Genomics Assembly and Analysis Training Module

Genomics Assembly and Analysis Training Module

Objectives

- Introduce the process of genome assembly and analysis, using Mate et. al. “Molecular Evidence of Sexual Transmission of Ebola Virus” as an example analysis.
- Starting with raw sequencing data, understand the steps for assembling a complete genome sequence.
- Compare multiple genome sequences.
- Use the output of the above analyses to draw conclusions about the biology of the samples.

Genomics Assembly and Analysis Training Module

Outline

- Brief review of Mate et. al. and Next Generation Sequencing.
- Step by step instructions for analyzing sequencing data using a Jupyter notebook.
- Glossary, FAQ, and complete breakdowns of all computational steps are provided at the end of this presentation.

ORIGINAL ARTICLE BRIEF REPORT

Molecular Evidence of Sexual Transmission of Ebola Virus

Suzanne E. Mate, Ph.D., Jeffrey R. Kugelman, Ph.D., Tolbert G. Nyenswah, L.L.B., M.P.H., Jason T. Ladner, Ph.D., Michael R. Wiley, Ph.D., Thierry Cordier-Lassalle, M.B.A., D.E.S.S., Athalia Christie, M.I.A., Gary P. Schroth, Ph.D., Stephen M. Gross, Ph.D., Gloria J. Davies-Wayne, R.N., M.P.H., Shivam A. Shinde, M.B., B.S., Ratnesh Murugan, M.B., B.S., et al.

- In Liberia, the partner of an Ebola survivor became sick.
- Did the partner contract Ebola through sexual transmission? Or through some other means?
- How can we tell?
- **These questions can be answered by sequencing.**

ORIGINAL ARTICLE BRIEF REPORT

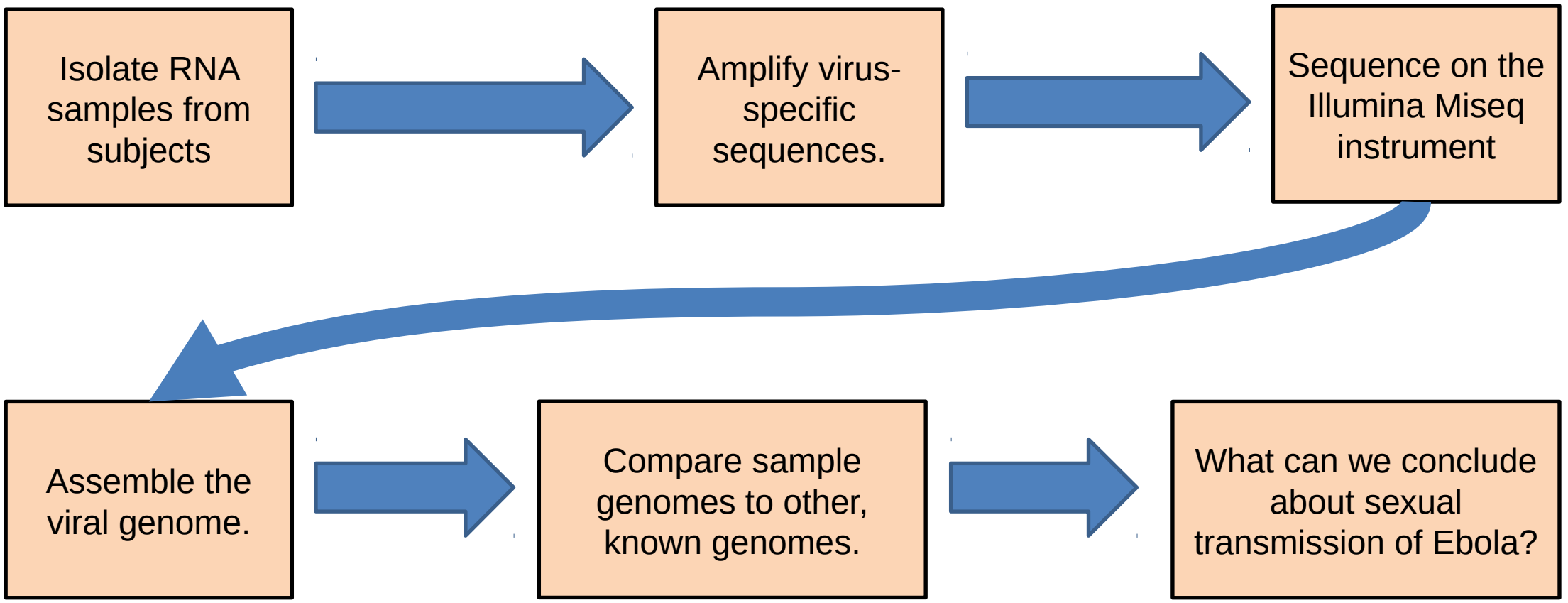
Molecular Evidence of Sexual Transmission of Ebola Virus

