暨南大学硕士学位论文

题名（中英对照）：广东东莞汉族 mtDNA 遗传多态性的研究

The research of mtDNA genetic polymorphism in Dong－Guan Han Population

作者姓名：　　成　　峰

指导教师姓名　　王　沙　燕
及学位、职称：　　博　士　　研　究　员

学科、专业名称：　　遗传学　　医学遗传学

论文提交日期：　　2007.05.

论文答辩日期：

答辩委员会主席：

论文评阅人：

学位授予单位和日期：　　暨 南 大 学　　2007.07

独 创 性 声 明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。据我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得暨南大学或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

学位论文作者签名：　　签字日期：　　年　月　日

学位论文版权使用授权书

本学位论文作者完全了解暨南大学有关保留、使用学位论文的规定，有权保留并向国家有关部门或机构送交论文的复印件和磁盘，允许论文被查阅和借阅。本人授权暨南大学可以将学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。

（保密的学位论文在解密后适用本授权书）

学位论文作者签名：　　导师签名：

签字日期：　　年　月　日　　签字日期：　　年　月　日

学位论文作者毕业后去向：

工作单位：　　电话：

通讯地址：　　邮编：

摘　　要

目的：通过研究广东东莞汉族的线粒体 DNA (mtDNA)突变位点，分析其单倍型类群，充实广东汉族线粒体 DNA 基因库。同时通过将东莞汉族线粒体 DNA 单倍型数据同其他地区的汉族以及少数民族的线粒体 DNA 数据比较，了解东莞汉族与其他地区汉族人群之间的关系，了解东莞汉族与其他民族之间的母系遗传关系，为有关广东汉族人群的起源与迁移的历史记载提供遗传学方面的依据。

方法：在知情同意原则下，采用 EDTA 抗凝收集广东东莞汉族三代之内无亲缘关系的 107 名健康个体，抽提其 DNA，建立基因库。对所有个体的线粒体高变区进行测序，通过与线粒体 DNA 标准序列（即剑桥序列，Cambridge Reference Sequence，CRS）进行比较，寻找突变位点，初步划分单倍型类群，再结合编码区的限制性片段多态性、部分序列突变位点等信息，确认样本的线粒体 DNA 进行单倍型分型。参考已经发表的 mtDNA 数据，计算东莞汉族与其他地区汉族和少数民族之间的遗传距离，并进行主成分分析。

结果：获得了东莞汉族线粒体 DNA 高变区序列信息和部分编码区的限制性片段多态性、部分序列数据。通过单倍型划分，发现在东莞汉族群体中存在多种（亚）单倍型类群，比较高频率的单倍型类群是 D、M7、B 和 F1，分别为 18.7%，18.7%，17.8%和 14.0%，而单倍型类群 A、C、G2、Z 的频率很低，分别是 2.8%，1.9%，1.9%和 1.9%。

结论：通过各地区汉族人群之间，以及东莞汉族和其他少数民族的比较，发现广东东莞汉族具有典型的南方汉族人群特点，并能代表广东地区汉族人群的母系遗传结构。通过东莞汉族与少数民族限制性片段多态性单倍型数据的分析，发现广东汉族与百越后裔、苗、瑶族等族关系密切，符合广东汉族是百越后代，在其形成的过程中融合了苗、瑶等少数民族的历史记载。

关键词： 东莞汉族 线粒体 DNA　剑桥序列 单倍型类群　主成分分析

Abstract

Objective: To amplify the mtDNA gene pool of Guang-Dong Han Populations which is a ancient population in Guang-Dong Province, the research the mtDNAs of the Dong-Guan Han Population and analysis the mtDNA haplogroups have been done. Comparing the mtDNA gene pool with other region Han Populations and ethnic Populations to know the Guang-Dong Han Population origins, evolving, migration and the genetic characters.

Methods: Blood samples of 107 unrelated individuals from Dong-Guan Han Population were collected with appropriate informed consent. Genomic DNA was extracted by. Sequencing the HVSⅠ, Ⅱ to find the mutation sites in the HVSⅠ, Ⅱ by comparing with the CRS. Haplotyping the mtDNAs by the mutation sites and other data include the RFLP etc. The genetic distance and principal component analysis have been done with these mtDNA data.

