2014 年西非埃博拉疫情病原体的系统发生分析

摘要：Baize 实验团队对近期在西非几内亚地区流行的埃博拉病毒进行系统发生分析，显示其在系统发生与分类上处于扎伊尔型埃博拉病毒之外，继而预测今年在西非流行的埃博拉是本土起源而非从中非传播过来。随后 Dudas 和 Rambaut 实验室根据 Baize 的数据更换了样本选取思路，剔除了亲缘关系较远的其他四种病毒型并且重新构建了系统发生树，得到了与 Baize 相反的结论：西非埃博拉病毒是从中非国家传播而来，并且在此基础上推测了进化分枝的时间。这一结论得到了 Calvignac-Spencer 进一步统计数据的支持，并且进一步推测了 2014 西非埃博拉与中非扎伊尔性埃博拉的分枝时间。本文较为详细地叙述了上述三份文章的主要实验内容与贡献，并对文章之间的关系脉络进行了梳理。

关键词：埃博拉病毒； 系统发生分析； 系统发生树

1. 引言

1.1. 埃博拉基因型

埃博拉病毒基因组由编码7个结构蛋白和2个非结构蛋白组成，基因顺序为3′端-NP-VP35-VP40-GP-VP30-VP24-L-5′端[1]，两端的非编码区含有重要的信号以调节病毒转录、复制和新病毒颗粒包装。除编码糖蛋白的基因外，所有基因均为一个单顺反子,编码一个结构蛋白。基因组所编译的蛋白中，NP是核衣壳蛋白，VP30和VP35是病毒结构蛋白[2]，VP35具有拮抗Ⅰ型干扰素作用[3]，GP 是跨膜糖蛋白,与病毒的入侵过程及细胞毒性有关，VP24 和VP40 与病毒的成熟释放有关，前者是小型膜蛋白[4]，后者构成病毒基质蛋白。RNA 依赖的RNA 聚合酶(L)，是病毒基因组转录成信使RNA 所必需的酶，它对病毒基因组的复制也有重要作用。另外，可溶性的糖蛋白(sGP)和小可溶性糖蛋白(ssGP)与跨膜GP 约有300个氨基酸相同，大小分别为60~70 kD 和全长度的150~170 kD，经过相同基因转录编辑过程，sGP 和ssGP 与埃博拉病毒细胞毒性有关。[5]

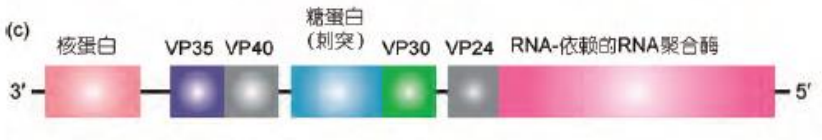

图 1. 埃博拉病毒基因组

1.2. 系统发生分析

系统发生分析是一种相当常用而且有效的从基因序列中获取进化信息的方法，其最终将能够输出一个概括进化历程的树状图。当树状图与时间关联并定根

后，我们可以从中进一步分析基因序列间的相关性并且将树状图变成系统发生树。[6]近十年以来，分子钟模型的应用使得系统发生分析方法本身与分析软件都有了长足的发展，分析方法上的飞跃继而让系统发生分析在进化研究中变得更加流行。BEAST、MrBayes 等平台使用 MCMC 法（马尔科夫链-蒙特卡洛）估计模型参数的后验分布，并根据这些参数能够得到系统发生树的可靠分枝组合。作为检验，这些分枝的可靠性都可以依靠分枝后验概率（RPP）作为统计量进行检验。