Suzanne E. Mate, Ph.D., Jeffrey R. Kugelman, Ph.D., Tolbert G. Nyenswah, L.L.B., M.P.H., Jason T. Ladner, Ph.D., Michael R. Wiley, Ph.D., Thierry Cordier-Lassalle, M.B.A., D.E.S.S., Athalia Christie, M.I.A., Gary P. Schroth, Ph.D., Stephen M. Gross, Ph.D., Gloria J. Davies-Wayne, R.N., M.P.H., Shivam A. Shinde, M.B., B.S., Ratnesh Murugan, M.B., B.S., et al.

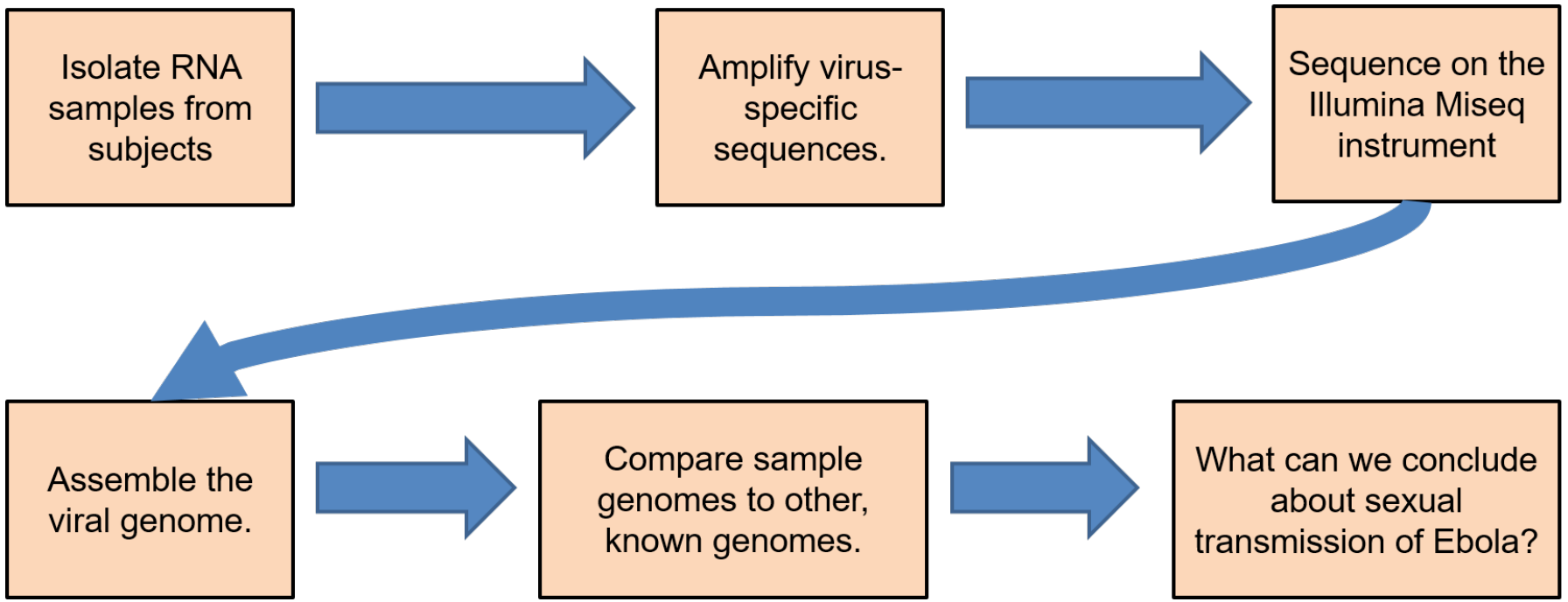
To answer these questions, we can:

- Isolate virus from the survivor and their partner.
- Sequence the virus to discover the complete, accurate genome of each sample.
- Compare these sequences to each other and to other virus samples from this outbreak.
- Is the partner sample more similar to the survivor sequence? Or to the other samples from this outbreak?

