

中国人群基因频率的主成分分析

肖春杰^{1)*} 杜若甫

(中国科学院遗传研究所, 北京 100101)

L. L. Cavalli-Sforza E. Minch

(Department of Genetics, School of Medicine, Stanford University, Stanford, CA 94305, USA)

摘要 对中国人群中 38 个基因座的 130 个等位基因的频率进行了主成分分析, 绘制了主成分分布地图和主成分综合地图。结果显示, 汉族人群的第 1 主成分分布反映了南方蒙古人种与北方蒙古人种间遗传结构差异的梯度变化, 而其第 2 主成分分布则反映了蒙古人种与高加索人种间的基因流动。中国少数民族人群的第 1 主成分分布与汉族中的第 2 主成分相似, 而第 2 主成分分布则与汉族中的第 1 主成分相似。当将中国汉族与少数民族的基因频率放在一起分析时, 所得出的主成分分布基本上与少数民族的相似。中国人群基因频率的第 3, 4 主成分的分布则可能反映了古代某些地区人口骤增而向外扩展的历史事件。中国人群基因频率主成分综合地图明显地显示出: 在亚洲的南、北方蒙古人种间有一明显的分界线——长江。猜测: 如果中国现代人确实起源于非洲, 则南、北方蒙古人种的祖先在进入亚洲以前已经分开, 南方蒙古人种的祖先自非洲或中东沿海岸向东迁徙、扩展、直至东南亚, 然后折向北, 至华南, 再直至长江, 甚至可能有一段时间曾占据过长江以北的一些地区。北方蒙古人种的祖先则在帕米尔以北地区自西向东, 直至东北亚, 再扩展至美洲、日本等地, 同时向南, 直至长江。在近二三千年内, 北方蒙古人种才不断向长江以南扩展、渗透。

关键词 中国人群 基因频率 主成分分析

人群遗传结构的差异, 实质上是基因频率的差异。分析人群基因频率的地理分布, 有助于了解人口迁移、人群融合、自然选择、漂变、地理和社会隔离、突变等对人类微进化的影响, 有助于了解现代人的起源与扩展。

在中国人群的基因频率方面, 目前已积累了较多的数据, 可以研究其地理分布。但是各个基因在各个人群中的频率分布变化十分复杂, 很难看出目前各基因频率地理分布的总趋势与造成这种分布的原因或事件。

主成分分析对于解决上述困难很有帮助。它是 Hotelling 于 1933 年首先提出来的, Rao (1948) 首先将它应用于人体测量学数据的分析, Cavalli-Sforza 和 Edwards 于 1949 年首先将它应用于人类基因频率的分析。

1999-07-03 收稿, 2000-01-13 收修改稿

* E-mail: cjxiao@public.km.yn.cn

1) 现地址: 云南大学生物系, 昆明 650091

本文采用人类基因频率地理分析的软件 Genography^[1]对中国人群 38 个基因座上 130 个等位基因的频率进行了主成分分析, 并绘制了主成分分布地图与综合地图。

1 材料与方法

进行基因频率主成分分析的基因座共 38 个, 与计算中国人群间的遗传距离^[2]时所用的相同, 它们是 ABO, MNSs, Rh, P, Diego, Duffy, Kell, Kidd, Lewis, Lutheran 等红细胞血型, 白细胞抗原系统中的 HLA(A, B, C, D), 酸性磷酸酶、腺苷脱氨酶、腺苷酸激酶、酯酶 D、葡萄糖-6-磷酸脱氢酶、谷丙转氨酶、乙二醛酶、葡萄糖磷酸变位酶-1、磷酸葡萄糖酸脱氢酶等红细胞酶, α -抗胰蛋白酶、补体第 2, 3, 4, 6, 7 组分、备解素因子 B、维生素 D 结合蛋白、结合珠蛋白、免疫球蛋白 Gm 因子与 Km 因子、转铁蛋白等血清蛋白质, 以及血型分泌型、苯硫脲味盲和盯眈类型等。其中 MNSs 与 Rh 都以 1 个基因座来对待, 而以各单倍型频率作为基因频率。对每一套(一个基因座上一个人群的基因频率算一套)数据先进行 c^2 分析, 不符合 Hardy-Weinberg 定律者剔除不用。

