用 38 个基因座的基因频率计算中国人群间遗传距离

杜若甫 肖春杰 L. L. Cavalli-Sforza①

(中国科学院遗传研究所,北京 100101;①Department of Genetics, Stanford University, CA94305, USA)

摘要 用 38 个基因座的基因频率计算了中国不同省、市、自治区汉族及少数民族相互间的遗传距离,并进行了聚类分析. 结果表明,中国汉族与少数民族都分为南方蒙古人种与北方蒙古人种两大类型,以长江为界. 因此,现代人非洲起源说与本地起源说都必须回答这两大类型是何时分开又如何发展而成的问题. 也确凿地从遗传学角度证明各地汉族与当地少数民族血缘相近,说明各地汉族中已融入有大量当地少数民族血缘,同时,各地少数民族也融入了部分汉族血缘.

关键词 遗传距离 聚类分析 中华民族 现代人起源

遗传学指标最大的优点是其稳定性,遗传物质 DNA 序列中的核苷酸以一定的频率发生变异,所产生的变异有相当大一部分是中性的,有一小部分虽也受环境选择的一定影响,却稳定地世代相传. 因此,遗传指标对研究民族的源与流及民族间的血缘关系,有其特殊的重要性. 用遗传标记的基因频率研究民族的源与流也存在一定的不准确性,主要是因为有些基因有时受外界环境选择作用的影响,另一方面,基因频率有时还会由于随机遗传漂变而发生不定的、不规则的变化. 削弱或消除这种不准确性的最有效方法之一,是用尽可能多的基因座的基因频率进行分析研究,这样,便可使分析得出的结果更接近实际.

中国不少学者曾用基因频率和人体测量学指标、姓氏频率、皮纹学指标等对中国汉族不同人群及少数民族间的遗传距离进行过计算,并根据遗传距离进行了聚类分析,以研究中华民族的源与流,可是,过去最多只用 4 个基因座的基因频率,而且,除用免疫球蛋白同种异型作遗传标记所研究的人群最多之外[1,2],所研究的人群一般也只是一二十个. 所以,这些研究虽然都有一定的意义,但是所利用的基因座太少,所研究的人群也有限[3~8],因此进展不大. 自 1980 年以来人类群体遗传学方面开展了大量的研究,积累了大量的基因频率数据,目前已有 38 个基因座上 130 个等位基因频率的数据可以利用,而且大部分省、市、自治区的汉族人群和大部分少数民族在一个基因座上的数据往往不止一份而是有好几份,甚至多达 29 份. 因此每一人群、每一基因座上的数据的准确性也大大提高了. 但迄今为止,尚无利用已有的全部基因频率数据对中国各民族人群进行遗传距离计算和聚类分析的报道. 本文收集了国内外刊物上有关中国汉族和各少数民族的基因频率,以及本室尚未发表的数据,用美国斯坦福大学 L. L. Cavalli-Sforza 实验室中的 Philip 软件,进行了遗传距离计算和聚类分析.

1996-11-15 收稿

1 材料与方法

1.1 材料

在国内外用人工查阅和计算机检索,共收集到中国人群 2 千多套基因频率数据(1 个人群在 1 个基因座上的基因频率数据算 1 套). 每 1 套数据都经 Hardy-Weinberg 遗传平衡吻合度检测,如 X^2 表明 $p<0.05$,则剔除不用. 最后共有 1 923 套数据可用,它们是 33 个省、市、自治区(包括中国台湾、香港、澳门)的汉族和 54 个少数民族在 38 个基因座上的基因频率. 但各人群有基因频率数据的基因座多少不一,如新疆、西藏、宁夏等地汉族只有 1~3 个基因座的数据,而北京、福建、广东、黑龙江等地汉族有 28~31 个基因座的数据. 55 个少数民族中,门巴族没有基因频率数据,而回、苗、壮、侗、满、瑶、白、黎、蒙古、维吾尔等族有 25~35 个基因座的数据.

用于进行分析的基因座(即遗传标记)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、血清 α-球蛋白、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、和耵聍类型、苯硫脲味觉等. 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.

1.2 统计与分析方法

遗传距离用的是 Fst 遗传距离[9]. 在计算各省、市、自治区的汉族人群以及各少数民族间的遗传距离时,每一基因座的基因频率是该省、市、自治区汉族或该少数民族所有取样人群的基因频率的加权平均值. 例如,湖北汉族 ABO 基因座基因频率有 29 套数据,就将这 29 套数据按其样本人数进行加权平均,得出平均数,作为湖北汉族 ABO 基因座的基因频率.

