Eliminating and Assessing Contamination during Ancient DNA Analyses

YUAN Yuan

MOE Key Laboratory of Contemporary Anthropology, School of Life Sciences, Fudan University, Shanghai 200433 China

ABSTRACT: Ancient DNA is recovered from post mortem materials, such as archaeological or historical specimens. Because samples tend to be sparse and highly damaged, ancient DNA is vulnerable to different sources of contamination during the process of manipulation. Thus, the authenticity of the DNA sequences retrieved is crucial to ancient DNA research. Recent advances in high-throughput DNA sequencing have made amplifying low-content ancient DNA molecules possible. However, contaminant DNA is amplified greatly in the mean time, which makes eliminating and assessing contamination more difficult. Authentication based on PCR reactions is also effective in high-throughput DNA sequencing; more approaches for high-throughput DNA sequencing have been developed, such as length distribution, nucleotide misincorporation patterning, nucleotide frequency at and around the ends of DNA fragments, and correlation of coverage with GC content. The level of contamination in the sample can, to some extent, be observed at the heterozygote. Overall, a wide range of methods are necessary to judge whether or not the retrieved DNA is usable.

Key words: ancient DNA; contamination; high-throughput sequencing

控制古 DNA 污染的方法

袁媛

复旦大学 生命科学学院 现代人类学教育部重点实验室，上海，200433

摘要：控制污染是古 DNA 研究的难题之一。当前，高通量测序提高了古 DNA 检测灵敏度的同时使污染分子大量扩增，这使控制和检测古 DNA 污染更加重要。先前的古 DNA 行业标准基于 PCR 反应，所以对现行的古 DNA 研究仍适用。现在针对高通量测序，古 DNA 研究又发展了新的减少和检测污染的手段。目前，证明古 DNA 可靠性的方法，主要有检测片段长度、DNA 的错配模式、片段末端 5'和 3'的碱基频率、测序覆盖率与 GC 含量的相关性。而对于样本的检测前污染，染色体的杂合态分析可以一定程度上判断污染状况。在现阶段的研究条件下，多方面的证据的支持更有利于判定所得的古 DNA 序列是否可靠。

关键字：古 DNA；污染；高通量测序

一直以来，控制污染就是古 DNA 研究的难题之一。如何防止以及检测古 DNA 的污染水平成为古 DNA 研究的重要课题。随着单分子测序时代的到来，古代样本中的 DNA 可以得到充分的扩增，使得我们可以获得大量古代样本序列。但在检测灵敏度提高的同时，污染序列也得到充分扩增，这使得消减污染和甄别污染越来越重要。

在过去的几十年里，基于 PCR 的古 DNA 研究建立了控制古 DNA 污染的标准[1]，主要是从实验室设施、实验器材、所用试剂及操作中，防止外源 DNA 的污染；在实验设计上也采取种种不同于现代样本的实验流程，检验污染的存在。这些方法被证明是有效的，并对现行的基于高通量测序的古 DNA 研究方法同样适用。当前，鉴定古 DNA 特有的分子特征被认为是甄别现代人污染的有效方法，而样本在到达实验者之前的污染(原初污染)控制仍是一个难题。在尼安德特人研究[2]中提出的检测原初污染的方法颇具启发性。然而，当前还没有一种方法可以证明古 DNA 序列是绝对可靠的。不过，多方面的证据更有利于判定获得的古 DNA 序列的可信度。

一、如何控制外源 DNA 的污染

控制污染首先需要先了解外源DNA的来源。外源DNA的来源可大致分为三种：环境中的污染、操作人员的污染和试剂的污染。

1. 环境中的污染

空气中的尘埃以及实验仪器和器材上都

收稿日期：2010 年 6 月 28 日 **修回日期：**2010 年 7 月 14 日 **联系人：**袁媛 yuanyuan.fudan@gmail.com
2010 年 7 月 16 日 http://comonca.org.cn/Abs/2010/015.htm

可能带有DNA甚至细胞，特别是在分子生物学操作密集的实验室中。例如，PCR产物是高浓度的DNA，其产生的气溶胶可以远远超过古代样本的DNA含量。PCR产物很容易在实验室和整个楼层间广泛散播。成功的PCR反应在不到50ml液体中就包含10^{12}-10^{15}个扩增了的DNA分子，而许多可扩增的古代样本中每克只含有10^{5}-10^{6}个拷贝的DNA [3,4]。因此，古DNA实验室必须在空间和材料上与其它分子实验室隔离，最好设置在一个没有任何分子生物学操作的地方。并且，实验者的每日流动只能从古DNA实验室到现代DNA实验室。

