A Virtual Machine to teach NGS data analysis

Andreas Gisel
CNR - ITB
Bari, Italy
The Virtual Machine

- A virtual machine is a tightly isolated software container that can run its own operating systems and applications as if it were a physical computer.
- A virtual machine behaves exactly like a physical computer and contains its own virtual (i.e., software-based) CPU, RAM hard disk and network interface card (NIC).
The Virtual Machine

A virtual machine is a tightly isolated software container that can run its own operating systems and applications as if it were a physical computer.

A virtual machine behaves exactly like a physical computer and contains its own virtual (i.e., software-based) CPU, RAM, hard disk, and network interface card (NIC).

**Benefits**

In general, VMware virtual machines possess four key characteristics that benefit the user:

- **Compatibility**: Virtual machines are compatible with all standard x86 computer
- **Isolation**: Virtual machines are isolated from each other as if physically separated
- **Encapsulation**: Virtual machines encapsulate a complete computing environment
- **Hardware independency**: Virtual machines run independently of the underlying hardware
The Virtual Machine
The Virtual Machine
The Virtual Machine

“Bioinformatics without borders”

www.embnet.org

Bioinformatics Virtual Machine
by
Andreas Gisel
Italian EMBnet Node, Bari Italy
and
Erik Bongcam-Rudloff
Swedish EMBnet Node, Uppsala Sweden

http://journal.embnet.org/
The Virtual Machine

apache2
MySQL
Bioperl
EMBOSS
Vienna RNA Package
blast2
ncbi-blast
R
BioConductor
CLUSTAL
MUSCLE
T-COFFEE
The Virtual Machine

apache2
MySQL
Bioperl
EMBOSS
Vienna RNA Package
blast2
ncbi-blast
R
BioConductor
CLUSTAL
MUSCLE
T-COFFEE

velvet
SOAPdenovo127mer
SOAPdenovo-Trans
bowtie
bwa
gmap
macs
tophat
cufflinks
fastx-toolkit
FastQC
samtools
The Virtual Machine

apache2
MySQL
Bioperl
EMBOSS
Vienna RNA Package
blast2
ncbi-blast
R
BioConductor
CLUSTAL
MUSCLE
T-COFFEE

velvet
SOAPdenovo127mer
SOAPdenovo-Trans
bowtie
bwa
gmap
macs
tophat
cufflinks
fastx-toolkit
FastQC
samtools

tablet
Apollo
Artemis
Artemis comparison tools
Galaxy
GbrowseV2
Application Level 1

The Virtual Machine gets installed on each participant's computer (desktop or notebook)

Every participant independent what OS he used is able to install it and the VM can be set to the physical properties or each computer.

BUT every Participant has exactly the same working environment!

ONLY the size of the data set has to be adapted!

The participant can take home HIS NGS data analysis platform!
Application Level 1

The Virtual Machine gets installed on each participants computer (desktop or notebook)

Personal computers:
• 1-2 CPU
• 2-4GB RAM

With ~30 participants we were analyzing NGS data; single-end 100k reads:
• quality control
• assembly
• mapping
Application Level 2

The Virtual Machine gets installed on a small portable server (i.e. MAC mini)

With little effort convert a small server into a bioinformatics platform
The mini server in the local network can be accessed via ssh of web services via http (Galaxy, Gbrowse etc)

BUT every Participant has exactly the same working environment!

ONLY the size of the data set has to be adapted!

The participant CANNOT take home the NGS data analysis platform!
Application Level2

The Virtual Machine gets installed on a small portable server (i.e. MAC mini)

Mac mini:
- 4CPU
- 16GB RAM

With 12 participants we were analyzing NGS data; paired-end 2.5m reads:
- assembly
- mapping
- miRNA search
Application Level3

The Virtual Machine gets installed on a computing farm

With little effort convert a larger server into a bioinformatics platform.
Via the local network one can accessed via ssh or web services via http (Galaxy, Gbrowse etc).

BUT every Participant has exactly the same working environment!

According on the resources available any kind of NGS data can be analyzed.

The participant CANNOT take home the NGS data analysis platform!
Application Level 3

The Virtual Machine gets installed on a computing farm

We were running two tutorials about ncRNA data analysis with each around 40 participants.

For these tutorials we prepared 4 independent VM with each:
• 16 CPU
• 64GB shared RAM

We were analyzing NGS data; single-end 2m reads:
• miRNA search with several mapping steps
Last Improvement
Local website with educational material
Last Improvement

Local website with educational material
Last Improvement

Local website with educational material

Steps of the workflow:
1. create work directory and load data
2. process the raw data for the mapping and miRNA search
3. calculate the fold change between the different biological sample
4. run the miRNA search using mirDeep2 (Friedländer et al, 2008)
5. run a mapping with BWA (Li and Durbin, 2008) on the mouse genome
6. combine all data
7. extract the data for the database
8. load the database (will be done for this tutorial)
9. access the graphical interface of the database

Detailed steps to follow the tutorial:
1. move to the working device
   - cd studentX ($X is your number)
2. we create a working directory in our working 'home'
   - mkdir work
3. enter into the working directory
   - cd work
4. copy the data from centralized directory (/Tutorial) to the working directory
   - cp /Tutorial/SRR3613* .
5. copy some software from centralized directory (/Tutorial) to the working directory
   - cp /Tutorial/fold.tgz .
6. unzip the the software
   - tar xzvf fold.tgz
Last Improvement
Local website with educational material
Future Improvement

- smaller 64bit VM for download
- more educational material
- more test data
- dedicated web site for download
Thanks!!

Andreas Gisel

andreas.gisel@ba.itb.cnr.it
www.ba.itb.cnr.it/gisel