北京大学

利用基因组学分析新型冠状病毒(SARS-CoV-2)的传播与演变

作者信息：	周津羽	1810301145	基础医学院
	杨礼铭	1800017720	元培学院
	崔博飞	1800017783	元培学院
	李韶威	1810307329	药学院
指导教师：	崔庆华		

二〇二〇 年 五 月

目　录

利用基因组学分析新型冠状病毒(SARS-CoV-2)的传播与演变

摘要：自 2019 年 12 月底新冠肺炎疫情首先报告于中国湖北武汉以来，截至 2020 年 5 月，疫情已扩散至 180 多个国家和地区，超过 300 万人感染，成为全球关注的紧急公共卫生事件。新型冠状病毒具有较强的传染性（R0 值在 2~3 之间），目前疫情仍在全球范围内传播。

我们团队对从 GISAID 等数据库下载的 579 个新冠病毒全基因组序列进行了系统进化网络分析，并在 Nextstrain 上将系统发生树和新冠病毒序列的进化枝分析结果进行了可视化。全部资料已上传 GitHub。

本研究中的 579 条序列的系统发生树可分为 4 个主要的进化枝分型 G,V,S,O：我们发现，英国、澳大利亚和中国的病毒株型在相对独立的进化枝上；中国大陆境内病毒序列主要为 S 型和 O 型；而其他国家如英国、美国等的序列中出现了新的突变类型，分布以 G 型和 V 型为主。一个武汉序列存在 G 型和 S 型两种突变。我们根据系统发生树分析和病毒变异规律推测：中国很可能不是或者不是疫情唯一的起源地，病毒存在于其他不同地区与国家发生最初传播的可能。

我们的研究忠实地记录了已知的新冠病毒的传播和变异情况，该方法同样可以用于对新冠病毒源头序列的系统进化分析。本研究对于继续监测新冠病毒传播与变异情况、对病毒进一步溯源和防止病毒再次爆发具有重要意义。

关键词：SARS-CoV-2;系统发生树；进化枝；Nextstrain；发源地

论文正文

第一章　引言

自 2019 年 12 月底新冠肺炎疫情首先报告于中国湖北武汉以来[1-3]，截至 2020 年 5 月，疫情已扩散至 180 多个国家和地区，超过 300 万人感染，[4]成为全球关注的紧急公共卫生事件[5]。新型冠状病毒具有较强的传染性，其基本传染数 R0 值在 2.2 左右[6-8]。并且，由于许多新冠病毒感染者表现为无症状或轻微症状（*Nature* 一篇研究指出 30%-60%的新冠感染者无症状或者症状轻微[9]；*NEJM* 发表的一篇对于临产妇进行新冠病毒的普查结果显示，无症状感染者是新冠肺炎患者的 7.25 倍[10]），容易与秋冬季节爆发的流感在诊断上发生混淆，这表明新冠病毒可在人群中隐匿传播,其总体的临床结果可能最终更类似于严重的季节性流感（病死率约为 0.1％）或大流行性流感（与发生在 1957 和 1968 年的相似），而不是类似于 SARS 或 MERS 的疾病，其病死率分别为 9%-10%和 36%。[11]这一特征使得针对 COVID-19 的疫情防控变得较为困难，客观上促成了疫情的全球性传播。

关于新冠病毒的原始宿主，目前学术界比较认可的是中科院病毒所石正丽团队发现中华菊头蝠体内冠状病毒基因序列（BatCoV RaTG13）与新冠病毒相似度达 96%，可以认为中华菊头蝠是新型冠状病毒的原始宿主。[12]但新冠病毒的中间宿主一直没有定论。目前影响力较大的有关中间宿主的研究是管轶团队报告的广东穿山甲冠状病毒和新冠病毒在 RBD 的 5 个关键残基上拥有相同的氨基酸，但只针对 RBD 的同义位点系统发育分析显示，广东穿山甲冠状病毒并非新冠病毒的最接近亲缘关系，更有可能是病毒发生了共同进化。[13]

关于新冠病毒的演化的研究，来自法国的研究者的最新研究将新冠病毒序列分为三个进化枝：G、S、V[14]，也有研究者将新冠病毒序列区分为 A、B 两型[15]。由于冠状病毒经常重组，这意味着一个单一的系统发育树可能并不总是充分地展现 SARS-CoV-2 的进化历史。 因此，本研究关于新冠病毒进化分支的讨论将着重于强调将其研究方法作为实时检测基因组流行病学的工具。

本研究是针对人类新冠病毒基因组序列的系统演化分析，能提供有关病毒起源和

传播的新证据，可以与现有的流行病学分析和宿主受体等相关研究相互补充，对于控制病毒新一轮爆发具有重要意义。

第二章　正文

2.1. SARS-CoV-2 的演化位置分析

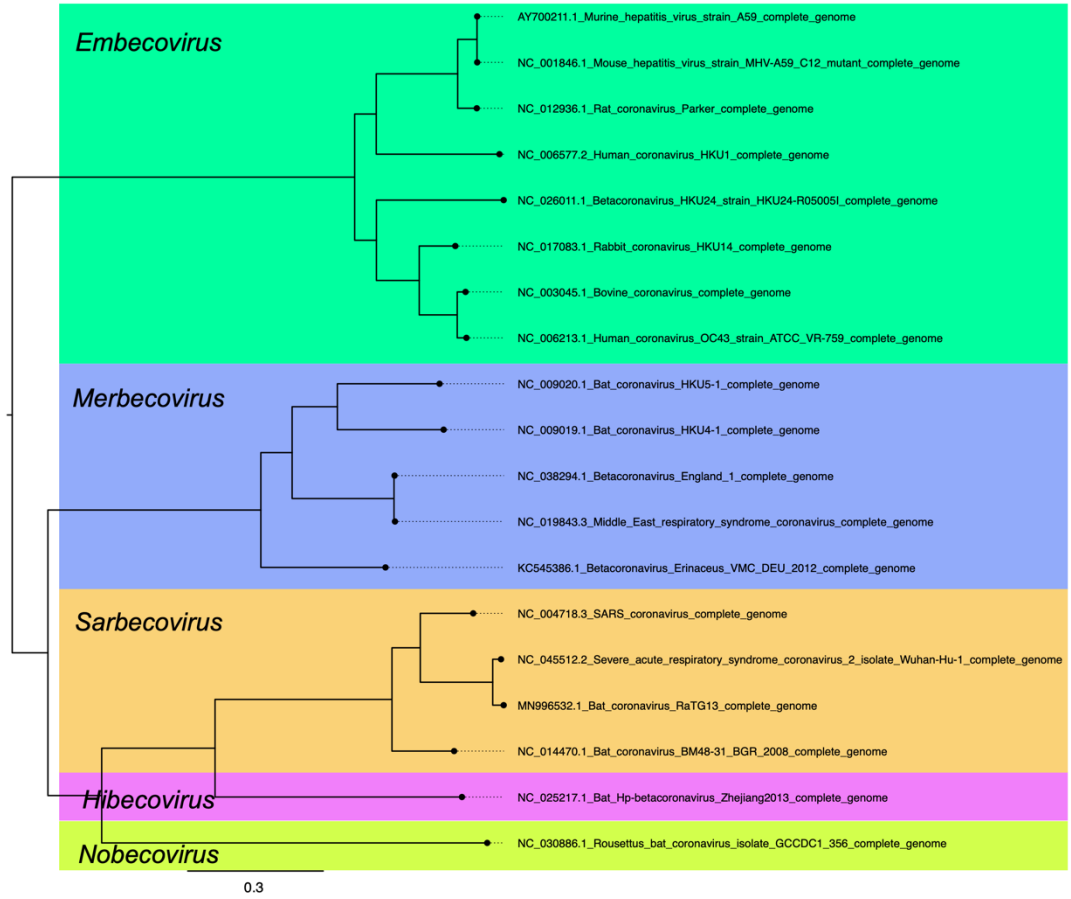


图 1：*Betacoronavirus* 主要物种的系统发生树。A.使用 mid-point root 方法确认系统发生树根部位置。图中标注了各病毒种序列所在的亚属。使用的序列来自 GenBank 数据库，获取编号在图中给出。B.其中 SARS-CoV-2 的序列使用了 GenBank 上的参考序列（NC_045512.2）。进化模型选用 GTR+I+G 模型，由 jModelTest 确定。C.所有分支的置信度均在 0.95 以上。

