

**研究生学位论文中期报告**

**报告题目 东非地区蝙蝠携带的脊椎动物病毒多样性的生**

**态学调查**

**学生姓名 杨凯心 学号 2020E8009588009**

**指导教师 王达希 职称 研究员**

**学位类别 专业型硕士**

**学科专业 生物与医药**

**研究方向 生态病毒学**

**研究所（院系） 生命科学学院（华大专项）**

**填表日期 2022年8月31日**

**中国科学院大学制**

目 录

[第一章 选题的背景及意义 2](#_Toc114529965)

[1.1 选题背景 2](#_Toc114529966)

[1.2 选题意义 2](#_Toc114529967)

[第二章 国内外本学科领域的发展现状与趋势 3](#_Toc114529968)

[2.1 蝙蝠携带的脊椎动物病毒检测 3](#_Toc114529969)

[2.2 生态病毒学的研究范畴和研究范式 3](#_Toc114529970)

[第三章 课题主要研究内容 6](#_Toc114529971)

[3.1 科学问题和研究对象 6](#_Toc114529972)

[3.2 研究方法和技术路线 6](#_Toc114529973)

[第四章 学位论文进展情况，存在的问题，已取得阶段性结果 12](#_Toc114529974)

[4.1 学位论文进展情况及存在的问题 12](#_Toc114529975)

[4.2 已取得的阶段性成果 12](#_Toc114529976)

[第五章 研究工作计划与进度安排 16](#_Toc114529977)

[参考文献 17](#_Toc114529978)

第一章 选题的背景及意义

1.1 选题背景

当前非洲正面临着由人畜共患病病毒引起的疾病暴发的风险，根据世界卫生组织的统计，与2001年至2011相比，2012年至2022年间，非洲人畜共患病暴发的数量增加了63%。其中由蝙蝠携带的脊椎动物病毒占这些暴发的70%以上，对人类生命健康造成了严重威胁。非洲四个大型蝙蝠种群中的三个都在东非地区(Bruce D Patterson, 2012)，而较大的蝙蝠种群所携带的病毒更有可能呈现出较高的多样性(Mollentze and Streicker, 2020)，但遗憾的是至今仍没有对东非地区蝙蝠携带的脊椎动物病毒多样性的大规模调查(Waruhiu et al., 2017)。

这个研究空缺很大程度上受制于病毒检测技术和采样偏好性的制约，传统的核酸或血清学检测技术十分依赖对病毒的先验了解，从而很难对未知病毒进行检测，进而很难对病毒多样性进行调查。幸运的是，目前科学家利用宏转录组测序等实验技术辅以无参考序列组装、机器学习识别病毒序列等生物信息学技术，已经在一定程度上克服了传统技术的不足，使得宏转录组技术已经发展为研究病毒多样性的主力工具(Zhang, 2018)。

传统的病毒学是将病毒作为致病病原来研究其分布、传播，所以采集的样本大部分来自于有明显疾病症状的样本。但宏转录组研究表明，可以导致明显疾病症状的病毒实际仅占全部病毒的很小一部分(Shi et al., 2016)，这说明研究重点需要从原来将病毒视为寄生病原转变为将病毒视为生态系统的自然组分，从而真正理解病毒的自然分布规律。学界在意识到这种观点转变的重要性之后发展出了生态病毒学，是指利用大规模无偏好性的采样以及高深度的宏转录组测序技术检测样本中的病毒群落组成，并分析病毒多样性与生态环境之间的关系，从而尽可能地还原病毒在真实生态环境中的生存状态，最终实现通过生态数据来预测病毒在空间分布上的多样性，从而提前预防病毒大流行(French and Holmes, 2020)。

1.2 选题意义

本课题填补了东非地区病毒多样性研究的空缺，同时也从生态病毒学的角度出发，分析了环境、宿主个体、宿主群体三个角度，多个维度的生态因素对病毒多样性的影响，为预防东非地区脊椎病毒大流行做出了贡献。

第二章 国内外本学科领域的发展现状与趋势

2.1 蝙蝠携带的脊椎动物病毒检测

蝙蝠是继啮齿类动物之后地球上多样性最丰富的第二个哺乳动物目，其物种总数约占所有哺乳动物的22%，研究表明，物种多样性越丰富的哺乳动物目，越有可能携带多样性更丰富的病毒(Mollentze and Streicker, 2020) ，过去50年以来，许多危害人类生命健康的病毒被证明与蝙蝠有关，例如埃博拉病毒、马尔堡病毒、尼帕病毒、亨德拉病毒、SARS冠状病毒、MERS冠状病毒、以及新冠病毒等(Letko et al., 2020)。

通常人类与蝙蝠的直接接触较少，但蝙蝠有可能与人类共享食物和水源，从而导致其排泄物中携带的病毒污染食物和水源，进而感染人类。研究表明，蝙蝠粪便或肛拭子样本所携带的病毒粒子数量和种类高于其它类型的样本(Lane et al., 2022)，且蝙蝠粪便具有容易采集，不对动物造成创伤等特点，所以我们以蝙蝠粪便及肛拭子样本作为了解蝙蝠携带病毒的指示剂。