Results: Blood samples of 107 unrelated individuals from Dong-Guan Han Population were collected with appropriate informed consent. The mtDNA mutation sites in HVSⅠ, Ⅱ and coding region have been funded. There are some (sub-) haplogroups existing in the Dong-Guan Han Population. Haploroup D, M7, B and F1 are existing with high frequency. These haplogroups including A, C, G2 and Z have a relative low frequency. By comparing the other regional Han Populations, Dong-Guan Han's maternal genetic construction is similar with the Guang-Zhou Han and Yun-Nan Han Populations. At the same time Dong-Guan Han Population have a proximate genetic relationship between the Yao, Miao and Dai ethnic Groups.

Conclusion: The Guang-Dong Han Populations have a typical south origin mtDNA haplogroup profile by analysis the Han Populations. The maternal genetic structure of Guang-Dong Han Population can be representing by the Dong-Guan Han Population. The Guang-Dong Han population have originate from Bai-Yuei Tribe and gene admixture with other ethic populations, such as Miao populations, Yao populations..

Key Words：Dong－Guan Han；mtDNA；CRS； haplogroup；principal component analysis (PCA)

目　录

1. 前　言

1.1 遗传标记在人类起源与进化研究中的应用

在20世纪下半叶以前，对人类种族起源和迁徙的研究主要采用考古学、语言学、民族学和人类学等方法。这些方法都存在一定的不足。人类的化石十分稀少，古文化遗物也不易得到，材料的欠缺使学者们很难得出肯定的结论，往往只是提出各种假说，以期望将来得到更多的证据证明[1]。而语言学和生活方式可以学习模仿，这种易变性语言学和民族学在研究人群起源和迁徙的问题上也很难得出肯定的结论。因此需要寻找更加稳定有效的研究手段以期补充并完善这些传统学科的发现。而从遗传学的角度来研究民族的源与流，优势就在于基因代代相传的稳定性，得出的结论也较传统方法更可靠[2, 3]。

1.1.1 遗传标记的类型和发展

遗传标记主要分为形态学标记、细胞学标记、生化标记和DNA分子标记4种类型。理想的遗传标记应具备多态性高、遗传稳定、遗传行为简单、能检测整个基因组、不受内外环境影响、操作经济简单等基本特征。

形态学标记是人们最早利用的遗传标记，是生物特定的、肉眼可见的特征特性。由于简单直观易于观察，长期以来，形态学中的体质遗传学得到了广泛的应用。但形态学标记的缺点是遗传表达有时不太稳定，易受环境及基因影响。

细胞学标记和生化标记在人类群体遗传也曾应用过，但它们的应用也非常有限。如ABO血型，最初是利用免疫学的手段来研究[4]。血型也可以用于研究民族之间存在的亲缘关系[5−7]，但是由于已知多态的蛋白质很少、多态的蛋白质等位基因的数目有限且无法获得足够的信息量、检测技术的烦琐等因素，限制了遗传分析工作，这促使人们设法从DNA上寻找分子多态标记[8]。

DNA分子标记大多是以DNA片段电泳谱带形式表现的，依其遗传特性可分为显性和共显性标记2种，依据多态性检测手段可分为以Southern杂交技术为核心的分子标记和以PCR技术为核心的分子标记。根据分子标记在基因组中出现的频率，又可分为低拷贝序列和重复序列标记。

DNA分子标记主要有以下几种：

1 限制性片段多态性 (Restriction Fragment Length Polymorphism, RFLP)是最早

发展的分子标记，至今仍被广泛应用。其基本原理是利用限制性内切酶酶切不同个体基因组DNA后，与同位素或非同位素探针标记杂交，从而显示与探针含同源顺序的酶切片段在长度上的差异。RFLP探针的来源主要是RG克隆和cDNA克隆,其中cDNA探针保守性较强。

2 随机扩增片段长度多态性标记(Random Amplified Polymorphic DNA, RAPD)技术是由Williams等[9]首先创立的一种DNA分子标记技术，利用单一的10个碱基寡核苷酸作为引物，对基因组DNA进行PCR扩增。经琼脂糖凝胶电泳来检测DNA序列多态性。RAPD技术与PCR技术相比有以下特点：(1)RAPD反应一般由1个10bp寡核苷酸组成随机引物，常规PCR需用2个20bp左右的特定设计引物；(2)RAPD反应退火温度低，一般是36℃左右，一方面保证引物与模板DNA的稳定配对，同时允许适当错误配对,提高多态性检出率；(3)常规PCR为特异扩增，而RAPD产物为随机扩增。