对汉族的 30 个人群进行了遗传距离计算与聚类分析,即中国 29 个省、市、自治区加上中国台湾,但不包括西藏,因西藏汉族的基因频率数据太少. 对少数民族则分析了 37 个民族,这些民族至少已有 7 个基因座的基因频率数据. 在计算人群或民族间的遗传距离时,凡是双方都有基因频率数据的基因座都利用,如果只有一方在该基因座上有基因频率数据,而另一方没有,则该基因座就不利用.

聚类分析用的是算术不加权平均值两两聚类法(Unweighted pair-group method using arithmatic averages, UPGMA)[10].

2 结果

2.1 汉族人群的聚类分析

根据汉族 30 个省、市、自治区人群的相互间遗传距离绘制的系统树,明显地显示出汉族分为南北两大群. 北方群包括长江以北等省、市、自治区及地跨长江两岸的安徽、江苏两省. 南方群分两支,一支是广东、广西、海南、福建、台湾,我们称之为典型南方人群;另一支包括浙江、江西、湖南、贵州等长江以南其他省和地跨长江两岸的湖北、四川、云南三省(图 1).

2.2 少数民族的聚类分析

根据 37 个少数民族相互间遗传距离绘制的系统树也明显地表明,少数民族可以分为南北

两大群，长江以南的民族为一群，长江以北的民族为另一群(图 2).

在南方群中，傣、德昂、壮、京、侗、黎、瑶等典型南方蒙古人种民族、紧密地聚在一起. 彝族相当一部分分布在长江以北，而且本来起源于北方，却也归属南方群. 同时，语言属彝语支的傈僳、纳西、哈尼、白等族都在南方群. 此外布依、畲、苗、景颇、土家、阿昌等也都在南方群.

在北方群中，新疆的 4 个民族，即哈萨克、塔吉克、维吾尔、柯尔克孜，紧密地聚在一起，并最后与其他北方民族相聚. 藏族在青海、四川等地也有一部分，但大部分在西藏，却明确地属于北方群. 羌族起源于西北地区，现在在四川北部，属北方群. 此外，鄂温克、鄂伦春、达斡尔、满、赫哲、锡伯、蒙古、回、东乡、保安、朝鲜等一直居住在北方的民族，无一例外，都明确地归在北方群.