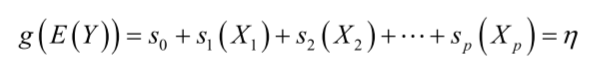
2.2 生态病毒学的研究范畴和研究范式

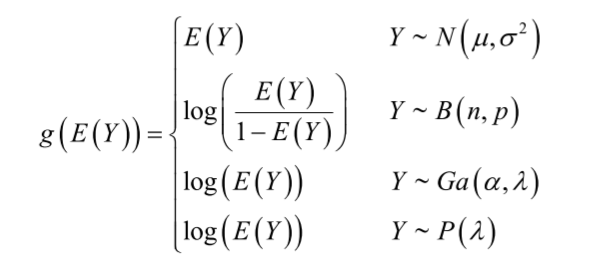
传统病毒学研究的一个核心观点是病毒通常是导致具有潜在破坏性后果的疾病的病原体。 这种观点是可以理解的，因为病毒性疾病给许多宿主物种带来了重大的疾病负担，特别是在跨物种传播之后，并且有大量的宿主基因致力于抵抗病毒感染(Enard et al., 2016)。 然而近期对不同宿主物种的宏基因组研究使学界对于病毒圈产生了新的理解，自然界中病毒的数量是非常庞大的(Geoghegan and Holmes, 2017)，正常情况下宿主会被多种病毒感染，即使在病毒丰度很高的情况下也很少表现出明显与疾病相关的症状(Zhang, 2018 , Shi et al., 2016)。 甚至在一些情况下病毒对宿主有益(Roossinck, 2015)。种种证据挑战着传统病毒学的核心观点，提示学界应该将病毒群落作为生态系统的天然组分来研究(French and Holmes, 2020)。

但事实上对病毒群落的研究颇具挑战性，特别是由于检测能力的限制导致通常只能对单个病毒物种进行研究。而宏基因组测序技术在一定程度上改善了这一现状，这种技术可以对样本中包含的整个病毒群落进行检测，从而为在群落层面上的病毒研究提供了可能性(Zhang, 2018)。然而目前大部分工作都是在利用宏基因组技术发现新病毒从而拓展病毒圈结构，较少有研究从将病毒视为生态系统的自然组成部分来讨论病毒群落-生态系统相互作用模式。近年来，通过大规模测序积累的数据，我们已经对常规环境中的病毒群落有了大致的了解，所以是时候更进一步来研究病毒群落与生态系统相互作用模式了(French and Holmes, 2020)。

对于病毒群落与生态系统的相互作用的研究直接引领了一个新学科的诞生——生态病毒学，该学科的主要目标在于回答三个科学问题：非生物的生态环境因素（如海拔、降水量、平均气温等）以及宿主性状因素（如宿主行为、饮食、营养水平、种类、种间相互作用以及种群密度等）如何影响病毒多样性和丰度，病毒多样性又如何反作用于这些生态因素以及病毒群落中的不同病毒物种之间是否存在一定程度上的相互作用以及相互作用模式。其中第一个问题最容易回答且最受到广泛关注，且已有许多证据表明生态环境和宿主性状能够影响到一些病毒的传播频率，从而影响到其分布(French and Holmes, 2020)。

在统计分析中，多变量线性回归模型是预测问题中最常用的工具，但它要求反应变量的期望与每个预测变量的关系都是线性的，如果这一假设不成立，就需推广到加性模型. 加性模型对预测变量的具体分布形式没有要求，而是用非参数的方法来拟合。广义线性模型也是多变量线性回归模型的扩展，它建立在 反应变量期望的某个函数和预测变量的线性关系基础之上。而广义加性模型(generalized additive model, GAM)是广义线性模型和加性模型的结合形式，是用一个连接函数来建立相应变量的期望与非参数形式的各解释变量之间的关系的(贾彬, 2005)。

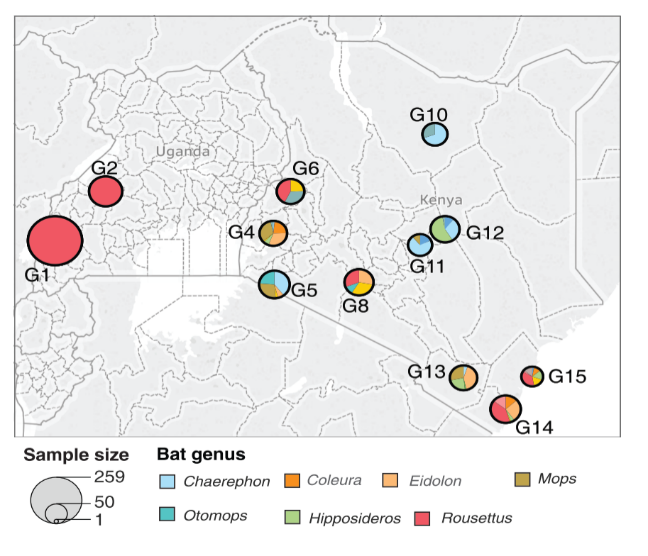
1985年，Stone提出了标准的加性模型，模型中的每一个加性项都使用单独的光滑函数来估计，避免了维数灾难的问题(STONE C. J., 1985)。1986年，Hastie和Tibshirani将加性模型的技术应用于广义线性模型(generalized linear model, GLM)，于是就产生了广义加性模型，广义加性模型是加性模型的推广，本质是利用连接函数把加性模型中响应变量的期望与加性部分联系起来(Trevor Hastie, 1986 , 李晓童, 2016)。广义加性模型的公式如下：

其中是的期望，是截距，是非参数光滑函数，满足，光滑函数是针对每一个解释变量的单变量光滑函数，它可以是样条函数、局部回归光滑函数或核函数等，本文使用的光滑函数是三次样条函数，它不仅具有较好的整体连续可导性，还能很好地适应数据和函数的变化。为连接函数，它是一个单调可微的非线性函数，对于不同分布类型的响应变量，连接函数的形式也不同，其具体对应关系如下(李晓童, 2016)：