COVID-19 的病原体 SARS-CoV-2 是一种冠状病毒，在演化上位于 *Betacoronavirus* 属 *Sarbecovirus* 亚属；与其同一亚属的成员还有引发 2003 年非典肺炎疫情的 SARS-CoV(NC_004718.3)及其相关病毒（NC_014470.1），以及目前报道的 SARS-CoV-2 的可能祖先——中华菊头蝠冠状病毒 RaTG-13(MN996532.1)[12]。同一属的成员包括中东呼吸综合征的病原体 MERS-CoV(NC_019843.3)等。

演化上的相近暗示，新冠病毒的感染机理、致病机理和潜在治疗方式，可能与曾经人类面对过的几种冠状病毒，尤其是与 SARS-CoV 相似。在此基础上，目前的一些结构生物学工作得出的结论[16]很可能同样适用于新型冠状病毒，这为已有药物应用和新型药物研发指明了方向。

并且，新冠肺炎疫情是 1900 年以来人类社会第三次爆发冠状病毒疫情，也是影响程度最为广泛的一次。现阶段，新冠病毒疫苗研究方兴未艾，而距离下一个适合病毒复制、传播的秋冬季节也为时不远，一些研究者悲观地认为新冠病毒将在未来 5 年内持续在人群中传播[17]，因此严格控制境外输入病例，警惕病毒再次爆发具有重要意义。

2.2. SARS-CoV-2 的基因组学传播情况分析

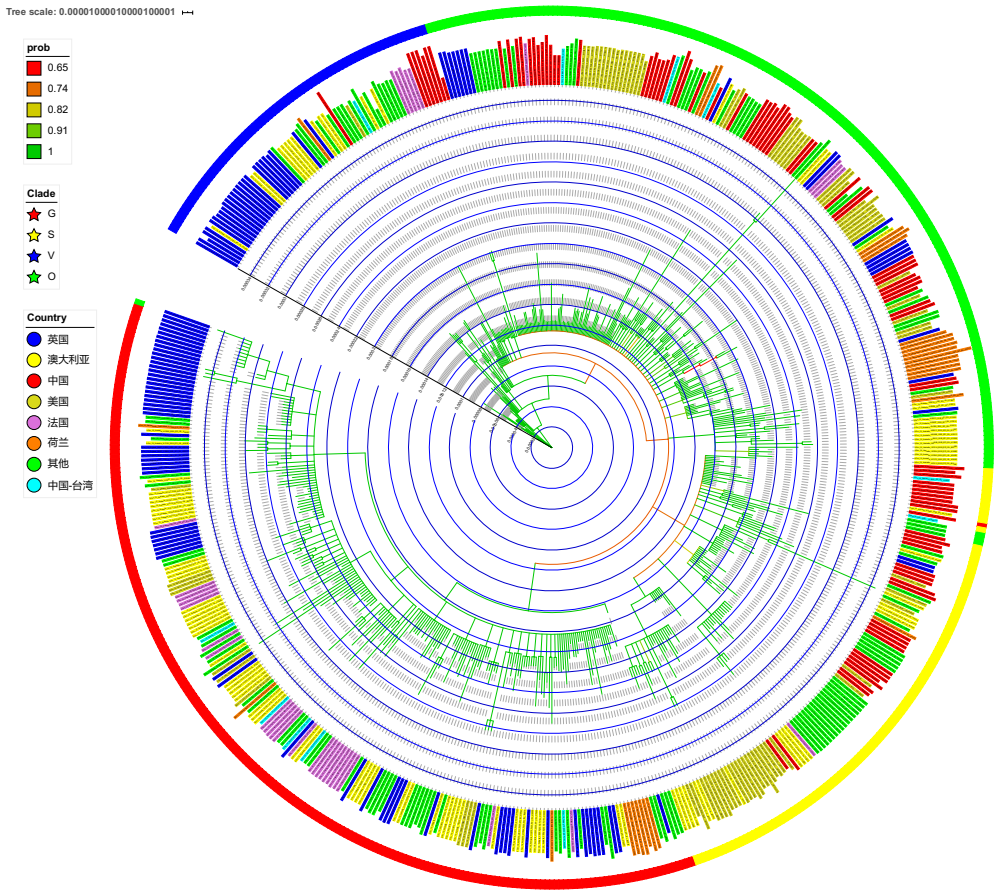


图 2:用 579 条 SARS-CoV-2 基因组序列构建的的系统发生树，未确定根部。A. 使用不同的颜色标注了序列来源地和序列进化枝分型。B. 进化模型选用 GTR+I+G 模型，由 jModelTest 确定。C. 分支置信度也在图中使用不同颜色标注。D. 考虑到中国台湾地区在疫情早期严格限制与中国大陆地区的出入境，在病毒的输入和输出情况上有所不同，因此为保证结果的准确性，单独分为一类。

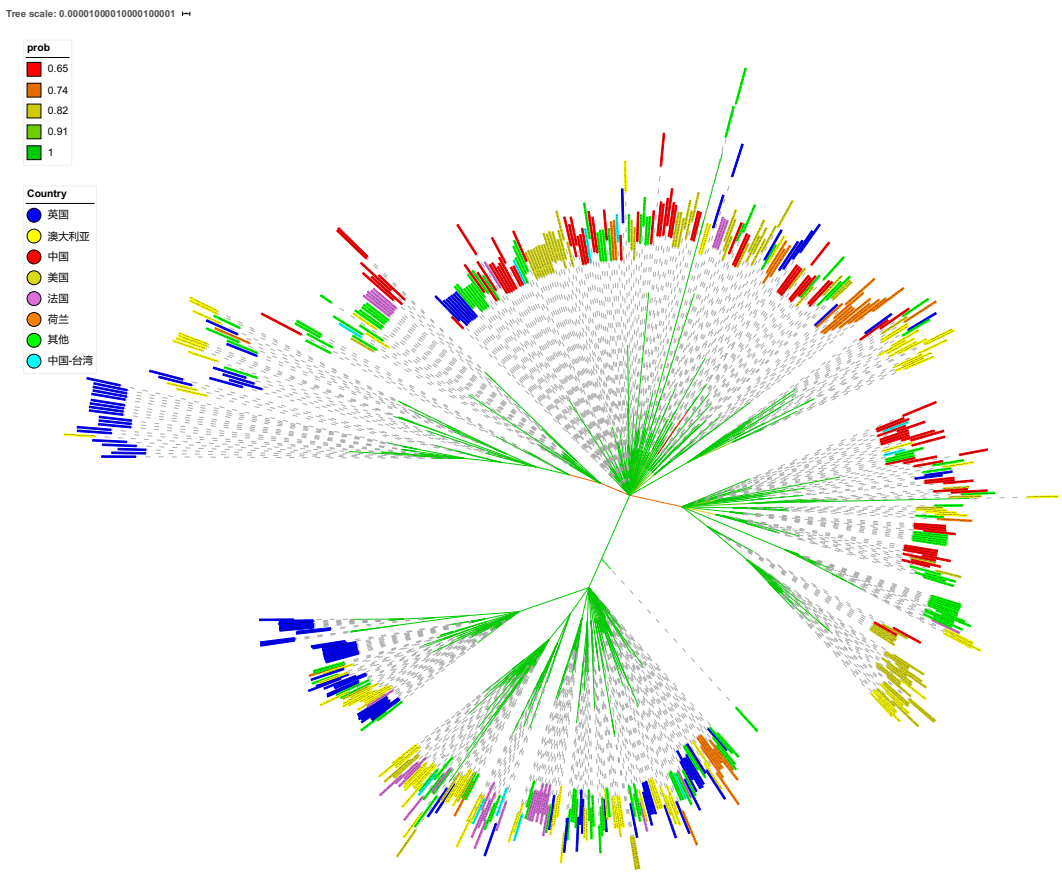


图 3：用 579 条 SARS-CoV-2 基因组序列构建的的系统发生树的无根部形式。A. 使用不同的颜色标注了序列来源地和序列进化枝分型。B. 进化模型选用 GTR+I+G 模型，由 jModelTest 确定。C. 分支置信度也在图中使用不同颜色标注。D. 考虑到中国台湾地区在疫情早期严格限制与中国大陆地区的出入境，在病毒的输入和输出情况上有所不同，因此为保证结果的准确性，单独分为一类。

