SARS-CoV-2 病毒进化的生物化学和数学教训：新型抗病毒战的途径

Nicolas Cluzel[1], Amaury Lambert[2,3], Yvon Maday[1], Gabriel Turinici[4], Antoine Danchin[5,6,*]

译者：尤从慧[7]

1. Tremplin Carnot SMILES, 4 Place Jussieu, 75005 Paris, France

2. Laboratoire de Probabilités, Statistique & Modélisation (LPSM), Sorbonne Université, Université de Paris, CNRS UMR8001, 4 place Jussieu, 75005 Paris, France

3. Centre Interdisciplinaire de Recherche en Biologie (CIRB), Collège de France, CNRS UMR7241, INSERM U1050, PSL Research University, 11 place Marcelin Berthelot, 75005 Paris, France

4. Ceremade, Université Paris Dauphine – PSL

5. Kodikos Labs / Stellate Therapeutics, Institut Cochin, 24 rue du Faubourg Saint-Jacques, 75014 Paris, France

6. 香港大学李嘉诚医学院生物医学科学学院, 21 沙宣道，薄扶林, 香港特区, 中国

7. 深圳大学生命与海洋科学学院，广东省深圳市南山区学苑大道 1066 号

邮 箱： adanchin@hkucc.hku.hk, antoine.danchin@normalesup.org

译者 => 邮 箱：cyou@szu.edu.cn

电话: +331 4441 2551; 传真机: +331 4441 2559

关键词

ddhCTP, D614G, F1757L, L37F, TN93, tRNA 核苷酸转移酶，非位似生长

摘要

在防止 COVID-19 传播的斗争中，重点是接种疫苗或重新激活用于其他目的的现有药物。然而病毒在繁殖时与其宿主的新陈代谢之间必然存在的紧密联系被系统地忽视了。在这里我们表明，所有细胞的新陈代谢都是由细胞基因组的核心构件—三磷酸胞苷（CTP）的可用性来协调的。这种代谢物也是病毒包膜合成及其基因组翻译成蛋白质的关键。这种独特的作用解释了为什么进化导致动物中很早就出现了一种抗病毒免疫酶--viperin，因为它能合成 CTP 的毒性类似物。这种依赖性所产生的制约因素指导了病毒的进化。考虑到这一点，我们用数学方法探索了发生在我们眼前的实时实验。我们因此几乎每天都在跟踪病毒基因组组成的进化，特别是在以引发病毒产生类似绽放烟花的伞形花序样的突变形式下将其与随着时间的推移而产生的后代联系起来。这其中的一些突变肯定会增加病毒的传播性。这样使我们弄清了病毒的几种蛋白质在这次进化中的关键作用，如它的核囊蛋白 N，并且更普遍地开始了解病毒是如何把宿主的新陈代谢联系起来，从而使之对自己有利。病毒在细胞中逃避 CTP 依赖性控制的一种方法是感染预期不会生长的细胞，如神经元。这可能是当前疫情中病毒在出乎意料的身体部位发展的原因。

介紹

关于 COVID-19 大流行的发展，有无数的文章在探讨。尽管数量如此之多，但由于我们人类的中心主义，这些研究都异常的集中在病毒本身上。当然，很多工作都在研究 SARS-CoV-2 病毒基因组的组成和结构、其编码的蛋白质及其感染动物的近亲病毒的细节。然而，主要集中于关于该病毒是如何利用宿主细胞新陈代谢的研究却很少。遏制该流行病的迫切性使研究者们重视疫苗接种或者更普遍地重视宿主免疫系统的参与。众所周知，遗憾的是，虽然有时相对容易产生一种既有效又无害的疫苗来对付一种广泛流行的疾病，但是也有情况相反的例子。现在仍然有一些非常严重且非常常见的疾病没有接种疫苗。接种疫苗有效的特别前提是病原体的后代能够保持足够长的时间，以防止疫苗引发的免疫反应逃逸。冠状病毒是由长基因组和包膜组成的病毒。长基因组可能导致非常高的突变率，但这些病毒启用了一种特殊的功能，即校对和纠正复制错误，从而避免了穆勒棘轮效应（Muller's ratchet）的普遍约束（见方框内的内容）[1]。这意味着虽然冠状病毒确实倾向于随着时间的推移产生遗传变异，但这些变异的数量仍然相当低。这种变异率看似非常有限，但在一次感染过程中产生的病毒颗粒数量是巨大的，而目前人类公认的感染人群超过两千万人。由此可见，每个核苷酸的突变率是每年每个位点大约有 8×10^{-4} 个变化 [2]。当然由于基因组中某些位置存在选择压力，突变率是非常不均一的。

在本文中，我们从 Fisher 提出的自然选择基本定理的角度出发对该病毒进行研究。该自然选择基本定理将环境适应性的进化和遗传变异联系起来 [3]。我们希望利用病毒的适应性进化所留下的痕迹来对该病毒进行研究。这种痕迹是以基因组序列的形式观察到的并且来源于生物化学约束条件下所产生的可供进化选择的偏差。然而，我们必须考虑到，具体的问题并不像人们所希望的那样明确：适应性是不知道的，时间标记也是不知道的（可以从系统发育树上估计的，或者直接用物理时间)。这促使我们采用足够稳健的程序来应对这些不确定性。尽管如此，这种分析的优点是能使我们对病毒的进化进行预测。因此，它是为流行病学或临床模型提供相关观测数据的一种明确手段。

在这种背景下，基于强调病毒与宿主代谢关系的前提下来探讨 SARS-CoV-2 是如何随着时间的推移在 COVID-19 传播的各个地方发生变异的细节似乎非常有趣。这应该可以让我们预见该病毒后代的一些变化，从而对该疾病的控制产生重要影响。通过对构成病毒基因组的核苷酸的代谢的制约因素的分析，我们发现其基因组中胞嘧啶（C）的含量受到强大的负性选择压力。导致随着时间的推移，胞嘧啶单磷酸出现系统性消耗 [4]。长期以来，人们一直认为这种系统性消耗与 APOBEC 脱氨酶家族对基因组中 C 含量的 "编辑 "形成主要的因果关系 [5,6]。现在我们知道正是动物细胞中嘧啶的代谢组成，尤其是三磷酸胞嘧啶的代谢组成 [7]，推动了相应的进化压力（图 1）。

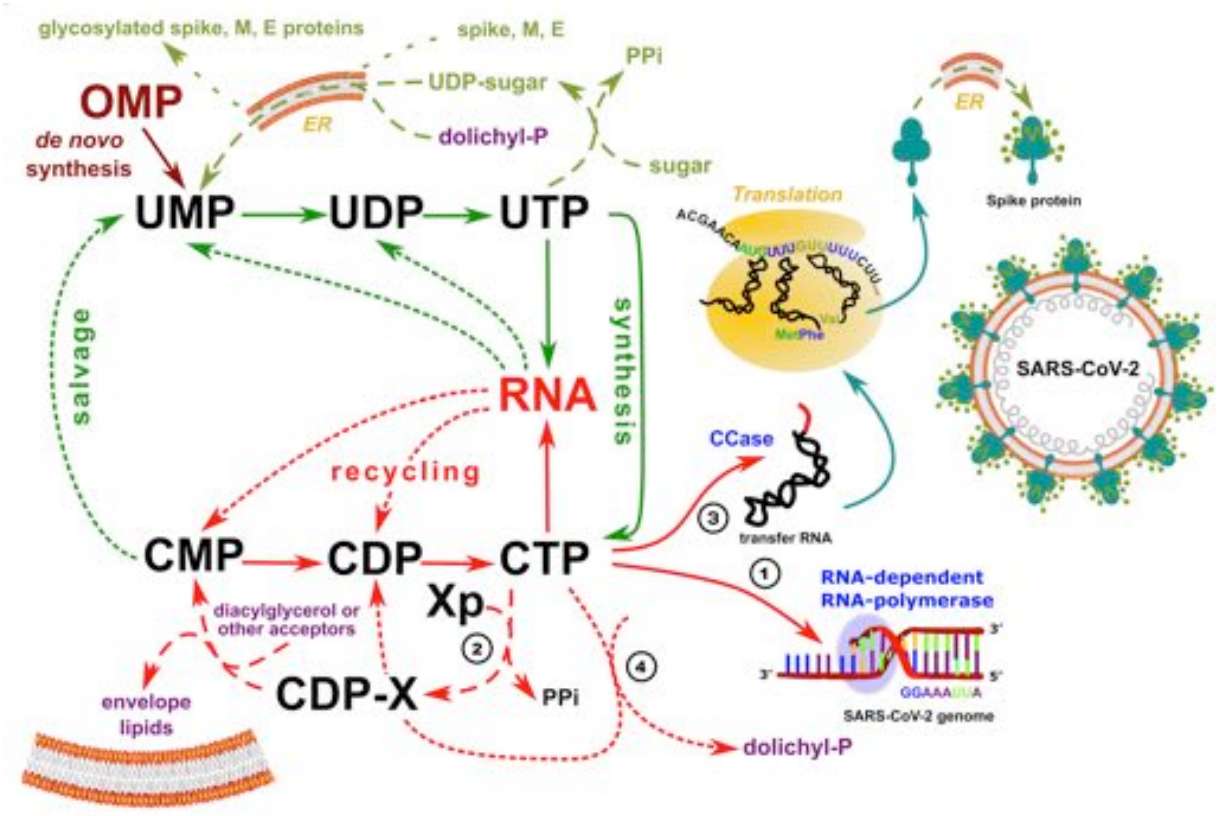


图 1.CTP 控制着构建活性 SARS-CoV-2 病毒所需的所有关键代谢步骤。1/ CTP 是病毒基因组的前体；2/ 其包膜的脂质来源于胞嘧啶基脂质核苷酸前体；3/ 宿主产生的所有成熟的转移 RNA 分子必须在其 3'OH 末端形成 CCA 三联体。4/病毒蛋白特别是其 S 蛋白的翻译后糖基化修饰需要内质网(ER)中的多酚磷酸锚定，而多酚激酶的活性特异性依赖于 CTP。详见正文及参考文献 [7]。