We can answer these questions by following this outline:

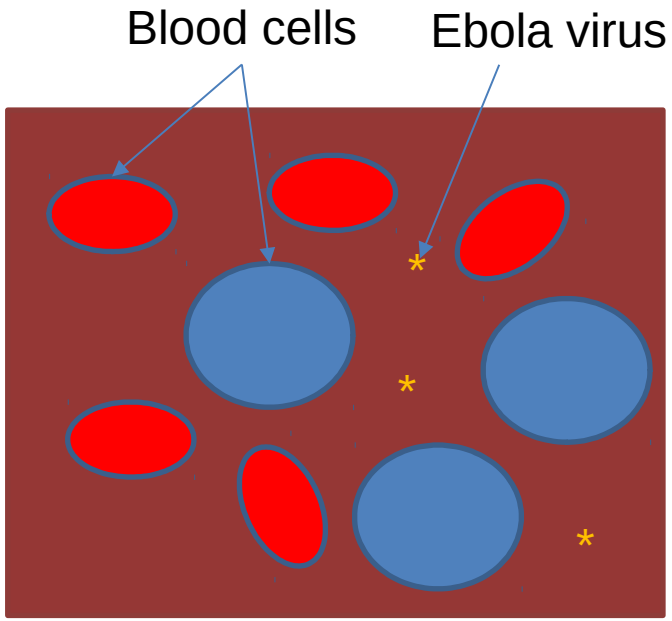


- This module focuses on the final three steps: analyzing the raw sequencing data.
- But it is important to first understand how the raw data is generated.

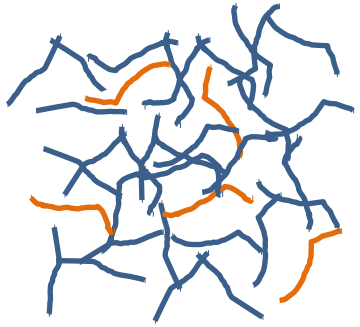


Isolate RNA samples from subjects

Collect samples



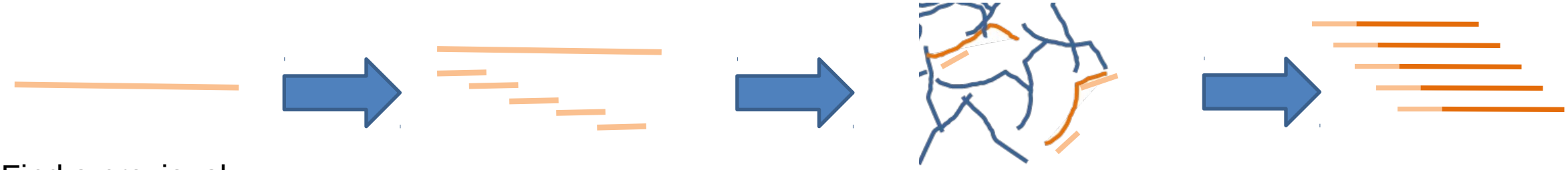
The samples contain mostly host cells.



Extract RNA, convert to DNA. It will be fragmented, and mostly from the host.

Amplify virus-specific sequences.

- Since most of the DNA will be from the host, we need to increase the proportion of viral DNA.



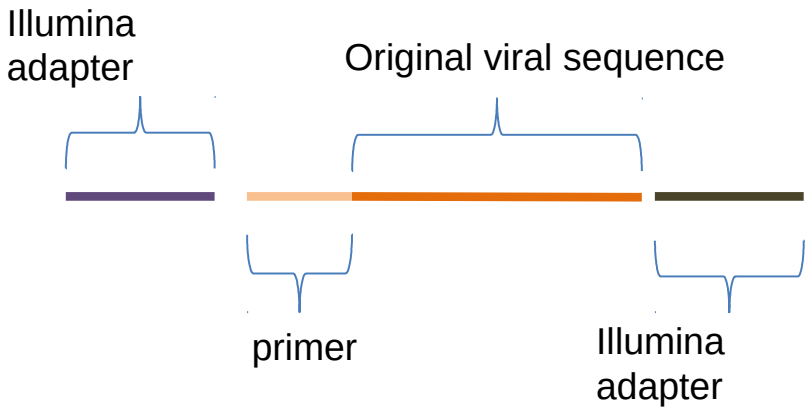
Find a previously sequenced Ebola genome (a “reference” sequence).

Make primers-short fragments of DNA-that match the reference genome.

The primers bind to the similar Ebola virus sequence found in our sample.

