

文章编号:1004-0374(2008)04-0540-09

人类线粒体DNA世系的系统发育关系研究

孔庆鹏^{1,2*}, 张亚平^{1,2}

(1 中国科学院昆明动物研究所遗传资源与进化国家重点实验室, 昆明 650223;

2 云南大学生物资源保护与利用重点实验室, 昆明 650091)

摘要: 本文以人类线粒体DNA为例, 回顾了其系统发育关系的重建的研究历史, 进而总结介绍了该分析方法在人类进化历史研究、线粒体DNA数据质量评估以及疾病相关线粒体DNA突变的甄别等方面的应用, 以期对该方法在国内的推广应用有所裨益。

关键词: 线粒体DNA; 系统发育关系; 世系; 突变

中图分类号: Q75; Q987 **文献标识码:** A

Phylogeny of human mitochondrial DNA lineages and its applications

KONG Qing-peng^{1,2*}, ZHANG Ya-ping^{1,2}

(1 State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650223, China; 2 Laboratory for Conservation and Utilization of Bio-resource,

Yunnan University, Kunming 650091, China)

Abstract: With the special intention to introduce the most widely adopted phylogenetic analysis human mitochondrial DNA (mtDNA) studies, the history of the reconstruction of mtDNA phylogeny was reviewed. And the applications of human mtDNA phylogeny in studying human evolution, estimating the quality of mtDNA data, and distilling the disease-associated mtDNA mutation were then summarized in the present review.

Key words: mitochondrial DNA; phylogeny; lineage; mutation

古人云:“工欲善其事,必先利其器。”显然,恰当的方法能够让人们事半功倍,在生活中如此,在研究中也是如此。以下主要就人类群体遗传学的线粒体DNA研究领域目前已经得到广泛接受和应用的系统发育研究方法的发展及其应用进行概述。

1 为什么要重建人类mtDNA世系的系统发育关系?

线粒体DNA(mitochondrial DNA, mtDNA)由于其自身特点,已经被证明是一个非常有效的用于研究人类起源与进化、追溯群体历史事件的遗传标记(genetic marker)。首先,由于mtDNA具有较高的突变速率(约为核基因的10倍),mtDNA它能够在较短时间内积累变异,从而可以有效地“记载”较为近期发生的人群动态事件。一般说来,当人类群体扩散定居到不同地区后,群体内存在的mtDNA世系(lineage)仍连续不断地积累突变,从而产生群体特有的衍生世系。通过分析古老及衍生世系的关系

及其分布范围与频率等信息,我们可以研究群体之间的关系以及重建群体事件。其次,由于mtDNA不存在重组,过去所发生过的变异均能忠实地连续地遗传下来,在不考虑频发突变(recurrent mutation)影响的情况下,mtDNA上积累的所有变异理论上存在时间上的先后顺序,也就是说,某些突变较为古老,而一些突变则相对年轻。因此,现代人群内的mtDNA世系之间理论上均存在着进化上的联系。当采用一定的技术手段提取到足够的变异信息并结合合理的方法进行分析时,则有可能重建某地区甚至于全世界人群内母系世系的系统发育关系(phylogeny)。

收稿日期:2008-07-01

基金项目:院长奖获得者科研启动专项资金

* 通讯作者: E-mail: kongqp@gmail.com

世系的系统发育关系树(phylogenetic tree, 以下简称“世系树”)的重建使得人们在研究人群关系及其起源等问题时不再受到建立在模型基础上的经典群体遗传学研究方法的限制,因为后者通常很容易受到诸多因素(如模型及参数的选择等)的影响,从而使通过软件等各种“黑箱”操作计算得到的结果也会存在一定的问题;另外,如何合理解释所得到的结果也是人们无法回避的问题。相反,世系树是人群内存在的各种世系的进化关系及其演化历程的最大可能的真实反映,由于mtDNA不存在重组且大部分突变都是选择中性的(neutral),各个世系在其自身演化过程中逐渐积累突变,因而世系间原本已经存在的系统发育关系并不因其所处群体后来所发生的历史动态事件(demography)而受到影响,基于这种世系间存在的进化上的“纽带”使得我们能够通过世系水平详尽探讨群体关系及群体事件等重要问题。然而,需要指出的是,任何事物都有其两面性,真实可靠的世系树当然能够对问题的探讨有帮助,但错误的世系树则会误导我们的认识。因此,如何尽可能地构建真实的或尽可能接近真实的世系树则是人们在研究中首先必须解决的问题。

2 人类mtDNA世系树的重建

2.1 构树方法的变革

通过mtDNA研究人类起源与进化主要始于20世纪80年代^[1,2]。但由于实验技术以及研究方法的限制,人们当时并不能充分地收集并提取出mtDNA中所蕴藏的信息。例如,当时所广泛采用的构树方法,如邻接法(neighbor joining)、最大似然法(maximum likelihood)及最大简约法(maximum parsimony)等,主要是基于种上水平的研究而发展起来的,当将其应用到种下水平(即群体水平)时,其功能则会由于以下几方面的原因而受到限制^[3]:(1)群体内世系的进化并不一定遵循二歧模式(bifurcation),尤其是在一个快速扩张的群体中,其新发生的世系间往往呈现多歧(multifurcate)或星形(star-like)关系;(2)相对种间关系而言,种内群体水平的变异度较低,因而传统构树方法只能获取少量的特征用于分析,这在一定程度上影响了其结果的分辨率及可靠性。尤其是在对现代人群的研究中,当人们仅仅集中分析mtDNA上的某一小段区域或一小部分特征时(由于技术等方面原因的限制,这是当时人们所能采取的最普遍的研究方式),所得到的信息则将更少。此时,采用传统构树方法进行分析,由于对信息的挖