2. 几内亚本土发生说

2.1. 概述

2014 年 3 月，世界卫生组织 WTO 通告了一种 2013 年 12 月开始在几内亚爆发的传染性疾病，该疾病的症状主要体现为发热、严重腹泻、呕吐并且伴随极高的死亡率。很快，以法国医学家 Sylvain Baize 为首的疫情研究团队率先对本次肆虐埃博拉的源头进行了全面的追查，包括针对本次疫情的诊断试验、病毒的分离与测序、电镜观察以及流行病学分析，并且在此基础上进行了病原体的系统发生分析。希望通过追踪流行病学证据推测出爆发初期的疫情发展状况；与此同时也通过检验本次疫情病原体与历史上中非爆发的埃博拉病毒的亲缘程度，推测了病毒的可能来源。其研究成果的初稿发表在了《新英格兰医学杂志》上。

2.2. 主要实验方法

2.2.1. 样本收集

研究团队通过向当地医疗系统请求人口与医疗数据的方式，最终收集了来自最初疫情爆发地（盖凯杜省、马桑达省、吉西杜古省）20 位疑似病人的血样，患者均伴随头疼、腹泻、呕吐与部分出血热症状。经过针对糖蛋白与核蛋白的 RT-PCR 扩增鉴定了其中 15 位为阳性，同时从其中 5 位患者的血样中分离得到细胞环境下的病毒，另从其中一位患者分离得到的埃博拉病毒实现了电镜观察。值得注意的是，该研究因为疫情爆发突然没有来得及与血样提供者签订知情同意书，这也给应对突发公共卫生事件与紧急状况下的科学研究伦理提出了新的问题与挑战。

2.2.2. 病毒基因测序

研究团队从 3 位患者血清样本中得到了较高浓度的病毒 RNA，继而通过提取 RNA 实时 RT-PCR 首次实现了 2014 几内亚流行埃博拉病毒的完整测序。采样得到的埃博拉病毒中大部分基因片段都使用丝状病毒特异性引物来对病毒 RNA 实现了 PCR 扩增，另外一部分重复序列片段则使用了埃博拉病毒特异引物，并且采用了 Sanger 传统测序方法从两端分别测序。

2.2.3. 系统分析

研究团队使用 jModelTest 软件，将近期流行的埃博拉病毒基因组序列与 GenBank 上 48 个不同发生时间地点的埃博拉病毒完整基因组序列进行最佳拟合分析（其中包括所有的 5 种类型的埃博拉病毒），通过序列进化时间可逆模型中 GTR 来解释系统发生数据。系统进化树 1 中，研究团队使用贝叶斯 MCMC 法（马尔科夫链-蒙特卡洛），进行系统进化分析（参数：two runs of four chains with 1 million steps with a burn-in rate of 25% and the GTR+gamma model）。另外一种系统进化树 2 中，则使用最大似然法（The maximum-likelihood method）与分子钟模型（molecular clock model）进行了构建。

2.3. 实验结论分析与讨论

2.3.1. 病毒核苷酸多态性

来自三位患者样本的埃博拉病毒样本通过 Sanger 测序法被完整测序。该基因组全长为 18959 个核苷酸，并且伴随少量核苷酸多态性，见下表：

位置	核苷酸多态性	原氨基酸	替换氨基酸
2124	G→A	甘氨酸	谷氨酸
2185	A→G	同义突变	
6909	A→T	精氨酸	色氨酸
9923	T→C	同义突变	
13856	A→G	天冬氨酸	甘氨酸
15660	T→C	同义突变	

表 1. 2014 西非埃博拉病毒核苷酸多态性

根据估算，在几内亚收集到的患者体内埃博拉病毒序列上与中非埃博拉病毒仅有 97%的一致性性，加以病毒 7×10-4 个/位点/年的突变速率，如此低的一致性给研究团队提出了以下问题：是否本次埃博拉疫情与中非的扎伊尔型（刚果、加蓬）存在于不同的进化树分枝上？是否本次疫情中的埃博拉病毒并非是中非传播过去的而是已在西非蛰伏已久却在近期激烈爆发？Baize 团队继而通过系统进化分析来验证自己的猜测。

