Analysis of plastid and mitochondrial DNA insertions in the nucleus (NUPTs and NUMTs) of six plant species: size, relative age and chromosomal localization

六种植物中线粒体假基因和叶绿体假基因的分析：大小、相对年龄和染色体定位

摘要：我们分析了六种测序和拼接完成的六种植物的细胞核中线粒体和叶绿体基因插入片段（其中线粒体插入片段简称为numt，叶绿体的简称为nupt）的长度、相对年龄和染色体定位。我们发现，在这六种植物中，所有的长插入片段比短插入片段有着更低的与细胞器基因组更低的差异性，可以用来判断他们的进化关系。在拟南芥和水稻这些基因组较小的植物中中，大片段的nupt和numt位于染色体着丝粒附近；在大豆和玉米中这些基因组较大的物种中，大片段的nupt和numt也出现在了染色体的其他区域。在拟南芥和高粱中，nupt与numt的分布与转座元件（TEs）的分布有着正相关；而在葡萄和大豆中有着负相关；在水稻和玉米种则无关。我们提出了一个可能的模型：新的叶绿体或线粒体序列先插入着丝粒区域后，通过TE转座元件或异位重组转移至其余位置。转座元件的转座的活性和速度的不同造成了nupt和numt在物种间分布的差异。

关键词：杂乱DNA，nupt，numt，着丝粒，植物

前言

植物细胞包含三种基因组：核基因组、线粒体基因组和质体基因组。在进化过程中，细胞器基因组发生了减少，并且部分DNA片段转移进入细胞核中。一些被称为杂乱DNA的DNA片段仍在三个种基因组发生转移。在片段整合进入核基因组后，由线粒体片段产生的称为numt，叶绿体的称为nupt，并且几乎在所有植物中均有发现。大约18%的拟南芥基因是从蓝藻祖先的质体获得的。对于nupt产生时间、长度、结构、染色体定位等等仅在水稻和拟南芥中进行过系统的分析。在水稻中，大片段nupt几乎都分布在着丝粒区域，但是由于关于它们定位、长度和相对年龄的系统性分析并没有在其他物种中进行，因此普遍规律并没有被揭示出来。

关于DNA从细胞器转移至并整合入核基因组的机制仍没有定论。有研究表明，这种插入是通过NHEJnonhomologous end-joining pathway非同源末端连接介导的双链损伤修复机制形成的。更多物种全基因组序列的可用性使得能够更详细地研究NUPT和NUMT的基因组结构。早期分析的四个物种中，检测到NUMTs的含量和核或线粒体基因组的大小之间没有相关性。然而，最近对几十种物种的系统研究表明，NUPT和NUMT的数量和累积长度与基因组大小和细胞中细胞器的数量相关。差异性最小的NUPT和NUMT具有更大的长度，这意味着原始插入是长片段的并且随时间而变短。然而，许多NUMT的存在可能不是因为插入而是因为插入后重复，大多数长片段NUPT在水稻似乎已在100万年内从核基因组中消除，表明NUPT和NUMTs的存在和积累受到核DNA扩增和缩减的影响。NUPT在玉米近交系之间的染色体定位的巨大变化还表明，质体DNA插入核基因组是经常发生的，并且是近期发生的。除了大片段的NUPT和NUMT，在植物核基因组中也发现了短的最近插入。这些短序列中的一些成簇出现，这可能表明它们在转移之前是有联系的。虽然有的时候，NUPT和NUMTs成为核中的功能基因，但大部分时候是突变的，有的被核基因组删除或加倍了。在这里，我们基于六个序列植物物种的分析发现，大片段的的NUPT和NUMT是最新插入的。它们在小基因组的物种（拟南芥和水稻）中的通常位于着丝粒区域；但是它们在具有更大基因组的物种（大豆和玉米）中更加碎片化并且沿着整个染色体分布。我们推测，转座因子（TE）的动力学与重组率一起决定了混杂DNA的更替和分布模式。