通常推荐古DNA实验室需要一些基本的防污染措施，如紫外照射系统、正压空气净化系统和预防积累灰尘的平滑墙角等。

2. 操作人员的污染

从样本出土的那一刻，所有与样本接触的人员都有可能成为污染源，如挖掘样本的考古工作者、在运送和保管过程中与之接触的人员和分子操作实验人员。由于古代样本的表面要经过去污染处理，前两种操作人员引入的污染比较容易控制。但分子实验的操作人员需要采取特殊措施，如需要在超净台内操作、穿着全密封的实验服、实验中尽量不说话等。所有实验人员的DNA还需要与所得DNA序列比对，以排除污染的可能性。如果测得的多个来自不同样本的DNA序列都与实验操作者的DNA序列相同，那么很可能存在操作者的污染。此外，如果研究男性样本的Y染色体，实验人员为女性无疑会大大减少污染的可能[5]。

3. 试剂的污染

实验器材和试剂标签上的无菌并不能保证没有细胞或者核酸。高压蒸汽灭菌不能完全消除短的DNA片段(小于或等于150bp)[1]。在古DNA实验室内对器材和试剂去污染很重要，针对不同的试剂和器材可以用超滤、紫外照射(45W, 72 hr)、HCl (2.5M HCl, 48 hr)、次氯酸钠(50%, 48 hr)等。

二、古 DNA 特殊的实验设计

Willerslev 等在 2005 年针对基于 PCR 的古 DNA 的实验方法建立了一套较完整的实验标准[1]。

1. 空白对照：应该分别设置抽提和 PCR 对照，对照和样本按照 1:5 和 1:1 的比例。
2. 独立重复实验：在另一个实验室做独立重复实验以排除实验室内的污染[6,7]。
3. 克隆和测序：PCR 产物应该通过克隆后测序，以检测损伤、污染和核插入；需要重复抽提和扩增。
4. 古 DNA 特有的分子行为特征：片段的扩增量和片段长度逆相关；核 DNA 和线粒体 DNA 及叶绿体 DNA 符合拷贝数的比例关系。
5. 模板数的定量：当抽提液中的 DNA 很少时，基于碱基替换需要对古 DNA 进行定量。

当测序方法进入到基因组时代，如单分子测序方法454和Solexa，这些传统的方法同样适用，同时针对这些新的方法也发展了相应的防污染措施，如在构建古DNA文库时使用特异的标签：这样，其它离开古DNA实验室的操作才安全且可以和其它样本区分开。如，Briggs等[8]对尼安德特人的基因组DNA进行测序时在3’端使用了特异的4个核苷标签(TGAC)，所得尼人文库的序列末端都带有这四个碱基。

三、如何证明所得 DNA 是古 DNA

从古代样本中获得DNA后，如何证明所得DNA就是古DNA，而不是现代人的污染呢？

2008 年，科学家们在丹尼索瓦洞穴(51°40′ N; 84°68′E，阿尔泰山南西伯利亚地区)发现了一个孩童的指骨，该未知古人被命名为丹人(Denisova hominin)。经年代测定，丹人生活在3~4万千年。Green等人[9]从该样本的30mg骨粉中提取DNA，使用带条形码(barcode)的接头转化成Illumina测序文库。然后使用引物延伸捕获(Primer Extension Capture，PEC)的方法，从全基因组文库中捕获mtDNA片段。捕获的片段在IlluminaGAII平台上双向测序。对于得到的序列，Green等人通过一系列方法证明所得DNA有古代DNA分子特征：检测片段的长度分布(图1)、DNA的错配模式(图2)、片段末端5’和3’的模式(图3)和覆盖率与GC含量的相关性(图4)[10]。结果显示，该丹人样本经过PEC的方