2.3.2. 系统发生分析

借助贝叶斯 MCMC 法与最大似然法对全基因组的系统发生分析显示，几内亚埃博拉病毒与中非扎伊尔型病毒在较为基部的位置就已经分枝，并且互为同源分枝。这一观点推论几内亚埃博拉与中非扎伊尔埃博拉来源于同一个祖先并且分别平行进化，而非从刚果、加蓬传播进入几内亚。

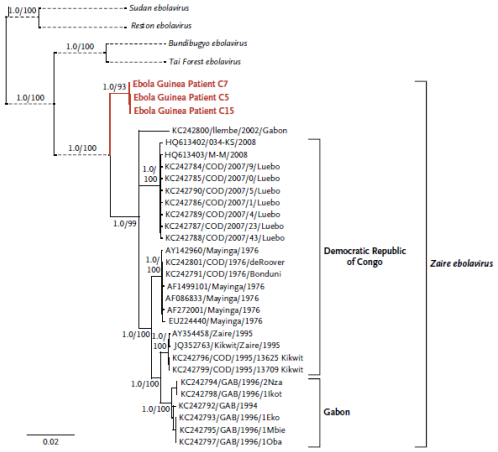

图 2. Baize 构建系统分析树

但是，因为缺乏对埃博拉病毒在自然界中进化速率的了解，Baize 团队表示并不能有效解释“病毒何时引入几内亚？”、“几内亚病毒系统发生源头是哪里？”这两个问题。

2.3.3. 疫情来源情况

通过系统发生分析得到几内亚埃博拉的较为独立分枝地位，Baize 推测了本次疫情的并非由中非疫区传播而来，而是从西非独立起源。同时，靠近基部的分枝也暗示了病毒在西非独立发展了较长的时间，经过数年的蛰伏后经过中间宿主动物与人类的长时间接触最终将病原体传播给人类，并且在西非地区的人类中大规模传播。

2.4. 该研究的重要意义

2.4.1. 首篇疫区详尽流行病学调查

该研究为首篇应对埃博拉疫情较为详实的流行病学、系统分析的科研论文，尤其是在临床与流行病学分析方面。Baize 团队深入开展了流行病追根溯源的研究，并且提出了本次埃博拉疫情最初爆发的过程：最初由患者 0（2 岁幼童）在 Meliandou 感染病死亡，随后该疾病在该地区继续小规模传播直至 S14 患者（医务工作者）携带病毒传播到了 Macenta、Nzerekore 和 Kissidougou，继而引发了逐渐大规模传播。为今后流行病学的传播模型与防控机制提供了一个很好的案例。

2.4.2. 引发后续深入研究

Baize 在本文中成功对 20 个患者中 3 个进行了病毒基因组测序，并且上传至 GenBank，为本次埃博拉爆发的后续科研提供了宝贵的数据。同时因为在系统发生分析中存在的诸多不确定性，未能完全确定本次病原体来源，也为后续进化领域的研究提供了方向。

3. 中非扎伊尔型发生说

3.1. 理论依据与实验目的

虽然 Baize 实验室对几内亚埃博拉起源理论作了较为详实的证明与论证，但是其文章发表之后还是遭到了来自其他是实验室与研究团队的质疑，其中包括爱丁堡大学进化生物研究所的 Dudas 和 Rambaut，他们认为几内亚的病毒并非一种与扎伊尔埃型非同源的埃博拉病毒。根据时间、地理距离以及基因距离的分析，他们认定：目前西非流行的埃博拉病毒与中非刚果、加蓬曾经爆发的扎伊尔型是同一种类型，本次疫情的罪魁祸首正是中非国家传播过去的扎伊尔型埃博拉病毒。

3.2. 实验原理与方法

3.2.1. 采样方法

相比 Baize 系统进化分析过程中将全基因组作为样本进行比对，Dudas 和 Rambaut 将 Baize 公布序列和 GenBank 中所有的埃博拉病毒基因组都分为 14647 个核苷酸的蛋白编码区与 4312 个核苷酸组成的基因间隔区。