材料和方法

使用blast2.2.25，e值0.01，没有过滤，错配罚分2，窗口9.长度小于100不算。

结论

拟南芥中的nupt通常很短，大多数都小于600bp；numt最长为3kb，位于3号染色体。虽然大多数numt的长度最大值达到2kb，但有16条长度超过2kb的片段，集中在2号染色体的着丝粒附近。拟南芥中所有长numt片段的总长度大于286kb。长numt片段与线粒体DNA的差别小于4%，说明它们可能是近期发生的转移；其余numt片段与线粒体DNA的差别则在20%左右。

水稻中nupt的长度一般大于拟南芥。在10号染色体中，共有5个细胞器基因组片段成簇地集中在一起，长度共为151kb。在4号染色体中，共有14个细胞器基因组片段成簇集中在一起，长度共为143kb。一样的，长片段的nupt通常位于着丝粒附近，并且和细胞器基因组相差很小。最长的numt位于12号染色体中，和其余的21个小片段numt集中在一起，长度共为233kb。类似的numt成簇存在的区域还在1号染色体（总长度28kb），4号染色体（总长度23kb），6号染色体（总长度25kb）和9号染色体（总长度43kb）中发现.

在葡萄中，大多数numt和nupt是小于600bp的小片段，但也存在13kb的nupt和21kb成簇的numt。在高粱中，插入片段之间差异很大，大部分很长，并且基本都分布在着丝粒附近。

在大豆中的nupt通常很短，最长片段为10kb。在大多数染色体中，着丝粒附近的短片段的nupt是减少的，但是长片段的nupt在整个染色体均有分布，仅有少部分在着丝粒附近。最长nupt与叶绿体DNA之间的差异有15%。

在玉米种，大部分numt和nupt小于10kb，但是，许多插入片段也较长，平均值达到34kb（位于2号染色体上）。类似的，numt最长为108kb。9号染色体上numt簇位于着丝粒附近，总长度为225kb。所以的插入均为近期发生，插入片段与线粒体DNA或叶绿体DNA的差异最大为4%。

这些结果说明了，在水稻、葡萄和玉米种，有着类似的numt、nupt的分布和长度差异规律。但是，在拟南芥和高粱中，这些规律是不同的：具有更多不同的NUMT数量较多且相对较近的NUPT数量较少。 总而言之，这些数据表明，这些种类的每种类型的细胞器DNA序列的周转模式可能是特异性的。

我们计算了nupt和numt的总长度，绘制它们与基因组大小之间的相关性图，发现插入总和与基因组大小相关，但与杂乱DNA所占基因组的比例无关。numt和nupt的最大长度和核基因组大小无关。长片段的nupt在小基因组的水稻中存在（151kb），在大基因组的玉米中也存在（61kb）；numt也一样。这些规律表明插入次数而不是其长度影响了基因组中混杂DNA的累积长度，并且插入在较大基因组中更加碎片化。

总之，我们发现在所有6个研究物种中，长片段转移片段与线粒体DNA或叶绿体DNA的差异较小，表明较大的插入片段是最近来源的。 其次，我们揭示了总NUPT / NUMT与基因组大小的相关性，但NUPT / NUMT形成的基因组比例与基因组大小之间无相关性。 第三，我们发现着丝粒是混杂DNA定位的优先插入区，特别是在具有小基因组的物种中。

numt和nupt的分布与转座元件的关系

这些结果表明，在基因组尺寸不同的物种和不同TE含量的物种中观察到类似的混杂DNA分布模式 - 例如在水稻（20-40％覆盖率，图4b）或玉米（70％覆盖率，图4f ）。 似乎混杂DNA的分布模式不依赖于TE在基因组中的丰度或定位，而是反映了转座元件的转座活力。

A putative recombination point was detected downstream of the cox2 pseudogene (pseudo-cox2) in potato mitochondrial DNA (mtDNA).一个可能的重结合位点位于线粒体cox2 假基因的下游。This sequence corresponds to a 32 bp sequence which appears to be well-conserved and is adjacent to the terminals of some mitochondrial genes in Citrullus lanatus, Beta vulgaris and Arabidopsis thaliana and is probably involved in the genic rearrangements.这个序列与拟南芥等生物中的一个保守32bp序列相一致，这个序列邻近于线粒体某些基因的末端，并且可能与基因重排有关。[[1]](#footnote-0)