GAM有许多优点，它能直接处理响应变量和多个解释变量之间的非线性关系，同时对于不同解释变量可以使用不同光滑函数。近年来，GAM凭借其优秀的可解释性以及灵活性受到了越来越多生态学家的青睐(Olival et al., 2017 , Eric J. Pedersen, 2019)。但在拟合GAM的实际数据中，共曲线性是一个普遍而不可忽视的问题。它类似于多变量线性回归模型中的共线性，是共线性在GAM中的非参数形式的体现。共曲线性的存在会低估模型参数的标准误，增大一类错误，同时会导致模型的解不惟一，使可GAM的拟合过程收敛的解受起始函数的影响，所以消除解释变量之间的共线性是在进行GAM建模过程中的难点之一(贾彬, 2005)。

第三章 课题主要研究内容

3.1 科学问题和研究对象

我们在2014-2019年间从东非乌干达和肯尼亚两个国家的12个区域收集了来自7个蝙蝠物种的637个蝙蝠粪便样本，首先利用高深度宏基因组技术从样本中挖掘脊椎动物病毒以及对应的丰度信息。然后进行生态病毒学分析，我们将所有病毒分为DNA病毒和RNA病毒两大类，分别统计了DNA、RNA病毒的物种数量，然后根据采样地点GPS、物种信息找到对应的到生态环境、宿主物种性状数据并利用广义加性模型分别对检测到的DNA、RNA病毒物种数量进行建模，最后分析模型来探究生态环境、宿主物种性状分别对DNA、RNA病毒多样性的影响。本课题拟回答的科学问题有两个，其一是东非乌干达和肯尼亚两国蝙蝠携带了哪些脊椎动物病毒，以及对应的地理、宿主物种分布情况如何。其二是生态因素与病毒群落多样性之间的关联关系。

由于两个科学问题的属性不同，对应的研究对象也不同，其一是东非乌干达和肯尼亚两国蝙蝠携带的脊椎动物病毒多样性，包括病毒的物种多样性和丰度。其二是非生物的生态环境因素（如海拔、降水量、平均气温等）以及宿主生态学因素（如宿主行为、饮食、营养水平、种类、种间相互作用以及种群密度等）如何影响东非乌干达和肯尼亚两国蝙蝠携带的脊椎动物病毒多样性。

3.2 研究方法和技术路线

根据研究对象以及研究对象之间的逻辑关系，整个研究过程可以分为两个部分——病毒多样性调查和生态病毒学分析。第一个部分采用抽样调查的方法采集蝙蝠粪便样本，利用宏转录组测序技术检测其中脊椎动物病毒群落组成来代表以样本为圆心，半径10公里区域内脊椎动物病毒的群落组成，从而回答病毒如何分布的问题。第二部分主要采用数学建模的方法，利用机器学习模型拟合生态环境因素（如海拔、降水量、平均气温等）以及宿主生态学因素（如宿主行为、饮食、营养水平、种类、种间相互作用以及种群密度等）与不同类别脊椎动物病毒丰度之间的关系。同时利用假设检验判断不同生态因素对不同蝙蝠属样本中检出的病毒多样性之间的显著性关系，从而回答哪些生态因素如何影响病毒多样性的问题。

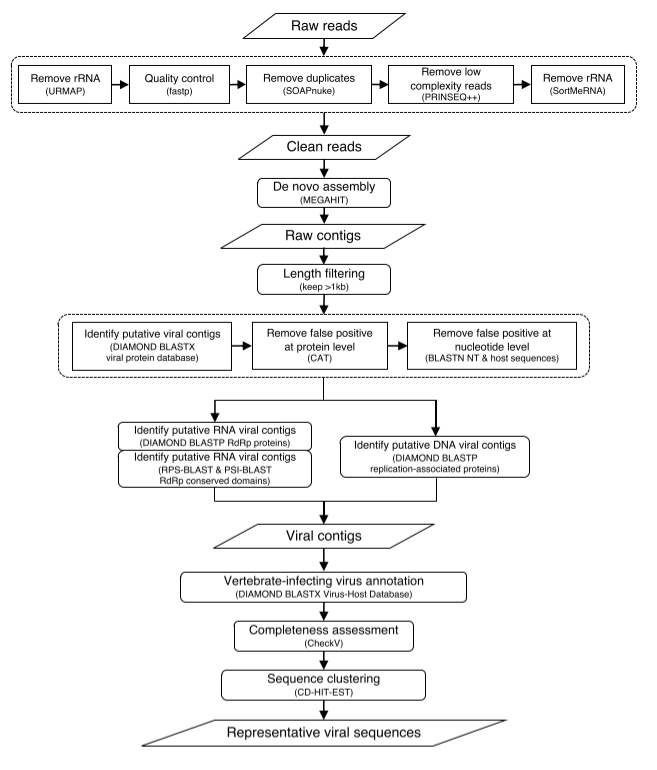
**病毒多样性调查**

蝙蝠粪便样本采集自蝙蝠栖息的洞穴、树林、废弃的建筑物等环境，为了收集新鲜粪便样本，我们将2m×2m的聚乙烯板放在蝙蝠巢穴底部，过夜后收集板上的蝙蝠粪便颗粒并浸泡在RNA稳定剂中(QIAGEN, Germany)，保存在-80℃冰箱里。之后利用QIAamp Viral RNA Mini Kit (QIAGEN, Hildon, Germany) 提取 RNA并使用 MGIEasy RNA Library Prep Kit V3.0 制备测序双端150bp的文库进行测序。