3.2.2. 实验方法与模型

Dudas 和 Rambaut 在进行病毒核酸系统发生分析时所使用模型与 Baize 研究团队相似均采用了 GTR+gamma 模型，并且借助 PhyML 和 MrBayes 软件加以实现。同时，借助非相关性分子钟（the uncorrelated relaxed molecular clock）与人口模型，在 BEAST 软件可以在时间尺度上可以建立几内亚埃博拉病毒从其他埃博拉病毒分枝的时间节点，并且验证几内亚流行埃博拉病毒的起源。

3.3. 实验结论与分析

3.3.1. 重复 Baize 实验室结果

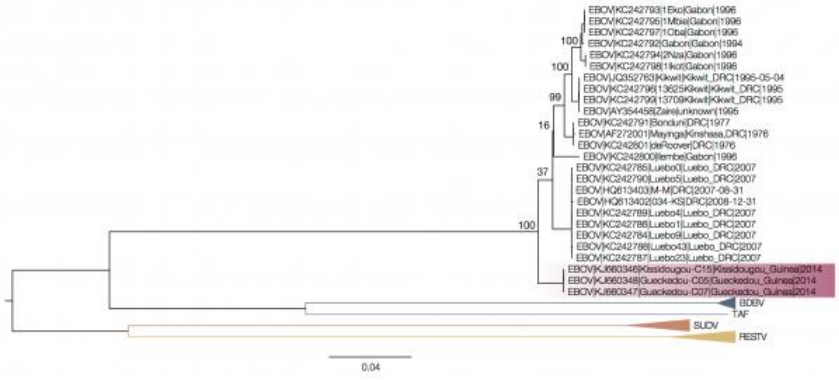

图 3. 全基因组 ML 系统发生树

Dudas 和 Rambaut 在重复 Baize 团队的系统发生树的时候发现，该结果仅仅在忽略呈离散 γ 分布的非均质率的情况下才可重复，而且统计数据并不是非常支持 Baize 的分枝结论。重复实验的困难主要来自于：（1）在分析过程中需要比对

不同长度的序列；（2）使用多种差异极大的埃博拉导致的无法确定可靠的根部位置；（3）不同的队列。

3.3.2. 编码区单独分析

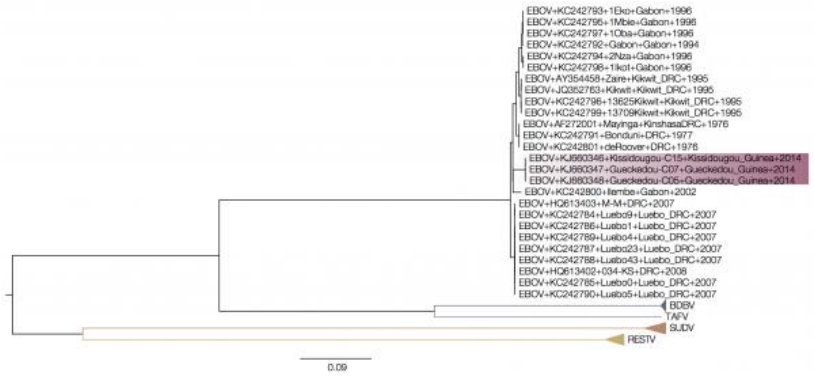

图 4. 仅计算蛋白编码序列的 MrBayes 系统发生树

由图 4 当仅有蛋白编码序列参与系统发生分析时，呈现几内亚病毒的序列分枝在众多加蓬/刚果扎伊尔型埃博拉病毒分枝内部。这一结果显示几内亚埃博拉病毒就是中非国家流行的扎伊尔型埃博拉病毒，与图 3 中的结论相悖。