3 扩增片段长度多态性 (Amplified Fragment Length Polymorphism, AFLP)是Zeabeau等[10]发明的一项技术,它既有RFLP的可靠性，又有RAPD的方便性。AFLP的基本原理是：通过PCR扩增基因组DNA片段，扩增产物的变性聚丙烯酰胺电泳显示扩增片段长度多态性，其中：引物＝接头＋酶切位点＋2～3个核苷酸。

4 微卫星 (Microsatellite)是指以几个(1～6bp)核苷酸为单位，多次串联重复序列,也称之为简单重复序列 (single sequence repeats, SSR)、短串联重复序列 (short tandem repeats, STR)或简单序列长度多态性 (single sequence length polymorphism, SSLP)。微卫星广泛分布于真核生物基因组中，大约每隔10～50bp就有一个，由于重复次数和重复程度的不完全相同而造成每一个位点的多态性[11]。

5 单核苷酸多态性 (Single Nucleotide Polymorphism, SNP)是指基因组内DNA中某一特定核苷酸在位置上存在置换、插入、缺失等变化，而且其中最少有一种等位基因在群体中突变频率不小于1%。SNP一般以转换为主，颠换与转换之比为1:2[12]。一般认为在人类基因组中平均1000个碱基对中存在一个SNP，在整个人类基因组中大约有3百万个SNPs[13, 14]。SNPs包含了已知DNA多态性的80%，为最常见的DNA变异类型。到2001年初人类基因组计划已鉴定出140万个SNPs[15]。SNP在CpG序列中出现的最为频繁,其标记在人群中只有两种等位型，故亦称双等位标记。在个体中多态信息量比目前常用微卫星标记等多，等位基因型简单，序列长度多态的

信息量少，但SNP二态性高频率稳定遗传的特性弥补了信息量上的不足。由于SNP二态性在检测时只需一个“+/-”或“全或无”的方式，无需像检测微卫星标记那样对片段长度作出测量，有利于发展自动化检测。正是因为SNPs具有高密度、遗传稳定和易于自动化分型等特点，在遗传分析中被广泛应用。

1.2 线粒体 DNA 中的遗传标记

人类基因组中包含 2 种类型的 DNA：细胞核内 DNA 和细胞核外 DNA。人细胞核内 DNA 又可分为：常染色体 DNA (autosome DNA)和性染色体 DNA (sex-chromosome DNA)。人细胞核外 DNA 是线粒体 DNA (mitochondrial DNA，mtDNA)。

常染色体是基因组的主要组成部分，包含了大部分的DNA分子多态。其中常染色中的微卫星多态性是一种广泛应用于研究人类迁移和遗传结构的标记，该标记广泛应用于各大州人群的遗传关系和渊源关系的研究[16, 17]。

人类Y染色体突变率很低，多数位点代表了人类进化史上独特的单一性突变，同一突变在人类进化史上重复发生的概率几乎为零[8]。研究现在人类Y染色体中保存的父系进化历程中发生的突变所形成的多态性，可用于重建父系的进化历史[18, 19]。

人类的线粒体DNA是独立于核基因之外的遗传物质。1981 年Anderson等人[20]首次测定并发表了全长为 16569bp的人mtDNA的全序列，该序列被公认为人mtDNA的标准序列，又称为剑桥序列 (Cambridge Reference Sequence，CRS, GenBank ＃NC_001807)。该序列的发表极大的促进了mtDNA研究。

mtDNA具有以下遗传特性：(1)因mtDNA是人细胞中唯一的核外遗传物质，在受精卵形成过程中只有卵细胞向合子提供mtDNA，所以呈现出严格的母系遗传方式。(2) 由于mtDNA的突变速率是核DNA的 5－10 倍[21]，因此在不同的人群中变异大。(3) mtDNA的多拷贝，使其比核DNA更容易被检测。(4) mtDNA缺乏重组。mtDNA的这些遗传特性使其成为了研究群体系统进化的有效工具。根据mtDNA构建的系统树能够从母系遗传的角度很好的反映人类的迁移历史。