COVID-19 已经呈现明显的全球化流行。我们选用了时间跨度四月余（附图 1）、来自 36 个国家（附图 2）不同病例的 579 条 SARS-CoV-2 基因组测序数据构建了系统发生树，并按照关键突变位置划定了分型。579 条基因组的序列信息可在 GitHub 的附录资料中获得（附表一）。

根据系统发生树的拓扑结构以及序列来源等信息，我们做出了如下的分析：

2.2.1. 来自中国大陆地区的序列在进化树上分布集中

中国大陆地区的序列集中于系统发生树的上部和右半部分，除下文所述的武汉的序列（hCoV-19/Wuhan/HBCDC-HB-06/2020）同时发生了 S 突变与 G 突变外，中国大陆地区的序列仅存在 S 型和 O 型；而其他国家如英国、美国等的序列较均匀地分散在系统发生树中，并且出现了新的突变类型，分布以 G 型和 V 型为主。特别地，以英

国、澳大利亚、法国等序列集中的系统发生树的左半部分没有中国大陆地区的序列分布。表明在中国大陆及香港地区传播的病毒，其基因进化的一致性相对于其他国家（如英国、美国等）更为明显。由此，我们可以认为中国政府在疫情发展早期采取果断的抗疫措施在很大程度上减缓了疫情的传播，减少了病毒的输入和输出。并且，考虑到各国间有相互输入病例的可能，根据病毒在进化中变异的规律，结合近期法国巴斯德研究院关于法国爆发的新冠疫情并非先前推测的来自意大利或中国的输入而是来自当地传播的来源不明的毒株的报道[14]，我们有理由推测：中国很可能不是或者不是疫情唯一的起源地，病毒存在于不同地区与国家发生最初传播的可能。

2.2.2. 系统发生树表明存在比武汉早期病毒序列更早的病毒基因组

在欧洲和北美的流行的 SARS-CoV-2 主要是 V 和 G 亚型。从整个无根进化树来看，还不能判断在欧洲和北美流行的 V 和 G 两个亚型与在中国境内流行的 S 和 O 亚型属于一个单系群。也就是说，我们在图中并没有找到这四种亚型共同指向的祖先序列。我们有理由认为在首例提供测序样本的患者之前 SARS-CoV-2 已经存在一定程度的流行。从病毒潜在起源地输出的 V 和 G 亚型早期患者可能因无症状或轻症状而没有确诊，或确诊但没有留下病毒基因组数据。但值得注意的是，有研究者报道，新冠病毒以一种高人类适应性的形式进入人类社会[18]，病毒由动物中间宿主传播到人类宿主的时间通过病毒“分子钟”分析等多手段分析均指向 2019 年底[19]（置信区间为 2019 年 10 月 6 日至 2019 年 12 月 11 日），排除了新冠病毒在更早期大规模隐匿的可能性。这些结论为研究者进一步寻找新冠病毒样祖先序列提供了理论依据。

2.3. 系统发生树中 SARS-CoV-2 的 G, V, S, O 亚型

根据新冠病毒的氨基酸序列位点突变情况，我们将 579 条新冠病毒全基因组序列分为 G, V, S, O 四个分型。其中，V 突变发生在 ORF3a 基因，核苷酸替换情况为 G26144T，氨基酸序列突变情况为 G251V；G 突变发生在 S 基因，核苷酸替换情况为 A23403G，氨基酸序列突变情况为 D614G；S 突变发生在 ORF8 基因，核苷酸替换情况为 T28144C，氨基酸序列突变情况为 L84S[14]。另外，我们在检测 579 条核酸序列时还发现了不属于上述突变的情况，我们将这三个位点均未发生突变得序列则统一归为一支，命名为 O 亚型。

其中，V 亚型最少，主要分布在英国和澳大利亚地区；G 亚型较多，分布的国家和地区范围也较广，主要集中分布于欧洲、美国、澳大利亚地区；S 亚型分布在系统

发生树的右半部，中国地区的序列大多为 S 分型，同属于 S 分型的国家还有美国、韩国、西班牙、澳大利亚等国，表明 S 分型传播范围较广；O 分型中也包含了部分中国地区的序列，其国家和地区的分布范围也较为广泛，包括了美国、荷兰、澳大利亚等。

一般而言，一个基因组数据只有一个保守性位点会发生突变，但是武汉的一个序列（hCoV-19/Wuhan/HBCDC-HB-06/2020，采集时间为 2020 年 2 月 7 日）同时发生了 G 和 S 两个关键位点的突变，这也是实验所选的 579 条序列中唯一一个发生了 G 突变的中国大陆境内测定序列，而且该序列与 G 支的序列在系统发生树上相距较远。由系统发生树可知中国境内的病毒主要为 S，O 两大分型。而根据法国研究者的报告，法国早期传播的新冠病毒主要为 G 分支[14]，这支持了法国研究者认为法国境内的病毒很可能并非来自于中国，而且来自于其他地区的结论。这一结论一定程度上支持了中国并不是病毒的发源地的观点。