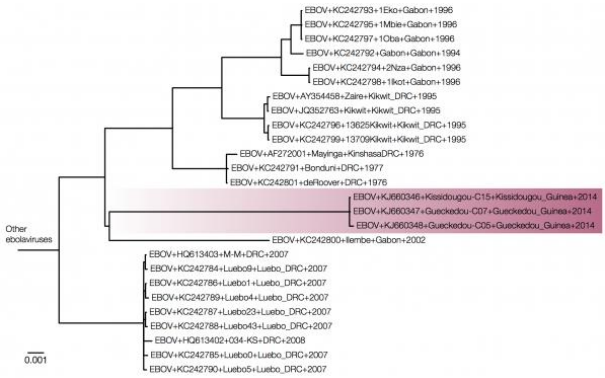

图 5. 剔除 outgroup 的蛋白编码序列的 MrBayes 系统发生树

将图 4 系统发生树中的与扎伊尔型非同型的 4 种其他种类删去即得图 5，从图 5 我们可以更加清晰地看到，几内亚埃博拉病毒分枝完全嵌入其他几种中非国家流行扎伊尔型病毒分枝中。这一结果也显示几内亚埃博拉病毒就是中非国家流行的扎伊尔型埃博拉病毒，与图 3 中的结论相悖。

3.3.3. 非编码区单独分析

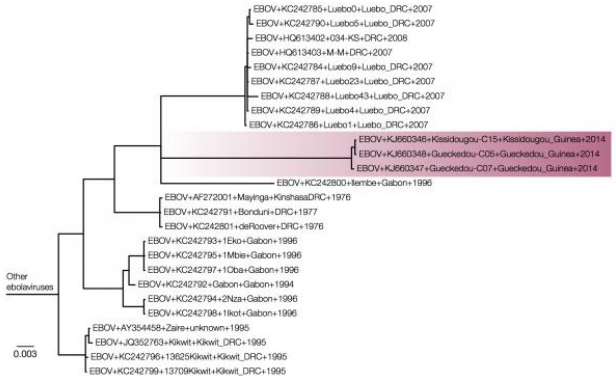

图 6. 仅计算基因间隔序列的 MrBayes 系统发生树

图 6 中基因间隔区域的系统发生树也反映了几内亚埃博拉病毒分枝完全嵌入其他几种中非国家流行扎伊尔型病毒分枝中，这一结果与图 4、图 5 结果较为相似，同样质疑了图 3 中 Baize 实验的结论。

图 5、图 6 中在把高度相异的埃博拉病毒从样本中剔除之后，分别使用了编码区与非编码区作分析，显示了与刚果爆发的扎伊尔型病毒的一致性，Dudas 和 Rambaut 继而怀疑将实验样本序列和高度相异的其他四型病毒种类一起进行比对是不合理的一种分析方法。

3.3.4. 病毒引入时间估计与最小二乘回归

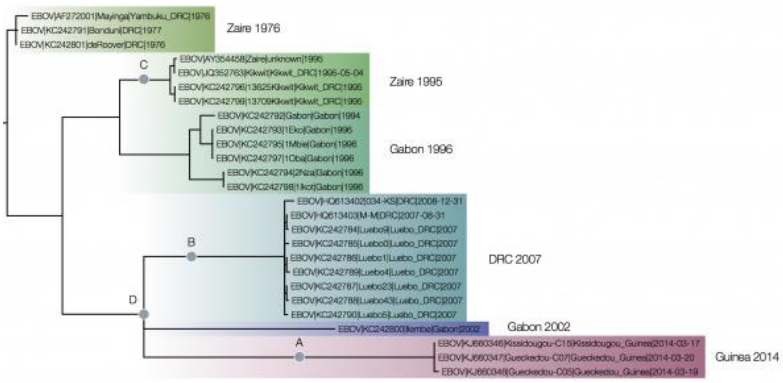

图 7. 蛋白编码序列通过最小二乘回归定根的 MrBayes 系统发生树

图 7 中对 2014 几内亚病毒与 2007 刚果扎伊尔型病毒、2002 加蓬扎伊尔型病毒归并分枝的贝叶斯检验为 1.0，这也就解释了为什么 Baize 无法完全确定几内亚埃博拉病毒在系统发生树中的位置。