掘严重不足,因而往往会导致许多枝系间的系统发育关系无法确定,且部分进化枝(clade)的支持率(bootstrap)往往会非常低;(3)mtDNA由于具有较高的突变速率,而且存在较多的频发突变,这些突变可能会模糊某些真实的进化事件或进化途径(evolutionary pathway),但这个问题在传统构树方法中却得不到有效的解决;(4)而当采用最大简约法构建系统树时,从许多均可能存在的进化途径中仅强制或主观地选择出一条(即构建一致树“consensus tree”)的做法显然不妥;(5)群体研究所涉及到样本数往往较多,当采用传统方法分析时不但计算时间会非常长,而且对计算机的数据处理能力也有较高的要求。

目前,已有一些关于群体水平研究的方法报道(mismatch analysis^[4]; cladistic analysis^[5,6]),但其中真正获得广泛承认及应用的(尤其是在人类mtDNA研究领域)则是德国汉堡大学的Hans-Jügen Bandelt等^[7-9]发展的中介网络法(mediate network)。该方法充分考虑到并弥补了传统构树方法在种下水平研究时的不足,其构建的网络图允许多重分歧以及网状结构(reticulation)的存在,因而能够最大程度地包含所有最可能存在的进化途径。但是,由于较多均可能的进化途径的存在,使得我们在分析时对于某些世系或节点之间的系统关系并不清楚,因此有必要对于所构建的网络图进行简化(reduce),即将某些不太可能存在的(至少得不到所研究数据支持的)进化途径进行删减,在简化过程中必须充分结合特征位点的权重(weigh)及存在冲突的节点(node)所代表样本的频率(frequency)以及多样性(diversity)或该节点的大致年龄(age estimation)等信息。该方法既可通过人工^[9]亦可通过软件(Network 4.500: <http://www.fluxus-engineering.com/>)进行。由于所得到的网络图能够充分反映出群体内的遗传结构(genetic structure, 包括世系关系及其频率信息),因而非常适合于群体水平的研究。在实际研究工作中,如果涉及到的样本量较大,直接采用手工方式进行网络图构建时工作量则会非常大,尤其当对研究样本的遗传背景(genetic background)一无所知时,在构建过程中可能会陷入困境。我们的经验是首先可以先采用相应的软件构建该样本的网络结构图,之后通过人工手段对其进行校正,尤其是对于某些网状结构的简化或某些进化途径的取舍,都要求研究人员对数据进行详尽的分析并充分结合其他可用的信息(如突变位

点的权重及其他数据信息等), 这些工作软件本身无法完成。

2.2 信息提取方法的变迁及世系树的重建

2.2.1 从低分辨率RFLP到高分辨RFLP

起初, 对mtDNA的研究主要采用低分辨率RFLP(low-resolution restriction fragment length polymorphism), 即通过5 - 6种限制性内切酶(restriction enzyme)对mtDNA全基因组进行酶切检测。由于该方法所能获取的信息量极少, 使得基于该方法的相关研究的一些结果存在较多的问题。鉴于此, 之后的研究^[2]主要采用高分辨率RFLP方法(high-resolution RFLP), 而该方法此后得到了美国Wallace研究小组的进一步发展, 他们通过9对引物重叠扩增mtDNA全基因组, 再采用14种限制性内切酶(*AluI*、*AvaII*、*BamHI*、*DdeI*、*HaeII*、*HaeIII*、*HhaI*、*HinfI*、*HincII*、*HpaI*、*HpaII/MspI*、*MboI*、*RsaI*及*TaqI*)对扩增片段分别进行酶切^[10,11]。相对于前期工作而言, 该方法的确能够提取更多的mtDNA突变信息, 且因此界定了一些位于编码区(coding region)的稳定的多态位点(polymorphism)。在此基础上, Torroni等^[12]将共享有某个或某些稳定突变的所有单倍型(haplotype)的集合定义为单倍型类群(haplogroup)。分析发现, 单倍型类群具有大洲或地理群体特异性, 例如, 类群L0、L1、L2及L3等^[13-18]的分布范围主要局限于非洲大陆, 绝大部分欧洲人归属于类群H、I、J、K、T、U、V、W及X^[19-27], 而类群A、B、R9及M(包括其亚类群M7 - M11和D及G等)则主要分布于亚洲^[28-42], 其中A - D是美洲人群的主要类群^[12,43,44]。显然, 当各地区人群的母系世系构成及其系统发育关系澄清之后, 通过系统发育地理学^[45]的方法研究地区间或内部人群之间的关系及其群体历史动态则成为可能。

由于高分辨率RFLP仅能检测mtDNA上约20%的突变, 基于该方法所建立的单倍型类群划分系统是否反映了各地区mtDNA世系真实的系统发育关系? 且该系统是否能经受来自其他区段(例如控制区“control region”), 甚至于全序列(complete sequence)信息的检验? 这些都是当时一直未能解决的问题。

2.2.2 控制区测序

mtDNA控制区或D-loop区(displacement loop)具有很高的突变速率(约为编码区的10倍), 该区域包含有三个高变区域(hypervariable segment, HVS): HVS-I、HVS-II及HVS-III, 其中HVS-I较另两个区域信息更为丰富, 故而大部分

的人类mtDNA测序研究通常局限于该区域。但是, 由于控制区内突变饱和和效应(saturation effect)及其较多的突变热点(hotspot)等因素的影响使得该区段内进化噪音较多, 从而增加了提取有效信息的难度。尤其是当采用传统构树方法仅基于该区段信息进行世系树构建时, 其结果也因此具有较大的不确定性, 使得相关推论的可信度也打了不少折扣。例如, Vigilant等^[46]试图基于控制区信息通过最大简约法构建现代人类mtDNA系统树, 期望通过该方法得到支持非洲起源的结果。但由于平行突变(parallel mutation)等因素的影响, Vigilant等所得到的系统树有成千上万棵, 且其中许多系统树并不支持非洲起源假说。显然, 控制区内大量平行突变的存在, 模糊了其原有的系统发育信息, 这也使得仅通过控制区信息进行构树或通过单个控制区突变进行单倍型类群划分时存在一定的问题, 尤其是当一部分研究人为地将所得到的片段统一裁剪到某一区域时(出于便于软件分析的目的^[47]), 这种问题则显得愈发突出^[36]。