事实上，由于病毒复制的极端不对称性（病毒从其互补模板复制 50 到 100 次[8]）。这些高度依赖宿主的酶的基因组编辑效应只有在阴性 RNA 模板上发生 C→U 修饰时才会显著，这将导致病毒基因组上主要富集 A，也可能是由于作用于双链 RNA 的另一类脱氨酶 ADAR 将腺嘌呤脱氨成肌苷而导致的 U→C 的转换而导致的[9]。此外，APOBEC 和 ADAR 都是高度特异性的酶，这很难解释我们不断观察到的在病毒进化过程中广泛存在的 C→U 的转换。在这里，我们重点研究了基因组中 C 的动态缺失过程，并寻求这些缺失位点以及驱动这种变化的原因。在第一段中，我们总结了在代谢水平上解释这一显著现象的原因。随后，在文章的正文中，我们表明基因组中 C 含量的限制导致了特定病毒株系的生成，而这可以用来揭示病毒存在的一些重要的功能并揭示宿主反应的作用。

胞苷三磷酸（CTP）的生物合成是一种普遍存在的代谢必需条件，它控制着病毒的进化。

我们对产生病毒颗粒（病毒体）的构件的合成了解多少呢？在病毒感染期间，细胞通常会停止繁殖。细胞的所有资源都被迅速转移到病毒的繁殖上。然而，生长是生命的普遍属性。这意味着，几乎总是这样， 除分化的神经元外，一旦出现繁殖的机会病毒所面对的细胞的新陈代谢就被组织起来以让细胞生长。当它感染一个细胞的那一刻，同样，除了那些不繁殖的细胞之外，任何病毒都将因此必

须通过统筹分配病毒所需构件的可用性来管理代谢压力。在我们通常的物理空间（三维）中，生长带来了一个不可避免的约束力。细胞必须把它的细胞质的生长（所以也是三维的），包围它的膜的生长（二维的）和基因组的生长（一维的，因为核酸是线性聚合物）放在一起。然而，主要是在细胞质中生成的同一种新陈代谢产生了构建这三大构件所需要的建筑材料。所以，这里我们有一个类似于经济学家提出 "非同构 "增长问题时提出的问题[10]。不幸的是，由于生命是在 35 亿年的时间里从原始的新陈代谢中分几个阶段发展起来的[10]，我们可能会担心许多生物体已经找到了解决这一约束的特异性办法。这一点经常在生命形式的巨大多样性中得到见证。出乎意料的是，解决这一难题的方法似乎是通用的：一种单一的代谢物，核苷酸三磷酸(CTP)，已被病毒征用来完成这一目的[4,7]。

CTP 的关键作用出现在细胞代谢的四个重要地方，这些地方对新病毒的形成至关重要。1/它是构成病毒基因组的四种核苷酸之一的直接前体；2/CTP 是合成病毒包膜的脂核苷酸前体所必需的；3/人类的转移 RNA 是由 415 个不编码其 3'OH-CCA 末端三联体的基因合成的，而这个三联体序列是由一种特殊的核苷酸转移酶从 CTP 合成的 [12]。最后 4/通过复杂的糖基化来 "装饰 "蛋白质，与它们在内质网(ER)中的翻译同时进行，通过由一种使用 CTP 而不是 ATP 作为其磷酸供体的激酶产生的多酚磷酸锚定底物 [13]。此外，中间代谢也是建立在嘧啶代谢的原始基础上，它通过三磷酸尿苷（UTP）系统地循环利用它们，这使得 CTP 成为代谢物的核心并大大限制了它的可用性（图 1）。因此，偶然的复制错误将倾向于用尿嘧啶取代基因组中的胞嘧啶。

SARS-CoV-2 病毒的总体进化情况

利用 SARS-CoV-2 GISAID 数据库(https://www.gisaid.org)中收集到的现有序列数据，和其他人一样[14,15]我们重新构建了病毒进化的系统发育树。由于每个病毒基因组的序列，以及这些序列的鉴定日期都是相当精确的，这棵系统发育树使我们有可能探索随着时间的推移而出现的突变病毒的有序族谱。特别是，除非我们可以怀疑是由于同一位患者被两种或两种以上的病毒感染而导致的重组事件，否则当树的不同分支中出现两个相同的突变时，我们可以认为这是趋同进化的结果 [16]。在分析每个相关突变时，我们会逐一讨论引起趋同进化的原因。另外一种观察需要建立在树的形状根本不是均一的观点上（见下文）。我们确实注意到了 "伞形花序 "的存在，即在树的某一节点上，出现了大量的分支，显示新突变呈 "爆炸性 "出现（图 2）。因此，我们设计了一种数学方法，使我们能够明确地描述它们的特征。

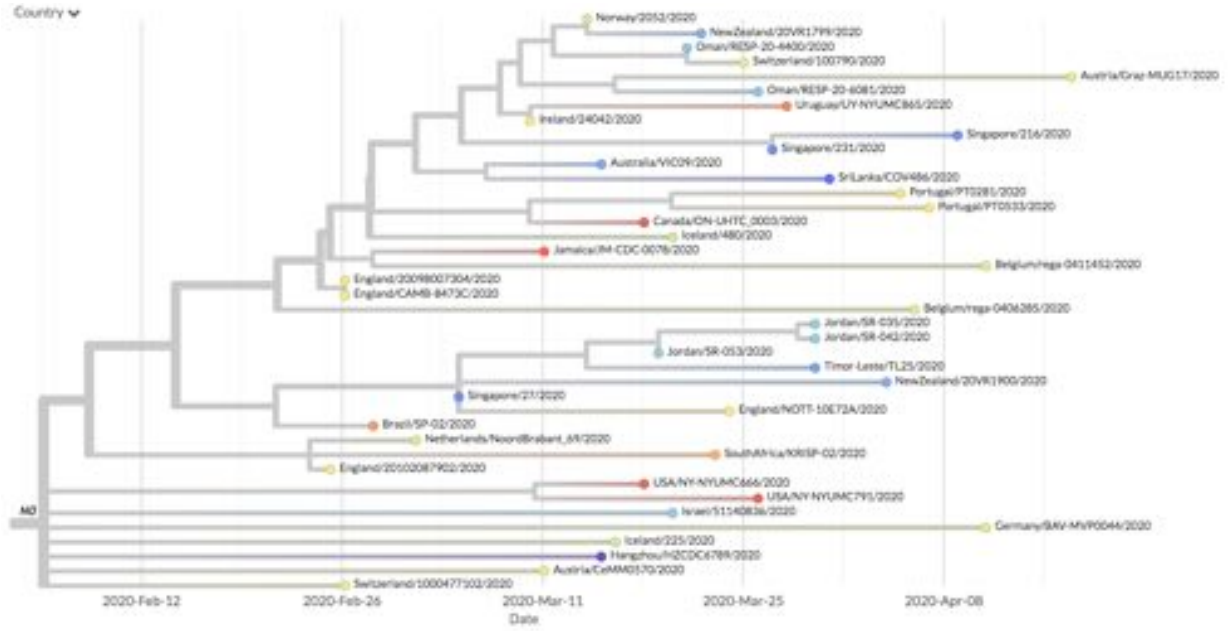


图 2. 一个由我们的统计方法检测到的伞形花序的例子。在节点 N_0，在子树的 40 个样本中，有 25 个不同的状态，并且有大量的分支。这种行为与其他子树的行为有很大不同。

引起这些伞形花序的原因是多方面的，但病毒重要功能的恶化可能是其根源，我们保留了几例这种情况进一步讨论（伞形花序的统计定义见材料和方法）。

描述和分析基因组中 C 含量的进化。

一般来说，冠状病毒基因组的进化倾向于使其 C 含量适应宿主的新陈代谢。更具体地说，SARS-CoV-2 随着疫情的发展，向着 C 含量较少的形式进化 [7]。然而，这种发展情况并不是均一的。

在感兴趣的两组数据中，77%的嘧啶之间的替换是由胞嘧啶到尿嘧啶的转换所代表。这些转换突变代表了第一组中所有确定的碱基替换的 48%（而在第二组中占 49%）。我们也注意到一个重要的不平衡发生在颠换突变的水平，有超过 73%的那些涉及到从嘌呤到嘧啶的颠换发生在第一组（74%在第二组）。然而，在这 73%中只有 20%导致胞嘧啶的发生（17%在第二组）。这再次表明了有利于尿嘧啶生成的趋势，并进一步证明了病毒突变过程的主要制约因素是细胞中每一个核苷三磷酸的可用性。这种不均一性在系统发育树的水平上也很突出。在分支 B4（包含样品的 20%）的水平上，与系统发育树的其他部分相比，C 的损失趋势非常明显（图 3）。

