**Abstract**

Virus Variation：web-based，包括专用数据库 + 搜索界面 + 整套工具

序列下载后可以后续用多种工具/方法进行数据处理和分析

序列的选择基于生物学的多样性和临床诊断的特征

近两年随着更多的病毒信息录入，一种新的多序列查看器被整合，增强了未来对新病毒分析的能力

**Introduction**

元数据丢失、从文献中查找困难、没有为众人所接受的统一标准、缺乏标准拓展、不同来源的信息对相同蛋白质序列注释不一致

数据众多，有传统数据存储模式有爆炸风险

**Virus Variation 2.0**

提高流感病毒数据可用性

发展出了基于元数据的搜索界面

术语存贮在数据库中，为同义词和拼写错误的术语提供单个术语

WNV在北美持续进化

共享后端促进高效数据流

目标：建立具有单一、增值数据模型但足够灵活以适应多种病毒的资源

核心：病毒变异数据库加载管道。负责对输入的核苷酸序列进行标准化注释，自动解析来自GenBank记录的元数据术语，并将解析后的术语映射到受控词汇表。本质为从INSDC记录中提取现有的注释并将其加载到数据库中

流感注释管道（FLAN）：从病毒变异网页和第一类序列通过BLAST比对一组病毒特异性核苷酸参考序列，使用针对每种病毒亚型的参考蛋白序列集注释蛋白编码区。具体来说，FLAN维护了一组参考核苷酸序列，这些参考核苷酸序列用于按类型（A，B或C）对输入的流感序列进行分类，识别特定的片段，并在适用时识别亚型的流感A血凝素和神经氨酸酶片段。然后，将相应的参考蛋白质组与翻译后的输入序列和使用“蛋白质与核苷酸比对工具”（ProSplign）预测的蛋白质编码区进行比对

缺陷：无法适应生物学的不断发展而更新蛋白质注释

新模型：所有功能都将使用改进版NCBI“蛋白质与核苷酸比对工具”（ProSplign 2）由内部管线直接注释。注释也将定期更新，并保持一致的注释，而不考虑序列提交日期或更改的注释标准

**New Features**

搜索界面：多标签关键词搜索。文件多类型下载

针对某些病毒/流感，过滤器对搜索序列长度进行限制。可diy准确性及匹配性

数据集可视化，便于蛋白质序列比对

**Future Direction**

增加病毒序列覆盖率

描述标准化

开发新的特定于病毒的序列注释模块

扩大用户范围

改进数据管理、参考序列分配和资源可用性