解决以上问题的一个可行的办法则是同时辅之以编码区信息。由于mtDNA编码区的突变速率远小于控制区, 其内的同质性事件因而也远较控制区少^[48], 从而使得通过分析编码区信息来构建清晰可靠的系统树成为可能。我们已经知道, mtDNA缺乏重组, 这意味着其上所有发生过的变异之间为完全连锁关系(不考虑频发突变的影响), 因此, 理论上通过编码区或控制区信息分别构建的mtDNA世系树应该是一致的。鉴于此, Torroni等^[19]比较了来自同一群体的RFLP信息及控制区突变, 其结果揭示属于同一类群(通过RFLP系统划分)的个体其控制区具有特有的呈单系(monophyletic)分布的突变, 从而表明基于RFLP酶切位点所构建的单倍型类群划分系统得到了控制区信息的支持, 进而能够区分控制区内古老且较为稳定的特征突变(characteristic mutation)和某些个体和类型所特有的稀有突变(rare mutation)。因此, 当仅仅依靠控制区信息对某一个体的类群归属无法作出正确判断时, 结合一定的编码区信息则可能有助于问题的解决^[29,35,37]。

另一个可行的办法则是借助于中介网络法进行分析判断^[7-9]。该方法在引入中介矢量(mediate vector)的辅助下, 将所有可能发生的进化路径通过图例的方式表示出来, 从而有效地界定出了模糊进化路径的变异(即判断出可能发生平行突变的位点)且

避免了人为地将一些可能存在的进化枝或进化途径抛弃的可能。在平行突变界定出来之后,所研究群体内部的世系关系则可能会变得更为清晰。

2.2.3 RFLP与控制区信息的合并 Torroni等^[19]的研究揭示了mtDNA控制区与RFLP信息的相容性,进而,Macaulay等^[26]将这两套系统结合起来对当时已揭示的世界范围内(重点为欧亚地区)的世系进行研究。由于RFLP信息量的限制(仅能检测mtDNA上约20%的突变),有些当时无法细分的类群在控制区信息的辅助下得到了较为妥善的解决。通过结合RFLP及控制区的突变信息,Macaulay等^[26]构建了当时最为完善的人类mtDNA世系树。为了便于今后相关研究的开展,Richards等^[49]及Macaulay等^[26]在Torroni等^[12]提出的类群划分(即类群A、B、C、D)的基础上统一规范了mtDNA世系的划分系统及其命名规则(nomenclature)。对于来自最近共同祖先(the most recent common ancestor, MRCA)并共享有某个或某些特征突变的所有单倍型的集合,即单倍型类群,给予一个特定名称以便和其他类群相区别。具体而言,主要类群由大写的罗马字母表示,而其子类群则根据需要通过交替附加正整数及小写罗马字母予以命名,例如M M7 M7b M7b1等。由于该系统是基于大量样本的RFLP信息及控制区(主要是HVS-I)突变建立的,因而能够较好地揭示现代人类母系基因库(matrilineal gene pool)的世系组成及其相互关系。

Macaulay等^[26]研究工作的重要性还在于他们对传统mtDNA研究方法进行了较大改进。由于高分辨率RFLP及控制区直接测序均有其各自的缺陷:前者工作量大花费较高,且由于受到内切酶识别位点的局限,RFLP能够检测到的变异量十分有限;而后者则无法有效排除频发突变的干扰。理论上,基于编码区的RFLP信息应能够较好地解决基部类群的系统发育关系(当然,由于信息量的不足,类群C与Z和K与U及T与J等之间的关系当时并没有完全清楚),而来自控制区的信息往往对一些新近产生的枝部类群的鉴别有帮助,将两者有效地结合起来则既能充分发挥优势,又能弥补各自的不足。由于世界范围内的世系关系及轮廓大致已经清楚,因而可以在对一个群体进行研究时首先通过直接测序方法获取样本控制区或HVS-I突变信息,通过控制区突变模式(motif)初步判断所研究样本的类群归属,从而选取特异的编码区突变位点通过RFLP检

测或直接测序的方式对之前的推断予以确证。例如,某中国人群内某个体的HVS-I突变模式为16223-16362,则该个体可能属于类群D或者G,由于类群D在中国人群中的分布频率总体大于类群G^[29,35-37,50],我们因而可以先检测5176AluI位点,如果该个体呈现-5176AluI模式,则其可以划分为类群D;如果是+5176AluI,则可能属于类群G,于是可通过检测4831HhaI来确证其是否归属为类群G;如果检测结果为+4831HhaI,则该个体确定属于类群G;如果是-4831HhaI,则该个体可能属于其他类群,因而需要对该个体进行进一步检测以确定其类群归属。通过这种控制区信息初步判断再结合RFLP检测进行证实的方式,可以在较短的时间内通过较少的工作量而达到个体类群识别的目的,因而现已在国际上得到广泛的应用^[29,33-37,39-41,51-53]。

2.2.4 mtDNA全序列与世系树的重建 考虑到mtDNA自身的遗传特性,可以预见当获得的mtDNA上的信息越多时构建出的世系树就会越精确,而全序列的测定是目前人们能够最大程度获取mtDNA中蕴涵的突变信息的唯一途径。随着技术的进步及测序成本的降低,越来越多的研究组将目光转向了mtDNA全基因组研究。mtDNA全序列的出现使得人们重建某地区的世系树成为可能,同时这也为检验通过合并RFLP及控制区信息构建的类群系统是否正确提供了一次绝佳的机会。在此方面,我们近期的全序列研究工作则是一个很好的范例^[30]。为了系统认识东亚母系世系的系统发育关系,我们从已经通过控制区及部分编码区信息初步确定了其系统地位的2000多个中国人样品中挑选出48个代表性个体进行mtDNA全序列测定,结果鉴别了一些新的单倍型类群并对原先通过RFLP及控制区信息构建的类群系统进行了修正和补充。但总的说来,我们的研究结果基本上支持原先的划分系统,因而表明Macaulay等^[26]提出的合并控制区及RFLP信息对于进行类群的划分是正确有效的。