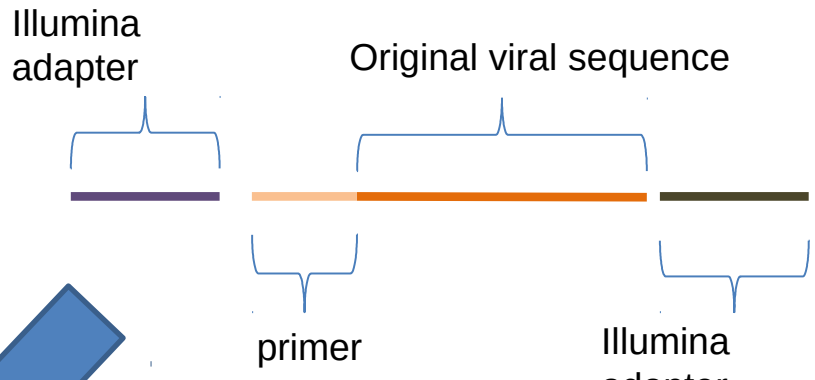
Primer-bound DNA is amplified, producing lots of viral DNA fragments.

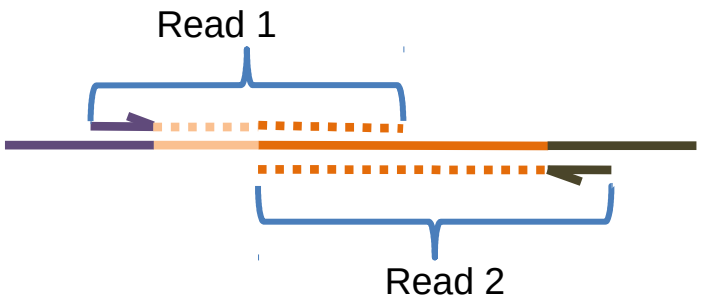
Sequence on the Illumina Miseq instrument



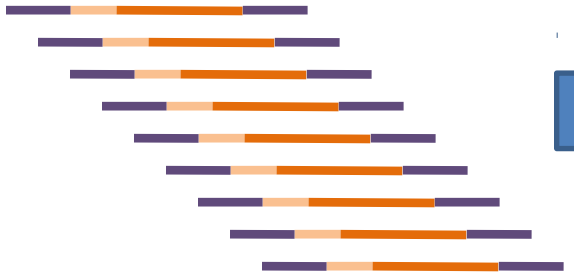
- **Add adapters to both ends of the DNA fragment to be sequenced.**
- **These are DNA sequences necessary for sequencing on the Illumina Miseq.**
- **The pool of DNA to be sequenced is known as a “library.”**

Sequence on the Illumina Miseq instrument





Each piece of DNA is sequenced twice, starting from either end (Paired end sequencing). The products of a sequencing run are called “reads.” The sequence and quality information are stored in .fastq files.

