南亚语系人群源于印度蒙达部落群

兰海
中国语言论坛, 苏州 215000

评论文献： Kumar V, Reddy AN, Babu JP, Rao TN, Langstieh BT, Thangaraj K, Reddy AG, Singh L (2007) Y-chromosome evidence suggests a common paternal heritage of Austro-Asiatic populations. BMC Evol Biol. 7:47.

编辑提要：南亚语系人群主要分布于东南亚的中南半岛和印度，可能是蒙古人种中最古老的族群，在东亚人群起源研究中极为关键。这个族群的体质特征非常多样，遗传背景也较为复杂。对其主要 Y 单倍群 O-M95 的多样性分析发现，南亚语系的分支，印度的蒙达语族群体多样性远远大于东南亚的另一个分支孟高棉语族群体。因此南亚语系群体可能起源于印度。但是该项研究中涉及的孟高棉语族样本较少，结果或许存在偏差。（李辉，耶鲁大学遗传学系）

关键词：南亚语系；蒙达部落；Y 染色体；东南亚人群起源

Austro-Asiatic Populations originated in Mundari Tribes of India

LAN Hai
Linguistic Forum of China, Suzhou 215000 China

Editor's Summary: The Austro-Asiatic speaking ethnic group mostly distributes in the Indo-China Peninsular in Southeast Asia and India, and may be the oldest group of the Mongoloid groups, therefore, plays a most important role in the studies on the origin of East Asian. The physical characters of the Austro-Asiatic group are quite various, and the genetic structure is relatively complicated. The diversity of the major Y-chromosome haplogroup of Austro-Asiatic group, O-M95, was analyzed. The result showed that the variance of the Mundari tribes in India, a subfamily of Austro-Asiatic, is much larger than that of another subfamily, Mon-Khmer in Southeast Asia. Therefore, Austro-Asiatic group might have originated from India. However, the sample size of Mon-Khmer populations in this study is rather small, which might have caused a certain misapprehension. (LI Hui, Department of Genetics, Yale University)

Key words: Austro-Asiatic; Mundari tribes; Y chromosome; Origin of Southeast Asian population

库玛等的论文《Y 染色体证据显示南亚语系人群有共同的父系起源》比较了南亚语系下面三个语族（蒙达、卡西-克木、孟高棉）人群的 Y 染色体多样性，发现 O-M95 是整个南亚语系人群中最重要的单倍群。其中印度的蒙达语人群的 STR 多样性最高，因此得出 O-M95 起源于蒙达语人群的结论。并通过比较三个语族人群 Y 染色体和 mtDNA 成分的不同，得出东南亚的南亚语系人群是从现在的蒙达语族人群地区经由印度东北走廊迁徙而来的结论。这个结论与语言学证据一致，后者认为蒙达语族是南亚语系中语法和语音都最复杂的语族。

该论文的推理过程的逻辑性很清晰。一个单倍群最有可能起源的地方可以用两个特征来推断: 高的频率和高的多样性。在东南亚的南亚语系人群中，O-M95 的最高频率为 35%(除了 3 个样本过小的人群)。而蒙达语族人群的样本量都很大(35~109,除了 2 个)，O-M95 的频率平均为 63%（排除了 3 个起源有争议的人群）。这明显比东南亚的南亚语系人群的要高。而且，STR 网络结构图显示，

收稿日期：2007 年 5 月 19 日 **修回日期：**2007 年 5 月 24 日 **联系人：**兰海 ranhaer@gmail.com

蒙达语族人群样本的多样性比东南亚的南亚语系人群要高得很多。

以上结论补充或动摇了之前相关研究得到的一些观点，如 Basu 等[1]发现蒙达人有很高频率的 K- M9，因此推断这些南亚语系人群是从非洲经由中亚迁入印度的。库玛等[2]通过研究印度和东南亚的 9bp 缺失/重复的多态性，认为印度的南亚语系人是多重起源的，特别是孟高棉语的亚洲起源和蒙达语的非亚洲起源。

此项研究得出的迁徙路线为：蒙达语地区(~65,000YBP)—印度东北(~57,000YBP)—东南亚（~8,000YBP）—尼科巴群岛(~17,000YBP)。这一结论将有助于对中南半岛人群的结构进行更深入的研究。我们注意到，作为南亚语系特征型的 O-M95 在侗台语，苗瑶语，藏缅语诸民族也有高频存在。比如壮族的 38.3%[3]，纳西族的 47.5%[3]，瑶族的 20%[4]。这一点提示我们语言与人群的复杂关系：在某些民族中，O-M95 是主要单倍群，但其语言并不属于南亚语系。另外，新近的一项研究提到，O-M119 的年代约为 20,000YBP[5]。比较本论文提供的 O-M95 单倍群和 O-M119 的分布：相对来说，在东南亚，O-M95 主要分布内陆而 O-M119 主要沿海岸线分布，而 O-M119 向西的迁徙发生得较为晚近的[5]。因此我们可以确定 O-M95 比 O-M119 的年代要古老得多。这个结果也将有助于研究澳泰语系的扩散对南亚语系的影响。

文中也有一些欠缺：1、得到的 O-M95（O2a*）的年代超过所有先前研究关于 O/N 得到的年代。文中有提到， 用全部的样本计算时，“发现随着 MCMC 循环次数的增加，$N_{posterior}$值下降了而扩张时间增大了，(计算结果)在 10^9时也不稳定”。所以没有使用全部的样本来计算。这有可能导致误差； 2、文中关于单倍群 N 的讨论没有利用到最新的信息。文中提到，“现代人可能从非洲迁往中亚，单倍群 N-LLY22g 和 O-M175 在那里起源”。而 Siiri Rootsi 等[6]的最新研究认为，“N 系下频率最高的 N3，可能起源于中国”。另外，从文中 N*的分布图我们可以看出，N 是由东南亚向北扩散的； 3、文中提到尼科巴人是从泰国和缅甸南部海岸迁往安达曼和尼科巴群岛的。但文中给出的证据不能完全排除经由苏门答腊岛的可能； 4、本文的语言分类与规范的语言学分类不一致[7]。在语言学界的传统分类法中，卡西－克木语仍属于孟高棉语族北高棉语支，整个孟高棉语族分为孟、京芒、北高棉、侏、东高棉、尼科巴、以及阿斯利等语支。但这种分类的不同对结果应该没有影响。

我们也很关注 STR 分析涉及到的样本结构。相对于东南亚的孟高棉群体庞大的人口数来说，它在该文中样本量很小，很多重要群体都没有分析。因为样本不全，孟高棉群体 Y 染色体的多样性有可能被低估因而在分析中显得边缘化。如果加大东南亚的样本量，结果会更令人信服。

Kumar *et al.* compared the Y-chromosome diversity of three subfamilies of Austro-Asiatic populations (Mundari, Mon-Khmer and Khasi-Khmuic), and found that O-M95 is the most important haplogroup in all Austro-Asiatic populations. In addition, the highest diversity of STR in Mundari population indicates that O-M95 originated from Mundari population in India. Comparing the differences of Y-chromosome and mtDNA among these three populations, they suggested that Austro-Asiatic populations in Southeast Asia migrated from the present-day Mundari region through the Northeast India corridor. This conclusion has a great agreement with the linguistic evidence, which suggested that the linguistic ancestors of the Austro-Asiatic populations have originated in India and then migrated to Southeast Asia。

The process of the ratiocination in this paper is most logistical. The most likely origin region of a certain haplogroup can be identified on the basis of two characteristics: the highest frequency and the highest diversity. Among the Southeast Asian Austro-Asiatic populations, the maximum frequency of O-M95 is only 35% after excluding three populations with small sample size. However, the sample size for Mundari populations is generally large (~35–109, except

for two) and the average frequency of the haplogroup is 63% (excluding the three populations with disputed origin), which is significantly higher than that of the Southeast Asian Austro-Asiatics. Besides，the M-J network shown that the haplotype diversity among the Mundari populations is much higher than that of Southeast Asian Austro-Asiatic.

These conclusions destabilized those relevant opinions given in the previous study. Basu *et al* [1] found high frequency of Haplogroup K-M9 among the Mundari populations and then inferred that the Austro-Asiatic populations have migrated from Africa to India via central Asia. Kumar *et al* [2] analyzed the samples from India and southeast Asia for mitochondrial DNA 9bp deletion/insertion polymorphism，and suggested multiple origins of Austro-Asiatic tribes in India, and especially the Asian and non-Asian origins of the Mon-Khmer and the Mundari populations. The possible route for the expansion of Austro-Asiatic populations given in this study is : the present-day Mundari region (~65,000YBP)- Northeast India (~57,000YBP)- Southeast Asia (~8,000YBP) / Nicobar Islands (~17,000YBP). This suggested route helps the further studies on the structure of populations in Southeast Asia.