不同研究者对同一经纬度的同一民族、同一基因座所报道的基因频率数据, 用加权法进行合并。

中国的汉族人群共有 704 套数据可供分析。在中国 55 个少数民族^[3]中共有 1 190 套数据可供分析。

在将汉族与少数民族放在一起进行分析时, 同一经纬度不同民族在同一基因座上的基因频率也以加权法先进行合并, 共有 1 606 套数据可供分析。

我们一共利用了 600 多篇文献中的基因频率数据, 由于文献太多, 无法在参考文献中一一列出, 故从略。

主成分分析的原理与方法参见文献^[1]。

2 结果

2.1 汉族

汉族的第 1 主成分的方差占总方差的 32.9%, 即包含了基因频率总变量的 1/3 左右的信息量, 其地理分布明显地显示出从东北向西, 而主要是向南的梯度变化。可以认为, 它反映了北方蒙古人种与南方蒙古人种间遗传结构的差异。值得注意的是在长江以北只有两个梯度, 说明北方地区的汉族人群间遗传结构的差异相当小。这是容易理解的, 因为北方地区主要是平原, 地理隔离少, 黄河冬天封冻, 可以自由行走, 而且近五千年来由战乱、灾荒所引起的大规模人口迁移, 且迁移频繁, 所以北方各地汉族人群在遗传结构上已融合得相当均匀了。而在长江以南地区, 第 1 主成分和下面要谈到第 2, 3, 4 主成分分布的梯度相当密, 表明其遗传结构变化大而剧烈。越往南, 南方蒙古人种血缘越多, 而同时北方蒙古人种血缘越少。南方山多, 河流多, 地理隔离要比北方严格得多, 同时由于战乱与灾荒引起的大规模的人口迁移也比北方少。可以看出, 在汉族中南方蒙古人种血缘成分最高的是海南、广西及广东西部(图 1(a))。

汉族人群基因频率的第 2 主成分的方差占总方差的 23.9%。它的梯度变化是从西部向东、向南的, 差异最大的是中国西部与海南、广东两省之间(图 1(b))。我们认为, 第 2 主成分反映了高加索人种与蒙古人种之间的基因流动。可见, 在汉族各地人群间, 高加索人种与蒙古人种