3 mtDNA世系树的应用

“The main interest of intraspecific phylogenies is not in themselves but rather in their applications”^[3]。诚然,人们研究mtDNA的最终目的并不只是重建世系树,而是试图通过重建真实的(或尽可能接近真实的)人类mtDNA世系树来对某地区人群关系、群体历史动态,甚至于甄别与疾病相关或致病性突变等方面进行研究探讨。

3.1 mtDNA世系树与人群起源及进化研究

mtDNA世系树的出现使得通过系统发育地理学的方法研究探讨人类起源、人群关系及其动态历史成为可能^[54]。例如, Yao等^[35]通过mtDNA控制区结合部分编码区信息的方式对来自中国6个地区共263个汉族个体进行了详尽的类群划分, 其结果表明, 不同汉族地理人群间的差异较大, 且观察到类群F1、B及D4的频率分布由南至北呈梯度变化。中国南部由于分布有较多古老类群(如R9、B等)以及一些可能的基部未定世系(即当时无法划分类群的单倍型, 记为M*、N*或R*), 提示现代人可能于旧石器时期(Paleolithic)由南向北迁入并定居东亚(尤其是中国大陆)。类似地, 根据已经重建的东亚^[28,30,34,42]及欧洲^[21-23,25,27,48]世系树信息, 我们通过控制区突变模式选择部分特定编码区位点进行检测, 对来自中国新疆5个少数民族群体(维吾尔、乌兹别克、哈萨克、蒙古及回族)共252个个体进行了详尽而有效的类群划分^[37]。在这些样品中, 除了8个个体的系统地位尚未完全清楚外, 其余所有个体均可明确地归属为类群M及N(包括R)的亚类群, 由于这些亚类群仅构成东亚和欧洲母系基因库(matrilineal gene pool)的一部分, 因而表明中亚实际上是东亚与欧洲的遗传混合之地。而对各个群体内部的世系组成的研究进一步提示, 来自同一地理区域的不同民族群体中的欧洲特有类群频率, 随着群体定居时间的缩短呈现递减趋势, 其中在较早的居民如维吾尔族(426%)和乌兹别克族(414%)中频率最高, 其次为哈萨克族(302%)和具有较近迁移历史的蒙古族(143%)与回族(6.7%), 而来自同一区域的最近迁移而来的汉族人群^[35]中则没有发现欧洲类群。巧合地是, 这种频率分布模式正好与这些民族群体各自的迁移历史相吻合, 因而暗示该地区群体的母系遗传结构中蕴涵着其迁移历史的印记。而对蒙古族及回族群体的进一步分析则表明族源及婚俗习惯分别对各自母系基因库的形成有着重要的作用。

通过以上例子我们可以清楚地认识到, mtDNA世系树已经发展为系统发育地理学研究的前提和基础, 它的出现使得人们能够通过研究世系时间上的(即进化上的联系及历程)和空间上的(即地理分布及频率)联系来探讨人群起源及其演化等重要问题, 从而避免了过去仅仅依靠软件等“黑箱”操作得到的粗略计算值或统计量而进行推测或比较的情况。显然, 后者是无法从世系角度详尽地对于以上问题进

行了解的, 其研究能力因而也受到极大的限制。

3.2 mtDNA世系树与遗传疾病研究

mtDNA世系树不但可以应用于人类起源及演化研究, 而且还有助于在mtDNA遗传疾病研究中对致病性或疾病相关的突变进行鉴别筛选。研究发现, mtDNA上发生的大多数变异都是中性的, 而有少数突变则为有害的, 由于其有害性的不同以及其他相关因素的影响, 具有这些突变的个体可能会呈现不同程度的病症^[11,55,56]。一般说来, 严重有害突变(severely deleterious mutation)会引发多系统的功能紊乱, 从而导致患者产生非常严重的病变。具有这种突变的个体往往在孩童时期就已发生病变, 因而该突变会很快地就被清除掉。而中度有害突变(moderately deleterious mutation)则主要会影响某个或某些组织或器官的正常功能, 其对个体生存的影响较严重有害突变为小。在个体孩童时期可能并未表现出任何症状, 但随着个体年龄的增长, 该突变在其某一组织或器官中所占比重增加并超过一定的阈值(threshold)时, 此时该组织或器官则可能会发生功能紊乱。由于中度有害突变可能会在个体生长中期或晚期表现出其致病性, 因而对个体的生育并未造成太大影响, 其所受到的较弱的选择强度使得该突变(或具有该突变的mtDNA)可能存在于有限的世代之中, 且在一些特殊的情况之下, 该突变可能会在群体中发生一定程度的扩散。轻微有害突变(mildly deleterious mutation)在其他因素的协同作用下可能会在个体生命的后期表达, 对携带个体的正常繁殖能力影响很小, 因而可能会在群体中以多态形式固定下来^[11,55-57]。

对于我们所获取的世系树, 其基部的古老突变由于经受过上万年的选择压力, 因而这些突变几乎都可以认为是中性的; 处于世系树中部的次古老突变因为也经受过长期的选择压力及其他事件, 因而大部分这种突变应该是中性, 而少量则为近中性(这里指轻微有害的); 对于处于世系树顶端枝系的特征突变, 则有可能存在部分的中度有害突变; 至于严重有害突变(如3243突变^[57])则不太可能存在于世系树上, 其往往会以稀有突变的形式存在于不同个体或世系之中。根据以上推论以及其他相关信息, 我们理论上可以对于所报道的mtDNA突变的致病性进行有效地甄别或判断。如果在某一母系遗传疾病患者的mtDNA上发现一个特异的编码区突变, 则可以通过检测与该个体属于同一单倍型类群的正常个体来