As an Austro-Asiatic-specific haplogroup, O-M95 can also be found in some populations of Daic, Tibeto-Burman, Hmong-Mien families. For example, the proportion of M95 is 38.3% in Zhuang [3], 47.5% in Naxi [3], and 20% in Mien [4]. This reminds us the complex relationships between linguistic and genetic structure. In some non- Austro-Asiatic populations, O-M95 is also the most dominant haplogroup. Furthermore, a new study suggested that O-M119 is more than 20,000 year old [5]. Comparing the geographical distribution of O-M119 and O-M95 from Southeast Asia to south China, it is clear that O-M95 approximately distributes in the west while O-M119 approximately distributes in the eastern coastal region. Considering that the westward spread of Daic populations happened in a very recent time, it can be concluded that O-M95 is much older than O-M119. This conclusion can also be applied to the study on influence of Austro-Tai's diffusion on Austro-Asiatic.

There are also some deficiencies in this paper: 1, the time to the most recent common ancestor (TMRCA) of O-M95 is much older than the estimated age of O/N in all others study. In the analysis, when using all samples, "we increased the MCMC cycle and found that the $N_{posterior}$ value and expansion time decrease and increase, respectively, with increase in MCMC cycle and do not stabilize even at 10^9 cycles." So the following analysis was basis on parts of samples which may result in misapprehension. 2, discussion about haplogroup N did not consistent with the most resent study. In this paper, "the modern man had probably migrated from Africa to Northeast Asia via Central Asia, where the haplogroups N-LLY22g and O-M175 might have originated." But in a resent study, Siiri Rootsi *et al* [6] suggests that the most frequent subclade N3, arose probably in the region of present-day China. We can also abstracted from the geographical distribution of N* and NO* that N expand northwards from Southeast Asia. 3, the authors suggest the Mon-Khmer people diffused from Southeast Asia through Thailand and coastal southern Burma to Andaman and Nicobar Islands. But the evidence given in this paper can not omit the possibility that the diffusion may be via Sumatra. 4，the linguistic classification in this paper is not consistent with the canonical classification [7]. In the conventional classification, Khasi-Khmuic belongs to the North Khmer branch of Mon-Khmer subfamily. However, this will exert no influence on the result.

We also concern with the sample coverage in the analysis of STR. Compared to the huge population of Mon-Khmer populations in Southeast Asia，the sample size in this paper was small and most of importance populations were

not included. The Y-chromosome diversity of Mon-Khmer population might be underestimated because of the incomplete sample. It will be more convictive for this study with a bigger sample size of Mon-Khmer populations in Southeast Asia.

参考文献

1. Basu A, Mukherjee N, Roy S, Sengupta S, Banerjee S, Chakroborty M, Dey B, Roy M, Roy B, Bhattacharyya NP, Roychoudhury S, Majumder PP (2003) Ethnic India: A genomic view, with special reference to peopling and structure. Genome Res 13:2277-2290.
2. Kumar V, Langsiteh BT, Biswas S, Babu JP, Rao TN, Thangaraj K, Reddy AG, Singh L, Reddy BM (2006) Asian and Non-Asian Origins of Mon-Khmer and Mundari Speaking Austro-Asiatic Populations of India. Am J Hum Biol 18:461-469.
3. 董永利,杨智丽,石宏,高路,鲁靖,程宝文,李开源,昝瑞光,肖春杰(2004) 云南18个民族Y染色体双等位基因单倍型频率的主成分分析. 遗传学报 31:1030-1036.
4. Su B, Xiao J, Underhill P, Deka R, Zhang W, Akey J, Huang W, Shen D, Lu D, Luo J, Chu J, Tan J, Shen P, Davis R, Cavalli-Sforza LL, Chakraborty R, Xiong M, Du R, Oefner P, Chen Z, Jin L(1999) Y chromosome Evidence for a northward migration of modern humans in East Asia during the last Ice Age. Am J Hum Genet 65: 1718-1724.
5. Li H (2005) Genetic Structure of Austro-Tai Populations. PhD Thesis of Human Biology, Fudan University.
6. Rootsi S, Zhivotovsky LA, Baldovic M, Kayser M, Kutuev IA, Khusainova R, Bermisheva MA, Gubina M, Fedorova SA, Ilumae AM, Khusnutdinova EK, Voevoda MI, Osipova LP, Stoneking M, Lin AA, Ferak V, Parik J, Kivisild T, Underhill PA, Villems R (2007) A counter-clockwise northern route of the Y-chromosome haplogroup N from Southeast Asia towards Europe. Eur J Hum Genet 15, 204–211.
7. Gordon RG Jr(ed.) (2005) Ethnologue: Languages of the World, Fifteenth edition. Dallas, Tex.: SIL International. Web version: http://www.ethnologue.com/

原文翻译 兰海 译

Y 染色体证据显示南亚语系人群有共同的父系起源

摘要 **背景** 南亚语系被认为是印度所有语系中最古老的，它在东南亚也有大量分布。但是，印度的南亚语系各语族间以及印度，东南亚两地的南亚语系人群之间可能的基因联系至今都没有被研究清楚。因此，为了追溯印度境内的南亚语系人群的起源以及历史上的扩张，我们分析了来自印度 25 个人群 1222 个样本的 Y 染色体 SNP 和 STR 数据。其中包含南亚语系部落的三大支系，即蒙达、卡西－克木、孟高棉，以及已经公开发表的亚洲和大洋洲的 214 个人群的数据。 **结果** 我们的结果显示，不但印度境内的南亚语系各群体之间，而且他们与东南亚的南亚语系人群之间都有明显的父系遗传联系。但是，基于 mtDNA 的母系研究没有得到类似证据。我们的结果也指出，单倍群 O-M95 约于~65,000 年前（95％CI: 25,442-132,230）起源于印度的南亚语系人群，然后通过印度东北走廊向东南亚扩散。而后，东南亚操孟高棉语的人群在相当晚近的一个时间到达了安达曼和尼科巴群岛。 **结论** 我们的结论与语言学证据一致。语言学证据认为南亚语系的祖先起源于印度然后向东南亚迁徙。

背景

印度次大陆现在居住着四个主要的语群：南亚语系，德拉维达语系，印欧语系和藏缅语族。他们在不同的时间进入印度。由于观察到其中南亚语系的名词分化程度最高[1]以及其他的一些语言学特征（稍后详细讨论），所以南亚语系被认为是四个语系[1,2]中最古老的。南亚语系在印度境内包括三个语族[3]： 1、蒙达语族，使用者为中部和东部印度焦达讷格布尔高原上的一些部族；2、孟高棉语族，使用者为安达曼和尼科巴群岛上的尼科巴人和匈蓬人；3、卡西－克木语族（早先被认为孟高棉语族的一部分），使用者为印度东北的卡西部落（图 1 ）。印度的卡西语支和克木语支人群有东亚人群的体质特征，而蒙达语族的一些体质特征与德拉维达语系相似。另外，除了蒙达语族使用者局限于印度外，其他两个语族的使用者包含大量的东南亚人群。尽管印度次大陆被认为是前往东南亚的迁徙重要走廊，印度的南亚语系各语族间的遗传联系，印度-东南亚两地的南亚语系人群之间的可能遗传联系却至今都没有被深入研究。