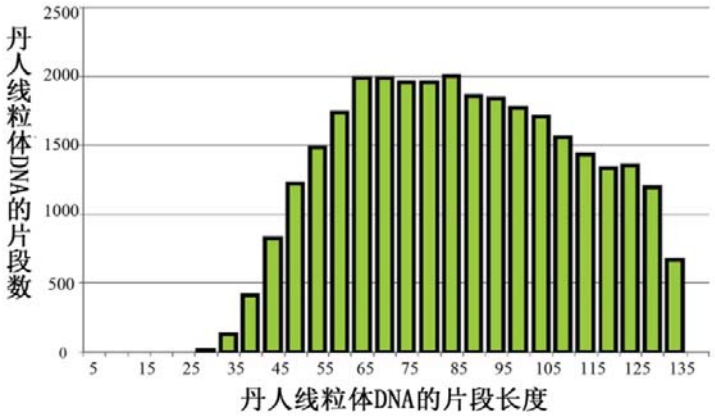

图 1 丹人 mtDNA 片段的长度分布 Fig.1. Length distribution of Denisova mtDNA fragments

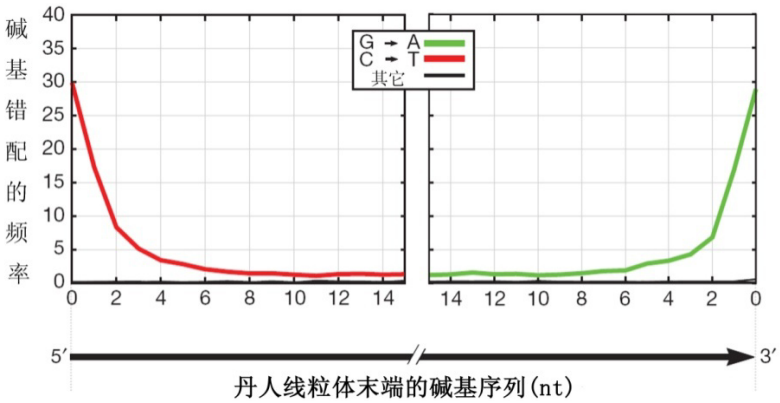

图2 丹人末端的碱基错配频率 Fig.2. Nucleotide frequencies of the aligned reference sequence

法捕获的序列(1)平均长度为85.3 bp；(2)片段的5’末端约30%胞嘧啶残基被胸腺嘧啶取代，同时，3’端也出现相应的取代——腺嘌呤取代鸟嘌呤；(3)DNA片段末端附近的嘌呤(A和G)显著高于其他碱基；(4)CG含量和覆盖率正相关。

另外一篇文章[11]，对一个三万年前的尼人样本的线粒体基因组分析时，也采用了相同的方法。该样本中的片段平均长度为54 bp；约38%的片段5’末端胸腺嘧啶取代胞嘧啶， 3’端腺嘌呤取代鸟嘌呤； 5’末端嘌呤显著较高，3’末端嘧啶显著较高，这表明在嘌呤处容易发生断裂；CG含量和覆盖率正相关。

四、如何检测样本的污染

然而，满足以上分子特征的序列并不能保证是完全可靠的。比如在博物馆存放很多年的样本，污染分子可能看起来也很古老，并且显示相应的分子行为[11]。因此，样本污染通常比实验室污染要难检测得多，这也是人类的古 DNA 材料研究普遍存在的问题[12]。

例如，有些情况下，观察获得序列的长度分布并不能排除存在污染的可能性[2]。图5显示了三个尼安德特人的样本的污染情况[13]。他们内源mtDNA超过80bp的片段的百分比分别为：西班牙El Sidron的尼人，11%；克罗地亚Vindija的尼人，27%；德国Feldhofer的尼人，约37%。

由于通过序列长度估计污染水平的方法只有在污染DNA和内源DNA的降解程度不一样时才有效，而El Sidron的样本中污染片段也很短；Vindija的污染片段有的长有的短；Feldhofer的污染片段的平均长度明显高于内源片段的，但也与内源DNA相互叠加。因此，片段长度并不是一个可靠地污染鉴别标准。

那么如何解决这个问题呢？对于与现代人有明显序列差距的样本，如尼安德特人和丹人，可以通过差异核苷酸外推污染水平，

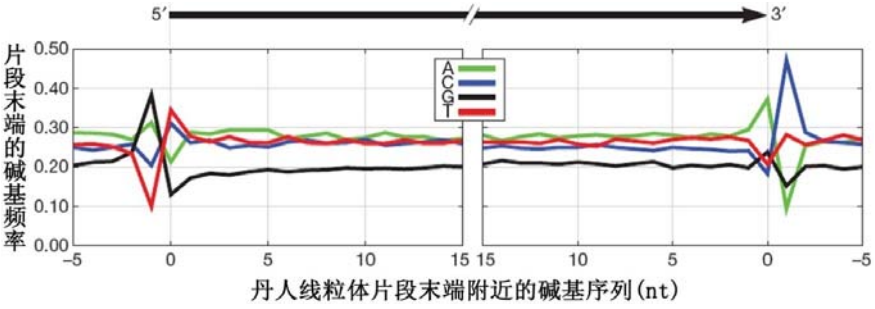

图3 丹人mtDNA片段末端附近的测序结果

Fig.3. At and around the ends of mtDNA fragments sequenced from the Denisova hominin