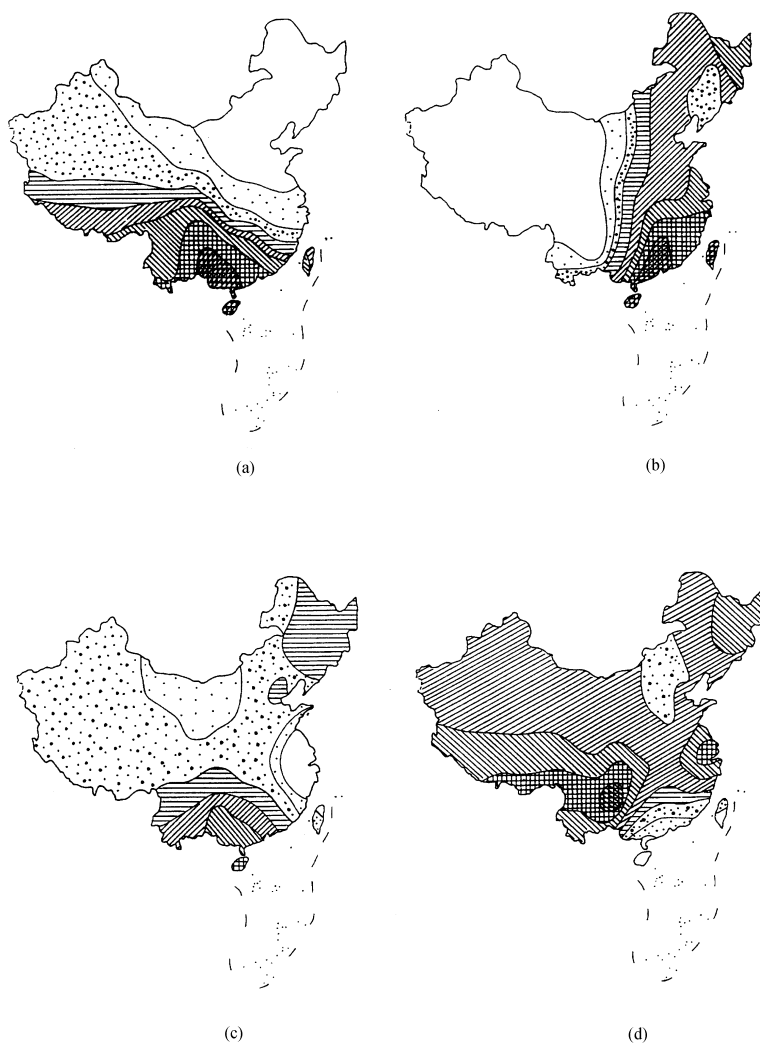


图 1 汉族人群基因频率主成分分布示意图

(a) ~ (d) 分别为第 1~4 主成分分布示意图

间基因流动所造成的遗传结构差异只居次要地位, 而南、北方蒙古人种间的基因流动所造成的差异却居主要地位. 这是因为, 尽管高加索人种与蒙古人种间的遗传差异要比南、北方蒙古人种间的大得多, 可是在汉族各地人群中, 蒙古人种血缘占绝大部分, 而高加索人种的血缘只占很少一部分, 也就是说, 高加索人种流入汉族人群的基因并不很多, 因此对各地汉族遗传结构的影响不大.

汉族各地人群基因频率的第 3, 4 主成分分别反映了 8.6% 与 5.5% 信息量. 在图 1(c) 与 (d) 上, 有一些颜色特别深或特别浅的区域, 如广东、广西与海南, 长江三角洲, 内蒙古中部、北部与山西、宁夏, 以及四川南部等, 这些地区很可能在古代某一时期因农业、牧业比较发达, 或

石器、青铜器、陶器等制造比较先进, 而曾出现过人口骤增、向邻近地区扩展的情况。由于基因频率调查的取样地点密度并不算大, 所以这些地区的定位可能还不够精确。而出现人口骤增、向邻近地区扩展的年代, 则需要与考古学、人类学、语言学、历史学等其他学科的研究结合起来, 才有可能得出结论。

2.2 少数民族

中国少数民族的基因频率主成分分布与汉族的有许多相似之处, 这再一次证明, 各地汉族人群中已融入了当地少数民族的大量血缘, 同时, 各地少数民族也从当地汉族吸收了相当多的血缘^[3]。可是两者也有明显的区别, 首先是少数民族基因频率的第1主成分反映了高加索人种与蒙古人种间的基因流动, 而第2主成分则反映了南、北方蒙古人种间的基因流动, 这正好与汉族人群中的情况相反。

少数民族基因频率分布的第1主成分自西北向东、向南呈梯度变化(图2(a)), 显然, 它是由高加索人种与蒙古人种之间的基因流动所造成的。从许多基因的等位基因频率及人类学性状早已明显看出, 在中国各民族中, 西北地区少数民族中高加索人种的血缘最多。在第1主成分分布地图上可看到, 西北地区少数民族处于一个极端, 而广东、广西及海南等省区少数民族处于另一极端, 即其高加索人种血缘最少。第1主成分的方差为总方差的30.1%。

少数民族基因频率分布的第2主成分占总方差的18.6%, 它呈现自东北向西、主要是向南的梯度变化。颜色最浅的是东北, 颜色最深的是广东、广西、海南、云南, 以及福建南部。可以明显看出, 它和汉族人群中的第1主成分一样, 反映了南、北方蒙古人种间的基因流动(图2(b))。

少数民族基因频率分布的第3主成分的方差占总方差的13.6%, 整个分布地图显示出东北与西南部灰度最浅, 而东南与西北部灰度最深(图2(c))。这也许是因为这些灰度最深或最浅的地区是我国今天少数民族祖先的几个主要发源地, 在历史上曾经一度人口骤增, 向中原地区扩展过。这只是初步推测, 尚有待有更多的基因频率数据和其他资料, 以便作深入的分析。

少数民族基因频率分布的第4主成分的方差占总方差的7.5%, 它十分明显地显示出, 是以黄河中游为核心, 向四面扩展的(图2(d))。黄河中游是五千年前以炎、黄两族为核心形成的华夏族(即汉族的祖先)的发源地, 因此, 第4主成分很可能反映了以华夏族为基础的汉族的基因向四周各少数民族扩散这一历史事件。

可是在汉族的第1~4主成分分布图中, 却没有显示出华夏族对各地汉族基因频率的明显影响。出现这一结果的最可能的解释是各地汉族基因频率受当地其他人群(古代少数民族)的影响大, 而受华夏族的影响小。这一点也已为对全国各地汉族和少数民族人群间遗传距离的分析所证明, 各地汉族人群与当地少数民族的遗传距离, 平均要比各地汉族间的小, 差异显著^[2]。