基于语言学，考古学和经典遗传标记的研究认为，南亚语系进入印度可能经由两种

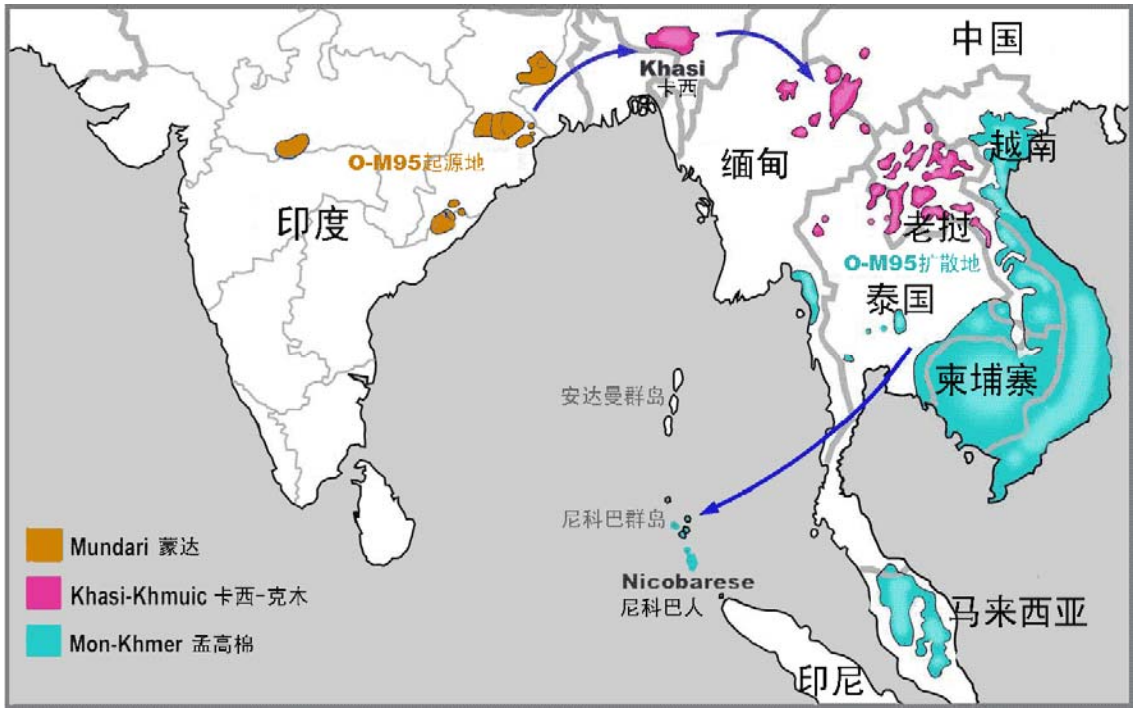


图 1 图中显示了现在南亚语系各群体的分布（修改自 van Driem[2]）以及印度各南亚语系支系的迁徙。

路线，引述如下：第一条迁徙路线是从非洲经由中亚进入印度，而第二条路线是从非洲到东北亚（译者按：可能是原文笔误，本文中的东北亚都是指东亚）然后进入印度。Basu 等[5]发现蒙达人有很高频率的 K-M9，因此推断这些南亚语系人群是从非洲经由中亚迁入印度的。但后来这个观点受到怀疑，因为这个单倍群在亚洲也普遍存在并在东亚有稳定的分布频率。另外，通过研究 mtDNA 的 9bp 缺失/插入的多态性， Thangaraj 等[6]以及 Prasad 等[7]报告了尼科巴人只有东亚特异的 mtDNA 单倍群，而 Roychoudury 等[8]及 Metspalu 等[9]在蒙达人中仅发现了印度特异的 mtDNA 单倍群。但以上推论仅是基于少量的基因证据，这些研究也仅包含很少的南亚语系人群（最多为 3 个）。尽管 Kumar 等[10]分析了大量的南亚语系人群，认为蒙达人、卡西－克木人、孟高棉人的起源和历史上的迁徙是不同的，但是该分析也是仅仅涉及 mtDNA 的 9bp 缺失/插入的多态性和特征。

我们对印度所有的南亚语系人群进行了采样，覆盖了所有地区和其内部支系（表一和图 S1[参见附件 1]）。其中包含东北印度的南亚语系卡西语族的基因数据－－东北印度被认为是人类迁往东南亚的重要走廊。我们分析了南亚语系人群的 Y 染色体 SNP 和 Y-STRs 的数据和先前报道的 214 个相关人群的数据，以追溯印度境内南亚语系人群的起源和历史上的迁徙过程。基于这些证据，我们认为，单倍群 O-M95 起源于印度的南亚语系人群，极可能是蒙达人，随后它的部分祖先携带这个单倍群迁徙到了东南亚。

结果

Y 染色体单倍群的分布及频率

分析了 16 个 Y-STRs 后，各人群的 Y 单倍群频率和多样性以及单倍群的多样性的结果列在表 2 中。总的说来，单倍群的多样性很高(98.87%)，从潘多人的 95.26%到卡西人、加若人、帕哈日人、那格西亚人和比日加人的 100%之间变动。本文检测了可能出现的 13 个双等位标记定义的单倍群，在这些人群中发现了 9 个单倍群（如图 2）。O-M95 的平均频率最高(52 %)，其次是 H-M69(26%)。南亚语系的三个语族平均，东北印度的蒙达人为 55%，卡西克木人为 41%，尼科巴人的 11 个样本全部属于 O-M95。为了了解未区分的 O-M95 样本是否包含下游支系，我们也检测了 O-M95 的下游单倍群 M88 标记，但在 O-M95 样本中都没有发现。

表 1　25 个被研究人群的地理分布、语言特征以及样本量

群体名称	语言系属	采样地点	样本量
Santhal	Munda, 北 Munda, Kherwari, (AA)	Jharkhand 邦 Jamshedpur 区; 西孟加拉邦 Purulia 区	109
Bhumij	Munda, 北 Munda, Kherwari, (AA)	Jharkhand 邦 Jamshedpur 区; 西孟加拉邦 Purulia 区	89
Mudi	Munda, 北 Munda, Kherwari, (AA)	Jharkhand 邦 Jamshedpur 区; 西孟加拉邦 Purulia 区和 Midnapore 区	37
Mahali	Munda, 北 Munda, Kherwari, (AA)	Jharkhand 邦 Jamshedpur 区; 西孟加拉邦 Purulia 区和 Midnapore 区	25
Asur	Munda, 北 Munda, Kherwari, (AA)	Jharkhand 邦 Lohardagga 区和 Gumla 区	55
Birjia	Munda, 北 Munda, Kherwari, (AA)	Jharkhand 邦 Lohardagga 区和 Gumla 区	24
Birhor	Munda, 北 Munda, Kherwari, (AA)	Jharkhand 邦 Lohardagga 区 Gumla 区和 Simdega 区; 西孟加拉邦 Purulia 区; Orissa 邦 Mayurbhanj 区	38
Munda	Munda, 北 Munda, Kherwari, (AA)	Jharkhand 邦 Lohardagga 区 Gumla 区和 Simdega 区; 西孟加拉邦 Purulia 区; Orissa 邦 Mayurbhanj 区	53
Ho	Munda, 北 Munda, Kherwari, (AA)	Orissa 邦 Mayurbhanj 区	79
Korwa	Munda, 北 Munda, Kherwari, (AA)	Jharkhand 邦 Simdega 区; Chattisgarh 邦 Surguja 区	42
Korku	Munda, 北 Munda, Korku, (AA)	Maharashtra 邦 Amravati 区	59
Juang	Munda, 南 Munda, Kharia-Juang, (AA)	Orissa 邦 Keonjhar 区	49
Kharia	Munda, 南 Munda, Kharia-Juang, (AA)	Jharkhand 邦 Simdega 区; 西孟加拉邦 Purulia 区; Orissa 邦 Mayurbhanj 区	36
Savar	Munda, 南 Munda, Sora-Juray-Gorum, (AA)	Jharkhand 邦 Jamshedpur 区; 西孟加拉邦 Purulia 区; Orissa 邦 Mayurbhanj 区	47
Lodha	Munda, 南 Munda, Sora-Juray-Gorum, (AA)	西孟加拉邦 Midnapore 区	47
Oraon	北支, Kurux, (Dra)	Jharkhand 邦 Lohardagga 区和 Gumla 区; Chattisgarh 邦 Surguja 区	91
Nagesia	北支, Kurux, (Dra)	Chattisgarh 邦 Surguja 区	14
Paharia	北支, Malto, (Dra)	Jharkhand 邦 Jamshedpur 区; 西孟加拉邦 Purulia 区;	11
Pando	中央带, Hindi (IE)	Chattisgarh 邦 Surguja 区	23
Kanwar	中央带, Rajasthani (IE)	Chattisgarh 邦 Surguja 区	41
Bhuiyan	东部带, Oriya (IE)	Orissa 邦 Keonjhar 区	81
Bathudi	东部带, Oriya (IE)	Orissa 邦 Mayurbhanj 区	36
Nicobarese	Mon-Khmer, Nico-Monic, Nicobar, (AA)	Andaman &Nicobar 领 Nicobar 岛	11
Khasi	Khasi-Khmuic, Khasian, (AA)	Meghalaya 邦 Ri-Bhoi 区 West Khasi 山区, East Khasi 山区, Jaintia 山区	92
Garo	Tibeto-Burman (TB)	Meghalaya 邦 South Garo 山	33