初步判断该突变是类群特有的或可能是致病的^[28,35,58]。例如,通过对来自全国各地共3 000个个体的综合分析表明,突变T12338C虽然导致了mtDNA ND5基因起始密码子的丧失(即将翻译起始氨基酸由蛋氨酸“methionine”转变为苏氨酸“threonine”),但该突变作为类群F2的特征突变之一,产生于大致42 000年前。T12338C(即F2类群)广泛分布于中国正常人群之中(虽然频率较低),且目前并没有关于F2类群与mtDNA遗传疾病相关的报道。以上证据表明T12338C不太可能是致病性突变^[31],该结果与关于突变T3308C(导致mtDNA ND1起始密码子丢失^[59])及A8527G(导致mtDNA ATP6起始密码子丢失^[60])的推论是一致的。关于该方面的详细论述,参见Wang等^[58]。

显然,在鉴别与疾病相关的mtDNA突变时,如果没有其他更为直接可信的证据(如进行功能检测等)时,结合现今的系统发育知识进行初步分析判断则是一个行之有效的方法,从而可以尽可能地避免出现一些仓促不严谨的推断或结论^[61]。

3.3 mtDNA世系树在数据质量评估中的应用

过去的研究表明,重建得到的mtDNA世系树对于检测已发表或未发表数据中潜在的错误非常有帮助。Bandelt等^[62]分析了大量已发表数据中出现的错误并将其总结为以下5种主要类型:碱基移位(base shift)、参考偏差(reference bias)、幻影突变(phantom mutation)、碱基误读(base misscoring)及人为重组(artificial recombination)。Bandelt等已经在其论文中通过大量范例详尽地区分论述了这5种错误类型,这里不再赘述。下面主要就如何在分析中通过世系树检测数据中潜在错误进行简要介绍。

要想检测某批数据中是否存在错误,首先,应该将序列中的突变以变异位点(序列标记以修正后的剑桥标准序列为参考^[63,64])形式输出出来。初步观测其内是否有大量稀有或奇特的(通过与已经发表的数据相比较)碱基颠换(transversion)存在。由于在过去所发表的大量的人群mtDNA控制区及编码区全序列信息表明^[16-18,21-25,27,28,30,34,42-44,52,65-82],mtDNA中碱基颠换较转换而言相当稀少,且其中突变为G的颠换最为少见。而在Herrnstadt等^[25]所报道的560条序列中所出现的大部分颠换(包括绝大部分G颠换),经其^[83]重新测序后证明都是错误的。因此,如果在所检测的数据中观察到较多的稀有颠换的存在,则可能含有较多潜在的错误,应该重新检查测序胶图

(electropherogram),甚至重新测序予以确证。然后,在世系树的指导下通过已获取信息将数据中所有个体划分到相应的类群之中。对于已经确定其系统地位的个体:如果某一稀有突变在不同遗传背景(即单倍型类群)中多次出现,则意味着该突变可能是有疑问的^[84];而个体mtDNA的某片段中的突变模式与其另一片段的变异模式相冲突时,则暗示着可能存在人为重组。为此,我们在已构建好的东亚地区mtDNA世系树的基础上,结合中介网络法发展了一种有效检测人为重组的方法,通过该方法我们在Tanaka等^[42]所报道的672条全序列中检测到了12条人为重组产生的序列(作者未发表数据)。对于尚未确定其系统地位的个体:该个体可能属于世系树中没有覆盖到的某一稀有的基部类型;该个体由于突变模式的不完整而导致类群划分失败。前者确实可能存在,因为目前的世界系树是建立在有限全序列数据的基础之上的,它能够较为容易地覆盖某地区(如东亚)分布范围较广且分布频率较高的绝大部分主要世系,但很可能出现某些基部的但分布范围很窄或分布频率很低的世系没有被检测到的情况。随着将来更多全序列数据的出现,世系树则会相应地变得更精细且覆盖面更广,代表性更强,因而出现无法划分的类型的情况则会大大减少。后者突变模式的不完整可能意味着:(1)该个体的部分突变模式与世系树某枝部类群(近似)匹配但却缺乏其基部应该具有的某个(些)特征突变;(2)所检测个体属于世系树中没有体现出来的某一中间类型。前者该个体突变之间的不相容可能是由频发突变或人为错误造成,对于该问题的判断需要重新进行实验或结合突变位点的保守性及类群频率等相关信息进行进一步分析;后者如果该中间类型只有一个代表性个体(通过分析所研究的数据或与其他已发表数据相比较),则其突变的缺失很可能是在试验中或数据处理中人为造成的,需要重新进行实验对该突变予以验证。

在具体分析过程中,不同的人对于如何检测数据中潜在的错误可能会有不同的心得和体会,但无论采用什么样的方法,大家的目标都是为了尽可能减少数据中出现的错误。因此,在数据收集期间,研究人员不但需要在实验过程中严格把关,在数据处理中也应小心谨慎。我们的研究经验表明,对于新得到的数据,首先通过系统发育分析进行自检可以有效减少数据中出现错误的可能,从而可以较好

地保证数据质量,避免在数据发表后出现一些不必要的尴尬。显然,数据质量的关键在于它不仅与研究结果及结论息息相关,甚至于还体现了研究者的工作态度。如果所报道的数据本身充满问题和错误,无论该研究采用多么复杂或花哨的分析方法或论文写作采用多么完美的论证逻辑,人们都有理由怀疑其研究结果及结论的可信度。对于 mtDNA 等缺乏重组的遗传标记的研究而言,研究者们是很幸运的,因为他们可以根据重建的该标记的系统发育关系对数据进行自检,但对于核基因的研究却无法进行类似的质量控制措施。考虑到 mtDNA 研究中已出现的大量错误的现象,我们有理由相信该现象并非 mtDNA 研究领域所独有,只不过是因为 mtDNA 数据中出现的错误能够通过有效的方法检测出来罢了。但是,对于已经报道的核基因数据中的错误情况,到目前为止人们还一无所知,因此,对于核基因研究领域应该采用什么样的措施来确保所获数据的可靠性则是今后人们在研究中必须要解决的问题。对于一项数据质量没有保证的研究,其所得到的关于核基因上某些突变位点的致病性或是是否受到正选择等研究推论都是有理由值得怀疑的。不可否认,随着信息的积累,某些现在看似正确的推测在将来可能会受到置疑,甚至会被完全否定,但发表过的数据却始终客观存在的。从这个角度来说,数据较结论而言也许更为重要,因为它也会直接影响到今后的其他相关研究。因此,如何保证数据质量将是每个研究者必须面对和解决的问题,而发表高质量数据则是研究者的责任和义务。