A Comparative Approach Shows Differences in Patterns of Numt Insertion During Hominoid Evolution

（numts）在真核生物中普遍存在，尽管它们在不同物种间的存在很大的差异。 大多数对于numt进化的了解来自于对单一物种的全基因组分析，最近也有人采用跨越广泛的系统发育距离的不同物种全基因组进行比较。 在这里，我们比较了人与黑猩猩的基因组数据，推断出numt形成的模式和过程。 我们发现了，从黑猩猩和人类分化以来，黑猩猩基因组中新形成的66个numt，这显著大于在人类中观察到的37个。

通过对100余物种numt的分析，我们确定了numt插入，缺失，重复的数量并量化了它们之间的微同源关系。 人类和黑猩猩的大多数numt是通过微同源和NHEJnonhomologous end-joining pathway介导的双链损伤修复机制形成的。人类特有的numt通常整合在转座因子缺少区，但在黑猩猩中没有看到类似现象。另外，系统发育分析表明，线粒体基因组长度至少为500bp，大多为大于1kb。Last, phylogenetic analyses indicate that mitochondrial-numt

alignments must be at least 500 bp, and preferably 1 kb in length, to accurately reconstruct hominoid phylogeny and recover the correct point of numt insertion.

核基因组整合线粒体DNA（numt）是转移并整合进入核基因组的线粒体DNA片段，在动物植物真菌中均有发现报道。它们常被称作“假基因”。

A recombination point is conserved in the mitochondrial genome of higher plant species and located downstream from the cox2 pseudogene in Solanum tuberosum L.

在土豆中，*cox3/sdh4/*pseudo-*cox2* 基因簇的存在已被证实。本实验证明了*sdh4* 和 *cox2* 基因在土豆、大豆、和桃子的线粒体中也是成簇出现的。

Analysis of Nuclear Mitochondrial DNA Segments of Nine Plant Species: Size, Distribution, and Insertion Loci

Young-Joon Ko, Sangsoo Kim等人对九种植物的numt的研究认为，1.在线粒体基因组大的植物中，长numt的比例大于线粒体基因组小的植物；2.全基因组加倍事件增加了短numt的比例；3.numt在外显子区域中更丰富。[[2]](#footnote-1)

前言

大体来说，核DNA的突变率小于线粒体DNA。在基因组测序完成后，完成了在一些物种中对于numt的分析，例如猫，牛，狗，果蝇，大猩猩，蚱蜢，鹅，马，马蹄蝙蝠，蜜蜂，人类，玉米，松鼠和鲸鱼等等。在对鲸鱼numt的分析中，作者对比了6种鲸鱼的numt，并由此建立了6种鲸鱼的进化关系。

结果和讨论

在绘制了基因组大小和线粒体DNA插入大小的相关性图表，表明了核基因组越大，线粒体DNA插入越多。在绿藻中也存在这个现象。numt的数量是植物的特征之一。除了绿藻（C.REINHARDTII AND C.subellipsoidea），由blast结果得出的numt数量，从拟南芥的770个到葡萄的14509个。另外，植物中的numt比鲸类（whale species）中更多。

对numt长度分布的研究：最短长度为25bp，最长为107kb。在物种间也有很大不同。超过70%的numt是短于400bp的。绿藻植物的线粒体基因组很小，超过80%的numt长度小于200bp；葡萄的线粒体基因组比较大，有30%的numt超过1kb；玉米有着最大的核基因组和第二大的线粒体基因组，其numt一半以上是超过5kb的；比玉米基因组稍小的白菜（B.rapa）的numt有着类似于玉米的分布。总之，线粒体基因组小的物种中短numt所占比例更高。对比单子叶植物和双子叶植物，并没有明显的差别。但藻类和陆地植物之间却有明显的差别，这种差别更多的是由于核基因组大小差异造成的。基因组加倍或三倍可能将长的numt分裂开，所以在加倍的植物中长numt减少。在G.max和V.vinifera中也观察到了这个现象。

