# 遗传学证实汉文化的扩散源于人口扩张*

李辉（译）
文波　李辉　卢大儒　宋秀峰　张锋　何云刚
李峰　高扬　毛显赟　张良　钱吉　谭婧泽
金建中　黄薇　Ranjan Deka　宿兵
Ranajit Chakraborty　金力

史载汉族源于古代中国北方的华夏部落，在过去的两千多年间，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）[4~5]。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[6~9]。两个亚群之间的方言和习俗差异也很显著[10]。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[11~12]，历史记载的汉族移民史[5]也与汉族的文化传播模式假说相矛盾。本研究对这两种假说进行了检验，证实汉文化的扩散中的确发生了大规模的人群迁徙（人口扩张模式）。

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作比较，其一是北方汉族，其二是南方原住民族，即现居于

---

* 原载于2004年8月出版的英国《自然》杂志。

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[13~16]，这些样本覆盖了中国绝大部分的省份（详见图 1 和补充信息表 1）。

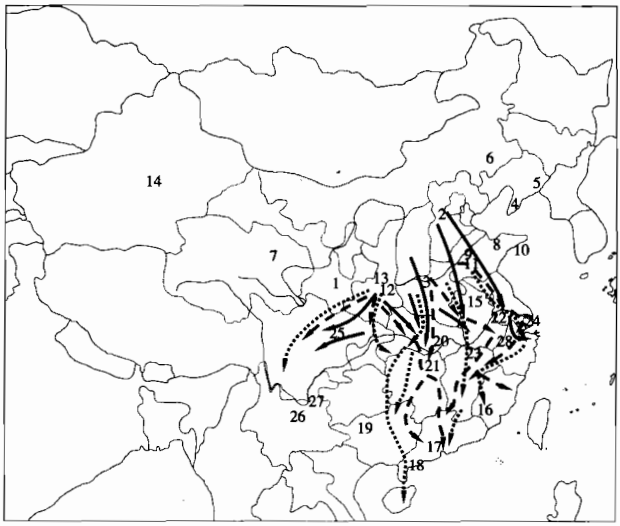


**图 1　调查群体的地理分布**

注：图中标出了历史记载中自北而南的三次迁徙浪潮。各群体的详细信息见补充材料 1。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三次迁徙浪潮。第一次发生于西晋时期（公元 265—316 年），迁徙人口约 90 万（大约当时南方人口的六分之一）；第二次发生于唐代（公元 618—907 年）规模比第一次大得多；第三次发生于南宋（公元 1127—1279 年），迁徙人口近 500 万。

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（见补充信息表 2），尤其是具有 M122－C 突变的单倍群（O3－M122 和 O3e－M134）普遍存在于我们研究的汉族群体中（北方汉族在 37%～71%之间，平均 53.8%；南方汉族在 35%～74%之间，平均 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 19%）高于北方汉族（1～10%，平均 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[17]，在南方汉族中低频存在（平均 4%），而北方汉族中却没观察到。如果我们假定起始于两千多年前的汉文化扩散[5]之前南方原住民族的 Y 类型频率与现在基本一致的话，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（Fst=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（补充信息表 3）。东亚北部的主要单倍群（A，C，D，G，M8a，Y，Z）在北方汉族中的频率（49～64%，平均 55%）比在南方汉族中（19～52%，平均 36%）高得多。另一方面，南方原住民族的主要单倍群（B，F，R9a，R9b，N9a）[12,14,18]在南方汉族中的频率（36～72%，平均 55%）要比在北方汉族（18－42%，平均 33%）高得多。线粒体类型的分布在南北汉族之间有