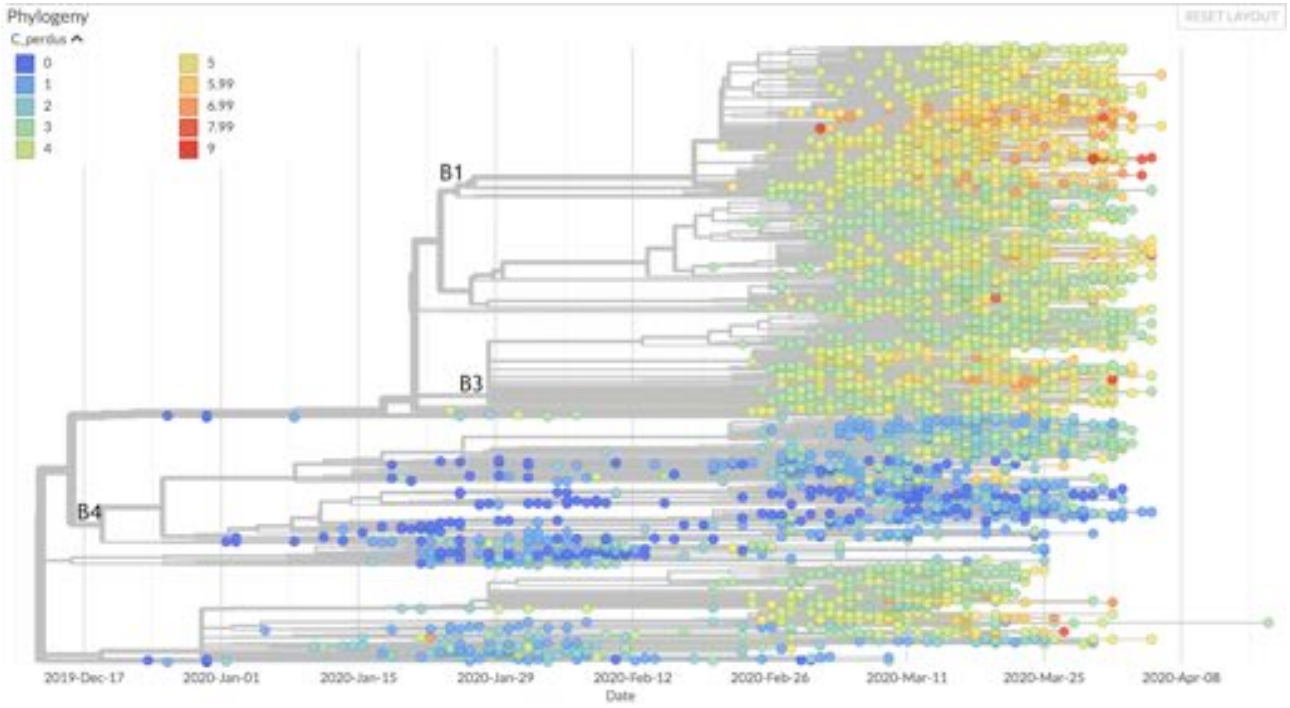


图 3：原始序列中 C 损失的热图。分支 1 和 4 可以很容易地通过其极端值进行区分。

在组成这个分支的 834 个样本中，平均每个样本损失了 1.5 个 C，比进化树的平均值少了 57%。有趣的是，在平均水平上，组成这个分支的病毒株与原始病毒株的差异最小。相比之下，在 B1 分支中，C 的损失尤为明显。在 1004 个样本中，平均每个样本丢失了 4.8 个 C，比系统发育树的平均值高出 42%。在这个分支中，病毒突变的速度似乎也在加快，颠换发生率比进化树的其他部分高出 20%（转换率也较高，但是比例差异不是很显著）。最后值得注意的是 B3 分支，也就是主要的伞形花序形成部位。与进化树的其他部位相比，它的嘧啶和嘌呤的转换率分别下降了 29%与 30%。

这种非均一性可能来源于许多限制因素：

1/ 基因组的结构本身。其必须折叠成一个紧凑的包膜并同时需要确保某些区域存在一些特定的 C 残基。例如控制复制 [8]或转录起源的 AACGAAC 区域[17]就是这种情况。在那些发生蛋白翻译的区域，C 存在的压力根据其在密码子三核苷酸中的位置而变化。当 C 位于密码子的第一个位置时，精氨酸、谷氨酰胺、组氨酸、亮氨酸或脯氨酸将被引入蛋白质。而组氨酸和谷氨酰胺由两个密码子家族编码，这在后面会讨论。对于精氨酸来说，其选择压力较小。因为 CGN 密码子可以用 AGR 密码子替代—我们在这里使用 IUPAC 惯例来标记核苷酸或氨基酸，例如 N 代表 aNy，R 代表 puRine 等 (https://www.bioinformatics.org/sms/iupac.html)。对亮氨酸含量的选择压力也较低，因为除了 CUN 密码子，该氨基酸还可以使用 UUR 密码子进行编码。在密码子的第二个位置上，C 再次被用来编码脯氨酸，也可以编码苏氨酸（ACN）、丙氨酸（GCN）和丝氨酸（UCN）。同样的，后一种氨基酸摆脱了 C 的可用性所带来的很大一部分限制因为它也可以使用 AGY 密码子。最后，密码子的第三个

位置受到的限制要小得多因为它可以被 U 取代。，但它在四个密码子（丙氨酸、脯氨酸、苏氨酸、缬氨酸）家族中也可以被 A 或 G 取代。UGY、AGY 和 NAY 这两个密码子家族是以嘧啶/嘌呤为中心来区分的。嘧啶用于保持编码残基的相同化学性质。当密码子使用 U 或 C 作为 3'端时可以编码化学性质类似的天冬氨酸、天冬酰胺、半胱氨酸、组氨酸和酪氨酸。最后，异亮氨酸由 3 个密码子(AUH)编码，涉及以 U 或 C 结尾的相关 tRNA [18]。

2/ 病毒蛋白的功能依赖于其序列中某些氨基酸的存在。例如，CCN 密码子编码的脯氨酸残基严格来说不是一个氨基酸，但其对病毒蛋白的关键结构域的折叠至关重要 [19]。

3 /进一步强调 CTP 的重要性。在进化过程中，先天性抗病毒免疫力招募了一种 viperin 酶的活性。它可以将 CTP 修饰成对病毒发育有毒性的形式，即 3'-脱氧-3′，4'-二脱氢-CTP（ddhCTP）[20]。这一途径的一个有趣的结果是通过降低基因组中 C 的含量从而使病毒在复制过程对这种核碱基的存在不那么敏感。由此可见，在一种相对富含 C 的病毒从动物宿主转移到人类的过程中向失去 C 的方向进化可能会短暂地伴随着其致病性的增加。但从长期来看，C 的丢失严重限制了病毒的进化出路，很可能会使病毒趋于衰减 [21]。

可以使我们对病毒蛋白功能提出建议的一些相关性的例子。

迄今为止，已经发现了数千种变异。我们可以沿着病毒的系统发育进化树来跟踪它们的出现，然后突出描述一些有趣的特征。这可以让我们预测它的一些未来变化。

导致翻译提前终止的突变

导致病毒蛋白质合成过早终止的突变预计会频繁出现。就本文的讨论内容，这种情况更有可能出现，因为翻译终止密码子 UAA、UAG 和 UGA 不含 C，因此有利于该核苷酸的消失。然而由于这些突变大多会导致无功能的多肽，所以一般来说，受影响的病毒很可能不会产生大量的后代。由此可见，当观察到这些非测序错误引起的突变时，这意味着该缺失蛋白的功能并不关键的或者说该缺失蛋白一直保持着可以使病毒繁殖的足够的功能从而得。然而，一些观察结果使我们能够为相关突变病毒可能存活下来的事实提供一个解释。下面是揭示该病毒一些有趣特征的三个例子。

例 1：在一株来自冰岛的菌株中，G1440A (Gly392Asp， Nsp2 蛋白) 和 G2891A (Ala876Thr，Nsp3 蛋白的泛素样域)的相继突变目前已在世界多个地方出现 [22]。这个序列最后以 C27661U 结束（在靠近蛋白质 Orf7a 的羧基端将氨基酸 Gln90 修饰成使翻译过早结束的终止子）。这种病毒蛋白存在于内质网、高尔基体和核周空间 [23]。在疫情流行的过程中，已经发现了几种类似的变异 [24]。值得注意的是，在该基因中已分离出几个缺失片段。这表明该缺失片段的功能对该基因来讲并不是必须的 [25]。然而，我们注意到在这些突变中许多像在这里讨论的突变保持了 Orf7a 下游的小型疏水性蛋白 Orf7b 基因的完整性。这种很小的蛋白存在于细胞器高尔基体中，并且在纯化的病毒

中也能找到它[26]。必须注意到，它在体内的合成是通过一个跨越 Orf7a 阅读框终止密码子的移码完成的（...GAA TGA TT...变成...GA ATG ATT...）。这可以解释为在这个区域中存在 Orf7a 和 Orf7b 的翻译冲突，从而这两个蛋白的表达都陷入了成本/效益的困境。因此，重要的是监测病毒的未来后代的这个区域，因为它可能导致形成有趣的毒性衰减的病毒。

例 2：另一个导致病毒蛋白过早翻译终止的连续突变是从 G11083U（蛋白 Nsp6，Leu37Phe）开始的。这种突变目前在世界范围内广泛分布。它很可能诱导蛋白与 ER 结合的更稳定并可能通过将病毒成份递呈到溶酶体进行降解来有利于冠状病毒的感染[27]；然后是 G1397A（Nsp2，Val378Ile），也可能有利于病毒的传播[28]。接着是 G29742U（病毒的 3'UTR）和 U28688C（同义突变）；随后我们有一对突变 C884U（又是 Nsp2，Arg207Cys [28]）和 G8653U（Nsp4，对包膜组装至关重要的蛋白[29]。相应的蛋白变化（Met2796Ile）位于该蛋白质的 ER 内腔域的边界。众所周知，ER 需要氧气才能正常的工作 [30]。而活性氧（ROS）与该腔室中蛋白质的错误折叠有关。Nsp4 有许多半胱氨酸残基，容易被氧化。蛋氨酸在母体病毒中的作用可能是作为对抗 ROS 的缓冲剂，所以突变体病毒会稍有减弱。这些突变之后是 A19073G（即 Nsp14 蛋白甲基酶域中的 Asp1869Gly 位点。该位点从 SARS-CoV-1 进化而来[31]。因此该突变很可能是中性突变），然后是其中一个突变能导致翻译终止的一对突变：G27915U，导致在 Orf8 的 N-末端产生翻译终止和 C29077U（同义突变）;接下来的突变以导致一对同义突变的 C19186U 和 G23608U 结束。SARS 相关的冠状病毒在这一区域是高度变异的。该区域在流行过程中发生变化，表明它受到持续的选择压力，有时会产生两个多肽 Orf8a 和 Orf8b [32]。它们属于在感染周期结束时表达的蛋白。因此监测它们在病毒毒力进化过程中的作用方式将非常重要。在这里展示的是在四个不同的国家和 7 个样本中出现的分支。从第一个突变到最后一个突变之间跨时六个星期。