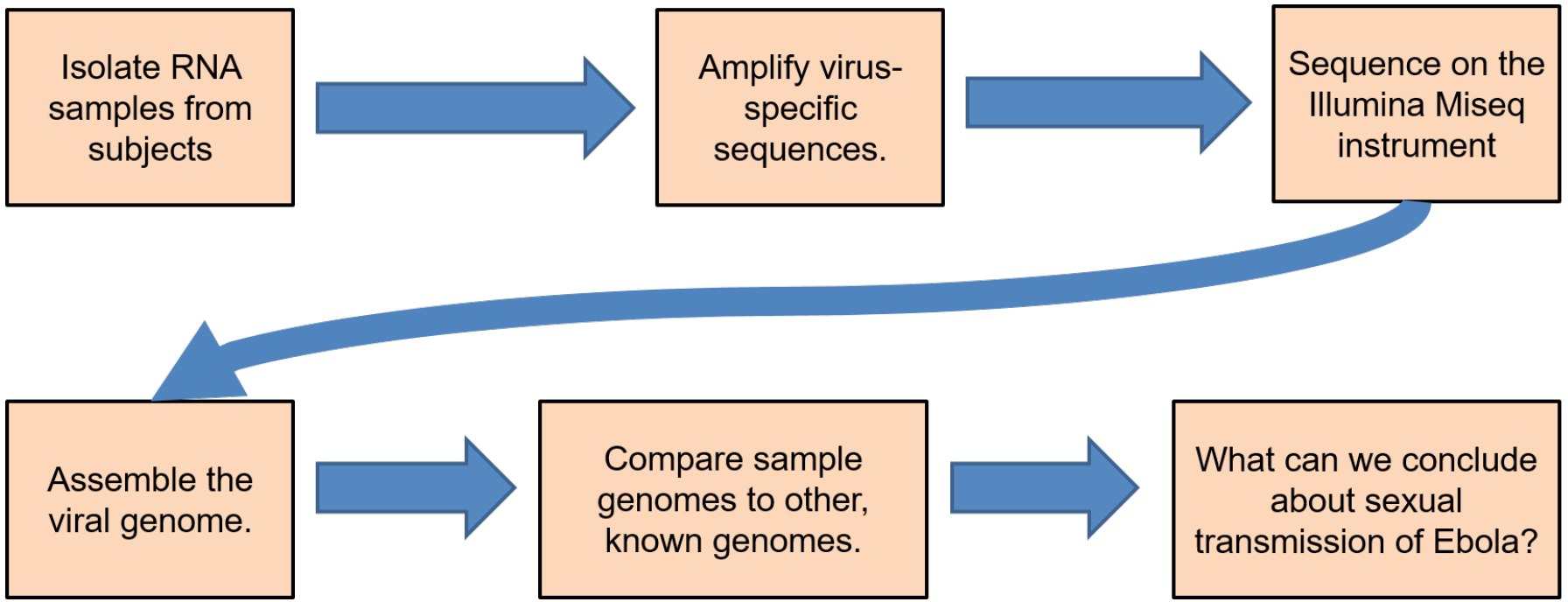


Library-DNA fragments ready to be sequenced

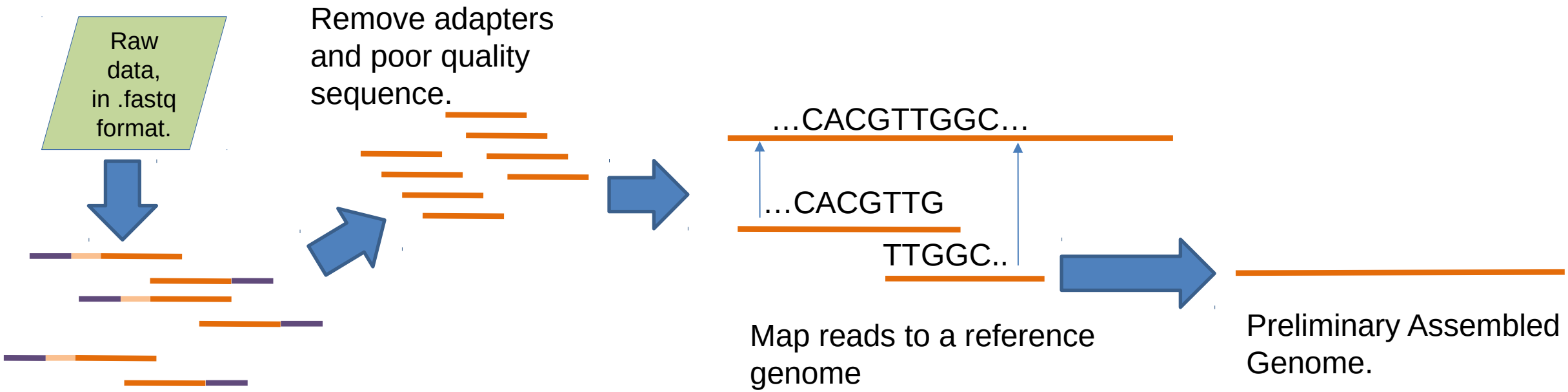
Illumina MiSeq

Raw data, in .fastq format.

Now we can assemble the raw data produced by the Miseq.

