1. 各位老师下午好，我是精准健康研究所病原组学研究中心的杨凯心，我的课题是：“东非地区蝙蝠携带的脊椎动物病毒多样性的生态学调查”
2. 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。据调查，非洲有四个大型蝙蝠群落，其中三个都在东非地区。有研究表明较大的蝙蝠群落所携带的病毒更有可能呈现出较高的多样性。
3. 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。
4. 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，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。
5. 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。
6. 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。
7. 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。
8. 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。
9. 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。
10. 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。
11. 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。