4 小结

mtDNA 全基因组信息的出现以及研究方法的变革使得世界各主要地理区域人群的母系世系关系日渐清晰,而据此重建的世系树亦已在 mtDNA 相关领域的研究中得到广泛应用,这使得人们能够从世系进化历史的视角研究探讨人类起源与演化等重要问题;能够深入了解 mtDNA 突变的演化历程,从而有助于在遗传性疾病研究中有效鉴别致病性或良性突变;能够有效地甄别 mtDNA 数据中潜在的错误,从而极大地提高数据质量。这种基于系统发育思想的分析方法在人类研究中取得的巨大成功也促进了它在其他研究领域(如家养动物等)的应用,而在研究中人们发展出的一些辅助方法,如匹配(matching)及近似匹配(near-matching)等则有助于对信息较为匮乏或难以获取的数据(如古 DNA 数据等)进行类群划分^[35,36],

从而进一步拓展了其适用范围。

致谢: 本文源自作者博士学位论文《东亚人群线粒体 DNA 系统发育基因组学研究》感谢对原论文提供帮助的所有同事及同仁。

[参 考 文 献]

- [1] Johnson MJ, Wallace DC, Ferris SD, et al. Radiation of human mitochondrial DNA types analyzed by restriction endonuclease cleavage patterns. *J Mol Evol*, 1983, 19: 255-71
- [2] Cann RL, Stoneking M, Wilson AC. Mitochondrial DNA and human evolution. *Nature*, 1987, 325: 31-6
- [3] Posada D, Crandall KA. Intraspecific gene genealogies: trees grafting into networks. *Trends Ecol Evol*, 2001, 16: 37-45
- [4] Rogers AR, Harpending H. Population growth makes waves in the distribution of pairwise genetic differences. *Mol Biol Evol*, 1992, 9: 552-69
- [5] Templeton AR, Crandall KA, Sing CF. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. *Genetics*, 1992, 132: 619-33
- [6] Templeton AR, Sing CF. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. IV. Nested analyses with cladogram uncertainty and recombination. *Genetics*, 1993, 134: 659-69
- [7] Bandelt HJ, Forster P, Rohl A. Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol*, 1999, 16: 37-48
- [8] Bandelt HJ, Forster P, Sykes BC, et al. Mitochondrial portraits of human populations using median networks. *Genetics*, 1995, 141: 743-53
- [9] Bandelt HJ, Macaulay V, Richards M. Median networks: speedy construction and greedy reduction, one simulation, and two case studies from human mtDNA. *Mol Phylogenet Evol*, 2000, 16: 8-28
- [10] Torroni A, Schurr TG, Yang CC, et al. Native American mitochondrial DNA analysis indicates that the Amerind and the Nadene populations were founded by two independent migrations. *Genetics*, 1992, 130: 153-62
- [11] Wallace DC. Mitochondrial DNA variation in human evolution, degenerative disease, and aging. *Am J Hum Genet*, 1995, 57: 201-23
- [12] Torroni A, Schurr TG, Cabell MF, et al. Asian affinities and continental radiation of the four founding native American mtDNAs. *Am J Hum Genet*, 1993, 53: 563-90
- [13] Chen YS, Olckers A, Schurr TG, et al. mtDNA variation in the South African Kung and Khwe-and their genetic relationships to other African populations. *Am J Hum Genet*, 2000, 66: 1362-83
- [14] Chen YS, Torroni A, Excoffier L, et al. Analysis of mtDNA variation in African populations reveals the most ancient of all human continent-specific haplogroups. *Am J Hum Genet*,