2.4. 原系统发生树 50 条序列抽样结果

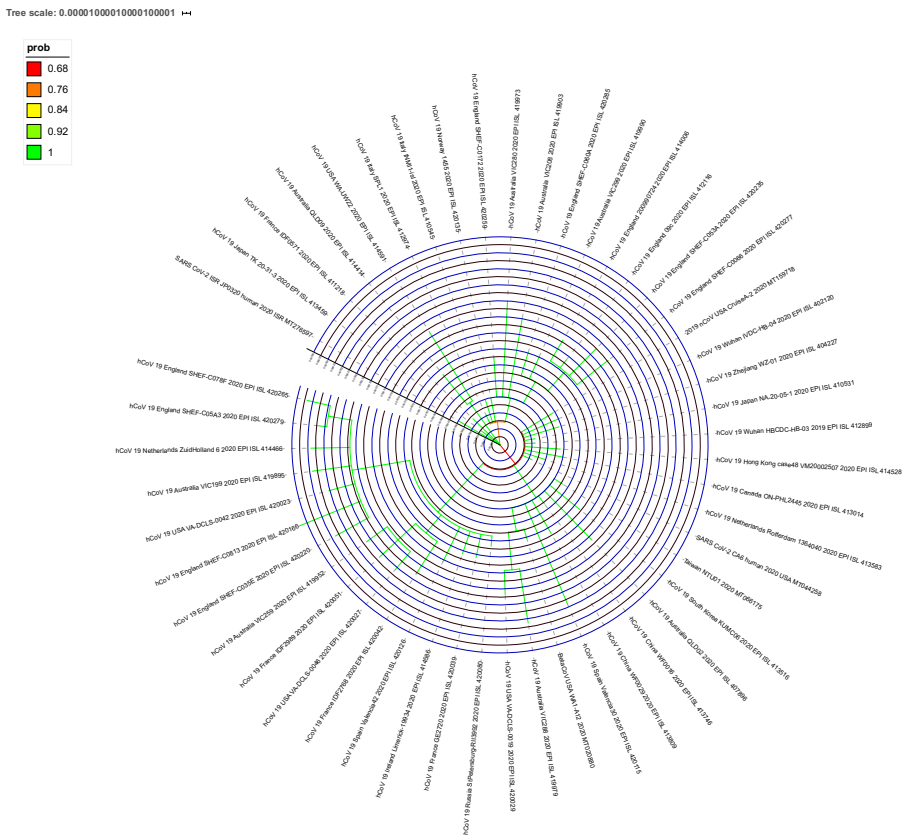


图 4：原系统发生树 50 条序列抽样结果的系统进化树。

在已有的 579 条序列绘制出的系统发生树的基础上，我们根据树的形状随机抽取了共 50 条序列进行第二次系统进化分析。该系统进化树共有三个主要分支。各分支的高置信度表明了我们所用的比对、分析模型对于新冠病毒各种全基因组序列分析的实用性。考虑到 50 条序列的样本偏小，存在较大的随机性，因此本文中不对其结果做更进一步的分析和说明。

第三章　结论

新冠病毒 SARS-CoV-2 作为一种在演化上位于 *Betacoronavirus* 属 *Sarbecovirus* 亚属的冠状病毒,是人类历史自 1900 年来第三次爆发的、影响程度最大的冠状病毒。自2019 年底新冠肺炎疫情首先报告于中国武汉以来，疫情已波及 180 多个国家，超过 300 万人确诊，并且很可能将在未来一段时间继续传播。

我们的研究发现中国大陆地区与英国、美国等其他国家和地区流行的病毒毒株在基因位点突变、系统发生树进化分析中存在较大差异，认为中国政府在疫情发展早期采取果断的抗疫措施在很大程度上减缓了疫情的传播。并且，由于在系统发生树中没有找到 G,V,S,O 四种亚型共同指向的祖先序列，表明存在比武汉早期病毒序列更早的病毒基因组；考虑到各国间有相互输入病例的可能，根据病毒在进化中变异的规律，我们有理由推测：中国很可能不是或者不是疫情唯一的起源地，病毒存在于不同地区与国家发生最初传播的可能。

本研究还将系统发生树及病毒传播情况、进化枝分型结果在 Nextstrain【20】中的可视化界面进行了展示，以帮助人们更好地理解目前的新冠病毒的传播与进化情况。

本研究中针对人类新冠病毒基因组序列的系统演化分析，提供了有关病毒起源和传播的新证据，可以与现有的流行病学分析和宿主受体等相关研究相互补充，对于未来继续监测新冠病毒传播与变异情况、寻找新冠病毒可能祖先序列与动物中间宿主、控制病毒新一轮爆发具有重要意义。

第四章　讨论与展望

4. 1. 系统发生树根部的确定

确定系统发生树的根部有助于对 SARS-CoV-2 传播的进一步分析，证实哪一个分类单元的分支先于其他的分类单元。确定树根的方法包括中点取根（mid-point rooting）、外群取根（outgroup rooting）以及基于分子钟（molecular clock）的取根方式等。【21】本研究中，我们尝试应用不同的确定树根的方法，判断各种方法在本问题中的优点和劣势。以下将详细阐明。

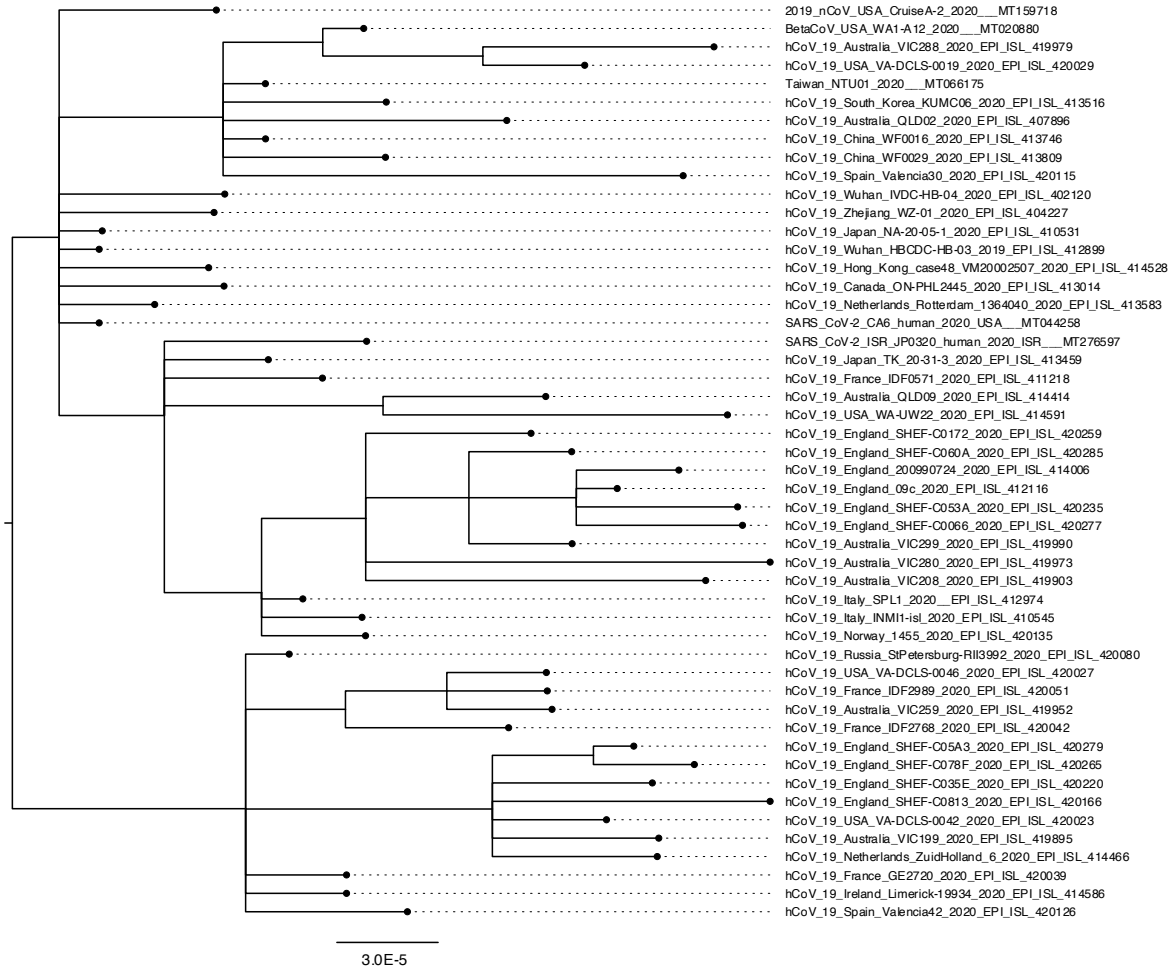


图 5.1：使用不同方式确定系统发生树根部。A.中点取根法。

在病毒的基因组学研究中，由于通常较难找到现存病毒在演化史上的祖先，即病毒形态微小，几乎不可能在化石中保存下来；且它们大都存在于宿主细胞内，插入宿主基因组中，难以区分。因此病毒学的遗传分析中常用中点取根方式分析病毒的演化地位（图一）。中点取根法：即在缺少一个合适的外群时，根大约可以置于两个分类操作单元间最长支的中点上，一般用于分析病毒演化地位时各枝是相对并列的系群（例如前述 *Betacoronavirus* 主要物种的系统发生树）[22]，当系统发生树中所有分支的进化速度大致相同而且实际的外群与其它分类群间的支的长度不太短时，这种方法相当准确。但是在本问题中，由于我们选择的各病毒基因组序列之间有潜在的代际关系，中点取根的方法对本问题研究的推动比较有限。并且由于中点取根法基于以下假设前提：假设两个最不同的谱系以相同的速率进化。显然，这个假设现实中很可能不成立。[18]（图 5.1）