同时由于 mtDNA 在生命活动中的重要作用及其基因组所具有的特点，使得对 mtDNA 的研究在很多的基础学科如细胞遗传学、分子遗传学、发育遗传学、群体

遗传学、人法医学上有重要的意义。

线粒体是真核细胞能量代谢的中心。人组织细胞mtDNA编码 2 种rRNA和 22 种tRNA，还编码 13 种多肽mRNA，这 13 种多肽产物分别参与构成氧化磷酸化(oxidative phosphorylation, OXPHOS)中的复合体Ⅰ、Ⅲ、Ⅳ、Ⅴ。因此mtDNA编码的多肽产物在人组织细胞能量代谢中具有重要作用，同时也可能直接影响细胞的分裂与增殖，研究发现线粒体的功能异常与许多疾病相关。与线粒体功能异常相关的疾病被称为线粒体疾病 (mitochondrial disorders)，如神经肌肉变性疾病、糖尿病、肌阵挛性癫痫和破损性红肌纤维病 (MERRF)、Leber遗传性视神经病 (Leber hereditary optic neuropathy, LHON) 等[22-26]。

在法医学应用方面，已有实验数据表明：4 代之内, 所有母系亲属的mtDNA序列相同，所以 mtDNA分析可以用于个体识别、母系鉴定、亲缘鉴定，尤其是对只有母系亲属的案例进行亲缘关系的鉴定。此外，对细胞核发生明显的转移，检测不到核染色体DNA而细胞质中的一些线粒体仍然存在的陈旧、腐败的检材、毛干、指甲等富含角化细胞的检材，可应用mtDNA多态性分析对此类检材进行个体识别和亲缘鉴定[27]。

在研究人类迁移进化方面，主要通过研究mtDNA单倍型类群 (Haplotype group 或 haplogroup，又称为单倍群或单倍型，是将具有相同特征性突变的mtDNA序列划分到一定的群体)。作为mtDNA系统发育树上主要的单元分支，单倍型类群是由mtDNA序列上特征性突变位点（mtDNA多态性）决定的，这些特征性的突变位点是mtDNA 系统发生的分化单位[28]。人类mtDNA单倍型类群分布表现出显著的地理变化，习惯上归因于遗传漂变，其根本原因是：由于mtDNA的母系遗传特点，原本的序列突变从这个家系建立者时期和其家系后人在地球上不同地理区域定居的过程中被积累，并稳定遗传[29]。因此根据现在的单倍型类群分布可以推测过去人群的起源与迁移模式，群体历史动态等情况。

线粒体 DNA 多态性研究主要经历了限制酶切多态性 (RFLP)、高变区 (HVS) 测序和全序列测序三个阶段。

上个世纪 80 年代初到 90 年代初，由于受到分子生物学技术发展技术中测序技术的限制，RFLP技术是研究者们选择的主要方法，虽然该方法的判断能力有限，仅包含了mtDNA一小部分的信息，但仍然产生了一些非常重要的结果。在提出“非

洲夏娃”学说的文章中，作者使用的方法即是限制性片段长度多态 (RFLP)。Wallace等人利用 9 个PCR扩增片段对线粒体全序进行了扩增，然后运用 14 种限制性内切酶对所扩增的PCR片段进行酶切分析，他们所建立的这一套方法一直沿用至今，成为了一直较为流行的RFLP分析系统[30, 31]。通过运用这个研究系统发现了很多具有大洲特异性的单倍群，为揭示世界各地人群的起源和迁移提供了重要的遗传学信息。

在线粒体DNA的结构上有一个独特的D-Loop环 (Diplacement-loop region)：位于tRNA-Pro和tRNA-Phe 基因之间，由少数碱基构成的一个突出结构。在线粒体DNA上，D- Loop 环是整个线粒体基因组序列和长度变异最大的区域，其进化速度最快，突变速率总体上约为核基因序列的 10 倍。在控制区序列中，大多数突变都发生在两段长度约为 300～400bp的区域[32]，被统称为第Ⅰ高变区(hypervariable segmentⅠ, HVSⅠ, 16024－16383bp) 和第Ⅱ高变区 (hypervariable segmentⅡ, HVSⅡ, 57－372bp)。目前，大多数对于D–环序列的测定都集中在这两个区域[33]，其中HVSⅠ能够提供更多、更有效的信息。随着测序技术的不断提高，20 世纪 90 年代初开始的高变区测序研究大大提高了mtDNA作为遗传标记的分辨率，能够对较近人群之间的亲缘关系做出更准确有效的评价[34]。

