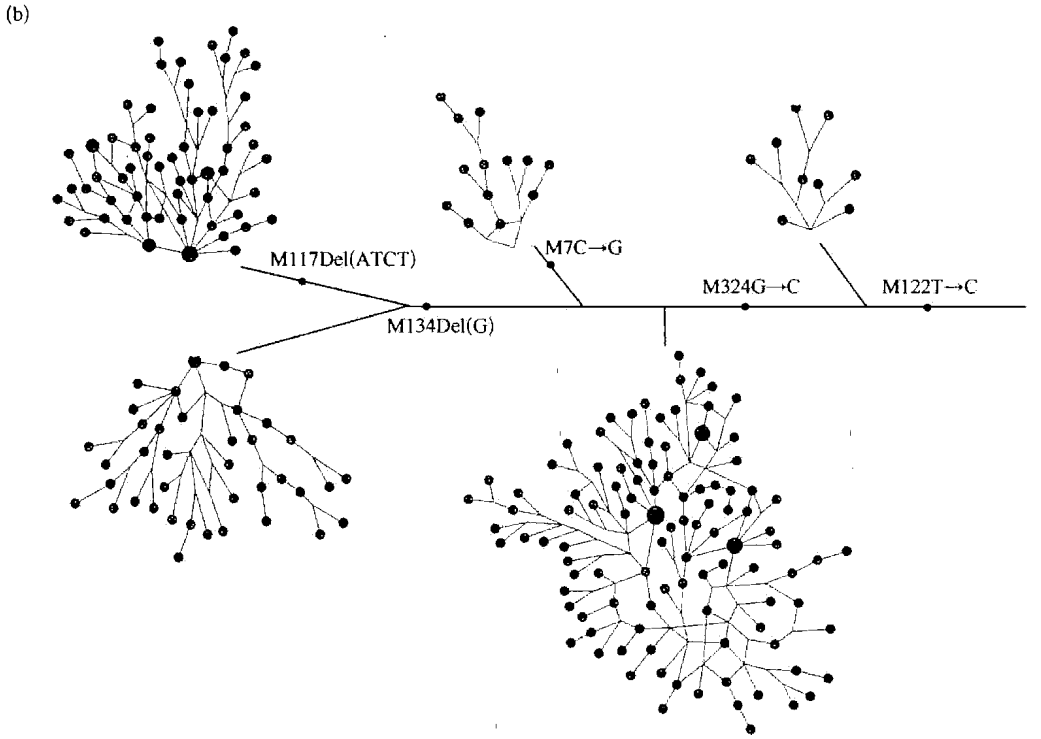


图 3-5 Y-SNP 背景下的 Y-STR 单倍型网络图

(a) 所有人群构建的网络图。蓝色为北方汉族，橙色为南方人群(侗傣、苗瑶和南亚)，绿色为南方汉族和藏缅人群。(b) 为避免近期人群混合影响而剔除南方汉族、阿尔泰、藏缅、苗瑶人群后构建的网络图。蓝色为北方汉族，橙色为侗傣和南亚人群。点的大小比例对应于单倍型频率，多颜色的点代表单倍型在人群中共享。

应当指出，遗传分化差异的缺乏，即南北方 O3-M122 支系的基因多样性差异不明显，表明 O3-M122 可能是最初往北迁徙人群中的主流类型。因此在 O3-M122 支系中并未观察到明显的瓶颈效应，不同于分布明显存在偏移的 O2-M95 和 O1-M119 支系[16,26]。但是，汉文化扩张引起的近期基因交流，也会在某种程度上导致 O3-M122 支系内部的同质性，尽管侗傣和南亚人群相比于藏缅和苗瑶人群受到汉族人群的影响更少[18,26,33]。

在 STR 数据的基础上，笔者估测了 O3-M122 主要亚单倍群的产生时间，用的是 Zhivotovsky 的联合方法[36]。表 3-3 列出了估计的分歧时间，所有的亚单倍群出现时间都早于新石器时代，范围为 2.5 万～3 万年前。

表 3-3 Y-STR 数据估计的 O3-M122 亚单倍型分歧时间

O3-M122 亚单倍型	分歧时间(年)		
	上限	下限	平均值±标准误
O3a_M324G	38 579	21 053	29 816 ±1 161
O3a5_M134Del	33 799	16 915	25 357 ±1 592

（续表）

O3－M122亚单倍型	分歧时间（年前）		
	上限	下限	平均值±标准误
O3a4_M7G	36 759	19 875	28 317 ±4 030
O3a5a2_M117Del	37 398	22 217	29 807 ±1 613

注：分歧时间估计依据TD方法[36,37]。估计上限时，假定$V_0=0$[9]；估计下限时，假定$V_0=Va$，Va为祖先人群的群体内方差。

3.1.4　讨论

O3－M122单倍型在东亚人群中的分布支持东亚人群的南方起源。该单倍型的时间估计和东亚发掘的化石记录一致，且未发现早于4万年的现代人化石[16,20,42]。有趣的是，O3－M122所有亚单倍群的分化似乎发生在一段相对较短的时间内（2.5万～3万年前），这意味着在东亚南部曾发生过一次古代人群扩张事件，并且引发了最初往北方的迁徙。因此，我们认为最初往北迁徙的时间应该与O3－M122的亚单倍群分歧时间相近，这个推测也与中国北方发现的最早化石记录（2.7万～3.9万年前）相一致[20,42]。

丁远春等[19]指出，南方地区人口稠密，而北方地区人口稀少。因而，发生在北方人群中的区域间迁徙，伴随着高频率的遗传漂变和谱系丢失，会造成谱系结构的不平衡，这也表明东亚人群的北方起源并不能被完全排除[11-22]。但是，在丁远春等[19]的研究中，所用的南部人群是有近期北方起源历史记载的藏缅语人群，因此正如许多遗传系统中所观察到的[11-22]，这会模糊南北方之间的分化差异。另一方面，在过去5 000年中，因为汉文化的扩张[33,43]，东亚主要的人群迁徙都是从北到南，这在最近的研究中也得到了证明[25]。当讨论新石器时代之前东亚最早的现代人迁徙时，我们所得的O3－M122多态性数据揭示，南方人群是较可能的祖先人群。来自另两个东亚特有的主要单倍群（O1－M119和O2－M95）的数据也支持北方人群的南方起源。这两个支系在东亚南方中普遍存在（表3－1），但是在东亚北方中较为罕见[26]。

Karafet等[22]发现，中亚、东亚北方和东亚南方共享一些Y单倍型，而东亚北方的多样性高于东亚南方。但是，当仔细查看Karafet等[22]所述的具体单倍型分布时发现，观察到的28个单倍群中的9个是M175的下游类型，而M175在所研究的北方和南方男性Y染色体中分别占30.5%（230/754）和79.1%（398/503）。已知M175是东亚特有的类型[2,28]。尽管东亚南方中发现的单倍群类型少于东亚北方，但这与M175支系内部东亚南方多样性高于东亚北方的模式是相反的。在东亚南方中，M175下游单倍群的基因多样性高于东亚北方（0.86比0.67），而在东亚北方中，下游单倍群的分布存在高度偏移。另一方面，从存在近期人群混合历史（<1 000年前[33]）的人群中分析解读数据，会导致错误的结果。例如，Karafet等[22]研究的回族和维吾尔族，他们都是晚近形成的东亚北方，同中亚存在不同程度的混合。蒙古族同中亚和北亚（西伯利亚）人群存在稳定的基因交流。因此，Karafet等声称的东亚北方中观察到更多的Y染色体单倍群类型[22]，事实上

只是近期人群混合造成的假象。同样的情况在东亚北方的 M45[27] 和 M89 中可以清晰地反映出来，然而这些类型在东亚南方中十分罕见[16]。当近期基因交流的遗传影响，即单倍群 M45 和 M89 从分析中消除后，Karafet 等研究中的东亚特有单倍群分布将会得出东亚南方是东亚北方祖先的结论。

根据已发表的数据和本研究的数据，可以认为大致有两个主要原因导致了目前可分辨的东亚南方和东亚北方之间的遗传差异：① 源于早期往北迁徙的奠基者效应和其后的地理隔离，可以通过 O1 - M119 和 O2 - M95 支系的分布看出[16,26]；② 可能源于东亚北方相对晚近的人群混合，引入了高加索和中北亚 Y 染色体类型[20,22]。

总之，东亚特有单倍群 O3 - M122 数据证明了 O3 - M122 的南方起源，进而支持东亚现代人的南方起源假说。最早往北的史前迁徙估计发生在 2.5 万～3 万年前。

参考文献

[1] Jobling M A, Tyler-Smith C. The human Y chromosome: an evolutionary marker comes of age. Nat Rev Genet, 2003, 4: 598 - 612.

[2] Underhill P A, Passarino G, Lin A A, et al. The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. Ann Hum Genet, 2001, 65: 43 - 62.

[3] Kayser M, Roewer L, Hedman M, et al. Characteristics and frequency of germline mutations at microsatellite loci from the human Y chromosome, as revealed by direct observation in father/son pairs. Am J Hum Genet, 2000, 66: 1580 - 1588.

[4] Nachman M W, Crowell S L. Estimate of the mutation rate per nucleotide in humans. Genetics, 2000, 156: 297 - 304.

[5] The Y Chromosome Consortium. A nomenclature system for the tree of human Y chromosomal binary haplogroups. Genome Res, 2002, 12: 339 - 348.

[6] Dupuy B M, Stenersen M, Egeland T, et al. Y chromosomal microsatellite mutation rates: differences in mutation rate between and within loci. Hum Mutat, 2004, 23: 117 - 124.

[7] Ramakrishnan U, Mountain J L. Precision and accuracy of divergence time estimates from STR and SNPTR variation. Mol Biol Evol, 2004, 21: 1960 - 1971.

[8] Di Giacomo F, Luca F, Popa L O, et al. Y chromosomal haplogroup J asa signature of the post-Neolithic colonization of Europe. Hum Genet, 2004, 115: 357 - 371.

[9] Rootsi S, Magri C, Kivisild T, et al. Phylogeography of Y chromosome haplogroup I reveals distinct domains of prehistoric gene flow in Europe. Am J Hum Genet, 2004, 75: 128 - 137.

[10] Semino O, Magri C, Benuzzi G, et al. Origin, diffusion, and differentiation of Y-chromosome haplogroups E and J: inferences on the neolithization of Europe and later migratory events in the Mediterranean area. Am J Hum Genet, 2004, 74: 1023 - 1034.

[11] Zhao T M, Zhang G L, Zhu Y M, et al. The distribution of immunoglobulin Gmallotypes in forty Chinese populations. Acta Anthropol Sin, 1986, 6: 1 - 8.

[12] Weng Z L, Yan Y D. Analysis of the genetic structure of human populations in China. Acta Anthropol Sin, 1989, 8: 261 - 268.

[13] Chu J Y, HuangW, Kuang S Q, et al. Genetic relationship of populations in China. Proc Natl Acad Sci USA, 1998, 95: 11763 - 11768.

[14] Du R F, Xiao C J, Cavalli-Sforza L L. Genetic distances between Chinese groups calculated on gene frequencies of 38 loci. Sci China Series C, 1998: 83 - 89.

[15] Piazza A. Towards a genetic history of China. Nature, 1998, 395: 636 - 639.

[16] Su B, Xiao J, Underhill P, et al. Y chromosome evidence for a northward migration of modern humans into Eastern Asia during the Last Ice Age. Am J Hum Genet, 1999, 65: 1718 - 1724.

[17] Su B, Jin L, Underhill P, et al. Polynesian origins: insights from the Y chromosome. Proc Natl Acad Sci USA, 2000a, 97: 8225 - 8228.

[18] Su B, Xiao C, Deka R, et al. Y chromosome haplotypes reveal prehistorical migrations to the Himalayas. Hum Genet, 2000b, 107: 582 - 590.

[19] Ding Y C, Wooding S, Harpending H C, et al. Population structure and history in East Asia. Proc Natl Acad Sci USA, 2000, 97: 14003 - 14006.

[20] Jin L, Su B. Natives or immigrants: modern human origin in East Asia. Nat Rev Genet, 2000, 1: 126 - 133.

[21] Capelli C, Wilson J F, Richards M, et al. A predominantly indigenous paternal heritage for the Austronesian-speaking peoples of insular Southeast Asia and Oceania. Am J Hum Genet, 2001, 68: 432 - 443.

[22] Karafet T, Xu L, Du R, et al. Paternal population history of East Asia: sources, patterns, and microevolutionary processes. Am J Hum Genet, 2001, 69: 615 - 628.

[23] Ballinger S W, Schurr T G, Torroni A, et al. Southeast Asian mitochondrial DNA analysis reveals genetic continuity of ancient Mongoloid migrations. Genetics, 1992, 130: 139 - 152.

[24] Yao Y G, Kong Q P, Bandelt H J, et al. Phylogeographic differentiation of mitochondrial DNA in Han Chinese. Am J Hum Genet, 2002, 70: 635 - 651.

[25] Wen B, Li H, Lu D, et al. Genetic evidence supports demic diffusion of Han culture. Nature, 2004a, 431: 302 - 305.

[26] Wen B, Xie X, Gao S, et al. Analyses of genetic structure of Tibeto-Burman populations reveals sex-biased admixture in southern Tibeto-Burmans. Am J Hum Genet, 2004b, 74: 856 - 865.

[27] Semino O, Passarino G, Oefner P J, et al. The genetic legacy of Paleolithic *Homo sapiens sapiens* in extant Europeans: a Y chromosome perspective. Science, 2000, 290: 1155 - 1159.

[28] Underhill P A, Shen P, Lin A A, et al. Y chromosome sequence variation and the history of human populations. Nat Genet, 2000, 26: 358 - 361.

[29] Wells R S, Yuldasheva N, Ruzibakiev R, et al. The Eurasian heartland: a continental perspective on Y chromosome diversity. Proc Natl Acad Sci USA, 2001, 98: 10244 - 10249.

[30] Lell J T, Sukernik R I, Starikovskaya Y B, et al. The dual origin and Siberian affinities of Native American Y chromosomes. Am J Hum Genet, 2002, 70: 192 - 206.

[31] Jin H J, Kwak K D, Hammer M F, et al. Y chromosomal DNA haplogroups and their implications for the dual origins of the Koreans. Hum Genet, 2003, 114: 27 - 35.

[32] Qian Y, Qian B, Su B, et al. Multiple origins of Tibetan Y chromosomes. Hum Genet, 2000,

106: 453 - 454.

[33] 王钟翰. 中国民族史. 北京: 中国社会科学出版社,1994.

[34] Hammer M F, Karafet T, Rasanayagam A, et al. Out of Africa and back again: nested cladistic analysis of human Y chromosome variation. Mol Biol Evol, 1998, 15: 427 - 441.

[35] Schneider S, Kueffer J M, Roessli D, et al. Arlequin: a software for population genetic analysis. Genetics and Biometry Laboratory, University of Geneva, Geneva, 1998.

[36] Zhivotovsky L A. Estimating divergence time with the use of microsatellite genetic distances: impacts of population growth and gene flow. Mol Biol Evol, 2001, 18: 700 - 709.

[37] Zhivotovsky L A, Underhill P A, Cinnioğlu C, et al. The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. Am J Hum Genet, 2004, 74: 50 - 61.

[38] Shen P, Wang F, Underhill P A, et al. Population genetic implications from sequence variation in four Y chromosome genes. Proc Natl Acad Sci USA, 2000, 97: 7354 - 7359.

[39] Ke Y, Su B, Li H, et al. Y chromosome evidence for no independent origin of modern human in China. Chinese Sci Bul, 2001a, 46: 935 - 937.

[40] Ke Y, Su B, Xiao C, et al. Y chromosome haplotype distribution in Han Chinese populations and modern human origin in East Asians. Sci China Series C, 2001b, 44: 225 - 232.

[41] Wen B, Li H, Gao S, et al. Genetic structure of Hmong-Mien speaking populations in East Asia as revealed by mtDNA lineages. Mol Biol Evol, 2005, 22: 725 - 734.

[42] Wu H C, Poirier F E, Wu X Z. Human evolution in China: a metric description of the fossils and a review of the sites. United Kingdom, Oxford: Oxford University Press, 1995.

[43] Cavalli-Sforza L L, Menozzi P, Piazza A. The history and geography of human genes. Princeton: Princeton University Press, 1994.

[44] Ramana G V, Su B, Jin L, et al. Y chromosome SNP haplotypes suggest evidence of gene flow among Caste, tribe, and the migrant Siddi populations of Andhra Pradesh, South India. Eur J Hum Genet, 2001, 9: 695 - 700.

3.2 汉文化的扩散源于人口扩张

3.2.1 研究背景

史载汉族源于古代中国北方的华夏部落,在过去的 2 000 多年间,汉文化(汉语和相关的文化传统)扩散到了中国南方,而中国南方原住民族则是说侗台、南亚和苗瑶语的人群(百越、百濮和荆蛮)[1,2]。经典遗传标记和微卫星位点研究显示,汉族和其他东亚人群一样都可以以长江为界分为两个遗传亚群:南方汉族和北方汉族[3-6]。两个亚群之间的方言和习俗差异也很显著[7]。这些现象看似支持文化传播模式,即汉族向南扩张主要是文化传播和同化的结果。然而,两个亚群之间有着许多共同的 Y 染色体和线粒体类型[8,9],历史记载的汉族移民史[2]也与汉族的文化传播模式假说相矛盾。本节对这两种假说进行检验,证实汉文化在扩散中的确发生了大规模的人群迁徙(人口扩张模式)。

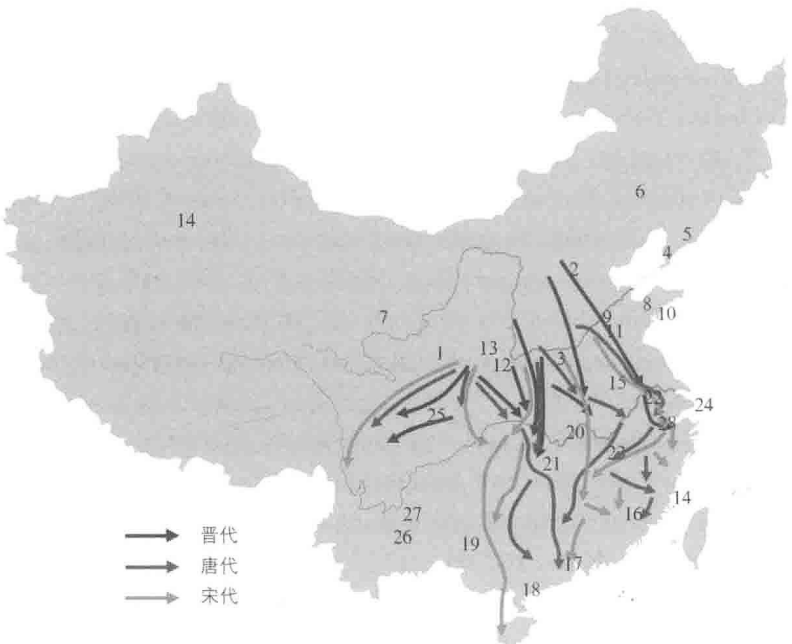


图 3-6　调查群体的地理分布

图中标出了历史记载中自北而南的3次迁徙浪潮。群体1～14是北方汉族，15～28是南方汉族。实线、段线和虚线依次表示3次迁徙浪潮。第1次发生于西晋时期(265—316年)，迁徙人口约90万(大约相当于当时南方人口的1/6)；第2次发生于唐代(618—907年)，规模比第一次大得多；第3次发生于南宋(1127—1279年)，迁徙人口近500万。

3.2.2　材料和方法

1. 样本

采集中国各地的17个汉族群体871个随机不相关个体的血样。用酚-氯仿法抽提基因组DNA。结合文献报道的Y染色体和线粒体多态性数据，总共分析的样本量是：Y染色体23个群体1 289人，线粒体23个群体1 119人。这些样本涉及中国大部分省份(图3-6)。

2. 遗传标记

通过聚合酶链反应-限制性片段长度多态性分析(PCR-RFLP)的方法[8]分型Y染色体上的13个双等位标记：YAP、M15、M130、M89、M9、M122、M134、M119、M110、M95、M88、M45和M120。根据国际Y染色体命名委员会的命名系统[9]，这些标记构成13

个单倍群，在东亚人群中具有较高的信息量[10]。

线粒体上，对高变1区（HVS-1）进行测序，对编码区8个多态位点做了分型（9-bp缺失，10397 AluI、5176 AluI、4831 HhaI、13259 HincII、663 HaeIII、12406 HpaI和9820 HinfI），有关方法已有报道[11]。根据东亚线粒体系统树[12]，用高变1区突变结构和编码区多态性构建单倍群。

3. 数据分析

根据线粒体和Y染色体单倍群频率，用SPSS10.0做主成分分析，研究群体间的关系。南北汉族的遗传差异用Arlequin[13]做AMOVA检验[14]。南方汉族中北方汉族和南方原住民族的混合比例估计用两种不同的统计方法[15,16]：Admix 2.0[17]和LEADMIX[18]。亲本群体的选择对混合比例的适当估计很重要[19,20]，我们通过扩大东亚的参考数据来减小偏差。分析中，10个北方汉族群体的各单倍群频率（Y染色体和线粒体标记分别分析）的算术平均作为北方亲本群体。南方原住民族的频率取了3个族群的平均：侗台语群（NRY，22群体；线粒体，11群体），南亚语群（NRY，6群体；线粒体，5群体），苗瑶语群（NRY，18群体；线粒体，14群体）。通过样本的混合比例与纬度[21-23]的线性回归分析揭示汉族群体的地理格局。

3.2.3 研究结果

为了验证这些假说，我们把南方汉族的遗传结构与两个亲本群体做比较，其一是北方汉族，其二是南方原住民族，即现居于中国境内和若干邻国的侗台、苗瑶和南亚语群体。我们分析了来自中国28个地区汉族群体的Y染色体非重组区（NRY）和线粒体DNA（mtDNA）遗传多态性[24-27]，这些样本覆盖中国绝大多数省份（图3-6）。

父系方面，南方汉族与北方汉族的Y染色体单倍群频率分布非常相近，尤其是具有M122-C突变的单倍群（O3-M122和O3e-M134）普遍存在于研究的汉族群体中（北方汉族为37%～71%，平均为53.8%；南方汉族为35%～74%，平均为54.2%）。南方原住民族中普遍出现的单倍群M119-C（O1）和M95-T（O2a）在南方汉族中的频率（3%～42%，平均为19%）高于北方汉族（1%～10%，平均为5%）。而且，南方原住民族中普遍存在的单倍群O1b-M110、O2a1-M88和O3d-M7[28]在南方汉族中低频存在（平均为4%），而北方汉族中却没观察到。如果我们假定起始于2 000多年前的汉文化扩散[2]之前，南方原住民族的Y类型频率与现在基本一致，南方汉族中南方原住民族的成分应该是不多的。分子方差分析（AMOVA）进一步显示北方汉族和南方汉族的Y染色体单倍群频率分布没有显著差异（$F_{ST}=0.006$，$P>0.05$），说明南方汉族在父系上与北方汉族非常相似。

母系方面，北方汉族与南方汉族的线粒体单倍群分布非常不同。东亚北部的主要单倍群（A、C、D、G、M8a、Y和Z）在北方汉族中的频率（49%～64%，平均为55%）比在南方汉族中（19%～52%，平均为36%）高得多。另一方面，南方原住民族的主要单倍群（B、F、R9a、R9b和N9a）[12,25,29]在南方汉族中的频率（36%～72%，平均为55%）要比在北方

汉族中(18%～42%，平均为 33%)高得多。线粒体类型的分布在南北汉族之间有极显著差异($F_{ST}=0.006$，$P<10^{-5}$)。虽然南北汉族之间线粒体和 Y 染色体的 F_{ST} 值相近，但线粒体的南北差异 F_{ST} 值占群体间总方差的 56%，而 Y 染色体仅占 18%。

用汉族群体的单倍群频率数据所做的主成分(PC)分析与以上结果相一致。NRY 分析发现，几乎所有的汉族群体都聚在图 3－7a 的右上方。北方汉族和南方原住民族在第二主成分上分离，南方汉族的第二主成分值处于北方汉族和南方原住民族之间，但是更接近北方汉族(北方汉族为 0.58 ±0.01；南方汉族为 0.46 ±0.03；南方原住民族为－0.32 ±0.05)，这表明南方汉族在父系上与北方汉族相近，受到南方原住民族的影响很小。就线粒体 DNA 而言，北方汉族和南方原住民族仍然被第二主成分分开(图 3－7b)，南方汉族也在两者之间但稍微接近南方原住民族(北方汉族为 0.56 ±0.02；南方汉族为0.09 ±0.06；南方原住民族为－0.23 ±0.04)，表明南方汉族的女性基因库比男性基因库有更多的混合成分。

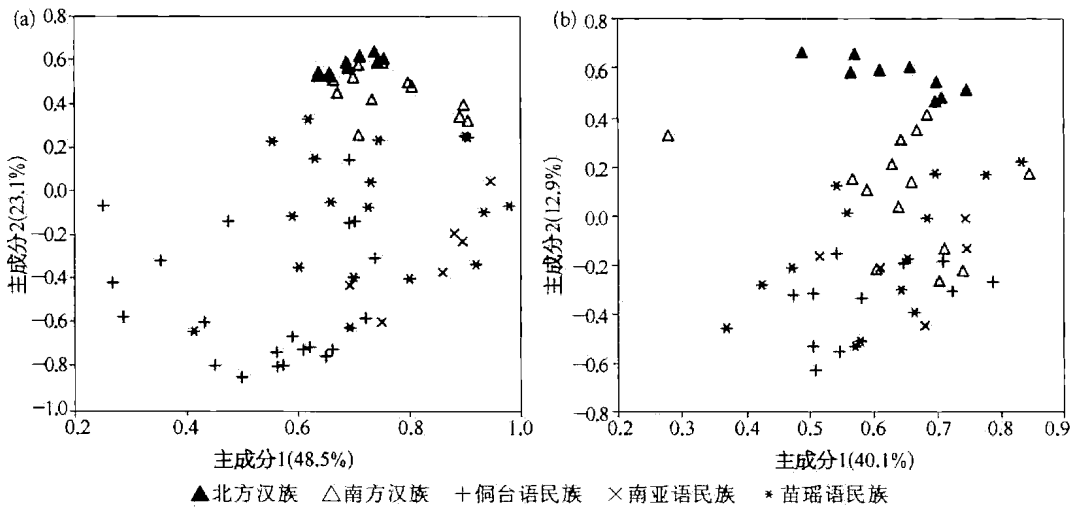


图 3－7　主成分散点图

(a) Y 染色体单倍群散点图；(b) 线粒体单倍群散点图

我们进一步用两种不同的统计方法[15,16]来估计两个亲本(北方汉族和南方原住民族)对南方汉族基因库的相对贡献(表 3－4)，这两个统计量用于单位点(single-locus)分析时比其他的方法更为准确[18]。两种方法得到的混合系数估计值(M，北方汉族的贡献比例)高度一致(Y 染色体，$r=0.922$，$P<0.01$；线粒体，$r=0.970$，$P<0.01$)。就 Y 染色体而言，所有的南方汉族都包含很高比例的北方汉族混合比率(M_{BE}：0.82 ±0.14，范围 0.54～1；M_{RH}：0.82 ±0.12，范围 0.61～0.97；M_{BE} 和 M_{RH} 的定义分别见参考文献[20]和[19])，这表明南方汉族男性基因库的主要贡献成分来自北方汉族。相反，南方汉族的线粒体基因库中北方汉族和南方原住民族的贡献比例几乎相等(M_{BE}：0.56±0.24[0.15，0.95]；M_{RH}：0.50 ±0.26[0.07，0.91])。总体上北方汉族对南方汉族的遗传贡献父系比母系高得多(t 检验，$P<0.01$)；各群体分别看也是这样：绝大部分南方汉族群体中北方汉族的贡献在父系上大于母系(M_{BE}，11/13，M_{RH}，13/13，$P<0.01$，零假设为男

女的贡献相等，为二项式分布），这表明南方汉族的群体混合过程有很强的性别偏向。南方汉族中北方汉族贡献的比例（M）呈现出由北向南递减的梯度地理格局。南方汉族线粒体的 M 值与纬度正相关（$r^2=0.569$，$P<0.01$），但 Y 染色体的相关性不显著（$r^2=0.072$，$P>0.05$），因为南方汉族父系的 M 值差异太小，不足以导致统计学上的显著性。

3.2.4 结论

综上所述，我们提出了两项证据支持汉文化扩散的人口扩张假说。首先，几乎所有的汉族群体的 Y 染色体单倍群分布都极为相似，Y 染色体主成分分析也把几乎所有的汉族群体都集合成一个紧密的聚类。其次，北方汉族对南方汉族的遗传贡献无论父系方面还是母系方面都是可观的，在线粒体 DNA 分布上也存在地理梯度。北方汉族对南方汉族的遗传贡献在父系（Y 染色体）上远大于母系（线粒体），表明这一扩张过程中汉族男性处于主导地位；换个角度看，在汉族和南方原住民的融合过程中有相对较多的当地女性融入南方汉族中。性别偏向的混合格局也同样存在于藏缅语人群中[11]。

据历史记载，受北方战乱和饥荒的影响，汉族人不断南迁，图 3-6 中画出了 3 次大规模移民的浪潮。在 2 000 多年间，除了这 3 次大潮，各个时期几乎都有小规模的南迁。所以，我们的遗传研究也与历史记载相吻合。大量的北方移民改变了中国南方的遗传构成，而汉族人口扩张的同时也带动了汉文化的扩散。除了大规模的人群迁徙，北方汉族、南方汉族和南方原住民族之间的基因交流造成的族群混合也在很大程度上改变了中国人群的遗传结构。

表 3-4 南方汉族中的北方汉族混合比例

群 体	Y 染色体		线粒体 DNA	
	M_{BE}（±s. e. m）	M_{RH}	M_{BE}（±s. e. m）	M_{RH}
安徽	0.868±0.119	0.929	0.816±0.214	0.755
福建	1	0.966	0.341±0.206	0.248
广东-1	0.677±0.121	0.669	0.149±0.181	0.068
广东-2	ND	ND	0.298±0.247	0.312
广西	0.543±0.174	0.608	0.451±0.263	0.249
湖北	0.981±0.122	0.949	0.946±0.261	0.907
湖南	0.732±0.219	0.657	0.565±0.297	0.490
江苏	0.789±0.078	0.821	0.811±0.177	0.786
江西	0.804±0.113	0.829	0.374±0.343	0.424
上海	0.819±0.087	0.902	0.845±0.179	0.833
四川	0.750±0.118	0.713	0.509±0.166	0.498
云南-1	1	0.915	0.376±0.221	0.245

（续表）

群 体	Y染色体		线粒体DNA	
	M_{BE}（±s. e. m）	M_{RH}	M_{BE}（±s. e. m）	M_{RH}
云南-2	0.935±0.088	0.924	0.733±0.192	0.645
浙江	0.751±0.084	0.763	0.631±0.180	0.540
平均	0.819	0.819	0.560	0.500

注：M_{BE}和M_{RH}分别为参考文献[16]和[15]所描述的统计量。M_{BE}的标准误通过1 000次自展（Bootstrap）获得。把南方原住民族和北方汉族作为南方汉族的亲本群体估计北方汉族的遗传贡献比例，假定2 000多年前开始的混合过程前后南方原住民族的等位基因频率基本不变，并且南北汉族之间的遗传交流不多。实际上，从北方汉族到南方原住民族的基因流动比反向的流动大得多，所以表中的估计值在没有适当调整前是低估的。因而汉族实际的人口扩张程度应该大于本项研究得出的数值。

参考文献

[1] 费孝通. 中华民族多元一体格局. 北京：中央民族大学出版社，1999.

[2] 葛剑雄，吴松弟，曹树基. 中国移民史. 福州：福建人民出版社，1997.

[3] Zhao T M, Lee T D. Gmand Kmallotypes in 74 Chinese populations: a hypothesis of the origin of the Chinese nation. Hum, Genet, 1989, 83: 101-110.

[4] Du R F, Xiao C J, Cavalli-Sforza L L. Genetic distances calculated on gene frequencies of 38 loci. Science in China Ser C, 1997, 40: 613.

[5] Chu J Y, Huang W, Kuang S Q, et al. Genetic relationship of populations in China. Proc Natl Acad Sci USA, 1998, 95: 11763-11768.

[6] Xiao C J, Cavalli-Sforza L L, Minch E, et al. Principal component analysis of gene frequencies of Chinese populations. Sci China C, 2000, 43: 472-481.

[7] Xu Y T. A brief study on the origin of Han nationality. J Centr Univ Natl, 2003, 30: 59-64.

[8] Su B, Xiao C, Deka R, et al. Y chromosome haplotypes reveal prehistorical migrations to the Himalayas. Hum Genet, 2000, 107: 582-590.

[9] The Y Chromosome Consortium. A nomenclature system for the tree of human Y chromosomal binary haplogroups. Genome Res, 2002, 12: 339-348.

[10] Jin L, Su B. Natives or immigrants: modern human origin in East Asia. Nature Rev Genet, 2000, 1: 126-133.

[11] Wen B, Xie X, Gao S, et al. Analyses of genetic structure of Tibeto-Burman populations revealed a gender-biased admixture in southern Tibeto-Burmans. Am J Hum Genet, 2004, 74: 856-865.

[12] Kivisild T, Tolk H V, Parik J, et al. The emerging limbs and twigs of the East Asian mtDNA tree. Mol Biol Evol, 2002, 19: 1737-1751.

[13] Schneider S, Roessli D, Excoffier L. Arlequin: Ver. 2.000. A software for population genetic analysis. Geneva: Genetics and Biometry Laboratory, Univ-of Geneva, 2000.

[14] Excoffier L, Smouse P E, Quattro J M. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics, 1992, 131: 479-491.

[15] Roberts D F, Hiorns R W. Methods of analysis of the genetic composition of a hybrid population. Hum Biol, 1965, 37: 38 - 43.

[16] Bertorelle G, Excoffier L. Inferring admixture proportions from molecular data. Mol Biol Evol, 1998, 15: 1298 - 1311.

[17] Dupanloup I, Bertorelle G. Inferring admixture proportions from molecular data: extension to any number of parental populations. Mol Biol Evol, 2001, 18: 672 - 675.

[18] Wang J. Maximum-likelihood estimation of admixture proportions from genetic data. Genetics, 2003, 164: 747 - 765.

[19] Chakraborty R. Gene admixture in human populations: models and predictions. Yb Phys Anthropol, 1986, 29: 1 - 43.

[20] Sans M, Weimer T A, Franco M H, et al. Unequal contributions of male and female gene pools from parental populations in the African descendants of the city of Melo, Uruguay. Am J Phys Anthropol, 2002, 118: 33 - 44.

[21] Cavalli-Sforza L L, Menozzi P, Piazza A. The history and geography of human genes. Princeton: Princeton Univ Press, 1994.

[22] Sokal R, Oden N L, Wilson C. Genetic evidence for the spread of agriculture in Europe by demic diffusion. Nature, 1991, 351: 143 - 145.

[23] Chikhi L, Nichols R A, Barbujani G, et al. Y genetic data support the Neolithic demic diffusion model. Proc Natl Acad Sci USA, 2002, 99: 11008 - 11013.

[24] Cavalli-Sforza L L, Feldman M W. The application of molecular genetic approaches to the study of human evolution. Nat Genet, 2003, 33: 266 - 275.

[25] Wallace D C, Brown M D, Lott M T. Nucleotide mitochondrial DNA variation in human evolution and disease. Gene, 1999, 238: 211 - 230.

[26] Underhill P A, Shen P, Lin A A, et al. Y chromosome sequence variation and the history of human populations. Nat Genet, 2000, 26: 358 - 361.

[27] Jobling M A, Tyler-Smith C. The human Y chromosome: an evolutionary marker comes of age. Nat Rev Genet, 2003, 4: 598 - 612.

[28] Su B, Xiao J, Underhill P, et al. Y chromosome evidence for a northward migration of modern humans into Eastern Asia during the Last Ice Age. Am J Hum Genet, 1999, 65: 1718 - 1724.

[29] Yao Y G, Kong Q P, Bandelt H J, et al. Phylogeographic differentiation of mitochondrial DNA in Han Chinese. Am J Hum Genet, 2002, 70: 635 - 651.

3.3 平话群体是汉族一致性遗传结构的例外

3.3.1 研究背景

汉族目前约有13亿人口，是世界人口最多的民族[1]，占世界总人口的19%。其人口绝大多数都分布于中国，或集居于中国东部农业发达的地区，或杂居于边疆少数民族中，或在世界各国散居。汉族的遗传结构对于研究大型民族的人口扩张有代表性意义。汉族有着悠久的历史，其起源和发展是一个持续不间断的过程。汉族灿烂的文化，对中国乃至世界

历史和文明都做出了伟大的贡献[2]。1 万～4 万年前汉藏语系的祖先由东亚南部迁徙到黄河中上游盆地，形成了最早的汉藏祖先人群，并在五六千年前分化出中国北方有史记载的华夏族，即汉族的祖先[3]。在此后的几千年里，汉族凭借其相对先进的生产力和领先的文化，不断发展壮大，汉族人口迅速扩张。在过去 2 000 多年间，汉族人群以及汉文化扩散到中国南方。南方的原住居民大多属于侗台、苗瑶和南亚语族人群，与汉族交流融合形成南方汉族。在此过程中，南北汉族在语言和文化等方面虽然发生了变化，但仍保留了相对的一致性。从之前的遗传研究中也可以观察到，大部分南方汉族的遗传结构仍与北方汉族基本一致，并未表现出过多的南方原住民族的影响。但是广西汉族的南方少数民族混合比例却较高，表现出与侗台、苗瑶群体相近的遗传结构[1]。汉族的语言统称汉语，汉语可以分成十大方言，即官话、吴语、湘语、客家话、赣语、粤语、闽语、晋语、徽

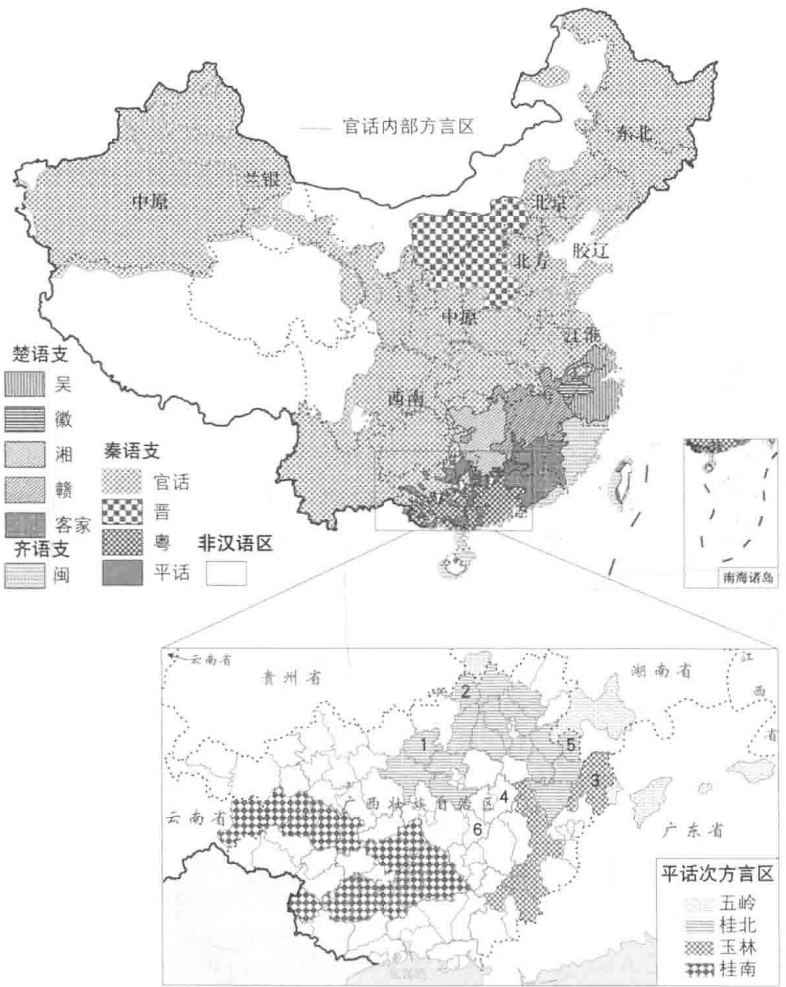


图 3-8　中国汉族分支、平话汉族分支及采样地区

在广西地图上的数字表示采样地点：1 指罗城，2 指三江，3 指贺州，4 指金秀，5 指富川，6 指武宣。

语和平话(图 3-8)。汉族按照这十大方言分成十类群体。广西的汉族群体主要有粤语人群、客家人、官话人群和平话人群。除平话人群外,其余汉族支系都在以往的分子人类学研究中进行过遗传结构的相关调查,并未体现出特殊性。所以广西汉族的特殊遗传结构可能主要源于平话人群。

平话汉族(plebeian Han)是汉族的一个古老支系族群,人口总数在 300 万~400 万,主要分布在广西,靠近广西的湖南南部、云南和广东西北等地也有部分散居。平话人的形成历史甚至比客家人、粤语人群、闽语人群还早。在文化上,平话汉族既保留了典型的汉族语言和文化特征,同时又具有当地少数民族的语言文化特色,包括语音、服饰和饮食习惯。平话人群分为四支,即桂北、桂南、五岭和玉林人群(图 3-8)[4]。平话在各地的名称不一,如南宁市郊区、邕宁、临桂称平话,阳朔叫平声,右江一带叫蔗园话,融安、融水称土拐话,贵港、横县叫土白话,平乐叫土话,横县叫村话,左江一带叫客话(不同于客家话),永福叫百姓话等[5]。

本节从父系和母系分别研究了桂北平话汉族的遗传结构,发现平话人群与南方少数民族在遗传结构上非常相近。这在汉族的起源方式中,除了人口扩张和人群分化之外,又增加了汉族同化外族群体的类型。本节也尝试总结和分析汉族所有支系之间的关系。

3.3.2 材料和方法

1. 群体样本

在广西壮族自治区贺州、富川县、罗城县、金秀县和武宣县的平话汉族群体中,随机采集了 197 份无可查亲缘关系的正常人血样,其中男性样本 101 份,女性样本 96 份。样本来源于不同的乡镇和村落,以避免采样偏差。同时作为对照,在平话人群附近采集了拉伽人、壮族、仫佬族、侗族、瑶族等广西本土少数民族样本,共 273 份无可查亲缘关系的正常人血样,其中男性样本 94 份,女性样本 179 份。所有研究对象都按规定签署知情同意书。采样采取静脉取血的方式,DNA 抽提使用经典的酚-氯仿法。

2. Y 染色体标记

在 195 份男性样本中对人群特异性的 Y 染色体非重组区 20 个位点用 PCR-RFLP 方法进行单倍型分型,它们包括 M130、Yap、M15、M89、M9、M175、M119、M110、M101、M268、M95、M88、M111、M122、M134、M117、M164、M159、M121 和 M7[3,6,7],并对六个 Y 染色体 STR 位点(DYS19、DYS389I、DYS390、DYS391、DYS392 和 DYS393)进行荧光扫描分型[6,7]。

3. 线粒体 DNA 标记

对所有 470 份样本进行线粒体高变 1 区(HVS-1)测序。线粒体 DNA 高变 1 区用引物 L15974 和 H16488 扩增[8]。纯化后的 PCR 产物用荧光标记末段终止法延伸,然后用 ABI 公司生产的 3130 测序仪进行测序分析。设计引物,扩增包含线粒体 DNA 编码区上能够定义单倍型的多态位点的 DNA 片段[1]。然后用限制性内切酶对 PCR 产物分型,这些酶切位点和对应的酶是:663HaeII、5176AluI、12406HpaI、4831HhaI、9824HinfI、

5417RsaI、10310NlaIII、13259HincII、9bp、3391HaeIII、10397 AluI和4715 HaeIII。综合测序分析结果和酶切结果按照经典的Kivisild定义的线粒体DNA单倍群对每一个体进行分型[8,9]。所有样本的线粒体高变1区序列都已经加入数据库GenBank（编号：EU277025－EU277489）。

4. 统计分析

所有Y－SNP单倍型分型以国际遗传谱系学会(ISOGG)2007年版的Y－DNA单倍群进化树为标准。线粒体DNA单倍型分型以文献[9]为标准。在分析中加入文献报道的相关群体数据作为参照[1,6,10-13]。根据线粒体DNA和Y染色体单倍群分布频率用SPSS13.0聚类分析方法中的furthest neighbor模式制作Y单倍型频率和线粒体DNA单倍型频率的树型聚类。同时使用主成分分析方法对以上数据进行分析，用第一主成分和第二主成分作图，分别得到Y染色体和线粒体DNA的主成分散点图。第一主成分和第二主成分在Y染色体的主成分图中总共占的比例为54.1%，线粒体DNA图中则占40.5%。使用Admix 2.0[14]和LEADMIX[15]来评价汉族和中国南方原住民族人群这两大祖先群体对平话人群的遗传贡献率。在此分析中将北方方言、吴方言、湘方言、客家方言、赣方言、粤方言、闽方言、晋方言和徽方言人群的单倍型个体数合并，作为汉族群体各单倍型的总个体数。同时将苗瑶语系和侗台语系的单倍型个体数合并，作为南方少数民族各单倍型的总个体数[1,6,11-13,16]。用Network4.201分析STR单倍型结构。用6个STR位点通过中点连接法(median-joining)构建O2a、O3和O3a5的网络结构。用线粒体DNA HVS－1的突变位点通过中点连接法构建B4＊、B4a＊、B5a＊、F1＊、F1a＊、M7＊、N9a＊、R9b＊和F＊的网络结构[16]。

表3－5　平话人群及其周边少数民族Y染色体单倍群频率分布

编号	人　群	样本量	单倍群比例（%）									
			C	D1	K	O＊	O3＊	O3a4	O3a5	O3a5a	O2＊	O2a＊
P3	平话-贺州	15		6.67		60	13.33			6.67	6.67	6.67
P5	平话-富川	28		17.86	3.57	14.29	3.57	3.57		3.57	3.57	50
P1	平话-罗城	21			4.76	4.76	9.52	9.52	23.81	42.86		4.76
P4	平话-金秀	6			16.67		16.67			33.33	16.67	16.67
P6	平话-武宣	31	3.23						3.23	6.45	3.23	83.88
D4	拉珈-金秀	23	4.35	52.17	4.35		26.1	4.35			8.7	
D6	北部壮族	21		8.92	14.29	4.76	14.29			9.25	4.76	42.86
D1	仫佬族-罗城	11			9.09	18.18	45.45	9.09	9.09		9.09	
D2	侗族-三江	28	21.43		7.14	3.57	7.14		3.57	7.14	21.43	28.57
H5	瑶族-富川	11		9.09	18.18		27.27	9.09			9.09	27.27

3.3.3 研究结果

1. Y 染色体单倍群分型

所有 195 份男性样本的 Y 染色体单倍型按照 ISOGG 最新的分型标准分型。平话人群以及其周边人群所属的单倍型频率列在表 3-5 中。可以看出平话汉族 Y 单倍群的主要类型是 O2a*、O3*、O*和 K*，与其周边民族普遍高频率的单倍型 O2a*、O3*和 K 很相近，仅罗城县和金秀县的平话汉族人群在 O3a5 这个单倍型下频率较高，与其余九个支系的汉族人群高频单倍群相同，而周边民族该单倍型频率普遍较低。一般认为，K、O1a 和 O2a*型是侗台语系人群的高频单倍型[16]，而 O3 和 O3a5 则是汉族的高频单倍型[1,3]；此外，苗瑶语系人群则以 K、O2a*和 O3*单倍型为特征[12]。从平话人 Y 染色体单倍型的分布特点可以看出，在父系遗传方面，平话人在遗传结构上保留了一定的汉族血统，但大部分已经被南方少数民族所取代。

2. 线粒体单倍群分型

所有 470 份样本的线粒体单倍群分型以文献[9]为标准，分为 43 个单倍型。各单倍型的频率列表如表 3-6。从表中可以看出，平话人群高频单倍型是 B4a、B5a、M*、F1a、M7b1 和 N*，与其周边原住民族尤其是侗台高频单倍型很接近。唯有金秀县平话人与众不同，它的 F3 单倍型频率明显高于其他类型。这可能是由于其样本量较少造成的结果偏倚。B、F、R9a、R9b、N9a 和 M7 单倍型是中国南方原住居民的主要单倍型。侗台语系人群线粒体 DNA 主要单倍群是 B4a、B5a、F1a、M7b1、M7b*、M*、R9A 和 R9b；苗瑶语系线粒体 DNA 单倍型则以 B4a、B5a、M*、M7b*、C、B4b1、M7b1、F1a、B4*和 R9b 为主。汉族的主要单倍型是 A、C、D、G、M8a、Y 和 Z。由这些数据可以看出，广西平话人的线粒体 DNA 单倍群分布更接近于中国南方原住居民。

3. 平话人群与其他族群的聚类分析

在图 3-9 中，将平话人群的 Y 染色体单倍群分布频率和线粒体 DNA 单倍群分布频

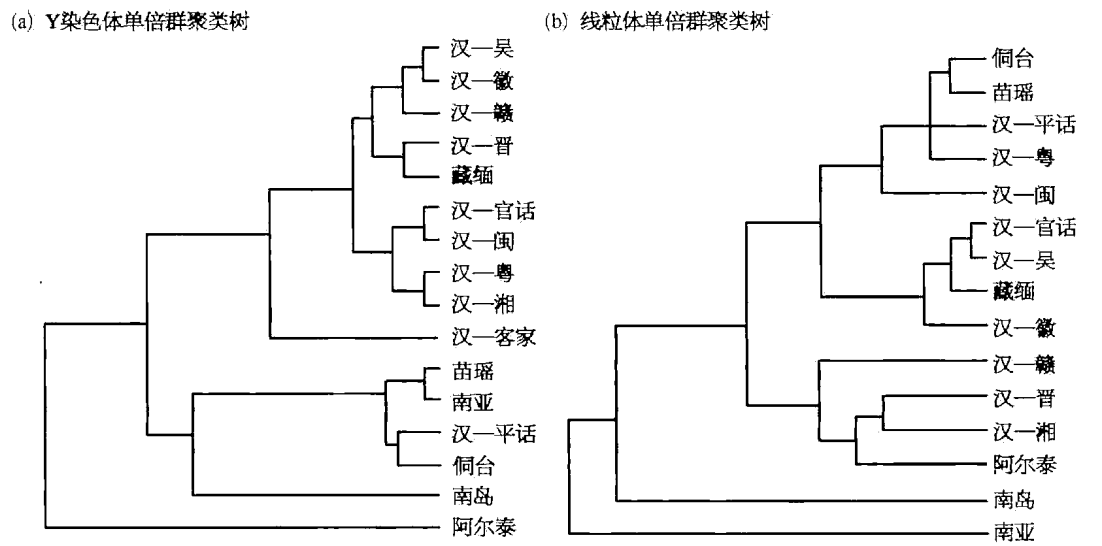


图 3-9 汉族各分支与其他东亚族群的聚类树

表3-6 平话人群及其周边少数民族线粒体DNA单倍群频率分布

编号	样本量	单倍群频率（%）													
		A	B	B4*	B4a	B4b1	B5a	B5b	C	C5	D*	D4	D4a	D5	D5a
P3	39	2.56		2.56	7.69	2.50	5.13				2.56				2.56
P5	48			4.17	4.17	2.08	4.17	6.25		2.08		2.08		2.08	
P1	51				3.92		7.84		1.96						
P4	13				7.69	7.69					7.69	7.69			
P6	46	2.17			2.17	2.17	8.70		2.17		2.17	2.17	4.35	6.52	
D4	67	4.48		2.99	5.97	2.99	16.42	1.49	5.97	1.49	1.49	1.49	1.49	2.99	
D6	78		1.28	6.41	6.41	7.69	7.69		3.84				2.57	5.13	
D1	27	7.41			7.41	3.70					3.70				
D2	72			1.39	11.11	2.78	5.56		5.56				1.39		
H5	29				10.34		6.90		3.45			3.45			

编号	样本量	单倍群频率（%）												
		F*	F1a	F1a1	F1a1a	F1b	F1c	F2*	F2a	F3	M*	G*	G1a	G2*	HV
P3	39	5.13	12.82						2.56	5.13	12.82				
P5	48	2.08	14.58			2.08			6.25	4.17	10.42				
P1	51	1.96	7.84	3.92		1.96			1.96	3.92	17.65				1.96
P4	13									23.08	30.77			7.69	
P6	46	4.35	8.70	2.17		2.17				4.35	19.57			2.17	
D4	67	2.99	11.94		4.48			1.49			14.93		1.49	1.49	

（续表）

编号	样本量	单倍群频率（%）													
		F*	F1a	F1a1	F1a1a	F1b	F1c	F2*	F2a	F3	M*	G*	G1a	G2*	HV	
D6	78	2.56	7.69			2.57	1.28		1.28	1.28	5.13	1.28		1.28		
D1	27		3.70	3.70			3.70			7.41	14.81					
D2	72	4.17	4.17			1.39				5.56	5.56					
H5	29		6.90						6.90	6.90	10.34					

编号	样本量	单倍群频率（%）													
		M7*	M7b*	M7b1	M7b2	M7c	M7c1*	M7c1a	M8a	M9a	N*	N9a	R*	R9b	R9c	Y
P3	39	5.13	5.13	10.26						2.56	7.69	2.56		2.56		
P5	48		6.25	8.33		2.08	2.08		4.17		2.08	8.33				
P1	51	11.76	5.88	9.80					3.92		1.96	1.96	1.96	3.92	1.96	1.96
P4	13										7.69					
P6	46		2.17	2.17							8.70	2.17		8.70		
D4	67	1.49		1.49	1.49		1.49				4.48			1.49		
D6	78	2.56	5.13	6.41	1.28		1.28		1.28		5.13	1.28	2.57	7.69		
D1	27	7.41	11.11	14.81	3.70				3.70					3.70		
D2	72	8.33	5.56	18.06		1.39	9.72	1.39				2.78		1.39		2.78
H5	29	3.45	6.90	17.24	3.45						6.90	3.45		3.45		

率数据分别结合文献报道的6大语系的相应数据，绘制了两幅语系人群聚类树状图。把不同来源的人群按照其所属的语系归为16类，其中除汉族细分为10个亚类外，其余都按照6大语系分类。分别取各语系或语族人群各个单倍型频率的平均值作为该语系该单倍型频率，然后与平话人数据进行聚类树分析。其中用于比较的Y染色体的数据来源于参考文献[1]、[6]、[11]、[17]；用于比较的线粒体DNA单倍群数据来自参考文献[6]、[12]、[13]和[17]。使用SPSS10.0软件的hierarchical cluster，聚类方法选择furthest neighbor。

从根据Y染色体单倍群分布频率绘制的聚类树中可以看出，除平话人一支外，汉族人群的其余9大语支都聚集在一块。这表明汉族遗传结构的高度一致性，且这种一致性突出体现在父系遗传上。同时我们还可以看到，地理分布上邻近的语支往往聚在一起，比如汉族的吴语、徽语和赣语人群，粤语和湘语人群。这提示依据语言学对人群的分类与人群的内部遗传结构相似性并不完全一致，地理分布对人群内部遗传结构有着显著的影响。平话人主要居住于广西，在地理分布上散居于南方原住民族之中。而广西主要为壮族聚居区，原住居民以侗台语系人群为主，间以苗瑶语系人群。从聚类树中可以看出，平话汉族人群与侗台语系人群聚集在一起，而与其他汉族支系关系较远。地理分布对人群内部遗传结构的影响更加突出地体现在据线粒体DNA单倍群分布频率绘制的聚类树中，地理分布上相邻的侗台、苗瑶语系、汉族粤语人群、汉族闽语人群和平话人群这几个南方人群聚集在一起。

4. 汉族和南方少数民族的主成分分析

主成分分析的数据来源同以上聚类分析。从图3-10中可以看出，南方汉族与北方汉族在第一主成分和第二主成分上都聚集在一起，尤其在根据Y染色体单倍群分布频率绘制的主成分分析图中这种趋势更加明显，而据线粒体DNA单倍群分布频率绘制的主成分分析图中北方汉族相对集中，南方汉族比较分散。但总的来说都体现了汉族内部遗

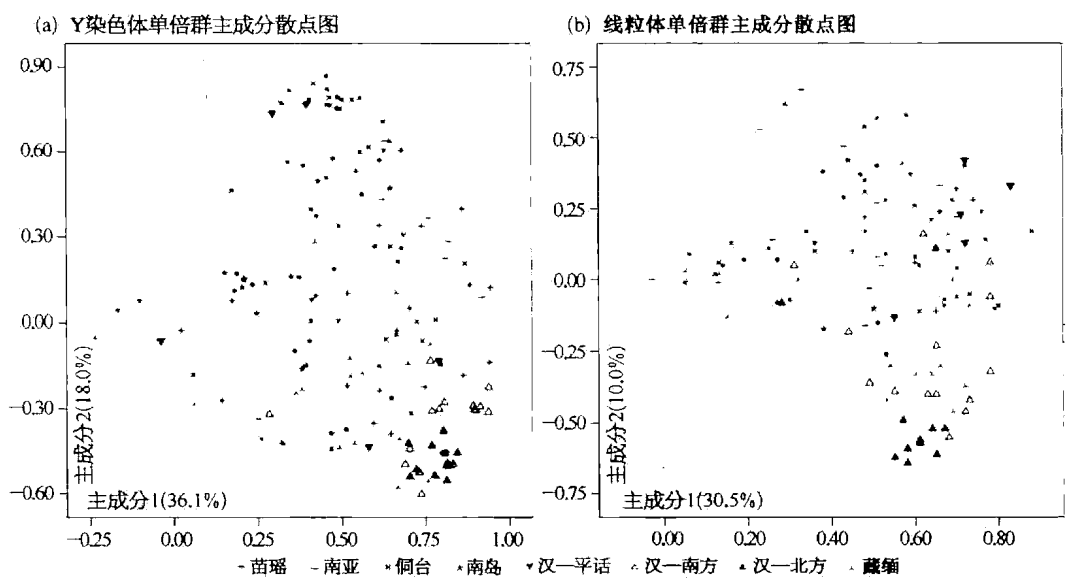


图3-10　汉族各分支与其他东亚族群的主成分分析图

传结构的高度一致性。南方原住少数民族以侗台、苗瑶和南语族为主，他们在第一主成分和第二主成分上非常分散，体现了南方少数民族内部遗传结构的多样性。本节的主要研究对象是广西平话人群，在图 3－10a 中仅罗城县和金秀县平话人群与汉族人群相对靠近，其余平话人群则散落于南方少数民族之中。而图 3－10b 中，几乎所有平话汉族人群都分散在南方少数民族之中。这体现了人群父系、母系遗传的差异，这说明女性在群体中的流动性要大于男性，所以平话人群的女性成分更多地受到其他民族影响。

5. 平话人群主要单倍群的网络结构分析

我们选择 Y 染色体单倍群中的 O2a、O3＊和 O3a5a 这 3 个主要类型，使用 Y 染色体的 6 个短串联重复序列位点 DYS19、DYS389、DYS390、DYS391、DYS392 和 DYS393 绘制平话人群与侗台、苗瑶以及其他汉族人群关系网络结构图，如图 3－11。用于比较分析的数据来源于参考文献[6]、[11]、[17]和[18]。同时选择平话人群线粒体 DNA 主要单倍型 B4＊、B4a、B5a、N9a、F＊、F1a、R9b 和 M7，根据线粒体 DNA 高变 1 区的突变类型，绘制平话人群与汉族其他人群以及其他语系人群的关系网络结构图，如图 3－12。用于

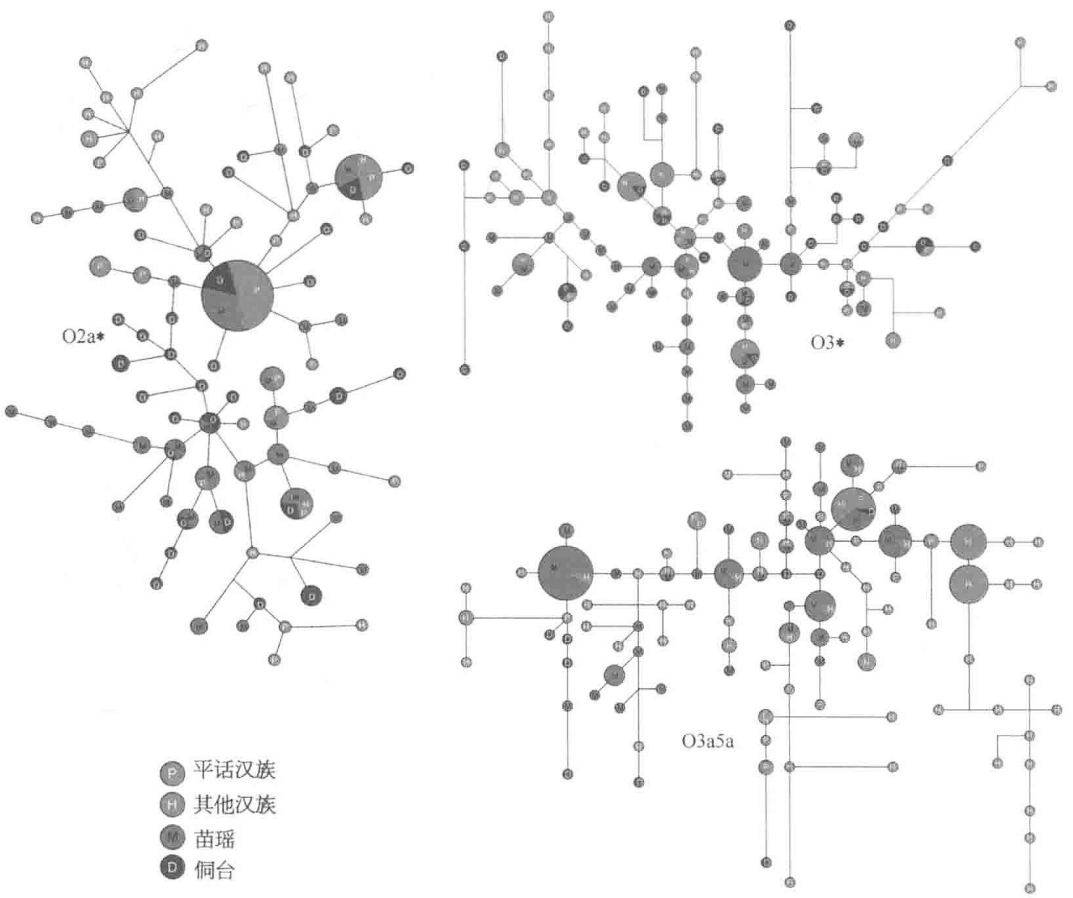


图 3－11 平话汉族与其他汉族群体以及南方原住民族 Y 染色体主要单倍群 O2a＊、O3＊和 O3a5a 的网络结构分析

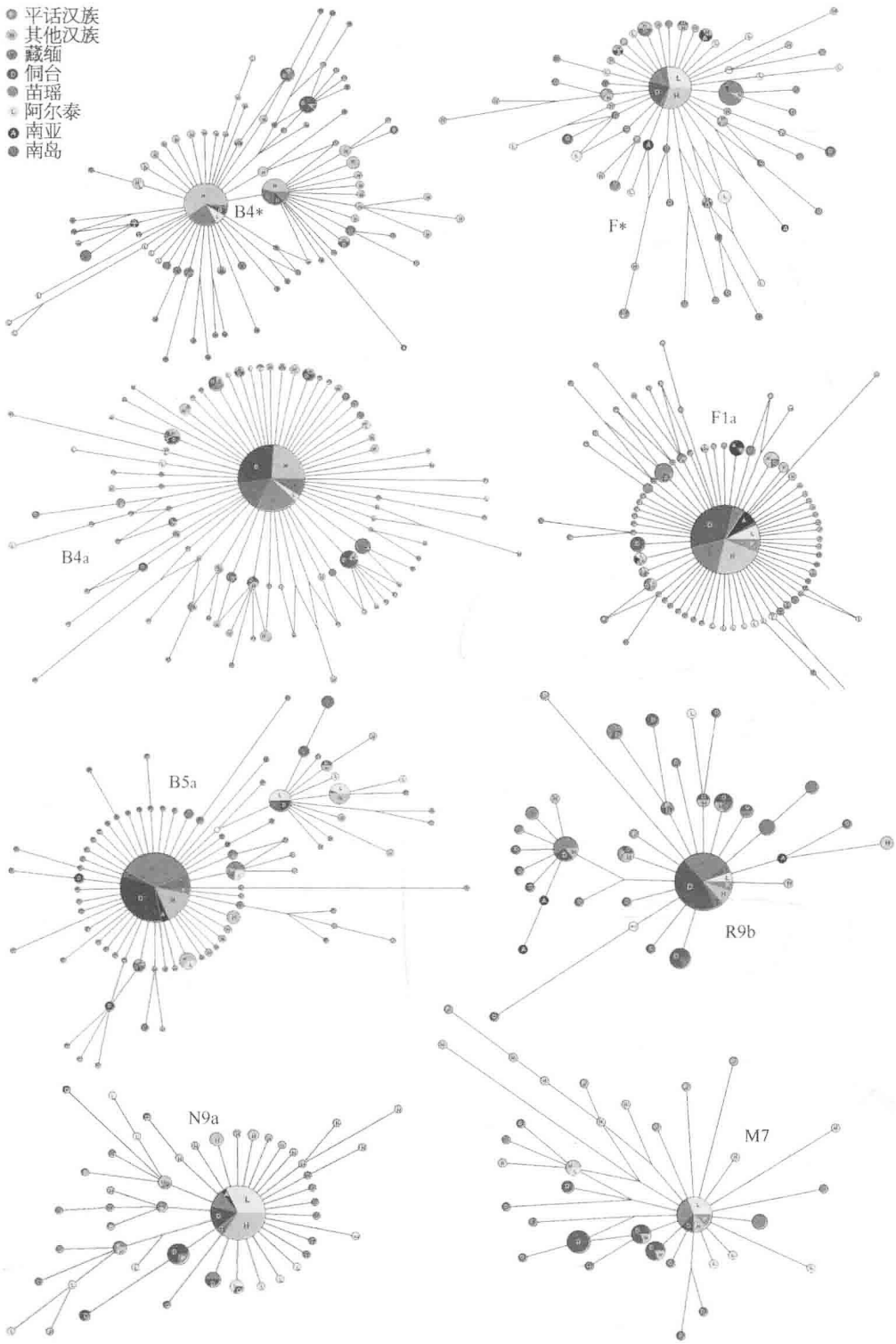


图 3－12　平话汉族与其他东亚群体线粒体主要单倍群网络结构分析

比较分析的数据来源于参考文献[6]、[12]、[16]和[19]。由图 3－11 可以看到，在 O2a 这个单倍群下绘制的网络图中，上下两半网络结构分别形成两个中心，上半部的中心是以侗台、苗瑶和平话人群为源头，汉族人群大多属于这个中心，并有部分在其下形成独特的分支，这种独特的分支绝大多数都是福建汉族人群。下半部中心则是以侗台和苗瑶为主。再往下细分，可以看出，侗台和苗瑶语系人群相对更接近源头。平话人与苗瑶、侗台人群共享的 STR 单倍型多于与其他汉族共享的单倍型，在相连关系上也是与苗瑶、侗台人群更为密切。除共享 STR 单倍型外，平话人其余 STR 单倍型更靠近网络图的支端末节，提示这部分平话人群由侗台、苗瑶人口分化出来的可能性。比较平话人群与两个原住人群的关系，可以发现其与苗瑶的关系更为密切，共享和连接的单倍型较多。以上这些特点在 O3＊ 和 O3a5a 单倍型下绘制的网络图中也有相应的体现。在 O3＊ 网络结构中，苗瑶人群 STR 单倍型处于中心位置，而汉族与侗台民族相对处于外围。平话人群在 O3＊ 网络结构中更多地与侗台民族共享单倍型。而在 O3a5a 网络图中平话人群则与其他汉族人群关系相对紧密，这部分平话人群主要来自广西罗城县，这个结果与前面分析的罗城县平话人在父系遗传结构上更接近汉族的结论是一致的。

图 3－12 为线粒体 DNA 网络结构图。各个网络结构形状多如太阳，由一个大的中心单倍型扩散出众多分支。几乎六大语系的人群都分享了每一个中心，而后由此中心发生大的扩散。各大群体之间都有一定程度的交流，这种交流并不局限于某两个或某几个群体之间。从图中看出，平话人群个体部分与汉族人群共享同一单倍型或者单倍型邻接。进一步验证女性在群体中的流动性大于男性，因此从母系遗传角度看，平话人群的内部遗传结构少部分保留了一些汉族特征，但大部分仍倾向于南方原住民族。

表 3－7 平话汉族人群的混合分析

体　　系	Y 染色体		线粒体 DNA	
亲本群体	其他汉族	南方民族	其他汉族	南方民族
混合比例(Admix)	－0.080 8	1.080 8	－0.150 3	1.150 3
混合比例(LEADMIX)	－0.305 5	1.305 5	0.188 8	0.811 2
网络结构中共享单倍型的个体数	2	29	8	11
网络结构中相连的个体数	25	65	58	61

6. 平话人群混合比例分析

通过单倍群频率可以估计祖先群体对混合群体的相对贡献。平话人的混合比率用 Admix2.0 和 LEADMIX 软件计算。设定混合发生的年代是 2 500 年。平话人群的祖先群体设为两类：一是汉族，将所有汉族个体分别按照不同单倍型汇总，Y 染色体数据共 1 693 个个体，线粒体 DNA 数据共 2 159 个个体；二是南方原住民，以侗台和苗瑶为主汇总，Y 染色体共 1 677 个个体，线粒体 DNA 共 2 374 个个体。分别从父系和母系的角度计算汉族和南方原住民对平话人群的相对贡献率。表 3－7 中可以看到两种软件计

算结果是基本一致的。即无论是Y染色体还是线粒体DNA的南方原住民混合比例都远远高于汉族，这说明平话人群极有可能来自南方原住民，而非汉族后裔。在网络结构中，与平话人群共享单倍型和相邻单倍型的人群也有群体差异性。在Y染色体网络结构中与南方少数民族共享或相邻单倍型的平话人群个体数远远高于与汉族共享或相邻单倍型的个体数。而在线粒体DNA中与两种群体相似的平话人群个体数差别不大。单倍群相似统计中的差异显著性用 t 检验，发现Y染色体差异显著（$P<10^{-12}$）；线粒体DNA差异不显著（$P=0.326$）。由此看出，在Y染色体上，平话人连接于南方民族显著多于连接于汉族。而线粒体则不那么明显。这些结果反映了汉族在平话人群中的比例极低，且仅限于女性。

3.3.4　讨论

1. 汉族遗传结构一致性所体现的大民族形成规律

考古学、历史学以及近年来的分子人类学的研究结果都表明，在距今5 000～6 000年，华夏族从汉藏语系群体中分化出来集居在黄河中上游盆地，这就是汉族前身。夏、商、周是华夏族形成的主要历史时期。随着生产力的发展，人口增长，必然导致向外扩张。汉族人口的扩张主要方向是中国东部和南部。在距今2 000多年间，有史记载的汉族人群大规模南迁不下3次，如西晋约90万人口的迁徙，唐代更大规模的南迁，以及宋代近500万人口的迁徙[20]。在汉族移民的过程中，不可避免地涉及汉族与中国东部、南部原住民族的同化与被同化问题以及民族融合问题。周代汉族融合了东夷，秦汉融合了楚越，南北朝五胡乱华，唐代突厥进犯，元代融合了契丹和女真族。但这些事件形成的外来成分都没有在汉族中占多数或局部多数。所以，汉族的遗传结构至今一致性较强[1]。由此看出，人口基础是同化外族的前提，大民族一般是靠人口扩张形成的，如果没有占优势的人口，就没有同化其他民族的可能。其他民族在人口占多数的情况下，文化上被人口占少数的民族同化的可能性极小。

2. 平话人群遗传结构特殊性的背景

本节报道平话人群的遗传结构与南方原住民更接近，体现出他们不是汉族人口扩张形成的分支，而是其他少数民族被同化形成的汉族人群。在广西，少数民族以侗傣语系人群为主，间以苗瑶语系人群，他们的人口占有局部优势。但是广西原住民在全国来说毕竟是少数，在中国主流的汉族文化的优势影响下，部分得以同化成为汉族。而这种同化是汉族同化外族的特例，是一种语言、文化以及自身认同感上的同化，在遗传结构上并没有占优势的汉族成分。历史上，汉族文化传播在广西地区非常悠久，壮、侗、瑶等族深受影响，语言中有很大成分的汉语借词，甚至一千多年前离开中国的泰族还是有很明显的汉族文化影响。平话人群处于汉族进入岭南的最初通道上，秦始皇开凿灵渠，通过桂北进入岭南。这条路线虽然不是汉族移民的最终留居地，但是途中的人群却汉化最深，成为汉族的一部分。这是平话人群被同化的社会历史背景。

3. 平话人群保持语言文化独特性的原因

自秦汉以来，几次较大规模的军事移民将中原汉族带到广西境内。如公元前214年，秦经略岭南；公元1053年狄青平农智高；1368年明代廖永忠平桂等[21]。但历代移居广西的汉族人口相对人口占绝对优势的原住民来说毕竟太少，他们最终未能取代原住民，反而淹没在广西少数民族人群之中。然而在全国占有绝对优势的汉语言文化却得以伴随着人口的迁徙和政治上的强势传入，并对广西原住民产生了深远的影响。从遗传结构上看，平话人群并非汉族移民的后裔，而是被汉族同化了的广西原住民。因此平话人群在文化上既有鲜明的汉族特征，又保持了南方少数民族的特色，如服饰、婚葬习俗等[22]。在语言上平话有着鲜明的南方少数民族的语言特征，平话中的边擦音声母[ɬ]不是汉语历史语音遗存，而是侗台语底层成分。桂南平话中普遍存在的入声四分也是壮侗语族语言底层的体现。同时平话中又带有典型汉语语音[23]。据此看来，平话形成的主体就是广西原住民，他们在学习汉语的过程中形成了以土著语言为底层的混合语。平话人独特的语言文化特征并不是一次形成的，而是不同历史时期的汉族移民所带来的具有不同时代特点的汉语言和文化的沉淀与本土民族语言文化的融合[24]。

参考文献

[1] Wen B, Li H, Lu D, et al. Genetic evidence supports demic diffusion of Han culture. Nature, 2004, 431: 302－305.

[2] 徐杰舜. 汉民族发展史. 成都：四川民族出版社，1992.

[3] Su B, Xiao C, Deka R, et al. Y chromosome haplotypes reveal prehistorical migrations to the Himalayas. Hum Genet, 2000, 107: 582－590.

[4] 徐杰舜. 平话人的形成及人文特征-华南汉族族群研究之一. 广西大学学报(哲学社会科学版)，1999,21(5)：103－108.

[5] 詹伯慧，崔淑慧，刘新中，等. 关于广西"平话"的归属问题. 语文研究，2003,88(3)：47－52.

[6] Li H. Genetic structure of Austro-Tai populations. Shanghai: PhD dissertation of Human Biology, Fudan University, 2005.

[7] Shi H, Dong Y L, Wen B, et al. Y chromosome evidence of southern origin of the East Asian-Specific haplogroup O3－M122. Am J Hum Genet, 2005, 77: 408－419.

[8] Yao Y G, Kong Q P, Bandelt H J, et al. Phylogeographic differentiation of mitochondrial DNA in Han Chinese. Am J Hum Genet, 2002, 70: 635－651.

[9] Kivisild T, Tolk H V, Parik J, et al. The emerging limbs and twigs of the East Asian mtDNA tree. Mol Biol Evol, 2002, 19: 1737－1751.

[10] Li H, Wen B, Chen S J, et al. Paternal relationships between Austronesian and Daic populations. BMC Evol Biol, 2008, 8: 146.

[11] 奉恒高. 瑶族通史. 北京：民族出版社，2007.

[12] Wen B, Li H, Gao S, et al. Genetic structure of Hmong-Mien speaking populations in East Asia as revealed by mtDNA lineages. Mol Biol Evol, 2004, 22: 725－733.

[13] Wen B, Xie X H, Gao S, et al. Analyses of genetic structure of Tibeto-Burman populations

reveals sex-biased admixture in southern Tibeto-Burmans. Am J Hum Genet, 2004, 74: 856 - 865.

[14] Dupanloup I, Bertorelle G. Inferring admixture proportions from molecular data: extension to any number or parental populations. Mol Biol Evol, 2001, 18: 672 - 675.

[15] Wang J. Maximum-likelihood estimation of admixture proportions from genetic data. Genetics, 2003, 164: 747 - 765.

[16] Li H, Cai X Y, Winograd-Cort E R, et al. Mitochondrial DNA diversity and population differentiation in Southern East Asia. Am J Phys Anthropol, 2007, 134: 481 - 488.

[17] 李辉. 分子人类学所见历史上闽越族群的消失. 广西民族大学学报(哲学社会科学版), 2007, 29(2): 42 - 47.

[18] Chen J, Li H, Qin Z D, et al. Y chromosome genotyping and genetic structure of Zhuang populations. Acta Genetica Sinica, 2006, 33: 1060 - 1072.

[19] Hill C, Soares P, Mormina M, et al. A mitochondrial stratigraphy for island Southeast Asia. Am J Hum Genet, 2007, 80: 29 - 43.

[20] 葛剑雄, 吴松弟, 曹树基. 中国移民史. 福州: 福建人民出版社, 1997.

[21] 袁善来, 黄南津. 广西强势语言(包括汉语方言、普通话)更替及其外部原因. 柳州师专学报, 2005, 20(3): 64 - 68.

[22] 徐杰舜. 平话人的形成及人文特征(续). 广西大学学报(哲学社会科学版), 1999, 21(6): 93 - 97.

[23] 李连进. 平话的历史. 中国语文, 2000, 6: 24 - 30.

[24] 潘悟云. 语言接触与汉语东南方言的形成//邹嘉彦. 香港语言接触圆桌会议论文集. 香港: 香港城市大学出版社, 2002.

3.4 藏族人群的双重起源

藏族是对居住在喜马拉雅山脉地区以及所用语言为汉藏语系藏缅语族藏语支的人群的通称[1]。藏语支主要分为3个方言区,分别是中部藏语、北部藏语和南部藏语。中部藏语和北部藏语的使用者主要分布在中国(包括西藏、四川和云南),而南部藏语使用者主要分布在喜马拉雅地区南部,如印度、不丹和尼泊尔(图3-13)。从历史上来看,藏族是从东亚北部起源的,这已经通过经典的常染色体遗传标记[2]以及微卫星标记[3]实验获得验证。但是,藏族群体中广泛存在YAP+,这是一种*Alu*插入多态性,除日本外在东亚群体中只零星出现[4,5],这一现象与先前的结论相矛盾,提示了藏族群体很可能受到从东南亚地区来的一批基因的明显影响[6]。我们最近一项关于3个藏族群体的研究通过运用19个Y染色体双等位基因标记揭示了藏族群体的多重遗传起源。

我们对3个藏族群体进行了采样,两个来自西藏中部(包括拉萨和日喀则),另一个来自云南西北部的香格里拉。前两个群体被称为卫藏,属于中部藏语分支的使用者;后一个是康巴藏,所用语言为北部藏语中的康巴藏语。总共19个Y染色体双等位基因标记被用于这3个人群的分型。这些标记的相关说明可以在1999年宿兵等发表的文献中找到[5]。

在这3个藏族群体中,我们观察到了10种Y染色体单倍型(表3-8)。在拉萨和日

喀则的人群中没有发现显著差别。明显可见，不论卫藏和康巴藏群体都由两种单倍群占主要成分，分别是 YAP+（单倍群 D）以及 M122C（单倍群 O3），这两种单倍群包括了他们基因库的 80%，也暗示了这两类群体的共同起源。

表 3-8 藏族和汉族群体中的 Y 染色体单倍群分布

群体	单倍群频率(%)													
	C-M130	D*-YAP	D1-M15	F-M89	K-M9	O3-M122	O3-M7	O3-M134	O1-M119	O2-M95	O2-M88	P-M45	Q1-M120	样本量
卫藏	8.7	23.9	17.4	4.3		4.3		34.8		2.2		2.2	2.2	46
康巴		14.8	29.6	3.7	14.8	7.4		29.6						27
南方汉族	8.1	0.4		1.4	12.7	25	1.8	27.6	17.3	3.5	0.7	1.4		283
北方汉族	8.5			2.4	21.9	29.3		23.2	9.8			4.9		82

YAP+是一种起源于西亚、从东南亚进入东亚的古老多态性[6]，对藏族人群有主要的遗传贡献（在卫藏中占 41.3%，在康巴藏中占 44.4%）。我们先前的研究表明 M122C（定义为单倍群 O3）是东亚群体，尤其是在中国汉族中（平均有 54.1%）的主要成分[5]。但是它在包括中亚的世界其他群体中几乎不存在（R. S. Wells，个人通讯）。因此，藏族

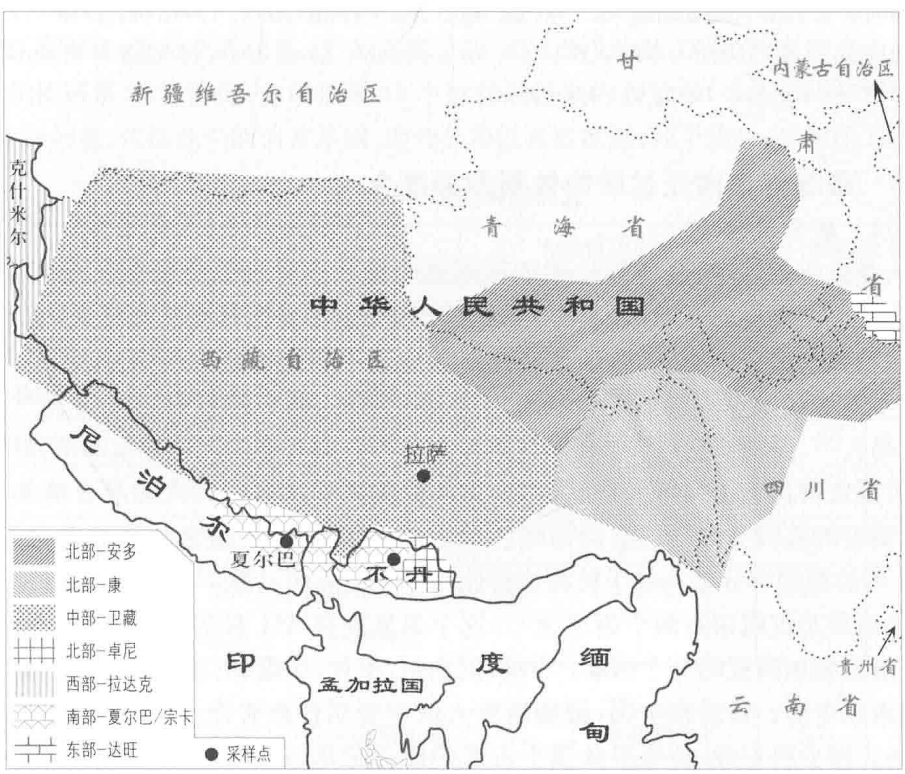


图 3-13 藏语各方言区的分布示意图

人群中M122C的高频存在(>35%)反映出他们具有一脉来自东亚的遗传根源,其中不包含YAP+。在康巴藏群体中存在的单倍群K很可能反映了他们与那些含有相当比例单倍群K的云南原住群体的密切交流[5]。

总之,藏族Y染色体很可能来源于两个不同的基因库:一个来自东南亚,另一个来自东亚。我们的前期工作为对汉藏语系人群进一步系统化地开展起源以及史前迁徙的遗传研究提供了逻辑基础。今后可以运用更多的位点进行研究,以便解释遗传漂变对于所观察的模式产生的可能影响。

参考文献

[1] Matisoff J A. Sino-Tibetan linguistics: present state and future prospects. Ann Rev Anthropol, 1991, 20: 469-504.

[2] Du R F, Xiao C J. Genetic distances between Chinese populations calculated on gene frequencies of 38 loci. Science in China Series C, 1997, 40: 613-621.

[3] Chu J Y, Huang W, Kuang S Q, et al. Genetic relationship of populations in China. Proc Natl Acad Sci USA, 1998, 95: 11763-11768.

[4] Hammer M F, Spurdle A B, Karafet T, et al. The geographic distribution of human Y chromosome variation. Genetics, 1997, 145: 787-805.

[5] Su B, Xiao J H, Underhill P, et al. Y chromosome evidence for a northward migration of modern humans in East Asia during the Last Ice Age. Am J Hum Genet, 1999, 65: 1718-1724.

[6] Altheide T K, Hammer M F. Evidence for a possible Asian origin of YAP+ Y chromosomes. Am J Hum Genet, 1997, 61: 462-466.

3.5 藏缅群体南迁过程的性别差异混合

3.5.1 研究背景

在人类进化过程中,遗传分化和地理隔绝的群体有时会发生接触从而形成混合群体。对这种混合事件的认识,在遗传学、人类学和历史学方面都有巨大意义,因为它既揭示了这些群体的遗传结构,又有助于通过混合连锁不平衡地定位疾病基因[1-5]。现已证明,遗传学方法在揭示人群的混合状况方面是严谨的,研究范例有吉普赛人[6]、土耳其人[7]、冰岛人[8]、美洲原住民[9-11]和非裔美洲人[12-15]。更有趣的现象是亲本群体男性与女性世系对混合群体有不均等的贡献,例如美洲原住民和非裔美洲人的混合群体,这被称为"性别偏差的基因交流"或"定向婚配"[9,16]。父系遗传的Y染色体和母系遗传的线粒体DNA单倍型频率分布有助于检测这种性别差异的混合过程[9,11-16]。

藏缅语族是汉藏语系两个语族之一,这个语族现存351种语言,其使用者主要分布在东亚、南亚和东南亚的9个国家:中国、尼泊尔、不丹、印度、巴基斯坦、缅甸、孟加拉国、泰国、越南和老挝。目前在中国,藏缅语族人群主要居住在青海,以及西藏、四川、云南、湖南等地。据史料记载,藏缅群体属于古老的氐羌部落,最初居住在中国西北。在春秋时期(距今约2 600年),藏缅群体展开了大规模南迁,通过藏缅走廊,进入南亚语系人群

的分布区，该区域也可能分布着侗傣语系和苗瑶语系的人群，这3个人群是南方原住人群[17]。这一记载与遗传学证据相一致，根据Y染色体标记，几乎所有藏缅群体都共享一个高频率的M122－C突变，和非常高频的来自M122－C突变的M134缺失[18]。根据这两个位点等位基因频率分布状况看来，这些南迁群体的混合状况略有差异，其中的遗传效应有待于进一步揭示。

本研究涉及23个藏缅群体的965个个体，分析了10种Y染色体SNP标记，以及21个藏缅群体754个个体的线粒体DNA高变1区（HVS－1）的序列变异和一系列编码区的分型变异。研究发现，来自亲本群体（北方移民和南方原住群体）的男女不均等贡献，在形成现存南部藏缅群体基因库的过程中发挥了重要作用。

3.5.2 材料和方法

1. 样本

收集来自中国云南、青海和湖南的15个藏缅群体622名不记名的血液样本，由酚-氯仿法提取基因组DNA。其他的数据采集自已公布的Y染色体[19-21]和线粒体DNA多样性[22-24]的报告。这些数据使最终的分析样本量增加到Y染色体样本965个（23个藏缅群体），线粒体DNA样本754个（21个藏缅群体）。这些样本涵盖了中国大多数藏缅群体。在本节中，把在云南、四川和湖南的藏缅群体（不包括藏族）称为“南部藏缅”。

此外，本研究使用的Y染色体数据还包括南亚语系4个群体、侗傣语系30个群体和苗瑶语系23个群体，以及北方汉族17个人群[18,19,21,25]，而线粒体DNA数据来自南亚语系4个群体、侗傣语系12个群体、苗瑶语系12个群体以及北方汉族10个人群[22-24,26-28]。人群研究的详细信息，包括他们的语言系属、地理分布以及数据源，列于表3－9和图3－14。

表3－9 研究人群信息

	人群	语言分支	所在地	Y染色体			线粒体DNA	
				样本量	参考		样本量	参考
藏族								
1	藏族-1	喜马拉雅语支	青海	91	本研究		56	本研究
2	藏族-2	喜马拉雅语支	西藏	75	[29]			
3	藏族-3	喜马拉雅语支	西藏	46	[20]			
4	藏族-4	喜马拉雅语支	云南迪庆	27	[20]		24	[22]
5	藏族-5	喜马拉雅语支	云南香格里拉	49	本研究		35	本研究
6	藏族-6	喜马拉雅语支	云南德钦				40	[23]
南部藏缅								
7	傈僳人	彝缅语支	云南西双版纳	52	本研究		50	本研究
8	白族-1	白语支	云南大理	63	本研究，[25]		68	本研究，[24]

（续表）

	人 群	语言分支	所 在 地	Y染色体			线粒体DNA	
				样本量	参考		样本量	参考
9	白族-2	白语支	云南西双版纳	20	本研究		19	本研究
10	哈尼族	彝缅语支	云南西双版纳	34	本研究		33	本研究
11	基诺族	彝缅语支	云南西双版纳	36	本研究,[18]		18	本研究
12	拉祜族-1	彝缅语支	云南思茅	13	[18]		32	[22]
13	拉祜族-2	彝缅语支	云南西双版纳	15	本研究		15	本研究
14	拉祜族-3	彝缅语支	云南澜沧				35	[22]
15	傈僳族-1	彝缅语支	云南福贡	49	本研究			
16	傈僳族-2	彝缅语支	云南贡山				37	[24]
17	纳西族	彝缅语支	云南丽江	40	本研究		45	本研究
18	怒族	彝缅语支	云南贡山	28	本研究		30	本研究
19	普米族	羌语支	云南宁蒗	47	本研究		36	本研究
20	土家族-1	土家语支	湖南西部	68	本研究		64	本研究
21	土家族-2	土家语支	湖南永顺	38	本研究		30	本研究
22	土家族-3	土家语支	湖南吉首	49	[29]			
23	彝族-1	彝缅语支	四川凉山	14	[19]			
24	彝族-2	彝缅语支	云南双柏	50	本研究		40	本研究
25	彝族-3	彝缅语支	云南西双版纳	18	本研究		16	本研究
26	彝族-4	彝缅语支	四川布拖	43	[29]			
27	彝族-5	彝缅语支	云南泸西				31	本研究

注：语言分支根据 Ethnologue 第 14 版。

2. Y 染色体标记

使用 PCR－RFLP 方法检测 10 个双等位基因 Y 染色体标记：YAP、M15、M130、M89、M9、M122、M134、M119、M95 和 M45[18]。这些标记在东亚人群中差异大[30]，按照国际 Y 染色体命名委员会(YCC,2002)的命名原则确认为 10 个单倍群。

3. 线粒体 DNA 标记

线粒体 DNA 的 HVS－1 区用引物 L15974 和 H16488 扩增[23]，纯化的 PCR 产物使用 BigDye Terminator 循环测序试剂盒和 ABI 3100 测序仪进行测序。为扩增包含编码区内的单倍群分型多态位点的多个片段设计了系列引物。PCR 产物用限制性内切酶：10397 AluI、5176 AluI、4831 HhaI、13259 HincII、663 HaeIII、12406 HpaI 以及 9820 HinfI[23,31]进行酶切反应。根据 Kivisild 等的分类[31]，用 HVS－1 区序列信息和编码区

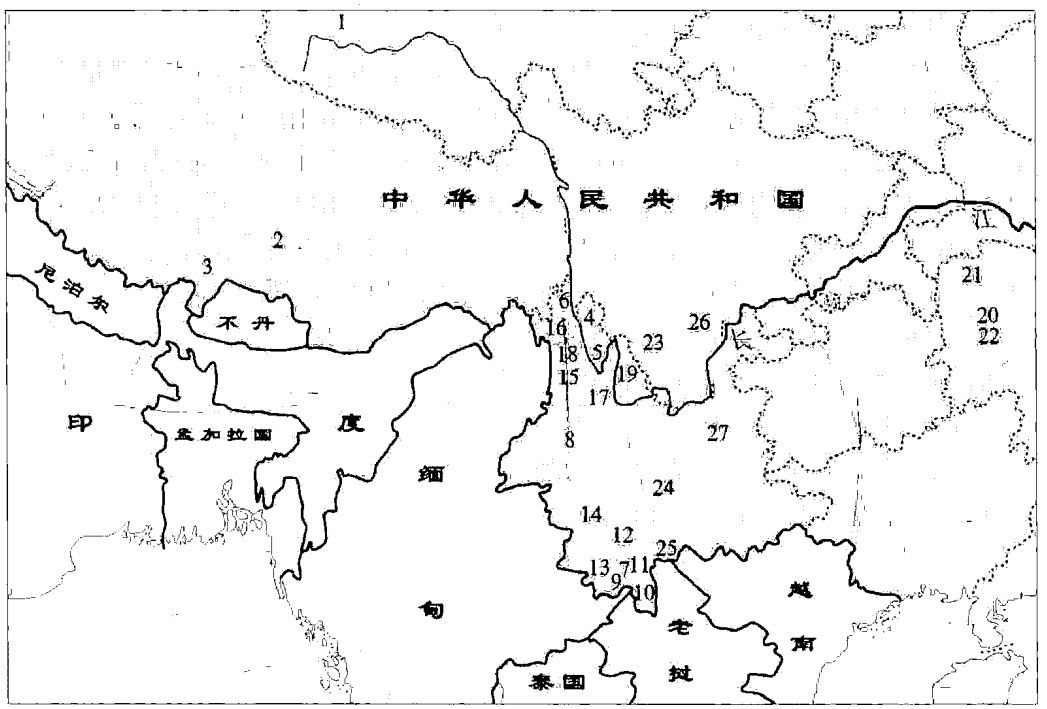


图 3-14 本文调查的藏缅族群的地理分布

编号对应的群体信息见表 3-9。

变异来推断单倍群。14 个藏缅群体 496 个个体的 HVS-1 区序列已提交到 GenBank(检索号 AY397784-AY398279)。

4. 数据分析

将线粒体 DNA 和 Y 染色体单倍群频率用 SPSS 10.0 软件进行主成分分析,主成分分析的结果用前两个主成分的坐标系表示,前两个主成分合计占这些人群 Y 染色体差异的 65.8%和线粒体 DNA 差异的 51.0%。群体的遗传结构由分子方差分析方法(AMOVA)[32]进行研究,使用 Arlequin 软件[33]。

使用 Admix 2.0[34](Admix 网站)和 LEADMIX[35]软件估算南北方人群在南部藏缅群体中的混合比例,分别用到 BE(Bertorelle 和 Excoffier[36])以及 RH(Roberts 和 Hiorns[37])的方法。Wang(2003)的模拟结果指出,这两种估算对单位点数据的偏差较少。亲本人群的选择对混合比例的相应估计是至关重要的[1,38],我们对此尤为注意。通过使用覆盖整个东亚地区的庞大数据集并结合历史记录,使亲本选择尽可能没有偏差。在分析中,藏族(西藏、青海和云南西北部藏族人群的算数平均数)和北方汉族人群的平均单倍群频率(Y 染色体和线粒体 DNA 分列)被作为北方亲本人群(东亚北方人群),而南亚语系、侗傣语系和苗瑶语系人群的平均单倍群频率作为南方亲本人群(东亚南方人群)。

3.5.3　研究结果

1. Y染色体和线粒体DNA单倍群的分布

藏缅群体的Y染色体单倍群频率分布列于表3-10中。除了纳西族和普米族，几乎所有的藏缅人群中，都出现高频率的O3*和O3e单倍群，两个单倍群都携带M122-C突变（单倍群频率之和为20%～86%，平均值为藏缅群体39.5%，南部藏缅群体40.7%），与宿兵等之前的结论一致[18]。在西藏和日本人群中非常频繁地出现的YAP+单倍群D（D*和D1等）在云南西北部的一些人群中普遍存在（纳西族38%，普米族70.2%，远高于藏缅群体中的15.7%和南部藏缅群体的9.8%）。与此相反，单倍群O1-M119和O2a-M95为东亚南方人群普遍的单倍群，在北方非常罕见[19,21]，它们出现在大部分藏缅群体中，频率却非常低（在藏缅群体中平均为10.5%，南部藏缅群体为12.9%），只有纳西族中的高频O2a-M95是例外。然而，单倍群F*-M89和P*-M45本是在中亚和东北亚地区占主流地位的单倍群，却只出现在一部分藏缅群体中，且频率很低（藏缅群体为6.7%，南部藏缅群体为6.1%），除了彝族-2（38%）和拉祜族-1（31%）中的F*-M89较多。大体上，藏缅人群中Y染色体单倍群的分布相较于其南部邻近的南亚语系、侗傣语系和苗瑶语系的人群来说，与东亚北方人群更为相似。

表3-10　Y-SNP藏缅群体中Y-SNP的单倍群频率

人　群	群体大小	突变/单倍群									
		M130	YAP	M15	M89	M9	M122	M134	M119	M95	M45
		C	D*	D1	F*	K*	O3*	O3e	O1	O2a	P*
藏族											
藏族-1	92	13	19	2	13	20	5	13	1		6
藏族-2	75	2	25	12	2	4	1	24			5
藏族-3	46	4	11	8	2		2	16		1	2
藏族-4	27		12		1	4	2	8			
藏族-5	49	1	14	4	1	5	5	17		2	
南部藏缅											
白族-1	61	5	1	3		11	10	21	3	7	
傈僳族-1	49				1	11	2	30		4	1
纳西族	40	1	15			3		1		20	
怒族	28		1			1	4	20		2	
普米族	47	3	33	1		3	1	3	2		1
彝族-1	14			2		6	3	1		2	
彝族-2	50	4	1	5	19	8	5	5	1	2	

（续表）

人 群	群体大小	突变/单倍群									
		M130	YAP	M15	M89	M9	M122	M134	M119	M95	M45
		C	D*	D1	F*	K*	O3*	O3c	O1	O2a	P*
彝族-4	43	1		7	2	15	2	12		4	
傈僳人	52	6			1	18	14	7	2	4	
白族-2	20	4				6	5	3	2	0	
哈尼族	34	4				12	11	5	1	1	
基诺族	36	5			2	13	7	7		2	
拉祜族-1	13	2			4	2	2	2		1	
拉祜族-2	15				1	3	5	1	3	2	
彝族-3	18	2			1	6	5	3		1	
土家族-1	68	10			2	7	20	18	5	6	
土家族-2	38	2		1		9	15	4	6		1
土家族-3	49	12		1		4	15	11	4		2
合 计	964	81	132	46	52	171	141	232	30	61	18

相比之下，线粒体 DNA 单倍群在藏缅群体中的分布更为复杂。几乎所有存在于东亚的单倍群都出现在藏缅群体中（表 3-11）。A(11%)、B(12%)、D(18%)、F(20%)和 M*(18%)是藏缅群体的主要单倍群类型，分别占藏缅群体和南部藏缅群体总线粒体 DNA 谱系的 77.6%和 77.7%。特别是单倍群 D 在大多数藏缅群体中都非常高频，其他的单倍群则表现出有差别的地域性分布。单倍群 A 和 M* 在青海（A 为 21%；M* 为 36%）和云南西北部（A 为 14%；M* 为 22%）的频率高于在云南南部（A 为 5%；M* 为 12%）和湖南（A 为 9%；M* 为 4%）的频率。单倍群 A 是东亚北方人群和西伯利亚人群中更为普遍的线粒体 DNA 单倍群之一[31]，而单倍群 M* 有可能是多个尚未明确亚单倍群的 M 谱系的混合。从由北到南频率递减差异的状况看来，大多数藏缅群体的 M* 型线粒体 DNA 可能属于某个未知单一谱系。单倍群 B 和 F 在东南亚是普遍存在的[23,31]。在藏缅群体中，这两个谱系（B 和 F）在云南南部（B 为 18%；F 为 30%）和湖南（B 为 17%；F 为 21%）的频率高于在青海（B 为 2%；F 为 6%）和云南西北部（B 为 10%；F 为 15%）的频率。

2. 主成分分析得到的人群聚类分析

Y 染色体单倍群的主成分分析结果见图 3-15。东亚北方人群（北方汉族和藏族）和东亚南方人群（南亚、侗傣和苗瑶群体）的第二主成分表现出明显不同的分布。北方人群第二主成分的值是 −0.41±0.05（−0.95～0.31），南方人群第二主成分的值是 0.39±

表 3-11 藏缅群体中线粒体 DNA 的单倍群分布

	人群																				
	1	4	5	6	7	8	9	10	11	12	13	14	16	17	18	19	20	21	24	25	27
A	12	6	4	2	1	4	1	4	1				11	4	9	5	9		4	3	3
B*					2																1
B4*			2	1	1		2	1		3	1	1	1	3	3	1	2	2	5	1	2
B4a	1				4		1	1	1	5	1			5			5		1	1	
B4b1						1		1			1								1		
B5*													1							1	
B5a			4		1	2	1	2	1					3			2	2		1	
B5b								1									3				
C	3		2		1	2		1			1		5	4		8	8	1	1		2
D*	4		3	20	7	9	3	6		2	4	5	3	2	6	6	9	3	4	3	8
D5				2	3	4		2	1	1							1	2	1		2
D5a					2			1									4		1		
F*					1	1	1					9					2	6	2		1
F1a	1	1			6	6	2	6	2	10	3	18	1	8			2	4	2	1	
F1b		1				2	2	1	1		1		2	2		1	2	1	2	2	3
F1c	2					4					1						1	2			
F2a			1			4				1			4	1	5	2				1	
G	2													3	1	2	1	1	1		1

（续表）

	人群																				
	1	4	5	6	7	8	9	10	11	12	13	14	16	17	18	19	20	21	24	25	27
G2	1					5											1	1	1		
G2a	2				2																
G3	1		2													1	1				
M*	20	6	15	12	7	15	3	4	5	6		1	7	7	1	7	3	1	8	1	3
M7*						1												1			
M7b*					1	2					1						1		2		
M7b1					4	2	1		3					1			1	1	3	1	2
M8a					1						1								1		
M9									1												
M9a	5	4	1	3		3	1										1	1			1
N*	1				1	2											3				2
N9a	1								2							1	2	2			
R10					1	2		1		4											
R9a							1					1		1			2				
Y								1													
Z		6	1		1								2	1	5	1					

注：人群编号与表3-9中一致。

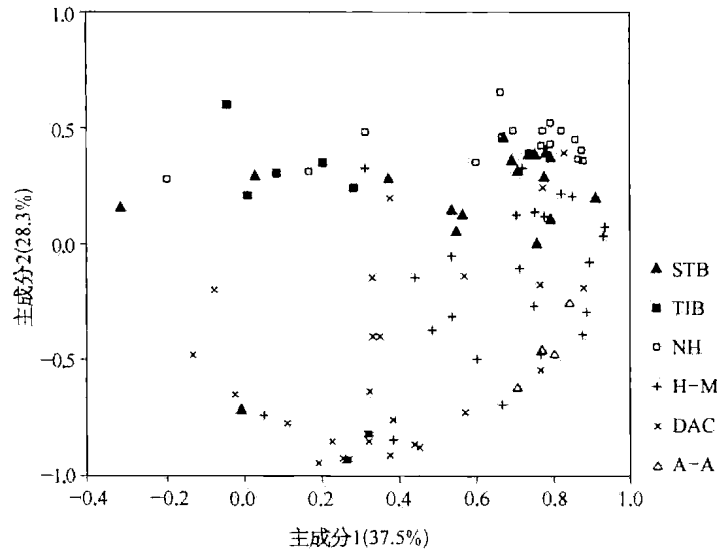


图 3-15　Y 染色体 SNP 单倍群频率的主成分分布图

STB 为南部藏缅；TIB 为藏族；NH 为北方汉族；H－M 为苗瑶语系人群；DAC 为侗傣语系人群；A－A 为南亚语系人群。

0.02(0.22～0.63)。南部藏缅群体在它们之间，第二主成分的值为 0.23±0.03(−0.04～0.43)，除去一个异常值(纳西族，主成分 2 = −0.71)，与这些人群混合的历史记录相符。

然而，线粒体 DNA 的主成分分析结果展示的情形不同(图 3－16)。东亚南方人群和东亚北方人群仍能凭借第二主成分分开，但南部藏缅群体在南方和北方人群中分散分

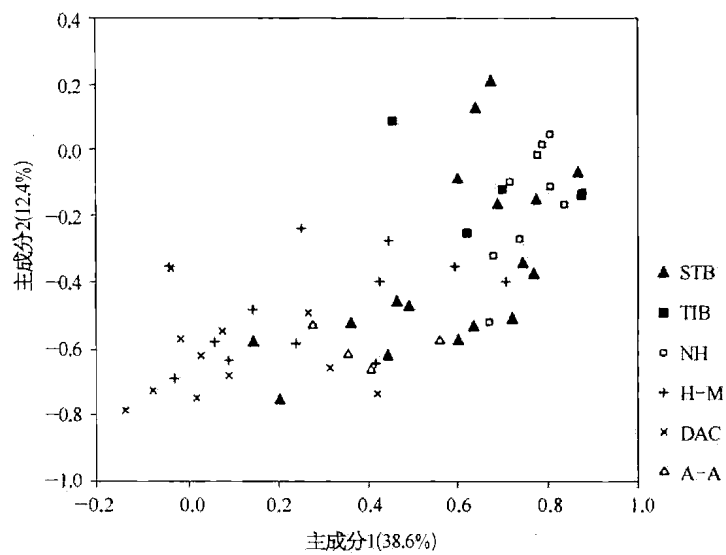


图 3-16　线粒体 DNA 单倍群频率的主成分分布图

STB 为南部藏缅；TIB 为藏族；NH 为北方汉族；H－M 为苗瑶语系人群；DAC 为侗傣语系人群；A－A 为南亚语系人群。

布。对线粒体 DNA 的这个观察也符合这些人群混合的性质，既有东亚北方人群也有东亚南方人群的贡献。

3. *利用分子方差分析估算分层结构*

应用分子方差分析估算南部藏缅群体和其他东亚人群之间的遗传分化程度，并研究南部藏缅内部的遗传结构(表 3－12)。Y 染色体的计算结果，南部藏缅群体和东亚南方人群之间的 F_{CT}(组间发散性)差异显著($P=0.002$)，而南部藏缅群体和北方人群(NEAs)之间的差异却不显著($P=0.221$)。南部藏缅群体和东亚南方人群之间的 F_{CT} 是南部藏缅群体和东亚北方人群的 F_{CT} 的 10 倍以上(0.029 对 0.002)，表明相比东亚南方人群，南部藏缅群体的亲本谱系与东亚北方人群更为相似。然而线粒体 DNA 的结果，南部藏缅/北方和南部藏缅/南方的 F_{CT} 值都显著且几乎等同(分别为 0.005 与 0.007；P 值分别为 0.003 和<0.001)，表明南部藏缅群体和两个东亚群体之间的相似的发散性。在藏缅群体中，Y 染色体的群体间发散性(F_{ST})约是线粒体 DNA 的群体间发散性 2 倍(0.115 和 0.057)，表明女性世系的群体间分化比相应的从夫居群体中的男性更广泛。这与先前对全球人群的研究相一致[39]，只是在藏缅群体中的反差程度较为温和。

表 3－12 分子方差分析结果

群 组	群组数量	人群数量	F_{CT}(p)	F_{CS}(p)	F_{ST}(p)
线粒体 DNA					
南部藏缅/东亚北方	2	31	0.005 (0.003)	0.035(0.000)	0.040 (0.000)
南部藏缅/东亚南方	2	45	0.007 (0.000)	0.035 (0.000)	0.042 (0.000)
南部藏缅/东亚北方/东亚南方	3	59	0.010 (0.000)	0.033 (0.000)	0.043 (0.000)
藏缅群体为一组	1	21	...	...	0.057 (0.000)
Y 染色体					
南部藏缅/东亚北方	2	40	0.002 (0.221)	0.084 (0.000)	0.086 (0.000)
南部藏缅/东亚南方	2	77	0.029 (0.002)	0.136 (0.000)	0.161 (0.000)
南部藏缅/东亚北方/东亚南方	3	98	0.035 (0.000)	0.110 (0.000)	0.141 (0.000)
藏缅群体为一组	1	23	...	...	0.115 (0.000)

注：F 统计的 P 值由 3 000 种排列组合得到，没有人口结构的假设。

4. *南部藏缅群体融合计算*

通过使用两种不同方法，根据线粒体 DNA 和 Y 染色体标记，在表 3－13 中列出了东亚南方人群在南部藏缅不同群体中所占的贡献比例。依据 Y 染色体或线粒体 DNA，东亚南方基因库对不同南部藏缅群体的贡献(M)有所不同。用两种方法估算的 M 是相当一致的[皮尔逊相关系数：Y 染色体结果为 0.931($P<0.01$)，线粒体 DNA 结果为 0.928

($P<0.01$)]。在表3-13中，M阴性表示为0，若大于100%表示为1。

表3-13　东亚南方和南部藏缅群体的混合

人　群	M_{BE}		M_{RH}	
	线粒体DNA	Y染色体	线粒体DNA	Y染色体
傈尼人	0.755±0.234	0.511±0.100	0.885	0.521
白族-1	0.878±0.219	0.542±0.134	0.849	0.429
白族-2	0.461±0.373	0.377±0.096	0.700	0.369
哈尼族	0.530±0.271	0.399±0.079	0.725	0.424
基诺族	0.403±0.408	0.327±0.103	0.795	0.363
拉祜族-1	1	0.123±0.163	1	0.078
拉祜族-2	1	0.677±0.186	1	0.831
拉祜族-3	1	ND	1	ND
傈僳族-1	ND	0.200±0.140	ND	0.169
傈僳族-2	0	ND	0	ND
纳西族	0.713±0.252	0.899±0.281	0.729	1
怒族	0	0.096±0.173	0	0.159
普米族	0	0	0	0
土家族-1	0.688±0.183	0.544±0.099	0.520	0.498
土家族-2	1	0.530±0.081	0.914	0.449
土家族-3	ND	0.556±0.089	ND	0.360
彝族-1	ND	0.501±0.199	ND	0.681
彝族-2	0.327±0.262	0.133±0.110	0.382	0.121
彝族-3	0.385±0.395	0.321±0.099	0.491	0.419
彝族-4	ND	0.129±0.114	ND	0.258
彝族-5	0.295±0.275	ND	0.218	ND
平均	0.555	0.381	0.601	0.396

注：南部混合人群（M）由（M_{BE}）[36]及（M_{RH}）[37]估算；BE方法的标准误差由1 000份样本再抽样得到；ND表示缺失数据。

Y染色体和线粒体DNA数据如表3-13所示，除了怒族和纳西族，在其余的南部藏缅群体中，东亚南方人群对女性世系的贡献高于对男性世系（表3-13）。也就是说，现存的藏缅群体中，南方原住民（东亚南方人群）对女性世系做出的贡献（$M_{BE}=55.5\%$；$M_{RH}=60.1\%$）比对男性世系（$M_{BE}=38.1\%$；$M_{RH}=39.6\%$）的大。另外，北方移民对男性世系（约61.9%）的贡献比对女性世系（约44.5%）的大。然而要重点指出，男性和女

性贡献的差异可能被低估。Y染色体单倍群在怒族和纳西族中的偏斜分布(在怒族中高频的O3e-M134,在纳西族中高频的O2a-M95)可能是强烈的瓶颈事件导致的,这也许可以解释其混合估计值的特殊倾向。

3.5.4 结果和讨论

通过研究Y染色体和线粒体DNA标记的单倍群分布,以及对其主成分分析,已经表明现存南部藏缅群体的遗传结构主要由两个亲本群体形成:北方移民和南方原住民(表3-10,表3-11)。南部藏缅群体与他们的亲本群体的线粒体DNA,以及南方藏缅群体和南方原住民的Y染色体已发生显著分歧;而南方藏缅群体和北方移民之间的Y染色体差异不显著(表3-12)。此外,群体混合有性别差异,因为在现存的南部藏缅群体中,北方移民对男性世系的影响更强烈,南方原住民对女性世系做出更大的贡献(表3-13)。

男性和女性不均等的混合现象之前已在美洲原住民和美洲非裔群体中观察到[11-15]。这种差异在过去几个世纪被解释为"定向婚配"的结果[9,16],这意味着从祖先到杂交后代的基因交流是定向的。例如,南美的美洲原住民的混合主要涉及欧洲男性和原住女性,受到欧洲殖民者不对称的性别比例和其他的政治因素的影响[10,40]。

在南部藏缅群体中,性别差异混合模式与美洲的不同,也许不仅在于差异程度,更在于混合机制。在来自美洲原住民和欧洲殖民者的混合群体中,原住民对线粒体DNA的贡献几乎是Y染色体的10倍[10,41],而中国南方原住民线粒体DNA和Y染色体对南部藏缅群体的贡献比例只有约1.5,这种差异远不如在当今美洲原住民中发现的那样极端。据史料记载,藏缅群体祖先向南方移民始于距今约2 600年,是由秦王朝的扩张引发[17,42]。然而,几乎没有历史文献证据表明藏缅群体移民中的男女比例不对称。藏缅群体中这种并不极端的性别差异可能表明藏缅群体向南方的迁徙是有两种性别参与的,而不是征服美洲大陆的仅由男性士兵组成的远征部队。可惜,历史记录中几乎找不到关于迁徙的描述,主要是因为中国历史文献往往以汉族为中心。

研究混合模式可以使我们更加透彻地认识藏缅群体迁徙的历史和人类学特征。例如,不同地理区域之间混合比例上的变化可能与移民到达的时期不同有关,也可能与混合初期该地区亲本群体的民族组成有关。据史料记载,云南南部和湖南南部的藏缅群体(傈僳人、哈尼族、基诺族、拉祜族和土家族)更加古老,而向云南西北部迁徙的藏缅群体(藏族、傈僳族、怒族和普米族)历史相对近期,这当中不包括白族、纳西族和彝族这些云南早期居民[42]。因此,除去纳西族和白族,南方原住民对云南西北部其他藏缅群体的贡献相对小于其对云南南部和湖南群体的贡献(表3-13),因为云南南部和湖南本来是侗傣、苗瑶、南亚语系人群占主流地位的地区。事实上,藏族是最后才移民到这个区域的[42],且研究中没有发现南方原住民对其的贡献。因此,遗传学研究结论与藏缅群体的历史记录相一致。

这项研究的结果也可为各个群体内部遗传结构研究提供信息。例如,不同地理区域

的彝族群体的混合水平有很大差异，说明这个族群的异质性高。这与历史学和语言学的观点相一致，彝族涵盖了一些很特别的分支。但是，拉祜族和土家族群体在不同区域的混合水平相似，表明这些族群的遗传均一性很高。纳西族在云南西北部各群体中南方原住居民的线粒体DNA最多，在所有研究的藏缅群体中南方原住居民的Y染色体的贡献最高，这反映了纳西族历史上来自南方群体的高比例的遗传输入，而这还未被历史学家和人类学家明确认识到。

对Y染色体和线粒体DNA混合水平的估算可能有较大误差，因为基于对单位点数据的估算存在较大方差。然而，三方面的证据表明我们的结论是有效的。首先，我们使用了两个不同的统计数据去估算东亚南方人群的贡献，并且它们互相一致（Y染色体和线粒体DNA结果的相关系数均为0.93）。其次，在所有可得到Y染色体和线粒体DNA数据的14个群体中，两种方法中M预测的相对数量级（线粒体DNA为1，Y染色体为1）是一致的。第三，若我们设定零假设是男女贡献相等的二项分布，能随机观察到女性世系贡献高于男性的概率在14个群体中有11个群体是0.01。此外，在分子方差分析中，南部藏缅群体与其北部祖先系之间在Y染色体上没有明显区分的结果（$F_{CT}=0.002$；$P=0.221$）也支持我们的结论。

还应当注意的是，亲本群体中单倍群的频率估算可能会不准确。在估算南部藏缅群体的混合水平时，亲本群体单倍群的频率（包括Y染色体和线粒体DNA）是由各个群体的单倍群频率的算术平均数估算的。我们没有尝试把群体规模考虑在内并据此计算加权平均值，这样做虽然很重要，但由于这些群体在历史上发生过剧烈的人口波动，所以它是无法实现的。然而，简化的处理不会影响我们观察到的男女世系的贡献差异，因为任何加权方案对男女世系的影响相同。另外，现存的东亚南方人群（亲本群体之一）可能也受到北方移民的影响，男性世系受到的影响比女性更多。这将导致南部藏缅群体中男性和女性遗传输入的比例差异更大的结果。因此，南部藏缅的性别差异混合，事实上可能比这项研究显示出的更加显著（表3-13）。

总之，本研究展现了南北融合的大致情形，证实了更有趣的男性和女性亲本群体在现存南部藏缅群体中的不对称贡献。南部藏缅群体保留了更多北方移民的男性血统，因此相比东亚其他南方人群，他们与东亚北方人群更近。与此同时，更多的南部线粒体DNA谱系出现在现存南部藏缅群体的基因库中。虽然美洲有明显性别差异的人群混合在世界其他人群中很少见，但是以南部藏缅群体为代表的较小的性别差异可能会较多出现。

参考文献

[1] Chakraborty R. Gene admixture in human populations: models and predictions. Yearb Phys Anthropol, 1986, 29: 1-43.