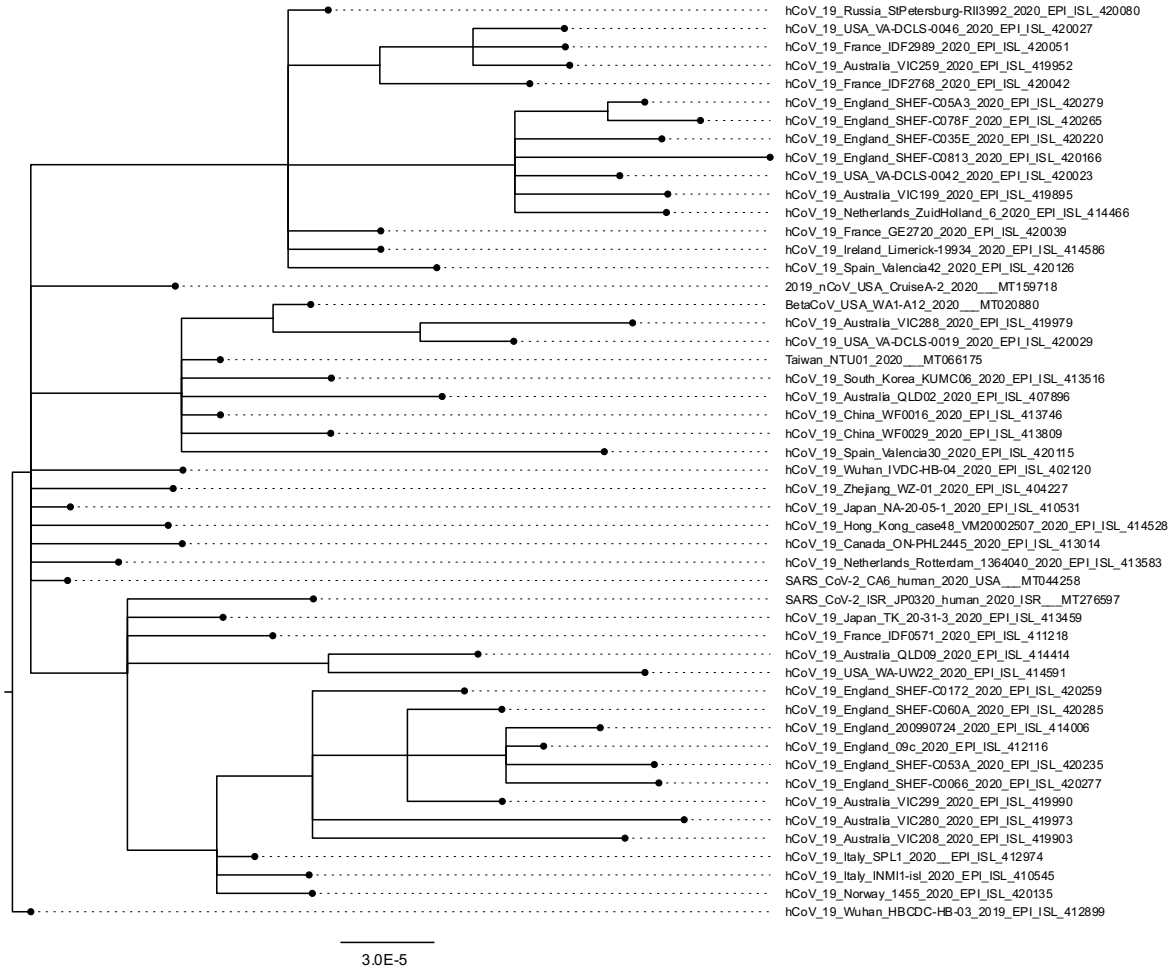


图 5.2：使用不同方式确定系统发生树根部。B.使用时间上最早的序列作为树根。

目前的部分针对 SARS-CoV-2 的基因组系统发生分析（如 Nextstrain[20]）选取了发现时间相对较早的武汉早期病例测序结果作为树根。这一方法先验地认为武汉早期病例的病毒基因组序列更接近病毒祖先。这一方法在一定程度上是合理的，但是由于目前仅知武汉是最早报告新冠疫情的疫情爆发地，科学家现仍不能确定武汉是否是疫情的发源地。[23]武汉早期病例数据（现有公开的最早核酸序列采集日期为 2019-12-26）只是我们收集到的最早序列，可以肯定的是自 2019 年 12 月 8 日武汉市卫健委通报了武汉首个新冠肺炎病例以来[24]，更早期的确诊病例的核酸序列曾经存在但未被收集。如果同期 SARS-CoV-2 已经分化为两个甚至更多支系[25]，通过这一方法得到的结论就将受到挑战。（图 5.2）

图 5.3：使用不同方式确定系统发生树根部。C.以 RaTG-13 作为外群取根。

最后，使用外群取根的方法虽然被系统发生学推崇，但外群的选择可能会对树的根部位置产生较大的影响。根据外群选取的一般性原则，我们选取了独立 SARS-CoV-2，同时在演化上与 SARS-CoV-2 相近的中华菊头蝠冠状病毒 RaTG-13【12】（图一）作为外群确定进化树根部（图 5.3）。但 RaTG-13 基因组序列与 SARS-CoV-2 基因组的差异仍然远大于不同时间、地区、病例的病毒基因组样本之间的差异，这会导致长枝吸引（long-branch attraction,LBA）【26】等问题，导致原本相距比较远的序列被推断成具有相同的祖先。这一方案有溯源和分析传播的可能，但结论的可信程度比较一般。

4. 2. 通过系统进化网络分析病毒宿主的可行性

基于系统发生学和基因组学同样可以进行 SARS-CoV-2 的宿主分析。有研究表明中华菊头蝠冠状病毒（BatCoV RaTG13）的基因组与新冠病毒相似度达 96%，是目前已知的冠状病毒中与 SARS-CoV-2 亲缘关系最接近的（图一），因此研究者推测中华菊头蝠是 SARS-CoV-2 最可能的原始宿主。【12】另有研究表明，SARS-CoV-2 刺突蛋

白上与血管紧张素转化酶 II（ACE2）的受体识别区域（Receptor Binding Domain,RBD）对应的基因序列与来自广东的从穿山甲样本中检测到的冠状病毒相应序列相近，5 个关键氨基酸残基突变相同，给出了穿山甲可能是 SARS-CoV-2 演化过程的中间宿主的观点，但针对 RBD 的同义位点系统发育分析无法排除病毒发生共同进化的可能。[13] 也有研究者通过新冠病毒序列与不同动物宿主 ACE2 受体的结合能力差异确定动物中间宿主的范围[27]。应该指出，病毒宿主分析的瓶颈在于新的冠状病毒的发现和测序，在目前已经公开全基因组序列的冠状病毒种类中，我们难以推断宿主信息，尤其是中间宿主信息。通过基因组学方法确定宿主，亟待更多冠状病毒基因组序列的获得。因此，为了追踪新冠病毒祖先病毒株，找出其中间宿主，我们提出了以下建议：

1、从市场、农场和野生动物中收集 SARS-CoV-2 样病毒；

2、检查 2019 年底前几个月的人类样本中 SARS-CoV-2 样病毒或 SARS-CoV-2 反应抗体，以检测病毒在人类社会流行的前体；

3、对更多早期的 SARS-CoV-2 分离株进行测序，特别是早期分离株，可以识别出人类适应性较弱的祖细胞的分支；

4、评估新冠疫情发生早期动物养殖者、食品经营者和动物贸易商的患病率相较普通人是否过高，特别是从事某种野生动物交易的人群的感染率是否偏高。

其中，第一项建议考虑到疫情早期爆发地华南海鲜市场已经多次进行了全面消杀清理以防止疫情再次爆发，从该地收集 SARS-CoV-2 样病毒的可能性较低。但由于已有研究表示华南海鲜市场不是病毒发源地[28]，因此关注武汉其他市场、农场以及武汉以外的市场、农场和野生动物养殖和交易场所都是有意义的。

