

中国 21 个人群的遗传拓扑学分析

谭 茜 杜若甫

(中国科学院遗传研究所, 北京 100101)

关键词 遗传拓扑学; 中国人; 群体遗传学

内 容 提 要

用脂酶 D, 酸性磷酸酶, 谷丙转氨酶和 6-磷酸葡萄糖酸脱氢酶等四个位点的基因频率对中国 16 个少数民族和 5 个汉族人群进行了遗传拓扑学分析, 绘出了他们地理位置二维同步排序图。根据所得结果, 对汉族和 16 个少数民族的起源, 彼此间血缘远近、历史上的迁移等进行了讨论。

在某一群体中某一位点各等位基因的频率, 即是该群体在该位点的遗传结构; 其全部位点的基因频率即是该人群的整个遗传结构。人们往往希望能形象地表示出各个群体的遗传结构及彼此间的异同。这在位点数少时很容易做到, 而当有基因频率数据的位点多时, 用直线坐标、双向坐标或三维立体坐标等等较简单的方法就无法形象地表示出群体的总的遗传结构, 需要特殊的能简化数据结构的方法。

早在 60、70 年代, 近代发展起来的一些处理数据的多元分析方法, 已被应用于研究人类群体结构, 如聚类和排序分析。有人把用这种方法分析得出的群体间的关系称为群体的空间结构。

这类简化数据结构的分析方法, 通常是在尽可能地保留原始信息量的前提下, 降低观察空间的维数(相异矩阵的维数), 从而可形象地、直观地在二维空间、至多在三维空间中显示群体在属性空间中的相对位置和变化趋势。群体结构研究中常用的降维处理方法有: 特征向量及主成分分析法, 遗传拓扑法、非参数标度、树形法等等(Lalouel, 1972)。

其中的遗传拓扑分析提供了一种比较两种相异矩阵之间关系的有效方法, 一般用于研究遗传距离与其他非遗传因素构成的矩阵之间的联系, 尤其是遗传距离与地

收稿日期: 1991-04-12

理距离间的关系。它通过对人群两个不同属性(遗传距离与地理距离)排序空间(如主坐标排序)的最大拟合, 在同一个二维空间中对两个不同属性实体进行排序, 故亦叫同步排序。这样便可直观地展示人群的遗传位置与地理位置的关系, 不仅可展示人群的遗传结构的相异程度, 还可在一定程度上展示人群历史上的迁移方向与距离远近。

Schonemann(1966)以及 Schonemann 与 Garroll(1970)提出的最小二乘法, 较好地解决了对两种排序空间进行拟合的问题。

国外已有一些学者用此方法研究过不同人群的群体结构(Morton *et al.*, 1968; Morton *et al.*, 1973; Balakrishnan *et al.*, 1968; Lee *et al.*, 1988)。

中国幅员辽阔, 人口众多, 而且是一个多民族的国家。虽然各民族的人种成分比较单纯, 绝大多数均属于蒙古人种, 但语言分化很大, 分属五个语系。同时, 各民族在风俗习惯、宗教信仰、文化发展水平等方面, 均存在极大的差异。在中国历史上曾发生过多次人口的大规模迁徙。各民族间的融合、地理隔离、环境因子的选择作用、突变等, 都影响着各个人群的遗传分化。研究中国人群遗传距离与地理距离之间的关系以及彼此间遗传距离的大小, 将有助于了解各民族的起源与发展、彼此间的血缘关系、历史上的迁徙等。

本文根据酯酶 D、酸性磷酸酶、谷丙转氨酶及 6-磷酸葡萄糖酸脱氢酶等四个位点的基因频率, 用遗传拓扑学的方法分析了中国 21 个人群的遗传距离及其与地理距离的关系。

一、 数据与分析方法

本文所分析人群的取样地点及其经纬度见表 1。21 个人群的 4 个位点的基因频率是中国科学院遗传研究所人类群体遗传实验室分析得出的(另文发表)。每一人群的被调查者约 200 人, 他们在取样地点均生活了三代以上, 三代以内无异族通婚, 而且彼此间无直系血缘关系。

在计算 21 个人群间遗传距离时, 用了 B^2 (Balakrishnam *et al.*, 1968)及 E^2 (Edwards, 1971)两种方法。两种方法得出的 21 个人群间遗传距离数据经检验呈高度相关, 相关系数为 0.9987。所以在本文中, 只用遗传距离 E^2 建立了遗传距离阵, 然后求出其前两个主坐标的二维排序阵。