[2] Chakraborty R, Weiss K M. Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. Proc Natl Acad Sci USA, 1988, 85: 9119-9123.

[3] Pritchard J K, Przeworski M. Linkage disequilibrium in humans: models and data. Am J Hum

Genet, 2001, 69: 1 - 14.

[4] Smith M W, Lautenberger J A, Shin H D, et al. Markers for mapping by admixture linkage disequilibrium in African American and Hispanic populations. Am J Hum Genet, 2001, 69: 1080 - 1094.

[5] Ardlie K G, Kruglyak L, Seielstad M. Patterns of linkage disequilibrium in the human genome. Nat Rev Genet, 2002, 3: 299 - 309.

[6] Gresham D, Morar B, Underhill P A, et al. Origins and divergence of the Roma (gypsies). Am J Hum Genet, 2001, 69: 1314 - 1331.

[7] Di Benedetto G, Erguven A, Stenico M, et al. DNA diversity and population admixture in Anatolia. Am J Phys Anthropol, 2001, 115: 144 - 156.

[8] Helgason A, Hickey E, Goodacre S, et al. mtDNA and the islands of the North Atlantic: estimating the proportions of Norse and Gaelic ancestry. Am J Hum Genet, 2001, 68: 723 - 737.

[9] Merriwether D A, Huston S, Iyengar S, et al. Mitochondrial versus nuclear admixture estimates demonstrate a past history of directional mating. Am J Phys Anthropol, 1997, 102: 153 - 159.

[10] Carvajal-Carmona L G, Soto I D, Pineda N, et al. Strong Amerind/ white sex bias and a possible Sephardic contribution among the founders of a population in Northwest Colombia. Am J Hum Genet, 2000, 67: 1287 - 1295.

[11] Mesa N R, Mondragon M C, Soto I D, et al. Autosomal, mtDNA, and Y chromosome diversity in Amerinds: pre-and post-Columbian patterns of gene flow in South America. Am J Hum Genet, 2000, 67: 1277 - 1286.

[12] Parra E J, Marcini A, Akey J, et al. Estimating African American admixture proportions by use of population-specific alleles. Am J Hum Genet, 1998, 63: 1839 - 1851.

[13] Parra E J, Kittles R A, Argyropoulos G, et al. Ancestral proportions and admixture dynamics in geographically defined African Americans living in South Carolina. Am J Phys Anthropol, 2001, 114: 18 - 29.

[14] Bortolini M C, Da Silva W A, Junior W, et al. African-derived South American populations: a history of symmetrical and asymmetrical matings according to sex revealed by bi-and uni-parental genetic markers. Am J Human Biol, 1999, 11: 551 - 563.

[15] Sans M, Weimer T A, Franco M H, et al. Unequal contributions of male and female gene pools from parental populations in the African descendants of the city of Melo, Uruguay. Am J Phys Anthropol, 2002, 118: 33 - 44.

[16] Chakraborty R. Molecular evidence of directional mating for gene migration in human populations. Abstract from the XVIIIth International Congress of Genetics, Beijing, August 10 - 15, 1998: 102.

[17] 王钟翰. 中国民族史. 北京: 中国社会科学出版社, 1994.

[18] Su B, Xiao C, Deka R, et al. Y chromosome haplotypes reveal prehistorical migrations to the Himalayas. Hum Genet, 2000b, 107: 582 - 590.

[19] Su B, Xiao J, Underhill P, et al. Y chromosome evidence for a northward migration of modern humans into Eastern Asia during the Last Ice Age. Am J Hum Genet, 1999, 65: 1718 - 1724.

[20] Qian Y, Qian B, Su B, et al. Multiple origins of Tibetan Y chromosomes. Hum Genet, 2000, 106: 453 - 454.

[21] Karafet T, Xu L, Du R, et al. Paternal population history of East Asia: sources, patterns, and microevolutionary process. Am J Hum Genet, 2001, 69: 615 - 628.

[22] Qian Y P, Chu Z T, Dai Q, et al. Mitochondrial DNA polymorphisms in Yunnan nationalities in China. J Hum Genet, 2001, 46: 211 - 220.

[23] Yao Y G, Kong Q P, Bandelt H J, et al. Phylogeographic differentiation of mitochondrial DNA in Han Chinese. Am J Hum Genet, 2002a, 70: 635 - 651.

[24] Yao Y G, Nie L, Harpending H, et al. Genetic relationship of Chinese ethnic populations revealed by mtDNA sequence diversity. Am J Phys Anthropol, 2002b, 118: 63 - 76.

[25] Su B, Jin L, Underhill P, et al. Polynesian origins: new insights from the Y chromosome. Proc Natl Acad Sci USA, 2000a, 97: 8225 - 8228.

[26] Kolman C J, Sambuughin N, Bermingham E. Mitochondrial DNA analysis of Mongolian populations and implications for the origin of New World founders. Genetics, 1996, 142: 1321 - 1334.

[27] Yao Y G, Lu X M, Luo H R, et al. Gene admixture in the silk road of China: evidence from mtDNA and melanocortin 1 receptor polymorphism. Genes Genet Syst, 2000, 75: 173 - 178.

[28] Yao Y G, Zhang Y P. Phylogeographic analysis of mtDNA variation in four ethnic populations from Yunnan province: new data and a reappraisal. J Hum Genet, 2002c, 47: 311 - 318.

[29] Hammer M F, Karafet T M, Redd A J, et al. Hierarchical patterns of global human Y chromosome diversity. Mol Biol Evol, 2001, 18: 1189 - 1203.

[30] Jin L, Su B. Natives or immigrants: modern human origin in East Asia. Nat Rev Genet, 2000, 1: 126 - 133.

[31] Kivisild T, Tolk H V, Parik J, et al. The emerging limbs and twigs of the East Asian mtDNA tree. Mol Biol Evol, 2002, 19: 1737 - 1751.

[32] Excoffier L, Smouse P E, Quattro J M. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics, 1992, 131: 479 - 491.

[33] Schneider S, Kueffer J M, Roessli D, et al. Arlequin version 2.000: a software for population genetic analysis. Geneva, Genetics and Biometry Laboratory, University of Geneva, 2000.

[34] Dupanloup I, Bertorelle G. Inferring admixture proportions from molecular data: extension to any number of parental populations. Mol Biol Evol, 2001, 18: 672 - 675.

[35] Wang J. Maximum-likelihood estimation of admixture proportions from genetic data. Genetics, 2003, 164: 747 - 765.

[36] Bertorelle G, Excoffier L. Inferring admixture proportions from molecular data. Mol Biol Evol, 1998, 15: 1298 - 1311.

[37] Roberts D F, Hiorns R W. Methods of analysis of the genetic composition of a hybrid population. Hum Biol, 1965, 37: 38 - 43.

[38] Sans M, Salzano F M, Chakraborty R. Historical genetics in Uruguay: estimates of biological

origins and their problems. Hum Biol, 1997, 69: 161 - 170.

[39] Seielstad M T, Minch E, Cavalli-Sforza L L. Genetic evidence for a higher female migration rate in humans. Nat Genet, 1998, 20: 278 - 280.

[40] Sans M. Admixture studies in Latin America: from the 20th to the 21st century. Hum Biol, 2000, 72: 155 - 177.

[41] Santos F R, Pandya A, Tyler-Smith C, et al. The central Siberian origin of native American Y chromosomes. Am J Hum Genet, 1999, 64: 619 - 628.

[42] 苍铭. 云南民族迁徙文化研究. 昆明: 云南民族出版社, 1997.

3.6 汉藏群体进入喜马拉雅东部的两条迁徙路线

3.6.1 研究背景

喜马拉雅山区东部位于早期人类进入东亚的入口处附近。相比于欧亚大陆东部的其他地区，喜马拉雅山区的人类活动出现得很晚，直至 5 000～7 000 年前，这个地区才开始有人类居住[1]。现在，汉藏语系的很多群体分布于此。在喜马拉雅山区西部居住着包括门巴族、不丹人、锡金人和尼泊尔的北部部落在内的藏语支群体[2]。这些群体都起源于约 2 000 年内的藏族群体的扩张[1]。位于中国西藏和印度阿萨姆邦之间的喜马拉雅东部山区分布着珞巴族和僜人[3]，他们的语言属于汉藏语系的北阿萨姆语支。珞巴族（又称阿迪人）主要分布在东起察隅、西至门隅之间的西藏珞瑜地区，目前总人口约 60 万，其中处于中国实际控制区内仅有 2 300 余人，其余的大部分珞巴族居住地因目前被印方占领而无法详细统计人口数量。僜人是生活在中国西藏和印度交界的察隅河流域，总人数 2 万～3 万，其中察隅县中国实际控制区内有 1 391 人（2006 年统计）[4]。分布在印占区的僜人被称为米什米。语言学的分类并未对珞巴和僜人的起源问题提供清晰的证据，因此我们调查了这些群体的遗传多样性，试图对他们的起源问题以及与其他汉藏群体的关系做深入探索。

Y 染色体和线粒体 DNA 是研究群体历史最有力的工具[5-7]，对青藏高原以及喜马拉雅山区的藏族群体的线粒体 DNA 研究结果显示，青藏高原的群体来源于多次进入高原的群体扩张[8]，但相关研究并未涉及群体进入喜马拉雅山区的具体路线。因此我们试图进行 Y 染色体多样性的研究，厘清喜马拉雅山区东部群体珞巴和僜人的起源问题。

在父系谱系中，Y 染色体单倍群 O3 是汉藏语系群体的主要单倍群[9]，是研究汉藏群体迁徙扩张历史最有力的父系谱系。本节分析了珞巴和僜人的 Y 染色体差异，并且试图还原汉藏群体进入喜马拉雅山区东部的迁徙路径。我们的研究数据显示这些群体至少可经两条线路从中国北部进入该地区。

3.6.2 材料和方法

1. 样本采集

本次研究一共采集了来自西藏林芝地区察隅县的 90 个僜人个体和林芝地区米林县

130个珞巴族个体的唾液样本。采样过程中遵循知情同意和匿名的原则，样本均来自无亲缘关系的健康个体。本次采样活动通过了复旦大学生命科学学院的伦理审查。我们采用了血液基因组DNA提取试剂盒提取基因组DNA。

为了进一步分析喜马拉雅山区群体与东亚群体之间的遗传关系，我们收集了75个群体的Y染色体数据作为参考数据[9-11]（表3-14），以便进行较为系统全面的比较和分析。这些群体数据涉及包括汉藏语系、阿尔泰语系、侗傣语系、苗瑶语系、南亚语系和印欧语系在内的各类群体。此外，汉藏语系北阿萨姆语支未公开发表的数据由基因地理组南亚研究中心提供，也包含在后续的分析中。

表3-14　样本信息和参考文献数据信息

	群　体	缩　写	语　类	样本量	参考文献
汉藏语系　藏缅语族					
1	阿迪(Adi)	TB1	山南语支	57	基因地理组未刊数据
3	阿帕塔尼(Apatani)	TB3	山南语支	73	基因地理组未刊数据
4	白马(Baima)	TB4	藏语支未定组	18	基因地理组未刊数据
6	僜人(北)(Deng)	TB5	山南语支	120	本研究
7	伽罗(Galo)	TB6	山南语支	43	基因地理组未刊数据
8	噶若(Garo)	TB7	库基-钦-那伽语支	46	基因地理组未刊数据
19	嘉戎(Jiarong)	TB8	羌语支	92	本课题组未刊数据
20	卡贝(Kabui)	TB9	库基-钦-那伽语支	65	基因地理组未刊数据
21	卡卓瑞(Kachori)	TB10	库基-钦-那伽语支		基因地理组未刊数据
22	卡比(Karbi)	TB11	库基-钦-那伽语支		基因地理组未刊数据
27	珞巴	TB12	山南语支	130	本研究
29	梅忒(Meitei)	TB13	库基-钦-那伽语支	61	基因地理组未刊数据
30	印占我国藏南地区(Arunachal)	TB14	库基-钦-那伽语支	18	基因地理组未刊数据
31	僜人(南)(Mishmi)	TB15	山南语支	12	基因地理组未刊数据
32	门巴	TB16	藏语支藏语组	34	本课题组未刊数据
34	木雅(Muyag)	TB17	羌语支	9	本课题组未刊数据
35	那伽(Naga)	TB18	库基-钦-那伽语支	34	基因地理组未刊数据
36	尼瓦尔(Newar)	TB19	藏语支马哈吉拉语组	66	[11]
37	诺特(Nocte)	TB20	景颇语支	45	基因地理组未刊数据
38	尼西(Nyishi)	TB21	山南语支	63	基因地理组未刊数据
40	普内(Phunoi)	TB22	缅彝语支	9	本课题组未刊数据

（续表）

	群　体	缩　写	语　类	样本量	参考文献
41	羌族	TB23	羌语支	152	本课题组未刊数据
42	却域(Queyu)	TB24	羌语支	15	本课题组未刊数据
44	塔金(Tagin)	TB25	库基-钦-那伽语支	15	基因地理组未刊数据
45	塔芒(Tamang)	TB26	藏语支藏语组	45	[11]
46	唐撒(Tangsa)	TB27	景颇语支	14	基因地理组未刊数据
47	安多藏	TB28	藏语支藏语组	101	本课题组未刊数据
48	卫藏	TB29	藏语支藏语组	297	本课题组未刊数据
49	昌都康巴	TB30	藏语支藏语组	77	本课题组未刊数据
50	尼泊尔藏族	TB31	藏语支藏语组	156	[11]
51	云南藏族	TB33	藏语支藏语组	50	本课题组未刊数据
53	万卓(Wancho)	TB32	景颇语支	12	[9]
55	阿昌族	TB34	缅彝语支	40	[9]
56	白族	TB35	缅彝语支	128	[9]
57	土家族	TB36	缅彝语支	100	[9]
58	怒族	TB37	缅彝语支	50	[9]
59	哈尼族	TB38	缅彝语支	41	[9]
60	拉祜族	TB39	缅彝语支	88	[9]
61	傈僳族	TB40	缅彝语支	49	[9]
62	彝族	TB41	缅彝语支	47	[9]
63	景颇族	TB42	景颇语支	17	[9]
64	普米族	TB43	羌语支	47	[9]
65	纳西族	TB44	缅彝语支	87	[9]
66	独龙族	TB45	山南语支	28	[9]
汉藏语系　汉语族					
9	广西汉族	H1	南方组	34	本课题组未刊数据
10	甘肃汉族	H2	北方组	259	本课题组未刊数据
11	山东汉族	H3	北方组	247	本课题组未刊数据
12	内蒙古汉族	H4	北方组	86	[9]
13	兰州汉族	H5	北方组	98	[9]

（续表）

	群　体	缩　写	语　类	样本量	参考文献
14	淄博汉族	H6	北方组	64	[9]
15	四川汉族	H7	南方组	39	[9]
16	云南汉族	H8	南方组	81	[9]
苗瑶语系					
17	苗族	HM2	苗语支	40	[9]
71	瑶族	HM3	勉语支	144	[9]与本课题组
印欧语系					
23	加德满都(Kathmandu)	IE1	印度伊朗语支	77	[11]
39	奥里萨(Orissa)	IE2	印度伊朗语支	40	基因地理组未刊数据
阿尔泰语系					
24	朝鲜族	Alt1	韩语族	32	本课题组未刊数据
28	满族	Alt2	通古斯语族	62	本课题组未刊数据
33	蒙古族	Alt3	蒙古语族	63	本课题组未刊数据
43	撒拉族	Alt4	突厥语族	28	本课题组未刊数据
52	土族	Alt5	蒙古语族	13	本课题组未刊数据
76	回族	Alt6	汉语族北方组	15	[9]
侗傣语系					
54	壮族	TK1	侗傣语族贝傣语支	86	[9]
67	布依	TK2	侗傣语族贝傣语支	48	[9]
68	水族	TK3	侗傣语族侗水语支	40	[9]
69	傣族	TK4	侗傣语族贝傣语支	132	[9]
70	泰国人	TK5	侗傣语族贝傣语支	60	[9]
2	阿霍姆人(Ahom)	TK6	侗傣语族贝傣语支	20	基因地理组未刊数据
25	万象老族	AA3	侗傣语族贝傣语支	23	本课题组未刊数据
26	占巴塞老族	AA4	侗傣语族贝傣语支	59	本课题组未刊数据
南亚语系					
5	碧(Bit)	AA1	孟高棉语族北高棉语支	10	本课题组未刊数据
18	门哒(Munda)	AA2	门哒语族	12	基因地理组未刊数据
72	佤族	AA5	孟高棉语族北高棉语支	31	[9]

（续表）

	群　体	缩　写	语　类	样本量	参考文献
73	布朗族	AA6	孟高棉语族北高棉语支	28	[9]
74	德昂族	AA7	孟高棉语族北高棉语支	16	[9]
75	柬埔寨人	AA8	孟高棉语族东高棉语支	14	[9]

2. Y染色体基因分型

所有样本均采用包含75个Y染色体SNP位点的7组试版进行基因分型检测，试版中各位点的具体组合如下。试版1（单倍群O）：M175、M119、P203、M110、M268、P31、M95、M176、M122、M324、M121、P201、M7、M134、M117、002611、P164、L127（rs17269396）、KL1（rs17276338）；试版2（非单倍群O）：M130、P256、M1、M231、M168、M174、M45、M89、M272、M258、M242、M207、M9、M96、P125、M304、M201、M306；试版3（单倍群C）：M217；试版4（单倍群D）：P47、N1、P99、M15、M125、M55、M64.1、M116.1、M151、N2、022457；试版5（单倍群N）：M214、LLY22g、M128、M46/Tat、P63、P119、P105、P43、M178；试版6（单倍群R）：M306、M173、M124、M420、SRY10831.2、M17、M64.2、M198、M343、V88、M458、M73、M434、P312、M269、U106/M405；试版7（单倍群Q）：P36.2。试版中各SNP位点的命名参照最新的Y染色体系统树[12]。对这些位点的检测采用SNaPshot的方法，SNaPshot多重反应试剂盒对用DNA模板进行荧光PCR反应后，在基因测序仪（3730 xl Genetic Analyzer）上读取结果。

此外所有的样本还进行了STR检测，所检测的位点包括：DYS19、DYS385a、DYS385b、DYS388、DYS389I、DYS389II、DYS390、DYS391、DYS392、DYS393、DYS426、DYS437、DYS438和DYS439。

3. 数据分析

所有单倍型均按照最新的Y染色体系统树[12]进行判定，基于STR数据，利用SPSS 15.0软件估算各群体的O3－M134单倍群和O3－M117单倍群的F_{ST}值[13]，并对F_{ST}值进行了多维尺度分析（MDS），以揭示各个群体之间的亲缘关系。利用中点连接法计算并绘制了网络结构图（network），该分析经软件Network 4.6完成[14]，以研究各个单倍群内部细枝之间的相互关系和在群体中的分布。同时估算网络结构图中某几个细枝的时间，计算过程中STR突变率选用进化突变率0.000 69[15]。

采用Arlequin 3.0软件包[16]在STR数据基础上计算各个群体的单倍型平均多样性，并用Surfer 7.0将各群体的多样性按群体位置的经纬度投射到地图上，并描画等值分布线图。用分子方差分析（AMOVA）估计群体之间差异程度，进一步了解差异程度在群体间和群体内的分布，该分析也利用Arlequin软件包计算。

表3－14所列参考文献中的群体被分为以下7组：西北藏缅（藏族）、东北部藏缅

(羌)、西南藏缅(印度东北部群体)、东南藏缅(华南和中南半岛群体)、中国汉族、北亚群体(阿尔泰语系)和东南亚群体(苗瑶语系、侗傣语系和南亚语系群体),并用上述分析方法,分析以上7组群体与珞巴族和僜人之间的亲缘关系。

3.6.3 研究结果

1. Y染色体SNP单倍群

群体的Y染色体单倍群多样性通常能够为群体的起源提供最直观的线索。研究结果显示珞巴族和僜人的Y染色体单倍群分布模式完全不同(表3-15)。在僜人中,单倍群O3为主要单倍群,而单倍群C、D和N则低频分布。这种分布模式和其他的汉藏群体的Y染色体分布模式相同[17]。而珞巴族的Y染色体单倍群的分布则更为复杂,单倍群O、N和D都高频存在,并且有单倍群J、R和Q的少量分布。单倍群O和D在藏族群体中高频存在[8],单倍群N在东亚群体中出现的频率并不高,但是在北欧的乌拉尔语族的群体中高频出现[18]。珞巴族的单倍群多样性是僜人的3倍多,通常较低的单倍群多样性是群体经历瓶颈效应后的结果。僜人和珞巴族或许都曾经历过瓶颈效应,而较晚的群体混合带入了新的单倍型而导致了多样性的上升。因此,Y染色体SNP单倍群分布模式显示,珞巴族群体与僜人相比较有着更为复杂的起源,并且在遗传上与藏族更为接近,而僜人则与中国南部的其他汉藏群体关系较近。

表3-15　喜马拉雅东部山区群体的Y染色体单倍群频率

群　体	单倍群(%)														多样性
	C3	D*	D1	D3	J	N1*	O*	O3*	O3a3c*	O3a3c1*	Q	R*	R1*	R1a1	
	M217	M174	M15	P99	M304	LLY22g	M175	M122	M134	M117	M242	M207	M173	M17	
珞巴族		1.54	3.08	16.15	0.77	34.62	1.54	0.77	2.31	30.77	0.77	2.31	1.54	3.85	0.7608
僜人	1.11		1.11	1.11		1.11	1.11		31.11	63.33					0.5072

Y染色体单倍群D可能在末次盛冰期之前就已经出现在藏族群体中[17],因此单倍群D也许并不是与汉藏群体扩张事件直接相关联的单倍群。相反,单倍群O3在汉藏群体中广泛存在,与汉藏群体的遗传历史则更为相关。因此本节的后续分析主要集中在O3单倍群内部的STR单倍型多样性上。

2. 聚类分析

为了研究不同群体中Y染色体单倍群O3的关系以及喜马拉雅山区群体O3的起源,我们收集了各群体已发表的单倍群O3-M117和O3-M134的STR数据,并进行聚类分析。由于参考数据中单倍群O3-M117在群体中的分布比O3-M134更为广泛,因而在O3-M117的分析中包含了更多的群体。

在单倍群O3-M117的MDS图(图3-17)中,藏族群体都分布于左上部的位置,羌族群体位于右半部分偏下的位置,珞巴族群体离藏族较近,而僜人则更接近羌族。印度

西南的藏缅群体跟珞巴族群体一样，显示了与藏族较近的亲缘关系，而中国东南的藏缅群体与羌族和僜人的距离较近。在单倍群 O3－M134 的 MDS 图(图 3－18)中，我们也观察到与单倍群 O3－M117 的 MDS 图中相类似的结果，僜人和珞巴族各自分别表现出与羌族和藏族较近的距离。

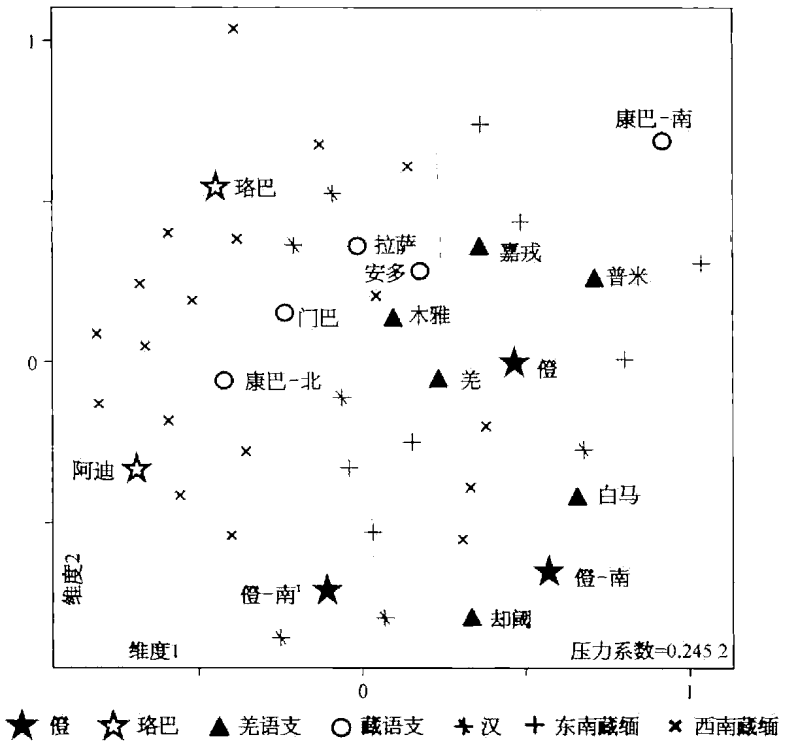


图 3－17 汉藏群体单倍群 O3－M117 的多维尺度分析图

3. *分子方差分析*

为了进一步分析珞巴族和僜人以及其他群体之间的差异，将僜人分别归入珞巴族、羌族、藏族、汉族、东南藏缅和西南藏缅这 6 大组，进行分子方差分析(图 3－19)。分子方差分析是一种专门针对遗传学数据估计群体之间差异程度的方法。在使用分子方差分析之前需要预先对群体进行分组，然后由该算法根据设定的分组把总体的方差分解为组间、组内群体间和群体内 3 个层次，最后通过计算机程序重排(permutation)模拟检验分组假设是否成立。通过聚类分析和多维尺度分析对群体的基本结构有了一个总体的认识之后，就可以通过分子方差分析进一步了解变异程度在群体间和群体内的分布，并在数量上估计其比例[19]。总体来说，组内群体之间的差异要大于组间的差异，并且单倍群 M117 的组间差异比 M134 的低，在单倍群 O3－M117 的分析中，将僜人归入珞巴组或者藏族组中，组内差异比其他组要高，而将僜人归入羌族组中，则组内差异较低。这反映了僜人与羌族的差异较小，而与珞巴族和藏族之间存在较大差异。

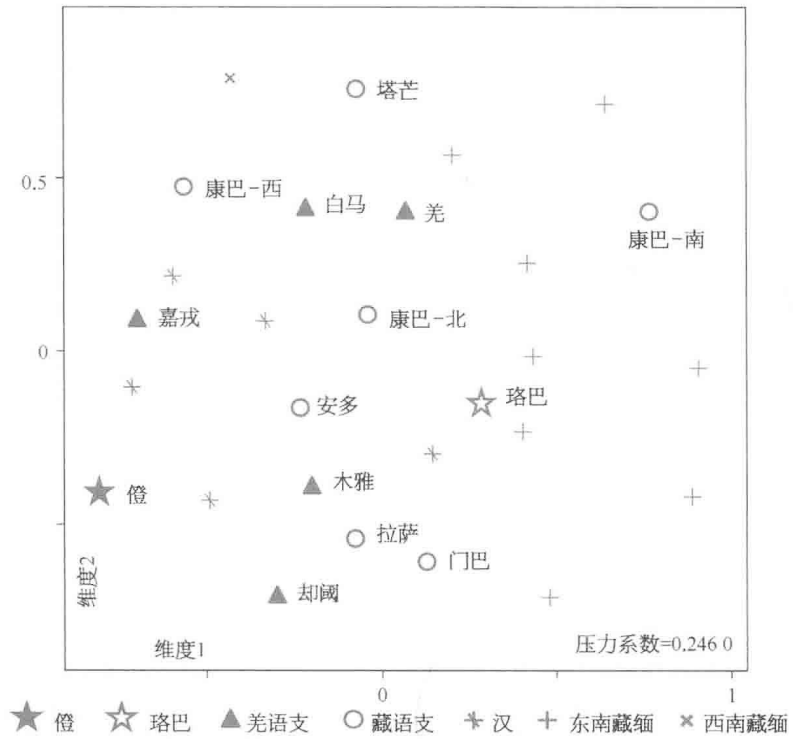


图 3-18　汉藏群体单倍群 O3-M134 的多维尺度分析图

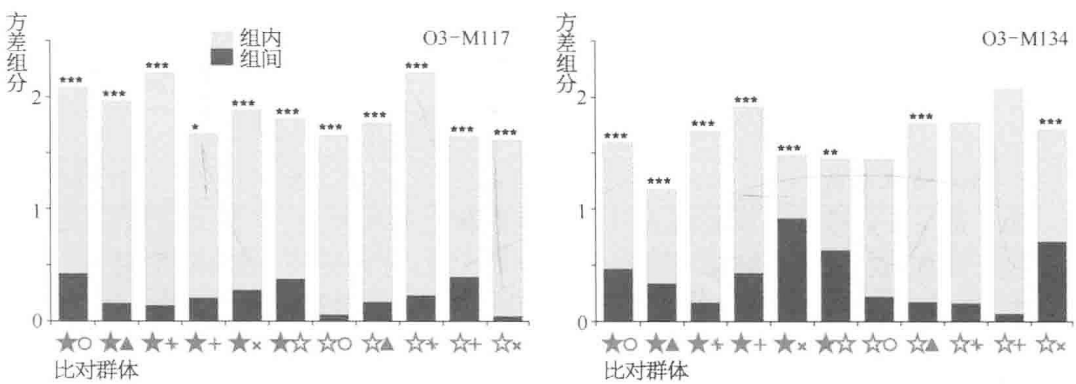


图 3-19　僜人分别与另一群体合并进行的分子方差分析结果

组内和组间的差异数值通过柱状图的形式表现在图中。

4. 遗传组分的结构分析

Structure 软件可以对一系列群体数据进行分析，估算个体水平上的各种遗传成分的比例。当遗传成分分组数量 K 值设为 8 的时候，从单倍群 O3-M1117 的 Structure 结果中观察到珞巴族和僜人的遗传结构差异明显。所有的群体样本根据 Structure 结果中遗传成分的不同大致分为两组，珞巴族与藏族以及西南部藏缅群体在遗传成分上类似，而

僜人更接近羌族、汉族和其他东南亚群体。单倍群 O3 - M134 的 Structure 分组虽然不明显，僜人和珞巴族还是能够很好地被区分开来，僜人的遗传成分与汉族群体更为相似（图 3 - 20）。

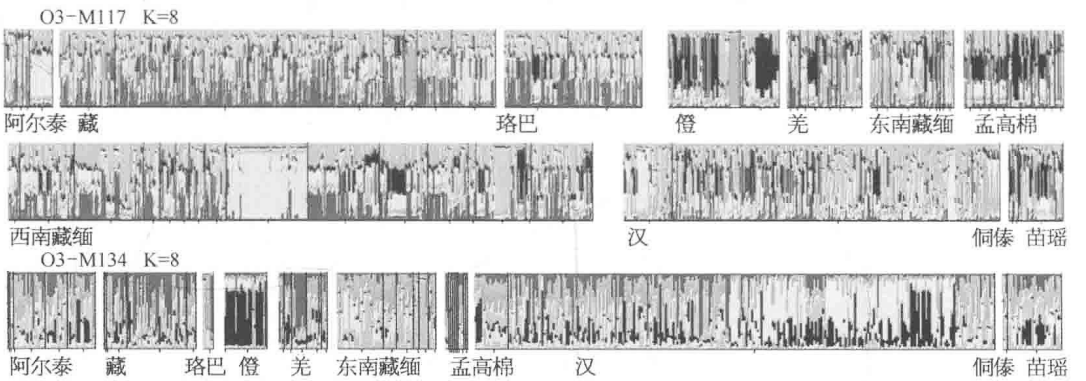


图 3 - 20 东亚群体的遗传组分结构分析

5. 单倍群网络结构分析

本研究对 M134 和 M117 单倍群进行了网络结构图的分析（图 3 - 21），图中虽然很难将各个群体一一区分出来，但还是可以观察到群体间的距离关系。在 O3 - M134 的网络结构图中大多数僜人个体与汉族个体共享单倍型，在 O3 - M117 单倍群的网络结构图中，珞巴族个体则大多数与藏族共享单倍型。同时估算了网络结构中僜人集中出现分支的时间，分别为 2 113 年±798 年和 2 052 年±1 122 年[1]。

6. 平均基因多样性

通常一个群体的多样性与群体的古老程度以及群体数量密切相关[13]。用 STR 数据计算了各个群体中普遍存在的单倍群 O3 - M117 的多样性。在等值线地图观察到羌族群体所在位置的多样性是所有群体中最高的，形成了一个中心（图 3 - 22），提示至少从单倍群 O3 - M117 来看，羌族群体可能是汉藏群体中最古老的。单倍群 O3 - M117 的多样性以羌族为中心，随后沿着两条路径呈递减状态流向喜马拉雅山区东部。一条呈逆时针方向穿过青海和西藏，另一条则沿顺时针方向经过云南，而珞巴和僜人分别位于这两条路线的终点位置。

3.6.4 讨论

1. 汉藏群体的起源

Y 染色体单倍群 O3 是汉藏群体的主要父系单倍群[9]，因此相比于其他单倍群，能更全面地展现群体历史。单倍群 O3 下游目前大约有 20 个亚单倍群，O3 - M117 是其中最高频的[20]。在云南和西藏相毗邻的地方的某些群体中，甚至所有单倍群 O3 的突变都属于 M117。因此在研究汉藏群体特别是西藏东南部（喜马拉雅山区东部）群体的遗传学起源时，对单倍群 O3 - M117 的分析至关重要。

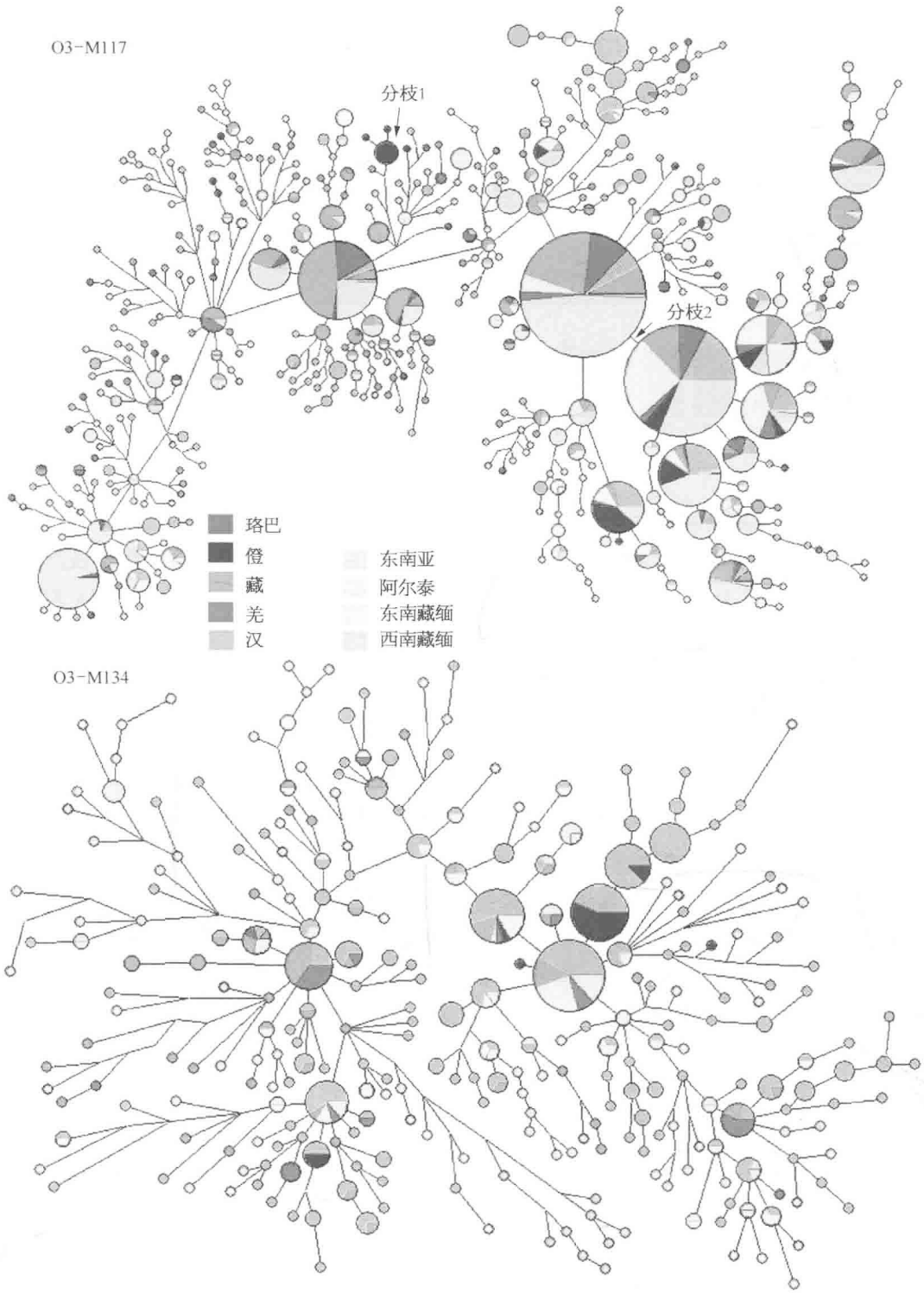


图 3-21　O3-M117 和 O3-M134 网络结果分析图(基于中间连接法的最简约树)

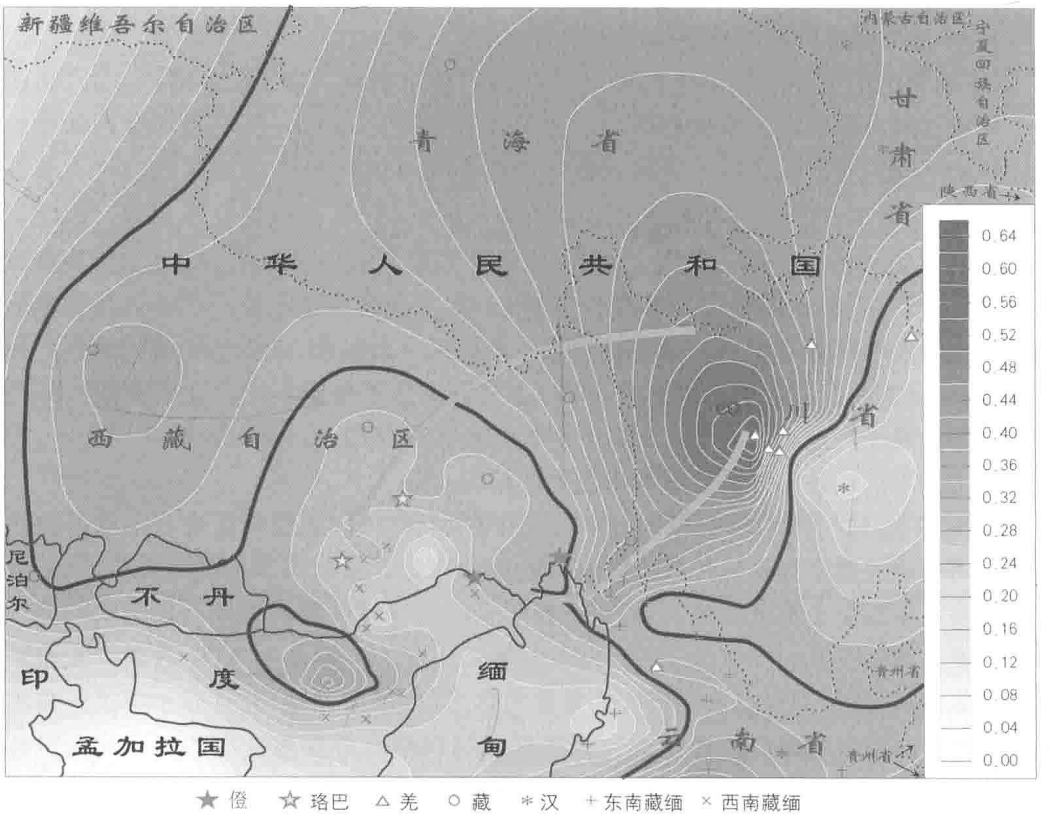


图 3-22 喜马拉雅山区东部群体的平均基因多样性和迁徙的两条路径

西藏以东的地区很有可能是所有汉藏群体的起源地，在这个地区的羌族群体中发现了 O3-M117 最高的基因多样性。汉族的古老传说中有明确的叙述，汉族的祖先可以追溯到羌族[21]。考古学的发现也指向大约 7 000 年前的仰韶文化起源于羌族所在的区域[22,23]。古书中也有记载，历史上藏缅语族下的大多数群体的名称都包含"羌"[24]。遗传学证据支持历史学和考古学的研究结论，认为羌族群体是汉藏群体的祖先。

2. 进入喜马拉雅山区东部的两条路径

比起汉藏群体现今所分布的其他区域，他们进入喜马拉雅山区东部的迁徙可能是晚近的事件，因为该地区群体的 O3-M117 单倍群 STR 多样性都相对很低。分析认为，汉藏群体迁徙进入喜马拉雅山区东部至少经历了两条路径。喜马拉雅山区东部的西半部分群体，包括珞巴族和印度东北部的大多数汉藏群体，与藏族的关系较近，直接从北方而来。而东半部分的群体，包括僜人以及靠近缅甸北部的群体则更接近羌族群体和汉藏东南部群体，自东边而来。笔者认为，喜马拉雅山区东部的各个群体是在不同的父系世系基础上，经历了不同的扩张过程而形成的。

图 3-22 中提出的迁徙路线只是大致的估计。由于地图上存在某些没有数据的区域，在等值图的绘制过程中采用了平滑化的方法，可能对等值图的精确性造成一定的

影响。但是，在本研究中，画出这两条路线所需要拥有的群体数据还是足够的。

3. 其他单倍群所反映的迁徙历史

O3谱系的历史并不能完全代表汉藏群体的整个历史。除O3单倍群之外，其他Y染色体单倍群也出现在分析的样本中。这些单倍群的出现也能反映出汉藏群体中的其他成分的不同历史。与乌拉尔语系群体密切相关的单倍群N[18]在该区域或许有着很久远的历史，因为东亚的早期人类经由该地区进入东亚，而N已经留在了当地。因此当地的单倍群N值得进一步深入研究。此外，Y染色体单倍群D是西藏和日本群体中普遍存在的单倍群[25-28]，单倍群D是否和单倍群O3一同从藏族中进入珞巴族抑或是在单倍群O3到达之前，单倍群D已经在该地区存在了很长的一段时间，是尚未厘清的问题。更有趣的是，珞巴族中出现了单倍群J、Q和R，这或许暗示了珞巴族中存在着欧亚大陆西部和北部群体的迁入或基因流动。因此，对分布在早期现代人进入东亚的入口之一，即喜马拉雅山区东部的各群体需要展开更细致的研究。

除Y染色体之外，包括线粒体DNA、常染色体DNA和免疫球蛋白在内的其他遗传标记也用来研究西藏和喜马拉雅山区群体的形成历史。青藏高原地区藏族线粒体DNA的研究显示藏族是多重来源的，包括旧石器时代的北亚群体起源和新石器时代的羌族起源[8]。僜人和珞巴族的线粒体DNA的研究尚未开展。常染色体STR数据显示藏族内部各群体都很类似，而珞巴族和僜人的常染色体STR却很独特，并且显示出了很强的奠基者效应[3]。免疫球蛋白的研究认为藏族群体与包括北方汉族在内的东亚北部群体有很高的相似度[29]，珞巴族和僜人也还未有此方面的研究。总的来说，对喜马拉雅山区东部群体的遗传学研究还很不充分，还需要更深入的研究。而目前已有的对其他遗传标记的研究结论，则多与本次研究的结论相符合。

参考文献

[1] Huang W. The prehistoric human occupation of the Qinghai-Xizang plateau. Gottinger Geographische Abhandlungen, 1994, 95: 201-219.

[2] Lewis M P. Ethnologue: languages of the world, sixteenth edition. Dallas, Tex: SIL International, 2009.

[3] Kang L, Li S, Gupta S, et al. Genetic structures of the Tibetans and the Deng people in the Himalayas viewed from autosomal STRs. J Hum Genet, 2010, 55(5): 270-277.

[4] 莫非. 即将消失的僜人(西藏少数派报告——僜人). 华夏地理, 2009, 69(3): 152-171.

[5] Jobling M A, Tyler-Smith C. The human Y chromosome: an evolutionary marker comes of age. Nat Rev Genet, 2003, 4(8): 598-612.

[6] Jin L, Su B. Natives or immigrants: modern human origin in East Asia. Nat Rev Genet, 2009, 1(2): 126-133.

[7] Underhill P A, Shen P, Lin A A, et al. Y chromosome sequence variation and the history of human populations. Nat Genet, 2000, 26(3): 358-361.

[8] 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(4): 555 - 569.

[9] Shi H, Dong Y L, Wen B, et al. Y chromosome evidence of southern origin of the East Asian-specific haplogroup O3 - M122. Am J Hum Genet, 2005, 77(3): 408 - 419.

[10] Shi H, Su B. Origin of modern humans in East Asia: clues from the Y chromosome. Front Biol China, 2009, 4: 241 - 247.

[11] Gayden T, Cadenas A M, Regueiro M, et al. The Himalayas as a directional barrier to gene flow. Am J Hum Genet, 2009, 80(5): 884 - 894.

[12] Karafet T M, Mendez F L, Meilerman M B, et al. New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. Genome Res, 2008, 18(5): 830 - 838.

[13] Nei M, Kumar S. Molecular evolution and phylogenetics. New York: Oxford University Press, 2000.

[14] Polzin T, Daneschmand S V. On steiner trees and minimum spanning trees in hypergraphs. Operations Research Letters, 2003, 31: 12 - 20.

[15] Zhivotovsky L A, Underhill P A, Cinnioglu C, et al. The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. Am J Hum Genet, 2009, 74(1): 50 - 61.

[16] Excoffier L, Laval G, Schneider S. Arlequin (version 3.0): an integrated software package for population genetics data analysis. Evol Bioinform Online, 2005, 1: 47 - 50.

[17] Shi H, Zhong H, Peng, et al. Y chromosome evidence of earliest modern human settlement in East Asia and multiple origins of Tibetan and Japanese populations. BMC Biol, 2008, 6: 45.

[18] Rootsi S, Zhivotovsky L A, Baldovic M, et al. A counter-clockwise northern route of the Y chromosome haplogroup N from Southeast Asia towards Europe. Eur J Hum Genet, 2007, 15(2): 204 - 211.

[19] 徐书华. 高密度常染色体 SNPs 揭示的现代人群遗传结构. 上海: 复旦大学博士学位论文, 2009.

[20] Yan S, Wang C C, Li H, et al. An updated tree of Y chromosome Haplogroup O and revised phylogenetic positions of mutations P164 and PK4. Eur J Hum Genet, 2009, 19(9): 1018.

[21] Wang Z H. History of nationalities in China. Beijing: China Social Sciences Press, 1994.

[22] Liu L. The Chinese Neolithic trajectories to early states. Cambridge, UK: Cambridge University Press, 2005.

[23] Zhao Y B, Li H J, Li S N, et al. Ancient DNA evidence supports the contribution of Di-Qiang people to the Han Chinese gene pool. Am J Phys Anthropol, 2011, 144(2): 258 - 268.

[24] Ge L. The historical relationship between ancient Qiang and Tibetans. Journal of Sun Yat-sen University, 1985, 2: 92 - 101.

[25] Hammer M F, Karafet T M, Park H, et al. Dual origins of the Japanese: common ground for hunter-gatherer and farmer Y chromosomes. J Hum Genet, 2006, 51(1): 47 - 58.

[26] Qian Y, Qian B, Su B, et al. Multiple origins of Tibetan Y chromosomes. Hum Genet, 2000,

106(4): 453-454.

[27] Su B, Xiao C, Deka R, et al. Y chromosome haplotypes reveal prehistorical migrations to the Himalayas. Hum Genet, 2000, 107(6): 582-590.

[28] Wen B, Xie X, Gao S, et al. Analyses of genetic structure of Tibeto-Burman populations reveals sex-biased admixture in southern Tibeto-Burmans. Am J Hum Genet, 2004, 74(5): 856-865.

[29] Matsumoto H. Characteristics of Mongoloid and neighboring populations based on the genetic markers of human immunoglobulins. Hum Genet, 1988, 80(3): 207-218.

3.7 摩梭人的遗传起源

3.7.1 研究背景

摩梭人是中国大陆现存唯一的母系社会群体，是人类学研究母系社会的最重要的典型部落之一[1]。1990年的全国人口普查显示摩梭人口为15 000人，他们主要分布在云南西北部的宁蒗县与四川西南的木里藏族自治县相接的泸沽湖区域。历史研究表明，摩梭人主要来源于古老的牦牛羌部落，是氐羌群体的一支。氐羌族群是所有藏缅语族群体的祖先，他们最初生活在中国西北部，在大约2 700年前迁徙到中国西南部[2,3]。摩梭人在20世纪50年代被认定为纳西族的一个分支。然而，摩梭人认为他们是一个独立的民族或者他们应属于藏族的一个分支，因为在风俗习惯上他们更接近藏族而不是纳西族[4]。对于摩梭人、藏族和纳西族的遗传学研究能帮助理解摩梭人的起源。

在过去20年里，遗传标记被广泛运用于推断人群的起源、迁徙和混合[5-8]。其中线粒体DNA和Y染色体在追溯人类进化历史上被证实更有信息量，因为它们分别只沿着母系和父系遗传[5,9]。对于线粒体DNA标记，D环区域的HVS-1片段比其他区域有更高的突变速率，在比较全世界范围内的大量群体数据时，它是最经常被使用来判断遗传关系标记的。随后，Y染色体被发现可以用来推断人群的遗传历史[9-11]。在Y染色体非重组区域上按时间顺序产生的由SNP组成的单倍型被广泛用于推断特定人群的历史。例如，Y染色体SNP被有效地用于揭示东亚和太平洋地区人群的起源和史前迁徙[12-19]。STR是Y染色体上的另一种遗传标记，由于较高的突变速率，它们被广泛用于构建遗传关系较近群体间的进化树并且估算进化事件的年代[20,21]。

摩梭人有着多样的文化和独特的风俗，他们的文化明显受到藏族、纳西族、普米族和彝族的影响[4]。本研究分析了摩梭人和现今居住在云南西北部的其他5个族群（纳西族、藏族、彝族、白族和普米族）在线粒体DNA的HVS-1区域，13个Y染色体SNP和8个Y染色体STR的多态性。应用主成分分析和进化树的分析方法揭示他们在父系和母系上的遗传关系。

3.7.2 材料和方法

1. 样本

研究中采集的6个群体成员都签署知情同意书。他们是来自以下地区的群体：宁蒗

县(摩梭人和普米族)、大理(白族)、丽江(纳西族)、双柏(彝族)和香格里拉(藏族)。研究中的个体是明确不相关的,而且每个个体的4个祖父母所属的民族与这个个体相同。每个个体取5 ml血样并用ACD抗凝。通过酚-氯仿标准方法提取基因组DNA并在-20℃下保存。Y染色体数据都来自男性。除了纳西族的5个没有亲缘关系的女性样本,大多数线粒体DNA来自男性样本。

2. DNA分析

线粒体DNA的HVS-1区域通过引物L15996和C-M1306401扩增[22]。PCR产物通过树脂纯化。纯化的PCR片段随后用ABI Big-Dye测序仪和ABI PRISM 377 DNA测序仪(Applied Biosystems)测序。用Sequence Analysis 3.3软件(Applied Biosystems)提取序列。所有测序的结果都通过正反两个方向的测序结果核对。摩梭人HVS-1序列信息已被提交到GenBank(编号:HM215521-67)。

根据宿兵等[14]研究的标准,本研究选择东亚人群里最有信息量的13个Y染色体SNP(YAP、M7、M9、M15、M45、M88、M89、M95、M110、M119、M122、M130和M134)。在3%~5%的NuSieve琼脂糖(FMC)上电泳观测YAP和M15的长度差异。其余11个位点用PCR-RFLP检测。除了含有一个天然限制性内切酶识别位点的M130,所有位点都通过引物设计引入限制性内切酶识别位点。表3-16是13个位点的引物、扩增环境和限制性酶。

表3-16 引物、PCR条件、限制性内切酶和Y-SNP的参数

位点	退火温度(℃)	引物序列(5′-3′)	限制性酶	碱基读数(片段大小bp)	
YAP	53	CAGGGGAAGATAAAGAAATA ACTGCTAAAAGGGGATGGAT	/	无*Alu*插入 150	*Alu*插入 400
M15	56	ACAAATCCTGAACAATCGC GTCTGGGAAGAGTAGAGAAAAG	/	9 bp插入 151	非插入 142
M89	52	GAAAGTAAAACCCACAGAAGGA GCAAATCAGGCAAAGTGAGACAT	NlaIII	C 79 +21	T 100
M9	55	GAAACGGCCTAAGATGGTTGGAT AAACTGAATCTTTTTTCCTCATTTTTG	Bam HI	C 190 +20	G 210
M119	55	AGGTAAATGACTCACCCTAAGGAAG GGGTTATTCCAATTCAGCATACACGC	BstuI	A 161	C 135 +26
M95	55	ATAAGGAAAGACTACCATATTAGCG TTTGAAGGCCCCAGTTGTGAG	HhaI	C 178 +24	T 202
M122	55	TAGAAAAGCAATTGAGATACTAATTCA GCGATGCTGATATGCTAGTTCAG	NlaIII	C 100 +22	T 122

（续表）

位点	退火温度(℃)	引物序列(5′-3′)	限制性酶	碱基读数(片段大小 bp)	
M134	55	AAGGACCAGGAAAGTATGATCG TTTGATGATTCTTCTTTGGGCTTC	NlaIII	无缺失 100 +22	1 bp 缺失 122
M7	56	TGTACCCTTGACCAATGCCTT TTGTAGTTGAGTTACTGTTCTTCTA	BfaI	C 103 +23	G 126
M130	56	TATCTCCTCTTCTATTGCAG CCACAAGGGGGAAAAAACAC	BslI	T 205	C 162 +43
M110	55	AACATTCTCTGTAGACTCACTGG ATTTAGCACTTCTTTTCCCC	NlaIII	T 200	C 88 +122
M88	55	TCTTATTCCTGCTTCTTCCGC CATGTGATGGTTCAGTAGGTGTGA	BstuI	A 146	G 125 +21
M45	55	ATTGGCAGTGAAAAATTATAGCTA TGCCTTTGCTACAACTCTCCTA	BfaI	G 140 +22	A 162

对Y染色体上的8个STR分型(DYS19、DYS388、DYS398－1、DYS389－2、DYS390、DYS391、DYS392和DYS393)。引物和Y染色体STR大小信息在GDB数据库(http://www.gdb.org)中。正向引物分别标记荧光6－FAM(DYS388、DYS389)、NED(DYS390、DYS392、DYS393)和HEX(DYS19、DYS392)。依据文献中的报道，Y染色体STR通过扩增并在ABI PRISM 377测序仪上电泳[21]。用Genescan™2.0和Genotype™1.1软件(Applied Biosystems)读取STR长度以区分基因型。通过直接测序确认了两个片段的长度，以此作为标准样本加到电泳里用作对照，测量不同凝胶下的基因型读数。

3. 数据分析

360 bp(线粒体区段16 024～16 083)长的HVS－1序列通过参照牛津标准序列(CRS)用ClustalW软件对位[23]。用Arlequin计算F_{ST}和Nei的遗传距离d_A[24]。在被比较的两个群体没有差异的零假设下，通过3 000次排列来检验F_{ST}的统计显著性。每个个体的单倍型和所在单倍群通过姚永刚等研究确定的HVS－1区域的诊断位点来推断[26]。通过宿兵等的研究中系统命名法推测Y－SNP单倍型[14]。对于Y－STR，通过Arlequin估算R_{ST}[27]，在群体间没有差异的零假设下通过3 000次排列来检验R_{ST}的显著性。由SPSS软件进行遗传距离的多维分析(MDS)、主成分分析(PCA)和关联分析。由MEGA软件构建邻接法(NJ)进化树[28]。

在Y－SNP单倍型的主成分分析中，包括藏族[15]、云南汉族[18]、四川汉族[18]、壮族[14]、蒙古族[14]、布依族[19]和维吾尔族(未发表)。研究中也包括云南汉族[26]、广东汉族[26]、青岛汉族[26]、新疆汉族[26]、维吾尔族[29]、哈萨克族[29]、土族[30]、蒙古族[30]、傣族[30]、壮族[30]和佤族[30]的线粒体DNA数据。

3.7.3 研究结果

1. Y-SNP 单倍型频率和线粒体 DNA 单倍群

表 3-17 显示 Y-SNP 单倍型在 6 个群体里的频率。研究中共观测到 10 种单倍型，摩梭人、白族和彝族中都出现了 9 种，藏族和普米族中出现了 8 种，纳西族中只有 5 种。K-M9、D*-YAP 和 O3a5-M134 是在摩梭人和藏族中的主要单倍型，它们分别占两个群体的 68%和 74%。D*-YAP 和 O2a-M95 是普米族(占 70.2%)和纳西族(占 47.5%)中最普遍的单倍型。单倍型频率在白族和彝族中分布相对分散，说明这两个群体有较高的多样性。比较有趣的是，在纳西族中频率最高的单倍型 O2a-M95(占 47.5%)没有在摩梭人中观察到。

表 3-17 Y-SNP 单倍型频率

群体	样本大小	标记突变及单倍型(%)												
		M130	YAP	M15	M89	M9	M122	M7	M134	M119	M110	M95	M88	M45
		C	D*	D1	F	K	O3*	O3a4	O3a5	O1a	O1a1	O2a	O2a1	P
白族	50	6.0	2.0	4.0		18.0	8.0	4.0	34.0	6.0		12.0	2.0	4.0
彝族	50	8.0	2.0	10.0	38.0	16.0	10.0		10.0	2.0		4.0		
摩梭人	47	6.4	23.4	4.3	2.1	31.9	4.3	6.4	12.8	8.5				
纳西族	40	2.5	37.5			10.0			2.5			47.5		
普米族	47	2.1	70.2	2.1	4.3	6.4	2.1		6.4	4.3				2.1
云南藏族	50	4.0	28.0	8.0		12.0	6.0	4.0	34.0			4.0		

表 3-18 基于线粒体 DNA 高变 1 区估算的单倍型频率

线粒体 DNA 单倍群	云南汉族①	青岛汉族①	新疆汉族①	广东汉族①	白族	摩梭人	纳西族	普米族	云南藏族	彝族
	(43)	(50)	(47)	(69)	(37)	(46)	(45)	(35)	(36)	(40)
A	4.7	4.0	10.6		5.4	4.3	8.9	14.3	13.9	19.5
B*	2.3	2.0								
B4	11.7	10.0	2.1	29.0		13.0	17.8	2.9	5.6	12.2
B5*					5.4					
B5a	4.7		4.3			17.4	6.6		11.1	
B5b	2.3		2.1	1.4						
B 合计	21.0	12.0	8.5	30.4	5.4	30.4	24.4	2.9	16.7	12.2
C	4.7		6.4			13.0	8.9	22.9	8.3	2.4

（续表）

线粒体DNA单倍群	云南汉族①	青岛汉族①	新疆汉族①	广东汉族①	白族	摩梭人	纳西族	普米族	云南藏族	彝族
	(43)	(50)	(47)	(69)	(37)	(46)	(45)	(35)	(36)	(40)
D4k	9.3	26.0	19.1	10.1	10.8	13.0	6.7	17.1	19.4	14.6
D5*	2.3	3.9	2.1	5.8						
D5a	2.3	6.0	4.3			8.7		5.7	5.6	2.4
D合计	13.9	35.9	25.5	15.9	10.8	21.7	6.7	22.8	25.0	17.0
F*	2.3		2.1	1.4						
F1a	11.6	4.0	4.3	17.4	16.2		17.8	2.9		4.9
F1b	4.7	4.0	2.1	1.4	5.4	17.4	4.4	2.9		4.9
F1c		2.0	2.1	1.4	2.7					
F2*				2.9						
F2a	2.3	2.0	4.3	1.4	8.1	4.3	2.2	5.7	2.8	2.4
F合计	20.9	12.0	14.9	25.9	32.4	21.7	24.4	11.5	2.8	12.2
G2		6.0	2.1	1.4	10.8		2.2			2.4
M10	2.3	2.0		2.9	8.1		8.0	2.9	2.9	9.8
M7*	2.3		2.1	1.4						
M7B	16.3	4.0	6.4	8.7	5.4		2.2			12.2
M7c			2.1	1.4						
M7合计	18.6	4.0	10.6	11.5	5.4		2.2			12.2
M8a		8.0	4.3	2.9						
M9		4.0	4.3	0.0	5.4			5.7	5.6	7.3
N9a	7.0	6.0		1.4						
R9a	2.3		4.3	1.4			2.2			
Y		2.0	2.1							
Z			2.1						2.8	
其他②	4.6	4.0	4.2	5.7	16.2	8.6	11.1	17.1	25	4.9

注：① 数据来自参考文献[26]。② 包含未定义的M*、N*及R*支系。

在分析的293个个体中，观察到177种HVS-1线粒体DNA的单倍型，其中47个摩梭人中含有39种单倍型。在摩梭人的39种单倍型中，29个是摩梭人特有的，5个与纳西族共有，3个与普米族共有，2个与白族、藏族和彝族共有。根据姚永刚等的标准，基

于 HVS－1 序列，这些单倍型被归类到不同的单倍群[26]。大部分单倍型(87%)都能被明确归到某一个单倍群，它们的频率如表 3－18 所示。B、D 和 F 是摩梭人中最常见的单倍群(占 73.8%)，单倍型 B、F 在纳西族中的频率与摩梭人非常相近。F 和 G 在白族中是主要的单倍群，单倍型 A、D 在藏族、普米族和彝族中最普遍(单倍群 C 在普米族中的频率非常高)。

2. 遗传距离

采用 Y－STR 数据估算 R_{ST} 距离。基于 8 个 Y－STR 的 R_{ST} 矩阵显示(表 3－19)，在成对比较中，摩梭人与藏族有最小的遗传距离，并且群体没有差异的零假设不能被否定。摩梭人和纳西族的 R_{ST} 距离是摩梭人和藏族 R_{ST} 距离的 6 倍。对于线粒体 DNA 的 HVS－1 数据，使用 F_{ST} 和 Nei 的网状遗传距离(d_A)来估算遗传距离(表 3－20)。线粒体 DNA 的两个遗传距离显示非常好的关联($r=0.974$，$P<0.01$)。有趣的是，摩梭人与纳西族的两个遗传距离远小于摩梭人与其他群体的遗传距离。由线粒体 DNA 估算的遗传距离显示出与 Y－STR 不同的结果($r=0.298$，$P=0.28$)，说明摩梭人父系和母系世系有着明显不同的遗传历史。

表 3－19 Y－STR R_{ST} 距离(下三角)及 P 值(上三角)

	白 族	彝 族	摩梭人	纳西族	普米族	云南藏族
白 族		0.003	0.012	0.000	0.000	0.005
彝 族	0.048		0.001	0.000	0.000	0.009
摩梭人	0.043	0.076		0.000	0.000	0.101
纳西族	0.260	0.191	0.138		0.003	0.006
普米族	0.410	0.334	0.238	0.111		0.000
云南藏族	0.065	0.057	0.021	0.084	0.193	

3. 主成分分析

基于 Y－SNP 单倍型频率的主成分分析(图 3－23)中的前 3 个主成分解释了总差异度的 70.3%。群体分为明显的 4 簇：A(云南藏族、藏族、摩梭人、普米族)，B(白族、傣族、四川汉族、云南汉族)，C(蒙古族、维吾尔族和彝族)和 D(壮族、布依族和纳西族)。第二个主成分(PC2)将 5 个云南西北部的群体(A 簇中的群体和纳西族)与其他群体分开。主成分(PC)的结果与 D*－YAP 的频率显著关联($r=0.844$，$P<0.01$)，显示高频率的 D*－YAP是云南西北部群体的主要特征。第一个主成分(PC1)和第三个主成分(PC3)将 A 簇和 D 簇分开，而且它们与 O3a5－M134($r=0.810$，$P<0.01$)和 O2a－M95($r=0.898$，$P<0.01$)的频率显著关联。正是 O3a5－M134 和 O2a－M95 的频率差异标记了摩梭人和纳西族的遗传差异。

在线粒体 DNA 单倍型的主成分分析中，B、D、F 和 M7 的子单倍型分别被并在一起(图 3－24)。前两个主成分解释了所有差异的 76.5%。图中可以观察到两个簇：北方簇

表 3-20 线粒体 DNA 高变 1 区序列 F_{ST}(下三角)及遗传距离 d_A(上三角)

		1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	白族		0.845	0.313	0.451	0.499	0.196	0.264	0.199	0.269	0.349	0.497	0.508	0.188	0.529
2	摩梭人	0.079		0.211	0.779	0.515	0.492	0.441	1.329	0.517	0.504	1.045	0.312	0.687	1.131
3	纳西族	0.031	0.019		0.569	0.353	0.203	0.108	0.630	0.201	0.214	0.572	0.137	0.263	0.572
4	普米族	0.052	0.071	0.054		0.438	0.305	0.661	0.939	0.680	0.445	0.60	0.927	0.376	0.504
5	云南藏族	0.057	0.048	0.034	0.048		0.273	0.514	1.162	0.481	0.271	0.642	0.540	0.394	0.624
6	彝族	0.022	0.046	0.020	0.033	0.030		0.215	0.702	0.282	0.049	0.250	0.400	0.140	0.579
7	云南汉族	0.032	0.043	0.011*	0.073	0.058	0.024		0.486	−0.019	0.145	0.545	−0.012	0.245	0.685
8	佤族	0.028	0.130	0.069	0.117	0.141	0.086	0.064		0.541	0.722	0.976	0.783	0.584	0.965
9	壮族	0.035	0.058	0.025	0.082	0.060	0.036	−0.002*	0.072		0.217	0.573	0.019	0.229	0.580
10	土族	0.045	0.050	0.022	0.055	0.034	0.006*	0.018	0.105	0.029		0.275	0.242	0.098	0.426
11	蒙古族	0.059	0.085	0.046	0.072	0.070	0.022	0.061	0.147	0.070	0.039		0.766	0.211	0.605
12	傣族	0.062	0.031	0.014*	0.104	0.063	0.046	−0.002*	0.106	0.003*	0.033	0.093		0.380	0.874
13	维吾尔族	0.026	0.071	0.030	0.049	0.050	0.018	0.032	0.086	0.031	0.014*	0.029	0.051		0.195
14	哈萨克族	0.069	0.108	0.059	0.063	0.076	0.068	0.084	0.141	0.075	0.062	0.089	0.111	0.029	

* 表示 F_{ST} $P>0.05$。

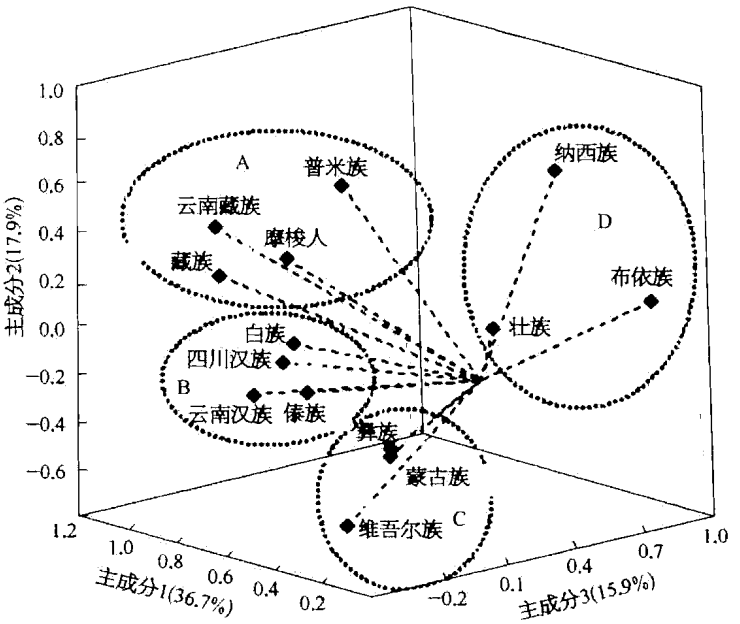


图 3-23 Y-SNP 单倍型主成分分析图

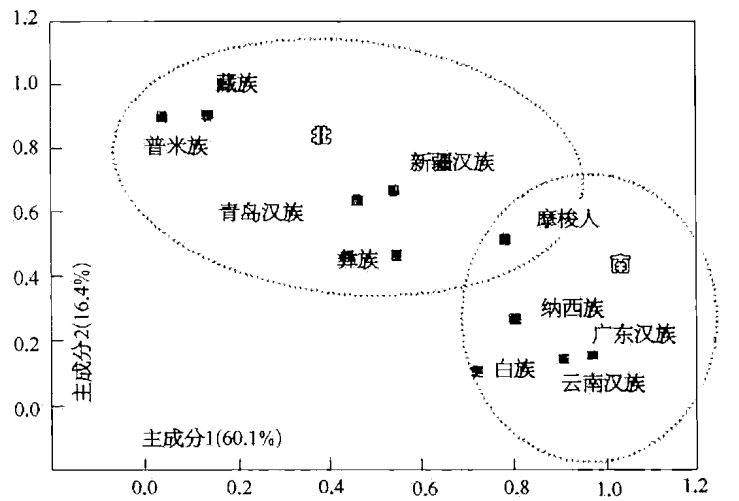


图 3-24 基于线粒体 DNA 单倍型频率构建的主成分分析图

包括云南藏族、普米族、彝族和两个北方汉族群体(青岛、新疆),而纳西族、白族和两个南方汉族群体(云南和广东)位于南方簇。PC1 分别与 F 单倍型($r=0.799, P<0.01$)和 B 单倍型($r=0.654, P<0.05$)的频率显著关联。PC2 与单倍型 D($r=0.721, P<0.05$)显著相关以及与单倍型 F($r=-0.876, P<0.01$)有强的负相关。这显示单倍群 B、D 和 F 造成北方簇和南方簇之间的差异。有趣的是,摩梭人位于北方簇和南方簇之间,它的 PC1 和 PC2 分别靠近南方簇和北方簇。用线粒体 DNA 的 F_{ST} 做主成分分析(图 3-25),

结果与用单倍型频率做主成分分析非常相似。南方簇和北方簇被 PC2 分开，并且摩梭人位于它们之间。

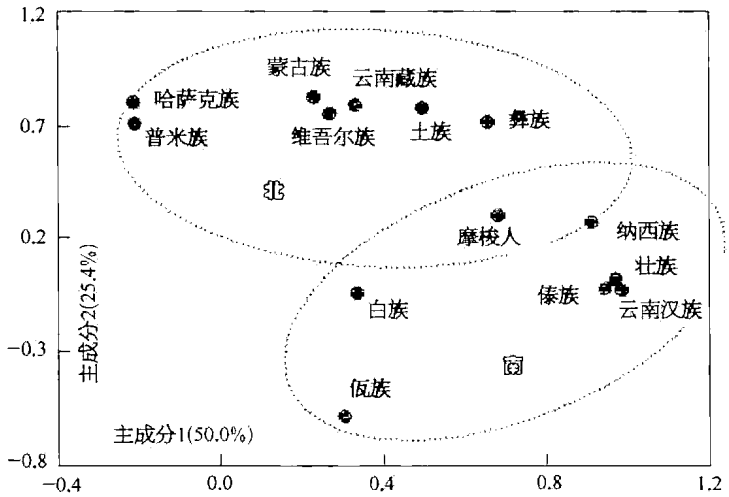


图 3－25　基于线粒体 DNA F_{ST} 距离矩阵构建的主成分分析图

4. 进化树分析

图 3－26 是进化树(邻接树)和由 Y－STR 的 R_{ST} 矩阵构建的 MDS 图像，它们都显示了一个相同的结构即纳西族和普米族形成一个簇，摩梭人、云南藏族、彝族和白族形成另一个簇。摩梭人和云南藏族显示较近的距离，支持由遗传距离观察到的结果。

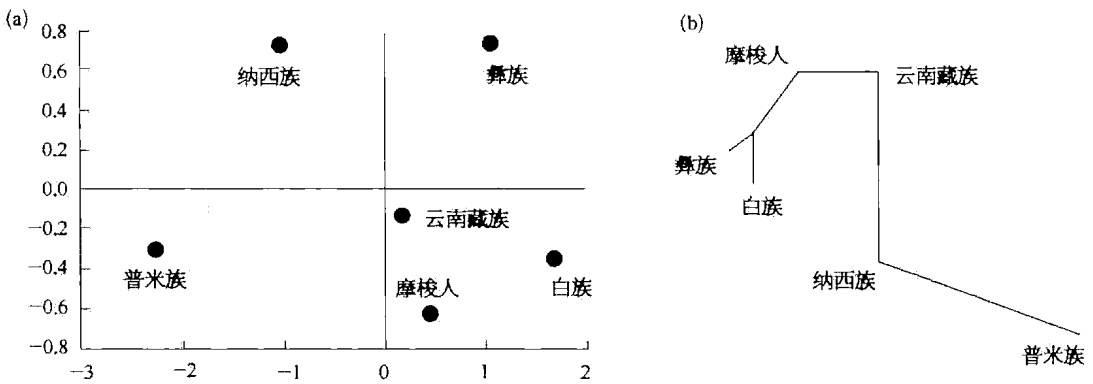


图 3－26　基于 Y－STR 的摩梭人与相关群体遗传距离分析

(a) 多维标度分析图；(b) 基于 Y－STR R_{ST} 距离构建的邻接树

相似的，基于线粒体 DNA F_{ST} 矩阵的进化树与主成分分析的结果一致(图 3－27)。群体被分成两个簇。北方簇包括土族、维吾尔族、哈萨克族、彝族、普米族和云南藏族，而其他族群都位于南方簇。在南方簇中摩梭人与纳西族的关系最接近。

3.7.4　讨论

在 Y－SNP 中，摩梭人主要的 Y 染色体单倍型包括 D＊－YAP、K－M9 和 O3a5－

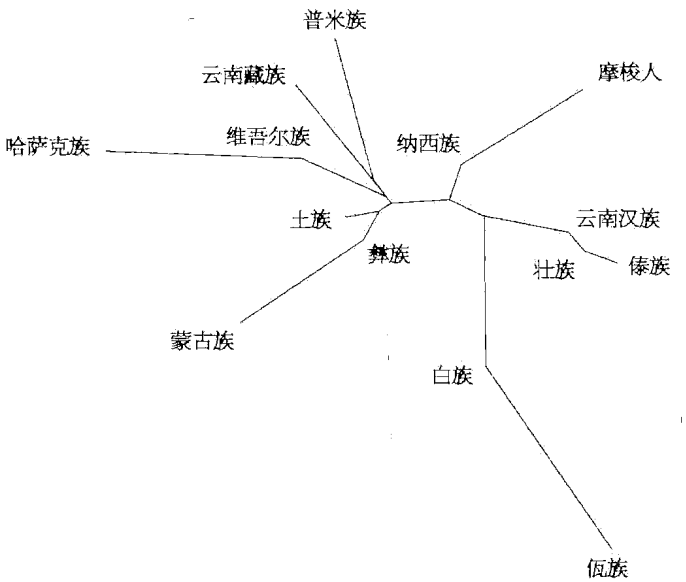


图 3-27 基于线粒体 DNA F_{ST} 距离矩阵构建的邻接树

M134。*Alu* 插入如 YAP(D*-YAP 和 D1-M15)被认为是最早到达东亚西南部的类型，而 O3a5-M134 是典型的东亚单倍型[15]。摩梭人和云南藏族 Y-SNP 单倍型频率分布的显著相关性($r=0.642$，$P<0.05$)显示他们的父系世系有相同的遗传历史。O2a-M95和 O1a-M119 在东亚南部的人群中广泛分布但在北部非常稀少，因此它们的出现标记着南方人群的特征。在摩梭人中只观察到 8.9%的 O1a-M119，而 O2a-M95 完全缺失，显示摩梭人中缺少南部的 Y 染色体世系。相反，纳西族有较高频率的 O2a-M95(47.5%)，它是摩梭人和纳西族父系遗传结构差异的主要原因，主成分分析和进化树分析也支持这一点。

对于显示母系世系的线粒体 DNA，单倍群 B (30.4%)、F(21.7%) 和 D(21.7%)在摩梭人中频率最高。单倍群 B 和 F 的频率分布在东亚人群中由南向北减少，显示高频率的 B 和 F 是南方人群的特征[26,31]。在摩梭人和纳西族中，这两种单倍群的频率非常相近，并且所有单倍群的频率分布在两个人群中显著相关($r=0.827$，$P<0.01$)，显示这两个母系支系有显著的相似度，因此可能有相同的遗传历史。单倍群 D 在人群中的频率分布与单倍群 F 成显著负相关($r=-0.785$，$P<0.05$)，并且在汉族中由南向北降低[26](表 3-19)。这个观察显示高频率的单倍群 D 可能是北方人群的一个特征。摩梭人线粒体 DNA 的频率分布类型显示其基因库受到南方和北方世系的共同影响，这同样也在主成分分析和进化树的结果中显示出来。

综合父系和母系世系得出的观察，我们认为摩梭人的父系与云南藏族有着较强的遗传亲缘关系，其中有少量的南方世系。而摩梭人的母系似乎与纳西族有较强的亲缘关系，其中主要是南方世系和相当数量的北方世系。摩梭人的父系和母系的基因特征即遗

传历史有着明显的差别。在完整的基因特征谱上纳西族和云南藏族与摩梭人都没有充分的遗传相似度。

单倍群(单倍型)归类是一种根据基因亲缘关系对线粒体DNA和Y染色体分类的基本方法,并且它的可靠性依赖于分型标准和单倍型之间真实进化关系之间的一致性。为进一步解析单倍群(单倍型)分析的结果,用Y-STR和HVS-1直接进行主成分分析和进化树分析。这些结果与用单倍群分析得到的结果非常一致。由于使用不同的基因标记和分析方法得出的结果一致,结论的可靠性更强。

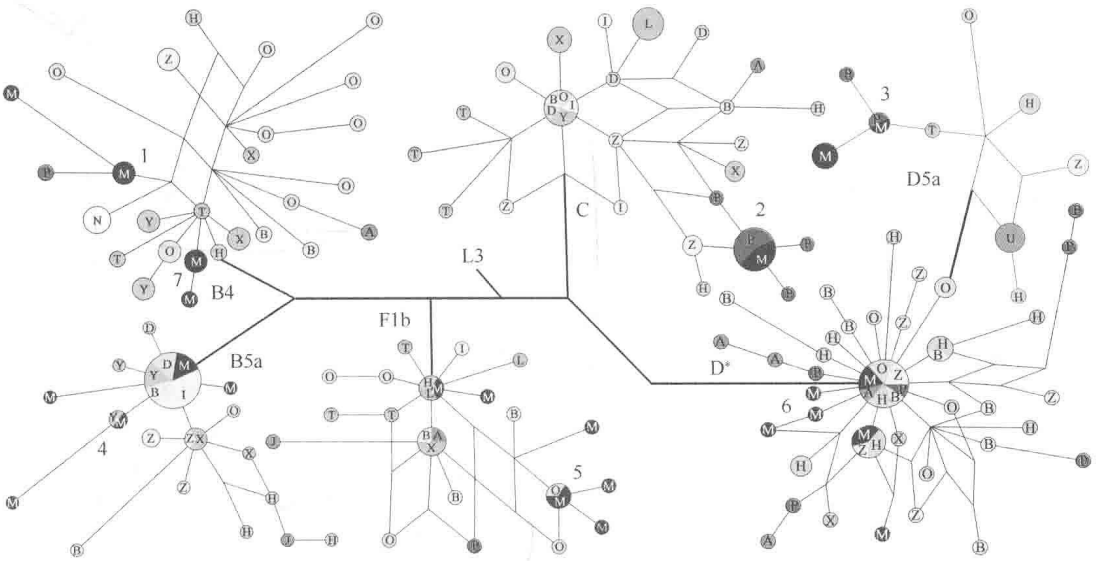
人群中父系和母系遗传历史的差别此前已有报道[32,33],但摩梭人中这种显著的差别是异常的。性别偏向的迁徙类型是最有可能的原因。根据历史记载,云南省的藏缅群体包括摩梭人来自最初居住于中国西北地区的古代氐羌族群,从2 700年前开始长途迁徙进入云南之后又重新分布[34]。不同的批次和群体的迁移、不同群体间的基因混合,以及迁入者和早期居民间的基因混合很可能造成了今天藏缅群体多样性的复杂局面。历史记载中摩梭人和纳西族有着相同的氏族名(Mosha),而且都起源于古代的牦牛羌。在大部分历史中,他们的地理位置有很大的重合部分[35]。而且,虽然他们使用不同的方言但都可能属于纳西族语言。他们母系世系的相似处与相似的语言和历史一致。他们基因库中的南部线粒体DNA世系可能是Mosha人在早期迁入时与当地原住民混合时引入的。摩梭人和藏族来源于古代氐羌不同的分支,藏族是发羌的后裔[34]。摩梭人和藏族的语言分别属于缅彝语支和喜马拉雅语支。虽然摩梭人和藏族在父系世系上有很近的遗传关系,历史和语言的证据并不支持摩梭人和藏族的共同祖先比摩梭人和纳西族的共同祖先更近。

摩梭人是中国大陆唯一的母系族群,而且仍然保留着独特的走婚制度:男人晚上到女方的屋里,黎明时回到他母亲的家里。根据1998年的调查,大约75%的摩梭人仍然实行这种婚姻制度[4]。这样的社会结构相比于其他母系社会和男性入赘的人群共同体提高了男性基因的贡献,同时也提高了父系基因结构的复杂性。人们对走婚制度的具体历史了解得很少,但一些人认为这是氐羌部落留下的传统[3]。因此这种制度可能实行了很长的时间。藏族文化对摩梭文化产生了重大的影响,并且由藏族传入的喇嘛教成为摩梭社会的主导宗教[36]。喇嘛以及摩梭妇女的走婚制度成为一个宗教仪式[4]。尽管藏族来源较少,随着父系世系的引入,摩梭人的父系遗传结构可能与藏族更接近。总的来说,来自其他族群的混合可能是摩梭人遗传结构变得更加复杂的主要原因,同时也伴随着早期混入的南部线粒体DNA世系和近期混入的父系世系。

人类群体的研究需要基于多学科的方法,遗传分析只提供众多方面中的一个层面。一个群体的完整历史只能通过历史学、语言学、民族学、人类学、考古学和遗传学的综合分析获得。这项研究的目的只限于提供有关摩梭人遗传历史的基因证据。

针对摩梭人的族源问题和母系社会的历史来源,对摩梭人的线粒体多样性进行了深入的分析[37]。因为婚后女性仍旧居住在母家,母系社会中的线粒体基因的流动性普遍较低,所以母系社会中的线粒体多样性应该较低。但是摩梭人的单倍群多样性并不比周边

的父系群体低，说明摩梭人的母系社会习俗可能开始得较晚，至少不是旧石器时代的残留。前面的单倍群频率聚类分析显示摩梭人接近纳西族，但是基于个体序列单倍型的网络结构分析显示大多的摩梭人世系与普米族相连（图 3－28）。所以，我们认为摩梭人与普米族的母系关系最紧密，这两个群体的女性服饰几乎完全一样（图 3－29）。



分支的共祖年代（年）

1:30270±12358	3:12108±12108	5:8072±5708	7:6727±6727
2:5045±2912	4:7338±2427	6:8649±3223	

图 3－28 中国西南若干群体的线粒体 HVS－1 单倍型网络结构

A 为拉祜族；B 为白族；D 为傣族；H 为哈尼族；I 为苗族；J 为基诺族；L 为傈僳族；M 为摩梭人；N 为怒族；O 为彝族；P 为普米族；T 为土族；U 为土家族；X 为纳西族；Y 为瑶族；Z 为藏族。

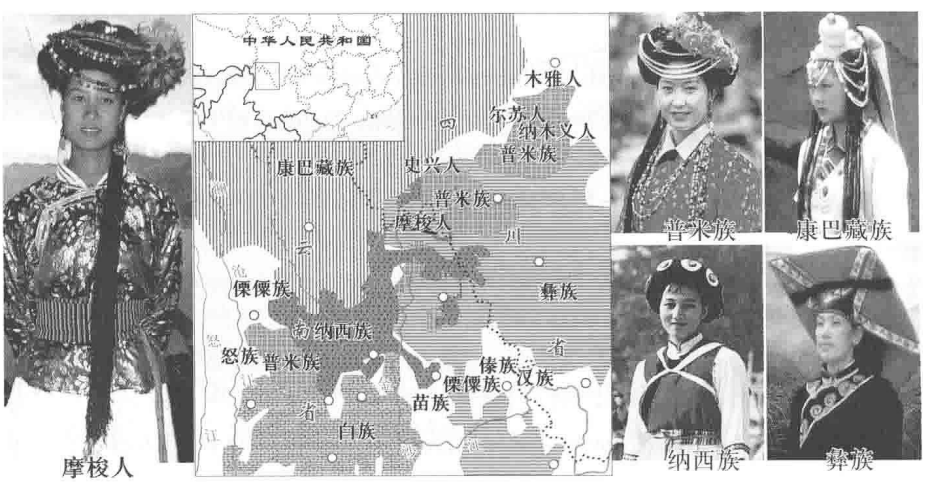


图 3－29 摩梭人和周边民族的女性服饰和地理分布

参考文献

[1] Gotteer-Abendroth H. The structure of matriarchal societies. Revision, 1999, 21: 31-35.

[2] You Z. The history of nationalities in Yunnan (in Chinese). Kunming: Yunnan University Press, 1997: 19-39.

[3] He J. A brief discussion on the essence and existing value of Mosuo culture. Xue Shu Tan Suo, 1999, 5: 60-62.

[4] Xu J. Study on folk culture of Mosuo people. Yunnan Geographic Environment Research (in Chinese), 1998, 10: 89-95.

[5] Cann R L, Stoneking M, Wilson A C. Mitochondrial DNA and human evolution. Nature, 1987, 325: 31-36.

[6] Cavalli-Sforza L, Menozzi P, Piazza A. The history and geography of human gene. Princeton: Princeton Univ Press, 1994: 733-734.

[7] Xiao C J, Cavalli-Sforza L L, Minch E, et al. Principle component analysis of gene frequencies of Chinese populations. Sci in China Ser C, 2000, 43: 472-481.

[8] Du R H, Xiao C J. Genetic distances between Chinese populations calculated on gene frequencies of 38 loci. Sci in China Ser C, 1997, 40: 613-621.

[9] Jobling M A, Tyler-Smith C. Father and sons: the Y chromosome and human evolution. Trends Genet, 1995, 11: 449-456.

[10] Underhill P A, Passarino G, Lin A A, et al. The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. Ann Hum Genet, 2001, 65: 43-62.

[11] Underhill P A, Shen P, Lin A A, et al. Y chromosome sequence variation and the history of human populations. Nature, 2000, 26: 358-361.

[12] Su B, Xiao C, Deka R, et al. Y chromosome haplotypes reveal prehistorical migrations to the Himalayas. Human Genetics, 2000, 107: 582-590.

[13] Jin L, Su B. Natives or immigrations: modern human origin in East Asia. Nature Genetics Reviews, 2000, 1: 126-132.

[14] Su B, Xiao J, Underhill P, et al. Y chromosome evidence for a northward migration of modern humans in East Asia during the Last Ice Age. American Journal of Human Genetics, 1999, 65: 1718-1724.

[15] Qian Y, Qian B, Su B, et al. Multiple origins of Tibetan Y chromosomes. Human Genetics, 2000, 106: 453-454.

[16] Su B, Jin L, Underhill P, et al. Polynesian origins: new insights from the Y chromosome. Proc Natl Acad Sci USA, 2000, 97: 8225-8228.

[17] Ke Y H, Su B, Song X F, et al. African origin of modern humans in East Asia: a tale of 12 000 Y chromosomes. Science, 2001, 292: 1151-1153.

[18] Ke Y, Su B, Xiao J, et al. Y chromosomes haplotype distribution in Han Chinese populations and modern human origin in East Asians. Sci in China Ser C, 2001, 44: 225-232.

[19] Li Y, Zuo L, Wen B, et al. Origins and migrations of Bouyei people in China-insights from Y chromosome and mitochondrion. Acta Genetica Sinica (in Chinese), 2002, 29: 196-200.

[20] Kayser M, Caglia A, Corach D, et al. Evaluation of Y chromosomal STRs: a multicenter study. Int J Legal Med, 1997, 110: 125 - 133.

[21] Kayser M, Krawczak M, Excoffier L, et al. An extensive analysis of Y chromosomal microsatellite haplotypes in globally dispersed human populations. Am J Hum Genet, 2001, 68: 990 - 1018.

[22] Lum J K, Cann R L, Martinson J J, et al. Mitochondrial and nuclear genetic relationships among Pacific Island and Asian populations. Am J Hum Genet, 1998, 63: 613 - 624.

[23] Anderson S, BAnkier A T, Barrell B G, et al. Sequence and organization of the human mitochondrial genome. Nature, 1981, 290: 457 - 465.

[24] Nei M, Kumar S. Molecular evolution and phylogenetics. New York: Oxford University Press, 2000: 256.

[25] Schneider S, Roessli D, Excoffier L. Arlequin Ver. 2.0: a software for population genetic data analysis. Switzerland: University of Geneva, 2000.

[26] Yao Y G, Kong Q P, Bandelt H J, et al. Phylogeographic differentiation of mitochondrial DNA in Han Chinese. Am J Hum Genet, 2002, 70: 635 - 651.

[27] Slatkin M. A new measure of population subdivision based on microsatellite allele frequencies. Genetics, 1995, 139: 457 - 462.

[28] Kumar S, Tamura K, Ingrid B, et al. MEGA2: molecular evolutionary genetics analysis software. Arizona State University, Tempe, Arizona, USA, 2001.

[29] Yao Y G, Lu X M, Luo H R, et al. Gene admixture in the silk road region of China: evidence from mtDNA and melanocortin I receptor polymorphism. Genes Genet Syst, 2000, 75: 173 - 178.

[30] Yao Y G, Nie L, Harpending H, et al. Genetic relationship of Chinese ethnic populations revealed by mtDNA sequence diversity. Am J Phys Anthropol, 2002, 118: 63 - 76.

[31] Wallace D C, Brown M D. Lott MT mitochondrial DNA variation in human evolution and disease. Gene, 1999, 238: 211 - 230.

[32] Gibbons A. The women's movement. Science, 1997, 278: 805.

[33] Pennisi E. Tracking the sexes by their gene. Science, 2001, 291: 1733 - 1734.

[34] Cang M. Study on the migration culture of the ethnic groups in Yunnan (in Chinese). Kunming: Yunnan Nationality Press, 1997: 10 - 11.

[35] Wang Z H. History of nationalities in China (in Chinese). Beijing: China Social Science Press, 1994: 289 - 291.

[36] He S. A brief discuss on the inter-action of Tibetan and Naxi culture. Nationality Research (in Chinese), 1995: 289 - 291.

[37] Lu Y, Wang C C, Qin Z D, et al. Mitochondrial origin of the matrilocal Mosuo people in China. Mitochondrial DNA, 2012, 23(1): 13 - 19.