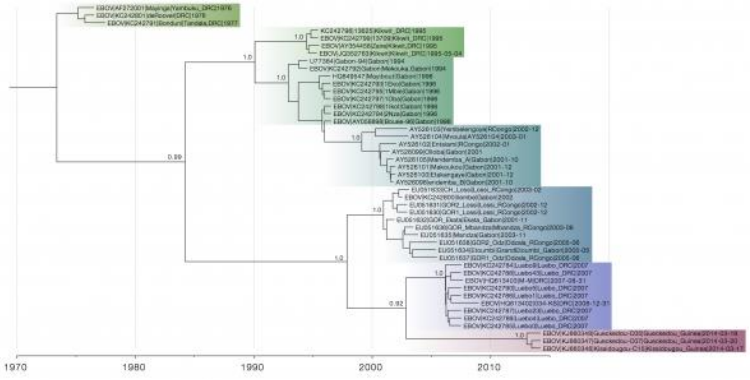

图 8. 最大置信度的糖蛋白

图 8 中通过时序分析估计了在几内亚爆发的病毒从中非刚果与加蓬爆发病毒中分枝的时间，约为 2002 年年末（95%置信区间为 2000~2006）。这个分析结果为从埃博拉病毒从中非地区引入几内亚这一事件提供了时间上的边界，因为 Dudas 和 Rambaut 从统计分析中得到较好支持了几内亚病毒与刚果、加蓬中非扎伊尔型病毒有相同的祖先（后验概率 posterior probability=1.0）

3.4. 结论与分析

因为病毒的在 40 年间多次自然爆发，埃博拉病毒的采样工作一直不是很完善，而且很取得一致的病毒序列。

虽然在 Baize 实验结果以及 Dudas 和 Rambaut 实验结果中几内亚流行的埃博拉病毒的都分枝很长，但是这并不意味着它们与中非埃博拉病毒是相异的分枝，而是因为最近爆发的病毒往往具有更多的时间来累计变异。Baize 方法中将另外四型埃博拉病毒与样本以及扎伊尔型病毒一起进行系统发生分析是不合理的，并且最终导致不可靠的病毒发源推测。

Dudas 和 Rambaut 认为在单独使用编码区、非编码区进行分析的同时，应该将与亲缘差异较大的病毒类型从样本中剔除就能够得到统计支持度较高的结论：本次在几内亚爆发的埃博拉病毒正是从近十年来从中非船舶进入几内亚的扎伊尔型埃博拉病毒，并非是一种新显的地区性埃博拉疫情。

虽然 Dudas 和 Rambaut 通过剔除 outgroup 对结果取得了较好的统计支持，但是他们也仅仅是给出了结论，并没有给出剔除的具体数理原因，这也就使得进一步统计分析显得很有必要。

4. 进一步的统计分析

4.1. 实验原理与方法

Calvignac-Spencer 对 Baize、Dudas 和 Rambaut 的数据使用分子钟模型进行了进一步分析，完全确定了 2014 年在几内亚爆发的埃博拉病毒是扎伊尔型的一员，并且认定 Baize 的团队在分析过程中的错误是源于没有注意到长枝吸引效应。

Calvignac-Spencer 的系统发生分析通过 GTR+gamma 核苷酸替换模型的 BEAST 软件来实现，加入了三个不同的人口学变量（恒定的人口、指数形式的出生率、贝叶斯模型），并且假设了末端矫正的非相关对数正态分布的分子钟。

作为对分枝的显著性检验，Calvignac-Spencer 采用了分枝定根后验概率 RPP（Branch root posterior probability）作为统计量，每当分析中有一个新的分枝可能出现（没有在系统发生树中体现），RootAnnotator 就会构建一棵系统发生树并且计算分析这个分枝的 RPP 值。