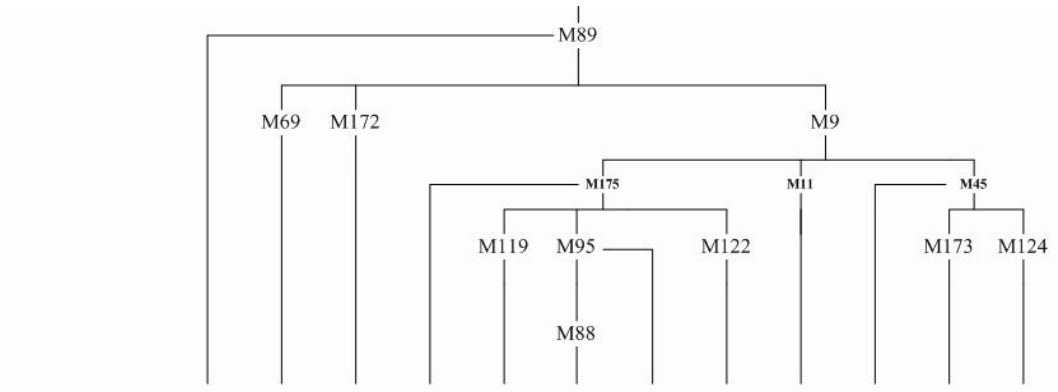
“AA”代表南亚语系, “Dra”代表德拉维达语系”, “TB”代表藏缅语族, “IE”代表印欧语系。

除了卡西语支人群(29%)和科尔库人的 1 例，在印度其他南亚语系人群中没有发现 O-M122。另外，加若人中，O-M122 最普遍(55%)，其次是 O-M95(8%)。由于在梅加拉亚，南亚语系的卡西人和藏缅语族的加若人居住地接近而且有已知的婚姻交流，所以我们检测了来自加若人和卡西人所有的 O-M122 样本，以确定卡西人的 O-M122 并非来自与加若人的基因交流。结果在我们用到的 8 个单倍群中，发现了三种: O-M133*, O-M134*, O-M122*。O-M134*在卡西人(56%)和加若人(67%)中的频率最高，其次是 O-M133*。卡西人 27 个样本中的 3 个以及加若人 18 个样本中的 3 个都属于未定义的 O-M122*。因此 O-M122 的支系在卡西人和加若人的分布是相似的(X^2=1.597；p=0.45)。

分子方差分析和 Y-STR 分析

基于 Y-STRs(表 3)的分子方差分析(AMOVA)显示，印度和东南亚的南亚语系人群之间的分化程度很大（F_{ST}=0.203）。蒙达人和东南亚的南亚语系人群之间的 F_{ST} 值比这个值大 3%，显示他们之间有更多的差异性。但是卡西人与东南亚的南亚语系人群之间的 F_{ST} 值(0.045)很小，相对于卡西人和蒙达人之间的 F_{ST} 值(0.099)，证明前两者有更多的基因相似性，。因全部属于 O-M95,尼科巴人的样本没有包含到这个分析来。基于 Y-STRs 的 AMOVA 显示蒙达人与卡西人之间有很高的 F_{ST} 值(0.175)，但低于卡西人与尼科巴人的值 F_{ST}(0.289)以及蒙达人与尼科巴人之间的 F_{ST} 值(0.442)。

对 O-M95 样本分析 16 个 Y-STR 标记得



群体	N	F-M89*	H-M69	J-M172	O-M175*	O-M119	O-M88	O-M95*	O-M122	L-M11	P-M45*	R1-M173	R-M124
蒙达人	789	3.30	25.35	4.44	0			54.88	0.13		1.52	5.45	4.94
卡西人	92	10.87	6.52	0	2.17			41.30	29.35		4.35	5.43	0
尼科巴人	11	0	0	0	0			100	0		0	0	0
南亚语系群体	892	4.04	23.00	3.92	0.22			54.04	3.14		1.79	5.38	4.37
过渡群体	297	3.70	35.02	4.71	0			51.52	0		0.67	3.37	1.01
加若人	33	0.09	0.09	0	3.03			18.18	54.55		0	6.06	0
总计	1222	4.09	25.61	4.01	0.25			52.46	3.76		1.47	4.91	3.44

图 2　由双等位标记定义的单倍群的有根最大简约树以及不同群体中的单倍群频率。

到的网络结构图（图 4）显示了两个差别较大的分枝，其一代表了蒙达人群，另一个代表卡西人和尼科巴人。尽管尼科巴人分枝在卡西人的分枝之下，但这个分枝包含两个不同的支系，表现了独立的特性。正如我们期望得到的，因为相当程度的混合，网络结构图显示加若人的样本是卡西人分枝的一部分，而不是独立的分枝。另外，O-M122 下游单倍群的各网络结构图分别都说明了以下事实：不管在 O-M133*（图 S2,附件 1），或 O-M134*下，加若人和卡西人的样本都没有形成不同的分支，证实了他们之间不同程度的基因交流。

最晚共祖时间（TMRCA）

因为 O-M95 被发现是最普遍的单倍群，我们通过 BATWING 程序分析 16 个 Y-STR，计算了这个单倍群的 TMRCA。我们选择了以下人口地理模式：人口规模在一段时间内保持不变并产生分化，然后发生指数增长。尽管有其他的两种模式，我们选择的模式被认为是最适合于人类人口规模的。这个模式假设人口规模保持不变，直到他们开始定居或半定居－－可能直到发明了农业，然后人口规发生了几何级数的增长（详见**方法**中的程序部分）。经 10^6 次 MCMC 循环得到的所有南亚语系群体的 TMRCA 的中值是～68,000 年前（95% C.I. 25,442~132,230）。 蒙达人的 TMRCA（～66,000YBP）和卡西人的（～57,000 YBP）相似，而尼科巴人的 TMRCA 则很晚（～17,000 YBP）。此外，各蒙达群体和过渡群体的 TMRCA 在～70,000 年前到～30,000 年前之间波动（结果未发表），平均值很大（～48,000 YBP），暗示单倍群 O-M95 起源得很早，可能在旧石器时代时期。

单倍群等频率分布图

所有单倍群的等频率分布图都已绘制，但仅两个相关的 O-M122 和 O-M95 被列在图 5 中。我们的数据以及另外 214 个相关人群的数据显示，O-M95 在东南亚人群普遍存在，而在印度则局限于南亚语系人群存在的地方。这一点强烈暗示印度的南亚语系人群与东南亚的南亚语系人群不但有语言学的关系，更有基因联系。现在南亚语系人群与 O-M95 的分布是显著相关的（表 5 及图 5）。例如，O-M95 在印度西北和中亚的频率分别为 3.4%，0.1%，在那里没有南亚语系人群，而它的频率在东南亚的南亚语系人群（38%）和相邻的非南亚语系人群（14.7%）的都很高。此外，南亚语系人群的这个频率明显比非南亚语系人群的要高（X^2=22.77; p<0.001）。O-M95 的频率从印度到东南亚有下降的趋