第4章　南方原住民族的Y染色体

东亚的Y染色体谱系大部分来源于南方。东南亚的群体被认为是东亚群体的祖先，这些群体的语言可以分为苗瑶语系、南亚语系、侗傣语系和南岛语系4种。尼格利陀人种的安达曼语系和澳大利亚人种的新几内亚和澳大利亚诸语系不能算作东亚人群的类型。东亚南方4种语系群体的分子人类学调查对于东亚人群起源研究尤为重要，可能揭示现代人由东南亚进入东亚的过程及独特的遗传结构的形成。

南亚语系的孟高棉语族和苗瑶语系语言文化中有诸多相似性。本章调查了东南亚和东亚周边地区的原住族群：孟高棉和苗瑶族群的47个群体的1 652个个体的Y染色体，结果显示Y染色体单倍群O2a－M95在多个群体中都存在高频现象，而单倍群O3a4－M7为孟高棉和苗瑶族群所特有，暗示了这两个族群较近的遗传关系。基于O3a4－M7单倍型的STR网络结构图显示出分层扩散的结构，南亚语系中的孟高棉族群所在单倍型位于网络结构的中心位置，苗瑶族群处于孟高棉的外圈，汉藏出现在最外围。这样的分层结构可能源于早期现代人从东南亚进入东亚过程中非常缓慢和均匀的，由无数个并行的小的“瓶颈”效应组成的“丛林过滤”效应。单倍群O3a4－M7整体年龄约为2.7万年前，早于1.6万年的末次冰川期。东亚人群的各种遗传特质可能源于这种“丛林过滤”效应。

南岛语系传统上分为4个语族：泰雅语族、排湾语族、邹语族、马来-波利尼西亚语族。前3个语族都在台湾岛内，而马来-波利尼西亚语族在东南亚、太平洋和印度洋有着非常广泛的分布，西至马达加斯加岛，东到复活节岛，南抵新西兰，北达夏威夷，是欧洲殖民时代之前分布最广的语族，内部群体之间非常相近。在语言学中，根据语言的相似性推导出了颇有争议的南岛语系起源假说，即中国台湾是马来-波利尼西亚人群的发源地。这一假说一直被民族学家、语言学家、考古学家以及遗传学家所争论。东部南岛语系群体（密克罗尼西亚和波利尼西亚人）来源于西部南岛群体（东南亚岛屿和台湾人群）这一观点已被广泛接受，而整个南岛语系族群的起源据推测是中国大陆的侗傣族群。

关于东部南岛群体波利尼西亚人的定居历史问题较复杂。有假设认为中国台湾是波利尼西亚人的故乡，而另有假设认为美拉尼西亚是波利尼西亚人的来源地，这两个假设都颇有影响。在本章中，笔者调查了生活在东南亚、中国台湾、密克罗尼西亚、美拉尼

西亚和波利尼西亚36个人群551个男性个体，分析了Y染色体样本上的19个双等位基因多态性构成的单倍型的分布。在美拉尼西亚和波利尼西亚群体中几乎没有发现台湾人的Y染色体单倍型。同样，美拉尼西亚特异性的单倍型在波利尼西亚人群中不存在。然而，所有的波利尼西亚人、美拉尼西亚人和台湾人的单倍型存在于现在的东亚大陆人群中。显然，Y染色体数据不支持目前任何一种流行的假说。准确地说，我们推测东亚大陆为两个不相关的移民过程提供了基因来源，基因流传方向一是台湾，一是通过东南亚群岛到达波利尼西亚。

东部南岛群体源于西部，而西部南岛群体由何种路线源于东亚大陆还需进一步研究。为此，笔者研究了包括中国大陆30个侗傣族群、印尼和越南23个马来-波利尼西亚群体以及中国台湾11个原住民群体，共计1 509份个体样本的Y染色体非重组区。这3组群体在父系遗传上有许多相似性。混合分析表明，侗傣族群很少受到汉族的基因影响，而他们对印尼人群有最大比例的遗传贡献。大部分群体样本中含有高频的单倍群O1a－M119，这在其他系统的民族群体中几乎没有。O1a＊的STR网络图显示了印尼人群并非按照语言学研究所推测的那样来源于台湾少数民族，而是来自侗傣族群。从父系遗传角度看，东南亚岛屿群体并没有台湾来源。另外，台湾少数民族以及印尼人群很可能是分别起源于侗傣族群。此后这两个群体似乎各自独立发展。我们的结果也从遗传上支持包括台湾少数民族、侗傣人群和马来-波利尼西亚人在内的一个超类群的存在，也就是说侗傣语系和南岛语系人群有着最近的父系遗传关系。

民族学研究普遍认为，东亚南方的4个语系人群的先民在新石器时代可能都居住于长江流域。这可以通过研究长江流域新石器时代的人类遗骸来验证。从远古遗骸中提取线粒体以及细胞核DNA来进行古DNA研究，这一领域发展至今已超过20年。进入21世纪以后，关于Y染色体的基因分析也开始应用于古DNA研究。在本章中，笔者对古代东亚人群DNA的Y染色体单倍群进行了检测分析，以求从遗传角度寻找古代人群与现代人群的关系。从长江沿岸5个考古遗址中采集了56份遗骸样本，遵循严格标准以避免污染。大部分样本中成功扩增出了Y染色体上5个SNP，至少62.5%的类型属于单倍群O，这一结果与东亚现代人群的频率相似。在长江口附近的良渚文化遗址中发现了高频的O1a－M119单倍群，这将良渚文化的先民与现代南岛族群与侗台族群联系在了一起。长江中游地区的大溪遗址发现一种罕见的单倍型O3d－M7，现今主要在苗瑶族群和孟高棉族群中有少量发现，提示大溪先民很可能是现代苗瑶人群的祖先。史前文化中所发现的明显遗传隔离，为中华文明多重起源提供了遗传基础。

侗傣语系分为黎语族、卡岱语族、侗水语族和壮傣语族4个语族。在侗傣语系人群中，海南岛原住民黎族是最早分化出的类群，也有着最原始的语言文化和遗传特征。而海南岛也处于东亚的南方东线入口处，通过此处的早期移民过程影响了大多数东亚人群的Y染色体多样性。为了探索海南岛隔离的遗传结构以及东亚南方入口处的原初遗传结构，我们对海南岛所有6个原住民群体的405名男性个体进行了Y染色体多样性研究，这些群体几乎未受到大陆人群回迁与基因交流的影响。这些群体的主要单倍群是O1a＊和O2a＊。

另外，大陆最主流的单倍群O3在海南岛原住民群体中的频率非常低，这表明海南岛原住民的遗传结构相对隔离。聚类分析显示，海南岛原住民的两个主流单倍群在3.1万年～3.6万年前进入东亚以后，从2万年前起就分离出来了。也就是说，海南岛原住民与大陆的隔离已经持续约2万年，其独特的遗传结构可作为许多其他人群遗传学研究的重要对照。

海南岛最西端还有一个特殊群体仡隆人，其起源一直是个谜。中国确认的民族共有56个，很多具有某种相似特征或者近缘关系的群体在民族识别时需要被归为一族。但事实上由于中国的族群关系极其复杂，在相关研究缺乏的情况下，民族识别工作非常艰难。而遗传学关系应当是民族识别的重要参考之一。本章试图用遗传学手段，对海南岛仡隆人的来源问题做出回答。仡隆人的语言属于侗傣语系下的卡岱语族，而该语族的其他人群（主要是仡佬族和布央人）都分布在遥远的中国西南地区。因为仡隆人的家谱中记载他们是福建迁来的汉族后裔，据此仡隆人被划归为汉族。笔者对78个仡隆个体样本进行Y染色体分型，分析了20个单核苷酸多态（SNP）和7个短串联重复序列（STR），结果共发现8种单倍群，其中O1a*为主要类型。通过主成分分析、聚类树分析和STR网络结构分析，将仡隆人与中国南部其他人群的Y染色体单倍群相比，发现在遗传结构上与其最接近的是仡佬族和黎族。与黎族的遗传相似性可能源于两个群体在海南岛上千百年来相邻而居所产生的基因交流，而相似的基因交流由于地理隔离在仡隆和仡佬之间不太可能发生，所以与仡佬族的遗传相似性更多是源于他们共同的祖先起源。鉴于语言学和遗传性证据均支持仡隆和仡佬的同质性，我们认为仡隆人应该起源于仡佬族。

海南岛三亚还分布着回族的一个独特类群回辉人。他们信仰伊斯兰教，语言属于南岛语系，与中国回族其他群体完全不同。一些历史文献和语言学分析表明，回辉人和越南占城人有着密切的文化关系，但是两者之间的遗传关系尚未可知。为了深入了解回辉人群的遗传历史，我们对102个回辉人样本的Y染色体和母系线粒体DNA进行分型。结果发现，Y染色体单倍群O1a*-M119和线粒体DNA单倍群D4、F2a、F1b、F1a1、B5a、M8a、M*、D5及B4a高频出现，其遗传结构类型与邻近的当地原住民十分类似。回辉、占城和其他东亚人群之间的聚类分析（主成分分析和网络结构分析）表明，相比占城和其他中南半岛人群，回辉人与海南的原住民更为接近。这些结果表明，回辉人的起源过程中可能伴随着对原住民大规模的同化。在同化过程中，回辉人的语言在结构上变成了有声调的孤立语，而他们的文化传统和自我认同却借由伊斯兰信仰得以保存。

4.1　苗瑶与孟高棉人群的遗传同源

4.1.1　研究背景

人类自诞生以来，从未停止过对自身的探索。现代人类的起源问题，也总能引起人们的探索热情和无休止的争论。对基因组研究的深入，为人类进化的探索注入了新的活力。对人类线粒体DNA和Y染色体非重组区上突变位点的研究推动了遗传物质在进

化领域的研究，得以分别从母系和父系遗传结构角度构建系统发生树，从而阐明世界各地不同人群之间的系统演化关系。

东亚地区一直在人类进化研究中占据非常重要的地位，其人群与东南亚和太平洋的人群关系相当紧密，与南亚也有着一定的联系，而且在当地挖掘出大量具有相当连续性的古人类化石。但是对于人群走出非洲后，是由北方还是南方进入东亚地区仍存在很大争议。走出非洲以后，需要绕过青藏高原才能到达东亚，因而不可避免会存在不同的途径，有人认为是从北边进入[1]，而更多对东亚地区人群 Y 染色体非重组区的研究则支持现代东亚人群走出非洲后，从东亚南部进入东亚地区，随后分别向南向北扩张[2-8]。

距今 3 万年左右，产生了对东亚地区现代人口分布影响最为深远的 Y 染色体 O－M175突变。其下游突变形成的单倍群 O3－M122、O2a－M95 和 O1a－M119 在东亚人群中所占的频率达 57% 以上，由此不难推断出 O 型为东亚的主要单倍型，并且 Y－SNP 和 Y－STR 的数据也证明 O 单倍型下面主要的 3 个分支起源于南方[5,7,9-11]。所以，目前可以确定的是东亚人群的主体从南部经过东南亚进入东亚。研究东南亚和东亚连接处的群体遗传结构可以探索东亚人群的起源发生过程。

由于 O2a－M95 和 O3－M122 主干单倍群起源太早，在东亚人群扩散之前就已经有了很高的多样性，所以分布广泛而且杂乱，分析的清晰度并不足以分辨群体之间的关系，因而得出的南方起源的迁徙路线略显粗略。对于较晚产生的单倍群的分析可能得到更清晰的结果。再者，先前对南方入口处人群的研究尚不够深入，对于东亚大陆入口处的两大关键原住群体——孟高棉族群（MK）和苗瑶族群（HM）（图 4－1）调查非常少。孟高

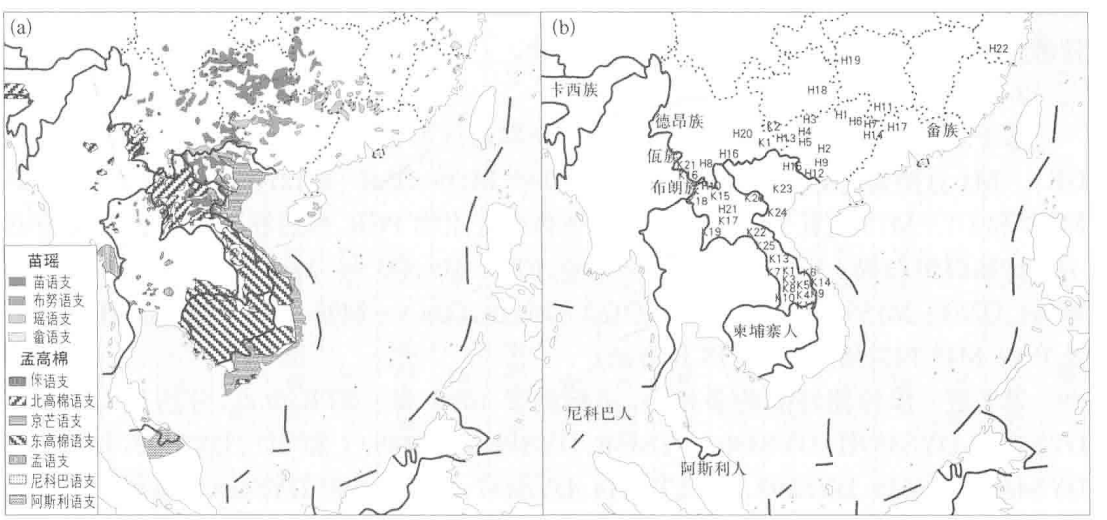


图 4－1 苗瑶族群和孟高棉族群调查研究

（a）苗瑶和孟高棉族群的分布；（b）采样群体的分布 b 图中群体编码与表 4－1 对应。

棉语族包含约100种语言，族群约3 500万人，主要生活在中南半岛地区，分布于印度东北部到柬埔寨和越南等东南亚许多地区。而苗瑶族群则主要分布在中国南部及东南亚地区，包括32种语言，族群在中国的人口超过1 200万人(2000年人口普查数据)。相比较而言，东南亚的汉藏和侗台族群则是较晚来到的外来群体[9-12]，他们分别从东亚北部和东部迁入东南亚的历史不超过3 000年。因而孟高棉和苗瑶群体在现代人进入东亚过程的研究探讨中格外重要。

本节通过对这两个族群Y染色体SNP和STR位点的完整调查，以及相关群体Y染色体信息的比较，发现单倍群O2a－M95在多个群体中都存在高频现象，仅单倍群O3a4－M7为这两个族群所特有，从O3a4的STR多样性结构上看到了现代人进入东亚过程中的扩散方式信息。

4.1.2　材料和方法

1. *群体采样*

本研究调查的群体涵盖范围较广，共采集22个苗瑶群体和25个孟高棉群体总共1 652个无关男性个体(表4－1)的外周静脉血或是指尖血。采样遵循知情同意的原则，随机采集无可查亲缘关系的健康男子血样。同一群体的样本在该群体的分布区内尽量分散，以确保样本的涵盖面。这些群体几乎覆盖了东南亚和东亚的交界处。然而有些群体人口很少，因而我们的样本量也相应较小(如木柄瑶)，在研究中也有一定意义。

外周静脉血样样品采用常规酚-氯仿法提取DNA，－70℃冻存。指尖血样品则保存在滤纸上。DNA提取时，先在滤纸上切下带血样的5 mm^2小纸片放入EP管，用TE浸泡过夜，然后经10 000转/min离心20 min弃上清。再加入100 μl TE震荡混合均匀，每管加入1 U蛋白酶K，混匀后55℃水浴2 h。之后95℃变性10 min。最后每管取2 μl液体进行全扩，全扩产物可用于后续的分型实验。

2. *遗传标记*

基于东亚和东南亚人群的特殊性，我们一共进行了20个Y－SNP位点的基因分型。DE＊－M1直接采用PCR产物电泳分型。O＊－M175、O3a1－M121、O3a5－M134、O3a5a－M117和D1－M15这五个位点采用Y染色体荧光引物PCR，然后在ABI 3130xl测序仪上进行基因组扫描。C－M130、F＊－M89、K＊－M9、O3＊－M122、O3a4－M7、O3a2－M164、O3a3－M159、O1a＊－M119、O1a2－M110、O2a＊－M95、O2a1－M88、Q1－M120及P＊－M45均采用RFLP－PCR方法。

为了进一步检测样品的多样性，还检测了16个Y－STR位点，分别是DYS456、DYS389I、DYS389II、DYS390、DYS458、DYS19、DYS385(I和II)、DYS393、DYS391、DYS439、DYS635、DYS392、YGATAH4、DYS437、DYS438和DYS448。采用多重降落式PCR反应条件，PCR产物在ABI 3130xl测序仪上进行基因组扫描，运用Genemapper软件进行基因分型数据分析处理。根据实验结果及2007年版Y－DNA单倍群进化树确定各个体的Y染色体单倍型。

表 4-1 苗瑶和孟高棉族群中 Y 染色体的 SNP 频率

编号①	语系②	群 体	ISO639-3	样本量	单倍群频率(%)														
					C	DE*	D1	F*	K*	O*	O3*	O3a4	O3a5	O3a5a	O1a*	O1a2	O2a*	O2a1	Q1	P*
H1	HM-N	巴哼	PHA	31					19.35	32.26	12.90	3.20	6.50	12.90			12.90			
H2	HM-N	布努-大化	BWX	20	5.00					20.00	15.00	10.00			20.00		10.00	20.00		
H3	HM-N	布努-南丹	BWX	10	50.00					20.00		30.00								
H4	HM-N	木柄瑶	BWX	6	50.00		16.67				16.67							16.67		
H5	HM-M	蓝靛瑶-田林	MJI	28				3.57		3.57		3.57	39.29	32.14			17.86			
H6	HM-M	平地瑶	MJI	31	3.23		19.35	3.23			12.90			16.13			16.13	29.03		
H7	HM-M	蓝靛瑶-贺州	MJI	41	2.44		7.32	2.44	4.88	4.88	14.63		7.32	12.20	12.20		24.39	7.32		
H8	HM-M	山瑶	MJI	32	3.13					3.13	6.25		50.00	6.25			31.25			
H9	HM-M	盘瑶-防城	IUM	31					9.68	32.26	6.45	6.45		6.45	6.45		19.35	12.90		
H10	HM-M	顶板瑶	IUM	11	9.09		9.09	9.09			9.09	9.09		27.27	9.09		9.09	9.09		
H11	HM-M	过山瑶	IUM	20					10.00		25.00	5.00	10.00	35.00	10.00			5.00		
H12	HM-M	花头瑶	IUM	19						21.05	15.79	5.26		31.58			21.05	5.26		
H13	HM-M	盘瑶-田林	IUM	33	3.03					9.09	21.21		12.12	18.18	9.09		21.21	6.06		
H14	HM-M	土瑶	IUM	41			12.20				58.54	7.32	4.88	7.32			7.32	2.44		
H15	HM-M	细板瑶	IUM	11				9.09		18.18	9.09			27.27	18.18		9.09	9.09		
H16	HM-M	盘瑶-红河	IUM	47			4.30		8.51	14.89	23.40	2.10	4.30	23.40	4.30		14.90			
H17	HM-M	八排瑶	BPN	37	5.41				2.70	5.41	16.22	32.43	2.70	13.51				21.62		

（续表）

编号①	语系②	群 体	ISO639-3	样本量	单倍群频率(%)														
					C	DE*	D1	F*	K*	O*	O3*	O3a4	O3a5	O3a5a	O1a*	O1a2	O2a*	O2a1	Q1	P*
H18	HM-H	苗-贵州	HMQ	49	8.16		2.04		4.08	24.49	24.49	4.08		2.04	12.24		16.33	2.04		
H19	HM-H	苗-湖南	MMR	100	14.00		1.00	2.00	9.00	9.00	24.00	4.00	7.00	9.00	7.00		9.00	5.00		
H20	HM-H	苗-云南	HMD	49	6.12		6.12		2.04	12.24	18.37	12.24		6.12	6.12		30.61			
H21	HM-H	苗-老挝香旷	MWW	51	25.49		7.84		5.88	3.92		33.33		7.84	1.96		5.88	5.88	1.96	
H22	HM-S	畲-温州	SHX	56	3.57			1.79		3.57	30.36	25.00	16.07	7.14	8.93		3.57			
K1	MK-B	花傈	BBH	32					3.13	31.25			6.25	9.38			50.00			
K2	MK-B	傣	PLY	30			3.33		3.33	10.00	30.00		6.67	10.00	10.00	3.33	23.33			
K3	MK-E	Alak	ALK	30						6.67		13.33		3.33			56.67	16.67		3.33
K4	MK-E	Brau	BRB	32				3.13		3.13	3.13	25.00					62.50	3.13		
K5	MK-E	Inh	IRR	34						5.88	2.94	8.82		2.94			79.41			
K6	MK-E	Jeh	JEH	32						6.25		46.88					46.88			
K7	MK-E	Suy	KDT	39						5.13	5.13	2.56					56.41	30.77		
K8	MK-E	Kataang	KGD	37						5.41	21.62	16.22		16.22			10.81	27.03		2.70
K9	MK-E	Katu	KUF	45						2.22		22.22			6.67		68.89			
K10	MK-E	Laven	LBO	50	2.00	24.00		4.00	4.00	4.00	2.00	12.00		2.00	2.00		42.00	2.00		
K11	MK-E	Ngeq	NGT	35								48.57					48.57			2.86
K12	MK-E	Oy	OYB	50						2.00		34.00		2.00			60.00	2.00		

（续表）

编号①	语系②	群　体	ISO639－3	样本量	单倍群频率(%)														
					C	DE*	D1	F*	K*	O*	O3*	O3a4	O3a5	O3a5a	O1a*	O1a2	O2a*	O2a1	Q1	P*
K13	MK－E	So	SSS	50					12.00	18.00	6.00	8.00		2.00			42.00	12.00		
K14	MK－E	Talieng	TDF	35	2.86					5.71		22.86		2.86			62.86	2.86		
K15	MK－N	Bit	BGK	28					3.57				10.71	32.14			53.57			
K16	MK－N	Blang	BLR	52	15.38				5.77	9.62	21.15		5.77	11.54			30.77			
K17	MK－N	Khmu	KJG	51	5.88				3.92	3.92	3.92	1.96		13.73			60.78	5.88		
K18	MK－N	Lamet	LBN	35	5.71						2.86	5.71					85.71			
K19	MK－N	Mal	MLF	50						4.00	4.00	2.00	2.00	14.00			66.00	8.00		
K20	MK－N	Xinhmul	PUO	29			3.45			3.45				6.90			17.24	68.97		
K21	MK－N	Ava	WBM	29	6.90				6.90	10.34	44.83		3.45	27.59						
K22	MK－V	Bo	BGL	28						14.29	3.57	7.14			3.57		64.29	3.57		3.57
K23	MK－V	京族	VIE	15				6.67		6.67	33.33		6.67	6.67	6.67		33.33			
K24	MK－V	芒族	MTQ	12				8.33		8.33	25.00		8.33	8.33			41.67			
K25	MK－V	Aheu	THM	38				2.63		13.16	15.79						52.63	15.79		

注：① 群体样本号和图 4－1b 一致。② 语系缩写如下：HM－N(苗瑶语系，布努语支)；HM－M(苗瑶语系，瑶语支)；HM－H(苗瑶语系，苗语支)；HM－S(苗瑶语系，畲语支)；MK－B(孟高棉语族，布甘语支)；MK－E(孟高棉语族，东孟高棉语支)；MK－N(孟高棉语族，北孟高棉语支)；MK－V(孟高棉语族，京芒语支)。

3. 数据分析

用 SPSS13.0 软件做主成分分析[13]，并进行聚类树分析，观察各群体间的亲缘远近。并使用 Golden Software Sufer7.0 绘制主成分地理分布图。

把研究对象群体中的 O3a4 - M7 个体的 Y - STR 单倍型数据用 Network4.2.0.0 软件绘制网络结构图[14]，并且把 O2a * - M95 个体的 Y - STR 单倍型数据与相关群体中同样单倍型的 STR 数据用该软件画网络结构图，分析各族群的来源[15]。由于参考的文献报道数据[7,11,15,16]不完整，分析中只用了 DYS19、DYS389I、DYS389II、DYS390、DYS391、DYS392 和 DYS393。

在 O3a4 单倍群中用 7 个 Y - STR 数据，运用 BATWING 对单倍群的分化时间做出估算[17,18]。计算中假设的有效群体大小 $N_e = 2\ 000$，突变率采用 Gusmao 等的数值 $(0.753\times10^{-3}\sim3.507\times10^{-3})$[18]。

4.1.3　研究结果

1. Y 染色体 SNP 单倍群分布

本项研究样本所涉及的单倍群频率见表 4 - 1。表中几乎观察不到 O1a2 - M110 的分布，仅在受侗台群体影响较深的侏人(Palyu)群体含有较少比例。P、Q 的频率也极低。苗瑶群体里 C、D 单倍群的分布比孟高棉稍多。显而易见，最普遍存在的单倍群是 O2a - M95，在苗瑶类群中最高频率可达 45.16%(O2a * 与 O2a1 之和)，在孟高棉类群中最高频率甚至达到 87.18%，但是这个单倍群在其他群体中也是广泛出现的。而 O3a4 - M7 在其他群体中则鲜见，为这两个群体所特有，存在族群特色[7,19]，为我们深入研究东亚入口处的族群及其在东亚族群形成中的作用提供了契机。

2. 群体的聚类分析

为了进一步了解这两个群体之间的亲疏关系，以及他们与东亚其他人群之间的亲缘关系，我们加入了已发表的其他 6 个群体的单倍型频率数据进行分析[5,9,16]。从主成分分析所绘制出的散点图(图 4 - 2a)中可以清晰地发现，东南亚的群体(南亚和侗台)与东亚的群体分向两个极端。东亚内部的群体又分为沿海(III)和内陆(II)两大群。而苗瑶和南岛处于这两端之间。说明苗瑶群体在研究人群从东南亚进入东亚过程中的作用举足轻重。

从同样用 SPSS13.0 绘制的聚类树(图 4 - 2b)分析产生的结果中，印度的 Dravidian 和 Indo - Aryan 作为外群，成为整个聚类树的根；汉族和藏缅语族人群距离最近，与我们设想的一样。南亚族群和苗瑶族群总体上非常接近，而且他们的亚群是交错的，其紧密关系可见一斑。

3. 两大主要单倍群 M7 和 M95 的地理分布

M7 和 M95 是孟高棉语族和苗瑶语族群体最主要的两个突变型遗传标记，深入研究这两个单倍群是探讨这两个族群遗传结构的关键。根据这两个遗传标记的突变型频率绘制地理分布图(图 4 - 3)，以期看到两个遗传标记的频率在族群和地理上的一些规

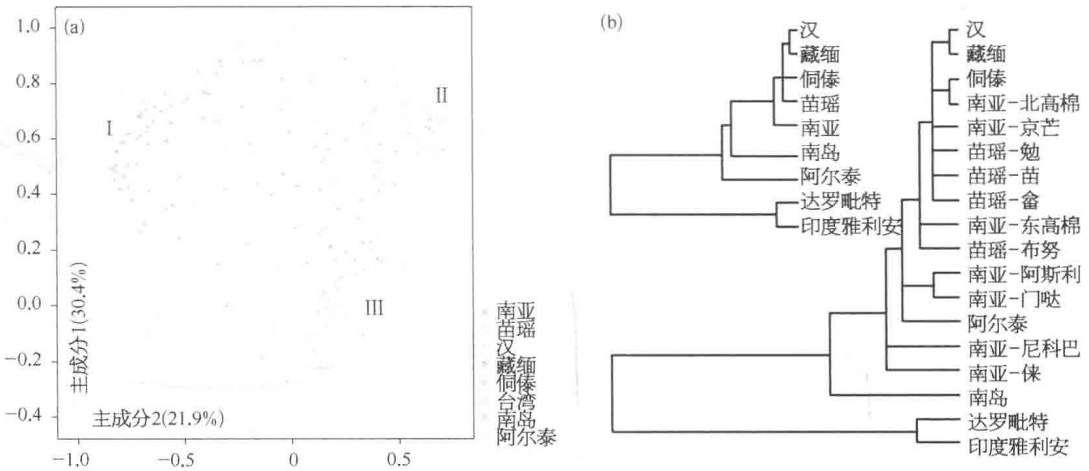
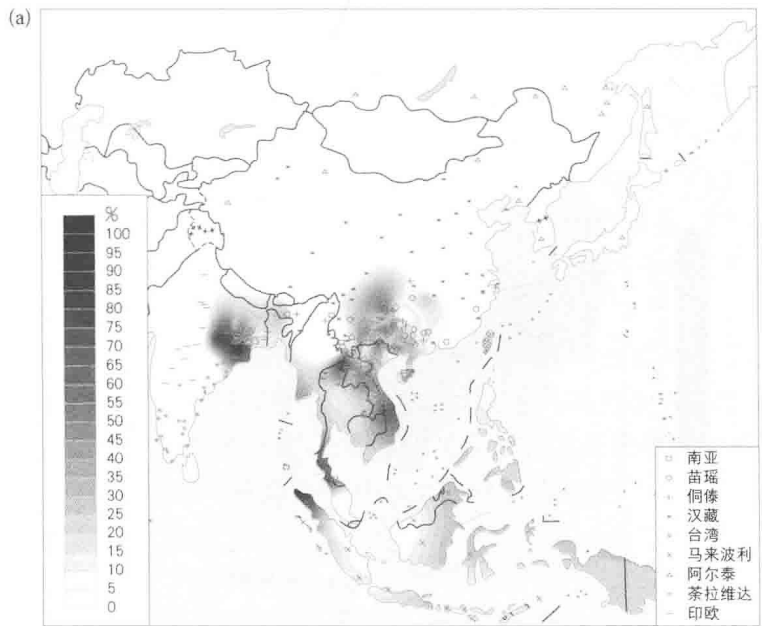


图 4-2 基于 Y 染色体 SNP 频率数据的群体聚类分析

(a) 主成分分析图;(b) 群体聚类分析



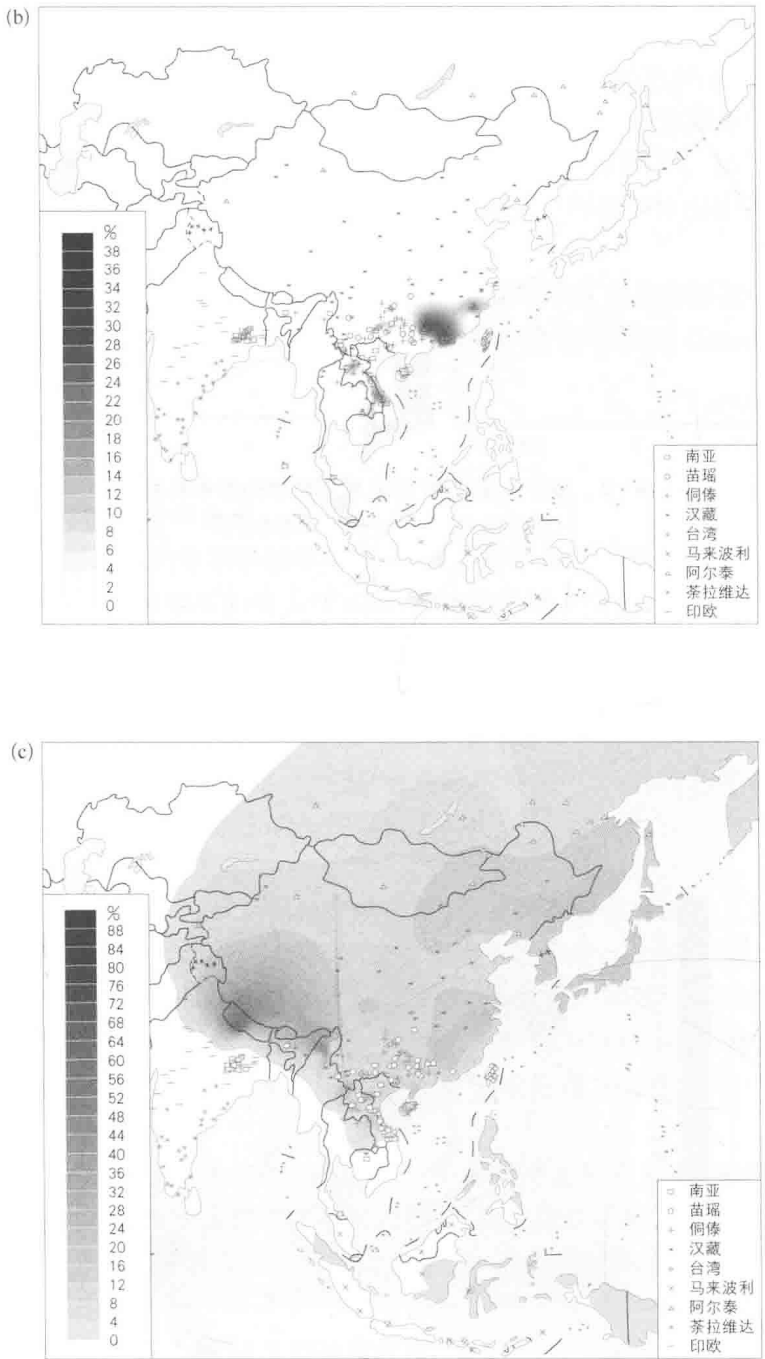


图 4-3　单倍群 M95、M7、M117 的地理分布及频率

(a) M95;(b) M7;(c) M117

律。从表 4-1 和图 4-2a 中可以清楚地看到，M7 突变型具有较高的民族性，在苗瑶中的畲族、布努中频率较高，在东高棉群体中也是相对高频。但是在侗台、汉族等其他民族系统中则几乎不存在，所以地理分布上被侗台等群体分割成零碎的块状。而 M95 突变型分布在从印度东北部到中国西南部的广大地理区域内，在各种族群中都具有较高的频率，并没有那么强烈的民族性。这提示 M95 突变型相对于 M7 突变型出现的时间要早得多，至少在整个南亚语系人群形成之初就出现了，并很早就影响了其后形成的整个东南亚和周边的群体。所以 M95 在印度的南亚语系人群中也有发现，并且多样性最高[20]。而 M7 突变型可能只是在人群将要从东南亚进入东亚的时候产生的，并且没有扩散到过多的群体中。

4. 两大主要单倍群 M7 和 M95 的网络结构图

对于特定遗传标记的起源和扩散过程，可以通过它所在单倍群的内部多样性的分析进行探讨。对观察到的所有 O3a4-M7 和 O2a-M95 个体进行 Y-STR 分析，并加入文献发表的数据[7,11,15,16]，分别构建单倍型网络图(图 4-4)。

全世界的 Y 染色体分型数据中，带有 M7 突变的 O3a4 单倍型在大量孟高棉和苗瑶人群中均有分布，而在其他人群中则极其罕见，显示出强烈的群体特异性。基于 O3a4-M7 单倍型的 STR 构建网络结构显示出极其明显的分层扩散结构。南亚语系中的孟高棉族群所在的单倍型位于网络结构中心，苗瑶族群的单倍型大多处于孟高棉单倍型的外圈，汉藏族群单倍型则出现在最外围处。这一点暗示了孟高棉人群的 O3a4 更为古老。这 3 个层次之间的分界线比较明显，很少有不同族群单倍型交错出现的情况。这说明整个 O3a4 单倍群在扩散过程中基本是单向地从孟高棉族群流向苗瑶族群，进而流向缅彝族群。并且这种扩散是缓慢的，以至于有足够的时间在新的群体中形成新的单倍型。同时扩散过程也是人口均匀增长的过程，所以在新的群体中多样性(单倍型种类)不断增加。在 O3a4 网络结构中，侗台族群所在单倍型的分布相对没有规律，零星地出现在中间和周边各处，并不构成独特的分支，说明侗台族群中少量 O3a4 单倍群很可能是与苗瑶和孟高棉族群较晚近交流过程中获得的。

基于 O2a-M95 单倍型构建的 Y-STR 网络图中，各种族群共享了大多数的单倍型，并没有类似 O3a4 单倍群的中心扩散的结构。这可能是因为该单倍群过于古老，在族群分化之前就已经有了高度的多样性，各种单倍型分散到了后期形成的各个族群中。值得关注的是，印度的门达群体的单倍型相对处于网络结构的中心位置，这与 O2a 单倍群的早期西部起源的观点一致。

5. 单倍群 O3a4-M7 的时间估计

O3a4-M7 的网络结构体现出其多态性有明显的族群辨析度，不同族群在网络结构中聚类在不同的分支，对这些分支估算年龄可以推算有关早期群体分化事件的大致年代。表 4-2 中是根据 BATWING 计算得到的单倍群整体年龄和几个重要分支的年龄。O3a4-M7 单倍群的整体年龄约为 27 000 年，这与东亚人群最初进入中国南方的时间大致相符，东亚地区最早发现的现代人是靠近东南亚的广西柳江人，距今大约 3 万年[21]。

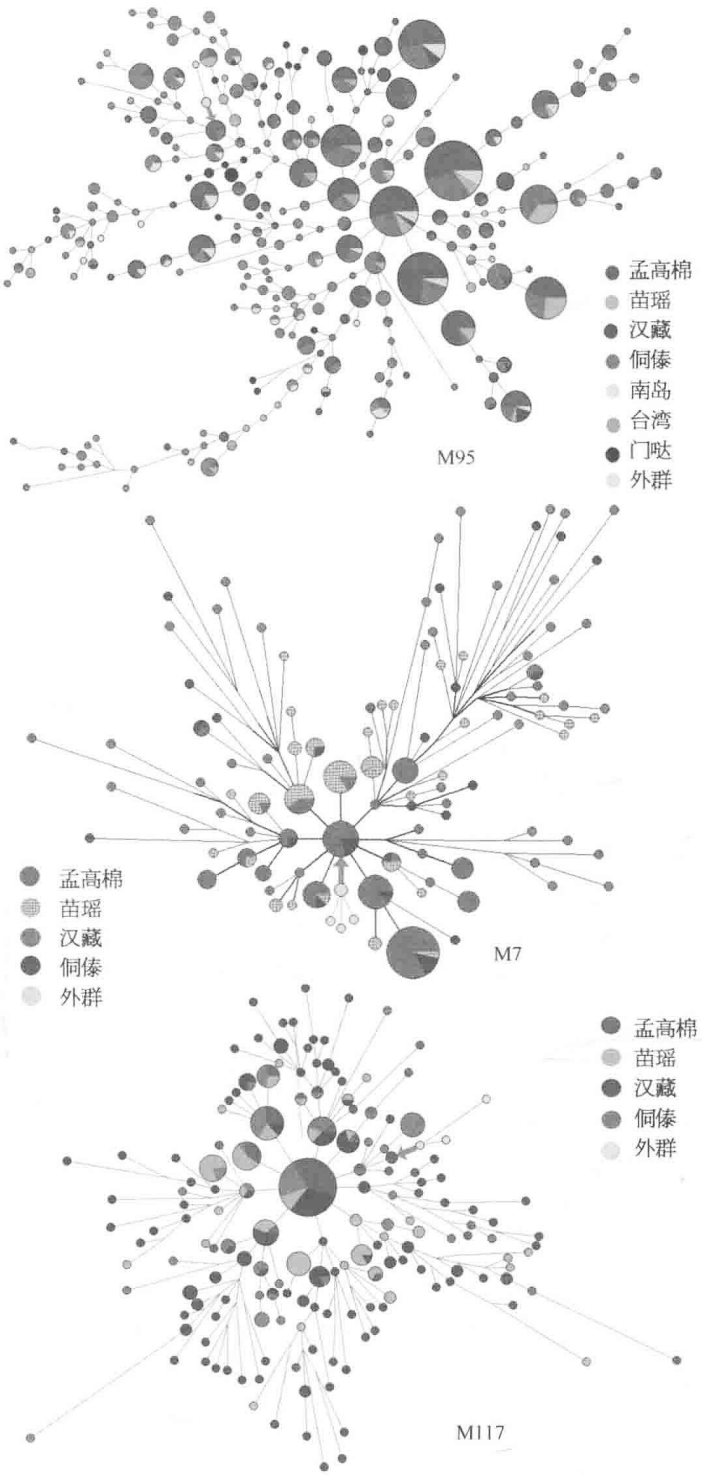


图 4－4　Y 染色体单倍群 O3－M7、O2－M95 和 O3－M117 的 STR 网络结构图

鹅黄色代表孟高棉群体(K)，紫红色代表苗瑶群体(H)，浅紫色代表汉藏群体(L)(本研究中，主要是 Lolo－Burman)，绿色代表侗台群体(T)，浅蓝色代表南岛群体(P)，橙色代表南亚门达群体(M)，青色代表台湾群体(W)。

在网络结构中上方有两个分支包含的苗瑶和缅彝单倍型比较集中，在图 4－4a 中分别标注为分支 I 和分支 II。表 4－2 中估计了这两个分支中苗瑶部分和缅彝部分的年龄。可见苗瑶部分的大致年龄在 14 000～16 000 年，而缅彝部分的年龄在 7 000～12 000 年。这些年龄与民族学、考古学和语言学估计的苗瑶、汉藏人群的发生历史比较吻合。

表 4－2 Y 染色体单倍群 O3a4 的时间估计

网络结构分支	年龄(年)	95%置信区间
整体	26 555	15.145～51.9
苗瑶-右	15 503	10.288～27.487
苗瑶-左	13 652	7.520～30.499
汉藏-右	12 464	7.555～22.322
汉藏-左	7 102	4.110～13.104

根据古气候学的资料，线粒体 DNA 的所有民族特异性(特别是苗瑶族群)亚单倍型都是在雨木(Yürm)冰期的最后一个冰退阶段[22]，而 Y 染色体单倍群 O3a4 中苗瑶部分的年龄，也处于这一时期。考古学和历史学的一些看法认为苗瑶族群至少在 1 万～1.7 万年前从“汉藏-苗瑶共同祖先”人群中分化并形成统一体[23]；古人类学证据表明，原汉藏祖先估计至少已经生存了 6 000 年[13,24]。因而这些数据也与古人类学证据和考古学年代基本吻合。

4.1.4 讨论

1. 苗瑶族群的孟高棉起源

本节研究的两个族群，孟高棉语族和苗瑶语族是中南半岛和中国西南部的原住民族，与东亚和东南亚的其他族群比较，在遗传结构上这两个族群的相似度非常高。这在聚类分析中有所体现，在 Y 单倍群的分布上更为明显。苗瑶和孟高棉族群的单倍群虽然都以 O2a＊为主，但它们都有区别于其他族群的特色单倍群 O3a4。O3a4 单倍群在除了孟高棉语族和苗瑶语族之外的族群中频率非常低，甚至不存在，而在孟高棉语族和苗瑶语族中却有相对较高的频率，这说明这两个族群的遗传关系必然比较亲密。O3a4 的 STR 网络结构体现出明显的分层扩散结构，苗瑶的单倍型分布在孟高棉单倍群的外围，而且年龄在 1 万年以上，这说明 O3a4 在这两个族群中同时出现并不是晚近时期内族群交流的结果，而是因为这两个族群有发生关系。苗瑶族群是在 1 万多年前从孟高棉族群中分化出来的一支，是从东南亚进入东亚的现代人的一支先锋。而汉藏族群则可能是稍后在苗瑶族群中分化出来的，所以缅彝分支的单倍型又处于苗瑶单倍群的外围。更北方的汉藏族群分支藏、羌、汉中很少见 O3a4 单倍群，可能是在族群迁徙过程中由于遗传漂变导致这一低频的单倍群丢失了。

在长江中游的考古学遗址中采集的 5 000 年前的人类样品中就发现有 O3a4 单倍

群[19]。这一地区是古代苗瑶祖先族群的主要聚居地。这说明至少在 5 000 年前，这一单倍群已经在古代苗瑶祖先族群中高频存在了，与我们的年龄估计也是吻合的。

苗瑶和孟高棉族群同源的观点也得到了语言学证据的支持。有观点认为苗瑶语族和孟高棉语族源于同一个大语系——原始长江语系(Proto-Yangzian Phylum)[25,26]。因此，苗瑶族群起源于南亚语系的孟高棉族群的观点在遗传学和语言学上都有相关的证据。但是在群体文化方面，苗瑶与孟高棉族群差异非常大，这可能是因为他们的分化在 1 万年以上，不同的生活环境和较快的文化发展速度使得两者向不同的方向发展，以至于相互的文化面貌截然不同，只有在基因组和语言词汇方面保留着两者关系的痕迹。

2. 东亚入口处的单向朝北通道

Y 染色体的各种单倍群由于年龄的差异，其内部的多样性结构差异也很大。本节分析了 O2a * - M95 和 O3a4 - M7 的 STR 网络结构，发现 O2a * 的结构比较不规则，在族群之间的分布也没有规律，而 O3a4 则体现出明显的单中心扩散的结构，不同族群处于不同的扩散层次上。对南亚地区门达族群的研究发现，O2a * 起源于这个族群，年龄将近 65 000 年[20]。在网络结构中，门达族群所在的单倍群的确处于结构的中心，符合起源族群的特征。在网络结构中门达和东南亚其他族群之间的分界比较明显，也符合中心扩散的层次结构。但是东南亚和东亚族群在 O2a * 的网络结构中则完全没有结构性。现代人从南亚迁徙到东南亚是 5 万～6 万年前，而从东南亚迁徙到东亚是 2 万～3 万年前。O2a * 的产生正是在现代人从南亚迁往东南亚的过程中，所以在其内部多样性结构中可以观察到南亚向东南亚的分化。而 O2a * 的大量多样性形成在东南亚人群向东亚迁徙之前，对东亚人群的分析就没有足够的清晰度了。而 O3a4 单倍群的年龄为大约 27 000 年，是最合适分析人群从东南亚向东亚迁徙过程的单倍群之一。

在东亚和东南亚之间有大量的山岭和丛林，这些是造成族群分化的地理隔离因素。比如安南山脉是分隔侗台和孟高棉的障碍物[16]。而云贵高原的丛林和峡谷是孟高棉和苗瑶之间的过渡层。根据我们的计算，群体至少在 16 000 年之前开始从中南半岛穿越这片丛林向东北方向进发。这一时期是末次冰川期的高峰期，高山的阻隔作用比现在更强，云贵高原北缘的几座雪山是人类无法逾越的障碍。从网络结构分析中可以推断，群体在穿越这个区域时伴随着人口增长的扩散。由于受到冰川期气候的影响和复杂地形结构的阻隔，扩散过程非常缓慢和均匀，新生的单倍型只出现在新生的群体中，而不会回流到较早的母群体中，而旧的单倍型在新生群体中却会丢失，以至于在迁徙方向越前方的群体的单倍型处在整个网络结构的越外围。这种现象就像是遗传标记被过滤过一样，新的单倍群才会被滤过，所以可以称为“丛林过滤”效应(图 4 - 5)。实际上，“丛林过滤”效应是由无数个并行的小的“瓶颈”效应组成的。东亚人群形成应该正是这种“丛林过滤”效应的结果，正是这种效应造成了东亚人群独特的体质特征和遗传结构，与东南亚的群体有了很大的不同。

研究发现的东亚南方人群的基因多样性高于北方人群，而且南方的单倍型种类包括

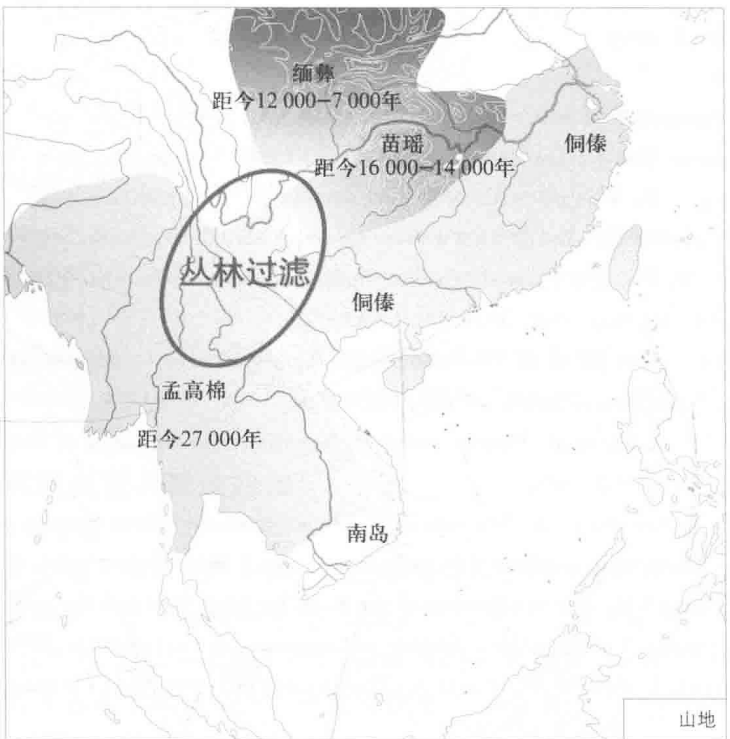


图 4-5 东亚入口处的"丛林过滤"假说

北方所有的种类[5]，这都无不揭示人类进入东亚始于南方，东南亚可能是早期由非洲前来的人群进入东亚的第一站。本研究又发现了人类当初进入东亚过程中的"丛林过滤"效应，这不但使东亚人类的起源过程越来越清楚，也有助于东亚人群的各种遗传特质性研究。

当然，从东南亚进入东亚南部的途径可能不止穿过云贵高原丛林的孟高棉-苗瑶这一条，在其东侧可能还有沿海迁徙的侗台-南岛这一条线路[16]，在中国西北部也可能有其他人群迁入[27]。通过合适的单倍群的分析，这些过程必然可以一一厘清。

参考文献

[1] Nei M, Roychoudhury A K. Evolutionary relationships of human populations on a global scale. Mol Biol Evol, 1993, 10(5): 927-943.

[2] Turner C G. Late Pleistocene and Holocene population history of East Asia based on dental variation. Am J Phys Anthropol, 1987, 73(3): 305-321.

[3] Ballinger S W, Schurr T G, Torroni A, et al. Southeast Asian mitochondrial DNA analysis reveals genetic continuity of ancient mongoloid migrations. Genetics, 1992, 130(1): 139-152.

[4] Chu J Y, Huang W, Kuang S Q, et al. Genetic relationship of populations in China. Proc Natl Acad Sci USA, 1998, 95(20): 11763-11768.

[5] Su B, Xiao J, Underhill P, et al. Y chromosome evidence for a northward migration of modern humans into Eastern Asia during the Last Ice Age. Am J Hum Genet, 1999, 65(6): 1718-1724.

[6] Yao Y G, Kong Q P, Bandelt H J, et al. Phylogeographic differentiation of mitochondrial DNA in Han Chinese. Am J Hum Genet, 2002, 70(3): 635-651.

[7] Shi H, Dong Y L, Wen B, et al. Y chromosome evidence of southern origin of the East Asian-specific haplogroup O3-M122. Am J Hum Genet, 2005, 77(3): 408-419.

[8] Zhang F, Su B, Zhang Y P, et al. Genetic studies of human diversity in East Asia. Philos Trans R Soc Lond B Biol Sci, 2007, 362(1482): 987-995.

[9] Su B, Xiao C, Deka R, et al. Y chromosome haplotypes reveal prehistorical migrations to the Himalayas. Hum Genet, 2000, 107(6): 582-590.

[10] Wen B, Li H, Lu D, et al. Genetic evidence supports demic diffusion of Han culture. Nature, 2004, 431(7006): 302-305.

[11] Wen B, Xie X, Gao S, et al. Analyses of genetic structure of Tibeto-Burman populations reveals sex-biased admixture in southern Tibeto-Burmans. Am J Hum Genet, 2004, 74(5): 856-865.

[12] Liang M, Zhang J R. The relationship of the Kam-Tai languages and the original dwelling areas and the migration of their peoples. Studies in Languages and Linguistics, 2006, 26(4): 8-26.

[13] Cavalli-Sforza L L, Menozzi P, Piazza A. The history and geography of human genes. Princeton, NJ: Princeton University Press, 1994.

[14] Bandelt H J, Forster P, Rohl A. Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol, 1999, 16(1): 37-48.

[15] Thanseem I, Thangaraj K, Chaubey G, et al. Genetic affinities among the lower castes and tribal groups of India: inference from Y chromosome and mitochondrial DNA. BMC Genet, 2006, 7: 42.

[16] Li H. Genetic Structure of Austro-Tai Populations. Ph. D. dissertation of Human Biology, Fudan University, 2005.

[17] Wilson I J, Weale M E, Balding D J. Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities. J R Stat Soc Ser A Stat Soc, 2003, 166(20): 155-188.

[18] Gusmao L, Sanchez-Diz P, Calafell F, et al. Mutation rates at Y chromosome specific microsatellites. Hum Mutat, 2005, 26(6): 520-528.

[19] Li H, Huang Y, Mustavich L F, et al. Y chromosomes of prehistoric people along the Yangtze River. Hum Genet, 2007, 122(3-4): 383-388.

[20] Kumar V, Reddy A N, Babu J P, et al. Y chromosome evidence suggests a common paternal heritage of Austro-Asiatic populations. BMC Evol Biol, 2007, 7: 47.

[21] Wu R C. Human fossils found in Liujiang Guangxi. Acta Anthropolagica Sinica, 1959, 1(3): 97-104.

[22] Li H, Cai X, Winograd-Cort E R, et al. Mitochondrial DNA diversity and population differentiation in southern East Asia. Am J Phys Anthropol. in press, 2007.

[23] Feng H. General history of Yao Nationality (Yaozu Tongshi). Beijing: Minzu Press, 2007.

[24] Martisoff J A. Sino-Tibetan linguistics: present state and future prospects. Annu Rev Anthropol, 1991, 20: 469 - 504.

[25] Starosta S. Proto-East Asian and the origin and dispersal of the languages of East and Southeast Asia and the pacific//Sagart L, Blench R, Sanchez-Mazas A. The peopling of East Asia: putting together archaeology, linguistics and genetics. London: Routledge Curzon, 2005: 182 - 197.

[26] Sagart L, Blench R, Sanchez-Mazas A. Introduction//Sagart L, Blench R, Sanchez-Mazas A. The peopling of East Asia: putting together archaeology, linguistics and genetics. London: Routledge Curzon, 2005: 1 - 14.

[27] Wells R S, Yuldasheva N, Ruzibakiev R, et al. The Eurasian heartland: a continental perspective on Y chromosome diversity. Proc Natl Acad Sci USA, 2001, 98: 10244 - 10249.

4.2 波利尼西亚人群的起源

4.2.1 研究背景

波利尼西亚史前人类定居的主体过程已经得到各学科领域的检验，由此提出了两种不同的人群迁徙模式。第一种被称为“快车模式”[1]，这一个模式主要建立在考古学和语言学证据上[2]，认为在 4 000～5 000 年前，在中国的南方发生了一次向东的人类迁徙，这一迁徙活动将南岛语系以及相关的拉皮塔文化传入太平洋岛屿，最终定居波利尼西亚。在这个模式中，毗邻亚洲大陆的中国台湾成为第一个定居点。这个假设由最近的线粒体 DNA 数据[3-6]支持，也将台湾少数民族和波利尼西亚人联系到了一起。第二种假说由 Terrell[7]提出，这个假说认为美拉尼西亚是波利尼西亚人的“邻近故乡”，在那里，波利尼西亚人和在此定居的更早的太平洋岛民之间形成了一种复杂的交流联系。

尽管大多数遗传证据支持前一种假说，争论一直存在，并且其他貌似可信的假说也被验证。值得关注的是，Richards 等[8]最近认为，来自线粒体 DNA 数据的证据实际与东印度尼西亚起源假说吻合。

最近几年，Y 染色体标记在解决人类群体进化历史方面的能力已经得到公认[9,10]。这是因为 Y 染色体非重组区的标记可以构建完整的单倍型，这样男性参与的迁徙可以很容易得到辨认。Y 染色体上大量的双等位基因标记的鉴定[12,13]使其应用效力增强。本项研究通过研究来自东南亚、中国台湾、密克罗尼西亚、美拉尼西亚和波利尼西亚 36 个群体 551 名男性个体的 Y 染色体样本，追溯了波利尼西亚人的父系来源。

4.2.2 材料和方法

研究群体的名称、样本大小和地理位置列于表 4 - 3 和图 4 - 6 中。作为对照，本项研究也包括了部分已经发表的东南亚人群的相关数据[14]。19 个双等位基因标记的细节、PCR 扩增步骤、单倍型的构建和术语都在之前的报道中提供[14]。单倍型多样性是通过 Nei[15]方法计算，基因距离通过 Nei[15]和 Reynolds 等的方法计算[16]。

表4-3 亚洲大陆和海岛人群Y染色体单倍型频率分布

群体	样本量	单倍型														
		H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11	H12	H14	H16	H17
		C	D	D1	F	K	O3	O3d	O3e	O1	O1a	O2	O2a	P	R1	M
东南亚																
土家(1)	10	10				20	30	10		20			10			
瑶(2)	20	35		15		10	5	15				20				
侗(3)	10	20					10		20	20	10	20				
彝(4)	14			14.3		42.9	21.4		7.1			14.3				
畲(5)	11	18.2				9.1	18.2	27.3	18.2			9.1				
黎(6)	11								9.1	27.3		54.5	9.1			
壮(7)	28	3.6		3.6	7.1	3.6	3.6		25	17.9		25	10.7			
北泰(8)	20		20				5		30			20	20	5		
东北泰国(9)	20				5	5	5	5		5	5	45	20	5		
索马里(10)	5	20	20						40				20			
柬埔寨(11)	26	3.8		3.8	11.5	11.5	3.8		15.4	3.8	3.8	23.1	11.5	3.8	3.8	
原住民(12)	17					23.5	5.9	5.9				64.7				
马来人(13)	27				3.7	18.5	33.3		22.2	3.7	14.8	3.7				
巴塔克(14)	18	5.6			5.6	11.1	11.1	16.7		22.2		27.8				
爪哇人(15)	11	9.1			9.1	27.3	9.1			18.2	9.1	18.2				
哥打基(16)	19	10.5				5.3	10.5			31.6	10.5	26.3		5.3		
中国台湾岛																
布农族(17)	9									11.1	66.7		22.2			
泰雅族(18)	24						29.2	4.2	4.2	54.2	8.3					
雅美族(19)	8									25		75				

（续表）

群 体	样本量	单倍型														
		H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11	H12	H14	H16	H17
		C	D	D1	F	K	O3	O3d	O3e	O1	O1a	O2	O2a	P	R1	M
排湾族(20)	11								18.2	54.5	27.3					
亚美族(21)	6									100						
美拉尼西亚																
班克斯，托雷斯(22)	6	33.3				33.3	16.7									16.7
迈沃岛(23)	10					60	20									20
桑托岛(24)	4					100										
纳肖(25)	3															100
新几内亚(26)	90	15.5			2.2	43.3										38.9
密克罗尼西亚																
特鲁克(27)	17	5.9				64.7	5.9		5.9					5.9		11.8
马朱罗(28)	9				11.1	66.7					22.2					
基里巴斯(29)	11						63.6		9.1				27.3			
关岛(30)	6	16.7	16.7			33.3	33.3									
帕劳群岛(31)	13	7.7			7.7	61.5	23.1									
丰栢(32)	10	30				70										
瑙鲁(33)	7					28.6	71.4									
波利尼西亚																
卡平阿马朗伊(34)	10	30				70										
汤加(35)	1						100									
萨摩亚人(36)	29	48.3				6.9	41.4							3.5		

注：括号内的数字编号对应图 4-6 中人群的地理位置。

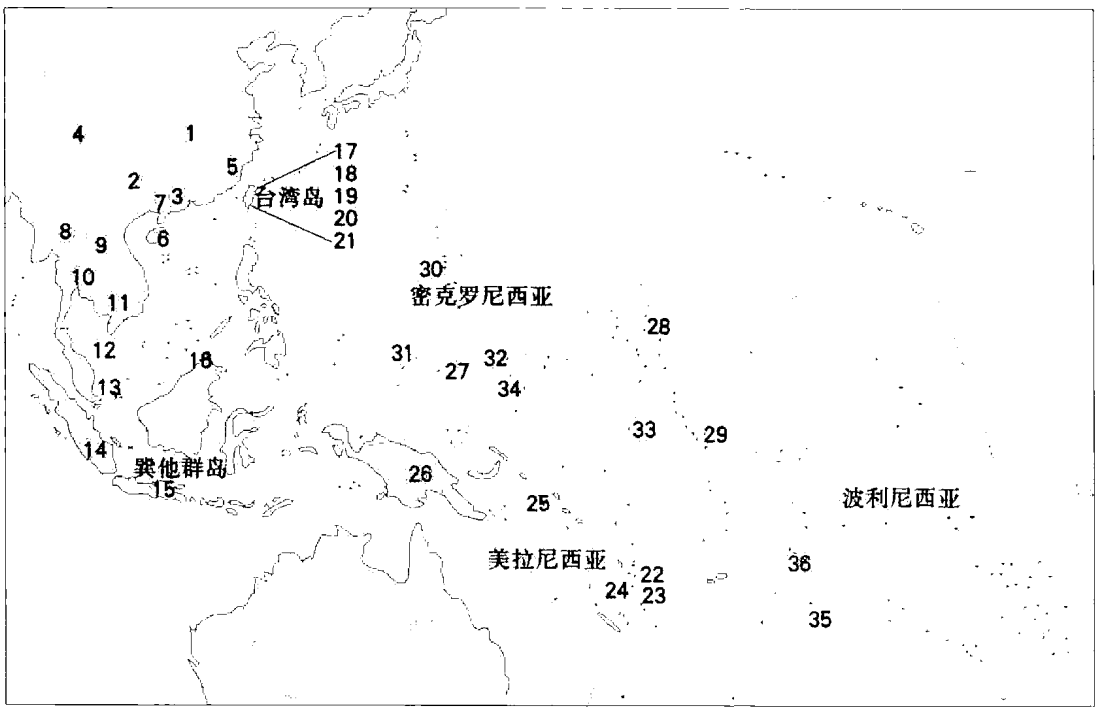


图 4－6　研究人群的地理位置

人群名称的数字编号和表 4－3 括号中的数字编号对应。波利尼西亚人群 Kapingamarangi(标记为 34)地理上位于密克罗尼西亚。

4.2.3　结果和讨论

使用 19 个双等位基因标记，在 551 个 Y 染色体样本中确定了 15 个单倍型。不同群体中的单倍型频率列于表 4－3。此前我们已经基于这 19 个双等位基因的单倍型[14]构建了简单的系统发生树，在这棵系统发生树中 H1(包括正式定名的 A/B/C 3 大单倍群，样本中仅包括 C)被认为是原始的单倍型，因为它出现在黑猩猩中。在其他单倍型中，H2(正式定名 D、E)也是相关的祖先型，因为它同时在非洲和非洲以外人群中出现；H5(正式定名 K)是所有其他非洲以外人群单倍型的共同祖先，非洲以外单倍型呈地域特异性分布。东南亚共有 14 种单倍型，多样性达 0.88，是目前为止所有被研究的群体中最多样化的群体。唯一缺失的单倍型是 H17(正式名 M)，这一单倍型特异性地存在于美拉尼西亚人群中。台湾少数民族总共 58 个男性中有 7 个单倍型(H6～H12)，多样性为 0.70。台湾的 7 种单倍型中两种(H6 和 H7)只存在于泰雅族中。在所有的 113 个密克罗尼西亚和波利尼西亚样本中，确定了 10 个单倍型(单倍型多样性为 0.72)，其中的 9 种(H1、H2、H4、H5、H6、H8、H10、H12 和 H14)也存在于东亚人群中。第 10 个单倍型 H17 只存在于两个特鲁科个体中。

将密克罗尼西亚和波利尼西亚人群中的单倍型与台湾人群中的做比较，发现差异非常明显。除了 H6，这两个人群存在两组不相干的单倍型。H1、H2、H4 和 H5 只存在于

密克罗尼西亚和波利尼西亚人群中，而不存在于台湾人群中。然而，台湾人单倍型 H7、H8、H9、H10、H11 和 H12 不存在于波利尼西亚人群中，在密克罗尼西亚人群中也相对稀少。实际上，密克罗尼西亚和波利尼西亚人群与台湾少数民族共有 4 个单倍型，其中 3 个单倍型只存在于密克罗尼西亚人群中而不存在于波利尼西亚人群中，这 3 个单倍型可能反映了从东南亚到密克罗尼西亚的近期基因传播。很明显，H1（祖先单倍型）和 H5（非洲以外单倍型共有的祖先单倍型）都不存在于台湾少数民族中，然而在密克罗尼西亚和波利尼西亚人群中频率很高。这证明台湾少数民族在一条迁徙路线上，而密克罗尼西亚和波利尼西亚人群在另一条路线上，他们分别接受了来自东南亚的不同类单倍型。遗传距离估算（表 4-4）显示，在台湾人和密克罗尼西亚/波利尼西亚人群之间的距离是这两个群体中任一个与东南亚人群之间距离的 2 倍。

表 4-4 遗传距离

群　　体	D_m	D_{ST}	F_{ST}
东南亚-中国台湾	0.108	0.268	0.098
东南亚-密克罗尼西亚/波利尼西亚	0.103	0.279	0.097
中国台湾-密克罗尼西亚/波利尼西亚	0.250	0.876	0.203

注：D_m 见参考文献[15]，D_{ST}、F_{ST} 见参考文献[16]。D_m 为根井距离，D_{ST} 为距离参数，F_{ST} 为固定参数。

“快车模式”的表述认为台湾少数民族是从中国南方沿海迁来的原始南岛语群体，出发时间在 5 000～6 000 年前。虽然中国大陆南方已经不再有人群用南岛语言，但是我们的数据不排除这样一种可能性，即中国南方沿海是现代台湾少数民族的故乡。就像上面提到的那样，台湾人 Y 染色体单倍型集合只是包括华南在内的广义东南亚人群集合中的子集。本研究中提到的中国沿海人群包括土家族、瑶族、侗族、畲族、黎族和壮族，其中一些人群受到汉族移民的影响，在过去 2 000 年内已经迁徙到中国西南部（云南、贵州、四川），他们具有与其他东南亚人群相似的 Y 染色体特性。Y 染色体单倍型分布的证据，明确显示出包括中国南方、东南亚大陆区域和东南亚海岛在内的广义东南亚地区的基因交流的连续性。这个连续体很明显地涵盖了跨越语系的各个人群，包括汉藏语、苗瑶语、南亚语和南岛语的使用者。虽有这些发现，Y 染色体数据不支持波利尼西亚人的故乡在中国台湾。不过，本研究结果不能否认“快车模式”的全部内容，这个模式主张南岛语系和拉皮塔文化来自东南亚。此外，和其他人群相比，东南亚人群的单倍型多样性程度最高，也表明东南亚是个发源地。对以上认识的最合理解释是，台湾人群和波利尼西亚人群的祖先起源于东南亚。但是，波利尼西亚的迁徙好像是通过一条与台湾路线完全不同的路线走出东南亚的。不过依据这些数据，还不能确定波利尼西亚人群祖先的中心位置。还有一种可能，Y 染色体单倍型从中国大陆传播到台湾，然后到达密克罗尼西亚/波利尼西亚；在这个模式中，因为 Y 染色体类群的随机灭绝，从而使不同地理区域的单倍型形成差异。但是，这个模式看来并不合理，因为事实上，研究中来自相同地区的多个人群有着类似的单倍

型分布。值得注意的是，两种主要的单倍型(H1和H5)在5个台湾人群中都是缺失的。

美拉尼西亚人群对波利尼西亚和密克罗尼西亚人群的贡献程度也颇有争议[7,17]。研究发现H17几乎局限于美拉尼西亚人群，虽然也在别的地方出现，但是仅仅出现在密克罗尼西亚一个人群中(在特鲁科中是12%)，这个现象非常值得注意。波利尼西亚人群中H17缺失的现象说明美拉尼西亚Y染色体单倍型向波利尼西亚扩张的程度很低或者微不足道，并不像美拉尼西亚人群核基因组或者线粒体DNA的等位基因流入波利尼西亚的程度那么大[3,18]。造成不同遗传物质的贡献差别的原因还不是很清楚，但是遗传漂变的作用强度在不同的基因区段有差异，尤其在涉及波利尼西亚人群定居的群体瓶颈作用中[19]。而且，单一性别主导的迁徙在太平洋移民的过程中很关键[17]，在其他地区的人群地理扩散中也有发现[20,21]。

就波利尼西亚的近期历史而言，Hurles等的Y染色体研究值得关注[11]。他们发现在波利尼西亚存在广泛的欧洲人群混血。然而，本研究没有发现显著的欧洲人群影响。欧洲人群特有的单倍型H14[14]只发现2例，一个密克罗尼西亚人和一个波利尼西亚人。这两项研究的结果不同可能由于所研究的人群不同。Hurles等研究的波利尼西亚样本来自库克群岛的拉罗汤加岛。必须说明，欧洲人的混血是相对晚近的历史事件，可能没有到达区域内的所有人群，不能掩盖波利尼西亚人最早的史前迁徙的基因轨迹。

线粒体DNA数据支持从中国台湾通过菲律宾和印度尼西亚"走廊"到波利尼西亚的人群扩张。线粒体DNA基因组调控区的核苷酸变异组合成的一种序列被称为"波利尼西亚序列"[5,6]，它在菲律宾-印尼走廊中频率很高，在波利尼西亚频率最高。这个序列的亲缘类型在这个地区的分布频率也很高，在台湾人群中达到最高的多样性。在这些观察的基础上，波利尼西亚序列的来源被追溯到中国台湾[6]，这似乎为"快车模式"假说提供了强有力的遗传依据。然而，最近对已经发表的线粒体DNA数据[3,5]的一项重新分析[8]质疑了这一推断。Richards等[8]认为，基于这一序列的分化时间的估算和相关群体的年龄估计，线粒体DNA数据不能支持波利尼西亚人来源于中国台湾。相反，这个证据与东南亚岛屿起源说相吻合，其故乡应该在印度尼西亚东部。尽管我们的发现与这个观点更加一致，但是Y染色体数据不能明确地指出波利尼西亚人群的"核心起源地"，而东南亚岛屿更可能是通往波利尼西亚线路的中途站。另外，就波利尼西亚序列而言，研究发现有一种突变型(CAT变体)广泛分布于包括中国南部的东南亚人群中。所以，一方面波利尼西亚序列并没有严格地与中国台湾和东南亚岛屿人群关联，再加上本项研究展现的Y染色体单倍型的分布格局，太平洋移民的祖先应该来自广义的东南亚人群分布区域。

参考文献

[1] Diamond J M. Express train to Polynesia. Nature (London), 1988, 336: 307-308.

[2] Bellwood P. Man's conquest of the Pacific: the prehistory of Southeast Asia and Oceania. New York: Oxford Univ Press, 1978.

[3] Sykes B C, Leiboff A, Low-Beer J, et al. The origins of the Polynesians: an interpretation from mitochondrial lineage analysis. Am J Hum Genet, 1995, 57: 1463 - 1475.

[4] Melton T, Peterson R, Redd A J, et al. Polynesian genetic affinities with Southeast Asian populations as identified by mtDNA analysis. Am J Hum Genet, 1995, 57: 403 - 414.

[5] Redd A J, Takezaki N, Sherry S T, et al. Evolutionary history of the COII/tRNALys intergenic 9 base pair deletion in human mitochondrial DNAs from the Pacific. Mol Biol Evol, 1995, 12: 604 - 615.

[6] Melton T, Clifford S, Martinson J, et al. Genetic evidence for the proto-Austronesian homeland in Asia: mtDNA and nuclear DNA variation in Taiwanese aboriginal tribes. Am J Hum Genet, 1998, 63: 1807 - 1823.

[7] Terrell J. History as a family tree, history as an entangled bank: constructing images and interpretations of prehistory in the South Pacific. Antiquity, 1988, 62: 642 - 657.

[8] Richards M, Oppenheimer S, Sykes B. mtDNA suggests Polynesian origins in Eastern Indonesia. Am J Hum Genet, 1998, 63: 1234 - 1236.

[9] Jobling M A, Tyler-Smith C. Fathers and sons: the Y chromosome and human evolution. Trends Genet, 1995, 11: 449 - 455.

[10] Hammer M F. A recent common ancestry for human Y chromosomes. Nature (London), 1995, 378: 376 - 378.

[11] Hurles M E, Irven C, Nicholson J, et al. European Y chromosomal lineages in Polynesians: a contrast to the population structure revealed by mtDNA. Am J Hum Genet, 1998, 63: 1793 - 1806.

[12] Underhill P A, Jin L, Zemans R, et al. A pre-Columbian Y chromosome-specific transition and its implications for human evolutionary history. Proc Natl Acad Sci USA, 1996, 93: 196 - 200.

[13] Underhill P A, Jin L, Linn A A, et al. Polynesian origins: insights from the Y chromosome. Genome Res, 1997, 7: 996 - 1005.

[14] Su B, Xiao J, Underhill P, et al. Y chromosome evidence for a northward migration of modern humans into Eastern Asia during the Last Ice Age. Am J Hum Genet, 1999, 65: 1718 - 1724.

[15] Nei M. Estimation of average heterozygosity and genetic distance from a small number of individuals. Genetics, 1978, 89: 583 - 590.

[16] Reynolds J, Weir B S, Cockerham C C. Estimation of the coancestry coefficient: basis for a short-term genetic distance. Genetics, 1983, 105: 767 - 779.

[17] Lum J K, Cann R L, Martinson J J, et al. Mitochondrial and nuclear genetic relationships among Pacific Island and Asian populations. Am J Hum Genet, 1998, 63: 613 - 624.

[18] Roberts-Thomson J M, Martinson J J, Norwich J T, et al. An ancient common origin of aboriginal Australians and New Guinea highlanders is supported by alpha-globin haplotype analysis. Am J Hum Genet, 1996, 58: 1017 - 1024.

[19] Flint J, Boyce A J, Martinson J J, et al. Population bottlenecks in Polynesia revealed by minisatellites. Hum Genet, 1989, 83: 257 - 263.

[20] Seielstad M T, Minch E, Cavalli-Sforza L L. Genetic evidence for a higher female migration rate

in humans. Nat Genet, 1998, 20: 278 - 280.

[21] Pe'rez-Lezaun A, Calafell F, Comas D, et al. Sex-specific migration patterns in Central Asian populations, revealed by analysis of Y chromosome short tandem repeats and mtDNA. Am J Hum Genet, 1999, 65: 208 - 219.

4.3　南岛西群和侗傣人群的紧密父系遗传关系

4.3.1　研究背景

南岛语系是世界上最重要的语系之一，分布在东南亚岛屿、太平洋以及印度洋的大部分区域，包含世界五分之一以上的语言[1]。这一语系最初是由Murdock[2]基于两组人群语言的相似性建立的一个单独的系统，两组人群是：马来-波利尼西亚人（包括东南亚岛屿人群、马达加斯加、密克罗尼西亚和波利尼西亚人）以及中国台湾少数民族[3,4]。之后Benedict发现东亚的另一语系，侗傣语系与南岛语系有许多相似点，因此将两者合称为澳台语系[5]。侗傣语系是位于东南亚岛屿类群以北的一个语系，主要位于中国南部。一些侗傣族群散布到了老挝、泰国甚至远到印度[1]。台湾少数民族、马来-波利尼西亚人以及侗傣语系人群存在明显相似之处，这被民族学家[6-10]和语言学家[11-15]所证实。相关研究将台湾少数民族和马来-波利尼西亚人的祖先追溯到中国东南沿海人群，主要就是侗傣语系人群以及他们的祖先——百越先民。

南岛语系的发源在语言学以及其他相关领域一直是个有争议性的话题。“快车模式”是一个被广为接受的关于南岛语系发源的语言学理论[3,4,16,17]。该模式提出，5 000～6 000年前，原始南岛语人群从台湾发源，通过菲律宾以及印度尼西亚东部向南扩散。他们向东到达密克罗尼西亚和波利尼西亚，向西到达印度尼西亚西部和马达加斯加。“快车”是比喻从印度尼西亚东部扩散至现今南岛语系分布范围的速度十分迅速。关于台湾是所有南岛语系群体起源的假说（台湾起源论）主要是由于观察到台湾少数民族语言的多样性要比马来-波利尼西亚人群的高得多[3,4]。但是，一些语言学家找到的证据反驳台湾起源论，并且猜测加里曼丹岛或苏拉威西岛才是南岛语系群体的发源地[15,18,19]。台湾起源论也进一步受到民族学家[6-9]、考古学家[10]和遗传学家的质疑[20-25]。

遗传方面的证据同样存在争议，一些线粒体DNA研究显示波利尼西亚人群是从台湾起源[20-22]。一项最近关于台湾少数民族线粒体DNA的研究发现，“波利尼西亚序列”在台湾存在原始型，说明台湾起源论或许在母系遗传上是成立的[26]。另一方面，这项理论受到父系遗传证据的挑战，Y染色体研究显示波利尼西亚人群与台湾少数民族之间缺少相似性，波利尼西亚人群的Y染色体大部分来自美拉尼西亚的棕色人种的C单倍群[23]。有一些线粒体DNA的研究也对其提出质疑，认为波利尼西亚人群其实起源于印度尼西亚[24,25]。这些遗传学证据的冲突可以归因于两个关键区域的人群或证据的缺失：一是3个南岛语系群体（台湾少数民族、马来人群、波利尼西亚人）起源的东南亚沿海人群；二是作为波利尼西亚群体发源地的马来语群体，包括印度尼西亚人群。

南岛语系人群遗传结构的另一个重要特点是南岛语系东部人群（波利尼西亚和密克

罗尼西亚)与西部人群(马来人群和台湾少数民族,如图 4-7)呈现明显不同的遗传类型,两者以华莱士线为界。常染色体 STR 突变研究[27]显示,在波利尼西亚人群与南岛语系西部人群间有一个明显的遗传分化。这些研究表明,波利尼西亚人群可能已经经历了自然选择或与美拉尼西亚人群融合。这些过程改变了他们的遗传结构[16,20,28]。也有可能波利尼西亚人群在扩散过程中经历了遗传漂变或奠基者效应。南岛语系西部人群的遗传结构,尤其是马来群岛一带,对于南岛语系的起源更为关键(图 4-7)。印度尼西亚(巴厘岛和松巴岛)人群 Y 染色体的高度多样性暗示了该区域的群体可能从旧石器时代以来就存在了[29,30]。由于这些高度的遗传多样性,说明马来人群尤其是印度尼西亚人群并不是近期从台湾起源的。

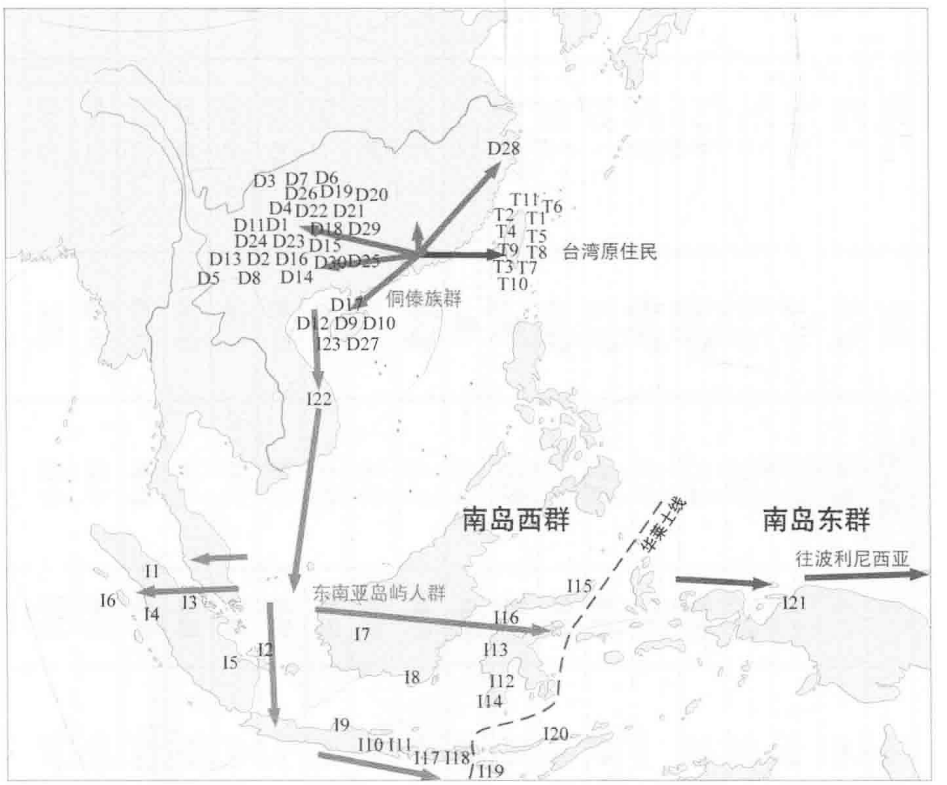


图 4-7 人群标本地理分布以及迁徙路线图

人群样本编码见表 4-5。绿色箭头代表侗傣群体的扩张;蓝色箭头代表台湾;橙色箭头代表马来群体。波利尼西亚人群的起源用紫色箭头表示,目前在父系角度仍存在争议。

本节通过研究所有相关群体 Y 染色体的多样性来检验关于马来人群的台湾起源论。相关人群包括侗傣人群、印度尼西亚的马来语人群以及台湾少数民族。研究发现马来群体以及台湾少数民族的父系遗传均来源于侗傣人群,但是两者的起源相对独立。所以,从父系结构来讲台湾并非马来人群的来源。

表4-5 群体样本的分类和分布信息

编号	部 族	ISO639-3	语系	语 族	语支	人 口	国 家	省 份	县 区
D1	侏人	ply	南亚	孟高棉	弗琉	10 000	中国	广西	隆林
D2	夜郎	yrn	侗傣	仡央	布-郎	400	中国	广西	那坡
D3	羿人	gio	侗傣	仡央	仡-基	3 000	中国	贵州	毕节
D4	仡佬-青	giq	侗傣	仡央	仡-基	1 700	中国	广西	隆林
D5	拉基	lbt	侗傣	仡央	仡-基	9 016	中国	云南	马关
D6	木佬		侗傣	仡央	仡-基	30 000	中国	贵州	麻江
D7	仡佬-红	gir	侗傣	仡央	仡-基	1 500	中国	贵州	大方
D8	仡佬-白	giw	侗傣	仡央	仡-基	1 200	中国	云南	麻栗坡
D9	黎族-杞	lic	侗傣	仡央	黎	747 000	中国	海南	通什
D10	加茂	jio	侗傣	仡央	黎	52 300	中国	海南	保亭
D11	布央	byu	侗傣	仡央	央-标	3 000	中国	云南	广南
D12	仡隆	cuq	侗傣	仡央	央-标	70 000	中国	海南	东方
D13	普标	laq	侗傣	仡央	央-标	307	中国	云南	麻栗坡
D14	高栏	mlc	侗傣	壮侗	贝-傣	114 000	中国	广西	防城
D15	僮-北	ccx	侗傣	壮侗	贝-傣	10 000 000	中国	广西	武鸣
D16	僮-南	ccy	侗傣	壮侗	贝-傣	4 000 000	中国	广西	崇左
D17	临高	onb	侗傣	壮侗	贝-傣	520 000	中国	海南	临高
D18	五色	eee	侗傣	壮侗	贝-傣	30 000	中国	广西	融水
D19	锦家	aih	侗傣	壮侗	侗水	2 300	中国	贵州	荔波
D20	侗族	doc	侗傣	壮侗	侗水	907 560	中国	广西	三江
D21	水族	swi	侗傣	壮侗	侗水	345 993	中国	广西	融水

（续表）

编号	部 族	ISO639－3	语系	语 族	语支	人 口	国 家	省 份	县 区
D22	莫家	mkg	侗傣	壮侗	侗水	10 000	中国	贵州	荔波
D23	仫佬	mlm	侗傣	壮侗	侗水	159 328	中国	广西	罗城
D24	毛南	mmd	侗傣	壮侗	侗水	37 000	中国	广西	环江
D25	标	byk	侗傣	壮侗	侗水	20 000	中国	广东	怀集
D26	佯僙	tct	侗傣	壮侗	侗水	20 000	中国	贵州	平塘
D27	畲家		侗傣	未定		1 000 000	中国	海南	陵水
D28	傜傣		侗傣	未定		500 000	中国	上海	闵行
D29	草苗	cov	侗傣	壮侗	侗水	63 632	中国	广西	融水
D30	拉伽	lbc	侗傣	壮侗	侗水	12 000	中国	广西	金秀
T1	阿美	ami	南岛	台湾	排湾	130 000	中国	台湾	花莲
T2	巴则海	uun	南岛	台湾	排湾	300	中国	台湾	卓兰
T3	西拉雅-玛卡道	fos	南岛	台湾	排湾	10 000	中国	台湾	花莲
T4	邵	ssf	南岛	台湾	排湾	248	中国	台湾	南投
T5	排湾	pwn	南岛	台湾	排湾	53 000	中国	台湾	台东
T6	泰雅	tay	南岛	台湾	泰雅	63 000	中国	台湾	宜兰
T7	鲁凯	dru	南岛	台湾	排湾	8 007	中国	台湾	屏东
T8	卑南	pyu	南岛	台湾	排湾	8 132	中国	台湾	台东
T9	邹	tsu	南岛	台湾	邹	5 797	中国	台湾	嘉义
T10	布农	bnn	南岛	台湾	排湾	34 000	中国	台湾	花莲
T11	赛夏	xsy	南岛	台湾	排湾	4 194	中国	台湾	宜兰
I1	巴答	bbc	南岛	马来-波利	西部	5 800 000	印尼	北苏门	
I2	邦加	mly	南岛	马来-波利	西部	500 000	印尼	南苏门	邦加

（续表）

编号	部 族	ISO639-3	语系	语 族	语支	人 口	国 家	省 份	县 区
I3	马来	mly	南岛	马来-波利	西部	2 000 000	印尼	黎牙	
I4	民昂卡宝	min	南岛	马来-波利	西部	4 000 000	印尼	西苏门	
I5	帕林邦	plm	南岛	马来-波利	西部	1 100 000	印尼	南苏门	
I6	尼亚斯	nia	南岛	马来-波利	西部	600 000	印尼	北苏门	尼亚斯
I7	达雅克	dyk	南岛	马来-波利	西部	2 100 000	印尼	中加里	
I8	班加	bjn	南岛	马来-波利	西部	3 000 000	印尼	南加里	
I9	爪哇	jav	南岛	马来-波利	西部	75 500 000	印尼	中爪哇	
I10	登加	tes	南岛	马来-波利	西部	500 000	印尼	东爪哇	
I11	巴厘	ban	南岛	马来-波利	西部	3 800 000	印尼	巴厘	
I12	布其斯	bug	南岛	马来-波利	西部	3 500 000	印尼	南苏拉	
I13	托拉加	sda	南岛	马来-波利	西部	500 000	印尼	南苏拉	
I14	马卡萨	mak	南岛	马来-波利	西部	1 600 000	印尼	南苏拉	
I15	民那哈撒	tom	南岛	马来-波利	西部	200 000	印尼	北苏拉	
I16	凯利	lew	南岛	马来-波利	西部	471 000	印尼	中苏拉	
I17	沙萨	sas	南岛	马来-波利	西部	2 100 000	印尼	西努沙	龙博
I18	松巴哇	smw	南岛	马来-波利	西部	400 000	印尼	西努沙	松巴哇
I19	松巴	xbr	南岛	马来-波利	中部	234 574	印尼	东努沙	松巴
I20	阿洛尔	aol	南岛	马来-波利	中部	25 000	印尼	东努沙	阿洛尔
I21	伊里安		吉文湾			20 806	印尼	伊里安加亚	
I22	占	cjm	南岛	马来-波利	西部	99 000	越南	平定	
I23	回辉	huq	南岛	马来-波利	西部	4 500	中国	海南	三亚

注：详细信息可以 ISO639-3 编号在 www.ethnologue.com 网上查询。

4.3.2 研究方法

1. 样本

本研究使用的血液样本是从中国南部30个侗傣族群中，通过FTA卡(Whatman® Inc)采集而得，这些样本几乎包括了中国境内所有的侗傣群体。采集的11个台湾少数民族群体样本包括台湾平埔族与高山族。23个马来-波利尼西亚群体中，21个来自印度尼西亚，1个来自越南平定，1个来自中国海南。每个群体的样本量显示在表4-6中，本研究中共使用了来自1 509个个体的样本，互相没有亲缘关系，每个个体都签署了知情同意书。每一群体的个体样本来自其分布区的不同地点，这增加了样本的多样性。从文献中获取东亚以及东南亚70个其他群体的参考数据(包括侗傣语系[23]、马来-波利尼西亚语族[23]、台湾少数民族[23]、藏缅语族[31-33]、汉族[31,34]，以及阿尔泰语系[31])，参考样本总计1 348个。在主成分分析中，包括新研究的以及之前报道过的样本，总共计134个群体。

虽然某些群体样本数量相对较少，我们并不认为需要扩大他们的样本量，因为他们本来就是从Y染色体多样性较低的小群体，例如侗水语支的锦家人以及吉文湾伊里安人群中收集的样本。Y染色体的有效群体大小通常不到常染色体的1/4。因此，Y染色体多样性研究只需要比常染色体遗传标记研究更小的群体样本。所以一般情况大约几十万的群体需要30个左右的样本就足够了。如果群体更小，需要的样本数量更少。这里对大部分群体都恒定采样约30个，对于小群体则采15个左右。

2. 遗传标记

选用20个双等位基因Y染色体标记(SNP)如下：YAP、M15、M130、M89、M9、M5、M122、M134、M7、M117、M121、M111、M17、M175、M119、M110、M95、M88、M45和M120，基于PCR限制性片段长度多态性方法分类所得[31]。这些标记在东亚人群中大都十分显著，并且按照YCC系统定义为19个单倍群[35]。

Y染色体上7个微卫星标记(STR)如下：DYS19、DYS388、DYS389-1、DYS390、DYS391、DYS392和DYS393，通过荧光标记引物分型[36]。

3. 数据分析

利用SPSS11.0软件对人群Y染色体单倍型频率数据进行主成分分析，观察群体间遗传关系。一些SNP位点，如M175与M117，并没有归入先前报道过的群体。因此，在分析时将O*-M175数据归入单倍群K的类型，将O3a5a-M117归入O3a5*的类群。单倍群与主成分间的关联分析也使用SPSS11.0软件进行。

使用Admix2.0软件[37]进行混合分析，评估中国汉族对于侗傣族群的遗传影响。我们猜测潜在的混合因素是2 500年前秦朝军队进入岭南侗傣区域所导致的。印度尼西亚人群的混合比例通过软件也计算得出，混合历史大约发生在5 000年前。

侗傣人群、台湾少数民族以及马来-波利尼西亚群体间的遗传距离通过R_{ST}以及线性R_{ST}[38]计算方法，使用Arlequin软件[39]估计得出。3个群体的多样性通过平均基因多样性、单倍群多样性[40]以及STR等位长度的方差[41]来进行估算。

用Network4.1软件(Fluxus Technology Ltd)为O1a*下STR单倍群绘制由中点

连接法构建的网络结构图。从网络图中能估算出 O1a * 的年龄，用于时间估算的突变率约为 1.932×10^{-4}/年[42]，在网络图中使用了所有 STR 位点突变率的总和(假设每一世代大约 25 年)。

4.3.3　结果和讨论

为了确定侗傣人群和南岛语系西部人群间的遗传关系，在 Y 染色体非重组区内选了 20 个 SNP 和 7 个 STR 位点进行研究，从 30 个侗傣族群、23 个马来群体以及 11 个台湾少数民族群体中选取了 1 509 份样本(图 4-7 中人群分布以及表 4-5 中人群信息)。这次研究几乎取样了中国所有的侗傣群体以及所有的台湾少数民族群体。

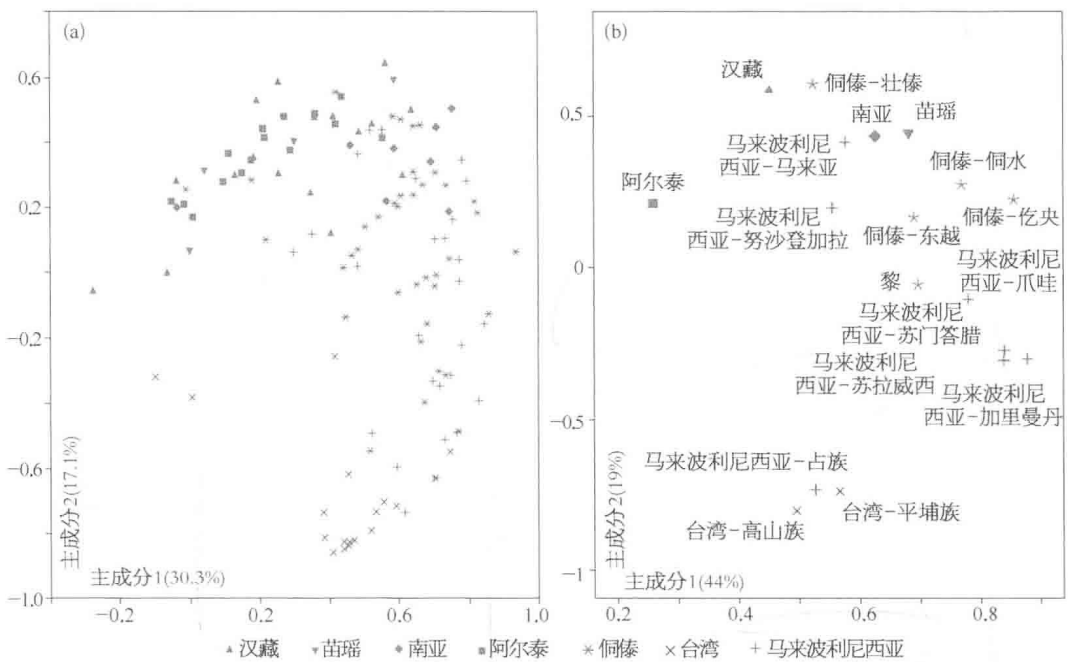


图 4-8　主成分分析图

(a) 所有样本群体。DC(绿星)与 MP(紫色十字)和 TA(蓝叉)最接近。其他所有群体，包括 ST、HM、AA、AT(红色三角形、正方形和菱形)，都与 MP 和 TA 相隔较远，表明 DC 是唯一与 MP 和 TA 有关的群体。(b) 合并样本。ST、HM、AA、AT 样本根据语系混合。DC 样本根据不同语族混合。MP 和 TA 样本根据不同地理位置混合。

此外，针对包含东亚及东南亚所有语系的 134 个群体，用 SNP 单倍型频率数据进行主成分分析。结果显示，侗傣人群比其他东亚和东南亚人群离南岛语系西部人群都更近(图 4-8)，证明侗傣人群与南岛西部群体间强烈的遗传关系。侗傣-马来-台湾群体簇与其他族群间的分离主要是由于第二主成分，而非第一主成分；单倍型 O1a * 与第二主成分显示最强相关性($r^2 = -0.875$，$P < 10^{-4}$)。另外，O1a-M119 是台湾少数民族的主要单倍群类型，频率从 54%～100%不等，平均为 77%(表 4-6，O1a * 与 O1a2 之和)。这一单倍群在侗傣语系人群与马来群体中也呈现较高的频率，侗傣约 20.5%、马来约 21.2%，

表 4-6 新研究群体的 Y-SNP 单倍群频率

群 体	样本	单倍群频率(%)															
		C	D*	D1	F	M	K	O*	O1a*	O1a2	O2a*	O2a1	O3*	O3a1	O3a4	O3a5	O3a5a	P
侠人	30		3.3				3.3	10.0	10.0	3.3	23.3		30.0			6.7	10.0	
夜郎	16										62.5	6.3	18.8			12.5		
羿人	13		15.4				7.7	23.1			15.4		30.8				7.7	
仡佬-青	30						3.3	13.3	60.0		16.7		3.3			3.3		
拉基	30	3.3		3.3	13.3		13.3	16.7	6.7		10.0		3.3			6.7	23.3	
木佬	30	10.0					3.3	13.3	3.3	3.3	63.3		3.3					
仡佬-红	31	3.2					6.5	22.6	22.6		16.1		12.9				16.1	
仡佬-白	14							35.7	14.3		42.9					7.1		
黎族-杞	34							35.3	32.4		29.4						2.9	
加茂	27							25.9	51.9		22.2							
布央	32		3.1		6.3		6.3	9.4	3.1		71.9							
仡隆	31	3.2					6.5	9.7	38.7				38.7				3.2	
普标	25							32.0	4.0		60.0					4.0		
高栏	30	10.0					10.0	53.3	3.3		20.0					3.3		
僮-北	22							13.6		4.6	72.7			4.6			4.6	
僮-南	15							13.3	20.0		60.0	6.7						
临高	30						3.3	16.7	26.7		13.3		3.3			10.0	26.7	
五色	31	3.2			3.2		9.7	16.1	6.5		54.8		3.2			3.2		
拉伽	23	4.4	52.2				4.4				8.7		26.1		4.4			
侗族	38	21.1					5.3	10.5			39.5		10.5			2.6	10.5	
水族	50				8.0		10.0		18.0		44.0					20.0		

（续表）

群 体	样本	单倍群频率(%)																
		C	D*	D1	F	M	K	O*	O1a*	O1a2	O2a*	O2a1	O3*	O3a1	O3a4	O3a5	O3a5a	P
莫家和锦家	40						2.5				87.5		5.0			2.5		2.5
仫佬	40	2.5		12.5	7.5		5.0		5.0	25.0	30.0		7.5			5.0		
毛南	32	9.4			9.4		15.6				56.3		9.4					
标	34	2.9						5.9	14.7		17.7		52.9					5.9
佯僙	30		3.3					3.3	33.3		50.0						6.7	3.3
蚩家	40	20.0	5.0		2.5		7.5	17.5	7.5	5.0	17.5					2.5	15.0	
傜傣-南	74	2.1		6.3					39.6	12.5	8.3		4.2			27.1		
傜傣-北	51	5.9	2.0				2.0	31.4	29.4		2.0		2.0			11.8	13.7	
草苗	33						8.2		10.0		3.0		66.7			12.1		
阿美	28							7.1	42.8	17.8	7.1		21.4			3.6		
巴则海	21						14.3		38.1	19.1	14.3		14.3					
玛卡道	37	2.7					2.7	5.4	70.3		5.4					13.5		
邵	22						4.6		81.8	4.6			9.1					
排湾	22								63.6	27.3						9.1		
泰雅	22								95.5				4.5					
鲁凯	11								81.8	18.2								
卑南	11								72.7	9.1			9.1					9.1
邹	18								88.9	5.6			5.6					
布农	17						5.9		17.6	58.8		17.6						
赛夏	11								45.5	9.1	9.1	9.1	27.3					
巴答	13						11.6	19.3	23.1		15.4		23.1					7.7

（续表）

群　体	样本	单倍群频率(%)																
		C	D*	D1	F	M	K	O*	O1a*	O1a2	O2a*	O2a1	O3*	O3a1	O3a4	O3a5	O3a5a	P
邦加	13	7.7					7.7		30.8		23.1		23.1		7.7			
马来	13				7.7		7.7	7.7	38.5		7.7		23.1					7.7
民昂卡宝	15				6.7		20.0	20.0			13.3		20.0					20.0
帕林邦	11	9.1							63.6		18.2		9.1					
尼亚斯	12											8.3	91.7					
达雅克	15				6.7		26.7		20.0	20.0	6.7	6.7	13.3					
班加	15	13.3			6.7				26.7		26.7		26.7					
爪哇	15						26.7	26.7	20.0		13.3		13.3					
登加	12	16.7					8.3		33.3		33.3				8.3			
巴厘	14						28.6	14.3	7.1		28.6		14.3		7.1			
布其斯	15				13.3		20.0		33.3				26.7				6.7	
托拉加	15				13.3		13.3	13.3	13.3	6.7	33.3				6.7			
民那哈撒	14					7.1	50.0		21.4		7.1		14.3					
马卡萨	13	23.1							30.8	15.4	7.7		23.1					
凯利	15	6.7					33.3		20.0		6.7		26.7					6.7
沙萨	15	13.3					13.3	26.7	6.7		20.0		20.0					
松巴哇	18						16.7						83.3					
松巴	14				14.3		78.6						7.1					
阿洛尔	13	38.5					30.7						23.1					7.7
伊里安	11	45.5				36.4	18.2											
占	11							9.1	90.9									
回辉	31	12.9						16.1	58.1		3.2					6.5	3.2	

但在其他东亚人群中较少(<5%)[23,31-34]。因此我们期待从O1a-M119上获得更多信息来分析侗傣族群与西部南岛语系间的关系。

图4-8的主成分分析显示一部分侗傣群体与汉藏群体关系密切,这或许是因为侗傣群体与汉藏群体有着共同的祖先,导致了他们遗传的相似性。但是,这一结果的另一种解释是东亚大陆上的侗傣群体受到汉族遗传结构的影响,因为从大约2 500年前开始两者就比邻共存。混合分析可以估算现今侗傣人群中可能的侗傣祖先或汉族祖先的比例,并且一些与汉族相互隔离的侗傣群体能够被用作这项分析的亲本群体。海南岛(黎族、加茂人和仡隆人)与台湾岛的原住民相对孤立,因为他们的文化很少受到大陆外来文化的影响。因此,这些岛屿原住民的遗传结构最接近侗傣祖先群体[43]。

为了估算汉族群体对大陆侗傣族群的遗传影响,利用大陆侗傣人群、海南原住民、台湾少数民族以及汉族群体[34]的Y-SNP数据进行混合分析。设定后3个群体作为大陆侗傣族群的亲本群体,结果显示,海南原住民的遗传贡献非常高(2.145±0.927),而汉族(−0.314±0.422)和台湾少数民族(−0.831±0.662)的贡献则几乎检测不到。这里用Admix软件估算的遗传贡献显示负值表明汉族群体对现今侗傣人群几乎没有影响。这一结果说明侗傣族群的父系遗传结构相对没有受到干扰,侗傣与西部南岛群体间的遗传关系也几乎没有受到混合人群的影响。

马来群体也可能是混合群体。本研究假设马来群体由3种可能的亲本群体混合成:侗傣族群、台湾少数民族和巽他群岛的原住民,就是与巴布亚人和美拉尼西亚人相类似的群体。对印度尼西亚人群进行混合分析,将文献中[44,45]巴布亚人的相关数据作为亲本群体的结构之一。分析显示如下的混合比例:侗傣(0.713±0.124)、台湾(0.143±0.125)、巴布亚(0.144±0.050),说明印度尼西亚人群来自侗傣祖先的贡献比率是最主要的。正如我们假设的,这些数据还有一些不确定性,无法检验马来人群是否是混合群体。

由于O1a*是侗傣与南岛西部族群中最独特的单倍群,在O1a*个体中通过7个STR位点来估算侗傣族群、印度尼西亚人群和台湾少数民族两两之间的遗传分化距离(表4-7)。研究发现台湾少数民族与印度尼西亚人群间的分化最大,大约是侗傣群体与台湾群体间差异的3倍。侗傣群体与印度尼西亚群体间的分化差异和侗傣群体与台湾群体间分化差异大致相当。这些发现表明印度尼西亚人群和台湾少数民族分别与侗傣族群的遗传关系比起他们相互之间的要更近。另外,基于O1a*所携带的7个STR位点分析得出的结果,侗傣族群STR多样性比印度尼西亚以及台湾少数民族的高(表4-7)。虽然多样性最高的群体不一定总是历史最悠久的,也可能是其他相邻群体混合的结果。但是,侗傣群体中O1a*单倍群的高频率应该是源自该群体最悠久的历史,因为这一单倍群在相邻群体中几乎没有分布,不可能存在增加多样性的混合。结合多样性以及遗传分化的结论,侗傣族群很可能是印度尼西亚人群以及台湾少数民族的父系来源祖先。另一些Y染色体单倍群(如O3-M122、O2a-M95)显示与O1a*相似的格局,证明侗傣群

体从遗传角度离印度尼西亚人群和台湾人群要比后两者相互之间更近(表 4-7)。有趣的是,最近一篇研究显示 O2a 或许能够进一步追溯到南亚语系人群[46]。

表 4-7 单倍型 O1a、O2a、O3 的 Y-STR 多样性分析

组间多样性(遗传距离)

	R_{ST}			线性 R_{ST}		
	O1a	O2a	O3	O1a	O2a	O3
侗傣-台湾	0.109 ($P<10^{-5}$)	0.012 ($P=0.271$)	0.019 ($P=0.187$)	0.122	0.012	0.019
侗傣-马来	0.108 ($P<10^{-5}$)	0.093 ($P<10^{-5}$)	0.049 ($P=0.001$)	0.121	0.102	0.052
台湾-马来	0.269 ($P<10^{-5}$)	0.318 ($P<10^{-5}$)	0.285 ($P<10^{-5}$)	0.368	0.466	0.398

组内多样性

	样本量			平均基因多样性			平均方差		
	O1a	O2a	O3	O1a	O2a	O3	O1a	O2a	O3
侗傣	140	292	145	0.601	0.518	0.658	0.938	1.041	1.494
马来	75	38	64	0.547	0.397	0.498	0.897	0.320	0.634
台湾	147	12	14	0.503	0.543	0.621	0.656	0.685	1.220

在这 3 类单倍群样本中,台湾少数民族与马来群体 Y-STR 之间的遗传距离总是最大的,并且是他们分别与侗傣族群相比距离的 2 倍还多。两类统计 R_{ST} 与线性 R_{ST} 出自 Slatkin 等的研究[38]。组间 Y-STR 的多样性总是侗傣族群最大。平均基因多样性出自 Nei 的研究[40],方差分析来自 Fimmel 等的研究[41]。

用 7 个 STR 位点通过中点连接法构建了 3 个族群之间 O1a* 的网络结构(图 4-9)。如果马来人群的"台湾起源论"是正确的,也就是说,马来人群主要来自台湾少数民族,那么网络图上马来群体谱系与台湾少数民族谱系间会出现共同连接。在图 4-9 中,侗傣谱系构成网络图的中心。所有的马来群体谱系和台湾少数民族谱系都是直接或间接地与侗傣谱系共享或连接。相反,没有一支台湾谱系与马来谱系直接相连。这一结果暗示马来群体并非直接起源于台湾少数民族群体,两者分别独立起源于侗傣族群。

研究中还观察了网络图中与马来群体谱系相连的侗傣谱系。有趣的是,大部分与马来群体相连的侗傣群体的单倍型都来自海南或广西,最远到达海南的西北部。这些海南和广西群体位于北部湾附近。尤其是占族,一个位于越南南部的马来-波利尼西亚语族群体,以及海南回辉人——占族的一个分群[11,47],在网络图中同时与侗傣人群以及马来人群相连。因此,猜测马来人群可能起源于北部湾附近的某个地方,向南迁徙通过中南半岛到达马来半岛,随后才扩张至太平洋及印度洋的大部分岛屿。

通过网络图估计 O1a* 单倍群存在的时间,总年龄为 33 765 年±5 221 年,对应到冰河时代末期。所有侗傣群体样本的年龄为 33 193 年±5 577 年,与 O1a* 相近。台湾群

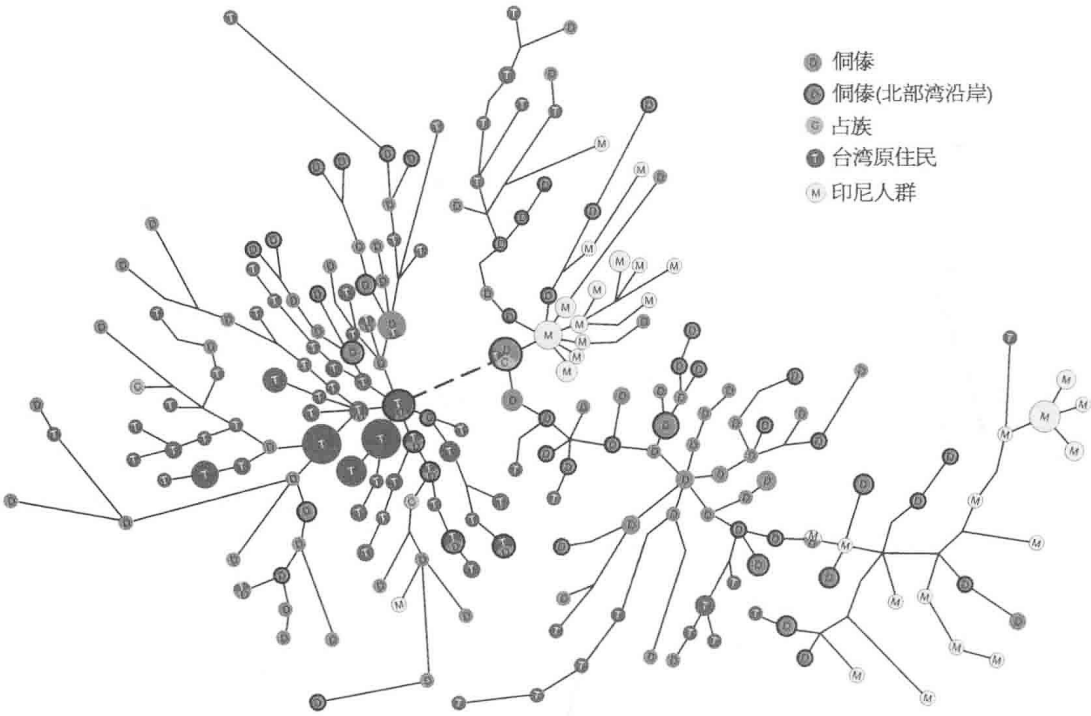


图 4-9　Y 染色体单倍群 O1a* 的网络结构图

由于原来的网络图过于复杂难以显示，这里我们尽可能地从网络结构上选取了最短树(由 Network 软件新近开发功能获取)。每个节点都代表一个 O1a* STR 单倍型。线条的长度与突变步骤成比例。虚线代表只有一步。节点的大小与所含样本的频率成正比。可以看到，几乎没有马来单倍群与台湾少数民族单倍群直接相连；他们都是直接或间接地与侗傣单倍群共享或连接，侗傣群体占据了网络图的中心(绿色的大节点)。

体的真实年龄很难估算，因为他们很大程度上和侗傣群体的单倍型重叠。这些重叠也暗示了大陆侗傣群体与台湾群体间多次交流迁徙。估计网络图左侧部分的台湾群体大致年龄为 14 659 年±3 110 年，所有台湾样本年龄为 21 268 年±3 148 年。有趣的是，后者的年龄与台湾发现的最早人类遗骸——左镇人[48]的时间非常接近。因此，研究认为 O1a* 个体大约是在旧石器时代从大陆向台湾迁徙的。

在网络图上能够看到两个比较特别的马来单倍群簇，笔者对两者都进行了时间估算。左侧群体距今 9 895 年±2 393 年，右侧距今 25 880 年±7 137 年。语言学估计马来-波利尼西亚人群的起源比本研究估计的要晚，在 5 000～6 000 年前[16]。此外，网络图显示侗傣谱系与马来谱系间几乎没有重叠，说明马来群体在离开侗傣祖先群体向外迁徙的过程中很可能遭遇了瓶颈效应从而形成两类群体。从地理上看，瓶颈效应很可能发生在越南狭窄的沿海地区。因此，单倍群 O1a* 很可能是在距今 7 500 年以前，即马来-波利尼西亚语族发源的时候进入马来人群的。不过，近期迁徙到马来群体的 O1a 个体也不能被忽视，因为遗传时间的估计并没有精确到能够排除这种可能性。

应该指出的是，在快车假说中，存在两个不同层面的论点：① 迁徙的起源地，即台湾起源论；② 迁移的模式，即从印度尼西亚开始迅速扩散。在这项研究中，笔者针对“台湾起源论”分析了侗傣语族以及南岛语系西部群体包括马来人群这两类在以往研究中多被忽视的群体。研究发现，台湾不太可能是印度尼西亚的马来群体的祖先，至少在父系血统主体上不是。虽然台湾少数民族与印度尼西亚马来人群都起源于侗傣群体，他们的迁移是相对独立的事件，也说明整个南岛语系西部人群（台湾和马来）的父系遗传结构并不一致。

有趣的是，东南亚群岛和太平洋地区家猪的扩散与研究中的南岛语系西部群体的扩散方式大致相同。台湾地区以及直到密克罗尼西亚地区的猪都是直接来源于东亚大陆，而那些东南亚群岛以及波利尼西亚的品种则来自中南半岛。可以猜测家猪是因为人类群体的早期迁徙而被带入的，也提示人类进入东南亚群岛以及太平洋区域是通过两条不同的路径[49]。

事实上，本研究的观察结果与澳泰族群（其亚群体包括侗傣语系群体、马来-波利尼西亚语族以及台湾少数民族）[5]为单一起源整体的理论是相一致的。研究结果证明，要研究南岛语系的起源问题，将侗傣语系群体以及马来语群体涵括在内是十分必要的。如果没有这两类群体，波利尼西亚人群与台湾少数民族会显示出某些明显的相似性，从而产生南岛语系群体起源于台湾这样的结论。

研究结果表明，北部湾很有可能是马来人群父系遗传的发源地。鉴于人群从印度尼西亚东部向太平洋岛屿[23,50-55]迁徙的复杂性，以及南岛语系东西部之间显著遗传分化[27]，分析时没有将波利尼西亚人群的数据包括进来。相反，研究中只分析了南岛语系西部群体。波利尼西亚人群中 O1a - M119 的缺失值得思考，不能简单地归因于瓶颈效应[21-25]，因为波利尼西亚群体中出现 Y 染色体单倍群的高度多样性[23,50]。

与父系结构的研究结果一致，一项关于马来西亚半岛群体线粒体 DNA 的研究也表明，马来原住民的祖先来自末次盛冰期的印度支那附近[56]。这些祖先随后通过马来半岛分散到东南亚岛屿各处[56]。马来人群的线粒体 DNA 研究也证明，如果南岛语系真的是从台湾起源迁徙的话，在人口比例上是偏少的[57]。

本研究大部分的结论都是基于 O1a * 的分析数据，不过这只是这些人群中发现的 Y 染色体谱系的一部分。这一谱系在台湾群体中频率十分高，但在马来-波利尼西亚人群以及侗傣人群中没有那么高频。马来和侗傣族群很可能是因为某些群体事件而融合了其他 Y 染色体谱系，也可能存在另一些不同来源的父系血统，例如南岛语系形成前的印度尼西亚原始原住民群体，或者更多的是近期来自南亚的迁徙[29]。东亚与东南亚的遗传结构远比预期的复杂。

4.3.4 结论

研究结果显示，侗傣语族群体与南岛语系西部群体在父系遗传结构上比东亚其他族群都更近。Y 染色体单倍型 O1a - M119 也是侗傣族群与南岛西部人群中最主要的单倍

型。它内部的STR多样性显示了台湾和马来这两类南岛语系群体，是分别由侗傣族群独立起源的。因此，最有可能的情况是马来-波利尼西亚人群主要是从侗傣族群的发源地——北部湾起源，通过越南走廊迁徙到印度尼西亚。相反，台湾少数民族则是直接从中国大陆迁移的。本研究结果证明了在遗传上一个超语系的族群的存在，这一族群包括台湾少数民族、侗傣人群和马来-波利尼西亚人。

参考文献

[1] Grimes B F. Ethnologue: languages of the world fourteenth edition. Dallas: International Academic Bookstore, 2002.