在建立地理距离阵时, 先按 Morton 等的方法, 将各人群所处经、纬度坐标转换成以公里数为单位的直角坐标(Morton *et al.*, 1968、1973)。然后求出 21 个人群间地理距离的前两个主坐标的二维排序阵。

最后一个步骤是将遗传距离的主坐标二维排序阵与地理距离的主坐标二维排序阵进行最大一致的拟合。所用的方法是 Schonemann(1966)及 Schonemann 与

Carroll(1970)提出的最小二乘法(Lalouet, 1972)。

表 1 各人群采样地点及其经纬度

Sampling sites of populations and their longitude and latitude

| 人 群 | 采 样 地 点 | 东 经 | 北 纬 | 人 群 | 采 样 地 点 | 东 经 | 北 纬 |
|-------|----------|-------|------|-----------|---------|-------|------|
| 维吾尔族 | 新疆乌鲁木齐市 | 87.6 | 43.8 | 侗 族 | 广西三江县 | 109.5 | 25.7 |
| 朝 鲜 族 | 吉林延吉市 | 129.4 | 42.7 | 壮 族 | 广西武鸣县 | 108.2 | 23.1 |
| 满 族 | 辽宁岫岩县 | 123.2 | 40.2 | 布 依 族 | 贵州安顺市 | 105.7 | 26.0 |
| 蒙 古 族 | 内蒙古呼和浩特市 | 111.7 | 40.8 | 黎 族 | 海南保亭县 | 109.7 | 18.6 |
| 彝 族 | 四川布拖县 | 102.7 | 27.7 | 瑶 族 | 广西马山县 | 107.2 | 24.1 |
| 回 族 | 宁夏同心县 | 105.9 | 36.9 | 哈 尔 滨 汉 族 | 黑龙江哈尔滨市 | 126.6 | 45.7 |
| 藏 族 | 西藏拉萨市 | 91.1 | 29.6 | 西 安 汉 族 | 陕西西安市 | 108.9 | 34.1 |
| 苗 族 | 湖南吉首市 | 109.7 | 28.3 | 成 都 汉 族 | 四川成都市 | 104.1 | 30.6 |
| 土 家 族 | 湖南吉首市 | 109.7 | 28.3 | 贵 阳 汉 族 | 贵州贵阳市 | 106.7 | 26.6 |
| 白 族 | 云南剑川县 | 99.8 | 26.5 | 漳 州 汉 族 | 福建漳州市 | 117.6 | 24.5 |
| 哈 尼 族 | 云南元江县 | 101.9 | 23.5 | | | | |

二、结果与讨论

表 2 列出了根据四个位点的基因频率计算得出的 21 个人群间的遗传距离。从表 2 可以大致地看出,一些典型的南方民族(如瑶、黎、壮等族)和一些北方的民族(如朝鲜、满、蒙古等族)及起源于北方的民族(如藏、彝等族)之间,遗传距离最大。如遗传距离在瑶族与藏、蒙古、满、朝鲜等族间分别是 0.2097、0.1262、0.1231 与 0.1100,在黎、布依等族与藏族之间分别是 0.1639 与 0.1393,黎族和满族之间是 0.1002。

表 2 21 个人群的遗传距离 E^2 ($\times 10^{-4}$)

Cenetic distances E^2 between 21 populations ($\times 10^{-4}$)

维吾尔族与其它民族遗传距离都比较大, 但与哈尼族却最相近($E^2 = 0.0064$), 这也许是偶然的。汉族亚群之间、北方民族之间、或南方民族之间的遗传距离往往较小。另外, 汉族亚群和当地的一些少数民族的遗传距离往往也较小。但由于数据很多, 有210个, 人群有21个, 分布在全国各地, 仅从遗传距离分析就很费力, 因此, 必须借助于拓朴学方法。

根据21个人群间的地理距离和遗传距离 E^2 , 经过最大一致性的拟合后绘出的二维同步排序图, 其遗传距离前两个主坐标所保留的信息量为84%, 这是相当高的。

遗传位置主坐标排序阵和地理位置主坐标排序阵之间的相关系数是0.628。这一数值也是相当高的, 说明所研究的21个人群间在四个位点上的遗传差异大小和地理距离的远近关系十分密切。这暗示地理距离对于迁徙的影响和地理隔离等地理因素对人群间四个位点的遗传差异影响相当大, 而突变、遗传漂变、选择作用等的影响相对来说都小一些。