- 1995, 57: 133-49
- [15] Watson E, Forster P, Richards M, et al. Mitochondrial footprints of human expansions in Africa. *Am J Hum Genet*, 1997, 61: 691-704
- [16] Behar DM, Vilems R, Soodyall H, et al. The dawn of human matrilineal diversity. *Am J Hum Genet*, 2008, 82: 1130-40
- [17] Kivisild T, Shen P, Wall DP, et al. The role of selection in the evolution of human mitochondrial genomes. *Genetics*, 2006, 172: 373-87
- [18] Torroni A, Achilli A, Macaulay V, et al. Harvesting the fruit of the human mtDNA tree. *Trends Genet*, 2006, 22: 339-45
- [19] Torroni A, Huoponen K, Francalacci P, et al. Classification of European mtDNAs from an analysis of three European populations. *Genetics*, 1996, 144: 1835-50
- [20] Torroni A, Lott MT, Cabell MF, et al. mtDNA and the origin of Caucasians: identification of ancient Caucasian-specific haplogroups, one of which is prone to a recurrent somatic duplication in the D-loop region. *Am J Hum Genet*, 1994, 55: 760-76
- [21] Achilli A, Rengo C, Battaglia V, et al. Saami and berbers—an unexpected mitochondrial DNA link. *Am J Hum Genet*, 2005, 76: 883-6
- [22] Achilli A, Rengo C, Magri C, et al. The molecular dissection of mtDNA haplogroup H confirms that the Franco-Cantabrian glacial refuge was a major source for the European gene pool. *Am J Hum Genet*, 2004, 75: 910-8
- [23] Coble MD, Just RS, O'Callaghan JE, et al. Single nucleotide polymorphisms over the entire mtDNA genome that increase the power of forensic testing in Caucasians. *Int J Legal Med*, 2004, 118: 137-46
- [24] Finnila S, Lehtonen MS, Majamaa K. Phylogenetic network for European mtDNA. *Am J Hum Genet*, 2001, 68: 1475-84
- [25] Herrnstadt C, Elson JL, Fahy E, et al. Reduced-median-network analysis of complete mitochondrial DNA coding-region sequences for the major African, Asian, and European haplogroups. *Am J Hum Genet*, 2002, 70: 1152-71
- [26] Macaulay V, Richards M, Hickey E, et al. The emerging tree of West Eurasian mtDNAs: a synthesis of control-region sequences and RFLPs. *Am J Hum Genet*, 1999, 64: 232-49
- [27] Palanichamy MG, Sun C, Agrawal S, et al. Phylogeny of mitochondrial DNA macrohaplogroup N in India, based on complete sequencing: implications for the peopling of South Asia. *Am J Hum Genet*, 2004, 75: 966-78
- [28] Kong QP, Bandelt HJ, Sun C, et al. Updating the East Asian mtDNA phylogeny: a prerequisite for the identification of pathogenic mutations. *Hum Mol Genet*, 2006, 15: 2076-86
- [29] Kong QP, Yao YG, Liu M, et al. Mitochondrial DNA sequence polymorphisms of five ethnic populations from northern China. *Hum Genet*, 2003, 113: 391-405
- [30] Kong QP, Yao YG, Sun C, et al. Phylogeny of east Asian mitochondrial DNA lineages inferred from complete sequences. *Am J Hum Genet*, 2003, 73: 671-6
- [31] Kong QP, Yao YG, Sun C, et al. Phylogeographic analysis of mitochondrial DNA haplogroup F2 in China reveals T12338C in the initiation codon of the ND5 gene not to be pathogenic. *J Hum Genet*, 2004, 49: 414-23
- [32] Kivisild T, Bamshad MJ, Kaldma K, et al. Deep common ancestry of Indian and Western-Eurasian mitochondrial DNA lineages. *Curr Biol*, 1999, 9: 1331-4
- [33] Kivisild T, Rootsi S, Metspalu M, et al. The genetic heritage of the earliest settlers persists both in Indian tribal and caste populations. *Am J Hum Genet*, 2003, 72: 313-32
- [34] Kivisild T, Tolk HV, Parik J, et al. The emerging limbs and twigs of the East Asian mtDNA tree. *Mol Biol Evol*, 2002, 19: 1737-51
- [35] Yao YG, Kong QP, Bandelt HJ, et al. Phylogeographic differentiation of mitochondrial DNA in Han Chinese. *Am J Hum Genet*, 2002, 70: 635-51
- [36] Yao YG, Kong QP, Man XY, et al. Reconstructing the evolutionary history of China: a caveat about inferences drawn from ancient DNA. *Mol Biol Evol*, 2003, 20: 214-9
- [37] Yao YG, Kong QP, Wang CY, et al. Different matrilineal contributions to genetic structure of ethnic groups in the silk road region in China. *Mol Biol Evol*, 2004, 21: 2265-80
- [38] Yao YG, Nie L, Harpending H, et al. Genetic relationship of Chinese ethnic populations revealed by mtDNA sequence diversity. *Am J Phys Anthropol*, 2002, 118: 63-76
- [39] Wen B, Li H, Gao S, et al. Genetic structure of Hmong-Mien speaking populations in East Asia as revealed by mtDNA lineages. *Mol Biol Evol*, 2005, 22: 725-34
- [40] Wen B, Li H, Lu D, et al. Genetic evidence supports demic diffusion of Han culture. *Nature*, 2004, 431: 302-5
- [41] Wen B, Xie X, Gao S, et al. Analyses of genetic structure of Tibeto-Burman populations reveals sex-biased admixture in southern Tibeto-Burmans. *Am J Hum Genet*, 2004, 74: 856-65
- [42] Tanaka M, Cabrera VM, Gonzalez AM, et al. Mitochondrial genome variation in Eastern Asia and the peopling of Japan. *Genome Res*, 2004, 14: 1832-50
- [43] Achilli A, Perego UA, Bravi CM, et al. The phylogeny of the four pan-American mtDNA haplogroups: implications for evolutionary and disease studies. *PLoS ONE*, 2008, 3: e1764
- [44] Tamm E, Kivisild T, Reidla M, et al. Beringian standstill and spread of native American founders. *PLoS ONE*, 2007, 2: e829
- [45] Avise JC. *Phylogeography: The history and formation of species*[M]. Cambridge: Harvard University Press, 2000
- [46] Vigilant L, Stoneking M, Harpending H, et al. African populations and the evolution of human mitochondrial DNA. *Science*, 1991, 253: 1503-7
- [47] Wang L, Oota H, Saitou N, et al. Genetic structure of a 2500-year-old human population in China and its spatiotemporal changes. *Mol Biol Evol*, 2000, 17: 1396-400
- [48] Finnila S, Hassinen IE, Ala-Kokko L, et al. Phylogenetic network of the mtDNA haplogroup U in Northern Finland based on sequence analysis of the complete coding region by conformation-sensitive gel electrophoresis. *Am J Hum Genet*, 2000, 66: 1017-26
- [49] Richards MB, Macaulay VA, Bandelt HJ, et al. Phylogeography of mitochondrial DNA in Western Europe. *Ann Hum Genet*, 1998, 62: 241-60
- [50] Yao YG, Zhang YP. Phylogeographic analysis of mtDNA variation in four ethnic populations from Yunnan province:

- new data and a reappraisal. *J Hum Genet*, 2002, 47: 311-8
- [51] Loogvali EL, Roostalu U, Malyarchuk BA, et al. Disuniting uniformity: a pied cladistic canvas of mtDNA haplogroup H in Eurasia. *Mol Biol Evol*, 2004, 21: 2012-21
- [52] Hill C, Soares P, Mormina M, et al. A mitochondrial stratigraphy for island Southeast Asia. *Am J Hum Genet*, 2007, 80: 29-43
- [53] Hill C, Soares P, Mormina M, et al. Phylogeography and ethnogenesis of aboriginal Southeast Asians. *Mol Biol Evol*, 2006, 23: 2480-91
- [54] Richards M, Macaulay V. The mitochondrial gene tree comes of age. *Am J Hum Genet*, 2001, 68: 1315-20
- [55] Wallace DC. Mitochondrial DNA sequence variation in human evolution and disease. *Proc Natl Acad Sci USA*, 1994, 91: 8739-46
- [56] Wallace DC, Brown MD, Lott MT. Mitochondrial DNA variation in human evolution and disease. *Gene*, 1999, 238: 211-30
- [57] Torroni A, Campos Y, Rengo C, et al. Mitochondrial DNA haplogroups do not play a role in the variable phenotypic presentation of the A3243G mutation. *Am J Hum Genet*, 2003, 72: 1005-12
- [58] Wang CY, Kong QP, Zhang YP. Application of the phylogenetic analysis in mitochondrial disease study. *Chn Sci Bull*, 2008, In press
- [59] Rocha H, Flores C, Campos Y, et al. About the "pathological" role of the mtDNA T3308C mutation. *Am J Hum Genet*, 1999, 65: 1457-9
- [60] Dubot A, Godinot C, Dumur V, et al. GUG is an efficient initiation codon to translate the human mitochondrial ATP6 gene. *Biochem Biophys Res Commun*, 2004, 313: 687-93
- [61] Bandelt HJ, Achilli A, Kong QP, et al. Low "penetrance" of phylogenetic knowledge in mitochondrial disease studies. *Biochem Biophys Res Commun*, 2005, 333: 122-30
- [62] Bandelt HJ, Lahermo P, Richards M, et al. Detecting errors in mtDNA data by phylogenetic analysis. *Int J Legal Med*, 2001, 115: 64-9
- [63] Anderson S, Bankier AT, Barrell BG, et al. Sequence and organization of the human mitochondrial genome. *Nature*, 1981, 290: 457-65
- [64] Andrews RM, Kubacka I, Chinnery PF, et al. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat Genet*, 1999, 23: 147
- [65] Derenko M, Malyarchuk B, Grzybowski T, et al. Phylogeographic analysis of mitochondrial DNA in Northern Asian populations. *Am J Hum Genet*, 2007, 81: 1025-41
- [66] Fagundes NJ, Kanitz R, Eckert R, et al. Mitochondrial population genomics supports a single pre-Clovis origin with a coastal route for the peopling of the Americas. *Am J Hum Genet*, 2008, 82: 583-92
- [67] Fraumene C, Belle EM, Castri L, et al. High resolution analysis and phylogenetic network construction using complete mtDNA sequences in sardinian genetic isolates. *Mol Biol Evol*, 2006, 23: 2101-11
- [68] Friedlaender J, Schurr T, Gentz F, et al. Expanding south-west pacific mitochondrial haplogroups P and Q. *Mol Biol Evol*, 2005, 22: 1506-17
- [69] Gonder MK, Mortensen HM, Reed FA, et al. Whole-mtDNA genome sequence analysis of ancient African lineages. *Mol Biol Evol*, 2007, 24: 757-68
- [70] Hudjashov G, Kivisild T, Underhill PA, et al. Revealing the prehistoric settlement of Australia by Y chromosome and mtDNA analysis. *Proc Natl Acad Sci USA*, 2007, 104: 8726-30
- [71] Ingman M, Gyllensten U. Mitochondrial genome variation and evolutionary history of Australian and New Guinean aborigines. *Genome Res*, 2003, 13: 1600-6
- [72] Macaulay V, Hill C, Achilli A, et al. Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. *Science*, 2005, 308: 1034-6
- [73] Merriwether DA, Hodgson JA, Friedlaender FR, et al. Ancient mitochondrial M haplogroups identified in the Southwest Pacific. *Proc Natl Acad Sci USA*, 2005, 102: 13034-9
- [74] Soares P, Trejaut JA, Loo JH, et al. Climate change and postglacial human dispersals in Southeast Asia. *Mol Biol Evol*, 2008, 25: 1209-18
- [75] Sun C, Kong QP, Palanichamy MG, et al. The dazzling array of basal branches in the mtDNA macrohaplogroup M from India as inferred from complete genomes. *Mol Biol Evol*, 2006, 23: 683-90
- [76] Thangaraj K, Chaubey G, Kivisild T, et al. Reconstructing the origin of Andaman islanders. *Science*, 2005, 308: 996
- [77] Trejaut JA, Kivisild T, Loo JH, et al. Traces of archaic mitochondrial lineages persist in Austronesian-speaking Formosan populations. *PLoS Biol*, 2005, 3: e247
- [78] Volodko NV, Starikovskaya EB, Mazunin IO, et al. Mitochondrial genome diversity in Arctic Siberians, with particular reference to the evolutionary history of Beringia and Pleistocenic peopling of the Americas. *Am J Hum Genet*, 2008, 82: 1084-100
- [79] Wang CY, Wang HW, Yao YG, et al. Somatic mutations of mitochondrial genome in early stage breast cancer. *Int J Cancer*, 2007, 121: 1253-6
- [80] Ingman M, Kaessmann H, Paabo S, et al. Mitochondrial genome variation and the origin of modern humans. *Nature*, 2000, 408: 708-13
- [81] Derbeneva OA, Sukernik RI, Volodko NV, et al. Analysis of mitochondrial DNA diversity in the Aleuts of the Commander Islands and its implications for the genetic history of Beringia. *Am J Hum Genet*, 2002, 71: 415-21
- [82] Starikovskaya EB, Sukernik RI, Derbeneva OA, et al. Mitochondrial DNA diversity in indigenous populations of the southern extent of Siberia, and the origins of native American haplogroups. *Ann Hum Genet*, 2005, 69: 67-89
- [83] Herrnstadt C, Preston G, Howell N. Errors, phantoms and otherwise, in human mtDNA sequences. *Am J Hum Genet*, 2003, 72: 1585-6
- [84] Yao YG, Macauley V, Kivisild T, et al. To trust or not to trust an idiosyncratic mitochondrial data set. *Am J Hum Genet*, 2003, 72: 1341-6