例 3：这里我们有一个连续突变。该连续突变始于病毒基因组的 5'端内，C241U。接着是复制酶 Nsp12 中锌指末端的突变 C14408U（Pro314Leu），该突变出现在病毒进化树的许多分支中。在下文将详细讨论它（即伞形花序的起源）。在这个突变之后是一个在 S 蛋白中广泛存在的 A23403G（Asp614Gly）突变（下文也有讨论）、C3037U 突变（同义突变）、在钾通道蛋白 Orf3a 中的 G25563U（Gln57His）突变（该突变可能抑制了该蛋白的功能[33]）、前面讨论过的蛋白质 Nsp2 中的 C1059U (Thr265Ile)突变、蛋白酶 Nsp3 的 SUD-N 域中的三倍体 G4181A (Ala1305Thr)突变、 G4285U (Glu1340Asp)突变以及引起蛋白质 Orf8 在 106 个氨基酸（谷氨酸）处发生终止翻译的 G28209U 的突变。如前所述，许多突变包括 Orf8 的缺失经常被观察到。这再次表明，我们应仔细监测该区域的进化以寻找病毒的减弱形式。这个引起翻译终止的特殊突变是很重要的，因为在克罗地亚的一个样本以及另一个来自泰国的样本发现该突变。这两个样本位于两个显著分离的分支上，并且有一个月的差异。这里的突变序列与泰国样本相对应。

扭转病毒基因组失去胞嘧啶残基的趋势。

我们在进化树的上游分支点这里保留了 2 个这样的例子。这些病毒的后代似乎不再失去胞嘧啶，甚至倾向于重新获得胞嘧啶。这些例子如下（图 4）。

图 4. 在第一组数据集中基因组失去胞嘧啶残基的趋势获得逆转的两个子树。上图为第一个子树：展示的子样本是那些获得了最多 C 的样本。能够区别它们和其他分支上的几个孤立的样本。具有 M1 突变的节点直接连着那些分别与 C8782U 和 U28144C 突变相关的节点。**下图为第二个子树：**这棵树

包含了大多数 C 为中性突变的病毒株（包括获得的和丢失的）以及 3 株获得的 C 多于丢失的 C 的病毒株。

在数据集 1 中，有两个子树。其中第一个子树更像是一个亚洲子树，其节点根部与 M2 和 M3 突变有关。第二个子树包含来自北美洲和大洋洲的样本，其节点根部与 M6 和 M7 突变有关。第一个子树源于 Orf8 蛋白中一系列的 C8782U（同义突变）和 U28144C（Leu84Ser）突变，其功能已在上文讨论过。它确定了病毒变种的主要进化支[24]，包括 C24034U（同义突变）以及最后又是 Orf8 蛋白中的 U26729C（同义突变）与 G28077C（Val62Leu）双突变。由于这些是我们观察到的伞形花序现象的起源，所以我们认为这是 Orf8（8a 或 8b）作用的改变造成的。Orf8 区域的变异特别大，并且该区域已经明确地牵涉到种间传播 [34]。对此一个常见的假说是，该基因的改变对应着其在翼手目动物祖先活性功能的丧失 [35]。由于这些基因的突变体一般比人类来源基因的胞嘧啶含量更丰富 [21]，人们可能会问调节 CTP 合成酶的活性是否是该蛋白的功能之一。

事实上，第二个分支来自同一起源。在此基础上增加了控制病毒 RNA 特异性翻译的 Nsp1 蛋白的 U490A（Asp75Glu）突变 [36]，其与多功能 Nsp3 蛋白[37]中没有任何明确功能的酸性域中的突变 C3177U（Pro971Leu）系统相关。最后是病毒 3'UTR 区的外切酶、N7-甲基转移酶 Nsp14 的 A29700G 与 U18736C（Phe1757Leu）双突变。Phe1757Leu 修饰位于 Nsp3 蛋白两个结构域间界面的锌离子结合位点中间。因此可以推测，这种突变可以巧妙地改变校对过程纠正复制错误的方式，使其不太容易校正与病毒负链模板中的 A 相对应的 UTP 的插入。我们注意到，在 5 个样本中有 3 个样本获得了最多的 C，是通过从 U 到 C 的转变实现的。第一个样本，HongKong/HKPU1_2101，在 12877 和 23857 的位置同时出现了两个转换突变。这些突变是同义突变，它们不太可能改变复制-校正机制。第二个样本 USA_SC_3571 和第三个样本 Australia/VIC209 分别在 11635 和 19713 的位置上显示了相同类型的转换突变，也都是同义突变。最后两个样品，USA/WI-33 和 USA/WI-28，来自于 ORF9b 末端的突变，即 29567 位的 A 到 C 的颠换突变。

就数据集 2 而言，这种趋势的逆转主要涉及拉丁裔美国人的病毒株。这几个连续的突变（C241U、C14408U 以及之前讨论的并且与病毒基因翻译产生终止密码子相关的 A23403G）之后是 C3037U（同义突变），以及与 N 核帽 N 基因 203-204 位置密码子重叠的 G28881A、G28882A 及 G28883C 三突变。它们将一个精氨酸-甘氨酸二肽突变为赖氨酸-精氨酸二肽。这改变了蛋白质的正电荷，可能有助于提高其在病毒衣壳中组装病毒基因组的功能，这正如下文所述的与伞状花序的出现有关（36）。在这三重突变之后，我们看到几个在基因组中失去 C 的趋势的逆转。在核壳 N 基因中又再次发现 U29148C（Ile292Thr），然后又在 Orf6 基因中发现 U27299C（Ile33Thr），结果是产生这样的一组样品：它们在最坏的情况下获得的 C 与失去的一样多。在子树的 39 个样本中，也有 3 个样本获得的 C 比失去的 C 多一个（巴西/RJ-763、阿根廷/赫里塔斯_HG007、阿根廷/赫里塔斯_HG006）。每一次，最后一个 C 的获得都来自于病毒基因组 3'UTR 最后面的一个腺嘌呤的颠换

突变，分别发生在 14164（Met233Leu）、16076（Asp870Ala）和 29782 的位置上。总的来说，似乎是核衣壳的变化最有利于扭转失去 C 的趋势，确实，这种蛋白在病毒感染过程中高水平表达，调节着病毒的复制/转录过程，这可能是导致这一显著观察结果的原因 [39]。

伞状花序的出现

C3037U，（C241U，A23403G）以及 C14408U 一系列突变出现在 10 个子树的上游。我们认为这非常显著（见图 5 和材料与方法）。

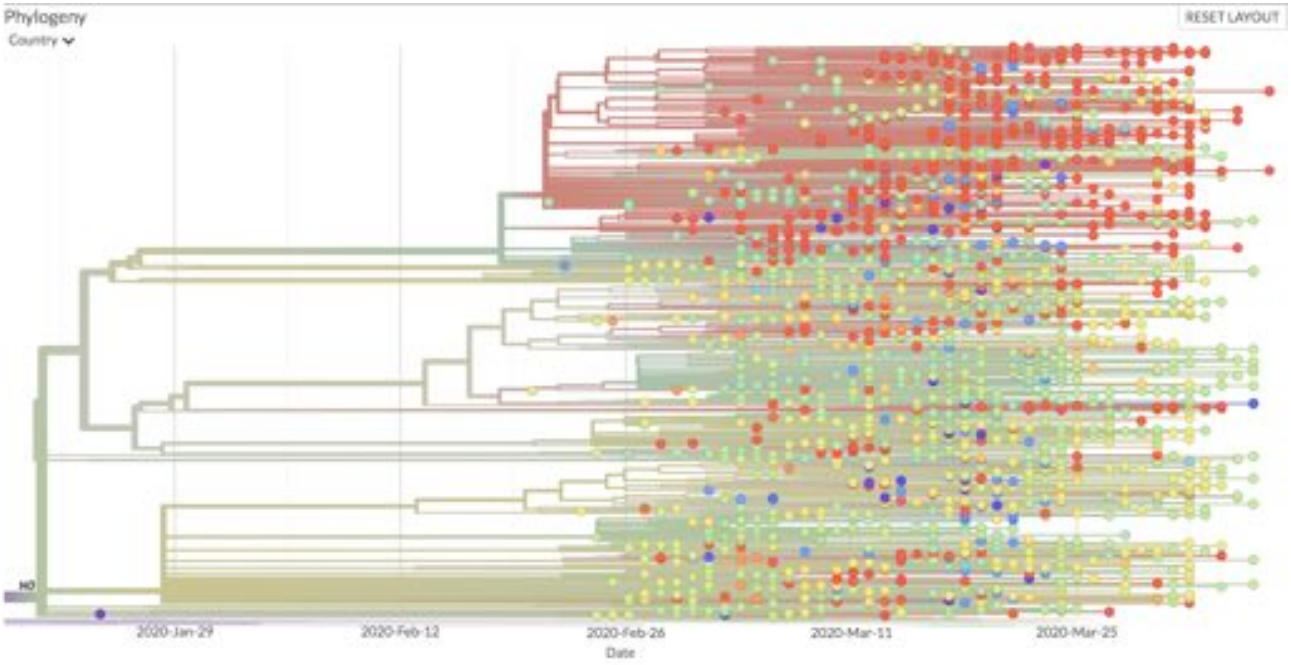


图 5：伞状花序的例子 这里显示的子树包含了用我们使用的方法所定义的 20 个最显著伞状花序中的 10 个。N0 节点是突变 C14408U 出现的地方。

同义突变 C3037U 位于蛋白质 Nsp3 的泛素样结构域 1 的末端。这产生了一个促使阅读框发生变化以及可能会降低 ORF1a 区蛋白质翻译效率的序列（UUUUUU）。我们非常频繁的观察到 C241U 突变 [40]。它存在于启动病毒复制的区域。因此我们可以认为这个突变可能会改变复制的频率。

A23403G 是一个广泛分布的非同义突变，它导致 S 蛋白中 614 位的天冬氨酸被甘氨酸所替代，而该天冬氨酸位点被病毒用来结合宿主细胞的受体。由于这个原因，以前的一些分析表明，该突变在病毒的传播中具有重要作用 [41,42]。在这里，它是组成主要伞形花序的一部分的这一事实可以被认为是支持这一解释的额外论据。在以 "锌指 "结尾的 Nsp12 蛋白质的 NiRAN 结构域（尼德病毒 RdRp 相关核苷酸转移酶）末端之后的 C14408U 突变将脯氨酸变为亮氨酸 （Pro314Leu）结束。NiRAN 结构域是病毒复制所必需的，作为核苷酸转移酶，其优先选择 UTP 作为底物的功能尚未明确 [43]。