为了可以更清楚地看出各民族的遗传位置与地理位置, 我们把汉族五个人群与16个少数民族的遗传位置与地理位置分别表明在图1与图2上。这两个图反映处的人群间亲缘关系与历史上的迁移、融合等, 是十分有意义的。

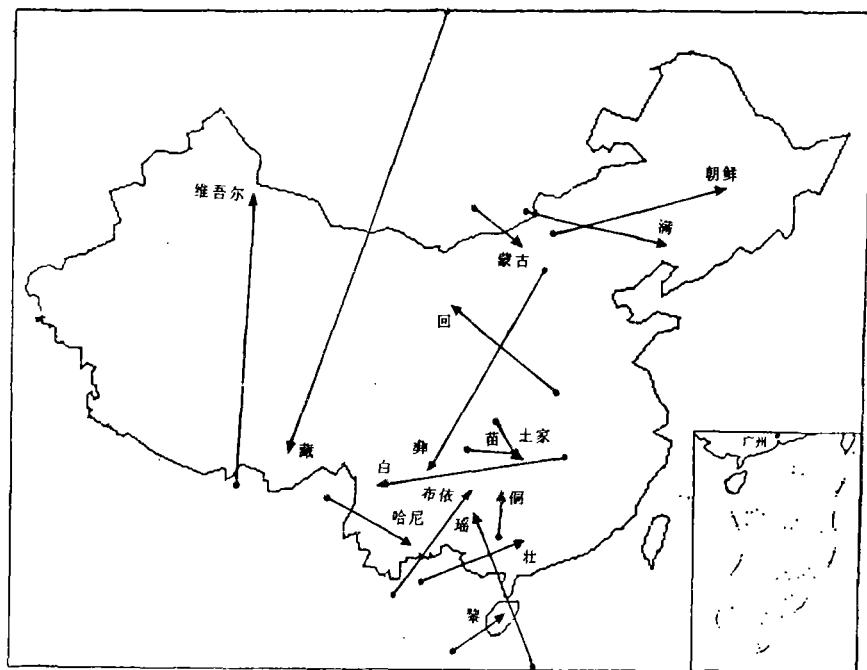


图1 16个少数民族的二维同步排序图
Simultaneous representation of geographic and genetic positions for 16 national minorities

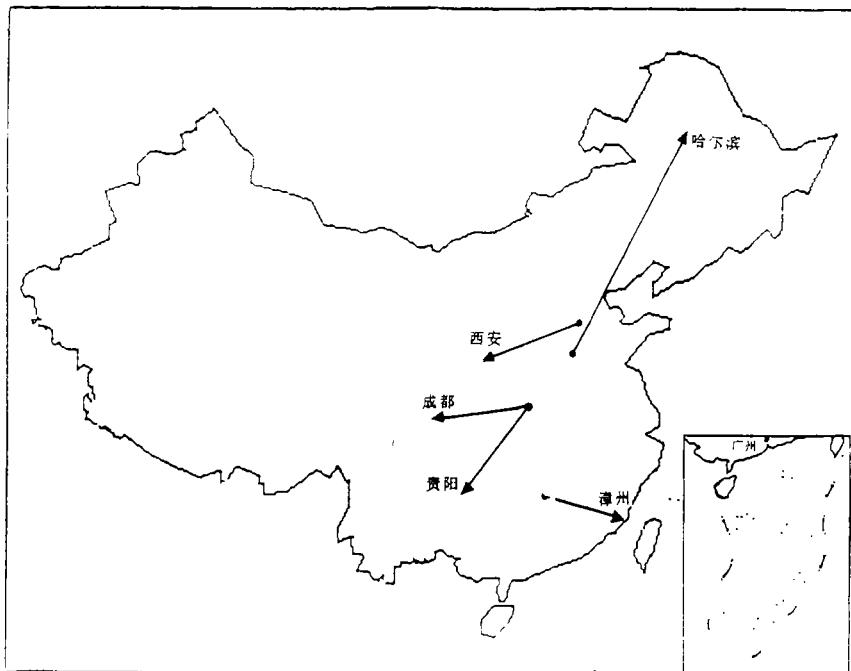


图 2 汉族 5 个人群的二维同步排序图
Simultaneous representation of geographic and genetic positions for 5 Han subpopulations

1、汉族

从图 1 可以清楚地看出, 五个汉族人群所在的地理位置虽十分遥远, 他们的遗传位置却明显地呈现向中原地区聚拢的趋势。这说明各地汉族人群确实有共同的血缘, 而且共同的血缘占有相当大的成份。这共同血缘可以追溯到古代在陕西、山西、河南一带融合许多部落所形成的华夏族。