鲸类平均核基因组大小为2.5Gb，平均线粒体基因组大小为16kb。在鲸类中numt长度大于5kb的比例小于2%，但是在植物中，比例却大于4%，在拟南芥中大于17%。这说明了，线粒体基因组即使很大，加上染色体加倍事件，在植物中长numt片段仍然存在。

下一个分析是关于numt插入位点的区域类型。在陆地植物中，一部分numt插入在基因内部，但在绿藻中，超过70%的numt在非基因区（genic boundaries）。在基因区内的numt，有90%与外显子区域有重叠。这与numt在动物中的分布是相反的：例如在鲸中，有很少的numt位于基因区，其中位于外显子区的就更少了。我们分析了各个物种中，numt在不同位置的分布，发现位于外显子区域的numt明显多于其他位置。鉴于外显子对于生物的重要意义，大量numt插入到外显子区域，可能影响了植物的表型。

关于numt的研究有很多，但并没有关于基因组大小、numt大小、numt大小分布比例、numt插入位置分类等等之间的关系研究。虽然我们的数据显示了它们之间可能的关系，但并不能not enough to infer the biological meaning说明它们就一定有什么生物学意义。进一步研究，还需要更好的测序数据。

Cytogenetic and Sequence Analyses of Mitochondrial DNA Insertions in Nuclear Chromosomes of Maize

本研究检查了多种Zea mays ssp染色体内的mtDNA插入。 mays（玉米）近交系通过使用荧光原位杂交。在9条染色体（9L）的长臂上的相对大的NUMT在四个近交系（B73，M825，HP301和Oh7B）的大致相同的位置被鉴定。对B73和M825相似位置的NUMT的进一步检查表明，这些位点的长片段是由于大部分线粒体基因组的存在；然而，在9号染色体的核基因组B73中仅发现了该NUMT（共252kb）的部分片段。荧光原位杂交分析估计9号染色体B73上的NUMT的大小为〜1.8Mb，并且显示NUMT被甲基化。两个mtDNA区（2.4kb和3.3kb）没有出现在NB物种的B73线粒体基因组中。不过，这些2.4kb和3.3kb片段存在于其他玉米线粒体基因组中，包括一种养殖玉米的祖先Zea mays ssp。 parviglumis中。

Environmental stress increases the entry of cytoplasmic organellar DNA into the nucleus in plants

使用两种不同的转基因烟草品系，我们显示从温和热应激显着增加从叶绿体到细胞核的DNA迁移。此外，我们显示，在诱导的双链断裂的修复期间线粒体DNA片段的插入通过热应激增加。 实验证明，细胞核DNA的核内流是潜在的核基因组突变的来源，其对于温度波动高度敏感，这在温度波动范围内是很好的。

鉴于所研究的环境并不是极端的，这个结果对必须承受环境温度的生物体中的核基因组演化具有深远的影响。 在理想环境中生长的烟草中细胞质细胞DNA的进入核基因组已经被认为是非常高的，并且等于或超过任何其他原因导致突变的速率。 显然，真核生物在高度可变生长条件的现实世界中演变，并且它们必须受到细胞质细胞器DNA整合时带来的一系列影响。

Factors Affecting the Relative Abundance of Nuclear Copies of Mitochondrial DNA (Numts) in Hominoids

尽管线粒体的任何部分都有可能进入细胞核，但线粒体的一些区域，例如线粒体控制区（mitochondrial control region MCR）,很少会进入细胞核。然而，是哪部分线粒体序列性质影响了其转移尚未得到探究。结果表明：在所有三个物种中，HV2和MCRF MCR子域的numts的明显不足。 这些MCR亚结构域有着最高比例的差异率，并且numt的数量最少。