2.3 汉族与少数民族合在一起分析

汉族与少数民族人群的数据放在一起分析得出的全中国人群基因频率第1~4主成分的方差, 分别占总方差的30.4%, 17.2%, 12.2%及6.0%, 这些主成分的分布地图明显地与少数民族主成分分布地图相似(图2, 3)。有两点主要的不同之处: 一是在反映南、北方蒙古人种间基因流动的第2主成分分布地图(图3(b))上, 西北地区与东北地区的差异减小了。这是可以理解的, 因为西北地区的汉族主要是从中国北部其他地区迁去的, 由于汉族数据的加入, 自然会少

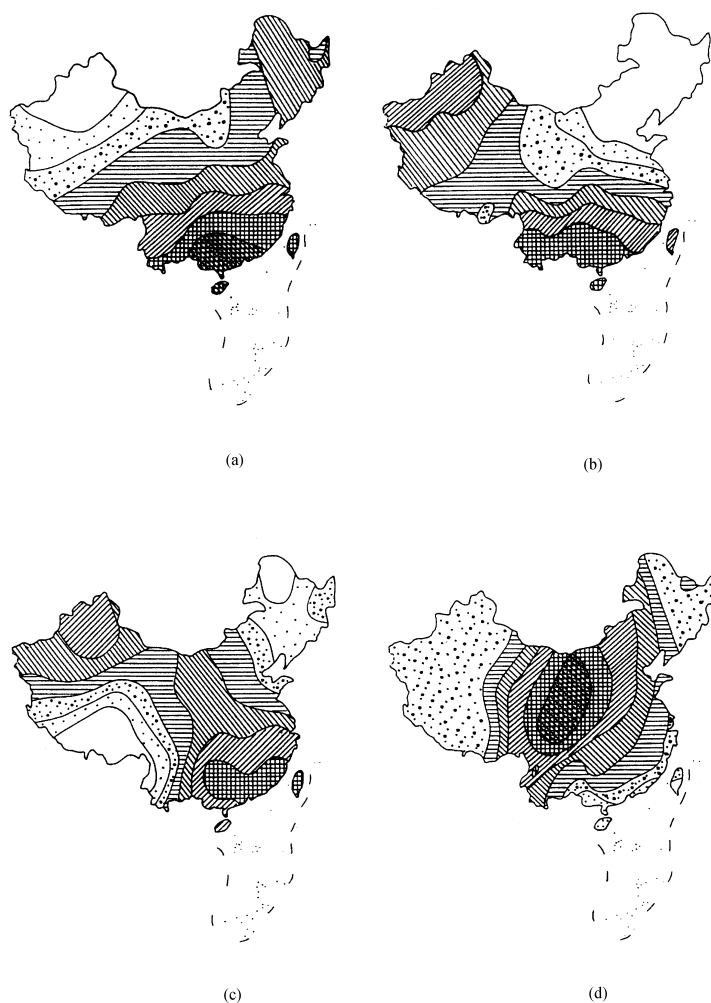


图 2 中国少数民族人群基因频率主成分分布示意图

(a)~(d) 分别为第 1~4 主成分分布示意图

数民族基因频率原第 2 主成分分布地图上西北地区与东北地区的差异减小。二是两张第 3 主成分分布地图虽然图形十分相似，可是深色区与浅区相互掉换了，全中国主成分分布地图中的深色区相当于少数民族分布地图上的浅色区，而其浅色区则相当于另一张图上的深色区(图 2(c)、3(c))。这是由于制图时的技术问题造成的，并无实质上的意义。

2.4 主成分综合图

利用我们收集的全部基因频率数据，加上 Cavalli-Sforza 实验室原来收集的亚洲人群基因频率数据，绘制了亚洲人群基因频率第 1~3 主成分的综合地图。从图版 I (附本刊后，下同)可以显著看出，华南南方蒙古人种居住的绿色区与东南亚等绿色区相连，而长江以北的北方蒙古人种居住区呈现为紫色，与整个西伯利亚连成一片。只在中亚地区紫色多有蓝、绿等色调，