[2] Murdock G P. Genetic classification of the Austronesian language: a key to Oceanic culture history. Ethnology, 1964, 3: 117-126.

[3] Blust R. Subgroup, circularity and extinction: some issues in Austronesian comparative linguistics//Zeitoun E, Li P J K. Selected papers from the eight international conference on Austronesian linguistics: 28-30 December 1997, Taipei. 1999.

[4] Bellwood P A. Mans conquest of the Pacific: the prehistory Southeast Asia and Oceania. Auckland: Williams Collins Publishers LTD, 1978.

[5] Benedict P K. Austro-Thai: language and culture. Human Relations Area Files Press, 1975.

[6] 宋蜀华. 百越. 长春：吉林教育出版社，1991.

[7] 林惠祥. 人类学论著. 福州：福建人民出版社，1981.

[8] 陈桥驿. 吴越文化论从. 北京：中华书局，1999.

[9] Shaffer L N. Maritime Southeast Asia to 1500. New York: M. E. Sharpe Inc, 1996.

[10] 张光直. 中国东南海岸考古与南岛语族的起源. 南方民族考古1. 成都：四川大学出版社，1987.

[11] Benedict P K. Austro-Tai parallel, a tonal chamic language on Hainan. Computational Analyses of Asian and African Languages, 1984, 22: 83-86.

[12] 倪大白. 南岛语与百越诸语的关系. 民族语文，1994，3：21-35.

[13] 蒙斯牧. 侗泰语与南岛语的历时比较研究. 贵州民族研究，1995，62：76-91.

[14] 何平. 南岛语民族的起源及其与中国南方民族的历史关系. 云南民族大学学报(哲学社会科学版)，2003，20：45-48.

[15] 吴安其. 台湾原住民的语言及其历史——兼论南岛语数词反映的南岛语史. 世界民族，2004，45：68-78.

[16] Bellwood P A. Prehistory of the Indo-Malaysian Archipelago. Sydney: Academic Press, 1985.

[17] Bellwood P A, Fox J J, Tryon D. The Austronesians: historical and comparative perspectives. Canberra: The Australian National University, 1995.

[18] Solheim W G II. South-east Asia and Korea from the beginnings of food production to the first states//De Laet S J. The history of humanity. London: Routledge, 1994: 468-481.

[19] Oppenheimer S J, Richards M. Polynesian origins. Slow boat to Melanesia? Nature, 2001, 410: 166-167.

[20] Diamond J M. The express train to Polynesia. Nature, 1988, 336: 307-308.

[21] Melton T, Peterson R, Redd A J, et al. Polynesian genetic affinities with Southeast Asian

populations as identified by mtDNA analysis. Am J Hum Genet, 1995, 57: 403 - 414.

[22] Sykes B, Leiboff A, Low-Beer J, et al. The origins of the Polynesians: an interpretation from mitochondrial lineage analysis. Am J Hum Genet, 1995, 57: 1463 - 1475.

[23] Su B, Jin L, Underhill P, et al. Polynesian origins: insights from the Y chromosome. Proc Natl Acad Sci USA, 2000, 97: 8225 - 8228.

[24] Richards M, Oppenheimer S, Sykes B. mtDNA suggests Polynesian origins in Eastern Indonesia. Am J Hum Genet, 1998, 63: 1234 - 1236.

[25] Lum K J, Cann R L. mtDNA lineage analyses origins and migrations of Micronesians and Polynesians. Am J Phys Anthropol, 2000, 113: 151 - 168.

[26] Trejaut J A, Kivisild T, Loo J H, et al. Traces of archaic mitochondrial lineages persist in Austronesian-speaking formosan populations. PLoS Biol, 2005, 3: e247.

[27] Shepard E M, Chow R A, Suafo'a E, et al. Autosomal STR variation in five Austronesian populations. Hum Biol, 2005, 77: 825 - 851.

[28] Kayser M, Lao O, Saar K, et al. Genome-wide analysis indicates more Asian than Melanesian ancestry of polynesians. Am J Hum Genet, 2008, 82(1): 194 - 198.

[29] Karafet T M, Lansing J S, Redd A J, et al. Balinese Y chromosome perspective on the peopling of Indonesia: genetic contributions from pre-neolithic hunter-gatherers, Austronesian farmers, and Indian traders. Hum Biol, 2005, 77: 93 - 114.

[30] Lansing J S, Cox M P, Downey S S, et al. Coevolution of languages and genes on the island of Sumba, eastern Indonesia. Proc Natl Acad Sci USA, 2007, 104: 16022 - 16026.

[31] Su B, Xiao J, Underhill P, et al. Y chromosome evidence for a northward migration of modern humans into Eastern Asia during the Last Ice Age. Am J Hum Genet, 1999, 65: 1718 - 1724.

[32] Su B, Xiao C, Deka R, et al. Y chromosome haplotypes reveal prehistorical migrations to the Himalayas. Hum Genet, 2000, 107: 582 - 590.

[33] Wen B, Xie X, Gao S, et al. Analyses of genetic structure of Tibeto-Burman populations reveals sex-biased admixture in southern Tibeto-Burmans. Am J Hum Genet, 2004, 74: 856 - 865.

[34] Wen B, Li H, Lu D, et al. Genetic evidence supports demic diffusion of Han culture. Nature, 2004, 431: 302 - 305.

[35] The Y Chromosome Consortium. A nomenclature system of the tree of human Y chromosomal binary haplogroup. Genome Res, 2002, 12: 339 - 348.

[36] Kayser M, Caglià A, Corach D, et al. Evaluation of Y chromosomal STRs: a multicenter study. Int J Legal Med, 1997, 110: 125 - 133.

[37] Dupanloup I, Bertorelle G. Inferring admixture proportions from molecular data: extension to any number or parental populations. Mol Biol Evol, 2001, 18: 672 - 675.

[38] Slatkin M A. measure of population subdivision based on microsatellite allele frequencies. Genetics, 1995, 139: 457 - 462.

[39] Schneider S, Roessli D, Excoffier L. Arlequin: Ver. 2.000. A software for population genetic analysis. Geneva, Genetics and Biometry Laboratory, Univ of Geneva, 2000.

[40] Nei M. Molecular evolutionary genetics. New York: Columbia University Press, 1987.

[41] Kimmel M, Chakraborty R. Measures of variation at DNA repeat loci under a general stepwise mutation model. Theor Popul Biol, 1996, 50: 345 - 367.

[42] Zhivotovsky L A, Underhill P A, Cinnioğlu C, et al. The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. Am J Hum Genet, 2004, 74: 50 - 61.

[43] Li D, Li H, Ou C, et al. Paternal genetic structure of Hainan aborigines isolated at the entrance to East Asia. PLoS One, 2008, 3: e2168.

[44] Mona S, Tommaseo-Ponzetta M, Brauer S, et al. Patterns of Y chromosome diversity intersect with the Trans-New Guinea hypothesis. Mol Biol Evol, 2007, 24: 2546 - 2555.

[45] Kayser M, Brauer S, Weiss G, et al. Reduced Y chromosome, but not mitochondrial DNA, diversity in human populations from West New Guinea. Am J Hum Genet, 2003, 72: 281 - 302.

[46] Kumar V, Reddy A N, Babu J P, et al. Y chromosome evidence suggests a common paternal heritage of Austro-Asiatic populations. BMC Evol Biol, 2007, 7: 47.

[47] Thurgood G. Phan Rang Cham and Utsat: Tonogenetic themes and variants//Edmondson J A, Gregerson K J. Tonality in Austronesian languages. Oceanic Linguistics Special Publication, 24. Honolulu: University of Hawaii Press, 1993: 91 - 106.

[48] Shikama T, Ling C C, Shimoda N, et al. Discovery of fossil Homo sapiens from Chochen in Taiwan. J Anthrop Soc Nippon, 1976, 84: 131 - 138.

[49] Larson G, Cucchi T, Fujita M, et al. Phylogeny and ancient DNA of Sus provides insights into neolithic expansion in Island Southeast Asia and Oceania. Proc Natl Acad Sci USA, 2007, 104: 4834 - 4839.

[50] Capelli C, Wilson J F, Richards M, et al. A predominantly indigenous paternal heritage for the Austronesian-speaking peoples of Insular Southeast Asia and Oceania. Am J Hum Genet, 2001, 68: 432 - 443.

[51] Underhill P A, Shen P, Lin A A, et al. Y chromosome sequence variation and the history of human populations. Nat Genet, 2000, 26: 358 - 361.

[52] Hurles M E, Irven C, Nicholson J, et al. European Y chromosomal lineages in Polynesians: a contrast to the population structure revealed by mtDNA. Am J Hum Genet, 1998, 63: 1793 - 1806.

[53] Lum J K, Cann R L, Martinson J J, et al. Mitochondrial and nuclear genetic relationships among Pacific island and Asian populations. Am J Hum Genet, 1998, 63: 613 - 624.

[54] Kayser M, Brauer S, Weiss G, et al. Melanesian origin of Polynesian Y chromosomes. Curr Biol, 2000, 10: 1237 - 1246.

[55] Terrell J E, Kelly K M, Rainbird P. Foregone conclusions? In search of 'Papuans' and 'Austronesians'. Curr Anthropol, 2001, 42: 97 - 124.

[56] Hill C, Soares P, Mormina M, et al. Phylogeography and ethnogenesis of aboriginal Southeast Asians. Mol Biol Evol, 2006, 23: 2480 - 2491.

[57] Hill C, Soares P, Mormina M, et al. A mitochondrial stratigraphy for island Southeast Asia. Am J Hum Genet, 2007, 80: 29 - 43.

4.4 长江沿岸史前人群的Y染色体

4.4.1 研究背景

从1984年樋口(Higuchi)关于斑驴(Quagga)[1]以及1985年Pääbo关于埃及木乃伊[2]的研究开始,古DNA的研究已超过20年。虽然古DNA研究结果的真实性一直受到质疑,但它研究的合理性已开始被接受。制定严格的处理程序,以求最大限度减少潜在外源DNA的污染,进行仔细地验证[3-5],可以确保报道的古DNA数据的可靠性。已经有大量的古DNA研究成果被报道出来,大部分都是关于线粒体DNA的。因为在单个样本中线粒体DNA比核DNA有更多的拷贝数,更易提取。甚至千百万年前尼安德特人的线粒体DNA数据也被确认了[6,7]。关于古代核DNA的研究[8,9]也有成功的,说明从遗骸中提取的各类DNA都有可能获得信息,以解答人类学感兴趣的问题。

不同考古文化的人群之间,以及古代与现代人群间的遗传关系,是东亚考古学家十分感兴趣的一个话题[10]。与民族类群高度相关的Y染色体多样性是分析民族关系的最好材料之一[11,12]。不同族群中Y染色体单倍群的分布模式有很大的不同。例如,单倍群O1在南岛语族和侗台语族人群中是主要成分[13]。关于古DNA的Y染色体研究也有报道[14,15],大部分都集中在短串联重复(STR)的研究上。关于东亚遗迹的两份Y-STR研究报道揭示了古代匈奴人的社会结构[16]以及美洲印第安人和古代西伯利亚人之间的关系[17]。然而,单倍群不是仅仅依据Y-STR数据决定的,还要依据单核苷酸多态性(SNP)。最早被报道的古代Y-SNP数据是从美洲印第安人一个已灭绝部落的样本中检测出来的[18]。另有文献报道了古代南部西伯利亚11个样本的Y-SNP数据[19]。在本节中,笔者列出了中国48个古代样本的Y-SNP数据,大部分来自长江沿岸,为这一地区史前人群的遗传多样性提供一个概况。

距今3 000～9 000年前的新石器时代,有几种不同的考古文化区系出现在东亚。整个新石器时代,各区系之间的文化差异一直持续,直到青铜时代。在长江流域,有两类截然不同的新石器文化。在长江中游的三峡地区,文化系列[20]包括楠木园文化(前6000—前5000)、柳林溪文化(前5000—前4400)、大溪文化(前4400—前3300)、屈家岭文化(前3300—前2500),以及石家河文化庙坪类型(前2500—前2200)。在长江口区域[21]的文化类型包括跨湖桥文化(约前6000)、马家浜文化(前5100—前3900)、崧泽文化(前3900—前3300)、良渚文化(前3300—前2100)以及马桥文化(前1900—前1200)。虽然不同时期的文化在同一系列中显示出延续性,但跨文化间的交流证据尚未找到。这两处区系之间还有一些孤立的文化系列[22],包括仙人洞文化(约前8000)、石碾山文化(前4000—前2500)、山背文化(前2800—前2000)和吴城文化(前1500—前1100)。类似的,在中国北部的黄河流域也有一些同时期的文化系列。最重要的包括裴李岗文化(前7000—前5000)、仰韶文化(前5000—前3000)、龙山文化(前3000—前2200)、二里头文化(前2200—前1500)以及殷商文化(前1500—前1100),这些被认为是中华文明的主要来源。有些观点认为长江沿岸的新石器文化也是中华文明的起源之一。这促使人类学家运用遗传学方法去探寻东亚不同类型的新石器文化先民是否现代中国或者其他国家人群的祖先,了

解史前的文化多样性是否与遗传多样性相符合。

4.4.2　材料和方法

1. 考古遗址和样品

本研究中采集的样本分别来自 5 个遗址：马桥、新地里、吴城、大溪和陶寺。具体位置如图 4-10 中标记。大部分样本属于 4 个不同的文化：最早的是大溪文化；其次是龙山文化和良渚文化，两者大约在同一时期；最后是青铜时期商代的吴城文化。这些文化是史前中国最有代表性的类型。另外还收集了一些历史时期(晚于公元前 841)的与新石器时代样本相同地点出土的样本用于比较，其中大部分属于汉代。同种文化的样本取自不同墓地，以防相关样本存在偏向性。

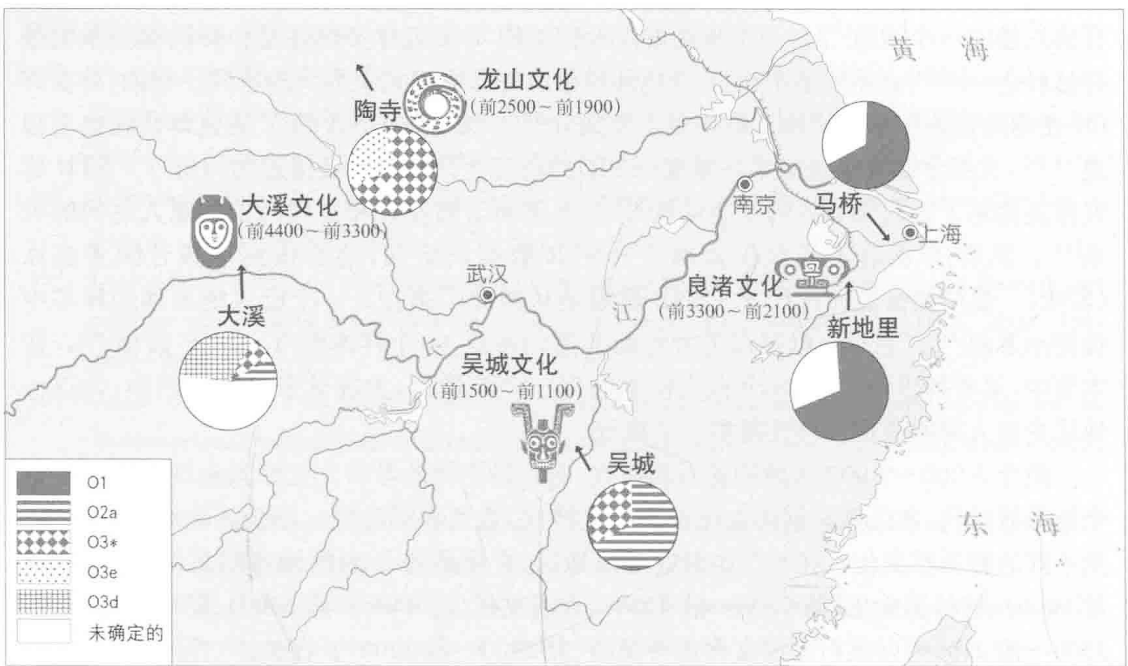


图 4-10　考古遗址、文化以及 Y-SNP 单倍群的分布

这些遗骨被直接埋在黏性黄土中，没有棺椁保护。大部分长江流域挖掘出的骨骼已经腐烂，但一些建造在高地上的墓地出土的样本情况较好，被黄土紧密覆盖并保护着。我们选择完整以及坚硬的骨骼作为样本。每具骨头直到被运进专用古 DNA 实验室前才会将包覆其约 2 in(1 in=0.025 4 m)的黄土去除，以使挖掘与搬运途中受到污染的风险降到最低。对每个样本 DNA 的抽提与分型要在挖掘后一个月内进行，确保样品的新鲜度以及扩增的最佳时机[23]。每具骨骼的性别鉴定依据 Murail 等在 1999 年发明的方法[24]。只对男性样本进行 Y-SNP 基因分型。总体而言，本次实验包括 56 个个体样本，在相同墓地用同样方法还收集了一些木片材料用于阴性对照。虽然动物遗骸是更好的

比对材料，但没有在人骨附近发现动物骨骼。

2. 避免污染以确保真实性的措施

笔者有两个专门开展古DNA研究工作的独立实验室。正如之前提到的，样本被墓地的黄土包裹着，在转移到实验室之前没有任何人接触过。即使这样，还要由尽量少的人类学家戴着面具与帽子用手套进行处理。在运输过程中，样品被包裹在密封的塑料袋中。

古DNA实验室按照之前古DNA的研究标准[5]严格进行控制，比如不同方法的常规灭菌(DNA酶降解、正向气压、次氯酸钠及紫外照射)、空气过滤、不同实验步骤分别在隔离室内进行(3个单间分别供样本清洗、DNA抽提和PCR体系配制)。每个实验室最多供3名研究员工作，穿着全身防护服，使用专门的设备与试剂。PCR准备工作在实验室指定房间进行，将体系配制好并仔细密封。进行PCR的房间远离其余房间，并且避免它们之间的空气流动。这样，实际上是将PCR扩增产物与PCR前的准备过程分离开来。

只有女性研究者参与PCR准备工作，以避免现代人Y染色体DNA可能造成的污染。每具骨骼分成几个部分进行采样：主要是牙齿、距骨、跟骨、颈椎等。选择的牙齿都是完整的(无龋损)，仍然连着下颌。只选用具有厚层皮质的骨骼，因为与松质骨相比，高密度骨质有两个优点：一是DNA附着的羟基磷灰石矿物晶体的数量比松质骨中要高；二是它能帮助防止外源污染。骨头的外表面会被磨去大约2 mm厚。两个实验室会对相同骨架的不同部分样品进行同样的操作来做重复比对。对于每个SNP，每个实验室至少重复3轮扩增。还要进行阴性提取对照。木片与人骨对照也采用相同的步骤处理。完全空白抽提与扩增也仍然作为阴性对照。

笔者没有对PCR产物进行克隆，或者按照某些研究者的标准[25,26]对反应的起始模板进行量化，因为损伤或者跳跃PCR并不会造成Y-SNP等位基因的判定错误。

3. DNA的提取和扩增

根据Fily等于1998年发表的操作步骤[27]进行DNA提取，没有额外的改动。骨骼样品在装有液氮的冷冻研磨机中粉碎，用硅胶吸附法提取DNA[28]。相关步骤已经在很多文献中报道过，此处不再赘述。抽提DNA的超净工作台有正向气压以及紫外照射装置，在处理每两个样本之间进行清洁消毒。一次只能提取一个样品以防交叉污染。

用于分型的SNP位点有M119、M95、M122、M7和M134。它们按照2002年制定的YCC命名规定[29]构成了5个单倍群(O1、O2a、O3*、O3d和O3e)。扩增步骤与之前发表的[11,30]相一致。PCR反应退火循环增加到60次，PCR产物的长度大都在100～200 bp，短于长度为100～500 bp[31]的普通古DNA。

4.4.3 结果和讨论

对照样本(木片)以及8个个体骨骼的抽提物没有得到扩增产物。这8个个体可能已经严重降解(在表4-8中显示为缺失数据)。抽提成功的样本与抽提不成功的样本之间没有体质特征差别。因为那些实验失败的样本染色体上决定性别的片段也无法扩增

出来，因此不能判断它们是否存在性别的形态学鉴定错误。有部分样本，不是所有5个SNP都能扩增出来，如果没有发现突变的等位基因则不能确定其单倍型（表4-8中显示为未确定）。不过，大部分个体都成功扩增了，也确认了其中半数的单倍型。至少有62.5%的个体（30/48）属于单倍群O，在今天仍然是东亚人群主要的单倍群。因此，这些古代结果与现代人群没有明显差别。这些所得的DNA类型存在"谱系合理性"（Y染色体单倍群结构），也进一步确认古DNA的真实性。先前报道过的古Y染色体结果都是从保存在冰冻环境里的样本中获取的[16-19]。但是，本研究的样本并没有埋在真正冰冻的环境中，因此，相对较高的扩增率是基于对保存最完好遗体的严格挑选而得。

表4-8　考古遗址样本Y-SNP单倍群统计数据

遗　址	文化类型	样本量	O1	O2a	O3*	O3d	O3e	未确定	缺失
马桥	良渚时期	6	4					2	
	历史时期	3	2					1	
新地里	良渚时期	9	5					3	1
	历史时期	4	3						1
吴城	吴城时期	4		2	1				1
大溪	大溪时期	20		1	1	5		9	4
	历史时期	5			2			3	
陶寺	龙山时期	5			3		1		1

现代人群中的频率（%）

语系	文献	样本量	O1	O2a	O3*	O3d	O3e		
侗傣	[33]	1 465	34.87	26.52	10.05	0.19	9.36		
南岛	[33]	381	24.07	16.25	22.41	2.01	2.32		
汉藏	[11]	281	1.57	8.49	17.24	0.53	32.74		
苗瑶	[34]	934	8.49	21.87	17.19	10.45	26.38		
南亚	[33]	140	1.79	31.40	18.49	6.25	19.73		
阿尔泰	[11]	303	1.74		4.11		6.60		

在良渚文化的两个遗址中，只发现了O1-M119单倍群，并且两个遗址中O1单倍群的频率几乎是相同的。这说明这两处遗址中的古人可能属于同一个群体。从相同遗址出土的历史时期的样本在单倍群类型上与史前的样本并没有什么不同。甚至上海周围同一地区的现代人群也包含很大一部分O1单倍群[32]。这一地区从新石器时代直到现代Y单倍群类型的一致性，说明该处的人群可能没有被替代过。O1在中国台湾少数民族以及中国西南部的侗傣语系人群中出现的频率最高[33]。因此，台湾少数民族、侗傣族群以

及良渚文化先民之间很可能有着某种紧密的联系。

高频率的O3d-M7仅仅发现于大溪文化遗址。O3d在中国现代人群中十分罕见，苗瑶人群中有一小部分O3d[34]。在这些苗瑶人群中，畲族和布努人中O3d的频率最高[11]。由于O3d只在苗瑶人群中有少量的频率，大溪文化的先民很可能是现代苗瑶人群的祖先。大溪遗址历史时期样本中O3d的缺失(也可能是由于样本量太小未被发现)，以及现代苗瑶人群向西南方向的迁徙，或许表明三峡地区的史前人群已被取代。

O3*-M122和O2a-M95存在于吴城、大溪、陶寺3个遗址中。由于O3*是现代东亚人群中最常见的单倍型，O2a在中国西南的不同人群中都有发现，我们无法根据这些共享的单倍群来推断这些遗址人群之间的亲近关系或是基因交流。长江流域与黄河流域的人群之间还是有着一些基因差异的，比如O2a没有在陶寺遗址中发现。

O1没有在良渚文化以外被发现，表明中国沿海和内陆人群间有着明显的遗传差异。这也在现代人群中有体现。O1在东亚沿海岸有所分布，从中国东北到马来西亚与印度尼西亚[11,13,33]。这种分布或许证明了早期东亚人群至少有两条迁徙路线：沿海路线与内陆路线。两条路线的史前人群之间几乎没有任何基因交流。

总之，东亚地区史前人群的遗传多样性揭示了不同考古文化类型是基于不同来源的人群发展形成的。遗传隔离或许远远早于新石器文化的多样性发展，并且一直持续到那些史前文明慢慢融合成中华文明。

参考文献

[1] Higuchi R, Bowman B, Freiberger M, et al. DNA sequences from the quagga, an extinct member of the horse family. Nature, 1984, 312: 282-284.

[2] Pääbo S. Molecular cloning of ancient Egyptian mummy DNA. Nature, 1985, 314: 644-645.

[3] Pääbo S. Ancient DNA: extraction, characterization, molecular cloning and enzymatic amplification. Proc Natl Acad Sci USA, 1989, 86: 1939-1943.

[4] Stoneking M. Ancient DNA: how do you know when you have it and what can you do with it? Am J Hum Genet, 1995, 57(6): 1259-1262.

[5] Pääbo S, Poinar H, Serre D, et al. Genetic analyses from ancient DNA. Annu Rev Genet, 2004, 38: 645-679.

[6] Serre D, Langaney A, Chech M, et al. No evidence of Neandertal mtDNA contribution to early modern humans. PLoS Biol, 2004, 2(3): e57.

[7] Dalton R. Neanderthal DNA yields to genome foray. Nature, 2006, 441(7091): 260-261.

[8] Lawlor D A, Dickel C D, Hauswirth W W, et al. Ancient HLA genes from 7500-year-old archaeological remains. Nature, 1991, 349(6312): 785-788.

[9] Béraud-Colomb E, Roubin R, Martin J, et al. Human P-globin gene polymorphisms characterized in DNA extracted from ancient bones 12000 years old. Am J Hum Genet, 1995, 57: 1267-1274.

[10] 苏秉琦. 中国文明起源新探. 上海：三联书店，1999.

[11] Su B, Xiao J, Underhill P, et al. Y chromosome evidence for a northward migration of modern

humans into Eastern Asia during the Last Ice Age. Am J Hum Genet, 1999, 65: 1718 - 1724.

[12] Shi H, Dong Y L, Wen B, et al. Y chromosome evidence of southern origin of the East Asian-specific haplogroup O3 - M122. Am J Hum Genet, 2005, 77(3): 408 - 419.

[13] Zhang F, Su B, Zhang Y P, et al. Genetic studies of human diversity in East Asia. Philos Trans R Soc Lond B Biol Sci, 2007, 362: 987 - 995.

[14] Hummel S, Herrmann B. Y chromosome-specific DNA amplified in ancient human bone. Naturwissenschaften, 1991, 78(6): 266 - 267.

[15] Schultes T, Hummel S, Herrmann B. Amplification of Y chromosomal STRs from ancient skeletal material. Hum Genet, 1999, 104: 164 - 166.

[16] Keyser-Tracqui C, Crubezy E, Ludes B. Nuclear and mitochondrial DNA analysis of a 2000 - year-old necropolis in the Egyin Gol Valley of Mongolia. Am J Hum Genet, 2003, 73(2): 247 - 260.

[17] Ricaut F X, Fedoseeva A, Keyser-Tracqui C, et al. Ancient DNA analysis of human neolithic remains found in northeastern Siberia. Am J Phys Anthropol, 2005, 126(4): 458 - 462.

[18] Kuch M, Grocke D R, Knyf M C, et al. A preliminary analysis of the DNA and diet of the extinct Beothuk: A systematic approach to ancient human DNA. Am J Phys Anthropol, 2007, 132(4): 594 - 604.

[19] Bouakaze C, Keyser C, Amory S, et al. First successful assay of Y-SNP typing by SNaPshot minisequencing on ancient DNA. Int J Legal Med, 2007, 121(6): 493 - 499.

[20] 国家文物局. 三峡考古之发现：巫山大溪遗址第三次发掘. 武汉：湖北科技出版社，1998.

[21] 浙江省文物考古研究所. 良渚文化研究. 北京：科学出版社，1999.

[22] 彭明瀚. 吴城文化研究. 北京：文物出版社，2005.

[23] Pruvost M, Schwarz R, Correia V B, et al. Freshly excavated fossil bones are best for amplification of ancient DNA. Proc Natl Acad Sci USA, 2007, 104(3): 739 - 744.

[24] Murail P, Bruzek J, Braga J. A new approach to sexual diagnosis in past populations: practical adjustments from Van Vark's procedure. Int J Osteoarchaeol, 1999, 9: 39 - 53.

[25] Cooper A, Poinar H N. Ancient DNA: do it right or not at all. Science, 2000, 289: 1139.

[26] Gilbert M T, Bandelt H J, Hofreiter M, et al. Assessing ancient DNA studies. Trends Ecol Evol, 2005, 20: 541 - 544.

[27] Fily M L, Crubézy E, Courtaud P, et al. Analyse paléogénétique des sujets de la grotte sépulcrale d'Elzarreko Karbia (Bronze ancien, Pays Basque). CR Acad Sci Ⅲ 321: 79 - 85.

[28] Gilbert M T, Willerslev E, Hansen A J, et al. Distribution patterns of postmortem damage in human mitochondrial DNA. Am J Hum Genet, 2003, 72: 32 - 47.

[29] The Y Chromosome Consortium. A nomenclature system of the tree of human Y chromosomal binary haplogroup. Genome Res, 2002, 12: 339 - 348.

[30] Ke Y, Su B, Xiao J, et al. Y chromosome haplotype distribution in Han Chinese populations and modern human origin in East Asians. Sci China C Life Sci, 2001, 44: 225 - 232.

[31] Hofreiter M, Serre D, Poinar H N, et al. Ancient DNA. Nat Rev Genet, 2001, 2: 353 - 359.

[32] Wen B, Li H, Lu D R, et al. Genetic evidence supports demic diffusion of Han culture. Nature,

2004, 431: 302－305.
[33] 李辉. 澳泰族群的遗传结构. 复旦大学人类生物学博士学位论文, 2005.
[34] 奉恒高. 瑶族通史. 北京: 民族出版社, 2007.

4.5 海南原住民父系遗传结构

4.5.1 研究背景

众所周知,因为与欧亚大陆西部群体隔离时间的久远,东亚群体的遗传结构中产生了许多特征性变异[1]。东亚人群和他们特殊的遗传特征的起源一直是人类群体向东亚扩张的研究中最有趣的部分,在学术界被深入探讨。然而,现代人具体何时何地进入东亚一直存在争议。许多研究观察到东亚人群的遗传结构呈现出南北向的梯度变化[2-4],所以不同的研究者分别提出了南方起源[5]和北方起源[6]的假说。也有人提出东亚人群是南北方移民的混合结果。稳定的遗传物质Y染色体的研究有助于解决争议。Y染色体单倍群研究中,部分单倍群由北方起源,可以把一些东亚的成分追溯到中亚群体中[7]。而大多数的单倍群,尤其是东亚的主流单倍群O[8,9],起源于欧亚大陆东部的南方[2,10,11]。由于东亚群体中南方移民是主流,所以南方入口对于东亚群体的形成非常重要,认识到这一点很重要。南方入口可能处于中国和印度支那半岛各国(缅甸、老挝、越南)的边界处。

然而,由于东亚人群历史上发生了大量的群体回流迁徙事件,南方入口处的群体初始多样性结构已经被深埋起来了。这些事件中最显著的是汉族[4]和藏缅民族[8,12]向中国南方和东南亚的回迁。大部分的南方群体,甚至是东南亚岛屿和新几内亚的人群,都被这些北方回流者"搅乱"了[13-17]。所以,我们已经很难了解群体进入东亚时的原初遗传结构以及不同Y单倍群途径的不同路径。在东亚入口处隔离的群体对解决这一问题最有帮助。

本节报道了在东亚最南端的海南岛(图4－11a)发现的几个具有相对原始并未被"搅乱"的Y染色体遗传结构的原住民人群。这些海南岛原住民父系的遗传结构与东亚大陆的其他群体有明显的差异。

海南岛是位于东亚和东南亚之间的北部湾中的一个大岛。在末次冰川期中当海平面比现在低得多的时期,海南岛与大陆相连[18,19],处于现代人类从东南亚向东亚迁徙的路径上(图4－11a)。海南的6个原住民群体世代居住在海南岛的中部和南部山区(图4－11b),可能是现代人最初迁移群体的直接后裔。自从7 000～11 000年的海进[19]把海南岛与大陆分离开来,这些原住民群体就可能开始隔离发展,延续了数千年。海南原住民的族群分类与语言学分类完全一致[20]。他们分为两个类群:黎族和仡隆人,他们的语言都属于侗傣语系(也称台-卡岱语系)的外围类群,在有些方面与南岛语系的马来类群有诸多相似处[21]。据2000年统计,黎族人口有120多万,文化多样,分为5个支系(图4－11b),分别为:侾黎、杞黎、润黎、美孚黎和加茂黎。每个支系人口都不少,其中最小的美孚黎也达到6万人口。5个支系各有不同的语言,相互不能通话[22]。仡隆人群体有8万人口,他们与黎族有截然不同的语言和文化特征。因为目前可供参考的海南考古学研究太少,海南原住民人群的历史还不太清楚。但其中一项关于三亚落笔洞的重要发现,

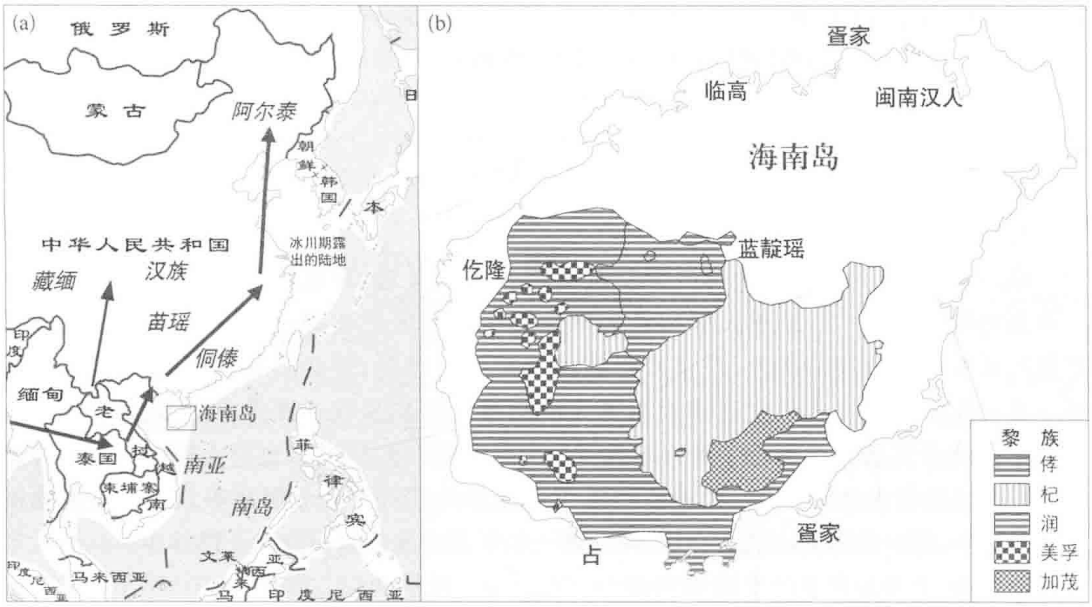


图 4－11　海南岛在东亚人群迁徙历史中的重要性和海南原住民的分布

(a) 在末次盛冰期海南岛与大陆相连，是现代人类从东南亚向东亚迁移的入口之一。图中箭头代表着现代人通过南方入口进入东亚的可能迁移路线。(b) 原住民黎族的 5 个支系在海南岛的分布；仡隆人是海南岛的另一个原住民；临高是侗水语支的一个古老的迁移群体；占人是来自越南的南岛语系群体；蓝靛瑶和汉族群体均来自大陆；畲家是一个没有迁移记录且居住在海上的渔民部落。

证实现代人类至少在大约 1 万年前的旧石器时代就已经居住在海南岛了[23]，远远早于其他东亚人群的回迁年代。海南最早的新石器时代遗址是距今 6 000 年的东方新街贝丘遗址[24]，位于仡隆人聚居地的海南岛东方市北黎河入海口 2.5 km 处。这说明新石器文化在 6 000 年以前就已经进入海南了。因此，有理由推测海南岛的原住民可能保持着最接近东亚祖先最初的遗传结构。然而至今为止对海南岛原住民没有做过什么直接的遗传学研究。相比之下，同样隔离的台湾少数民族的相关研究特别深入[9,17,25]，虽然他们距离东亚的入口相对较远(无论是印度支那半岛的南方入口还是中亚的北方入口)。

4.5.2　材料和方法

从海南原住民的 405 个男性志愿者上臂静脉血采集样本。每个群体的样本量列于表 4－9，群体的地理分布在图 4－11b 中。更多的群体详细信息可以通过 ISO639－3 编码[lic]、[jio]和[cuq]在民族语言网站(http：//www. ethnologue. com)搜索。志愿者都是健康人，来自不同村庄，姓氏不同，确保样本个体之间没有近缘关系。所有志愿者都签署了知情同意书。本研究通过了国家人类基因组南方研究中心伦理委员会的批准。

利用 PCR－RFLP 方法对 Y 染色体非重组区进行了 15 个 SNP 位点(M130、M89、M9、M45、M120、M119、M110、M101、P31、M95、M88、M122、M164、M159 和 M7)分型。

利用 Taqman 方法对 6 个 SNP(M210、M208、M48、M8、M217 和 M356)进行了分型。另外 7 个 SNP 位点(YAP、M15、M175、M111、M134、M117 和 M121)和 7 个 Y-STR 位点(DYS19、DYS389I、DYS389II、DYS390、DYS391、DYS392 和 DYS393)采用荧光引物 PCR 扩增后,变性产物由 ABI 3100 测序仪进行毛细管电泳扫描自动分型。选取的所有位点都是东亚人群中高多样性的[8-10,25,26]。Y 染色体单倍群确立的依据是由 YCC(Y 染色体委员会[27])发展而来的 ISOGG(国际遗传谱系学会[28])的 2007 年版人类 Y-DNA 单倍群进化树。

聚类树分析和主成分分析利用 SPSS13.0 软件进行。STR 中点连接算法网络结构利用 Network 4.201 构建[29],网络结构中的年龄估算用的突变率是 Zhivotovsky 等的数据(6.9×10^{-4}/25 年)[30]。在年龄的估计中,7 个 Y-STR 的总突变率是 1.932×10^{-4}/年。假定每代为 25 年,所以网络中每个突变需要 5 176 年。时间的估计由 BATWING 验证[31]。

4.5.3 结果和讨论

本项研究中,笔者对海南岛原住民所有 6 个人群进行了 Y 染色体多样性研究,分析了 405 个男性个体的 22 个 Y-SNP 和 7 个 Y-STR,并根据 YCC[27]和 ISOGG[28]国际命名系统进行单倍群确定。研究发现这些群体样本的 Y-SNP 单倍群频率非常接近(表 4-9)。在各个群体中频率最高的是单倍群 O1 和 O2,所以这两个单倍群可能是海南原住民最初的单倍群。海南原住民中的杞黎,这两个单倍群的总频率高达 100%,说明该群体可能在历史上经历了强烈的瓶颈效应。事实上杞黎的聚居地位于海南岛的五指山深处,他们很可能是由一个小的群体发展形成,这一点也可由杞黎群体 Y-STR 的低多样性印证。单倍群 O1 和 O2 在台湾少数民族和东亚大陆最南端的原住民族中也呈现较高的频率,但均低于海南原住民。在南中国海的南方和东方的南岛语系群体,如婆罗洲和菲律宾人群,也有高频率的 O1 和 O2 单倍群[9,13,17],这种遗传相似性可与侗傣和南岛语系之间的语言相似性相印证[21]。除了侗傣和南岛语系的人群,其他群体的 O1 和 O2 单倍群

表 4-9 Y 染色体单倍群频率

语言	群体	样本量	单倍群频率(%)										
			C3*	F*	K*	O1a*	O2*	O2a*	O3*	O3a1	O3a3	O3a5*	O3a5a*
黎	侾	74	4.05			22.97	5.41	59.46	5.41			1.35	1.35
黎	美孚	66				28.79		66.67	3.03			1.52	
黎	杞	62				8.06		91.94					
黎	润	75			1.33	32.00	5.33	58.67	1.33		1.33		
黎	加茂	50	2.00	2.00		40.00		46.00	6.00		2.00		2.00
仡央	仡隆	78	5.13			57.69	17.95	3.85	8.97	1.28		3.85	1.28

频率是相当低的，尤其是印度支那半岛西部原住民即南亚语系群体[26,32]。单倍群O3是汉藏语系人群常见的单倍群[10]，如汉族(50.51%)和藏缅群体(54.70%)，但在海南原住民中是少见的(6.91%)，而在台湾少数民族中达到11.36%(0～37.6%)[17,25,26]，在大陆的南方原住侗傣群体(19.60%)和中部原住苗瑶群体(54.02%)频率更高[32]。这表明相比台湾及大陆南方的原住民群体，海南原住民有较少的汉族男性基因的混合影响。另外与大陆群体不同，单倍群D、P、N和Q在海南原住民中全部都不存在。

为了了解海南原住民与其他人群的遗传关系，笔者用海南原住民和东亚其他群体[2,4,8-10,12,13,26]的Y-SNP单倍群频率进行了两种方法的聚类分析(图4-12)。在图4-12a的聚类树中，东亚群体聚类成两个组：南方组(侗傣等)和北方组(汉族等)。所有的大陆侗傣群体与南亚语系群体聚类。侗傣和南亚语人群散布在东南亚地区，他们的分布区是重合的(图4-11a)。笔者推测是由于这些群体之间发生了大量的基因交流，导致Y-SNP频率的类似，从而使他们聚成一类。黎族的各个支系形成了南方组的外围分支，历经瓶颈效应的隔离群体杞黎支系处于分支的最外围。这些结果表明，黎族与混合大陆群体有很大区别。在北方组，苗瑶、汉族和藏缅族群彼此非常接近，而南岛语系和阿尔泰语系人群都受到了汉藏群体的遗传影响。研究发现这一聚类组的外围分支是台湾少数民族与海南原住民仡隆人，他们的单倍群多样性都比较低。但是，仡隆人为什么与台湾少数民族相似，原因尚有待探讨。

主成分分析(图4-12b)也显示，被研究的群体同样聚类成了南方组和北方组。根据这一主成分图，台湾少数民族接近北方组，而两种海南原住民聚类进入了不同的组。仡隆人归进了北方组，与台湾少数民族在一起。所有的黎族支系群体都特别接近南北分化方向的南方末端，再次证明他们遗传结构的隔离性。这一分类对于分类中处于南方组最南端的杞黎尤为合理。

根据SNP分析，海南原住民同北方的群体一直保持隔离，尤其是与携带高频率O3单倍群迁移到中国南方的汉族群体保持隔离。同时南迁的苗瑶群体也对南方O3频率的增加有作用。最近的研究揭示，苗瑶和汉藏语系的群体共同起源于中国西南部，而O3单倍群可能是他们最近共同祖先群体中的主体单倍群。正如前文所述，在海南原住民中O3单倍群几乎是缺失的。在海南原住民中占优势的单倍群O1和O2，在大陆侗傣语系群体和台湾少数民族中同样高频。为此SNP单倍群分析无法排除海南原住民与大陆侗傣语系群体及台湾少数民族之间的基因流动。除了聚类树分析和主成分分析外，笔者又在海南原住民、台湾少数民族和大陆侗傣族群中进行了基于O1和O2单倍群的STR网络分析(图4-13)。

在网络分析中，参考样本包括台湾少数民族和海南北面的广西、广东两省的侗傣群体[26]。在图4-13两个网络图的顶端，海南原住民群体(显示为黑色的节点)都形成了几乎独立的分支(粗线条显示)，只有很少数个体来自其他群体，说明海南原住民群体与其他侗傣群体和台湾少数民族是长期隔离的。此外，大部分黎族群体的单倍型均处在顶端的分支，而仡隆人的单倍型在两个网络中都形成了两个较小的分支(网络中较下方的黑色

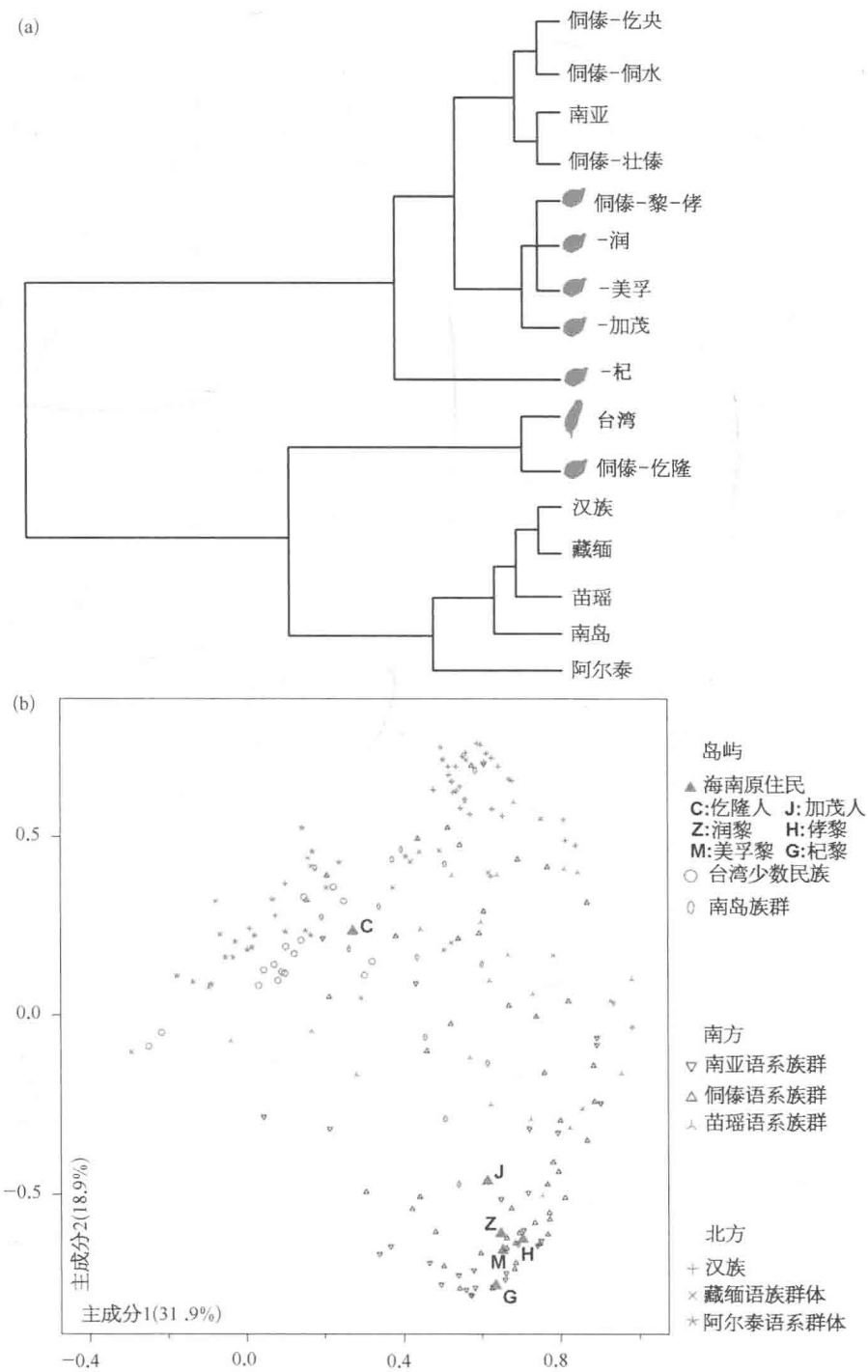


图 4-12 海南原住民和其他东亚族群的聚类分析

(a) 分子系统树展示了黎族与仡隆人的不同，黎族形成了侗傣外围的分支；(b) 主成分分析体现了明显的南北分化，黎族和其他侗傣族群、南亚语系群体处于南端，且黎族处于最南端；仡隆人接近于台湾少数民族及南岛语系群体。

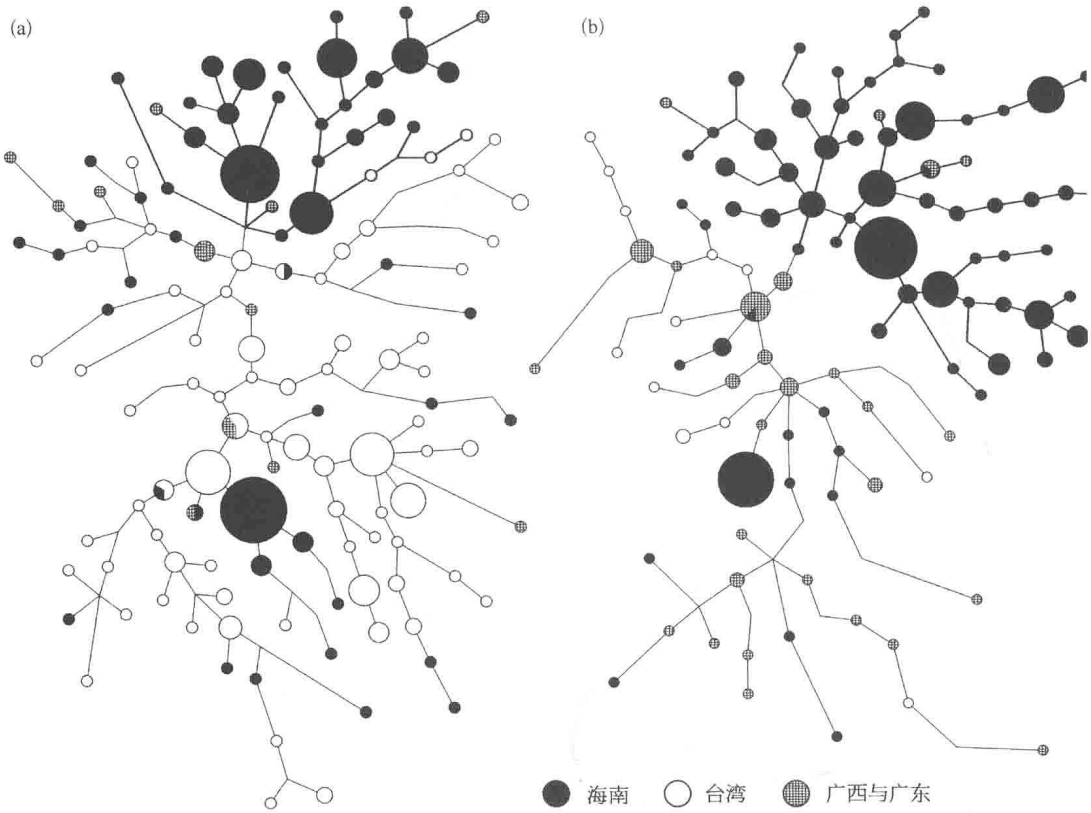


图4-13　Y-STR网络结构

(a) O1a*网络分析;(b) O2a*网络分析

由于STR网络结构的原始状态过于复杂,为了便于分析,截取最短的分子系统树图。在这个系统树中海南原住民占据了较多的分支。

节点)。黎族群体的分支相对较大,几乎占据网络图的1/3。在O2a*网络中,黎族的分支比大陆侗傣及台湾的群体大得多,说明黎族比大陆的侗傣群体和台湾少数民族更古老,而不是大陆侗傣群体的分化亚群。形成如此大的分支需要STR相当长时间的突变积累。估计相关的群体年龄,O1a*的年龄是3.6万年,其中海南分支大约是1.9万年。O2a*形成的年代是3.2万年前,其中海南分支形成年代约是2.6万年前(表4-10)。年

表4-10　单倍群O1a*和O2a*的年龄估计

	Network		Batwing	
	年龄(万年)	标准误	年龄(万年)	95%置信区间
O1a*	3.61	0.70	3.95	2.29~7.53
O1a*海南	1.89	0.56	1.45	0.72~3.30
O2a*	3.17	0.73	3.04	2.08~5.35
O2a*海南	2.57	0.76	1.96	1.13~4.45

注:Batwing所用有效群体大小为1 300。

代分析是建立在中国南方人群的基础上，不同于世界人群的 O2a* 年龄。由于 O1a* 和 O2a* 单倍群在中国北方几乎是缺失的，中国南方的 O1a* 和 O2a* 单倍群的年龄可能就是这两种单倍群进入东亚的时间。海南原住民 O1a* 和 O2a* 单倍群的年龄比较接近，都落在 1.8 万～2.6 万年。这个时间段正好与末次盛冰期（约 2 万年前）[18]一致。当时中国海域的大陆架高出海平面之上，正好为现代人向东亚的迁徙提供了一个捷径。

当然，进入东亚可能还有其他迁徙路线。比如东亚与东南亚边境的西段（从缅甸到云南乃至中国内地，图 4-11a）。笔者估计单倍群 O3 是汉藏和苗瑶族群的祖先通过西线携带进入东亚的，虽然目前证据还不充分（最近通过对苗瑶和孟高棉群体的调查提供了西侧路线的线索）。当然，O3 单倍群也可能同样来源于或许是唯一入口的东线，进入东亚的时间可能晚于 O1a* 和 O2a* 单倍群。进入东亚北方后，O3 单倍群通过向南回迁影响了大多数东亚群体。海南岛隔离人群的 O3 单倍群低缺，说明侗傣祖先最初通过海南进入东亚的时候并未携带 O3 单倍群。而 O2a* 应该是最早的移民沿着东线进入东亚的时候携带的最古老的单倍群之一（超过 4 万年）。O2a* 也可能是最早到达海南的单倍群。O2a* 海南分支的年龄估计是 2.6 万年，这个时间远早于考古学发现的海南最早出现人类的落笔洞遗址的年代。研究认为这个年代是海南原住民形成并开始隔离的年代，将来在海南可能会发现更早的考古遗址。

O1a* STR 网络分析显示，台湾少数民族的节点比海南原住民的节点更接近于网络的中心，说明台湾在地理上可能更接近于单倍群 O1a* 的起源。而且，O1a* 的网络图中海南分支比 O2a* 中的年龄要小一些，因此 O1a* 可能起源于海南岛以东的地区，之后回流到海南。在中国东部沿海地区 5 000 多年的新石器时代人类样本中也发现了 O1a* 单倍群[32]。东亚新石器时期大约开始于 8 000 年前。单倍群 O1a* 有可能随着新石器文化扩散到海南岛。但海南 O1a* 的年龄（约 1.8 万年）比在海南发现的最早新石器遗址要早（约 6 000 年）[24]。可能的解释是，估计的年龄是携带 O1a* 的人群从 O1a* 祖先群体（可能是台湾少数民族祖先）分离之后到达海南之前。台湾少数民族和海南原住民 O1a* 分歧时间大概是 2.2 万年（95%置信区间：1.25～4.68）。

综上所述，海南原住民源于末次盛冰期进入东亚的早期移民，并从此与大陆发生了隔离。海南原住民几乎未受到东亚大陆人群回迁的影响，为此海南原住民的 Y-SNP 单倍群组分最接近早期移民的最初遗传结构。建议海南原住民不仅可用于揭示东亚群体及其独特遗传结构的起源，而且也能作为进行东亚人群遗传学研究的模式人群。

众所周知，很多遗传学研究都需要在隔离群体中开展。例如，在隔离群体中复杂疾病的致病因素明显减少，这使分析变得更容易。另外，作为具有较大人口和悠久历史的群体，海南原住民的大多数群体可能并未经历明显的遗传漂变，所以虽然 Y-SNP 多样性偏低，但 Y-STR 多样性偏高。相比东亚其他群体，海南原住民应该有较高的常染色体多样性，较低的连锁不平衡，这些对于在进行疾病相关研究中避免因高连锁不平衡引

起的假阳性结果具有重要价值。建议利用海南原住民群体作为东亚隔离群体模式进行更多的遗传学研究。

参考文献

[1] Li H, Mukherjee N, Soundararajan U, et al. Geographically separate increases in the frequency of the derived ADH1B* 47His allele in Eastern and Western Asia. Am J Hum Genet, 2007, 81: 842-846.

[2] Su B, Xiao J, Underhill P, et al. Y chromosome evidence for a northward migration of modern humans into Eastern Asia during the Last Ice Age. Am J Hum Genet, 1999, 65: 1718-1724.

[3] 肖春杰,Cavalli-Sforza L L, Minch E, et al. 中国人群的等位基因地理分布图. 遗传学报,2000, 27: 1-6.

[4] Wen B, Li H, Lu D, et al. Genetic evidence supports demic diffusion of Han culture. Nature, 2004a, 431: 302-305.

[5] Jin L, Su B. Natives or immigrants: modern human origin in East Asia. Nat Rev Genet, 2000, 1: 126-133.

[6] Nei M, Roychoudhury A K. Evolutionary relationships of human populations on a global scale. Mol Biol Evol, 1993, 10: 927-943.

[7] Wells R S, Yuldasheva N, Ruzibakiev R, et al. The Eurasian heartland: a continental perspective on Y chromosome diversity. Proc Natl Acad Sci USA, 2001, 98: 10244-10249.

[8] Su B, Xiao C, Deka R, et al. Y chromosome haplotypes reveal prehistorical migrations to the Himalayas. Hum Genet, 2000b, 107: 582-590.

[9] Su B, Jin L, Underhill P, et al. Polynesian origins: insights from the Y chromosome. Proc Natl Acad Sci USA, 2000a, 97: 8225-8228.

[10] Shi H, Dong Y L, Wen B, et al. Y chromosome evidence of southern origin of the East Asian-specific haplogroup O3-M122. Am J Hum Genet, 2005, 77: 408-419.

[11] Zhang F, Su B, Zhang Y P, et al. Genetic studies of human diversity in East Asia. Philos Trans R Soc Lond B Biol Sci, 2007, 362: 987-995.

[12] Wen B, Xie X, Gao S, et al. Analyses of genetic structure of Tibeto-Burman populations reveals sex-biased admixture in southern Tibeto-Burmans. Am J Hum Genet, 2004b, 74: 856-865.

[13] Karafet T M, Lansing J S, Redd A J, et al. Balinese Y chromosome perspective on the peopling of Indonesia: genetic contributions from pre-neolithic hunter-gatherers, Austronesian farmers, and Indian traders. Hum Biol, 2005, 77: 93-114.

[14] Kayser M, Brauer S, Cordaux R, et al. Melanesian and Asian origins of Polynesians: mtDNA and Y chromosome gradients across the Pacific. Mol Biol Evol, 2006, 23: 2234-2244.

[15] Mona S, Tommaseo-Ponzetta M, Brauer S, et al. Patterns of Y chromosome diversity intersect with the Trans-New Guinea hypothesis. Mol Biol Evol, 2007, 24: 2546-2555.

[16] Hurles M E, Sykes B C, Jobling M A, et al. The dual origin of the Malagasy in Island Southeast Asia and East Africa: evidence from maternal and paternal lineages. Am J Hum Genet, 2005,

76：894－901.

[17] Capelli C, Wilson J F, Richards M, et al. A predominantly indigenous paternal heritage for the Austronesian-speaking peoples of insular Southeast Asia and Oceania. Am J Hum Genet, 2001, 68：432－443.

[18] 施雅风，崔之久，李吉均. 中国东部第四纪冰川与环境问题. 北京：科学出版社，1989.

[19] 邢关英. 从海南岛最初设治说起. 海南档案，2004，3：43－47.

[20] 王学萍. 中国黎族. 北京：民族出版社，2004.

[21] Thurgood G. Tai-Kadai and Austronesian：the nature of the historical relationship. Oceanic Linguistics, 1994, 33：345－368.

[22] Gordon R G Jr. Ethnologue：languages of the world, fifteenth edition. Dallas, Tex.：SIL International, 2005. Available：http：//www.ethnologue.com/.

[23] 郝思德，黄万波. 三亚落笔洞遗址. 广州：南方出版社，1998.

[24] 郝思德，王大新. 海南考古的回顾与展望. 考古，2003，4：291－299.

[25] Chen S J, Chen Y F, Yeh J I, et al. Recent Anthropological Genetic Study of Taiwan Indigenous Populations//Proceedings of the International Symposium of Anthropological Studies at Fudan University. Shanghai：Center for Anthropological Studies at Fudan University, 2002：52－60. Available：http：//comonca.org.cn/SSA/2002ISASFU/052.pdf.

[26] 李辉. 澳泰族群的遗传结构. 上海：复旦大学，人类生物学博士学位论文，2005.

[27] The Y Chromosome Consortium. A nomenclature system for the tree of human Y chromosomal binary haplogroups. Genome Res, 2002, 12：339－348.

[28] International Society of Genetic Genealogy. Y-DNA Haplogroup Tree 2007, Version：2.08 Date：31 October 2007. Available：http：//www.isogg.org/tree/.

[29] Bandelt H J, Forster P, Rohl A. Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol, 1999, 16：37－48.

[30] Zhivotovsky L A, Underhill P A, Cinnioglu C, et al. The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. Am J Hum Genet, 2004, 74：50－61.

[31] Wilson I J, Weale M E, Balding D J. Inferences from DNA data：population histories, evolutionary processes and forensic match probabilities. J R Stat Soc Ser A Stat Soc, 2003, 166：155－188.

[32] Li H, Huang Y, Mustavich L F, et al. Y chromosomes of prehistoric people along the Yangtze River. Hum Genet, 2007a, 122：383－388.

4.6 海南岛卡岱语群体仡隆人的遗传起源

按照中国的民族识别政策，大陆共有56个民族。有一些人群虽然被归入某个大型民族之中，但从遗传起源的角度进行分析，它的遗传结构会与该大型民族主体人群的遗传结构非常不一样，从而显示出与该民族主体人群不一样的遗传起源。聚居在海南岛西海岸昌化江下游（图4－14）讲“村话”的仡隆人就是一个很好的例

子[1-3]。由于大量仡隆人的家谱资料称其祖先是来自福建的闽南人，仡隆人被归为汉族，但是他们的语言（村话）属于侗傣语系下的卡岱语族仡央语支，与汉语并无关系。通过对仡隆人与现有少数民族的比较，可能在现有少数民族中找到仡隆人的遗传起源。

由于仡隆人的语言属于卡岱语族，他们可能起源于两个与他们语言相近的民族：仡佬族和黎族。黎族是海南岛的主要民族[4]，他们的语言最初被归为卡岱语族，但是最近又作为独立的一个语族从卡岱语族中分离出来[5]。尽管村话和黎语十分不同，他们仍共享某些特征。据此，20 世纪 50 年代的一些语言学家曾建议将仡隆人归为黎族。但是，由于两个族群间文化和语言的较大差异，仡隆人无法接受这一建议。除去黎族，其他使用卡岱语的人群，主要都归在仡佬族下。另外还有某些原因特殊的归类。如布央人被归入壮族（较大的侗傣语人群）；夜郎被归入瑶族（苗瑶语人群）；拉基和普标被归入彝族（藏缅语人群）。仡隆人的问题在于，包括仡佬族在内的其他所有的卡岱语人群都分布于中国西南（贵州、云南、广西和越南边境），远离海南岛（图 4－14a），大多数仡隆人甚至从未听说过仡佬族。那么，仡隆人与仡佬族是否具有遗传起源方面的相关性？

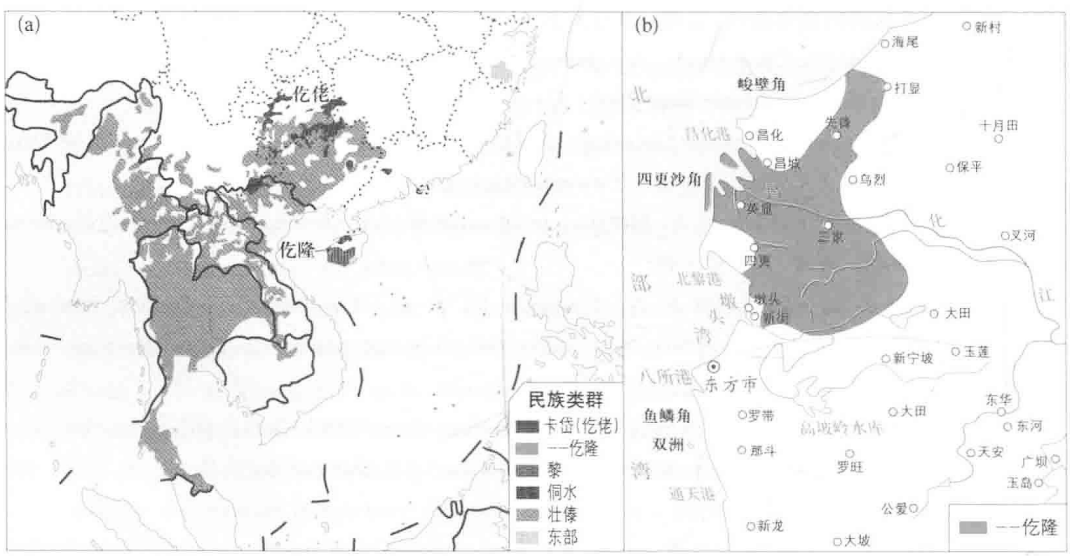


图 4－14　侗傣语系人群（a）和仡隆人的分布（b）

有趣的是，仡隆人指称男性亲属用汉语中的词汇，而指称女性亲属用村话中的词汇。有学者称，某些家谱记录的证据进一步表明，仡隆人主要由汉族的男性祖先和当地的女性祖先通婚形成[2,3]。但是从来没有确凿的证据支持仡隆人的父系祖先源于汉族。最新的遗传学证据可能有助于解决这个问题，在汉、黎、仡佬或者其他民族中寻找到仡隆人的遗传起源。

现有3种可用的遗传材料：母系线粒体DNA、父系Y染色体DNA、双谱系的常染色体和X染色体DNA。Y染色体非重组区(NRY)严格遵循父系遗传，是追寻人群父系祖先的最佳研究材料。较小的有效人群大小、低突变率、充足的单核苷酸多态(SNP)标记和短串联重复序列(STR)标记，这些特点都使得NRY成为区分和界定民族的有效工具[6-8]。本节对一仡隆人群体样本中相关的Y染色体标记进行分型，并分析仡隆人的父系起源。

4.6.1 材料和方法

1. 群体样本

2007年，在中国海南省东方市和昌江县的传统仡隆人村庄中进行了一次人口普查(表4-11)。在东方市采集了78份仡隆男性血液样本，所有样本无可查亲缘关系，三代以内祖先均为仡隆人。所有受试者均签署了知情同意书。

2. DNA抽提和Y染色体分型

DNA抽提和分型的实验方案与之前的研究相同[4]。遵照Y染色体委员会的命名原则(http://ycc.biosci.arizona.edu)[9,10]进行分型分析。Y染色体非重组区上的14个SNP(M130、M89、M9、M45、M119、M110、M101、P31、M95、M88、M122、M164、M159和M7)通过PCR-RFLP分型。4个SNP(M48、M8、M217和M356)通过Taqman(Applied Biosystems)分型。7个SNP(YAP、M15、M175、M111、M134、M117和M121)和7个STR多态性(DYS19、DYS389I、DYS389II、DYS390、DYS391、DYS392和DYS393)通过荧光标记PCR分型。变性产物通过丙烯酰胺凝胶电泳分离，使用ABI 3100测序仪区分等位基因。对每个标记进行不少于5次的个体样本重复以控制质量，重复结果未见不一致。

3. 数据分析

将仡隆人的Y染色体单倍群类型同其他群体样本的类型进行比较[4,11-15]。参照的仡佬族样本[15]取自广西隆林县，那里被认为是仡佬族扩张的起源地[16]。参照的黎族样本取自海南岛黎族的不同支系[4]。用SPSS 13.0软件进行相关性分析、主成分分析(PC)和树状聚类分析(组间聚类)。使用Network 4.510[17]画出O1a*单倍群内的STR单倍型系统树。

4.6.2 研究结果

1. 仡隆人口普查

本研究调查了位于昌化江下游的东方市和昌江县中仡隆人传统居住区的人口规模(表4-11)。仡隆人主要聚居在汉族(闽南人)和黎族移民较少的区域，他们的居住地南北邻闽南人、东接黎族，不同民族间的通婚偶有发生。

表4-11　2007年仡隆人村落的人口数

市/县乡镇	村(数)	户　数	人　口　数
东方市/县			
新街	报坡	224	1 264
	田庄	181	795
	益兴	139	558
	文通	184	1 082
	那等	415	2 409
	老官(2)	286	1 195
	平岭	196	938
	昌义	11	60
	玉章	291	1 600
四更	土地(2)	207	1 242
	长山	448	2 557
	来南	234	1 015
	赤坎(大/小)	299	1 705
	大新	362	2 181
	日新	248	1 357
	居多	234	1 248
	英显	450	2 260
	沙村	90	449
	旦场	248	1 343
	旦场园(2)	247	1 035
三家	窑上	215	931
	居侯	365	1 669
	酸梅	1 002	4 542
	小酸梅	164	735
	水流东	273	1 500
	代鸠	212	892
	红草	604	2 707
	三家	1 140	5 038
	旺老	260	1 172

（续表）

市/县乡镇	村(数)	户　数	人　口　数
	玉雄	1 492	6 331
	上荣	98	542
	下荣	75	508
昌江县			
昌化	旧县	564	3 104
	耐村	576	3 166
	大风	654	3 594
	浪炳	269	1 481
	光田	246	1 355
	黄姜	157	863
	先锋	74	409
海尾	进懂	114	624
	白沙	153	843
合计			68 299

注：村中非仡隆家庭和个体不计在内；居住在城市中的仡隆人家庭未统计在内。

也有部分仡隆人住在这些传统居住区之外，主要分布在东方市和昌江县城的郊区，仡隆人口总数约大于 8 万人。中国仡佬族的登记人口总数为 579 357(2000 年人口普查)，远多于仡隆人。

2. 仡隆人 Y 染色体 SNP 单倍群与其他人群的比较

根据 Y 染色体委员会的命名规则[10]，从 78 个仡隆人个体样本中共确定了 8 个 SNP 单倍群，其中 O1a*、O2* 和 O3* 依次为最主流的单倍群(表 4-12)。超过一半的样本被确认为 O1a*，这是侗傣语和南岛语人群的典型特征[15,18]，暗示仡隆人有清晰的侗傣遗传背景。

表 4-12　仡隆人的 Y 染色体 SNP 单倍群频率

单倍群	C3	O1a*	O2*	O2a	O3*	O3a1	O3a3c*	O3a3c1
诊断 SNP	M217	M119	P31	M95	M122	M121	M134	M117
频率(%)	5.1	57.7	17.9	3.8	9.1	1.3	3.8	1.3

本次分型所依据的几个 SNP 在以往涉及中国人群的文献中很少被报道，为了使数据更具可比性，几个从样本里区分的单倍群根据上下游关系做了归并。例如，O2*-P31 作为 K-M9 下游分支，O3a3c1-M117 作为 O3a3c-M134 下游分支。在 Y 染色体单倍

群谱系树中，O2是O的一个支系，而O又是K的一个支系。如果某个体样本实际上属于O2*，而分型时没有分析单倍群O2和O的决定位点P31和M175，只分析了M9这个位点，那么该个体样本就只能被统计在单倍群K中。合并O3a3c1也是出于这个原因。另外，因为O3a3c1和O3a3c*的频率在仡隆人、仡佬族和黎族中都非常低，所以这样做在分析中不会导致严重的偏差。最近，从汉族和仡佬族的K-M9个体中分析了P31位点，但是在汉族中却没有发现任何O2*个体。仡佬族则与仡隆人的情况类似，有单倍群O2*，但并没有K*或者O*。仡隆人和其他参照人群的各单倍群频率分布如表4-13所示。可以清楚地看到，仡佬族和仡隆人单倍群类型最接近，台湾少数民族也与他们相当类似。

表4-13　仡隆人与其他群体的Y-SNP单倍群频率比较

群　体	样本量	单倍群频率(%)											
		C	D	F	K	P*	O1a*	O1a2	O2a*	O2a1	O3*	O3a3b	O3a3c
		M130	YAP	M89	M9	M45	M119	M110	M95	M88	M122	M7	M134
仡隆人	79	5.1			17.9		57.7		3.8		10.4		5.1
仡佬族	30				16.7		60.0		16.7		3.3		3.3
黎族	11						27.3		54.6	9.1			9.1
泰雅	24						54.2	8.3			29.2	4.2	4.2
排湾	11						54.6	27.3					18.2
布依族	49	7.0		4.4	17.7		4.4		46.7	11.1	4.4	2.2	2.2
侗族	10	20.0					20.0	10.0	20.0		10.0		20.0
水族	92				5.4		31.5		58.7				4.4
苗族	49	8.2	2.1		28.6		12.3		16.3	2.1	24.5	4.1	2.1
布努人	10	20.0	30.0						40.0		10.0		
壮族	28	3.6	3.6	7.1	3.6		17.9		25.0	10.7	3.6		25.0
汉族-江西	21	4.8	4.8	9.5	19.1		14.3		4.8		19.1		23.8
汉族-福建	13	7.7			7.7					7.7	38.5		38.5
汉族-海南	15	6.7			6.7		6.7		13.3		33.3		33.3
汉族-河南	28	7.1		3.6	25.0	7.1	10.7				32.1		14.3
泰国	20			5.0	5.0	5.0	5.0	5.0	45.0	20.0	5.0	5.0	
柬埔寨	26	3.8	3.8	11.5	11.5	7.7	3.8	3.8	23.1	11.5	3.8		15.4
马来西亚	13			7.7	7.7		7.7	23.1	7.7		30.8		15.4

3. SNP单倍群类型的相关性分析

为了估计人群样本单倍群类型的相似度，笔者进行了群体间相关性分析(表4-14)。

仡隆人和仡佬族的单倍群类型显著相关($r=0.962, P<0.001$),暗示这两者几乎有相同的Y染色体SNP类型。台湾少数民族也与仡隆人显著相关。仡隆人和其他被研究人群之间未发现显著相关性,包括仡隆的邻近人群黎族。与仡隆人相关度最弱的是福建汉族和广西瑶族。

表4-14 仡隆人与其他群体的Y染色体单倍群频率的相关性分析

配　对	相关系数 r	P 值
仡隆人-仡佬族	0.962	<0.001
仡隆人-黎族	0.329	0.296
仡隆人-壮族	0.321	0.309
仡隆人-布依族	0.003	0.993
仡隆人-侗族	0.414	0.180
仡隆人-水族	0.401	0.196
仡隆人-泰雅	0.866	<0.001
仡隆人-排湾	0.776	0.003
仡隆人-汉族江西	0.443	0.149
仡隆人-汉族福建	−0.023	0.944
仡隆人-苗族	0.417	0.177
仡隆人-瑶族	−0.187	0.561

4. SNP单倍群类型的主成分分析

主成分分析能够根据频率数据揭示人群之间的关系。本研究的Y染色体区段不发生重组,所以Y染色体可以被看作是单基因座。因主成分分析不能处理单基因座上的个体数据,因此我们用群体样本的单倍群频率数据来做分析。按前四个成分做出的人群分布图如图4-15所示。前两个主成分解释了61.9%的数据方差,前四个主成分总共能够解释86.6%的数据方差。每个主成分都与不同的民族群体相关。例如,第一主成分与侗傣人群,第二主成分与台湾少数民族,第三主成分与汉族,第四主成分与苗瑶人群相关。两个图中,仡隆人和仡佬族都非常接近,尤其是在第三和第四主成分绘制的图中,两个人群几乎重叠。但在第一、第二主成分所做的图中,仡隆人与台湾少数民族(排湾族和泰雅族)和江西汉族也很接近。因此,多重分析比较之后得出的结论更为可靠。

5. 人群聚类树分析

聚类树分析有别于主成分分析,能够从另一个角度揭示人群关系,给出整体框架。用相同的数据进行聚类树分析共聚成四大类(图4-16)。侗傣人群组成的一个聚类群包括水、黎、布依和泰国民族,另两个侗傣人群壮、侗与高棉人合为一群,汉、苗与马来人合为一群。最后一群包括仡隆人、仡佬族和两个台湾民族。图中仡隆人和仡佬族聚在一起,这与相关性分析、主成分分析得出的结果一致。

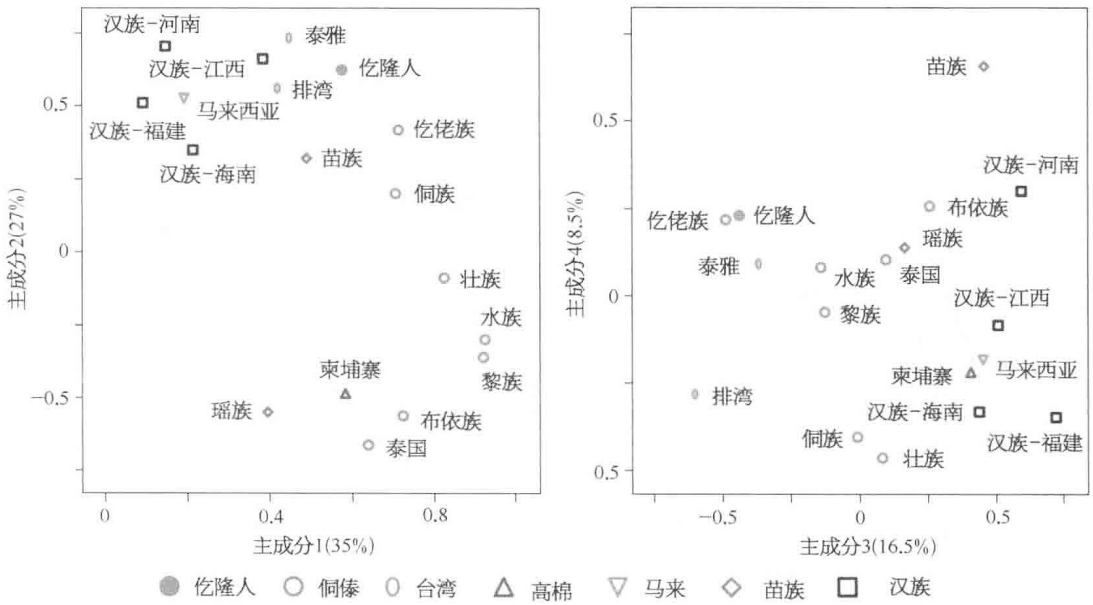


图 4-15　相关群体的主成分分析

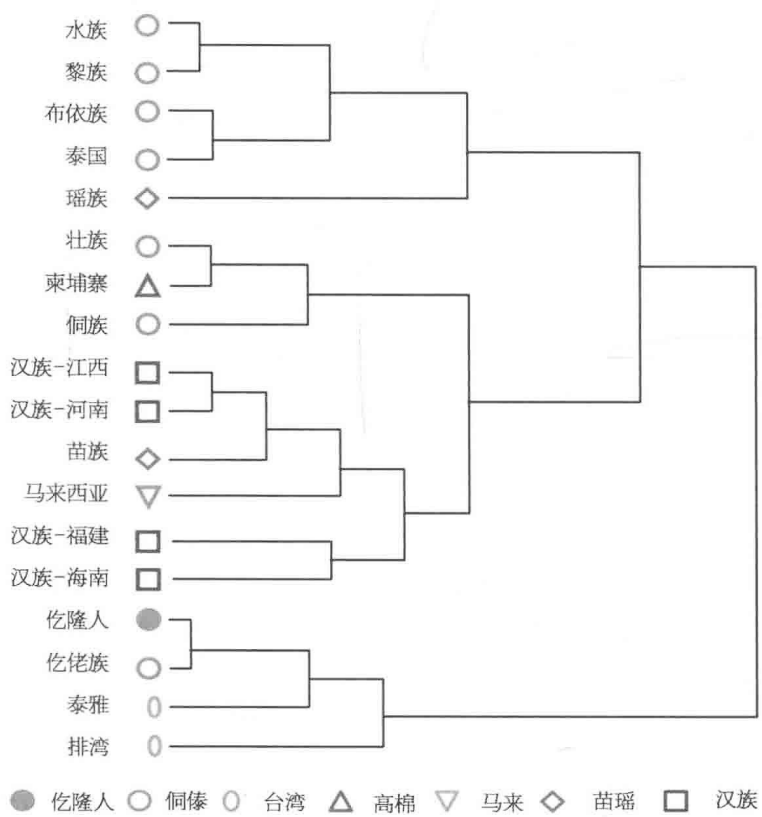


图 4-16　根据 Y-SNP 单倍群频率绘制的群体聚类树

6. STR 多样性和网络图

7 个 STR 多态位点的分型结果和相应 SNP 单倍群的整合信息如表 4-15 所示。不同 SNP 单倍群的 STR 单倍型均不相同，这表明 7 个 STR 多态已经提供了足够多的单倍型多样性信息。STR 单倍型在群体样本中并不是等同分布的，O1a * 和 O2 * 下有两种单倍型的总和超过了样本的一半。这两个 SNP-STR 单倍型可能是仡隆人的起始单倍型，需要进一步详细分析。近来在人群样本中分析了 O2 * -P31，并在仡佬、黎和其他侗傣人群中发现突变型的等位基因，为进一步分析提供了比较数据。

表 4-15 仡隆人 STR 单倍群

编号	样本量	DYS19	DYS389-1	DYS389-2	DYS390	DYS391	DYS392	DYS393	诊断 SNP	单倍群
Hap01	1	15	12	28	26	10	14	12	M117	O3a3c1
Hap02	28	15	13	28	23	10	14	13	M119	O1a *
Hap03	4	15	13	29	23	10	14	13	M119	O1a *
Hap04	1	15	14	29	23	10	14	13	M119	O1a *
Hap05	2	15	14	30	23	10	14	13	M119	O1a *
Hap06	2	16	12	27	23	10	12	13	M134	O3a3c *
Hap07	1	16	12	28	22	9	14	13	M119	O1a *
Hap08	4	16	12	28	23	10	14	13	M119	O1a *
Hap09	2	16	12	28	25	11	15	13	M119	O1a *
Hap10	1	16	12	29	23	10	12	12	M134	O3a3c *
Hap11	3	16	12	29	23	11	13	13	M119	O1a *
Hap12	1	16	13	28	25	10	13	12	M122	O3 *
Hap13	1	16	13	29	23	10	14	13	M119	O1a *
Hap14	1	16	13	29	26	11	13	14	M95	O2a *
Hap15	1	16	13	30	25	10	11	13	M217	C3
Hap16	1	16	13	30	24	11	11	14	M217	C3
Hap17	1	16	13	31	23	11	13	12	M121	O3a1
Hap18	1	16	14	30	23	11	11	14	M217	C3
Hap19	11	16	14	30	24	10	13	14	P31	O2 *
Hap20	2	16	14	30	24	11	13	14	P31	O2 *
Hap21	1	17	13	29	25	10	13	14	M95	O2a *
Hap22	4	17	12	27	25	10	13	12	M122	O3 *

（续表）

编号	样本量	DYS19	DYS389 - 1	DYS389 - 2	DYS390	DYS391	DYS392	DYS393	诊断 SNP	单倍群
Hap23	2	17	12	28	25	11	13	12	M122	O3 *
Hap24	1	17	13	29	23	10	11	15	M217	C3
Hap25	1	17	13	30	24	11	13	14	M95	O2a *

我们使用 Network 程序构建了两株最短网络树（图 4 - 17）。这些树显示了 O2 * 和 O1a * 单倍群下的 STR 单倍型之间的最可能关系。在 O1a * 树中，大多数的黎族单倍型在远离主体单倍型的分支中。有趣的是，仡隆的最大节点（Hap02）介于黎族分支和主体分支之中，连接了黎族和卡岱族群。在 O2 * 树中，卡岱和黎的节点分散在不同支系中，仡隆最大节点（Hap19）介于卡岱族群和水族（侗傣分支中）节点之间。在两株树中，仡隆节点都与卡岱族群的节点较为接近。我们紧接着统计了仡隆和其他侗傣人群分支间共享和邻近的单倍型（表 4 - 16）。由表中可知，仡隆个体中有 48 个拥有与其他卡岱族群相似的单倍型，与黎族相似的较少，但很少有和其他分支类似的。仡隆和黎之间的相似性也不能被忽视，毕竟这两个族群千百年来相邻而存，之间必然存在着基因交流。总之，相比于其他侗傣人群，仡隆与卡岱最为接近。由于两个族群地理上相距甚远，这种相似性无法简单地用近期的基因交流解释。因此，仡隆人和其他卡岱族群在分散到不同地区之前可能有着共同的祖先起源。

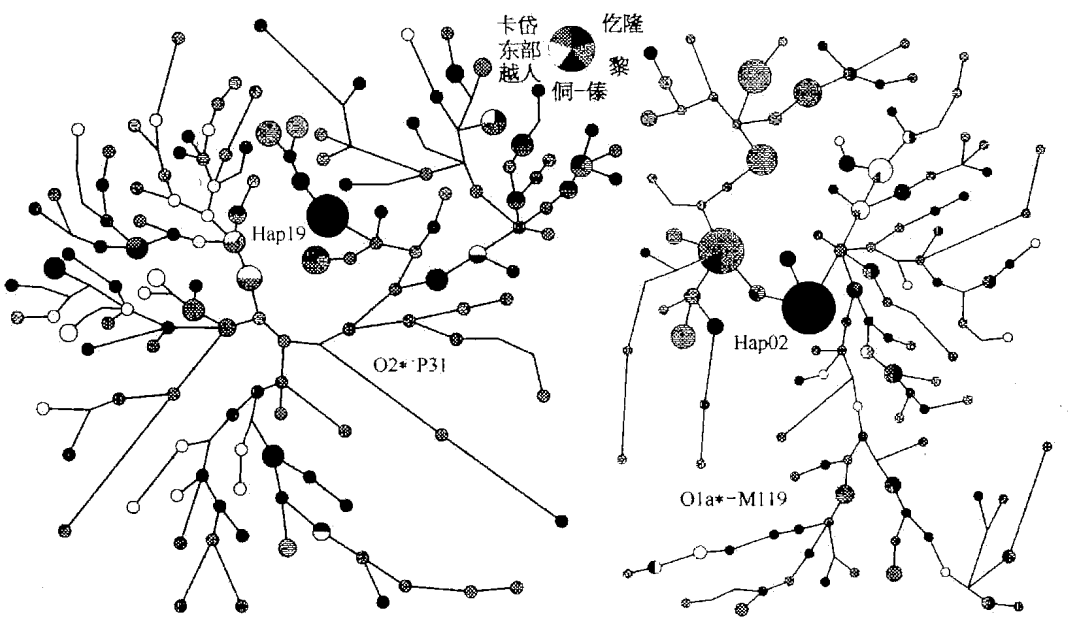


图 4 - 17　STR 单倍型的网络结构最短树

节点间的连线长度等比于突变步数。

表 4-16 网络结构中仡隆人和其他侗傣群体之间 STR 单倍型的相似度统计

	共享单倍型		邻接单倍型	
	O1a*	O2*	O1a*	O2*
卡岱	1	1	31	15
侗水	1	0	2	1
壮傣	0	0	0	0
东越	1	0	1	2
黎	7	0	35	0

4.6.3 讨论

1. 近期基因交流抑或共同祖先起源

仡隆人的 Y 染色体结构与黎族、卡岱族群(仡佬族)两个人群都很接近,尽管仡隆人目前民族身份定为汉族,但是他们的遗传结构与汉族并不相近。汉族的主流单倍群 O3a3c*-M134 和 O3a3c1-M117[12,19,20]在仡隆人中只占 5%,而且这个比例很可能来源于与近期汉族移民的基因交流,这种情况与 O3a3c 在黎族[4]和中国许多其他少数民族中的比例相似[21]。尽管仡隆人对父系亲属有一套完整的汉语称谓,但是他们似乎并不具有明显的汉族父系起源。

两个人群间的遗传相似性可能源于不同的人群历史事件,例如,邻近人群的近期基因交流,或者是两个人群源自同一祖先,甚至随机的遗传漂变也能导致该结果。由于 Y 染色体标记系统的复杂性,由随机的遗传漂变导致的概率极小,那么只剩下近期基因交流和共同祖先起源可以解释仡隆人和相关群体的遗传相似性。近期基因交流只有在改变大多数成员遗传比例的前提下,才有可能改变人群内部组分的主从关系。另一方面,如果两个人群源于同一祖先,他们的遗传结构就会很相似,被分到一组也是很合理的。

根据研究结果,仡隆人的祖先起源必然是黎族和仡佬族之一,或者两者都是。中国的一些语言学学者认为他们的语言应该是黎语的一支,因为两者之间共享 42.1%的词汇[2,3]。一些民族学家在 20 世纪 50 年代曾建议将仡隆人登记为黎族。但是仡隆人自己认为同黎族之间存在巨大差异,并不认同这个民族身份。本研究显示出仡隆和黎族之间的遗传相似性。经过千百年的邻居历史,两个人群间的基因交流必然产生了强烈影响。尽管如此,相比于黎族,仡隆人仍然更接近于仡佬族。因此,虽然部分仡隆人有黎族渊源,但他们之间的遗传相似性并不能作为黎族起源的证据。

与此相反,由于地理隔离,仡隆人和仡佬族之间的遗传相似性不可能来源于近期基因交流。仡佬族和仡隆人生活的区域至少相隔 600 km,其间隔着海峡、山脉以及与仡隆

人截然不同的其他侗傣人群。至今并无证据支持两个群体间存在近期基因交流。因此，共同祖先起源是仡隆人和仡佬族之间遗传相似性的最合理解释。另外，有语言学证据支持他们的相似性，将两者的语言均划归于卡岱语族[1,16]。综上，仡隆人应该是仡佬族的远房亲戚，他们共享一个祖先起源。

2. 文化还是遗传

在界定不同人群间的相互关系时，到底是文化还是遗传背景占有更重要的位置，这场争论在中国已经持续了几十年。但是在仡隆人的例子中，文化（语言[1]和葬俗[22]）和遗传证据均支持他们同仡佬族同源之间的关系。他们还有着相近的族名。

另一方面，仡隆人和黎族的文化很少有相同的方面，尽管他们都生活在海南岛上。海南岛的考古研究发现，仡隆人在6 000年前可能就到达了他们现今的家园，留下了海南岛最早的新石器时代遗址——东方新街贝丘遗址[23]。该遗址发掘的文化同黎族所在地区发现的考古文化截然不同。遗传学估计，黎族已经在海南岛独自生活了一两万年[4]，因此晚近来到的仡隆人必然带来了一种全新的文化。仡隆人和仡佬族分开并相互隔离生活了至少6 000年，因此文化和遗传上的差异不能被忽略。但是，我们仍然能清楚地看到两者之间的相似性。这可能跟这两个族群相对隔离的生活状态有关。仡佬族生活在深山中，而仡隆人生活在海岛一角，相比其他侗傣人群，他们都不易被外来人群所影响。

参考文献

[1] Gordon R G Jr Ethnologue. Languages of the world. Fifteenth edition. SIL International, Dallas, Texas, 2005.

[2] 符昌忠. 海南村话. 广州：华南理工大学出版社，1996.

[3] 欧阳觉亚. 村语研究. 上海：上海远东出版社，1998.

[4] Li D, Li H, Ou C, et al. Paternal genetic structure of Hainan aborigines isolated at the entrance to East Asia. PLoS One, 2008, 3: e2168.

[5] Burusphat S. A comparison of general classifiers in Tai-Kadai languages. Mon-Khmer studies: a journal of Southeast Asian languages and cultures, 2007, 37: 129－153.

[6] Jobling M A, Tyler-Smith C. Father and sons: the Y chromosome and human evolution. Trends Genet, 1995, 11: 449－456.

[7] Jobling M A, Tyler-Smith C. New uses for new haplotypesthe human Y chromosome, disease and selection. Trends Genet, 2000, 16: 356－362.

[8] Underhill P A, Shen P, Lin A A, et al. Y chromosome sequence variation and the history of human populations. Nat Genet, 2000, 26: 358－361.

[9] The Y Chromosome Consortium. A nomenclature system for the tree of human Y chromosomal binary haplogroups. Genome Res, 2002,12, 339－348.

[10] Karafet T M, Mendez F L, Meilerman M B, et al. New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. Genome Res, 2008, 18, 830－

838.

[11] Su B, Xiao J, Underhill P, et al. Y chromosome evidence for a northward migration of modern humans in East Asia during the Last Ice Age. Am J Hum Genet, 1999, 65: 1718 - 1724.

[12] Ke Y, Su B, Xiao J, et al. Y chromosome haplotype distribution in Han Chinese populations and modern human origin in East Asians. Sci China C Life Sci, 2001, 44: 225 - 232.

[13] 李永念,左丽,文波,等.中国布依族人的起源及迁移初探——来自Y染色体和线粒体的线索.遗传学报,2002,29: 196 - 200.

[14] 何燕,文波,单可人,等.贵州三都水族Y染色体单倍型频率分析.遗传,2003,25: 249 - 252.

[15] Li H, Wen B, Chen S J, et al. Paternal genetic affinity between Western Austronesians and Daic populations. BMCEvol Biol, 2008, 8: 146.

[16] 李锦芳,周国炎.仡央语言探索.北京:中央民族大学出版社,1999.

[17] Bandelt H J, Forster P, Rohl A. Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol, 1999, 16: 37 - 48.

[18] Li H, Huang Y, Mustavich L F, et al. Y chromosomes of Prehistoric People along the Yangtze River. Hum Genet, 2007, 122: 383 - 388.

[19] Wen B, Li H, Lu D, et al. Genetic evidence supports demic diffusion of Han culture. Nature, 2004, 431: 302 - 305.

[20] Gan R J, Pan S L, Mustavich L F, et al. Pinghua population as an exception of Han Chinese's coherent genetic structure. J Hum Genet, 2008, 53: 303 - 313.

[21] Shi H, Dong Y L, Wen B, et al. Y chromosome evidence of southern origin of the East Asian specific haplogroup O3 - M122. Am J Hum Genet, 2005, 77: 408 - 419.

[22] 李辉,李冬娜.海南仡隆人院子里的墓地.现代人类学通讯,2010,4: 1 - 4.

[23] 郝思德,王大新.海南考古的回顾与展望.考古,2003,4: 291 - 299.