表 2　单倍群频率以及 Y 染色体多样性

群体名称	样本量	F-M89*	H-M69	J-M172	O-M175*	O-M122	O-M95	P-M45*	R-M124	R-M173	Haplogroup Diversity±SE	Y-STR Haplotype Diversity±SE
Santhal	109	0.03	0.39	0.03	0	0	0.47	0.02	0.02	0.05	0.627±0.028	0.9913±0.0039
Bhumij	89	0.02	0.27	0	0	0	0.63	0	0.04	0.03	0.534±0.046	0.9950±0.0031
Mudi	37	0.03	0.43	0.03	0	0	0.43	0.03	0.03	0.03	0.640±0.049	0.9910±0.0090
Mahali	25	0.04	0.4	0.08	0	0	0.12	0	0.12	0.24	0.777±0.057	0.9855±0.0179
Asur	55	0.04	0.22	0.09	0	0	0.64	0	0	0.02	0.547±0.064	0.9970±0.0045
Birjia	24	0	0	0	0	0	0.96	0.04	0	0	0.083±0.075	1.0000±0.0120
Birhor	38	0	0.24	0	0	0	0.71	0	0	0.05	0.448±0.077	0.9640±0.0189
Munda	53	0.02	0.25	0	0	0	0.45	0.08	0.08	0.13	0.719±0.044	0.9898±0.0062
Ho	79	0	0.25	0.01	0	0	0.66	0.04	0.04	0	0.506±0.051	0.9949±0.0034
Korwa	42	0	0.33	0.02	0	0	0.6	0	0.05	0	0.545±0.054	0.9906±0.0090
Korku[¶]	59	0.07	0.08	0	0	0.02	0.81	0.02	0	0	0.331±0.076	0.9989±0.0053
Juang	49	0.02	0	0	0	0	0.98	0	0	0	0.041±0.039	0.9745±0.0114
Kharia	36	0.17	0.33	0.03	0	0	0.39	0	0	0.08	0.722±0.041	0.9982±0.0077
Savar	47	0.09	0.32	0.13	0	0	0.15	0	0	0.32	0.767±0.030	0.9677±0.0123
Lodha	47	0.02	0.15	0.32	0	0	0.09	0	0.43	0	0.702±0.039	0.9815±0.0115
Oraon	91	0.02	0.57	0	0	0	0.32	0.01	0.03	0.04	0.575±0.039	0.9980±0.0022
Nagesia	14	0	0.36	0.07	0	0	0.57	0	0	0	0.582±0.092	1.0000±0.0302
Paharia	11	0	0.64	0	0	0	0.36	0	0	0	0.509±0.101	1.0000±0.1265
Pando	23	0.04	0.22	0.09	0	0	0.65	0	0	0	0.542±0.101	0.9526±0.0335
Kanwar	41	0.17	0.29	0.15	0	0	0.39	0	0	0	0.729±0.034	0.9892±0.0102
Bhuiyan	81	0.01	0.11	0.04	0	0	0.84	0	0	0	0.285±0.062	0.9959±0.0031
Bathudi	36	0	0.39	0.06	0	0	0.36	0.03	0	0.17	0.706±0.041	0.9964±0.0082
Nicobarese[#]	11	0	0	0	0	0	1	0	0	0	0.000±0.000	0.9643±0.0772
Khasi	92	0.11	0.07	0	0.02	0.29	0.41	0.04	0	0.05	0.730±0.030	1.0000±0.0031
Garo	33	0.09	0.09	0	0.03	0.55	0.18	0	0	0.06	0.669±0.077	1.0000±0.0171
Total	1222	0.04	0.26	0.04	0	0.04	0.52	0.01	0.03	0.05	0.533[§]	0.9887[§]

[¶]来自 Dr. V. R. Rao 未发表数据,;[#]SNP 数据来自 Thangaraj 等[40]；[§]为平均值。

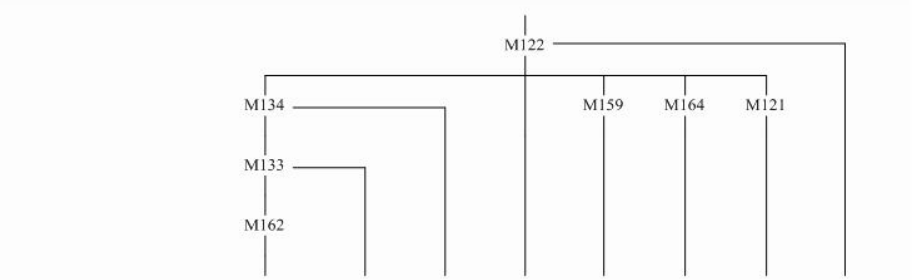
势。但图中所示并不太明显，因为 45 个东南亚人群中，有 7 个人群的 O-M95 的频率从 50%变化到 70%。但是，这 7 个人群中有 6 个人群的样本数小于 20，有的更少。无论如何，相比于东南亚人群，不管是东南亚的南亚语系人群（38%；X^2=68.89；p<0.001），还是东南亚的非南亚语系人群（14.7%；X^2=330.68；p<0.001），印度的南亚语系人群中，O-M95 的平均频率要高得多（54%）。另外，正如图 5 描绘的陡峭边界线所示，单倍群 O-M122 在印度的分布仅局限于东北印度，而它在东北亚和东南亚非常普遍（表 5）。单倍群 H-M69 的分布仅局限于印度次大陆的边界，因此，强烈暗示它起源于印度次大陆。

讨论

南亚语系群体的共同基因遗传

蒙达人群中，O-M95 的频率最高，22 个人群中仅 3 个人群不同于这个总体趋势。这三个分别是 Lodha，Savara 和 Mahali，可能是因为他们的起源有争议。这个单倍群在卡西人和尼科巴人中的频率也相当高。因此在最近的一项研究中，尼科巴群岛匈蓬人的 12 个样本全部属于 O-M95，与当地同语系的尼科巴人一样。这说明，蒙达人、卡西－克木人、孟高棉人之间不但有语言学的关系，更有基因联系，很可能有单一且广泛的父系基因遗传。这个单倍群在印度的其他语系中不存在或频率很低，显示印度的南亚语系人群与他们有不同的基因特征。在另外一个方面，东南亚的南亚语系人群有高频的 O-M95（平均 38%），与他们相邻的人群中也有一定比例的 O-M95(平均 14.7%)。但这个单倍群在印度西北和中亚几乎不存在。所以，O-M95 在印度和东南亚的南亚语系人群中的存在和在亚洲其他人群的缺失，显示了这个语系的人群有共同的父系遗传。

印度的藏缅语族人群完全没有 O-M95，显示这些人群迁入印度时没有伴随单倍群 O-M95。因此，梅加拉亚的加若部落中 O-M95



群体	N (O-M122)	O-M162	O-M133	O-M134	O-M7	O-M159	O-M164	O-M121	O-M122
卡西人	27		33.33	55.56					11.11
加若	18		16.67	66.67					16.67
总计	45		26.67	60.00					13.33

图 3　O-M122 下游单倍群的最简系统树以及它们在卡西和加若样本中的频率

的存在，应该是来自相邻卡西人高比例的基因交流－－两个部落之间的婚姻传统提高了这个比例。在南亚语系卡西人中 O-M122 相当高频的存在，也可以认为是来自与加若人的基因交流，这与他们之间 O-M122 下游支系有相似的频率和组成是一致的(X^2=1.597；p＝0.45)。现在，在这些 O-M122 下游单倍群的 M-J 网络结构图中，区分不出独立的 Y-STRs 分支(图 S2，附件 1)。比较数据得到，靠近东北印度的东南亚的南亚语系人群[33]都包含 O-M133*或 O-M134*（63%），而靠近中国东南、柬埔寨的南亚语系人群[24,33]则包含 O-M159(65%)。以上事实说明，不同地区的南亚语系人群含有不同的 O-M122 下游单倍群，而这些是相邻的非南亚语系人群的特征单倍群，明显是来自外部的混合。因此，O-M133*/O-M134*在南亚语系的卡西人和东北印其他的藏缅语族（包括加若人）中的存在说明，卡西人中的 O-M122 可能来自加若人。尽管前述的分析认为印度的南亚语系人群分享了相同的基因联系，但各语族的比较分析也显示，他们分离得很早，现在已经深度分化，正如 AMOVA（表 3），M-J 网络结构图（图 4）和 TMRCA 的结果（表 4）所示的那样。

表 3　分子方差分析 AMOVA

群体	F_{ST}	
	Y-SNP	Y-STR
蒙达人与卡西人	0.099	0.175
蒙达人与尼科巴人	—	0.442
卡西人与尼科巴人	—	0.289
蒙达人与 S.E. AA	0.227	—
卡西人与 S.E. AA	0.045	—
印度 AA 与 S.E. AA	0.203	—

S.E.AA=东南亚的南亚语系人群。注：P 值都小于 0.05。因尼科巴人的样本全部属于一种单倍群，故不计算 AMOVA。没有得到 S.E.AA 的 Y-STR 数据。

单倍群 O-M95 的起源和南亚语系的扩张

既然 O-M95 在南亚语系人群中有相当高的频率，那么，O-M95 很可能就起源在他们之间。但问题是，它起源于印度还是东南亚？一个单倍群最有可能起源的地方可以用两个特征来确定: 高的频率和高的多样性。8 个东南亚的南亚语系人群中，排除了样本数过小的 3 个人群后，O-M95 的最高频率为 35%。而另一方面，蒙达语族人群的样本量都很大(35～109，除了 3 个)，O-M95 的频率从 39%～89%变化，平均为 63%（排除了 3 个起源有争议的人群）。这明显比东南亚的南亚语系人群要高。而且，蒙达语族人群样本的多样性高达 99%。考虑这个以及该单倍群在印度的其他地方以及西亚和中亚低频存在的事实，我们可以谨慎地得到结论：如 AMOVA 所得到的，O-M95 约于 65,000 年前（95%CI: 25,442-132,230）起源于蒙达语族人群。因此，可能在更新世，现今蒙达语族的祖先在 O-M95 产生之前来到印度。这与考古学证据是一致的，后者说明人类在新石器时代早期就已经居住在印度次大陆。