HVS极高的突变率同时也带来了负面后果，即增加了回复突变 (recurrent mutation)和平行进化 (parallel evolution)的频率，具有相同序列的mtDNA可能具有不同的来源，因而造成了分子系统学研究上的混乱。但是，由于RFLP和HVS具有很程度的一致性，单倍群的特征RFLP多态位点往往与HVS的某些多态位点关联。综合利用这两种方法能够更精确的描述单倍群之间及单倍群内部mtDNA的系统发生关系[35]，而且通过RFLP特征位点的“筛选”，可以很大程度上消除平行进化和重复突变造成的不利影响。

进入 21 世纪后，一些研究者相继发表了世界各地人群mtDNA全序列的数据[36-40]。这些研究者对mtDNA全序列的分析不仅验证了通过RFLP和HVS得到的结果，而且还为建立高分辨率的世界人群mtDNA系统关系提供了大量的更为有效的信息。

1.3 世界范围内的人群迁移

1987 年在《Nature》发表的文章《Mitochondrial DNA and Human Evolution》[41]对来自世界各地 147 例个体的mtDNA 进行了RFLP数据的分析，结果显示所有个体总共分为 2 支，最古老的支L全部为非洲人所特有，其他大洲和一部分非洲人的mtDNA分布在衍生的另外一支L3 里。他们就此提出了著名的“非洲夏娃”学说，认为现存各大洲mtDNA的共同祖先在 10－20 万年前起源于非洲，在 6－7 万年前迁入亚洲，约在 3－5 年前迁入欧洲，随后(1－3 万年前)从亚洲北部或可能由欧洲迁入美洲。

由 L3 衍生出的所有非洲以外的 mtDNA 谱系分为 M 和 N 两大分支。N 包括了所有的西部欧亚特异的谱系 (H、I、J、K、T、U 等)，东亚特异的 (A、B、R9、N9 等)以及大洋洲特异的 P 谱系。M 分支下游的谱系分布在东非 (M1)，南亚(M1-M6)，大洋洲 (Q/M12)和东亚 (C、D、G、M7、M8、M9 等)的人群中。

1.3.1 亚洲人群的迁移和 mtDNA 变异

1 亚洲人群单倍型分布与迁移

Yao[42], Kivisild[38]等对 7 个东南亚和东亚群体共 153 个个体，其中 54 个藏族个体以及 758 个西伯利亚人的mtDNA 进行了高分辨率RFLP 研究，结果表明，几乎所有的群体都存在 10394 DdeⅠ和 10397 AluⅠ突变。通过是否有 10394 DdeⅠ突变位点判断，所有的单倍型都可分在两个大的聚类群M和N中，其中M聚类群含+10394 DdeⅠ和+10397 AluⅠ突变，包含有C、D、G和E等单倍型类群；另一个聚类群中没有这两个突变，包含A、B和F等类群。A、C 和D单倍型类群在西伯利亚人中检测到的频率最高，分别可达到 68%、84%和 28% [43]。我国西北地区的一些民族群体如维吾尔族、哈萨克族中还存在欧洲单倍型类群H、J、T、W、U2、U4、U5 等，这方面也支持古丝绸之路地区存在广泛的欧亚人群基因交流[44, 45]。

根据文波对 101 个东亚人群 mtDNA 单倍型频率分布进行的分析发现：M 和 N 两大分支各自包括了大约各 50％的东亚 mtDNA，其中 B、D、R9 是最重要的单倍群，A，M7，M8 具有中等频率，G 和 N9 是比较少的单倍群，频率在 5％以下。

现代人走出非洲以后的迁徙，特别是进入东亚的路线还存在着争议。目前普遍认同的看法是经过了 2 条路线：第一条从非洲北部进入地中海东部地区，并以

此作为中转站沿东西走向分别进入中亚和欧洲，该路线被称为“北线”；另外一条路线从东非进入阿拉伯半岛，然后沿着南部海岸线进入南亚、东亚和大洋洲，该路线被称为“南线”[46]。

1.3.2 中国人群 mtDNA 单倍型分布与迁移

对我国民族人群的mtDNA研究始于 1987 年，贺林[47]和俞民彭等[48]对汉族、藏族、回族和维吾尔族mtDNA进行了全序列酶切。随后依照同样的方法，针对不同地区不同民族的群体也进行了研究。