4.7 海南岛的占城回辉人世系被本土成分替换

4.7.1 研究背景

回辉人人口相对较少,主要生活在中国大陆南部的海南岛最南端。虽然他们有独特的语言,但并不被认为是一个单独的民族,而是划入回族。他们的起源并不明确。回辉人的民间故事认为他们的祖先一直是穆斯林,起源于中亚,一如中国其他回族。但回辉人同时也被认为是占城王国(7—18世纪)为躲避越南入侵而流亡的占城人后裔[1]。根据口述历史材料,占城王子和大约1 000占城人在越南占领占城后迁到了海南,并获得明朝政府允许在海南建立了流亡政权[2]。据中国历史文献记载,早在宋代,占城的首都于公元982年陷落之后,占城难民就开始流落海南[3]。

回辉人的语言回辉话和占语一样,同属于南岛语系下的马来-波利尼西亚语族[4,5]。回辉话与北拉格莱语十分相像,并被归为占北语支[6]。但是,回辉话在海南岛上却犹如一块"语言飞地",其周围均非南岛语系(如侗傣语系和汉语)。由于长期接触汉语和黎语,以及有方向性的内部漂变,回辉话从结构上变得类似汉语和黎语。例如,回辉话发展出了马来-波利尼西亚语少有的固定声调[7]。

在过往研究中，笔者发现东亚的语言与Y染色体父系谱系有着很强的关联[8-11]。因此，回辉话结构上的改变也有可能反映在遗传上。笔者在2008年报道过31个回辉人样本的Y染色体数据，其中高频出现的O1a-M119(58.1%)和Y-STR网络结构中显示的其与侗傣人群的联系，表明回辉人很可能有侗傣遗传背景[10]。这些结果说明，回辉人的起源很可能伴随着对原住民的同化，或者近期的基因交流。但是，Y染色体数据只能从父系角度提供证据，而且研究所涉及的样本量较小，也可能导致偏差。此外，缺乏同占城人的数据进行比较，使回辉人起源问题仍存在争议。为了解决该问题，在本研究中，笔者对102个回辉人样本(72男和30女)进行母系遗传的线粒体DNA和相关Y染色体标记的分型，以期对回辉人的起源有更深入的理解。

4.7.2　材料和方法

1. 群体样本

该研究获得了复旦大学生命科学学院伦理委员会的批准。102份回辉人外周血样本均采自海南三亚。受试者均获得了相关研究的充分信息，并签署了知情同意书。所有研究对象均健康，5代以内无可查亲缘关系。

2. Y染色体标记

根据最新的Y染色体谱系树[12,13]，对样本在Y染色体非重组区上的14个单核苷酸多态(SNP)(M130、M89、M9、M45、M119、M110、M101、P31、M95、M88、M122、M164、M159和M7)使用聚合酶链反应-限制性片段长度多态分析进行分型。4个SNP(M48、M8、M217和M356)使用Taqman(Applied Biosystems)进行分型。7个STR多态(DYS19、DYS389I、DYS389II、DYS390、DYS391、DYS392和DYS393)用荧光标记引物PCR进行分型。变性产物用丙烯酰胺凝胶电泳分离，在ABI 3730xl测序仪上区分等位基因。

3. 线粒体DNA标记

线粒体高变1区(HVS-1)使用引物L15974和R16488进行扩增[14]。PCR产物经虾碱酶和外切酶(Roche Diagnostics, Shanghai)纯化后，使用Big-Dye Terminator Cycle Sequencing试剂盒(Applied Biosystems)进行测序反应。用软件Sequence Analysis 3.3(Applied Biosystems)读取序列。根据修订的剑桥标准序列[15]，使用软件DNASTAR(DNASTAR, Madison)对HVS-1序列进行编辑和排列。编码区上的22个多态(3010、7598、663、10400、10310、4216、4491、12308、10646、11719、4715、4833、8271、5301、70287、13263、14569、5417、5178、12705、15607和9824)根据谱系使用SNaPshot(ABI SNaPshot Multiplex Kit; Applied Biosystems)进行分型。PCR产物也在ABI 3730xl测序仪上电泳分离。每段线粒体DNA的单倍群谱系关系根据HVS-1基序和编码区多态综合分析推断得出[16,17]。

4. 统计分析

Y染色体STR和线粒体DNA的HVS-1基序的网络结构根据中点连接法[18]，使

用软件 Network version 4.510 (http://www.Fluxus-engineering.com)构建。回辉人基因型数据由本次研究得到，其他邻近人群数据来源于已有文献[8,9,14,19-27]。使用 Arlequin 3.11 计算 Y-STR 的 Slatkin 线性 F_{ST}(R_{ST})遗传距离[28]。使用 SPSS 18.0 软件进行主成分分析(PCA)和多维尺度分析(MDS)。

4.7.3 结果和讨论

1. Y 染色体

根据国际 Y 染色体命名委员会的命名规则[12,13]，从 72 个回辉人个体样本中共确定了 8 个 SNP 单倍群。尽管回辉人的语言被归为占北语，但在他们的父系遗传结构中单倍群 O1a*-M119 占高频，这与占城人并不相似。而在占城人中占主流的 O2a1* 和其下游单倍群 O2a1a 只占回辉人的 4.17%(图 4-18)。在回辉人中频率中等的古老东南亚支系 C-M130 和 F*-M89，可能源于特定祖先贡献之后发生的遗传漂变。另外，发现了汉藏的典型支系 O3a2c1a-M117[29,30]在回辉人中低频存在，占 4.17%。这可能源于汉族移民的近期基因交流。

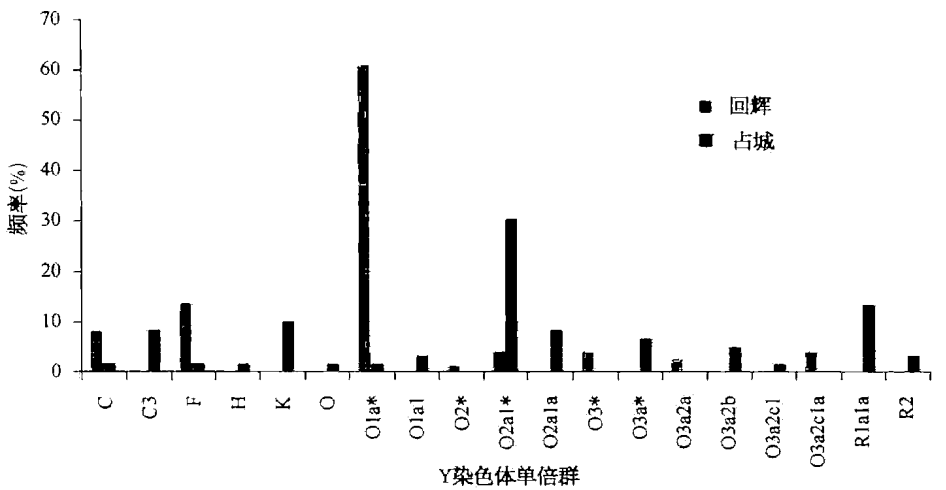


图 4-18 回辉人和占城人的 Y 染色体单倍群频率

使用其他已发表的 Y 染色体数据，可以比较分析回辉人、占城人和东亚其他人群的详细父系遗传结构类型。对回辉人和其他 43 个东亚人群的 Y 染色体频率数据进行主成分分析(图 4-19)，发现在第二主成分上，来自印度支那半岛的人群和来自海南岛的人群分别聚类成两组。其中，回辉人位于海南组，和海南原住民、中国南部人群侗水等聚在一起。而占城人与印度支那组非常接近。

基于 52 个人群在 6 个通用 Y-STR 位点(DYS19、DYS389I、DYS390、DYS391、DYS392 和 DYS393)上的遗传距离 R_{ST}所做的 MDS 图也显示，回辉人与海南岛人群较为接近(图 4-20)。形成这种模式的主要原因是回辉人中高频的 O1a*-M119 单倍群和低

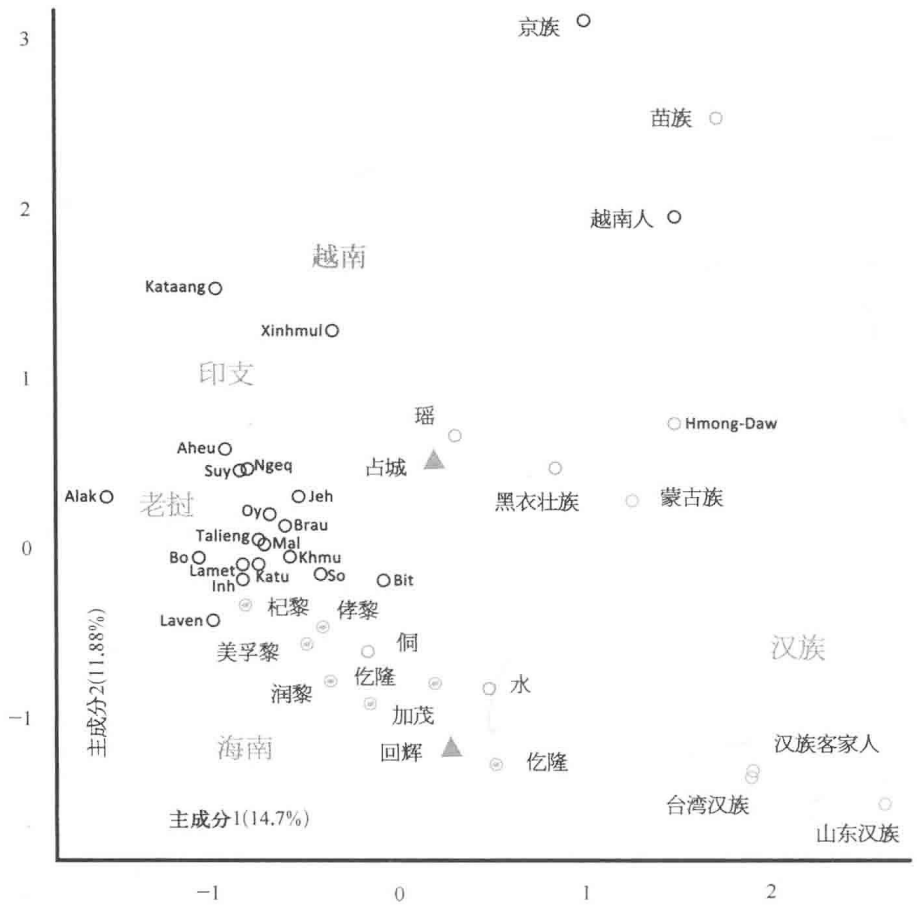


图 4－19　44 个群体 Y 染色体单倍群的主成分分析图

频的 O2a1 *－M95 单倍群。

回辉人、海南原住民和印度支那人群在 O1a *－M119 单倍群上的具体距离，能够清楚地揭示回辉人的主流父系遗传起源。因此，研究中找到相关人群中的 O1a *－M119 个体，在 6 个 STR 多态位点(DYS19、DYS389I、DYS390、DYS391、DYS392 和 DYS393)上构建中点连接网络结构(图 4－21)。图中所示，海南原住民除了少数散布在其他人群中，其他均形成了若干个几近孤立的分支，这表明海南原住民早已隔离于中国南部的其他侗傣人群和台湾少数民族。几乎所有的回辉人样本都聚类于海南原住民的孤立分支中，而印度支那的样本倾向于和中国南部聚类到一起。这些结果说明，回辉人的主要父系单倍群来源于海南本土的民族群体，而不是占城人或者其他印度支那人群。

2. 线粒体 DNA

目前从 102 个回辉人样本中共发现 19 个线粒体 DNA 单倍群，其中较为高频的是 D4、F2a、F1b、F1a1、B5a、M8a、M *、D5 和 B4a(按降序排列)。D4 和 F2a 为回辉人的两个主要单倍群，分别占 16.67%和 15.69%，但是这两个单倍群在其他海南原住民和印度

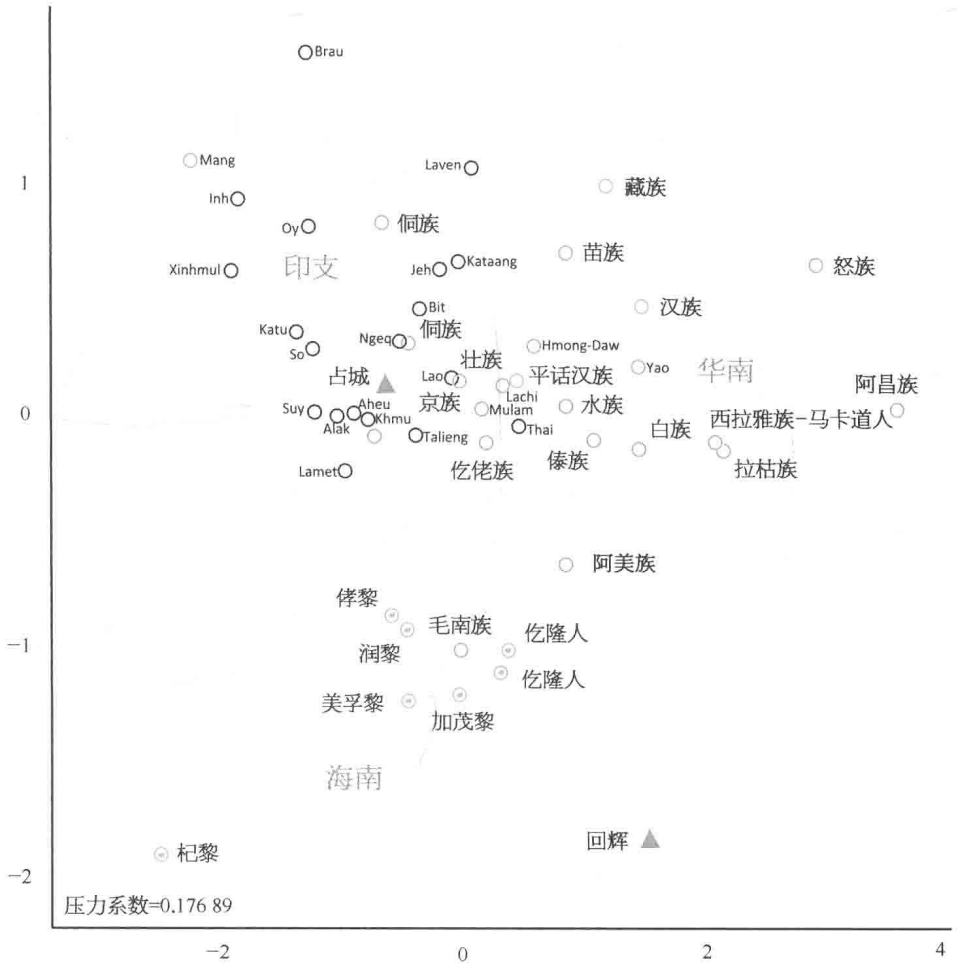


图 4-20 基于 6 个通用 Y-STR（DYS19、DYS389I、DYS390、DYS391、DYS392 和 DYS393）的 52 个人群 MDS 聚类图

支那人群中没有发现或者低频出现。在单倍型水平上将回辉人的这两个单倍群同其他相关人群进行比较，发现大部分回辉人的 D4 样本共享几个相同的 HVS-1 基序位点，为 16223、16316 和 16362，但是这种单倍型在东亚和印度支那人群中较为罕见。回辉人的 F2a 单倍型仅仅在部分汉族和云南的一些小型人群（拉祜族、彝族、摩梭人）中发现[8,9,19,27,31]。回辉人这种 B 和 F 单倍群占高频的分布模式，与邻近人群和南方其他人群十分相似。此外，用回辉人和其他所有 30 个人群的线粒体 DNA 单倍群分布频率做 PCA 分析（图 4-22），发现台湾少数民族、印度支那人群和海南原住民在第一主成分上形成 3 个聚类。单倍群 E、F5 和 B4 为台湾一组的主要类型，而单倍群 G、A、C、M9 和 M8 为海南和汉藏一组的主要贡献。回辉人倾向于和海南原住民聚在一起，而占城人则和印度支那人群聚为一类。综上，频率分布模式揭示了回辉人和海南原住民的遗传相似性。

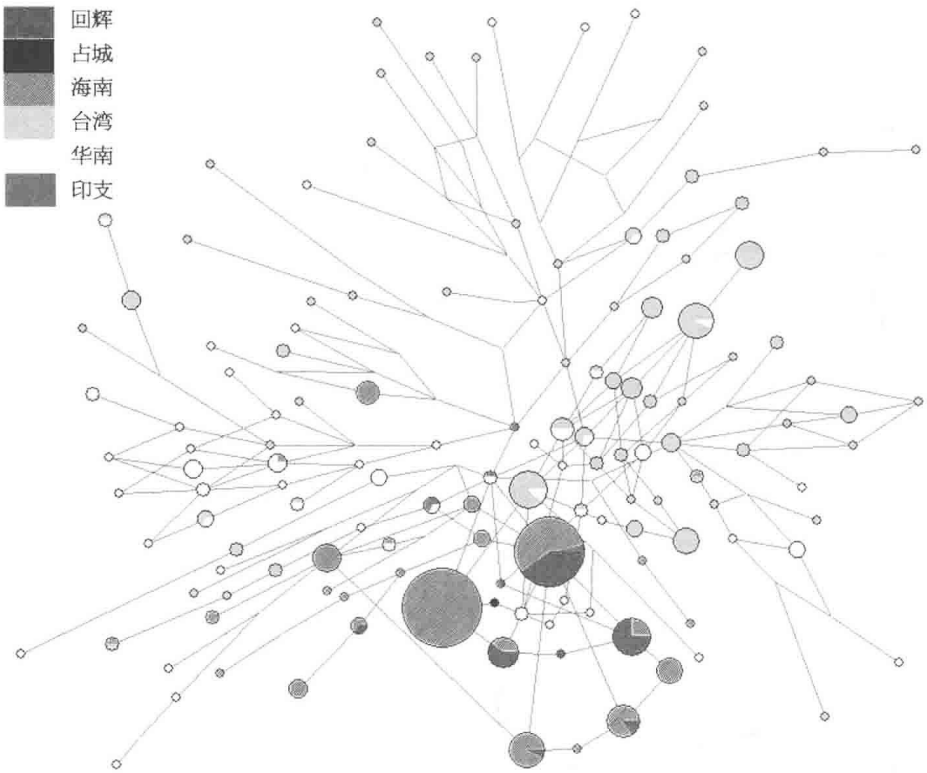


图 4-21　根据 O1a*-M119 内部 6 个 STR 单倍型构建的中点连接网络结构
节点之间的连线长度等比于突变步数。

然而，由于正向选择或遗传漂变造成线粒体 DNA 谱系变化频率较高，仅仅基于单倍群频率比较得出的结果可能存在误导[27,32]。而基于个体谱系的网络结构分析能够更好地提供一个关于回辉人、占城人和其他人群的关系[23,25]。根据线粒体 DNA 的 HVS-1 基序和 SNP 确定的单倍群，构建线粒体 DNA 单倍群 D4、F2a、F1b、F1a1、M8a、D5 和 B4a 的网络结构(图 4-23)。这些线粒体 DNA 单倍群在回辉人和印度支那人群中均占高频或中频，总和占回辉人总数的 72.55%。在 D4 单倍群网络结构中，回辉人只有一个单倍型和泰族人共享，其他形成了一个较大的独立分支；在 F2a 和 F1b 中也能见到独立分支。在 B4a 和 M8a 图中，回辉人只和海南原住民的样本聚在一起。在 F1a1 图中，回辉人只和中国大陆南部人群聚为一类。在 B5a 和 D5 图中，回辉人、海南原住民和来自中国大陆南部、中国台湾岛、印度支那的人群一起，聚在较大或中等大小的分支中。但无一回辉人样本直接和占城人聚为一类。总体而言，回辉人的母系遗传谱系与海南和中国南部的族群更为接近，而不是印度支那人群。

在本研究中，Y 染色体父系遗传谱系和线粒体 DNA 母系遗传多样性表明，相比于占城人和其他印度支那人群，回辉人与海南原住民最为接近。这说明回辉人的形成过程中伴随着对原住民的大量同化。在同化过程中，回辉人的语言从结构类型上变得更像汉语

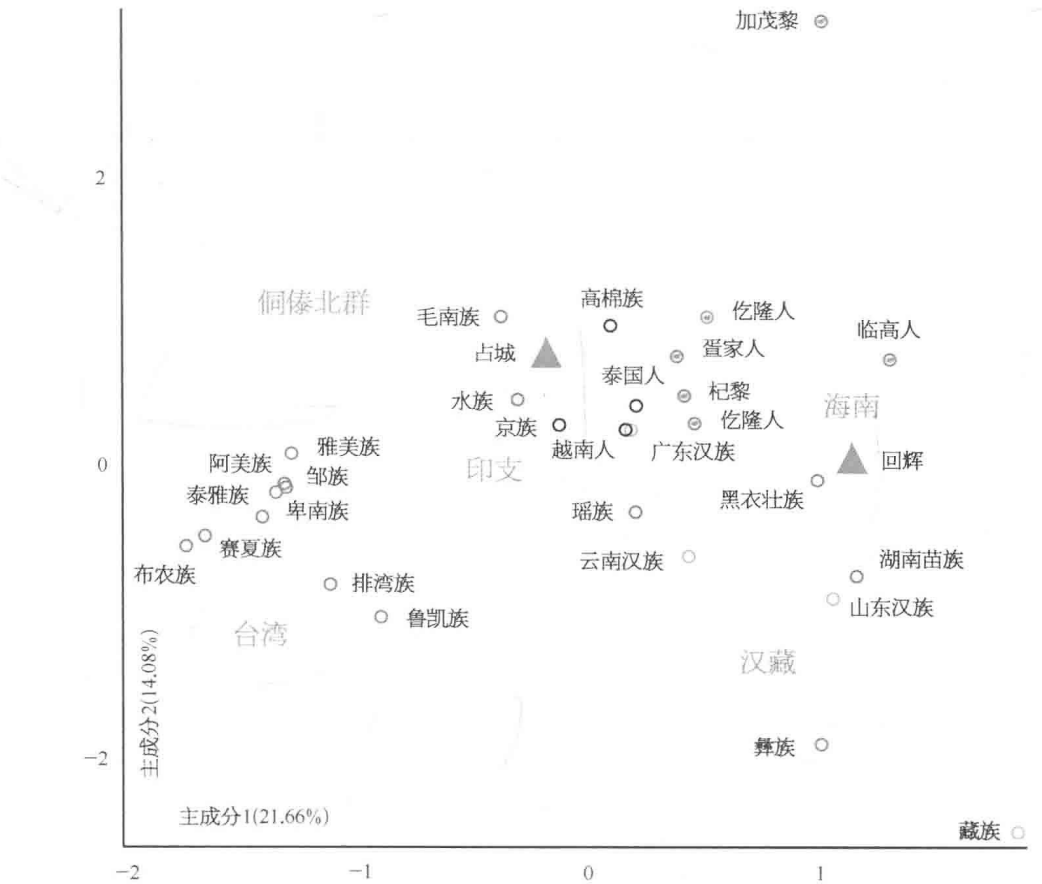


图 4－22 东亚 31 个群体线粒体单倍群频率的主成分分析图

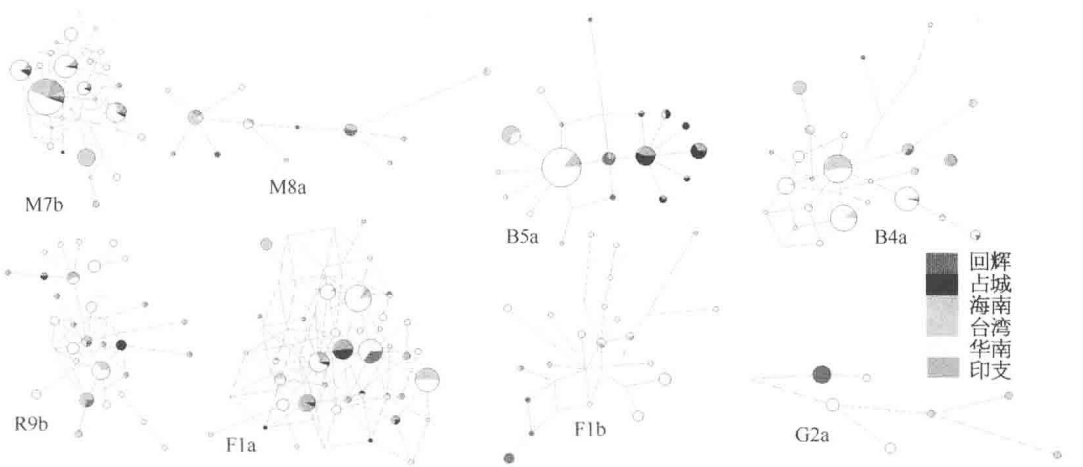


图 4－23 线粒体 DNA 的 HVS－I 序列单倍群网络结构

节点之间的连线长度等比于突变步数。

或侗傣语系的语种。然而最有意思的是，回辉人的文化和自我认同仍然保留了占城人的传统。作为穆斯林的回辉人依照伊斯兰教义和原则来处理日常生活中的各种事务，比如卫生、饮食、斋戒，甚至是祷告的确切时间。这些伊斯兰信仰对保存他们的生活方式和自我认同起着至关重要的作用，这种作用更多的是在社群意义层面上，而不是生物学层面。我们可以称之为遗传替换的"宗教决定"机制：一小群迁入移民被当地原住民接纳后，在遗传成分上被当地人群替换，但是这一小群移民带来的宗教信仰却使他们保存了他们根植于宗教的文化传统和自我认同。

参考文献

[1] Olson J S. An ethnohistorical dictionary of China. Westport: Greenwood Publishing Group, 1998.

[2] Tran N T, Reid A. Viet Nam: borderless histories. Wisconsin: Univ of Wisconsin Press, 2006.

[3] Andaya L Y. Leaves of the same tree: trade and ethnicity in the Straits of Melaka. Honolulu: University of Hawaii Press, 2008.

[4] Pang K F, Ian M. Tone in Utsat//Edmondson J A, Gregerson K J. Tonality in Austronesian languages. Oceanic Linguistics Special Publication 24. Honolulu: University of Hawaii Press, 1993.

[5] Graham T. Sociolinguistics and contact-induced language change: Hainan Cham, Anong, and Phan Rang Cham. In: Tenth International Conference on Austronesian Linguistics, 17 - 20 Jan, 2006. Palawan: Linguistic Society of the Philippines and SIL International, 2006.

[6] Lewis M P. Ethnologue: languages of the world, sixteenth edition. Dallas: SIL International, 2009.

[7] Graham T. Language contact and the directionality of internal drift: the development of tones and registers in chamic. Language, 1996, 72: 1 - 31.

[8] Wen B, Li H, Lu D, et al. Genetic evidence supports demic diffusion of Han culture. Nature, 2004, 431: 302 - 305.

[9] Wen B, Xie X, Gao S, et al. Analyses of genetic structure of Tibeto-Burman populations reveals sex-biased admixture in southern Tibeto-Burmans. American Journal of Human Genetics, 2004, 74: 856 - 865.

[10] Li H, Wen B, Chen S J, et al. Paternal genetic affinity between Western Austronesians and Daic populations. BMC Evolutionary Biology, 2008, 8: 146.

[11] Cai X, Qin Z, Wen B, et al. Human migration through bottlenecks from Southeast Asia into East Asia during Last Glacial Maximum revealed by Y chromosomes. PLoS One, 2011, 6: e24282.

[12] Karafet T M, Mendez F L, Meilerman M B, et al. New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. Genome Research, 2008, 18: 830 - 838.

[13] Yan S, Wang C-C, Li H, et al. An updated tree of Y chromosome Haplogroup O and revised phylogenetic positions of mutations P164 and PK4. European Journal of Human Genetics, 2011,

19: 1013 - 1015.

[14] Yao Y G, Kong Q P, Bandelt H J, et al. Phylogeographic differentiation of mitochondrial DNA in Han Chinese. American Journal of Human Genetics, 2002, 70: 635 - 651.

[15] Andrews R M, Kubacka I, Chinnery P F, et al. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nature Genetics, 1999, 23: 147.

[16] Kivisild T, Tolk H V, Parik J, et al. The emerging limbs and twigs of the East Asian mtDNA tree. Molecular Biology and Evolution, 2002, 19: 1737 - 1751.

[17] Kong Q P, Yao Y G, Sun C, et al. Phylogeny of East Asian mitochondrial DNA lineages inferred from complete sequences. American Journal of Human Genetics, 2003, 73: 671 - 676.

[18] Bandelt H J, Forster P, Rohl A. Median-joining networks for inferring intraspecific phylogenies. Molecular Biology and Evolution, 1999, 16: 37 - 48.

[19] Yao Y G, Zhang Y P. Phylogeographic analysis of mtDNA variation in four ethnic populations from Yunnan Province: new data and a reappraisal. Journal of Human Genetics, 2002, 47: 311 - 318.

[20] Trejaut J A, Kivisild T, Loo J H, et al. Traces of archaic mitochondrial lineages persist in Austronesian-speaking Formosan populations. PLoS Biology, 2005, 3: e247.

[21] Li D, Li H, Ou C, et al. Paternal genetic structure of Hainan aborigines isolated at the entrance to East Asia. PLoS One, 2008, 3: e2168.

[22] Li D, Sun Y, Lu Y, et al. Genetic origin of Kadai-speaking Gelong people on Hainan island viewed from Y chromosomes. Journal of Human Genetics, 2010, 55: 462 - 468.

[23] Li H, Cai X, Winograd-Cort E R, et al. Mitochondrial DNA diversity and population differentiation in southern East Asia. American Journal of Physical Anthropology, 2007, 134: 481 - 488.

[24] Gan R J, Pan S L, Mustavich L F, et al. Pinghua population as an exception of Han Chinese's coherent genetic structure. Journal of Human Genetics, 2008, 53: 303 - 313.

[25] 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. American Journal of Physical Anthropology, 2010, 143: 555 - 569.

[26] He J D, Peng M S, Quang H H, et al. Patrilineal perspective on the Austronesian diffusion in Mainland Southeast Asia. PLoS One, 2012, 7: e36437.

[27] Lu Y, Wang C, Qin Z, et al. Mitochondrial origin of the matrilocal Mosuo people in China. Mitochondrial DNA, 2012, 23: 13 - 19.

[28] Excoffier L, Laval G, Schneider S. Arlequin ver. 3.0: An integrated software package for population genetics data analysis. Evolutionary Bioinformatics Online, 2005, 1: 47 - 50.

[29] Su B, Xiao C, Deka R, et al. Y chromosome haplotypes reveal prehistorical migrations to the Himalayas. Human Genetics, 2000, 107: 582 - 590.

[30] Kang L, Lu Y, Wang C, et al. Y chromosome O3 haplogroup diversity in Sino-Tibetan populations reveals two migration routes into the eastern Himalayas. Annals of Human Genetics, 2012, 76: 92 - 99.

[31] Wang D, Su L Y, Zhang A M, et al. Mitochondrial DNA copy number, but not haplogroup, confers a genetic susceptibility to leprosy in Han Chinese from Southwest China. PLoS One, 2012, 7: e38848.

[32] Yang K, Zheng H, Qin Z, et al. Positive selection on mitochondrial M7 lineages among the Gelong people in Hainan. Journal of Human Genetics, 2011, 56: 253 - 256.

第5章　北方原住民族的Y染色体

东亚北方包括西伯利亚的民族，其语言分为乌拉尔语系、叶尼塞语系、阿尔泰语系和古西伯利亚语系4个类群。乌拉尔语系人群的主流Y染色体单倍群是N1-TAT。N起源于汉藏族群，在汉族中有N1和N2的各种亚型，藏缅族群中的N主要是N2，而乌拉尔的N1-TAT只是一种下游的类型。所以乌拉尔族群的父系来自东亚中西部是毫无疑问的。乌拉尔语系分为萨摩耶德语族、尤卡吉尔语族、乌戈尔语族、芬语族。萨摩耶德语族在俄罗斯东扩前，从阿尔泰山北麓一直分布到北冰洋，其东侧沿北冰洋是尤卡吉尔语族，西侧乌拉尔山南部和匈牙利是乌戈尔语族，乌拉尔山北部到芬兰、爱沙尼亚大致是芬语族。叶尼塞语系目前仅存的是叶尼塞河下游的Ket语两种方言，但在俄罗斯东进前的西伯利亚汗国时期，沿着叶尼塞河分布着十多种叶尼塞语系的语种，分布在萨摩耶德语族的东侧，是一个较大的语系。其中有些族称很有意思，比如Assan、Kotch，或许来自乌孙、月氏，也有人认为Ket一名来自羯族。古代的匈奴也最可能属于叶尼塞语系。现代Ket人的Y染色体几乎全是Q，而匈奴遗骸的Y染色体也以Q为主。如果这些古代民族都是叶尼塞语系，那么这个语系应该是汉藏语系的北邻，长期广泛分布于中国北方草原。有趣的是，叶尼塞语系是北方四语系中唯一一种有调形声调的分析语，与汉藏语系、苗瑶语系、侗傣语系一样，是东亚地区独有的现象。古西伯利亚语系分布在黑龙江入海口和西伯利亚最东北角的楚克奇堪察加地区，与叶尼塞语系有部分同源词，所以可能是近缘的，但不是声调语言。古西伯利亚语系的人群Y染色体也以Q型为多，也有N和C等北亚常见类型。相比之下，阿尔泰语系上古时期可能仅分布于黑龙江上游，其居民Y染色体以C型为主，在西扩过程中渐渐融入了大量其他语系的部族，形成了现在极不均衡的Y染色体类型分布，特别是分布在中国西北部和中亚地区的突厥语族。

中国西北地区是欧亚大陆东西方交流的最前沿，来自东亚、西亚、南亚、北亚甚至欧洲的人群在这里混合交融。笔者调查了1 514例男性无关个体的样本的Y染色体，发现东亚南方起源的四种单倍群C、D、O、N的频率占该地区群体所有Y染色体单倍群频率的64.36%，其他Y染色体单倍群的比例为35.64%，说明东亚Y染色体特征谱系在中国西北占主体地位。单倍群E、F、G、H、I、J在西北地区人群中的分布反映了该地区的群体与西部欧亚大陆的基因交流历史。西方特征谱系的频率自西向东呈递减的梯度。中国西北地区

群体内部各族群在主成分图上彼此分散的状态体现了阿尔泰语系人群的遗传非同源性。

阿尔泰语系虽然被认可为一个可以成立的语系，但目前多数语言学家倾向于认为所有的阿尔泰语相似性是由接触和借贷关系导致的，而不是同源关系的表现。笔者总结了目前关于欧亚大陆北部几乎所有人群的父系遗传的Y-SNP遗传结构。结果显示，阿尔泰语人群的主要父系类型C3-M217、N-M213、R1a1a-M17以及Q-M242都有完全不同的独立起源。并且，除了C3外，这些父系类型在某些非阿尔泰语的人群中的比例比在任何阿尔泰语人群中都要高。母系方面的数据也表明，有阿尔泰语3个主体语族（满-通古斯、蒙古、突厥）的人群有较大的差异。这表明，绝大多数阿尔泰语人群在遗传上没有共同的起源，这一结果支持语言学的观点。实际上，满蒙两语族可能接近阿尔泰语系的核心，而突厥语族更可能是大融合的结果。

有研究认为，日本、朝鲜和阿伊努3个语族从类型上可归为阿尔泰语系，但更多观点认为他们都是孤立的语系，比如日本语是扶余-百济语系的遗存。作为一个长期隔离于东亚大陆的岛屿群体，日本人的起源迁徙路径一直有争议。考古学证据证明在日本的史前时期至少有两批人群从东亚大陆迁徙前往日本，分别是绳文人（Jomonese）与弥生人（Yayoiese）。但是，绳文人与弥生人对现代日本人的遗传贡献比例并不清晰。史学界对绳文人和弥生人的演变过程有替代说、维持说和混合说3种观点。利用Y染色体和线粒体DNA多样性数据检验日本人起源的不同假说，认为生活在日本主岛上的现代日本人是绳文人、弥生人与中国移民混合的结果，符合日本人起源的混合模型。

5.1　中国西北人群的父系遗传结构

5.1.1　研究背景

中亚，顾名思义是欧亚大陆的中心地带。它位于东亚、西亚和东欧之间，南接南亚，北接西伯利亚。中亚地区的群体由多种人种和民族成分构成，具有较高的遗传多样性。有关该地区人群多样性起源的研究形成了两个相互对立的观点："混合假说"认为中亚地区是欧亚大陆东部和西部人群混合的结果；而"源泉假说"则提出中亚地区的群体是欧亚大陆群体的基因源泉[1]，是一系列人口迁徙事件的主要来源[2]。混合假说得到了线粒体DNA研究的支持，对中亚现有人群的线粒体类型分析显示，中亚群体的线粒体都可以分别归入欧亚东部、欧亚西部和欧亚南部3大类群[3-5]。最近也有来自常染色体的研究结果支持"混合假说"[6]。而源泉假说得到了Y染色体研究的支持，特别是P系单倍群起源于中亚，最终R亚群西迁成为印欧语系人群最大类群，而Q亚群东迁成为叶尼塞语系、古亚语系和美洲诸语系人群的主体类群。

现代人走出非洲进入东亚有两条主要迁徙路线[7,8]：一条由中东进入中亚和东亚北部，称为北线；另一条是沿着亚洲大陆南部海岸线进入南亚并到达东亚南部的南线。其中南线由于气候条件优越，在大约5万年前就开始迁徙事件。而北线的迁徙可能仅仅是1万年以内的事件。中国西北地区在中亚和东亚之间，虽然也位于北线上，但是由于南线迁徙远早于北线，南线人群与北线人群可能大致同时到达这一地区，所以中国西北人群可能受走出非洲后分别走不同线路的现代人的基因影响。

早期有研究认为东亚北方的 Y 染色体多样性高于南方，因为中国西北地区紧邻中亚，北方的群体可能与中亚群体的关系更接近[9]。考古资料证明，在先秦时期，东西方之间就存在着经济贸易交流。在德国南部斯图加特及乌克兰的克里米亚半岛都曾发掘到我国春秋战国时期生产的丝绸[10]。张骞通使西域以后，一条东起长安，经陇西、河西走廊，然后沿塔里木盆地南北两缘，进而连接中亚、南亚、西亚和欧洲的中西交流通道正式建立起来了。由于当时在这条交通通道上输出的物品中主要是丝织品，所以欧洲的学者首先把它称为“丝绸之路”(图 5－1)，这条道路最远曾经到达地中海沿岸地区[11]。以丝绸之路为标志的古代东西交通网络在很大程度上影响了中国西北人群的遗传构成。

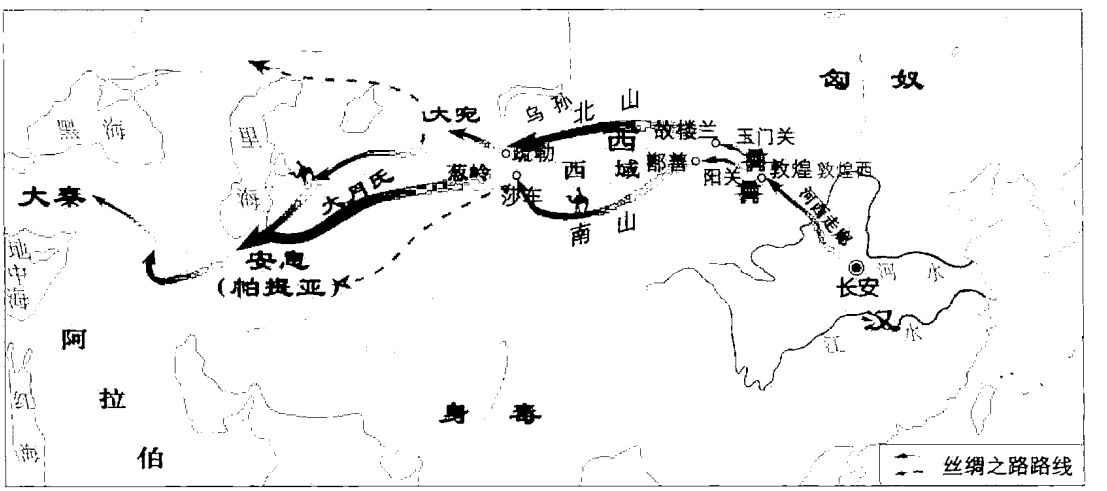


图 5－1 丝绸之路示意图

历史上，中国西北地区经历了多次文化更替、人群迁徙以及基因交流，是一个多民族聚居，多种宗教(萨满教、祆教、摩尼教、佛教、景教、伊斯兰教、道教等)并存的地区。本节中，笔者对中国西北地区分属于 6 个民族的 10 个人群进行了 Y 染色体遗传结构分析。这些群体中除讲汉语的回族外，还包括分属阿尔泰语系的突厥语族和蒙古语族。对该地区群体的遗传学研究，有助于了解西北人群的起源和形成历史，及其与欧亚大陆东西部人群之间的关系和基因交流的程度。

5.1.2 材料和方法

1. 群体样本

本研究共采集了中国西北地区 1 514 例男性无关个体的样本。其中包括哈萨克族样本 268 份(新疆哈密地区 101 份、新疆伊犁地区 87 份、新疆昌吉地区 80 份)、维吾尔族样本 621 份(新疆和田地区 478 份、新疆吐鲁番地区 143 份)、回族样本 240 份(宁夏回族自治区 65 份、新疆昌吉地区 175 份)、青海省循化县的撒拉族样本 135 份、青海省互助县的土族样本 121 份，以及青海德令哈市蒙古族样本 129，群体样本信息及其分布见图 5－2。

其中哈萨克族、维吾尔族和新疆回族的样本为外周血样本，宁夏回族为纱布干血点

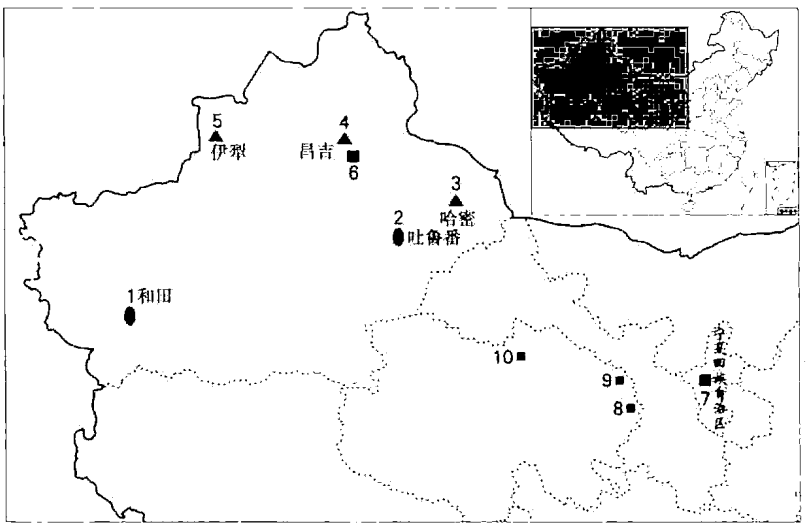


图5-2　本研究的群体样本信息和群体分布

维吾尔族(1指新疆和田地区,2指新疆吐鲁番地区);哈萨克族(3指新疆哈密地区,4指新疆昌吉地区,5指新疆伊犁地区);回族(6指新疆昌吉地区,7指宁夏回族自治区);8指撒拉族;9指土族;10指蒙古族。

样本,撒拉族、土族和蒙古族样本为面颊拭子刮取的口腔上皮细胞。采样过程中遵循知情同意和匿名的原则,样本均来自无亲缘关系的健康个体,采样活动通过了复旦大学生命科学学院的伦理审查。针对以上不同形式的样本,分别采用了不同的方法提取DNA。

2. *实验方法*

所有样本均采用包含92个Y染色体SNP位点的如下7组版块进行基因分型检测,分别是:根部单倍群、单倍群O、单倍群C、单倍群D、单倍群N、单倍群Q、单倍群J和单倍群R(表5-1)。

SNaPshot多重反应试剂盒对DNA模板进行荧光PCR反应后,在ABI 3730 xl测序仪上读取结果。对序列长度受限制或者在多重PCR反应体系中受其他引物的影响而导致最终实验效果不理想的个别位点,本研究采用了Taqman的方法进行检测。

此外所有的样本还进行了17个位点的STR检测,采用了Applied Biosystems公司的AmpFLSTR® YfilerTM PCR扩增试剂盒,所检测的位点包括:DYS19、DYS385a、DYS385b、DYS389I、DYS389II、DYS390、DYS391、DYS392、DYS393、DYS426、DYS437、DYS438、DYS439、DYS448、DYS456、DYS635和YGATA H4。

3. *数据分析*

收集并整理了已发表的包括西亚、欧洲、中亚、北亚和东亚等地区的110个欧亚大陆群体[1,2,12-16]的单倍型频率数据。利用SPSS15.0软件对Y染色体单倍型频率进行主成分分析,研究中国西北地区群体与欧亚大陆群体间的亲缘关系。

汇集已发表文献的参考文献数据[2,12-26],利用中点连接法计算并绘制了网络结构图[27]。该分析经软件Network 4.6完成,以研究各个单倍群内部细支之间的相互关系和

表 5-1 具体版块内包含的 SNP 位点及其所在单倍群信息

版1	SNP	M130	P256	YAP	M231	M168	M174	M45	M89	M272	M258		
	单倍群	C	M	DE	N	CR	DE	P	F-R	T	I		
	SNP	M242	M207	M9	M96	M175	P125	M304	M201	M306			
	单倍群	Q	R	K-R	E	O	IJ	J	G	R			
版2	SNP	M268	M7	M88	M324	P203	M164	M176	M110	M122			
	单倍群	O2	O3a3b	O2a1	O3a	O1a1	O1a2	O2b	O1a2	O3			
	SNP	P31	M95	2611	P201	P164	PK4	L127	M134	M121			
	单倍群	O2	O2a	O3a4	O3a3	O3a2c*	O2a*	O3a1*	O3a2c1*	O3a1a			
版3	SNP	P54	M105	M48	M208	M407	P33	M93	P39				
	单倍群	C2a2	C1	C3c	C2a	C3d	C2a1	C3a	C3b				
	SNP	P92	P53.1	M217	M38	M210	M356	P55	M347				
	单倍群	C5a	C3e	C3	C2	C4a	C5	C6	C4				
版4	SNP	P42	P12	N2	M125	M55	P53.2	P99	P47	M15	M151	22457	P120
	单倍群	D2a1a	D2a1a1	D1a1	D2a1	D2	D2a1b1	D3	D3a	D1	D2a2	D2a1b	D2a3
版5	SNP	M242	M3	M120	MEH2	M378	M25	N14	P36.2	M346	L53	M323	
	单倍群	Q	Q3	Q1a1	Q1a	Q1b	Q1a2	Q1a1	Q1	Q1a3	Q1a3	Q1a6	
版6	SNP	M306	M173	M124	M420	M64.1	M17	M198					
	单倍群	R	R1	R2	R1a	R1a1a3	R1a1a	R1a1a					
	SNP	M343	V88	M458	M434	P312	U106	M269					
	单倍群	R1b	R1b1a	R1a1a7	R1a1a6	R1b1b2a1a2	R1b1b2g	R1b1b2					
版7	SNP	P56	M62	M410	P58	M68	M280	M92					
	单倍群	J1d	J1a	J2a	J1e	J2a3	J2b2b	J2a2a					
	SNP	M47	M267	M172	M390	M12	M67	M158					
	单倍群	J2a1	J1	J2	J1c	J2b	J2a2	J2a5					

在群体中的分布。各个单倍群的分支年代估算也是由 Network 4.6 完成，STR 的突变率选用进化突变率为 0.000 69[28]，每 25 年一世代。由于参考文献数据之间所涉及的 STR 位点数目不尽相同，在网络结构图的绘制中，根据各个单倍群 STR 数据的位点实际信息，选择不同的 STR 位点组合。单倍群 C3c 的网络结构图选择了 6 个 STR 位点：DYS389I、DYS389Ib、DYS390、DYS391、DYS392 和 DYS393。单倍群 C3 *、D1、D3、O3a2c1 * 和 O3a2c1a 的网络结构图选择 7 个 STR 位点：DYS19、DYS389I、DYS389Ib、DYS390、DYS391、DYS392 和 DYS393。单倍群 J 和 R1a1 选取 8 个 STR 位点：DYS19、DYS389I、DYS389b、DYS390、DYS391、DYS392、DYS393 和 DYS439。单倍群 N1 *、N1c 的网络结构图采用 10 个 STR 位点：DYS19、DYS389I、DYS389b、DYS390、DYS391、DYS392、DYS393、DYS437、DYS438 和 DYS439。同时，基于以上 STR 数据，采用 Arlequin 3.0 软件包[29]计算各个群体的单倍型平均多样性。

5.1.3　研究结果

1. Y 染色体 SNP 单倍群频率分布

在中国西北地区的群体中共检测出了包括 C、D、E、F、G、H、I、J、L、N、O、P、Q、R、T 在内的 15 种 Y 染色体 SNP 的根部主干单倍群，以及主干单倍群下游的 41 个亚单倍群，单倍群在各群体中的频率分布见表 5－2。只有非洲的 A、B 和大洋洲的 K、S、M 没有发现。

表 5－2　中国西北地区群体 Y 染色体 SNP 单倍型频率分布(%)

群　体		哈萨克族			维吾尔族		回族		撒拉族	土族	蒙古族
		哈密	伊犁	昌吉	和田	吐鲁番	宁夏	昌吉	青海	青海	青海
样本量		101	87	80	478	143	65	175	135	121	129
C总	M130	79.21	32.18	78.75	8.37	16.78	9.23	9.71	5.93	13.22	27.13
C1	M1										2.33
C3 *	M217	71.29	27.59	7	7.32	13.29	9.23	8.00	2.96	7.44	9.30
C3c	M48	7.92	1.15	3.75	1.05			0.57			13.18
C3d	M407		3.45	5.00		0.70		1.14			
C3e	P53.1					2.10				0.83	0.78
D总	M174	6.93		1.25	2.72	2.80	1.54	4.00	4.44	10.74	34.11
D1	M15				1.26	2.10	1.54	1.14	2.22	7.44	4.65
D3	P99	6.93		1.25	1.46	0.70		2.86	2.22	3.31	29.46
E	M96				2.51	2.80	1.54	1.14			
F *	M89				0.84						
G	M201			1.25	5.65	0.70		1.71	2.22	2.48	0.78
H	M69	0.99			3.56			1.71	5.19		4.65

（续表）

群体		哈萨克族			维吾尔族		回族		撒拉族	土族	蒙古族
		哈密	伊犁	昌吉	和田	吐鲁番	宁夏	昌吉	青海	青海	青海
样本量		101	87	80	478	143	65	175	135	121	129
I	M258				1.26			1.14	2.96		
J总	M304	2.97	1.15	2.50	15.69	12.59	10.77	11.43	2.96	5.79	1.55
J1*	M267				1.67		1.54		nd	nd	nd
J1e	P58				2.93	0.70	1.54		nd	nd	nd
J2*	M172				1.26	3.50	3.08		nd	nd	nd
J2a	M410			1.25	7.95	6.29	3.08	9.14	nd	nd	nd
J2b	M12						1.54		nd	nd	nd
L	M20				3.56	0.70					
N总	M231	1.98	2.30	5.00	5.02	4.90	6.15	4.57	1.48	12.40	2.33
N1*	LLY22g	1.98		1.25	1.67	2.80	nd	nd		2.48	
N1a	M128				0.21		nd	nd	1.48	7.44	1.55
N1c1	M178		2.30	3.75	3.14	2.10	nd	nd		2.48	0.78
O总	M175	4.95	59.77	8.75	14.23	10.49	38.46	46.29	22.22	35.54	7.75
O1a*	M119						3.08				
O1a1	P203				0.21		7.69	2.86	4.44	1.65	0.78
O2	P31				0.63			1.14	0.74	9.09	0.78
O2a	M95		1.15		0.84			4.00	nd	nd	nd
O2b	M176							0.57			
O3*	M122		1.15				1.54	1.71		2.48	
O3a*	M324		1.15					0.57	2.96		4.65
O3a1*	KL1		1.15		0.21				nd	nd	nd
O3a1c	2611.00				0.63		7.69	10.86	1.48	3.31	
O3a2*	P201		3.45					1.14			
O3a2b	M7		1.15				1.54		4.44	3.31	2.33
O3a2c*	P164	1.98	2.30	1.25	3.14			6.29			
O3a2c1*	M134	2.97	48.28	7.50	3.14	9.09	6.15	8.00	3.70	4.96	0.78
O3a2c1a	M117				5.44	1.40	10.77	9.14	3.70	7.44	3.10
P	M45								5.19	0.83	0.78

（续表）

群　体		哈萨克族			维吾尔族		回族		撒拉族	土族	蒙古族
		哈密	伊犁	昌吉	和田	吐鲁番	宁夏	昌吉	青海	青海	青海
样本量		101	87	80	478	143	65	175	135	121	129
Q总	M242	1.98	1.15	1.25	7.74	15.38	1.54	4.57	2.96	1.65	5.43
Q1*	P36.2		1.15		1.46	3.50		1.14			1.55
Q1a1	M120				1.05	2.80		1.71	0.74		1.55
Q1a2	M25					0.70		0.57			0.78
Q1a3	M346				5.02	2.80		1.14		0.83	1.55
Q1b	M378					4.20			1.48		
R总	M207	0.99	3.45	1.25	28.24	32.87	30.77	13.71	44.44	17.36	15.50
R1a*	M420					0.70					
R1a1a*	M17		3.45	1.25	21.55	19.58	16.92	9.14	27.41	14.05	4.65
R1a1a7	M458				0.21						
R1b*	M343	0.99			2.30	7.69	6.15	0.57	5.19	0.83	6.20
R1b1b	P312				0.21						
R1b1b2	M269				0.84	0.70		2.29			2.33
R2	M124				1.88	2.80	4.62	1.14	2.96		
T	M272				0.63						

注：J单倍群的总频率包括下游已细分亚单倍群以外未能分类的部分。

根据平均分布频率的高低，将Y染色体单倍群大致分为3类。第1类是中国西北地区群体中较为高频的单倍群，平均频率为15%～30%，包括单倍群C、O和R，平均分布频率分别为28.05%、24.84%和18.86%。在这3个较为高频的单倍群中也各有高频亚群，单倍群C的下游单倍群C3*的频率占22.64%，O的下游单倍群O3a2c*平均频率为9.46%，R的下游单倍群R1a1a平均频率为11.8%。第2类是中度频率分布的Y染色体SNP单倍群，频率为4%～10%，如单倍群D（平均频率6.85%）、J（平均频率6.74%）、N（平均频率4.61%）和Q（平均频率4.37%）。第3类是极低频的单倍群，如单倍群E、F、G、H、I、T，在群体中的平均频率不超过2%。

之前的研究认为Y染色SNP单倍群C、D和O从南方进入东亚[17,23,30]，是东亚的特征单倍群，而单倍群E、F、G、J、I、T等是欧亚西部特征单倍群，在西亚、中东和欧洲有着高频分布[16,31]。欧亚大陆东西部特征单倍群在本研究的中国西北人群中同时存在，说明该地区的群体同时受到欧亚大陆东部和西部的影响，并且从单倍群频率来看，东部的特征单倍群频率要大于西部的单倍群频率。

Y 染色体 SNP 单倍群频率在所研究的中国群体内部也有很大的频率差异。单倍群 C 是哈萨克族的主体单倍群，该单倍群在哈密地区和昌吉地区的哈萨克族中的分布频率高达 79.21%和 78.75%。伊犁地区的哈萨克族的情况则有所不同，该群体中主要单倍群为 O，单倍群 C 的频率为 32.18%，低于 O 的频率。在维吾尔族群体中单倍群 R 的频率最高，该单倍群在吐鲁番维吾尔族中为 32.87%，和田维吾尔族中为 28.24%。此外，和田地区的维吾尔族中存在的欧亚西部特征单倍群 E、F*、H、I 和 T 在吐鲁番地区的维吾尔族中并未出现。回族群体中的主要单倍群为 O 和 R，这两个单倍群的频率总和占 60%以上，新疆回族群体中低频分布的单倍群 G、H 和 I 未在宁夏回族群体中发现，除此之外，这两个地区的回族群体在单倍群频率分布上无明显差异。单倍群 R 在青海撒拉族中的频率高达 44%，是此次研究的中国西北地区群体中最高的。土族群体中高频的单倍群为 O，频率为 35.54%，单倍群 C、D、N 和 R 的频率为 10%～20%。青海蒙古族中有 34.11%的单倍群 D 的分布，可能是由于该群体受到青海地区藏族的影响，其次是单倍群 C，频率为 27.13%。

2. 群体聚类

本研究收集了已发表的欧亚大陆的 110 个群体的 Y 染色体 SNP 单倍群频率数据进行主成分分析，在第一主成分(PC1)和第二主成分(PC2)绘制的散点图(图 5－3)中，绿色部分代表包括东亚的汉藏和侗傣语系群体在内的欧亚大陆东部群体，红色代表欧亚大陆西部，包含中亚和欧洲的印欧和阿尔泰语系群体。蓝色部分则包含了西伯利亚南部以及中国东北部的阿尔泰群体，将之归为北亚群体。图 5－3 显示，本次研究的各个群体并未聚在一起而是分别与欧亚西部、北亚和欧亚东部聚类，显示出了彼此不同的亲缘关系。维吾尔族的两个群体和撒拉族与欧亚大陆西部群体距离较近，青海的蒙古族和哈萨克族的两个群体靠近北亚的阿尔泰群体，伊犁地区的哈萨克族接近东亚的群体，宁夏回族、新疆回族和土族则位于欧亚东部和欧亚西部群体之间。

3. 中国北方群体的 Y 染色体 SNP 单倍群类型

Y 染色体 SNP 单倍群具有较强的民族特异性，并且在欧亚大陆的分布存在显著的差异。通过研究群体中这些单倍群内部 STR 的分化，可以了解中国西北部群体与欧亚大陆群体之间的人群融合和迁移历史。通过 STR 网络结构图构建和观察分布于不同群体的同一种 Y 染色体 SNP 单倍群内的 STR 单倍型的共享和连接关系，有助于研究群体的起源和分化。

(1) 单倍群 C　Y 染色体 SNP 单倍群 C 在东亚[32,33]、东南亚、大洋洲[34,35]有着广泛的分布，它的下游分支单倍群在地理上呈现边缘离散的分布格局，暗示该单倍群的古老起源，随后又被其他单倍群所挤压而形成了现在的分布格局[36]。由于在 STR 的多样性上观察到的从南往北、自东向西的递减趋势，因而单倍群 C 被认为是南方起源，并且走东部的沿海路线从东南亚迁徙往东亚的[17]。

单倍群 C3* 是单倍群 C 下游在亚洲较为高频的亚单倍群。图 5－4 是按语系划分绘制的 C3* 的 STR 网络结构图，该网络结构图可以分为上下两个部分，下半部分中心主要

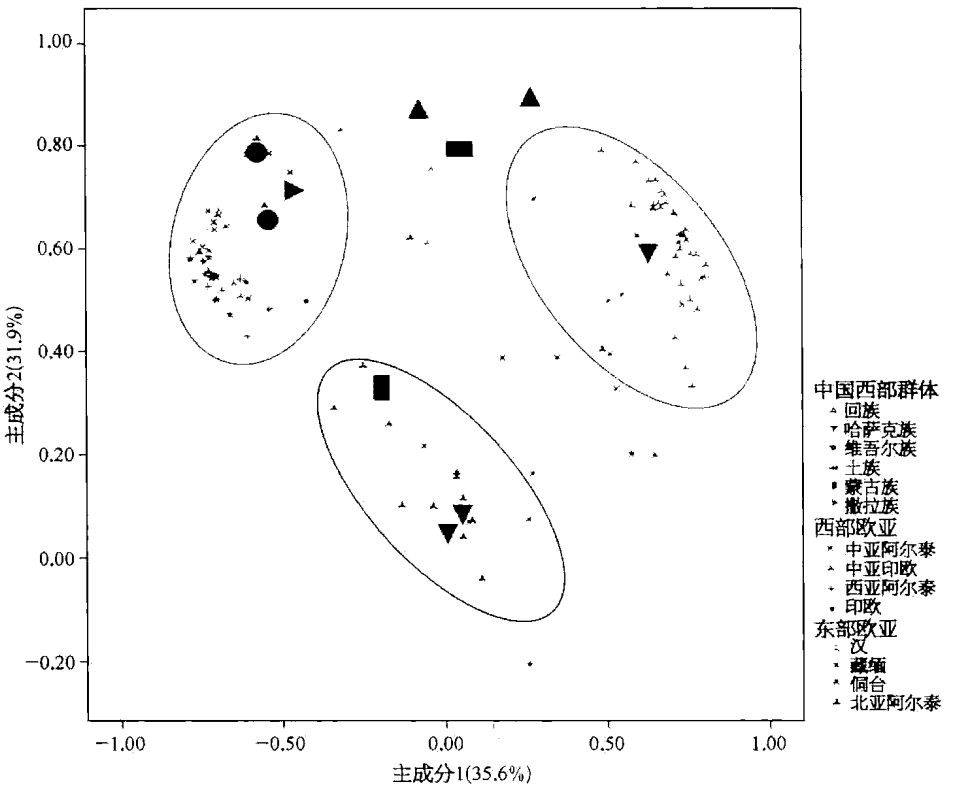


图 5－3　基于 Y 染色体 SNP 单倍群频率的东亚人群主成分分析图

分布着汉藏群体和苗瑶群体，阿尔泰群体则出现在边缘末端的位置，这暗示了单倍群 C3＊从苗瑶、汉藏群体到阿尔泰群体自南往北的迁徙。网络结构图的上半部分中，有一种高频的 STR 单倍型包含了很多阿尔泰族群的个体，这一现象与之前文献报道的欧亚大陆约占 8%频率的一类特殊的 STR 单倍型[37]相一致，之前的研究猜测该单倍型可能与蒙古帝国的扩张相关。计算 STR 网络结构图中各语系群体的 STR 平均多样性，在苗瑶群体中多样性为 0.490 8，侗傣为 0.377 1，藏缅为 0.342 9，汉族为 0.450 0，中国西北地区群体所在的阿尔泰语系的平均多样性最低，为 0.320 8。说明 C3＊也是从南到北迁徙的。

为了了解本次研究的各个群体之间的关系，进一步对网络结构图 5－4a 中涉及的阿尔泰群体进行细化，如图 5－4b。网络结构图的上半部分，维吾尔族、哈萨克族和青海蒙古族与阿尔泰语系突厥语族、蒙古语族和通古斯语族的其他群体共享同一个 STR 单倍型的情况较普遍，说明单倍群 C3＊在阿尔泰语系诸族群中频繁的融合和迁徙。回族中既有与阿尔泰群体共享或者相连接的 STR 单倍型，也有少部分与汉族相近的 STR 单倍型。撒拉族和土族群体所在的 STR 单倍型大多与汉族相近。

西北地区群体单倍群 C3＊的两个分支年代分别为：哈萨克族和维吾尔族（分支 1）18 818 年±6 024 年，回族和哈萨克族（分支 2）为 7 762 年±3 952 年。

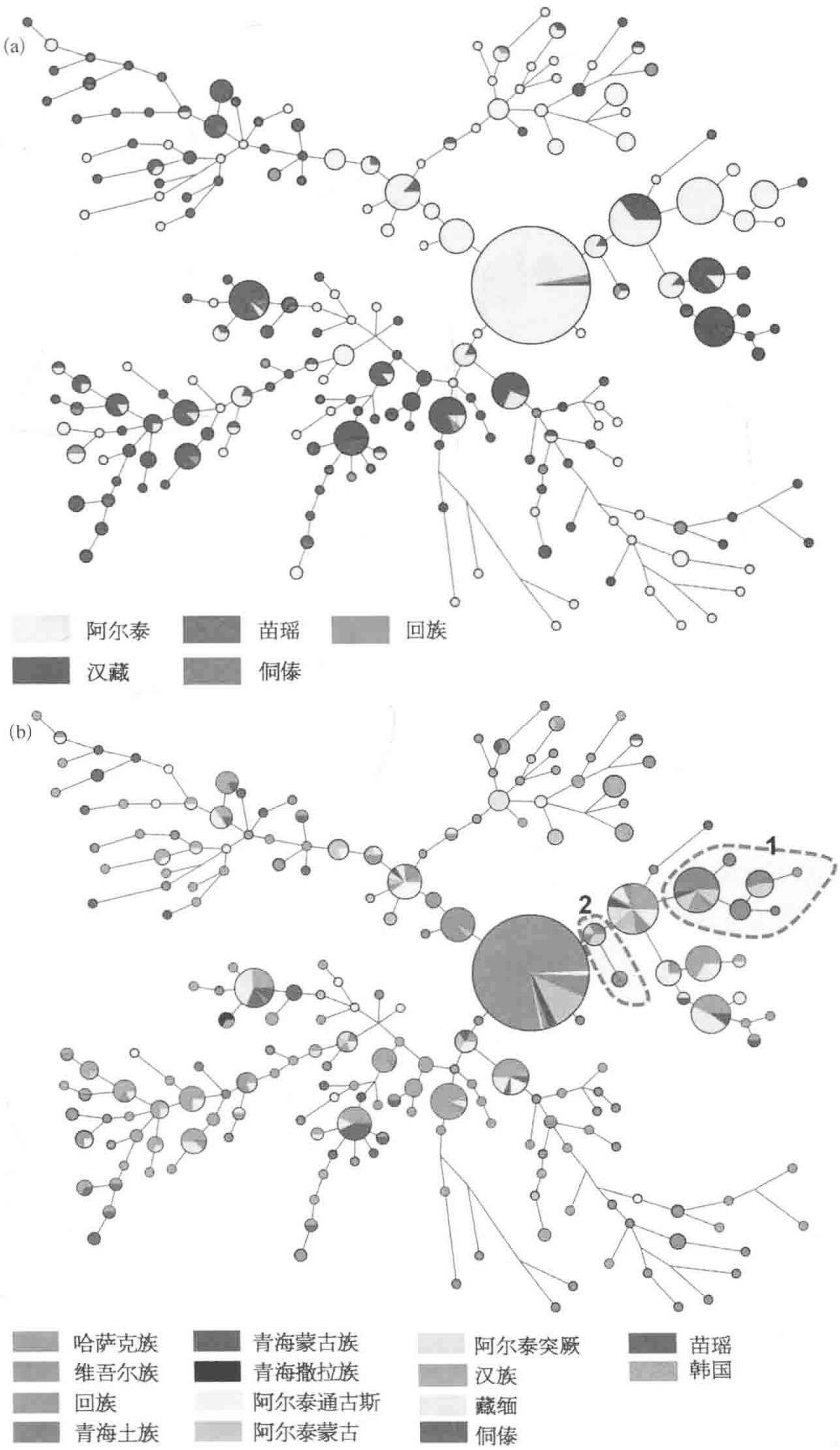


图 5-4 单倍群 C3* 的网络结构的最简树

(a) 族系区分示意图;(b) 西北民族区分示意图

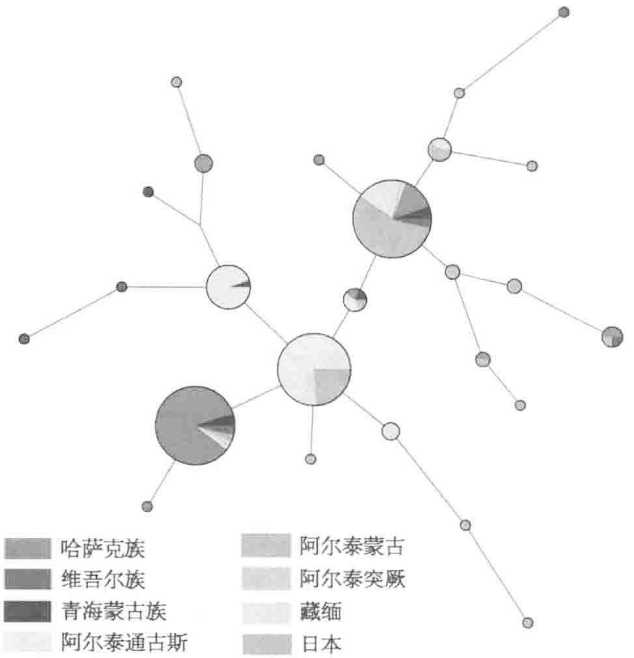


图5-5　单倍群C3c网络结构的最简树

单倍群C3的下游单倍群C3c几乎只出现在阿尔泰群体中，在汉藏以及韩国群体中有零星分布[17]。C3c的网络结构图显示出了星状结构，从图5-5中可以看到C3c的STR单倍型种类很少，提示该单倍型晚近的起源。该单倍型的整体时间为7 223年±2 927年，左下方哈萨克的分支年代为3 254年±3 299年。

(2) 单倍群D　单倍群D在亚洲也呈现和单倍群C相同的边缘分布格局，主要高频分布在印度洋的安达曼群岛[38]、中国的羌藏群体和日本[19]。单倍群D是起源于南方的单倍群，在南北人群中观察到了深度的分化，是东亚地区现代人中极其古老的支系[30]。

本研究构建了单倍群D1和D3的网络结构图。在D1的网络结构图(图5-6)中，总体来看，南方的苗瑶、侗傣和藏缅群体有自己的分支，处于网络结构图的上半部分。所研究的中国西北地区群体不与这些南方的群体直接连接，而是出现在和藏族群体相连接的下半部分。青海的土族和蒙古族与藏族群体共享单倍型，往北的维吾尔族个体分布在网络结果的外围，反映了相对藏区的距离远近而发生的由南往北的人群交流。单倍群D1网络结构图中西北地区群体一个分支的年代为9 610年±3 545年。

在D3的网络结构图(图5-7)中可以观察到，D3中大部分单倍型属于藏族，并且藏族群体拥有独有的分支。本研究的青海蒙古族群体拥有较高比例的D3。图中绝大多数的青海蒙古族都只与藏族群体共享最主要的单倍型。哈萨克族的样本从藏族所在的单倍型中分出，并处于网络图的外围区域，汉族和维吾尔族的样本绝大部分与藏族共享单倍型，反映了来自历史时期人群的融合。回族的样本与汉族样本相连或者共享单倍型。

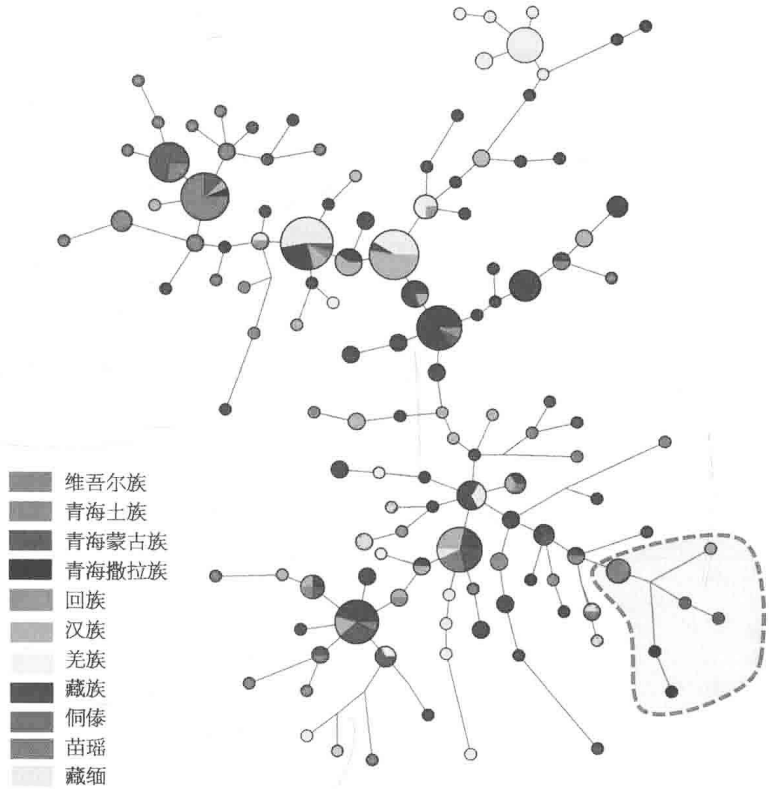


图 5-6 单倍群 D1 网络结构的最简树

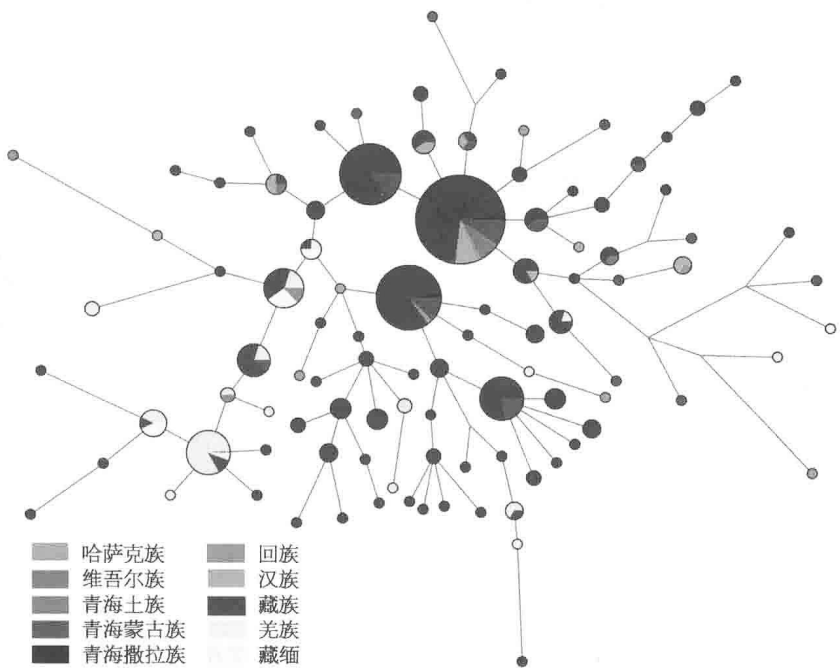


图 5-7 单倍群 D3 网络结构的最简树

（3）单倍群O　Y染色体单倍群O-M175是东亚地区现代人中频率最高、影响最为深远的单倍群。其下游突变形成的单倍群O3-M122、O2a-M95和O1a-M119在东亚人群中所占的频率可达57%以上。Y-SNP和Y-STR的数据也证明O单倍群下面主要的3个分支起源于南方[23,25,32,39,40]。其中单倍群O1a-M119和O2a-M95沿着东部沿海路线从东南亚进入东亚[41]，O3-M122则可能通过内陆路线进入东亚[42,43]。从单倍群O3a2c1*的网络结构图（图5-8）和O3a2c1a的网络结构图（图5-9）显示，汉藏群体的个体占据很高的比例，该单倍群是汉藏群体的主要父系特征单倍群。

从STR单倍型的平均多样性来看，西北部群体O3a2c1a的平均基因多样性为0.369 8，小于南方的群体：苗瑶为0.410 4，侗傣为0.430 5，汉藏为0.458 4。

西北部群体的总体O3a2c1*的平均多样性为0.351 8。把其中群体分开计算得到的结果显示出西北部各群体之间存在差异。哈萨克族为0.097 7，维吾尔族为0.485 5，回族为0.454 8，撒拉族为0.171 4，土族为0.304 8。维吾尔族和回族的STR单倍型较高的多样性可能是发生晚近混合的结果。

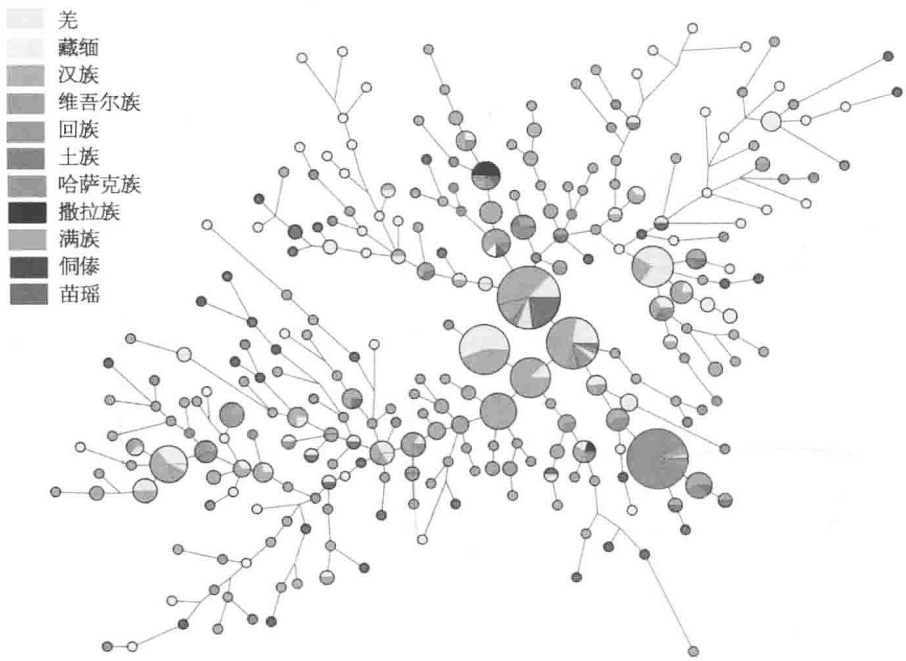


图5-8　单倍群O3a2c1*网络结构的最简树

（4）单倍群N　N1*在珞巴族群体中达到34.62%的高频分布，在僜人中有1.11%的分布频率，在东南亚也普遍存在，提示单倍群N的南方起源，N1*极有可能诞生于喜马拉雅山区东部的河谷地带。在N1*的网络结构图（图5-10）中，一些维吾尔族个体与珞巴族个体共享单同一种STR单倍型，单倍群N1*可能很早就分布在新疆地区，随后融入维吾尔族群体中。在N1c的网络构图（图5-11）上，N1c主要分布在乌拉尔群体以及阿尔泰群体中，维吾尔族和哈萨克族连接在一起，并且形成独立的分支，该分支的年代估算为12 378年±4 161年。

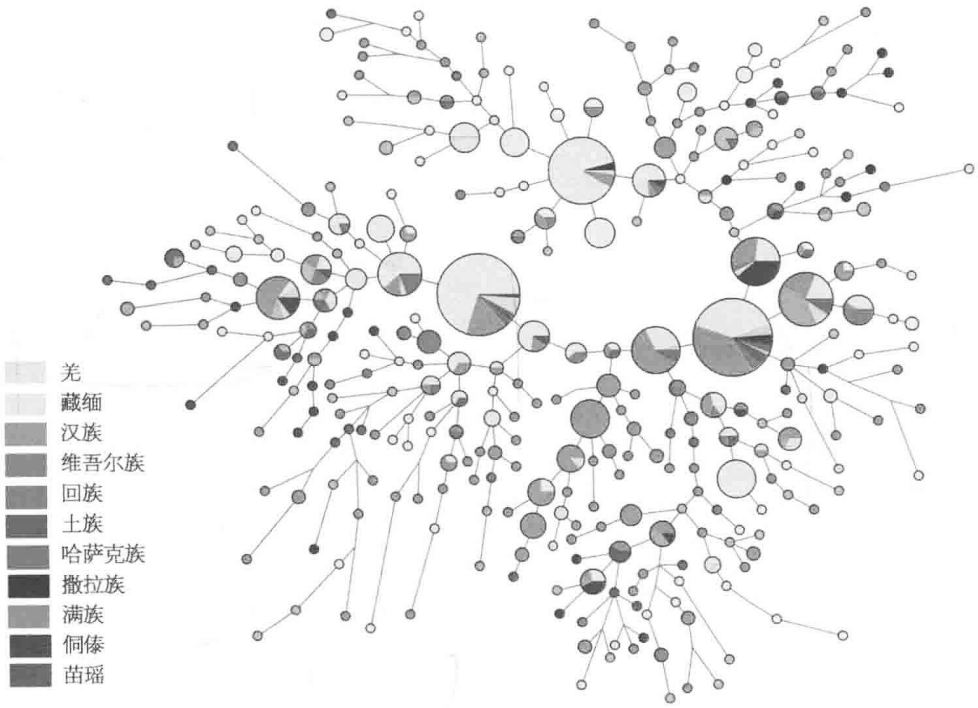


图 5-9 单倍群 O3a2c1a 网络结构的最简树

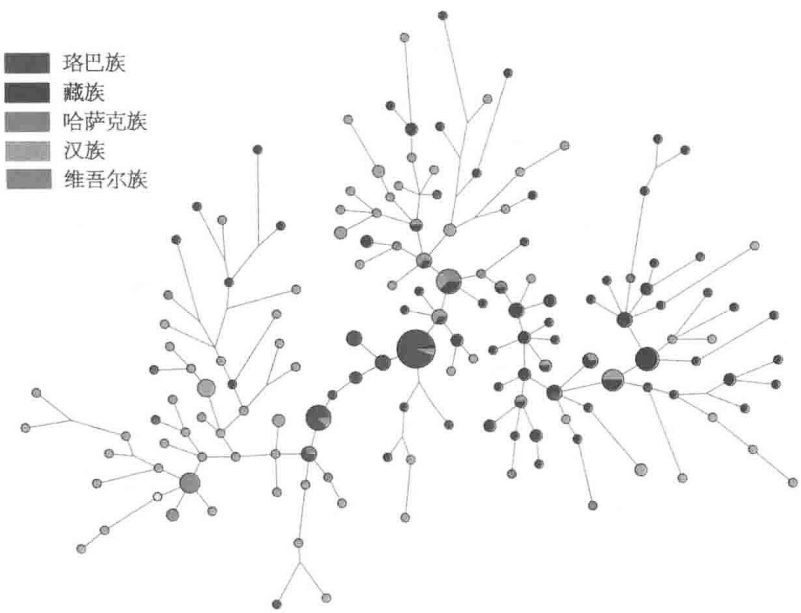


图 5-10 单倍群 N1* 网络结构的最简树

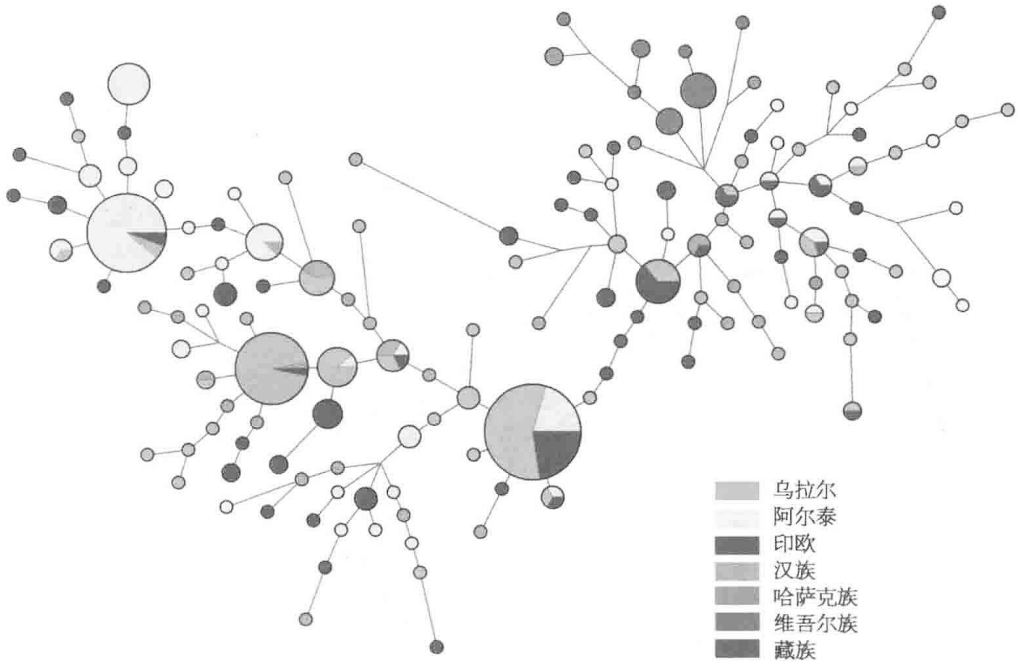


图 5－11　单倍群 N1c 网络结构的最简树

(5) 单倍群 J　Y 染色体单倍群 J 在中东地区分布广泛。该单倍型的迁徙跟西亚的农业文明人群扩张有着明显的联系[44]。下游单倍群 J2 在西亚较为常见，并且在中亚地区也有一定频率的分布[2]。中亚地区的 J2 来源于新石器时期西亚地区向东的农业扩张[45]。单倍群 J 的网络结构图(图 5－12)中，西北部群体中的维吾尔族和回族，有一部分直接与西亚的群体共享单倍群，另一部分与中亚的群体相连接，提示中国西北部群体中存在直接来自西亚的基因流动的可能性。网络结构图上，回族和维吾尔族分支 1 和分支 2 的时间分别为 2 716 年±2 014 年和 2 264 年±1 687 年。

(6) 单倍群 R　单倍群 R 在欧亚大陆有着非常广泛的分布，遍布欧洲、中亚、南亚和西亚等地区[1,21,46]。由于多次复杂的迁徙和融合叠加，在 R1a1a 的网络结构图(图 5－13)上，各地区群体所在的单倍型相互混杂，散落在各个位置，没有明显的规律可循。还存在整个欧亚大陆所有地区的群体共享同一种 STR 单倍型的情况。网络结构图中，少数回族个体与西亚的群体共享单倍型，显示了与西亚群体之间的基因联系。网络结构图中的一个维吾尔族、回族和土族群体共同的分支的时间估算为 13 018 年±4 802 年。另一个撒拉族的分支 1 年代相对年轻，为 3 801 年±3 460 年。

本研究计算了网络结构图中 STR 数据的平均基因多样性。南亚为 0.458 8，西亚为 0.419 3，中亚为 0.314 8，西伯利亚南部为 0.357 7，中国西北地区群体总的平均基因多样性为 0.425 2，其中维吾尔族为 0.424 6，回族为 0.431 1，土族为 0.410 3，蒙古族为 0.350 0，哈萨克族为 0.312 5，撒拉族为 0.302 8。中国西北部群体的 STR 多样性高于中亚群体的多样性。

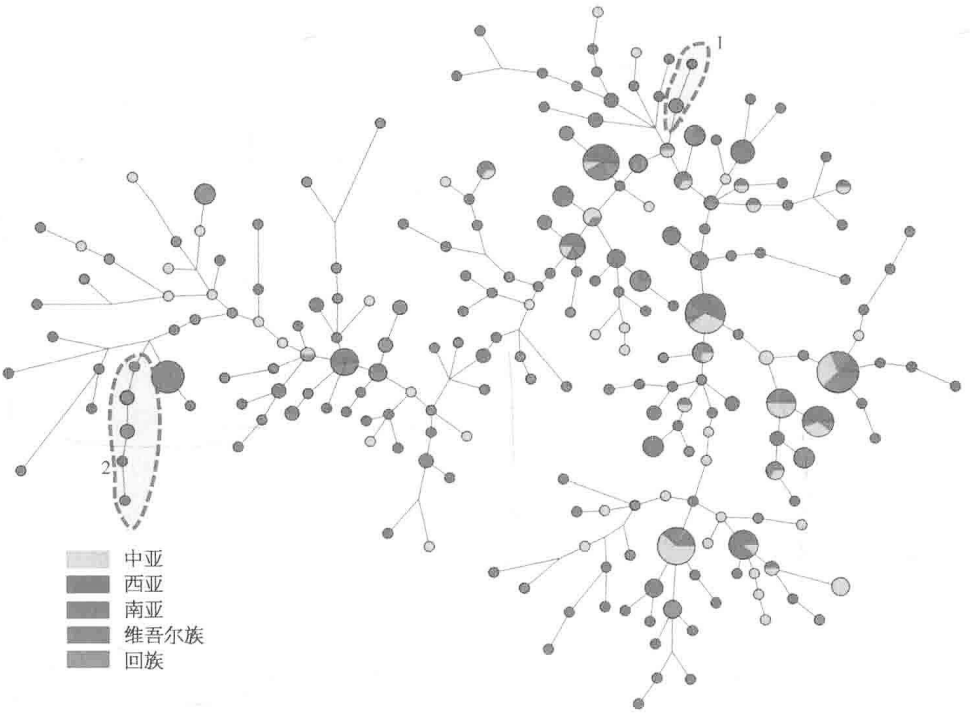


图 5-12 单倍群 J 网络结构的最简树

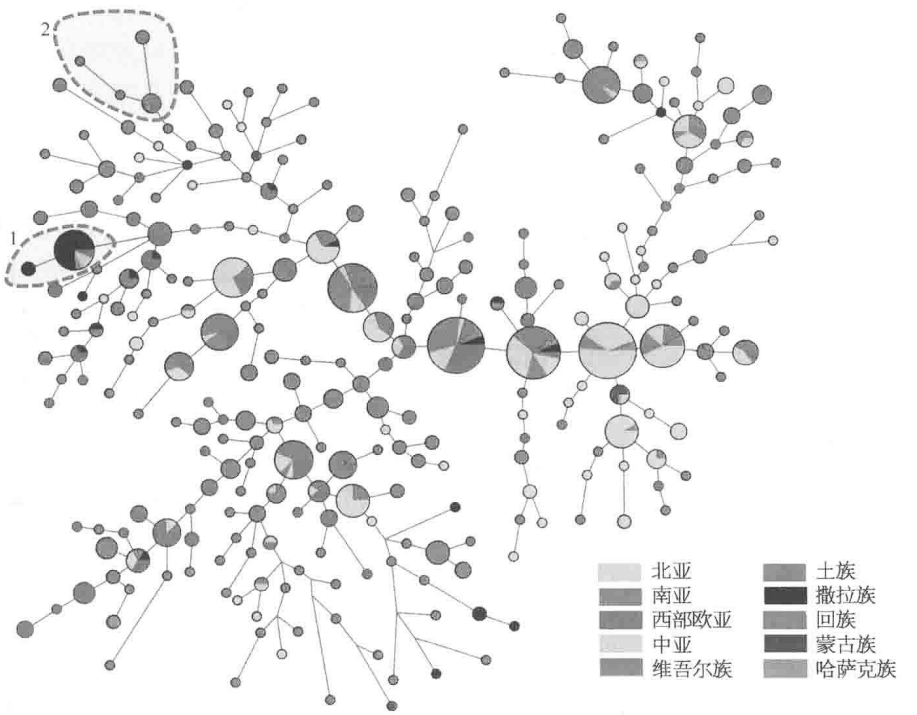


图 5-13 单倍群 R1a1a*(M17)网络结构的最简树

5.1.4　讨论

1. 欧亚大陆东西部特征谱系在西北人群中的混合比例

分子人类学早期研究提出了现代人走出非洲进入东亚的两条主要迁徙路线。一条由中东进入中亚和东亚北部，称为北线；另一条是沿着印度洋海岸线进入南亚并到达东亚南部的南线[8,47]。对东亚和东南亚的Y染色体[32,41-43]、线粒体DNA[48,49]和常染色体DNA[50,51]的研究认为，南线的成分是构成东亚人群基因库的主体，东亚地区的现代人主要是通过南方入口由东南亚进入东亚的。Y染色体SNP单倍群的研究证实单倍群O-M175、C-M130、D-M174和N-M231都是南方起源，并且在东亚地区呈现由南往北的多样性递减[17,23,30,52]。钟华等统计了这四个南方起源的Y染色体SNP单倍群在整个东亚地区群体中的频率高达92.87%，再一次证实现代人走出非洲向东亚迁徙的南线成分在东亚地区占绝对主流地位[53]。

在本研究的中国西北人群中，4种南方起源的单倍群C、D、O、N的频率占Y染色体所有单倍群频率的64.36%，其他Y染色体单倍群的比例为35.64%。虽然该地南方单倍群频率低于东亚人群的总体水平，却不难看出即便是在东西方交流频繁发生的中国西北地区，欧亚大陆东部的Y染色体特征谱系依旧占主导地位，所以西北地区现代人的主体成分也是经由南线进入东亚来到西北的。

线粒体DNA的研究发现，中国北方以及北亚的群体同时具有欧亚大陆东部和西部特征的单倍型，关于西北地区群体线粒体DNA的未发表数据显示，东西部谱系在该地区群体中的分布频率分别为67.6%和32.4%，这与Y染色体的频率比例基本一致。在线粒体DNA和Y染色体DNA水平上，中国西北地区人群中源于欧亚大陆东部与西部成分的比例约为7∶3，欧亚大陆东部的谱系在中国西北地区群体中具有较高的比重。

2. 欧亚大陆的基因交流

从南方起源的单倍群C、D、O和N的STR单倍型的网络分布图看，虽然西北人群单倍群的分布比较离散，但各群体所在的STR单倍型大多处于网络图的外围，这是一个较为普遍的状态。中国西北地区群体中的单倍群C主要源于东边蒙古族的扩张，单倍群D来自藏族的基因流入，汉藏群体将单倍群O带入西北地区群体中，单倍群N在从中国西南地区往北欧的逆时针迁徙时，也在西北地区的群体中遗留下了痕迹。

除此之外，西北地区的群体中还有低频的单倍群E、F、G、H和I的分布，这些单倍群是典型的欧亚大陆西部的特征谱系，在东亚南部群体中几乎未有发现[53]。单倍群I-M170基本出现在欧洲，在其他地区几乎没有分布，该单倍群可能起源于巴尔干地区[54]。在本研究的西北地区群体中，单倍群I-M170在新疆昌吉地区的回族与和田地区的维吾尔族中分别有1.14%和1.26%的分布。单倍群G-M201起源于亚洲西南部，在西亚和高加索地区有高频的分布[55,56]。新疆昌吉地区的哈萨克族、两个维吾尔族群体，以及青海的土族、蒙古族和撒拉族中均有单倍群G的分布，其中和田地区的维吾尔族中G的频率最高，为5.65%。单倍群H-M69主要分布在南亚次大陆，并被认为具有印度本土起源[57]。单倍群E分布于北非、中东地区和欧洲[31]，新石器时期农业文明的人口扩张将中

东地区的 E 带入欧洲地区[58]。这些西方特异的单倍群在中国西北地区群体中的出现反映了该地区的群体与欧亚大陆西部的基因交流历史。

对单倍群 J 和 R1a1a * 的网络结构图中的几个分支年代进行估算，西北地区群体中单倍群 J 的两个分支的时间都为大约 2 500 年。单倍群 R1a1a * 时间约为 13 000 年。钟华等认为，东亚北方群体中的单倍群 Q 和 R 代表了大约 1.8 万年前末次冰川后期来自北线的迁徙，单倍群 J2a 等则代表了近 3 000 年内的混合[53]。对新疆地区且末先民的古 DNA 样本研究认为，欧亚大陆东西部人群在中国新疆地区开始混合的时间大概在 2 500 年前[59]，这与单倍群 J 在西北地区群体中的时间估算相一致。单倍群 Q 在西伯利亚群体中有较高频的分布[60]。在西北群体中，Q 单倍群下游 Q1a1 - M120、Q1a3 - M346 和 Q1b - M378 均有少量分布，大多数群体中发现有单倍群 Q1 * 的分布。单倍群 R1a1a * 多样性在南亚地区最高，中国西北地区群体的多样性仅次于南亚，随后是西亚和南西伯利亚，中亚的 R1a1a * 多样性是最低的。中国西北地区群体中很高的 R1a1a * 多样性可能来自多次相互重叠的迁徙和融合事件，也可能因为单倍群 R1a1a * 很早就已经进入中国西北地区。

西方特征的单倍群 E、F、G、H、I、J 和 T 的频率总和自西向东呈递减趋势。新疆地区和田维吾尔族为 29.5%，吐鲁番维吾尔族为 16.08%，新疆昌吉回族为 17.14%，宁夏回族为 10.77%，在青海的撒拉族、土族和蒙古族中进一步降低为 13.3%、8.26% 和 6.98%。地理位置越靠近西方的群体，欧亚西部特征单倍群的频率越高。这种趋势在新疆地区的群体中表现得尤为明显，新疆自南向北分布着昆仑山、天山和阿尔泰山，塔里木盆地和准格尔盆地分别夹在天山和昆仑山、天山和阿尔泰山之间，复杂的地形条件使得西方人群进入新疆地区的扩散极其缓慢，气候环境的恶劣也使西方基因往东扩散的道路变得异常艰辛，因而在 Y 染色单倍群上观察到西方基因频率自西向东递减的趋势。

3. 西北地区各群体 Y 染色体的来源和历史

在 Y 染色体 SNP 频率的主成分图散点上，中国西北人群并没有聚类在一起，而是呈现出较为分散的状态，分别与欧亚大陆东部、西部和北部较为接近。显示了各群体不同的来源和族群历史。除回族外，这些群体分属于阿尔泰语系的突厥语族和蒙古语族，它们在遗传成分上彼此分散的状态也体现了阿尔泰语系人群的遗传非同源性[61]。

新疆地区的哈萨克族的主要 Y 染色体单倍群类型为 C3 *、C3c 和 O3a2c1 *。哈密和昌吉地区的哈萨克族中 C 的频率在 79%左右，该单倍群在哈萨克族中的高频分布被认为是与西伯利亚和蒙古境内的阿尔泰群体向西的扩张有关[62]。13 世纪成吉思汗西征以后的五六百年间，新疆及其周边广大地域一直在蒙古贵族统治之下。哈萨克的单倍群 C3c 显示出了与蒙古群体的联系。伊犁地区的哈萨克族中，单倍群 C 的频率只有 32%，单倍群 O 有 60%的分布，其中大部分都属于 O3a2c1 * - M134，对该单倍群的 STR 网络结构分析显示，伊犁哈萨克族与汉族共享 STR 单倍型，所以伊犁哈萨克族的O3a2c1 * - M134来源于汉族，这可能源于清代伊犁地区曾经经历过的大规模移民活动[63]。

新疆的两个维吾尔族群体在 Y 染色体单倍群分布上也有差别，位于南疆的和田维吾

尔族比北部吐鲁番维吾尔族具有更丰富的Y染色体SNP单倍型类型，是所有群体中受西方影响最大的。

对两个回族群体的研究结果显示，回族群体的主要来源是东亚地区的人群，在单倍群C3*、O3a2c1*和O3a2c1a的STR单倍型上与汉族有紧密的联系。宁夏和新疆地区的回族中检测到了低频的西方特征的单倍群。在单倍群J中与西亚群体共享STR单倍型。中东起源的单倍群E在回族中有低频的分布。中国西北地区的回族大部分人口源于受伊斯兰教影响的东亚原住民群体。该地区的回族主要源于宗教文化的转变，伴随有部分西方的基因流入。

参考文献

[1] Wells R S, Yuldasheva N, Ruzibakiev R, et al. The Eurasian heartland: a continental perspective on Y chromosome diversity. Proc Natl Acad Sci USA, 2001, 98(18): 10244 - 10249.

[2] Zerjal T, Wells R S, uldasheva N, et al. A genetic landscape reshaped by recent events: Y chromosomal insights into central Asia. Am J Hum Genet, 2002, 71(3): 466 - 482.

[3] Comas D, Calafell F, Mateu E, et al. Trading genes along the silk road: mtDNA sequences and the origin of central Asian populations. Am J Hum Genet, 1998, 63(6): 1824 - 1838.

[4] Yao Y G, Lu X M, Luo H R, et al. Gene admixture in the silk road region of China: evidence from mtDNA and melanocortin 1 receptor polymorphism. Genes Genet Syst, 2000, 75(4): 173 - 178.

[5] Comas D, Plaza S, Wells R S, et al. Admixture, migrations, and dispersals in Central Asia: evidence from maternal DNA lineages. Eur J Hum Genet, 2004, 12(6): 495 - 504.

[6] Xu S, Jin W, Jin L. Haplotype-sharing analysis showing Uyghurs are unlikely genetic donors. Mol Biol Evol, 2009, 26(10): 2197 - 2206.

[7] Underhill P A, Passarino G, Lin A A, et al. The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. Ann Hum Genet, 2001, 65 (Pt 1): 43 - 62.

[8] Underhill P A. Inferring human history: clues from Y chromosome haplotypes. Cold Spring Harb Symp Quant Biol, 2003, 68: 487 - 493.

[9] Karafet T, Xu L, Du R, et al. Paternal population history of East Asia: sources, patterns, and microevolutionary processes. Am J Hum Genet, 2001, 69(3): 615 - 628.

[10] 苗普生. 新疆历史研究需要阐明的几个问题. 西域研究, 2004, (2): 1 - 13.

[11] 田澍, 李勇锋. 世界遗产视野中的丝绸之路. 西北师大学报(社会科学版), 2007, 44(6): 12.

[12] Xue Y, Zerjal T, Bao W, et al. Male demography in East Asia: a north-south contrast in human population expansion times. Genetics, 2006, 172(4): 2431 - 2439.

[13] Derenko M, Malyarchuk B, Denisova G A, et al. Contrasting patterns of Y chromosome variation in South Siberian populations from Baikal and Altai-Sayan regions. Hum Genet, 2006, 118(5): 591 - 604.

[14] Wen B, Li H, Lu D, et al. Genetic evidence supports demic diffusion of Han culture. Nature,

2004, 431(7006): 302 - 305.

[15] Wen B, Xie X, Gao S, et al. Analyses of genetic structure of Tibeto-Burman populations reveals sex-biased admixture in southern Tibeto-Burmans. Am J Hum Genet, 2004, 74(5): 856 - 865.

[16] Cinnioglu C, King R, Kivisild T, et al. Excavating Y chromosome haplotype strata in Anatolia. Hum Genet, 2004, 114(2): 127 - 148.

[17] Zhong H, Shi H, Qi X B, et al. Global distribution of Y chromosome haplogroup C reveals the prehistoric migration routes of African exodus and early settlement in East Asia. J Hum Genet, 2010, 55(7): 428 - 435.

[18] Dulik M C, Osipova L P, Schurr T G. Y chromosome variation in Altaian Kazakhs reveals a common paternal gene pool for Kazakhs and the influence of Mongolian expansions. PLoS One, 2011, 6(3): e17548.

[19] Hammer M F, Karafet T M, Park H, et al. Dual origins of the Japanese: common ground for hunter-gatherer and farmer Y chromosomes. J Hum Genet, 2006, 51(1): 47 - 58.

[20] Giacomo F D, Luca F, Popa L O, et al. Y chromosomal haplogroup J as a signature of the post-neolithic colonization of Europe. Hum Genet, 2004, 115(5): 357 - 371.

[21] Sengupta S, Zhivotovsky L A, King R, et al. Polarity and temporality of high-resolution Y chromosome distributions in India identify both indigenous and exogenous expansions and reveal minor genetic influence of Central Asian pastoralists. Am J Hum Genet, 2006, 78(2): 202 - 221.

[22] Rootsi S, Zhivotovsky L A, Baldovic M, et al. A counter-clockwise northern route of the Y chromosome haplogroup N from Southeast Asia towards Europe. Eur J Hum Genet, 2007, 15(2): 204 - 211.

[23] Shi H, Dong Y L, Wen B, et al. Y chromosome evidence of southern origin of the East Asian-specific haplogroup O3-M122. Am J Hum Genet, 2005, 77(3): 408 - 419.

[24] Wen B, Xie X, Gao S, et al. Analyses of genetic structure of Tibeto-Burman populations reveals sex-biased admixture in southern Tibeto-Burmans. Am J Hum Genet, 2004, 74(5): 856 - 865.

[25] Su B, Xiao C, Deka R, et al. Y chromosome haplotypes reveal prehistorical migrations to the Himalayas. Hum Genet, 2000, 107(6): 582 - 590.

[26] Shi H, Zhong H, Peng Y, et al. Y chromosome evidence of earliest modern human settlement in East Asia and multiple origins of Tibetan and Japanese populations. BMC Biol, 2008, 6: 45.

[27] Bandelt H J, Forste P R, Rohl A. Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol, 1999, 16(1): 37 - 48.

[28] Zhivotovsky L A, Underhill P A, Cinnioglu C, et al. The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. Am J Hum Genet, 2004, 74(1): 50 - 61.

[29] Excoffier L, Laval G, Schneider S. Arlequin (version 3.0): an integrated software package for population genetics data analysis. Evol Bioinform Online, 2005, 1: 47 - 50.

[30] Shi H, Zhong H, Peng Y, et al. Y chromosome evidence of earliest modern human settlement in East Asia and multiple origins of Tibetan and Japanese populations. BMC Biol, 2008, 6: 45.

[31] Semino O, Magri C, Benuzzi G, et al. Origin, diffusion, and differentiation of Y chromosome haplogroups E and J: inferences on the neolithization of Europe and later migratory events in the Mediterranean area. Am J Hum Genet, 2004, 4(5): 1023 - 1034.

[32] Su B, Xiao J, Underhill P, et al. Y chromosome evidence for a northward migration of modern humans into Eastern Asia during the Last Ice Age. Am J Hum Genet, 1999, 65 (6): 1718 - 1724.

[33] Xue Y, Zerjal T, Bao W, et al. Male demography in East Asia: a north-south contrast in human population expansion times. Genetics, 2006, 172(4): 2431 - 2439.

[34] Hudjashov G, Kivisild T, Underhill P A, et al. Revealing the prehistoric settlement of Australia by Y chromosome and mtDNA analysis. Proc Natl Acad Sci USA, 2007, 104(21): 8726 - 8730.

[35] Kayser M, Brauer S, Cordaux R, et al. Melanesian and Asian origins of Polynesians: mtDNA and Y chromosome gradients across the Pacific. Mol Biol Evol, 2006, 23(11): 2234 - 2244.

[36] 李辉. 澳泰族群的遗传结构. 上海: 复旦大学, 2005.

[37] Zerjal T, Xue Y, Bertorelle G, et al. The genetic legacy of the Mongols. Am J Hum Genet, 2003, 72(3): 717 - 721.

[38] Thangaraj K, Chaubey G, Kivisild T, et al. Reconstructing the origin of Andaman Islanders. Science, 2005, 308(5724): 996.

[39] Wen B, Xie X, Gao S, et al. Analyses of genetic structure of Tibeto-Burman populations reveals sex-biased admixture in southern Tibeto-Burmans. Am J Hum Genet, 2004, 74(5): 856 - 865.

[40] Wen B, Li H, Lu D, et al. Genetic evidence supports demic diffusion of Han culture. Nature, 2004, 431(7006): 302 - 305.

[41] Li D, Li H, Ou C, et al. Paternal genetic structure of Hainan aborigines isolated at the entrance to East Asia. PLoS One, 2008, 3(5): e2168.

[42] Cai X Y, Qin Z D, Wen B, et al. Human migration through Bottlenecks from Southeast Asia into East Asia during Last Glacial Maximum revealed by Y chromosomes. PLoS One, 2011, 6(8): e24282.

[43] 蔡晓云. Y染色体揭示的早期人类进入东亚和东亚人群特征形成过程. 上海: 复旦大学, 2009.

[44] Semino O, Magri C, Benuzzi G, et al. Origin, diffusion, and differentiation of Y chromosome haplogroups E and J: inferences on the neolithization of Europe and later migratory events in the Mediterranean area. Am J Hum Genet, 2004, 74(5): 1023 - 1034.

[45] Shou W H, Qiao E F, Wei C Y, et al. Y chromosome distributions among populations in Northwest China identify significant contribution from Central Asian pastoralists and lesser influence of western Eurasians. J Hum Genet, 2010, 55(5): 314 - 322.

[46] Semino O, Passarino G, Oefner P J, et al. The genetic legacy of Paleolithic Homo sapiens sapiens in extant Europeans: a Y chromosome perspective. Science, 2000, 290 (5494): 1155 - 1159.

[47] Underhill P A, Passarino G, Lin A A, et al. The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. Ann Hum Genet, 2001, 65 (Pt 1): 43 - 62.

[48] 文波. Y 染色体、mtDNA 多态性与东亚人群的遗传结构. 上海：复旦大学，2004.

[49] 覃振东. 东亚西南部人群线粒体 DNA 研究：藏族和孟高棉语族人群的母系遗传多样性. 上海：复旦大学，2010.

[50] Abdulla M A, Ahmed I, Assawamakin A, et al. Mapping human genetic diversity in Asia. Science, 2009, 326(5959): 1541 - 1545.

[51] Chu J Y, Huang W, Kuang S Q, et al. Genetic relationship of populations in China. Proc Natl Acad Sci USA, 1998, 95(20): 11763 - 11768.

[52] Rootsi S, Zhivotovsky L A, Baldovic M, et al. A counter-clockwise northern route of the Y chromosome haplogroup N from Southeast Asia towards Europe. Eur J Hum Genet, 2007, 15(2): 204 - 211.

[53] Zhong H, Shi H, Qi X B, et al. Extended Y chromosome investigation suggests postglacial migrations of modern humans into East Asia via the northern route. Mol Biol Evol, 2011, 28(1): 717 - 727.

[54] Rootsi S, Magri, Kivisild T, et al. Phylogeography of Y chromosome haplogroup I reveals distinct domains of prehistoric gene flow in europe. Am J Hum Genet, 2004, 75(1): 128 - 137.