但 Kayser 等[16]认为 O-M95 起源自东南亚，然后那里的南亚语系人群迁徙到了印度。基于印度境内非蒙达语族的南亚语系人群

（尼科巴人和卡西人）有东亚特异的 mtDNA, Kayser 等[16]和 Thangaraj 等[6]认为他们来自东南亚。但是，在蒙达部落与东南亚的人群间没有相似的母系基因联系[8~10]。我们分析了 1147 个样本，包含了大部分蒙达部落以及过渡群体，完全没有发现东亚特异的 mtDNA 单倍群[Kumar V, Reddy BM and Langstieh BT 等未发表的数据]，显示印度境内南亚语系的三个语族有不同的母系遗传历史，这一点与早期共同的父系遗传是不同的。我们是怎样得到这个结论的？既然南亚语系各语族有共同的单倍群 O-M95 而 mtDNA 没有发现类似情况，那么，一个男性占主导的前往东南亚的迁徙就是极有可能的。但更重要的是，O-M122 被认为是东南亚人群的特征单倍群，在蒙达部落中它完全不存在。而任何分析关于始自东南亚的迁徙都是要以 O-M122 为基础的[18,20,33]。

单倍群计算得到 O-M122 的年代在 15,000~60,000 年前[18]之间，而东南亚的 O-M95 计算结果为～8,000 年前[16]。如果印度的南亚语系人群确实来自东南亚，那么他们应该包含有 O-M122，而且印度的南亚语系人群中，O-M95 的 TMRCA 计算结果应该比现在得到的～65,000 年前要低(表 4)。虽然说这个值的置信区间很大（25,442～132,230），应该谨慎对待。无论如何，蒙达人这个值的下限依然比 Kayser 等[16]得到的上限要大，而且没有重叠。因为 Kayser 等[16]应用了 16 个位点中突变速率相当快的 7 个，我们用这 7 个位点和与 Kayser 相同的突变速率重新计算了我们的数据，发现得到的 TMRCA 结果相似（～65,000 年前）。这个结果显示，本文的计算并不是由于大数目的位点和低的突变速率造成的人为后果。另外，蒙达人被认为是传统的食物采集狩猎者，现在他们的居住地不适合于耕作，这可能反映了他们传统的生存模式。因此，尼科巴人在新石器时代农业人口的扩张时期进行了迁徙[16]，这对蒙达人来说是不可能的。基于这些证据，我们认为，现今蒙达人的祖先从印度迁徙到了东南亚，而不是相反。这个情形与以下事实是一致的：蒙达语族是南亚语系中语法和语音都最复杂的语族，它比其他的语族更接近于原始南亚语。因此这个语系的祖先起源于印度。更多的分析显示，O-M95 最有可能起源于现今蒙达人群的祖先，而后携带着它迁往东南亚。

表 4 计算得到的各语族的 O-M95 的 TMRCA

群体	TMRCA (Years)	95%置信区间（Years）	
		下限	上限
蒙达人	65,730	25,442	132,320
卡西人	57,252	27,664	92,201
尼科巴人	16,578	4,565	51,377
南亚语系人群	68,098	25,992	146,883

AMOVA 结果（表 3），M-J 网络结构图（图 4）和 O-M95 的 TMRCA（表 4）显示，蒙达人和其他的南亚语系人群很早就分开了。因为分离得很早，我们希望这些群体的 O-M95 样本至少形成单独的支系。但是，没有发现一例 O-M88（图 2）。直到现在，O-M88 仅在柬埔寨和老挝发现一例[40]，暗示这个支系的频率很低并局限那个地区。因此如果有下游支系存在，它应该由其他未知的双等位标记点定义。因为卡西人有相当高频的 O-M122（29%），且卡西－克木语族各语支的分布局限在缅甸和泰国北部，我们推断卡西语族人群是从印度迁往东南亚的。然而，卡西人中 O-M122 的存在是来自加若人的基因交流，说明这个人群最早并不含有这个单倍群。此外，印度的 mtDNA 单倍群贡献了卡西人 30%的 mtDNA 成分（Reddy 等未发表的数据），而这些单倍群在他们的藏缅语族邻居中完全缺失[8,41]。因此，卡西人中的东亚特异的 mtDNA 可以认为是来自相邻的加若人和其他的藏缅语族群体，后两者事实上只含有东亚特异的 mtDNA。这一点加强了以下观点：蒙达人和卡西人分离得很早，后者可能经由东北印度走廊迁往东南亚，正如地图所示的那样(图 1)。

由基于 Y-STRs 的 AMOVA 结果（表 3）和 M-J 网络结构图（图 4）显示，尼科巴人与蒙达人以及卡西部落都很不一样。尼科巴人只含有东亚特异的 mtDNA[6,7]和父系的 O-M95（图 2 及表 2）。我们也对属于孟高棉语族的匈蓬人进行了基于 16 个 STR 的 AMOVA 分析。结果显示，匈蓬人和尼科巴

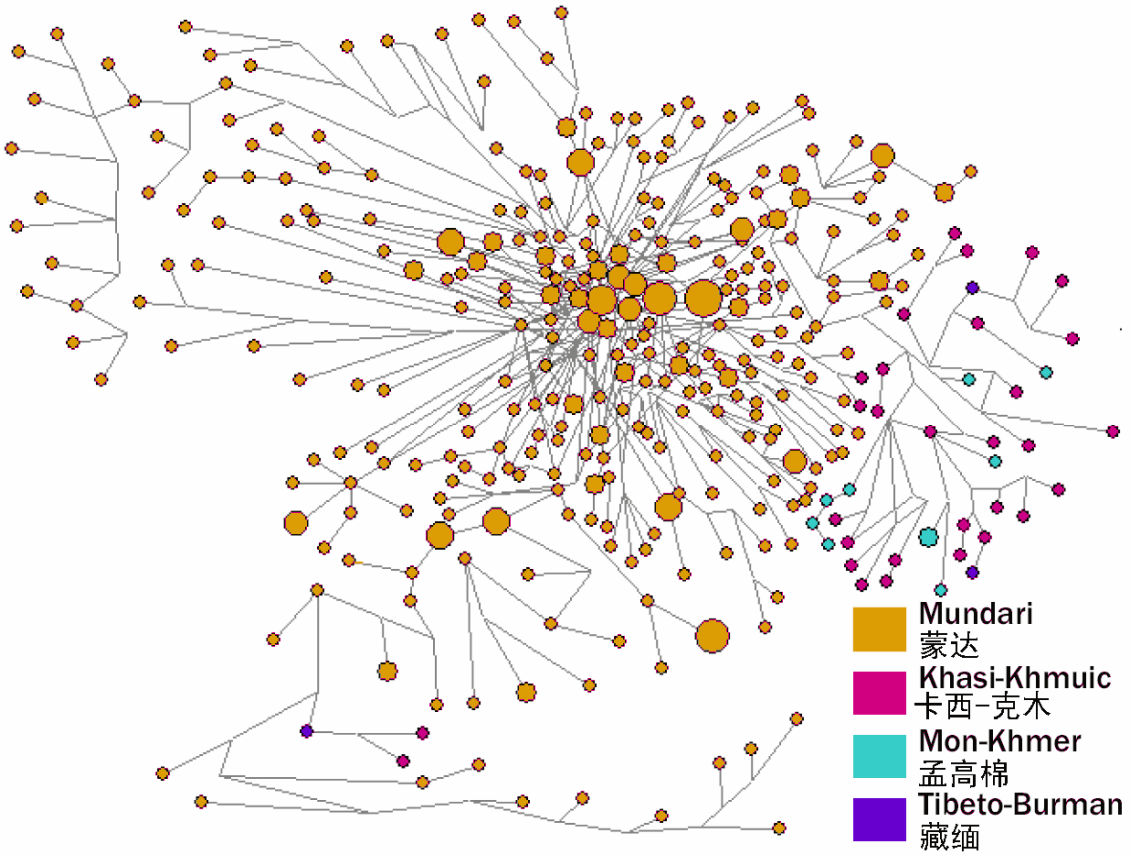


图 4 单倍群 O-M95 的 Y-STR 的 M-J 网络结构图 剔除无 STR 数据的样本后分析了余下的 564 个样本。圆圈大小正比于单倍型的频率。微卫星位点的突变用黑线表示。