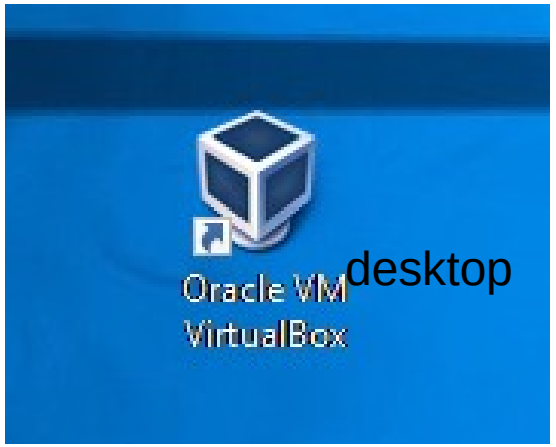


Assembly: Broad Overview of the Computational Steps

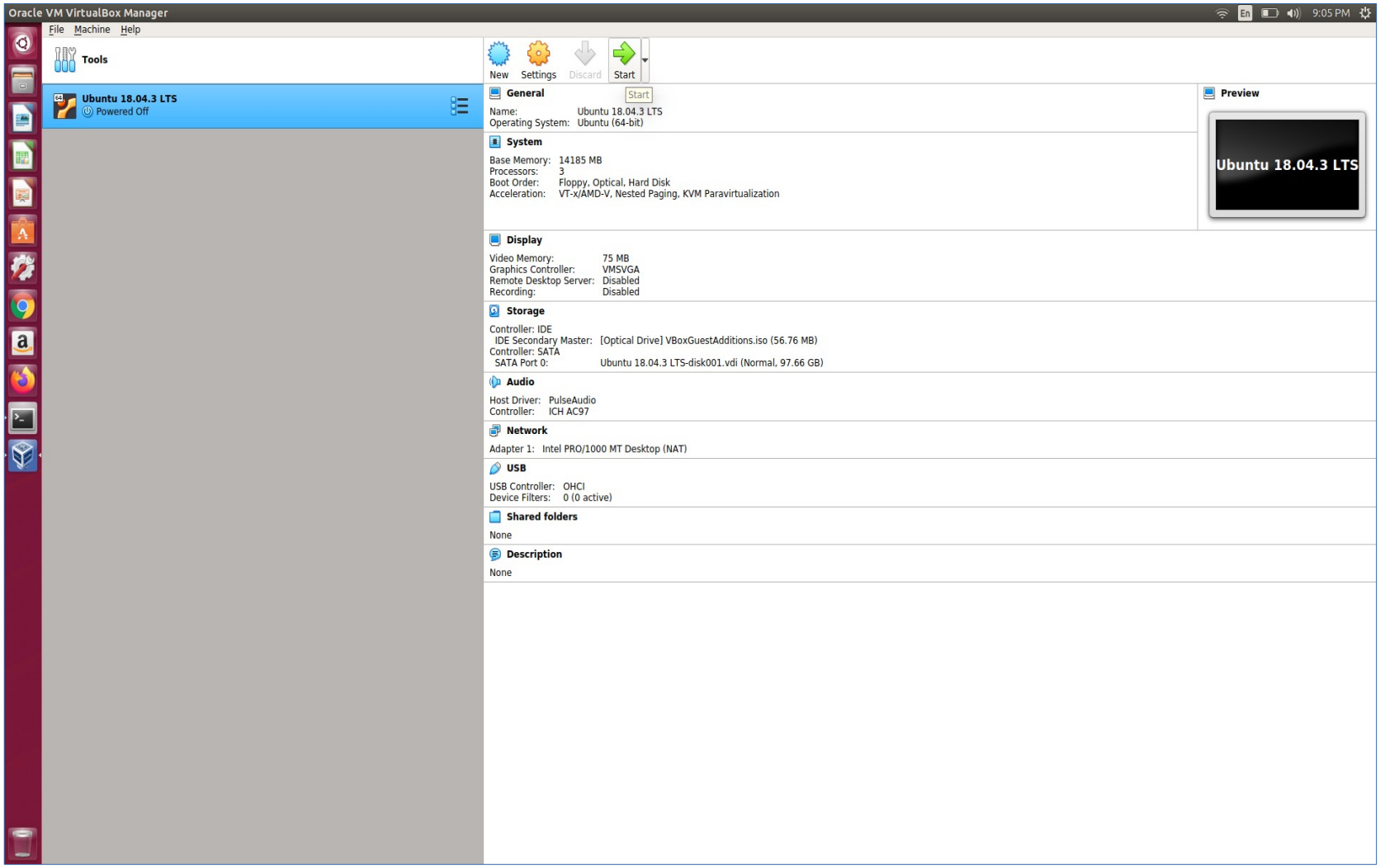


Raw data consists of sequences containing fragments of the Ebola genome. Ultimately, we need to take these fragments and assemble them into the complete genome.

Prepare for assembly: Open a virtual machine, which contains all of the data and programs you need to complete the module.



Click on the Oracle VM VirtualBox icon



Open the Jupyter notebook.

- Run through each step of the Jupyter notebook, examining any slides embedded in the step.
- Steps that advance the pipeline will be accompanied by an explanation slide, which is detailed next.