我们利用URMAP (v1.0.1480) 对每个样本的测序数据进行快速比对，从而去除其中大部分比对上rRNA的reads，随后利用fastp (v0.20.1) 去掉测序接头和低质量的reads，接着使用SOAPnuke (v2.1.5)、PRINSEQ++ (v1.2) 和SortMeRNA (v4.3.2)去除了重复、低复杂性和剩余比对上 rRNA 的reads从而得到clean reads。

得到 clean reads之后我们用MEGAHIT (v1.2.9) 进行denovo组装得到组装contigs序列，将大于1000bp的contigs用 DIAMOND (v0.9.36.137) 的BLASTX 模式，设置较为严格的E-value (< 10e-5) 与NCBI非冗余蛋白数据库 (NR) 、NCBI RefSeq 和 IMG-VR 数据库的病毒蛋白质进行比较，挑选其中比对到病毒蛋白的contigs。为了去除非病毒contigs，我们用CAT (v5.2) 参考NR数据库对上一步选出的contigs进行注释，将注释为病毒界的contigs与NCBI 非冗余核酸 数据库(NT)和翼手目的参考基因组进行了进一步比较，从而去除宿主序列。

为了筛选宿主为脊椎动物的病毒contigs，我们从病毒-宿主关联数据库(Virus-Host Database)中找到每条contigs的最近源病毒的宿主，作为该contigs能够感染的潜在宿主。之后我们用CheckV对病毒contigs完整性进行评估，并用CD-HIT对病毒contigs进行属（序列相似性大于80%）和物种（序列相似性大于95%）层面的聚类，挑选出物种的代表序列（每个物种聚类簇中最长的那条contig）作为检出的病毒序列。

为了测定检出病毒的丰度，我们用Bowtie2 (v2.4.2)将每个样本的clean reads比对到所有检出的病毒序列上，保留覆盖率大于40%，覆盖长度大于1000bp以及错赔率小于6%的病毒序列计算其相对丰度，保留相对丰度大于0.0001%的病毒丰度作为每个样本中检测到的对应病毒的丰度。

**生态数据的获取**

生态数据按照获取方式分为两类：从公共数据库获取或从现场采样测序数据中获取。

从公共数据库中获取的数据根据​​查询关键词可以分为两类：一类是根据采样位置的GPS坐标查询的环境特征，包括海拔高度(Jeffrey W. Hollister, 2022)，哺乳动物数量(IUCN, 2015)、附近栅格内的人口密度、耕地面积、放牧面积(Klein Goldewijk, 2017)；另一类是根据样本对应的宿主物种查询的宿主特征，包括个体宿主特征（体重、前臂长度、窝产仔数、摄食习性等）和宿主群体特征（群体规模大小、全球栖息地总面积等）(Kate E. Jones, 2009)。

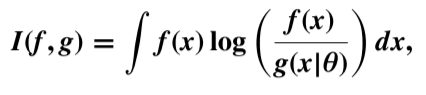
通过现场采样和测序获得的数据有4个字段。野外采样信息包括在每个样本所在的采样点周围10km范围内采集到的其它蝙蝠物种数量、从周围10km范围内其它样本中检测出的DNA/RNA病毒数量、每个样本的测序数据碱基总数以及组装结果的总长度。

**广义加性模型的建模**

我们利用广义加性模型来分析各类生态因素（以下称为解释变量）对样本中检测到的DNA、RNA以及所有病毒数量（以下称为响应变量）的影响。为了避免解释变量之间的多重共线性问题，我们将测量相似因素的解释变量分成了9个组，分别是：宿主性状（包括重量、臂长、翼荷载、窝产仔数、年产窝数）；宿主食性（包括食谱广度、营养等级）；宿主群体（种群大小、全球栖息地总面积）；人类密度（包括附近栅格内的人口总数、城市人口总数、农村人口总数以及采样点距离最近的人口聚居地的距离）；人类土地利用（种植、畜牧面积）；环境生态因素（海拔）；生物生态因素（同域哺乳动物数量，是指同时出现在同一地点哺乳动物物种数量）；测序数据质量（clean碱基数、Q30、组装contigs长度）；样本背景信息（样本附近采集到的其它样本数量、附近采集到的其它蝙蝠属数量、附近样本中检测到的平均DNA，RNA病毒数量）。为了检查宿主物种所在的属对响应变量的影响，基于样本之间属的分类互斥性，我们将属考虑为虚拟变量并用0-1矩阵代替分类，当虚拟解释变量的取值为0时，其系数对相应变量没有影响，而当虚拟解释变量的取值为1时，其系数会改变模型截距。我们通过遍历每组连续型解释变量的组合对应三种响应变量（DNA病毒、RNA病毒、所有病毒）生成了5760个模型公式，然后利用R包mgcv中的gam函数对所有模型进行拟合(Simon N Wood, 2011)。

为了保证模型整体的光滑性和较小的计算所需的时间开销，我们将连续型解释变量的样条函数设置为三次样条函数并设置了较小的节点数量（k=3）。同时，我们利用双重光滑收缩技术来消除模型中几乎没有预测能力的连续型解释变量，这种技术可以确定哪些连续型解释变量对响应变量的影响最大，从统计学的角度来看，它代表了一种在拟合效果和模型简约性之间取得平衡的方法。通过双重光滑收缩技术有效识别重要的连续型解释变量的子集，既可以增强模型的可解释性，又可以提高模型预测精度(Giampiero Marra, 2011)。