突变体中修饰的脯氨酸是双脯氨酸二肽的一部分，它的功能是作为分离 NiRAN 结构域和后面相继结构域之间的铰链。

第二个伞形花序与前一个伞形花序有几个共同的部分，从相同的序列 C3037U、(C241U、A23403G)和 C14408U 开始。然而它接着发生了一系列连续突变，正如我们之前看到的那样，导致核帽 N 的变化（G28881A，G28882A，G28883C）。值得注意的是，这种变化可能对病毒基因组在帽壳中的组装有相分离的作用 [38]。这可能会提高病毒的传播效率，从而促进伞形花序的形成。事实上，这是一个涉及 G 的突变群，这非常有趣的。这可能源于这个三重突变跨越了一个 GGGG 序列。

我们之前已经看到，突变 G11083U（蛋白 Nsp6，Leu37Phe）已经启动了另一个连续突变并导致病毒蛋白的翻译过早终止。在这里，这种突变在伞形花序的根部广泛分布。如前所述，它可能是通过将病毒成份递呈到溶酶体进行降解来有利于冠状病毒的感染，。这肯定会有利于伞形花序的形成。在产生第一个伞形花序序列中的突变之后是蛋白质 Orf3a 中的 G26144U（Gly251Val）突变。尽管该蛋白质的确切功能仍有待商榷， 但其形成了对先天免疫反应很重要的钾通道 [44]。后续的突变是 C14805U（同义突变）和 U17247C（同义突变）。这种接连不断的突变表明，蛋白质 Nsp6 的第一个突变，或许还有第二个突变，是导致伞形花序形成的主要原因 [27]。第一个突变的作用被后续的第二个伞形花序进一步证实。在第二个伞形花序处连着一个四重突变： 在多结构域蛋白酶 Nsp3 的 G2M 结构域之前的域间区域的 C6312A （Thr2016Lys），然后与三个 C→U 突变相关联，因此预计其发生突变的频率更高；Nsp12 蛋白质的 NiRAN 结构域中的 C13730U （Ala88Val）；C23929U （同义突变）；最后是位于核帽蛋白 N 起始序列的 C28311U （在 4 个 C 的序列中，Pro13Leu）。

导致产生伞形花序的第二种连续突变是 Orf8 蛋白质中的 C8782U （同义突变）和 U28144C （Leu84Ser）（其功能在前文中已经讨论过并且定义了一个重要的病毒变种支系 [24]），并最终以 C26088U（同义突变）结束。Leu84Ser 突变与上面讨论的 S 蛋白的 Asp614Gly 突变有显著的共同进化 [37]，这使得它成为正向选择的另一个可能的候选者并增强病毒的传染性，因此形成伞状花序。

转换/颠换突变频率的改变

在分支上游的突变中，显示出显著的转换/颠换持续改变的变化是突变 C17747U，它将 Nsp13 蛋白中的一个脯氨酸残基突变成亮氨酸残基（图 6 及材料与方法）。

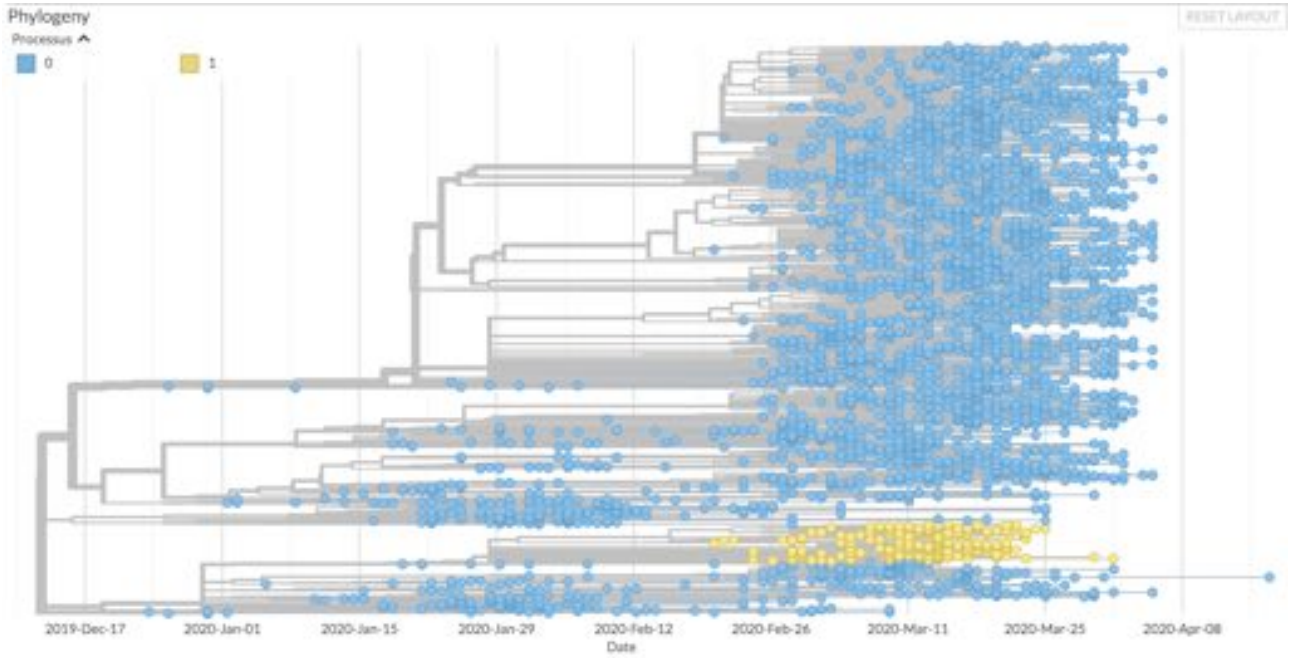


图 6.分子进化过程中变化显著的后代之一。C17747U 突变产生的后代用黄色显示，其进化过程由 6 参数的 TN93 模型建模（过程 1）。树的其余部分（蓝色叶片）由 3 参数的 TN93 模型建模。

这个突变影响了该蛋白质的具有核苷三磷酸酶活性的结构域，其确切作用虽然未知但与复制校对活性相关[45]。我们可以提出它的功能是参与病毒复制产物的质量控制，例如通过稳定核苷酸的 "反式 "形式，从而避免错配导致的颠换突变。事实上，这种蛋白已经被确定为导致病毒基因组多样性显著改变的蛋白之一 [16]。因此，位于该树下游的突变类型存在明显的变化，是该蛋白这个相应区域具有决定性作用的有力论据。此外，如果这种突变以一种偏向性的方式增加了突变的频率，我们可以预期接下来的产生的病毒后代会导致病毒的衰减。然而，由于这改变了病毒进化的局面，这种进化可能会导致 "创新性 "的突变从而改变病毒的致病性。尤其是在由于共同感染而导致的病毒重组的条件下会有利于病毒的进化。这也是支持选择强有力的公共卫生政策的另一个论据，这种政策倾向避免产生病毒群的感染。

结论和观点

COVID-19 疫情是一个实时发生的且全球化的病毒进化实验。值得注意的是，我们既不知道该病毒的真正起源 [46]，也不知道它将把我们引向何方。这就解释了为什么绝大多数关于 SARS-CoV-2 病毒及其进化的研究基本上都是描述性的。在这里，我们尝试利用该病毒正在发生的进化过程并利用假说驱动的数学方法来研究与其相关的一些制约因素。我们以为病毒颗粒扩增提供必备物质的宿主细胞的代谢组成为基础，指出了病毒后裔进化模式的具体变化。随着时间的推移，病毒基因组组成的变化见证了这一变化。以该基因组组成中广泛分布的 C 到 U 的变化为基础，我们找出了变化放向转向的节点。该变化有利于颠换而不是转换，最终将 C 到 U 的趋势逆转为 U 到 C 的富集或产生伞状花序也就是在进化树中突然出现多个分支的现象。这使我们能够指出病毒的一系列功能正在朝向更

有利于其传播的方向演化（如之前发现的 S 蛋白的 Asp214Gly 突变，也包括 Orf3a 钾通道的 Gln57His 突变）。我们还注意到 Orf8 是一个可能持续性竞争表达两个依赖移码的重叠蛋白 Orf8a 和 Orf8b 的位点。同样 Orf7 上的不稳定区域可能促使一个极小膜蛋白 Orf7b 的合成，但其功能至今仍未知。最后，趋势的反转变化倾向于 U 而非 C。这说明核帽蛋白 N 可能参与了宿主 CTP 合成的控制并提示其是未来控制病毒发展的一个有趣的靶点。我们希望这种数学和生化知识的结合能帮助我们设计出更多的战略方针来对抗 COVID-19 的可怕后果。我们注意到，病毒在细胞中逃避 CTP 依赖性控制的可能方式之一是感染预期不会生长的细胞，如神经元。这就可能解释了在目前疫情中观察到的病毒在出乎意料的身体部位发展的原因。

致谢

AL 感谢来自法国大学生物交叉学科研究中心（CIRB）的经费以及 CIRB 中心 SMILE（Stochastic Models for Inference of Life Evolution）研究小组成员们对构建 COVID-19 流行病模型提出的富有成效的讨论。AD 感谢 Stellate Therapeutics 对其实验室的支持。