但不同地区汉族人群的遗传位置却有不同, 西安与哈尔滨两个人群的遗传位置十分接近, 而且位在北方, 说明他们之间在遗传结构上比较接近, 并且蒙古人种北亚类型的血缘占有很大的成分。成都与贵阳两地人群的遗传位置几乎重叠在一起, 位在中央, 即在西安、哈尔滨两个人群遗传位置之南, 漳州人群遗传位置之北, 反映出这两地人群的遗传结构中, 主要是蒙古人种东亚类型的成分。漳州人群的遗传位置在最南方。, 反映出在漳州人群中融入了一定的蒙古人种南亚类型的成分。

赵桐茂等用免疫球蛋白同种异型频率、Lee 等用 HLA 基因频率分析得出的结果都表明南、北汉族遗传结构有明显差异(赵桐茂等, 1987; Lee *et al.*, 1988)。本文的结果除显示南、北汉族有明显差异外, 还暗示成都、贵阳等长江流域汉族应是第三大群, 与南北汉族并列。

从各人群的遗传位置到地理位置间所联的线,反映出这些人群历史上的迁移方向。哈尔滨汉族大多是从山东、河北等地迁移过去的,在黑龙江居住一般只几百年,其遗传结构与中原汉族十分接近。成都与贵阳汉族有相当大部分是从北向南,从东向西迁徙过去的,所以四川汉族与贵阳汉族遗传位置相近当属预料之中。漳州汉族有从中原向南、东迁移过去的汉族血缘,也吸收了相当一部分当地古代越人的血缘,所以他们在遗传结构上与北方的汉族相差较远,与南方的少数民族则比较接近。

2. 少数民族

朝鲜族、满族和蒙古族的遗传位置相互间十分接近,说明这三个民族的遗传结构是十分相似的,存在较密切的血缘关系。而且三个民族的遗传位置多在地理位置之西北,这也反映历史上这三个民族曾住在比今天居住地西边更北的地方,后来才逐渐南迁。(在图 2 中,朝鲜族的居住地是延吉市,不过在实际上延吉市的朝鲜族是两百多年内才逐渐从朝鲜半岛迁去的。)蒙古、满、朝鲜等族在古代的起源地可能比他们在图 2 中的遗传位置更靠北,由于他们已融入汉族等一些其他民族的血缘,因此使他们的遗传位置向南移了。

藏族的遗传位置在最北方,所以藏族无疑起源于北方。至于藏族的遗传位置甚至比蒙古、满、朝鲜等民族的遗传位置还要靠北,并不一定意味着藏族的起源地点比后三个民族在更北方,更可能的是说明藏族在南迁以后,与中原汉族及南方少数民族极少发生融合,而蒙古、满、朝鲜等却融入汉族等其他民族的血缘较多。同时由于藏族在南迁后,一直在青藏高原的高寒气候条件下生活,与寒冷的北方差别不大,因此也看不出自然条件通过选择作用对于藏族这四个位点的基因频率有什么影响。可能正是由于这些原因,使藏族的遗传位置更靠北,即与汉族的差异比蒙古、满、朝鲜等族与汉族的差异更大。

彝族的遗传位置在北方,也说明他们起源于北方,同时彝族在南迁后融入了一定的南方人群的血缘,可能比藏族融入的更多,因此其遗传位置在藏族遗传位置之南也是符合预期的。如果所分析的遗传位点更增多,可能彝族的遗传位置会更向南移一些。

维吾尔族的遗传位置在南方,这也许可以解释为取样地乌鲁木齐市的维吾尔族原来生活在新疆南部,但这种解释是似是而非的,因为据史书记载,在两千多年前,维吾尔族曾生活在贝加尔湖及巴尔喀什湖一带。而且,维吾尔族中还有相当多的白种人血缘成分。这些在图 2 上的维吾尔族的遗传位置都未显示出来。同时,图 2 却显示出维吾尔族在遗传结构上与哈尼族十分相似,这也是无法解释的。总之,用四个位点的基因频率计算得出的二维同步排序图上的遗传位置,并不都是非常合