**广义加性模型的筛选和检验**

模型筛选是指在给定的一组拟合同一响应变量的模型集合中选择更符合现象背后过程的一组模型子集，优秀的模型通常具有较好的拟合优度(goodness of fit)和简单性(Shirangi, 2016)，统计模型的拟合优度描述了模型预测值与一组观察值的拟合程度。拟合优度度量观察值与相关模型下预期值之间的差异。此类差异可用于统计假设检验，例如检验残差的正态性，检验两个样本是否来自相同分布，或者检验预测结果频率是否遵循特定分布等。简单性是指参数的个数，参数越少的模型越简单，通常来讲，模型参数越多，越复杂，拟合优度越好，但模型可能存在过拟合问题，反之模型可能存在欠拟合问题(Kenneth P. Burnham, 2002)。最简单的模型筛选准则是奥卡姆剃刀准则(Barry C. M., 2014)，该准则直接选择模型集合中参数个数最少的模型，但未考虑拟合优度，所以不适用于过程较为复杂的生态学模型。1974年日本统计学家赤池弘次基于Fisher最大化对数似然推导出了Kullback-Leibler散度的相对期望估计量，称为Akaike信息准则 (AIC)(H. Akaike, 1974)，Kullback-Leibler散度是一个信息论领域的术语，描述的是两个分布之间的统计距离(Kenneth P. Burnham, 2002)，表示用近似时丢失的信息，也可以通俗的理解为是从模型到模型的距离。

但是如果完全不了解每个候选模型中的每个参数以及最优模型就无法计算K-L散度，幸运的是，Akaike找到了一种基于模型最大似然对数函数来严格估计K-L散度的方法，并且这种方法非常简单，公式中为模型中参数的个数，为模型的似然函数的最大值，系数设为-2是由于某些“历史遗留原因”(Kenneth P. Burnham, 2002)。

计算备选模型集合中每个模型对应的AIC之后我们按照AIC从小到大的顺序对集合进行排序，并依据经验值选择其中小于3的模型进行后续分析(Kenneth P. Burnham, 2002)。

为了确保模型没有发生过拟合，我们使用十折交叉验证来评估所有模型的拟合优度。将数据集平均分成十份，随机选取其中的一折作为验证集，根据其他九折数据重新拟合模型，通过进行非参数置换检验来评估拟合优度，检验一折验证数据集预测值与实际值的差异，若差异不显著则说明模型没有过拟合。

第四章 学位论文进展情况，存在的问题，已取得阶段性结果

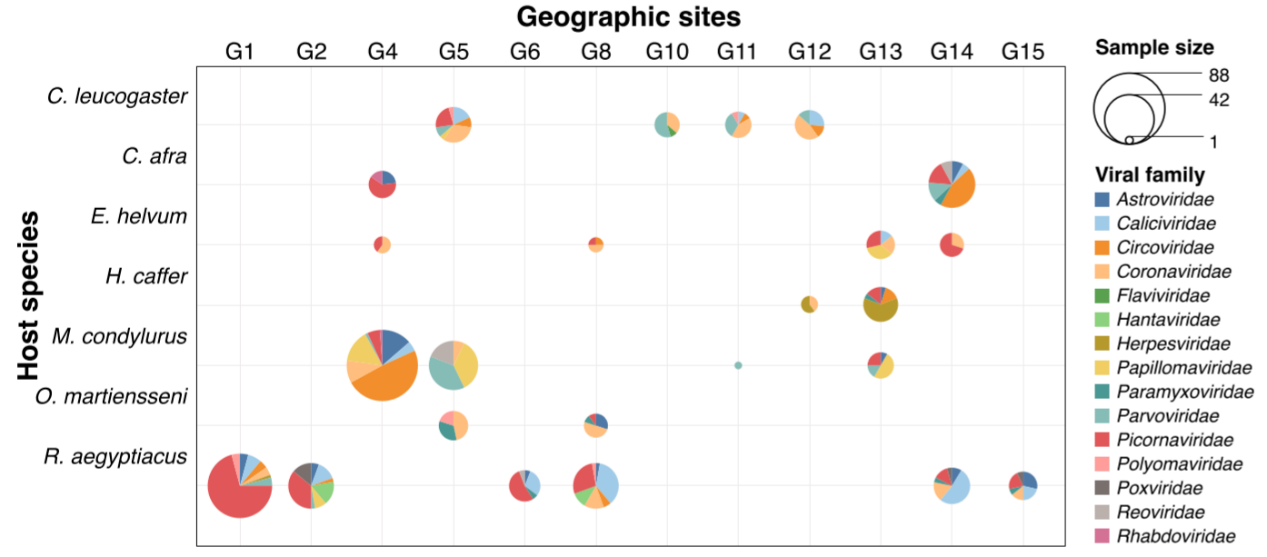
4.1 学位论文进展情况及存在的问题

所有分析任务基本已经完成，正在总结梳理结论，目前研究背景部分已经完成，尚未开始撰写结果部分。

存在的问题主要有三点，其一是没有充分利用到病毒基因组数据，没有对不同物种、采样地域之间的病毒基因组差异做详细分析；其二是RNA病毒的GAM模型仅能解释21.8%的总变异，说明可能仍未找到对于RNA病毒检出数量具有预测能力的解释变量；其三是生态建模使用的模型筛选准则为AIC准则，近期有学者指出使用该准则可能无法筛选到能揭示因果关系的因素，如需要揭示因果关系，则需使用backdoor准则(Arif and MacNeil, 2022)。