材料和方法

数据处理

截至 2020 年 4 月 17 日，第一数据集共从 GISAID[47]数据库中恢复了 4792 条 SARS-CoV-2 序列。只有长于 25,000 bp 的 SARS-CoV-2 人类宿主病毒的基因组被保留。取样日期不够明确的序列（没有采集日期，有时是没有采集月份）也被丢弃。对于多次出现的序列，只保留第一个分离株的序列。我们还重用了 Nextstrain 团队的工作，并丢弃了那些分歧较大或不稳定的样本，而他们自己也将其排除在外（github.com/nextstrain/ncov/blob/master/defaults/exclude.txt）。以 NC_045512 为参考序列，鉴定出 26 个编码区（Nsp1、Nsp2、Nsp3、Nsp4、Nsp5、Nsp6、Nsp7、Nsp8、Nsp9、Nsp10、Nsp11、Nsp12、Nsp13、Nsp14、Nsp15、Nsp16、S、ORF3a、E、M、ORF6、ORF7a、ORF7b、ORF8、N 和 ORF10）的序列。数据处理后保留的序列总数为 4088 个。于 2020 年 7 月 6 日利用 Nextstrain API[48]检索到第二个数据集的 3246 个序列，其中 510 个序列与第一个数据集共同。

在此我们注意到，随着时间的推移，可用的数据不断发生改变，一些序列被从样本中删除，另一些则进入数据库。另外，提取大样本的序列一般是很困难的，这使得建立一个可以实施正确统计方法的统一的数据存储库变得异常困难。非常遗憾的是，尽管一些主要研究机构提出了建议，但大多数具有全球重要意义的病毒序列还没有存储在国际序列数据库中 [49,50]。

系统发育重建

重建过程首先将所有序列与参考序列进行序列比对。在这里不考虑核苷酸插入和缺失的问题，只研究潜在的碱基替换。我们使用程序 MAFFT[51]来进行序列比对。在比对的过程中一些比对不明确的区域被突显出来。例如，基因组的一些区域可能会显示出高不稳定性和变异性，这取决于用于比对的算法的参数。为了克服这个问题，我们使用了与 Nextstrain 团队使用的相同的掩码。因此在替换过程中不考虑 18529、29849、29851、29853 位点以及基因组的前 130 个位点和后 50 个位点。然后，我们使用一般时间可逆（GTR）模型，利用 IQTREE 软件[52]来推断实际的替代过程。由于没有考虑到进化的时间，第一棵系统发育树是一个相对粗糙的版本。我们使用 Treetime 软件[53]通过考虑序列的采样日期来完善这棵树。然后，重建与采样序列相关的具有最大似然的树。该软件还通过最大似然推断出样本原始序列的组成，以及生成这些共同原始序列最可能日期的 90%置信区间。一旦建立了系统发育树，那么我们就可以在最大似然意义上重建每个样本的突变出现的顺序。为了实现系统发育树的可视化以及绘制图 2 至图 6，我们使用了 Nextstrain 开发的 Auspice 程序并做了一些修改来展示我们感兴趣的参数。为此，我们开发了一个 Python 脚本来修改用来输入 Auspice 程序的 JSON 文件。这使我们能够通过添加参数，如一个样本与参考序列相比获得或失去的 C 的数量，来丰富该软件的可视化功能并能够演示生成的原始系统发育树。

识别伞形花序

我们在识别伞形花序时面临的主要难题是从系统发育树上选择样本时存在偏差。例如，由于不同国家的卫生政策和资源不同，一些医院的样本可能比其他医院多。为了避免选择因过度采样而导致出现的伞形花序的节点，我们选择开发了一种量身定做的统计方法。

子树是以主树的一个节点为根的任何节点和叶子的集合。我们的想法是利用每个子树所代表的国家的特性所提供的信息：一种病毒株越容易传播，预计观察到它的国家数量就越多。为了实施这一启发式方法，需要控制两个因素：树的大小（生根于不同日期的两棵深度不等的树自然会显示出不同的国家多样性）和采样的异质性（采样和测序强度不同的国家出现在某个子树中的概率不同）。
这两个因素是相互影响的，因为一棵树的大小（如叶子的数量）显然会随着采样强度而变化。控制这种相互作用的一种方法是用树的总长度，或枝条长度的总和，并以时间为单位来衡量树的大小。事实上，上述做法对过度采样并不十分敏感，因为同时同地存在的多个采样序列，导致子树的长度接近于零。

为了控制长度因子 L 对所代表的国家数量 N 的影响，我们尝试研究典型系统发育树中 N=f(L)的关系，这样我们就可以找出在已知长度 L 的情况下，所代表的国家数量超过预期 f(L)的子树。一个简单的统计模型是假设在一棵长度为 L 的树上，第 i 个国家的出现次数是参数 $\theta_i L$ 的泊松分布，而且这些数字

是独立的。如果 K 是 Nextstrain 引用的国家总数，那么长度为 L 的树所代表的国家数 N 因此就是 K 个不依赖于参数 1-exp(-θ_iL)的伯努利变量之和。例如，如果我们假设将国家分为两组，k_1 "频繁 "强度 θ_1，k_2 "罕见 "强度 θ_2，且 $\theta_2 <<< \theta_1$，则 N 有一个平均 K - $k_1$exp(-θ_1L)- $k_2$exp(-θ_2L)。

当 L 大时，其结果为 K-$k_2$exp(-θ_2L)。

此外，当 L 较大时，假设 θ_2L=O(1)，N 的分布近似等于 k_1+N_2，其中 N_2 遵循参数 k_2(1-exp(-θ_2L))的泊松定理。

因此，我们采用参数化方法：

N = a - bexp(-cL)，对参数的解释如下：a 是最大国家数，b 是采样/测序强度低的国家数，c 是每单位树长的这些国家的存在密度。在零假设下，N 的分布为 a - b + N_1，其中 N_1 遵循参数 b(1 - exp(-cL))的泊松定理。最后，我们选取了 20 个最显著的伞形花序，即与我们的推算式所预期的结果偏差最大的那些。这样，我们就可以重建族谱以及重建在每个产生伞形花序的节点上游连续出现的突变。这使我们能够确定其中形成一些节点常见的连续突变序列，从而确定那些引起大多数具有统计学意义的伞形花序的突变序列。此外，我们还限制了节点的自动选择，使被选择的节点不存在于另一个节点的族系中。因此，所选的伞形花序是相互独立的，即使明显它们可能有共同的祖先。为对存在于同一族系中的两个节点之间做出仲裁选择，我们系统地保留了最古老的节点，从而保留了最密集的树。

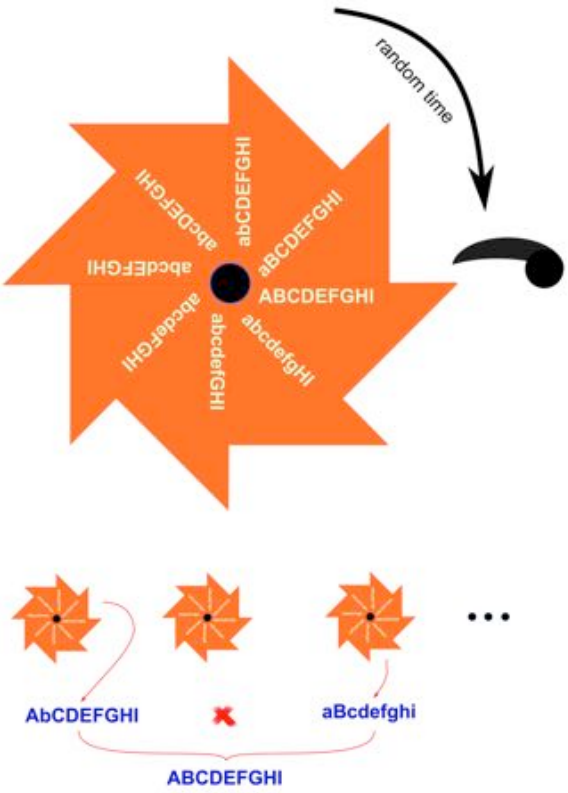
检测分子进化过程中的变化

我们对一些子树中的碱基替换过程与树的其他部分中观察到的过程相比是否有统计学差异的问题进行了研究。为此，我们使用了经典的 Tamura 和 Nei[54]的 3 参数（嘌呤转换速率、嘧啶转换速率和颠换速率）TN93 模型，并允许这三个参数值在候选节点 N_i 的下游采取与它们在树的其他部分所采取的不同的值。接下来，我们使用第二个（6 参数）模型。第二个模型嵌套在第一个（3 参数）模型中，检验统计量为似然比 2ΔI=2(l_1 - l_0)，其中 l_0 为 H_0 假设下的对数似然（3 参数 TN93 模型估计树的所有因素），l_1 为 H_1 假设下的对数似然（6 参数 TN93 模型，对每个选定节点下游的参数进行局部区分）。然后我们将似然比与 3 个自由度的/chi² 分布进行比较，其在具有 5%显著性时的阈值为 7.81。然后，我们可以找出其进化过程变化显著的节点，并量化不同碱基替代率的变化率。也就是说我们可以拒绝 H_0 假设的 3 参数 TN93 模型比 6 参数 TN93 模型预估出更好的树替代速率的节点。我们选择自己用 Python 实现这些模型，就是为了拥有这种参数化的灵活性。该程序使我们能够确定感兴趣节点下游存在的一组节点和叶子，并通过计算似然比和不同的替代速率来进行假设检验。

文本框

穆勒棘轮效应（Muller’s Ratchet）

生物学是建立在物理学规律之上的。因为它在大约 300K 的温度下发展，所以它受到普遍存在的热噪声的压力，其涉及的能量与生物化学的化学键所涉及的能量相差不大。由此可见，正在发生的以及组成生物的反应不能以严格的可重复性来发展。反应过程中不可避免的误差会使反应的产物与它的本来面目有所不同。基因组复制无法摆脱这种约束。这样引起的后果是，在病毒的后代中，总会出现一些变异体。当它们携带了基因组的改变时，就被命名为突变体。在大多数情况下，与这些突变体对应的是四个核苷酸之一被突变为不同的核苷酸。粗略的来估计，这个过程是随机的，即突变体的位置可以在基因组的任何地方，同时一个核苷酸可以被其他三个核苷酸中的任何一个所替换。随着时间的推移，基因组的所有核苷酸都有可能突变成其他核苷酸。这将影响到病毒繁殖所需要的功能，有些突变会继续传播，而有些突变则会最终没有传递下去。某种变异回复突变到原始状态的概率非常低。这就迫使进化总是向前发展，没有回头的可能。于 1932 年，这一过程被赫尔曼-穆勒注意到。此后被称为 "穆勒棘轮效应" [55]。显然，大多数突变极有可能导致基因组中被改变的区域所编码蛋白的功能部分或全部丧失。因此，从长期来看而不是从短期来看，这通常会导致病原体繁殖和毒力功能的减弱。这就是为什么 Louis Pasteur 和他的继承者们能够幸运地分离出减毒的生物体。在一些罕见的情况下，这些生物体可以被用来给感染者接种疫苗 [56]。然而，一旦在可能发生重组的情况下发生不同突变体的共同感染，接种疫苗就会变得毫无成效。因为两种不同的突变体可以重组为病原体的原始祖先形式而抹杀了病毒衰减的全部好处。这样就更有害了，因为病毒的古老形式往往也是最容易传播的形式。



方框文字图。穆勒棘轮效应和重组 该图转载自参考文献 [57]。基因（大写字母）以不同的形式（小写字母）随机突变。突变以棘轮状积累，因为还原到起初母体形式的概率可以忽略不计。这对于不同来源的病毒来说是独立发生的。然而，如果不同来源的病毒碰巧存在于同一个细胞中，它们可以重新组合。这使得它们可以重新创造出病毒的原始祖先形式。

参考文献

[1] M. Romano, A. Ruggiero, F. Squeglia, G. Maga, R. Berisio, A structural view of SARS-CoV-2 RNA replication machinery: RNA synthesis, proofreading and final capping, Cells. 9 (2020) 1267. https://doi.org/10.3390/cells9051267.