4.2. 实验结果

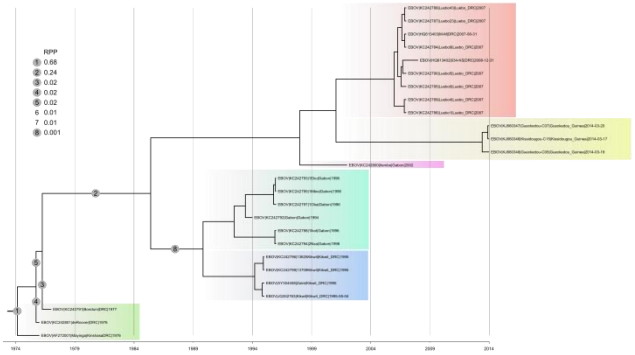

图 9. 在扎伊尔型内最高置信度的蛋白编码序列的系统发生树

图 9 中的带有 RPP 值的系统发生树来源于对恒定人口模型和非相关分子钟的分析，图片左上角为每个分枝的 RPP 显著值。所有同种颜色的内部分枝都得到了极好的验证（RPP=1.0），而 2014 几内亚病毒与刚果、加蓬以往流行的病毒有关联，但统计支持并不十分强（RPP 在 0.56 到 0.68 左右）。

4.3. 结论与分析

根据分析，扎伊尔型病毒的系统发生树有相当多的分枝，在非相关分子钟模型的假设分析下 2014 几内亚爆发的埃博拉病毒极大可能处在扎伊尔型的众多分枝里。所有的统计数据都有力的支持了 Dudas 和 Rambaut 的提出的结论。并且根据三种人口学变量（恒定的人口、指数形式的出生率、贝叶斯模型）给出了与 Dudas 和 Rambaut 结论相近的分枝时间，见下表

分枝时间	95%置信区间	模型
1999	1996-2004	贝叶斯模型
2001	1996-2003	恒定人口&指数出生率
2002	2000-2006	糖蛋白分析(Dudas 和 Rambaut)

表 2. Calvignac-Spencer 与 Dudas 对于西非埃博拉分枝时间估计

5. 分析与小结

5.1. 三个相关研究的比较

篇名	Emergence of Zaire Ebola Virus Disease in Guinea	Phylogenetic Analysis of Guinea 2014 EBOV Ebolavirus Outbreak	Clock Rooting Demonstrates that Guinea 2014 EBOV is a Member of the Za ŕe Lineage
发表期刊	N Engl J Med	PLOS Currents Outbreaks	
第一作者	Sylvain Baize,	Gytis Dudas &Andrew Rambaut	Sebastien Calvignac-Spence
取样	血样病毒 RNA 序列、GenBank	使用 Baize 公布的数据；糖蛋白序列；GenBank 中公布的各地埃博拉序列	
样本选择	不剔除较远分枝，全基因组比对	单独比较编码区与基因间隔区，剔除\非剔除	全基因组 剔除较远分枝
系统发生模型	GTR+gamma 模型	GTR+gamma 模型 非相关分子钟模型 人口学模型	GTR+gamma 替换模型 非相关分子钟模型
系统发生分析方法	MCMC 最大似然法	根-尖端回归 最小二乘回归	MCMC
系统发生分析结论	不能归类于中非爆发扎伊尔型	属于中非扎伊尔型，且亲缘较近	证实了 Dudas 的结论，属于中非扎伊尔型
病毒起源	较早分枝，病毒在西非独立进化	由中非国家（刚果、加蓬）爆发在 2002 年左右之后传播进入西非，进化	