人一样，与蒙达人以及卡西人差异明显（FST 分别为 0.402，0.476）。计算得到，尼科巴人的 TMRCA 为 17,000 年前，匈蓬人的 TMRCA 为 19,000 年前。孟高棉人群局限于泰国，越南和柬埔寨的南部，暗示尼科巴人在新石器时代农业人口的扩张时期迁徙到了印度的尼科巴群岛。尼科巴人中 O-M122 的完全缺失可能是因为奠基者效应加以及后来的基因漂移，尽管更确定的意见因为较少的样本数而未能得到。

南亚语系人群进入印度的两种可能路线

Kumar 和 Reddy[4]认为，经由东北印度走廊从东南亚进入，或经由西印度走廊从中亚进入，都有可能是印度的南亚语系部落从非洲迁入印度的路径。单倍群 O-M175 的姊妹单倍群，也就是 N-LLY22g 仅局限于东北亚，包括俄罗斯和西伯利亚，而在中亚，南亚，东南亚不存在或频率低到可以忽略。同样，单倍群 O-M175 以及它的下游单倍群在南亚（除了南亚语系）和中亚不存在或是频率极低，而它存在于整个东亚。我们面对两个可选的图景：1、现代人可能从非洲迁往中亚，单倍群 N-LLY22g 和 O-M175 在那里起源[43,44]。结果，携带 O-M175 的人群迁往印度，在那里诞生了 O-M95，然后扩散到东南亚。只是，由于 100%的蒙达人和 30%的卡西样本属于印度特异的 mtDNA，而所有东南亚的南亚语系人群只有东亚特异的 mtDNA[45]，所以这个迁徙可能是男性占主导的。2、注意到这一点，南亚语系人群的祖先从中亚经由西北印度走廊迁往印度的可能性也不能完全忽视，它可以说明现在印度和东南亚的南亚语系人群中的 mtDNA 差异。尽管在中亚没有报道过 N-LLY22g 和 O-M175 的存在，但很多研究观察到了相当频率的 K-M9*[15,24,26],有可能这些样本属于由连接了 N 和 O 的双等位标记 M214 定义的单倍群[46]。但是，这个标记在中亚人群没有被检测。一部分人群迁往东北亚，并在那里诞生了 N, 还有一部分人群通过西北印度走廊迁入印度，并在那里诞生了 O-M175。后来，O-M95 作为主要父系连同印度特异的 mtDNA

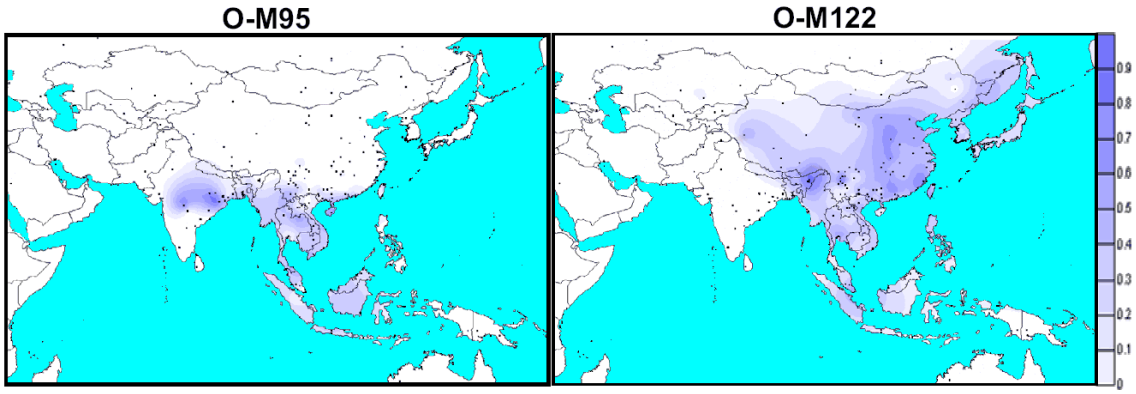


图 5 等频率分布图 描绘了亚洲和大洋洲单倍群 O-M122 和 O-M95 的频率[14~16,18~30]。对于 O-M95，尼科巴人的数据被排除。黑点表示采样的人群和地点。

表 5 中亚和东亚共有的单倍群的频率分布

	人群数*		不同单倍群的平均比例		
地区	O-M95/122	N-LLY22g	O-M95	O-M122	N-LLY22g
中亚	21	35	0.1 (0 – 1.6)	2.7 (0 – 12.1)	0.9 (0 – 9.5)
东北亚	83	25	3.4 (0 – 50)	34.1 (0 – 85.1)	9.2 (0 – 42.9)
东南亚非南亚语系人群	37	5	14.7 (0 – 75)	30.5 (0 – 100)	0.5 (0 – 2.4)
东南亚南亚语系人群	8	5	38.0 (3 – 68)	34.3 (0 – 70.2)	0

*单倍群频率来自图 5 的参考文献。

一起发展。而后，这些人群以男性为主导且快速地由印度东北迁往东南亚，导致东南亚的南亚语系没有印度特异的 mtDNA 而仅有东亚的成分。东亚发掘的解剖学意义的现代人类化石的年代没有比 40,000 年更早的[18,47,48]，这暗示南亚语系人群最早的前往东南亚的迁徙可能在 40,000 年前或稍晚。因此，蒙达人看来是最早的南亚语系群体之一，卡西－克木人和孟高棉人很早就从这个群体分离后，迁往东北亚并定居在那里。而很久以后，另外一次迁徙浪潮，由现在孟高棉人群的分布可以推断，操孟高棉语的人群是从泰国和缅甸南部海岸迁往安达曼和尼科巴群岛的。

结论

总之，我们总结如下：因为很高的频率和高的多样性，单倍群 O-M95 就起源在印度的南亚语系人群中，极可能是蒙达人中，而不是之前认为的东南亚。由于计算得到的 TMRCA 很大，我们认为蒙达人群是印度次大陆最早的定居者。很有可能，这些人群从中亚经由西北印度走廊迁往印度，然后迁徙到东南亚。当然，我们需要更多的 Y 染色体和 mtDNA 的数据以便得到更确定的结论。

方法

样本 我们采集了 1222 份健康无关个体的静脉血液样本，属于 25 个部落群体，其中 17 个是南亚语系人群而 7 个是过渡人群。这 7 个人群被认为在地理上和历史上都与前者有密切联系，并且原来是操南亚语的。我们用以下方式采集了蒙达人，卡西－克木人以及孟高棉人的血样，以便完整地反映整个南亚语系部落人群的基因种类：样本来自不同南亚语系人群不同的方言种类以及同一部落不同的地理单元——因为某些人群的分布很广。我们的样本包含了梅加拉亚省操藏缅语的加若人。加若人与卡西人居住地接近，并有已知的婚姻交流。各省采样地点如图 S1 所示(见附件 1)。人群名称，样本数量以及详细地点列在表 1 中。采样之前，献血者知情并同意。

基因分析 按照 Sambrook 的方法提取样本 DNA[49]。分型了用以检测亚洲人变异的以下 20 个 Y 染色体单核苷酸多态位点(SNPs)：M89, M69, M172, M9, M11, M175, M95, M88,

M122, M119, M45, M173, M124, M134, M159, M164, M7, M121, M133 和 M162。这些标记点在别处有描述[44]。大部分样本检测了所有的双等位基因标记点，以便内部检查定型的可靠性以及检测重复突变。我们遵从 YCC 的命名系统[50]。

我们也对以下的 20 个短序列重复(STRs)基因座作了分型：DYS19, DYS385a, DYS385b DYS388, DYS389Ⅰ, DYS389b, DYS390, DYS391, DYS392, DYS393, DYS426, DYS437, DYS438, DYS439, DYS447, DYS448, DYS460,H4，YCAⅡa，YCAⅡb。Y-STRs 经多次 PCR 扩增[51]，并在 ABI3730 分析仪上分析。GENOTYPER 程序被用来分析片段长度。等位片段的长度按照 Butler 等[51]的方法转换为重复单元。