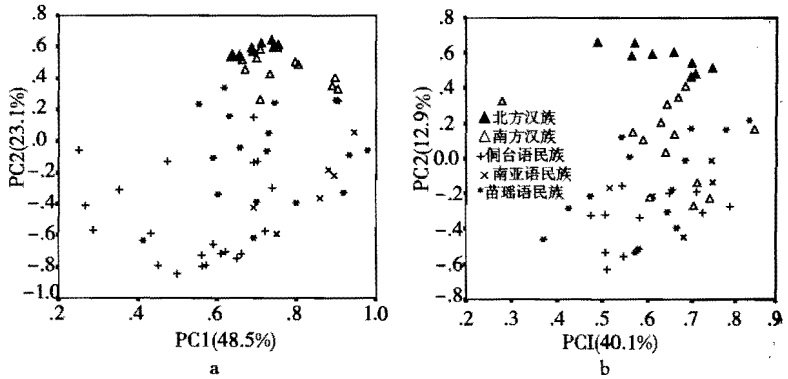


**图 2　主成分散点图**

a. 为 Y 染色体单倍群散点图；b. 为线粒体单倍群散点图

极显著差异（Fst=0.006，$P<10^{-5}$）。虽然南北汉族之间线粒体和Y染色体的Fst值相近，但线粒体的南北差异Fst值占群体间总方差的56%，而Y染色体仅仅占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。对NRY分析发现，几乎所有的汉族群体都聚在图2a的右上方。北方汉族和南方原住民族在第2主成分上分离，南方汉族的第2主成分值处于北方汉族和南方原住民族之间，但是更接近于北方汉族（北方汉族0.58±0.01；南方汉族0.46±0.03；南方原住民族-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mtDNA而言，北方汉族和南方原住民族仍然被第2主成分分开（图2b），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0.56±0.02；南方汉族0.09±0.06；南方原住民族-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。

我们进一步用两种不同的统计方法[19~20]来估计两个亲本（北方汉族和南方原住民）对南方汉族基因库的相对贡献（表1），这两个统计量用于单位点（single-locus）分析时比其它的方法更为准确[21]。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-test，$P<0.01$）；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，不足以导致统计上的显著性。

**表 1　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 | .868 ±.119 | .929 | .816 ±.214 | .755 |
| 福建 | 1 | .966 | .341 ±.206 | .248 |
| 广东 1 | .677 ±.121 | .669 | .149 ±.181 | .068 |
| 广东 2 | ND | ND | .298 ±.247 | .312 |
| 广西 | .543 ±.174 | .608 | .451 ±.263 | .249 |
| 湖北 | .981 ±.122 | .949 | .946 ±.261 | .907 |
| 湖南 | .732 ±.219 | .657 | .565 ±.297 | .490 |
| 江苏 | .789 ±.078 | .821 | .811 ±.177 | .786 |
| 江西 | .804 ±.113 | .829 | .374 ±.343 | .424 |
| 上海 | .819 ±.087 | .902 | .845 ±.179 | .833 |
| 四川 | .750 ±.118 | .713 | .509 ±.166 | .498 |
| 云南 1 | 1 | .915 | .376 ±.221 | .245 |
| 云南 2 | .935 ±.088 | .924 | .733 ±.192 | .645 |
| 浙江 | .751 ±.084 | .763 | .631 ±.180 | .540 |
| 平均 | .819 | .819 | .560 | .500 |

注：$M_{BE}$和 $M_{RH}$分别为参考文献［20］和［19］所描述的统计量。$M_{BE}$的标准误通过 1000 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 2000 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。

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。再有，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[22]。

据历史记载，受北方战乱和饥荒的影响，汉人不断的南迁，图 1 中画出了三次大规模移民的浪潮。在两千多年间，除了这三次大潮，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。

## 方法

### 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 1289 人，线粒体 23 个群体 1119 人。这些样本涉及了中国的大部分省份（图 1 和补充材料表 1）。

### 遗传标记

通过聚合酶链式反应—限制性片断长度多态性（PCR - RFLP）的方法[11]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，M120。根据Y染色体委员会的命名系统(YCC)[24]，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[23]。