第二、第三项建议已经被越来越多的世界各国研究者们执行，现不断有新的结果被报道，地区首例病人的时间也被不断提前[14]。

第四项建议中关于患病人群的调查可在流行病学上对新冠病毒中间动物宿主的溯源提供支持。

另外，由于动物中间宿主尚未确定，研究中也发现了新冠病毒与猫[29]、狗[30]等常见动物具有相对较好的亲和力，病毒很有可能再次通过动物宿主传播到人类。因此减少人类与野生动物的接触，阻断可能的传播途径，对于防止疫情再次爆发具有重要意义。

4.3. 本研究样本的局限性和未来研究的展望

本研究共选取了 579 条序列，这也是为了保证数据完整与高质量进行筛选后的结果。但新冠疫情截至成稿前已经在人群中广泛流行五月余，波及 180 多个国家，确诊感染人数超过 300 万人；相对于全球庞大的感染人群，本研究选取的样本数量显得偏小，采集国家和地区相对有限，但本研究所使用的方法已经被证明具有准确性和可重复性。因此，在未来的下一步研究中，我们将根据疫情严重程度不同进一步收集新冠病毒核酸序列信息，持续对新冠病毒传播和演化情况进行检测，继续寻找病毒起源和传播的新证据。

第五章　材料与方法

5.1. 数据来源与数据情况

用于 SARS-CoV-2 在 Betacoronavirus 属中演化位置分析的序列来自 GenBank, 获取编号(accession number)已经在图一中给出。用于 SARS-CoV-2 基因组学进化与传播分析的 579 条全基因组序列来自国家生物信息中心 2019 新型冠状病毒信息库（2019nCoVR）（https://bigd.big.ac.cn/ncov?from=timeline）。这些序列数据的原始来源主要是 GISAID（https://www. gisaid.org/）、GenBank 等世界各国研究者们普遍认可并共同维护的开源数据库。

5.2. 进化树构建方法

5.2.1. 数据筛选

为了确保实验结果的可靠性，我们在从数据库获得序列时选用了序列完整，质量评级为高（表明该序列已经通过了基于未知碱基（N）数量、简并碱基（非 ATGCN 的碱基）数量两项质控检测）的 600 余个数据，选用的绝大多数数据还通过了与参考序列（MN908947）比对后出现的 gap（deletion、insertion、indel）数量、变异总数、变异密度（变异数/区间长度，区间长度<=20nt）等质量评估。[32] 随后，我们使用 MEGA-X[34] 对序列按照质量情况进行了再一次筛选，删除了未知碱基偏多的不合格序列。最

终我们选择了质量较好的 579 条序列用于系统发生树分析。这些数据分别来自 36 个国家和地区（附图 1），样本采集时间范围为 2019 年 12 月 26 日至 2020 年 4 月 1 日（附图 2）。

5.2.2. 序列对齐、分型和进化模型选择

实现如此庞大的全基因组序列比对对算法的选择和计算机的运算能力都是很大的挑战。[35]经过反复试验研究，我们团队序列对齐使用 MAFFT 7.037[36]的 FFT-NS-2 算法，参数是： GOP 是 1.53，GEP 是 0.123，矩阵是 BLOSUM62，这是 FFT-NS-2 策略下的最佳参数。在 MAFFT 7.037 现有的算法中，FFT-NS-2 可以同时满足比对精度高、运行时间相对较短的要求。其后，为了确定用于构建系统发生树的最佳进化模型。序列分型使用原创的 Python 程序对关键突变位点进行检测，Python 程序可在 GitHub 上获取（附件一）。

对齐后的数据被导入到 jModelTest2.1.10[37]进行进化模型检验，部分工作在开源服务器 CIPRES Science Gateway 上进行。[38]

5.2.3. 进化树构建和数据可视化

接着，在 jModelTest2.1.10 检验的 88 种进化模型中我们使用了软件推荐的 GTR+I+G 模型进行 SARS-CoV-2 的演化位置分析和人类新冠病毒序列的系统发生树分析。系统发生树分析使用 MrBayes3.2.7[39]的蒙特卡洛马尔可夫链方法（MCMC）进行，得到了 579 个新冠病毒序列系统发生树的结果。我们使用在 FigTree.v1.4.4[40]和 iTOL v5[41]（https://itol.embl.de/）上对系统发生树进行注释(annotation)和美化。文中涉及到的数据和资料已经上传到 GitHub 上，网址是 https://github.com/chunfenri/Transmission_decoder_SARS_CoV_2;我们还在全球研究者们共同关注的可视化平台 Nextstrain 上建立了自己的社区，文中提及的系统发生树可以在网站的交互式页面中看到，Nextstrain 社区可以通过上文提及的 GitHub 界面进入。

参考文献

【1】The 2019-nCoV Outbreak Joint Field Epidemiology Investigation Team, Li Q. Notes from the field: an outbreak of NCIP (2019-nCoV) infection in China — Wu- han, Hubei Province, 2019–2020. China CDC Weekly 2020; 2: 79-80.

【2】Tan WJ, Zhao X, Ma XJ, et al. A novel coronavirus genome identified in a cluster of pneumonia cases — Wuhan, China 2019–2020. China CDC Weekly 2020; 2: 61-2.

【3】Zhu N, Zhang D, Wang W, et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N Engl J Med*. 2020;382(8):727-733. doi:10.1056/NEJMoa2001017

【4】Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect Dis*; published online Feb 19. https://doi.org/10.1016/S1473-3099(20)30120-1.

【5】 Topcuoglu, Nursen. (2020). Public Health Emergency of International Concern: Coronavirus Disease 2019 (COVID-19). The Open Dentistry Journal. 14. 71-72. doi:10.2174/1874210602014010071.

【6】Li Q, Guan X, Wu P, et al. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. *N Engl J Med*. 2020;382(13):1199-1207. doi:10.1056/NEJMoa2001316

【7】Lin Q, Zhao S, Gao D, et al. A conceptual model for the coronavirus disease 2019 (COVID-19) outbreak in Wuhan, China with individual reaction and governmental action. *Int J Infect Dis*. 2020;93:211-216. doi:10.1016/j.ijid.2020.02.058

【8】Zhao S, Lin Q, Ran J, et al. Preliminary estimation of the basic reproduction number of novel coronavirus (2019-nCoV) in China, from 2019 to 2020: A data-driven analysis in the early phase of the outbreak. *Int J Infect Dis*. 2020;92:214-217. doi:10.1016/j.ijid.2020.01.050

【9】 Qiu J. Covert coronavirus infections could be seeding new outbreaks [published online ahead of print, 2020 Mar 20]. *Nature*. 2020;10.1038/d41586-020-00822-x. doi:10.1038/d41586-020-00822-x

【10】Sutton D, Fuchs K, D'Alton M, Goffman D. Universal Screening for SARS-CoV-2 in Women Admitted for Delivery [published online ahead of print, 2020 Apr 13]. *N Engl J Med*. 2020;NEJMc2009316. doi:10.1056/NEJMc2009316

【11】 Guan WJ, Ni ZY, Hu Y, et al. Clinical Characteristics of Coronavirus Disease 2019 in China. *N Engl J Med*. 2020;382(18):1708-1720. doi:10.1056/NEJMoa2002032

【12】 Zhou P, Yang XL, Wang XG, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020;579(7798):270-273. doi:10.1038/s41586-020-2012-7