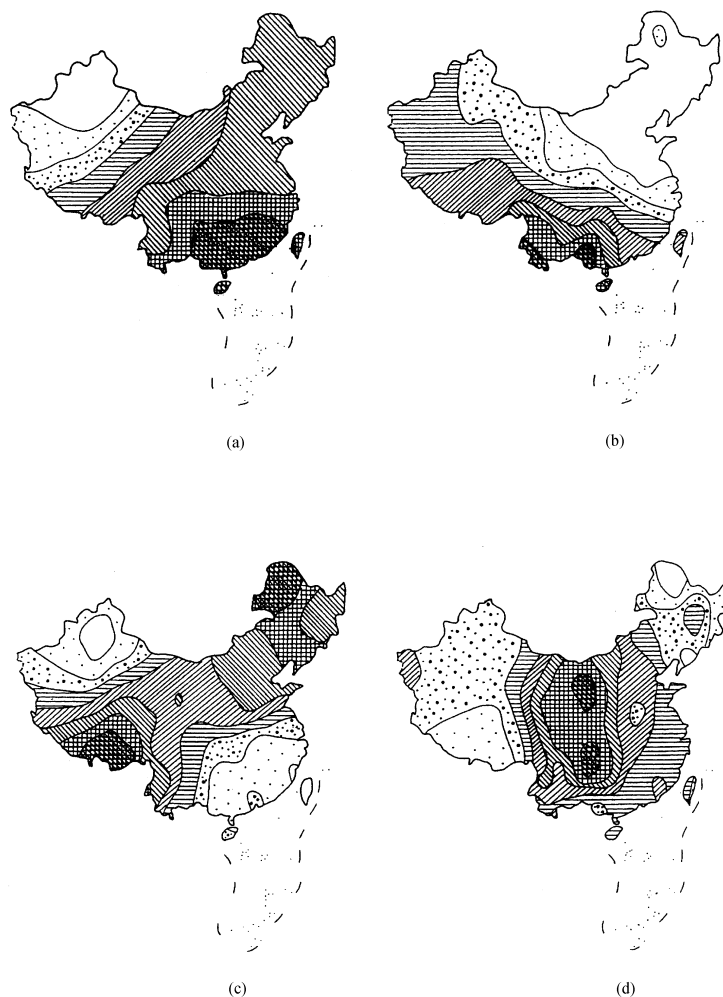


图 3 中国人群基因频率主成分分布示意图

(a) ~ (d) 分别为第 1~4 主成分分布示意图

而且略深, 与高加索人种居住的欧洲相连, 说明在中亚地区, 人群中高加索人种的血缘有所增加(图版 I)。在 Cavalli-Sforza 等人所绘制的世界人群基因频率的主成分综合地图上^[1], 东西伯利亚的紫色区与美洲相连, 而印度尼西亚、菲律宾、新几内亚等的黄色区与澳大利亚相连。从以上叙述可以看出, 中国长江附近的南、北方蒙古人种居住区的分界线, 在世界范围内显得何等地突出、明显。

3 讨论

3.1 南、北蒙古人种间的基因流动

尽管至少在近二三千年内, 主要是由于战乱和灾荒, 中国人口迁移的规模很大, 可是令

人惊奇的是, 至今在亚洲基因频率主成分综合地图上, 南、北方蒙古人种间的界线仍非常清楚, 基本上以长江为界. 这与应用线粒体、微卫星、Y 染色体多态以及肤纹研究的结果一致^[4~7].

近四五千年内, 中国人口迁移的主要方向是以黄河中、下游为中心, 向四面扩散, 主要方向是向南, 在过了长江以后, 继续向南, 并向西. 但是在史前时期, 来自南方的南方蒙古人种很可能曾越过长江, 在长江以北占据了一些地方, 可是后来又被北方蒙古人种赶回至长江以南. 中国很早的历史文献就曾记载, 汉族传说中的始祖黄帝, 曾与南方苗人的领袖蚩尤大战于河北涿鹿一带, 蚩尤战败, 率众越过长江, 居住在今鄱阳湖至洞庭湖一带^[8, 9].

已有比较有力的考古学证据证明, 根据新石器时代的古人类化石的人体测量学数据进行的聚类分析, 明显地表明, 当时已存在南、北方蒙古人种之间的差异^[10]. 于是自然地产生一个问题: 这种差异可以追溯到多远? 也就是说, 南、北方蒙古人种究竟是从什么时候分开的? 这就涉及到现代人的起源问题.