4.2 已取得的阶段性成果

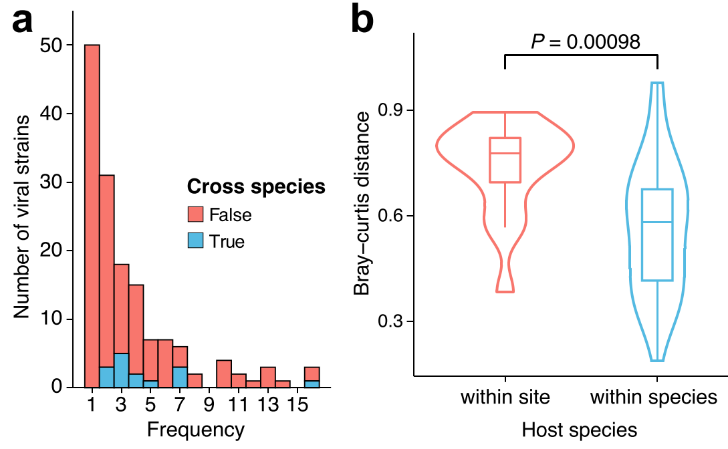
**病毒分布**

通过对637个蝙蝠粪便和肛拭子样本的宏转录组测序，我们获得了17.96Tb个碱基，过滤核糖体RNA后每个样本平均剩余7.9G个碱基，通过组装、比对等技术，我们在287个样本中鉴定到了包括腺病毒科(n=2)、星状病毒科(n=31)、杯状病毒科(n=36)、圆环病毒科(n=21)、冠状病毒科(n=18)、黄病毒科(n=1)、汉坦病毒科(n=1)、疱疹病毒科 (n=3)、乳头瘤病毒科(n=21)、副粘病毒科(n=6)、细小病毒科(n=20)、小核糖核酸病毒科(n=41)、多瘤病毒科(n=7)、痘病毒科(n=1)、呼肠孤病毒科(n=7)和弹状病毒科(n=1)，共计16个病毒科的217个病毒物种。

了解病毒在宿主物种和地理区域之间的分布情况是病原监测的重要步骤，我们首先将所有样本按照不同蝙蝠物种进行分类，我们发现不同蝙蝠物种来源的样本检出病毒阳性率不同，均值为50%，极差为45%。同时不同蝙蝠物种来源的阳性样本中平均检出的病毒数量也不同，均值为1.65，极差为2.06。

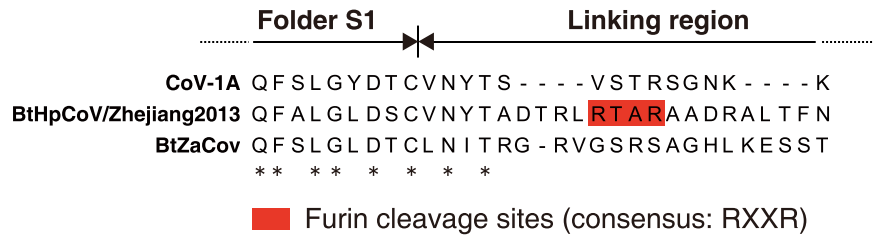
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **蝙蝠物种** | **样本数量** | **检出病毒数量** | **阳性率** | **阳性样本中平均检出病毒数量** |
| *Chaerephon leucogaster* | 111 | 23 | 34.23% | 1.58 ± 0.95 |
| *Coleura afra* | 26 | 23 | 73.08% | **2.68 ± 2.40** |
| *Eidolon helvum* | 82 | 11 | 31.71% | 1.27 ± 0.96 |
| *Hipposideros caffer* | 60 | 8 | 31.67% | 1.37 ± 0.60 |
| *Mops condylurus* | 62 | 42 | 69.35% | **3.33 ± 1.95** |
| *Otomops martiensseni* | 25 | 7 | 76.00% | 1.32 ± 0.48 |
| *Rousettus aegyptiacus* | 271 | 52 | 45.39% | 1.61 ± 0.91 |

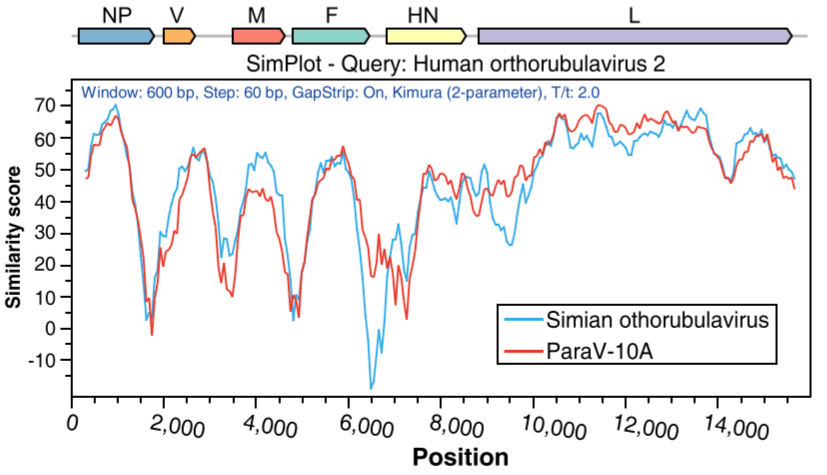
**Table 1. 不同蝙蝠物种样本中检出的病毒情况**

我们发现，在超过三个样本中检测到的病毒株占所有病毒记录的69.0%，说明了病毒群落在样本间的分布是不均匀的。在287份（45.1%）病毒阳性样本中，123份样本携带不止一种病毒株，其中大部分（83.7%）携带了不同科的病毒株。然而对于病毒来说，只有15株病毒出现了跨宿主物种感染的情况，说明病毒具有较高的宿主特异性。