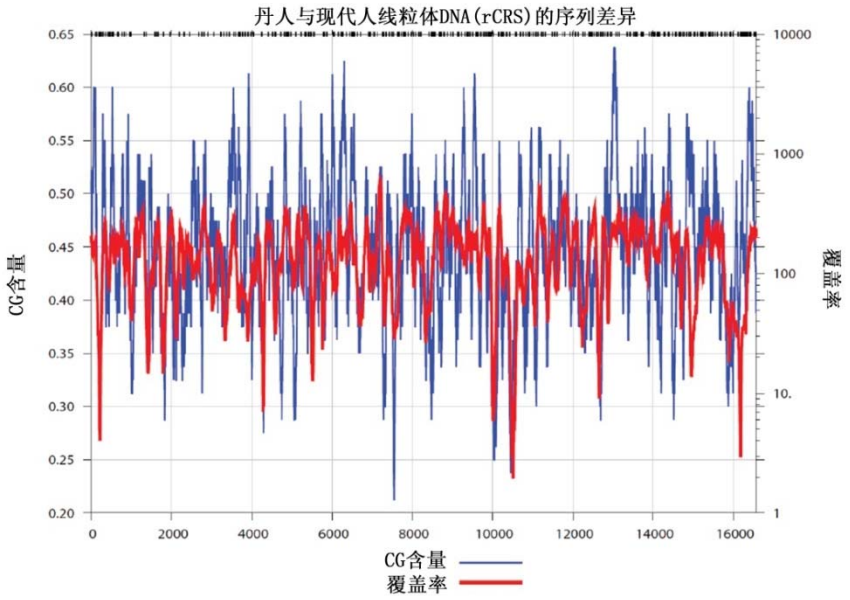

图4 与丹人现代人DNA的参考序列(rCRS)的CG差异(蓝色)

Fig.4. Sequence coverage of the Denisova mtDNA (red, after clustering unique molecules) and GC content (blue) for the complete mtDNA. Above, along the second x-axis, the nucleartide differences along the Denisova mtDNA to a present-day human mtDNA (rCRS) are shown.

然而这需要积累大量的尼人样本序列，鉴别到较固定的尼人和现代人的不同位点后才可行。

Richard 等人[2]讨论了几种颇具启发性的尼人样本检测污染的方法，这些方法涉及到了染色体 DNA 污染水平检测，然而它们也各有局限。以下两种方法可能应用到与现代人差距不太大的古代样本的核染色体污染水平检测。

1. Y 染色体的污染估计

对于女性的样本，可以通过检测Y染色体序列判断男性个体对该样本的污染水平。由于女性不含有Y染色体，那么样本中任何Y染色体的序列肯定都是源于男性个体的污染，通过比较这些污染序列和基因组其它序列的数量，有可能估计在该女性样本中男性个体的污染水平。然而，由于Y染色特含有很多区域和X染色体相同或高度同源，需要避免将X染色体序列误认为Y染色体。同时，也要认识到男性个体的污染并不代表所有外源污染。

2. X 染色体污染的估计

男性样本可以采取类似的策略。因为男性只有一条X染色体，因此，男性DNA样本X染色体的叠加区段应该观察不到杂合性。如果观察到X染色体呈杂合状态，那么必定存在污染。虽然这个策略在理论上很吸引人，但在应用中有几个局限。首先，由于碱基错配

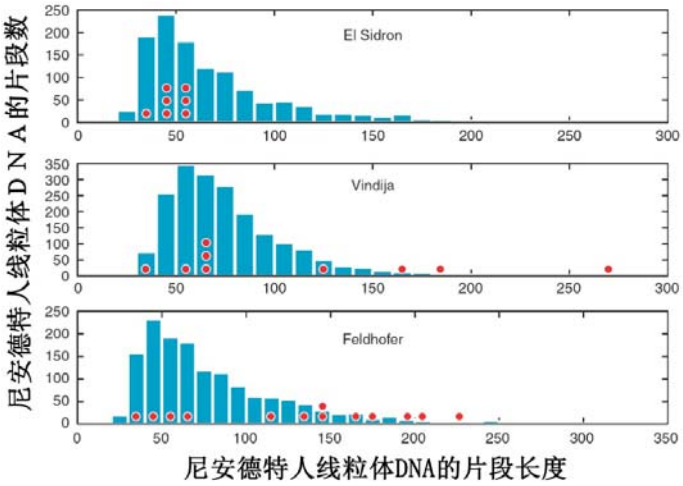

图5 不同尼人样本中的序列长度分布和污染片段的长度。每个红色的点都代表了一个现代人mtDNA的污染。

Fig. 5. Lengths of Neandertal and human mtDNA fragments. Distributions of mtDNA fragments carrying Neandertal diagnostic positions are shown in blue for three Neandertal fossils. Each red dot represents a single contaminating human mtDNA fragment of the indicated length (data from Briggs et al, 2009).