[2] A. Lai, A. Bergna, C. Acciarri, M. Galli, G. Zehender, Early phylogenetic estimate of the effective reproduction number of SARS-CoV-2, J. Med. Virol. 92 (2020) 675–679. https://doi.org/10.1002/jmv.25723.

[3] R.A. Fisher, The genetical theory of natural selection, Clarendon Press, Oxford, 1930. https://doi.org/10.5962/bhl.title.27468.

[4] A. Danchin, P. Marlière, Cytosine drives evolution of SARS-CoV-2, Environ. Microbiol. 22 (2020) 1977–1985. https://doi.org/10.1111/1462-2920.15025.

[5] P. Simmonds, Rampant C→U hypermutation in the genomes of SARS-CoV-2 and other coronaviruses: causes and consequences for their short- and long-term evolutionary trajectories, MSphere. 5 (2020) e00408-20. https://doi.org/10.1128/mSphere.00408-20.

[6] P.C.Y. Woo, B.H.L. Wong, Y. Huang, S.K.P. Lau, K.-Y. Yuen, Cytosine deamination and selection of CpG suppressed clones are the two major independent biological forces that shape codon usage bias in coronaviruses, Virology. 369 (2007) 431–442. https://doi.org/10.1016/j.virol.2007.08.010.

[7] Z. Ou, C. Ouzounis, D. Wang, W. Sun, J. Li, W. Chen, P. Marlière, A. Danchin, A path towards SARS-CoV-2 attenuation: metabolic pressure on CTP synthesis rules the virus evolution, 2020. https://doi.org/10.1101/2020.06.20.162933.

[8] I. Sola, F. Almazán, S. Zúñiga, L. Enjuanes, Continuous and discontinuous RNA synthesis in coronaviruses, Annu Rev Virol. 2 (2015) 265–288. https://doi.org/10.1146/annurev-virology-100114-055218.

[9] S. Di Giorgio, F. Martignano, M.G. Torcia, G. Mattiuz, S.G. Conticello, Evidence for host-dependent RNA editing in the transcriptome of SARS-CoV-2, Sci Adv. 6 (2020) eabb5813. https://doi.org/10.1126/sciadv.abb5813.

[10] J. Alonso-Carrera, C. de Miguel, B. Manzano, Economic growth and environmental degradation when preferences are non-homothetic, Environ Resource Econ. 74 (2019) 1011–1036. https://doi.org/10.1007/s10640-019-00357-4.

[11] Freeman J Dyson, Origins of life, Cambridge University Press, Cambridge, UK, 1985.

[12] K. Wellner, H. Betat, M. Mörl, A tRNA's fate is decided at its 3' end: Collaborative actions of CCA-adding enzyme and RNases involved in tRNA processing and degradation, Biochim Biophys Acta Gene Regul Mech. 1861 (2018) 433–441. https://doi.org/10.1016/j.bbagrm.2018.01.012.

[13] P. Shridas, C.J. Waechter, Human dolichol kinase, a polytopic endoplasmic reticulum membrane protein with a cytoplasmically oriented CTP-binding site, J. Biol. Chem. 281 (2006) 31696–31704. https://doi.org/10.1074/jbc.M604087200.

[14] C. Wang, Z. Liu, Z. Chen, X. Huang, M. Xu, T. He, Z. Zhang, The establishment of reference sequence for SARS-CoV-2 and variation analysis, J. Med. Virol. 92 (2020) 667–674. https://doi.org/10.1002/jmv.25762.

[15] X. Yang, N. Dong, E.W.-C. Chan, S. Chen, Genetic cluster analysis of SARS-CoV-2 and the identification of those responsible for the major outbreaks in various countries, Emerg Microbes Infect. 9 (2020) 1287–1299. https://doi.org/10.1080/22221751.2020.1773745.

[16] L. van Dorp, M. Acman, D. Richard, L.P. Shaw, C.E. Ford, L. Ormond, C.J. Owen, J. Pang, C.C.S. Tan, F.A.T. Boshier, A.T. Ortiz, F. Balloux, Emergence of genomic diversity and recurrent mutations in SARS-CoV-2, Infect. Genet. Evol. 83 (2020) 104351. https://doi.org/10.1016/j.meegid.2020.104351.

[17] Y. Yang, W. Yan, B. Hall, X. Jiang, Characterizing transcriptional regulatory sequences in coronaviruses and their role in recombination, BioRxiv. (2020). https://doi.org/10.1101/2020.06.21.163410.

[18] H. Grosjean, V. de Crécy-Lagard, C. Marck, Deciphering synonymous codons in the three domains of life: co-evolution with specific tRNA modification enzymes, FEBS Lett. 584 (2010) 252–264. https://doi.org/10.1016/j.febslet.2009.11.052.

[19] S.S. Rout, M. Singh, K.S. Shindler, J. Das Sarma, One proline deletion in the fusion peptide of neurotropic mouse hepatitis virus (MHV) restricts retrograde axonal transport and

neurodegeneration, J. Biol. Chem. 295 (2020) 6926–6935. https://doi.org/10.1074/jbc.RA119.011918.
[20] E.E. Rivera-Serrano, A.S. Gizzi, J.J. Arnold, T.L. Grove, S.C. Almo, C.E. Cameron, Viperin reveals its true function, Annu. Rev. Virol. 7 (2020) annurev-virology-011720-095930. https://doi.org/10.1146/annurev-virology-011720-095930.
[21] J. Armengaud, A. Delaunay-Moisan, J.-Y. Thuret, E. van Anken, D. Acosta-Alvear, T. Aragón, C. Arias, M. Blondel, I. Braakman, J.-F. Collet, R. Courcol, A. Danchin, J.-F. Deleuze, J.-P. Lavigne, S. Lucas, T. Michiels, E.R.B. Moore, J. Nixon-Abell, R. Rossello-Mora, Z.-L. Shi, A.G. Siccardi, R. Sitia, D. Tillett, K.N. Timmis, M.B. Toledano, P. van der Sluijs, E. Vicenzi, The importance of naturally attenuated SARS-CoV-2in the fight against COVID-19, Environ. Microbiol. 22 (2020) 1997–2000. https://doi.org/10.1111/1462-2920.15039.
[22] S. Liu, J. Shen, L. Yang, C.-D. Hu, J. Wan, Distinct genetic spectrums and evolution patterns of SARS-CoV-2, Health Informatics, 2020. https://doi.org/10.1101/2020.06.16.20132902.
[23] C.A. Nelson, A. Pekosz, C.A. Lee, M.S. Diamond, D.H. Fremont, Structure and intracellular targeting of the SARS-Coronavirus Orf7a accessory protein, Structure. 13 (2005) 75–85. https://doi.org/10.1016/j.str.2004.10.010.
[24] J.-S. Kim, J.-H. Jang, J.-M. Kim, Y.-S. Chung, C.-K. Yoo, M.-G. Han, Genome-wide identification and characterization of point mutations in the SARS-CoV-2 genome, Osong Public Health Res Perspect. 11 (2020) 101–111. https://doi.org/10.24171/j.phrp.2020.11.3.05.
[25] A. Addetia, H. Xie, P. Roychoudhury, L. Shrestha, M. Loprieno, M.-L. Huang, K.R. Jerome, A.L. Greninger, Identification of multiple large deletions in ORF7a resulting in in-frame gene fusions in clinical SARS-CoV-2 isolates, J. Clin. Virol. 129 (2020) 104523. https://doi.org/10.1016/j.jcv.2020.104523.
[26] S.R. Schaecher, J.M. Mackenzie, A. Pekosz, The ORF7b protein of severe acute respiratory syndrome coronavirus (SARS-CoV) is expressed in virus-infected cells and incorporated into SARS-CoV particles, J. Virol. 81 (2007) 718–731. https://doi.org/10.1128/JVI.01691-06.
[27] D. Benvenuto, S. Angeletti, M. Giovanetti, M. Bianchi, S. Pascarella, R. Cauda, M. Ciccozzi, A. Cassone, Evolutionary analysis of SARS-CoV-2: how mutation of Non-Structural Protein 6 (NSP6) could affect viral autophagy, J. Infect. 81 (2020) e24–e27. https://doi.org/10.1016/j.jinf.2020.03.058.
[28] S. Angeletti, D. Benvenuto, M. Bianchi, M. Giovanetti, S. Pascarella, M. Ciccozzi, COVID-2019: The role of the nsp2 and nsp3 in its pathogenesis, J. Med. Virol. 92 (2020) 584–588. https://doi.org/10.1002/jmv.25719.
[29] M.C. Hagemeijer, I. Monastyrska, J. Griffith, P. van der Sluijs, J. Voortman, P.M. van Bergen en Henegouwen, A.M. Vonk, P.J.M. Rottier, F. Reggiori, C.A.M. de Haan, Membrane rearrangements mediated by coronavirus nonstructural proteins 3 and 4, Virology. 458–459 (2014) 125–135. https://doi.org/10.1016/j.virol.2014.04.027.
[30] B. Short, A call for oxygen in the ER, The Journal of Cell Biology. 203 (2013) 552–552. https://doi.org/10.1083/jcb.2034iti3.
[31] C. Selvaraj, D.C. Dinesh, U. Panwar, R. Abhirami, E. Boura, S.K. Singh, Structure-based virtual screening and molecular dynamics simulation of SARS-CoV-2 Guanine-N7 methyltransferase (nsp14) for identifying antiviral inhibitors against COVID-19, J. Biomol. Struct. Dyn. (2020) 1–12. https://doi.org/10.1080/07391102.2020.1778535.
[32] S. Chen, X. Zheng, J. Zhu, R. Ding, Y. Jin, W. Zhang, H. Yang, Y. Zheng, X. Li, G. Duan, Extended ORF8 Gene Region Is Valuable in the Epidemiological Investigation of Severe Acute Respiratory Syndrome-Similar Coronavirus, J. Infect. Dis. 222 (2020) 223–233. https://doi.org/10.1093/infdis/jiaa278.
[33] E. Issa, G. Merhi, B. Panossian, T. Salloum, S. Tokajian, SARS-CoV-2 and ORF3a: nonsynonymous mutations, functional domains, and viral pathogenesis, MSystems. 5 (2020) e00266-20. https://doi.org/10.1128/mSystems.00266-20.
[34] M. Bolles, E. Donaldson, R. Baric, SARS-CoV and emergent coronaviruses: viral determinants of interspecies transmission, Current Opinion in Virology. 1 (2011) 624–634. https://doi.org/10.1016/j.coviro.2011.10.012.
[35] V.M. Corman, H.J. Baldwin, A.F. Tateno, R.M. Zerbinati, A. Annan, M. Owusu, E.E. Nkrumah, G.D. Maganga, S. Oppong, Y. Adu-Sarkodie, P. Vallo, L.V.R.F. da Silva Filho, E.M. Leroy, V. Thiel, L. van der Hoek, L.L.M. Poon, M. Tschapka, C. Drosten, J.F. Drexler, Evidence for an