我们认为, 无论现代人非洲起源说和本地起源说, 都必须对蒙古人种今天如此明显地分为南、北两大类型这一事实作出解释.

在亚洲的基因频率主成分综合地图上可看到, 中国的南方蒙古人种分布区与东南亚等连成一片, 而长江以北的北方蒙古人种分布区则与广袤的西伯利亚、东北亚都连成一片, 因此, 很可能南、北方蒙古人种早在蒙古人种发展的早期就已分开了, 而且, 如果中国的北方蒙古人种是从中国南部逐渐扩展过去的^[4], 或者中国南方蒙古人种是从北方逐渐扩展过去的^[11], 似乎都不太可能形成他们之间今天在长江附近如此明确而显著的分界线. 所以, 如果站在现代人非洲起源说的立场, 则最可能的是蒙古人种的祖先由两条路线从非洲或中东到中国, 并在长江汇合. 一条是沿着亚洲大陆南部的海岸直达南亚、东南亚, 然后往北, 直达长江, 正如上面已提到的, 在史前时期, 还可能曾越过长江, 到达华北的东部地区, 后来又被迫退回长江以南; 另一条是从帕米尔高原以北地区进入今天的中国境内, 向东、向北, 直抵东北亚, 向南直抵长江, 后又经朝鲜半岛到日本和经白令海峡到美洲.

我们并不完全否定蒙古人种现代人从一方(南方或北方)进入中国大陆的可能性. 我们只是认为, 从南方或北方一方进入中国大陆, 很难解释在主成分综合图上所显示出来的南、北蒙古人种在遗传结构上如此显著而且突出的差异. 因此, 我们提出第 3 种可能性, 当然, 这也只是一种猜测. 蒙古人种起源及迁徙的最后阐明, 尚有待积累更多的人类群体遗传学研究结果, 也有待于更多的考古学的新发现, 特别是距今 10 万~2 万年的考古新发现, 同时还应结合语言学、人类学、古地理学、历史学等研究成果, 才能得出正确的结论.

3.2 高加索人种与蒙古人种间的基因流动

本研究可得出两点肯定的结论:

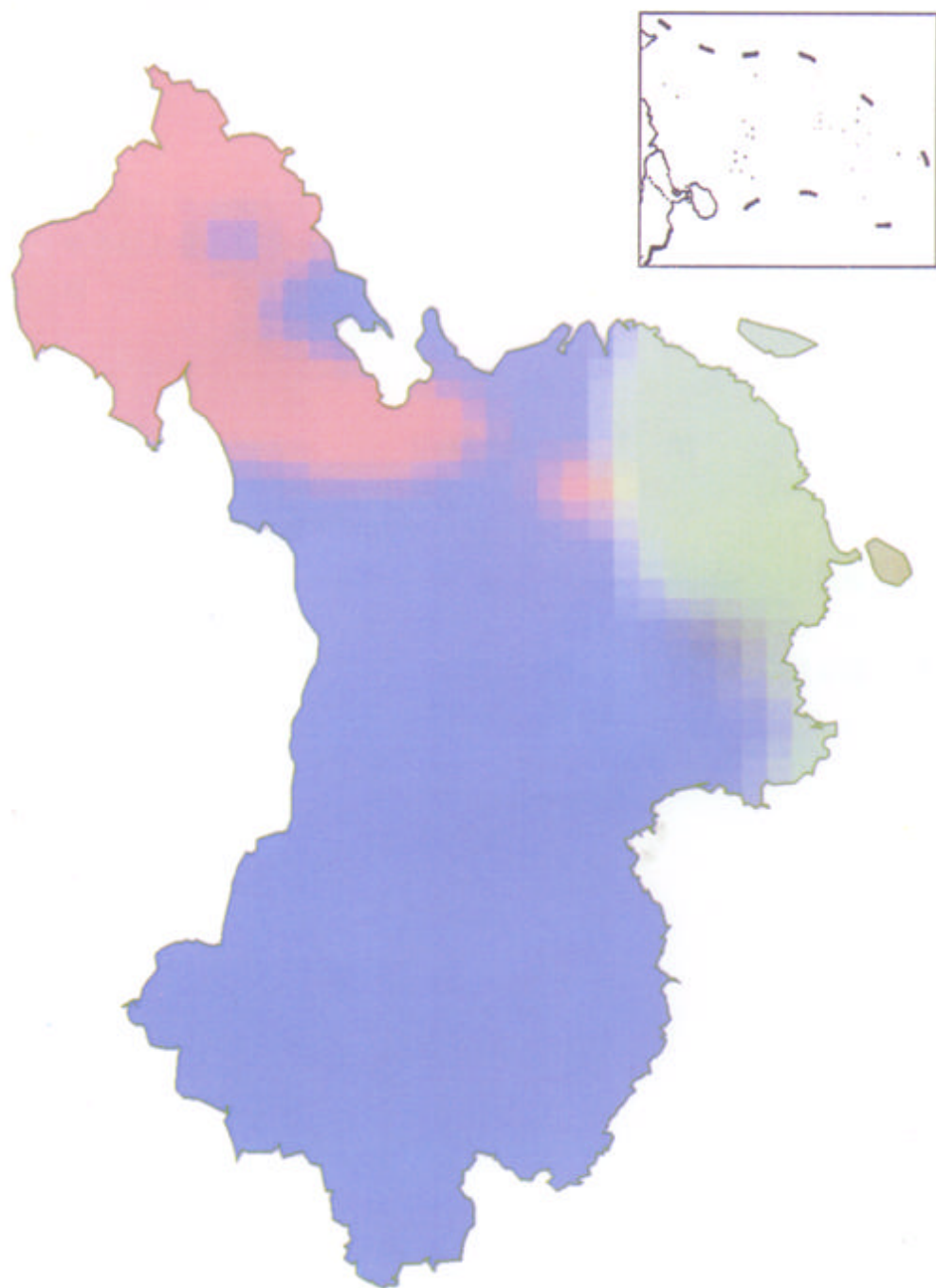
第 1 点是高加索人种与蒙古人种间基因流动所造成的人群间遗传结构的差异, 在汉族中只占第 2 位, 占第 1 位的是南、北方蒙古人种间基因流动所造成的差异; 而在少数民族人群中当少数民族与汉族人群放在一起分析时, 高加索人种与蒙古人种间基因流动所造成的差异却居第 1 位.