[55] Al-Zahery N, Semino, Benuzzi G, et al. Y chromosome and mtDNA polymorphisms in Iraq, a crossroad of the early human dispersal and of post-Neolithic migrations. Mol Phylogenet Evol, 2003, 28(3): 458 - 472.

[56] Nasidze I, Sarkisian T, Kerimov A, et al. Testing hypotheses of language replacement in the Caucasus: evidence from the Y chromosome. Hum Genet, 2003, 112(3): 255 - 261.

[57] Sahoo S, Singh A, Himabindu G, et al. A prehistory of Indian Y chromosomes: evaluating demic diffusion scenarios. Proc Natl Acad Sci USA, 2006, 103(4): 843 - 848.

[58] Semino O, Passarino G, Brega A, et al. A view of the neolithic demic diffusion in Europe through two Y chromosome-specific markers. Am J Hum Genet, 1996, 59(4): 964 - 968.

[59] 徐智. 中国西北地区古代人群的 DNA 研究. 上海：复旦大学，2008.

[60] Karafet T M, Osipova L P, Gubina M A, et al. High levels of Y chromosome differentiation among native Siberian populations and the genetic signature of a boreal hunter-gatherer way of life. Hum Biol, 2002, 74(6): 761 - 789.

[61] 韦兰海. 遗传学证据不支持阿尔泰语人群的共同起源. 现代人类学通讯，2011，(5)：98 - 106.

[62] Dulik M C, Osipova L P, Schurr T G. Y chromosome variation in Altaian Kazakhs reveals a common paternal gene pool for Kazakhs and the influence of Mongolian expansions. PLoS One, 2011, 6(3): e17548.

[63] 赖洪波. 论清代伊犁多民族移民开发及其历史意义. 伊犁师范学院学报(社会科学版)，2010，(4)：33 - 42.

5.2 阿尔泰语人群没有共同起源

阿尔泰语人群广泛分布在欧亚大陆北部、中亚以及近东。这些人群的形成和扩张，对整个欧亚大陆族群的分布产生了重大的影响。欧亚大陆中部地带的人群与欧亚大陆边沿的各种古代文明的交往史，是人类历史中最重要的一部分。特别的，在欧亚大陆东

部，东亚主要的居民汉族与北亚的各个族群的交流与斗争，构成了整个东亚历史中最重要的一部分。对戎狄、匈奴、突厥、蒙古和满-通古斯等族群的研究，一向是中国学术界乃至世界学术界关注的焦点。

5.2.1　阿尔泰语系是否成立的争论

在语言学方面，阿尔泰语最初是被包含在“乌拉尔-阿尔泰语”之中的。经过G. J. Ramstedt、W. Kotwicz、N. Poppe和P. Pelliot等学者的研究，阿尔泰语被认可为可以成立的一个语系。但反对的声音从一开始就没有停止过，如W. Kotwicz、G. Clauson、G. Doerfer和A. Rona-tas等一大批语言学家，总的观点是：“阿尔泰诸语言之间的所有相同和相似的成分都是由接触和借贷关系导致的，而不是同源关系的表现。”[1]

目前主张“接触关系”观点占据上风，但并不妨碍“Altaic”这个词汇被频繁地使用于指代突厥-蒙古-满-通古斯诸语言的共同体，其原因如下。

(1) 阿尔泰诸语言的同构和对应现象是明显的，这种亲缘关系的紧密程度，远远大于这些语言与其他语言的亲密程度。

(2) 虽然已建立的有同一起源的印欧语系是历史比较语言学的典型范例，但目前还没有一个语言学家明确主张“仅仅是接触关系就一定不可以成立一个语系”这样的观点，反对阿尔泰诸语言是同源关系的研究者也没有给出，如果阿尔泰语不能成立，那么这些语言应该怎样重新划分。

(3) 由于历史上的紧密联系，在有关突厥-蒙古-满-通古斯诸民族的学术研究过程中，“Altaic阿尔泰语人群”这个概称的使用，仍然是非常方便的。

在民族学方面，突厥语诸民族、蒙古语诸民族和满-通古斯语诸民族的独立起源还是比较清晰的。但是无论是在诞生阶段还是在后期的发展阶段，3大民族群体间的交流都非常频繁。例如，米努辛斯克盆地上自远古以来的蒙古人种和印欧人种的深度接触，哈萨克中帐(克烈、乃蛮)的蒙古-通古斯起源，蒙古人中众多通古斯起源的部落等。

5.2.2　阿尔泰语人群父系遗传结构

乌拉尔语-阿尔泰语人群中的主要Y染色体单倍群是C3、C3c、N、Q和R1a1。本研究收集了涉及几乎所有欧亚北部人群的相关研究文献中的数据，然后根据各个单倍群在人群中的比例，用Surfer软件画出各个单倍群的地理分布。据此探讨这些单倍群的扩散和现代北亚人群的起源关系。

1. Y-SNP单倍群C3-M217的起源与扩散

C3在阿尔泰语诸族中，是一个主要的Y染色体单倍群。主要出现在东部哈萨克人(中玉兹)、卡尔梅克人、蒙古人和通古斯人中。它在通古斯人中尤为高频。

C的年代非常古老，是现代人走出非洲(约6万年)后诞生的最初几个父系单倍群之一，经过沿海岸线的长途迁徙，分别诞生了印度南部的C5，印度尼西亚/南岛的C2，日本的C1，

澳大利亚的 C4 和东亚的 C3。C3 在东南亚、华南的部分现代人群中也有较高的比例。由于年代极其古老，我们相信末次盛冰期对于 C3 的分布有过强烈的影响，并模糊了它的史前迁徙，大致推断如图 5－14 的红色箭头所示。本节中主要讨论末次盛冰期以后的迁徙。

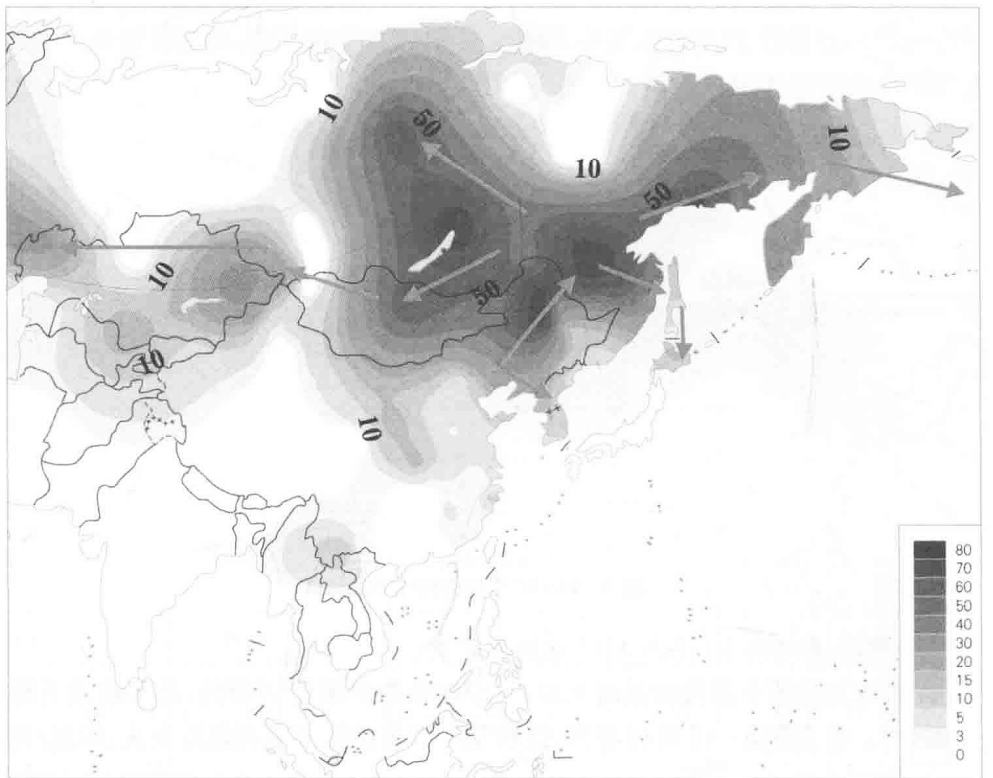


图 5－14 Y－SNP 单倍群 C3－M217 在北亚的扩散

从图 5－14 可以得出如下结论。

(1) 内蒙古兴安盟附近的人群中 C3 高频存在。这里居住着蒙古族、鄂温克族和鄂伦春族。C3 的比例接近或超过 80%。这是一个很高的比例，因此可以说 C3 是蒙古-通古斯语人群的一个典型单倍群，在哈萨克人、乌兹别克人和卡尔梅克人中也有较高的比例。这些人群的起源都与古代的蒙古部落有关。

(2) C3 在库页岛的尼夫赫人(Nivkh，38%)和北海道的阿伊努人(Ainu，25%)中也有存在。这提示 C3 沿黑龙江下游来到库页岛，最后到达北海道的迁徙路径。尼夫赫人的语言不属于阿尔泰语。

总的说来，在末次盛冰期以后，C3 在黑龙江中下游得到扩散，迁徙的末端包括纳-德内语人群和北海道阿伊努人。在新石器时代以后，C3 伴随蒙古-通古斯人的扩张得到强烈的扩散，基本奠定了现有的分布格局。

2. Y－SNP 单倍群 C3c－M48 的起源与扩散

C3c 是 C3 的一个下游支系，C3c 是在北亚经历过强烈扩张的一个单倍群。C3c 的分

布(图 5 - 15)表明,C3c - M48 是通古斯语人群的主要父系,但也高频出现在说突厥语的哈萨克人和说蒙古语的卡尔梅克人中。这些人群的起源与鞑靼人有关。根据历史记载,鞑靼人可能是与蒙古部落有不同起源的人群。

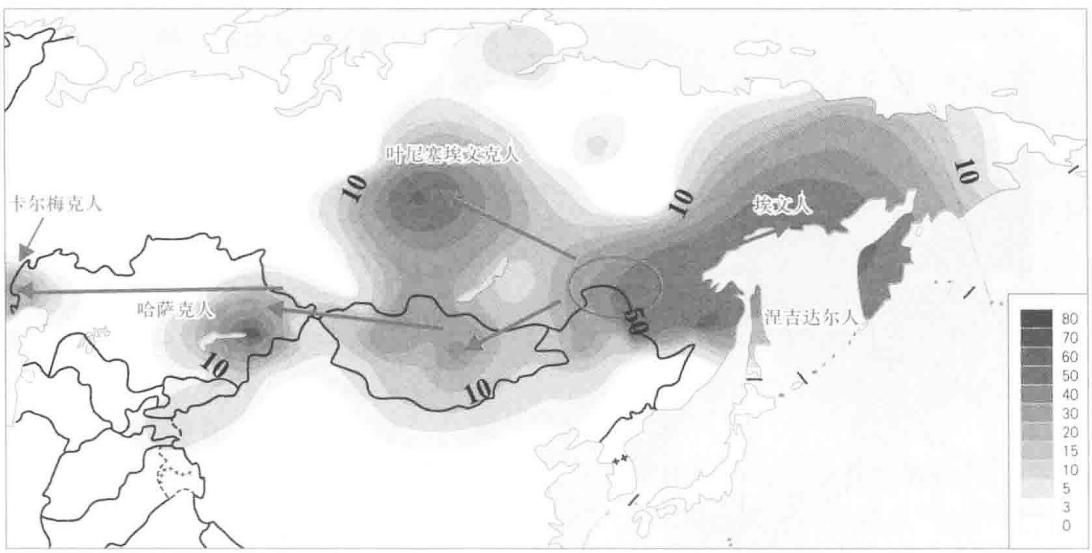


图 5 - 15　C3c 的地理分布图

3. Y - SNP 单倍群 R1a1a - M17 起源与扩散

R1a1 被认为起源于里海和黑海北岸,与印欧语的扩散密切相关,是印欧语人群的特征单倍群[2-4]。根据图 5 - 16 可以看到,这种父系类型高频出现在斯拉夫人、印度-雅利安

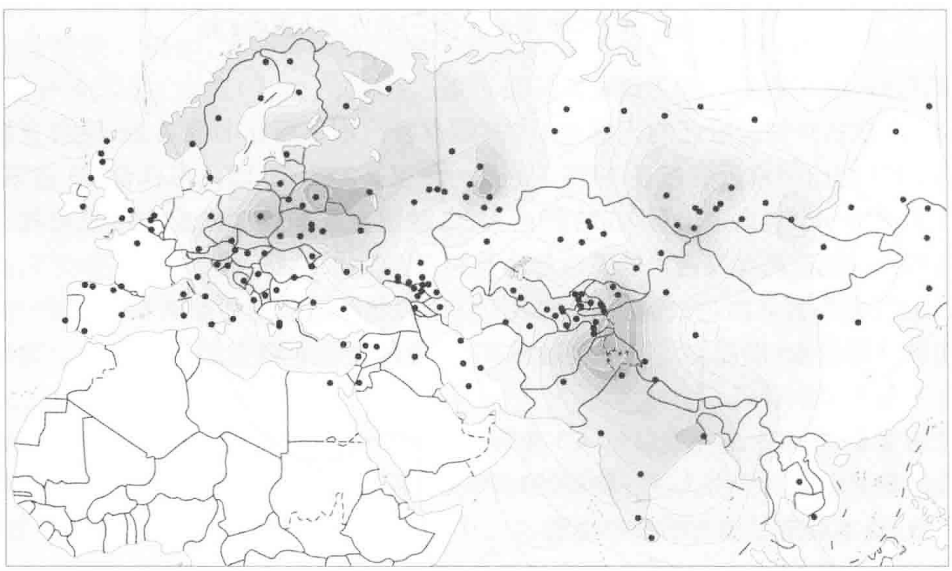


图 5 - 16　印欧语人群中高频的父系类型 R1a1 的分布(源自 Kivisild 的资料)

人、德拉维德人，以及突厥语人群和芬-乌戈尔语人群中。

在阿尔泰山区，R1a1 也出现在阿勒泰人、Altai-Kizhi 人、索尔(Shors)人、黠戛斯(Khakassian)人、Todjin 人、索约特(Sojot)人和帖良古特(Teleut)人等突厥语人群中(表 5-3)[5]。

体质人类学的证据表明，阿凡纳羡沃文化和安德罗诺沃文化的遗骸的颅骨，均属于欧罗巴大人种的原始欧洲人种类型。而后出现了东部地中海人种类型和中亚两河类型。分子人类学也给出了确定的答案。C. Bouakaze 在他的两个研究报告里，报道了阿尔泰山附近 Krasnoyarsk 地区的 9 例古代 DNA 的 Y-SNP 测试结果[6,7]。在这中间，有 8 例包含 R1a1，分别来自 Andronovo、Tagar 和 Tachtyk 文化的遗骸。

表 5-3 南西伯利亚地区诸人群的 Y-SNP 结构[5]

人群	样本数	单倍群(%)						
		P*	R1*	R1a1	N*	N3	C	K*
Altaians-Kizhi 人	92	28.3	1.1	41.3	2.2	5.4	13.0	0.0
帖良古特人	47	0.0	12.8	55.3	0.0	10.6	8.5	0.0
Khakassian 人	53	7.6	7.6	28.3	28.3	13.2	5.7	5.7
索尔人	51	2.0	19.6	58.8	13.7	2.0	2.0	0.0
Todjins 人	36	22.2	2.8	30.6	2.8	11.1	8.3	13.9
索约特人	34	8.8	0.0	23.5	8.8	11.8	17.6	26.5
布里亚特人	238	1.7	0.8	2.1	1.3	18.9	63.9	8.8
喀尔梅克人	68	11.8	2.9	5.9	2.9	0.0	70.6	4.4
鄂温克人	50	0.0	6.0	14.0	18.0	16.0	40.0	0.0
图法拉尔人	32	3.1	12.5	12.5	34.4	25.0	6.3	3.1
图瓦人	113	35.4	0.9	17.7	14.2	9.7	7.1	8.9
蒙古人	47	4.3	4.3	2.1	6.4	2.1	57.4	21.3
韩国人	83	0.0	0.0	0.0	0.0	0.0	12.0	86.7
俄罗斯人	414	2.2	6.8	48.3	0.2	14.0	0.2	1.7

4. Y-SNP 单倍群 N-M231 的起源与扩散

Y-SNP 单倍群 N-M231 主要分布在欧亚北部人群，比如因纽特人、雅库特人、黠戛斯人、图法拉尔(Tofalar)人、恩加纳桑(Ngnasan)人、汉特-曼西人以及东北欧所有的乌戈尔语人群。可见，N 完全适应了寒带的气候，并且在欧亚北部甚至北极地区得到强烈的扩散，其主要支系是 N1b-P43 和 N1c-M178(图 5-17)。

Rootsi 等[8]基于强有力的证据说明，N 很有可能起源于中国境内，1.2 万～1.4

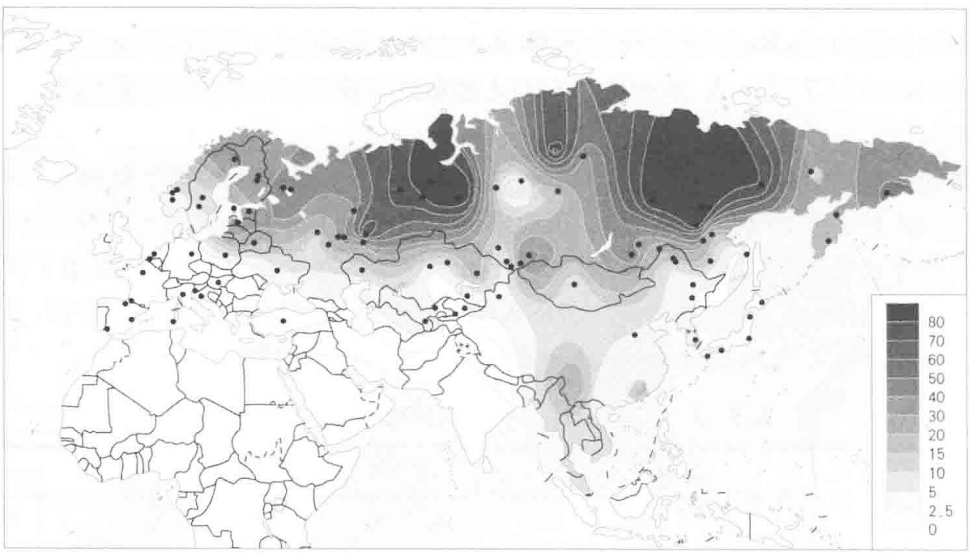


图5-17　N-M231在欧亚大陆北部的分布[8]

万年前从西伯利亚南部开始迁徙，最后高频出现于东北欧的乌拉尔语人群中。频率最高的下游单倍群N3可能起源于今天的中国，然后在西伯利亚经历多次瓶颈效应，最后扩散到东欧。另一个下游单倍群N2主要分布于Nenet、Khants-Mansi、恩加纳桑、阿尔泰山北麓的人群和部分突厥语人群中。在东北部的雅库特人中，N1c1-M178接近100%。

5. Y-SNP单倍群Q-M242的起源与扩散

Y-SNP单倍群主要出现在阿尔泰山周围人群以及中亚人群中，在羯人(Kets)(93.8%)和Selkups(66.4%)中比例较高[9]。羯人和Selkups被认为是来自目前居住地的南方——萨彦岭斜坡地区(羯人所说的语言是一种孤立语，又有分类法称之为叶尼塞语系)。从Seielstad等[10]报道的欧亚大陆各人群中Q的比例看，Q在突厥语人群中的比例也是不小的，在图瓦(Tuvinian)人、土库曼(Turkmen)人和乌兹别克(Uzbek/Tashkent)人中分别达17%、10%和14%。与中亚的可萨汗国有起源关系的Ashkenazi Jews[11]中，Q也达到5.2%。土耳其的一个人群中也有高达6%的Q。因此认为，突厥人的扩张，伴随着一定比例的Q的扩张。

6. 阿尔泰语诸人群的父系遗传结构有很大差异

从图5-14至图5-17可以看出，现代阿尔泰语人群各语族人群的父系遗传结构有重大的差异。C3c主要在通古斯人中极端高频，可以认为是这个语族人群的主要父系。C3d和C3c构成了蒙古语族人群的主要父系，但是其他类型也存在一定的比例。在突厥语人群中，除了哈萨克人以外，主要的父系是N、R1a1、Q以及其他西部欧亚单倍群。M. Derenko等的研究表明，南西伯利亚地区的部分突厥语和蒙古语人群虽然居住地非常临近，但父系Y-SNP的类型比例的差别非常大。这表明了这些群体起源上的重大差异。

重要的是，这些单倍群有完全不同的起源地和发展历史。根据现有的研究，C3、R1、N 和 Q 这些单倍群，至少已经分离了 3 万年。

此外，欧亚大陆北部人群的非阿尔泰语群体，在父系遗传上与阿尔泰语人群并非截然不同，但没有参与阿尔泰语共同体的形成过程中。这些人群主要有乌拉尔语人群和古亚细亚语人群。乌拉尔语人群的主要父系是 N－M231 的下游支系。说叶尼塞语的羯人的主要父系是 Q。西北太平洋沿岸的古亚细亚语人群有楚克奇人、科里亚克人、伊捷尔缅人、尤卡吉尔人和尼夫赫人。这些人群的主要父系类型是 N、Q 和 C。这些父系类型，同样也是构成人群阿尔泰语诸民族的主要父系。

所以，从父系遗传的 Y－SNP 证据来看，现代阿尔泰语人群的父系是漫长的历史中来的不同地区的人群混合的结果。绝大多数阿尔泰语人群在父系遗传上没有共同的起源。

5.2.3 阿尔泰语人群母系遗传结构

研究表明，通古斯语人群的母系与突厥语人群的母系差异巨大。埃文人、鄂温克人和鄂伦春人的线粒体 DNA 类型几乎完全由 C、D 和 G 组成。中亚的突厥语人群都有很高（部分超过 50%）的西部欧亚母系类型[12,13]。蒙古人的母系类型，则正好位于通古斯人和中亚突厥语人群的中间位置[14]。

通过初步研究，找到了一个伴随着阿尔泰语人群扩张的小单倍群，如图 5－18 所示。

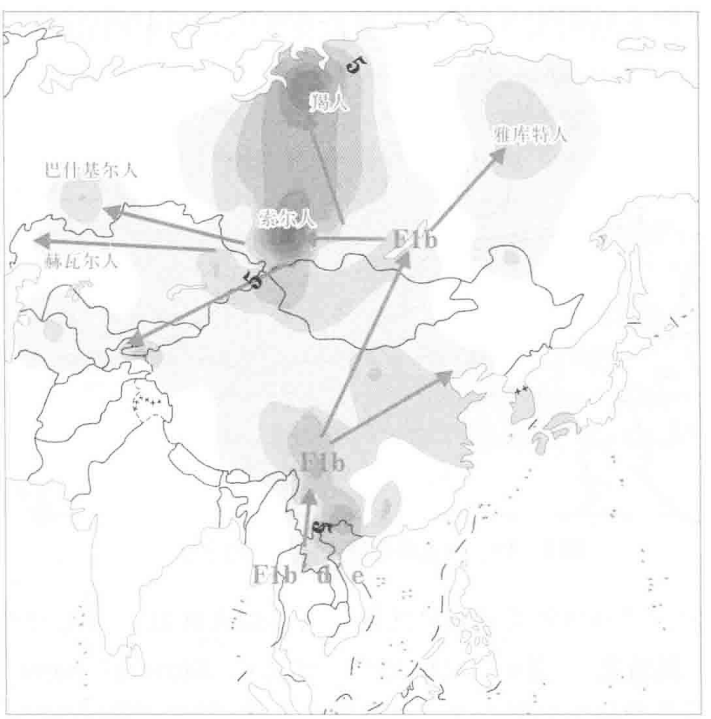


图 5－18 F1b′d′e 的分布和扩散

F1b′d′e是北亚人群中一个比例很小的母系类型。它仅仅大比例地出现在以下人群中：俄罗斯哈卡斯共和国的黠戛斯人(Khakassians,17%)和索尔人(Shors,40.2%)，叶尼塞河中下游的羯人(Kets，38%)。其次，它也较高频地出现在阿勒泰人(Altaians,7.3%)、巴什基尔人(Bashkirs,5.8%)、图瓦人(Tuvinians,7.6%)、图法拉尔人(Tofalar,8.7%)、塔吉克人(Tajikistan,9.4%)、蒙古人(5%～8%)以及克罗地亚赫瓦尔(Hvar)岛上的赫瓦尔人(8.3%)中。在其他人群中，F1b的比例很低。在蒙古人中，有约5%的比例。在通古斯语人群中，F1b的比例极少，或者不存在。

根据F1b′d′e的高变1区序列分析，可以得出推论：F1b在中亚的扩散，伴随着突厥语的扩散；而且这种扩散源自阿尔泰山与贝加尔湖之间的人群。F. P. Mooder等研究指出，贝加尔湖西北岸Lokomotiv墓葬的7 000年前的采集-狩猎人群中也存在高频的F(48.4%)，表明还可以将这种扩散追溯到更久远的时代。

F1b仅仅是北亚母系中的极小一部分，但也反映了它在阿尔泰语3个语族人群的重大差异。

5.2.4　阿尔泰语诸人群的来源与阿尔泰语系的起源

综合前述诸多数据和论述推测“阿尔泰语”诸人群，即突厥-蒙古-通古斯语诸人群的父系起源大致如图5-19所示。红色字体标记了不同的父系来源在漫长的历史时期融合成为现代人群的大致方向。现代说“阿尔泰语”的各个人群，并没有如同说印欧语的人群那样，有一个共同的父系起源(尽管后者在晚期也曾经历过跨人群的扩散)[2-4]。

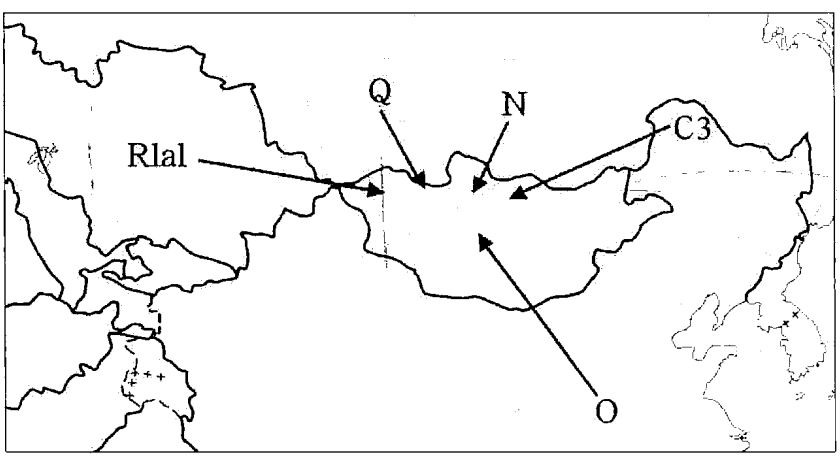


图5-19　阿尔泰语诸人群父系的大致起源

研究人员测试了2 000年前匈奴时代额金河墓地人群的Y染色体单倍群[16]。共测试6例，有3例得到结果，分别是N3、Q和C。根据C. Keyser-Tracqui等[17]提供的Y-STR，加上各个Y染色体单倍群下Y-STR的特异性，笔者根据经验值推测墓葬第三部分的父系是C3*，而且C. Keyser-Tracqui等所提到70号遗骸属于R1a1。可见图5-19

所揭示的这样一种混合，最迟在匈奴时代全然已经存在了。

通过父系单倍群的分布和 STR 网络图，说明“阿尔泰语”诸人群的主要父系（C3、C3c、N、Q 和 R1a1）有着各自独立的起源，并且与该语系之外的古亚细亚语人群、乌拉尔语人群和印欧语人群有关。

语言学的研究表明，现代人类的语言，绝大多数情况下是继承自父系群体的。在欧亚大陆北部，特别是阿尔泰语诸人群中，目前还没有证据表明存在过纯粹的母系社会群体。大致可以认为北亚人群的父系遗传结构与语言谱系有显著的相关性。既然目前构成阿尔泰语诸人群的主要父系类型有着完全不同的起源地和扩散历史，我们有理由相信，所谓阿尔泰语系是包含着各种远古语言的类群，即阿尔泰语系没有一个共同的起源。

综上所述，分子人类学的数据表明，遗传学证据不支持阿尔泰语人群的共同起源。

参考文献

[1] 孟达来. 北方民族的历史接触与阿尔泰诸语言共同性的形成. 北京：中国社会科学出版社，2001.

[2] Wells R S, Yuldasheva N, Ruzibakiev R, et al. The Eurasian heartland: a continental perspective on Y chromosome diversity. Proc Natl Acad Sci USA, 2001, 98: 10244 - 10249.

[3] Gimbutas M. Proto-Indo-European culture: the Kurgan culture during the fifth, fourth, and third millenia BC//Cardona G, Hoenigswald H M, Senn A. Indo-European and Indo-Europeans. Philadelphia: University of Pennsylvania Press, 1970: 155 - 195.

[4] Underhill P A, Myres N M, Rootsi S, et al. Separating the post-Glacial coancestry of European and Asian Y chromosomes within haplogroup R1a. Eur J Hum Genet, 2009, 18: 479 - 484.

[5] Derenko M, Malyarchuk B, Denisova G A, et al. Contrasting patterns of Y chromosome variation in South Siberian populations from Baikal and Altai-Sayan regions. Hum Genet, 2006, 118(5): 591 - 604.

[6] Bouakaze C, Keyser C, Amory S, et al. First successful assay of Y-SNP typing by SNaPshot minisequencing on ancient DNA. Int J Legal Med, 2007, 121: 493 - 499.

[7] Keyser C, Bouakaze C, Crubézy E, et al. Ancient DNA provides new insights into the history of south Siberian Kurgan people. Hum Genet, 2009, 126(3): 395 - 410.

[8] Rootsi S, Zhivotovsky L A, Baldovic M, et al. A counter-clockwise northern route of the Y chromosome haplogroup N from Southeast Asia towards Europe. Eur J Hum Genet, 2007, 15: 204 - 211.

[9] Karafet T M, Osipova L P, Gubina M A, et al. High levels of Y chromosome differentiation among native Siberian populations and the genetic signature of a boreal hunter-gatherer way of life. Hum Biol, 2002, 74: 761 - 789.

[10] Seielstad M, Yuldasheva N, Singh N, et al. A Novel Y chromosome variant puts an upper limit on the timing of first entry into the Americas. Am J Hum Genet, 2003, 73: 700 - 705.

[11] Behar D M, Garrigan D, Kaplan M E, et al. Contrasting patterns of Y chromosome variation in Ashkenazi Jewish and host non-Jewish European populations. Hum Genet, 2004, 114:

354 - 365.

[12] Irwin J A, Ikramov A, Saunier J, et al. The mtDNA composition of Uzbekistan: a microcosm of Central Asian patterns. Int J Legal Med, 2010, 124: 195 - 204.

[13] Comas D, Plaza S, Wells R S, et al. Admixture, migrations, and dispersals in Central Asia: evidence from maternal DNA lineages. Eur J Hum Genet, 2004, 12: 495 - 504.

[14] Pakendorf B, Wiebe V, Tarskaia L A, et al. Mitochondrial DNA evidence for admixed origins of central Siberian populations. Am J Phys Anthropol, 2003, 120: 211 - 224.

[15] Mooder K P, Schurr T G, Bamforth F J, et al. Population affinities of Neolithic Siberians: a snapshot from prehistoric Lake Baikal. Am J Phys Anthropol, 2006, 129: 349 - 361.

[16] Petkovski E. Polymorphismes ponctuels de sequence et identification génétique. Thèse présentée pour obtenir le grade de Docteur, de l'Université Louis Pasteur Strasbourg I, 2006.

[17] Keyser-Tracqui C, Crubézy E, Ludes B. Nuclear and mitochondrial DNA analysis of a 2 000-year-old Necropolis in the Egyin Gol Valley of Mongolia. Am J Hum Genet, 2003, 73: 247 - 260.

5.3 日本列岛人类遗传多样性分布

在日本的史前时期，至少有过两批不同人群向日本列岛迁徙。第一批人群的迁徙开始于约5万年前，在日本渐渐发展，并在约1万年前达到群体文化的顶峰，形成了日本的绳文文化。这一批移民的代表是出土于琉球的港川人[1]。第二批移民约在距今2 300年前到达日本列岛，并形成了日本的弥生文化。根据化石和人类遗骸证据，弥生文化接下来主导了日本列岛，在公元300年左右完成了其扩张[2]。在同一时期，除了北海道最北边以及琉球最南边之外的几乎所有人类定居点内都发现了农业的痕迹，这与弥生人在这一时期的广泛扩张相吻合。但是，体质人类学对颅形的研究并不支持弥生人完全取代绳文人的解释模型[3]。因而，从19世纪起许多学者便提出了若干关于现代日本人起源与扩张的不同解释模型[4]。

自20世纪80年代发明聚合酶链反应起，人类遗传学及分子人类学快速发展。分子人类学最成功的案例之一便是利用Y染色体、线粒体DNA及常染色体证据确证了现代生活在非洲以外的现代人均起源于非洲，并有少量来自早期生活在欧亚大陆上的智人（即尼安德特人及丹尼索瓦人）的遗传混合，解开了现代人（*Homo sapiens sapiens*）的起源之谜[5,6]。另一个成功案例是通过对12 000份来自东亚地区的男性样本的Y染色体进行分析，进一步确认了生活在东亚的现代人均起源于非洲，反驳了基于古人类学材料推测的东亚地区人群独立起源假说[7]。近年来，基于常染色体的分析方法使得分子人类学家能够检验更加复杂的人群扩张模型。在本节中将利用这些分子人类学方法对日本人的父系与母系起源进行分析与检验。

5.3.1 Y染色体与线粒体DNA：研究人类进化的有力工具

分子人类学的研究对象是人类的遗传物质，其传递是连续的，且在代际传递时能够

保持相对稳定。与之相对，历史学、考古学与体质人类学的研究对象包括文化遗存、化石及人类遗骸，具有不连续、容易受环境影响、经常难以辨认等特点。另外，与 DNA 标记不同，遗骸和文物表面上的相似特征并不直接意味着发生学关系。因此，从证据可靠性的角度上来看，分子人类学得出的结论要比历史学、考古学及体质人类学得出的结论更加可靠。

人类的遗传物质包括细胞核内的染色体及细胞核外的线粒体 DNA 两类。Y 染色体和线粒体 DNA 是分子人类学家最常使用的两类研究材料。它们的优点包括不发生重组、分别随父系与母系遗传、有效群体大小较小等，因而较易形成群体特异的遗传标记，且能够方便地揭示被研究群体的父系与母系遗传结构[8]。

1. Y 染色体非重组区(NRY)的随父遗传

与常染色体及 X 染色体不同，Y 染色体的绝大部分在代际传递时都不会发生重组，被称为 Y 染色体非重组区[9]。此外，由于只有男性才携带 Y 染色体，理论上其有效群体数量是常染色体的 1/4。因而与由常染色体揭示的群体遗传结构相比，由 Y 染色体揭示的群体遗传结构将更易受遗传漂变的影响。

如前所述，由于较易受遗传漂变影响，Y 染色体上较易形成群体特异的遗传标记。Y 染色体单核苷酸多态性(Y-SNP)是一类较常使用的 Y 染色体遗传标记，在 Y 染色体上分布广泛，且非常可靠。利用高效液相色谱法，Underhill 等[10]发现了约 160 个二态(bi-allelic)的 Y-SNP。近年来，通过对各个 Y 染色体单倍群进行测序，有超过 15 000 个 Y-SNP 被发现。由于 Y 染色体的碱基突变率较低，Y-SNP 可能难以反映较为晚近的人群扩张事件。我们可以使用突变速率较高的 Y 染色体短串联重复(Y-STR)来填补 Y-SNP 留下的空缺，用以绘制进化的网络图及计算 Y-SNP 的产生时间等。

综上所述，将 Y-SNP 与 Y-STR 信息相结合，Y 染色体非重组区可以用来研究人类群体的扩张与迁徙。

2. 线粒体 DNA 的随母传递

人类的线粒体 DNA 是环状的，包含 16 659 个碱基对，并有一个长度约为 1 100 个碱基对的非编码控制区(control region)，该区域突变率较高，又被称为高变区[11,12]。尽管高变区相对较高的突变率为研究人类进化提供了方便，但其较高的突变率也可能引入回复突变，可能在进化分析中造成偏差[13,14]。

在受精过程中，由于携带线粒体的精子尾部是不进入卵细胞的，这就意味着父系的线粒体 DNA 无法遗传给后代。也就是说，线粒体 DNA 遵循严格的随母遗传规律，它在人类进化研究中的特性与随父遗传的 Y 染色体非重组区是非常类似的。

5.3.2 人类向日本列岛迁徙的历史

1. 石器时代的绳文文化

人类向日本列岛的首次迁徙开始于距今约 5 万年前。日本列岛上出土的人类化石较少，生活在距今约 18 000 年前的港川人是目前出土的存在时间最早的人类化石[15]。对港川一号男性头骨的形态学分析显示，它与中国广西的柳江人间的关系比与山顶洞人

的关系更近[16]。实际上，观察最有特征性的眶骨形状和鼻根形态，港川人更像山顶洞人（图 5－20）。这些证据揭示了绳文人可能就是港川人的后代[17]。

图 5－20　东亚旧石器时代 3 种现代人头骨比较

绳文文化开始于末次冰川期末期，结束于距今约 2 300 年前，在巅峰时期其范围覆盖了日本列岛的大部。值得注意的是，因为日本列岛在约 12 000 年前由于海平面的上升才与欧亚大陆分开，人类学家并不确定绳文人起源于东亚的南方还是北方。

基于体质人类学证据，绳文人与现代及石器时代的中国人均有较大差异。有意思的是，如图 5－21，一项对绳文人、现代中国人、古代中国人、韩国人、现代日本人、阿伊努人、琉球人、北亚人、弥生人、波利尼西亚人及密克罗尼西亚人的头骨测量数据进行的研究[3]显示，除阿伊努人与琉球人（这两个人群是已知的绳文人的后代）外，绳文人与现代密克罗尼西亚人及波利尼西亚人的关系最近，说明他们在人种上接近与新几内亚有交流的南岛族群。

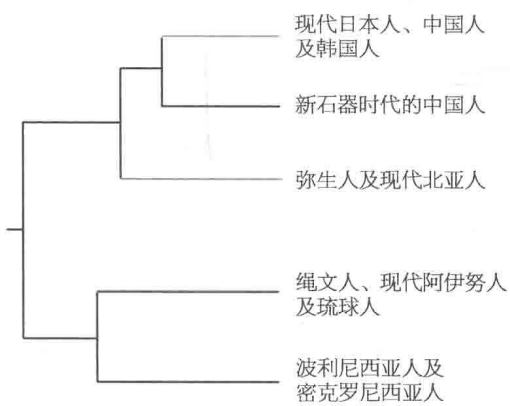


图 5－21　基于 9 种颅骨形态特征的 Q 模式回归矩阵绘出的 21 个群体的聚类图

（改自 K. Hanihara，1984）

2. 红铜时代的弥生文化

人类向日本列岛的第二次大规模迁徙开始于距今约 2 300 年前，这批人群在 3 世纪

左右完成了在日本列岛上的扩张。这次迁徙到日本列岛上的人群形成了弥生文化，这个文化对此后日本的社会发展起到了重要作用。在弥生人抵达日本前，绳文人一般生活在较小的群体中，每个人类定居点平均生活着约 24 人。与之对应的是，弥生人生活在更大的群体中，每个人类定居点平均生活着约 57 人，是绳文人的 2 倍以上[16,18]。

弥生人的起源及其迁徙的驱动力是两个颇具争议的话题。考古学及人类化石证据显示迁徙开始于绳文时代末期，约为公元前 3 世纪，并在接下来的 600 年中经历了大规模的人群扩张。迁徙的时间与气候变冷及东亚大陆发生社会动荡的时期一致。因此，弥生人迁徙的驱动力可能来自东亚大陆上的动乱以及变冷的气候。此外，在弥生人遗存中发现的类似东北亚人使用的金属工具揭示着弥生人可能来自东北亚。

与考古学证据相吻合，对颅骨形态的分析也显示弥生人与蒙古人、西伯利亚人及中国东北人群较为相似。值得注意的是，这些人群都对寒冷的气候有较好的适应。这些证据提示弥生人可能由库页岛及北海道进入日本列岛，而不一定由朝鲜半岛进入。

继弥生时代之后，古坟时代(Kofun Period)由 3 世纪持续到 6 世纪。在此期间，受日本皇室出台的鼓励引入中国及韩国更先进的文化及技术的政策影响，来自东亚大陆的移民逐渐增多。这一政策最具代表性的证据是在 5 世纪左右引入日本的中国汉字[19]。

5.3.3 现代日本人起源的 3 种理论

与其他地区人类起源的争议相类似，现代日本人起源的争议可以归结为 3 种模型：替代说、维持说及混合说(图 5-22)。这 3 种模型争议的焦点在于绳文人及弥生人对现代日本人遗传结构的贡献比例。

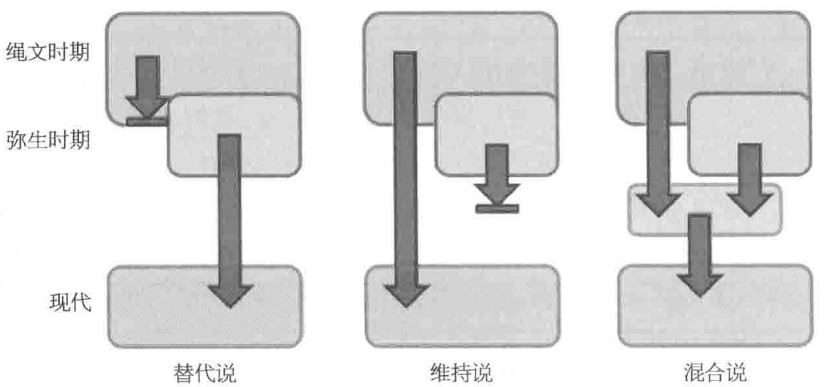


图 5-22 现代日本人起源的三种理论

图中长方形为基因库，箭头为基因流。

替代说认为从弥生文化取代绳文文化，在日本列岛占统治地位开始，绳文人就对日本人的遗传结构没有贡献了[20,21]。

与替代说相对立，维持说认为尽管弥生文化取代了绳文文化，但这种取代仅发生在文化上，并不发生在遗传结构上。因此依据维持说，现代日本人是绳文人的直接后代，与

弥生人并无遗传结构上的联系[4,22]。

混合说结合了替代说和维持说的论点，基于考古学与人类学证据，认为现代日本人是绳文人、弥生人及更加晚近的来自东亚的移民相混合的结果[19]。

上述3种理论均有遗传学或考古学证据支持，其中最近的证据倾向于支持混合说。支持替代说的证据如Cavalli-Sforza等[23]，支持维持说的证据如Nei等[24]，支持混合说的证据包括Hammer等[25]、Horai等[27]、Omoto等[28]、Sokal等[29]及Tajima等[30]。

5.3.4　来自Y染色体分析的证据

Hammer等[26]研究了日本列岛、东北亚、东南亚、中亚、南亚及大洋洲群体的Y染色体单倍群类型及分布(表5-4)。这些群体中的单倍群有C、D、N和O等，其中98.9%的日本人被划入C、D、N和O这4个单倍群中的一个，这4个单倍群也就是日本列岛人群的主要Y染色体单倍群。

表5-4　各Y染色体单倍群在日本人群体及其他参考群体中的频率分布(%)

单倍群①	日　本	东北亚	东南亚	中　亚	南　亚	大洋洲
C2a1-P33						17.2
C2*-M38			2.2			8.1
C*-RPS4Y			3.1	0.2	2.6	8.6
C3*-M217	1.9	21.3	2.7	17.4		
C3c-M86	1.2	23.6		7.9		
C1-M8②	5.4					
D1-M15			2.6	4.3		
D*-M174				2.1		
D3a-P47				9.1		
D2*-P37.1②	3.9	0.2				
D2a*-M116.1②	16.6					0.5
D2a1*-M125②	12.0	0.5				
D2a1a-P42②	2.3					
N1a-M128		0.7				
N1*-LLY22g	1.2	1.6	3.8	1.4		
N1b-P43		1.1		3.1		
N1c-M178	0.4	10.0	0.3	1.0		
NO-M214②	2.3	0.2	0.4	0.2		
O2a1a-M111			8.1			

（续表）

单倍群①	日 本	东北亚	东南亚	中 亚	南 亚	大洋洲
O2a1 * - M95	1.9	0.5	6.3		7.9	
O2 * - P31		2.3	3.2	0.5		
O2b * - SRY465	7.7	6.3	0.4			0.5
O2b1 - 47z②	22.0	0.7	0.6			
O * - M175		0.7	0.1	0.5		
O3a1c - 002611	3.1	5.2	19.8	1.9	0.2	0.5
O3 * - M122	6.6	4.8	8.8	1.4		9.1
O3a2c1 - M134	10.4	10.7	12.4	15.5	0.2	1.4
O1a * - M119		1.1	11.4	0.7		2.9
O1a2 - M110			2.3			0.5
其他	1.1	8.5	11.5	32.8	89.1	50.7

注：① Y染色体单倍群的命名依照 T. M. Karafet 等[31]（单倍群 C、D 及 N）和 S. Yan 等[32]（单倍群 O）。②日本人群特有的 Y 染色体单倍群，依照 M. F. Hammer 等[26]。

1. 单倍群 O

单倍群 O 的频率在日本主岛人群中为 46.2%～62.3%，在琉球人中约为 37.8%。在生活于北海道北部的阿伊努人中未发现单倍群 O 的分布。单倍群 O 的日本人主要分为两支，分别是携带 SNP 位点 M31 的单倍群 O2 和携带 M122 的单倍群 O3。

在单倍群 O2 中，支系 O2b1（由 SNP 47z 定义）仅高频分布于日本人及韩国人中，该支系在日本主岛人群中的分布频率比在琉球人中的频率要高。根据 Y - STR 数据用 Stumpf 等[33]提出的算法估算，单倍群 O2b1 出现于距今 5 720～12 630 年前[26]。这一年代远早于弥生时代。因为单倍群产生于东亚大陆，可能是在韩半岛，人口扩张以后才进入日本，所以计算的年代是其在韩半岛产生的年代，应该早于其进入日本的年代。

单倍群 O3 是日本人中的另一主要单倍群，同时也是现代汉族中国人中的最大的单倍群[32]。日本人中单倍群 O3 的分化年代非常晚近，约为距今 500 年或更晚。由此可以推测，日本人中的单倍群 O3 可能是由最近几百年来向日本列岛迁徙的东亚大陆人口所引入的。

2. 单倍群 D

日本人群中频率仅次于单倍群 O 的单倍群为单倍群 D，约占日本主岛人群的 34.7%。单倍群 D 的高频分布仅限于亚洲，并且在日本人和中国康藏人群中频率较高，可能提示日本人与中国康藏区有较近的亲缘关系。

单倍群 D2（由 SNP P37.1 定义）的高频分布仅限于日本人中，在韩国人及东亚其他地区人群中都未见或偶见分布。与单倍群 O 分布趋势相反，单倍群 D 在琉球人及北海道

北部的阿伊努人中的频率较高，高于日本主岛人群中的频率（图5－23，图5－24）。

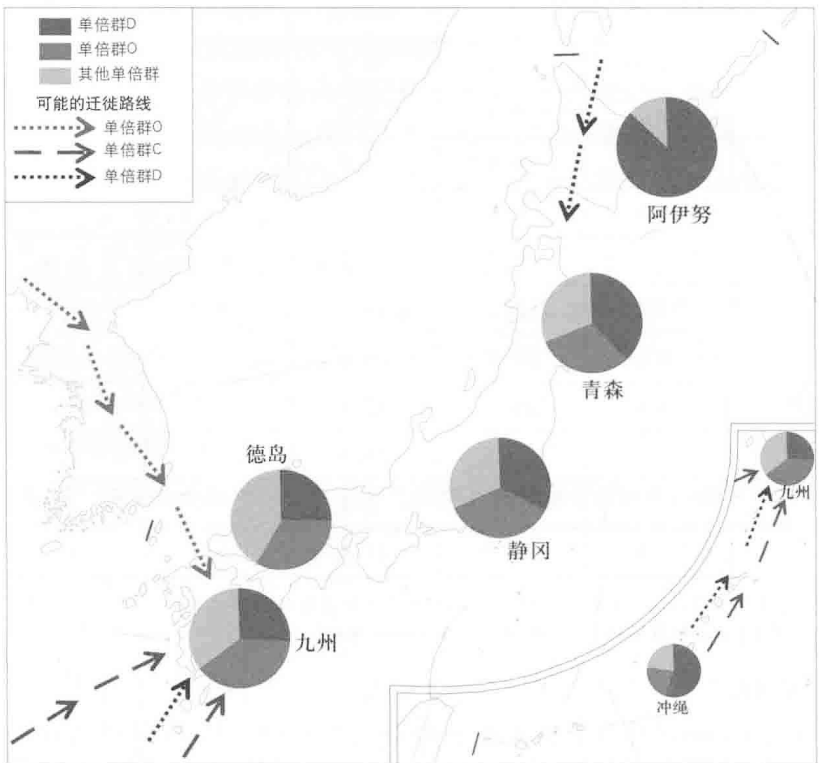


图5－23　Y染色体单倍群D和O在不同日本人群中的频率分布

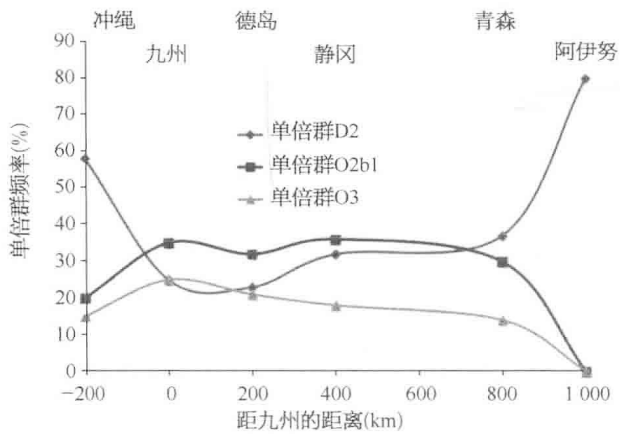


图5－24　据Y染色体单倍群频率与距九州的距离绘出的U形及倒U形的趋势图

（改自M. F. Hammer等，2006）

单倍群D2出现于距今14 060～31 050年前[26]，比单倍群O2b1的年代要早许多。这一较早的年代与距今1万～5万年前的人类向日本列岛首次迁徙的时间相吻合。因

此，单倍群 D2 与单倍群 O2b1 可能分别代表的是两个不同时期向日本列岛迁徙的人群。

3. 单倍群 C 和 N

单倍群 C 和 N 是在日本人群中频率分布第三、第四的单倍群，其频率分别为 8.5% 和 1.5%。在单倍群 C 中也可以找到一个仅在日本人中高频分布的单倍群 C1（由 SNP M8 定义）。有意思的是，单倍群 C1 未见于阿伊努人中，而在四国岛上的德岛（Tokushima）人中频率最高，这一分布趋势与单倍群 D2 和 O2b1 均不同。单倍群 C1 的出现年代距今 8 460～18 690 年前[26]。

5.3.5 来自线粒体与常染色体分析的证据

1. 来自线粒体分析的证据

根据线粒体全序列数据，现代日本人与韩国人最为接近[34]。这一结果与可能来自朝鲜半岛的弥生人及古坟人的大规模扩张相吻合。此外，也发现了来自古代东亚南方群体与北方群体混合的证据。例如，线粒体单倍群 M12 与 Y 染色体单倍群 D 在分布上较为相似，也同时高频出现于日本人及中国西藏人中，且其多样性均较高。

现代日本人、古代日本人、其他东亚人及西伯利亚人的线粒体单倍群频率分布如表 5－5，主要基于线粒体高变区测序数据计算。与 Y 染色体单倍群 C1 和 D2 相似，线粒体单倍群 N9b 和 M7a 被认为是绳文人特有的线粒体单倍群[34-36]。在表 5－5 的所有日本人群体中，阿伊努人和琉球人中的线粒体单倍群 N9b 和 M7a 的频率是最高的，而常见于现代中国人和韩国人的线粒体单倍群 A 和 D（除 D1）的频率则明显低于日本主岛人群。考虑到海平面上升使得日本列岛脱离欧亚大陆这一事件发生于绳文时代早期，线粒体单倍群 N9b 和 M7a 在西伯利亚东南人群中的高频分布也有助于分析现代人类向日本列岛的第一次迁徙过程与路径。但是，线粒体单倍群 D1 在关东（Kantō）绳文人与礼文岛船舶地区（Funadomari）绳文人中不同的频率分布提示现代人类在绳文时期可能通过不同路径向日本列岛迁徙，这也与 Y 染色体单倍群 C 和 D 的不同地理分布趋势相吻合。

表 5－5 线粒体 DNA 单倍群在绳文人群体及其他参考群体中的频率分布（%）

单倍群	礼文岛船舶绳文人①	关东绳文人①	阿伊努①	主岛日本人①	琉球人①	北方中国人①	韩国人①	Udegey 西伯利亚人①
N9b②	64.3	0	7.8	1.9	4.6	0	0.3	30.4
D1	28.6	0	0	0	0	0	0	0
M7a②	7.1	3.7	15.7	10.0	23.3	0	1.3	19.6
A	0	7.4	3.9	9.0	6.8	5.0	8.8	0
D（除 D1）	0	18.5	17.6	41.2	38.3	51.5	32.5	0
其他	0	70.4	55.0	37.9	27.0	43.5	57.1	50.0

注：① 数据来自文献[36－44]。② 绳文人特有的线粒体 DNA 单倍群。

综上所述，现代日本主岛人群中既出现了绳文人特有的线粒体单倍群N9b和M7a，又出现了东亚常见线粒体单倍群A和D(除D1)，为绳文人和弥生人曾发生人群混合提供了有力的证据。与日本主岛人群相比，线粒体单倍群N9b和M7a在北海道阿伊努人及琉球人中更高的分布频率提示，这两个人群有更少的弥生人混合，而保留了更多的绳文人的成分，这也与Y染色体分析的证据相吻合。

2. 来自常染色体分析的证据

由于常染色体携带的遗传物质远多于Y染色体和线粒体DNA，常染色体是检验复杂群体遗传学模型的理想材料。一项对东亚地区人类常染色体遗传结构的综合性研究(The HUGO Pan-Asian SNP Consortium)显示[45]，琉球人的遗传结构与其他人群有较为明显的不同。在结构分析中体现为缺少东亚大陆人群的成分，且包含较多由东亚、中亚、东南亚及大洋洲人群共享的成分。这一证据揭示了琉球人可能与日本主岛人群具有不同的起源，可能与琉球绳文人的南方来源相吻合。

此外，Yamaguchi-Kabata等[46]对多个日本人群的140 387个常染色体SNP进行了主成分分析并计算了遗传距离。从主成分分析图上可以看到，日本人群个体分布于两个不同于中国汉族人的簇中，还有一小部分日本人群个体与汉族中国人分布较为接近。琉球人分布的簇与日本主岛人群分布的簇(Hondo cluster)之间的遗传距离与日本主岛不同人群间的遗传距离相比明显要高。这一证据也支持琉球人与日本主岛人群的不同起源，同时也是混合说的有力证据。

但是，使用常染色体数据分析仍然有许多技术限制，上述两份研究中分析得出的遗传成分(如结构分析中得出的东亚特有成分)并不能被归属于某一个现代或古代群体。因此，我们无法利用常染色体数据计算绳文人与弥生人在现代日本人中的混合比例。

5.3.6　结论与展望

根据前人对东亚、东南亚及南亚群体遗传结构的分析[47-52]，可以得出以下结论。

据图5-24，以距九州的距离为横轴，单倍群D的频率分布呈U形，而单倍群O的频率分布呈倒U形。早在发现Y染色体上的详细遗传标记之前，Sokal等[29]就曾预测，假如日本人的起源符合混合说，某些特征的分布将会呈U形，而另一些特征的分布将会呈倒U形。基于大规模的Y染色体分析，这种U形与倒U形的分布趋势已被确认，成为支持混合说的证据。

日本列岛人群中单倍群C和D的年代估计与绳文时期相符，而单倍群O的年代估计与弥生时期相符。

在单倍群D频率较高的阿伊努人和琉球人中单倍群C1的频率较低的原因目前尚未明确。由于单倍群C1的年代估计(8 460～18 690年)显然晚于单倍群D2的年代估计(14 060～31 050年)，且均早于单倍群O2b1的年代估计，我们可以确定单倍群C1/D2和O2b1是由两次不同的人群迁徙所引入的。但是，考虑到单倍群C和D在日本列岛的频率分布有差异，单倍群C1和单倍群D2是由同一批人群引入还是由两批人群分别引入尚

待考证。

Y-STR 多样性和下游 SNP 标记均确认单倍群 O2 和 O3 在韩国人中的多样性高于日本人，因此可以推断弥生人向日本列岛的迁徙是经由朝鲜半岛的，而非经由库页岛北下。

在排除采样误差和采样点的不均匀分布后，基于 Y 染色体单倍群 D 和 O 的频率分布，Hammer 等[26]计算出绳文人和弥生人对现代日本人的遗传贡献分别为 40.3%与 51.9%。

线粒体 DNA 证据也支持日本人起源的混合说，但其提供的证据没有 Y 染色体清晰。通过线粒体 DNA 计算出绳文人对现代日本人的遗传贡献约为 80%。这可能是由于群体之间的战争或其他活动所引起。在战争期间，战败方的男性通常会被杀死，而女性则更有可能活下来，并且与战胜方的男性产下后代。这一假说已经在密克罗尼西亚人中得到证实。密克罗尼西亚人中有较低频率的日本人特有 Y 染色体单倍群 D2b1 分布，这可能是由于在第二次世界大战前日本对密克罗尼西亚的长期统治所致。

此外，基于线粒体 DNA 高变区数据得出的推断是非常有限且不可靠的，其标准差较大。因此，使用高变区数据估算的绳文人-弥生人混合比例也可能存在问题。为解决这些问题，需要使用多群体的线粒体全序列数据进行分析，以便提高使用线粒体 DNA 进行群体历史推断的可靠性，方能得出较为可靠的结论。

利用常染色体得出的结论也较为薄弱。我们仅能确定日本人在遗传结构上是由多个群体混合而成的。越来越多的基于 Y 染色体、线粒体 DNA 和常染色体的证据解释了现代日本列岛人群起源最符合混合说的描述。

尽管已经能够粗略地描绘出日本列岛人类起源的路线，但仍有许多谜团有待分子人类学家解决。例如，使用线粒体 DNA 估算得出的绳文人对现代日本人的贡献比例是否可靠？Y 染色体单倍群 C 和 D 是否在同时抵达日本列岛？

参考文献

[1] Hisao B, Hajime S, Shuichiro N. Human skeletal remains of Jomon period from Minamitsubo shell-mound site in Ibaraki Prefecture, East Japan. Bulletin of the National Science Museum, Tokyo, 1989, Serial D, 15: 1-40.

[2] Chard C S. Northeast Asia in prehistory. Madison, WI: University of Wisconsin Press, 1974.

[3] Hanihara K. Origins and affinities of Japanese viewed from cranial measurements. Acta Anthropogenetica, 1984, 8: 149-158.

[4] Mizoguchi Y. Contributions of prehistoric Far East populations to the population of modern Japan: a Q-mode path analysis based on cranial measurements//Akazawa T, Aikens C M. Prehistoric hunter-gatherers in Japan. Japan: University of Tokyo Press, 1986: 107-136.

[5] Vigilant L, Stoneking M, Harpending H, et al. African population and the evolution of human mitochondrial DNA. Science, 1991, 253: 1503-1507.

[6] Bowcock A M, Ruiz-Linares A, Tomfohrde J, et al. High resolution of human evolutionary trees

with polymorphic microsatellites. Nature, 1994, 368: 455 - 457.

[7] Ke Y, Bing S, Song X, et al. African origin of modern humans in East Asia: a tale of 12 000 Y chromosomes. Science, 2001, 292: 1151 - 1153.

[8] Zhang F, Su B, Zhang Y P, et al. Genetic studies of human diversity in East Asia. Philosophical Transactions of the Royal Society B, 2007, 362: 987 - 995.

[9] 王传超,严实,李辉.姓氏与Y染色体.现代人类学通讯,2010,4: 26 - 33.

[10] Underhill P A, Jin L, Lin A A, et al. Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography. Genome Research, 1997, 7: 996 - 1005.

[11] Anderson S, Bankier A T, Barrell B G, et al. Sequence and organization of the human mitochondrial genome. Nature, 1981, 290: 457 - 465.

[12] Pakendorf B, Stoneking S. Mitochondrial DNA and human evolution. Annual Review Genomics Human Genetics, 2005, 6: 165 - 183.

[13] Arctander P. Mitochondrial recombination? Science, 1999, 284: 2090 - 2091.

[14] Yao Y G, Lu X M, Luo H R, et al. Gene admixture in the Silk Road region of China: evidence from mtDNA and melanocortin 1 receptor polymorphism. Genes & Genetic Systems, 2000, 75: 173 - 178.

[15] Matsu'ura S, Kondo M. Relative chronology of the Minatogawa and the Upper Minatogawa series of human remains from Okinawa Island, Japan. Anthropological Science, Advance Publication, 2010.

[16] Suzuki H. Skulls of the Minatogawa man//Suzuki H, Hanihara K. The Minatogawa man: the upper pleistocene man from the island of Okinawa. Japan: University of Tokyo Press, 1982.

[17] Hisao B, Shuichiro N, Seiho O. Minatogawa hominid fossils and the evolution of Late Pleistocene humans in East Asia. Anthropological Science, 1998, 106, 27 - 45.

[18] Koyama S. Jomon subsistence and population. Senri Ethnological Studies, 1978, 2: 1 - 65.

[19] Hanihara K. Dual structure model for the population history of Japanese. Japan Review, 1991, 2: 1 - 33.

[20] Howells W W. The Jomon people of Japan: a study by discriminant analysis of Japanese and Ainu crania. Paper of the Peabody Museum of Archeology and Ethnology, Harvard University, 1966, 57: 1 - 4.

[21] Tuner C G. Dental evidence on the origins of the Ainu and Japanese. Science, 1976, 193: 911 - 913.

[22] Suzuki H. Racial history of the Japanese//Schwidetzky I. Rassengeschichte der Menschheit. Germany: Lieferung Oldenbourg, 1981.

[23] Cavalli-Sforza L L, Menozzi P, Piazza A. The history and geography of human genes. Princeton, NJ: Princeton University, 1994.

[24] Nei M. The origins of human population: genetic, linguistic, and archeological data//Brenner S., Hanihara K. The origin and past of modern humans as viewed from DNA. Singapore and London: World Scientific, 1995.

[25] Hammer M F, Horai S. Y chromosomal DNA variation and the peopling of Japan. American Journal of Human Genetics, 1995, 56: 951 - 962.

[26] Hammer M F, Karafet T M, Park H. Dual origins of the Japanese: common ground for hunter-gatherer and farmer Y chromosomes. Journal of Human Genetics, 2006, 51: 47 - 58.

[27] Horai S, Murayama K, Hayasaka K, et al. mtDNA polymorphism in East Asian populations, with special reference to the peopling of Japan. American Journal of Human Genetics, 1996, 59: 579 - 590.

[28] Omoto K, Saitou N. Genetic origins of the Japanese: a partial support for the dual structure hypothesis. American Journal of Physical Anthropology, 1997, 102: 437 - 446.

[29] Sokal R R, Thomson B A. Spatial genetic structure of human population in Japan. Human Biology, 1998, 70: 1 - 22.

[30] Tajima A, Pan I H, Fucharoen G, et al. Three major lineages of Asian Y chromosomes: implications for the peopling of east and southeast Asia. Human Genetics, 2002, 110: 80 - 88.

[31] Karafet T M, Mendez F L, Mellerman M B. New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroups tree. Genome Research, 2008, 18: 830 - 838.

[32] Yan S, Wang C C, Li H, et al. An updated tree of Y chromosome Haplogroup O and revised phylogenetic positions of mutations P164 and PK4. European Journal of Human Genetics, 2011, 19: 1013 - 1015.

[33] Stumpf M P H, Goldstein D B. Genealogical and evolutionary inference with the human Y chromosome. Science, 2001, 291: 1738 - 1742.

[34] Tanaka M, Cabrera V M, Gonzalez A M, et al. Mitochondrial genome variation in Eastern Asia and the peopling of Japan. Genome Research, 2004, 14: 1832 - 1850.

[35] Kivisild T, Tolk H V, Parik J, et al. The emerging limbs and twigs of the East Asian mtDNA tree. Molecular Biology and Evolution, 2002, 19: 1737 - 1751.

[36] Umetsu K, Tanaka M, Yuasa I, et al. Multiplex amplified product-length polymorphism analysis of 36 mitochondrial single-nucleotide polymorphisms for haplogrouping of East Asian populations. Electrophoresis, 2005, 26: 91 - 98.

[37] Adachi N, Shinoda K I, Umetsu K, et al. Mitochondrial DNA analysis of Jomon skeletons from the Funadomari site, Hokkaido, and its implication for the origin of Native American. American Journal of Physical Anthropology, 2009, 138: 255 - 265.

[38] Shinoda K. DNA analysis of the Jomon skeletal remains excavated from Shimo-Ohta shell midden, Chiba prefecture. Report for Sohnan Research Instutite for Cultual Properties (in Japanese), 2003, 50: 201 - 205.

[39] Shinoda K, Kanai S. Intracemetry genetic analysis at the NakazumaJomon site in Japan by mitochondrial DNA sequencing. Anthropological Science, 1999, 107: 129 - 140.

[40] Tajima A, Hayami M, Tokunaga K, et al. Genetic origins of the Ainu inferred from combined DNA analyses of maternal and paternal lineages. Journal of Human Genetics, 2004, 49: 187 - 193.

[41] Maruyama S, Minaguchi K, Saitou N. Sequence polymorphisms of the mitochondrial DNA control region and phylogenetic analysis of mtDNA lineages in the Japanese populations. Internal Journal of Legal Medicine, 2003, 117: 218 - 225.

[42] Yao Y G, Kong Q P, Bandelt H J, et al. Phylogenetic differentiation of mitochondrial DNA in Han Chinese. American Journal of Human Genetics, 2002, 70: 635 - 651.

[43] Lee H Y, Yoo J E, Park M J, et al. East Asian mtDNA haplogroup determination in Koreans: haplogroup-level coding region SNP analysis and subhaplogroup-level control region sequence analysis. Electrophoresis, 2006, 27: 4408 - 4418.

[44] Starikovskaya E B, Sukernik R I, Derbeneva O A, et al. Mitochondrial DNA diversity in indigenous populations of the southern extent of Siberia, and the origins of native American haplogroups. Annual Human Genetics, 2005, 69: 67 - 89.

[45] The HUGO Pan-Asian SNP Consortium. Mapping Human Genetic Diversity in Asia. Science, 2009, 326: 1541 - 1545.

[46] Yamaguchi-Kabata Y, Nakazono K, Takahashi A, et al. Japanese population structure, based on SNP genotypes from 7003 individuals compared to other ethnic groups: effect on population-based association studies. American Journal of Human Genetics, 2009, 83: 445 - 456.

[47] Jobling M A, Tyler-Smith C. The human Y chromosome: an evolutionary marker comes of age. Nature Reviews Genetics, 2003, 4: 598 - 612.

[48] Tu J F, Yip V F. Ethnic groups in China. Beijing and New York: Science Press, 1993.

[49] Wen B, Xie X, Gao S, et al. Analyses of genetic structure of Tibeto-Burman population reveals sex-biased admixture in southern Tibeto-Burmans. American Journal of Human Genetics, 2004, 74: 856 - 865.

[50] Thangaraj K, Singh L, Reddy A G, et al. Genetic affinities of the Andaman Islanders, a vanishing human population. Current Biology, 2004, 13: 86 - 93.

[51] Deng W, Shi B, He X, et al. Evolution and migration history of the Chinese population inferred from Chinese Y chromosome evidence. Journal of Human Genetics, 2004, 49: 339 - 348.

[52] Karafet T M, Zegura S L, Posukh O, et al. Ancestral Asian sources(s) of New World Y chromosome founder haplotypes. American Journal of Human Genetics, 1999, 64: 817 - 831.

第6章　Y染色体与家族传承

人们的姓氏大多继承自父亲，而Y染色体是严格的父子相传的基因组片段。所以姓氏与Y染色体的遗传应该是平行的，有共同姓氏的男性可能有相同或相近的Y染色体类型。然而，多起源、改姓、非亲生、从母姓等社会因素弱化了某些姓氏与Y染色体的关联，此时家谱研究可为厘清父系血缘提供线索。Y染色体上稳定的SNP突变可以永远在父系后代中流传，可以构建可靠的父系基因谱系；而其上突变较快的STR位点又可以用以估算时间。因此，Y染色体可用以研究很多姓氏宗族的历史，甚至千百年前的历史疑案。重建姓氏、家谱与Y染色体的关系必将成为历史人类学研究的重要内容。

大跨度家系对于研究Y染色体进化有着极其重要的意义，但家系的可信度却需仔细甄别。笔者用Y染色体分型比对的方法确认了若干有1 800多年历史、延续70～100代的大跨度家系，这些家系宣称是魏武帝曹操的后裔。单倍型O2－F1462是唯一在宣称是曹操后裔的众多家族里频率显著升高的单倍型（$P=9.323\times10^{-5}$，$OR=12.72$），因此也极可能是曹操的Y染色体单倍型。分析结果还显示曹操的Y染色体单倍型与其自称的先祖曹参的单倍型O3－002611并不一致。因为曹操族谱中的祖父是大太监曹腾，曹操父亲的来源一直是历史学界争议的问题。东吴编写的《曹瞒传》宣称曹操父亲曹嵩从夏侯家过继而来，但是在现代夏侯家族中没有检出O2－F1462。为了进一步确认曹操的血缘，笔者检验了曹操祖辈的Y染色体。在安徽亳州发现的元宝坑一号墓中出土了大量的文物，包括诸多铭砖和人骨。从铭砖材料看，墓主很可能是曹操祖父曹腾的弟弟，河间相曹鼎。对人骨牙齿的Y染色体分析发现，其单倍型是O2－F1462，与之前发现的曹操后代的Y染色体类型一致。进一步分析Y染色体STR的多样性，发现元宝坑墓主与安徽现代的数个曹操后裔家族最接近。所以笔者认为，元宝坑墓主应该与曹操有特别近的血缘关系，曹操的父亲是本族过继的。Y染色体与姓氏的研究是Y染色体和谱牒分析相结合的成功探索，为遗传学用于历史学研究提供一个范例。

操姓主要有鄱阳郡操姓和重庆长寿操姓两大分支。据传，鄱阳郡操姓源自逃难的曹操后人。Y染色体调查则显示鄱阳郡操姓与曹操家族，乃至其他曹姓均无关系。

中国的许多穆斯林都将赛典赤·瞻思丁作为自己的祖先，明代杰出的航海家郑和就是赛典赤的后代。赛典赤的祖源虽是回族历史研究的重要话题，但至今仍未有明确的解

读。我们依据谱牒材料对赛典赤和郑和的后裔——云南的纳姓和马姓进行父系Y染色体分型，发现他们属于单倍群L1a-M76，这一类型集中分布在南亚西部，揭示了赛典赤和郑和的波斯祖源。

6.1 Y染色体与姓氏

现代社会中，几乎每人都有自己的姓氏。一个人的姓氏不仅仅是简单的符号，还有着丰富的文化、历史和宗族背景。以血缘为脉络的姓氏记录着各家族甚至各民族的源流，常被用于寻根溯源、族群识别和婚姻关系等相关研究。数千年来大部分姓氏都从父传递，而人类基因组中的Y染色体更严格地遵循父系遗传，因此姓氏与Y染色体有很好的平行对应关系。随着Y染色体上众多遗传标记的发现，用Y染色体结合姓氏分析人类学问题的方法，在分子人类学领域发挥了重要作用。

6.1.1 姓氏在遗传学上的应用

姓氏最早在中国产生，其历史可追溯到5 000年前，主要来源于远古时代各种图腾和地名，“氏”为“姓”的分支，“姓”以别婚姻，“氏”以分贵贱。秦汉以后，姓氏合一，数量大增[1]。据最新统计，拥有13亿人口的中国目前有4 100个姓氏[2]。在日本，姓氏于5世纪晚期产生，但直到1875年才被普通平民使用。至1997年，日本仅1亿左右的人口却拥有约30万个姓氏[3]。在西班牙，子女姓氏沿承父母双方的姓氏，因而会逐代变化。尽管姓氏的起源与发展演变在不同国家和地区有明显差异，但大多数情况下姓氏从父遗传却是相同的。

姓氏的基本作用是明血缘、别婚姻，《左传·僖公二十三年》载：“男女同姓，其生不蕃。”姓氏制度也在不同角度反映了宗法制度的某些内容，起到了维系和延续宗法制度的社会作用。姓氏最早用于遗传研究是在1875年，George Darwin通过分析堂(表)婚得出了英国同姓通婚率和不同阶层的堂(表)近亲通婚率[4]。George Darwin的工作不断被后人丰富和发展，Crow等通过同姓率(isonymy)来计算近亲结婚率[5]。由于居民出生、结婚和死亡等大量相关数据的易得性，姓氏分布与同姓率被广泛用于研究群体遗传结构、迁徙率等[6-32]。姓氏在流行病学方面也得到应用，如Abbotts以姓氏为线索从生理、体质特征和社会经济因素等方面分析了移居到苏格兰的爱尔兰人后裔的高死亡率，Polednak发现了美国康涅狄格州的西班牙姓氏女性有较低乳腺癌发病率等[33-35]。

6.1.2 姓氏与Y染色体的父系遗传

虽然姓氏在宏观上被用于分析群体遗传结构，但是姓氏并不完全遵从父系遗传。就中国的社会情况而言，收养、继养、入赘，甚至直接改姓，都会影响姓氏与父系血统的关联程度。另一方面，中国大多数姓氏起源于春秋时期的各个封国，当封国内的百姓都以国为姓的时候，这些同国百姓的血统可能本来就不一致。这就造成了很多比较大的姓氏内部遗传结构不一致。即便这样，当我们不拘泥于群体中同一姓氏的研究，而是针对有着

明确的历史记载的宗族进行研究，姓氏无疑还是一个很好的遗传标记。

与姓氏不同，人类的Y染色体直接代表着父系遗传，永远是父子相传的，不会受到任何社会文化和自然因素的影响。人体内有23对染色体，其中22对常染色体中，每一对染色体都有一条来自父系，一条来自母系，两条染色体在传代过程中对应的部分会发生交换，从而造成混血的效应，就是遗传学上说的重组。另一对性染色体包括X染色体和Y染色体。在女性体内，X染色体也是成对的，分别来自父母双方，所以也不能避免混血的影响。而在男性体内，却只有一条来自母亲的X染色体和一条来自父亲的Y染色体，也就是说男性的Y染色体只能来源于父亲，所以人体性染色体的遗传方式决定了Y染色体遵从严格的父系遗传(图6-1)。

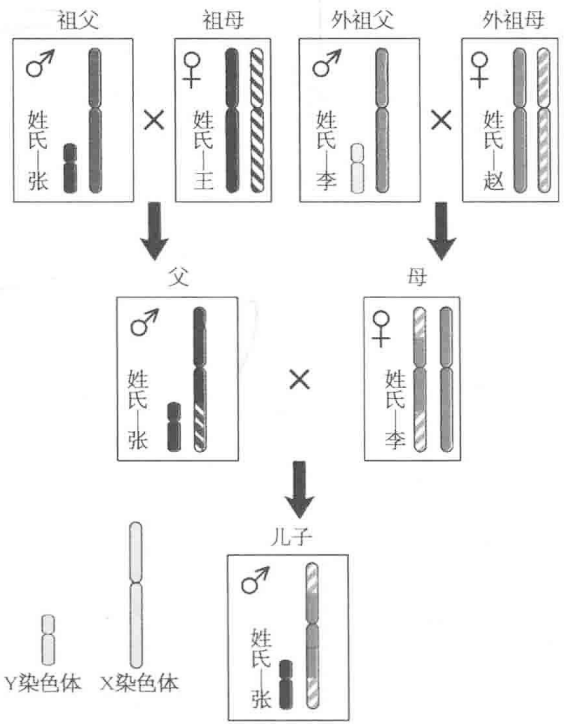


图6-1 姓氏和Y染色体的共同传递

Y染色体与X染色体之间是否会发生重组呢？要回答这个问题，必须先了解Y染色体的结构。人类Y染色体DNA大约包含6 000万个碱基对，其中染色体两端的5%为拟常染色体区域(pseudoautosomal region)，在传代过程中与X染色体相应区段会发生重组，而主干部分的95%为非重组区域(non-recombinating portion of Y chromosome, NRY)，不与任何染色体发生重组(图6-2)[36,37]。Y染色体主干部分的这一特性，保证了子代能完整地继承父代的Y染色体主干而不受混血影响，保证了Y染色体主干的严格父系遗传。

当姓氏已经无法作为追寻祖先的可靠标记的时候，以现代分子生物学技术为基础，

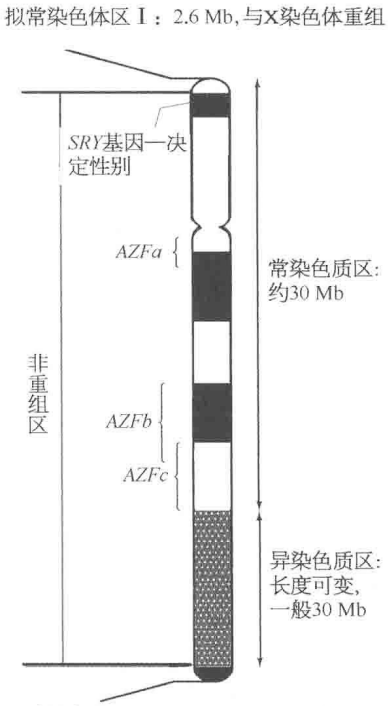


图 6-2　Y 染色体模式图

研究 Y 染色体主干的类型分布是直接追溯群体或者家族父系起源的最佳方法，是验证祖先与后代父系关联的唯一手段。例如，在曹操的后人中分析 Y 染色体特征，我们就可以了解曹操本人的 Y 染色体特征。实际上，在一段有较可信历史记录的时期内，整个家族的姓氏与父系遗传的关联是可以保证的，所以家族的姓氏往往与固定的 Y 染色体类型共同传递，紧密关联。

6.1.3　稳定中变化着的 Y 染色体

在一代一代父子相承的传递过程中，Y 染色体也在慢慢地积累着变化。正是因为遗传突变的积累，使得人类父系遗传体系中，距离越远的个体 Y 染色体差异越大。Y 染色体上的突变形成的个体差异主要有两大类：单核苷酸多态(SNP)和短串联重复(STR)。DNA 分子由四种碱基(A、T、C 和 G)按照一定的顺序连接而成，SNP 是仅仅一个位置上的碱基类型变化。Y 染色体上的同一个 SNP 在人群中一般只有两种类型。STR 则是在染色体的特定区段，由几个碱基组成一个单位重复出现，不同的 Y 染色体上的同一个 STR 位置往往有不同的重复拷贝数。SNP 和 STR 由于突变性质和突变速度不同，在分析中有着不同的用途。

要确立父系遗传体系，最重要的前提是祖先的突变可以稳定地保留在后代的 Y 染色体上。SNP 突变因为突变速率极低，可以做到在后代中永久地保留，后代只能在祖先的突变基础上积累新的突变，而不会丢失祖先的突变特征。通过比较人类与黑猩猩的 Y 染色体差异[38]，以及大家系中 Y 染色体的差异程度[39]，Y 染色体上的 SNP 突变的速率被计算了出来。每出生一个男子，一个染色体位置上发生 SNP 突变的概率大约为三千万分之一。实际上由于 Y 常染色质区的保守性，以及人类历史上大量男子都没有男性后代保留至今的事实，实际的群体中突变率应该低几个数量级。而我们通常研究的是 Y 染色体非重组区大约 3 000 万个碱基对的常染色质区[40,41]，按照每个碱基对三千万分之一的突变率，这个区段内每个男子平均都会有一个新的突变。这个新的突变随机地出现在 Y 染色体常染色质区的任意一个点上，如果这个突变了的点上再发生一次突变，那么这个突变就在后代中丢失了，我们就无法通过后代确定祖先的 Y 染色体突变谱。但是同一个点上先后发生两次突变的概率，按照概率计算方法就是三千万分之一的平方，也就是九百万亿分之一，相对于人类自古以来的人口，这个概率就近似于零。可以说，绝大多数情况下，祖先的 Y 染色体上出现的 SNP 突变特征在后代中能够找到，而后代只能在祖先 Y

染色体突变谱的基础上增加新的突变(图6-3)。

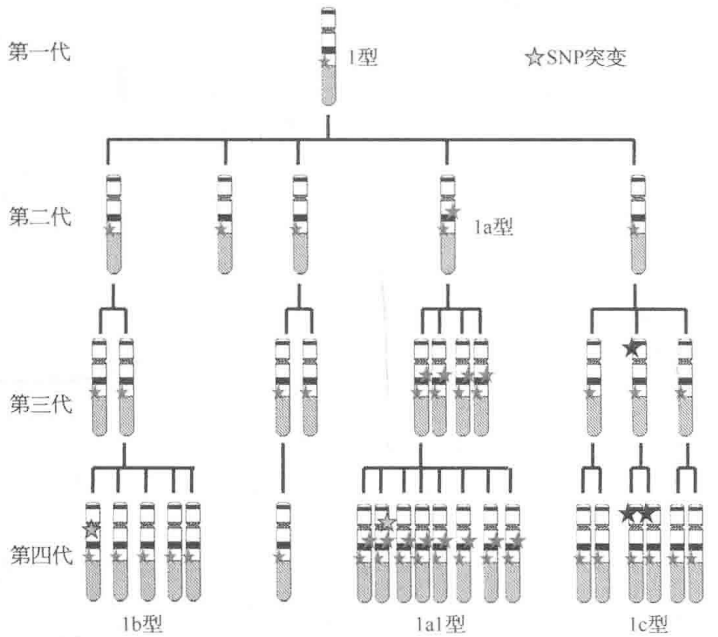


图6-3 Y染色体突变谱构成单倍型

由多个SNP突变构成的一种突变谱被称为一种单倍型。例如图6-3中就有5个SNP突变,陆续构成5种单倍型。其中1型是其他单倍型的祖先型,其他单倍型都是后代型。祖先型与所有后代型合称为一个单倍群。一个家族的所有Y染色体理论上都属于一个单倍群,因为其中所有的男性都应该来自同一个祖先。

当然,单倍群的概念可大可小。大而言之,全世界的Y染色体都属于一种单倍群,都来自二十多万年前的一个东非晚期智人男子[42,43]。进而,全世界又可以分为20种主干单倍群,编号从A到T。最古老的A和B单倍群都没有走出非洲,C和D单倍群最早来到大洋洲和亚洲,E单倍群来到亚洲又回到非洲,F单倍群衍生出G、H、I和J等单倍群在西方形成欧罗巴人种,衍生出K单倍群并形成N、O、P和Q等单倍群在东方形成蒙古人种,其中O单倍群成为中国人的主流,而Q单倍群成为美洲印第安人的主流。所以Y染色体的谱系构建出了全人类的一部大家谱。

6.1.4 Y染色体上的时钟

利用Y染色体上稳定遗传的SNP,我们可以构建出个体或家族之间明确的遗传渊源。而且,既然SNP有稳定的突变速率,当我们统计出不同人的Y染色体之间的突变差异数,将差异数除以速率,经过换算就可以估算两条Y染色体之间的分化时间。但是,由于SNP的突变速率实在太低,个体之间的突变差异散布在Y染色体的各处,只能使用Y染色体全测序来寻找,而目前全测序的成本太高,尚不能普遍应用。这一缺点被Y染色

体上的另一遗传标记STR弥补了。一些STR位点分布在Y染色体上的固定位置，每一个STR位点内部的重复单位在传代过程中改变着拷贝数，这种改变也有着固定的速率。而STR突变速率要比SNP大得多，在家系中每出生一个男子每个STR位点突变概率大约是三百分之一（表6-1）。一般的Y染色体分析中，调查15个STR位点，总体突变率大约是二十分之一。而Y染色体上大约有150个4～6个核苷酸重复的STR[52]，如果分析全部的STR位点，那么总突变率大约就是二分之一。这一高突变率就非常有利于估算不同Y染色体之间的分化时间，因此STR位点成为Y染色体上的“时钟”。

表6-1 不同文献估算的STR突变率

突变率类型	参考文献	发表时间	突变率（位点/代）
家系突变率	[44]	1997年	2.1×10^{-3}
家系突变率	[45]	1997年	$(0.27\sim1.1)\times10^{-3}$
家系突变率	[46]	1999年	2.1×10^{-3}
家系突变率	[47]	1997年	3.2×10^{-3}
家系突变率	[48]	2000年	2.8×10^{-3}
进化突变率	[49]	2000年	2.6×10^{-4}
家系突变率	[50]	2001年	4.0×10^{-3}
进化突变率	[51]	2004年	6.9×10^{-4}

STR的突变是双向性的，拷贝数可以增加或减少。出自同一祖先的不同个体的同一STR位点，可能有不同突变方向和重复数。同SNP一样，数个不同位置上的STR也可以构成单倍型[47,53]。在群体中分析STR单倍型的多样性程度可以计算群体的共祖时间。假设一个STR每次突变都只增加或者减少一个重复单位，也就是一步（single-step）突变模型，且群体有着恒定的有效群体大小，就可由公式 $t=-Ne\ \ln(1-V/Ne\mu)$ 推算出某特定Y-SNP发生的大致时间。公式中，Ne 是有效群体大小，μ 是突变率，ln是自然对数，V 是观察到的群体中的某一STR数值的方差，计算得到的 t 是经历的世代数，再乘以每一世代的年数即可得到时间[54]。

以Y染色体上STR的总突变率二分之一来估算，几乎每个人都可以构成独特的单倍型。但是，由于突变是一步一步发生的，父系亲缘关系越近的个体之间的STR单倍型越相似，一个纯粹由父系传递的姓氏应有相近的STR单倍型。理论上，有了足够数量的Y染色体SNP和STR后，通过调查一个姓氏宗族内的男性单倍型，就能够很清楚地构建其家族Y染色体的谱系树，乃至编写一部清晰的基因家谱。

6.1.5 姓氏与Y染色体关联实践分析

在实际应用中，姓氏与Y染色体是否具有基本相同的和平行的表现还要看姓氏传递

是否连续和稳定。多项研究证实各国姓氏的传承是相对稳定的。袁义达等[55-58]分析和比较了宋代、明代和当代中国姓氏的分布曲线、同姓率和地域人群间的亲缘关系，发现1 000多年前的姓氏分布和当代基本一致。Legay等[59]比较了法国19世纪末和1975年的若干姓氏的分布，也发现了其传递的稳定性。这些研究为姓氏和Y染色体关联分析提供了理论依据。Giraldo等[60]发现哥伦比亚4个有"卫星式"Y染色体[61]的家族有3个是同姓氏的。Sykes等[62]检测DYS19、DYS390、DYS391和DYS393四个Y-STR位点，对48名随机的Sykes姓氏英国男性个体进行Y-STR单倍型分析，发现21名个体(43.8%)的4个Y-STR单倍型相同，而该单倍型在对照组的139名随机的英国男性个体中却未检出，这反映了Sykes姓氏比较单一的起源。其余的Sykes姓氏个体的Y-STR单倍型与对照组差异不明显，若不考虑Y-STR基因变异的累积，Sykes家族的姓氏变更率为每代1.3%。与Sykes等只选一个姓氏做研究不同，King等[63]用150对随机选出的男性(每对两人有一个共同的英国姓氏)来分析，发现同一姓氏出现相同Y染色体单倍型的概率提高；并且随着姓氏在人群中分布频率的减少，Y染色体单倍型相同的概率增加。Hill等[64]和Wilson等[65]的研究也表明爱尔兰、奥克尼群岛男性的Y染色体单倍型分类与姓氏显著相关。

利用Y染色体来检测历史上的家族关系疑案有多项成功的案例，较有意思的是Foster[66]和Skorecki[67]的研究。1802年，美国第三任总统托马斯·杰弗逊因被怀疑与女仆Sally Hemings有过孩子而遭起诉。此后，人们一直对此事争论不休，而Foster用Y染色体回答了这个问题。Foster比较了Jefferson的叔叔、Sally的大儿子和最小儿子的男性后代Y染色体YAP、一些SNP、STR及小卫星MSY1等多态位点，得出结论杰弗逊是Sally的最小儿子的生父。Y染色体不但能够解决数百年的疑案，还能追溯到数千年前的历史。Skorecki等就证实了圣经中的传说。圣经中记载犹太人中的祭司是由犹太教的第一祭司长Aaron开始按血缘代代相传，而身为德系犹太人祭司的Skorecki发现他与一个西班牙系犹太人祭司的体质特征差别很大，这让他寝食难安。Skorecki就和研究Y染色体的专家Hammer教授合作以YAP和DYS19来分析犹太教祭司的单倍型，结果显示德系和西班牙系犹太祭司们与非祭司的犹太人相比有较近的亲缘关系，也就是说祭司们可跨越3 300年追溯到一个共同的父系祖先。Y染色体的分析与圣经故事的完美契合着实让人吃惊。

对于中国的姓氏与Y染色体的相关性，也有许多研究见诸报道。吴东颖等[68]应用Y染色体的DYS19、DYS390、DYS391和YCAII四个STR，分析50例中国汉族王姓男性DNA样本，结果与随机样本的单倍型频率和分布均无显著差异。邓志辉等[69]通过对深圳无偿献血人群中李姓、王姓和张姓无关男性个体Y-STR单倍型遗传多态性分析表明，该地区此三姓氏无关男性个体Y-STR单倍型的遗传多态性丰富，与以往的汉族无关男性群体遗传资料比较差异不显著。这说明，汉族的大姓内部基本没有同源性，相关Y染色体研究只能在明确的姓氏宗族中开展。

汉族大姓氏内部的不一致有很多可能的原因。在理想的情形下，每种姓氏都有一个

唯一来源，即该姓氏的奠基者只是一人或是有相同Y染色体单倍型的多人，在姓氏传承过程中没有发生过干扰(改姓、非亲生等)，此时一种姓氏可以被一种SNP和STR的单倍型来鉴定[70]。但是中国的大多数姓氏起源并不单一。周朝的姓氏大多是以封国为氏，后改为姓。比如曹国的王室后代姓曹，但是其仆役后人也可以姓曹，甚至整个封国内所有百姓后代都可以姓曹。而曹国内的百姓来源本来就是多样的，有着各种各样的Y染色体。所以中国的姓氏总体上内部父系血缘不一致。另外，犹如STR单倍型随时间而演化出越来越多的类型一样，姓氏在传承过程中经历的时间越长，其受到的社会干扰越多，显示出的差异也越大。在中国，姓氏有近5 000年的历史，来源复杂，且存在避祸改姓、避讳改姓、过继改姓、皇帝赐姓与贬姓、少数民族用汉姓等问题。举个简单的例子，中国的100个大姓中有53个据称起源于姬姓[58]。如此，研究中国的姓氏难度极大，但是中国编修家谱的传统对厘清这纷繁复杂的血缘关系有很大帮助。

家谱是一种以表谱形式记载某一同宗共祖以血缘关系为主体的家族世系繁衍兼及其他方面情况的特殊图书体裁。也就是说，入谱者必须是同宗共祖，即使同姓，若不同祖，也不能修入一部家谱之中[71]。在中国的广大农村，人们一直有着同姓聚居的习俗，加上婚姻半径较小，由家谱确定的某一地域内同姓人群，可以认为是有相同或相近Y染色体的父系隔离群体，这也就为分子人类学分析Y染色体DNA多样性提供了极好的研究模型。然而，某些家谱里有假托、借抄的内容[72]，因此对于家谱资料的应用必须审慎。但是在Y染色体检验这种无可辩驳的科学证据面前，任何家谱都可以得到检验和修正。姓氏、家谱和Y染色体的关联研究必然成为研究中国人起源与演变的重要方式，开创历史人类学研究的新进程。

参考文献

[1] 范又琪. 姓氏起源. 武汉：长江出版社，1999：3-4.

[2] 袁义达，邱家儒. 中国姓氏·三百大姓. 上海：华东师范大学出版社，2007.

[3] 丹羽基二. 日本苗字大辞典. 东京：芳文馆，1997.

[4] Darwin G H. Marriages between first cousins in England and their effects. J Statist Soc, 1875, 38: 153-184.

[5] Crow J F, Mange J F. Measurement of inbreeding from the frequency of marriages between persons of the same surname. Eugenics Q, 1965, 12: 199-203.

[6] Roberts D F, Rawling C P. Secular trends in genetic structure: an isonymic analysis of Northumberland parish records. Ann Hum Biol, 1974, 1: 393-410.

[7] Morton N E, Smith C, Hill R, et al. Population structure of Barra (Outer Hebrides). Ann Hum Genet, 1976, 39: 339-352.

[8] Shin E H, Yu E Y. Use of surnames in ethnic research: the case of Kims in the Korean-American population. Demography, 1984, 21: 347-360.

[9] Lasker G W, Kaplan B A. Surnames and genetic structure: repetition of the same pairs of names in married couples, a measure of subdivision of the population. Hum Biol, 1985, 55: 431-440.

[10] Mascie-Taylor C G N, Lasker G W. Geographical distribution of common surnames in England and Wales. Ann Hum Biol, 1985, 12: 397 - 401.

[11] Pinto-Cisternas J, Pineda L, Barrai I. Estimation of inbreeding by isonymy in Iberoamerican populations: an extension of the method of crow and mange. Am J Hum Genet, 1985, 37: 373 - 385.

[12] Barrai I, Barbujani G, Beretta M, et al. Surnames in ferrara: distribution, isonymy and levels of inbreeding. Ann Hum Biol, 1987, 14: 415 - 423.

[13] Piazza A, Rendine S, Zei G, et al. Migration rates of human populations from surname distributions. Nature, 1987, 329: 714 - 716.

[14] Barrai I, Formica G, Barale R, et al. Isonymy and migration distance. Ann Hum Genet, 1989, 53: 249 - 262.

[15] Holloway S, Sofaer J A. Coefficients of relationship by isonymy within and between the regions of Scotland. Hum Biol, 1989, 61: 87 - 97.

[16] Mascie-Taylor C G N, Lasker G W. The distribution of surnames in England and Wales: a model for genetic distribution. Man New Series, 1990, 25: 521 - 530.

[17] Sokal R, Harding R, Mascie-Taylor C G N. A spacial analysis of 100 surnames in England and Wales. Ann Hum Biol, 1992, 19: 445 - 476.

[18] Rodríguez-Larraldea A, Pavesia A, Scapolia C, et al. Isonymy and the genetic structure of Sicily. J Biosoc Sci, 1994, 26: 9 - 24.

[19] Mourrieras B, Darlu P, Kochez J. Surname distribution in France: a distance analysis by distorted geographical map. Ann Hum Biol, 1995, 22: 183 - 198.

[20] Barrai I, Scapolia C, Beretta M, et al. Isonymy and the genetic structure of Switzerland I. The distribution of surnames. Ann Hum Biol, 1996, 23: 431 - 455.

[21] Barrai I, Scapoli C, Beretta M. Isolation by distance in Germany. Hum Genet, 1997, 100: 684.

[22] Rodríguez-Larralde A, Barrai I, Nesti C. Isonymy and isolation by distance in Germany. Hum Biol, 1998, 70: 1041 - 1056.

[23] Barrai I, Rodríguez-Larralde A, Mamolini E. Isonymy and isolation by distance in Italy. Hum Biol, 1999, 71: 947 - 961.

[24] Barrai I, Rodríguez-Larralde A, Mamolini E. Elements of the surname structure of Austria. Ann Hum Biol, 2000, 27: 607 - 622.

[25] Rodríguez-Larralde A, Morales J, Barrai I. Surname frequency and isonymy structure of Venezuela. Am J Hum Biol, 2000, 12: 352 - 362.

[26] Barrai I, Rodríguez-Larralde A, Mamolini E. Isonymy structure of USA population. Am J Phys Anthropol, 2001, 114: 109 - 123.

[27] Barrai I, Rodríguez-Larralde A, Manni F. Isonymy and isolation by distance in the Netherlands. Hum Biol, 2002, 74: 263 - 283.

[28] 徐立群,李辉,奚慧峰,等.上海郊区姓氏和通婚分析.遗传学报,2002,29: 666 - 673.

[29] Rodríguez-Larralde A, Gonzales-Martin A, Scapoli C, et al. The names of Spain: a study of the isonymy structure of Spain. Am J Phys Anthropol, 2003, 121: 280 - 292.

[30] Dipierri J E, Alfaro E L, Scapoli C, et al. Surnames in Argentina: a population study through

isonymy. Am J Phys Anthropol, 2005, 128: 199 - 209.

[31] Jorde L B, Morgan K. Genetic structure of Utah Mormons: isonymy analysis. Am J Phys Anthropol, 2005, 72: 403 - 412.

[32] Tay J S H, Yip W C L. The estimation of inbreeding from isonymy: relationship to the average inbreeding coefficient. Ann Hum Genet, 2007, 48: 185 - 194.

[33] Cook D, Hewitt D, Milner J. Uses of the surname in epidemiologic research. Am J Epidemiol, 1972, 95: 38 - 45.

[34] Abbotts J, Williams R, Smith G D. Association of medical, physiological, behavioural and socio-economic factors with elevated mortality in men of Irish heritage in West Scotland. J Publ Health Med, 1999, 21: 46 - 54.

[35] Polednak A P. Estimating breast cancer incidence in Hispanic women in Connecticut, 1989 - 1991. Ethn Health, 1996, 1: 229 - 235.

[36] Quintana-Murci L, Krausz C, McElreavey K. The human Y chromosome: function, evolution and diease. Forensic Sci Intl, 2001, 118: 169 - 181.

[37] Jobling M A, Pandya A, Tyler-Smith C. The Y chromosome in forensic analysis and paternity testing. Int J Legal Med, 1997, 110: 118 - 124.

[38] Kuroki Y, Toyoda A, Noguchi H, et al. Comparative analysis of chimpanzee and human Y chromosomes unveils complex evolutionary pathway. Nat Genet, 2006, 38: 158 - 167.

[39] Xue Y L, Wang Q J, Long Q, et al. Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree. Curr Biol, 2009, 19: 1453 - 1457.

[40] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature, 2001, 409: 860 - 921.

[41] Venter J C. The sequence of the human genome. Science, 2001, 291: 1304 - 1351.

[42] Jobling M A, Tyler-Smith C. The human Y chromosome: an evolutionary marker comes of age. Nat Rev Genet, 2003, 4: 598 - 612.

[43] Underhill P A, Shen P, Lin A A, et al. Y chromosome sequence variation and the history of human populations. Nat Genet, 2000, 26: 358 - 361.

[44] Heyer E, Puymirat J, Dietjes P, et al. Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees. Hum Mol Genet, 1997, 6: 799 - 803.

[45] Caglià A, Novelletto A, Dobosz M, et al. Y chromosome STR loci in Sardinia and continental Italy reveal islander-specific haplotypes. Eur J Hum Genet, 1997, 5: 288 - 292.

[46] Jobling M A, Bouzekri N, Taylor P G. Hypervariable digital DNA codes for human paternal lineages: MVR-PCR at the Y-specific minisatellite, MSY1 (DYF155S1). Hum Mol Genet, 1998, 7: 643 - 653.

[47] Kayser M, CagliàA, Corach D, et al. Evaluation of Y chromosomal STRs: a multicenter study. Int J Legal Med, 1997, 110: 125 - 133.

[48] Kayser M, Roewer L, Hedman M, et al. Characteristics and frequency of germline mutations at microsatellite loci from the human Y chromosome, as revealed by direct observation in father/son pairs. Am J Hum Genet, 2000, 66: 1580 - 1588.

[49] Forster P, Röhl A, Lünnemann P, et al. A short tandem repeatbased phylogeny for the human Y chromosome. Am J Hum Genet, 2000, 67: 182 - 196.

[50] Holtkemper U, Rolf B, Hohoff C, et al. Mutation rates at two human Y chromosomal microsatellite loci using small pool PCR techniques. Hum Mol Genet, 2001, 10: 629 - 633.

[51] Zhivotovsky L A, Underhill P A, Cinnioğlu C, et al. The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. Am J Hum Genet, 2004, 74: 50 - 61.

[52] Ayub Q, Mohyuddin A, Qamar R, et al. Identification and characterisation of novel human Y chromosomal microsatellites from sequence database information. Nucleic Acids Res, 2000, 28(2): e8.

[53] Jobling M A, Tyler-Smith C. Fathers and sons: the Y chromosome and human evolution. Trends Genet, 1995, 11: 449 - 456.

[54] Su B, Xiao J, Underhill P, et al. Y chromosome evidence for a northward migration of modern humans in East Asia during the Last Ice Age. Am J Hum Genet, 1999, 65: 1718 - 1724.

[55] 袁义达,张诚,马秋云,等.中国人姓氏群体遗传Ⅰ.姓氏频率分布与人群遗传分化.遗传学报,2000,27: 471 - 476.

[56] 袁义达,张诚,杨焕明.中国人姓氏群体遗传Ⅱ.姓氏传递的稳定性与地域人群的亲缘关系.遗传学报,2000,27: 565 - 572.

[57] 袁义达,金锋,张诚.宋朝中国人的姓氏分布与群体结构分化.遗传学报,1999,26: 187 - 197.

[58] 袁义达,张诚.中国姓氏:群体遗传与人口分布.上海:华东师范大学出版社,2002.

[59] Legay J M, Vernay M. The distribution and geographical origin of some French surnames. Ann Hum Biol, 2000, 27: 587 - 605.

[60] Giraldo A, Martinez I, Guzmán M. A family with a satellited Yq chromosome. Hum Genet, 1981, 57: 99 - 100.

[61] Schmid M, Haaf T, Eva Solleder, et al. Satellited Y chromosomes: structure, origin and clinical significance. Hum Genet, 1984, 67: 72 - 85.

[62] Sykes B, Irven C. Surname and the Y chromosome. Am J Hum Genet, 2000, 66: 1417 - 1419.

[63] King T E, Ballereau S J, Schurer K E, et al. Genetic signatures of coancestry within surnames. Curr Biol, 2006, 16: 384 - 388.

[64] Hill E W, Jobling M A, Bradley D G. Y chromosome variation and Irish origins. Nature, 2000, 404: 351 - 352.

[65] Wilson J F, Weiss D A, Richards M. Genetic evidence for different male and female roles during cultural transitions in the British Isles. Proc Natl Acad Sci USA, 2001, 98: 5078 - 5083.

[66] Foster E A, Jobling M A, Taylor P G, et al. Jefferson fathered slave's last child. Nature, 1998, 396: 27 - 28.

[67] Skorecki K, Selig S, Blazer S, et al. Y chromosomes of Jewish priests. Nature, 1997, 385: 32.

[68] 吴东颖,马素参,刘明,等.应用Y染色体多态标记对汉族王姓亲缘关系的研究.人类学学报,2000,19: 132 - 137.

[69] 邓志辉,李茜,王大明,等.深圳无偿献血人群中李姓、王姓和张姓无关男性个体Y-STR单倍型

遗传多态性. 遗传,2007,29: 1336-1344.

[70] Jobling M A. In the name of the father: surnames and genetics. Trends Genet, 2001, 17: 353-357.

[71] 王鹤鸣. 中国家谱总目. 上海: 上海古籍出版社,2009.

[72] 赵世瑜. 祖先记忆、家园象征与族群历史——山西洪洞大槐树传说解析. 历史研究,2006,1: 49-64.