【13】Lam TT, Shum MH, Zhu HC, et al. Identifying SARS-CoV-2 related coronaviruses in Malayan pangolins [published online ahead of print, 2020 Mar 26]. Nature. 2020;10.1038/s41586-020-2169-0. doi:10.1038/s41586-020-2169-0

【14】Introductions and early spread of SARS-CoV-2 in France.Fabiana Gámbaro, Sylvie Behillil, Artem Baidaliuk, Flora Donati, Mélanie Albert, Andreea Alexandru, Maud Vanpeene, Méline Bizard, Angela Brisebarre, Marion Barbet, Fawzi Derrar, Sylvie van der Werf, Vincent Enouf, Etienne Simon-Loriere

bioRxiv 2020.04.24.059576; doi: https://doi.org/10.1101/2020.04.24.059576

【15】A dynamic nomenclature proposal for SARS-CoV-2 to assist genomic epidemiology

Andrew Rambaut, Edward C. Holmes, Verity Hill, Áine O'Toole, JT McCrone, Chris Ruis, Louis du Plessis, Oliver G. Pybus bioRxiv 2020.04.17.046086; doi: https://doi.org/10.1101/2020.04.17.046086

【16】Zumla A, Chan JFW, Azhar EI, Hui DSC, Yuen KY. Coronaviruses-drug discovery and therapeutic options. Nat Rev Drug Discov. 2016;15(5):327-347. doi:10.1038/nrd.2015.37

【17】 Kissler SM, Tedijanto C, Goldstein E, Grad YH, Lipsitch M. Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period [published online ahead of print, 2020 Apr 14]. Science. 2020;eabb5793. doi:10.1126/science.abb5793

【18】SARS-CoV-2 is well adapted for humans. What does this mean for re-emergence?Shing Hei Zhan, Benjamin E. Deverman, Yujia Alina Chan.bioRxiv 2020.05.01.073262; doi:https://doi.org/10.1101/2020.05.01.073262

【19】Lucy van Dorp, Mislav Acman, Damien Richard, Liam P. Shaw, Charlotte E. Ford, Louise Ormond, Christopher J. Owen, Juanita Pang, Cedric C.S. Tan, Florencia A.T. Boshier, Arturo Torres Ortiz, François Balloux,Emergence of genomic diversity and recurrent mutations in SARS-CoV-2,Infection, Genetics and Evolution,2020,104351,ISSN 1567-1348,https://doi.org/10.1016/j.meegid.2020.104351.

【20】Hadfield J, Megill C, Bell SM, et al. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics. 2018;34(23):4121-4123. doi:10.1093/bioinformatics/bty407

【21】Kinene T, Wainaina J, Maina S, Boykin LM. Rooting Trees, Methods for. Encycl Evol Biol. 2016;3(c):489-493. doi:10.1016/B978-0-12-800049-6.00215-8

【22】Estimating Phylogenetic Trees from Distance Matrices.James S. Farris.The American Naturalist 1972 .doi:106:951, 645-668

【23】Forster P, Forster L, Renfrew C, Forster M. Phylogenetic network analysis of SARS-CoV-2 genomes. Proc Natl Acad Sci U S A. 2020;117(17):9241-9243. doi:10.1073/pnas.2004999117

【24】第一财经. 假如武汉的警铃有机会被拉响，可以是哪天? Https://mp.weixin.qq.com/s/TQj7IIUZkwIf0M3I8PquA, 2020

【25】Lucy van Dorp, Mislav Acman, Damien Richard, Liam P. Shaw, Charlotte E. Ford, Louise Ormond, Christopher J. Owen, Juanita Pang, Cedric C.S. Tan, Florencia A.T. Boshier, Arturo Torres Ortiz, François Balloux,Emergence of genomic diversity and recurrent mutations in SARS-CoV-2,Infection, Genetics and Evolution,2020,104351,ISSN 1567-1348,https://doi.org/10.1016/j.meegid.2020.104351.

【26】Zou H, Jakovlić I, Zhang D, et al. Architectural instability, inverted skews and mitochondrial phylogenomics of Isopoda: outgroup choice affects the long-branch attraction artefacts. R Soc Open Sci. 2020;7(2):191887. Published 2020 Feb 5. doi:10.1098/rsos.191887

【27】Functional and Genetic Analysis of Viral Receptor ACE2 Orthologs Reveals Broad Potential Host Range of SARS-CoV-2

【28】Yu, Wen-Bin,Tang, Guang-Da,Zhang, Li,Corlett, Richard T..(2020).Decoding evolution and transmissions of novel pneumonia coronavirus using the whole genomic data.[ChinaXiv:202002.00033]

【29】Shi J, Wen Z, Zhong G, et al. Susceptibility of ferrets, cats, dogs, and other domesticated animals to SARS-coronavirus 2 [published online ahead of print, 2020 Apr 8]. Science. 2020;eabb7015. doi:10.1126/science.abb7015

【30】Xia X. Extreme genomic CpG deficiency in SARS-CoV-2 and evasion of host antiviral defense [published online ahead of print, 2020 Apr 14]. Mol Biol Evol. 2020;msaa094. doi:10.1093/molbev/msaa094

【31】Benson DA, Cavanaugh M, Clark K, et al. GenBank. Nucleic Acids Res. 2018;46(D1):D41-D47. doi:10.1093/nar/gkx1094

【32】Zhao WM, Song SH, Chen ML, et al. The 2019 novel coronavirus resource. Yi Chuan. 2020;42(2):212-221. doi:10.16288/j.yczz.20-030

【33】Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data - from vision to reality. Euro Surveill. 2017;22(13):30494. doi:10.2807/1560-7917.ES.2017.22.13.30494

【34】Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Mol Biol Evol. 2018;35(6):1547-1549. doi:10.1093/molbev/msy096

【35】 Maier, D. The complexity of some problems on subsequences and supersequences[J]. Journal of the ACM (JACM), 1978, 25(2): 322-336.

【36】Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30(4):772-780. doi:10.1093/molbev/mst010

【37】 Darriba D, Taboada GL, Doallo R, Posada D. jModelTest 2: more models, new heuristics and parallel computing. Nat Methods. 2012;9(8):772. Published 2012 Jul 30. doi:10.1038/nmeth.2109

【38】Miller MA, Pfeiffer W, Schwartz T. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. 2010 Gatew Comput Environ Work GCE 2010. 2010. doi:10.1109/GCE.2010.5676129

【39】Ronquist F, Teslenko M, van der Mark P, et al. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol. 2012;61(3):539-542. doi:10.1093/sysbio/sys029

【40】 https://github.com/rambaut/figtree/releases

【41】Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. 2019;47(W1):W256-W259. doi:10.1093/nar/gkz239

Decoding the transmissions of the novel coronavirus(SARS-CoV-2)using whole genomic data

ABSTRACT：An outbreak of novel coronavirus (SARS-CoV-2) that first reported in Wuhan, China, has spread rapidly, with more than 3 million cases confirmed in more than 180 countries as of May 2020. WHO Director-General has declared it a Public Health Emergency of International Concern.The novel coronavirus is highly infectious (R0 is between 2 and 3), and the pandemic is still spreading worldwide.

In a phylogenetic network analysis of 579 complete SARS-Cov-2 genomes freely available at databases such as GISAID,we analyzed its clades as well as visualized the phylogenetic tree on the nextstrain .All materials are available at github.

The phylogenetic tree can be divided into four main clades: we found that the virus strains of the United Kingdom, Australia, and China belong to relatively independent clades;The virus sequences from mainland China are mainly S and O types; while other countries and areas such as the United Kingdom, the United States possess new mutation types, mainly distributed in G and V types.Notably,a sequence from Wuhan enjoys two types of mutations:G and S.Based on our phylogenetic tree analysis and the law of virus variation,we speculated that: China is probably not or not the only place of birthplace of the novel coronavirus outbreak. The virus may originated from other regions rather than Wuhan,China.