ancestral association of human coronavirus 229E with bats, J. Virol. 89 (2015) 11858–11870. https://doi.org/10.1128/JVI.01755-15.
[36] M. Thoms, R. Buschauer, M. Ameismeier, L. Koepke, T. Denk, M. Hirschenberger, H. Kratzat, M. Hayn, T. Mackens-Kiani, J. Cheng, C.M. Stürzel, T. Fröhlich, O. Berninghausen, T. Becker, F. Kirchhoff, K.M.J. Sparrer, R. Beckmann, Structural basis for translational shutdown and immune evasion by the Nsp1 protein of SARS-CoV-2, Molecular Biology, 2020. https://doi.org/10.1101/2020.05.18.102467.
[37] S. Laha, J. Chakraborty, S. Das, S.K. Manna, S. Biswas, R. Chatterjee, Characterizations of SARS-CoV-2 mutational profile, spike protein stability and viral transmission, Infect. Genet. Evol. 85 (2020) 104445. https://doi.org/10.1016/j.meegid.2020.104445.
[38] C. Iserman, C. Roden, M. Boerneke, R. Sealfon, G. McLaughlin, I. Jungreis, C. Park, A. Boppana, E. Fritch, Y.J. Hou, C. Theesfeld, O.G. Troyanskaya, R.S. Baric, T.P. Sheahan, K. Weeks, A.S. Gladfelter, Specific viral RNA drives the SARS CoV-2 nucleocapsid to phase separate, Biochemistry, 2020. https://doi.org/10.1101/2020.06.11.147199.
[39] Y. Cong, M. Ulasli, H. Schepers, M. Mauthe, P. V'kovski, F. Kriegenburg, V. Thiel, C.A.M. de Haan, F. Reggiori, Nucleocapsid protein recruitment to replication-transcription complexes plays a crucial role in coronaviral life cycle, J Virol. 94 (2019) e01925-19, /jvi/94/4/JVI.01925-19.atom. https://doi.org/10.1128/JVI.01925-19.
[40] O.M. Ugurel, O. Ata, D. Turgut-Balik, An updated analysis of variations in SARS-CoV-2 genome, Turk. J. Biol. 44 (2020) 157–167. https://doi.org/10.3906/biy-2005-111.
[41] Z. Daniloski, X. Guo, N.E. Sanjana, The D614G mutation in SARS-CoV-2 Spike increases transduction of multiple human cell types, 2020. https://doi.org/10.1101/2020.06.14.151357.
[42] R. Lorenzo-Redondo, H.H. Nam, S.C. Roberts, L.M. Simons, L.J. Jennings, C. Qi, C.J. Achenbach, A.R. Hauser, M.G. Ison, J.F. Hultquist, E.A. Ozer, A unique clade of SARS-CoV-2 viruses is associated with lower viral loads in patient upper airways, 2020. https://doi.org/10.1101/2020.05.19.20107144.
[43] C.C. Posthuma, A.J.W. Te Velthuis, E.J. Snijder, Nidovirus RNA polymerases: Complex enzymes handling exceptional RNA genomes, Virus Res. 234 (2017) 58–73. https://doi.org/10.1016/j.virusres.2017.01.023.
[44] S.-Y. Fung, K.-S. Yuen, Z.-W. Ye, C.-P. Chan, D.-Y. Jin, A tug-of-war between severe acute respiratory syndrome coronavirus 2 and host antiviral defence: lessons from other pathogenic viruses, Emerg Microbes Infect. 9 (2020) 558–570. https://doi.org/10.1080/22221751.2020.1736644.
[45] K.A. Ivanov, J. Ziebuhr, Human coronavirus 229E nonstructural protein 13: characterization of duplex-unwinding, nucleoside triphosphatase, and RNA 5'-triphosphatase activities, J. Virol. 78 (2004) 7833–7838. https://doi.org/10.1128/JVI.78.14.7833-7838.2004.
[46] M. Letko, S.N. Seifert, K.J. Olival, R.K. Plowright, V.J. Munster, Bat-borne virus diversity, spillover and emergence, Nat. Rev. Microbiol. 18 (2020) 461–471. https://doi.org/10.1038/s41579-020-0394-z.
[47] S. Elbe, G. Buckland-Merrett, Data, disease and diplomacy: GISAID's innovative contribution to global health: Data, Disease and Diplomacy, Global Challenges. 1 (2017) 33–46. https://doi.org/10.1002/gch2.1018.
[48] J. Hadfield, C. Megill, S.M. Bell, J. Huddleston, B. Potter, C. Callender, P. Sagulenko, T. Bedford, R.A. Neher, Nextstrain: real-time tracking of pathogen evolution, Bioinformatics. 34 (2018) 4121–4123. https://doi.org/10.1093/bioinformatics/bty407.
[49] R.I. Amann, S. Baichoo, B.J. Blencowe, P. Bork, M. Borodovsky, C. Brooksbank, P.S.G. Chain, R.R. Colwell, D.G. Daffonchio, A. Danchin, V. de Lorenzo, P.C. Dorrestein, R.D. Finn, C.M. Fraser, J.A. Gilbert, S.J. Hallam, P. Hugenholtz, J.P.A. Ioannidis, J.K. Jansson, J.F. Kim, H.-P. Klenk, M.G. Klotz, R. Knight, K.T. Konstantinidis, N.C. Kyrpides, C.E. Mason, A.C. McHardy, F. Meyer, C.A. Ouzounis, A.A.N. Patrinos, M. Podar, K.S. Pollard, J. Ravel, A.R. Muñoz, R.J. Roberts, R. Rosselló-Móra, S.-A. Sansone, P.D. Schloss, L.M. Schriml, J.C. Setubal, R. Sorek, R.L. Stevens, J.M. Tiedje, A. Turjanski, G.W. Tyson, D.W. Ussery, G.M. Weinstock, O. White, W.B. Whitman, I. Xenarios, Toward unrestricted use of public genomic data, Science. 363 (2019) 350–352. https://doi.org/10.1126/science.aaw1280.
[50] I. Karsch-Mizrachi, T. Takagi, G. Cochrane, International Nucleotide Sequence Database Collaboration, The international nucleotide sequence database collaboration, Nucleic Acids Res. 46 (2018) D48–D51. https://doi.org/10.1093/nar/gkx1097.

[51] K.D. Yamada, K. Tomii, K. Katoh, Application of the MAFFT sequence alignment program to large data-reexamination of the usefulness of chained guide trees, Bioinformatics. 32 (2016) 3246–3251. https://doi.org/10.1093/bioinformatics/btw412.
[52] B.Q. Minh, H.A. Schmidt, O. Chernomor, D. Schrempf, M.D. Woodhams, A. von Haeseler, R. Lanfear, IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era, Mol. Biol. Evol. 37 (2020) 1530–1534. https://doi.org/10.1093/molbev/msaa015.
[53] P. Sagulenko, V. Puller, R.A. Neher, TreeTime: Maximum-likelihood phylodynamic analysis, Virus Evol. 4 (2018) vex042. https://doi.org/10.1093/ve/vex042.
[54] Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees., Molecular Biology and Evolution. (1993). https://doi.org/10.1093/oxfordjournals.molbev.a040023.
[55] H.J. Muller, Some genetic aspects of sex, The American Naturalist. 66 (1932) 118–138. https://doi.org/10.1086/280418.
[56] K.A. Smith, Louis Pasteur, the father of immunology?, Front Immunol. 3 (2012) 68. https://doi.org/10.3389/fimmu.2012.00068.
[57] A. Danchin, K. Timmis, SARS-CoV-2 variants: Relevance for symptom granularity, epidemiology, immunity (herd, vaccines), virus origin and containment?, Environ. Microbiol. 22 (2020) 2001–2006. https://doi.org/10.1111/1462-2920.15053.