6.2 现代Y染色体揭示曹操的身世

6.2.1 研究背景

Y染色体上的绝大部分是从父遗传且缺乏重组,所以可以通过研究历史人物现存后代的Y染色体来揭示历史人物之间的父系关系[1-3]。近年来,国际上的成功例子有美国第三任总统托马斯·杰弗逊的私生子的确认[4]、犹太教祭司的Y染色体单倍型的推定[5]等。实际上,Y染色体推测历史人物的深度还可以往更古老的年代推进。通过可靠的家谱信息可以重建跨度极大的家系,这样的家系可用来推断其遥远祖先的Y染色体单倍型,为解决历史悬案提供新途径。这样的大跨度家系对研究Y染色体的突变率及其进化机制也都有重要意义[6]。修著家谱是中国人的传统,一些家谱甚至跨越3 000年将现代人与其祖先相连,虽然它们的真实性还有待确认。我们通过研究宣称是曹操后代的现代家族的男性Y染色体来揭示曹操的士族出身是否属实[7],为探索将遗传学、谱牒资料等用于进化和历史学研究提供范例。

魏武帝曹操(155—220)因小说《三国演义》的风靡而成为东亚最有名的历史人物之一。陈寿的《三国志》记载曹操是西汉第二任相国曹参(?—190)的后代[8],而曹操自称其祖源可远溯至古曹国(前11世纪—前487)的曹叔振铎[9],即曹操宣称自己是皇室贵族后裔。是否具有显赫的士族出身对于曹操争夺政治利益有很重要的意义。然而曹操的祖父曹腾在东汉为宦官之首,曹操的父亲曹嵩是曹腾的养子,曹操的政敌袁绍在攻曹的檄文中写到"父嵩乞丐携养"[10],说他是路边捡来的乞丐。关于曹操身世的各种说法由此流传,且纷纷扰扰争论了近2 000年。

曹操家族的起源争议颇大,现存家谱是了解曹操后裔分布情况的一条重要线索。家谱的先世传说并不十分可靠。曹氏家谱的起点,我们选择陈寿(西晋时期史学家,《三国志》作者)以来言之有据的传承,那就是始于曹参的后裔宗族,而不是再往前的西周乃至夏朝先祖。本研究课题组对上海图书馆收藏的118件曹氏族谱进行了全面查阅和筛选,将家谱所载世系同历史记载相比较,挑选出比较可靠的家谱,其中包含自称是曹操后裔的家谱。把曹氏取样的面铺得广一些,多取一些样本,能够更加全面地反映自古以来多支曹氏家族的基因状况。家谱调查反映出与曹参或者曹操有关系的曹氏,大量分布于长江流域,与史料所见五胡十六国以后曹操家族迁居江南的情况是吻合的。据此,可大致确定几处最接近曹操后裔的居住地:一是位于安徽、山东、河南、江苏四省的交界处,以安徽亳州为中心,这里是曹操政权的发源地,也是安阳汉墓被发现的地区;二是江东地区,包括安徽泾县、繁昌、歙县、绩溪,浙江金华、东阳、绍兴、余姚、萧山,江西赣州等地;三是

湖南等省沿江地区，如湖南新化、郴州、益阳、长沙等地。

6.2.2 材料和方法

1. 样本

本项目由复旦大学生命科学学院伦理委员会审核通过，根据知情同意的原则，在全国各地采集了79个曹姓家族的280个男性和其他姓氏的446个男性外周血样本，大部分谱牒资料也在家族成员知情同意后影印保存。捷瑞GK1042型试剂盒提取DNA。

2. 遗传标记

本项目所选取的101个SNP位点如下。版1（单倍群O内）：M175，M119，P203，M110，M268，P31，F1462，M95，M176，M122，M324，M121，P201，M7，M134，M117，002611，P164，L127（rs17269396）和KL1（rs17276338）；版2（单倍群O外）：M130，P256，M1，M231，M168，M174，M45，M89，M272，M258，M242，M207，M9，M96，P125，M304，M201和M306；版3（单倍群C）：P54，M105，M48，M208，M407，P33，M93，P39，P92，P53.1，M217，M38，M210，M356，P55和M347；版4（单倍群D）：P47，N1，P99，M15，M125，M55，M64.2，M116.1，M151，N2和022457；版5（单倍群N）：M214，LLY22g，M128，M46/Tat，P63，P119，P105，P43和M178；版6（单倍群R）：M306，M173，M124，M420，SRY10831.2，M17，M64.1，M198，M343，V88，M458，M73，M434，P312，M269和U106/M405；版7（单倍群Q）：P36.2，M3，M120，MEH2，M378，N14/M265，M25，M143，M346，L53和M323。

这些位点涵盖了最新Y染色体谱系树上东亚的所有单倍群[11]。基因分型采用SNaPshot（ABI SNaPshot®多重试剂盒）和荧光引物PCR相结合的方法[12]，PCR产物纯化后在ABI 3730测序仪上分析。

3. 统计分析

为确认曹操最可能的Y染色体单倍型，笔者使用Fisher精确检验分析每种单倍型在不同分类的家族中的频率差异[13,14]、双侧显著性检验分析分类家族间的差异、右侧检验分析每种单倍型在宣称是曹操后裔家族中是否显著高频出现。同时还计算了每种单倍型在宣称是曹操后裔的家族和对照家族间的*OR*值。*OR*值可用来描述两组变量间的相关性，常被用在流行病学的病例-对照研究中[15]。本项目中，*OR*值是用来描述Y染色体单倍型和宣称的祖源之间的关系，用N_c^H和N_r^H分别代表在宣称的家族和对照家族中是某一单倍型H的家族的数目，T_c和T_r则分别代表宣称的家族和对照家族的所有家族，那么单倍型H的*OR*值为$OR_H = \frac{N_c^H \times (T_c - N_c^H)}{N_r^H \times (T_r - N_r^H)}$。

6.2.3 研究结果

用Y染色体上的100个单核苷酸位点（SNP）对全国各地曹姓79个家族的280个男

性以及其他姓氏的 446 个男性进行分型。所选取的 100 个 SNP 位点涵盖了最新 Y 染色体谱系树上的 O、N、C、D、Q、R 和 J 等东亚地区可见的所有单倍群[11]。如果同一家族的个体间 Y 染色体单倍型不同，该家族则是由一些不同来源的单纯家族组成的复合家族，因此实际上总共研究了 111 个曹姓单纯家族(图 6－4a)。同时在家族成员知情同意后将大部分家谱资料影印保存。根据家谱资料，15 个曹姓家族宣称是曹操后代，他们中共有 6 种 Y 染色体单倍型(图 6－4a)，其中仅有一种单倍型是曹操的，而其他 5 种可能是由于收养、从母姓、非亲生等原因引入的[16]。需要通过分析不同单倍型在家族间的分布来确认哪个才最可能是曹操的 Y 染色体单倍型。

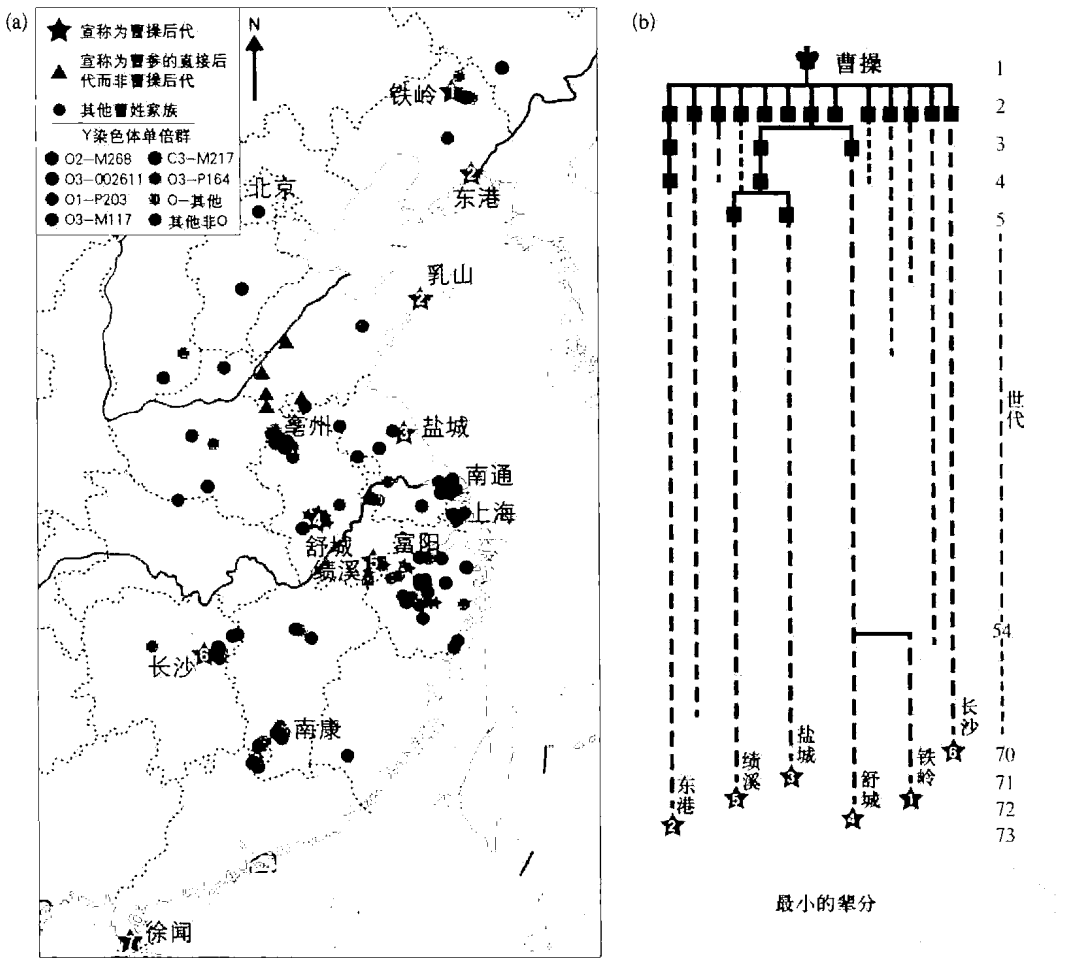


图 6－4　曹姓家族的 Y 染色体单倍型分布(a)和已确认的大跨度家系(b)

本研究完成后，在山东乳山发现了与辽宁东港同一家谱的家族，在广东徐闻的曹家发现了记录祖先来自曹操的家谱。

把这些曹姓家族分成三类：宣称是曹操后代的、未宣称是曹操后代或宣称是其他祖源的、其他姓氏的普通对照人群。首先，两两比较每种单倍型在这三类家族间的频率差

异，在未宣称为曹操后代的家系与普通对照人群间并未发现任何单倍型有显著差异（表6-2），这表明曹姓与其他姓氏一样也是多起源的[17]，所以在之后的分析中把未宣称的家族和普通人群合并在一起作为对照组。非常有趣的是，O2-F1462（最初报道中分型到M268）是唯一在宣称是曹操后代的家族中显著高频出现的单倍型（Fisher精确检验，$P=9.323\times10^{-5}$），很有可能这就是曹操的Y染色体单倍型。某种单倍型是否是曹操的单倍型不能仅从其在宣称的家族里频率来判断，还要考虑其在宣称的家族与对照人群的频率差异。用*OR*值估算O2-F1462是曹操单倍型的可能性，*OR*值（带置信区间）可判断两组变量间关系。O2-F1462在宣称为曹操后代的家族和对照人群间的*OR*值为12.72（表6-2），所以其为曹操单倍型的可能性为92.71%。还有一点需注意，有8个曹姓家族明确认为不是曹操后代，这8个家族的Y染色体单倍型也都不是O2-F1462，这也增加了O2-F1462为曹操单倍型的可能性。

表6-2 不同类别的曹姓家族间Fisher精确检验

单倍型	家族的数目			两两比较的*P*值		宣称的家族和对照组间的*OR*值[95%置信区间]
	宣称家族	未宣称家族	其他姓氏	未宣称家族对普通人群①	宣称家族对对照组②	
O2-F1462	6	5	22	0.801	9.32×10^{-5}*	12.72[4.22～38.32]
O3-002611	1	21	79	0.384	0.952	0.32[0.04～2.43]
O1-P203	2	6	65	0.030	0.607	1.02[0.23～4.62]
O3-M117	3	15	67	0.876	0.408	1.40[0.39～5.08]
C3-M217	2	5	25	1.000	0.211	2.63[0.57～12.17]
O3-P164	1	2	13	1.000	0.358	2.51[0.31～20.34]
其他	0	42	175	0.423	1.000	0

注：未宣称曹操后裔的家族和普通人群合并为对照组。*表示$P<0.01$；①为双侧检验；②为右侧检验。

同时，笔者也估算了反概率，即假定所有的曹操家谱都是伪造的，那么出现6个家族都是相同的O2-F1462单倍型的巧合概率。由于这些家族在历史中都没有联络，所以他们如果来源不同而基因类型相同就只能是由巧合造成的。我们知道，两个事件同时发生的巧合概率是两个事件单独发生的概率的乘积。那么可以这样估计曹操家谱相关家族的基因巧合概率：出现两个家族基因相同的事件计为一次巧合事件，概率就是O2-F1462在全国汉族人中的频率约5%[18]；出现第三个相同基因的家族则是两次巧合事件，概率为5%乘以5%；那么6个分布在不同地域的曹姓家族同时是O2-F1462的概率是$(5\%)^5=3.125\times10^{-7}$，大约为一千万分之三，也就是几乎不可能。

单倍型O3-002611是在其他曹姓家族中出现频率最高，也是唯一在宣称是曹参直系后裔的5个曹姓家族中都唯一出现的单倍型（图6-4a）。继续使用推断曹操Y染色体

单倍型的方法对不同曹姓家族进行分类去推定曹参最可能的单倍型：5个自称是曹参后裔的O3-002611家族作为宣称组，79个未宣称的O3-002611家族和458个其他单倍型的家族作为对照组。经分析，O3-002611就是曹参最可能的单倍型(Fisher精确检验，$P=1.968\times10^{-4}$)，可能性近于100%(*OR*值很高)。综合分析结果，笔者认为魏武帝曹操不太可能是曹参的后裔，遗传学证据并不支持曹操自称的士族出身。

6.2.4　结论

结合数据资料，研究结果也确认了O2-F1462的曹操后裔和O3-002611的曹参后裔的谱牒资料的可靠性(图6-4b)，这些谱牒将现代曹姓族人上溯70～100代与其2 000多年前的祖先相连，这对于Y染色体的进化研究有重要意义。此次曹姓家族Y染色体调查是将遗传学用于古代史研究的一个成功范例，且提供了两个重要的契机：一是促成了历史学和分子生物学的深层次合作研究，建立了国内第一个以分子生物学为主要研究工具的历史人类学新学科；二是加快了人类基因调查从以民族向以家族为对象的转变。

参考文献

[1] Sykes B, Irven C. Surnames and the Y chromosome. Am J Hum Genet, 2000, 66: 1417-1419.

[2] Underhill P A, Shen P, Lin A A, et al. Y chromosome sequence variation and the history of human populations. Nat Genet, 2000, 26: 358-361.

[3] Jobling M A, Tyler-Smith C. The human Y chromosome: an evolutionary marker comes of age. Nat Rev Genet, 2003, 4: 598-612.

[4] Foster E A, Jobling M A, Taylor P G, et al. Jefferson fathered slave's last child. Nature, 1998, 396: 27-28.

[5] Skorecki K, Skorecki K, Selig S, et al. Y chromosomes of Jewish priests. Nature, 1997, 385: 32.

[6] 王传超，严实，李辉. 姓氏与Y染色体. 现代人类学通讯，2010，4：e5：27-34.

[7] Fang A. The chronicle of the Three Kingdoms, vol. I. Cambridge MA: Harvard University Press, 1952.

[8] 西晋·陈寿. 三国志. 卷一魏书一. 武帝纪第一.

[9] 西晋·陈寿. 三国志. 蒋济传. 注引裴松之.

[10] 西晋·陈寿. 三国志. 卷六魏书六. 董二袁刘传第六. 袁绍字本初. 注引《魏氏春秋》所载袁绍讨伐曹操檄文. 197.

[11] Karafet T M, Mendez F L, Meilerman M B, et al. New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. Genome Res, 2008, 18: 830-838.

[12] Howard T D, Bleecker E R, Stine O C. Fluorescent allele-specific PCR (FAS-PCR) improves the reliability of single nucleotide polymorphism screening. Biotechniques, 1999, 26: 380-381.

[13] Lehmann E L, Romano J P. Testing Statistical Hypotheses. 3rd ed. Springer, 2005.

[14] Weir B S. Genetic Data Analysis. Methods for Discrete Population Genetic Data. Sunderland, MA: Sinauer Associates, Inc. Publishers, 1990.

[15] Bland J M, Altman D G. The odds ratio. Brit Med J, 2000, 320: 1468.

[16] King T E, Jobling M A. What's in a name? Y chromosomes, surnames and the genetic genealogy revolution. Trends Genet, 2009, 25: 351 - 360.

[17] King T E, Ballereau S J, Schürer K E, et al. Genetic Signatures of Coancestry within Surnames. Curr Biol, 2006, 16: 384 - 388.

[18] Yan S, Wang C C, Li H, et al. An updated tree of Y chromosome haplogroup O and revised phylogenetic positions of mutations P164 and PK4. Eur J Hum Genet, 2011, 19 (9): 1013 - 1015.

6.3 曹操叔祖的古 DNA 结果与曹操后世子孙相符

《人类遗传学报》第 57 卷第 3 期发表了全国曹氏 Y 染色体调查的结果[1]，笔者用 Y 染色体上 100 个单核苷酸位点(SNP)对全国各地曹姓 79 个家族的 280 个男性以及其他姓氏的 446 个男性进行分型研究。结合家谱材料，发现 O2 - F1462 是唯一在宣称是曹操后代的家族中显著高频出现的 Y 染色体单倍群(Fisher 精确检验，$P=9.323\times10^{-5}$)，有 92.71%的可能性就是曹操的 Y 染色体类型。而 O3 - 002611 是在其他曹姓家族中出现频率最高，也是唯一在宣称是西汉丞相曹参直系后裔的 5 个曹姓家族中都出现的单倍群，极可能是曹参的类型。笔者认为魏武帝曹操不太可能是曹参的后裔，遗传学证据并不支持曹操自称的士族出身。

虽然此次曹姓家族 Y 染色体调查是将遗传学用于古代史研究的一个成功范例，但并没有最终探明曹操的身世之谜。曹操的祖父曹腾在东汉为宦官之首，位高权重，且出自谯县旧家，他的养子可以继承官爵封地，决不会随便过继。按照当时过继承宗祧的基本原则，曹腾的养子，也就是曹操的父亲曹嵩，应该是从本宗他房中过继[2]。然而，曹操的政敌袁绍在攻曹的檄文中写到“父嵩乞丐携养”[3]，说他是路边捡来的乞丐。关于曹操身世的各种说法由此流传，且纷纷扰扰争论了近 2 000 年。在通过现代的曹姓家族明确了曹操与曹参的关系后，笔者又通过亳州曹操家族墓群的古人遗骸来辨析曹操身世。

曹氏宗族墓群，位于安徽省亳州市，是曹操家族一个规模宏大的墓群。主要包括元宝坑汉墓、董园汉墓、马园汉墓、袁牌坊汉墓、曹四孤堆、刘园孤堆和观音山孤堆等，曹操的祖父曹腾及父亲曹嵩的墓均在其中。这些墓葬大都发掘于 20 世纪 70 年代，大部分墓的墓主人还没有确定。其中元宝坑一号墓出土了许多铭文字砖，对考证墓主身份提供了一定线索，依据出土自元宝坑一号墓的一颗牙齿的磨损度推断墓主的年龄在 50 岁甚至 55 岁以上[4]，由此判断墓主人最可能是曹腾之弟、曹操叔祖曹鼎。曹鼎的 DNA 结果将为揭开曹操身世提供最直接的线索。为此，笔者选取了元宝坑一号墓的一枚牙齿进行了后续的古 DNA 实验。

实验时严格按照古 DNA 实验的操作流程对样本进行处理[5]：① 去污染预处理；

② 样品的钻孔、粉碎；③ DNA 抽提；④ 使用美国 ABI 的 Y - Filer™试剂盒检测 Y 染色体微卫星串联重复位点(STR)。

笔者成功扩增和检测了元宝坑牙齿的 Y 染色体上 12 个 STR 位点：DYS19：15，DYS389I：13，DYS389II：30，DYS390：23，DYS391：10，DYS393：14，DYS437：14，DYS456：16，DYS458：16，DYS635：21，DYS385a/b 为 13/14 或者 12/16。使用除 DYS385a/b 之外的 10 个 STR 位点利用贝叶斯频率法进行单倍群预测[6]，依照 Y 染色体数据库，最可能的单倍群类型是 O2 *（M268 +，F1462 +，PK4 −和 M176 −），可能性为 60.18%，其次是 C3 * - M217 (13.97%)，其余单倍群的可能性都低于 11%。这 10 个位点的数值还与数据库中一例来自安徽的 O2 *（F1462 +，PK4 −和 M176 −）的样本完全一致，更增加了上述单倍群判读的可信性。

在之前的研究中，笔者推定来自辽宁铁岭、安徽舒城、安徽绩溪、江苏盐城、湖南长沙和辽宁东港的六支 Y 染色体单倍群为 O2 * - F1462 的曹氏可能是曹操后代。在本次实验中，笔者又对 F1462 的一个下游位点 PK4 进行了检测[7]，发现长沙以及舒城的少数几例 O2 * - F1462 曹姓为 O2a * - PK4 +，或与曹操无关。值得一提的是，笔者对山东乳山河南村宣称为曹操后裔的曹氏进行调查，发现其 Y 染色体类型也是 O2 *（F1462 +，PK4 −和 M176 −），或许也是曹操之后。

为了更准确解析 Y 染色体为 O2 * 的几支曹氏的远近关系，笔者使用 15 个 STR 位点绘制邻接法网络图进行分析(图 6 - 5)。安徽的 O2 * 曹氏在网络图中倾向于紧密地聚在一起，元宝坑牙齿(YBK)与安徽绩溪曹氏和亳州曹氏邻接，显示了较近的亲缘关系。一部分山东乳山曹氏与辽宁东港曹氏 STR 位点完全一致，另一部分与其仅差一步突变，均显示了极近的父系亲缘关系，这也与他们的家谱记载相吻合。O2 * 的曹氏很可能就源

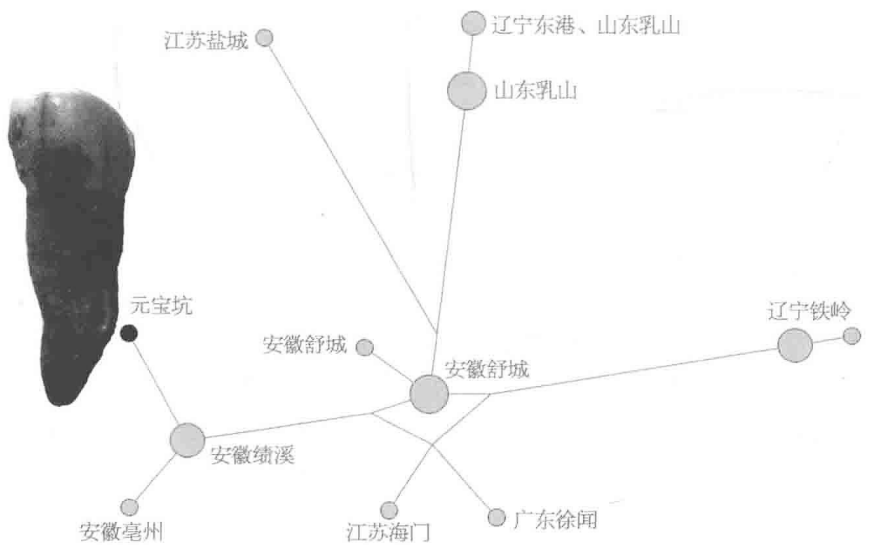


图 6 - 5　全国范围的 O2 * 曹氏家族的 15 个 Y 染色体 STR 位点(Y - filer 除去 DYS385a/b)邻接法网络图分析

自安徽，并且由于曹魏政权的辉煌而经历过轻度的人口扩张。

曹操家族墓葬群的元宝坑一号墓的古 DNA 结果支持之前通过现代曹氏家族所做出的推断：O2 * - F1462 应该就是曹操的 Y 染色体类型，同时也为进一步解开曹操身世之谜提供了线索，因元宝坑墓主人的 Y 染色体类型极可能与曹操一致，所以曹操的父亲应是从曹腾本宗室过继而非抱养自街头的乞丐。

参考文献

[1] Wang C, Yan S, Hou Z, et al. Present Y chromosomes reveal the ancestry of Emperor CAO Cao of 1 800 years ago. J Hum Genet, 2012, 57: 216 - 218.

[2] 韩昇. 曹操家族 DNA 调查的历史学基础. 现代人类学通讯, 2010, 4: 46 - 52.

[3] 西晋·陈寿. 三国志. 卷六魏书六. 董二袁刘传第六. 袁绍字本初. 注引《魏氏春秋》所载袁绍讨伐曹操檄文. 197.

[4] 李淑元，李辉. 从牙齿磨损度推断安徽亳州元宝坑一号墓墓主身份. 现代人类学通讯, 2010, 4: 53 - 56.

[5] Pääbo S, Poinar H, Serre D, et al. Genetic analyses from ancient DNA. Annu Rev Genet, 2004, 38: 645 - 679.

[6] Athey T W. Haplogroup prediction from Y - STR values using a bayesian-allele-frequency approach. J Genetic Genealology, 2006, 2: 34 - 39.

[7] Yan S, Wang C C, Li H, et al. An updated tree of Y chromosome haplogroup O and revised phylogenetic positions of mutations P164 and PK4. Eur J Hum Genet, 2011, 19: 1013 - 1015.

6.4 鄱阳操姓血缘上并非出自曹操

2009 年末，河南省安阳市宣布发现了曹操墓，出土了一男两女三具尸骨，考古人员推测男尸可能是曹操。消息一发表，立即在国内舆论界掀起轩然大波，质疑者甚众，引起对曹操墓真伪的争议。当时由于没有可以作为标准的曹操家族 DNA 特征来作为骨骸验证的对照，遗传学无法就骨骸的身份做出判断。能否找到明确的曹操家族 DNA 特征，从而分析曹操的身世，对曹操及汉魏历史的研究都有极为重要的意义。

同时，复旦大学正在积极推进历史人类学学科的发展，期望使用遗传学工具来准确定位历史上的民族和家族，解决单纯的历史学方法所无法解决的问题。曹姓作为一个重要的姓氏，同时又具有不大不小的适当规模，被选择为第一个深入研究的姓氏。2010 年 1 月 26 日，复旦大学现代人类学教育部重点实验室和历史学系联合启动复旦大学文科科研推进计划项目——曹操后代的历史人类学调查（B200902），利用姓氏和 Y 染色体相关性原理，调查分析曹氏 Y 染色体，进而给曹操身世血统的研究提供科学的证据，也开始了人类基因调查从民族分析向家族分析的转变，从史前时期向历史时期的迈进[1]。鉴于操姓源自曹姓的说法由来已久，在项目具体实施过程中，笔者也把操姓纳入调查范围，对操姓的渊源做了详细的分析。

操姓主要有如下两支[2-4]。

6.4.1　鄱阳郡操氏

据操氏家谱谱序记载：西晋泰始二年(266 年)，司马炎废魏帝，建立晋政权后，疯狂地杀害曹魏皇族。曹操嫡孙曹休举家逃往鄱阳郡新蛊(今江西鄱阳)，为避免被司马氏政权斩尽杀绝，遂以曹操名为姓，改曹为操。

自南唐时期起，由于人口繁盛，操氏各宗支纷纷从鄱阳外迁，最大的一支迁往金华府、甬城(今浙江宁波)和绍兴诸地。有的迁往北方的河北、山东、甘肃和陕西西安；有的迁往安徽池州、徽州、亳州，及江苏扬州、无锡；还有的迁到我国台湾。

例如，河南鹿邑操氏就是于明代天顺二年(1458 年)从江西饶州府鄱阳县迁至陈州，

图 6－6　光绪年操氏家谱中所载康熙五十年(1711 年)反驳曹操来源说文章

再迁到现居地。

6.4.2 重庆长寿操氏

明正德年间(1506—1521年)的族谱记载：操公讳节，祖籍山西牟道县，先人明永乐年(1403—1424年)迁燕，操节膺五经魁，在朝为官，后因戍边有功，明正德九年(1514年)，柱奉谕赐祭不留停候、继授车骑将军、终任两湖总兵官。其子操洁清亦同朝为官，封晋赠太傅、太师、太保，刑部侍郎，父子当朝，班班可考。后因奸宦陷害，操节弃官入蜀，定居"小歌山"(今重庆长寿)数百余载。相传操节在临终遗嘱曰：我操氏始祖乃周武王姬发之后代、周昭考公第十三子之第二十七代子孙，因先祖在当时社会地位崇高，于孔子所著《漪兰操》一书，尤能独解其奇妙，世人恭称其为琴操家，他的后人始以操为氏。

为厘清操姓和曹操的关系，笔者所在的课题组赴安徽潜山和浙江嵊州采集操姓血样。潜山操姓源自鄱阳，均称祖先为妙荣公。但其对自身源流有两种说法：一是与鄱阳操姓一致，源自曹操的四十二代孙改姓，但是在光绪年的家谱中专门撰文批驳了这种说法(图6-6)；二是源自鄡(Zào)姓，在新编家谱中开宗明义地说明(图6-7)。鄡姓或源于姬姓，出自春秋时期郑国鄡邑，或源于狄族，出自春秋时期狄族分支鄋瞒国。浙江嵊州操姓则明确宣称来自鄱阳，出自曹操。笔者选取Y染色体上涵盖东亚所有单倍群的100个SNP位点[5]，基因分型采用SNaPshot，PCR产物纯化后在ABI 3730测序仪上分析。为更清楚勾画各地操姓之间的亲缘关系，还用YFiler进行了17-STR位点的检测。

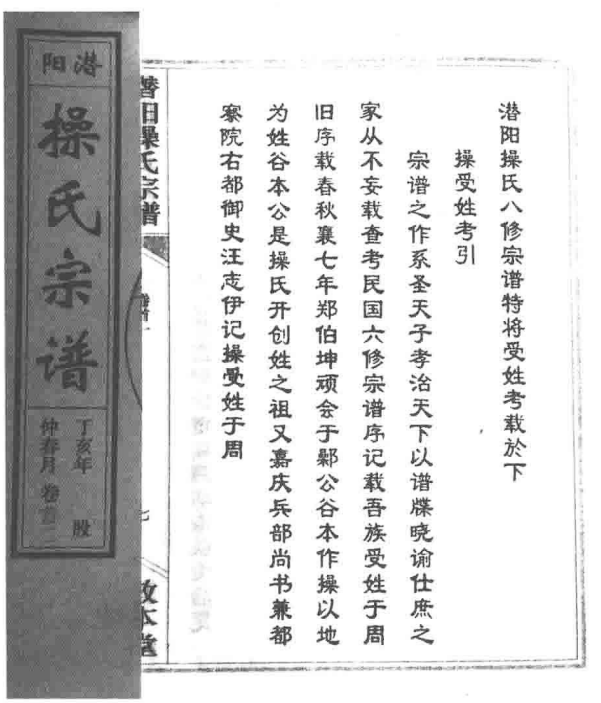


图6-7 新编家谱中记载操姓源于春秋时期的鄡公

实验结果表明，安徽潜山和浙江嵊州的操姓 Y 染色体均属于单倍群 O3a3 * - P164 + 和 M134 −。该单倍群在全国汉族中约占 3.3%[6]，除出现在操姓中，还极其少量地出现于曹、张、陈、王等姓氏中。鄱阳郡操姓的 Y 染色体与笔者推定的曹操 Y 染色体 O2 - M268 并不一致[5]。为进一步检验极少数(约 1%)的曹姓 O3a3 * 与操姓的关系，笔者用 15 个 STR 位点(YFiler 除去 DYS385a、b)做了所在实验室所有 O3a3 * 样本 Network 分析(图 6 - 8)。在 Network 图中，鄱阳郡操姓呈较小的星状扩散趋势，表明操姓内部的遗传距离非常小，他们都有着非常相近的血缘。但是操姓与包括曹姓在内的其他姓氏均相距较远，推断他们的共同祖先远在史前时期，所以在历史时期中没有血缘关系。鄱阳郡操姓 Y 染色体显出了极好的单一源流，这在姓氏的 Y 染色体分析中非常少见。在全国范围内，操姓是否也属单起源，就需要看对重庆长寿操姓的进一步分析结果。

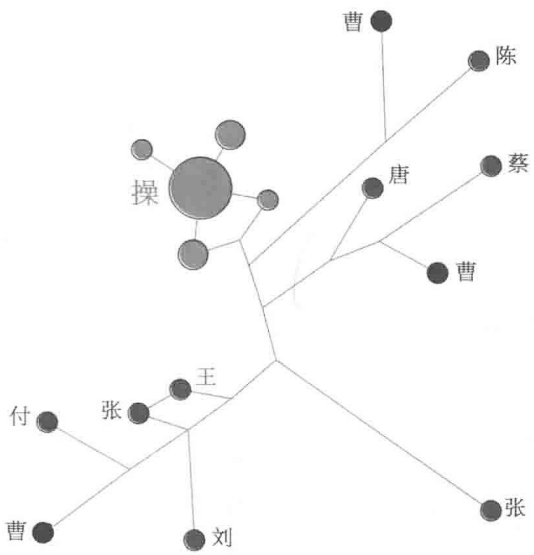


图 6 - 8　单倍群 O3a3 * 的 Network 分析

综上所述，鄱阳郡操姓应该不是曹操的后代。而与姬姓及其衍生的大量姓氏的 Y 染色体类型也不吻合，应该也不是姬姓的衍生姓氏。另外，O3a3 * - P164 这种 Y 染色体类型并不见于北方少数民族，狄族起源的观点也不可取。所以彻底解决操姓起源问题还有待于中华民族姓氏遗传分析的全面完成。

参考文献

[1] 韩昇. 曹操家族 DNA 调查的历史学基础. 现代人类学通讯，2010，4：46 - 52.

[2] 中华操氏宗亲网 http：//www. cszq. com. cn/.

[3] 华人百家姓论坛 http：//www. cnsurname. com/.

[4] 谈洁. 操姓. 环球人文地理，2010，3：239.

[5] Wang C, Yan S, Hou Z, et al. Present Y chromosomes reveal the ancestry of Emperor CAO Cao of 1800 years ago. J Hum Genet in press, 2011.

[6] Yan S, Wang C C, Li H, et al. An updated tree of Y chromosome haplogroup O and revised phylogenetic positions of mutations P164 and PK4. Eur J Hum Genet, 2011, 19(9): 1013-1015.

6.5 赛典赤·瞻思丁和郑和的波斯祖源

因为姓氏与Y染色体有很好的平行对应关系，笔者在前面的研究中就结合姓氏家谱材料，利用Y染色体追溯三国曹操的身世[1,2]。对于同一个民族内的家族的追溯，需要较复杂的统计分析来识别。而对于不同民族，甚至不同种族来源的家族的辨别，由于涉及Y染色体类型差异可能会很大，反而要容易得多。这里，我们继续用这一方法揭开赛典赤和郑和的身世之谜。

赛典赤·瞻思丁为我国元代杰出的政治家，他为昆明乃至云南的发展做出了重大贡献。赛典赤是中亚布哈拉(今乌兹别克斯坦)人，成吉思汗西征时，率部归元，后被派到云南任行省平章政事[3]。中国的许多穆斯林都把赛典赤作为祖先，特别是明代著名航海家郑和、现今在云南的纳姓等[4]。历史文献和家谱均记载纳姓源于赛典赤的长子纳速拉丁。郑和本姓马，也被认为是纳速拉丁的后裔。然而，赛典赤的祖源还一直有疑问，他出生在布哈拉，但他的祖先却是走过了西亚、北非、欧洲和波斯之后才定居中亚的[5]。还有一些谱牒材料却记载赛典赤的祖先在宋代就来到了中国[6]，所以赛典赤是否有波斯祖源还需要进一步印证。

纳剑波在自己的硕士论文中就试图去回答这一问题[7]。他从云南大通搜集了40份纳姓男性样本，检测了Y染色体上的9个STR位点，发现35人有着相同的单倍型(DYS19, 14; DYS389I, 12; DYS389b, 16; DYS390, 22; DYS391, 11; DYS392, 14; DYS393, 11)，2人在此相同单倍型的基础上仅有一步突变。但纳剑波没有检测Y染色体上的SNP位点，所以不能对赛典赤的祖源下确切结论。笔者依照家谱在云南昆明和大通采集了赛典赤和郑和后裔的纳姓和马姓男子的DNA样本，按照最新Y染色体谱系树上的位点对其Y染色体进行分型，还检测了17个Y染色体STR位点。三个纳姓样本中有一个的STR单倍型与纳剑波论文中的纳姓共享单倍型一致(DYS19, 14; DYS389I, 12; DYS389b, 16; DYS390, 22; DYS391, 11; DYS392, 14; DYS393, 11; DYS437, 15; DYS438, 11; DYS439, 12; DYS448, 19; DYS456, 17; DYS458, 16; DYS635, 24; H4, 12; DYS385a, 13; DYS385b, 17)，一个马姓样本也仅在DYS19位点上有一步突变，这两个样本的Y染色体单倍群是L1a-M76。

Y染色体单倍群L1a-M76主要分布在伊朗东部、巴基斯坦南部和印度，低频分布于沙特阿拉伯、尼泊尔和中亚(图6-9a)。为进一步精细解析L1a的纳姓和马姓与其他相关人群的关系，笔者对L-M11单倍群下的样本用Y-STR位点来构建邻接法网络图(图6-9b)。纳姓L1a样本与印度南部的德拉威人和印度裔马来西亚人、4个巴基斯坦样本(3个俾路支人和1个莫克兰人)、4个阿富汗样本(1个巴格拉姆人，1个阿拉伯人，1个Balush和1个乌兹别克人)聚在了L1a-M76的根部。大多数纳姓样本与根部单倍型

仅差一步突变，他们与一些印度裔马来西亚样本聚在一起。纳姓样本的Y染色体显示出的这种近乎排他的聚类暗示了该支系很可能经历了很强的奠基者效应和后续的人口扩张。

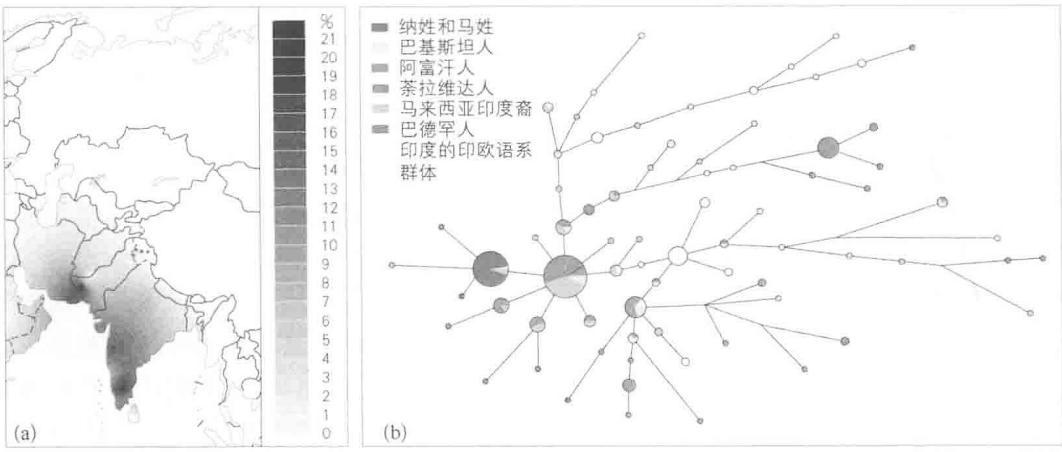


图6-9　Y染色体单倍群L1a-M76的分布

(a) Y染色体单倍群L1a-M76的分布(该地图是基于google地图使用Surfer软件的Kriging方法绘制的)　(b) 基于7个Y-STR(DYS19, DYS389I, DYS389b, DYS390, DYS391, DYS392, and DYS393)位点的邻接构建的L-M11网络图(数据取自参考文献[8-10])

Y染色体类型属于L1a-M76的纳姓和马姓穆斯林的起源或可追溯到南亚的西部，而该地区在历史上又长期被波斯占领。非常严格的伊斯兰宗谱把纳姓和马姓穆斯林和几百年前赛典赤、郑和联系在一起，所以遗传学证据支持赛典赤和郑和的波斯祖源。

参考文献

[1] Wang C, Yan S, Hou Z, et al. Present Y chromosomes reveal the ancestry of Emperor CAO Cao of 1 800 years ago. J Hum Genet, 2012, 57: 216-218.

[2] Wang C C, Yan S, Yao C, et al. Ancient DNA of Emperor CAO Cao's granduncle matches those of his present descendants. J Hum Genet, 2013, 58: 238-239.

[3] Arnold S T W. The preaching of Islam: a history of the propagation of the Muslim faith. Daryaganj New Delhi: Adam Publishers, 1913.

[4] Dillon M. China's Muslim Hui community: migration, settlement and sects. New York: Routledge Curzon, 1999.

[5] 纳为信. 赛典赤·瞻思丁波斯身世考略. 回族研究, 2004, 54: 19-24.

[6] 李清升. 所非尔入宋与赛典赤归元考略——谱牒与史志记载的比较研究. 回族研究, 2004, 54: 25-29.

[7] 纳剑波. 中国两个穆斯林民族和四个汉族群体的遗传多样性研究. 北京: 中国协和医科大学, 2002.

[8] Pamjav H, Zalán A, Béres J, et al. Genetic structure of the paternal lineage of the Roma people.

Am J Phys Anthropol, 2011, 145: 21 - 29.

[9] Lacau H, Gayden T, Regueiro M, et al. Afghanistan from a Y chromosome perspective. Eur J Hum Genet, 2012, 20: 1063 - 1070.

[10] Thanseem I, Thangaraj K, Chaubey G, et al. Genetic affinities among the lower castes and tribal groups of India: inference from Y chromosome and mitochondrial DNA. BMC Genet, 2006, 7: 42.

第7章　Y染色体与相关学科发展

Y染色体是男性的性染色体，所以Y染色体上的各种突变，很有可能与男性的生育能力有关联。Y染色体*AZFc*区域内的多个基因家族对于人类精子发生具有重要作用。在以前针对荷兰人、西班牙人、意大利人的研究中，一类*AZFc*区域的部分缺失——"gr/gr缺失"会减少*AZFc*基因家族的拷贝数，并被认为是精子发生障碍的一个重要风险因子。然而，有研究发现可育的健康法国和德国男性中也有gr/gr缺失，这与先前的"gr/gr缺失会导致男性不育"的观点不一致。在本章中，笔者发现gr/gr缺失并不会增加中国汉族男性精子发生障碍的风险。对来自东亚8个人群中886个个体的调查表明：gr/gr缺失是常见的(约占8%)。进一步分析表明，*DAZ*基因家族中*DAZ1/DAZ2*拷贝的缺失是东亚人gr/gr缺失的主要亚型，而*DAZ1/DAZ2*拷贝在欧洲人中却对正常的精子发生十分重要。不同类型的*AZFc*部分缺失对精子发生有不同的影响，*AZFc*中不同拷贝之间的功能差异可能是gr/gr缺失在不同群体中呈现出差异性的精子发生作用的可能原因。

*AZFc*区域的完全缺失是最常见的导致男性不育的原因之一；然而针对*AZFc*区域的部分缺失在精子发生中的作用存在争议。为了研究*AZFc*区域部分缺失在精子发生障碍中的作用以及*AZFc*区域完全缺失和部分缺失之间的关系，对上述缺失进行分型，并对*DAZ*基因拷贝数进行定量分析以及Y染色体的单倍型分析。所研究的样本包括7个*AZFc*完全缺失的家系、296个男性不育患者和280个健康中国男性。实验发现，gr/gr和b2/b3的缺失都与精子形成失败相关联。在一个谱系中，观察到*AZFc*区域的完全缺失是由于gr/gr缺失引起的，说明*AZFc*区域的完全缺失是由于部分缺失导致的。另外，定义了一个新的gr/gr缺失的Y染色体单倍型类群和证实了以前被报道的在单倍型类群N的b2/b3稳定缺失。从这些数据中第一次证明*AZFc*区域的部分缺失可以增加*AZFc*区域完全缺失的风险。*AZFc*区域部分缺失变成完全缺失的敏感性需要更多的检验，尤其是在群体中或者*AZFc*区域部分缺失富集的Y染色体。

Y染色体在人群间的多样性分布与语言语系的分类有着最紧密的相关性。世界语言多样性的研究最近有了突破性进展，不同区域语言的语音多样性差异明显。欧亚大陆，特别是东亚的语言语音复杂度最高，而非洲稍低，美洲和大洋洲最低。全世界的语音复杂度呈现出从里海南岸向外的最大斜率的衰减，这与Y染色体超单倍群CT(CDEF)

扩张的历史地理学过程是吻合的。全世界语言的元音系统有着明显的分布规律。大多数语言具有5～7个元音，只有日耳曼语族和吴语方言发展出了12个以上的元音，吴方言中的奉贤傷傣话达到了最高20个元音音位。类型学上元音系统可以按照包含元音的性质分为数种形式。非洲只有3部式，欧洲、西亚、大洋洲和美洲发展出4部式，只有东亚的汉藏语系、阿尔泰语系、乌拉尔语系发展出明确的5部式。而只有汉语发展出了最复杂的6部式。元音系统类型学的分布规律提示，语言的演化是有规律可循的。

古代样本的Y染色体研究有助于详细分析Y染色体演变和人群演化的历史。古DNA研究以分子生物学技术为基础，以古生物的DNA为研究对象，是一个新兴领域。20余年来，古DNA实验技术不断发展。分子克隆、PCR、下一代测序技术、引物延伸捕获和芯片杂交捕获等扩增和测序技术的不断涌现，分别引领了古DNA研究的三次革命，极大地推动古DNA研究发展、成熟。从229 bp斑驴的线粒体DNA到尼安德特人基因组草图，古DNA研究取得一系列的突出成果，解开了众多千万年来的谜题。

7.1 东亚人群中 *AZFc* 部分缺失并不提高精子发生障碍风险

7.1.1 研究背景

人类Y染色体长臂的缺失能够导致精子发生障碍，这是男性不育的主要原因之一[1,2]。迄今已经有至少3个精子发生相关位点（*AZFa*，*AZFb* 和 *AZFc*）被定位到Yq11，其中 *AZFc* 是在包括东亚人的多个人群中缺失最频繁的区域[3-5]。需要特别指出的是，*AZFc* 由多个长片段的重复序列组成（图7-1a），这使得这个区域更容易受到重复序列之间的非等位同源重组的影响，并导致不同类型 *AZFc* 缺失的发生[6,7]。*AZFc* 完全缺失（b2/b4缺失）会引起无精子症和少精子症[6,8]。因此，*AZFc* 区域内睾丸特异表达的

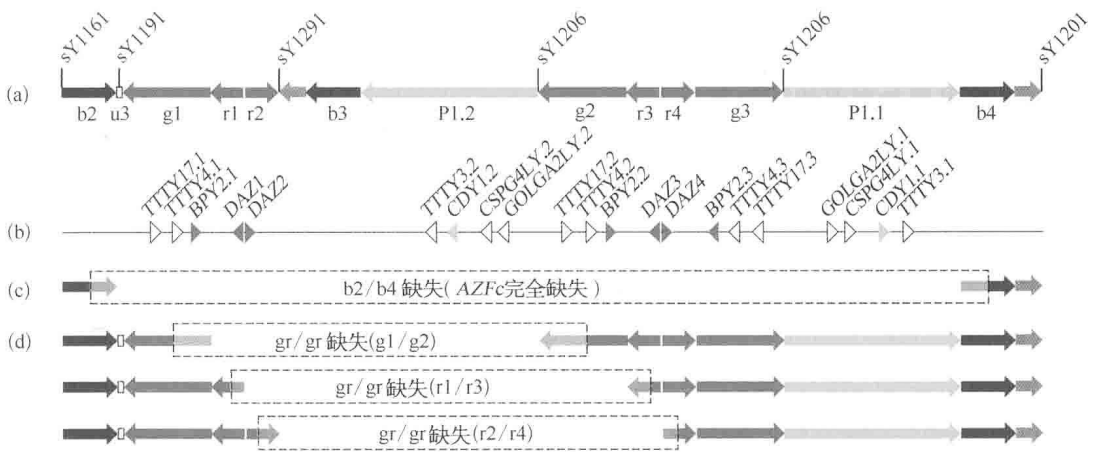


图7-1 *AZFc* 区域基因和重复序列的结构，以及常见的基因组重排形式

(a) GenBank参考序列中 *AZFc* 基因区域重复序列结构[6]；(b) 3个编码蛋白的基因家族（实心三角形）和5个非编码基因家族（空心三角形），箭头方向即5′→3′[30]；(c) *AZFc* 的完全缺失（b2/b4缺失）常导致无精子症或少精子症[6]；(d) 3种不同形式的gr/gr缺失引起 *AZFc* 区域基因家族的拷贝数差异[9]

基因[9]在人类精子发生过程中有重要的作用。所有的 *AZFc* 基因都是多拷贝的(图7-1b),并且能够分成8个基因家族[6,9]。尽管 *DAZ*、*CDY1* 以及其他 *AZFc* 基因家族与精子发生有密切关系[6,10-12],但是对每个基因家族中各个成员的生精作用还了解很少。最近观察到的 *AZFc* 部分缺失和对其在精子发生过程中的作用的研究可以为探索这个问题提供有效的研究途径。

长度为1.6Mb的gr/gr缺失并没去除任何一个 *AZFc* 基因家族,而只是减少它们的拷贝数。在荷兰、西班牙和意大利等人群中这一缺失被鉴定为精子发生障碍的风险因子[9,13,14],这提示 *AZFc* 基因家族足够剂量的基因拷贝数可能对于正常的精子发生过程是必要的。但其他研究发现,在健康的法国人和德国人中gr/gr缺失也发生[12,15];尽管这个缺失在不育男性中大量出现,但是在不育患者和健康对照之间并没有明显的分布差异[12,15]。Repping等[9]发现被检测的所有Y染色体D2b单倍群(*hg*D2b)男性都是gr/gr缺失。除此之外,Machev等[12]通过 *DAZ* 和 *CDY1* 基因家族序列多样性(SFV)分析可将gr/gr缺失分为四类,其中在法国人中只有 *DAZ3/4 - CDY1a* 型与不育有密切关系。因此,为了探索部分 *AZFc* 缺失在精子发生障碍中的作用,需要在不同群体中进行更为严谨的实验。

7.1.2 材料和方法

1. 群体样本

在群体研究中,笔者采集了无亲缘关系的886例个体,他们来自东亚的8个群体(汉族、苗族、瑶族、蒙古族、纳西族、土家族、维吾尔族和壮族)。根据种族和地理位置,这些样本分为如下15个群体:蒙古族-1(内蒙古)、蒙古族-2(新疆)、维吾尔族(新疆)、北汉-1(甘肃)、北汉-2(内蒙古)、北汉-3(山东)、北汉-4(山东)、南汉-1(广东)、南汉-2(四川)、南汉-3(云南)、土家(湖南)、苗族(湖南)、瑶族(广西)、壮族(广西)、纳西族(云南)。

2. 病例组和对照组

87例无亲缘关系的精子发生障碍患者都采集自上海仁济医院。根据精液浓度的不同,将这些患者分为三类:非阻碍性无精子症(精液中无精子,49人),严重少精子症(<500万个/ml,16人)和少精子症(500万~2 000万个/ml,22人)。作为对照组的89例健康对照(精子浓度>6 000万个/ml,且精子形态活力正常)为采集自上海仁济医院的捐精者。对精液中精子的数量、活力和形态进行分析都依据世界卫生组织的标准[16]。所有的患者和精液捐献者都来自中国东部的汉族人。

3. Y染色体单倍型分析

基因组DNA提取自人体外周血。为了进行Y染色体谱系分析,按照前人的研究[17]对M9、M89、M95、M117、M119、M122、M130、M134、M175和YAP 10个双等位遗传标记进行检测。这些标记在东亚人中具有很高的信息量[18,19]。根据Y染色体谱系定义了10个单倍群(图7-2)。

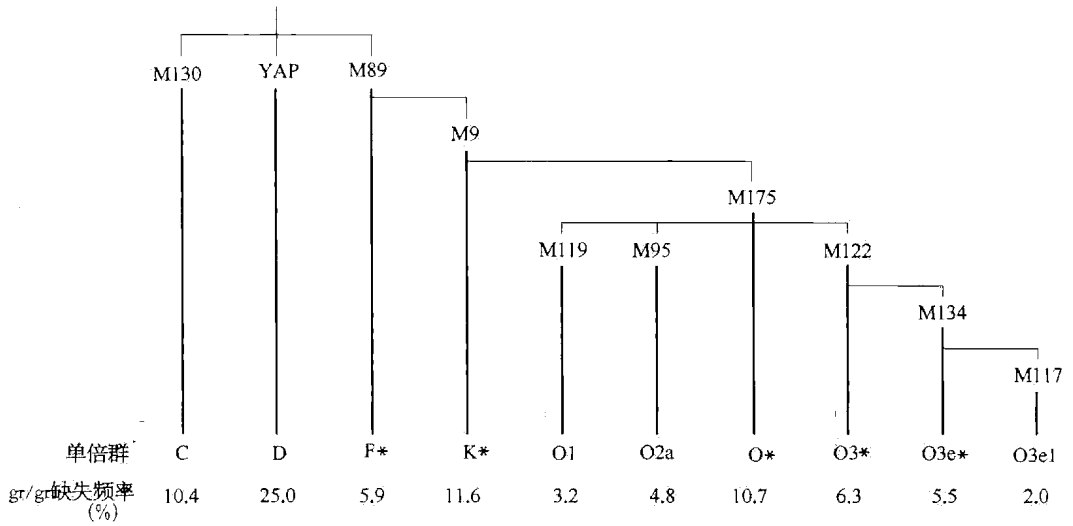


图 7-2 调查群体中东亚人的 Y 染色体谱系关系(基于国际 Y 染色体命名委员会,2002)
分枝上标明了各个单倍群所对应的突变位点名称,底部的数字是 Y 染色体单倍群中 gr/gr 的缺失频率。

4. STS 分析

根据 Repping 等[9]方案中的+/— STS 模式:SY1161(+)、SY1191(+)、SY1201(+)、SY1206(+)和 SY1291(—)对 gr/gr 缺失进行鉴定。这些 STS 位点的位置见图 7-1a。SY1291 位点的缺失至少通过三次重复的 PCR 实验来验证。

5. 实时定量 PCR

用 SYBR-Green 染料去检测 *DAZ* 基因的拷贝数。M159 是一个 *AZFc* 区域外的单拷贝 Y 染色体位点,用来作为参考位点。分别位于 *DAZ* 基因两端的位点 *DAZ*-SNV 1 和 *DAZ*-SNV 5 作为测试位点。通过重复扩增 24ng、12ng、6ng、3ng 和 1.5ng 标准 DNA(无 gr/gr 缺失)各三次来绘制 M159、SNV 1 和 *DAZ*-SNV 5 的标准曲线(图 7-3)。

因为循环数目(C_t)的阈值是初始模板量(N_t)的一个变量[20,21]。通过 SDS 软件(Applied Biosystems)计算 C_t 值和标准曲线得到测试模板和标准模板的相对值($N_{t(\text{test})}/N_{t(\text{standard})}$)。Y 染色体位点的 N_t 值是通过拷贝数(N_c)和用来扩增的 Y 染色体数目(N_y),根据公式 $N_{t(\text{test})}/N_{t(\text{standard})} = N_{c(\text{test})}/N_{c(\text{standard})} \times N_{y(\text{test})}/N_{y(\text{standard})}$ 计算。与 *DAZ* 基因不同,在 M159 位点没有拷贝数变异,即 $N_{c(\text{test})}/N_{c(\text{standard})} = 1$。因此,对于 M159 $N_{y(\text{test})}/N_{y(\text{standard})} = N_{t(\text{test})}/N_{t(\text{standard})}$。因为用来分析 M159 和 *DAZ*-SNV 的 DNA 量是相等的,故这两个位点的 $N_{y(\text{test})}/N_{y(\text{standard})}$ 值应该相等。因此根据 *DAZ*-SNV 和 M159 的 $N_{t(\text{test})}/N_{t(\text{standard})}$ 比值可以估计测试和标准样品在 *DAZ*-SNV 区域的相对拷贝数。

用来进行实时定量 PCR 的引物是:M159(正向:5′-ATTGGATTGATTTCAGCCTTC-3′;反向:5′-ATTTTATTTTCTGTGTTCCTTGC-3′);*DAZ*-SNV 1(正向:5′-CATCCAACCCCTTATATCTCA-3′;反向:5′-CACAATCACTGTTAGCGTCAC-

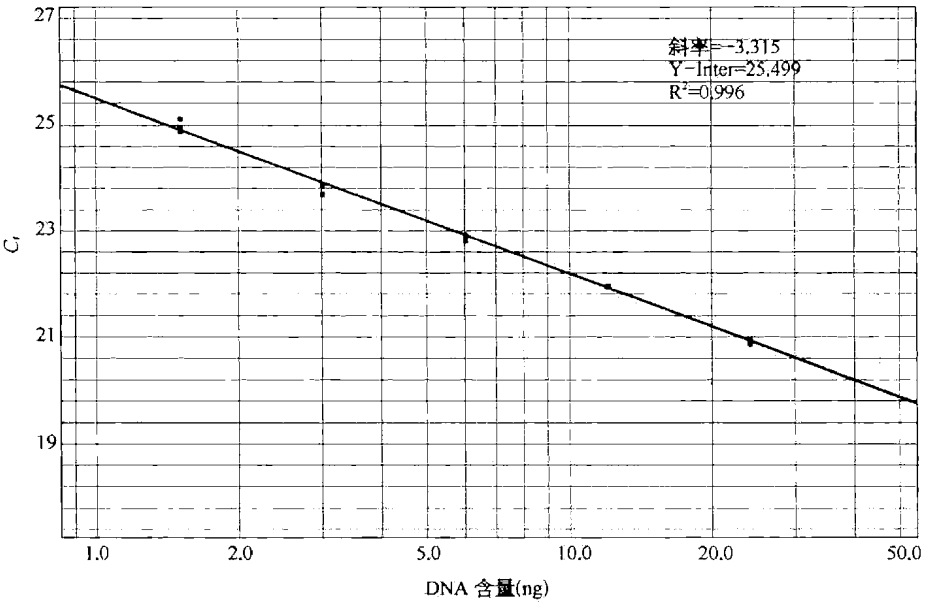


图 7-3　M159 位点的标准曲线(通过 C_t 值来确定待检测 DNA 的含量)

3′)；*DAZ* - SNV 5（正向：5′- CTGGAACTTGTGTGTATCCCT - 3′；反向：5′- TTACAAACATTATGCTGAGTGAAA - 3′)。

实时定量 PCR 是在 ABI Prism™ 7900 高通量序列检测系统和 ABI Prism™ 384 孔优化反应板中进行的。每一个位点都进行独立扩增，反应体积为 10 μl，其中包含上下游引物各 0.5 mmol，2.5 μl ABgene Absolute™ QPCR SYBR Green Rox Mix 和 4 μl DNA 溶液。所有的扩增在相同的条件下进行：95℃ 15 min，然后 40 个循环的 95℃ 15 s 加 60℃ 1 min。扩增后从 60～95℃以一个线性速率(0.1℃/s)进行熔解曲线分析以排除非特异产物的出现。扩增结果用 SDS 软件进行分析(版本 2.1，Applied Biosystems)。

6. *DAZ* - SNV 分析

为了研究 *DAZ* 基因的拷贝数在患者和正常者的差异，利用 5 个双等位的 *DAZ* - SNV 位点(Ⅰ～Ⅴ)和一个 *DAZ3* 拷贝特异的 STS (Y - *DAZ3*)。Fernandes 等[22]在 *DAZ* - SNV 分析中已经对等位基因 A/B 进行了命名和分析。笔者在研究中做了一些改变。设计了如下的新引物。*DAZ* - SNV Ⅰ(正向：5′- CACAGGCACTCAGTAACTATCTC - 3′；反向：5′- CCAAAGGCTAAGGTGTAAGG - 3′；产物长度：630 bp)，*DAZ* - SNV Ⅳ（正向：5′- AGGAATGGCGGTATTAACT - 3′；反向：5′- AACGTCTTCCAGGTATCAAC - 3′；产物长度：300 bp)。并且用 Hha Ⅰ对 *DAZ* - SNV Ⅰ进行限制性酶切。GenBank 中的拷贝特异性等位基因参考序列如下：等位基因 A 对应 *DAZ1/DAZ2/DAZ3*，等位基因 B 对应 *DAZ4*，*DAZ* - SNV Ⅱ(等位基因 A 对应 *DAZ1*，等位基因 B 对应 *DAZ2/DAZ3/DAZ4*)，*DAZ* - SNV Ⅲ(等位基因 A 对应 *DAZ2*，等位基因 B 对应 *DAZ1/DAZ3/DAZ4*)，*DAZ* - SNV Ⅳ(等位基因 A 对应 *DAZ2*，等位基因 B 对应 *DAZ1/DAZ3/DAZ4*)，*DAZ* - SNV Ⅴ(等位基因 A 对应 *DAZ3/DAZ4*，等位基因 B 对应 *DAZ1/DAZ2*)[22]。

7.1.3 结果和讨论

1. 东亚人群体调查

笔者调查了 gr/gr 缺失在东亚人群中的分布情况，通过十/一STS 分析发现总共有 71 个 gr/gr 缺失的男性(表 7-1)，并且东亚人的 gr/gr 缺失频率(约 8%)明显高于美国人群(2.2%)，甚至高于美国或欧洲人群精子发生障碍患者中的缺失频率(3.7%～5.1%)[9,12,13,15]。这个数字也明显高于中国人群中不育患者的频率(1.6%～2.5%)[23]。男性不育相关的 Y 染色体缺失在普通人群中是罕见的，例如 b2/b4 缺失[6]，而在东亚人群中 gr/gr 缺失的高频分布以及 gr/gr 缺失在 D2b 单倍群中被固定(100%存在)[9]。这些事实表明 *AZFc* 部分缺失对东亚人精子发生障碍的作用有限。gr/gr 缺失不仅在白种人和东亚人之间存在明显的分布差异，而且 gr/gr 缺失在多个东亚人群之间也存在很明显的频率差异，土家族和南汉-1 最高(15.3%)，瑶族最低。另外，gr/gr 缺失在所有 10 个 Y 染色体单倍群中都有分布，但是其频率差异明显(2.0%～25.0%)(表 7-1，图 7-1)。需要指出的是，年代古老的单倍群往往具有高频率的 gr/gr 缺失。考虑到重复性的突变事件能够增加 gr/gr 缺失的频率，而净化选择(如精子发生障碍)能够降低缺失频率，在东亚人群中随时间不断积累的缺失提示 gr/gr 缺失所导致的自然选择压力是很有限的。

表 7-1 所研究的群体中 886 个东亚人的 Y 染色体单倍群分布

人群	群体样本量	Y 染色体单倍群和样本量										缺失频率(%)
		C	D	F*	K*	O*	O1	O2a	O3*	O3e*	O3e1	
		115	48	34	121	56	31	104	205	73	99	
蒙古-1	48	0/9	0/2	0/0	1/10	1/3	0/0	0/0	0/11	0/7	0/6	4.2
蒙古-2	56	1/25	0/4	0/8	0/17	0/0	0/0	0/1	0/0	0/1	0/0	1.8
维吾尔	55	0/4	0/2	0/15	1/20	0/1	0/0	0/0	0/5	0/4	0/4	1.8
北汉-1	59	2/7	1/5	0/2	1/11	1/4	1/4	0/1	3/15	0/3	0/7	15.3
北汉-2	58	2/12	0/4	1/2	0/6	0/2	0/1	0/1	1/15	1/5	0/10	8.6
北汉-3	70	3/11	1/4	0/3	4/9	0/3	0/0	0/1	1/30	1/6	0/3	14.3
北汉-4	69	0/6	1/2	1/2	3/13	0/4	0/0	0/1	1/23	0/6	0/12	8.7
南汉-1	57	1/3	0/1	0/0	0/3	0/4	0/4	0/11	1/13	0/5	0/13	3.5
南汉-2	61	0/3	2/2	0/0	1/6	0/4	0/5	2/8	0/16	1/14	0/3	9.8
南汉-3	49	0/5	1/2	0/0	1/8	0/8	0/0	0/2	2/18	0/1	0/5	8.2
土家	72	1/12	2/2	0/0	1/6	1/3	0/5	1/6	3/20	0/3	2/15	15.3
苗族	100	2/14	0/1	0/2	1/9	3/11	0/7	2/14	0/25	0/9	0/8	8.0
瑶族	52	0/1	0/0	0/0	0/0	0/6	0/3	0/14	0/11	0/6	0/11	0.0

（续表）

人群	群体样本量	Y染色体单倍群和样本量										缺失频率（%）
		C	D	F*	K*	O*	O1	O2a	O3*	O3e*	O3e1	
		115	48	34	121	56	31	104	205	73	99	
壮族	40	0/2	0/2	0/0	0/0	0/3	0/2	0/24	1/3	1/2	0/2	5.0
纳西	40	0/1	4/15	0/0	0/3	0/0	0/0	0/20	0/0	0/1	0/0	10.0

注：表中数据表示缺失个体数/单倍群个体数。

2. 关联分析

由于群体样本中所有人的精子发生的表型不明，研究中抽取了87例有精子发生障碍和89例健康精液捐献者来研究gr/gr缺失在精子发生中的作用。通过STS分析鉴定得到9例不育患者携带gr/gr缺失，以及9例健康的捐献者携带gr/gr缺失（表7-2）。这一结果与先前在欧洲人群中的关联分析结果不同[9,12-15]，笔者在病例组和对照组之间中并没有发现明显的gr/gr缺失频率的差异（分别为10.3%，10.1%）。进一步研究发现，在健康捐献者中的9例gr/gr缺失个体的精液浓度（16 090万±5 210万个/ml）和非缺失个体（16 420万±7 050万个/ml）并没有明显差别。根据精液浓度把患者分为三类：非障碍性无精子症、严重无精子症和少精子症，gr/gr缺失频率在三类人群中分别占10.2%、6.3%和13.6%，这和精子发生障碍的严重程度并没有相关性。在两位患者（24号和54号）的父亲中也检测到gr/gr缺失，STS分析表明这两位父亲都是gr/gr缺失，这表明父亲把gr/gr缺失成功传递给了他们的儿子。因为不同Y染色体单倍群男性可能有不同的精子发生能力[24]，为此用Arlequin软件[25,26]对病例和对照样本间可能存在的群体分层进行了Rousset精确检验。结果表明在病例组和对照组之间并没有明显的Y染色体单倍群频率差异，病例组和对照组至少在Y染色体基因型上并没有遗传结构上的分层。结合前述的结果，我们能知道gr/gr缺失并没有增加东亚群体的精子发生障碍的风险。

Machev等[12]在Y染色体单倍群J中发现了和sY1291相对应的O1084/O1085多态性，因此为了使gr/gr缺失的鉴定结果更加可靠而不受遗传多态性的影响，笔者对所有经STS分析鉴定得到的gr/gr缺失样本都进一步进行了*DAZ*-CNV和*DAZ*-SNV分析。

3. *DAZ*基因拷贝的定量分析

笔者用SYBR-Green染料进行了实时定量PCR，分析*DAZ*基因的拷贝数。因为M159在基因组中是个单拷贝的特异性STS位点，因而被选为参考位点。M159的标准曲线见图7-3。在GenBank的参考序列中，在每一个*DAZ*基因中有一个是*DAZ*-SNV1位点，有一个是*DAZ*-SNV 5位点[22]。因为gr/gr缺失去除了原有4个*DAZ*基因拷贝中的2个拷贝[9]，所以测试位点的相对拷贝数在没有影响的个体中是1，在gr/gr缺失的个体中是0.5，利用实时定量PCR对所有18例gr/gr缺失个体（9例患者，9例精液捐献者）以及18例非缺失个体（9例患者，9例精液捐献者）的*DAZ*拷贝数进行了定量分析，结果见图7-4。

表 7-2 Y 染色体单倍群在不育患者和健康精液捐献者中的分布

分组		群体样本量	Y 染色体单倍群									
			C	D	F*	K*	O*	O1	O2a	O3*	O3e*	O3e1
仅 gr/gr 缺失	患者	9	1			2	1			3	1	1
	捐献者	9				3	3	1			1	1
全部	患者	87	8	1		9	4	15	2	26	9	13
	捐献者	89	5		1	7	8	13	4	19	14	18

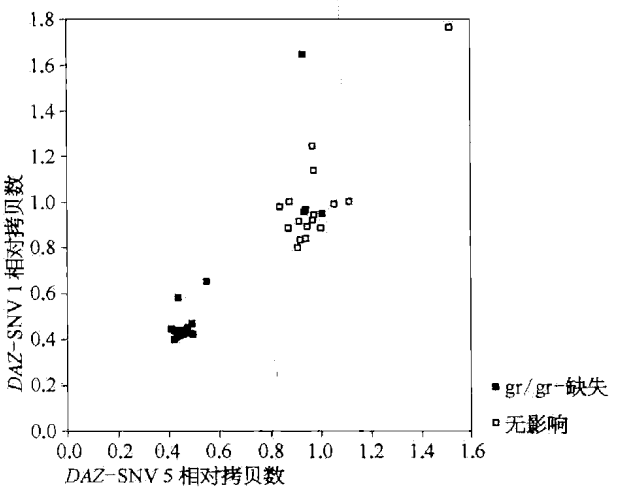


图 7-4 测试 DNA 中 *DAZ*-SNVs 拷贝数相对于标准 DNA 的散点图

在大部分的 gr/gr 缺失样本中，不管是患者还是捐精者，都聚在坐标(0.5,0.5)附近，而大部分非缺失个体聚在坐标(1.0,1.0)附近。同时也观察到 3 个例外：59 号患者(非缺失个体，相对拷贝数：*DAZ*-SNV1 1.76，*DAZ*-SNV5 1.51)，48 号患者(gr/gr 缺失，相对拷贝数：*DAZ*-SNV1 1.65，*DAZ*-SNV5 0.93)，19 号捐精者(gr/gr 缺失，相对拷贝数：*DAZ*-SNV1 0.95，*DAZ*-SNV5 1)(图 7-4)。结果表明，非缺失的 59 号患者可能有 *AZFc* 基因的重复并且携带有多于 4 份 *DAZ* 基因拷贝[9,27]。之后的 *DAZ*-SNV 分析[22]表明，48 号患者和 19 号捐精者都失去了 SNV 多态性，然而属于同一单倍群的非缺失个体却没有 SNV 多态性的丢失。因此，在 48 号患者和 19 号捐精者中 *AZFc* 基因的重复可能补偿了由于 gr/gr 基因缺失导致的 *DAZ* 基因的丢失。除了这 3 个例外，相对拷贝数在非缺失样本(17 例个体)中是分别是 0.95±0.11(*DAZ*-SNV 1)，0.95±0.07(*DAZ*-SNV 5)，然而在 gr/gr 缺失样本(16 例个体)中分别是 0.46±0.07(*DAZ*-SNV 1)，0.46±0.04(*DAZ*-SNV 5)。

4. *DAZ*-SNV 分析

对白种人的观察结果表明，*AZFc* 部分缺失并不总会增加精子发生障碍的风险。考虑到 gr/gr 缺失有 3 种不同的亚型(g1/g2，r1/r2 和 r2/r4)，每一个都可能去除不同的

AZFc 区(图 7－1d)，不同缺失基因拷贝间功能上差异也许是导致不同精子发生表型的原因之一。已有证据表明，*DAZ* 基因的不同拷贝存在功能上的差异。在精子发生障碍的欧洲人群中，已发现高频率的 *DAZ1/DAZ2* 基因的缺失[14,22,28]，而 *DAZ3/DAZ4* 缺失可 100%存在于单倍群 N 中，且该单倍群在欧亚大陆的人群中较为常见[27,29]。上述结果提示，可能只有 *DAZ1/DAZ2* 拷贝对正常的精子发生过程才是重要的，而 *DAZ3/DAZ4* 拷贝可能不是精子发生所必需的。

在这个背景下，根据 Fernandes 等[22]的方法对来自不同 Y 染色体单倍群的所有 18 个 gr/gr 缺失的患者和健康捐献者以及一部分非缺失的个体进行了 *DAZ*－SNV 分析(表 7－3)。因为 *DAZ*－SNV 等位基因的缺失可能源自缺失或者遗传多态性[22,27,29]，根据 Y 染色体的进化关系分为 10 个 Y 单倍群，并且进化上关系相近的单倍群可能表现出类似的 SNV 模式，这样有助于减少在缺失鉴别过程中单倍群特异性遗传多态的影响。在本研究中所有的 gr/gr 缺失，除了那些单倍群 K＊(2 例患者，3 例捐精者)，都是 *DAZ1/DAZ2* 缺失。在单倍群 K＊中，*DAZ1* 缺失由于 *DAZ*－SNV II 的多态性而无法被确认。因此，在单倍群 K＊中的 gr/gr 缺失可能是 *DAZ1/DAZ2* 缺失。在对照组中 *DAZ1/DAZ2* 成对缺失表明这个群体中这对拷贝并不是精子发生所必需的，尽管这对拷贝在欧洲人群精子发生中是必需的[14,22,28]。这表明 *DAZ1/DAZ2* 拷贝在不同群体中具有不同的功能。涉及 *DAZ1/DAZ2* 的 gr/gr 缺失可能属于不同的亚型(g1/g2 或者 r1/r3)，每一个 *AZFc* 区域的特定基因家族中都携带不同的拷贝数。Machev 等[12]发现 *DAZ3/4*－*CDY1a* 的缺失以及 *CDY1a* 的多样性和不育是相关的。因此，除了 *DAZ* 以外的其他 *AZFc* 基因拷贝数的功能性差异也可能是不同人群的 gr/gr 缺失表型差异的原因。

总之，通过对来自东亚 8 个不同民族的 1 062 例男性(包括 176 例临床样本)的研究表明，东亚人存在高频率的 gr/gr 缺失。对有精子发生障碍的不育样本以及在遗传上与之相匹配的健康精液捐献者的 gr/gr 缺失进行鉴定，并进一步对 *DAZ* 基因进行了定量分析。这些结果表明，gr/gr 缺失并没有显著影响东亚人群的精子发生，这和之前有关欧洲人群的报道[9,13,14]不同。有趣的是，大部分的 gr/gr 缺失去除了 *DAZ1/DAZ2* 基因拷贝，而 *DAZ1* 和 *DAZ2* 拷贝在欧洲人群中对正常的精子发生有重要作用[14,22,28]。在不同人群中 gr/gr 缺失的精子发生表型不一致可能是由于 *AZFc* 基因拷贝之间的功能差异所致。如果确实如此，那么鉴定不同群体的 gr/gr 缺失所涉及的基因拷贝有可能揭示 *AZFc* 基因在精子发生过程中的作用。

表 7－3　*DAZ*－SNV 分析一览表

个　体		gr/gr 缺失	Y－DAZ3	SNV I	SNV II	SNV III	SNV IV	SNV V	gr/gr 缺失影响的 DAZ 拷贝数
单倍群 C	患者－22	非	＋	A＋B	A＋B	A＋B	B	A＋B	
	患者－23	非	＋	A＋B	A＋B	A＋B	B	A＋B	
	患者－68	非	＋	A＋B	A＋B	A＋B	B	A＋B	

（续表）

个体		gr/gr缺失	Y-DAZ3	SNV I	SNV II	SNV III	SNV IV	SNV V	gr/gr缺失影响的DAZ拷贝数
单倍群C	捐献者-03	非	+	A+B	A+B	A+B	B	A+B	
	捐献者-62	非	+	A+B	A+B	A+B	B	A+B	
	捐献者-72	非	+	A+B	A+B	A+B	B	A+B	
	患者-48	是	+	A+B	Bb	B	B	A	*DAZ1+DAZ2*
单倍群K*	患者-04	非	+	A	A+B	A+B	A+B	B	
	患者-55	非	+	A	A+B	A+B	A+B	B	
	患者-73	非	−	A	A	A+B	A+B	B	
	捐献者-06	非	+	A	A+B	A+B	A+B	B	
	捐献者-40	非	+	A	A	A+B	A+B	B	
	捐献者-68	非	+	A	A	B	A+B	B	
	患者-24	是	−	A	A	A+B	B	B	*DAZ2*
	患者-33	是	−	A	A	A+B	B	B	*DAZ2*
	捐献者-14	是	+	A	A	B	B	B	*DAZ2*
	捐献者-42	是	+	A	A	B	B	B	*DAZ2*
	捐献者-46	是	+	A	A	A+B	B	B	*DAZ2*
单倍群O*	患者-59	非	+	A	A+B	A+B	A+B	A+B	
	患者-87	非	+	A	A+B	A+B	A+B	A+B	
	捐献者-33	非	+	A	A+B	A+B	A+B	A+B	
	捐献者-69	非	+	A	A+B	A+B	A+B	A+B	
	捐献者-88	非	+	A	A+B	A+B	A+B	A+B	
	患者-56	是	+	A	B	B	B	A	*DAZ1+DAZ2*
	捐献者-37	是	+	A	B	B	B	A	*DAZ1+DAZ2*
	捐献者-38	是	+	A	B	B	B	A	*DAZ1+DAZ2*
	捐献者-61	是	+	A	B	B	B	A	*DAZ1+DAZ2*
单倍群O1	捐献者-49	非	+	A	A+B	A+B	A+B	A+B	
	捐献者-64	非	+	A	A+B	A+B	A+B	A+B	
	捐献者-71	非	+	A	A+B	A+B	A+B	A+B	
	捐献者-65	是	+	A	B	B	B	A	*DAZ1+DAZ2*
单倍群O3*	患者-30	非	+	A	A+B	A+B	A+B	A+B	
	患者-52	非	+	A	A+B	A+B	A+B	A+B	

（续表）

个　体		gr/gr 缺失	Y-DAZ3	SNV I	SNV II	SNV III	SNV IV	SNV V	gr/gr 缺失影响的 DAZ 拷贝数
单倍群 O3 *	患者-65	非	+	A	A+B	A+B	A+B	A+B	
	捐献者-21	非	+	A	A+B	B	B	A+B	
	捐献者-28	非	+	A	A+B	A+B	A+B	A+B	
	捐献者-67	非	+	A	A+B	A+B	B	A+B	
	患者-28	是	+	A	B	B	B	A	*DAZ1+DAZ2*
	患者-54	是	+	A	B	B	B	A	*DAZ1+DAZ2*
	患者-76	是	+	A	B	B	B	A	*DAZ1+DAZ2*
单倍群 O3e *	患者-16	非	+	A	A+B	A+B	A+B	A+B	
	患者-19	非	+	A	A+B	A+B	A+B	A+B	
	患者-75	非	+	A	A+B	A+B	A+B	A+B	
	捐献者-39	非	+	A	A+B	A+B	A+B	A+B	
	捐献者-45	非	+	A	A+B	A+B	A+B	A+B	
	捐献者-60	非	+	A	A+B	A+B	A+B	A+B	
	患者-80	是	+	A	B	B	B	A	*DAZ1+DAZ2*
	捐献者-50	是	+	A	B	B	B	A	*DAZ1+DAZ2*
单倍群 O3e1	患者-27	非	−	A	A+B	A+B	A+B	A+B	
	患者-53	非	+	A	A+B	A+B	A+B	A+B	
	捐献者-56	非	+	A	A+B	A+B	A+B	A+B	
	捐献者-77	非	+	A	A+B	A+B	A+B	A+B	
	捐献者-83	非	+	A	A+B	A+B	A+B	A+B	
	患者-42	是	+	A	B	B	B	A	*DAZ1+DAZ2*
	捐献者-19	是	+	A	B	B	B	A	*DAZ1+DAZ2*

参考文献

[1] Tiepolo L, Zuffardi O. Localization of factors controlling spermatogenesis in the nonfluorescent portion of the human Y chromosome long arm. Human Genetics, 1976, 34(2): 119-124.

[2] Reijo R, Lee T Y, Salo P, et al. Diverse spermatogenic defects in humans caused by Y chromosome deletions encompassing a novel RNA-binding protein gene. Nature Genetics, 1995, 10(4): 383-393.

[3] Kim S W, Kim K D, Paick J S. Microdeletions within the azoospermia factor subregions of the Y

chromosome in patients with idiopathic azoospermia. Fertility and Sterility, 1999, 72: 349 - 353.

[4] Foresta C, Moro E, Ferlin A. Y chromosome microdeletions and alterations of spermatogenesis. Endocrine Reviews, 2001,22(2): 226 - 239.

[5] Sawai H, Komori S, Koyama K. Molecular analysis of the Y chromosome AZFc region in Japanese infertile males with spermatogenic defects. Journal of Reproductive Immunology, 2002, 53(1 - 2): 37 - 44.

[6] Kuroda-Kawaguchi T, Skaletsky H, Brown L G, et al. The AZFc region of the Y chromosome features massive palindromes and uniform recurrent deletions in infertile men. Nature Genetics, 2001, 29(3): 279 - 286.

[7] Yen P. The fragility of fertility. Nature Genetics, 2001, 29(3): 243 - 244.

[8] Tyler-Smith C, McVean G. The comings and goings of a Y polymorphism. Nature Genetics, 2003, 35(3): 201 - 202.

[9] Repping S, Skaletsky H, Brown L, et al. Polymorphism for a 1.6 - Mb deletion of the human Y chromosome persists through balance between recurrent mutation and haploid selection. Nature Genetics, 2003, 35(3): 247 - 251.

[10] Slee R, Grimes B, Speed R M, et al. A human DAZ transgene confers partial rescue of the mouse Dazl null phenotype. Proceedings of the National Academy of Sciences of the United States of America, 1999, 96(14): 8040 - 8045.

[11] Kleiman S E, Yogev L, Hauser R, et al. Members of the CDY family have different expression patterns: CDY1 transcripts have the best correlation with complete spermatogenesis. Human Genetics, 2003, 113(6): 486 - 492.

[12] Machev N, Saut N, Longepied G, et al. Sequence family variant loss from the AZFc interval of the human Y chromosome, but not gene copy loss, is strongly associated with male infertility. Journal of Medical Genetics, 2004, 41(11): 814 - 825.

[13] de Llanos M, Ballesca J L, Gazquez C, et al. High frequency of gr/gr chromosome Y deletions in consecutive oligospermic ICSI candidates. Human Reproduction, 2005, 20(1): 216 - 220.

[14] Ferlin A, Tessari A, Ganz F, et al. Association of partial AZFc region deletions with spermatogenic impairment and male infertility. Journal of Medical Genetics, 2005, 42(3): 209 - 213.

[15] Hucklenbroich K, Gromoll J, Heinrich M, et al. Partial deletions in the AZFc region of the Y chromosome occur in men with impaired as well as normal spermatogenesis. Human Reproduction, 2005, 20(1): 191 - 197.

[16] Organization W H. WHO laboratory manual for the examination of human semen and sperm cervical mucus interaction. Oxford: Cambridge University Press, 1992.

[17] Underhill P A, Shen P, Lin A A, et al. Y chromosome sequence variation and the history of human populations. Nature Genetics, 2000, 26(3): 358 - 361.

[18] Jin L, Su B. Natives or immigrants: modern human origin in east Asia. Nature Reviews Geneticsl, 2000, (2): 126 - 133.

[19] Jobling M A, Tyler-Smith C. The human Y chromosome: an evolutionary marker comes of age.

Nature Reviews Genetics, 2003, 4(8): 598 - 612.

[20] Wilke K, Duman B, Horst J. Diagnosis of haploidy and triploidy based on measurement of gene copy number by real-time PCR. Human Mutation, 2000, 16(5): 431 - 436.

[21] Boehm D, Herold S, Kuechler A, et al. Rapid detection of subtelomeric deletion/duplication by novel real-time quantitative PCR using SYBR - green dye. Human Mutation, 2004, 23(4): 368 - 378.

[22] Fernandes S, Huellen K, Goncalves J, et al. High frequency of DAZ1/DAZ2 gene deletions in patients with severe oligozoospermia. Molecular Human Reproduction, 2002, 8(3): 286 - 298.

[23] Zhang G. Issues on male in fertility. Chinese J Androl, 2000, 14: 147 - 148.

[24] Krausz C, Quintana-Murci L, Rajpert-DeMeyts E, et al. Identification of a Y chromosome haplogroup associated with reduced sperm counts. Human Molecular Genetics, 2001, 10(1873 - 1877): pe40.

[25] Raymond M, Rousset F. An exact test of population differentiation. Evolution, 1995, 49: 1280 - 1283.

[26] Excoffier L, Laval G, Schneider S. Arlequin (version 3.0): an integrated software package for population genetics data analysis. Evolutionary Bioinformatics Online, 2005, 1: 47 - 50.

[27] Repping S, van Daalen S K, Korver C M, et al. A family of human Y chromosomes has dispersed throughout northern Eurasia despite a 1.8 - Mb deletion in the azoospermia factor c region. Genomics, 2004, 83(6): 1046 - 1052.

[28] Ferlin A, Moro E, Rossi A, et al. A novel approach for the analysis of DAZ gene copy number in severely idiopathic infertile men. Journal of Endocrinological Investigation, 2002, 25(1): RC1 - 3.

[29] Fernandes S, Paracchini S, Meyer L H, et al. A large AZFc deletion removes DAZ3/DAZ4 and nearby genes from men in Y haplogroup N. American Journal of Human Genetics, 2004, 74(1): 180 - 187.

[30] Skaletsky H, Kuroda-Kawaguchi T, Minx P J, et al. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature, 2003, 423(6942): 825 - 837.

7.2 *AZFc* 部分缺失可诱发全缺失导致男性不育

7.2.1 研究背景

Y 染色体的缺失是造成男性不育的一个主要原因[1,2]。3 个无精子症区域(*AZFa*、*AZFb* 和 *AZFc*)已经被定位为到 Yq11[3]，并且 *AZFc* 区域是最常发生缺失的区域[4,5]。由于 *AZFc* 包含一些区域性的长片段重复序列家族(扩增子)(图 7 - 5a)，它很容易在扩增子之间发生非等位同源重组，并引起多种类型的缺失[6,11,12]。

两种类型的 *AZFc* 区域局部缺失已确定。一种局部缺失是“gr/gr 缺失”，它是由于两个 g 或者两个 r 重复序列之间的同源重组导致的(图 7 - 5b)[8]。该缺失位于 *AZFc* 区域内部，长度 1.6 Mb，并在荷兰、西班牙、意大利和澳大利亚人群的研究中被认为是导致精子发生障碍的重要原因[8,13-16]。然而，这种相关性并没有在法国、德国、巴西、日本和斯里兰卡人群或者我们先前在中国人群的研究中得到验证[17-24]。需要特别指出的是，常

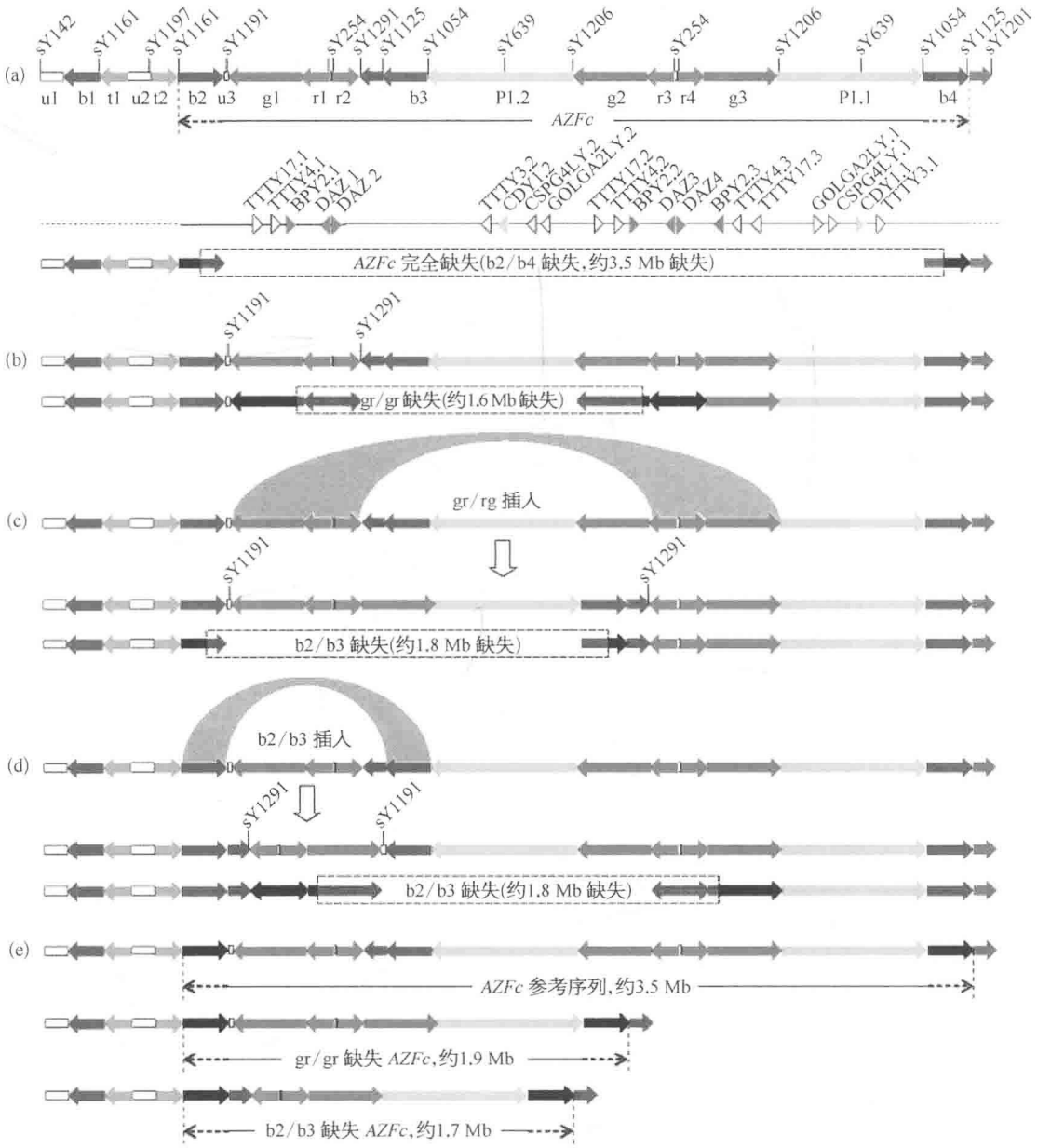


图 7-5 *AZFc* 区域的扩增子结构和 DNA 重排方式

(a) 基于 GenBank 参考序列的 *AZFc* 区域的扩增子的结构[6,7],序列标签位点,3 个编码蛋白的基因家族和 5 个不编码蛋白的基因家族,两个 b 扩增子之间的重组可以导致 *AZFc* 区域的完全缺失;(b) gr/gr 缺失使包括 SY1291 在内的局部 *AZFc* 区域缺失[8];(c) b2/b3 缺失使包括 SY1191 在内的局部 *AZFc* 区域缺失[9];(d) b2/b3 倒位使包括 SY1191 在内的出于 b2～b3 的 *AZFc* 区域发生颠倒,并通过后续的重组导致缺失[9,10];(e) *AZFc* 区域参考序列,gr/gr 缺失后的 *AZFc* 区域和 b2/b3 缺失后的 *AZFc* 区域之间的 b 扩增子距离的对比。

见的 Y 染色体单倍群 D2b 中的男性都携带 gr/gr 缺失[8]。

另外一个局部缺失是“b2/b3 缺失”(也叫 g1/g3 缺失或 u3 - gr/gr 缺失),由于 b2/b3 的倒位或 gr/gr 的倒位,它能够去除 *AZFc* 区域内 1.8 Mb 的片段(图 7 - 5c、d)[9,10,17]。b2/b3 缺失在 N 单倍群(一种在欧亚大陆北部广泛分布的单倍群)中是被固定的(100%)[25,26]。对中国人群的研究显示 b2/b3 的缺失和男性不育存在关联,但是在其他人群中没有检测出倾向性[9,10,16-18,24]。

相比较而言,*AZFc* 区域的完全缺失(去除所有 8 个在 *AZFc* 区域中的睾丸特异性表达基因)(图 7 - 5a)已经被确认是引起无精子症和少精子症的原因[6,27]。只存在极少数的例外。实际上,该缺失占男性不育相关 Y 染色体缺失的一半以上[4,5]。

本研究中,笔者在中国群体中检测了 296 个生精障碍的患者和 280 个健康的志愿者,研究了这些个体 *AZFc* 区域的部分缺失和完全缺失,并探索这些缺失在精子发生和男性不育中的作用。同时还用 19 个遗传标记进行了 Y 染色体单倍群的分型。研究 Y 染色体上缺失的分布能够帮助揭示不同遗传背景下 Y 单倍群在引发缺失过程中的作用。另外,采用定量分析来研究 *AZFc* 区域中的 *DAZ* 基因拷贝数,进一步鉴定 *AZFc* 区域的部分缺失。

7.2.2　材料和方法

1. 样本

本研究得到南京医科大学和上海仁济医院伦理审查委员会的批准,并获得了参与者的知情同意。通过南京医科大学附属医院及上海仁济医院不孕不育门诊,共采集 296 个非梗阻性无精子症和少精子症的样本,且样本间没有亲缘关系。笔者选取了 280 个捐精者作为对照组,这些人精液浓度、活力和形态正常(90 人),是一个或者多个孩子的父亲(190 人)。这些样本同样来自上述两家医院。另外,7 个 *AZFc* 区域完全缺失的家系采集自仁济医院。采用世界卫生组织的标准来对精液的浓度、活力和形态进行分析[28]。这些患者和对照组人员全部为来自华东的汉族人。

2. 缺失鉴定

以前的研究中已经详细描述了实验过程的细节[23],这里只做简要的描述。从外周血样本中提取基因组 DNA。通过序列特异性位点(STS)来检测 *AZFc* 区域的局部缺失:gr/gr 缺失(SY1161、SY1191、SY1201 和 SY1206 扩增结果为阳性,SY1291 扩增结果为阴性)和 b2/b3 缺失(SY1161、SY1201、SY1206 和 SY1291 扩增结果为阳性;SY1191 扩增结果为阴性)(图 7 - 6)[8,9]。通过 STS 手段分析的 *AZFc* 区域的完全缺失已经在以前的研究中进行了描述[6]。笔者采用的 STS 的位置展示在图 7 - 5a 中。用定量分析研究 *DAZ* 基因的拷贝数来进一步鉴定 *AZFc* 区域的局部缺失。运用 Machev 等[17]提出的方法来检测在 SY587 上的序列家族变异(SFV),也就是 *DAZ* 基因上的单核苷酸多态(SNP)Ⅱ。它可用来区分 *DAZ1/2* 和 *DAZ3/4*。实时定量 PCR 也用来测定 *DAZ* 基因的拷贝数。在 *AZFc* 区域外的 Y 染色体上,笔者选择了一处参考位点,用来校正测量结

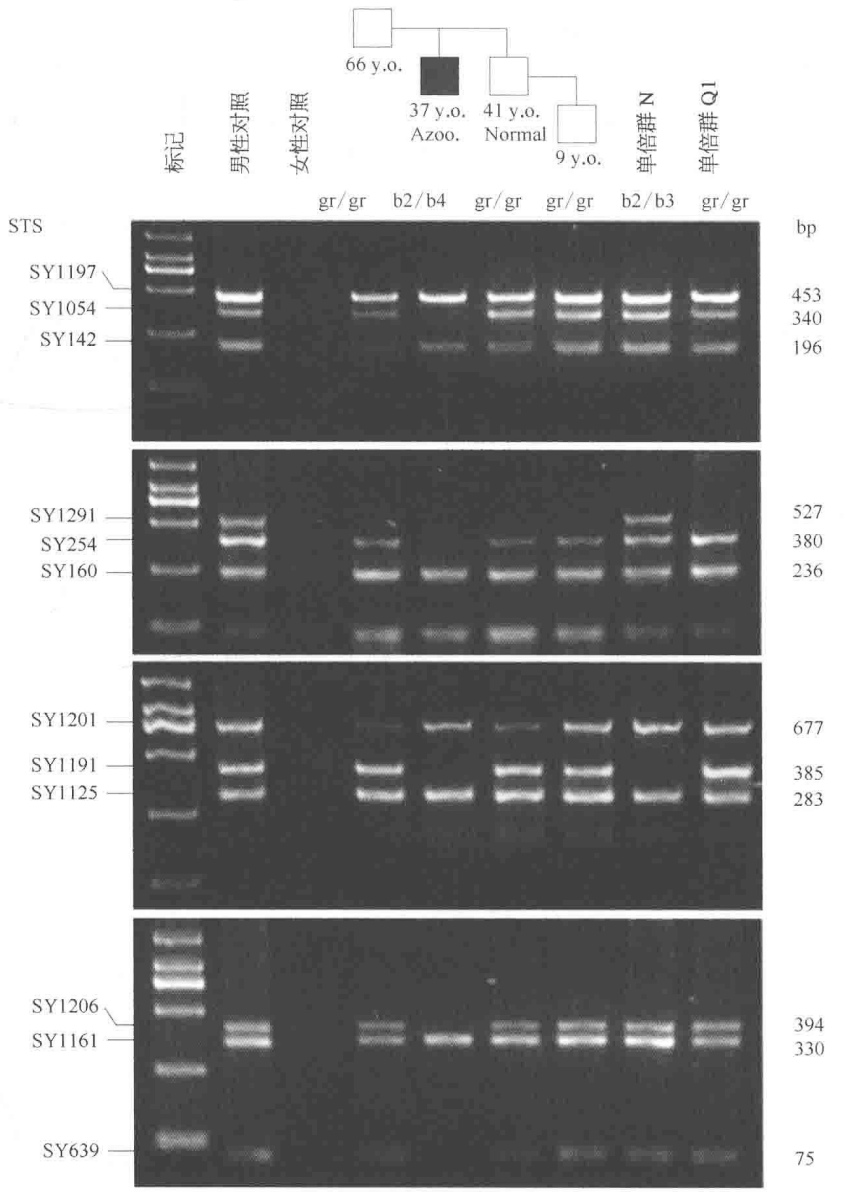


图 7-6 STS +/−的不同模式确定 *AZFc* 区域的局部缺失和完全缺失[6,8,9]

胶图的左侧给出了 STS 的名称，右侧给出了 STS 位点扩增产物的大小。黑色方块代表患有无精子症的患者。在方块下面还提供了年龄的信息。

果。被检测位点 DAZ-SNV1 和 DAZ-SNV5 位于 *DAZ* 基因的两端。

运用 16 个遗传标记来进行 Y 染色体单倍群分型：M9、M89、M95、M117、M119、M120、M122、M130、M134、M175、M176（SRY + 465）、M214、M231、M268、LLY22g 和 YAP（M1）[29-31]。LLY22g 的分型参照了 Y. Xue 和 C. Tyler-Smith 提供的

方法。这 16 个标记确定了基于 Y 染色体协作组(YCC)命名法则的 16 个 Y 染色体单倍群(图 7－7a)[31-33]。与早期报道的 M231 在系统发育上和 LLY22g 等价的结果不同[26]，笔者发现 M231 确定了一个新的单倍群 N＊。所以，根据 YCC 的命名法则，重新命名这个由 LLY22g 定义的单倍体群为 N1[33]。单倍体型 N1 能够用另外三个标记(N1a－M128，N1b－P43 和 N1c－Tat)来区分 4 个单倍群亚群[25,26,29,34]。

3. 统计学分析

笔者用 Arlequin 软件[35]通过 Rousset 精确检验来检测不育病例组和健康对照组之间的差异。采用 1 万次马尔可夫链和统计显著水平 $P<0.05$。利用 Fisher 精确检验检测在不同 Y 单倍群和不育病例组和健康对照组之间不同的缺失频率(统计上的显著设定在 $P<0.05$)(图 7－7b)。

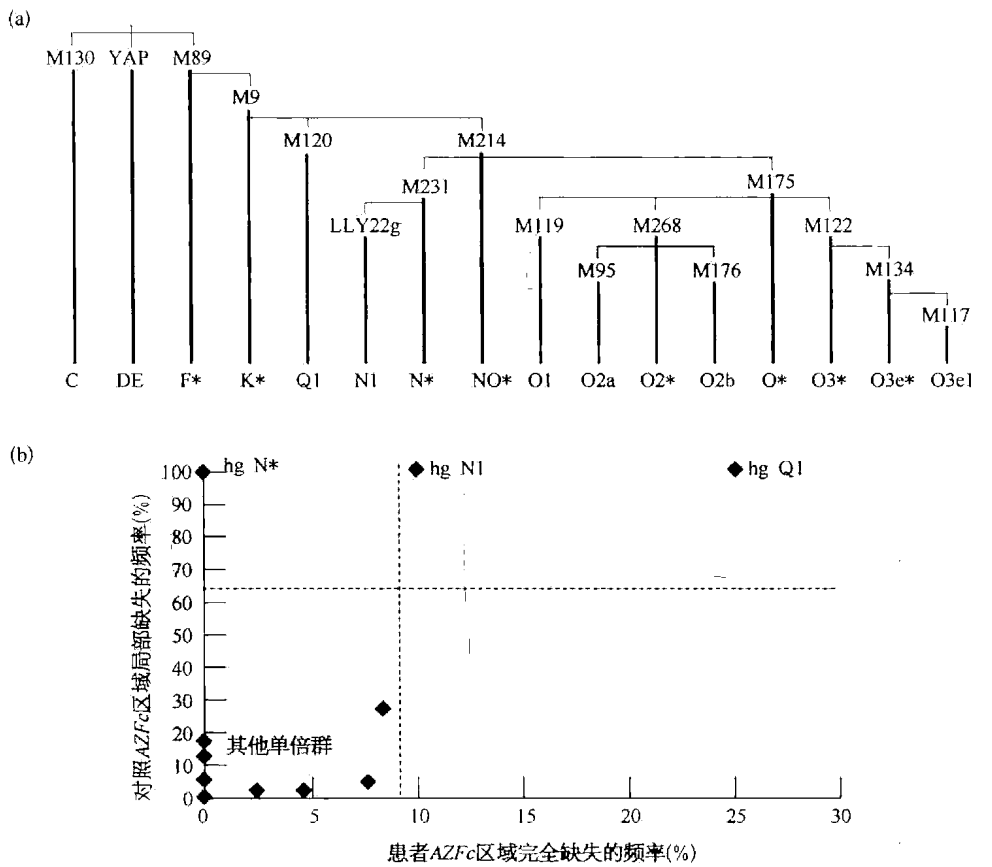


图 7－7　Y 染色体单倍群与 *AZFc* 缺失的关联分析

(a) Y 染色体的系统发生树。分支上标出了实验中用来测定的遗传标记。(b) Y 单倍群缺失频率的离散图。患者中 *AZFc* 区域完全缺失的频率(X 轴)和对照组健康人中 *AZFc* 区域局部缺失的频率(Y 轴)。因为在对照组健康人中没有发现单倍群 K＊和 NO＊，所以这两个单倍型类群没有在图中显示。

7.2.3 研究结果

1. 家系中 *AZFc* 区域完全缺失的基因分析

笔者研究了 7 个 *AZFc* 完全缺失的家系。在这些家系中，患者完全缺失 *AZFc* 区域，但是他们的父亲或者父系亲属并没有 *AZFc* 完全缺失。在一个家系中，研究者发现患者的父系亲属是 gr/gr 缺失。在其他 6 个家系中没有发现 gr/gr 和 b2/b3 的缺失。出于上述 gr/gr 缺失家系的患者有精子发生障碍。他的兄弟有正常的精子发生功能，而他的侄子(9 岁)和父亲(66 岁，认为是健康的)并没有参与精液分析(图 7-6)。Y 染色体单倍群分析表明，这个家系的 Y 染色体属于单倍群 O1。通过 SFV 对 *DAZ* 基因的分析表明，gr/gr 缺失造成了 *DAZ1/2* 基因的缺失。在父系亲属中进行 Y 染色体分型后，提示该家系的 *AZFc* 区域完全缺失源于一个携带 gr/gr 缺失的 Y 染色体。

2. *AZFc* 区域缺失在患者和对照组健康人中的分布

在 296 个男性不育患者和 280 个健康人中调查 *AZFc* 区域局部缺失的分布情况(表 7-4，表 7-5)。病例组中共发现有 24 个(8.1%)gr/gr 和 26 个(8.8%)b2/b3 缺失。而在健康对照组中发现 20 个(7.1%)gr/gr 缺失和 18 个(6.4%)b2/b3 缺失。在病例组和对照组中，这两种缺失频率没有显著的差异。

表 7-4 Y 染色体单倍群中 *AZFc* 全部/部分缺失以及样本的分布

分组		样本例数	Y 染色体单倍群															
			C	DE	F*	K*	Q1	N1	N*	NO*	O1	O2a	O2*	O2b	O*	O3*	O3e*	O3e1
全部	病例	296	25	3	2	1	8	20	3	1	52	5	12	3	1	82	35	43
	对照	280	24	1	2	0	10	13	1	0	49	8	11	3	1	70	39	48
仅 gr/gr 缺失	病例	24	1	1			6				2		3			4	5	2
	对照	20	1				10				2	1	3				2	1
仅 b2/b3 缺失	病例	26	3					18	3							2		
	对照	18	3					13	1							1		
仅 *AZFc* 全缺失	病例	14					2	2		1	4		1			2		2
	对照	0																

表 7-5 部分 *AZFc* 缺失中 DAZ 基因的 SFV 缺失

分组	样本例数	gr/gr 缺失		b2/b3 缺失	
		DAZ1/2 缺失	*DAZ3/4* 缺失	*DAZ1/2* 缺失	*DAZ3/4* 缺失
病例	50	17	7	1	25
对照	38	10	10	0	18

根据以前发表的对于 *DAZ* 基因的 SFV 或者 SNV 分析数据，可能只有当 *AZFc* 区域局部缺失涉及 *DAZ1/2* 基因时才与精子发生障碍相关联，而 *DAZ3/4* 基因的缺失可