第 2 点是在中国境内, 高加索人种血缘成分最多的是西北部的人群, 而成分最少的是最南部地区, 即广东、广西与海南. 高加索人种与蒙古人种间的基因流动所造成的遗传结构梯度变化呈自西北向东、向南, 直至中国的最南部. 这说明, 唐、宋以后到广州、泉州等地经商

的波斯商人等所带入的高加索人种血缘甚少, 而自印度几乎没有高加索人种的血缘传入中国, 高加索人种血缘基本上是中国西北部向东、向南扩散的, 而且高加索人种血缘流向中国主要可能还是在史前时代, 而不是在有了丝绸之路以后。

参 考 文 献

- 1 Cavalli-Sforza L L, Menozzi P, Piazza A. The History and Geography of Human Genes. Princeton: Princeton Univ Press, 1994. 541~548
- 2 杜若甫, 肖春杰, Cavalli-Sforza L L. 用 38 个基因座的基因频率计算中国人群间遗传距离. 中国科学, C 辑, 1998, 28(1): 83~89
- 3 Du R, Vincent F Y. Ethnic Groups in China. Beijing: Science Press, 1993. 318
- 4 Chu J Y, Huang W, Kuang J M, et al. Genetic relationship of populations in China. Proc Natl Acad Sci USA, 1998, 95: 11 763~11 768
- 5 肖君华, 胡 芳, 徐红岩, 等. 3 个与艾滋病程相关的基因的多态位点在中国不同地区人群中的分布研究. 中国科学, C 辑, 1999, 29(5): 556~560
- 6 许丽萍, 徐玖瑾, 朱苏玲, 等. 中国 10 个人群中 Y 染色体 Alu 序列的多态分布. 科学通报, 1998, 43(8): 843~846
- 7 张海国, 丁 明, 焦云萍, 等. 中国人肤纹研究Ⅲ. 中国 52 个民族的肤纹聚类. 遗传学报, 1998, 25(5): 381~391
- 8 伍新福, 龙伯亚. 苗族史. 成都: 四川民族出版社, 1992. 2~38
- 9 范文澜. 中国通史简编. 修订本. 第一编. 北京: 人民出版社, 1965. 88~94
- 10 张振标. 中国新石器时代人类遗骸. 见: 吴汝康主编. 中国远古人类. 北京: 科学出版社, 1989. 62~80



中国人群基因频率主成分综合分布示意图