此外，我们比较了同一宿主物种和属于同一地理区域的样本中检出病毒科的组内Bray−curtis距离，发现同一宿主物种显著高于同一地理区域的样本中检出病毒科的组内Bray−curtis，说明宿主的生物学差异对病毒组成的影响大于地理距离。

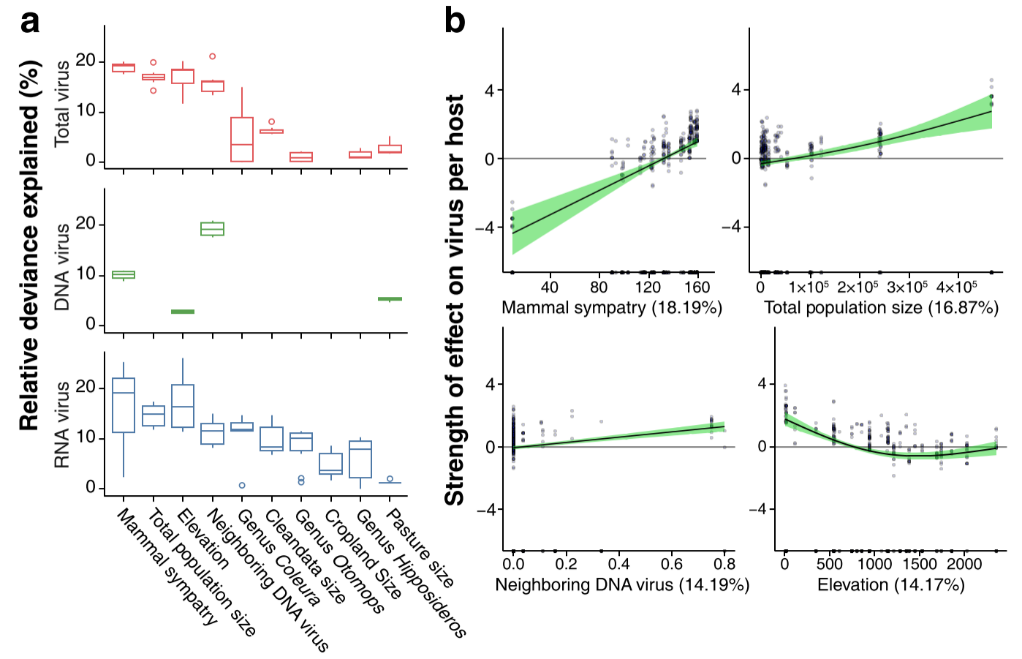
**重要病毒发现**

对于冠状病毒科，我们从蹄蝠中鉴定出一种新的冠状病毒，根据保守蛋白序列比对分类到Coronaviridae，Betacoronavirus属的一个亚属Hibecovirus，我们根据基因组聚类结果将其命名为CoV-1A，与同一亚属的Hp-betacoronavirus Zhejiang2013的氨基酸一致性(RAP-AAI)为87.8%，与尼日利亚Hipposideros commersoni样本中检测到的BtZaCoV的氨基酸一致性为88.4%，与中国蹄蝠中检测到的SARS-CoV-2的氨基酸一致性为74.2%。CoV-1A与Hibecovirus亚属的其他成员一样，在M和N基因之间也包含两个插入的ORF。与Hp-betacoronavirus Zhejiang2013相比，CoV-1A和BtZaCoV的刺突蛋白均不含弗林蛋白酶切割基序，这一证据支持了Hp-betacoronavirus Zhejiang2013在与CoV-1A和BtZaCoV分离后独立变异出了弗林蛋白酶切割位点。

对于副粘病毒科，我们从埃及果蝠样本中检测到了一种与人副流感病毒2近缘(78.8% RAP-AAI)的orthorubulavirus，根据基因组聚类结果将其命名为ParaV-10A，这是在野生动物样本中发现的与人副流感病毒2最近缘的orthorubulavirus。ParaV-10A与人副流感病毒2具有最高相似性的基因组区域主要位于L基因内，说明复制相关基因更为保守。

**生态建模分析**

我们根据样本采样点的GPS坐标和对应的宿主物种标签从公共数据库中查询到了一系列生态因素，然后利用广义加性模型将生态因素作为解释变量，分别构建三个模型用于拟合每个样本中检测到的DNA、RNA、所有病毒株的数量，我们通过解释变量之间的分组排列组合建模来避免解释变量之间的曲线相关，并利用AIC筛选到拟合优度较好的模型集合，最后用交叉验证过滤掉过拟合的模型（详见研究方法部分）。

最终最优模型（AIC最小）分别解释了每个样本中检测到的DNA、RNA、所有病毒株数量总变异的21.8%、48%、32%。通过分析三类模型中解释变量对应的权重、拟合曲线以及偏残差分布，我们发现同域哺乳动物物种数量、采样点附近的人口总数和采样点海拔是每个样本中检测到的所有病毒株数量的最强预测因子。DNA病毒模型中只有同域哺乳动物物种数量、采样点海拔和采样点附近其他样本中检出的DNA病毒株的平均数量是有预测能力的解释变量。

第五章 研究工作计划与进度安排

2022年9月-2022年11月 整理汇总所有分析结果，完成毕业论文大纲

2022年11月-2023年1月 完成毕业论文初稿

2022年1月-2022年2月 完成毕业论文终稿

参考文献

ARIF S, MACNEIL M A 2022. Predictive models aren't for causal inference. Ecol Lett [J], 25: 1741-1745.