理和有深刻含义的，因为所用的位点有限，如果对图 2 每一结果都去解释，有时会十分勉强。

回族在图 2 中的遗传位置向中原汉族靠拢，这显示他们融入了相当多的汉族的血缘。据历史记载，元朝以前，尤其是元朝时大批来中国的中亚细亚人、波斯人、阿拉伯人绝大多数未带眷属，他们主要是和汉族女子结婚而繁衍生息的。明洪武五年(公元 1372 年)的诏书还规定：“蒙古色目人现居中国，许与中国人结婚，不许与本类自相嫁娶”。不过回族在图 2 的遗传位置，没有充分显示出他们含有的中亚、波斯、阿拉伯人的血缘。这可能和我们分析基因频率的位点数比较有限有关。

云南白族的遗传位置也显示出他们向汉族靠拢，这也反映出白族中融入了相当多的汉族血缘。与回族相比，白族先民原来是南方的人群，所融入的汉族也大部分是长江流域的汉族，而回族所融入的主要也是北方汉族的血缘，因此白族的遗传位置在回族的遗传位置之南，是符合预期的。

苗族和土家族在图 2 的遗传位置十分靠近，与其地理位置也均相距不远，苗族是古老的民族，三千年以前可能生活在江淮一带，后来逐渐向南、向西迁徙，形成今天的分布。土家族一说来源于川东一带的巴人，一说可能是唐中叶进入湘西的乌蛮的一部分，还有一说认为是湘西土著与巴人、汉人融合而成的。从群体遗传学数据的初步分析来看，苗族和土家族都是在长江流域居住很久的人群，他们不同于蒙古、满、朝鲜等蒙古人种中典型的北亚类型，也不同于壮、黎等典型的南方类型，他们应是东亚类型。

在目前生活在南方的一些人群中，壮、侗、黎、瑶、布依等族的遗传位置都在其地理位置之南，说明他们都起源于南方，均应属于蒙古人种南亚类型。其中布依族和壮族的语言同属壮侗语族壮傣语支，据历史学家分析，认为起源于古代百越中的骆越，他们在图 2 中的遗传位置也十分靠近。侗族语言属壮侗语族中的侗水语支。本文所分析的广西巴马县瑶族讲布努语，可是从根据基因频率数据绘成的二维同步排序图看，瑶族和苗族差异较大。虽然过去一直认为苗瑶关系密切，解放前甚至不少人认为瑶族原来就是苗族，但根据人类群体遗传学数据分析结果，可以初步认为，湖南吉首苗族与广西巴马瑶族差异甚大，苗族应属东亚类型，而瑶族应属南亚类型。黎族语言属壮侗语黎语支。他们肯定是从大陆迁到海南岛的。从人类群体遗传学分析结果看，他们属典型的南亚类型。

参 考 文 献

- 赵树茂等，1987. 免疫球蛋白同种异型Gm因子在四十个中国人群中的分布。人类学学报，6(1):1—9。
- Balakrishnan, V. and L. D. Sanghvi, 1968. Distance between populations on the basis of attribute data. *Biometrics*, 24: 859—865.
- Edwards, A. W. F., 1971. Distances between populations on the basis of gene frequencies. *Biometrics*, 27: 873—881.

- Lalouel, J. M., 1972. Topology of population structure. In: *Genetic Structure of Populations*. Ed. N.E. Morton. 139—152. University of Hawaii Press, Honolulu.
- Lee, T. D., et al., 1988. The polymorphism of HLA antigens in the Chinese. *Tissue Antigens*, 32: 188—208.
- Morton, N. E. and J. M. Lalouel, 1973. Topology of Kinship in Micronesia. *Am. J. Hum. Gen.*, 25: 422—432.
- Morton, N. E., C. Miki, and S. Yee, 1968. Bioassay of population structure under isolation by distance. *Am. J. Hum. Gen.*, 20: 411—419.
- Schonemann, P. H., 1966. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31: 1—10.
- Schonemann, P. H. and R. M. Carroll, 1970. Fitting one matrix to another under choice of a central dilation and a rigid motion. *Psychometrika*, 35: 245—255.

GENETIC-TOPOLOGICAL ANALYSIS OF 21 POPULATIONS IN CHINA

Tan Qian Du Ruofu

(Institute of Genetics, Chinese Academy of Sciences, Beijing 100101)

Key words Genetic topology; Chinese; Population genetics

Abstract

By using the gene frequencies of esterase D, acid phosphorase, glutamate—pyruvate—transaminase and 6—phosphogluconate—dehydrogenase, the genetic topological analysis was carried out on 16 national minorities and 5 subpopulations of the Han ethnic group in China. The dimensional eigenvectorial representations calculated on the basis of genetic distances between 21 populations were fitted to maximum congruence with those calculated on the basis of geographical distances, and the origin, kinship and migration of these populations were discussed.