线粒体上，对高变1区（HVS－1）进行测序，对编码区8个多态位点作了分型（9－bp缺失，10397 *Alu*Ⅰ，5176 *Alu*Ⅰ，4831 *Hha*Ⅰ，13259 *Hinc*Ⅱ，663 *Hae*Ⅲ，12406 *Hpa*Ⅰ，9820 *Hinf*Ⅰ），有关方法已有报道[22]。根据东亚线粒体系统树[18]，用高变1区突变结构和编码区多态性构建单倍群。

**数据分析**

根据线粒体和Y染色体单倍群频率，用SPSS10.0软件(SPSS公司）作主成分分析，研究群体间关系。南北汉族的遗传差异用ARLEQUIN软件[26]做AMOVA检验[25]。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[19~20]：ADMIX 2.0[27]和LEADMIX[21]软件。亲本群体的选择对混合比例的适当估计很重要[28~29]，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平均了三个族群：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[1,3]的线性回归分析揭示汉族群体的地理格局。

**参考文献**

[1] Cavalli－Sforza, L. L., Menozzi, P. & Piazza, A. The History and Geography of Human Genes (Princeton Univ. Press, Princeton, 1994).

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

[3] Chikhi, L. et al. Y genetic data support the Neolithic demic diffusion model. Proc. Natl Acad. Sci. USA 99, 11008 - 11013 (2002) .

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

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

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

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

[8] Chu, J. Y. et al. Genetic relationship of populations in China. Proc. Natl Acad. Sci. USA 95, 11763 - 11768 (1998) .

[9] Xiao, C. J. et al. Principal component analysis of gene frequencies of Chinese populations. Sci. China C, 43, 472 - 481 (2000) .

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

[11] Su, B. et al. Y chromosome haplotypes reveal prehistorical migrations to the Himalayas. Hum. Genet. 107, 582 - 590 (2000) .

[12] Yao, Y. G. et al. Phylogeographic differentiation of mitochondrial DNA in Han Chinese. Am. J. Hum. Genet. 70, 635 - 651 (2002) .

[13] Cavalli - Sforza, L. L. & Feldman, M. W. The application of molecular genetic approaches to the study of human evolution. Nature Genet. 33, 266 - 275 (2003) .

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

[15] Underhill, P. A. et al. Y chromosome sequence variation and the history of human populations. Nature Genet. 26, 358 - 361 (2000) .

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

(2003) .

[17] Su, B. et al. Y - chromosome evidence for a northward migration of modern humans into eastern Asia during the last ice age. Am. J. Hum. Genet. 65, 1718 - 1724 (1999) .

[18] Kivisild, T. et al. The emerging limbs and twigs of the East Asian mtDNA tree. Mol. Biol. Evol. 19, 1737 - 1751 (2002) .

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

[20] Bertorelle, G. & Excoffier, L. Inferring admixture proportions from molecular data. Mol. Biol. Evol. 15, 1298 - 1311 (1998) .

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

[22] Wen, B. et al. Analyses of genetic structure of Tibeto - Burman populations revealed a gender - biased admixture in southern Tibeto - Burmans. Am. J. Hum. Genet. 74, 856 - 865 (2004) .

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

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

[25] 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 131, 479 - 491 (1992) .

[26] Schneider, S. , et al. Arlequin: Ver. 2.000. A software for population genetic analysis. (Genetics and Biometry Laboratory, Univ. of Geneva, Geneva, 2000) .

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

[28] Chakraborty, R. Gene admixture in human populations: Models and predictions. Yb. Phys. Anthropol. 29, 1 - 43 (1986) .

[29] Sans, M. 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. 118, 33 - 44 (2002).

**补充信息** 本论文的补充信息置于www. nature. com/nature.

**致谢** 感谢所有为本研究提供样品的志愿者。样本采集得到 NSFC 和 STCSM 基金给予复旦大学的资助。金力、Ranjan Deka 和 Ranajit Chakraborty 得到 NIH 和 NSF 基金的资助。