BARRY C. M. 2014. Who sharpened Occam's Razor. Irish Philosophy. [J].

BRUCE D PATTERSON P W. Keys to the Bats (Mammalia: Chiroptera) of East Africa[C]//,2012

ENARD D, CAI L, GWENNAP C, et al. 2016. Viruses are a dominant driver of protein adaptation in mammals. Elife [J], 5.

ERIC J. PEDERSEN D L M, GAVIN L. SIMPSON, NOAM ROSS 2019. Hierarchical generalized additive models in ecology: an introduction with mgcv. PeerJ [J].

FRENCH R K, HOLMES E C 2020. An Ecosystems Perspective on Virus Evolution and Emergence. Trends Microbiol [J], 28: 165-175.

GEOGHEGAN J L, HOLMES E C 2017. Predicting virus emergence amid evolutionary noise. Open Biol [J], 7.

GIAMPIERO MARRA S N W 2011. Practical variable selection for generalized additive models. COMPUTATIONAL STATISTICS & DATA ANALYSIS [J].

H. AKAIKE 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control [J], 19.

IUCN C, COLUMBIA UNIVERSITY, NASA 2015. Gridded Species Distribution: Global Mammal Richness Grids, 2015 Release. NASA Socioeconomic Data and Applications Center [J].

JEFFREY W. HOLLISTER 2022. Accessing elevation data in R with the elevatr package. R package [J].

KATE E. JONES J B, MARCEL CARDILLO,SUSANNE A. FRITZ,JUSTIN O'DELL,C. DAVID L. ORME,KAMRAN SAFI,WES SECHREST,ELIZABETH H. BOAKES,CHRIS CARBONE,CHRISTINA CONNOLLY,MICHAEL J. CUTTS,JANINE K. FOSTER,RICHARD GRENYER,MICHAEL HABIB,CHRISTOPHER A. PLASTER,SAMANTHA A. PRICE,ELIZABETH A. RIGBY,JANNA RIST,AMBER TEACHER,OLAF R. P. BININDA-EMONDS,JOHN L. GITTLEMAN,GEORGINA M. MACE,ANDY PURVIS 2009. PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals.

KENNETH P. BURNHAM D R A 2002. Model Selection and Multimodel Inference : A Practical Information-Theoretic Approach. 2nd ed. New York: Springer [J].

KLEIN GOLDEWIJK K, BEUSEN, A., DOELMAN, J., AND STEHFEST, E. 2017. Anthropogenic land use estimates for the Holocene – HYDE 3.2. Earth Syst. Sci [J], Data: 927–953.

LANE J K, NEGASH Y, RANDHAWA N, et al. 2022. Coronavirus and Paramyxovirus Shedding by Bats in a Cave and Buildings in Ethiopia. Ecohealth [J], 19: 216-232.

LETKO M, SEIFERT S N, OLIVAL K J, et al. 2020. Bat-borne virus diversity, spillover and emergence. Nat Rev Microbiol [J], 18: 461-471.

MOLLENTZE N, STREICKER D G 2020. Viral zoonotic risk is homogenous among taxonomic orders of mammalian and avian reservoir hosts. Proc Natl Acad Sci U S A [J], 117: 9423-9430.

OLIVAL K J, HOSSEINI P R, ZAMBRANA-TORRELIO C, et al. 2017. Host and viral traits predict zoonotic spillover from mammals. Nature [J], 546: 646-650.

ROOSSINCK M J 2015. Plants, viruses and the environment: Ecology and mutualism. Virology [J], 479-480: 271-277.

SHI M, LIN X D, TIAN J H, et al. 2016. Redefining the invertebrate RNA virosphere. Nature [J], 540: 539-543.

SHIRANGI M G D, LOUIS J. 2016. A general method to select representative models for decision making and optimization under uncertainty. Computers & Geosciences [J], 96.

SIMON N WOOD 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. Journal of the Royal Statistical Society [J], (B): 3-36.

STONE C. J. 1985. ADDITIVE REGRESSION AND OTHER NONPARAMETRIC MODELS. ANNALS OF STATISTICS [J].

TREVOR HASTIE R T 1986. Generalized Additive Models. Statist. Sci. [J], August 1: 297-310.

WARUHIU C, OMMEH S, OBANDA V, et al. 2017. Molecular detection of viruses in Kenyan bats and discovery of novel astroviruses, caliciviruses and rotaviruses. Virol Sin [J], 32: 101-114.

ZHANG Y S, M; HOLMES, EC 2018. Using Metagenomics to Characterize an Expanding Virosphere.

贾彬 王, 王琳娜, 陈长生 2005. 广义可加模型共曲线性及其在空气污染问题研究中的应用. 第四军医大学学报 [J], 26(3): 280-283.

李晓童 张 2016. 基于广义加性模型的北京市PM2.5浓度影响 因素分析. 数据挖掘 [J].