Infrastructure clouds, microbial genomics, and the Cloud Virtual Resource project (CloVR)

Sam Angiuoli

sangiuoli@som.umaryland.edu

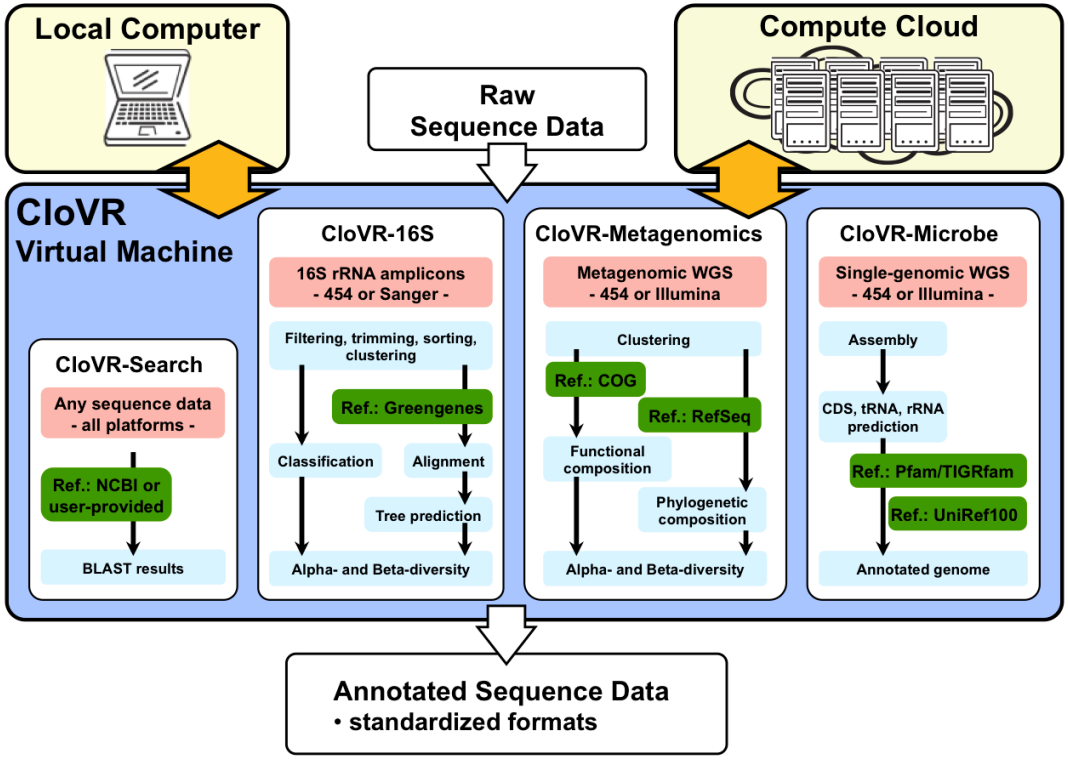
Institute for Genome Sciences
University of Maryland School of Medicine

Our community

- Microbial genomics
  - Biologists, MDs, bioinformaticians
  - Many backgrounds; increasingly inter-disciplinary
- Democratization of sequencing
  - Data acquisition is less of a bottleneck
- Many datasets
  - Including many small to medium in size
- Diverse analysis needs
  - Often exploratory
  - Evolving protocols

CloVR

Our project



- Integrate analysis software and automated pipelines into a portable VM
- VM runs on a local PC and can (optionally) access the Cloud seamlessly on-demand

Funding NSF NIH NHGRI

What drew our development team to infrastructure clouds?

- Reliability of Amazon EC2
  - Things just worked
- Flexibility
  - Pipelines and protocols change frequently
  - Customizable VMs
- Independence
  - Root access
  - On-demand

How does our community benefit?

- Increased throughput
  - Many small to medium scale needs but more than a desktop can handle
- Easy to use
  - Zero software install
  - CloVR VM is prebuilt, ready to go
  - No need for building, managing local cluster
  - In CloVR, cloud utilization is built-in and seamless
- Improved control
  - CloVR VM runs on the desktop as a stand-alone application
  - Cloud used for temporary processing only; data is downloaded back to the PC
  - DIY nature is appealing to many groups
  - An alternative to uploading data to a third-party web server for analysis

What are the challenges in using clouds?

- Poor reliability
  - Must just work, end-users cannot debug
  - Cloud and software must be very robust

  Ongoing need for testing and good software engineering
- Poor availability
  - What exists outside of commercial space is small

  Need more machine types
- Limited portability
  - Despite APIs, not simply write once, run anywhere

  Need for more standardization
- Authentication

  Need for simplified sign-on
- Big data
  - Internet transfers can be prohibitive