导致的测序的错误在古 DNA 中很普遍，或者机器的误读也会导致出现杂合性，被误认为是污染。其次，许多污染分子和样本的序列相同，因此无法检测。并且，序列在染色体上的错误定位，如一些平行进化的序列也可能被误认为是污染。因此这个方法可操作性相对较弱。

五、展望

高通量测序时代的到来极大地推动了古 DNA 的发展，但同时也带来了挑战。由于种种原因，如古 DNA 的高度降解，长时间的存放等，大多数情况下古代样本包含的并不是单一个体的基因组序列，再加上实验试剂和操作中外源 DNA 的进入，使得检测古 DNA 污染越来越重要。现阶段下，还没有一个单一的验证方法能断定是否存在污染或评估污染的水平，但是多个证据，如 DNA 的片段长度，3’和 5’的碱基错配，女性样本中不含 Y 染色体序列等特点都满足古 DNA 特征时，在现阶段没有更好的方法的情况下，我们基本可以判定所得序列是可靠的。然而，古 DNA 区别现代人 DNA 的特征还有哪些，有没有更可靠易行的检测古 DNA 污染的方法等问题值得更深入的研究，这些问题的解决将使古 DNA 研究在人类学、遗传学、微生物学等领域发挥更大的作用。

致谢

本研究得到复旦大学文科科研推进计划资金支持。

参考文献

1.Willerslev E, Cooper A (2005) Ancient DNA. Proc Biol Sci 272: 3-16.
2.Green RE, Briggs AW, Krause J, Prüfer K, Burbano HA, Siebauer M, Lachmann M, Pääbo S (2009) The Neandertal genome and ancient DNA Authenticity. EMBO J 28: 2494-2502.
3.Handt O, Krings M, Ward RH, Pääbo S (1996) The retrieval of ancient human DNA equences. Am J Hum Genet 59: 368-376.
4.Cooper A, Lalueza-Fox C, Anderson S, Rambaut A, Austin J, Ward R (2001) Complete mitochondrial genome sequences of two extinct moas clarify ratite evolution. Nature 409: 704-707.
5.Li H, Huang Y, Mustavich LF, Zhang F, Tan JZ, Wang LE, Qian J, Gao MH, Jin L (2007) Y chromosomes of prehistoric people along the Yangtze River. Hum Genet 122: 383-388
6.Cooper A, Poinar HN (2000) Ancient DNA: do it right or not at all. Science 289:1139.
7.Willerslev E, Hansen AJ, Poinar HN (2004) Isolation of nucleic acids and cultures from ice and permafrost. Trends Ecol Evol 19:141-147.
8.Briggs AW, Good JM, Green RE, Krause J, Maricic T, Stenzel U, Pääbo S (2009) Primer Extension Capture: Targeted Sequence Retrieval from Heavily Degraded DNA Sources. J Vis Exp 31:1573.
9.Krause J, Fu Q, Good JM, Viola B, Shunkov MV, Derevianko AP, Pääbo S (2010) The complete mitochondrial DNA genome of an unknown hominin from southern Siberia. Nature 464:894-897.
10.Krause J, Fu Q, Good JM, Viola B, Shunkov MV, Derevianko AP, Pääbo S (2010) A Complete mtDNA Genome of an Early Modern Human from Kostenki, Russia. Curr Biol 20:231-236.
11.Handt O, Höss M, Krings M, Pääbo S (1994) Ancient DNA: methodological challenges. Experientia 50:524-529.
12.Serre D, Hofreiter M, Pääbo S (2004) Mutations induced by ancient DNA extracts? Mol Biol Evol 21:1463-1467.
13.Briggs AW, Good JM, Green RE, Krause J, Maricic T, Stenzel U, Lalueza-Fox C, Rudan P, Brajkovic D, Kucan Z, Gusic I, Schmitz R, Doronichev VB, Golovanova LV, de la Rasilla M, Fortea J, Rosas A, Pääbo S (2009) Targeted retrieval and analysis of five Neandertal mtDNA genomes. Science 325:318-321.