随着全自动DNA序列测定技术的完善，中国各地区人群的HVSⅠ区序列被大量报道，但是这些数据在数目和地域上缺乏代表性，没有构建出一棵分辨率很高的系统发育关系树。2002 年，Yao等[42]通过对中国汉族mtDNA高变区和编码区的研究构建了一个东亚人群特异的聚类树（图 1－3）。同年Kivisild等[38]利用东亚人群的mtDNA数据，也构建了一棵分辨率很高的东亚特异单倍群系统树（图 1－4）。2006 年Kong等[49]通过利用特有单倍型类群的界定位点对庞大的样本库进行了“模式筛选 (motif-search)”，并对新亚单倍型的mtDNA进行了全序测定，修订了东亚人群mtDNA系统发育关系树。

在我国南北方人群中，mtDNA 单倍型分布有明显的差别，A、D、G、M8、M9 和 Y 在北方的频率高于南方，为表示方便，称“北方谱系”。B、M7、N9 和 R9 在南方的频率比较高，称“南方谱系”。北方谱系中的 C、D、M8、M9 都是 M 单倍群的下游分支，M 单倍群只存在于东非、南亚、大洋洲、东亚和中亚群体，而在西部欧亚地区几乎完全缺失，这一分布和“南线”吻合很好。此外南方单倍型群体多样性比较高。

据文波估算[50]，南方谱系的扩张时间将近 3 万年，北方谱系的扩张时间大约在 2 万年左右或更短，且认为“北方谱系”大部分也是来自南线的成分，从而认为：现代人进入东亚南部后，逐渐向北迁徙，随后在北方发生了群体扩张，产生了C、D、M8、M9 等谱系，同时南方单倍群B、R9 和M也发生了扩张。经过群体扩张，初步形成中国南北方人群的mtDNA遗传结构。

1.4 广东地区汉族人群的形成历史和遗传结构研究

1.4.1 广东地区汉族人群形成历史

广东三大民系的语言分别是粤方言(简称粤语)、潮汕方言(简称潮语)和客家方言。根据语言作为划分的标志，广东汉族基本上由广府、潮汕和客家等三大民系组成。广府人主要在珠江三角洲及粤北、粤西等地。潮汕人主要在粤东之东南部即清代潮州府主要辖境。粤西沿海各县市居民所操之雷州话，与潮语接近，同属闽方言，故将其归入潮汕民系。客家人相对分散，其中心在粤东的东北部各县市和粤北山区，在粤中、粤西各县市的山区也有零散的分布[51]。

1 广府民系的出现和发展

广府民系的先民是由粤地土著南越和西瓯（即百越的部分族人）融合于汉族而出现的。广东自古为百越聚居之地。中原华夏族（汉代之后称汉族）入粤约始于西周中期。秦始皇统一岭南后，秦推行“移民实边”政策。这政策得到后来统治者的沿用[52]。从西汉中期至东汉末年，番禺的南越、广信的西瓯等相当部分族人逐渐融合于汉族，成为汉朝的编民。粤方言系粤地土著百越族的语言不断接受汉语的影响并相融合而形成的一种方言，它的出现是越汉融合的标志。现代粤语仍保留着若干古土著百越部落的语音、语法的特点和某些词语，足以证明广府人与古越族的关系[53]。

对广府民系发展起决定性作用的是广东的俚汉融合。在越汉融合之后，史籍出现“俚”的记载。俚是百越未融合于汉族者的后裔。由于各地汉族涌入，促进俚汉文化、生活交流。俚汉融合，始于晋，发展于南朝，隋唐之间进入高潮，而结束于唐初。

经过俚汉融合，土著百越族的遗裔已为数不多。由于大量俚人融合为汉族，广东民族的结构因之发生了根本性的变化：原来占人口少数的汉族一跃而成为占人口的大多数，成为广东人口最多的民族。这种态势即使元明两代瑶族大量入迁亦未能逆转[54]。除粤东之外，俚人基本上都融合到广府民系之中，其对广府民系的作用，主要表现在由于人口的增加、分布范围的扩大和稳定、以及语言的初步形成，这使广府民系成为较为稳定的、自成一体的民系，其分布格局已隐约可见。

2 潮汕民系的出现及其发展

潮汕民系同广府民系一样，其先民系由百越的一支即闽越融合于汉族形成。粤东与毗连的福建省无山川阻隔，兼有水陆之便，因此，其与福建的关系远较其与粤中的关系密切[55]。