尽管 Y 染色体变异很多，但由于非重组区双等位标记的低频率的平行突变和回复突变，它在重建和鉴别可靠的父系遗传时非常有用，并可以追溯到数千年前。此外，Y 染色体的有效人群很小，使得人群迁徙总是伴随着显著的瓶颈效应，这使它成为研究早期人类迁徙的最佳基因工具。尽管自然选择也可能是影响整个 Y 染色体的重要因素，并可能使某个单倍群的频率比仅有基因漂移的影响增长得更快，但相关经验的证据至今没有得到总结[46]。

数据分析 由于等位片段 DYS389Ⅱ的长度也包含了 DYS389Ⅰ，因此在所有的数据分析中，我们从 DYS389Ⅱ的长度中简单地减去了 DYS389Ⅰ的长度，以避免 DYS389Ⅰ长度的重复计算。简化的 DYS389Ⅱ命名为 DYS389b。由于 DYS385a 与 DYS385b，YCAⅡa 与 YCAⅡb 难以分辨，所以这些 STR 在进一步的计算中被排除，只基于余下的 16 个 STR。分析 Y-SNP 和 Y-STR 数据分别得到了单倍群和单倍型的多样度，还使用了 ARLEQUIN3.01 软件计算出的它们对应的标准误差(SE)值。用 ARLEQUIN3.01 也检测了 Y-SNP 和 Y-STR 的分子方差。各 Y- STR 的单倍型根据单倍群进行分组，并用 NETWORK3.0 构建了 M-J 网络结构图。权重的设置依据各 STR 位点的方差——权重与方差成反比。权重值设定为 2～8。

单倍群的最晚共祖时间有一个重要的局限性：该单倍群在这个时间之后开始扩散，并且携带它的人群在这个单倍群诞生之前就已经居住在这个地区。因此我们如 Wilson 等[13]那样用 BATWING 计算了最近共祖时间。这个软件用 MCMC 进程根据输入数据(一系列单倍型)产生系统树和其他相关数值，同时参考了基因模式或统计人口学模式。基因模式假设 STR 是逐步突变的，而统计人口学模式假设人口规模在一段时间内保持不变并产生分化，然后发生指数增长。基于 Zhivotovsky 等[54]给出的进化突变速率，我们在对所有 16 个 STR 计算时应用的是符合 γ 分布(1.47,2130)的突变速率。在 γ 分布之前，也分别尝试过 α (2,400), β (2,1)以及 N(1,0.001)分布 [15,17]。我们取一代人为 25 年。运算输出了 13,000 个样本，但因为数据偏离，最早的 3，000 个被剔除。因此，所有的结果都是基于这 10,000 个样本。各样本的 MCMC 循环次数在 10^2 到 10^5 间变化，所以总的循环次数在 10^6 到 10^9 之间。我们一开始对所有人群的 O-M95 样本进行了 10^6 次 MCMC 循环，但是得到的 $N_{posterior}$（人群扩张前的有效人口数量）值为～13，000。这个数字比全球的～5000 人大了很多[15,17]。我们增加了 MCMC 循环次数,发现随着 MCMC 循环次数的增加,$N_{posterior}$ 值下降了而扩张时间增大了,在 10^9 时也不稳定。因此，我们任意选择了各个人群 O-M95 样本中的 3～5 个染色体一起作为总样本，结果在 10^2 到 10^7 次循环时得到了收敛的 $N_{posterior}$ 值和扩张时间。参数显示，这个地区的有效人口数量为～1，000，与先前关于东亚的研究相符[17,18]。

来自公开发表的整个亚洲的 214 个人群的数据(来源如图 5)，包括印度次大陆，大洋洲以及澳大利亚，加上我们的数据一起通过应用 GIS 软件的 ArcView 程序产生了等位频率分布图。数据地点在图中用黑点表示。由于尼科巴人是当地人群的唯一样本，势必会对等频线产生强烈影响，所以计算等位频率分布图时排除了它。

原文参考文献

1. Gadgil M et al (1998) In Balasubramanian D et al (ed.) The Human Heritage. Hyderabad: Hyderabad University Press. 100-129.

2. van Driem G (2001) Languages of the Himalayas: An ethnolinguistic handbook of the Greater Himalayan region with a brief introduction to the symbiotic theory of language Volume 2. Leiden, Brill.
3. Diffloth G (2005) In: Sagart L et al (ed.) The Peopling of East Asia: Putting Together Archaeology, Linguistics and Genetics. London: Routledge Curzon. 77-81.
4. Kumar V & Reddy BM (2003) J Biosci 28:507-522.
5. Basu A et al (2003) Genome Res 13:2277-2290.
6. Thangaraj K et al (2005) Hum Genet 116:507-517.
7. Prasad BV et al (2001) Hum Biol 5:715-725.
8. Roychoudhury S et al (2001) Hum Genet 3:339-350.
9. Metspalu M et al (2004) BMC Genet 5:e26.
10. Kumar V et al (2006) Am J Hum Biol 18:461-469.
11. Langstieh BT & Reddy BM (1999) J Indian Anthro Soc 34:265-275.
12. Langstieh BT & Reddy BM (2004) The NEHU J 2:15-42.
13. Wilson IJ et al (2003) J R Stat Soc Ser A Stat Soc 166:155-188.
14. Qamar R et al (2002) Am J Hum Genet 70:1107-1124.
15. Zerjal T et al (2002) Am J Hum Genet 71:466-482.
16. Kayser M et al (2003) Am J Hum Genet 72:281-302.
17. Xue Y et al (2003) Genetics 4:2431-9.
18. Su B et al (1999) Am J Hum Genet 65:1718-1724.
19. Qian Y et al (2000) Hum Genet 106:453-454.
20. Su B et al (2000) Hum Genet 107:582-590.
21. Su B et al (2000) Proc Natl Acad Sci USA 97:8225-8228.
22. Capelli C et al (2001) Am J Hum Genet 68:432-443.
23. Hammer MF et al (2001) Mol Biol Evol 18:1189-1203.
24. Karafet T et al (2001) Am J Hum Genet 69:615-628.
25. Ramana GV et al (2001) Eur J Hum Genet 9:695-700.
26. Wells RS et al(2001)Proc Natl Acad Sci USA 98:10244-10249.
27. Kivisild T et al (2003) Am J Hum Genet 72:313-332.
28. Cordaux R et al (2004) Mol Biol Evol 21:1525-1533.
29. Cordaux R et al (2004) Curr Biol 14:231-235.
30. Wen B et al (2004) Nature 431:302-305.
31. Ray BC (1989) In: The Changing Socio-economic Profile New Delhi, Gian Publishing House.
32. Trivedi R et al (2006) J Hum Genet 51:217-226.
33. Shi H et al (2005) Am J Hum Genet 77:408-419.
34. Quintana-Murci L et al (2001) Am J Hum Genet 68:537-542.
35. Lal BB (1956) Ancient India 12:58-92.
36. Mohapatra GC (1975) J Archaeol Soc Nippon 60:4-18.
37. Mohapatra GC (1985) In: Deo SB et al (ed.) Recent advances in Indian Archaeology. Poona: Deccan College. 23-73.
38. Zide NH & Anderson GDS (2001) In: Subbarao KV et al (ed.) Yearbook of South Asian Languages and Linguistics. Sage Publications. 517-540.
39. Pinnow HJ (1963) In: Shorto HL (ed.) Linguistic Comparison in Southeast Asia and the Pacific. London: SOAS.140-152.
40. Underhill PA et al (2000) Nat Genet 26:358-361.
41. Cordaux R et al (2003) Eur J Hum Genet 3:253-264.
42. Thangaraj K et al (2003) Curr Biol 13:86-93.
43. Ding YC et al (2000)Proc Natl Acad Sci USA25:14003-14006.
44. Underhill PA et al (2001) Ann Hum Genet 65:43-62.
45. Fucharoen G et al (2001) J Hum Genet 3:115-125.
46. Jobling MA & Tyler-Smith C (2003)Nat Rev Genet 8:598-612.
47. Wu XZ, Poirier FE: Human evolution in China Oxford, Oxford University Press; 1995.
48. Jin L & Su B (2000) Nat Rev Genet 1:126-133.
49. Sambrook J et al (1989) Molecular cloning: A laboratory manual Cold Spring Harbor, Cold Spring Harbor Press.
50. Y Chromosome Consortium (2002) Genome Res 12:339-348.
51. Butler JM et al (2002) Forensic Sci Int 129:10-24.
52. Excoffier L et al (2005) Evol Bioinfo Online 1:47-50.
53. Bandelt H et al (1999) Mol Biol Evol 16:37-48.
54. Zhivotovsky LA et al (2004) Am J Hum Genet 74:50-61.