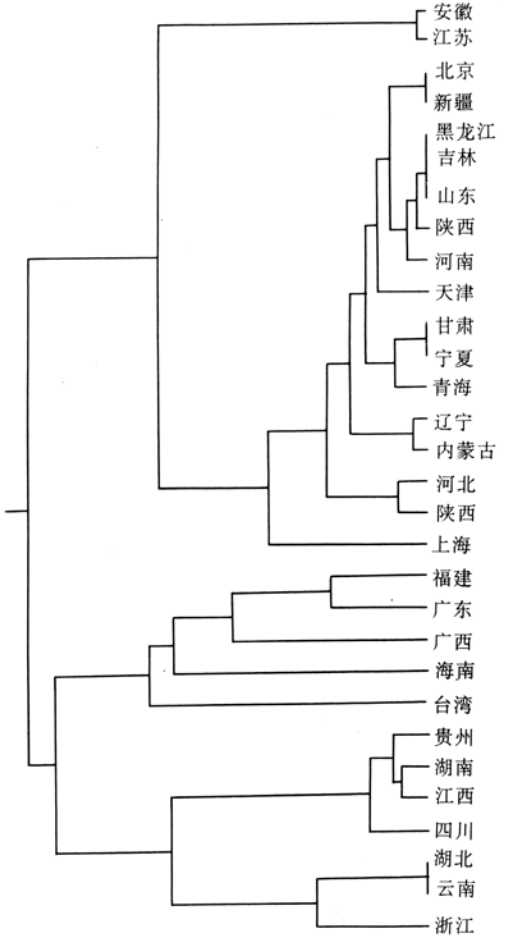

图 1　30 个省、市、自治区汉族人群的聚类图

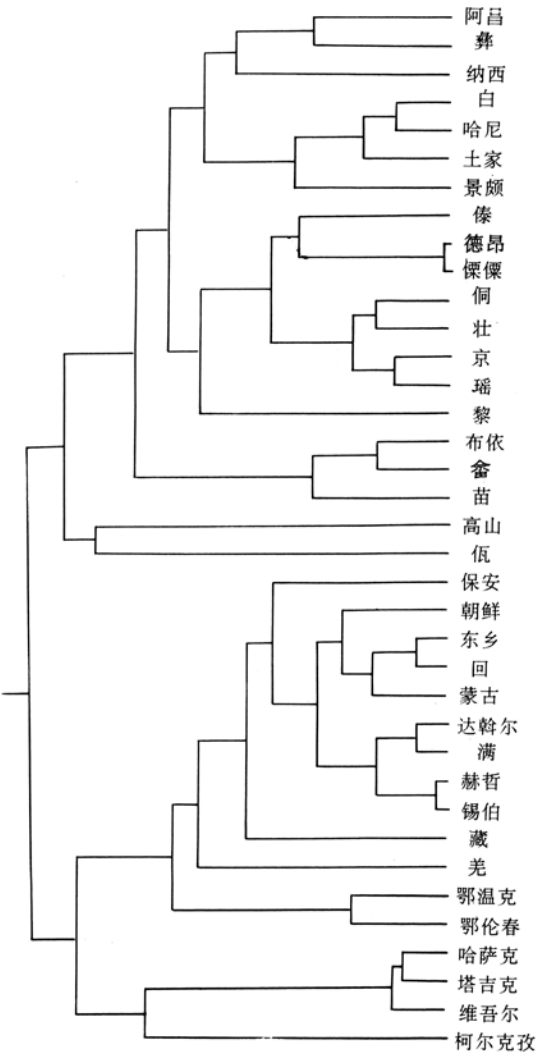

图 2　我国 37 个少数民族的聚类图

2.3　汉族人群与少数民族间的遗传距离

有 10 个汉族人群和 14 个少数民族的基因频率数据较多，均达 27 个基因座以上，我们把这 24 个人群放在一起，计算他们彼此间的遗传距离，结果见表 1.

表 1　10 个汉族人群和 14 个少数民族相互间的遗传距离

	甘肃汉族	黑龙江汉族	吉林汉族	内蒙古汉族	蒙古族	陕西汉族	回族	满族	朝鲜族	藏族	鄂伦春族	维吾尔族	广东汉族	广西汉族	壮族	瑶族	苗族	侗族	黎族	贵州汉族	湖南汉族	四川汉族	彝族
黑龙江汉族	55																						
吉林汉族	60	0																					
内蒙古汉族	42	42	33																				
蒙古族	87	31	69	63																			
陕西汉族	98	88	68	70	129																		
回族	91	75	68	108	88	60																	
满族	92	120	206	106	87	98	167																
朝鲜族	226	102	63	163	101	156	128	45															
藏族	139	259	349	225	230	215	245	96	274														
鄂伦春族	367	460	791	334	356	386	463	318	392	197													
维吾尔族	552	303	1 200	565	341	307	255	472	439	544	849												
广东汉族	427	287	721	371	356	261	291	466	367	590	925	549											
广西汉族	992	636	1 046	1 036	679	495	605	903	852	944	1 425	1 053	49										
壮族	962	755	1 595	961	666	474	612	717	683	801	1 065	775	174	63									
瑶族	487	598	447	620	538	384	426	338	346	433	534	587	155	137	87								
苗族	738	629	1 474	672	535	513	478	727	701	725	1 146	826	249	372	289	355							
侗族	857	656	1 340	751	614	498	486	806	637	890	1 111	711	149	211	149	280	193						
黎族	750	714	555	852	772	366	627	527	792	634	663	953	472	172	148	291	393	224					
贵州汉族	506	478	421	499	519	222	484	726	633	739	830	686	387	129	733	777	679	679	1 010				
湖南汉族	813	552	554	694	613	468	444	880	706	1 095	1626	878	89	177	230	106	263	304	224	31			
四川汉族	297	228	451	235	285	246	268	413	428	513	752	532	136	84	301	338	258	289	407	43	35		
彝族	368	330	635	377	304	157	264	387	423	431	729	545	182	96	245	299	195	216	358	517	92	166	
土家族	285	269	214	255	269	126	174	310	321	417	463	427	200	234	259	329	324	254	248	116	101	187	113

为了更清楚地看出 24 个人群间遗传距离中的某些规律，我们把表 1 中 276 个遗传距离分为 9 种情况，分别计算其平均值与标准误，得出的结果见表 2．从表 2 可以得出：