能对精子发生的作用有限[9,10,14,15]。为了对 *AZFc* 区域的局部缺失进行深入分析，笔者对 SY587 位点的 SFV 进行分型，它可用来区分 *DAZ1/2* 和 *DAZ3/4*（表 7－5）。相对于 gr/gr 缺失，b2/b3 缺失更多地涉及 *DAZ3/4* 的缺失。这个结论与早期的报道和 b2/b3 缺失的一系列机制是相符的（图 7－5c、d）[10]。尽管在患者中有比正常人更多的 *DAZ1/2* 缺失，但是在这两种局部缺失亚型的频率上并无显著的差异。

笔者还观察到了 14 个（4.7%）*AZFc* 区域完全缺失的患者，然而在对照组中并没有发现这种情况。这与早期报道的 *AZFc* 区域完全缺失可以导致无精子症和少精子症是相符的[6,27]。

3. *具有高频率 AZFc 局部缺失的单倍群*

为了探索缺失和 Y 染色体的关系，首先测定了 16 个 Y 染色体遗传标记，并在患者和对照人群中区分得到 16 个单倍群。缺失的频率在 Y 染色体单倍群之间变化很大（表 7－4）。高频率的 *AZFc* 区域局部缺失在单倍群 Q1、N1 和 N＊中被发现。

单倍群 Q1 被发现携带有 gr/gr 缺失。所有 10 个 Q1 对照组健康人是 gr/gr 缺失。8 个 Q1 患者中的 6 个（75%）是 gr/gr 缺失，另外两个是 *AZFc* 区域完全缺失。因此，我们推测单倍群 Q1 的祖先是 gr/gr 缺失，而所发现的两例 *AZFc* 完全缺失是在 gr/gr 缺失的基础上发生进一步的缺失后导致的。

在单倍群 N1 中，所有被测试的人都携带 *AZFc* 区域完全缺失或者部分缺失。所有 13 个对照组的健康人和 20 个患者中的 18 个是 b2/b3 缺失，而其他两个是完全缺失。这个结果符合以前报道的在单倍群 N1 中 b2/b3 缺失是稳定存在的[9,10]。从而再次推测这些完全缺失是由于 N1 单倍群祖先的 b2/b3 缺失引起的。在本研究中，笔者还检测了另外三个遗传标记（M128、P43 和 Tat）来对单倍群进行更细致的区分。B2/b3 缺失在单倍群 N1 中是广泛分布的。这些表明他们起源是相同的并且是从 b2/b3 缺失的 N1 祖先中衍生而来的。

和 N1 相似的是所有的 N＊单倍群的三个患者和一个对照组健康人都含有 b2/b3 缺失。因为 N1 是从 N＊中系统发育而来的，b2/b3 缺失在这两个单倍群中是有共同的缺失祖先引起的。

相比较 *AZFc* 区域的缺失在 N＊、N1 和 Q1 中是近乎 100%稳定存在的，*AZFc* 区域局部缺失的频率在其他单倍群中是比较低的（在患者中平均为 8.7%，在对照组中是 5.5%）。

4. *AZFc 局部缺失中 DAZ 基因的拷贝数变异*

早期研究表明，在大多数 Y 染色体 *AZFc* 区域存在 4 个拷贝的 *DAZ* 基因[36]。*AZFc* 区域的缺失可以改变 *DAZ* 基因的拷贝数。gr/gr 缺失和 b2/b3 缺失通常将 *DAZ* 的拷贝数由 4 减少为 2[8-10]，而 *AZFc* 区域局部的重复可以增加 *DAZ* 的拷贝数使其数目大于 4 或者对部分缺失进行中 *DAZ* 基因拷贝数减少的补偿[36,37]。*DAZ* 基因拷贝数的数目检测是利用实时定量 PCR 分析 *DAZ* 基因的两个位点完成的（图 7－8）。

在 28 个随机选择的没有缺失个体中总共有 26 个人（13 个患者和 13 个对照组健康

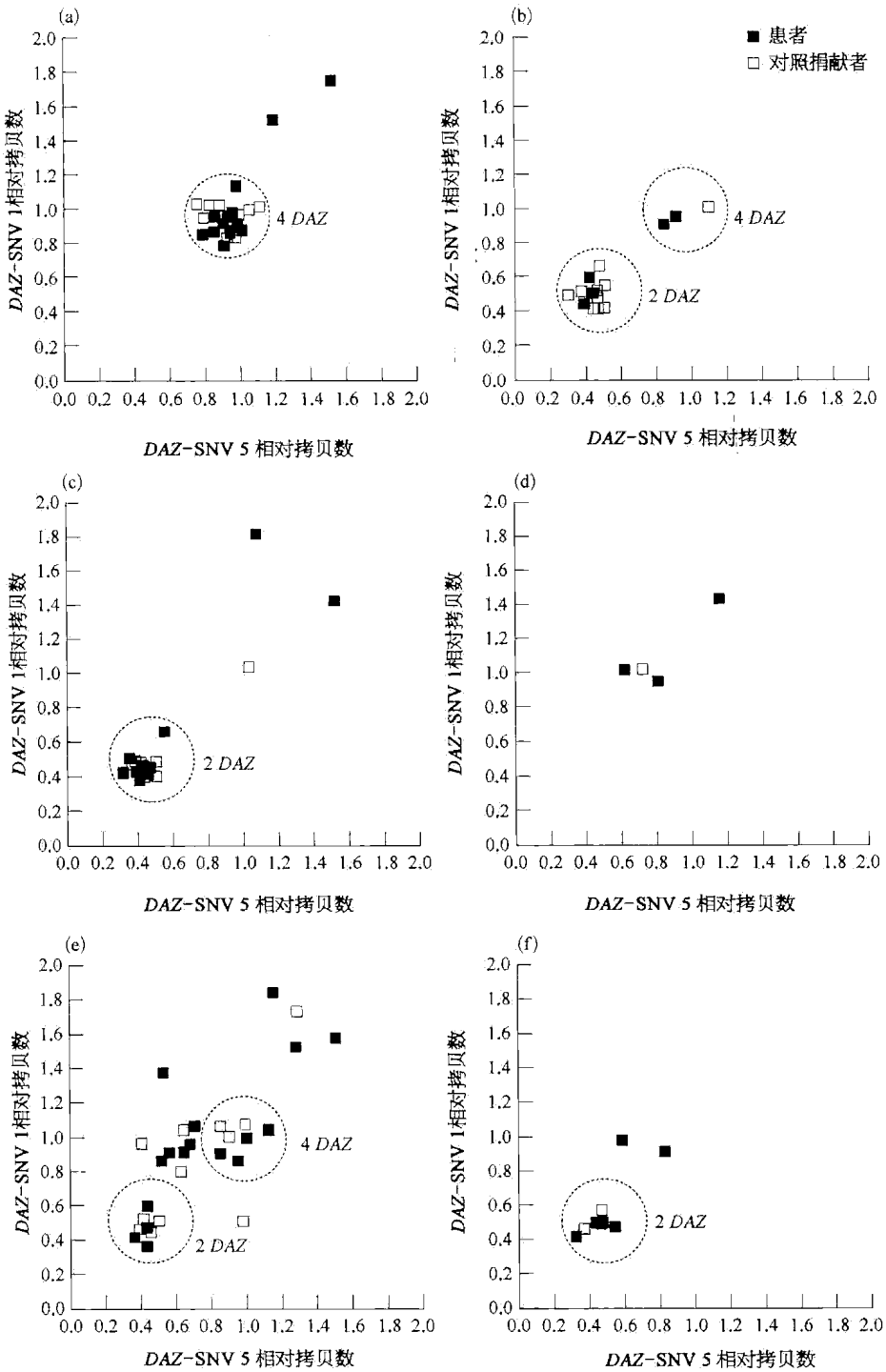


图 7-8 样本中 *DAZ* 拷贝数的离散图

(a) 没有缺失的样本;(b) 单倍群 Q1 的 gr/gr 缺失;(c) 非单倍群 Q1 的 gr/gr 缺失;(d) 单倍群 N* 的 b2/b3 缺失;(e) 单倍群 N1 的 b2/b3 缺失;(f) 非 N 单倍群的 b2/b3 缺失

黑方块代表患者,白方块代表对照组健康人。

人)拥有相同标准的 *DAZ* 拷贝数(图7-8a)。考虑到以前的研究,认为Y染色体上 *DAZ* 基因是4拷贝的[36],这些26个没有缺失和标准假定为 *DAZ* 基因存在4个拷贝。另外,在不携带缺失的两个患者中发现了 *DAZ* 拷贝数增加的情况,但是在对照组人群中并没有发现这种情况。在没有缺失的患者中并没有发现 *DAZ* 基因拷贝数减少的情况。

共有44个gr/gr缺失被检测到。在单倍型类群Q1中,16个gr/gr缺失的男人中13人(4个患者和9个对照健康人)含有两个 *DAZ* 拷贝,这与事实是相符的,通常4个拷贝的 *DAZ* 基因由于gr/gr缺失而减半(图7-8b)[8]。其他两个患者和一个对照健康人拥有4个拷贝的 *DAZ* 基因,这说明gr/gr缺失随着部分重复并且补偿由于gr/gr缺失所引起的 *DAZ* 基因拷贝数减少[37]。实验中有28例(18个患者和10个对照组健康人)gr/gr缺失不属于Q1单倍群,其中25人的 *DAZ* 基因为两个拷贝(图7-8c)。*DAZ* 基因的重复还在另外三个男性个体身上发现,其中一个对照组的健康人为 *DAZ* 基因4个拷贝,一个患者是6拷贝,另一个患者是复杂的 *DAZ* 基因亚结构(4拷贝的 *DAZ*-SNV 1和大于4拷贝的 *DAZ*-SNV 5)。*DAZ*-SNV 1和 *DAZ*-SNV 5在 *DAZ* 基因相反的两端,它们的数目不相等可能意味着 *DAZ* 基因的不完全重复。

相对于这个事实,*DAZ* 基因的拷贝数会因为gr/gr的缺失而减少到两个,局部 *DAZ* 基因的重复可能发生在很多b2/b3缺失中。在单倍群N*中,4个b2/b3缺失的男性个体(三个患者和一个正常的健康人)全部至少拥有4个拷贝。如 *DAZ*-SNV 1位点(图7-8d)。在单倍群N1中,31个缺失的患者中10个拥有两个 *DAZ* 基因的拷贝,而另外21个人拥有大于两个的拷贝,例如 *DAZ*-SNV 1或者 *DAZ*-SNV 5位点(图7-8e)。在非N单倍群中的9个b2/b3缺失个体中,三个患者和4个对照健康人拥有两个 *DAZ* 拷贝数,两个患者在 *DAZ*-SNV 1位点拥有大于两个的拷贝(图7-8f)。

最近在台湾汉族人群中的报道认为,在单倍群N-LLY22g中三分之一的b2/b3缺失发生了部分重复(对应于本研究中的N1)。这个频率是低于本研究所发现的三分之二的。然而,在两个研究中N1单倍群的部分重复的频率都是相对比较高的。这些差异可能是由于两个原因造成的。一个是台湾汉族人和中国东部汉族人的N1单倍群的结构差异。另一个是 *DAZ*-SNV 1和 *DAZ*-SNV 5位点的不平衡重复或许不能被前述的研究方法检测出来[37]。我们观察到的局部重复也支持了这个可能,在单倍型类群N1中31个b2/b3缺失中11个是 *DAZ*-SNV 1和 *DAZ*-SNV 5在一起,这是和以前的报道相符的[37]。

5. Y染色体上 *AZFc* 区域完全缺失的分布

在7个Y单倍群的患者中总共发现了14个 *AZFc* 区域的完全缺失,频率在2.4%~100%,其中频率最高的是单倍群NO*、Q1和N1。考虑到 *AZFc* 区域的部分缺失在单倍群N*、N1和Q1中是稳定存在的,这些结果表明在局部缺失富集的单倍群中呈现出高频率的完全缺失(图7-7b)。由于本研究中NO*单倍群只包含一个样本(*AZFc* 区域完全缺失),部分缺失的频率是没有办法确定的,所以在相关性研究中NO*单倍群没有被纳入。

N*、N1、Q1 和非 N*、N1、Q1 单倍群的 *AZFc* 区域完全缺失频率存在显著的差异（$P<0.04$，单侧检验；$OR=4.20$，95%置信区间为 1.21～14.5）（表 7-6）。相比较单倍群 Q1 和 N1，单倍群 N* 中所有的 b2/b3 缺失都会引起部分重复。当 *AZFc* 区域的部分缺失但没发生部分重复的个体也被包含在内，单倍群 N1、Q1 和非单倍群 N1、Q1 的完全缺失频率存在更加显著的差异（$P<0.03$，单侧检验；$OR=4.78$，95%置信区间为 1.37～16.7）。前面所观察到的结果说明，在 *AZFc* 区域的部分缺失相对于非缺失染色体更有可能引起完全缺失。

表 7-6 不同频率的部分 *AZFc* 缺失的 Y 染色体单倍群中全部 *AZFc* 缺失的病例分布

单倍群	全部 *AZFc* 缺失	非全部 *AZFc* 缺失	*OR*(95% 置信区间)	*P* 值①
N1 和 Q1（部分 *AZFc* 缺失高频②）	4 (14.3%)	24 (85.7%)	4.78 (1.37～16.7)	0.026
非 N1Q1（部分 *AZFc* 缺失低频）	9 (3.4%)	258 (96.6%)		
N*，N1，Q1（全体部分 *AZFc* 缺失③）	4 (12.9%)	27 (87.1%)	4.20 (1.21～14.5)	0.036
非 N* N1Q1（非全体部分 *AZFc* 缺失）	9 (3.4%)	255 (96.6%)		

注：① Fisher 精确检验，单尾。② 排除部分 *AZFc* 缺失伴随部分重复。③ 包括部分 *AZFc* 缺失伴随部分重复。

7.2.4 讨论

1. 一个新的 gr/gr 缺失单倍群：Q1

早期的报道认为 gr/gr 缺失在 D2b 单倍群中的稳定存在挑战了它在男性不育中的作用[8]。它认为单倍群 D2b 的祖先是 gr/gr 缺失的，并且这个缺失可代代相传。这与 gr/gr 缺失与精子发生障碍的关联相矛盾。

在本研究中，笔者发现了一个新的 gr/gr 缺失单倍群 Q1，所有被测试的 Q1 对照组健康人都是 gr/gr 缺失的，所有的男性不育患者是 gr/gr 缺失的或者是 *AZFc* 完全缺失的（估计是由 gr/gr 缺失导致的）。

作为 gr/gr 缺失特征的 SY1291 的缺乏也可以在单倍群 J 中由 O1084/O1085 的缺失引起[17]。笔者运用 *DAZ* 拷贝数检测来确定 gr/gr 缺失[23]。多数单倍群 Q1 的 gr/gr 缺失被发现含有两个拷贝[8]。这也是与 4 个 *DAZ* 基因的拷贝中两个由于 gr/gr 缺失而缺失的事实相符的。部分重复发生在另外三个 Q1 单倍群的 gr/gr 缺失个体中来补偿 *DAZ* 基因拷贝数的减少。

单倍群 Q1 是在汉藏人口中广泛存在的，频率为 1.8%～7.1%[38]。没有发现 Q1 单倍群中男性存在不育的倾向。因此，需要检测更多的 Q1 单倍群男性不育病例以揭示 gr/gr 缺失在精子发育中的作用。

2. *AZFc* 区域完全缺失的两个过程

鉴于 *AZFc* 区域的重复序列结构，*AZFc* 区域完全缺失可以源自一个正常的 Y 染色体或者一个携带有 *AZFc* 部分缺失的 Y 染色体。在本研究所涉 7 个家系中，其中一个家

系所携带的 *AZFc* 完全缺失被认为是在 gr/gr 缺失的基础上进一步突变而导致的。这种现象说明一些 *AZFc* 完全缺失是有两个突变途径的。在部分缺失 *AZFc* 区域内发生完全缺失是否比在没有缺失的 *AZFc* 区域内容易？答案还无从知晓。如果是更容易的话，在 *AZFc* 区域部分缺失的单倍群中相对于没有缺失的单倍群，将会有更高频率的 *AZFc* 完全缺失。

3. *AZFc* 部分缺失增加完全缺失风险

在 296 个少精子症和无精子症患者中共发现 14 个 *AZFc* 完全缺失的个体，这与估计的频率(5%～6%)相符[12]。研究发现，在不同的单倍群中缺失的分布是有差异的。在含有高频率 *AZFc* 部分缺失的 N1 和 Q1 单倍群中明显存在更多的 *AZFc* 完全缺失。这个现象说明 *AZFc* 部分缺失可导致更多的 *AZFc* 完全缺失。

尽管 *AZFc* 部分缺失导致完全缺失的风险机制还不清楚，对比 *AZFc* 部分缺失的结构和正常的 *AZFc* 区域后，我们得出一个产生这种倾向的可能原因。*AZFc* 完全缺失是由于姐妹染色单体在位于 *AZFc* 区域末端的 b 扩增子的同源重组造成的[6]。两个 b 扩增子在 *AZFc* 区域内的距离为 3.5 Mb，存在 gr/gr 缺失时 b 扩增子间距为 1.9 Mb，存在 b2/b3 缺失时 b 扩增子间距为 1.7 Mb(图 7－5e)。最近在人类基因组缺失多态性调查中发现，缺失的长度相对于缺失的频率服从 L 型分布，这说明小的缺失比大的缺失更容易发生[39]。因此，重组位点间距的减少可能增加了 *AZFc* 完全缺失的发生及男性不育的风险。

最近，来自意大利北部人群的报道称，单倍群 E 可能与 *AZFc* 全部缺失有关[40]。由于在 E3b 单倍群(单倍群 E 在欧洲的主要类型)中 gr/gr 缺失(40% 单倍群 E3b2 与 20% 单倍群 E3b3) 的频率据报道比其他欧洲单倍群(16.7%)更高[8,41]，所以部分 *AZFc* 缺失的效应可能是这一关联性的可能原因。

4. 过量的 *DAZ* 基因可能导致精子发育不全的易感性

在最近对台湾汉族的研究中[37]，发现 *AZFc* 区域的部分重复导致六拷贝 *DAZ* 是男性不育的风险因素。在本研究中，大多数被测试的男性个体不携带缺失并且含有 4 个 *DAZ* 拷贝，而没有缺失但含有大于 4 个 *DAZ* 拷贝数的男性个体只在患者中发现了一例。相比于没有缺失的男性个体，通常在部分缺失后含有两个拷贝的 *DAZ* 基因。*DAZ* 基因的拷贝数重复可以在一些部分缺失中发现。4 个 gr/gr 缺失(两个患者，两个对照组健康人)和 21 个 b2/b3 缺失(13 个患者和 8 个对照组正常人)含有 4 个拷贝的 *DAZ*－SNV 1 或 *DAZ*－SNV 5，但是在患者组和对照组没有明显的差异。

这些观察表明，在缺失后 *AZFc* 区域部分重复可使 *DAZ* 拷贝数恢复到 4 个拷贝，但这一过程对于精子发生的改善作用有限。然而，两个 gr/gr 缺失和 6 个 b2/b3 缺失的个体含有大于 4 个拷贝数的 *DAZ*－SNV 1 或者 *DAZ*－SNV 5。*DAZ* 基因的高拷贝数(大于 4)更多地在患者中出现，这一现象说明过量的 *DAZ* 基因可能导致精子发生障碍的易感性。

5. *AZFc* 区域的部分缺失对精子发育的不同影响

正如引言所描述的那样，*AZFc* 部分缺失在不同人群中的作用是有差异的。早期研

究中，gr/gr 缺失在东亚人群中不会增加精子发生障碍的风险[23]。在本研究中，笔者扩大了样本量，并且还检测了 b2/b3 缺失，同时还分析了群体差异性，但是并没有发现明显的群体差异[35]，并且在病例组和对照组之间没有发现明显 Y 单倍群分布差异。在本研究中，gr/gr 缺失和 b2/b3 缺失都与精子发生障碍没有关联。

7.2.5 结论

综上，通过研究 296 例精子发生障碍患者和 280 例健康人，笔者发现一个携带有 gr/gr 缺失的新的单倍群 Q1，没有发现 gr/gr 缺失和 b2/b3 缺失与精子发生障碍的关联性。所有这些结果揭示在群体之间 *AZFc* 区域的部分缺失具有表型差异。由于 Y 染色体在地区分布上的特异性[32]，不同 Y 单倍群的遗传背景可能是导致人群间表型差异的因素之一。因此，Y 单倍群可以部分地解释 *AZFc* 部分缺失在精子发生过程中作用的争议。更为重要的是发现了一例患者的 gr/gr 缺失可以进一步突变产生 *AZFc* 完全缺失，这说明 *AZFc* 完全缺失可以分两步完成，而不一定是一步突变产生的。

尽管 *AZFc* 部分缺失后发生的后续重复在病例组和对照组之间没有明显的频率差异，但是患者中存在过量的 *DAZ* 基因，而且无法排除其造成精子发生障碍的风险。*AZFc* 部分缺失与 *AZFc* 完全缺失的发生增加是相关联的，这为研究 *AZFc* 部分缺失在男性不育中的作用提供了一个新的视角。对更多人群体进行的相关研究将会促进我们对 Y 染色体 *AZFc* 完全缺失的突变机制的认识。

参考文献

[1] Tiepolo L, Zuffardi O. Localization of factors controlling spermatogenesis in the nonfluorescent portion of the human Y chromosome long arm. Hum Genet, 1976, 34: 119 - 124.

[2] Reijo R, Lee T Y, Salo P, et al. Diverse spermatogenic defects in humans caused by Y chromosome deletions encompassing a novel RNA-binding protein gene. Nat Genet, 1995, 10: 383 - 393.

[3] Vogt P H, Edelmann A, Kirsch S, et al. Human Y chromosome azoospermia factors (AZF) mapped to different subregions in Yq11. Hum Mol Genet, 1996, 5: 933 - 943.

[4] Foresta C, Moro E, Ferlin A. Y Chromosome microdeletions and alterations of spermatogenesis. Endocr Rev, 2011, 22: 226 - 239.

[5] Krausz C, Forti G, McElreavey K. The Y chromosome and male fertility and infertility. Int J Androl, 2003, 26: 70 - 75.

[6] Kuroda-Kawaguchi T, Skaletsky H, Brown L G, et al. The AZFc region of the Y chromosome features massive palindromes and uniform recurrent deletions in infertile men. Nat Genet, 2001, 9: 279 - 286.

[7] Skaletsky H, Kuroda-Kawaguchi T, Minx P J, et al. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature, 2003, 423: 825 - 837.

[8] Repping S, Skaletsky H, Brown L, et al. Polymorphism for a 1.6Mb deletion of the human Y

chromosome persists through balance between recurrent mutation and haploid selection. Nat Genet, 2003, 35: 247 - 251.

[9] Repping S, van Daalen S K, Korver C M, et al. A family of human Y chromosomes has dispersed throughout northern Eurasia despite a 1. 8Mb deletion in the azoospermia factor c region. Genomics, 2004, 83: 1046 - 1052.

[10] Fernandes S, Paracchini S, Meyer L H, et al. A large *AZFc* deletion removes DAZ3/DAZ4 and nearby genes from men in Y haplogroup N. Am J Hum Genet, 2004, 74: 180 - 187.

[11] Yen P. The fragility of fertility. Nat Genet, 2001, 29: 243 - 244.

[12] Noordam M J, Repping S, et al. The human Y chromosome: a masculine chromosome. Curr Opin Genet Dev, 2006, 16: 225 - 232.

[13] de Llanos M, Ballesca J L, Gazquez C, et al. High frequency of gr/gr chromosome Y deletions in consecutive oligospermic ICSI candidates. Hum Reprod, 2005, 20: 216 - 220.

[14] Ferlin A, Tessari A, Ganz F, et al. Association of partial *AZFc* region deletions with spermatogenic impairment and male infertility. J Med Genet, 2005, 42: 209 - 213.

[15] Giachini C, Guarducci E, Longepied G, et al. The gr/gr deletion(s): a new genetic test in male infertility? J Med Genet, 2005, 42: 497 - 502.

[16] Lynch M, Cram D S, Reilly A, et al. The Y chromosome gr/gr subdeletion is associated with male infertility. Mol Hum Reprod, 2005, 11: 507 - 512.

[17] Machev N, Saut N, Longepied G, et al. Sequence family variant loss from the *AZFc* interval of the human Y chromosome, but not gene copy loss, is strongly associated with male infertility. J Med Genet, 2004, 41: 814 - 825.

[18] Hucklenbroich K, Gromoll J, Heinrich M, et al. Partial deletions in the *AZFc* region of the Y chromosome occur in men with impaired as well as normal spermatogenesis. Hum Reprod, 2005, 20: 191 - 197.

[19] Carvalho C M, Zuccherato L W, Bastos-Rodrigues L, et al. No association found between gr/gr deletions and infertility in Brazilian males. Mol Hum Reprod, 2006, 12: 269 - 273.

[20] Carvalho C M, Zuccherato L W, Fujisawa M, et al. Study of *AZFc* partial deletion gr/gr in fertile and infertile Japanese males. J Hum Genet, 2006, 51: 794 - 799.

[21] Fernando L, Gromoll J, Weerasooriya T R, et al. Y chromosomal microdeletions and partial deletions of the azoospermia factor c (*AZFc*) region in normozoospermic, severe oligozoospermic and azoospermic men in Sri Lanka. Asian J Androl, 2006, 8: 39 - 44.

[22] Ravel C, Chantot-Bastaraud S, El Houate B, et al. gr/gr deletions within the azoospermia factor c region on the Y chromosome might not be associated with spermatogenic failure. Fertil Steril, 2006, 85: 229 - 231.

[23] Zhang F, Li Z, Wen B, et al. A frequent partial *AZFc* deletion does not render an increased risk of spermatogenic impairment in East Asians. Ann Hum Genet, 2006, 70: 304 - 313.

[24] Wu B, Lu N X, Xia Y K, et al. A frequent Y chromosome b2/b3 subdeletion shows strong association with male infertility in Han-Chinese population. Hum Reprod, 2007, 22: 1107 - 1113.

[25] Zerjal T, Dashnyam B, Pandya A, et al. Genetic relationships of Asians and Northern

Europeans, revealed by Y chromosomal DNA analysis. Am J Hum Genet, 1997, 60: 1174 - 1183.

[26] Rootsi S, Zhivotovsky L A, Baldovic M, et al. A counter-clockwise northern route of the Y chromosome haplogroup N from Southeast Asia towards Europe. Eur J Hum Genet, 2007, 15: 204 - 211.

[27] Tyler-Smith C, McVean G. The comings and goings of a Y polymorphism. Nat Genet, 2003, 35: 201 - 202.

[28] World Health Organization. WHO laboratory manual for the examination of human semen and sperm-cervical mucus interaction. Cambridge, United Kingdom: Cambridge University Press, 1992.

[29] Underhill P A, Shen P, Lin A A, et al. Y chromosome sequence variation and the history of human populations. Nat Genet, 2000, 26: 358 - 361.

[30] Shinka T, Tomita K, Toda T, et al. Genetic variations on the Y chromosome in the Japanese population and implications for modern human Y chromosome lineage. J Hum Genet, 1999, 44: 240 - 245.

[31] Sengupta S, Zhivotovsky L A, King R, et al. Polarity and temporality of high-resolution Y chromosome distributions in India identify both indigenous and exogenous expansions and reveal minor genetic influence of Central Asian pastoralists. Am J Hum Genet, 2006, 78: 202 - 221.

[32] Jobling M A, Tyler-Smith C. The human Y chromosome: an evolutionary marker comes of age. Nat Rev Genet, 2003, 4: 598 - 612.

[33] Y Chromosome Consortium. A nomenclature system for the tree of human Y chromosomal binary haplogroups. Genome Res, 2000, 12: 339 - 348.

[34] Karafet T M, Osipova L P, Gubina M A, et al. High levels of Y chromosome differentiation among native Siberian populations and the genetic signature of a boreal hunter-gatherer way of life. Hum Biol, 2002, 74: 761 - 789.

[35] Raymond M, Rousset F. An exact test of population differentiation. Evolution, 1995, 49: 1280 - 1283.

[36] Repping S, van Daalen S K, Brown L G, et al. High mutation rates have driven extensive structural polymorphism among human Y chromosomes. Nat Genet, 2006, 38: 463 - 467.

[37] Lin Y W, Hsu L C, Kuo P L, et al. Partial duplication at *AZFc* on the Y chromosome is a risk factor for impaired spermatogenesis in Han Chinese in Taiwan. Hum Mutat, 2007, 28: 486 - 494.

[38] Su B, Xiao C, Deka R, et al. Y chromosome haplotypes reveal prehistorical migrations to the Himalayas. Hum Genet, 2000, 107: 582 - 590.

[39] Conrad D F, Andrews T D, Carter N P, et al. A high-resolution survey of deletion polymorphism in the human genome. Nat Genet, 2006, 38: 75 - 81.

[40] Arredi B, Ferlin A, Speltra E, et al. Y chromosome haplogroups and susceptibility to *AZFc* microdeletion in an Italian population. J Med Genet, 2007, 44: 205 - 208.

[41] Semino O, Magri C, Benuzzi G, et al. Origin, diffusion and differentiation of Y chromosome

haplogroups E and J: inferences on the neolithization of Europe and later migratory events in the Mediterranean area. Am J Hum Genet, 2004, 74: 1023 - 1034.

7.3　世界语音多样性分布格局与人类扩张

全世界有 7 000 多种语言。语言起源于哪里？又是怎样演化成那么多不同的样子？最极端形态的语言是怎么样的？这些语言人类学研究了上百年的问题在 2012 年有了初步的答案。国际顶级学术期刊 *Science* 上连续发表了数篇论文，全面探讨了世界各种语言的语音多样性分布与语言起源问题[1-7]。笔者也在 *Science* 上参与了分析讨论[3]，分析了全世界各个语系中最具代表性的 500 多种语言的语音系统，得出了语音分布的世界规律。在全世界任意选择一个地点，计算从这个地点向四周的语音复杂度的降低速率，发现降低速率最高的是里海南岸，而不是非洲。所以，除了非洲南部的科伊桑语系，全世界大部分语言都源于大约 4 万年前中东的早期现代人大扩张。这一研究得到了全世界语言人类学界的极大关注。这一研究结果与 Y 染色体的谱系分析是完全吻合的。Y 染色体中，科伊桑语系的布须曼人所拥有的 A 单倍群没有走出过非洲，所以与中东的扩张无关。俾格米人的 B 单倍群也没有走出非洲，但是他们的语言已经被西非黑人的班图语取代。其余所有的 Y 染色体单倍群应该都是来自中东的。

这一项研究的最关键工作是建立语种样本数据库，并对数据库进行挖掘和分析。在收集数据的时候，面临着各个语种的调查者的调查标准偏差带来的数据不一致性。这是包括语言在内的各种数据库都会出现的数据误差问题。在数据库分析中，误差是永远无法避免的，关键的问题是把误差控制在不影响分析结果的范围内。本次讨论的发起者阿特钦森所依据的数据库，就出现了把误差人为扩大的问题，而使得结果完全偏差了。比如，他把元音的数量分为三档：少(2～4 个)、中等(5～6 个)、多(大于 7 个)。而实际上世界各语言的元音数量从 2～20 个不等，他的分类把 10 个元音以上的数据差异都抹平了，引入的人为误差造成了结果的失真(图 7 - 9)。

笔者的做法是把全世界语言的语音系统整理成新的数据库，保留各个语种使用者所认知的所有真实语音，利用这一没有被简化归并的数据库分析世界语言的语音分布规律。这一数据库中不存在数据库整合引入的人为误差，但是仍然存在调查标准误差。这种误差无法在建库之前估算，但是通过建库以后对数据库的分析，可以做出估算。数据库中语音多样性的分布呈现规律性，语音多样性的变化都是渐变的，由不同调查者整理的亲缘关系相近各语种的语音资料体现出明显的相似性，同一语系内的相似程度也远远大于不同语系间的相似程度。诸多证据显示，调查者误差并没有影响数据库的有效性。

在这一基础上，对数据库进行深入分析，发现各种语音的分布规律比较明显，特殊或者极端的语音丰富度的分布集中在个别语系中。所以通过这一研究，也发现了几种最极端的语音系统。语音大致分为声调、辅音和元音三部分。其中声调的分布最简单，很多

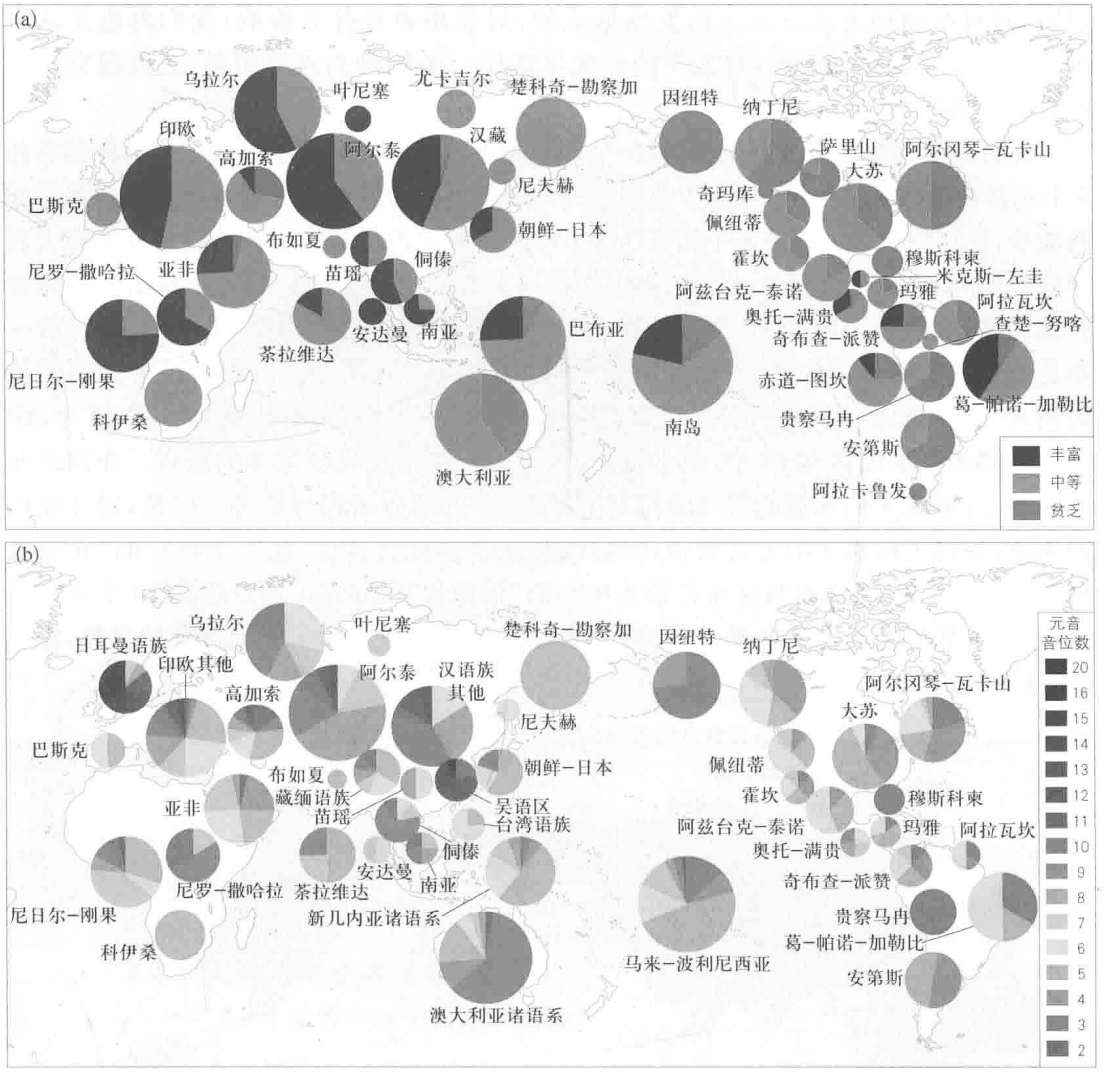


图 7-9 世界元音分布规律在分档观察(a)和不分档观察(b)时显示出明显的差异

语言都不区分声调,有声调的语言集中分布在西非、东亚和北美西北,其中东亚语言的声调最复杂。西非和北美的声调多是以调值的高低来区分的,而东亚的语言既有调值高低区分,又有调形区分。东亚的 4 个语系都发生了声调,包括侗傣、苗瑶、汉藏和叶尼塞。声调最多的语言是中国贵州锦屏县和剑河县之间的高坝南侗语,有 15 个声调,说话就像唱歌一样。辅音种类很多,有数百种,分布也比较复杂。辅音较多的地方有多处,包括南非、高加索山区和中国川西等,涉及科伊桑语系(影响相邻的尼日尔刚果语系)、西北高加索语系和汉藏语系(羌语支)等。世界上辅音最多的语言是西北高加索地区的优必语(Ubykh),达到 180 个辅音[8]。但是在 1864 年 3 月,优必人被沙皇亚历山大二世赶出家园,逃到了土耳其,从此开始流散。优必语也渐渐式微,到 1992 年 10 月,最后一个使用者逝世,这种奇异的语种也就灭绝了。这使得全世界的相关学者唏嘘不

已，这种复杂的辅音系统是如何发展形成的，对使用者有什么影响，我们再也无从研究。所以一种语言的灭绝，对人类的损失可能比一个物种的灭绝更甚，更何况它是一种极端语言。

元音是语音中最重要的部分，是一个音节的主体部分。世界上大部分的语言都有约 5 个元音音位(a、e、i、o、u)，只有少数语言的元音音位数量发生了大的变化。有些语言元音减少，比如阿拉伯语只有 3 个元音(a、i、u)，而高加索西北部的阿布哈兹语以及澳大利亚中北部的 Aranda 语和 Anindilyakwa 语都只有两个元音(a、ə)。另外一些语言的元音开始增多，少数甚至超过 12 个。世界上一共只有两大类语言达到 12 个元音音位，一类是北欧的日耳曼语族，另一类是东亚的吴语方言区，这两个地方成为世界元音系统的两大“高原极地”(图 7－10)。还有两个语种元音也比较多，喀山鞑靼语有 13 个，印度果阿邦的孔卡尼语有 14 个，属于特例，但它们都还不是元音最多的语言。在两个元音高原上，原来人们知道的最高峰是有 16 个元音的瑞典语和丹麦语。但是，通过我们的调查，发现上海郊区的方言普遍达到或超过 16 个元音音位，是一个真正的“元音极地”(图 7－11)。而以奉贤区金汇镇为中心的“傷傣话”(Dônđăc)甚至达到 20 个元音音位，成为世界元音无以企及的最高峰。“傷傣”[dɔ̃˧˩˧ đæ˧˩]属于当地居民的自称，意思是“这个地区”。

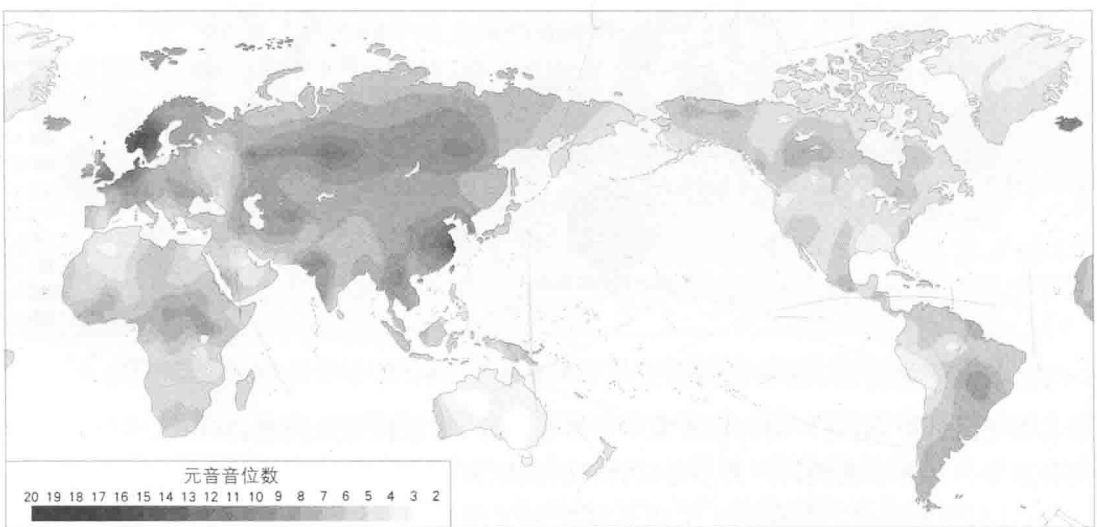


图 7－10　世界语言元音多样性分布格局

从元音的类型分析，我们发现世界各个语种的元音系统类型更有规律。我们可以根据前或后、圆唇或非圆唇、央元音、舌尖元音，把元音分为 6 个部类。大多数语言的元音都只有三部，即“非圆唇的前元音、圆唇的后元音、央元音”。少数语言向着复杂化或简化的方向进化。我们可以把所有的元音系统类型分为以下 6 种形式。

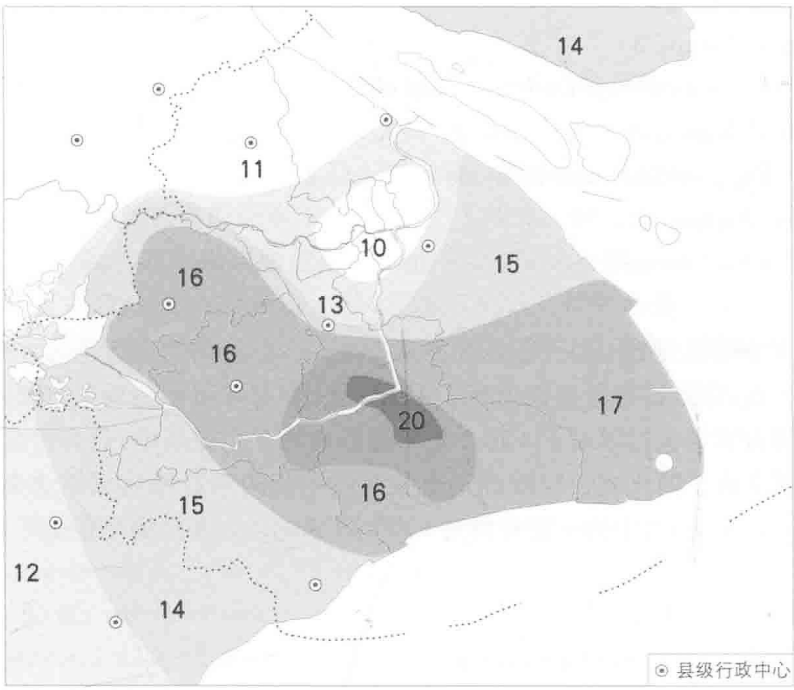


图 7-11 上海地区方言元音多样性的大致分布格局

最初形式：3 部
非圆前＋低央＋圆后
简化形式：1 部
高＋低
复杂 1 式：(日耳曼语)4 部
非圆前＋央＋圆后＋圆前
复杂 2 式：4 部
非圆前＋央＋圆后＋非圆后
更复杂式：5 部
非圆前＋央＋圆后＋圆前＋非圆后
最复杂式：(汉语)6 部
非圆前＋央＋圆后＋圆前＋非圆后＋舌尖

再看世界上各个语系(语群)的元音系统类型，我们可以发现同一语系的类型一般相似。非洲南部的 3 个语系都是最普通的 3 部式，没有进化出其他形式。非洲北部和西亚的亚非语系也是 3 部式的。具体如下。

Khoisan：3

ie+a+ou

Niger-Congo：3

ie+a+ou→ieɛ+a+ɔou→iɪeɛ+a+ɔoʊu→iɪeɛ+aʌə+ɔoʊu

Nilo-Saharan：3

ieɛ+a+ɔou→iɪeɛ+a+ɔoʊu→iɪeɛ+aɑ+ɔoʊu

Afro-Asiatic：3

i+a+u←iə+a+u←ie+a+u←ie+a+ou→ieɛ+a+ɔou→ie+aɐɨ+ou

欧洲的3个语系开始出现元音系统类型的演变。但是绝大多数还是3部式的，有些发展出4部1式，印欧语中极少数出现了4部2式（葡萄牙语和高地盖尔语）。印欧语中还有中国的塔吉克语也发展成了4部2式。高加索的3个语系中，只有西北高加索语系的元音系统演变成了简化式。与欧洲语言关系较密切的茶拉维达语系也有3部式和4部1式两种元音系统，其中的y很可能是ɨ演变过来的。布如夏语系也是3部式的。

Basque：3－4

ie+a+ou→ie+a+ou+ø

Caucasians：1－4

aə←e+aə←ie+a+ou→ieæ+ɑə+ou→ieɛæ+aə+ɔu+œy

Indo-European：3－4

i+a+u←ie+a+ou→ieɛ+a+ɔou→ieɛ+aə+ɔou→ie+aəi+ou+øy→

ieɛ+aə+ɔou+ʏɯ→iɪeɛæ+aɑə+ɔoʊu+y→ieɛæa+ɑʌə+ɒɔou+Œœøy→

iɪeɛa+ɑɵu+ɔoʊu+œøʏy

Dravidian：3－4

ie+a+ou→ieæ+ɑɵi+ou+y

Burushaski：3

ie+a+ou

两大洋土著的各种语言的元音系统都比较简单，基本上都是3部式的。澳大利亚的两种语言退化成为简化式。只有恩果语（Ngkoth）演变出了4部1式，其中的ø很可能是ə演变过来的。

Andamanese：3

ie+aə+ou→ieæ+aə+ou

Australians：1/3/4

ia←i+a+u←ie+a+u←ie+a+ou→ieæ+a+ou+ø

Papuans：3

ie+a+u←ie+a+ou→ie+aə+oʊu→ieɛ+aə+ɔou

东亚的诸语系进化出了最复杂多样的元音系统。但是最早的南亚语系和苗瑶语系并没有在元音系统上有所发展，都是3部式的。南岛语系的大部分语言都是3部式的，只有密克罗尼西亚的沃利延语发展出了4部1式，而新喀里多尼亚的雅爱语是一个特例，处于非常远缘的几种语言的影响下，发展出了5部式。侗台语都是4部2式的，除了侗水语支是3部式的。藏缅语的演化形式非常丰富，喜马拉雅山南的语言多为3部式的，藏语是4部1式的，羌语支发展出4部2式或5部式。汉语的各种方言元音系统的演化形式最为丰富，从3部式到6部式的各种类型都出现了，也是唯一发展出最复杂式的元音系统的类群。

Austro-Asiatic：3

高棉语：ieɛ+aʁəi+ɔou，越南语：ieɛ+aəi+ɔou

Hmong-Mien：3

畲语：iɛ+aɑə+ɔou

Austronesian：3－5

泰雅语：ie+aə+ou，他加禄语：iɛ+a+ou，萨摩亚语：iɪeɛ+aə+ɔoʊu，沃利延语：ie+aɑ+ou+øy，雅爱语(特例)：ieæ+a+ɔou+œøy+ʏ

Daic：4

黑泰语：ieɛ+aəi+ɔou+ɯ

Tibeto-Burman：3－5

库宋达语：ie+aə+ou，藏语：ieɛ+a+ou+øy，嘉戎语：ie+a+ou+ʏɯ，普米语 ɿ+ieɛ+aɑə+ou+ʉy

Chinese：3－6

闽南语 ie+aə+ɔu→广东话 iɛ+ɑʁ+ɔu+œy→福州闽东语 ieɛ+aɑ+ɔou+œøy→梅县客家语 ɿ+iɪɛ+aə+ɔu→长沙湘语 ɿ+ieɜæ+aɑ+o+y→抚州赣语 ɿ+ieɜ+aə+ou+y→平阳吴语 ɿ+ieɛ+a+ɔou+œy→平遥普语 ɻɻʯ+iɛæ+ɑə+ɔu+y→高淳吴语 ɿ+iɪɛ+aə+ɔou+yʏ→普通话 ɻɻ+iɛ+a+ou+yʏ→兰溪吴语 ɿ+ieæ+aɑ+ɔou+y+ʏɯ→平湖吴语 ɻʯ+ieɛ+aə+ɔou+øy+ɯ→宁波吴语 ɻʯ+iɪɛ+aə+ɔou+œøy+ɤ→奉贤傍傣话 ɿ+iɪɛɜæ+aɑə+ɒɔoʊu+œøʏy+ʌʏ

北亚的4个语系中，叶尼塞语系和古亚语系都是3部式的。阿尔泰语系中满语和蒙语是3部式的，而朝鲜语族、突厥语族大多是5部式的。乌拉尔语系中同样出现了从3部式到5部式的各种形态，但是没有出现舌尖元音。乌拉尔语系和阿尔泰语系元音体系普遍出现了更复杂式，与汉藏语系相似，这与使用这两种语言的人群起源于汉藏人群可能

有关系。

Yeniseian：3

羯语：iɛ+aəɨ+ɔu

Palaeosiberian：3

Nivkh：ɪeæ+ou+ʏ，Chukchi：ie+a+ou

Uralic：3－5

科米语：ie+aə+ou，尤卡吉尔语：ie+a+ou+ø，涅涅兹语：ieæ+aə+ou+ø，芬兰语：ieæ+ɑ+ou+øy，马瑞语：ieæ+ɑə+ou+øy+ɯ

Altaic：3/5

满语：ie+a+oʊu，蒙古语：ie+a+ɔoʊu，朝鲜语：ieɛ+a+ou+ø+ʌɯ，图瓦语：ie+a+ou+øy+ɯ，喀山鞑靼语：ieɛæ+ɑə+ou+œøy+ʌɯ

美洲诸语系的元音系统比欧亚大陆简单得多。爱斯基摩语系和纳丁尼语系都只有3部式。印第安诸语系的元音系统大多是3部式的，少数语种演化出了4部式，例如Arawakan语系的几种方言。Hopi语据记录是5部式的，但是元音比较少，其中的ɯ很可能是ə，所以这种5部式是很可疑的。

Eskimo-Aleut：3

i+a+u

Na-Dene：3

Navajo：iɛ+a+o，Haan：ie+aɑə+ou

Amerinds：

3部式 Cheyenne：e+a+o，Creek：i+a+o，Quechua：i+a+u，Mayan：iɛ+aə+ou，Cherokee：ie+aə+ou，Lakota：ie+a+ou，Pomo：ie+a+ou，Maidu：ie+aəɨ+ou

4部2式 St"at"imcets：eɛæ+aə+ɔo+ʌ，Wayuu：iɛ+a+ɔu+ɯ，Carib：ie+a+ou+ɯ

4部1式 Otomi：ieɛ+ɑʉ+ɔou+ø

5部式（特例）Hopi：iɛ+a+o+ø+ɯ

从全世界的各个语系的元音系统形态复杂性分布看来，元音系统的形态演化是相对保守的，发展出一种新的更复杂的形态需要很苛刻的条件，人口少语种少的小语系一般不会发展。非洲和大洋洲的形态基本没有什么演化，欧亚大陆西部的形态演变也比较简单，美洲也并不复杂，只有东亚和北亚的语系演化出比较复杂的形态，而最复杂的6部式形态只有汉语中出现。4部式的元音系统不可能包括超过18个元音，所以东亚和北亚以

外的语言都不可能发展出超过吴语的元音数量。北亚的两个语系虽然发展出5部式，但是元音数量都非常少。

对于吴语“傷傣话”的研究，实际上早在1998年就开始了。为了调查上海各地区人群的来源，笔者选择了13个乡镇（包括金汇镇），分析了当地居民的语言、文化、体质、遗传和家谱方志等[9]。在体质和遗传上，金汇居民有着明显的古代百越民族特征，而在语言上，金汇方言也保留着百越后裔语言侗傣语系的口音和部分词汇[10]。金汇的“傷傣话”就此进入了笔者的研究视野。金汇方言最早是钱乃荣教授调查记录的[11]，当时已经记录了至少18个元音。为了分析“傷傣话”的元音音位是否是世界上最多的，笔者在世界各地的图书馆中收集语言资料，总结语音的分布规律。最终了解到，世界上大部分地区都是元音“平原”，甚至“低谷”，只有北欧日耳曼语区和中国吴语区是元音“高原”。元音的音位数量在语言系统中是渐变的，不可能突然升高。所以元音“最高峰”只可能出现在“高原”上，不可能出现在“平原”或者“低谷”。因此只可能在日耳曼语族和吴语区寻找最高峰。所幸的是，这两大类语言都在经济最发达地区，都已经得到了最详细的研究，所以我们能够确定上海地区的方言，特别是“傷傣话”真的是世界元音“最高峰”。

上海方言的元音音位增多，实际上是入声的特殊性质造成的。汉语方言的入声，大多是根据音节的延时短促来识别的，有的是根据塞音韵尾来识别的，但是大多数的北部吴语的入声音节既不必须延时短促，也不一定有塞音韵尾，而是根据元音音值的低央化来判识的。事实上，日耳曼语族元音增多的原因与吴语的情况相似，长短元音实际上并不一定有延时上的差异，而是在元音音值上有差别。在上海郊区的各个方言中，阴声韵和入声韵通过不同的元音音值来区分。而上海的松江片方言的入声韵本来就种类繁多，19种入声是很常见的，所以在这一区域内元音音位就普遍增多。而金汇镇是这个区域内5～7条同言线的交叉点[12]，所以字音有更多的分化，使得元音音位进一步增多。有20个元音的地区可能并不局限于金汇镇，奉贤很多地区以及周边区县可能都有分布，是否都有音位对立，这有待于进一步调查。

作为世界上元音音位最多的语言，“傷傣话”是全世界人类的宝贵文化财富，是值得人类学、语言学和社会学等领域深入研究的对象，对于人类的发展可能有着非常的意义。这种复杂的语音系统，也可能影响使用者的生理和心理状态。我们现在不知道，这一语种还会对其他学科产生什么样的启发和影响，但是我们必须把这一人类文化遗产保护下来，传承下去。因此，笔者专门编写了教材《傷傣话》，从2012年9月起，在上海市奉贤区的小学中开始了方言课教学[13]。教材中使用的音系与传统的汉语方言调查规范不同，而是完全照顾到方言使用者对该方言语音的认知和辨识（图7－12），得到当地师生和学者的广泛认可。这是一次有益的尝试，希望在未来看到其良好影响。

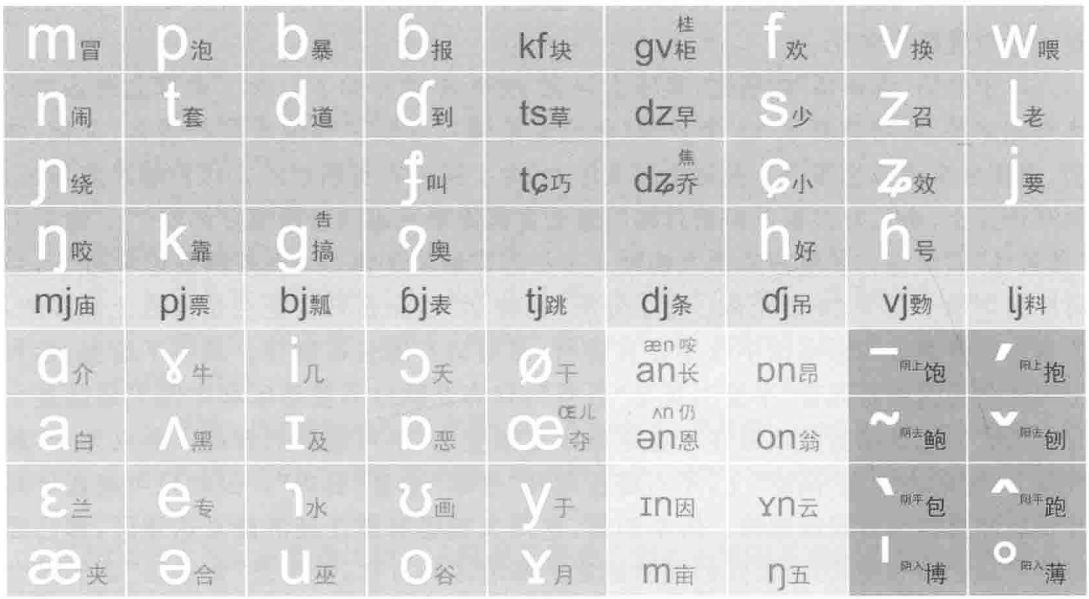


图 7－12　奉贤傍傣话实用音系表

注：表中送气符号省略，如 p 的实际读音是[p‘]，ts 是[ts‘]。而 dz 是[ts]，dz 是[tɕ]或[dz]，以不同声调区分。入声声调保留以标示特殊语气下的短促发音。

参考文献

[1] Quentin D Atkinson. Phonemic diversity supports a serial founder effect model of language expansion from Africa. Science, 2011, 332: 346－349.

[2] Cysouw M, Dediu D, Moran S, et al. Comment on ‘phonemic diversity supports a serial founder effect model of language expansion from Africa’. Science, 2012, 335: 657b.

[3] Wang C C, Ding Q L, Tao H, et al. Comment on ‘phonemic diversity supports a serial founder effect model of language expansion from Africa’. Science, 2012, 335: 657c.

[4] Van T R, Pereltsvaig A. Comment on ‘phonemic diversity supports a serial founder effect model of language expansion from Africa’. Science, 2012, 335: 657d.

[5] Quentin D Atkinson. Response to comments on ‘phonemic diversity supports a serial founder effect model of language expansion from Africa’. Science, 2012, 335: 657e.

[6] Jaeger T F, Pontillo D, Graff P. Comment on ‘phonemic diversity supports a serial founder effect model of language expansion from Africa’. Science, 2012, 335: 1042a.

[7] Quentin D Atkinson. Response to Comment on ‘phonemic diversity supports a serial founder effect model of language expansion from Africa’. Science, 2012, 335: 1042b.

[8] Hans Vogt. Dictionnaire de la langue oubykh. Oslo: Universitetsforlaget, 1963.

[9] 林凌，李辉，张海国，等. 上海郊区人群的体质特征和遗传关系. 人类学学报，2002，21(4)：293－306.

[10] 郑张尚芳. 浙南和上海方言中的紧喉浊塞音声母初探//中国语言文学研究所吴语研究室编. 吴语论丛. 上海：上海教育出版社，1988：232－237.

[11] 钱乃荣. 奉贤东西乡的语言同言线//李振麟编. 语言研究集刊 I. 上海：复旦大学出版社，1987：297－308.

[12] 钱乃荣. 奉贤语音的内部差异. 中国语学研究·开篇. 东京：早稻田大学，1994.

[13] 李辉，洪玉龙. 傣傣话——世界上元音最多的语言. 上海：复旦大学出版社，2012.

7.4 古 DNA 分析技术发展的三次革命

古 DNA 是指从考古遗迹和古生物化石标本中获取的古生物的遗传物质[1]。古 DNA 研究是以分子生物学技术为基础发展起来的一个新兴领域，通过古 DNA 研究能够分析古代生物的谱系、分子演化理论、人类的起源和迁徙、动植物的家养和驯化过程等[2]。自 20 世纪 80 年代开始古 DNA 研究以来，研究者们一直在为探索古 DNA 实验技术和建立古 DNA 研究标准而努力[3]。近年来，随着新的古 DNA 扩增和测序技术的出现及不断成熟，古 DNA 研究已逐渐成为一个用途广泛、极有发展前景的领域。

7.4.1 古 DNA 研究的第一次革命

早在 1980 年，湖南医科大学的研究人员就从马王堆汉墓的女尸中提取出了 DNA 和 RNA，这是世界上最早的古 DNA 研究[4]。古 DNA 研究的真正起步是利用了分子克隆这一技术，即将提取出的古 DNA 构建入测序载体，在宿主菌中增殖后进行测序(图 7－13)。

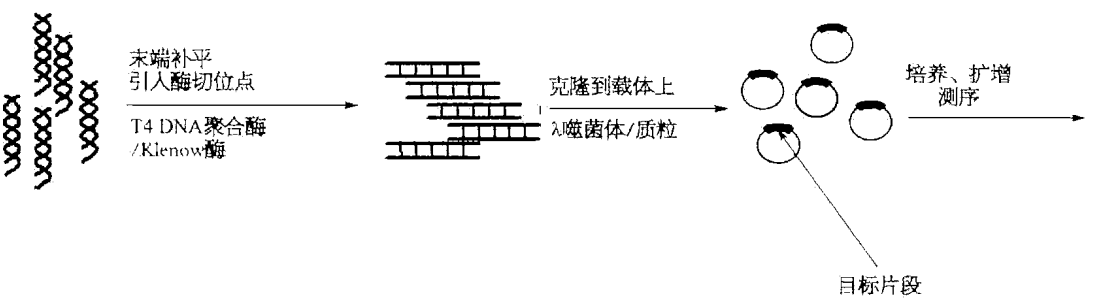


图 7－13 古 DNA 的分子克隆

1984 年，Higuchi 等[5]从保存约 140 年的斑驴(*Equus quagga quagga*，已于 1883 年灭绝)皮肤中得到了 229 bp 的线粒体 DNA 序列，并通过与其近亲——马、驴、斑马的相关线粒体 DNA 序列进行比对，得出斑驴与斑马的亲缘关系最近，而与马或驴的较远。1985 年，Pääbo 等[6]对 23 具埃及木乃伊进行了分析，从一具 2 400 年前的木乃伊孩童上克隆得到 3.4 kb 的 DNA 片段。分子克隆方法在以上两项目中的成功运用，充分证明了古 DNA 研究的可行性和重要性，开创了古 DNA 研究的先河。

随着研究的不断深入，分子克隆方法的缺陷也逐渐暴露：分子克隆需要的古 DNA 量较大，而多数古生物材料数量较少并有其他研究价值，且古生物样品中古 DNA 的含量极低，Handt 等[7-9]的研究表明，在保存状态较好的情况下，每毫克古代样品中约含 2 000 个线粒体 DNA 分子，比新鲜组织中的含量至少低六个数量级；而保存状态一般的样品

中的古 DNA 含量更低，每毫克仅 10～40 个线粒体 DNA 分子。保存下的古 DNA 由于漫长年代中水解作用、氧化作用及环境微生物降解作用等的存在而被严重破坏[10]，其线粒体 DNA 片段长度短于 400 bp，核 DNA 片段则不超过 150 bp[11]，残存的这些片段内部还广泛存在着单链缺口、碱基转换、碱基脱落或分子交联等各种各样的损伤[12,13]。古 DNA 含量极低且损伤严重导致克隆效率低，克隆后的分子在宿主内也可能会受到某些修饰而无法完全真实地反映古代生物的遗传信息。能否成功解决这些问题成为古 DNA 研究能否持续发展的关键。

7.4.2　PCR 技术掀起古 DNA 研究的第二次革命

1983 年，聚合酶链反应(PCR)问世[14,15]，研究人员可以用少量甚至单分子 DNA 模板，通过体外扩增的方式，获得大量的目标 DNA 拷贝。它给生命科学的各个领域带来了革命性的突破，极大地促进了分子生物学的发展。此后，这一突破性技术被广泛应用于古 DNA 研究，PCR 技术使微量的古 DNA 在短时间内大量扩增成为现实，使古 DNA 研究摆脱了较为繁琐的重组及克隆实验过程而真正迅速开展起来。

PCR 技术的原理是：DNA 聚合酶以单链 DNA 为模板，借助一小段双链 DNA 来启动合成，通过一个或两个人工合成的寡核苷酸引物与单链 DNA 模板中的一段互补序列结合，形成部分双链。在适宜的温度和环境下，DNA 聚合酶将脱氧单核苷酸(dNTP)加到引物 3′-OH 末端，并以此为起始点，沿模板 5′→3′方向延伸，合成一条新的 DNA 互补链。PCR 反应的基本成分包括：模板 DNA、引物、dNTP、DNA 聚合酶和适宜的缓冲液。PCR 反应中，通过双链 DNA 的高温变性(denaturation)、引物与模板的低温退火(annealing)和适温延伸(extension)这三步反应反复循环。每一循环中所合成的新链，又都可作为下一循环中的模板。PCR 合成的特定 DNA 序列产量随着循环次数呈指数增加，从而达到迅速大量扩增的目的。

1988 年，Pääbo 等[16]首先将 PCR 技术运用到古 DNA 研究中，从距今约 7 000 年的人颅脑中提取出 92 bp 的线粒体 DNA。Golenberg 等[17]从距今 1 700 万年前的木兰属(*Magnolia*)植物化石中获取到叶绿体 DNA 的 *rbcL* 序列，Woodward 等[18]从距今约 12 000 万年前的恐龙骨骼中提取出 DNA，Desalle 等从琥珀中提取 DNA[19,20]等。

PCR 技术并不是万能的。尽管用 PCR 理论上可以扩增单分子的 DNA，但是很多古代材料中的 DNA 常常无法通过 PCR 被检测到[21-23]。究其原因，主要是从很多古生物样本中抽提到的 DNA 含量太低。相比于现代生物材料产生的 DNA 量，尤其是 PCR 所产生的指数增长的目标片段 DNA 分子，古代样品产生的 DNA 在数量上是没有竞争力的。而 PCR 过程是 DNA 分子和聚合酶碰撞的热力学过程，数量上的劣势会导致古代 DNA 分子在 PCR 时被扩增的概率降低甚至不被扩增[24]。所以在用 PCR 技术扩增极低含量的古 DNA 分子片段的同时，也会将环境中的污染 DNA 放大。如 Gutierrez 等[25]就证明了 Cano[26]于 1993 年报道的取自琥珀中的象鼻虫的 226 bp 的 *ITS* 序列是真菌污染。

面对以上问题，研究者们不断改进 PCR 技术，创立了乳化 PCR、多重 PCR 等方法，

使之更适用于古 DNA 研究。

1. 乳化 PCR

乳化 PCR(emulsion-PCR, emPCR)技术[27,28]是指将 PCR 体系分散在有机相中乳化,以达到单分子扩增的目的。乳化 PCR 分两步:首先把包含模板 DNA 的反应混合物加入硅(氧烷)油(silicone oil)中进行乳化,接下来进行 PCR 反应,这样每个液滴中或者扩增一个分子的 DNA,或者因没有 DNA 模板而无法反应。第一步结束时,包含模板 DNA 的液滴中的 PCR 底物被耗尽;然后,对第一步产物进行离心处理,使小液滴混合为一体,使得第一步中没有模板的小液滴中的 PCR 底物继续进行第二次 PCR(图 7-14)。

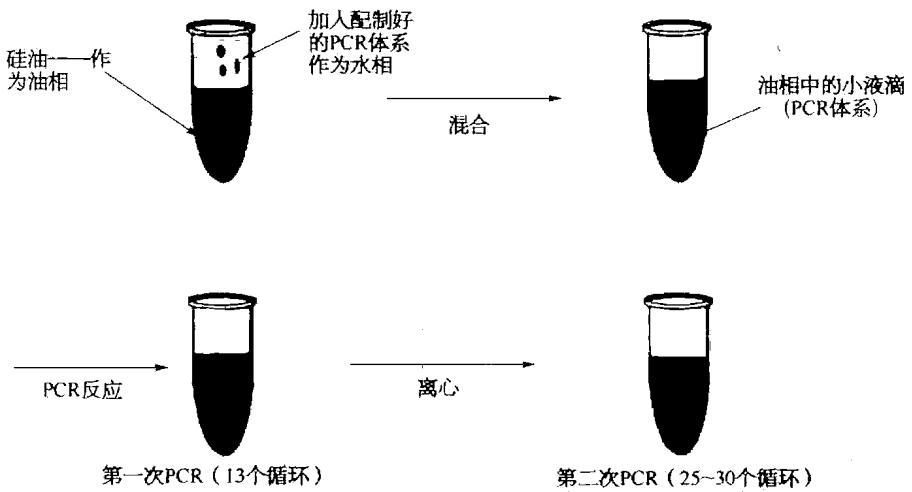


图 7-14 乳化 PCR

2. 多重 PCR 与微测序技术

保存条件相对理想的古 DNA 实验材料一般都来之不易,如何从有限的材料中得到大量的遗传信息以用于后续研究,也一直是研究者颇为关注的问题。古 DNA 分子的长度一般不会超过 500 bp,要得到古生物完整的基因或基因组,需要设计合成大量特异性引物对,将扩增所得序列进行拼接。该如何解决较多的模板需求与有限的材料供给之间的矛盾? Hofreiter 等在 2006 年提出了古 DNA 的分步扩增方法(图 7-15)[29]:先在第一步 PCR 反应中加入多对引物对古 DNA 模板进行多重扩增(multiplex amplification),然后在第二步 PCR 反应中再用第一步反应的引物或者巢式引物(nested primer)对目标片段进行 30~40 个循环的单一扩增(simplex amplification)。相比于普通的 PCR,分步扩增法省时,节约模板 DNA,且灵敏度和通量上都有很大提高。Krause 等设计了 46 对引物,采用多重 PCR 扩增技术,成功地从 200 mg 猛犸象(*Mammuthus primigenius*)骨粉的抽提液中扩增出 16 700 bp 的 mtDNA 基因组全序列[30]。

与多重 PCR 紧密联系的是微测序技术。微测序是指首先扩增出含有 SNP 位点的一段 DNA,然后在 PCR 扩增产物中加入一检测引物进行微测序反应。微测序引物 3′-OH

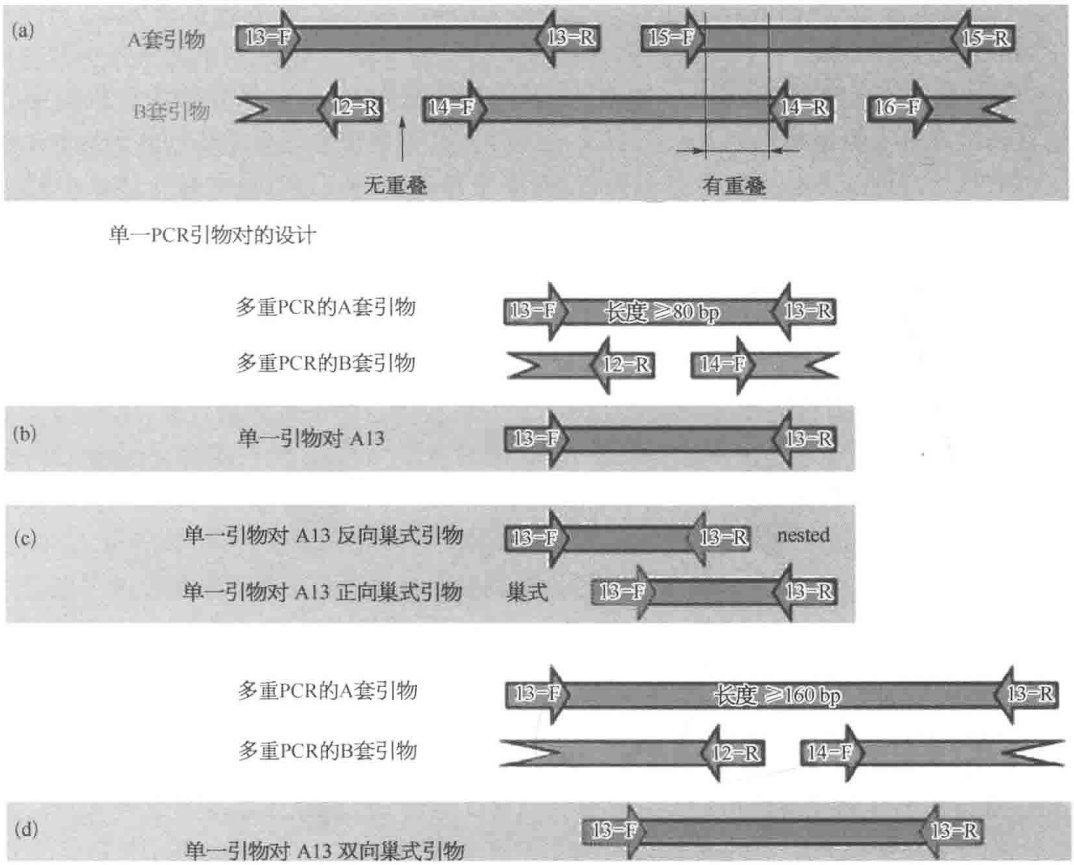


图 7-15　多重 PCR 的引物设计原则

(a) 多重 PCR 的第一步需要两套引物，这里分别称为 A 和 B，这两套引物可覆盖我们想要的序列。由单独一套引物扩增出的产物片段不会重叠，而在两套引物之间的 PCR 片段则会出现重叠。(b-d) 第二步单一 PCR 是将第一步多引物 PCR 产物片段单独进行再次扩增。图 b 所示是用与多引物扩增中相同的单一引物对，图 c 表示仅在所得 PCR 片段的一端用巢式引物，b 和 c 的方法均可以用于大于 80 bp 的 PCR 产物的再扩增。图 d 表示在 PCR 产物的两端均使用巢式引物，这种方法仅在 PCR 产物大于 160 bp 时才可使用。

末端碱基紧挨于多态性碱基，加入 DNA 聚合酶及荧光标记的 dNTP 后，只进行一个碱基的延伸反应，延伸的这个碱基就是多态性碱基，最后经毛细管电泳检测荧光信号，通过不同的荧光检测结果达到对 SNP 位点的分析。

Endicott 等将多重 PCR 和微测序技术用于古 DNA 研究，实现对 20 个线粒体 DNA 编码区的信息位点同时进行检测，弄清了安达曼人线粒体 DNA 单倍群的起源[31]。

7.4.3　下一代测序技术引领古 DNA 研究的第三次革命

PCR 技术将古 DNA 研究推向巅峰。在巨大的成功面前，一些学者指出，现有的古 DNA 技术是难以获得灭绝生物的核基因组全序列的，尽管多重 PCR 可以快速获得线粒体 DNA 全序列，但对核基因组来说却完全是另外一回事，古 DNA 研究或许就此止步[30,32]。而下一代测序技术的兴起却让人们看到了希望。在这之前，古 DNA 测序使用

传统方法先通过克隆或PCR来达到一定的模板数量。然后这些模板被逐一测序，并通过毛细管电泳将测序产物分离。新的测序方法不是通过毛细管电泳来测序，而是通过序列的合成过程来进行测序，主要有454焦磷酸测序法(454 pyro- sequencing)、Solexa合成测序法(Solexa SBS sequencing)、寡核苷酸聚合酶群落测序法(SOLiD polony sequencing)和单分子测序法(single molecule sequencing)。

454焦磷酸测序法[33]首先是将待测DNA处理成小于500 bp的片段并制备成单链DNA文库，使相同长度和碱基组成的DNA连上相同的接头，再用含配对接头的吸附珠吸附特定DNA分子后进行乳化PCR，洗去乳胶物质并使双链DNA变性成单链后转入多微孔板中，再开始由4种酶催化的同一反应体系中的酶级联反应，4种酶分别为：DNA聚合酶、ATP硫酸化酶、荧光素酶和双磷酸酶，反应底物为腺苷酰硫酸(APS)、荧光素。向反应体系中只加入一种dNTP，若其能与DNA模板的下一个碱基配对，则会在DNA聚合酶的作用下，添加到测序引物的3′-OH末端，同时释放出一分子的焦磷酸(PPi)；在ATP硫酸化酶的作用下，生成的焦磷酸可以和APS结合形成ATP；在荧光素酶的催化下，生成的ATP又可以和荧光素结合形成氧化荧光素，同时产生可被光学系统所检测的可见光而获得一个特异的检测峰；每个峰值的高度与反应中掺入的核苷酸数目成正比。反应体系中剩余的dNTP和残留的少量ATP在双磷酸酶的作用下发生降解。然后加入下一种dNTP，继续DNA链的合成，因而是边合成边测序(sequencing by synthesis)。

Solexa测序法[34]首先需要将待测DNA片段分别接上接头，DNA片段通过接头与芯片上引物序列的碱基互补锚定到芯片上的特定位置，继而进行桥联扩增(bridge amplification)形成数百万的单分子阵列。测序时向芯片上加入4种被不同可逆终止化合物标记的碱基，其中能与芯片上各单分子阵列中单链DNA的第一个碱基配对的核苷被加到引物的3′-OH末端，其自身的3′-OH末端被保护以阻止下一个核苷的加入，多余的核苷被移走。同阵列中新加入的核苷释放出一致的荧光信号，该信号被图像采集系统采集后，荧光基团从新加入的核苷碱基上脱落，同时其3′-OH末端去保护成游离状态为下一个核苷的连接做准备。以上步骤可以重复几十个循环，得到的大量短片段用相关软件程序进行拼接和处理。Solexa合成测序法除了高效快速，主要优势在于大大降低了测序成本。

SOLiD测序法[35]是在测序反应前将随机打断的单个DNA分子与含不同标签序列的片段通过A-T互补连接，得到含配对的特异标签的DNA文库，然后用内含聚合酶阵列的吸附珠吸附DNA分子并进行乳化PCR扩增，离心去除未发生扩增反应的吸附珠后将含数百万个相同DNA分子拷贝的单层吸附珠密集排布在固体介质表面，测序仪根据介质表面不同位点所发出的不同荧光信号来确定不同的碱基。

单分子测序法[36]也是一种合成测序法，但是不需要模板的预先扩增。这种技术利用高清晰度的光学设备检测测序反应中单链DNA或RNA模板中单个碱基的加入，避免PCR所造成的偏向性，可用于定量。但该方法的错误率较高，尚不实用[37]。

这些高通量的测序技术在古DNA中的应用，突破了以前的古DNA研究主要基于细

菌中随机的分子克隆或者目的 DNA 片段 PCR 扩增的低通量的技术瓶颈。Noonan 等[39,40]首先尝试了不经 PCR 扩增而是通过构建古 DNA 宏基因组文库(metagenomic library)来直接克隆古 DNA。宏基因组[38]是特定环境全部生物遗传物质的总和,宏基因组研究是以生态环境中全部 DNA 作为对象,通过克隆、异源表达来筛选相关基因,研究其功能和彼此之间的相互作用等,而基因组文库(library)是指包含特定 DNA 片段的重组子集合。Noonan 采用末端补平-平末端连接-克隆的方法(图 7-16)构建宏基因组文库(metagenomic library),成功获得了洞熊(*Ursus spelaeus*)的全基因组序列。

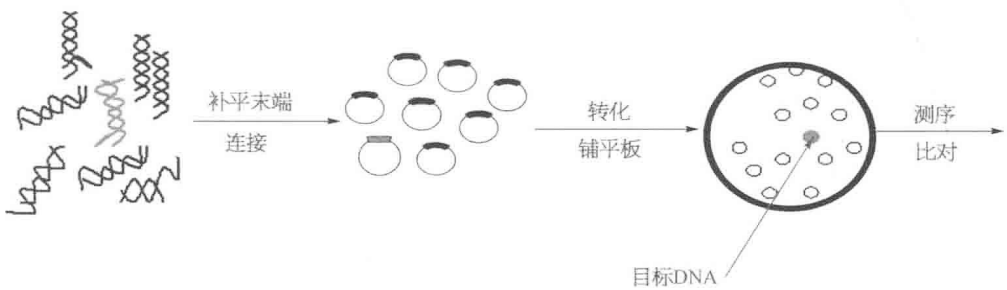


图 7-16　宏基因组文库的构建和分析

2006 年,Grenn 等[41]在宏基因组法的基础上结合乳化 PCR 和焦磷酸测序得到并分析了一个 3.8 万年前的尼安德特人的全基因组序列中的 1Mb(图 7-17)。Poinar 等[42]用同样方法进行西伯利亚猛犸象下颚骨的核 DNA 和线粒体 DNA 宏基因文库构建和大规模测序,共获得 28Mb,其中 13Mb(45.4%)是属于猛玛象的,剩余部分 DNA 则是污染

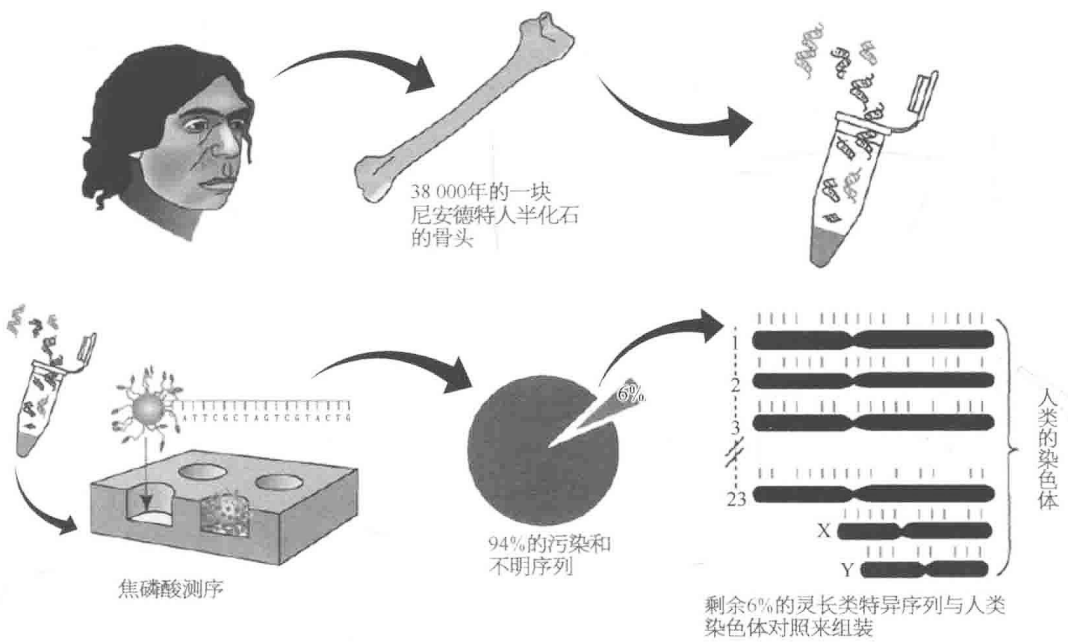


图 7-17　使用焦磷酸测序法对尼安德特人核 DNA 和线粒体 DNA 测序

或者环境 DNA。

2008 年，Blow 等[43]采用接头介导的 emPCR，结合 Solexa 测序技术，对 4.5 万年和 6.9 万年前的古哺乳动物样品进行全基因组测序，共获得 100Mb，也证实高通量法研究核 DNA 的可行性。之后，针对全基因组的古 DNA 研究不断开展起来[44-49]。

在这个过程中，古 DNA 扩增和测序技术不断被优化。2009 年，Briggs 等[50]创立了引物延伸捕获法(primer extension capture，PEC)。PEC 方法简单、快速且特异性好，非常适合捕获小目标片段(图 7-18)。

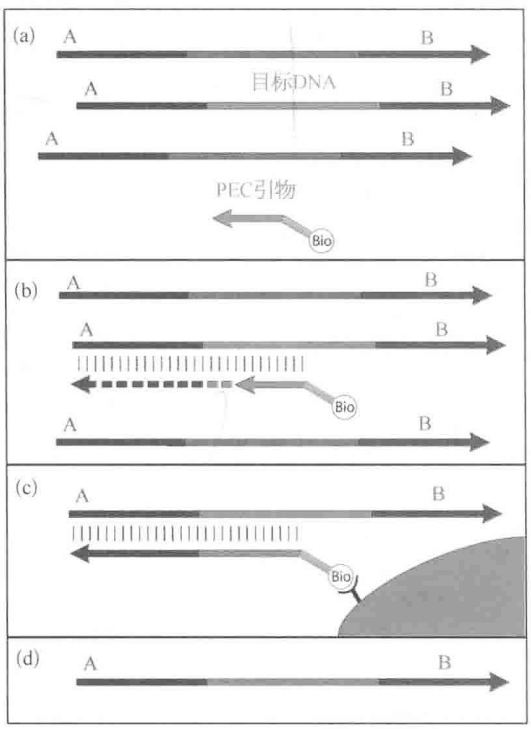


图 7-18 引物延伸捕获技术

(a) 将 5′端带有生物素的寡核苷酸引物加入到 454 文库中(接头分子 A 和 B 中带有特殊的标签)，退火时引物可与它们各自的目标序列结合；(b) *Taq* DNA 聚合酶介导的单向延伸反应使得引物和带有 5′端接头的目标序列结合成双链；(c) 纯化，去除多余的 PEC 引物。生物素标记的引物与目标序列被带有链霉亲和素的磁珠捕获。在 PEC 引物的变性温度下用洗脱液洗涤磁珠来保证发生延伸反应的模板 DNA 优先与引物结合；(d) 从珠子上洗脱下捕获的目标 DNA 并在其接头位点处进行扩增。扩增产物能被用于第二轮的捕获，或者直接用于单分子乳化 PCR 以供 454 测序。

Briggs 等用该方法重建了来自不同地理区域的 5 个尼安德特人的线粒体 DNA 全序列，分析出晚期的尼安德特人线粒体 DNA 的遗传多态性约为现代人的 1/3，且尼安德特人的有效群体比现代人和现存的大猩猩都要小。Krause 等[51]用同样方法从出土于西伯利亚 Denisova 洞穴的一块指骨化石中提取出线粒体 DNA 全序列，进一步分析发现这一

线粒体 DNA 序列同任何已知的古人种都不一样。这一人种(丹人)约在 100 万年前走出非洲,行至西伯利亚后逐渐消失。

PEC 法获取线粒体 DNA 这样的小片段的能力毋庸置疑,但它难以实现对较大的目标片段,诸如外显子、大的染色体区域或者低覆盖的鸟枪法基因组中的目的片段等的抓取。Burbano 等将 Hodges 创立的芯片捕获外显子的方法(图 7-19)[52,53]用于古 DNA 研究,从而成功解决了这个难题。

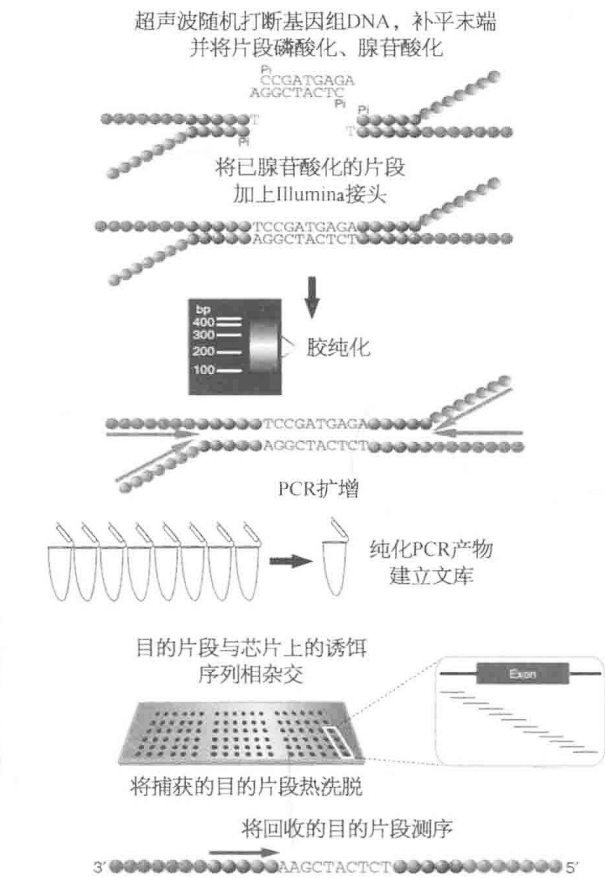


图 7-19　芯片杂交捕获方法简图

Burbano 等[54]用 AHC 方法来分析发现于西班牙的一块距今 49 000 年的尼安德特人的遗骨。他们将目光聚焦在 14 000 个蛋白编码区,通过与现代人、黑猩猩和红猩猩等的对比,发现了 88 个在现代人与尼安德特人分开之后才产生的替代氨基酸。这为研究基因的进化和多样性提供了极其重要的线索。

同一期的 *Science* 杂志上又刊出下一代测序技术在古 DNA 领域的运用取得的最引人注目的成就。Pääbo 领导的研究组以 454 焦磷酸测序和 Solexa 测序技术从 3 个女性尼安德特人骨骼中获得了 4 000Mb,做出了尼安德特人的基因组草图,他们将测序结果与来自世界 5 个地区的现代人基因组进行比较后发现,现代人有 1%～4%的 DNA 源自

尼安德特人，即是说现代人与尼安德特人非常可能在小范围内发生过基因交流，时间是现代人走出非洲后在中东遇到尼安德特人之时[55]。尼安德特人基因组序列首个版本的获得，揭开了尼安德特人与现代人是否有基因交流的谜题。

从 1984 年 229 bp 斑驴的线粒体 DNA 到 2010 年的尼安德特人基因组草图，分子克隆、PCR、下一代测序技术、PEC 和 AHC 等引领古 DNA 研究蓬勃发展，许多以前因技术问题而无法解决的谜题将被一一揭开。随着标准的不断细化、技术的不断发展，古 DNA 研究在古病理学、进化速率估计、已灭绝生物进化谱系、群体历史、动植物驯化、人类历史和迁徙等研究领域中必将发挥更大的作用。

参考文献

[1] Poinar H N. DNA from fossils: the past and the future. Acta Paediatrica, 1999, 88: 133 - 140.

[2] Hofreiter M, Serre D, Poinar H N, et al. Aneient DNA. Nat Rev Genet, 2001, 2: 353 - 359.

[3] Willerslev E, Cooper A. Ancient DNA. Proc Biol Sci, 2005, 272: 3 - 16.

[4] Hunan Medical College. Study of an ancient cadaver in Mawangtui tomb No. 1 of the Han Dynasty in Changsha. New York: Ancient Memorial Press, 1981: 184 - 187.

[5] Higuchi R, Bowman B, Freiberger M, et al. DNA sequences from the quagga, an extinct member of the horse family. Nature, 1984, 312: 282 - 284.

[6] Pääbo S. Molecular cloning of Aneient Egyptian mummy DNA. Nature, 1985, 314: 644 - 645.

[7] Handt O, Richards M, Trommsdorff M, et al. Molecular genetic analyses of the Tyrolean Ice Man. Science, 1994, 264: 1775 - 1778.

[8] Handt O, Krings M, Ward R H, et al. The retrieval of ancient human DNA sequences. Am J Hum Genet, 1996, 59: 368 - 376.

[9] Graziano P, Cecilia S. A novel method for estimating substitution rate variation among sites in a large dataset of homologous DNA sequences. Genetics, 2001, 157: 859 - 865.

[10] Wandeler P, Smith S, Morin P A, et al. Patterns of nuclear DNA Degeneration over time-a case study in historic teeth samples. Mol Ecol, 2003, 12: 1087 - 1093.

[11] Pääbo S, Higuchi R G, Wilson A C. Ancient DNA and the polymerase chain reaction. The emerging field of molecular archaeology. J Biol Chem, 1989, 264: 9709 - 9712.

[12] Greer S, Zamenhof S. Studies on depurination of DNA by heat. J Mol Biol, 1962, 4: 123 - 141.

[13] Poinar H N. The genetic secrets some fossils hold. Acc Chem Res, 2002, 35: 676 - 684.

[14] Saiki R K, Scharf S, Faloona F, et al. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. Science, 1985, 230: 1350 - 1354.

[15] Mullis K B, Faloona F A. Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. Methods Enzymol, 1987, 155: 335 - 350.

[16] Pääbo S, John A Gifford, et al. Wilson Mitochondrial DNA sequences from a 7000-year old brain. Nucleic Acids Res, 1988, 16: 9775 - 9787.

[17] Golenberg E M, Giannasi D E, Clegg M T, et al. Chloroplast DNA sequence from a miocene Magnolia species. Nature, 1990, 344: 656 - 658.

[18] Woodward S R, Weyand N J, Bunnell M. DNA sequence from Cretaceous period Bone fragments. Science, 1994, 266: 1229 - 1232.

[19] Desalle R, Gatesy J, Wheeler W, et al. DNA sequences from a fossil termite in Oligo-Miocene amber and their phylogenetic implications. Science, 1992, 257: 1933 - 1936.

[20] Desalle R, Bareia M, Wray C. PCR jumping in clones of 30-million-year-old DNA fragments from amber preserved termites (Mastotermes electrodominicus). Experientia, 1993, 49: 906 - 909.

[21] Handt O, Höss M, Krings M, et al. Aneient DNA: methodological challenges. Experientia, 1994, 50: 524 - 529.

[22] Poinar H N, Höss M, Bada J L, et al. Amino acid racemization and the preservation of ancient DNA. Science, 1996, 272: 864 - 866.

[23] Kumar S S, Nasidze I, Walimbe S R, et al. Brief communication: discouraging prospects for aneient DNA from India. Am J Phys Anthropol, 2000, 113: 129 - 133.

[24] 徐智. 中国西北地区古代人群的DNA研究. 人类生物学博士毕业论文. 复旦大学, 2008.

[25] Gutierrez G, Marin A. The most ancient DNA recovered from an amber-preserved specimen may not be as ancient as it seems. Mol Biol Evol, 1998, 15: 926 - 929.

[26] Cano R J, Poinar H N, Pieniazek N J, et al. Amplification and sequencing of DNA from a 120 - 135-million-year-old weevil. Nature, 1993, 363: 536 - 538.

[27] Ohuchi S, Nakano H, Yamane T. In vitro method for the generation of protein libraries using PCR amplification of a single DNA molecule and coupled transcription/translation. Nucleic Acids Res, 1998, 26(19): 4339 - 4434.

[28] Nakano M, Komatsu J, Matsuura S, et al. Single-Molecule PCR using water-in-oil emulsion. J Biotechnol, 2003, 102: 117 - 124.

[29] Römpler H, Dear P H, Krause J, et al. Multiplex amplification of ancient DNA. Nat Protoc, 2006, 1: 720 - 728.

[30] Krause J, Dear P H, Pollack J L, et al. Multiplexed amplification of the mammoth mitochondrial genome and the evolution of Elephantidae. Nature, 2006, 439: 724 - 727.

[31] Endicott P, Metspalu M, Stringer C, et al. Multiplexed SNP typing of ancient DNA clarifies the origin of andaman mtDNA haplogroups amongst South Asian tribal populations. PLoS One, 2006, 1: e81.

[32] Cooper A. The year of the mammoth. PLoS Biol, 2006, 4: e78.

[33] Margulies M, Egholm M, Altman W E. Genome sequencing in microfabricated high density picolitre reactors. Nature, 2005, 437: 376 - 380.

[34] Ju J, Kim D H, Bi L, et al. Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. Proc Natl Acad Sci USA, 2006, 103: 19635 - 19640.

[35] Shendure J, Porreca G J, Reppas N B, et al. Accurate multiplex polony sequencing of an evolved bacterial genome. Science, 2005, 309: 1728 - 1732.

[36] Mitchelson K R. New high throughput technologies for DNA sequencing and genomics. Capitalbio Corporation Beijing China, 2007: 1 - 20.

[37] Korlach J, Marks P J, Cicero R L, et al. Selective aluminum passivation for targeted

immobilization of single DNA polymerase molecules in zero mode waveguide nanostructures. Proc Natl Acad Sci USA, 2008, 105: 1176 - 1181.

[38] Handelsman J, Rondon M R, Brady S F, et al. Molecular biological access to the chemistry of unkown soil microbes: a new frontier for natural products. Chem Biol, 1998, 5(10): 245 - 249.

[39] Hoelzel A R. Ancient genomes. Genome Biol, 2005, 6: 239.

[40] Noonan J P, Hofreiter M, Smith D, et al. Genomic sequencing of Pleistocene cave bears. Science, 2005, 309: 597 - 599.

[41] Green R E, Krause J, Ptak S E, et al. Analysis of one million base pairs of Neanderthal DNA. Nature, 2006, 444: 330 - 336.

[42] Poinar H N, Schwarz C, Qi J, et al. Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. Science, 2006, 311: 392 - 394.

[43] Blow M J, Zhang T, Woyke T, et al. Identification of the source of ancient remains through genomic sequeneing. Genome Res, 2008, 18: 1347 - 1353.

[44] Bentley D R, Balasubramanian S, Swerdlow H P, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature, 2008, 456: 53 - 59.

[45] Green R E, Malaspinas A S, Krause J, et al. A complete Neandertal mitochondrial genome sequence determined by high-through put sequencing. Cell, 2008, 134: 416 - 426.

[46] Gilbert M T, Drautz D I, Lesk A M, et al. Intraspecific phylogenetic analysis of Siberian woolly mammoths using complete mitochondrial genomes. Proc Natl Aead Sci USA, 2008, 105: 8327 - 8332.

[47] Mardis E R. Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet, 2008, 9: 387 - 402.

[48] Millar C D, Huynen L, Subramanian S, et al. New developments in ancient genomics. Trends Ecol Evol, 2008, 23: 386 - 393.

[49] Rasmussen M, Li Y R, Lindgreen S, et al. Ancient human genome sequence of an extinct Palaeo-Eskimo. Nature, 2010, 463: 757 - 762.

[50] Briggs A W, Good J M, Green R E, et al. Targeted retrieval and analysis of five Neandertal mtDNA genomes. Science, 2009, 325: 318 - 321.

[51] Krause J, Fu Q, Good J M, et al. The complete mitochondrial DNA genome of an unknown hominin from southern Siberia. Nature, 2010, 464: 894 - 897.

[52] Hodges E, Xuan Z Y, Balija V, et al. Genome-wide in situ exon capture for selective resequencing. Nat Genet, 2007, 39: 1522 - 1527.

[53] Hodges E, Rooks M, Xuan Z Y, et al. Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. Nat Protoc, 2009, 4: 960 - 974.

[54] Burbano H A, Hodges E, Green R E, et al. Targeted investigation of the Neandertal genome by array-based sequence capture. Science, 2010, 328: 723 - 725.

[55] Green R E, Krause J, Briggs A W, et al. A draft sequence of the Neandertal genome. Science, 2010, 328: 710 - 772.

索 引

A

阿尔泰语系 9,243,244,261,266,273,343
阿伊努人 279,281,282
爱斯基摩语系 344
安达曼人 18
奥莫现代人 45
澳大利亚人 60,63,67
澳大利亚人种 17,50,59
澳台语系 190

B

白马氐人 18
白种人 19,50
百濮 22
百越 19,22,190,345
班图黑人 17
傍人 42
北方汉族 103,105,107
俾格米人 49,338
波利尼西亚 183
波利尼西亚人 187,188,191
波利尼西亚序列 188,190
布须曼人 17,49,338

C

曹操 67,287,298,300－303,305,308
常染色质 26
橙色人种 49
"丛林过滤"效应 180

D

大坌坑文化 19
大汶口文化 19
大溪文化 11,19,208,211
傣族 9
丹尼索瓦人 34,43,45
单倍群 17,46,291
单倍型 6,291
僜人 89,139,144,150
氐羌族群 152
地猿 41
侗傣 20,84,339
侗傣语系 9,125,166,167,203,345
侗傣族群 23,200
侗台 119
侗台语 343
多地区起源说 37

F

非重组区 26
非洲起源说 35
菲律宾人 9
分子方差分析 145
分子人类学 1,13
弗洛勒斯人 43

G

港川人 274－276
高加索人种 50,59
格鲁吉亚人 43
古 DNA 207,313,347
古汉族 22
古人类 57
古西伯利亚语系 243
古亚细亚语 271
古亚语系 343
国际 Y 染色体命名委员会 17

H

哈萨克族 245,251,252,254,261
海德堡人 44
海南原住民 213,219,236－238
汉藏 20,144,150,200,339
汉藏语系 9,120,124,139,339,343
汉藏族群 19
汉族 9,11,22,84,88,109,119
河姆渡文化 11
黑猩猩 30,41
黑种人 19
宏基因组 352
华夏族 19,22,110,120
黄种人 19,50
回辉 233
回辉人 168,201,236－238,240
回族 245,251,254,261,262

J

家谱 294,298,300
柬埔寨人 9
匠人 43
羯人 270,271
荆蛮 19,23
精子发生障碍 313,317,319,320,333,335
净化选择 29－31

K

卡岱语族 168,222
卡岱族群 230,231
科伊桑语系 338,339
空间遗传自相关性 85
快车模式 183,187

L

拉祜族 136
黎族 20,168,213,222,231,232
李家村文化 19
历史人类学 1,302
良渚文化 10,19,22,208,211
琉球人 281,282
柳江人 275
龙山文化 11,19,20,23,208
卢道夫人 42,43
罗得西亚人 34,44,45
珞巴族 89,139,144,150

M

马来-波利尼西亚 204
马来人群 191,203
马来西亚人 9
满族 9
美拉尼西亚人 188
蒙古利亚人 60,67
蒙古人种 50,59
蒙古语 266
蒙古语族 252,261,270
蒙古族 9,246,251,252,254
孟高棉 166,170,179
孟高棉族群 19,22,23
弥生人 244,274,277,282,283
密克罗尼西亚人 188
缅彝语支 22
苗瑶 20,23,119,166,170,179,339
苗瑶语系 9,125,166,343

苗族　9
摩梭人　90,152,161－163

N

纳丁尼语系　344
纳卡里猿　40
纳西族　90,136,152,161,162
男性不育　324,326,333
南部藏缅　125,135,136
南岛语系　9,166,190,203,343
南方汉族　103,105,107
南方原住民　119,120
南方智人　34
南亚语系　9,166,343
南猿　42
能人　42,43
尼安德特人　3,34,43,45,353,355
尼格利陀人　18,49,60,64,67
尼格利陀人种　59
尼格罗人　49
拟常染色体区　26

P

裴李岗文化　19
平话汉族　111
平话人　88,89,116,119－121
瓶颈效应　15
普米族　90,163

Q

千禧人　41
羌族　90
青莲岗文化　19

R

人口扩张假说　107
人类基因组　13
人族　41

S

撒拉族　246,251
赛典赤·瞻思丁　287,309
山顶洞人　275
畲族　9
社会选择　30,67
生物人类学　3
绳文人　50,244,274,276,277,281－283
树居人　43

T

台湾起源论　190,203
台湾少数民族　23,183,187,200,204
傣傣话　340,345
通古斯　19
通古斯语族　252,266
同源重组　16
突厥语　266,270
突厥语族　252,261
茶拉维达语系　342
土家族　136
土族　246,251,254
托塔维尔人　43

W

晚亚洲人　17,19,20
维吾尔族　9,245,251,252,254,261
文化人类学　3
文化系列　207
沃利延语　343
乌拉尔语系　243,343
乌拉尔族群　23
吴城文化　208

X

西布兰诺人　43
虾夷人　18
先驱人　44

现代人　3，34，43，45，51，57，67，244，260，355
小黑人　18
性别偏向　28
匈奴族　20
选择性清除　27

Y

Y染色体谱系树　28
雅爱语　343
亚美利加人种　50
仰韶文化　11
瑶族　9
叶尼塞　20，339
叶尼塞语系　20，243，343
遗传搭车效应　27
遗传漂变　15
彝族　136
仡佬族　168，222，231，232
仡隆人　20，168，213，221，230－232
异染色质　26
印度尼西亚人　9
印度尼西亚人群　200
印度支那人　238
优必语　339
有效群体　29
元音音位　340
原始长江语系　180
越南人　9

Z

早亚洲人　17，19
藏缅　89
藏缅语族　124
藏语支　122
藏族　9，89，122，124，135，152，161，162
占城人　233，236－238，240
占族　201
长者智人　45，46
真人属　42
正向选择　31
郑和　287，309，310
直立人　43
智人　44
中华民族　1
壮族　9
自然选择　29，30，37
棕色人种　17，50
棕种人　19
左镇人　202

后 记

这本书的编写，耗时不可谓不长，耗力不可谓不多。追溯起来，本书是复旦大学人类学实验室二十年科研成果的积累。在这二十年时间中，本实验室在各种学报、期刊上发表了大量的科研报告或者综述文章，不过大多数是英文的。这些文献对于很多对分子人类学有兴趣的业内外读者来说，阅读的难度颇大。要在这么多论文的基础上进行系统总结，把它编撰成可读性强的中文专著，并不是一件容易的事情。首先，实验室的众多师生把英文论文翻译成中文，参与翻译工作的包括：第二章第一节王晶，第三节肖晗，第四节王传超，第五节沈溧冰；第三章第一节王凌翔，第二节李辉，第三节甘瑞静，第四节张曼菲，第五节佟欣竹，第六节陆艳，第七节胡雅；第四章第一节蔡晓云，第二节陈丹丹，第三节张曼菲，第四节张曼菲，第五节李冬娜，第六节王凌翔，第七节王凌翔；第五章第三节丁琦亮；第六章第二节王传超，第三节王传超；第七章第一节刘港彪，第二节董征。其次，由于撰写于不同年代、不同文献里的科技名词并不统一，有的还需要重新实验分析以达到一致标准，这也是很繁重的工作。特别是Y染色体系统树的更新，使得Y染色体单倍群名称需要调整，相关内容根据最新的更新系统树调整改编。而且在最初发表论文时，由于数据有限，有些结论不太准确，在本书编写中对这些分析和结论做了必要的补充或修改。所以本书的素材虽然来源于多年来已发表的论文，实际上是对成果的重新整理和总结，所呈现的是最新的数据和观点。另外，在民族学方面，姚大力教授提出了很多宝贵的建议，使我们能更好地解读数据。

当然，随着今后研究的深入，我们对东亚族群的演化历程会越来越清晰。这本书的总结，正好是Y染色体单倍群分析向全染色体序列分析的时候，是一个时代结束、一个时代开始的时候。在这个时机做一个总结，具有特殊的意义。这也是我们出版这本书的最重要的原因。

李辉　金力

2015年1月