Our research faithfully records the spread and mutations of documented novel coronavirus,indicating that our phylogenetic methods can likewise be successfully used to help trace undocumented COVID-19 infection sources, which is significant for motoring the dynamics of SARS-CoV-2 in the future, tracing the ancestral virus type and potential intermediate animal hosts as well as preventing recurrent spread of the disease worldwide.

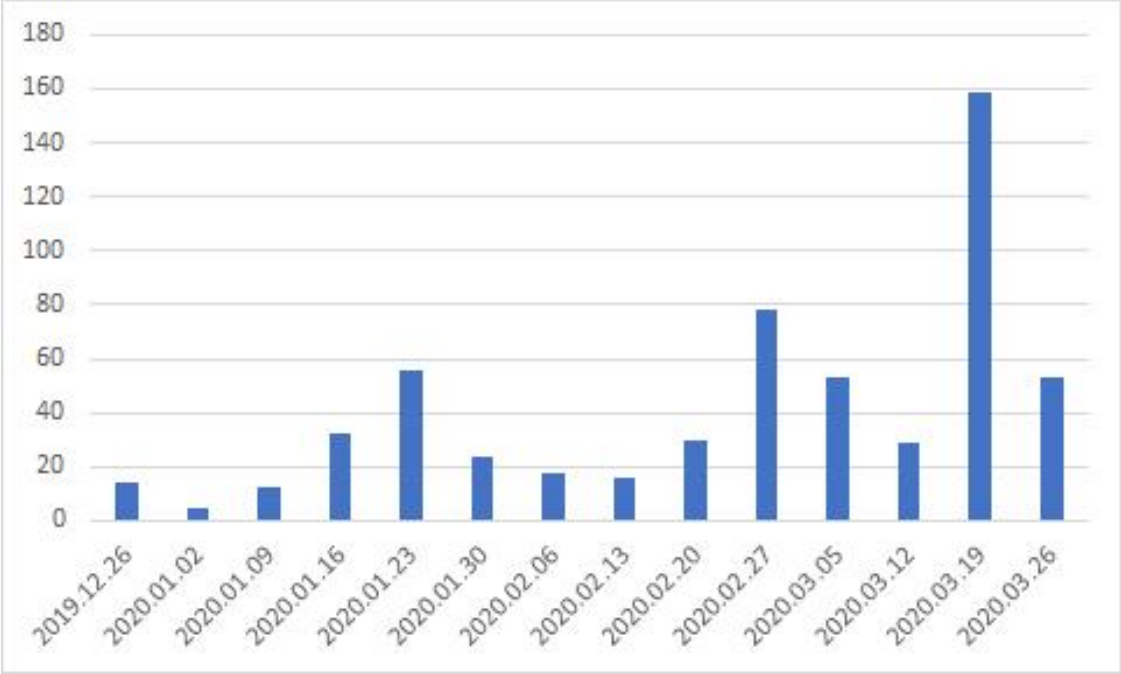
KEY WORDS: SARS-CoV-2; phylogenetic tree; clade; nextstrain; birthplace

致　谢

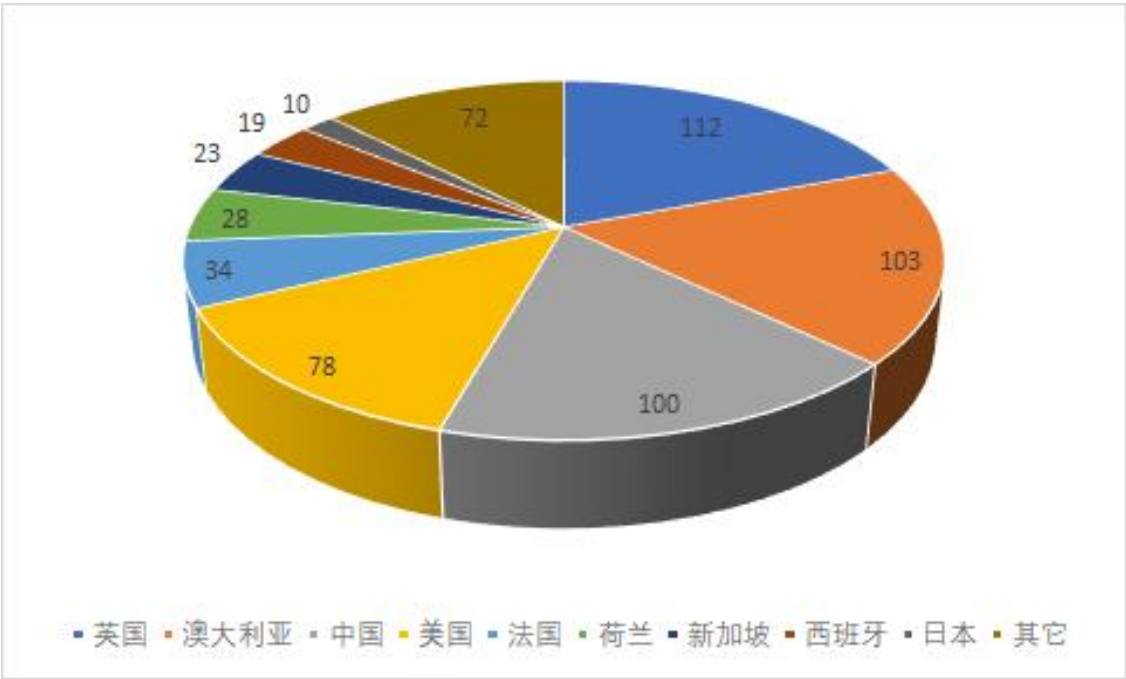
论文的最后，我想向两个多月以来一直并肩作战的挑战杯团队的队友们致以最诚挚的感谢。从选题到阅读文献、寻找模型、分析、构图、成文，尽管挑战杯之外还有种种课业的压力，但每个人一直都以最饱满的热情全力以赴，提出了一个个精彩的想法，完成一个个任务，怀着对新冠肺炎疫情热切地关注和探索科学真相的热情不断地在我们也许尚未熟悉的领域里学习新知，共同成长，最终得到了可喜的成果！

我也要感谢团队的指导老师基础医学院生物信息学系的崔庆华老师以及公共卫生学院的黄捷老师和公共卫生学院的博士生梁志生师兄，在我们的研究遭遇瓶颈和困难时热情地给予指导和帮助。在论文的写作过程中，华南农业大学的刘苹教授、香港大学的 Tommy Tsan-Yuk Lam 教授，法国巴斯德研究院的 Etienne Simon-Loriere 教授和 Artem Baidaliuk 教授通过邮件对我们在研究中遇到的问题给予了耐心的指导，在此向各位老师们表示感谢！最后我要特别感谢那些慷慨地将新冠病毒序列的数据及时上传到 GISAID、GenBank、NCBI、中国国际生物信息中心等数据库的世界各国的研究者们，本研究正是基于他们提供的实验室数据信息才得以完成。

附　录



附图 1：579 条新冠病毒全基因组序列采集时间情况。A. 以图中所示采集日期后七天内的序列为一组数据。B. 数据的最早采集时间为 2019 年 12 月 26 日，最晚采集日期为 2020 年 4 月 1 日。



附图 2：579 条新冠病毒全基因组序列采集国家分布情况。A. 共有 36 个国家和地区；B. 图中序列采集数目小于 10 条的国家归入“其它”一类。

附表一：579 条新冠病毒全基因组序列数据信息（可在

https://github.com/chunfenri/Transmission_decoder_SARS_CoV_2 获得）

附件一：对关键突变位点进行检测的原创 Python 程序，可在 GitHub 上获取

附件二：A.Betacoronavirus 主要物种的系统发生树;B.579 条序列的系统发生树；C.50 条序列的系统发生树的 mafft 文件、jmodeltest 文件、tree 文件