表 2　汉族人群及少数民族相互间的平均遗传距离

人　群	平均遗传距离($X \pm SE$)
北方汉族人群之间	55.6±8.9
南方汉族人群之间	116.0±33.8
北方少数民族之间	289.9±42.3
南方少数民族之间	250.0±18.0
北方汉族人群与北方少数民族之间	244.7±40.6
南方汉族人群与南方少数民族之间	304.1±38.2
北方汉族人群与南方汉族人群之间	517.4±49.8
北方少数民族与南方少数民族之间	598.9±31.9
北方少数民族与南方汉族人群以及南方少数民族与北方汉族人群之间	653.1±38.8

(1)从表 2 所列平均值可以看出，北方汉族人群间的平均遗传距离(55.6)最近，其次是南方汉族人群之间的平均遗传距离(116.0)，这说明各地汉族间确有相当一部分的共同血缘．

(2)可以看出，北方少数民族间的平均遗传距离(289.9)要比南方少数民族间的(250.0)略大一些，这是因为北方少民族中包含了新疆的民族．而北方汉族人群与北方少数民族间的平均遗传距离(244.7)则比南方汉族人群与南方少数民族之间的(304.1)近．上面已经提到，北方汉族人群间的遗传距离也比南方汉族人群间的近．这可能是因为北方是平原，黄河冬天封冻而且经常改道，因此人口流动比多山的南方容易，几千年来多次因战乱与灾荒造成的人口大迁移与民族大融合，也主要发生在北方．

(3)北方汉族人群与南方汉族人群之间的平均遗传距离(517.4)，要比南方汉族人群间的(116.0)或北方汉族人群间的(55.6)都大得多．同样，南、北方少数民族间的平均遗传距离(598.9)，也要比南方少数民族间的(250.0)或北方少数民族间的(289.9)大得多．这说明无论汉族或少数民族，都可分为南北两大群，和上面汉族及少数民族分别计算遗传距离及聚类的结果相一致．

(4)非常有意义的是：北方汉族与北方少数民族的平均遗传距离(244.7)，和北方少数民族间的差不多(289.9)．南方汉族和南方少数民族间的平均遗传距离(304.1)，也和南方少数民族间(250.0)比较接近．可是南、北方汉族间的平均遗传距离(517.0)和南、北方少数民族间的平均遗传距离(598.9)却大得多．这充分说明，无论在南方还是北方，汉族与当地的少数民族间都已有了许多基因流动，他们的遗传结构已相互接近了．

(5)最大的是南方汉族与北方少数民族以及北方汉族与南方少数民族间的遗传距离(653.1)，因为这些遗传距离同时包涵了南、北两大人群间的差异以及汉族与少数民族间的差异．

2.4　汉族与少数民族的聚类分析

根据表 1 所列的遗传距离，用 UPGMA 法聚类，绘出系统树(图 3)．从图 3 可以看出：

(1)全部 24 个人群明显地分为南、北两大群．北方群包括甘肃、黑龙江、吉林、内蒙、陕西的汉族人群以及蒙古、回、满、朝鲜、藏、鄂伦春、维吾尔等 7 个少数民族．南方群包括广东、广西、贵州、湖南、四川的汉族人群以及壮、瑶、苗、侗、黎、彝、土家等 7 个少数民族．这与上述汉族人群与少数民族分别聚类时的结果是完全一致的．