表 3. 有关 2014 几内亚埃博拉系统发生分析 3 篇研究的比较

5.2. 对于科研的启示

第一篇文献《Emergence of Zaire Ebola Virus Disease in Guinea》大而全地涉及了流行病学、症状医学、样本测序与系统发生分析，作为首篇应对埃博拉疫情较为详实的流行病学、系统分析的科研论文，尤其是在临床与流行病学分析方面。Baize 团队深入开展了流行病追根溯源的研究，并且提出了本次埃博拉疫情最初爆发的过程。然而在基于测序结果的系统发生过程中并没有考虑十分全面，没有尝试将蛋白编码区域、基因间隔区域独立分析，也没有意识到长指吸引效应对系统发生分析的影响，最终得到了一个并不确定而且错误的结论。但是，这并不能抹杀 Baize 对于埃博拉的巨大贡献，尤其在详实的流行病学研究以及实现几内亚埃博来病毒首次测序方面，都为后续研究提供了非常好的数据支持与待解决的科研问题。

随后的两篇文献《Phylogenetic Analysis of Guinea 2014 EBOV Ebolavirus Outbreak. PLOS Currents Outbreaks》和《Clock Rooting Further Demonstrates that Guinea 2014EBOV is a Member of the Za ïre Lineage. PLOS Currents Outbreaks》则是通过文献一提供的数据，针对病毒进化与系统发生分枝这一专门问题进行了深入的研究，提出了埃博拉是由中非爆发病毒传播到西非并流行的，补充了文献一中的缺陷，使得学术界对这一问题有了更加全面的认识。

参考文献：

1. Baize S, Pannetier D, Oestereich L, Rieger T, Koivogui L, Magassouba N, Soropogui B, Sow MS, Ke ïta S, DeClerck H, Tiffany A, Dominguez G, Loua M, Traor éA, Koli éM, Malano ER, Heleze E, Bocquin A, M édy S, Raoul H,Caro V, Cadar D, Gabriel M, Pahlmann M, Tappe D, Schmidt-Chanasit J, Impouma B, Diallo AK, Formenty P, VanHerp M, G ünther S. Emergence of Zaire Ebola Virus Disease in Guinea - Preliminary Report. N Engl J Med. 2014Apr 16. PubMed PMID:24738640.

2. Dudas G, Rambaut A. Phylogenetic Analysis of Guinea 2014 EBOV Ebolavirus Outbreak. PLOS Currents Outbreaks. 2014 May 2. Edition 1. doi: 10.1371/currents.outbreaks.84eefe5ce43ec9dc0bf0670f7b8b417d.

3. Calvignac-Spencer S, Schulze JM, Zickmann F, Renard BY. Clock Rooting Further Demonstrates that Guinea 2014EBOV is a Member of the Za ïre Lineage. PLOS Currents Outbreaks. 2014 Jun 16. Edition 1. doi:10.1371/currents.outbreaks.c0e035c86d721668a6ad7353f7f6fe86.

[1] Sanchez A, Geisbert T W, Feldmann H. Filoviridae: Marburg and Ebola viruses. In: Knipe D M, Howley P M, eds. Fields Virology. Philadelphia:Lippincott Williams & Wilkins, 2006. 1409–1448

2 Kiley M P, Bowen E T, Eddy G A, et al. Filoviridae: A taxonomic home for Marburg and Ebola viruses? Intervirology, 1982, 18: 24–32

3 Muhlberger E, Weik M, Volchkov V E, et al. Comparison of the transcription and replication strategies of Marburg virus and Ebola virus by using artifi cial replication systems. J Virol, 1999, 73: 2333–2342

4 Basler C F, Wang X, Muhlberger E, et al. The Ebola virus VP35 protein functions as a type I IFN antagonist. Proc Natl Acad Sci USA, 2000, 97: 12289–12294

5 Reid S P, Leung L W, Hartman A L, et al. Ebola virus VP24 binds karyopherin alpha1 and blocks STAT1 nuclear accumulation. J Virol,2006, 80: 5156–5167

6 Hall BG. Building phylogenetic trees from molecular data with MEGA. Mol Biol Evol. 2013 May;30(5):1229-35.PubMed PMID:23486614.