Intracellular mitochondrial DNA transfers to the nucleus in human cancer cells

值得注意的是，数以千计的人类基因组的调查显示，体细胞染色体重排通常与转移到细胞核中的线粒体DNA（mtDNA）片段组合。高转移率和集中断点的特征为理解这些事件潜在机制提供了线索，并提供了转移到细胞核中的mtDNA片段的作用的见解。在这篇文章中，我讨论了我们目前对线粒体DNA体细胞核转移到人类癌细胞核基因组中的了解。

Mammalian NUMT insertion is non-random

由于人类全基因组序列已经完成，许多生物信息学的研究已经确定了可能的NUMT，并且从试图找出NUMT形成中涉及的因素。这些研究得出结论，NUMT是线粒体基因组随机选择的区域。关于NUMT的核插入位点的共识较少 - 以前的研究已经讨论了反转录转座子的可能作用，但是最近有一些报道它们之间没有关联，甚至在NUMT位点和反转录转座子之间负相关性。这些研究通常使用BLAST和默认参数定义了NUMT。我们仔细考虑并重新设定了参数，形成了人类NUMT集合。我们发现NUMT的插入点具有DNA高曲率的强烈倾向，通常发生在实验定义的开放染色质区域中，并且经常紧邻A + T寡聚体发生。

我们还找到了了明确的证据表明，他们的侧翼区确实富含反转录转座子。

最后，我们显示线粒体基因组D环的一部分在灵长类动物进化中作为NUMT的来源是不足的。

BLAST参数设置

Instead, we used the scoring scheme of +1 for matches, -1 for mismatches, 7 for gap-open penalty and 1 for gap-extension penalty, which is suitable for detecting distant homology (24).

关于E值取值

我们得出结论，10-3应该是一个安全的门槛，几乎没有错误的风险。事实上，当在做线粒体和核基因组的匹配时，E值在10-3和10-4之间没有发现匹配，所以即使我们的NUMTs集合中最差的E值（10-3）应该被认为是结果非常显着的。

对于哺乳动物中numt插入区域的总结

（1）NUMTs倾向于插入反转录转座子附近，但不偏好反转录转座子的方向。

（2）NUMTs往往不能插入反转录转座子内。

（3）NUMTs倾向于插入具有高局部DNA曲率和/或弯曲性的区域。

（4）NUMTs倾向于插入具有高A + T的低聚物的区域，特别是TAT。

（5）NUMTs倾向于插入开放的染色质区域。

线粒体假基因研究综述

首例关于线粒体基因转入核的报道是Du Buy和Riley(1967)的研究，他们纯化小鼠线粒体基因并将之与核基因进行杂交，证实两者间具有同源序列，说明小鼠中存在线粒体基因转人核的现象。1994年Lopez等在家猫核染色体中发现7．9 kb的线粒体基因片段的插入，这些片段在转人核后失去了功能，Lopez等将之定义为线粒体假基因(nuclear mitochondrial psedogene)，简称Nutms。之后，Numts被证实在生物中普遍存在，目前已证实含有Numts序列的物种已超过80种，涵盖了所有研究较多的生物类群(Bensasson et a1．2001；Hazkani—Covo et a1．2010)。Numts在不同生物类群中的分布并不均匀，甚至在同一物种的不同种群都有差异，表现出物种、谱系、种群特异性(species一，lineage一，population—specificity)(Song et a1．2008)。随着基因组研究的发展，研究者开始探讨Numts这种特异性分布的机理。

Correlation between Nuclear Plastid DNA Abundance and Plastid Number Supports the Limited Transfer Window Hypothesis

核基因组中核质体DNA样序列（NUPT）的丰度可能有很大差异；然而，对产生这种现象的原因知之甚少。“有限转移窗口假说”假设，每个细胞只有一个质体的物种的NUPT数量将少于每个细胞的质体数量较较多的物种，但缺乏来自单一种质的物种的基因组序列数据使得该假设难以测试。在这篇文章中，通过分析来自不同单倍和多形态分类单元的新获得的基因组序列，我们证明了假说的情况。平均来说，我们多质体物种中的numt比单质体物种中多80倍。此外，NUPT含量与核基因组大小正相关，表明除了质体数量外，NUPT受到控制非编码核DNA扩增和收缩的力的影响。这些发现与线粒体起源（NUMT）核DNA数据一致，表明类似的过程决定了了NUPT和NUMTs的丰度。