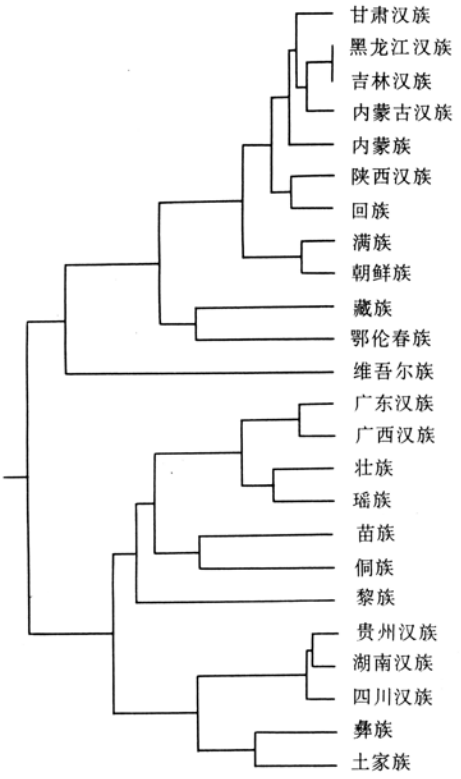

图3 10个汉族人群和14个少数民族的聚类图

(2)在南、北两大群中,往往是地理相近的人群先相聚,而且往往是地理相近的汉族先相聚,然后再与地理相近的少数民族相聚,最后才与地理远的汉族及少数民族人群相聚. 在南方群中,广东、广西汉族与壮、瑶、苗、侗、黎等先相聚,而贵州、湖南、四川汉族则与彝、土家先相聚,然后再聚在一起. 在北方群中,吉林、内蒙、黑龙江、甘肃汉族先相聚,然后与蒙古族相聚,再与已相聚在一起的陕西汉族及回族相聚,然后再与满、朝鲜、藏、鄂伦春等少数民族相聚,最后与维吾尔族相聚.

3 讨论

(1)本文用大量基因频率数据分析证明,今天中国汉族与少数民族都分为南方蒙古人种与北方蒙古人种. 在中国,南方蒙古人种与北方蒙古人种的分界线在哪里? 过去,有人提是北纬30度,虽然北纬30度与长江中、下段相当接近,但是,我们认为,提以长江为界更为合理,因为纬度只是一条直线,而长江是"天堑",是形成人群间隔离的天然屏障.

过去,在文献中曾提出过蒙古人种东亚类型和南亚类型、北亚类型并列. 现在看来,东亚类型并不单独存在,他们只是南亚类型及北亚类型的混合类型,即是南、北两大类型融合而成的.

目前中国人的南、北两大群,实质上是新石器时代时就已存在的南、北两大类型[11]的延续. 于是,可以提出一个问题:亚洲蒙古人种中南、北两大类型是从什么时候起开始存在的? 这自然涉及到现代人的起源问题. 我们认为,无论是现代人非洲起源说或者本地起源说,都必须回答蒙古人种南、北两大类型是什么时候分开然后发展而成的这一问题.

(2)关于中国汉族与少数民族间的融合,已有许多历史学的论述[12,13],本文确凿地从遗传学的角度证明各地汉族与当地少数民族血缘相近,而南、北汉族间血缘却更远. 各地汉族中融入了大量当地少数民族血缘,同时,汉族也有一部分血缘融入了当地的少数民族. 汉族与少数民族是根连根的,这里所说的根是指祖根,即血缘. 所以,中国目前虽然有56个民族,但合在一起称中华民族是完全合乎科学的,因为中华各民族相互间有相当大一部分共同的血缘.

参 考 文 献

1 赵桐茂,张工梁,朱永明,等. 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,1987,6(1):1~9

2 赵桐茂,张工梁,朱永明,等. 中国人免疫球蛋白同种异型的研究:中华民族起源的一个假设. 遗传学报,1991,18(2):

97～108
3 袁义达，杜若甫．中国十七个民族间的遗传距离的初步研究．遗传学报，1983，10(5)：398～405
4 赵桐茂，张工梁，袁义达，等．用 HLA 基因频率计算人群间的遗传距离．人类学学报，1984，3(2)：165～170
5 张振标．现代中国人体质特征及其类型的分析．人类学学报，1988，7(4)：314～323
6 潘　犁，刘祖洞．中国十四个群体中 Gm 和 Km 因子的分布．复旦学报(自然科学版)，1988，27(4)：381～389
7 张海国．肤纹参数在 52 个中国人群中的分布．人类学学报，1988，7(1)：39～45
8 Du R, Yuan Y, Hwang J, et al. Chinese surnames and the genetic differences between north and south China. Journal of Chinese Linguistics, Monograph Series, 1992, (5): 93
9 Cavalli-Sforza L L, Menozzi P, Piazza A. The History and Geography of Human Genes Princeton. New Jersey: Princeton Univ Press, 1994. 535, 8, 518
10 Sneath R H A, Sokal R R. Numerical Taxonomy. San Francisco: W.H. Fruman, 1973. 201～213
11 张振标．中国新石器时代人类遗骸．吴汝康主编．中国远古人类．北京：科学出版社，1989．62～80
12 杜若甫，叶福升．中国的民族．北京：科学出版社，1994．318
13 陈育宁，吴　金．中华民族凝聚力的历史探索．昆明：云南人民出版社，1994．404