粤东闽越融合于汉族很可能在东汉中晚期，比广府民系先民出现的时间稍后。潮汕民系先民出现之后，虽然发展缓慢，但汉末至南北朝都不断有汉族移入。中原和闽南汉族不断迁入，使潮汕民系人口增加，同时也使潮语得到进一步的发展。两宋时期，由于航海业的发达，汉族移民从水陆两路涌入，令潮州户口骤增。元代时，福建汉族继续迁居潮州，他们从福建带来的闽南话，进一步与先本地居民所操的方言相融合，逐步形成后来的潮州话。至此,潮汕民系居广东东部的分布格局已基本形成[54]。

3 客家民系的出现及其发展

客家民系的出现与发展都较前两者简单。客家民系与其他民族(如畲族)或民系都有密切的关系或融合关系，但基本上是集团性移民的结果。客家语言,基本上也是集团性的人群迁徙而形成的“移民集团”的方言。

就迄今所见，客家先民在两晋之交已经出现。西晋末年的连续动乱，引发了以流民为主的起义，数量足以置县的流民在粤东的东北一带立足，在相对封闭的环境中保留着原有的语言和文化，形成最早的纯客家人聚居区。此后，客家民系便以此为生长点，随着中原汉族的不断入迁而发展壮大。

由于客家人进入之时，广府、潮汕两大民系已基本形成东、西分布的态势，留给客家人的，只有粤北、粤东的两片相连的山区地带。因此,这两块相连的地区也就成了广东境内客家的基本分布地[54]。

1.4.2 广东汉族的最终形成

明清二代，广东再次出现民族大融合。世居山区的瑶、壮、畲三族的大多数陆续融合于广府、潮汕和客家民系之中。瑶、壮、畲三族融合于汉族的成分，从地望来看，多数融合于广府民系，少量融合于潮汕民系和客家民系。经过这次民族融合，三大民系由于人口大幅增加、分布范围进一步扩大而最终形成了广东汉

族[54]。

综上所述，广东汉族不是简单的来自全国各地的汉族的复合体，而是以广府民系和潮汕民系的先民为基础或者生长点，然后继续融合百越后裔、世居广东农村的少数民族、迁入的苗、瑶等少数民族，以及不断吸纳或融合全国各地入粤的汉族而逐渐形成的复杂的融合体。

1.4.3 广东地区汉族人群的遗传结构研究

1 广东地区汉族经典遗传标记研究

高雅等对19个不同地区汉族人群9个STR 间的分子遗传学关系的研究结果显示：系统发生树和聚类分析结果与19个汉族人群地理分布基本一致。广东、广西、湖南、成都、宁波、浙江、江苏和湖北均属于长江水系地域而聚为一类，而与深圳、山东、辽南、上海、天津、青岛、福建等的距离较远。深圳虽然地处南方，但外来人口密集，所以其汉族的群体遗传结构与广东本地汉族相差较远。上海、天津、青岛等地均是经济发达的城市，外来人口流动量大，所以其整体遗传结构与其他地区差异较大[56]。

班贵宏等通过对中国 13 个群体(云南汉族、广东汉族、山东汉族、白族、傣族、拉祜族、黎族、纳西族、撒拉族、畲族、土族、佤族和云南藏族）共 577 例无亲缘关系的研究对象的DNA 样本进行MICA 基因微卫星扫描分型，发现：MICA基因微卫星在 13 个群体中的分布存在显著性差异，即使在同一民族的不同群体之间亦有显著性差异，如山东汉族与云南汉族和广东汉族之间。而不同民族之间的差异更为明显，如拉祜族和佤族与其他群体之间存在显著性差异，黎族和畲族也与多数群体之间存在显著性差异[57]。

李焱等对 562 个广东汉族人群HLA-B 位点基因多态性进行了分析，并和中国香港人群的该等位基因频率进行比较，发现：中国广东汉族人群HLA-B 等位基因频率总体分布与中国香港人群差异均无统计学意义 (P> 0.05)，由于地缘接近，这样的结果也在预料之中[58]。

台运春等[59]应用荧光标记复合扩增系统, 对广东地区 10071 名汉族无关个体的 15 个等位基因分布频率进行了调查，并与浙江汉族、青岛汉族 2 个频率资料进行了比较发现：与张幼芳研究[60]的浙江群体资料比较，共 7 个等基因座有显著性