Mitochondrial pseudogenes: evolution’s misplaced witnesses

最近有证据表明，numt的丰度在各个物种中是不同的，例如在植物中numt会比在动物中常见，在人类中也比在果蝇中多得多。

第一个线粒体DNA整合进入核基因组的报道是在发现细胞器也存在遗传物质后不久的1967年。1 Du Buy, H.G. and Riley, F.L. (1967) Hybridization between the nuclear and kinetoplast DNA’s of Leishmania enriettii and between nuclear and mitochondrial DNA’s of mouse liver. Proc. Natl. Acad. Sci. U. S. A. 57, 790–797

BLAST参数

默认值，E值0.0001

Mobile DNA elements in the generation of diversity and complexity in the brain

移动元件是可以改变其在基因组内的位置（逆转录）的DNA序列。尽管其生物学功能在很大程度上未得到证实，但是来自移动元件的DNA包含了近一半的人类基因组。 长期以来，神经元基因组被认为是不变的；然而，最近的研究已经证明，移动元件在神经发生过程中主动逆转，从而在神经元之间产生基因组多样性。 此外，数据的不断增加表明，移动元件在某些神经系统障碍（包括Rett综合征和精神分裂症）中被错配。

Molecular Poltergeists: Mitochondrial DNA Copies (numts) in Sequenced Nuclear Genomes

线粒体DNA向核基因组的转移产生的片段（numt），是基因组序列证实的持续进化过程。 在人类中，五种不同的numt引起遗传性疾病。 在实验室和自然界中，线粒体DNA片段通过双链断裂（DSB）的非均质末端连接（NHEJ）进入核DNA。 85个真核生物基因组中的数字插入次数显示，numt含量与基因组大小密切相关，numt插入率可能受到DSB频率的影响。

blast参数

BlastN, e-值 = 0.0001

NUMTs in Sequenced Eukaryotic Genomes

线粒体DNA序列经常转移到细胞核中，形成了所谓的核线粒体假基因（NUMT）。 对具有测序线粒体和核基因组的13种真核生物物种的分析揭示了NUMT数量和大小和种间差异。 在按蚊，秀丽杆菌属，疟原虫，果蝇和富古的numt数量很少，在人，水稻和拟南芥中超过500个。 平均大小分别为62kb（面包酵母）和647 bps（Neurospora）之间。 NUMT的丰度与核或线粒体基因组的大小或核基因密度之间的相关性并不明显。其他因素，如种系中线粒体的数量和稳定性，或控制核DNA累积/丧失的物种特异性机制，可能是NUMT积累的种间多样性的原因。

本篇文章中，探讨了E值对BLAST结果的影响。作者将E值分别设为10-4、10-7、10-10……10-46、10-49、<10-50，总的来说，随着E值的减少，blast所得结果减少。但对C.elegnas来说，结果没有变化；对于P.falciparum来说，只在E值为10-4时有结果，减小E值后，没有blast结果。

每个细胞中线粒体的数量会影响numt的数量，此外，从合子到减数分裂的体细胞分裂的数量，也会影响线粒体DNA到核的转移（Walbot, V., and M. M. Evans. 2003. Unique features of the plant life cycle and their consequences. Nat. Rev. Genet. 4:369–379.）

Origin, evolution and genetic effects of nuclear insertions of organelle DNA

1. Tada SFS, Souza AP. A recombination point is conserved in the mitochondrial genome of higher plant species and located downstream from the cox2 pseudogene in Solanum tuberosum L.[J]. Genet.mol.biol, 2006, 29(29):83-89. [↑](#footnote-ref-0)
2. Ko Y J, Kim S. Analysis of Nuclear Mitochondrial DNA Segments of Nine Plant Species: Size, Distribution, and Insertion Loci[J]. Genomics & Informatics, 2016, 14(3):90. [↑](#footnote-ref-1)